The battle of neighborhoods

IBM Data Science Professional Certificate - Capstone Project Report

Leonardo Fernández
leohfermamdez@gmail.com

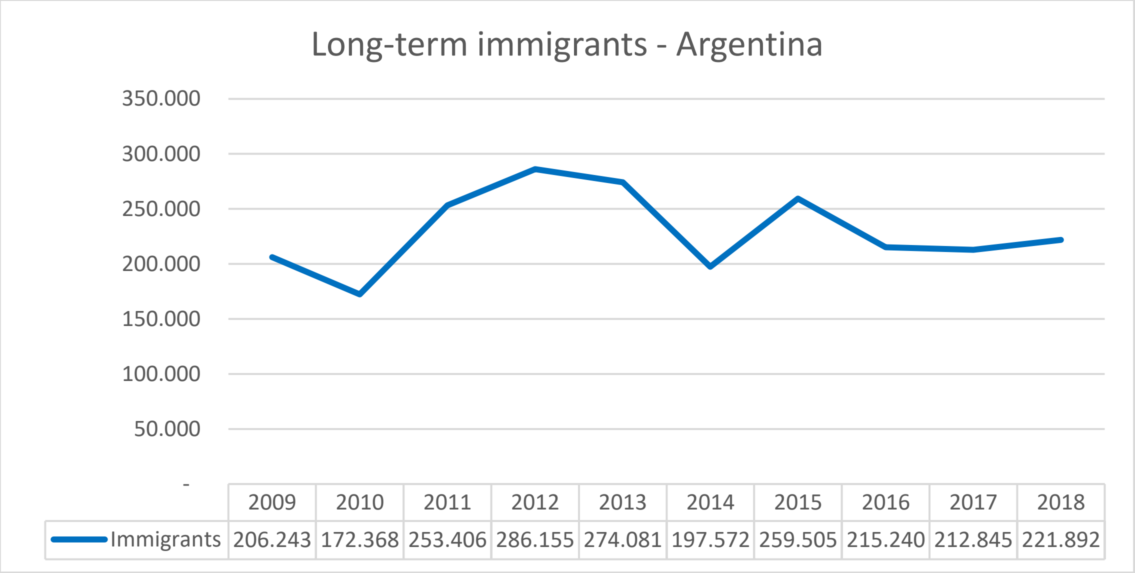
Table of contents

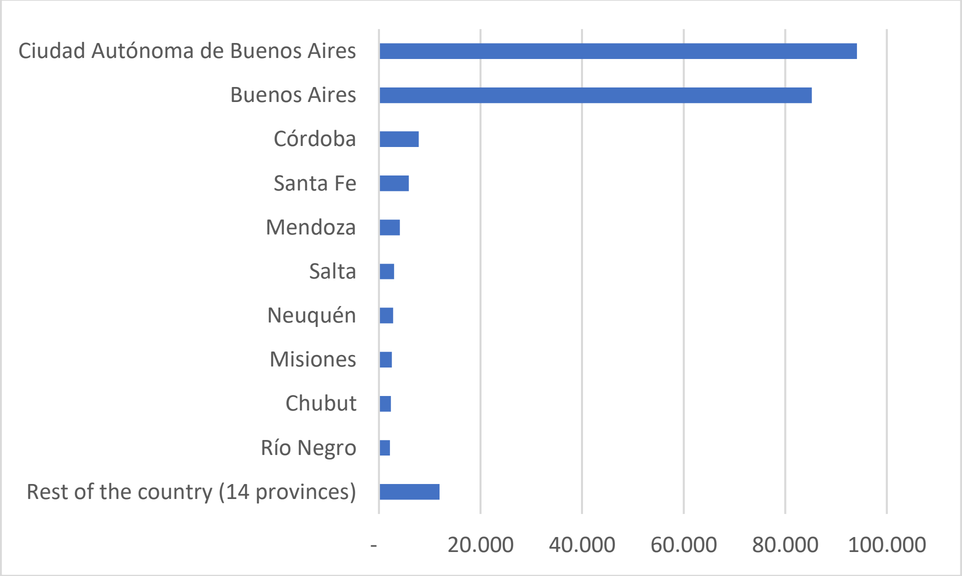
Introduction ... 2
Background ... 2
Problem ... 3
Project objective ... 3
Data ... 4
Data sources ... 4
General city information ... 4
Venues in the area ... 4
Prices and characteristics of current rentals. ... 4
Data cleansing ... 4
Methodology ... 5
Results ... 6
Neighborhoods ... 6
Clustering ... 7
Housing pricing ... 8
Discussion ... 11
Datasets ... 11
Population density ... 11
Venue category hits ... 11
Clustering ... 12
Housing results ... 12
Linear regression ... 13
Conclusion ... 15

Introduction

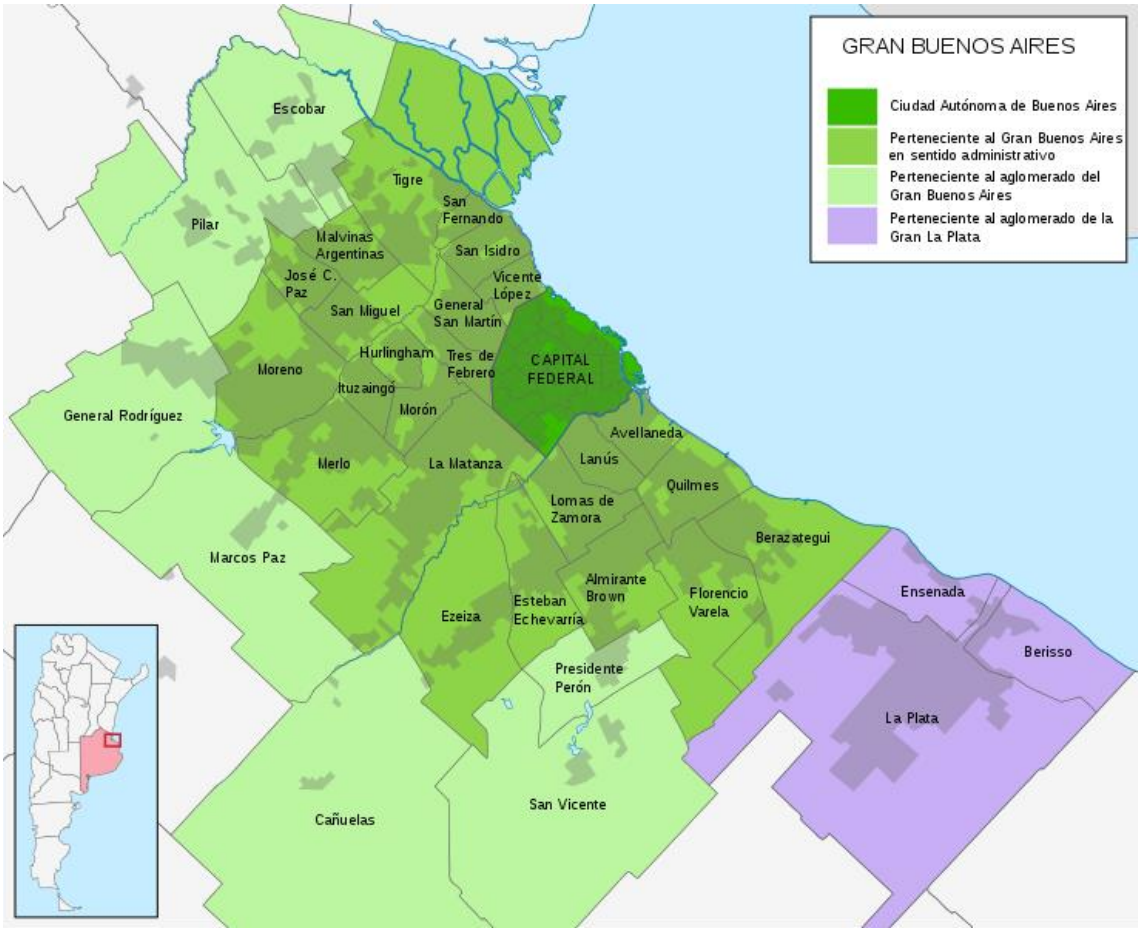
Background

According to the Ministry of Interior of Argentina, the average amount of long-term immigrants to the country for the last ten years is 230k, and around 80% of them are moving to the Buenos Aires Metropolitan Area (AMBA - Buenos Aires City plus surrounding suburbs)[1][2].





[1] http://www.migraciones.gov.ar/pdf/estadisticas/radicaciones_resueltas_2018.pdf
[2] https://www.indec.gob.ar/dbindec/folleto_gba.pdf



Problem

Moving to another city usually is a major change in people’s life and a decision that should not be taken lightly. Signing a rental contract is a commitment for a long period of time and making a mistake could be costly to fix.

Characteristics and venues of the chosen neighborhood could affect habits and daily activities. In addition, every mayor city has zones safer than others, which might be unknown prior to moving to the new city. Finally, rental cost is one of the biggest monthly expenses that cannot be easily modified and therefore it can have a big impact in personal monthly budget.

Project objective

The aim of this project is to elaborate and analyze an algorithm capable of selecting the neighborhoods in a destination city that are most similar to the surroundings of an origin address according to venues in the area and population density, pick the 10 safest and predicting the rent value of an apartment or house.

A particular case of a person moving from Barcelona, Spain, to Buenos Aires, Argentina will be used as an example, but the technics and methods could be extrapolated to any different case.

Data

Data sources

General city information

For the origin city, Barcelona, I scrapped **borough** and **neighborhood** names and **population density** from Barcelona's townhall website using **Selenium** library. After cleaning I added each centroid **coordinates** with **Geocoder** library.

The case of the destination city, Buenos Aires, is much more complex as several sources are used and need to be carefully combined after cleaning each one.

First I get the province, **borough**, **neighborhood** names and its **coordinates** from the Ministry of Interior website in a **JSON** file which I process with **Pandas** (I need the province for filtering as borough names are repeated in different provinces)

For **population density** the best granularity available is at the borough level, but it can be used as an estimate for each neighborhood inside each borough with a good level of confidence. Data is collected from the National Institute of Statistics website via **Selenium** webdriver.

Finally, crime rate information is extracted from a **CSV** file available at the Security Ministry of Argentina website.

Venues in the area

The list of the venues per area are gathered through the Foursquare Explore API.

Prices and characteristics of current rentals.

The dataset of characteristics and prices of currently available rental places will constructed by scrapping through **BeautifulSoup4** the MercadoLibre website, one of the most used ones in the country. For the simplicity of this analysis, I will concentrate just on the zone and living area of each apartment/house.

Data cleansing

The Barcelona dataset table acquired through Selenium has several rows at the top and bottom that must be erased. The first column is repeated 3 times, and it comprises borough and neighborhood in the same column considering borough as a grouping row. These rows should be erased and the borough separated into a new column.

After all Argentina information is gathered and correctly merged, different filters are applied to leave just the relevant information regarding the Buenos Aires Metropolitan Area.

The las row contains NaN data and must be erased, and the first column comprises borough and neighborhood as a comma separated value and should be separated into 2 different columns.

Finally, a 25km radius from the city center is applied for neighborhoods in AMBA, as it would be the most realistic case if somebody would be to move to Buenos Aires.

Methodology

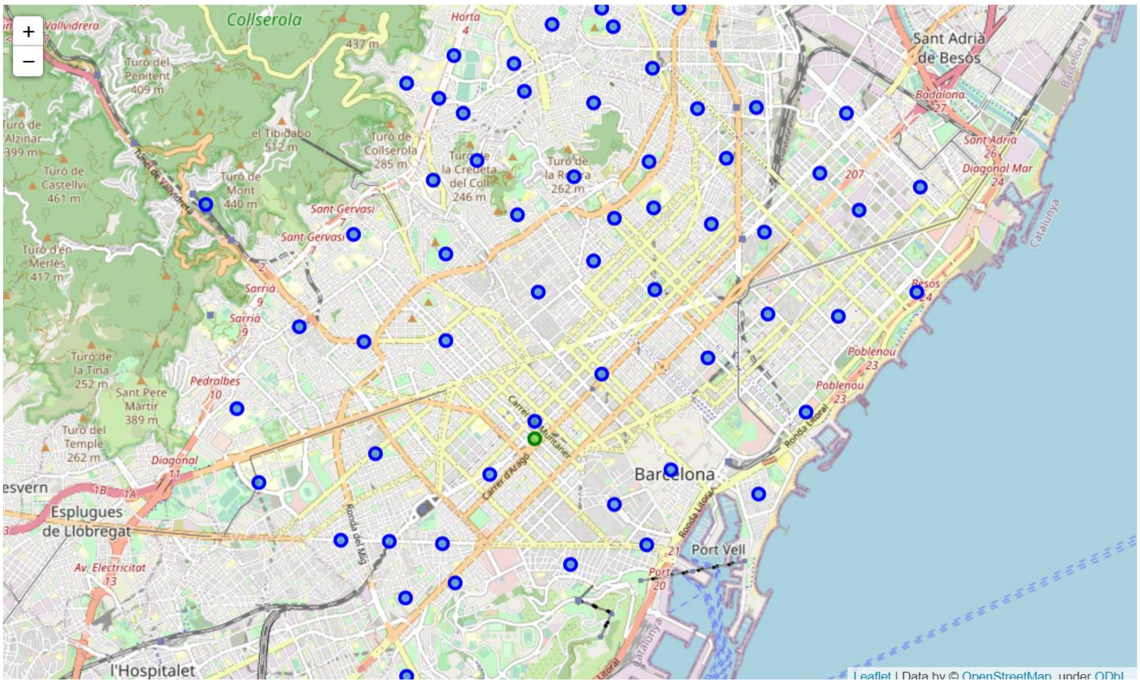
The first step should be to characterize the origin neighborhood and the different ones in the destination city. This will be done by categorizing the most popular venues through Foursquare City Guide APIs and the population density of each city based on official census. After getting this information, a **k-means clustering** algorithm is applied and the k is selected through an **elbow curve** analysis.

Once the similarity between neighborhoods in AMBA and the origin address has been established, the top 5 safest ones will be chosen for the next step, and over this subset the currently available rental places will be scrapped from the corresponding website. With this information at hand, a **linear regression algorithm** will be used to construct a prediction formula for the rental price based on zone and living area.

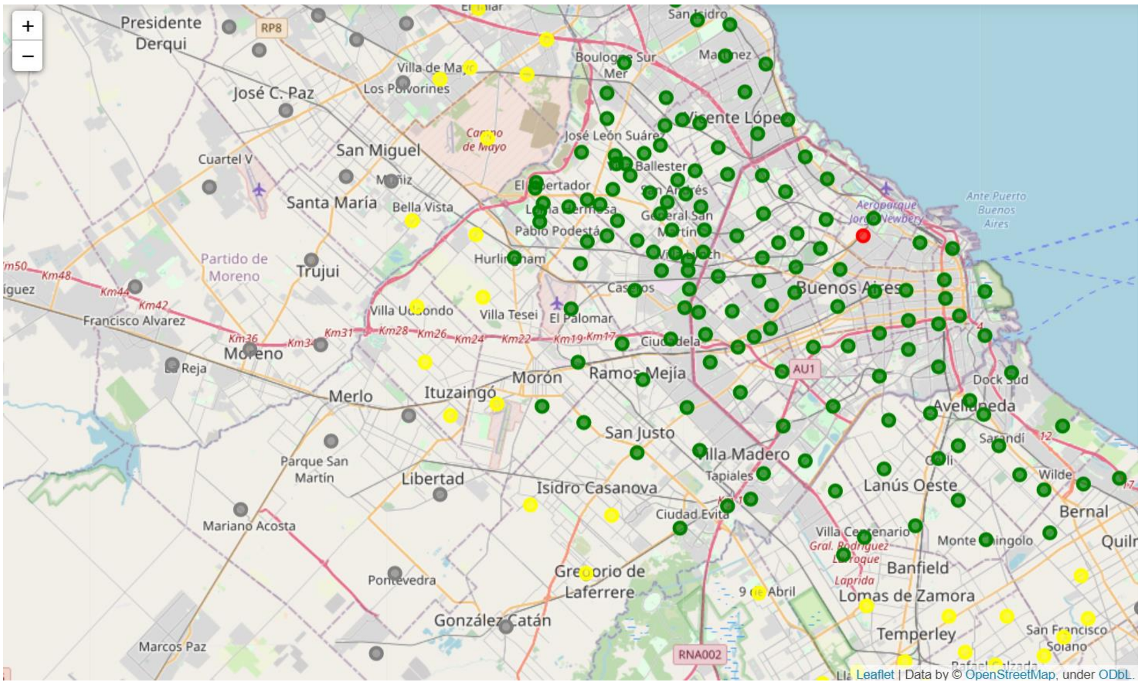
Finally, housing rent price will be predicted based on housing requirements and current prices of available places. For this example, I will predict the rental price of a 70 m^2 place.

Results

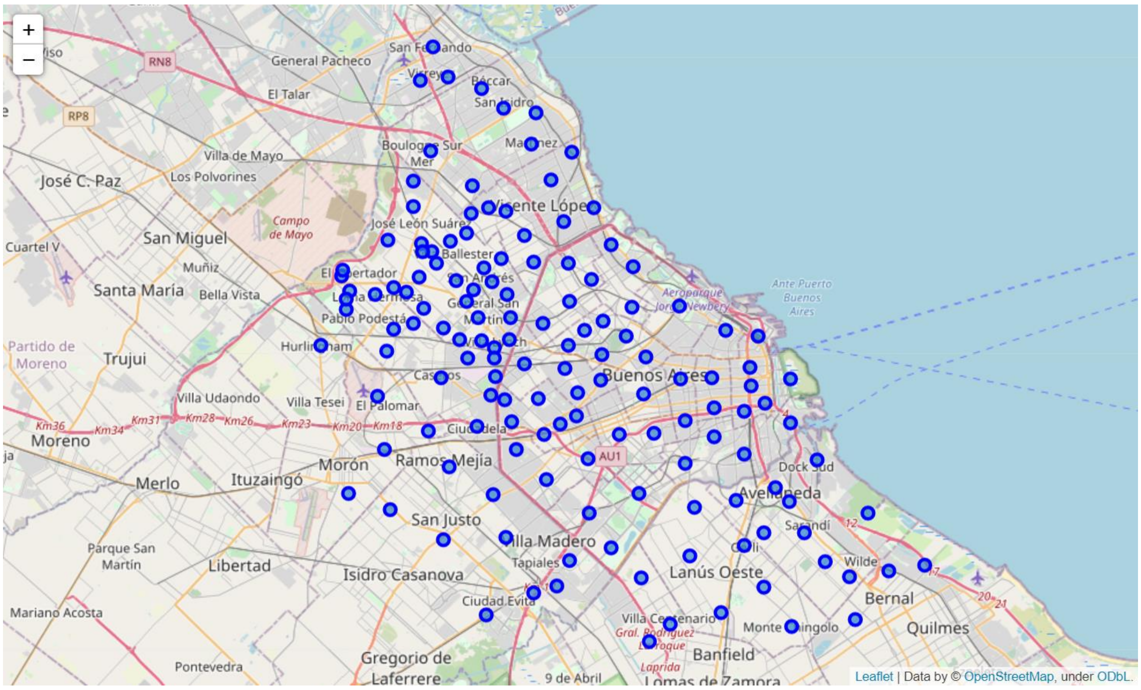
Neighborhoods



Blue dots are all Barcelona Neighborhoods, the green dot is the *origin address*.

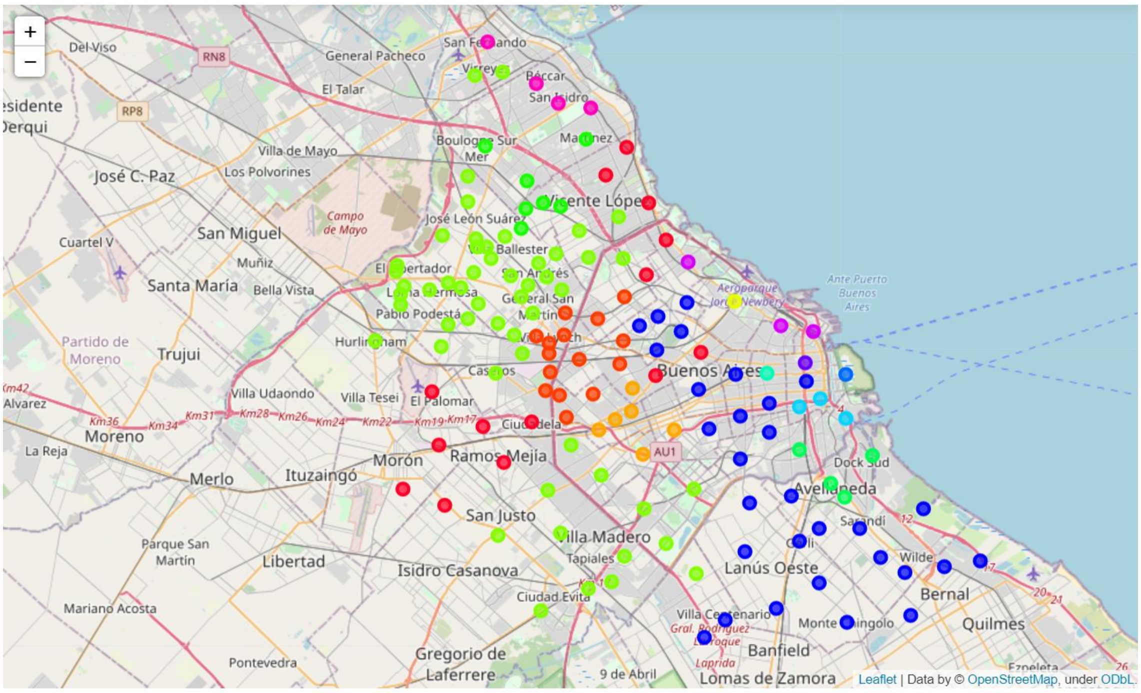


The red dot is the city center and green dots are the neighborhoods inside the radius filter that are analyzed in this project. Yellow dots where considered but finally discarded together with gray dots that where discarded from the beginning.

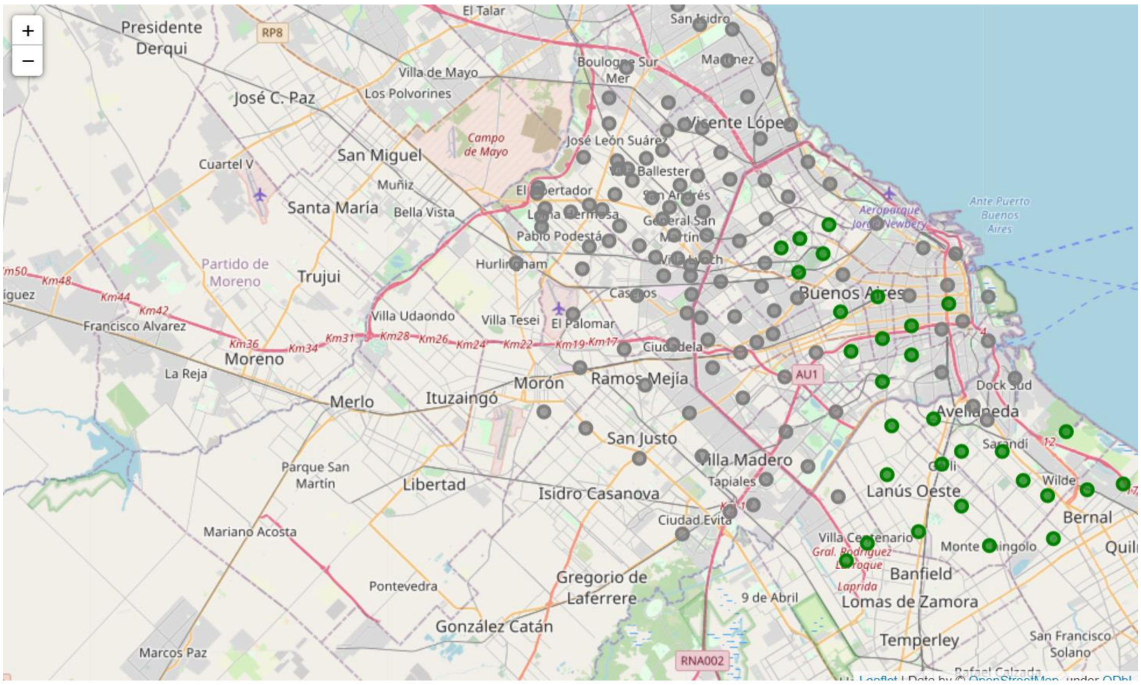


This is the filtered set for analysis.

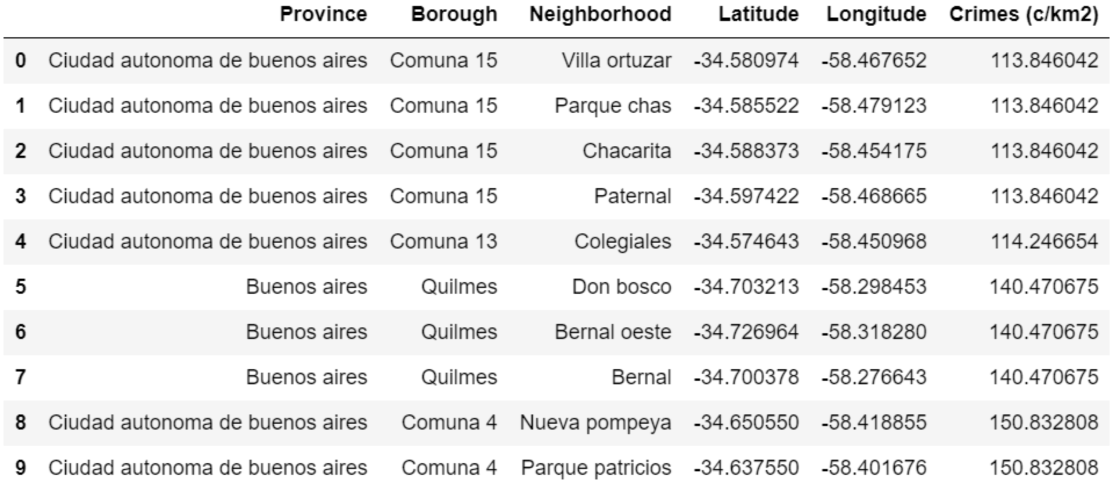
Clustering



Clustered neighborhoods



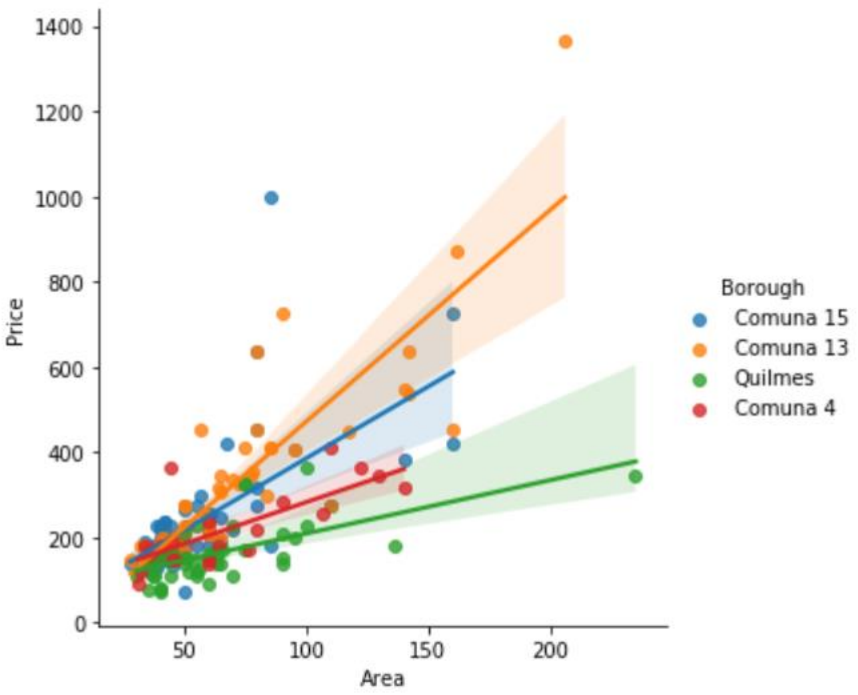
Selected cluster that has bigger similarity with the origin address in green

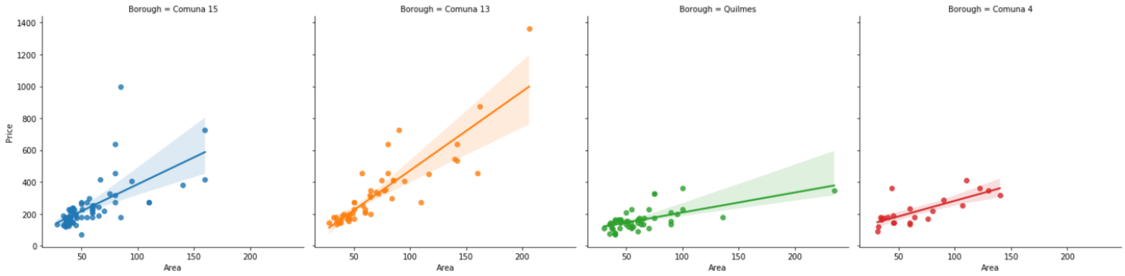


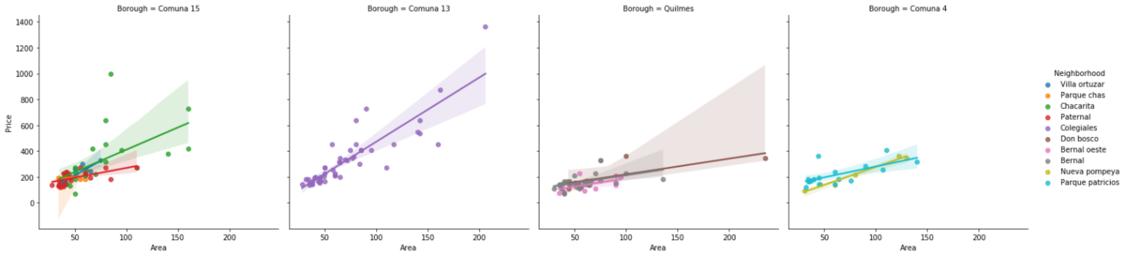
	Province	Borough	Neighborhood	Latitude	Longitude	Crimes (c/km2)
0	Ciudad autonoma de buenos aires	Comuna 15	Villa ortuzar	-34.580974	-58.467652	113.846042
1	Ciudad autonoma de buenos aires	Comuna 15	Parque chas	-34.585522	-58.479123	113.846042
2	Ciudad autonoma de buenos aires	Comuna 15	Chacarita	-34.588373	-58.454175	113.846042
3	Ciudad autonoma de buenos aires	Comuna 15	Paternal	-34.597422	-58.468665	113.846042
4	Ciudad autonoma de buenos aires	Comuna 13	Colegiales	-34.574643	-58.450968	114.246654
5	Buenos aires	Quilmes	Don bosco	-34.703213	-58.298453	140.470675
6	Buenos aires	Quilmes	Bernal oeste	-34.726964	-58.318280	140.470675
7	Buenos aires	Quilmes	Bernal	-34.700378	-58.276643	140.470675
8	Ciudad autonoma de buenos aires	Comuna 4	Nueva pompeya	-34.650550	-58.418855	150.832808
9	Ciudad autonoma de buenos aires	Comuna 4	Parque patricios	-34.637550	-58.401676	150.832808

Housing pricing

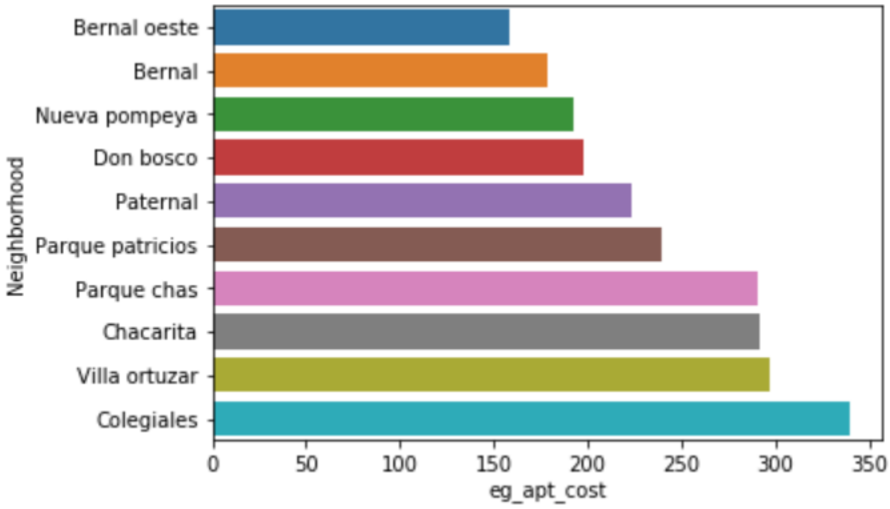
The housing results from the scrapping are as follows:



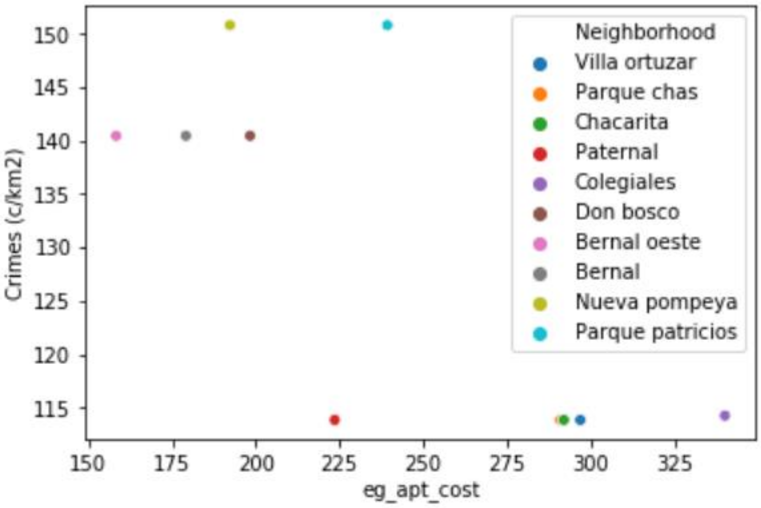




And the rental proce prediction according to the linear regression for each neighborhood are:



Which combined with the crime rate data provides the information needed to pick the right neighborhood in Buenos Aires to move to:



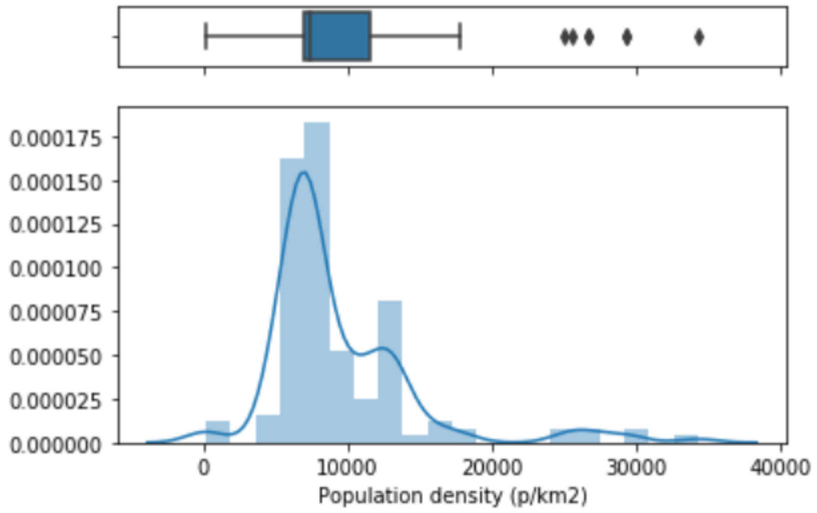
Discussion

Datasets

In regards with the data picked for this analysis, there are two variables that might input noise as some values are offset like outliers but they are real relevant point:

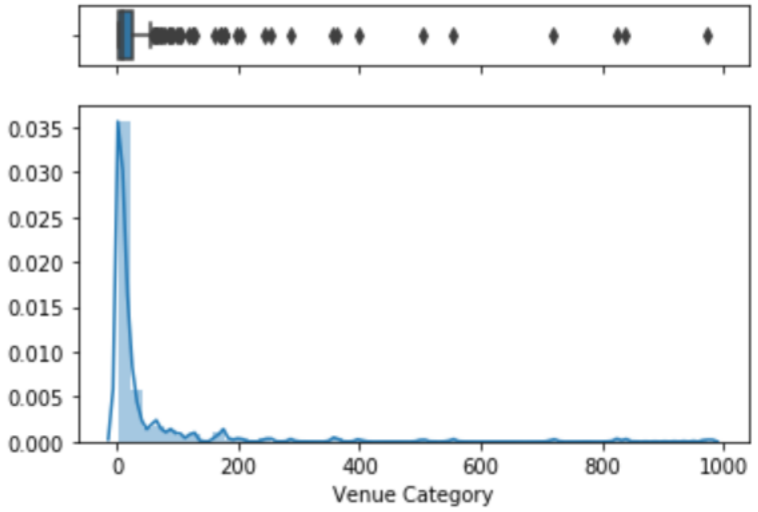
Population density

High population density neighborhoods

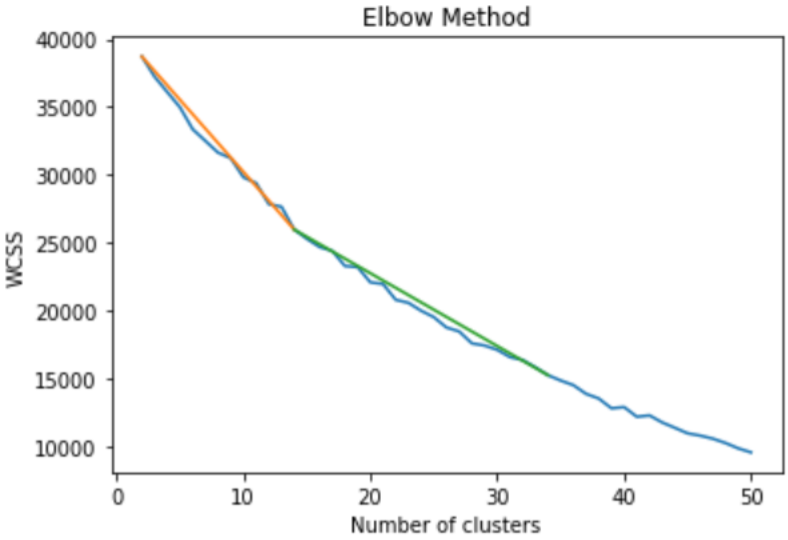


Venue category hits

There are lots of venue categories that appear only once, so this will definitely affect the one-hot encoding and therefore the k-means clustering.



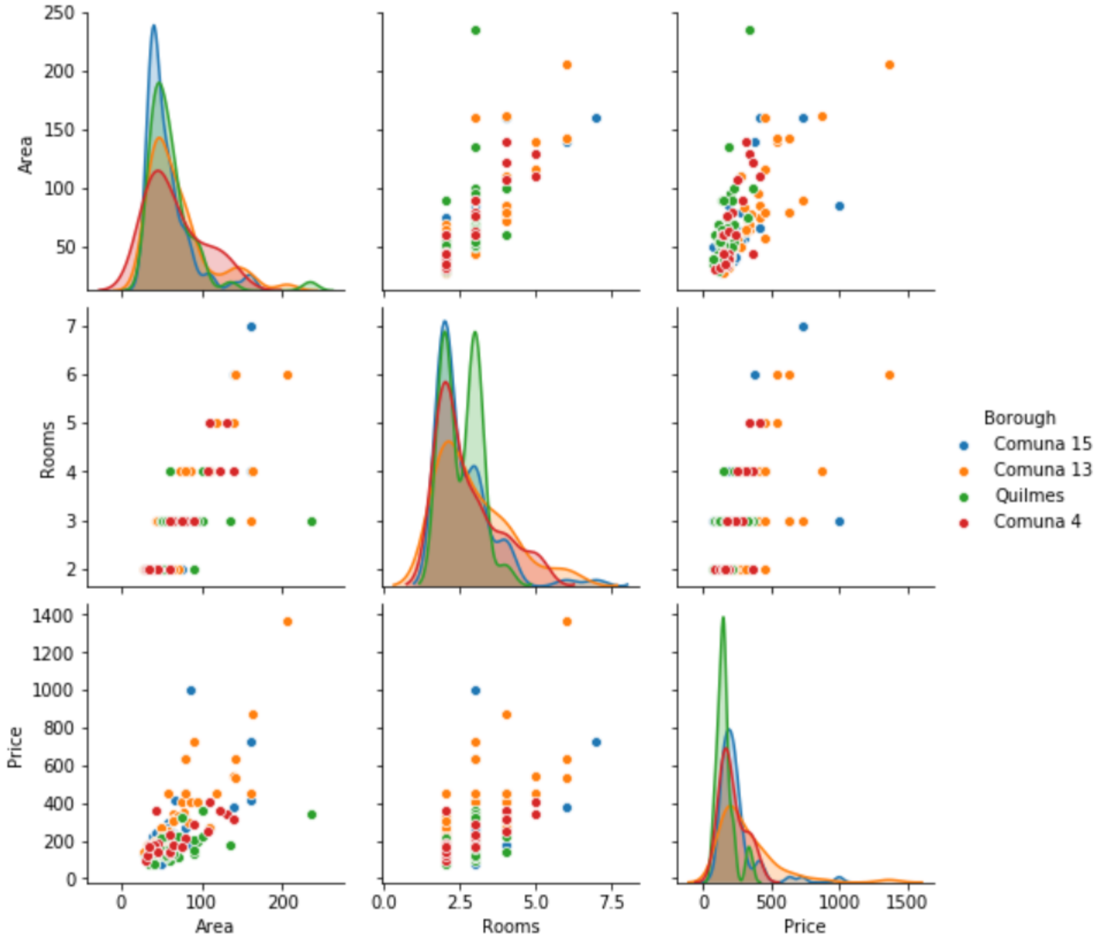
Clustering



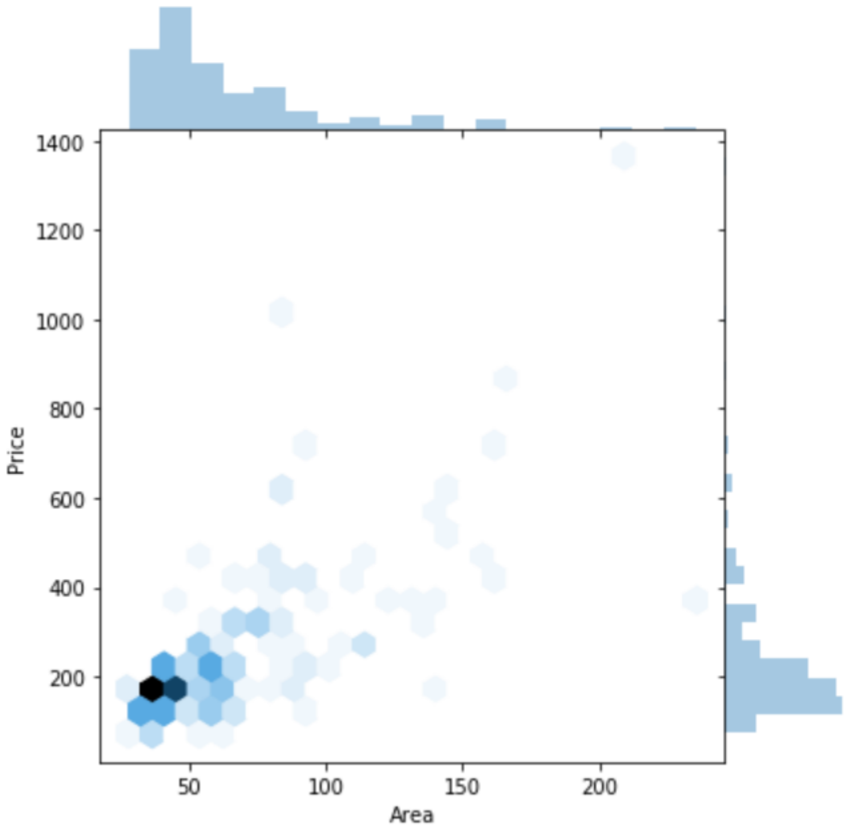
There is no clear elbow shape on the WCSS graph, but 13 clusters seems to be the most cost efficient option for this model

Housing results

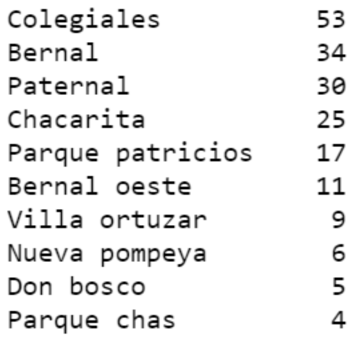
After the outliers filter the dataset looks good.



But most od the data point are concentrated in a small region



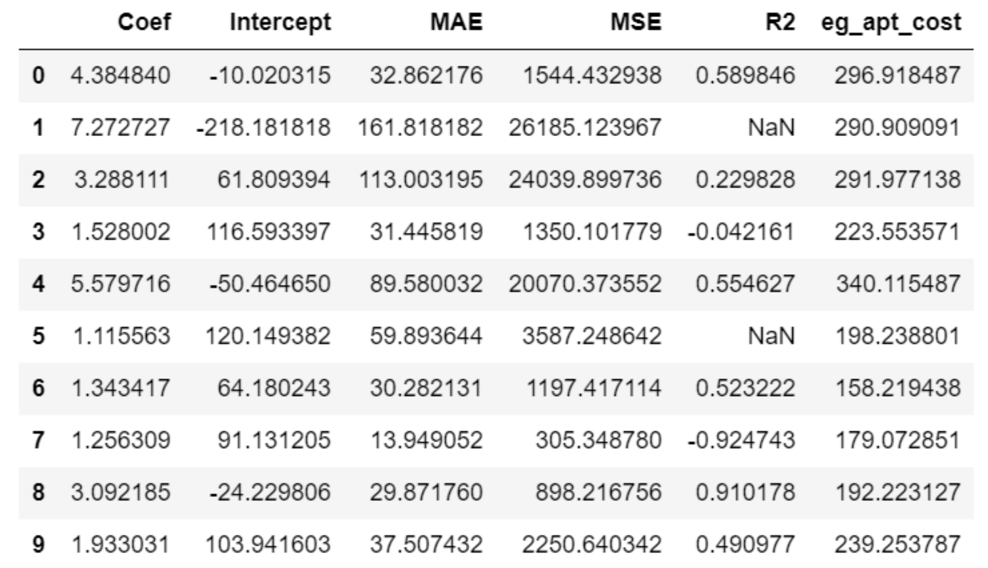
And some neighborhoods had too few hits for a regression to be efficient



```
Colegiales          53
Bernal              34
Paternal            30
Chacarita           25
Parque patricios    17
Bernal oeste        11
Villa ortuzar        9
Nueva pompeya        6
Don bosco            5
Parque chas          4
```

Linear regression

As expected from the low sample quantity, some of the models have a high error.



	Coef	Intercept	MAE	MSE	R2	eg_apt_cost
0	4.384840	-10.020315	32.862176	1544.432938	0.589846	296.918487
1	7.272727	-218.181818	161.818182	26185.123967	NaN	290.909091
2	3.288111	61.809394	113.003195	24039.899736	0.229828	291.977138
3	1.528002	116.593397	31.445819	1350.101779	-0.042161	223.553571
4	5.579716	-50.464650	89.580032	20070.373552	0.554627	340.115487
5	1.115563	120.149382	59.893644	3587.248642	NaN	198.238801
6	1.343417	64.180243	30.282131	1197.417114	0.523222	158.219438
7	1.256309	91.131205	13.949052	305.348780	-0.924743	179.072851
8	3.092185	-24.229806	29.871760	898.216756	0.910178	192.223127
9	1.933031	103.941603	37.507432	2250.640342	0.490977	239.253787

Conclusion

Despite some error due to low data samples, this methodology has proven to be efficient in solving the proposed problem. I should work properly in a more dense destination city with more rentals (at the moment Buenos Aires is going through a rental and economic crisis) and where Foursquare were more popular and widely used.