I. Introduction

a. Background

The capital of Spain has 3.27 million inhabitants and in the metropolitan area with approximately 6.5 million of citizens. The muncipality covers 604.3 km2 and it's divided in 21 boroughs with a huge unequality in average income per capita between them.

But that kind of business are inside each borough? Does Madrid meets the cliché that banks are in the areas with higher income per capita and lottery retailers and betting shops are in the poorest areas? Are any other criteria to set up in a specific place any of these kind of businesses in the city?

This analysis trries to classify Madrid neighborhoods based on the concentration of banks and betting shops they have around, taking in account as well the average income per capita in each neighborhood and the average price per square feet for each borough.

In order to reach a conclussion will be used one of the most well known unsupervised machine learning algorithmns: K-Means clustering.

b. Data description

To consider this problem, will be used:

- **Foursquare API**, as this is one of the most accurate free APIs to get information about these particular venues in each borough in Madrid
- **Geojson files** to plot the different neighborhoods in Madrid and show each cluster. Geojson will be used alongside the **Folium** library.
- The latest information about average square feet prices of Madrid, which can be obtained in the official website of the city.
- The average income of Madrid inhabitants, split by neighborhoods, which was obtained from the Instituto Nacional de Estadística (INE), the official statistical entity in Spain.
- And... lot of Python code, where you can see the whole analysis in this Github's repository.

II. Methodology

Following the information obtained from the INE and Madrid.org, the income per capita and the average square feet price in each neighborhood is as follows:

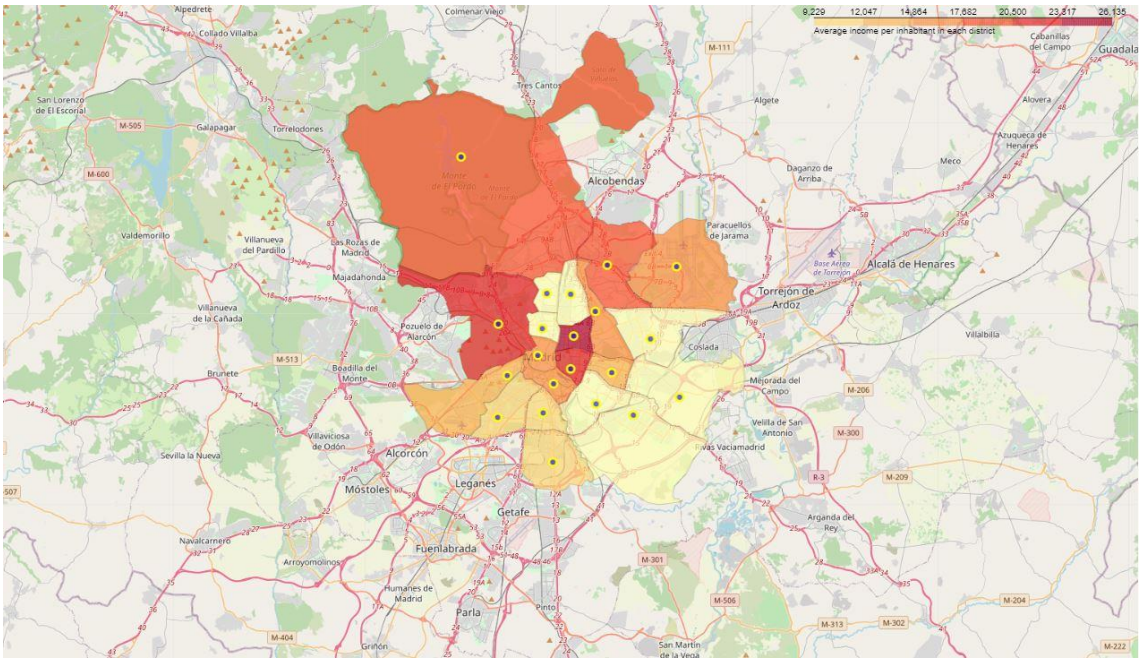
	Borough	Income per capita	Avg Square Feet price	Ratio
0	Centro	16,147	5,037.67	0.31
1	Arganzuela	17,306	4,044.0	0.23
2	Retiro	21,504	4,600.42	0.21
3	Salamanca	24,433	5,846.67	0.24
4	Chamartín	25,969	5,034.08	0.19
5	Tetuán	14,970	3,701.92	0.25
6	Chamberí	22,499	5,301.33	0.24
7	Fuencarral-El Pardo	18,573	3,466.75	0.19
8	Moncloa-Aravaca	22,152	3,932.17	0.18
9	Latina	12,232	2,295.67	0.19
10	Carabanchel	10,872	2,188.75	0.20
11	Usera	9,395	2,042.0	0.22
12	Puente Vallecas	9,545	1,931.5	0.20
13	Moratalaz	13,944	2,538.08	0.18
14	Ciudad Lineal	15,408	3,066.0	0.20
15	Hortaleza	18,277	3,698.58	0.20
16	Villaverde	9,756	1,715.17	0.18
17	Villa Vallecas	11,925	2,414.67	0.20
18	Vicálvaro	11,695	2,274.25	0.19
19	San Blas-Canillejas	13,404	2,521.17	0.19
20	Barajas	17,641	3,176.33	0.18

There's a disbalance in "Centro" borough, where the average square feet is pretty higher in comparison to the income per capita. This has lots of sense, taking in account that is the center of Madrid and the most touristic part, where the space is limited and there are huge investors putting their money there.

As expected, there's a strong linear relationship between the average income per capita and the average square feet price. In concrete, the R-Squared value returned is 0.90, which means that the 90% of the variability of one variable can be explained by the other. Surrounded by a red circle we can see "Centro" borough and how is the farest point from the line.

If we plot all the boroughs based on the average income per capita, there's the evidence that the borough with lower incomes per capita are located in the South. Each small blue and white point represents the geographic coordinates for each district and the opacity of the colors, going from

yellow to maroon, means the average income per capita, being yellow the lowest and maroon the highest.



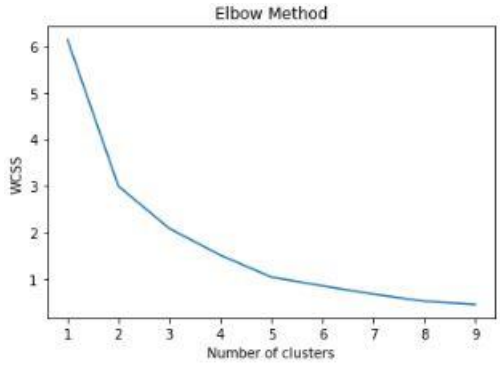
Each of these blue and yellow point will be use as the coordinates to make queries to Foursquare API. Due to the limit on the results that this API returns while be used a loop to make the queries in a radius of 1km, 3km and 5km, which will return 17.350 results. Taking in consideration that Foursquare API doesn't have classified the businesses by borough but by radius from the coordinate, lots of results are duplicated. After classifying them we get 1.000 businesses.

As expected, the ratio between banks and gambling places (considering them lottery retailers, betting shops and casinos) is favorable for the bank side. Nonetheless we find the highest ratio (and also the most number of gambling places) in the centric boroughs: "Centro" and "Chamberí". It seems that the preferable place to set up a gambling business are the most touristic ones. In this order, the boroughs with more lottery retailers, betting shops and casinos are "Centro", "Chamberí", "Barrio de Salamanca", "Tetuán" and "Ciudad Lineal", with a huge different between the two first ones to the other.

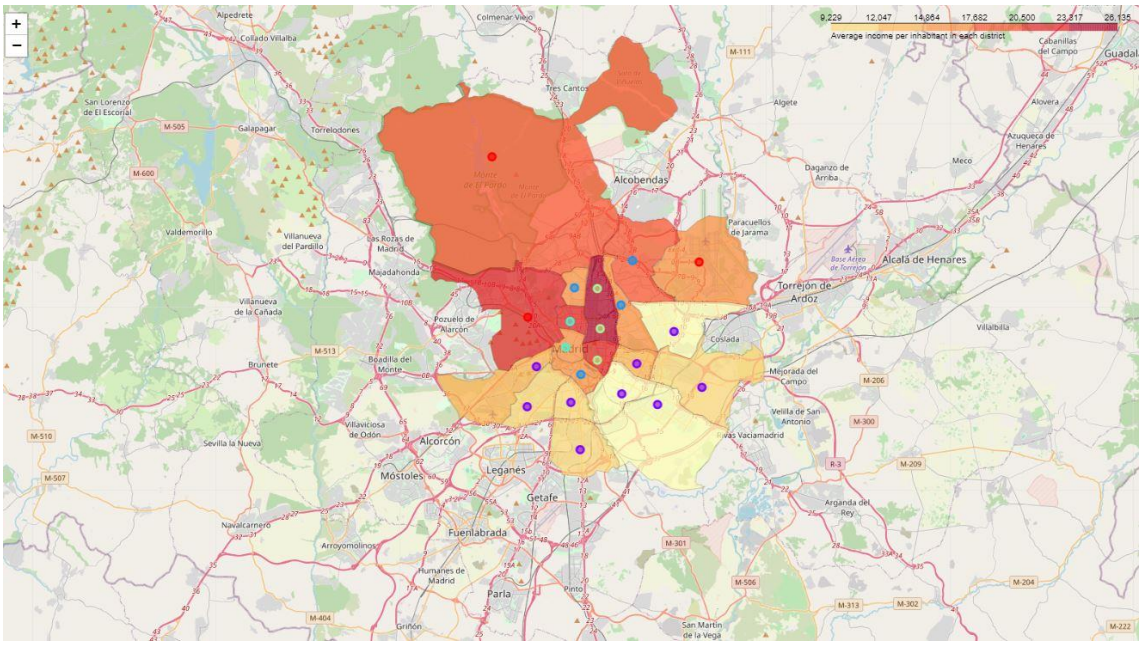
On the contrary, the borough with less gambling businesses are in the suburb of Madrid. The district which has less gambling venues is "Fuencarral-El Pardo", followed by "Villa de Vallecas", "Moratalaz", "Villaverde" and "Barajas".

The results shows a clear pattern that gambling places are mainly located in the center of Madrid and there are not more gambling business based on the average income per capita, which could be thought beforehand.

In that sense, if we ulse the know as "elbow method" to determinate which is the optimal number of clusters in Madrid to clasify the neighborhoods between average income per capita, average square feet price and these two typologies of businesses around the neighborhoods, we get that we can differentiate in Madrid between 2 and 5 groups of boroughs.



In my opinion, after analyzing them, I think that 5 groups or clusters are the optimal split. Each one with their main characteristics:



- **Group Red:** Agglomerates northern districts in the suburbs of Madrid. This group has the lowest number of gambling places and banks and they also have a high average income per citizen.
- **Group Purple**: It is made of the southern districts in the suburbs of Madrid. They have a very few number of gambling places and the lowest average income per citizen.
- **Group Light Blue**: This cluster agglomerates the district of the center of Madrid and has the most of the touristic places. They have more gambling places than any other district and a huge average price per square foot. The average income per capita is not high as much as the districts on the right side of "Paseo de la Castellana" or group green.
- **Group Green**: They are the districts located in the right part of Madrid center. This group has the highest average income per citizen and the second highest average price per square foot, but less gambling places than group light blue. Most of the banks are in this place.
- **Group Blue**: Due to their characteristics they are hybrids that cannot be clasified in other groups because they don't have pretty much gambling places, although they have some and significantly more than cluster red, and the income per capita and price per square feet is on average.

III. Summary

With all this information, we can conclude that gambling places in Madrid are located in the most touristic and center boroughs and the average income is not as much significant as it could be thought beforehand to have this kind of businesses in a specific neighborhood.

On the other side, most of the banks are located in the right side of Madrid center, which is the main financial area, specially the neighborhoods of "Salamanca" and "Chamartin".

As long as we go to the suburbs the number of gambling places decreases and there's a huge different between average income if we are in the North or in the South of the city, differentiating two clear clusters.

If you have read till this point, if you know someone who has in mind to open a business of this typology, this analysis could be useful to advice where to open the business.

https://www.linkedin.com/pulse/battle-neighborhoods-approach-banking-gambling-madrids-javier-moreno