IBM Applied Data Science Capstone

Opening a new office building for an engineering company in Madrid, Spain.

Raúl Cabanas Contreras

June 2020

Introduction

For many companies the decision of where to place a new office building for the company is not an easy process. They have to take into account many factors that are very important such as prize, venues, quick access for clients, employees preferences… and so on. This decisión have to be taken by analyzing data and discusión with all the parts involved.

One of the big factors, is the location according the point of view of the employees. Many studies show that one of the most valuable factors of people to work in a company is its location.

People want to live near the place of work, so they could minimize the transport time. This make the location of the company a key factor to be eligible for the best employees.

Business Problem

In this project we are going to analyze the case of an engineering company, in Spain, that currently have one office building in Madrid, and because the business increase, they have to establish a new location to place more employees.

The current location is in the north of Madrid, in the district of Charmartin as is shown in the above map. Employees surveys have been done about the current location, and the results are very good. Most of the employees rate the current location with a grade 5/5.

For the decision many factors will be reviewed, but this report is focused in determinate which other districts in Madrid are equivalent to Charmartin district in order to maintain the employees satisfaction.

Data

In order to work in the project, we will use the following data:

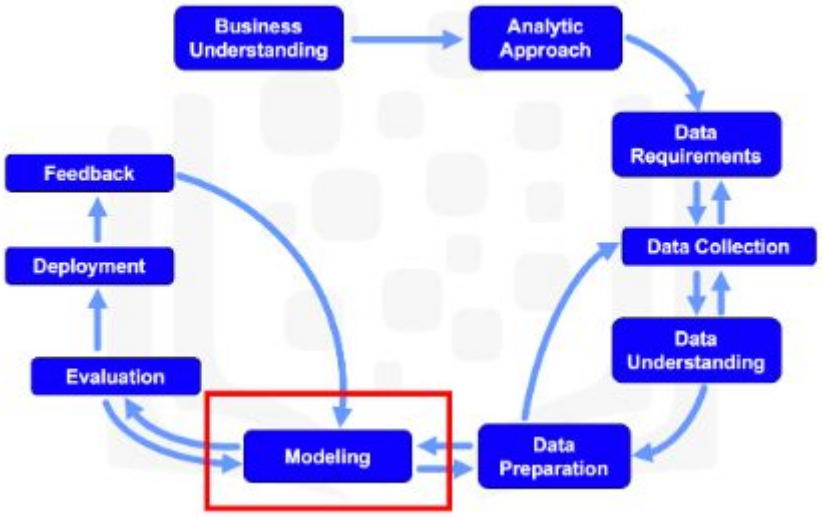
- Districts information of Madrid: The source found is the government official data web page in Spain: https://datos.gob.es/en
  Here we can find many sata set, and it is possible to download the files in json or csv formats.
  In the .csv dataset downloaded, we can find information about the 20 districts in madrid: name, latitude, longitude, population, area and district id.

- Venue data from Foursquare API, which as one of the largest database of places and it is used by many developers.

In this project, we will use many Data Science skills, such as loading csv file, data cleaning, data wrangling, working with API, and machine learning algorithm.

Methodology

For this project we will use the following methodology of data science:



Firstly, we have understand the business problem which is described in the previous sections.

The target of the business problem is the following: Find all districts in Madrid that could be labelled as similar as Chamartin, the one in which currently are placed the company offices.

So the data required must show all characteristics of the districts such as, population, location, most important venues… by feeding the machine learning algorithm with this information we should get the final output required.

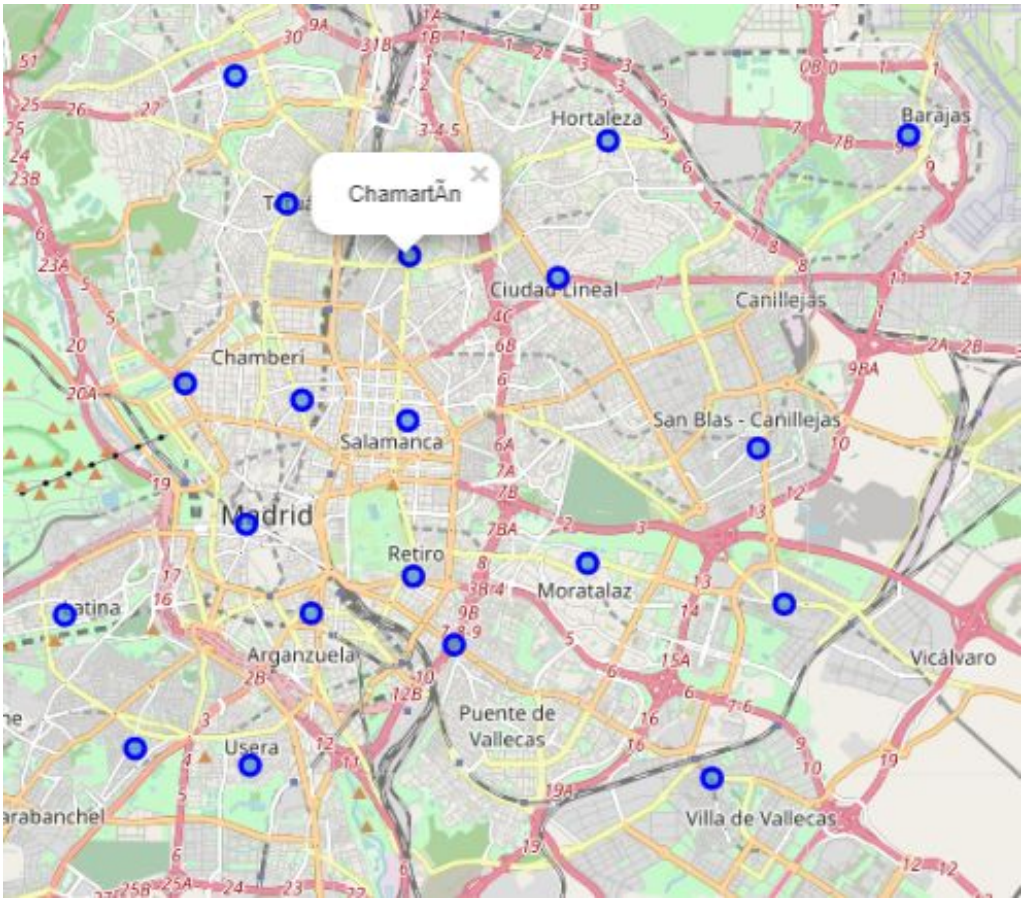
So we will collect the data of two main sources:

- Districts information of Madrid: all information is in a .csv file with the following columns:

	District_id	District	city_id	city	area_km2	population_dens	Latitude	Longitude
0	79601	Centro	796	Madrid	5.21	25340.69	40.415347	-3.707371
1	79602	Arganzuela	796	Madrid	6.52	23306.44	40.402733	-3.695403
2	79603	Retiro	796	Madrid	5.42	21867.53	40.408072	-3.676729
3	79604	Salamanca	796	Madrid	5.36	26830.78	40.430000	-3.677778

The file is downloaded from the original site and loaded in the project through pandas, and a encoding='latin1'.

We position the district in the map to visualize them:



No more data wrangling is needed for this DataFrame.

- Foursquare API: we will use it to get the top 100 venues that are in a radius of 500 meters. For getting the information we had to register in Foursquare Developer Account in order to get an ID and a Password to be able to complete the calls. A for loop is programmed in order to get all information of each district.

  Foursquare output of venue data is in JSON format, so we have to extract the information of name, venue category, latitude and longitude.

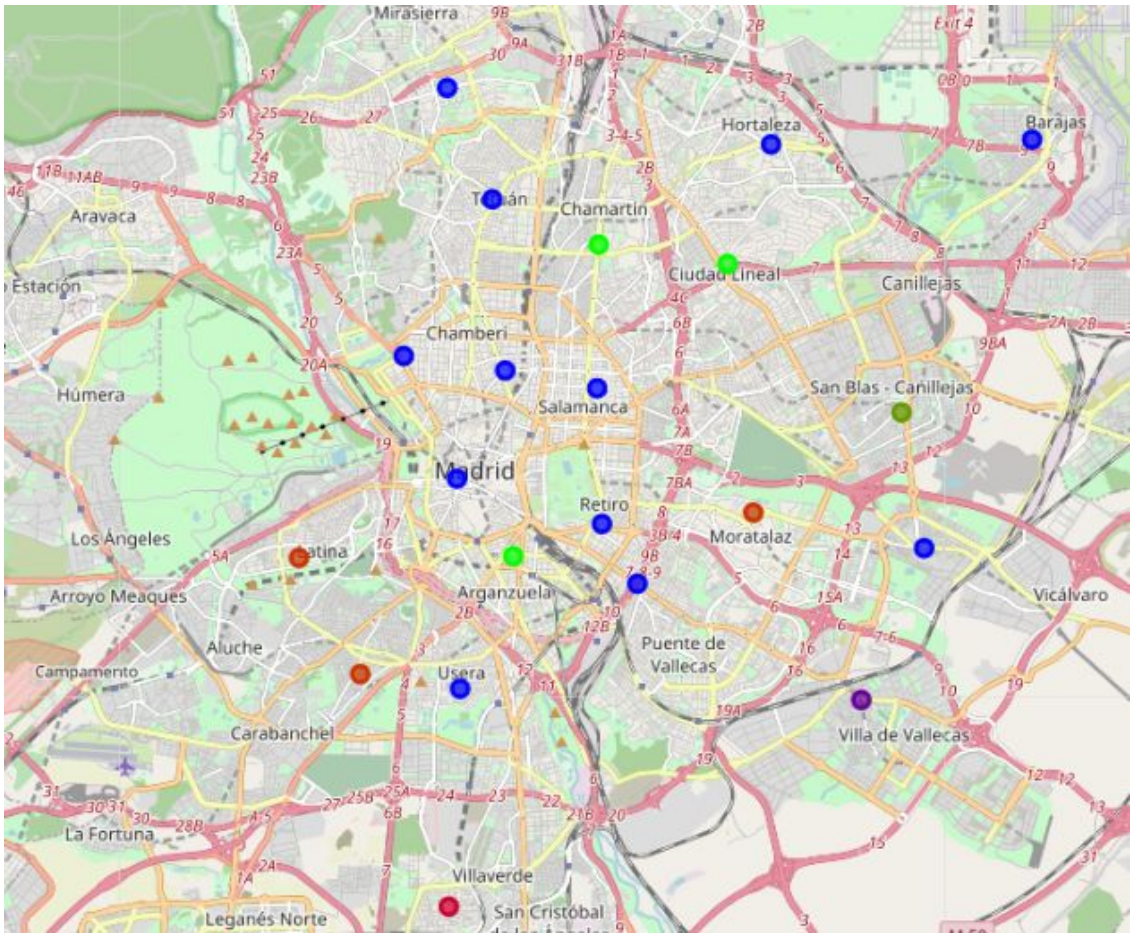
In the data preparation part, we will analyse each district by grouping them and determine the top 10 venues per each. In this operation we have use pandas tools such as get_dummies to change from categorical data to numeric data.

Finally, in modeling phase, we will perform the clustering of the data by using k-means method. This method is an unsupervised method of machine learning and is specially suited to solve problems like this one, in which we have to categorize data.

We have try with different values of k, selecting k=6 as the value that fits better with the current characteristic problem.

Results

The results of the k-means clustering algorithm by finally setting up the value of k=6, shows the following scenario:



In the case the districts that are similar to the district target, Chamartin, are Ciudad Linea, and Arganzuela.

Discussion

The data of the clusters shows that most of the districts can be included in cluster number 1 (blue) and are places more related to residential areas.

The second big group is the red one, which include outer districts like Latina, Carabanchel or Moratalaz with similar characteristics.

Some districts have unique characteristics, such us Villa de Vallecas, San Cristobal and San Blas Canillejas. The three of them are far from the city centre.

And finally, the green cluster, which contains the districts of Chamartin, Ciudad Lineal, and Arganzuela. The three of them are between the city centre and the most outer districts.

In the k-means algorithm the population density is also included and this 3 district are also one of the most clouwded.

	Cluster_labels	population_dens
0	0	19161.750000
1	1	14848.466667
2	2	2026.820000
3	3	7059.130000
4	4	13998.073333
5	5	6934.370000

Conclusion

In this section we will answer the business problem:

Determinate which other districts in Madrid are equivalent to Charmartin district in order to maintain the employees satisfaction.

The two districts that are similar according the output data of the K-means algorithm are: Ciudad Lineal and Arganzuela.

Both of these districts have similar characteristics to Charmartin and placing the new offices in any of these two districts, the satisfaction of employees based on the venues and the population density should be reached.