**1.Introduction**

**1.1 Problem Statement and a discussion of the background**

New York City is the world’s liveliest citiy where the world flocks to for the best food, theatre, museum and business activity. New York city is home to a large working population with a huge disposable income and a large number of tourists. This makes New York one of the ideal locations to open a new restaurant. Having said that New York city also has one of the largest number of restaurants per capita. With its high real estate prices and stiff competition, its quite a challenge to decide the best neighborhood to open a restaurant and what type of restaurant to open

I would like to use data science to guide the decision making. I would like to take Manhattan which is the borough with the highest GDP. I would like to identify the top ten neighborhoods. Once I have identified the top ten neighborhoods in Manhattan, I would like to perform a search for the different types of restaurants in each neighborhood. I would like to identify the most popular category of restaurants and also the neighborhood in the top ten neighborhoods with the least density of restaurants.

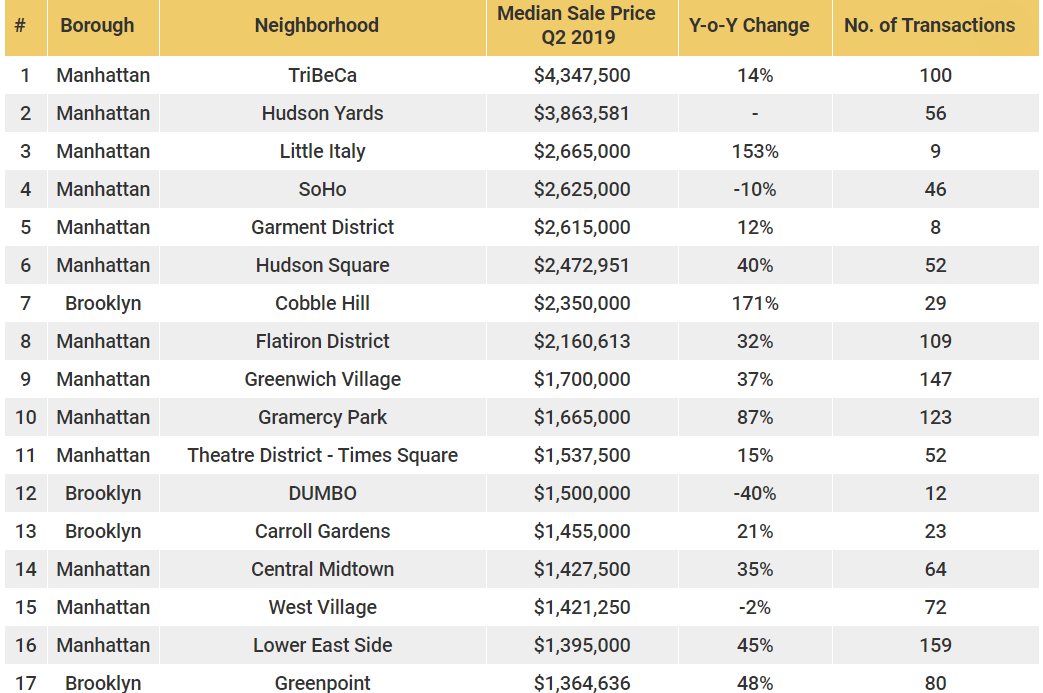
**1.2 People who will be interested in this project**

This would be a huge opportunity for any potential investors looking to open a new restaurant. This analysis would be useful not just for investors interested in opening restaurant but also other consumer-related businesses. The data analysis will also be useful to realtors who can leverage the data to attract potential home buyers or renters.

**2. Data description and analysis**

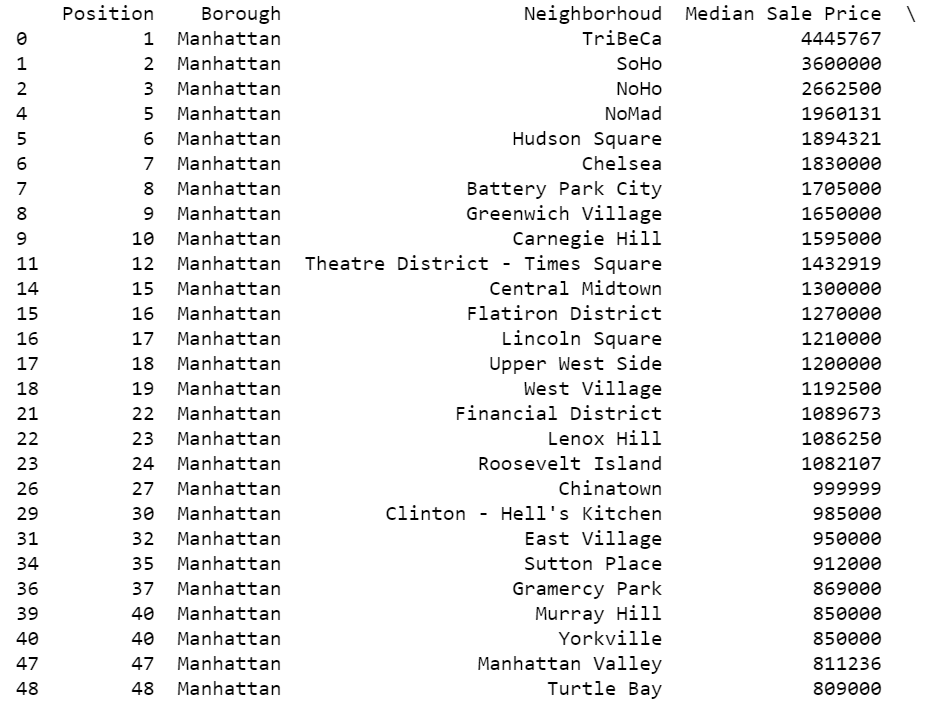
**2.1 Data Source**

I will use the table from the following database <https://infogram.com/2019q2priciestneighborhoods-1h984wgnwp0z2p3> to identify the most priciest neighborhoods in New York. This database lists all the top 50 neighborhoods in New York City with data including meidan sale price, number of transactions and year on year price change.

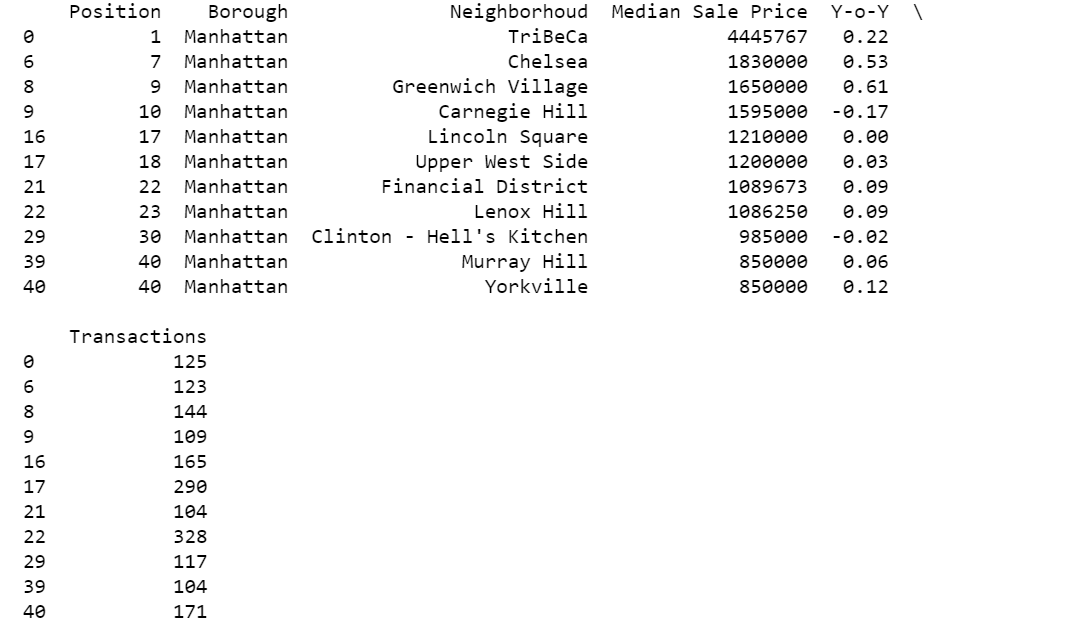


**2.2 Data Cleaning**

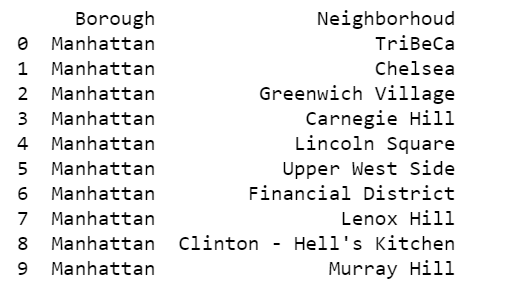
The dataset included all the Boroughs in New York City. I wanted to limit the neighborhood search to Manhattan. I dropped all the other Boroughs from the dataset.



With the data now limited to Manhattan, I wanted to identify the most active neighborhoods with the highest real estate prices. I first filtered the neighborhoods with a minimum of 100 real estate transactions per year



I then sorted the list and picked the top 10 neighborhoods by median sale price. This list would now give me the top 10 neighborhoods in Manhattan by real estate activity and median sale price. In my opinion these would be the top neighborhoods for a restaurant. All my data analysis would be based on this dataset. With the dataset already identified based on Borough, number of transactions and median sale price, I dropped the Median sale price, year on year change and number of transactions from the final data to be used for analysis

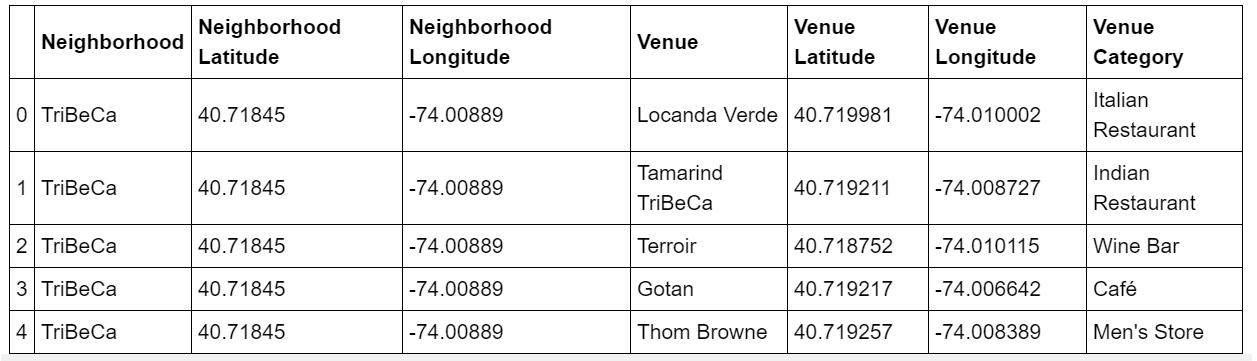


**3. Methodology**

I first used geocoder package to add latitude and longitude information to the dataset.



With the latitude and longitude information, I used the Foursquare API to get the top 100 venues in each of the neighborhoods. The List would contain all the venue categories



I filtered the venue category to only include the restaurants and drop all the other categories.



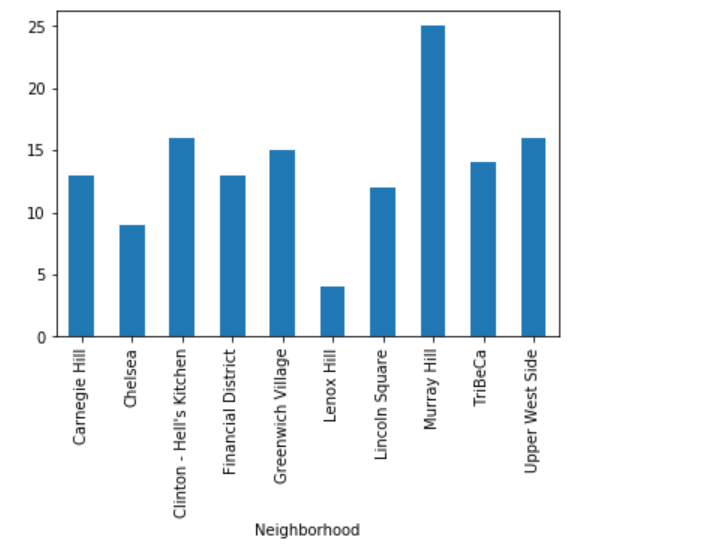
With the dataset containing all the top restaurants in the neighborhoods of interest. I can visualize the data in different ways to give me the most meaningful interpretations.

1. I wanted to first visualize the number of restaurant categories (eg. Italian, Indian, American) in each neighborhood. This information provides a lot of insight on how diverse the interests are. This also provides clues on neighborhoods which have low diversity in restaurants and potential for more growth for other restaurant categories
2. I wanted to also visualize the number of restaurants in each neighborhood. This information provides inputs to any potential investor on the competition. This helps me also identify the neighborhoods with the least number of restaurants.
3. The third set of analysis was to identify the most popular restaurant categories including all the neighborhoods. This analysis would provide information on what is popular and what categories are least represented
4. The fourth set of analysis was to identify the top 5 represented restaurant categories in each neighborhood. This information provides insights on the neighborhood preferences and what is least represented out of the popular categories

With the information on restaurants in all the different neighborhoods I used the machine learning algorithm (k-mean) to cluster the neighborhoods into three different clusters. The clusters would provide information on different neighborhood and restaurant trends

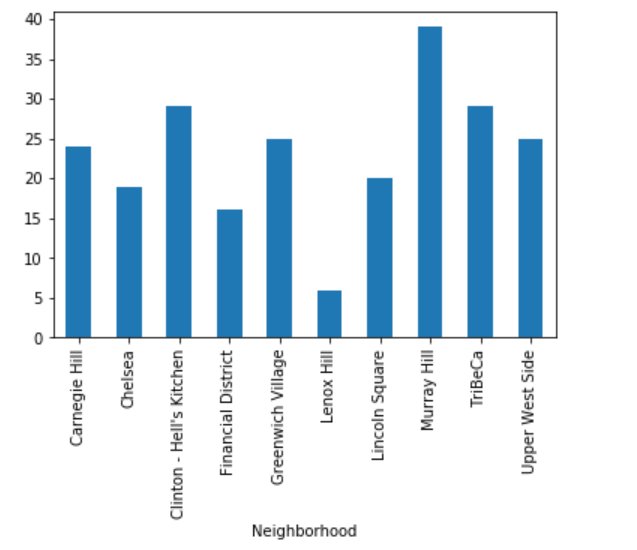
**Results**

1. Number of Restaurant category types by Neighborhood



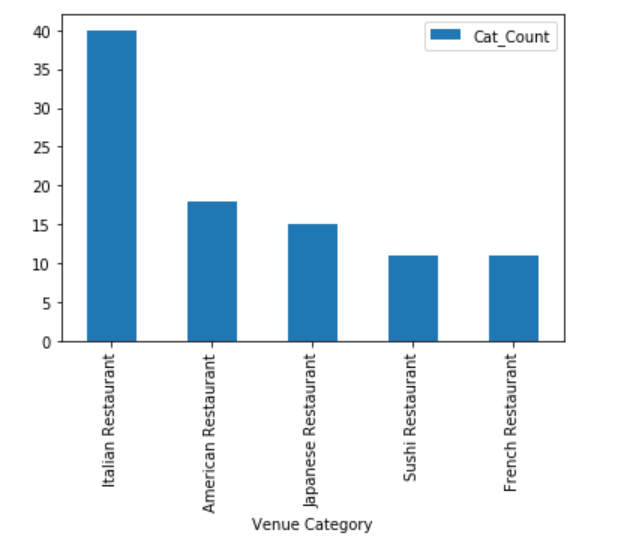
This plot shows that Murray Hill has the most diversity in restaurants, while Lenox Hill and Chelsea have the least diversity.

1. Number of restaurants by Neighborhood



The plot again shows Murray Hill has the highest number of restaurants, while Lenox Hill has the least number of restaurants, followed by Financial district and Chelsea

1. Most popular restaurant categories



Plotting the top 5 restaurant categories combining all the neighborhoods shows Italian is the most popular restaurant type by a very wide margin. If you combine Japanese and Sushi restaurants together, they would be the second popular followed by American and French cuisine.

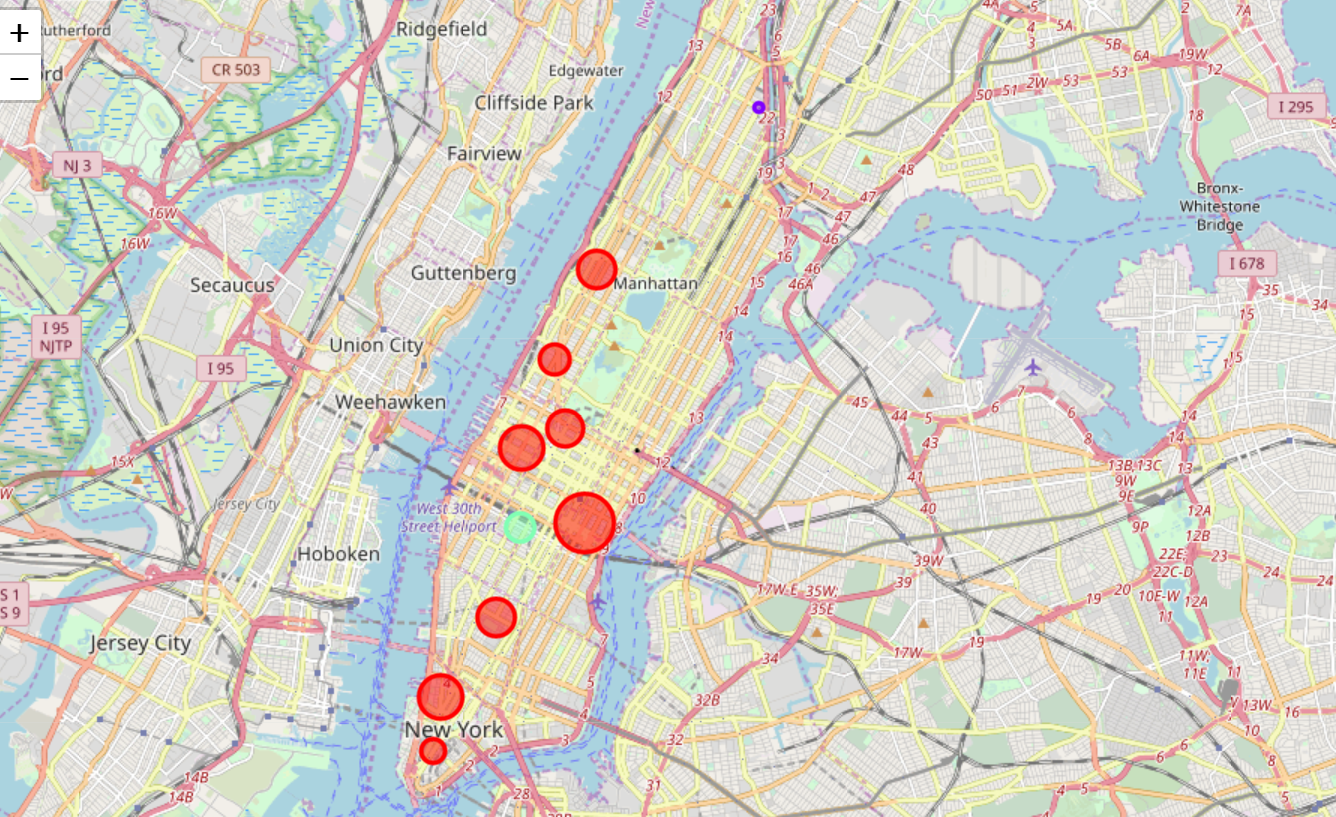
1. Most represented restaurant category by neighborhood



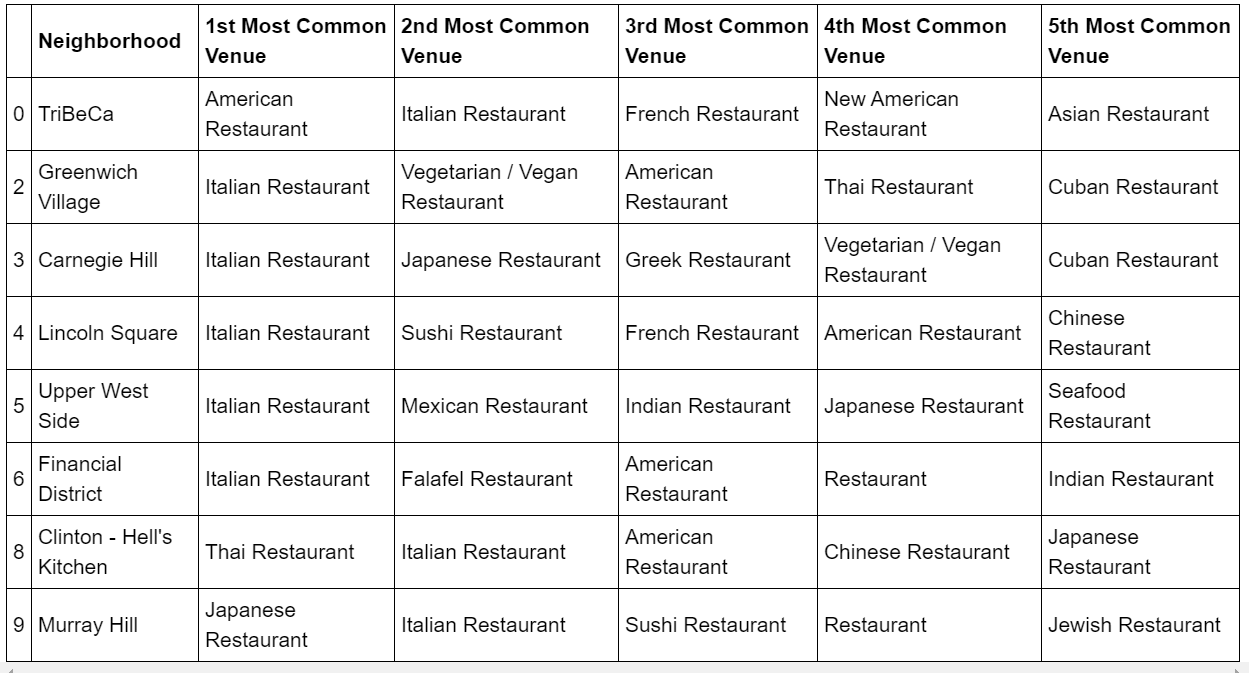
This table shows the most common restaurant types in each neighborhood. As expected Italian is the top cuisine in most neighborhoods. The analysis also shows Lenox Hill is missing all the top restaurant categories (Italian, Japanese, American and French)

1. K-mean Clustering and Visualization of the clusters using Folium package

The k-mean clustering algorithm clustered the ten neighborhoods into three clusters. I used the folium package to visualize the clusters in the map. The radius for the circle in the folium map represents the total number of restaurants at each neighborhood. This gives a clear visualization of the neighborhoods that have the maximum and minimum number of restaurants.



1. Cluster 1

The first cluster combines neighborhoods which have a trend of being filled with the most restaurants and most popular restaurant types. Italian and Asian restaurants were the most popular types in this cluster.

1. Cluster 2



Cluster 2 picked the neighborhood with the least restaurants and the least representation of the popular restaurants

1. Cluster 3



Cluster 3 picked the neighborhood with a representation of mostly restaurants from the lesser popular categories

The clustering from the machine learning algorithm is well aligned with the detailed analysis done on the data. The data clearly shows the following two observations for any potential investor

1. Lenox Hill has the least number of restaurants, least diversity in restaurants and has less representation of the popular categories (Italian, Japanese, American and French)
2. Chelsea has a good number of restaurants, but is less represented for the most popular category – Italian

**Conclusion**

In this analysis I used the data for the top neighborhoods in NYC by real estate sales. I identified the neighborhoods in Manhattan with the most real estate activity and median price. I used foursquare API to pull all the restaurant information. The data was then analyzed to identify the number of restaurants and diversity of restaurants by neighborhood. I also identified the popular categories of restaurant across all neighborhoods and the most represented restaurants in each neighborhood. I also used a machine learning algorithm to cluster the neighborhoods. The clustering from the machine learning algorithm and the data analysis showed the same conclusions.

The conclusion from the analysis showed the most desired neighborhoods for any new investor would be Lenox Hill or Chelsea and the recommendation would be to open an Italian restaurant in either of these two neighborhoods.