**The Battle of Neighborhood**

**Nikhil Garware**

**August 2020**

**1. Introduction :**

In this assignment we are going to study the dataset of New York City. We will convert the address in the form of latitude and longitude. After converting the location in the longitude and latitude form we will explore the neighborhood using the Foursquare API. The explore function would be used to explore the data and get the most common venue and sort the venues in the neighborhood according to the categories and their features. we will also use the K-Means algorithm to find the location nearby with categorical data and according to the features provides by the user review available form the Foursquare API.

**2. Problems which we tried to solve:**

1. Clustering the places according to their Features.
2. Plotting them on the map using the Matplotlib library
3. Making the Map available so that we can plot them easily.
4. Forming clusters and highlighting them.
5. Exploring location by making use of the explore function.

**3. Locality Selection Section :**

As the New-York city is world wide famous. So I decided to do data analysis on the same city. Many of the visitor visit the New-York city throughout the Year, so providing them the data useful is one of the most important factor. Making research and providing them the appropriate data depends totally on the data analysis. As the dataset of new-york city was available, i decided to Visualize the data of the same city.

**4. The Foursquare API:**

The foursquare API is a very well designed platform which is very useful for developers and data analysis. This API can provide a locality by tracking them by their longitude and latitude. By making use of this API we can easily plot the data clusters which may be used for data analysis and data visualization purpose. The API also provides a review of the customers through which we can select the location with good reviews and better quality of offerings .

**5. Libraries Used For the project.**

1. Pandas
2. Matplotlib
3. Scikit learn
4. Folium
5. Dependencies for K-Means Algorithm.

**6. Data acquisition and cleaning.**

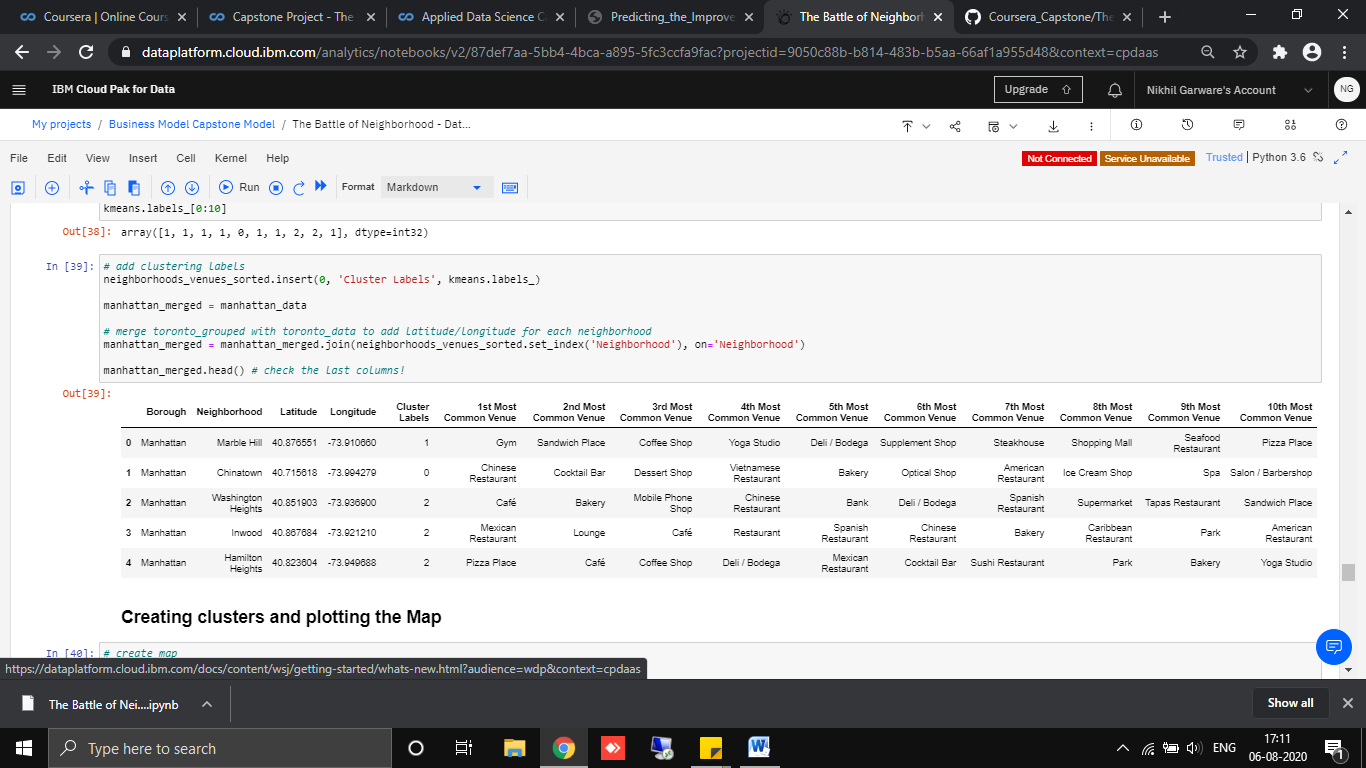
The dataset which we used in this assignment was the dataset of the new York city. The data set consists of many of the data like the longitude and latitude of the neighborhood. The data consists of the places and venues which are nearby to the location. Data downloaded or scraped from multiple sources were combined into one table. There were a lot of missing values from earlier seasons, because of lack of record keeping. There are several problems with the datasets but we managed to solve these problems and we made the data set more clean and valuable. The data needs to be cleaned because if there are some mistakes in the values then the dataset may generate wrong working result, which make conclude in generating wrong output. These output may totally ruin the purpose of data analysis. For the data analysis to be perfect we have to clean the data and then the data has to be sorted. After the sorting of the data is done then the main phase comes. In this phase we reduce all the redundant data and the data which is only valuable and may be considered for the data analysis phase is considered and taken for the data analysis phase. If the whole data is taken under consideration then there might arise a situation of data over fitting and the condition of data over fitting leads to generating incorrect values and incorrect results.

7 . **Methodology used for data analysis:**

The methodology used for the data analysis was quite simple & complex at the same time. We first collected the data. Then the data was cleaned, after cleaning the data the data was checked for data redundancy. All the redundant data was removed and then the final data was considered for data analysis and data processing. Only the data which was required was taken under consideration and the processed .

**8. Algorithm used for the data analysis:**

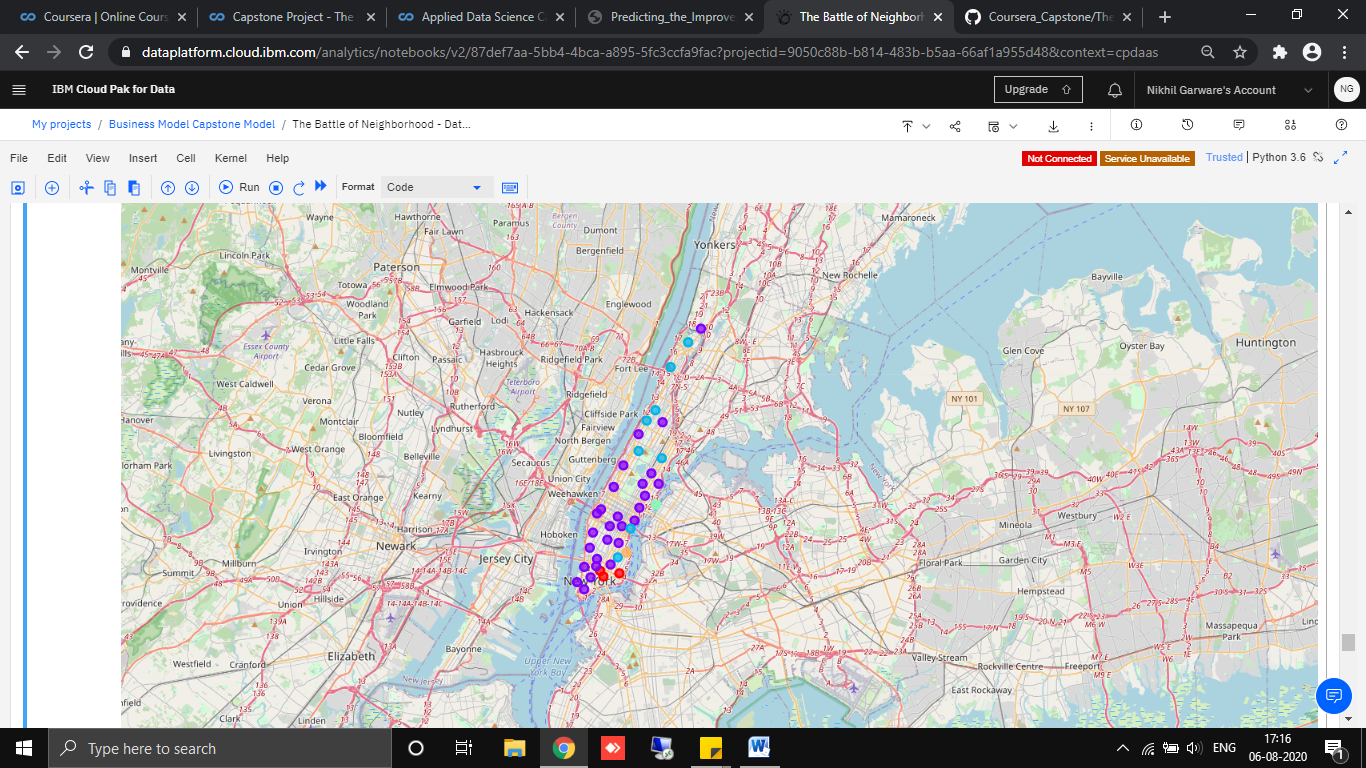
In this project we used the K-Means Algorithm for the processing of the data. The K-Means algorithm is a very powerful tool which is used for finding the nearest neighbourhood from the location we provided. Following data shows the details use of the K-Means Algorithm which we used in the project.



**Fig(1)**

The Above figure shows that we have create the clusters and then sorted the head of the clusters. By using the head function only the first data rows are shown in the output. When there is a large data set at that time this process plays an very important role. The clustering helps in sorting the various kinds of data in the various formats. Once the data is been sorted we can plot the data clusters on the map. These data cluster are shown In the form of blue indicators.

**Ploting the different clusters on the Map**



Fig(2)

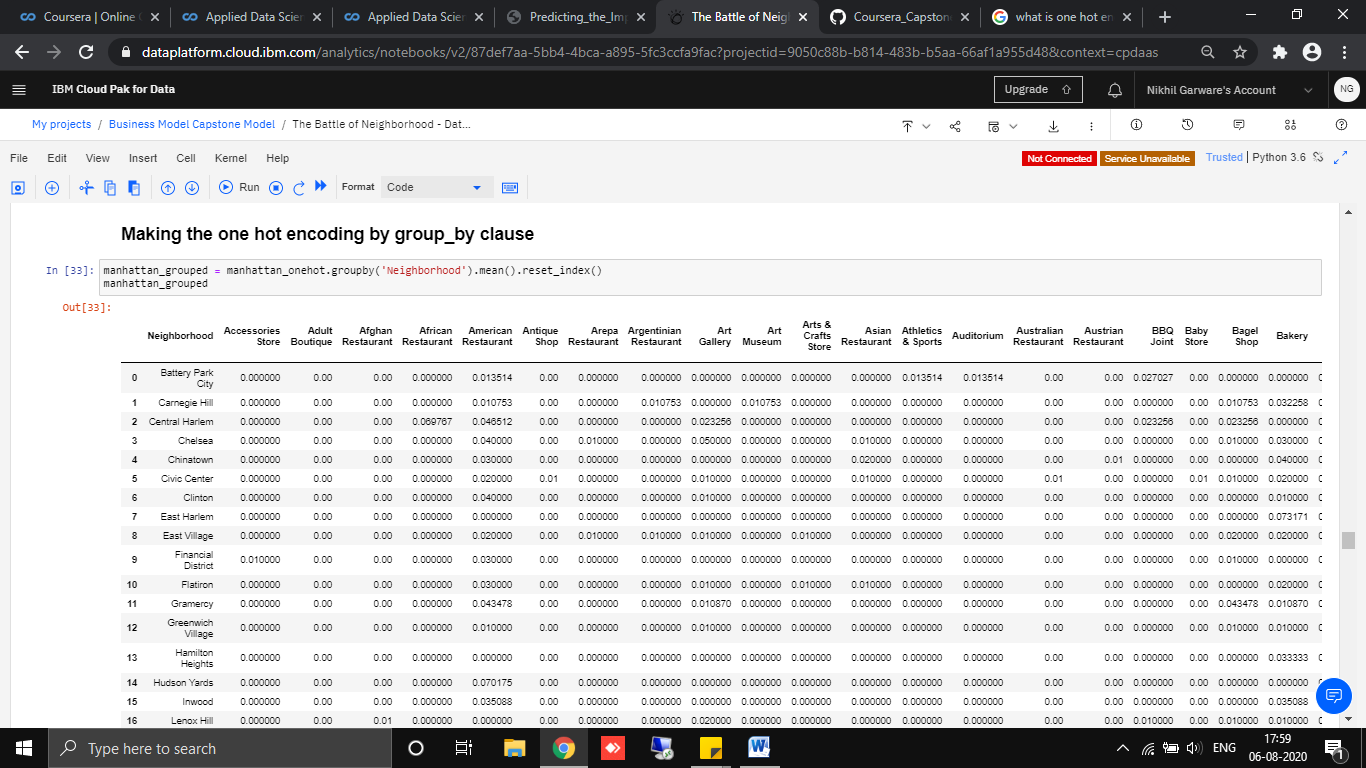
The above figure shows the clusters of each city. These clusters are the localities that are chosen according the neighbourhood and the features of the dataset. Every cluster has its unique value. This unique values depends of the features of those clusters.

Here these clusters represent different localities and these localities are the features which we have taken under consideration of the dataset.

**One Hot-Encoding:**

One hot encoding is a process by which categorical variables are converted into a form that could be provided to ML algorithms to do a better job in prediction. We have used the following technique to convert the data into the one hot encoding format. The one-hot encoding is very powerful and often used for the better understanding of the data and making the prediction well for a dataset.

The following figure shows the one-hot encoding technique and how we have implemented and what is the desired output of the commands that we have implemented on the system.



Fig(3)

**Conclusion:**

Thus we have studied and implemented the operation on the dataset of new-york city. By this data analysis operation we have developed an basic idea about the data of the neighbourhood in the New-York data set.