# Introduction

Touring a big city such as the New York City in the United States of America is definitely a must do for any tourist visiting the state from abroad. However, for tourists on a limited time vacation, this can be a little bit stressful given how huge the city is, and the relatively little time to have to make the most out of this experience. Luckily, the field of Data Science is here to make our lives, and most importantly the tourist’s, a lot easier.

In this project, we will mainly assist a hypothetical tourist trying to effectively select the best places to visit given the little time they have. This will save the extensive research and decision-making time especially for indecisive tourists! We will adopt a structured top-down analysis, starting with the selection of the best NYC Borough, to the identification of the top venues of the potential neighborhoods to be selected.

The report will be structured as follows:

First, the topic and the problem are introduced in the introduction section of this report. Second, the data that will be used in this project and the source of the data will be described in the Data section. Third, we will elaborate the methodology adopted in this project in the methodology section. In this section, we will discuss and describe the exploratory data analysis that we did, and what machine learnings we used and why. Fourth, we will present the obtained results, followed by a discussion section. And finally, we will conclude our findings.

# Data

## Data Background

The project extensively relies on the dataset of the New York City made available by the course instructor in the labs (Week 3 of the capstone project course). The Data is first obtained as a JSON file that was later loaded and read into a Pandas Dataframe. Moreover, critical data about the venues in each neighborhood in the NYC will be retrieved from the FourSquare API.

The acquired data will be manipulated via exploratory data analysis tools such as pandas library tools, and will be visualized using matplotlib (for plots) and folium (for maps).

The data is to be cleaned and filtered to acquire only the necessary information and then used and manipulated for a clear data analysis.

This of course will necessitate importing the necessary libraries and dependencies which will be elaborated in the methodology section of this report.

## **Data** Content

The data used in this project and acquired from the available NYC Dataset on the IBM server mainly consist of the different neighborhoods of the city, categorized under 5 main Boroughs, along with their corresponding latitude and longitude. There are 306 neighborhoods in the City of New York, distributed over 5 Boroughs, as we will see in the next sections of this report.

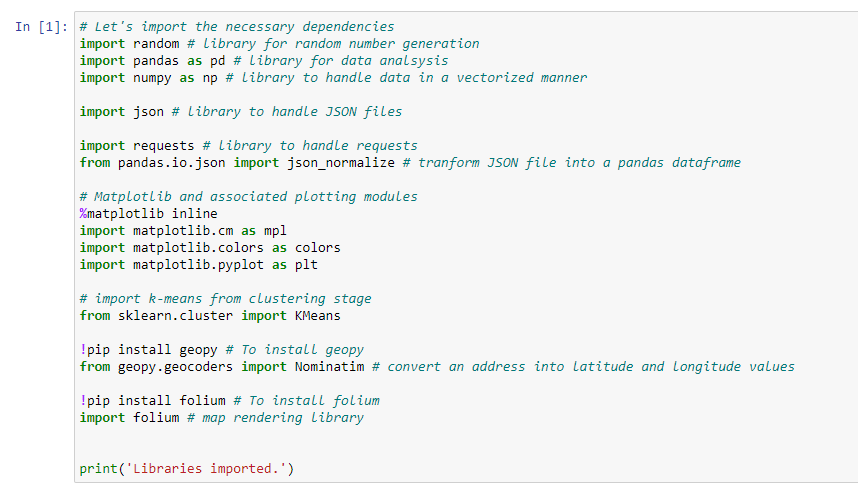
With respect to the data obtained from the Foursquare API, it mainly comprises the different venues in each neighborhood of each borough in the city, along with the venue’s coordinates, name, and category.

The obtained data is extremely rich in content which will allow for an extensive data analysis.

# Methodology

## Importing Libraries

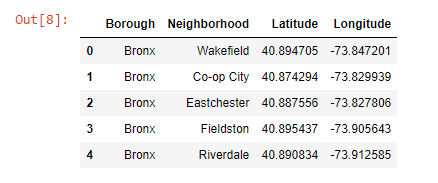
Before going about retrieving, cleaning and manipulating our data, it is necessary to import the right tools and libraries that will allow us to do so. Throughout this project, the main libraries used were *pandas* data frame analyses, *matplotlib* for data visualization, *geopy* for geographical necessities, and folium for map generations. Other libraries were used as well and are outlined in the following figure showing the executed python code. For more details, please refer to the enclosed jupyter notebook accompanying this report.



## Retrieving the necessary NYC data

The next step would be retrieving the data using the imported tools and installed packages. The dataset corresponding to the neighborhoods of the city of New York was made available for use by the course instructor on the servers. The dataset is in JSON format, and the JSON library tools were imported to be able to download the file, and read it.

The dataset was then filtered and only the necessary features were selected and translated into a *pandas* data frame. The below figure demonstrates the first five rows of the created data frame.



We can see that the selected features are the Boroughs of NYC, along with their corresponding Neighborhoods and their coordinates. These are key features that were used in the data analysis.

The dataframe indicates that the City of New York comprises a total of 5 boroughs and 306 neighborhoods.

## Generating the map of New York

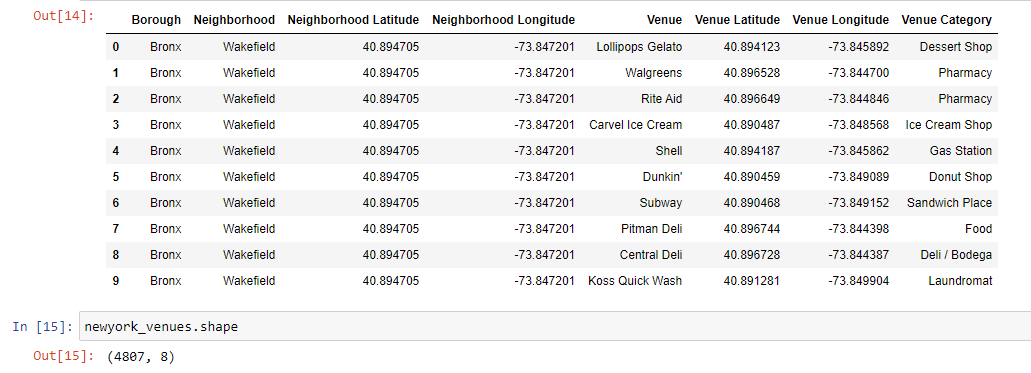
The third step was to generate the map of NYC to get a better understanding of the retrieved data discussed above. But in order to do so, we first needed to acquire the coordinates of the city. We used *geopy* for that. Next, we used *folium* to generate the desired map.

The neighborhoods of the city were labeled on their map based on their respective coordinates and the borough they belong to. Each borough was assigned a specific color. The below figure shows the generated map.



## Exploring the city

The next step would be exploring the city. That’s when the Foursquare API comes in handy. After generating the necessary credentials, we made calls to the API to acquire the necessary information about the top 20 venues within 500 meters of the center of each neighborhood of each borough in the city. An extensive list of data was retrieved and summarized in the below figure:



The data was then wrangled to obtain the necessary insights which will be discussed in the next sections of this report.

## Data visualization

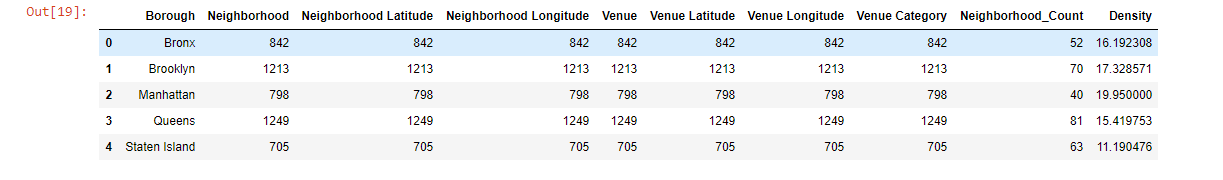
Data visualization was used to better visualize the obtained results. This was mainly done through the use of the *pyplot* method of the *matplotlib* library. Bar charts were mostly used in this project given the nature of the conducted analysis.

## Machine Learning Model

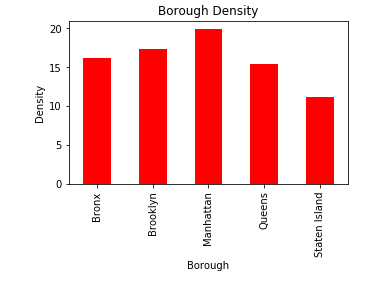
The analysis conducted in this project also uses a machine learning model, the K-means model, to cluster the neighborhoods based on their similarities. The best number of clusters was also determined for better results, as will be discussed in the later sections of this report.

# Results and Discussion

The first step in guiding our tourist towards a rich tourism experience in the city of New York was to guide them to select the borough of the city that is the richest in attraction venues and the most condensed. For that, we wrangled all the obtained data to find the necessary insights to answer this first problem. The obtained datafame for that is shown below:



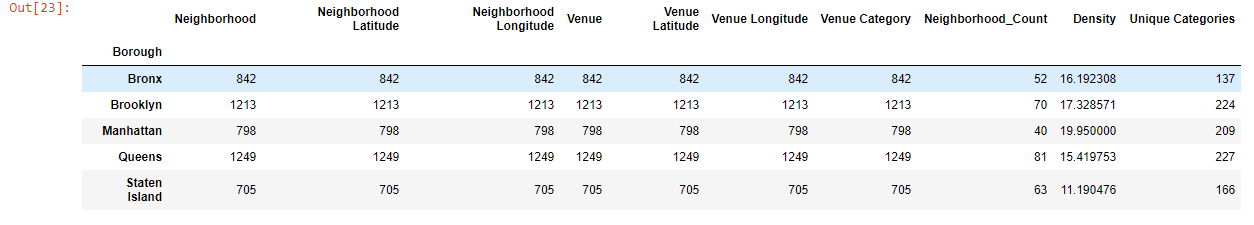
We have created a parameter called Density, which referes to the density of venues in each borough. This parameter mainly indicated the average number of venues per neighborhood in each borough. Having a denser borough means that our tourists will never run out of options, and that they’ll always find something to entertain them upon roaming the streets of the city. The results of the Density parameters were displayed via a bar chart shown below:



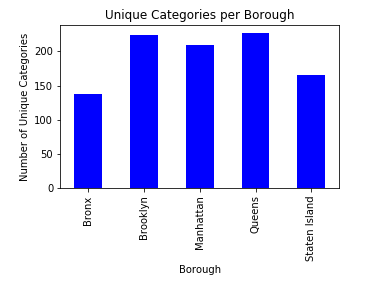
Clearly Manhattan has higher venue density compared to the other boroughs, but not by much. This can be a key indicator to our tourist who's looking to have a very rich touristic experience.

But having a high density of venues does not necessarily mean the tourist will have a rich experience. Richness comes with diversity. Therefore, we have to determine the borough with the most diversified set of venues.

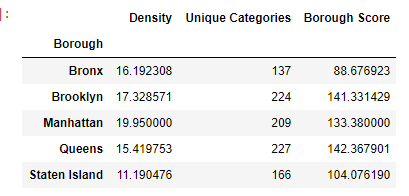
For that, another parameter was created to test this point, and was called “Unique Categories”, indicating the number of unique categories of venue comprised inside each borough.



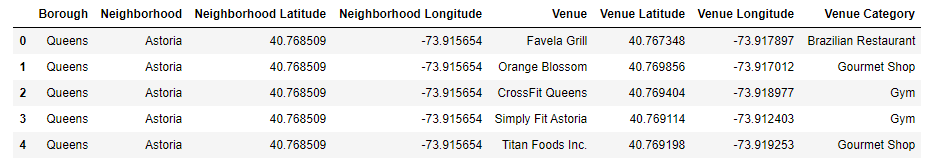
The results of this parameter were visualized in the below bar chart:



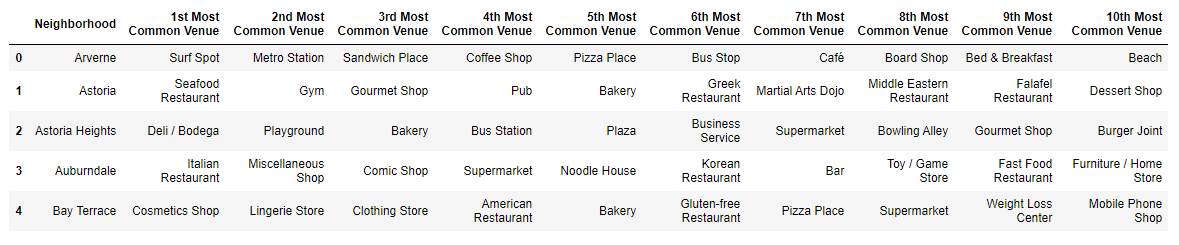
So we have now two assessment parameters at hand to guide our tourist in selecting the best borough to visit (or visit first should they wish to visit a second one or third ones). To make an informed decision, a score for each borough was given by calculating the weighted average of the two parameters used. The uniqueness of the venue was given a higher weight (60%) as it is more important for the tourist that the density of venues(40%) (you can have 1000 venue of the same type, that doesn’t give you a rich touristic experience). The results of this comparison are tabulated in the below dataframe:



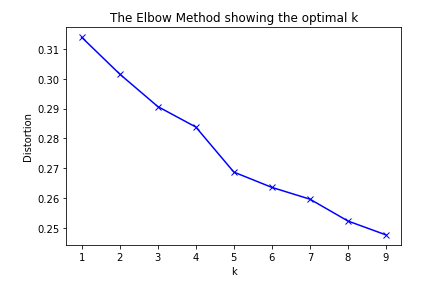
Based on the above results, Queens has the highest score, and that is our borough selection. From now on, we will be dealing with the dataframe corresponding to Queens. This dataframe was extracted from the bigger dataframe we discussed earlier, and is presented below:



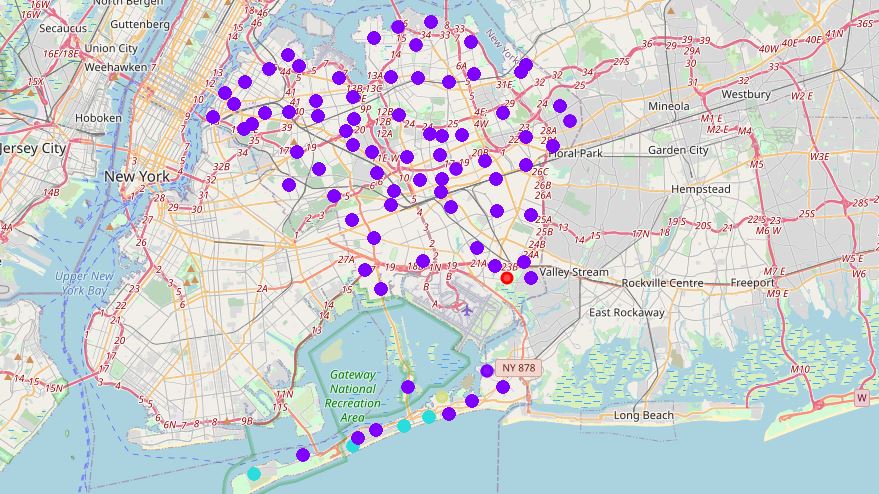
Next, one-hot-encoding was used to determine the categories of venues in each neighborhood of Queens, and the 10 most common venues in each neighborhood were reported. This will aid our tourist in selecting the neighborhoods to visit. The results of this analysis are reported below:



Next, we should cluster the neighborhoods based in their venue similarities. This will help the tourist determining which cluster of neighborhoods ha their themes of interests. To do that, we will be using the K-Means machine learning model. But first, we must determine the optimal number of clusters to be used in our study. You can refer to the enclosed notebook for that, and here we will demonstrate the result of this analysis in the below picture:



Using the elbow method, we can see that our neighborhoods can be clustered into 4 clusters. After determining the optimal number of clusters, the K-means model was applied to our Queens dataset, and the necessary dataframe was generated, including the cluster label of each neighborhood in Queens. The results are visualized using the below map generated using *folium:*



The above map clearly shows that most of the neighborhoods demonstrate high similarities and are clustered under one cluster, except a few which are distributed over 3 other clusters. We can see that this is beneficial to our tourist, as they can select neighborhoods randomly from the biggest cluster, without the fear of missing out on anything. They can also visit the remaining clusters, perhaps all of them, given that they’re unique, and therefore intriguing to the tourist. Of course, they can build their decision on the 10 most common places in each neighborhood we discussed earlier.

# Conclusion

In this project, we guided our tourist in selecting the places to visit during his tour in the City of New York. The results of our analysis should help the tourist selecting the best borough, neighborhoods and venues to visit, and should make his decision-making process an easy and clear process. Data science techniques were used to reach the above results, and extract the hidden insights behind the provided datasets.