**Analyzing the Optimal Districts**

**to Start a Restaurant Business in Paris**

October 19, 2020

*By: Nabih Saleh*

Table of Contents

[Introduction 3](#_Toc47451679)

[Data 3](#_Toc47451680)

[Methodology 6](#_Toc47451684)

[Results 7](#_Toc47451688)

[Discussion 8](#_Toc47451689)

[Conclusion 8](#_Toc47451690)

**Introduction**

Paris is the capital and the most populous city in France. The city is known to be a very touristic city welcoming millions of travelers per year from all over the world. It is home to many cultural and architectural landmarks. People from all different ethnicities and cultures come to Paris on holidays and it has not stopped being one of the most important traveling hubs in Europe. As for the food, it is widely known that Paris has a number of the most decorated chefs in the world serving all different cuisines from all over the globe. I have personally traveled to Paris and there was no doubt in my head that I had eaten the best food of my life there. I chose the location to see, if I were to open a restaurant business there, which district would be optimal. To do so, I have to study each district in Paris and check how the food business is doing there. The project could be extremely useful for investors who are looking to indulge in the restaurant business in the city of Paris. The objective is to collect appropriate data, process and analyze it, and finally give recommendations to where stakeholders should invest their money (which districts in Paris).

**Data**

To deal with this king of project, we will have to gather data from different sources. We will first need to get the Paris district data. We will then need the respective latitude and longitude coordinates for each district. Finally, we will have to collect the data of the venues for each district in Paris.

* **District Data**

I have collected the Paris district data from the following Wikipedia list <https://en.wikipedia.org/wiki/Arrondissements_of_Paris#Arrondissements> . The data will then be converted into a pandas data frame either by using the read\_html() method or manually inserting the 20 districts. Figure 1 shows the names of the Paris districts.



Figure 1: Paris district data taken from Wikipedia.

* **Districts’ Geographical Coordinates**

We can find each districts’ geographical location by using Python’s GeoPy library. To get the latitude and longitude coordinates of each district, we will use the geocoder library. The code for getting the geographical coordinates of Paris is shown in Figure 2. This data also allows us to plot the map of Paris and marking the different districts using the Folium library. Figure 3 then shows the updated data frame with the latitude and longitude columns added.

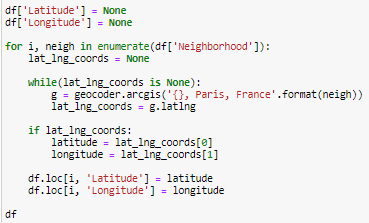


Figure 2: Obtaining geographical coordinates of Paris.



Figure 3: Adding geographical coordinate columns in the district dataframe.

* **Districts’ Venue Data**

We will use the Foursquare API to extract the venue data for each Paris district. We will want to get the venue recommendations for each district and we will study the popular venues further. The venue recommendations were obtained using a maximum of 200 venue recommendations per district and a radius of 2.2 km around the district’s coordinates. Figure 4 shows some of the venue data obtained from the Foursquare API.

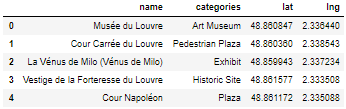


Figure 4: Data obtained from Foursquare API after cleaning.

**Methodology**

I used feature extraction to obtain the appropriate Foursquare API data which was then used for clustering. In order to do this, the “Venue Category” column had to be converted to a numeric value in order to build the model. This was completed using the One-hot Encoding method which creates a column for each unique category. If the venue belonged to the category, it would get a value of 1, and if it does not, it gets a value of 0. This had to be done to all the venues. I then grouped the data frame by the district name and an average value was found for all the categories. Figure 5 shows the result of the top 10 rows. One can notice that most of the values are 0 and that is due to the large number of unique cuisine categories.

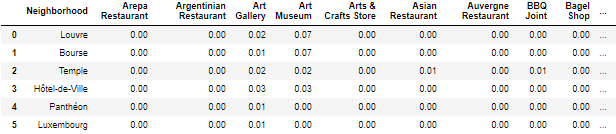


Figure 5: One-hot Encoding data frame.

Dropping the district name column, we use this data for our k-means unsupervised learning model to cluster the neighborhoods based on its category of venues. The first step is to find the optimal number of clusters determined by the Silhouette score of each k value which turns out be to 2.

A data frame was also created which contained the top 10 most common venues of all the districts shown in Figure 6. This table will be used to combine the results from the model.



Figure 6: Top 10 most common venues for each district.

**Results**

The k-means model will find a label for each district in Paris which represents the cluster this district belongs to. The labels column, along with the latitude and longitude columns, is then added to Figure 6’s data frame as shown in Figure 7.



Figure 7: Clustering districts in Paris.

We then extract the districts for each cluster.

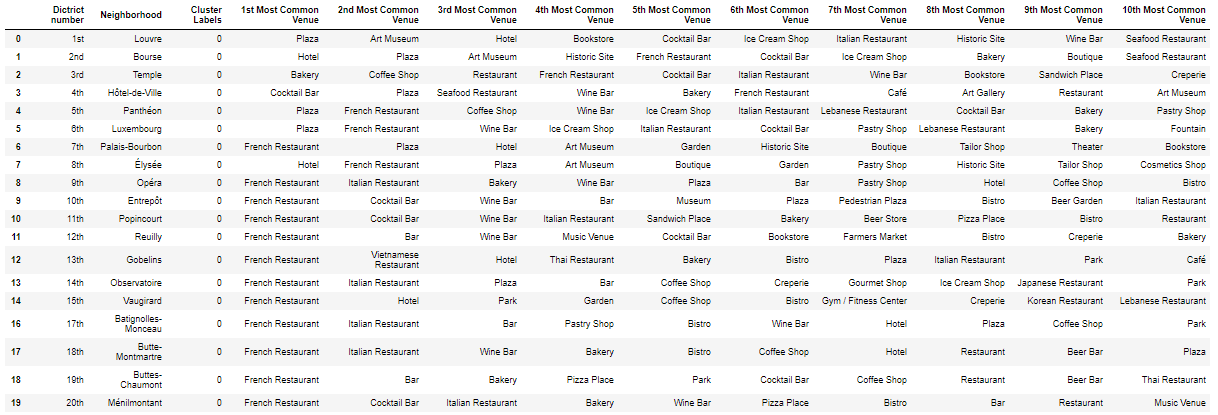


Figure 8: Cluster 1.



Figure 9: Cluster 2.

**Discussion**

Analysing the different clusters, we can see that cluster 1 is more suited for restaurants and hotels than the other. Cluster 2contain a small percentage of restaurants in the top 10 common venues compared to the other districts.Thus, the 16th district (Passy) is not well suited to open up a restaurant. On the other hand, clusters 1contains a large number of restaurants, hotels, cafes, bistros, bakeries and bars which indicates the districts in these clusters are suited for opening a restaurant.



Figure 8: Visualizing the appropriate districts in Paris.

**Conclusion**

We have successfully determined what would be the most appropriate districts to open a new restaurant business. We have found that all the districts belonging to cluster 1would be appropriate locations. However, investors should also consider other factors before choosing the location such as the cost associated with the location.