THE BATTLE OF THE NEIGHBOURHOODS: DISCOVERING CASABLANCA

By: Ahmed KCHIKECHE

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 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 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.

DATA SOURCES (1)

- The data on postal codes in Casablanca was obtained from a webpage 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.

DATA SOURCES (2)

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.

DATA SOURCES (3)

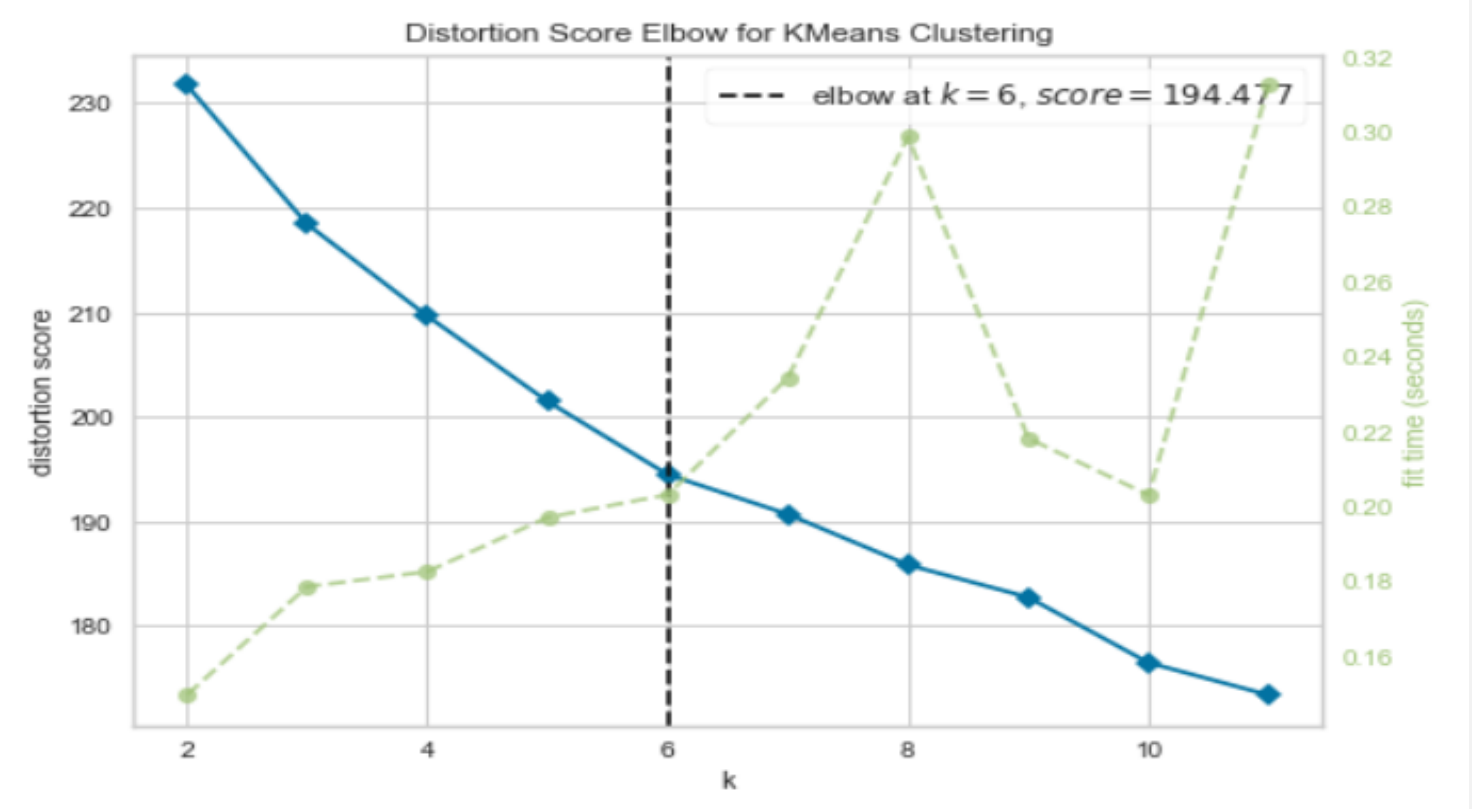
Finally, we used the Foursquare API to get the nearby venues for each neighbourhood in the processed dataset. This data was then processed and made available for clustering. The final dataset consisted of the neighbourhoods names, coordinates and the 10 most frequent nearby venues for each neighbourhood.

FEATURE ENGINEERING

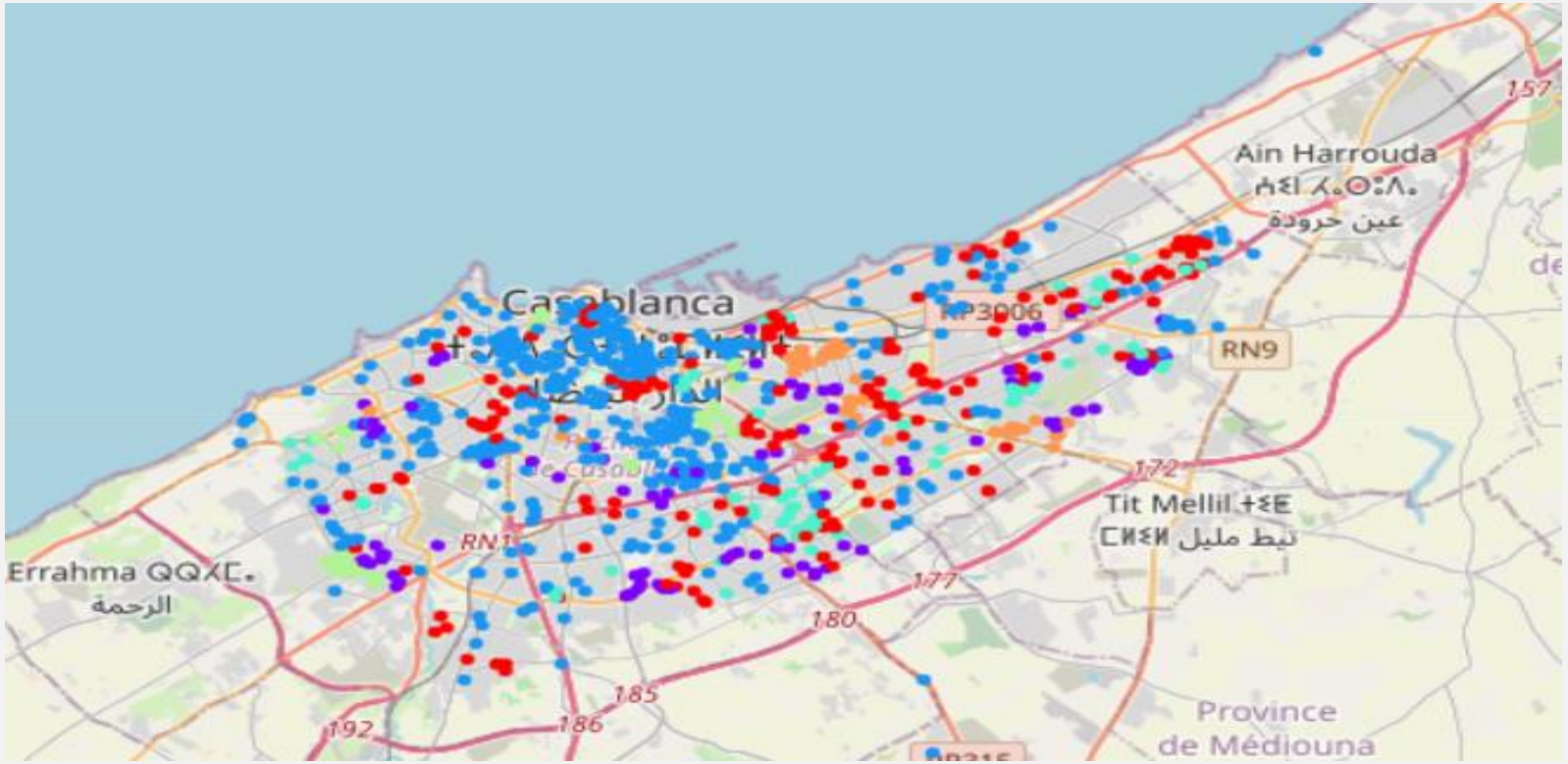
To apply the clustering algorithm to our dataset, we did the following

- Used one-hot encoding to get dummies for our categorical variables;
- Grouped rows by neighbourhood and by taking the mean of the frequency of occurrence of each category;
- Sorted the venues in descending order of venues categories frequency;
- Created a new data frame displaying the top 10 venues for each neighbourhood.

CLUSTERING : THE BEST K



CLUSTERING: RESULTS



RESULTS: NEIGHBORHOODS PER CLUSTER

The cluster	The number of neighbourhoods
0	216
1	121
2	463
3	76
4	54
5	41

RESULTS: CLUSTER 0

The most common venues in cluster 0	Venues categories	frequency
1st Most Common Venue	Coffee Shop	74.5%
	Zoo Exhibit	3.7%
	Fast Food Restaurant	2.3%
2nd Most Common Venue	Coffee Shop	34.3%
	Fast Food Restaurant	7.4%
	Hotel	3.2%
3rd Most Common Venue	Yoga Studio	13.0%
	Coffee Shop	16.7%
	Fast Food Restaurant	6.9%

RESULTS: CLUSTER 1

The most common venues in cluster 1	Venues categories	frequency
1st Most Common Venue	Coffee Shop	61.2%
	Zoo Exhibit	18.2%
	Convenience Store	4.1%
2nd Most Common Venue	Coffee Shop	44.6%
	Fast Food Restaurant	9.9%
	Snack Place	5.0%
3rd Most Common Venue	Coffee Shop	24.0%
	Yoga Studio	9.9%
	Pizza Place	9.9%

RESULTS: CLUSTER 2

The most common venues in cluster 2	Venues categories	Frequency
1st Most Common Venue	Coffee Shop	24.2%
	Fast Food Restaurant	11.2%
	Hotel	6.0%
2nd Most Common Venue	Yoga Studio	10.8%
	Fast Food Restaurant	8.9%
	Hotel	7.3%
3rd Most Common Venue	Coffee Shop	5.4%
	Pizza Place	4.5%
	Bakery	4.1%

RESULTS: CLUSTER 3

The most common venues in cluster 3	Venues categories	Frequency
1st Most Common Venue	Coffee Shop	100.0%
2nd Most Common Venue	Yoga Studio	51.3%
	Ice Cream Shop	6.6%
	Bistro	5.3%
3rd Most Common Venue	Donut Shop	59.2%
	Yoga Studio	32.9%
	Doner Restaurant	3.9%

RESULTS: CLUSTER 4

The most common venues in cluster 4	Venues categories	Frequency
1st Most Common Venue	Shopping Mall	27.8%
	Snack Place	13.0%
	Coffee Shop	18.5%
2nd Most Common Venue	Yoga Studio	10.8%
	Fast Food Restaurant	8.9%
	Hotel	7.3%
3rd Most Common Venue	Shopping Mall	35.2%
	Yoga Studio	11.1%
	Pool Hall	11.1%

RESULTS: CLUSTER 5

The most common venues in cluster 5	Venues categories	Frequency
1st Most Common Venue	Tram Station	46.3%
	Coffee Shop	31.7%
	Hot Dog Joint	12.2%
2nd Most Common Venue	Coffee Shop	51.3%
	Tram Station	6.6%
	Pizza Place	5.3%
3rd Most Common Venue	Donut Shop	31.7%
	Yoga Studio	24.4%
	Doner Restaurant	12.2%