Wojciech Stach

Clustering Paris and Marseille districts: restaurant location

1. Introduction

• Background

Choosing a location for a business is difficult yet critical decision that may have a huge impact on the business profitability. This is especially important for restaurants as customers often use proximity as a very important factor in making decisions where to eat out. Therefore, for a restaurant to be successful it is crucial to understand the neighbourhood profile to be able to adjust the service and menu to meet the clients' needs and, in turn, to make the restaurant profitable.

This project focuses on two largest cities in France: Paris and Marseille. The city of Paris is divided into twenty administrative districts, *arrondissements municipaux*, referred to later in this document interchangeably as arrondissements or districts. Marseille is divided into sixteen districts. These districts will be described using different features and then clustered to provide useful insights about their profiles and to find similar/dissimilar groups.

• Problem

This project aims at segmenting and clustering the arrondissements of Paris and Marseille to find similar districts in terms of characteristic of venue types in district.

• Audience

The results of this project may be useful to:

- New/future business owners to help them deciding in which district they should open a business or how to choose the restaurant type and menu to best address the local needs
- Existing business owners that are considering either moving to a different arrondissement or opening another location. If their current business is successful, they may want to consider opening a restaurant in a similar district. If not, clustering may provide useful insights on possible modification in the target market to make the business more profitable or identify potentially better locations.

2. Data

• Overview

I searched in the Internet for potentially useful data. I identified a few data sources listed below. I will be performing web scraping or copy/paste to extract the relevant features. In case of Foursquare places database, I will use my developer account to retrieve relevant venues.

Below, there is a list of data sources I found to be used in this project:

a) https://opendata.paris.fr/explore/dataset/arrondissements/table - This data source includes districts' number and name, longitude and latitude, area, and perimeter
b) https://en.wikipedia.org/wiki/Arrondissements_of_Paris - This data source includes each district's population from 1999 and 2005 as well as population density from 2005
c) https://frenchmoments.eu/arrondissements-of-paris/ - This data source includes each district's population from 2013, population density, and median household income

d) https://public.opendatasoft.com/explore/dataset/arrondissements-millesimes0/table/ - This data includes district names, longitude and latitude for a few large cities in France, including Paris and Marseille
e) https://developer.foursquare.com/ - Foursquare Places Database includes community-based venues data which can be queried for each arrondissement using longitude and latitude values

- Selection

I initially planned on using venues data along with population and socio-economic data to describe each district. However, with limited time for completing this project I ran into issues with obtaining the same type of data for Paris and Marseille. Consequently, I ended up using data from sources d) and e).

3. Methodology

- Exploratory data analysis
    - Data pre-processing

I downloaded data from the websites listed above (excluding Foursquare places data, which are described later in this section) . As the first step, I cleaned the data removing irrelevant features and formatting the features in a consistent manner, especially regarding city names (capital/upper case letters), district names and numbers (capital/upper case letters, spelling) to avoid issues with merging multiple tables by city and district names. In addition, I split latitude and longitude data into two separate columns as these two numbers were stored in a single column using comma separated strings. At each step I looked at first few rows of the data sets (using pandas *head* function) to make sure it looks correct. In addition, I checked dimensions of the tables (especially after merging multiple tables) to avoid extraneous data and/or missing rows. After pre-processing was done, I used *describe* method of Pandas Dataframes to check some basic statistical details of the data tables and verify they look correct.

    - Data visualization

In the next step, I used Folium Python library to visualize the districts of Paris and Marseille and make sure they align with their real locations. Please see Figure 1 and Figure 2 below to see the corresponding maps with added markers to show location of each district/arrondissement.

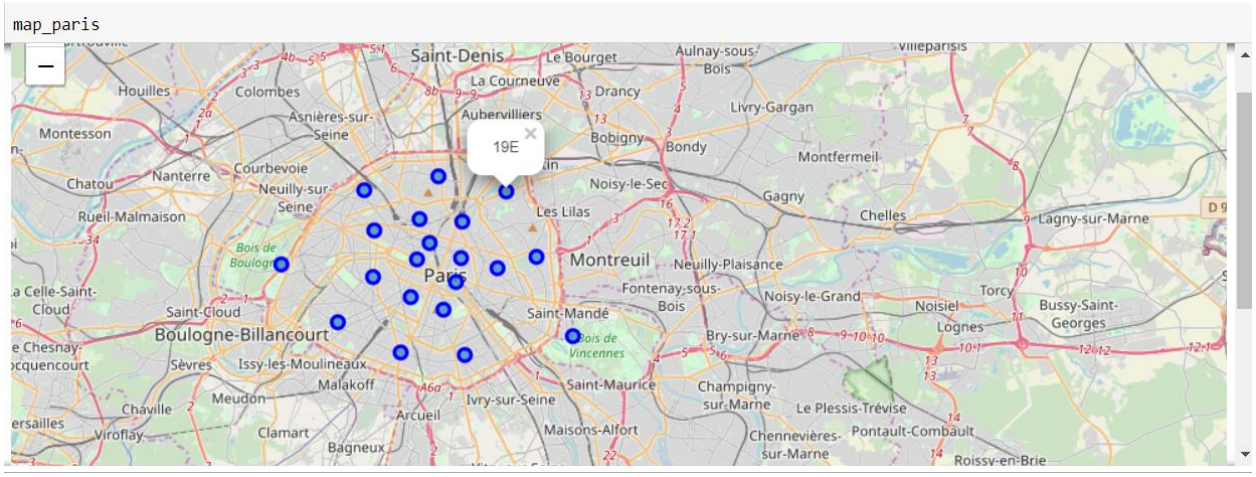

Figure 1 Districts of Paris

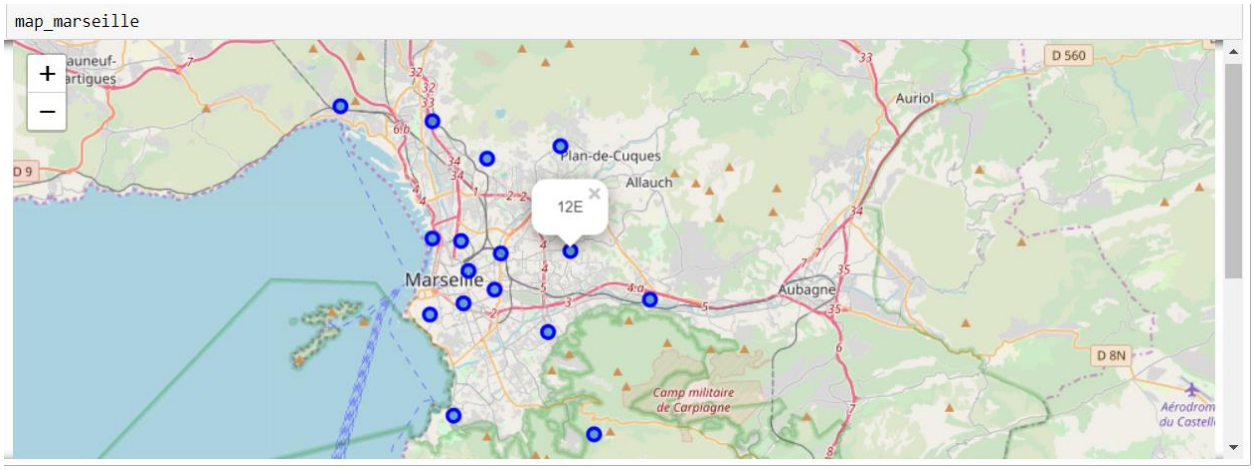

Figure 2 Districts of Marseille

- Foursquare places data

I obtained data about places in each district from using Foursquare Developer account and their API to prepare queries and retrieve results as JSON objects. In the following step, JSON objects were converted to Pandas Dataframes. I also validated the shapes of the dataframes and listed the number of distinct venues categories to make sure they look correct, see Figure 3 below.

Show number of different venues in each city

```
In [59]: print('There are {} uniques categories in Paris.'.format(len(paris_venues['Venue Category'].unique())))
         There are 203 uniques categories in Paris.

In [60]: print('There are {} uniques categories in Marseille.'.format(len(marseille_venues['Venue Category'].unique())))
         There are 68 uniques categories in Marseille.

In [61]: print('There are {} uniques categories in Paris and Marseille.'.format(len(paris_marseille_venues['Venue Category'].unique())))
         There are 223 uniques categories in Paris and Marseille.
```

Figure 3 Distinct venues per city

Next step was to transform the data to a form that is suitable for clustering. To this end one-hot encoding and grouping by means transformations were applied. As a result, each district had a single row in the Dataframe and each column stored a floating point value that correspond to the popularity of given venue type in that district. Lastly, for each district a top 10 common venue types were listed. While this data was not needed for the clustering algorithm, it was later used to describe clusters after clustering process was done.

- Methods

Because of the nature of problem being addressed (find similar/dissimilar districts), clustering methods of unsupervised learning are well suited as they aim to categorize/group similar data items. K-means clustering has been chosen as a method that is relatively simple, well documented, popular and powerful. One of the challenges using K-means method is to select the optimal number of clusters. I ran the elbow method for a range of values for K from 1 to15 and then for each value of K computed an average sum of squared distances for all clusters and plot this value as a function of K. If the line chart

resembles an “elbow” (the point of inflection on the curve), then it is a good indication that the underlying model fits best at that point.

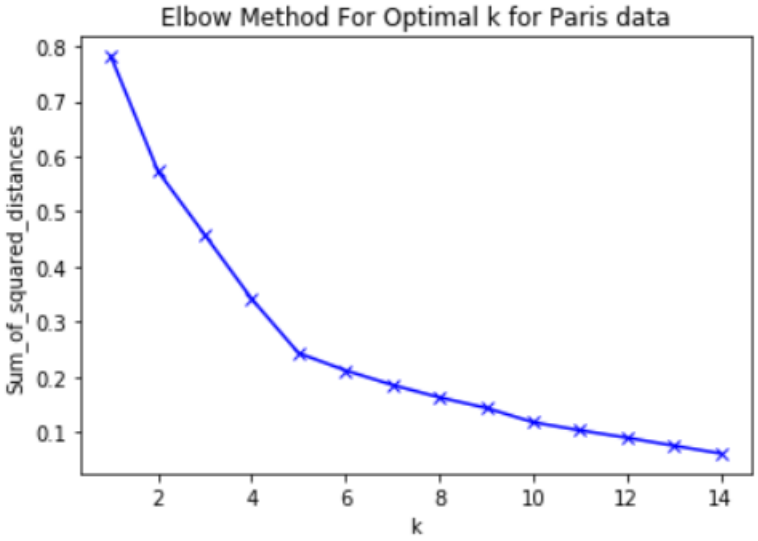

Figure 4 Elbow method for K-Means with Paris data

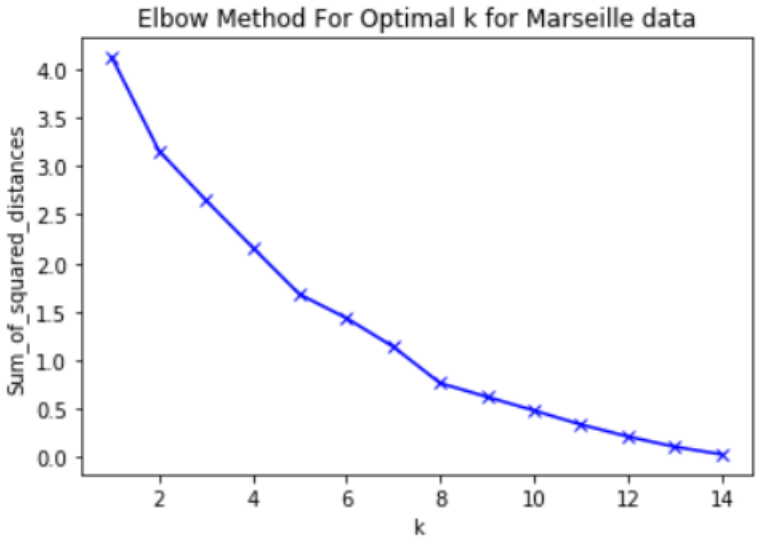

Figure 5 Elbow method for K-Means with Marseille data

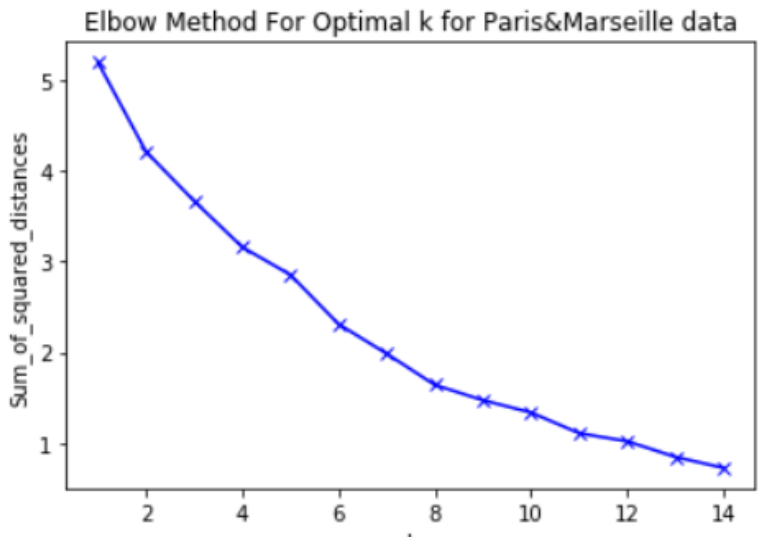

Figure 6 Elbow method for K-Means with Paris and Marseille data

Optimal K for Paris data is relatively easy to visually notice, see Figure 4. The value is equal to 5. On the contrary, for Marseille as well as Paris&Marseille datasets the elbow is not as clearly visible. Looking closer, I chose the values of K to be 5 and 8, respectively.

4. Results

- Paris

Figure 7 shows clustering results for Paris data that was carried out with 5 clusters.

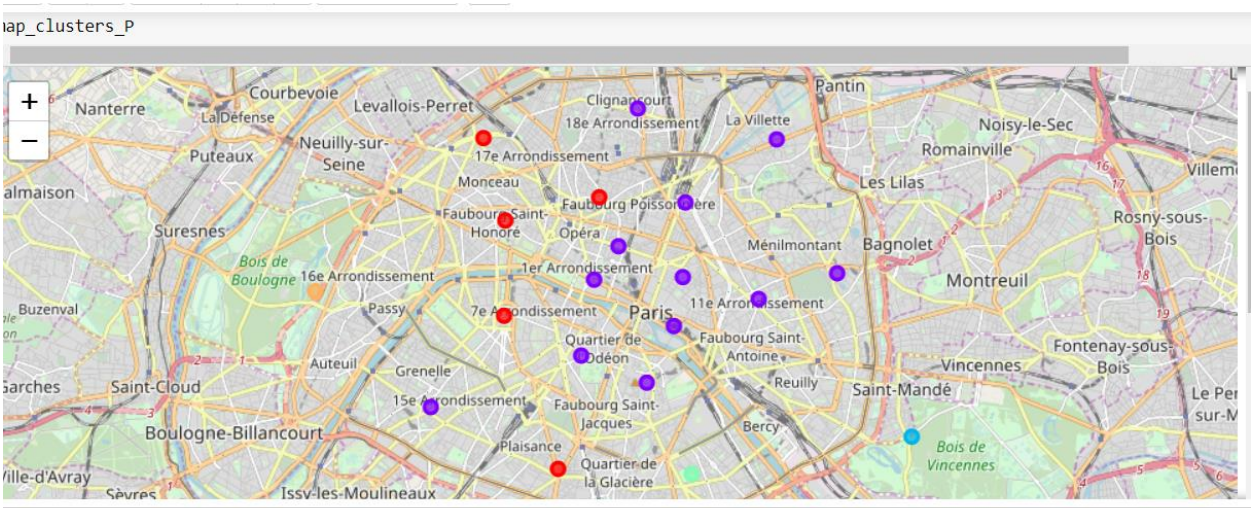

Figure 7 Clustering results for Paris data

Below, are a few observations based on the clustering results and most popular venues in each cluster:

- Cluster 0 - red: Central-West districts of Paris (major venues: French restaurants and hotels)

- Cluster 1 - purple: Central-East districts of Paris and 15th district (that is in the West of Paris). Similar profile to Central-East district except for non-French restaurants showing quite high in terms of most popular venues
- Cluster 2 - blue: residential districts. Main venues include markets, stores, zoo
- Cluster 3 - green: multicultural districts, specifically including lots of Asian restaurants
- Cluster 4 - orange: upscale districts. Main venues include plaza, pool, lake.

- Marseille

Figure 8 shows clustering results for Marseille data that was carried out with 5 clusters.

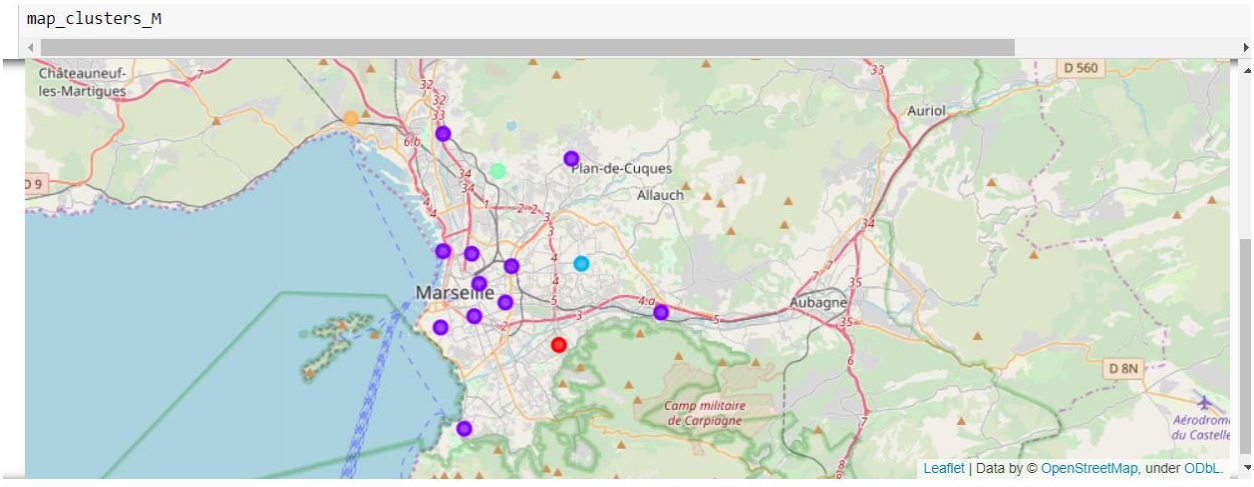

Figure 8 Clustering results for Marseille data

Below, are a few observations based on the clustering results and most popular venues in each cluster:

- Cluster 1 - purple: Center area of Marseille. Most popular venues include hotels and restaurants. Interestingly, districts 8th, 11th , 13th , and 15th that are further from the center of Marseille are in this cluster, too
- Clusters 0, 2, 3, 4 – other colors: Other venues are more popular, including shops tram stations, etc. More knowledge about the city is necessary to provide better distinctions between these clusters

- Paris and Marseille combined

Figure 9 shows clustering results for Paris and Marseille data that was carried out with 9 clusters.

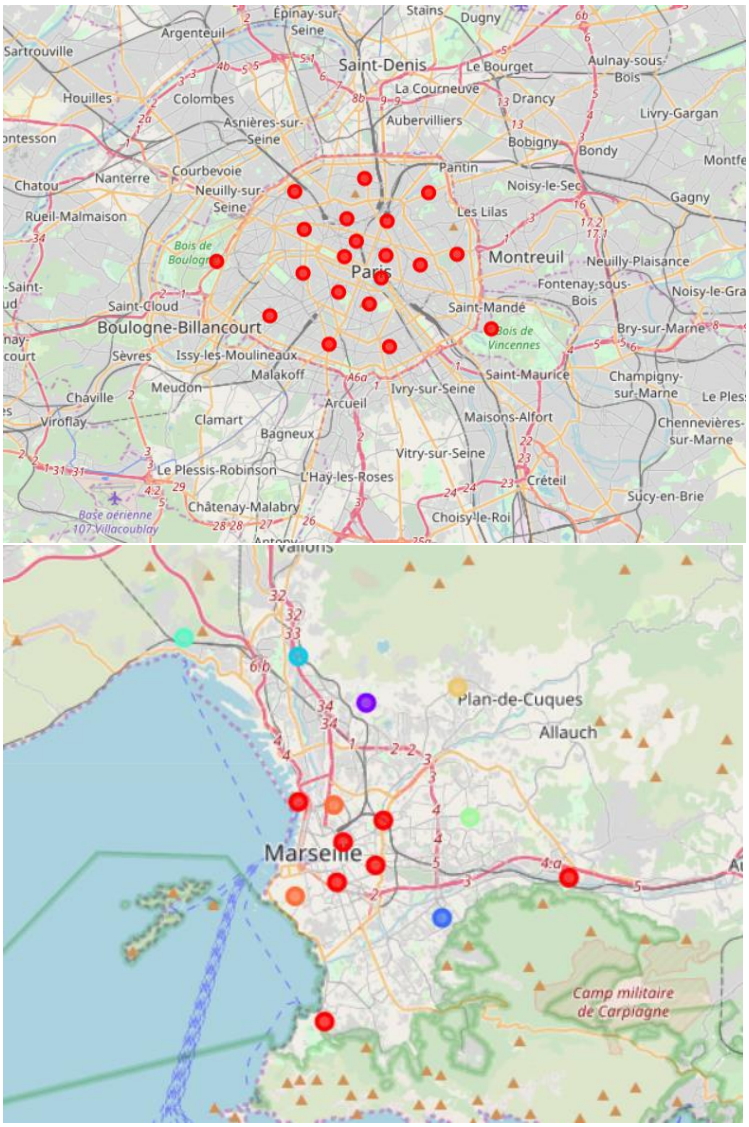

Figure 9 Clustering results for Paris&Marseille data

Below, are a few observations based on the clustering results and most popular venues in each cluster:

- Cluster 0 - red: Includes all districts from Paris and seven districts from Marseille (all of them from cluster 1 in Marseille clustering). This suggests that all Paris districts are quite like each other when compared to Marseille districts
- Clusters 1-7: Outskirts of Marseille. More knowledge about the city is necessary to provide better distinctions between these clusters

5. Discussion

Clustering results for both Paris and Marseille seemed to be intuitively correct for the districts near city centers as they are clustered together. In Paris, the there seems to be vertical line that splits the central East districts and central West districts as they were assigned to different clusters (with very few exceptions). In Marseille, on the other hand, all central districts were assigned to the same cluster.

Interestingly, a few other, geographically far districts, were assigned to the same cluster, which suggests they have similar distribution of the most popular venues. The third clustering done on Paris and Marseille data combined suggests that Paris districts are much more like each other than Marseille districts.

Knowledge of similarities between districts in terms of most popular venue types may be useful for restaurant owners that want to expand their business. For instance, if they run a successful business in Marseille city center (1st district) and they want to open another location in Marseille, they may investigate districts in the same clusters. The clustering results suggest for instance district 15th, which seems to have similar venue types despite being quite far from the city center. A different example may involve a successful owner that has a restaurant Paris and want to open a location in Marseille. Again, looking at red markers in Figure 9 he/she can find similar districts in Marseille.

It would be very beneficial to have knowledge about these two cities in order to provide better description of the clusters. Looking at the most popular venues (the data is shown in the Jupyter notebook project) it is difficult to find clear explanation of differences between clusters 0 and 1 in Figure 7 or to understand why in Marseille some clusters that are far from city center seems to be very similar to clusters from the city center.

6. Conclusion

This project uses K-means clustering on Foursquare places data to find similarities between districts in two biggest cities in France. The clustering results identified similar districts within the cities and between the cities. Consequently, the results may help making decisions about finding a perfect location for a restaurant or other type of business.

As an example of a future work that builds upon this project, it could be very interesting to redo the clustering with added population density as well as socio-economic and demographic features of each district. It would allow for more comprehensive similarity analyses between different districts.