**“A Delightful Tale of Two Cities”**

*A data science report for work submitted*

*By*

**VENKATA SHANMUKESWAR TATA**

*as a part of*

**IBM Professional Certification Program with Coursera**

Under the supervision of

|  |  |
| --- | --- |
| **Coursera** | **IBM** |
| Coursera is an American massive open online course provider founded in 2012 by Stanford University's computer science professors Andrew Ng and Daphne Koller that offers massive open online courses, specializations, degrees, professional and master track courses. | International Business Machines Corporation is an American cloud platform and cognitive solutions multinational technology and consulting company headquartered in Armonk, New York, with more than 350,000 employees serving clients in 170 countries. |

A picture containing chart

Description automatically generated A picture containing drawing

Description automatically generated

**Project Details:**

|  |  |
| --- | --- |
| Author | Venkata Shanmukeswar Tata |
| Published | Oct 15, 2020 |
| Version | 1.0 |
| Contributions | NA |

# Cover Page

# Table Of Contents

# Abstract

# Introduction

# Business Problem

# Data Description

## London

## ArcGIS API

## Paris

## Foursquare API Data

# Methodology

## Data Collection

## Data Preprocessing

## Feature Selection

## Feature Engineering

## Visualizing the Neighborhoods of London and Paris

### Neighborhoods of London

### Neighborhoods of Paris

# Model Building

## Model Building for London neighborhood

## Model Building for Paris Neighborhood

# Results and Discussion

# Conclusion

# References

Table of Contents

1. Cover Page 1

[2. Table Of Contents 3](#_Toc53755952)

[3. Abstract 4](#_Toc53755953)

[4. Introduction 5](#_Toc53755954)

[5. Business Problem 6](#_Toc53755955)

[6. Data Description 7](#_Toc53755956)

[6.1 London 7](#_Toc53755957)

[6.2 ArcGIS API 7](#_Toc53755958)

[6.3 Paris 8](#_Toc53755959)

[6.4 Foursquare API Data 8](#_Toc53755960)

[7. Methodology 9](#_Toc53755961)

[7.1 Data Collection 10](#_Toc53755962)

[7.2 Data Preprocessing 11](#_Toc53755963)

[7.3 Feature Selection 11](#_Toc53755964)

[7.4 Feature Engineering 12](#_Toc53755965)

[7.5 Visualizing the Neighborhoods of London and Paris 13](#_Toc53755966)

[7.5.1 Neighborhoods of London 13](#_Toc53755967)

[7.5.2 Neighborhoods of Paris 14](#_Toc53755968)

[8. Model Building 16](#_Toc53755969)

[8.1 Model Building for London neighborhood 16](#_Toc53755970)

[8.2 Model Building for Paris Neighborhood 16](#_Toc53755971)

[9. Results and Discussion 18](#_Toc53755972)

[10. Conclusion 23](#_Toc53755973)

[References 24](#_Toc53755974)

**Abstract**

As a part of Coursera IBM Professional data science project, we will be analyzing the similarities and dissimilarities of two cities. **London (capital of the United Kingdom) and Paris (capital of France)** has been taken as an example for this study

Analyzing cities using venue based data from Foursquare lead to an overall understanding of the type of venues in each neighborhood and presented some of the key features of the cities but the level of data is not adequate to provide a comprehensive analysis for a city-to-city comparison. For a potential interested person (job-seeker or person deciding to move to either of the cities) or a bigger clientele like a business corporation or city planners, one would need to do a more detailed analysis adding features such as rents, salaries, transportation, cost of living, growth rate, economy, etc.

The capstone project provided a medium to understand in depth about how real life data science projects work and what all steps go in building a data science methodology. All steps from understanding the business problem, data understanding to data preparation, and model building were discussed in detail here.

The reports highlights the methodology and analysis used for the final capstone project in the [IBM Data Science Professional](https://www.coursera.org/professional-certificates/ibm-data-science) course. Full source code is available in [GitHub](https://github.com/alone-survivour/IBM-DS-Professional/blob/CapstoneFinalProject/Battle_Of_Neighbourhoods.ipynb)

1. **Introduction**

Picking a city, when it comes to [London](https://en.wikipedia.org/wiki/London) and [Paris](https://en.wikipedia.org/wiki/Paris) is always a hard decision as both these cities are truly global, multicultural, and cosmopolitan cities found at the heart of two European nations. Along with being two of Europe’s most important diplomatic centres, they are major centers for finance, commerce, sciences, fashion, arts, culture and gastronomy. Both London (capital of the United Kingdom) and Paris (capital of France) have a rich history and are two of the most visited and sought-after cities in Europe. London is the largest city within the UK and stands on River Thames in South East England. Paris, on the other hand, is located in the north-central part of the nation. Similar to London, the city also stands along a river, commonly known as the Seine River.

A Tale of Two cities, a novel written by Charles Dickens was set in London and Paris which takes place during the French Revolution. These cities were both happening then and now. A lot has changed over the years and we now take a look at how the cities have grown.

London and Paris are quite the popular tourist and vacation destinations for people all around the world. They are diverse and multicultural and offer a wide variety of experiences that is widely sought after. We try to group the neighborhoods of London and Paris respectively and draw insights to what they look like now.

A Tale of Two Cities was published in 1859, right between Little Dorrit and Great Expectations, two of other Dicken's novels. These are novels set in Dicken's own century, if several decades before the years of their publication. And they're focused on individual stories of crime, poverty and class in 19th Century England and Europe. Tale of Two Cities, by contrast, is one of only two works Dickens wrote, that could properly be considered historical novels. The other is Barnaby Rudge, set in the 1770s and the early 1780s. A Tale of Two Cities has always been considered one of the least Dickensian of Dickens' novels. For example, a scholar writing for the Cambridge History of English Literature, published in 1941, wrote, many people who do not care for the rest of Dickens like it greatly; many who are enthusiastic about Dickens refuse to give it a second reading. It is the least Dickensian of all the tales. And it does have an odd relationship to his copious.

1. **Business Problem**

Our goal is to perform a comparison of the two cities to see how similar or dissimilar they are. Such techniques allow users to identify similar neighbourhoods among cities based on amenities or services being offered locally, and thus can help in understanding the local area activities, what are the hubs of different activities, how citizens are experiencing the city, and how they are utilising its resources.

**What kind of clientele would benefit from such an analysis?**

1. A potential job seeker with transferable skills may wish to search for jobs in selective cities which provide the most suitable match for their qualifications and experience in terms of salaries, social benefits, or even in terms of a culture fit for expats.
2. Further, a person buying or renting a home in a new city may want to look for recommendations for locations in the city similar to other cities known to them.
3. Similarly, a large corporation looking to expand its locations to other cities might benefit from such an analysis.
4. Many within-city urban planning computations might also benefit from modelling a city’s relationship to other cities.
5. **Data Description**

We require geographical location data for both London and Paris. Postal codes in each city serve as a starting point. Using Postal codes, we use can find out the neighborhoods, boroughs, venues and their most popular venue categories.

* 1. **London**

To derive our solution, We scrape our data from [here](https://en.wikipedia.org/wiki/List_of_areas_of_London)

This Wikipedia page has information about all the neighbourhoods, we limit it London.

1. borough: Name of Neighborhood
2. town: Name of borough
3. postcode: Postal codes for London.

This Wikipedia page lacks information about the geographical locations. To solve this problem we use ArcGIS API.

* 1. **ArcGIS API**

ArcGIS Online enables you to connect people, locations, and data using interactive maps. Work with smart, data-driven styles and intuitive analysis tools that deliver location intelligence. Share your insights with the world or specific groups.

More specifically, we use ArcGIS to get the geo locations of the neighbourhoods of London. The following columns are added to our initial dataset which prepares our data.

1. latitude: Latitude for Neighborhood
2. longitude: Longitude for Neighborhood
   1. **Paris**

To derive our solution, We leverage JSON data available at [here](https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e)

The JSON file has data about all the neighbourhoods in France, we limit it to Paris.

1. postalcode : Postal codes for France
2. nomcomm : Name of Neighborhoods in France
3. nomdept : Name of the boroughs, equivalent to towns in France
4. geo\_point\_2d : Tuple containing the latitude and longitude of the Neighborhoods.
   1. **Foursquare API**

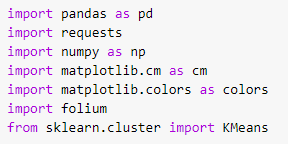
We will need data about different venues in different neighbourhoods of that specific borough. In order to gain that information we will use "Foursquare" locational information. Foursquare is a location data provider with information about all manner of venues and events within an area of interest. Such information includes venue names, locations, menus and even photos. As such, the foursquare location platform will be used as the sole data source since all the stated required information can be obtained through the API.

After finding the list of neighbourhoods, we then connect to the Foursquare API to gather information about venues inside each and every neighbourhood. For each neighbourhood, we have chosen the radius to be 500 meters.

The data retrieved from Foursquare contained information of venues within a specified distance of the longitude and latitude of the postcodes. The information obtained per venue as follows:

1. Neighborhood : Name of the Neighborhood
2. Neighborhood Latitude : Latitude of the Neighborhood
3. Neighborhood Longitude : Longitude of the Neighborhood
4. Venue : Name of the Venue
5. Venue Latitude : Latitude of Venue
6. Venue Longitude : Longitude of Venue
7. Venue Category : Category of Venue
8. **Methodology**

Based on all the information collected for both London and Paris, we have sufficient data to build our model. We cluster the neighbourhoods together based on similar venue categories. We then present our observations and findings. Using this data, our stakeholders can take the necessary decision.



Package breakdown:

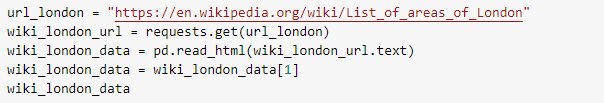
1. Pandas : To collect and manipulate data in JSON and HTMl and then data analysis
2. requests : Handle http requests
3. matplotlib : Detailing the generated maps
4. folium : Generating maps of London and Paris
5. sklearn : To import Kmeans which is the machine learning model that we are using.

The approach taken here is to explore each of the cities individually, plot the map to show the neighbourhoods being considered and then build our model by clustering all of the similar neighbourhoods together and finally plot the new map with the clustered neighbourhoods. We draw insights and then compare and discuss our findings.

* 1. **Data Collection**

In the data collection stage, we begin with collecting the required data for the cities of London and Paris. We need data that has the postal codes, neighbourhoods and boroughs specific to each of the cities.

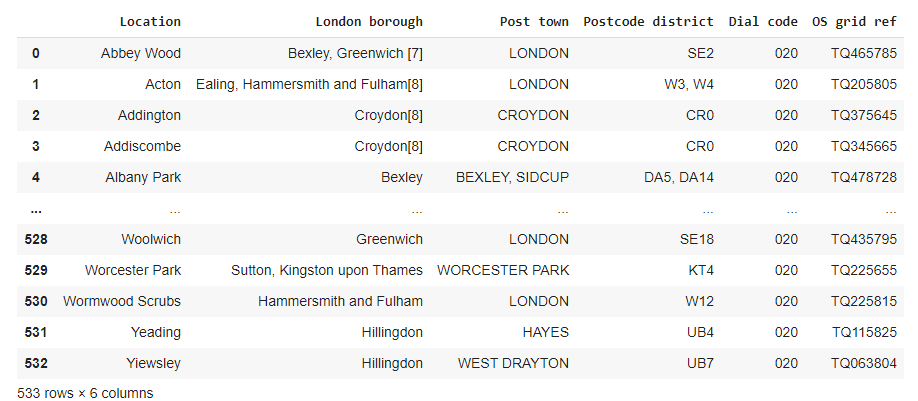
To collect data for London, we scrape the List of areas of London Wikipedia page to take the 2nd table using the following code:



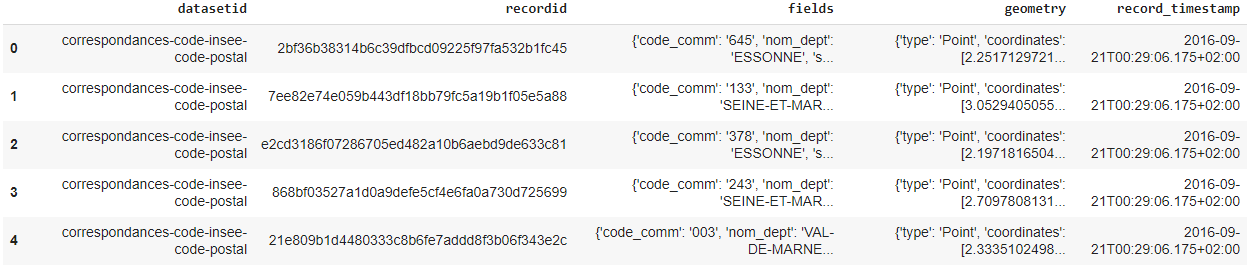
To collect data for Paris, we download the JSON file containing all the postal codes of France from [datasets](https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e)



The resultant dataframes are:



**Table 1:** London city boroughs data frame



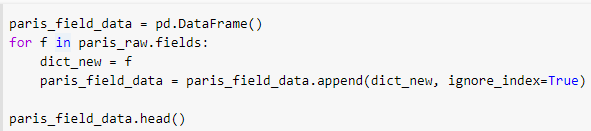
**Table 2:** Paris city boroughs data frame

* 1. **Data pre-processing**

For London, We replace the spaces with underscores in the title.The borough column has numbers within square brackets that we remove using:

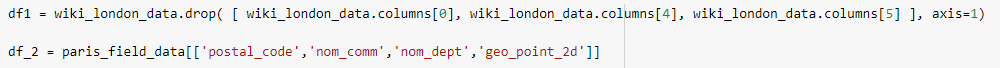


For Paris, we break down each of the nested fields and create the dataframe that we need:



* 1. **Feature Selection**

For both of our datasets, we need only the borough, neighborhood, postal codes and geolocations (latitude and longitude). So we end up selecting the columns that we need by:



* 1. **Feature Engineering**

Both of our Datasets actually contain information related to all the cities in the country. We can narrow down and further process the data by selecting only the neighbourhoods pertaining to 'London' and 'Paris'. Looking over our London dataset, we can see that we don't have the geolocation data. We need to extrapolate the missing data for our neighbourhoods. We perform this by leveraging the ArcGIS API. With the Help of ArcGIS API, we can get the latitude and longitude of our London neighbourhood data. Defining London arcgis geocode function to return latitude and longitude. Passing postal codes of London to get the geographical co-ordinates. Extracting the latitude from our previously collected coordinates.

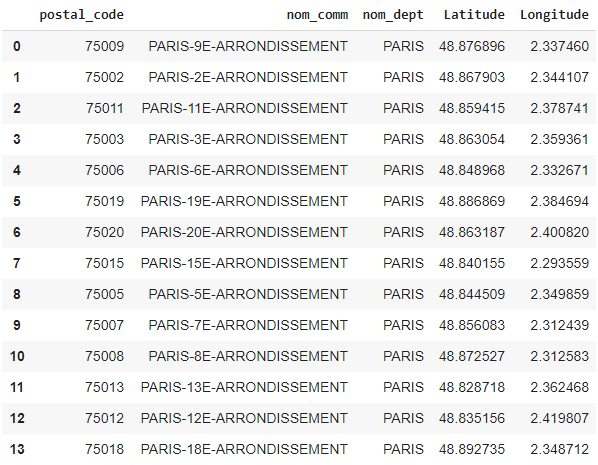
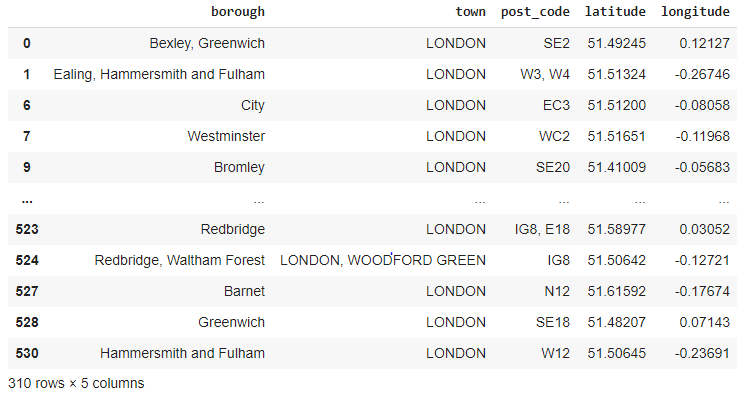
London:



Paris:



Final data frames will be as below:

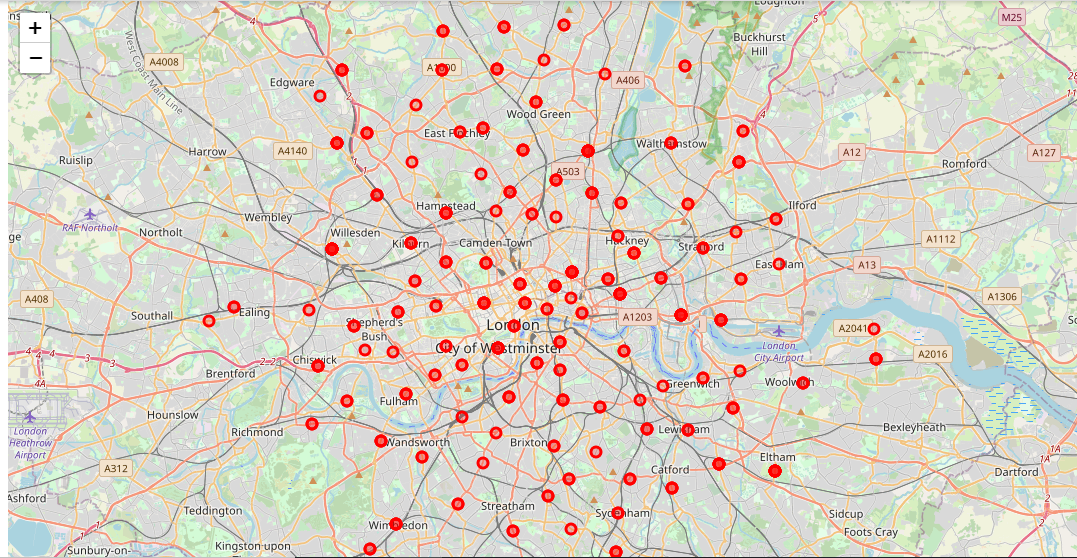
 

**Table 3:** Paris featured engineered data frame(left) and London featured engineered data frame(Right)

* 1. **Visualizing the neighborhoods**
     1. London

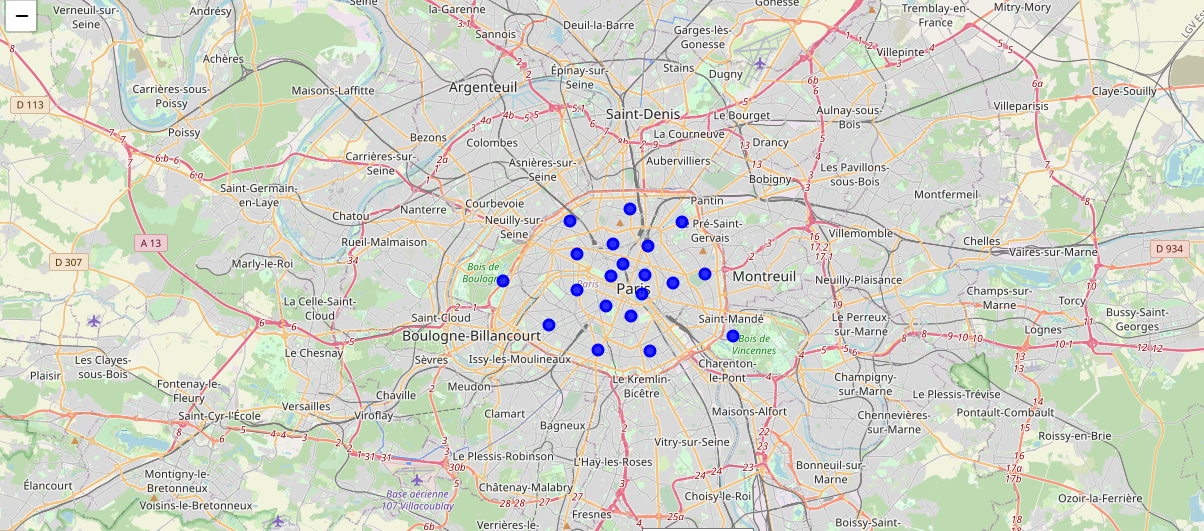


Map:



* + 1. Paris

Applying the above function to Paris neighborhood, we will see below map



Top venue categories of neighborhoods in London:



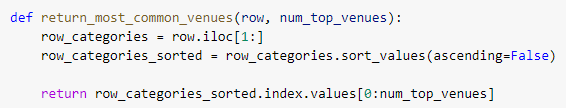
Getting the venues in London:



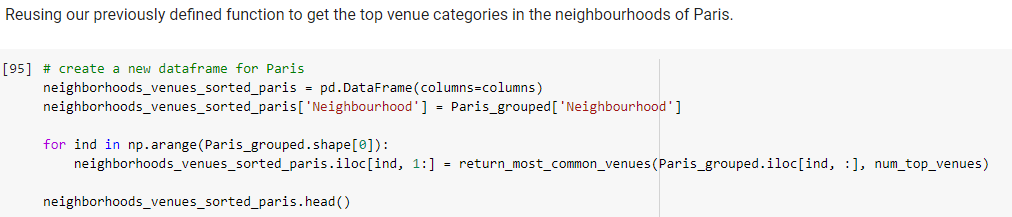
Venue categories:



Top most common venue categories can be retrieved by defining a function



Above concepts are applied for retrieving the top venue categories of neighborhoods in Paris:



Top venue locations in London:

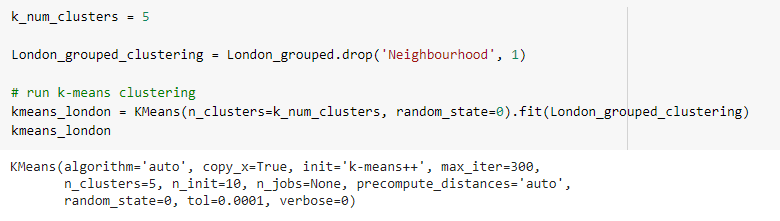


Top Venue locations in Paris:



1. **Model Building**
   1. **Model Building for London neighborhoods**

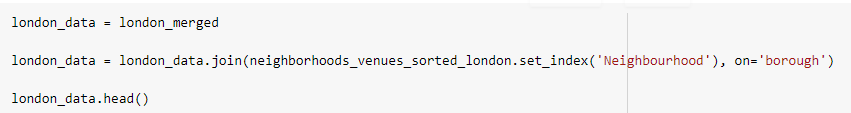
We use K-Means clustering algorithm for building our model



Labelling the data:

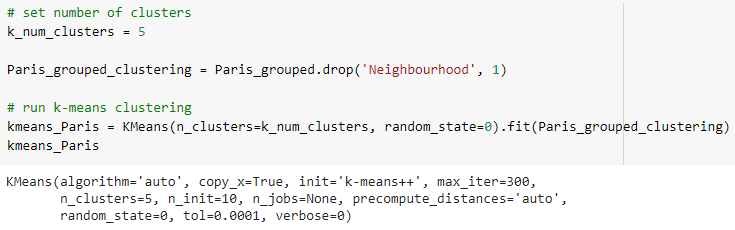


Merging the data:

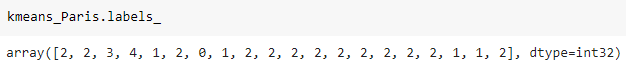


* 1. **Model building for Paris neighborhood**

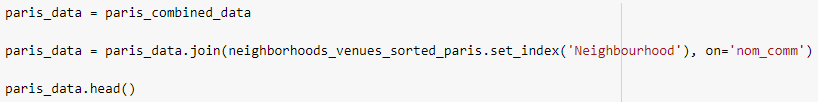
We use K-Means clustering algorithm for the paris neighborhoods as well.



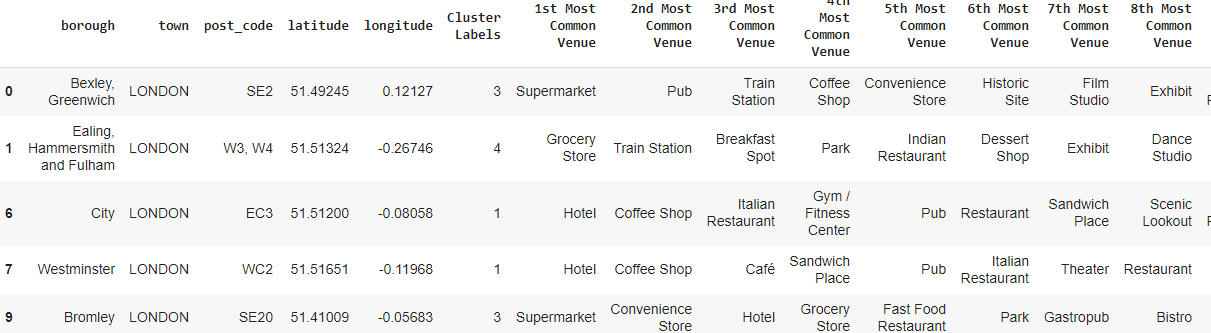
Labelling the data and merging to form data frame



Join paris\_combined\_data with our neighbourhood venues sorted to add latitude & longitude for each of the neighborhood to prepare it for plotting



London Neighborhood dataframe



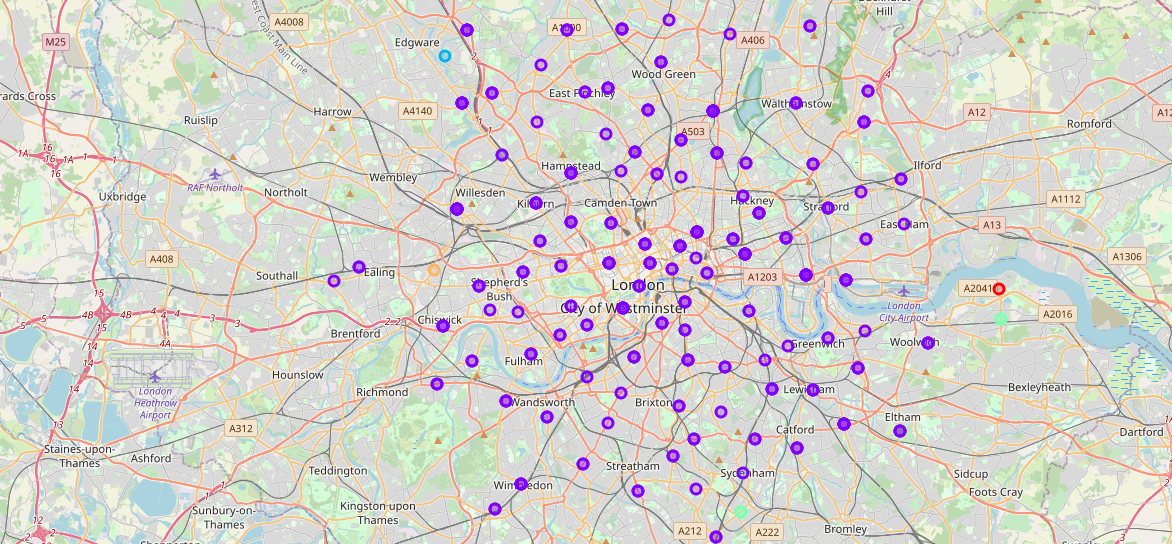
Paris Neighborhood dataframe



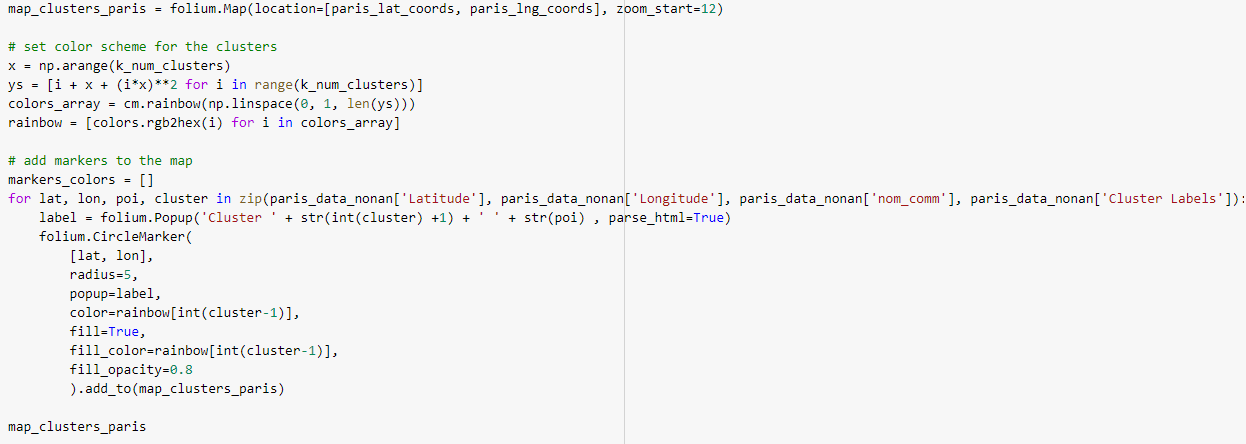
1. **Results and Discussions**

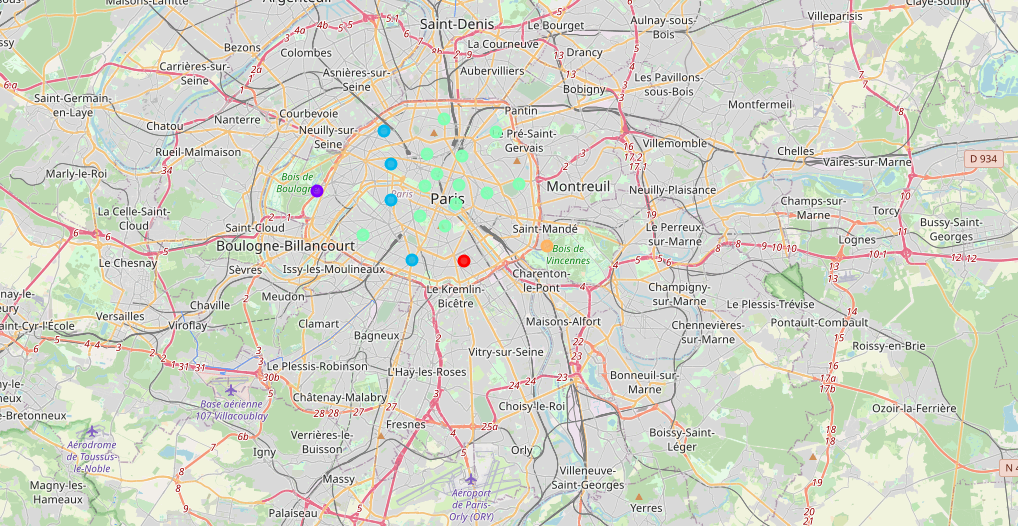
Plotting the clusters on London map:





Plotting the clusters on Paris map:



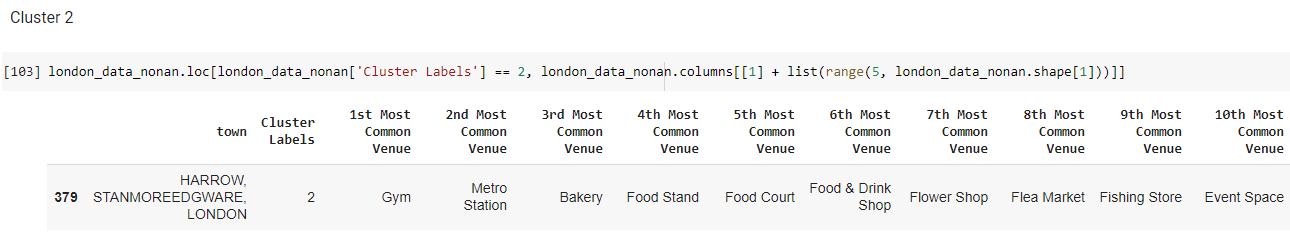


**Clusters Results of London:**

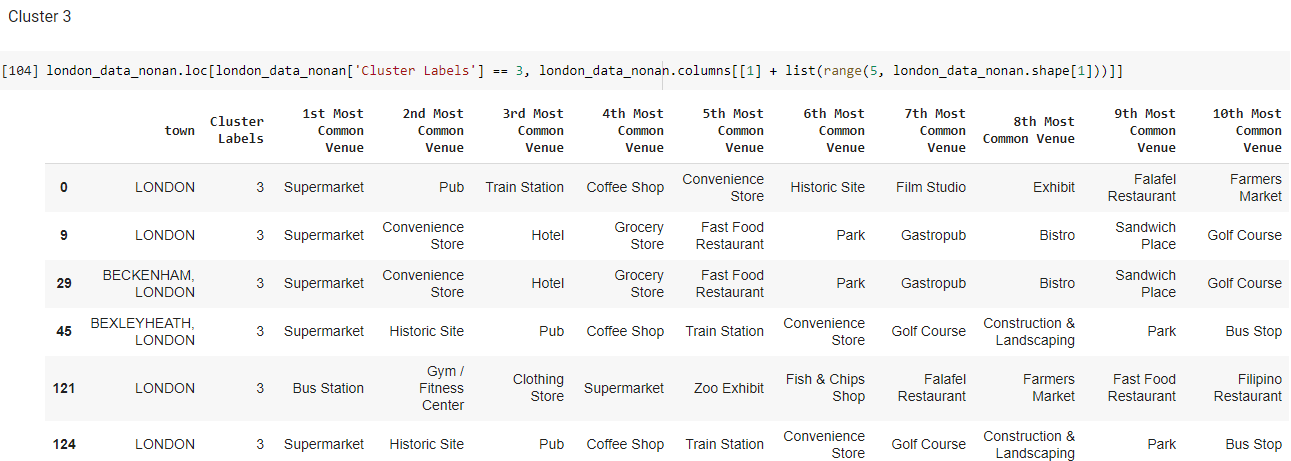
Cluster 1:



Cluster 2:



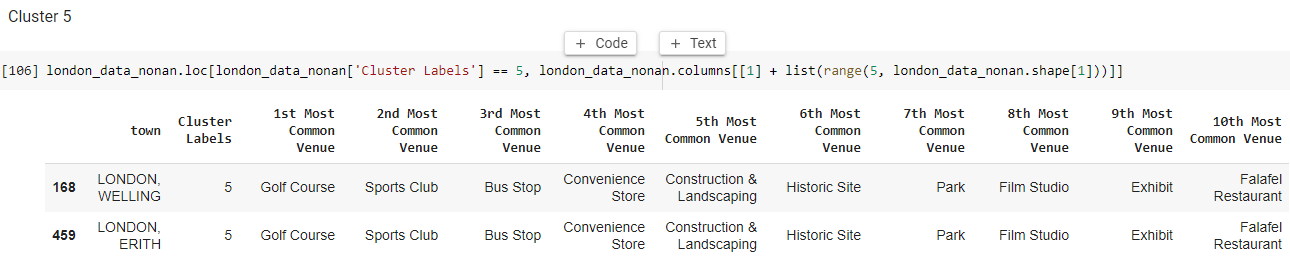
Cluster 3:



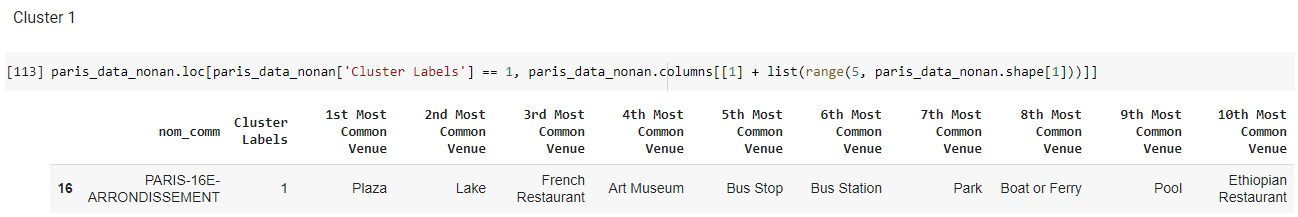
Cluster 4:



Cluster 5:

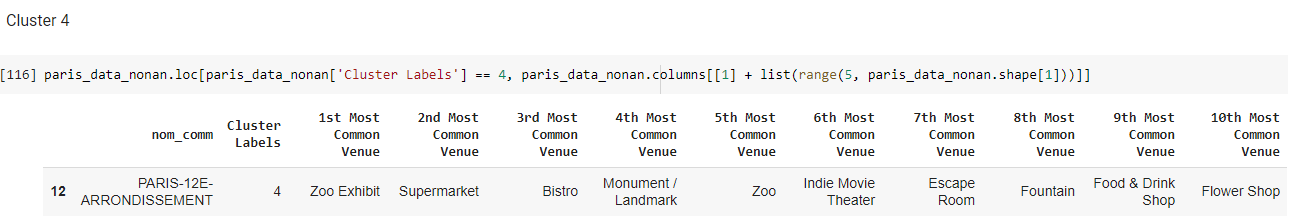


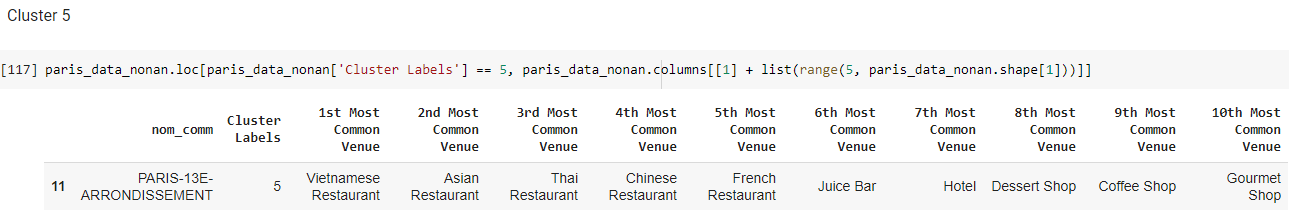
**Clusters results on Paris:**











The neighbourhoods of London are very mulitcultural. There are a lot of different cusines including Indian, Italian, Turkish and Chinese. London seems to take a step further in this direction by having a lot of Restaurants, bars, juice bars, coffee shops, Fish and Chips shop and Breakfast spots. It has a lot of shopping options too with that of the Flea markets, flower shops, fish markets, Fishing stores, clothing stores. The main modes of transport seem to be Buses and trains. For leisure, the neighbourhoods are set up to have lots of parks, golf courses, zoo, gyms and Historic sites.

Overall, the city of London offers a multicultural, diverse and certainly an entertaining experience.

Paris is relatively small in size geographically. It has a wide variety of cusines and eateries including French, Thai, Cambodian, Asian, Chinese etc. There are a lot of hangout spots including many Restaurants and Bars. Paris has a lot of Bistro's. Different means of public transport in Paris which includes buses, bikes, boats or ferries. For leisure and sight seeing, there are a lot of Plazas, Trails, Parks, Historic sites, clothing shops, Art galleries and Museums. Overall, Paris seems like the relaxing vacation spot with a mix of lakes, historic spots and a wide variety of cusines to try out.

1. **Conclusions**

The purpose of this project was to explore the cities of London and Paris and see how attractive it is to potential tourists and migrants. We explored both the cities based on their postal codes and then extrapolated the common venues present in each of the neighbourhoods finally concluding with clustering similar neighbourhoods together.

We could see that each of the neighbourhoods in both the cities have a wide variety of experiences to offer which is unique in its own way. The cultural diversity is quite evident which also gives the feeling of a sense of inclusion.

Both Paris and London seem to offer a vacation stay or a romantic gateway with a lot of places to explore, beautiful landscapes and a wide variety of culture. Overall, it's up to the stakeholders to decide which experience they would prefer more and which would more to their liking.

The capstone project provided a medium to understand in depth about how real life data science projects work and what all steps go in building a data science methodology. All steps from understanding the business problem, data understanding to data preparation, and model building were discussed in detail here. Many drawbacks of the current analysis and further ways to improve the analysis were also mentioned. This was an initial attempt to understand and solve the business problem at hand. However, there still exists a huge potential to extend this project in real life scenarios.

**References**

1. <https://en.wikipedia.org/wiki/List_of_London_boroughs>
2. <https://en.wikipedia.org/wiki/Arrondissements_of_Paris>
3. <https://foursquare.com/>
4. <https://developers.arcgis.com/python/>
5. <https://python-visualization.github.io/folium/modules.html>
6. <https://pandas.pydata.org/pandas-docs/stable/>
7. <https://scikit-learn.org/stable/index.html>
8. <https://medium.com/@yrnigam/how-to-write-a-data-science-report-181bd49d8f4d>
9. <https://towardsdatascience.com/a-tale-of-two-cities-e693c15b3ddb>
10. <https://www.coursera.org/lecture/historical-fiction/dickens-and-the-french-revolution-a-tale-of-two-cities-oVjhu>