Coursera Capstone

Opening a new shopping mall in Madrid, Spain

Mikel Hernández Jiménez

April, 2021

1 Introduction

1.1 Background

For many shoppers, visiting shopping malls is a great way to relax and enjoy themselves during weekends and holidays. They can do grocery shopping, dine at restaurants, shop at the various fashion outlets, watch movies and perform many more activities. Shopping malls are like a one-stop destination for all types of shoppers. For retailers, the central location and the large crowd at the shopping malls provides a great distribution channel to market their products and services. Property developers are also taking advantage of this trend to build more shopping malls to cater to the demand. As a result, there are many shopping malls in the city of Madrid and many more are being built. Opening shopping malls allows property developers to earn consistent rental income. Of course, as with any business decision, opening a new shopping mall requires serious consideration and is a lot more complicated than it seems. Particularly, the location of the shopping mall is one of the most important decisions that will determine whether the mall will be a success or a failure.

1.2 Business question

Having this topic in mind, the objective of this capstone project is to analyse and select the best locations in the city of Madrid, Spain to open a new shopping mall. Using data science methodology and machine learning techniques like clustering, this project aims to provide solutions to answer the following business question:

In the city of Madrid, Spain, if a property developer is looking to open a new shopping mall, where would you recommend that they open it?

1.3 Target audience

This project is particularly useful to property developers and investors looking to open or invest in new shopping malls in the capital city of Madrid.

2 Data

2.1 Needed data

To answer the business question the following data will be needed:

- List of neighbourhoods in Madrid, Spain.
- Latitude and longitude coordinates of those neighbourhoods.
- Venue data, particularly data related to shopping malls.

2.2 Data gathering

First, to gather the list of neighbourhoods in Madrid a table has been downloaded from a public government data source, available at: https://datos.gob.es/en/catalogo/l01280796-barrios-municipales-de-madrid. The dataset contains a list of all 131 neighbourhoods of Madrid. Then, to get the geographical coordinates (latitude and longitude) of the neighbourhoods the Python Geocoder package will be used. After that, the Foursquare API is going to be used to get the venue data for those neighbourhoods. This API can provide many categories of venue data, but the Shopping Mall category is the most interest one for this project.

3 Methodology

3.1 Load and clean the dataframe of neighborhoods in Madrid

In this step the dataframe of neighborhoods in Madrid presented in Section 2.2 has been first loaded with pandas. Then, the unnecesary columns have been dropped, to obain only the names of the neighborhoods in Madrid. Next, it has been verified that there are no nan values in the dataframe. Finally, a dataframe of 131 neighborhoods has been obtained.

3.2 Get the geographical coordinates of the neighborhoods

After getting a dataframe of all neighborhoods in Madrid, the geographical coordinates (latitude and longitude) of them has been obtained using the Python Geocoder package. A function has been defined to obtain a the geographical coordinates giving the name of the neighborhood. Then, this function has been used to get a list of geographical coordinates of all the neighborhoods in Madrid. Finally, the coordinates of all neighborhoods hace been merged into the original dataframe, obtaining a dataframe with the neighborhood names and the latitude and longitude of each of them.

The obtained dataframe has been used to create a map of Madrid with neighborhoods superimposed on top. For that, first the geographical coordinates of Madrid have been found with the Python Geocoder package, and them, a map has been created with folium package with circle markers in each neighborhood. The map can be seen in Figure 1.

3.3 Obtain the venue data for the neighborhoods

In this step, the venue data for all neighborhoods in Madrid has been obtained using the Foursquare API. First, it has been nedeed to register a Foursquare Developer Account in order to obtain the Foursquare ID and Foursquare secret key. Then, API calls have been made to Foursquare passing in the geographical coordinates of the neighbourhoods in a Python loop.

The Foursquare API has been used to get the top 100 venues that are within a radius of 2000 meters. The API returns the venue data in JSON format and the venue name, venue category, venue latitude and longitude have been extracted. With this data, how many venues have been returned for each neighbourhood has been analyse and how many unique categories can be curated from all the returned venues has been examined.

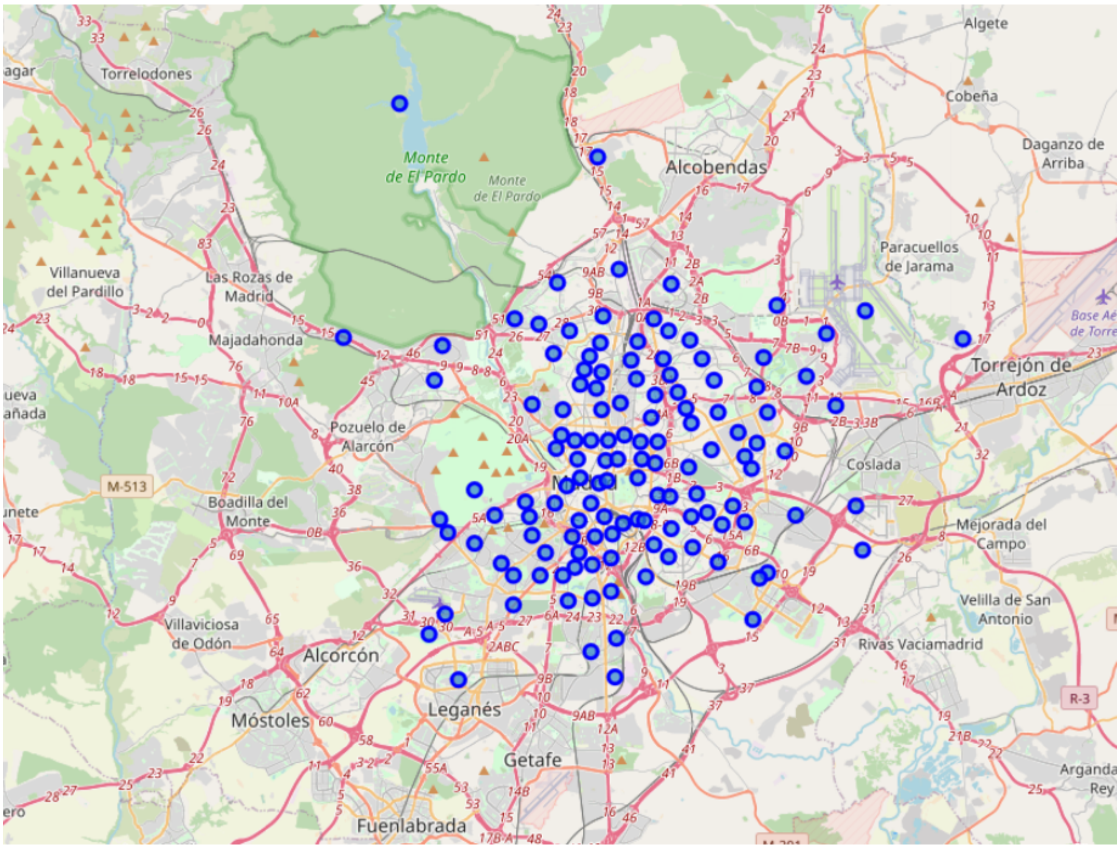

Figure 1: Map of Madrid with neighborhoods superimposed on top.

3.4 Analyse and cluster the neighborhoods

Once the venue data of all neighborhoods has been obtained, the venue categories has been one-hot encoded. Then, each neighbourhood has been analysed grouping the rows by neighbourhood and taking the mean of the frequency of occurrence of each venue category. By doing so, the data has been prepared for use in clustering. Since the target of the project is to analyse the Shopping Mall data, the obtained dataframe has been filtered using "Shopping Mall" as venue category for the neighbourhoods. This way, the obtained dataframe is composed of the frequency of occurrence of Shopping Mall venue category fo each neighborhood. In total, 42 neighborhoods with a frequency of occurrence for a Shopping Mall higher than 0 have been obtained.

Next, a clustering on the data has been performed using k-means clustering. This way, the neighborhoods has been clustered into 3 clusters based on their frequency of occurrence for Shopping Mall venues. Then, a map of Madrid has been created to visualize the clusters using the folium package.

The results of this clustering can be useful to identify which neighborhoods have higher concentration of Shopping Malls while which neighborhoods have fewer number of Shopping Malls. Based on the occurrence of Shopping Malls in different neighbourhoods, it can help to answer the question as to which neighbourhoods are most suitable to open a new Shopping Mall.

3.5 Examine the clusters and select the best neighborhood to open a new Shopping Mall

Finally, the clusters has been analysed to select the best neighborhoods to open a new Shopping Mall in Madrid, Spain. For this, the values of the three clusters has been analysed and the neighborhoods corresponding to the cluster with low amount of Shopping Malls has been considered as potential neighborhoods to open a new Shopping Mall.

4 Results

The results from the k-means clustering show that the neighbourhoods can be categorized into 3 clusters based on the frequency of occurrence for a Shopping Mall:

- **Cluster 0**: Neighbourhoods with moderate number of Shopping Malls.
- **Cluster 1**: Neighbourhoods with no existence of Shopping Malls.
- **Cluster 2**: Neighbourhoods with high concentration of Shopping Malls.

The results of the clustering can be visualized in Figure 2 with Cluster 0 in red colour, Cluster 1 in purple colour, and Cluster 2 in mint green colour.

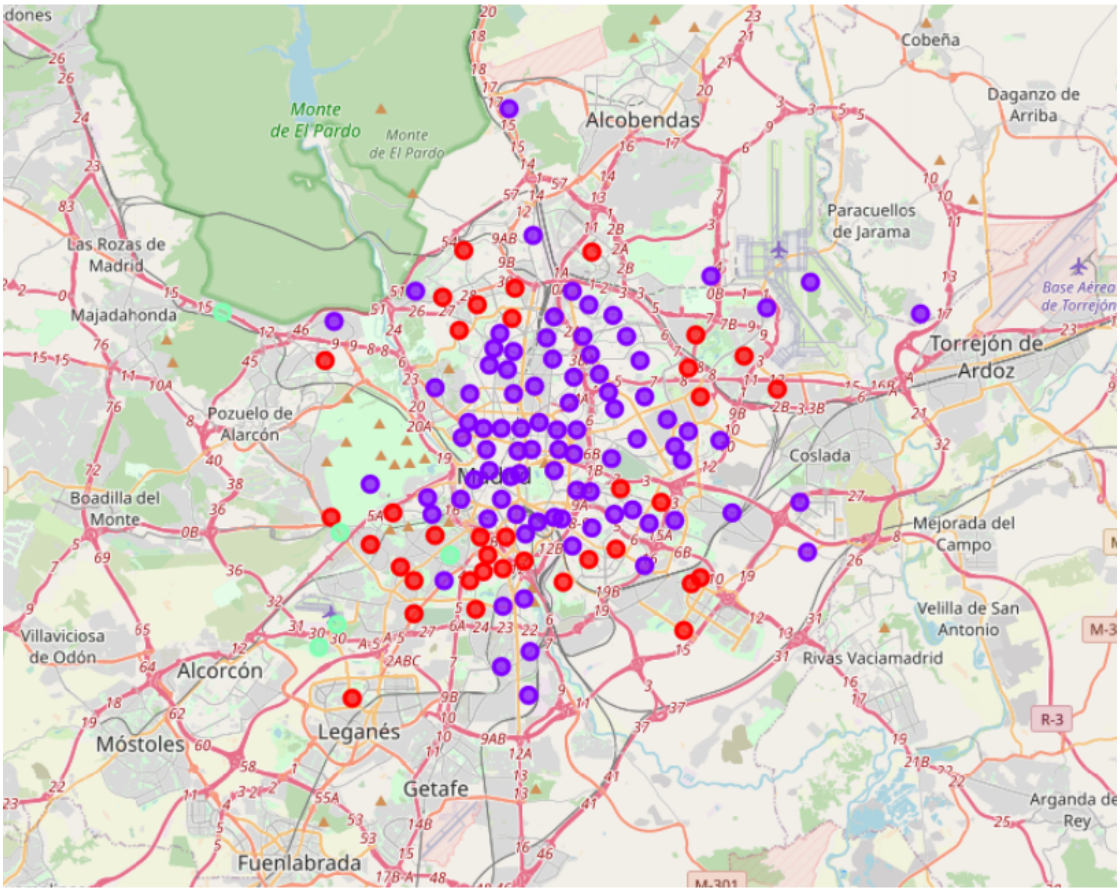

Figure 2: Map of Madrid with the resulting clusters based on the frequency of occurrence for a Shopping Mall center.

The cluster results shows that neighborhoods in Cluster 1 represents a great opportunity and high potential areas to open new shopping malls as there is no competition from existing malls. Meanwhile, Shopping Malls in Cluster 2 are likely suffering from intense competition due to oversupply and high concentration of shopping malls. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighborhoods in Cluster 0 with moderate competition.

In total, 88 potential neighborhoods have been found to open a new Shopping Mall in Madrid, Spain, which are superimposed on top in the map shown in Figure 3. From the map it can be concluded that the best neighborhoods to open a Shopping Mall in Madrid are those located in the city centre, instead of in the suburns.

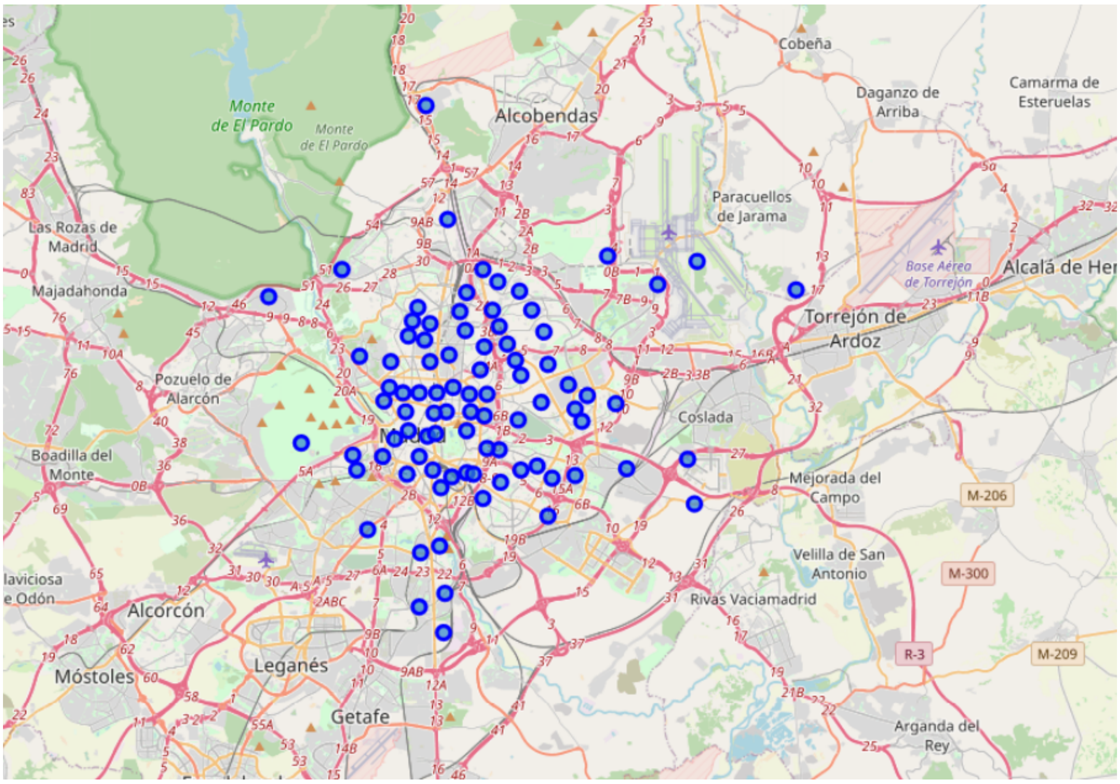

Figure 3: Map of Madrid with the potential neighborhoods to open a Shopping Mall superimposed on top.

5 Discussion

5.1 Findinds

As observations noted from the maps in Section 4, most of the Shopping Malls are concentrated in four peripheral neighborhoods of Madrid city. However, moderate number of Shopping Malls can be found in the surroundings of the central area of Madrid city. On the other hand, the neighborhoods with a great opportunity and high potential to open a new Shopping Mall are located in the city center and some peripheral neighborhoods.

Therefore, this project recommends property developers to capitalize on these findings to open a new Shopping Mall in neighbourhoods in Cluster 1 with no competition. Property developers with unique selling propositions to stand out from the competition can also open new shopping malls in neighbourhoods in Cluster 0 with moderate competition. Lastly, property developers are advised to avoid neighbourhoods in Cluster 2 which already have high concentration of shopping malls and suffering from intense competition.

5.2 Limitations

In this project, only factor has been considered, i.e. frequency of occurrence of Shopping Malls, there are other factors such as population and income of residents that could influence the location decision to open a new Shopping Mall. However, to the best knowledge of this researcher such data has not been available to the neighbourhood level required by this project.

5.3 Future Research

Future research could devise a methodology to estimate such data to be used in the clustering model to determine the preferred locations to open a new Shopping Mall. In addition, this project made use of the free Sandbox Tier Account of Foursquare API that came with limitations with the number of API calls and results returned. Future research could make use of paid account to bypass these limitations and obtain more trustuful results.

6 Conclusion

In this project, the process of identifying a business problem, specifying the data required, extracting and preparing the data, performing machine learning by clustering the data into 3 clusters based on their similarities, and providing recommendations to the relevant stakeholders has been developed.

The target audience of the project may be property developers and investors regarding the best locations to open a new Shopping Mall in Madrid, Spain.

To answer the business question raised in the introduction section, the answer proposed by this project has been the following: The neighbourhoods in cluster 1 are the most preferred locations to open a new shopping mall. The findings of this project will help the relevant stakeholders to capitalize on the opportunities on high potential locations while avoiding overcrowded areas in their decisions to open a new shopping mall.