**Exploring the Taste of NYC Neighborhoods**

Machine Learning

Eklavya Saxena

October 21, 2019

# Introduction

## Background

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.

As quoted in an article - [What Food Tells Us About Culture](https://freelymagazine.com/2017/01/07/what-food-tells-us-about-culture/)  
“Traditional cuisine is passed down from one generation to the next. It also operates as an expression of cultural identity. Immigrants bring the food of their countries with them wherever they go and cooking traditional food is a way of preserving their culture when they move to new places.”

## Problem

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. Exploratory Data Analysis (EDA) will help to discover further about the culture and diversity of the neighborhood.

## **Stakeholders**

This quantifiable analysis can be used to understand the distribution of different cultures and cuisines over ‘the most diverse city on the planet – New York City’. Also, this project can be utilized by a new food vendor who is willing to open his or her restaurant. Or by a government authority to examine and study their city's culture diversity better.

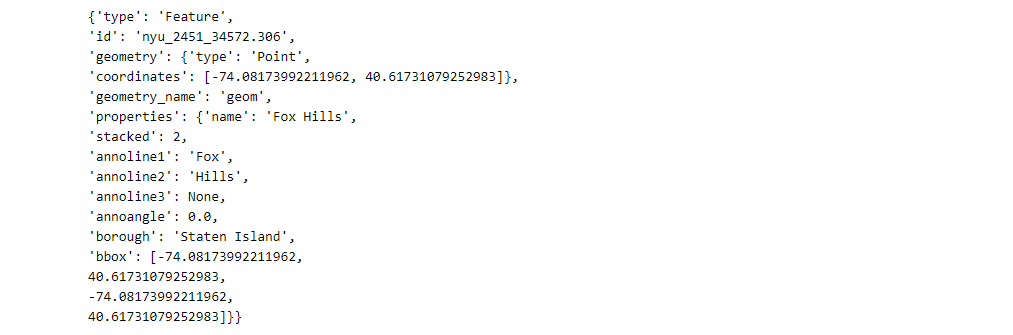
# Data

To examine the above said, following data sources will be used:

## New York City Dataset

Link: <https://geo.nyu.edu/catalog/nyu_2451_34572>

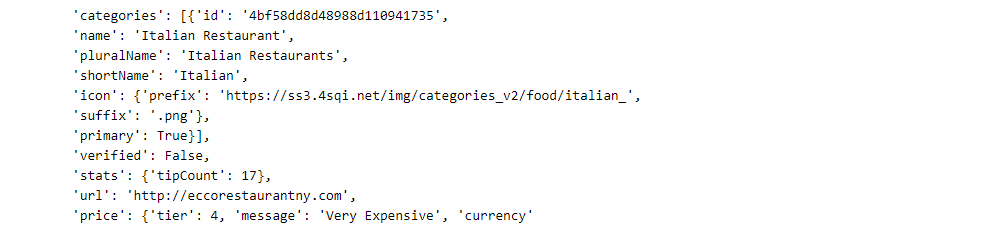
This New York City Neighborhood Names point file was created as a guide to New York City’s neighborhoods that appear on the web resource, ‘New York: A City of Neighborhoods.’ Best estimates of label centroids were established at a 1:1,000 scale, but are ideally viewed at a 1:50,000 scale. This dataset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



## Foursquare API

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

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](https://developer.foursquare.com/docs/resources/categories). Venues retrieved from all the neighborhoods are categorized broadly into ‘Arts & Entertainment’, ‘College & University’, ‘Event’, ‘Food’, ‘Nightlife Spot’, ‘Outdoors & Recreation’, etc. An extract of an API call is as follows:

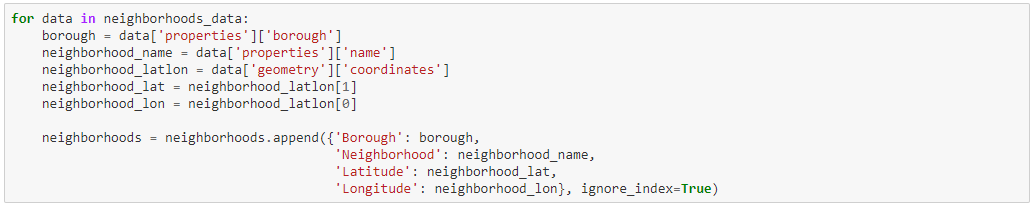


# Methodology

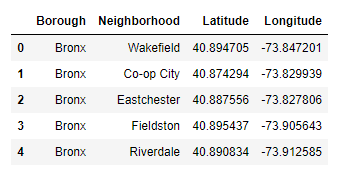
## Download and Explore New York City Dataset

In order to segment the neighborhoods of New York City, a dataset is required that contains the 5 boroughs and the neighborhoods, that exist in each borough, with respective latitude and longitude coordinates. This dataset is downloaded using the mentioned URL.

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 using the following depicted loop.



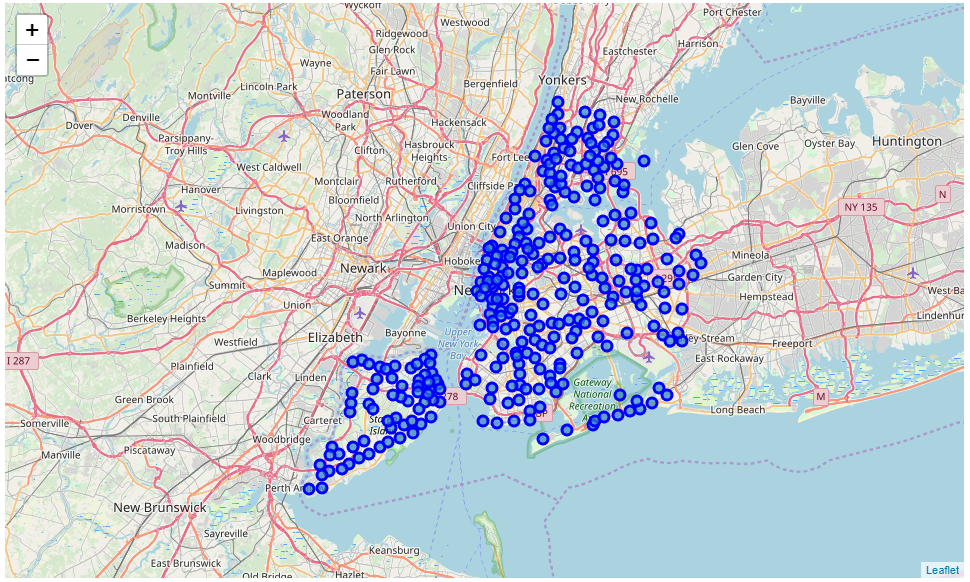
As a result, following dataframe is created with Borough, Neighborhood, Latitude and Longitude details of the New York City’s neighborhood.



Upon analysis, it is found that the dataframe consists of 5 boroughs and 306 neighborhoods.

Further, ‘geopy’ library is used to get the latitude and longitude values of New York City, which was returned to be Latitude: 40.7127281, Longitude: -74.0060152.

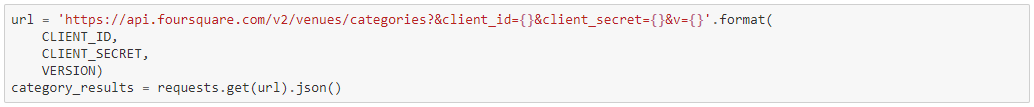
The curated dataframe is then used to visualize by creating a map of New York City with neighborhoods superimposed on top. The following depiction is a map generated using python ‘folium’ library.



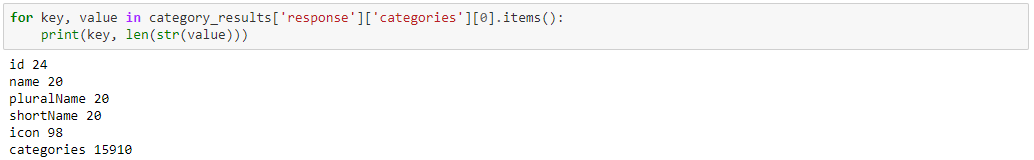
## RESTful API Calls to Foursquare

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

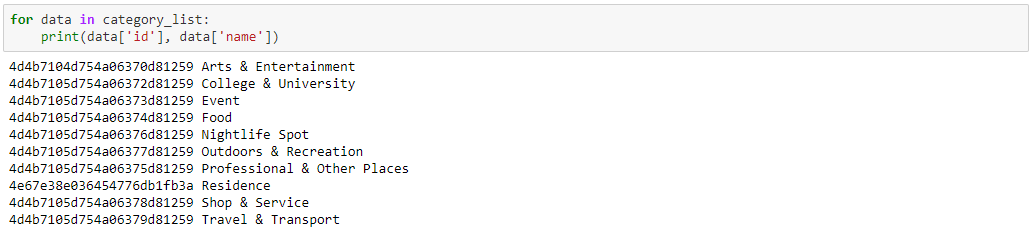
There are many endpoints available on Foursquare for various GET requests. But, to explore the cuisines, it is required that all the venues extracted are from ‘Food’ category. Foursquare Venue Category Hierarchy is retrieved using following code block:



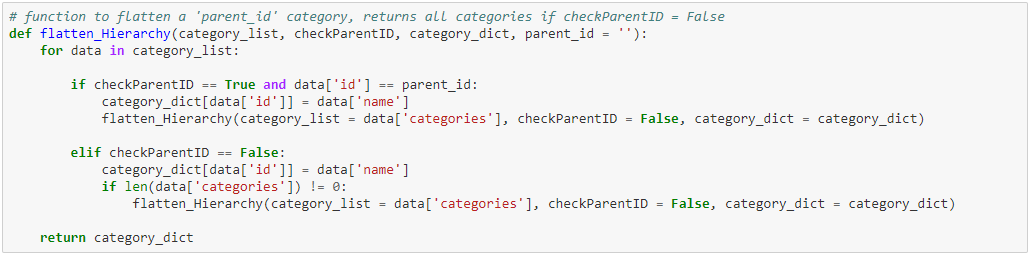
The returned request is further analyzed:



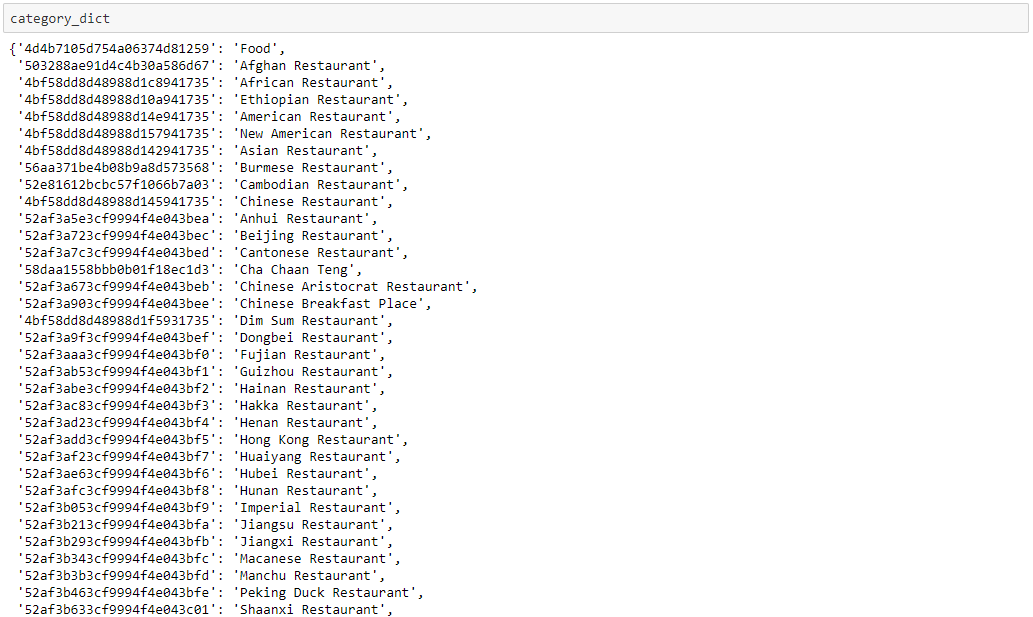
Upon analysis, it is found that there are 10 major or parent categories of venues, under which all the other sub-categories are included. Following depiction shows the ‘Category ID’ and ‘Category Name’ retrieved from API:



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



This above function takes the parent ‘Category ID’ and returns the ‘Category Name’ and ‘Category ID’ of all the sub-categories.



To further understand the results of GET Request, the first neighborhood of the ‘New York City’ dataset is explored. The first neighborhood returned is ‘Wakefield’ with Latitude 40.89 and Longitude -73.85.

Then, a GET request URL is created to search for Venue with ‘Category ID’ = '4d4b7105d754a06374d81259', which is the ‘Category ID’ for ‘Food’, and radius = 500 meters.



The returned request is then examined, which results in following:

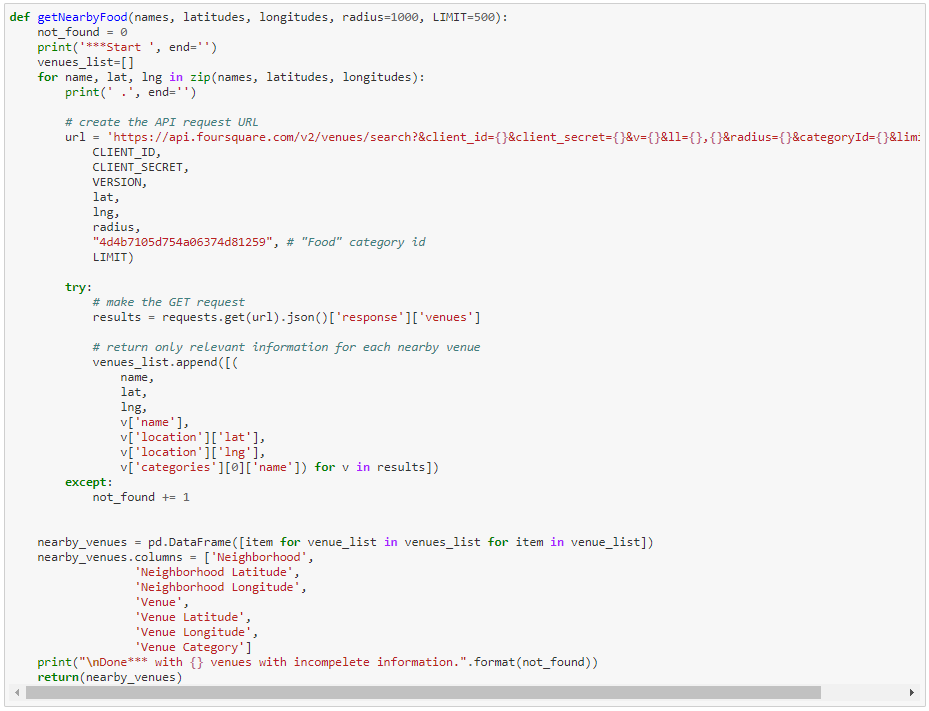


The request returned the ‘Category Name’ of the venue as 'Carvel Ice Cream' is of 'Food' category.

As, the aim is to segment the neighborhoods of New York City with respect to the ‘Food’ in its vicinity, it is further required to fetch this data from all the 306 neighborhoods' venues.

To overcome the redundancy of the process followed above, a function ‘getNearbyFood’ is created. This functions loop through all the neighborhoods of New York City and creates an API request URL with radius = 500, LIMIT = 100. By limit, it is defined that maximum 100 nearby venues should be returned. Further, the GET request is made to Foursquare API and only relevant information for each nearby venue is extracted from it. The data is then appended to a python ‘list’. Lastly the python ‘list’ is unfolded or flattened to append it to dataframe being returned by the function.

It is inquisitive that Foursquare API returns all the sub-categories, if a top-level category is specified in the GET Request.



## 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.



The returned ‘dataframe’ is as follows:

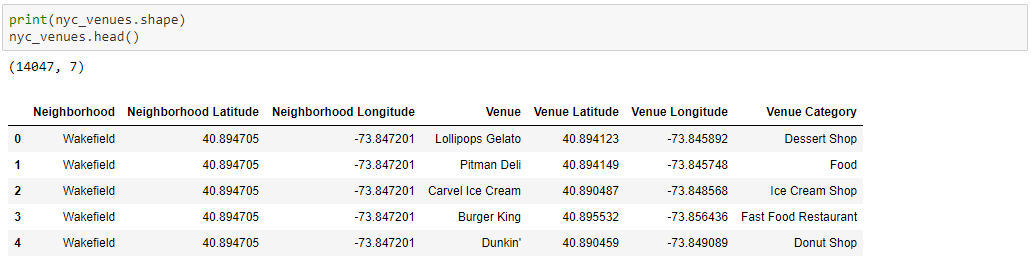


As of now, two python ‘dataframe’ are created:

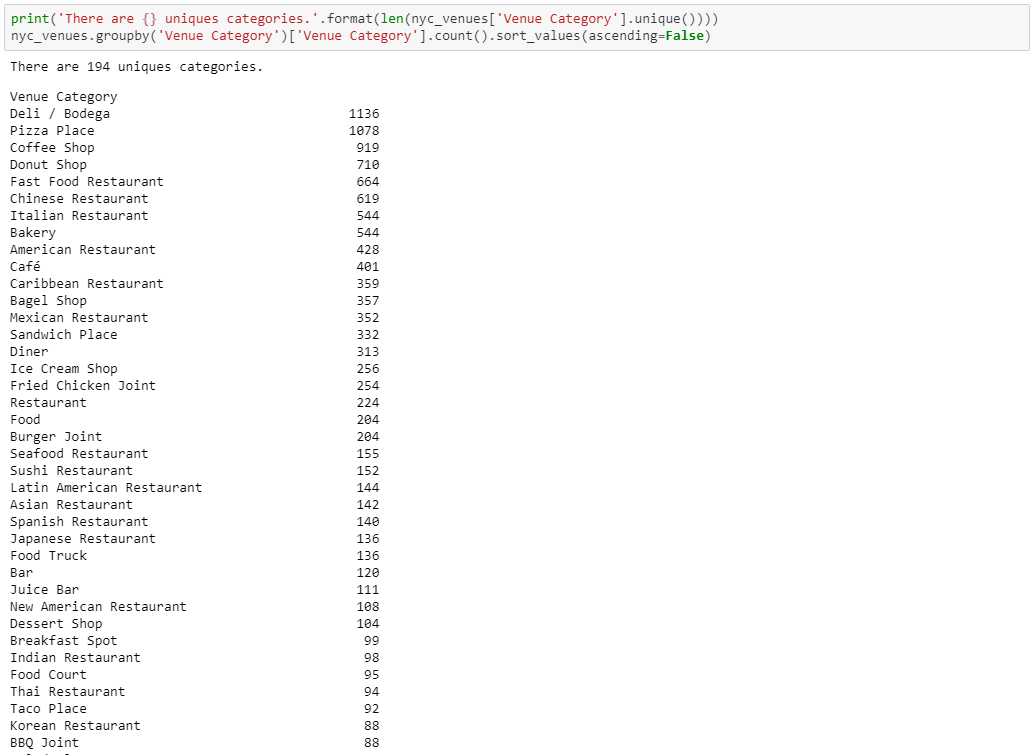
1. ‘neighborhoods’ which contains the Borough, Neighborhood, Latitude and Longitude details of the New York City’s neighborhood, and
2. ‘nyc\_venues’ which is a merger between ‘neighborhoods’ dataframe and its ‘Food’ category venues searched with ‘Radius’ = 500 meters and ‘Limit’ = 100. Also each venue has its own Latitude, Longitude and Category.

# Exploratory Data Analysis

The merged dataframe ‘nyc\_venues’ has all the required information. The size of this dataframe is determined, and it is found that there are total 14,047 venues.



Now, it is important to find out that how many unique categories can be curated from all the returned venues. There are 194 such categories, with most occurring venues as follows:



## 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. Example of categories of these types of venues are Coffee Shop, Cafe, etc.

So, firstly all the unique categories are fed into a python ‘list’.



Then, manually the categories are determined to be ‘general’ (as explained above). This data pre-preparation totally depends upon the ‘Data Analyst’ discretion and can be modified as required. Following are the categories listed as ‘general’:

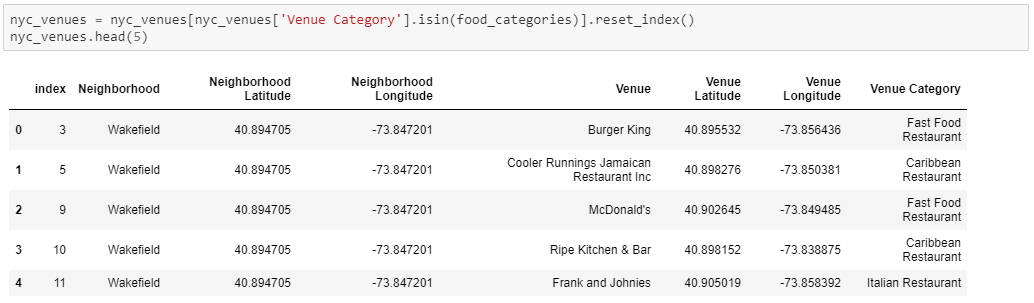


A simple subtraction of two python ‘list’ i.e ‘unique\_categories’ and ‘general\_categories’ gives a ‘list’ of all the categories which are required for further analysis.

Following image depicts the result of the above activity:



The python ‘list’ curated above, is used to remove all the venues with categories not in ‘food\_categories’, and the following dataframe is retrieved:



Again, the number of unique categories is examined, and it is found that there are only 92 of them, as compared to 194 earlier. That means, almost 50% of the data was a noise for the analysis. This essential step, data cleaning, helped to capture the data points of interest.

## Feature Engineering

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

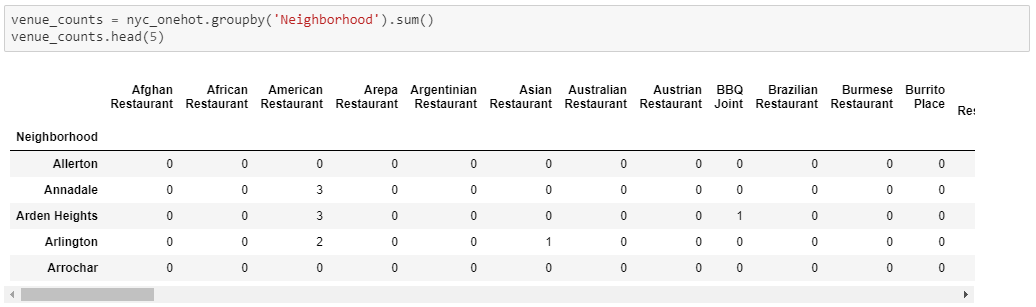


Upon converting the categorical variables, as shown above, ‘Neighborhood’ column is added back which results into the following:



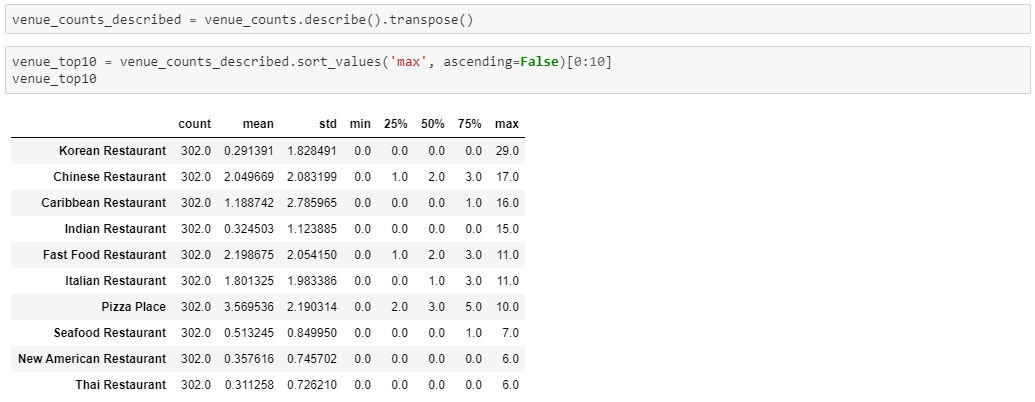
The size of the new dataframe ‘nyc\_onehot’ is examined and it is found that there are around 6,846 data points all together.

Further, number of venues of each category in each neighborhood are counted.



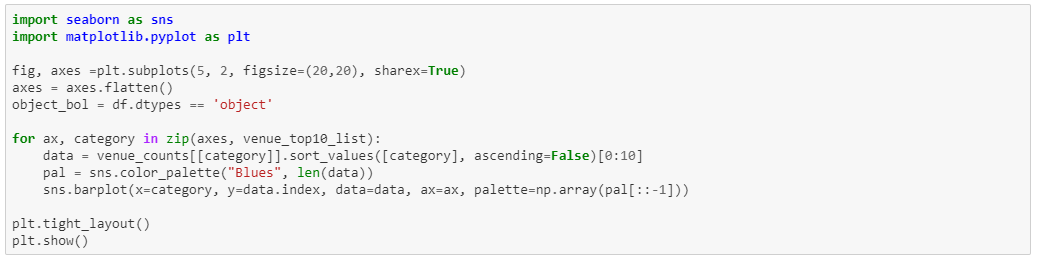
It is observed that, in the first five neighborhoods of the dataframe, ‘Annadale’, ‘Arden Heights’ and ‘Arlington’ has 3, 3, 2 ‘American Restaurant’ in its 500 meters vicinity.

The top 10 ‘Venue Categories’ can also be found by counting their occurrences. This analysis is depicted below which shows that ‘Korean Restaurant’, ‘Chinese Restaurant’, ‘Caribbean Restaurant’, ‘Indian Restaurant’, and ‘Fast Food Restaurant’ are among the top 5.

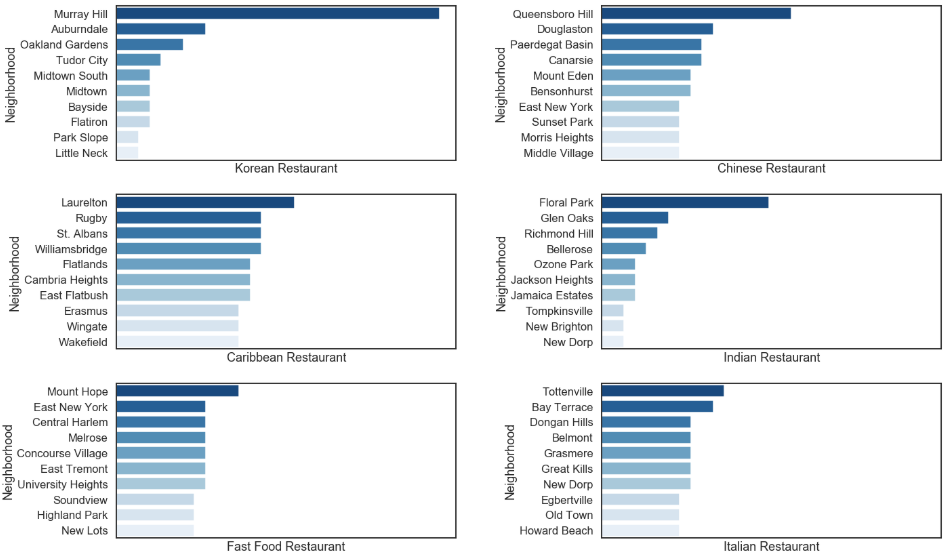


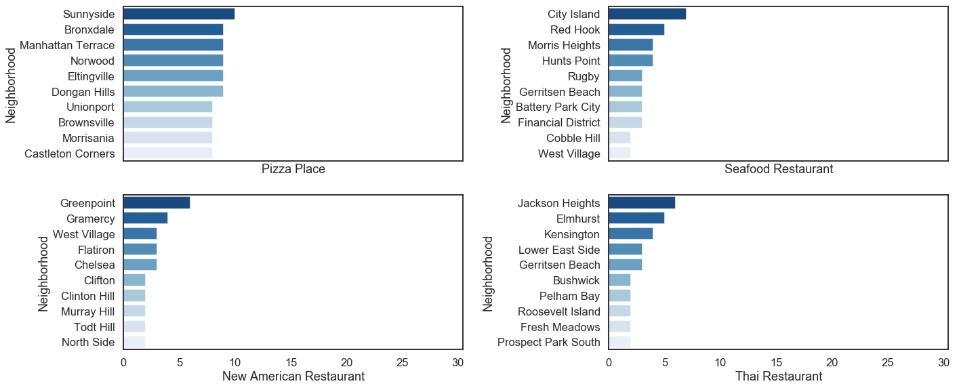
## Data Visualization

These top 10 categories are further plotted individually on bar graph using python ‘seaborn’ library. The following code block creates the graph of top 10 neighborhoods for a category.



The result of the above code block returns the following bar graphs:





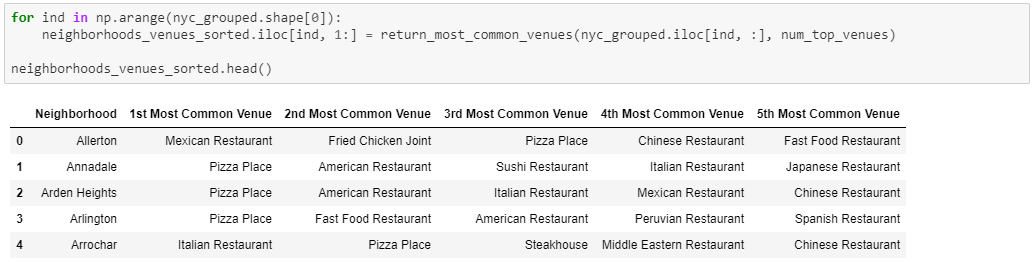
Next, the rows of the neighborhood are grouped together and the frequency of occurrence of each category is calculated by taking the mean.



As the limit is set to be 100, there will be many venues being returned by the Foursquare API. But a neighborhood food habit can be defined by the top 5 venues in its vicinity. Following ‘for’ loop creates a dataframe to record the above-mentioned data points:



Further, the above created dataframe is fed with the top 5 most common venues categories in the respective neighborhood.



# Machine Learning

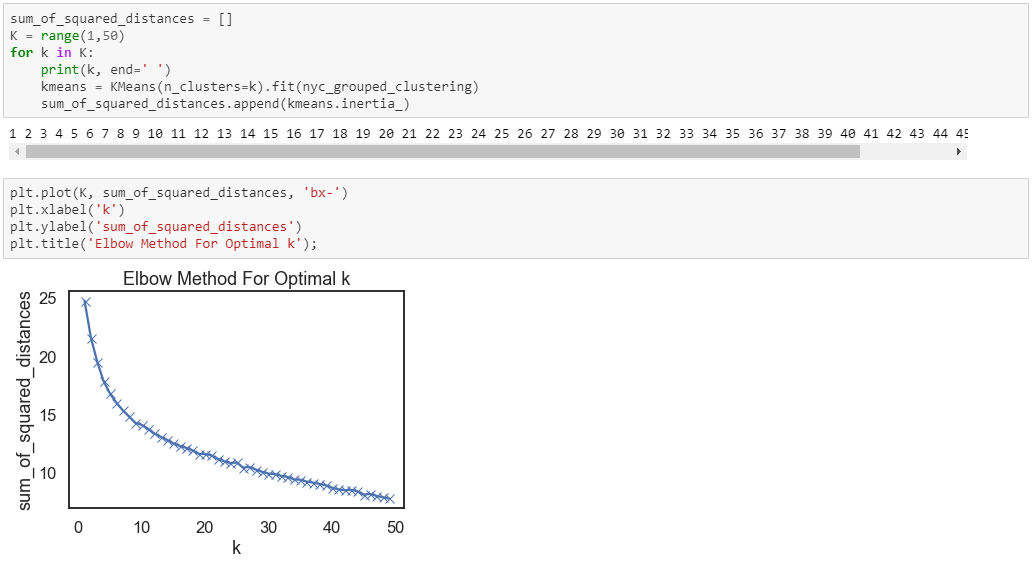
‘k-means’ is an unsupervised machine learning algorithm which creates clusters of data points aggregated together because of certain similarities. This algorithm will be used to count neighborhoods for each cluster label for variable cluster size.

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.

Following is an implementation of this method (with varying number of clusters from 1 to 49):

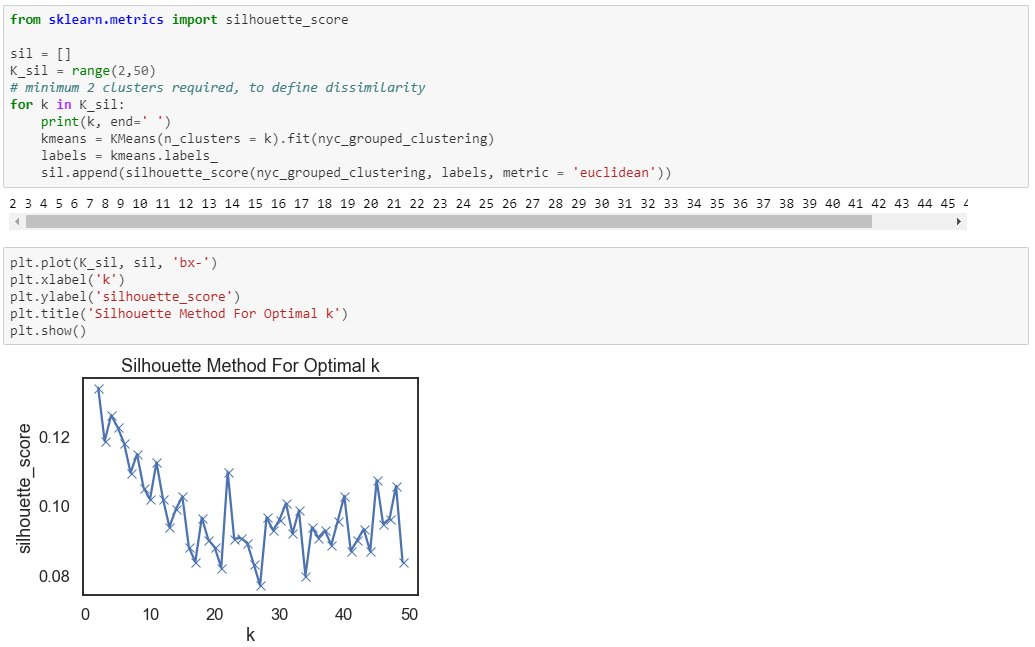


Sometimes, Elbow method does not give the required result, which happened in this case. As, there is a gradual decrease in the sum of squared distances, optimal number of clusters can not be determined. To counter this, another method can be implemented, as discussed below.

## The Silhouette Method

As quoted in Wikipedia – “The Silhouette Method measures how similar a point is to its own cluster (cohesion) compared to other clusters (separation).”

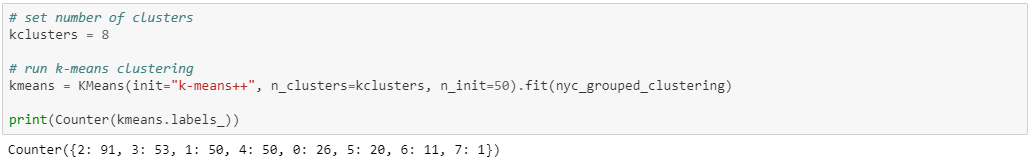
Following is an implementation of this method. As it requires minimum 2 clusters to define dissimilarity number of clusters (i.e. ‘k’) will vary from 2 to 49):



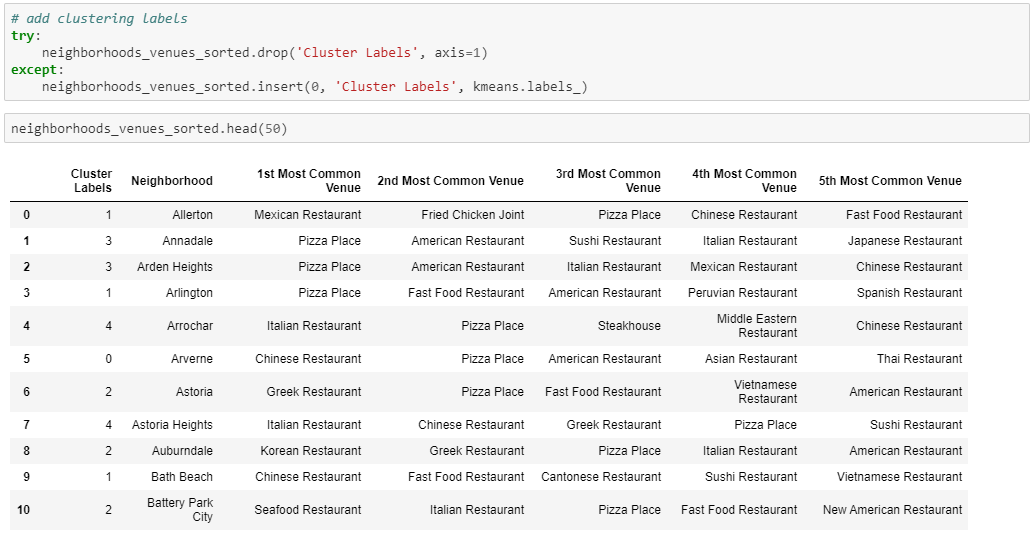
There is a peak at k = 2, k = 4 and k = 8. Two and four number of clusters will cluster the neighborhoods very broadly. Therefore, number of clusters (i.e. ‘k’) is chosen to be eight.

## k-Means

Following code block runs the k-Means algorithm with number of clusters = 8 and prints the counts of neighborhoods assigned to different clusters:



Further the cluster labels curated are added to the dataframe to get the desired results of segmenting the neighborhood based upon the most common venues in its vicinity:



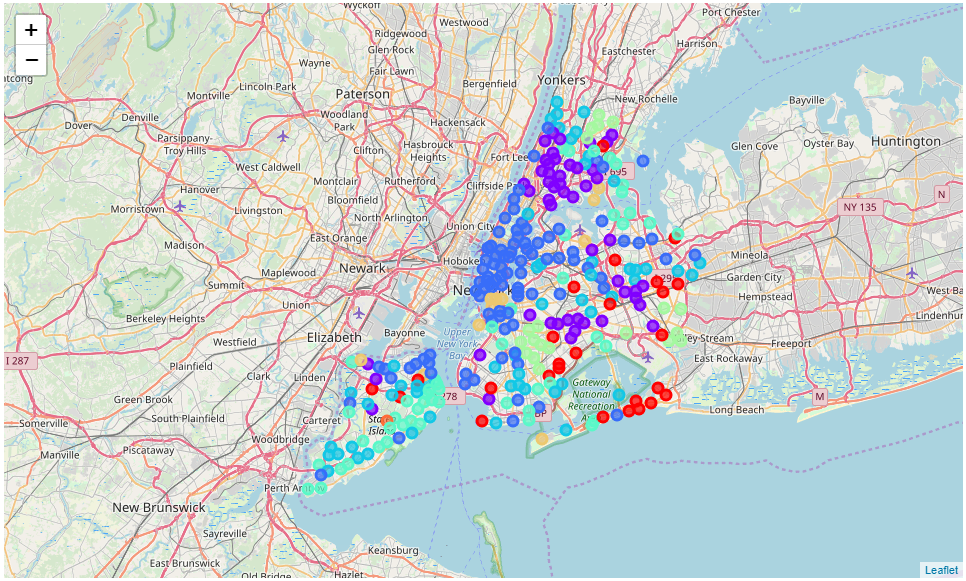
Now, ‘neighborhoods\_venues\_sorted’ is merged with ‘nyc\_data’ to add the Borough, Latitude and Longitude for each neighborhood.



Again, the New York City’s neighborhoods are visualized by using the code block as shown, which utilizes the python ‘folium’ library.



Following map is generated which show the desired segmentation of the New York City’s neighborhoods:



# Results