Applied Data Science Capstone

Final Assignment

Analysing Neighbourhoods of Paris, France For Advising A Contractor To Start A New Restaurant

BRAUD Lucas

25 novembre 2020

Table des matières

1 Introduction

The aim fo this project is to analyse the neighbourhoods of Paris in France in order to advise a contractor where to start a new restaurant. Paris is famous for the variety of the cuisines and its restaurants, and starting a business in this cooking field is a very difficult task. Indeed, Paris is full of hundreds of restaurants in many districts and the food commodities are enormous. Hence why it is very important to study the market and suggest the contractor a good location in one of the many neighbourhoods in Paris which would lead to the sucess or failure of his business. The goal here is to determine, extract and analyse the right data about neighbourhoods in Paris using various data science tools in order to suggest to the investor, the best location to start his new business.

2 Use of data

To provide right recommendations for the contractor, we used the following data :

- **Districts of Paris (Wikipedia Source)** : https ://en.wikipedia.org/wiki/Category :Districts_of_Paris. This data comes from Wikipedia and was extracted using Beautiful Soup's Python library. The result gave us a dataframe containing all neighbourhoods of Paris.
- **Geographical Coordinates of the neighbourhoods.** The coordinates are useful for plotting maps during the visualization phase and were extracted using the Python library GeoPy. At the end, the dataframe created before was completed with latitude and longitude of each neighbourhood.
- **Venues data from Foursquare**. This data was provided by Foursquare to give us information about venues in the Paris neighbourhoods such as restaurants and markets. It was the determining source of data because at the end it gave us features of restaurants in order to better understand the competition.

The goal was to scrape the Wikipedia page and wrangle the data, clean it, and then read it into a pandas dataframe so that it was in a structured format like the New York dataset that we saw in the lesson. Once the data was in a structured format, we replicated the analysis that we did to the New York City dataset to explore and cluster the neighbourhoods for the new restaurant in Paris.

3 Methodology section

3.1 Analysing each neighbourhood

For this part, we replicated the same method that we saw in the lab : the one hot encoding. This technique converts each category that belongs to a venue into a binary number which reprensents whether or not a category is found in the venue. Moreover, the venues are then grouped by neighbourhoods and the mean of the frequency of occurrence of each category is taken. Finally, we could print each neighborhood along with the top 5 most common venues.

3.2 K-Means Clustering

For this part, we used an unsupervised machine learning algorithm to cluster neighbourhoods in order to find out their similarities. This algorithm helped us looking for clusters within the data in order to minimize the data dispersion for each cluster. Next, we could better observe the pattern inside the multidimensional features of the set of data. To implement this algorithm, we first needed to know the number of clusters (input of the algorithm) so we used the elbow method saw in the lab to select the right k. At the end, the correct number of clusters found was 3 as shown in the graph below.

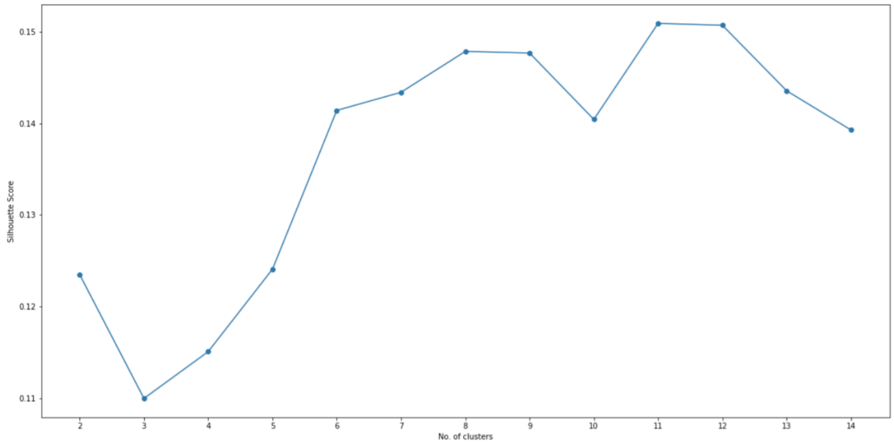

FIGURE 1 – Elbow methode to find optimal value for k

4 Results

4.1 Visualization using Folium library

To visualize the data and see the results, we essentially used the Folium diversity because it was very much suited for our project. Indeed, it alllowed us to see and understand the positions of the clusters by plotting maps and markers on them to categorize the neighbourhoods.

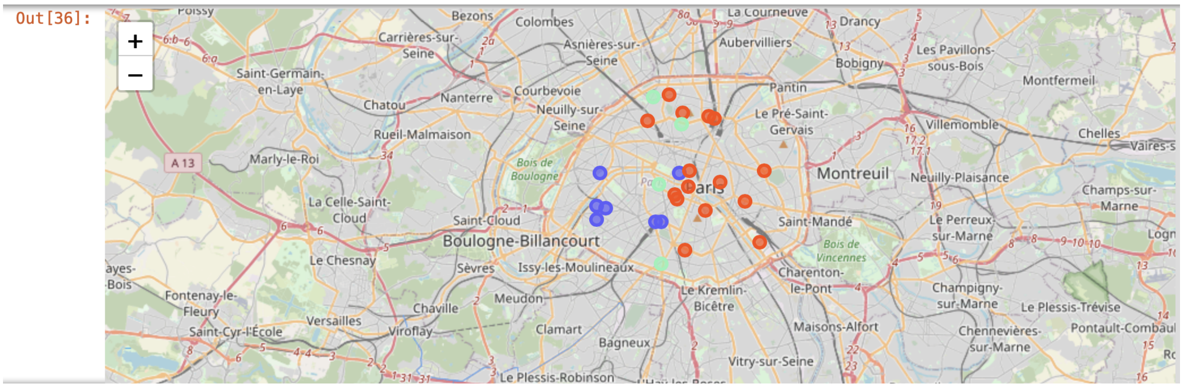
Out[36]:

FIGURE 2 – Clusters in Paris using the Folium Library

4.2 Clusters examination

Now, we could examine each cluster and determine the discriminating venue categories that distinguish each cluster. An important result occur to us when plotting the final clusters dataframes : the neighbourhood that had the most number of restaurants was cluster with $k = 1$ (see code).

4.3 Discussion

The most suitable neighbourhood is the cluster 1 for starting a new restaurant. Cluster 1 represent indeed the heart of Paris, with the historic center. We can conclude that our K-Means clustering technique worked because in reality we can easily know that the historic district is the one with the most restaurants in it. Nevertheless, we looked at the price and availibility for places in this cluster and it is very hard to find something good. This means that opening a restaurant is a hard task here because it relies on the money you can invest in a new property.

5 Conclusion

To conclude, data analysis with python and applied data science tools such as capstone were very useful for this business problem. Libraries such as Beautiful Soup, pandas and Folium were extremely useful to respectively extract, transform and visualize the data in order to achieve our initial goal : give a recommendation to a contractor for starting a new restaurant in Paris.