ELABORATO FINALE

METODI STATISTICI PER LA BUSINESS ANALYSIS A.a. 2024/2025

Mattia Dalla Fontana 908793

Titolo dell'elaborato

Analisi del dataset "**New York Airbnb Open Data 2024**"

Sommario

INTRODUZIONE

Il presente elaborato ha come obiettivo l'analisi statistica dei fattori che influenzano i prezzi degli alloggi Airbnb a New York City. I dati utilizzati provengono da un dataset aggiornato al 5 gennaio 2024.
Per ciascun annuncio, il dataset include diverse variabili, tra cui il tipo di alloggio (intero appartamento, stanza privata o condivisa), la posizione geografica (latitudine e longitudine, quartieri),), il prezzo per notte, il numero di camere da letto, bagni e letti disponibili, nonché il punteggio medio delle recensioni ricevute dagli ospiti.
L'analisi parte dall'osservazione dei dati per provare a captare l'esistenza di variabili chiave in grado di spiegare le differenze di prezzo.
Il lavoro si sviluppa attraverso diverse fasi. Inizialmente, viene presentata una panoramica del dataset e viene effettuata una fase di pre-processing, seguita poi da una pulizia di eventuali dati mancanti. Successivamente, viene svolta un'analisi descrittiva per comprendere la distribuzione delle variabili e le loro relazioni con il prezzo. Segue poi la costruzione di un modello predittivo, volto a stimare il prezzo degli alloggi sulla base delle loro caratteristiche. Prima della costruzione del modello viene effettuata un'analisi della correlazione per individuare eventuali problemi di multicollinearità e i predittori più rilevanti.

Obiettivo: Determinare quali fattori influenzano maggiormente il prezzo degli alloggi Airbnb a New York, con l'intento di comprendere le dinamiche di tale mercato.
Previsioni: Qual è il prezzo di un alloggio a New York che presenta determinate caratteristiche chiave?

PRE-PROCESSING E PULIZIA DEI DATI

R ci informa della presenza nel dataset di dati riguardanti 20758 alloggi a New York, con 17 variabili di cui 1 come variabile risposta.
Inizialmente il numero di variabili ammontava a 22 ma sono state eliminate “neighbourhood” e “neighbourhood_group”, sostituite con la creazione di un Factor chiamato “Region” formato da 4 livelli (North, South, East, West), all’interno del quale sono stati inseriti tutti i quartieri di NY, così da facilitare l’analisi. Sono state poi eliminate le variabili “host_name”, “name”, ”host id”, ”id” perché ritenute irrilevanti ai fini dell’analisi. Per “latitude” e “longitude” la logica è quella di tentare prima un’analisi generale basandosi sulle 4 macro aree e, nel caso si vedesse un’influenza rilevante di queste sul prezzo, intraprendere in futuro ulteriori analisi a riguardo.

Variabili quantitative continue- Num:

1. **price**: prezzo per notte dell'appartamento. Indica quanto costa affittare la proprietà per una notte.
2. **reviews_per_month**: numero medio di recensioni ricevute dall'appartamento ogni mese. Aiuta a comprendere la frequenza delle recensioni.
3. **rating**: valutazione media dell'appartamento, che rappresenta la qualità complessiva basata sulle recensioni degli utenti. Questa veniva inserita sotto forma di Factor. È stata modificata in variabile quantitativa continua.
4. **baths**: numero di bagni presenti nell'appartamento. Questa si presentava come Factor. È stata modificata in variabile quantitativa continua. I bagni che prima venivano identificati come livello 0 sono diventati NA e quindi eliminati.

Variabili quantitative discrete- Int:

5. **minimum_nights**: numero minimo di notti richieste per poter affittare l'appartamento.
6. **number_of_reviews**: numero totale di recensioni che l'appartamento ha ricevuto, riflettendo così la popolarità dell'appartamento e la sua visibilità sulla piattaforma.
7. **calculated_host_listings_count**: numero di appartamenti gestiti dallo stesso host.
8. **availability_365**: numero di giorni all'anno nei quali l'appartamento è disponibile per la prenotazione. Valore che va da 0 a 365.
9. **number_of_reviews_ltm**: numero di recensioni ricevute dall'appartamento nell'ultimo mese, il che aiuta a capire l'attività recente dell'appartamento.
10. **beds**: numero totale di letti nell'appartamento. Può includere letti singoli, matrimoniali e altri tipi di letti.
11. **bedroom**: numero di camere da letto nell'appartamento. Rappresenta una delle caratteristiche di capienza dell'immobile. Questa veniva inserita sotto forma di Factor, con numeri che indicavano quante stanze erano presenti. Inoltre, uno dei livelli veniva indicato come “studio”. Per semplificare l’analisi quest’ultimo viene considerato come una stanza e la variabile è stata modificata in quantitativa discreta

Variabili identificate come Factor:

12. **region**: le quattro zone geografiche della città nelle quali sono stati raggruppati i quartieri.
13. **room_type**: Il tipo di stanza offerta nell'appartamento: "Entire home/apt", "Hotel room", "Private room", "Shared room"

14. **last_review**: La data dell'ultima recensione ricevuta dall'appartamento. Indica quando è stata lasciata l'ultima valutazione. Per facilitare l'analisi i dati sono stati raggruppati in tre livelli al posto di 1870, ovvero "2011-2015", "2016-2020","2021-2024"
15. **license**: licenza necessaria per alloggi con tempo massimo di pernottamento sotto i 30 giorni. Raggruppata in un Factor formato da tre livelli: "License", "No license"," Exempt"

Dopo aver modificato le variabili è necessario entrare più nello specifico all'interno del dataset per analizzare i diversi valori mancanti indicati come NA's.
Una volta pulito il dataset secondo le modalità che verrà indicata in ciascuna variabile, il dataset presenta la seguente dimensione:

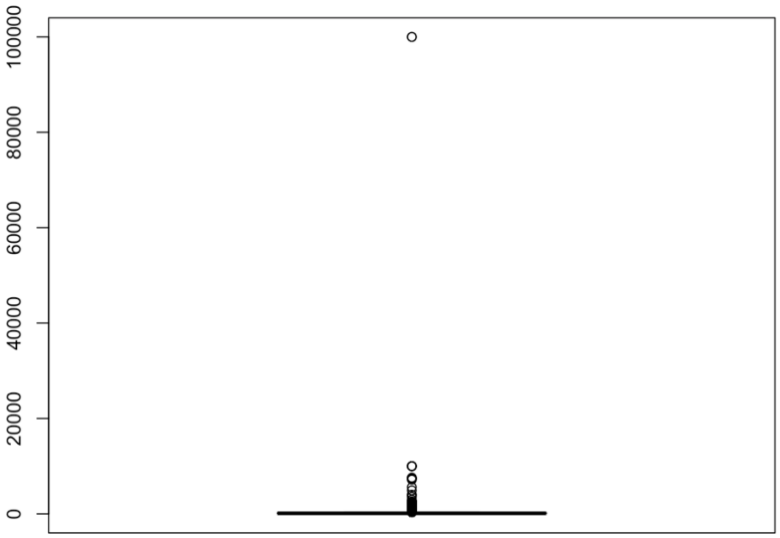
```
> dim(NewYork)
[1] 12897    15
```

```
> str(NewYork)
'data.frame':   12897 obs. of  15 variables:
 $ room_type                     : Factor w/ 3 levels "Entire home/apt",..: 1 2 1 1 2 1 1 2 2 2 ...
 $ price                         : num  187 90 292 160 196 220 84 60 90 106 ...
 $ minimum_nights                : int  2 30 30 30 30 30 30 30 30 30 ...
 $ number_of_reviews             : int  6 19 12 49 5 12 4 9 7 8 ...
 $ last_review                   : Factor w/ 3 levels "2011-2015","2016-2020",..: 3 3 3 3 3 2 2 3 2 3 ...
 $ reviews_per_month             : num  1.67 0.24 1.71 0.67 0.1 0.21 0.08 0.18 0.09 0.22 ...
 $ calculated_host_listings_count: int  1 2 1 1 12 1 1 1 3 1 ...
 $ availability_365              : int  343 5 365 0 0 0 0 2 0 15 ...
 $ number_of_reviews_ltm         : int  6 2 12 7 0 0 0 0 0 5 ...
 $ license                       : Factor w/ 2 levels "License","Exempt": 2 2 2 2 2 2 2 2 2 2 ...
 $ rating                        : num  4.17 4.79 4.67 4.71 4.8 5 4.75 4.89 4.43 5 ...
 $ bedrooms                      : int  1 1 1 1 1 2 1 1 1 1 ...
 $ beds                          : int  2 1 1 1 1 2 1 1 1 1 ...
 $ baths                         : num  1 1 1 1 1 1 1 1 1 1 ...
 $ region                        : Factor w/ 4 levels "East","North",..: 3 3 3 3 3 3 3 2 3 3 ...
```

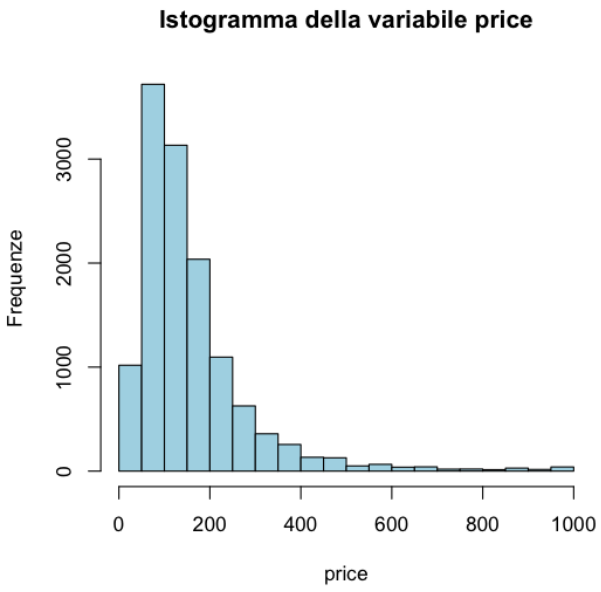
ANALISI DESCRITTIVA

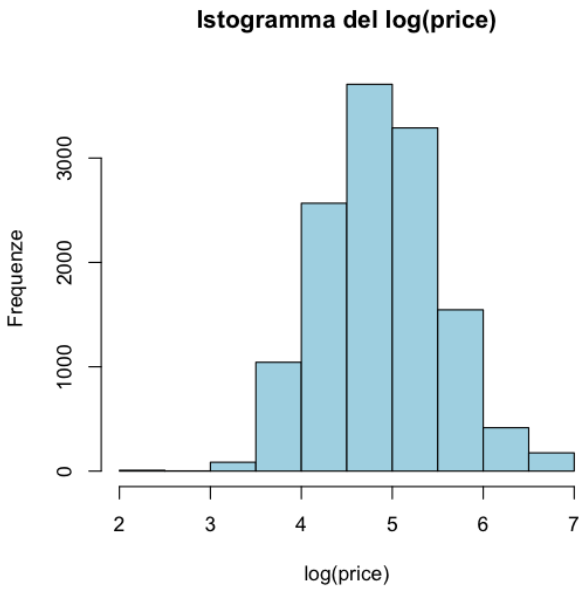
Variabili Quantitative Continue

Price



Il seguente box plot riguardante la variabile dipendente "price" presenta outliers. La funzione summary ci riporta una media di 194 dollari circa.
Dopo essere andati a identificare le case che escono dal trend generale, si decide di calcolare lo "Z-score", ovvero una misura che indica di quante deviazioni standard un valore si discosta dalla media, così da individuare i valori con |Z|>3, eliminarli, e poter normalizzare la distribuzione.





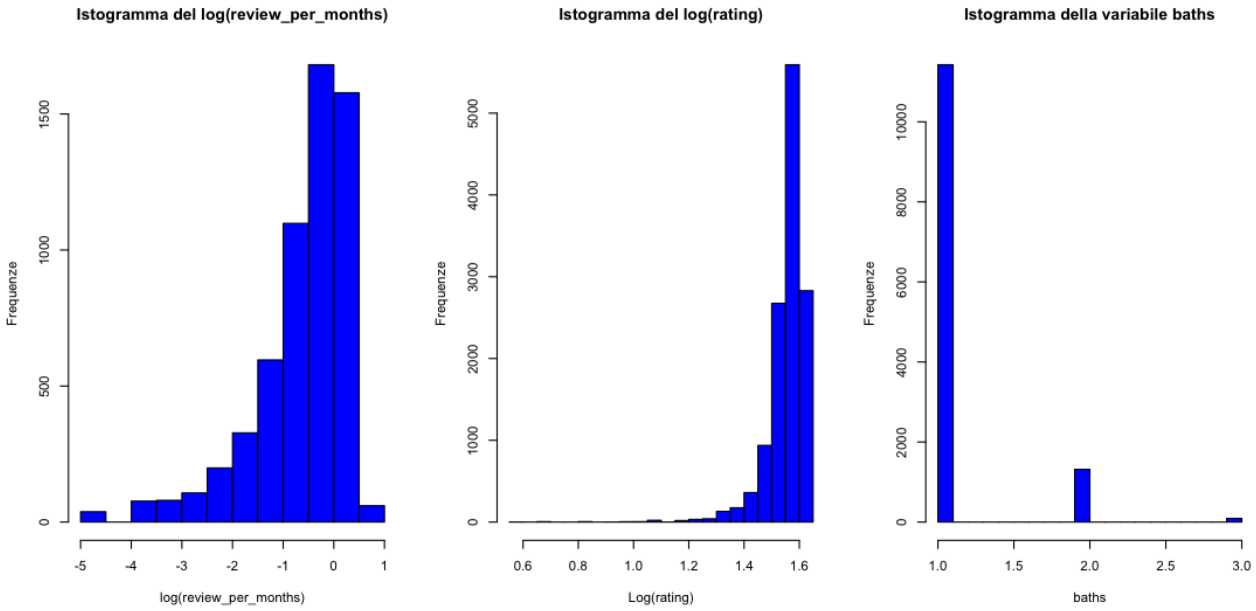
Una volta fatto questo passaggio, possiamo notare come il prezzo minimo sia 10 dollari, mentre il prezzo massimo raggiunge i 1000, indicando una notevole variazione. Il 25% delle inserzioni ha un prezzo inferiore a 85, mentre il 50% ha un prezzo inferiore a 129. Per il 75% il valore è di 200, il che suggerisce che la maggior parte degli alloggi ha un prezzo relativamente contenuto. Tuttavia, si mantiene u.s. con un prezzo massimo di 1000 perché potrebbero rappresentare inserzioni di lusso o con caratteristiche particolari.

Review x Months - Ranking - Baths

L'analisi delle possibili variabili esplicative inizia con queste prime tre variabili quantitative, per le quali si cerca di individuare la distribuzione e la presenza di eventuali outlier, prima attraverso un box plot e poi un istogramma.
Per la variabile "review per months" si nota, attraverso la funzione table() che dopo le 6 recensioni medie il numero di osservazioni che presenta un valore superiore a 1 inizia a diminuire drasticamente, con larghi salti da un numero di recensioni all'altro. Non avendo una distribuzione uniforme, si decide quindi di calcolare Z-score in modo tale da identificare come outlier i valori che superano la soglia di 3. Vengono eliminate tutte le osservazioni che superano le 7 recensioni mensili.
La variabile presenta valori al di sotto dell'1 e inizialmente questo potrebbe risultare un'anomalia. L'eliminazione di questi potrebbe aiutare a normalizzare ancora di più la distribuzione. Sono state cercate allora informazioni riguardanti il significato delle variabile sul sito "Airbnb open data" in modo tale da confermare l'anomalia, ma non è stato trovato niente. Si è quindi ipotizzato che la variabile in questione riguardi il numero di recensioni medie mensili negli anni. Un appartamento potrebbe quindi non ricevere recensioni per mesi, ottenendone poi un numero elevato in un periodo particolare, comportando così una media inferiore a 1. Per questo motive non verranno apportate ulteriori modifiche. La trasformazione logaritmica permette di normalizzare abbastanza la distribuzione
Anche per la variabile "baths" presentavano valori molto elevati, comportando quindi l'eliminazione di outliers, oltre che di osservazioni con valore 0.
Per provare a normalizzare il più possibile la distribuzione ed inserirla nel modello di regressione è stata cambiata la natura della variabile, passando da una quantitativa continua ad una discreta, racchiudendo le osservazioni con valore 1.5 in quelle da 1, quelle da 2.5 in quelle da 2, quelle da 3.5 in quelle da 3 e quelle da 4.5 in quelle da 4. A questo punto, la numerosità elevata del dataset ha permesso di intervenire nuovamente ed eliminare tutte le osservazioni con numero di bagni oltre i 3, essendo poi queste perlopiù caratterizzate dalla presenza di situazioni anomale, come la seguente:

8840 0 Exempt 4.83 1 1 4.5 (Un appartamento con una camera, un letto e 4.5 bagni.)

Per la variabile "rating" si è deciso di mantenere tutti valori essendo questi compresi tra 1 e 5. Si è però deciso di calcolare i percentili 1 e 99 per sostituirli ai valori estremi in modo tale da ridurne l'effetto. Dopo aver trasformato la variabile in una logaritmica ed averne calcolato la deviazione standard, è stato effettuato un test di assimetria (Skewness) per valutare la forma della distrubuzione, ed è stato ottenuto un valore pari a -2,045. Si ha quindi una distribuzione dei dati significativamente asimmetrica.

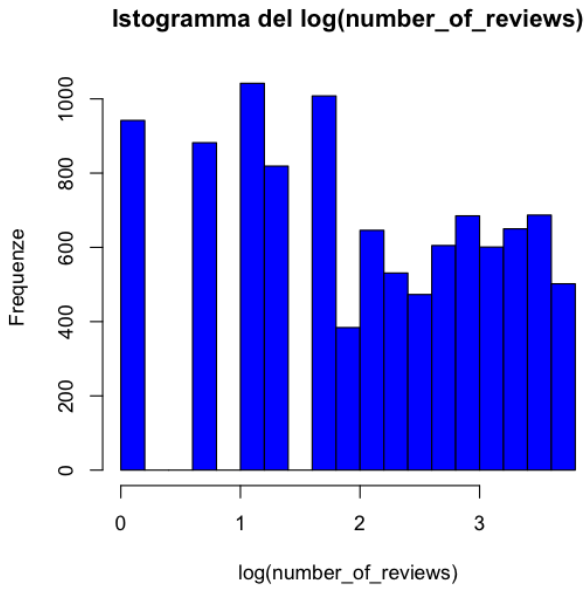


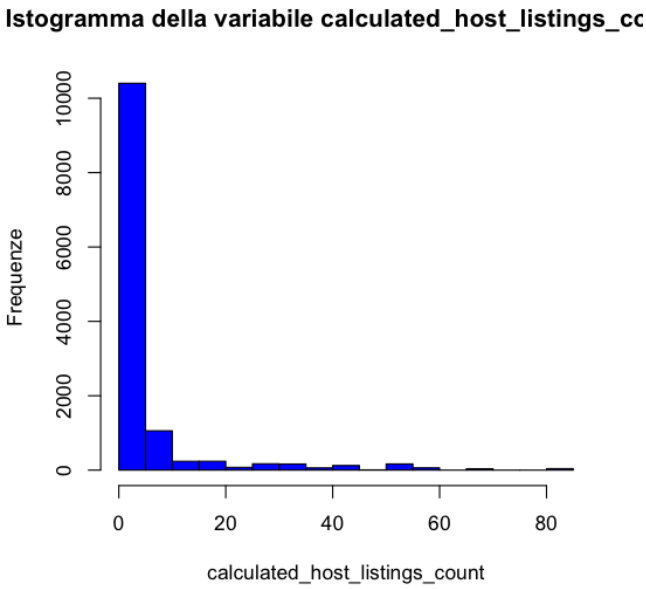
La variabile “reviews per month” mostra una distribuzione asimmetrica positiva, con una mediana di 0.85 e una media di 1.236, indicando la presenza di valori estremi superiori. Il 75% degli alloggi riceve meno di 1.88 recensioni mensili, mentre il massimo di 7.00 evidenzia una notevole variabilità nell'attività di recensione.
L'indagine sui rating degli alloggi evidenzia una marcata tendenza verso valutazioni positive. La maggioranza degli alloggi riceve punteggi elevati, come indicato dalla mediana di 4.81 e dal terzo quartile di 4.94, suggerendo un elevato standard qualitativo. Tuttavia, la presenza di un valore minimo di 1.75 segnala alcune eccezioni negative, indicando una certa eterogeneità nella soddisfazione degli ospiti.
Il numero di bagni presenti in un alloggio sembra essere principalmente 1.

Variabili Quantitative Discrete

Number of Reviews e Number of Reviews last month

Un box plot per la variabile “number of reviews” ha permesso di individuare un’asimmetria nella distribuzione, confermata anche dalla differenza nei valori di media e mediana.
Dopo la pulizia dei dati attraverso l’eliminazione degli outliers, il minimo di review ottenute da un appartamento è stato di 2, contro un massimo di 294.
Attraverso una trasformazione logaritmica è stata ottenuta una differenza tra media e mediana ora praticamente nulla.
Situazione diversa invece per la variabile “number of review ltm”, la quale presenta un valore massimo molto elevato. È stato calcolato lo Z-score e sono stati individuati i valori che uscivano dalla soglia di 2, così da poterli eliminare ed applicare una trasformazione logaritmica per normalizzare la distribuzione. Neanche dopo questi accorgimenti la situazione migliora.
Un dubbio riguardante queste due variabili è che con “review per months”, pur presentando dati riguardanti situazioni diverse, possano avere al loro interno informazioni simili e che si ripetono.
I coefficienti di correlazioni calcolati per le tre variabili sono effettivamente molto elevati, il che potrebbe voler dire multicollinearità. Verranno comunque mantenute nel dataset e l’assunzione appena riportata verrà confermata o smentita durante la formulazione del modello di regressione.



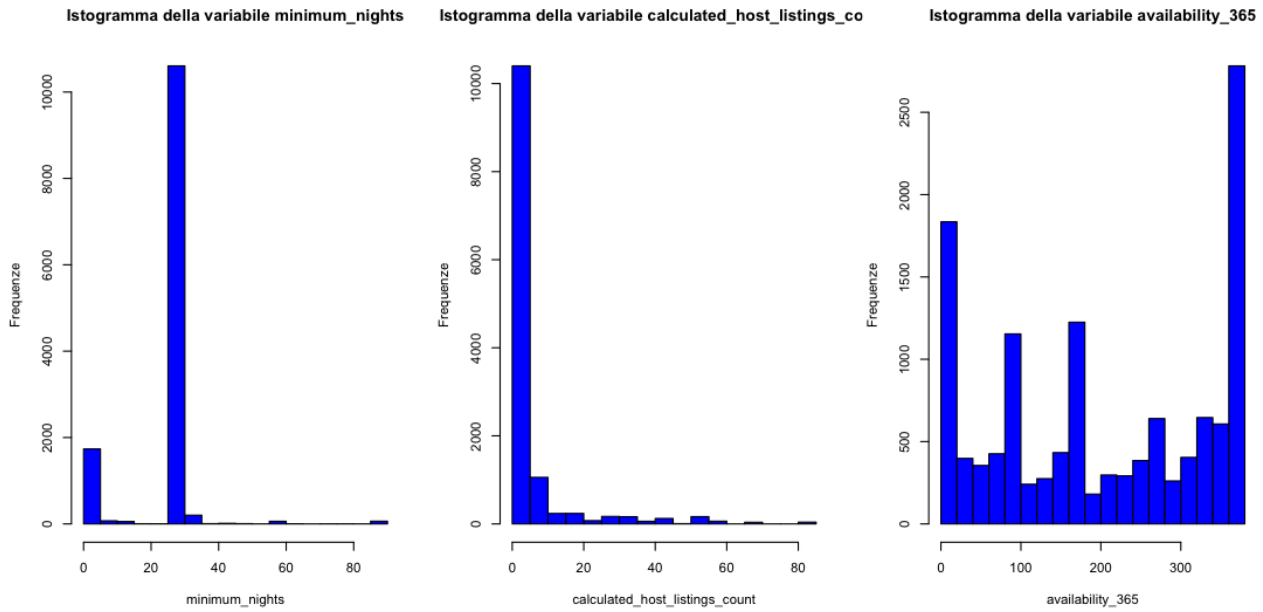


L'analisi del numero di recensioni dell’ultimo mese presenta una mediana di 5 e una media di 10.07, indicando la presenza di valori estremi superiori. Il 75% degli alloggi riceve meno di 16 recensioni, mentre il massimo di 44 evidenzia una notevole variabilità nell'attività di recensione. La presenza di valori nulli (0) riguarda alloggi senza recensioni recenti, contribuendo all'asimmetria osservata.
L'analisi del numero totale di recensioni mostra una notevole eterogeneità tra gli alloggi. Sebbene la maggioranza degli alloggi abbia ricevuto un numero di recensioni relativamente basso, come indicato dalla mediana di 21, alcuni alloggi spiccano con un numero significativamente più alto, raggiungendo

un massimo di 294. La media di 42.18, superiore alla mediana, suggerisce una distribuzione asimmetrica positiva, confermata dalla differenza tra il terzo quartile (54) e il valore massimo.

Minimum Nights - Calculated host listing counts – Availability 365

Per quanto riguarda la prima variabile, sono stati individuati outliers attraverso un box plot e si è deciso di entrare più nello specifico. Dato che la maggior parte delle osservazioni relative al numero minimo di notti risultava essere circa di 30 gg, si è deciso di calcolare lo Z-score in modo tale da individuare come outlier i valori che superano la soglia di 3. A questo punto, sono state tenute all'interno del dataset solamente alloggi che richiedevano un numero minimo di 90 gg.
Per la seconda variabile il problema era lo stesso. Questa volta però i valori estremi risultavano essere fortemente anomali, dato che avrebbe voluto dire, ad esempio, che un host possedeva più di 1000 appartamenti a New York. Gli outliers sono stati quindi calcolati sempre trovando Z-score ma riducendo la soglia a 2. A questo punto sono state eliminate tutte le osservazioni con un valore sopra 83 appartamenti.
Per l'ultima variabile sopracitata le azioni correttive non esistono. Essendo riferita alla disponibilità dell'appartamento durante l'anno è perfettamente legittima la presenza di numerose osservazioni con valore 0 e 365 (i due estremi).
Le variabili in questione verranno mantenute quindi come trovate in origine.

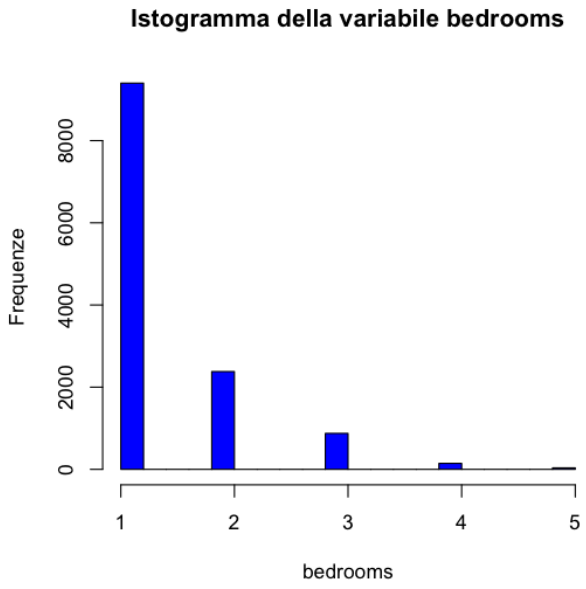


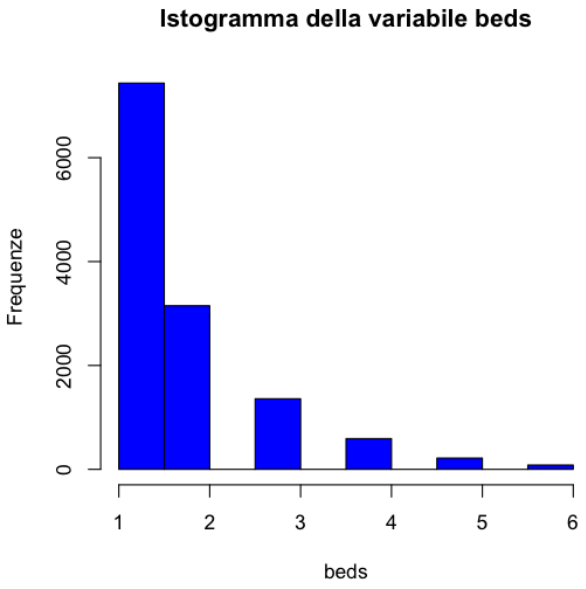
L'analisi del numero di alloggi gestiti da ciascun host rivela una distribuzione altamente asimmetrica. La maggior parte degli host gestisce un numero limitato di alloggi, con una mediana di 2 e il 75% che ne gestisce 4 o meno. Tuttavia, la presenza di un valore massimo di 83 evidenzia alcuni host con un portafoglio significativamente più ampio, suggerendo una notevole disparità nella gestione degli alloggi.
L'analisi della disponibilità annuale degli alloggi rivela una distribuzione bimodale. La presenza di un minimo pari a 0 e un massimo pari a 365 indica una polarizzazione tra alloggi costantemente

disponibili e quelli con disponibilità nulla. La mediana di 188 e la media di 201.8 suggeriscono una tendenza verso una disponibilità parziale, con una notevole variabilità tra gli alloggi, come evidenziato dai quartili.
L'analisi del numero minimo di notti richieste per la prenotazione rivela una distribuzione altamente concentrata sul valore di 30, con una mediana e quartili coincidenti. La media di 26.53, inferiore alla mediana, indica la presenza di valori minimi inferiori, come 1, che influenzano la media. La notevole differenza tra il terzo quartile (30) e il valore massimo (90) evidenzia una variabilità significativa nelle politiche di soggiorno minimo, suggerendo una disomogeneità nell'offerta degli alloggi.

Beds e Bedrooms

La prima variabile sottoposta a pulizia è stata “bedrooms”, la quale presentava la maggior parte delle osservazioni nelle u.s. con uno, due, tre o quattro bagni. Essendo la variabile soggetta ad outliers la prassi è stata la stessa delle precedenti variabili, andando così ad eliminare le u.s. che presentavano più di 5 camere da letto.
Per la variabile “beds” la questione è molto simili. È stata prevista l’eliminazione dei valori ritenuti outliers (oltre 6 bagni).





La composizione degli alloggi in termini di numero di camere da letto evidenzia una netta predominanza di unità con una sola camera, rappresentando circa il 73% del totale. La presenza di alloggi con due camere da letto, seppur in misura minore (circa il 18%), conferma la tendenza verso unità abitative di dimensioni contenute. La progressiva diminuzione della frequenza con l'aumentare del numero di camere da letto, con solo una minima percentuale di alloggi con quattro o cinque camere, sottolinea la rarità di unità abitative di grandi dimensioni all'interno del campione.
La distribuzione del numero di letti negli alloggi evidenzia una forte concentrazione su unità con un solo letto, che rappresentano circa il 58% del campione. La presenza di alloggi con due letti, seppur in misura minore (circa il 24%), conferma ancora una volta la tendenza verso unità abitative di dimensioni contenute.

Variabili Qualitative

Region

Si tratta di un Factor all'interno del quale sono stati raggruppati tutti i quartieri di New York inizialmente indicati separatamente. Ora i livelli sono 4: North, South, East and West.
Per questa variabile non si ritengono necessarie ulteriori modifiche.
Il grafico a barre mostra una chiara disparità nella distribuzione delle recensioni per regione. La regione South domina nettamente, mentre le altre regioni presentano numeri significativamente inferiori. Questo suggerisce che la maggior parte delle attività recensite si concentra nella regione meridionale.

License

Inizialmente la variabile è stata modificata raggruppando tutte le licenze di tipo OSE (licenza per affitti a breve termine, cioè sotto i 30gg, a New York) in un'unica categoria denominata "license". È stata quindi creata una variabile Factor da 3 livelli: License, No License e Exempt.
Il livello No license non presentava però valori al proprio interno, rendendo quindi necessario un cambiamento in un Factor a 2 livelli: License e Exempt.
Il grafico a torta indica che la stragrande maggioranza delle attività recensite sono esenti da licenza (Exempt). Questo viene confermato anche dal fatto che la maggior parte degli alloggi richieda un numero minimo di notti inferiore a 30 gg.

Room Type

Si tratta di una variabile Factor da 4 livelli: Entire home/apt, Hotel room, Private room e Shared room. La numerosità di Hotel è molto bassa e, dato che potrebbe darci la stessa informazione delle ultime due variabili, si decide di escluderla e modificare la variabile in un Factor a 3 livelli.
Il grafico a barre mostra che la categoria "Entire home/apt" è la più frequente, seguita da "Private room", mentre "Shared room" è significativamente meno comune, cosa che ipoteticamente potrebbe dimostrare una preferenza degli individui per la privacy che quindi viene offerta maggiormente dagli host.

Last Review

Altra modifica apportata è stata la trasformazione di questa variabile da un Factor a più di 1000 livelli ad un Factor da 3 livelli: 2011-2015, 2016-2020, 2021-2025 cercando di mantenere uniforme il periodo temporale. Non sono state apportate ulteriori modifiche.
Il grafico a torta evidenzia che la maggior parte delle recensioni, il 90% circa, sono state effettuate nel periodo 2021-2024, dimostrando un forte livello di attività recente negli alloggi.

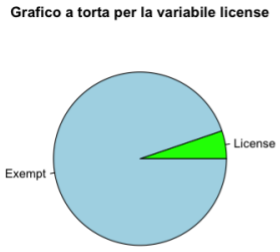


Grafico a torta per la variabile license
License
Exempt

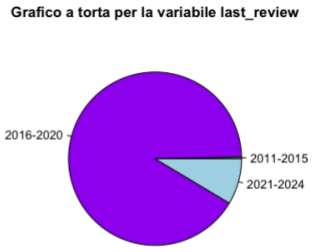


Grafico a torta per la variabile last_review
2016-2020
2011-2015
2021-2024

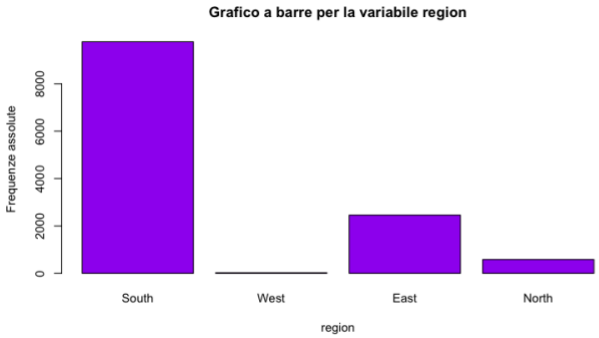


Grafico a barre per la variabile region
Frequenze assolute
0
2000
4000
6000
8000
South
West
East
North
region

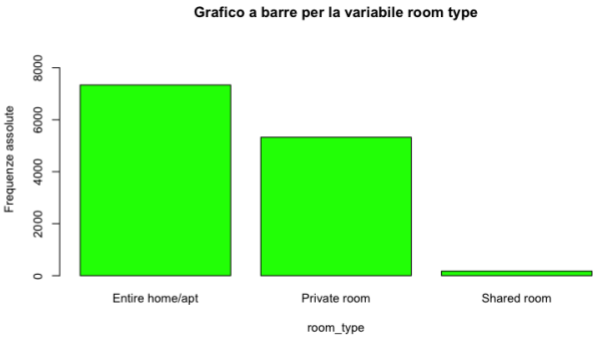


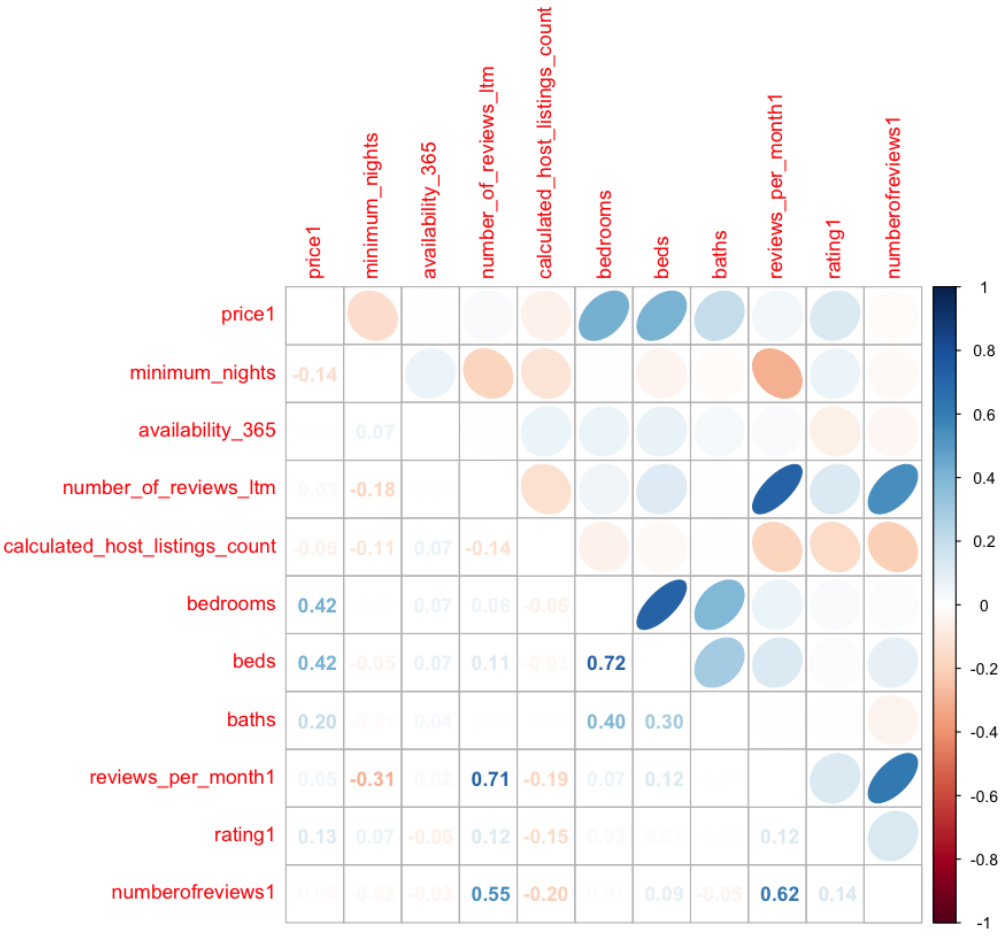
Grafico a barre per la variabile room type
Frequenze assolute
0
2000
4000
6000
8000
Entire home/apt
Private room
Shared room
room_type

ANALISI DI CORRELAZIONE

Una volta fatta la parte di pre-procesing, pulizia dei dati e un'analisi descrittiva delle variabili, in modo tale da verificare la bontà e l'adeguatezza di queste per il futuro modello di regressione lineare, si passa all'analisi di correlazione tra variabili e tra la variabile risposta e le variabili esplicative.
Questa parte ha come finalità quella di garantire l'assenza di multicollinearità, ovvero evitare che le variabili varino insieme, rendendo difficile isolare l'effetto individuale di ciascuna variabile sulla variabile dipendente.
Per fare ciò si presenterà prima una situazione generale dell'intensità delle relazioni tra variabili con un grafico corrplot, individuando allo stesso tempo l'intensità della relazione espressa tramite un coefficiente di correlazione.
Questo ci permetterà di analizzare situazioni particolari tra variabili che potrebbero presentare multicollinearità, oltre che permettere un confronto tra le variabili ed il prezzo.
Verrà studiata la situazione per quanto riguarda le variabili categoriali, in modo tale da poter trarre le prime conclusioni con riferimento alle variabili che si sono dimostrate buone per il modello di regressione lineare.

- Corrplot delle Variabili Numeriche

Le variabili in questione sono: price, minimum_nights, number_of_reviews, calculated host listings count, reviews_per_month, availability_365, number_of_reviews_ltm, rating, bedrooms, beds, baths.



Coefficienti di correlazione:

```
> print(cor_matrix)
                                price1 minimum_nights availability_365 number_of_reviews_ltm calculated_host_listings_count     bedrooms        beds       baths reviews_per_month1      rating1 numberofreviews1
price1                      1.000000000   -0.140901103      0.006007146          0.026543523                 -0.0595233362  0.423282915  0.41707594  0.2004455371        0.048869316  0.128202769      -0.01783669
minimum_nights             -0.140901103    1.000000000      0.066162864         -0.181855841                 -0.1104815323  0.007216721 -0.04643805 -0.0130149778       -0.309571361  0.065593628      -0.02085685
availability_365            0.006007146    0.066162864      1.000000000         -0.003234616                  0.0689264200  0.067326050  0.07282924  0.0352816759        0.024084867 -0.061137604      -0.03174653
number_of_reviews_ltm       0.026543523   -0.181855841     -0.003234616          1.000000000                 -0.1367216874  0.059072576  0.11136937  0.0044753358        0.711851643  0.121869005       0.54918298
calculated_host_listings_count -0.059523336 -0.110481532     0.068926420         -0.136721687                  1.0000000000 -0.057931010 -0.02955844  0.0008367021       -0.185481636 -0.153297144      -0.20328394
bedrooms                    0.423282915    0.007216721      0.067326050          0.059072576                 -0.0579310097  1.000000000  0.71801607  0.3972028228        0.068036445  0.027960766       0.01093615
beds                        0.417075942   -0.046438050      0.072829243          0.111369367                 -0.0295584418  0.718016069  1.00000000  0.2996380530        0.122546048  0.014419753       0.08570248
baths                       0.200445537   -0.013014978      0.035281676          0.004475336                  0.0008367021  0.397202823  0.29963805  1.0000000000       -0.008429533 -0.004034557      -0.04650185
reviews_per_month1          0.048869316   -0.309571361      0.024084867          0.711851643                 -0.1854816359  0.068036445  0.12254605 -0.0084295329        1.000000000  0.120056981       0.62130686
rating1                     0.128202769    0.065593628     -0.061137604          0.121869005                 -0.1532971440  0.027960766  0.01441975 -0.0040345572        0.120056981  1.000000000       0.13901639
numberofreviews1           -0.017836687   -0.020856850     -0.031746531          0.549182976                 -0.2032839356  0.010936148  0.08570248 -0.0465018485        0.621306865  0.139016386       1.00000000
```

La relazione che cattura immediatamente l'attenzione:

1. Beds e Bedrooms: con un coefficiente di correlazione pari a 0,72 potremmo avere una situazione caratterizzata da multicollinearità. Abbiamo quindi una possibile ridondanza informativa che rende instabili le stime dei coefficienti.
2. Reviews per month e Number of reviews ltm: anche in questo caso il coefficiente di correlazione è molto elevato (0,76) e potrebbe presentarsi multicollinearità.

I coefficienti di correlazione inferiori rispetto a quello sopra riportato, anche se con correlazione moderata, non vengono ritenuti problematici.
Il tutto verrà comunque controllato durante la stepwise regression.

Variabili Qualitative

Anche per queste procediamo ad individuare l'eventuale presenza di relazioni molto forti.
Le variabili in questione sono: Region, Room_type, last_review, license.

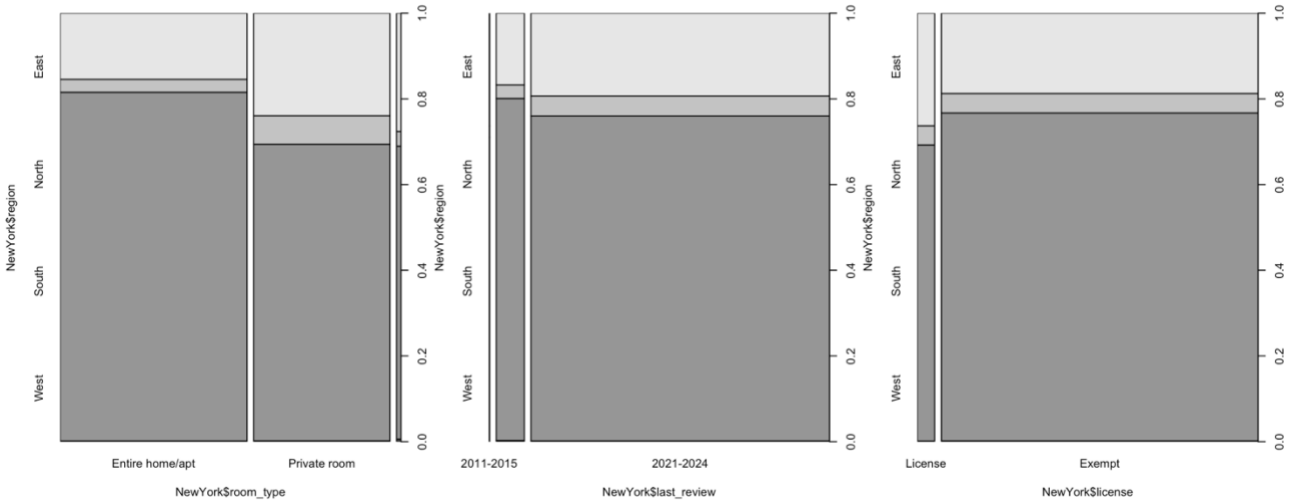
Iniziamo osservando la relazione della variabile Region rispetto alle altre tre.

```
> test1

        Pearson's Chi-squared test

data:  tab1
X-squared = 277.15, df = 6, p-value < 2.2e-16
```

```
> test2

        Pearson's Chi-squared test

data:  tab2
X-squared = 12.876, df = 6, p-value = 0.04504
```

```
> test3

        Pearson's Chi-squared test

data:  tab3
X-squared = 23.467, df = 3, p-value = 3.227e-05
```

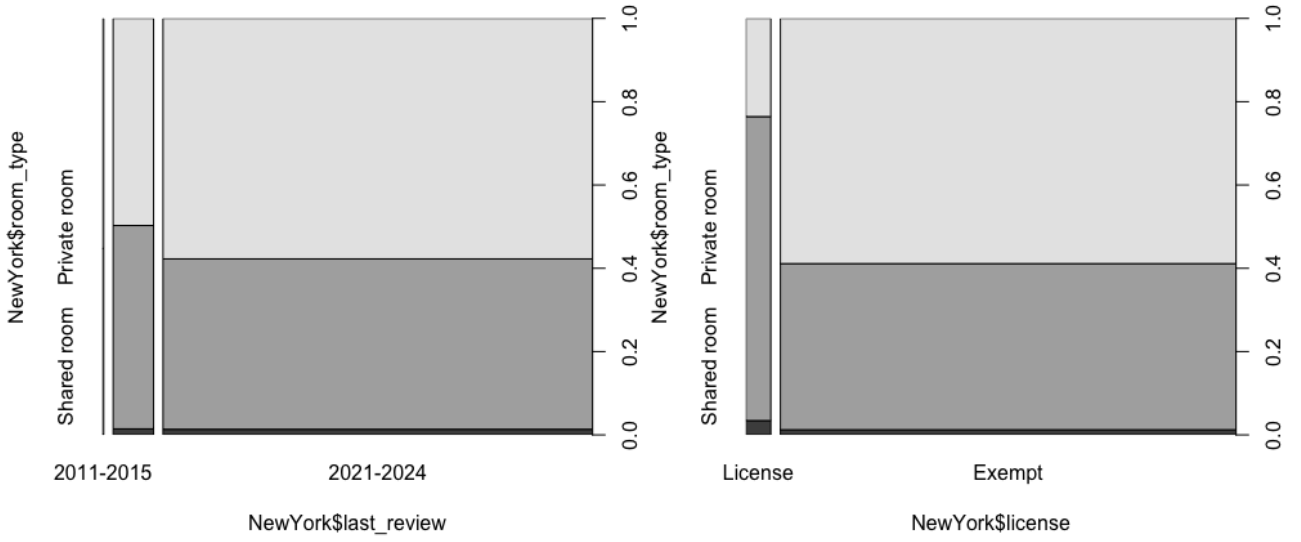
Attraverso una funzione di R è stato svolto il test χ 2, così da poter affermare se esiste dipendenza tra le due variabili.
Il p-value restituiti sono tutti e tre molto piccoli (<0.05), il che mi permette di rifiutare l'ipotesi nulla di indipendenza tra "region" e le variabili, affermando che c'è una relazione statisticamente significativa tra queste, osservabile anche nei grafici a barre contrapposte allegati di seguito.



Ora analizziamo la relazione tra la variabile Room type rispetto a last review e license.

```
> test4

        Pearson's Chi-squared test

data:  tab4
X-squared = 27.295, df = 4, p-value = 1.732e-05
```

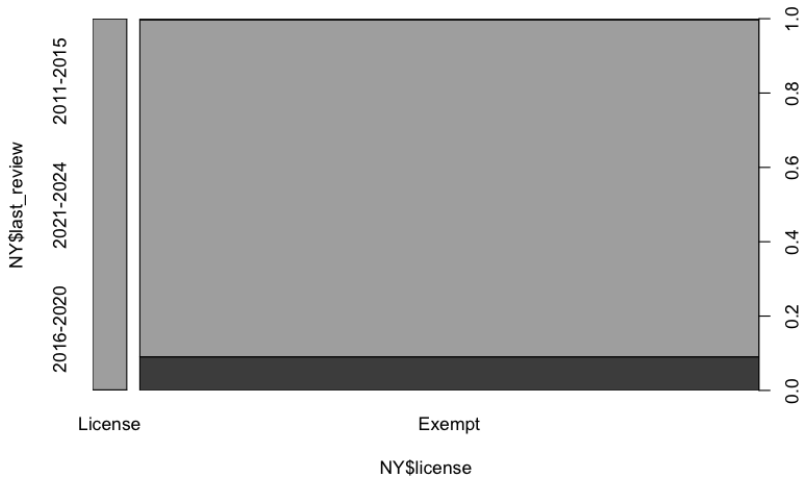
```
> test5

        Pearson's Chi-squared test

data:  tab5
X-squared = 328.76, df = 2, p-value < 2.2e-16
```



R ci restituisce due valori più bassi della soglia di 0.05, permettendoci di affermare che la variabili room type presenta una forte associazione con le due variabili considerate.

Per ultima, analizziamo la relazione esistente tra la variabile Last review e License

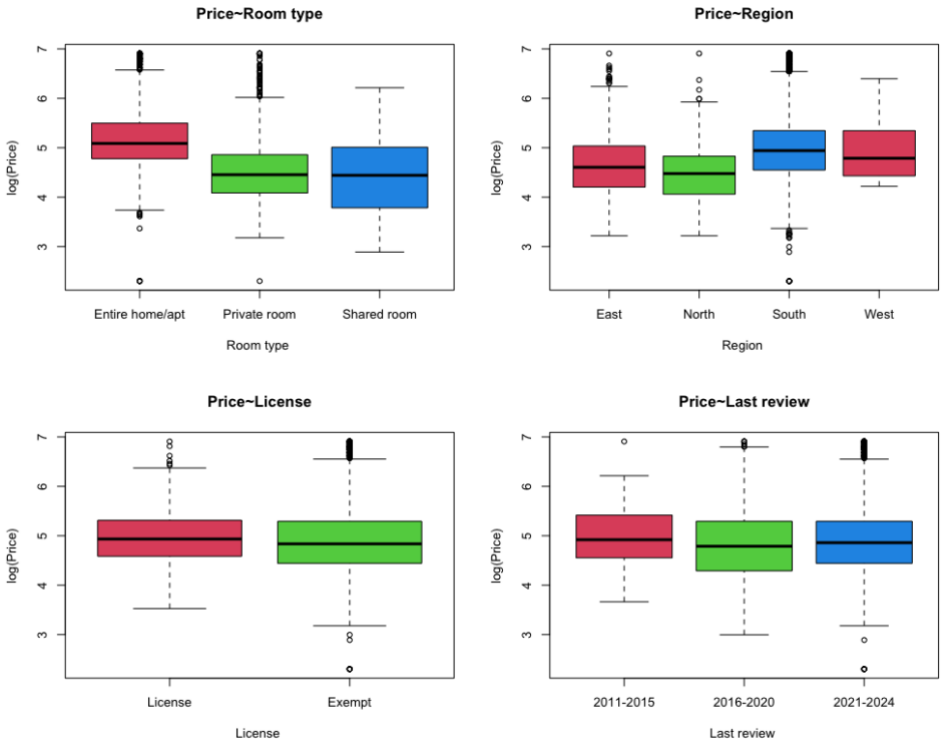
```
> test6

        Pearson's Chi-squared test

data:  tab6
X-squared = 66.574, df = 2, p-value = 3.496e-15
```



Ancora una volta, esiste una forte associazione tra le due variabili in questione.

Prezzo e Variabili Qualitative

Una volta evidenziato lo stato della relazione tra le variabili quantitative, si passa ad evidenziare la relazione tra il Prezzo e quest'ultime.

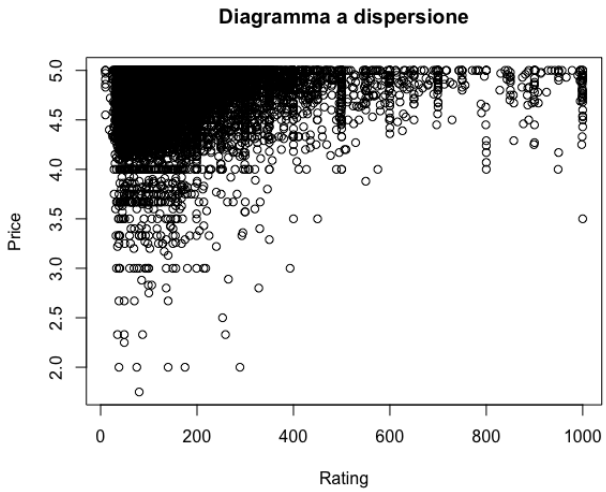


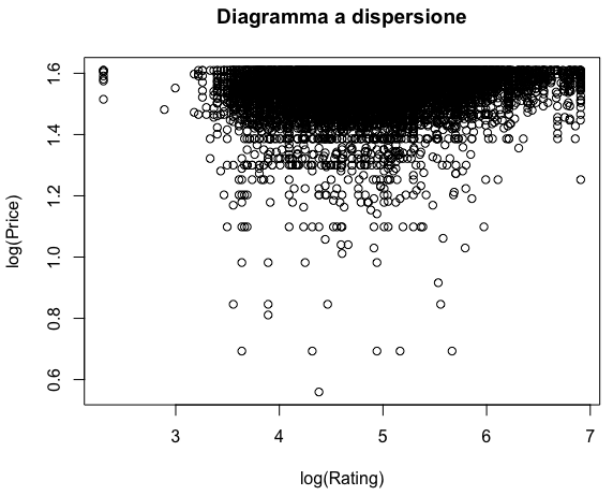
Il grafico "Price~Room type" mostra chiaramente che il tipo di stanza ha un impatto significativo sul prezzo. La mediana del logaritmo del prezzo è più alta per gli appartamenti/case intere, seguita dalle stanze private e infine dalle stanze condivise. Questo si traduce in una differenza percentuale sostanziale nei prezzi: gli appartamenti/case intere tendono ad essere significativamente più costosi delle stanze private e condivise, aspetto comprensibile visto la maggiore privacy ed il maggior spazio a disposizione.
La dispersione dei prezzi è maggiore per gli appartamenti/case intere, indicando più variabilità percentuale nei prezzi di questa categoria. Anche la presenza di outlier suggerisce che ci sono alloggi con prezzi percentualmente molto diversi dalla media, sia verso l'alto che verso il basso.
La relazione "Price~Region" rivela differenze meno marcate tra le regioni. La regione South mostra una mediana del logaritmo del prezzo leggermente più alta rispetto alle altre, mentre la regione North mostra quella leggermente più bassa. A south possiamo trovare i più importanti quartieri come Soho, Upper East e West Side, ecc... In ogni caso, queste differenze si traducono in variazioni percentuali relativamente modeste nei prezzi medi tra le regioni.
Il terzo grafico ci riporta la relazione "Price-License". Si può individuare una differenza moderata tra gli alloggi con licenza e quelli esenti. La mediana del logaritmo del prezzo è più alta per gli alloggi con licenza, il che significa che questi tendono ad avere prezzi percentualmente più alti rispetto a quelli esenti.
Ricordando che la licenza serve per gli alloggi a breve termine, si potrebbe ipotizzare una politica di sconti da parte degli host per soggiorni di lungo periodo.
La dispersione dei prezzi è maggiore per gli alloggi con licenza, suggerendo una maggiore varietà percentuale di offerte in questa categoria.
L'ultimo grafico, "Price-Last review", riporta una leggera tendenza all'aumento dei prezzi nel tempo. Gli alloggi con recensioni più recenti (2021-2024) tendono ad avere una mediana del logaritmo del prezzo leggermente più alta rispetto a quelli con recensioni più datate (2011-2015). Questo si traduce in un leggero aumento percentuale dei prezzi nel tempo.
Il fattore da tenere in considerazione è però la numerosità del sottocampione di osservazioni riguardanti il periodo 2011-2015. Un numero così piccolo potrebbe non essere adeguato a spiegare perfettamente la situazione.

Prezzo e Variabili Quantitative

- Rating

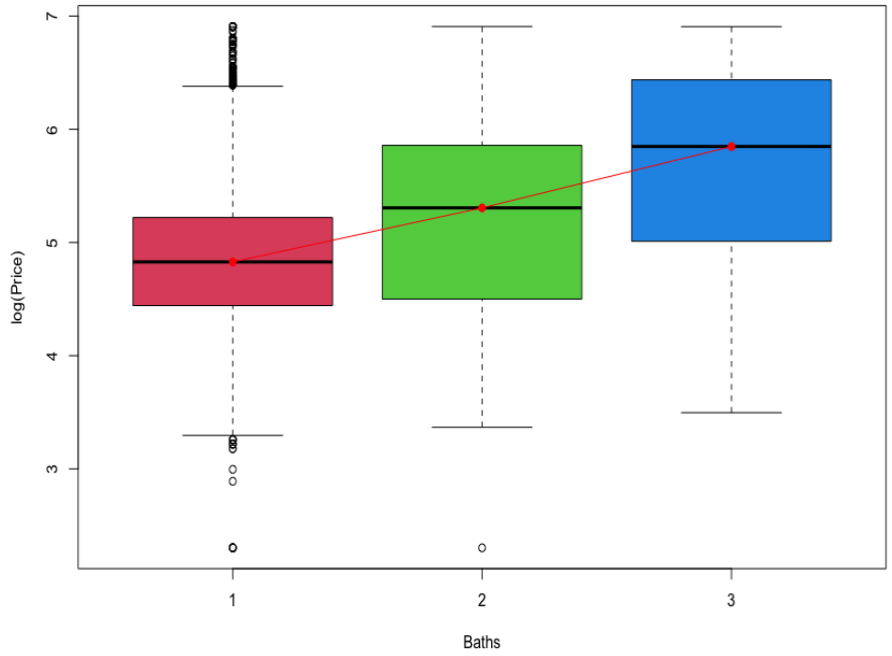
Nel grafico di sinistra sotto riportato si osserva una notevole dispersione dei punti, suggerendo una relazione debole tra le due variabili. Sembra esserci una leggera tendenza all'aumento del prezzo con l'aumentare del rating, ma è tutt'altro che lineare. Queste infatti, presentavano un problema di eteroschedasticità. Si è quindi provato a trasformare in logaritmo prima solo la variabile risposta e successivamente entrambe, così da ottenere il secondo grafico.
Questo passaggio pare abbia compresso la scala dei dati, ma la dispersione rimane elevata. Ora la relazione tra le variabili trasformate sembra ancora più debole rispetto al grafico originale.
Con un coefficiente di correlazione tra log(Prezzo) e log(Rating) di 0.1297345 possiamo affermare che la variabile Rating ha un impatto molto limitato sul prezzo.





- Baths

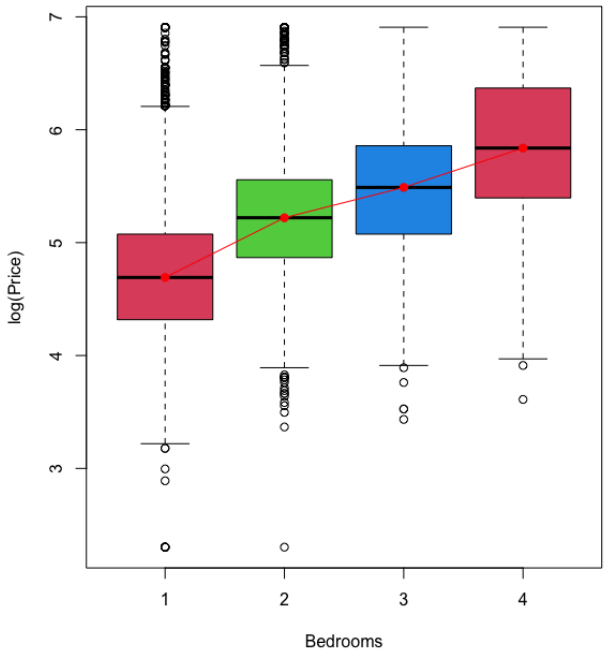
Il logaritmo del prezzo in relazione alla variabile “baths” restituisce il seguente box plot condizionato.

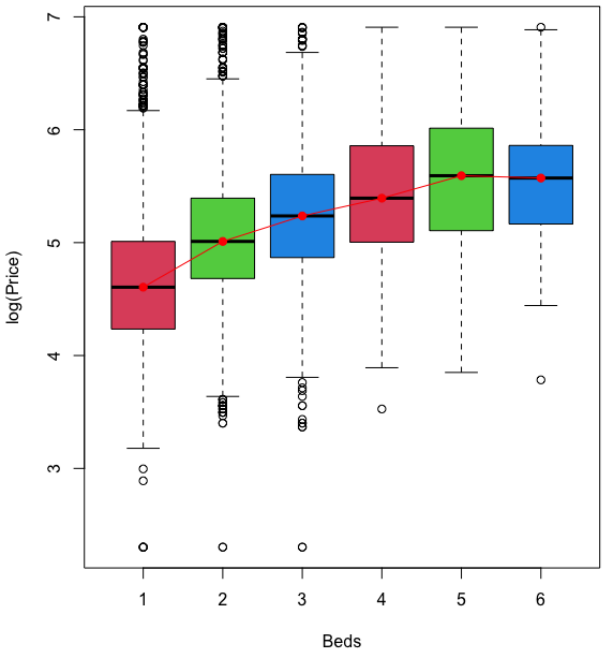


Ci mostra una chiara tendenza all'aumento del logaritmo del prezzo all'aumentare del numero di bagni. In altre parole, gli alloggi con più bagni tendono ad avere prezzi più elevati. La dispersione dei prezzi sembra aumentare leggermente con il numero di bagni. Questo suggerisce che la variabilità dei prezzi è maggiore per gli alloggi con più bagni. La tendenza appena osservata viene confermata da un coefficiente di correlazione pari a 0.2004455.

- Bedrooms e Beds

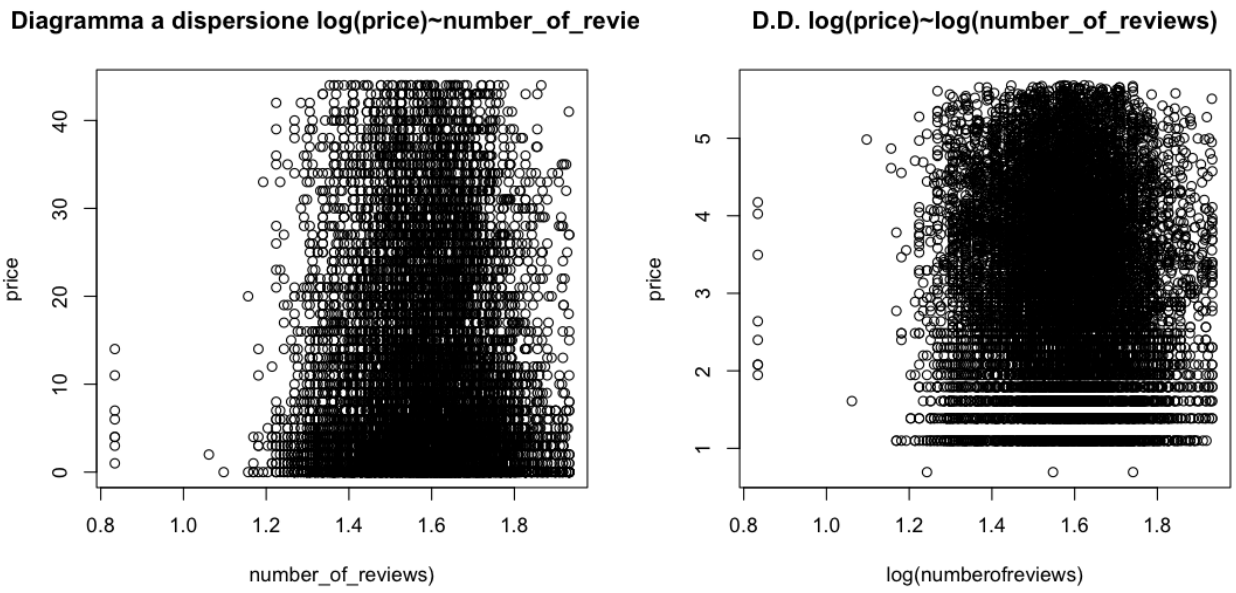
Anche “bedrooms”, come “baths”, presenta chiara tendenza all'aumento del logaritmo del prezzo all'aumentare del numero di camere da letto. La situazione cambia però per alloggi con 5 camere, dove il prezzo nuovamente subisce un crollo. Osservando più nello specifico le osservazioni interessate si osserva che appartengono quasi tutte ad alloggi situati nella zona sud della città, che prima abbiamo appurato essere quella con i prezzi più elevati. Inoltre, circa la metà degli alloggi presenta un letto solo, il che stona con la numerosità delle camere. Queste considerazioni, affiancate al fatto che la numerosità di osservazioni con 5 camere è molto ridotta e al fatto che il campione è ancora ottimo in termini di numerosità (12884 osservazioni), ci permettono di eliminare le u.s. con 5 camere e rappresentare graficamente la nuova situazione.
La variabile “beds” viene mantenuta invariata. Il logaritmo del prezzo aumenta sempre all’aumentare del numero di letto a disposizione, per poi stabilizzarsi quando raggiunge numerosità pari a 6 letti. Questo rappresenta quindi il limite oltre il quale il numero di letti non rappresenta più un valore aggiunto per un alloggio.
I coefficienti di correlazioni delle due variabili con la variabile risposto sono entrambe molto elevate: 0.4372122 (bedrooms) e 0.4142195 (beds).





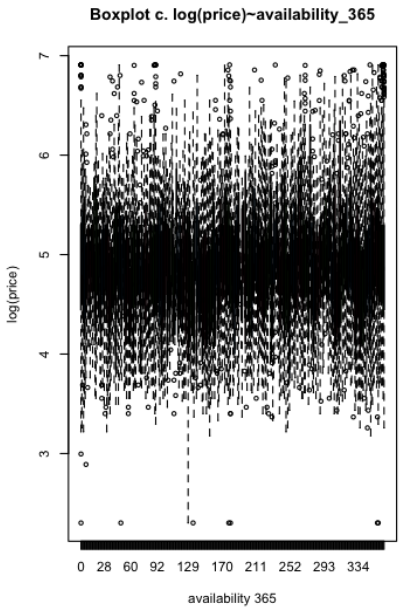
- Number of reviews:

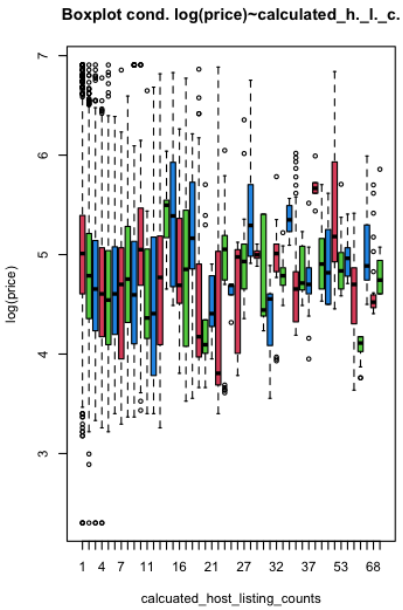
Come per la variabile “rating”, anche “number of reviews” non riesce a beneficiare della trasformazione logaritmica, presentando ancora forte concentrazione sulla metà destra senza però assumere un andamento lineare.
Il coefficiente di correlazione pari -0.01616541.
Questo significa che il numero di recensioni non fornisce informazioni utili per prevedere o spiegare le variazioni del logaritmo del prezzo.

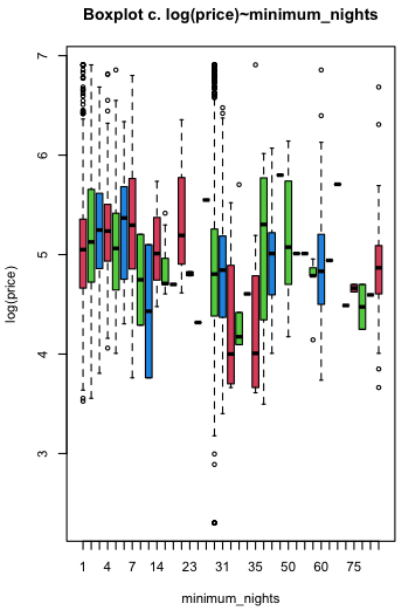


- Minimum nights; Availability 365; Calculated host listing counts

Per tutte e tre le variabili non sono state apportate modifiche ulteriori rispetto a quelle fatte durante la fase iniziale del lavoro.
Per prima cosa è stata rappresentata la relazione tra prezzo logaritmico rispetto alla disponibilità dell'alloggio su 365 giorni. La distribuzione appare molto densa e sparsa, suggerendo che non esista una chiara relazione tra le due variabili. Questo è confermato dal coefficiente di correlazione -1,607056e-05, praticamente nullo. Significa che la disponibilità dell'alloggio durante l'anno non influenza significativamente il prezzo.
Si è poi passati ad identificare il legame tra il prezzo e il numero di appartamenti posseduti dallo stesso host. Anche qui si osserva una distribuzione con molta variabilità, ma senza un trend chiaro. Il coefficiente di correlazione è -0,06633018, indicando una leggera correlazione negativa: chi possiede più alloggi potrebbe applicare prezzi più bassi, probabilmente per ottenere un'occupazione più alta e garantire un cash flow costante. Tuttavia, il valore della correlazione è basso, quindi la relazione non è particolarmente forte.
L'ultimo grafico mostra la distribuzione dei prezzi in scala logaritmica rispetto al numero minimo di notti richiesto. Notiamo che la maggior parte delle osservazioni si concentra nei primi valori di "minimum nights" (1-10 notti), mentre per valori più elevati i dati sono più sparsi e la variabilità aumenta. Il coefficiente di correlazione è -0,1411074, indicando una correlazione negativa debole, ovvero all'aumentare del numero minimo di notti richiesto, il prezzo tende a diminuire, ma la relazione non è forte.

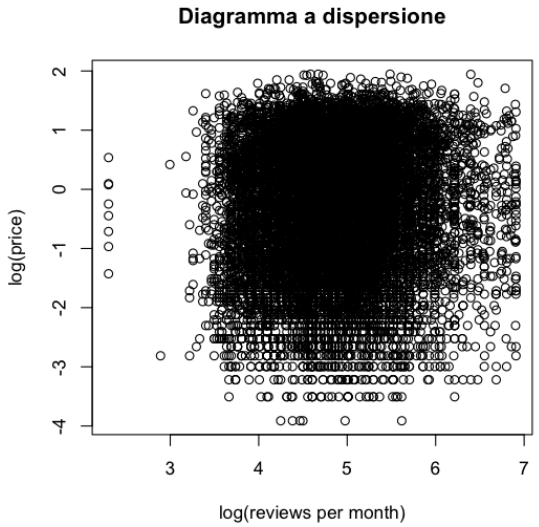


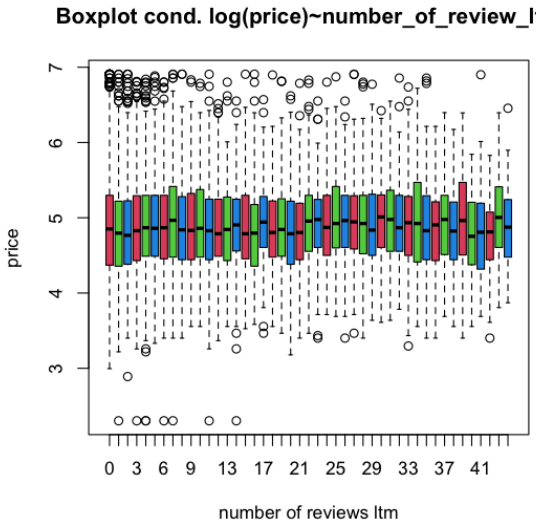




- Reviews per months e Number of reviews ltm

Il diagramma a dispersione tra il logaritmo del prezzo e il logaritmo delle recensioni mensili (log(reviews per months)) mostra una notevole dispersione dei dati, suggerendo una relazione debole. Il coefficiente di correlazione di 0.04886932 conferma questa osservazione, indicando una correlazione positiva molto debole tra le due variabili.
Analogamente, il boxplot condizionato che mette in relazione il logaritmo del prezzo con "number of reviews ltm" rivela una dispersione significativa dei prezzi per ogni livello di recensioni. Non emerge una chiara tendenza all'aumento o alla diminuzione del prezzo in funzione del numero di recensioni. Il coefficiente di correlazione di 0.02654352 conferma la debolezza della relazione, indicando una correlazione positiva molto debole.





MODELLO DI REGRESSIONE LINEARE

Grazie alle fasi di pre proceesing, pulizia dei dati e analisi descrittiva sono state individuate come potenziali variabili esplicative “baths”, “beds”, “bedrooms”, “room type”, “region”, “license” per la variabile risposta “price”.
Si inizia con una stima del modello di regressione attraverso la funzione stepwise come anticipato in precedenza, sia con il metodo backward che con quello forward.

Prima di iniziare sono state eliminate le variabili price, number of reviews e reviews per month mantenendo solamente i corrispettivi trasformati in logaritmo. Per la variabile rating è stato fatto invece il procedimento inverso, dato che la trasformazione logaritmica non avevamo comunque permesso di normalizzare la distribuzione.

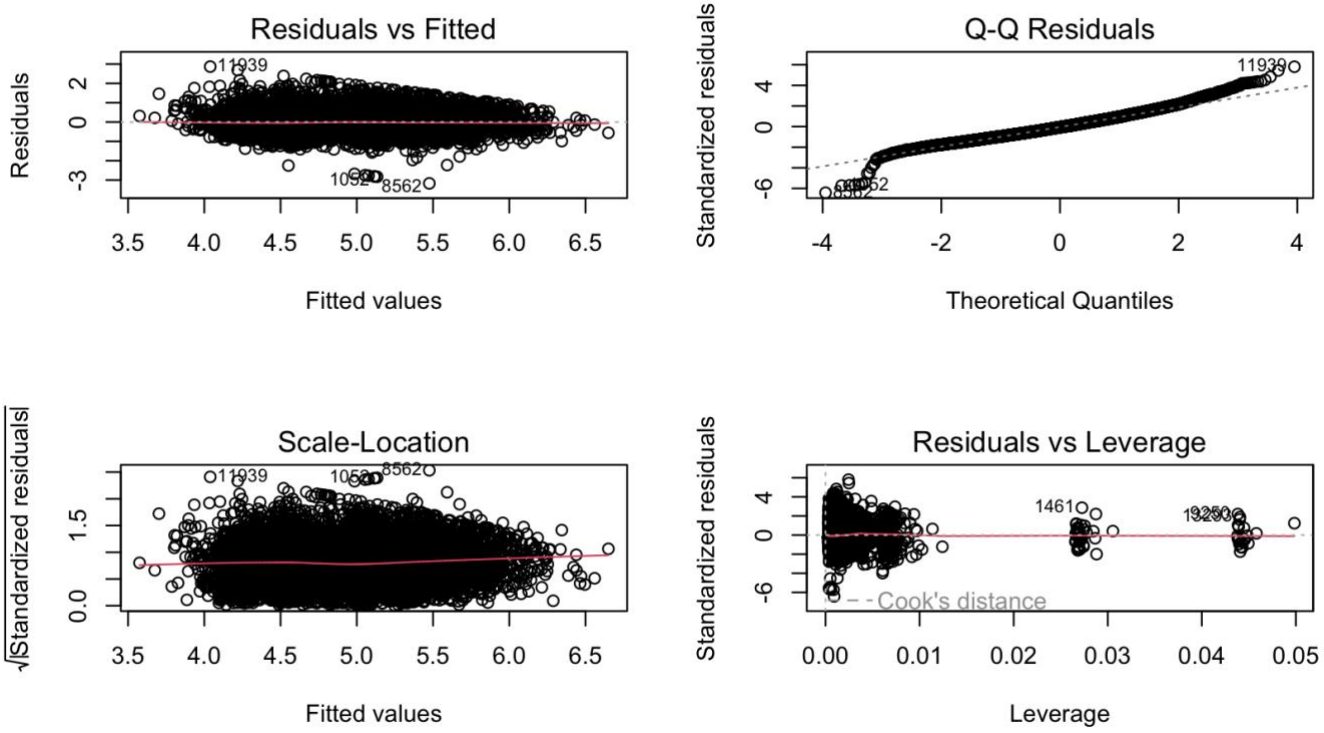
Si inizia con la modalità backward:

```
> summary(mod.backward)

Call:
lm(formula = price1 ~ room_type + minimum_nights + last_review +
    calculated_host_listings_count + availability_365 + rating +
    bedrooms + beds + baths + region + numberofreviews1 + reviews_per_month1,
    data = NY)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1731 -0.3386 -0.0287  0.2984  2.8689

Coefficients:
                               Estimate Std. Error t value Pr(>|t|)
(Intercept)                   4.018e+00  1.110e-01  36.209  < 2e-16 ***
room_typePrivate room        -5.041e-01  9.959e-03 -50.616  < 2e-16 ***
room_typeShared room         -6.350e-01  3.816e-02 -16.639  < 2e-16 ***
minimum_nights               -1.109e-02  4.422e-04 -25.073  < 2e-16 ***
last_review2016-2020         -2.563e-01  8.181e-02  -3.133  0.00174 **
last_review2021-2024         -3.332e-01  8.116e-02  -4.105 4.07e-05 ***
calculated_host_listings_count -6.003e-03  4.131e-04 -14.532  < 2e-16 ***
availability_365              1.681e-04  3.378e-05   4.977 6.55e-07 ***
rating                        2.235e-01  1.513e-02  14.773  < 2e-16 ***
bedrooms                      1.787e-01  9.614e-03  18.587  < 2e-16 ***
beds                          7.546e-02  6.456e-03  11.690  < 2e-16 ***
baths                         1.313e-01  1.401e-02   9.367  < 2e-16 ***
regionNorth                  -6.492e-02  2.282e-02  -2.845  0.00445 **
regionSouth                   2.671e-01  1.151e-02  23.209  < 2e-16 ***
regionWest                    2.381e-01  1.036e-01   2.297  0.02164 *
numberofreviews1             -2.692e-02  4.889e-03  -5.506 3.73e-08 ***
reviews_per_month1            1.095e-02  5.849e-03   1.873  0.06111 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4946 on 12821 degrees of freedom
Multiple R-squared:  0.4242,    Adjusted R-squared:  0.4235
F-statistic: 590.3 on 16 and 12821 DF,  p-value: < 2.2e-16
```

Il modello ottenuto attraverso la modalità backward (AIC) non presenta più due variabili: license e number of reviews ltm.
È stato calcolata poi il VIF (variance inflation factor), una misura statistica che ha permesso di confermare l'assenza di multicollinearità nel modello.

```
> vif(mod.backward)
                                    GVIF Df GVIF^(1/(2*Df))
room_type                       1.259842  2        1.059446
minimum_nights                  1.251750  1        1.118816
last_review                     1.276181  2        1.062865
calculated_host_listings_count  1.148678  1        1.071764
availability_365                1.073978  1        1.036329
rating                          1.062281  1        1.030670
bedrooms                        2.282462  1        1.510782
beds                            2.258108  1        1.502700
baths                           1.212240  1        1.101018
region                          1.081710  3        1.013177
numberofreviews1                1.825413  1        1.351079
reviews_per_month1              2.359124  1        1.535944
```

A questo punto è stato svolto un controllo dei residui:



Il grafico in alto a sinistra mostra la distribuzione dei residui rispetto ai valori predetti. La nuvola di punti, seppur molto concentrata, si presenta abbastanza omogenea e senza pattern evidenti, confermando che l'ipotesi di linearità è rispettata e che non sono presenti segnali preoccupanti di eteroschedasticità. Anche la linea rossa di tendenza si mantiene vicina allo zero lungo tutto l'intervallo dei valori stimati.
Il grafico Scale-Location conferma quanto osservato, con una dispersione dei residui standardizzati abbastanza costante su tutto il range. L'assenza di un evidente allargamento o restringimento della nuvola di punti indica la presenza di omoschedasticità.
Il Q-Q plot presenta invece una situazione meno ideale. I residui seguono la linea teorica normale nella parte centrale, ma le code deviano verso l'esterno, in particolare la coda superiore dove i punti si allontanano dalla bisettrice. Questo suggerisce la presenza di asimmetria a sinistra, con i residui che quindi non sono perfettamente normali. Durante l'analisi descrittiva sono state comunque analizzate variabili che verranno escluse successivamente per problemi legati alla normalità della distribuzione, così da poter migliorare il modello.
Infine, il grafico Residuals vs Leverage mostra la presenza di alcune osservazioni potenzialmente influenti, contrassegnate da etichette numeriche. Tuttavia, nessuno di questi punti oltrepassa le linee di Cook's distance, suggerendo che anche se il leverage di queste osservazioni risulta più elevato, non sono presenti outlier in grado di influenzare in modo significativo la stima dei coefficienti.

Modalità forward:

Il modello di partenza prevedeva come variabili esplicative del prezzo degli alloggi: bedrooms, baths, beds, room type e region.
Il risultato ottenuto è lo stesso di quello sopra riportato. Anche in questo caso ad essere escluse dal modello finale sono le variabili license e number of review last month.

```
> summary(mod.forward)

Call:
lm(formula = price1 ~ bedrooms + baths + beds + room_type + region +
    minimum_nights + calculated_host_listings_count + rating +
    numberofreviews1 + last_review + availability_365 + reviews_per_month1,
    data = NY)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1731 -0.3386 -0.0287  0.2984  2.8689

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.018e+00  1.110e-01  36.209  < 2e-16 ***
bedrooms                        1.787e-01  9.614e-03  18.587  < 2e-16 ***
baths                           1.313e-01  1.401e-02   9.367  < 2e-16 ***
beds                            7.546e-02  6.456e-03  11.690  < 2e-16 ***
room_typePrivate room          -5.041e-01  9.959e-03 -50.616  < 2e-16 ***
room_typeShared room           -6.350e-01  3.816e-02 -16.639  < 2e-16 ***
regionNorth                    -6.492e-02  2.282e-02  -2.845  0.00445 **
regionSouth                     2.671e-01  1.151e-02  23.209  < 2e-16 ***
regionWest                      2.381e-01  1.036e-01   2.297  0.02164 *
minimum_nights                 -1.109e-02  4.422e-04 -25.073  < 2e-16 ***
calculated_host_listings_count -6.003e-03  4.131e-04 -14.532  < 2e-16 ***
rating                          2.235e-01  1.513e-02  14.773  < 2e-16 ***
numberofreviews1               -2.692e-02  4.889e-03  -5.506 3.73e-08 ***
last_review2016-2020           -2.563e-01  8.181e-02  -3.133  0.00174 **
last_review2021-2024           -3.332e-01  8.116e-02  -4.105 4.07e-05 ***
availability_365                1.681e-04  3.378e-05   4.977 6.55e-07 ***
reviews_per_month1              1.095e-02  5.849e-03   1.873  0.06111 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4946 on 12821 degrees of freedom
Multiple R-squared:  0.4242,    Adjusted R-squared:  0.4235
F-statistic: 590.3 on 16 and 12821 DF,  p-value: < 2.2e-16
```

Durante le analisi descrittiva e di correlazione erano state evidenziate variabili che, per diversi motivi, non potevano essere inserite all'interno del modello di regressione lineare per spiegare il prezzo degli alloggi.
Si parte allora dal modello ottenuto attraverso la stepwise regression e si procede a ritroso eliminando le variabili inadatte, in modo tale da giungere ad una situazione ideale.
Inoltre, si modifica l'ordine delle due variabili factor "region" e "last review", in modo tale da definire come level di riferimento "South" per region, mentre far assumere la posizione intermedia a "2021-2024" per last review.

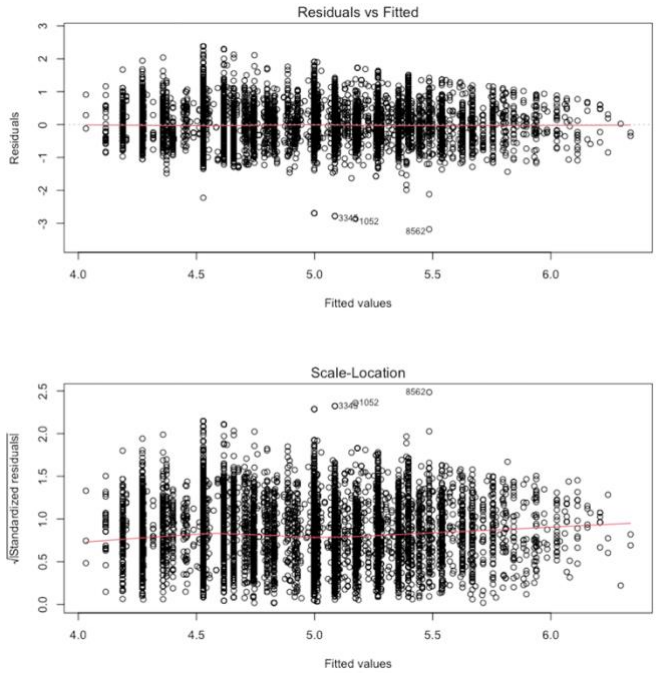
```
> summary(mod.backward)

Call:
lm(formula = price1 ~ room_type + minimum_nights + last_review +
    calculated_host_listings_count + availability_365 + rating +
    bedrooms + beds + baths + region + numberofreviews1 + reviews_per_month1,
    data = NY)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1731 -0.3386 -0.0287  0.2984  2.8689

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                     4.018e+00  1.110e-01  36.209  < 2e-16 ***
room_typePrivate room          -5.041e-01  9.959e-03 -50.616  < 2e-16 ***
room_typeShared room           -6.350e-01  3.816e-02 -16.639  < 2e-16 ***
minimum_nights                 -1.109e-02  4.422e-04 -25.073  < 2e-16 ***
last_review2016-2020           -2.563e-01  8.181e-02  -3.133  0.00174 **
last_review2021-2024           -3.332e-01  8.116e-02  -4.105 4.07e-05 ***
calculated_host_listings_count -6.003e-03  4.131e-04 -14.532  < 2e-16 ***
availability_365                1.681e-04  3.378e-05   4.977 6.55e-07 ***
rating                          2.235e-01  1.513e-02  14.773  < 2e-16 ***
bedrooms                        1.787e-01  9.614e-03  18.587  < 2e-16 ***
beds                            7.546e-02  6.456e-03  11.690  < 2e-16 ***
baths                           1.313e-01  1.401e-02   9.367  < 2e-16 ***
regionNorth                    -6.492e-02  2.282e-02  -2.845  0.00445 **
regionSouth                     2.671e-01  1.151e-02  23.209  < 2e-16 ***
regionWest                      2.381e-01  1.036e-01   2.297  0.02164 *
numberofreviews1               -2.692e-02  4.889e-03  -5.506 3.73e-08 ***
reviews_per_month1              1.095e-02  5.849e-03   1.873  0.06111 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4946 on 12821 degrees of freedom
Multiple R-squared:  0.4242,    Adjusted R-squared:  0.4235
F-statistic: 590.3 on 16 and 12821 DF,  p-value: < 2.2e-16
```

Il modello iniziale si presentava così.

Le prime ad essere state eliminate sono: "reviews per month", "calculated host listings count", "minimum nights", "availability 365" e "number of reviews". Questo per due motivi principali: la mancata possibilità, attraverso mezzi adeguati, di normalizzare la distribuzione, o perché presentavano una influenza quasi inesistente sul prezzo.

La variabile "region" assumeva nel level West un p-value di 0,8789, quindi molto alto.
Questo significa che, rispetto alla regione di riferimento (South) la zona West non ha un effetto significativo sul prezzo.
È stata analizzata allora la situazione senza considerare nel modello tale livello, ma il risultato è rimasto pressoché invariato. Dato che la variabile region nel complesso è utile, si è deciso di continuare senza variazioni.

Una volta esclusa anche “rating” (per gli stessi motivi sopracitati) restano variabili principalmente categoriche o con pochi valori distinti. Questo fa si che i valori predetti non risultino più distribuiti in maniera continua ma si concentrino su pochi livelli distinti (linee verticali nei grafici in basso e alto a sx).

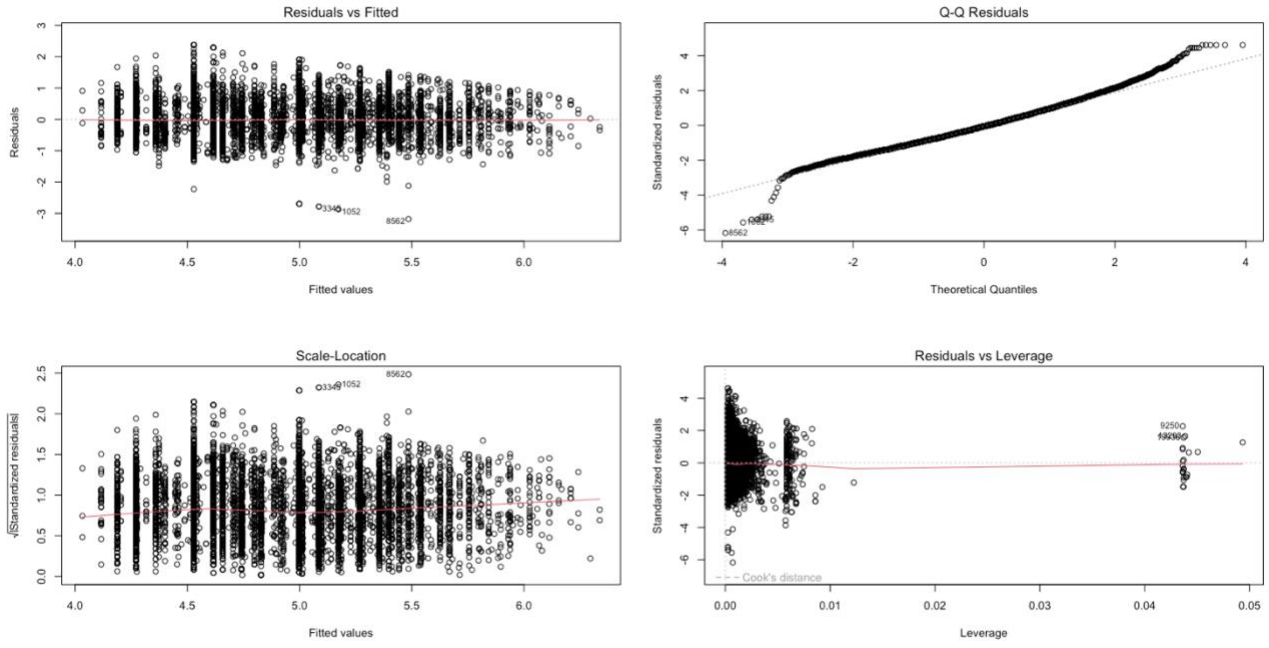


Per provare a migliorare questa situazione si è tentato di modificare da factor a quantitativa discreta la variabile “last review”. Essendo che originariamente le osservazioni si presentavano come nel seguente esempio: “2024-01-01”, è stata data prima l’indicazione di prendere come riferimento 2024 come anno 1 e, man mano che si procedeva a ritroso negli anni, le unità statistiche dovevano assumere un valore pari a “1(2024) + il numero di anni che passava da questo”.
Tale trasformazione ha permesso di ottenere una nuvola di punti nei grafici Residual vs Fitted e Scale-Location, ma ha causato l’accentuarsi della deviazione della coda superiore nel grafico QQ plot (i residui più grandi sono fuori scala. Arrivano a oltre 10 deviazioni standard).
Questo indica forti outlier e una forte non-normalità dei residui.
Dato che l’obiettivo di questo elaborato è fare inferenza sulle relazioni tra variabili e prezzo, si accetta che i valori predetti non risultino più distribuiti in maniera continua ma si concentrino su pochi livelli distinti. Si decide quindi di continuare con la variabile nella forma inizialmente prevista.

Una volta ottenuto il modello finale si interviene sugli outlier indicati da R durante il controllo dei residui. Nella coda inferiore del QQ plot, così come nella metà inferiore dei grafici Residual vs Fitted e Scale-Location, si presentano svariate u.s. riferite a case/appartamenti interi con prezzo=10 $.
La media dei prezzi per questo level del factor è di 100, ma per quanto possa sembrare anomala la situazione, ben 7272 u.s. presentano un prezzo uguale a 10 per questo level. Si decide quindi di non eliminare niente dal dataset.

Modello finale:

```
> summary(lm.NY7)

Call:
lm(formula = price1 ~ room_type + bedrooms + beds + baths + region,
    data = NY)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1827 -0.3573 -0.0289  0.3155  2.3790

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             4.599253   0.017216 267.153   <2e-16 ***
room_typePrivate room -0.469850   0.010180 -46.156   <2e-16 ***
room_typeShared room  -0.625189   0.039695 -15.750   <2e-16 ***
bedrooms               0.182761   0.009969  18.333   <2e-16 ***
beds                   0.087360   0.006625  13.186   <2e-16 ***
baths                  0.129207   0.014581   8.861   <2e-16 ***
regionWest            -0.016476   0.107737  -0.153    0.878
regionEast            -0.257434   0.011797 -21.823   <2e-16 ***
regionNorth           -0.341176   0.022113 -15.429   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5159 on 12828 degrees of freedom
Multiple R-squared:  0.3725,    Adjusted R-squared:  0.3721
F-statistic:   952 on 8 and 12828 DF,  p-value: < 2.2e-16
```



Il modello finale, dopo l’eliminazione delle variabili ritenute inappropriate, risulta semplificato rispetto a quello iniziale ottenuto tramite stepwise regression, mantenendo però una buona capacità esplicativa.
Ora sono incluse solamente le variabili più significative: “room type”, “bedrooms”, “beds”, “baths” e region. L’Adjusted R^2 scende leggermente rispetto al modello iniziale, passando da 0.4235 a 0.3721.

Questo è probabilmente dovuto alla riduzione del numero di variabili esplicative, ma viene compensato da un miglioramento della stabilità complessiva del modello.
Analizzando i grafici, quello dei residui rispetto ai valori predetti mostra una distribuzione alquanto omogenea attorno alla linea rossa, senza evidenti pattern sistematici. Anche se la dispersione risulta leggermente più ampia rispetto al modello iniziale, soprattutto per quanto riguarda i valori più bassi di fitted values, non si evidenzia alcun segnale preoccupanti di eteroschedasticità.
Il grafico in basso a sinistra conferma questa impressione: la varianza dei residui appare pressoché stabile lungo tutto il range dei valori stimati, e la linea di tendenza rossa si presenta più regolare rispetto a quella osservata nel modello iniziale, segno che l'omoschedasticità è stata migliorata.
Il Q-Q plot continua a mostrare qualche deviazione dalle code della distribuzione normale, cosa dovuta probabilmente all'elevata numerosità del campione di riferimento. Le u.s. evidenziate come outlier da R sono stati analizzati separatamente e si è deciso di mantenerle nel dataset. Si nota comunque una maggiore aderenza alla linea teorica nella parte centrale della distribuzione. Questo rappresenta un piccolo passo avanti rispetto al modello di partenza, anche se la non perfetta normalità dei residui nelle code rimane un elemento da tenere presente.
Infine, il grafico dei residui rispetto al leverage segnala la presenza di alcune osservazioni potenzialmente influenti (Private room con valori più elevati della media, ma comunque possibili) ma nessun punto supera la linea di Cook's distance.
Non sembrano quindi esserci osservazioni in grado di alterare in maniera significativa la stima dei coefficienti, e la situazione appare nel complesso più sotto controllo rispetto alla situazione iniziale.
È stato quindi ottenuto un modello più snello e interpretabile, che presenta però una buona capacità di spiegare la variabilità dei dati. Se da una parte si registra una lieve perdita in termini di R^2 aggiustato, dall'altro migliorano alcuni aspetti legati alla stabilità della varianza dei residui e alla semplicità del modello, che ora risulta più robusto e facilmente comunicabile.

Interpretazione del modello

```
> coefficients(lm.NY7)
      (Intercept) room_typePrivate room  room_typeShared room            bedrooms                beds               baths 
       4.59925282          -0.46984964          -0.62518888           0.18276125           0.08735973           0.12920712 
       regionWest           regionEast          regionNorth 
      -0.01647604          -0.25743369          -0.34117636 
```

Tenendo conto che l'intercetta indica il valore stimato di log(price) quando tutti gli altri coefficienti sono pari a 0 (cosa impossibile dato che ci sarà almeno una camera da letto in ogni alloggio), l'interpretazione di questa viene considerata non significativa.

1. b1: il coef. associato al level "Private room" della variabile "Room type" è -0,47. Questo significa che, a parità di altre condizioni, il prezzo di una camera privata in media costa (≈) il 47% in meno rispetto a quello di una casa/appartamento intero
2. b2: il coef. associato al level "Shared room" della variabile "Room type" è -0,63. Questo significa che, a parità di altre condizioni, il prezzo di una camera condivisa in media costa (≈) il 63% in meno rispetto a quello di una casa/appartamento intero
3. b3: il coef. associato alla variabile numero di camere è 0,18. Per ogni aumento dell'1% del numero di camere presenti in un alloggio il prezzo in media aumenta (≈) del 18%
4. b4: il coef. associato alla variabile numero di letti è 0,09. Per ogni aumento dell'1% del numero di letti presenti in un alloggio il prezzo in media aumenta (≈) del 9%

5. b5: il coef. associato alla variabile numero di bagni è 0,13, per ogni aumento dell'1% del numero di bagni presenti in un alloggio il prezzo in media aumenta (≈) del 13%
6. b6: il coef. associato al level "West" della variabile "Region" è -0,02. Questo significa che, parità di altre condizioni, un alloggio situato nella zona ovest costa (≈) il 2% in meno rispetto ad uno situato nella zona sud
7. b7: il coef. associato al level "East" della variabile "Region" è -0,26, a parità di altre condizioni un alloggio situato nella zona est costa (≈) il 26% in meno rispetto ad uno situato nella zona sud
8. b8: il coef. associato al level "North" della variabile "Region" è -0,34, a parità di altre condizioni un alloggio situato nella zona nord costa (≈) il 34% in meno rispetto ad uno situato nella zona sud

Previsioni

Quanto costa un alloggio a New York in tre diverse ipotetiche situazioni?

1) Casa/appartamento intero, spazioso e ben posizionato:

- Tipo di stanza: Entire home/apt
- Bedrooms: 4
- Beds: 5
- Baths: 3
- Regione: South (zona con prezzo più elevato)

→ Prezzo previsto ≈ **471,88** dollari a notte

2) Camera privata, con tutto il necessario e ben posizionata:

- Tipo di stanza: Private room
- Bedrooms: 1
- Beds: 3
- Baths: 1
- Regione: East

→ Prezzo previsto ≈ **85,29** dollari a notte

3) Camera condivisa, spazio ridotto e zona periferica:

- Tipo di stanza: shared room
- Bedrooms: 1
- Beds: 1
- Baths: 1
- Regione: North

→ Prezzo previsto ≈ **56,39** dollari a notte

CONCLUSIONI

Il seguente elaborato aveva come obietti determinare quali fattori influenzavano maggiormente i prezzi degli alloggi a New York, in modo tale da provare a prevedere i prezzi di questi in diversi scenari possibili.
Dopo aver eliminato dal dataset variabili ritenute ininfluenti (es. host id) è stata fatta un'analisi descrittiva che ha permesso di eliminare osservazioni anomale e modificare la natura di alcune variabili per ottenere il miglior risultato possibile. Successivamente si è passati ad un'analisi volta a comprendere la relazione tra prezzo e variabili in modo tale da capire l'influenza esercitata da queste sulla nostra variabile dipendente.
Tutto questo ha permesso di costruire un modello di regressione lineare composto dalle variabili "room_type", "bedrooms", "beds", "baths" e "region", il quale rispetta tutte le assunzioni necessarie per costruire questa tipologia di regressione ed è in grado di spiegare i prezzi degli alloggi.
Tuttavia, questo permette di spiegare solo il 37% della nostra variabile dipendente, indicando implicitamente che esistono altre variabili risposta rilevanti. È possibile pensare, ad esempio, alla distanza dalle attrazioni turistiche principali, o al punteggio guadagnato dall'host sulla piattaforma, oltre che alla metratura precisa dell'alloggio o alle condizioni di rimborso in caso di annullamento.
Bisogna inoltre ricordare le diverse modifiche apportate per semplificare le variabili. Si potrebbe, ad esempio, effettuare un'analisi mantenendo tutti i quartieri di New York come singoli e vedendo se l'influenza di questi cambia sulla variabile prezzo oppure ha lo stesso effetto della variabile complementare aggregata; oppure si potrebbe studiare più attentamente la certificazione OSE (per affitti a breve termine) e vedere se è uguale per tutti o presenta diverse tipologie al suo interno, studiando poi solamente la situazione per gli affitti a breve termine. La bontà del dataset permetterebbe quindi numerosissime strade alternative di analisi, entrando più nello specifico in determinate variabili.

SITOGRAFIA

New York Airbnb Open Data 2024
https://www.kaggle.com/datasets/vrindakallu/new-york-dataset

Inside Airbnb: Adding data to the debate
https://insideairbnb.com/get-the-data/

NYC Short-Term Rental Registration
https://strr-portal.ose.nyc.gov/s/?language=en_US

Materiale Moodle del Corso Metodi Statistici per la Business Analysis
https://moodle.unive.it/course/view.php?id=21119

R Programming 101 YouTube channel. Playlist: "R programming for beginners"
https://www.youtube.com/@RProgramming101

CODICE R

```
NewYork<-read.csv("new_york_listings_2024.csv",stringsAsFactors = TRUE)
table(NewYork$neighbourhood)

NewYork$region <- NA
NewYork$region[NewYork$neighbourhood %in% c("Marble Hill", "Riverdale", "Spuyten Duyvil",
"Fieldston", "Kingsbridge", "Norwood", "Woodlawn", "Wakefield", "Williamsbridge", "Olinville",
"Edenwald", "Baychester", "Eastchester", "Bronxdale", "Morris Park", "Schuylerville", "Pelham
Bay", "Pelham Gardens", "City Island", "Throgs Neck", "Co-op City", "Hunts Point", "Port
Morris", "Mott Haven", "Morrisania", "Longwood", "Melrose", "Mount Eden", "Mount Hope",
"Tremont", "University Heights", "Highbridge", "Concourse", "Concourse Village", "Claremont
Village", "East Morrisania", "West Farms")] <- "North"

NewYork$region[NewYork$neighbourhood %in% c("Battery Park City", "Financial District",
"Tribeca", "Chinatown", "Two Bridges", "Lower East Side", "SoHo", "Nolita", "Little Italy",
"Greenwich Village", "West Village", "NoHo", "East Village", "Gramercy", "Kips Bay", "Murray
Hill", "Stuyvesant Town", "Chelsea", "Flatiron District", "Midtown", "Hell's Kitchen", "Theater
District", "Upper West Side", "Upper East Side", "Morningside Heights", "Harlem", "Civic Center",
"DUMBO", "Brooklyn Heights", "Cobble Hill", "Boerum Hill", "Carroll Gardens", "Red Hook",
"Gowanus", "Park Slope", "Prospect Heights", "Prospect-Lefferts Gardens", "Windsor Terrace",
"South Slope", "Greenpoint", "Williamsburg", "Bushwick", "Bedford-Stuyvesant", "Clinton Hill",
"Fort Greene", "Downtown Brooklyn", "Vinegar Hill", "Navy Yard", "Borough Park",
"Bensonhurst", "Bath Beach", "Dyker Heights", "Bay Ridge", "Sunset Park", "Breezy Point",
"Brighton Beach", "Coney Island", "Sea Gate", "Gravesend", "Sheepshead Bay", "Manhattan
Beach", "Mill Basin", "Bergen Beach", "Canarsie", "Flatlands", "East Flatbush", "Brownsville",
"East New York", "Springfield Gardens", "Rosedale", "Laurelton", "Cambria Heights", "St.
Albans", "South Ozone Park", "Ozone Park", "Howard Beach", "Lighthouse Hill", "Arden
Heights", "Great Kills", "Tottenville", "Rossville", "Huguenot", "Prince's Bay", "Woodrow",
"Eltingville", "New Dorp Beach", "Grant City", "Emerson Hill", "Todt Hill", "Dongan Hills",
"Annadale", "Richmondtown", "Oakwood", "South Beach", "Midland Beach", "Shore Acres",
"Rosebank", "St. George", "Tompkinsville", "Stapleton", "Arrochar", "New Brighton", "Silver
Lake", "Grymes Hill", "Randall Manor", "Port Richmond", "West Brighton", "Mariners Harbor",
"Howland Hook", "Graniteville", "New Springville", "Travis", "Bloomfield", "Chelsea, Staten
Island", "Lighthouse Hill", "Willowbrook")] <- "South"

NewYork$region[NewYork$neighbourhood %in% c("Astoria", "Long Island City", "Ditmars
Steinway", "Woodside", "Sunnyside", "Elmhurst", "Jackson Heights", "Corona", "Rego Park",
"Middle Village", "Maspeth", "Glendale", "Ridgewood", "Cypress Hills", "East Elmhurst",
"College Point", "Whitestone", "Bayside", "Little Neck", "Fresh Meadows", "Kew Gardens Hills",
"Kew Gardens", "Briarwood", "Jamaica Hills", "Jamaica Estates", "Jamaica", "Queens Village",
"Bellerose", "Holliswood", "Hollis", "Springfield Gardens", "Rosedale", "Laurelton", "Cambria
Heights", "St. Albans", "South Ozone Park", "Ozone Park", "Howard Beach", "Edgemere",
"Arverne", "Rockaway Beach", "Far Rockaway", "Bayswater", "Belle Harbor", "Neponsit")] <-
"East"

NewYork$region[NewYork$neighbourhood %in% c("Columbia St", "Concord", "Clifton",
"Westerleigh")] <- "West"

NewYork$region <- as.factor(NewYork$region)
table(NewYork$region)
```

```
View(NewYork$region)

NewYork$neighbourhood <- NULL
NewYork$neighbourhood_group <- NULL
NewYork$host_name <- NULL
NewYork$name <- NULL
NewYork$host_id <- NULL
NewYork$id <- NULL
NewYork$longitude <- NULL
NewYork$latitude <- NULL
str(NewYork)

options(scipen=999)

table(NewYork1$baths)

NewYork$bedrooms <- as.numeric(as.character(NewYork$bedrooms))
NewYork$bedrooms[is.na(NewYork$bedrooms)] <- 1
NewYork$bedrooms <- as.integer(as.character(NewYork$bedrooms))
NewYork$baths <- as.numeric(as.character(NewYork$baths))

NewYork$rating <- as.numeric(as.character(NewYork$rating))

NewYork$last_review <- as.Date(NewYork$last_review)
NewYork$last_review <- cut(NewYork$last_review,breaks = as.Date(c("2011-01-01", "2016-01-
01", "2021-01-01", "2025-01-01")),labels = c("2011-2015", "2016-2020", "2021-2024"),right =
FALSE)
library(dplyr)
NewYork$license <- ifelse(grepl("OSE-ST[Rr]REG-", NewYork$license), "License",
                          ifelse(NewYork$license == "License", "No License", "Exempt"))
NewYork$license <- factor(NewYork$license, levels = c("License", "No License", "Exempt"))
str(NewYork)
NewYork<-na.omit(NewYork)
is.na(NewYork)

###### VISUALIZZAZIONE DEI DATI

# PRICE

boxplot(NewYork$price,main="Boxplot della variabile Prezzo",col="light blue")
summary(NewYork$price)
z_scores_Price <- scale(NewYork$price)
z_scores_Price
outliers_z_Price <- NewYork$price[abs(z_scores_Price) > 3]
print(outliers_z_Price)
index_outliers<-which(NewYork$price>1831)
print(index_outliers)
NewYork[index_outliers,]
hist(NewYork$price,main="Istogramma del Prezzo", xlab="Prezzo",ylab="Frequenza",col="light
blue")
```

```
hist(log(NewYork$price),main="Istogramma del log(Prezzo)",
xlab="log(Prezzo)",ylab="Frequenza",col="light  blue")
summary(log(NewYork$price))
NewYork<- NewYork[NewYork$price<=1000, ]
price1<-log(NewYork$price)
NewYork_lr$price<-log(NewYork_lr$price)

quantile(NewYork$price)
par(mfrow=c(1,1))
sd(log(NewYork$price))
table(NewYork$price)

# REVIEW PER MONTHS

table(NewYork$reviews_per_month)
summary(NewYork$reviews_per_month)
boxplot(NewYork$reviews_per_month,main="Boxplot della variabile Review x Months")
hist(NewYork$reviews_per_month,main="Istogramma delle Review x Months", xlab="Review x
Months",ylab="Frequenze",col="orange")
which(NewYork$reviews_per_month==0)
z_scores_Reviewpermonths <- scale(NewYork$reviews_per_month)
z_scores_Reviewpermonths
outliers_z_Reviewpermonths <- NewYork$reviews_per_month[abs(z_scores_Reviewpermonths) >
3]
print(outliers_z_Reviewpermonths)
index_outliers1<-which(NewYork$reviews_per_month>7.82)
print(index_outliers1)
NewYork<- NewYork[NewYork$reviews_per_month<=7.00, ]


boxplot(NewYork$reviews_per_month,main="Boxplot della variabile Review x Months")
hist(log(NewYork$reviews_per_month),main="Istogramma del log(Review per Months)",
xlab="log(Review per Months)",ylab="Frequenze",col="light blue")
summary(log(NewYork$reviews_per_month)+0.1)
quantile(NewYork$reviews_per_month)


# RATING

table(NewYork$rating)
summary(NewYork$rating)
boxplot(NewYork$rating,main="Boxplot della variabile Rating")
hist(NewYork$rating,main="Istogramma delle rating", xlab="rating",ylab="Frequenza",col="light
blue")
hist(log(NewYork$rating),main="Istogramma del log(rating)",
xlab="Log(rating)",ylab="Frequenze",col="light blue")
summary(log(NewYork$rating))
NewYork$rating <- NewYork$rating[!is.na(NewYork$rating)]
percentile_1R <- quantile(NewYork$rating, 0.01)
percentile_1R
percentile_99R <- quantile(NewYork$rating, 0.99,na.rm=TRUE)
```

```
Rating_winsorizzata <- ifelse(NewYork$rating < percentile_1R, percentile_1R,
                      ifelse(NewYork$rating > percentile_99R, percentile_99R, NewYork$rating))
Rating_winsorizzata
Rating1<-log(Rating_winsorizzata)
hist(log(Rating1),main="Istogramma del log(Rating)",
xlab="log(Rating)",ylab="Frequenze",col="light blue")
summary(log(Rating1))
sd(log(Rating_winsorizzata))
install.packages("moments")
library(moments)
skewness(Rating1)


#BATHS


table(NewYork$baths)
summary(NewYork$baths)
NewYork[NewYork$baths==4.5,]
boxplot(NewYork$baths,main="Boxplot della variabile Baths")
hist(NewYork$baths,main="Istogramma di baths", xlab="Baths",ylab="Frequenze",col="light
blue")
NewYork<- NewYork[NewYork$baths<=4.5, ]
indici_zeri <- which(NewYork$baths == 0)
NewYork<- NewYork[-indici_zeri, ]
summary(NewYork$baths)
boxplot(NewYork$baths,main="Boxplot della variabile Baths")
hist((log(NewYork$baths)+0.1),main="Istogramma del log(Baths)",
xlab="Log(Baths)",ylab="Frequenze",col="light blue")
summary(log(NewYork$baths)+0.1)
sd(log(NewYork$baths))
Baths1<-(log(NewYork$baths)+0.1)

NewYork$baths <- ifelse(NewYork$baths == 1.5, 1,
                        ifelse(NewYork$baths == 2.5, 2,
                               ifelse(NewYork$baths == 3.5, 3,
                                      ifelse(NewYork$baths == 4.5, 4, NewYork$baths))))
table(NewYork$baths)
NewYork$baths <- as.integer(as.numeric(NewYork$baths))
NewYork <- NewYork[NewYork$baths<=3, ]
NewYork_lr <- NewYork_lr[NewYork_lr$baths<=3, ]

par(mfrow=c(1,3))
hist(log(NewYork$review_per_months),main="Istogramma del log(Review per Months)",
xlab="log(Review x Months)",ylab="Frequenze",col="light blue")
hist(Rating1,main="Istogramma del log(Rating)", xlab="log(Rating)",ylab="Frequenze",col="light
blue")
hist(NewYork$baths,main="Istogramma del baths", xlab="baths",ylab="Frequenze",col="light
blue")

summary(NewYork$reviews_per_month)
```

```
summary(NewYork$rating)
summary(NewYork$baths)
quantile(NewYork$reviews_per_month)
quantile(NewYork$rating)
quantile(NewYork$baths)



#  MINIMUM NIGHTS

table(NewYork$minimum_nights)
summary(NewYork$minimum_nights)
boxplot(NewYork$minimum_nights,main="Boxplot della variabile minimum nights")
hist(NewYork$minimum_nights,main="Istogramma della variabile minimum nights",
xlab="minimum nights",ylab="Frequenze",col="green")
z_scores_Minnights <- scale(NewYork$minimum_nights)
z_scores_Minnights
outliers_z_Minnights <- NewYork$minimum_nights[abs(z_scores_Minnights) > 3]
print(outliers_z_Minnights)
index_outliers2<-which(NewYork$minimum_nights>120)
print(index_outliers2)
NewYork<- NewYork[NewYork$minimum_nights<=90, ]



#CALCULATED HOST LISTINGS COUNT

boxplot(NewYork$calculated_host_listings_count,main="Boxplot della variabile calculated host
listing count")
hist(NewYork$calculated_host_listings_count,main="Istogramma della variabile calculated host
listing count", xlab="calculated host listing count",ylab="Frequenze",col="green")
table(NewYork$calculated_host_listings_count)
summary(NewYork$calculated_host_listings_count)
z_scores_calculated <- scale(NewYork$calculated_host_listings_count)
z_scores_calculated
outliers_z_calculated<- NewYork$calculated_host_listings_count[abs(z_scores_calculated) > 2]
print(outliers_z_calculated)
table(outliers_z_calculated)
NewYork<-NewYork[NewYork$calculated_host_listings_count<=83,]

#AVAILABILITY 365

boxplot(NewYork$availability_365,main="Boxplot della variabile availability_365")
hist(NewYork$availability_365,main="Istogramma della variabile availability 365",
xlab="availability 365",ylab="Frequenze",col="green")
summary(NewYork$availability_365)
table(NewYork$availability_365)

par(mfrow=c(1,3))
hist(NewYork$minimum_nights,main="Istogramma della variabile minimum nights",
xlab="minimum nights",ylab="Frequenze",col="blue")
```

```
hist(NewYork$calculated_host_listings_count,main="Istogramma della variabile calculated host listing count", xlab="calculated host listing count",ylab="Frequenze",col="blue")
hist(NewYork$availability_365,main="Istogramma della variabile availability 365", xlab="availability 365",ylab="Frequenze",col="blue")



summary(NewYork$minimum_nights)
summary(NewYork$calculated_host_listings_count)
summary(NewYork$availability_365)
quantile(NewYork$minimum_nights)
quantile(NewYork$calculated_host_listings_count)
quantile(NewYork$availability_365)


#BEDROOMS

boxplot(NewYork$bedrooms,main="Boxplot della variabile Bedrooms")
hist(NewYork$bedrooms,main="Istogramma della variabile Bedrooms",xlab="Bedrooms",ylab="Frequenze",col="blue")
summary(NewYork$bedrooms)
table(NewYork$bedrooms)
prop.table(table(NewYork$bedrooms))

NewYork[NewYork$bedrooms>7,]
z_scores_bedrooms <- scale(NewYork$bedrooms)
z_scores_bedrooms
outliers_z_bedrooms<- NewYork$bedrooms[abs(z_scores_bedrooms) > 3]
print(outliers_z_bedrooms)
table(outliers_z_bedrooms)
NewYork <- NewYork[NewYork$bedrooms<=5, ]
NewYork <- NewYork[NewYork$bedrooms<=4, ]

NewYork_lr <- NewYork_lr[NewYork_lr$bedrooms<=5, ]


#BEDS
summary(NewYork)
boxplot(NewYork$beds,main="Boxplot della variabile Beds")
hist(NewYork$beds,main="Istogramma della variabile Beds",xlab="Beds",ylab="Frequenze",col="blue")
summary(NewYork$beds)
z_scores_beds <- scale(NewYork$beds)
z_scores_beds
outliers_z_beds<- NewYork$beds[abs(z_scores_beds) > 3]
print(outliers_z_beds)
table(outliers_z_beds)
NewYork <- NewYork[NewYork$beds<=6, ]
NewYork_lr <- NewYork_lr[NewYork_lr$beds<=6, ]
```

```
table(as.factor(NewYork$beds))
prop.table(table(NewYork$beds))
table(NewYork$beds)
par(mfrow=c(1,2))

#NUMBER OF REVIEWS

boxplot(NewYork$number_of_reviews,main="Boxplot della variabile number of reviews")
summary(NewYork$number_of_reviews)
table(NewYork$number_of_reviews)
head(NewYork$number_of_reviews)
tail(NewYork$number_of_reviews)
z_scores_numberofreviews <- scale(NewYork$number_of_reviews)
z_scores_numberofreviews
outliers_z_numberofreviews<- NewYork$number_of_reviews[abs(z_scores_numberofreviews) >
3]
print(outliers_z_numberofreviews)
table(outliers_z_numberofreviews)
NewYork <- NewYork[NewYork$number_of_reviews<=294, ]
hist(NewYork$number_of_reviews,main="Istogramma della variabile Number of
Reviews",xlab="Number of Reviews",ylab="Frequenza",col=110)
hist(log(NewYork$number_of_reviews),main="Istogramma del log(Number of
Reviews)",xlab="log(Number of Reviews)",ylab="Frequenza",col="blue")
summary(log(NewYork$number_of_reviews))
numberofreviews1<-log(NewYork$number_of_reviews)

#NUMBER OF REVIEWS LTM

boxplot(NewYork$number_of_reviews_ltm,main="Boxplot della variabile Number of Reviews
ltm")
summary(NewYork$number_of_reviews_ltm)
table(NewYork$number_of_reviews_ltm)
hist(NewYork$number_of_reviews_ltm,main="Istogramma della variabile Number of Reviews
ltm",xlab="Number of Reviews ltm",ylab="Frequenza",col="blue")
z_scores_number_of_reviews_ltm <- scale(NewYork$number_of_reviews_ltm)
z_scores_number_of_reviews_ltm
outliers_z_number_of_reviews_ltm<-
NewYork$number_of_reviews_ltm[abs(z_scores_number_of_reviews_ltm) > 2]
print(outliers_z_number_of_reviews_ltm)
table(outliers_z_number_of_reviews_ltm)
NewYork <- NewYork[NewYork$number_of_reviews_ltm<=44, ]
summary(NewYork$number_of_reviews_ltm)


par(mfrow=c(1,1))

summary(NewYork$number_of_reviews)
summary(NewYork$number_of_reviews_ltm)

quantile(NewYork$number_of_reviews)
```

```
quantile(NewYork$number_of_reviews_ltm)


cor(NewYork$number_of_reviews,NewYork$reviews_per_month,use="complete.obs")
cor(NewYork$number_of_reviews_ltm,NewYork$reviews_per_month,use="complete.obs")
cor(NewYork$number_of_reviews,NewYork$number_of_reviews_ltm,use="complete.obs")

NewYork$number_of_reviews_ltm <- NULL
NewYork$reviews_per_month <- NULL


par(mfrow=c(2,2))
#ROOM TYPE

table(NewYork$room_type)
NewYork <- NewYork[NewYork$room_type != "Hotel room", ]
NewYork$room_type <- factor(NewYork$room_type)
levels(NewYork$room_type)
barplot(table(NewYork$room_type),main="Grafico a barre per Room type", xlab="Room
type",ylab="Frequenze assolute",ylim=c(0,9000),col=c("yellow"))
prop.table(table(NewYork$room_type))


#REGION
table(NewYork$region)
barplot(table(NewYork$region),main="Grafico a barre per Region",
xlab="Region",ylab="Frequenze assolute",ylim=c(0,12000),col=c("orange"))
by(NewYork[,1],NewYork$region,summary)

#LICENSE

table(NewYork$license)
NewYork$license <- factor(NewYork$license, levels = c("License", "Exempt"))
pie(table(NewYork$license),main="Grafico a torta per License",labels =c("License", "Exempt"),col
= c("yellow","orange"))
NewYork[NewYork$license == "Exempt",2]
prop.table(table(NewYork$license))

#LAST REVIEW

table(NewYork$last_review)
pie(table(NewYork$last_review),main="Grafico a torta per Last Review",labels =c( "2011-
2015","2016-2020","2021-2024"),col = c("red","orange","yellow"))
prop.table(table(NewYork$last_review))


dim(NewYork)
str(NewYork)
```

```
####### ANALISI DI CORRELAZIONE

library(corrplot)

NewYork$price1 <- log(NewYork$price)
NewYork$rating1 <- log(Rating_winsorizzata)
NewYork$numberofreviews1 <- log(NewYork$number_of_reviews)
NewYork$reviews_per_month1 <- log(NewYork$reviews_per_month)

Quant_corr <- NewYork[, c("price1","minimum_nights", "availability_365",
"number_of_reviews_ltm",
                          "calculated_host_listings_count", "bedrooms", "beds", "baths",
                          "reviews_per_month1", "rating1", "numberofreviews1")]
cor_matrix <- cor(Quant_corr, use = "complete.obs")
corrplot.mixed(cor_matrix, use="complete.obs",upper="ellipse", tl.pos="lt")
print(cor_matrix)




par(mfrow=c(1,1))
par(mfrow=c(1,3))
tab1<-table(NewYork$region,NewYork$room_type)
tab1
plot(NewYork$region~NewYork$room_type,data=NewYork)
test1 <- chisq.test(tab1,correct=TRUE)
test1


tab2<-table(NewYork$region,NewYork$last_review)
tab2
plot(NewYork$region~NewYork$last_review,data=NewYork)
test2 <- chisq.test(tab2)
test2

tab3<-table(NewYork$region,NewYork$license)
tab3
plot(NewYork$region~NewYork$license,data=NewYork)
test3 <- chisq.test(tab3)
test3
table(NewYork$license)
is.na(NewYork$license)
par(mfrow=c(1,1))
tab4<-table(NewYork$room_type,NewYork$last_review)
tab4
plot(NewYork$room_typeNewYork$last_review,data=NewYork)
test4 <- chisq.test(tab4)
test4
```

```
tab5<-table(NewYork$room_type,NewYork$License1)
tab5
plot(NewYork$room_type,NewYork$license,data=NewYork)
test5 <- chisq.test(tab5)
test5

tab6<-table(NewYork$last_review,NewYork$license)
tab6
plot(NewYork$last_review,NewYork$license,data=NewYork)
test6 <- chisq.test(tab6)
test6

# TRA PREZZO E LE VARIABILI

#QUANTITATIVE

plot(log(NewYork$price),NewYork$rating,xlab="rating",ylab="price", main="Diagramma a
dispersione")
plot(log(NewYork$price),log(NewYork$rating),xlab="log(rating)",ylab="log(price)",
main="Diagramma a dispersione")
cor(log(NewYork$price),log(NewYork$rating),use="complete.obs")
cor(log(NewYork$price),NewYork$rating,use="complete.obs")



par(mfrow=c(1,1))
boxplot(log(NewYork$price)~NewYork$beds,main="boxplot condizionato log(price)~beds",
xlab="Beds",ylab="log(price)",col=2:4)
boxplot(log(NewYork$price)~NewYork$bedrooms,main="boxplot condizionato
log(price)~bedrooms",xlab="Bedrooms",ylab="log(price)", col=2:4)
cor(log(NewYork$price),NewYork$beds,use="complete.obs")
cor(log(NewYork$price),NewYork$bedrooms,use="complete.obs")
table(NewYork$bedrooms)
points(by(NewYork$price1, NewYork$bedrooms,mean),col="red")
lines(by(NewYork$price1, NewYork$bedrooms,mean),col="red")
points(by(NewYork$price1, NewYork$bedrooms,mean),col="red")
lines(by(NewYork$price1, NewYork$beds,mean),col="red")


par(mfrow=c(1,1))
boxplot(log(NewYork$price)~NewYork$baths,main="boxplot condizionato
log(price)~baths",xlab="Baths",ylab="price",col=2:4)
cor(log(NewYork$price),NewYork$baths,use="complete.obs")
table(NewYork$baths)
points(by(log(NewYork$price), NewYork$baths,mean),col="red")
lines(by(log(NewYork$price), NewYork$baths,mean),col="red")

plot(log(NewYork$price),NewYork$number_of_reviews, main="Diagramma a dispersione
log(price)~number_of_reviews",xlab="number_of_reviews)",ylab="price")
plot(log(NewYork$price),NewYork$numberofreviews1, main="D.D.
log(price)~log(number_of_reviews)",xlab="log(numberofreviews)",ylab="price")
```

```
cor(log(NewYork$price),NewYork$numberofreviews1,use="complete.obs")

par(mfrow=c(1,3))

plot(log(NewYork$price),NewYork$reviews_per_month,xlab="reviews per months",ylab="Price",
main="Diagramma a dispersione")
plot(log(NewYork$price),log(NewYork$reviews_per_month),xlab="log(reviews per
month)",ylab="log(price)", main="Diagramma a dispersione")
cor(log(NewYork$price),log(NewYork$reviews_per_month),use="complete.obs")
cor(log(NewYork$price),log(NewYork$reviews_per_month),use="complete.obs")

boxplot(log(NewYork$price)~NewYork$number_of_reviews_ltm,xlab="number of reviews
ltm",ylab="price", main="Boxplot c. log(price)~number_of_review_ltm",col=2:4)
cor(log(NewYork$price),NewYork$number_of_reviews_ltm,use="complete.obs")
table(NewYork$number_of_reviews_ltm)

table(NewYork$availability_365)
boxplot(NewYork$price1~NewYork$availability_365,xlab="availability 365",ylab="log(price)",
main="Boxplot c. log(price)~availability_365",col=2:4)
cor(NewYork$price1,NewYork$availability_365,use="complete.obs")


table(NewYork$calculated_host_listings_count)
boxplot(NewYork$price1~NewYork$calculated_host_listings_count,xlab="calcuated_host_listing_
counts",ylab="log(price)", main="Boxplot cond. log(price)~calculated_h._l._c.",col=2:4)
cor(NewYork$log_price,NewYork$calculated_host_listings_count,use="complete.obs")

table(NewYork$minimum_nights)
boxplot(NewYork$price1~NewYork$minimum_nights,xlab="minimum_nights",ylab="log(price)",
main="Boxplot c. log(price)~minimum_nights",col=2:4)
cor(NewYork$log_price,NewYork$minimum_nights,use="complete.obs")


##### QUALITATIVE

table(NewYork$room_type)
par(mfrow=c(2,2))
boxplot(log(NewYork$price)~NewYork$room_type,xlab="Room type ",ylab="Price",col=2:4)
boxplot(log(NewYork$price)~NewYork$license,xlab="license ",ylab="Price",col=2:4)
boxplot(log(NewYork$price)~NewYork$neighbourhood_group,xlab="neighbourhoud
",ylab="Price",col=2:4)
boxplot(log(NewYork$price)~NewYork$last_review,xlab="last review ",ylab="Price",col=2:4)



####### MODELLO DI REGRESSIONE LINEARE

corrplot.mixed(cor(NewYork[,c("log_price","minimum_nights","log_numberofreviews","availabilit
y_365","calculated_host_listing_count","rating","bedrooms","beds","baths")],use="complete.obs"),
upper="ellipse",tl.pos="lt")
```

```
NY<-NewYork
NY$price<- NULL
NY$rating1<- NULL
NY$number_of_reviews<-NULL
NY$reviews_per_month<- NULL

lm.full<-lm(price1~.,data=NY)
summary(lm.full)
mod.backward<-step(lm.full,direction="backward")
summary(mod.backward)
mod.backward$anova
cor(NY$price1,NY$reviews_per_month1,use="complete.obs")
install.packages("car")
library(car)
vif(mod.backward)
par(mfrow=c(2,2))
plot(mod.backward)
lm.base<-lm(price1~bedrooms+baths+beds+room_type+region,data=NY)
mod.forward<-step(lm.base,scope=list(upper=lm.full),direction="forward")
summary(mod.forward)
mod.forward$anova

contrasts(NY$region)
NY$region<-factor(NY$region,levels=c("South","West","East","North"))
NY$last_review<-factor(NY$last_review,levels=c("2011-2015","2021-2024","2016-2020"))

lm.NY<-
lm(price1~room_type+minimum_nights+last_review+calculated_host_listings_count+availability_
365+rating+bedrooms+beds+baths+region+numberofreviews1+reviews_per_month1,data=NY)
summary(lm.NY)

lm.NY1<-
lm(price1~room_type+minimum_nights+last_review+calculated_host_listings_count+availability_
365+rating+bedrooms+beds+baths+region+numberofreviews1,data=NY)
summary(lm.NY1)
par(mfrow=c(2,2))
plot(lm.NY1)

lm.NY2<-
lm(price1~room_type+minimum_nights+last_review+availability_365+rating+bedrooms+beds+bat
hs+region+numberofreviews1,data=NY)
summary(lm.NY2)
plot(lm.NY2)

lm.NY3<-
lm(price1~room_type+minimum_nights+last_review+rating+bedrooms+beds+baths+region+numb
erofreviews1,data=NY)
summary(lm.NY3)
plot(lm.NY3)
```

```
lm.NY4<-
lm(price1~room_type+last_review+rating+bedrooms+beds+baths+region+numberofreviews1,data=
NY)
summary(lm.NY4)
plot(lm.NY4)

lm.NY5<-
lm(price1~room_type+last_review+bedrooms+beds+baths+region+numberofreviews1,data=NY)
summary(lm.NY5)
plot(lm.NY5)

lm.NY6<-lm(price1~room_type+last_review+bedrooms+beds+baths+region,data=NY)
summary(lm.NY6)
plot(lm.NY6)
vif(lm.NY6)


NY_noWest <- subset(NY, region != "West")
NY_noWest$region <- droplevels(NY_noWest$region)
NY_noWest$region <- relevel(NY_noWest$region, ref = "South")
lm.noWest <- lm(price1 ~ room_type + last_review + bedrooms + beds + baths + region, data =
NY_noWest)
summary(lm.noWest)
plot(lm.noWest)

lm.NewYork6<-lm(price~room_type+last_review+bedrooms+beds+baths+region,data=NewYork)
summary(lm.NY6)
plot(lm.NY6)
vif(lm.NY6)
residuals_std <- rstandard(lm.NewYork6)
outliers <- which(abs(residuals_std) > 3)
outliers
NY[outliers, ]


NY$last_review <- as.character(NY$last_review)
NY$last_review_year <- as.numeric(substr(NY$last_review, 1, 4))
anno_max <- max(NY$last_review_year, na.rm = TRUE)
NY$last_review_numeric <- anno_max - NY$last_review_year + 1
lm.NY.7 <- lm(price1 ~ room_type+last_review_numeric + bedrooms + beds + baths + region, data
= NY)
summary(lm.NY.7)


lm.NY7<-lm(price1~room_type+bedrooms+beds+baths+region,data=NY)
summary(lm.NY7)
plot(lm.NY7)
par(mfrow=c(2,2))
```

```
library(dplyr)
NY %>%
  filter(room_type %in% c("Entire home/apt", "Private room")) %>%
  group_by(room_type) %>%
  summarise(media_price = mean(price, na.rm = TRUE))

NY[NY$room_type == "Entire home/apt" & NY$price <= 10, ]
NY[NY$room_type == "Private room" & NY$price == 10000, ]

NY <- NY[!rownames(NY) %in% c("11939"), ]


coefficients(lm.NY7)
```