

Università degli studi di Padova



Facoltà di scienze statistiche

corso di laurea in

statistica e tecnologie informatiche

tesi di laurea

Studio del comportamento di visita a due siti web.

relatore: dott. Bruno Scarpa

Laureando: Yabre Yacouba

matricola: 523827

Anno accademico: 2008-2009

Voglio ringraziare la mia famiglia, per avermi sostenuto in questa straordinaria avventura.

Ringrazio inoltre il mio relatore il dott. Bruno Scarpa e la prof. Patrizia Amaglio dell' IPSS Bartolomeo Montagna di Vicenza per la pazienza che hanno avuto nel correggere gli errori dovuti alla lingua.

Infine ringrazio le signore dell' “Alliance française de Vicenza” senza le quale non sarei qui.

Indice

Introduzione.....	7
Analisi delle componenti principali: procedure di calcolo.....	9
•1. Introduzione.....	9
•2. Metodologia di calcolo delle componenti.....	10
a. Matrice dei dati, delle medie e di varianze-covarianze.....	10
b. Determinazione della varianza nell'ACP.....	11
c. Proprietà della varianza nell'ACP.....	15
d. Standardizzazione delle variabili di partenza.....	17
e. Scelta del numero delle componenti principali.....	18
f. Interpretazione delle componenti principali.....	19
Analisi dei dati relativi al dataset gazelle.....	23
•1. introduzione.....	23
•2. Analisi descrittive.....	28
•3. Differenze tra i buoni e i cattivi clienti.....	31
•4. Applicazione dell'ACP sui dati	35
a. Matrice di varianze-covarianze o matrice di correlazione?.....	35
b. Analisi dei dati relativi ai buoni clienti.....	35
i. Percentuale di varianza spiegata da ogni componente.....	35
ii. Interpretazione delle componenti trovate.....	38
Analisi dei comportamenti di visita al sito di Microsoft.....	45
•1. Introduzione.....	45
a. Breve presentazione degli obiettivi primari dell'autore.....	45
b. Presentazione dei dati.....	46
•2. Analisi esplorativa dei dati.....	48
• 3. Applicazione dell'analisi delle componenti principali sui dati.....	52

Bibliografia.....	61
-------------------	----

Introduzione

Gli ultimi decenni sono stati caratterizzati dall'esplosione delle nuove tecnologie. All'occorrenza, in questi ultimi anni assistiamo alla proliferazione della rete internet e alla creazione di tanti servizi ad esso connesso tale "social network", videoconferenze, commercio online etc.

Infatti in questi ultimi anni, si può dire che nessuno può fare almeno di usare internet, il che ovviamente contribuisce alla creazione di vere e proprie aziende online, e quindi di usare particolari tecniche per poter trarre informazioni su ad esempio il comportamento degli utenti che visitano particolari siti, in modo da sviluppare specifiche azioni di marketing.

Queste tecniche conosciute con il termine di "web mining" costituiscono oggi un potente strumento per analizzare dei dati avente tante variabili. Infatti il web mining raggruppa tanti strumenti statistici tale l'analisi dei gruppi (cluster analysis), l'analisi delle componenti principale etc.

Questi strumenti che oggi caratterizzano della statistica moderna sono usati per analizzare dati provenienti da tanti settori d'attività, perché permettono una sintesi accurata e facilmente interpretabile, anche da non statistici.

In questa sede useremo solo uno di questi strumenti, ovvero useremo solo l'analisi delle componenti principali per poter sintetizzare l'informazione contenuta nei dati. Cercheremo quindi nel primo capitolo di illustrare i metodi di calcoli delle componenti per passare negli ultimi due capitoli all'analisi dei dati provenienti da due siti di commercio online.

Capitolo I:

Analisi delle componenti principali: procedure di calcolo.

1. Introduzione

La disponibilità di grandi "basi di dati" permette oggi di archiviare grandi quantità di informazioni, e quindi la necessità di usare specifiche tecniche per poter oltrepassare il limite della semplice analisi descrittiva, che di solito si limita all'analisi di una o due variabili.

Questi metodi, tra cui l'analisi delle componenti principali fanno parte dell'ambito della statistica multivariata.

L'analisi delle componenti principali (ACP) introdotta da K. Pearson nel 1901 e sviluppati da H. Hotelling nel 1933, ha avuto un grande successo, perché oltre a presentare i risultati in modo chiaro e facilmente interpretabile, ha il pregio di avvalersi di rappresentazioni grafiche, che rendono il suo uso immediato anche da non esperti e in ambiti diversi.

Lo scopo primario di questa tecnica è la riduzione di un numero elevato di variabili in alcune variabili latenti. Questo avviene tramite l'individuazione di particolari combinazioni lineari dette appunto "componenti principali". Tali combinazioni saranno tra loro intercorrelate ed ordinate in modo tale che la prima combinazione sintetizza la quota massima possibile della variabilità totale, la seconda sintetizza la quota massima della variabilità totale dopo la prima, e così via.

L'individuazione delle variabili, come abbiamo detto, consiste in una trasformazione lineare delle variabili che proietta quelle originarie in un nuovo sistema cartesiano nel quale le variabili vengono ordinate in ordine decrescente di varianza spiegata. Pertanto, la variabile che spiega la maggior percentuale di varianza viene proiettata sul primo asse, la seconda sul secondo e così via.

In questo capitolo, vedremo le procedure e i risultati salienti di questa analisi per poi passare nei prossimi capitoli a delle applicazioni vere e proprie.

2. Metodologia di calcolo delle componenti.

a. Matrice dei dati, delle medie e di varianze-covarianze.

Supponiamo di voler studiare le caratteristiche di una certa popolazione, a questo proposito si rilevano p variabili su un insieme di n unità. Questi dati possono essere sintetizzati in una matrice detta matrice dei dati. Chiamiamo X questa matrice; una sua rappresentazione è la seguente.

$$X = \begin{pmatrix} x_{11}, x_{12}, \dots, x_{1p} \\ \dots\dots\dots \\ \dots\dots\dots \\ x_{n1}, x_{n2}, \dots, x_{np} \end{pmatrix}$$

Ogni colonna rappresenta una variabile e ogni riga un'unità.

Una prima sintesi dei dati consiste nell'individuare il vettore delle medie ovvero $\bar{x}$.

$$\bar{X} = (\bar{X}_1, \dots, \bar{X}_n)$$

Un'idea della relazione che esiste tra le variabili ovvero una sintesi delle informazioni relative alla distribuzione congiunta delle variabili può essere ottenuta attraverso i momenti secondi centrati, che costituiscono la matrice di varianze-covarianze.

$$S = \begin{pmatrix} s_{11}, s_{12}, \dots, s_{1p} \\ \dots\dots\dots \\ \dots\dots\dots \\ s_{p1}, s_{p2}, \dots, s_{pp} \end{pmatrix}$$

Tale matrice risulta simmetrica (ovvero $S'=S$) e semi-definita positiva. Va notato però che questa matrice presenta a volte degli inconvenienti, infatti, essa risente dell'unità di misura. Nel caso in cui volessimo fare un'analisi nella quale tutte le variabili non sono espresse nella stessa unità di misura, conviene considerare una sua versione standardizzata, cioè la matrice di correlazione.

Nel seguito, considereremo la matrice di varianze-covarianze ovvero ci mettiamo in un contesto in cui tutte le variabili hanno la stessa unità di misura.

b. Determinazione della varianza nell'ACP

L'ACP, come abbiamo detto, ha per obiettivo la determinazione di combinazioni lineari tra di loro indipendenti delle variabili di partenza, in modo da avere, dopo alcune semplificazioni ed eliminazioni delle ridondanze, una sintesi facilmente interpretabile dei dati.

Sia allora Y la matrice delle variabili trasformate linearmente,

cioè:

$$\boxed{Y = XA}$$

Dove con A indichiamo una particolare matrice di pesi di dimensione (p x p), e Y la matrice delle componenti principali di dimensione (n x p).

La forma matriciale della matrice y è quindi:

$$y = \begin{pmatrix} x_{11}, x_{21}, \dots, x_{n1} \\ x_{12}, x_{22}, \dots, x_{n2} \\ \dots \\ x_{1p}, x_{2p}, \dots, x_{np} \end{pmatrix} \begin{pmatrix} a_{11}, a_{12}, \dots, a_{1p} \\ a_{21}, a_{22}, \dots, a_{2p} \\ \dots \\ a_{p1}, a_{p2}, \dots, a_{pp} \end{pmatrix}$$

Per semplicità d'esposizione consideriamo la prima componente.

Abbiamo quindi che:

$y_1 = x a_1$ ove: $a_1 = (a_{11}, \dots, a_{1s}, \dots, a_{1p})$, ovvero il vettore a_1 rappresenta la prima colonna della matrice A, e x indica la prima riga di X.

Vogliamo che la prima componente sia, per definizione, la combinazione lineare di massima varianza. Dato che la varianza totale di una trasformazione lineare di x è esprimibile in funzione della matrice di varianze-covarianze S, abbiamo:

$$\boxed{\text{var}(y_1) = a_1' S a_1}$$

Il vettore dei coefficienti dovrà essere in modo tale da massimizzare l'espressione precedente. Le soluzioni di tale problema di massimizzazione sono però infinite, perché la combinazione lineare contiene un fattore di scala arbitrario. Ovvero le soluzioni sono del tipo, scelto un k qualsiasi:

$$(ka)'S(ka)$$

si capisce che k può aumentare all'infinito. Per ovviare a questo, si introduce quindi un vincolo, cioè si impone che la norma del vettore di coefficienti sia uguale ad una costante. Solitamente, si sceglie la condizione di normalizzazione, cioè che tale norma sia uguale ad uno, nel nostro caso: $a_1' a_1 = 1$.

Il problema diventa quindi un problema di massimizzazione vincolata e va risolto con il **metodo dei moltiplicatori di Lagrange**, ovvero richiede l'impiego di un moltiplicatore di Lagrange λ .

Poniamo:

$$f = a'Sa - \lambda(a'a - 1)$$

Massimizzare f equivale a determinare le soluzioni della derivata prima di f rispetto agli "a", ovvero:

$$\frac{\partial f}{\partial a} = 2S - 2\lambda a = 0 \quad \Rightarrow \quad 2(S - \lambda I)a = 0$$

$$\Rightarrow \quad (S - \lambda I)a = 0$$

(1)

L'equazione (1) ha soluzione se e solo se (teorema di ROUCHE-CAPELLI) è singolare, ovvero, se il suo determinante è nullo. Questa equazione è conosciuta come equazione caratteristica. La soluzione di essa comporta la ricerca di p valori di λ , dette radici caratteristiche o semplicemente radici oppure autovalori (in inglese Eigenvalues)

Abbiamo posto:

$$\begin{aligned}\Rightarrow (S - \lambda_1 I)a_1 &= 0 \\ \Rightarrow Sa_1 &= \lambda_1 a_1\end{aligned}$$

Moltiplicando entrambi i membri per a_1 si ottiene

$$a_1' Sa_1 = \lambda_1 a_1' a_1$$

ma sappiamo che $a_1' a_1 = 1$ allora si ottiene:

$$\text{var}(y_1) = a_1' Sa_1$$

$a_1' Sa_1 = \lambda_1$ ma sappiamo ancora che:

$$\text{var}(y_1) = a_1' Sa_1 = \lambda_1$$

allora:

$$\boxed{\text{var}(y_1) = \lambda_1}$$

Abbiamo così mostrato che la varianza della prima componente è l'autovalore associato ad essa, che sappiamo essere il massimo di tutti gli autovalori.

Passiamo adesso a determinare la seconda componente:

$$y_2 = x a_2 ,$$

affinché essa sia ortogonale alla prima componente, il vettore dei coefficienti a_2 deve soddisfare il seguente vincolo:

$$a_1 a_2 = 0, \text{ oltre al vincolo di normalizzazione: } a_2' a_2 = 1.$$

Operando in maniera analoga a prima si ottiene:

$$(S - \lambda_2 I)a_2 = 0$$

per cui si sceglie il secondo autovalore, in ordine decrescente dalla matrice di varianze-covarianze, e lo si indica con λ_2 .

In generale, si dice i -esima componente principale di p variabili, espresse in termini di scostamenti dalla media, la combinazione lineare:

$$y_i = x a_i, \text{ per } i=1. \dots k \leq p$$

In cui a_i è l'autovettore associato all' i -esimo autovalore λ_i , in ordine decrescente, della matrice di varianze-covarianze.

Ogni autovalore λ_i è uguale alla varianza della corrispondente componente.

c. Proprietà della varianza nell'ACP

- Due componenti qualsiasi y_i e y_j sono linearmente indipendenti, ovvero i rispettivi vettori sono ortogonali:

Quindi per ogni $i \neq j$ abbiamo $y_i \perp y_j$, il che equivale alla condizione $a_i' a_j = 0$.

Siccome la matrice S è simmetrica e gli autovettori sono a due a due ortogonali, si deduce che la matrice di varianze-covarianze delle componenti è una matrice diagonale avendo sulla diagonale le rispettive varianze.

- La somma degli autovalori è uguale alla traccia, ossia alla somma degli elementi diagonali della matrice. Se per la determinazione delle componenti principali abbiamo usato la matrice di varianze-covarianze allora la somma

degli autovalori è uguale alla somma delle varianze delle variabili osservate.

$$\sum_i^r \lambda_i = tr(s) = \sum_j^p s_j^2$$

Se invece le componenti sono state ottenute a partire dalla matrice di correlazione, allora:

$$\sum_i^p \lambda_i = tr(R) = p$$

- Il prodotto degli autovalori è uguale al determinante della matrice di partenza.

$$\prod_{i=1}^p \lambda_i = |s|$$

Dal teorema di scomposizione spettrale (Salce 2005), si deduce che:

- È sempre possibile scomporre una matrice di varianze covarianze o una matrice di correlazione, in un numero di componenti principali mai superiore al numero di variabili osservate.
- È sempre possibile ricomporre una matrice di varianze-covarianze, conoscendo autovalori e autovettori.
- Una matrice di varianze-covarianze può essere approssimata con gli autovalori e gli autovettori più "significativi".

- Le componenti non sono indipendenti dalla scala di misura delle variabili: se si moltiplica una variabile osservata per un valore costante, la matrice di varianze-covarianze cambia, e così pure le componenti generate. Esse non variano se le variabili hanno uguali varianze, oppure sono state standardizzate.

d. Standardizzazione delle variabili di partenza

ACP, essendo una procedura esplorativa multivariata viene adoperata su un insieme di dati che ha di solito tante variabili, inoltre queste variabili non sono a volte tra di loro confrontabili. In tale caso, non è possibile lavorare direttamente con la matrice di varianze-covarianze, ma occorre una sua trasformata, ovvero lavorare con la matrice di correlazione.

La **standardizzazione** è un procedimento che riconduce una variabile aleatoria con media μ e varianza σ^2 , ad una variabile aleatoria con distribuzione "standard", ossia di media zero e varianza pari ad uno.

Il procedimento è il seguente a partire dalla matrice di partenza X , e note le medie e le varianze delle popolazioni di appartenenza delle variabili, il processo di standardizzazione permette di normalizzare le variabili:

$$\left(\frac{x_1 - \mu_1}{\sigma_1}, \frac{x_2 - \mu_2}{\sigma_2}, \dots, \frac{x_p - \mu_p}{\sigma_p} \right)$$

Ed ottenere un nuovo campione che ha la particolarità di avere le variabili con media nulla e varianza unitaria. Si mostra che la matrice di varianze-covarianze del nuovo campione così ottenuto coincide con la matrice di correlazione del campione d'origine. Ecco il perché dell'uso diretto della matrice di correlazione. Va notato però che gli autovalori ottenuti con la matrice di correlazione sono diversi da quelli ottenuti con la matrice di varianze-covarianze iniziale.

e. Scelta del numero delle componenti principali

Come abbiamo visto precedentemente la varianza di una componente è uguale al relativo autovalore, e la varianza totale è quindi uguale alla somma di questi ultimi.

Per ottenere la proporzione di variabilità spiegata da ciascuna componente, è sufficiente, quindi, avere il rapporto tra l'autovalore relativo alla componente e la varianza totale:

$$\lambda_i / \sum_{i=1}^P \lambda_i$$

Resta allora da definire il numero di componenti da scegliere, ovvero, quali sono i criteri per la scelta del numero di componenti?

I criteri sono in generale di tre tipi:

- Il primo consiste nel prendere le prime componenti le cui percentuali cumulate spiegano tra l'80% e il 90% della variabilità dei dati.
- Oppure seguire **"la regola di Kaiser"** che ci dice di prendere solo quelle componenti che hanno un autovalore maggiore o uguale a uno (se l'analisi è stata fatta sulla matrice di correlazione), oppure, in modo equivalente, prendere le componenti che hanno varianza maggiore della media degli autovalori (se invece l'analisi è stata condotta su quella di varianze-covarianze).
- Oppure usare il metodo grafico detto "screen-plot", dove si sceglie il numero di componenti in corrispondenza del quale il grafico presenta un "gomito".

f. Interpretazione delle componenti principali.

Dopo aver scelto il numero delle componenti e quindi le componenti stesse, viene la parte più importante dell'analisi ovvero l'interpretazione delle componenti trovate.

Interpretare una componente o fattore significa attribuirgli "un nome", ma non è facile a volte trovare un nome ad una combinazione di variabili. Questo può risultare più agevole se è stata ben fatta prima un'accurata analisi esplorativa dei dati in modo da accertarsi che le variabili siano davvero unite da relazioni così forti da rendere plausibile l'ACP. Quindi oltre a limitarsi all'analisi delle componenti trovate conviene considerare anche la forma del test grafico (screen-plot) sugli autovalori; se il grafico presenta un gomito, allora l'ipotesi dell'esistenza di fattori è plausibile, se invece la forma è quasi rettilinea allora le componenti trovate sono di solito una semplice trasformazione delle variabili di partenza.

Poi, viene la fase vera e propria dell'analisi delle componenti trovate. Questo può essere fatto in diverse maniere elencate di seguito:

- **Saturazione di una componente:** Il coefficiente di correlazione tra un componente e una variabile prende il nome di peso fattoriale o saturazione, questo valore risulta particolarmente interessante perché permette di quantificare l'importanza di ogni variabile nello spiegare la componente. Infatti, tanto grande è in valore assoluto, tanto è importante nello spiegare il fattore stesso. Non esistono però regole universali per fissare il limite dopo il quale una saturazione è rilevante. Si possono accettare alcuni suggerimenti, come quello di Overall e Klett (1972), di prendere come valore minimo in assoluto 0.35, oppure, tenendo conto del fatto che il livello della correlazione è legato alla variabilità della componente, si può scalare la soglia in funzione della radice quadrata degli autovalori, si può partire ad esempio con

il primo fattore da 0.8 e arrivare per l'ultimo con un valore prossimo a quello suggerito da Overall e Klett.

- A volte per interpretare agevolmente le componenti, occorre appoggiarsi ai mezzi grafici che abbiamo a disposizione. Infatti, la rappresentazione grafica dei pesi fattoriali sui piani definiti da coppie di fattori mette in evidenza le variabili che più saturano le componenti.

Questa rappresentazione permette di percepire forme di fattori e posizioni particolari di classi di entità. Tra le forme più comuni si evidenziano: la concentrazione di nuvole di punti in posizioni isolate dello spazio fattoriale e le configurazioni curve, dette "EFFETTO GUTTMAN".

- L'individuazione di un insieme di punti in posizione isolate nello spazio fattoriale sta a indicare che le variabili che fanno parte di una nuvola si riferiscono alla stessa componente, e quindi in questo caso conviene ruotare gli assi finché uno non vi passa o, ovviamente, nelle vicinanze.
- **Effetto Guttman:** è caratterizzato da forme generalmente curvilinee. La più nota è sicuramente quella che presenta una configurazione detta a forma di "ferro di cavallo", con eventuale involuzione (rientro verso l'interno) dei punti terminali. Questo è dovuto generalmente dalla mancata standardizzazione delle variabili, oppure dall'aver dato alle variabili peso uguale quando in realtà hanno peso diverso.

Un fattore curvilineo di solito implica la necessità di estrarre un numero di componenti generalmente superiore a quello effettivamente estratto: la rappresentazione di una componente non lineare richiede l'estrazione di due assi fattoriali.

In alcuni casi la forma a ferro di cavallo è una distorsione della rappresentazione eliminabile cambiando la scala degli assi, se la ragione dell'arcuarsi degli assi sta nelle compressioni dei punti terminali; può anche essere

ineliminabile se il pattern rappresentato è complesso oppure i dati analizzati sono effetti di vari errori di rilevazione.

Nei prossimi capitoli analizzeremo dei casi relativi al comportamento degli utenti che visitano alcuni siti.

CAPITOLO II:

Analisi dei dati relativi al dataset gazelle

1. introduzione

Il sito web "www.Gazelle.com" è un portale di commercio elettronico che vende prodotti di calzetteria e di prodotti per la cura delle gambe (sito in realtà ora chiuso, ma qui presumiamo che sia ancora attivo). L'ufficio Marketing dell'azienda che gestisce il portale è interessato a conoscere meglio i navigatori ed i loro comportamenti per poter predisporre meglio il disegno e l'ordinamento delle diverse pagine di cui si compone il sito, eventualmente adottando strumenti per la personalizzazione delle pagine.

Con lo scopo quindi di identificare i navigatori (o meglio loro comportamenti) che sono clienti buoni, sono stati raccolti alcuni dati per ciascun cliente sulle varie sedute in cui ha acquistato o è entrato nel sito. I dati riguardano (ove presenti) le risposte ad un questionario somministrato ai clienti al primo accesso al sito, e informazioni legate alla navigazione e agli ordini di acquisto effettuati. Le variabili disponibili sono descritte nel seguito. Si osservi che questi dati non sono frutto di un esperimento controllato e possono essere influenzati da fattori esterni (come campagne di marketing ecc.) non controllabili.

I dati sono composti di 1781 unità e 136 variabili, per fare l'analisi abbiamo deciso di dividere la matrice in sottomatrici, una composta dalle variabili qualitative e

un'altra composta da sole variabili quantitativi, della quale riportiamo di seguito il significato delle variabili:

VARIABILI	DESCRIZIONI
V42	Percentuale di Prodotti Acquistati di Giovedì .
V43	Percentuale di Prodotti Acquistati di Sabato .
V44	Percentuale di Prodotti Acquistati di Martedì .
V45	Percentuale di Prodotti Acquistati di Domenica .
V46	Percentuale di Prodotti Acquistati di Mercoledì .
V47	Percentuale di Prodotti Acquistati di Venerdì .
V48	Percentuale di Prodotti Acquistati di Lunedì .
V49	Percentuale di Prodotti Acquistati dalle 18.00 alle 24.00 .
V50	Percentuale di Prodotti Acquistati dalle 3.00 alle 7.00 .
V51	Percentuale di Prodotti Acquistati dalle 0.00 alle 3.00 .
V52	Percentuale di Prodotti Acquistati dalle 15.00 alle 18.00 .
V53	Percentuale di Prodotti Acquistati dalle 11.00 alle 15.00 .
V54	Percentuale di Prodotti Acquistati dalle 7.00 alle 11.00 .
V55	Percentuale acquisti calzetteria taglia 3x-4x .
V56	Percentuale acquisti calzetteria taglia Q-Petite .
V57	Percentuale acquisti calzetteria taglia 1/2 .
V58	Percentuale acquisti calzetteria taglia AB .
V59	Percentuale acquisti calzetteria taglia 5x-6x .
V60	Percentuale acquisti calzetteria taglia 1 .
V61	Percentuale acquisti calzetteria taglia 2 .
V62	Percentuale acquisti calzetteria taglia 3 .
V63	Percentuale acquisti calzetteria taglia 2+ .

V64	Percentuale acquisti calzetteria taglia 4 .
V65	Percentuale acquisti calzetteria taglia 5 .
V66	Percentuale acquisti calzetteria taglia 1P .
V67	Percentuale acquisti calzetteria taglia A .
V68	Percentuale acquisti calzetteria taglia 1X .
V69	Percentuale acquisti calzetteria taglia B .
V70	Percentuale acquisti calzetteria taglia C .
V71	Percentuale acquisti calzetteria taglia D .
V72	Percentuale acquisti calzetteria taglia E .
V73	Percentuale acquisti calzetteria taglia F .
V74	Percentuale acquisti calzetteria taglia I .
V75	Percentuale acquisti calzetteria taglia L .
V76	Percentuale acquisti calzetteria taglia M .
V77	Percentuale acquisti calzetteria taglia P .
V78	Percentuale acquisti calzetteria taglia S .
V79	Percentuale acquisti calzetteria taglia T .
V80	Percentuale acquisti calzetteria taglia S/M .
V81	Percentuale acquisti calzetteria taglia 2P .
V82	Percentuale acquisti calzetteria taglia CD .
V83	Percentuale acquisti calzetteria taglia 2X .
V84	Percentuale acquisti calzetteria taglia L/XL .
V85	Percentuale acquisti calzetteria taglia 3P .
V86	Percentuale acquisti calzetteria taglia C/D .
V87	Percentuale acquisti calzetteria taglia 3X .
V88	Percentuale acquisti calzetteria taglia 3/4 .

V89	Percentuale acquisti calzetteria taglia EF .
V90	Percentuale acquisti calzetteria taglia 6-8 1/2 .
V91	Percentuale acquisti calzetteria taglia 4X .
V92	Percentuale acquisti calzetteria taglia MAXI .
V93	Percentuale acquisti calzetteria taglia 9-11 .
V94	Percentuale acquisti calzetteria taglia P1 .
V95	Percentuale acquisti calzetteria taglia 1x-2x .
V96	Percentuale acquisti calzetteria taglia P2 .
V97	Percentuale acquisti calzetteria taglia P3 .
V98	Percentuale acquisti calzetteria taglia M/L .
V99	Percentuale acquisti calzetteria taglia One Size .
V100	Percentuale acquisti calzetteria taglia XL .
V101	Percentuale acquisti calzetteria taglia XT .
V102	Percentuale acquisti calzetteria taglia 5-6 1/2 .
V103	Percentuale acquisti calzetteria taglia PP .
V104	Percentuale acquisti calzetteria taglia A/B .
V105	Num Hanes Product Views .
V106	Num Orobolu Product Views .
V107	Num AmericanEssentials Product Views .
V108	Num Danskin Product Views .
V109	Num DonnaKaran Product Views .
V110	Num Givenchy Product Views .
V111	Num EvanPicone Product Views .
V112	Num NicoleMiller Product Views .
V113	Num HotSox Product Views .

V114	Num EllenTracy Product Views .
V115	Num Legwear Product Views .
V116	Num LegCare Product Views .
V117	Num Cotton Product Views .
V118	Num Nylon Product Views .
V119	Num Coolmax Product Views .
V120	Num Lycra Product Views .
V121	Num Luxury Product Views .
V122	Num Rayon Product Views .
V123	Num Cotton/Acrylic Product Views .
V124	Num Silk Product Views .
V125	Num Opaque Product Views .
V126	Num Herringbone Pattern Views .
V127	Num Stripe Pattern Views .
V128	Num Plaid Pattern Views .
V129	Num Conversational Pattern Views .
V130	Num Solid Pattern Views .
V131	Num Pique Pattern Views .
V132	Num Pin Dot Pattern Views .
V133	Num Floral Pattern Views .
V134	Num BrandOrder Assortment Views .
V135	Num UniqueBoutiques Assortment Views .
V136	Num Brands Assortment Views .
V137	Num Departments Assortment Views .
V138	Num LifeStyles Assortment Views .

V139	Num SaleAssortments Assortment Views .
V140	Num Welcome Assortment Views .
V141	Num Seasonal Assortment Views .
buon	Ha speso in media più di \$12 per ordine: True,False.

In questa sede ci limiteremo all'analisi della matrice relativa ai dati quantitativi.

2. Analisi descrittive

Un primo approccio all'analisi dei dati consiste nel farne una sintesi in modo da evidenziare le loro principali caratteristiche. Di seguito riportiamo una tabella che riassume la loro sintesi, dove abbiamo indicato con:

- N: il numero totale di unità.
- Media: le medie delle variabili.
- Std.Dev: le devianze standard.

	N	Media	Std.dev		N	Media	Std. Dev
V42	1781	17.18	37.46	V92	1781	0.09	2.62
V43	1781	8.10	27.20	V93	1781	0.03	0.79
V44	1781	18.44	38.54	V94	1781	0.16	3.64
V45	1781	5.09	21.81	V95	1781	0.38	5.41
V46	1781	29.21	45.23	V96	1781	0.18	3.29
V47	1781	14.74	35.22	V97	1781	0.14	2.82
V48	1781	6.96	25.36	V98	1781	0.00	0.18
V49	1781	8.74	28.12	V99	1781	60.98	47.07
V50	1781	20.92	40.45	V100	1781	0.24	4.55
V51	1781	2.87	16.68	V101	1781	0.25	4.67
V52	1781	25.33	43.33	V102	1781	0.00	0.16
V53	1781	21.61	40.92	V103	1781	0.14	3.13
V54	1781	20.25	39.92	V104	1781	0.45	6.69
V55	1781	0.48	6.28	V105	1781	0.61	1.63
V56	1781	0.26	4.90	V106	1781	0.22	1.64
V57	1781	0.11	2.28	V107	1781	1.21	2.45
V58	1781	2.01	12.28	V108	1781	0.48	4.12
V59	1781	0.14	3.55	V109	1781	0.22	1.38
V60	1781	0.82	8.73	V110	1781	0.19	1.23
V61	1780	1.79	12.61	V111	1781	0.33	1.18
V62	1781	3.10	16.55	V112	1781	0.39	1.34
V63	1781	0.07	2.42	V113	1781	0.45	1.27
V64	1781	0.57	6.77	V114	1781	0.64	1.68
V65	1781	0.52	6.94	V115	1781	5.71	10.47
V66	1781	0.20	4.27	V116	1781	0.47	4.01
V67	1781	0.31	4.39	V117	1781	2.11	3.96
V68	1781	0.13	3.47	V118	1781	0.72	4.31
V69	1781	0.86	8.48	V119	1781	0.01	0.18
V70	1781	1.05	9.19	V120	1781	0.09	0.47
V71	1781	0.59	6.85	V121	1781	0.06	0.33
V72	1781	0.23	4.08	V122	1781	0.12	0.52
V73	1781	0.03	1.18	V123	1781	0.03	0.27
V74	1781	1.68	11.96	V124	1781	0.01	0.15
V75	1781	1.77	12.27	V125	1781	0.00	0.05
V76	1781	4.58	19.40	V126	1781	0.10	0.53
V77	1781	0.02	0.79	V127	1781	0.01	0.13
V78	1781	3.45	17.12	V128	1781	0.03	0.29
V79	1781	2.58	15.19	V129	1781	0.88	2.30
V80	1781	0.43	6.33	V130	1781	1.92	5.41
V81	1781	0.32	5.12	V131	1781	0.02	0.21
V82	1781	2.18	13.25	V132	1781	0.00	0.06
V83	1781	0.27	4.85	V133	1781	0.10	0.44
V84	1781	0.24	4.40	V134	1781	1.34	1.68
V85	1781	0.33	5.13	V135	1781	10.03	15.26
V86	1781	0.18	4.14	V136	1781	11.43	22.31
V87	1781	0.15	3.49	V137	1781	7.25	10.55
V88	1781	0.14	3.00	V138	1781	0.40	2.88
V89	1781	0.95	8.80	V139	1781	0.09	0.76
V90	1781	0.27	4.30	V140	1781	0.38	1.21

A questo punto, analizziamo questa tabella in dettaglio in modo da identificare le pagine più visitate. Per fare ciò, ordiniamo le pagine in ordine decrescente rispetto al numero medio divisibile e ci limitiamo alle prime diciassette.

V99	V46	V52	V53	V50	V54	V44	V42	V47	V136
60.98	29.21	25.33	21.61	20.92	20.25	18.44	17.18	14.74	11.43
V135	V49	V43	V137	V48	V115	V45			
10.03	8.74	8.10	7.25	6.96	5.71	5.09			

Analizzando questa tabella, sembra che la pagina v99 sia la pagina più visitata, a seguire vengono le altre confrontando questa tabella con la sua omologa relativa alle percentuali di valori maggiore di zero:

V115	V134	V135	V137	V136	V99	V130	V117	V107	V46	V129	V52
0.97	0.85	0.79	0.75	0.69	0.64	0.56	0.53	0.40	0.30	0.30	0.26
V140	V118	V114	V53	V105	V50	V54	V44	V113	V42	V47	V108
0.25	0.24	0.23	0.22	0.22	0.21	0.21	0.19	0.19	0.18	0.15	0.14

L'analisi di questa tabella mette in luce le pagine più visitate, infatti, vediamo che il 97% degli utenti visitano la pagina v115, a seguire, le pagine v134, v135, v136, v99, e v130 che sono visitate da più del 50% degli utenti, si può allora intuire che queste pagine sono le pagine di più interesse ovvero le pagine le più visitate.

Si può inoltre verificare in corrispondenza a quali pagine si hanno dei massimi:

V136	V115	V135	V118	V130	V137	V108	V116	V42
548	271	206	162	162	162	160	154	100

A questo punto possiamo dire che le pagine V136, V115, V135, V118, V130, V137, V108, V116 sono le pagine più visitate.

Restringiamo la nostra analisi alla sotto-matrice composta dalle variabili relative al numero di articoli visti dagli utenti nei diversi giorni in modo da vedere se esistono delle differenze sostanziali tra i buoni e i cattivi clienti.

3. Differenze tra i buoni e i cattivi clienti

In questa parte, siamo interessati a mettere in evidenza le differenze che esistono tra i buoni e i cattivi clienti, identificati dalla variabile disponibile "buono".

A questo proposito, consideriamo una nuova variabile che misura il numero di tutti i siti visitati.

Effettuiamo prima un'analisi esplorativa dei dati. Siamo allora interessati a sapere se mediamente i buoni clienti visitano più spesso il sito rispetto agli altri.

La sintesi dei dati trasformati relativi ai due gruppi è la seguente:

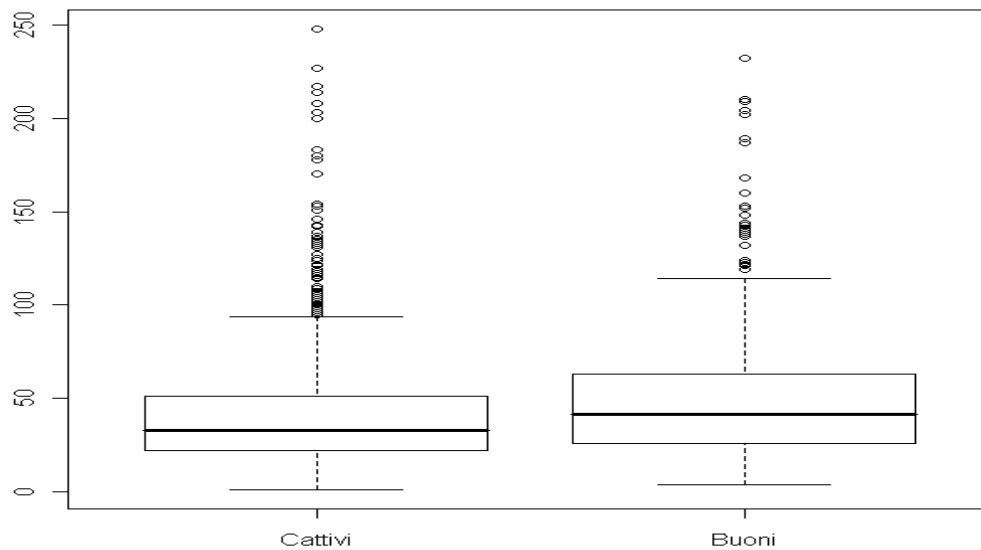
- Buoni clienti:

Min.	primo Qu.	Mediana	Media	3° Qu.	Max.
4	26	40	49.79	62.00	232

- Cattivi clienti:

Min.	1st Qu.	Mediana	Media	3° Qu.	Max.
1	22	34	41.98	52	248

Sembra che i buoni clienti visitino più siti rispetto ai clienti cattivi; infatti, il primo quantile, la mediana e il terzo quantile relativi ai buoni clienti sono leggermente superiori a quelli dei cattivi. Il boxplot, diviso in maniera tale da identificare i buoni e i cattivi clienti, è riportato di seguito:



- **Boxplot relativo ai buoni e cattivi clienti**

Il boxplot, in entrambi i casi, presenta delle asimmetrie a destra, il che significa che i dati non si possono considerare normali, come ci conferma il test di Shapiro (shapiro 1985) che verifica la normalità di un insieme di dati.

Test di normalità dei dati

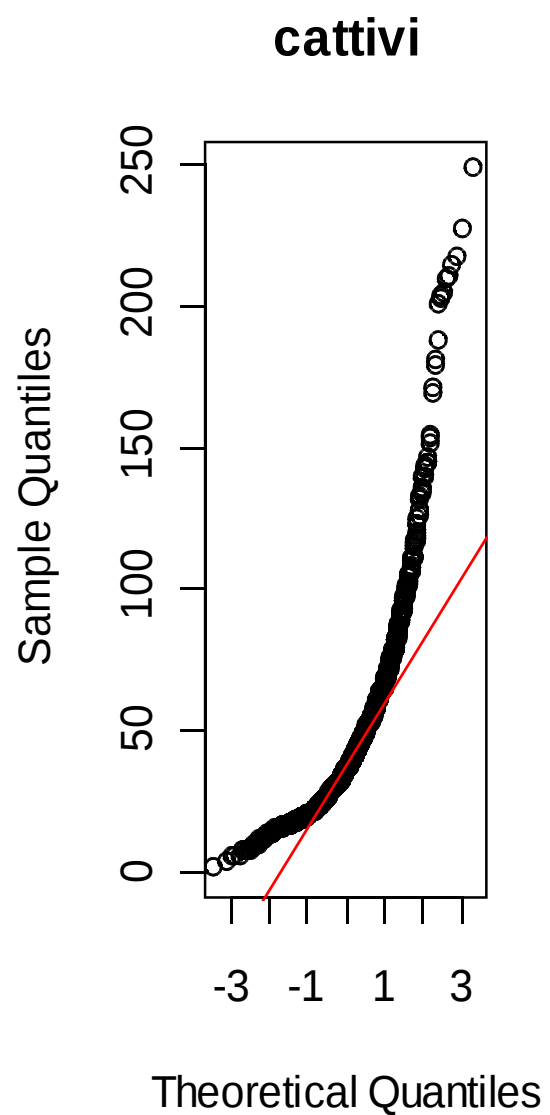
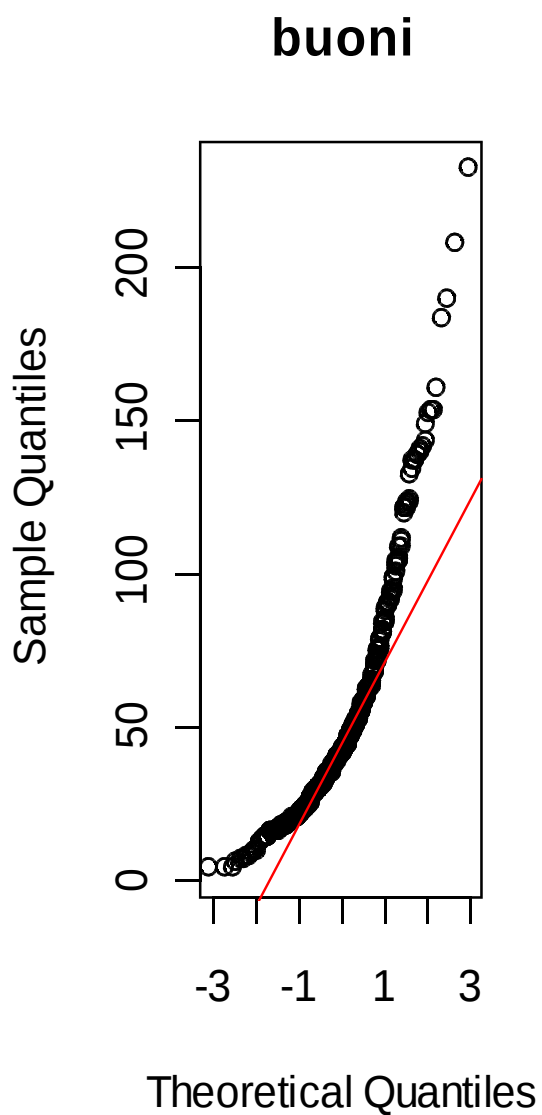
- data: **buoni clienti**

W = 0.8466, p-value < 2.2e-16

- data: **cattivi clienti**

$W = 0.7849$, $p\text{-value} < 2.2e-16$

In entrambi i casi il test rifiuta l'ipotesi di normalità dei dati al livello 5% e questo si vede anche dai grafici di confronto dei quantili delle distribuzioni osservate con quelli teorici della normale.



- grafici di confronto dei quantili delle distribuzioni osservate con quelli teorici della normale.

Va notato però, che abbiamo un'elevata numerosità di dati, il nostro dataset è composto da 1768 unità, numero talmente grande da permetterci di applicare il test senza preoccuparci della loro normalità.

Il risultato del test t-student (Laboratorio di statistica con R, Stefano M. Iacus, Guido M.) per la verifica delle medie nei gruppi è il seguente

t.test a due campioni

data: buoni e cattivi

t = 4.0866, df = 1766, p-value = 4.573e-05

ipotesi alternativa: le due medie sono diverse da zero

intervalli di confidenza al 95%

0.07777575 0.22132360

stime delle medie:

media di x media di y

3.682171 3.532621

Il valore del test è 4.0866, con un valore di significatività osservato di 4.573e-05; rifiutiamo quindi l'ipotesi nulla ovvero che le due medie sono uguali.

In conclusione, possiamo affermare che mediamente esistono delle differenze tra i buoni e i cattivi clienti nel visitare le pagine relative ai diversi articoli; in effetti, una persona prima di acquistare un prodotto tende a visitare più articoli.

Questo è un fatto interessante, perché suggerisce di classificare rigorosamente le pagine in base all'interesse di particolari gruppi di clienti, in modo da attirare le loro

attenzioni, e quindi “costringerli” a spendere anche più del dovuto.

Per agevolare questa classificazione, è opportuno avere un’idea sull’importanza dei prodotti visti dagli utenti. Con questo scopo applichiamo sulle variabili così ottenute l’analisi delle componenti principali.

4. Applicazione dell’ACP sui dati

a. Matrice di varianze-covarianze o matrice di correlazione?

Come abbiamo visto nel primo capitolo, la scelta della matrice è molto importante ai fini dei calcoli e dell’interpretazione dei risultati.

In questo caso, notiamo che i nostri dati sono tutti relativi al numero di prodotti visti da un utente. Possiamo considerare le osservazioni nella stessa unità di misura, quindi, è possibile confrontarle senza procedere ad una loro trasformazione. Optiamo allora per la matrice di varianze-covarianze invece di quella di correlazione.

b. Analisi dei dati relativi ai buoni clienti

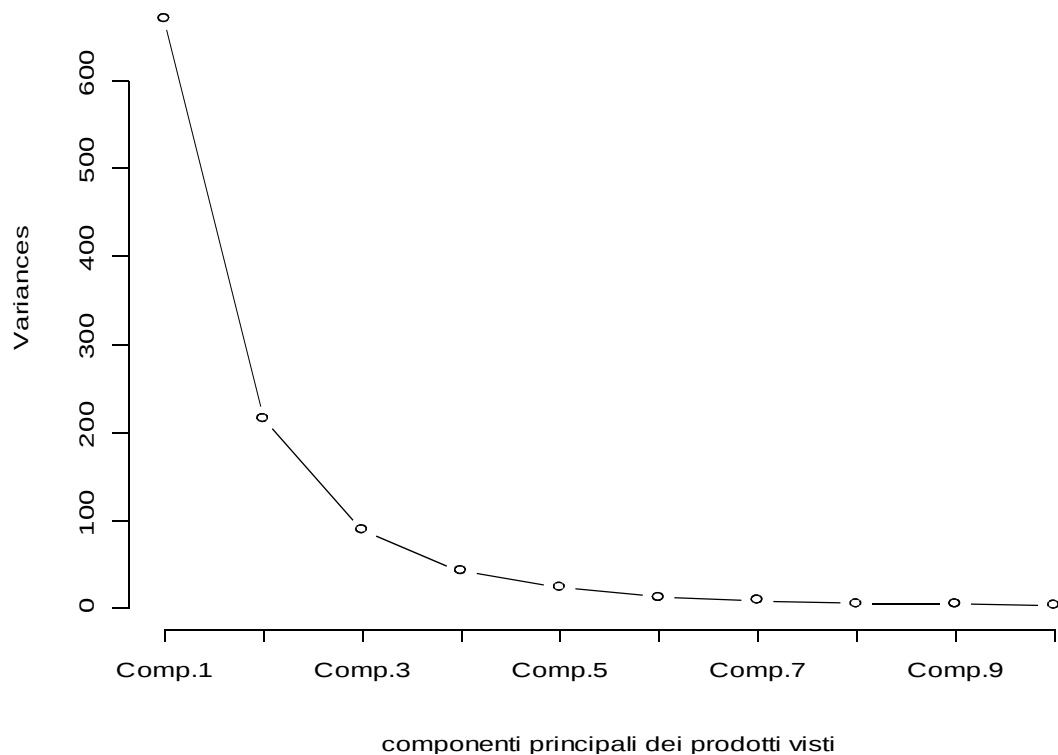
i. Percentuale di varianza spiegata da ogni componente.

La tabella sottostante riporta lo standard error, le percentuali di varianza spiegata da ogni componente e le percentuali cumulate di ogni componente.

	std. error	P. di varianza	P. cumulata
Comp.1	25.91	61.61	61.65
Comp.2	14.69	19.80	81.42
Comp.3	9.40	8.12	89.54
Comp.4	6.50	3.87	93.42
Comp.5	4.84	2.15	95.58
Comp.6	3.55	1.16	96.74
Comp.7	2.96	0.80	97.54
Comp.8	2.29	0.48	98.03
Comp.9	2.13	0.42	98.45
Comp.10	1.74	0.27	98.73
Comp.11	1.46	0.19	98.92
Comp.12	1.44	0.19	99.12
Comp.13	1.24	0.14	99.26
Comp.14	1.18	0.12	99.39
Comp.15	1.15	0.12	99.51
Comp.16	1.00	0.093	99.60
Comp.17	0.96	0.085	99.69
Comp.18	0.77	0.055	99.74
Comp.19	0.72	0.048	99.79
Comp.20	0.69	0.044	99.84
Comp.21	0.60	0.037	99.87
Comp.22	0.58	0.031	99.90
Comp.23	0.49	0.022	99.92
Comp.24	0.44	0.018	99.94
Comp.25	0.36	0.012	99.96
Comp.26	0.33	0.010	99.97
Comp.27	0.26	0.0066	99.97

Comp.28	0.26	0.0063	99.98
Comp.29	0.25	0.0059	99.98
Comp.30	0.17	0.0029	99.99
Comp.31	0.15	0.0023	99.99
Comp.32	0.14	0.0019	99.99
Comp.33	0.125	0.00145	99.99
Comp.34	0.11	0.00124	99.99
Comp.35	0.06	0.00034	99.99
Comp.36	0.04	0.0002	100

Vediamo che i risultati sono soddisfacenti; infatti, si vede dal riassunto che per spiegare più dell'80% della variabilità dei dati abbiamo bisogno solo di tre componenti su un totale di trentasei. I grafici mostrano, infatti, che dopo la terza componente il guadagno in termini di variabilità risulta insignificante.



Vediamo da questi grafici che dopo la terza componente il "guadagno" in termini di variabilità è irrilevante; scegliamo allora tre componenti per poter spiegare e interpretare i dati.

ii. Interpretazione delle componenti trovate.

L'interpretazione delle componenti principali consiste nell'analisi delle combinazioni trovate e quindi dei coefficienti di correlazione o autovalori.

Queste combinazioni sono le seguenti (riportiamo solo le prime tre componenti):

Pesi:

	Comp.1	Comp.2	Comp.3
V108	-0.108		
V115	-0.359		0.160
V116		0.116	
V117		0.100	
V118	-0.114		
V130	-0.163		
V135	-0.255	-0.922	-0.223
V136	-0.827	0.356	-0.270
V137	-0.222		0.898

In questa tabella sono stati omessi gli autovalori in valore assoluto inferiore a 0,1 e le variabili in cui nessuna delle componenti ha valori maggiore di 0,1. Per interpretare agevolmente questi risultati è meglio avvalersi di una nuova

tabella nella quale da una parte mettiamo le variabili con segno + e dall'altra quelle con segno (-).

- Prima componente

Variabili con segno (+)	Variabili con segno (-)
	•V108 Num Danskin Product Views (-0.108) •V115 Num Leg wear Product Views (-0.359) •V118 Num Nylon Product Views (-0.114) •V130 Num Solid Pattern Views (-0.163) •V135 Num Unique Boutiques Assortment Views (-0.255) •V136 Num Brands Assortment Views (-0.827) • V137 Num Departments Assortment Views (-0.222)

Nell'analisi di questa componente, vediamo che tutti gli autovalori hanno segno negativo; questo vuole dire che essi si collocano solo sul quadrante relativo all'asso negativo.

Vediamo inoltre che la variabile che ha il più grande valore (in valore assoluto) è la variabile v136, relativa alle nuove collezioni, poi viene la variabile v115 (che è la categoria delle calze).

Tornando all'analisi effettuata sulla tabella relativa alla sintesi dei dati si nota che queste variabili sono le variabili più importanti, ovvero le pagine che sono più

visitate. Si può allora dire che questa componente mette in luce le pagine più importanti.

• **Seconda componente:**

Variabili con segno (+)	Variabili con segno (-)
• V136 Num Brands Assortment Views (0.356)	• V135 Num Unique Boutiques Assortment Views (-0.922)

Questa componente sceglie solo due delle variabili importanti e le mette a confronto, ovvero oppone i prodotti della "unique boutique" a quelli dei nuovi assortimenti.

Abbiamo quindi da un lato i clienti attenti alla moda, e dall'altra parte quelli che sono interessati ai prodotti della "unique boutique". Andando quindi a indagare a fondo sulle variabili vediamo che i prodotti della "unique boutique" sono i cosiddetti "prodotti chic" mentre i nuovi assortimenti sono i prodotti intermedi.

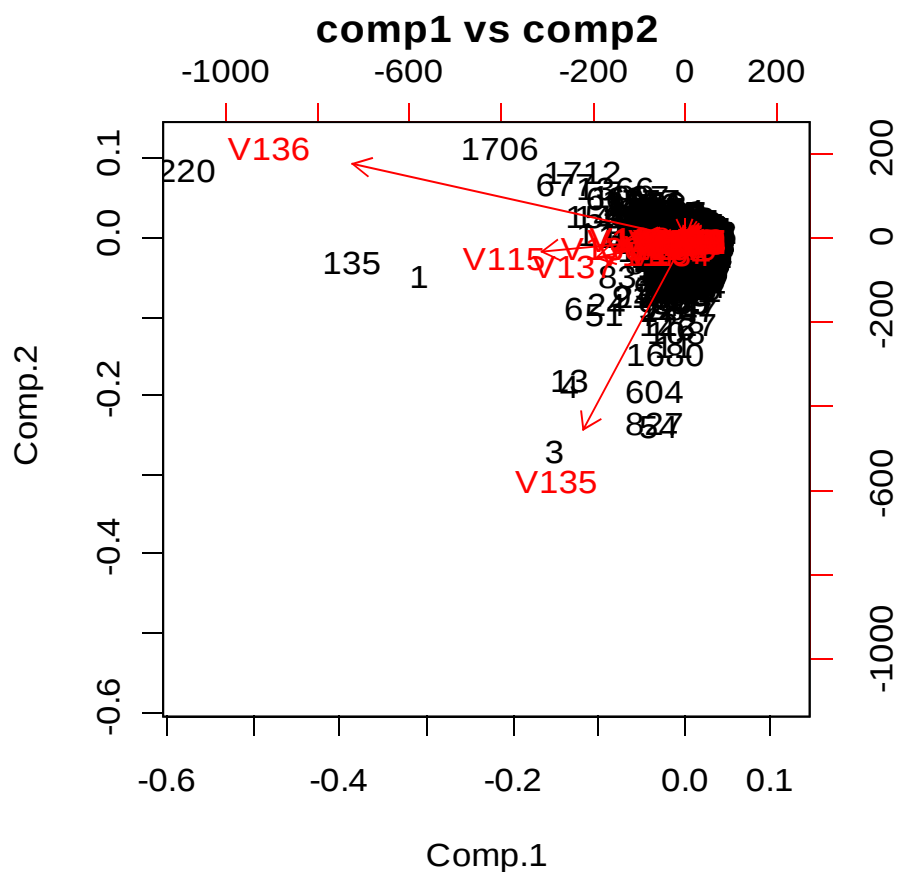
• **Terza componente :**

Variabili con segno +	Variabili con segno -
•V115 Num Leg wear Product Views (0.160) •V116 Num Leg Care Product Views (0.116) •V117 Num Cotton Product Views (0.100) •V137 Num Departments Assortment Views (0.898)	•V135 Num Unique Boutiques Assortment Views (-0.223) •V136 Num Brands Assortment Views (-0.270)

Qui abbiamo da un lato i prodotti relativi alle gambe insieme ai prodotti in cotone e i prodotti del dipartimento, e all'opposto abbiamo i prodotti relativi alla "unique boutique" e i nuovi assortimenti.

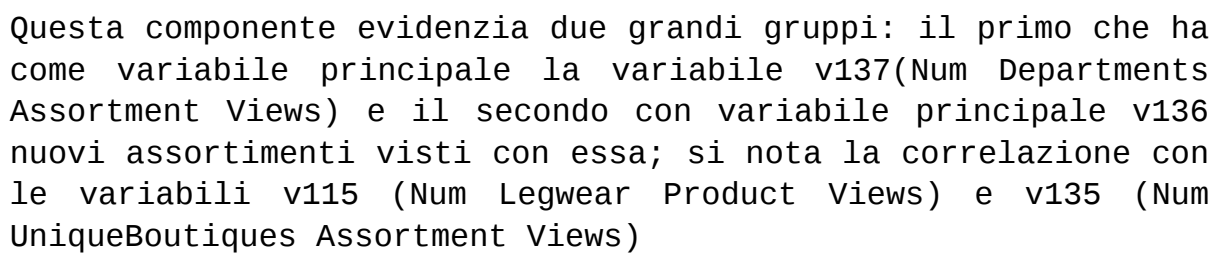
Questa componente divide quindi la popolazione dei clienti in due: abbiamo in primis i clienti a basso profilo interessati solo ai prodotti legati alle gambe senza tener conto della marca e dall'altra parte abbiamo l'insieme dei clienti interessati sia alla marca che ai prodotti trend.

Dopo queste analisi conviene a questo punto andare a vedere le correlazioni che esistono tra le componenti, analizziamo i grafici sottostanti:



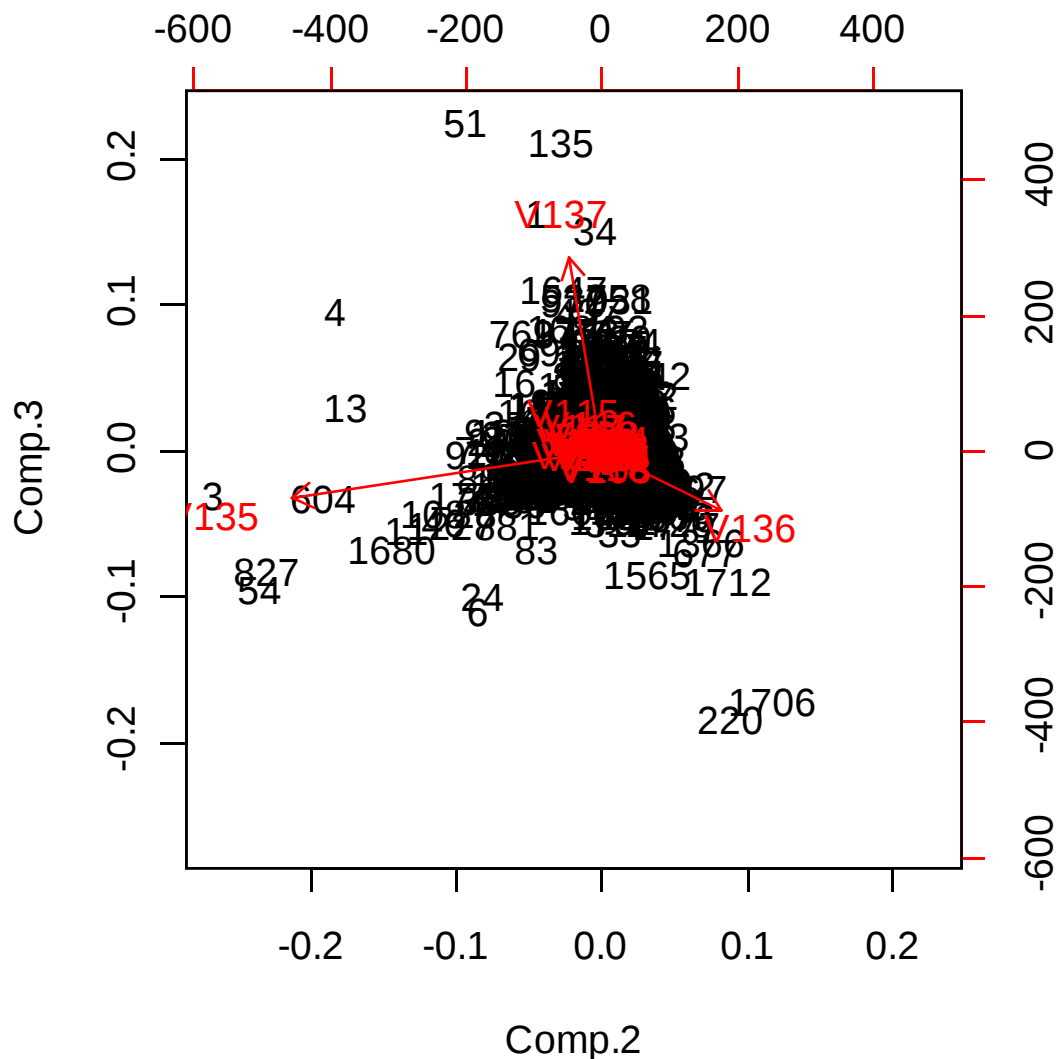
Questo grafico evidenzia le prime due componenti, la prima che ha come variabile principale la v136 (i nuovi assortimenti) e la

Il grafico della prima componente verso la terza è riportato di seguito



Infine, l'ultimo grafico è relativo alle componenti due e tre in cui si evidenziano tre grandi gruppi:

il primo caratterizzato dalla variabile v137, il secondo dalla variabile v135, e infine l'ultimo composto dalla variabile v136



Questo esempio è uno tra i molteplici esempi che mostrano l'utilità dell'uso dell'ACP. Infatti, vediamo che quando abbiamo dei dati di molte variabili non sempre è facile analizzarli e delineare il profilo dell'informazione contenuta in essi. A questo proposito l'ACP si presta bene a ridurre come si è visto la dimensionalità dei dati in modo da rendere agevole la loro interpretazione. Partiti da un insieme di 1781 unità e trentasei variabili, ci siamo ricondotti ad un insieme di tre componenti facilmente interpretabili.

L'analisi del dataset gazelle, ci ha fatto vedere come il comportamento dei clienti di questa società diverge in funzione della loro disponibilità economica e dei loro gusti; abbiamo mostrato che chi è realmente interessato a comprare un articolo generalmente visita più pagine di chi invece non lo è. E tra quelli che comprano possiamo tipicamente distinguere tre classi: quelli non interessati alla marca (che ho chiamato clienti a profilo basso), quelli che sono interessati alla moda e quindi che generalmente visitano gli articoli trend e quindi classificati nel profilo medio e infine quelli che sono classificati nel profilo alto e che visitano tipicamente i prodotti della "unique boutique".

Capitolo III:

Analisi dei comportamenti di visita al sito di Microsoft.

1. Introduzione.

a. Breve presentazione degli obiettivi primari dell'autore

I dati che consideriamo in questo capitolo sono già stati studiati da studiosi come: Cadez et al. (2000), da un punto di vista bayesiano da Giudici e Castello, e Giudici (2005).

Lo scopo di Giudici (2005) consiste nella classificazione dei visitatori in gruppi omogenei, in base ai rispettivi profili comportamentali. Ovvero l'analisi è orientata ad ottenere una regola che assegni ogni singolo utente a un determinato cluster relativo al comportamento di navigazione sul sito; in questo modo si ottiene una segmentazione dell'utenza da poter impiegare nelle successive decisioni di marketing relazionale. Con una tale analisi si è, inoltre, in grado di monitorare l'evolversi del comportamento della clientela, osservando in diversi istanti di tempo la distribuzione dell'utenza nei diversi segmenti comportamentali. In tal modo è anche possibile valutare l'impatto, in termini di aumento o riduzione delle visite, di una qualsiasi azione effettuata sul sito, come ad esempio il restyling delle pagine oppure le azioni pubblicitarie e promozionali.

Nel seguito, faremo prima un'analisi descrittiva dei dati in modo da trarre le prime informazioni, e applicheremo l'analisi delle componenti principali.

b. Presentazione dei dati.

Il dataset contiene dei dati relativi alle pagine del sito: www.microsoft.com; queste pagine sono state visitate da parte di 32711 utenti. Per ciascun visitatore viene indicata la pagina del sito web visitata nella prima settimana di febbraio 1998.

L'individuazione di ogni visitatore avviene tramite assegnazione di un codice univoco che va da 10001 a 42711 e non viene fornita alcuna informazione di carattere personale. Il numero totale delle pagine iniziali ammonta a 196. Queste pagine sono identificate con un numero, un titolo e l'indirizzo corrispondente.

Come si è visto nell'esempio precedente abbiamo già tante unità (32711), e il fatto di avere oltre a queste unità tante variabili fa sì che l'analisi dei dati diventi difficile, e quindi il mezzo più facile per ovviare a questo consiste nel ridurre il numero delle variabili.

Infatti, come fa notare Giudici, è possibile che non tutte le pagine vengano visitate, e quindi considerare le pagine che non sono mai o raramente visitate non fa guadagnare niente in termine di informazione.

E stata fatta un'elaborazione dei dati di partenza in modo da ottenere una matrice dei dati meno complessa. Da un totale di 294 pagine, abbiamo adesso 13 variabili che raggruppano le pagine visitate, di cui diamo una breve descrizione:

- Initial: comprende tutte le pagine di accesso generale e tutte le pagine dedicate alla ricerca nel sito.
- Support: riguarda le pagine relative alla richiesta di supporto.
- Entertainment: comprende le pagine relative ai software di intrattenimento, ai giochi e alla cultura.
- Office: comprende tutte le pagine relative ai software del pacchetto office.
- Windows: raggruppa le pagine relative al sistema operativo "Windows".

- Othersoft: riguarda le pagine relative a tutti i software per professionisti, diversi da quelli del pacchetto office.
- Download: comprende le pagine dedicate allo scaricamento di software e aggiornamenti.
- Otherint: riguarda le pagine dedicate ai servizi via internet per professionisti di information technology, diverse da quelle di download.
- Development comprende tutte le pagine dedicate agli sviluppatori professionisti (per esempio java).
- Hardware: riguarda le pagine relative ai prodotti hardware della Microsoft.
- Business: comprende le pagine dedicate alle aziende.
- Info: comprende tutte le pagine che forniscono informazioni sui nuovi prodotti e servizi.
- Area: comprende le pagine che si riferiscono all'accesso locale, in base alla propria lingua.

Questo nuovo dataset è tale che per ogni pagina indichiamo il numero di volte che è stata visitata da un utente. Si capisce allora che il dataset è composto da variabili quantitative discrete.

Visto che abbiamo provveduto ad una riduzione del numero di variabili, ci si aspetta che le prime tre o quattro componenti principali spieghino solo una minima parte della variabilità dei dati. Nel senso che quando uno effettua un'ACP su dati molto numerosi, generalmente ci si aspetta che la maggiore parte dell'informazione sia concentrata nelle prime componenti (generalmente tre), mentre, eseguendo l'ACP su un numero minore di variabili, ci si aspetta risultati ancora migliori.

A questo punto, eseguiamo una prima analisi esplorativa in modo da evidenziare le prime caratteristiche dei dati.

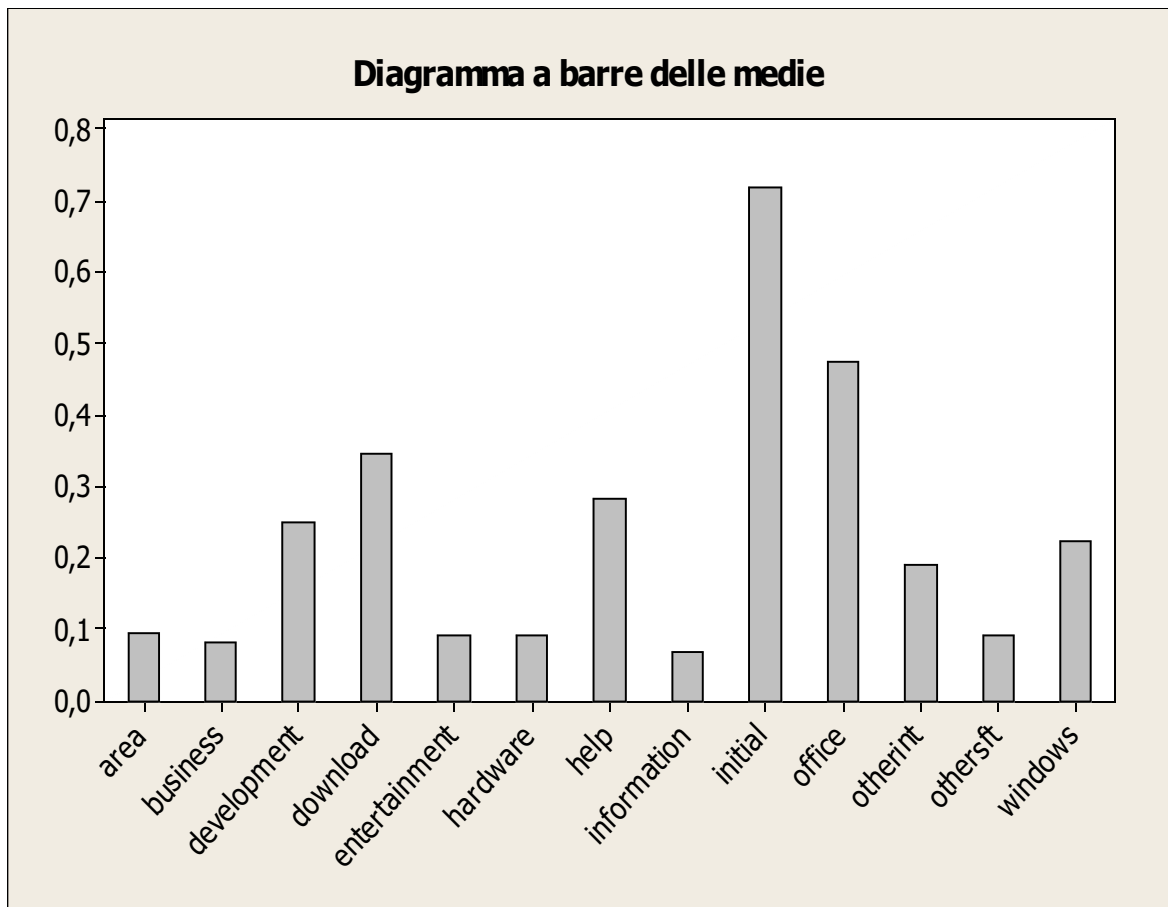
2. Analisi esplorativa dei dati.

Facciamo gli stessi ragionamenti di prima, ovvero in questa prima parte siamo interessati a sapere quali sono le pagine più visitate. La tabella di sintesi dei dati è la seguente:

	N	medie	Dev. Stand.	Massimi	Mediane
initial	32711	0.71	0.88	7	1
help	32711	0.28	0.66	8	0
entertainment	32711	0.09	0.32	4	0
office	32711	0.47	0.70	9	0
windows	32711	0.22	0.53	4	0
othersft	32711	0.09	0.38	7	0
download	32711	0.34	0.48	3	0
otherint	32711	0.19	0.46	6	0
development	32711	0.25	0.67	9	0
hardware	32711	0.09	0.36	6	0
business	32711	0.08	0.30	4	0
information	32711	0.07	0.26	4	0
area	32711	0.09	0.32	5	0

•Tabella 2 :sintesi dei dati.

Notiamo dapprima, analizzando questa tabella, che, oltre alla pagina iniziale, tutte le altre hanno una mediana di zero; questo ci dice già che la pagina iniziale è una pagine cruciale del sito, ovvero che la maggior parte dei visitatori inizia la propria navigazione partendo dalla pagina iniziale. Per poter classificare le pagine in funzione del numero medio delle visite disegniamo un diagramma a barre:

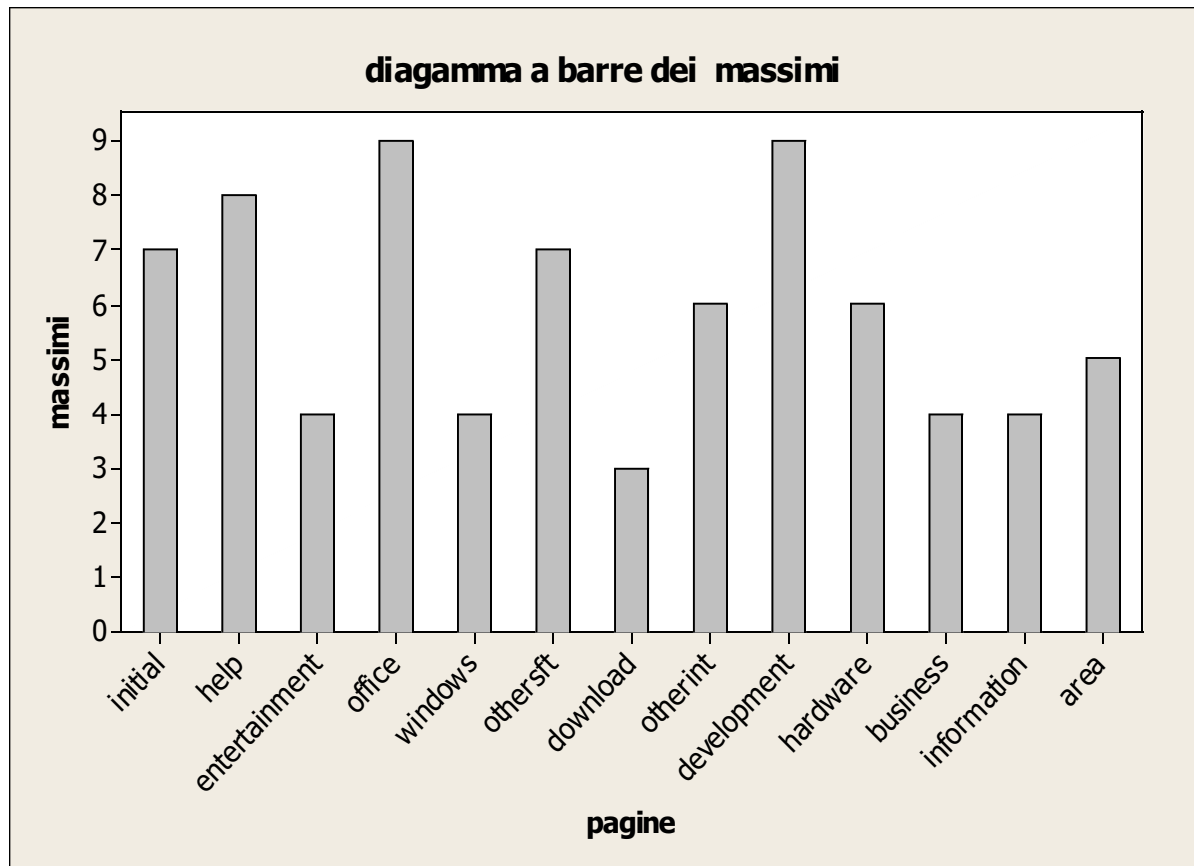


•Grafico 1: diagramma a barre delle medie

come già detto, la pagina più visitata è la pagina iniziale. Dopo di che viene la pagina di office, questo è logico visto che la popolarità della Microsoft risiede oltre al sistema operativo, anche al potente pacchetto "Office". La pagina dei download viene classificata in terza posizione dopo le pagine "iniziale" e "office", e prima delle pagine "help" e "development".

Analizzando la colonna relativa al terzo quartile della tabella 2, osserviamo un fatto molto importante: vediamo che il terzo quartile di tutte le pagine è nullo, ad eccezione delle pagine "initial", "office" e "download", che hanno il terzo quartile pari ad uno. Dato che un terzo quartile nullo vuole dire che il 75% dei navigatori non va a visitare queste pagine, possiamo allora dire che queste tre pagine, sono le

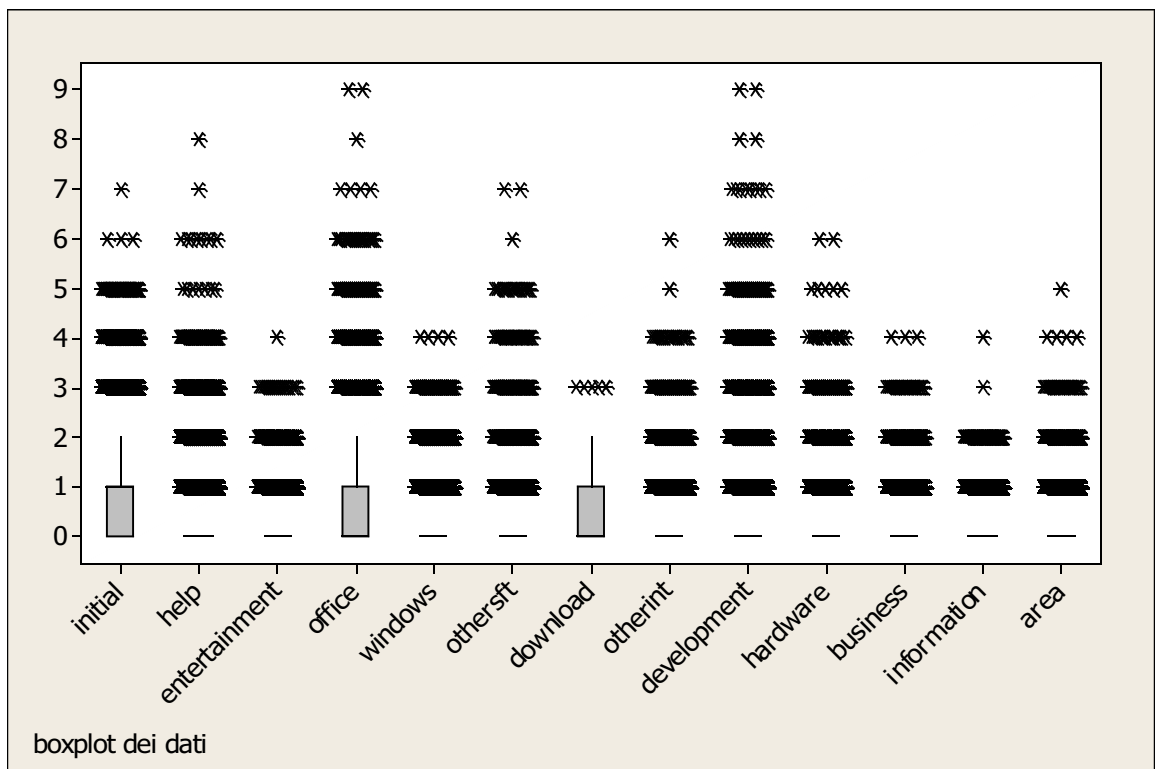
pagine con una frequenza di visita alta, ovvero il 75% dei visitatori guarda almeno una volta queste pagine. Guardando anche il grafico del numero massimo di visite delle pagine, vediamo che: in corrispondenza alle tre pagine appena citate abbiamo anche dei massimi, a cui aggiungiamo la pagina development. Questo conferma quanto appena detto



•Grafico 2 diagramma a barre dei massimi

Questo si vede anche dai boxplot sottostanti dove, oltre alla diversità delle variabilità dei gruppi, osserviamo che a parte le tre pagine appena citate grande parte della popolazione dei clienti non visita queste pagine.

Questo sta a dire che possiamo distinguere già tre pagine importanti ovvero la pagina iniziale, quella del pacchetto office e quella dei download come pagine che più attirano l'attenzione dei visitatori.



a.Grafico 3: boxplot dei dati.

3. Applicazione dell'analisi delle componenti principali sui dati.

Anche in questo caso conviene lavorare con la matrice di covarianza invece che con quella di correlazione o un'altra matrice standardizzata. Infatti, osserviamo che tutte le variabili, anche se hanno variabilità diversa come si vede dal boxplot precedente, sono tutte relative al numero di pagine visitate dagli utenti e quindi nella stessa unità.

L'applicazione della procedura dell'ACP sui dati ha prodotto i dati riassunti nella tabella seguente.

	std. error	P. di varianza	P. cumulate
Comp.1	0.97	26.59	26.59
Comp.2	0.75	16.08	42.67
Comp.3	0.70	13.79	56.47
Comp.4	0.55	8.60	65.07
Comp.5	0.53	7.98	73.06
Comp.6	0.43	5.29	78.35
Comp.7	0.39	4.47	82.83
Comp.8	0.36	3.64	86.47
Comp.9	0.34	3.32	89.79
Comp.10	0.32	2.92	92.72
Comp.11	0.32	2.88	95.61
Comp.12	0.30	2.69	98.306
Comp.13	0.24	1.69	100

I. Tabella 3: riassunto dell'ACP

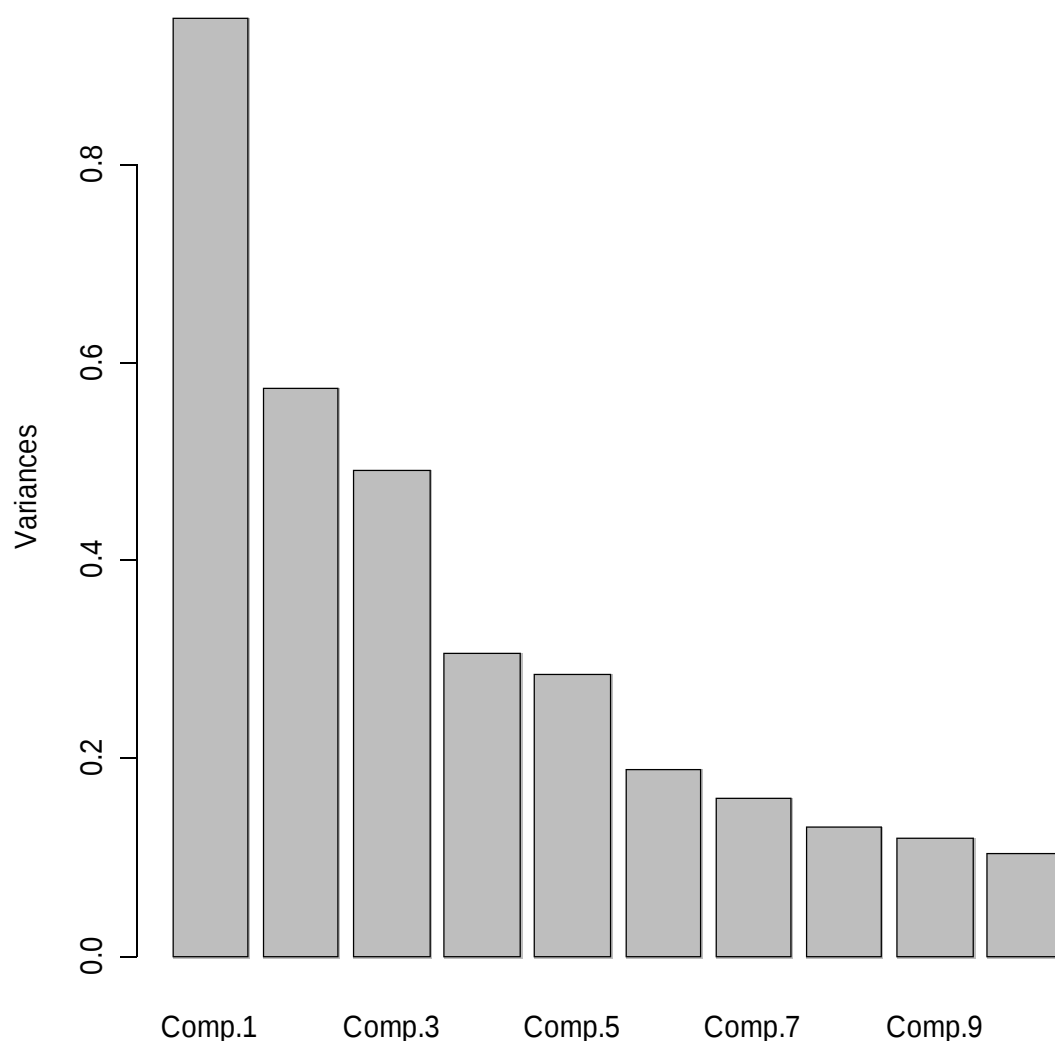


grafico delle varianze delle componenti

Come ci si aspettava, per spiegare l'82% della variabilità dei dati abbiamo bisogno di un totale di 7 variabili su 13, mentre sarebbe stato preferibile che la maggior parte della variabilità fosse concentrata nelle prime tre componenti e che lo screen-plot presentasse un gomito evidente. Questo ci porterebbe quindi a dire erroneamente, che l'ACP non è stata ben eseguita, e quindi occorrerebbe cercare magari una versione standardizzata della matrice in modo da migliorare i risultati.

Però bisogna anche tener presente che siamo partiti da un dataset avente un totale di 194 variabili e che ci siamo ricondotti a 13, aggregando variabili.

Questo è uno dei motivi per cui l'analisi dell'ACP in questo caso non ha prodotto risultati tanto soddisfacenti quanto quelli dell'esempio precedente, perché tende a considerare tutte le variabili come importanti e come abbiamo fatto notare sono importanti perché elaborazioni o sintesi di altre variabili.

Proviamo quindi ad interpretare i risultati prodotti da questa analisi. Per l'interpretazione ci imiteremo a cinque componenti, perché con cinque componenti riusciamo a spiegare più del 70% della variabilità dei dati.

Pesi:					
	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
initial	0.859			0.467	-0.141
help	0.455	0.181		-0.696	0.520
entertainment					
office		-0.698	-0.623		0.183
windows	0.162			-0.457	-0.749
othersft		-0.140	0.165		
download		-0.223	-0.208	-0.286	-0.317
otherint		-0.241	0.293		
development		-0.586	0.669		
hardware					
business					
information					
area					

Anche in questo contesto, per poter spiegare il senso delle variabili trovate, aiutiamoci con le tabelle.

- Componente 1:

Variabili con segni positivi	Variabili con segni negativi
I.initial 0.859	
II.help 0.455	
III.windows 0.162	

Osserviamo che la prima componente ha valori alti solo in corrispondenza di tre variabili. E queste variabili sono a coefficiente positivo. Vediamo che questa componente è relativa agli utenti standard, ovvero gli utenti alle prime armi con l'informatica e il mondo virtuale, come si vede dalla variabile più importante "initial"; questi utenti usufruiscono tanto della pagina iniziale, ma hanno anche bisogno della pagina di aiuto ("help" seconda pagina importante). Notiamo anche che questi utenti sono anche interessati alle pagine relative a Windows.

-Componente 2:

Variabili con segni positivi	Variabili con segni negativi
I.help 0.181	II.office -0.698
	III.othersft -0.140
	IV.download -0.223
	V.otherint -0.241
	VI.development -0.586

Analizzando le variabili più importanti di questa componente (office , development), vediamo che questa componente mette da una parte i professionisti e i semi professionisti e all'opposto la pagina di aiuto; questo sta a dire che è vero che i professionisti e semi professionisti visitano particolari pagine a loro dedicate, ma hanno anche bisogno della pagina di aiuto. Questo fatto risulta molto importante, perchè sta a indicare che la pagina di aiuto è una pagina visitata dalla quasi totalità della popolazione, quindi una sua buona organizzazione porterebbe ad un miglioramento dei servizi della società.

-Componente 3:

Variabili con segni positivi	Variabili con segni negativi
. Othersft 0.165	. office -0.623
. otherint 0.293	. download -0.208
. development 0.669	

L'analisi di questa componente, invece, oppone i professionisti ai semi professionisti. Infatti, abbiamo sull'asse positivo tutte le pagine relative ai software di sviluppo per professionisti e dall'altra, i prodotti office e la pagina di download, tipiche pagine visitate, di solito, solo da chi è interessato al prodotto office della Microsoft.

-Componente 4:

Variabili con segni positivi	Variabili con segni negativi
. initial 0.467	. help -0.696
	. windows -0.457
	. download -0.286

Questa componente, invece, ci oppone la pagina iniziale a quelle di help, windows e download, ovvero sembra dividere i principianti in due, cioè sembra fare la differenza tra quelli che usano solo la pagina iniziale e quelli con un livello leggermente più alto che visitano le pagine di aiuto, windows download.

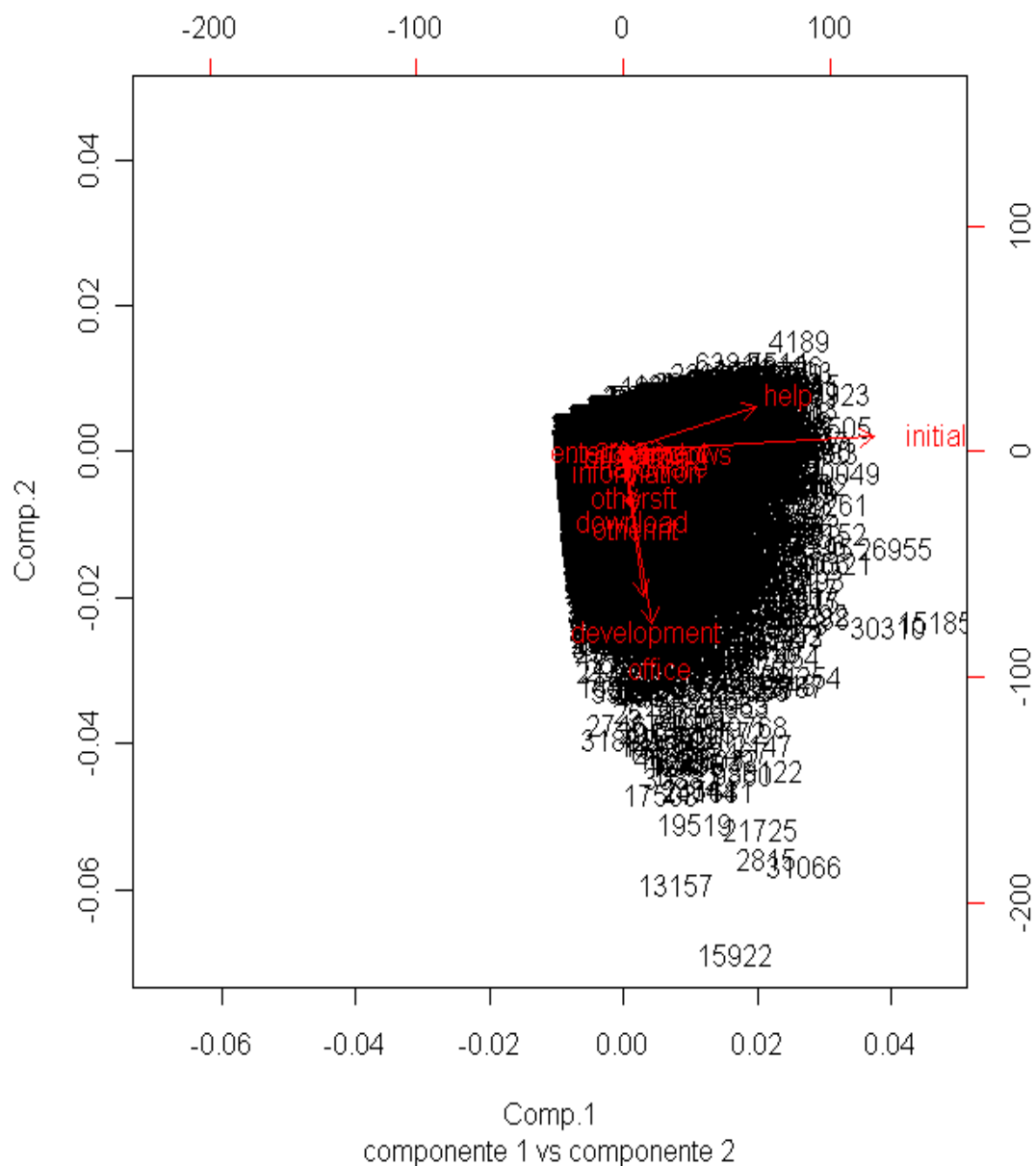
-Componente 5 :

Variabili con segni positivi	Variabili con segni negativi
. help 0.520	. initial -0.141
. office 0.183	. windows -0.749
	. download -0.317

Quest'ultima componente confronta i semi professionisti con i non professionisti.

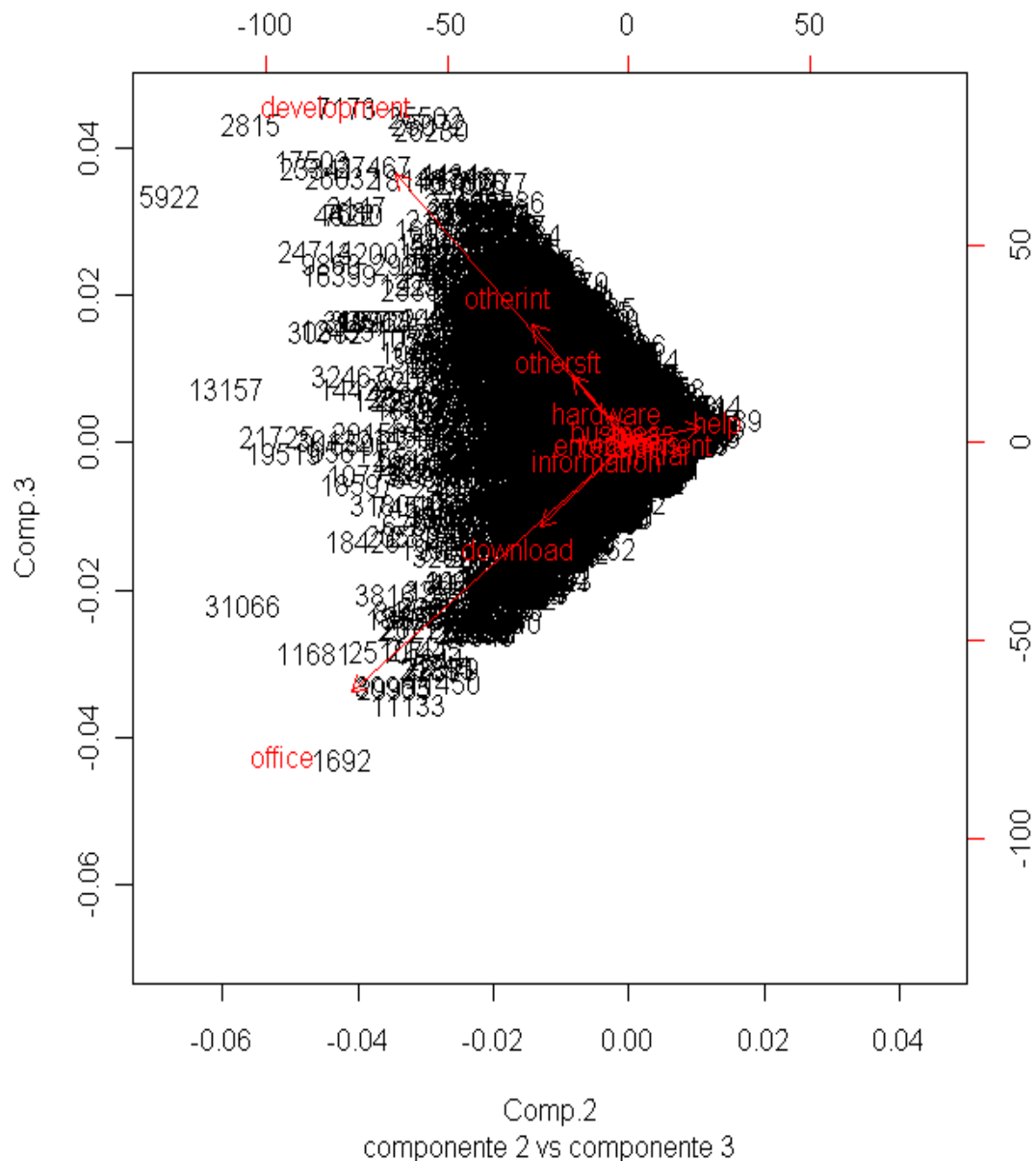
L'analisi delle correlazioni tra le variabili si può fare attraverso l'analisi dei grafici sottostanti chiamati biplot.

Il biplot delle due prime componenti è il seguente:



Da questo biplot si evidenziamo due grandi gruppi: il primo composto dalle unità per cui variabili initial e help assumono valori elevati, mentre il secondo composto da quelle con valori alti per office e development. Vediamo che queste due pagine sono correlate con le variabili otherdft,download, otherint, information e hardware. Questo significa che, generalmente, chi visita le pagine di office e development generalmente visita anche le pagine ad essa correlate.

Il biplot della seconda componente contro la terza è il seguente:



Qui si evidenziano tre grandi gruppi: il primo composto dalle pagine relative ai professionist development, otherint, othersoft; il secondo ai semi professionisti office download, e infine la pagina di aiuto.

L'analisi di questo dataset ci ha permesso, una volta di più, di mostrare come i comportamenti dei visitatori di un sito web differiscano tra di loro. In particolare la visita delle pagine del sito della Microsoft da parte dei clienti si fa in modo ben chiaro, ovvero si sono evidenziati tre grandi gruppi che abbiamo deciso di chiamare principiante, ovvero quelli che iniziano la loro navigazione dalla pagina principale e utilizzano tanto la pagina di aiuto, poi il gruppo dei semi professionisti, cioè quelli interessati ai prodotti del pacchetto office, e alla pagina di download, e infine i professionisti, quelli che visitano specifiche pagine legati allo sviluppo di software.

Inoltre si è anche visto come in questo caso l'analisi, pur non avendo prodotto risultati straordinari, ha permesso di dividere la popolazione in sotto-popolazioni, in base al loro comportamento o meglio conoscenza nel riguardo dei servizi della Microsoft.

Conclusione

Abbiamo visto come attraverso specifici strumenti statistici è possibile trarre utili informazioni da un insieme di dati anche complessi.

Infatti, l'analisi dei dati relativi ai due siti considerati in questa sede, ci ha permesso di mostrare come il comportamento di visita degli utenti a questi siti si fa in maniera non casuale, ovvero abbiamo visto, ad esempio, che possiamo classificare i visitatori del sito della Microsoft in funzione della loro conoscenza e del loro interesse alle nuove tecnologie. Oppure che, gli utenti che visitano il sito della Gazelle, possono anch'essi essere divisi in classi in funzione del loro "portafoglio" e del fatto che siano interessati o meno alla moda.

Queste conoscenze ci porta quindi ad usare specifiche tecniche di marketing per poter ottimizzare i profitti delle vendite. E questa ottimizzazione inizia da un ordinamento rigoroso fatto sulla base dei risultati ottenuti dalle analisi.

Bibliografia

- Azzalini A., Scarpa B. (2004), Analisi dei dati e data mining: Springer.
- Dunleman G. H. (1989), Principals components analysis : Newbury park[etc].
- Fabbris L. (1983), Analisi dei dati multidimensionali: Cleup.
- Fabbris L. (1997), Analisi multivariata, analisi esplorativa dei dati: McGraw-Hill
- Marchetti Giovanni (1984), Analisi in componenti principali e approssimazioni di matrici: alcune interpretazioni e legami con altri metodi multivariati: Pubblicazione del dipartimento statistico dell'università di Firenze
- Giudici P (2005), Data mining: Metodi informatici, statistici e applicazioni, seconda edizione: McGraw-Hill.
- Iacus S. M., Masarotto G. (2007), Laboratorio di statistica con R: McGraw-Hill.
- Shapiro S. S. (1985), Come provare la normalità ed altre ipotesi per le distribuzioni di frequenza: Editoriale Itaca

