Analyze and cluster Algiers neighborhoods

Miloud DAMENE

March 17, 2020

1. Introduction / Business Problem :

I am an Algerian citizen, and as you may know Algeria is a touristic country, specially the capital “Algiers”, but a lot of people how come to visit Algiers from inside or outside of Algeria don’t know where to go and the neighborhoods where they can find places to where to spend their stay like Hotels, restaurants and so on. So the purpose of this project is to Analyze the neighborhoods of Algiers, and extract different venues in each neighborhoods, then cluster all similar neighborhoods in the same clusters and display this neighborhoods in a map using folium. Also display different venues that exist in a cluster by selecting the cluster to help travelers and employees find and select the venue that they want in the cluster of neighborhoods that they want.

2. Data description and features selection :

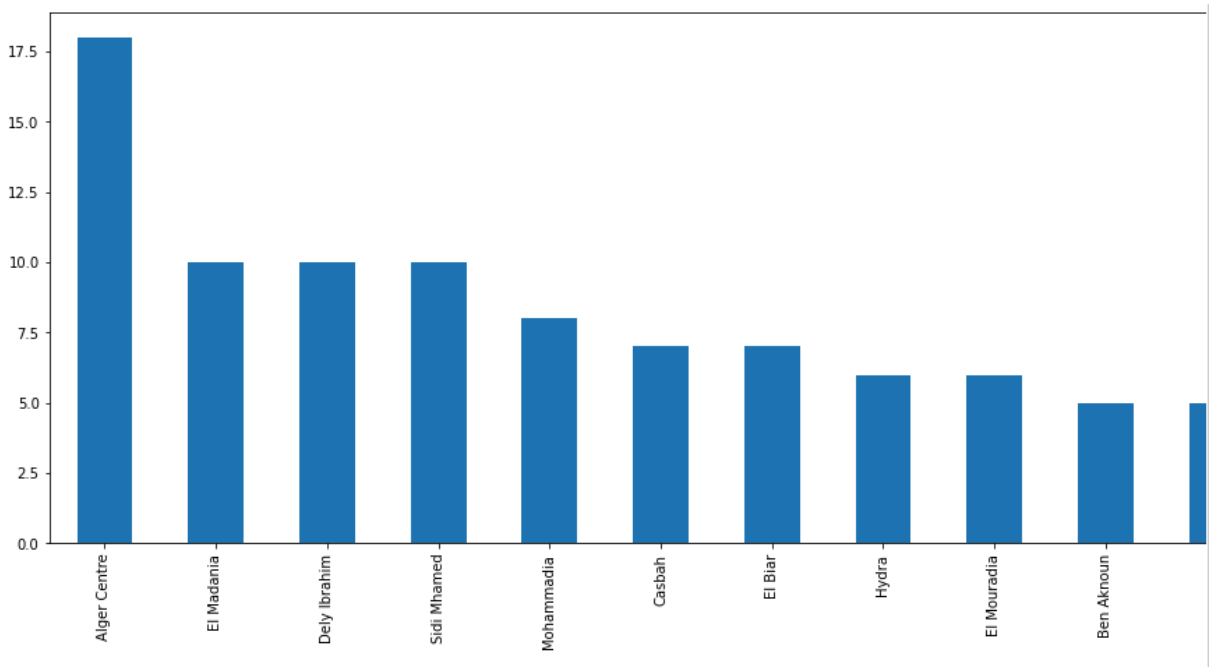
I will use in this project a json file that contain each neighborhood in Algeria, the city id of this neighborhood, postal code, latitude and longitude of the neighborhood.
The purpose of our study is only the analysis of Algiers neighborhoods, so I will select only the neighborhoods that exist in the city of Algiers to create my data frame by filtering in rows with column ‘wilaya_id’ is 16, then make a feature selection to let only the “nom, latitude, longitude” columns in my data.
After that I am going to use the Foursquare API to get the venues in each neighborhood in this data to do all my analysis in the resulting data.

3. Methodology :

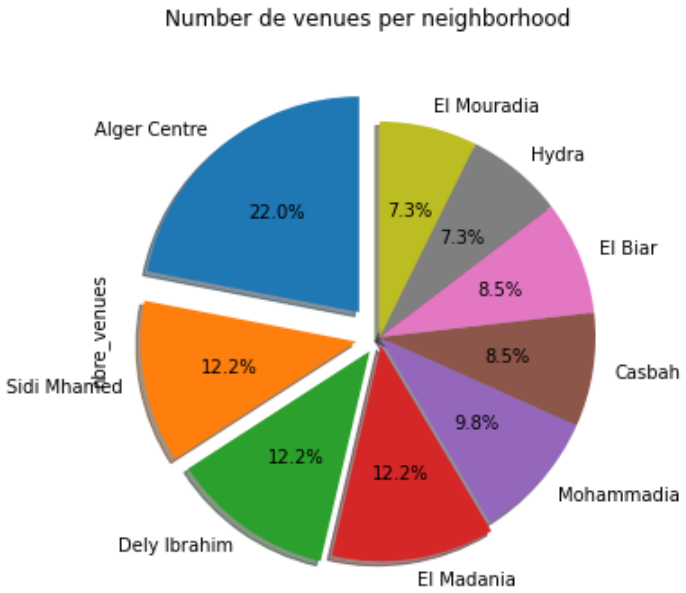
My strategy after getting the venues in each neighborhood is to do some data exploratory to more understand the data and differences between neighborhoods, so I started by classing neighborhoods by descending order of venues number, and the result was as follows:

Neighborhood	
Alger Centre	18
El Madania	10
Dely Ibrahim	10
Sidi Mhamed	10
Mohammadia	8
Casbah	7
El Biar	7
Hydra	6
El Mouradia	6
Ben Aknoun	5
Staoueli	5
Hamma Anassers	5
Cheraga	5
Kouba	5

To visualize this data more clearly, I create a bar plot of neighborhood's venues using matplotlib library to get this plot :



And to show the percentage of each neighborhood I created a pie plot :



Now that I got an Idea about the data, I started the preprocessing and structuring of the data to use it for the clustering of neighborhoods. Firstly, by the one hot encoding in venues categories I created a data frame that for each row (venue) have 1 in the column that represent the category of this venue, 0 else. And then by this data frame I calculated the mean of existence of each venue category in each neighborhood.

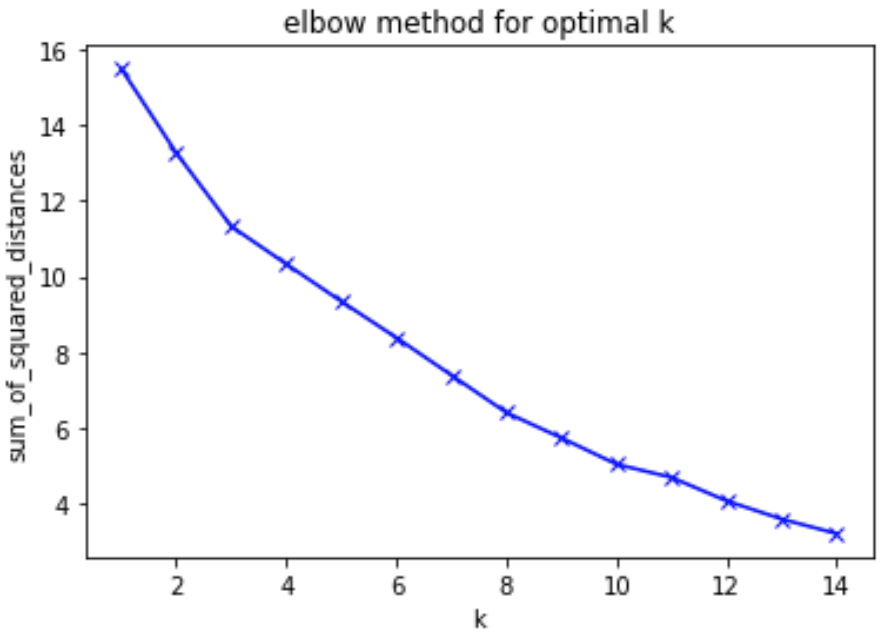
	Neighborhood	African Restaurant	American Restaurant	Art Museum	BBQ Joint	Bar	Beach	Burger Joint	Bus Station	Cafeteria	Caf
0	Alger Centre	0.0	0.0	0.055556	0.055556	0.0	0.0	0.0	0.0	0.0	0.16666
1	Bab Azzouar	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00000
2	Bab El Oued	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00000
3	Baba Hassen	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	0.0	0.0	0.00000
4	Bachedjerah	0.0	0.0	0.000000	0.000000	0.0	0.0	0.0	1.0	0.0	0.00000

Secondly, I created a function that return the top venues of a neighborhood, and I called this function for each neighborhood to get the top 10 venues for each neighborhood :

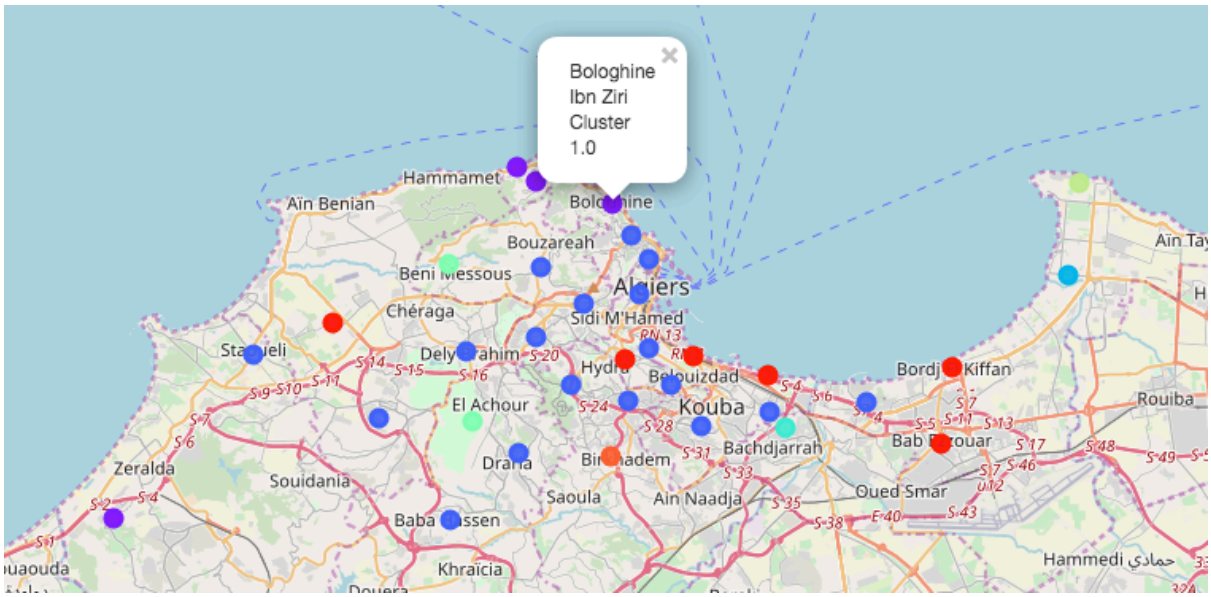
	Neighborhood	1st Most Common Venue	2nd Most Common Venue	3rd Most Common Venue	4th Most Common Venue	5th Most Common Venue	6th Most Common Venue	7th Most Common Venue	8th Most Common Venue	9
0	Alger Centre	Hotel	Café	Restaurant	Historic Site	Italian Restaurant	Lounge	French Restaurant	Mediterranean Restaurant	F R
1	Bab Azzouar	Hotel	Halal Restaurant	Cosmetics Shop	Harbor / Marina	Gym / Fitness Center	Gym	Garden	Furniture / Home Store	R
2	Bab El Oued	Fast Food Restaurant	Park	Plaza	Video Game Store	Historic Site	Halal Restaurant	Gym / Fitness Center	Gym	
3	Baba Hassen	Restaurant	Middle Eastern Restaurant	Fast Food Restaurant	Construction & Landscaping	Cosmetics Shop	Halal Restaurant	Gym / Fitness Center	Gym	
4	Bachedjerah	Bus Station	Video Game Store	Cosmetics Shop	Harbor / Marina	Halal Restaurant	Gym / Fitness Center	Gym	Garden	F

This data will be used to fit the kmeans model to cluster neighborhoods by the top venues category that exist in each neighborhood.

For the clustering, I used kmeans with different number of clusters in each iteration from 1 to 15 and calculate the sum of squared distances of samples to the nearest cluster center in each clustering, then use the elbow method to decide which number of clusters to choose. In my case I decided that number of clusters be 9.



Finally, I used folium to display clusters of neighborhood in the map :



4. Results and discussion section:

- ‘Alger Centre neighborhood’ has more venues than all other neighborhood.
- Hotels and restaurants are the more occurred categories of venues occurred in different neighborhood.
- There is a big similarity between neighborhoods of Algiers in the categories of venues.
- Big part of Algiers neighborhoods are in the cluster 2.
- Neighborhoods of cluster 2 have more hotels, restaurants, coffee and Pizza places.
- Neighborhoods of cluster 1 have beaches, churchs, touristic places, video game stores.
- Neighborhoods of cluster 2 are more about living and working.
- Neighborhoods of cluster 1 are more about going on vacation.

5. Conclusion section:

For Travelers:

- I recommend the neighborhoods of cluster 1, so **Bologhine**, **Bains Romains**, **Rais Hamidou** and **Zeralda** because they all have the good beaches in Algiers and touristic places.

For Employees:

- I recommend the neighborhoods of cluster 2, like **Alger centre, Casbah, Bab El Oued, Mohammadia…etc.** because they will find all what they need like restaurants, subways, gyms… etc.