Coursera Capstone Project: The Battle of Neighborhoods (Week 2)

Predicting the best place to open a shopping mall in Casablanca, Morocco

1. Introduction:

1.1 Background

Nowadays there are many ways to go shopping, however locating one that matches a modern way of living can be complicated. We can go shopping in towns, markets as well as on the net, but the most practical is the shopping mall. The mall culture has become a big business, as they have become multi-story structures that house a large number of shops selling various products and services. Shoppers prefer shopping malls to stand-alone shops for various reasons. They can have their own parking facility, compare between a large varieties of products from different producers before making purchases, enjoy meals at food courts with a wide variety of cuisine, or see a movie at cinemas in shopping centers. All these features make shopping a fun-filled and satisfying experience.

Since shopping centers are the most sought-after shopping destinations, it is beneficial for a retail store owners to rent shop space in a mall because they are usually located in prime locations which are easily accessible, It enables him also to build a clientele by attracting clients of competitors who have shops in the mall, and to focus on his business without having to direct time and efforts towards the maintenance of the shop. Property developers and investors are also taking advantage of it by investing and building more shopping centers to cater to the demand. With so many benefits of shopping malls to shoppers as well as businessmen we can conclude that shopping malls will only rise in popularity with time and that's why there are more and more shopping malls in the big city of Casablanca.

There are many points to consider before opening a shopping center as with any business decision, and one of the most important decisions is the location of the shopping center which will determine its success or failure

1.2 Business Problem

The objective of this capstone project is to analyze and determine the best place to open a new shopping mall in the city of Casablanca, Morocco. Using Data Science methodology and tools, It will give an answer to the business question: In the city of Casablanca, Morocco, if a property developer is looking to open a shopping mall, where would I recommend that they open it?

1.3 Interest

Obviously, property developers and investors would be very interested by this project for competitive advantage and business values if they are looking to open or invest in new shopping malls in the biggest city of Morocco i.e. Casablanca.

2. Data acquisition and cleaning:

2.1 Data requirements:

- List of neighborhoods in Casablanca. This defines the scope of this project which is limited to the city of Casablanca.
- Latitude and longitude of those neighborhoods. This is useful to get the venue data and build map.
- Venue data. This is useful to perform clustering on the neighborhoods, especially data related to shopping malls.

2.2 Data sources and extraction methods:

This Wikipedia page https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Quartier_de_Casablanca contains a list of neighborhood in Casablanca. I'll use web scraping techniques to extract data from it using Beautifulsoup and Requests packages of Python. Then, I'll use Geocoder Python package to get the geographical coordinates of the neighborhoods (latitude and longitude).

After that, I'll use the Foursquare API to get the venue data of those neighborhoods. Foursquare API will provide many categories of the venue data, I'm particularly interested in the Shopping mall category to solve the business problem put forward.

This project will allow me to use many data science skills, from web scraping, working with API, data cleaning, data wrangling, to machine learning and map visualization.

In the next section, I'll present the methodology section where I'll discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

3. Methodology:

First of all, I needed to get the neighborhoods in the city of Casablanca. I found a list of the neighborhoods available on Wikipedia (https://fr.wikipedia.org/wiki/Cat%C3%A9gorie:Quartier_de_Casablanca).

I did web scraping using Python Requests and Beautilfusoup packages to extract the list of neighborhoods data. After that, I needed to get the geographical coordinates (longitude and latitude) of the extracted neighborhoods to be able to use Foursquare API.

For that, I used the Geocoder package which allow to convert address into geographical coordinates in the form of latitude and longitude. After gathering data, I stored the data into a pandas Dataframe and visualized the neighborhoods on a map using Folium package in order to perform a sanity check and make sure that the geographical coordinate's data returned by Geocoder are correctly plotted in the city of Casablanca.

After that, I used Foursquare Api to get the 100 venues that are within a radius of 2000 meters using my Foursquare Developer Account's credentials.

I made API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a loop. Foursquare API returned the venue data in a JSON format from where I extracted the venue names, venue categories, venue latitude and longitude. With those data, I was able to check how many venues were returned for each neighborhood and examine how many unique categories can arranged from all the returned venues.

Then, I analyzed each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, I was also preparing the data to use it in clustering. Since I was analyzing the “Shopping Mall” data, I filtered the “Shopping Mall” as venue category for the neighborhoods.

Finally, I performed clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. I clustered the neighborhoods into 3 clusters based on their frequency of occurrence for “Shopping Mall”. The results allowed me to identify which neighborhoods have higher concentration of shopping malls and which one have fewer number of shopping malls. Based on the occurrence of shopping malls in different neighborhoods, it helped to answer the question as to which neighborhoods are most suitable to open new shopping malls.

4. Results:

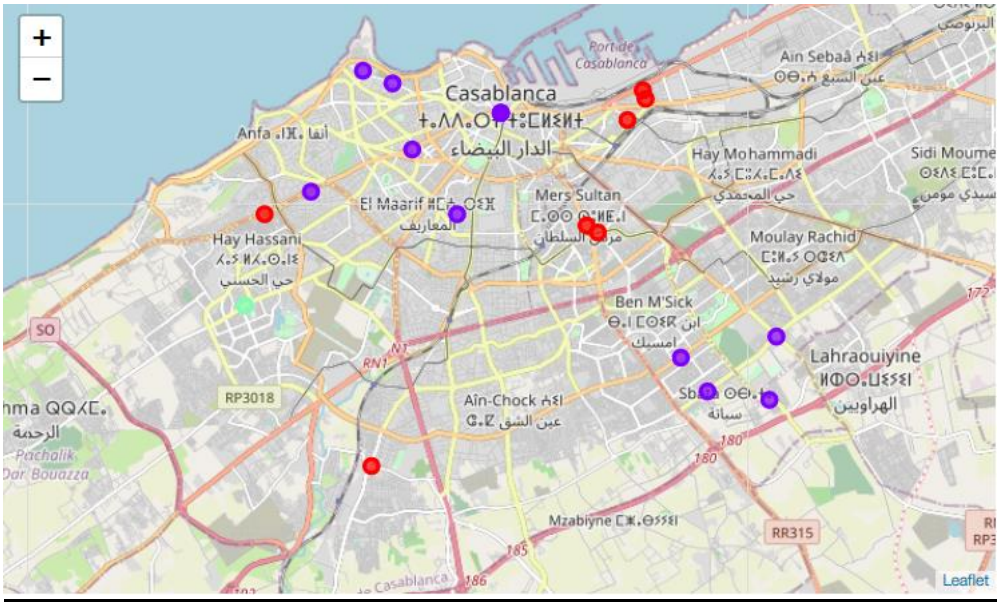
The results from the k-means clustering show that we can categorize the neighborhoods into 3 clusters based on the frequency of occurrence for “Shopping Mall”:

Cluster 0: Neighborhoods with moderate number of shopping malls

Cluster 1: Neighborhoods with high concentration of shopping malls

Cluster 2: Neighborhoods with low number to no existence of shopping malls

We can visualize the results of the clustering in the map below:



5. Discussion:

As we can see in the map, most shopping malls are concentrated in the central area of the city of Casablanca, with the highest number in cluster 1 and a moderate number in cluster 0. In contrast, cluster 2 has a very low or totally non-existent number in districts. This represents an excellent opportunity and areas with high potential to open new shopping centers, as there is very little or no competition from existing shopping centers. Meanwhile, cluster 1 shopping centers are likely to suffer from intense competition due to oversupply and high concentration of shopping centers. From another point of view, it also shows that the oversupply of malls mainly occurred in the central area of the city, with the suburb still having very few malls. Therefore, this project recommends that real estate developers capitalize on these results to open new shopping centers in cluster 2 neighborhoods with little or no competition. Real estate developers with unique selling propositions to set themselves apart from the competition can also open new malls in cluster 0 neighborhoods with moderate competition. Finally, real estate developers are advised to avoid neighborhoods in cluster 2 which already have a high concentration of shopping centers and which suffer from intense competition.

6. Conclusion:

In this project, I followed the process of identifying the business problem, specifying the required data, extracting and preparing the data, performing the machine learning by grouping the data into 3 clusters based on their similarities, and finally by providing recommendations to relevant stakeholders i.e. real estate developers and investors regarding the best locations to open a new shopping center. To answer the commercial question that was raised in the introductory section, the answer proposed by this project is: The neighborhoods of cluster 2 are the most privileged places to open a new shopping center. The results of this project will help relevant stakeholders capitalize on opportunities at high potential sites while avoiding overcrowded areas in their decisions to open a new shopping center.
And finally, I can add that by knowing Casablanca well, the neighborhoods of cluster 2 know more and more inhabitants and there is not enough shopping malls in this area. In addition, the inhabitants of these neighborhoods are rather consumers and go regularly to the malls in the city center even if it's way too far from them. So being able to have a mall near their home can not only generate gains for owners and investors but also save transport costs as well as travel time to consumers and reduce crowds in the shopping malls of the cluster 1.

7. Limitations and Suggestions for Future Research

In this project, I only considered the frequency of occurrence of shopping malls as a factor, but there are other factors such as population and income of residents that could influence the choice of location for a new shopping mall. However, to the knowledge of this research, this data is not available for the neighborhoods required by this project. Future research could design a methodology to estimate this data for use in the clustering algorithm to determine preferred locations to open a new mall. Additionally, this project used the free Foursquare API sandbox level

account, with limits on the number of API calls and the results returned. Future research might use the paid account to bypass these limitations and get more results.