Capstone Project Report - Moroccan Cities Clustering

March 14, 2019

1 Introduction

In this project, we will try to cluster the cities of my own country Morocco based on their characteristics. In order to know the similar cities and to recommend new cities from people's preferences. This recommender systems may be utilized in a variety of areas. It may be used, for example, as a method for assessing tourists. It can also be extended to different compare different cities from different countries. The data collected will also be used for clustering the cities based on the presence of a single venue. To decide, for example where to invest his money on a project.

2 Data Description

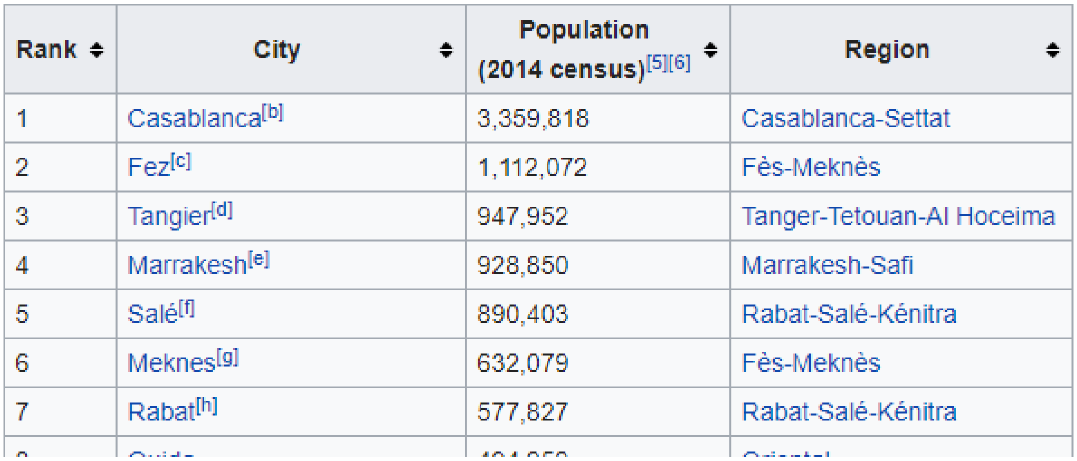
The data that we'll be using is constructed of the cities names and the characteristics of each city based on the Foursquare location data. We will use a Wikipedia page to extract the Moroccan cities with their population, latitudes and longitudes. Then, we will use the foursquare API to get the venues of each city. The dataset created will be used in the recommender system.

3 Methodology

3.1. Get the list of Moroccan cities

```
In [11]:
```

```
Out[11]:
```

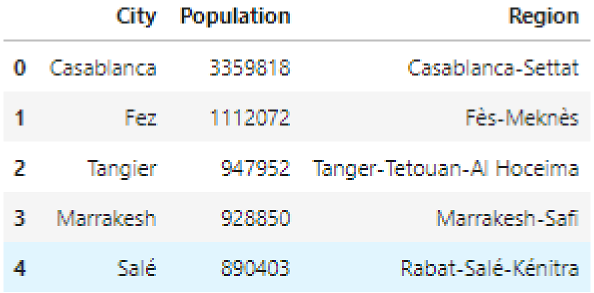


Rank	City	Population (2014 census)[5][6]	Region
1	Casablanca[b]	3,359,818	Casablanca-Settat
2	Fez[c]	1,112,072	Fès-Meknès
3	Tangier[d]	947,952	Tanger-Tetouan-Al Hoceima
4	Marrakesh[e]	928,850	Marrakesh-Safi
5	Salé[f]	890,403	Rabat-Salé-Kénitra
6	Meknes[g]	632,079	Fès-Meknès
7	Rabat[h]	577,827	Rabat-Salé-Kénitra

3.2. Convert data to numeric

In [13]:

Out[13]:

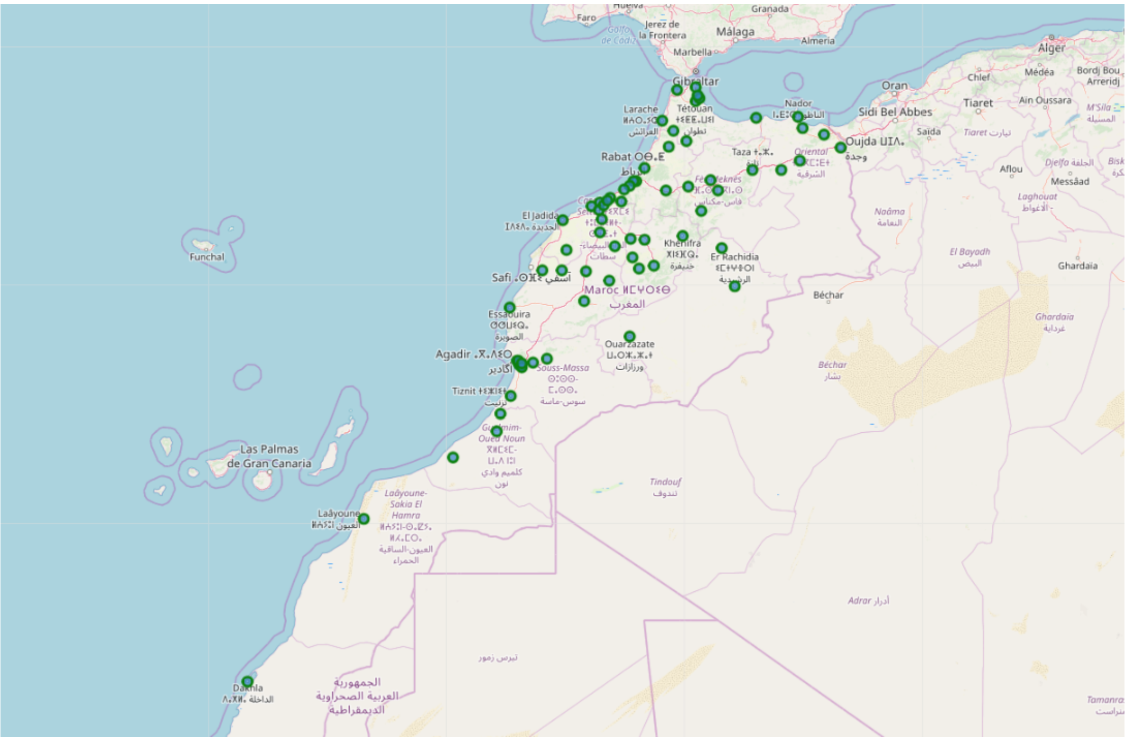


	City	Population	Region
0	Casablanca	3359818	Casablanca-Settat
1	Fez	1112072	Fès-Meknès
2	Tangier	947952	Tanger-Tetouan-Al Hoceima
3	Marrakesh	928850	Marrakesh-Safi
4	Salé	890403	Rabat-Salé-Kénitra

3.3. Get latitude and longitude from address

In [6]:

Out[6]:



2

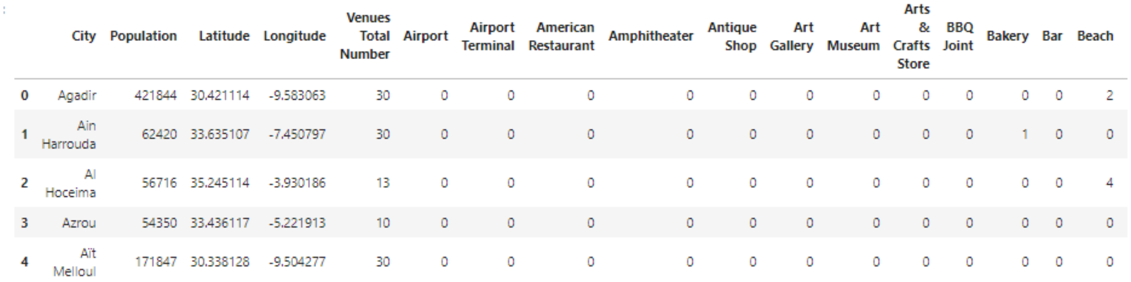
3.4. Define Foursquare Credentials and Version

3.5. Get the venues of each city

3.6. Finalize the dataset by regrouping the population the venues and the total venues of each city

```
In [7]:
```

```
Out[7]:
```



	City	Population	Latitude	Longitude	Venues Total Number	Airport	Airport Terminal	American Restaurant	Amphitheater	Antique Shop	Art Gallery	Art Museum	Arts & Crafts Store	BBQ Joint	Bakery	Bar	Beach
0	Agadir	421844	30.421114	-9.583063	30	0	0	0	0	0	0	0	0	0	0	0	2
1	Ain Harrouda	62420	33.635107	-7.450797	30	0	0	0	0	0	0	0	0	0	1	0	0
2	Al Hoceima	56716	35.245114	-3.930186	13	0	0	0	0	0	0	0	0	0	0	0	4
3	Azrou	54350	33.436117	-5.221913	10	0	0	0	0	0	0	0	0	0	0	0	0
4	Aït Melloul	171847	30.338128	-9.504277	30	0	0	0	0	0	0	0	0	0	0	0	0

As you can see the data we'll be using is given by the data frame "data". It contains the the Moroccan cities their population, latitude, longitude and the number of venues with their types. This data is used for clustering in the next section

3.7. Data clustering After creating the dataset, in this section we will use it for clustering the cities based on their characteristics. We will be using the *k*-means which is vastly used for clustering in many data science applications.

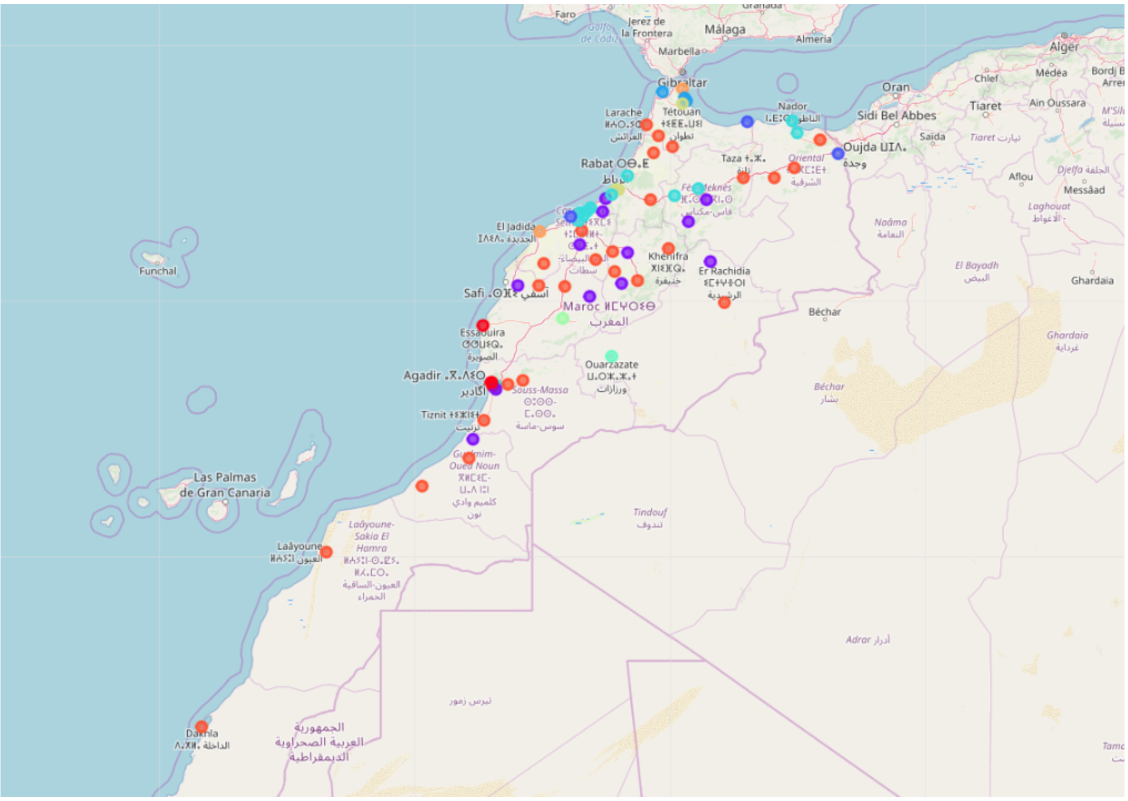
The KMeans class has many parameters that can be used, but we will use these three:

- init : Initialization method of the centroids. Value will be: "k-means++". k-means++ selects initial cluster centers for k-means clustering in a smart way to speed up convergence.
- n_clusters : The number of clusters to form as well as the number of centroids to generate. Value will be: 10
- n_init : Number of times the k-means algorithm will be run with different centroid seeds. The final results will be the best output of n_init consecutive runs in terms of inertia. Value will be: 12

Finally, we visualize the resulting clusters

```
In [8]:
```

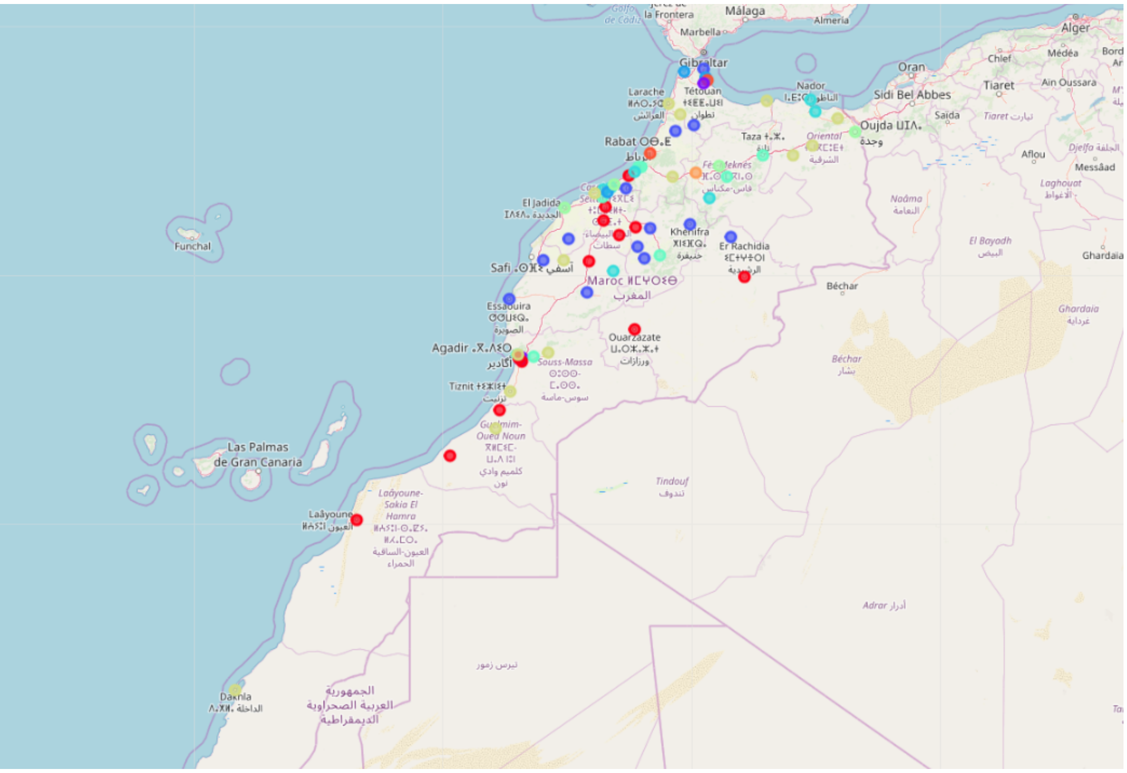
```
Out[8]:
```



3.4. Moroccan cities clustering based on the presence of coffees We can use the data also to cluster the moroccan cities based on the presence of a single venues. In order, for example, to decide where to invest on a projet. In the next part, we will use the data to cluster the cities based on the presence of coffees.

In [9]:

Out[9]:



4 Results and Discussion

The first use of the dataset was for clustering the Moroccan cities based on their venues. We can see the association between the similar cities on the map. For the next part, our analysis shows that there exist a lot of cities with a low number of coffees. Even if the data is not 100 % correct, it is significant enough from my knowledge about the Moroccan cities. The result of the analysis is 14 cities with a low number of coffees, which may be good places for such projects. However, it may not be a good idea to invest in some cities, taking into account the population and its culture. Recommended zones should therefore be considered only as a starting point for more detailed analysis which could eventually result in locations which have not only no nearby competition but also other factors taken into account and all other relevant conditions met.

5 Conclusion

Purpose of this project was to cluster the cities of my own country Morocco based on their characteristics. The to cluster the cities based on the presence of single venue. We used a Wikipedia page to extract the Moroccan cities, their populations, latitudes and longitudes. Then we used the foursquare API to get the venues in each city. The data collected was used to cluster the cities to see the distribution of the venues in Morocco. The dataset was also use to cluster the cities based on the presence of coffees. We've grouped the cities into 10 clusters from cities with low number of coffees to cities with high number of coffees. The results is 14 cities with low number of coffees what means low competition. However, further analysis may be needed to make a more precise decision where to invest his money.