Optimum Location for a Restaurant in Medellin, Antioquia, Colombia

1. Introduction

Medellín is a city in Colombia that is divided in zones called "Comunas", each one of them has several neighborhoods associated with it. The neighborhoods that belong to the same comuna share some socioeconomic profiles related to people's purchasing power, education, mean of trasportation, etc.

When trying to find a location for a restaurant it is important to take into account people's ability to buy from the restaurant and the density of restaurants that the location has, because generally the more restaurants a zone has, the more people will recognize that specific location as a place to eat, but in contrast a very high restaurant density means that there is a lot of competition and rent prices for businesses can be higher than in zones with not so high restaurant density.

The goal of this project is to determine a great location for a restaurant in Medellín based in the number of restaurants and current density of restaurants in certain zones of the city, and a socioeconomic index of each zone.

2. Data

The first set of data to be used contains information about the neighborhoods of Medellín, containing: name and information about a corresponding socioeconomic index.

In colombia there is a socioeconomic index called "estrato" (stratum) which is associated with people's wealth. This indicator goes from 1 to 6, 1 meaning poverty (Low) and 6 meaning wealth (High).

1 = Low - Low
2 = Low
3 = Low - Medium
4 = Medium
5 = High - Medium
6 = High

The table containing neighborhoods information has the name of each neighborhood, the number of homes in that neighborhood associated with each one of the socioeconomics and the total of homes. The neighborhood id, which is used as index in this case, contains first the number of the comuna (1 – 16) to which the neighborhood belongs followed by a consecutive id for each of the neighborhoods in the comuna.

The data in this table was available in the url: https://www.medellin.gov.co/irj/go/km/docs/pccdesign/SubportaldelCiudadano_2/PlandeDesarrollo_0_17/Publicaciones/Shared%20Content/sisben/03_ViviendasComunaBarrioVeredaEstrato_Certificada_17122015.pdf, in the webpage of Medellin's government.

The second dataset contains information for each comuna. It is available in the url: https://www.medellin.gov.co/irj/portal/medellin?NavigationTarget=navurl://40245fdc67f729e064c3ca24924bea6c (again the Medellin's goverment webpage).

The report in that webpage contains the names and ID of each comuna and the superficial extension in km^2 of each comuna, which is used used to stablish a radius in which locations (restaurants) are going to be searched for each comuna.

The third and last piece of data used is available in: https://www.medellin.gov.co/irj/go/km/docs/pccdesign/SubportaldelCiudadano_2/PlandeDesarrollo/ObservatoriodePolticasPblicas/Shared%20Content/Boletin%20Mercado%20Inmobiliario%20Trimestre%202%20de%202014.pdf.

The last report contains basic info about the real state market in Medellin, from there the average rent price in each comuna is used.

3. Methodology

After the datasets were imported and cleaned some analysis was done to try to better understand Medellin.

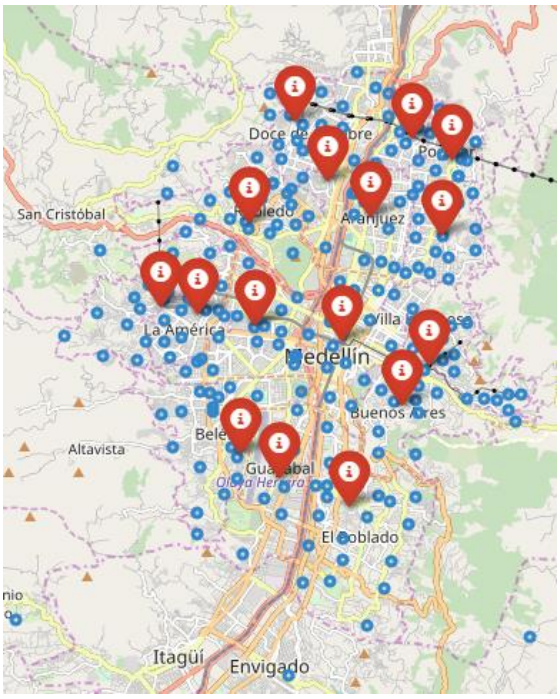
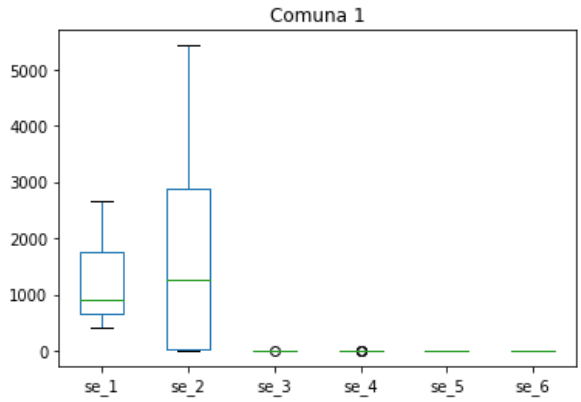
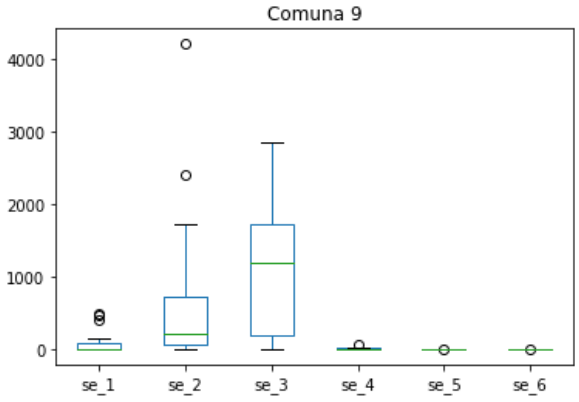
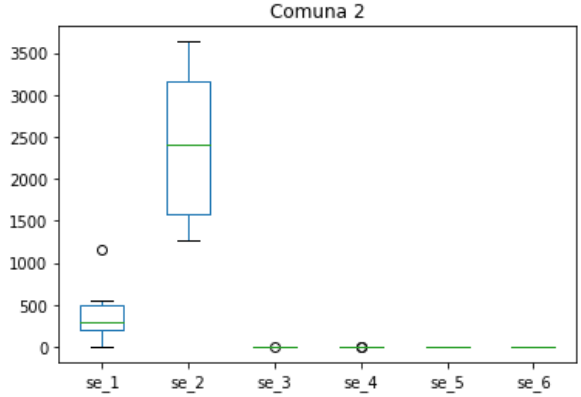
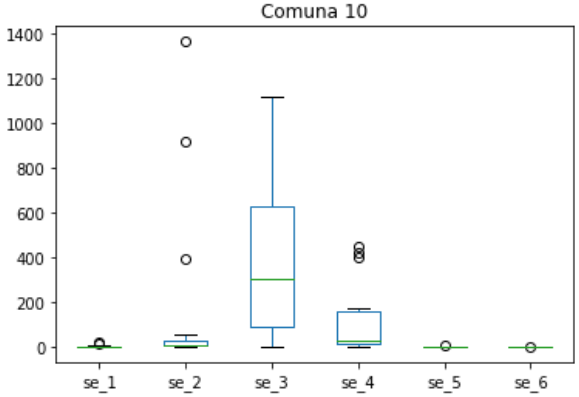


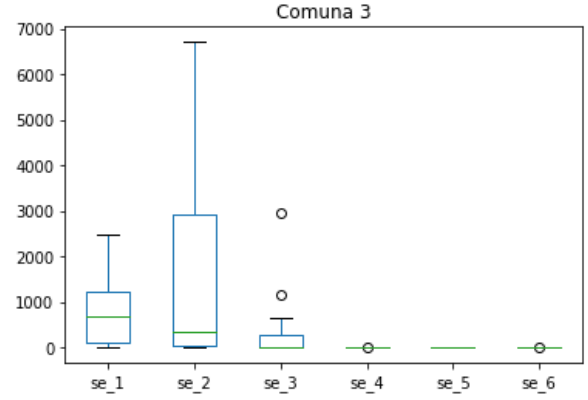
Fig 1. Map of Medellin with neighborhoods and comunas.

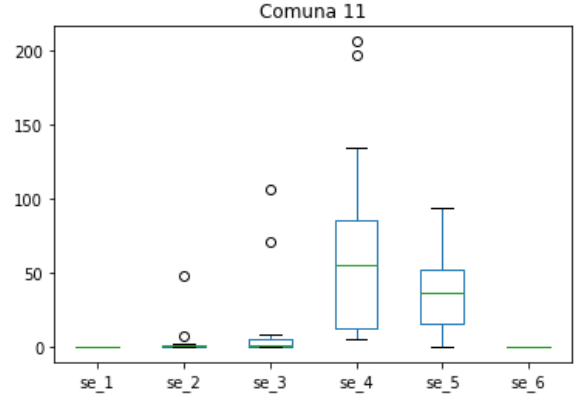

Comuna 1
5000
4000
3000
2000
1000
0
se_1
se_2
se_3
se_4
se_5
se_6

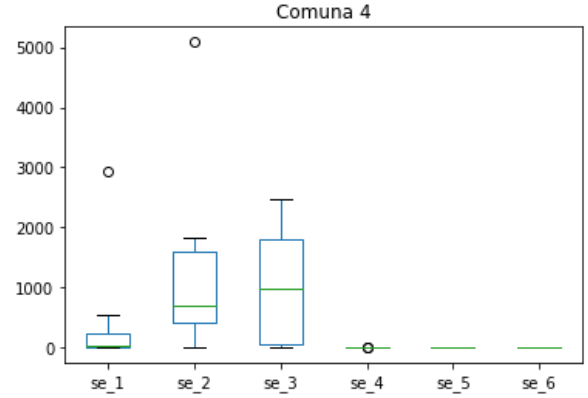

Comuna 9
4000
3000
2000
1000
0
se_1
se_2
se_3
se_4
se_5
se_6

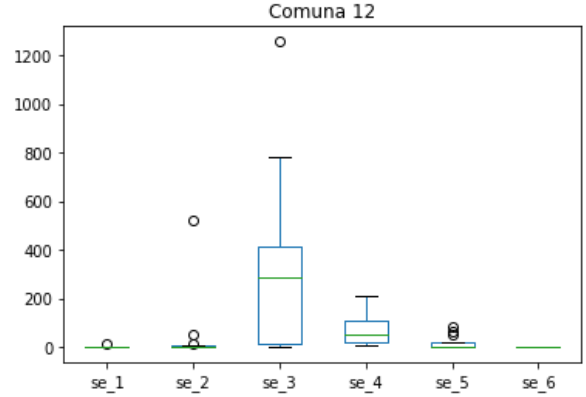

Comuna 2
3500
3000
2500
2000
1500
1000
500
0
se_1
se_2
se_3
se_4
se_5
se_6


Comuna 10
1400
1200
1000
800
600
400
200
0
se_1
se_2
se_3
se_4
se_5
se_6


Comuna 3
7000
6000
5000
4000
3000
2000
1000
0
se_1
se_2
se_3
se_4
se_5
se_6


Comuna 11
200
150
100
50
0
se_1
se_2
se_3
se_4
se_5
se_6


Comuna 4
5000
4000
3000
2000
1000
0
se_1
se_2
se_3
se_4
se_5
se_6


Comuna 12
1200
1000
800
600
400
200
0
se_1
se_2
se_3
se_4
se_5
se_6

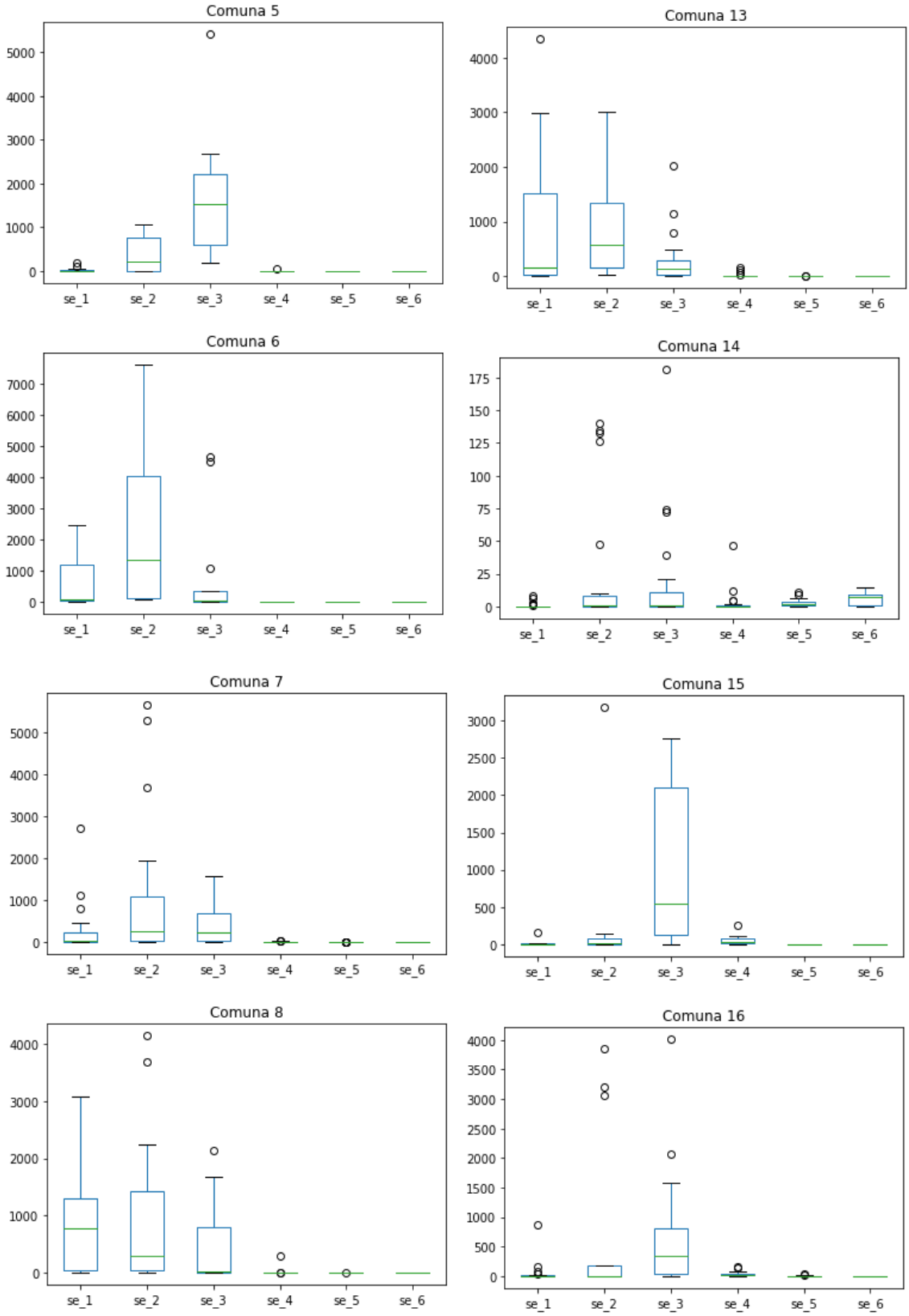


Fig 2. Distribution of socioeconomic index per neighborhood in each comuna.

In figure 1 we can see the map of Medellin with the blue circles being the neighborhoods and the red markers the comunas.

In figure 2 the socioeconomic index distribution can be seen for each comuna. The data plotted in those boxplots represent the summary of the number of homes associated with each socioeconomic in each one of the neighborhoods belonging to a comuna.

Comuna 14 is El Poblado, where Medellin's welthiest live. But the boxplot shows some points having a high number of homes with socioeconomics of 1 and 2 which is not normal. Those values affect the mean socioeconomic of the comuna but do not help to represent the socioeconomic profile of the comuna in a correct way so data for socioeconomic 1 was ignored and neighborhoods having more than 20 homes with a socioeconomic of 2 in El Poblado were ignored too.

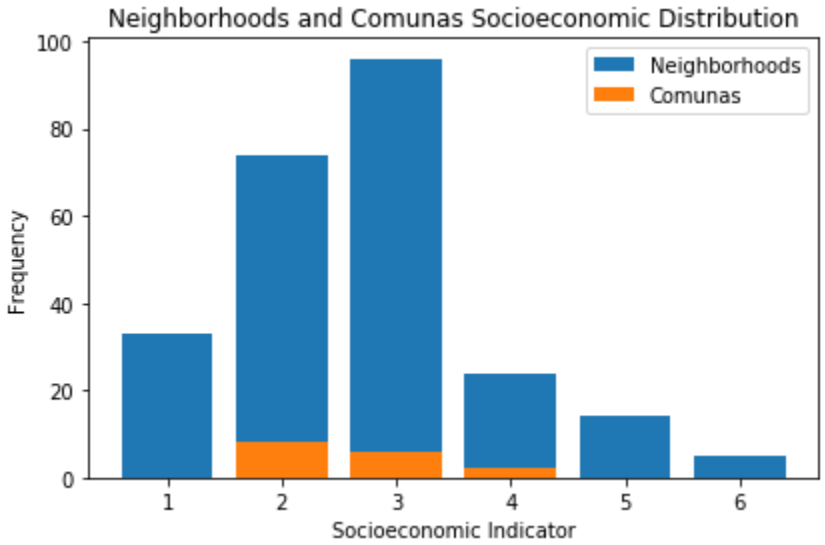


Fig 3. Distribution of the socioeconomic index in Medellin's neighborhoods and comunas.

From the information available on the socioeconomic index per neighborhood a mean socioeconomic index was calculated for each one of the neighborhoods and comunas and the distribution of those values can be seen in figure 3.

Figure 3 shows how the majority of homes in Medellin are located in zones associated with a Low – Low to Low – Medium socioeconomic and that there is only a small fraction of homes with a high socioeconomic index.

The same goes for the distribution of the comunas, the majority of homes have low socioeconomic indexes and only a fraction reach a medium socioeconomic. Also, it is possible to note that in average not a single comuna reaches a high socioeconomic index in Medellin.

After doing a brief description of the data, the Foursquare API was used to get the number of restaurants associated with each comuna and the restaurant density was calculated.

$$restaurant_{density} = \frac{restaurant\_count}{comuna\_extenson\ [km^2]}\ [=]\ restaurants\ per\ km^2$$

id	comuna	count	restaurant_density	se	avg_rent
4	Aranjuez	6	1.229508	2	0.5
7	Robledo	12	1.268499	2	0.6
8	Villa Hermosa	6	1.048951	2	0.4
9	Buenos Aires	8	1.322314	3	0.4
10	La Candelaria	44	5.978261	3	0.6
11	Laureles Estadio	62	8.378378	4	1.0
12	La America	23	5.793451	3	0.7
14	El Poblado	66	4.576976	4	1.3
15	Guayabal	46	6.310014	3	0.6
16	Belen	29	3.273138	3	0.8

Table 1. Summary of comunas data to be used for clustering.

Table 1 shows a summary of the key data used to describe each comuna. The table contains the count that represents all restaurants or food places obtained using the API call, the restaurant density as explained before, the socioeconomic index calculated at the beginning and the average rent price in each comuna expressed in Colombian pesos millions.

id	comuna	count	restaurant_density	se	avg_rent	count_norm	rest_density_norm	se_norm	avg_rent_norm	score
14	El Poblado	66	4.576976	4	1.3	1.000000	0.481351	1.0	1.000000	1.096270
11	Laureles Estadio	62	8.378378	4	1.0	0.933333	1.000000	1.0	0.666667	1.040000
15	Guayabal	46	6.310014	3	0.6	0.666667	0.717800	0.5	0.222222	0.599116
10	La Candelaria	44	5.978261	3	0.6	0.633333	0.672537	0.5	0.222222	0.576730
16	Belen	29	3.273138	3	0.8	0.383333	0.303460	0.5	0.444444	0.491803
12	La America	23	5.793451	3	0.7	0.283333	0.647322	0.5	0.333333	0.476131
7	Robledo	12	1.268499	2	0.6	0.100000	0.029954	0.0	0.222222	0.134880
9	Buenos Aires	8	1.322314	3	0.4	0.033333	0.037297	0.5	0.000000	0.120793
4	Aranjuez	6	1.229508	2	0.5	0.000000	0.024635	0.0	0.111111	0.049371
8	Villa Hermosa	6	1.048951	2	0.4	0.000000	0.000000	0.0	0.000000	0.000000

Table 2. Summary of comunas data with normalized columns and corresponding score.

Table 2 shows the same summary showed in table 1 but with five new columns. Four of them represent the same values of the count, restaurant density, socioeconomic index and average rent columns but normalized. The fifth column has a weighted score for each comuna calculated from the normalized values. To calculate the score more importance was given to restaurant count and average rent.

Having reviewed and described the available data a final DataFrame is created to be used in the clustering algorithm. In this case the K-Means cluster algorithm is used to cluster the comunas and find which of them are similar.

In order to get the optimum number of clusters (K) the elbow method was used and the result is shown in figure 4.

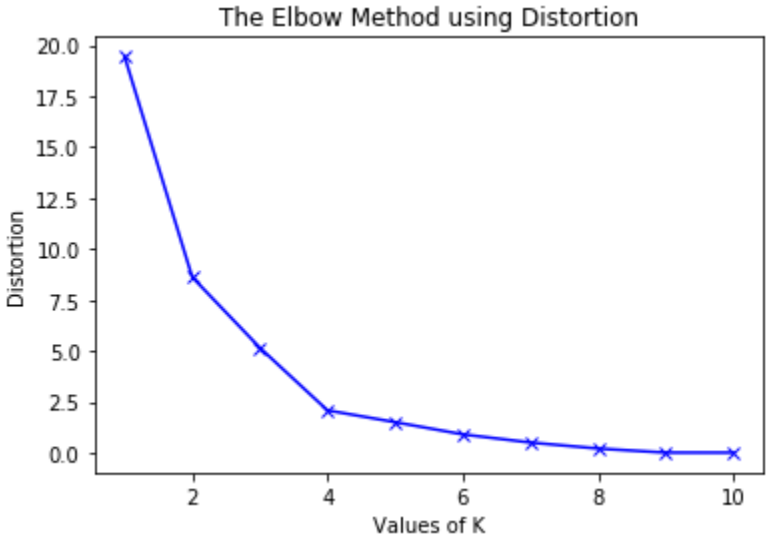


Fig 4. Elbow graph used to determine optimum K value.

4. Results

From the graph showing the results of the distortion for each number of K from 1 to 10, it can be noted that the optimum number of clusters for which the error is not significantly lower after it, is 4.

So 4 is used as the number of clusters and the results of the clustering algorithm are plotted in figure 7. Given that there are four features to use in the clustering algorithm, principal component analysis was done for dimensional reduction to be able to graph the results from the clustering.

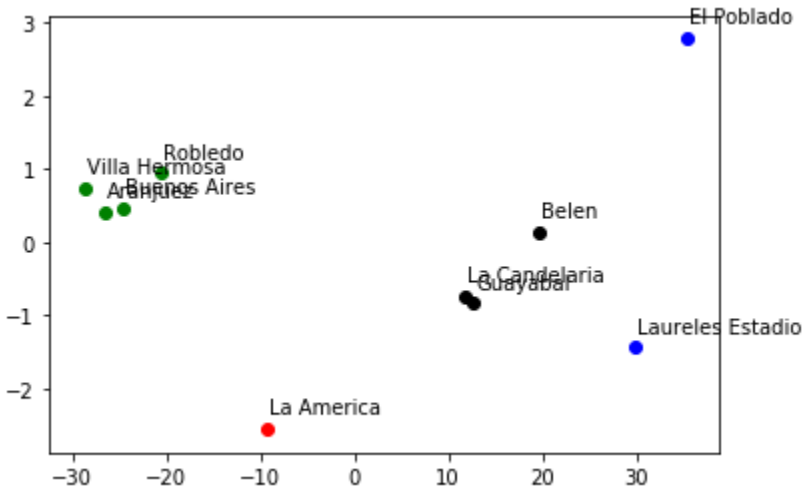


Fig 5. Clusters calculated with K-means.

Clusters can be described as follows:

- Villa Hermosa – Robledo – Aranjuez – Buenos Aires: comunas with low average rent price, low to medium socioeconomic index, low restaurant density and low restaurant count.
- La America: comuna with low – medium average rent price, medium socioeconomic index, medium restaurant density and medium restaurant count.
- La Candelaria – Guayabal – Belen: comunas with low to médium average rent price, medium socioeconomic index, medium restaurant density and medium – high restaurant count.
- El Poblado – Laureles Estadio: high average rent price, high socioeconomic index, medium to high restaurant density and high restaurant count.

5. Discussion

id	count	restaurant_density	se	avg_rent	count_norm	rest_density_norm	se_norm	avg_rent_norm	kmeans_labels	score	comuna
14	68	4.715673	4	1.3	1.000000	0.523028	1.0	1.000000	1	1.104606	El Poblado
11	62	8.378378	4	1.0	0.906250	1.000000	1.0	0.666667	1	1.029167	Laureles Estadio
16	52	5.869074	3	0.8	0.750000	0.673228	0.5	0.444444	3	0.712423	Belen
15	45	6.172840	3	0.6	0.640625	0.712786	0.5	0.222222	3	0.587696	Guayabal
10	44	5.978261	3	0.6	0.625000	0.687447	0.5	0.222222	3	0.576378	La Candelaria
12	23	5.793451	3	0.7	0.296875	0.663380	0.5	0.333333	0	0.484759	La America
7	12	1.268499	2	0.6	0.125000	0.074123	0.0	0.222222	2	0.153714	Robledo
9	8	1.322314	3	0.4	0.062500	0.081131	0.5	0.000000	2	0.141226	Buenos Aires
4	6	1.229508	2	0.5	0.031250	0.069046	0.0	0.111111	2	0.070754	Aranjuez
8	4	0.699301	2	0.4	0.000000	0.000000	0.0	0.000000	2	0.000000	Villa Hermosa

Table 3. Summary of comunas after clustering.

Table 3 show all the information about the comunas including the label obtained through the clustering algorithm. Table 3 is sorted by score and it can be noted that clusters correspond to groups of scores because in some way the score represents a summary of all the features of each comuna.

From the clustering results it can be concluded that the more attractive comunas to locate a restaurant in Medellín are those in cluster with label 1, namely, El Poblado and Laureles Estadio.

Using that result and taking into account only neighborhoods with socioeconomic index of 5 and 6, table 4 shows the list of neighborhoods in which it is better to locate a restaurant in Medellin according to the data used and the analysis done here.

id	comuna	name	latitude	longitude	se
1406	14	Las Lomas No.1	6.210818	-75.561149	6
1413	14	El Diamante No.2	6.285250	-75.585219	6
1420	14	Astorga	6.210516	-75.573519	6
1422	14	La Aguacatala	6.199026	-75.577885	6
1423	14	Santa Maria de los Angeles	6.190038	-75.579340	6
1102	11	Suramericana	6.255346	-75.584433	5
1105	11	Los Conquistadores	6.242355	-75.581523	5
1108	11	Laureles	6.251594	-75.588799	5
1109	11	Las Acacias	6.240964	-75.601167	5
1110	11	La Castellana	6.240500	-75.607715	5
1113	11	Estadio	6.260128	-75.593892	5
1404	14	Castropol	6.216140	-75.566970	5
1405	14	Lalinde	6.212935	-75.567698	5
1407	14	Las Lomas No.2	6.209378	-75.558436	5
1414	14	El Castillo	6.190502	-75.572791	5
1415	14	Los Balsos 2	6.195581	-75.569881	5

Table 4. Summary of neighborhoods corresponding to the top 2 comunas.

6. Conclusion

The best conclusion that can be drawn from this exercise is that this information is useful for someone trying to make an educated and informed decision about the location for a new restaurant, and supposing the same (or similar) information and data is available for any city, this can be replicated in lots of places to choose locations in a more efficient manner.

For future work on similar projects, at least for the city of Medellin, it would be great to have more detailed information about each neighborhood in order to be more specific with the neighborhoods that result as great locations at the end of the analysis and filtering even more the results with the use of data.