**Madrid Data Analysis**

**Average income vs Venues per Neighborhood**

**Mauro Pérez Manfredini**

**INDEX**

**1. Introduction**

**1.1. Description & Discussion of the Background**

**1.2. Data Description**

**2. Methodology**

**2.1 Cleanning Neighborhood Coordinates data**

**2.2 Cleanning Average income data.**

**2.3 Data visualization.**

**2.4 Venues data from Forsquare API.**

**3. Results**

**4. Discussion**

**5. Conclusion**

**6. References**

**1. Introduction**

**1.1. Description & Discussion of the Background**

This project is about Madrid city. I am currently living in it, so I know it very well and I can see the results with other perspective. Madrid is the capital of Spain and the largest city in the country. It has a population over 6.5 million people in all his region, an about 3.5 million of people in the city, according INE 2019, with a high density of population 5321,05 people/km². Regarding the update data, it is the 5th most populated in Europe [1].

Located in the centre of the Iberian Peninsula, over 646m above sea level, Madrid is divided into 21 boroughs with 126 neighbourhoods around 604,3 km². It preserved one of the most important historical centres among the great European cities, which is harmoniously merged with the most modern infrastructures, a complete offer of accommodation and services and the most advanced technology in audiovisual and communication media. These conditions, together with the drive of a dynamic and open society, but also joyful and welcoming, have made this metropolis one of the great capitals of the western world.

Madrid citizen are mainly young-adult age profile, 44.4% are between 16 and 44 years old (INE 2006). The city has 230.018 M€ of GDP, and a nominal GDP per capita of 34.916€, being the 1st economical metropolis of Spain and the 10th of Europe, behind London, Paris, Rin-Ruhr, Amsterdam, Milan, Brussels, Moscu, Francfort del Meno and Munich.

Due to the economic potential city of Europe, when we think of it by the investor, we expect from them to prefer the districts where there is a high average income cost and the type of business they want to install is less intense. If we think of the city residents, they may want to know which type of social places they have around their neighbourhood to spend money on. However, it is difficult to obtain information that will guide investors in this direction, nowadays.

When we consider all these problems, we can create a map and information chart where the average income per neighbourhood is placed on Madrid and each district is clustered according to the venue density.

**1.2. Data Description**

To consider the problem we can list the data as below:

* I found the Average Income per capita by postal code and neighbourhood of Madrid city [2] The .json file has the name of the neighbourhood and the postal code with the average income I used it to create choropleth map of Average Income per Capita.
* I used Forsquare API to get the most common venues of given Neighbourhood of Madrid [3].
* There are many public data related to demographic and social parameters for the city of Madrid. In this case, I collected the geographical area point for each Neighbourhoods [4][5].

**2. Methodology**

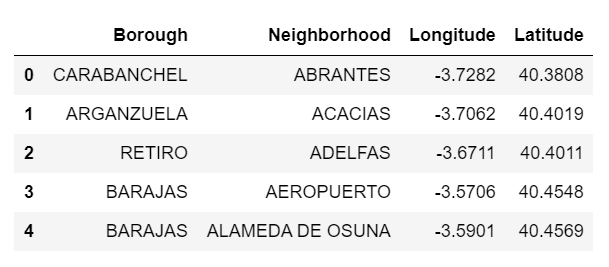
As recommended by the course, I used Github repository in my study to save all my data. I start from the data of the coordinates of each neighbourhood of Madrid. I get it form the statistical website of Madrid, where I get all the limitation point of the area of each neighbourhood.

**2.1 Cleanning Neighborhood Coordinates data**

I checked all the data involve about the neighbourhood to validate and put it in a correct shape to be manipulated. Below are the steps and the manipulations done:

* I found some neighbourhood names with special characters where I changed form a common and generic one.
* The coordinates for all the points where located in the same cells, so I split the information into unique cells.
* I took the average for all the longitude and latitude points of each area to get the center of the neighbourhood.
* Finally, I adequate the units of coordinates from minutes to numerical.
* I transform all the info from the .csv file into a dataframe.

After cleaning all the data, I have information about 21 Borough that contains a total of 126 neighborhood.

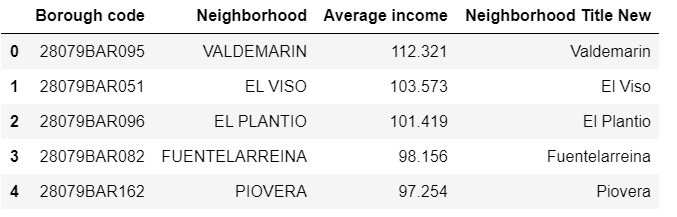


**2.2 Cleanning Average income data.**

I get the data of the average income by neighborhood. from the statistical website of Madrid city and in a .csv file that was already order and clean, so put it into a dataframe.

I have information about 21 Borough code that contains a total of 129 Neighborhood. But I don’t have information for 4 of the Neighborhood (*El Cañaveral, Valderriba, Valdebernardo and Ensanche de Vallecas)*, so I took the decision to delete them form the dataframe.

I create a new column to adequate the name of the Neighborhood to the geojson.



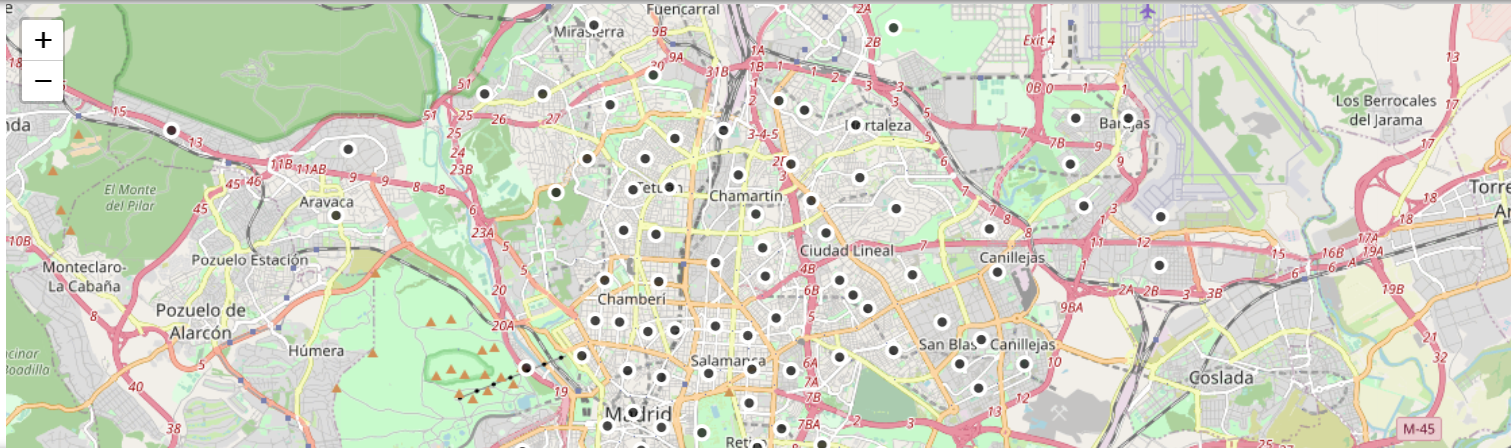
Now I have all the information for all the neighborhood that I am going to study.

**2.3 Data visualization.**

To continue the analyse I get the coordinates of Madrid city using the import function geocode to use as center of my study.

***The geographical coordinates of Madrid are 40.4167047, -3.7035825***

I used python folium library to visualize geographic details of Madrid and its boroughs and I created a map of Madrid with the neighborhoods superimposed on top. I used latitude and longitude values to get the visual as below:



**2.4 Venues data from Forsquare API.**

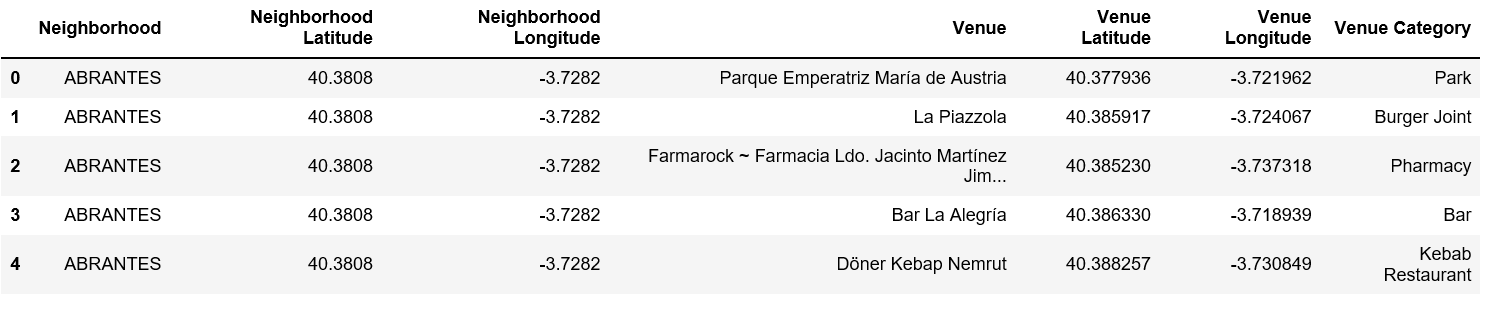
I used the Foursquare API to explore the neighborhoods and segment them. To select the data conditions for the API I use the map, doing a visual estimation with my knowledge about the city, I decided the limit as 100 venue and the radius 750 meter for each neighborhood from their given latitude and longitude information.

To assign each venue information I use the function that was explained in the course. Here is a head of the list Venues name, category, latitude and longitude information from Forsquare API

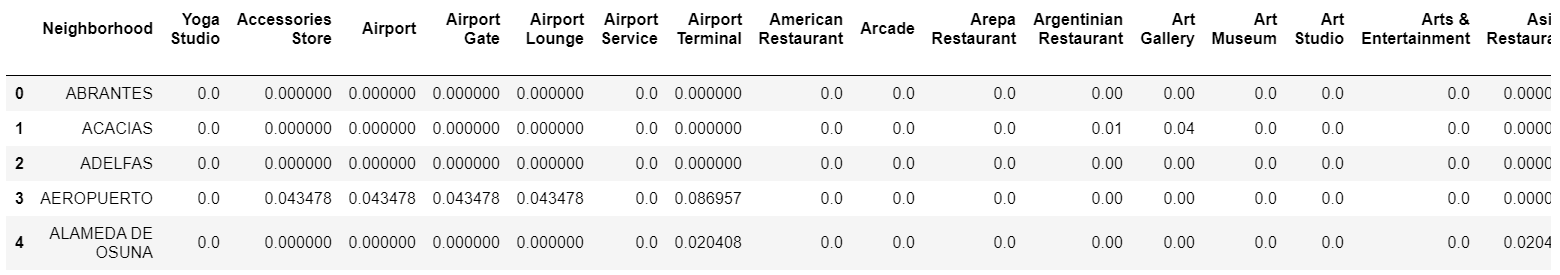


Analysing the venues get, I count the venue by neighbourhood and about 20% of it get the limit of venue, in the contrary I see that around 30% has less than 20 Venue. The result doesn’t mean that inquiry run all the possible results in boroughs. It depends on given Latitude and Longitude information and here is we just run single Latitude and Longitude pair for each borough. We can increase the possibilities with Neighbourhood information with more Latitude and Longitude information. Also, it must be considered that the representative longitude and latitude of each neighbourhood was calculate without considering the distance between neighbourhood.

In summary 293 unique categories were returned by Foursquare, with an average of 58 venues per neighborhood, with a total of 7501 venues found, that is quite good data to evaluate.



Now, I normalize the data taking the mean of the frequency of occurrence of each category grouped by neighbothoods.

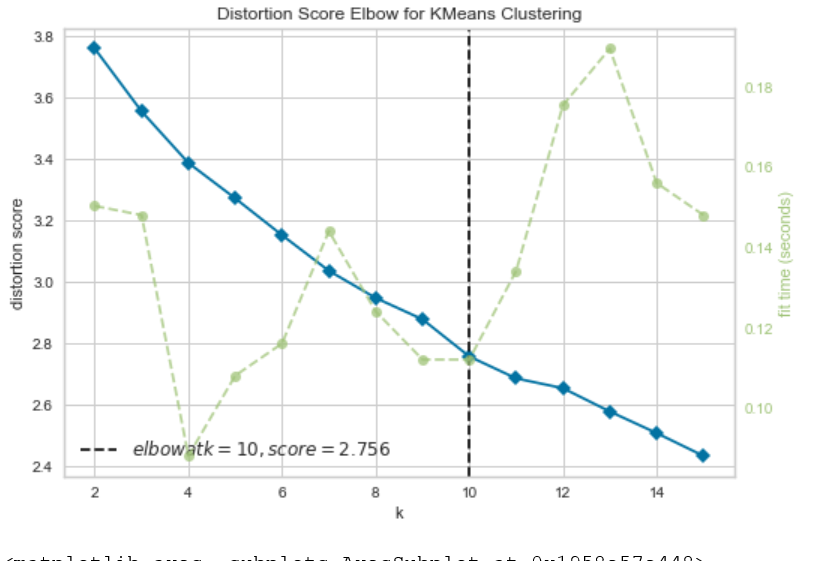


Then I created a table which shows list of top 10 venue category for each borough in below table.



I choose the unsupervised learning **K-means algorithm** to cluster the boroughs. K-Means algorithm is one of the most common cluster methods of unsupervised learning.

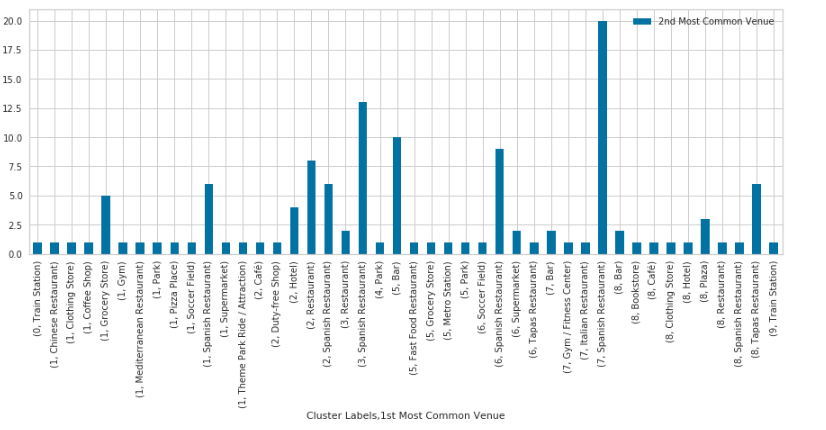
The first thing was to calculate which is the most appropriate number of clusters to my data. I used the KElbowVisualizer function, and the result show that K=10 it is a good number of clusters.



So, I fit my model with the data and the set K=10 clusters. I merged all the information into a dataframe with neighborhood, borough, longitude, latitude, cluster label and the 10st most common venue.



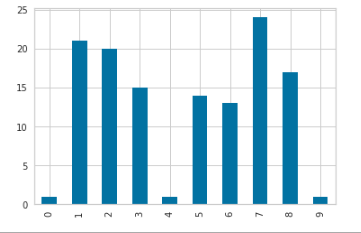
We can also estimate the number of **1st Most Common Venue** in each cluster. Thus, I can analyze them from a pivot table where I sum up in the next table. So, the most appropriate label for each cluster is determine.



From the chart above I rename the cluster to the next labels considering the main venues of each cluster and trying to get the best description.

|  |  |  |
| --- | --- | --- |
| Cluster | Main Venue | New Label |
| 0 | Train Station | Transport area |
| 1 | Spanish Restaurant | Commercial Area |
| 2 | Restaurant | Accommodation & Food Area |
| 3 | Spanish Restaurant | Spanish Restaurant |
| 4 | Park | Green Area |
| 5 | Bar | Bar Area |
| 6 | Spanish Restaurant | Gastronomic Area |
| 7 | Spanish Restaurant | Spanish Restaurant |
| 8 | Tapas Restaurant | Tourist Area |
| 9 | Train Station | Transport area |

Getting these new labels, I check what is the frequency of the number of neighborhoods by clusters. Showing that main neighborhoods are classify as a Multiple Social Venues, follow by Restaurants areas, Spanish restaurant and Multiple Entretainment Venues.



We can also examine that what is the frequency of average income in different ranges. Thus, histogram can help to visualization:



The shape of the data makes this appropriate classification:

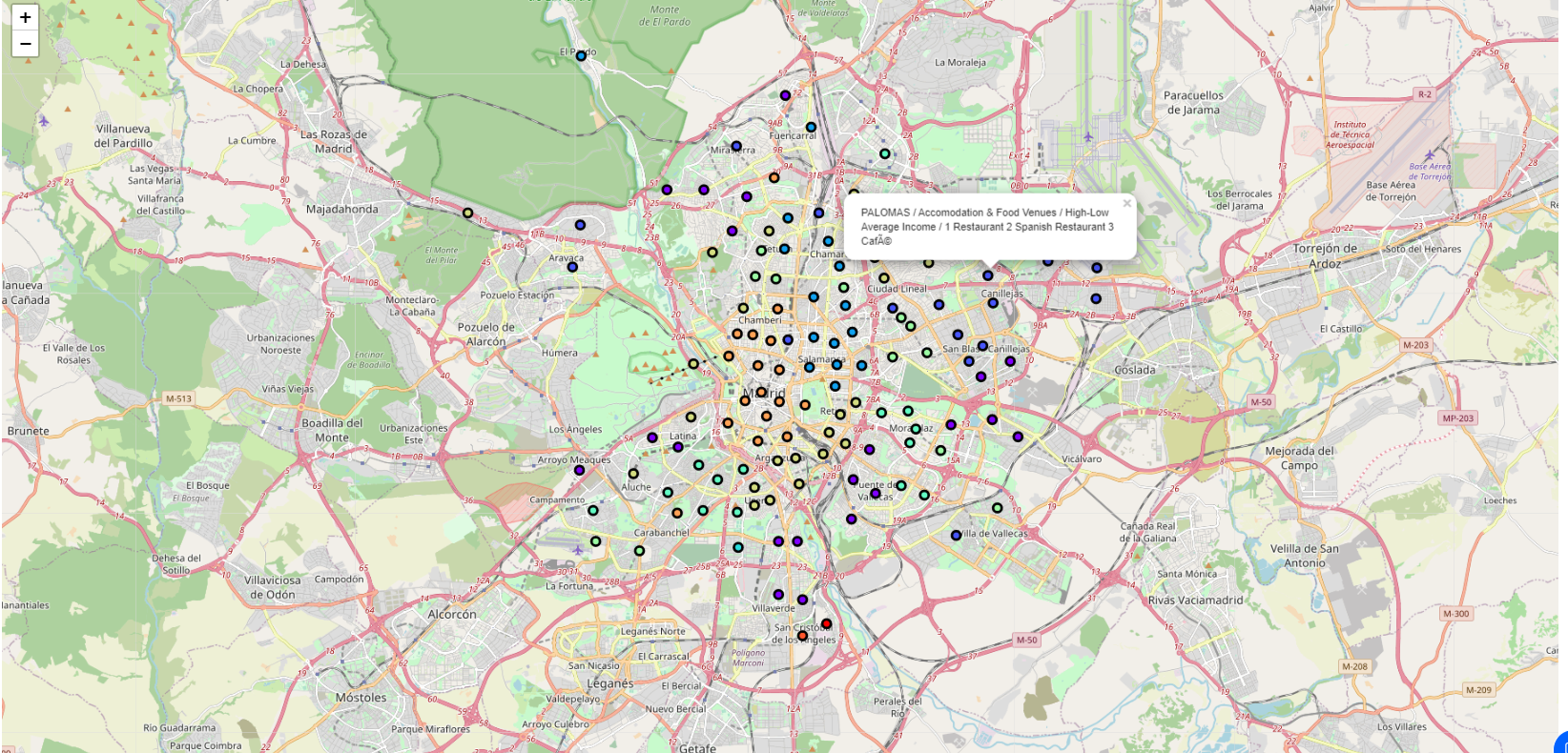
* 19.000 - 30.000€ 🡪 Low average income
* 30.000 - 41.000€ 🡪 Low-High average income
* 41.000 - 52.000€ 🡪 Middle-Low average income
* 52.000 - 63.000€ 🡪 Middle average income
* 63.000 – 74.000€ 🡪 Middle-High average income
* 74.000 – 85.000€ 🡪 High-Low average income
* 85.000 – 96.000€ 🡪 High average income
* 96.000 – 107.000€ 🡪 High-High average income

3. **Results**

All this information was merged in one dataframe to have all the information In one. I also create a join column to get a label of the 1st, 2nd, and 3rd most common venue. I also introduce a column with the label of the correspondent average income group.



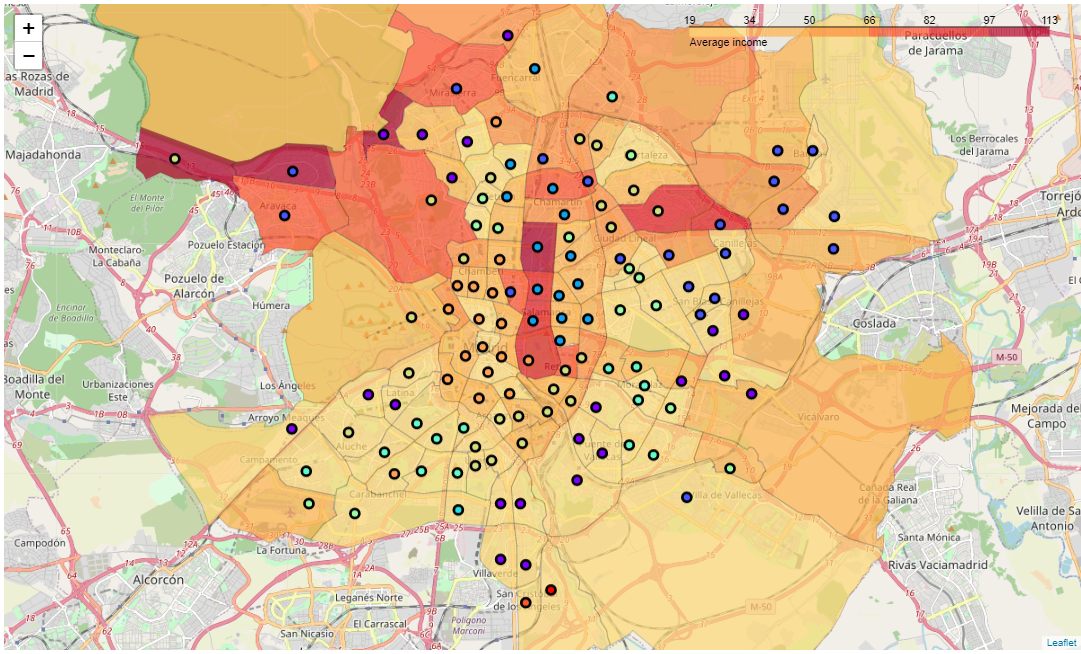
You can now see Join, Labels and Level\_labels columns as the last three ones in above table. Below, you can also see a clustered map borough of Madird.



In summary section, one of my aim was also visualize the Average Income per Neighborhood with choropleth style map. Thus, first I downloaded a geojson file from a Madrid data github repository and created it.

In the las part, I created choropleth map which also has the below information for each borough:

* Neighborhood name,
* Cluster name,
* Average income level,
* Top 3 number of venue



**4. Discusion**

As I mentioned at the introduction, Madrid is a city with lot of neighborhoods, a total of 126. The income distribution show that 9% of the neighbourhood are considered as negihborhood with high and high-low income, the main percentage, around 58% are considered as Low income.

As there is such a complexity, very different approaches can be tried in clustering and classification studies. Moreover, it is obvious that not every classification method can yield the same high-quality results for this metropoly.

I used the Kmeans algorithm as part of this clustering study. When I tested the Elbow method, I set the optimum k value to 10. However, 126 neighborhood coordinates were used. For more detailed and accurate guidance, the data set can be expanded and the details of the neighborhood or street can also be drilled. As I get the center of the neighbourhood by mean of the polygon coordinates, It is possible to adjust the exact center to the geometrical point or to the optimal consideration of the center regarding main streets of more venues coordinates.

I also performed a data analysis through this information by adding the coordinates of districts and average income per neighborhood as static data on GitHub. In future studies, these data can also be accessed dynamically from specific platforms or packages.

I ended the study by visualizing the data and clustering information on the Madrid map. In future studies, web or telephone applications can be carried out to direct investors.

Analyzing the businesses of the neighborhoods, I found a huge diversification, becoming more representative bars and spanish restaurants in most of the neighborhoods. This is logicial, due to Madrid, has a lot of gastronomic culture and social gatherings in bars and take away meals.

**5. Conclusion**

As a result, form the analysis, there is a common trend in the negihborhood, the main businesses are focus on meals, mainly Spanish restaurants, bar or other cuisine restaurants. Rich neighborhood has plenty of market, food and leisure business, that is obvious to spend more money.

Middle average income are more relaxing areas where main venues regard supermarkets, pharmacies, Parks and bars. That’s why people want to settle in this kind of neighborhood where they still have lot of possibilities, but their basic needs are cover in the proximity.

In the future, it can also be done a further analysis merging the rental price per square meter of the neighborhood, age segmentation or other data that can be interesting. It so helpful to have this kind of analysis, wheter you are a investor or a client, through their access to the platforms where such information is provided.

**6. References:**

* [1] [Madrid — Wikipedia](https://en.wikipedia.org/wiki/Madrid)
* [2] [Average Income of Madrid Neighbourhood -- Epdata](https://www.epdata.es/datos/datos-graficos-estadisticas-municipio/52/madrid/4245)
* [3] [Forsquare API](https://developer.foursquare.com/)
* [4] <http://centrodedescargas.cnig.es/CentroDescargas/index.jsp>
* [5][Geographical information of Neighbourhoods of Madrid](https://datos.madrid.es/portal/site/egob/menuitem.c05c1f754a33a9fbe4b2e4b284f1a5a0/?page=0&vgnextoid=b3c41f3cf6a6c410VgnVCM2000000c205a0aRCRD&vgnextchannel=374512b9ace9f310VgnVCM100000171f5a0aRCRD&vgnextfmt=default)