**The Battle of Neighborhoods – Opening an Indian Restaurant in New York City, NY**

# 1. Introduction/Business Problem

* **Background**

**New York City (NYC), often simply called New York, is the most populous city in the United States. With an estimated 2019 population of 8,336,817 distributed over about 302.6 square miles (784 km2), New York City is also the most densely populated major city in the United States.** (New York City, n.d.)

**NYC is also known as financial and cultural capital of the world. The city has incredible cultural diversity which has brought food from all around the world.**

* **Problem**

NYC is famous for its cuisine. Its cuisine reflects the city's immigrant background, with a diverse range of different cuisines. Therefore, to survive in such a competitive market choosing a correct location is very important. Indian food with has received a lot of appreciation for its remarkable use of herbs and spices. The food is famous for its great variety of vegetarian dishes. As a result, Indian restaurants have gained popularity all over US

In this project, I will investigate and identify the best potential neighborhoods for opening a new Indian restaurant in NYC.

* **Interest**

This project will provide data-based evidence for anyone interested in opening an Indian restaurant in New York City, NY.

# 2. Data

* **Data Sources**

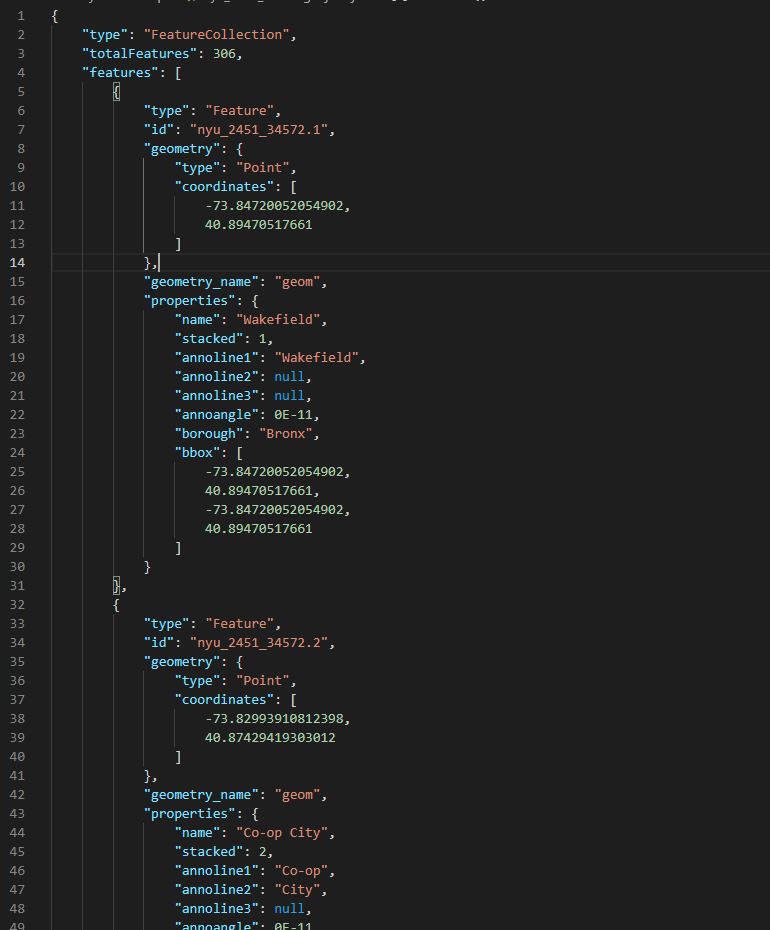
**The following datasets and information are considered for analysis to build a restaurant recommendation model.**

1. Get Neighborhood data of NYC along with latitudes, and longitudes from <https://cocl.us/new_york_dataset>.
2. **Obtained geographical coordinates of the New York using Python Geocoder package.**
3. **Foursquare API** **venues explore** method to get the venues of NYC neighborhoods.

This data will be used to analyze which neighborhood has the highest and lowest concentration of Indian restaurant and other nearby unique **venues** to provide a location where a new Indian restaurant could thrive.

# 3. Methodology

* GitHub repository was used for version control.
* Neighborhood data of NYC along with latitudes, and longitudes was extracted from <https://cocl.us/new_york_dataset> as JSON. After that data was cleaned and processed into a dataframe to start the analysis.

Figure 1: Neighborhood data of NYC

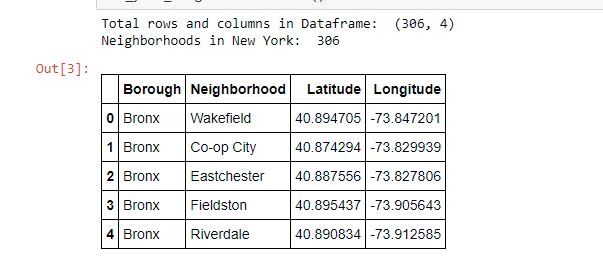
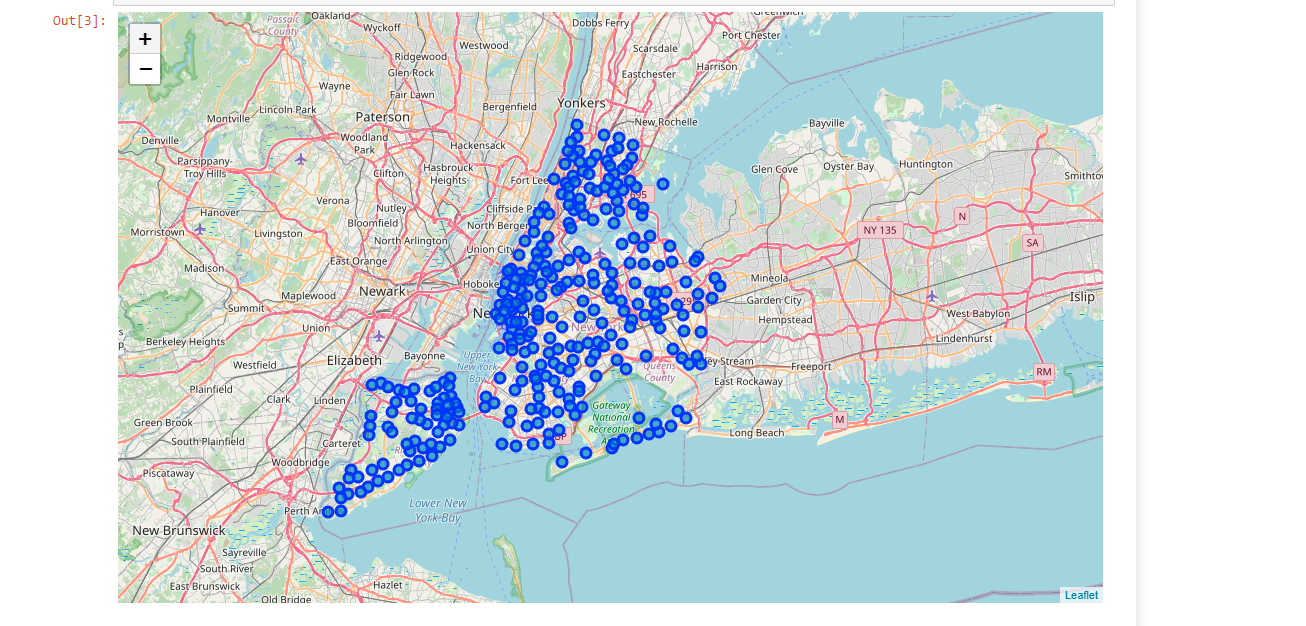


Figure 2: NYC Neighborhood DataFrame

* **Using folium, geographical coordinates of the New York obtained using Python Geocoder package, and the NYC neighborhoods dataframe map of New York with NYC neighborhoods superimposed on top was created.**

Figure 3: Map of New York with NYC Neighborhoods



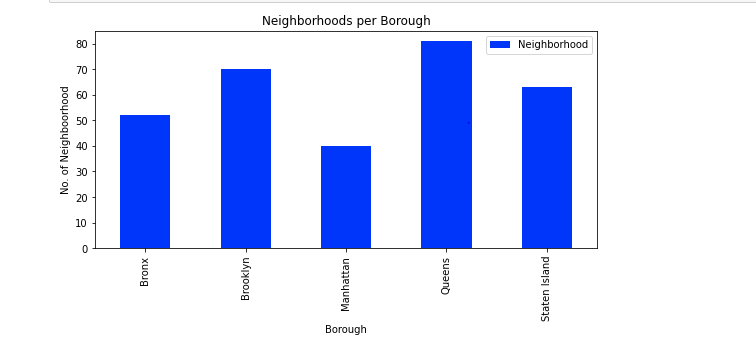
* Next, we plot the bar graph to visualize neighbored per borough.
* Then we used the Foursquare API to get a list of all the Venues in NYC and merged it with neighborhood data to know the nearest Venue for each of the neighborhoods.

Figure 4: NYC Venues

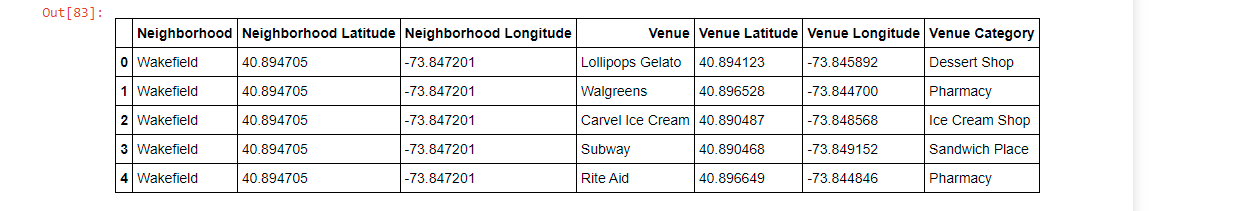
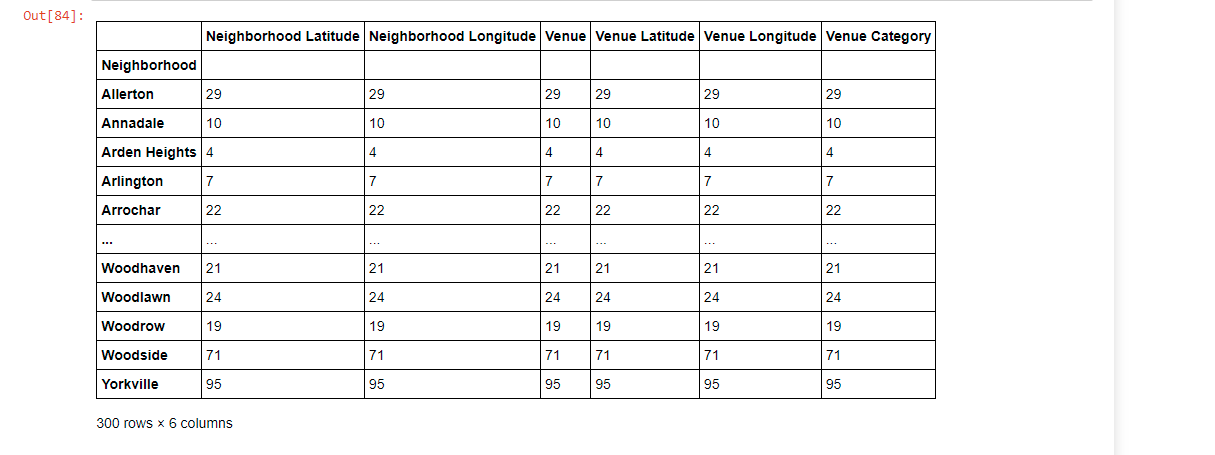
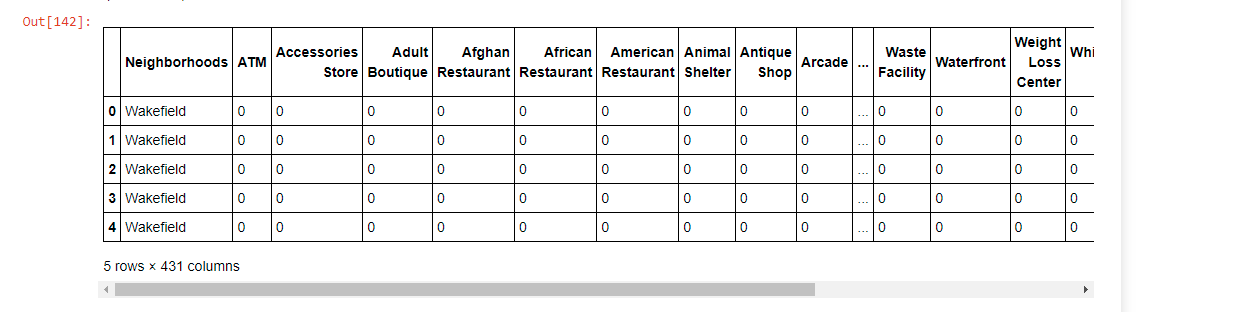


Figure 5 NYC Venues Grouped



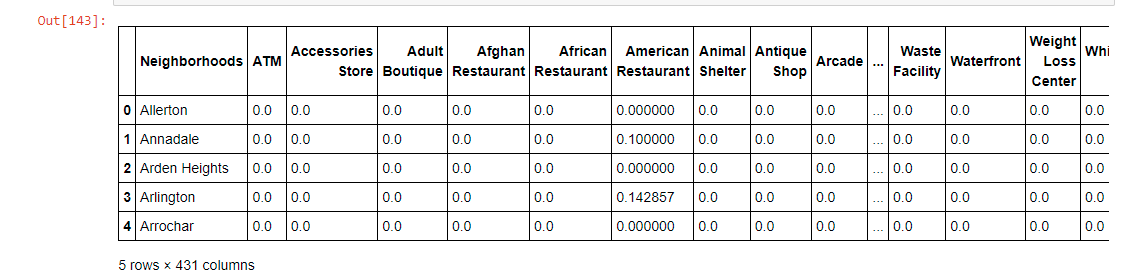
* Performed One hot encoding to change Categorical Data to Numerical Data for Machine Learning algorithms. Individual venues were converted into the frequency of how many of them were in each neighborhood for each of the neighborhoods.

Figure 6: NYC One Hot Encoding



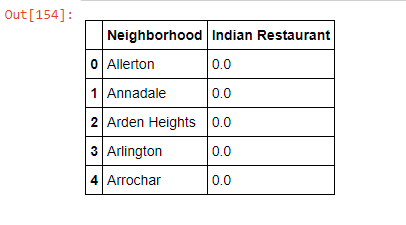
* Then we grouped those rows by Neighborhood and averaged the frequency of each Venue Category's occurrence.

Figure 7: Mean of the frequency of occurrence of each category



* After that created a subset of the previously grouped with only neighborhood name and mean frequency of Indian Restaurants in that Neighborhood.

Figure 8: Indian Restaurant Frequency



* Used K-Mean clustering to cluster neighborhoods into 4 clusters which has same similar mean frequency of Indian Restaurants.

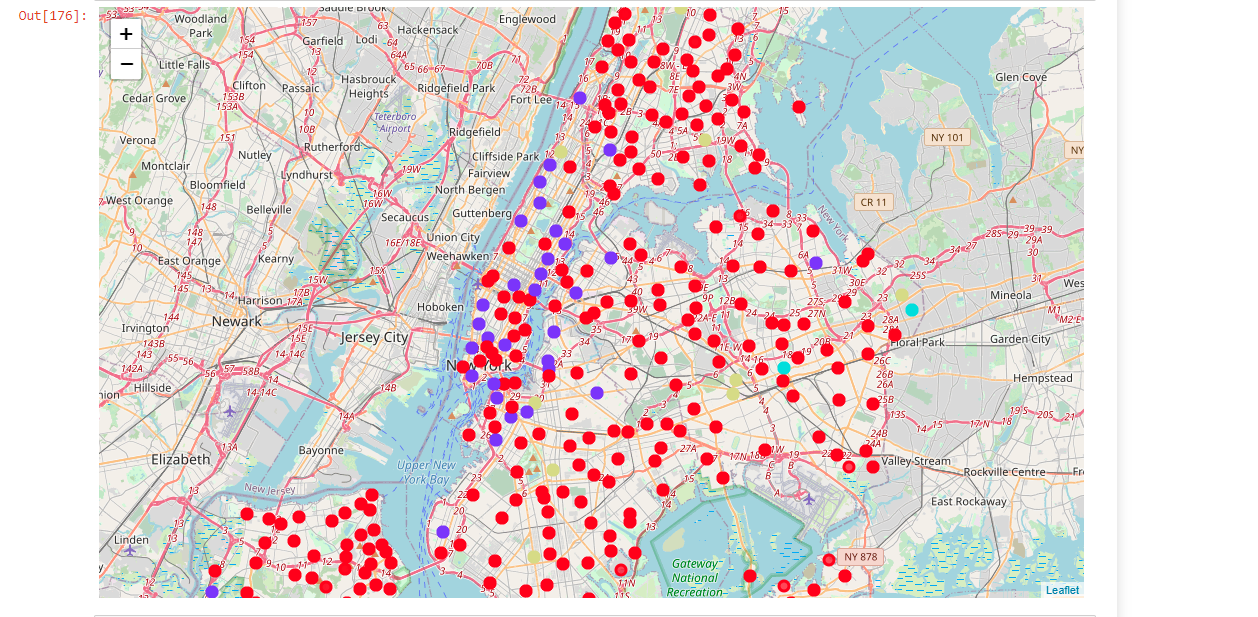
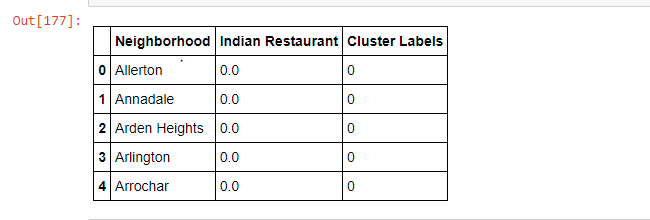
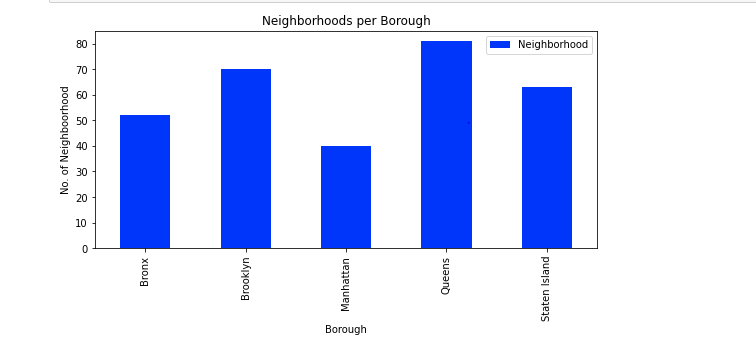


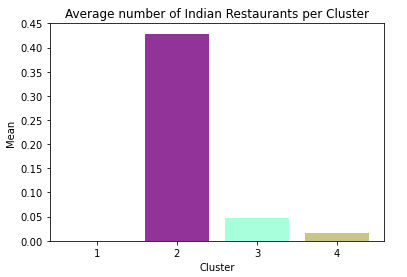
Figure 9: Indian Restaurant Frequency Cluster

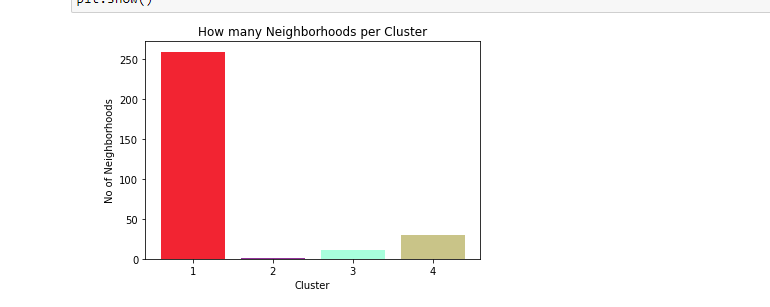


# 4. Analysis/Results

* Queens has the highest number of neighborhoods.

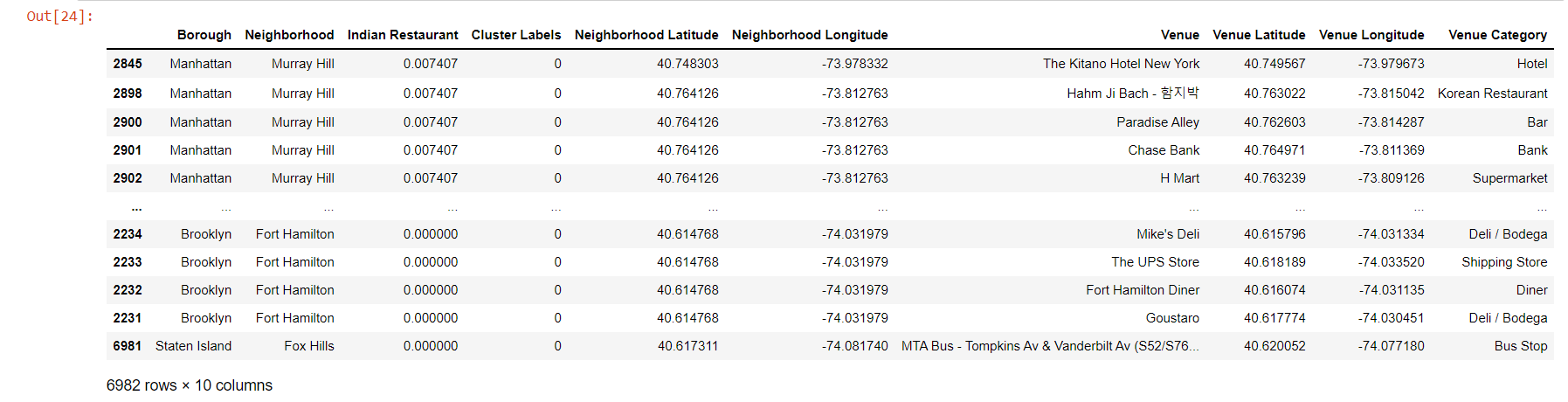


* We have a total of 4 clusters (0,1,2,3) represented by color Red, Purple, Turquoise and Dark-Khaki

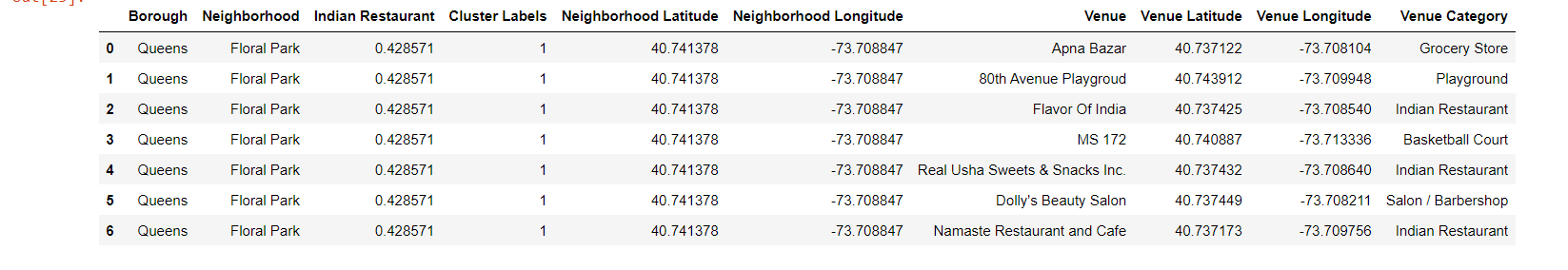


* + **Cluster 1 (Red):**

It has 259 neighborhoods, 394 unique Venue Category but only two Indian restaurants. Cluster 1 has lowest average of Indian Restaurants equating to 0.0003.

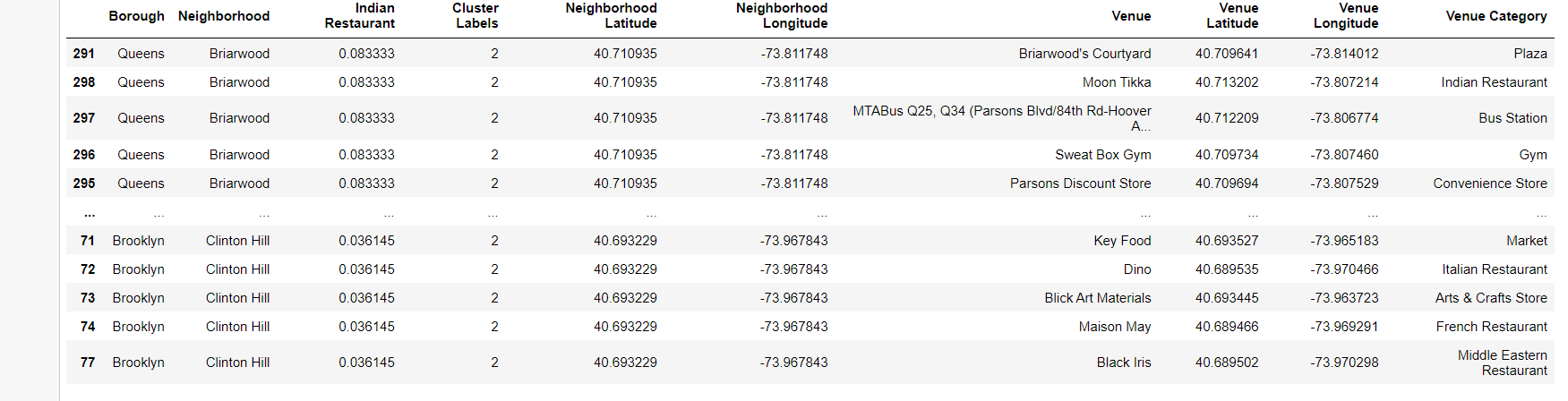


* + **Cluster 2 (Purple):**

It has 7 neighborhoods, 5 unique Venue Category out of which 3 were Indian restaurants. Cluster 2 has highest average of Indian Restaurants equating to 0.43.

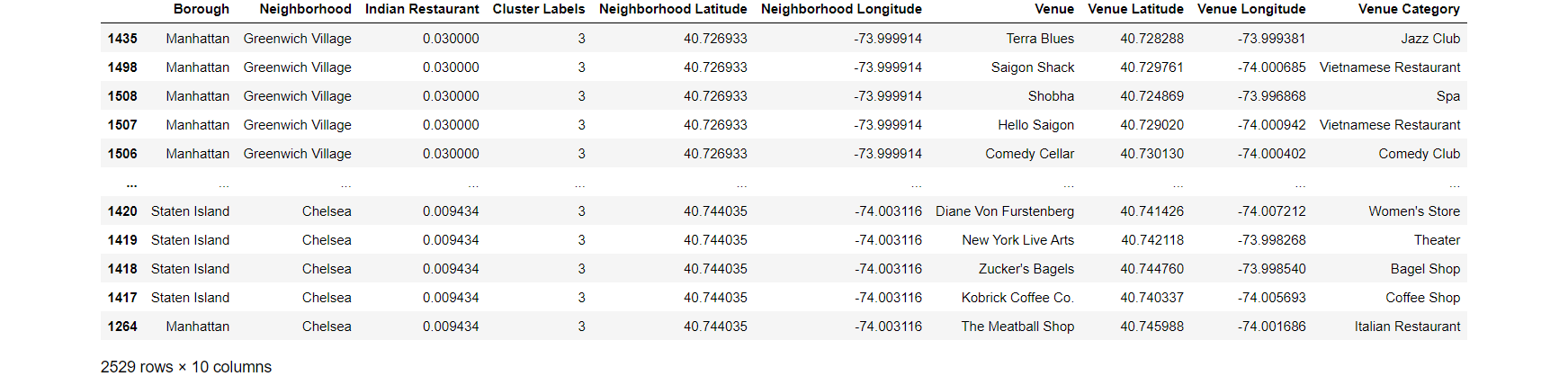
* + **Cluster 3 (Turquoise):**

It has 361 neighborhoods, 114 unique Venue Category out of which 17 were Indian restaurants. Cluster 3 has 2nd lowest average of Indian Restaurants equating to 0.047.



* + **Cluster 4 (Dark-Khaki):**

It has 30 neighborhoods, 293 unique Venue Category out of which 40 were Indian restaurants. Cluster 3 has 2nd highest average of Indian Restaurants equating to 0.047.



# 5. Discussion

Cluster 1 has huge number of neighborhoods and distinct venues but there is little to no Indian Restaurants as a result Murray Hill in Manhattan would be optimal place to open an Indian restaurant.

All the above analysis is based on data obtained from Foursquare API. Moreover, analysis does not take into Indian population which could play an important role while choosing the to open an Indian Restaurant.

# 6. Conclusion

In this project we had an opportunity to analyses a business problem and use data science methodology to solve that business problem. We utilized various Python libraries and Foursquare API to extract, prepare, and visualize data, perform data analysis, and finally recommend a solution to a business problem.

At the end of this project, we can answer the question: “What is the optimal location to open an Indian restaurant in NYC.” This answer was based on the competition aspect of the Indian restaurant business.