Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in
Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Analisi di un dataset sulle performance dei piloti del Motomondiale sull'Autodromo del Mugello dal 2016 ad oggi

Relatore: Prof. Livio Finos
Dipartimento di Scienze Statistiche

Laureando: Alessandro Cristanelli
Matricola: N° 2072768

Anno Accademico 2024/2025

Indice

1 Introduzione **3**
- 1.1 Cos'è il Motomondiale? . . . 3
- 1.2 Le classi del Motomondiale . . . 3
- 1.3 Le piste del Motomondiale . . . 5
- 1.4 Come si organizza un weekend di gare? . . . 6
- 1.5 Su cosa ci focalizzeremo? . . . 7

2 EDA **10**
- 2.1 Preprocessing . . . 10
- 2.2 Variabili quantitative . . . 11
- 2.3 Variabili qualitative . . . 14
- 2.4 Rapporto tra variabili qualitative e quantitative . . . 16

3 Riduzione della dimensionalità del dataset con variabili quantitative **28**
- 3.1 PCA (Analisi delle componenti pricipali) . . . 28
- 3.2 PCA: applicazione al dataset con variabili numeriche . . . 28
- 3.3 Factor analysis . . . 31
- 3.4 Applicazione della factor analysis al dataset con variabili numeriche . . . 32

4 CCA **36**
- 4.1 Cos'è la CCA? . . . 36
- 4.2 Applicazione al dataset con variabili numeriche . . . 37

5 Cluster Analysis **40**
- 5.1 Cos'è il processo di clustering? . . . 40
- 5.2 Cluster analysis con metodi gerarchici . . . 41
- 5.3 Cluster analysis con il metodo delle k-medie . . . 63

6 Previsione del tempo sul giro in qualifica **67**
- 6.1 Contesto . . . 67
- 6.2 Modello saturo . . . 69
- 6.3 Modello di regressione lineare stepwise backward . . . 71
- 6.4 Modello di regressione lineare stepwise forward . . . 73
- 6.5 Modello di regressione lineare stepwise both . . . 74
- 6.6 Modello di regressione con metodo di regolarizzazione Ridge . . . 76
- 6.7 Modello di regressione con metodo di regolarizzazione Lasso . . . 78
- 6.8 Conclusione: qual è il modello migliore? . . . 79

7 Conclusioni **81**

8 Bibliografia e sitografia **83**

1 Introduzione

1.1 Cos'è il Motomondiale?

Il Motomondiale è la massima espressione del motociclismo sportivo agonistico presente su questo pianeta.
E' un campionato che si disputa ogni anno a partire dal 1949 e consiste nello svoglimento di 22 Gran Premi (ad oggi) in giro per tutto il mondo, tra America, Oceania, Europa e Asia, dove i migliori piloti motociclistici del mondo si sfidano a suon di sorpassi e giri veloci.
Come affermato anche da questa fonte[1], le moto su cui i piloti corrono sono prototipi costruiti dalle migliori case costruttrici di moto appositamente per le corse, quindi non sono omologate per la circolazione su strada.
Come riportato anche da questo articolo[2], tutte le moto corrono con pneumatici di tipo slick, ovvero con superficie liscia, quando la gara è dichiarata "con asfalto asciutto" dai direttori di corsa e da bagnato, con superficie piena di scanalature, se la pista è dichiarata "con asfalto bagnato", sempre dai direttori di gara. Fino al 2017 erano presenti anche le opzioni di gomma intermedia, per asfalto non troppo bagnato, tuttavia esse non erano gradite dai piloti. Le gomme slick possono essere di 3 mescole, sia per i pneumatici anteriori che posteriori: morbide (soft, le mescole che permettono di avere più aderenza con l'asfalto da subito ma si consumano più in fretta), dure (hard, le mescole che funzionano all'esatto opposto delle soft) e medie (medium, il compromesso tra le altre 2 mescole). Le gomme da bagnato, invece, possono essere di solo 2 mescole, ovvero soft e media.

1.2 Le classi del Motomondiale

Ci sono 4 classi di competizione, basate su cilindrata e tipo di motore[3]:

- **MotoE**, una categoria monomarca (ad oggi Ducati) di motociclette elettriche nata nel 2019, che corre con pneumatici prodotti dal costruttore Michielin e che non si disputerà più fino a data da destinarsi a partire dal 2026[4] (e che quindi non citeremo più minimamente nel corso dell'indagine);

- **Moto3**, categoria entry level per i giovani agli inizi della carriera agonistica mondiale, dove le moto sono tutte con motori a 4 tempi monocilindrici da 250 cc (cc = centimetri cubici) di cilindrata prodotte principalmente dalle case costruttrici Honda e KTM (che corre anche con il marchio CFMoto e che fino a poco tempo fa correva anche con i sottomarchi GASGAS e Husqvarna).
  La categoria sostiuisce dal 2012 la classe 125, dove le moto erano tutte con motori 2 tempi da, appunto, 125 centimetri cubici e c'erano anche moto prodotte da costruttori come Aprilia e Derbi.
  I piloti sono per la maggior parte ragazzini che vanno dai 17 ai 20 anni di etè e la classe in cui competono è l'unica categoria che, oltre ad avere un'età minima di accesso di 17 anni, ha dal 2022 anche un'età massima di accesso, ovvero 28 anni.
  Il peso minimo combinato tra moto e pilota deve essere di 148 kg e dal 2024 le moto corrono con pneumatici forniti dal brand Pirelli (fino al 2023 erano forniti dal brand Dunlop)[5];

- **Moto2**, categoria intermedia per piloti un po' più esperti, con un'età media pressochè di 22-23 anni dove le moto sono tutte con motore a 4 tempi, dal 2019 derivato dalla motocicletta Triumph Street Triple, tricilindrici con 765 cc di cilindrata.

[1] Wikipedia, *Motomondiale*, https://it.wikipedia.org/wiki/Motomondiale
[2] GripMoto, *Quali gomme si usano MotoGP™?*, https://www.gripmoto.it/consigli-utili/news/quali-gomme-si-usano-motogp.html
[3] Wikipedia, *Motomondiale*, https://it.wikipedia.org/wiki/Motomondiale#Le_classi
[4] P300, *MotoE — Il campionato elettrico si ferma alla fine del 2025*, https://www.p300.it/motoe-il-campionato-elettrico-si-ferma-alla-fine-del-2025/
[5] FormulaPassion, *Moto2-Moto3 / Ufficiale: Pirelli è il nuovo fornitore*, https://www.formulapassion.it/moto/moto2-moto3-ufficiale-pirelli-e-il-nuovo-fornitore/

Generalmente è considerata la categoria "successiva" alla Moto3 per un pilota, difatti di solito i piloti che in una stagione si giocano il campionato del mondo Moto3, o comunque i migliori 5 a fine campionato, tendenzialmente corrono la stagione successiva nel campionato del mondo Moto2. Inoltre è considerata la categoria propedeutica alla classe regina, ovvero la MotoGP, dove tendenzialmente corrono i piloti che hanno vinto o per lo meno hanno ottenuto risultati importanti anche in Moto2.
A sfidarsi, oltre che ai piloti, sono soprattutto anche i telaisti, come Kalex, Boscoscuro e Forward (fino al 2023 MV Agusta, la quale ha deciso di tirarsi indietro con i finanziamenti del progetto).
La categoria, che dal 2012 al 2018 ha corso con moto con motore a 4 tempi quadricilindrici Honda da 600 cc di cilindrata, sostituisce dal 2012 la classe 250 dove i piloti correvano su moto con motori a 2 tempi da 250 cc di cilindrata e non prevedeva un monomarca motoristico: i telaisti erano anche motoristi, motivo per il quale correvano moto prodotte da costruttori come Honda, Aprilia, Gilera e KTM.
Come per la Moto3, fino al 2023 i pneumatici erano forniti dal brand Dunlop, dal 2024 sono forniti da Pirelli;

- **MotoGP**, la classe regina, quella dove a sfidarsi sono i piloti più esperti e forti mentalmente e fisicamente. I piloti corrono ad oggi su moto prodotte da 5 costruttori diversi: Aprilia, Ducati, Honda, KTM e Yamaha. Fino al 2022 competeva anche un sesto costruttore, Suzuki, ritiratosi "a fronte della necessità di riallocare risorse su altre iniziative per la sostenibilità"[6].
  Le moto in sè sono fornite di motori a 4 tempi quadricilindrici con 1000 cc (990 dal 2002 al 2006, 800 dal 2007 al 2011 e 850 dal 2027) di cilindrata, mentre i pneumatici sono forniti da Michielin (in sostituzione di Bridgestone dal 2016 e sostituiti da Pirelli dal 2027).
  E' la categoria, dal 2002, sostituiva della defunta classe 500, dove i piloti si sfidavano su moto con motori a 2 tempi con, appunto, 500 cc di cilindrata e non c'era un'unico fornitore di pneumatici ma molteplici (Michielin, Dunlop, Bridgestone, ecc.).
  Nel corso del tempo le motociclette hanno avuto un enorme evoluzione tecnologica: dall'utilizzo di appendici aerodinamiche per migliorarne la velocità in curva e in rettilineo[7] all'introduzione di abbassatori dell'altezza di guida del pilota che, se attivati all'uscita di una curva che porta ad un rettilineo, riducono di molto la tendenza della moto ad impennarsi e quindi ne migliorano accelerazione e stabilità in rettilineo; per non parlare dei miglioramentei elettronici, effettuati anche nelle altre classi ma soprattutto in questa, quali controllo di trazione, sistemi anti-impennata, controlli di stabilità (dal Gran Premio d'Austria del 2025), rilevatori di pressione delle gomme, ecc. Tutto questo dal 2016 è gestito per ogni moto da una centralina, detta ECU, fornita esclusivamente dal 2016 da Magneti Marelli (fino al 2015 ogni costruttore produceva la propria).
  La MotoGP, così come la defunta classe 500, è la categoria dove più si sono create le leggende per i tifosi di tutto il mondo del motociclismo: piloti come Valentino Rossi, Marc Marquez, Giacomo Agostini, Mick Doohan, Mike Hailwood, Casey Stoner e Jorge Lorenzo sono piloti che hanno più lasciato il segno, nel bene ma delle volte purtroppo anche nel male, non solo della disciplina ma proprio dello sport in generale. Sono piloti che la gente ha amato e odiato, oppure ama o odia tuttora, ma che non si può negare abbiano portato allo sport un'identità, facendo cose sempre più nuove sulle motociclette, cose che persino alcuni di essi si chiedano come esse possano essere state fatte.
  E' l'unica categoria dove i costruttori hanno ognuno la propria squadra ufficiale, mentre alle altre squadre vengono fornite moto dai costruttori attraverso accordi reciproci principalmente monetari, ma non solo (ad esempio la casa costruttrice di moto Ducati fornisce ad altri team diversi dal proprio team ufficiale delle moto in leasing, a quanto dice il CEO del brand Claudio Domenicali[8]). A queste squadre, spesso, vengono fornite moto di uno o due anni più vecchie,

[6] MotoGrandPrix, *MotoGP — Ufficiale, la Suzuki conferma il ritiro a fine 2022*, `https://motograndprix.motorionline.com/motogp-ufficiale-la-suzuki-conferma-il-ritiro-a-fine-2022/`

[7] InSella, *L'aerodinamica della MotoGP tra alettoni e carene: come funziona*, `https://www.insella.it/da-sapere/tecnica/laerodinamica-della-motogp-alettoni-carene-come-funziona-46067`

[8] Chapeau_Project, *Ecco QUANTO COSTA Creare Una Moto Da MOTO GP*, `https://www.youtube.com/shorts/WWR4qbLOjtc`

tranne alcuni casi sempre più recenti dove vengono date moto completamente nuove, ed esse vengono chiamate **squadre clienti** o **team satelliti**.

1.3 Le piste del Motomondiale

Le piste dove corre il Motomondiale sono piste omologate direttamente dalla **Fédération Internationale de Motocyclisme**[9], federazione organizzatrice dei campionati motociclistici mondiali di maggior rilievo quali, oltre che al Motomondiale, Superbike (comprese categorie propedeutiche come Supersport, dove le moto sono derivate dalla produzione serie) e Mondiale Endurance, creato appositamente per le gare di durata (gare di molteplici ore per intenderci dove a sfidarsi su moto derivate dalla produzione di serie sono squadre e non piloti individuali). Dal 1949 ad oggi sono stati corsi Gran Premi del Motomondiale su più di 70 circuiti diversi attorno al Mondo[10], dai più iconici come Assen in Olanda o l'Isola di Man in Gran Bretagna, ai più recenti, come Buriram in Thailanda o Balaton Park in Ungheria (novità del 2025). Se c'è però un circuito su cui l'atmosfera è completamente diversa da tutte le altre è quello dove si corre il Gran Premio d'Italia, ovvero il **Mugello, in Toscana**.
Ad ogni Gran Premio sono presenti decine se non centinaia di migliaia di tifosi durante tutto il week end in cui esso si disputa (dopo ci arriveremo), a seconda anche delle culture motociclistiche del posto. Essendo situato poco sotto la Motor Valley della Romagna, il circuito del Mugello è quindi il posto quasi perfetto dove poter svolgere il Gran Premio d'Italia: durante tutto il fine settimana assistono tifosi da tutte le parti d'Italia e non solo, basti pensare alle tifoserie dei piloti non italiani. Inoltre è tra gli unici Gran Premi del calendario che vanta sia tifoserie di piloti che tifoserie di costruttori; questo perché l'Italia è uno dei pochi paesi che può vantare sia di avere piloti italiani forti, come Rossi, Biaggi o più recentemente Bagnaia, Bezzecchi e Bastianini, sia di avere i costruttori produttori delle moto probabilmente adesso più performanti di tutta la concorrenza, ovvero Ducati e Aprilia. Addirittura pensate che Ducati ha degli spalti situati sulla quart'ultima curva del circuito che puntualmente durante il fine settimana si riempiono di tifosi vestiti di colore rosso (colore simbolico di Ducati) a tifare per le loro moto e i loro piloti[11].

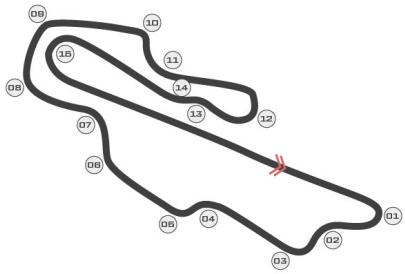


Figura 1: Mappa del circuito del mugello: i tifosi della Ducati si posizionano sulla curva numero 12

Ma un Gran Premio non si svolge in un solo giorno, come appena accennato. O meglio il Gran Premio in sè si disputa in un solo giorno, la domenica, ma prima di esso vengono disputate dai piloti sessioni di prova e qualifica distribuite nei 2 giorni precedenti.

[9] FIM, *FIM STANDARDS FOR CIRCUITS 2024*, https://www.fim-moto.com/fileadmin/user_upload/Documents/2024/2024_Circuit_Standards.pdf?t=1707410999

[10] Wikipedia, *Categoria:Circuiti del motomondiale*, https://it.wikipedia.org/wiki/Categoria:Circuiti_del_motomondiale

[11] Ducati, *Tribuna Ducati MotoGP Mugello 2025*, https://www.ducati.com/it/it/experience/tribuna-ducati/motogp-mugello

1.4 Come si organizza un weekend di gare?

Ecco l'organizzazione del weekend a partire dal 2023, anno di introduzione delle Sprint Races, gare di breve durata per la sola classe MotoGP[12]:

- **il venerdì** vengono disputate 2 sessioni di prove libere, una al mattino e una al pomeriggio, per ciascuna categoria. Entrambe le sessioni della Moto3 dureranno 35 minuti ciascuna, quelle della Moto2 40 minuti ciascuna e quelle della MotoGP rispettivamente 45 e 60 minuti.
  Sono state rinominate alquanto spesso negli ultimi 2 anni per via dell'introduzione delle Sprint Races per la sola classe MotoGP dal 2023: fino a quella stagione la durata era di 45 minuti per entrambe le sessioni della MotoGP e di 40 minuti per entrambe le sessioni sia di Moto3 che di Moto2. Per comodità noi le chiameremo **P1 e P2**;

- **il sabato** viene disputata una terza sessione di prove libere al mattino per ogni categoria dalla durata di 30 minuti ciascuna, che noi chiameremo per comodità **P3**. Dopodiché la MotoGP disputerà 2 sessioni di qualifica da 15 minuti l'una: la prima sessione, detta **Q1**, viene disputata da coloro che non si sono classificati nelle prime 10 posizioni della P2. I primi 2 classificati al termine della Q1 potranno accedere alla seconda sessione di qualifica, detta **Q2** dove invece saranno presenti anche i primi 10 classificati della P2 e assieme ai quali si giocheranno la pole position (ovvero la partenza dalla prima posizione in griglia) sia per la Sprint Race sia per il Gran Premio. Gli altri piloti saranno invece classificati dalla 13esima posizione in giù sulla griglia di partenza, sia della Sprint Race che del Gran Premio. Anche Moto3 e Moto2 disputeranno successivamente le proprie sessioni di qualifica una dopo l'altra: la struttura è la stessa della MotoGP, via che nelle Q2, per ogni categoria, saranno presenti non i primi 10 piloti delle P2 ma i primi 14 e ad aggiungersi a loro non saranno i primi 2 classificati delle Q1 ma i primi 4.
  Dopo le sessioni di qualifica, nel primo pomeriggio, si svolgerà la Sprint Race della MotoGP, una gara della durata di circa metà Gran Premio (sia in termine di giri che quindi in termine di tempo), al termine della quale verranno assegnati dei punti validi per la classifica mondiale ai primi 9 piloti al traguardo nella seguente maniera[13]:

Posizione	Punti
1°	12
2°	9
3°	7
4°	6
5°	5
6°	4
7°	3
8°	2
9°	1

- **la domenica** è giorno di gara: si parte con una sessione di riscaldamento per la sola MotoGP, detta Warm-Up. Questo tipo di sessioni venivano disputate anche da Moto2 e Moto3, una dopo l'altra, ma in favore di una parata di tutti i piloti della MotoGP prima del Gran Premio della Moto3 esse sono state rimosse dal 2023.
  La domenica si conclude tutto il weekend con i piatti forti: vengono corsi in successione, con una pausa di riti e festeggiamenti sul podio tra la fine di uno e l'inizio di quello dopo, nell'ordine i Gran Premi di Moto3, Moto2 e MotoGP.

[12] Wikipedia, *Motomondiale*, `https://it.wikipedia.org/wiki/Motomondiale#Svolgimento_dei_Gran_Premi`
[13] Wikipedia, *Monomondiale*, `https://it.wikipedia.org/wiki/Motomondiale#Punteggio_in_vigore`

Per ogni classe il sistema di punteggio, basato sull'ordine di arrivo dei piloti, è il seguente:

Posizione	Punti
1°	25
2°	20
3°	16
4°	13
5°	11
6°	10
7°	9
8°	8
9°	7
10°	6
11°	5
12°	4
13°	3
14°	2
15°	1
≥ 16	0

1.5 Su cosa ci focalizzeremo?

Per la nostra analisi noi ci focalizzeremo sulle prestazioni dei piloti della sola MotoGP nel corso delle prove libere di venerdì e sabato (P1, P2 e P3) e nel corso delle sessioni di qualifica (Q1 e Q2) sul circuito del Mugello, in occasione dei Gran Premi d'Italia disputatisi dal 2016 ad oggi.
Ho organizzato così un dataset da più di 50 variabili, con dati raccolti dal sito ufficiale del Motomondiale nelle sezione Results[14], dove le unità statistiche non sono altro che i migliori giri che ogni pilota partecipante realizza in ogni sessione di qualifica.
Le variabili rilevate sono:

- **Year**, variabile quantitativa discreta relativa all'anno di rilevazione delle performance, con dominio pari a {2016;2017;...;2025};
- **Nat**, fattore con 13 livelli relativo alla nazionalità del pilota di cui sono state rilevate le performance;
- **Exp**, variabile quantitativa continua relativa all'esperienza, misurata in anni, che un pilota ha con una MotoGP su quel circuito al momento della rilevazione (ad esempio se Exp=7 vuol dire che in quel momento il pilota è al settimo anno che corre con una MotoGP sul circuito del Mugello);
- **Bike**, fattore a 6 livelli con livelli pari alla marca della moto con cui il determinato pilota sta correndo al momento delle rilevazioni. L'insieme delle modalità modalità è {Aprilia, Ducati, Honda, KTM, Suzuki, Yamaha};
- **Bike_year**, variabile quantitativa che indica l'anno della moto e che useremo per distinguere anche i piloti che dispongono di moto ufficiale dai piloti che corrono per i cosiddetti "team saltelliti", di cui ho parlato prima;
- **Factory_rider**, fattore a 2 livelli (Yes e No) che indicano se il pilota è titolare (Yes), ovvero sta disputando tutta la stagione come gli altri, oppure se egli è un pilota sostitutivo di un pilota infortunato o una cosiddetta "wild card" (No in entrambi i casi), ovvero un pilota tester per un costruttore che partecipa al weekend di gara con lo scopo di testare nuovi componenti della moto da eventualmente poter trasferire, con la sua approvazione, sulle moto dei piloti titolari più avanti nel corso della stagione;

[14] `https://www.motogp.com/it/gp-results/2025/ita/motogp/q2/classification`, si possono cambiare dal sito la sessione, il gran premio e l'anno di cui si vogliono vedere i risultati.

- **Q_session**, fattore a 2 livelli (Q1 e Q2) relativo alla sessione di qualifica di cui sono state rilevate le sole performance in qualifica. I piloti che disputano sia Q1 che Q2 disporranno quindi di 2 unità, ovvero quelle con le rilevazioni di qualifica relativa alle Q1 (dove la nomenclatura dell'unità sarà nome, cognome e anno di rilevazione scritti nel formato NCOGAA_1) e quelle con le rilevazioni di qualifica relativa alle Q2 (dove la nomenclatura dell'unità sarà scritta nel formato NCOGAA_2). Gli altri piloti avranno nomenclatura classica NCOGAA;

Ad esse si aggiungono poi le variabili delle performance vere e proprie per ogni sessione:

- **Q_Time**, **P1_Time**, **P2_Time** e **P3_Time**, variabili quantitative continue, si riferiscono ai migliori tempi sul giro effettuati da ogni unità rispettivamente nelle sessioni di qualifica, nelle P1, nelle P2 e nelle P3. Sono misurati in secondi;
- **Q_Lap**, **P1_Lap**, **P2_Lap** e **P3_Lap**, variabili quantitative continue, si riferiscono al conteggio di giri di pista che sono serviti al pilota per effettuare il proprio miglior giro nella relativa sessione;
- **Q_Top_speed**, **P1_Top_speed**, **P2_Top_speed** e **P3_Top_speed**, variabili quantitative continue, si riferiscono alla velocità di punta raggiunta dal pilota alla fine del rettilineo principale nella relativa sessione;
- **Q_Front_tyre**, **P1_Front_tyre**, **P2_Front_tyre** e **P3_Front_tyre**, fattori con massimo 3 livelli (SS = slick soft, SM = slick media e SH = slick hard, può capitare che alcune di esse abbiano solo 2 modalità) che indicanono le mescole della gomma anteriore usate dai piloti per effettuare il proprio miglior giro in ogni sessione;
- **P1_Rear_tyre**, **P2_Rear_tyre** e **P3_Rear_tyre**, fattori con massimo 3 livelli (SS = slick soft, SM = slick media e SH = slick hard, può capitare che alcune di esse abbiano solo 2 modalità) che indicanono le mescole della gomma posteriore usate dai piloti per effettuare il proprio miglior giro in ogni sessione di prova libera. **Non è presente la variabile Q_Rear_tyre** perché sarebbe presente la sola modalità SS, dato che nel dataset tutti i giri in qualifica sono stati effettuati usando una gomma posteriore di mescola morbida;
- **Q_Front_tyre_age**, **P1_Front_tyre_age**, **P2_Front_tyre_age** e **P3_Front_tyre_age**, variabili quantitative continue, che indicano per ciascuna sessione il numero di giri che sono stati effettuati dalla gomma anteriore usata per effettuare il proprio miglior tempo di sessione dal pilota;
- **Q_Rear_tyre_age**, **P1_Rear_tyre_age**, **P2_Rear_tyre_age** e **P3_Rear_tyre_age**, variabili quantitative continue, che indicano per ciascuna sessione il numero di giri che sono stati effettuati dalla gomma posteriore usata per effettuare il proprio miglior tempo di sessione da ogni pilota;
- **Q_Sky_cond**, **P1_Sky_cond**, **P2_Sky_cond** e **P3_Sky_cond**, fattori con massimo 4 livelli (Clear, Partly cloudy, Cloudy e Light rain, può essere che alcuni non presentino tutte le modalità), che indicano le condizioni del cielo con cui ogni pilota ha avuto a cha fare nella relativa sessione, quando ha registrato il proprio miglior tempo;
- **Q_Humidity**, **P1_Humidity**, **P2_Humidity** e **P3_Humidity**, variabile quantitativa continua ma con dominio [0:1] per indicare l'umidità con cui ogni pilota ha avuto a cha fare in ogni sessione quando ha registrato il proprio miglior tempo;
- **Q_Air_temp**, **P1_Air_temp**, **P2_Air_temp** e **P3_Air_temp**, variabili quantitative continue, usate per indicare le temperature dell'aria (misurate in gradi centigradi) con cui ogni pilota ha avuto a che fare in ogni sessione, quando ha registrato il proprio miglior tempo;
- **Q_Ground_temp**, **P1_Ground_temp**, **P2_Ground_temp** e **P3_Ground_temp**, variabili quantitative continue, usate per indicare le temperature dell'asfalto (misurate in gradi centigradi) con cui ogni pilota ha avuto a cha fare in ogni sessione, quando ha registrato il proprio miglior tempo;

Il dataset iniziale comprenderà 227 unità statistiche e 50 variabili totali ed ogni unità presenterà la seguente struttura:

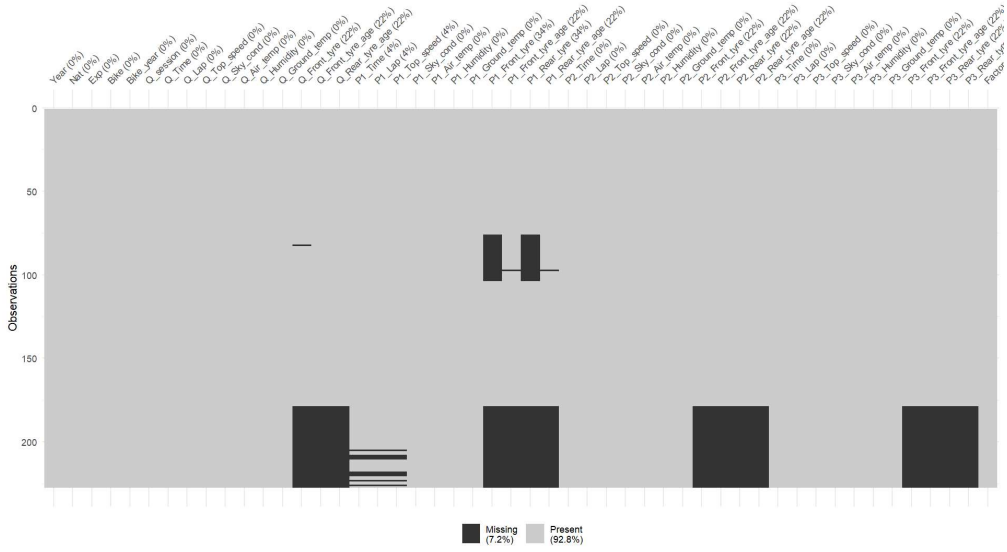
Rider	Year	Nat	Exp	Bike	Bike_year	Q_session	Q_Time	Q_Lap	Q_Top_speed	.....
MMAR25	2025	Spa	12	Ducati	2025	Q2	104.169	5	349.5	.....

Rider, che nell'esempio proposto si riferisce al giro in qualifica del pilota Marc Marquez nel 2025, presenterà la nostra unità statistica. I dati non comprendono rilevazioni nell'anno 2020 poiché in quell'anno il Gran Premio non si disputò a causa della pandemia del virus SARS-CoV-2.

2 EDA

2.1 Preprocessing

Il seguente grafico mostra la presenza di dati mancanti per alcune variabili:



In particolar modo, ad essere assenti, sono principalmente i dati relativi alle mescole e ai consumi di gomme anteriori e posteriori, sia in qualifica che in prova, degli anni 2016 e 2017 e i dati relativi ai consumi delle gomme anteriori e posteriori dei piloti nelle P1 del 2022. Abbiamo poi valori mancanti nei tempi sul giro nelle P1 del 2016, dato che alcuni piloti in quella specifica occasione non sono scesi in pista per via delle condizioni avverse e anche uno relativo al consumo della gomma anteriore in qualifica del 2022.
Detto ciò, rimuovendo quindi le osservazioni con dati mancanti, ci ritroviamo ad avere il dataset con le osservazioni dall'anno 2018 in poi, 2020 e 2022 esclusi, con le unità statistiche che diventano quindi 148, quasi 80 in meno (circa il 35 %).
A seguito di ciò viene poi eseguita una trasformazione ad una variabile, ovvero Nat: se io infatti vado a vedere le frequenze assolute di quella variabile, ottengo che ho veramente tante frequenze delle modalità "Spa" (relativo alla nazionalità spagnola) e "Ita" (relativo alla nazionalità italiana), con "Fra" (relativo alla nazionalità francese) che ha la terza frequenza più elevata ed è l'unico oltre alle altre 2 ad avere più di 10 conteggi:

Tipo frequenza	Aus	Bel	Cze	Fra	Gbr	Ger	Ita	Jap	Mal	Por	Rsa	Spa	Tha
Assoluta	8	1	2	11	4	1	44	7	2	5	4	58	1
Relativa	0.054	0.007	0.014	0.074	0.027	0.007	0.297	0.047	0.014	0.034	0.027	0.392	0.007

Per questo motivo decido di trasformare la variabile in un fattore a 4 livelli con modalità "Spa" per piloti di nazionalità spagnola, "Ita" per piloti di nazionalità italiana, "Fra" per piloti di nazionalità francese e "Oth" per piloti di altre nazionalità. così facendo ho le seguenti distribuzioni di frequenza:

Tipo frequenza	Fra	Ita	Oth	Spa
Assoluta	11	44	35	58
Relativa	0.074	0.297	0.236	0.392

Dopodiché posso iniziare la mia EDA.

2.2 Variabili quantitative

Per analizzare il rapporto tra le mie variabili quantitative parto dall'analisi di un grafico detto corrplot, che non è altro che un grafico delle correlazioni tra le mie variabili numeriche:

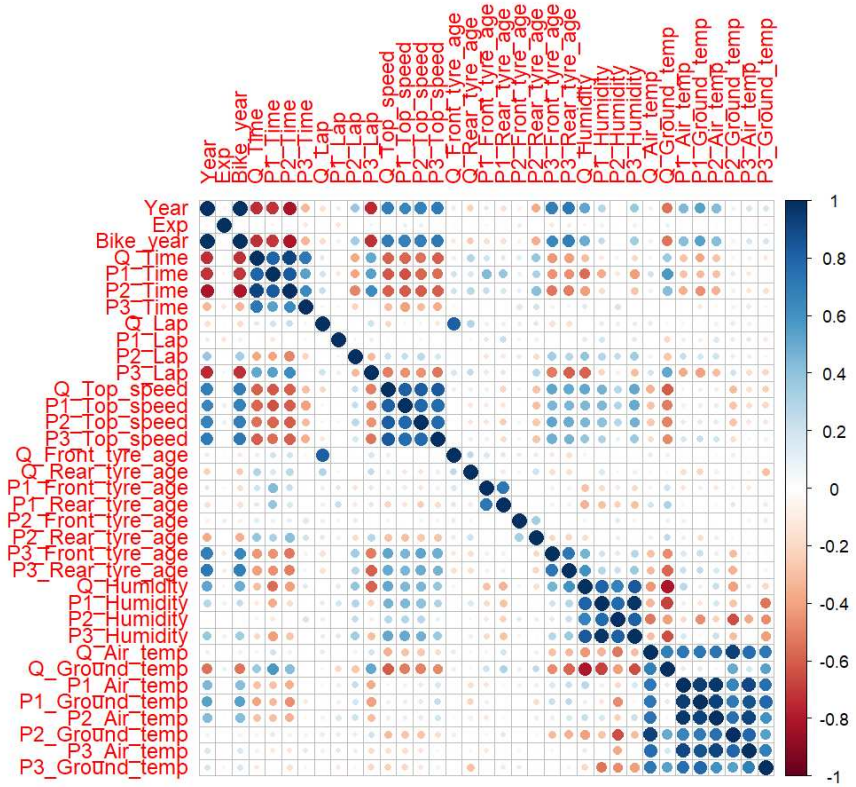


Figura 2: Corrplot delle variabili numeriche

Nel grafico i pallini rappresentano le correlazioni: più un pallino è grande e più le variabili sono correlate tra loro. Se il pallino è tendente al rosso vuol dire che sono correlate negativamente, se è tendente al blu sono correlate positivamente.
Ad esempio, quindi, noto grande correlazione positiva reciproca tra tutte le variabili relative ai tempi sul giro, alle velocità massime registrate, alle umidità e, tranne nel caso della temperatura dell'asfalto in qualifica, anche tra le temperature di aria e asfalto; tutto questo sia per le rilevazioni delle qualifiche che quelle delle prove.
Noto poi che le variabili **Year** e **Bike_year** hanno lo "stesso schema di correlazioni": oltre ad essere molto correlate positivamente tra loro, lo sono anche con le velocità massime, sia in qualifica che nelle prove, con i consumi delle gomme, sia anteriori che posteriori, delle P3, con le temperature di aria e asfalto delle P1 e con le temperature dell'aria delle P2, mentre molto correlate negativamente con i tempi sul giro, sia in prova che in qualifica, con il numero di giri necessari per ottenere il proprio miglior tempo nelle P3 e anche leggermente con le temperature dell'asfalto in qualifica. Questo vuol dire che con il passare del tempo c'è stata un'evoluzione tecnica delle moto tale queste ultime sono diventate sempre più veloci (sia in rettilineo che nei tempi sul giro, poiché l'abbassarsi dei tempi sul giro è ovviamente sintomo di miglioramento) e più gentili per quanto riguarda il consumo gomme nelle P3, mentre che le temperature dell'asfalto in qualifica sono diminuite, probabilmente perché dal 2023 le sessioni di qualifica si disputano al mattino anziché al pomeriggio.
Noto, poi, che i tempi sul giro sono correlati negativamente con le velocità massime: questo indica che più una moto è veloce in rettilineo più il tempo sul giro sarà basso.

Ciò è testimoniato anche dai seguenti grafici di dispersione:

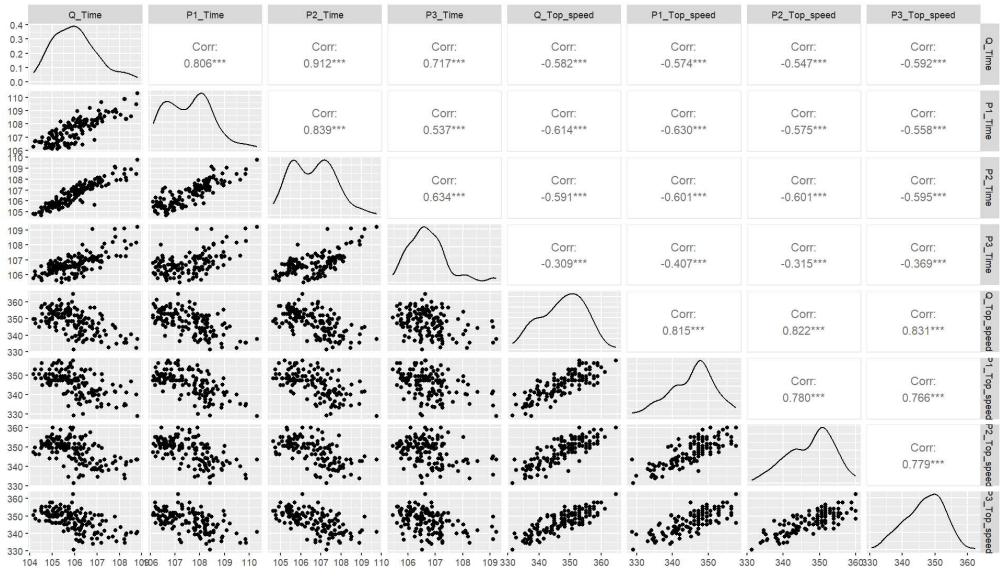


Figura 3: Grafici di dispersione con velocità di punta e tempi sul giro

Inoltre i tempi sul giro, rilevati nelle P1 e nelle P2, sono correlati negativamente con il consumo delle gomme nelle P3, sia anteriore che posteriore, mentre sono correlati positivamente con il numero di giri necessari per registrare il proprio miglior tempo nelle P3; questo quindi mi dice che i piloti più gentili con le gomme nelle P3 e che impiegano più giri degli altri a registrare il proprio miglior tempo nelle P3, sono tendenzialmente i più veloci di tutto il fine settimana, quindi quelli da tenere più d'occhio per i rivali, come testimoniato anche dai seguenti grafici di dispersione:

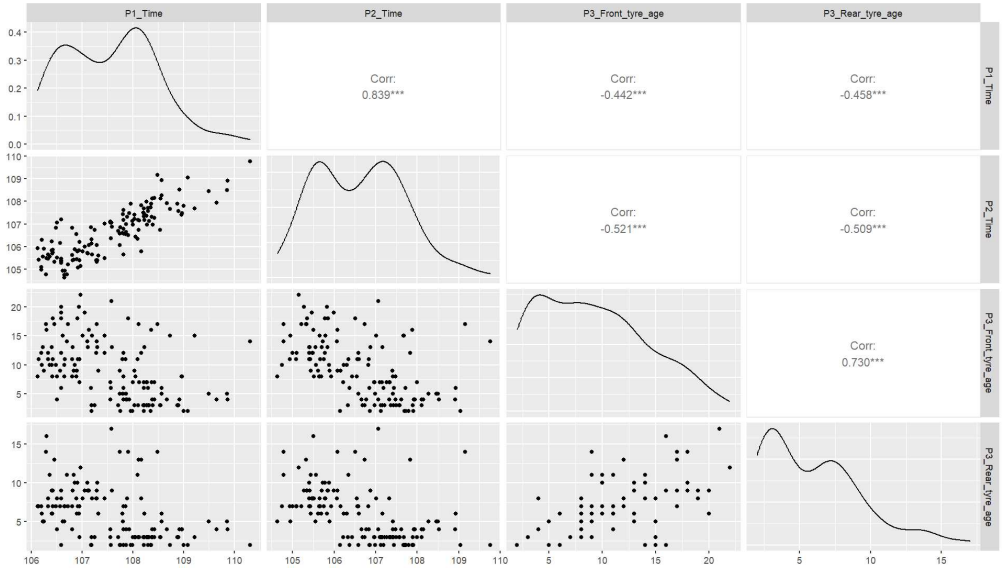


Figura 4: Grafici di dispersione con tempi sul giro di P1 e P2 e consumi gomme nelle P3

Le velocità massime, oltre ad essere correlate con variabili già descritte in precedenza, presentano correlazione negativa con la variabile **P3_Lap**, quindi i piloti con le moto più veloci in rettilineo sono

quelli che ci mettono meno tempo a trovare il limite nelle P3; inoltre, sempre le velocità massime, sono correlate positivamente con i consumi gomme delle P3 e le umidità di tutte le sessioni, segno che probabilmente con un'umidità più alta e con gomme più consumate, almeno nelle P3, le moto possono essere più veloci in rettilineo. Aggiungo poi che, sempre le velocità, sono correlate negativamente con la temperatura dell'asfalto in qualifica, segno che almeno in qualifica più la temperatura dell'asfalto è alta e più la moto può essere lenta sul rettifilo.
I consumi gomme, anteriore e posteriore, delle P1 sono correlati solamente tra loro, positivamente, mentre quelli delle P3, oltre a presentare le correlazioni già descritte in precedenza, presentano anche correlazione positiva con l'umidità in qualifica e negativa con la temperatura dell'asfalto, sempre in qualifica.
Le umidità, sia in prova che in qualifica, presentano, oltre che alle correlazioni positive reciproche descritte in precedenza, anche correlazioni negative con le temperature di aria e asfalto in qualifica. Inoltre, l'umidità delle P2 è correlata negativamente con la temperatura dell'asfalto delle P2 e, assieme anche a quella delle P1 e delle P3, anche con la temperatura dell'asfalto nelle P3. Viste questi risultati posso affarmare che, tendenzialmente, le temperature durante il weekend di gara aumentano al diminuire dell'umidità.
Infine noto che l'unica variabile relativa a qualsiasi temperatura che si differenzia dalle altre nelle correlazioni, è quella relativa alla temperatura dell'asfalto: oltre a presentare le correlazioni già descritte in precedenza, è anche correlata positivamente con la temperatura dell'aria in qualifica e con le temperature dell'asfalto delle P2 e delle P3, ma non è correlata con nessun'altra variabile misurante qualsiasi tipo di temperature.
Per ogni variabile numerica, poi, mostro un "resoconto", con medie, mediane e quantili:

Variabile	Minimo	Quantile 0.25	Mediana	Media	Quantile 0.75	Massimo
Year	2018	2019	2023	2021.689	2024	2025
Exp	1	2.75	5	5.297	7	18
Bike_year	2016	2019	2022	2021.426	2024	2025
Q_Time	104.169	105.235	105.951	106.025	106.570	108.794
P1_Time	106.121	106.802	107.651	107.571	108.218	110.315
P2_Time	104.634	105.684	106.655	106.613	107.403	109.757
P3_Time	105.456	106.271	106.644	106.727	107.098	109.224
Q_Lap	2	5	6	5.453	7	8
P1_Lap	3	8.75	15	13.135	17	20
P2_Lap	3	16	19	17.818	21	26
P3_Lap	3	6	13	11.885	17	20
Q_Top_speed	331.5	342.6	348.3	347.682	352.9	364.8
P1_Top_speed	328.8	341.075	347.2	345.478	349.5	357.6
P2_Top_speed	331.4	343.425	349.5	347.672	351.7	360.0
P3_Top_speed	330.6	342.8	347.2	346.832	350.825	362.4
Q_Front_tyre_age	2	3	6	5.189	7	13
Q_Rear_tyre_age	2	2	3	2.574	3	4
P1_Front_tyre_age	2	3	5	6.993	9	20
P1_Rear_tyre_age	2	3	6	7.818	12	20
P2_Front_tyre_age	2	3	5	5.5	6	18
P2_Rear_tyre_age	2	2	3	3.338	4	19
P3_Front_tyre_age	2	4	8.5	9.122	12.25	22
P3_Rear_tyre_age	2	3	5	5.912	8	17
Q_Humidity	0.3	0.365	0.51	0.501	0.583	0.74
P1_Humidity	0.45	0.56	0.61	0.598	0.63	0.72
P2_Humidity	0.32	0.37	0.48	0.435	0.5	0.53
P3_Humidity	0.38	0.58	0.64	0.6	0.65	0.74
Q_Air_temp	20	23	25	25	26	30
Q_Ground_temp	29	34	44	41.027	46	48
P1_Air_temp	19	20	20	21.953	23	28

Variabile	Minimo	Quantile 0.25	Mediana	Media	Quantile 0.75	Massimo
P1_Ground_temp	28	29	30.5	32.27	33	43
P2_Air_temp	21	23	25	25.764	28	33
P2_Ground_temp	36	43	43	43.568	46	49
P3_Air_temp	18	21	23	22.946	25	28
P3_Ground_temp	28	29	30	32.39	34	40

Una cosa importante da notare è che generalmente i tempi sul giro e le velocità di punta sono più alte nelle qualifiche che nelle altre sessioni: questo perché le qualifiche sono le sessioni che contano più di tutte, quindi sono quelle in cui i piloti danno di più il massimo e dove anche le moto presentano messe a punto dell'elettronica, relative soprattutto all'erogazione della potenza del motore, tali per cui i mezzi riescono ad essere più veloci in rettilineo. Inoltre vediamo come i piloti tendano ad usare pneumatici posteriori più usurati di quelli anteriori: questo perché si vede che quelli anteriori si consumano più in fretta, o sono quelli la cui usura è peggiore da gestire per i piloti, rispetto a quelli posteriori, quindi vengono cambiati più spesso nel corso della sessione.

2.3 Variabili qualitative

Partiamo dal vedere, attraverso l'uso di tabelle di frequenza relativa, quali sono le combinazioni di mescole di gomme più "suggestionate" per ogni sessione, a partire dalle **P1**:

	P1_Rear_tyre		
P1_Front_tyre	SH	SM	SS
SH	0	0.727	0.273
SM	0	0.706	0.294
SS	0.035	0.837	0.128

Sembra di vedere che, a prescindere dalla mescola usata al pneumatico anteriore, la mescola del pneumatico posteriore prevalente sia la mescola media. Ha più probabilità di essere usata quando il pneumatico anteriore è di mescola morbida, per cui viene da dire che la combinazione SS davanti e SM dietro sia la più efficiente nelle P1. Vediamo cosa ci dicono le tabelle di frequenza delle singole varibili:

P1_Front_tyre	SH	SM	SS
Frequenza assoluta	11	51	86
Frequenza relativa	0.074	0.345	0.581

P1_Rear_tyre	SH	SM	SS
Frequenza assoluta	3	116	29
Frequenza relativa	0.02	0.784	0.196

Da queste 2 tabelle capisco che effettivamente le modalità appena descritte in precedenza sono anche le più frequenti per le rispettive variabili. Questo conferma che per le P1 è più suggestionabile l'utilizzo di una gomma anteriore di mescola morbida e di una gomma posteriore di mescola media.
Passiamo ora alle combinazioni di pneumatici nelle **P2**:

	P2_Rear_tyre	
P2_Front_tyre	SM	SS
SH	0	1
SM	0.083	0.917
SS	0.2	0.8

Nelle P2 noto che a prescindere dalla mescola del pneumatico anteriore, la mescola del pneumatico posteriore prevalente sia quella morbida. Ha più probabilità di essere scelta quando il pneumatico anteriore è di mescola dura. Tuttavia, se vado a guardare le tabelle di frequenza per le singole variabili, ottengo un risultato diverso:

P2_Front_tyre	SH	SM	SS
Frequenza assoluta	34	109	5
Frequenza relativa	0.23	0.736	0.034

P2_Rear_tyre	SM	SS
Frequenza assoluta	10	138
Frequenza relativa	0.068	0.932

Praticamente tre quarti dei tempi sul giro nelle P2 sono stati effettuati con mescola media all'anteriore. Per cui dire che la scelta più azzeccata sia combinare mescole dura per il pneumatico davanti e morbida per il pneumatico dietro è un po' una forzatura. Sicuramente però capisco che se un pilota sceglie di usare una mescola dura del pneumatico anteriore, allora è molto consigliato l'utilizzo di una mescola morbida del pneumatico posteriore. Per quanto riguarda la combinazione più azzeccata, sembrerebbe essere la combinazione di gomme media davanti e morbida dietro.
Passiamo ora alle **P3**:

	P3_Rear_tyre	
P3_Front_tyre	SM	SS
SH	0.158	0.842
SM	0.041	0.959
SS	0	1

La gomma morbida anteriore viene scelta solo in coppia con una gomma morbida posteriore, a quanto sembra. Quest'ultima sembrerebbe essere la più azzeccata delle scelte, per il pneumatico posteriore. Se andiamo a guardare poi le tabelle di frequenza singole:

P3_Front_tyre	SH	SM	SS
Frequenza assoluta	19	121	8
Frequenza relativa	0.128	0.818	0.054

P3_Rear_tyre	SM	SS
Frequenza assoluta	8	140
Frequenza relativa	0.054	0.946

Capisco che per le P3 la combinazione ideale darebbe la stessa delle P2, ovvero media anteriore - morbida posteriore.
Spostiamoci infine sulle **qualifiche**, dove analizzeremo solo la mescola del pneumatico anteriore, come spiegato nell'introduzione:

Q_Front_tyre	SH	SM	SS
Frequenza assoluta	28	107	13
Frequenza relativa	0.189	0.723	0.088

Sapendo che il pneumatico posteriore usato dai piloti in qualifica è sempre stato il pneumatico di mescola morbida, attraverso la tabella di frequenza della variabile relativa alla mescola della gomma anteriore da qualifica, capisco che per quest'ultima sessione la combinazione migliore sarebbe teoricamente media anteriore - morbida posteriore. La scelta migliore delle gomme per quanto riguarda la velocità la vedremo al **paragrafo 2.4**.
Spostiamoci ora sulle variabili relative alle condizioni del cielo; in particolar modo, sapendo che P1 e P2 si svolgono nello stesso giorno, il venerdì, così come P3 e qualifiche si svolgono entrambe al sabato, vediamo quali sono le potenziali combinazioni che si possono trovare nei 2 giorni:

	P2_Sky_cond	
P1_Sky_cond	Clear	Partly cloudy
Clear	0.658	0.342
Partly cloudy	0	1

	Q_Sky_cond		
P3_Sky_cond	Clear	Cloudy	Partly cloudy
Clear	0.602	0	0.398
Partly cloudy	0	1	0

Per quanto riguarda le prove del venerdì, ovvero P1 e P2, noto che se nelle P1 il cielo presenta qualche nuvola (Partly_cloudy) allora queste rimarranno anche nelle P2. Se invece il cielo nelle P1 è limpido allora è molto probabile che lo resti anche nelle P2, senza escludere però l'arrivo di qualche nuvola.
Al sabato, invece, se ho qualche nuvola nelle P3 è sicuro che queste aumenteranno di numero fino a coprire completamente il cielo in qualifica (modalità Cloudy), mentre se si cielo nelle P3 è limpido, allora, come tra P1 e P2, è molto probabile che resti così anche in qualifica, non escludendo però il potenziale arrivo di nuvole.

2.4 Rapporto tra variabili qualitative e quantitative

Partiamo con il rapporto tra tempi sul giro e nazionalità del pilota, attraverso anche l'uso di boxplot. Come visto al **paragrafo 2.1**, useremo la variabile Nat trasformata per effettuare questa analisi e quelle successive includenti le nazionalità. Partiamo quindi dai tempi sul giro delle prove del **venerdì**, ovvero di P1 e P2:

Nat2	Fra	Ita	Oth	Spa
Tempo sul giro medio nelle P1	107.242	107.527	107.923	107.453

Nat2	Fra	Ita	Oth	Spa
Tempo sul giro medio nelle P2	106.324	106.561	107.077	106.428

Già dalle tabelle dei tempi medi sul giro capisco che potenzialmente i piloti francesi possono essere i più veloci nelle sessioni del venerdì. Questo probabilmente è dettato anche dalla numerosità ridotta delle rilevazioni relative ai piloti francesi (vedere sempre **paragrafo 2.1**). I piloti di nazionalità diversa da francese, italiana e spagnola sono tendenzialmente, invece, i meno veloci (dire "più lenti" ai loro livelli sarebbe offensivo) il venerdì. Questa spiegazione viene in parte confermata dai seguenti boxplot:

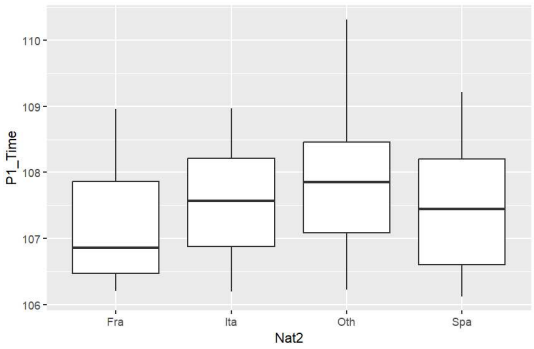


Figura 5: Boxplot relativo ai tempi nelle P1

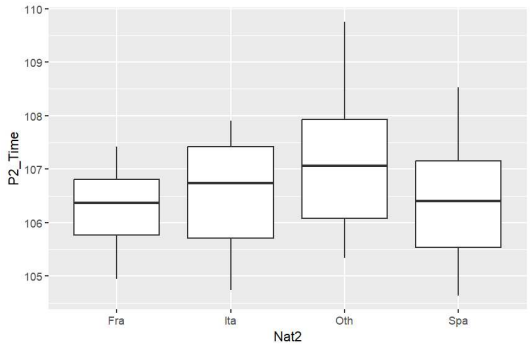


Figura 6: Boxplot relativo ai tempi nelle P2

In entrambe le sessioni notiamo come i piloti spagnoli abbiano i minimi più bassi rispetto ai piloti francesi. Tuttavia, avendo anche i massimi più alti, questi ultimi influiscono in negativo sul tempo sul giro medio. Possiamo però dire con certezza che francesi e spagnoli se la giocano per la velocità nel venerdì, gli italiani (nonostante corrano sulla pista di casa) ed i piloti di altra nazionalità sono più attardati.
Vediamo adesso le performance del **sabato**, ovvero di P3 e qualifiche:

Nat2	Fra	Ita	Oth	Spa
Tempo sul giro medio nelle P3	106.479	106.603	107.11	106.637

Nat2	Fra	Ita	Oth	Spa
Tempo sul giro medio nelle qualifiche	105.677	105.941	106.503	105.867

A primo impatto mi sembra una ripetizione del trend visto per le prove del venerdì: i francesi con un tempo medio più basso degli altri e i piloti di altra nazionalità i meno veloci di tutti. Gli italiani sembrano avvicinarsi leggermente, tanto che nelle P3 hanno un tempo medio minore di quello degli spagnoli.

Vediamo dai boxplot se la superiorità numerica di questi ultimi li penalizza come nelle prove del venerdì:

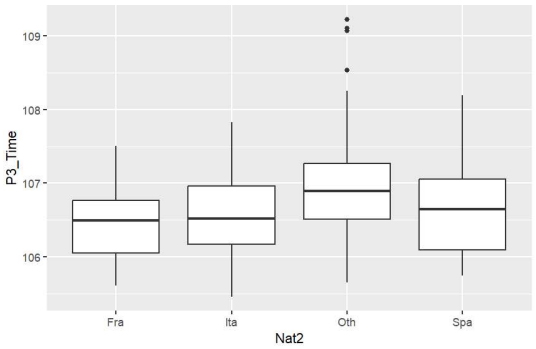


Figura 7: Boxplot relativo ai tempi nelle P3

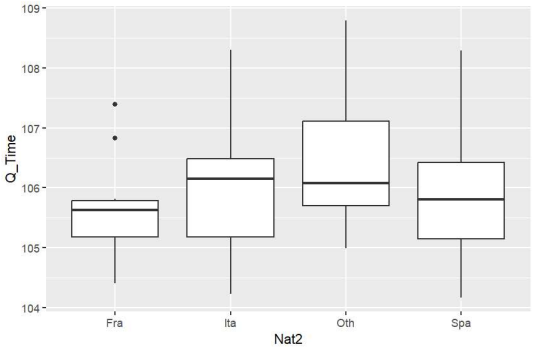


Figura 8: Boxplot relativo ai tempi nelle qualifiche

La superiorità numerica degli spagnoli ne penalizza il tempo medio sul giro solo nelle qualifiche, dove hanno, come nelle prove del venerdì, i minimi più bassi ma i massimi più alti rispetto ai francesi, che nelle qualifiche presentano persino 2 outliar.
Nelle P3, invece, i più penalizzati dalla propria superiorità numerica sembrano essere gli italiani, che hanno i minimi più bassi di tutti, ma i massimi più alti di quelli dei francesi.
In sostanza dire che i francesi sono i più veloci di tutti sarebbe un po' azzardato, visto la loro inferiorità numerica che nel loro caso, avendo piloti comunque molto veloci, può giocare a loro vantaggio.
Vediamo adesso il rapporto tra tempi sul giro e costruttore di motociclette, a partire dalle prove del venerdì. Prima, però, vediamo le frequenze relative della variabile Bike, per vedere se si possono ripetere casi come per la variabile Nat dove una modalità, essendo in superiorità o inferiorità numerica, può avere svantaggi o vantaggi nel calcolo del tempo medio:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Frequenza assoluta	23	49	25	24	6	21
Frequenza relativa	0.155	0.331	0.169	0.162	0.041	0.142

Notiamo che abbiamo pochissime rilevazioni per il costruttore Suzuki e molte per il costruttore Ducati, motivo per il quale staremo molto attenti a quelle 2 modalità quando legheremo Bike ad una qualsiasi variabile numerica. Partiamo, quindi, con il legame tra Bike e i tempi sul giro delle prove del **venerdì**:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle P1	107.554	107.401	107.803	107.684	107.797	107.514

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle P2	106.425	106.469	106.884	106.746	107.09	106.546

Notiamo che nelle P1 le Ducati sono tendenzialmente le moto più veloci, mentre le Honda le meno veloci. Tra le 2 medie però da notare una differenza di soli 0.4 secondi, il che indica che non c'è una differenza abissale.
Nelle P2 invece ad essere più veloci sembrerebbero essere le Aprilia, con Ducati che insegue a pochi centesimi. I tempi delle Suzuki sono i più lenti, non tanto per via dell'inferiorità numerica delle loro rilevazioni quanto per la mancanza di rilevazioni a partire dal 2022.

Se infatti calcolassimo il tempo sul giro medio fino all'anno 2021, ultimo anno delle rilevazioni con moto Suzuki, otterremmo il seguente risultato:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle P2	107.578	107.51	107.332	107.705	107.09	107.182

Le Suzuki, fino ad allora, sarebbero state potenzialmente le moto più veloci, mentre le Aprilia tra le moto più lente!
Nel complesso sembrerebbe quindi esserci equilibrio tra le case a seconda degli anni di rilevazione. Una cosa certa è che le 3 case europee, ovvero Aprilia, Ducati e KTM hanno avuto miglioramenti di ben 1 secondo (che nelle competizioni è veramente tanto a questi livelli) per quanto riguarda i tempi medi nelle P2 considerando anche le rilevazioni successive al 2021, mentre Honda e Yamaha solo di 5 o 6 decimi di secondo. Questa è quindi una probabile testimonianza della maggiore evoluzione che le case europee hanno avuto nello sviluppo delle proprie moto, nel corso degli ultimi 4 anni.

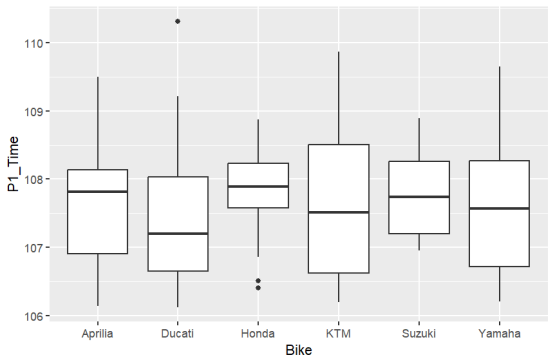


Figura 9: Boxplot relativo ai tempi nelle P1

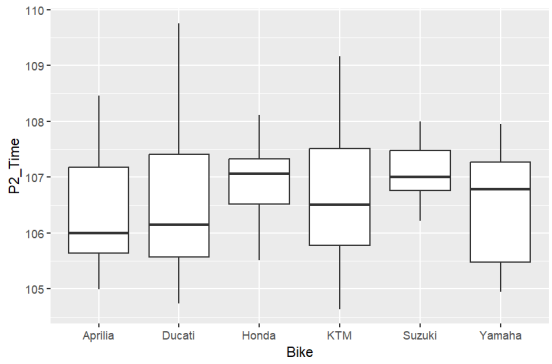


Figura 10: Boxplot relativo ai tempi nelle P2

Come possiamo poi notare dai boxplot c'è un buon equilibrio tra le case: Aprilia, Ducati, KTM e Yamaha, infatti, hanno i minimi molto simili tra loro nelle P1, mentre Honda e Suzuki (ricordiamolo, "penalizzata"), soprattutto la prima, sono quelle più faticanti. Honda ha infatti addirittura 2 outliar verso però il basso, quindi stiamo parlando di 2 tempi sul giro che sono stati eseguiti, probabilmente, oltre quello che è il potenziale della moto.
Nelle P2 ad avere più potenziale sembrerebbero essere Ducati e KTM, con Aprilia e Yamaha leggermente più attardate. Da notare, quindi, che i boxplot confermano come i tempi medi sul giro, se usati come indicatori di performance, siano un po' da prendere con le pinze (tranne nel caso di Honda), poiché è vero che Aprilia ha la media più bassa (e anche la mediana, come si vede dal boxplot), ma è anche vero che probabilmente le KTM sono quelle con più potenziale, se usate dai giusti piloti e messe a punto nel modo più consono.
Spostiamoci ora alle sessioni di prova e qualifica del **sabato**:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle P3	106.874	106.591	106.827	106.919	106.708	106.551

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle qualifiche	106.082	105.754	106.297	106.203	106.564	105.916

Nelle P3 sembra esserci molto equilibrio, a primo impatto: infatti il gap tra il tempo medio di Yamaha, la più veloce, e quello di KTM, la più lenta, è di soli 3 decimi di secondo, quindi per trarre conlusioni bisognerebbe vedere i boxplot.
Nelle qualifiche, invece, notiamo ancora una volta come Suzuki sia penalizzata dalla sua mancanza di rilevazioni recenti: il suo tempo medio è di ben 8 decimi superiore rispetto a quello di Ducati, la più

veloce. Togliendo quindi Suzuki, le più lente sembrerebbero essere Honda e KTM. Proviamo però a vedere cosa succederebbe se le rilevazioni si fermassero al 2021, come abbiamo fatto per le P2:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro medio nelle qualifiche	107.277	106.553	106.655	107.056	106.564	106.345

Suzuki see la sarebbe giocata persino con Ducati, che qui sarebbe la seconda più veloce dopo Yamaha. Un'altra volta, però, come dalle P2, notiamo come le 3 case europee, soprattutto Aprilia, si siano evolute con il passare del tempo, abbassando i propri tempi sul giro dai 0.8 agli 1.2 secondi, mentre Honda e Yamaha (giapponesi) si siano evolute relativamente poco, abbassando di soli 4 decimi di secondo i propri tempi medi sul giro. Ma cosa di dicono i boxplot?

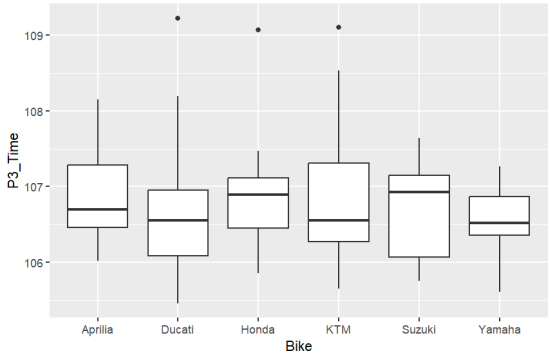


Figura 11: Boxplot relativo ai tempi nelle P3

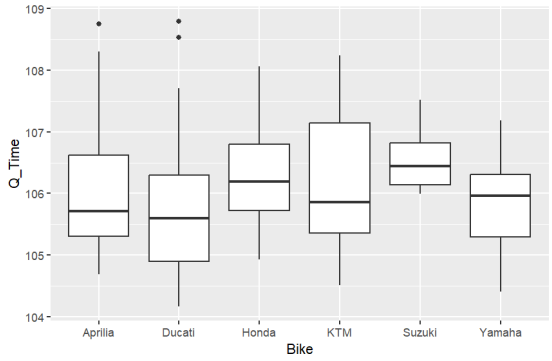


Figura 12: Boxplot relativo ai tempi nelle qualifiche

Il boxplot delle P3 parla di una leggera superiorità di Ducati e Yamaha, con la prima delle 2 che ha il minimo più basso quindi sembrerebbe essere quella con più potenziale. Per il resto sembra confermarsi l'equilibrio, con addirittura Ducati, Honda e KTM che presentano degli outliar verso l'alto, segno di tempi veramente troppo alti, in relazione al potenziale della moto.
Il boxplot delle qualifiche, invece, conferma la superiorità di Ducati, con Yamaha che invece presenta un tempo mediano superiore rispetto a quelli di Aprilia e KTM:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Tempo sul giro mediano nelle qualifiche	105.718	105.598	106.195	105.858	106.443	105.959

Tolta Suzuki, la casa con più difficoltà su questo circuito nelle qualifiche, ma a quanto abbiamo visto non solo, sembra quindi essere Honda: ha il tempo medio e mediano più alto nelle qualifiche (sempre tolta Suzuki), ma curiosamente non è quella con il massimo più alto. Quest'ultimo sembrerebbe appartenere a Ducati ed è però un outliar; tolti gli outliar sembrerebbe appartenere ad Aprilia. Da notare infine che le dimensioni delle "scatole" nei grafici, per quanto riguarda le 3 case europee sono più elevate: ciò indica più variabilità e può essere un altro segno dei miglioramenti tecnici effettuati negli ultimi anni.
Cambiando invece argomento, essendo il tempo sul giro correlato positivamente con le velocità massime, **quali sono poi le moto più veloci in rettilineo?** Per rispondere a questa domanda verificheremo sia le top speed medie che quelle massime, per ogni costruttore, a partire come sempre dalle sessioni del **venerdì**:

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Top speed medie nelle P1	345.857	346.767	344.3	347.083	340.283	343.105
Top speed massime nelle P1	349.5	356.4	355.2	357.6	347.2	357.6

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Top speed medie nelle P2	349.274	348.2	347.66	349.508	341.817	344.2715
Top speed massime nelle P2	358.8	355.9	360	360	348.3	354

Sia per le P1 che per le P2 noto una cosa in comune: KTM ha sia la top speed media che la top speed massima più alte. Quindi questo vuol dire che, motoristicamente, sembrerebbero essere le moto meglio dotate. Noto poi un miglioramento di Aprilia dalle P1 alle P2, mentre le Suzuki patiscono, così come le Yamaha (queste ultime meno), il fatto di essere dotate di motori con 4 cilindri disposti in linea, al contrario delle altre case che li dispongono a V[15]. Questa "pensalizzazione" si nota soprattutto nelle P2. Ma come queste eventuali penalizzazioni si riflettono sui tempi sul giro?

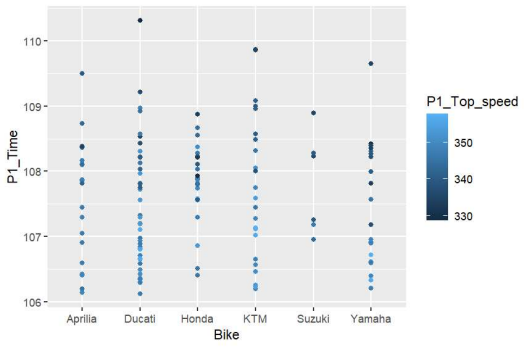


Figura 13: Grafico relativo alle top speed delle P1

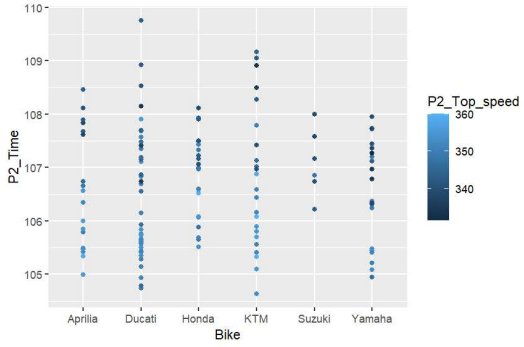


Figura 14: Grafico relativo alle top speed delle P2

I tempi sul giro di un po' tutte le case potrebbero essere influenzate anche dalle top speed, quelle però meno influenzate sembrerebbero essere Honda e Suzuki, quest'ultima per i motivi già spiegati in precedenza.
E nelle sessioni del **sabato** invece?

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Top speed medie nelle P3	346.835	347.857	345.888	350.129	340.183	343.69
Top speed massime nelle P3	357.6	355.2	351.7	362.4	344.7	354

Costruttore	Aprilia	Ducati	Honda	KTM	Suzuki	Yamaha
Top speed medie nelle qualifiche	349.235	348.896	347.088	349.808	340.267	343.548
Top speed massime nelle qualifiche	358.8	358.8	358.8	364.8	346.1	355.2

Le conclusioni sono le stesse di prima; per quanto Aprilia si avvicini molto a KTM in qualifica, in termini di top speed media, quest'ultima è la più convincente in termini di velocità di punta. Aprilia, invece, condivide con Ducati e Honda la stessa top speed massima in qualifica. Yamaha e Suzuki restano le più lente.

[15]Corsedimoto, *MotoGP, la tecnica: ecco perché Yamaha esita tra il V4 e il 4 in linea*, `https://www.corsedimoto.com/motogp/motogp-tecnica-perche-yamaha-esita-tra-v4-e-4-in-linea`

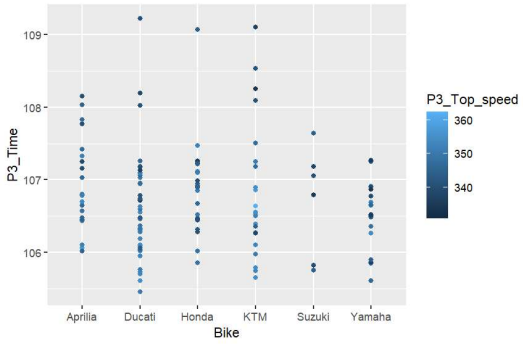


Figura 15: Grafico relativo alle top speed delle P3

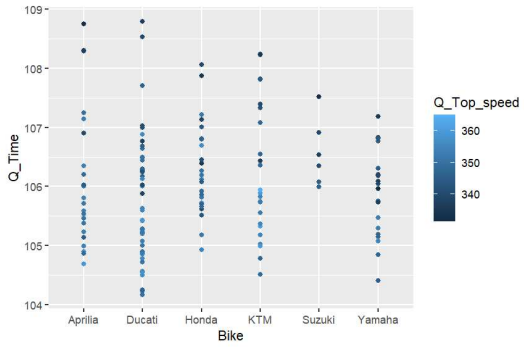


Figura 16: Grafico relativo alle top speed delle qualifiche

Sia nelle P3 che nelle qualifiche, noto poi come KTM sia quella i quali tempi sul giro siano più influenzati dalle velocità massime a fine rettilineo. Per cui si vede che eventuali lacune da compensare, nel loro caso, non siano di tipo motoristico, bensì di tipo ciclistico, telaistico o aerodinamico. I tempi sul giro delle altre case non sono influenzati così tanto dalle velocità di punta, soprattutto nel caso dei 3 costruttori giapponesi (Honda, Suzuki e Yamaha).
Focalizziamoci ora sul rapporto tra tempi sul giro e mescole di pneumatici e vediamo se il loro rapporto è influenzato dall'usura di questi ultimi, a partire dai tempi delle sole **P1**:

Mescola della gomma anteriore delle P1	SH	SM	SS
Tempo sul giro medio nelle P1	106.879	107.927	107.448

Mescola della gomma posteriore delle P1	SH	SM	SS
Tempo sul giro medio nelle P1	108.265	107.678	107.069

Queste tabelle confermano come nelle P1 le scelte di gomme migliori sembrino essere definitivamente le mescole dura e morbida, rispettivamente per pneumatici anteriore e posteriore. Verifichiamo se l'usura influenza tale rapporto:

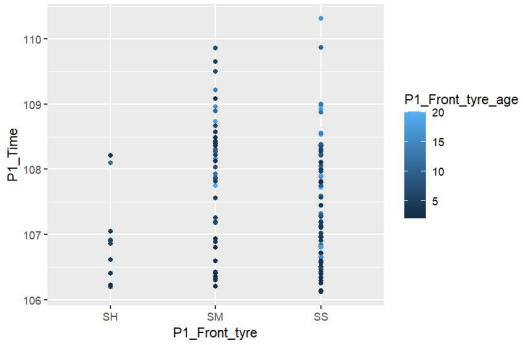


Figura 17: Grafico dei tempi sul giro nelle P1, in relazione alle gomme anteriori

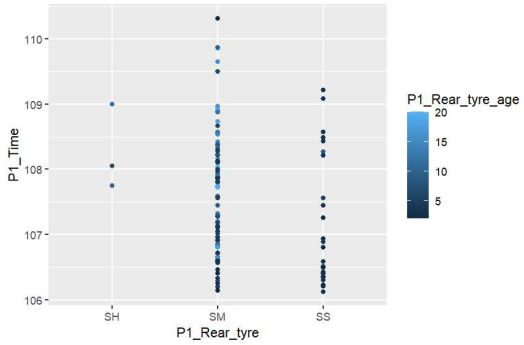


Figura 18: Grafico dei tempi sul giro nelle P1, in relazione alle gomme posteriori

L'effetto dell'usura degli pneumatici si nota bene quando la gomma anteriore è media o morbida, oppure quando quella posteriore è media: in tutti e tre i casi il tempo sul giro è più lento quando la gomma è più consumata.
Concentramoci ora sui dati delle **P2**:

Mescola della gomma anteriore delle P2	SH	SM	SS
Tempo sul giro medio nelle P2	106.795	106.581	106.073

Mescola della gomma posteriore delle P2	SM	SS
Tempo sul giro medio nelle P2	107.567	106.544

Queste tabelle confermano come, nelle P2, le scelte di gomme migliori sembrino essere definitivamente le mescole morbide per entrambi gli pneumatici. Per quanto riguarda lo pneumatico anteriore, come si può vedere anche dalle tabelle relative alle P2 (**paragrafo 2.3**), abbiamo un caso di vantaggio per inferiorità numerica: solo 5 tempi sul giro nelle P2 sono stati effettuati con la gomma morbida anteriore. Verifichiamo quindi se la mescola media possa essere un'opzione migliore e se l'usura influenzi tale rapporto:

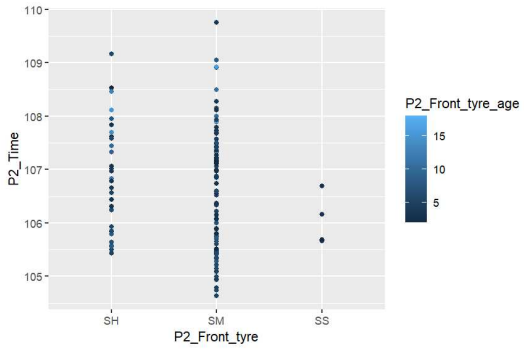


Figura 19: Grafico dei tempi sul giro nelle P2, in relazione alle gomme anteriori

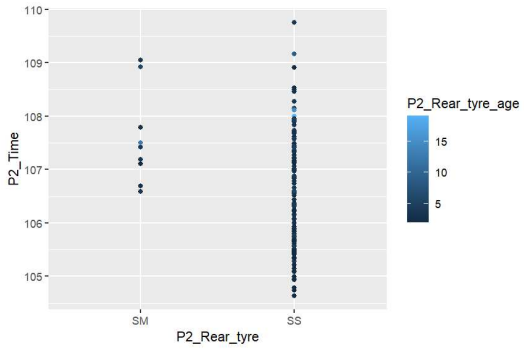


Figura 20: Grafico dei tempi sul giro nelle P2, in relazione alle gomme posteriori

Esattamente come pensavo: per quanto i tempi sul giro medi siano più bassi con le gomme morbide anteriori, con le gomme medie si possono effettuare tempi, potenzialmente, ben più rapidi. Sia poi nel loro caso che nel caso delle gomme dure, il tempo sul giro è influenzato dall'usura del pneumatico. Per la scelta dello pneumatico posteriore, quello morbido è decisamente il migliore e c'è un leggerissimo legame anche con la propria usura.
Ora guardiamo i dati relativi alle **P3**:

Mescola della gomma anteriore delle P3	SH	SM	SS
Tempo sul giro medio nelle P3	106.916	106.69	106.837

Mescola della gomma posteriore delle P3	SM	SS
Tempo sul giro medio nelle P3	107.433	106.687

Secondo le tabelle, la migliore combinazione di gomme per far registrare un buon tempo sul giro sembrerebbe essere mescola media anteriore e mescola morbida posteriore. Vediamo se però come prima c'è un vantaggio o uno svantaggio in termini di frequenze assolute:

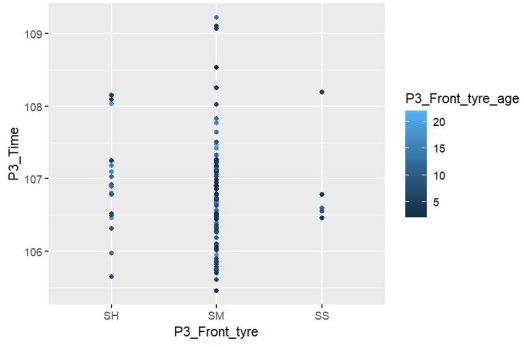


Figura 21: Grafico dei tempi sul giro nelle P3, in relazione alle gomme anteriori

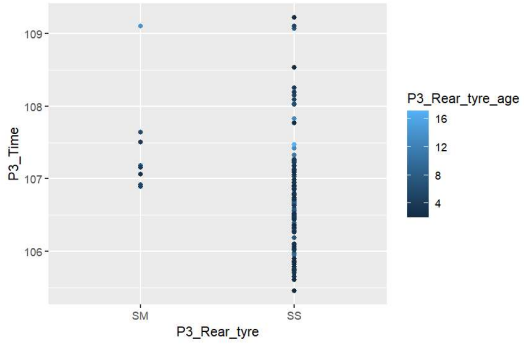


Figura 22: Grafico dei tempi sul giro nelle P3, in relazione alle gomme posteriori

Non sembra esserci, per cui posso dire che la combinazione scritta prima sia effettivamente la migliore. C'è una leggera influenza dell'usura degli pneumatici sull'effetto che la mescola media anteriore ha sul tempo sul giro, ma nulla di più.
Osserviamo infine cosa succede in **qualifica**, dove analizzeremo la sola gomma anteriore:

Mescola della gomma anteriore nelle qualifiche	SH	SM	SS
Tempo sul giro medio nelle qualifiche	106.267	106.004	105.679

A prima vista sembra che la mescola dura sia quella più efficace, avendo un tempo medio di poco più di 3 decimi più basso rispetto a quello della mescola media. Tuttavia, se vediamo il rispettivo grafico:

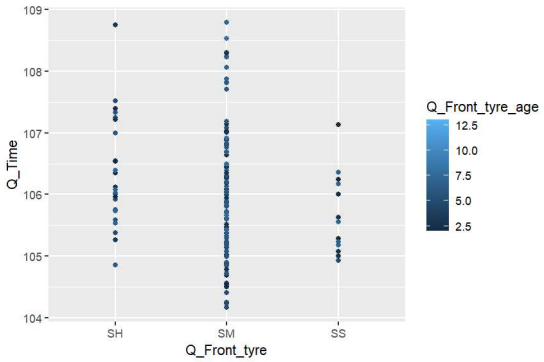


Figura 23: Grafico dei tempi sul giro nelle qualifiche, in relazione alle gomme anteriori

Esattamente come pensavo: anche in questo caso la frequenza assoluta di osservazioni bassa ha giovato in positivo sul tempo medio effettuato con la gomma di mescola dura. Quella con i tempi sul giro più bassi è definitivamente quella media, per cui se fossi un pilota professionista, ma con poca esperienza nel mondiale, tenderei a scegliere quella per effettuare un buon tempo sul giro in qualifica. Non sembra esserci un grande effetto dell'usura sul legame tra mescola e tempo sul giro.
Vediamo infine se ci possono essere legami tra le variabili atmosferiche quantitative (coloro che misurano temperature di asfalto, aria e umidità) **e le variabili atmosferiche qualitative** (nel nostro caso le condizioni del cielo). Partiamo dalle rilevazioni effettuate nelle **P1**:

Condizioni del cielo nelle P1	Clear	Partly cloudy
Temperatura media dell'aria nelle P1	22.63	21.293
Temperatura media dell'asfalto nelle P1	34.696	29.987
Umidità media nelle P1	0.542	0.653

A primo impatto sembra che, al riempirsi il cielo di nuvole, nelle P1, le temperature di asfalto e aria diminuiscano, mentre l'umidità aumenti. Vediamo con dei boxplot se effettivamente può essere così:

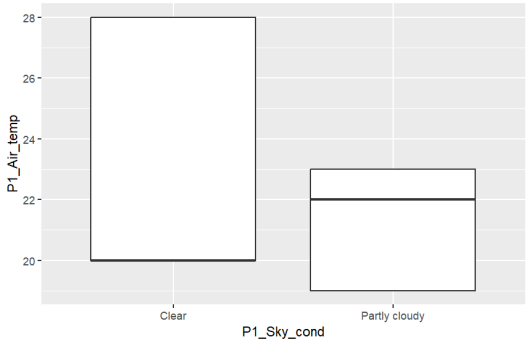


Figura 24: Boxplot del rapporto tra temperature dell'aria e condizioni del cielo, nelle P1

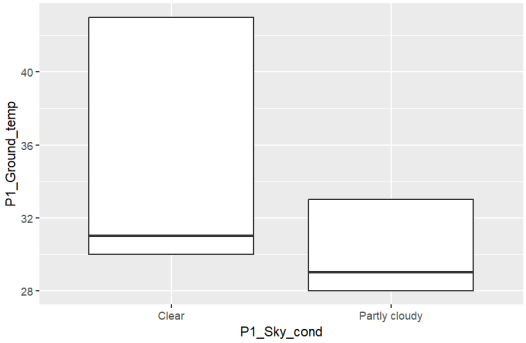


Figura 25: Boxplot del rapporto tra temperature dell'asfalto e condizioni del cielo, nelle P1

Figura 26: Boxplot del rapporto tra umidità e condizioni del cielo, nelle P1

I boxplot confermano le considerazioni appena fatte. Da notare che, avendo tante unità con gli stessi valori nelle 3 variabili (così come nelle altre variabili atmosferiche, per via della parità di condizioni nelle loro rilevazioni), in nessun grafico è presente un outliar. Potrebbe essere così anche nelle variabili atmosferiche relative alle altre sessioni. Concentriamoci ora alle rilevazioni effettuate nelle **P2**:

Condizioni del cielo nelle P2	Clear	Partly cloudy
Temperatura media dell'aria nelle P2	29	24.21
Temperatura media dell'asfalto nelle P2	47.5	41.68
Umidità media nelle P2	0.39	0.457

Nelle P2 il trend sembrerebbe essere lo stesso delle P1.

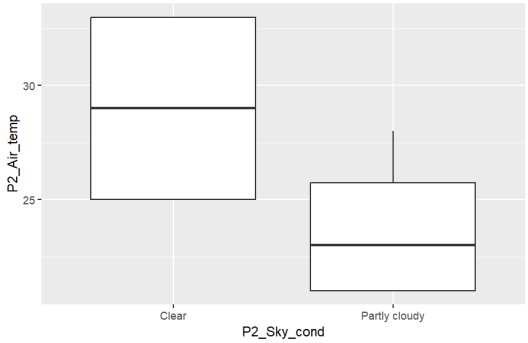


Figura 27: Boxplot del rapporto tra temperature dell'aria e condizioni del cielo, nelle P2

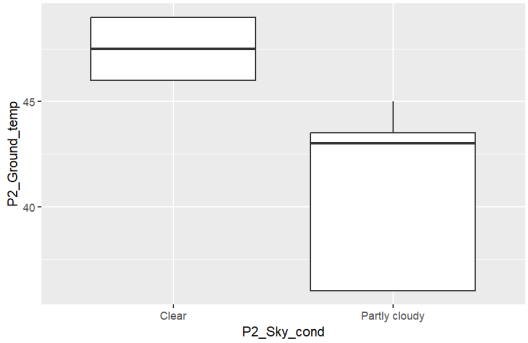


Figura 28: Boxplot del rapporto tra temperature dell'asfalto e condizioni del cielo, nelle P2

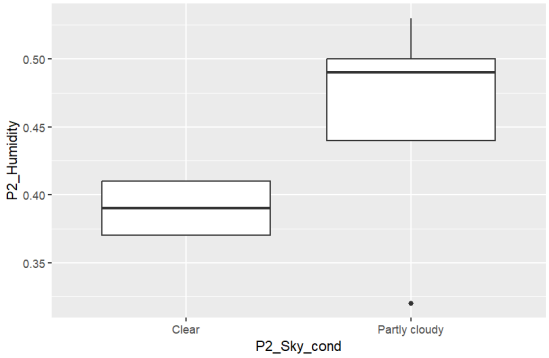


Figura 29: Boxplot del rapporto tra umidità e condizioni del cielo, nelle P2

I boxplot confermano il trend. Da notare che la temperatura dell'asfalto è molto variabile quando si presentano delle nuvole nel cielo, il che è un fattore da tenere in considerazione per un pilota che sta affrontando tali condizioni atmosferiche.
Concentriamoci ora sulle rilevazioni effettuate nelle **P3**:

Condizioni del cielo nelle P3	Clear	Partly cloudy
Temperatura media dell'aria nelle P3	22.528	25
Temperatura media dell'asfalto nelle P3	32.959	29
Umidità media nelle P3	0.572	0.74

Si nota come, apparte per la temperatura dell'aria dove avviene il contrario rispetto al solito, il trend sembrerebbe essere lo stesso anche nelle P3. Vediamo cosa ci dice il boxplot:

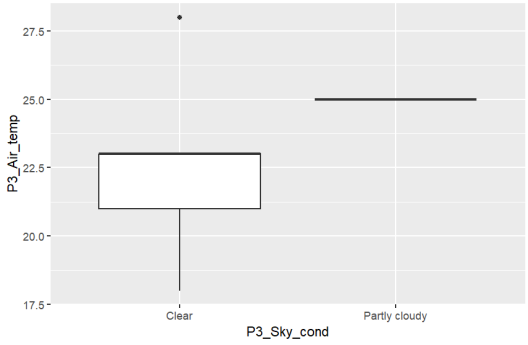


Figura 30: Boxplot del rapporto tra temperature dell'aria e condizioni del cielo, nelle P3

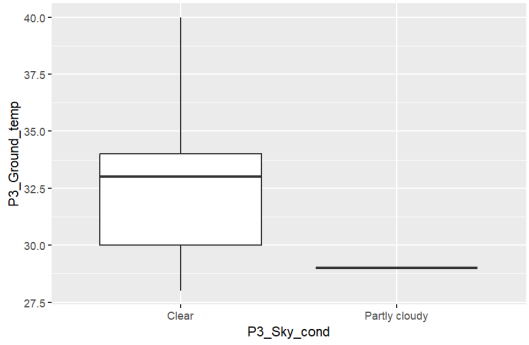


Figura 31: Boxplot del rapporto tra temperature dell'asfalto e condizioni del cielo, nelle P3

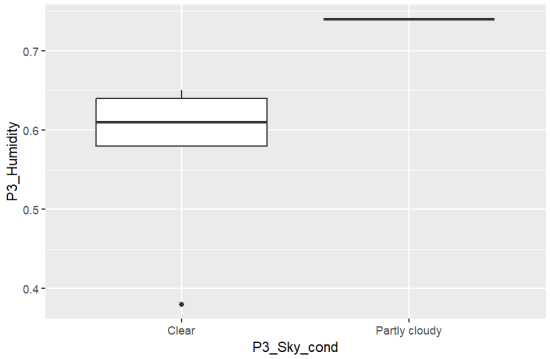


Figura 32: Boxplot del rapporto tra umidità e condizioni del cielo, nelle P3

Dai boxplot capisco che il rapporto tra la variabile atmosferica qualitativa e quelle quantitative è difficile da valutare, perchè abbiamo un solo valore per ogni variabile relativo al cielo con nuvole sparse. Lo possiamo capire anche dalle frquenze assolute della variabile P3_Sky_cond:

P3_Sky_cond	Clear	Partly cloudy
Frequenza assoluta	123	25

25 osservazioni sono bene o male il numero di osservazioni per un anno. Se infatti costruisco una tabella di frequenze assolute relativa alla variabile **Year**, ovvero l'annata della rilevazione:

Year	2018	2019	2021	2023	2024	2025
Frequenza assoluta	24	25	24	25	26	24

Le annate con 25 osservazioni sono il 2019 e il 2023, per cui posso dire con certezza che in una sola annata ho avuto delle nuvole sparse nel cielo nelle P3. Sarebbe, quindi, alquanto azzardato trarre conclusioni riguardo al rapporto spiegato in precedenza.
Spostiamoci quindi alle sessioni di **qualifica**, dove prima, però, vediamo la tabella dellle frequenze assolute per la relativa variabile delle condizioni del cielo:

P3_Sky_cond	Clear	Cloduy	Partly cloudy
Frequenza assoluta	74	25	49

Attenzione: anche in questo caso avrò, probabilmente, valori uguali per le variabili numeriche, quando la variabile qualitativa sarà di una modalità, ovvero, in questo caso, Cloudy. Continuiamo comunque l'analisi, ma con un punto di domanda:

Condizioni del cielo nelle qualifiche	Clear	Cloudy	Partly cloudy
Temperatura media dell'aria nelle qualifiche	25.189	23	25.735
Temperatura media dell'asfalto nelle qualifiche	42.432	29.52	44.776
Umidità media nelle qualifiche	0.533	0.735	0.335

C'è un legame strano: quando il cielo passa da limpido a parzialmente nuvoloso le temperature salgono leggermente, mentre l'umidità si abbassa; quando invece il cielo diventa interamente nuvoloso, le temperature riscendono vorticosamente e la temperatura si innalza alla grande. Vediamo, quindi, se il legame tra le variabili atmosferiche quantitative e quelle qualitative varia a seconda della sessione di qualifica, ovvero tra Q1 e Q2:

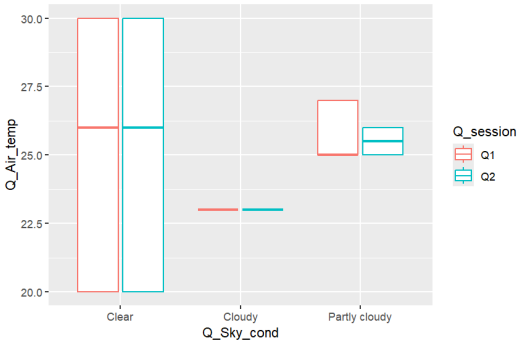


Figura 33: Boxplot della temperatura dell'aria, in relazione alle condizioni del cielo, con l'influenza della sessione di qualifica

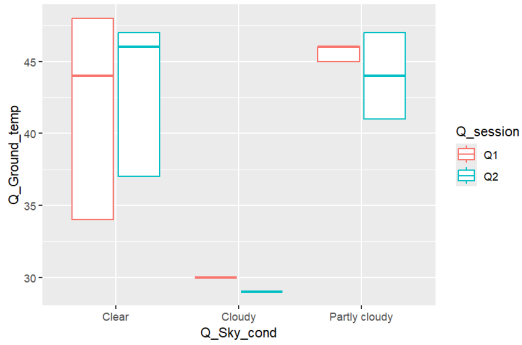


Figura 34: Boxplot della temperatura dell'aria, in relazione alle condizioni del cielo, con l'influenza della sessione di qualifica

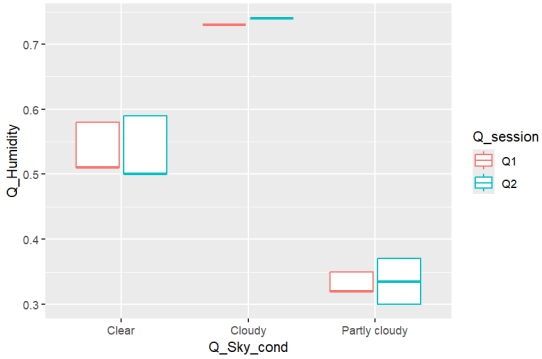


Figura 35: Boxplot dell'umidità, in relazione alle condizioni del cielo, con l'influenza della sessione di qualifica

Come possiamo vedere, è difficile dare una spiegazione al rapporto tra condizioni del cielo e variabili atmosferiche numeriche, in qualifica: la temperatura dell'aria presenta variazioni al cambiare delle condizioni del cielo e come possiamo vedere, quando il cielo è leggermente nuvoloso, essa presenta più variabilità nelle Q1 rispetto alle Q2. La temperatura dell'asfalto precipita repentinamente allo scurirsi del cielo e resta tendenzialmente più elevata nelle Q1, in questo caso, mentre semrberebbe che quando il cielo è limpido la temperatura sia più variabile, soprattutto quando si disputano le Q1. L'umidità è, invece, enormemente più elevata quando il cielo è nuvoloso, mentre come le temperature di aria e asfalto, presenta più variabilità a cielo limpido.
Questi esiti sono dati dal fatto che abbiamo una sola annata dove si sono corse le qualifiche con cielo completamente nuvoloso, per cui abbiamo valori molto simili tra loro, per quanto riguarda le variabili atmosferiche numeriche. L'anno in questione è il 2023.

3 Riduzione della dimensionalità del dataset con variabili quantitative

Per queste analisi considereremo il dataset con sole variabili numeriche: esso contiene 148 osservazioni e 35 variabili. Vogliamo vedere se quindi si può ridurre di dimensionalità il nostro dataset e, se sì, che trasformazioni subiscono le nostre variabili.

3.1 PCA (Analisi delle componenti pricipali)

Cominciamo quindi con il primo metodo, ovvero la PCA (Principal Component Analysis). E' un metodo di riduzione della dimensionalità che, data una matrice dei dati X di dimensione n*p (n = numero di unità statistiche e p = numero di variabili), prevede l'esistenza di una trasformazione lineare $Y = XE$, sempre di dimensione n*p, dove E è la matrice di dimensione p*p degli autovettori di S, matrice delle covarianze di X.
Ogni i-esima colonna di Y (Y_i, i=1,...,p), è detta **componente** ed ha varianza pari a λ_i i-esimo autovalore della matrice S. Le componenti principali sono le prime q colonne di Y, che spiegano una percentuale della sua varianza totale pari a $\sum_i^p \lambda_i$, che va dal 60 all'80 % a seconda del dataset che sto analizzando. In sostanza voglio che $\frac{\sum_i^q \lambda_i}{\sum_i^p \lambda_i} > 0.6/0.8$, per cui andrò a selezionare da Y le prime q colonne che, in sostanza, meglio riescono a riassumere X.
Quando poi, nel dataset originale, quindi la matrice X, abbiamo variabili che hanno varianze troppo diverse tra loro, oppure che hanno unità di misura diverse tra loro, conviene effettuare l'analisi su Z, matrice dei dati standardizzati, al posto della matrice X, cosicchè le variabili che hanno varianza più elevata non abbiano troppa influenza sulle componenti.
Ogni componente risulterà infatti correlata, positivamente o negativamente, con delle variabili del dataset originale. Per vedere quali esse siano analizzeremo un grafico, detto **biplot**, che vedremo al **paragrafo 3.2**.

3.2 PCA: applicazione al dataset con variabili numeriche

Nel nostro caso utilizzeremo come matrice X il dataset con le sole variabili numeriche, per via della gran maggioranza numerica di esse rispetto alle variabili qualitative (35 variabili numeriche, a fronte 15 variabili qualitative). Avrò quindi n = 148 e p = 35. Data la differenza nelle unità di misura delle variabili originarie (alcune misurano tempi sul giro in secondi, altre temperature in gradi celsius, altre ancora sono variabili di conteggio, ecc.), svolgeremo l'analisi sulla matrice dei dati standardizzati Z. Il risultato, per le prime 10 componenti, è il seguente:

Componente	**Varianza** λ_i	$\frac{\lambda_i}{\sum_{j=1}^p \lambda_j}$	$\frac{\sum_{j=1}^i \lambda_j}{\sum_{j=1}^p \lambda_j}$
PC1	10.632	0.304	0.304
PC2	6.646	0.190	0.494
PC3	3.238	0.093	0.586
PC4	2.094	0.060	0.646
PC5	1.939	0.055	0.701
PC6	1.569	0.045	0.746
PC7	1.273	0.036	0.783
PC8	1.161	0.033	0.816
PC9	0.996	0.028	0.844
PC10	0.913	0.026	0.870

Le ultime 2 colonne della tabella spiegano rispettivamente la percentuale di varianza spiegata e la percentuale di varianza cumulata spiegata per ogni componente. Attraverso quest'ultimo valore, capisco quante q componenti bastano per spiegare il fenomeno.
Nel nostro caso sembra che 3 o 4 componenti riassumino il nostro dataset a sufficienza, avendo alla quarta componente una porzione di varianza spiegata di 0.646.

Effettuo un'altra verifica, stavolta di tipo grafico:

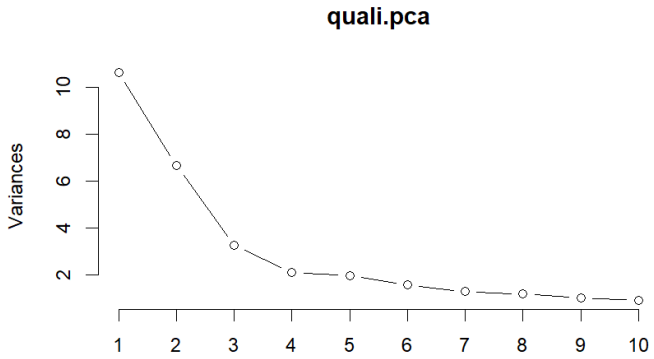


Figura 36: Grafico relativo alle varianze di ogni componente

Un metodo grafico per capire quante componenti sono sufficienti a riassumere il dataset è il "metodo del gomito": dal grafico delle varianze di ogni componente, come quello in **figura 36**, selezioniamo il numero di componenti, dall'asse delle x, dove si forma visivamente un "gomito" sulla curva.
Nel nostro caso il "gomito" si andrebbe a formare quando x=3, quindi sulla componente numero 3: potremmo, quindi, affermare che q=3 componenti principali sarebbero sufficienti a riassumere il nostro dataset.
Andiamo poi a vedere, attraverso i biplot introdotti alla fine del **paragrafo 3.1**, quali sono le variabili più correlate con ogni componente principale:

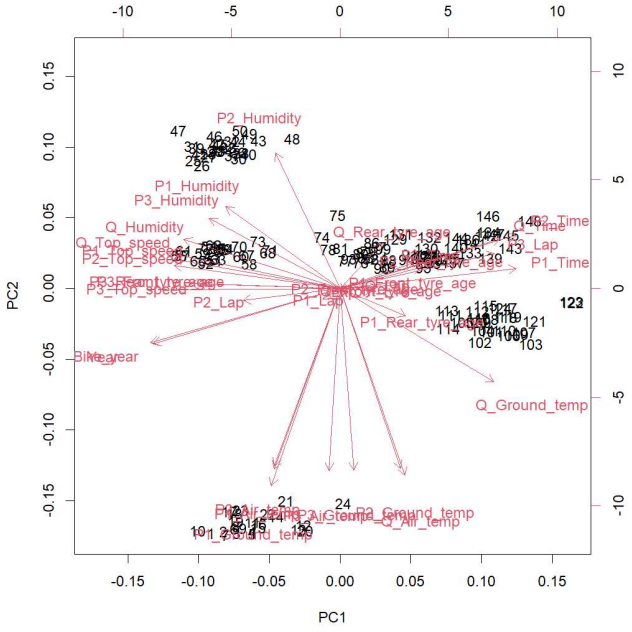


Figura 37: Biplot relativo alle prime 2 componenti principali

Come si legge un grafico di questo tipo? Le frecce centrali al grafico sono relative alle variabili del dataset originale. Più la frecca relativa ad una variabile è parallela all'asse di una componente (PC1 o PC2 nel nostro caso), più la variabile è correlata con la componente. Se tende verso la zona negativa dell'asse sarà correlata negativamente, se tende verso la parte positiva sarà correlata positivamente. Invece, più la freccia è perpendicolare all'asse più è la variabile è incorrelata.
I numeri al centro del grafico, sono relativi alle unità statistiche del dataset: più un numero è vicino alla freccia di una variabile e più quell'unità statistica avrà il valore della propria componente influenzato da quella variabile.
Parlando quindi del biplot in **figura 37**, notiamo che:

- le variabili relative alle umidità di P3, P1 e qualifica, alle top speed di tutte le sessioni, agli anni di rilevazione e delle motociclette, ai consumi delle gomme nelle P3 (anteriore e posteriore) e al numero di giri necessari per far registrare il proprio miglior tempo nelle P2, sono correlate negativamente con la prima componente;
- le variabili relative ai tempi sul giro di tutte le sessioni, alla temperatura dell'asfalto in qualifica, al numero di giri necessario per far registrare il proprio miglior tempo nelle P3 e ai consumi delle gomme posteriori nelle qualifiche, nelle P2 (nascosto assieme a P3_Time, sotto il cluster di osservazioni comprendente la numero 146) e nelle P1, sono correlate positivamente con la prima componente;
- le temperature dell'asfalto di P1, P2 e P3 e le temperature dell'aria di tutte le sessioni, sono correlate negativamente con la seconda componente, alla quale è invece correlata positivamente l'umidità delle P2.

Noto, inoltre, che sono presenti circa 5 cluster ben definiti: questo perché tra le variabili incluse sono presenti anche le varibaili Year, Bike_year e tutte le variabili atmosferiche numeriche, che variano solo con il variare dell'anno di rilevazione (umidità, temperatura dell'aria e temperatura dell'asfalto). Essendoci poche modalità per ogni variabile, nonostante esse siano numeriche, ciò porta ad una distinzione ben specifica anche nei valori delle componenti, calcolati con la trasformazione lineare $Y = XE$.
Spostiamoci ora sul biplot relativo alla terza componente, comprendente anche la seconda:

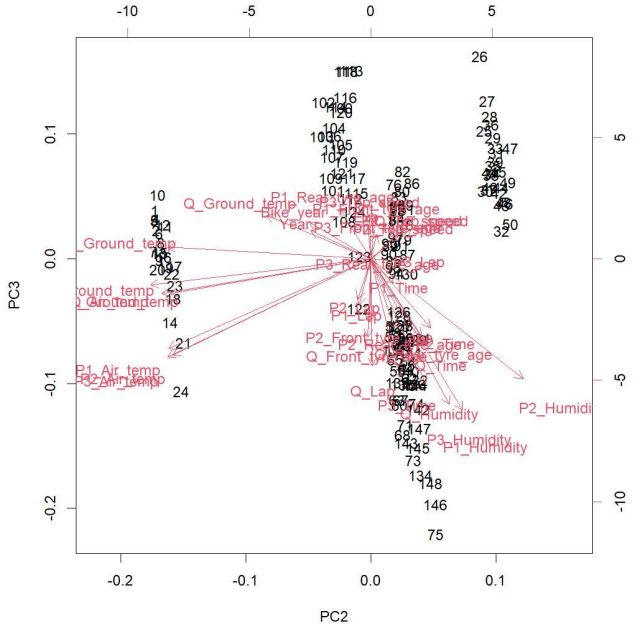


Figura 38: Biplot relativo alla seconda e alla terza componente principali

Da questo biplot capiamo che:

- le variabili relative alle umidità di tutte le sessioni, al numero di giri necessari a realizzare il proprio miglior tempo nelle qualifiche, nelle P1 e nelle P2, ai tempi sul giro di qualifiche, P3 e P2 (nascosto, sula destra del cluster di punti comprendente la 126esima osservazione) e ai consumi delle gomme di P2 e qualifiche, sono correlate negativamente con la terza componente;
- un gruppetto di variabili comprendente anche le annate di motociclette e di rilevazione, sono correlate positivamente con la terza componente, seppur non in maniera così influenzante come le variabili descritte in precedenza.

Anche in questo grafico notiamo che ci sono 5 cluster ben definiti, per lo stesso identico motivo per cui si formano anche in **figura 38**. Ciò lo dovrò, quindi, tenere in considerazione quando eseguirò la cluster analysis, al **capitolo 5**.

3.3 Factor analysis

L'analisi fattoriale, o factor analysis, è un altro metodo di riduzione della dimensionalità, simile alla PCA, che si basa sul cosiddetto **modello fattoriale**: quest'ultimo prevede che, data l'esistenza di un vettore aleatorio X di dimensione p*1, appartenente ad una distribuzione p-variata ($X \sim \mathrm{f_p}(\mu, \Sigma)$), questi dipenda linearmente da $F_1, ..., F_m$ *fattori latenti* (m arbitrario) e da altre p variabili $\epsilon_1, ..., \epsilon_p$, dette *fattori specifici*. Sia i fattori latenti che quelli specifici non sono osservabili. Il modello, quindi, diventa:

$$X - \mu = LF + \epsilon$$

dove $X - \mu$ è di dimensione p*1, L, matrice dei pesi di ogni variabile all'interno di ogni fattore latente, ha dimensione p*m, F ha dimensione m*1 ed ϵ ha dimensione p*1.
Il modello rispetta le seguenti assunzioni:

1: $E(F) = 0$, di dimensione m*1, mentre $Cov(F) = I$, di dimensione m*m;

2: $E(\epsilon) = 0$, di dimensione p*1, mentre $Cov(\epsilon) = \Psi = diag(\psi_1, .., \psi_p)$, di dimensione p*p;

3: $F \perp \epsilon$, quindi $Cov(F, \epsilon) = 0$ ed ha dimensione m*p.

Se vengono rispettate l'equazione del modello e le assunzioni **1** e **3** viene dato vita al **modello a fattori ortogonali**. Esso implica una struttura specifica per Σ, matrice di covarianza di X, ovvero:

$$\Sigma = LL' + \Psi$$

Inoltre da ciò ne consegue che $Cov(X, F) = L$.
In conseguenza alla formula di Σ, otteniamo quindi che la varianza di ogni variabile è ottenibile come:

$$V(X_i) = \Sigma_{ii} = l_{i1}^2 + ... + l_{im}^2 + \psi_i, i = 1, ..., p$$

I valori $l_{i1}^2 + ... + l_{im}^2 = h_i^2$ formano la cosiddetta *comunalità di una variabile*, ovvero la porzione di varianza di X_i associata agli m fattori comuni, mentre ψ_i è la *specificità di una variabile*, ovvero la porzione di varianza specifica di X_i.
Siccome m<<p, per stimare Σ correttamente si aggiunge al modello un'ortogonalizzazione, essendo esso non unico. Il modello, con l'aggiunta di una matrice ortogonale G di dimensione m*m, tale che $GG' = G'G = I$, diventerà quindi:

$$X - \mu = LGG'F + \epsilon = L^*F^* + \epsilon$$

dove $L^* = LG$ e $F^* = G'F$.
Per stimare quindi L e Ψ, dopo la quale stima posso selezionare la mia matrice G di rotazione (o "di ortogonalizzazione", come dirsi voglia) migliore, posso usare 2 metodi:

- **il metodo delle componenti principali** consiste nella stima di L attraverso le matrici E e Λ, matrici rispettivamente degli autovettori e degli autovalori di Σ. Dopo aver opportunamente deciso il numero m di fattori, andrò a selezionare i primi m valori sulla diagonale di Λ e le prime m colonne di E, per poi stimarwe L attraverso la formula $L = E\Lambda^{\frac{1}{2}}$. Di solito seleziono m quando gli ultimi p-m autovalori di Σ sono vicini a 0, oppure quando la percentuale di varianza spiegata dai primi m autovalori è pari a circa il 60-80 % di quella totale, quindi quando $\frac{\sum_{j=1}^{m}\lambda_j}{\sum_{j=1}^{p}\lambda_j} = 0.6-0.8$. Da lì vedo, per ogni fattore, quali sono le variabili con i loadings più alti e stimo Ψ, usando la formula $\Psi = \Sigma - LL'$.
  Se ho unità di misura diverse tra le variabili, oppure esse hanno varianze ben diverse tra loro, allora dovrò svolgere l'analisi standardizzando i dati e potrei quindi usare la matrice di correlazione R al posto di Σ, poiché quando i dati sono standardizzati $\Sigma = R$;

- **il metodo di massima verosimiglianza** assume che F ed ϵ siano distribuiti normalmente, implicando che quindi $X_j \sim \mathrm{f_p}(\mu, \Sigma)$, $\Sigma = LL' + \Psi$ e j=1,..,n (n = numerosità campionaria). Dovrò quindi stimare il parametro $\hat{\theta} = (\hat{\mu}, \hat{L}, \hat{\Psi})$, massimizzando la funzione di log-verosimiglianza:

$$l(\theta) = log \prod_{j=1}^{n} \phi_p(x_j; \mu, \Sigma)$$

  dove $\phi_p(x_j; \mu, \Sigma)$ è la funzione della distribuzione normale p-variata, quindi è pari asintoticamente a:

$$\phi_p(x_j; \mu, \Sigma) = -\frac{n}{2} log|\Sigma| - \sum_{j=1}^{n}(x_j - \mu)'\Sigma^{-1}(x_j - mu)/2$$

  Siccome il modello è soggetto a problemi di non unicità rispetto a rotazioni ortogonali di L, bisogna quindi aggiungere un **vincolo di unicità**:

$$L'\Psi^{-1}L = \Delta, \Delta matricediagonale$$

3.4 Applicazione della factor analysis al dataset con variabili numeriche

Nel caso specifico del dataset, quello con soli dati numerici, che stiamo esaminando, il determinante della matrice di covarianza, scritto come $|\Sigma|$ nella formula di $\phi_p()$, è pari a 0: pertanto non si può usare il metodo di massima verosimiglianza per stimare L e Ψ, bensì il metodo delle componenti principali. Nel nostro caso, avendo già svolto la PCA in precedenza basandoci sulla matrice di covarianza dei dati standardizzati, possiamo usare gli autovalori già trovati durante quell'analisi. Fermandoci ai primi 10 fattori, otteniamo il seguente risultato, già ottenuto al **paragrafo 3.3** durante la PCA:

Numero di fattori (m)	**Varianza λ_i**	$\frac{\lambda_i}{\sum_{j=1}^{p}\lambda_j}$	$\frac{\sum_{j=1}^{i}\lambda_j}{\sum_{j=1}^{p}\lambda_j}$
1	10.632	0.304	0.304
2	6.646	0.190	0.494
3	3.238	0.093	0.586
4	2.094	0.060	0.646
5	1.939	0.055	0.701
6	1.569	0.045	0.746
7	1.273	0.036	0.783
8	1.161	0.033	0.816
9	0.996	0.028	0.844
10	0.913	0.026	0.870

Per quanto riguarda la FA sembrerebbe essere più sensato prendere in considerazione i primi m=6 fattori, dato che con m=6 la varianza $\lambda_m = 1.569$ è pari ad una volta e mezza quello che è di base il valore medio di λ: già, perché la somma degli autovalori di una matrice (calcolabili solo se la matrice è quadrata, quindi con lo stesso numero di righe e colonne) è pari al suo numero di righe, quindi anche di colonne. Per cui il valore medio del vettore degli autovalori λ è pari a $\bar{\lambda} = \frac{\sum_{i=1}^{p}\lambda_i}{p} = \frac{p}{p} = 1$.

Fermandoci quindi a m=6 fattori, che spiegano poco più del 74 % della varianza totale, ottengo i seguenti valori di L, usando la formula $L = E\Lambda^{\frac{1}{2}}$:

Variabile	F1	F2	F3	F4	F5	F6
Year	0.897	0.256	0.120	0.012	-0.171	-0.022
Exp	-0.017	0.014	0.137	-0.173	0.256	-0.328
Bike_year	0.885	0.261	0.154	0.002	-0.140	-0.047
Q_Time	-0.756	-0.235	-0.365	0.214	-0.157	-0.038
P1_Time	-0.834	-0.095	-0.102	0.053	-0.316	-0.114
P2_Time	-0.833	-0.253	-0.292	0.085	-0.118	-0.047
P3_Time	-0.343	-0.122	-0.500	0.539	-0.326	-0.154
Q_Lap	-0.179	-0.006	-0.452	-0.627	-0.164	-0.174
P1_Lap	0.081	0.049	-0.195	-0.037	-0.318	0.601
P2_Lap	0.459	0.055	-0.170	-0.057	0.028	-0.039
P3_Lap	-0.731	-0.164	-0.013	-0.270	0.215	0.061
Q_Top_speed	0.822	-0.171	0.126	-0.166	-0.043	-0.030
P1_Top_speed	0.785	-0.140	0.126	-0.308	-0.018	-0.071
P2_Top_speed	0.785	-0.109	0.099	-0.233	-0.011	-0.237
P3_Top_speed	0.769	0.005	0.187	-0.239	-0.058	-0.091
Q_Front_tyre_age	-0.156	0.012	-0.331	-0.689	-0.238	-0.253
Q_Rear_tyre_age	-0.198	-0.212	-0.326	-0.338	0.120	0.093
P1_Front_tyre_age	-0.275	-0.018	0.166	-0.240	-0.732	0.126
P1_Rear_tyre_age	-0.310	0.129	0.206	-0.200	-0.711	0.143
P2_Front_tyre_age	-0.058	0.004	-0.269	0.008	-0.088	-0.604
P2_Rear_tyre_age	-0.384	-0.100	-0.293	-0.005	0.037	-0.353
P3_Front_tyre_age	0.677	-0.024	0.102	0.201	-0.320	-0.213
P3_Rear_tyre_age	0.712	-0.023	-0.021	0.306	-0.357	-0.182
Q_Humidity	0.740	-0.232	-0.528	0.222	-0.031	0.032
P1_Humidity	0.541	-0.387	-0.640	-0.101	0.090	0.218
P2_Humidity	0.306	-0.643	-0.508	0.058	0.041	-0.104
P3_Humidity	0.619	-0.333	-0.615	-0.075	0.055	0.122
Q_Air_temp	-0.308	0.883	-0.148	-0.148	0.074	0.014
Q_Ground_temp	-0.729	0.443	0.184	-0.081	0.061	-0.214
P1_Air_temp	0.312	0.854	-0.379	0.071	-0.003	-0.040
P1_Ground_temp	0.327	0.932	-0.110	0.037	-0.051	-0.012
P2_Air_temp	0.311	0.837	-0.408	-0.036	0.035	0.102
P2_Ground_temp	-0.285	0.850	-0.148	-0.189	0.109	0.199
P3_Air_temp	0.052	0.860	-0.416	0.098	0.059	0.092
P3_Ground_temp	-0.065	0.859	0.050	0.205	-0.081	-0.307

Dalla tabella dei loadings capisco quali sono le variabili più influenzanti per ogni fattore (prenderemo in considerazione una variabile quando il valore assoluto del suo loading, nel relativo fattore, sarà maggiore di 0.5): quando un loading è negativo, vuol dire che la variabile è correlata negativamente con il fattore:

- **Primo fattore:** molta influenza da parte delle variabili relative alle annate, sia di rilevazione che della moto, da parte delle velocità massime in tutte le sessioni, dei consumi delle gomme nelle P3 e di tutte le umidità, meno quelle rilevate nelle P2, correlate positivamente.
  Inoltre molta influenza anche da parte dei tempi sul giro di tutte le sessioni e della temperatura dell'asfalto in qualifica, correlate invece negativamente;

- **Secondo fattore:** molta influenza da parte delle variabili relative alle temperature dell'aria e dell'asfalto, rilevate in tutte le sessioni, tranne la temperatura dell'asfalto in qualifica; con esse c'è una correlazione positiva.
  L'unica variabile alquanto influente e correlata negativamente con il fattore è l'umidità registrata nelle P2;

- **Terzo fattore:** le variabili che più influenzano il fattore sono tutte correlate negativamente e sono le umidità in tutte le sessioni ed il tempo sul giro nelle P3;
- **Quarto fattore:** influenzato principalmente dalle variabili relative all'usura della gomma anteriore in qualifica e dal numero di giri necessari per realizzare il proprio miglior tempo nelle qualifiche, correlate negativamente. Il fattore è poi correlato positivamente dal tempo sul giro nelle P3;
- **Quinto fattore:** influenzato principalmente dalle usure degli pneumatici, sia anteriore che posteriore, nelle P1, correlate negativamente;
- **Sesto fattore:** influenzato principalmente dalla variabile relativa al numero di giri necessari per realizzare il proprio miglior tempo nelle P1, correlata positivamente, oltre che dall'usura del pneumatico anteriore nelle P2, correlata negativamente.

Dopo quest'analisi, svolta attraverso l'osservazione dei loadings, possiamo quindi calcolarci le comunalità ($h_i^2 = l_{i1}^2 + ... + l_{im}^2$) e le specificità ($\psi_i = \Sigma_{ii} - h_i^2$), per ogni variabile (i=1,..,p, m=5). Come valori di Σ_{ii} userò sempre 1, avendo svolto l'analisi con la matrice dei dati standardizzati, quindi otterrò che ogni $\psi_i = 1 - h_i^2$:

Variabile	**comunalità** h_i^2	**specificità** ψ_i
Year	0.914	0.086
Exp	0.222	0.778
Bike_year	0.897	0.103
Q_Time	0.832	0.168
P1_Time	0.831	0.169
P2_Time	0.867	0.133
P3_Time	0.803	0.197
Q_Lap	0.687	0.313
P1_Lap	0.511	0.489
P2_Lap	0.248	0.752
P3_Lap	0.684	0.316
Q_Top_speed	0.751	0.249
P1_Top_speed	0.752	0.248
P2_Top_speed	0.748	0.252
P3_Top_speed	0.695	0.305
Q_Front_tyre_age	0.729	0.271
Q_Rear_tyre_age	0.328	0.672
P1_Front_tyre_age	0.713	0.287
P1_Rear_tyre_age	0.721	0.279
P2_Front_tyre_age	0.448	0.552
P2_Rear_tyre_age	0.369	0.631
P3_Front_tyre_age	0.657	0.343
P3_Rear_tyre_age	0.762	0.238
Q_Humidity	0.931	0.069
P1_Humidity	0.918	0.082
P2_Humidity	0.781	0.219
P3_Humidity	0.896	0.104
Q_Air_temp	0.924	0.076
Q_Ground_temp	0.818	0.182
P1_Air_temp	0.977	0.023
P1_Ground_temp	0.992	0.008
P2_Air_temp	0.977	0.023
Variabile	**comunalità** h_i^2	**specificità** ψ_i
P2_Ground_temp	0.913	0.087
P3_Air_temp	0.937	0.063
P3_Ground_temp	0.887	0.113

Da questa tabella possiamo capire che le variabili relative all'esperienza del pilota, al numero di giri necessari a svolgere il proprio miglior tempo nelle P2, alle usure degli pneumatici posteriori delle qualifiche ed alle usure degli pneumatici anteriori e posteriori delle P2, non condividono una grossa porzione di varianza con i fattori comuni F, bensì con i fattori specifici ϵ_i.

4 CCA

Supponiamo di voler verificare se ci sono relazioni tra le performance, le condizioni di pista e le condizioni meteo in qualifica e quelle invece registrate nelle prove. Creiamo quindi i due dataset *quali.q*, relativo alle rilevazioni delle qualifiche, e *quali.p*, relativo alle rilevazioni delle prove, derivati entrambi dal dataset di sole variabili numeriche e svolgiamo un'**analisi delle correlazioni canoniche**, anchesì detta **CCA**.

4.1 Cos'è la CCA?

La CCA è un tipo di analisi che si occupa di scovare i vettori a_i e b_i che massimizzano la correlazione tra le due trasformazioni lineari, dalla varianza unitaria, $W_i = X1 * a_i$ e $Z_i = X2 * b_i$, con $X1$ e $X2$ "sotto-dataset" di rispettivamente $p1$ e $p2$ variabili, derivati dal dataset principale X ($i = 1, .., p$, $p = 1, .., min(p1, p2)$).
Date le matrici di varianze e covarianze $var(X1) = S_{11}$, $var(X2) = S_{22}$ e $cov(X1, X2) = S_{12}$, la correlazione tra W e Z da massimizzare è la seguente:

$$r_{iWZ} = \frac{a_i' S_{12} b_i}{\sqrt{a_i' S_{11} a_i}\sqrt{b_i' S_{22} b_i}}$$

per la quale deve essere rispettato il vincolo $a_i' S_{11} a_i = b_i' S_{12} b_i = 1$.
Per cui il valore da massimizzare definitivamente sarà $a_i' S_{12} b_i$.
Si è riuscito a dimostrare come i vettori a_i e b_i siano derivati dalla decomposizione spettrale di $C = S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}}$: dopo aver effettuato la decomposizione spettrale di C, ovvero

$$C = UDV'$$

con U matrice degli autovettori di CC', D matrice diagonale degli autovalori di CC' (o $C'C$, gli autovalori sarbbero uguali per entrambi) e V matrice degli autovettori di $C'C$, pongo le prime colonne di U e V rispettivamente pari ai vettori $a0_1$ e $b0_1$; i valori, quindi, di a_1 e b_1 li ricavo con le formule

$$a_1 = S_{11}^{-\frac{1}{2}} a0_1$$

$$b_1 = S_{22}^{-\frac{1}{2}} b0_1$$

I vettori $a_2, .., a_p$ e $b_2, .., b_p$, li ricavo ponendo le successive colonne di U e V pari ad $a0_2, .., a0_p$ e $b0_2, .., b0_p$, trascrivendole nelle formule al posto di $a0_1$ e $b0_1$.
Sostituendo a_i e b_i nella formula del valore da massimizzare, ottengo che quest'ultimo sarà definitivamente:

$$a0_i' S_{11}^{-\frac{1}{2}} S_{12} S_{22}^{-\frac{1}{2}} b0_i$$

per il quale dovrà essere rispettato il vincolo $a0_i' a0_i = b0_i' b0_i = 1$.
Tornando alla formula $C = UDV'$, decomposizione spettrale di C, la diagonale della matrice $D = diag(\lambda_1, .., \lambda_k)$ è composta dai quadrati delle correlazioni canoniche tra W e Z. Quindi la correlazione r_{iWZ} tra $W_i = X1 * a_i$ e $Z_i = X2 * b_i$, che da ora in poi chiameremo ρ_i, sarà pari a $\sqrt{\lambda_i}$.
Di solito non tutte le p correlazioni canoniche sono significative; per verificare ciò, si esegue un test, dove l'ipotesi nulla è la seguente:

$$H0 : \rho_1 \neq 0, ... = \rho_k \neq 0, \rho_{k+1} = ... = \rho_p = 0$$
$$H1 : \rho_i \neq 0, i \geq k + 1$$

Per verificare tale ipotesi, dovrò calcolarmi la statistica test

$$w_B^k = -(n - 1 - \frac{1}{2}(p1 + p2 + 1))log \prod_{i=k+1}^{p} (1 - \hat{\rho_i}^2)$$

con n pari alla numerosità campionaria dei dataset $X1$ e $X2$ (che, logicamente, sarà la stessa, provenendo essi dallo stesso dataset X). Fissato un livello di significatività α, che noi porremo pari a 0.05, rifiuteremo l'ipotesi nulla $H0$ se $w_B^k > \chi^2_{(p1-k)(p2-k)}(1 - \alpha)$.

4.2 Applicazione al dataset con variabili numeriche

Dai due dataset *quali.q* e *quali.p* ho rimosso le variabili relative alle temperature dell'aria, poiché linearmente dipendenti da altre variabili: come possiamo in effetti notare dal corrplot in **figura 2**, al **paragrafo 2.2**, esse sono molto correlate tra loro e con le variabili relative alle temperature dell'asfalto, per cui possiamo ipotizzare che siano dipendenti linearmente da loro. *quali.q* presenterà quindi $p1 = 7$ variabili, mentre *quali.p* ne conterrà $p2 = 21$.
Il vettore ρ delle correlazioni canoniche è risultato essere il seguente:

$$\rho = [0.999; 0.971; 0.936; 0.836; 0.500; 0.339; 0.180]$$

Tra queste $p = 7$ correlazioni sono risultate significative solo le prime $k = 4$:

k	w_B^k	$\chi^2_{(p1-k)(p2-k)}(0.95)$
1	1694.287	0
2	872.168	24.996
3	493.271	46.194
4	217.339	68.669
5	58.619	92.808

Trarremo le nostre conclusioni, di conseguenza, in 4 dimensioni. A noi, più che i vettori canonici, interessano le correlazioni, oltre che tra W e Z (contenute nel vettore ρ), anche tra X e W e tra Y e Z. Le correlazioni in questione sono le seguenti:

	W_1	W_2	W_3	W_4
Q_Time	-0.315	-0.410	0.847	-0.014
Q_Lap	-0.003	-0.170	0.175	-0.050
Q_Top_speed	0.505	0.451	-0.310	-0.655
Q_Front_tyre_age	-0.066	-0.117	0.139	-0.194
Q_Rear_tyre_age	-0.031	0.092	0.312	0.115
Q_Humidity	0.999	0.043	0.032	0.003
Q_Ground_temp	-0.774	-0.611	-0.156	-0.011

	Z_1	Z_2	Z_3	Z_4
P1_Time	-0.553	-0.395	0.602	-0.028
P2_Time	-0.422	-0.362	0.799	-0.021
P3_Time	0.145	-0.442	0.743	-0.043
P1_Lap	0.130	0.133	0.159	0.060
P2_Lap	0.405	0.005	-0.309	0.135
P3_Lap	-0.587	-0.180	0.318	0.066
P1_Top_speed	0.442	0.388	-0.359	-0.588
P2_Top_speed	0.452	0.337	-0.349	-0.636
P3_Top_speed	0.382	0.413	-0.405	-0.612
P1_Front_tyre_age	-0.286	-0.004	0.138	-0.179
P1_Rear_tyre_age	-0.343	-0.079	0.032	-0.036
P2_Front_tyre_age	0.077	-0.258	0.075	-0.070
P2_Rear_tyre_age	-0.158	-0.279	0.274	0.012
P3_Front_tyre_age	0.461	0.247	-0.239	-0.197
P3_Rear_tyre_age	0.605	0.218	-0.174	-0.091
P1_Humidity	0.795	0.065	0.177	-0.150
P2_Humidity	0.672	-0.222	0.231	-0.184
P3_Humidity	0.834	-0.018	0.058	-0.209
P1_Ground_temp	0.118	-0.242	-0.564	0.021
P2_Ground_temp	-0.390	-0.291	-0.228	0.093
P3_Ground_temp	-0.165	-0.587	-0.514	0.074

Cosa possiamo capire dalle seguenti tabelle? Semplicemente il legame tra le variabili del dataset *quali.q* e quelle del dataset *quali.p*, attraverso le concordanze e le discordanze nei segni delle correlazioni:

- 1^a **dimensione**: c'è una correlazione positiva elevata tra le umidità in qualifica e quelle registrate in tutte le sessioni di prove libere, così come tra la temperatura dell'asfalto in qualifica e le variabili relativa al tempo sul giro nelle P1 ed al numero di giri necessari per far registrare il proprio miglior tempo nelle P3.
  Tra i 2 gruppi di variabili appena descritti c'è la correlazione negativa più elevata, soprattutto tra l'umidità in qualifica e la variabile P3_Lap: quindi possiamo dire che più è elevata l'umidità del sabato (comprendendo quindi anche quella delle P3, correlata anch'essa negativamente a P3_Lap) e prima i piloti fanno registrare il proprio miglior tempo nelle P3;

- 2^a **dimensione**: le correlazioni positive più elevate sono quelle tra le velocità massime, in tutte le sessioni, e quelle tra le temperature dell'asfalto delle qualifiche e delle P3, il che è alquanto attendibile, svolgendosi le 2 sessioni nello stesso giorno.
  Tra questi 2 gruppi, invece, soprattutto tra la temperatura dell'asfalto in qualifica e le velocità massime in prova, ci sono le correlazioni negative più elevate: questo vuol dire che le moto, soprattutto al sabato, sono più veloci in rettilineo, in qualifica, quando le temperature dell'asfalto in prova sono più basse;

- 3^a **dimensione**: balzano all'occhio le correlazioni positive tra i tempi sul giro, di tutte le sessioni, quelle tra tutte le velocità di punta e anche quelle tra tutte le temperature dell'asfalto.
  Tutte le variabili di questi ultimi 2 gruppi sono correlate, bene o male, negativamente con tutte le variabili relative ai tempi sul giro, per cui possiamo affermare che avere delle temperature dell'asfalto elevate e una moto veloce in rettilineo, entrambe in prova, implicano tendenzialmente un tempo sul giro in qualifica più basso, quindi una probabilità più elevata di partire da una posizione più importante, sulla griglia di partenza, nelle gare di sabato e domenica;

- 4^a **dimensione**: l'ultima dimensione significativa conferma ancora una volta la correlazione positiva elevata tra tutte le velocità massime in rettilineo, mentre sembrerebbe mostrare una correlazione negativa, seppur bassa, tra la velocità di punta in qualifica e la variabile P2_Lap: questo, quindi, sembrerebbe implicare che i piloti con la moto più veloce in rettilineo, in qualifica, tendono a far registrare il proprio miglior tempo nelle P2 prima degli altri, indice che essi quindi arrivano prima al limite nelle P2 rispetto ai piloti con moto meno performanti in rettilineo, come Yamaha e Suzuki (guardare grafici nelle **figure 15**, **16**, **17** e **18**, presenti al **paragrafo 2.4**).

Le tabelle delle correlazioni appena spiegate, possono essere poi rappresentate nei seguenti 2 grafici; il primo si riferisce alle prime 2 dimensioni, il secondo alla terza e alla quarta:

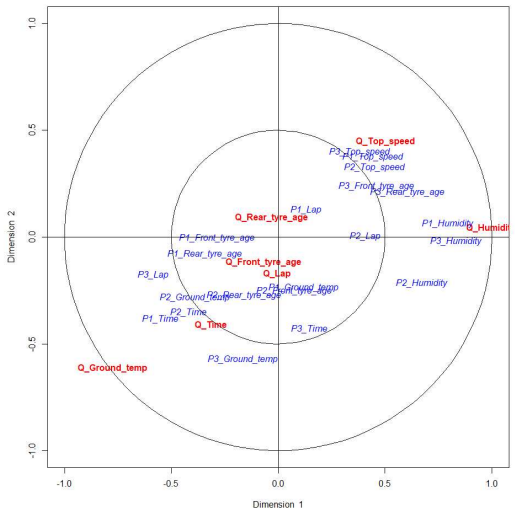


Figura 39: Grafico relativo alla CCA, nelle dimensioni 1 e 2

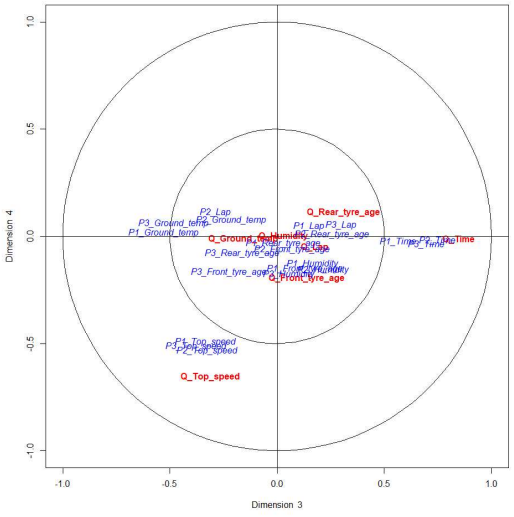


Figura 40: Grafico relativo alla CCA, nelle dimensioni 3 e 4

Vediamo, poi, i grafici dei punteggi delle combinazioni lineari W e Z nelle prime 4 dimensioni:

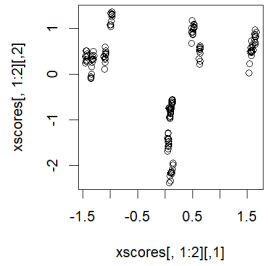


Figura 41: Grafico relativo ai pounteggi di W, nelle dimensioni 1 e 2

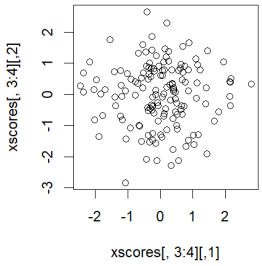


Figura 42: Grafico relativo ai pounteggi di W, nelle dimensioni 3 e 4

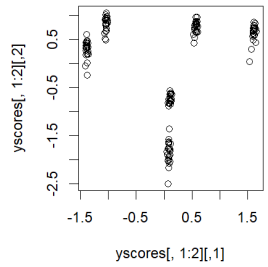


Figura 43: Grafico relativo ai pounteggi di Z, nelle dimensioni 1 e 2

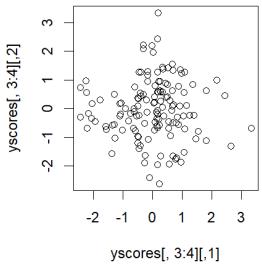


Figura 44: Grafico relativo ai pounteggi di Z, nelle dimensioni 3 e 4

Da questi grafici noto un dettaglio molto importante, già notato anche nei biplot della PCA (**figure 37** e **38**, al **paragrafo 3.2**): nelle prime 2 dimensioni si formano, sia per i punteggi di W che per i punteggi di Z, almeno 5 o 6 cluster di unità ben definiti. Ciò è dato dal fatto che tra le variabili incluse sono presenti quelle atmosferiche, relative alle umidità ed alle temperature dell'asfalto, che variano solo di anno in anno, che è lo stesso motivo per cui si formano i cluster nei biplot delle componenti principali. Da queste considerazioni, quindi, decido che per la cluster analysis eliminerò le variabili relative all'annata, sia di rilevazione che della motocicletta, assieme a tutte le variabili atmosferiche numeriche, che variano solo al variare dell'annata.

5 Cluster Analysis

5.1 Cos'è il processo di clustering?

Il clustering, ovvero il raggruppamento delle unità statistiche (dall'inglese "cluster" = "gruppo", "ammasso"[16]), può essere di 2 tipi:

- **clustering con metodi gerarchici**: si basano sulla **distanza di Minkowski** tra unità statistiche, la quale, date 2 unità x e y, parte di base come:

$$d_m(x,y) = \sum_{h=1}^{p} [|x_h - y_h|^m]^{\frac{1}{m}}$$

  Se $m = 1$ la distanza diventa pari a quella di **Manhattan**, dove $d(x,y) = \sum_{h=1}^{p} |x_h - y_h|$; se $m = 2$, invece, diventa pari a quelle **euclidea**, dove $d(x,y) = \sum_{h=1}^{p} \sqrt{|x_h - y_h|^2}$; se invece $m = \infty$, si avrà a che fare con la distanza di **Lagrange**, dove $d(x,y) = \max_{h=1,..,p} |x_h - y_h|$.
  Dopo aver stabilito, inizialmente, la presenza di n gruppi (n=numero di unità statistiche) ed il tipo di distanza tra unità che si vuole usare, si sceglie opportunamente quale metodo di clustering gerarchico voler utilizzare per raggrupparle tra loro. Ne esistono principalmente 4:

  - **clustering con legame singolo**: dati 2 gruppi G_1 e G_2, di numerosità rispettivamente n_{G_1} e n_{G_2}, la distanza tra essi è pari alla distanza tra le unità dei 2 cluster più vicine tra loro:
$$d(G_1, G_2) = min(d(x_i, x_j) : x_i \in G_1, x_j \in G_2)$$
  - **clustering con legame completo**: dati 2 gruppi G_1 e G_2, di numerosità rispettivamente n_{G_1} e n_{G_2}, la distanza tra essi è pari alla distanza tra le unità dei 2 cluster più lontane tra loro:
$$d(G_1, G_2) = max(d(x_i, x_j) : x_i \in G_1, x_j \in G_2)$$
  - **clustering con legame medio**: dati 2 gruppi G_1 e G_2, di numerosità rispettivamente n_{G_1} e n_{G_2}, la distanza tra essi è pari alla distanza media tra le unità dei 2 cluster:
$$d(G_1, G_2) = \frac{1}{n_{G_1} n_{G_2}} \sum_{x_i \in G_1} \sum_{x_j \in G_2} d(x_i, x_j)$$
  - **clustering con legame di Ward**: metodo che sceglie i cluster in base a quelli che perdono meno informazione, definita come la somma delle devianze interne $ESS_{tot} = \sum_{j=1}^{k} ESS_j$, dove $ESS_j = \sum_{x_i \in G_j} (x_i - \bar{x}_{G_j})'(x_i - \bar{x}_{G_j})$. $\bar{x_{G_j}}$ è pari alla media delle osservazioni nel gruppo G_j.

  Ovviamente questi metodi possono essere utilizzati per raggruppare più di $k = 2$ gruppi.
  Il processo di clustering con metodi gerarchici è un processo verificabile anche visivamente, attraverso l'uso di un grafico detto **dendrogramma**, che vedremo bene al **paragrafo 5.2**.

- **clustering con metodi non gerarchici**: non si basano sulle distanze tra ogni unità, ma su altri criteri. Il più noto di tutti è il **clustering con il metodo delle k-medie**. Questo metodo parte dalla creazione di un numero k di gruppi a priori; dopodiché, per ogni gruppo, viene calcolato un centroide (le medie per ogni variabile) e si calcola la distanza, di solito euclidea, tra ogni unità ed il centroide di ogni gruppo. Dopodiché si riassegna ogni unità al cluster (o gruppo, come dirsi voglia) con il centroide più vicino. Si ripete il processo fino a quando non si vedono più cambiamenti nei cluster.

Per verificare l'efficienza di un processo di clustering, si può verificare la cosiddetta **silhouette** di ogni unità: date, per l'i-esima unità, le distanze $a(i)$, ovvero quella media tra l'unità e gli altri punti del

[16] Wordreference, `https://www.wordreference.com/enit/cluster`

suo cluster, e $b(i)$, ovvero quella media tra l'unità e i punti del cluster più vicino al suo, la silhouette si definisce come:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$s(i)$ è compreso tra -1 e 1:

- se tende ad 1 allora l'unità è stata ben assegnata al cluster;
- se tende a 0 allora è al confine tra 2 cluster;
- se tende a -1 è stata assegnata al cluster sbagliato.

Infine, se ho a che fare con dataset che contengono variabili che hanno diverse unità di misura, come nel caso del nostro dataset con variabili numeriche (sul quale baseremo l'analisi) è meglio effettuare l'operazione di clustering con il dataset standardizzato.
E' esattamente quello che andremo a fare noi, con l'aggiunta del fatto che le distanze, così come le k-medie, verranno calcolate sul dataset numerico con la rimozione delle variabili relative alle annate, sia di rilevazione che delle motociclette (quindi **Year** e **Bike_year**), assieme a tutte le variabili relative alle umidità e alle temperature di aria e asfalto che variano solo al variare dell'annata: questo perché, come spiegato anche alla fine del **capitolo 4**, sebbene siano variabili quantitative, hanno comunque frequenze assolute elevate per ogni loro valore, il che rischierebbe quindi di fare in modo che il clustering si basi solamente su di esse.
Il nostro obiettivo finale sarà quello di riuscire a raggruppare al meglio le nostre unità in k=3 gruppi, per riuscire a scovare bene le variabili che meglio spiegano il clustering, senza però semplificare troppo l'analisi.

5.2 Cluster analysis con metodi gerarchici

Cominciamo dai processi di **clustering con legame singolo** e vediamo attraverso i dendrogrammi quale tipo di distanza esprime meglio il processo, **a partire da quella euclidea**:

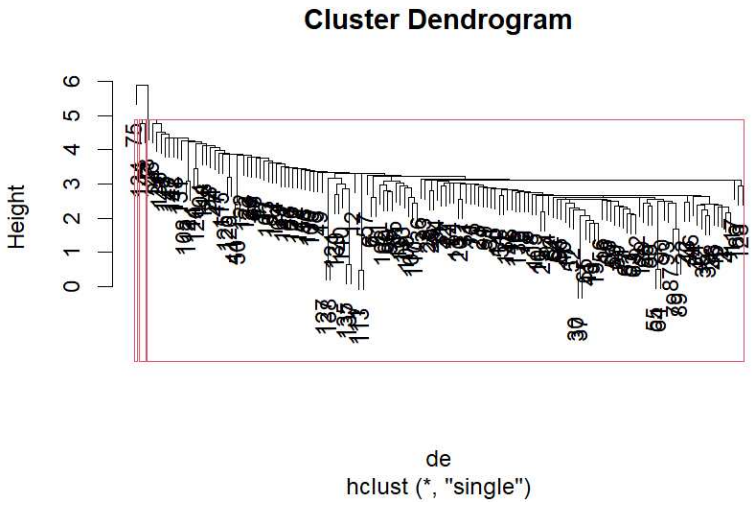


Figura 45: Dendrogramma di un clustering con legame singolo basato su distanza euclidea

Ma a cosa serve e come si legge un dendrogramma? Il dendrogramma dà un'idea visiva di quanti cluster si possono creare e in che maniera vanno a crearsi, ovvero quali unità entrano per prime nei vari clsuter. Tendenzialmente meglio le unità sono distribuite in tutto il grafico e più sono distribuite anche all'interno dei cluster.
Nel caso, quindi, del grafico in **figura 45** possiamo capire che, vedendo che le unità sono quasi tutte distribuite nel terzo gruppo, non possiamo dire che si tratti di un buon metodo di clustering.

Vediamo cosa succede se uso la **distanza di Manhattan**:

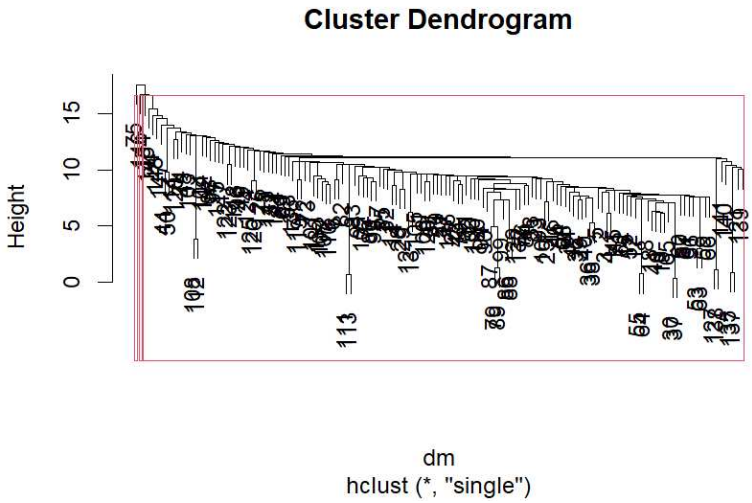


Figura 46: Dendrogramma di un clustering con legamme singolo basato su distanza di Manhattan

Usando la distanza di Manhattan non sembrerebbe esserci una variazione rispetto a quella euclidea, segno che forse il legame singolo non è il miglior metodo di clustering.
Vediamo infine cosa succede se uso la **distanza di Lagrange**:

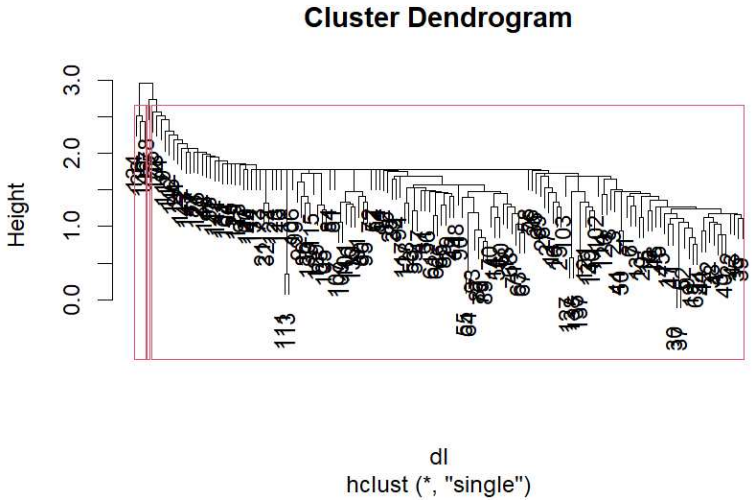


Figura 47: Dendrogramma di un clustering con legame singolo basato su distanza di Lagrange

Le conclusioni sembrerebbero essere le stesse degli altri 2 clustering, ovvero che le unità sono mal distribuite. Quindi potrei affermare che il clustering con legame singolo non funziona per raggruppare le mie unità in 3 gruppi. Per confermare ciò, conviene quindi effettivamente vedere i cluspolot, ovvero i grafici dei gruppi nelle prime 2 componenti principali, dove ogni gruppo è rappresentato da un'ellisse e più le ellissi sono distinguibili tra loro, meglio è stato effettuato il clustering. Da notare che le prime 2 componenti principali spiegano circa il 44 % della varianza totale, per cui, nel nostro caso, non sono grafici che spiegano tutto e al meglio, ma possiamo comunque usarli per farci ancor di più un'idea sul lavoro che svolgono i processi di clustering.

Il risultato sarebbe il seguente:

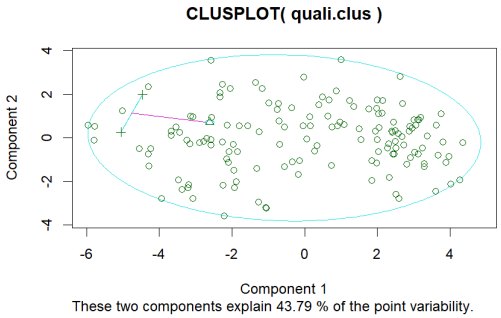


Figura 48: Clusplot del clustering con legame singolo basato su distanze euclidee

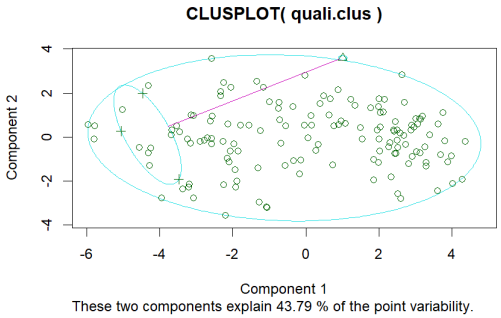


Figura 49: Clusplot del clustering con legame singolo basato su distanze di Lagrange

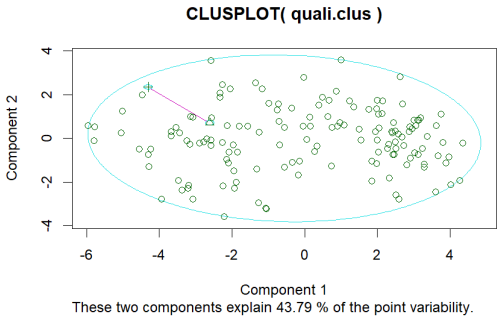


Figura 50: Clusplot del clustering con legame singolo basato su distanze di Manhattan

Tutti e 3 i clsuplot mostrano un gruppone che contiene gli altri 2 gruppetti, quindi confermo che i clustering con legame singolo non sono dei buoni metodi di raggruppamento. Vediamo cosa succede se uso dei metodi di **clustering con legame di tipo completo**, a partire da quello basato sulle **distanze euclidee**:

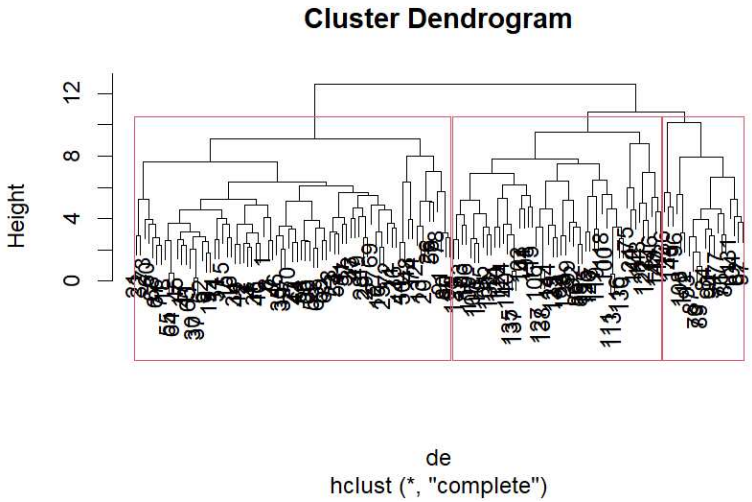
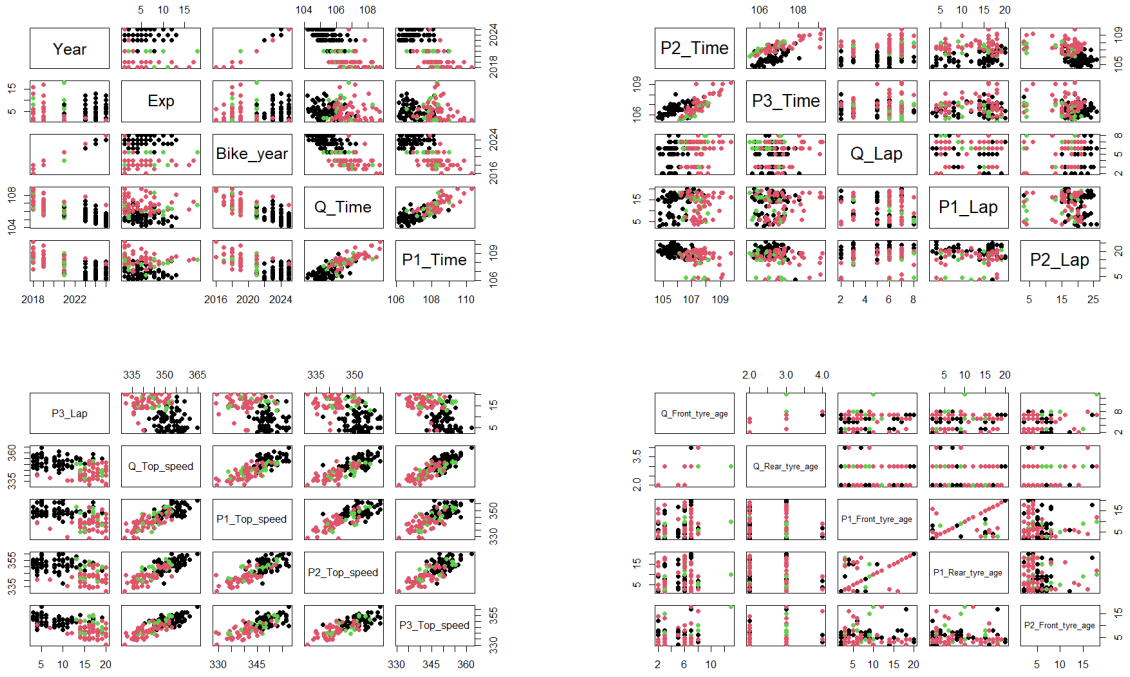
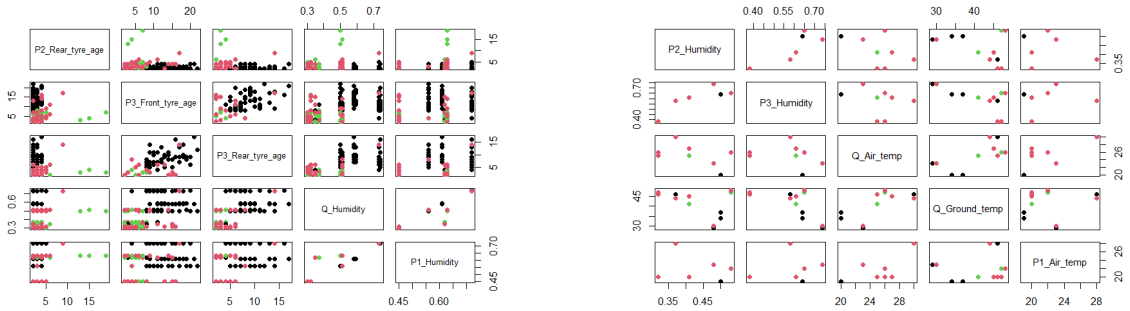


Figura 51: Dendrogramma di un clustering con legame completo basato, su distanze euclidee

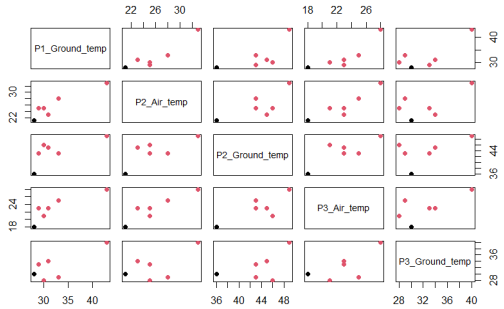
Finalmente iniziamo a vedere un grafico più bilanciato. A primo impatto, le unità sembrano distribuirsi secondo qualche variabile, vediamo cosa ci dicono i grafici di dispersione:



Tutte le prime 10 variabili fanno un discreto, ma non eccellente, lavoro nello spiegare i 3 cluster. Le variabili Year e Bike_year, fanno fatica a far distinguere bene le unità del cluster meno numeroso (di colore verde nei grafici), mentre spiegano molto meglio quelli con numerosità più elevata. Lo stesso vale per le variabili relative alle top speed, così come la variabile P3_Lap.



A sorpresa sono le variabili relative ai consumi delle gomme posteriori nelle P2 e nelle P3 a fare in modo che le unità del terzo gruppetto siano distinguibili, oltre che continuare a spiegare bene i 2 cluster più numerosi. Tutte le altre non ci dicono quasi niente, invece.



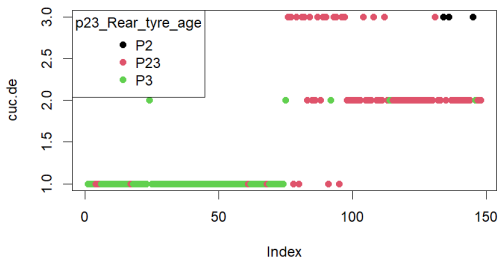
Addirittura le ultime variabili atmosferiche non spiegano nemmeno il terzo cluster.
Quindi, potremmo dire che questo clustering esegua un buon lavoro nel distinguere le unità in base ai consumi degli pneumatici posteriori. Creiamo quindi la variabile **p23_Rear_tyre_age**, che avrà valori:

- **P2** quando P2_Rear_tyre_age ≥ 9 e P3_Rear_tyre_age < 6, ovvero il pilota consuma molto di più gli pneumatici posteriori nelle P2, piuttosto che nelle P3;
- **P23** quando P2_Rear_tyre_age ≤ 9 e P3_Rear_tyre_age < 6, ovvero il pilota non gira con pneumatici consumati in bene o male nessuna sessione, tra P2 e P3;
- **P3** quando P2_Rear_tyre_age > 9 e P3_Rear_tyre_age ≥ 6, ovvero il pilota consuma molto di più gli pneumatici nelle P3, piuttosto che nelle P2.

La tabella di frequenze incrociate presenta i seguenti valori:

Cluster	**P2**	**P23**	**P3**	**totale**
1	0	9	68	77
2	0	46	5	51
3	3	17	0	20
totale	3	72	73	148

Vengono mal classificate 9+17+5 = 31 unità, circa il 20.9 % delle unità totali, che non sono proprio poche. Se poi effettuiamo anche un'analisi grafica capiamo che potrebbero essere usati criteri migliori, per classificare le unità, più che altro quelle del gruppo meno numeroso:



Verifichiamo ulteriormente, attraverso un test d'indipendenza ed il calcolo di un valore fondamentale, detto Rand Index, se il clustering è significativamente dipendente da quella variabile: il test, usato per verificare l'ipotesi nulla $H0 : cluster \perp p23_Rear_tyre_age$, utilizza la statistica test

$$X^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

dove O_{ij} è la frequenza osservata sull'i-esima riga alla j-esima colonna ed E_{ij} è invece la frequenza osservata, calcolata come $E_{ij} = \frac{(n_{i\cdot})(n_{\cdot j})}{n}$, ovvero la media tra l'i-esimo totale di riga ed il j-esimo totale di colonna. X^2 si distribuisce come una χ^2 con gradi di libertà $(r-1)(c-1)$, con r numero di righe e c numero di colonne. $H0$ viene riifiutato quando $X^2 > \chi^2_{(r-1)(c-1)}$.
Nel nostro caso X^2 ha un valore pari a 114.09, con gradi di libertà (3-1)(3-1) = 4, a fronte di un valore di χ^2_4 pari a 9.488. Rifiuterò quindi $H0$, con un p-value finale del test minore di $2.2*10^{-16}$, e affermerò che c'è una forte evidenza statistica della dipendenza tra i 3 cluster ed i consumi degli pneumatici posteriori nelle P2 e nelle P3.
L'altro indice che possiamo utilizzare per verificare la somiglianza tra clustering e variabile, ovvero il **Rand Index**, si basa sul conteggio delle coppie di elementi e su come queste vengono assegnate ai cluster. Per calcolarlo abbiamo bisogno di:

- **a**: numero di coppie assegnate allo stesso cluster in entrambe le partizioni;
- **b**: numero di coppie assegnate a cluster diversi in entrambe le partizioni;
- **c**: coppie assegnate allo stesso cluster in una sola delle due partizioni;
- **d**: coppie assegnate a cluster diversi in una sola delle due partizioni.

La formula finale sarà:

$$RI = \frac{a+b}{a+b+c+d} = \frac{a+b}{\binom{n}{2}}$$

Più il Rand Index sarà vicino ad 1 e più ci sarà somiglianza tra la variabile ed il clustering.
Nel nostro caso viene calcolato un valore pari a 0.521, quindi più vicino ad 1 che a 0, ma non di troppo. Quindi, per quanto p23_Rear_tyre_age sia significativamente buona nel dividere le mie unità in 3 gruppi, non sembra essere il miglior criterio.
Vediamo, quindi, cosa ci dicono i grafici delle silhouette e il clusplot:

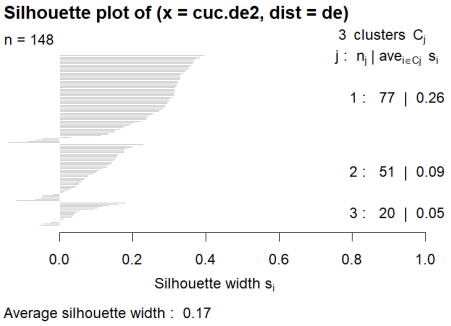


Figura 52: Grafico delle silhouette del clustering con legame completo, basato su distanze euclidee

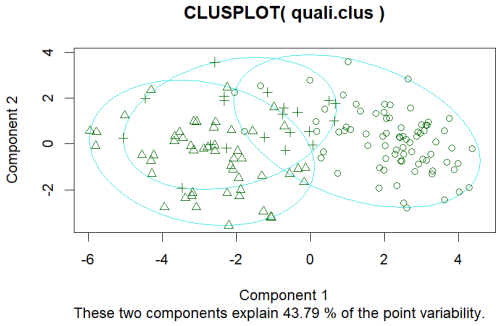


Figura 53: Clusplot delle unità raggruppate con un metodo di clustering con legame completo, basato su distanze euclidee

Le unità con silhouette negativa sono molte meno rispetto a quelle mal classificate con la variabile prima spiegata (circa la metà), oltre che essere distribuite diversamente nei 3 gruppi; per cui posso dire che, anche a fronte di un clusplot dove i 3 gruppi non sono distinguibili in maniera netta, questo metodo di clustering, in generale, **non è il migliore se voglio dividere le mie unità in 3 gruppi**.

Vediamo quindi cosa succede se uso un metodo di **clustering con legame di tipo completo basato sulle distanze di Manhattan**:

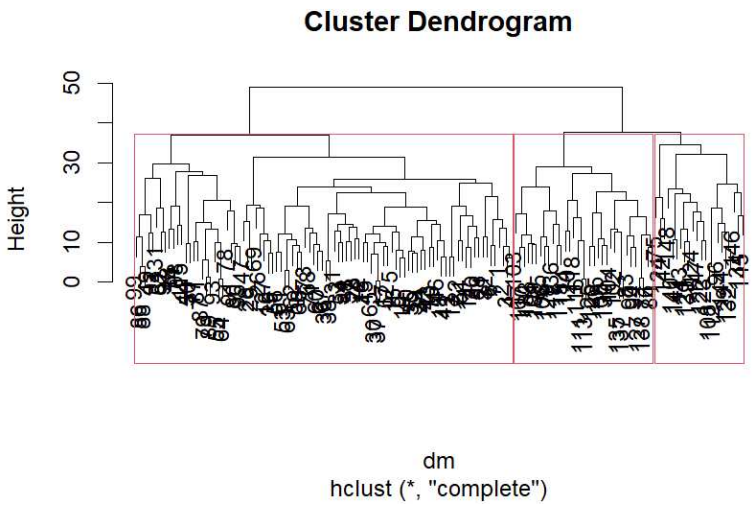
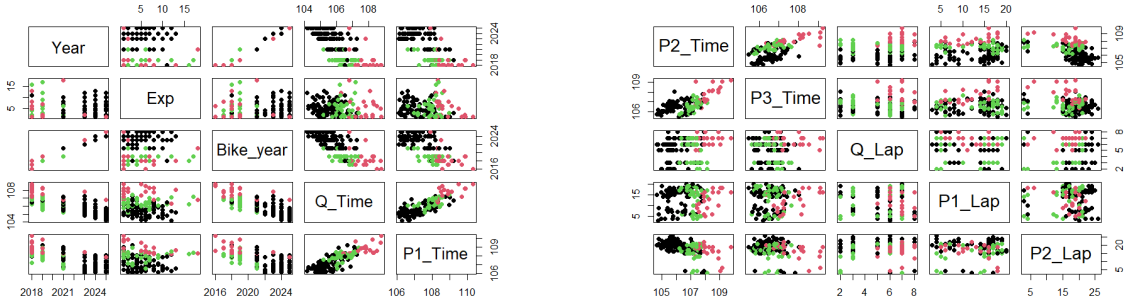
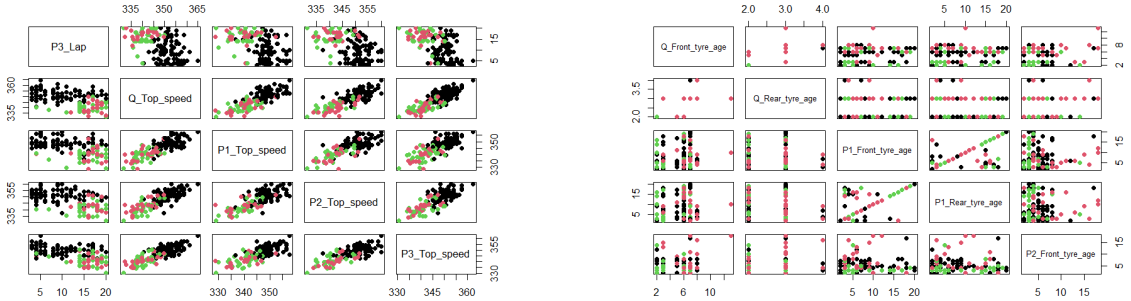


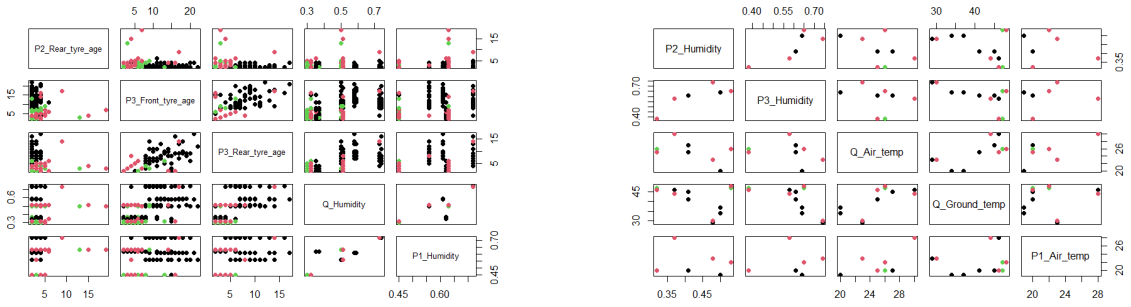
Figura 54: Dendrogramma di un clustering con legame completo, basato su distanze di Manhattan

A grandi linee sembra, a primo impatto, molto simile al clustering visto in precedenza. Conviene guardare quindi i grafici di dispersione per capire quale/i variabile/i meglio spiegano il clustering:

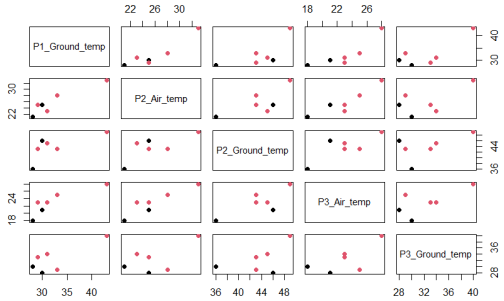


Questo clustering sembrerebbe basarsi principalmente sui tempi sul giro e sull'annata della moto. Soprattutto la combinazione tra Bike_year e Q_Time è molto interessante, fa pensare ad una classificazione basata su una variabile categoriale dipendente da queste 2.





Le top speed non sembrano distinguere bene le unità dei 2 gruppi meno numerosi, per cui non posso dire che siano delle buone variabili. Inoltre le variabili relative ai consumi gomme, sia anteriori che posteriori, di P2 e P3 non fanno un buon lavoro di distinzione. Non vedo nessun'altra variabile, oltre quelle spiegate in precdenza, in grado di poter raggruppare bene le mie unità in 3 gruppi.



Addirittura queste ultime variabili non mostrano uno tra i 3 gruppi, quindi scarto definitivamente anche loro.
Procedo quindi alla creazione di una nuova variabile, basata sui tempi in qualifica e l'annata della moto: la chiamerò **q_bike_year** e presenterà 3 modalità:

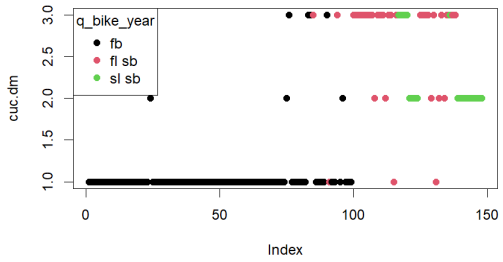
- **fb** (fast bike), per tutte le unità che presentano tempi sul giro con moto veloci, ovvero con moto di annata superiore o pari al 2021;
- **sl sb** (slow lap slow bike), per tutte le unità che presentano tempi sul giro con moto "lente", quindi di annata precedente al 2021 ed un tempo sul giro maggiore di 107 secondi;
- **fl sb**, per tutte le unità che presentano tempi sul giro con moto "lente", quindi di annata precedente al 2021 ed un tempo sul giro minore di 107 secondi.

La tabella di frequenze incrociate tra le modalità della variabile ed i cluster risulta essere la seguente:

Cluster	**fl**	**fl sb**	**sl sb**	**totale**
1	89	3	0	92
2	3	5	14	22
3	4	25	5	34
totale	96	33	19	148

Vengono commessi 3+4+3+5+5 = 20 errori, quindi il 13.5 % delle unità viene mal classificato, il che è buono. Il test d'indipendenza, calcolato per l'ipotesi nulla $H0 : cluster \perp q_bike_year$, risulta avere una statistica test X^2 pari a 149.19, con 4 gradi di libertà, che a fronte di un valore di χ^2_4 pari a 9.488 mi porta a rifiutare l'ipotesi nulla, con un p-value minore di $2.2*10^{-16}$. Il Rand Index risulta pari

a 0.713, il che indica che le unità vengono generalmente ben classificate. Graficamente, però, si può vedere come il risultato non sia così soddisfacente:



Le unità del secondo e del terzo cluster, in particolare, sono quelle dove le unità sono peggio distribuite, forse per via della poca numerosità del secondo.

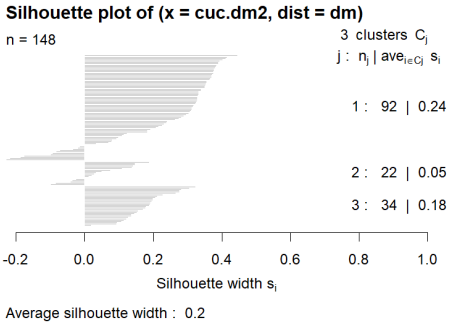


Figura 55: Grafico delle silhouette del clustering con legame completo, basato su distanze di Manhattan

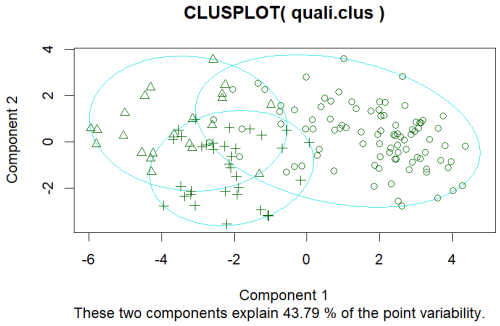


Figura 56: Clusplot delle unità raggruppate con un metodo di clustering con legame completo, basato su distanze di Manhattan

Le unità con silhouette negativa sarebbero distribuite un po' diversamente da quelle mal classificate nella tabella di frequenze incrociata: circa poco più di 10 unità con silhouette negative nel primo gruppo, a fronte di sole 3 mal classificate nella tabella, mentre nessuna unità con silhouette negativa per il terzo, a fronte di ben 4+5=9 mal classificate nella tabella. Le unità del secondo gruppo, invece, coinciderebbero abbastanza. Il clusplot non distingue così bene i cluster, ma non esegue neanche un pessimo lavoro. Visto, quindi, il buon valore del Rand Index, posso dire che **la variabile q_bike_year è significativamente una buona variabile per raggruppare le mie unità statistiche in 3 gruppi**.

Vediamo cosa succede se uso le **distanze di Lagrange con il legame completo**:

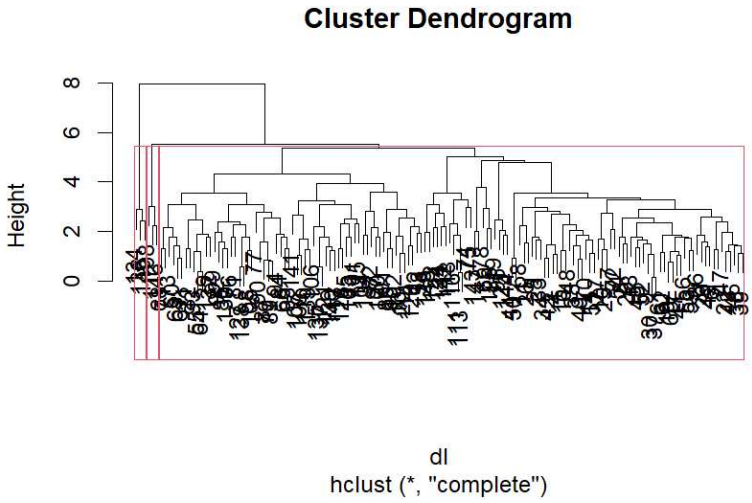


Figura 57: Dendrogramma di un clustering con legame completo basato, su distanze di Lagrange

Il dendrogramma indica che questo processo di clustering è pessimo: abbiamo 2 cluster con pochissime unità ed il terzo che ne ha decisamente troppe, segno che il clustering è eseguito in maniera pessima. Per questo motivo non vedremo il grafico delle silhouette, ma solo il clusplot:

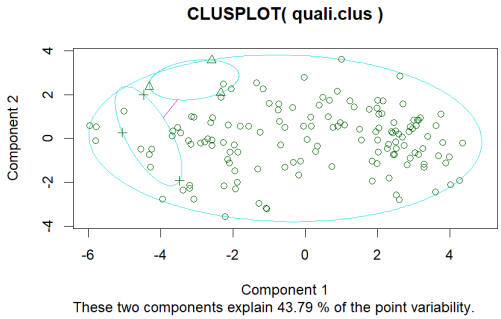


Figura 58: Clusplot delle unità raggruppate con un metodo di clustering con legame completo, basato su distanza di Lagrange

Quest'ultimo va a confermare quanto appena affermato. Non ha quindi senso procedere ulteriormente con l'analisi di questo metodo.

Analizziamo quindi i cluster che si vanno a creare con il **legame medio**, a partire da quello basato sulle **distanze euclidee**:

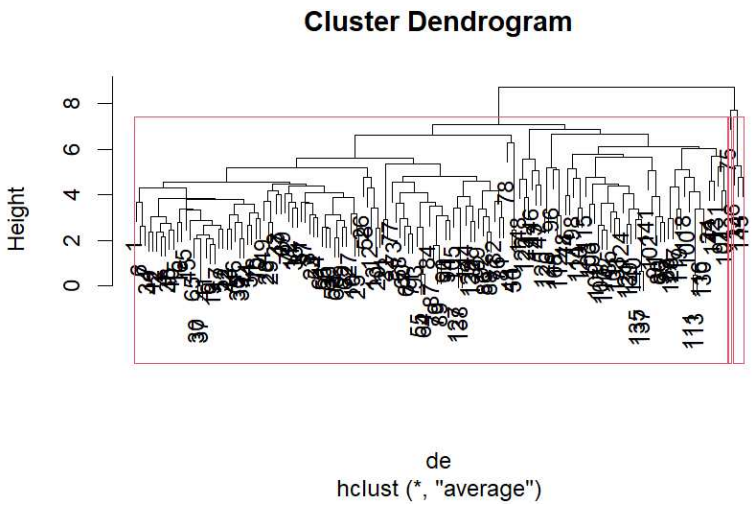


Figura 59: Dendrogramma di un clustering con legame medio, basato su distanze euclidee

Capisco da subito che non sembra essere un buon metodo: sembra quasi un clustering con legame singolo da quanto raggruppa male le unità. Ci sono 2 gruppi che insieme comprenderanno circa 4 o 5 unità e tutte le rimanenti verrebbero assegnate al terzo cluster. Inutile quindi dire, quindi, che questo metodo di clustering non funzioni.

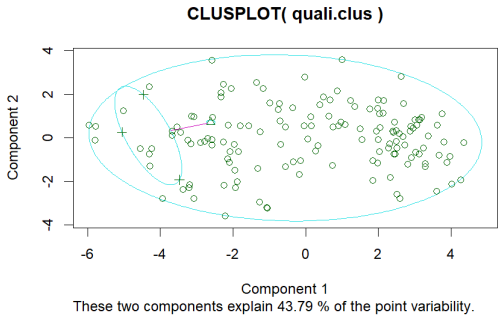


Figura 60: Clusplot delle unità raggruppate con un metodo di clustering con legame medio, basato su distanza euclidea

Il clusplot conferma ancor di più questa tesi, essendo 2 gruppetti poco numerosi contenuti all'interno di un gruppone enorme.

Vediamo se cambia qualcosa usando le **distanze di Manhattan**:

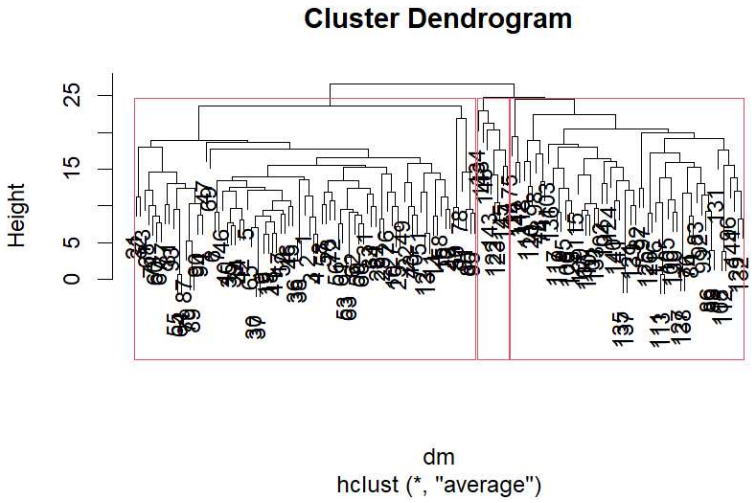


Figura 61: Dendrogramma di un clustering con legame medio, basato su distanze di Manhattan

Neanche questo metodo sembra svolgere un buonissimo lavoro, abbiamo un gruppo molto poco numeroso. Se però andiamo, tuttavia, a vedere il clusplot otteniamo un risultato molto curioso:

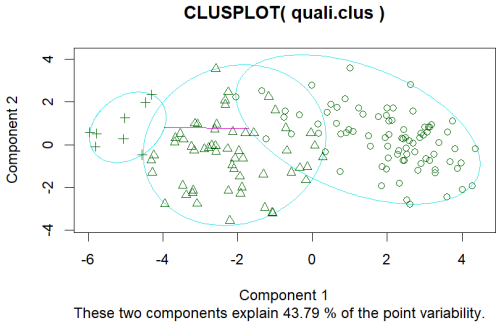
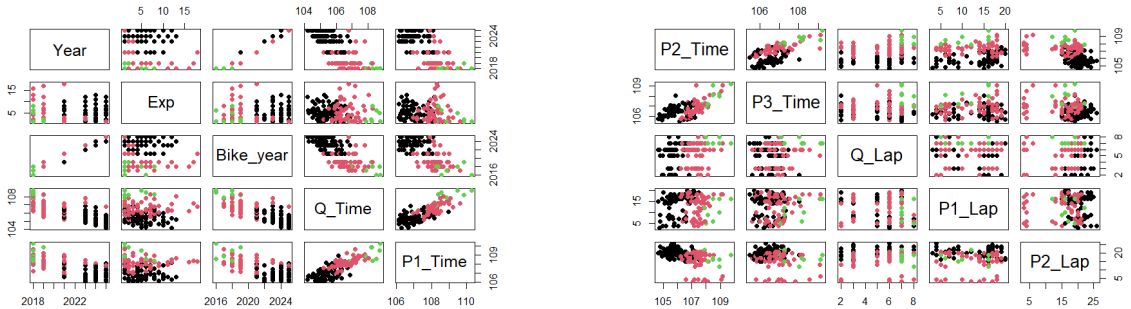
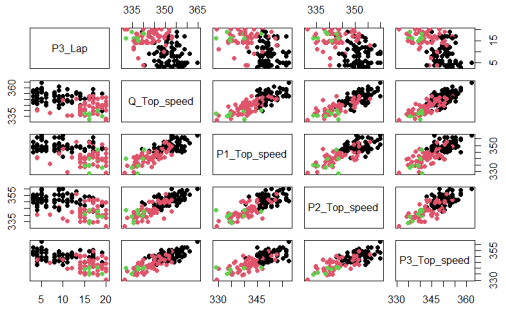


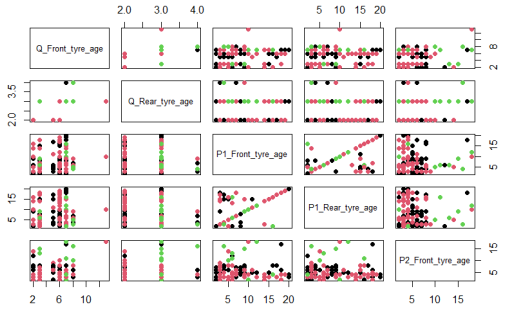
Figura 62: Clusplot delle unità raggruppate con un metodo di clustering con legame medio, basato su distanze di Manhattan

Il gruppo meno numeroso è molto distinguibile dagli altri 2, nelle prime 2 componenti principali: ci devono essere quindi delle variabili in grado di spiegare il clustering. Vediamo quali possono essere:

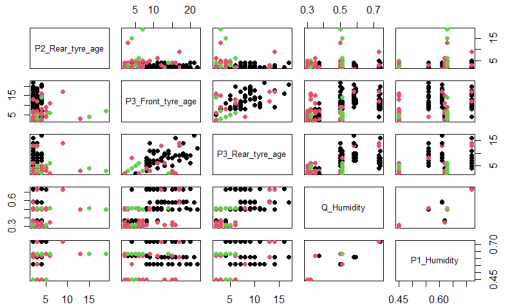


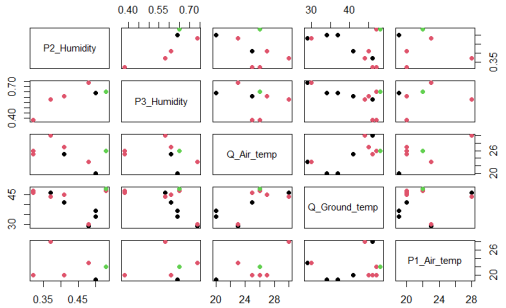
Tra le prime 10 variabili noto che c'è grande distinzione per quel che riguarda i tempi sul giro di tutte le sessioni, soprattutto quelli di qualifiche e P1, oltre che per quanto riguarda le annate di rilevazione e delle moto che, per quanto non vengano rappresentate al 100 % dal clustering, comunque lo spiegano il processo di clustering alquanto egregiamente. Le altre variabili, soprattutto Exp, funzionerebbero bene sono se accompagnate, in un grafico di dispersione, alle variabili precedentemente citate.

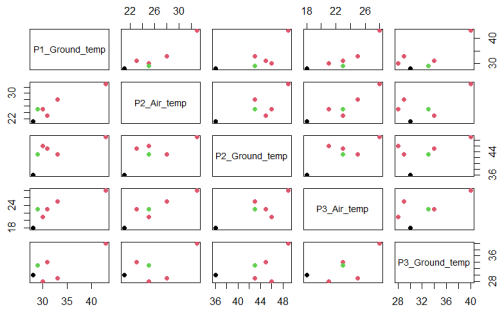




Le variabili relative alle top speed, di tutte le sessioni, e la variabile P3_Lap spiegano altrettanto bene i clusteri. Forse li spiegherebbero ancora meglio se ad essere creati fossero 2 gruppi, anzichè 3. Le variabili relative ai consumi delle gomme posteriori nelle P1 e delle gomme anteriori nelle P2 non riescono a fare una buona distinzione tra i 2 cluster più numerosi, mentre la fanno decentemente per quello meno numeroso.







La stessa cosa appena detta vale per le variabili relative ai consumi delle gomme anteriori e posteriori nelle P3. Dalle variabili atmosferiche, invece, viene svolto un discreto lavoro, per quanto difficile a dirsi, dalle temperature dell'asfalto nelle P1 e nelle P2, così come dalla temperature dell'aria nelle P2. Le altre non svolgono un lavoro altrettanto eccellente.

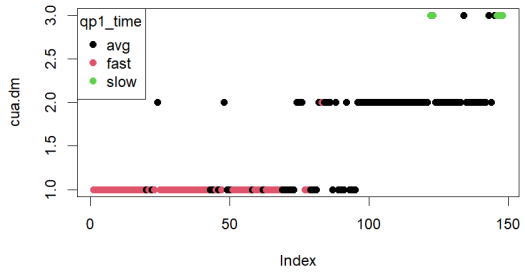
Provo quindi a creare una nuova variabile, basata sulle 2 variabili che meglio spiegano il clustering, ovvero Q_Time e P1_Time, per verificare il clustering: la variabile **qp1_time**, con 3 modalità:

- **fast**, se il tempo sul giro delle P1 è minore di 107.5 secondi e quello delle qualifiche è minore di 106 secondi;
- **slow**, se entrambi i tempi sul giro sono stati maggiori di 108 secondi;
- **avg** (average), se il tempo sul giro delle P1 è compreso tra i 107.5 ed i 108 secondi e quello delle qualifiche è compreso tra i 106 ed i 108 secondi.

La tabella di frequenze incrociate con i cluster è la seguente:

Cluster	**avg**	**fast**	**slow**	**totale**
1	24	59	0	83
2	56	1	0	57
3	3	0	5	8
totale	83	60	5	148

Il clustering sbaglierebbe a classificare 24+1+3 = 28 unità, quindi poco più del 18 % del totale. Se effettuiamo anche una piccola verifica grafica, possiamo notare che le unità sono anche semi-decentemente distribuite:



Nel test d'indipendenza, la statistica test X^2 ha 4 gradi di libertà ed un valore pari a 160.48, che a fronte del nostro solito riferimento χ^2_4=9.488 ci porta a rifiutare il test d'indipendenza, con un p-value minore di $2.2*10^{-16}$; tuttavia il Rand Index ottenuto è pari a 0.422, quindi molto basso. Questo, secondo me, è dato dal fatto che un cluster, ovvero il numero 3, ha numerosità molto bassa. In sostanza, quindi, il clustering effettua un buon lavoro se volessi classificare le mie unità in 3 gruppi, ma non è detto che la variabile migliore per spiegare il clustering sia qp1_time. Vediamo, quindi, cosa ci dice il grafico delle silhouette:

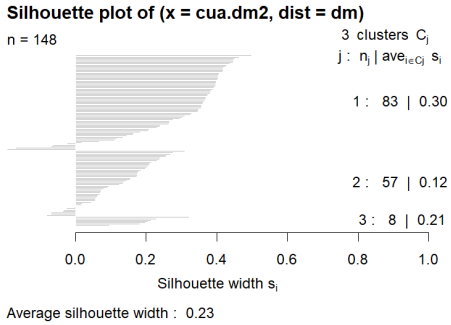


Figura 63: Grafico delle silhouette del clustering con legame medio basato su distanza di Manhattan

Abbiamo poche unità con silhouette negative, per cui confermo come, sicuramente, il clustering si basi su un critero leggermente diverso dalla variabile introdotta in precedenza. Possiamo comunque dire, in definitiva, che **i cluster dipendono significativamente dai tempi sul giro, con una maggior influenza da parte di quelli registrati in qualifica e nelle P1**.
Spostiamoci ora al metodo di **clustering con legame medio basato sulle distanze di Lagrange**:

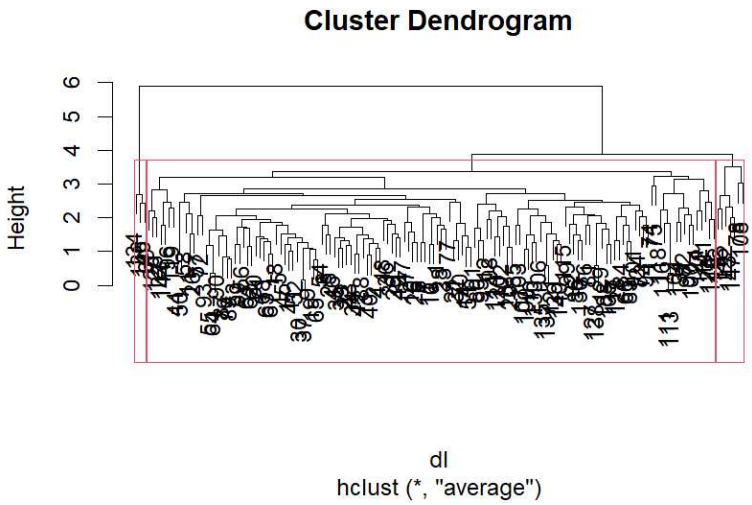


Figura 64: Dendrogramma di un clustering con legame medio basato su distanza di Lagrange

Anche questo metodo non sembra funzionare bene, come quello basato sulle distanze euclidee. Vediamo se ci possono essere sorprese, però, come visto in precedenza:

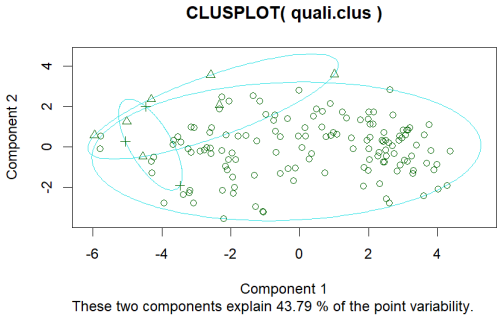


Figura 65: Grafico delle silhouette del clustering con legame medio, basato su distanze di Lagrange

Questa volta non abbiamo sorprese, il clustering non fa un bel lavoro. Nelle prime 2 componenti principali, faccio molta fatica a distiguere le ellissi dei gruppi l'una dall'altra. Per cui possiamo affermare con certezza che **l'unico clustering con legame medio funzionante è quello basato sulle distanze tra unità di Manhattan**.

Spostiamoci ora sui metodi gerarchici di **clustering con legame di Ward**, partendo da quello basato su **distanze euclidee**:

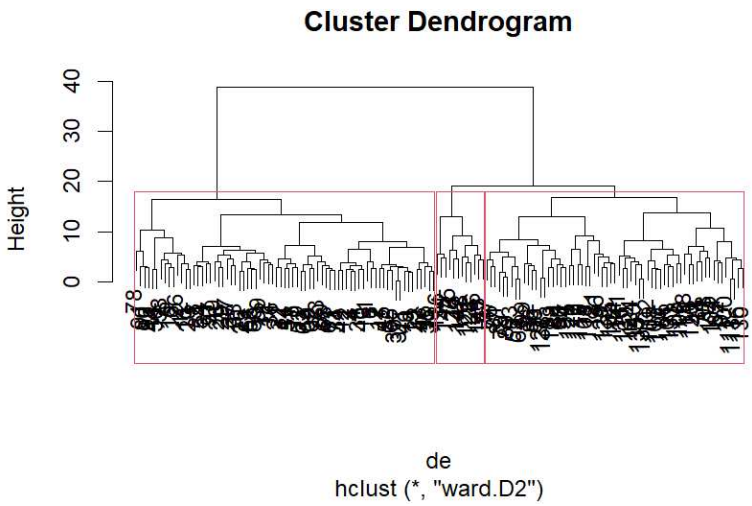


Figura 66: Dendrogramma di un clustering con legame di Ward, basato su distanze euclidee

Il metodo con legame di Ward, basato su distanze euclidee, sembra dessere un metodo dove le unità vengono comunque ben distribuite, ma tra i 3 gruppi noto che ce n'è uno molto meno numeroso. Verifichiamo, quindi, dal clusplot se possiamo potenzialmente trovare dei criteri di clustering:

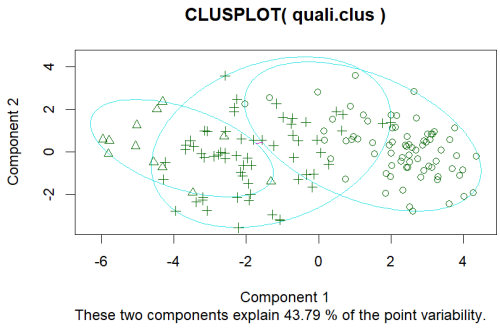
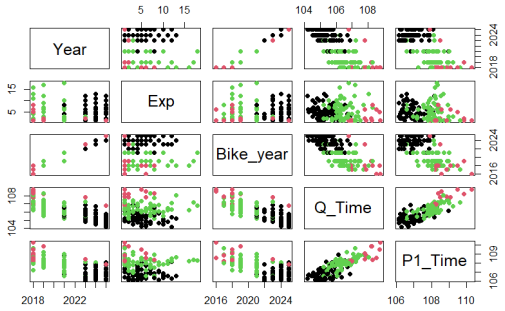
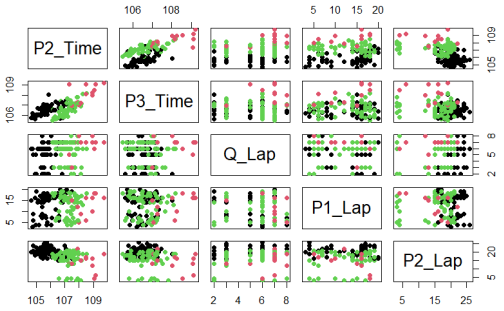
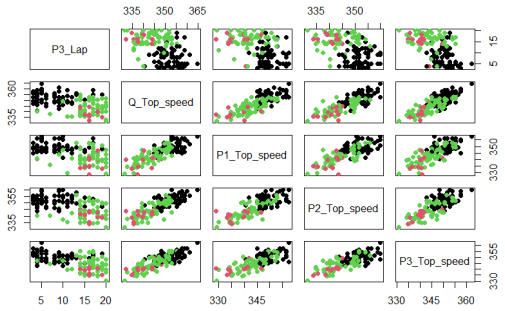


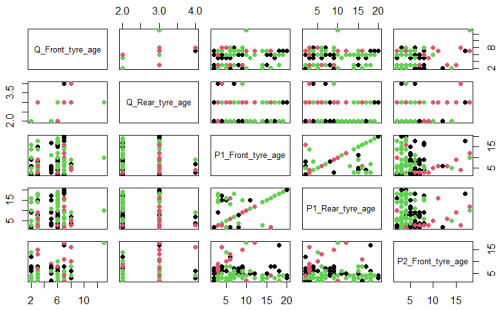
Figura 67: Clusplot delle unità raggruppate con un metodo di clustering con legame di Ward, basato su distanze euclidee

Il clusplot ci mostra 3 gruppi abbastanza ben distinti, ma non nel migliore dei modi. Quasi sicuramente il clustering si basa su un criterio, che noi andremo a vedere, dai grafici di dispersione, quale possa essere. Tuttavia non mi aspetto di vedere una buonissima dispersione dei cluster negli scatterplot, proprio perché nei clusplot le 3 ellissi sono distinte, ma non perfettamente.

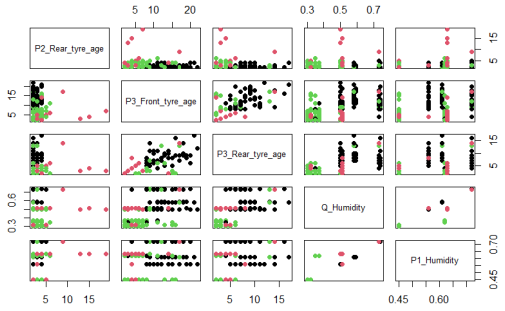


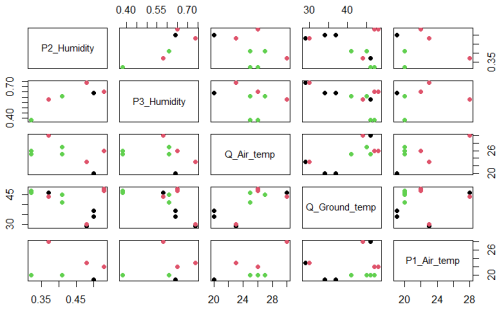


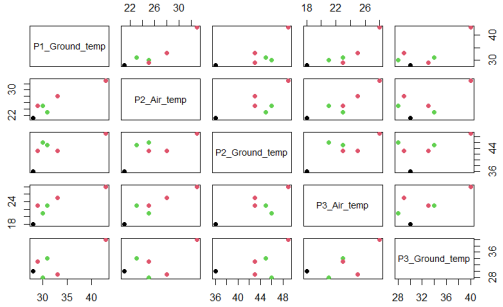




Anche in questo caso abbiamo una distinzione in base alle variabili Year e Bike_year, assieme soprattutto alle variabili relative ai tempi sul giro di tutte le sessioni, soprattutto quelli di P2 e P3 (secondo me le migliori), a quelle relative alle velocità massime di tutte le sessioni e alla variabile P3_Lap.



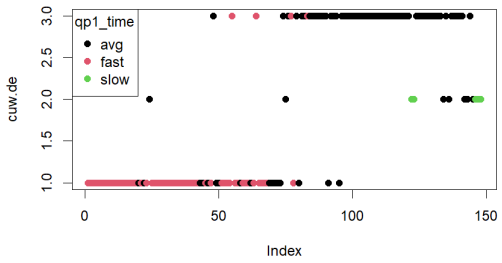




Sia la variabile relativa al consumo delle gomme posteriori nelle P2 che quelle relative ai consumi delle gomme, anteriori e posteriori, nelle P3, sono a loro volta molto buone; sembrano quasi dividere le unità in soggetti più e meno bravi a far registrare un buon tempo con gomme usurate. Sia le temperature dell'aria che quelle dell'asfalto, nelle 3 sessioni di prove libere, fanno a loro volta un egregissimo lavoro.
Vista la similarità visiva, confrontiamo anche questo clustering con la stessa variabile usata per il clustering con legame medio basato su distanze di Manhattan, ovvero **qp1_time**:

clm dm	**avg**	**fast**	**slow**	**totale**
1	17	56	0	73
2	7	0	5	12
3	59	4	0	63
totale	73	12	63	148

17+7+4+5 = 33 unità sono distribuite diversamente, circa il 22.3 % che non è poco. Se effettuiamo un verifica grafica notiamo ulteriormente che, per i 2 gruppi più numerosi, le classificazioni sono più o meno giuste, mentre per il meno numeroso si potrebbe fare meglio:



Eseguendo comunque un test d'indipendenza, per l'ipotesi $H0 : cluster \perp qp1_time$, la mia statistica test X^2, con 4 gradi di libertà, ha valore pari a 133.12, che essendo maggiore del valore di riferimento χ^2_4=9.488, mi porta a rifiutare $H0$ con un p-value minore di $2.2*10^{-16}$. Quindi, comunque, il clustering è significativamente dipendente dalla variabile. Tuttavia, un valore del Rand Index pari a 0.448, mi porta a dire che esistano variabili migliori per spiegarlo.

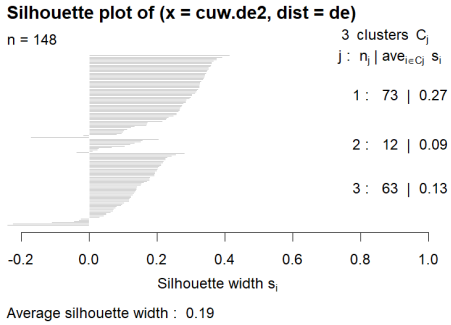


Figura 68: Grafico delle silhouette del clustering con legame di Ward, basato su distanze euclidee

Abbiamo circa una decina di unità con silhouette negativa, ben di meno rispetto alle 33 mal classificate dal clustering rispetto alla variabile. Per cui possiamo dire definitivamente che si tratti di un metodo che, **con la creazione di una nuova variabile qualitativa basata sui tempi sul giro, avente una maggior influenza da parte dei tempi registrati nelle P2 e nelle P3, si potrebbe effettuare un buon clustering**.

Vediamo se usando le **distanze di Manhattan nel clustering con legame di Ward** cambia qualcosa:

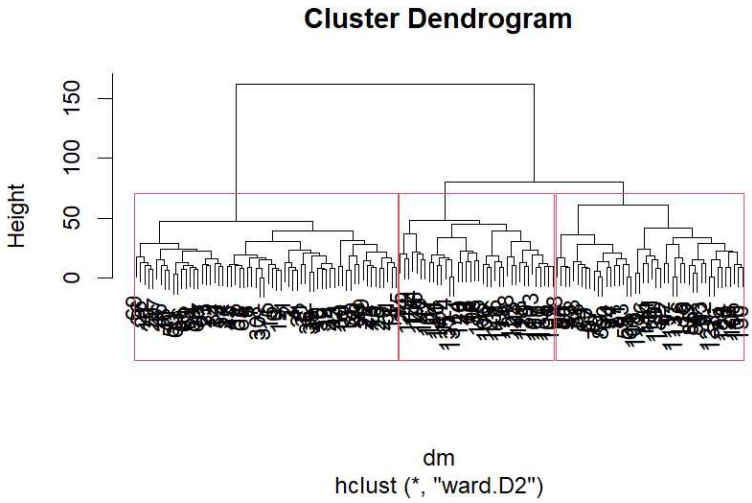


Figura 69: Dendrogramma di un clustering con legame di Ward, basato su distanze di Manhattan

E', finora, il metodo di clustering dove le unità sono state divise nella maniera più equa. Inoltre il clusplot suggerisce una buona distinzione tra i gruppi nelle prime 2 componenti principali:

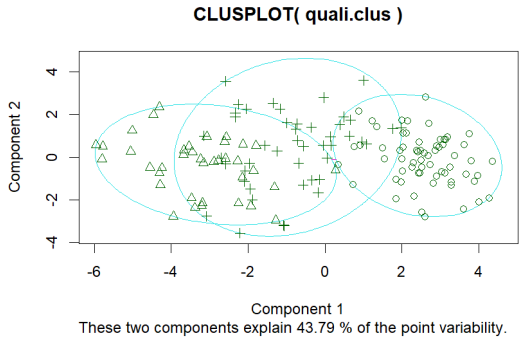
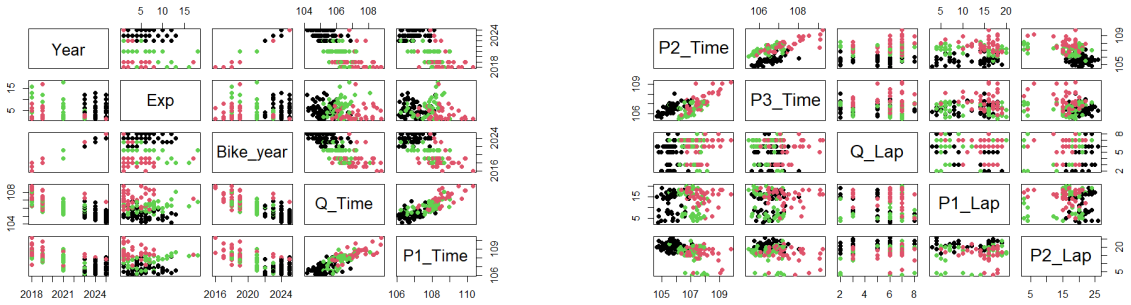
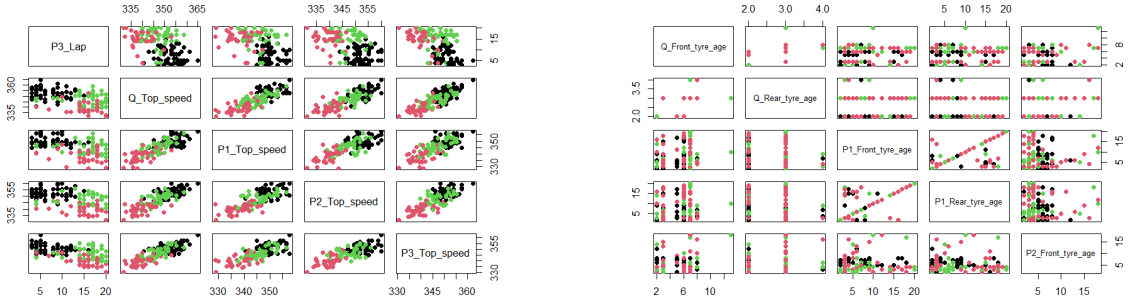


Figura 70: Clusplot delle unità raggruppate con un metodo di clustering con legame di Ward, basato su distanze di Manhattan

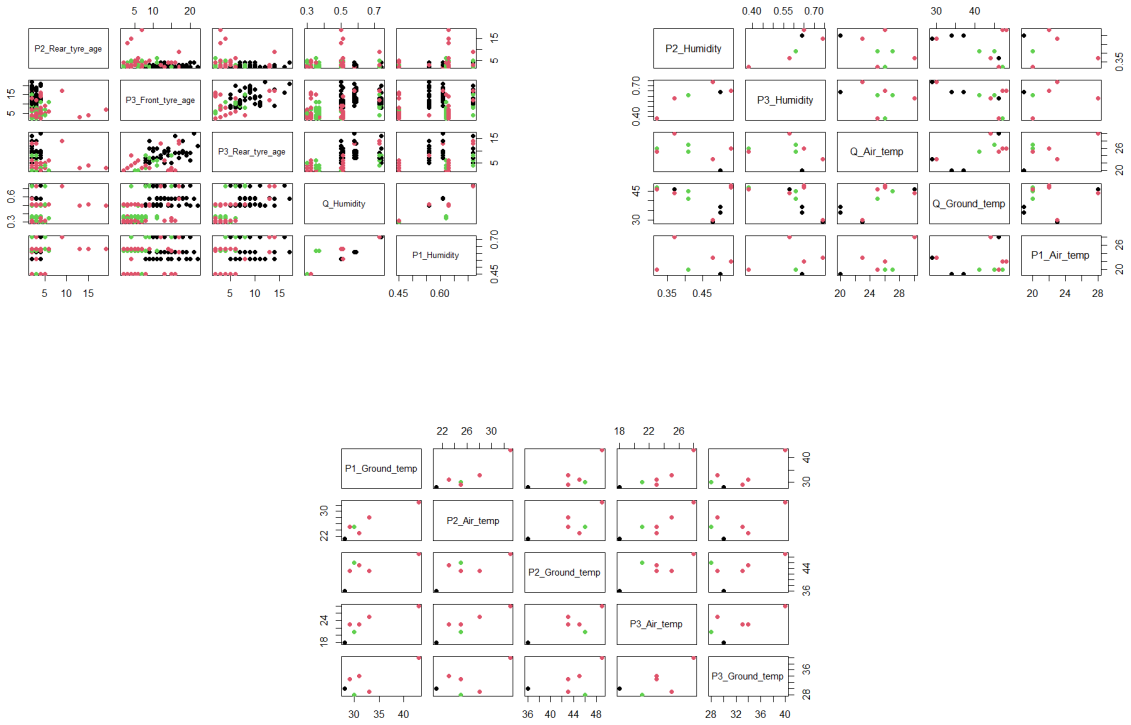
Vediamo, quindi, dai grafici di dispersione che criteri di classificazione vengono usati:



Tra le prime 10 variabili, quelle che fanno il miglior clustering sono sicuramente quelle relative alle annate, sia di rilevazione che delle motociclette, mentre quelle relative ai tempi sul giro mescolano un po' troppe unità tra di loro per essere considerate ottimali.



Da questi grafici capisco, inoltre, che questo metodo di clustering si basa principalmente sulle velocità di punta di tutte le sessioni, oltre che soprattutto sulla variabile P3_Lap, che secondo me può essere la migliore, tra tutte quelle viste finora, nel distinguere i 3 cluster.



Tra le altre variabili nessuna riesce a spiegare veramente bene il clustering, per cui possiamo affermare che quest'ultimo dipende da una variabile categoriale che riassume le velocità di punta ed il numero di giri impiegati dal pilota per far registrare il proprio miglior tempo nelle P3. Potremmo, a questo punto, creare una nuova variabile basata sulla top speed delle P3 e su P3_Lap, chiamata **P3_summary**, con 3 modalità:

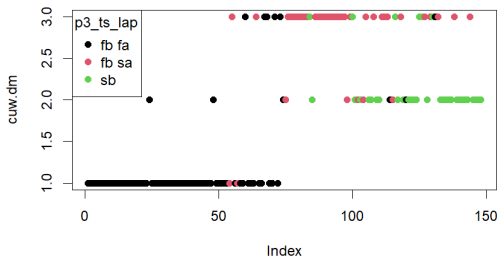
- **sb** (slow bike), per tutte le veolcità di punta registrate nelle P3 minori di 342 km/h;

- **fb fa** (fast bike fast adaptaion), per tutte le veolcità di punta registrate nelle P3 maggiori o uguali a 342 km/h e un valore di P3_Lap minore di 14 (quest'ultimo indice di un veloce adattamento alle condizioni della pista nelle P3);
- **fb sa** (fast bike slow adaptaion), per tutte le veolcità di punta registrate nelle P3 maggiori o uguali a 342 km/h ed un valore di P3_Lap maggiore o uguale di 14 (indice di un adattamento più lento alle condizioni della pista).

La tabella di frequenze incrociate con i cluster sarebbe la seguente:

Cluster	**fb fa**	**fb sa**	**sb**	**totale**
1	62	2	0	64
2	5	5	28	38
3	6	34	6	46
totale	73	41	34	148

Vengono commessi veramente pochi errori, ovvero 5+6+5+2+6=24, circa il 16.2 %, che per una variabile nuova creata sul momento non è male.



Graficamente viene confermata la presunta bontà di adattamento del clustering rispetto alla variabile, fermo restando che **il clustering non viene svolto come metodo di classificazione a scopo di prediction**.
Nel test d'indipendenza, volto a verificare l'ipotesi nulla $H0 : cluster \perp P3_Summary$, la statistica test X^2, con 4 gradi di libertà, ha valore pari a 163.22, a fronte del valore di riferimento χ^2_4 pari a 9.488. Ciò ci porta al rifiuto del test d'indipendenza, con un p-value minore di $2.2*10^{-16}$ e ad affermare, quindi, che il clustering è significativamente dipendente dalla variabile P3_Summary.
Tuttavia, il Rand Index risulta pari a circa 0.6, che non è così vicino ad 1, per cui potrei dire che la variabile effettivamente funziona per raggruppare le mie unità, ma non al meglio. Andiamo a vedere infine il grafico delle silhouette, per farci un'ulteriore idea sul metodo che stiamo analizzando:

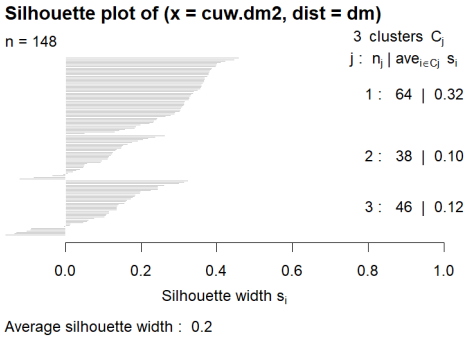


Figura 71: Grafico delle silhouette del clustering con legame di Ward basato su distanza di Manhattan

Il numero di unità con silhouette negativa è circa la metà del numero di unità mal classificate dal clustering, tuttavia con una struttura simile: per il primo cluster non ci sono unità con silhouette negative, a fronte di sole 2 mal classificate; per il secondo ce ne sono 4 con silhouette negativa, anziché 10 mal classificate; nel terzo ce ne sono 7 con silhouette negativa, a fronte di 12 mal classificate. **Posso quindi affermare che, sicuramente, esiste una variabile legata alle velocità di punta di tutte le sessioni e a P3_Lap, in base alla quale le unità vengono clusterizzate in 3 gruppi ben distinti**.
Se io volessi infine basare il **clustering con legame di Ward sulle distanze di Lagrange**?

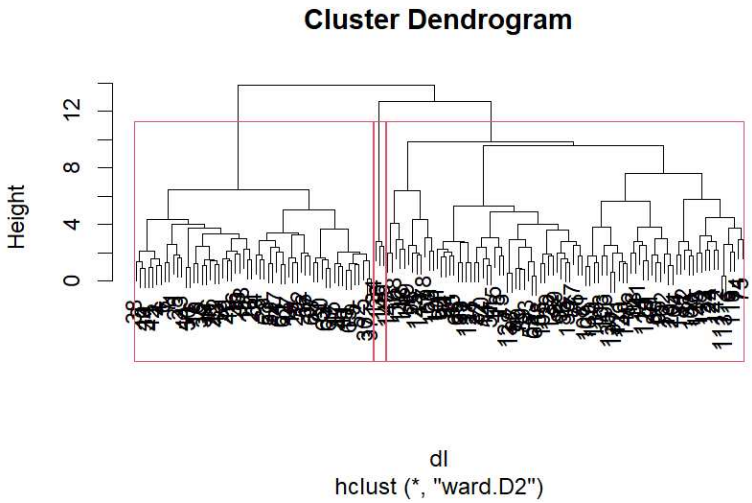


Figura 72: Dendrogramma di un clustering con legame di Ward, basato su distanze di Lagrange

Il tutto peggiorerebbe esponenzialmente: ci sarebbe un cluster, posto visivamente nel mezzo, che sarebbe composto solo da 3 o 4 unità; le rimanenti 144 sarebbero spartite tra gli altri 2, ma non è quello che voglio. Per cui potrei scartare, anche solo a prima vista, questo metodo come metodo ottimale di clustering.

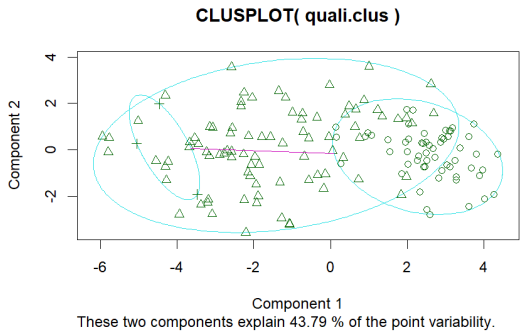


Figura 73: Clusplot delle unità raggruppate con un metodo di clustering con legame di Ward, basato su distanze di Lagrange

Il clusplot conferma quanto appena scritto: il gruppo con numerosità inferiore non è ben distinguibile sul piano delle prime 2 componenti principali.

5.3 Cluster analysis con il metodo delle k-medie

Prima di vedere i risultati che uscirebbero usando il clustering con il metodo delle k-medie, con k=3, vediamo alcuni dettagli sulle devianze che ne uscirebbero dal clustering stesso, ovvero:

- **devianza totale nei gruppi**, calcolata con la formula

$$SSW = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (y_{ij} - \bar{y_j})$$

  dove n_j è la numerosità del j-esimo gruppo, y_{ij} è l'i-esimo vettore di valori presente nel j-esimo gruppo e $\bar{y_j}$ è il vettore delle medie del j-esimo gruppo. La formula $\sum_{i=1}^{n_j}(y_{ij} - \bar{y_j})$ è la formula della devianza interna di ciascun gruppo;

- **devianza totale tra gruppi**, data dalla formula

$$SSB = \sum_{j=1}^{k} n_j \|\bar{y_j} - \bar{y}\|^2$$

  dove $\bar{y}$ è il vettore delle medie dell'intero dataset;

- **devianza totale**, data dalla somma delle devianze interne e le devianze tra gruppi ($SST = SSW + SSB$);

- **rapporto tra devianza tra gruppi e devianza totale** $\frac{SSB}{SST}$. Di solito più alto è, più i gruppi sono distinti tra loro.

SSW	SSB	SST	$\frac{SSB}{SST}$
2059.426	1027.574	3087	0.333

Il rapporto $\frac{SSB}{SST}$ ha valore abbastanza basso, segno che i gruppi non sono così tanto distinti tra loro. Non essendo questo un metodo di clustering gerarchico, non utilizzeremo il dendrogramma, bensì guarderemo direttamente il clusplot:

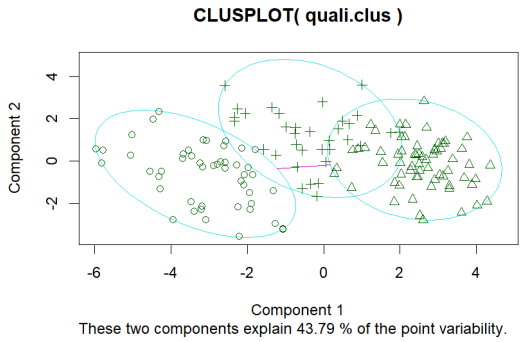
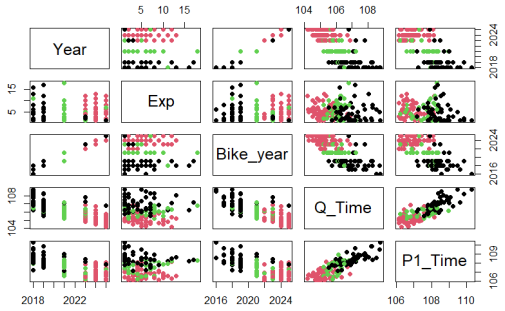
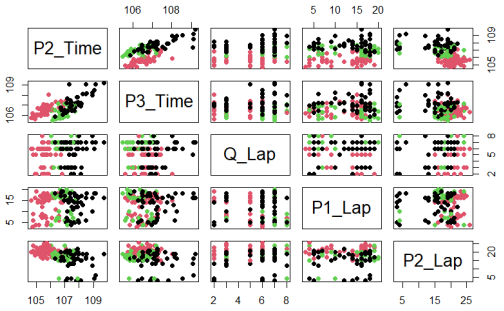


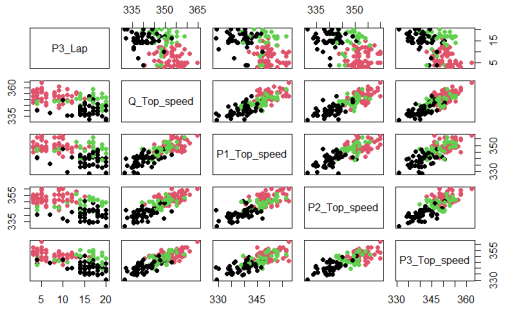
Figura 74: Clusplot del clustering con metodo delle k-medie, con k=3

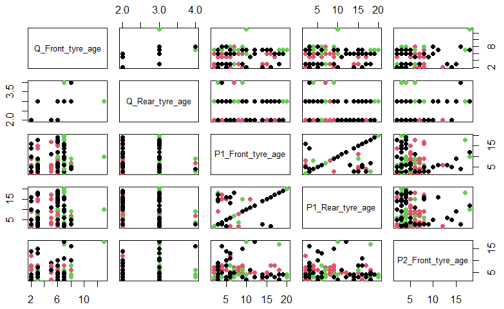
Il clusplot ci mostra 3 gruppi molto ben distinti nelle prime 2 componenti principali. Per cui, mi aspetto di trovare sicuramente una o più variabili che spiegano alquanto bene il clustering.



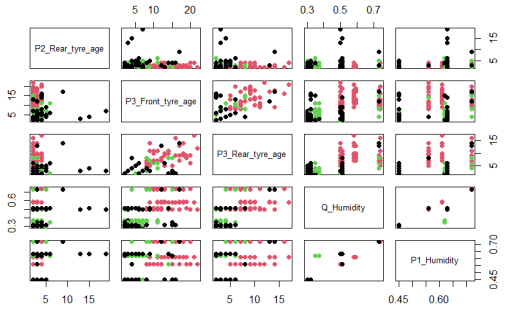


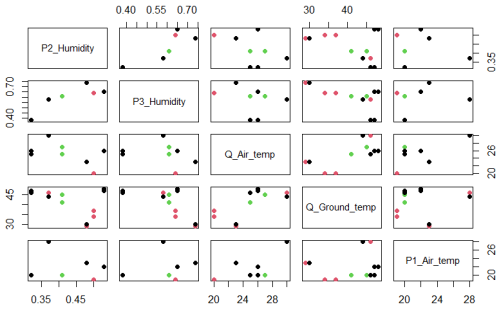
In base alle prime 10 variabili affermo con quasi certezza che questo metodo fa un bellissimo lavoro a dividere le unità per periodo di rilevazione, così come un bel lavoro lo fa anche la variabile relativa all'annata della moto. I tempi sul giro, di tutte le sessioni, per quanto facciano un lavoro non male, non replicano i risultati delle 2 variabili appena citate.



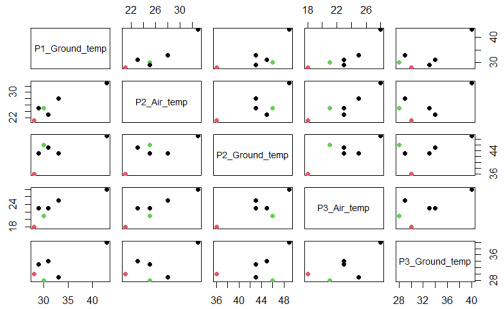


La variabile P3_Lap fa anch'essa un bel lavoro, mentre le velocità di punta, di tutte le sessioni, fanno fatica a distinguere bene le unità di 2 gruppi tra loro.





I consumi delle gomme posteriori nelle P2, così come quelli degli pneumatici anteriori e posteriori rilevati nelle P3, fanno fatica a distinguere 2 gruppi (visivamente di colore nero e verde) tra di loro, mentre tra le variabili atmosferiche viste fino a qui solo P1_Air_Temp sembra svoglere un bel lavoro, ma non come le annate di rilevazione e della moto



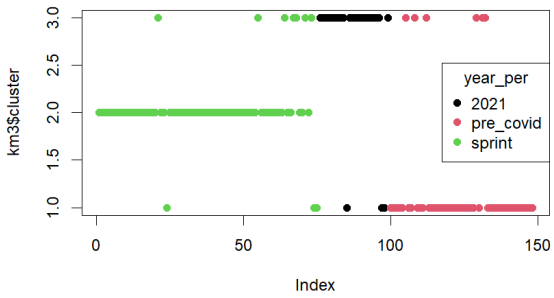
Anche P3_Air_Temp sembra svoglere un lavoro discreto, ma non eccelso come quello svolto dalle annate di rilevazione e della moto.
A questo punto potrei dividere le unità in 3 gruppi, secondo una nuova variabile chiamata **period**, che presenta le seguenti modalità:

- **pre-covid**, relativa alle rilevazioni precedenti al 2020, quindi 2018 e 2019;
- **2021**, relativa all'anno stesso;
- **sprint**, relativa alle annate 2023, 2024 e 2025 nelle quali, a differenza delle altre, si disputano anche le Sprint Races del sabato pomeriggio.

La tabella di frequenze incrociata con i cluster risulta essere la seguente:

Cluster	**2021**	**pre-covid**	**sprint**	**totale**
1	3	43	3	49
2	0	0	65	65
3	21	6	7	34
totale	24	49	75	148

Sembra veramente essere un buon metodo: sbaglia a classificare solo 3+6+7+3 = 16 unità, circa il 10.8 %, quindi veramente poche. Ecco anche una prova grafica della sua funzionalità:



Nel test d'indipendenza tra il clustering e la variabile, la statistica test X^2, con 4 gradi di libertà, ha valore pari a 181.76, a fronte di un valore di χ^2_4 pari a 9.488; ciò ci porta a rifiutare l'ipotesi nulla di indipendenza tra la variabile ed il clustering, con un p-value minore di $2.2*10^{-16}$. Per cui posso dire che il raggruppamento delle unità è significativamente dipendente dal periodo di rilevazione. Il Rand Index ha valore pari a 0.69, quindi più basso di quanto ci si possa aspettare, ma non in maniera drastica. Infatti è tra i più alti, fra quelli calcolati finora nei vari confronti tra clustering e variabili categoriali.

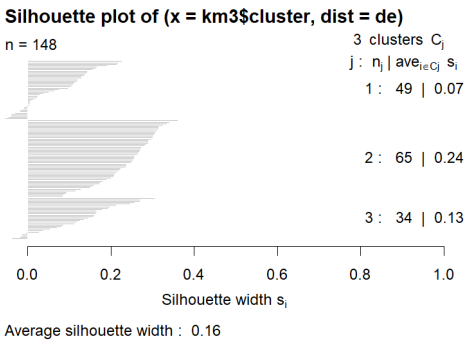


Figura 75: Grafico delle silhouette del clustering con metodo delle k-medie, con k=3

Il numero di unità con silhouette negativa è quasi lo stesso numero di unità mal classificate, con strutture però un po' diverse: il numero di unità mal classificate è 6 nel primo gruppo, a fronte di 10 unità con silhouette negative, mentre nel terzo abbiamo 13 unità mal classificate a fronte di sole 5 con silhouette negative. Tuttavia, quando le silhouette sono negative, hanno comunque valore assoluto basso, segno che non sono state mal classificate di troppo.
In conclusione: questo metodo di clustering è il migliore tra quelli analizzati, se si vuole considerare la classificazione delle unità in periodi di rilevazione. Tuttavia, se si vuole raggruppare in base ad altri criteri, alcuni metodi di clustering gerarchico, come visto in precedenza, sono nettamente migliori di questo.

6 Previsione del tempo sul giro in qualifica

6.1 Contesto

Arriviamo allo scopo dell'esistenza delle variabili qulitative in questo dataset: **la previsione del tempo sul giro in qualifica di un pilota**. La variabile in questione, **Q_Time**, è detta **variabile risposta**, mentre le altre variabili sono dette **variabili concomitanti**.
Per eseguire tale operazione, svilupperò una serie di modelli lineari normali a scopo di prediction, basati sulla seguente formula che spiega la relazione tra la variabile risposta e le variabili concomitanti:

$$Y = f(X) + \epsilon = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p + \epsilon$$

dove ϵ vettore dei residui, calcolato come $\epsilon = Y - \hat{Y}$ ($\hat{Y}$ è la stima di Y, calcolata con la formula $\hat{Y} = \hat{f}(X) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$) e p è il numero di coefficienti, esclusa l'intercetta β_0, presenti nel modello.
Ognuno dei seguenti modelli si baserà su parte del dataset, circa il 75 % delle unità (111 unità), chiamato **dataset di training**, mentre il rimanente 25 % circa (37 unità), chiamato **dataset di test**, verrà usato come dataset per testare la bontà di adattamento del modello, attraverso il calcolo dell'MSE, abbreviativo di Mean Squared Error (errore quadratico medio):

$$MSE = \sum_{i=1}^{m} (y_i - f(x_i))^2$$

dove y_i è il valore della variabile risposta per l'i-esima unità del dataset di test, $f(\cdot)$ è pari alla funzione del modello lineare normale, usato per stimare il vettore $\hat{Y}$, mentre x_i è pari al vettore delle variabili concomitanti per l'i-esima unità del dataset di test.
Più l'MSE è basso, meglio il modello si adatterà anche a dati diversi rispetto a quelli su cui si basa e quindi meglio riuscirà ad effettuare previsioni, riducendo la probabilità di overfitting (sovradattamento ai dati su cui si basa, con la conseguenza di una cattiva prediction).
Il numero di coefficienti, detti "regressori" (espressi con il vettore β), in un modello, può essere maggiore del numero di variabili effettivamente introdotte: ciò succede quando almeno una di esse è di tipo qualitativo. Nel loro caso viene creato un coefficiente per ogni modalità di esse, meno una (poiché se no la matrice X del modello non sarà a rango pieno), che verrà moltiplicato per 1 quando l'unità avrà quella modalità come valore della variabile e 0 altrimenti. Se tutti i coefficienti saranno moltiplicati per 0, vuol dire che l'unità presenterà la modalità per la quale non è presente il coefficiente nel modello.
Per ogni modello verificheremo, quindi, l'MSE e, ove possibile:

- **la normalità dei residui contenuti nel vettore** ϵ, che si dovrebbe distribuire sotto ipotesi nulla come $\epsilon \sim N(0, \sigma^2)$, dove σ^2 è la varianza residua. Per farlo eseguiro il test di normalità di Shapiro-Wilk, dove la relativa statistica test W assumerà valori compresi tra 0 e 1. Più il valore è vicino ad 1 e più posso non rifiutare l'ipotesi di normalità dei residui, anche attraverso il calcolo del relativo p-value;

- **il coefficiente** $\bar{R}^2$, correzione del coefficiente R^2 che spiega la porzione di varianza della variabile risposta spiegata dal modello.
  E' considerabile un indice di adattamento del modello ai dati e quindi della capacità di prediction, con valore variabile tra 0 e 1. Più il coefficiente è vicino ad 1 e meglio spiega la variabilità dei dati osservati, quindi esegue una migliore previsione. Data la numerosità n del dataset su cui si basa il modello e il numero k di regressori nel vettore β, l'$\bar{R}^2$ si calcola con la formula:

  $$\bar{R}^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2) = 1 - \frac{n-1}{n-k-1}\frac{SSR}{SST}$$

  dove SSR è la devianza residua, calcolata con la formula $SSR = \sum_{i=1}^{n}(Y - \hat{Y})^2$, mentre SST è la devianza totale, calcolata con la formula $SST = \sum_{i=1}^{n}(Y - \bar{Y})^2$, $\bar{Y} = \sum_{i=1}^{n} \frac{Y_i}{n}$;

- **il coefficiente AIC**, o coefficiente d'informazione di Akaike, che è un tipo di coefficiente che è utile solo se confrontato con i coefficienti di altri modelli; da solo non dice niente.
  Piuù basso è, meglio il modello si adatta ai dati. Si calcola con la formula:

  $$AIC = 2n - 2\ln(L(\beta; X))$$

  dove n è la numerosità campionaria e $L(\cdot)$ è la funzione di massima verosimiglianza del modello, usata per stimare i regressori $\hat{\beta}$.

- **il coefficiente BIC**, o coefficiente d'informazione bayesiano, il quale valore ha la stessa funzione dell'AIC (più basso è, meglio il modello si adatta ai dati) ma si calcola con la formula:

  $$BIC = -2\log(L(\beta; X)) + k\log(n)$$

  dove, oltre ai valori spiegati in precedenza, abbiamo anche k, ovvero il numero di regressori da stimare.
  La differenza con l'AIC, oltre che nella formula, sta nel fatto che il BIC è un migliore indice di overfitting, mentre l'AIC è un migliore indice di underfitting (più basso è e meglio il modello si adatta sia ai dati di training che ai dati di test). Per cui, rispetto all'AIC, forse, è leggermente meno versatile.

Per ogni regressore β_i, ove possibile, rileveremo invece:

- **Stima di β_i, ovvero** $\hat{\beta}_i$, contenuta nel vettore $\hat{\beta}$ calcolato con la formula dei minimi quadrati:

  $$\hat{\beta} = (X'X)^{-1}X'Y$$

  con X matrice del modello e Y vettore delle variabili risposta. Da notare che, essendo un valore stimato, si distribuisce come $\hat{\beta} \sim t_{n-p}(\hat{\beta}, s^2(X'Y)^{-1})$, con n numerosità campionaria, p numero di regressori e s^2 stima della varianza residua σ^2, calcolata con la formula $s^2 = \frac{SSR}{n-1}$;

- **il valore** $sd(\hat{\beta}_i)$, stima dello standard error di $\hat{\beta}_i$ calcolato con la formula $\sqrt{(s^2(X'Y)^{-1}_{ii})}$ ($s^2(X'Y)^{-1}$ è una matrice di varianze e covarianze), dove s^2 è la stima della varianza dei residui (differenze tra valori stimati dal modello e variabile risposta);

- **Valore della stima della statistica test** T_i, per calcolare la significatività dei coefficienti $\hat{\beta}$ ($H0 : \beta_i = 0, H1 = \bar{H}0$, i=1,...,p), calcolata con la formula

  $$T_i = \hat{\beta}_i / sd(\hat{\beta}_i)$$

  Da ricordarsi che sotto l'ipotesi nulla $T_i \sim t_{n-p}(0, 1)$;

- **Il p-value** p, indicatore della significatività del test $H0$ per cui si calcola la statistica test T_i. p si trova con la formula

  $$p = 2 * P(T > T_i)$$

  Se $p > 0.05$ allora il nostro coefficiente β_i non sarà significativo, ovvero potrà essere considerato come pari a 0 e non fondamentale per la stima di $\hat{Y}$.

I valori delle significatività, tuttavia, non possono essere presi così seriamente: questo perché essi non sono indipendenti. Infatti all'aumentare o diminuire del numero di regressori i loro valori cambiano, di conseguenza non sono la cosa primaria da guardare per la bontà di adattamento di un modello.
Se un coefficiente $\hat{\beta}_i$ presenterà come valori appena elencati il valore NA, vuol dire che la variabile relativa a $\hat{\beta}_i$ è linearmente dipendente da altre variabili, oppure che il modello è troppo saturo e che quindi non è stato possibile stimare i valori di $\hat{\beta}_i$. In quel caso, quindi, non li trascriveremo neanche.
Non verranno usati, infine, modelli con interazioni tra variabili poiché meno precisi nella prediction, rispetto ai modelli senza interazioni.

6.2 Modello saturo

Il modello lineare normale saturo è un tipo di modello che comprende tutte le variabili del dataset:

$$\hat{Y} = f(X) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$$

Essendoci 51 variabili per 148 unità statistiche si verifica il fenomeno spiegato alla fine del **paragrafo 6.1** di esagerata saturazione del modello. I coefficienti con valore NA, quindi, come accennato, non li trascriveremo.
La numerosità campionaria del dataset di training è di molto maggiore di 30, per cui automaticamente, per il teorema del limite centrale mi viene da non rifiutare l'ipotesi sia di normalità dei dati che quindi anche l'ipotesi di normalità dei residui.
Vediamo comunque un grafico di confronto tra quantili empirici e pratici dei residui, detto **qqplot**, per vedere se il test di normalità dei residui può essere accettato definitivamente:

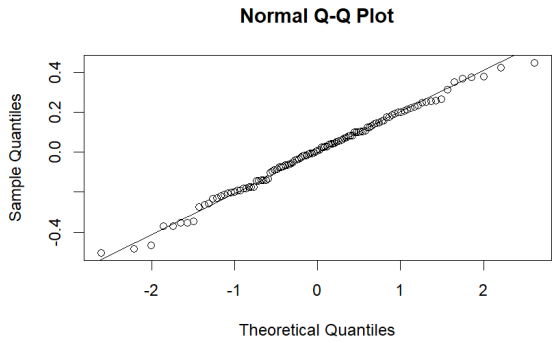


Figura 76: qqplot residui del modello saturo

Come possiamo vedere la linea, indicante i quantili empirici dei residui, e i puntini, indicanti i quantili pratici dei residui, coincidono abbastanza, quindi posso dire che i residui sono normali.
Vediamo i valori del vettore di regressori $\hat{\beta}$:

Coefficiente	$\hat{\beta}_i$	$sd(\hat{\beta})$	T_i	p
Intercetta ($\hat{\beta}_0$)	237.6	529.8	0.448	0.6554
Year	-0.0885	0.2609	-0.339	0.7355
Exp	0.0008	0.0101	0.081	0.936
BikeDucati	-0.2538	0.1483	-1.711	0.0924
BikeHonda	0.0495	0.1325	0.373	0.7103
BikeKTM	0.0254	0.1354	0.188	0.8516
BikeSuzuki	-0.0276	0.1743	-0.158	0.8748
BikeYamaha	-0.1326	0.131	-1.012	0.3158
Bike_year	-0.0274	0.0926	-0.296	0.7684
Q_sessionQ2	-0.2339	0.0913	-2.562	0.013*
Q_Lap	-0.0491	0.0368	-1.334	0.1872
Q_Top_speed	-0.0269	0.0107	-2.505	0.015*
Q_Sky_condCloudy	-1.72	0.802	-2.144	0.0361*
Q_Sky_condPartly cloudy	2.993	3.839	0.78	0.4387
Q_Air_temp	-0.1073	0.1497	-0.716	0.4766
Q_Humidity	8.738	7.279	1.2	0.2348
Q_Ground_temp	0.0367	0.0427	0.860	0.3931
Q_Front_tyreSM	-0.0581	0.1205	-0.482	0.6315
Q_Front_tyreSS	-0.1328	0.1947	-0.682	0.4979
Q_Front_tyre_age	8.08	0.0333	0.243	0.8091
Q_Rear_tyre_age	0.1858	0.0662	2.808	0.0067**

Coefficiente	$\hat{\beta}_i$	$sd(\hat{\beta})$	T_i	p
P1_Time	0.1319	0.1039	1.269	0.2094
P1_Lap	0.0142	0.0086	1.648	0.1046
P1_Top_speed	0.0308	0.0122	2.515	0.0147*
P1_Sky_condPartly cloudy	0.902	4.312	0.209	0.835
P1_Air_temp	0.2573	0.5183	0.497	0.6214
P1_Front_tyreSM	0.0008	0.1567	0.005	0.9962
P1_Front_tyreSS	-0.0753	0.1335	-0.564	0.575
P1_Front_tyre_age	0.0057	0.0109	0.520	0.6049
P1_Rear_tyreSM	-0.1888	0.3086	-0.612	0.5431
P1_Rear_tyreSS	-0.2605	0.3629	-0.718	0.4758
P1_Rear_tyre_age	-0.0062	0.0097	-0.644	0.5218
P2_Time	0.3132	0.1159	2.702	0.00897**
P2_Lap	0.0063	0.011	0.576	0.5667
P2_Top_speed	-0.0044	0.0116	-0.379	0.7061
P2_Front_tyreSM	0.1265	0.1186	1.067	0.2902
P2_Front_tyreSS	-0.0714	0.2357	-0.303	0.763
P2_Front_tyre_age	0.0085	0.0163	0.523	0.6032
P2_Rear_tyreSS	-0.1597	0.197	-0.81	0.4209
P2_Rear_tyre_age	-0.0089	0.0186	-0.477	0.6351
P3_Time	0.4355	0.097	4.49	< 0.0001***
P3_Lap	0.0039	0.0119	0.327	0.7446
P3_Top_speed	-0.0044	0.0115	-0.382	0.7036
P3_Front_tyreSM	0.0142	0.1661	0.086	0.9319
P3_Front_tyreSS	-0.0417	0.2149	-0.194	0.8467
P3_Front_tyre_age	0.0136	0.0102	1.332	0.188
P3_Rear_tyreSS	0.2259	0.1797	1.257	0.2136
P3_Rear_tyre_age	-0.0158	0.0196	-0.806	0.4234
Factory_RiderYes	-0.0409	0.1584	-0.258	0.7972
Nat2Ita	0.254	0.1555	1.633	0.1077
Nat2Oth	0.2216	0.1562	1.419	0.16126
Nat2Spa	0.1712	0.1479	1.157	0.2518

Le variabili che più hanno influenza nel modello sono quelle con i p-value minori di 0.05 e posso dire che siano le velocità di punta rilevate in qualifica e nelle P1, l'usura del pneumatico posteriore in qualifica ed i tempi sul giro rilevati nelle P2 e nelle P3. Per quanto riguarda Q_Sky_cond, direi di non includerla nella lista, visto che è significativo solo il regressore relativo alla modalità Cloudy.
Il nostro modello presenta poi i seguenti valori, per gli indici di adattabilità del modello ai dati:

Coefficiente	**Valore**
R^2 corretto	0.9215
AIC	62.4
BIC	206.005
MSE (s^2)	0.204

Posso dire dall'R^2 che, nonostante alcuni coefficienti non siano definiti nel valore, il modello di per sé è anche molto buono, essendo l'R^2 alquanto vicino ad 1. L'MSE di 0.204 s^2 (secondi quadrati) è considerabile anche buono, se teniamo conto del fatto che il dataset di test su cui si basa il calcolo dell'MSE presenta ben 37 unità statistiche, che sono circa poco più di una dozzina in più del numero di piloti che ogni anno prende parte al gran premio (nel senso che se ci fossero meno unità probabilmente sarebbe ancora più basso).

6.3 Modello di regressione lineare stepwise backward

E' un modello lineare trascrivibile sempre con la formula

$$\hat{Y} = f(X) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

solo che, a partire dal modello saturo, vengono rimosse da esso, passo dopo passo, le variabili che comporterebbero un aumento dell'AIC, del BIC o dell'MSE oppure una riduzione del coefficiente $\bar{R}^2$. Infatti, ad essere presenti nel modello, non sono più tutti i p regressori, bensì solo i q "migliori". Da notare che, probabilmente, avremo nel modello regressori non significativi, presenti per il fatto che la loro rimozione avrebbe comportato un aumento di AIC, BIC o MSE oppure una riduzione dell'$\bar{R}^2$.

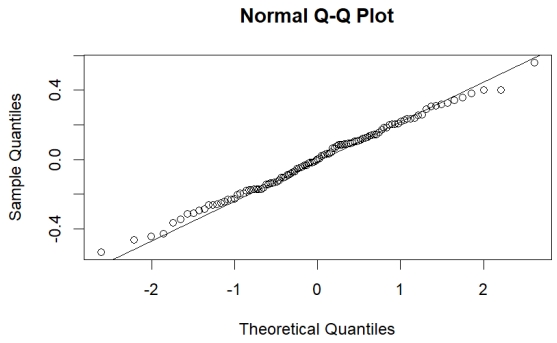


Figura 77: qqplot dei residui del modello di tipo stepwise backward

Anche per questo modello non rifiuto l'ipotesi nulla di normalità dei residui.
Diamo un'occhiata ai regressori:

Coefficiente	$\hat{\beta}_i$	$sd(\hat{\beta})$	T_i	p
Intercetta ($\hat{\beta}_0$)	367	81.47	4.504	< 0.0001***
Year	-0.1784	0.039	-4.571	< 0.0001***
BikeDucati	-0.1656	0.0869	-1.906	0.06
BikeHonda	-0.0207	0.0934	-0.221	0.8254
BikeKTM	0.0565	0.0941	0.6	0.5498
BikeSuzuki	-0.0497	0.1293	-0.384	0.7017
BikeYamaha	-0.1605	0.0921	-1.743	0.085
Q_sessionQ2	-0.2355	0.0639	-3.685	0.0004***
Q_Lap	-0.0588	0.0166	-3.54	0.0006***
Q_Top_speed	-0.0331	0.0073	-4.535	< 0.0001***
Q_Sky_condCloudy	-1.668	0.4804	-3.473	0.0008***
Q_Sky_condPartly cloudy	1.848	0.5043	3.664	0.0004***
Q_Humidity	8.996	2.387	3.769	0.0003***
Q_Front_tyreSM	0.0008	0.0765	0.010	0.992
Q_Front_tyreSS	-0.209	0.1213	-1.723	0.0885
Q_Rear_tyre_age	0.2096	0.0493	4.253	< 0.0001***
P1_Time	0.1403	0.0652	2.152	0.0342*
P1_Lap	0.0107	0.0054	1.976	0.0513
P1_Top_speed	0.0262	0.0081	3.232	0.0017**
P1_Air_temp	0.0868	0.0262	3.31	0.0014**
P2_Time	0.2936	0.0733	4.007	0.0001***
P3_Time	0.4557	0.0674	6.764	< 0.0001***
P3_Front_tyre_age	0.0168	0.0076	2.231	0.0283*
P3_Rear_tyreSS	0.3022	0.1092	2.768	0.0069**
P3_Rear_tyre_age	-0.23	0.0144	-1.594	0.1147

I coefficienti relativi alle motociclette (Bike) non sono significativi, tuttavia la loro mancata rimozione dal modello implica che quest'ultimo si sarebbe adattato peggio ai dati.
Come possiamo leggere, quindi, i coefficienti di un modello lineare normale? Facciamo un esempio, basandoci su quelli appena ricavati dal modello backward:

- **L'intercetta** può tendenzialmente indicare una sorta di valore medio della mia variabile risposta, che però qua è veramente troppo alto per essere considerato tale;
- il regressore relativo a **Year** è negativo, per cui al progredire del tempo il valore della variabile risposta, Q_Time, diminuisce;
- i coefficienti relativi alle moto (**Bike**) ci dicono che rispetto alla modalità di riferimento (Aprilia) tutte le motociclette, a parità di pilota e a parte KTM, sarebbero in grado di effettuare tempi sul giro in qualifica più bassi, soprattutto Ducati e Yamaha;
- **Q_session** ci dice che i tempi sul giro nella sessione di qualifica Q2 sono tendenzialmente un paio di decimi di secondo più bassi rispetto ai tempi della Q1, modalità di riferimento;
- **Q_Lap** ci dice che in media, in qualifica, i piloti si migliorano ad ogni giro di pista, senza tener conto del consumo delle gomme, di circa mezzo decimo di secondo;
- **Q_Top_Speed** ci dice che ad ogni chilometro orario in più alla fine del rettilineo si guadagnerebbero in media circa 3 centesimi di secondo. Al contrario, i piloti che hanno fatto registrare velocità di punta più alte nelle P1, come possiamo vedere da **P1_Top_speed** hanno avuto tempi sul giro di circa 0.026 secondi più alti, rispetto alla media, per ogni chilometro orario in più;
- i regressori relativi a **Q_Sky_cond** ci dicono che se, in qualifica, il cielo è nuvoloso, allora i tempi sono in media 1.67 secondi più veloci rispetto a quando il cielo è limpido (Clear è modalità di riferimento), mentre se è parzialmente nuvoloso i tempi sono 1.8 secondi più lenti. **Attenzione**: considerazione a prendere con le pinze, poiché abbiamo sì 25 rilevazioni con modalità di Q_Sky_cond uguale a Cloudy, ma sono tutte della stessa annata, ovvero il 2023, dove i tempi sul giro in qualifica sono stati molto rapidi;
- siccome **Q_Humidity** ha dominio pari a [0:1], il suo regressore ci indica che ad ogni aumento centesimale dell'umidità in qualifica, il tempo sul giro di quella sessione incrementa di circa 0.09 secondi (divisione del coefficiente per 100);
- dai coefficienti relativi a **Q_Front_tyre**, capiamo che la gomma soft anteriore è la migliore scelta: i tempi sul giro effettuati con quella sono circa 2 decimi di secondo più bassi rispetto a quelli effettuati con la gomma dura (modalità di riferimento), i quali tempi però sono di circa un millesimo di secondo più rapidi rispetto a quelli con la gomma media (nelle competizioni ogni millesimo è fondamentale);
- **Q_Rear_tyre_age** ci dice che, in qualifica, più lo pneumatico posteriore è usurato e più il tempo sul giro si alza: in specifico ogni giro sarebbe 2 decimi più lento del precedente, in media, man mano che il pneumatico si usura, per cui conviene eseguire il tempo quando il pnuematico è nuovo, o comunque il meno usurato possibile;
- **P1_Time** ci dice che per ogni secondo che un pilota ha eseguito un tempo sul giro nelle P1 più lento, il suo tempo in qualifica ne ha risentito, ovvero è stato più lento, di circa 140 millesimi di secondo, mentre **P2_Time** ci dice che per ogni secondo che il tempo sul giro di un pilota è stato più lento nelle P2, il tempo in qualifica ne ha risentito di circa 3 decimi di secondo (294 millesimi). Per quanto riguarda le P3 (**P3_Time**) la perdita sarebbe di quasi mezzo secondo (456 millesimi) per ogni secondo che il pilota è stato più lento rispetto alla media;
- per ogni grado centigrado che la temperatura dell'aria nelle P1 (**P1_Air_temp**) è stata più alta, i tempi sul giro delle qualifiche ne hanno risentito di circa 9 centesimi di secondo;

- **P3_Front_tyre_age** mi dice che più un pilota ha fatto registrare tempi sul giro con pneumatici consumati nelle P3 e più ne ha risentito anche il suo tempo in qualifica: in specifico si parla di 0.017 secondi per ogni giro di usura. **P3_Rear_tyre_age** mi dice che invece il contrario: in specifico, per ogni giro di gomma posteriore in più, con cui un pilota ha effettuato un tempo sul giro nelle P3, quel pilota potrebbe effettuare tempi sul giro in qualifica di più di 2 decimi di secondo più bassi;
- infine **P3_Rear_tyre** ci dice che effettuare un tempo sul giro nelle P3 con uno pneumatico posteriore morbido influenza il tempo sul giro in qualifica in negativo, specificamente di 3 decimi di secondo rispetto all'utilizzo di un pneumatico posteriore medio (modalità di riferimento).

I coefficienti, in un modello di regressione lineare, vanno sempre letti nella maniera appena spiegata, per cui la lettura nei modelli successivi sarà effettuata con lo stesso modus operandi.
Diamo un'occhiata agli indicatori di bontà di adattamento del modello:

Coefficiente	**Valore**
R^2 corretto	0.9388
AIC	22.54
BIC	92.988
MSE (s^2)	0.165

Il modello è, in tutto e per tutto, migliore del modello saturo, ovviamente: oltre a non contenere coefficienti dal valore indefinito ha dei migliori valori per gli indicatori di bontà di adattamento (AIC, BIC ed MSE più bassi, $\bar{R}^2$ più alto e vicino ad 1). Per cui, per ora, consideriamo questo come il modello migliore.

6.4 Modello di regressione lineare stepwise forward

Funziona esattamente all'opposto del modello backward: a partire dal modello nullo, ovvero il modello con la sola intercetta β_0, si aggiungono le variabili che man mano vanno ad abbassare l'AIC, il BIC o l'MSE oppure ad alzare l'$\bar{R}^2$. La formula del modello è sempre la stessa:

$$\hat{Y} = f(X) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

Diamo un'occhiata al test d'ipotesi sulla normalità dei residui:

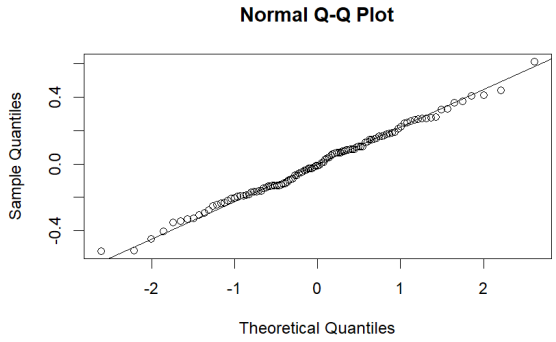


Figura 78: qqplot dei residui del modello di tipo stepwise forward

Anche questo modello ha dei residui distribuibili normalmente.

I regressori presentano i seguenti valori:

Coefficiente	$\hat{\beta}_i$	$sd(\hat{\beta})$	T_i	p
Intercetta ($\hat{\beta}_0$)	362.3505	83.5776	4.335	< 0.0001***
P2_Time	0.2791	0.0737	3.788	0.0003***
P3_Time	0.4828	0.0719	6.715	< 0.0001***
Q_Front_tyreSM	-0.0075	0.0662	-0.114	0.9094
Q_Front_tyreSS	-0.1758	0.119	-1.478	0.1431
Q_Rear_tyre_age	0.2213	0.0488	4.53	0.0001***
Q_Front_tyre_age	-0.0036	0.0214	-0.166	0.8687
P3_Rear_tyre_age	-0.0274	0.0146	-1.873	0.0644
Q_sessionQ2	-0.2532	0.0658	-3.847	0.0002***
Year	-0.1723	0.04	-4.421	< 0.0001***
P2_Sky_condPartly cloudy	-0.5188	0.1564	-3.317	0.0013**
Q_Air_temp	-0.1141	0.0373	-3.054	0.003**
Q_Top_speed	-0.0327	0.0074	-4.425	< 0.0001***
P1_Top_speed	0.0249	0.0082	3.041	0.0031**
P1_Time	0.1373	0.0614	2.237	0.0279*
P1_Air_temp	0.0737	0.0305	2.415	0.0179*
Q_Lap	-0.0572	0.0261	-2.196	0.0308*
P1_Lap	0.0113	0.0054	2.076	0.0409*
P3_Rear_tyreSS	0.2733	0.1066	2.565	0.0121*
P3_Front_tyre_age	0.0164	0.0076	2.166	0.0331*
BikeDucati	-0.1668	0.0882	-1.891	0.062
BikeHonda	-0.0278	0.0946	-0.294	0.7698
BikeKTM	0.0391	0.094	0.416	0.6781
BikeSuzuki	-0.0716	0.1318	-0.543	0.5885
BikeYamaha	-0.1734	0.0932	-1.861	0.0661

Apparte i coefficienti relativi a Q_Front_tyre e Bike, tutti i coefficienti sono significativi, quindi le relative variabili hanno effetto sulla variabile risposta.
Diamo un'occhiata ai coefficienti di bontà di adattamento del modello:

Coefficiente	**Valore**
R^2 corretto	0.9371
AIC	25.629
BIC	96.077
MSE (s^2)	0.169

E' molto simile al modello backward, ma quest'ultimo, oltre che ad avere l'$\bar{R^2}$ più alto di poco più di un centesimo, ha soprattutto sia l'AIC che il BIC più bassi. Inoltre l'MSE del modello forward è di 4 s^2 più alto rispetto a quello del modello backward, per cui tendo ancora a preferire il modello backward se voglio effettuare prediction.

6.5 Modello di regressione lineare stepwise both

Ha di per sè la stessa funzione di regressione:

$$\hat{Y} = f(X) = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q$$

Le migliori variabili, tuttavia, vengono aggiunte o tolte, a partire dal modello nullo, in base all'impatto che hanno su $\bar{R^2}$, AIC e BIC. E' un modello, quindi, che mescola il procedimento del modello forward con quello del modello backward.

Diamo quindi un'occhiata al grafico relativo all'ipotesi di normalità dei residui:

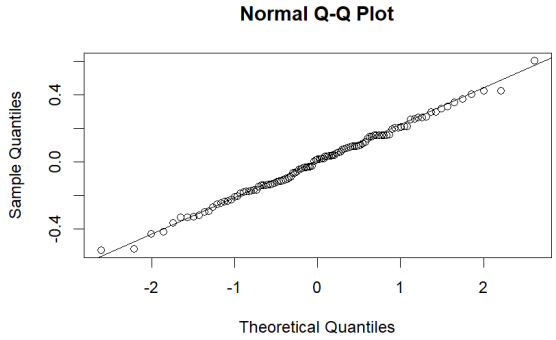


Figura 79: qqplot dei residui del modello di tipo stepwise both

Non rifiuto nemmeno per questo modello l'ipotesi di normalità dei residui. I regressori presentano i seguenti valori:

Coefficiente	$\hat{\beta}_i$	$sd(\hat{\beta})$	T_i	p
Intercetta ($\hat{\beta}_0$)	357.9	83.87	4.267	< 0.0001***
P2_Time	0.2778	0.07316	3.797	0.0003***
P3_Time	0.4857	0.0714	6.804	< 0.0001***
Q_Rear_tyre_age	0.2175	0.0479	4.539	< 0.0001***
Q_sessionQ2	-0.2537	0.0647	-3.919	0.0002***
Year	-0.1708	0.0392	-4.357	< 0.0001***
P2_Sky_condPartly cloudy	-0.5034	0.1546	-3.257	0.0016**
Q_Air_temp	-0.108	0.0372	-2.907	0.0046**
Q_Top_speed	-0.0323	0.0072	-4.47	< 0.0001***
P1_Top_speed	0.02574	0.0081	3.194	0.002**
P3_Rear_tyreSS	0.3167	0.1058	2.993	0.0036**
Q_Lap	-0.0598	0.0166	-3.595	0.0005***
P1_Air_temp	0.0692	0.0305	2.268	0.0258*
P1_Time	0.1428	0.06	2.378	0.0196*
P1_Lap	0.0109	0.0054	2.034	0.045*
BikeDucati	-0.1861	0.0835	-2.229	0.0284*
BikeHonda	-0.0478	0.0934	-0.51	0.6113
BikeKTM	0.0352	0.0922	0.381	0.7038
BikeSuzuki	-0.0732	0.1329	-0.551	0.5833
BikeYamaha	-0.184	0.0922	-1.996	0.049*
P3_Front_tyre_age	0.0153	0.0076	2.019	0.0465*
P3_Rear_tyre_age	-0.0274	0.0144	-1.907	0.0598
P3_Front_tyreSM	-0.0007	0.0766	-0.01	0.9924
P3_Front_tyreSS	-0.2306	0.1345	-1.715	0.0899

I risultati, sia a livello di significatività che a livello di valori dei coefficienti (e quindi anche di quello che è il legame tra variabili concomitanti e variabile risposta), sono molto simili agli altri 2 modelli di regressione stepwise. Noto che, curiosamente, come nel modello forward, viene tenuto conto delle condizioni meteo delle P2; noto poi che è l'unico modello, tra quelli di tipo stepwise, che tiene conto di tutte le variabili relative agli pneumatici, sia anteriori che posteriori, nelle P3 (sia mescole delle gomme che usura di esse). Per cui, per vedere la sua efficacia, devo andare a controllare i coefficienti di adattamento del modello ai dati:

Coefficiente	Valore
R^2 corretto	0.9383
AIC	22.741
BIC	90.479
MSE (s^2)	0.18

L'MSE è il più alto tra quelli visti nella stepwise regression. Tuttavia, l'$\bar{R}^2$ è molto poco più basso di quello della regressione backward, mentre l'AIC è molto poco più alto, segno di una similarità nei 2 modelli. Il BIC è il più basso di tutti, per cui questo modello lo considererei migliore almeno rispetto al modello forward.

6.6 Modello di regressione con metodo di regolarizzazione Ridge

Ci spostiamo ora sui modelli basati su metodi di regolarizzazione: essi comprendono di base tutte le variabili, come i modelli saturi, solo che applicando un termine di penalità λ, detto **parametro di tuning**, alla stima dei coefficienti β, cercano di fare in modo che il maggior numero di esse abbia regressore pari a 0. La formula di stima di β è sempre quella dei minimi quadrati, ma con la penalità inclusa diventa la seguente:

$$\hat{\beta} = (X'X + \lambda I)^{-1} X'Y$$

Come si decide λ? Lo si decide selezionandolo da una lista di λ, per ognuno dei quali calcolo l'MSE penalizzato. Nel caso del metodo di regolarizzazione Ridge, la penalità applicata all'MSE, detta **shrinkage penalty**, è di tipo quadratico. L'MSE verrà quindi calcolato con la seguente formula:

$$MSEp = \sum_{i=1}^{n} (y_i - \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

dove la shrinkage penalty è pari a $\lambda \sum_{j=1}^{p} \beta_j^2$.
Verrà selezionato, infine, il termine λ che più minimizza l'MSE penalizzato e di conseguenza verrà scelto il corrispondente vettore β dei regressori penalizzati.
Nel nostro caso, applicando il metodo di regolarizzazione Ridge ai nostri dati di training, è risultato che il valore di λ che più minimizza, da una lista di 2000 possibili valori, l'MSE penalizzato è 1872, con un MSE di 0.174. Graficamente i vettori β si evolvono nella seguente maniera:

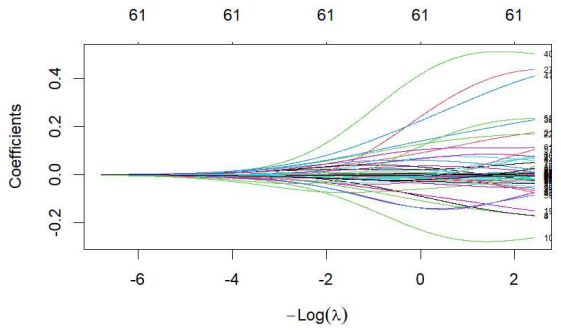


Figura 80: Evoluzione dei coefficienti, al variare del negativo del logaritmo di λ, nel modello con metodo di regressione ridge

All'aumentare del valore di λ, il negativo del suo logaritmo diminuisce, quindi ci sono sempre più coefficienti tendenti a 0: in quel caso, si parla di **regolarizzazione forte**. Al contrario, al diminuire del valore di λ, il negativo del suo logaritmo aumenta, quindi ci sono sempre meno coefficienti tendenti a 0: in quel caso, si parla di **regolarizzazione debole** L'ideale sarebbe ottenere il maggior numero di valori di regressori del vettore β, quindi, più vicini possibile alla sinistra del grafico, nella **figura 80**.

I valori di β ottenuti, alla fine, sono stati i seguenti:

Coefficiente	$\hat{\beta}_j$	**Coefficiente**	$\hat{\beta}_j$
Intercetta (β_0)	0	P1_Rear_tyreSM	-0.0091
Year	-0.0345	P1_Rear_tyreSS	-0.0505
Exp	0.0039	P1_Rear_tyre_age	-0.0009
BikeDucati	-0.1579	P2_Time	0.2078
BikeHonda	-0.0148	P2_Lap	0.0007
BikeKTM	0.0681	P2_Top_speed	-0.0018
BikeSuzuki	0.0839	P2_Sky_condPartly cloudy	-0.0099
BikeYamaha	-0.157	P2_Air_temp	0.0003
Bike_year	-0.0233	P2_Humidity	0.5105
Q_sessionQ2	-0.2738	P2_Ground_temp	0.0082
Q_Lap	-0.0276	P2_Front_tyreSM	0.0581
Q_Top_speed	-0.0132	P2_Front_tyreSS	-0.1026
Q_Sky_condCloudy	0.0012	P2_Front_tyre_age	0.0062
Q_Sky_condPartly cloudy	0.0419	P2_Rear_tyreSS	-0.0449
Q_Air_temp	-0.0027	P2_Rear_tyre_age	-0.0029
Q_Humidity	-0.0116	P3_Time	0.3714
Q_Ground_temp	0.0049	P3_Lap	0.0052
Q_Front_tyreSM	-0.0129	P3_Top_speed	-0.0088
Q_Front_tyreSS	-0.1342	P3_Sky_condPartly cloudy	0.0004
Q_Front_tyre_age	-0.006	P3_Air_temp	0.0025
Q_Rear_tyre_age	0.1585	P3_Humidity	0.2269
P1_Time	0.1651	P3_Ground_temp	-0.0071
P1_Lap	0.0082	P3_Front_tyreSM	-0.0158
P1_Top_speed	0.0058	P3_Front_tyreSS	-0.0525
P1_Sky_condPartly cloudy	0.0223	P3_Front_tyre_age	0.0037
P1_Air_temp	-0.0029	P3_Rear_tyreSS	0.0705
P1_Humidity	0.4195	P3_Rear_tyre_age	-0.0167
P1_Ground_temp	-0.007	Factory_RiderYes	-0.1055
P1_Front_tyreSM	0.0349	NatIta	0.0524
P1_Front_tyreSS	-0.0426	NatOth	0.1111
P1_Front_tyre_age	0.0011	NatSpa	0.0083

Come possiamo notare, apparte l'intercetta nessun regressore è stato posto pari a 0. Secondo il metodo Ridge, noto comunque per essere un metodo che "shrinka" (ovvero "taglia" o "riduce") molto poco, tutte le variabili possono avere una funzione "vitale" per la stima del tempo sul giro in qualifica di un pilota. Ad esempio, guardando il coefficiente relativo a **Factory_Rider**, posso notare come la differenza media di tempo sul giro, a parità di condizioni, tra un pilota titolare ed un pilota sostitutivo, o comunque un tester, è di circa 1 decimo di secondo a vantaggio del pilota ufficiale.
Così come possiamo notare, dai coefficienti relativi alla variabile **Nat**, come tutti i piloti di nazionalità diversa da quella francese (modalità di riferimento) siano tendenzialmente più lenti, seppur di poco, rispetto ai francesi stessi.
Noi, tuttavia, vorremmo un modello che sintetizzi di più le nostre variabili. Da notare che, non essendo i metodi di regolarizzazione dei modelli normali lineari a tutti gli effetti, non è possibile ricavare i coefficienti AIC, BIC o $\bar{R}^2$; inoltre, per modelli di questo tipo, non andremo a verificare, sempre per lo stesso motivo, l'ipotesi nulla di normalità dei residui. Quindi, l'unico parametro di cui possiamo effettivamente tenere conto è l'MSE, come scritto in precedenza pari a 0.179, quindi più alto del modello di tipo stepwise backward (0.165).

6.7 Modello di regressione con metodo di regolarizzazione Lasso

Il modello con metodo di regolarizzazione Lasso prevede, come shrinkage penalty, un valore di tipo assoluto. L'MSEp, infatti, con questo metodo, si calcolerebbe con la seguente formula:

$$MSEp = \sum_{i=1}^{n}(y_i - \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

Ovviamente il metodo di selezione di λ, e di conseguenza il metodo di stima di β, sono gli stessi del modello Ridge.
Il valore di λ, proveniente da una lista di 2000 possibili valori, minimizzante l'MSE penalizzato, nella Lasso, è risultato essere 902, con un MSE di 0.161 (il più basso di tutti!), mentre i valori dei regressori sono risultati essere i seguenti:

Coefficiente	$\hat{\beta}_j$	**Coefficiente**	$\hat{\beta}_j$
Intercetta (β_0)	0	P1_Rear_tyreSM	0
Year	-0.0454	P1_Rear_tyreSS	0
Exp	0	P1_Rear_tyre_age	0
BikeDucati	-0.1315	P2_Time	0.3483
BikeHonda	-0.0175	P2_Lap	0
BikeKTM	0	P2_Top_speed	0
BikeSuzuki	0.0632	P2_Sky_condPartly cloudy	-0.0288
BikeYamaha	-0.0944	P2_Air_temp	0
Bike_year	0	P2_Humidity	0.4704
Q_sessionQ2	-0.2096	P2_Ground_temp	0
Q_Lap	-0.0342	P2_Front_tyreSM	0.0255
Q_Top_speed	-0.0127	P2_Front_tyreSS	0
Q_Sky_condCloudy	0	P2_Front_tyre_age	0
Q_Sky_condPartly cloudy	0	P2_Rear_tyreSS	0
Q_Air_temp	0	P2_Rear_tyre_age	0
Q_Humidity	0	P3_Time	0.4098
Q_Ground_temp	0	P3_Lap	0.0026
Q_Front_tyreSM	0	P3_Top_speed	-0.0003
Q_Front_tyreSS	-0.1119	P3_Sky_condPartly cloudy	0
Q_Front_tyre_age	-0.019	P3_Air_temp	0
Q_Rear_tyre_age	0.1709	P3_Humidity	0
P1_Time	0.1371	P3_Ground_temp	-0.0061
P1_Lap	0.0043	P3_Front_tyreSM	0
P1_Top_speed	0.0066	P3_Front_tyreSS	-0.0508
P1_Sky_condPartly cloudy	0	P3_Front_tyre_age	0
P1_Air_temp	0	P3_Rear_tyreSS	0.1196
P1_Humidity	0.0138	P3_Rear_tyre_age	-0.0155
P1_Ground_temp	0	Factory_RiderYes	0
P1_Front_tyreSM	0.0366	NatIta	0.0125
P1_Front_tyreSS	0	NatOth	0.0169
P1_Front_tyre_age	0	NatSpa	0

Il metodo Lasso taglia decisamente più variabili: secondo questo modello, come nei modelli stepwise regression, non conta se il pilota sia titolare o sostitutivo (il coefficiente relativo a **Factory_Rider** è pari a 0), ma conta molto di più, ad esempio, l'umidità rilevata nelle P2: se un pilota riscontrasse umidità massima (quindi 100 %, traducibile nella variabile come il massimo 1), il suo tempo sul giro in qualifica ne risentirebbe di quasi mezzo secondo.
Come in tutti gli altri modelli, noto poi che tra Q1 e Q2 (variabile **Q_Session**) ci sono circa 2 decimi di secondo di differenza in media a vantaggio della seconda, a parità di valore nelle altre variabili.
Se andiamo poi a vedere i tempi sul giro, durante anche il resto del fine settimana, notiamo che quello che più ha impatto di tutti è quello registrato nelle P3 (**P3_Time**): per ogni secondo che esso è stato

più lento, il tempo sul giro in qualifica ne risente di più di 4 decimi di secondo.
La moto più efficace sembra essere Ducati: con essa i piloti, a parità di condizioni, tendono ad essere più veloci in qualifica di poco più di un decimo, rispetto ai colleghi che corrono con Aprilia (modalità di riferimento della variabile **Bike**); Yamaha è il costruttore che più si avvicina alla velocità di Ducati, in qualifica. KTM, invece, è sullo stesso livello di Aprilia, mentre Suzuki, come già visto nel **paragrafo 2.4**, a causa dell'assenza di rilevazioni recenti, risulta essere la moto più lenta, essendo il suo tempo sul giro in qualifica mediamente 6 centesimi di secondo più lento di quello di Aprilia e quasi 2 decimi di secondo più lento rispetto a Ducati.
Infine, come visto anche nel modello backward, più lo pneumatico posteriore in qualifica è consumato e più il tempo sul giro in qualifica ne risente: si pagano poco meno di 2 decimi di secondo ad ogni giro che passa.

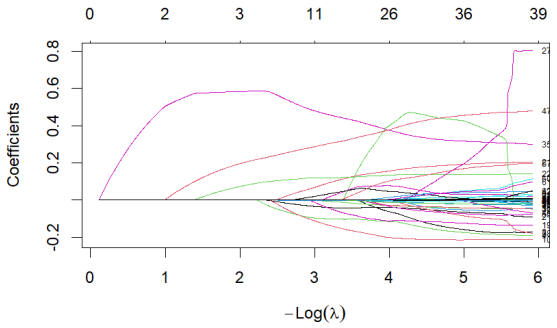


Figura 81: Evoluzione dei coefficienti al variare del logaritmo di λ nel modello con metodo di regressione lasso

All'aumentare di λ, si vede subuto come uno dei coefficienti non tendi quasi mai a 0. Inoltre, come possiamo ulteriormente notare, al diminuire di λ, i coefficienti sono molto meno lineari nell'evoluzione dei propri valori. Questo è tipico del modello con metodo di regolarizzazione Lasso.

6.8 Conclusione: qual è il modello migliore?

L'unico metodo comune a tutti i modelli, per verificare quale di essi si adatta meglio ai dati, è il confronto tra gli MSE. Escluderemo da subito, quindi, il modello normale saturo, che è l'unico ad avere coefficienti dal valore indefinito ed è il peggiore, di gran lunga, per quanto riguarda AIC, BIC e $\bar{R}^2$:

MODELLO	Backward	Forward	Both	Ridge	Lasso
MSE	0.165	0.169	0.18	0.174	0.161

Come possiamo vedere, a prima vista il modello migliore è il modello con metodo di regolarizzazione Lasso.
Tuttavia non voglio basarmi solo sul dataset di test per trarre le mie conclusioni, ma voglio vedere come si comportano i vari modelli nella prediction di tempi sul giro in qualifica in 3 annate distinte, per vedere se le conclusioni sono le stesse. Le 3 annate in questione, scelte abbastanza casualmente, sono 2018, 2021 e 2025.
Vediamo quindi cosa ci dicono i vari modelli, a partire dalla **prediction per i dati del 2018**. Per essa sono risultati, a partire da 2000 valori possibili, dei valori di λ nei metodi di regolarizzazione pari rispettivamente a 1946 per il metodo Ridge e 985 per il metodo Lasso: questo comporterà anche dei cambiamenti nei valori dei regressori β, che però non vedremo. Concentriamoci sugli MSE:

MODELLO	Backward	Forward	Both	Ridge	Lasso
MSE	0.108	0.11	0.113	0.126	0.128

Come possiamo notare, nella "previsione" (o stima) di dati passati i metodi di stepwise regression fanno un lavoro tendenzialmente migliore, soprattutto il modello di tipo backward che ha l'MSE più basso di tutti.
Vediamo cosa succede con i **dati del 2021**, per i quali sono risultati dei valori di λ nei metodi di regolarizzazione pari a 2000 per il metodo Ridge, quindi il massimo, e 1260 per il metodo Lasso:

MODELLO	Backward	Forward	Both	Ridge	Lasso
MSE	0.047	0.046	0.047	0.046	0.041

In questo caso, invece, notiamo che i metodi sono tutti quanti molto molto simili tra loro. La regressione Lasso tende ad essere la migliore, tuttavia anche gli altri metodi restano molto molto buoni. Da notare poi che gli MSE sono tutti quanti diminuiti, segno che forse i vari modelli si adattano meglio ai dati futuri che ai dati passati, il che è esattamente lo scopo della loro esistenza.
Vediamo, infine, cosa succede per i **dati del 2025**, per i quali sono risultati dei valori di λ, nei metodi di regolarizzazione, pari a 1838 per il metodo Ridge e 836 per il metodo Lasso:

MODELLO	Backward	Forward	Both	Ridge	Lasso
MSE	0.046	0.045	0.046	0.033	0.035

Rispetto ai dati del 2018 c'è un'inversione di trend: i modelli con metodi di regolarizzazione tendono a svolgere un lavoro migliore, seppur di poco, soprattutto il metodo Ridge (stranamente, dato che è quello con probabilmente più regressori di tutti).
In conclusione: siccome i modelli con metodi di regolarizzazione tendono a svolgere un lavoro migliore con dati più "contemporanei", mi viene da dire che siano quindi anche i migliori, senza sceglierne uno tra i 2 in specifico, nell'eseguire prediction. Tuttavia, anche i modelli basati sulla stepwise regression restano dei buonissimi modelli, per cui se usassi quelli non compierei un errore.

7 Conclusioni

Il dataset, per quanto mancanti numerosi dati, è utile a farci capire, attraverso anche semplici analisi esplorative numerosi dettagli: tra questi, ad esempio, il fatto che nonostante sia il loro circuito di casa, gli italiani non siano i piloti che più si esaltano. Al contrario, i piloti francesi abbiano un potenziale enorme su questa pista, addirittura probabilmente superiore a quello degli spagnoli, considerati da tutti tendenzialmente i migliori negli ultimi anni, e che allo stesso tempo i promotori della disciplina, ovvero dal 2025 Liberty Media[17], dovrebbero spingere sull'investimento da parte delle aziende nelle carriere di piloti di nazionalità diversa da italiana e spagnola: quasi il 70 % dei giri registrati nel nostro dataset è stato effettuato da piloti italiani o spagnoli. Per cui, con un investimento maggiore da parte delle aziende, ci sarebbe meno difficoltà, soprattutto dal punto di vista economico, per altri piloti nell'entrare a far parte di un campionato del mondo o comunque nel fare della passione per la guida della moto in pista un mestiere.
Durante il fine settimana del Gran Premio d'Italia, i piloti più veloci sono tendenzialmente sempre gli stessi, sia in prova che in qualifica e sono anche quelli con le moto più evolute e più veloci in rettilineo, mentre le condizioni meteo non variano più di tanto nè da una sessione all'altra, nè tantomeno da un giorno all'altro e sono molto correlate tra loro durante tutto il fine settimana.
La Ducati è il costruttore che storicamente ha avuto le moto più veloci su questa pista, mentre tralasciando Suzuki, ritiratasi a fine 2022 e comunque fino ad allora molto competitiva, Honda sembra essere, al contrario, il costruttore con le moto storicamente meno competitive.
Il dataset è riducibile in 3 componenti:

- **Velocità generica durante il fine settimana, in relazione alle condizioni atmosferiche e della pista ed al consumo degli pneumatici**;
- **Condizioni atmosferiche in prova**;
- **Competitività sul time attack**, quindi con l'esclusione delle P1, date le condizioni atmosferiche e degli pneumatici durante il weekend.

Oppure può essere spiegato da 6 fattori latenti:

- **Performance del pilota durante il weekend**, in base all'annata di rilevazione e della moto ed in base all'umidità riscontrata durante tutto il weekend, oltre che al consumo gomme nelle P3 e alla temperatura dell'asfalto in qualifica;
- **Temperature dell'aria e dell'asfalto in prova**;
- **Preparazione del pilota sul giro secco, in base alle umidità registrate durante il fine settimana**; **Preparazione del pilota alla sessione di qualifica, in base al consumo gomme e alla propria organizzazione della sessione** (quanti giri fare, quando spingere e che gomme usare);
- **Usure degli pneumatici nelle P1**;
- **Preparazione delle prove del venerdì**, soprattutto dal punto di vista dell'organizzazione e del consumo dello pneumatico anteriore (molto importante su un circuito dove saper frenar forte fa la differenza).

Capiamo, poi, che c'è un legame significativo tra le performance registrate in prova e quelle registrate in qualifica, riassumibile con le seguenti caartteristiche:

- più è elevata l'umidità del sabato e prima i piloti fanno registrare il proprio miglior tempo nelle P3;

[17] La Repubblica, *UE approva acquisizione di Dorna da parte di Liberty Media senza condizioni*, https://finanza.repubblica.it/News/2025/06/23/ue_approva_acquisizione_di_dorna_da_parte_di_liberty_media_senza_condizioni-133/

- le moto, potenzialmente, sono più veloci in rettilineo in qualifica quando le temperature dell'asfalto in prova sono più basse;
- avere in prova temperature dell'asfalto elevate ed una moto veloce in rettilineo, implicano la registrazione di un tempo sul giro in qualifica più basso;
- i piloti che arrivano prima al limite delle proprie performance e di quelle della propria moto nelle P2, hanno anche la moto più veloce in rettilineo in qualifica.

Le unità del dataset sono divisibili in 3 cluster, basati su numerosi criteri:

- **consumi degli pneumatici posteriori nelle P2 e nelle P3**, con un metodo di clustering gerarchico con legame completo, basato su distanze euclidee;
- **tempi sul giro in qualifica e annate delle moto**, con un metodo di clustering gerarchico con legame completo, basato su distanze di Manhattan;
- **tempi sul giro in qualifica e nelle P1**, con un metodo di clustering gerarchico con legame medio, basato su distanze di Manhattan, oppure con un metodo di clustering gerarchico con legame di Ward, basato su distanze euclidee;
- **numero di giri impiegati per registrare il proprio miglior tempo e velocità di punta nelle P3**, con un metodo di clustering gerarchico con legame di Ward, basato su distanze di Manhattan;
- **periodo di rilevazione**, con un metodo di clustering non gerarchico delle k-medie.

Le variabili più influenti per la previsione, in tempo reale, del tempo sul giro in qualifica potrebbero essere, oltre al tipo di sessione di qualifica, le semplici mescole ed usure degli pneumatici nelle le sessioni, appunto, di qualifica e di terze prove libere, così come i tempi sul giro di tutto il resto del fine settimana e l'umidità registrata nelle prove libere 2. Il tempo sul giro conviene, inoltre, prevederlo usando modelli basati su metodi di regolarizzazione, anzichè semplici modelli lineari normali o modelli lineari di tipo stepwise regression, poiché compiono un lavoro migliore con la previsione di tempi sul giro registrati più avanti nel tempo.

8 Bibliografia e sitografia

Le formule e le applicazioni pratiche sono state ricavate dalle seguenti fonti:

- **Capitolo 3:** Johnson, Wichern, *Applied Multivariate Statistical Analysis*, 2014:
  - **Paragrafo 3.1:** pagg. 430-455;
  - **Paragrafo 3.3:** pagg. 481-526;
- **Capitolo 4:** Johnson, Wichern, *Applied Multivariate Statistical Analysis*, 2014: pagg. 539-565;
- **Capitolo 5:** Johnson, Wichern, *Applied Multivariate Statistical Analysis*, 2014: pagg. 671-702;
- **Capitolo 6:**
  - **Paragrafo 6.2:** Johnson, Wichern, *Applied Multivariate Statistical Analysis*, 2014: pagg. 360-417;
  - **Paragrafi 6.3-6.6:** James, Witten, Hastie, Tibshirani *An Introduction to Statistical Learning with Applications in R*, 2021: pagg. 225-282.

I dati analizzati, invece, sono stati raccolti e riorganizzati a partire dalle rilevazioni registrate sui seguenti siti e file:

- `https://www.motogp.com/en/gp-results/2025/ita/motogp/q2/classification`
- `https://resources.motogp.com/files/results/2025/ITA/MotoGP/Q2/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2025/ita/motogp/q1/classification`
- `https://resources.motogp.com/files/results/2025/ITA/MotoGP/Q1/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2025/ita/motogp/fp2/classification`
- `https://resources.motogp.com/files/results/2025/ITA/MotoGP/FP2/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2025/ita/motogp/pr/classification`
- `https://resources.motogp.com/files/results/2025/ITA/MotoGP/PR/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2025/ita/motogp/fp1/classification`
- `https://resources.motogp.com/files/results/2025/ITA/MotoGP/FP1/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2024/ita/motogp/q2/classification`
- `https://resources.motogp.com/files/results/2024/ITA/MotoGP/Q2/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2024/ita/motogp/q1/classification`
- `https://resources.motogp.com/files/results/2024/ITA/MotoGP/Q1/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2024/ita/motogp/fp2/classification`
- `https://resources.motogp.com/files/results/2024/ITA/MotoGP/FP2/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2024/ita/motogp/pr/classification`
- `https://resources.motogp.com/files/results/2024/ITA/MotoGP/PR/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2024/ita/motogp/fp1/classification`
- `https://resources.motogp.com/files/results/2024/ITA/MotoGP/FP1/Analysis.pdf`
- `https://www.motogp.com/en/gp-results/2023/ita/motogp/q2/classification`

- https://resources.motogp.com/files/results/2023/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2023/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2023/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2023/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2023/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2023/ita/motogp/pr/classification
- https://resources.motogp.com/files/results/2023/ITA/MotoGP/PR/Analysis.pdf
- https://www.motogp.com/en/gp-results/2023/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2023/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2022/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2022/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2022/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2022/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2022/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2022/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2022/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2022/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2022/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2022/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2021/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2021/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2021/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2021/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2021/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2021/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2021/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2021/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2021/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2021/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2019/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2019/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2019/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2019/ITA/MotoGP/Q1/Analysis.pdf

- https://www.motogp.com/en/gp-results/2019/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2019/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2019/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2019/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2019/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2019/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2018/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2018/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2018/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2018/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2018/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2018/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2018/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2018/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2018/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2018/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2017/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2017/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2017/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2017/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2017/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2017/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2017/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2017/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2017/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2017/ITA/MotoGP/FP1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2016/ita/motogp/q2/classification
- https://resources.motogp.com/files/results/2016/ITA/MotoGP/Q2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2016/ita/motogp/q1/classification
- https://resources.motogp.com/files/results/2016/ITA/MotoGP/Q1/Analysis.pdf
- https://www.motogp.com/en/gp-results/2016/ita/motogp/fp3/classification
- https://resources.motogp.com/files/results/2016/ITA/MotoGP/FP3/Analysis.pdf
- https://www.motogp.com/en/gp-results/2016/ita/motogp/fp2/classification
- https://resources.motogp.com/files/results/2016/ITA/MotoGP/FP2/Analysis.pdf
- https://www.motogp.com/en/gp-results/2016/ita/motogp/fp1/classification
- https://resources.motogp.com/files/results/2016/ITA/MotoGP/FP1/Analysis.pdf