The battle of the neighbourhoods - Discovering Casablanca

Introduction to the business problem

As the largest city in Morocco, Casablanca is one of the best investment destinations in north Africa. Casablanca is located in the centre of the Casablanca-Settat region who according to the ministry of finance[i] contribute 26.5 per cent to the nation's GDP.

Casablanca's strategic location, the availability of undertrial and logistical infrastructure and its attractive business climate makes the city the main destination for startups and large-scale investment project. However, the publicly available microdata on the locations of venues is extremely scarce. This can make it hard to choose the best location for an investment project.

The objective of this project is to facilitate the choice of businesses locations in Casablanca based on the frequency of nearby venues.

The goals of this project are:

- Identify the geographical position of all the neighbourhoods in the city of Casablanca
- Identify the nearby venues to each neighbourhood by frequency
- Cluster the neighbourhoods based on each neighbourhood by frequency

Data

To achieve the goals of our project, we will need to get the following datasets

1. The names and postal codes of all neighbourhoods in the city of Casablanca;
2. The coordinates of each neighbourhood;
3. The nearby venues data for each neighbourhood.

The data on postal codes in Casablanca was obtained from a webpage[ii] from the postal service of morocco website. We scraped the webpage using the .read_html() pandas method. After processing the data, we obtained a table containing two columns (The neighbourhood name and its corresponding postal code) and 3048 rows.

The second step of our data gathering process was to get the coordinates (Latitude, Longitude) of each neighbourhood. To do so, we used ArcGIS geocoder. After extracting the coordinates of each neighbourhood, we added two new columns to our previous dataset. This version of the dataset was cleaned to fill missing data and remove duplicate and redundant information.

Finally, we used the Foursquare API to get the nearby venues for each neighbourhood in the processed dataset. This data was then processed and made available for clustering. The final dataset consisted of the neighbourhoods names, coordinates and the 10 most frequent nearby venues for each neighbourhood.

[i] https://www.finances.gov.ma/Publication/depf/2019/profils-regionaux.pdf

[ii] http://www.codepostal.ma/search_mot.aspx?keyword=CASABLANCA