**Introduction**

New York City is the most populous city in the United States, home to the headquarters of the United Nations and an important center for international diplomacy. It just might be the most diverse city on the planet, as it is home to over 8.6 million people and over 800 languages. Undoubtedly, Food Diversity is an important part of an ethnically diverse metropolis. The idea of this project is to categorically segment the neighborhoods of New York City into major clusters and examine their cuisines. A desirable intention is to examine the neighborhood cluster's food habits and taste. Further examination might reveal if food has any relationship with the diversity of a neighborhood.This project will help to understand the diversity of a neighborhood by leveraging venue data from Foursquare’s ‘Places API’ and ‘k-means clustering’ machine learning algorithm.

**Problem**

The idea of this project is to categorically segment the neighborhoods of New York City into major clusters and examine their cuisines. Intention is to examine the neighborhood cluster’s food habits and taste. This project will help to understand the diversity of a neighborhood by leveraging venue data from Foursquare’s ‘Places API’ and ‘k-means clustering’ unsupervised machine learning algorithm. Exploratory Data Analysis (EDA) will help to discover further about the culture and diversity of the neighborhood.

**Data**

New York City Dataset

Link: https://geo.nyu.edu/catalog/nyu\_2451\_34572

Description: This New York City Neighborhood Names point file was created as a guide to New York City’s neighborhoods that appear on the web resource, “New York: A City of Neighborhoods.” Best estimates of label centroids were established at a 1:1,000 scale, but are ideally viewed at a 1:50,000 scale. This dataset will provide the addresses of neighborhood of NYC in json format.

Foursquare API:

Link: https://developer.foursquare.com/docs

Description: Foursquare API, a location data provider, will be used to make RESTful API calls to retrieve data about venues in different neighborhoods. This is the link to Foursquare Venue Category Hierarchy. Venues retrieved from all the neighborhoods are categorized broadly into "Arts & Entertainment", "College & University", "Event", "Food", "Nightlife Spot", "Outdoors & Recreation", etc

**Methodology**

**Download and Explore New York City Dataset**

Once the .json file is downloaded, it is analyzed to understand the structure of the file. A python dictionary is returned by the URL and all the relevant data is found to be in the features key, which is basically a list of the neighborhoods. The dictionary is transformed, into a pandas dataframe, by looping through the data and filling the dataframe rows one at a time.

As a result, a dataframe is created with Borough, Neighborhood, Latitude and Longitude details of the New York City’s neighborhood

Further, ‘geopy’ library is used to get the latitude and longitude values of New York City, which was returned to be Latitude: 40.71, Longitude: -74.01.

The curated dataframe is then used to visualize by creating a map of New York City with neighborhoods superimposed on top. The following depiction is a map generated using python ‘folium’ library.

**RESTful API Calls to Foursquare**

The Foursquare API is used to explore the neighborhoods and segment them. To access the API, ‘CLIENT\_ID’, ‘CLIENT\_SECRET’ and ‘VERSION’ is defined.

There are many endpoints available on Foursquare for various GET requests. But, to explore the cuisines, it is required that all the venues extracted are from ‘Food’ category. Foursquare Venue Category Hierarchy is retrieved and returned request is further analyzed.

Upon analysis, it is found that there are 10 major or parent categories of venues, under which all the other sub-categories are included.

As said earlier, the ‘FOOD’ category is the matter of interest. A function is created to return a dictionary with ‘Category ID’ & ‘Category Name’ of ‘Food’ & it’s sub-categories.

**Pickle**

Pickle is a very important and easy-to-use library. It is used to serialize the information retrieved from GET requests, to make a persistent ‘.pkl’ file. This file can later be deserialized to retrieve an exact python object structure. This is a crucial step as it will counter any redundant requests to the Foursquare API, which is chargeable over the threshold limits.

**Exploratory Data Analysis**

The merged dataframe ‘nyc\_venues’ has all the required information

**Data Cleaning**

It is crucial to understand that the point of interest in the project is to understand the cultural diversity of a neighborhood by clustering it categorically, using the venues’ categories. Thus, it is important to remove all the venues from the ‘dataframe’ which have generalized categories. Here, by generalized, it means that these categorized venues are common across different cultures and food habits. Examples of categories of this type of venues are Coffee Shop, Cafe, etc. So, firstly all the unique categories are fed into a python ‘list’.

**Feature Engineering**

Now, each neighborhood is analyzed individually to understand the most common cuisine being served within its 500 meters of the vicinity.

The above process is taken forth by using ‘one hot encoding’ function of python ‘pandas’ library. One hot encoding converts the categorical variables (which are ‘Venue Category’) into a form that could be provided to ML algorithms to do a better job in prediction.

**Machine Learning**

‘k-means’ is an unsupervised machine learning algorithm which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

To implement this algorithm, it is very important to determine the optimal number of clusters (i.e. k). There are 2 most popular methods for the same, namely ‘The Elbow Method’ and ‘The Silhouette Method’.

**The Elbow Method**

The Elbow Method calculates the sum of squared distances of samples to their closest cluster center for different values of ‘k’. The optimal number of clusters is the value after which there is no significant decrease in the sum of squared distances.

**The Silhouette Method**

The following is an implementation of this method. As it requires minimum 2 clusters to define dissimilarity number of clusters (i.e. ‘k’)

**k-Means**

Following code block runs the k-Means algorithm with the number of clusters

**Results**

**Cluster — 0**

Chinese is dominant

**Cluster — 1**

Following are the results of the Cluster — 1 analysis:

Fast Food Restaurant 26

Pizza Place 14

**Cluster — 2**

Following are the results of the Cluster — 2 analysis:

Cluster 3

Italian Restaurant 33

Italian is dominant

Follow the rest of analysises in Notebook

**Conclusion**

One application of *Clustering Algorithm*, k-Means or others, to a multi-dimensional dataset, a very inquisitive result can be curated which helps to understand and visualize the data. The neighborhoods of New York City were very briefly segmented into eight clusters and upon analysis, it was possible to rename them basis upon the categories of venues in and around that neighborhood. Along with American cuisine, Italian and Chinese are very dominant in New York City and so is the diversity statistics.