Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

L'impatto della pandemia di Covid-19 sul business di Airbnb: un'analisi per la città di Venezia.

Relatrice Prof. Mariangela Guidolin
Dipartimento di Scienze Statistiche

Laureanda: Francesca Bosin
Matricola N. 1234265

Anno Accademico 2021/2022

Indice

Capitolo 1

Il fenomeno di Airbnb: caratteristiche salienti

1.1 Introduzione

Questo elaborato nasce con l'obiettivo di approfondire alcuni aspetti relativi all'impatto sul mercato dei viaggi della pandemia di Covid-19, che sta colpendo il contesto globale dalla fine del 2019. Si è voluto focalizzare l'attenzione sul commercio online e sulla *sharing economy* e, in particolare, sul business di "Airbnb". Si è deciso di analizzare il contesto italiano e, nello specifico, la città di Venezia, concentrandosi sui prezzi degli alloggi. Più precisamente, con il seguente lavoro di tesi, si desidera analizzare l'andamento dei prezzi degli annunci di "Airbnb" in tre anni consecutivi, per indagare quali siano i fattori che influenzano eventuali variazioni avvenute in tempi di pandemia. La città di Venezia è stata scelta sia per caratterizzare il fenomeno considerato nel territorio locale, sia perché è da considerarsi una città particolare per il turismo in Italia. Si ritiene, infatti, che coloro che ricercano un alloggio in affitto a Venezia provengano da diversi contesti globali e appartengano a diverse categorie socio-economiche, con esigenze e motivazioni differenti per giustificare la loro permanenza in questa città.
A tale scopo, sono stati utilizzati i dati disponibili sul sito "insideairbnb.com" per la città di Venezia, aggiornati nei mesi di giugno di tre anni consecutivi,

ovvero 2019, 2020 e 2021. Si tratta di un sito non ufficiale in cui sono raccolti e aggiornati periodicamente i dati del sito e dell'applicazione di "Airbnb".

La *sharing economy* è un fenomeno in crescita, la cui analisi a partire dai dati raccolti dal sito di "Airbnb" è di grande interesse in tempi recenti. La letteratura riporta casi di studio riguardanti numerose realtà globali, in cui si analizzano le molteplici sfaccettature dell'argomento. In particolare, risulta essere di interesse l'analisi del successo di questa forma di commercio di recente adozione nel mercato degli hotel, in cui ha avuto un forte impatto, poiché era completamente diversa da quelle esistenti. Il fenomeno "Airbnb" è frequentemente analizzato rispetto al numero di utilizzatori che usufruiscono di questo servizio, rispetto ai sentimenti nutriti nei confronti di esso da parte degli utenti e ci si focalizza spesso sulla reazione del tradizionale mercato degli hotel all'innovazione che "Airbnb" ha costituito.
Il modello economico adottato è quello di un'economia condivisa (*sharing economy*), o economia collaborativa, nota anche come economia *peer-to-peer* (*P2P*). Si tratta di un modello in cui i singoli individui condividono beni mobili o immobili, come una casa, e le transazioni economiche sono facilitate da Internet. Esso permette, tra le altre cose, ai clienti di avere dei vantaggi in termini di prezzo e di accesso al servizio e di utilizzare la modalità "fai-da-te" per l'acquisto di un bene. "Airbnb" mette a disposizione un'applicazione che fornitori e clienti utilizzano per acquistare o vendere beni o servizi.
Il fenomeno dell'economia *peer-to-peer* nel mercato dei viaggi, al giorno d'oggi ha un forte impatto nella società e presenta diverse sfaccettature. Tuttavia, quando si parla di economia *P2P* si intende spesso un modello teorico o una strategia di marketing, e non qualcosa di effettivamente attuabile, come testimonia, nel nostro caso, la notevole quantità di "host" che in realtà sono agenzie, che gestiscono gli affitti degli immobili.
Da uno studio di Lee & Kim (2019) emerge che il successo dell'economia *peer-to-peer* è un indicatore della qualità dei siti che forniscono lo spazio per questo genere di rapporto tra gli affittuari e i loro clienti. Tali siti dimostrano credibilità e ricevono l'apprezzamento e la fiducia degli utenti nella scelta di adoperarli come una migliore alternativa rispetto ai tradizionali metodi di prenotazione di un alloggio.
Nell'articolo di Cesarani & Nechita (2017), inoltre, si espongono alcune ca-

ratteristiche della *sharing economy*, in quanto si tratta di un servizio condiviso tra "hosts" e "guests", guidato dal progresso tecnologico e dall'ampia accessibilità di una innovazione. Si evidenziano gli aspetti di *trust*, *togetherness*, *technology* e *trasformation* ed elementi che hanno favorito lo sviluppo di questo commercio, quali la crisi economica, i pagamenti digitali, l'utilizzo diffuso degli *smartphone* e la cultura odierna. Qusto articolo analizza, inoltre, la differente numerosità degli utilizzatori di "Airbnb" nelle diverse città e località italiane. Si osserva, poi, una caratteristica della società, che si fida dello scambio di un bene tra persone e lo considera una modalità vantaggiosa per effettuare acquisti.
L'articolo di Contu, Conversano, Frigau & Mola (2019) è un altro degli esempi di trattazione del fenomeno "Airbnb" nel contesto italiano. Vengono identificate le differenze con il turismo tradizionale secondo i criteri di sostenibilità sociale, economica e ambientale. Si osservano, inoltre, le diverse numerosità di turisti tradizionali e innovatori in molteplici città utilizzando indicatori percentuali sul totale della popolazione.
In Cheng & Jin (2019), invece, vengono messe in evidenza le motivazioni per cui l'innovazione apportata da "Airbnb" costituisce un vantaggio per i clienti. Attraverso un'analisi di *text-mining*, emergono il prezzo e la vicinanza ai servizi come caratteristiche salienti per le scelte dei consumatori.
L'argomento viene trattato anche nell'articolo Oskam & Boswijk (2016). Dopo aver messo in evidenza le caratteristiche della nuova forma di commercio proposta, gli autori pongono l'accento sulle particolarità di tale innovazione. Infatti, essa costituisce un vantaggio per il turismo, in quanto le città sono maggiormente frequentate. Risulta, tuttavia, sconveniente per gli hotel tradizionali. Dal momento che si prevede una crescita di "Airbnb", essi dovranno ricorrere ad un'offerta sempre più personalizzata e meno standardizzata e verso un cambiamento dei canali di distribuzione. Restano, secondo gli autori, molte questioni in via di definizione, tra cui le modalità di tassazione del nuovo business, la misurazione dell'andamento del turismo, la conformità delle regolamentazioni della piattaforma con le leggi delle singole località, gli aspetti di sicurezza, protezione delle transazioni di denaro, *pricing*, competizione con gli hotel. Si osserva, quindi, un cambiamento nella domanda di mercato e nella priorità attribuita alla convenienza e alla comodità degli

acquisti, oltre che alla ricerca di un commercio etico, che operi in un contesto di assenza di discriminazione sociale.
Un altro esempio in cui è analizzato il cambiamento nel mercato è Tussyadiah & Park (2018), nel quale, attraverso un'analisi di *text-mining* in quattordici città degli Stati Uniti, viene trattata la fiducia dei consumatori verso gli "hosts" e si sottolinea l'importanza dei forum di domande online tra gli utilizzatori di un servizio.
Si segnala, a partire dalla letteratura disponibile, infine, che nelle motivazioni con cui i clienti si affidano alla piattaforma non è ben delineato il confine tra la fiducia al contenuto degli annunci e la convenienza dell'offerta e che gli acquirenti sono per la maggior parte di età inferiore ai quarant'anni. Questo dato dimostra un'adesione non completa della società all'*e-commerce*.
Nell'articolo di Bernardi & Guidolin (2022), sono trattati gli elementi determinanti i prezzi degli alloggi di "Airbnb", con specifico riferimento alla città di New York. Ne emerge che tali prezzi sono molto eterogenei e cambiano al variare delle caratteristiche degli alloggi e della zona in cui essi sono situati. In tale analisi sono stati considerati i diversi quartieri della città e applicati metodi di regressione prima di tipo lineare, poi secondo un approccio semiparametrico basato sui quantili. In questo modo, si ricerca un collegamento tra il prezzo degli alloggi e i vantaggi che caratterizzano ogni annuncio. In particolare, per quanto riguarda le caratteristiche del quartiere, sono analizzate proprietà quali la vicinanza ad attrazioni e servizi, la vicinanza alle strutture alberghiere (che sembrerebbe essere un fattore di aumento del prezzo, contrariamente a quanto si suppone, data la competizione tra i due tipi di struttura), la distanza da punti di ristoro e la criminalità. Giocano, infine, un ruolo importante nella determinazione del prezzo le recensioni degli altri utenti, le caratteristiche degli alloggi, il numero di ospiti consentiti e la flessibilità delle cancellazioni. Questo elaborato si propone uno scopo simile all'articolo citato, ovvero quello di studiare il prezzo degli alloggi in una specifica città, Venezia.
In Gambatesa (2020), invece, particolare spazio è dato all'analisi delle recensioni, volta a capire l'impatto che l'economia *peer-to-peer* ha sulla società. I risultati ottenuti riguardano la realtà italiane di Milano, Roma e Napoli, e si utilizzano metodi di regressione per stimare il numero di recensioni del sito,

effettuando, poi, una analisi dei testi di queste ultime.

La tesi si articola nei seguenti capitoli:

- Si presentano, di seguito, alcuni casi di studio di interesse per comprendere meglio le sfaccettature del fenomeno "Airbnb" e una descrizione delle principali caratteristiche del servizio offerto e dei cambiamenti avvenuti in tempo di pandemia.

- Nel secondo capitolo sono presentati i dati a disposizione e sono riportate le operazioni di preparazione dei dataset. Inoltre, si esplicitano le variabili che compongono i dataset finali, utilizzate per le analisi. Vengono esposte brevemente le caratteristiche delle loro distribuzioni, riportando alcuni risultati dell'analisi descrittiva dei dati.

- Nel terzo capitolo si riportano i risultati dell'applicazione di modelli di regressione al dataset. Ci si pone l'obiettivo di studiare le variabili più rilevanti per la descrizione dell'andamento del prezzo e ci si focalizza sulle differenze riscontrate nelle tre annualità considerate. Si presenta un approfondimento dell'applicazione dei modelli ritenuti più utili per le analisi di interesse.

- Nel quarto capitolo si riportano i risultati dell'applicazione di modelli lineari gerarchici ai dataset e il confronto di questi con modelli di regressione lineare. L'obiettivo è quello di studiare le variazioni nella stima del prezzo in base alla zona geografica in cui si collocano le abitazioni, servendosi di alcuni predittori scelti tra le informazioni disponibili.

- Nel quinto capitolo si riportano alcune osservazioni di carattere conclusivo sull'intero elaborato, sui risultati ottenuti, sulle conclusioni tratte, sui limiti e sulle potenzialità del lavoro svolto.

- Alcune analisi descrittive ulteriori e una parte del codice utilizzato sono riportati in tre appendici finali.

1.1.1 Il caso dei *multihosts* a Venezia

Sul sito "insideairbnb.com", da cui provengono i dati per le analisi, sono disponibili degli studi effettuati riguardanti il coinvolgimento di "Airbnb" in differenti tematiche, relativamente a diverse città.
In particolare, per quanto riguarda il contesto italiano, è presente un articolo di Corona (2019). Esso riporta un resoconto degli affitti mensili di "Airbnb" a Venezia, aggiornato al 12 agosto 2019, soffermandosi in particolare sulle irregolarità della effettiva tipologia di commercio attuata da alcuni "hosts" presenti sulla piattaforma. Tale tematica è molto discussa.
L'articolo tratta di come il *business* di "Airbnb sia monopolizzato da "host" che non sono privati cittadini, che condividono la loro casa con gli ospiti, ma appartengono o possiedono agenzie che gestiscono l'affitto di immobili a livello locale e nazionale. Questo costituisce una concorrenza sleale agli hotel, i quali svolgono un'attività uguale o equivalente. Molte di queste persone, infatti, nel pubblicare il loro annuncio, inseriscono una serie di servizi aggiuntivi, comparabili con quelli di una struttura professionale. La maggior parte di coloro che gestiscono gli affitti, tuttavia, spesso trovano il modo di superare i controlli, prestando attenzione a far figurare come host il proprietario dell'immobile o un privato cittadino. Questo tipo di "host" gestisce più della metà degli annunci del sito.
L'articolo preso in considerazione è basato sui dati provenienti dal sito "insideairbnb.com" e, in particolare, il dataset di riferimento si basa su 8907 annunci, di cui il 75% riguardanti intere case o appartamenti. Di questi, 6832 possono considerarsi annunci attivi, in quanto sono stati prenotati almeno una volta e hanno ricevuto almeno una recensione negli ultimi sei mesi. È emerso come il 63% appartenga a host con più annunci sul sito (*multihost*) e il 61% è occupato per più di sessanta notti all'anno. Dallo studio si evince, poi, che il 59% degli annunci riguarda, in realtà, annunci commerciali e la maggior parte di essi è situata nel centro storico di Venezia. Un'evidenza del fenomeno trattato nell'articolo si ha dalla disparità dell'incasso medio e mediano di un "host", tramite solo il sito "Airbnb". Infatti, l'incasso medio è di 33095 euro, mentre quello mediano ammonta a 20300 euro. Quest'ultimo varia per tipologia e ubicazione dell'alloggio. L'incasso mediano per l'affitto di

un intero appartamento è di 15096 euro (per un'intera casa nel centro storico è di 17026 euro, nelle isole è di 7634 euro e sulla terraferma è di 11337 euro), mentre per una singola stanza è di 9016 euro. La forte disparità tra incasso medio e mediano indica che la maggior parte degli annunci appartiene a un numero contenuto di "host". I *multihost* gestiscono la parte più consistente dei ricavi e a loro appartiene la quota di mercato più rilevante.
Nei dataset analizzati in questo elaborato si osserva, a conferma di quanto asserito dall'articolo, che diversi annunci presentano un prezzo elevato e che una parte non irrilevante degli alloggi è adeguata ad ospitare un numero abbastanza elevato di persone.

1.1.2 L'impatto del Covid-19 sull'economia di Airbnb

Da un recente articolo di Boros, Dudás & Kovalcsik (2020), si traggono alcune interessanti informazioni, utili per contestualizzare questo elaborato. In seguito alla pandemia globale di Covid-19, il turismo è diminuito e Airbnb non fa eccezione. Questo si evince nell'andamento dei prezzi e delle prenotazioni. Infatti, molti utenti hanno cancellato prenotazioni già effettuate o non hanno prenotato nuovamente dopo la prima ondata dei contagi.
I cambiamenti a causa del Covid-19, relativamente alle prenotazioni, variano molto da città a città e l'articolo considerato indaga sui fattori che influiscono. La maggior parte delle città ha dichiarato lo stato di emergenza l'11 marzo 2020 e questo ha coinciso con le restrizioni di viaggio. I dati su cui si fonda lo studio sono tratti da "insideairbnb.com", nel 2020, compatibilmente con la disponibilità degli stessi, in quindici diverse città, ovvero Londra, New York, Parigi, Sydney, Los Angeles, Pechino, Rio de Janeiro, Copenaghen, Roma, Città del Capo, Madrid, Barcellona, Praga, Tokyo e Milano. Inizialmente si è considerato il primo periodo della pandemia, per osservare il processo che ha influenzato il mercato dei viaggi senza l'intervento degli Stati.
I dati di "Airbnb" sono comparati confrontando i numeri e gli identificativi delle prenotazioni. Si è osservato, in questo modo, che il vero cambiamento si è avuto dopo il terzo mese delle fasi di viaggio. Nel contesto europeo e globale, un alto impatto della pandemia si è avuto a Parigi e a Praga. Inoltre, a Barcellona e Madrid il numero di cancellazioni ha superato quello di nuove

prenotazioni. In particolare, il numero di prenotazioni è diminuito molto da marzo a maggio. Tuttavia, in queste ultime quattro città non sono state rilevate cancellazioni per la fine di aprile 2020. Una decrescita è stata riscontrata anche a Londra e New York, mentre a Pechino le prenotazioni da febbraio a maggio sono state addirittura sospese. A Los Angeles, nonostante si sia verificato un calo di prenotazioni, esso è iniziato dopo la metà di marzo 2020 ed è stato abbastanza costante. Nell'emisfero sud, il picco delle prenotazioni è calato a Sydney, a Rio de Janeiro e a Città del Capo. Si è riscontrato, inoltre un calo delle prenotazioni *last minute*. La pandemia in questa zona, tuttavia, è arrivata più tardi e il calo è dovuto in parte al fatto che in Giappone il turismo fosse legato in dimensioni massicce alla popolazione cinese, la quale, con la pandemia, è stata bandita.
Nella maggior parte dei casi, la fascia di prezzo più colpita dal calo è quella alta, senza che questo significasse una perdita per "Airbnb". Fa eccezione a tale situazione l'Africa con Città del Capo. Gli effetti della pandemia nelle varie città cambiano in base alle caratteristiche dei mercati locali, alla diversa fase dell'emergenza in quello stesso periodo, alle reazioni e alle politiche dei governi. Le caratteristiche locali, inoltre, hanno un ruolo significativo nel dare forma a molti aspetti dei trend delle prenotazioni. I dati dimostrano, poi, che la popolazione ha reagito rapidamente alla pandemia, cancellando le prenotazioni ed evitando di farne di nuove. Quando il turismo è ripartito, i governi hanno supportato gli hotel "tradizionali" in termini finanziari e di regolamento e li hanno favoriti rispetto alle economie *peer-to-peer*. Questo fatto poteva portare al declino dei *multihost*. Le maggiori restrizioni hanno, oltretutto, influito sulla percezione del denaro da parte dei turisti. Infine, anche gli effetti della seconda ondata dei contagi, come riscontrato per la prima ondata, variano in base alla località. Si osserva, per concludere, che la durata della crisi e dei suoi effetti sul lavoro e sui guadagni influenzerà come il mercato e le politiche di governo evolveranno in futuro.

1.2 Airbnb

"Airbnb" è una società che nasce nel 2007 e fonda la sua politica sulla *sharing economy*, ovvero su un modello economico che prevede un rappor-

to diretto tra cliente e fornitore. In particolare, quest'ultimo, detto "host", mette a disposizione degli ospiti il suo alloggio tramite un apposito spazio all'interno di un sito o di un'applicazione. Le informazioni sulla sistemazione sono condivise dal proprietario stesso e gli aspetti cruciali delle abitazioni, per cui si debbano rispettare degli standard minimi, vengono uniformati attraverso opzioni stabilite da "Airbnb". La società oggi comprende circa quattro milioni di "hosts" e un miliardo di ospiti, di provenienza mondiale. Ogni giorno, sono messi a disposizione numerosi alloggi ed esperienze, che permettono agli utilizzatori di entrare in contatto con ambienti diversi da quelli in cui vivono.

Di seguito sono descritte alcune caratteristiche proprie del business di "Airbnb".

1.2.1 Caratteristiche del servizio offerto al cliente di Airbnb

Il cliente di "Airbnb" può effettuare, sul sito, una ricerca degli alloggi adeguati alle sue esigenze. Vi è, inoltre, la possibilità di vedere le recensioni di altri utenti e di mettersi in contatto con l'"host" e con l'organizzazione sia prima sia dopo aver effettuato la prenotazione, per ricevere consigli e assistenza. Queste operazioni si possono svolgere facilmente tramite un'apposita applicazione.

1.2.2 Caratteristiche degli Hosts di Airbnb

La piattaforma "Airbnb" è un'entità. Gli iscritti si chiamano "Members" e sono di fatto gli utenti della piattaforma. Essi si dividono in "Hosts", ovvero coloro che pubblicano e offrono servizi e "Ospiti", ovvero coloro che usufruiscono dei servizi. Gli "hosts" possono proporre "Alloggi" ed "Esperienze". L'offerta e i servizi costituiscono complessivamente un "Annuncio". "Airbnb", inoltre, non possiede nessun annuncio o servizio aggiuntivo. Questi sono unicamente pubblicati dagli "hosts". Sono gli "hosts" a impostare i termini di cancellazione sui loro annunci, seguendo le linee guida stabilite dal sito.

1.2.3 La risposta di Airbnb all'emergenza del Covid-19

Nel contesto della pandemia di Covid19, sono state istituite delle attenuanti per il rimborso delle prenotazioni cancellate, per un certo periodo di tempo. Inoltre, gli "hosts" si sono dovuti impegnare in prima persona per rispettare le norme igieniche e le nuove disposizioni sulla pulizia imposte da "Airbnb." Le "Esperienze" sono state particolarmente colpite dalle restrizioni imposte dai governi, legate al nuovo Coronavirus e, in particolare, è stato necessario sospenderle in molti Paesi.
"Airbnb" ha, inoltre, predisposto dei requisiti di sicurezza per i suoi soggiorni. Tali norme sono obbligatorie sia per gli "host", sia per gli ospiti e sono formulate sulla base delle indicazioni date dall' "Organizzazione Mondiale della Sanità" e dai "Centri per la prevenzione e il controllo delle malattie degli Stati Uniti". Sono sempre in vigore, inoltre, le linee guida generali per la sicurezza, legate all'emergenza e quanto predisposto va accompagnato dalle norme di viaggio di ogni governo. Oltre all'obbligo di indossare la mascherina e del distanziamento di due metri, infatti, è stato predisposto il "Processo di pulizia in 5 fasi", a cui l' "host" deve aderire in tutti gli annunci. Vengono fornite delle vere e proprie linee guida e dei manuali di pulizia. Le precauzioni vanno accompagnate da una tempestiva segnalazione di contagio o di contatto con soggetti contagiati e dal rispetto dei periodi di quarantena.

Capitolo 2

Presentazione del dataset e analisi descrittive

I dataset utilizzati per le analisi provengono dal sito "insideairbnb.com" e si riferiscono agli annunci pubblicati su "Airbnb" per la città di Venezia. Si è scelto di effettuare un confronto tra i dati aggiornati al mese di giugno in tre anni consecutivi, ovvero 2019, 2020, 2021. L'obiettivo è quello di verificare l'impatto della pandemia di Covid-19, che ha caratterizzato il periodo preso in considerazione, con particolare attenzione all'andamento dei prezzi degli alloggi a cui si riferiscono le offerte.

2.1 Analisi preliminari

2.1.1 Preparazione dei dataset

Innanzitutto, si è reso necessario effettuare una preparazione dei dataset, che rendesse possibili i passaggi successivi.
Per prima cosa, sono state eliminate alcune variabili, non ritenute di interesse per le analisi specifiche. Si tratta, ad esempio, di variabili che si riferiscono a recensioni, descrizioni o altri tipi di testo, oppure di alcune informazioni

personali degli "hosts", come nomi e indirizzi email dei padroni di casa.
A questo punto, si è proceduto con l'imputazione di alcuni valori mancanti, valori anomali o considerabili "outliers". Per quanto riguarda le variabili indicanti il numero massimo e minimo di notti prenotabili, sono state considerate solo due delle informazioni disponibili nel file di dati, una indicante il minimo e una il massimo, per evitare che il dataset contenesse informazioni ridondanti. Per evitare, poi, che valori troppo elevati o troppo bassi di queste variabili, distorcessero eccessivamente le stime, si sono considerati solo degli intervalli di valori, ritenuti utili. Nello specifico, per il minimo numero di notti prenotabile è stato fissato un estremo superiore pari a sei notti, ossia circa una settimana e per il massimo numero di notti prenotabile sono previsti un valore minimo di tre notti e un massimo di 1125, che corrisponde circa a tre anni.
La variabile "property_type", che indica a quale tipo di alloggio ci si riferisce, è stata ricodificata in modo da ridursi ad un numero contenuto di modalità ("Apartment", "House", "Bed and breakfast", "Boat", "Camping", "Hotel", "Hostel", "Other"). Per quanto riguarda la collocazione territoriale degli alloggi, è stata conservata solo una delle variabili relative al quartiere. Essa è stata ricodificata in modo da presentare solo due possibili modalità, ovvero "Venezia e isole" e "fuori Venezia". La variabile relativa al numero di bagni, data la codifica originale di difficile interpretabilità, è stata ricodificata con due modalità: "fino a 1", "almeno 2". Tra le variabili che danno informazione sulla disponibilità dell'alloggio si è deciso di mantenere nei dataset solo quella relativa alla disponibilità annuale. In merito alle variabili inerenti le voci di prezzo, i conteggi e le voci di valutazione, si è deciso in questa fase di mantenerle tutte nei dataset. Tuttavia, per rendere possibili le analisi successive, esse sono state in buona parte escluse, a causa di alcune loro caratteristiche, come l'elevata correlazione.
È stata, infine, effettuata una corretta codifica delle variabili presenti nei dataset come numeriche, fattori o date.

2.2 Composizione dei dataset

Il numero di unità statistiche presenti nei tre dataset, in seguito alle operazioni di pulizia degli stessi, è pari a 8637 per il 2019, a 8651 per il 2020 e a 7706 per il 2021. L'unità statistica è il singolo annuncio. Ogni annuncio e ogni "host" sono identificati da una chiave univoca, la quale è ricorrente nei diversi dataset se l'unità statistica a cui ci si riferisce è la stessa in anni diversi. Questo consente di capire quali annunci si ripetono, quali sono nuovi e quali vengono cancellati nell'anno successivo.
I dataset relativi agli anni 2019 e 2020 si compongono all'incirca delle stesse variabili, mentre quello relativo al 2021 è costruito in modo diverso. Pertanto, i primi due file di dati presi in considerazione sono stati facilmente uniformati per quanto riguarda le variabili, ma questa operazione non è stata completamente effettuabile per il terzo.

2.2.1 Variabili

Si riporta la tabella contenente le variabili a disposizione per le analisi, in seguito alla preparazione dei dataset, e la loro descrizione. Tuttavia, non in tutti i file di dati sono presenti esattamente le stesse informazioni. Si vedano le Tabelle 2.1, 2.2, 2.3, 2.4, 2.5, 2.6.

2.3 Analisi descrittiva

Si riporta un'analisi descrittiva del dataset, per visualizzare graficamente le caratteristiche delle variabili a disposizione.

2.3.1 Prezzo per notte

Come specificato in precedenza, in questo elaborato si desidera valutare l'andamento del prezzo degli annunci di Airbnb, tenendo in considerazione l'impatto della pandemia di Covid-19. Osservando l'andamento del prezzo a notte per ogni annuncio, si nota un progressivo aumento. Infatti, nel 2019 e nel 2020, le frequenze maggiori si trovano intorno ai 100 euro a notte, con

Tabella 2.1: Variabili relative alle caratteristiche dell'host

Caratteristiche dell'host		
Nome della variabile	Tipo di variabile	Descrizione
Id	Qualitativa sconnessa	Identificativo annuncio
Host_id	Qualitativa sconnessa	Identificativo "host"
Host_since	Qualitativa sconnessa	Data di inizio dell'attività degli "host"
Host_response_time	Qualitativa ordinale	Velocità di risposta per la prenotazione
Host_response_rate_perc	Quantitativa continua	Percentuale di risposta degli "host"
Host_acceptance_rate_perc	Quantitativa continua	Percentuale di accettazione delle prenotazioni
Host_is_superhost	Qualitativa sconnessa	Indicatore dei "superhost"
Host_total_listings_count	Quantitativa continua	Numero di annunci per "host"
Host_has_a_profile_pic	Qualitativa sconnessa	Indicatore della presenza di foto profilo dell'"host"
Host_identity_verified	Qualitativa sconnessa	Indicatore della verifica dell'identità dell'"host"

Tabella 2.2: Variabili relative alle caratteristiche geografiche

Caratteristiche geografiche		
Nome della variabile	Tipo di variabile	Descrizione
Neighbourhood_cleansed	Qualitativa sconnessa	Indicazione zona: "Venezia e isole", "fuori Venezia"
Market	Qualitativa sconnessa	Indicazione mercato di riferimento: "Venezia" e "Altro"
Latitude	Quantitativa continua	Latitudine
Longitude	Quantitativa continua	Longitudine
Is_location_exact	Qualitativa sconnessa	Indicatore dell'esattezza della posizione dell'alloggio

Tabella 2.3: Variabili relative alle caratteristiche dell'alloggio

Caratteristiche dell'alloggio		
Nome della variabile	Tipo di variabile	Descrizione
Property_type	Qualitativa sconnessa	Indicazione del tipo di alloggio
Room_type	Qualitativa sconnessa	Indicazione del tipo di stanza
Accommodates	Quantitativa discreta	Numero massimo di persone potenzialmente ospitabili
Bathrooms	Qualitativa sconnessa	Indicazione del numero di bagni: "fino a uno", "almeno due"
Bedrooms	Quantitativa discreta	Numero di camere da letto
Beds	Quantitativa discreta	Numero letti
Bed_type	Qualitativa sconnessa	Indicazione sul tipo di letti: "Bed", "Altro"

Tabella 2.4: Variabili relative alle caratteristiche del prezzo

Caratteristiche del prezzo		
Nome della variabile	Tipo di variabile	Descrizione
Price_dollars	Quantitativa continua	Prezzo complessivo per notte (dollari)
Weekly_price_dollars	Quantitativa continua	Prezzo complessivo a settimana (dollari)
Monthly_price_dollars	Quantitativa continua	Prezzo complessivo al mese (dollari)
Security_deposit_dollars	Quantitativa continua	Prezzo del deposito di sicurezza (dollari)
Cleaning_fee_dollars	Quantitativa continua	Prezzo delle pulizie (dollari)
Guests_included	Quantitativa discreta	Numero di ospiti previsti dall'annuncio
Extra_people_dollars	Quantitativa continua	Prezzo per le persone aggiuntive (dollari)

Tabella 2.5: Variabili relative alle caratteristiche dell'annuncio

Caratteristiche dell'annuncio		
Nome della variabile	Tipo di variabile	Descrizione
Minimum_nights	Quantitativa continua	Minimo di notti prenotabili
Maximum_nights	Quantitativa continua	Massimo di notti prenotabili
Availability_365	Quantitativa continua	Numero di giorni in cui l'alloggio è prenotabile in un anno
Calendar_last_scraped	Qualitativa sconnessa	Data di ultimo aggiornamento dell'annuncio
Number_of_reviews	Quantitativa discreta	Numero di revisioni complessive dell'annuncio
Number_of_reviews_ltm	Quantitativa discreta	Numero di revisioni dell'annuncio nell'ultimo mese
First_review	Qualitativa sconnessa	Data del primo aggiornamento dell'annuncio
Last_review	Qualitativa sconnessa	Data, dell'ultimo aggiornamento dell'annuncio
Review_scores_rating	Quantitativa continua	Punteggio attribuito dagli ospiti all'esperienza complessiva
Review_scores_accuracy	Quantitativa continua	Punteggio attribuito dagli ospiti all'accuratezza delle informazioni
Review_scores_checkin	Quantitativa continua	Punteggio attribuito dagli ospiti alla semplicità del checkin
Review_scores_cleanliness	Quantitativa continua	Punteggio attribuito dagli ospiti alla pulizia e alle condizioni dell'alloggio
Review_scores_communication	Quantitativa continua	Punteggio attribuito dagli ospiti alla comunicazione tra host e ospite
Review_scores_location	Quantitativa continua	Punteggio attribuito dagli ospiti alla posizione dell'alloggio
Review_scores_value	Quantitativa continua	Punteggio attribuito dagli ospiti al rapporto qualità prezzo dell'alloggio
Instant_bookable	Qualitativa sconnessa	Indicatore della possibilità di prenotare da subito
Cancellation_policy	Qualitativa sconnessa	Indicazione della politica di cancellazione adottata
Require_guest_profile_picture	Qualitativa sconnessa	Indicatore della richiesta di un'immagine profilo dell'ospite
Require_guest_phone_verification	Qualitativa sconnessa	Indicatore della richiesta di una verifica telefonica dell'ospite
Calculated_host_listings_count	Quantitativa continua	Numero di annunci dell'"host"
Calculated_host_listings_count_entire_homes	Quantitativa continua	Numero di annunci dell'"host" per un'intera casa
Calculated_host_listings_count_private_rooms	Quantitativa continua	Numero di annunci dell'"host" per stanze private
Calculated_host_listings_count_shared_rooms	Quantitativa continua	Numero di annunci dell'"host" per stanze condivise
Reviews_per_month	Quantitativa continua	Numero di revisioni al mese

Tabella 2.6: Altre variabili

Altre variabili		
Nome della variabile	Tipo di variabile	Descrizione
Periodo	Quantitativa continua	Indicatore del dataset
Price_dollars_norm	Quantitativa continua	Prezzo per notte per persona
Weekly_price_dollars_norm	Quantitativa continua	Prezzo settimanale per persona
Monthly_price_dollars_norm	Quantitativa continua	Prezzo mensile per persona
Cleaning_fee_dollars_norm	Quantitativa continua	Prezzo per le pulizie per persona
Security_deposit_dollars_norm	Quantitativa continua	Prezzo del deposito di sicurezza per persona
Prom_week	Qualitativa sconnessa	Indicatore di uno sconto sul prezzo settimanale per persona rispetto al prezzo giornaliero
Prom_month	Qualitativa sconnessa	Indicatore di uno sconto sul prezzo mensile per persona rispetto al prezzo giornaliero

un calo progressivo per le somme più elevate. Nel 2021, questa disparità è di poco meno accentuata e si osserva che le fasce di prezzo superiori sono più frequenti rispetto agli anni precedenti.
Per una corretta interpretazione è opportuno considerare la diversa numerosità dei tre dataset, in particolare il 2021 comprende significativamente meno annunci degli altri due file di dati e questo spiega una curva delle frequenze generalmente più bassa. Si veda la Figura 2.1. Si riporta una tabella di sintesi dell'andamento della variabile relativa al prezzo. Si veda la Tabella 2.7.

Voci di prezzo

Il prezzo settimanale, disponibile solo per i dataset relativi al 2019 e 2020, evidenzia un aumento dei prezzi tra i primi due anni considerati. In questo caso, le osservazioni disponibili per il 2020 sono di poco superiori a quelle del 2019. Si osserva, quindi, che le fasce di prezzo settimanale più elevate sono più frequenti nel secondo periodo considerato, mentre nel primo periodo, si hanno frequenze più elevate per cifre inferiori. Si veda la Figura 2.2
Considerazioni analoghe a quelle fatte per il prezzo settimanale valgono per il prezzo mensile negli anni 2019 e 2020. Anche in questo caso, la variabile in questione non è rilevata per l'anno 2021. Si veda la Figura 2.3

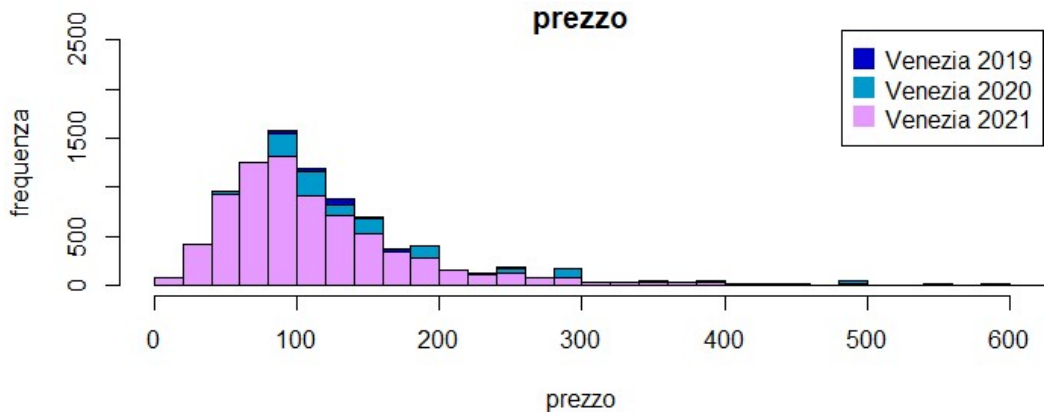

Figura 2.1: Prezzo per notte degli annunci

Altre voci di prezzo disponibili solo per i primi due periodi oggetto di analisi sono il prezzo del deposito di sicurezza, le spese delle pulizie e quelle

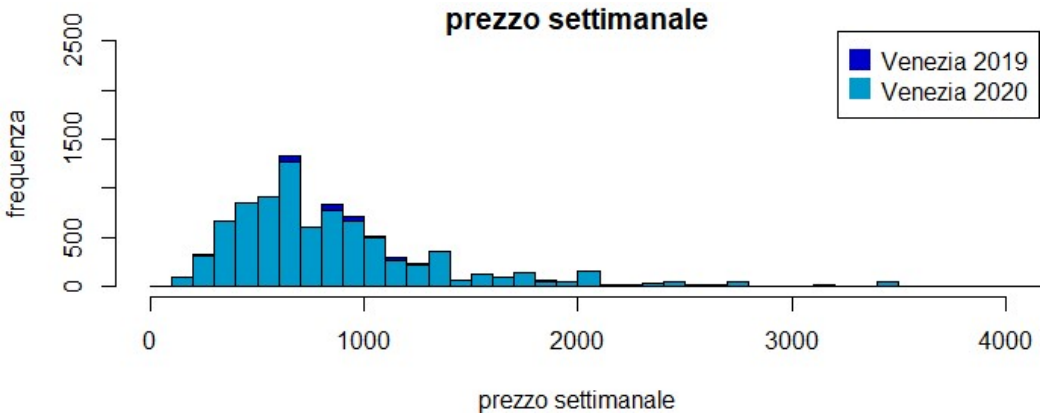

Figura 2.2: Prezzo settimanale degli annunci

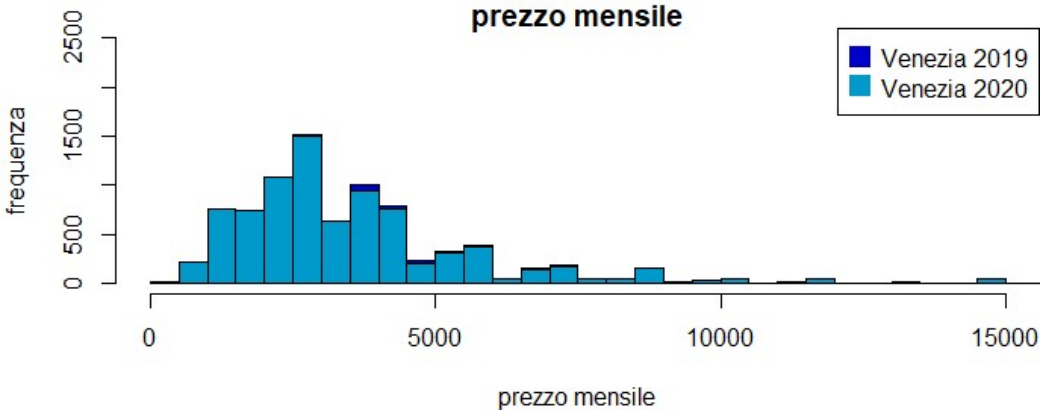

Figura 2.3: Prezzo mensile degli annunci

per gli ospiti aggiuntivi.
Relativamente al deposito di sicurezza, si osserva un andamento praticamente identico alle variabili precedentemente descritte, leggermente più elevato per il 2019, nei due anni e, in entrambi i casi, il costo sembra essere contenuto, se non prossimo allo zero. Un'informazione rilevante è data dal prezzo per le pulizie, di cui si ha più informazione per l'anno 2020 e che, in tale periodo, sembra essere più elevato. Questo è, probabilmente, legato alla necessità di rispettare nuove norme per limitare la diffusione del virus. Infine, l'andamento del prezzo per le persone aggiuntive negli anni 2019 e 2020 non risulta cambiare da un anno all'altro. Si riporta una tabella di sintesi dell'andamento della variabile relativa alle varie voci di prezzo. Si veda la Tabella 2.8.
Si è, a questo punto, voluto indagare se ci fossero delle promozioni settimanali o mensili e quanto frequenti fossero. Quanto ne è emerso non è risultato interessante per analisi successive, ma le operazioni svolte vengono riportate per completezza. Sono state create, a questo scopo, due variabili dicotomiche per ogni dataset. La prima di queste indica la presenza di una promozione settimanale ("1"= "sì", "0"= "no") ed è pari a zero se il prezzo settimanale è uguale o superiore al prezzo giornaliero moltiplicato per il numero dei giorni di una settimana, mentre è pari a uno se è inferiore. Con lo stesso criterio è stata costruita la variabile relativa alla promozione mensile. Si evidenzia che il numero di promozioni è molto simile tra settimanali e mensili e, in entrambi i casi, la loro frequenza è superiore nel 2019 rispetto al 2020. Questo fa pensare ad una situazione di minore ricchezza per gli hosts, che, avendo riscontrato un calo di prenotazioni e dovendo sostenere spese superiori per garantire la sicurezza dei loro ospiti sono meno propensi ad effettuare degli sconti. Un'altra possibile interpretazione è che si siano riscontrate generalmente prenotazioni per intervalli di tempo più corti e che, quindi, non vi fosse una effettiva convenienza per gli hosts nell'effettuare una operazione di marketing con questo genere di promozioni.
Queste congetture non sono, tuttavia, verificabili perché sono disponibili unicamente le informazioni relative agli annunci e non alle prenotazioni.

2.3.2 Annunci su Airbnb

Relativamente agli annunci, inoltre, non è presente il numero effettivo di notti a cui essi si riferiscono, ma solo alcune misure di massima o minima disponibilità degli alloggi.
Relativamente alla tipologia di alloggi affittabili, si osserva che essi sono suddivisi in tre o quattro categorie, a seconda del dataset. La maggior parte degli annunci si riferisce a case o appartamenti interi. Una parte abbastanza contenuta di essi, ma significativa considerando la politica degli scambi forniti da Airbnb, è relativa alle stanze in hotel. Inoltre, una percentuale importante riguarda stanze private e una percentuale molto contenuta camere condivise. Nello specifico, si osserva per l'anno 2019 che circa il 76% degli annunci riguarda intere case o appartamenti, circa il 23% stanze private e circa lo 0.74% sistemazioni condivise. Nel 2020, invece, circa il 75% si riferisce a interi alloggi, circa il 3.56% a stanze in hotel, circa il 20% a camere private e circa il 0.75% a camere condivise. Infine, nel 2021, circa il 76% degli annunci è relativo a case o appartamenti interi, circa il 3.53% a camere di hotel, circa il 20% a stanze private e circa il 0.57% a stanze condivise. Si riportano di seguito i grafici relativi a queste informazioni. Si veda la Figura 2.4.
Osservando il numero di letti a cui si riferiscono gli annunci, si nota, innan-

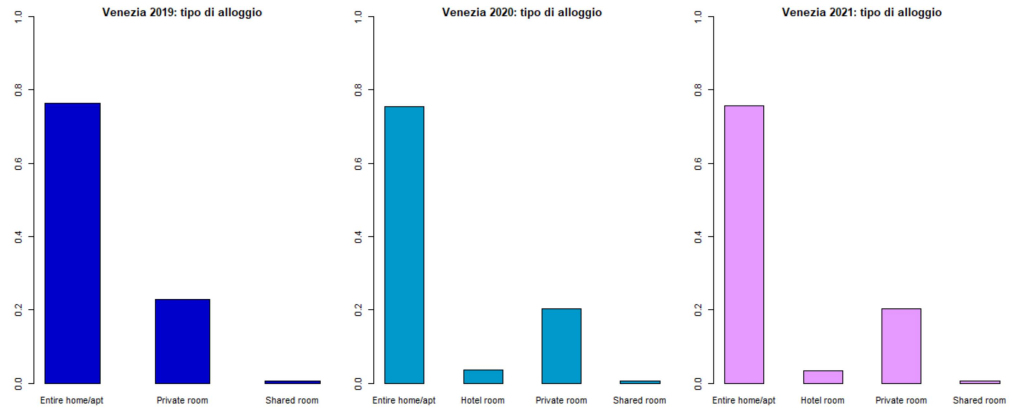

Figura 2.4: Tipo di alloggio: *entire home/apt, private room, shared room, hotel room*

zitutto, che vi è una piccola percentuale di offerte con 0 letti prenotabili. Si

tratta di un valore molto basso, ma in crescita, ossia 0.2% nel 2019, 1.9% nel 2020 e 2.3% nel 2021. Si possono interpretare come errori o come alloggi effettivamente non disponibili. L'aumento, in questo secondo caso, si può spiegare con la scelta di un "host" di non affittare la sua proprietà in seguito alla pandemia. La maggior parte degli "hosts", mette a disposizione da uno a quattro letti, con percentuali che vanno da circa il 30% per gli annunci relativi a un solo letto a circa il 11% per annunci con quattro letti. Numeri di letti superiori sono molto meno frequentemente disponibili. Le percentuali per ogni modalità di questa variabile sono molto simili nei tre dataset con valore leggermente più alto nel 2019, un lievissimo calo nel 2020 e una crescita molto contenuta nel 2021.
Una situazione molto simile si ha per quanto riguarda il numero di camere disponibili per ogni alloggio. Vi è una bassissima percentuale di alloggi con zero camere disponibili nel 2019 e nel 2020, ma non nel 2021. In questo caso l'informazione non è interpretabile. La maggior parte degli alloggi ha una o due camere disponibili con percentuali vicine al 56% e al 31% in tutti e tre gli anni. Non si osservano differenze importanti per il numero di camere offerte, che si possano ricondurre all'effetto dell'emergenza.
Nemmeno per quanto riguarda il numero di bagni si osservano cambiamenti nei tre periodi considerati. Infatti, in tutti e tre gli anni la percentuale di "hosts" che mette a disposizione fino a un bagno è circa il 24%, mentre circa il 76% degli annunci prevede almeno due bagni. Questo è comprensibile, considerando che, nella maggior parte dei casi, ci si riferisce a intere case o appartamenti.
Ogni annuncio riguarda un determinato numero di ospiti. Tale informazione è disponibile solo per gli anni 2019 e 2020. È opportuno sottolineare che questo valore non coincide con il numero massimo di persone ospitabili, per il quale è predisposta una apposita variabile, presente, invece, in tutti e tre i dataset. Per quanto riguarda il numero di ospiti, si osserva che in entrambi gli anni gli annunci riguardano per la maggior parte una o due persone e comunque raramente più di sei. Infatti, nel 2019 il 48% di essi si riferisce a una sola persona, mentre il 34% a due persone. Nel 2020, invece, tali percentuali sono rispettivamente il 47% e il 36% circa. Per quanto riguarda il numero massimo di persone ammissibili per ogni alloggio, le percentuali maggiori in tutti e tre

i dataset sono comprese tra le due e le sei persone, con i valori più elevati per gli alloggi con al massimo due prenotazioni possibili. Questa percentuale corrisponde circa al 27% nei tre anni consecutivi. Si osserva che per il 2021 è prevista anche una modalità "zero", molto poco frequente, per quest'ultima variabile. Potrebbe, pertanto trattarsi di un errore.
Coerentemente con la politica di Airbnb, la quale prevede che l'host metta a disposizione una sua proprietà, la maggior parte degli annunci, ossia circa l'85% per ogni dataset a disposizione, riguarda gli appartamenti, una percentuale compresa tra il 4% e il 5% riguarda bed and breakfast, le case sono circa il 7%, per gli hotel si hanno percentuali tra il 2% o il 4% circa, mentre le altre modalità della variabile considerata, ovvero barche, campeggi, ostelli e altre sistemazioni, sono rappresentate in quantità molto inferiore. Non si osservano differenze significative nei tre anni. Si vedano le Figure 2.5, 2.6, 2.7
Nel 2019 e nel 2020, inoltre, circa l'82% delle persone metteva a disposizione

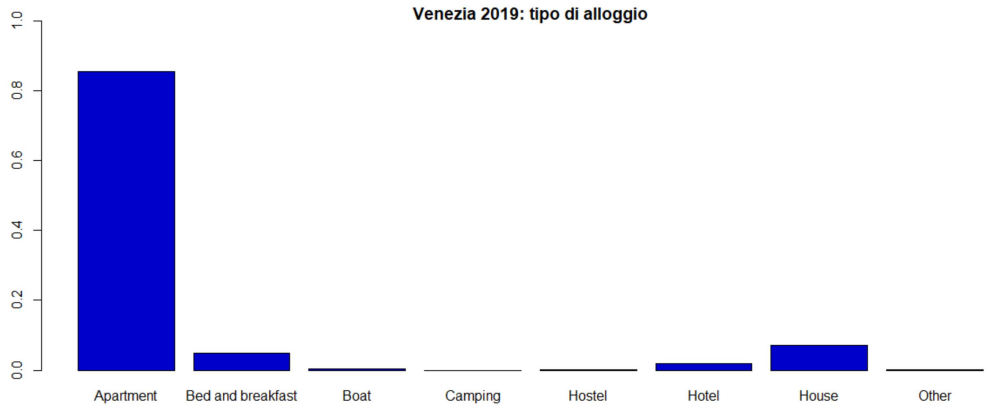

Figura 2.5: Tipo di alloggio nel 2019: *apartment, bed and breakfast, boat, camping, hostel, hotel, house, other*

abitazioni a Venezia e nelle isole, mentre circa il 18% fuori dalle isole. Nel 2021 le percentuali cambiano leggermente, diventando rispettivamente circa 83% e 17%, ma tali valori non risultano interpretabili alla luce dell'emergenza che gli "hosts" hanno dovuto affrontare. Si veda la Figura 2.8
Si osserva, poi, come il numero massimo e il numero minimo di notti prenotabili tendano ad abbassarsi in modo molto contenuto. In particolare, nel 2019 si tendeva a fissare più spesso un numero minimo pari a due o tre notti,

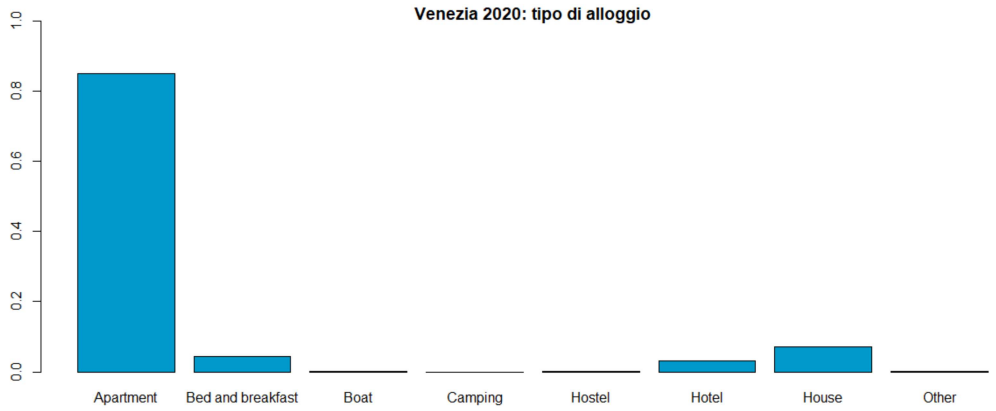

Figura 2.6: Tipo di alloggio nel 2020: *apartment, bed and breakfast, boat, camping, hostel, hotel, house, other*

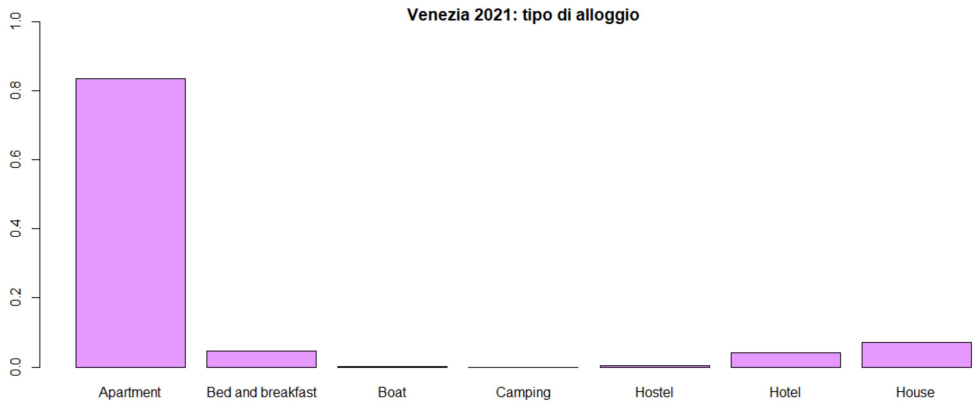

Figura 2.7: Tipo di alloggio nel 2021: *apartment, bed and breakfast, boat, camping, hostel, hotel, house, other*

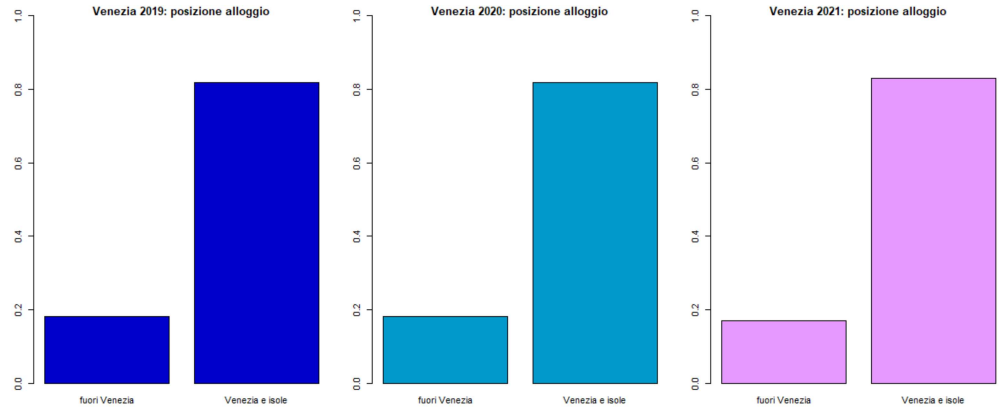

Figura 2.8: Posizione dell'alloggio:fuori Venezia, Venezia e isole

mentre negli anni successivi, aumenta la percentuale di alloggi prenotabili per un minimo di una notte. La disponibilità annuale degli alloggi, inoltre, non sembra subire modifiche rilevanti, osservabili graficamente.
In tutti e tre i periodi, alla maggior parte degli annunci sono associati punteggi elevati di valutazione dell'esperienza. Si segnala che nel 2021 sembrerebbe essere cambiata la scala di valutazione, e il massimo punteggio assegnabile risulta essere notevolmente più basso rispetto agli anni precedenti. Tuttavia, l'andamento della variabile resta immutato. I prezzi più alti degli annunci si trovano in corrispondenza a punteggi molto bassi o molto elevati.
Osservando i valori delle correlazioni per l'anno 2019, si nota che i punteggi di recensione dei vari aspetti relativi agli annunci sono positivamente correlati tra loro e negativamente correlati con il prezzo. Si tratta di valori rispettivamente pari a 0.97, 0.98, 0.99 e -0.11, -0,12. Questo mette in evidenza come siano state recensite meglio sistemazioni con prezzo inferiore. Si tratta di una specifica caratteristica degli utenti di Airbnb. Infatti, chi preferisce questo tipo di affitti, punta alla convenienza delle offerte e non a servizi particolari o a strutture lussuose e questo influenza i punteggi delle recensioni. Una situazione analoga, per quanto riguarda l'analisi della correlazione, si ha nel dataset relativo al 2020, in cui sono disponibili le stesse variabili. Tuttavia, per una approfondita analisi delle correlazioni si rimanda alla Appendice 1. Per ulteriori grafici relativi alle analisi descrittive, invece, si rimanda alla Ap-

pendice 2.
Si riportano, infine, alcune tabelle di sintesi dell'andamento delle variabili quantitative, delle variabili relative alle revisioni da parte degli ospiti e delle variabili relative ai conteggi dei tipi di alloggio. Si vedano le Tabelle 2.9, 2.10, 2.11.

Tabella 2.7: Distribuzione della variabile prezzo

Prezzo

Anno	Variabile	min	quartile 1	mediana	media	quartile 3	max	errore std
2019	Price_dollars	9	79	110	137.3971	150	8234	176.9558
2020	Price_dollars	9	77	108	146.976	150	8459	200.9529
2021	Price_dollars	5	70	100	139.5067	15	9999	269.5842

Tabella 2.8: Distribuzione delle variabili relative alle voci di prezzo

Voci di prezzo

Anno	Variabile	min	quartile 1	mediana	media	quartile 3	max	errore std
2019	Weekly_price	63	553	763	956.4	1050	57638	1233.5
2020	Weekly_price	63	539	750	1026.3	1050	59213	1404.6
2019	Monthly_price	270	2250	3100	4037.9	4500	247020	5316.8
2020	Monthly_price	270	2250	3030	4342.9	4500	253770	6039.6
2019	Security_dep	0	0	0	85.99	150	4333	215.1
2020	Security_dep	0	0	0	84.00	100	4400	216.0
2019	Cleaning_fee	0	0	30	34.42	50	600	33.30
2020	Cleaning_fee	0	0	30	36.09	50	2222	42.70
2019	Extra_people	0	0	5	11.01	20	278	17.09
2020	Extra_people	0	0	5	10.95	20	278	16.54

Si riportano la mappa di Venezia, con la suddivisione in sestieri e un'immagine del territorio limitrofo, con l'indicazione di alcune delle località. Si vedano le Figure 2.9 e 2.10.

Tabella 2.9: Distribuzione delle variabili relative alle caratteristiche dell'alloggio

Caratteristiche dell'alloggio

Anno	Variabile	min	quartile 1	mediana	media	quartile 3	max	errore std
2019	Accommodates	1	2	4	3.91	5	26	1.92
2020	Accommodates	1	2	4	3.92	5	26	1.97
2021	Accommodates	0	2	4	3.93	5	16	2.00
2019	Bedrooms	0	1	1	1.57	2	10	0.84
2020	Bedrooms	0	1	1	1.57	2	14	0.87
2021	Bedrooms	1	1	1	1.61	2	14	0.89
2019	Beds	0	1	2	2.59	3	16	1.64
2020	Beds	0	1	2	2.47	3	19	1.59
2021	Beds	0	1	2	2.47	3	19	1.64
2019	Minimum_nights	1	1	2	2.04	3	6	0.96
2020	Minimum_nights	1	1	2	2.00	2	6	1.01
2021	Minimum_nights	1	1	2	2.03	2	6	1.07
2019	Maximum_nights	3	30	365	572.34	1125	1125	534.10
2020	Maximum_nights	3	30	365	555.88	1125	1125	529.17
2021	Maximum_nights	3	30	365	552.66	1125	1125	525.40
2019	Availability_365	0	140	257	222.30	313	365	109.97
2020	Availability_365	0	162	302	244.82	357	365	126.92
2021	Availability_365	0	90	243	211.75	343	365	135.28
2019	Number_reviews	0	4	20	53.35	72	667	76.98
2020	Number_reviews	0	3	24	59.57	81	899	85.02
2021	Number_reviews	0	2	20	58.17	79	765	87.15
2019	Num_reviews_ltm	0	1	8	18.06	27	276	23.33
2020	Num_reviews_ltm	0	0	6	13.01	20	291	17.24
2021	Num_reviews_ltm	0	0	0	3.91	3	222	9.10

Tabella 2.10: Distribuzione delle variabili relative alle voci di revisione

Voci di revisione

Anno	Variabile	min	quartile 1	mediana	media	quartile 3	max	errore std
2019	Reviews_sc_rating	0	86	94	81.54	98	100	31.27
2020	Reviews_sc_rating	0	84	94	78.79	98	100	34.08
2021	Reviews_sc_rating	0	42.1	47.1	38.38	49.0	50.0	18.35
2019	Reviews_sc_accuracy	0	9	10	8.42	10	10	3.22
2020	Reviews_sc_accuracy	0	9	10	8.12	10	10	3.51
2021	Reviews_sc_accuracy	0	44.3	48.2	39.05	49.5	50.0	18.67
2019	Reviews_sc_cleanliness	0	9	10	8.39	10	10	3.21
2020	Reviews_sc_cleanliness	0	9	10	8.09	10	10	3.50
2021	Reviews_sc_cleanliness	0	43.8	48.1	39.01	49.5	50.0	18.65
2019	Reviews_sc_checkin	0	9	10	8.46	10	10	3.23
2020	Reviews_sc_checkin	0	9	10	8.17	10	10	3.52
2021	Reviews_sc_checkin	0	45.0	48.6	39.34	49.7	50.0	18.77
2019	Reviews_sc_communication	0	9	10	8.44	10	10	3.23
2020	Reviews_sc_communication	0	9	10	8.15	10	10	3.52
2021	Reviews_sc_communication	0	44.8	48.6	39.30	49.7	50.0	18.77
2019	Reviews_sc_location	0	9	10	8.54	10	10	3.22
2020	Reviews_sc_location	0	10	10	8.27	10	10	3.53
2021	Reviews_sc_location	0	46.0	48.7	39.64	49.7	50.0	18.78
2019	Reviews_sc_value	0	9	9	8.12	10	10	3.13
2020	Reviews_sc_value	0	9	9	7.86	10	10	3.41
2021	Reviews_sc_value	0	42.5	46.7	38.09	48.3	50.0	18.23

Tabella 2.11: Distribuzione delle variabili relative alle voci di conteggio

Voci di conteggio

Anno	Variabile	min	quartile 1	mediana	media	quartile 3	max	errore std
2019	Calc_host_listings_count	1	1	3	10.81	8	114	20.77
2020	Calc_host_listings_count	1	1	3	9.98	8	121	18.99
2021	Calc_host_listings_count	1	1	3	9.74	7	121	20.10
2019	Listings_count_entire_homes	0	1	2	9.20	6	114	20.01
2020	Listings_count_entire_homes	0	1	2	8.22	6	121	18.52
2021	Listings_count_entire_homes	0	1	2	7.74	5	121	18.39
2019	Listings_count_private_rooms	0	0	0	1.57	1	28	4.23
2020	Listings_count_private_rooms	0	0	0	1.21	1	26	3.23
2021	Listings_count_private_rooms	0	0	0	1.46	1	41	5.03
2019	Listings_count_shared_rooms	0	0	0	0.03	0	5	0.32
2020	Listings_count_shared_rooms	0	0	0	0.04	0	7	0.40
2021	Listings_count_shared_rooms	0	0	0	0.02	0	7	0.30

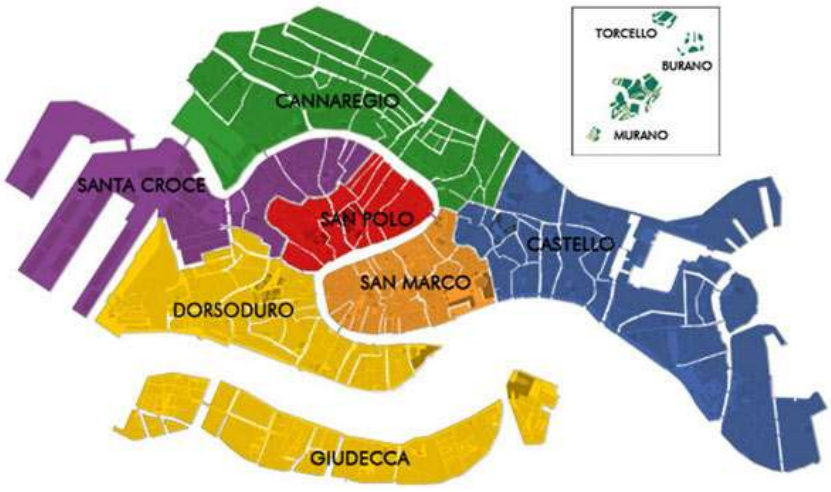

Figura 2.9: Mappa di Venezia con suddivisione in sestieri
https://evenice.it/venezia/storie-tradizioni/venezia-i-suoi-sestieri

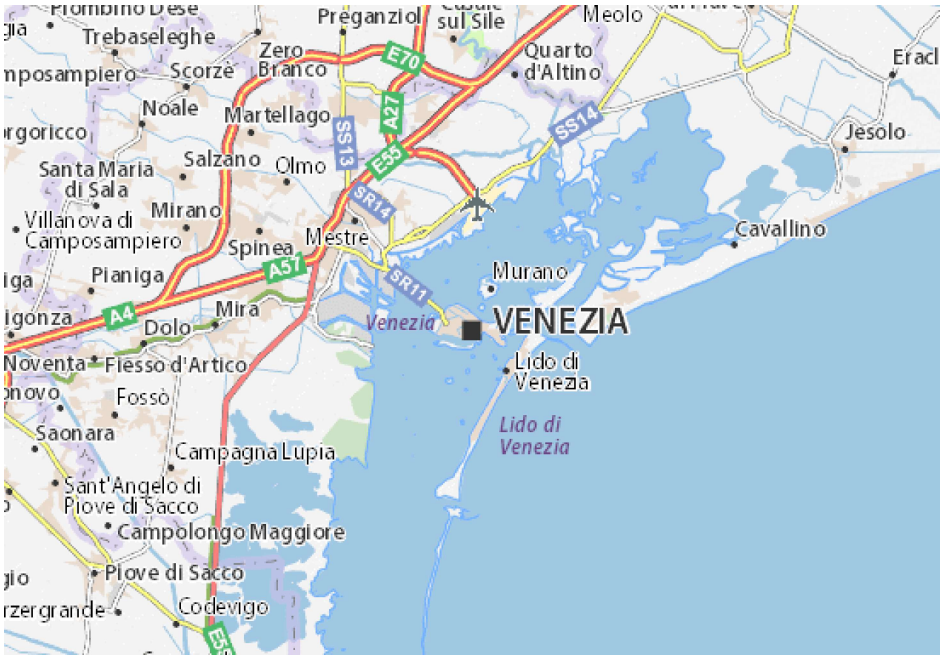

Figura 2.10: Venezia e il territorio limitrofo
https://www.viamichelin.it/web/Mappe-Piantine/Mappa_Piantina-Venezia-30121-Venezia-Italia

Capitolo 3

Il prezzo delle case su Airbnb: analisi di regressione

In questo capitolo si riportano i risultati dell'applicazione di modelli di regressione ai diversi dataset. Lo scopo di tale analisi è quello di capire quali di questi modelli possono essere utilizzati per stimare l'andamento del prezzo con più elevata precisione. Si desidera, inoltre, indagare quali siano le variabili che influenzano maggiormente il comportamento del prezzo nei tre anni e in che modo esse siano rilevanti. Infine, si effettua un confronto dei risultati dell'applicazione dei modelli nei diversi dataset.

3.1 Preparazione del dataset

Innanzitutto, si è resa necessaria un'ulteriore preparazione dei dataset, in modo che le variabili fossero utilizzabili nei modelli.
Le operazioni effettuate sono le seguenti:

- Eliminazione delle variabili più fortemente correlate tra loro..
- Codifica delle variabili qualitative ordinali come numeriche per un più ampio utilizzo di queste all'interno dei modelli proposti.

- Eliminazione delle variabili con un numero di valori mancanti ritenuto eccessivamente elevato. Si tratterebbe di una fonte di distorsione, che non potrebbe apportare rilevanti informazioni nella regressione.

- Imputazione di alcuni valori mancanti delle variabili.

- Eliminazione di variabili ritenute poco utili per le analisi e di variabili costituenti una fonte di informazione ridondante.

- Trasformazione logaritmica della variabile relativa al prezzo per notte, a causa della sua distribuizione asimmetrica. Trattandosi della variabile risposta, i modelli applicati operano su scala trasformata. Tuttavia, dal momento che si applica il logaritmo sia nell'insieme di stima, sia in quello di verifica, il calcolo della funzione obiettivo ("MSE") resta invariato. Esso è ottenuto calcolando la media dei quadrati delle differenze tra la trasformazione logaritmica della previsione e la trasformazione logaritmica della variabile risposta nell'insieme di verifica.

Al termine di questa fase, il dataset relativo al 2019 si compone delle seguenti ventisette variabili: "host_response_time", "host_response_rate_perc", "host_is_superhost", "host_total_listings_count", "host_has_a_profile_pic", "host_identity_verified", "neighbourhood_cleansed", "is_location_exact", "property_type", "room_type", "accommodates", "bathrooms", "bedrooms", "beds", "bed_type", "price_dollars", "guests_included", "minimum_nights", "maximum_nights", "availability_365", "number_of_reviews", "review_scores_rating", "review_scores_value", "instant_bookable", "cancellation_policy", "calculated_host_listings_count", "reviews_per_month".
I set di dati riguardanti i periodi successivi si compongono rispettivamente di ventisette e ventitre variabili. Dal momento che l'informazione disponibile è leggermente differente nei vari periodi considerati, i risultati delle tre analisi non sono perfettamente confrontabili, ma mettono in evidenza le differenze tra le variabili che influiscono sui prezzi in anni successivi.
Il dataset relativo al 2020 differisce dal precedente per l'aggiunta di

"host_acceptance_rate_perc", e l'assenza di "reviews_per_month", mentre quello relativo al 2021 per l'aggiunta di "host_acceptance_rate_perc" e l'assenza di "is_location_exact", "bed_type", "cancellation_policy", "guests_included" e "reviews_per_month".
Per effettuare al meglio le analisi, si suddividono i datastet in un insieme di stima e in uno di verifica, composti rispettivamente dal 75% e dal 25% delle osservazioni dell'intero set di dati.

3.2 Modelli applicati

I modelli utilizzati per le analisi sono i seguenti:

- Modello di regressione lineare: tutte le variabili
- Modello di regressione lineare con selezione delle variabili tramite procedura *stepwise*
- Alberi di regressione
- MARS
- GAM
- Bagging
- Foreste casuali (*Random forest*)
- Gradient boosting

Si riportano di seguito una indicazione sintetica della performance dei modelli nei tre dataset in termini di errore quadratico medio ("MSE") e una descrizione più dettagliata dei risultati ottenuti dal modello di regressione lineare e dal modello *Gradient Boosting*. Infatti, si tratta rispettivamente del modello più facilmente interpretabile e di quello più efficiente, ritenuto, pertanto, il più adatto per la stima del prezzo. Si vedano le Tabelle 3.1, 3.2 e 3.3.

Tabella 3.1: Performance dei modelli applicati al dataset del 2019 in termini di errore quadratico medio (MSE)

Modello applicato	Errore quadratico medio (MSE)
Modello lineare	0.1613889
Modello lineare stepwise	0.1616326
Alberi di regressione	0.2285655
MARS	0.2192893
GAM	0.2162799
Bagging	0.1778015
Foreste casuali	0.1210568
Gradient boosting	0.1249851

Tabella 3.2: Performance dei modelli applicati al dataset del 2020 in termini di errore quadratico medio (MSE)

Modello applicato	Errore quadratico medio (MSE)
Modello lineare	0.2502140
Modello lineare stepwise	0.2507255
Alberi di regressione	0.2848401
MARS	0.2816277
GAM	0.2957768
Bagging	0.2306915
Foreste casuali	0.1611946
Gradient boosting	0.1599949

Tabella 3.3: Performance dei modelli applicati al dataset del 2021 in termini di errore quadratico medio (MSE)

Modello applicato	Errore quadratico medio (MSE)
Modello lineare	0.2228500
Modello lineare stepwise	0.2232441
Alberi di regressione	0.3180029
MARS	0.2690975
GAM	0.2792688
Bagging	0.2466694
Foreste casuali	0.1742114
Gradient boosting	0.1788557

3.3 Modello di regressione lineare

Il primo modello applicato ai dati è il modello di regressione lineare. Dal valore dell'errore quadratico medio, si osserva che non è il miglior modello per prevedere il valore del prezzo, in nessun dataset. Si ritiene opportuno applicare questo modello all'intero set di stima all'inizio delle analisi, in quanto costituisce il metodo più semplice per stimare il contributo di ogni variabile al valore del prezzo, mediante una relazione di tipo lineare, con coefficienti stimati con i minimi quadrati, dalla cui significatività si traggono informazioni importanti su quali siano le variabili più rilevanti per le stime.

3.3.1 Cenni teorici sul modello

Si descrivono qui, in sintesi, alcuni aspetti teorici relativi al modello lineare, tratti da Azzalini & Scarpa (2009).
Il modello di regressione lineare è una funzione dei dati e dei parametri del tipo $y = f(x, \beta) + \varepsilon$, in cui $y = (y_1, ..., y_n)^T$ è il vettore delle variabili risposta,

$X = (x_1, ..., x_n)$ è la matrice contenente i vettori di osservazioni sulle variabili esplicative, detta anche "matrice di regressione", $\beta = (\beta_1, ..., \beta_p)^T$ è il vettore dei parametri e $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)^T$ è il vettore dei termini di errore, con p pari al numero di parametri e n pari al numero di osservazioni. Per ipotesi, gli errori seguono una distribuzione normale con media nulla e varianza costante e ignota σ^2. Inoltre, la funzione f è del tipo $f(x, \beta) = (f(x_1, \beta), ..., f(x_n, \beta))$. Il vettore dei parametri β è stimato con il criterio dei minimi quadrati e corrisponde ai valori che minimizzano la seguente funzione obiettivo:

$$D(\beta) = \sum_i (y_i - f(x_i, \beta))^2$$

Si tratta della devianza, che fornisce una quantificazione della discrepanza fra i valori stimati e osservati.
Si ottiene, così, il modello di regressione lineare $y = X\beta + \varepsilon$. Da esso, è possibile ricavare la forma esplicita del vettore β, come soluzione del problema di minimizzazione con i minimi quadrati, ossia:

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

Si ottiene, inoltre, il vettore dei valori stimati $\hat{y} = X\hat{\beta} = Py$, dove P è la matrice di proiezione (nxn), tale che $P = X(X^T X)^{-1} X^T$, $P = P^T$, $PP = P$ e $tr(P) = rk(P) = p$. A questo punto, è possibile ottenere una stima di σ^2, ovvero $s^2 = D(\beta)/(n - p)$, da cui, poi, si ricava la stima della varianza dei parametri stimati, $\hat{V}(\hat{\beta}) = s^2(X^T X)^{-1}$.
Imponendo la normalità dei termini di errore, le quantità presentate sono fondamentali per definire i risultati del modello. Infatti, dall'applicazione della regressione lineare, si ottengono:

- Le stime dei coefficienti $\hat{\beta}$
- Le stime degli errori standard delle componenti di $\hat{\beta}$, che corrispondono alla radice quadrata degli elementi diagonali della matrice $\hat{V}(\hat{\beta})$
- I valori normalizzati delle stime di β: t = *stima/errore standard*
- I "valori-p" o "livelli di significatività" delle stime

Il modello così costruito si intende lineare nei parametri. Le variabili, invece, possono subire trasformazioni anche di tipo non lineare.
Il modello proposto si presta favorevolmente all'applicazione di metodi di selezione automatica iterativa delle variabili di tipo *stepwise*. Essi producono una variante più parsimoniosa rispetto al modello descritto, in quanto viene selezionata automaticamente una parte del dataset di stima, scegliendo le variabili esplicative più significative tramite tre diverse varianti della procedura. Si tratta di un inserimento iterativo delle variabili partendo dal modello con la sola intercetta fino ad una modello più completo (*stepwise forward*), di una eliminazione iterativa delle variabili partendo dal modello applicato all'intero dataset di stima verso uno più semplice (*stepwise backward*) e di una procedura di selezione ibrida delle due appena descritte (*stepwise both*).

3.3.2 Applicazione del modello

Si riportano i risultati del modello di regressione lineare applicato a tutte le variabili. In questo modo, si ha una visione generale su quali informazioni siano maggiormente significative per la definizione del prezzo. La scelta di proporre il modello completo è legata anche al fatto che esso è leggermente più efficiente delle sue versioni *stepwise*, benché meno parsimonioso. Nel trarre conclusioni a partire dai modelli proposti, si analizzano le variabili nella significatività e nel segno, poiché il valore assoluto dei coefficienti non risulta essere interpretabile.
Viene riportato un confronto tra due annualità, 2019 e 2020, in quanto si ritiene che esse siano le più indicative per un eventuale cambiamento legato all'emergenza del Covid-19 nella definizione del prezzo. Il dataset relativo all'anno 2021 ha una composizione differente rispetto agli altri due e risulta, pertanto, meno adeguato ad essere confrontato con il dataset che si riferisce ad un periodo antecedente allo scoppio della pandemia. Tuttavia, a partire dal dataset relativo al 2021, si possono trarre considerazioni molto simili a quelle che si evincono dall'anno precedente.

Definizione del prezzo nel 2019

Nell'applicare il modello di regressione lineare al dataset relativo nel 2019, si acquisiscono informazioni su quali fossero le variabili più utili per stimare il prezzo prima dello scoppio della pandemia di Covid-19.
Alcune delle variabili più significative riguardano le caratteristiche dell' "host" che pubblica l'annuncio. In particolare, la variabile "host_response_time" è presente con segno negativo. Dal momento che essa è costruita in modo tale che a valori più elevati corrispondano "host" più celeri nella risposta, si deduce che coloro che rispondono più velocemente tendono ad avere prezzi più contenuti, forse perchè essi hanno un maggior numero di prenotazioni. Inoltre, si osserva che laddove si ha una maggiore conoscenza dell' "host" si tendono ad avere prezzi più elevati, dato il segno positivo di "host_identity_verified" e di "host_is_superhost". Questi ultimi, in particolare, sono gli "host" più attivi e meglio valutati nelle recensioni.
Si osserva, poi, che le caratteristiche geografiche degli alloggi sono molto importanti nella definizione del prezzo. Infatti, "neighbourhood_cleansed" risulta essere una delle variabili più significative del modello e si deduce che una sistemazione a Venezia o nelle isole porta a stimare un prezzo più elevato rispetto a una casa situata fuori da Venezia.
Anche il tipo di alloggio offerto e le sue caratteristiche sono tra le informazioni che maggiormente influiscono sui prezzi degli annunci. Osservando la variabile (fattore) "property_type", si deduce che strutture che offrono alloggio in modo sistematico e professionale comportano prezzi più elevati rispetto alla modalità di riferimento, ossia gli appartamenti. Occorre, però, considerare che le modalità hotel e bed and breakfast compaiono con una frequenza molto bassa. La variabile (fattore) "room_type", inoltre, indica che la stanze, siano esse private o condivise, portano a stimare prezzi più contenuti rispetto alle case intere. Questo mette in luce come venga data particolare importanza alla quantità di spazio messa a disposizione, ma potrebbe essere legato anche al numero di persone a cui si riferisce l'annuncio.
Informazioni più accurate si hanno dai dati che riguardano le caratteristiche dell'alloggio. Tra le variabili più significative del modello, infatti, compaiono "accommodates", "bathrooms" e "bedrooms" con segno positivo. Quindi,

all'aumentare del numero di persone potenzialmente ospitabili, non tutte necessariamente incluse nell'offerta, del numero di bagni e di camere da letto, si hanno prezzi più elevati. Tuttavia, si deduce dal segno negativo di "beds" e "guests_included", che all'aumentare del numero di letti e del numero di persone a cui si riferisce l'annuncio il prezzo tende ad abbassarsi. Una possibile spiegazione è che lo spazio sia concepito in relazione al numero di ospiti a cui è rivolta l'offerta. Spazi più ampi a disposizione di meno inquilini sono valutati di più in termini di prezzo rispetto a spazi più ristretti o proposti a un maggior numero di persone.
Tra le caratteristiche degli annunci, la più significativa è la disponibilità annuale "availability_365", presente con segno positivo. Questo porta a pensare che gli "hosts" che mettono a disposizione gli alloggi per più tempo siano quelli che fissano i prezzi più elevati. Si tratta, probabilmente di persone che utilizzano la loro proprietà al solo scopo di affitarla e non di viverci per una parte dell'anno. Si osservano, infine, "number_of_reviews", "review_scores_value" e "reviews_per_month" con segno negativo e "review_scores_rating" con segno positivo. Questo porta e dire che alloggi maggiormente recensiti, quindi più frequentati, con un rapporto "qualità-prezzo" percepito inferiore portano a stimare prezzi più bassi, mentre sistemazioni in cui l'esperienza complessiva viene valutata meglio tendono ad avere un prezzo più alto. Si veda la Tabella 3.4.

Definizione del prezzo nel 2020

Il dataset relativo all'anno 2020 si riferisce ad un periodo successivo allo scoppio della pandemia di Covid-19. Analizzare le variabili più significative per la stima del prezzo e rilevare le differenze rispetto all'anno precedente è utile per cogliere aspetti legati all'emergenza sanitaria, che possono aver portato a criteri diversi per la definizione del prezzo degli annunci.
Anche in questo caso, tra le variabili più significative compaiono alcune delle caratteristiche degli "hosts". Come per il 2019, si osservano le variabili "host_response_time" e "host_is_superhost" con segno rispettivamente negativo e positivo. Inoltre, compaiono "host_acceptance_rate_perc" con segno negativo e "host_response_rate_perc" con segno positivo. Questo si-

gnifica che coloro che accettano più spesso le richieste di affitto tendono ad avere prezzi più bassi, quindi le persone cercano generalmente sistemazioni più economiche. Gli "hosts" che rispondono più spesso, tuttavia, tendono a fissare prezzi più elevati. Si tratta, probabilmente di coloro per cui l'affitto di una o più proprietà coincide con la principale fonte di reddito.
Come per l'anno precedente, la posizione geografica della sistemazione appare molto rilevante e le case a Venezia e nelle isole sono le più care. Inoltre, si osserva ancora una volta come hotel e bed and breakfast portino a stimare prezzi più elevati degli appartamenti e come le stanze, sia private sia condivise, inducano a prezzi più bassi rispetto alle intere case. Anche in questo caso, comunque, hotel e bed and breakfast sono modalità che compaiono molto poco frequentemente nel dataset.
Per quanto riguarda le caratteristiche delle abitazioni, tra le variabili più significative si osservano "accommodates", "bathrooms" e "bedrooms" con segno positivo, mentre perdono la significatività "beds" e "guests_included". Questa evidenza potrebbe portare a conclusioni rilevanti legate allo scoppio della pandemia. Una possibile spiegazione, infatti, è legata al fatto che per mantenere gli annunci attivi, gli "hosts" abbiano dovuto garantire uno spazio minimo, adeguato al numero di ospiti considerati. Partendo da questo presupposto, è stato possibile valutare di più, in termini di prezzo, lo spazio messo a disposizione in maniera assoluta, non più in rapporto al numero di persone ospitate. Per questo, abitazioni più ampie, ossia con un un numero di individui potenzialmente ammissibile maggiore, con più bagni e più camere da letto portano a stimare prezzi più elevati, indipendentemente dal numero di ospiti considerato.
Le caratteristiche degli annunci che vengono messe in evidenza sono, anche in questo caso "availability_365" con segno positivo, "number_of_reviews" e "reviews_scores_value" con segno negativo.
Tra le variabili più significative compaiono anche "cancellation_policy" con segno negativo e "calculated_host_listings_count" con segno positivo. Questo risultato appare molto interessante poiché è interpretabile alla luce della pandemia di Covid-19 e costituisce una differenza rispetto all'anno precedente. Per quanto riguarda "cancellation_policy", essendo la variabile costruita in modo tale che a valori più elevati corrispondano politiche di cancellazione

più flessibili, si osserva che il prezzo tende ad abbassarsi all'aumentare della flessibilità della cancellazione. Tale evidenza può essere interpretata supponendo che gli "hosts" prevedano cancellazioni delle prenotazioni più frequenti e un calo della domanda. Per questo, tendono a proporre offerte più accattivanti, con prezzi più contenuti e la possibilità di disdire con maggiore semplicità e minore perdita di denaro, sperando di garantirsi un certo numero di prenotazioni. La significatività di "calculated_host_listings_count" indica che "hosts" con un numero maggiore di annunci portano a stimare prezzi più alti. Questo fa pensare che gli "hosts" più attivi su "Airbnb" siano quelli che maggiormente hanno potuto adattarsi alle condizioni imposte dall'emergenza e fissare prezzi più elevati. Si veda la Tabella 3.5.
Le conclusioni che è possibile trarre dal risultato dell'applicazione del modello lineare per l'anno 2021 sono molto simili a quelle esposte per il 2020. Si sottolinea, però, che la composizione del dataset relativo al periodo più recente si discosta da quella degli altri due, che sono quasi uguali.

Conclusioni sul modello di regressione lineare

Nonostante i principali determinanti del prezzo degli annunci siano rimasti invariati prima e dopo lo scoppio della pandemia di Covid-19, la necessità di fronteggiare l'emergenza ha portato ad alcune differenze nella stima del prezzo, da cui si possono trarre le seguenti conclusioni:

- Nel 2020 il prezzo è influenzato dall'ampiezza degli spazi messi a disposizione, che tende ad essere valutata positivamente come nel 2019 e non risente del numero di persone ospitate, probabilmente perchè sono garantiti degli standard minimi di spazio.

- Dopo lo scoppio della pandemia i prezzi risentono delle offerte più economiche con politiche di cancellazione più flessibili, probabilmente pubblicate per incentivare le prenotazioni.

- Gli "hosts" più attivi, che sono forse quelli che hanno potuto adattarsi meglio alla nuova situazione nel 2020, tendono a fissare prezzi più elevati nei loro annunci.

Tabella 3.4: Risultato dell'applicazione del modello lineare nel 2019

Variabile	Coefficiente	Errore std	Valore-t	Pr(>|t|)	Significatività
(Intercept)	3.879	0.1335	29.055	< 2e-16	***
host_response_time	-0.05067	0.01217	-4.164	3.17e-05	***
host_response_rate_perc	0.02516	0.03980	0.632	0.52726	
host_is_superhost	0.08048	0.01226	6.563	5.69e-11	***
host_total_listings_count	0.0001909	0.0001379	1.384	0.16631	
host_has_profile_pic	0.07791	0.1124	0.693	0.48841	
host_identity_verified	0.05871	0.01208	4.859	1.21e-06	***
neighbourhood_cleansed-Venezia e isole	0.6039	0.01448	41.723	< 2e-16	***
is_location_exact	0.01490	0.01101	1.354	0.17594	
property_type-Bed and breakfast	0.2078	0.02699	7.699	1.57e-14	***
property_type-Boat	-0.009333	0.07823	-0.119	0.90505	
property_type-Camping	-0.7881	0.4165	-1.892	0.05854	.
property_type-Hostel	-0.1348	0.1142	-1.180	0.23811	
property_type-Hotel	0.2707	0.03979	6.803	1.12e-11	***
property_type-House	0.01052	0.02050	0.513	0.60784	
property_type-Other	0.2264	0.2411	0.939	0.34780	
room_type-Private room	-0.2526	0.01642	-15.384	< 2e-16	***
room_type-Shared room	-0.7823	0.06463	-12.105	< 2e-16	***
accommodates	0.07692	0.005444	14.129	< 2e-16	***
bathrooms	0.1500	0.01444	10.390	< 2e-16	***
bedrooms	0.1376	0.01020	13.493	< 2e-16	***
beds	-0.03701	0.005525	-6.698	2.29e-11	***
bed_type-Bed	0.07321	0.06335	1.156	0.24789	
guests_included	-0.01810	0.004438	-4.077	4.62e-05	***
minimum_nights	-0.0006818	0.006156	-0.111	0.91182	
maximum_nights	-0.000001772	0.000009854	-0.180	0.85734	
availability_365	0.0004990	0.00004922	10.138	< 2e-16	***
number_of_reviews	-0.0004232	0.00009492	-4.459	8.38e-06	***
review_scores_rating	0.003905	0.0009463	4.127	3.72e-05	***
review_scores_value	-0.05842	0.009452	-6.180	6.80e-10	***
instant_bookable	0.04249	0.01312	3.239	0.00120	**
cancellation_policy	-0.02334	0.007177	-3.253	0.00115	**
calculated_host_listings_count	0.0001650	0.0003122	0.529	0.59714	
reviews_per_month	-0.01704	0.003807	-4.477	7.69e-06	***

Tabella 3.5: Risultato dell'applicazione del modello lineare nel 2020

Variabile	Coefficiente	Errore std	Valore-t	Pr(>|t|)	Significatività
(Intercept)	4.297	0.1566	27.434	< 2e-16	***
host_response_time	-0.06444	0.01248	-5.163	2.50e-07	***
host_response_rate_perc	0.1415	0.03528	4.011	6.11e-05	***
host_acceptance_rate_perc	-0.1512	0.02537	-5.958	2.69e-09	***
host_is_superhost	0.08999	0.01422	6.327	2.67e-10	***
host_total_listings_count	0.00005707	0.0001220	0.468	0.639952	
host_has_profile_pic	-0.06724	0.1303	<-0.516	0.605803	
host_identity_verified	0.03212	0.01359	2.364	0.018131	*
neighbourhood_cleansed-Venezia e isole	0.5140	0.01706	30.130	< 2e-16	***
is_location_exact	0.02080	0.01295	1.607	0.108202	
property_type-Bed and breakfast	0.1188	0.03495	3.399	0.000681	***
property_type-Boat	-0.1878	0.1306	-1.438	0.150577	
property_type-Camping	-0.6902	0.2808	-2.458	0.014002	*
property_type-Hostel	0.01354	0.1392	0.097	0.922524	
property_type-Hotel	0.1866	0.04313	4.327	1.54e-05	***
property_type-House	-0.01766	0.02401	-0.735	0.462075	
property_type-Other	-0.1288	0.1988	-0.648	0.516982	
room_type-Hotel room	-0.03301	0.04288	-0.770	0.441539	
room_type-Private room	-0.2709	0.01926	-14.070	< 2e-16	***
room_type-Shared room	-0.8993	0.07255	-12.395	< 2e-16	***
accommodates	0.05568	0.006392	8.710	< 2e-16	***
bathrooms	0.1966	0.01699	11.572	< 2e-16	***
bedrooms	0.1220	0.01186	10.284	< 2e-16	***
beds	-0.01899	0.006543	-2.902	0.003716	**
bed_type-Bed	0.01902	0.07474	0.255	0.799100	
guests_included	-0.01114	0.005069	-2.198	0.027974	*
minimum_nights	-0.002079	0.006667	-0.312	0.755125	
maximum_nights	-0.00002363	0.00001163	-2.032	0.042155	*
availability_365	0.0003086	0.00004937	6.251	4.34e-10	***
number_of_reviews	-0.0005337	0.00008048	-6.631	3.61e-11	***
review_scores_rating	0.001502	0.001163	1.292	0.196442	
review_scores_value	-0.03896	0.01157	-3.368	0.000761	***
instant_bookable	0.02178	0.01535	1.419	0.155982	
cancellation_policy	-0.03378	0.007913	-4.269	2.00e-05	***
calculated_host_listings_count	0.005489	0.0003738	14.687	< 2e-16	***

3.4 Gradient Boosting

Il modello più indicato per prevedere l'andamento del prezzo risulta essere il *Gradient Boosting*, poichè è uno dei più validi in termini di mean squared error ("MSE") e opera secondo un algoritmo basato sulla progressiva riduzione dell'errore di stima. La proprietà che rende questo modello il migliore tra quelli applicati è la capacità di lavorare su sottoinsiemi dei dati, selezionando i punti del dataset di stima in cui il modello commette un errore più grande e può essere maggiormente migliorato. Attraverso un algoritmo che unisce le procedure di *Gradient Descent* e di *Boosting*, si mediano tanti alberi cresciuti su versioni pesate dei dati di stima.

3.4.1 Cenni teorici sul modello

Si descrivono qui, in sintesi, alcuni aspetti teorici relativi al modello *Gradient Boosting*, tratti da Hastie, Tibshirani & Friedman (2009).
Dato un insieme di dati $(x_1, y_1) \dots (x_n, y_n)$ e una funzione di questi, $y_i = f(x_i)$, che minimizza una funzione di perdita, si migliora la performance del modello con l'aggiunta di un albero di regressione $h(x_i)$, ottenendo la funzione $y_i = f(x_i) + h(x_i)$ e si aggiornano le previsioni. Ad ogni iterazione della procedura si calcolano i residui del modello, $r(x_i) = y_i - f(x_i)$ e su di essi si stima un nuovo albero di regressione, migliorando progressivamente $f(x_i)$. Questo significa che, per M iterazioni della procedura, ad ogni passo m, $1 < m < M$, si ha che $f_{m+1}(x) = f_m(x) + h_m(x)$ e $r_{m+1}(x) = y_{m+1} - f_{m+1}(x)$.
La funzione di perdita minimizzata dall'algoritmo può essere di varia natura e questo è reso possibile dall'utilizzo della procedura di *Gradient Descent*. Con questo metodo, i residui vengono interpretati come gradienti negativi e la minimizzazione della funzione di perdita avviene in direzione opposta al gradiente. Nei modelli utilizzati, si considera la funzione di perdita quadratica $L(y, f(x)) = (y - f(x))^2$ e si minimizzano i residui, ovvero i gradienti negativi, aventi tale forma:

$$-g(x_i) = -\frac{dL(y_i, f(x_i))}{df(x_i)} = y_i - f(x_i)$$

A partire da un insieme di stima $(x_1, y_1) \dots (x_n, y_n)$, da un numero di iterazioni "M" e dalla funzione di perdita $L(y, f(x))$ appena descritta, si implementa l'algoritmo di *Gradient Tree Boosting*. Esso si compone delle seguenti fasi:

- Inizializzazione del modello con un valore costante, il quale, in questo caso, coincide con la media di y, poiché la funzione che si sceglie di minimizzare è la funzione di perdita quadratica.
- A ogni iterazione $1 < m < M$, si calcolano i gradienti negativi $-g(x_i)$, con $i = 1, 2, \dots, n$.
- Si stima un albero di regressione $h_m(x)$ su $-g(x_i)$, assegnando regioni terminaliR_{jm}, dove j indica le funzioni di perdita $j = 1, 2, \dots, J_m$.
- A ogni passo si calcola la minimizzazione della funzione di perdita,

  $$\gamma_{jm} = argmin_\gamma \sum_{x_i \in R_{jm}} L(y_i, f_{m-1}(x_i) + \gamma)$$

  con $j = 1, 2, \dots, J_m$.
- Si aggiorna, quindi, il modello:

  $$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

- Si ottiene il risultato finale: $\hat{f(x)} = f_M(x)$

Per migliorare ulteriormente le performance del modello, si possono introdurre tecniche di regolarizzazione per ridurre il sovradattamento del modello ai dati. In particolare, lo *shrinkage*, modifica l'algoritmo in modo tale che:

$$f_m(x) = f_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \in R_{jm})$$

Il parametro v $(0 < v < 1)$ controlla il tasso di apprendimento della procedura di *boosting* e si ha un *trade-off* tra v e il numero di iterazioni M, dal momento che per valori piccoli di v lo *shrinkage* è maggiore e M aumenta.
Il *Gradient Boosting* consente di utilizzare dati di natura sia quantitativa sia qualitativa, ha una buona capacità previsiva ed è robusto in presenza di valori anomali. Infine, gode di una certa interpretabilità.

3.4.2 Applicazione del modello

Vengono applicate quattro varianti del metodo e, per ogni dataset, si considera quella che porta un errore più contenuto. La prima stima è caratterizzata da 5000 alberi con profondità 1, la seconda da 5000 alberi con profondità 4, la terza da 5000 alberi con profondità 1 e con uno shrinkage di 0.01, che regola il sovradattamento, comportando un maggior numero di iterazioni della procedura. Infine, la quarta stima è caratterizzata da 5000 alberi con profondità 4 e con uno *shrinkage* di 0.01.
Basandosi sulla minimizzazione dell'errore di stima si sceglie la seconda versione del modello. Anche in questo caso, si effettua un confronto dei risultati dell'applicazione del Gradient Boosting ai dataset del 2019 e del 2020, con le stesse motivazioni fornite per il modello di regressione lineare.

Definizione del prezzo nel 2019

L'applicazione del *Gradient Boosting* al dataset relativo al 2019, consente di osservare quali sono le variabili più importanti nella definizione del prezzo in una situazione precedente allo scoppio della pandemia. Si ottiene, infatti, una indicazione dell'influenza relativa delle variabili esplicative sulla variabile risposta, che consente di quantificare, in termini percentuali, quanto esse influiscano sulla stima del prezzo e quali proporzioni ci siano tra loro in termini di rilevanza per rispondere alla domanda di ricerca. In particolare, si osserva che "neighourhood_cleansed", "availability_365" e "reviews_per_month" influiscono di percentuali superiori al 10%, con una notevole differenza rispetto alle altre variabili.
Sono confermate le conclusioni tratte a partire dall'applicazione del modello lineare. In particolare, le prime cinque variabili selezionate in ordine di importanza sono: "neighbourhood_cleansed",
"availability_365",
"reviews_per_month",
"accommodates" e
"bedrooms".
Questo denota che influiscono sul prezzo innanzitutto la posizione geografica

delle abitazioni, la quantità di tempo in cui un alloggio viene messo spesso a disposizione all'interno di un anno, la notorietà dell'"host" e della casa offerta e le caratteristiche spaziali della struttura, quantificate dal numero di persone potenzialmente ospitabili e dal numero di camere da letto presenti. Si vedano la Tabella 3.6 e la Figura 3.1.
Per ottenere informazioni aggiuntive è opportuno approfondire gli effetti marginali delle variabili citate sulla variabile risposta, a partire dai *Partial Dependence Plots*. Essi riportano l'effetto di una variabile sulla risposta, considerando, tuttavia, la presenza di tutte le informazioni nel modello. Anche in questo caso, come in tutte le applicazioni dei modelli, ci si riferisce ad una trasformazione logaritmica dei prezzi degli annunci. Si osserva che la modalità "Venezia e isole" della variabile "neighbourhood_cleansed" influisce sui prezzi più elevati, mentre la modalità "fuori Venezia" sui prezzi più bassi. Si nota, inoltre, che all'aumentare della disponibilità delle strutture, definita da "availability_365", il prezzo tende a crescere. "Reviews_per_month", invece, ha un effetto non costante sulla variabile risposta, la quale tende a stabilizzarsi su un valore intermedio intorno alle dieci recensioni al mese. Per quanto riguarda la variabile "accommodates", si osserva che essa comporta una crescita del prezzo, che si stabilizza su un valore massimo quando gli alloggi possono ospitare tra le dieci e le quindici persone. Infine, si osserva che anche "bedrooms" porta il prezzo verso un aumento progressivo ed esso tende a stabilizzarsi su un valore elevato quando sono presenti almeno cinque o sei letti nella struttura, mentre "number_of_reviews" ha un effetto decrescente e la risposta si stabilizza intorno alle 400 unità. Si veda le Figura 3.2.

Definizione del prezzo nel 2020

L'applicazione del Gradient Boosting al dataset del 2020 permette di osservare quali sono le variabili principali nella stima del prezzo degli annunci in un contesto globale segnato dalla pandemia di Covid-19. Considerando le differenze con quanto emerso dalle analisi effettuate per l'anno precedente, è possibile capire quali variabili hanno acquisito importanza nella definizione del prezzo e di quali ne hanno persa.

Si osservano, innanzitutto, i risultati riferiti all'influenza relativa delle variabili. Si nota che "calculated_hosts_listings_count" e "availability_365" influiscono con percentuali superiori al 10%. Tali proporzioni tendono, poi, a calare per le altre informazioni. Le prime cinque variabili in ordine di importanza, sono:
"calculated_hosts_listings_count",
"availability_365",
"number_of_reviews",
"neighbourhood_cleansed" e
"room_type".
Nonostante vi siano delle differenze rispetto al 2019, si ha una conferma della rilevanza della posizione geografica, della disponibilità annuale degli alloggi e della notorietà della sistemazione, quantificata dal numero di recensioni. Tuttavia, la variabile più importante risulta essere
"calculated_hosts_listings_count", il che porta a dedurre che gli "hosts" con un maggior numero di annunci siano quelli che hanno saputo adattarsi meglio alle condizioni imposte dall'emergenza e hanno maggiore possibilità di influire sulla definizione dei prezzi. Acquisisce importanza, poi, la tipologia di alloggio, rappresentata dalla variabile "room_type". Questo interessante risultato lascia intendere una maggiore differenza di prezzo in base alla tipologia di struttura, che può essere un'intera casa o appartamento, una stanza di hotel, una stanza privata o una stanza condivisa. Si vedano la Tabella 3.7 e la Figura 3.3.
Si analizzano, quindi, gli effetti marginali delle prime sei variabili esplicative sulla variabile risposta, tenendo in considerazione anche la presenza delle altre variabili nel modello. Si osservano, pertanto, i *Partial Dependence Plots*. Si nota che "calculated_hosts_listings_count" comporta prezzi crescenti, fino al valore di circa sessanta annunci. A questo punto, il prezzo si stabilizza su un valore elevato. "Availability_365" e "number_of_reviews" hanno effetto non costante e difficilmente interpretabile sulla risposta. Invece, si può constatare che "neighbourhood_cleansed" ha lo stesso effetto nelle due annualità considerate. Infine, per quanto riguarda la variabile "room_type", si osserva che intere abitazioni e stanze di hotel comportano prezzi elevati, camere private portano a prezzi intermedi e camere condivise a prezzi molto

bassi, mentre "host_total_listings_count" ha un effetto discontinuo e la risposta si stabilizza prima delle 500 osservazioni. Si veda la Figura 3.4.
I risultati dell'applicazione del *Gradient Boosting* al dataset relativo al 2021, confermano quanto detto per il 2020, malgrado la differenza della composizione dei due set di dati.

Conclusioni sul modello *Gradient Boosting*

Dall'applicazione del modello *Gradient Boosting* ai dataset relativi al 2019 e al 2020, è possibile trarre le seguenti conclusioni:

- La zona di ubicazione degli alloggi, la loro disponibilità durante un anno e la notorietà dell'"host" sono rimaste delle determinanti importanti per il prezzo prima e dopo lo scoppio della pandemia di Covid-19.
- Perdono, invece, importanza le caratteristiche dell'alloggio.
- Nella situazione di crisi, ha ottenuto molta rilevanza il numero di annunci pubblicati dagli "hosts". Questo significa che chi meglio si è adattato alle condizioni imposte dall'emergenza con la sua offerta, ha possibilità maggiore di fissare i prezzi.
- Acquisisce, infine, importanza, la caratteristica di affittare una intera abitazione o una camera in hotel, rispetto ad una stanza privata o condivisa.

3.5 Conclusioni

- Il miglior adattamento dei modelli al dataset relativo al 2019, rispetto agli anni successivi è dovuto a caratteristiche intrinseche dei dati. Esso potrebbe essere legato alla presenza di valori più diversificati per le variabili, di un maggior numero di osservazioni mancanti, di più frequenti errori di inserimento delle informazioni o di altre peculiarità. Il peggiore riscontro negli anni più recenti è indice di un'offerta più disomogenea e con più anomalie e può essere attribuito alle condizioni difficili determinate dall'emergenza del Covid-19.

Tabella 3.6: Influenza relativa delle variabili nel modello *Gradient Boosting* - anno 2019

Variabile	Influenza relativa (%)
neighbourhood_cleansed	11.279210906
availability_365	10.989366531
reviews_per_month	10.242799043
accommodates	6.713426361
bedrooms	6.141382065
number_of_reviews	5.839021870
maximum_nights	5.666472526
room_type	5.515686412
host_total_listings_count	5.014865295
calculated_host_listings_count	4.576513575
review_scores_rating	4.136711906
host_response_rate_perc	3.472990302
minimum_nights	3.095569892
property_type	2.789620828
guests_included	2.724005365
bathrooms	2.612601723
cancellation_policy	2.229592176
beds	2.057318458
host_response_time	1.216939749
review_scores_value	1.175435211
instant_bookable	0.712268124
is_location_exact	0.622979587
host_identity_verified	0.612003478
host_is_superhost	0.528420956
bed_type	0.030526027
host_has_profile_pic	0.004271634

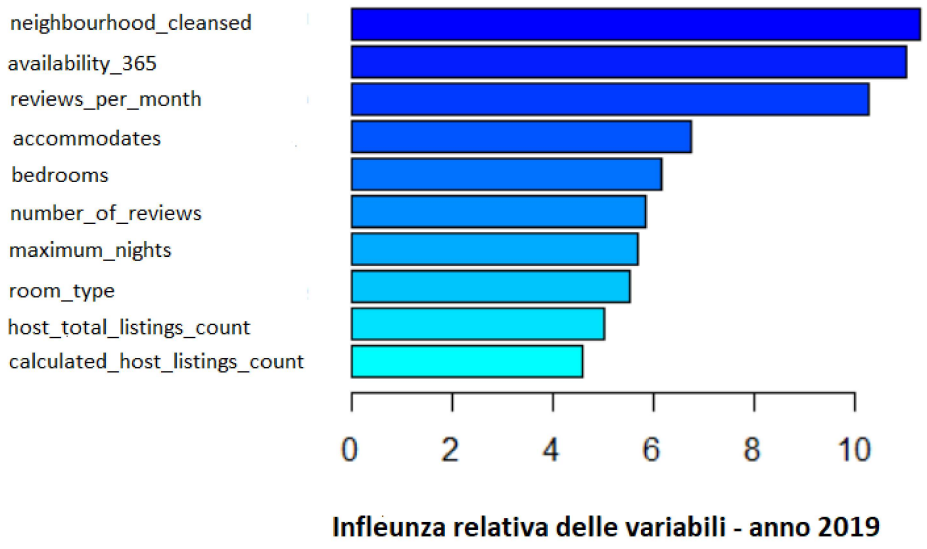

Figura 3.1: Influenza relativa delle variabili per la stima del prezzo - anno 2019

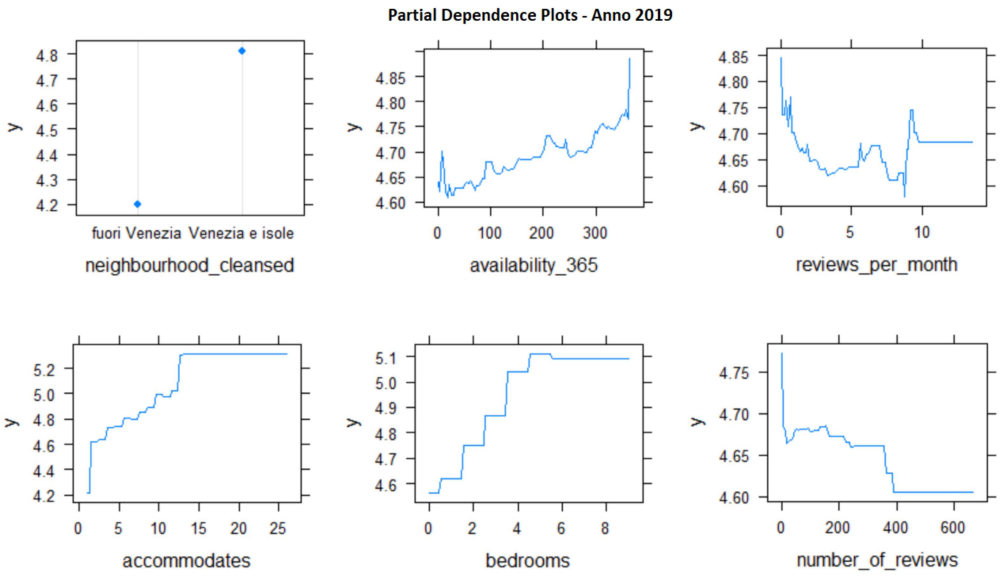

Figura 3.2: Partial Dependence Plots - anno 2019

Tabella 3.7: Influenza relativa delle variabili nel modello *Gradient Boosting* - anno 2020

Variabile	Influenza relativa (%)
calculated_host_listings_count	10.233847186
availability_365	10.095753132
number_of_reviews	8.516947352
neighbourhood_cleansed	7.867969363
room_type	6.758369003
host_total_listings_count	6.322240205
bedrooms	5.275474838
accommodates	5.214474787
host_acceptance_rate_perc	5.145258708
maximum_nights	4.868022980
review_scores_rating	4.235501129
property_type	3.967748867
bathrooms	3.460667621
cancellation_policy	2.978931018
minimum_nights	2.749279632
guests_included	2.185883021
beds	2.138330796
host_response_rate_perc	2.073842336
host_response_time	1.586848412
review_scores_value	1.563480847
is_location_exact	0.799090977
instant_bookable	0.668244744
host_identity_verified	0.637939040
host_is_superhost	0.601985197
bed_type	0.052145707
host_has_profile_pic	0.001723101

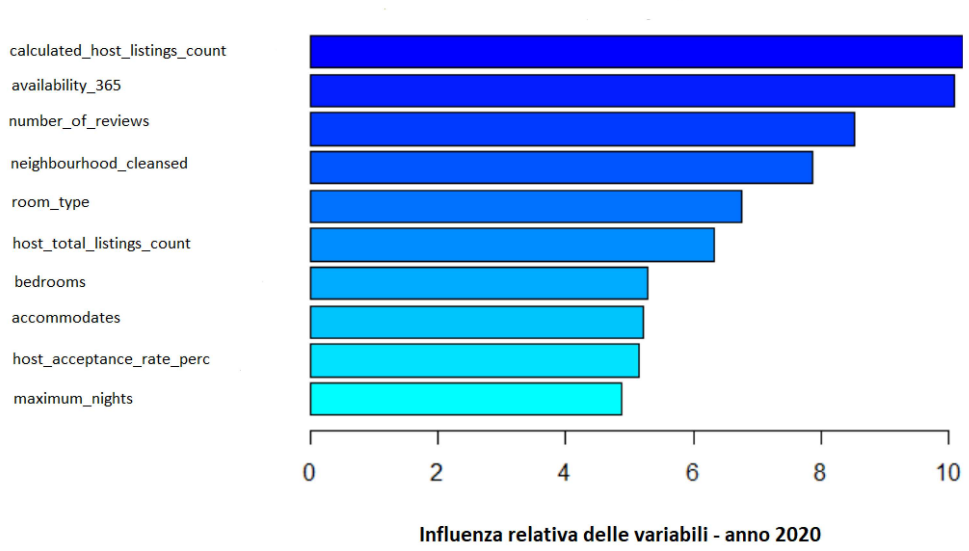

Figura 3.3: Importanza relativa delle variabili per la stima del prezzo - anno 2020

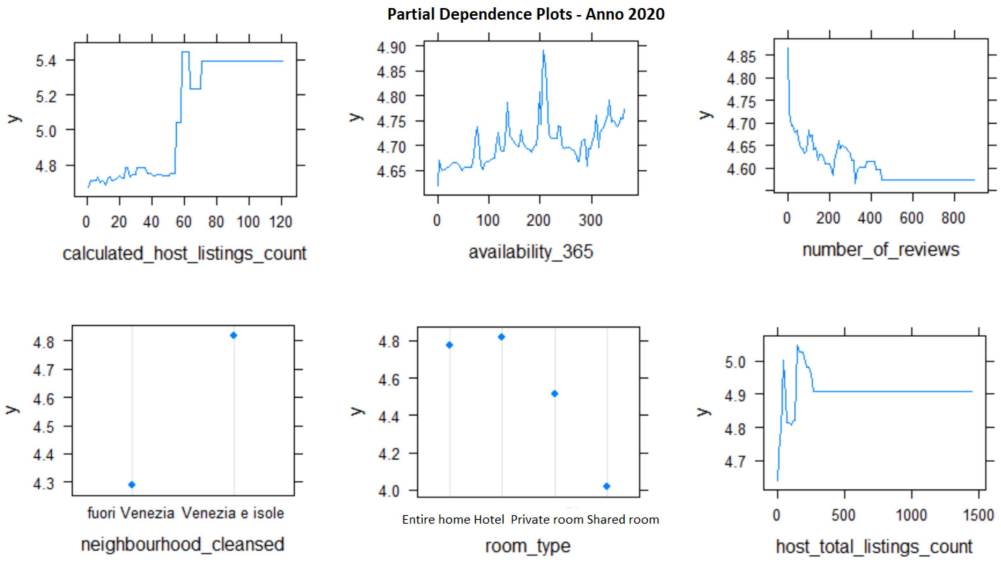

Figura 3.4: Partial Dependence Plots - anno 2020

- Alcune delle variabili principali per la stima del prezzo, come il posizionamento geografico degli alloggi, la disponibilità delle strutture nell'arco di un anno e la notorietà dell'"host", restano invariate prima e dopo lo scoppio della pandemia di Covid-19.

- L'ampiezza degli ambienti messi a disposizione, quantificata in termini di persone potenzialmente ospitabili e di numero di stanze presenti, è una caratteristica importante nella definizione del prezzo. Tuttavia, nel 2019 essa viene considerata solo in relazione al numero di individui indicato nell'annuncio, mentre nel 2020 non si osserva particolarmente questa informazione, forse perchè sono garantiti degli standard minimi di spazio per ogni persona. Inoltre, nel 2020, vi sono altri aspetti degli annunci ritenuti più importanti.

- Nel 2020 acquisisce maggiore importanza nella stima del prezzo la tipologia di alloggio affittato. Perciò, intere abitazioni o camere di hotel, queste ultime molto poco frequenti, vengono valutate maggiormente di stanze private o condivise.

- Dopo lo scoppio della pandemia influiscono maggiormente sul prezzo le politiche di cancellazione delle prenotazioni più flessibili, contenute in offerte più economiche.

- In uno stato di crisi acquisisce notevole importanza il numero di annunci pubblicato dagli "hosts". Questo, probabilmente, è legato al fatto che coloro che sono molto attivi su "Airbnb" hanno saputo adattarsi meglio alle condizioni imposte dalla situazione di emergenza, il che ha concesso loro di stabilire prezzi più elevati per le proprie offerte.

Capitolo 4

La posizione delle case su Airbnb: un'analisi

Dal momento che i modelli di regressione evidenziano una elevata significatività della variabile relativa alla zona di collocazione degli edifici in tutte le annualità considerate, si desidera approfondire quanto le caratteristiche geografiche influiscano sulla stima del prezzo. Nel precedente capitolo, si considera solamente se la sistemazione è situata a Venezia o nelle isole, oppure fuori da Venezia. Si desidera, ora, avvalersi di informazioni più dettagliate e si propone una modellazione con indicatori dei quartieri in cui si collocano le abitazioni e "availability_365", "accommodates", "number_of_reviews" e "room_type" come variabili esplicative.

4.1 Preparazione del dataset

Per quanto riguarda la composizione dei dataset sono state effettuate scelte diverse rispetto a quanto fatto nel capitolo precedente. Infatti, vengono mantenute le variabili qualitative originali, che forniscono informazioni sulle zone in cui sono situate le abitazioni offerte negli annunci. Non sono, inoltre, imputati eventuali dati mancanti a tali variabili.

Viene scelto di analizzare i risultati ottenuti con la variabile che si ritiene porti a conclusioni più interessanti.
Le informazioni disponibili sono:

- "neighbourhood": indicazione del quartiere dove sono situati gli alloggi. Nel caso dei dataset relativi al 2019 e al 2020, si tratta di un fattore con 9 livelli, corrispondenti ai sestieri di Venezia (San Marco, San Polo, Santa Croce, Cannaregio, Dorsoduro, Castello) e ad altre tre modalità (Giudecca, Lido, Murano). Nel 2021, invece, questa informazione è un fattore con 120 livelli, le cui modalità sono spesso incerte. In tutti e tre i casi, la variabile presenta una quantità non indifferente di valori mancanti.

- "neighbourhood_cleansed": indicazione della zona in cui si collocano le abitazioni. Oltre ai sestieri, sono presenti informazioni anche per gli edifici situati sulle isole o nei quartieri e comuni collocati fuori da Venezia. La variabile non presenta valori mancanti per nessuna annualità ed è rappresentata da un fattore con 62 livelli nei dataset relativi al 2019 e al 2020 e con 58 livelli in quello relativo al 2021.

- "neighbourhood_group_cleansed": fattore con due livelli che indica se l'alloggio è situato sulla terraferma o su un'isola; non presenta valori mancanti in nessun dataset.

Queste variabili vengono descritte per completezza. Tuttavia, si studiano solo i modelli lineari gerarchici con "neighbourhood_cleansed" come indicatore della zona geografica.
Per quanto riguarda le rimanenti informazioni, le operazioni di preparazione del dataset sono analoghe a quelle effettuate per l'applicazione dei modelli di regressione. Tuttavia, in questo caso, si è ritenuto opportuno selezionare solo alcune specifiche variabili, che fossero particolarmente rilevanti per la stima del prezzo nelle diverse annualità. Pertanto, oltre alle informazioni riguardanti la collocazione degli alloggi, le variabili utilizzate sono le seguenti:

- "price_dollars": variabile quantitativa continua, che corrisponde al prezzo per notte definito negli annunci. La trasformazione logaritmica viene usata come risposta dei modelli proposti in questo capitolo.

- "availability_365": variabile quantitativa continua, che riporta il numero di giorni in un anno in cui la sistemazione considerata nell'annuncio è resa disponibile dall'"host". Dai modelli di regressione emerge che in tutte le annualità questo aspetto è tra i più rilevanti per prevedere il prezzo e la variabile viene utilizzata come predittore nei modelli proposti in questo capitolo.

- "accommodates": variabile quantitativa discreta, che costituisce un'informazione riguardo alle caratteristiche dell'alloggio. In particolare, essa corrisponde al numero di individui potenzialmente ospitabili nelle sistemazioni e, pertato, dà indicazioni sulla spaziosità delle strutture. Emerge dalla regressione che si tratta di un aspetto importante nello stimare il prezzo e in questo contesto la variabile viene utilizzata come predittore.

- "number_of_reviews": variabile quantitativa discreta che rappresenta il numero di recensioni ricevute dall'alloggio. Dal momento che essa costituisce una indicazione della notorietà dell'"host" e che la regressione fa emergere la sua rilevanza nella stima del prezzo, essa viene inserita come predittore nei modelli proposti in questo capitolo.

- "room_type": variabile qualitativa sconnessa che indica la tipologia di struttura a cui ci si riferisce nell'annuncio. Essa è presente con modalità "Entire home/ appartment", "Private room" e "Shared room" nel dataset relativo 2019. Negli anni successivi compare anche la modalità "Hotel room", che è molto poco frequente e non risulta mai significativa nei modelli. Tuttavia, questa evidenza potrebbe indicare l'ingresso degli hotel nel business di "Airbnb" per sopperire ad una perdita di clienti dovuta alla condizione di crisi legata al Covid-19. La variabile "room_type" è rilevante per la stima del prezzo secondo i modelli di regressione. Pertanto, essa viene utilizzata come predittore nei modelli proposti in questo capitolo.

Per ogni annualità, sia nel caso dei modelli lineari gerarchici, sia in quello dei modelli di regressione lineare con cui questi sono confrontati, si applicano i modelli all'intero dataset. I modelli non sono, infatti, implementati a fini

previsivi, ma per confrontare le diverse possibilità di stima e per analizzare le caratteristiche dei dataset a cui essi sono applicati. I modelli vengono confrontati sulla base dell'indicatore *AIC*, Azzalini & Scarpa (2009).

4.2 Modelli

Si sceglie di analizzare i risultati ottenuti con "neighbourhood_cleansed" in qualità di variabile indicatrice dei gruppi perchè si ritiene che essa possa fornire indicazioni più precise. Tra le variabili riguardanti l'ubicazione degli alloggi, infatti, essa appare la più completa in termini di numero di modalità e assenza di dati mancanti e gode della proprietà di essere confrontabile in tutte le annualità considerate. Si studiano i risultati dello studio dei modelli con "neighbourhood_cleansed" come indicatore della zona e "availability_365", "accommodates", "number_of_reviews" e "room_type" come variabili esplicative. Essi sono applicati ai dataset relativi al 2019 e al 2020. Lo scopo di tali analisi è di avere un confronto della situazione osservata prima e dopo lo scoppio della pandemia di Covid-19.

4.2.1 Cenni teorici sui modelli lineari gerarchici

Si descrivono qui, in sintesi, alcuni aspetti teorici relativi ai modelli gerarchici, tratti da Gelman & Hill (2006).
I modelli lineari gerarchici sono estensioni del modello di regressione. Essi vengono applicati quando i dati sono strutturati in gruppi e i coefficienti possono variare tra tali gruppi. In alcuni casi limite, questa metodologia e la regressione arrivano a coincidere. Nei modelli lineari gerarchici si considerano un campione di n osservazioni $i = 1, ..., n$, una variabile risposta $y = (y_1, ..., y_n)$, uno o più predittori lineari contenuti in una matrice X, di dimensioni (nxk), tale che $\hat{y}_i = X_i\beta$ sia la previsione per l'unità i e J gruppi, $j = 1, ..., J$. Questi ultimi sono identificati da una variabile categoriale, un fattore con J livelli, corrispondenti a predittori nei modelli di regressione. Questa variabile è tale che $j[i]$ soano gli indici di appartenenza ai gruppi. I J coefficienti sono a loro volta modellabili.
Si distinguono tre possibilità: un modello a intercetta variabile $y_i = \alpha_{j[i]} +$

$\beta x_i + \varepsilon_i$, un modello a pendenza variabile $y_i = \alpha + \beta_{j[i]} x_i + \varepsilon_i$ e un modello a pendenza e intercetta variabili $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \varepsilon_i$. La pendenza variabile esplicita un'interazione tra il predittore lineare e l'indicatore di gruppo. Si possono distinguere gli effetti fissi e gli effetti casuali degli elementi variabili. I modelli lineari gerarchici sono utili per considerare le variazioni a livello individuale e di gruppo nella stima dei coefficienti di regressione nei gruppi. Sono, inoltre, adoperati per fare previsioni per nuovi gruppi e per stimare coefficienti di regressione per gruppi piccoli o con caratteristiche particolari. Questi modelli costituiscono un compromesso tra una situazione di *complete pooling*, $y_i = \alpha + \beta x_i + \varepsilon_i$, in cui la variabile categoriale non viene inserita nel modello e tutte le unità si considerano appartenenti allo stesso gruppo e una situazione di *no pooling*, $y_i = \alpha_{j[i]} + \beta x_i + \varepsilon_i$. Col metodo *no pooling* si stimano modelli separati per ogni livello della variabile categoriale, sovrastimando, così, la variabilità dei singoli gruppi. Nel caso del *complete pooling*, invece, non vengono considerate in nessun modo le variazioni tra i gruppi.

Pooling parziale con predittori

Si tratta di modelli in cui, nel caso più semplice, y segue una distribuzione normale $N(\alpha_{j[i]} + \beta x_i, \sigma_y^2)$, con $i = 1, ..., n$ e α_j segue una distribuzione normale $N(\mu_\alpha, \sigma_\alpha^2)$, con $j = 1, ..., J$. Si indicano con n_j il numero di unità all'interno del gruppo j, con σ_y^2 la varianza all'interno del gruppo, assunta costante e con σ_α^2 la varianza tra i gruppi. Per quanto riguarda σ_α, si individuano due casi limite. Infatti, se esso tende a ∞, si ha il modello *no pooling*, mentre se tende a zero si ottiene il modello *pooling completo*. La stima di α_j approssima la media pesata della stima *no pooling*, ossia $(\bar{y}_j - \beta \bar{x}_j)$, e della media μ_α. Alternativamente, il modello può essere stimato per ciascun gruppo e in ogni gruppo possono variare l'intercetta, il coefficiente angolare o entrambi.

4.2.2 Applicazione dei modelli lineari gerarchici - anno 2019

Come per le analisi di regressione, si desidera verificare, innanzitutto, le caratteristiche degli annunci nel mese di giugno 2019, in una situazione precedente allo scoppio della pandemia di Covid-19.

Modello *partial pooling* con intercetta variabile e quattro predittori

Si applica un modello con "availability_365", "accommodates", "number_of_reviews" e "room_type" come predittori e con intercetta variabile. L'intercetta fornisce un'indicazione del quartiere e gli effetti casuali stimati per essa esplicitano come ogni località presente nel dataset influenzi diversamente il prezzo degli alloggi.
Tutte le stime dei coefficienti sono significative. Ne emerge che le variabili inserite nel modello come esplicative sono, effettivamente, utili per determinare l'andamento del prezzo. Poiché il modello opera su scala logaritmica, non si può dare una interpretazione in termini quantitativi ai valori dei coefficienti. Tuttavia, osservando il segno degli effetti fissi, si deduce che una disponibilità annuale maggiore comporta un aumento del prezzo e lo stesso vale per alloggi più spaziosi, che potrebbero potenzialmente ospitare più persone. Si nota, invece, che un maggior numero di revisioni comporta un abbassamento nella stima del prezzo. Quindi, gli alloggi più recensiti sono quelli più economici, probabilmente perchè essi vengono prenotati più di frequente. Inoltre, se un alloggio è costituito da una stanza, privata o condivisa, il prezzo è inferiore rispetto a quello di una intera casa o appartamento. La modalità di riferimento della variabile "room_type", a cui si riferiscono gli altri valori delle stime, infatti, è "Entire home/appartment". Per costruzione, i coefficienti sono costanti in tutte le zone per le variabili esplicative, mentre per l'intercetta variano da quartiere a quartiere, assumendo valori superiori a 3, fino a valori di poco superiori a 4. Per quanto riguarda gli effetti casuali relativi al quartiere, infine, si nota che essi possono essere positivi o negativi e in valore assoluto di poco inferiori allo zero. Un elemento interessante è che i coefficienti casuali relativi ai sestieri e ad alcune altre località della città e delle isole di Venezia sono maggiori di zero. Questo conferma che la collocazione degli alloggi a Venezia e sulle isole influisce positivamente sulla stima del prezzo, mentre le osservazioni che influiscono negativamente riguardano abitazioni situate, per lo più, fuori da Venezia. Si vedano le Tabelle 4.3 e 4.8. Il modello presenta un AIC pari a 9979.479. Si decide di effettuare un confronto con il modello di regressione lineare sulla base di questo indicatore. In generale, sipreferiscono modelli in cui il valore dell'AIC è più basso, poichè

esso rappresenta un indice di complessità.

Confronto del modello lineare gerarchico con un modello di regressione lineare, sulla base dell'*AIC*

Si applica ai dati un modello di regressione lineare con intercetta per la trasformazione logaritmica del prezzo. L'obiettivo del confronto proposto è quello di capire se, in termini di efficienza, un modello lineare gerarchico è effettivamente valido, oppure se è più opportuno considerare ogni quartiere come un'entità a sè stante per la stima del prezzo. Le variabili esplicative utilizzate nel modello di regressione lineare sono "availability_365", "accommodates", "number_of_reviews", "room_type" e "neighbourhood_cleansed". Quest'ultima è la variabile originariamente presente nei dataset e non trasformata. Essa è costituita da un fattore con 62 livelli, indicanti le specifiche zone di ubicazione degli edifici (la modalità di riferimento per le stime dei coefficienti è "Aeroporto"). Il modello di regressione lineare stimato presenta un AIC pari a 9820.397. Si osserva che le varibili "availability_365", "accommodates", "number_of_reviews", "room_type" sono altamente significative per la definizione del prezzo e permettono di trarre conclusioni analoghe a quelle rilevate a partire dal modello lineare gerarchico. Per quanto riguarda, invece, l'indicazione del quartiere in cui l'edificio è situato, i risultati possono essere sia positivi sia negativi e raramente sono significativi. I casi il cui lo sono generalmente rappresentano delle eccezioni. Ne è un esempio l'Isola di Santa Cristina in cui è presente una sola abitazione. La località, quindi, influenza completamente la variabile risposta, in quanto è definito un unico prezzo. Una situazione simile riguarda anche altri quartieri con pochissime osservazioni, come Ca′ Sabbioni e Chirignago. Confrontando il modello di regressione lineare descritto con il modello lineare gerarchico in termini di AIC, il modello lineare gerarchico sembra essere meno efficiente di quello di regressione lineare, il quale, strutturalmente, è meno complesso. Il confronto dei modelli basato sull'AIC porta a concludere che ogni zona può essere considerata come un'entità a sè stante nella definizione del prezzo. Tuttavia, si potrebbero ottenere risultati poco informativi. A livello interpretativo, infatti, si osserva una scarsa significatività dei coefficienti

del modello di regressione lineare. Si vedano le Tabelle 4.7 e 4.5.

4.2.3 Applicazione dei modelli lineari gerarchici - anno 2020

Si applicano i modelli lineari gerarchici al dataset relativo all'anno 2020. In questo modo è possibile analizzare le caratteristiche delgi annunci in una situazione successiva allo scoppio della pandemia di Covid-19 e osservare le differenze rispetto l'anno precedente.

Modello *partial pooling* con intercetta variabile e quattro predittori

Si applica un modello con "availability_365", "accommodates", "number_of_reviews" e "room_type" come predittori e con intercetta relativa al quartiere variabile. Rispetto all'anno precedente, si osserva che gli effetti fissi stimati delle variabili esplicative e dell'intercetta presentano gli stessi segni. Tuttavia, si nota esserci una modalità in più relativamente alla variabile "room_type", ovvero "Hotel room", la quale, però, non risulta essere significativa. Essa indica che una camera in hotel porta a stimare un prezzo più elevato rispetto ad un'intera casa o appartamento. Si ricorda, in ogni caso, che tale modalità è molto poco frequente all'interno del dataset e, pertanto, non si ritiene di poter trarre conclusioni utili dal risultato emerso. Si può, tuttavia, dedurre che gli hotel entrino nel buisness di "Airbnb" in un contesto di crisi, legato al Covid-19, probabilmente in seguito ad una perdita considerevole di clienti. Nel modello lineare gerarchico, i coefficienti stimati dell'intercetta variano da zona a zona, mentre i coefficienti dei predittori sono costanti per tutti i quartieri. Gli effetti casuali dell'intercetta, infine, hanno le stesse caratteristiche osservate nel 2019. Questo modello, data l'elevata significatività delle variabili esplicative, conferma che, anche nel 2020, esse costituiscono elementi importanti per la stima del prezzo. Solamente la significatività della variabile relativa alla disponibilità annuale degli alloggi subisce un abbassamento. Il risultato è, probabilmente, legato al fatto che le persone tendono a prenotare soggiorni più brevi, in seguito allo scoppio della pandemia, probabilmente per motivazioni legate ai contagi o per avere perdite di denaro inferiori con un'eventuale cancellazione improvvisa. Si veda la Tabella 4.4. Il valore dell'AIC del modello è pari a 13095.81.

Confronto del modello lineare gerarchico con un modello di regressione lineare, sulla base dell'*AIC*

Come per il dataset relativo all'anno 2019, anche per il 2020, si effettua un confronto tra le applicazioni di un modello di regressione lineare con intercetta ed il modello lineare gerarchico con intercetta variabile e quattro esplicative. Le variabili scelte per le analisi sono le stesse utilizzate per l'anno precedente. Il valore dell'AIC per il modello di regressione lineare è pari a 12976.2. Anche in questo caso, si osserva che le variabili "availability_365", "accommodates", "number_of_reviews", "room_type" e "neighbourhood_cleansed" sono altamente significative, malgrado per la disponibilità annuale la significatività sia più bassa rispetto all'anno precedente. Tale risultato conferma che, dopo lo scoppio della pandemia di Covid-19, le persone tendono a prenotare per periodi più brevi. Si osserva, inoltre, che la modalità corrispondente alla camera in hotel per il tipo di stanza affittata è molto poco frequente nel dataset e la stima del coefficiente corrispondente non è significativa. Anche nel 2020, i coefficienti relativi ai quartieri sono quasi sempre non significativi. In tutti i casi, quindi, segno delle stime conduce alle medesime osservazioni ottenute per il 2019. Infine, osservando l'AIC, sembrerebbe essere migliore il modello di regressione lineare, anche se dal modello lineare gerarchico si possono ottenere risultati interpretativi validi. Quindi, il confronto dei modelli basato sull'AIC porta a concludere che ogni zona può essere considerata come un'entità a sè stante per la stima del prezzo. Tuttavia, data la scarsa significatività dei coefficienti del modello lineare, è probabile che si ottengano risultati poco informativi.

L'applicazione dei modelli proposti al dataset relativo all'anno 2021 porta a risultati simili rispetto al dataset dell'anno 2020, con i valori delle stime che cambiano di poco. Le conclusioni che si possono trarre relativamente agli anni successivi allo scoppio della pandemia di Covid-19 sono analoghe per i due periodi considerati. Si veda la Tabella 4.6

4.2.4 Confronto tra un modello lineare gerarchico e un modello di regressione lineare nei sestieri di Venezia

Si propone, infine, un confronto tra un modello di regressione lineare con una sola variabile esplicativa (detto anche modello *complete pooling*) e un modello lineare gerarchico con intercetta casuale e un predittore (lo stesso scelto per il modello lineare). Queste restrizioni consentono di ottenere una rappresentazione grafica in due dimensioni. La variabile risposta dei modelli è la trasformazione logaritmica del prezzo. In entrambi i casi la variabile esplicativa inserita è "number_of_reviews". La scelta del numero di revisioni come predittore è dettata dal fatto che si tratta di una informazione significativa in tutte le annualità e che essa influisce in modo negativo sulla variabile risposta. Quindi, più un alloggio è conosciuto, più il suo prezzo tende ad essere basso. Questa evidenza è in linea con il business di "Airbnb", in cui gli utenti cercano soluzioni che consentano, quanto più possibile, di risparmiare denaro. Si propongono, di seguito, i risultati ottenuti per il 2019 e il 2020, relativamente ai sestieri di Venezia. Si ricorda che tali risultati si riferiscono al confronto tra un modello di regressione lineare e un modello lineare gerarchico in cui l'indicatore della zona è "neighbourhood_cleansed", che non si riferisce, pertanto, solamente ai sestieri.
Si ottengono risultati molto simili per il 2021. Tuttavia, a causa della composizione leggermente diversa di quest'ultimo dataset, appare più efficace effettuare il confronto delle altre due annualità.
Si osserva che, in seguito allo scoppio della pandemia di Covid-19, i prezzi tendono, in generale, ad abbassarsi. Si nota, poi, come il modello di regressione lineare, rappresentato in blu, sottostimi in tutti i casi il modello lineare gerarchico, indicato in rosso. Per la maggior parte dei sestieri, le rette sono molto vicine tra loro, quasi sovrapposte. Il risultato suggerisce che il modello lineare gerarchico può essere utile per la definizione del prezzo. Infatti, è molto importante tenere in considerazione la posizione degli edifici. Questo si evince effettuando un confronto dei modelli sulla base dell'AIC. Tale indicatore suggerisce che in tutti e due gli anni sia preferibile il modello lineare gerarchico. Se ne conclude che la zona in cui l'edificio è collocato sia importante per la stima del prezzo. Si vedano le Tabelle 4.1 e 4.2.

Per quanto riguarda l'anno 2019, osservando i grafici, si può notare come le case si concentrino nei sestieri di Cannaregio, Castello e Dorsoduro. La distanza tra le due rette, inoltre, è diversa nei quartieri. In particolare, essa è molto ridotta a Cannaregio, Castello e Santa Croce, mentre è più marcata a Dorsoduro, San Marco e San Polo. In questi ultimi tre il modello lineare gerarchico sembra essere più utile perchè si deduce che i prezzi degli alloggi situati in queste zone risentano molto della posizione degli edifici. In generale, si osserva che a San Marco i prezzi sono decisamente più elevati che negli altri sestieri. Per quanto riguarda la pendenza delle rette, infine, si ha conferma che gli annunci più recensiti si riferiscono a sistemazioni più economiche. Si suppone che vi siano alloggi poco recensiti molto più costosi di altri, che ricevono numerose recensioni. Inoltre, si osserva che a San Polo vi è una minore concentrazione di case.
Nell'anno 2020 si può notare che vi sono meno case disponibili rispetto al 2019 a Cannaregio, Catello e Dorsoduro. Invece, a San Marco, il numero di annunci sembra essere aumentato. In questo sestiere, inoltre, i prezzi sono considerevolmente più bassi dell'anno precedente e questo si riscontra, in misura minore, anche a Santa Croce. Dopo lo scoppio della pandemia, inoltre, la distanza tra le rette sembra essere pressocché invariata e si accentua lievemente solo a San Marco. Dopo l'inizio dell'emergenza, infine, la rilevanza della notorietà degli alloggi è aumentata nella stima del prezzo.
Per concludere, si può affermare che il confronto proposto conferma l'importanza per la definizione del prezzo della zona in cui sono situate le abitazioni, sia prima sia dopo lo scoppio della pandemia di Covid-19. Infatti, quartieri diversi hanno caratteristiche differenti per quanto riguarda la numerosità degli annunci, i prezzi delle case e la loro dipendenza da altre variabili. Tali caratteristiche cambiano in modo diverso da zona a zona prima e dopo l'inizio della situazione di emergenza. Si vedano le Figure 4.1 e 4.2.

4.3 Conclusioni

Dall'analisi effettuata con modelli lineari gerarchici si possono trarre le seguenti conclusioni.

Tabella 4.1: Confronto tra un modello di *complete pooling* e un modello lineare gerarchico sulla base dell'*AIC* - anno 2019

Modello	AIC
Modello di regressione lineare (*complete pooling*)	15307.95
Modello lineare gerarchico	12830.59

Tabella 4.2: Confronto tra un modello di *complete pooling* e un modello lineare gerarchico sulla base dell'*AIC* - anno 2020

Modello	AIC
Modello di regressione lineare (*complete pooling*)	16954.44
Modello lineare gerarchico	15158.1

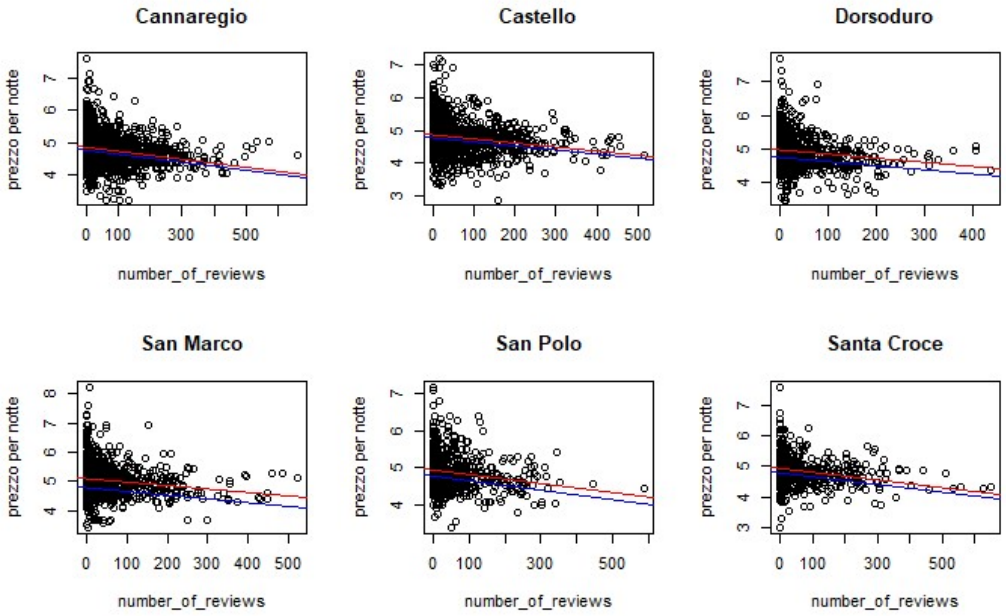

Figura 4.1: Confronto tra un modello di regressione lineare e un modello lineare gerarchico - anno 2019

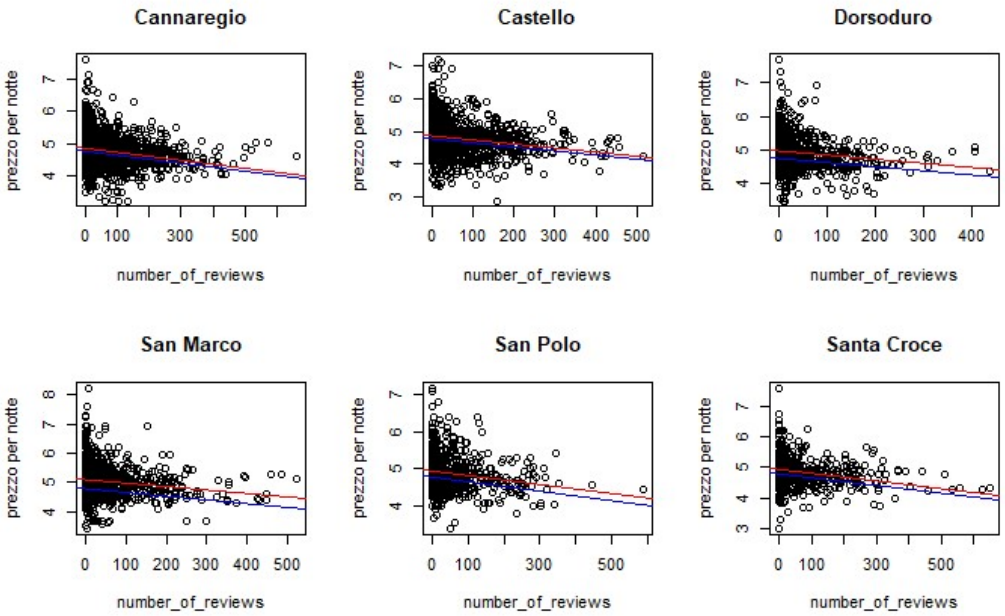

Figura 4.2: Confronto tra un modello di regressione lineare e un modello lineare gerarchico - anno 2020

- La zona di ubicazione degli edifici influenza la stima del prezzo sia prima sia dopo lo scoppio della pandemia di Covid-19.
- Sembrerebbe opportuno considerare ogni quartiere come un'entità a sé stante per la definizione del prezzo. Tuttavia, è probabile che si otterrebbero risultati poco informativi.
- I risultati confermano che una abitazione situata a Venezia o sulle isole tende a fare alzare la stima del prezzo rispetto a una sistemazione che si trova fuori da Venezia.
- La disponibilità annuale degli alloggi influisce molto sulla stima del prezzo prima dello scoppio della pandemia di Covid-19 e molto meno dopo. Questo risultato è collegabile al fatto che le persone hanno cominciato a viaggiare meno e a cercare affitti con prenotazioni più brevi e flessibili dopo l'inizio dell'emergenza.
- Si ha evidenza dell'importanza nella determinazione del prezzo della spaziosità degli alloggi, misurata in termini di persone potenzialmente ospitabili, della notorietà degli annunci e della tipologia di stanza proposta. Abitazioni che potrebbero ospitare più persone tendono ad avere prezzi maggiori, come anche le intere case e appartamenti rispetto alle stanze private e condivise. Invece, gli alloggi più conosciuti sono quelli con prezzi più bassi.

Tabella 4.3: Modello lineare gerarchico con intercetta casuale - anno 2019

Variabile	Coefficiente	Errore std	Valore-t	Pr(>|t|)	Significatività
(Intercept)	3.910	0.05254	74.43	< 2e-16	***
availability_365	0.0004635	0.00004227	10.97	< 2e-16	***
accommodates	0.1116	0.002665	41.87	< 2e-16	***
number_of_reviews	-0.0009639	0.00006039	-15.96	< 2e-16	***
room_type - Private room	-0.1661	0.01256	-13.23	< 2e-16	***
room_type - Shared room	-0.8233	0.05406	-15.23	< 2e-16	***

Tabella 4.4: Modello lineare gerarchico con intercetta casuale - anno 2020

Variabile	Coefficiente	Errore std	Valore-t	Pr(>|t|)	Significatività
(Intercept)	4.018	0.04763	84.359	< 2e-16	***
availability_365	0.0002629	0.00004412	5.958	3.92e-09	***
accommodates	0.1055	0.003079	34.281	< 2e-16	***
number_of_reviews	-0.001105	0.00006554	-16.854	< 2e-16	***
room_type - Hotel room	0.03908	0.03078	1.270	0.22098	
room_type - Private room	-0.2321	0.01557	-14.903	< 2e-16	***
room_type - Shared room	-0.8608	0.06429	-13.389	< 2e-16	***

Tabella 4.5: Confronto dei modelli sulla base dell'AIC - anno 2019

Modello	AIC
Modello di regressione lineare	9820.397
Modello lineare gerarchico	9979.479

Tabella 4.6: Confronto dei modelli sulla base dell'AIC - anno 2020

Modello	AIC
Modello di regressione lineare	12976.2
Modello lineare gerarchico	13095.81

Tabella 4.7: Modello di regressione lineare - anno 2019

Variabile	Coefficiente	Errore std	Valore-t	Pr(>|t|)	Significatività
(Intercept)	4.142	0.1750	23.673	< 2e-16	***
zona - Alberoni	-0.02440	0.1902	-0.128	0.897910	
zona- Altobello	-0.3125	0.1923	-1.626	0.104069	
zona - Asseggiano	-0.2640	0.4599	-0.574	0.566001	
zona - Bissuola	-0.5208	0.1814	-2.871	0.004098	**
zona - Burano	-0.04658	0.2103	-0.221	0.824740	
zona - Ca' Brentelle	-0.2364	0.3475	-0.680	0.496422	
zona - Ca' Emiliani	-0.2.338	0.3012	-0.776	0.437616	
zona - Ca' Sabbioni	-1.146	0.3480	-3.293	0.000995	***
zona - Campalto	-0.2369	0.2007	-1.180	0.237961	
zona - Campalto Bagaron	-0.2076	0.3009	-0.989	0.322703	
zona - Campalto CEP	-0.3766	0.2370	-1.589	0.112069	
zona - Campalto Cimitero	-0.3315	0.4598	-0.721	0.470999	
zona - Campalto Gobbi	-0.3026	0.2371	-1.277	0.201756	
zona - Cannaregio	0.1882	0.1743	1.080	0.280191	
zona - Carpenedo	-0.5396	0.1801	-2.997	0.002737	**
zona - Case Dosa	-0.1688	0.4599	-0.367	0.713625	
zona - Castello	0.2032	0.1743	1.165	0.243898	
zona - Chirignago	-0.6415	0.1892	-3.390	0.000701	***
zona - Cipressina	-0.6674	0.2199	-3.034	0.002417	**
zona - Dese	-0.05429	0.3475	-0.156	0.875845	
zona - Dorsoduro	0.3055	0.1748	1.747	0.080626	.
zona - Favaro	-0.5726	0.1817	-3.152	0.001629	**
zona - Ferrarese	-0.09538	0.3477	-0.274	0.783861	
zona - Gatta - Bondu	-0.7703	0.3478	2.215	0.026782	*
zona - Gazzera	-0.6357	0.2022	-3.145	0.001669	**
zona - Giudecca	0.2535	0.1765	1.436	0.150917	
zona - Giustizia	-0.4078	0.4598	-0.887	0.375138	
zona - Isola Santa Cristina	2.000	0.4608	4.341	0.0000143	***
zona - La Favorita	-0.4463	0.2244	-1.989	0.046760	*
zona - Lido	0.07169	0.1756	0.408	0.683086	
zona - Macello	-0.7465	0.4598	-1.624	0.104512	
zona - Malamocco	0.06958	0.2130	0.327	0.743936	
zona - Malcontenta	-0.8429	0.3475	-2.426	0.015302	*
zona - Marghera	-0.4310	0.1767	-2.439	0.014751	*
zona - Marghera Catene	-0.5165	0.1844	-2.801	0.005105	**
zona - Marghera Zona Industriale	-0.5287	0.2077	-2.545	0.010940	*
zona - Marocco Terraglio	-0.3639	0.2750	-1.324	0.185664	
zona - Murano	0.03611	0.1805	0.200	0.841457	
zona - Pellestrina	-0.5962	0.3477	-1.715	0.086420	.
zona - Piave 1860	-0.4046	0.1747	-2.317	0.020541	*
zona - Pra' Secco	-0.3931	0.4599	-0.855	0.392620	
zona - Quartiere Pertini	-0.5654	0.3015	-1.875	0.060807	.
zona - Sacca Fisola	-0.1140	0.2578	-0.442	0.658452	
zona - San Lorenzo XXV Aprile	-0.3579	0.1761	-2.032	0.042156	*
zona - San Marco	0.4282	0.1746	2.453	0.014189	*
zona - San Pietro in Volta	-0.4341	0.4599	-0.944	0.345251	
zona - San Polo	0.2757	0.1748	1.577	0.114832	
zona - Sant'Elena	0.9680	0.1797	0.539	0.590193	
zona - Sant'Erasmo	-0.2552	0.2245	-1.137	0.255739	
zona - Santa Barbara	-0.4562	0.1900	-2.401	0.016387	*
zona - Santa Croce	0.2853	0.1748	1.632	0.102756	
zona - Scaramuzza	-0.6681	0.4598	-1.453	0.146212	
zona - Tessera	0.1354	0.1952	0.693	0.488101	
zona - Torcello	-0.5724	0.3478	-1.646	0.099877	.
zona - Trivignano	-0.5036	0.4599	-1.095	0.273501	
zona - Tronchetto	0.5919	0.2750	2.153	0.031372	*
zona - Villabona	-0.6074	0.3012	-2.017	0.043758	*
zona - Villaggio San Marco	-0.3046	0.1905	-1.599	0.109821	
zona - Villaggio Sartori	-0.7687	0.4600	-1.671	0.094736	.
zona - Zelarino	-0.5378	0.3013	-1.785	0.074279	.
zona - Zona Commerciale via Torino	-0.4305	0.2081	-2.069	0.038582	*
availability_365	0.0004642	0.00004231	10.970	< 2e-16	***
accommodates	0.1114	0.2675	41.651	< 2e-16	***
number_of_reviews	-0.0009622	0.00006041	-15.930	< 2e-16	***
room_type-Private room	-0.1654	0.01259	-13.133	< 2e-16	***
room_type-Shared room	-0.8221	0.05411	-15.192	< 2e-16	***

Tabella 4.8: Modello lineare gerarchico - anno 2019

Variabile	Coefficiente	Effetti fissi intercetta	Effetti casuali intercetta
zona - Acroporto	4.095519	3.9102598378	0.185259328
zona - Alberoni	4.107350	3.9102598378	0.197090006
zona- Altobello	3.833824	3.9102598378	-0.076436276
zona - Asseggiano	3.897496	3.9102598378	-0.012764070
zona - Bissuola	3.627318	3.9102598378	-0.282942145
zona - Burano	4.075531	3.9102598378	0.165271499
zona - Ca' Brentelle	3.908174	3.9102598378	-0.002085393
zona - Ca' Emiliani	3.908644	3.9102598378	-0.001615933
zona - Ca' Sabbioni	3.391777	3.9102598378	-0.518483197
zona - Campalto	3.905702	3.9102598378	-0.004557901
zona - Campalto Bagaron	3.866871	3.9102598378	-0.043389066
zona - Campalto CEP	3.791162	3.9102598378	-0.119097789
zona - Campalto Cimitero	3.870649	3.9102598378	-0.039610559
zona - Campalto Gobbi	3.851932	3.9102598378	-0.058328293
zona - Cannaregio	4.329766	3.9102598378	0.419506094
zona - Carpenedo	3.607836	3.9102598378	-0.302423960
zona - Case Dosa	3.934736	3.9102598378	0.024475826
zona - Castello	4.344634	3.9102598378	0.434373864
zona - Chirignago	3.518423	3.9102598378	-0.391836512
zona - Cipressina	3.53256	3.9102598378	-0.377699062
zona - Dese	4.011112	3.9102598378	0.100852058
zona - Dorsoduro	4.446031	3.9102598378	0.535771525
zona - Favaro	3.577034	3.9102598378	-0.333226163
zona - Ferrarese	3.987075	3.9102598378	0.076815498
zona - Gatta - Bondu	3.605046	3.9102598378	-0.305214195
zona - Gazzera	3.539931	3.9102598378	-0.370329172
zona - Giudecca	4.391909	3.9102598378	0.481648755
zona - Giustizia	3.840668	3.9102598378	-0.069591471
zona - Isola Santa Cristina	4.791337	3.9102598378	0.881077531
zona - La Favorita	3.726980	3.9102598378	-0.183279733
zona - Lido	4.212047	3.9102598378	0.301786992
zona - Macello	3.706705	3.9102598378	-0.203554788
zona - Malamocco	4.177398	3.9102598378	0.267138321
zona - Malcontenta	3.564512	3.9102598378	-0.345747361
zona - Marghera	3.712917	3.9102598378	-0.197342406
zona - Marghera Catene	3.634401	3.9102598378	-0.275858507
zona - Marghera Zona Industriale	3.642692	3.9102598378	-0.267568290
zona - Marocco Terraglio	3.814367	3.9102598378	-0.095892610
zona - Murano	4.172858	3.9102598378	0.262597739
zona - Pellestrina	3.703573	3.9102598378	-0.206686748
zona - Piave 1860	3.737990	3.9102598378	-0.172270102
zona - Pra' Secco	3.846371	3.9102598378	-0.063888686
zona - Quartiere Pertini	3.689550	3.9102598378	-0.220709919
zona - Sacca Fisola	4.000463	3.9102598378	0.090203611
zona - San Lorenzo XXV Aprile	3.784948	3.9102598378	-0.125311433
zona - San Marco	4.568978	3.9102598378	0.658717759
zona - San Pietro in Volta	3.830426	3.9102598378	-0.079834296
zona - San Polo	4.416339	3.9102598378	0.506078794
zona - Sant'Elena	4.233201	3.9102598378	0.322941534
zona - Sant'Erasmo	3.890596	3.9102598378	-0.019664067
zona - Santa Barbara	3.696281	3.9102598378	-0.213978607
zona - Santa Croce	4.425954	3.9102598378	0.515694319
zona - Scaramuzza	3.738080	3.9102598378	-0.172179582
zona - Tessera	4.254750	3.9102598378	0.344489851
zona - Torcello	3.716682	3.9102598378	-0.193577821
zona - Trivignano	3.802505	3.9102598378	-0.107755294
zona - Tronchetto	4.506079	3.9102598378	0.595819655
zona - Villabona	3.661386	3.9102598378	-0.248874013
zona - Villaggio San Marco	3.841119	3.9102598378	-0.069141299
zona - Villaggio Sartori	3.697873	3.9102598378	-0.212386484
zona - Zelarino	3.707247	3.9102598378	-0.203012890
zona - Zona Commerciale via Torino	3.730795	3.9102598378	-0.179464465
availability_365	0.000463524		
accommodates	0.1115951		
number_of_reviews	-0.0009639473		
room_type-Private room	-0.1661338		
room_type-Shared room	-0.8232508		

Capitolo 5

Conclusioni

A conclusione di questo elaborato, si propongono alcune considerazioni in merito al lavoro di tesi svolto.
L'obiettivo principale delle analisi è quello di indagare quali sono gli elementi che maggiormente influenzano il prezzo degli annunci di "Airbnb" per la città di Venezia e se essi hanno subito variazioni in seguito allo scoppio della pandemia di Covid-19. Le informazioni utilizzate a tale scopo provengono dal sito "insideairbnb.com" e i dataset sono aggiornati al mese di giugno degli anni 2019, 2020 e 2021.
Dopo aver effettuato una preparazione dei dataset e una prima fase di analisi descrittive, si passa all'applicazione di modelli di regressione per il prezzo degli annunci ai dati delle diverse annualità. Si riportano i risultati relativi al 2019 e al 2020, con lo scopo di poter effettuare alcune osservazioni sulla situazione precedente e successiva allo scoppio della pandemia. Si nota, infatti, che dalle due annualità che seguono l'inizio della situazione di emergenza si ottengono risultati analoghi. Per quanto riguarda la scelta dei modelli, invece, si propongono il modello lineare e il modello *Gradient Booosting* perchè sono ritenuti i più efficienti e facilmente interpretabili. Da queste analisi emerge che la zona di ubicazione degli alloggi è importante per la definizione del prezzo e che una sistemazione a Venezia o sulle isole tende ad avere un prezzo maggiore di una collocata fuori da Venezia. Si evince, inoltre, che la spaziosità delle abitazioni e altre caratteristiche strutturali sono molto più rilevanti prima dello scoppio della pandemia e si deduce che, in seguito all'emergen-

za siano garantiti degli standard minimi riguardo ad esse. Nel 2020, invece, acquisisce importanza la tipologia di alloggio affittato e le intere abitazioni vengono valutate di più delle singole stanze in termini di prezzo. Anche le politiche di recessione dagli affitti pesano maggiormente sulle stime e sembrerebbero esserci offerte con prezzi più bassi e cancellazioni più flessibili. Infine, appare che, dopo lo scoppio della pandemia, "hosts" con più notorietà e con maggiore attività sulla piattaforma abbiano la possibilità di fissare prezzi più elevati.

Dal momento che la zona di ubicazione degli edifici sembra influire pesantemente sulla definizione del prezzo, si propone un approfondimento di questo aspetto attraverso l'applicazione di modelli lineari gerarchici ai diversi dataset. A tale scopo, si decide di utilizzare una variabile originale dei dataset iniziali, indicante i quartieri. Per le analisi precedenti, essa era stata trasformata. Si decide, inoltre, di selezionare solo alcune delle variabili esplicative, poiché, sulla base dei risultati emersi dalla regressione, esse appaiono particolarmente rilevanti per la stima del prezzo. Si propone un confronto tra gli anni 2019 e 2020. Il modello utilizzato è un modello lineare gerarchico con intercetta variabile. Esso viene confrontato con un modello di regressione lineare sulla base dell'AIC. Il modello lineare gerarchico sembra essere meno efficiente del modello di regressione lineare e questo risultato indica che ogni quartiere può essere considerato una entità a sé stante nella definizione del prezzo. Tuttavia, osservando la significatività dei coefficienti del modello di regressione lineare, si deduce che, stimando il prezzo sotto tale assunzione, si otterrebbero risultati poco informativi. Si conferma la rilevanza del quartiere di ubicazione degli edifici nella definizione del prezzo e che una sistemazione a Venezia o sulle isole tende a costare di più di una fuori da Venezia. Si conferma, inoltre, la validità delle esplicative scelte, anche se la disponibilità annuale perde importanza dopo lo scoppio della pandemia. Probabilmente, coloro che viaggiano cercano ora prenotazioni più brevi e flessibili. Si osserva, infine, che le abitazioni che potrebbero ospitare più persone tendono ad avere prezzi più elevati, come anche le intere case e appartamenti rispetto alle stanze private e condivise. Invece, gli alloggi più conosciuti e recensiti sono quelli con prezzi più bassi.

Le analisi proposte si potrebbero migliorare considerando anche alcune delle

variabili che erano state inizialmente escluse dai dataset in una fase preliminare perché erano state valutate di poco interesse, effettuando delle trasformazioni che le rendano maggiormente utilizzabili. Si potrebbero, inoltre, arricchire le analisi con uno studio sulle variabili esplicative, utilizzando esse stesse come variabili risposta nei modelli, per indagare il loro andamento prima e dopo lo scoppio della pandemia.
Alcuni possibili sviluppi del lavoro proposto sono quelli di confrontarlo con i risultati delle stesse analisi effettuate per altre città italiane, oppure per località con legislazioni differenti. I risultati proposti, potrebbero essere paragonati, inoltre, con analisi sulla stima dei prezzi effettuate per periodi antecedenti e successivi all'entrata in vigore di alcune specifiche normative per il contenimento del virus o ad annualità che si collocano prima e dopo la diffusione dei vaccini.

Bibliografia

[1] Azzalini, A., Scarpa, B. (2009). Analisi dei dati e data mining. Springer Science Business Media.

[2] Bernardi, M., Guidolin, M. (2022). The determinants of Airbnb prices in New York city: a spatial quantile regression approach. (in corso di revisione)

[3] Boros, L., Dudás, G., Kovalcsik, T. (2020). The effects of COVID-19 on Airbnb. Hungarian Geographical Bulletin, 69(4), 363-381.

[4] Oskam, J., Boswijk, A. (2016). Airbnb: the future of networked hospitality businesses. Journal of tourism futures.

[5] Contu, G., Conversano, C., Frigau, L., Mola, F. (2019). The impact of Airbnb on hidden and sustainable tourism: the case of Italy. International Journal of Tourism Policy, 9(2), 99-130.

[6] Cesarani, M., Nechita, F. (2017). Tourism and the sharing economy. An evidence from Airbnb usage in Italy and Romania. Symphonya, 32-47.

[7] Cheng, M., Jin, X. (2019). What do Airbnb users care about? An analysis of online review comments. International Journal of Hospitality Management, 76, 58-70.

[8] Gambatesa, R. (2020). Airbnb e l'importanza delle parole: Analisi di text mining per valutare l'attrattività di un annuncio in tre città italiane. Tesi di laurea. Università di Padova. Laurea Magistrale in Scienze Statistiche.

[9] Gelman, A., Hill, J. (2006). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.

[10] Lee, K. H., Kim, D. (2019). A peer-to-peer (P2P) platform business model: The case of Airbnb. Service Business, 13(4), 647-669.

[11] Hastie, T., Tibshirani, R., Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

[12] Tussyadiah, I. P., Park, S. (2018). When guests trust hosts for their words: Host description and trust in sharing economy. Tourism Management, 67, 261-272.

Sitografia

Corona, A. (2019). Airbnb numbers in Venice. InsideAirbnb Infokit. http://insideairbnb.com/venice/report_en.html

http://insideairbnb.com/, aggiornato al 29/07/2021 - 11/05/2021 - 02/04/2021

https://www.viamichelin.it/web/Mappe-Piantine/Mappa_Piantina-Venezia-30121-Venezia-Italia, aggiornato al 01/02/2021

https://evenice.it/venezia/storie-tradizioni/venezia-i-suoi-sestieri, aggiornato al 01/02/2021

https://www.airbnb.it/help/article/2908/termini-del-servizio, aggiornato al 12/04/2021

https://www.airbnb.it/help/article/2909/termini-del-servizio-di-pagamento, aggiornato al 12/04/2021

https://www.airbnb.it/help/article/2859/informativa-sulla-privacy-di-airbnb-archivio#sec201910_2, aggiornato al 12/04/2021

https://it.wikipedia.org/wiki/Economia_collaborativa, aggiornato al 22/09/2021

https://news.airbnb.com/about-us/, aggiornato al 22/09/2021

http://insideairbnb.com/about.html, aggiornato al 22/09/2021

https://www.airbnb.it/help/article/2701/termini-delle-circostanze-attenuanti-per-la-pandemia-da-coronavirus-covid19, aggiornato al 15/09/2021

https://www.airbnb.it/help/article/2809/il-processo-avanzato-di-pulizia-in-5-fasi-di-airbnb, aggiornato al 15/09/2021

https://www.airbnb.it/resources/hosting-homes/a/how-to-tell-your-guests-about-your-cleaning-process-190, aggiornato al 15/09/2021

https://blog.atairbnb.com/health-and-safety-guidelines-for-hosting-experiences-in-reopened-regions-it/, aggiornato al 15/09/2021

https://www.airbnb.it/help/article/2830/quando-e-dove-ripartiranno-le-esperienze-airbnb,
aggiornato al 15/09/2021

https://www.airbnb.it/help/article/1593/cancella-o-riprogramma-unesperienza-airbnb, aggiornato al 15/09/2021

https://www.airbnb.it/help/article/1320/termini-delle-circostanze-attenuanti, aggiornato al 15/09/2021

https://www.airbnb.it/resources/hosting-homes/a/answers-for-travelers-about-covid-19-153, aggiornato al 15/09/2021

https://www.airbnb.it/, aggiornato al 15/09/2021

Appendice A

Analisi delle correlazioni

Si propone, per ciascuno dei tre dataset, un'analisi delle correlazioni tra alcune delle variabili quantitative o qualitative ordinali a disposizione. Tali informazioni sono raggruppate sulla base di alcune tematiche e sono tutte confrontate con la variabile risposta, ovvero il prezzo totale per notte di ogni annuncio. Lo scopo di questa analisi esplorativa è quello di ottenere una prima indicazione del legame tra i vari dati disponibili ed in particolare di constatare quali variabili potrebbero influire maggiormente sui prezzi e in che modo e di identificare la presenza di variabili *leaker*.

A.1 Analisi delle correlazioni nel dataset relativo al 2019

Caratteristiche dell'alloggio

Per questa prima analisi sono state prese in considerazione le variabili "price_dollars", "host_total_listings_count", "accommodates", "bathrooms", "beds", "guests_included". Si osservano solo valori positivi. In particolare, il prezzo ha una correlazione di 0.265 con il numero di persone ospitabili ("accommodates") e di 0.207 con la numerosità dei bagni (variabile dicotomica pari a 0 per meno di due bagni e a uno per la presenza di almeno due

bagni).
Sono rilevanti, inoltre le correlazioni tra "accommodates" e le altre variabili cosiderate (0.101 con "host_total_listings_count", 0.459 con "bathrooms", 0.790 con "beds" e 0.464 con "guests_included") e quelle tra "bathrooms" e "guests_included", tra "bathrooms" e "bed"s e tra "guests_included" e "beds", pari rispettivamente a 0.208, 0.401 e 0.358. Si vedano la Tabella A.1, la Figura A.1 e la Figura A.6.

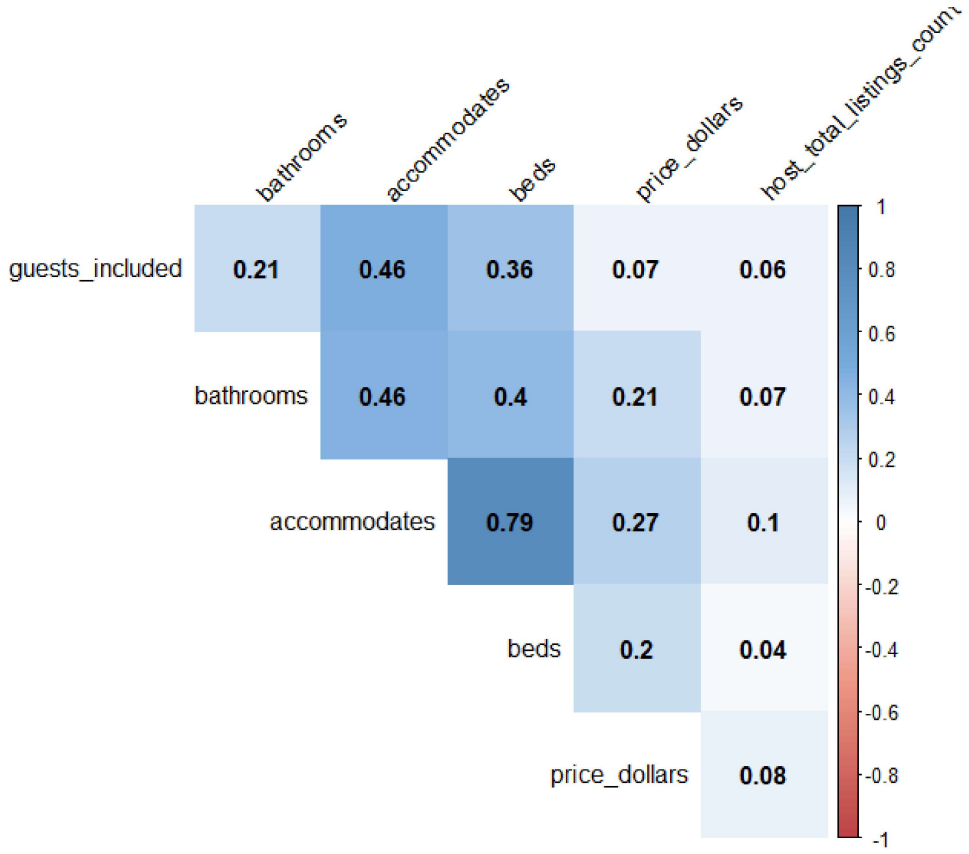

Figura A.1: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2019

Voci di revisione

La correlazione del prezzo con tutte le voci di revisione è dell'ordine di grandezza di circa -0.100 con tutte le variabili considerate. Questo permette di osservare che gli alloggi meno cari hanno punteggi più elevati nelle recensioni

e un maggior numero di persone ha espresso un giudizio relativamente ad essi. Si segnala una correlazione elevata, pari a 0.758, tra "number_of_reviews" e "number_of_reviews_ltm" e si osserva che esse si comportano in modo analogo nei confronti delle altre variabili. Potrebbe, pertanto, trattarsi di variabili *leaker*. Infine, si osserva che le voci relative agli specifici aspetti di revisione hanno tra loro una correlazione superiore a 0.978. Anche in questo caso, si suppone possano portare informazione ridondante e occorre, pertanto, prestare attenzione al loro utilizzo all'interno dei modelli. Si vedano la tabella Tabella A.2, la Figura A.2 e la Figura A.7.

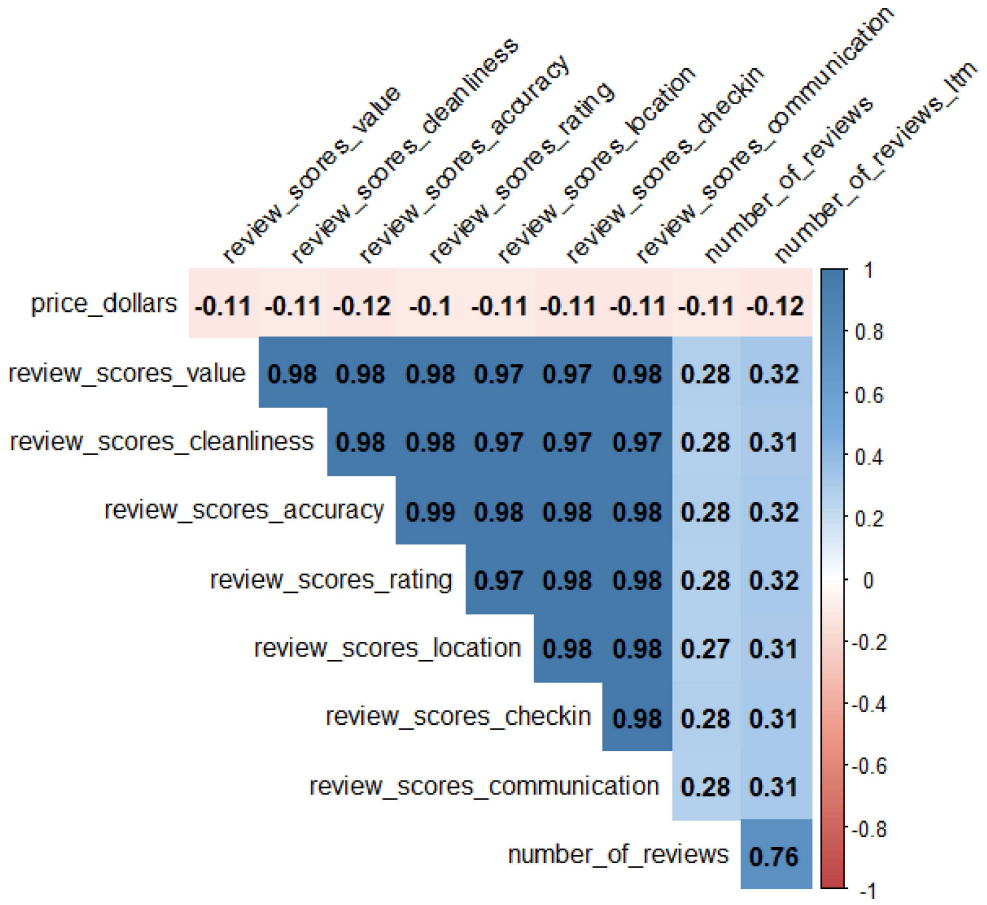

Figura A.2: Correlazioni delle voci di revisione relativamente al 2019

Conteggi

La correlazione del prezzo con le voci di conteggio è, in generale, molto bassa, positiva solo nel caso del numero totale di annunci e del numero di annunci relativi a case intere, per cui è circa 0.112. Tali evidenze possono essere ricondotte a una diminuzione del prezzo con l'aumento delle proprietà affittate, poichè chi gestisce più annunci tende ad affittare singole stanze e non intere abitazioni. Gli unici casi di correlazioni elevate sono quelli di "calculated_host_listings_count" con il conteggio degli annunci per intere case (0.979) e con il conteggio degli annunci per stanze private (0.282). Si vedano la tabella Tabella A.3 e la Figura A.3.

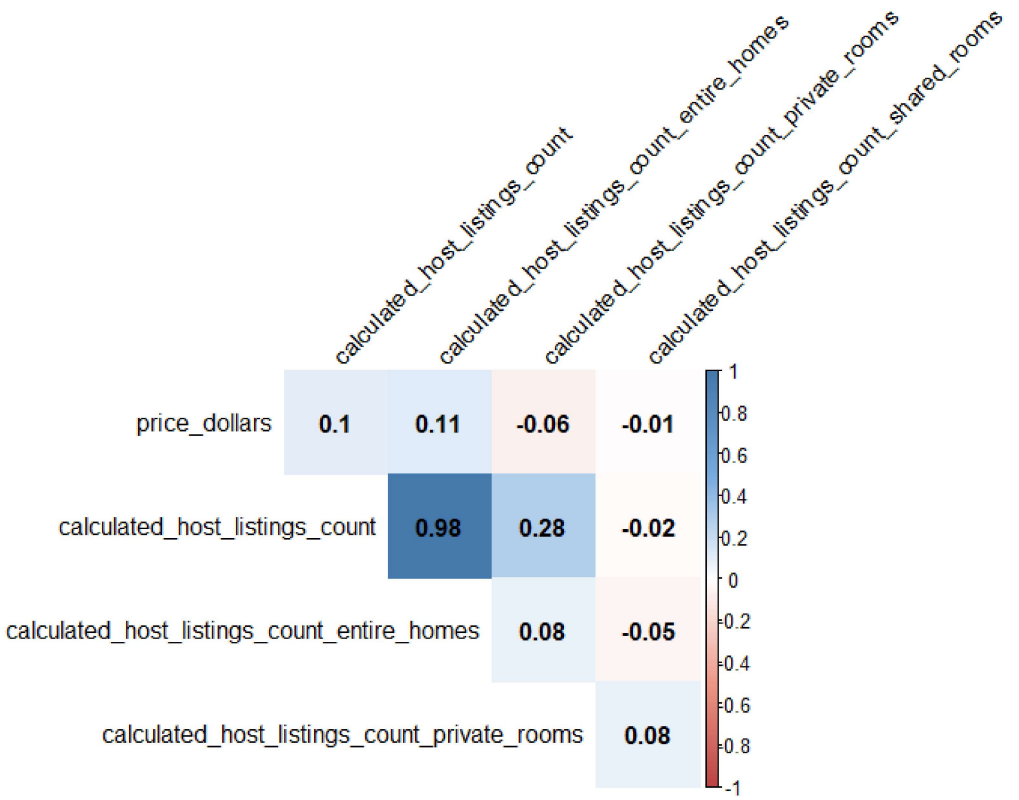

Figura A.3: Correlazioni delle variabili di conteggio nel 2019

Voci di prezzo

Si osserva che la correlazione tra il prezzo per notte e i prezzi settimanali e mensili è molto elevata, pari a 0.997 in entrambi i casi. Le tre variabili presentano, poi, correlazioni analoghe nei confronti delle altre voci di prezzo, il che spinge a pensare che portino informazione ridondante. Si può notare, poi, che il deposito di sicurezza e il prezzo delle pulizie hanno correlazioni con le altre variabili che si aggirano rispettivamente attorno a 0.305 (0.311 con il prezzo per notte) e 0.217 (0.221 con il prezzo per notte). Fanno eccezione le correlazioni con "extra_people_dollars", che sono molto basse, inferiori a 0.090 per tutte le variabili (0.068 con il prezzo a notte). Si vedano la tabella Tabella A.4 e la Figura A.4.

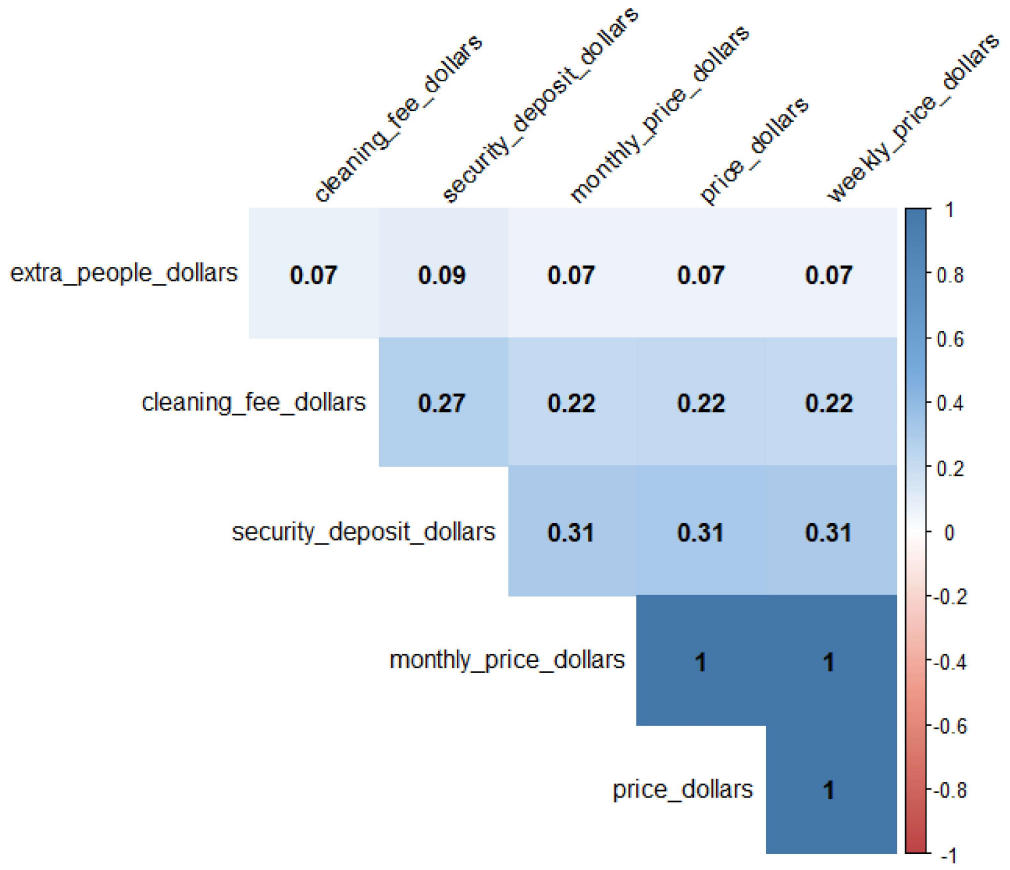

Figura A.4: Correlazioni delle voci di prezzo nel 2019

Caratteristiche degli "host"

Le correlazioni del prezzo con le variabili relative alle caratteristiche degli "host" forniscono solo una indicazione di un possibile legame tra le varibili, in quanto si tratta di variabili per lo più dicotomiche. Esse non sono interpretabili, tuttavia, sono, in generale, molto basse. Si vedano la tabella Tabella A.5 e la Figura A.5.

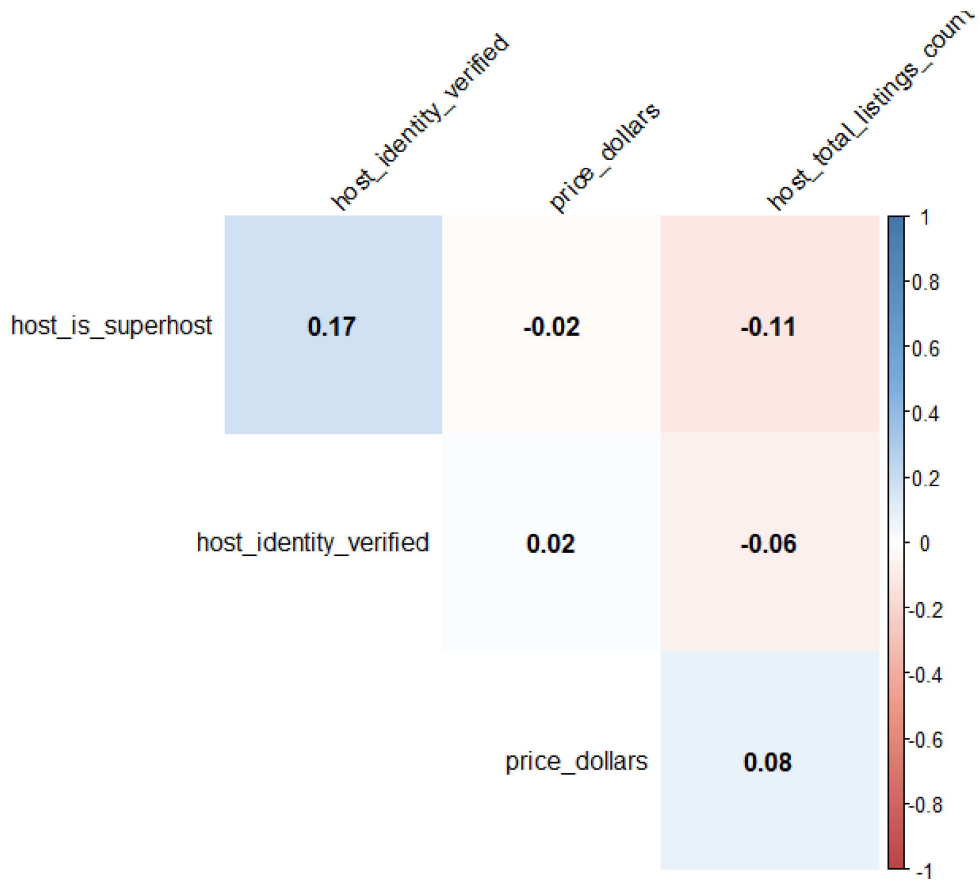

Figura A.5: Correlazioni delle caratteristiche degli "host" nel 2019

A.2 Analisi delle correlazioni nel dataset relativo al 2020

Caratteristiche dell'alloggio

Nel 2020, il prezzo per notte ha correlazioni molto simili rispetto al 2019 con le altre variabili, anche se meno elevate. Si tratta, anche in questo caso, di correlazioni positive. In particolare si segnalano le correlazioni con "accommodates" pari a 0.225, con "bathrooms" pari a 0.195 e con "beds" pari a 0.175. "Host_total_listings_count" presenta con tutte le variabili valori inferiori a 0.075 e "accommodates" valori di circa 0.480 con "beds" e 0.788 con "guests_included". La correlazione di "bathrooms" è di 0.478 con il numero di letti e di 0.240 con il numero di ospiti a cui si riferisce l'annuncio, mentre quella tra "beds" e "guests_included" è di 0.371. Si segnalano correlazioni molto basse di "guests_included" con "price_dollars" e "host_total_listings_count". Si vedano la tabella Tabella A.6, la Figura A.8 e la Figura A.9.

Voci di revisione

Il prezzo e in generale tutte le variabili relative alle voci di revisione presentano valori della correlazione molto simili al 2019 degli stessi ordini di grandezza. Si segnalano i valori elevati delle correlazioni di "number_of_reviews" con le altre variabili (pari almeno a 0.311) e a 0.682 con "number_of_reviews_ltm". Si vedano la tabella Tabella A.7 , la Figura A.10 e la Figura A.11.

Conteggi

Rispetto al 2019, le voci di conteggio sono più fortemente correlate tra loro. Probabilmente lo scenario emerso durante la pandemia nel mercato dei viaggi ha reso possibile la sopravvivenza di annunci con caratteristiche molto simili o molto contrastanti e le tipologie di offerta hanno teso a uniformarsi. Si segnala una correlazione dell'ordine di grandezza di 0.200 tra il numero totale di annunci di un "host" e le altre variabili, e valori pari a 0.969 tra "host_total_listings_count" e "host_total_listings_count_entire_homes",

pari a -0.114 tra quest'ultima variabile e il totale di stanze private di un "host" e di 0.134 tra il totale delle stanze private e condivise degli "host". Si vedano la tabella Tabella A.8 e la Figura A.12.

Voci di prezzo

Si osserva che le correlazioni tra le voci di prezzo aumentano sensibilmente rispetto al 2019. In particolare, il prezzo per le persone aggiuntive e quello per le pulizie pesano di più nella definizione delle altre tipologie di prezzi, con valori delle correlazioni compresi nel primo caso tra 0.123 e 0.145 e nel secondo tra 0.145 e 0.470. Per quanto riguarda la variabile risposta, la correlazione del prezzo per notte con il prezzo per persone aggiuntive è di 0.126 e quella con il prezzo per le pulizie è di 0.470. Tali risultati sono coerenti con una maggiore cura verso le norme igieniche, i distanziamenti e il numero di ospiti accettati in seguito alla pandemia di Covid-19. Questo implica costi aggiuntivi per i proprietari, che comportano un aumento dei prezzi per i clienti. Si vedano la tabella Tabella A.9 e la Figura A.13.

Caratteristiche degli "host"

Valgono le considerazioni fatte per il 2019. I valori sono, inoltre, molto simili nel 2019 e nel 2020, generalmente poco più bassi nel 2020. Si vedano la tabella Tabella A.10 e la Figura A.14.

A.3 Analisi delle correlazioni nel dataset relativo al 2021

Data la diversa composizione del dataset rispetto ai precedenti due, non si è ottenuta per il 2021 la stessa quantità di risultati, nè è possibile considerare quanto emerso da questo dataset perfettamente confrontabile con i valori del 2020 e del 2019. Tuttavia, si possono trarre alcune conclusioni relativamente all'influenza delle variabili, le une rispetto alle altre, all'interno dello stesso set di dati.

Caratteristiche dell'alloggio

Le correlazioni tra il prezzo e le altre varaibili considerate sono, in generale, più basse rispetto agli altri dataset e, in ogni caso, sempre positive. In particolare, la correlazione con il numero totale di annunci per "host" è di 0.010, quella con "accommodates" è di 0.199 e quella con "beds" di 0.155. Anche le correlazioni tra le variabili riguardanti le caratteristiche degli alloggi sono in generale più basse. Si segnala solo un valore elevato, pari a 0.754, tra "accommodates" e "beds". Si vedano la tabella Tabella A.11 , la Figura A.15 e la Figura A.16.

Voci di revisione

Le correlazioni tra le voci di revisione sono in generale sensibilmente più basse rispetto agli altri dataset. Esse conservano, tuttavia, gli stessi segni di quelle del 2019. Si vedano la tabella Tabella A.12 , la Figura A.17 e la Figura A.18.

Conteggi

Per quanto riguarda le voci di conteggio, si osservano, nel 2021, valori molto bassi e difficilmente interpretabili, soprattutto dal punto di vista del segno. Si vedano la tabella Tabella A.13 e la Figura A.19.

Tabella A.1: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2019

	price_dollars	host_total_listings_count	accommodates	bathrooms	beds	guests_included
price_dollars	1.000	0.076	0.265	0.207	0.199	0.069
host_total_listings_count	0.076	1.000	0.101	0.066	0.037	0.061
accommodates	0.265	0.101	1.000	0.459	0.790	0.463
bathrooms	0.207	0.066	0.459	1.000	0.401	0.208
beds	0.199	0.037	0.790	0.401	1.000	0.358
guests_included	0.069	0.061	0.464	0.208	0.358	1.000

Tabella A.2: Correlazioni delle voci di revisione relativamente al 2019

	price_dollars	number_of_reviews	number_of_reviews_ltm	review_scores_accuracy	review_scores_checkin	review_scores_cleanliness	review_scores_communication	review_scores_location	review_scores_rating
price_dollars	1.000	-0.107	-0.117	-0.115	-0.110	-0.108	-0.112	-0.106	-0.102
number_of_reviews	-0.107	1.000	0.758	0.283	0.282	0.276	0.280	0.275	0.277
number_of_reviews_ltm	-0.117	0.758	1.000	0.318	0.312	0.309	0.311	0.306	0.316
review_scores_accuracy	-0.115	0.283	0.318	1.000	0.981	0.980	0.980	0.978	0.985
review_scores_checkin	-0.110	0.282	0.312	0.981	1.000	0.973	0.984	0.976	0.979
review_scores_cleanliness	-0.108	0.276	0.309	0.980	0.972	1.000	0.973	0.973	0.983
review_scores_communication	-0.112	0.280	0.311	0.980	0.984	0.974	1.000	0.976	0.981
review_scores_location	-0.106	0.275	0.306	0.978	0.976	0.973	0.976	1.000	0.974
review_scores_rating	-0.102	0.277	0.316	0.985	0.979	0.984	0.981	0.974	1.000
review_scores_value	-0.113	0.283	0.322	0.980	0.973	0.976	0.975	0.972	0.984

Tabella A.3: Correlazioni delle variabili di conteggio nel 2019

	price_dollars	calculated_host_listings_count	calculated_host_listings_count_entire_homes	calculated_host_listings_count_private_rooms	calculated_host_listings_count_shared_rooms
price_dollars	1.000	0.095	0.112	-0.063	-0.005
calculated_host_listings_count	0.095	1.000	0.979	0.282	-0.016
calculated_host_listings_count_entire_homes	0.112	0.979	1.000	0.080	-0.049
calculated_host_listings_count_private_rooms	-0.063	0.282	0.080	1.000	0.078
calculated_host_listings_count_shared_rooms	-0.005	-0.016	-0.049	0.078	1.000

Tabella A.4: Correlazioni delle voci di prezzo nel 2019

	price_dollars	weekly_price_dollars	monthly_price_dollars	extra_people_dollars	security_deposit_dollars	cleaning_fee_dollars
price_dollars	1.000	0.997	0.997	0.068	0.311	0.221
weekly_price_dollars	0.997	1.000	0.997	0.067	0.309	0.218
monthly_price_dollars	0.997	0.997	1.000	0.065	0.305	0.217
extra_people_dollars	0.068	0.066	0.065	1.000	0.090	0.072
security_deposit_dollars	0.311	0.308	0.305	0.090	1.000	0.271
cleaning_fee_dollars	0.221	0.218	0.217	0.072	0.271	1.000

Tabella A.5: Correlazioni delle caratteristiche degli “host” nel 2019

	price_dollars	host_is_superhost	host_identity_verified	host_total_listings_count
price_dollars	1.000	-0.018	0.016	0.076
host_is_superhost	-0.018	1.000	0.173	-0.115
host_identity_verified	0.016	0.173	1.000	-0.064
host_total_listings_count	0.076	-0.115	-0.064	1.000

Tabella A.6: Correlazioni tra il prezzo e le caratteristiche dell’alloggio nel 2020

	price_dollars	host_total_listings_count	accommodates	bathrooms	beds	guests_included
price_dollars	1.000	0.083	0.225	0.195	0.175	0.075
host_total_listings_count	0.083	1.000	0.069	0.044	0.038	0.074
accommodates	0.225	0.069	1.000	0.478	0.788	0.481
bathrooms	0.195	0.044	0.478	1.000	0.409	0.240
beds	0.175	0.038	0.788	0.409	1.000	0.371
guests_included	0.074	0.074	0.481	0.240	0.371	1.000

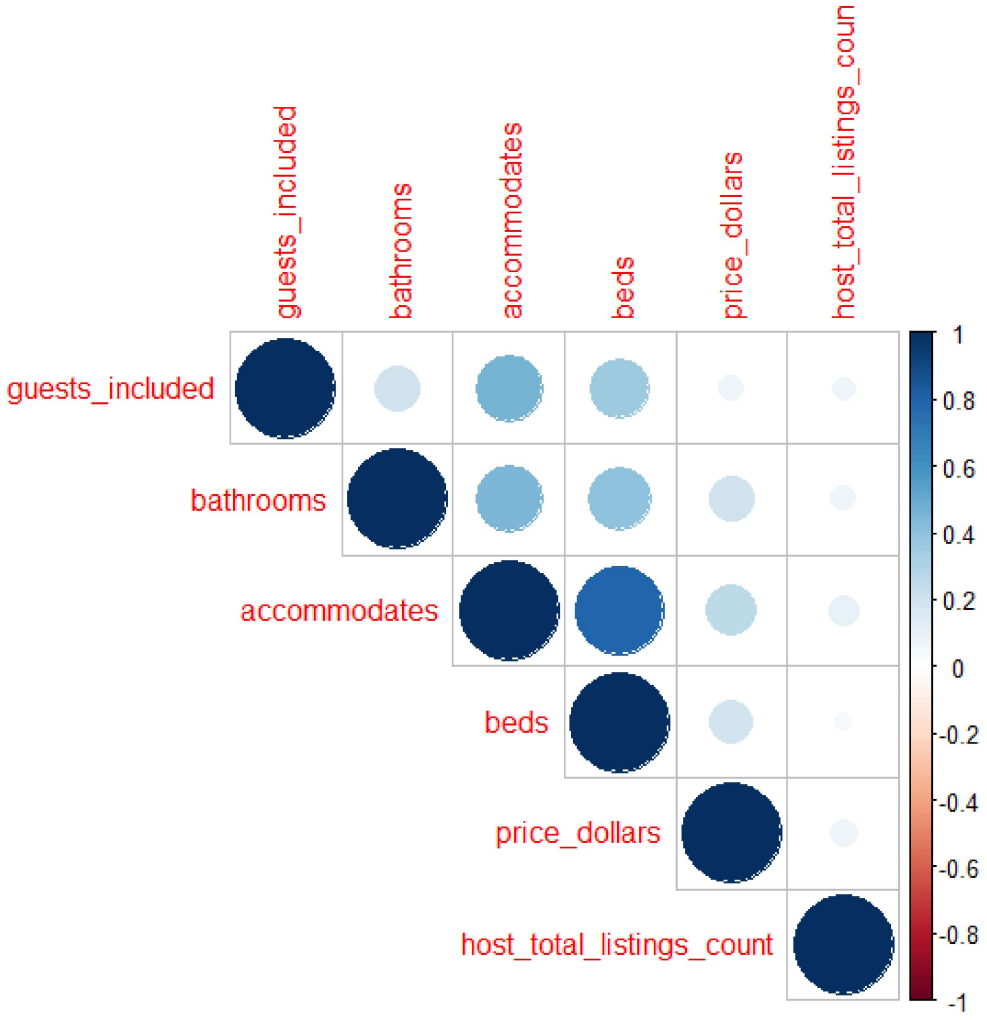

Figura A.6: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2019

Tabella A.7: Correlazioni delle voci di revisione relativamente al 2020

	price_dollars	number_of_reviews	number_of_reviews_ltm	review_scores_accuracy	review_scores_checkin	review_scores_cleanliness	review_scores_communication	review_scores_location	review_scores_rating
price_dollars	1.000	-0.127	-0.126	-0.174	-0.171	-0.167	-0.175	-0.169	-0.165
number_of_reviews	-0.127	1.000	0.682	0.322	0.318	0.312	0.316	0.312	0.317
number_of_reviews_ltm	-0.126	0.682	1.000	0.346	0.339	0.337	0.339	0.337	0.345
review_scores_accuracy	-0.174	0.322	0.346	1.000	0.985	0.982	0.984	0.982	0.989
review_scores_checkin	-0.171	0.318	0.339	0.985	1.000	0.977	0.987	0.982	0.984
review_scores_cleanliness	-0.167	0.312	0.337	0.982	0.977	1.000	0.977	0.978	0.986
review_scores_communication	-0.175	0.316	0.339	0.984	0.987	0.977	1.000	0.981	0.985
review_scores_location	-0.169	0.312	0.336	0.982	0.982	0.978	0.981	1.000	0.981
review_scores_rating	-0.165	0.317	0.345	0.989	0.984	0.986	0.985	0.981	1.000
review_scores_value	-0.173	0.316	0.348	0.984	0.980	0.980	0.980	0.979	0.988

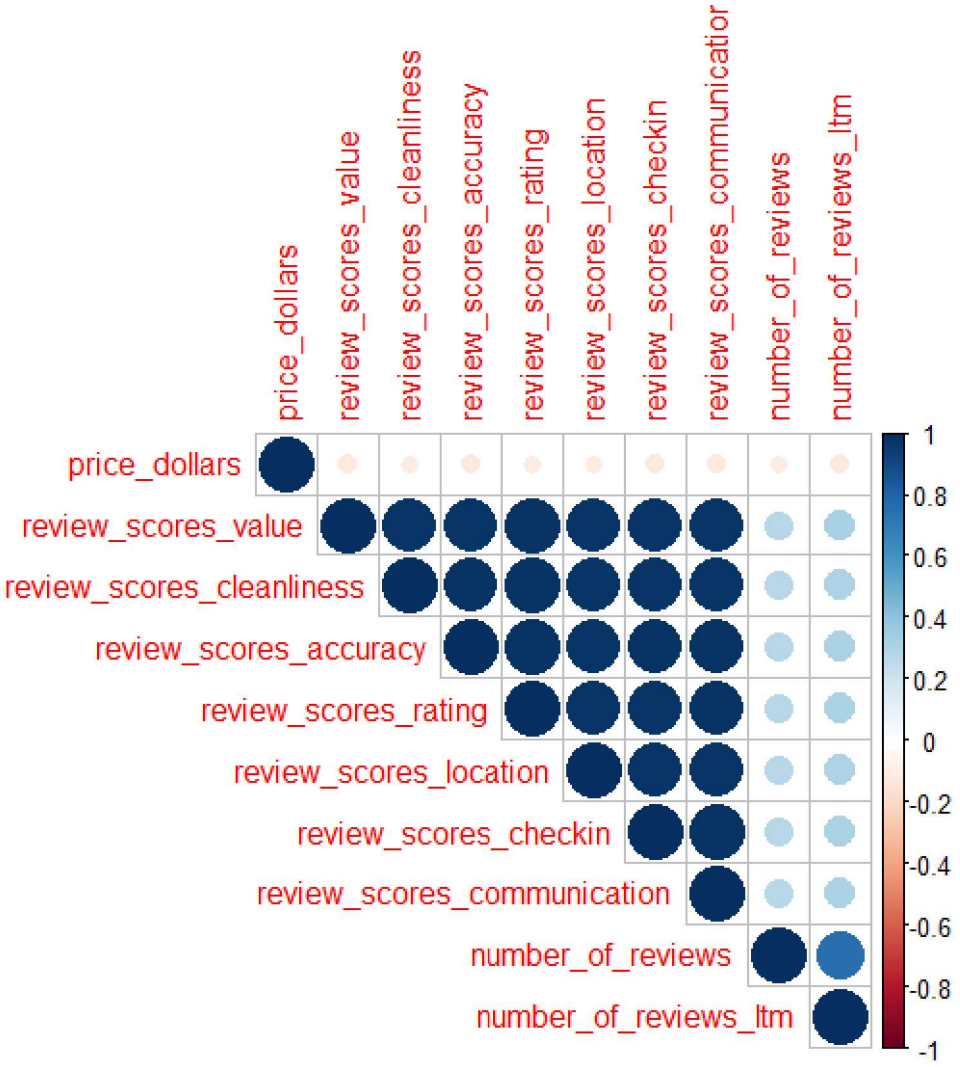

Figura A.7: Correlazioni delle voci di revisione relativamente al 2019

Tabella A.8: Correlazioni delle variabili di conteggio nel 2020

	price_dollars	calculated_host_listings_count	calculated_host_listings_count_entire_homes	calculated_host_listings_count_private_rooms	calculated_host_listings_count_shared_rooms
price_dollars	1.000	0.198	0.219	-0.062	-0.016
calculated_host_listings_count	0.198	1.000	0.969	0.054	-0.005
calculated_host_listings_count_entire_homes	0.219	0.969	1.000	-0.114	-0.047
calculated_host_listings_count_private_rooms	-0.062	0.054	-0.114	1.000	0.134
calculated_host_listings_count_shared_rooms	-0.016	-0.004	-0.047	0.134	1.000

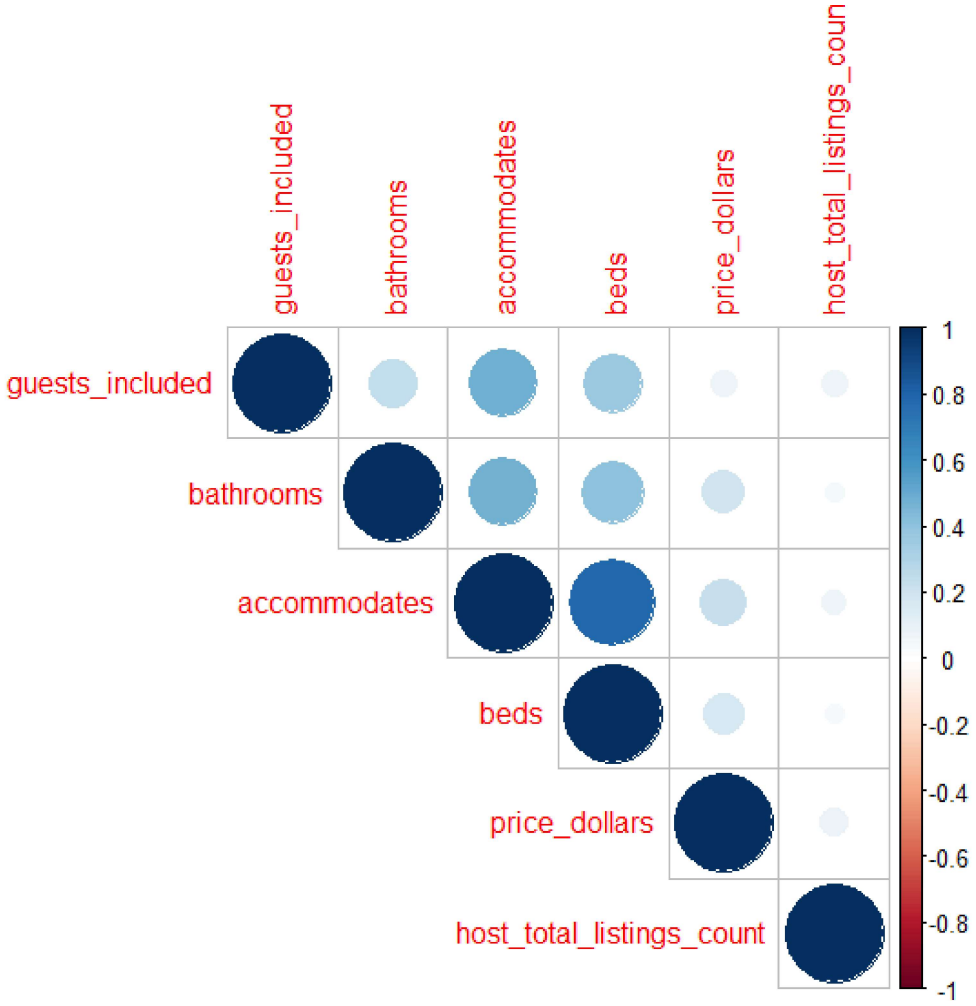

Figura A.8: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2020

Tabella A.9: Correlazioni delle voci di prezzo nel 2020

	price_dollars	weekly_price_dollars	monthly_price_dollars	extra_people_dollars	security_deposit_dollars	cleaning_fee_dollars
price_dollars	1.000	0.997	0.997	0.126	0.320	0.470
weekly_price_dollars	0.997	1.000	0.997	0.125	0.319	0.469
monthly_price_dollars	0.997	0.997	1.000	0.124	0.317	0.467
extra_people_dollars	0.126	0.125	0.124	1.000	0.111	0.145
security_deposit_dollars	0.320	0.319	0.317	0.111	1.000	0.316
cleaning_fee_dollars	0.470	0.469	0.467	0.145	0.316	1.000

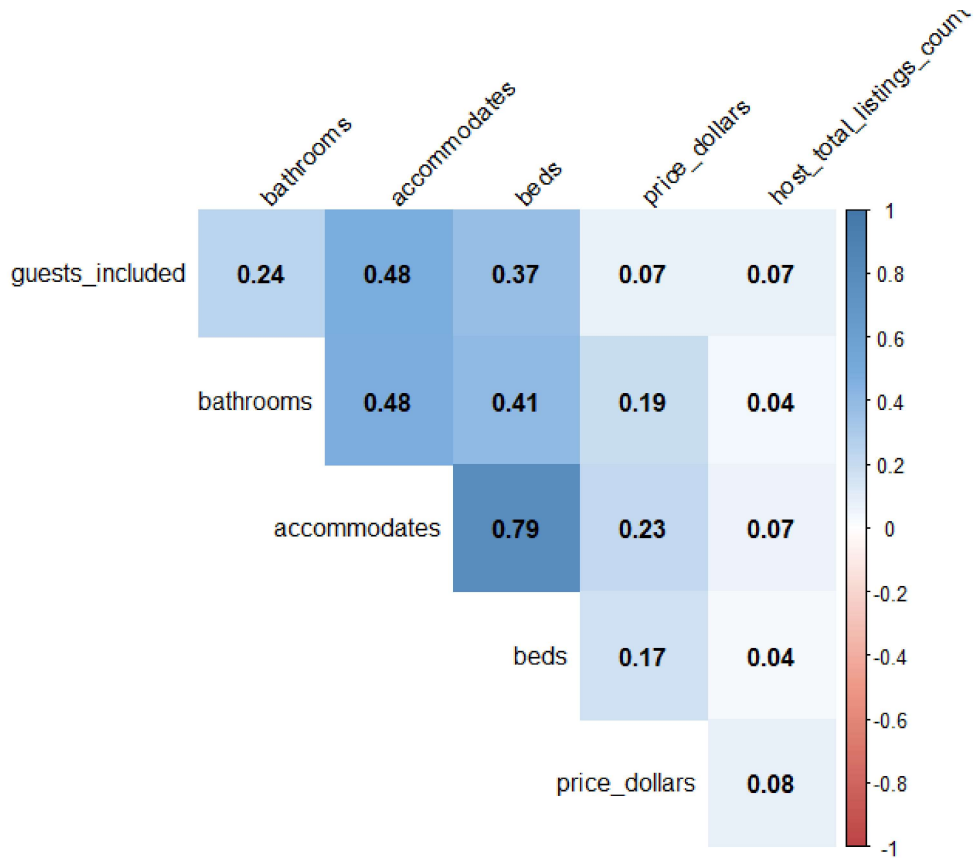

Figura A.9: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2020

Tabella A.10: Correlazioni delle caratteristiche degli "host" nel 2020

	price_dollars	host_is_superhost	host_identity_verified	host_total_listings_count
price_dollars	1.000	-0.038	-0.024	0.083
host_is_superhost	-0.038	1.000	0.143	-0.106
host_identity_verified	-0.024	0.143	1.000	-0.039
host_total_listings_count	0.083	-0.106	-0.039	1.000

Tabella A.11: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2021

	price_dollars	host_total_listings_count	accommodates	beds
price_dollars	1.000	0.010	0.199	0.155
host_total_listings_count	0.010	1.000	0.045	0.003
accommodates	0.199	0.045	1.000	0.754
beds	0.155	0.003	0.754	1.000

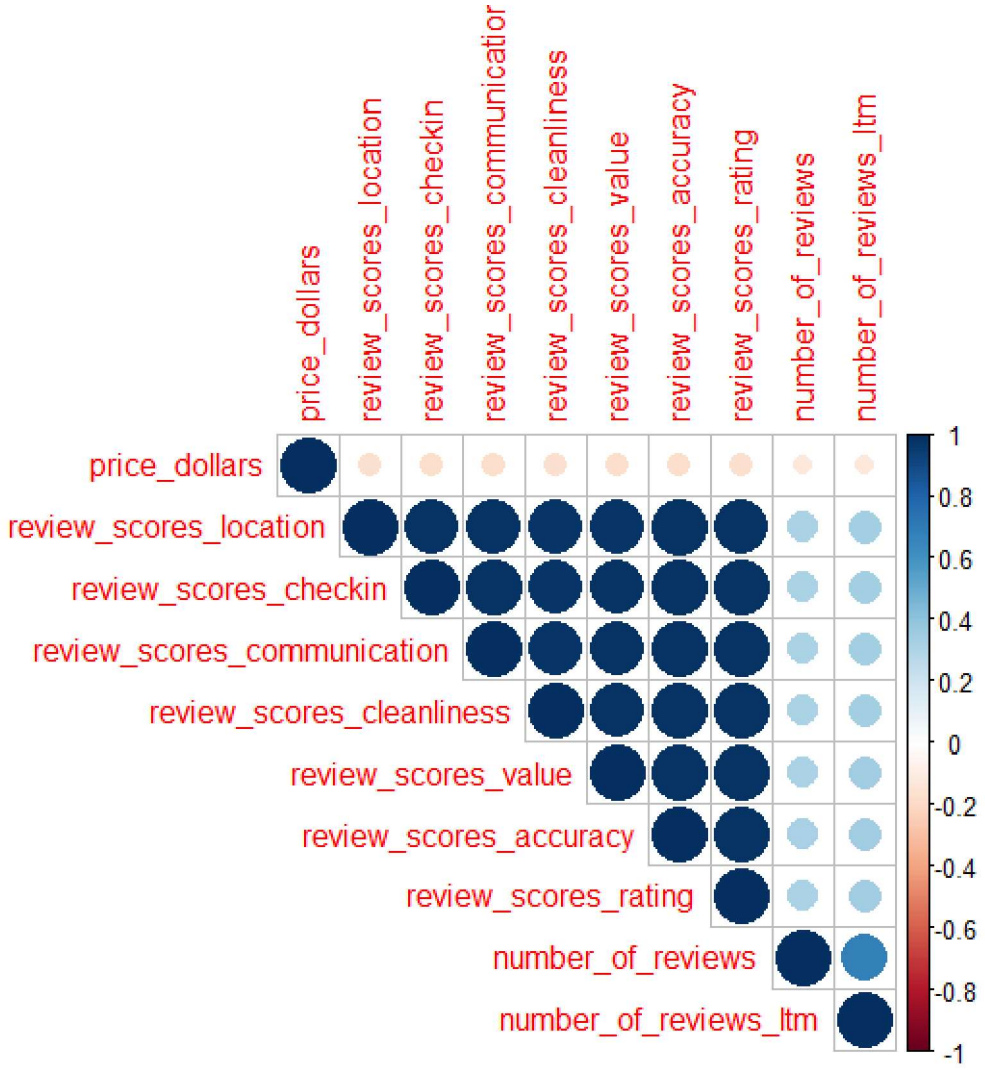

Figura A.10: Correlazioni delle voci di revisione relativamente al 2020

Tabella A.12: Correlazioni delle voci di revisione relativamente al 2021

	price_dollars	number_of_reviews	number_of_reviews_ltm	review_scores_accuracy	review_scores_checkin	review_scores_cleanliness	review_scores_communication	review_scores_location	review_scores_rating
price_dollars	1.000	-0.070	-0.058	-0.086	-0.086	-0.083	-0.088	-0.083	-0.075
number_of_reviews	-0.070	1.000	0.431	0.331	0.326	0.328	0.327	0.316	0.329
number_of_reviews_ltm	-0.056	0.431	1.000	0.213	0.213	0.209	0.214	0.211	0.216
review_scores_accuracy	-0.086	0.331	0.213	1.000	0.990	0.990	0.989	0.988	0.991
review_scores_checkin	-0.086	0.326	0.213	0.990	1.000	0.985	0.992	0.988	0.987
review_scores_cleanliness	-0.083	0.328	0.209	0.990	0.985	1.000	0.986	0.987	0.989
review_scores_communication	-0.088	0.327	0.214	0.989	0.992	0.986	1.000	0.989	0.988
review_scores_location	-0.083	0.316	0.211	0.988	0.988	0.987	0.989	1.000	0.985
review_scores_rating	-0.075	0.329	0.216	0.991	0.987	0.989	0.988	0.985	1.000
review_scores_value	-0.087	0.330	0.218	0.991	0.988	0.989	0.988	0.988	0.991

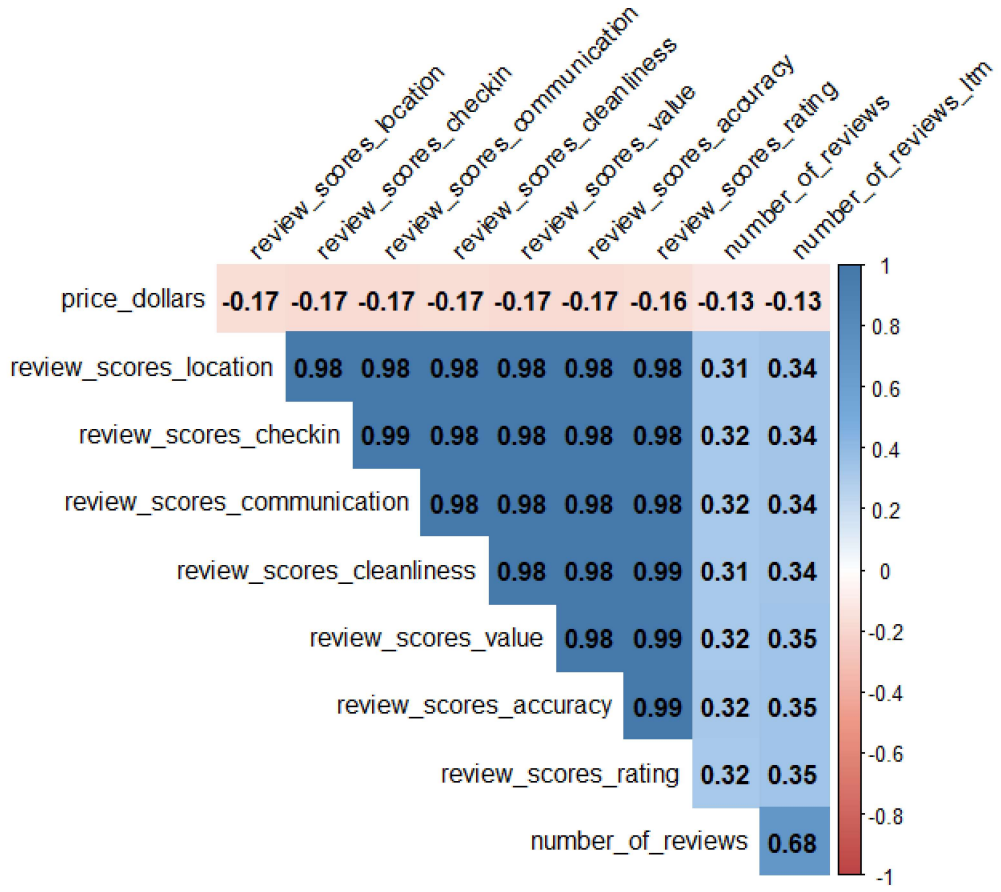

Figura A.11: Correlazioni delle voci di revisione relativamente al 2020

Tabella A.13: Correlazioni delle variabili di conteggio nel 2021

	price_dollars	calculated_host_listings_count	calculated_host_listings_count_entire_homes	calculated_host_listings_count_private_rooms	calculated_host_listings_count_shared_rooms
price_dollars	1.000	0.081	0.084	-0.025	0.021
calculated_host_listings_count	0.081	1.000	0.954	0.344	-0.017
calculated_host_listings_count_entire_homes	0.084	0.954	1.000	0.109	-0.031
calculated_host_listings_count_private_rooms	-0.025	0.344	0.109	1.000	-0.009
calculated_host_listings_count_shared_rooms	0.021	-0.017	-0.031	-0.009	1.000
host_total_listings_count	0.010	0.307	0.310	0.043	-0.010

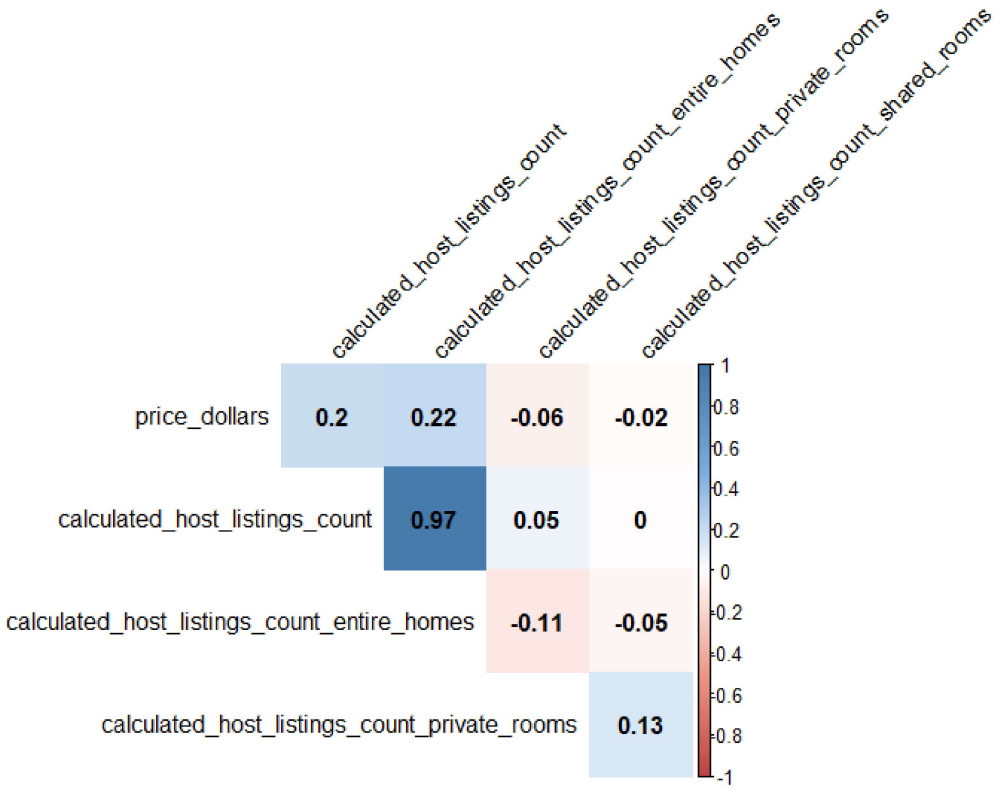

Figura A.12: Correlazioni delle variabili di conteggio nel 2020

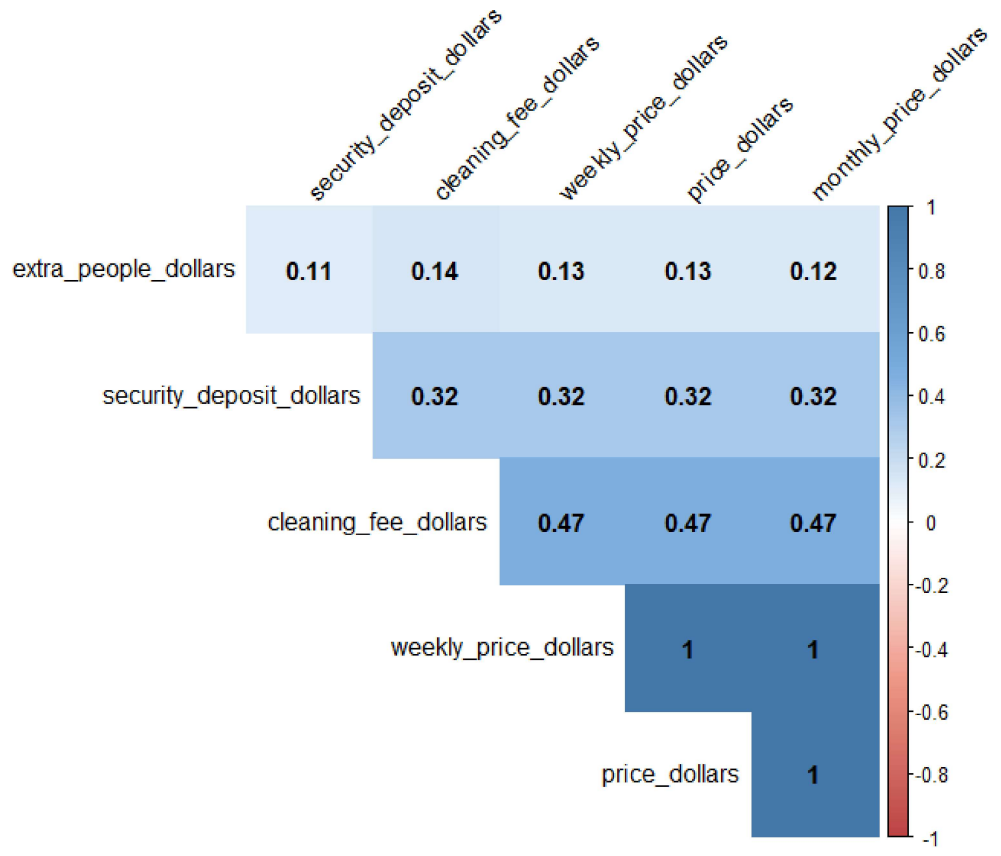

Figura A.13: Correlazioni delle voci di prezzo nel 2020

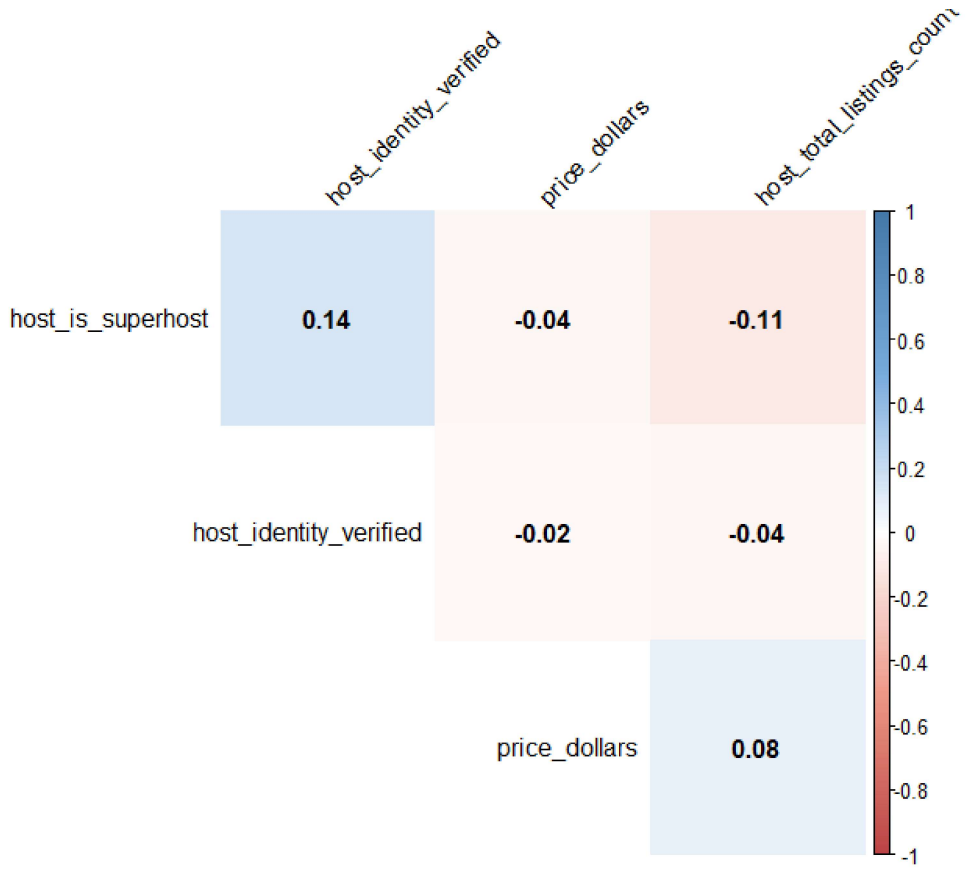

Figura A.14: Correlazioni delle caratteristiche degli "host" nel 2020

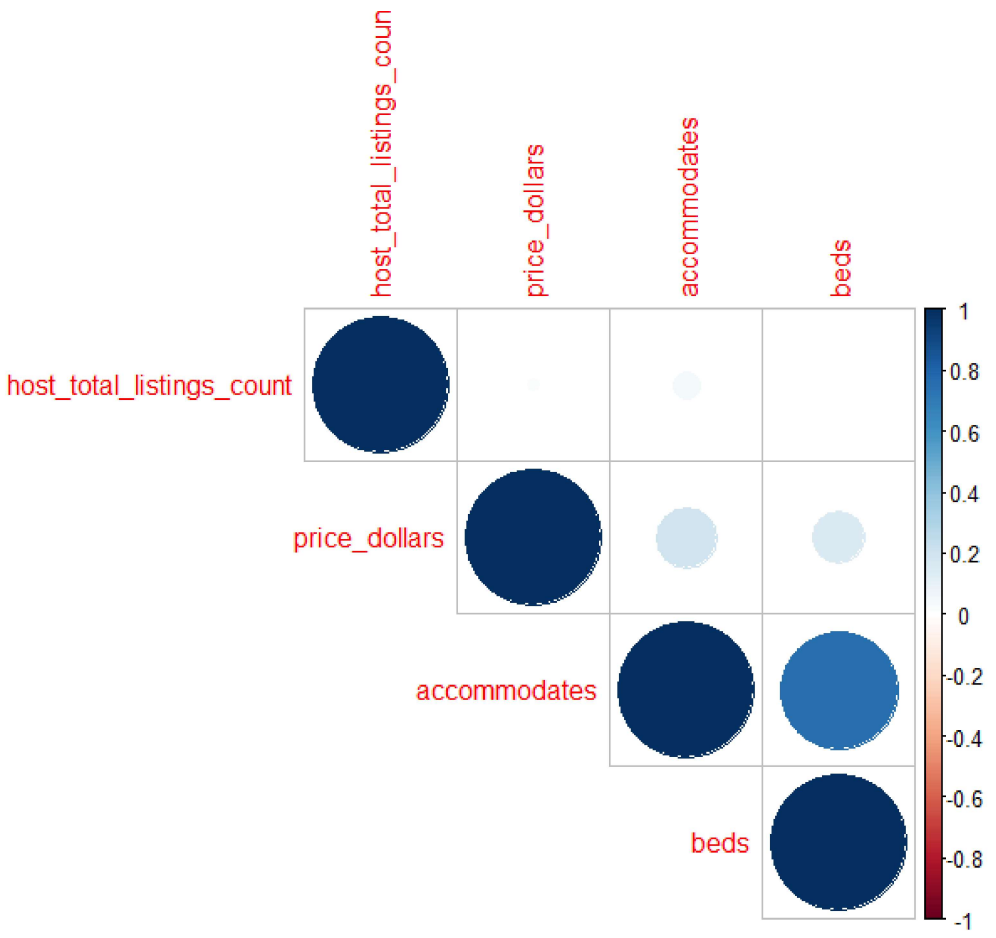

Figura A.15: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2021

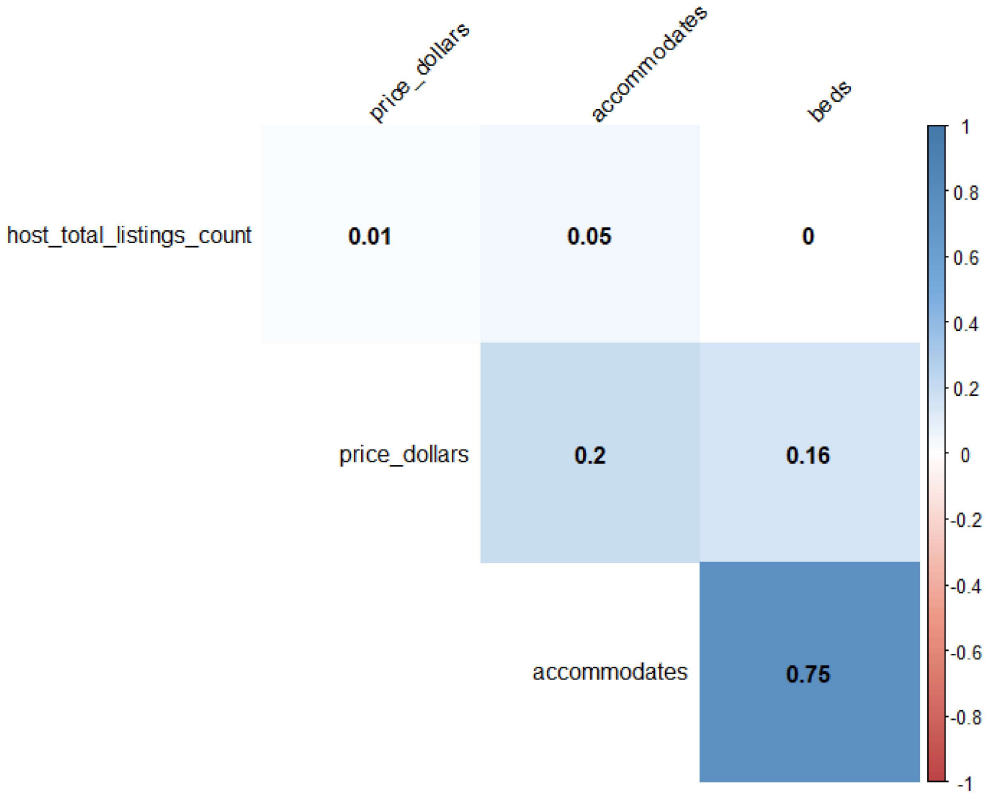

Figura A.16: Correlazioni tra il prezzo e le caratteristiche dell'alloggio nel 2021

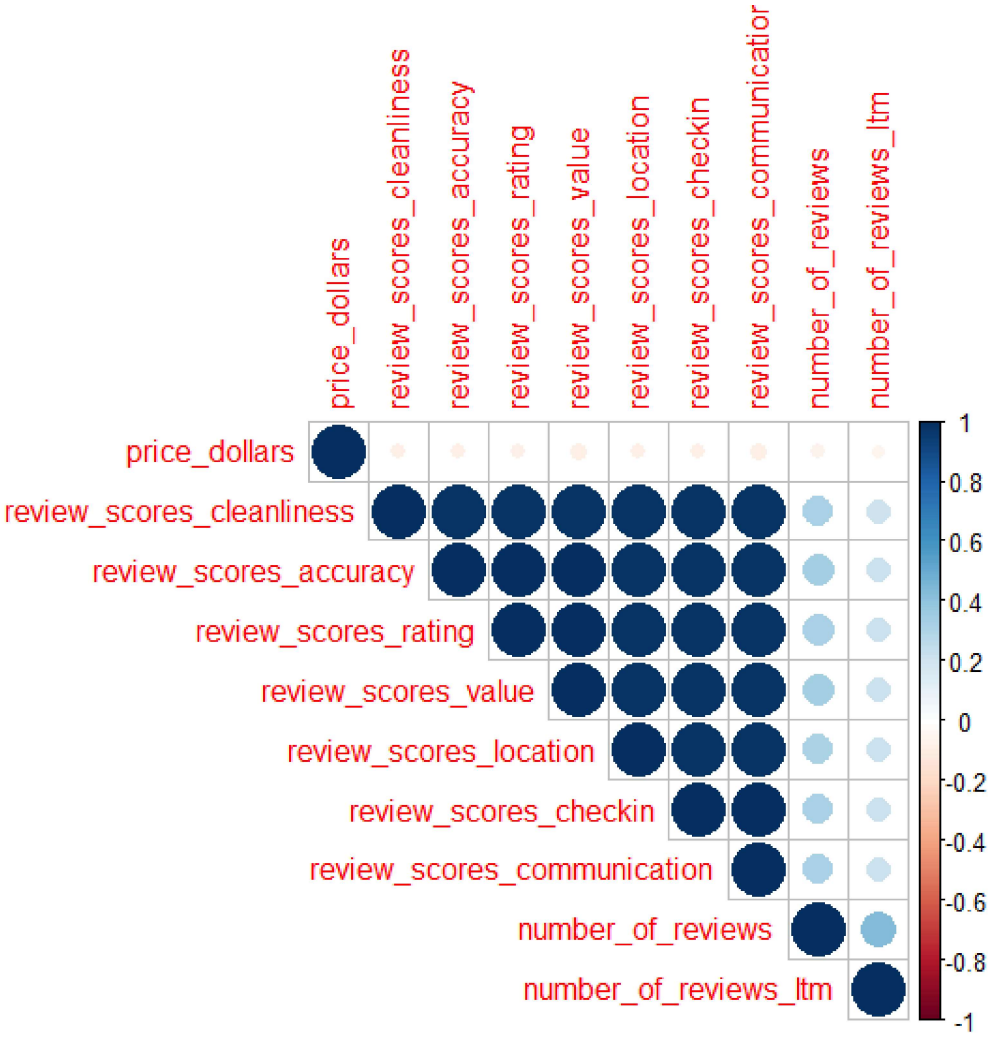

Figura A.17: Correlazioni delle voci di revisione relativamente al 2021

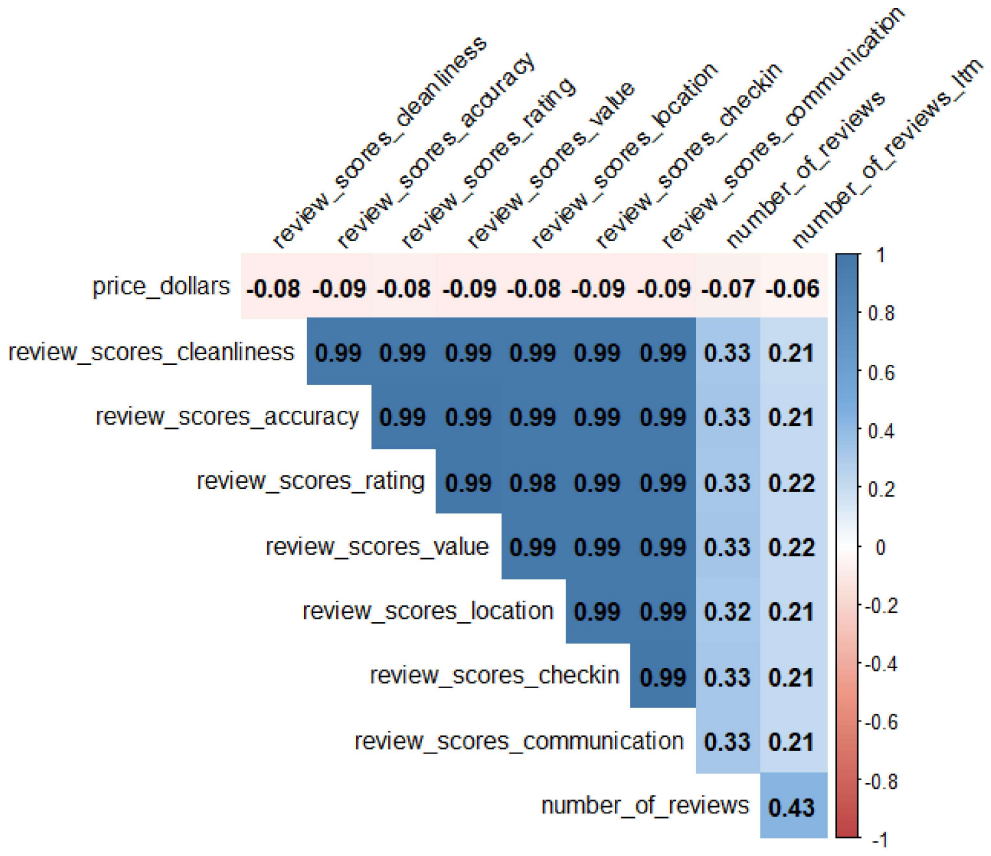

Figura A.18: Correlazioni delle voci di revisione relativamente al 2021

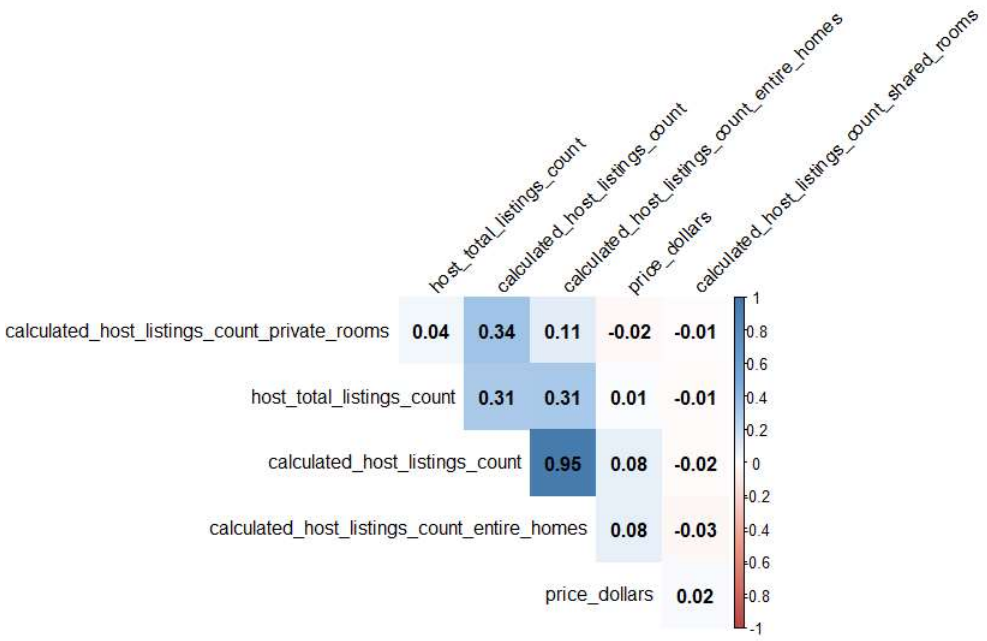

Figura A.19: Correlazioni delle variabili di conteggio nel 2021

Appendice B

Ulteriori grafici utilizzati per le analisi descrittive

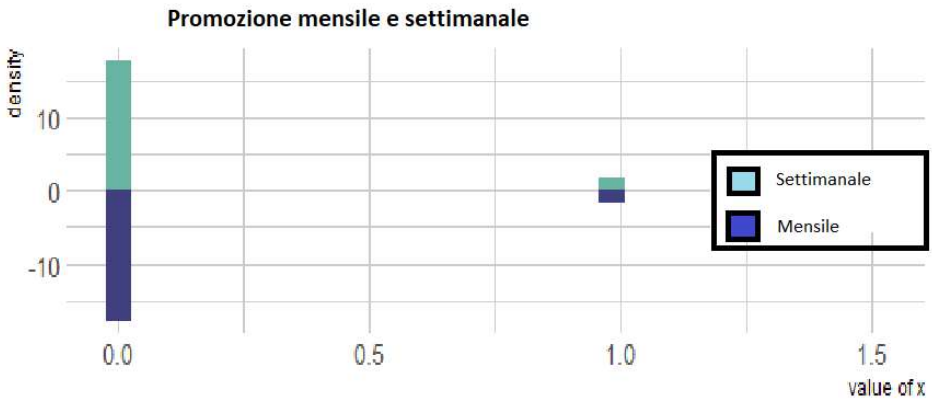

Figura B.1: Promozione mensile e settimanale

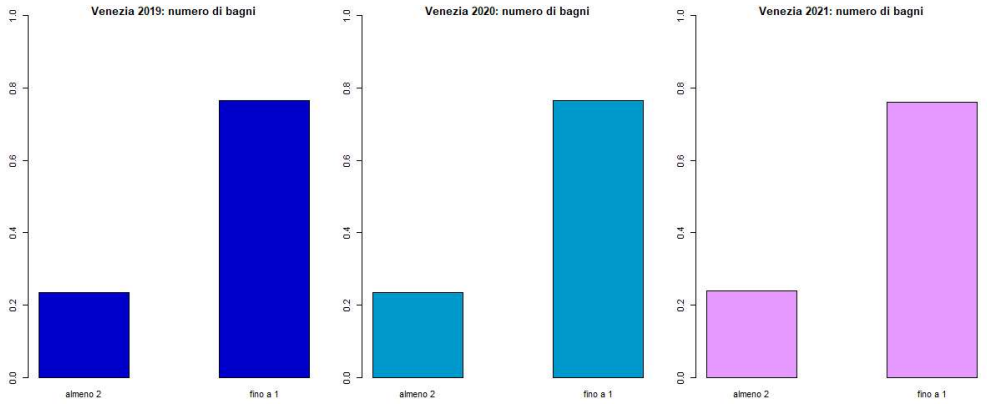

Figura B.2: Numero di bagni

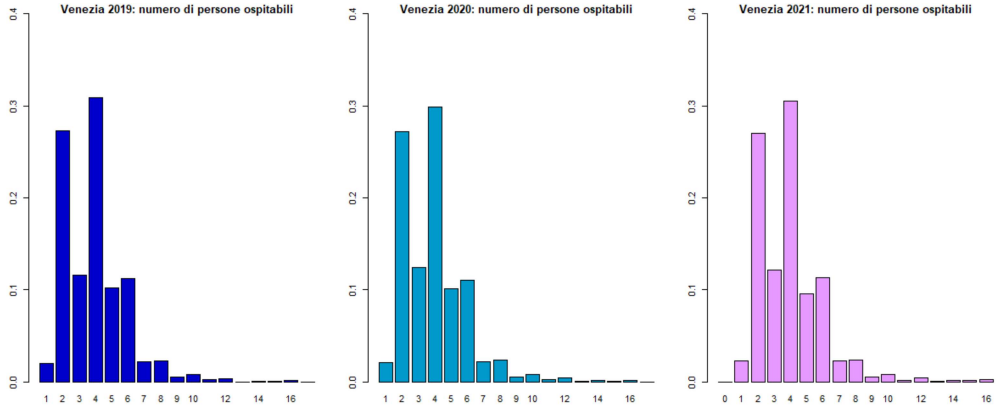

Figura B.3: Numero di persone

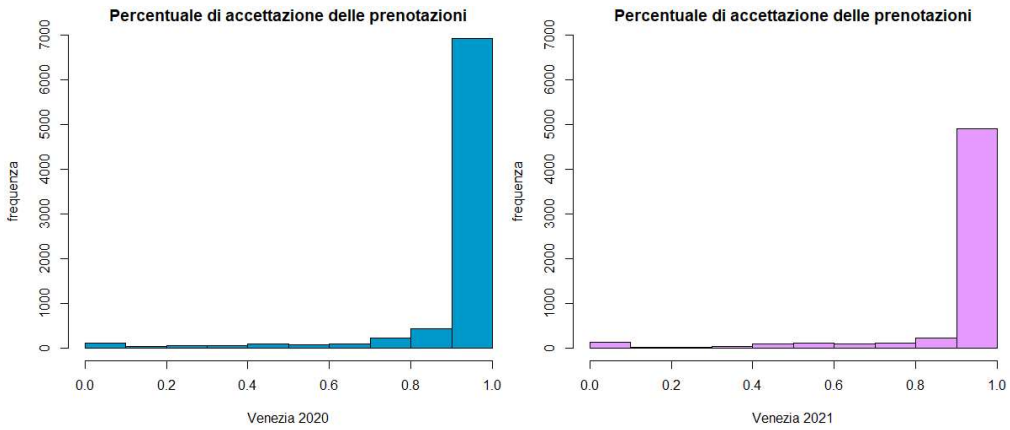

Figura B.4: Percentuale di accettazione delle prenotazioni

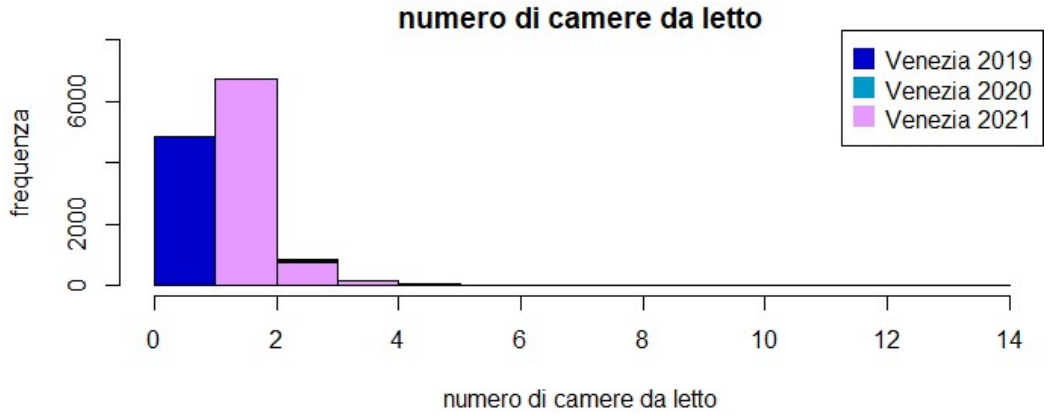

Figura B.5: Numero di camere da letto

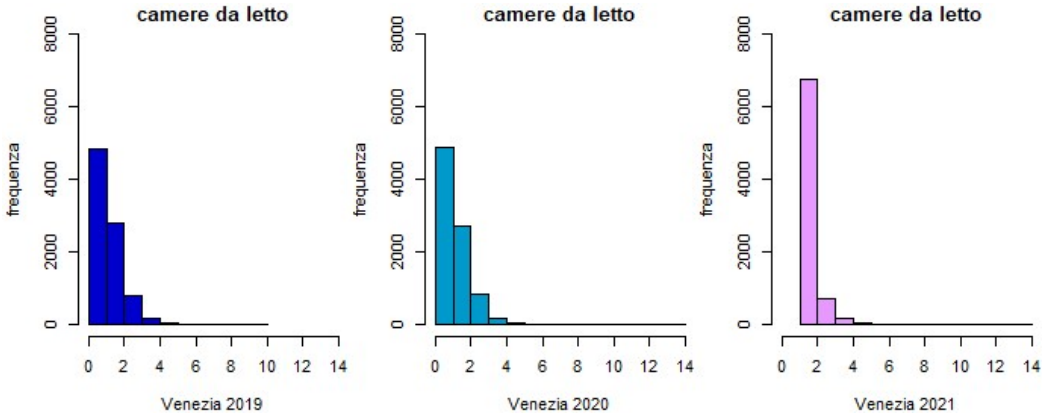

Figura B.6: Numero di camere da letto

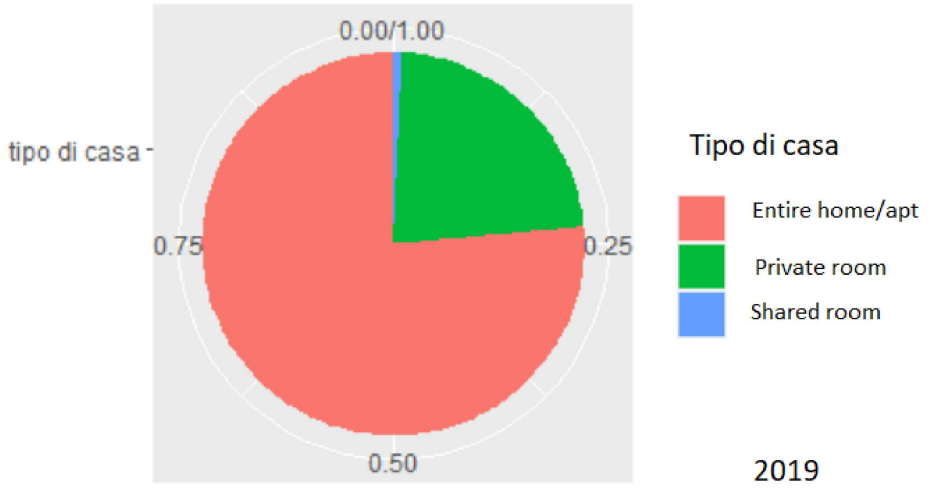

Figura B.7: Tipo di casa nel 2019

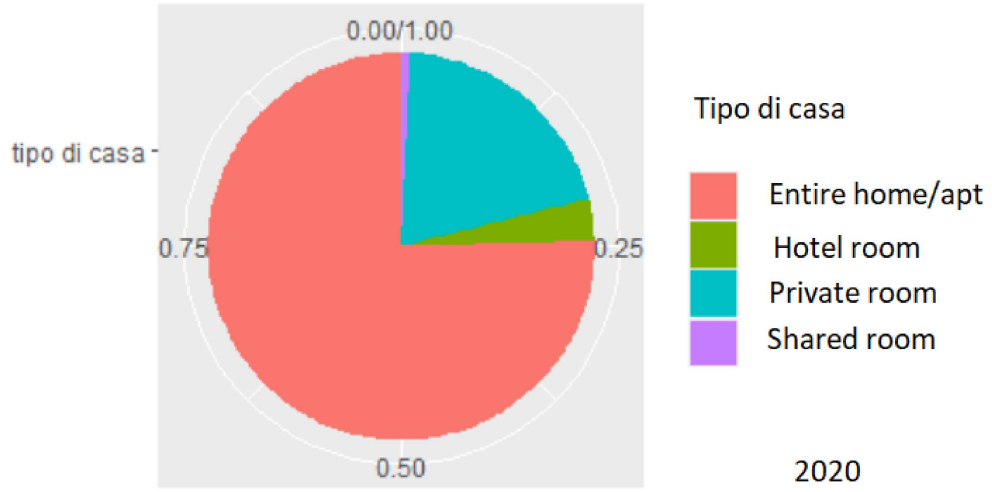

Figura B.8: Tipo di casa nel 2020

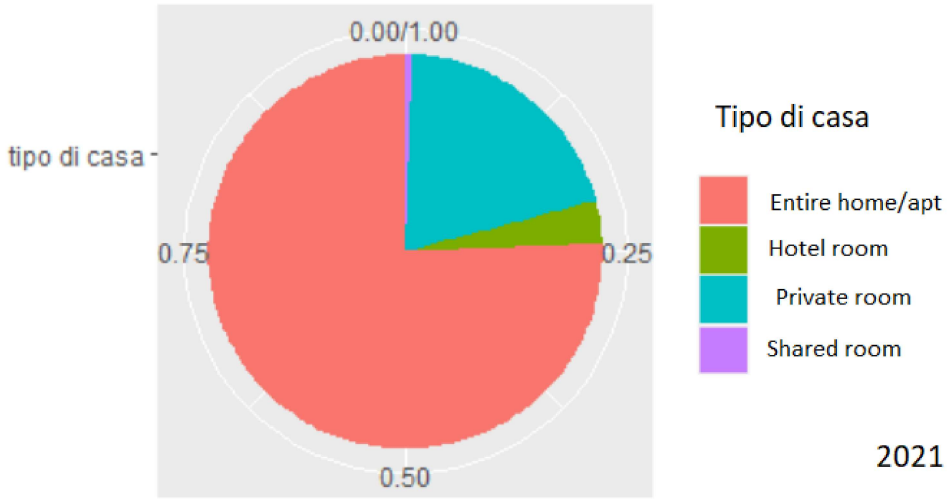

Figura B.9: Tipo di casa nel 2021

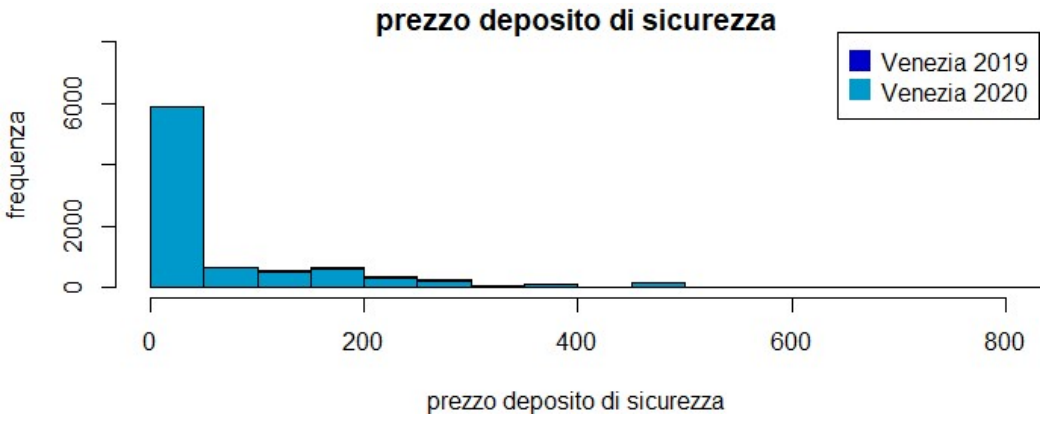

Figura B.10: Prezzo del deposito di sicurezza

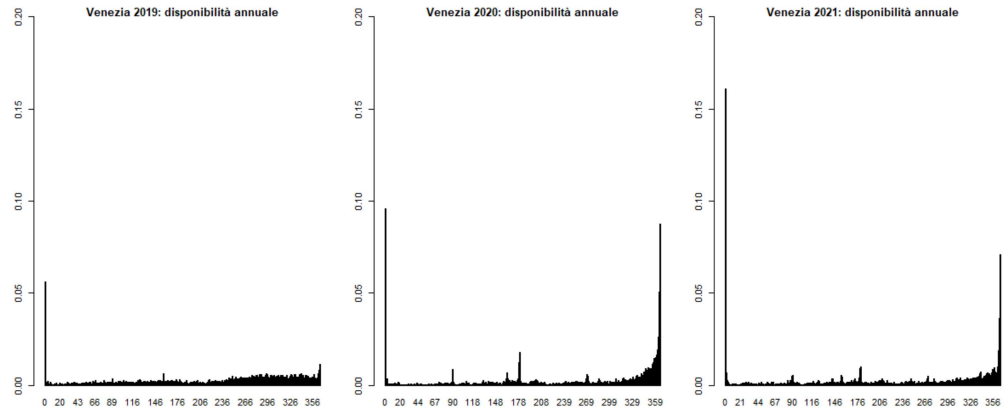

Figura B.11: Disponibilità annuale degli annunci

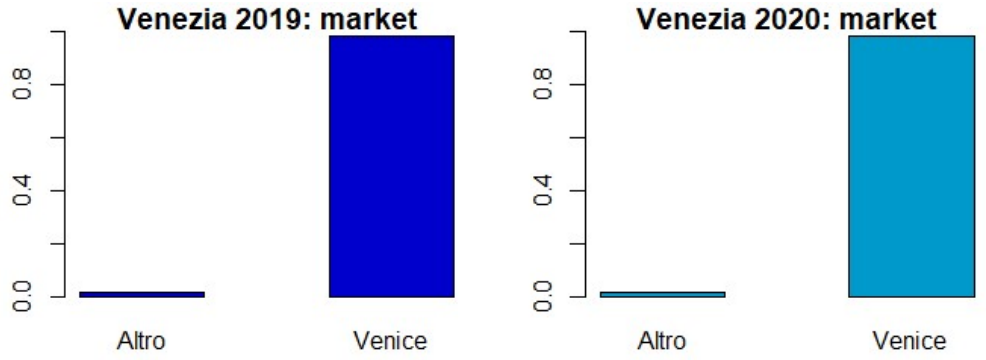

Figura B.12: Mercato di riferimento

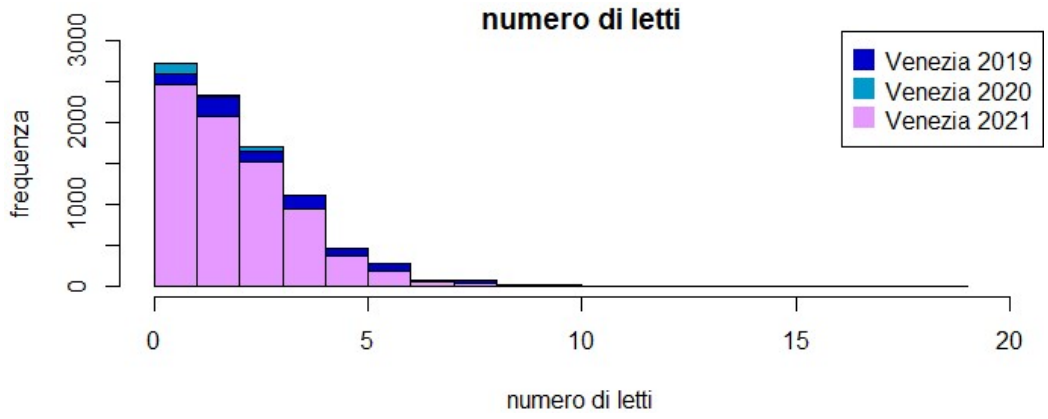

Figura B.13: Numero di letti per annuncio

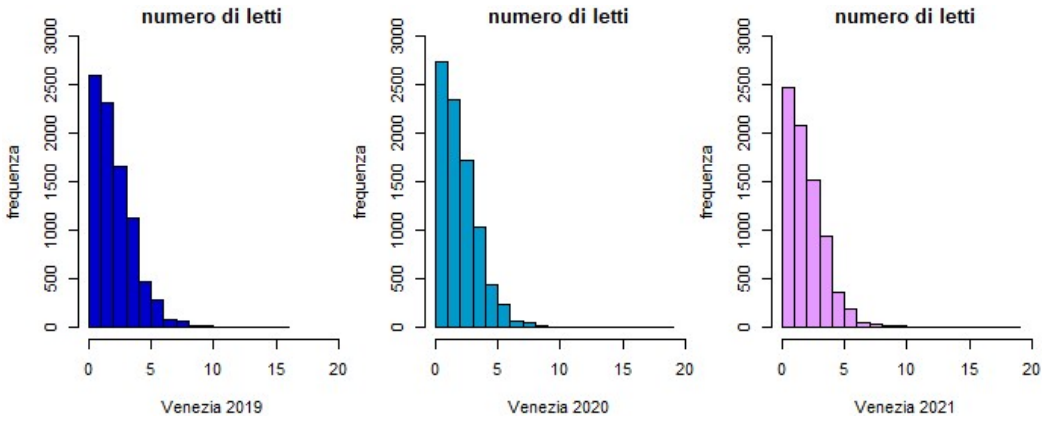

Figura B.14: Numero di letti per annuncio

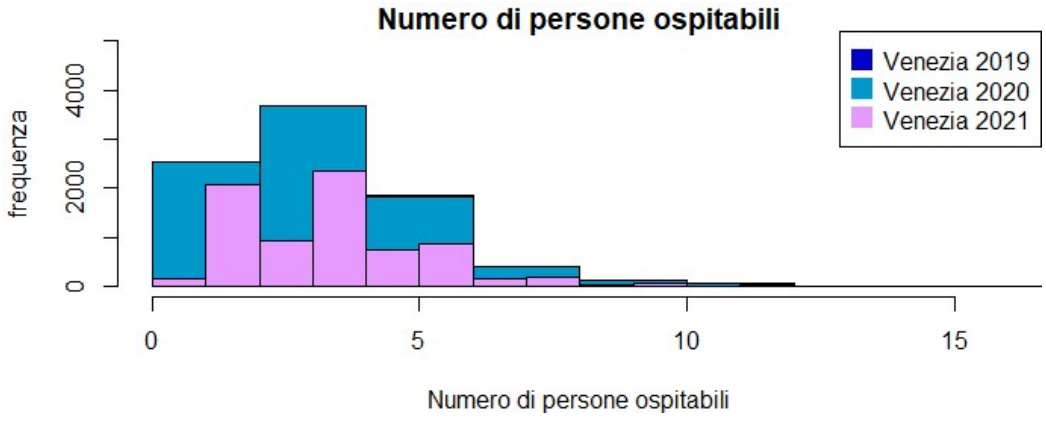

Figura B.15: Numero di persone ospitabili

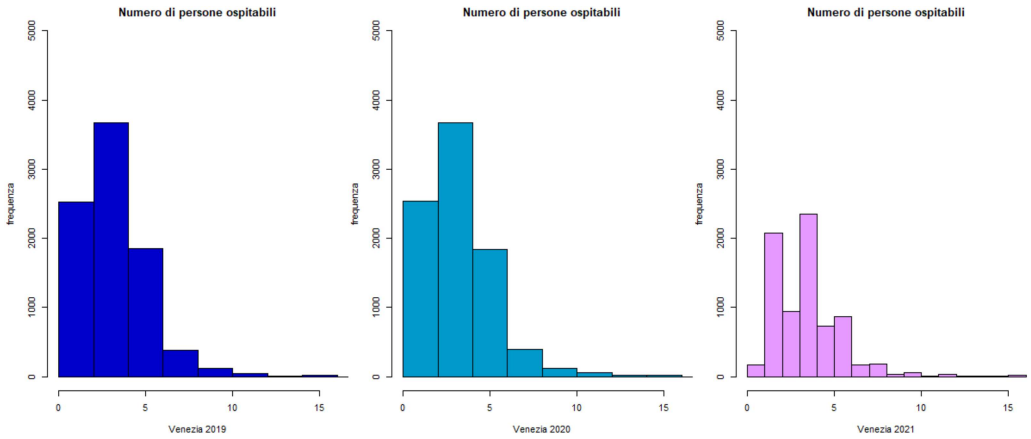

Figura B.16: Numero di persone ospitabili

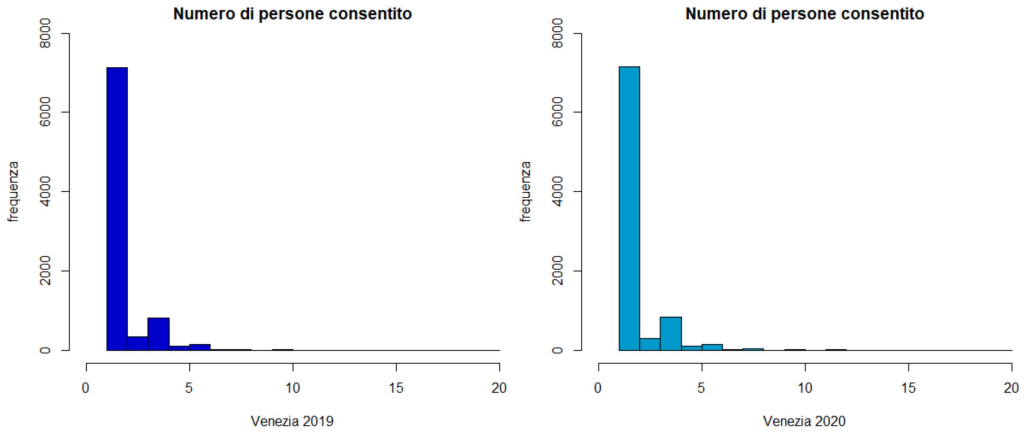

Figura B.17: Numero di persone consentito

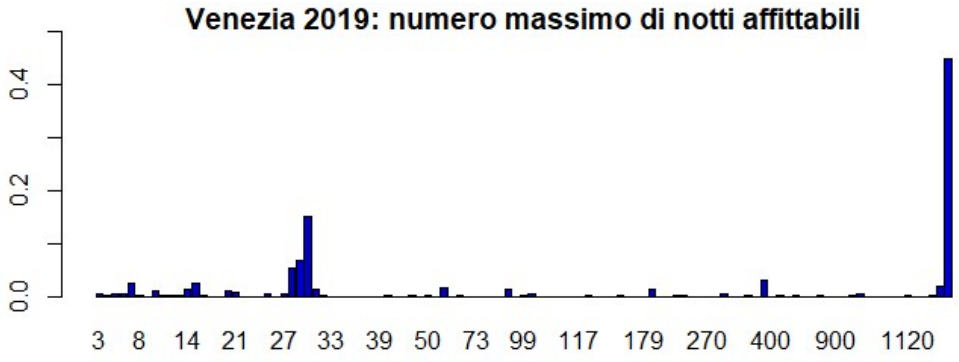

Figura B.18: Numero massimo di notti affittabili nel 2019

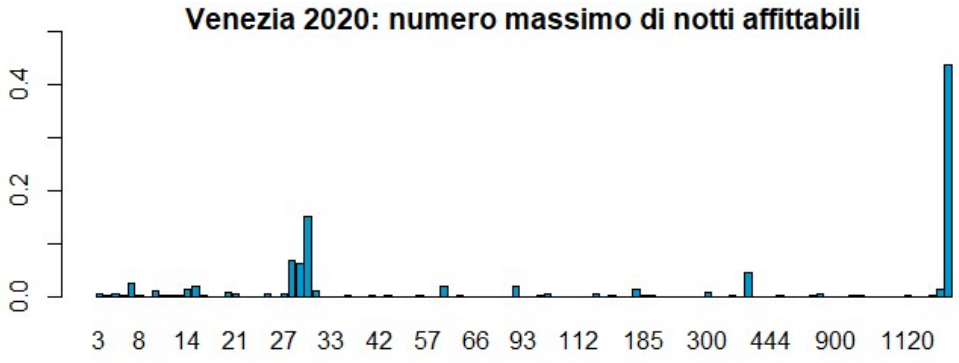

Figura B.19: Numero massimo di notti affittabili nel 2020

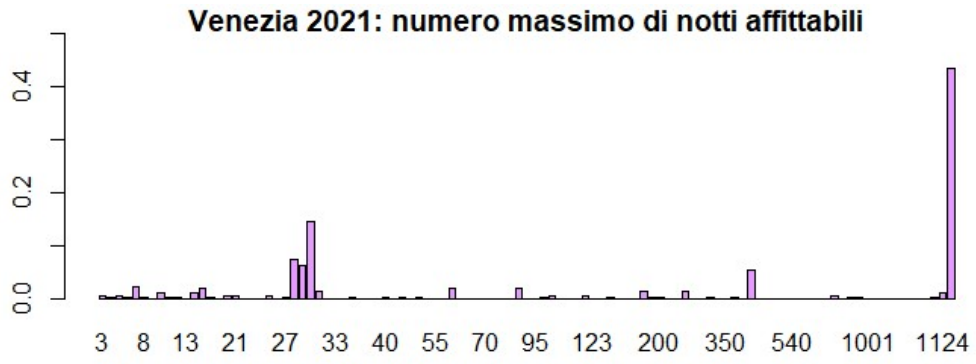

Figura B.20: Numero massimo di notti affittabili nel 2021

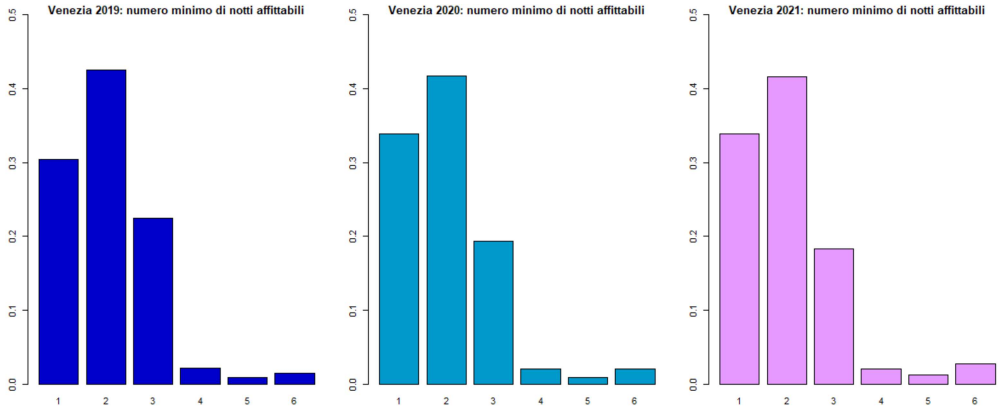

Figura B.21: Numero minimo di notti affittabili

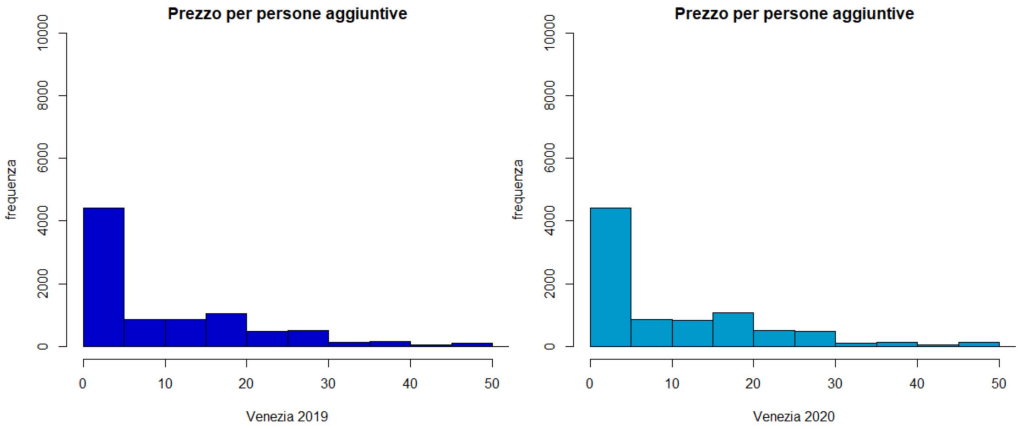

Figura B.22: Prezzo delle persone aggiuntive

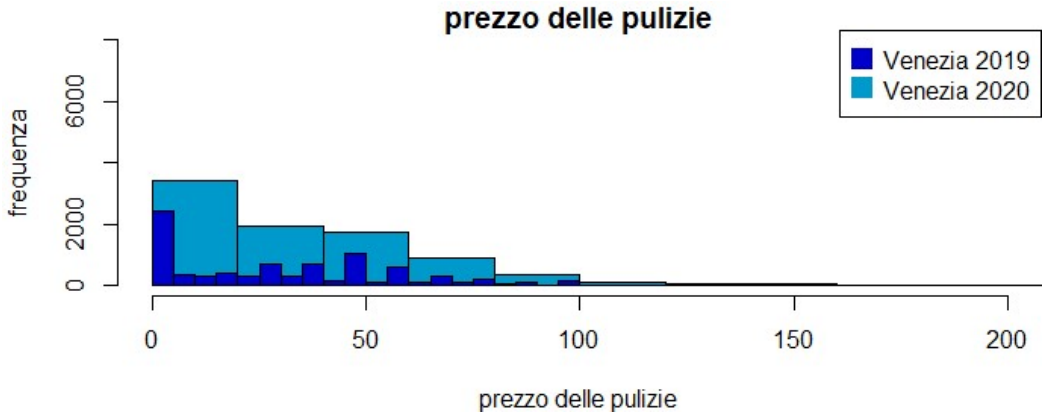

Figura B.23: Prezzo delle pulizie

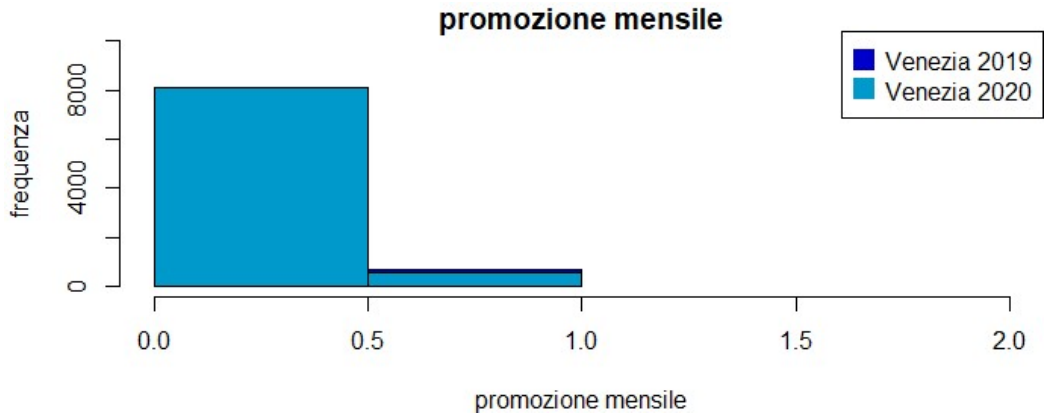

Figura B.24: Promozione mensile

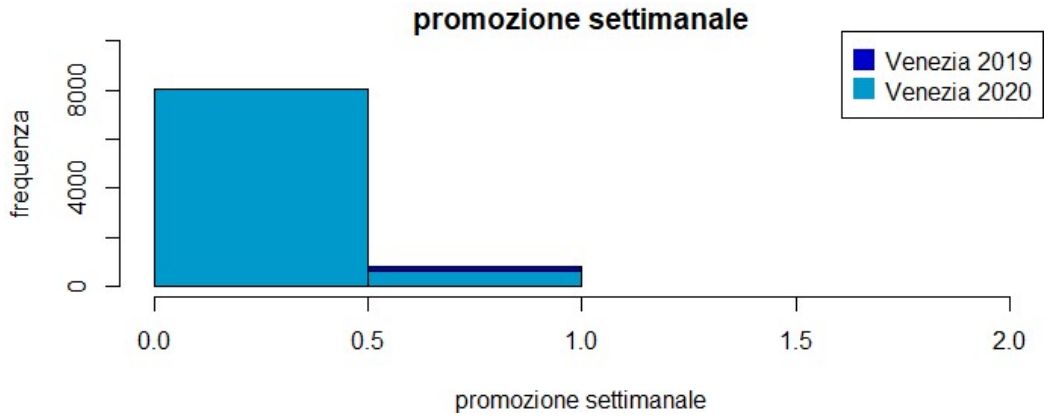

Figura B.25: Promozione settimanale

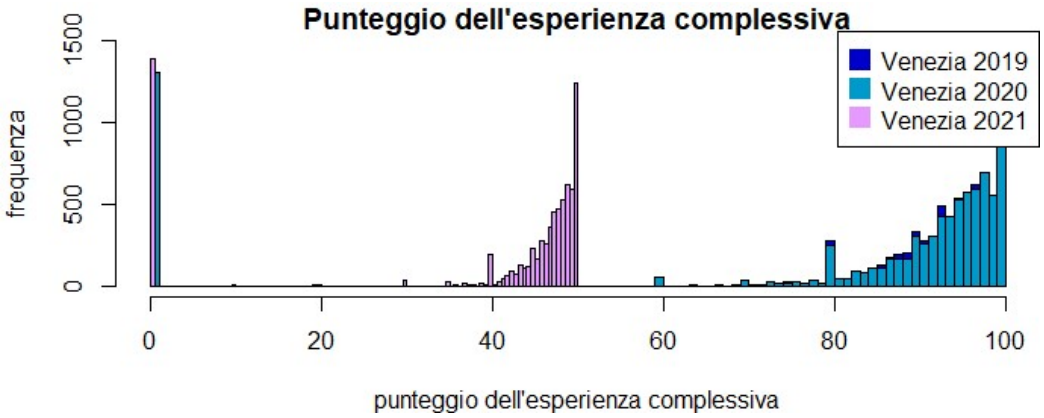

Figura B.26: Punteggio dell'esperienza complessiva

Figura B.27: Andamento del prezzo rispetto alle recensioni dell'esperienza nel 2019

Figura B.28: Andamento del prezzo rispetto alle recensioni dell'esperienza nel 2020

Figura B.29: Andamento del prezzo rispetto alle recensioni dell'esperienza nel 2021

Appendice C

Codice

Si riporta di seguito il codice R che si riferisce ai modelli oggetto di approfondimento in questo elaborato. Si fa riferimento al dataset relativo al 2019. Gli stessi comandi sono stati applicati anche agli altri dataset.

Codice reletivo ai modelli approfonditi nel terzo capitolo

```
#MODELLO LINEARE
str(sss$price_dollars)
l<-lm(price_dollars~.,data=sss)
summary(l)
pl=pmax(predict(l,newdata=vvv),0.5)
mse_modello_lineare<-mean((pl-vvv$price_dollars)^2)

#MODELLO GRADIENT BOOSTING
library (gbm)
set.seed(1234)
boost.price1=gbm(price_dollars ~ ., data=sss,
                 distribution="gaussian", n.trees=5000, interaction.depth=4)
#
#grafico degli errori
plot(boost.price1$train.error, type="l")
#
```

```
summary(boost.price1, las=1, cBar=10,
main="Importanza delle variabili - anno 2019")
#cBar definisce quante variabili disegnare
#
# previsione su test set
yhat.boost1=predict(boost.price1 ,newdata=vvv,n.trees=1:5000)
err1 = apply(yhat.boost1,2,function(pred) mean((vvv$price_dollars-pred)^2))
plot(err1, type="l")
best=which.min(err1)
par(
  mfrow=c(3,3)
)
names(data)
plot(boost.price1, i.var=7, n.trees = best)
plot(boost.price1, i.var=19, n.trees = best)
plot(boost.price1, i.var=26, n.trees = best)
plot(boost.price1, i.var=11, n.trees = best)
plot(boost.price1, i.var=13, n.trees = best)
plot(boost.price1, i.var=20, n.trees = best)

min(err1)
```

Codice reletivo ai modelli approfonditi nel quarto capitolo

```
#MODELLO LINEARE
y <- data$price_dollars #logaritmo del prezzo
x <- data$availability_365
zona<- data$neighbourhood_cleansed
x1<-data$accommodates
x2<-data$number_of_reviews
x3<-data$room_type
str(zona)
str(x3)
#modello lineare
```

```
l<-lm(y~zona+x+x1+x2+x3)
summary(l)
AIC(l)

#MODELLO A INTERCETTA VARIABILE CON QUATTRO PREDITTORI
y <- data$price_dollars #logaritmo del prezzo
zona<- data$neighbourhood_cleansed
x <- data$availability_365
x1<-data$accommodates
x2<-data$number_of_reviews
x3<-data$room_type
library(lme4)
M1_1 <- lmer(y ~ x + x1 + x2 + x3 + (1 | zona))
summary(M1_1)
table(x3)
AIC(M1_1)
coef(M1_1)
# effetti fissi e casuali
fixef(M1_1)
ranef(M1_1)

#MODELLO LINEARE (COMPLETE POOLING)
l1<-lm(y~x2)
summary(l1)
coef(l1)
AIC(l1)

#MODELLO A INTERCETTA VARIABILE CON UN PREDITTORE
library(lme4)
M1_1 <- lmer(y ~  x2 + (1 | zona))
summary(M1_1)
AIC(M1_1)
coef(M1_1)
# effetti fissi e casuali
fixef(M1_1)
ranef(M1_1)
```

Codice reletivo ai grafici studiati nel quarto capitolo

```
#GRAFICO
a_hat <- coef(l1)[1]                # primo elemento intercetta
b_hat <- coef(l1)[2]                # secondo elemento pendenza
a_hat_M1_1 <- coef(M1_1)$zona[,1]               # primo elemento intercetta
b_hat_M1_1 <- coef(M1_1)$zona[,2]               # secondo elemento pendenza

par(mfrow = c(2,3))
plot(x2[zona=="Cannaregio"], y[zona=="Cannaregio"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main ="Cannaregio")
abline(a_hat_M1_1[15], b_hat_M1_1[15],  col = "red")
abline(l1,  col = "blue")
plot(x2[zona=="Castello"], y[zona=="Castello"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main="Castello")
abline(a_hat_M1_1[18], b_hat_M1_1[18],  col = "red")
abline(l1,  col = "blue")
plot(x2[zona=="Dorsoduro"], y[zona=="Dorsoduro"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main="Dorsoduro")
abline(a_hat_M1_1[22], b_hat_M1_1[22],  col = "red")
abline(l1,  col = "blue")
plot(x2[zona=="San Marco"], y[zona=="San Marco"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main="San Marco")
abline(a_hat_M1_1[46], b_hat_M1_1[46],  col = "red")
abline(l1,  col = "blue")
plot(x2[zona=="San Polo"], y[zona=="San Polo"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main="San Polo")
abline(a_hat_M1_1[48], b_hat_M1_1[48],  col = "red")
abline(l1,  col = "blue")
plot(x2[zona=="Santa Croce"], y[zona=="Santa Croce"],
xlab = "number_of_reviews", ylab = "prezzo per notte", main="Santa Croce")
abline(a_hat_M1_1[52], b_hat_M1_1[52],  col = "red")
abline(l1,  col = "blue")
```

Ringraziamenti

Desidero ringraziare la Professoressa Mariangela Guidolin, con la quale ho scritto questa tesi di Laurea Magistrale nonché la relazione finale del percorso di Laurea Triennale. Grazie per aver deciso di accompagnarmi nell'esecuzione di entrambi gli elaborati, dimostrandomi l'importanza del rapporto umano tra docente e studente, che porta a collaborazioni proficue e durature. Grazie per la disponibilità che ha dimostrato nei miei confronti, per l'aiuto ricevuto e per avermi dato la possibilità di approfondire argomenti di mio interesse.
Desidero ringraziare, inoltre, la mia famiglia, che mi ha supportata e incoraggiata durante tutto il percorso di studi e che ha creduto nella formazione universitaria.
Grazie, infine, a tutti coloro che hanno vissuto con me questi importanti anni della mia vita.