IBM Data Science Professional Certificate

Capstone Project – The Battle of Neighborhoods

Report

Executive Summary

The Capstone Project for the IBM Data Science Professional Certificate is to create a program to leveraging the FourSquare location data to explore or compare neighborhoods or cities of my choice. The program created attempts to solve the problem faced by travelers or migrants when they decide on a neighborhood to stay while traveling to another city. The program compares the neighborhoods of Singapore (Planning Areas) and Paris (Quartiers). The analysis uses K Means to cluster the neighborhoods in each city, to find which neighborhoods have the same amenities. The analysis has found 3 Parisian neighborhoods, in 2 of the clusters, have similar amenities to neighborhoods in Singapore. These neighborhoods are proposed as places that migrants should consider staying. For travelers, 2 other clusters are proposed to address the issue this program is designed to address.

Contents

1. Introduction

France, and her capital Paris, is one of the premier travel destinations in the world. This can also be said of an international city like Singapore. With Singapore and Paris being two cities that are major centers of finance, commerce and arts in their respective regions, a comparison of the venues in each neighborhood at both cities can provide unique insights. As a resident of Singapore and having a desire to travel to Europe and see the ‘City of Lights’, a comparison of both cities would give any traveler or migrant a nice overview of the neighborhoods in the city.

Recently, travelers desire a more authentic view of countries, as can be seen by the rise of AirBnB. This program aims to provide travelers with data to make a choice on which neighborhood they can consider staying in. Therefore, the target audience would be travelers and migrants moving from Singapore to Paris. For example, a traveler travelling to Paris from Singapore might want to know if a neighborhood is able to provide a more authentic travel experience with immersion to French culture yet provide similar amenities as their neighborhoods back in Singapore. Another example would be to advise an immigrant from Singapore to Paris if a neighborhood provides similar amenities like another neighborhood in Singapore.

This program would provide a quick macro view of the neighborhoods in Singapore and Paris and cluster the neighborhoods by the venues or amenities provided. Specifically, the program leverages on the Foursquare API and K Means Clustering to help travelers or immigrants decide on neighborhoods to stay in Paris that are like neighborhoods in Singapore by comparing venues in the neighborhoods. The clusters generated using the neighborhood data for both Singapore and Paris will help provide guidance to the stakeholders of this project for their final decision.

2. Data Description

a. Data Sources

The data used in this program was obtained from the World Bank website as a CSV file, from the Wikipedia website using 'BeautifulSoup' and location data is obtained using the 'FourSquare' API.

b. Data Cleaning (Europe Travel Data)

To build the business case for selecting France, and by extension its capital Paris, we look at the data provided by the World Bank. The data is contained in 2 files that are downloaded from the website. This data is provided in CSV format and shows the international arrivals to France from 1995 to 2018, and the country information. The first file consists of the country names and its regional and income information. The second file consists of the travel data from 1960 to 2019.

The data contained in the files consists of country, regional and continental data. As the regional and continental data is not required, these data are removed from the files. To make the dataset more manageable, we choose only countries that are in the 'Europe & Central Asia' region, as well as those that are in the High-Income group. The files are then merged. Null values are replaced with a '0' value in this dataset. This is to prevent the data being skewed unnecessarily.

We note that data relating to international arrivals from 1960 to 1995 and data pertaining to 2019 are not included into the dataset. Therefore, these years with no data are not considered for our analysis and are dropped from the dataset. There are 38 countries in the dataframe we created. To make the dataset more manageable and to make data visualization clearer, we will take the top 10 countries with the highest total international arrivals.

Before addressing the data obtained from the World Bank, the file has a total of 217 entries and 64 rows, before processing to make the dataset more manageable. After merging and processing the datasets, the data frame contains a total of 38 entries (rows) and 27 columns.

c. Data Cleaning (Neighborhood Data)

The neighborhood data is obtained from Wikipedia pages detailing the planning areas of Singapore and the breakdown of the Paris 'Arrondissement' system. The 'BeautifulSoup' package is used to parse the required data from the separate webpages into Python. The 2 Wikipedia pages contain a table detailing the respective neighborhoods. For the Wikipedia page containing the Singapore neighborhood data, the table consists of 9 columns. We pull the data from the 'Name', 'Region', 'Area' and 'Population' columns. As the data obtained under the 'Name' column is for the 'Planning Areas' in Singapore, we will be using it as the list of Neighborhoods. For the Wikipedia page containing the Paris neighborhood data, the table consists of 5 columns.

The parsed data is then pre-processed to a Data Frame of 4 columns. The 4 columns are 'Neighborhood', 'Region', 'Area' and the 'Population'. The 'Neighborhood' column details the name of the neighborhood or planning area reviewed. The 'Region' details the planning region used by the city. The 'Area' column details the size of the Neighborhood reviewed and are measured in kilometers squared, km2. The 'Population' column details the population living in the area per the last census conducted.

The Data obtained for Singapore contains areas where there are no residential areas, these are seen by the null value in the population column. These values are dropped from the dataset as they provide no meaningful information. Additionally, as Singapore is a dense urban environment, areas that are below 2000 in population are removed from the data as these areas are either areas with large water catchment zones or industrial areas.

The Data obtained for the Paris neighborhoods do not contain any NaN data. The 'Arrondissement' data is equated to the 'Region' data for Singapore. While the Quarter or 'Quartiers' data is equated to the

‘Neighborhood’ data. The population data is processed to make it an integer and the Area data is processed into float data type.

Based on the problem definition, we choose 2 specific criteria that would influence our decision. The first is the similarity of venues at each neighborhood. The second the presence of neighborhoods from the base city of Singapore in each cluster.

3. Methodology

For this project, the methodology would be to direct our efforts in determining 2 aspects of the problem presented. The first would be to determine if selecting France is appropriate and if there is a business case to solve this problem. The second aspect would be to determine what is the number of venues at each neighborhood location and the similarities of each neighborhood.

To build the business case, we attempt to visualize the international arrivals data presented. We then use a simple regression analysis to see the trend of the international arrivals to France.

Following this, the venue type and location data relating to each neighborhood in both Singapore and Paris are collected into a dataframe. The search is performed using a limit of 500m radius around the center of each neighborhood. Simple data visualisations of the neighborhoods are produced. This would be used to put some context into the problem and our solution.

The second step consists of processing the venue data regarding each neighborhood to how common it is. The analysis of the data is performed at this step. One hot encoding is used to obtain the frequency of the venues at each neighborhood.

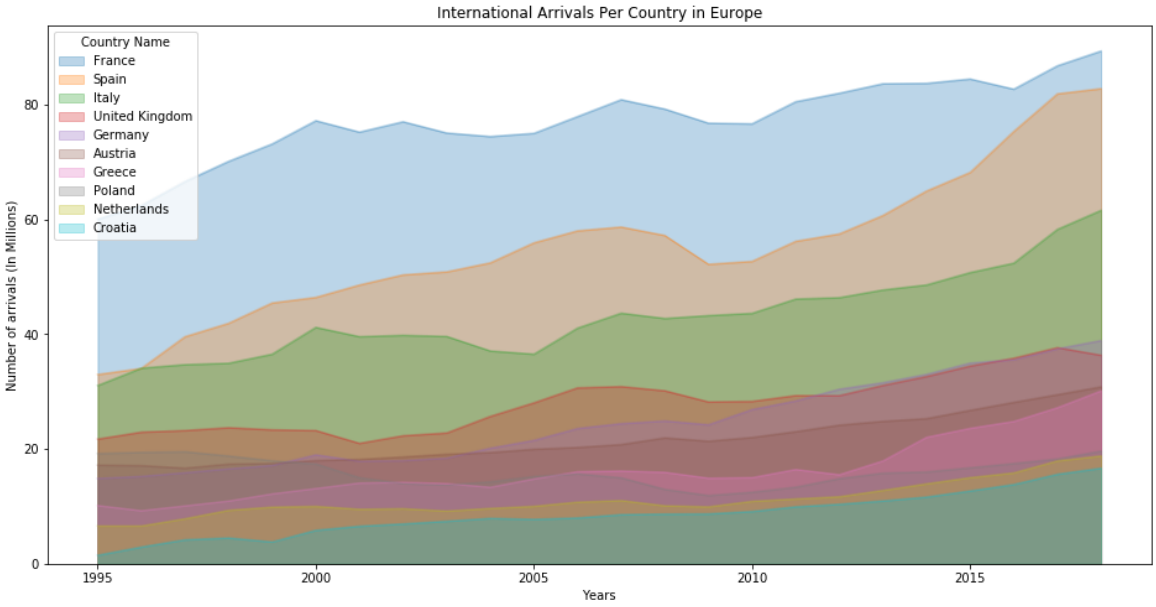
The final step is the use of K Means clustering to cluster the neighborhoods in Singapore and Pairs. This clustering accounts for the venue locations at each neighborhood and would give a general overview of the neighborhood and provide a starting point for the target audience to solve their issue. The elbow method would be used to determine the optimal K cluster value used.

4. Exploratory Data Analysis

a. International Arrivals (World Bank Data)

We begin the exploratory data analysis on the international arrivals data obtained from the world bank. The European regional data is visualized with an Area Plot. To make our case, we analyze the data for the country in the European region with the highest international arrivals. The capital of the country with the most international arrivals is selected as the baseline city due to its position as the country's capital and as an international transport hub.

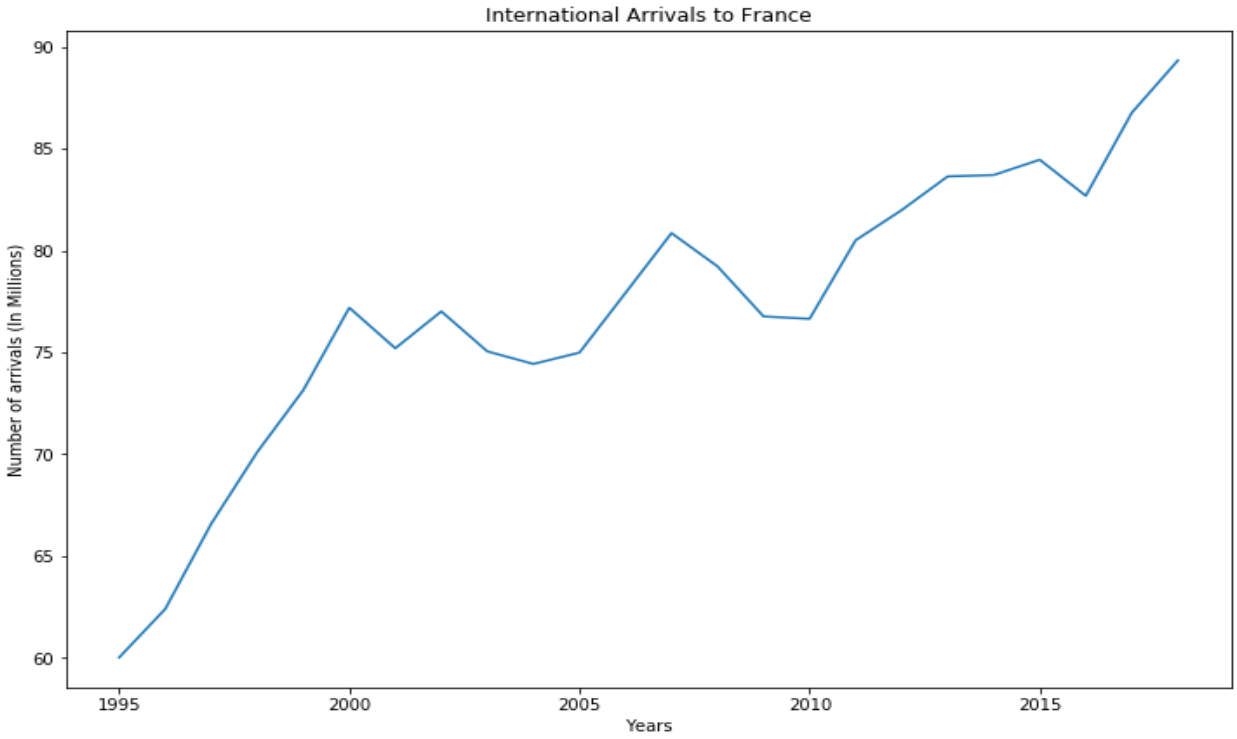
The data is sorted to show the country with the largest international arrivals since 1995. This data is plotted on an area graph to see growth trends in international travelers for the region. (See Fig. 1).



(Figure 1: International Arrivals Per Country in Europe (Top 10))

We can see that international arrivals to France is the largest amount the countries selected from Europe. In fact, France has the highest international arrivals since 1995, with Spain only catching up in around 2015. However, the chart is not clear when it comes to whether there is an increasing trend of new international arrivals. The growth in International Arrivals to France seems modest when compared to other countries such as Spain and Italy.

To address this, we create 2 charts. The first would be a line chart to visualize the yearly international arrivals to France. The second would be a scatter plot, with the regression analysis performed on the data on France provided. This second chart will be included into the analysis section of the report. We now select the data specifically from France to see if there is a positive trend. (See Fig. 2)

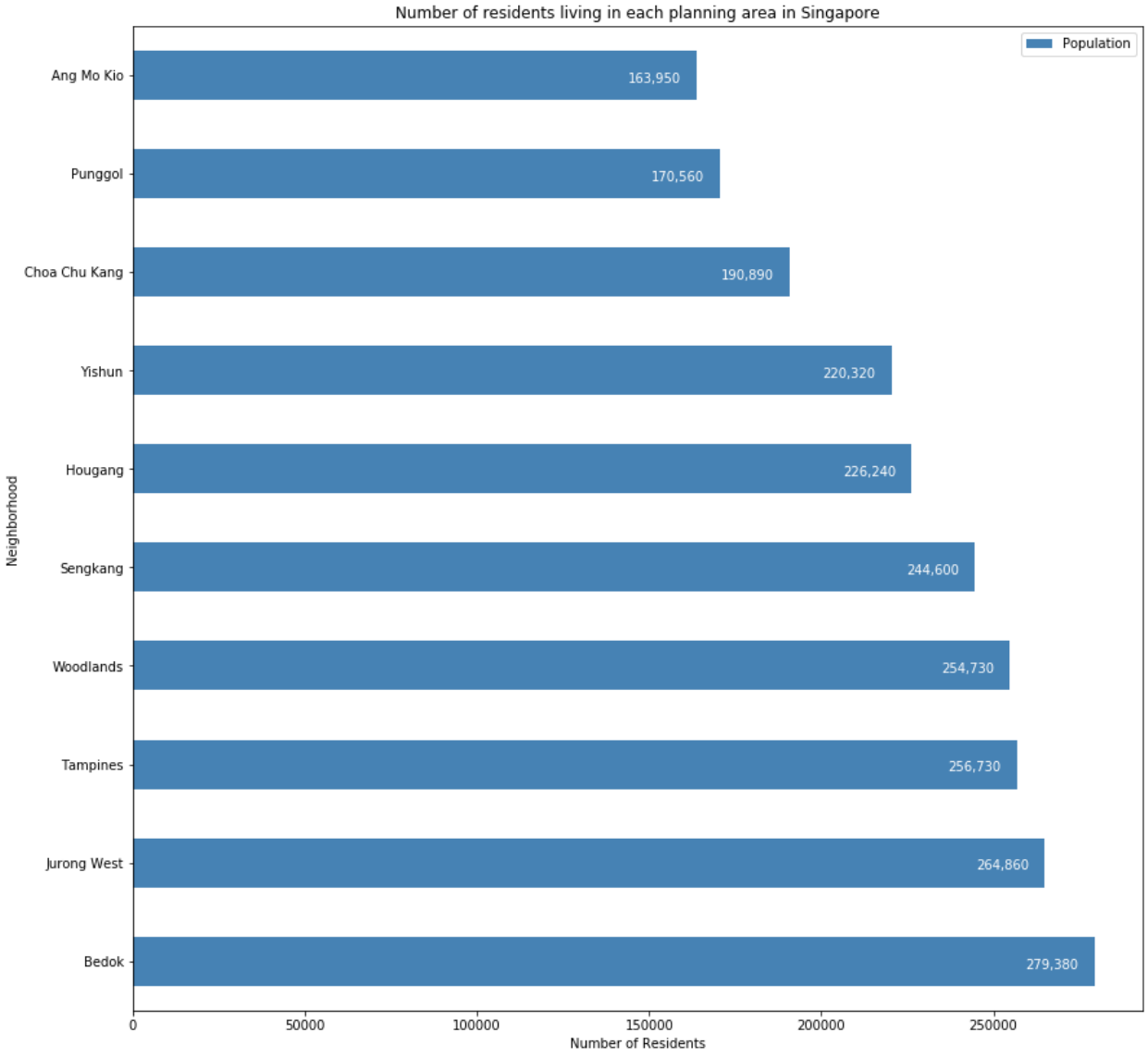


(Figure 2: International Arrivals to France)

We can see that the range of international arrivals entering France from the year 2000 to 2010 is relatively stable at 75 to 80 million. From 2011 onwards, there is a significant increase in international arrivals to France. This further substantiates our choice of France as our destination.

b. Population and Population Density Data (Singapore)

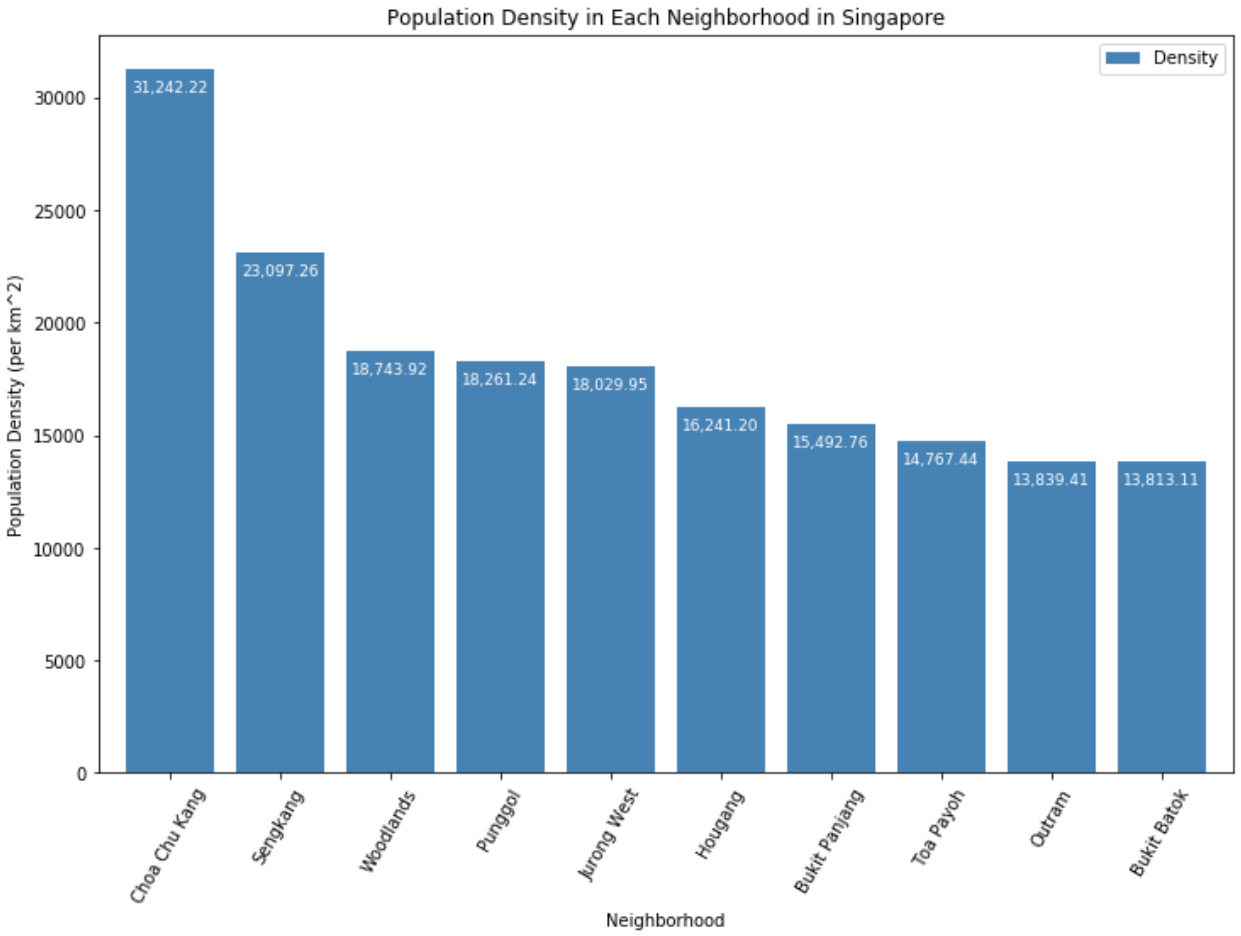
Next we explore the data scrapped from the Wikipedia page on Singapore's 'Planning Areas' and their associated data. We plot the top 10 most populous neighborhoods in Singapore. A Horizontal Bar Chart is chosen to bring across clearly the population statistic for the chosen neighborhoods. (See Fig. 3)



(Figure 3: Number of Residents Living in Each Planning Area in Singapore (Top 10))

From the figure above, we see that Bedok has the highest neighborhood population in Singapore.

We plot the population density of the top 10 neighborhoods in Singapore. For this chart, it was determined that the use of a Vertical Bar Chart is more appropriate. One reason for this choice is the significant changes in population density for the first 2 neighborhoods when compared to the remaining neighborhoods. The use of the Vertical Bar Chart would bring across the information clearly. (See Fig. 4)

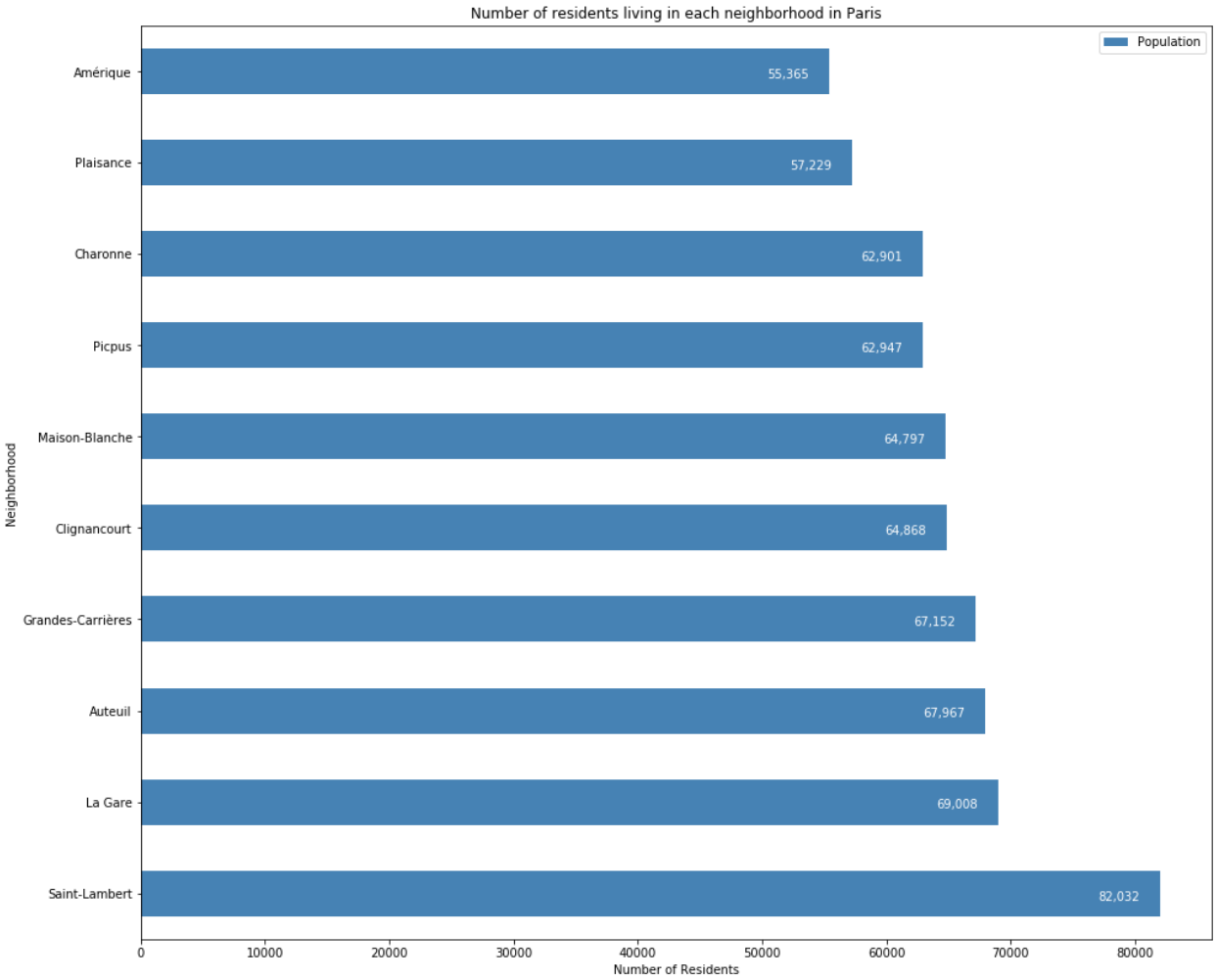


(Figure 4: Population Density in Each Neighborhoods in Singapore (Top 10))

We can observe that the most populous neighborhood, Bedok, is not the densest. However, the neighborhood 'Chua Chu Kang' is the most population dense neighborhood. The chart observation hints that there may be a need to compare the venues generated for both the most populous and most population dense neighborhoods to determine if there are structural differences on the venues or amenities provided.

c. Population and Population Density Data (Paris)

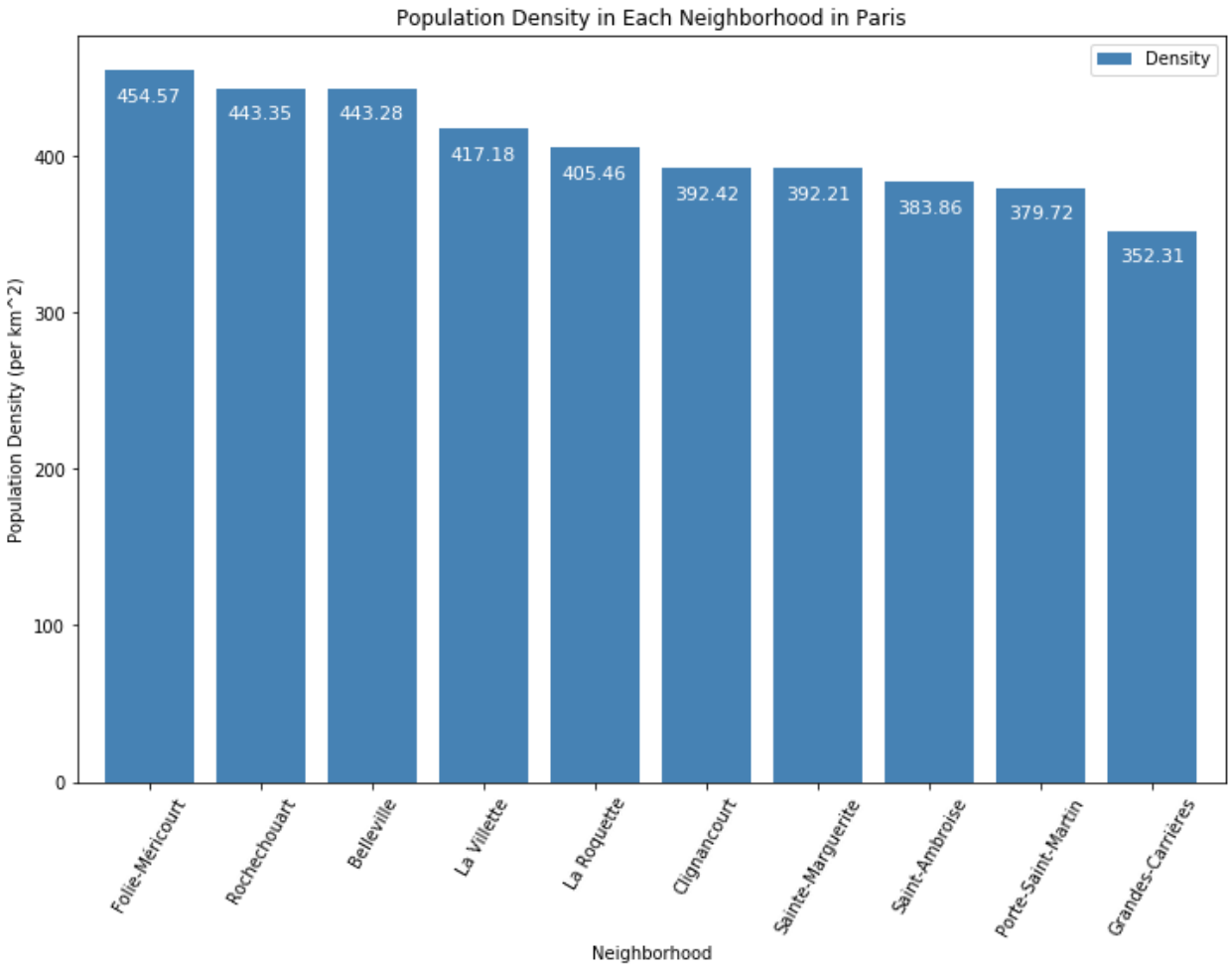
Next we explore the data scrapped from the Wikipedia page on Paris's 'Arrondissement', 'Quartiers' and their associated data. We plot the top 10 most populous neighborhoods in Paris. (See Fig. 5)



(Figure 5: Number of Residents Living in Each Neighborhood in Paris (Top 10))

From the figure above, we see that Sainte-Lambert has the highest neighborhood population in Paris.

We plot the population density of the top 10 neighborhoods in Paris.



(Figure 5: Population Density in Each Neighborhood in Paris (Top 10))

An observation of note is that the neighborhoods in Paris is significantly lesser in terms of population and population density. It may be possible to guess that the frequency of venues in Singapore may be higher than that in Paris.

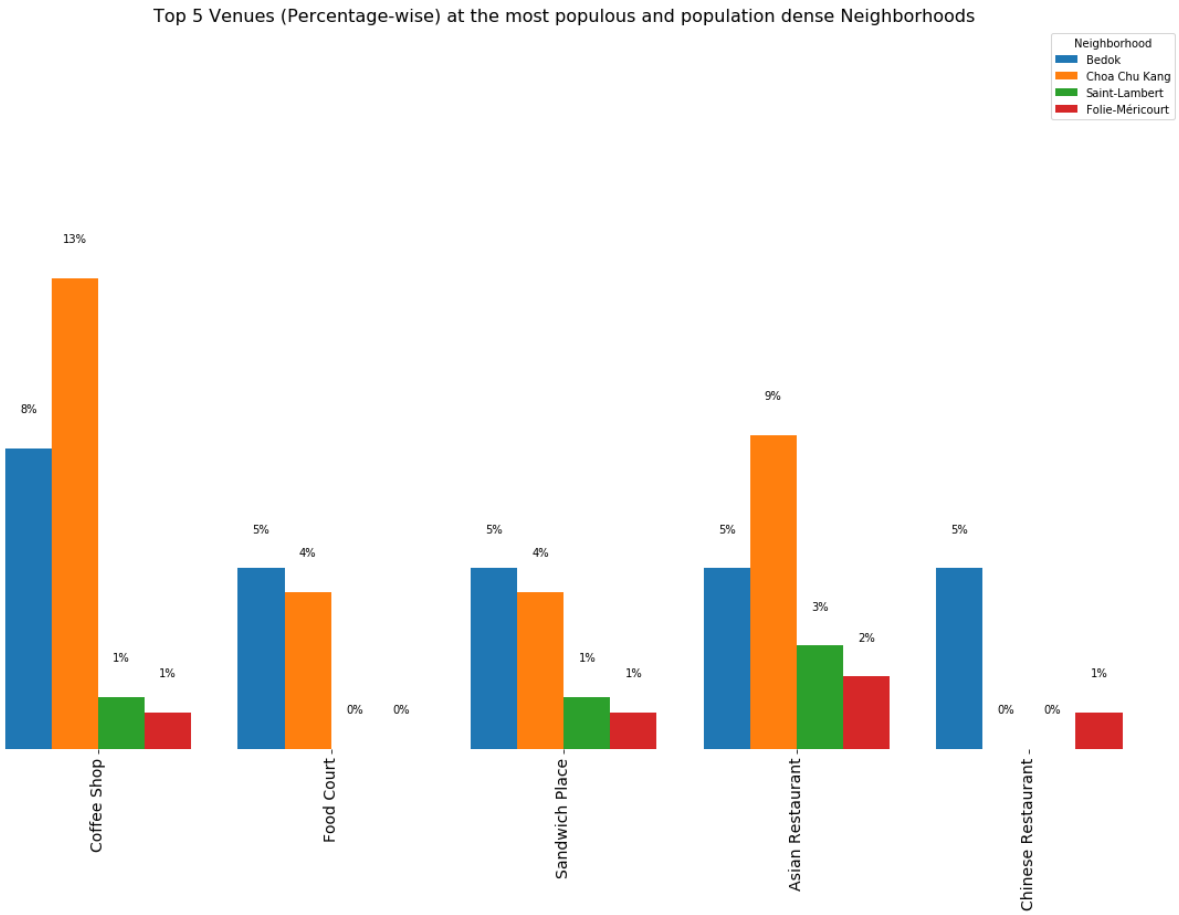
We can see that Folie-Mericourt is the most population dense neighborhood in Paris. Like the Exploratory charts for Singapore, we can note that the most populous neighborhood, Saint-Lambert, is not in the top 10 most population dense neighborhood. With this discrepancy, it is reasoned that a comparison of the venues and amenities in the top neighborhoods of Singapore and Paris, must account for both the top populous and the top population dense neighborhoods.

d. Venues in Singapore and Paris

We explore the venue data generated by the FourSquare API and attempt to gain some initial insights from the data. We note that the FourSquare APi generated a total of 352 unique venue categories for all the neighborhoods listed in our dataset. The data set has a total of 6969 entries, with a maximum of 100 venues pulled per neighborhood.

The venue data is shortened to account for the highest populous and highest population dense neighborhoods in Singapore and Paris. This would allow us to see the differences in the neighborhoods about the most common venues and amenities. Such a comparison would show if there are significant structural differences in venues between high population neighborhoods and high population density neighborhoods.

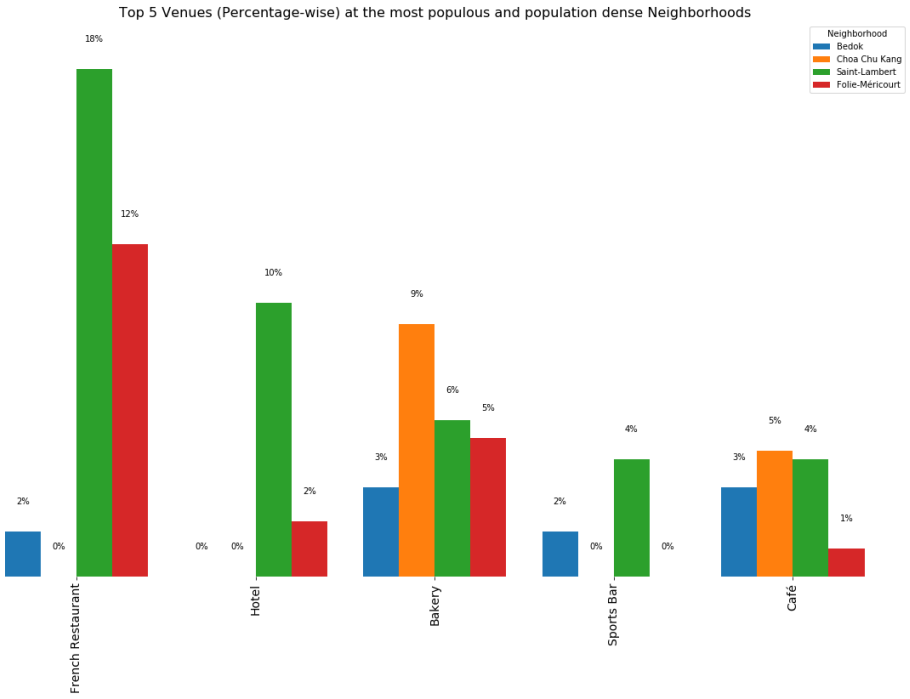
One hot encoding was performed on the venue dataset to process the data, and ready it for visualization and K Means analysis. This initial data would give us some insights on the venue data generated by the API. (See Fig. 6)



(Figure 6: Top 5 Venues (Percentage-wise) at the most populous and population dense Neighborhoods (Sorted by Singapore))

We can immediately see that the Singaporean neighborhoods analyzed are roughly similar in terms of amenities, regardless of population density. This is seen by the high similarity of venues. Additionally, it can be surmised that the amenities favored by Singaporeans are starkly different from the French as seen by the low frequency of venues like Coffee Shops and Food Courts.

While the above chart plots the top 5 venues as sorted using the most populous neighborhood in Singapore, to contrast this, we sort using the most populous neighborhood in Paris. (See Fig. 7)



(Figure 7: Top 5 Venues (Percentage-wise) at the most populous and population dense Neighborhoods (Sorted by Paris))

We can observe that the top 5 venues for the Parisian neighborhood are all different from the Singaporean neighborhood. Additionally, the frequency of the top venue, French Restaurant, is significantly higher when compared to that of Singapore. As France is largely a monoculture when compared to highly diverse Singapore, the high frequency of the 'French Restaurant' venue is not surprising. Additionally, travelers to Paris would be keener on French cuisine rather than their own native cuisine. This is against the initial assumption that there would be higher venue frequency of neighborhoods in Singapore.

Nonetheless, the notable appearance of a venue like 'Asian Restaurant' when the data was sorted by the Singaporean neighborhoods, is interesting. We can deduce that there may be a significant minority of Asian populations in the most populous and population dense neighborhoods of Paris.

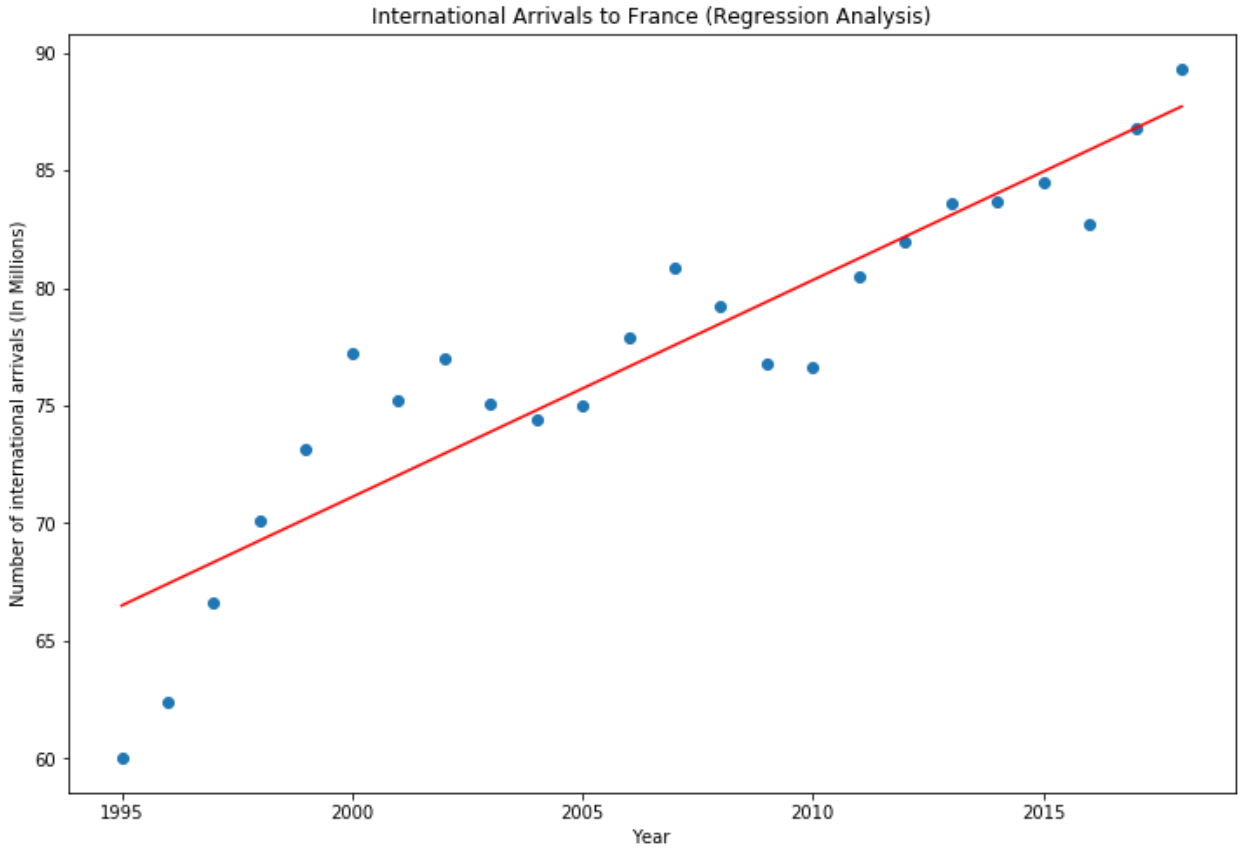
5. Analysis

Two types of analysis were performed on the datasets that are used in this program. The first would be a linear model on the International Arrivals dataset for France. This was conducted to address the issue of a target market for this analysis. The second is the use of the K Means Clustering Method for the venue data generated using the FourSquare API.

a. International Arrivals analysis

For this dataset, we start by plotting a scatterplot for the international arrival data obtained from world bank. The linear regression model is then applied to the data for France. We are clearly able to see a trend of increasing International Arrivals for France.

While the line chart used in the Exploratory Data Analysis section does show an increase in international arrivals in France, regression analysis can be applied to the dataset to see the trend. We now plot the France international arrival data in a scatter plot and perform regression analysis on the data. (See Fig. 8)

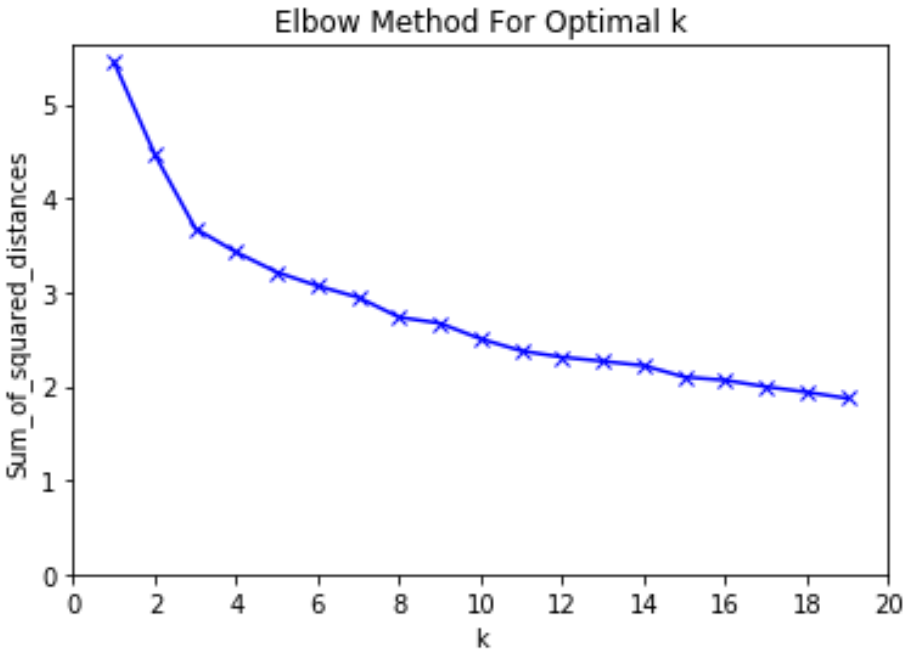


(Figure 8: International Arrivals to France (Regression Analysis))

b. Venue Data Analysis

We start by obtaining the top venue data generated by the FourSquare API. Following this we start sorting and processing the data frame generated. One hot encoding was performed on the dataset. Following this, the data is normalized and readied for the analysis. K Means Cluster is proposed to cluster both the Singapore and Paris neighborhoods.

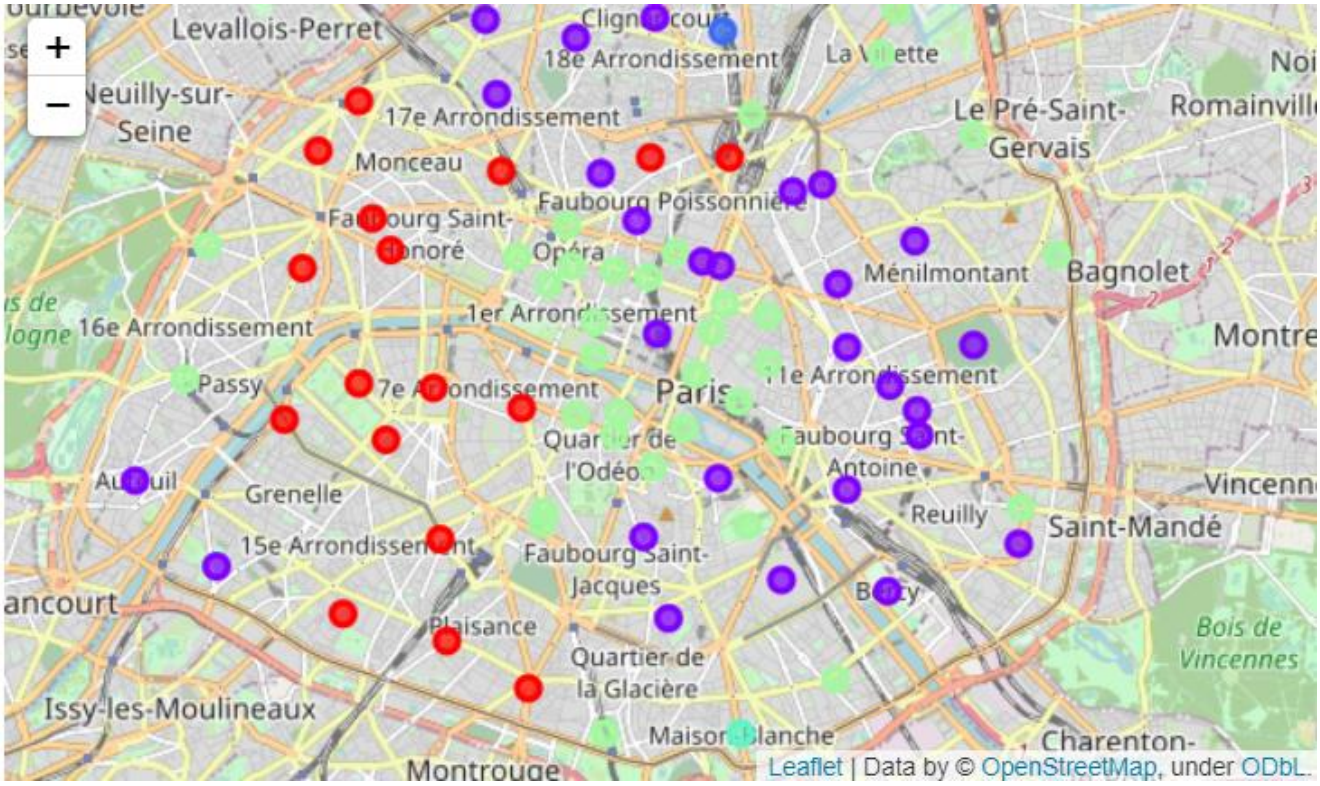
To obtain the optimal number of K clusters, the Elbow method is used. The Elbow method uses the sum of squared distances to calculate the optimal number of K clusters. From the graph below, we can see that the appropriate number of clusters is 8. (See Fig 9)



(Figure 9: Elbow Method for Optimal K)

Using the optimal K cluster number, we use the KMeans clustering method provided by the SciKit Learn Library to perform the analysis.

We now visualize the clustered neighborhoods in Paris. (See Fig. 10)



(Figure 10: Neighborhood Clusters in Paris)

6. Results and Discussion

With the neighborhoods clustered using K Means clustering method, we can see that there are 5 major clusters out of the 8 clusters selection to run K Means. These clusters are largely representative of the cities that they are in, with Cluster 0, Cluster 1 and Cluster 5 containing mostly neighborhoods from Paris and the remaining clusters containing mostly neighborhoods from Singapore.

For a first-time traveler to Paris, the neighborhoods in Cluster 5 would be an attractive location to stay in, as they are like 5 neighborhoods in Singapore. This indicates that the Singapore neighborhoods in Clusters 5 share similar characteristics to their French counterparts and would not cause too much of a culture shock to new visitors. This is seen by the majority Parisian neighborhoods in Cluster 5, but still similar enough to Singapore to contain 5 local Neighborhoods.

Travelers that have been to Paris, from Singapore, may choose the neighborhoods in Cluster 0 for their stay. The neighborhoods in Cluster 0 contain many hotels and French restaurants. Many of the neighborhoods in Cluster 0 are very indicative of French culture, with the most popular venue being French Restaurants. The second most common venue is the hotel. This allows for the traveler to easily find accommodation, immerse themselves into French culture and provide a sufficiently different travel experience.

Long term travelers and migrants may consider choosing the neighborhoods in Clusters 2 & 4, to rent a medium/long term rental or apartment. The French Neighborhoods Goutte-d'Or and Maison-Blanche are most like popular residential neighborhoods in Singapore. Additionally, we can see that Cluster 2 contains the most populous and the most population dense neighborhoods in Singapore.

Interestingly, we note that there are 3 single neighborhood clusters after the completion of the analysis. The neighborhoods in these clusters are all Singaporean neighborhoods. Looking into the venue data, the 3 neighborhoods contain unique top common venues. These might be due to the neighborhood coordinates obtained using the geocoder library. As the activity center of neighborhoods may not correspond to the geographical center of the neighborhood, these unique venues, like ‘Intersection’, are more likely.

7. Conclusion and Future Analysis

This project was created to identify similar neighborhoods between Singapore and Paris. The project aims to address the concern travelers or migrants from Singapore to France have when picking a place to stay. The project starts out with visualizing data from the World Bank to determine if there is an increasing trend in travel from Singapore to France. We can see that the overall trend for travel to European nations are on an increase. With France having the largest international arrivals compared to other high-income countries in Europe.

The venue information collected from the FourSquare API is used to cluster the neighborhoods. The clustering is performed using the K Means method. This clustering of the neighborhoods helped determine the similarity of the neighborhood and help decide for a traveler or migrant reaching Paris from Singapore.

From Cluster 1, the Parisian neighborhoods are quite like each other and tend to be significantly different from Singaporean neighborhoods. Nonetheless, it can be seen from the results that Parisian neighborhoods like Maison-Blanche and Goutte-d'Or, are good for long term migrants.

Further decision on an optimal location is determined by the stake holders on the characteristics of the neighborhoods and any additional requirements. Additional factors that would affect the decision would be rental cost, proximity of culture specific cuisine, or proximity to amenities like parks or cinemas. Further Analysis would require additional criteria and would benefit from including rental information into the analysis.