Capstone Project - The Battle of Neighborhoods (Week 2):

Clustering locations in Paris based on Restaurant Types

Maaouia Ben Traki

April 21,2020

1. Introduction

1.1. Background

Faced with the shortage of computer engineers(Developers, specialists in cyber-security, data science/analyst,…)in recent years, France turns to Maghreb countries such as Morocco and Tunisia to have the profiles sought in order to succeed in their digital transformations. The majority of the demands are mainly based in the capital, Paris[1].

This context would provide great opportunity for opening restaurants with Maghreb foods in Paris.

1.2. Problem

Putting a food business in the proper location, might be the single most important thing to do at startup in order to promote its success.

This project, aims to find the best location for opening a new restaurant by leveraging the Foursquare location data to explore Paris neighborhoods restaurants.

1.3. Interest

The group of people that would be interested by this project are:

- Investors or business people with the ambition of starting up a restaurant in Paris. This analysis will provide them a comprehensive guide of neighborhoods classification based on food category.
- Workers who want to find desirable food places close to their offices.
- Students, or Data Science Enthusiast interested in intelligence location.

2. Data Acquisition and Cleaning

2.1. Data sources

Paris has a total of 20 boroughs and 80 neighborhoods. In order to segment the neighborhoods and explore them, we will essentially need a dataset that contains the 20 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

This dataset named quartier_paris.csv, exists for free on the web. Here is the link to the dataset[2]: https://www.data.gouv.fr/fr/datasets/quartiers-administratifs/

You find below snapshot of the initial dataset:

```
[61]: df.head()
```

[61]:

	N_SQ_QU	C_QU	C_QUINSEE	L_QU	C_AR	N_SQ_AR	PERIMETRE	SURFACE	Geometry X Y	Geometry
0	750000036	36	7510904	Rochechouart	9	750000009	2862.450525	5.004354e+05	48.8798119198,2.344861291	{"type": "Polygon", "coordinates": [[[2.349708...
1	750000047	47	7511203	Bercy	12	750000012	6155.005036	1.902932e+06	48.8352090499,2.38621008421	{"type": "Polygon", "coordinates": [[[2.391141...
2	750000002	2	7510102	Halles	1	750000001	2606.417128	4.124585e+05	48.8622891081,2.34489885831	{"type": "Polygon", "coordinates": [[[2.349365...
3	750000015	15	7510403	Arsenal	4	750000004	2878.559656	4.872649e+05	48.851585175,2.36476795387	{"type": "Polygon", "coordinates": [[[2.368512...
4	750000018	18	7510502	Jardin-des-Plantes	5	750000005	4052.729521	7.983894e+05	48.8419401934,2.35689388962	{"type": "Polygon", "coordinates": [[[2.364561...

2.2. Data cleaning

We start by formatting the column names and add Longitude and Lattitude columns, the new dataframe looks as following:

In france the Borough est named by the their number followed by a name, so I kept the number as Borough identifer for simplicity:

```
[16]: neighborhoods.head()
```

[16]:

	Borough	Neighborhood	Latitude	Longitude
0	9	Rochechouart	48.879812	2.344861
1	12	Bercy	48.835209	2.386210
2	1	Halles	48.862289	2.344899
3	4	Arsenal	48.851585	2.364768
4	5	Jardin-des-Plantes	48.841940	2.356894

The dataframe contains 20 boroughs and 80 neighborhoods.The boroughs are identified by their number from 1 to 20.

2.3. Foursquare API Location Data :

In order to get data about different Restaurants of Paris, we will use the Foursquare API to explore their neighborhoods venues.

Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos.

Paris neighborhoods geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each.

We will limit to 100 venues and at radius of 500 meters for each borough from their given latitude and longitude information. Here is the header of the result, adding venue id, venue name, category, latitude, and longitude information from Foursquare API.

[30]:
```
Paris_restaurants.head()
```

[30]:

	Neighborhood	Neighborhood Latitude	Neighborhood Longitude	Venue	Venue Latitude	Venue Longitude	Venue Category
0	Rochechouart	48.879812	2.344861	Mamiche	48.880112	2.343699	Bakery
1	Rochechouart	48.879812	2.344861	Lakshmi Bhavan	48.881077	2.344232	Indian Restaurant
2	Rochechouart	48.879812	2.344861	Le Barbe à Papa	48.879654	2.347438	French Restaurant
3	Rochechouart	48.879812	2.344861	Corso Trudaine	48.881794	2.345264	Italian Restaurant
4	Rochechouart	48.879812	2.344861	Pizza di Loretta	48.880634	2.344011	Pizza Place

The total dataframe ,named as Paris_restaurants, has 3741 rows and 7 columns with 153 unique categories.

3. Methodology

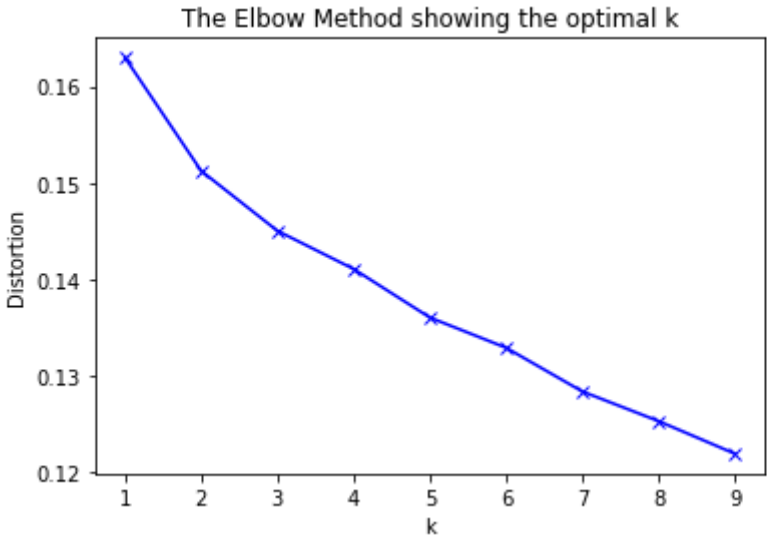
3.1. Cluster Neighborhoods Restaurants :

We will use the KMeans algorithm to cluster our dataframe. The elbow method helps us to find the optimal value for number of clusters k.

We can see in the figure below ,when k increases, the centroids are closer to the clusters centroids.

The improvements will decline, at some point rapidly, creating the elbow shape.

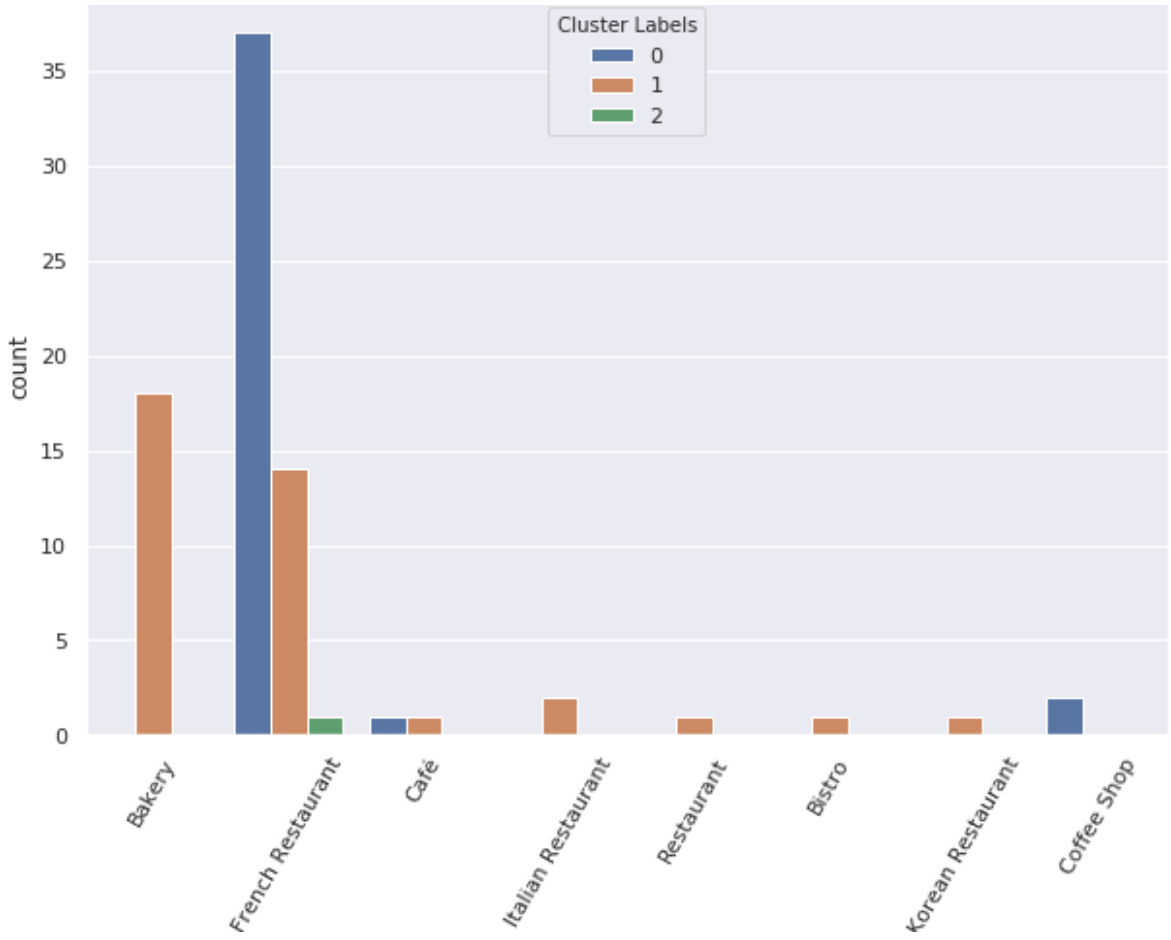
The plot looks like an arm with an elbow at k = 3 ,so that point is the optimal value for k.



4. Results

Using k-means we have segmented Paris neighborhoods into three groups since we specified the algorithm to generate 3 clusters. The neighborhood in each cluster are similar to each other in terms of the features included in the dataset.

Now we can create a profile for each group, considering the common characteristics of each cluster according to the following figure:

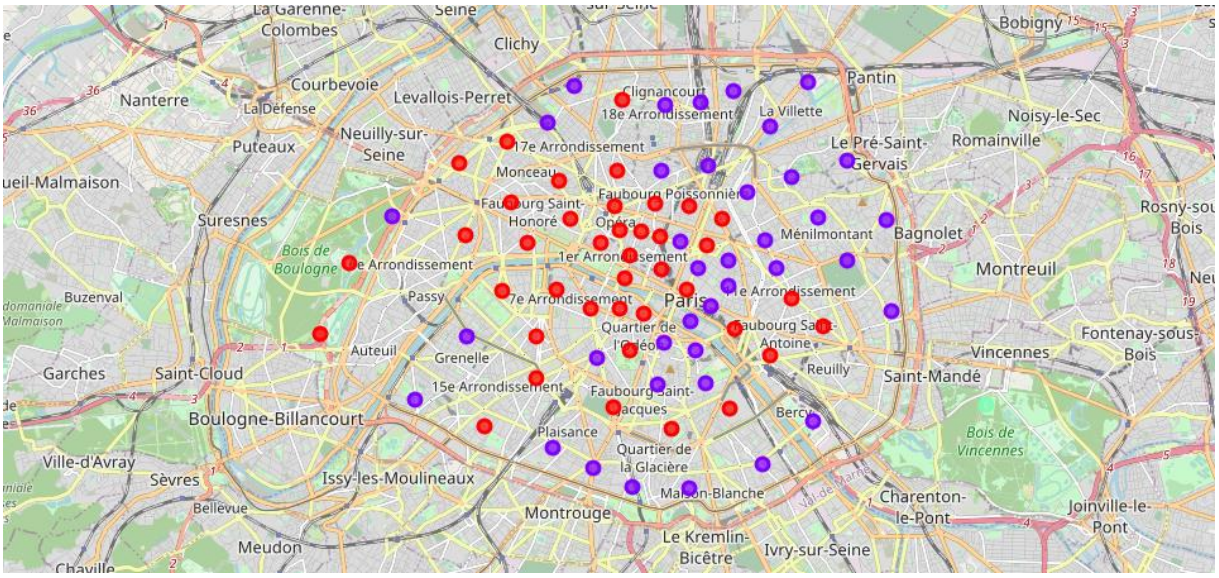


The clusters are named according to their most common restaurant category:

Cluster 1 - French Restaurants:
In this cluster, we can notice that locations usually have French restaurants being the top most common restaurants. People in this cluster prefer more French style foods.
Cluster 2- Bakery & Divers Restaurants:
In this cluster, people prefer Bakery and others food styles(Italian, Korean,..) together with the French food.
Cluster 3- Bel-Air:
It is an isolated cluster, it contain only Bel-Air neighborhood.

Using Folium, a Python library for making leaflet maps, we'll create a map with circle markers that depicts the locations in our data frame. The locations will also be color-coded to indicate what cluster they belong to.

- Cluster 1 being colored red .
- Cluster 2 being colored purple.
- Cluster 3 being colored sky-blue.



5. Discussion

As a recommendation to those who want to find location for new restaurants providing Maghreb foods in Paris, they should consider locations in Cluster 2 neighborhoods where there is less competitions from French restaurants. Analysis also show that people in this cluster have divers eating styles or food preferences.

The analysis can help to gain insight on competitors geographical partitions . But when it come de decide about the right location for such business, additional analyzes including neighborhood traffic generators, demographic and lifestyle data, and competitors are highly recommended.

Conclusion

In this project, using k-means cluster algorithm we have classified the 80 Paris neighborhoods into 3(three) different clusters according to their similar food categories.

This report may be used for future restaurant owners who are planning to open a restaurant within Paris City. This may also help them in deciding where to build such restaurant .

Though this study has given us results based on our data. If you'd like to improve this study, others features, such as, location costs, Maghreb workers data can be added. Another thing that may improve this model is by adding traffic pattern information, demographic and lifestyle data.

6. Reference:

[1]: https://www.agoravox.fr/tribune-libre/article/penurie-d-informaticiens-en-france-212880

[2]: https://www.data.gouv.fr/fr/datasets/quartiers-administratifs/