# IBM Data Science Professional Certificate

Picking the right location for a new fast food restaurant in Paris

## Capstone Project – Final Report

Cédric Verone

April 19, 2020

# 1. Introduction

### a. Business Problem

For this project I have created a hypothetical business problem.

I would like to start a fast food restaurant, and therefore I am looking to identify which areas in Paris would be the best to do so. Hence, I could use datascience techniques to analyse paris boroughs environements to determine which boroughs are the most suitable. The ideal location will depend on the target market (families, young adults…).

According to Statistita, there are around 180,000 restaurants in France including 20% located in Paris. The French restaurant industry sales increased by 1.6% to reach EUR50.4 billion in 2018. The fast food industry appears to be the most dynamic segment. It represents 37% of restaurants in France and generates around EUR19 billion of sales. The fast-food average ticket increased by 1.4%. Based on the figures displayed by the Paris Chamber of Commerce, 23% Parisian restaurants are fast-food restaurant only topped by traditional French restaurants. It is the fastest growing segment with a growth of 11% (in terms of number of restaurants) between 2014 and 2017.

### b. Target Audience

This report could be useful for business entrepreneur looking to open a new fast food restaurant.

# 2. Data

In order to perform this analysis, I will use the following datasets :

- The list of Paris metro stations
- The annual trafic for each metro stations
- The Foursquare API

The metro stations' datasets are available on the RATP (Paris Public Transport Operating Company) website. https://dataratp.opendatasoft.com/explore/

We will use the Foursquare API to get the venues around metro stations / neighbordhoods. Based on the venue type distributions, I will use a clustering algorithm (Kmeans) to group metro stations / neighborhoods. It would then be possible to determine which cluster would be suitable to start the business.

# 3. Methodology

### a. Overview

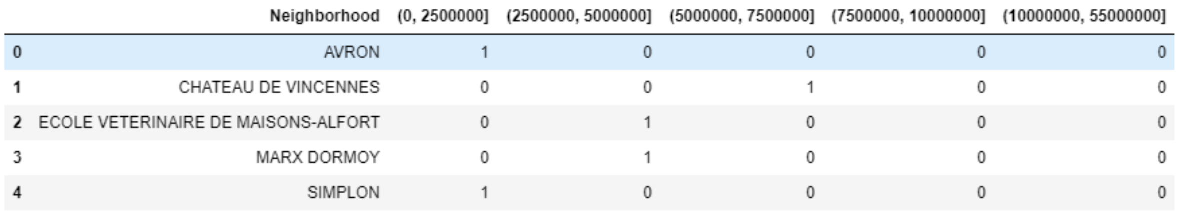
We will start with some data cleaning and exploration. We will then apply an unsupervised learning algorithm (kmeans) in order to create clusters of metro stations. We will use the silhouette score to determine the optimal number of clusters.

Paris metro is the most efficient way to travel in the city. Hence areas nearby metro stations could be very insteresting places to open a restaurant.

There are around 300 metro stations in Paris. In this analysis, we will focus on a sample of 179 stations.

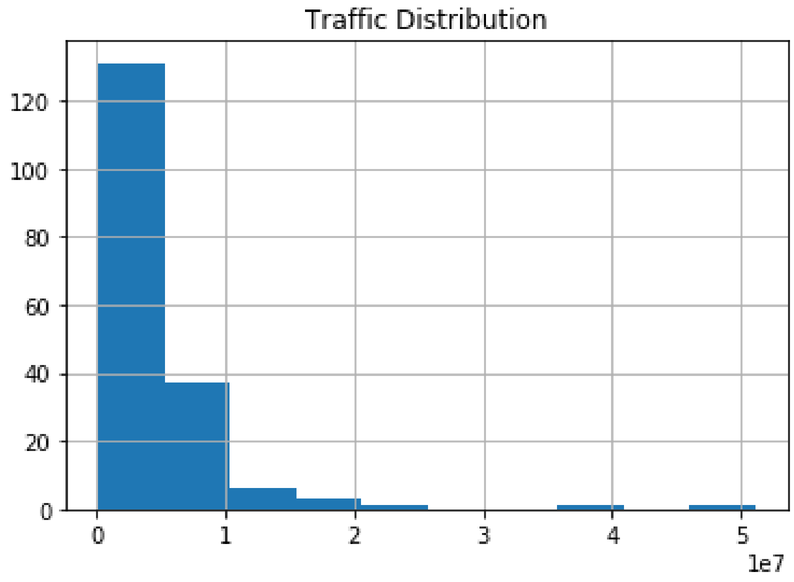
### b. Data Preparation

We download two datasets, in csv format, respectively consisting of public transport stations' traffic and coordinates. First we need to exclude the stations belonging to other public transport network (bus, RER etc…) from the public transport stations' traffic dataset. We also segmented and sorted stations into 5 bins based on their annual traffics , using the pandas' cut function. The output of this operation is the dataframe below:

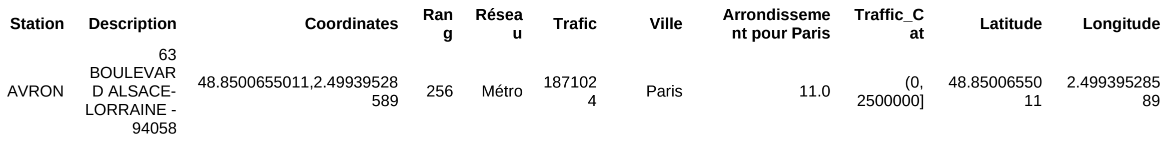


| | Neighborhood | (0, 2500000] | (2500000, 5000000] | (5000000, 7500000] | (7500000, 10000000] | (10000000, 55000000] |
|---|---|---|---|---|---|---|
| 0 | AVRON | 1 | 0 | 0 | 0 | 0 |
| 1 | CHATEAU DE VINCENNES | 0 | 0 | 1 | 0 | 0 |
| 2 | ECOLE VETERINAIRE DE MAISONS-ALFORT | 0 | 1 | 0 | 0 | 0 |
| 3 | MARX DORMOY | 0 | 1 | 0 | 0 | 0 |
| 4 | SIMPLON | 1 | 0 | 0 | 0 | 0 |

We create an histogram of the traffic distribution where we can see that the majority of the stations have an annual traffic below 5 million passengers :



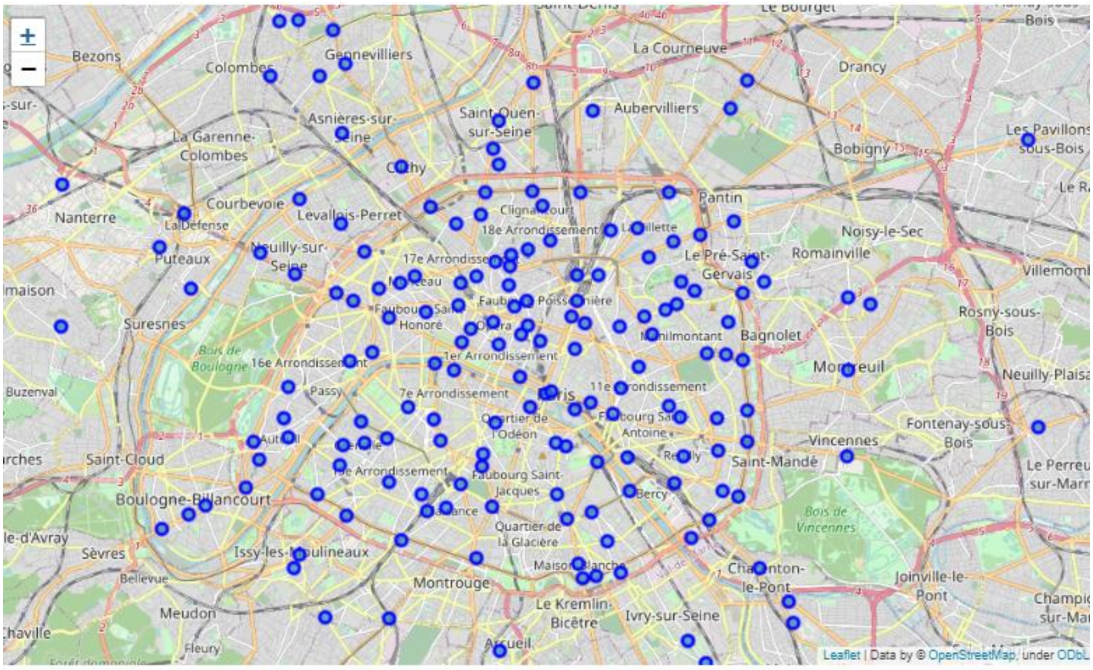
We also have to format the coordinates. After performing these tasks, we merge the datasets into a dataframe. You can see below a screenshot of the top five rows of the data frame.



| Station | Description | Coordinates | Rang | Réseau | Trafic | Ville | Arrondissement pour Paris | Traffic_Cat | Latitude | Longitude |
|---|---|---|---|---|---|---|---|---|---|---|
| AVRON | 63 BOULEVARD ALSACE-LORRAINE - 94058 | 48.8500655011,2.49939528589 | 256 | Métro | 1871024 | Paris | 11.0 | (0, 2500000] | 48.8500655011 | 2.49939528589 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| CHATEAU DE VINCENNES | PISTE GARE ROUTIERE - 75112 | 48.8442170813,2.44079723454 | 50 | Métro | 6353285 | Vincennes | NaN | (5000000, 7500000] | 48.8442170813 | 2.44079723454 |
| ECOLE VETERINAIRE DE MAISONS-ALFORT | 31-35 AVENUE DU GENERAL LECLERC - 94046 | 48.8147969334,2.42270643461 | 157 | Métro | 3193857 | Maisons-Alfort | NaN | (2500000, 5000000] | 48.8147969334 | 2.42270643461 |
| MARX DORMOY | AVENUE FRANCOIS MITTERRAND - 91027 | 48.7036035904,2.37143564263 | 151 | Métro | 3359946 | Paris | 18.0 | (2500000, 5000000] | 48.7036035904 | 2.37143564263 |
| SIMPLON | 46 BOULEVARD ORNANO - 75118 | 48.8948064764,2.34715016514 | 218 | Métro | 2366858 | Paris | 18.0 | (0, 2500000] | 48.8948064764 | 2.34715016514 |

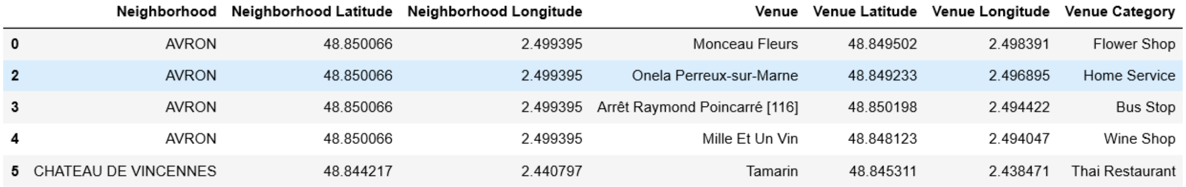
We also create a visualisation of the map of Paris with location markers on metro stations.



In the next step of the data analysis, we explore the venues around metro stations using the Foursquare API and obtain a dataframe consisting of the metro stations, the metro stations' coordinates, the venue names, the venues' coordinates and the venue categories. The search was focused on the top 100 venues within 500 meters of each station.

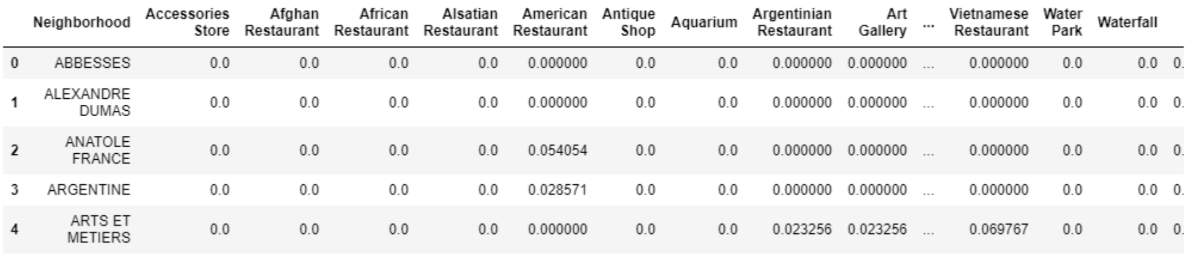
We decide to exclude the French restaurant and the hotels because they are very common venues in Paris.

You can see below a screenshot of the top 5 rows of the dataframe.



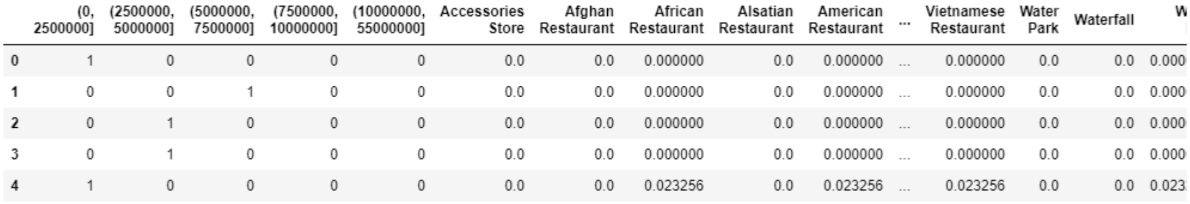
| | Neighborhood | Neighborhood Latitude | Neighborhood Longitude | Venue | Venue Latitude | Venue Longitude | Venue Category |
|---|---|---|---|---|---|---|---|
| 0 | AVRON | 48.850066 | 2.499395 | Monceau Fleurs | 48.849502 | 2.498391 | Flower Shop |
| 2 | AVRON | 48.850066 | 2.499395 | Onela Perreux-sur-Marne | 48.849233 | 2.496895 | Home Service |
| 3 | AVRON | 48.850066 | 2.499395 | Arrêt Raymond Poincarré [116] | 48.850198 | 2.494422 | Bus Stop |
| 4 | AVRON | 48.850066 | 2.499395 | Mille Et Un Vin | 48.848123 | 2.494047 | Wine Shop |
| 5 | CHATEAU DE VINCENNES | 48.844217 | 2.440797 | Tamarin | 48.845311 | 2.438471 | Thai Restaurant |

The, we perform a one hot encoding and created a dataframe with the frequency of each venue for every stations.



| | Neighborhood | Accessories Store | Afghan Restaurant | African Restaurant | Alsatian Restaurant | American Restaurant | Antique Shop | Aquarium | Argentinian Restaurant | Art Gallery | ... | Vietnamese Restaurant | Water Park | Waterfall | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | ABBESSES | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0. |
| 1 | ALEXANDRE DUMAS | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0. |
| 2 | ANATOLE FRANCE | 0.0 | 0.0 | 0.0 | 0.0 | 0.054054 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0. |
| 3 | ARGENTINE | 0.0 | 0.0 | 0.0 | 0.0 | 0.028571 | 0.0 | 0.0 | 0.000000 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0. |
| 4 | ARTS ET METIERS | 0.0 | 0.0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.0 | 0.023256 | 0.023256 | ... | 0.069767 | 0.0 | 0.0 | 0. |

We then merged this dataframe with the dataframe containing the traffic categories created earlier.



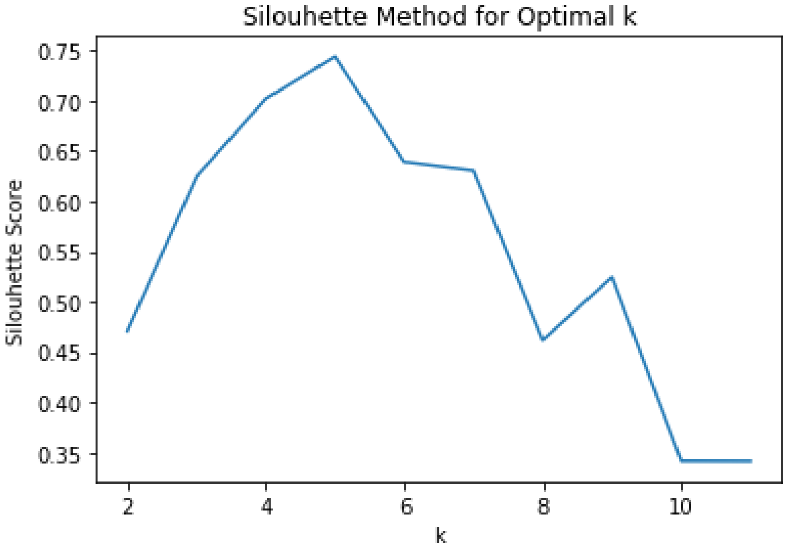
| | (0, 2500000] | (2500000, 5000000] | (5000000, 7500000] | (7500000, 10000000] | (10000000, 55000000] | Accessories Store | Afghan Restaurant | African Restaurant | Alsatian Restaurant | American Restaurant | ... | Vietnamese Restaurant | Water Park | Waterfall | W |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | 0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0.000 |
| 1 | 0 | 0 | 1 | 0 | 0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0.000 |
| 2 | 0 | 1 | 0 | 0 | 0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0.000 |
| 3 | 0 | 1 | 0 | 0 | 0 | 0.0 | 0.0 | 0.000000 | 0.0 | 0.000000 | ... | 0.000000 | 0.0 | 0.0 | 0.000 |
| 4 | 1 | 0 | 0 | 0 | 0 | 0.0 | 0.0 | 0.023256 | 0.0 | 0.023256 | ... | 0.023256 | 0.0 | 0.0 | 0.023 |

This dataframe will be the input of our clustering model.

### c. Clustering

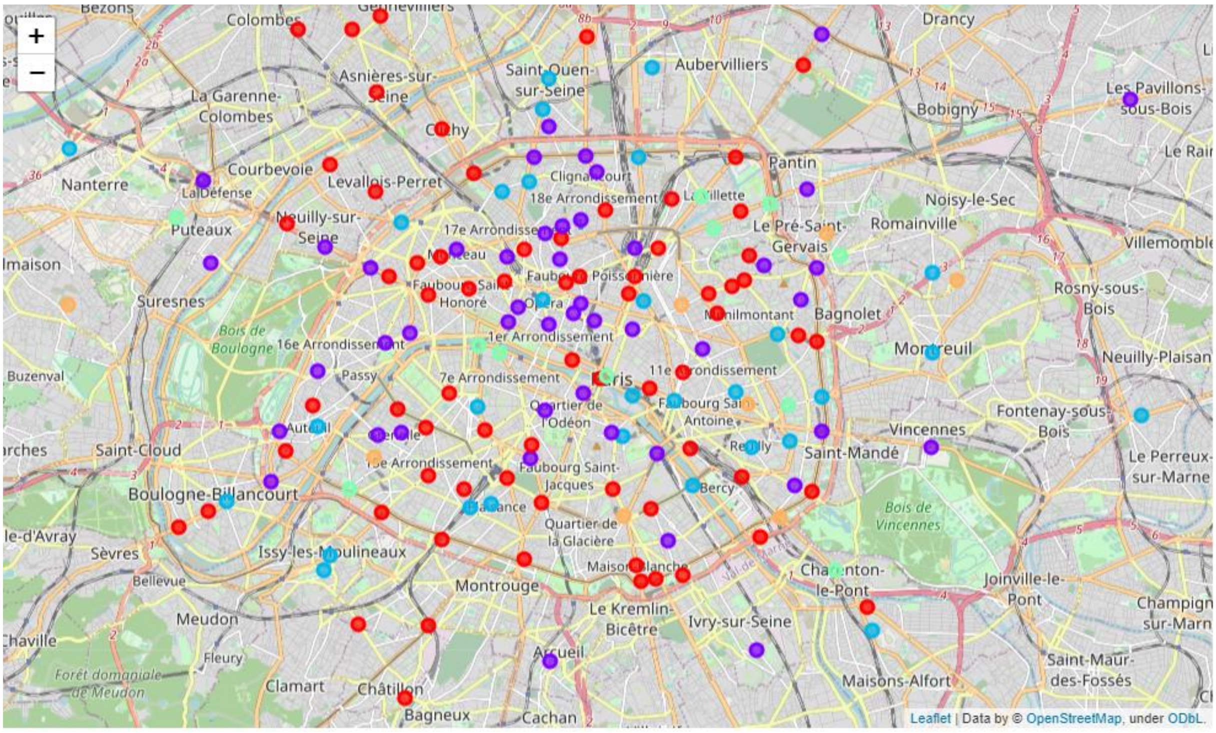
Now that we have the dataset that will be used to train our model, we needed to choose a clustering algorithm. We decided to use the K-means algorithm which is an unsupervised learning. It is easy to implement and it is quite computationally fast.

The first step is to decide how many clusters we want build. For this purpose we use the silouhette analysis which can be used to study the separation distance between the resulting clusters. The graph below displays the silouhette scores for each number of clusters.



In this graph, we can see that the optimal cluster number is five since it has the highest silouhette score.

We can see the clusters distribution on the following map :



Markers color: Cluster 0: red, CLuster 1: purple, Cluster 2: blue, Cluster 3: green, Cluster 4: orange.

## 4. Results

We look at the results we can see the following cluster distribution :

- Number of metro stations in cluster 0: 77
- Number of metro stations in cluster 1: 51
- Number of metro stations in cluster 2: 31
- Number of metro stations in cluster 3: 12
- Number of metro stations in cluster 4: 8

We can also display the distribution by traffic ranges for each cluster.

| | **(0, 2500000]** | **(2500000, 5000000]** | **(5000000, 7500000]** | **(7500000, 10000000]** | **(10000000, 55000000]** |
|---|---|---|---|---|---|
| **0.0** | 26% | 48% | 19% | 0% | 6% |
| **1.0** | 31% | 33% | 16% | 14% | 6% |
| **2.0** | 29% | 55% | 6% | 3% | 6% |
| **3.0** | 25% | 33% | 33% | 0% | 8% |
| **4.0** | 38% | 25% | 25% | 0% | 13% |

Cluster 0, cluster 1 are the biggest clusters in terms of metro stations counts.

They are surrounded by a lot of restaurants, the most fequent venue being italian restaurants. We can also see a competition from Japanese restaurants which appeared in the top four venues as well as from pizza places, vietnamese restaurants and sandwich places. In terms of traffic cluster 1 consist of bigger stations with 20% of them having more than 7.5 million tavellers. La Defense which is the main business district and the third biggest metro station falls within this cluster.

Cluster 2 is quite similar, but consist of stations with less traffic.

Cluster 3 consists of 12 stations. Three stations could be interesting spots because they are located on the Seine banks, in central Paris. Three stations are located in the canal de l'Ourcq neighborhood which is in the process of gentrification with a young population.

Cluster 4 is quite interesting with most of its stations located in the eastern part of Paris. Place d'Italie could be a good location since it is big a transport hub.

## 5. Discussion

Based on this analysis, we recommend to consider cluster 3 or 4 to start a new fast food restaurants. These clusters gather metro stations located in areas where the competition is less intense than in other clusters.

This analysis can be improved with additional data such as demography, population youth, offices locations etc.

## 6. Conclusion

In this report, we conducted an analysis aiming for determining the best postential locations for a new fast food restaurant. In order to do so, we performed data cleaning/preparation using common python libraries such as pandas, matplotlib and numpy. In order to implement the k-means algorithm, we used the scikit learn package. The result of the analysis can serve as a basis for business decision making.

## 7. References

You will find below the link to the Jupyter notebook related to this analysis.

https://github.com/Kadrik87/IBM-Capstone
Project/blob/master/Fast%20Food%20Restaurant.ipynb