ANTONIO J. ROLDÁN CASTRO
IBM Professional Certificate on Data Science

Coursera Capstone

Opening a Multiplex or Movie Theater in Madrid, Spain

Introduction

Despite the current trend of appearing new VOD platforms in Spain and worldwide, such as Netflix, Disney +, HBO, etc., the business of entertainment and theaters still generates an average of Box Office of 600 million € per year in Spain with an upward trend since few years ago (without contemplating the current Covid situation).
The city of Madrid gathers itself approx. the 20% of the total Box Office of films in the national territory.

Business Problem

With the mentioned situation, we'll intent to realize an initial approximation to analyze the best (if any) location of a new profitable Multiplex or Movie Theaters in Madrid (Spain) using data science tools and methodologies and machine learning techniques such as clustering data.

Target of the project

This report will be useful as a first study to analyze the different neighborhoods of Madrid in terms of density of theater nearby each one of them, to identify the best initial position to open a new theater.

Data and Documentation Needed

- List of neighborhoods in Madrid, Spain.
- Location of each neighborhood: latitude and longitude.
- Venue data (theaters and Multiplexes) to enumerate the amount of theaters around each node (neighborhood).

Sources of data and software

- Wikipedia pages to extract information of the Neighborhoods of Madrid:
  - https://en.wikipedia.org/wiki/Districts_of_Madrid.
- Python, Jupyter Notebooks and packages related to:
  - Python Geocoder (to extract coordinates)
  - Plotting maps (Folium)
  - Reading of websites (html)
  - K-means and Cluster analysis
- Foursquare API
  - To get the information and location of the theaters

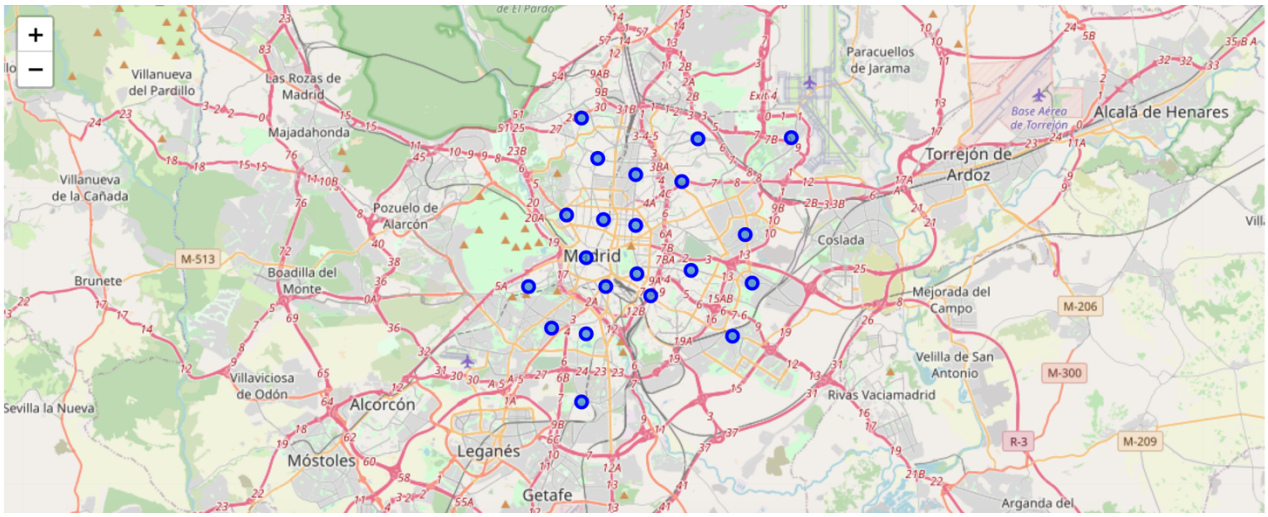
Methodology and Steps

Using IBM Watson and Jupyter Notebooks we have used Python to develop our code which will perform the analysis. We have found the necessary data of the Neighbors in Madrid on Wikipedia page (https://en.wikipedia.org/wiki/Districts_of_Madrid) and we'll scrap the data using Python without using the Beautysoap packages, just simply with the "pd_read_html" code of the Pandas Package since the table of contents is quite simple.

We cleaned the data, renaming columns and adding information separated in different columns to order the data. After that and by using Geopy we tried to get the coordinates of each one of the neighbourhoods. Checking the information manually we find different errors on the coordinates. Neighbors' coordinates received like "Distrito Centro" or "Distrito Retiro" differ from the correct coordinates so we manually create a new csv just with the name and the coordinates of each neighbors that we find on Geohack.com.

We upload this csv to a Google Drive folder and we upload the document onto the Notebook Jupyter. We merge the dataframes to get a final dataframe that we can visualize with the Folium Map package of Python. The result is the following:

[14]:



Once we have a cleaned dataframe with the correct coordinates we can start searching for the theaters around each node by using the Foursquare API making calls to receive the information data (json file) of theaters around 2.000meters from the center of the node. Instead of searching for any kind of Venue we directly want the category of theaters so we receive only that kind of place around each node.

	Neighborhood	Latitude	Longitude	VenueName	VenueLatitude	VenueLongitude	Venue
0	Centro	40.415347	-3.707371	Yelmo Cines Ideal	40.413775	-3.703745	
1	Centro	40.415347	-3.707371	Cines Callao - Callao City Lights	40.420235	-3.706089	Mov
2	Centro	40.415347	-3.707371	Cine Doré	40.411833	-3.699267	Mov
3	Centro	40.415347	-3.707371	Cines Golem	40.424703	-3.713700	Mov
4	Centro	40.415347	-3.707371	Renoir Plaza España	40.424343	-3.713386	Mov
...	...	...	...	...	...	...	
129	Vicálvaro	40.404200	-3.608060	Centro Comercial Manoteras	40.398313	-3.605202	
130	Vicálvaro	40.404200	-3.608060	Cine 3D	40.391545	-3.610527	Mov
131	Vicálvaro	40.404200	-3.608060	Cinesa Las Rosas 3D	40.418567	-3.620932	
132	San Blas-Canillejas	40.426001	-3.612764	Cinesa Las Rosas 3D	40.418567	-3.620932	
133	San Blas-Canillejas	40.426001	-3.612764	Sala de proyecciones Deluxe	40.432081	-3.632247	I

Now we analyze and study the data by using:

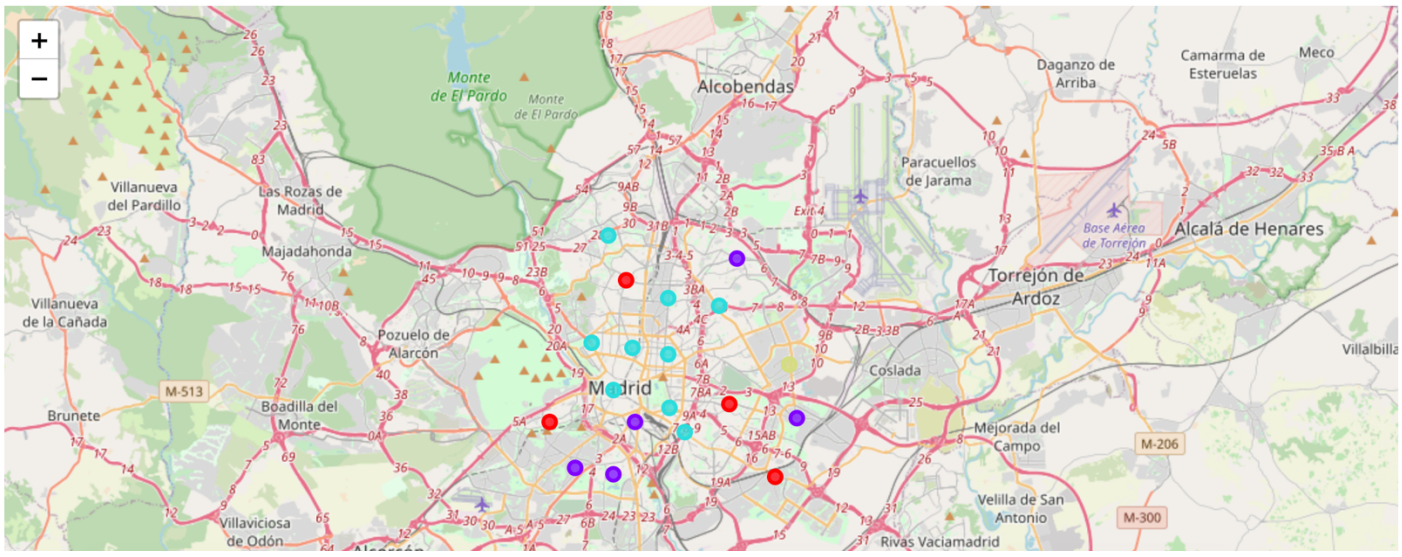
- One hot encoding to stablish the mean of frequency of the Theaters in each neighborhood.
- K-means clustering algorithms: we create 4 clusters of frequency of occurrence of Theaters in each neighborhood.

Results

The results from the K-mean algorithms show the four categories of the neighbors of Madrid in terms of frequency of occurrence for Movie Theaters:

- Cluster 0: District with the highest number of movie theaters. Red color
- Cluster 1: Low to Moderate number of Movie Theaters. Purple.
- Cluster 2: Moderate to high number of Movie Theaters. Green.
- Cluster 3: Low or no number of Movie Theaters occurrence. Yellow

[34]:



Discussions of results

There is a clear concentration of cinemas at the center of the city (cluster 2) despite the Multiplexes around the city and located in nearby populations. According to these results there may be an opportunity of the Cluster 3 (Yellow) on the District of San Blas-Canillejas with a population of over 150K citizens.

Limitations and Suggestions for Next Reports

Since the scope of this report was just the analysis of the neighbors in terms of number of frequency and existence of Movie Theaters nearby we should consider this first approach as an initial guideline to cross more data to interpret the results. We suggest:

- Population
- Average income of the neighbors
- Performance of the theaters of each node (Are they getting profits or losing money?)

Conclusion

There is only one neighbor in Madrid with a market opportunity for a new movie theater considering the lack of those nearby it: District of San Blas. We should take this as a premise to drill down the information of this district and obtain necessary data (we can't currently access to) in order to complete the business report such as average income, social habits, etc.