Capstone Project

Introduction/Business Problem

Buenos Aires is the capital of Argentina and one of the best city to live in Latin America , it has a population of at least 2.89 million people spread in 48 neighborhoods and 13 communes, in this project we will study Buenos Aires because I've the intention of moving over the this year.

To move over there, first we will do an analysis based on the cost of rent and if the area is considered safe or not, here we look for the minimum cost of rent having the lowest crime rate possible, this is very difficult due to safe areas are expensive and there is a budget to consider. Another aspect to keep in mind is the amount of jobs sources close to the neighborhood and the ease to move around the city.

A preselection will be done just based on crime rate and rent price (Cluster Analysis), then of the remaining neighborhoods we seek the ones with the most job offers and transportation methods.

Data

Data Description

For the Map the of the neighborhoods of Buenos Aires we used the geojson data provided by the portal of the mayoralty of Buenos Aires data(https://data.buenosaires.gob.ar/dataset/barrios), the rest of the data was also extract from this portal:

- Rent (https://www.estadisticaciudad.gob.ar/eyc/?p=62993)
- Crimes (https://e-8ef01e7270.cognitiveclass.ai/download/data/delitos.csv), to handle missing data, we seek alternative in real state web pages.

Exploratory Data Analysis

First we check the column names and the first rows of the dataset; consistency in the names of the columns is required for future merges between data frames, so if a name is not proper we change it. After calculate the amount of missing data, if is significant that column must be dropped, also other columns that won't be use in the analysis have to be drop.

	Neighbourhood	commune	lat	lng
0	CHACARITA	15	-34.588372	-58.454175
1	PATERNAL	15	-34.597421	-58.468665
2	VILLA CRESPO	15	-34.598834	-58.442722
3	VILLA DEL PARQUE	11	-34.604247	-58.490677
4	ALMAGRO	5	-34.609227	-58.421745

There data frame has 48 neighborhoods, let's use the coordinates of Buenos Aires to center the map, and the coordinates of the neighborhoods to point where they are.

The coordinates of Buenos Aires are -34.6075616, -58.437076.

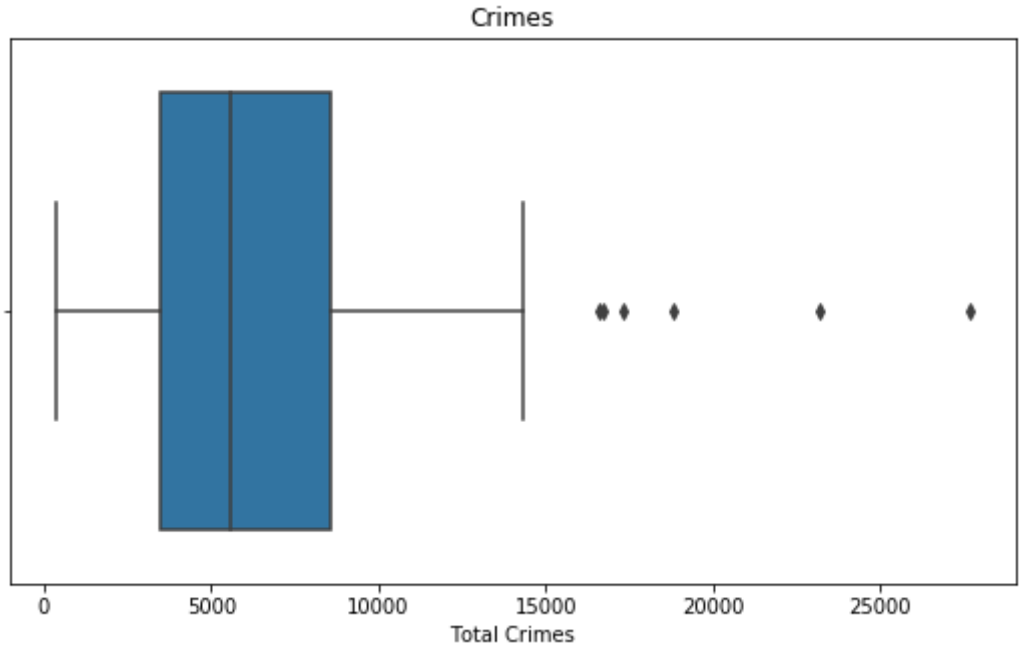
Check the first rows of the crime dataset.

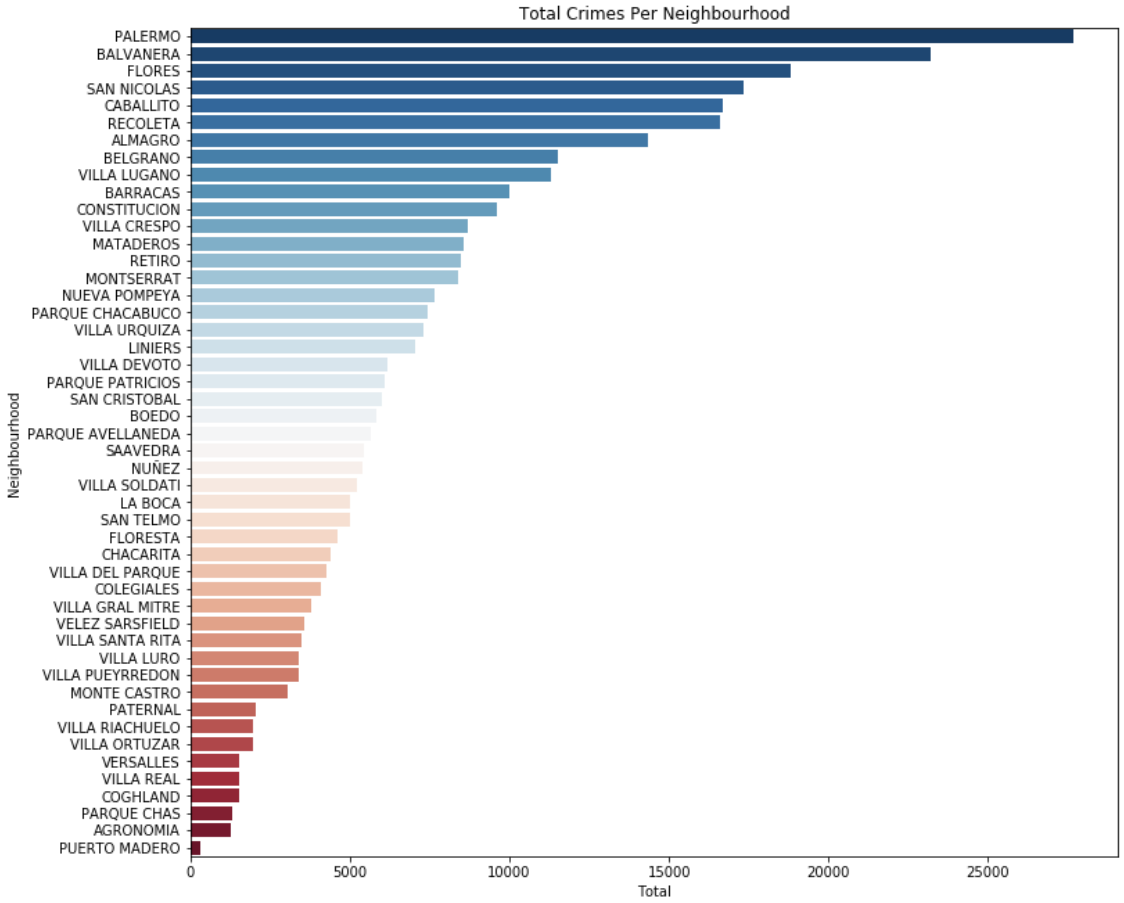
	id	comuna	barrio	latitud	longitud	fecha	hora	uso_arma	uso_moto	lugar	origen_dato	tipo_delito	cantidad_vehiculos	cantidad_victimas
0	68313	Comuna 7	FLORES	-34.647598	-58.442600	2016-01-31	19:50:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Homicidio Doloso	0	0
1	130524	Comuna 9	LINIERS	-34.639801	-58.529701	2016-01-31	12:00:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	0
2	130582	Comuna 9	LINIERS	-34.638802	-58.527500	2016-01-31	08:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	0
3	130586	Comuna 10	VILLA REAL	-34.618000	-58.527302	2016-01-31	20:30:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Hurto Automotor	0	0
4	130615	Comuna 9	LINIERS	-34.645599	-58.526501	2016-01-31	20:24:00	SIN USO DE ARMA	SIN MOTO	NaN	NaN	Robo (Con violencia)	0	0

To know which neighborhoods is the dangerous, we group by the column "Neighborhood" and apply the size method to count how many crimes were committed, then we rename the new column, and estimate a new column call "Pct" that is the percentage of crimes, we sort the values from most dangerous to least.

	Total	Pct
count	48.000000	48.000000
mean	7250.375000	0.020658
std	5905.397202	0.016825
min	338.000000	0.000963
25%	3482.250000	0.009922
50%	5549.000000	0.015810
75%	8590.250000	0.024475
max	27689.000000	0.078891

Neighbourhood	Total	Pct
PALERMO	27689	0.078891
BALVANERA	23195	0.066086
FLORES	18792	0.053542
SAN NICOLAS	17319	0.049345
CABALLITO	16703	0.047590



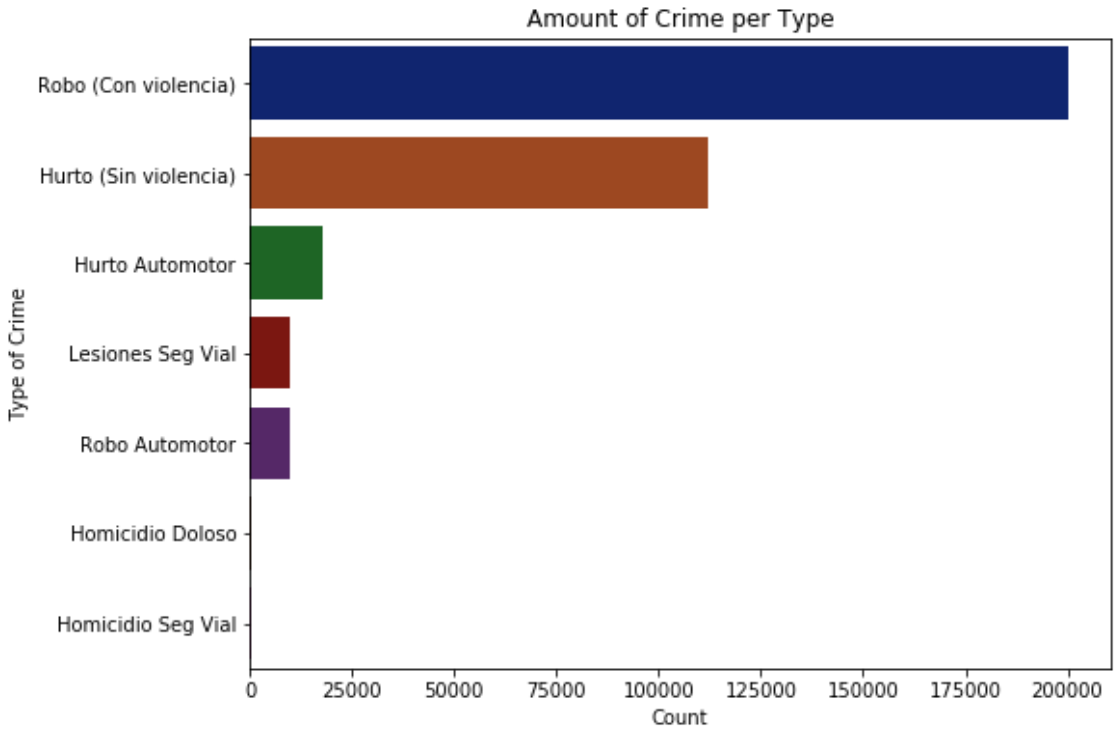


We can see that Palermo is the most dangerous neighborhood in Buenos Aires with 27698 crimes and Puerto Madero is the least dangerous with just

338, around 90 times less than Palermo. The median, mean and standard deviation of crimes per neighborhood are 5549, 7250, and , 5905.

After determine which are the most dangerous and safe neighborhoods, it's important to know which type of crimes are committed the most, to do this we group by the column "type_crime" and apply the size method and sort in descending order.

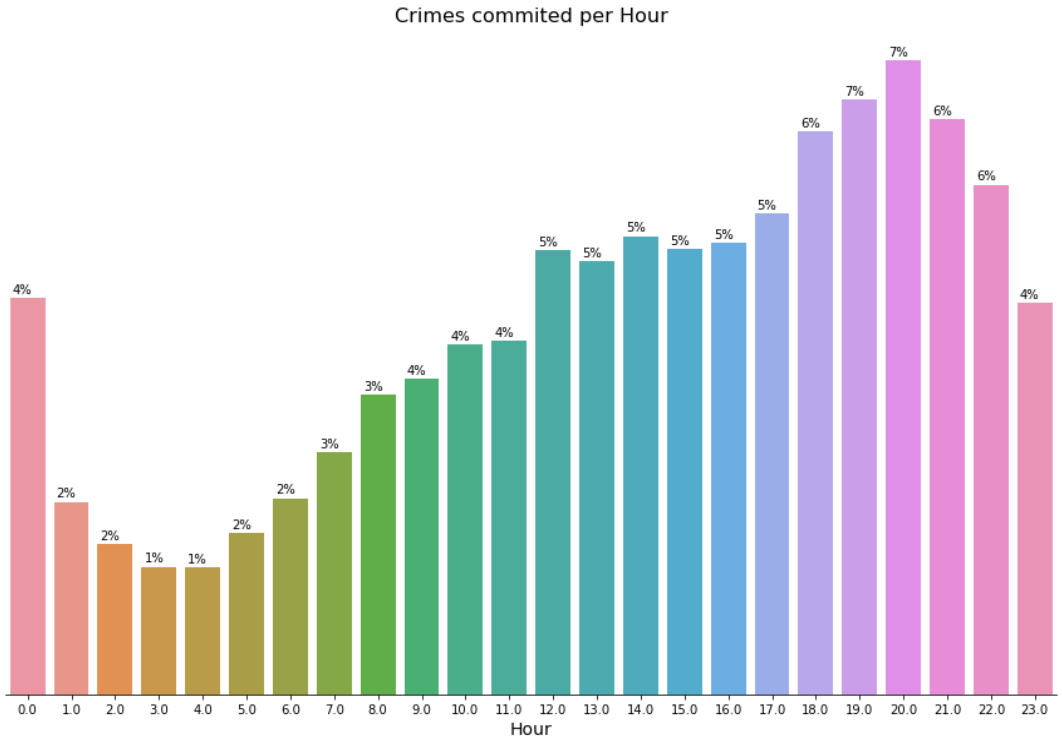
	type_crime	Total	Pct
0	Robo (Con violencia)	200374	0.570899
1	Hurto (Sin violencia)	112301	0.319964
2	Hurto Automotor	18010	0.051313
3	Lesiones Seg Vial	9833	0.028016
4	Robo Automotor	9793	0.027902
5	Homicidio Doloso	405	0.001154
6	Homicidio Seg Vial	264	0.000752



From the previous data we can see that "Theft with violence" is the most common crime and "Homicide road Safety" is the least common, we can also say that the rate of murder("Homicide" per 100000 inhabitants) in Buenos Aires is low compare to other main cities in Latin America like Brazil, Honduras, Venezuela, Mexico, among others.

Finally we look at which are the most commons crimes in each neighborhood, we repeat the process of group by, in this case two columns "type_crime" and “Neighborhood" and applying the size method.

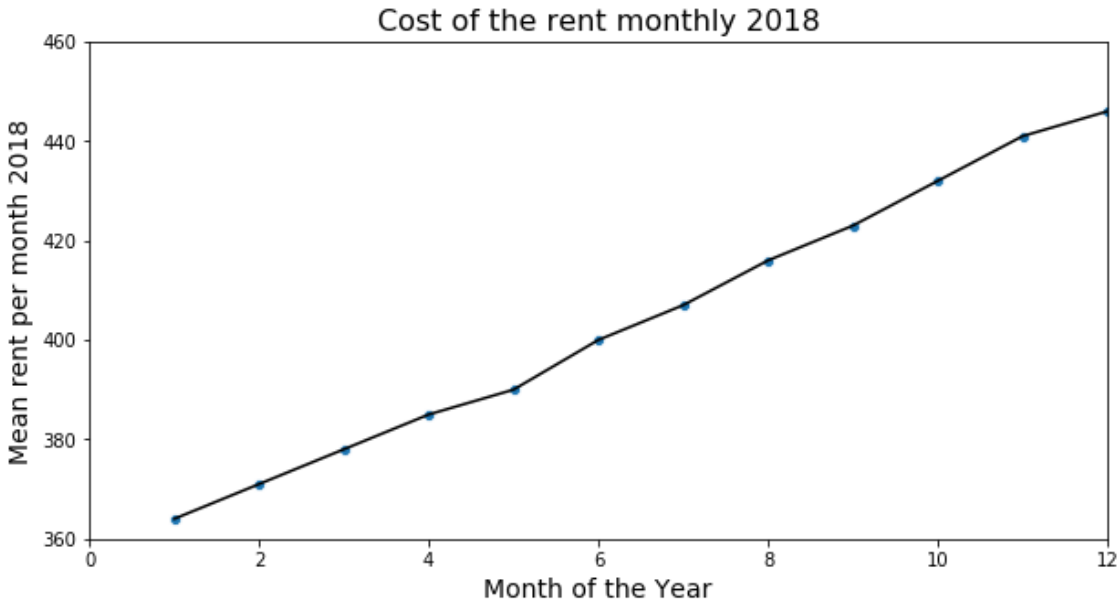
To see which hour is the most dangerous in the city, first check if the column type is appropriate to work with, if isn't we convert to a date type, then extract the hour and assign it in to a list, then we create a new column call "hour", finally we repeat the group by process and method size.

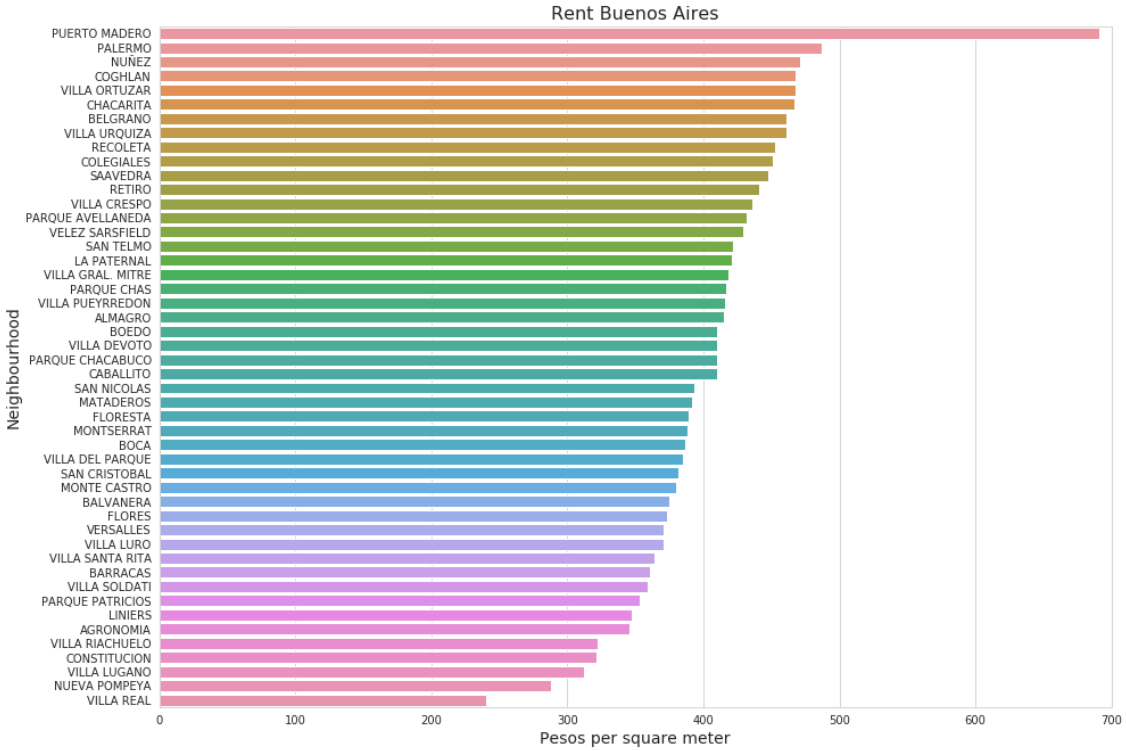


We can see that 64% of the crimes are happened in the second half of the day, and between six and 10 p.m. 32% of the crimes are committed, clearly

the preferred time is by the night, when usually more people go out for distraction.

Now we check the change of the price of rent in the last year.



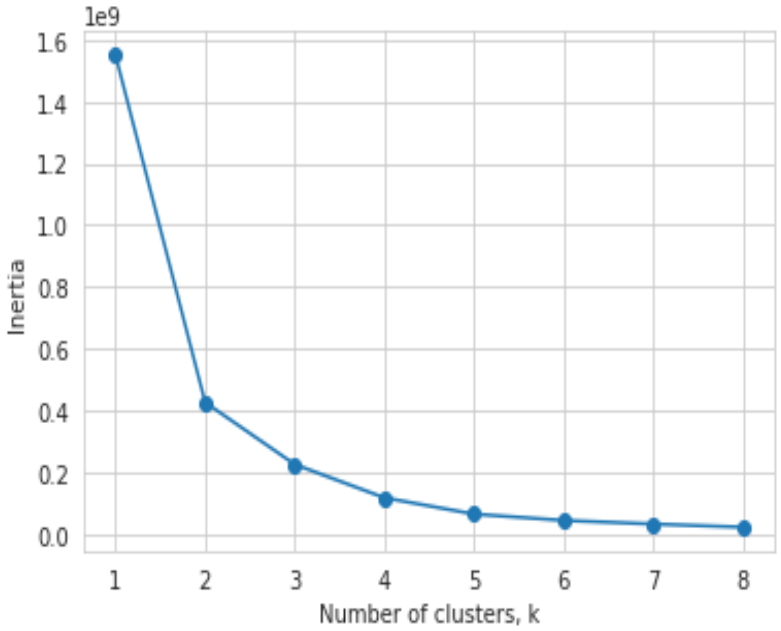


Here we can appreciate how all most statistical indicator have gone up from January to December, in the first chart we can see that the trend is up for all months.

Also we can see that Puerto Madero is by more than 200 pesos per square meter is the most exclusive neighborhood in all Buenos Aires, this area of the city is full of new and modern real estate development and is one of the most popular neighborhood for tourism, also happened to be one the safes test area in all the city.

Then we can say that the rest of the neighborhoods are between 500 and 300 pesos per square meter, 22 in the range of 499 to 400 and 23 in the range of 399 to 300.

	Neighbourhood	Total_Crimes	Psm	commune	lat	lng
0	PALERMO	27689	487.0	14	-34.573902	-58.422435
1	BALVANERA	23195	375.0	3	-34.609099	-58.403062
2	FLORES	18792	373.0	7	-34.636803	-58.458268
3	SAN NICOLAS	17319	393.0	1	-34.603666	-58.380515
4	CABALLITO	16703	410.0	6	-34.616825	-58.443603



	Neighbourhood	Total_Crimes	Psm	commune	lat	lng	Labels
0	PALERMO	27689	487.0	14	-34.573902	-58.422435	0
1	BALVANERA	23195	375.0	3	-34.609099	-58.403062	0
2	FLORES	18792	373.0	7	-34.636803	-58.458268	0
3	SAN NICOLAS	17319	393.0	1	-34.603666	-58.380515	0
4	CABALLITO	16703	410.0	6	-34.616825	-58.443603	0

	Labels	Total_Crimes	Psm	Count
0	0	13934.866667	388.266667	15
1	1	338.000000	691.000000	1
2	2	1525.750000	400.500000	4
3	3	3208.555556	382.222222	9
4	4	5963.615385	407.846154	13

Results

To classify the neighborhoods, first we merge the data frames containing the crimes, rent cost and locations of each neighborhood, then we drop the column "commune", and extract the values of these columns.

Is necessary apply a machine learning classification method, in this case KMeans, first the "Elbow Method" was applied, and select the number where the slope start to decrease her decline the most, in this case is 5 so that will be the number of cluster to use. After that sort out the neighborhood, and attach their label to the data frame and convert to a category type.

To obtain the characteristic of each category, the mean method was applied:

Category 0: There are 15 neighborhoods, that have a mean price per square meter of 405.67 and a mean rate crime of 6219 per year.

Category 1: There is 1 neighborhood, Palermo, it has a mean price per square meter of 691 (max rent price) and a mean rate crime of 338 (min total crime) per year.

Category 2: There are 4 neighborhoods, which have a mean price per square meter of 400.5 and a mean rate crime of 1526 per year.

Category 3: There are 9 neighborhoods, which have a mean price per square meter of 382.22 and a mean rate crime of 3209 per year.

Category 4: There are 13 neighborhoods, which have a mean price per square meter of 387.77 and a mean rate crime of 14867.15 per year.

The only "outlier" is Puerto Madero it has the highest rent and the lowest crime rate in all the city, then deviation in rent price for the rest of the clusters is just 133 pesos per square meter looking to the crime stats the range is much wider than for the rent, here the deviation is 5427 crimes, so it can be said that the crime rate is the biggest difference between the clusters, and it's a fundamental factor, also the rent prices are packed in big groups, they don't differ that much from neighborhood to neighborhood.

Finally a representation of the category is made with an interactive bar chart, which provides the characteristics of each cluster.

	Neighbourhood	Total_Crimes	Psm	commune	lat	lng	Labels	dist (km)
36	VILLA ORTUZAR	1960	468.0	15	-34.580973	-58.467651	2	9.262694
37	VERSALLES	1528	371.0	10	-34.630123	-58.522417	2	14.227842
39	PARQUE CHAS	1336	417.0	15	-34.585523	-58.479121	2	10.165282
40	AGRONOMIA	1279	346.0	15	-34.592945	-58.488673	2	10.902523
27	CHACARITA	4392	467.0	15	-34.588372	-58.454175	3	7.861156
29	COLEGIALES	4103	451.0	13	-34.574642	-58.450968	3	8.058815
30	VELEZ SARSFIELD	3562	429.0	10	-34.631361	-58.493277	3	11.659750
31	VILLA SANTA RITA	3505	364.0	11	-34.616193	-58.482957	3	10.398895
32	VILLA LURO	3414	371.0	10	-34.636413	-58.502729	3	12.646533
33	VILLA PUEYRREDON	3379	416.0	12	-34.582103	-58.503499	3	12.432692
34	MONTE CASTRO	3035	380.0	10	-34.619298	-58.506581	3	12.592036
35	VILLA RIACHUELO	1972	322.0	8	-34.691858	-58.463350	3	12.945927
38	VILLA REAL	1515	240.0	10	-34.619493	-58.526039	3	14.364754

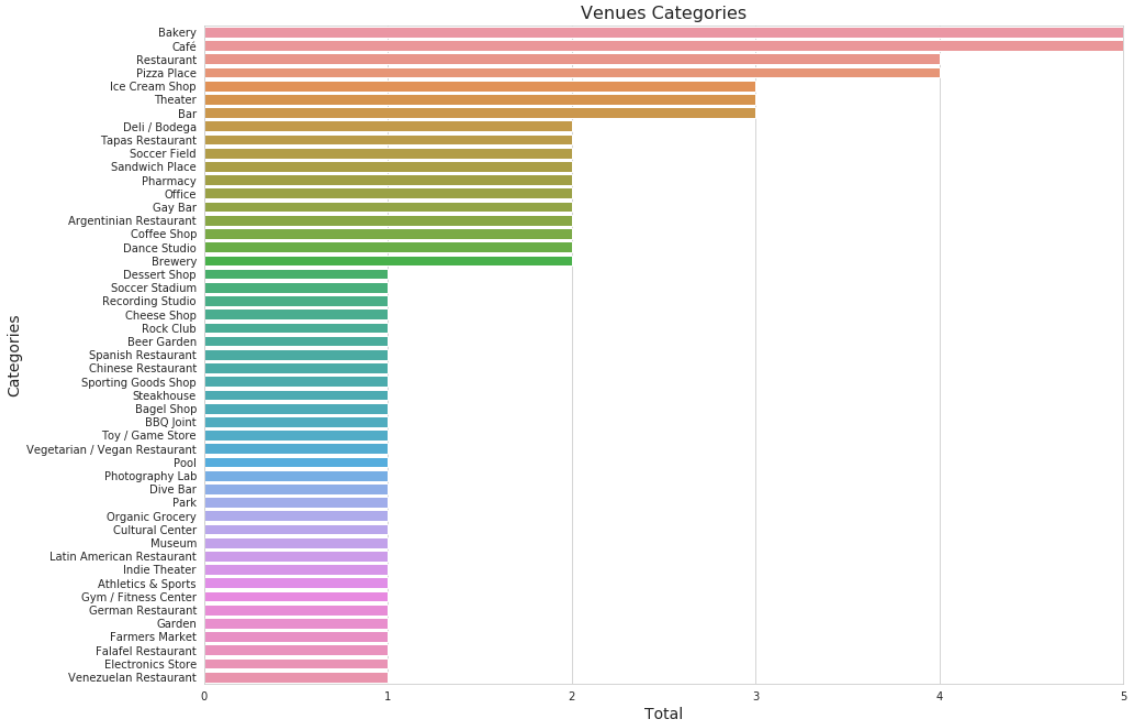
Then the cluster to be selected are the one we the lowest crime rate and price per square meter.

Chacarita is the closest neighborhood to the center of the city that is relative cheap and safe, in the top five closest to the downtown Chacarita is the one with the most crimes and higher rent, but in this case minimize distance is the main goal so this is the neighborhood selected to move.

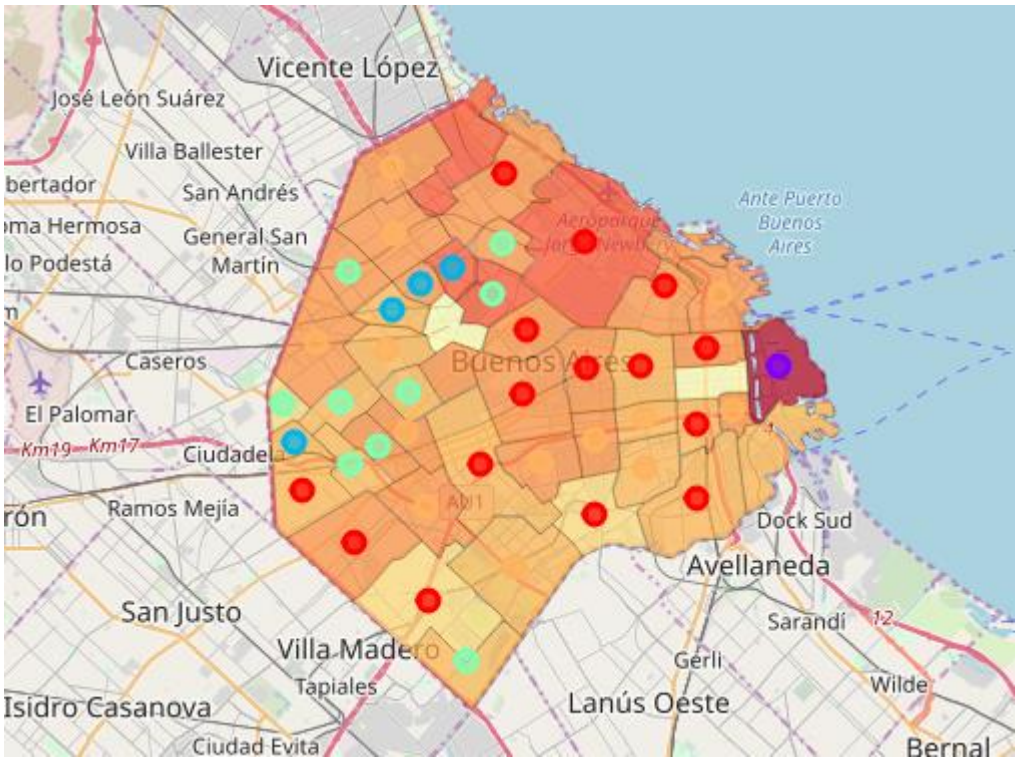
With the neighborhood coordinates start a search of the 100 venues within a radius of one kilometer.

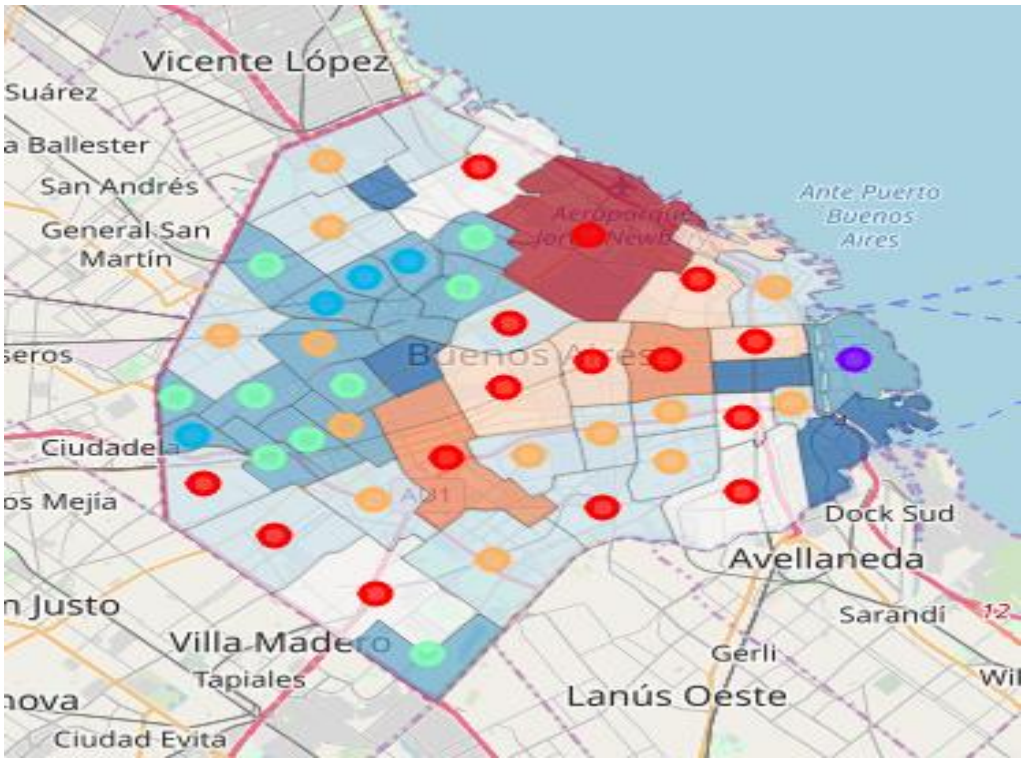
	name	categories	lat	lng
0	El Imperio de la Pizza	Pizza Place	-34.586890	-58.454967
1	Pizzería Santa María	Pizza Place	-34.587238	-58.454005
2	Santos 4040	Theater	-34.588822	-58.449863
3	Albamonte Ristorante	Argentinian Restaurant	-34.587803	-58.453075
4	Fábrica de Churros Olleros	Bakery	-34.586983	-58.453640

There are 79 venues close to Chacarita, in the next chart we can see which are the most common venues around Chacarita.



The neighborhood has all the basic store/locals that are needed pharmacy, restaurants, markers, bus stop, and swimming pool among others.





Discussion

One of the aspects that stand out the most is that the average rent in Buenos Aires increased 24% the last year, one of the main reason is the financial situation in Argentina where inflation is one of the highest in the world.

Most of the crimes were committed between 5 to 10 pm. so the law enforcement institutions should place more patrols at this hours in the most dangerous neighborhoods.

More variables can be used like public transportation, job index, or more specific matters depending on the person necessities, there are several neighborhoods that have similar characteristic to Chacarita and those can be contemplated to move.

The rent data is a bit ambiguous because the majority of the rent required advance deposits, so the prices are a bit higher.

Conclusion

Data Analysis is a great tool in this case we could detected which are the best neighborhoods to move, depending on the variables to analyze. It's important that government institutions apply these techniques to attack problems they have in their communities like insecurity, traffic, public transportation among others.

Also the information should be public in this case to make decision about which neighborhoods the best to live in or where a business can be started.