Real Estate Prices Analysis in Roma

Introduction

This project aims to analyze real estate market in the city of Roma. The stakeholders could be any individual or group of investors interested in real estate opportunities. In particular, we want to find out relevant differences in market prices between similar neighborhoods, to spot overestimated and underestimated market zones.

Data

Italian Agency for Territorial Services provides (on demand, for registered users only) a free dataset with real estate observed prices for each zone of a city in a given time interval. A .kml file for each city is provided as well with the borders of each zone. This file has been converted, for convenience, to geojson format using a free online converter service.

Latest available data are prices observations of the second half of 2019. For each zone there is a price range (min and max observed) for purchasing and for rentals, depending on the maintenance state and the usage destination (residential, office, shop) of the property. The source format is .csv, delimited by semicolon (;).

Data is originally split into two source files: the first one contains the list of all neighborhoods and the second one contains the observed prices.

In addition we used Foursquare data: for each neighborhood a list of nearby venues. The distribution of venue categories is used to build a classifier model for neighborhoods.

Methodology

First we proceed with an exploratory data analysis. In the source files the city is divided into 234 neighborhoods, for each the most frequent house type and manteinance state are reported. Some neighborhoods (23) have no purchase and rental data.

This suggests to focus on the most common combination of usage destination and state of maintenance: residential houses in a normal state, which is present in all the 211 neighborhoods with purchase and rental data.

The neighborhoods of Rome will be compared using Foursquare APIs. In order to perform the exploration we must provide geospatial coordinates of a representative point of each neighborhood, which are not included in the source files. We can calculate the centroid of each neighborhood extracting borders coordinates from geojson file and processing them with Shapely library.

Neighborhoods are represented as polygons of different shape and size, tipically smaller in the center of the city and much larger in the suburbs. Hence, we can’t use the same radius in the foursquare api calls for all the neighborhoods. We decided to calculate the area of each polygon and set the radius so that a circle of that given radius centered on the centroid of the neighborhood would have the same area as the neighborhood polygonal representation.

The source files have been merged in a single Pandas dataframe keeping only relevant information:

- **Zona**: a short code for the neighborhood
- **Compr_min, Compr_max, Compr_avg** = observed min, max and average for purchase
- **Loc_min, Loc_max, Loc_avg** = observed min, max and average for rental
- **Zona_Descr**: local name for the neighborhood
- **lat, lon, radius**: geospatial coordinates and approximate radius of the neighborhood

The Fourquare API call returned the top 100 venues for each neighborhood but unfortunately for some neighborhoods there were too few venues to perform a significant analysis. We decided to discard neighborhoods with less than 20 venues found.

Then, we categorized all the venues and calculated the relative distribution of venue categories for each neighborhood. In order to take in account of the number of venues found we decided to add a feature representing the number of venues found, divided by the max (100).

Finally, we used this dataset as input for a clustering model with k-means algorithm, grouping neighborhoods in 10 clusters.

Results

Here is the cluster list, with a brief interpretation of the results.

C1. Medium density (0.5) of venues. Hotel is the most common category.

C2. Quite low density of venues (0.3), the most common category is Pizza Place. Small shops and supermarkets are quite common, which suggest that these neighborhood are mainly residential.

C3. High density of venues (1.00). Restaurants, Cafè and Pizza Place are the most common venues.

C4. Quite high density of venues (0.7). Few hotels and shops, many restaurants and entertainment venues.

C5. Low density of venues (0.25). Beyond the usual Italian Restaurants and Cafè we can find some categories that are quite uncommon in other clusters. Most of them are related to sports: Soccer Stadium, Gym Pool, Tennis Court, Golf Course, Soccer Field etc.

C6. Commercial districts in the suburbs. Most common venues are stores and shops.

C7. Quite Low density (0.35) of venues. Restaurants and Cafe are mixed with urban venues like Parks, Stations, etc.

C8. Low density of venues (0.25-0.30), which suggest peripheral neighborhoods. Cafè is the most common venue category, restaurants are less common then other clusters.

C9. Very high density of venues (1.0). Hotels on top. No shops among most common venues categories.

C10. Medium density of venues (>0.5). Shops and stores are quite common, suggesting a significant commercial presence.

Some clusters look similar, with the only difference of venue density. For example: C1 and C9 (mid-central and central hotels), C2 and C10 (residentials neighborhoods with commercial activities), C5 and C7 (mainly residentials).

Finally, let's look for underrated neighborhoods. The most interesting cases are:

Cluster 3

	Purchase_avg (mq)	**Purchase_avg_cluster (other neighborhoods)**	**saving**
B18 ESQUILINO (PIAZZA VITTORIO)	3.250,00	4.897,00	33,63%
C10 GARBATELLA	3.450,00	4.875,00	29,23%
E34 OSTIA (VIA DELLE BALENIERE)	2.275,00	5.011,00	54,60%
D29 EUR (VIALE EUROPA)	3.600,00	4.858,00	25,90%

This cluster includes neighborhoods of the historic center, which obviously have a very high market value. So we have to be careful to compare prices. For this reason the really underestimated neighborhood seems to be B18 ESQUILINO (PIAZZA VITTORIO), which is a very central neighborhood itself, unlike the others reported.

Cluster 4

	Purchase_avg (mq)	**Purchase_avg_cluster (other neighborhoods)**	**saving**
C30 (PIGNETO)	2.875,00	4.336,00	33,69%

Cluster 7

	Purchase_avg (mq)	**Purchase_avg_cluster (other neighborhoods)**	**saving**
D14 (CENTOCELLE)	2.350,00	3.433,00	31,54%
E33 (ACILIA SUD)	2.150,00	3.450,00	37,68%

Cluster 4 and 7 otherwise are quite omogeneous so we can say that the higlighted neighborhoods seem really underestimated.

Discussion

The weight of the feature 'Venues', which represents the number of venues found, on one hand seems prevalent and pushes the model to cluster neighborhoods primarily on venues density but on the other hand gives us a better classification because peripheral neighborhoods have low venue density and it makes more sense to group them together.

The city of Roma covers a very wide territory, and the presence of historic sites, monuments or touristic attractions is very important for giving a proper rating to a neighborhood. A suggested improvement for the model could be a weighted clustering where some categories are weighted to be more significant and maybe some others can be grouped together, for example some categories of restaurants.

Conclusion

The better chance of investments are in the following four neighborhoods:

	purchase_avg	rental_avg	rental/purchase
CENTOCELLE (PIAZZA DEI MIRTI)	2.350,00	10,25	0,004362
ACILIA SUD (VIA DI PRATO CORNELIO)	2.150,00	9,40	0,004372
PIGNETO (PIAZZA DEL PIGNETO)	2.875,00	11,90	0,004139
ESQUILINO (PIAZZA VITTORIO)	3.250,00	12,00	0,003692

If the investment is made for renting the purchased house, the better rental/purchase ratio can be fouind in a peripheral and residential neighborhood as Acilia sud. Otherwise, if the investor wants to resell the house for a higher price, the better chance of getting a value increment is in a central or mid-central underestimated neighborhood as Esquilino or Pigneto, depending of the capital he is willing to invest.