Clustering Moroccan citites

A .Belcaid

Coursera Capstone

June 4, 2020

1. Introduction
2. Loading the data
3. Exploratory data analysis
4. Clusetring the data

Project description

The goal of this project it to cluster the cities of my country Morocco. The clustering algorithm will use geographical data such as population, number of hotels and number and type of industries. This clustering could serve for serveral purposes:

- Say I had to move from my current city, I would like to choose another city which is similar to my current city.
- For a foreign tourist, Say you visited a city A and you liked it but didn't like city B. In future visit to Morocco, you'll would like to avoid all the cities in the B cluster and try to discover more cities in the A cluster.

Recommander

Better application for this would be a Recommender system but can also use clustering.

Loading the data

The data was already prepared in the previous week. It is a list of Moroccan cities with a set of features

City	Population	Region	latitude	longitude	Café	Hotel	Moroccan Restaurant	Coffee Shop	Diner
Casablanca	3359818	Casablanca-Settat	33.595063	-7.618777	4.0	6.0	3.0	1.0	1.0
Fez	1112072	Fès-Meknès	34.034653	-5.016193	0.0	0.0	1.0	1.0	0.0
Tangier	947952	Tanger-Tetouan-Al Hoceima	35.777103	-5.803792	4.0	3.0	1.0	0.0	5.0
Marrakesh	928850	Marrakesh-Safi	31.625826	-7.989161	6.0	11.0	12.0	0.0	0.0
Salé	890403	Rabat-Salé-Kénitra	34.044889	-6.814017	1.0	0.0	0.0	0.0	0.0

Figure: Moroccan cities clustering data

Here we plot a

- Box plot
- Distribution plot

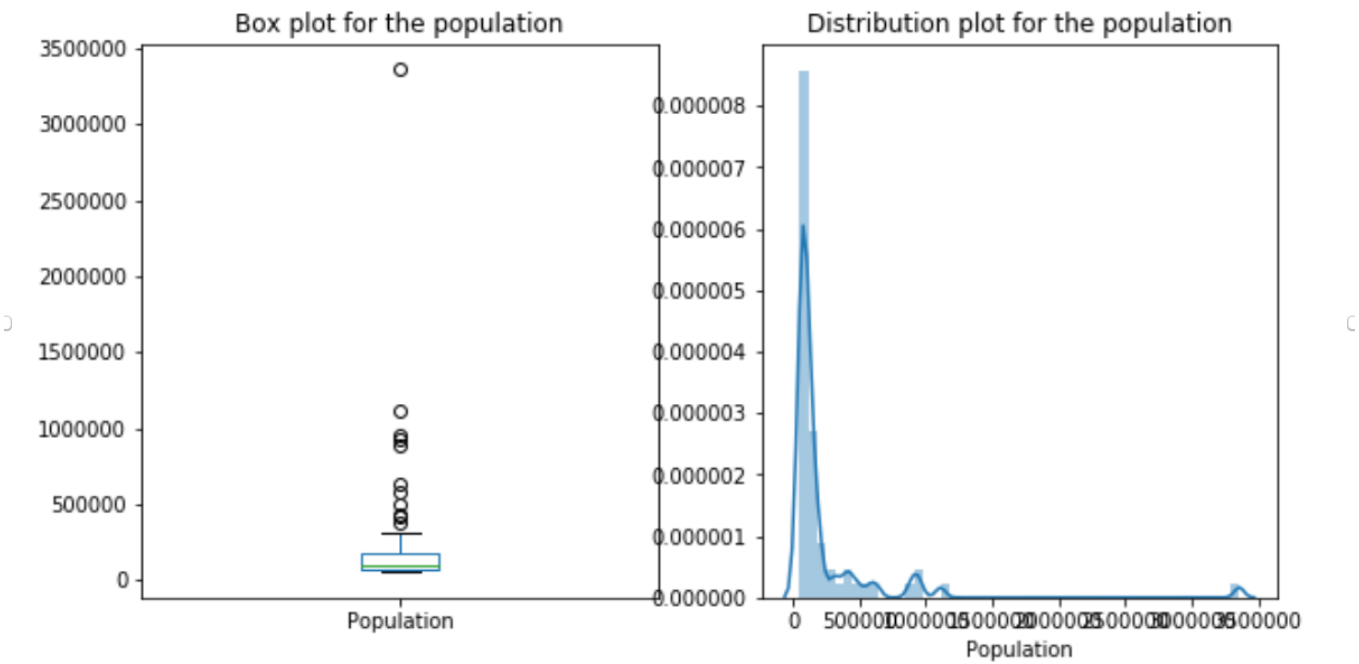

Figure: Disribution and box plot for the population

Café venue

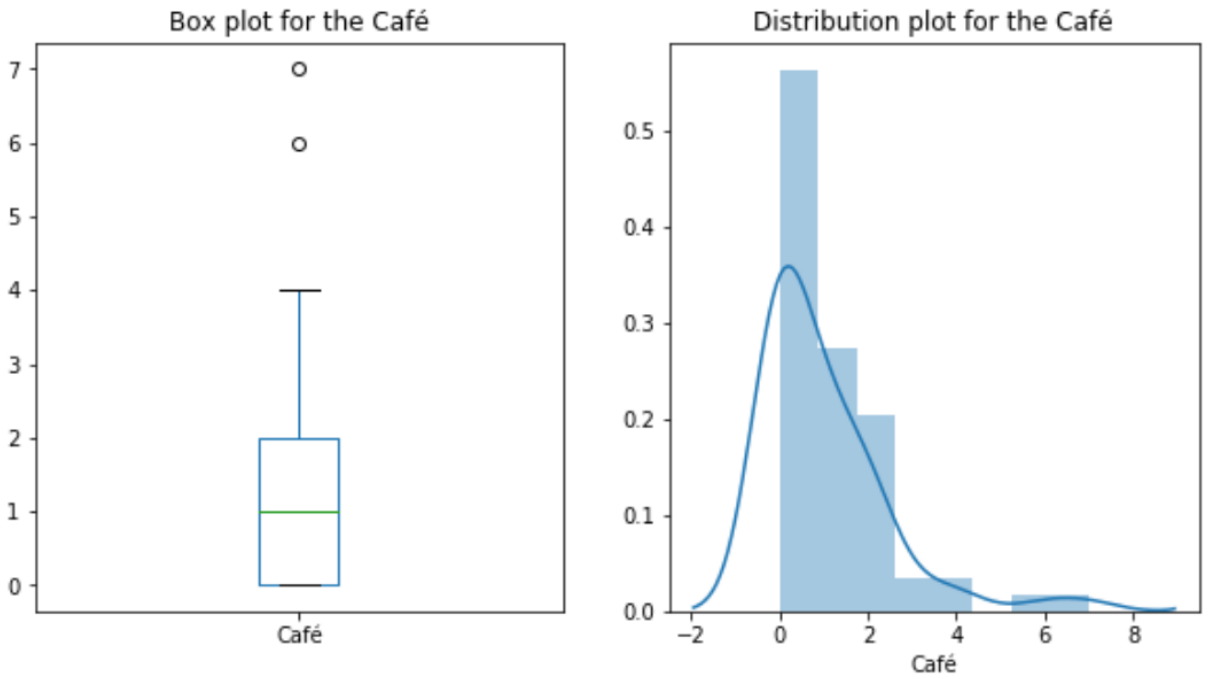

Figure: Box and distributin plot for the Café venue

Correlation

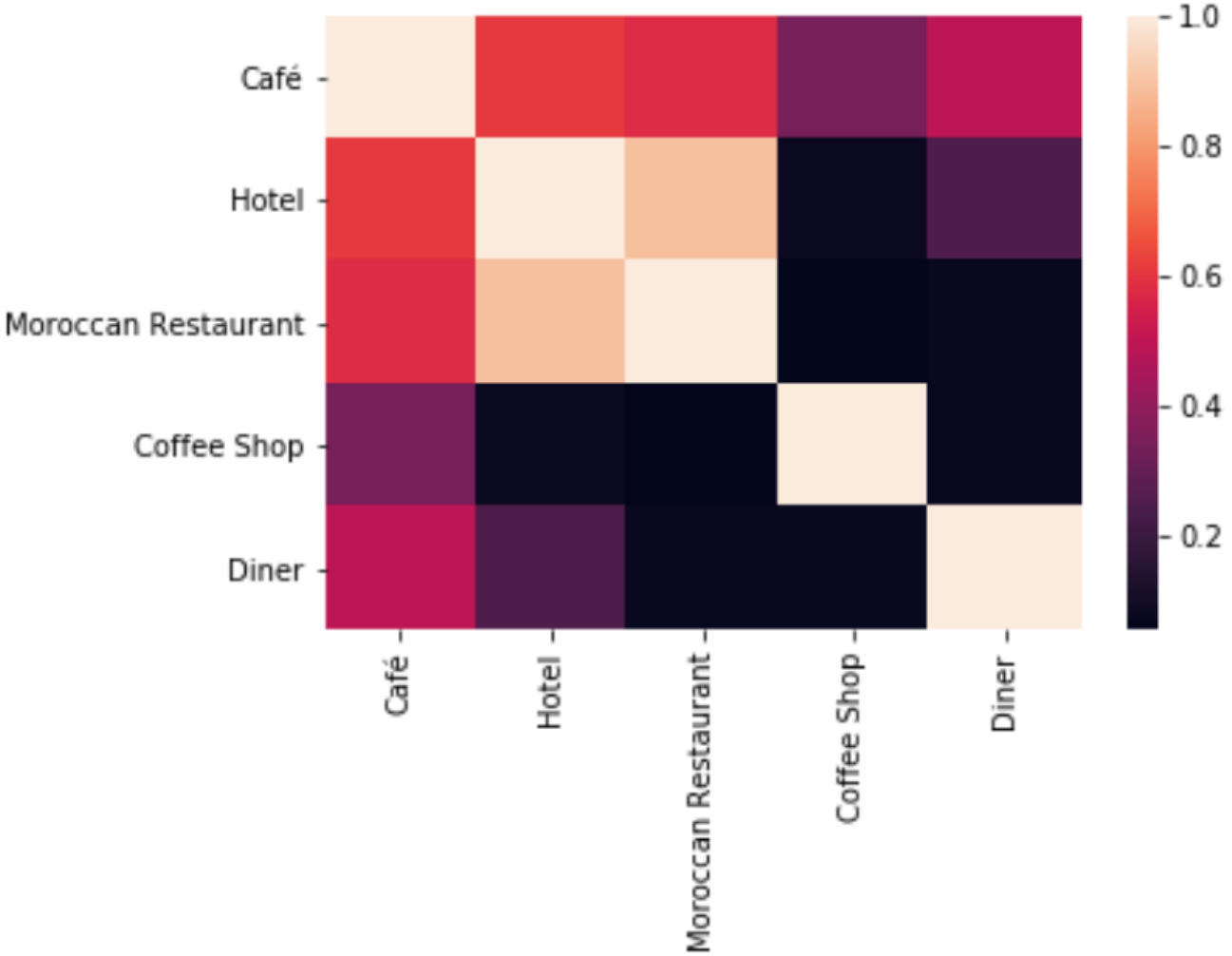

Figure: Correlation between the venues seeked from foursquare

Confirming the correlation

Let confirm this correlation by plotting the linear regression between **Café** and **Hotel** and **Moroccan Restaurant**

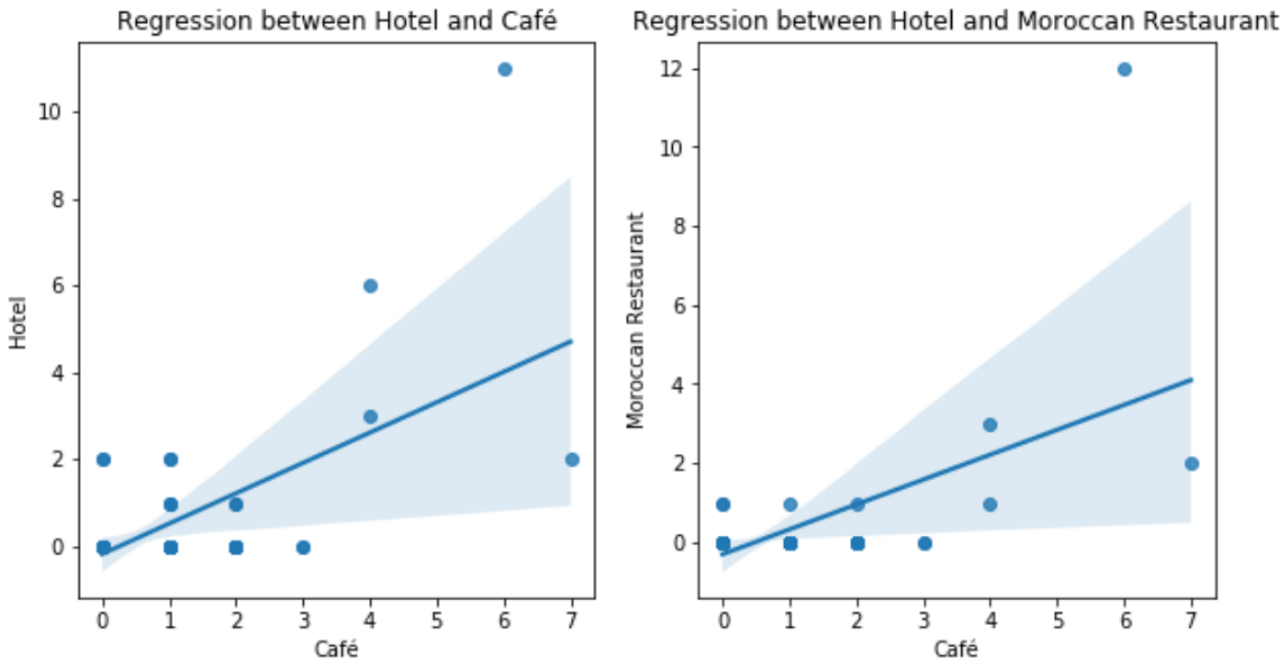

Figure: Linear regression between venues variables

Data Normalization

First, we need to normalize the data using `StandardNormalizer` from scikit learn

City	Population	latitude	longitude	Café	Hotel	Moroccan Restaurant	Coffee Shop	Diner
Casablanca	6.950718	-0.035651	-0.196372	2.137331	3.411629	1.757776	1.125446	0.956183
Fez	1.941305	0.097944	-0.034972	-0.712444	-0.325360	0.441899	1.125446	-0.378489
Tangier	1.575541	0.627486	-0.083815	2.137331	1.543135	0.441899	-0.478913	6.294871
Marrakesh	1.532969	-0.634115	-0.219342	3.562218	6.525787	7.679225	-0.478913	-0.378489
Salé	1.447285	0.101055	-0.146465	0.000000	-0.325360	-0.216040	-0.478913	-0.378489
...	...	...	...	...	...	...	...	...
M'diq	-0.411792	0.598997	-0.054012	1.424887	-0.325360	-0.216040	-0.478913	-0.378489
Sidi Bennour	-0.412710	-0.322625	-0.246321	0.712444	-0.325360	-0.216040	1.125446	0.956183
Midelt	-0.413849	-0.313639	-0.017837	0.000000	0.297472	-0.216040	1.125446	-0.378489
Azrou	-0.415975	-0.083955	-0.047730	0.000000	0.297472	-0.216040	-0.478913	-0.378489
Drargua	-0.423561	-1.012113	-0.311575	-0.712444	-0.325360	-0.216040	-0.478913	-0.378489

Figure: normalized data

Plotting the clustered data

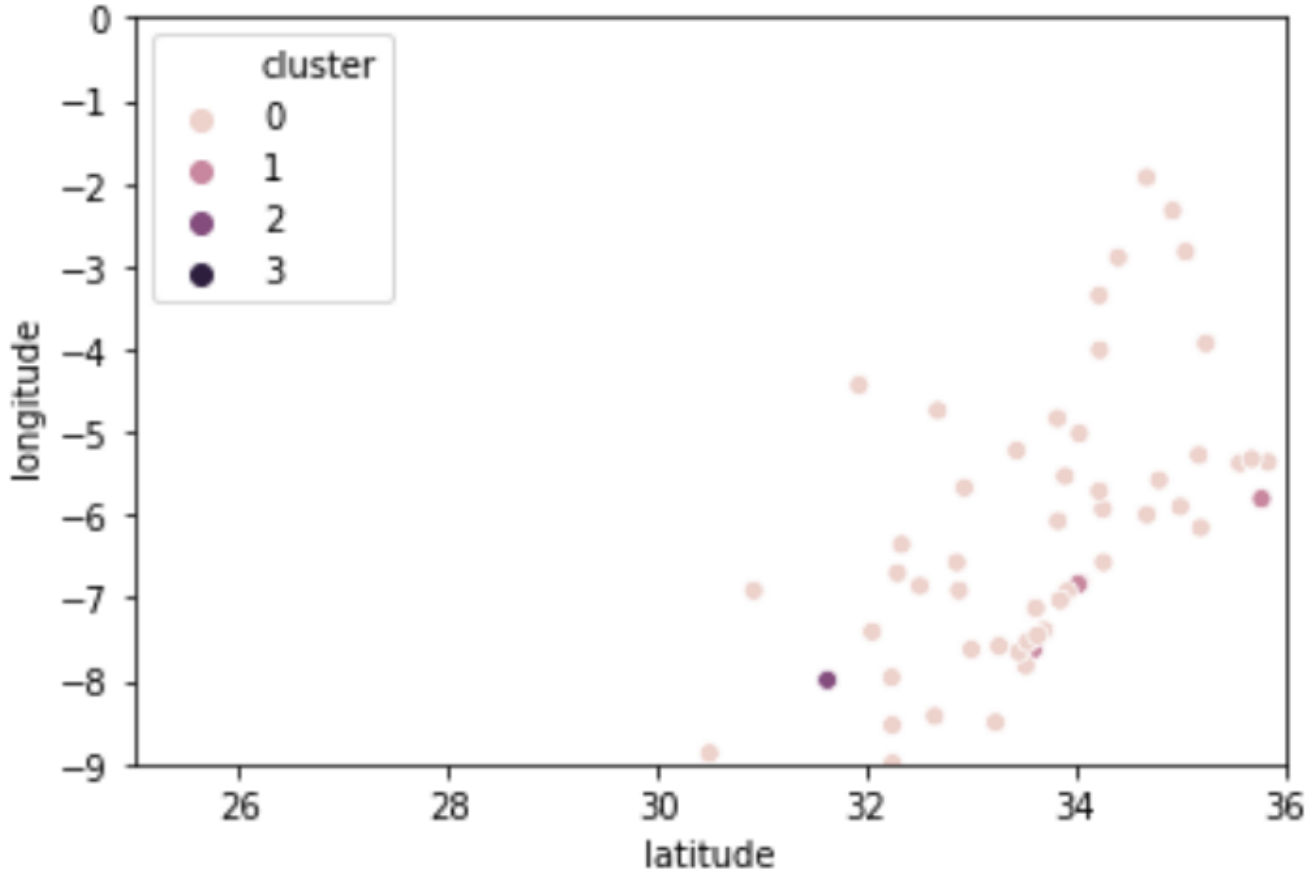

Figure: Results of the clustering, we remak except for the two big cities (Casablanca, and Rabat), all the cities have the same label 0.