# **Final Report for IBM Applied Data Science Capstone Project**

# **1. Introduction**

As time flies and cities develop, it is meaningful to discuss the differences between cities as decades passed. I put my focus on two cities Paris and London, A lot has changed over the years and we now take a look at how the cities have grown.

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

The aim is to help tourists to choose their destinations depending on the experience of the local neighborhoods. My findings will help stakeholders to make decisions and address any concerns they will have including the different kinds of cuisines, stores and what the city looks like right now.

**2. Data**

I acquire geographical location data for both London and Paris. Using postal codes in each city as a starting point.

## London

Please see the following link for the data

<https://en.wikipedia.org/wiki/List_of_areas_of_London>

I set up different names for different kind of data information. *borough* is the Name of Neighborhood; *town* is Name of borough; *post\_code* is the Postal codes for London.

There is no information about the geographical locations for London. To solve this, we derive the data of ArcGIS API, which can help us to connect people, locations, and data using interactive maps. We can obtain more detailed information of geographical locations of London. To simplify the data, i also set up different names for different information. *latitude* is the Latitude for Neighborhood; *longitude* is the Longitude for Neighborhood.

## Paris

Please see the following link for the data

<https://www.data.gouv.fr/fr/datasets/r/e88c6fda-1d09-42a0-a069-606d3259114e>

We limit this data to only see and analyze the city of Paris. Here as well, i set up different names for different information. *postal\_code* is the Postal codes for France; *nom\_comm* is the Name of Neighborhood in France; *nom\_dept* is the Name of the boroughs, equivalent to towns in France; *geo\_point\_2d* is the Tuple containing the latitude and longitude of the Neighbourhoods.

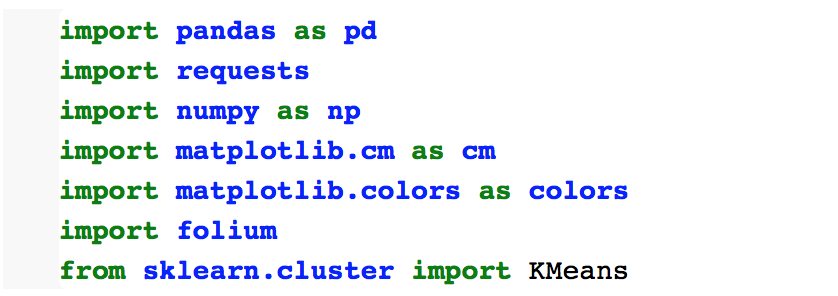
## Foursquare API Data

We will need data about different venues in different neighborhoods. In order to obtain it I will use "Foursquare" Location information. Information like venue names, locations, menus and even photos will be included. After obtaining the list of neighborhood, I will connect it to the Foursquare API to gather information about venues inside each and every neighborhood. For each neighborhood, I have chosen the radius to be 500 meters. In this data set, *Neighborhood* is the Name of the Neighborhood; *Neighborhood Latitude* is the Latitude of the Neighborhood; *Neighborhood Longitude* is the Longitude of the Neighborhood; *Venue* is the Name of the Venue; *Venue Latitude* is the Latitude of Venue; *Venue Longitude* is the Longitude of Venue; *Venue Category* is the Category of Venue

Based on the information i collected for both London and Paris, there will be sufficient data to build the model. After analyzing, the observations and findings can be used for people to make decisions.

1. **Methodology**

First of all, we need to import all packages that we will need.



In order to analyze each of the cities individually, we plot the map to show the neighborhoods being considered, then build our model by clustering all of the similar neighborhoods together. Finally we can plot a new map with the clustered neighborhoods. At last, we will go into the insights and discuss the findings.

1. **Data Processing**

At this stage, I begin with collecting the required data for the cities of London and Paris. We need a data set that includes the information of the postal codes, neighborhood and boroughs of each city.

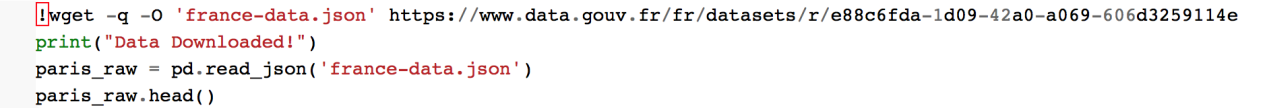
For the data of London, we scrape the List of areas of London from wikipedia page to take the 2nd table, by using the following code:



After running these codes, we get the following table

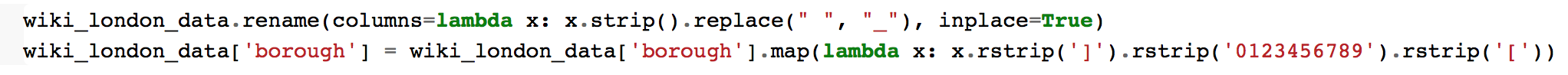


On the other hand, for the data of Paris, we use Pandas to load the table after reading the JSON file:

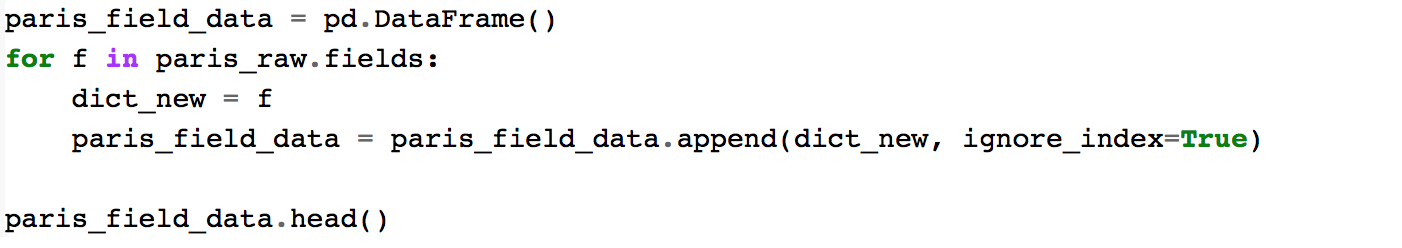




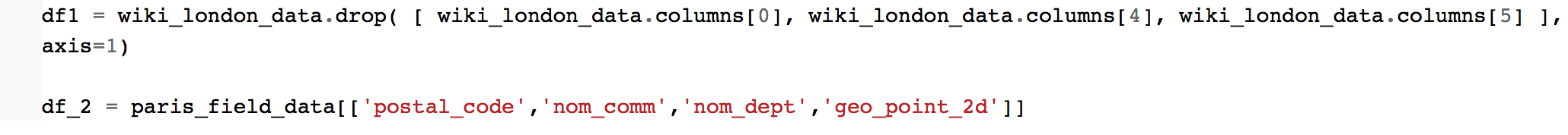
To process the data of London, we replace the spaces with underscores in the title. The *borough* column has numbers within square brackets that we need to remove by using:



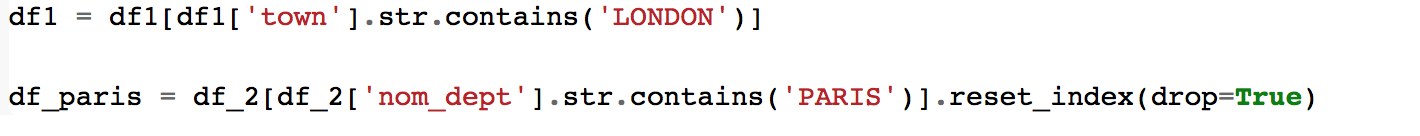
For Paris on the other hand:



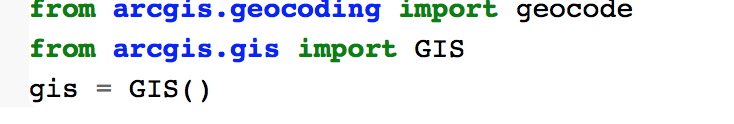
For both of our data sets, we only need the data of borough, neighborhood, postal codes and geographical locations (latitude and longitude). So we need to select the useful columns:

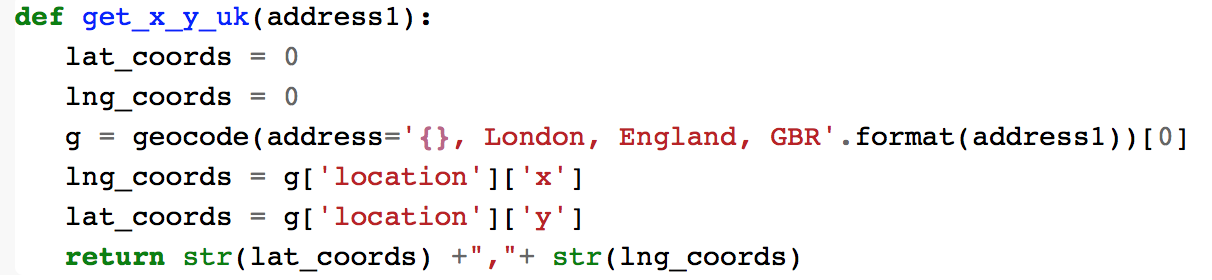


Both of the Data sets contain information that related to all other cities in the whole country. We can narrow down the data by selecting only the neighborhoods pertaining to 'London' and 'Paris'



When looking over the London data set, we can see that we don't have the geographical location data. So we need to extrapolate the missing data for our neighborhoods. We perform this by leveraging the ArcGIS API. With this we can get the latitude and longitude of our London neighborhood data.

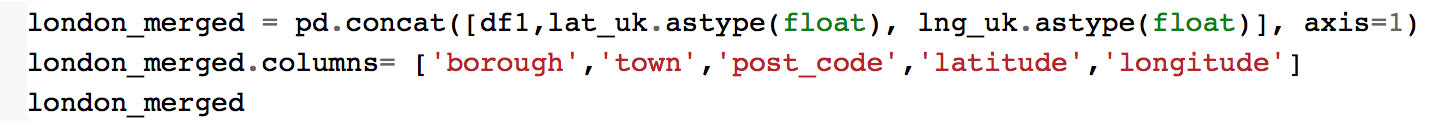




Then passing postal codes of London to get the geographical co-ordinates.

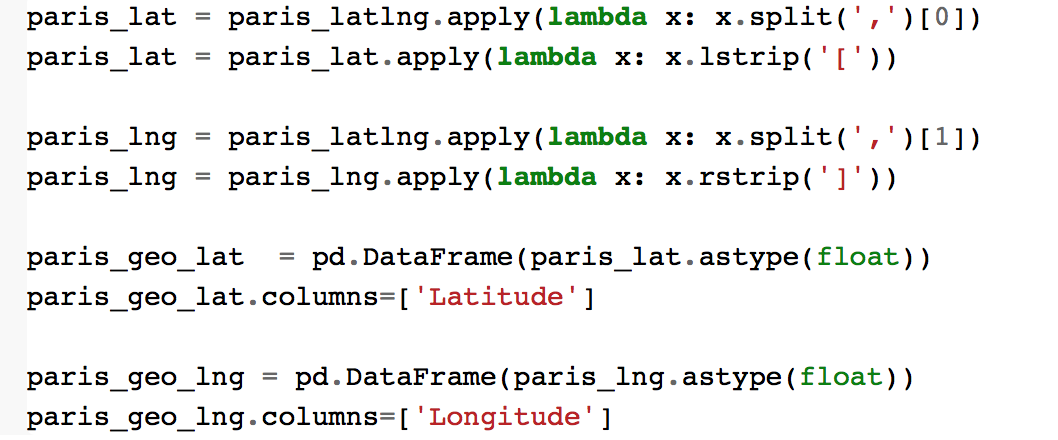
屏幕快照 2020-11-14 下午9.01.23

Then we merge our source data with the geographical co-ordinates to make our data set ready.

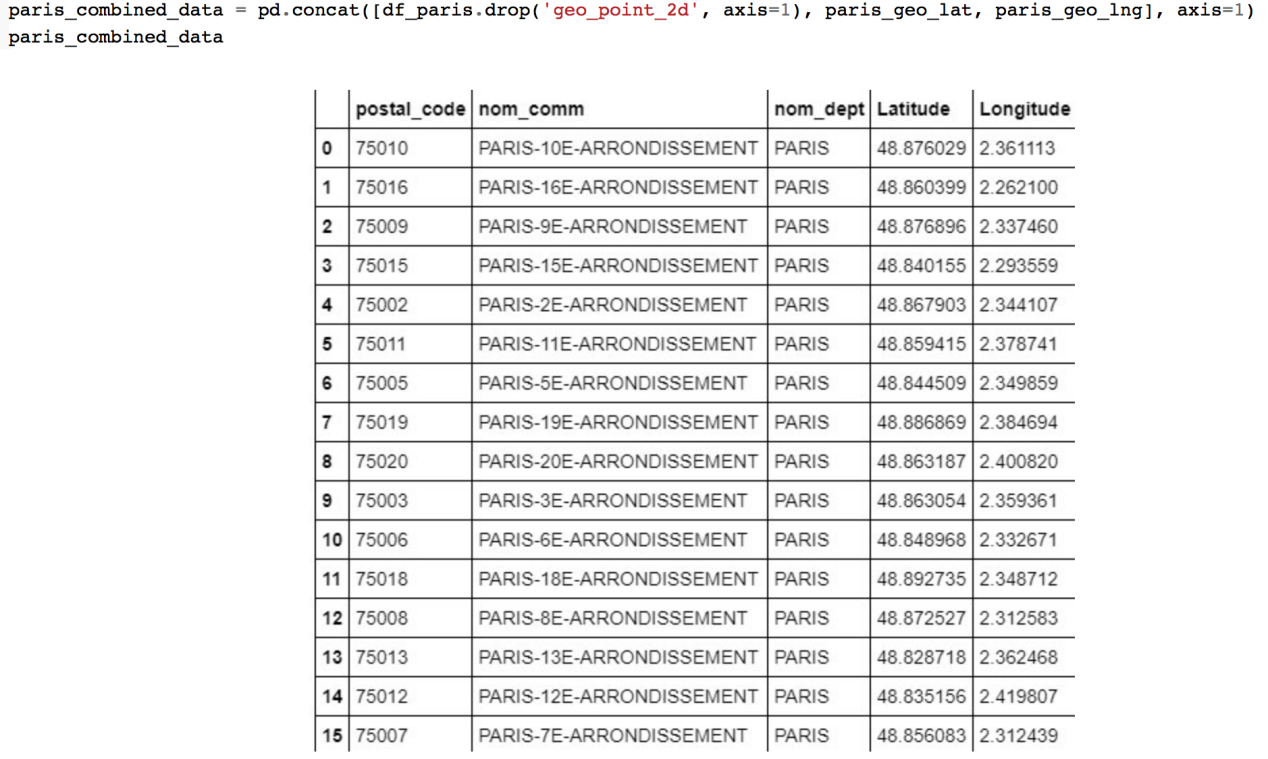




On the other hand for the Paris data set, we don't need to get the geographical coordinates of that, we just need to extract the latitude and longitude for the column:



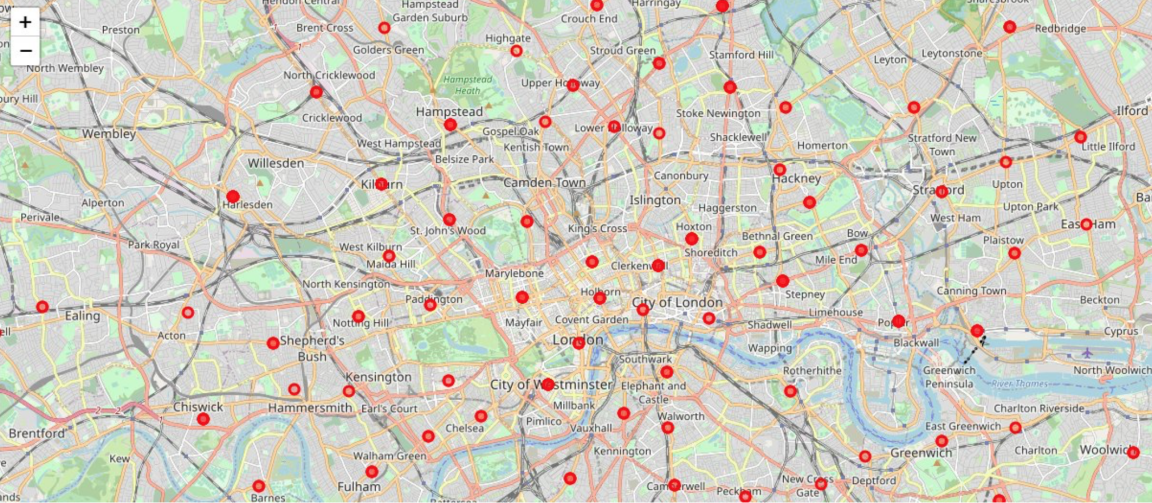
Then create our Paris data set with the required information:



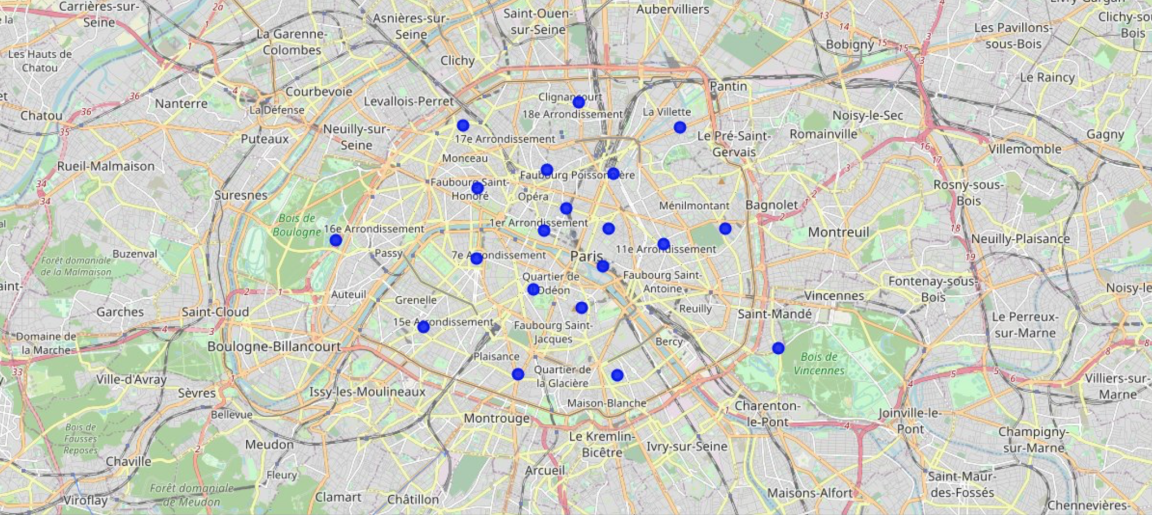
Now we can jump into the next stage of visualizing the plots.

By using the Folium package, we can visualize the maps of London and Paris with the neighborhoods that we collected.

Neighborhood map of London:



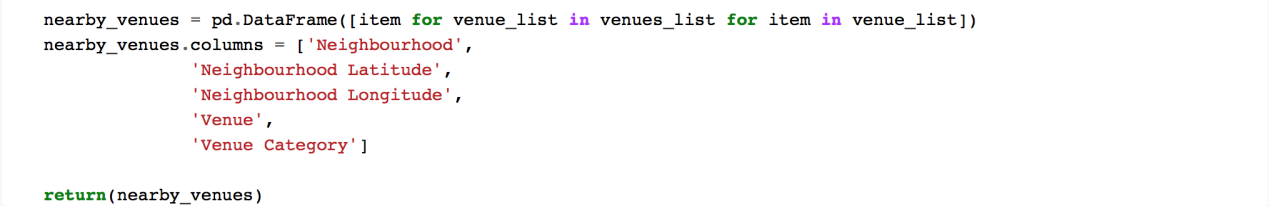
On the other hand for the city of Paris:



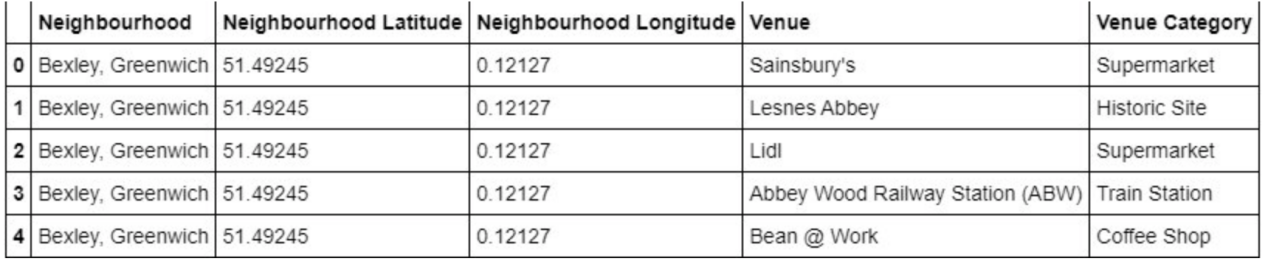
Now that we have visualized the neighborhoods of two cities, then we need to find out what each neighborhood looks like, and what are the common venues and venue categories within a 500m radius.

With the help of Foursquare we can define a function which collects information pertaining to each neighborhood including that of the name of the neighborhood, geographical coordinates, venue and venue categories.

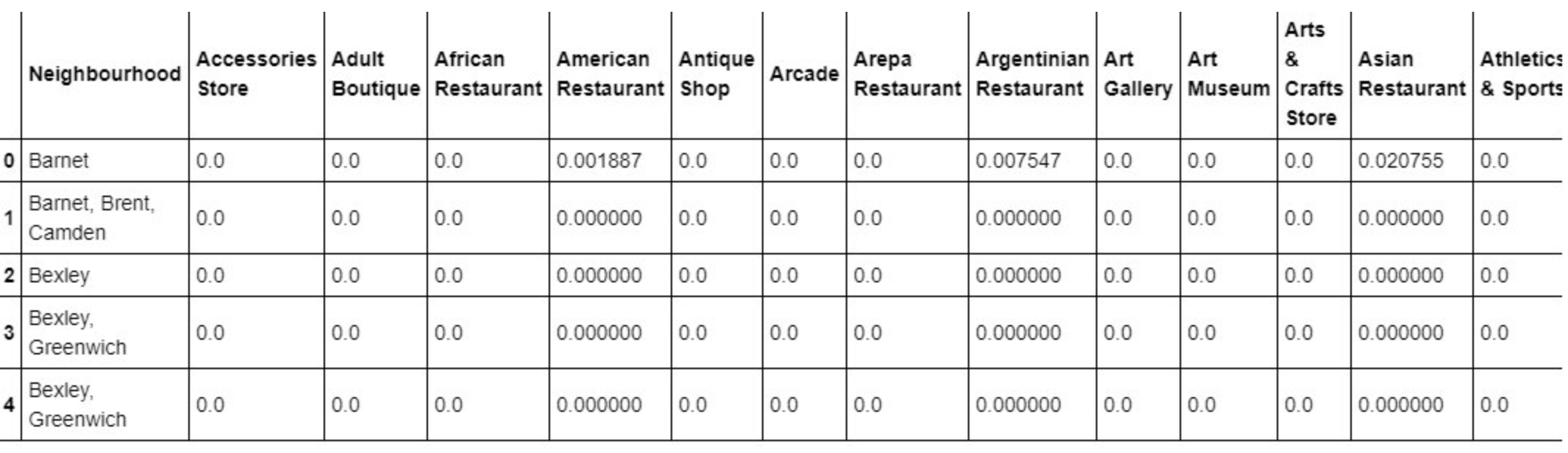


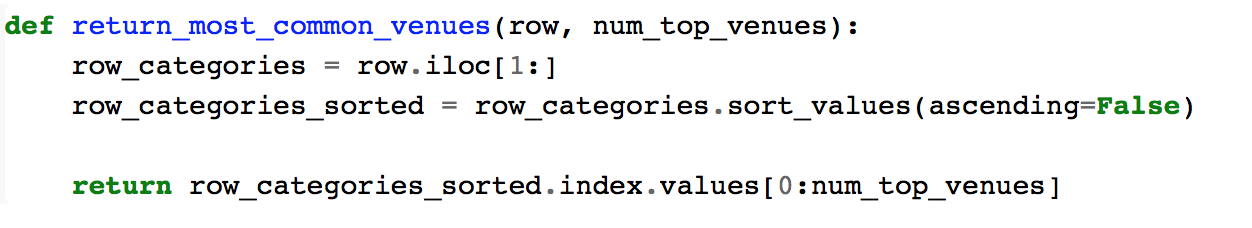


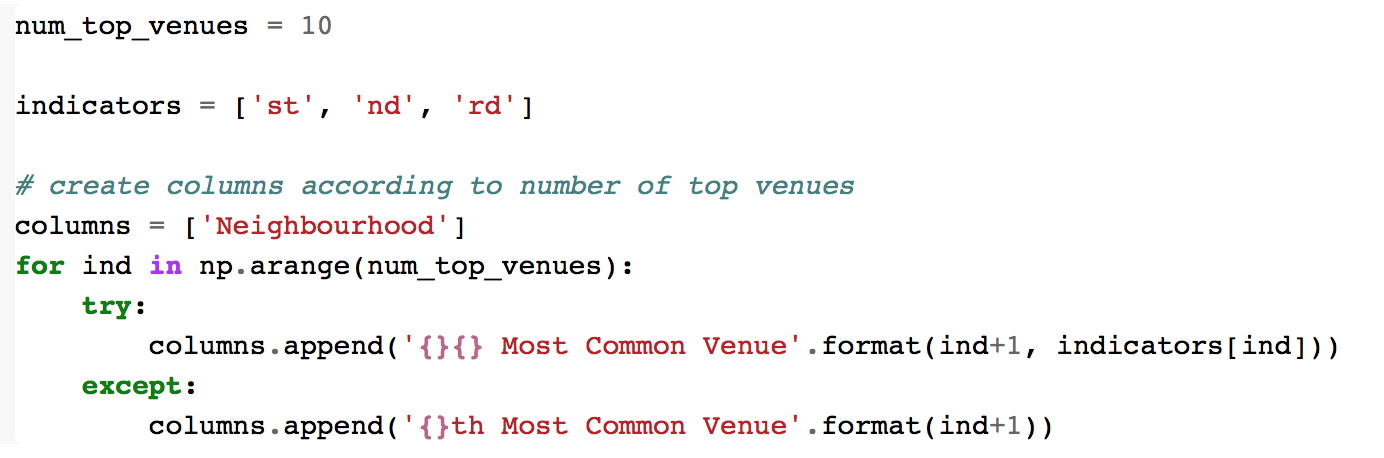
And the result will look like:



Since we are trying to find out what are the different kinds of venue categories in each neighborhood, we can use the One Hot Encoding to work with our categorical data type of the venue categories. We perform one hot encoding and then calculate the mean of the grouped venue categories for each of the neighborhoods.

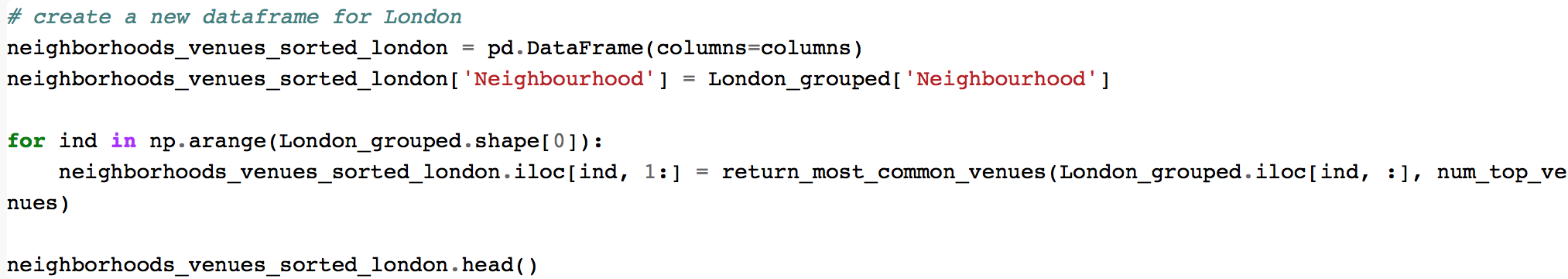


In the next step, we rank and label the top venue categories in our neighborhood.



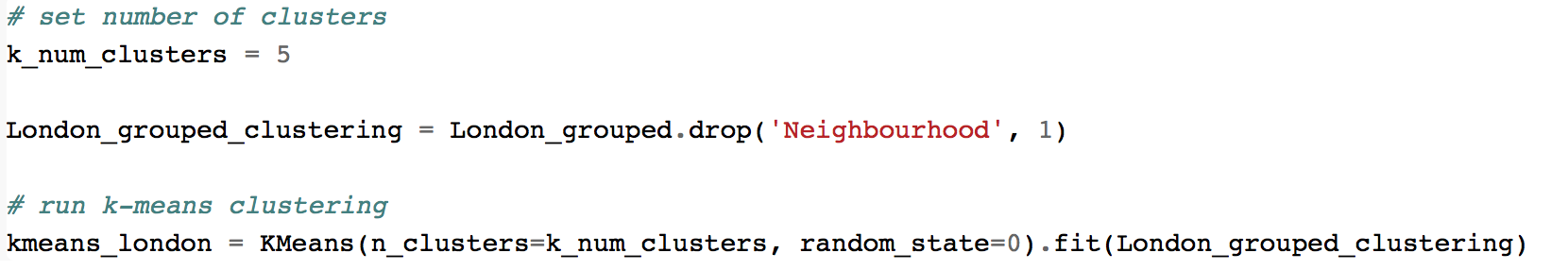
There are many categories, so we only consider top 10 categories of them.

Then we get the top venue categories in the neighborhood of London.



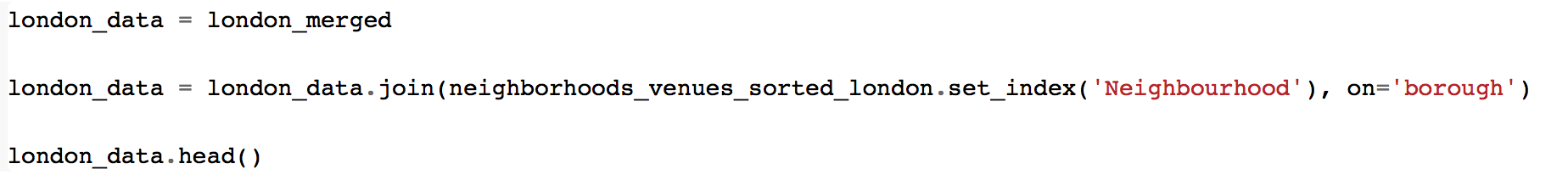


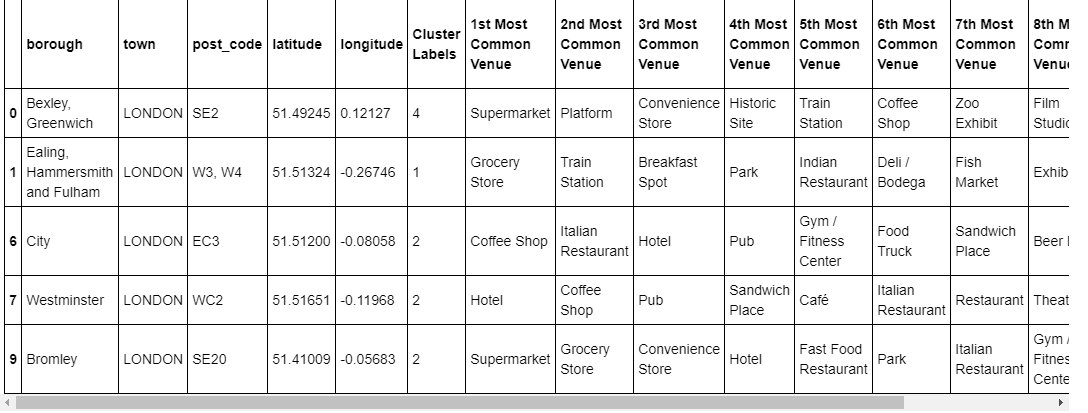
Then we move on to **Model Building part.** We will be using KMeans Clustering Machine learning algorithm to cluster similar neighborhood together. And we will be going with the number of clusters as 5.



屏幕快照 2020-11-14 下午9.52.06

We then join London\_merged with our neighborhood venues sorted to add latitude & longitude for each of the neighborhood to prepare it for visualization.

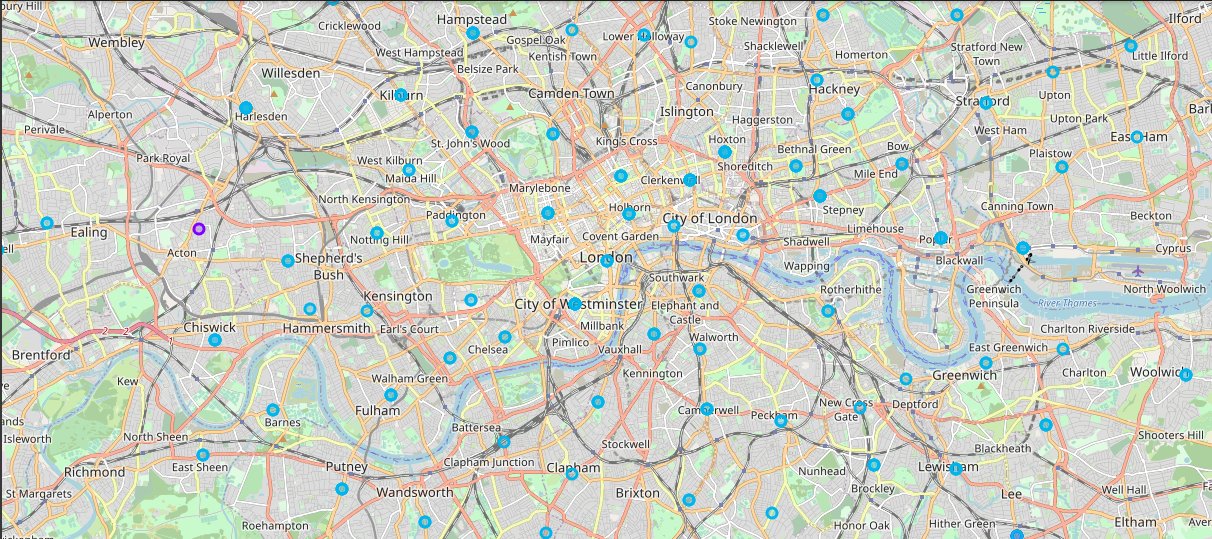




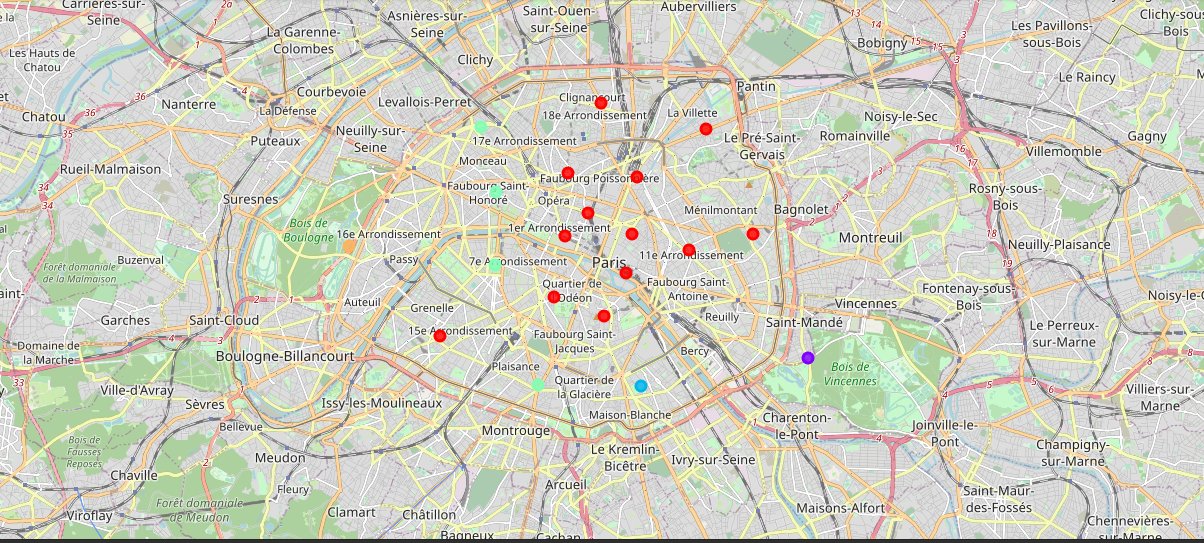
At the same time, missing data is collected and compiled. Although the Model is built. All that's remaining is to see the clustered neighborhoods on the map. We can use Folium package to do so. We need to drop all the NaN values to prevent data skew first.

屏幕快照 2020-11-14 下午9.56.14

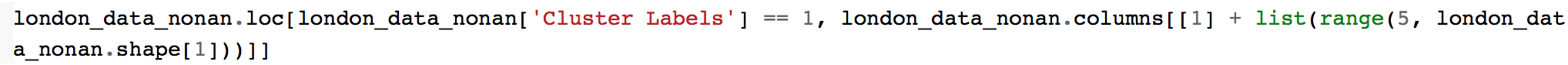
Map of clustered neighborhoods of London:



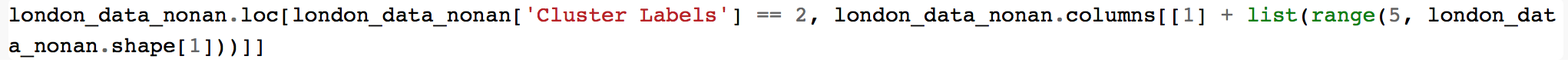
Map of clustered neighborhoods of Paris:



The last part of this section is to examine our clusters by expanding on our code using the Cluster Labels column:

Cluster1:

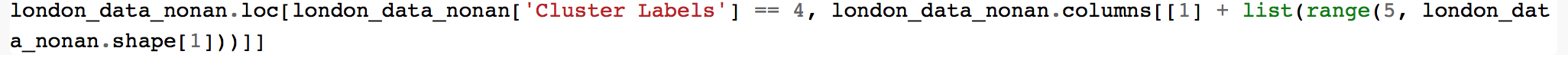
Cluster 2:



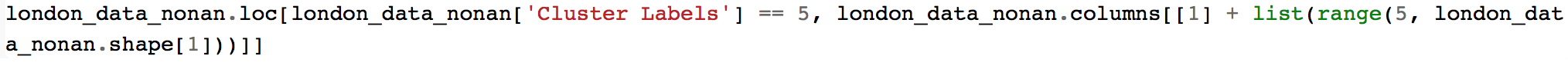
Cluster 3:



Cluster 4:



Cluster 5:



1. **Result and Discussion**

By comparing the data we obtained for two cities, we can see that the neighborhoods of London are very muliti-cultural. There are a lot of different cusines including Indian, Italian, Turkish and Chinese. Also, London has a lot of shopping options too, like markets, flower shops, fish markets, Fishing stores and clothing stores. The main modes of transport seem to be Buses and trains. For leisure, the neighborhoods 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 on the other hand, is relatively small. It has a wide variety of cuisine and eateries. 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 cuisine to try.

1. **Conclusion**

The purpose of this project is to analyze the cities of London and Paris, and to 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 neighborhoods, and finally concluding with clustering similar neighborhoods together.

We could see that each of the neighborhoods in both the cities have a wide variety of experiences to offer which is unique in it's 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 getaway with a lot of places to explore, there are beautiful landscapes, amazing foods and a wide variety of culture. These two cities both offer beautiful and impressive landscapes and neighborhoods for tourists to explore.