REPORT

Metodi Informatici per la Gestione Aziendale

Università degli Studi Milano-Bicocca
AA 2020/2021
Sanvito Alessio 844785

INDICE

DATA ACQUISITION

Il dataset scelto è stato il seguente
https://www.kaggle.com/shree1992/housedata
che indica i prezzi di 4600 case negli USA; gli attributi di questo dataset sono 18, ovvero la data in cui viene valutato il prezzo dell'abitazione, il numero di camere da letto, il numero di bagni, il numero di piedi quadrati della zona abitabile, la dimensione del terreno (in piedi quadrati), il numero di piani (compresi soppalchi e/o seminterrati), la posizione della casa ovvero se si trova vicino al mare, un voto alla vista (su una scala da 0 a 4), le condizioni della casa (su una scala da 1 a 5), la superficie del piano terra (in piedi quadrati), dimensione del piano interrato (se c'è), anno di costruzione, anno di ristrutturazione (se presente), indirizzo, città, CAP e nazione.

Tramite il comando read.csv viene importato il dataset e sono fatte le prime analisi su di esso con str, head, tail, summary, describe. I risultati ottenuti sono i seguenti

```
> str(case.df)
'data.frame':   4600 obs. of  18 variables:
 $ date          : Factor w/ 70 levels "2014-05-02 00:00:00",..: 1 1 1 1 1 1 1 1 1 1 ...
 $ price         : num  313000 2384000 342000 420000 550000 ...
 $ bedrooms      : num  3 5 3 3 4 2 2 4 3 4 ...
 $ bathrooms     : num  1.5 2.5 2 2.25 2.5 1 2 2.5 2.5 2 ...
 $ sqft_living   : int  1340 3650 1930 2000 1940 880 1350 2710 2430 1520 ...
 $ sqft_lot      : int  7912 9050 11947 8030 10500 6380 2560 35868 88426 6200 ...
 $ floors        : num  1.5 2 1 1 1 1 1 2 1 1.5 ...
 $ waterfront    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ view          : int  0 4 0 0 0 0 0 0 0 0 ...
 $ condition     : int  3 5 4 4 4 3 3 3 4 3 ...
 $ sqft_above    : int  1340 3370 1930 1000 1140 880 1350 2710 1570 1520 ...
 $ sqft_basement: int  0 280 0 1000 800 0 0 0 860 0 ...
 $ yr_built      : int  1955 1921 1966 1963 1976 1938 1976 1989 1985 1945 ...
 $ yr_renovated : int  2005 0 0 0 1992 1994 0 0 0 2010 ...
 $ street        : Factor w/ 4525 levels "1 View Ln NE",..: 1523 3900 2292 4264 4353 3522 2287 2039 3370 3847 ...
 $ city          : Factor w/ 44 levels "Algona","Auburn",..: 37 36 19 4 32 36 32 22 28 36 ...
 $ statezip      : Factor w/ 77 levels "WA 98001","WA 98002",..: 63 59 27 8 32 55 32 24 28 55 ...
 $ country       : Factor w/ 1 level "USA": 1 1 1 1 1 1 1 1 1 1 ...
> head(case.df)
                 date   price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
1 2014-05-02 00:00:00  313000        3      1.50        1340     7912    1.5          0    0
2 2014-05-02 00:00:00 2384000        5      2.50        3650     9050    2.0          0    4
3 2014-05-02 00:00:00  342000        3      2.00        1930    11947    1.0          0    0
4 2014-05-02 00:00:00  420000        3      2.25        2000     8030    1.0          0    0
5 2014-05-02 00:00:00  550000        4      2.50        1940    10500    1.0          0    0
6 2014-05-02 00:00:00  490000        2      1.00         880     6380    1.0          0    0
  condition sqft_above sqft_basement yr_built yr_renovated            street       city
1         3       1340             0     1955         2005     18810 Densmore Ave N Shoreline
2         5       3370           280     1921            0          709 W Blaine St    Seattle
3         4       1930             0     1966            0 26206-26214 143rd Ave SE       Kent
4         4       1000          1000     1963            0          857 170th Pl NE   Bellevue
5         4       1140           800     1976         1992        9105 170th Ave NE    Redmond
6         3        880             0     1938         1994          522 NE 88th St    Seattle
  statezip country
1 WA 98133     USA
2 WA 98119     USA
3 WA 98042     USA
4 WA 98008     USA
5 WA 98052     USA
6 WA 98115     USA
> tail(case.df)
                    date    price bedrooms bathrooms sqft_living sqft_lot floors waterfront view
4595 2014-07-09 00:00:00 210614.3        3      2.50        1610     7223      2          0    0
4596 2014-07-09 00:00:00 308166.7        3      1.75        1510     6360      1          0    0
4597 2014-07-09 00:00:00 534333.3        3      2.50        1460     7573      2          0    0
4598 2014-07-09 00:00:00 416904.2        3      2.50        3010     7014      2          0    0
4599 2014-07-10 00:00:00 203400.0        4      2.00        2090     6630      1          0    0
4600 2014-07-10 00:00:00 220600.0        3      2.50        1490     8102      2          0    0
     condition sqft_above sqft_basement yr_built yr_renovated            street      city statezip
4595         3       1610             0     1994            0 26306 127th Ave SE      Kent WA 98030
4596         4       1510             0     1954         1979      501 N 143rd St   Seattle WA 98133
4597         3       1460             0     1983         2009   14855 SE 10th Pl  Bellevue WA 98007
4598         3       3010             0     2009            0    759 Ilwaco Pl NE    Renton WA 98059
4599         3       1070          1020     1974            0  5148 S Creston St   Seattle WA 98178
4600         4       1490             0     1990            0  18717 SE 258th St Covington WA 98042
     country
4595     USA
4596     USA
4597     USA
4598     USA
4599     USA
4600     USA
```

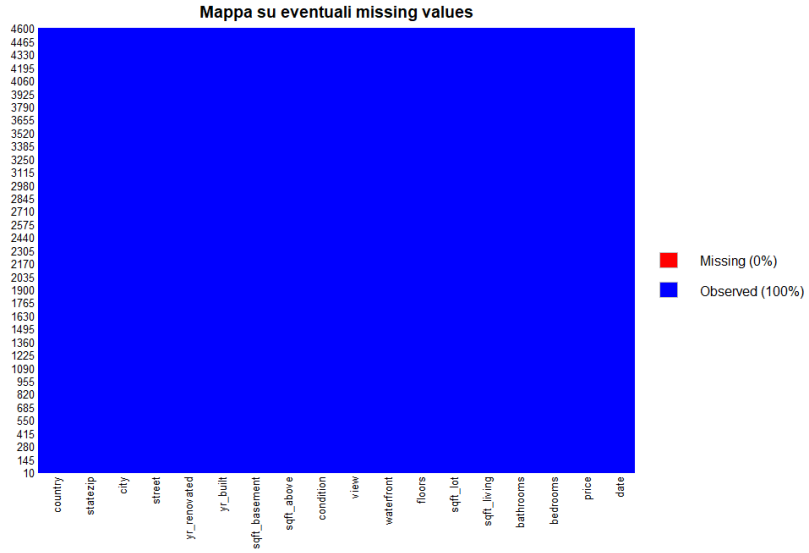
```
> summary(case.df)
                  date          price               bedrooms       bathrooms      sqft_living
 2014-06-23 00:00:00: 142   Min.   :       0   Min.   :0.000   Min.   :0.000   Min.   :  370
 2014-06-25 00:00:00: 131   1st Qu.:  322875   1st Qu.:3.000   1st Qu.:1.750   1st Qu.: 1460
 2014-06-26 00:00:00: 131   Median :  460943   Median :3.000   Median :2.250   Median : 1980
 2014-07-08 00:00:00: 127   Mean   :  551963   Mean   :3.401   Mean   :2.161   Mean   : 2139
 2014-07-09 00:00:00: 121   3rd Qu.:  654962   3rd Qu.:4.000   3rd Qu.:2.500   3rd Qu.: 2620
 2014-06-24 00:00:00: 120   Max.   :26590000   Max.   :9.000   Max.   :8.000   Max.   :13540
 (Other)            :3828
    sqft_lot           floors        waterfront           view          condition       sqft_above
 Min.   :    638   Min.   :1.000   Min.   :0.000000   Min.   :0.0000   Min.   :1.000   Min.   : 370
 1st Qu.:   5001   1st Qu.:1.000   1st Qu.:0.000000   1st Qu.:0.0000   1st Qu.:3.000   1st Qu.:1190
 Median :   7683   Median :1.500   Median :0.000000   Median :0.0000   Median :3.000   Median :1590
 Mean   :  14852   Mean   :1.512   Mean   :0.007174   Mean   :0.2407   Mean   :3.452   Mean   :1827
 3rd Qu.:  11001   3rd Qu.:2.000   3rd Qu.:0.000000   3rd Qu.:0.0000   3rd Qu.:4.000   3rd Qu.:2300
 Max.   :1074218   Max.   :3.500   Max.   :1.000000   Max.   :4.0000   Max.   :5.000   Max.   :9410

 sqft_basement       yr_built     yr_renovated                     street          city
 Min.   :   0.0   Min.   :1900   Min.   :   0.0   2520 Mulberry Walk NE:   4   Seattle :1573
 1st Qu.:   0.0   1st Qu.:1951   1st Qu.:   0.0   2500 Mulberry Walk NE:   3   Renton  : 293
 Median :   0.0   Median :1976   Median :   0.0   1018 NE 96th St      :   2   Bellevue: 286
 Mean   : 312.1   Mean   :1971   Mean   : 808.6   106 24th Ave E       :   2   Redmond : 235
 3rd Qu.: 610.0   3rd Qu.:1997   3rd Qu.:1999.0   11034 NE 26th Pl     :   2   Issaquah: 187
 Max.   :4820.0   Max.   :2014   Max.   :2014.0   1149-1199 91st Ave NE:   2   Kirkland: 187
                                                  (Other)              :4585   (Other) :1839

     statezip    country
 WA 98103: 148   USA:4600
 WA 98052: 135
 WA 98117: 132
 WA 98115: 130
 WA 98006: 110
 WA 98059: 106
 (Other) :3839
> describe(case.df)
               vars    n      mean        sd    median   trimmed       mad  min        max      range
date*             1 4600     37.14     19.71     39.00     37.39     25.20    1       70.0       69.0
price             2 4600 551962.99 563834.70 460943.46 489082.22 233509.50    0 26590000.0 26590000.0
bedrooms          3 4600      3.40      0.91      3.00      3.37      1.48    0        9.0        9.0
bathrooms         4 4600      2.16      0.78      2.25      2.12      0.74    0        8.0        8.0
sqft_living       5 4600   2139.35    963.21   1980.00   2038.67    845.08  370    13540.0    13170.0
sqft_lot          6 4600  14852.52  35884.44   7683.00   8473.19   4109.77  638  1074218.0  1073580.0
floors            7 4600      1.51      0.54      1.50      1.47      0.74    1        3.5        2.5
waterfront        8 4600      0.01      0.08      0.00      0.00      0.00    0        1.0        1.0
view              9 4600      0.24      0.78      0.00      0.00      0.00    0        4.0        4.0
condition        10 4600      3.45      0.68      3.00      3.33      0.00    1        5.0        4.0
sqft_above       11 4600   1827.27    862.17   1590.00   1718.39    726.47  370     9410.0     9040.0
sqft_basement    12 4600    312.08    464.14      0.00    225.44      0.00    0     4820.0     4820.0
yr_built         13 4600   1970.79     29.73   1976.00   1973.03     34.10 1900     2014.0      114.0
yr_renovated     14 4600    808.61    979.41      0.00    759.44      0.00    0     2014.0     2014.0
street*          15 4600   2266.39   1307.59   2264.50   2266.83   1681.27    1     4525.0     4524.0
city*            16 4600     26.67     11.98     33.00     28.03      4.45    1       44.0       43.0
statezip*        17 4600     39.74     20.92     42.00     40.19     26.69    1       77.0       76.0
country*         18 4600      1.00      0.00      1.00      1.00      0.00    1        1.0        0.0
                skew kurtosis      se
date*          -0.12    -1.17    0.29
price          24.77  1042.76 8313.29
bedrooms        0.46     1.23    0.01
bathrooms       0.62     1.86    0.01
sqft_living     1.72     8.28   14.20
sqft_lot       11.30   219.54  529.09
floors          0.55    -0.54    0.01
waterfront     11.68   134.34    0.00
view            3.34    10.45    0.01
condition       0.96     0.19    0.01
sqft_above      1.49     4.06   12.71
sqft_basement   1.64     4.07    6.84
yr_built       -0.50    -0.67    0.44
yr_renovated    0.39    -1.85   14.44
street*         0.00    -1.20   19.28
city*          -0.76    -0.79    0.18
statezip*      -0.15    -1.12    0.31
country*         NaN      NaN    0.00
```

Da queste analisi si può notare come la maggior parte delle colonne sia di tipo numerico (num o int) con qualche eccezione per indicatori come la via, la città, la data, lo stato di appartenenza e il CAP che sono factor.
Per chiarezza vengono esplicitati gli attributi:

- date (data in cui viene valutato il prezzo della casa)
- price (prezzo della casa)
- bedrooms (numero di camere da letto)
- bathrooms (numero di bagni)
- sqft_living (il numero di piedi quadrati della zona abitabile)
- sqft_lot (la dimensione in piedi quadrati del terreno)
- floors (il numero di piani)
- waterfront (un indicatore riguardo alla vicinanza della casa al mare)
- view (un voto da 0 a 4 alla vista)
- condition (la condizione in cui si trova la casa, espressa con un numero da 1 a 5)
- sqft_above (la superficie del piano terra in piedi quadrati)
- sqft_basement (la dimensione del piano interrato)

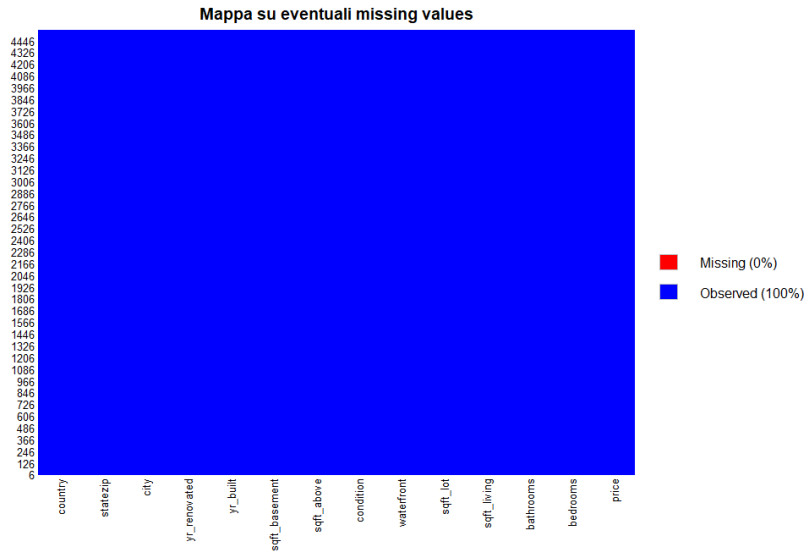
- yr_built (anno in cui è stata costruita la casa)
- yr_renovated (anno in cui è stata ristrutturata la casa)
- street (via in cui è situata la casa)
- city (città in cui è situata la casa)
- statezip (CAP della città in cui è situata la casa)
- country (nazione in cui è situata la casa)

In primis sono stati ricercati eventuali valori mancanti tramite la missmap che ha avuto come esito 0 valori mancanti



Poi è stato verificato che non ci fossero valori Na o NaN tramite l'istruzione any(is.na(case.df)) è si è notato che la funzione restituisce FALSE per indicare che non ci sono valori di questo tipo.
Successivamente sono state eliminate delle colonne ritenute poco funzionali allo studio di questo dataset come la data, l'indirizzo, il voto alla vista (la maggior parte ha valore 0), il numero di piani (in quanto in molti casi ha valore non intero), la nazione (in quanto poco utile all'analisi).

Sono state inoltre rimosse le righe che avevano un prezzo pari a 0.0 (case.df <- subset(case.df, price!= 0.0) perché poco utili ai fini del progetto.
Dopo queste operazioni viene controllato nuovamente (per scrupolo) tramite la missmap che non ci siano missing values, ottenendo nuovamente 0% di missing values.



Il dataset è quindi pulito ed è così costituito

```
'data.frame':   4551 obs. of  13 variables:
 $ price         : num  313000 2384000 342000 420000 550000 ...
 $ bedrooms      : num  3 5 3 3 4 2 2 4 3 4 ...
 $ bathrooms     : num  1.5 2.5 2 2.25 2.5 1 2 2.5 2.5 2 ...
 $ sqft_living   : int  1340 3650 1930 2000 1940 880 1350 2710 2430 1520 ...
 $ sqft_lot      : int  7912 9050 11947 8030 10500 6380 2560 35868 88426 6200 ...
 $ waterfront    : int  0 0 0 0 0 0 0 0 0 0 ...
 $ condition     : int  3 5 4 4 4 3 3 3 4 3 ...
 $ sqft_above    : int  1340 3370 1930 1000 1140 880 1350 2710 1570 1520 ...
 $ sqft_basement: int  0 280 0 1000 800 0 0 0 860 0 ...
 $ yr_built      : int  1955 1921 1966 1963 1976 1938 1976 1989 1985 1945 ...
 $ yr_renovated : int  2005 0 0 0 1992 1994 0 0 0 2010 ...
 $ city          : Factor w/ 44 levels "Algona","Auburn",..: 37 36 19 4 32 36 32 22 28 36 ...
 $ statezip      : Factor w/ 77 levels "WA 98001","WA 98002",..: 63 59 27 8 32 55 32 24 28 55 ...
```

Essendo la maggior parte delle variabili di tipo numerico si potranno fare delle analisi sui valori delle statistiche descrittive quali media, varianza, deviazione standard, kurtosis, skewness.

La media è la media aritmetica dei ritorni.
La varianza è una misura che esprime di quanto oscilla il valore rispetto al valore che può assumere la variabile.
La deviazione standard è matematicamente la radice quadrata della varianza, quindi è una misura simile alla varianza, ma ha un intervallo più ristretto.
La skewness misura la simmetria di una distribuzione intorno alla sua media; il suo valore può appartenere a tre intervalli:

- sk= 0, quindi è simmetrica (distribuzione normale)
- sk> 0 coda di destra più lunga rispetto alla distribuzione normale
- sk< 0 coda di sinistra più lunga rispetto alla distribuzione normale

La kurtosis misura lo spessore della coda di una distribuzione; il suo valore può appartenere a tre intervalli:

- ks > 0 coda più spessa di una coda di una distribuzione normale
- ks < 0 coda meno spessa di una coda di una distribuzione normale
- ks = 0 coda uguale alla coda di una distribuzione normale

I quartili sono vettori numerici che definiscono dove si trova la metà della popolazione e quindi la metà dei valori della distribuzione; il primo quantile indica il minimo, il secondo indica il 25%, il terzo è il valore che rappresenta la mediana della distribuzione (50%), il quarto indica il 75%, mentre l'ultimo indica l'ultimo valore.

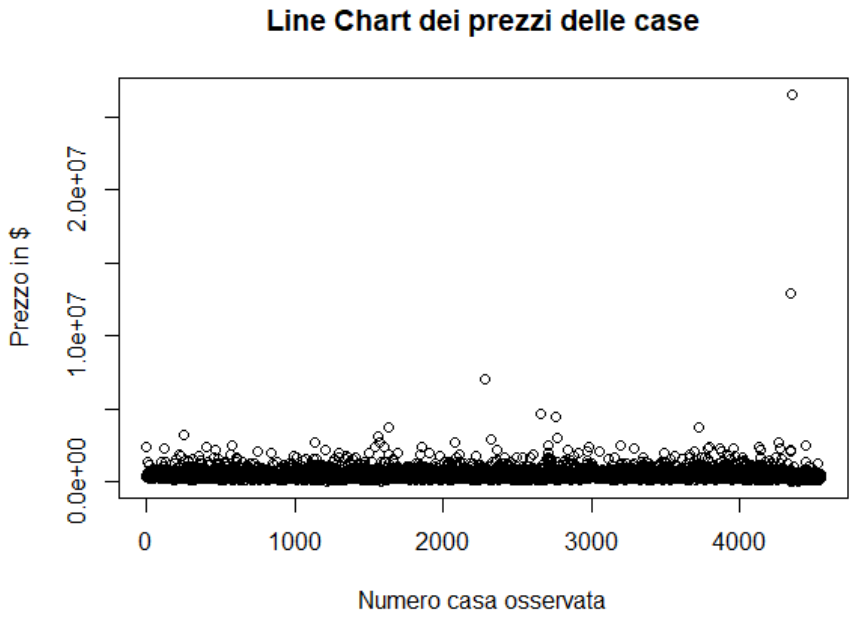
	price	bedrooms	bathrooms	sqft_living	sqft_lot	waterfront	condition	sqft_above	sqft_basement	yr_built	yr_renovated	city	statezip
	Min. : 7800	Min. :0.000	Min. :0.000	Min. : 370	Min. : 638	Min. :0.000000	Min. :1.000	Min. : 370	Min. : 0.0	Min. :1900	Min. : 0.0	Seattle :1561	WA 98103: 148
	1st Qu.: 326264	1st Qu.:3.000	1st Qu.:1.750	1st Qu.: 1460	1st Qu.: 5000	1st Qu.:0.000000	1st Qu.:3.000	1st Qu.:1190	1st Qu.: 0.0	1st Qu.:1951	1st Qu.: 0.0	Renton : 291	WA 98052: 135
	Median : 465000	Median :3.000	Median :2.250	Median : 1970	Median : 7680	Median :0.000000	Median :3.000	Median :1590	Median : 0.0	Median :1976	Median : 0.0	Bellevue: 281	WA 98117: 132
	Mean : 557906	Mean :3.395	Mean :2.155	Mean : 2132	Mean : 14835	Mean :0.006592	Mean :3.449	Mean :1822	Mean : 310.2	Mean :1971	Mean : 808.6	Redmond : 235	WA 98115: 129
	3rd Qu.: 657500	3rd Qu.:4.000	3rd Qu.:2.500	3rd Qu.: 2610	3rd Qu.: 10978	3rd Qu.:0.000000	3rd Qu.:4.000	3rd Qu.:2300	3rd Qu.: 600.0	3rd Qu.:1997	3rd Qu.:1999.0	Kirkland: 187	WA 98006: 109
	Max. :26590000	Max. :9.000	Max. :8.000	Max. :13540	Max. :1074218	Max. :1.000000	Max. :5.000	Max. :9410	Max. :4820.0	Max. :2014	Max. :2014.0	Issaquah: 186	WA 98059: 106
	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	(Other) :1810	(Other) :3792

Tramite la funzione kable(summary(case.df)) vengono calcolati i valori di minimo e massimo, media e mediana, primo e terzo quartile per le variabili numeriche, la frequenza dei vari fattori per le variabili di tipo factor.

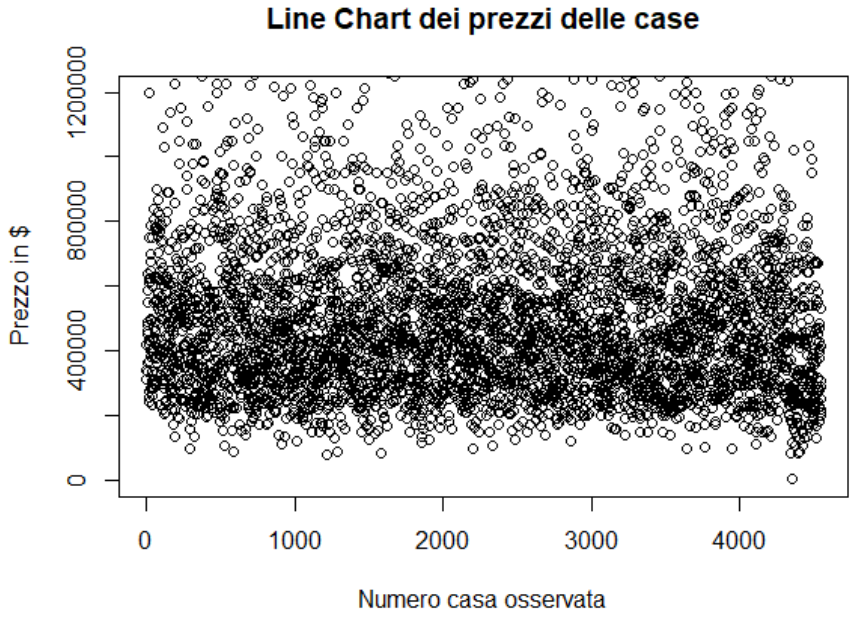
EXPLORATORY ANALYSIS

1. *PREZZO*

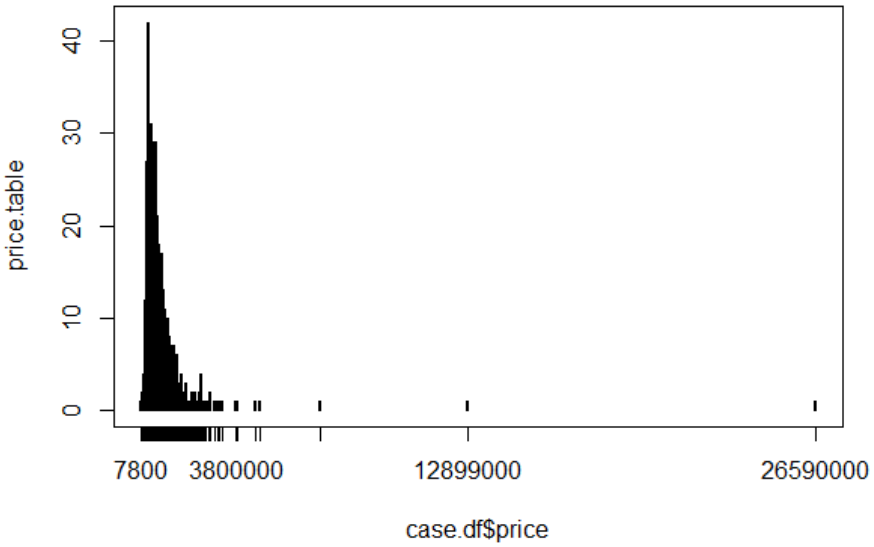
Come primo grafico per il prezzo viene realizzato il line chart che può essere utile per capire in quale fascia di prezzo siano più concentrati i prezzi.

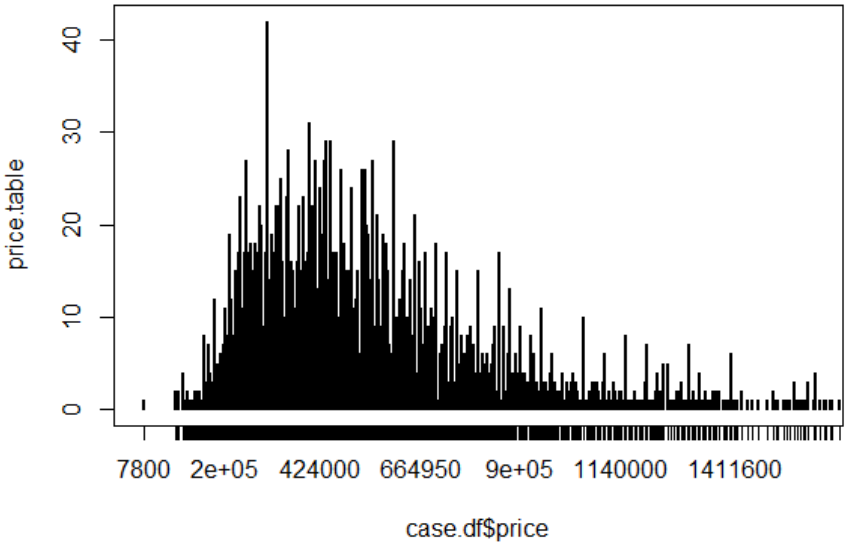


Si può notare che la concentrazione è massima a un livello di un ordine decimale inferiore a 1 miliardo, nonostante ci siano poche case con un prezzo a quell'ordine di grandezza; viene quindi deciso di fare uno zoom su un livello di prezzi attorno a 1 milione di $ e si può notare che la maggior parte delle case ha un prezzo inferiore anche a 1 milione di dollari, infatti, man mano che il prezzo sale, diminuisce di molto il numero di osservazioni per quel prezzo.



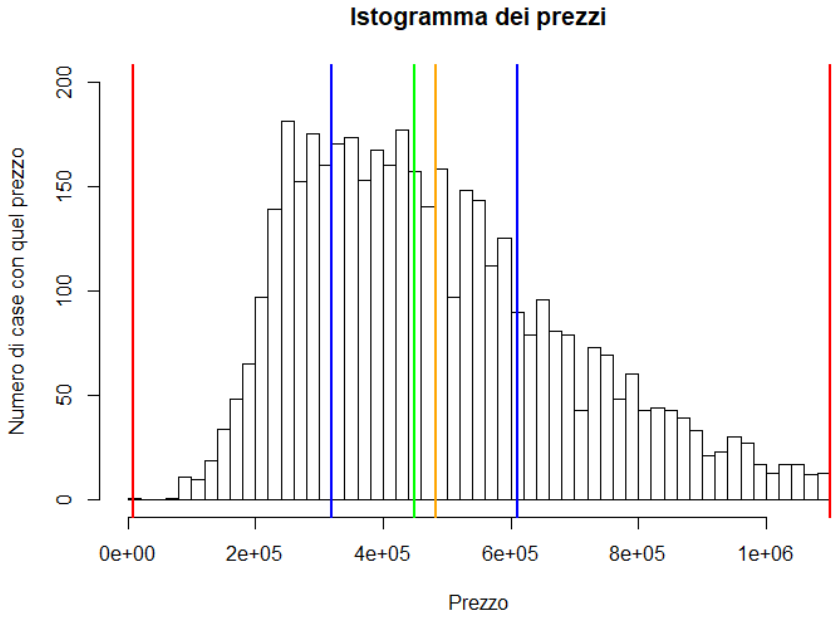
Queste osservazioni possono essere fatte anche plottando la price table creata selezionando la colonna price del dataset.

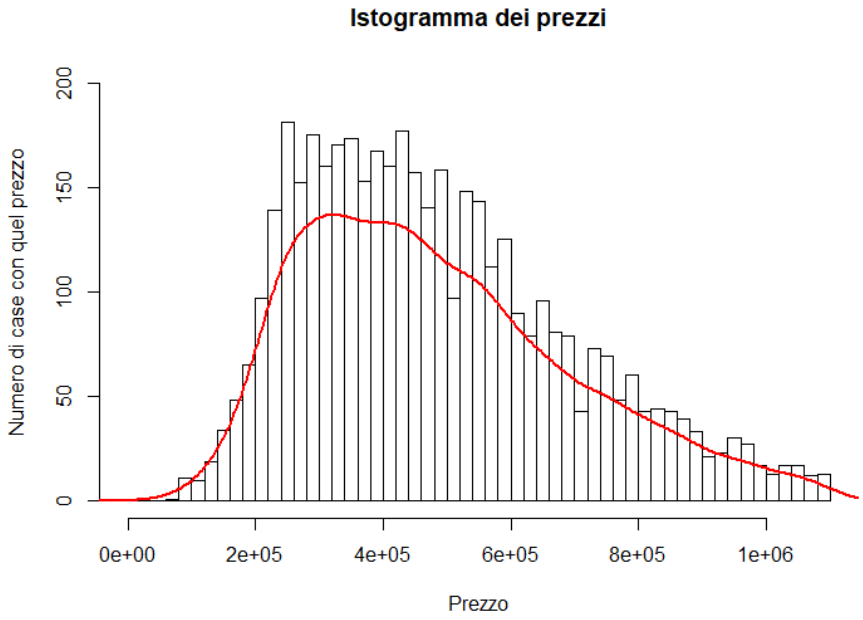




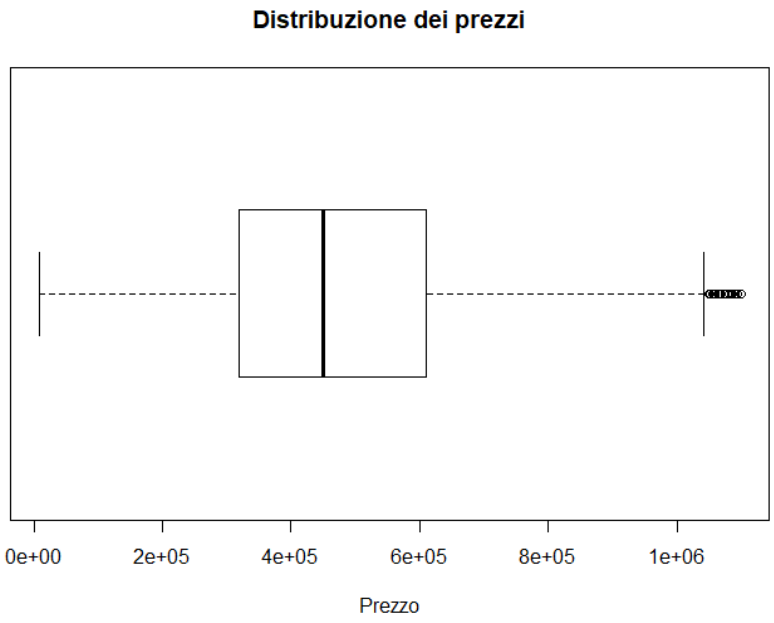
Si nota che la concentrazione maggiore dei prezzi delle case è nell'intervallo tra 200000$ e circa 1100000$, quindi viene deciso di eliminare le osservazioni con prezzo superiore a 1100000$ (questa operazione rende abbastanza veritiera anche l'analisi, in quanto potrebbe trattarsi di un'analisi fatta da un cittadino che vuole comprare una casa e ha un budget nella media). Tramite questa eliminazione vengono rimosse 268 case.

Viene quindi creato un istogramma con i prezzi delle 4283 case e vengono stampati anche tramite il comando abline i valori dei quartili (rosso, blu e verde) e il valore della media dei prezzi; si può notare che la distribuzione è maggiore nell'intervallo 200000-500000 $ circa e che la media dei prezzi si trova a un valore maggiore rispetto alla mediana, ma vicina a essa; in seguito viene plottata anche la curva della densità che è una curva che si discosta molto dalla curva di una distribuzione normale.



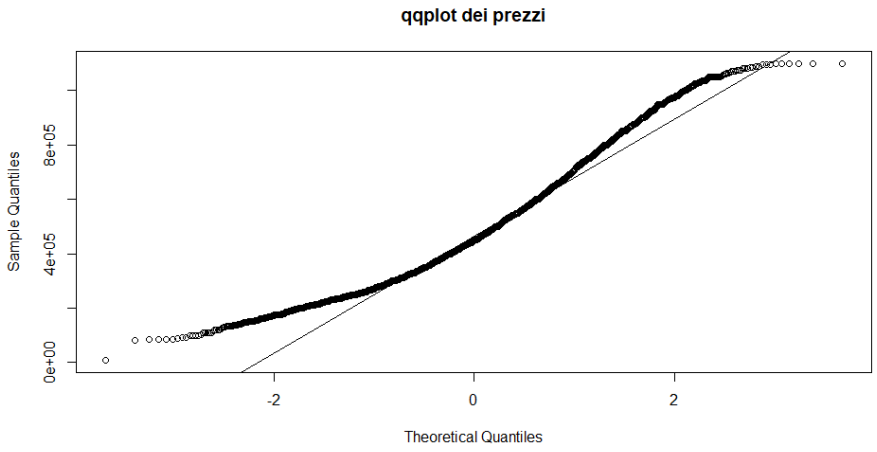


Infine, viene creato un boxplot che presenta numerosi outlier, ovvero un insieme di valori diverso da tutti gli altri per un valore di prezzi superiore a 1 mln di $.



Analizzando il valore di skewness si può concludere che, essendo un valore > 0, la coda di destra è più lunga della distribuzione normale; per la kurtosis (<0) invece si può dire che la coda è più sottile della coda di una distribuzione normale; infine essendo entrambi i valori diversi da 0 si può dire che la distribuzione non è normale.

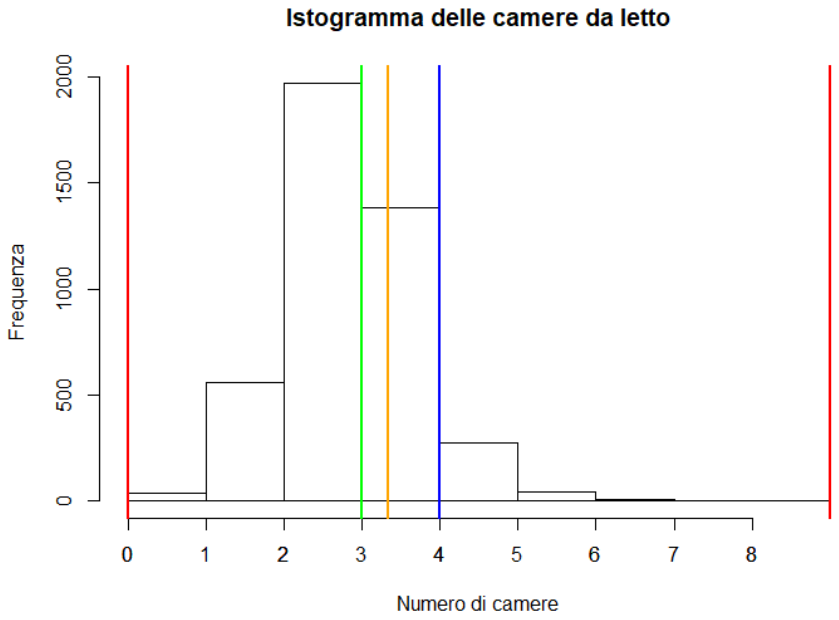
price_q	Named num [1:5] 7800 319975 449950 610000 1100000
price.ks	-0.134573287186344
price.mean	483232.186343714
price.sd	209912.457973992
price.sk	0.663343308395872
price.var	44063240012.6828



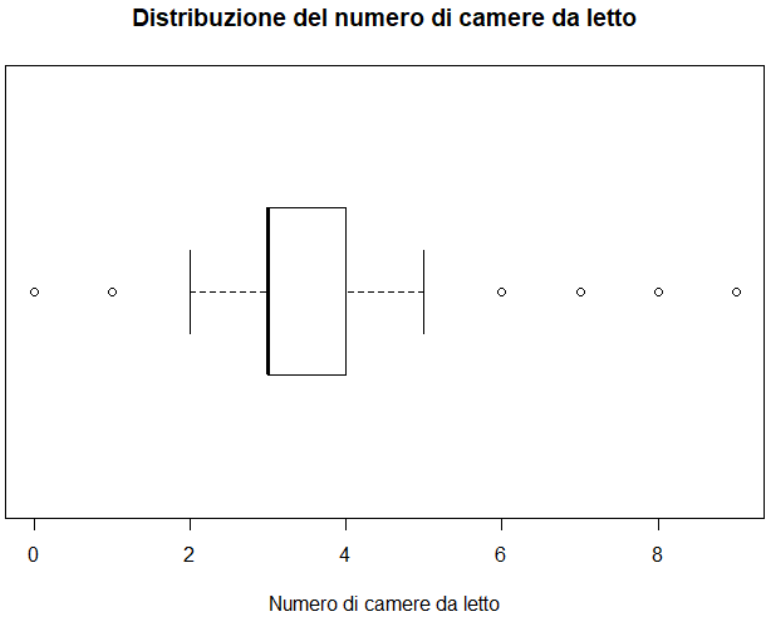
Anche il qqplot evidenzia il fatto che la distribuzione non sia normale dal momento che i punti stanno in prossimità della qqline e non su di essa.

2. NUMERO DI CAMERE DA LETTO

Analizzando il numero di camere da letto si può notare come il numero medio sia poco più di tre; in questo istogramma il primo e il secondo quartile coincidono, denotando il fatto che il quartile che rappresenta il 25% della distribuzione sia lo stesso del quartile che rappresenta la mediana (che ha un valore pari a 3)

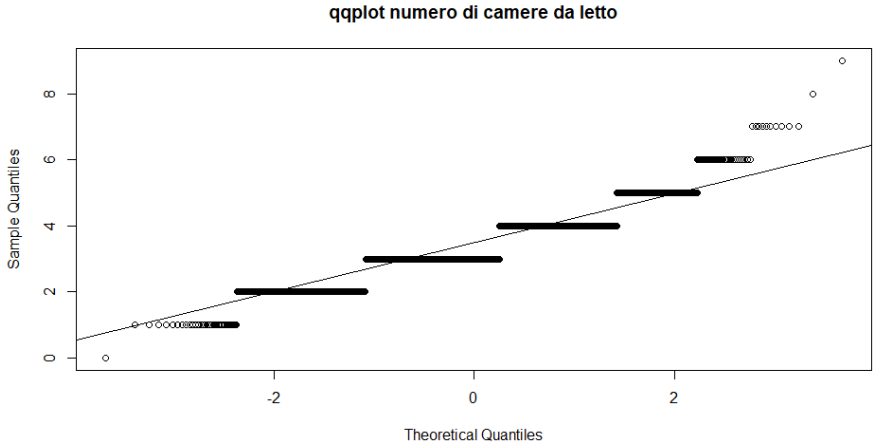


Questo comportamento si ripresenta anche all'interno del boxplot dove la linea della mediana coincide con il minimo del box; in questo boxplot si nota che la distribuzione è maggiormente concentrata nell'intervallo tra 3 e 4 e presenta outlier per valori pari a 0, 1, 6, 7, 8 e 9.



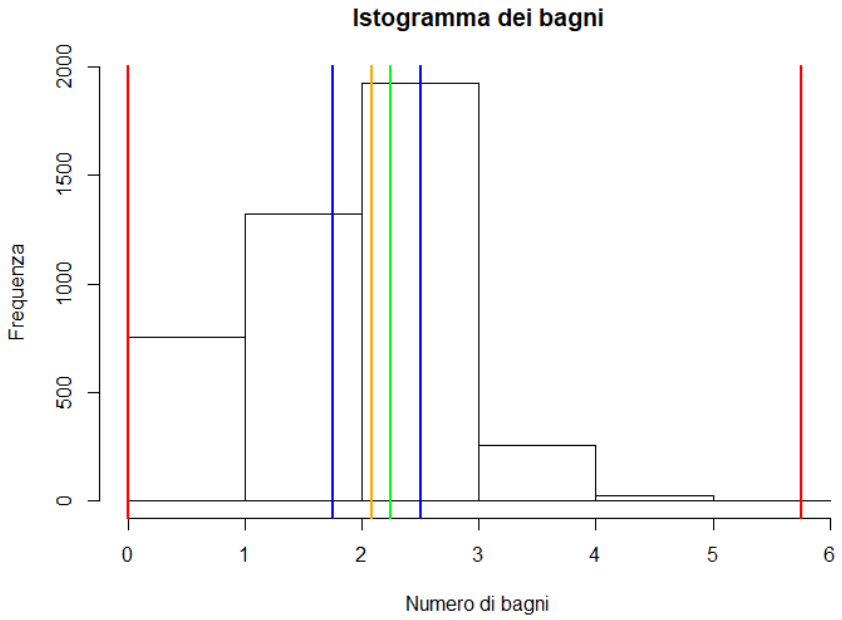
Infine, analizzando i valori delle statistiche descrittive, si nota che sia kurtosis che skewness hanno un valore maggiore di 0 che indicano rispettivamente che la coda è più spessa della coda della distribuzione normale e che la coda di destra è più lunga della distribuzione normale (infatti comprende tutti i valori da 4 a 9); essendo anche questi valori diversi da 0 si può dire che la distribuzione non è normale (lo si può capire anche dal qqplot dove la maggior parte dei punti non si trova sulla qqline).

bed_q	Named num [1:5] 0 3 3 4 9
bed.ks	1.26724689045678
bed.mean	3.34578566425403
bed.sd	0.877463361236746
bed.sk	0.434170842503727
bed.var	0.769941950312888

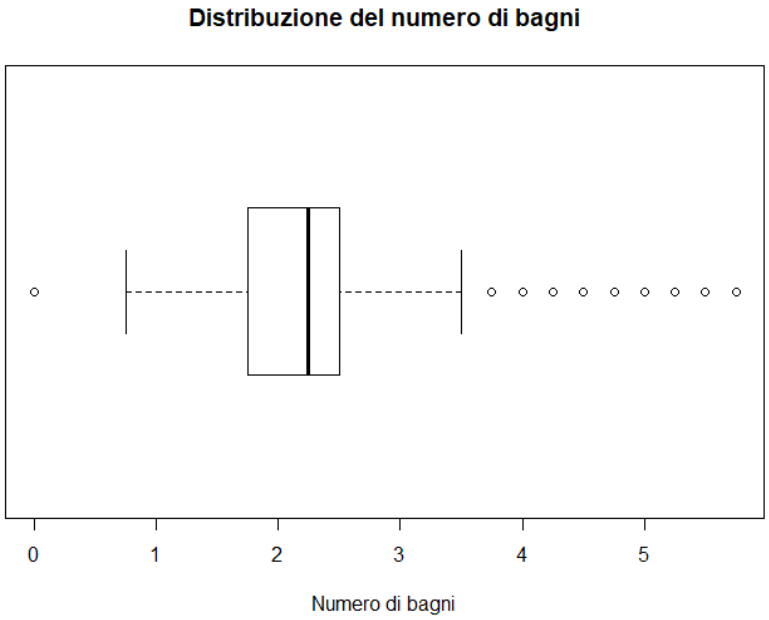


3. *NUMERO DI BAGNI*

Viene creato anche l'istogramma per il numero dei bagni che fa notare una maggiore distribuzione per un numero di bagni inferiore a 4, una media di poco maggiore di 2 e una mediana questa volta con un valore maggiore rispetto alla media.



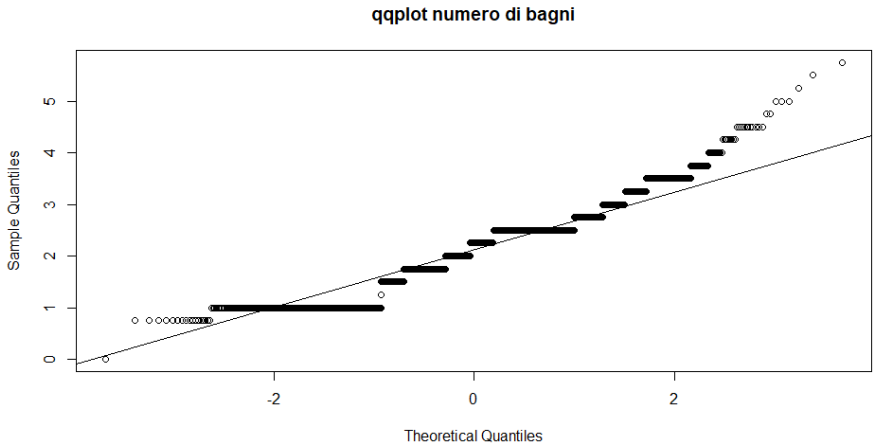
Nel boxplot si possono notare numerosi outlier per valori che sono lontani dai valori più frequenti e quindi con una distribuzione maggiore; in particolare si possono notare outlier per un numero di bagni maggiore di 4 o pari a 0.



I valori di skewness e kurtosis sono prossimi a zero (0.2) entrambi positivi, quindi valgono le considerazioni fatte per il numero di camere da letto; in questo caso però la distribuzione è più simile a quella normale rispetto a quella precedente in quanto i valori di sk e ks sono più vicini allo 0.

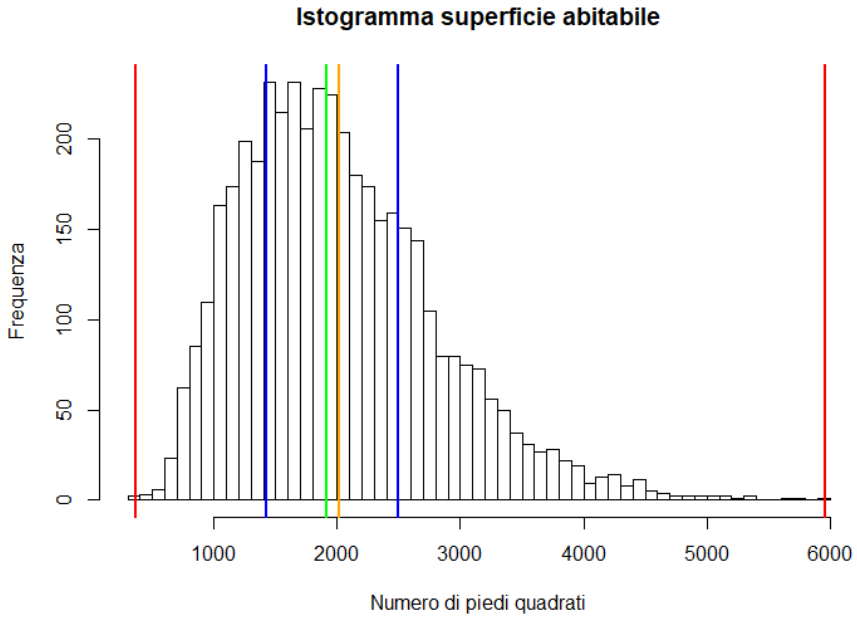
bath_q	Named num [1:5] 0 1.75 2.25 2.5 5.75
bath.ks	0.288702432088889
bath.mean	2.09012374503852
bath.sd	0.712529199056739
bath.sk	0.216288446583841
bath.var	0.507697859508438

Il qqplot evidenzia il fatto che la distribuzione del numero di bagni sia comunque abbastanza lontana dalla distribuzione normale.

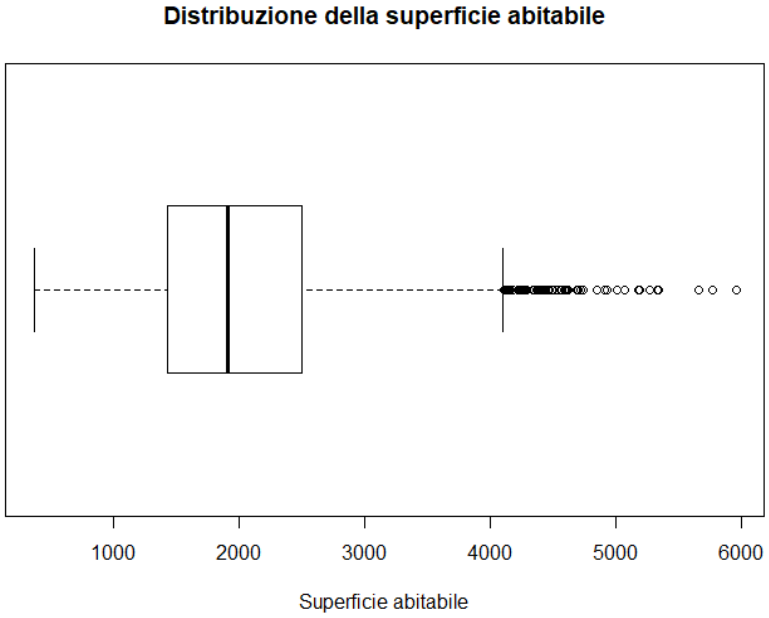


4. ESTENSIONE SUPERFICIE ABITABILE (sqft_living)

In questo istogramma si può notare come la dimensione della superficie abitabile abbia un valore di densità molto alto nell'intervallo tra circa 1400 (valore del secondo quartile, la prima retta blu) e poco più di 2000 (circa dove si posiziona la media, la retta arancione); inoltre si può vedere che la coda di destra è molto più lunga di quella di sinistra.



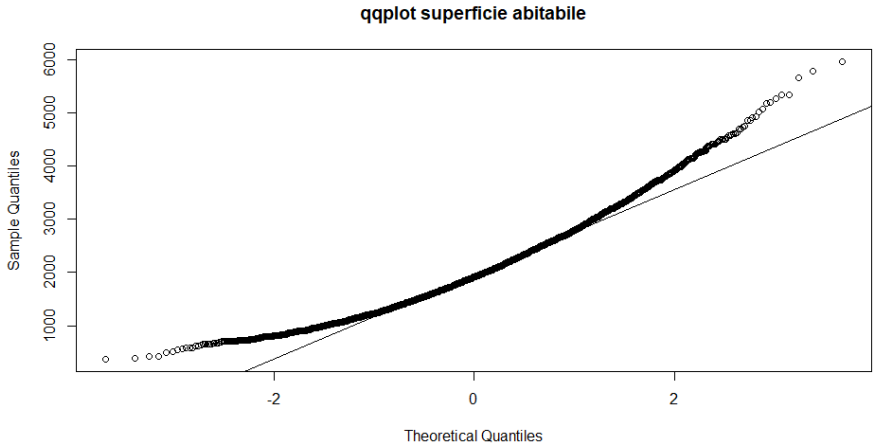
Dal boxplot si può notare che sono presenti numerosi outlier per valori maggiori o uguali a 4000, valori che però sono presenti con una bassa frequenza.



Analizzando i valori di skewness e kurtosis sono entrambi positivi e hanno un valore di circa 0.8; da queste informazioni si può capire che la coda di destra è più lunga rispetto alla coda di una distribuzione normale e inoltre che la coda è più spessa.

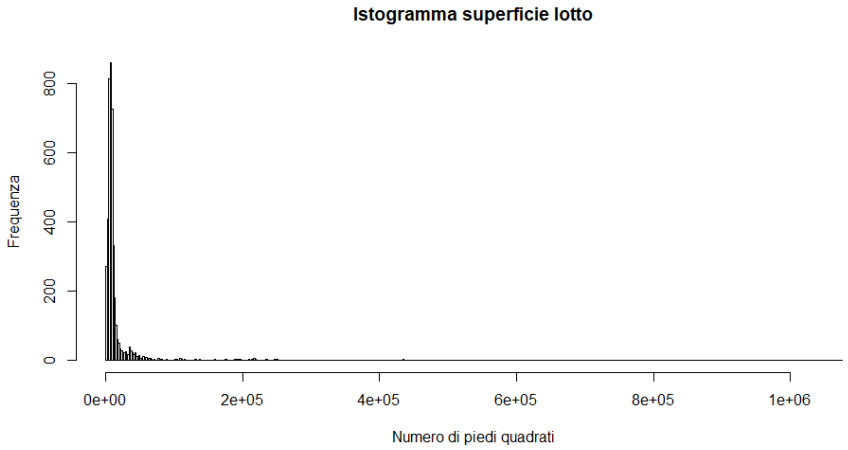
liv_q	Named num [1:5] 370 1430 1910 2500 5960
liv.ks	0.793794724627535
liv.mean	2019.92271772122
liv.sd	794.286894671024
liv.sk	0.810001492777893
liv.var	630891.671046139

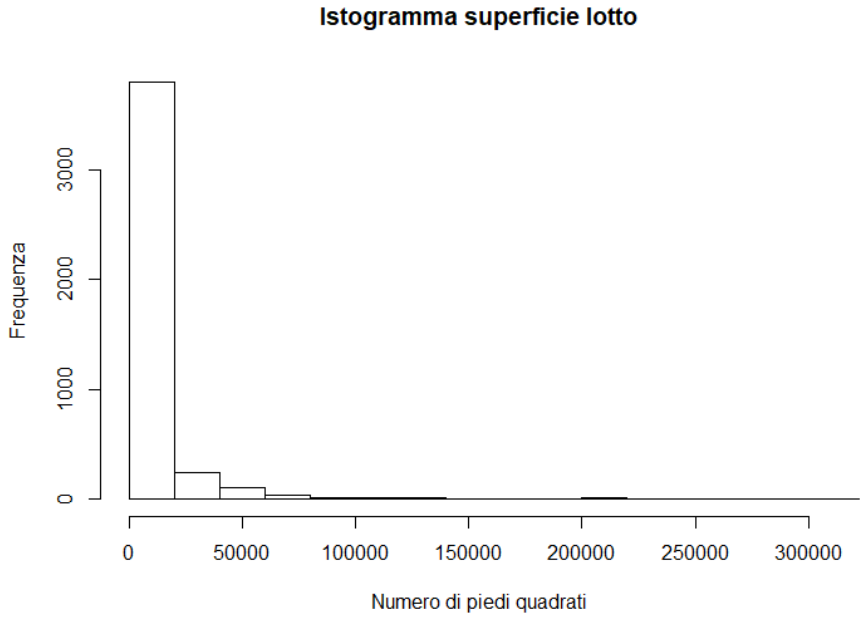
Infine, si può notare che essendo i valori di sk e ks diversi da 0 e non essendo i punti nel qqplot sulla qqline si può concludere che la distribuzione non sia una distribuzione normale.



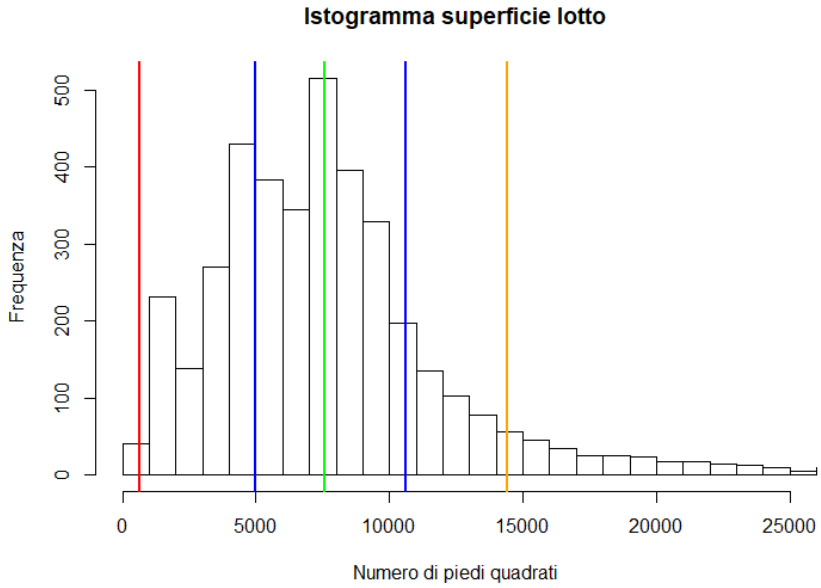
5. *ESTENSIONE SUPERFICIE DEL LOTTO (sqft_lot)*

Da questo istogramma si può notare come la maggiore concentrazione si ha per un valore compreso tra 0 e circa 25000 anche se ci sono valori molto meno frequenti a livelli successivi che rappresentano estensioni di gran lunga maggiori.





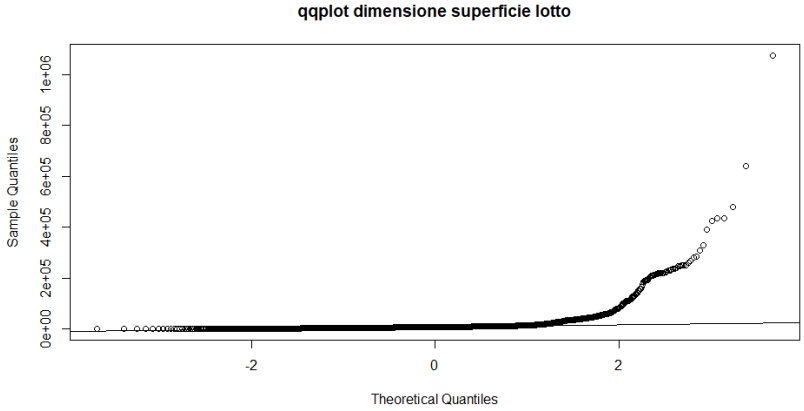
Si è deciso dunque di fare uno zoom imponendo un xlim tra 0 e 25000.
Si può notare che il valore medio è poco meno di 15000 ed è dovuto al fatto che ci siano dei valori poco frequenti che però alzano la media; inoltre, l'ultimo quartile non è visibile in quanto si trova a un valore molto elevato (1074218 piedi quadrati).



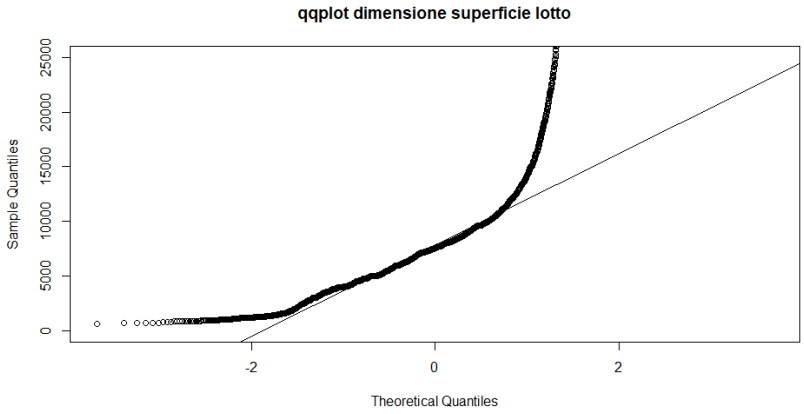
I valori di skewness e kurtosis sono decisamente alti, maggiori di 0 e si può concludere che la distribuzione sia molto lontana da una distribuzione normale.

lot_q	Named num [1:5] 638 5000 7560 10630 1074218
lot.ks	238.113834740342
lot.mean	14396.3509222508
lot.sd	35733.9989163781
lot.sk	11.8993935974527
lot.var	1276918678.55571

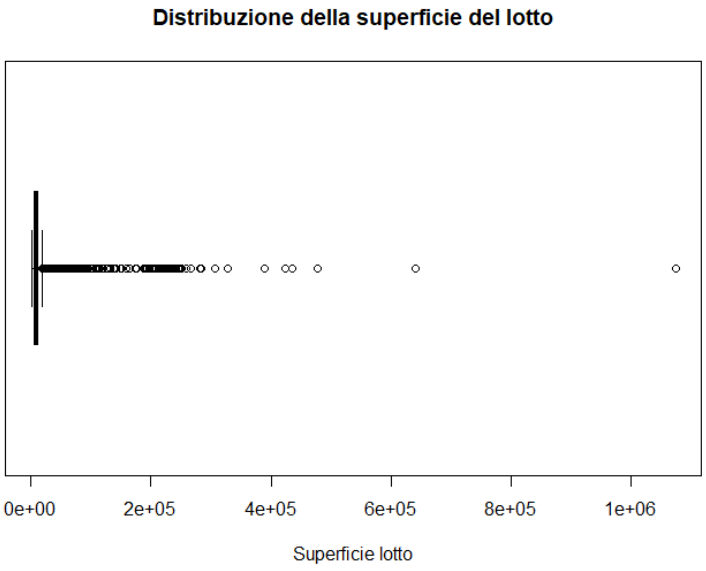
Lo si può notare anche dal qqplot dove ci sono i punti che rappresentano le case con un numero molto elevato di piedi quadrati sono molto distanti dalla qqline.

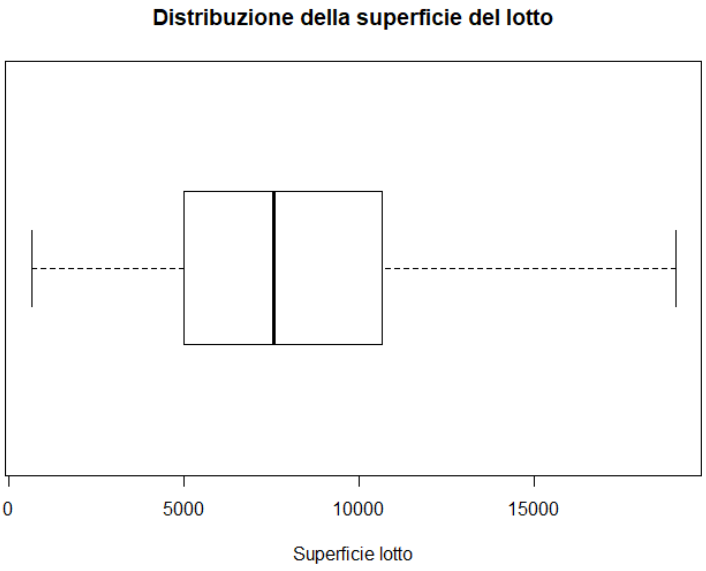


Per dei valori compresi tra 5000 e 10000 circa invece si può dire che la distribuzione è pressochè normale data l'ampia densità dei punti sulla qqline.



Infine, il boxplot si presenta inizialmente così, molto molto schiacciato a causa della presenza di numerosissimi outliers.

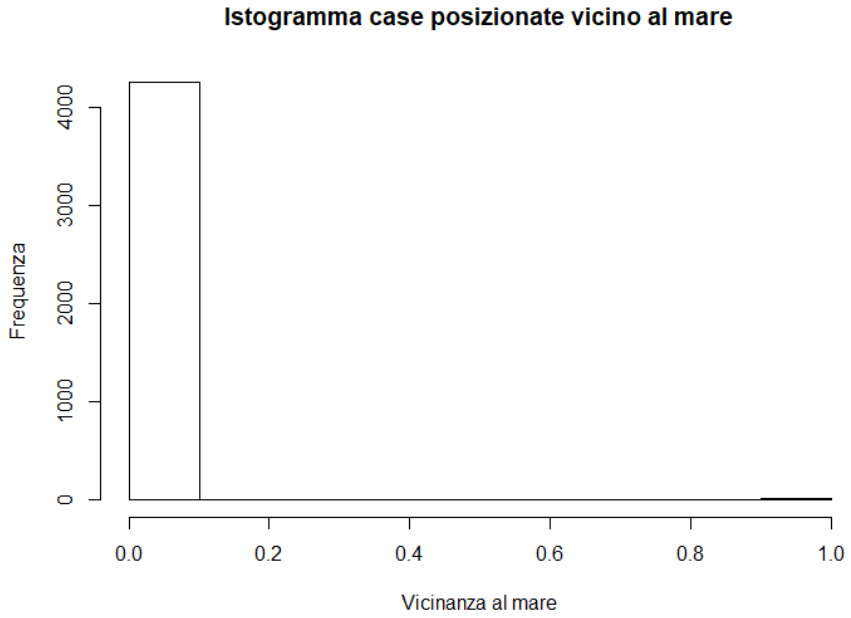


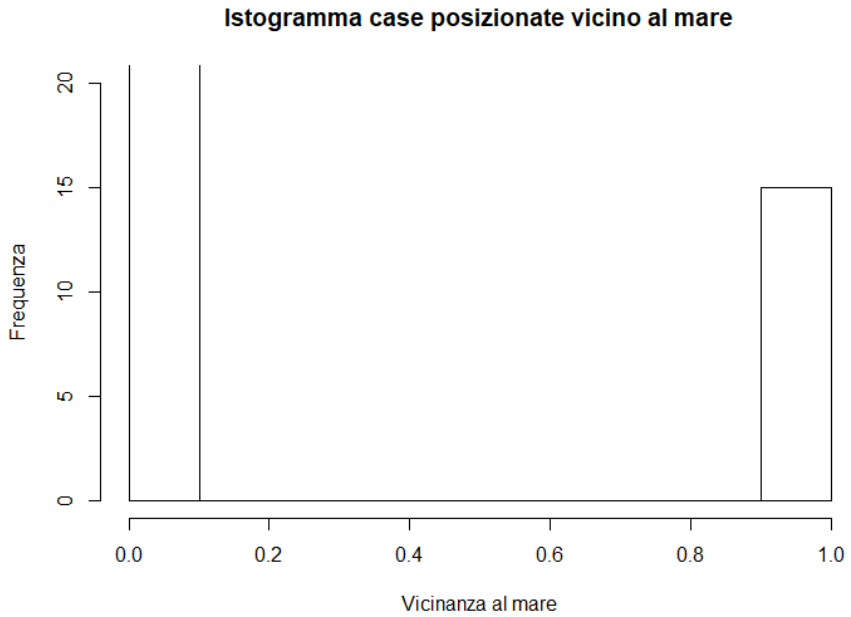


Se gli outliers vengono nascosti (outline = FALSE) il boxplot si presenterà in questa maniera con la mediana praticamente al centro del box e con la coda destra decisamente più allungata di quella sinistra.

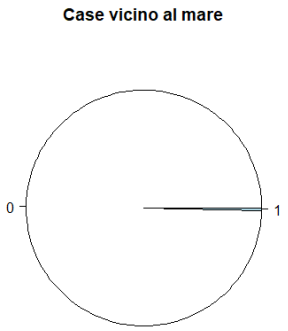
6. *VICINANZA AL MARE (waterfront)*

Questo istogramma contiene solo valori 0 o 1; in particolare si può notare la grandissima numerosità dei valori pari a zero (più di 4000 dati) e la numerosità molto bassa (15 case) delle righe che hanno il valore 1 nella colonna riguardante questa caratteristica.



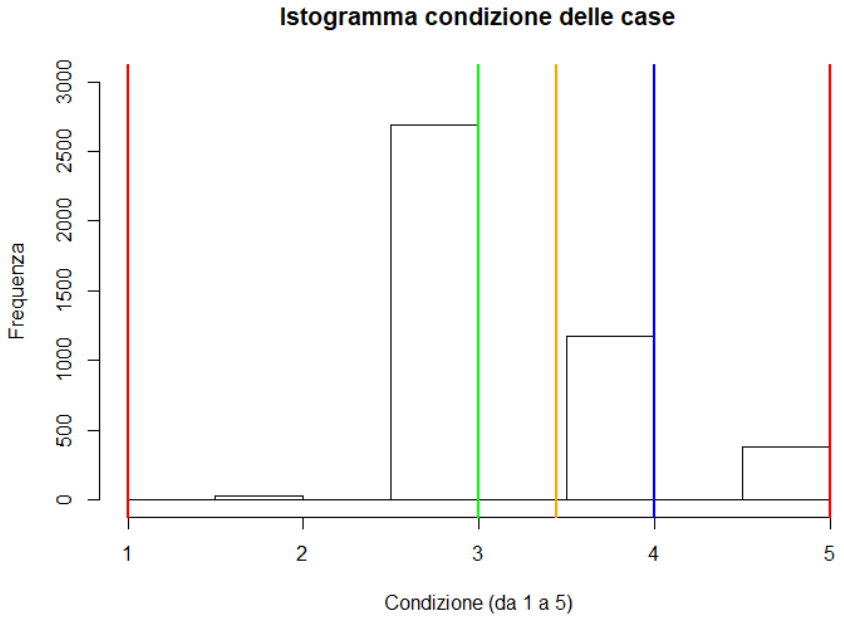


Viene anche creato un grafico a torta per mostrare che praticamente l'intero dataset ha un valore pari a 0 nella colonna chiamata waterfront e che quindi la maggior parte delle case in vendita in analisi si trovi in una città o in una posizione lontana dal mare.

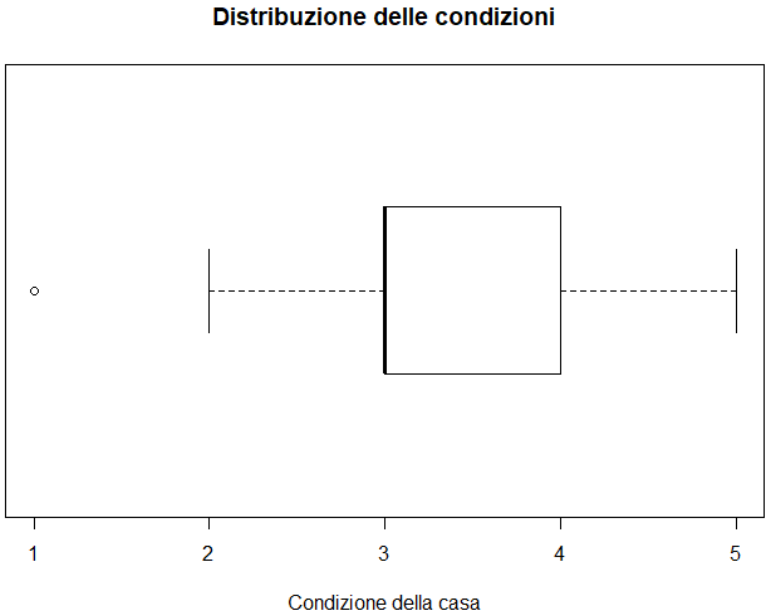


7. CONDIZIONE DELLA CASA

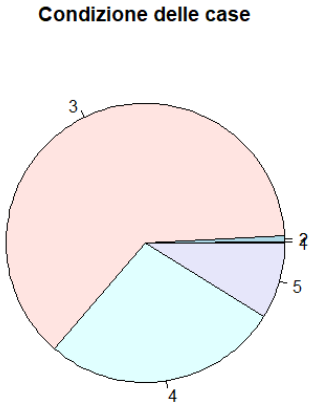
La condizione della casa viene espressa tramite una valutazione che va da 1 a 5; da questo istogramma si capisce come la maggior parte delle case abbia una condizione buona (valutazione pari a 3) o una condizione ancora migliore (valutazioni pari a 4 e 5). Molto basso è il numero di case con una valutazione scarsa sotto al 3 (quindi 2 o 1).
In questo istogramma il secondo e il terzo quartile (quindi quelli che corrispondono al 25% e al 50% della distribuzione) coincidono, quindi si può dire che la distribuzione è molto alta per il valore pari a 3.
La media inoltre ha un valori pari a 3.44 indicando il fatto che ci siano poche case con condizione inferiore a 3.



Dal boxplot si può notare la presenza di un outlier per un valore pari a 1, il che sta ad indicare che, nonostante siano in numero basso, esistono delle case in condizioni pessime.



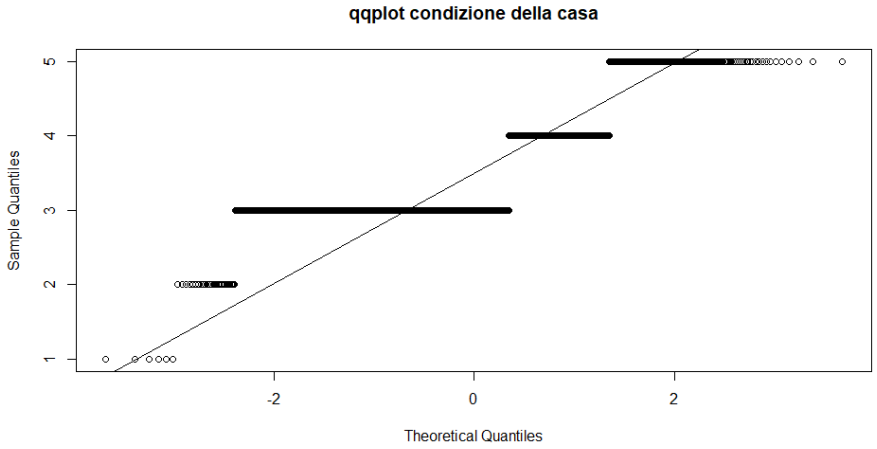
Viene creato un grafico a torta per rappresentare e visualizzare la distribuzione delle valutazioni.



Successivamente vengono analizzati i valori di skewness e kurtosis; entrambi questi valori positivi, quindi fanno capire rispettivamente che la coda di destra è più lunga rispetto alla distribuzione normale e che la coda è più spessa.

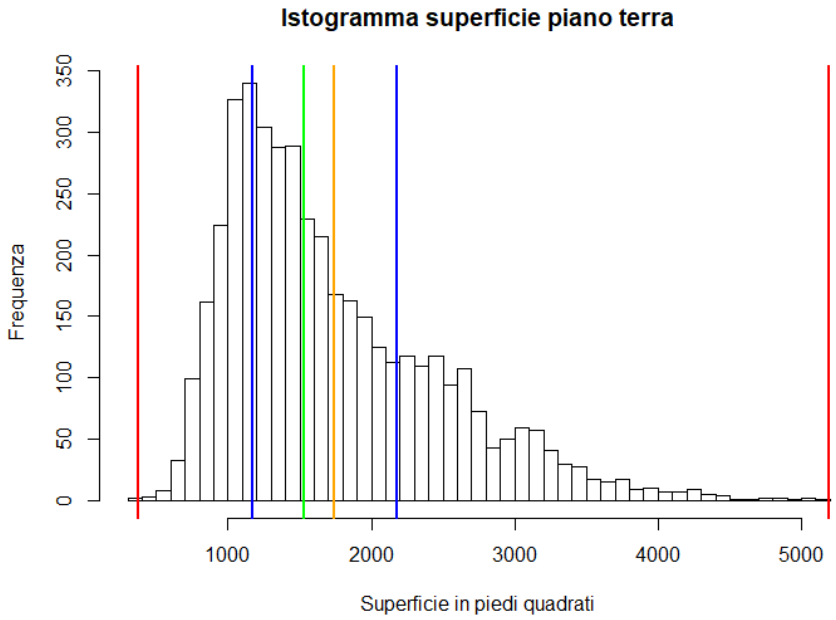
cond_q	Named num [1:5] 1 3 3 4 5
cond.ks	0.276251554187278
cond.mean	3.44221340182115
cond.sd	0.668300006634388
cond.sk	0.962872551038731
cond.var	0.446624898867523

Inoltre, analizzando il qqplot, si può notare che la distribuzione non è normale, ma multimodale in quanto sono pochi i punti sulla qqline; questo si può anche capire dal fatto che sk e ks hanno un valore diverso da 0.

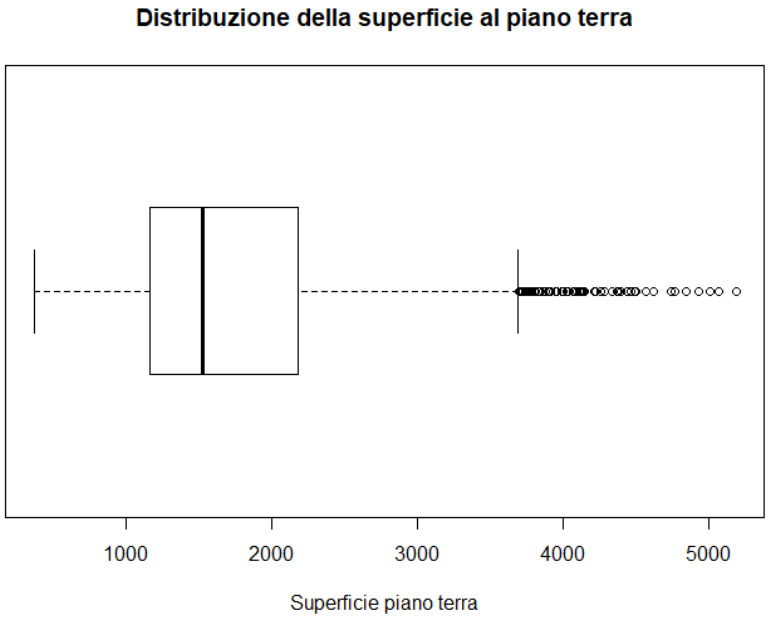


8. ESTENSIONE SUPERFICIE PIANO TERRA (sqft_above)

Da questo istogramma si può notare che la frequenza massima si trova per un valore pari circa a 1100 e la densità è massima nell'intervallo che va da 1000 a 1500 circa; si può inoltre dire che la media ha un valore pari a 1738 circa ed è compresa tra la mediana e il quarto quartile.



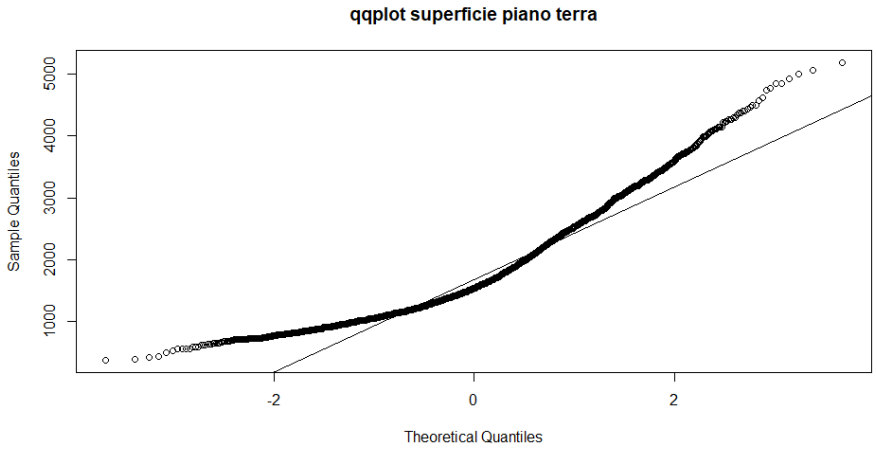
Il boxplot presenta numerosi outlier per valori maggiori di 3000 (3700 circa) e ha la coda di destra più allungata rispetto a quella di sinistra; la mediana invece non si trova perfettamente al centro del box rinforzando quindi l'affermazione precedente.



I valori di skewness e kurtosis sono entrambi maggiori di zero (e quindi diversi da zero); ciò porta a una conclusione simile a quelle fatte precedentemente ad esempio per la condizione o l'estensione della superficie del lotto.

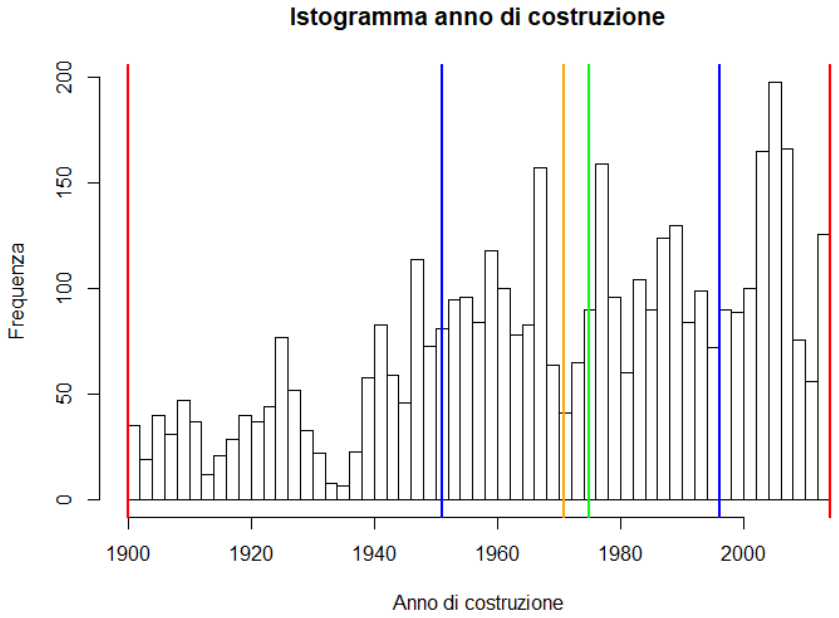
ab_q	Named num [1:5] 370 1170 1530 2180 5190
ab.ks	0.937580818623803
ab.mean	1738.91991594677
ab.sd	752.002143359806
ab.sk	1.07019991320628
ab.var	565507.223617742

Analizzando il qqplot, infine, si può dire che la distribuzione non sia simile a una distribuzione normale in quanto la maggior parte dei punti non si trova sulla qqline (o in prossimità di essa).

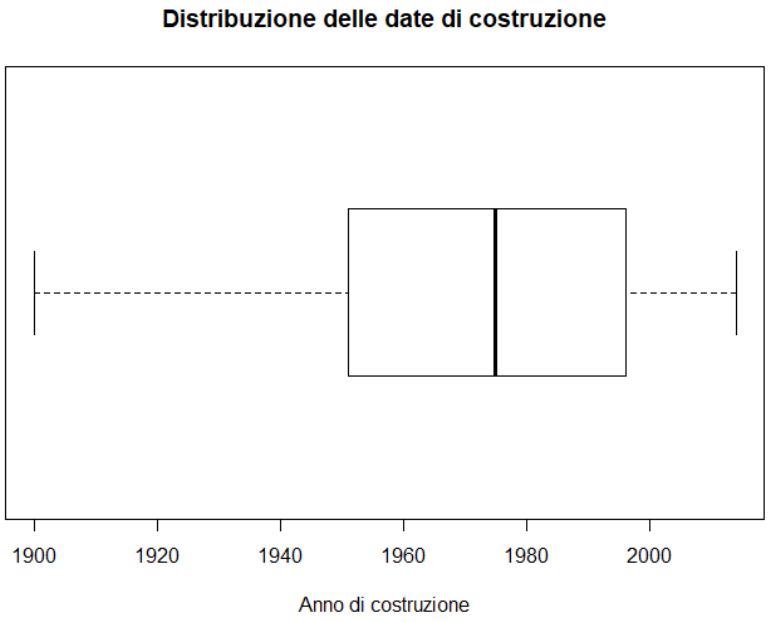


9. ANNO DI COSTRUZIONE

Questo istogramma descrive gli anni in cui sono state costruite le case in analisi, anni che vanno dal 1900 al 2014; si può notare che la distribuzione è pressoché simile nei vari anni, e si può anche notare che c'è un picco di costruzioni nei primi anni 2000.



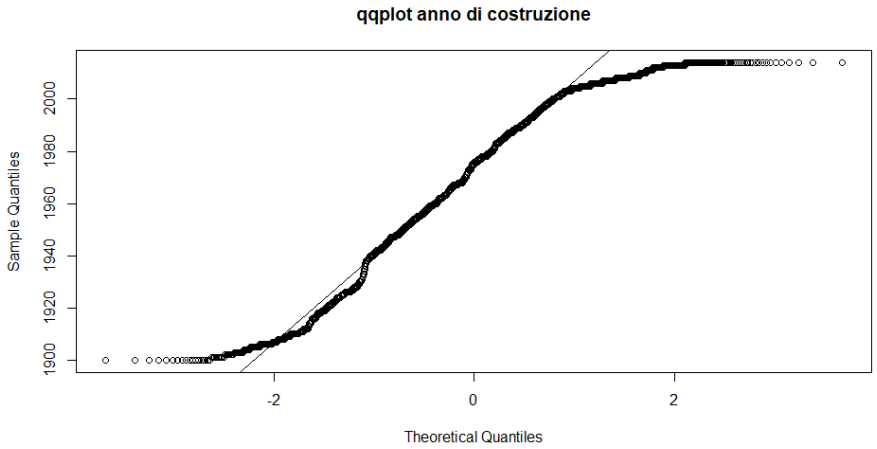
Il boxplot non identifica outliers e in esso il box si estende da poco prima di 1960 per arrivare fino a poco prima di 2000; la mediana si trova praticamente nel mezzo del box.



In questo caso i valori di skewness e kurtosis sono entrambi negativi, quindi rispettivamente indicano il fatto che la coda di sinistra è più lunga rispetto alla distribuzione normale e che la coda è più sottile della coda di una distribuzione normale.

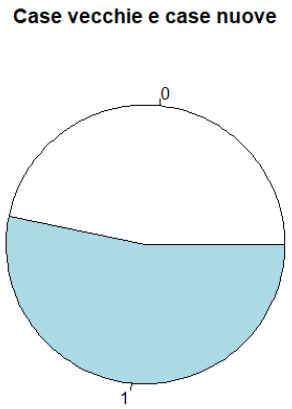
built_q	Named num [1:5] 1900 1951 1975 1996 2014
built.ks	-0.650985191192739
built.mean	1970.84730329208
built.sd	29.5026742047153
built.sk	-0.496486954588979
built.var	870.407785229571

I valori di sk e ks sono vicini allo 0 e dai qqplot si può notare che tanti punti stanno sulla qqline indicando che la distribuzione è abbastanza simile a una distribuzione normale.



Siccome la media degli anni di costruzione è 1971 circa, sono state considerate vecchie (0) le case costruite prima del 1971, nuove (1) quelle costruite dopo.

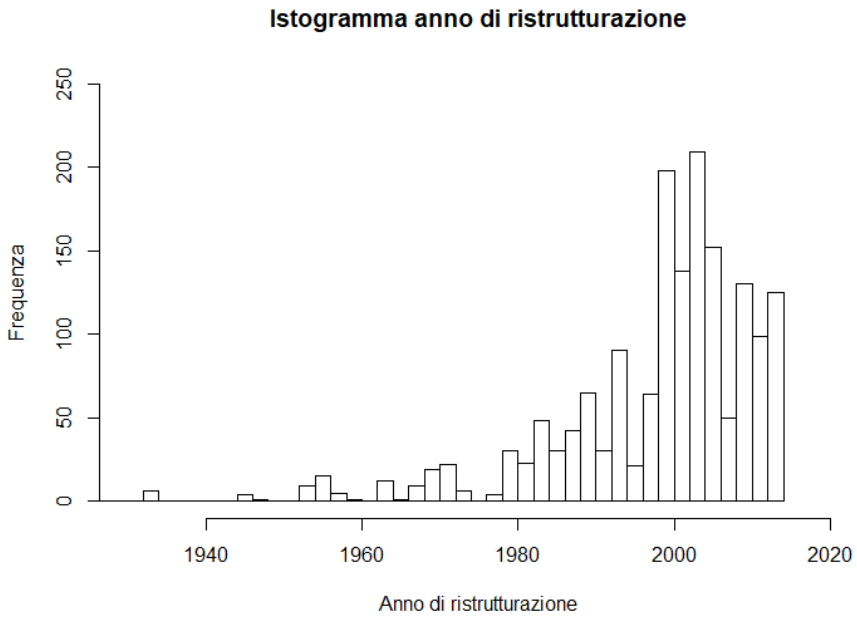
Si può notare che la maggior parte delle case sono case nuove, infatti nel grafico a torta più della metà corrisponde al valore 1.



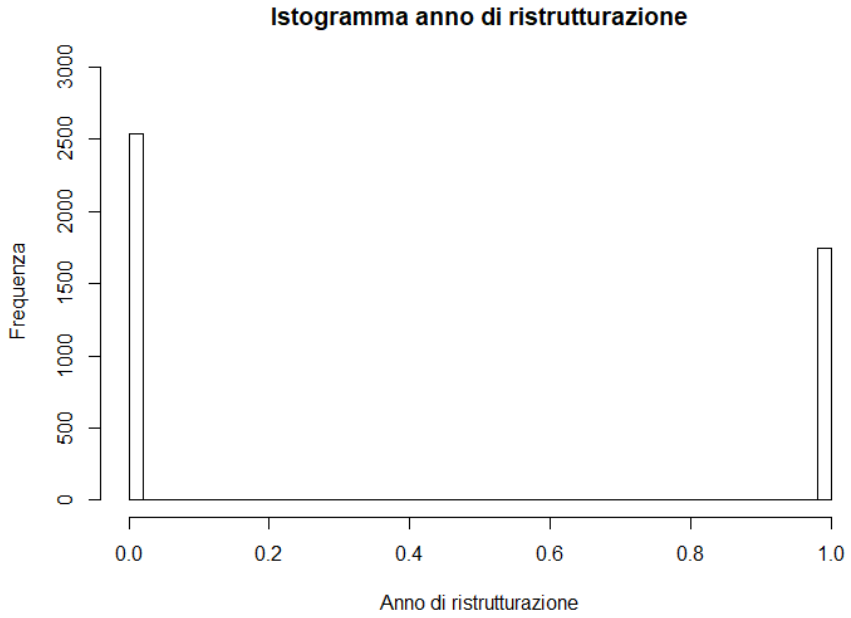
10. ANNO DI RISTRUTTURAZIONE

Questo istogramma mostra un grande numero di case non ristrutturate (hanno valore pari a 0), quindi si fa uno zoom nella parte dopo il 1900 per vedere quale sia l'intervallo temporale in cui sono concentrate maggiormente le operazioni di ristrutturazione.

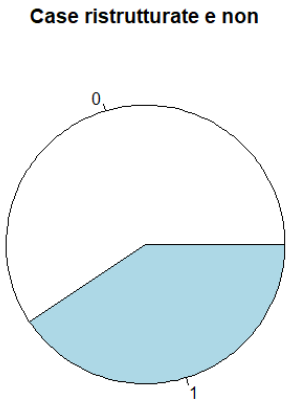
Si può capire che la maggior parte delle ristrutturazioni è avvenuta a cavallo degli anni 2000 e negli anni successivi.



Analizzare le statistiche descrittive avrebbe poco senso perché molti valori di questa colonna sono 0; si può notare che molte case non sono state ristrutturate, quindi vengono divise le case non ristrutturate (0) dalle case ristrutturate (1)

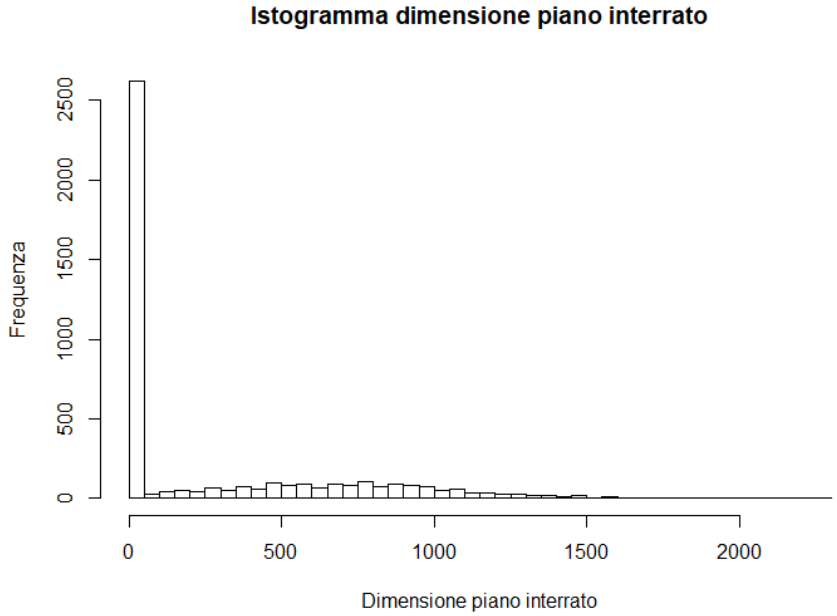


Infine, viene creato un grafico a torta che evidenzia che la maggior parte delle case non è stata ristrutturata, infatti la componente con valore 0 è maggiore di quella con valore 1.

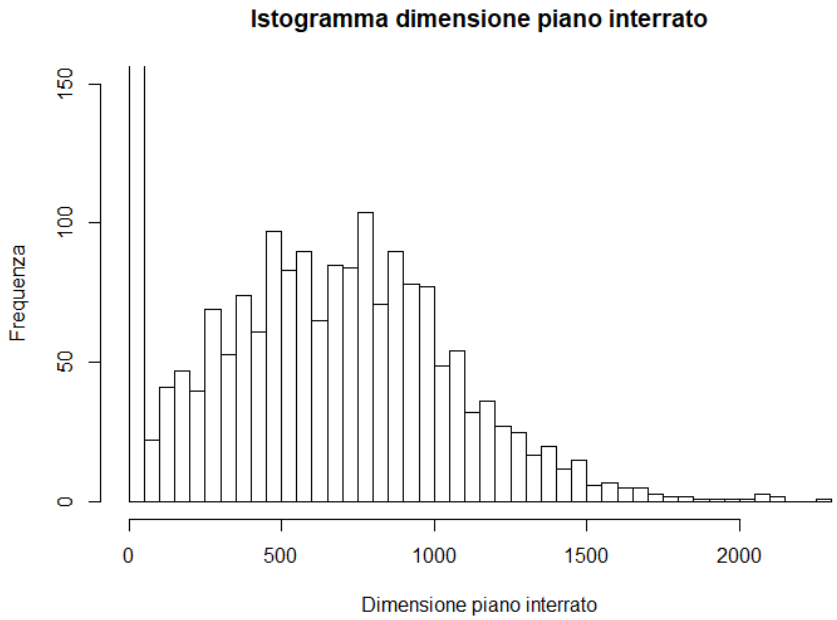


11. ESTENSIONE SUPERFICIE PIANO INTERRATO (sqft_basement)

Anche in questo istogramma si può notare che ci sono molti valori che corrispondono a zero, ma ci sono anche diverse occorrenze per valori fino a circa 2000

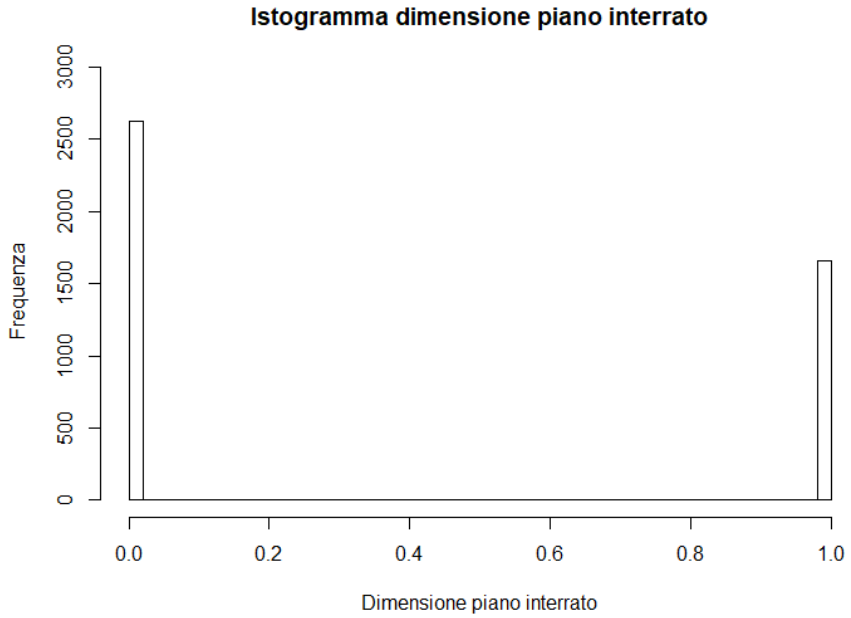


Eseguendo uno zoom si può comprendere meglio che la distribuzione è pressochè costante per i valori nell'intervallo tra 500 e 1000.

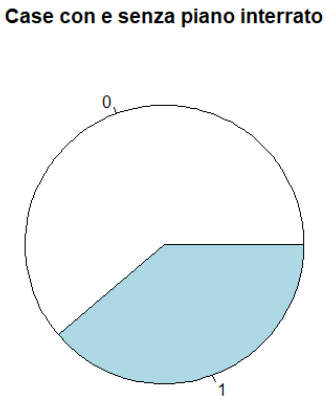


Si può però notare che molte case non hanno il piano interrato, quindi vengono divise le case senza piano interrato (0) dalle case con il piano interrato (1).

```
for (i in 1:4283){
  if(case.df$sqft_basement[i] > 0){
    case.df$sqft_basement[i]= 1
  }
}
```

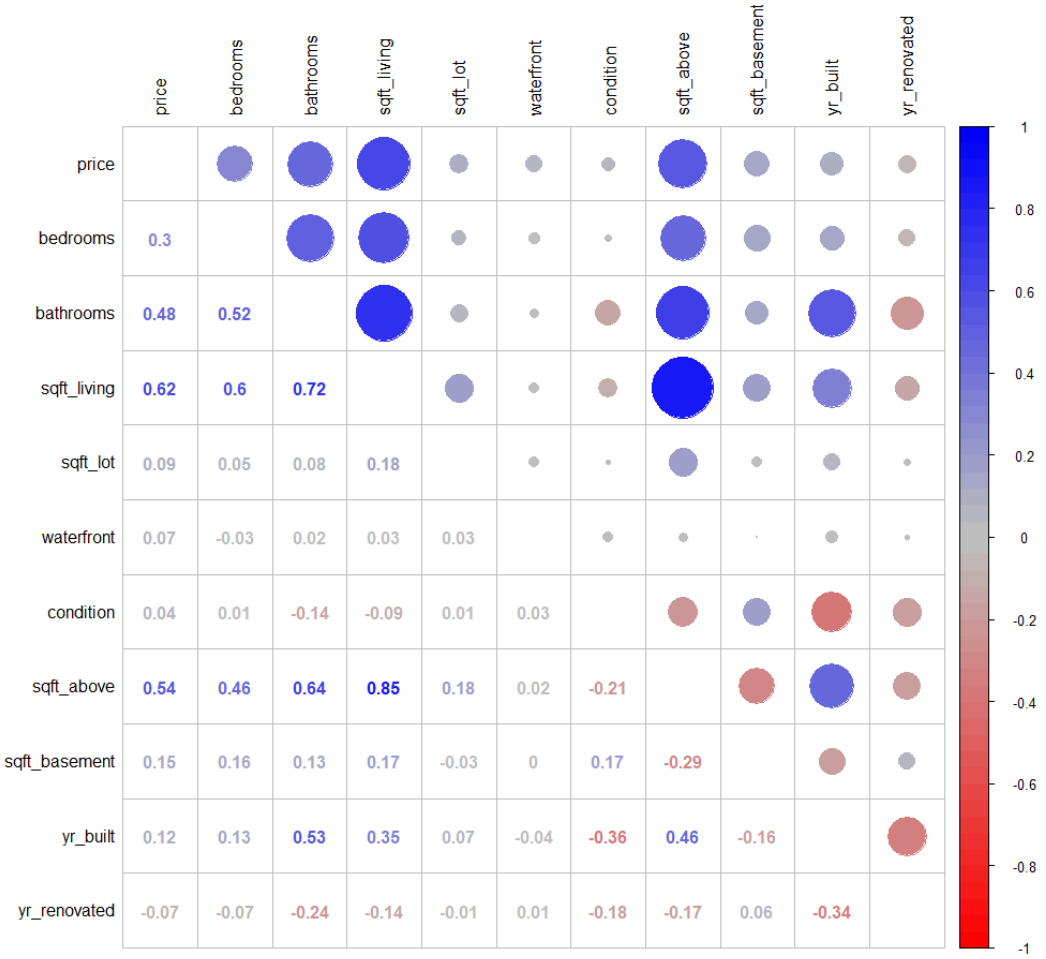


Creando il grafico a torta si nota che più della metà delle case non possiede il piano interrato (la variabile ha valore 0).



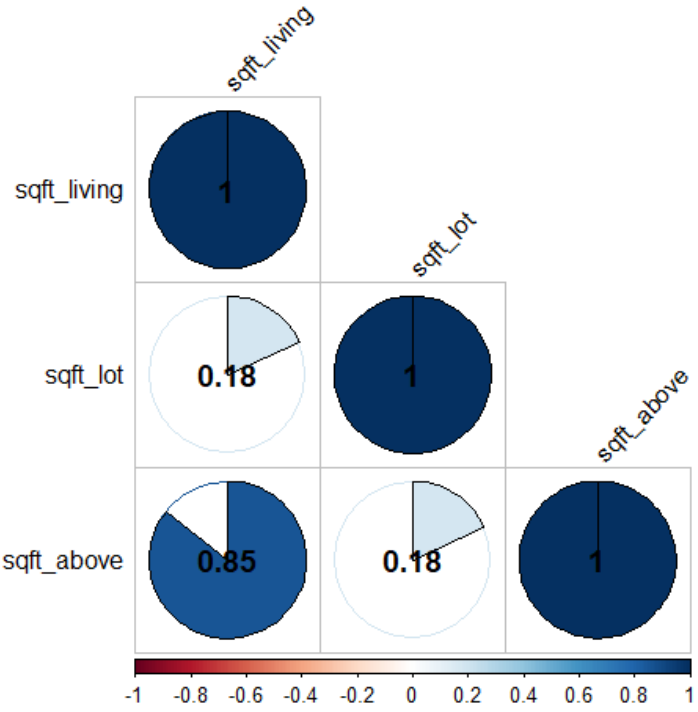
12. CORRELAZIONE TRA LE VARIABILI

Per stabilire la correlazione tra le variabili è stato creato un corrplot a cui sono state passate in input le 11 variabili precedentemente analizzate, ottenendo il risultato seguente.

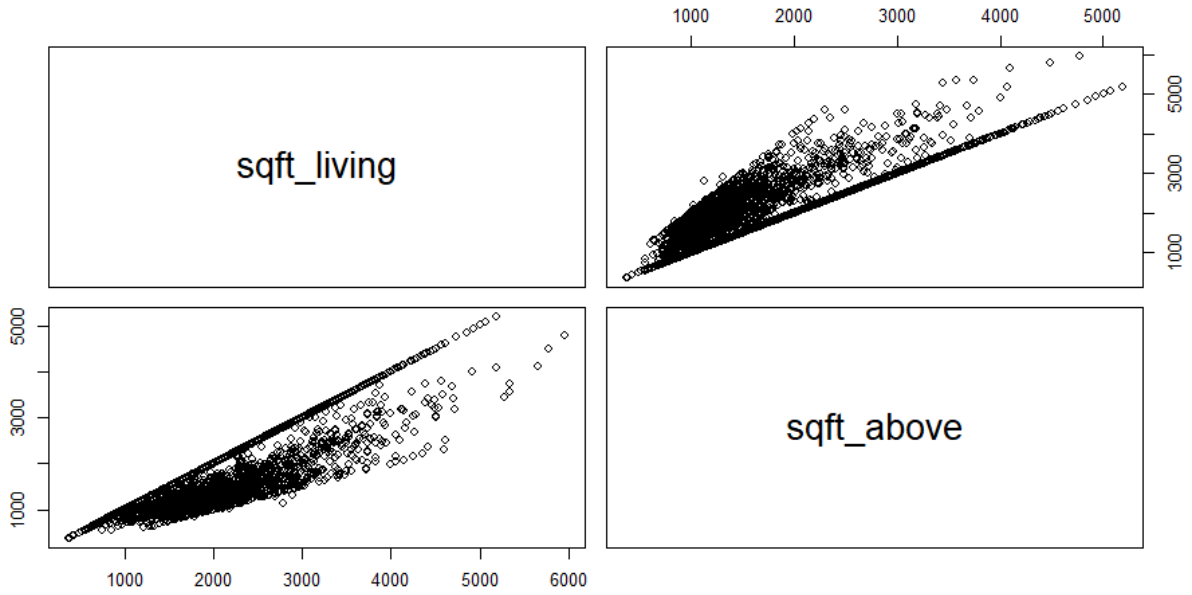


	price	bedrooms	bathrooms	sqft_living	sqft_lot	waterfront	condition	sqft_above	sqft_basement	yr_built	yr_renovated
price		0.3	0.48	0.62	0.09	0.07	0.04	0.54	0.15	0.12	-0.07
bedrooms	0.3		0.52	0.6	0.05	-0.03	0.01	0.46	0.16	0.13	-0.07
bathrooms	0.48	0.52		0.72	0.08	0.02	-0.14	0.64	0.13	0.53	-0.24
sqft_living	0.62	0.6	0.72		0.18	0.03	-0.09	0.85	0.17	0.35	-0.14
sqft_lot	0.09	0.05	0.08	0.18		0.03	0.01	0.18	-0.03	0.07	-0.01
waterfront	0.07	-0.03	0.02	0.03	0.03		0.03	0.02	0	-0.04	0.01
condition	0.04	0.01	-0.14	-0.09	0.01	0.03		-0.21	0.17	-0.36	-0.18
sqft_above	0.54	0.46	0.64	0.85	0.18	0.02	-0.21		-0.29	0.46	-0.17
sqft_basement	0.15	0.16	0.13	0.17	-0.03	0	0.17	-0.29		-0.16	0.06
yr_built	0.12	0.13	0.53	0.35	0.07	-0.04	-0.36	0.46	-0.16		-0.34
yr_renovated	-0.07	-0.07	-0.24	-0.14	-0.01	0.01	-0.18	-0.17	0.06	-0.34	

Dal momento che lo scopo del progetto è quello di trovare un modello valido per predire i prezzi delle case, viene analizzata la prima colonna di questo plot che mostra il livello di correlazione esistente tra il prezzo di una casa e le altre variabili; si può notare come il prezzo sia correlato positivamente con tutte le variabili (tranne yr_renovated) e in particolare è evidente che ci sia una correlazione abbastanza forte tra il prezzo e la dimensione della zona abitabile (sqft_living). La correlazione tra queste due variabili è la massima per quanto riguarda il prezzo anche se si registrano altri valori di correlazione moderata tra prezzo e numero di bagni (0.48), prezzo e dimensione del piano terra (sqft_above, correlazione = 0.54).
Analizzando gli altri valori della correlazione tra le altre variabili salta all'occhio la forte correlazione tra sqft_living e sqft_above, quindi si decide di analizzarla tramite un altro corrplot inserendo anche la variabile sqft_lot che indica la dimensione del lotto, del terreno.

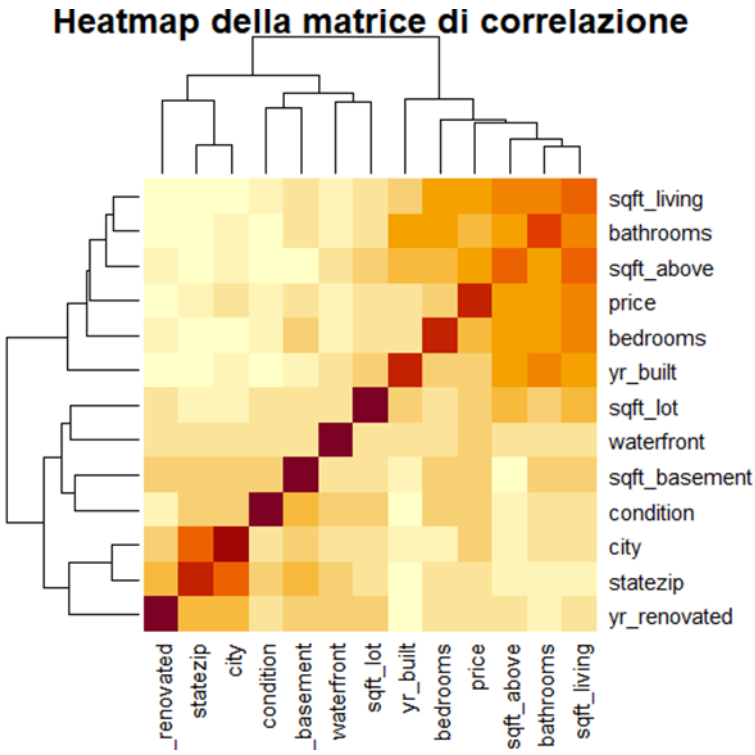


C'è una forte correlazione tra sqft_living e sqft_above, quindi si può dire che in molti casi le variabili assumono gli stessi valori; viene creato uno scatterplot tra queste due variabili.

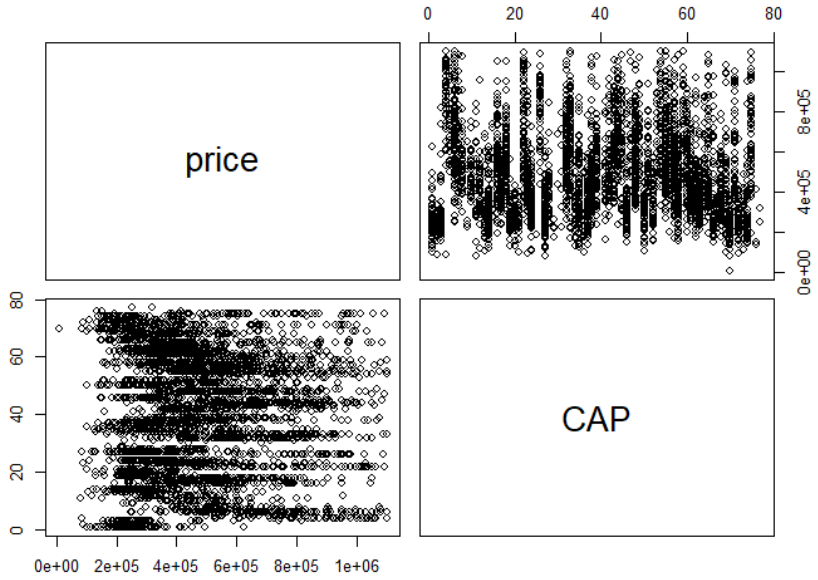


Viene rimossa una delle due colonne tra sqft_living e sqft_above (in particolare viene eliminata sqft_above), in quanto essendo molto correlate ci sono molti valori duplicati.

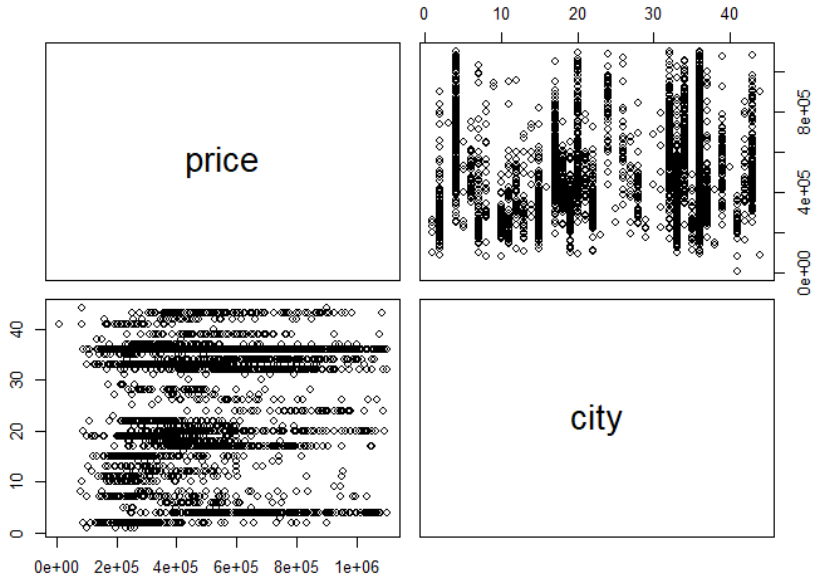
Viene creata quindi una heatmap della matrice di correlazione.



Vengono trasformate inoltre le ultime due variabili (ovvero city e statezip) da factor a num con as.numeric e viene valutato se esiste una correlazione tra la città in cui la casa si trova e il prezzo oppure tra la zona identificata dal CAP e il prezzo.



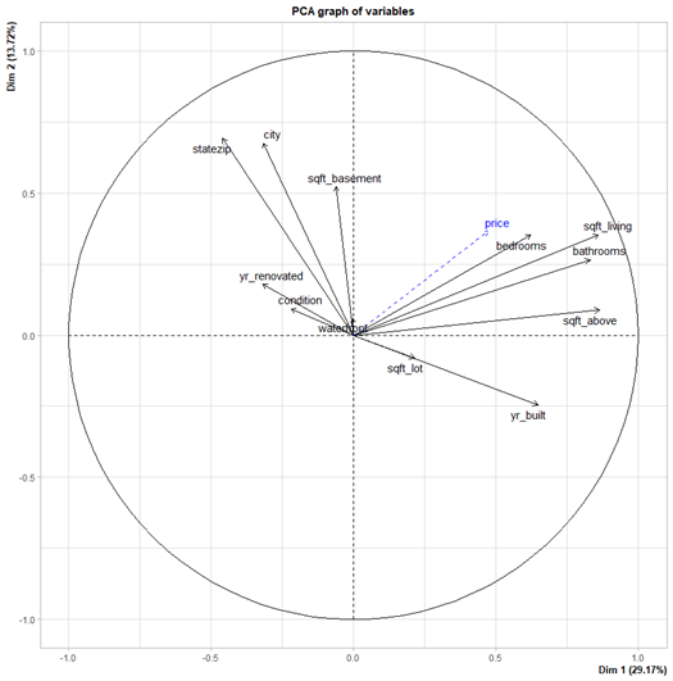
L'indice di correlazione tra queste due variabili è -0.02335188, quindi si può dire che le due variabili sono inversamente correlate.



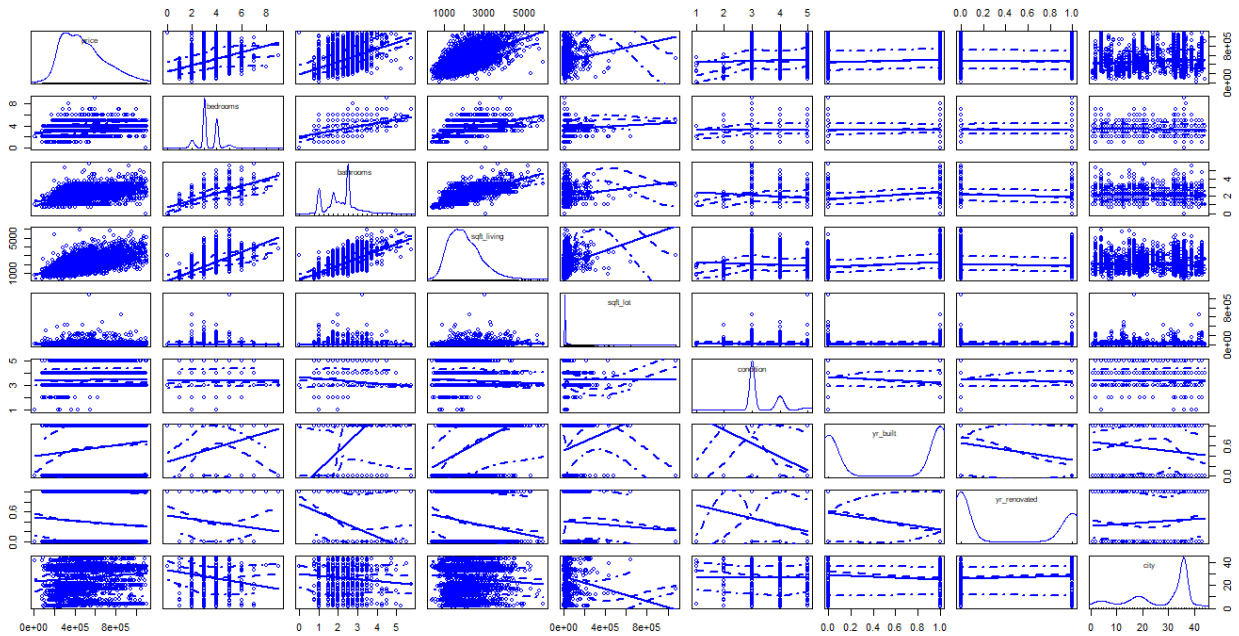
L'indice di correlazione tra queste due variabili è 0.1184739, quindi le due variabili sono molto poco correlate tra di loro.
Si può concludere quindi che il prezzo non è molto condizionato né dalla zona né dalla città dove la casa si trova.

Questo grafico indica la correlazione tra la città e il CAP della città, correlazione moderata in quanto il valore è pari a 0.67.

Inoltre, è stato creato questo grafico che mostra dei vettori che indicano quali siano le variabili che si muovono nella stessa direzione o in direzioni simili al prezzo.



Infine, è stata creata una scatterPlotMatrix con i valori dei prezzi, numero di camere, numero di bagni, dimensione della superficie abitabile, di quella del lotto, la condizione della casa, l'anno in cui è stata costruita, se è stata ristrutturata e la città in cui si trova la casa.



ML ALGORITHM

In questa sezione l'obiettivo è quello di creare dei modelli di predizione per predire il prezzo di una casa in base alle sue caratteristiche.

Il dataset viene diviso in due parti, ovvero training set e test set; in particolare il training set sarà formato dall'80% del dataset (3426 entries) e il test set sarà costituito dal restante 20%, ovvero 857 entries.

Sono stati usati due metodi principali di regressione, ovvero la regressione lineare e gli alberi di regressione e i modelli creati tramite questi due metodi sono stati validati utilizzando delle tecniche di cross validation.

Inizialmente è stata creata la InitFormula che contiene tutte e 12 le variabili del dataset e che utilizza come unica variabile dipendente quella del prezzo (price).
InitFormula = price ~ bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition + +sqft_basement + yr_built + yr_renovated + city + statezip

Viene creato quindi il modello denominato Casereg_train che utilizza la InitFormula e come dati considera quelli del training set; il risultato è il seguente.

```
Call:
lm(formula = InitFormula, data = training_val)

Residuals:
    Min      1Q  Median      3Q     Max
-952999 -109467   -8787   96109  704713

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.952e+04  2.300e+04   0.849   0.3962
bedrooms      -3.677e+04  3.937e+03  -9.341  < 2e-16 ***
bathrooms      5.112e+04  6.150e+03   8.312  < 2e-16 ***
sqft_living    1.741e+02  5.405e+00  32.206  < 2e-16 ***
sqft_lot      -1.493e-01  7.597e-02  -1.966   0.0494 *
waterfront     1.079e+05  4.321e+04   2.498   0.0125 *
condition      2.798e+04  4.609e+03   6.071 1.41e-09 ***
sqft_basement -2.333e+03  5.829e+03  -0.400   0.6891
yr_built      -5.795e+04  7.578e+03  -7.647 2.66e-14 ***
yr_renovated   8.794e+03  6.180e+03   1.423   0.1548
city           3.499e+03  2.979e+02  11.745  < 2e-16 ***
statezip      -7.933e+02  1.817e+02  -4.367 1.30e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 154400 on 3414 degrees of freedom
Multiple R-squared:  0.4588,    Adjusted R-squared:  0.4571
F-statistic: 263.1 on 11 and 3414 DF,  p-value: < 2.2e-16
```

Si può notare che 2 variabili non hanno una significatività molto alta (infatti hanno zero asterischi alla fine della riga), quindi verranno rimosse nella creazione del nuovo modello; inoltre analizzando il valore dell'adjusted R-squared si può concludere che il modello non è molto buono, efficace in quanto il suo valore non è molto vicino a 1 (0.4571).
Viene quindi effettuata la previsione usando questo modello e vengono calcolati i valori dei residui e degli errori di questa predizione, in particolare sono calcolati i valori del Root Mean Square Error (155328.7$) e del Mean Absolute Error (124472.1$) che verranno utilizzati in seguito per confrontarli con gli errori delle altre previsioni per capire quale sia la predizione più accurata.

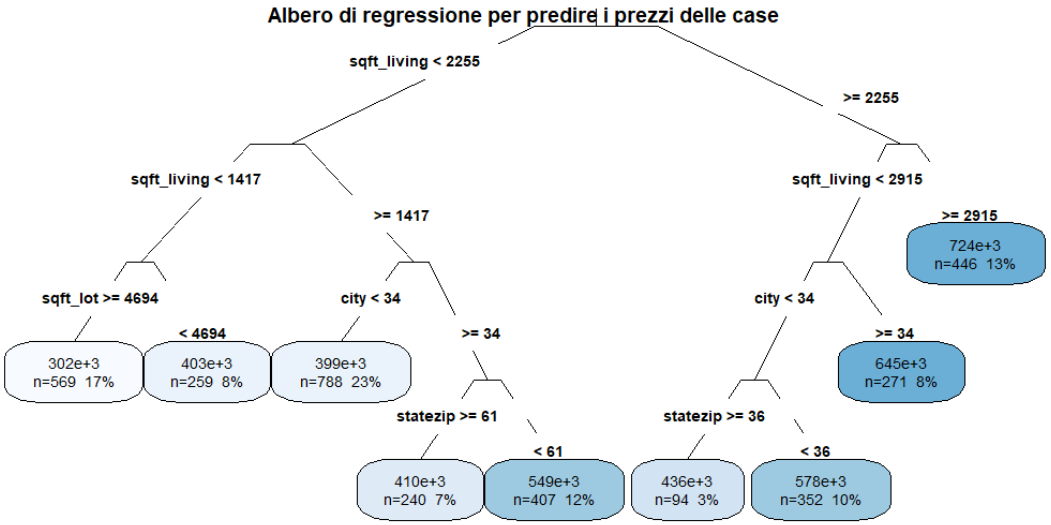
A questo punto viene creata una nuova formula, ovvero ImprovedFormula che è formata solo dalle colonne significative (sono state rimosse sqft_basement e yr_renovated).
ImprovedFormula = price ~ bedrooms + bathrooms + sqft_living + condition + waterfront + sqft_lot + +yr_built + city + statezip

Creando il modello che usa ImprovedFormula il risultato che si ottiene è il seguente.

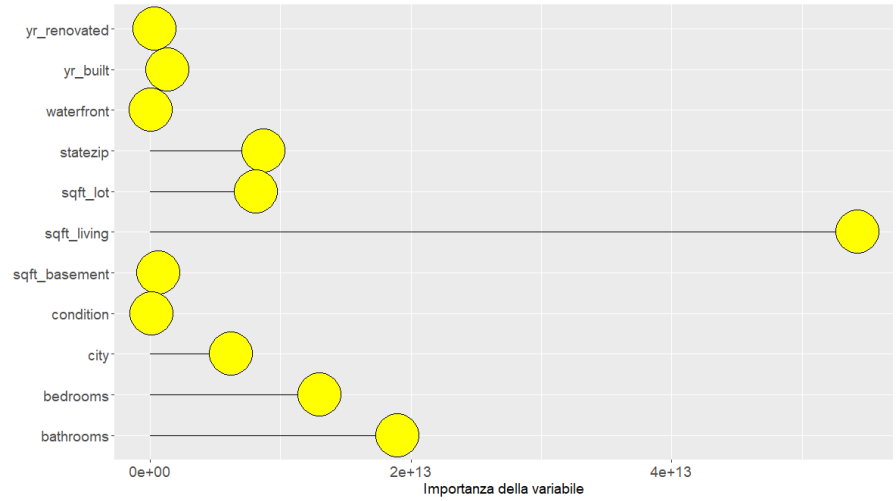
```
Call:
lm(formula = ImprovedFormula, data = training_val)

Residuals:
    Min      1Q  Median      3Q     Max
-933526 -110073  -10422   95587  700689

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.315e+04  2.107e+04   1.099  0.27195
bedrooms    -3.142e+04  3.917e+03  -8.020 1.43e-15 ***
bathrooms    5.260e+04  6.149e+03   8.554  < 2e-16 ***
sqft_living  1.662e+02  5.340e+00  31.126  < 2e-16 ***
condition    2.465e+04  4.286e+03   5.751 9.65e-09 ***
waterfront   1.344e+05  4.161e+04   3.229  0.00125 **
sqft_lot    -1.443e-01  7.260e-02  -1.988  0.04694 *
yr_built    -5.863e+04  7.025e+03  -8.345  < 2e-16 ***
city         3.811e+03  2.970e+02  12.834  < 2e-16 ***
statezip    -8.706e+02  1.812e+02  -4.805 1.62e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 154200 on 3416 degrees of freedom
Multiple R-squared:  0.4503,    Adjusted R-squared:  0.4489
F-statistic: 310.9 on 9 and 3416 DF,  p-value: < 2.2e-16
```

Si può quindi dire che questo modello creato è meno accurato e meno efficiente di quello creato precedentemente in quanto il valore dell'adjusted R-squared è più lontano da 1 rispetto al precedente; anche in questo caso sono stati calcolati i valori di MAE e RMSE ottenendo rispettivamente 122797.7$ e 155782.3$, valori più o meno simili rispetto a quelli precedenti (uno di poco superiore, uno di poco inferiore) che rendono la predizione poco più accurata.

Successivamente sono stati utilizzati gli alberi di regressione utilizzando le variabili del modello migliore precedentemente trovato. Il modello ottenuto è il seguente

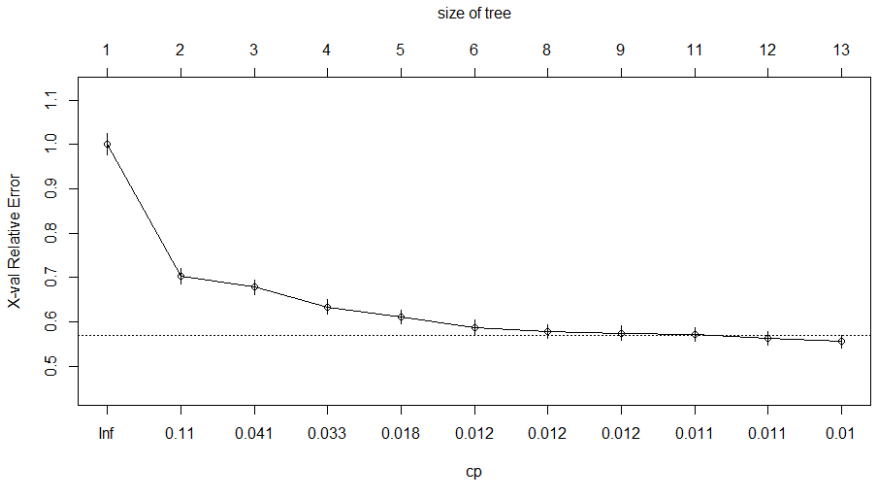


Dopo aver creato questo albero viene stampato anche un ggplot che mostra quali siano le variabili più influenti.



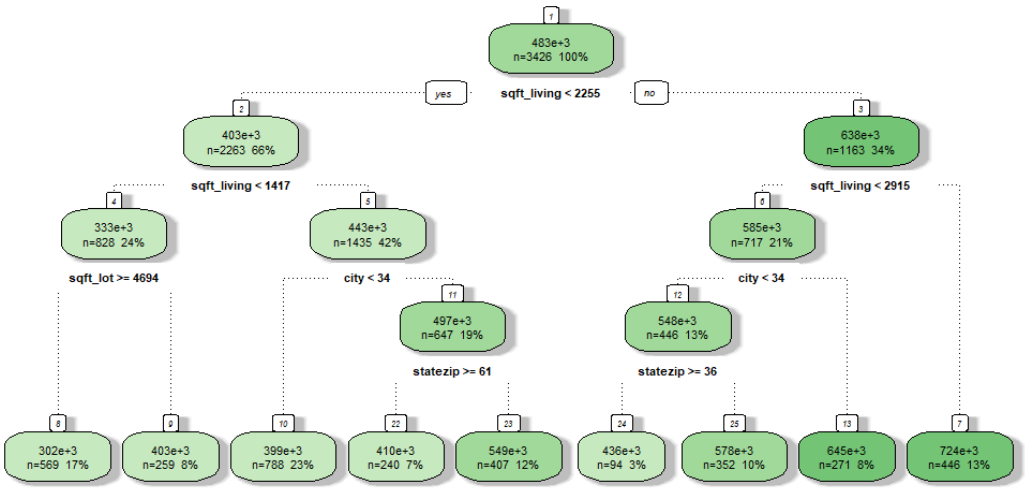
Questo modello ad albero viene utilizzato per effettuare una previsione sui valori del test set che produce un Mean Absolute Error pari a 125192.6$ e un Root Mean Square Error che ha valore 160512.9$; entrambi questi errori hanno un valore superiore rispetto ai rispettivi errori calcolati dalle due predizioni precedenti, quindi la predizione creta con questo albero di regressione è meno accurata rispetto alle precedenti.

In seguito, è stata eseguita una valutazione sull'albero e in particolare su quale fosse il valore di size tree che rendesse l'errore minimo.



Si può notare che il valore di cp (complexity parameter) che rende l'errore minimo è 0.01 quindi viene creato e successivamente stampato l'albero potato (utile in quanto permette di evitare l'overfitting) tramite la seguente istruzione

```
prunedTree = prune(fit, cp=0.01)
fancyRpartPlot(prunedTree)
```



In seguito sono stati usati altri tipi di alberi di regressione detti Weka classifier tra cui M5P, J48 e LMT; per gli ultimi due non è stato possibile trovare un valore di MAE e di RMSE in quanto non funzionano con classi di tipo numeriche, mentre con l'albero di regressione M5P sono stati ottenuti i valori degli errori pari a 95391.54$ per il MAE (minore dei valori dei MAE trovati precedentemente) e 132675.3$ per l'RMSE (20 mila $ in meno rispetto agli RMSE precedenti) che evidenziano il fatto che questa predizione sia molto più accurata e che sia quindi la migliore previsione utilizzata finora.

Infine, per validare il modello migliore sono state utilizzate tre tecniche di cross validation, ovvero la cross validation k-fold, la repeated cross validation e la leave one out cross validation (loocv).

La cross validation k-fold consiste nella suddivisione del dataset in in numero 'number' di parti di uguale numerosità e nell'utilizzo della parte numero 'number' come test set; viene utilizzato principalmente perché si allena il modello per ogni parte e si evitano problemi di overfitting e di campionamento asimmetrico.

Cross validation k-fold

```
set.seed(123)
train.control_k <- trainControl(method = "cv", number = 10)
model_k <- train(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition +
+sqft_basement + yr_built + yr_renovated + city + statezip,
data = case.df, method = "lm", trControl = train.control_k)
print(model_k)
```

La repeated cross validation è molto simile alla cross validation descritta precedentemente, ma risulta utile in quanto ripete la cross validation per un numero di volte stabilito (in questo caso la variabile repeats indica il numero di ripetizioni) e restituisce il risultato medio delle diverse iterazioni; per questo solitamente risulta più precisa della semplice k-fold.

Repeated cross validation

```
set.seed(123)
train.control_krep <- trainControl(method = "repeatedcv", number = 10, repeats = 3)
model_krep <- train(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition +
+sqft_basement + yr_built + yr_renovated + city + statezip,
data = case.df, method = "lm", trControl = train.control_krep)
print(model_krep)
```

La loocv, invece, non divide il dataset in due sottoinsiemi delle stesse dimensioni, ma si basa su una singola osservazione per la validazione e tutte le altre osservazioni costituiscono il training set; questa procedura è ripetuta tante volte quante sono le variabili nel dataset e gli errori vengono calcolati come la media dei singoli errori calcolati volta per volta.

Loocv (Leave one out cross validation)

```
set.seed(123)
train.control_loocv <- trainControl(method = "LOOCV")
model_loocv <- train(price ~ bedrooms + bathrooms + sqft_living + sqft_lot + waterfront + condition +
+sqft_basement + yr_built + yr_renovated + city + statezip,
data = case.df, method = "lm", trControl = train.control_loocv)
print(model_loocv)
```

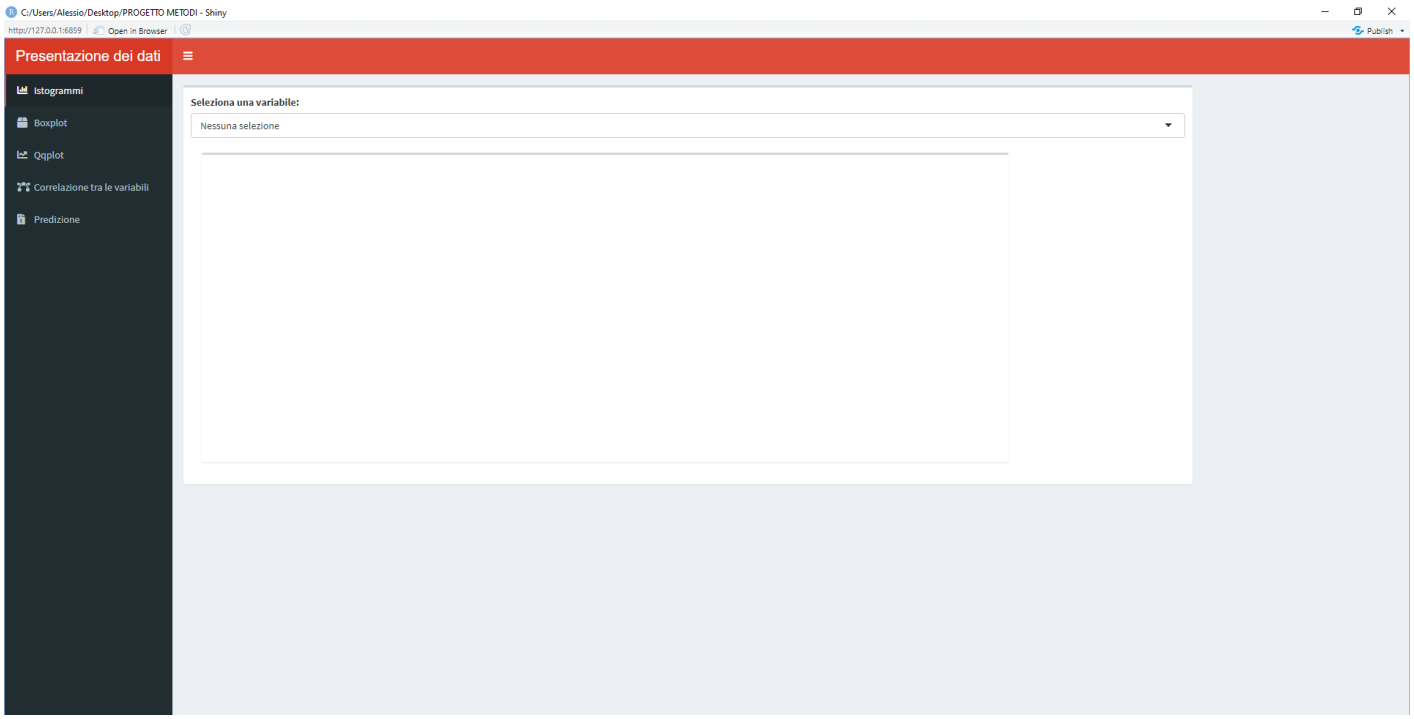
Analizzando i valori degli errori ottenuti da questi tre modelli di cross validation il risultato che si otterrà sarà il seguente

- cross validation k-fold: RMSE = 154684.8$, MAE = 121715.9$
- repeated cross validation: RMSE = 154698.7 $, MAE = 121669.2$
- Leave one out cross validation: RMSE = 154792.6$, MAE = 121647.4$

Si può dunque capire che anche il modello creato con l'albero di regressione M5P è il migliore possibile, è il modello ottimale, in quanto i valori di MAE e RMSE sono minori rispetto a quelli trovati tramite le tecniche di cross validation; quindi si può concludere che tramite le tecniche di cross validation è stato validato il modello ottimale che risulta essere quello creato con M5P, albero costruito ricostruendo l'algoritmo M5 di Quinlan e che consiste in due fasi: una prima fase, dove viene ottenuto l'albero che meglio descrive i dati a partire dalla costruzione dei possibili alberi ricavabili scegliendo il miglior compromesso tra minori dimensioni dell'albero e maggiore capacità descrittiva; una seconda fase di "potatura" per ridurre le dimensioni dello stesso, ottenendo un minor costo computazionale e una migliore interpretabilità delle regole, a discapito dell'accuratezza anche se in maniera poco significativa.

WEBAPP

L'applicazione è stata creata usando Rshiny e ha due funzionalità principali, ovvero quella di mostrare i vari grafici della sezione riguardante exploratory analysis e quella di predire il prezzo di una casa in base ai valori delle variabili che si passano in input.
Si presenta in questo modo.



Sono quindi presenti 5 sezioni, ovvero istogrammi, boxplot, qqplot, correlazione tra le variabili e predizione.

Nella prima, cliccando sul menu a tendina (selectInput), si può scegliere la variabile di cui visualizzare l'istogramma che verrà stampato nel riquadro bianco sottostante; lo stesso comportamento sarà replicato nelle due sezioni sottostanti in cui verranno stampati rispettivamente boxplot e qqplot delle variabili selezionate nel menu a tendina.

Nella quarta sezione, ovvero quella relativa alla correlazione tra le variabili, appare un menù a scelta multipla in cui si può selezionare, una alla volta, una delle opzioni presenti; una volta fatta la scelta verrà mostrato il grafico corrispondente, in particolare selezionando corr_mixed verrà stampato il grafico che descrive la correlazione tra tutte le variabili del dataset (tranne city e statezip), selezionando corr_superficie viene mostrata la correlazione esistente tra le tre variabili riguardanti le superfici rimaste all'interno del dataset (infatti sqft_above è stata precedentemente eliminata), selezionando heatmap viene stampata la heatmap che mostra la correlazione delle variabili, selezionando corr_citta verrà mostrato il corrplot che descrive la correlazione tra prezzo, città e CAP e infine selezionando corr_PCA verrà stampato il grafico che indica le variabili come vettori e ne mostra la direzione.

Infine, nella sezione riguardante la predizione, è possibile, tramite delle caselle di testo, modificare la quantità di una variabile (esempio numero di bagni, condizione della casa, città in cui si trova la casa, ecc.) e vedere come il prezzo predetto vari.
Per creare questa predizione è stato usato il modello migliore trovato e validato precedentemente, ovvero M5P che viene passato in input nel momento in cui viene eseguita la previsione e il risultato viene approssimato per difetto eliminando le cifre decimali.

CONCLUSIONS

L'obiettivo di questo progetto è stato quello di analizzare un dataset e le sue variabili, cercando di avere informazioni sulla loro distribuzione e sulla correlazione tra di esse; in particolare sono stati analizzati i gradi di correlazione tra il prezzo e le altre variabili in quanto sono state sviluppate delle predizioni sul prezzo in relazione alle altre variabili con diverse metodologie quali regressione semplice e alberi di regressione.

Inizialmente il dataset è stato analizzato nel dettaglio stampando le tipologie di variabili che lo compongono; successivamente sono stati cercati missing values all'interno del dataset (ricerca che ha dato esito negativo), sono state eliminate alcune colonne (variabili) ritenute poco utili ai fini dell'analisi, sono state analizzate nel dettaglio le variabili rimanenti operando anche su di esse con delle trasformazioni in modo che fossero più fruibili (per esempio alcune variabili sono state trasformate in variabili binarie, alcune sono state trasformate da variabili testuali, ovvero factor, a variabili numeriche).

In seguito all'analisi sulla correlazione tra le diverse variabili rimaste nel dataset è stato quindi evidente come il prezzo fosse moderatamente correlato con poche variabili quali sqft_living, sqft_above, bathrooms e correlato in maniera bassa o correlato negativamente con indice di correlazione <= 3 con il resto delle variabili del dataset.

Dall'applicazione dell'algoritmo di regressione invece è stato possibile concludere che, tra i modelli creati sia con regressione lineare che con alberi di regressione, il più accurato e quello che restituisse quindi l'errore (RMSE o MAE) minimo fosse quello creato con un albero di regressione di tipo M5P.

Infine, per validare il modello migliore trovato sono state usate tre differenti tipologie di cross validation, ovvero la cross validation k-fold, la repeated cross validation e la leave one out cross validation che hanno permesso di concludere che il modello trovato fosse il modello più valido in assoluto in quanto le cross validation hanno portato ad avere dei valori degli errori maggiori rispetto a quelli trovati con la previsione effettuata sul modello creato con M5P.

Come ultima operazione è stata creata una webapp che permettesse sia di visualizzare a livello grafico il comportamento delle diverse variabili del dataframe tramite grafici quali istogrammi, boxplot, qqplot e grafici di correlazione sia di poter effettuare una previsione sul prezzo cambiando il valore dei parametri in input.