Analyzing Algeria Metro Stations

1. Introduction

1.1 Background & Problem Statement

The Algiers Metro is a rapid transit system serving Algiers, the capital of Algeria. The metro has 19 stations and spanning across 18.5 km migrating passengers daily. For this project, we will look at the neighbourhoods surrounding the metro stations and classify them. Some neighbourhoods are mostly residential, others have more business or commercial spaces surrounding them. The venues closest to a station determine why and how people use it. Thus, by analyzing this data we can classify stations by their primary usage. This data is useful for city planners to determine where from and where to people are most likely to travel for work and leisure. This can help plan further extension of the network and find places for new development.

2. Data & Source

We will need data on the location of stations and on the venues closest to them. So, the below set of data will be used for this project.

1. List of metro stations and their geographical coordinates — scraped from this Wikipedia page.

	Station Name	Coordinates
0	Place des Martyrs	36.78556,3.06222
1	Ali Boumendjel	36.77917,3.05806
2	Tafourah - Grande Poste	36.77194,3.05806
3	Khelifa Boukhalfa	36.76639,3.05361
4	1er Mai	36.76056,3.05528

Stations Data (Sample data of 5 Stations)

2. Foursquare API to explore venue types surrounding each station. Foursquare outlines these high-level venue categories with more sub-categories.

    - Arts & Entertainment (4d4b7104d754a06370d81259)

    - College & University (4d4b7105d754a06372d81259)

    - Event (4d4b7105d754a06373d81259)

    - Food (4d4b7105d754a06374d81259)

    - Nightlife Spot (4d4b7105d754a06376d81259)

    - Outdoors & Recreation (4d4b7105d754a06377d81259)

    - Professional & Other Places (4d4b7105d754a06375d81259)

- Residence (4e67e38e036454776db1fb3a)

- Shop & Service (4d4b7105d754a06378d81259)

- Travel & Transport (4d4b7105d754a06379d81259)

We will query the number of venues in each category in a 1000m radius around each station. This radius was chosen because 1000m is a reasonable walking distance.

3. Methodology

We will use the Foursquare explore API with **category ID** (for each of the venue categories mentioned above) to query the number of venues of each category in a specific radius. The response contains a totalResults value for the specified coordinates, radius and category.

	Station Name	Coordinates	Arts & Entertainment	College & University	Event	Food	Nightlife Spot	Outdoors & Recreation	Professional & Other Places	Residence	Shop & Service	Travel & Transport
0	Place des Martyrs	36.78556,3.06222	5	1	0	3	0	4	6	1	3	4
1	Ali Boumendjel	36.77917,3.05806	3	3	0	10	2	3	4	1	6	5
2	Tafourah - Grande Poste	36.77194,3.05806	4	3	0	12	2	3	3	0	11	6
3	Khelifa Boukhalfa	36.76639,3.05361	4	4	0	9	7	2	4	2	13	6
4	1er Mai	36.76056,3.05528	4	4	0	6	6	4	4	2	10	5

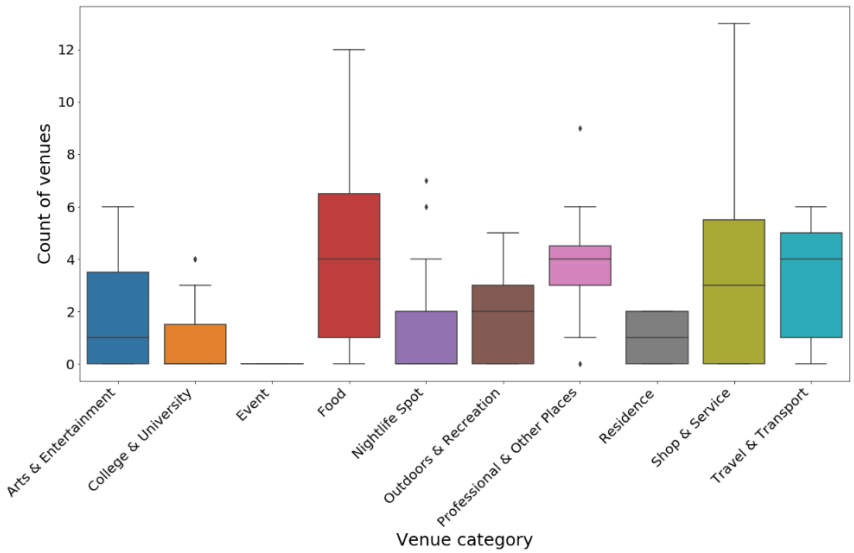
Count of venues of each category in a 1000m radius for each station (Sample data of 5 Stations)

3.1 Exploratory Data Analysis & Data Cleaning

Let's look at the data. From the output we see that **Amirouche** station has the highest number of Professional & Other Places (9) while **El Harrach Centre** station has (0).

Let's display the number of venues as boxplots (showing the average count, spread and outliers).

Boxplots of number of venues in each category

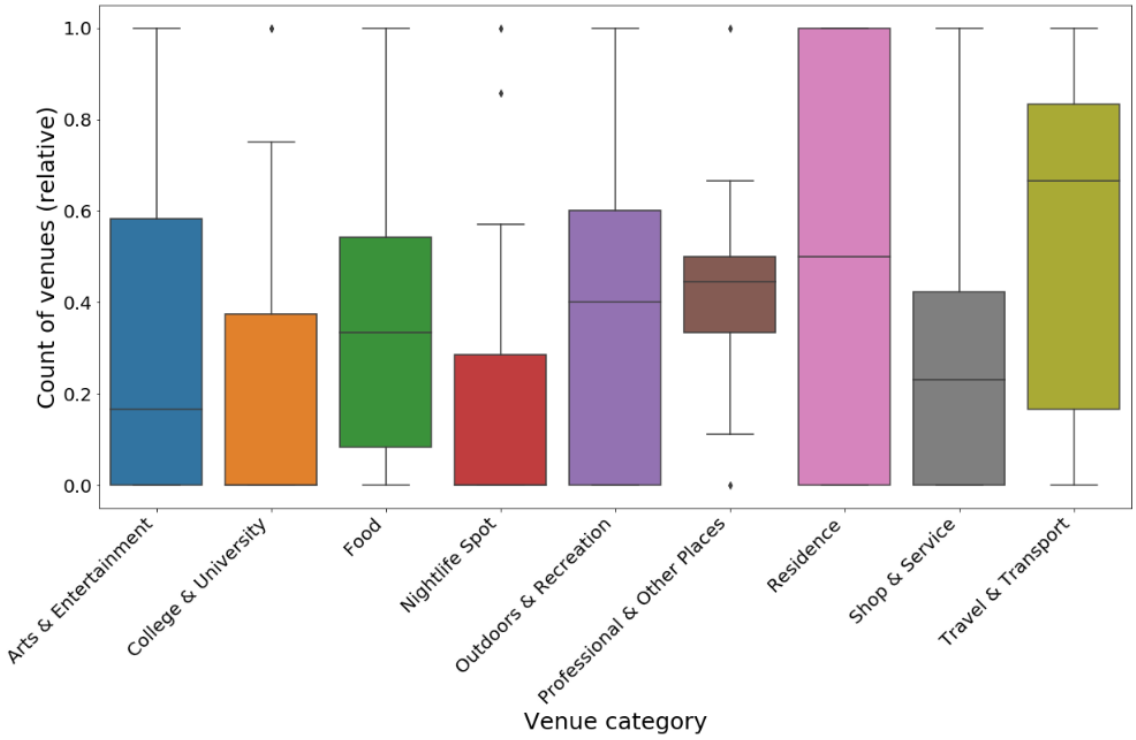


We can see that the most frequent venue categories are Food and Shop & Service. Event has very little data, so will discard it.

3.2 Data Preparation

Let's normalize the data using min-max scaling (scale count of venues from 0 to 1 where 0 is the lowest value in a set and 1 is highest). This both normalizes the data and provides an easy way to interpret the score at the same time. The scaled diagram looks like this:

Boxplots of number of venues in each category



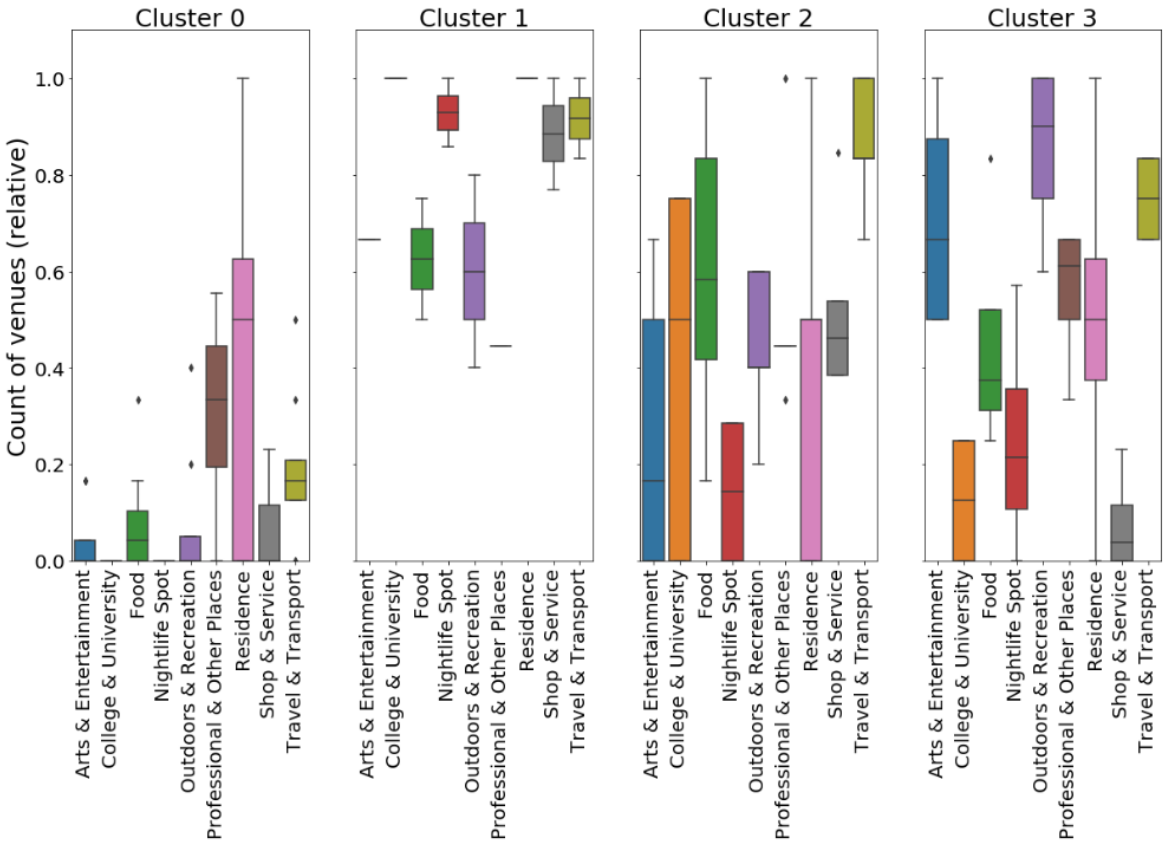
3.3 Clustering

We will be using k-means clustering. These were the preliminary results with different number of clusters:

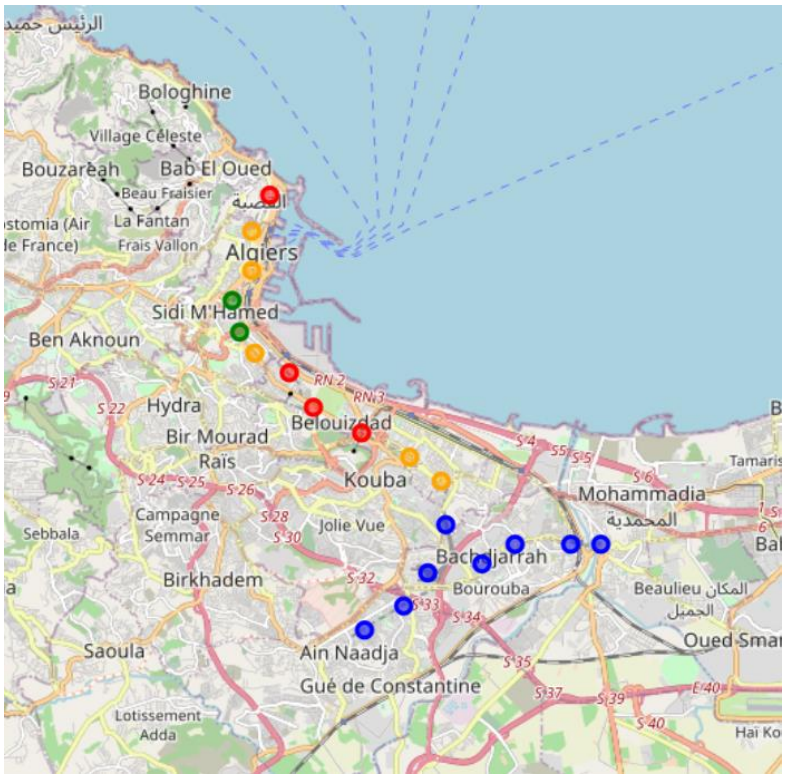
- K=2 clusters don't identify neighborhoods clearly
- K=4 clusters identify neighborhoods with low number of venues
- K=5 and more clusters are difficult to interpret

For the final analysis let's settle on 4 clusters. Let's visualize the clusters profiles using boxplots.

Clusters and their relative count of venues

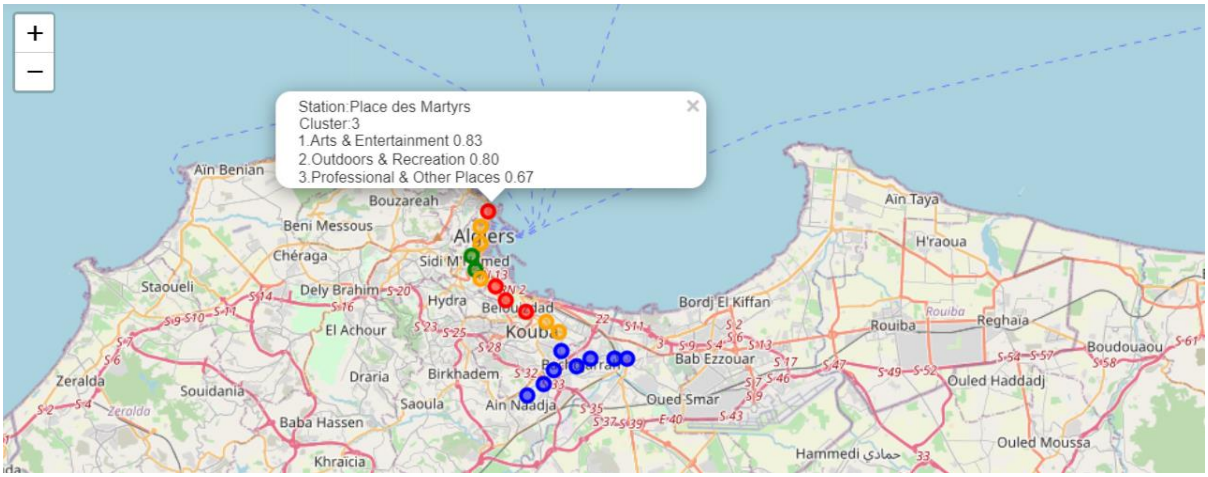


And plotting clusters on a map.



Clusters map. Cluster 0 is Blue, 1 is Green, 2 is Yellow and 3 is Red.

For each station we will display top 3 venue categories and their 0 to 1 scores for this category.



Sample showing scores in top 3 venue categories

4. Results

Here is how we can characterize the clusters by looking at venue scores:

Cluster 0 (Blue) scores high on Residence, Professional & other places. This appears to be residential area as well as business part of the city. Cluster 1 (Green) scores high on nightlife, shops & services and travel & transport. Cluster 2 (Orange) scores high on college & university, food and travel & transport. This appears to be the student hub of the city. Cluster 3 (Red) scores high on outdoor & recreational and art & entertainment.

Plotting the clusters on a map shows us that:

Most of the stations are located along the coastal line which appears to be old and developed part of the city. Cluster 0 is the most inside of the coastal line than other clusters indicating it as primary upcoming areas for residential and business development. The highest number of venues are in the Food and Shop & Service categories.

5. Discussion

The clustering in this analysis has been done basis venues obtained from Foursquare data which is not all encompassing. It does not consider aspects such as footfall in venues, venue's size which can considerably impact the results. Thus, this analysis is far from being conclusory. Furthermore, the results could also potentially vary if we use some other clustering techniques like DBSCAN.

6. Conclusions

Finally, to conclude this project we can say that Foursquare data has certainly given some preliminary information and brief insight into city' development. This data can be combined with other sources to provide more accurate results.