ANALYSIS OF THE CHARACTERISTICS OF ESTATES AND THEIR INFLUENCE ON HIS PRICE

Marco Speciale

Index

1. Introduction

During the years 2007-2008 a serious national and global economic crysis broke out which has been caused by the bursting of a real estate bubble in the United States of America. The crysis destabilized the markets until 2013 . The effects of what has been called "The Great Recession" have not been stemmed, despite an ups and downs of collapses and recoveries, to the point of affecting the economy of today. The situation was further aggravated by the global Sars-Cov-2 pandemic, which, from Wuhan, spread upon the world. Many families were hit hard by the rise in the price of raw materials and a credit-banking crisis and they have seen cut their income. They hardly managing buying real estate. In a historical period in which an excessive supply coexists with a contraction in demand, the professional figure of the Statistician acquires more relevance, since his technical knowledge allows to analyze and draw conclusions from countless amounts of useful data in guiding customers towards a conscious purchase choice.

The data that the team acquired relates to the municipality of Palermo; therefore, they cannot be generalized for the entire national territory, especially if we take into account the great socio-economic variety of the Peninsula.

The focus of the analysis was the impact of the characteristics have on the determination of an estate price.

Paying particular attention to all the elements most taken into consideration by a hypothetical buyer, such as the rental area, the building's floors, the number of rooms and bathrooms available.

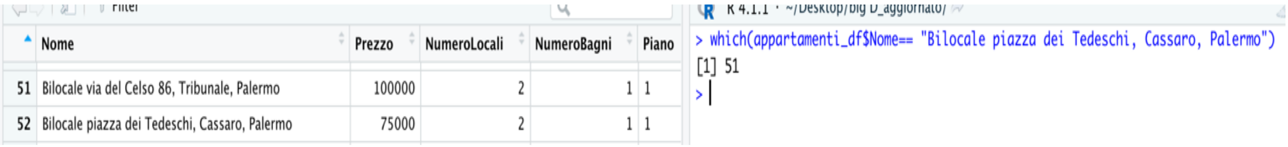
Information was extracted through the obtaining and observation of a number of real estate advertisements that refer to the name and description of the estates.

2. Data extraction and data clearing

Data were obtained from the Immobiliare.it website, using the Google Chrome **selector gadget** extension.
The "R" programming software was used to obtain the data and analyze them.
The extracted datasets are the following:

- Dataset of apartments for sale in excellent condition in the center of Palermo
- Dataset on attics for sale in the municipality of Palermo
- Dataset of the intersection between the two aforementioned datasets
- Dataset of an apartment chosen from among the intersection of attics and apartments datasets

After having taken the individual variables and merging them into a dataframe, was checked the presence of repetitions among the observations.
In the first dataset 6 repetitions were found on 164 observations. It was used the **unique** function, which eliminates repetitions. However, re-examining the dataset, it was clear that the repetitions were not completely removed, because two of them contained a slightly different **Description** from the others. Therefore, the aforementioned excess repetitions were removed by assigning rows to the dataset different from the original format. Furthermore, the unique function has changed the numbering of the observations in a tricky way, returning a different observation than the one entered in the code; to make things clear, an example is given:



	Nome	Prezzo	NumeroLocali	NumeroBagni	Piano
51	Bilocale via del Celso 86, Tribunale, Palermo	100000	2	1	1
52	Bilocale piazza dei Tedeschi, Cassaro, Palermo	75000	2	1	1

```
> which(appartamenti_df$Nome== "Bilocale piazza dei Tedeschi, Cassaro, Palermo")
[1] 51
>
```

Then statistical units were reordered.

The process reported was also used for the creation of the attics dataframe.
In this dataset 3 repetitions were found among 116 observations.

The dataset that consists in the intersection of the two previous ones was made by the **semi_join** function which takes into account all the observations that in the first dataset (apartments) have a correspondence in the second (attics).
It was chosen a fairly common type of apartment for the average customer.
To carry out this study, the attic located in Via Giuseppe Garibaldi, in the Kalsa district of Palermo was chosen as the object of the investigation.

In order to verify that the datasets had valid values, were applied rules, with particular reference to the variables floor - number of rooms - number of bathrooms - price; All the numerical variables were set as strictly positive, the price was set greater than the number of bathrooms and rooms, and the floors had to reflect the site's standards. (rules chart → figure 1 in appendix)

7 errors were founded in the apartment dataset (as shown in the graph). These values are plausible, so the errors highlighted are due to the presence of NA; For this reason the dataset has not been further modified. (rules violation chart → figure 2 in appendix)

No rule violations were found in the Attics dataset. (rules violation chart → figure 3 in appendix)

3. Exploratory analysis

In order to verify which factors had the greatest impact on the determination of the price, we proceeded with various graphical representations, which shows that in the distribution of the price of both datasets (apartments and attics) there is the presence of an anomalous value for each one of them.

The anomalous value in the apartments dataset corresponds to 3.7 € million (MLN), in the other one, the outlier value corresponds to € 2.2 million (MLN).

Synthetic measures were calculated for each aspect of the estates with and without the anomalous values.

The analysis started with the study of the variable **Floor**. The units were grouped for each **Floor** mode, then was calculated the price averages for each group. There were no substantial differences between the group means. Although the expectations were that a higher floor would correspond to a higher price, this did not happen. Then it was decided to not remove NA in order to not lose further information. (Table 1 in the appendix)

With regard to the **Number of Rooms**, it was adopted the same procedure as the previous one. In this case, however, results show that the price was directly proportional to the number of rooms in the apartments. (Table 5 in the appendix)

Then it was analyzed the **Number of Bathrooms**, the same results were found as the previous variable. It was clear from both the table and the graph that the price increased as the number of bathrooms available increased. (Table 3 and Figure 6 in the appendix)

In order to analyze the neighborhood variable, the units were grouped for each modality of the variable itself and the price averages for each group were then calculated, as done for the previous variables.

The S. Erasmo district is on average the one with a higher price, probably due to its proximity to the sea area. It was noticed that the average price of the Kalsa area was strongly influenced by the presence of the anomalous value; in fact, from the second position for the highest average price, it moved back by one, making the Castellammare district in second place. (Tables 9 - 10 in the appendix).

In order to verify the presence of any similarities between the advertisements descriptions we made a wordcloud graphic, from which it was deduced that the most frequent type of advertisement concerns three-room and two-room apartments located in Via Roma and in the Tribunal district. (Figure 4 in the appendix)

The same considerations were made for the Number of Rooms (table 6 in the appendix) and Floor (Table 2 see in the appendix) are even more marked if we analyze the Attics dataset, with the sole exception of the Number of Bathrooms (Table 4 and Figure 7 in the appendix) where the average price per number of bathrooms triples going from 2 to 3 bathrooms in the apartments' dataset, while in the other dataset this price just doubles.

With regards the neighborhood, on the other hand, differences were found with respect to the previous dataset. This happened for two reasons:

- There is a greater number of areas because the dataset in question takes into account the municipality of Palermo, as opposed to the apartment dataset (which only and exclusively considers the historical center of Palermo).
- There are fewer observations for each group, given that the sample size of the attic dataset is lower than that of the apartments. (Figures 8 - 9 in the appendix).

Nothing this, it can be deduced that the anomalous value has a greater weight on the average price of its reference group since it cannot be absorbed by the other values.

It was noticed that the average price of the Borgo Vecchio area is strongly influenced by the presence of the anomalous value, in fact, from the area with the highest average price it falls back by as many as 15 positions; it emerged that the first three areas with the highest average price, without the anomalous value, are in order: Politeama, Strasbourg and S. Erasmo. (Tables 7 - 8 and Figure 10 in the appendix).

The S. Erasmo district is reconfirmed as one of the areas with the highest average price, taking also into account each area of the municipality of Palermo.

In order to verify the presence of any similarities between the descriptions of the ads, a wordcloud graph was drawn, from which, it is clear that the most frequent type of ad describes a bright attic with a terrace with a panoramic view in the district of Via Lanza di Scalea and Corso Calatafimi.

(Figure 5 in the appendix)

4. Conclusions

The following analysis highlights how the elements that most affect the price are:

- Numbers of Rooms
- Numbers of Bathrooms
- Neighbrood

The floor is not so relevant for determining the price of a property. It also denotes how the types of properties most offered for sale are two-room and three-room apartments. Furthermore, the districts where are estates with the highest value are Politeama, Strasburgo and S. Erasmo.

The results of our analysis may not be entirely reliable because the statistical units that in possession are just a few.

Appendix

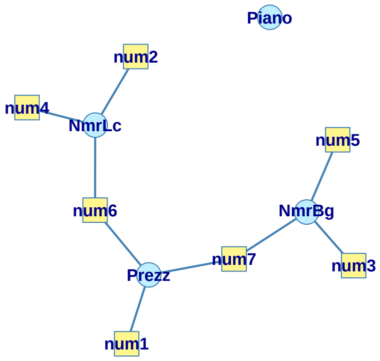


Image 1 rules

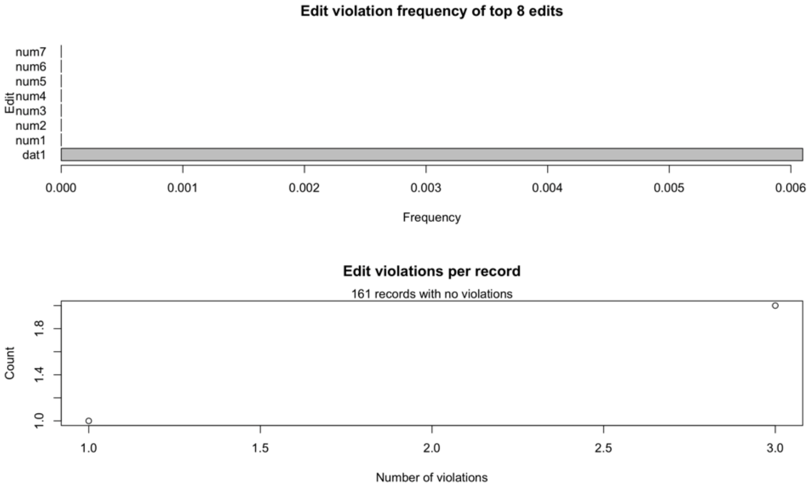


Image 2 df apartments rules violation

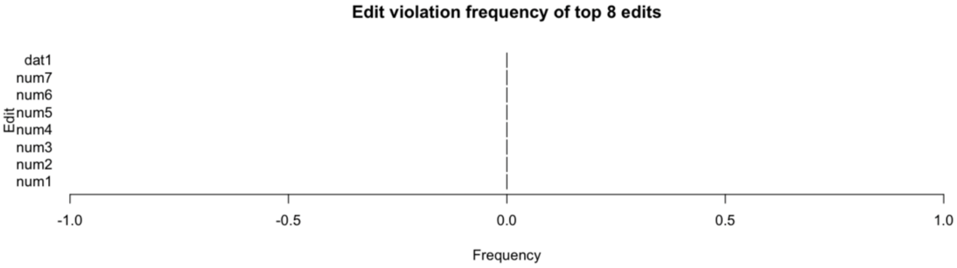
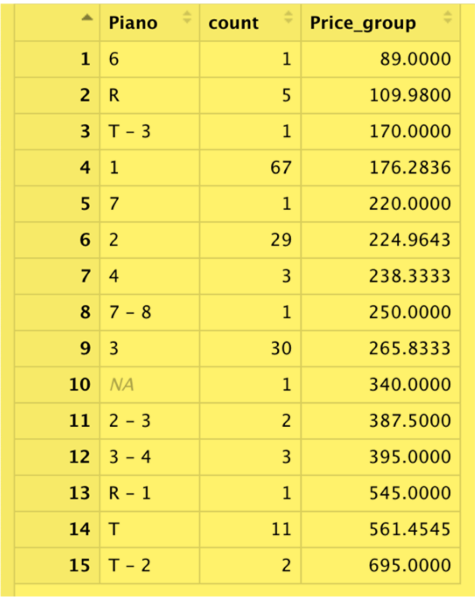
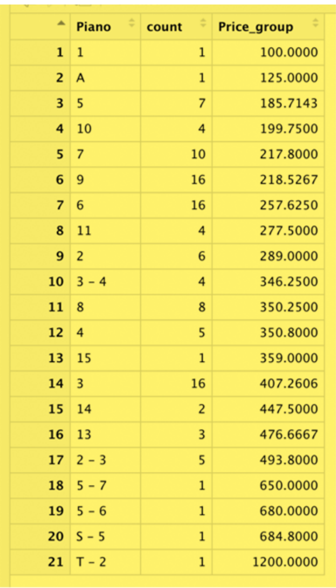


Image 3 df attics rules violation



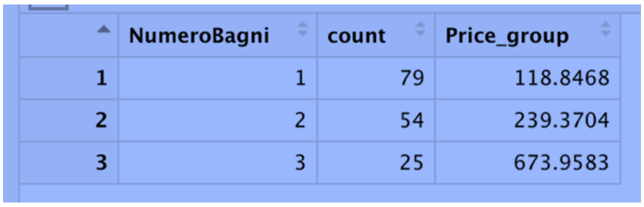
	Piano	count	Price_group
1	6	1	89.0000
2	R	5	109.9800
3	T - 3	1	170.0000
4	1	67	176.2836
5	7	1	220.0000
6	2	29	224.9643
7	4	3	238.3333
8	7 - 8	1	250.0000
9	3	30	265.8333
10	NA	1	340.0000
11	2 - 3	2	387.5000
12	3 - 4	3	395.0000
13	R - 1	1	545.0000
14	T	11	561.4545
15	T - 2	2	695.0000

Table 1 apartments df price average per floor



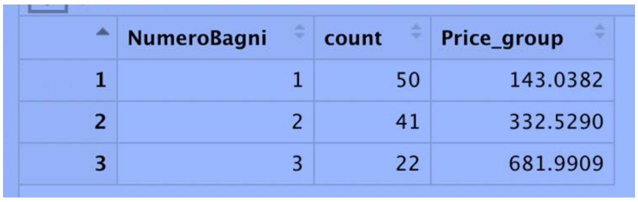
	Piano	count	Price_group
1	1	1	100.0000
2	A	1	125.0000
3	5	7	185.7143
4	10	4	199.7500
5	7	10	217.8000
6	9	16	218.5267
7	6	16	257.6250
8	11	4	277.5000
9	2	6	289.0000
10	3 - 4	4	346.2500
11	8	8	350.2500
12	4	5	350.8000
13	15	1	359.0000
14	3	16	407.2606
15	14	2	447.5000
16	13	3	476.6667
17	2 - 3	5	493.8000
18	5 - 7	1	650.0000
19	5 - 6	1	680.0000
20	S - 5	1	684.8000
21	T - 2	1	1200.0000

Table 1 attics df price average per floor



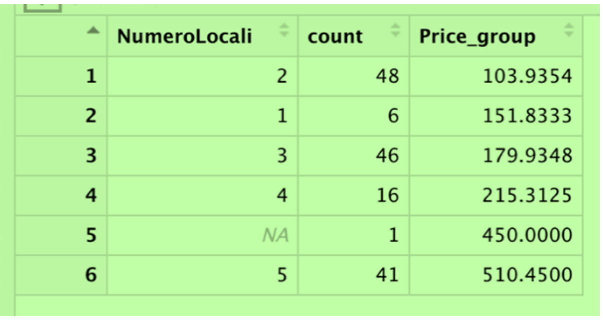
	NumeroBagni	count	Price_group
1	1	79	118.8468
2	2	54	239.3704
3	3	25	673.9583

Table 3 apartments df averages price per bathrooms



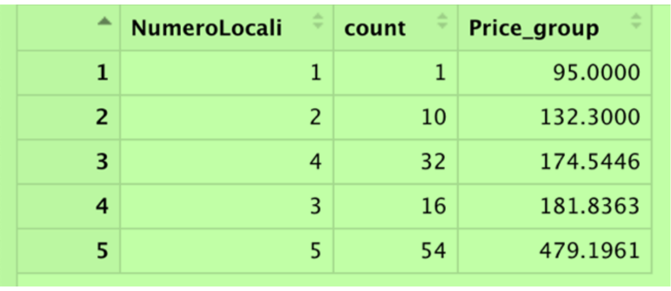
	NumeroBagni	count	Price_group
1	1	50	143.0382
2	2	41	332.5290
3	3	22	681.9909

Table 4 attics df averages price per bathrooms



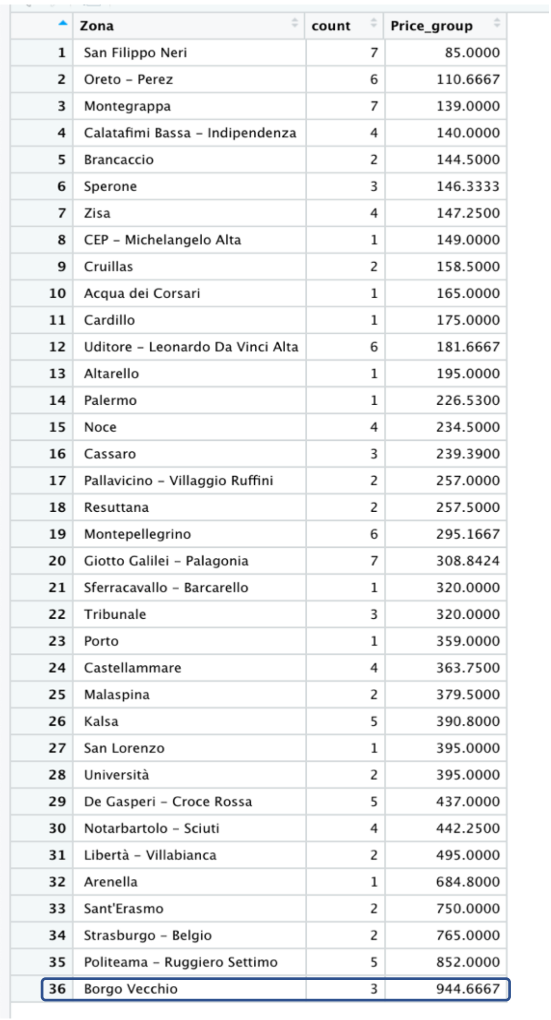
	NumeroLocali	count	Price_group
1	2	48	103.9354
2	1	6	151.8333
3	3	46	179.9348
4	4	16	215.3125
5	NA	1	450.0000
6	5	41	510.4500

Table 5 apartments df average price by number of rooms



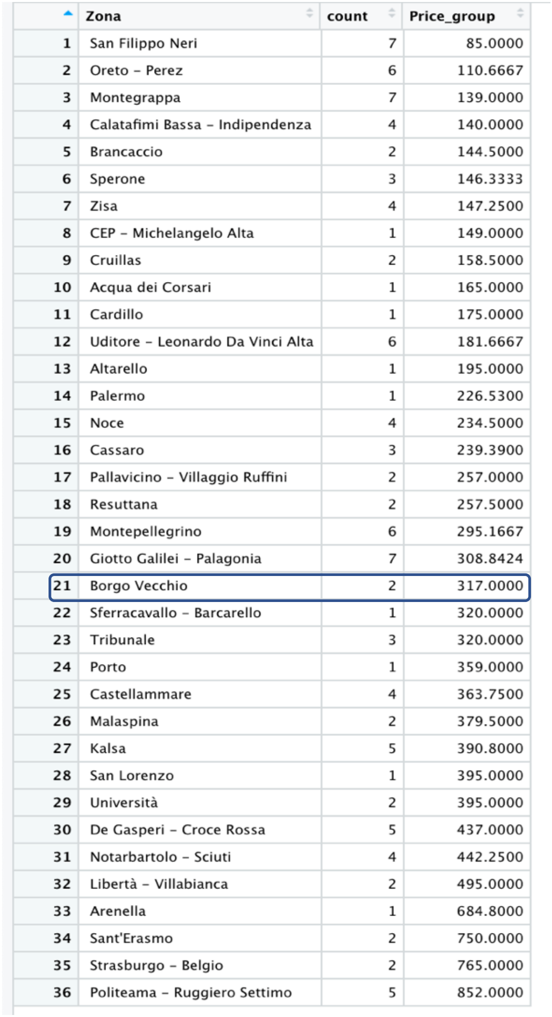
	NumeroLocali	count	Price_group
1	1	1	95.0000
2	2	10	132.3000
3	4	32	174.5446
4	3	16	181.8363
5	5	54	479.1961

Table 6 attics df average price by number of rooms



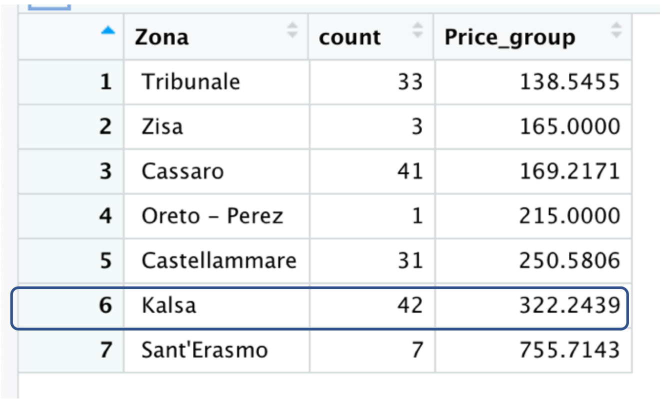
	Zona	count	Price_group
1	San Filippo Neri	7	85.0000
2	Oreto – Perez	6	110.6667
3	Montegrappa	7	139.0000
4	Calatafimi Bassa – Indipendenza	4	140.0000
5	Brancaccio	2	144.5000
6	Sperone	3	146.3333
7	Zisa	4	147.2500
8	CEP – Michelangelo Alta	1	149.0000
9	Cruillas	2	158.5000
10	Acqua dei Corsari	1	165.0000
11	Cardillo	1	175.0000
12	Uditore – Leonardo Da Vinci Alta	6	181.6667
13	Altarello	1	195.0000
14	Palermo	1	226.5300
15	Noce	4	234.5000
16	Cassaro	3	239.3900
17	Pallavicino – Villaggio Ruffini	2	257.0000
18	Resuttana	2	257.5000
19	Montepellegrino	6	295.1667
20	Giotto Galilei – Palagonia	7	308.8424
21	Sferracavallo – Barcarello	1	320.0000
22	Tribunale	3	320.0000
23	Porto	1	359.0000
24	Castellammare	4	363.7500
25	Malaspina	2	379.5000
26	Kalsa	5	390.8000
27	San Lorenzo	1	395.0000
28	Università	2	395.0000
29	De Gasperi – Croce Rossa	5	437.0000
30	Notarbartolo – Sciuti	4	442.2500
31	Libertà – Villabianca	2	495.0000
32	Arenella	1	684.8000
33	Sant'Erasmo	2	750.0000
34	Strasburgo – Belgio	2	765.0000
35	Politeama – Ruggiero Settimo	5	852.0000
36	Borgo Vecchio	3	944.6667

Table 7 attics df price averages by area with outlier



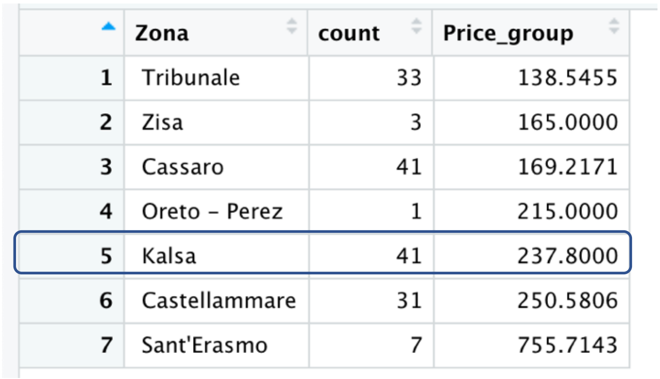
	Zona	count	Price_group
1	San Filippo Neri	7	85.0000
2	Oreto – Perez	6	110.6667
3	Montegrappa	7	139.0000
4	Calatafimi Bassa – Indipendenza	4	140.0000
5	Brancaccio	2	144.5000
6	Sperone	3	146.3333
7	Zisa	4	147.2500
8	CEP – Michelangelo Alta	1	149.0000
9	Cruillas	2	158.5000
10	Acqua dei Corsari	1	165.0000
11	Cardillo	1	175.0000
12	Uditore – Leonardo Da Vinci Alta	6	181.6667
13	Altarello	1	195.0000
14	Palermo	1	226.5300
15	Noce	4	234.5000
16	Cassaro	3	239.3900
17	Pallavicino – Villaggio Ruffini	2	257.0000
18	Resuttana	2	257.5000
19	Montepellegrino	6	295.1667
20	Giotto Galilei – Palagonia	7	308.8424
21	Borgo Vecchio	2	317.0000
22	Sferracavallo – Barcarello	1	320.0000
23	Tribunale	3	320.0000
24	Porto	1	359.0000
25	Castellammare	4	363.7500
26	Malaspina	2	379.5000
27	Kalsa	5	390.8000
28	San Lorenzo	1	395.0000
29	Università	2	395.0000
30	De Gasperi – Croce Rossa	5	437.0000
31	Notarbartolo – Sciuti	4	442.2500
32	Libertà – Villabianca	2	495.0000
33	Arenella	1	684.8000
34	Sant'Erasmo	2	750.0000
35	Strasburgo – Belgio	2	765.0000
36	Politeama – Ruggiero Settimo	5	852.0000

Table 8 attics df price averages by area without outlier



	Zona	count	Price_group
1	Tribunale	33	138.5455
2	Zisa	3	165.0000
3	Cassaro	41	169.2171
4	Oreto – Perez	1	215.0000
5	Castellammare	31	250.5806
6	Kalsa	42	322.2439
7	Sant'Erasmo	7	755.7143

Table 9 apartments df price average by Area with outlier outlier



	Zona	count	Price_group
1	Tribunale	33	138.5455
2	Zisa	3	165.0000
3	Cassaro	41	169.2171
4	Oreto – Perez	1	215.0000
5	Kalsa	41	237.8000
6	Castellammare	31	250.5806
7	Sant'Erasmo	7	755.7143

Table 10 apartments df price averages by area without

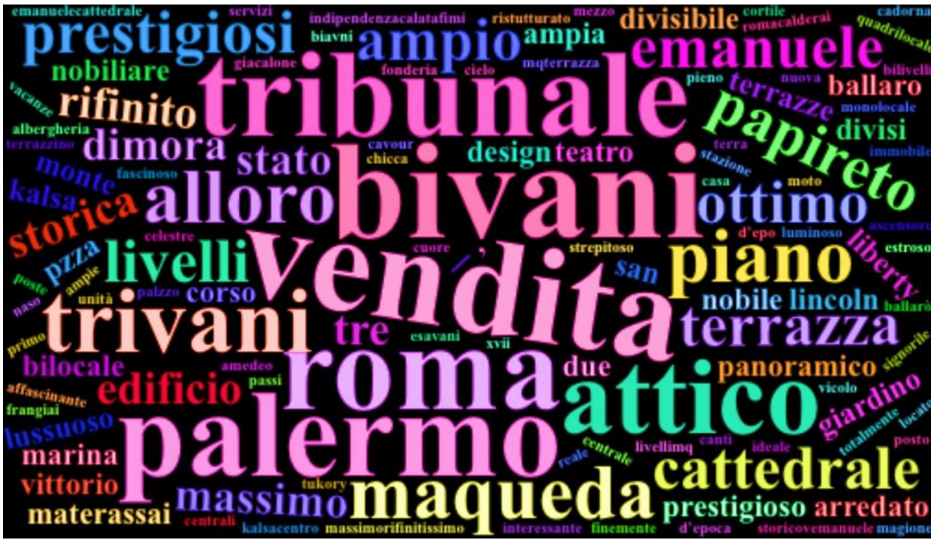


Image 4 wordcloud apartments df

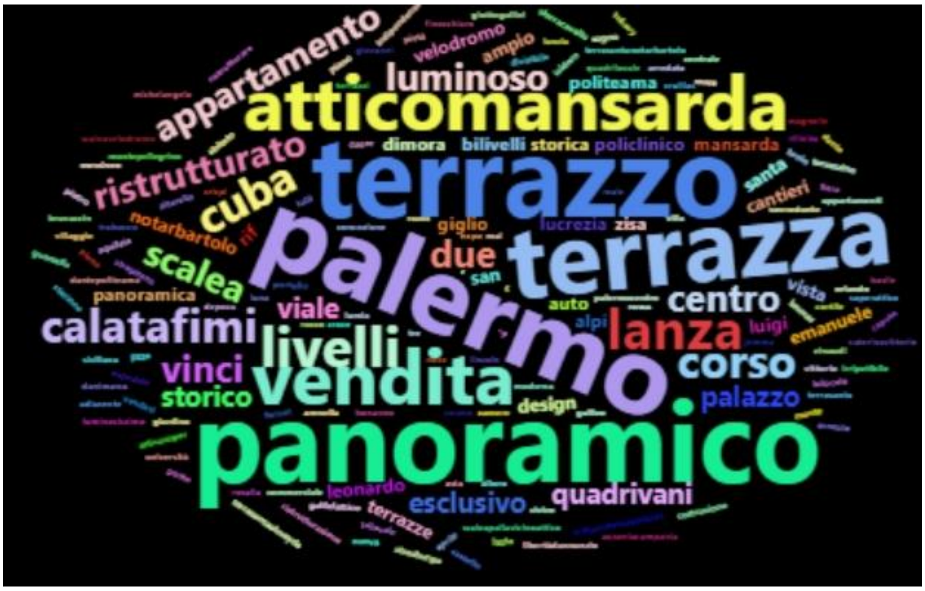


Image 5 attics df wordcloud

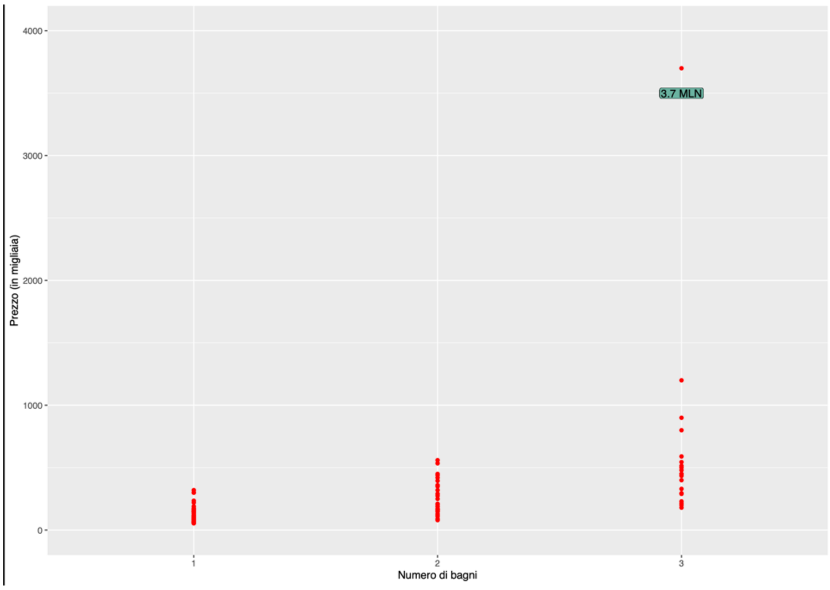


Image 6 apartments df price conditioned by the number of bathrooms

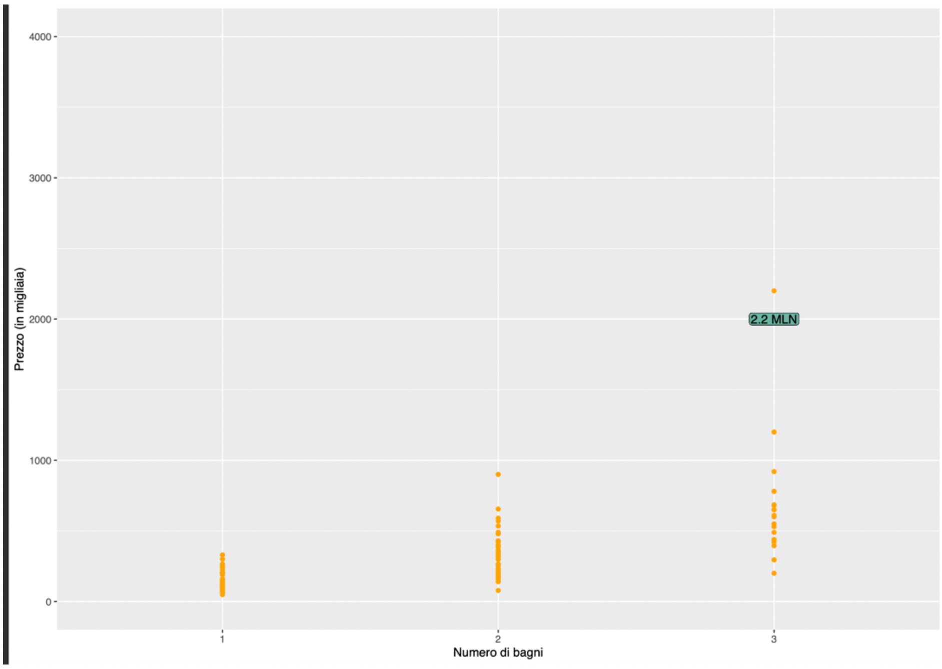


Image 7 attics df price conditioned by the number of bathrooms

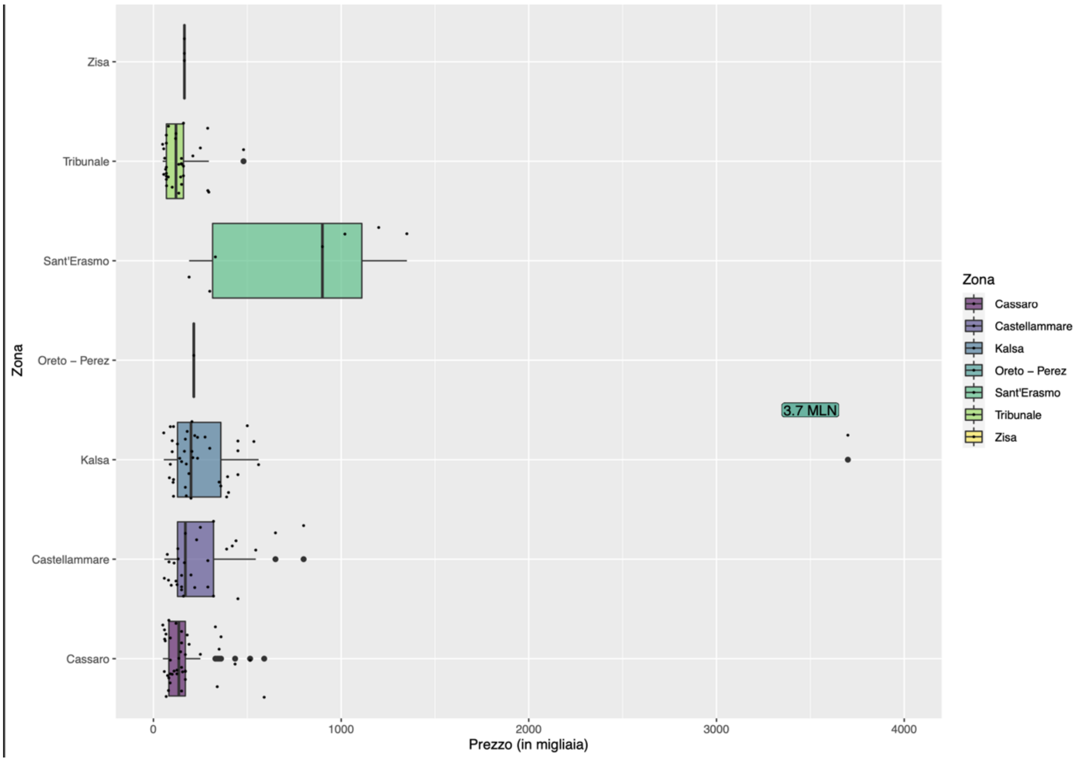


Image 8 apartments df price conditioned by neighborhood

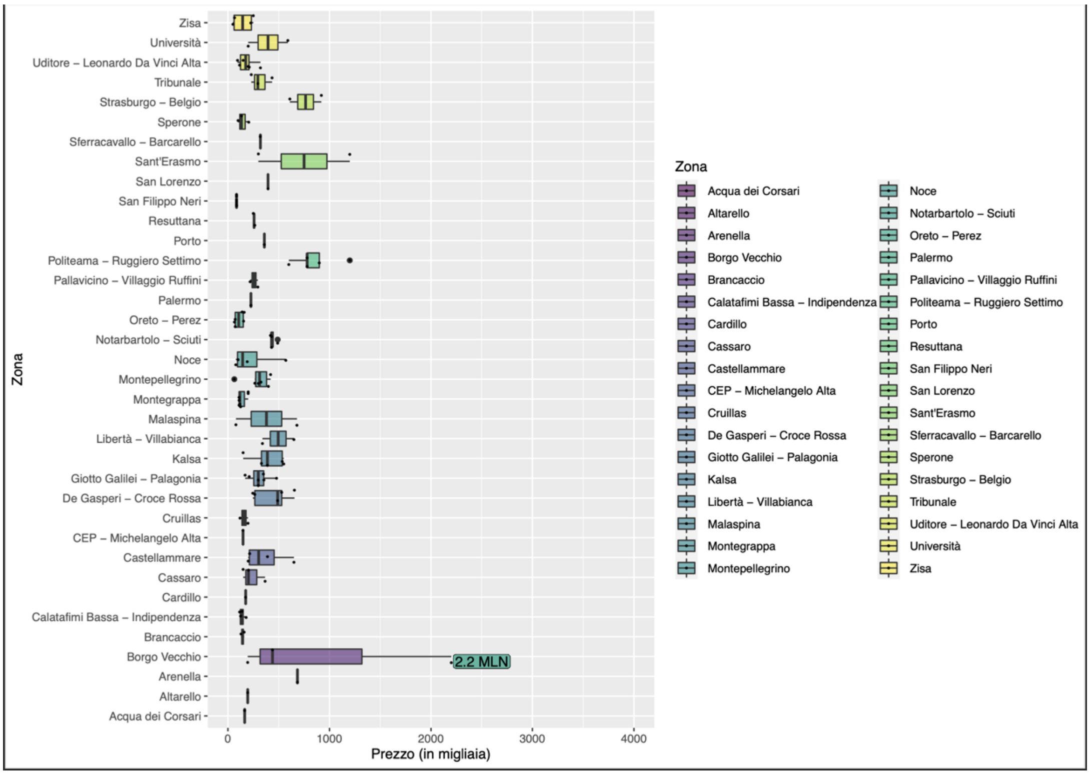


Image 9 attics df price conditioned by neighborhood with outlier

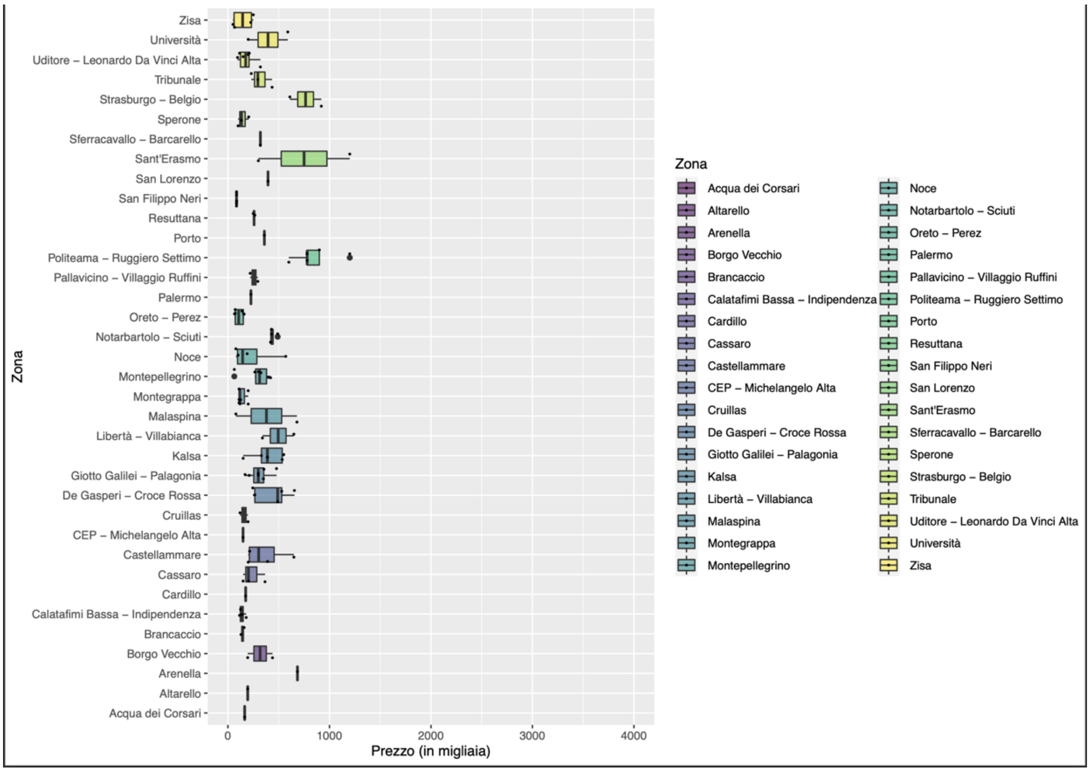


Image 10 attics df price conditioned by neighborhood without outlier