2. Data acquisition and Data wrangling

2.1 Data Acquisition

In this data science project we collect the data from three various data sources:

- New York City dataset will contains detail of the information about neighborhoods and boroughs. We can get it from the open data source: https://cocl.us/new_york_dataset.

- We need Latitude and longitude coordinates of those neighborhoods. This is required to plot the map and get the venue data. We will get the geographical coordinates of the neighborhoods (latitude and longitude) using Python Geocoder package.

- We acquired Venue data, particularly data related to restaurants. We are going to use this data to perform further analysis of the neighborhoods and boroughs. We will get this data by working with Foursquare API.

The data preparation and wrangling for each of the three sources of data is done separately in the project.

Open source New York City dataset contains the neighborhoods and boroughs. But in the project we should for visually map. We can get the geographical coordinates of the neighborhoods (latitude and longitude) by using Python Geocoder package.

Finally, we will use Foursquare API to get the venue data for the neighborhoods defined at the previous step. Foursquare has one of the largest databases of 105+ million places and over 125,000 developers use this application. Foursquare API provides many categories of the venue data; we are particularly interested in the restaurant data to solve the business problem defined above.

This project will require to using of many data science skills, such like web scrapping from open source dataset, working knowledge of Foursquare API, data cleaning, data wrangling on the many format of dataset, understanding the folium package for map visualization.

In the next section, we will discuss and describe any exploratory data analysis that we did, any inferential statistical testing that we performed, and what machine learning techniques were used to get the solutions for the our data science project problem.

2.2 Data Wrangling / Cleaning

From open data source: https://cocl.us/new_york_dataset

We extract the data in the tabular format by using geo location function of the geo coder package we can see will get in the form tabular format as shown in the in the following fig.

Out[5]:

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

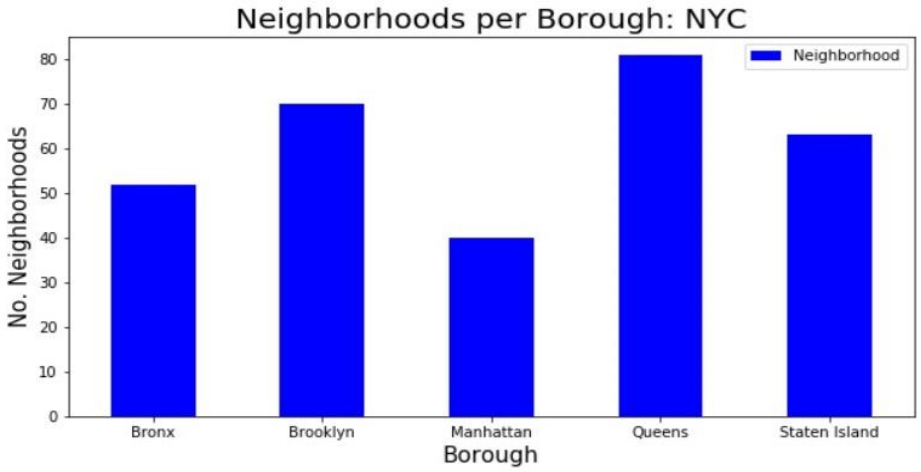
Initial Dataset get from sources

Dimension of Data

```
In [6]: new_york_data.shape
Out[6]: (306, 4)
```

There are total of 306 different Neighborhoods in New York.

Then we use the matplotlib package in the python language to plot the graph of data to visualize the dataset in the bar chart format so audience can easily what they see and what they can get idea about data science problem and the solutions. The following fig show the bar chart of the output



By using Foursquare API on the dataset we obtained from the open source, we can get the venues in the New York City neighborhoods such like Indian Restaurant information shown in the following fig.

Out[14]:

	Unnamed: 0	Borough	Neighborhood	ID	Name
0	0	Bronx	Woodlawn	4c0448d9310fc9b6bf1dc761	Curry Spot
1	1	Bronx	Unionport	4c194631838020a13e78e561	Melanies Roti Bar And Grill
2	2	Brooklyn	Bay Ridge	545835a1498e820edc6f636f	Bombay Grill
3	3	Brooklyn	Greenpoint	51a5445c498ee0f182370cb2	Agra Taj Mahal
4	4	Brooklyn	Bushwick	5169f5c4e4b0c7fcb77a0f3c	Agra Heights

```
In [15]: indian_ny.shape
Out[15]: (42, 5)
```

There are 42 Indian Restaurants in the New York City

As like shown above we use the various data science tools to acquiring, cleaning, wrangling the datasets to preparation for the analyzing the data. We will explain in the next section of this data science report.