IBM Applied Data Science Capstone Report

-Mark Alexander

Contents

1) Introduction/Business Problem

2) Data Sources, Gathering & Cleansing

3) Data Feature Selections

4) Methodology

5) Data Preparation Choices

6) Discussions

7) Conclusion

Introduction

Introduction / Business Problem

Mark an Investor is looking to start a new Indian restaurant in New York City a major metropolitan area with more than 8.4 million people living within city limits New York City is home to approximately 315,000 people from the Indian subcontinent.

The investor is interested in firstly finding out the best location(s) for Indian cuisine in New York City, in what Neighborhood and/or borough the restaurant will have the best chance of being successful.

Secondly the investor is looking to understand the different restaurants such as Chinese, Mexican, French, etc around the neighborhood

Data Sources, Gathering & Cleansing

This section describes the data sourced for this project, as well as the data cleansing and preparation for

Subsequent exploration.

New York City data containing the neighborhoods and boroughs, latitudes, and longitudes will be obtained from the data source: https://cocl.us/new\_york\_dataset

New York City data containing neighborhood boundaries will be obtained from the data source: https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

All data related to locations and quality of Indian restaurants will be obtained via the Foursquare API utilized via the request library in Python.

Methodology

• Data will be collected from https://cocl.us/new\_york\_dataset and cleaned and processed into a data frame.

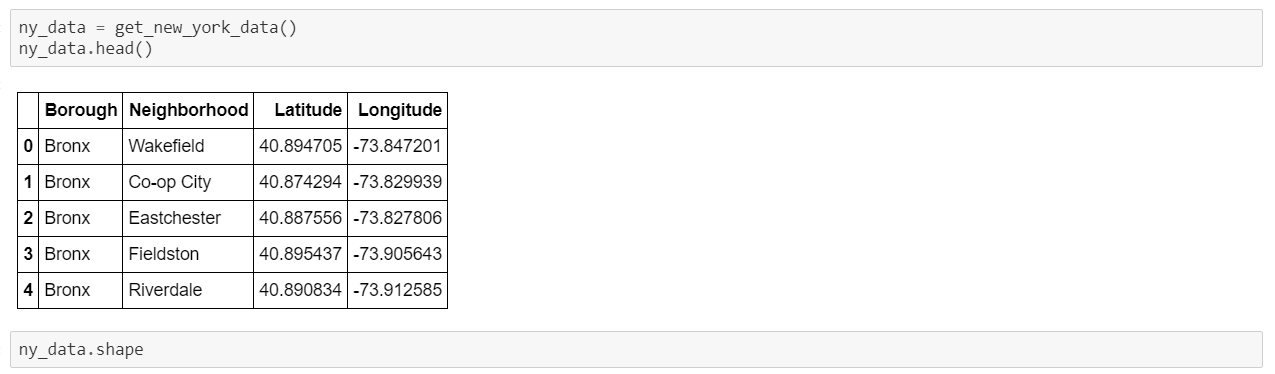
• Foursquare be used to locate all venues and then filtered by Italian restaurants. Ratings, tips, and likes by users will be counted and added to the data frame.

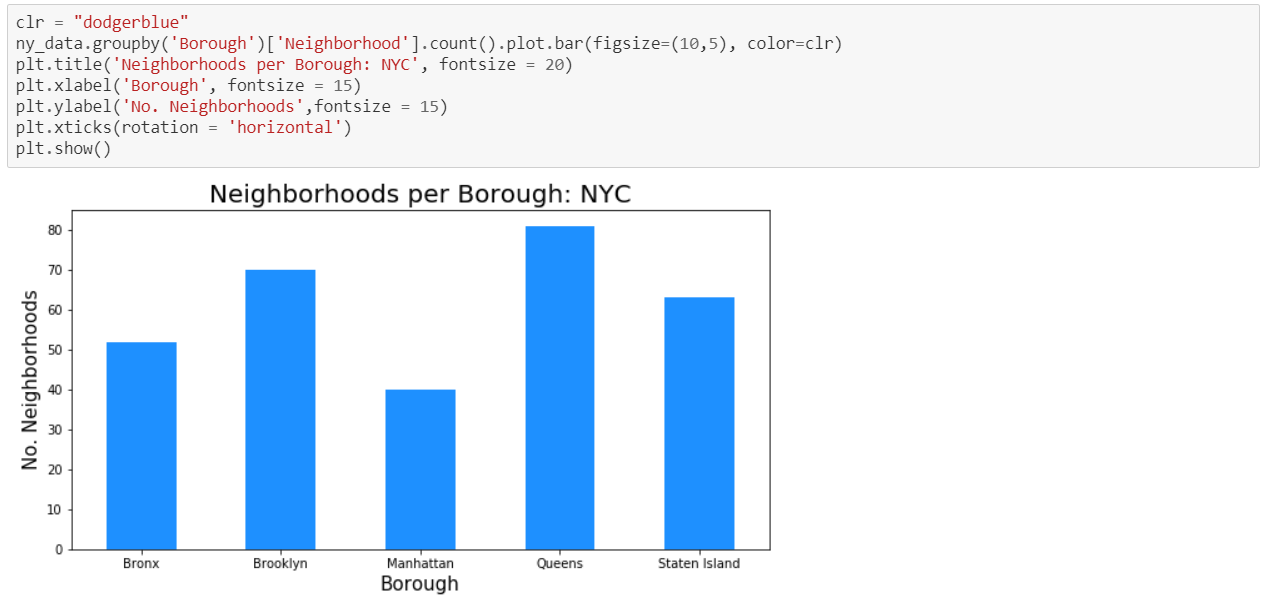
• Data will be sorted based on rankings

• Finally, the data be will be visually assessed using graphing from various Python libraries.

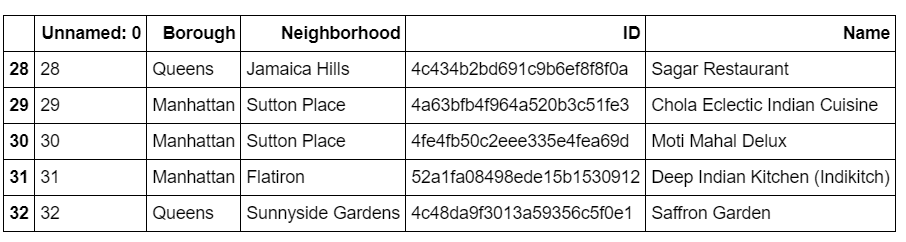
Data Preparation & histograms of neighborhoods

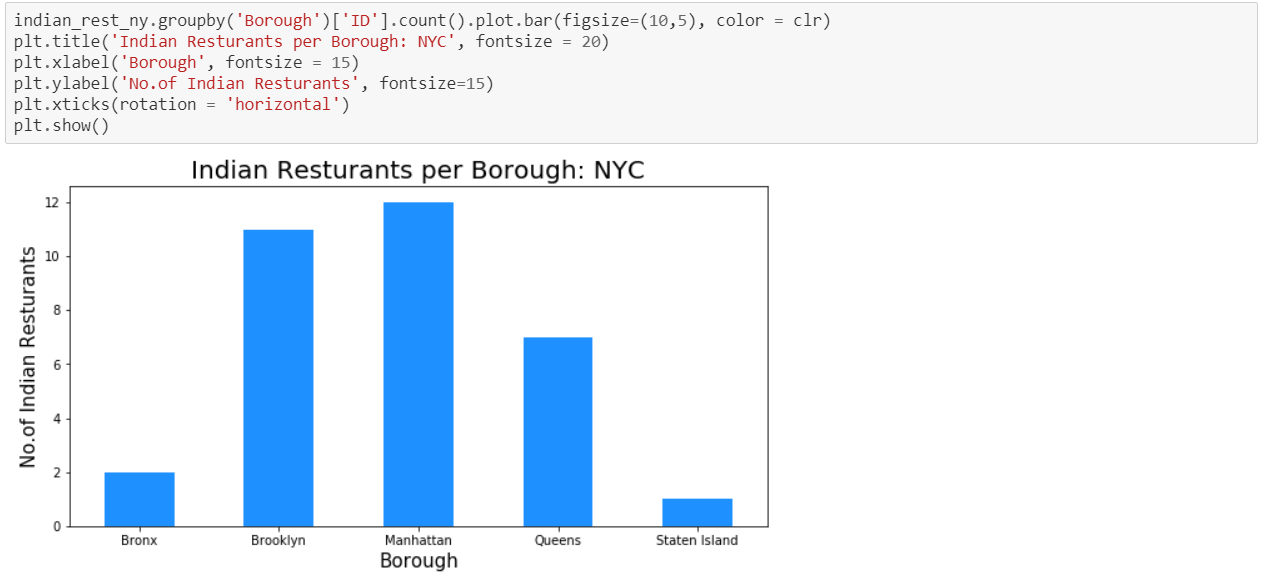
Using: <https://cocl.us/new_york_dataset>



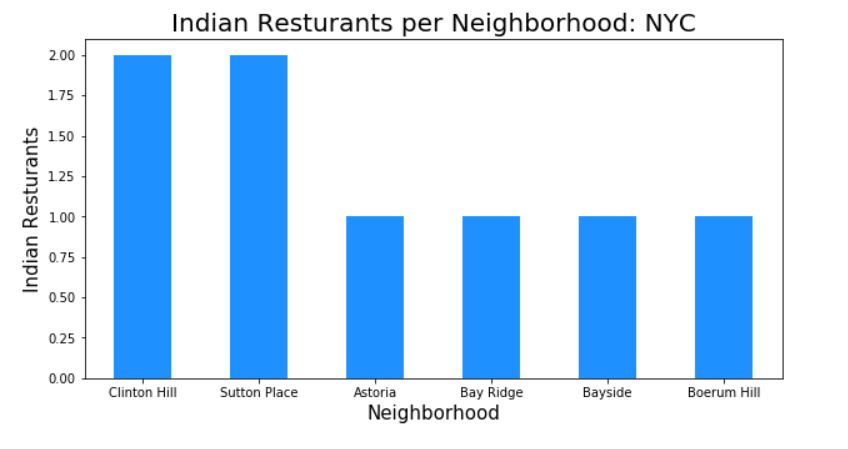


We See that Queens has the most number of neighborhoods, now using Foursquare Location Data we prepare a neighborhood list which contains all Indian restaurants in NYC

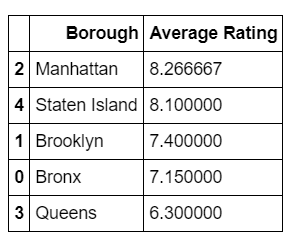


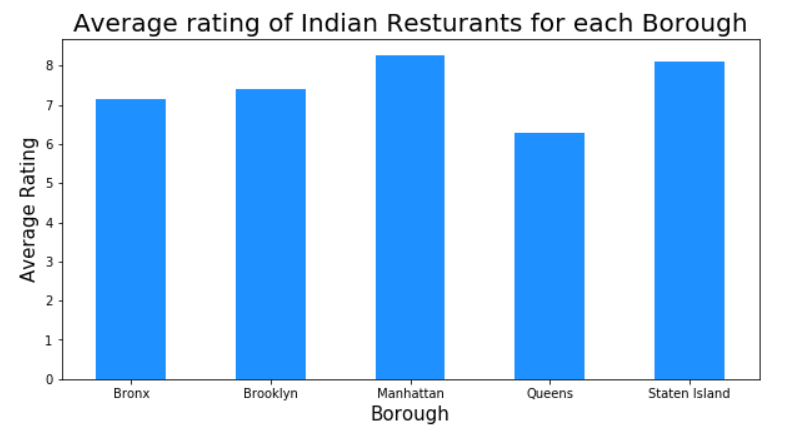


As we continue our analysis, we see that although Manhattan has the least number of neighborhoods, it does have the highest number if Indian restaurants. Additionally, we see how many restaurants the top 6 neighborhoods have. The neighborhood of Clinton has the highest number of Indian restaurants in all of NYC and is actually located in the borough of Manhattan.

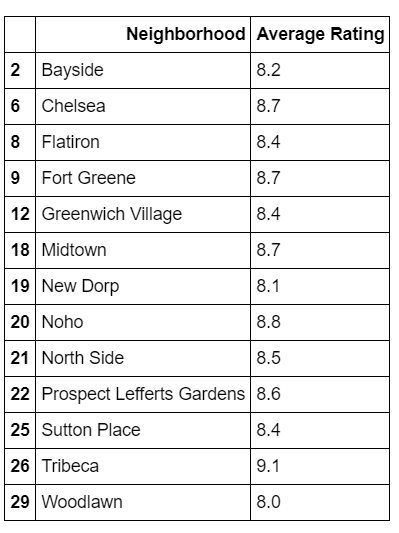


Now we search for all Indian restaurants in NYC based on Average Rating , we see that Manhattan have the highest Average rating based on Likes, Rating and Tips

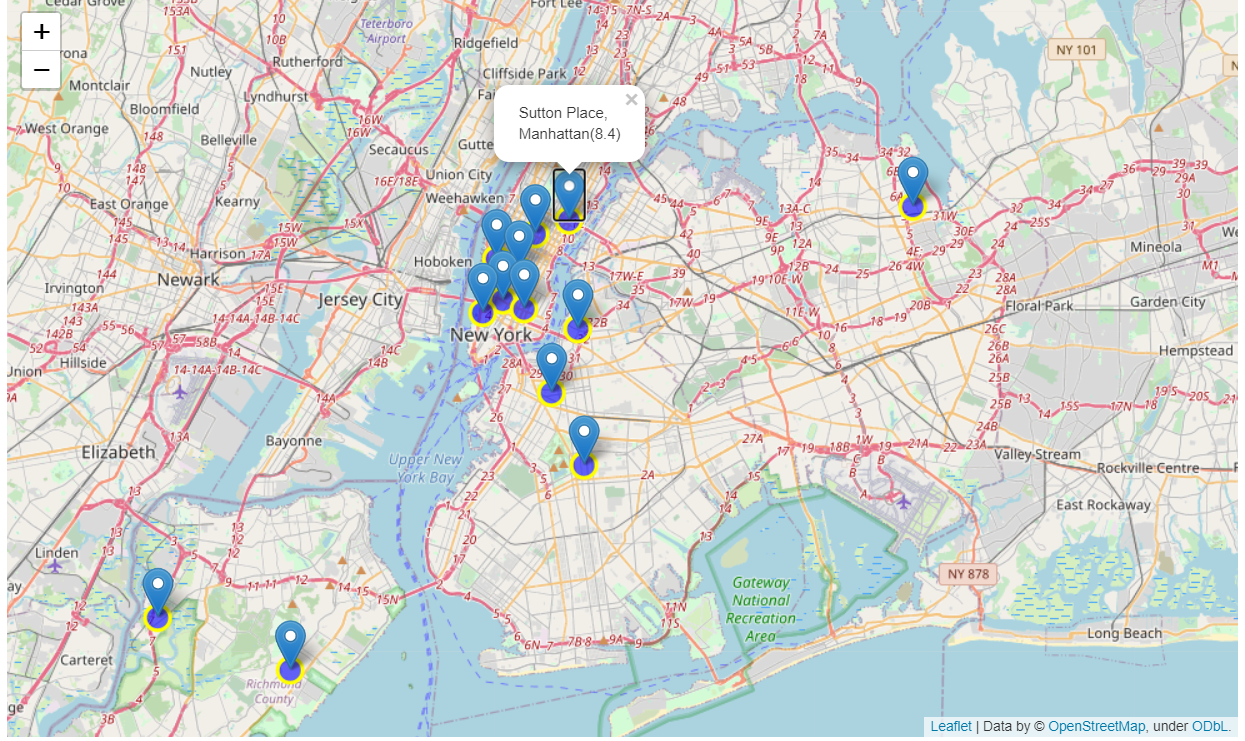




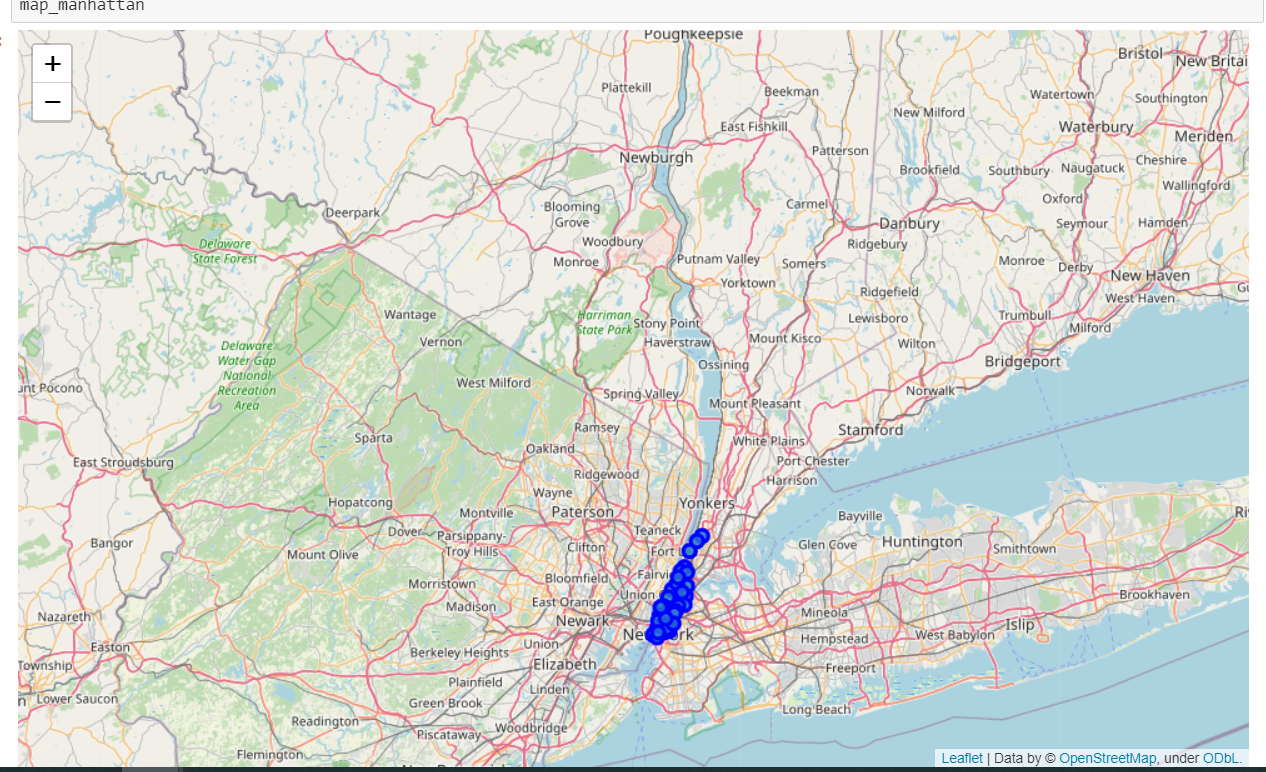
We found that under the Borough of Manhattan a neighborhood called Tribeca had the highest Average Rating.



Visualization of the Average rating of Indian Restaurant in Manhattan



Visualization of the map of Manhattan



We then look at the top five venues in the Neighborhood of Manhattan



We then look at the top ten most common venues in the Neighborhood of Manhattan



Using K-mean Clustering we see that

Cluster 1)



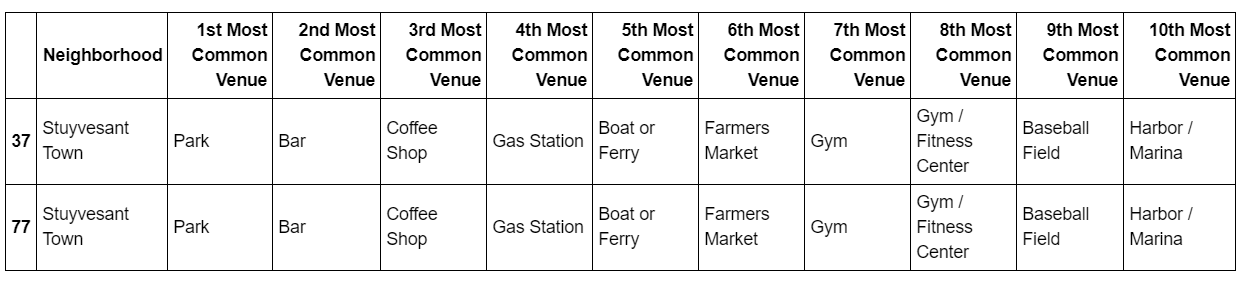
Cluster 2)



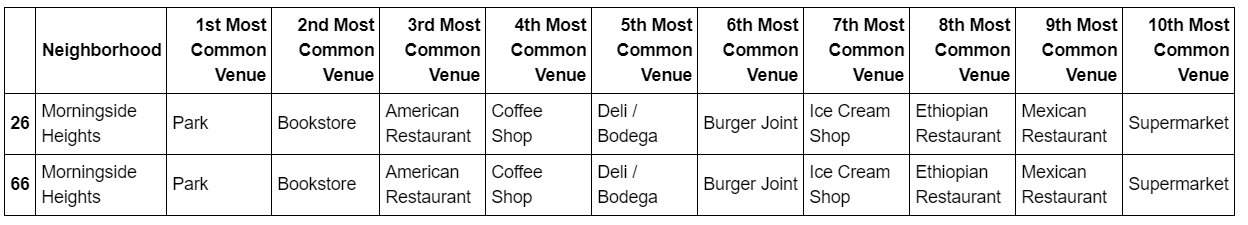
Cluster 3)



Cluster 4)



Cluster 5)



Discussion Recommendations

Finally, the most common venues across the four show some variety as observed in table 1 above. The

Top three most common venues share some similar venue categories, but none share exactly the same

Pattern. For example, Mexican cuisine is ranked first in cluster one,

The venue data – both the top three venue categories in the summary table above, as well as the top

Ten given in the detail charts in the workbook – offer valuable insights into market penetration options.

Depending on the type of venue and market penetration strategy, entrepreneurs investors can choose to align

With direct competitors operating in the same category

Conclusion

The restaurant industry is challenged with balancing a variety of demand factors that sometimes

compete, for example genuinely healthy and organic foods and low cost. Restaurateurs and other

stakeholders can use these data analyses and insights presented in this project to explore how to

effectively target new venue locations and effectively service the demand as a final all of the above analyses is depended on the adequacy and accuracy of Four Square data. A more comprehensive analysis and future work would need to incorporate data from other external databases.