**CUISINE RECOMMENDATION**

**Introduction**

New York is one of the most popular cities in the United States. It is also one of the most densely populated cities in the United States. New York City is composed of five boroughs, each of which is a county of the State of New York.

The goal of the project is to identify the different cuisine restaurants that are frequently visited in the Manhattan city and its neighborhood by converting the addresses into their equivalent latitude and longitude coordinates. This analysis will help people who are planning to open a restaurant within Manhattan city and its neighborhood. By identifying the frequently visited restaurants, and analyzing the data, we can recommend that a new restaurant can be opened in a particular neighborhood where it will be received better. I am using Foursquare location data to explore the neighborhoods of Manhattan city. With the help of Foursquare API, I get the most common venue categories in each neighborhood and filter to choose the restaurants among those venues and then use those features to group the neighborhoods into clusters by utilizing the K-means clustering algorithm. With the help of Folium library, neighborhoods in Manhattan city can be visualized along with their emerging clusters.

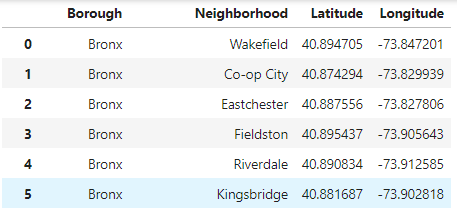
**Data**

New York dataset is in Json format. The dataset is available on the web for free access. I use ‘wget’ command to access the data. For this project I am using the Jupyter notebook from IBM Watson Cloud.

The New York dataset has total of 5 boroughs and 306 neighborhoods. The latitude and longitude coordinates of each neighborhood is needed to segment the neighborhoods and explore them. Among the five boroughs I am choosing the Manhattan borough since Manhattan is one of the busy places with most number of restaurants in New York when compared to other boroughs. It has the iconic sites like Empire State building, Times square and many other main tourist spots. My objective is to explore the different cuisines (e.g. American, Italian, Mexican, etc.,.) in that borough and neighborhood. The Foursquare API is used for obtaining the venues around the coordinates. Python is the primary language used for the analysis and by mainly utilizing a library called Pandas. Since the dataset is in JSON format, the values are in key-value format. Some of the main Keys are Features like ID, coordinates, name. I convert this Json file to pandas Dataframe for analysis.

**Data cleaning and formatting**

One important step is to make sure the data is cleaned and formatted in such a way that it can be utilized for analysis. The first step is to change the Json dataset into a pandas Dataframe and extracting the necessary columns needed. Feature selection was a major step as I needed to make sure the necessary columns were available for analysis and removed the columns not required.

*Fig 1: (Left) Screenshot of the JSON dataset (Right) Screenshot of the Json dataset converted into Dataframe*

For the next step, I used Geopy library to get the latitude and longitude values of New York city by using a module called Nominatim. Geopy library is a Python 2 and 3 clients for several popular geocoding web services. It makes it easy for locating the coordinates of addresses, cities, countries, and landmarks across the globe using third party geocoders and other data sources.

**Data Preprocessing**

My primary choice of location for analysis is the Manhattan city in New York. Foursquare API was utilized for getting the neighborhood venues using coordinates of Manhattan retrieved by the Geopy. Foursquare API can be accessed by using the unique developer’s credentials like Client\_ID and Client\_secret and a specific API link for the GET request. From the venues retrieved, I filtered the categories of the venues which are restaurants of different cuisine like American, Chinese, Indian, Italian, etc., and their name along with their coordinates for manipulation and analysis. I limited the venues to a radius of 5000.

**Exploratory Data analysis**

One-hot encoding technique is used to turn the categorical values into binary representation such as 1 or 0. I grouped the neighborhood based on the restaurants using the one hot encoding for getting a better data set for analysis purpose. The grouping is also done by finding the mean of the encoded categorical variable venues which gives us a dataset with the frequency of the restaurants in each neighborhood in Manhattan. With this information, we retrieve the first ten venues that have the highest mean.

For performing this analysis, I have implemented the K-Means clustering. K-means clustering is a process of partitioning n observations into k clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroids) serving as a prototype of the cluster. There are some common initialization methods like Forgy and Random Partition. The Forgy method randomly chooses k observations from the dataset and uses these as the initial means. The Random Partition method first randomly assigns a cluster to each observation and then proceeds to the update step thus computing the initial mean to be the center of the data set.

By using the Forgy and Random partition method, the best value for K was identified to be five. A set of 5 clusters are used for clustering. This is imported by using the scikit-learn library in python.

Finally, Folium maps are used to visualize data that has been manipulated in Python on an interactive leaflet map.

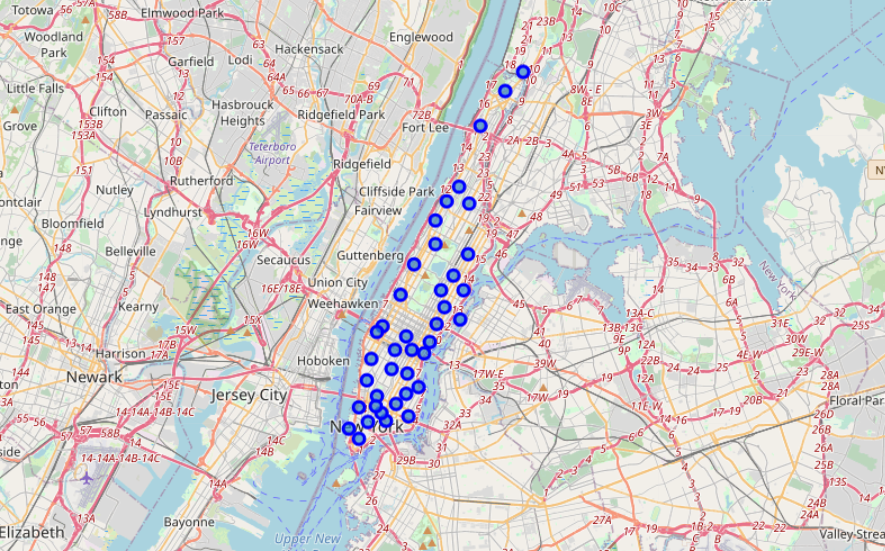
**Results**

Cluster 0 – It is the most populated cluster. Apparently Italian restaurants are the first common venue for this cluster.

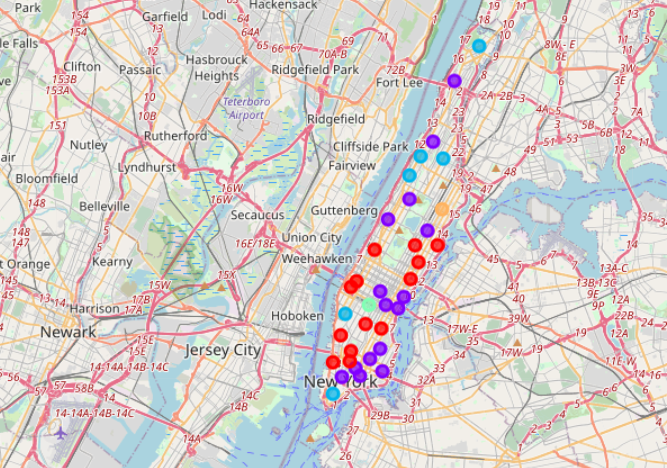
Cluster 1 and Cluster 4- It is very less populated and has just one neighborhood in the cluster.

Cluster 2 and Cluster 3 – It is highly populated with different variety of cuisines.

As Manhattan is a busy place with several famous landmarks and different tourists, we can see that the different cuisines like Italian, Chinese and Mexican are most common and have a higher frequency of being visited. The analysis recommends that it is best to try starting a different cuisine like Italian, Chinese or Mexican in one of the neighborhoods .



*Figure2: Folium map of the neighborhood in Manhattan city*



*Figure3: Clustered map after Analysis*



*Figure 4: Screenshot of the resulting Cluster1 data*

**References:**

1. <https://en.wikipedia.org/wiki/K-means_clustering>
2. <https://geopy.readthedocs.io/en/stable/>
3. <https://pandas.pydata.org/docs/>
4. <https://developer.foursquare.com/>
5. <https://python-visualization.github.io/folium/>