Choosing the Best Restaurant Location in New York City

Applied Data Science Capstone Report

Janrey Nevado

May 16, 2020

INTRODUCTION

Background

New York City is the most populous city in the United States and the New York metropolitan area is one of the most populous in the world. More than 800 languages are spoken in New York City, making it the most linguistically diverse city in the world. 4 in 10 households speak a language other than English. These make New York City an attractive location for restaurants because higher population and population density means higher foot traffic for business establishments and diversity means more people will be more open-minded when they are introduced to new things. However, foot traffic and diversity should not be just the factors to consider when choosing a location.

Finding a good restaurant location is harder than some people think. While a busy street may look like a perfect location for a restaurant, it may not be always the case. Food and service are important to the success of a restaurant, but the location is just as crucial. Some other factors to consider include the following: (1) parking, (2) accessibility, (3) number of competitors, and (4) crime rates among others. While these factors do not ultimately determine a restaurant's success, knowing these is a better starting point than not knowing them at all.

Problem Statement

A new market entrant is looking for opportunities to expand their restaurant business in New York City. They would like to try establishing a new brand of restaurant that is not yet in their portfolio. The company is looking to have answers to the following questions:

1. Which neighborhood(s) is/are the best location(s) to build a new restaurant considering the following factors:
    a. Historical crime rate
    b. Number of competitors
2. What cuisine dominates in each of the neighborhoods?
3. Knowing these things, what cuisine could be introduced to the location?

Target Audience

To answer these questions, the company reached out to our team. The company's objective is to locate the best neighborhood(s) to build a new restaurant in. The company also expects to understand the rationale behind the recommendations.

DATA REQUIREMENTS

For this study, we will be using the following data sets:

1. Location data
    a. Source: Foursquare API
    b. The data will be primarily used to obtain the number of restaurants by type in every neighborhood.
2. Five-year Historical Crime Data of New York City (2015-2019)
    a. Source: https://data.cityofnewyork.us/Public-Safety/NYC-crime
    b. This data will be used to obtain the number of crimes recorded for every neighborhood in New York City. Since this data only is categorized per borough, further data processing will be needed to determine the neighborhood in which the coordinates given belong to. We will use Mapbox API instead of Nominatim for reverse geocoding because we can directly extract the specific neighborhoods from the list of addresses, and it is much more reliable to use.

METHODOLOGY

Since we wanted to find the neighborhoods with least number of crimes and least number of competitors, *k-means clustering* will be used to cluster neighborhoods that have similar characteristics. What we would like to obtain are the neighborhoods that have the least number of crimes and least number of competitors. For us to have an insight on which neighborhoods have the highest number of crimes on average, we processed the data in order to get the average crime count from 2015 to 2019 by level of offense.

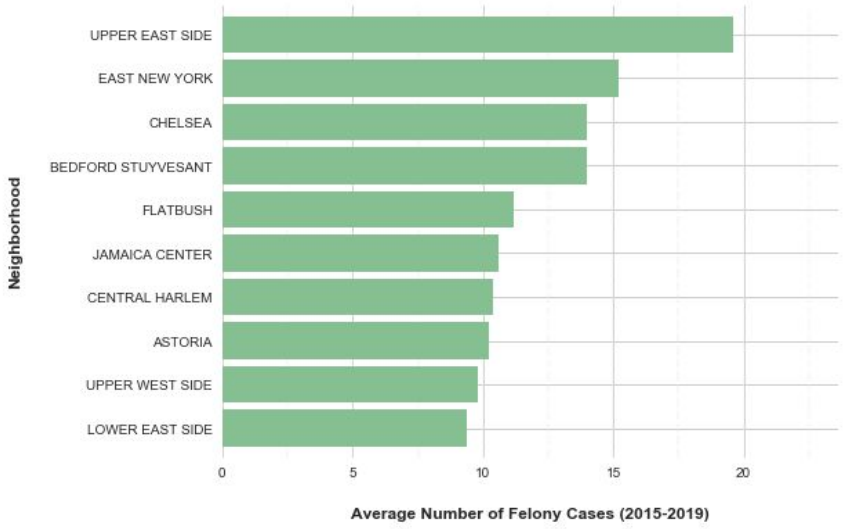

Fig.1: Top Neighborhoods by Average Number of Felony Cases (2015-2019)

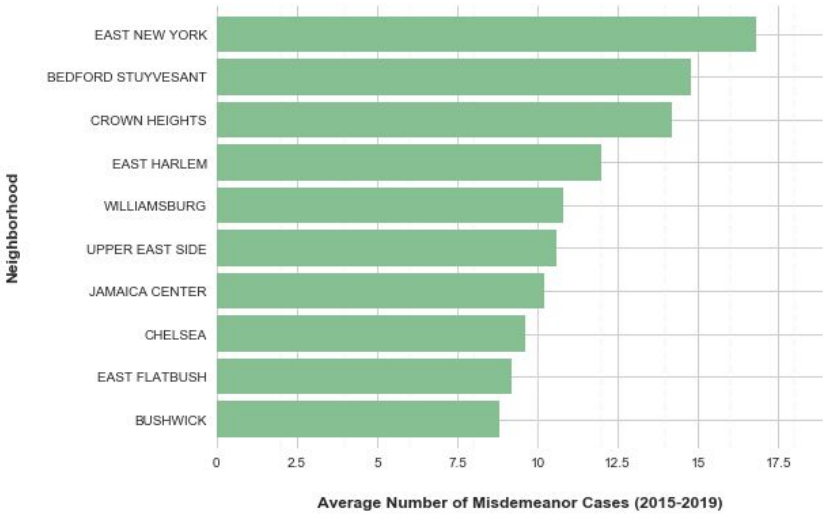

Fig. 2: Top Neighborhoods by Average Number of Misdemeanor Cases (2015-2019)

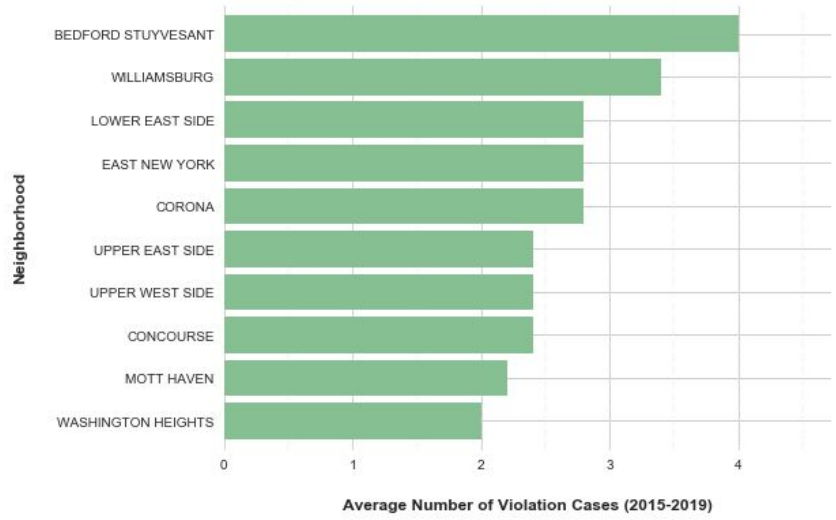

Fig. 3: Top Neighborhoods by Average Number of Violation Cases (2015-2019)

According to the charts, the top neighborhoods by average number of cases are mostly in Manhattan and Brooklyn (Bedford Stuyvesant, East New York, Lower East Side, Upper East Side, and Upper West Side). These are the neighborhoods that we should avoid based on our criteria.

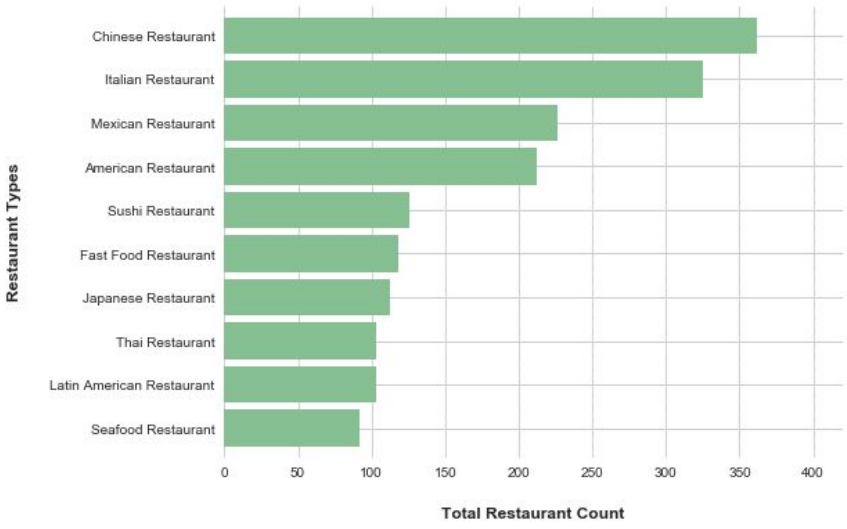

Fig. 4: Top Restaurant Types by Total Restaurant Count (2015-2019)

The most common restaurant types in New York City are Chinese, Italian, Mexican, Fast Food, Japanese, Thai, Latin American, and Seafood Restaurants.

The final dataset that we used for clustering contains 304 rows (304 neighborhoods) and 107 columns (106 features, 1 ID column). No standardization or normalization is done because the data points are very similar in context (counts).

Elbow method is used to find the optimal number of clusters *k*. The scoring parameter metric used is distortion score, which computes the sum of squared distances from each point to its assigned center. The optimal number of clusters *k* is 3 using elbow method.

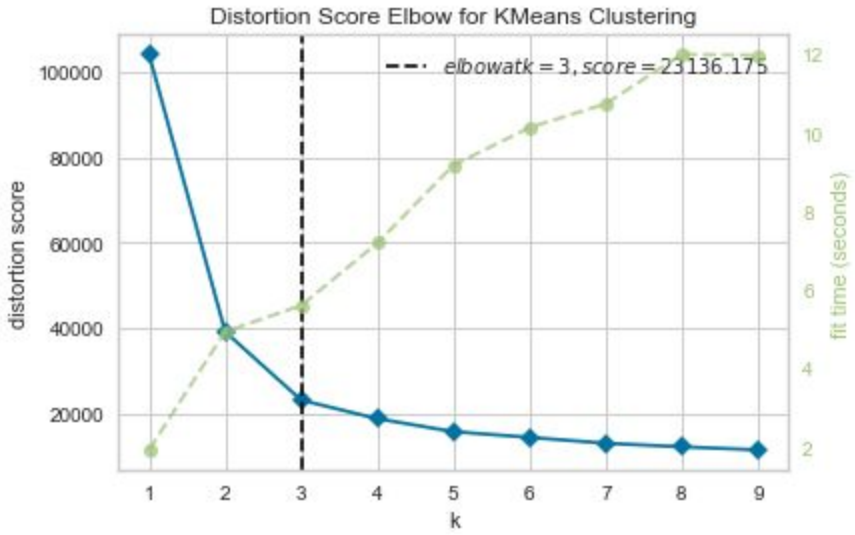

Fig. 5: Distortion Score Elbow for KMeans Clustering

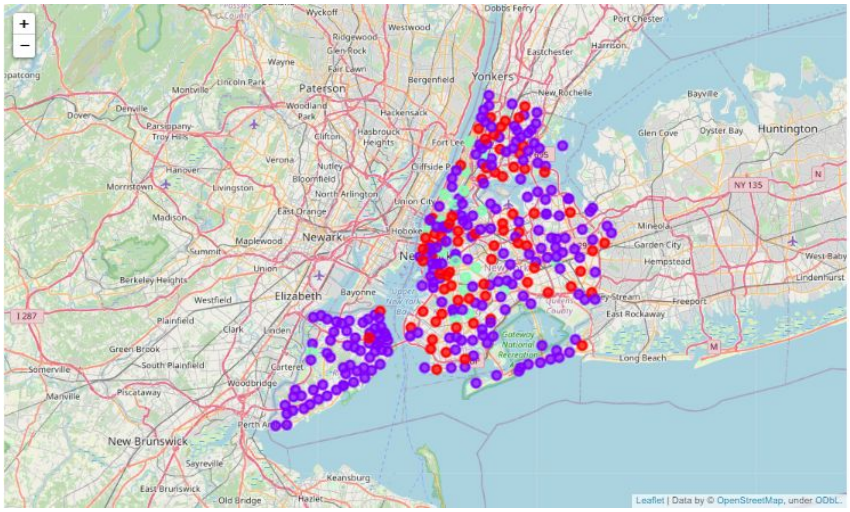

Fig. 6: Map of New York City with Clusters plotted

RESULTS

After clustering the neighborhoods into three clusters, here is the count of neighborhoods per cluster.

The k-means clustering algorithm was able to cluster the neighborhoods into three neighborhood groups with distinct characteristics. Most neighborhoods are under cluster 2 while few neighborhoods are under cluster 3.

Clusters	Neighborhood Count
1	74
2	212
3	18

Table 1: Clusters with Neighborhood Count

The following graphs show the average number of restaurants by type and average number of cases by level of offense per cluster.

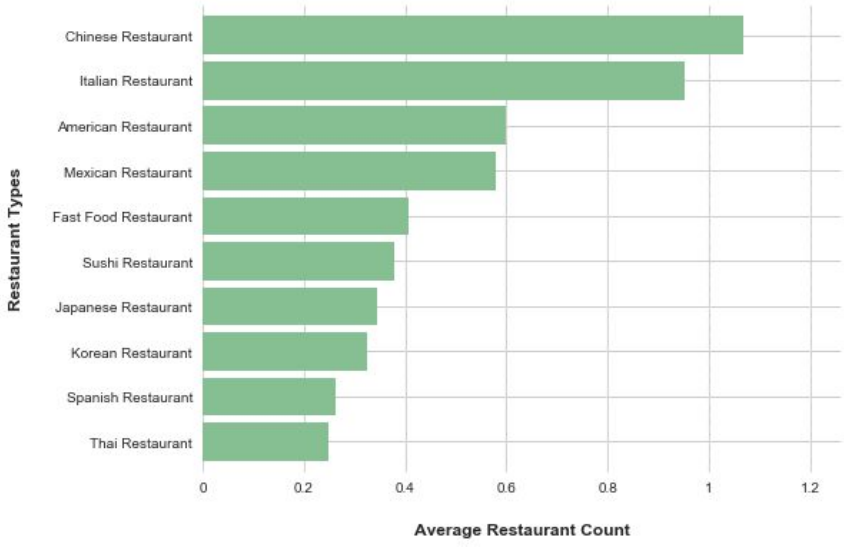

Fig. 7: Top Restaurant Types by Average Restaurant Count (Cluster 1)

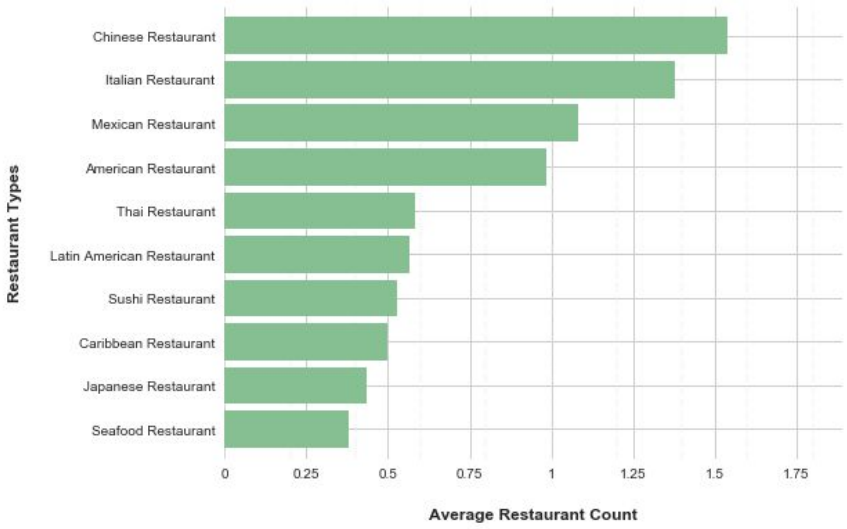

Fig. 8: Top Restaurant Types by Average Restaurant Count (Cluster 2)

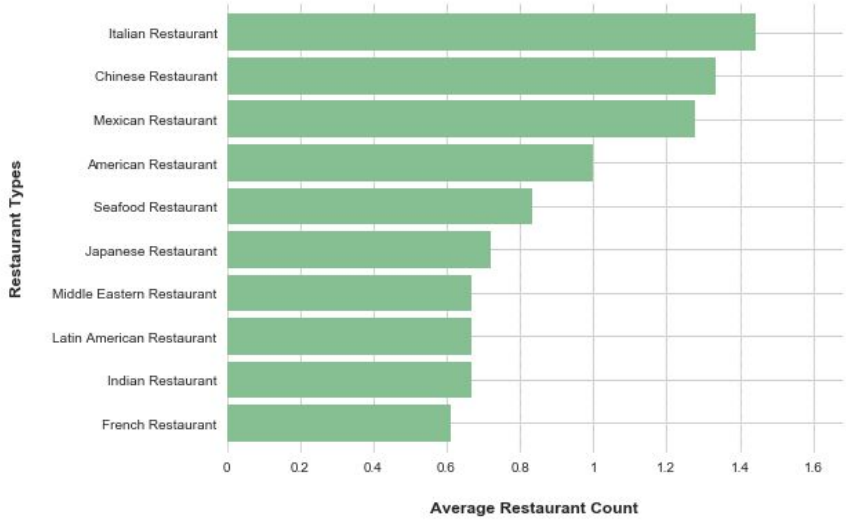

Fig. 9: Top Restaurant Types by Average Restaurant Count (Cluster 3)

The top restaurant types in all clusters are very similar to the one top restaurant types in all neighborhoods before clustering (see Fig. 4).

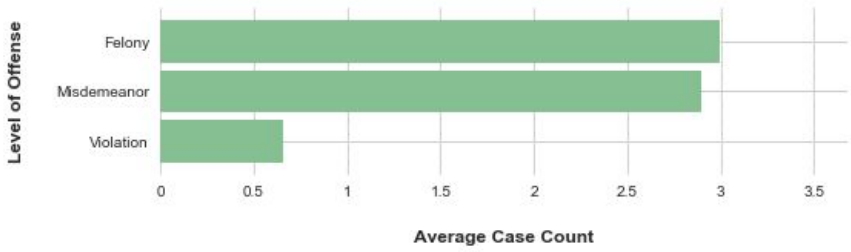

Fig. 10: Top Levels of Offense by Average Case Count (Cluster 1)

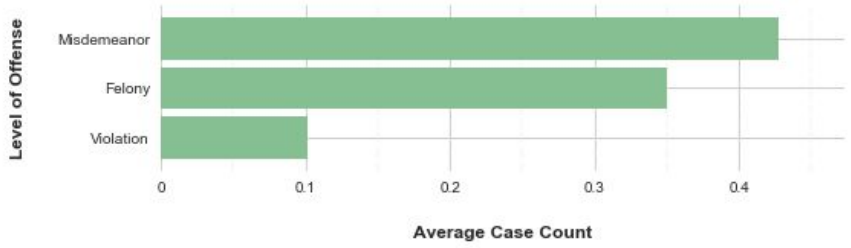

Fig. 11: Top Levels of Offense by Average Case Count (Cluster 2)

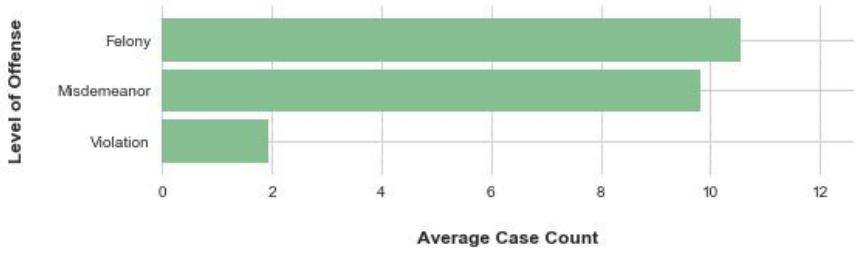

Fig. 12: Top Levels of Offense by Average Case Count (Cluster 3)

According to the graphs, Cluster 2 has the lowest average crime count while Cluster 3 has the highest average crime count. This suggests that neighborhoods in Cluster 3 are not that safe while neighborhoods in Cluster 2 are pretty safe. Cluster 1 also has a pretty low average crime count.

DISCUSSION

The study sought to answer the following questions:

1. Which neighborhood(s) is/are the best location(s) to build a new restaurant considering the following factors:
    a. Historical crime rate
    b. Number of competitors
2. What cuisine dominates in each of the neighborhoods?
3. Knowing these things, what cuisine could be introduced to the location?

To answer the first question, we can see from the clusters that Cluster 2 neighborhoods have the things that we are looking for because the average restaurant count is the lowest among three clusters, which suggests that competition is not that intense. Crimes are also almost nonexistent in these areas, which is good for the business in the long run.

Cluster 3 neighborhoods have things that we are *not* looking for: high crime rates and high number of competitors. Cluster 1 is just like cluster 2 but with a higher number of competitors and crimes; however, it is not as bad as Cluster 3, which still makes it a viable cluster for building a new restaurant.

Clusters	Neighborhood Count	Description
1	74	Relatively high average number of restaurants and low average number of crimes
2	212	Relatively low average number of restaurants and very low average number of crimes
3	18	Relatively high average number of restaurants and high average number of crimes

Table 2: Clusters with Neighborhood Count and Description

To answer the second question, Chinese, Italian, Mexican, and American Restaurants are the most common restaurants in the cluster, so creating a new restaurant that is similar to these restaurants may not appeal to customers.

To answer the third question, there are a lot of cuisine that can still be introduced to the neighborhoods; however, to reduce risks, it would be good to introduce something that is not completely new to the neighborhoods like European cuisine (French, Greek) or other Asian cuisine (Japanese, Korean, Thai, Indian, and Filipino). In this case, going with other Asian cuisine is recommended because the data shows that Asian restaurants are more common and familiar in the area, so it would be easier for the company to attract new customers.

CONCLUSION

In summary, analysis of the data used in this project shows that New York City is still a very attractive location to build a new restaurant because the majority of the neighborhoods have low crime rates and low competition. It is also reinforced that New York City is a really diverse place given the numerous restaurants offering different cuisine.

The analysis done is rather simple, and it will be better if other datasets are also incorporated in the study, like population growth, population density, rent prices, nearness to market, nearness to transportation, and number of tourists among others. Other clustering algorithms can also be considered given the new forms of data that will be added.

Overall, I really enjoyed doing all the courses in this specialization. As a newbie in data science, I really learned a lot, and I think I renewed my passion to solve problems and to come up with solutions that are actionable.