FINAL REPORT

Capstone Project – Optimal location for a Colombian restaurant in Madrid

Introduction:

Madrid is the capital and most populous city of Spain. The city has almost 3.3 million inhabitants and a metropolitan area population of approximately 6.5 million. It is the second-largest city in the European Union (EU), surpassed only by Berlin, and its monocentric metropolitan area is the second-largest in the EU, surpassed only by Paris. The municipality covers 604.3 km^2.

As the capital city of Spain, the city has attracted many immigrants from around the world, with most of the immigrants coming from Latin American countries. In 2020, around 76 % of the registered population was Spain-born, while, regarding the foreign-born population (24 %), the bulk of it relates to the Americas (around 16 % of the total population), and a lesser fraction of the population is born in other European, Asian and African countries.
With its diverse culture, comes diverse food items. There are many restaurants in Madrid, each belonging to different categories like Spanish, Argentinian, Peruvian, Ecuadorian, Chinese.

Problem description

The main task is to find the best possible location or the most optimal, for a Colombian restaurant in Madrid. To accomplish this task, an analytical approach will be used, based on advanced Machine Learning techniques and Data Analysis, mainly Clustering and some Data Visualization techniques.

During the process of analysis, several data transformations will be performed to find the best possible data format for the Machine Learning model. Once the data is set up and prepared, a modeling process will be carried out, and this statistical analysis will provide the best possible places to locate the Colombian restaurant.

Data Presentation

The data that will be used to develop this project is based on two sites:

- The Foursquare API, this data will be accessed via Python, and used to obtain the most common venues per neighborhood in the city of Madrid. It is possible to have a taste of how the city's venues are distributed, what the most common places are for leisure, and for the most part, it will provide an idea about the preferences of the inhabitants.

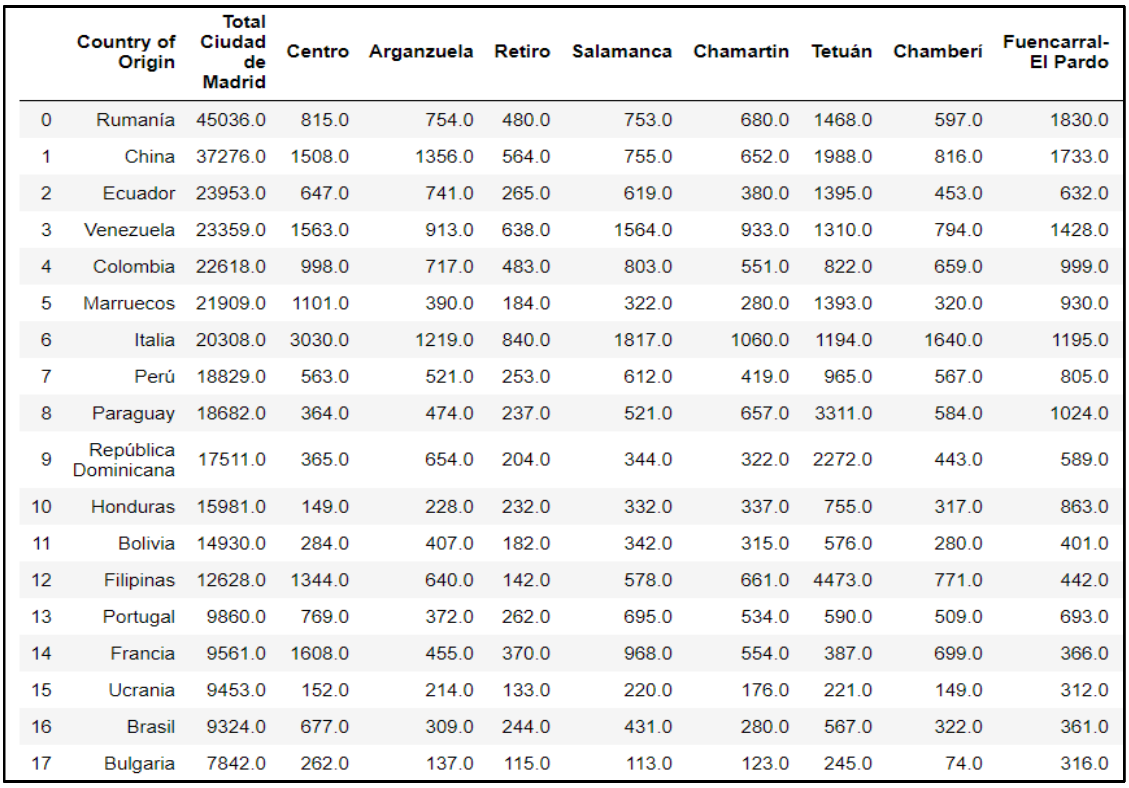
- The Madrid City Hall's Web Portal, this site provides several data sources of great utility to solve this problem. The files are provided in Excel format and are built over a statistical exploitation. The data contains updated information about the immigrant population per nationality. This data will be analyzed in such a way that one could determine the best location of a new venue/restaurant/other based on nationality. For make things simpler, it will be assumed for this exercise that people preferences vary according to their nationality, and people from one specific country will be more attracted to place that matches the environment and culture of their own countries, rather than the ones from foreign countries.

Methodology

The methodology used to approach this problem includes some statistical exploration of the data and visualizations. The main Machine Learning technique involved in the development of this project is Clustering where the K-Means algorithm was used via Python.

The main problem was how to obtain the necessary data to build a constructive approach. For the most part, to solve these kinds of optimal business location problems, bunch of consumer's data are needed; but for this case and for the sake of simplicity, the focus was mainly on the population nationality. A study was carried out over the inhabitants of Madrid, and it was assumed that the population from a certain country would prefer restaurants based on their national country and food, rather than restaurants from other countries or; specially when it comes to immigrant populations and certainly would like to usually have a taste of their food and original culture. In the end, it is not only about the food, it is also about having a piece of the country in the place you are currently based.

With all these things under consideration, the main goal to efficiently solve this problem was to define the target population; additionally, find the areas where this population is residing, and finally, examine the venues and restaurants in the chosen area to see if our product could work. Below, it is outlined a screenshot of the data used for this capstone project.

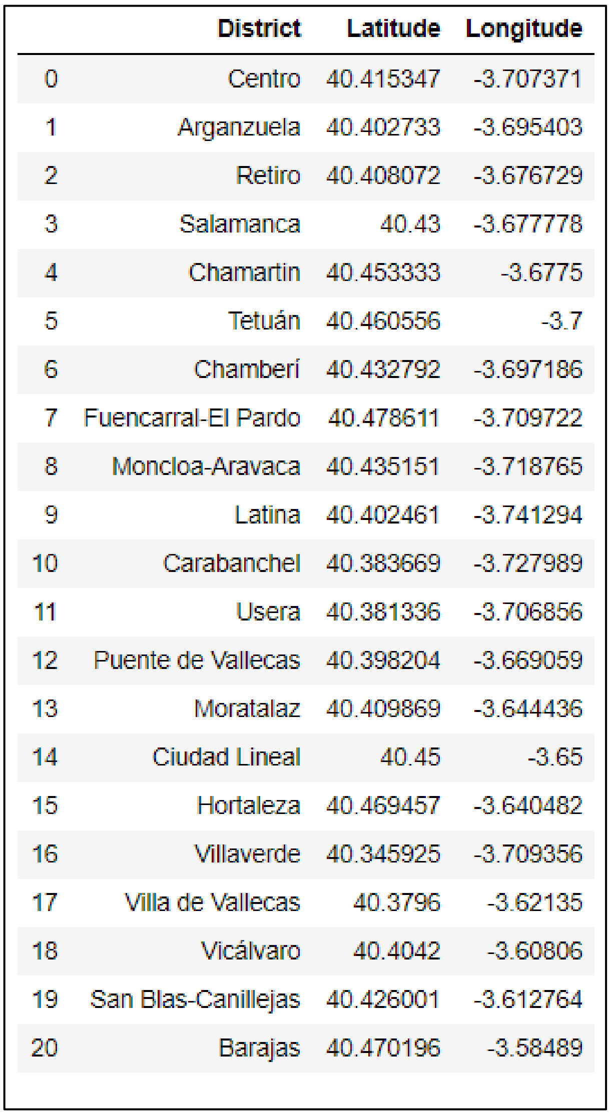


	Country of Origin	Total Ciudad de Madrid	Centro	Arganzuela	Retiro	Salamanca	Chamartin	Tetuán	Chamberí	Fuencarral-El Pardo
0	Rumanía	45036.0	815.0	754.0	480.0	753.0	680.0	1468.0	597.0	1830.0
1	China	37276.0	1508.0	1356.0	564.0	755.0	652.0	1988.0	816.0	1733.0
2	Ecuador	23953.0	647.0	741.0	265.0	619.0	380.0	1395.0	453.0	632.0
3	Venezuela	23359.0	1563.0	913.0	638.0	1564.0	933.0	1310.0	794.0	1428.0
4	Colombia	22618.0	998.0	717.0	483.0	803.0	551.0	822.0	659.0	999.0
5	Marruecos	21909.0	1101.0	390.0	184.0	322.0	280.0	1393.0	320.0	930.0
6	Italia	20308.0	3030.0	1219.0	840.0	1817.0	1060.0	1194.0	1640.0	1195.0
7	Perú	18829.0	563.0	521.0	253.0	612.0	419.0	965.0	567.0	805.0
8	Paraguay	18682.0	364.0	474.0	237.0	521.0	657.0	3311.0	584.0	1024.0
9	República Dominicana	17511.0	365.0	654.0	204.0	344.0	322.0	2272.0	443.0	589.0
10	Honduras	15981.0	149.0	228.0	232.0	332.0	337.0	755.0	317.0	863.0
11	Bolivia	14930.0	284.0	407.0	182.0	342.0	315.0	576.0	280.0	401.0
12	Filipinas	12628.0	1344.0	640.0	142.0	578.0	661.0	4473.0	771.0	442.0
13	Portugal	9860.0	769.0	372.0	262.0	695.0	534.0	590.0	509.0	693.0
14	Francia	9561.0	1608.0	455.0	370.0	968.0	554.0	387.0	699.0	366.0
15	Ucrania	9453.0	152.0	214.0	133.0	220.0	176.0	221.0	149.0	312.0
16	Brasil	9324.0	677.0	309.0	244.0	431.0	280.0	567.0	322.0	361.0
17	Bulgaria	7842.0	262.0	137.0	115.0	113.0	123.0	245.0	74.0	316.0

This data contains information about the immigrant population in Madrid within each Neighborhood. The main features are the country of origin which outlines the precedence of the inhabitants by country living in each neighborhood. So, with this is mind, it is already possible to have an idea of where our target population is located.

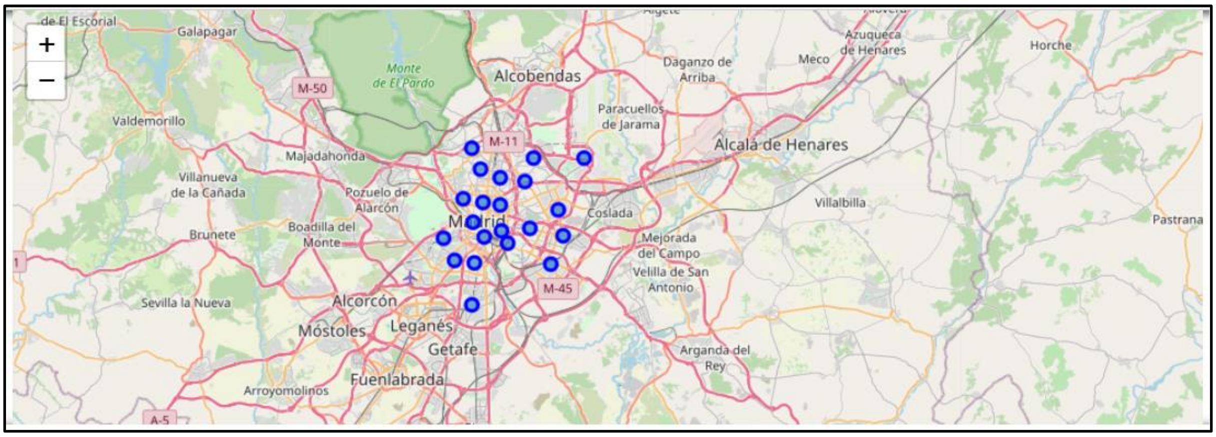
In this project, the idea is to open a Colombian restaurant in the city. With further analysis, this question will have an answer; nevertheless, this task could not be achieved only working with this raw data. It was also needed information about the most common venues in these neighborhoods, besides of the type of population residing on the different neighborhoods. It was also needed to determine somehow in what measure these neighborhoods were different or similar between them.

To make it possible, the Foursquare API was used to get the data regarding the venues in each neighborhood; First, it was first necessary to transform the raw data into a scheme the Foursquare API was capable to handle. Basically, the coordinates of each neighborhood were needed.

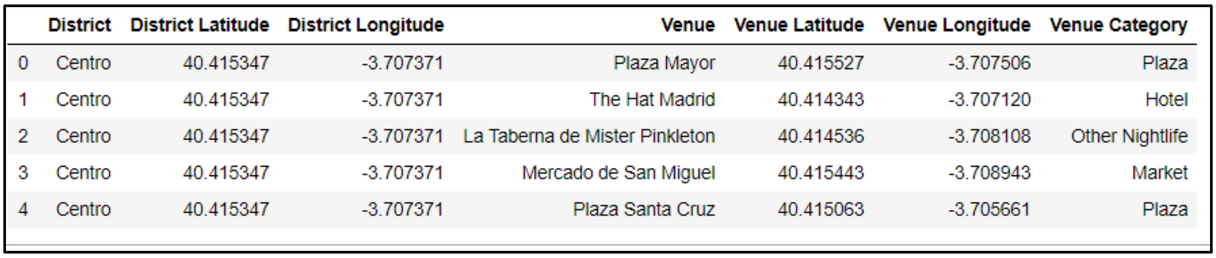


	District	Latitude	Longitude
0	Centro	40.415347	-3.707371
1	Arganzuela	40.402733	-3.695403
2	Retiro	40.408072	-3.676729
3	Salamanca	40.43	-3.677778
4	Chamartin	40.453333	-3.6775
5	Tetuán	40.460556	-3.7
6	Chamberí	40.432792	-3.697186
7	Fuencarral-El Pardo	40.478611	-3.709722
8	Moncloa-Aravaca	40.435151	-3.718765
9	Latina	40.402461	-3.741294
10	Carabanchel	40.383669	-3.727989
11	Usera	40.381336	-3.706856
12	Puente de Vallecas	40.398204	-3.669059
13	Moratalaz	40.409869	-3.644436
14	Ciudad Lineal	40.45	-3.65
15	Hortaleza	40.469457	-3.640482
16	Villaverde	40.345925	-3.709356
17	Villa de Vallecas	40.3796	-3.62135
18	Vicálvaro	40.4042	-3.60806
19	San Blas-Canillejas	40.426001	-3.612764
20	Barajas	40.470196	-3.58489

Later, the data was transformed into a format readable by the Foursquare API to get the information about the venues. The districts were then plotted into a map of Madrid, so it was possible to have an idea of their geographical location

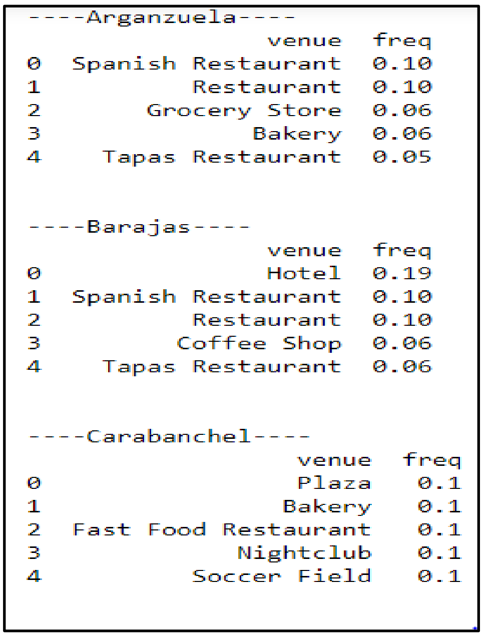


The following step was to obtain the nearby venues by district, together with their respective coordinates



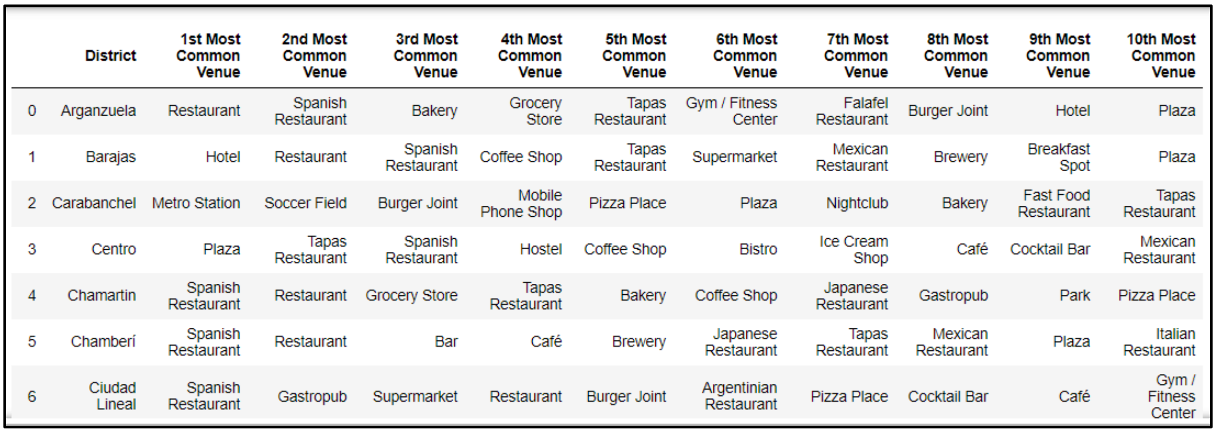
	District	District Latitude	District Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Centro	40.415347	-3.707371	Plaza Mayor	40.415527	-3.707506	Plaza
1	Centro	40.415347	-3.707371	The Hat Madrid	40.414343	-3.707120	Hotel
2	Centro	40.415347	-3.707371	La Taberna de Mister Pinkleton	40.414536	-3.708108	Other Nightlife
3	Centro	40.415347	-3.707371	Mercado de San Miguel	40.415443	-3.708943	Market
4	Centro	40.415347	-3.707371	Plaza Santa Cruz	40.415063	-3.705661	Plaza

Looking at this sample, it is possible to see the venues names, their coordinates and the category of each one. The results are ordered by district. This is a vital step in the segmentation process, since all the important data about the venues are obtained from this dataset. Once the venues per district were obtained, it was then needed to look at the mean occurrence of each venue:



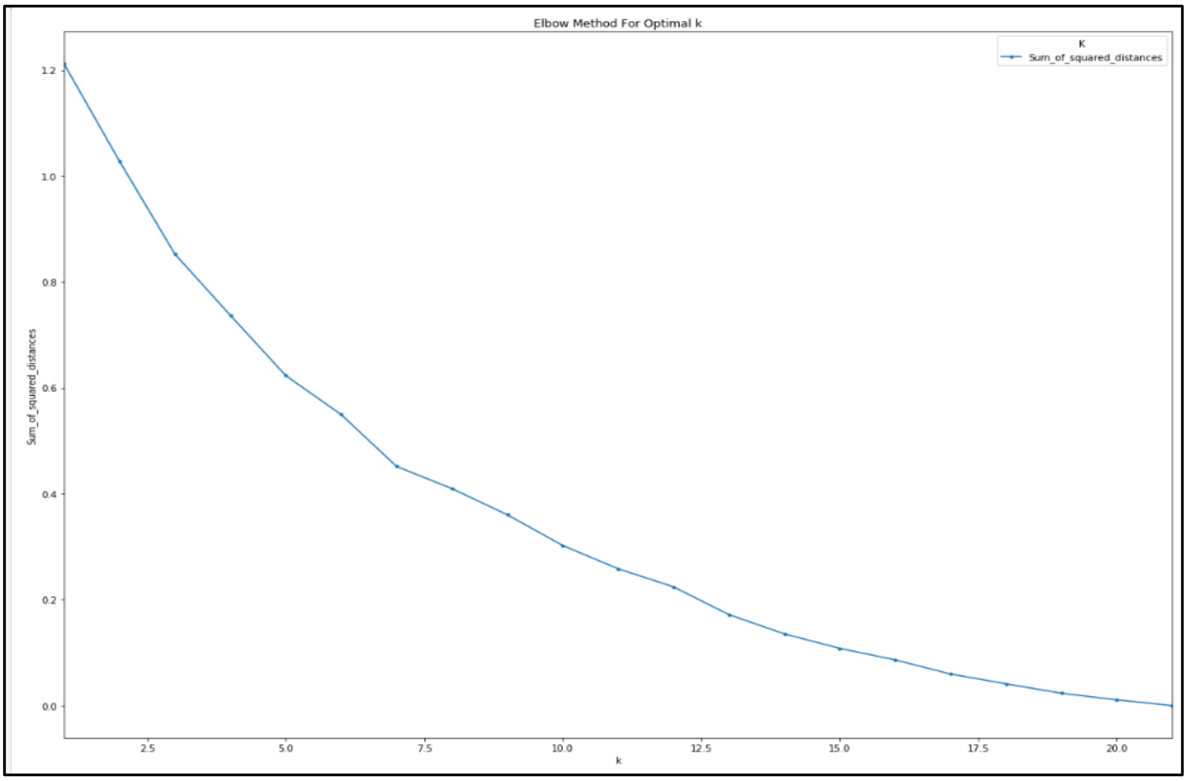
```
----Arganzuela----
               venue  freq
0  Spanish Restaurant  0.10
1          Restaurant  0.10
2       Grocery Store  0.06
3              Bakery  0.06
4    Tapas Restaurant  0.05


----Barajas----
               venue  freq
0               Hotel  0.19
1  Spanish Restaurant  0.10
2          Restaurant  0.10
3         Coffee Shop  0.06
4    Tapas Restaurant  0.06


----Carabanchel----
                 venue  freq
0                Plaza   0.1
1               Bakery   0.1
2  Fast Food Restaurant   0.1
3            Nightclub   0.1
4         Soccer Field   0.1
```

With the information shown above is possible to know which the most common venues are. Below, you can have a glimpse regarding the most common venues of each district

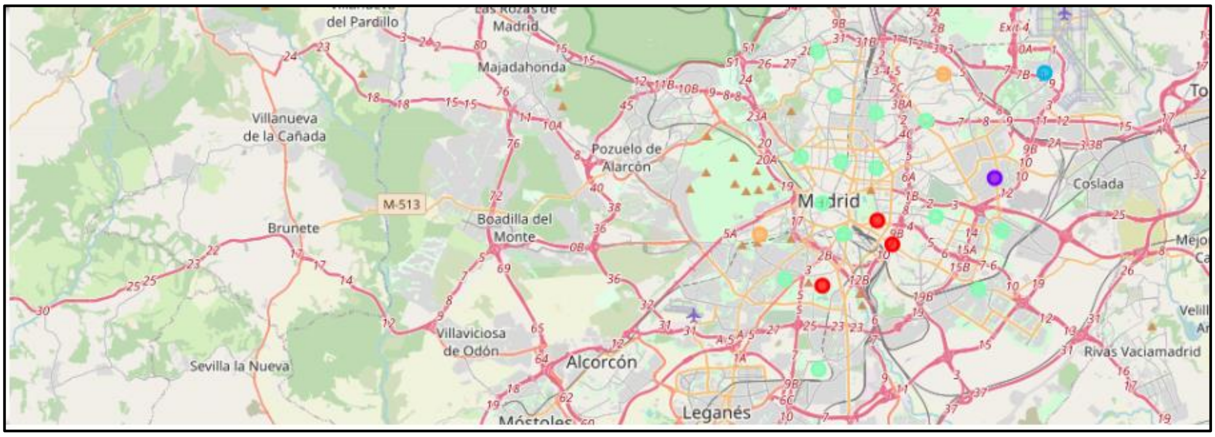


	District	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
0	Arganzuela	Restaurant	Spanish Restaurant	Bakery	Grocery Store	Tapas Restaurant	Gym / Fitness Center	Falafel Restaurant	Burger Joint	Hotel	Plaza
1	Barajas	Hotel	Restaurant	Spanish Restaurant	Coffee Shop	Tapas Restaurant	Supermarket	Mexican Restaurant	Brewery	Breakfast Spot	Plaza
2	Carabanchel	Metro Station	Soccer Field	Burger Joint	Mobile Phone Shop	Pizza Place	Plaza	Nightclub	Bakery	Fast Food Restaurant	Tapas Restaurant
3	Centro	Plaza	Tapas Restaurant	Spanish Restaurant	Hostel	Coffee Shop	Bistro	Ice Cream Shop	Café	Cocktail Bar	Mexican Restaurant
4	Chamartin	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Coffee Shop	Japanese Restaurant	Gastropub	Park	Pizza Place
5	Chamberí	Spanish Restaurant	Restaurant	Bar	Café	Brewery	Japanese Restaurant	Tapas Restaurant	Mexican Restaurant	Plaza	Italian Restaurant
6	Ciudad Lineal	Spanish Restaurant	Gastropub	Supermarket	Restaurant	Burger Joint	Argentinian Restaurant	Pizza Place	Cocktail Bar	Café	Gym / Fitness Center

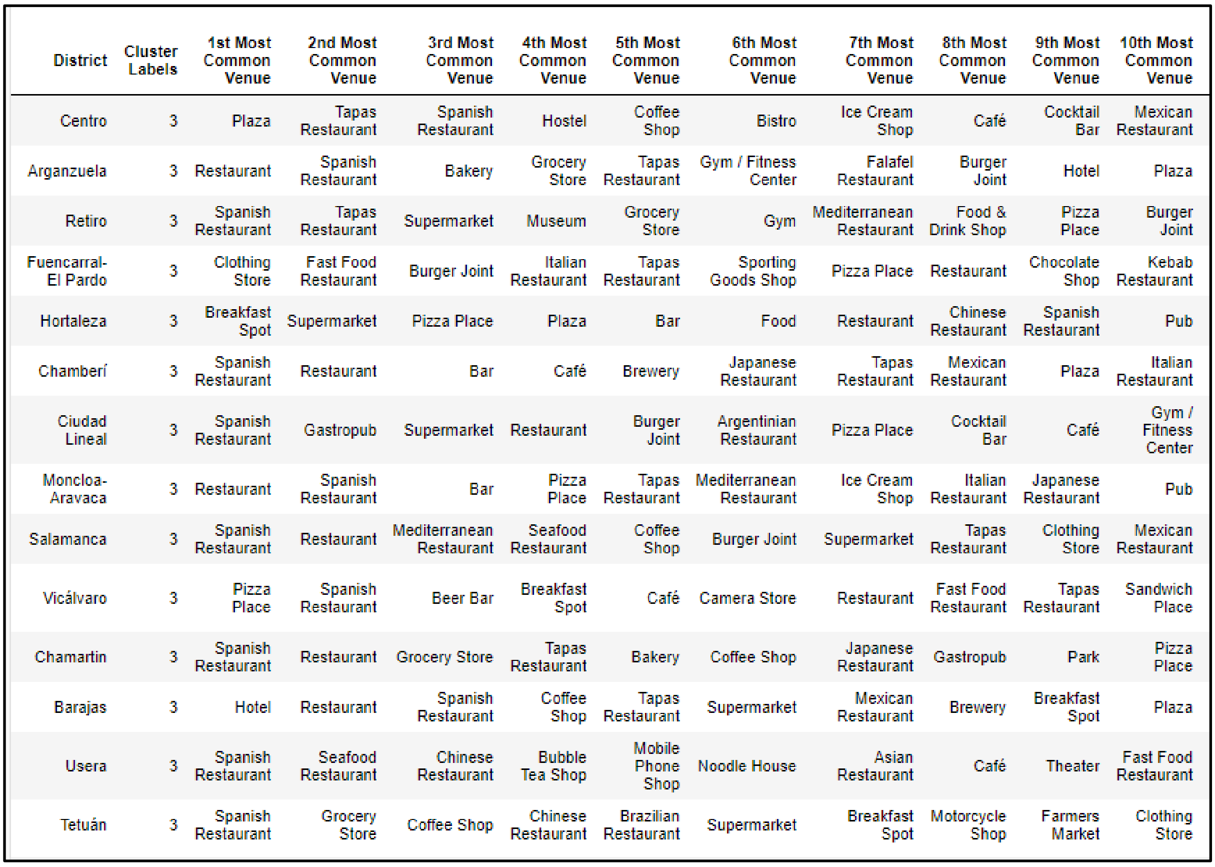
This process is sequential, once a piece of information is obtained, the coding allows to get the next one and so on. Once the whole information is processed completely, the segmentation can be made and the clusters created. First, it is necessary to determine somehow, what the appropriate number of clusters is. To get this number, the Elbow method which consists of in plotting a hypothetical and usually large number of clusters in our data, and draw a curve representing the squared distances between each cluster was used. At some point, the distances will descend to a point where there is no need to keep increasing them. This means that creating more divisions in the data (clusters) is pointless as the difference between groups starts being highly difficult to appreciate



The curve shown above where the distances start reducing importantly from cluster 5 on. So, it was determined that the optimal number of clusters for this case was 5 and it is possible to build the clusters now and have a look at them as shown below



These are the 5 clusters on the Madrid map, it can be seen how many districts belong to each cluster, which is also important information to know as it allows to examine the data of each cluster.



District	Cluster Labels	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9th Most Common Venue	10th Most Common Venue
Centro	3	Plaza	Tapas Restaurant	Spanish Restaurant	Hostel	Coffee Shop	Bistro	Ice Cream Shop	Café	Cocktail Bar	Mexican Restaurant
Arganzuela	3	Restaurant	Spanish Restaurant	Bakery	Grocery Store	Tapas Restaurant	Gym / Fitness Center	Falafel Restaurant	Burger Joint	Hotel	Plaza
Retiro	3	Spanish Restaurant	Tapas Restaurant	Supermarket	Museum	Grocery Store	Gym	Mediterranean Restaurant	Food & Drink Shop	Pizza Place	Burger Joint
Fuencarral-El Pardo	3	Clothing Store	Fast Food Restaurant	Burger Joint	Italian Restaurant	Tapas Restaurant	Sporting Goods Shop	Pizza Place	Restaurant	Chocolate Shop	Kebab Restaurant
Hortaleza	3	Breakfast Spot	Supermarket	Pizza Place	Plaza	Bar	Food	Restaurant	Chinese Restaurant	Spanish Restaurant	Pub
Chamberí	3	Spanish Restaurant	Restaurant	Bar	Café	Brewery	Japanese Restaurant	Tapas Restaurant	Mexican Restaurant	Plaza	Italian Restaurant
Ciudad Lineal	3	Spanish Restaurant	Gastropub	Supermarket	Restaurant	Burger Joint	Argentinian Restaurant	Pizza Place	Cocktail Bar	Café	Gym / Fitness Center
Moncloa-Aravaca	3	Restaurant	Spanish Restaurant	Bar	Pizza Place	Tapas Restaurant	Mediterranean Restaurant	Ice Cream Shop	Italian Restaurant	Japanese Restaurant	Pub
Salamanca	3	Spanish Restaurant	Restaurant	Mediterranean Restaurant	Seafood Restaurant	Coffee Shop	Burger Joint	Supermarket	Tapas Restaurant	Clothing Store	Mexican Restaurant
Vicálvaro	3	Pizza Place	Spanish Restaurant	Beer Bar	Breakfast Spot	Café	Camera Store	Restaurant	Fast Food Restaurant	Tapas Restaurant	Sandwich Place
Chamartin	3	Spanish Restaurant	Restaurant	Grocery Store	Tapas Restaurant	Bakery	Coffee Shop	Japanese Restaurant	Gastropub	Park	Pizza Place
Barajas	3	Hotel	Restaurant	Spanish Restaurant	Coffee Shop	Tapas Restaurant	Supermarket	Mexican Restaurant	Brewery	Breakfast Spot	Plaza
Usera	3	Spanish Restaurant	Seafood Restaurant	Chinese Restaurant	Bubble Tea Shop	Mobile Phone Shop	Noodle House	Asian Restaurant	Café	Theater	Fast Food Restaurant
Tetuán	3	Spanish Restaurant	Grocery Store	Coffee Shop	Chinese Restaurant	Brazilian Restaurant	Supermarket	Breakfast Spot	Motorcycle Shop	Farmers Market	Clothing Store

The information shown above, for example corresponds to the cluster No. 3 which it is the largest one. This type of data allows to analysis information of an entire city by analysing its venues and population.

For this case, the results obtained were five clusters of a diverse population and venues distribution. Below, it is outlined a description of the main features of each cluster.

Cluster One

Mostly inhabited by Ecuadorian, Bolivian and Filipino citizens. The most common venues are Fast-Food restaurants, South American restaurants, cafés, soccer fields, among many others.

Cluster Two

Only comprised of English citizens. The most common venues are Soccer Field, Grocery Store, Spanish Restaurants, etc.

Cluster Three

This cluster is only composed by Bangladeshi people. The most common places are Falafel restaurants, Fish markets, Pizza Place and Diner.

Cluster Four

This is a very diverse cluster; some of the main countries here are Rumania, France, Honduras, Philippines, Paraguay and Morocco among others. The most common venues do also vary, for example, Spanish Restaurants, Mexican restaurants, Chinese restaurants, Breweries, Seafood Restaurants, Coffee Shops, Mediterranean restaurants, etc.…

Cluster Five

This cluster is made up only by 2 nationalities, Ukrainian and Dominican Republic citizens. The most common venues are Metro Stations, Fast Food Restaurants, Pizza Place, Asian restaurants, Grocery stores and bakeries among others.

Discussions

It is interesting how the venues and people from different countries varies from one cluster to another. The main differentiation is located on these two variables, each cluster has its own features, but also common spots with the others clusters. If the results are examined with more detail, some conclusions can be made.

As a recommendation, to make good predictions about where to open a certain type of businesses or shops, more data is needed. For example, socio-demographic data about the population, like their income level, information about quantity of family members, the education level, what kind of job they've got, among others. Also, one of the most important data to examine carefully are those ones related to the people preferences about how they prefer to spend their leisure time, what kinds of food they like, or what their hobbies are. Once, all these data are gathered, a more in-depth analysis could be carried out and the segmentations would be more accurate.

Conclusions

According to data obtained from The Madrid City Hall's Web Portal, there is a considerable Colombian population registered in town which ranks five with 22.618 K as per survey. Furthermore, it can be seen the Colombian citizens are mostly located in the districts Caranbachel (3395 inhabitants), Ciudad Lineal (1792 inhabitants), Latina (1786 inhabitants), Usera (1752 inhabitants) and Puente de Vallecas (1733 inhabitants).

Regarding the districs outlined above, Carabanchel which belong to Cluster 1 is the one with the highest Colombian population and is also considered one of the most diverse neighborhoods in the country, with a large population of immigrants. Furthermore, if we carry out a deeper look of it about its most popular venues, it can be observed there isn't any Colombian and/or Latino restaurants; there are just Tapas and Fast-Food Restaurant. As a consequence, it could be a good opportunity to open a Colombian restaurant and, in the mid-term, some other Latino restaurants owing to the huge potential of the district.

Regarding the other districts with a considerable Colombian citizens population like Ciudad Lineal, Latina, Usera and Puente de Vallecas located in the clusters 1, 4, 5 its population are mostly Latinos, mixed with some other Europeans and Asians. Besides, having analyzed their most popular venues, there are several Fast Food, Argentinian, and South American restaurants. As a consequence, in these clusters, it can be seen the existing restaurants matches the population nationalities and food preferences.

If someone might be interested to open a new Colombian restaurant in the city or any kind of Latino restaurant, it would only be necessary to find a place where there are similar restaurants like the one considered to be opened, make a market research, and find similar clusters of population in the city that don't have them yet or have very few venues like the one to be created.

In conclusion, taking into consideration the explanations given above as well as the data, it is highly possible that the clusters 1, 4 and 5 could be a right place to open a Colombian restaurant. As explained above, the same logic could apply to open other type of restaurants or businesses in any other area of the city. It is just necessary to examine the existing businesses in the target area and study the population, then compare these 2 factors with the same ones in areas where there are existing businesses like the one to be opened, and finally verify if the matching is correct.

Finally, there is always room for improvement and hence the solution given above can be also improved for better results depending upon the data available.