Segmentation and Clustering of Casablanca tram stations

Kenza Boutaleb Houssaïni Joutei

June 6, 2020

1. Introduction

1.1. Background

Casablanca is a city in Morocco which, in the last century, has seen a remarkable evolution in terms of area, population and activities. However, despite becoming an economic capital and one of the big cities in North Africa, public transport has not changed much. Indeed, apart from buses and taxis, its first modern tram line *(T1)* was not put into service until 2012. This line had been so sought-after by citizens that Casa Tramway – *the company in charge of the tram network in Casablanca* – opened a second line (*T2)* in 2019 in order to serve other areas. Despite the progress of the network, several neighborhoods are not served by the tram and have only a few Bus stations and taxis which can be very costly for residents in terms of time and money. Indeed, the city has about fifty neighborhoods grouped into 16 boroughs; while the two lines serve only 16 neighborhoods. The company will continue with two new lines *(T3)* and *(T4)* which are under construction, however it is advantageous to analyze the places in which stations are located or should be located to optimize routes and lines.

1.2. Business problem

Casablanca had a real problem of public transport. Neither bus nor tram station are enough for these 4 million residents. In order to optimize tram routes and lines, we will use this analysis to categorize stations by their surroundings and possibly by district.

1.3. Interest

This project could interest the community service in order to target the neighborhoods that require more public transport in general whether it's a tram or a bus. Also, public transport companies could use this analysis to optimize their stops and stations namely the case of Casa Tramway.

2. Data acquisition and cleaning

2.1. Data sources

For this project we will need:

- The list of Casa Tramway stations with their coordinates,
- The names of Boroughs and Neighborhoods in which belongs the stations,
- The venues surroundings each station.

Unfortunately we didn't find any page or document on the web which provides the coordinates of the stations, this is why we created the *geodata_stations* file which contains the name of the station, the line to which it belongs, the coordinates and the neighborhoods.

We may need the boroughs that we can find on the following pages: *https://fr.wikipedia.org/wiki/Ligne_1_du_tramway_de_Casablanca* for line 1, *https://fr.wikipedia.org/wiki/Ligne_2_du_tramway_de_Casablanca* for line 2.

Finally, we will use the **Foursquare API** to get the venues next to the stations.

2.2. Data cleaning and features selection

Since the file *geodata_stations* was created by us, no cleaning process was done. The dataset is already cleaned and well-structured for our analysis; we just loaded it and change the column names from French to English.

Next, we scraped the previous web pages to get the two tables corresponding to the two tram lines; and we removed columns that we deemed unreliable or unnecessary. We renamed the columns with the same format as previously and we combined the previous dataset with the two new datasets.

We retrieved the venues data with Foursquare API. There are several Macro-categories on Foursquare, the main ones are:

- **Art & Entertainment:** 4d4b7104d754a06370d81259
- **College & University:** 4d4b7105d754a06372d81259
- **Event:** 4d4b7105d754a06373d81259
- **Food:** 4d4b7105d754a06374d81259
- **Outdoors & Recreation:** 4d4b7105d754a06377d81259
- **Professional & other places:** 4d4b7105d754a06375d81259
- **Residence:** 4e67e38e036454776db1fb3a
- **Shop & services:** 4d4b7105d754a06378d81259
- **Travel & transport:** 4d4b7105d754a06379d81259
- **Train station:** 4bf58dd8d48988d129951735
- **Hotel:** 4bf58dd8d48988d1fa931735
- **Bus station:** 4bf58dd8d48988d1fe931735

Unfortunately, the free version of this API didn't allow us to directly retrieve these categories from out dataset. Indeed, the loop exceeded the limit number of regular calls per day. So we imported the micro-categories specific to each place and then we grouped them into 9 macro-categories:

- Coffees & Restaurants
- Entertainments
- Arts & Cultures
- Sport Venues
- Shops
- Food stores
- Hotels & Bars
- Trains & Buses

Our dataframe contains now 433 rows which means 433 venues surrounding Tram Stations and it is shaped into 11 columns:

- **Station:** *name of the Tram station,*
- **Latitude:** *latitude coordinate of the station,*
- **Longitude:** *longitude coordinate of the station,*
- **Neighborhood:** *Neighborhood of the station,*
- **Borough:** B*orough of the stations,*
- **Venue:** *name of the venue,*
- **Venue Latitude:** *latitude coordinate of the station,*
- **Venue Longitude:** *longitude coordinate of the station,*
- **Venue Category:** *Micro-category,*
- **Venue Macro-category:** *Macro-category.*

Below is a view of the final dataset:

	Station	Line	Latitude	Longitude	Neighborhood	Borough	Venue	Venue Latitude	Venue Longitude	Venue Category	Venue Macro-Category
0	Sidi Moumen	T1	33.587459	-7.500901	Sidi Moumen	Sidi Moumen	Café Station Afriquia	33.588429	-7.498692	Coffee Shop	Coffees & Restaurants
1	Nassim	T1	33.585144	-7.504294	Sidi Moumen	Sidi Moumen	Mega dinde	33.585197	-7.499898	Sandwich Place	Coffees & Restaurants
2	Mohammed Zefzaf	T1	33.582373	-7.508647	Sidi Moumen	Sidi Moumen	Play Fitness	33.581276	-7.508755	Gym	Sports venues
3	Mohammed Zefzaf	T1	33.582373	-7.508647	Sidi Moumen	Sidi Moumen	Hadika	33.581673	-7.508580	Gym	Sports venues
4	Mohammed Zefzaf	T1	33.582373	-7.508647	Sidi Moumen	Sidi Moumen	Swag Coffee	33.581523	-7.509139	Coffee Shop	Coffees & Restaurants

Table 1: Cleaned dataset

3. Methodology

Before we start clustering, we did a descriptive statistical analysis on our cleaned data. Since the variables are almost all categorical, the analysis was limited to bar plots, density plots and cross tables. First, we represent the number of venues within each category *[Dominance of categories, fig 1]*; we did the same representation for neighborhoods *[Dominance of neighborhoods, fig 2]*; followed by the number of venues surrounding stations and their densities *[Density of venues by station, fig 3]* and finally we defined the top 5 most common venues for each neighborhood *[Neighborhoods and their top 5 common venues categories, table 2]*.

For clustering we used the K-means algorithm of scikit-learn with k=3.

3.1. Exploratory Analysis

Dominance of categories

Using a bar plot, we can see that the dominant category in our analysis is **Coffees & Restaurants**. However, very few tram stations are close (+ 500 meters) to other transport stations such as **Trains & Buses**. This can have at least one of the following meanings:

- Foursquare does not have enough information on Casablanca transport stations
- The city of Casablanca has few transport lines
- Tram company must add stations closer to train and bus lines.

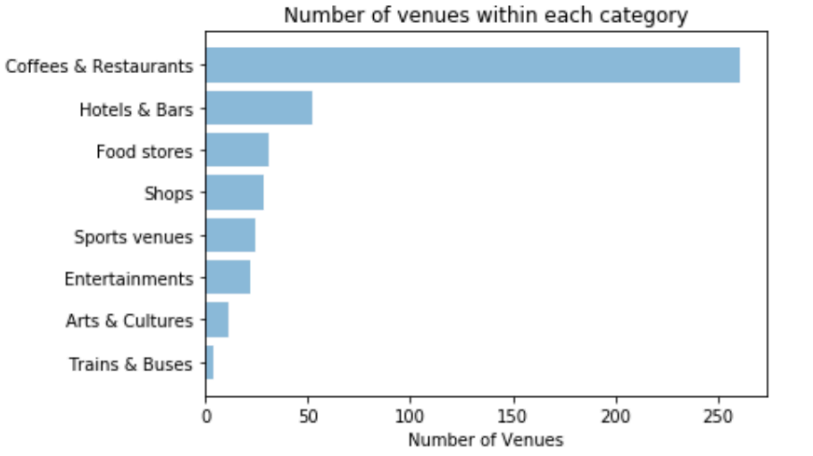

Figure 1: Dominance of categories

Dominance of neighborhoods

According to the number of venues within each neighborhood, the neighborhood "**Centre ville**" which literally means city center, is the most the most requested which is normal. The neighborhoods that follow it such as "**Derb Omar**" and "**Al Fida**" are popular districts where shops and coffees are very commons.

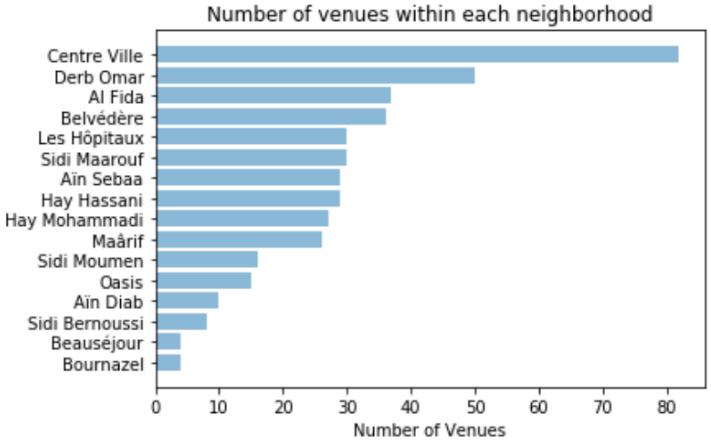

Figure 2 : Dominance of neighborhoods

Density of venues by stations

The following density plot shows that the majority of stations have a number of close venues between 1 and 5, which is very low. We can deduct from this analysis that:

- Either Foursquare doesn't have many registered venues of Casablanca,
- Or tram stations mainly pass through residential areas
- Or tram lines are not optimized.

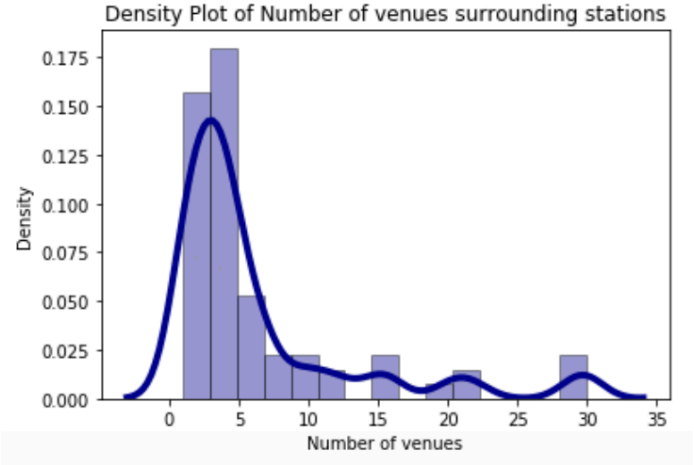

Figure 3: Density of venues by stations

Neighborhoods and their top 5 common venue categories

For more details, we have defined the top 5 common venue categories for each neighborhood. The result shows that the category **Coffees & Restaurants** is the 1st common venue category in almost all neighborhoods. For the rest of variables, the categories **Shops**, **Entertainment** & **Sport Venues** are very common.

Neighborhood	1st Most Common Venue category	2nd Most Common Venue category	3rd Most Common Venue category	4th Most Common Venue category	5th Most Common Venue category
Al Fida	Coffees & Restaurants	Shops	Food stores	Sports venues	Trains & Buses
Aïn Diab	Hotels & Bars	Coffees & Restaurants	Entertainments	Trains & Buses	Sports venues
Aïn Sebaa	Coffees & Restaurants	Shops	Entertainments	Food stores	Trains & Buses
Beauséjour	Coffees & Restaurants	Sports venues	Food stores	Trains & Buses	Shops
Belvédère	Coffees & Restaurants	Hotels & Bars	Shops	Sports venues	Food stores
Bournazel	Coffees & Restaurants	Shops	Entertainments	Trains & Buses	Sports venues
Centre Ville	Coffees & Restaurants	Hotels & Bars	Entertainments	Arts & Cultures	Sports venues
Derb Omar	Coffees & Restaurants	Hotels & Bars	Food stores	Entertainments	Trains & Buses
Hay Hassani	Food stores	Coffees & Restaurants	Sports venues	Entertainments	Shops
Hay Mohammadi	Coffees & Restaurants	Sports venues	Shops	Hotels & Bars	Arts & Cultures
Les Hôpitaux	Coffees & Restaurants	Shops	Hotels & Bars	Sports venues	Arts & Cultures
Maârif	Coffees & Restaurants	Shops	Sports venues	Food stores	Entertainments
Oasis	Coffees & Restaurants	Food stores	Entertainments	Sports venues	Trains & Buses
Sidi Bernoussi	Coffees & Restaurants	Trains & Buses	Sports venues	Shops	Hotels & Bars
Sidi Maarouf	Coffees & Restaurants	Shops	Food stores	Hotels & Bars	Sports venues
Sidi Moumen	Coffees & Restaurants	Sports venues	Trains & Buses	Food stores	Shops

Table 2 : Neighborhoods and their top 5 common venue categories

3.2. Clustering

Since we were going to use the clustering algorithm K-means, we had to define the most optimal K number of clusters for our dataset. We noticed that after 3 clusters, the groups no longer had meaning or were difficult to interpret; this is why we choose K = 3. The map of stations and clusters is as follow:

Figure 4: Map of Casablanca with tram stations and clusters on it

We can distinguish the 3 clusters on the map from the color of stations. At first glance, the red cluster appears to be the group with the most stations while the purple and green clusters appear almost equal.

4. Results

Cluster 0 Red Cluster: The size of the purple cluster represents 60% of the number of stations thus covering 10 neighborhoods out of 16. Its commons venues are **Coffees & Restaurants** and **Food stores**.

Cluster 1 Purple Cluster: The size of the purple cluster represents 14% of the number of stations thus covering 4 neighborhoods out of 16. Its commons venues are **Coffees & Restaurants** and **Hotels & Bars**.

Cluster 2 Green Cluster: The size of the purple cluster represents 26% of the number of stations thus covering 4 neighborhoods out of 16. Its commons venues are **Coffees & Restaurants** and **Shops**.

We can deduce that the most requested neighborhoods which needs more tram stations are the district which contains coffees, restaurants, shops and hotels.

5. Discussion

Foursquare API is a very interesting API but unfortunately not very popular in all countries. Indeed, according to our analysis, the venues and places of Casablanca existing in its database are mainly coffee shops & restaurants. We couldn't find any work areas which somewhat impacted oud analysis. Work areas require a lot of tram and bus stations which would be very useful for us. However, we cannot blame the API only. It is true that the city of Casablanca – despite being a large city and an economic capital of Morocco- had a lack of transport lines and places of entertainment.

6. Conclusion

Through this project, we have tried to group the tram stations according to their surroundings and neighborhoods in order to have an idea on optimizing routes for future stations. Data analysis and clustering allowed us to describe the similarities between stations and the most common venues in the city. The result shows that coffee shops, restaurants, hotels and shops and very present in Casablanca which can be beneficial for Casa Tramway. However, as mentioned in the previous section, many important information is missing due to the unpopularity of the Foursquare App in Morocco. The result could be more promising if residential areas and work areas where listed in Foursquare.