Data Section

To solve the problem, we will need the following data:

- List of neighborhoods in Paris. This defines the scope of this project, which is confined to the city of Paris.
- Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and to get the venue data.
- Venue data, particularly data related to Hotels. We will use this data to perform clustering on the neighborhoods.

Sources of data and methods of extraction

This Wikipedia page (https://en.wikipedia.org/wiki/Arrondissements_of_Paris) contains a list of neighborhoods in Paris, with a total of 20 neighborhoods. We will use web scraping techniques to extract the data from the Wikipedia page, with the help of Python requests and Pandas packages. Then we will get the geographical coordinates of the neighborhoods using Python Geocoder package which will give us the latitude and longitude coordinates of the neighborhoods.

After that, we will use Foursquare API to get the venue data for those neighborhoods. Foursquare has one of the largest databases of 105+ million places and is used by over 150,000 developers. Foursquare APIs provide many categories of the venue data, we are particularly interested in the Hotels category to help us to solve the business problem put forward. This is a project that will make use of many data science skills, from web scraping (Wikipedia), working with API (Foursquare), data cleaning, data wrangling, to machine learning (K-means clustering) and map visualization (Folium). In the next section, we will present the Methodology section where we will discuss the steps taken in this project, the data analysis that we did and the machine learning technique that was used.

Methodology

Firstly, we need to get the list of neighborhoods in the city of Paris. Fortunately, the list is available in the Wikipedia page (https://en.wikipedia.org/wiki/Arrondissements_of_Paris). We will do web scraping using Python requests and Pandas packages to extract the list of neighborhoods data. However, this is just a list of names. We need to get the geographical coordinates in the form of latitude and longitude to be able to use Foursquare API. To do so, we will use the wonderful Geocoder package that will allow us to convert address into geographical coordinates in the form of latitude and longitude. After gathering the data, we will populate the data into a pandas Data Frame and then visualize the neighborhoods in a map using Folium package. This allows us to perform a sanity check to make sure that the geographical coordinates data returned by Geocoder are correctly plotted in the city of Paris.

Next, we will use Foursquare API to get the top 100 venues that are within a radius of 2000 meters. We need to register a Foursquare Developer Account to obtain the Foursquare ID and Foursquare secret key. We then make API calls to Foursquare passing in the geographical coordinates of the neighborhoods in a Python loop. Foursquare will return the venue data in JSON format and we will extract the venue name, venue category, venue latitude and longitude. With the data, we can check how many venues were returned for each neighborhood and examine how many unique categories can be curated from all the returned venues. Then, we will analyze each neighborhood by grouping the rows by neighborhood and taking the mean of the frequency of occurrence of each venue category. By doing so, we are also preparing the data for use in clustering. Since we are analyzing the Hotels data, we will filter the “Hotel” as venue category for the neighborhoods.

Lastly, we will perform clustering on the data by using k-means clustering. K-means clustering algorithm identifies k number of centroids, and then allocates every data point to the nearest cluster, while keeping the centroids as small as possible. It is one of the simplest and popular unsupervised machine learning algorithms and is particularly suited to solve the problem for this project. We will cluster the neighborhoods into 3 clusters based on their frequency of occurrence for “Hotel”. The results will allow us to identify which neighborhoods have higher concentration of Hotels while which neighborhoods have fewer number of Hotels. Based on the occurrence of Hotels in different neighborhoods, it will help us to answer the question as to which neighborhoods are most suitable to open new Hotel.