IBM – COURSERA

DATA SCIENCE SPECIALIZATION

Capstone project – Final report

REAL ESTATE ANALYSIS: BEST MILAN NEIGHBORHOODS WHERE TO BUY/RENT YOUR HOUSE

by Federica Gadda,

contact: federica.gadda1@gmail.com

2020

Table of content:

1. INTRODUCTION

Milan is a city located in northern Italy, capital of Lombardy, and the second-most populous city in Italy after Rome. The official estimated population of the City of Milan was 1.4 million as of 01 January 2019, according to ISTAT, the official Italian statistical agency (source).
Milan is considered a leading global city, with strengths in the field of the art, commerce, design, education, entertainment, fashion, finance, healthcare, media, services, research and tourism. The city has been recognized as one of the world's four fashion capitals thanks to several international events and fairs, including Milan Fashion Week and the Milan Furniture Fair, which are currently among the world's biggest in terms of revenue, visitors and growth. It hosts numerous cultural institutions, academies and universities.
Whereas Rome is Italy's political capital, Milan is the country's industrial and financial heart. In 2019 GDP per-capita of Milan is estimated at €49.000, steadily increasing, and significantly higher than the Italian average of €26.000 (source).
Milan is the destination of 11 million visitors in 2019 (as reported in the city website (source), attracted by its museums and art galleries, that include some of the most important collections in the world, like the major works by Leonardo da Vinci. The city is served by many luxury hotels and dreamy restaurants.
Last but not least, Milan will host the 2026 Winter Olympics together with Cortina d'Ampezzo. In short, if you are looking for a city to live in, with events of every kind, where you will divinely eat and where you will never be bored, Milan is right for you!

The goal of this project is to help people who want to buy or to rent house in Milan, finding the characteristics of each neighborhood in terms of house prices and relevant venues in the surrounding area (like restaurants, gyms, parks...). By using data science methods and machine learning method, like clustering, this project will answer to the following question: if you want to move to Milan, what is the better neighborhood for you, according to your financial resources and your interests?

2. DATA

The data for this project has been retrieved from multiple sources, paying the utmost attention to the reliability of them. For this reason, the data was collected from:

1. **Milan borough dataset** and **house market and rental values dataset**: 2 csv filed, retrieved from the Italian Revenue Agency website (source), where the Milan borough list and the information about the market values and the rental values of the houses have been found, related to the 2nd half of 2019, depending on the house location and the state of the property.
   In order to access to these files, it's necessary to register to the website.
2. **Geo-locational information of Milan city center and the neighborhoods**: thanks to Google Maps Geocoding API, it has been possible to retrieve the geo-locational information (latitude and longitude) of Milan city center and the neighborhoods.
3. **Surrounding venues for each neighborhood**: obtained using FourSquare API platform. It has been set the limit of 100 venues and the radius of 5 Km.

These data have allowed to explore and achieve to the goal of this project. The neighborhood data has made possible to determine the value of the house, on the basis of the borough position and the state of the property. Neighborhoods locations have been fundamental understand the correlation between the neighborhood positions (in terms of distance from the Milan city center) and the value of the houses. These positions, together with venues data, have been essential to determinate the clusters and identify the most common venues for each of them.

3. METHODOLOGY

This section will illustrate the overall methodology that has been involved in this project.

The methodology will include:

- data retrieval, cleaning and exploration.
- Visualization of the neighborhoods' location, thanks by Folium library.
- Understand the correlation between market and rental values and the condition and position of the houses; through bar-plots, box-plots and regression.
- Determine the most common venues for each cluster. It has been made by performing K-means clustering algorithm to segment neighborhoods, based on the frequencies of the venues.

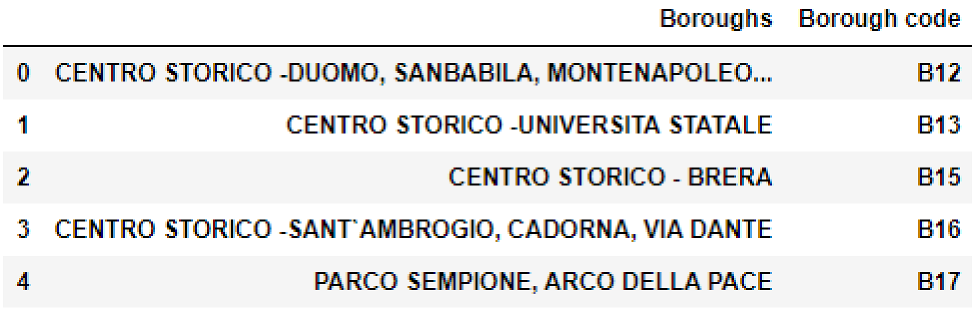
3.1 DATA RETRIEVAL, CLEANING AND EXPLORATION

The first and important step in data science is the data retrieval; indeed, there aren't reliable and precise analysis without using the best data and the most appropriate technique and algorithms.

This analysis starts with the data collection and cleaning, in order to get all the essential data to achieve the goal of this study.

3.1.1 MILAN BOROUGH DATASET AND HOUSE MARKET AND RENTAL VALUES DATASET

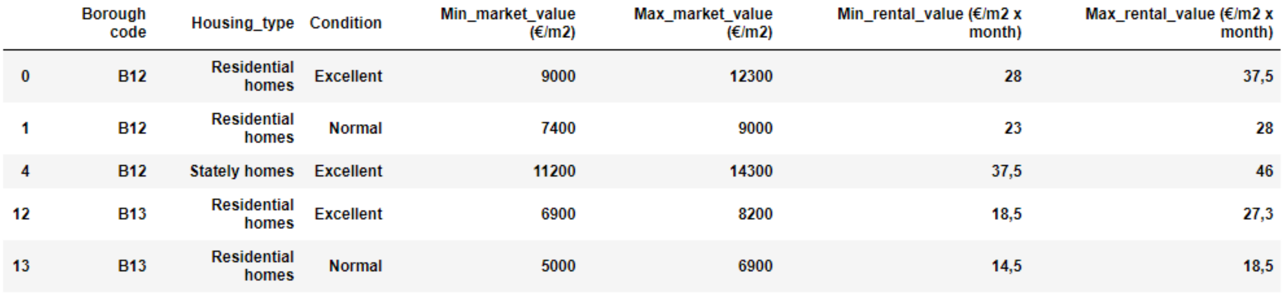
In first hand, the data was cleaned up to get, in the first case, the Milan boroughs names and the related borough codes (fig. 1).



	Boroughs	Borough code
0	CENTRO STORICO -DUOMO, SANBABILA, MONTENAPOLEO...	B12
1	CENTRO STORICO -UNIVERSITA STATALE	B13
2	CENTRO STORICO - BRERA	B15
3	CENTRO STORICO -SANT`AMBROGIO, CADORNA, VIA DANTE	B16
4	PARCO SEMPIONE, ARCO DELLA PACE	B17

Figure 1 - Milan borough and related zone.

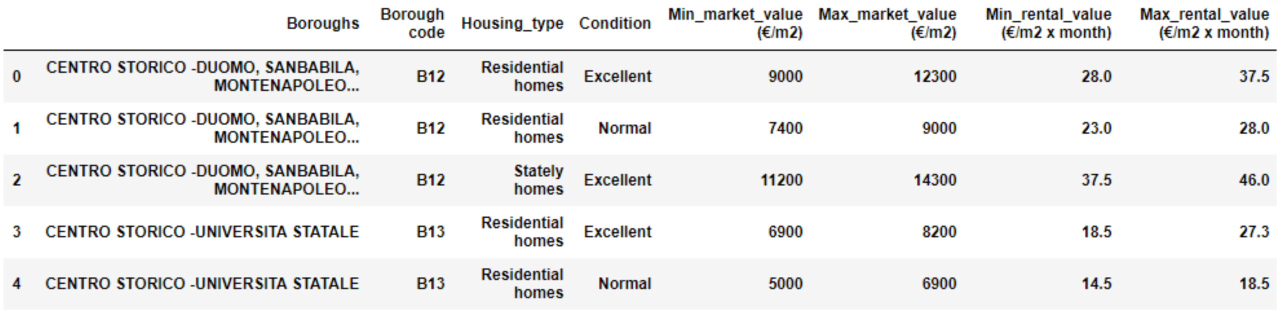
and in second case, for each borough, the minimum and maximum market and rental values of the houses (respectively in terms of €/m2 and €/m2 x moth), depending on the condition of the property (fig. 2).



	Borough code	Housing_type	Condition	Min_market_value (€/m2)	Max_market_value (€/m2)	Min_rental_value (€/m2 x month)	Max_rental_value (€/m2 x month)
0	B12	Residential homes	Excellent	9000	12300	28	37,5
1	B12	Residential homes	Normal	7400	9000	23	28
4	B12	Stately homes	Excellent	11200	14300	37,5	46
12	B13	Residential homes	Excellent	6900	8200	18,5	27,3
13	B13	Residential homes	Normal	5000	6900	14,5	18,5

Figure 2 - Market and rental values of Milan houses.

Then, the datasets were merged to obtain the market and rental values for each Milan borough (fig. 3).



	Boroughs	Borough code	Housing_type	Condition	Min_market_value (€/m2)	Max_market_value (€/m2)	Min_rental_value (€/m2 x month)	Max_rental_value (€/m2 x month)
0	CENTRO STORICO -DUOMO, SANBABILA, MONTENAPOLEO...	B12	Residential homes	Excellent	9000	12300	28.0	37.5
1	CENTRO STORICO -DUOMO, SANBABILA, MONTENAPOLEO...	B12	Residential homes	Normal	7400	9000	23.0	28.0
2	CENTRO STORICO -DUOMO, SANBABILA, MONTENAPOLEO...	B12	Stately homes	Excellent	11200	14300	37.5	46.0
3	CENTRO STORICO -UNIVERSITA STATALE	B13	Residential homes	Excellent	6900	8200	18.5	27.3
4	CENTRO STORICO -UNIVERSITA STATALE	B13	Residential homes	Normal	5000	6900	14.5	18.5

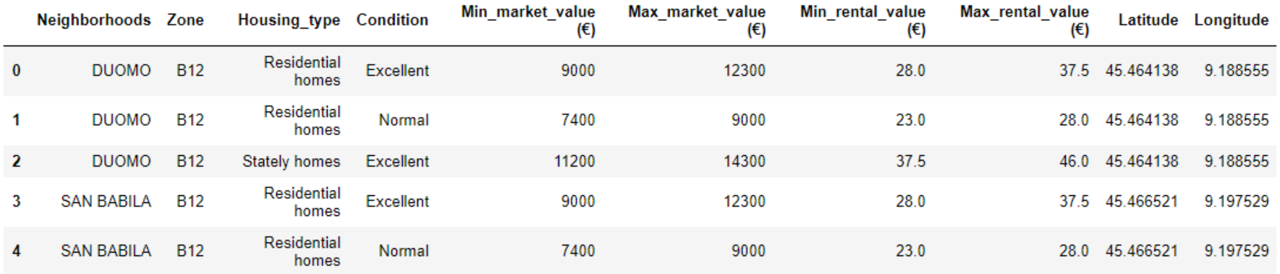
Figure 3 - Market and rental values for each Milan borough.

3.1.2 GEO-LOCATIONAL INFORMATION OF THE NEIGHBORHOOD

First of all, it has been observed that each borough is composed by many neighborhoods, divided by comma. To retrieve the latitude and longitude of each neighborhoods, each borough has been split into the neighborhood of which is composed.

Then, it has been necessary to modify some neighborhood names, in order to make the name recognizable to the Google Maps Geocoding AP.

Finally, all data retrieved by the API has been added to the neighborhood market and rental values data frame (fig. 4).



	Neighborhoods	Zone	Housing_type	Condition	Min_market_value (€)	Max_market_value (€)	Min_rental_value (€)	Max_rental_value (€)	Latitude	Longitude
0	DUOMO	B12	Residential homes	Excellent	9000	12300	28.0	37.5	45.464138	9.188555
1	DUOMO	B12	Residential homes	Normal	7400	9000	23.0	28.0	45.464138	9.188555
2	DUOMO	B12	Stately homes	Excellent	11200	14300	37.5	46.0	45.464138	9.188555
3	SAN BABILA	B12	Residential homes	Excellent	9000	12300	28.0	37.5	45.466521	9.197529
4	SAN BABILA	B12	Residential homes	Normal	7400	9000	23.0	28.0	45.466521	9.197529

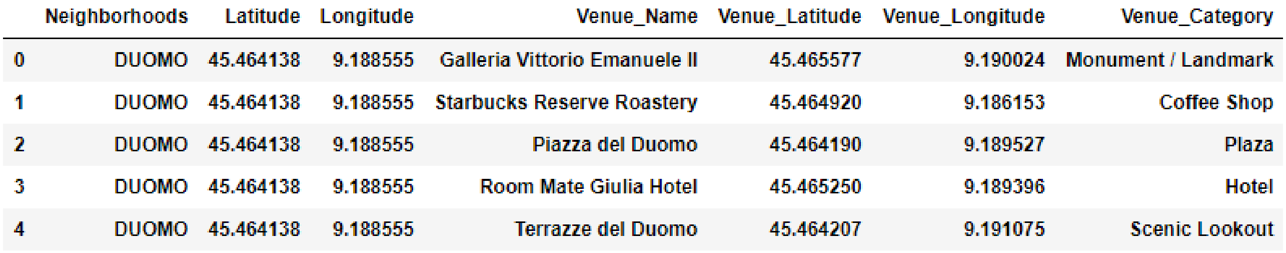
Figure 4 - Neighborhood market and rental values table, with geo-locational information.

3.1.3 SURROUNDING VENUES OF EACH NEIGHBORHOOD

It has been utilized FourSquare API platform to explore each Milan neighborhood, setting the limit of 100 venues and the radius of 5 Km.

The resulting dataset is a 2-dimension data frame (fig. 5), where:

- Each row represents a neighborhood
- Each column represents a venue



	Neighborhoods	Latitude	Longitude	Venue_Name	Venue_Latitude	Venue_Longitude	Venue_Category
0	DUOMO	45.464138	9.188555	Galleria Vittorio Emanuele II	45.465577	9.190024	Monument / Landmark
1	DUOMO	45.464138	9.188555	Starbucks Reserve Roastery	45.464920	9.186153	Coffee Shop
2	DUOMO	45.464138	9.188555	Piazza del Duomo	45.464190	9.189527	Plaza
3	DUOMO	45.464138	9.188555	Room Mate Giulia Hotel	45.465250	9.189396	Hotel
4	DUOMO	45.464138	9.188555	Terrazze del Duomo	45.464207	9.191075	Scenic Lookout

Figure 5 – Venues Dataset.

3.2 MAP OF MILAN NEIGHBORHOODS

To get a sense to the study, it is of primary importance to know the precise location of each neighborhoods. For this reason, it's essential to create a map of Milan, in which all the neighborhood positions are shown. To be more exhaustive, all the Milan areas (B, C, D and E) are differentiated by different colors.

Therefore, the map of Milan neighborhoods has been plotted using the Folium library, that makes it easy to visualize data, by manipulating in Python on an interactive leaflet map (fig. 6).

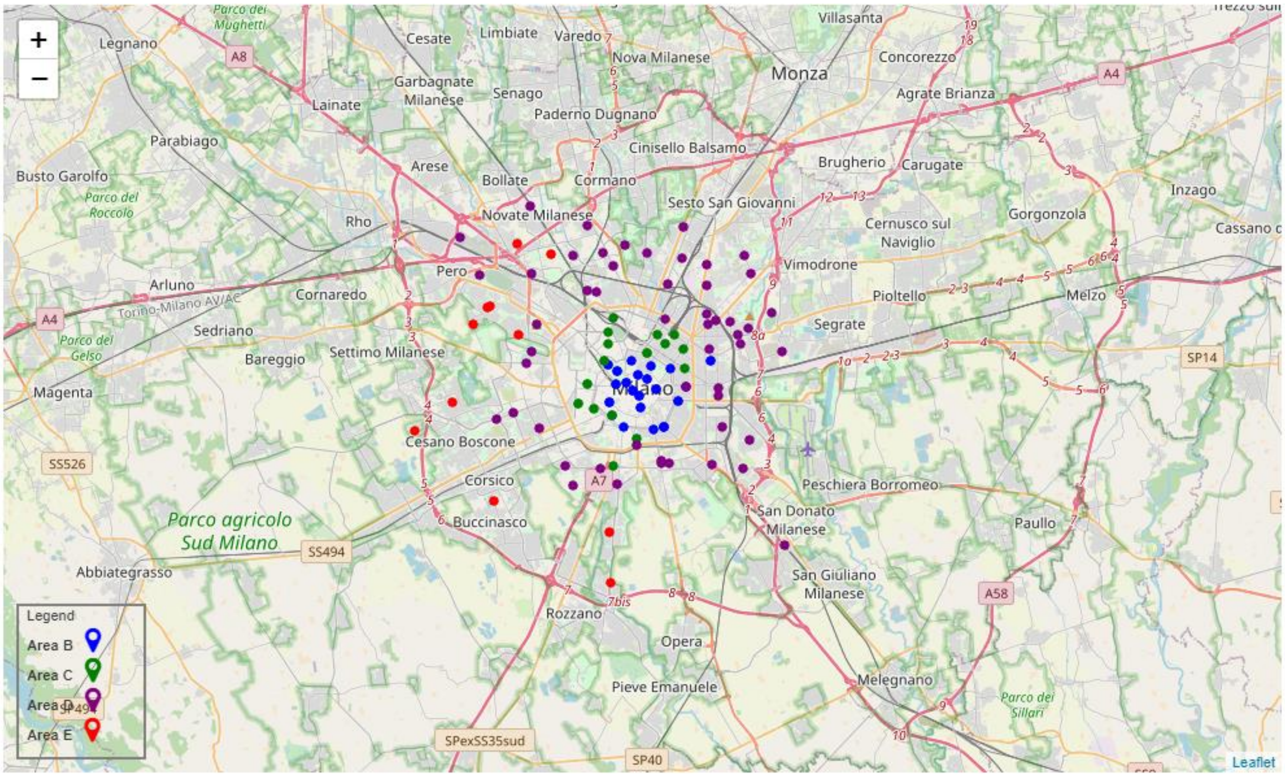


Figure 6 - Milan neighborhoods map.

3.3 COMPARISON OF THE MARKET AND THE RENTAL VALUES ACROSS EACH MILAN BOROUGH AND THE HOUSING TYPE/CONDITION

Bar charts have been created to determine the house price behaviors of each Milan borough, depending also on the housing type (residential or stately house) and the condition of residential homes (excellent or normal).

Since the table published on the Italian Revenue Agency website only shows the prices for each borough, it has been decided to work with them, assuming that the house prices of the neighborhoods located in the same zone don't considerably change.

First of all, it has been calculated the average of the maximum and minimum market and rental values. Then, multiple tables have been created, for each house type and the related condition. The tables that have been created are the following ones:

- market values of residential homes in excellent condition;
- rental values of residential homes in excellent condition;
- market values of residential homes in normal condition;
- rental values of residential homes in normal condition;
- market values of stately homes;
- rental values of stately homes.

Finally, as mentioned above, to compare the average prices, bar charts have been generated. This type of chart, that presents categorical data, consists in rectangular bars with heights proportional to the values that they represent. It is the better way to compare the prices for each borough visually.

3.4 DETERMINATION OF THE CORRELATION BETWEEN THE DISTANCE OF THE HOUSE FROM THE CITY CENTER AND AVERAGE HOME VALUE

Bar plots have shown that the boroughs closer to the city center are more expensive; on the contrary, the ones farthest to the city center are the most affordable.
To deepen into this discovery, it has been decided to extend the boundaries taken into account, examining the house prices for all the Milan areas (B, C, D and E).
Multiple box plots have been created, for the average market and rental values of each housing type and the related condition, considering also the Milan areas.
Therefore, regression have been generated, to estimate the relationships between a dependent variable (house prices) and the independent variables (distance to the city center).

3.4.1 BOX PLOTS: AVERAGE MARKET/RENTAL VALUES OF HOUSING TYPES AND THE RELATED CONDITIONS, CONSIDERING THE MILAN AREAS

A boxplot is a standardized way of displaying the dataset based on a five-number summary: the minimum, the maximum, the sample median, and the first and third quartiles. Any data not included between the whiskers should be plotted as an outlier.
In order to create the box plots, the 4 Milan areas have been taken in account and, as bar charts already generated, multiple plots have been created for each housing type and the related condition:

- market values of residential homes in excellent condition;
- rental values of residential homes in excellent condition;
- market values of residential homes in normal condition;
- rental values of residential homes in normal condition
- market values of stately homes;
- rental values of stately homes.

3.4.2 REGRESSION

Box plots have showed that it can be a correlation between the residential house prices and the distance from the city center, indeed it clearly visible that the most expensive residential homes are located in the B and C areas, and the most affordable ones are located in D and E areas. The precedent analysis highlight something different for stately house, indeed the prices of houses in B and C areas are in line with each other.
To clearly see if what supposed above is true, regression analysis has been done. In fact, as mentioned above, regression is the main tool to estimate the relationships between a dependent variable, in this case the house values, and the independent variables, the distance to the city center.

First of all, the distance of each neighborhoods has been calculated: retrieving Milan city center coordinates (Duomo square, where Duomo Cathedral is located) using to Google Maps Geocoding API at first, and then calculating the distances.
Subsequently, for each housing type and the related condition, two type of scatter plots have been created:

- in the first ones, the distance of all the Milan neighborhoods and the house prices are compared. Each neighborhood point has different colors, based on the belonging Milan area (the same colors of the above map of Milan neighborhoods have been chosen).
- In second ones, the curve has been fitted by applying curve_fit, which uses nonlinear least squares to fit the data. In these cases, the functions are all logarithmic, as it is possible to see form the scatter plots previously created.

3.5 VENUES ANALYSIS

After determining the surrounding venues of the Milan neighborhoods, the latter have been divided in clusters, as to grouping them according to the more common venue types, in such a way that neighborhoods in the same group are more similar to each other than to those in other groups. To do that, K-means algorithm has been use.

However, to use this algorithm it's necessary to specify the number of clusters. There are various methods to find the optimal number of clusters; for this project 2 methods have been used:

- <u>elbow method</u>: by plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the best number of clusters to use (fig. 7).

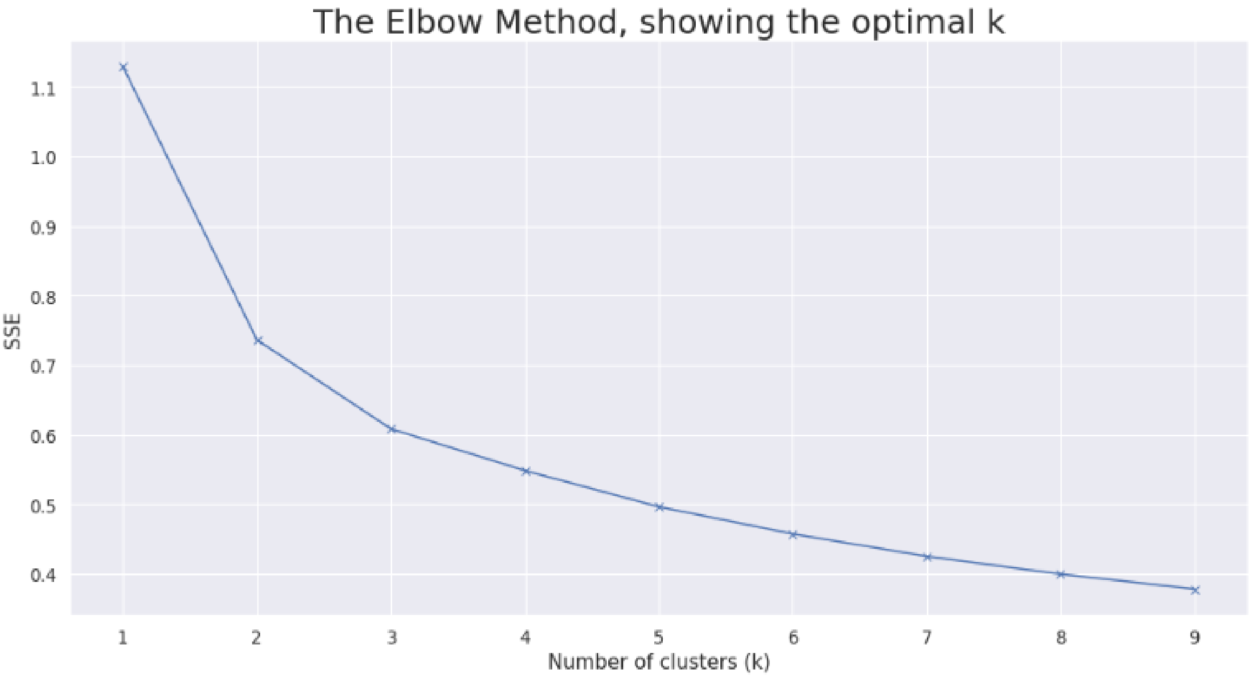
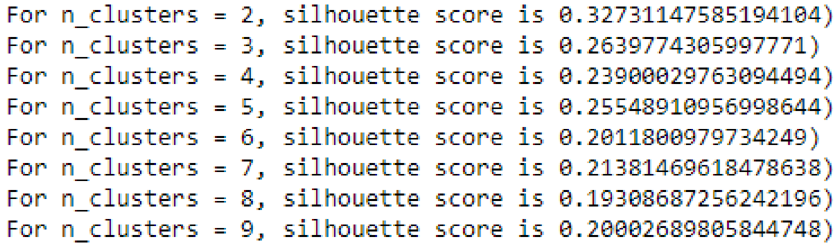


Figure 7 - Elbow method.

- <u>Silhouette method</u>: silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). The silhouette ranges from −1 to 1, where a high value indicates that the object is well matched to its own cluster and poorly matched to neighboring clusters. If most objects have a high value, then the clustering configuration is appropriate (fig. 8).



```
For n_clusters = 2, silhouette score is 0.32731147585194104)
For n_clusters = 3, silhouette score is 0.2639774305997771)
For n_clusters = 4, silhouette score is 0.23900029763094494)
For n_clusters = 5, silhouette score is 0.25548910956998644)
For n_clusters = 6, silhouette score is 0.2011800979734249)
For n_clusters = 7, silhouette score is 0.21381469618478638)
For n_clusters = 8, silhouette score is 0.19308687256242196)
For n_clusters = 9, silhouette score is 0.20002689805844748)
```

Figure 8 - Silhouette method.

Both methods have shown that the optimal number of clusters are 2.

The clustering has been performed using Euclidean distance metric, with the frequency of venue type as a feature.

To determinate the geographical position of the neighborhoods in both clusters, a Milan map has been generated, using Folium library, with different colors assigned basing on the cluster of belonging.

Finally, the top 10 venues for both clusters have been discovered and graphically visualized by horizontal bar charts (which show also the frequency percentages).

4. RESULTS

This section will illustrate the results obtained during the analysis and seeking to answer to the main question of this project: if a person wants to move to Milan, what is the better neighborhood for him, according to his financial resources and interests?

4.1 COMPARISON OF THE MARKET AND THE RENTAL VALUES ACROSS EACH MILAN BOROUGH AND THE HOUSING TYPE/CONDITION

Following the 6 bar plots that compare, for each Milan borough, the market and rental prices of:

- Stately homes (fig. 9).

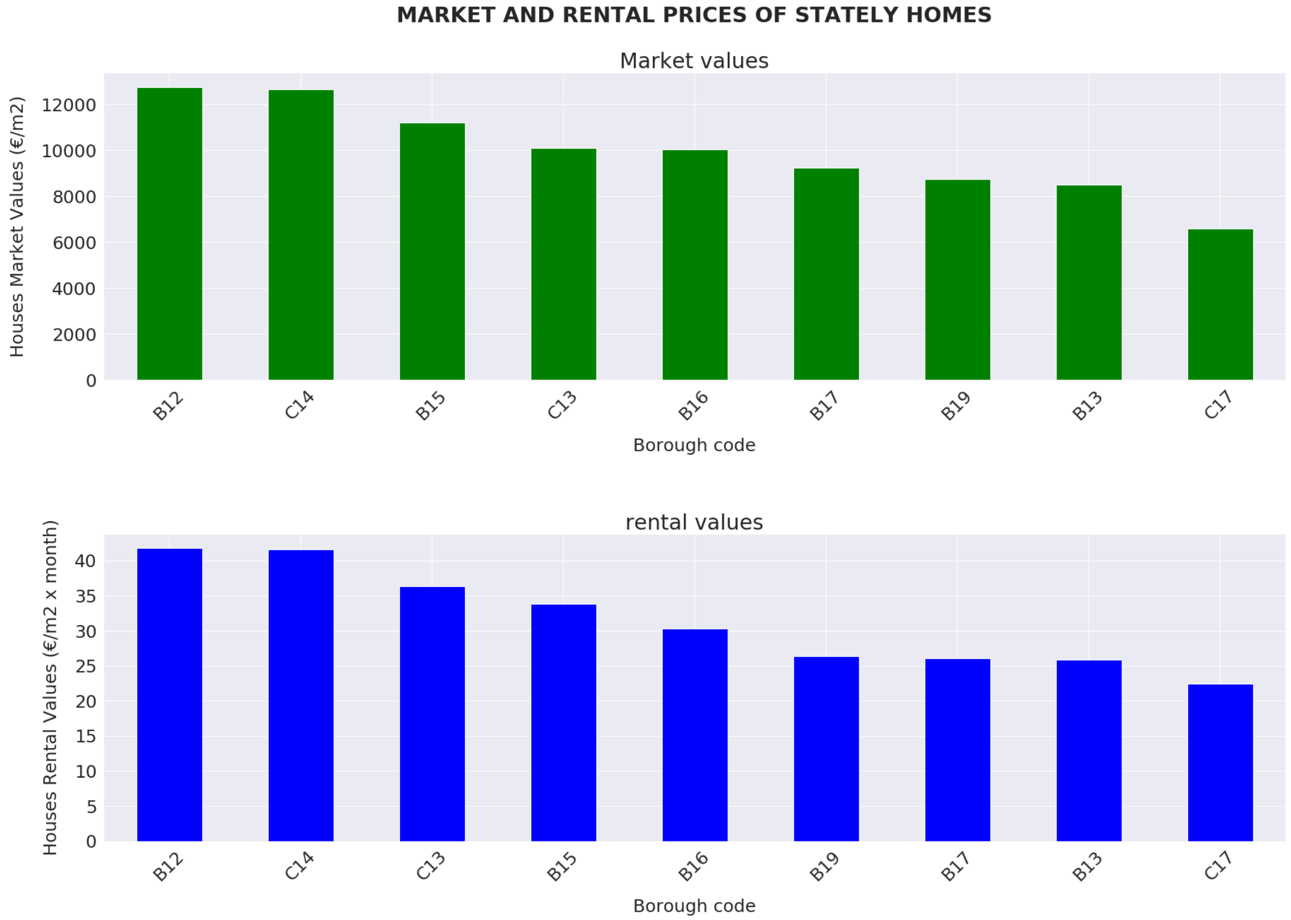


Figure 9 - Values of stately homes.

- residential homes in excellent condition (fig. 10);

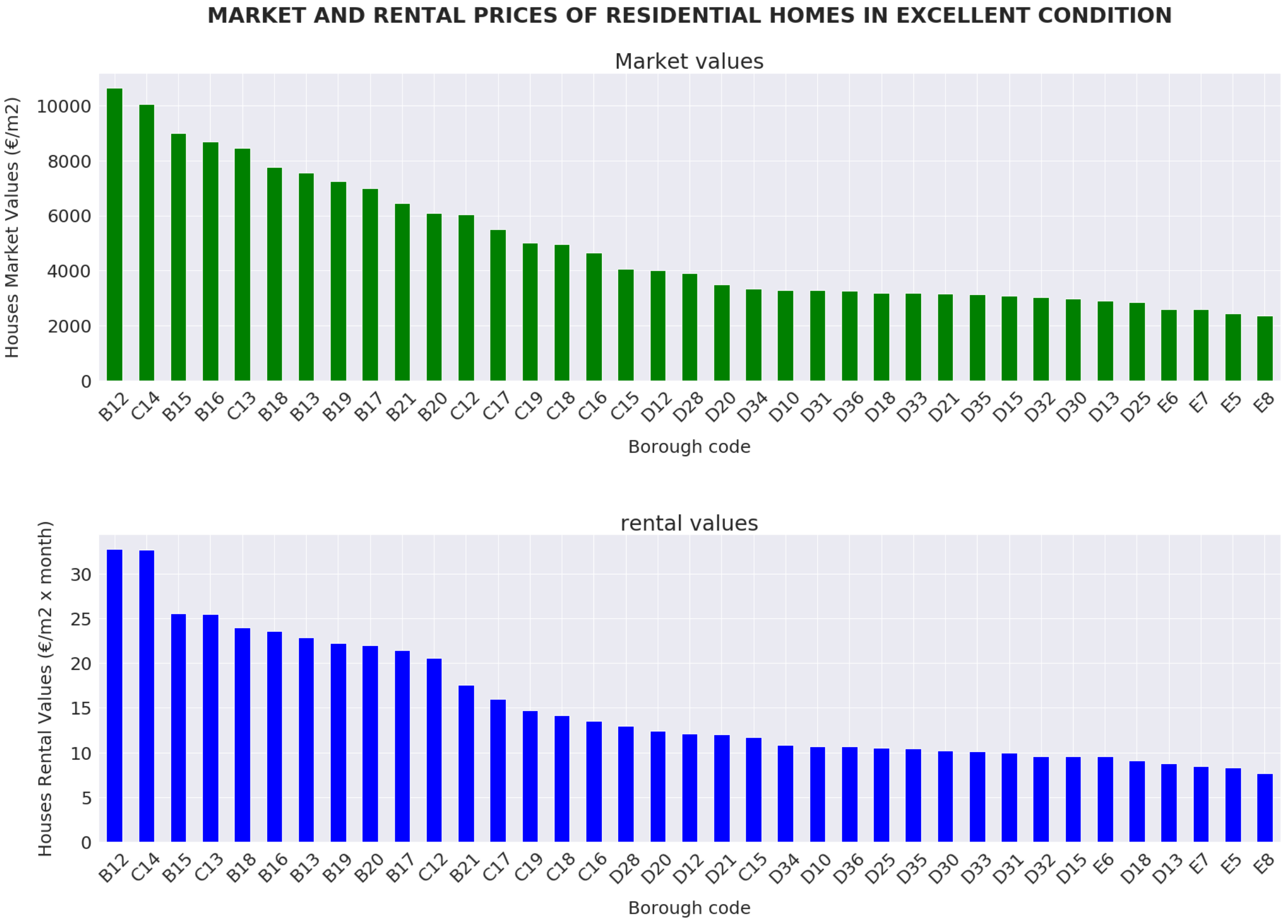


Figure 10 - Values of residential homes in excellent condition.

- residential homes in normal condition (fig. 11).

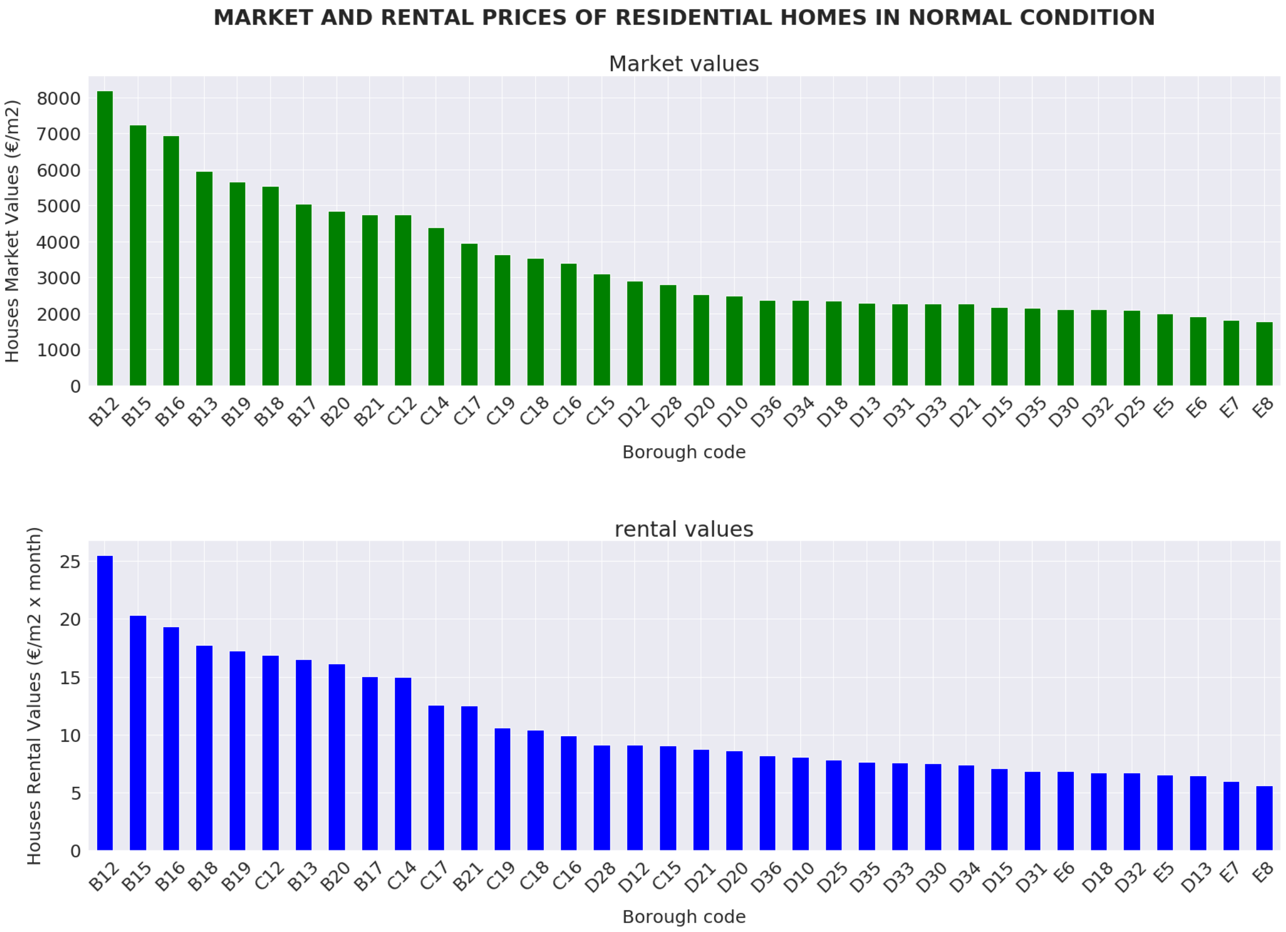


Figure 11 - Values of residential homes in normal condition.

As it can be seen from the results obtained, stately houses are more expensive than residential houses, moreover residential homes in normal condition are cheaper than the ones in excellent condition.

As regards the value of the houses located in the different Milan boroughs, in general the higher market and rental values of residential houses are of the ones located in B area, a little less in C area, than in D area and finally the houses located in E area are most affordable. This doesn't apply to stately houses, which prices don't change significatively between B and C areas.

In addition, market and rental values are quite aligned between the same house types in the same Milan area.

Broadly speaking, houses nearest to the city center are the most expensive, as they moving away from the center the prices decrease. However, this is not a general rule, for example the values of the residential houses in excellent condition located in the C13 and C14 boroughs are more expensive than some houses present in the B area; it is supposed that the reason of these exceptions are due to the borough themselves, since Citylife and Porta Nuova boroughs have been the subject of a vast urban regeneration project. This phenomenon can be seen in the stately houses chart, in fact C13 and C14 house values are higher than, for example, B15 and B16 houses.

4.2 DETERMINATION OF THE CORRELATION BETWEEN THE DISTANCE OF THE HOUSE FROM THE CITY CENTER AND AVERAGE HOME VALUE

The results obtained from the above bar plots suggest that the house market and rental values change according to the Milan areas. To confirm this hypothesis, firstly, box plots have been created taking into consideration the values of the houses in each area, then regression method has been applied.

4.2.1 BOX PLOTS: AVERAGE MARKET/RENTAL VALUES OF HOUSING TYPES AND THE RELATED CONDITIONS, CONSIDERING THE MILAN AREAS

Following the 6 box plots, which compare the market and rental values of the houses, depending on the type and state of the property and the Milan area where they are located:

- stately homes (fig. 12);

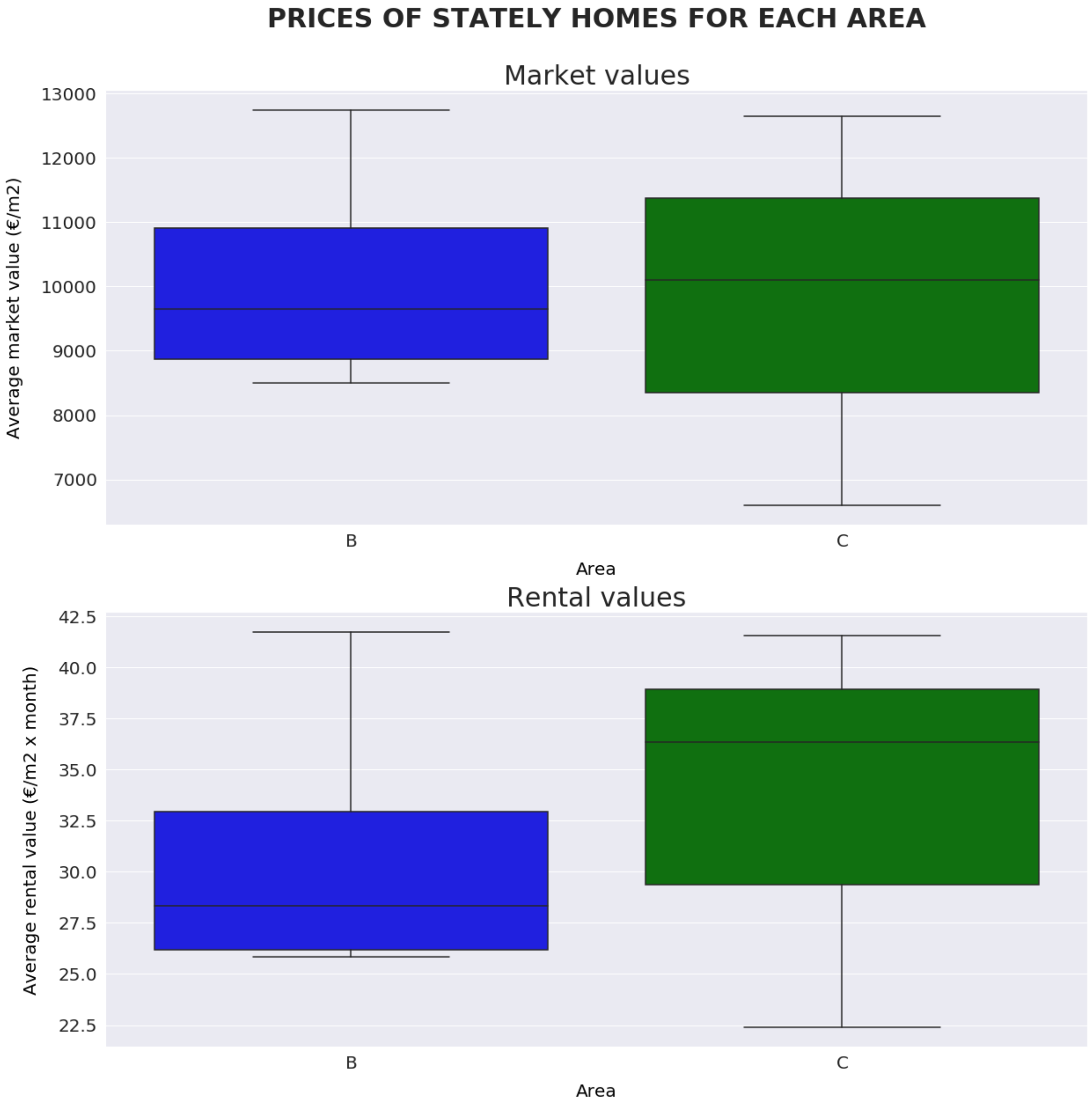


Figure 12 - Stately home values for Milan areas.

- residential homes in excellent condition (fig. 13);

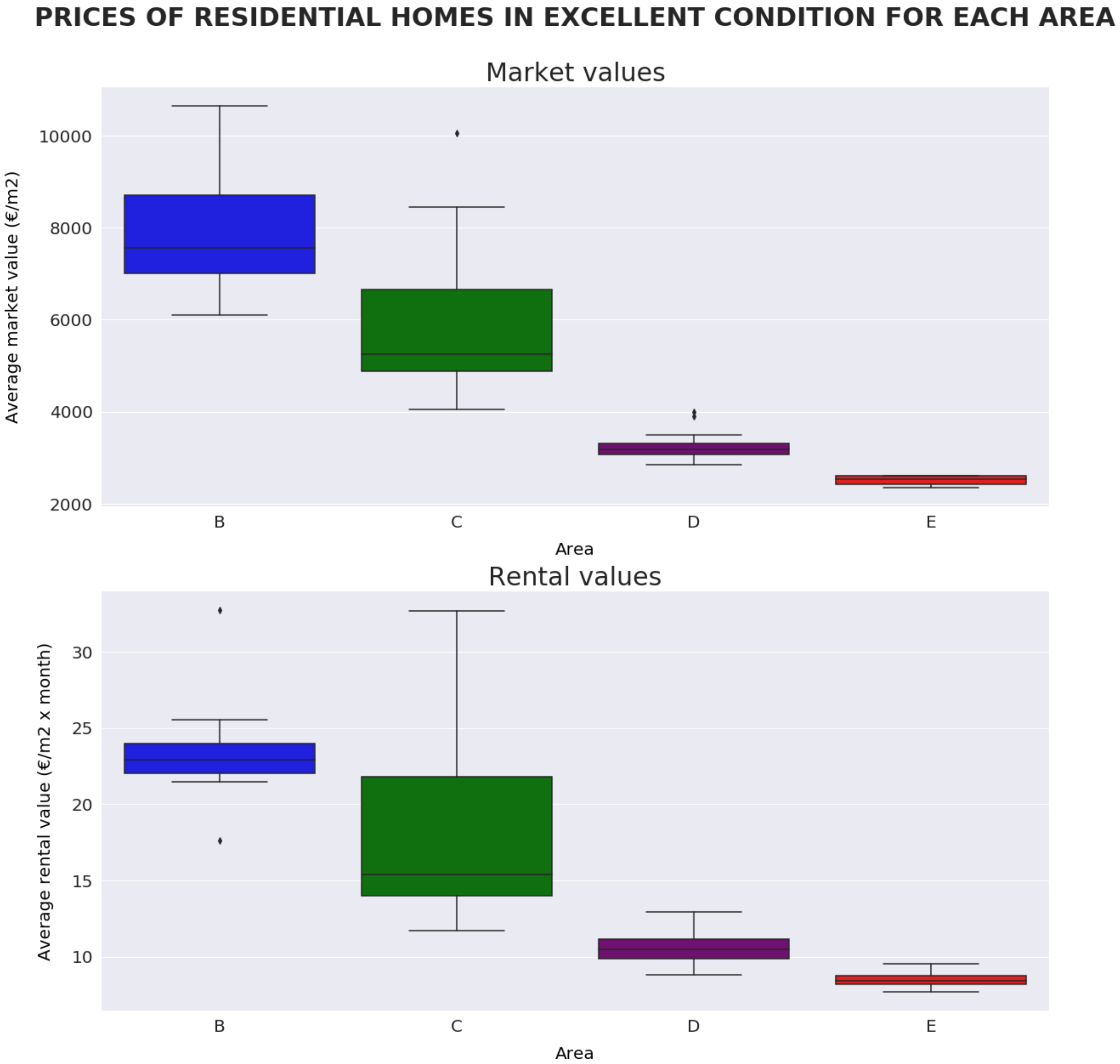


Figure 13 - Residential homes in excellent condition values for Milan areas.

- Residential homes in normal condition (fig. 14):

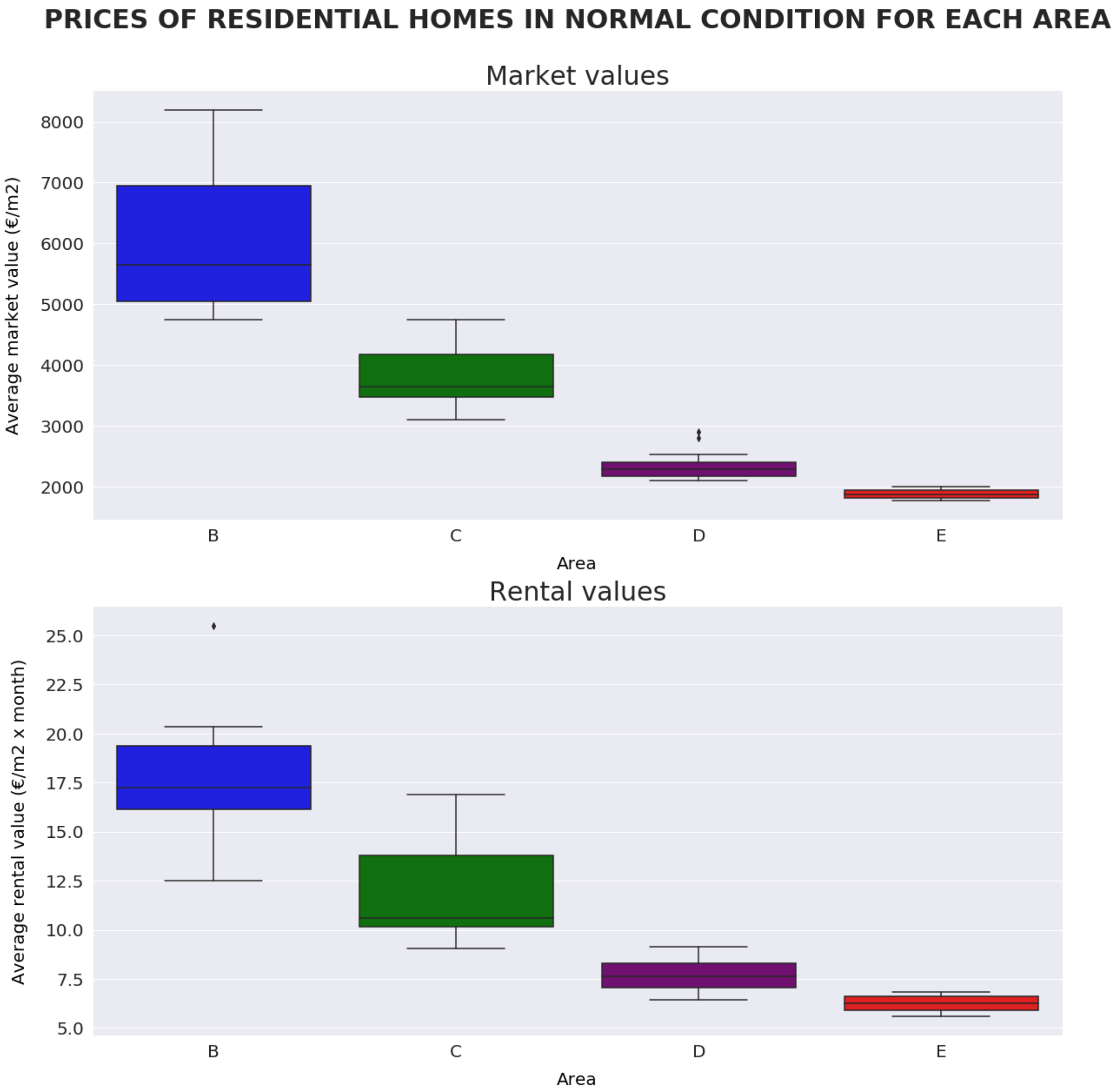


Figure 14 - Residential homes in normal condition values for Milan areas.

The above box plots show a comparable situation to the one observed through the bar plots. Indeed, significant differences between the market and rental values of stately houses are not observed, but rather, the median rental value of the houses in C area is higher than the rental value of the houses in B area. This particular condition could be due to the fact that, for example, the dimension and the impressiveness of the house more affect the values of stately house than the proximity to the city center.

It is possible to observe a different situation for residential homes, indeed there are significant differences between the market and the rental values, depending on the Milan areas. Moreover, it’s interesting to observe that the variation range of values of the residential houses located in the B and C areas is wider that the variation range of the houses located in D and E areas.

Finally, how it was possible to observe in depth through bar plots, residential houses in excellent condition located in C area have some market and rental values comparable to the ones of the house located in B area. The same situation was found for the rental values of the residential homes in normal condition. Anyway, the median values of these 2 areas are very different in all cases.

4.2.2 REGRESSION

Regression analysis is the last step of this correlation study and it was able to confirm or refute this hypothesis.
In this case, to compare the market and rental houses values with the distance of the neighborhoods to the city center, 12 scatter plots have been created:

- 6 where each neighborhood point has different colors, based on the belonging Milan area;
- 6 where the regression curve has been fitted, to best expresses the relationship between the price and the distance to the city center.

Following the resulting charts:

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for market values of stately houses (respectively fig. 15 and 16).

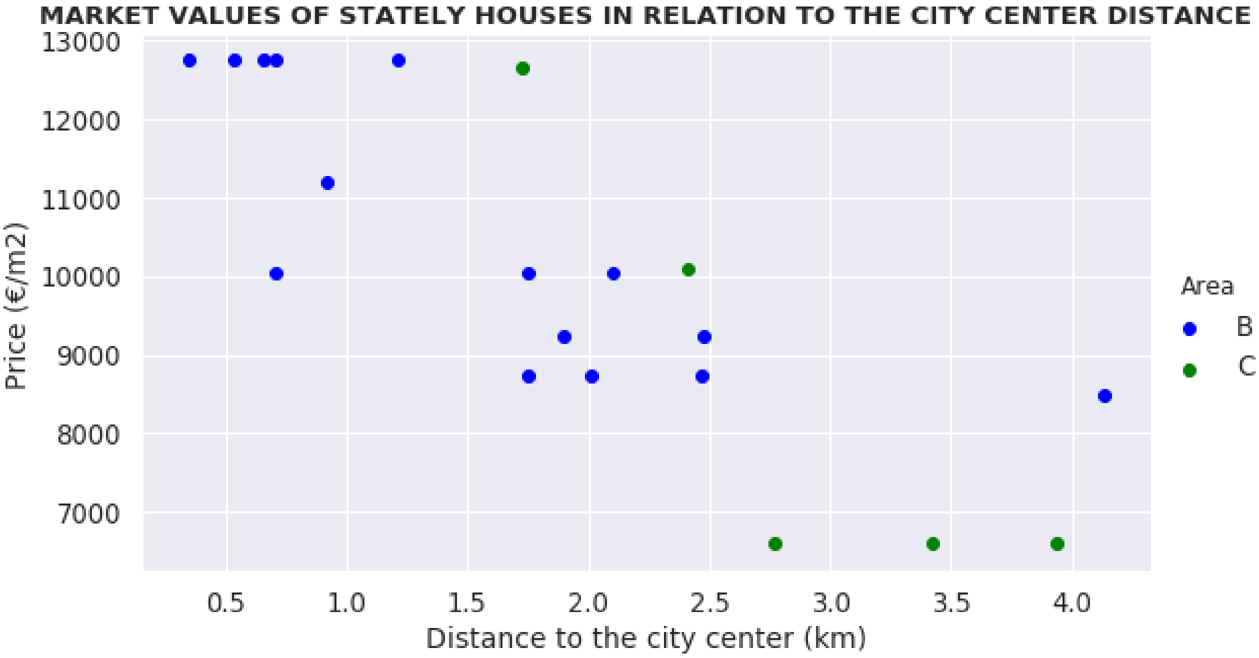


Figure 15 - Distribution of the market values of stately houses, depending on the city center distance and on the area location.

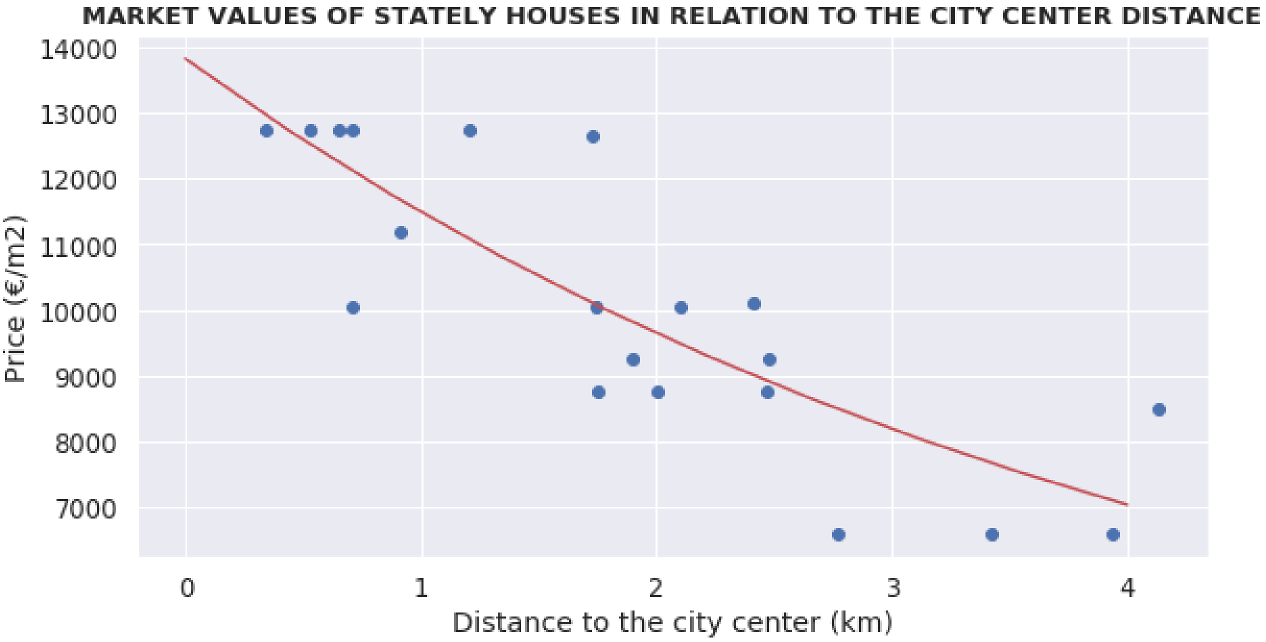


Figure 16 - Trend of the market values of stately houses, depending on the city center distance.

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for rental values of stately houses (respectively fig. 17 and 18).

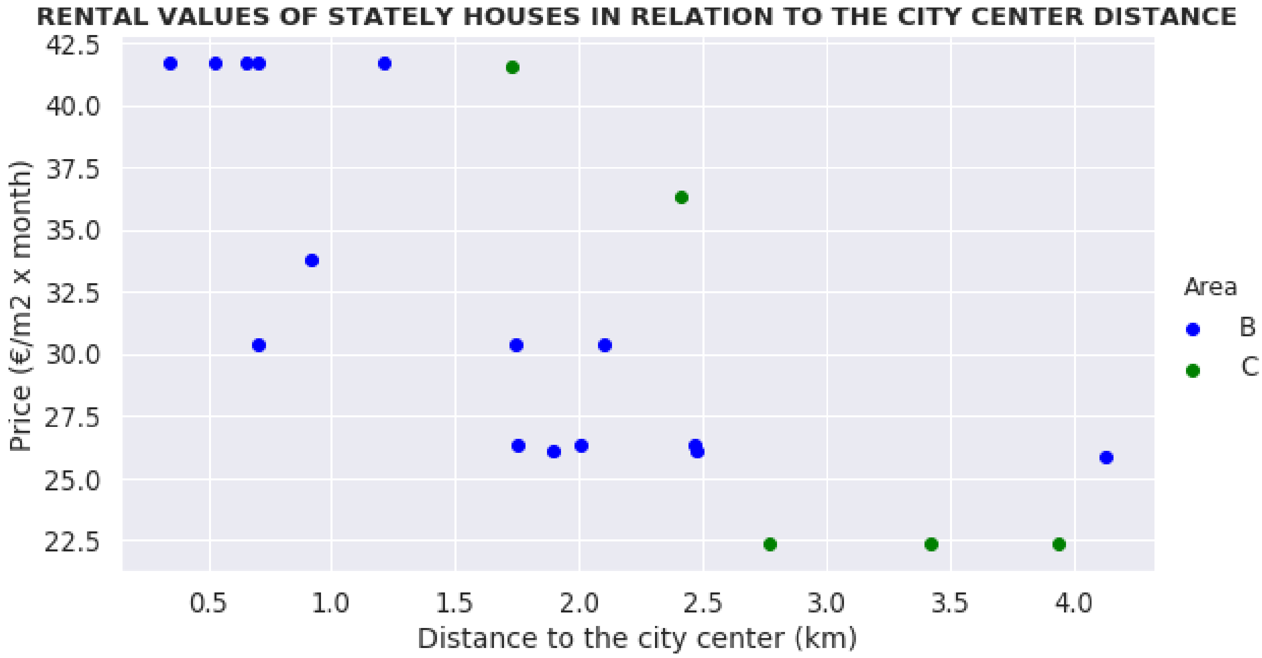


Figure 17 - Distribution of the rental values of stately houses, depending on the city center distance and on the area location.

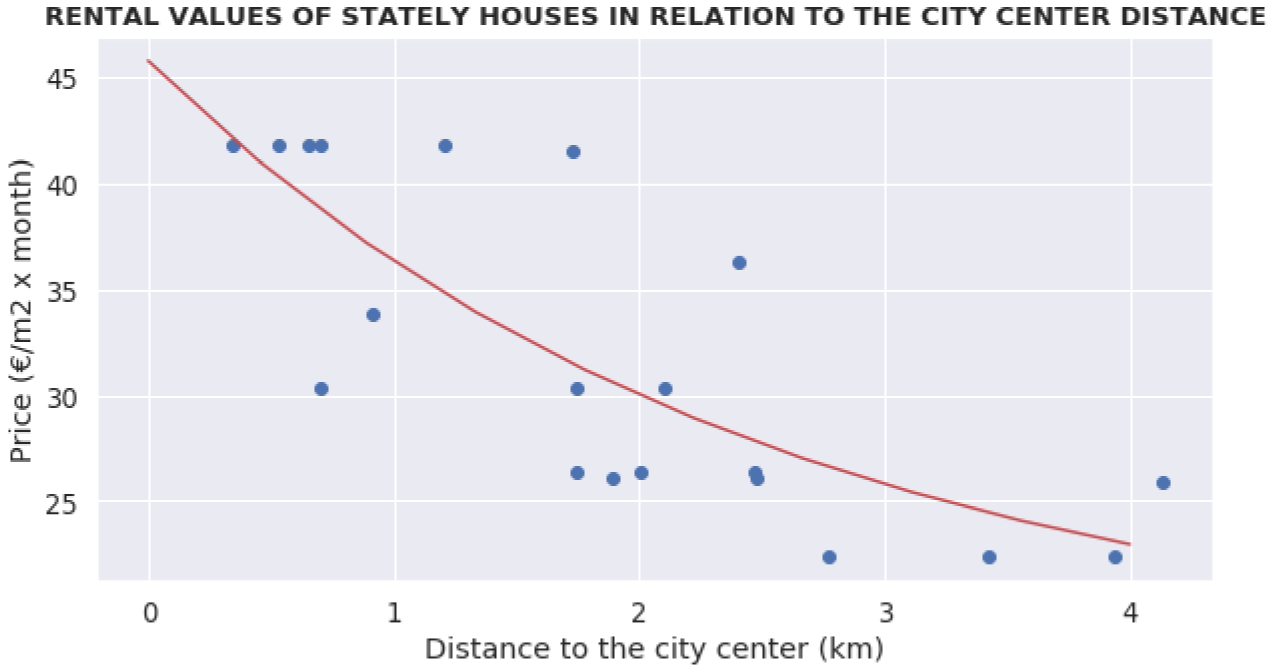


Figure 18 - Trend of the rental values of stately houses, depending on the city center distance.

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for market values of residential houses in excellent condition (respectively fig. 19 and 20).

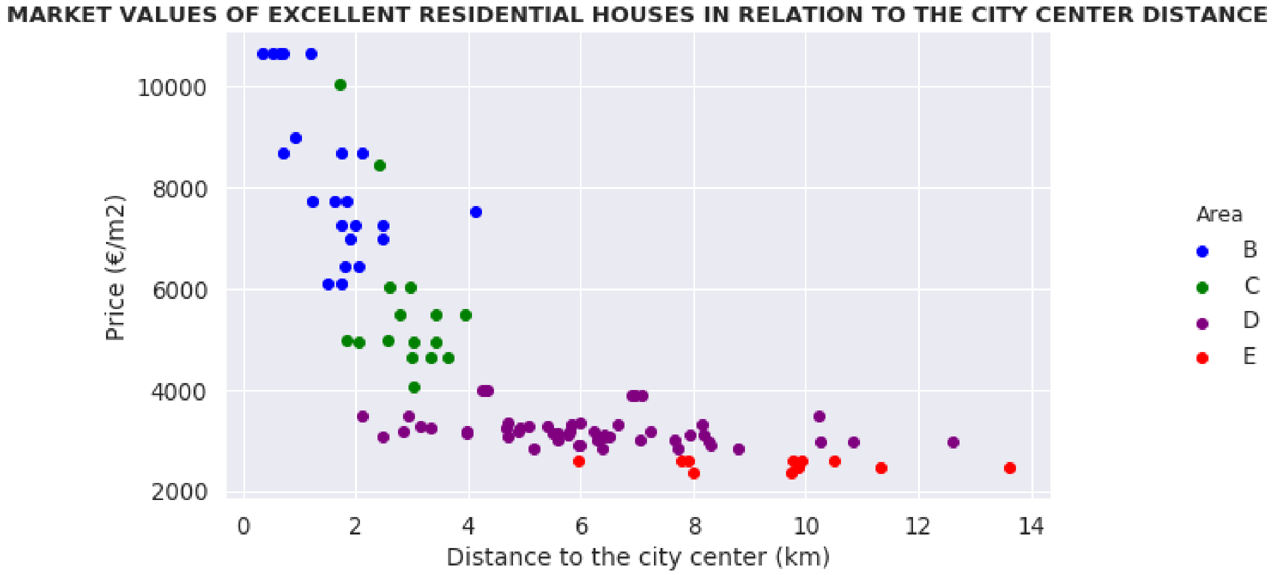


Figure 19 - Distribution of the market values of residential houses in excellent condition, depending on the city center distance and on the area location.

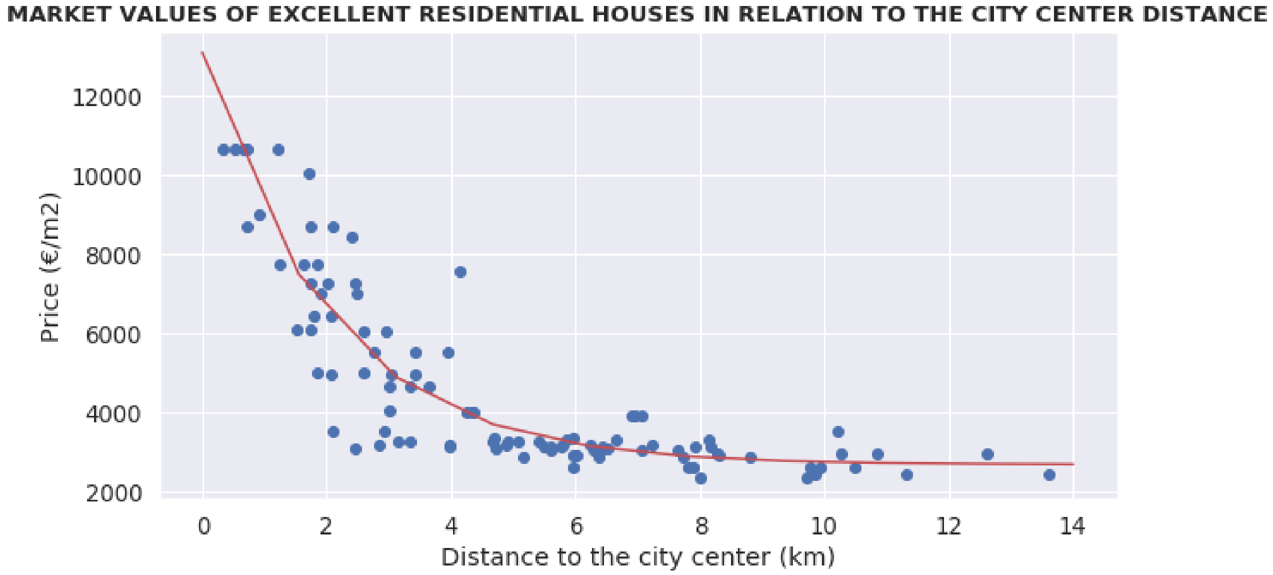


Figure 20 - Trend of the market values of residential houses in excellent condition, depending on the city center distance.

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for rental values of residential houses in excellent condition (respectively fig. 21 and 22).

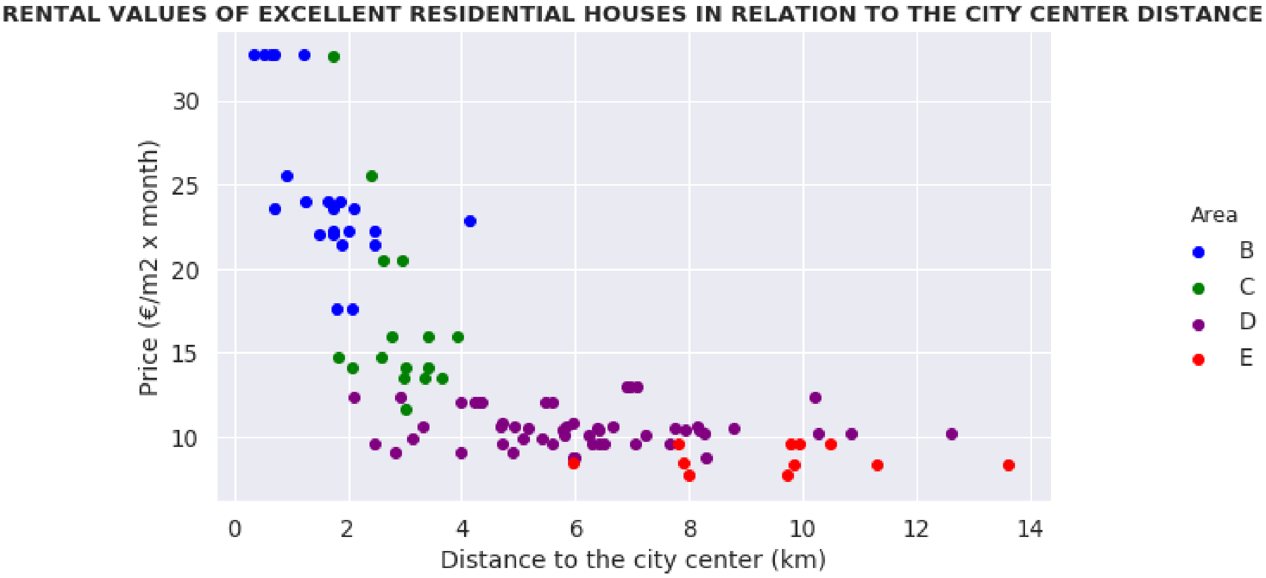


Figure 21 - Distribution of the rental values of residential houses in excellent condition, depending on the city center distance and on the area location.

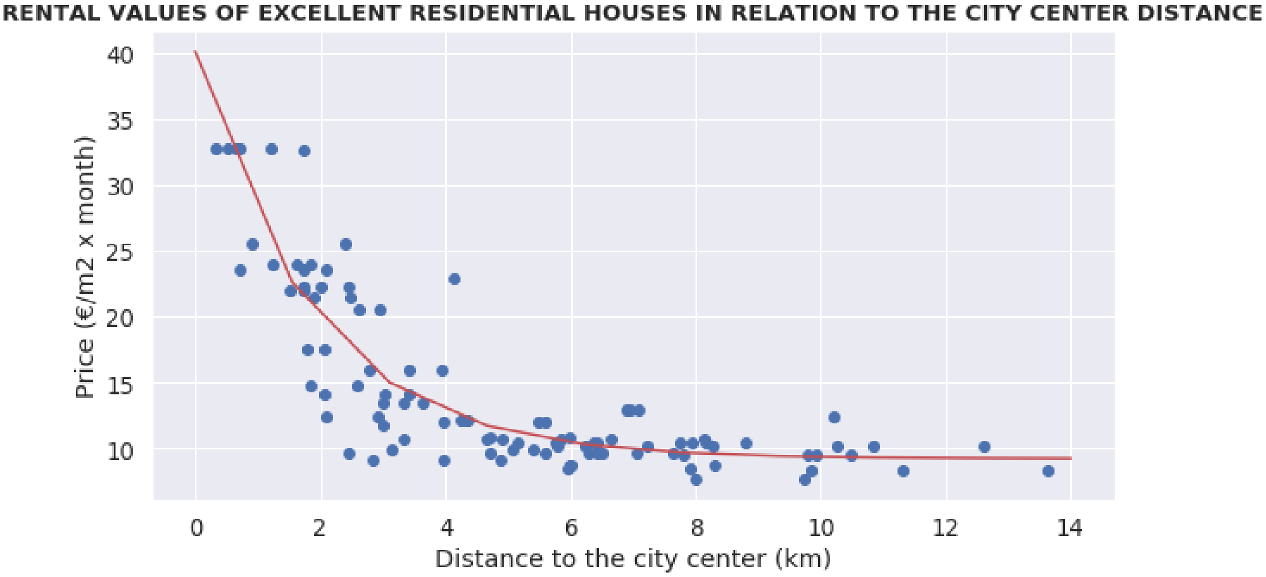


Figure 22 - Trend of the rental values of residential houses in excellent condition, depending on the city center distance.

Federica Gadda

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for market values of residential houses in normal condition (respectively fig. 23 and 24).

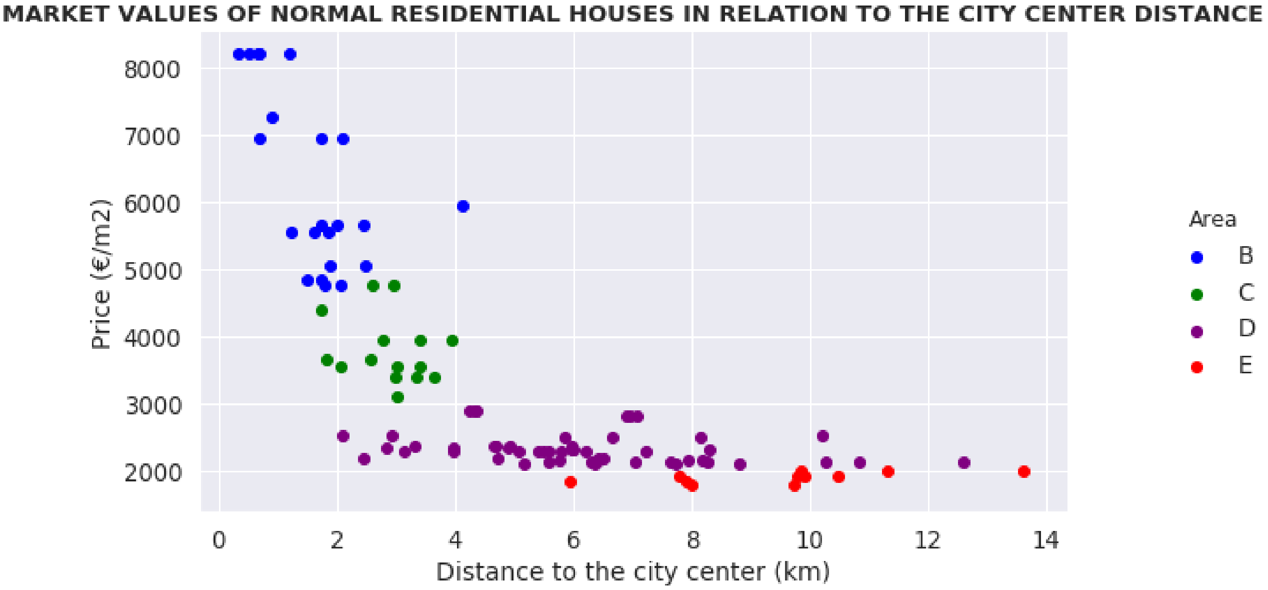


Figure 23 - Distribution of the market values of residential houses in normal condition, depending on the city center distance and on the area location.

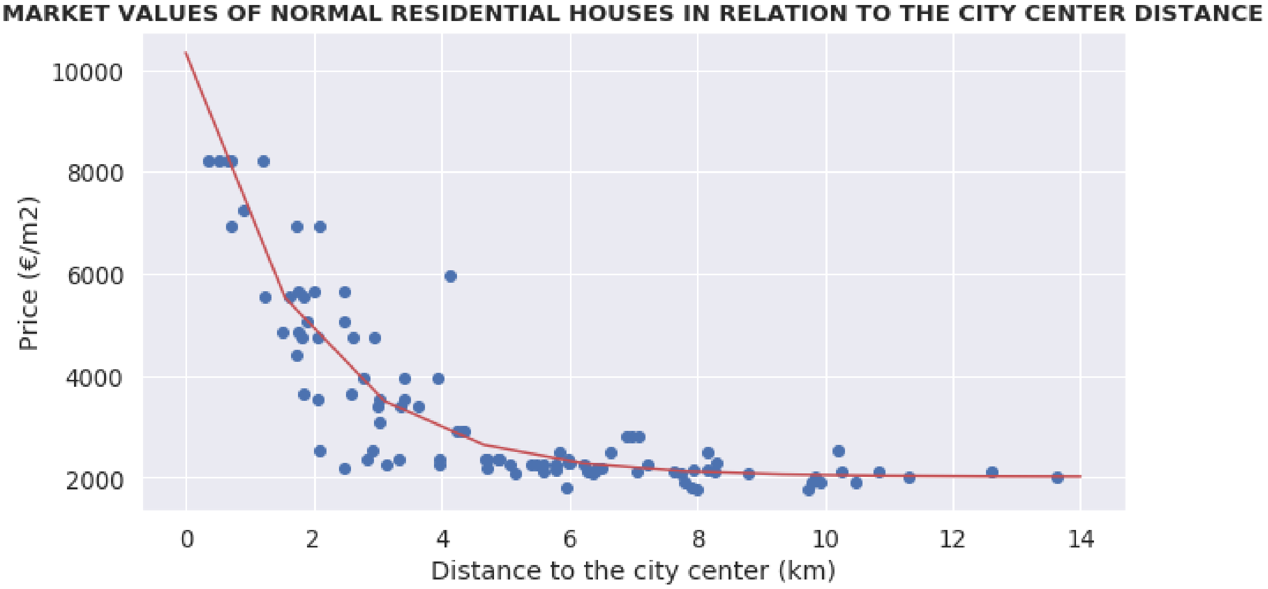


Figure 24 - Trend of the market values of residential houses in normal condition, depending on the city center distance.

- Scatter plots with different colors, based on the area of membership, and with the regression curve, for market values of residential houses in normal condition (respectively fig. 25 and 26).

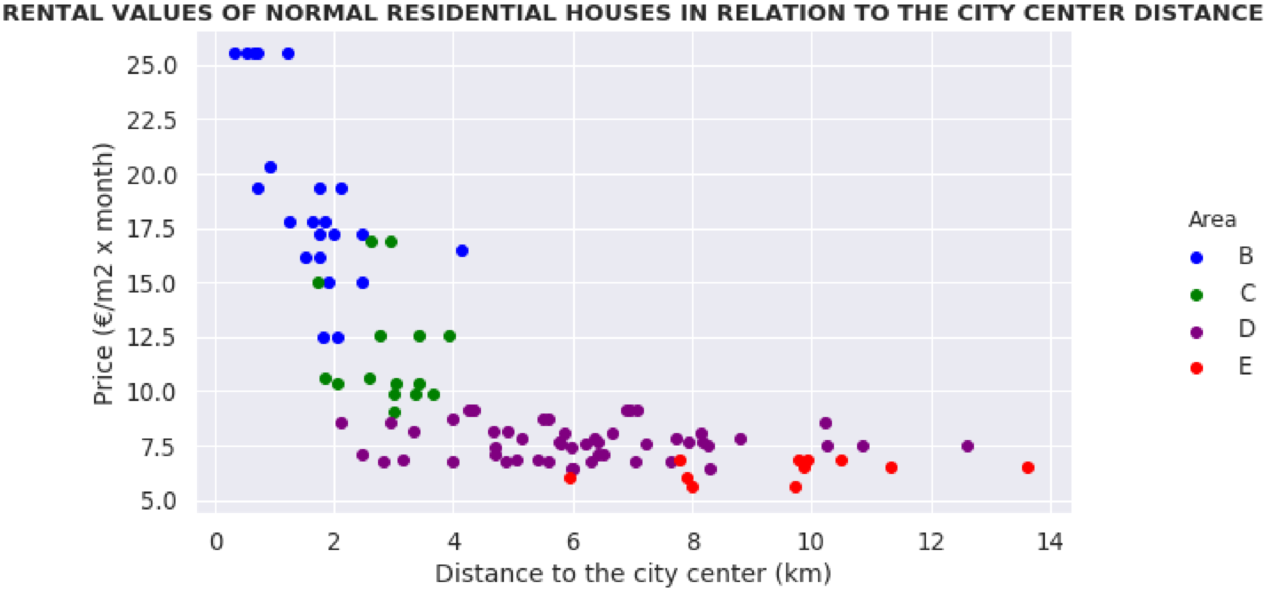


Figure 25 - Distribution of the rental values of residential houses in normal condition, depending on the city center distance and on the area location.

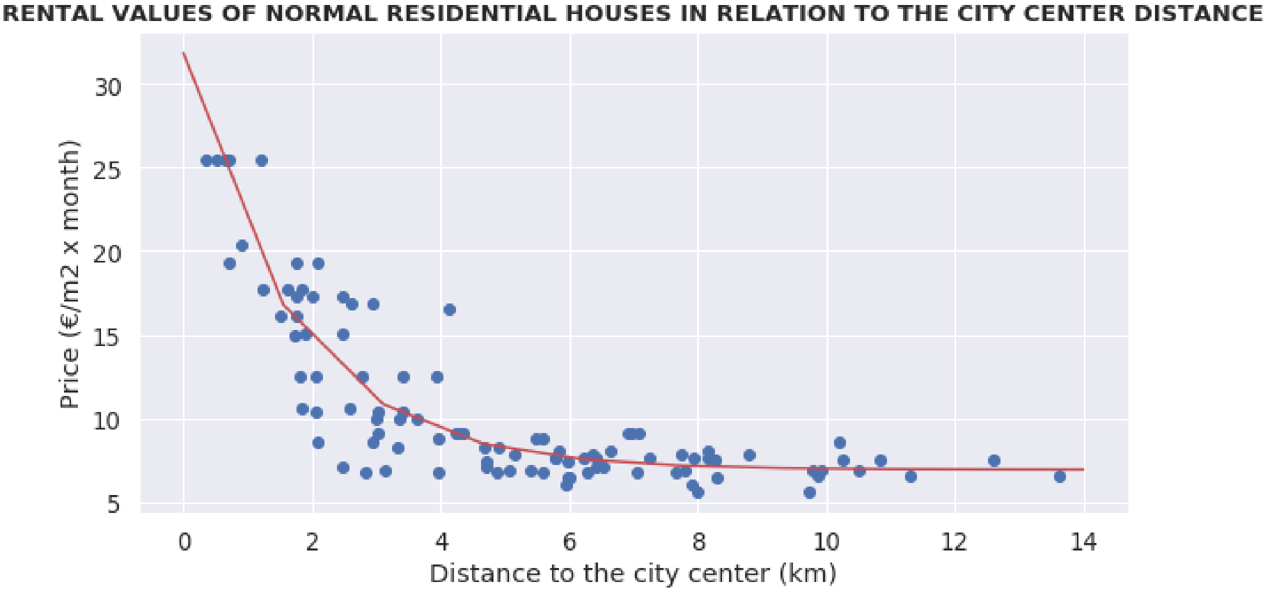


Figure 26 - Trend of the rental values of residential houses in normal condition, depending on the city center distance.

As it can be seen from the above charts, with the increasing of the distance of the neighborhoods from the city center, the residential house values decrease. Another important aspect is that the decrease of the market and the rental values is steeper until the proximity from the city center of 2 km, then the curve becomes flatter (the values decreases exponentially).

This rule is also true for stately house where, however, the value decreases is more regular.

From the charts that distinguish the 4 Milan areas by different colors, it is possible to notice that residential houses located in B area, which includes the neighborhoods closest to the center, are in

general the most expensive. Moving to the other Milan areas, the residential house values decrease in gradual manner, according to the distance from the center. This rule is not always true, indeed it is possible to observe some exception, for example C and D areas include some neighborhoods with the same distance from the city center, nevertheless the prices of the houses within D area are more convenient compared to the houses localized in C area.
As regards the stately houses, it cannot be seen a clear distinction between the market and rental values of this type of houses, depending on whether the home is located in the B and C areas.

4.3 VENUES ANALYSIS

The last step of this analysis aims to determine the most common venues that are present in the 2 neighborhood clusters.
The first important thing is to determine the location of the neighborhoods belonging to the 1st or 2nd cluster. For this reason, a map of Milan has been created, which shows the position of the neighborhoods, differentiated by clusters through 2 different colors (fig. 27).

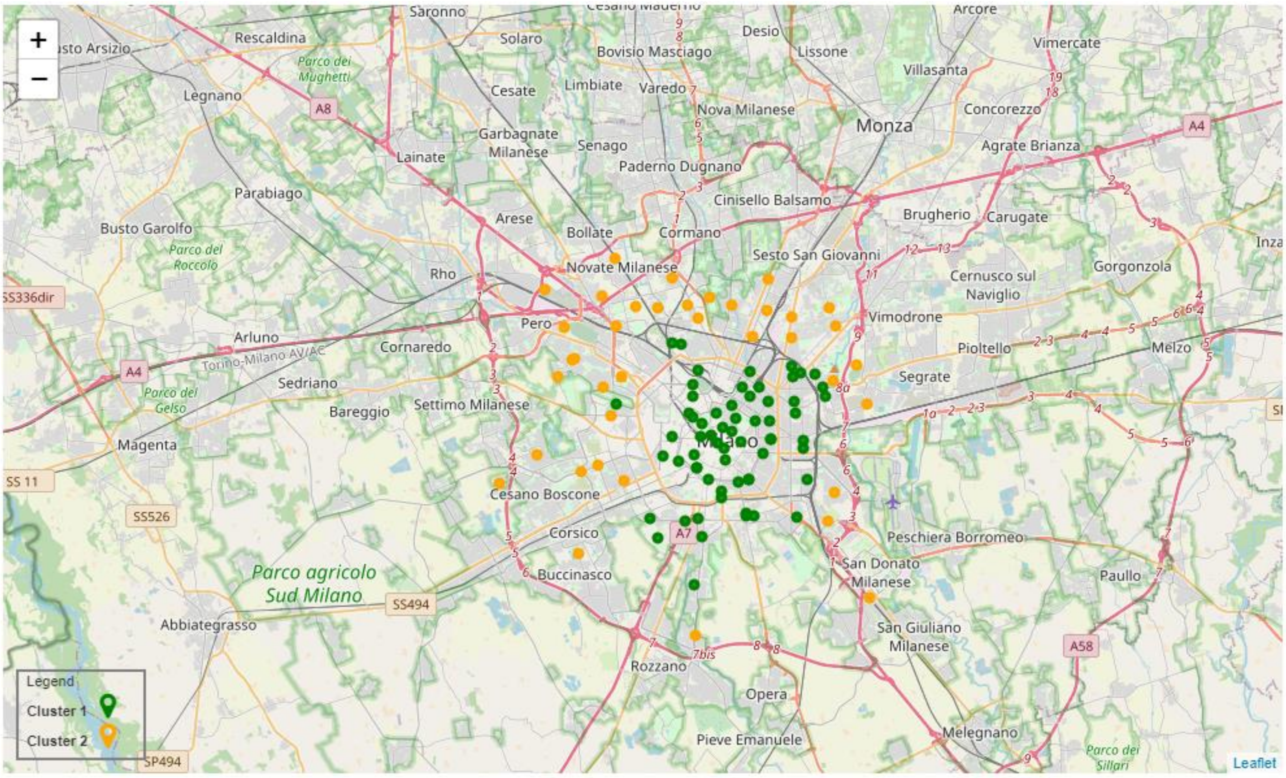


Figure 27 - Location of the neighborhoods belonging to cluster 1 or 2.

As it is possible to see from the above map, cluster 1 includes the neighborhoods closest to the Milan city center; on the contrary, cluster 2 covers the suburban neighborhoods.

As regards the top 10 most common venues for cluster 1 and 2, the following horizontal bar charts have been created (fig. 28 and 29):

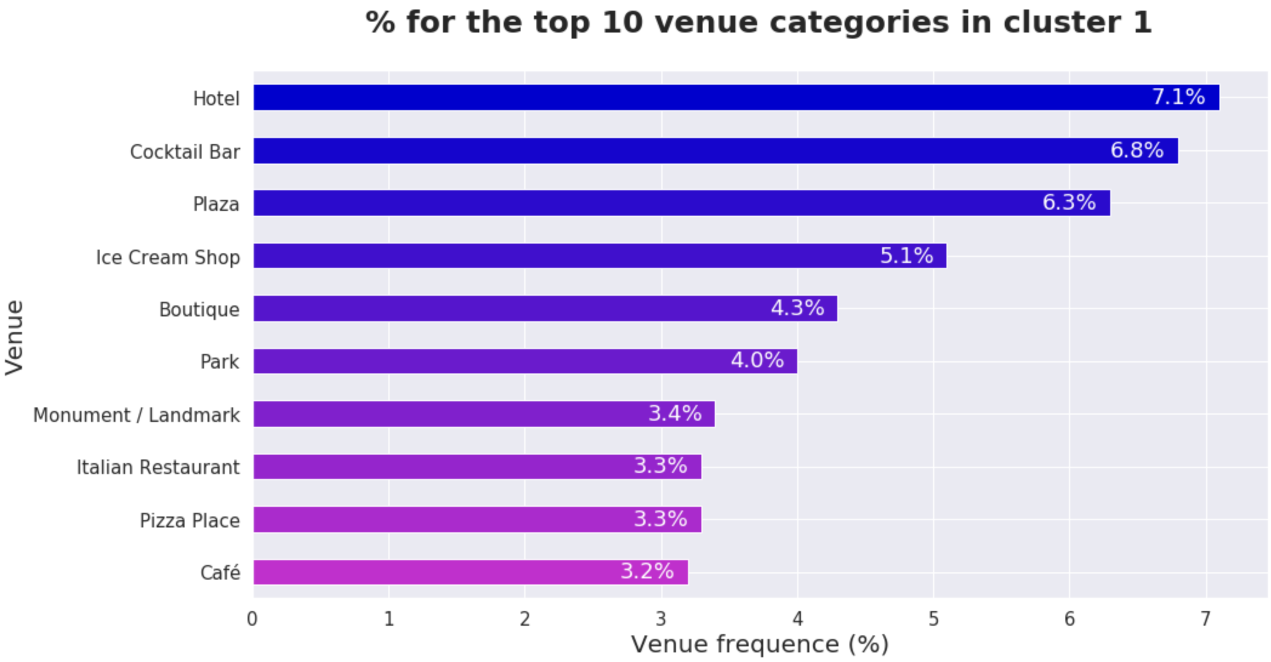


Figure 28 - Top 10 most common venues of neighborhoods belonging to cluster 1.

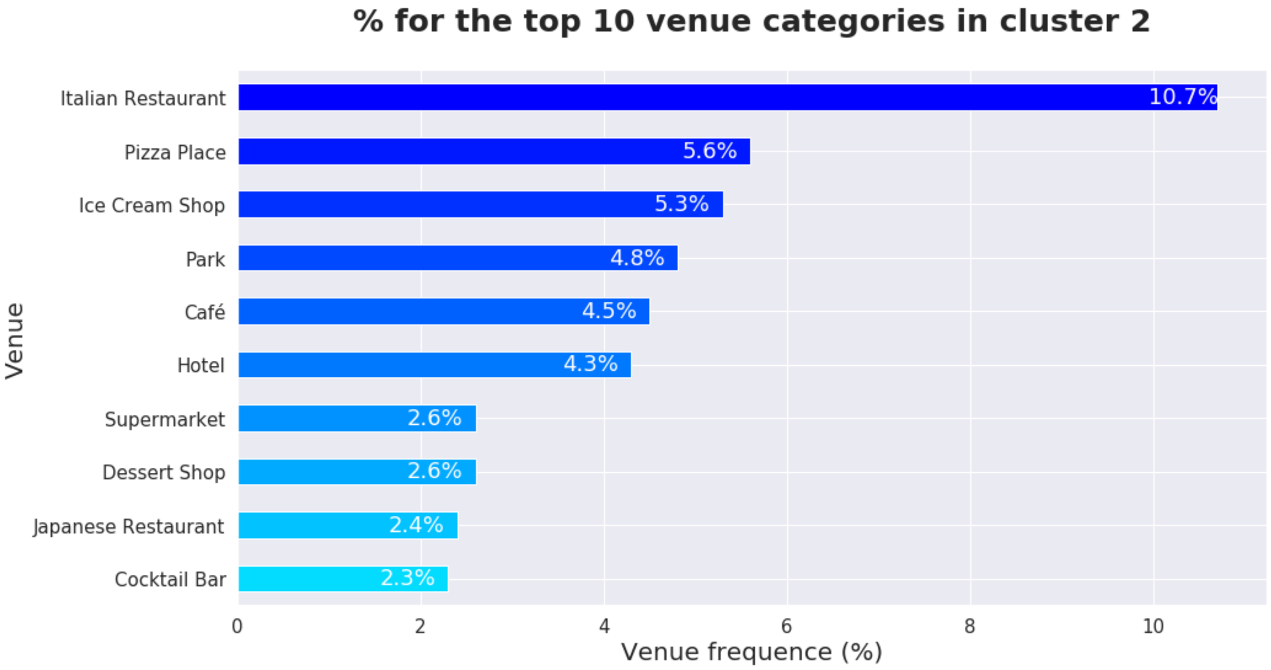


Figure 29 - Top 10 most common venues of neighborhoods belonging to cluster 2.

As it is possible to observe to the 2 horizontal bar charts, the most common venues present in the 2 clusters are quite different:

- cluster 1 is full of hotels, attractive plazas, boutiques, monuments, Italian restaurants and also art galleries, etc.;
- cluster 2, on the other hand, includes many places that serve food and drinks, in addition to park and hotels.

5. DISCUSSION

This analysis revealed some important aspects to take into account when someone decides to buy or to rent a house in Milan and they don't know in which neighborhood go to live.

First of all, the house price is one the main factors that affects the house choice. It's good to know which Milan areas have house prices in line with the budget of the subject: this will allow him to refine its choice and optimize the time of house research. For example, a person who intends to invest a large sum of money could focus on houses for sale/rent localized in B and C areas (the most expensive but, obviously, the most prestigious ones). On the contrary, a person that plans to spend a relatively small amount of money can focus on the house located in D and E areas (the most distant to the city center but, at the same time, the most affordable ones).
However, it's important to considerate the type and the state of the house: indeed, a house located in the same Milan areas but of different type and/or in different condition will have a different price. Certainly, a stately house is dearer than a residential house in excellent condition; this last one, in turn, is less convenient that a residential house in normal condition. It's better to have a clear idea about the condition of the house that it's looking for, because this aspect greatly affect the house value.
However, if this is not enough and someone wants to focus the attention on a smaller group of choice, they could consider the distance of the neighborhoods from the city center beyond which the price is no longer in line with the predefined budget, this will allow them to select the more aligned neighborhoods. In general, there is a rapid decrease of the market/rental values of the residential house, until a distance of 4 km from the city center, then the price decrease more slowly (beyond 4 km only neighborhoods belonging to D and E areas are present).
Different situation occurs for the stately houses, indeed the values decrease in a more uniform way.

Someone who want to buy or to rent a house in Milan would do well to choose the best neighborhoods to live also according to their interests. This will ensure less use of the car and the public transport, therefore less stress.
This study has found out that Milan neighborhoods can be split in 2 clusters: the first one includes the neighborhoods closer to the city center and the second one the neighborhoods located in the suburbs.
It has been observed that cluster 1 is rich in hotels, attractive plazas, boutiques, monuments, Italian restaurants, art galleries, etc.: all venues that attract tourists and people who love art. Cluster 2, on the other hand, includes many places that serve food and drinks, in addition to park, hotels, etc. The neighborhoods belonging to this cluster are the perfect ones for people that live in Milan and don't give up to enjoying social life.
Finally, it is possible to notice that in general Milan, even if it is a metropolis, is rich in parks. This is an important aspect, since it's able to offer relaxation and leisure to those who appreciate outdoors life and for people who have dogs.

What could be done better?

This project only considers 2 aspects that could affect the neighborhood choice i.e. house value and the proximity to the venues of interests. However, there are many more characteristics that may come into play, like the presence of public transports (it’s well known that traffic on the roads of big cities is particularly heavy) or the availability of the houses themselves. Future research could develop a methodology that allows to consider more aspects.

In addition, this project has been made by the use of Foursquare account that has limitation on the number of API call (in this case 100 for each neighborhood), this could have been affected the real venue frequencies. In the future this problem may be bypassed using a paid account.

Finally, it would be interesting to split the Milan neighborhoods in more than 2 clusters, as to improve the differentiation of the neighborhoods according to the most common venues. However, this would affect the K-means accuracy.

6. CONCLUSION

This project has been designed as a tool to help people who intend to buy or to rent a house in Milan, as to facilitate the choice of the zone where focus the house research.
First of all, data has been collected from the Italian Revenue Agency website, all the neighborhoods have been geolocated (by using Google Maps Geocoding API) and then mapped to determinate the exact position. Subsequently, the data were analyzed, as to determinate the correlation between the house location (Milan borough, area and finally neighborhood) and the related price, also depending on the type and the state of the property, and the most frequent venues that are present (by using Foursquare API).

It has been possible, on the whole, to find satisfactory answers to the question cornerstone of this study. Indeed, based on these findings, it has been concluded that there is a positive correlation between the house prices and the proximity of them to the city center; moreover, it has been possible to split the neighborhoods in different clusters, depending on the venue frequencies, and so determine the differences among them.
Obviously, there may be other variables at play that also contribute to the house choice. However, it is something that could help someone during the decision process of the future house location.

Ballroom
Seafood Restaurant
Sporting Goods
Mediterranean Restaurant
Abruzzo Restaurant
Cupcake Shop
Pool Hall
Restaurant Italian
Noodle House
Irish Pub
Shoe Store
Office Supplies
Church
Hotel Restaurant
Shop Italian
Comedy Club
Pool
Gallery Hotel
Restaurant Park
Rock Club
Site Cocktail
Hotel
Dim Sum
Shop Restaurant
Ramen
Hostel
Scenic Lookout
Art Museum
Kebab Restaurant
Fried Chicken
Dessert Shop
Boutique Men
Design Studio
Historic Site
Japanese Restaurant
Winery
Performing Arts
Paper Office
Sports Club
Vegetarian Vegan
Store
Soccer Stadium
Lombard
Art Gallery
House Art
Bakery
Restaurant Restaurant
American Restaurant
Coffee Shop
Hotel Art
Warehouse Store
Jewelry Store
Clothing Store
Game Store
Home Store
Men Store
Electronics Store
Trattoria Osteria
Food Store
Roman Restaurant
Fitness Center
South American
Plaza
Italian Restaurant
Cheese Shop
Furniture Home
Nightclub
Airport
Castle
Pub
Fountain
Canal
Food Truck
Water Park
Toy Game
Farmers Market
Gym
Asian Restaurant
Beer Bar
Gourmet Shop
Scandinavian
Cocktail
Bar
Steakhouse

Candy Store
Jazz Club
Pizza Place
Soccer Field
Restaurant Ice
Arcade
Playground
Wine Shop
Dive Bar
Cultural Center
Piadineria
Argentinian Restaurant
Airport Lounge
Health Food
Creperie
Campanian Restaurant
Movie Theater
Café
Sandwich Place
Golf Course
Monument Landmark
Gallery Opera
Bed Breakfast
Chicken Joint
Bistro
Gastropub
Cosmetics Shop
Gym Pool
Chocolate Shop
Brewery
Place Restaurant
Comic Shop
Burger Joint
Salon Barbershop
Breakfast Spot
Sum Restaurant
Ice Cream
Salad Place
Department Store
Planetarium
Chinese Restaurant
Sardinian Restaurant
Restaurant Bar
Bar Seafood
Plaza Plaza
Camera Store
Plaza Restaurant
Arts Venue
Other Nightlife
Museum
Opera House
Snack Place
Tram Station
Supplies Store
Boutique
Middle Eastern
Park
Gym Fitness
Bar Restaurant
Vegan Restaurant
Museum Ice
Shopping Plaza
Tuscan Restaurant
Amphitheater
Cemetery
Cream Shop
Pedestrian
Thrift Vintage
Gift Shop
Indie Movie
Supermarket
Music Venue
Pet Store
Bookstore
Pastry Shop
Event Space
Diner
Sushi Restaurant
Climbing Gym
Moroccan Restaurant