THE BATTLE OF THE NEIGHBORHOODS

Applied Data Science Capstone by IBM/Coursera

Opening a Grocery Store in Safest Borough of New York

Prepared By: Muhammad Ilyas

Table of Contents

Introduction:

Background

New York (NY), is the most populated city in the United States with an estimated population of 8.3 M (according to 2019 census data) distributed over about 302.6 square miles (784 km2). Besides its population, it is also been called both the world's leading financial center and the most financially powerful city in the world. It is the home to the world's two largest stock exchanges by total market capitalization, the New York Stock Exchange and NASDAQ.

It has very diversified and multicultural environment which attracts everyone to be part of it. New York is home to the largest ethnic Chinese population outside of Asia, with multiple distinct Chinatowns across the city. Not only for Chinese, this city is the most attractive location in the globe for everyone, especially for migrants and immigrants.

Because of its multicultural and financial position, although it provides many opportunities of doing new businesses, but this market is very much volatile and highly competitive also. As the opportunity cost of doing business is very high so any new business venture or expansion needs to be analyzed very carefully.

Problem Description

Beside its multicultural, dynamic and attractive environment, crime rate in some boroughs of NY is also high compare to other boroughs. This project will help the stakeholder to find a comparatively safe and secure location for opening a new business of Grocery Store in New York.

First, we will choose the **safest borough** by analyzing crime data and then will short list neighborhoods in the safest boroughs having **low competition but heavy population**.

By making use of our data science tools and analyzing data, most promising neighborhoods in the safest borough will be presented to the stakeholder for their final selection process.

Interest / Target Audience:

This project will be of interest for all those stakeholders who are intended to open a new grocery store business in New York.

Data Sources:

To satisfy our problem description, following are the factors which will influence stakeholder for final decision making.

- Finding the safest borough based on crime statistics:
- Gathering neighborhoods venues and population data within the safest borough:
- Choosing the most promising neighborhoods for final selection.

Following data sources will be needed to extract/generate the required information:

1. **Borough Crime Data**
   To find safest borough, real world data of NYPD complaint records will be collected from NYC Open Data source. This dataset includes all valid felony, misdemeanor, and violation crimes reported to the New York City Police Department (NYPD). The link to dataset which will be used is https://data.cityofnewyork.us/Public-Safety/NYPD-Complaint-Data-Current-Year-To-Date-/5uac-w243.

2. **Neighborhoods Data**
   Neighborhoods in each borough will be collected from open data source dataset. Along with neighborhoods detail this dataset is also contain geolocation of each neighborhood. Out of this dataset, neighborhoods of safest borough will be selected for further analysis based on population and venues. The link to dataset which will be used is https://geo.nyu.edu/catalog/nyu_2451_34572

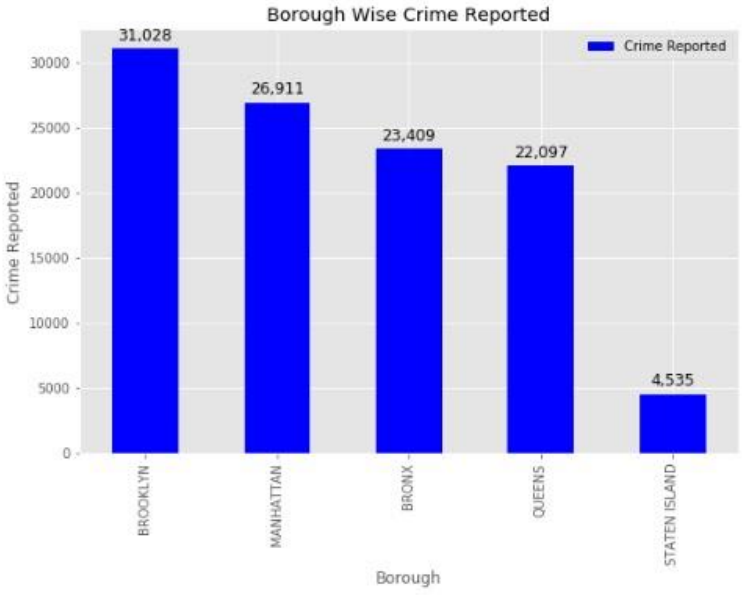
3. **Neighborhoods Population Data**
   To make grocery store business successful, population is one of the key features which should be taken into consideration for analysis. For our project, neighborhoods population data will be collected from NYC Open Data Source. This dataset contains 2010 census data at the Neighborhood Tabulation Area (NTA) level. We will use this data along with venues detail to compile our final neighborhoods selection. The link to dataset which will be used is https://data.cityofnewyork.us/City-Government/Census-Demographics-at-the-Neighborhood-Tabulation/rnsn-acs2

4. **Neighborhoods Venues Detail**
   Neighborhoods geolocation coordinates will be used in ***Foursquare API*** to explore and fetch the venues data.

Methodology & Analysis:

Step 1

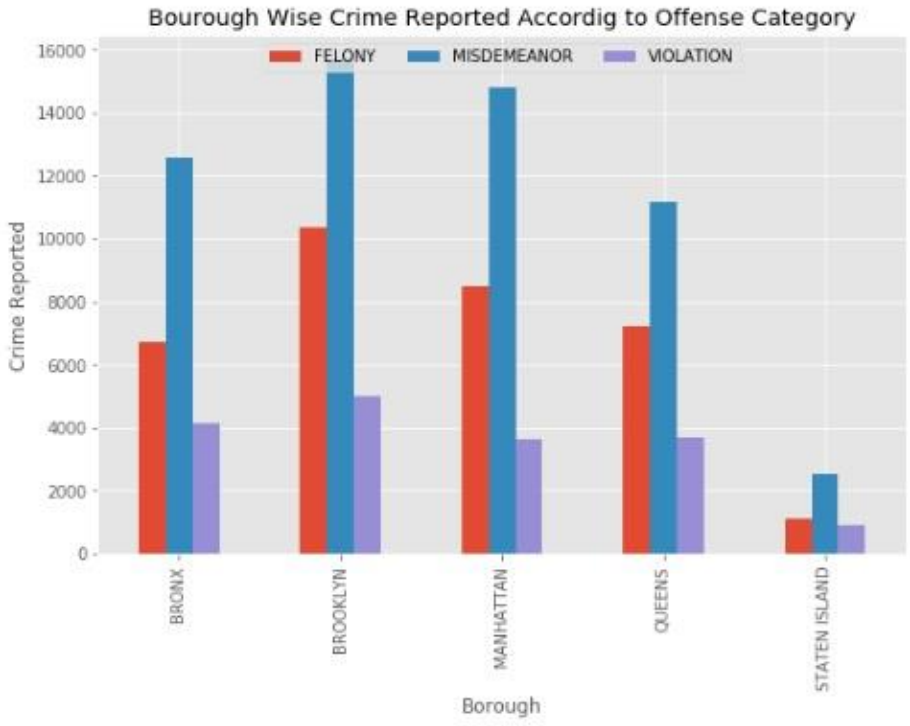
Real world data was collected for NYPD complaint records and manipulate it to get the borough wise crime detail. Although this data consists of more than 108K records but while exploring, it is being found that some of the records have missing values which was later removed to clean the dataset.



After necessary cleansing and processing aggregation functions, data has been put into the bar graph for better understanding.

To get the accurate picture of crime, it is important to see it at offense level. So, for further evaluation and analysis the dataset was grouped at

offense level for each borough and put it on the bar graph for better visualization as shown below.



To have a very high-level glimpse of crime data we have put it on the map by using Folium as shown blow.

After overall analysis of this step 1, we have selected two borough ***Queens & Staten Island*** for further location selection.

Step 2

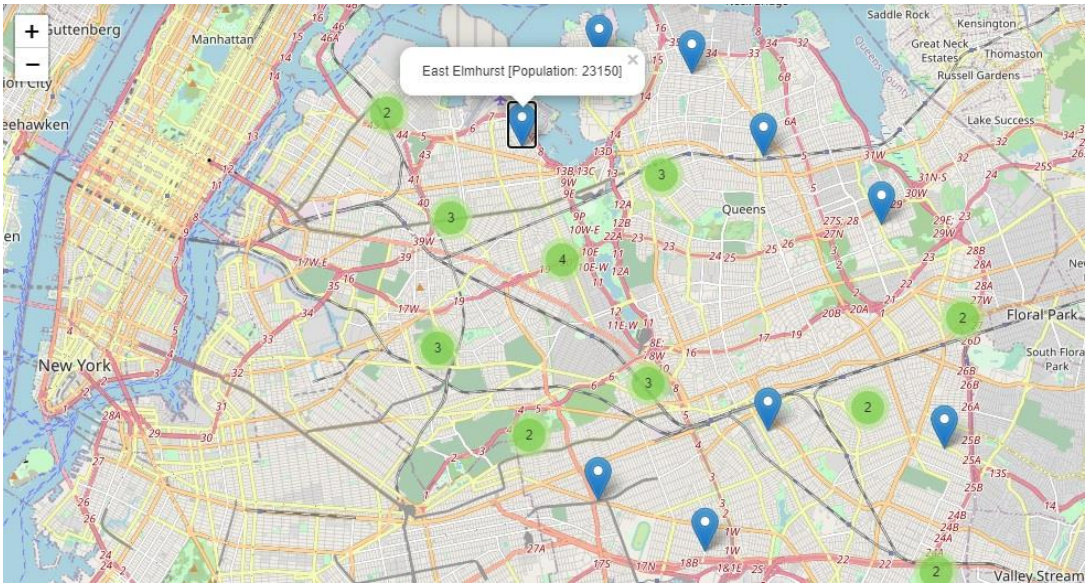
During second step we have collected the neighborhoods data of all five boroughs of New York. Along with neighborhood information for each borough this dataset has geo locations of each neighborhood also.

To keep our focus on selected two boroughs remaining data from the dataset has been deleted. After processing of 306 neighborhoods data, we got **81** for **Queens** and **63** for **Staten Island**.

Step 3

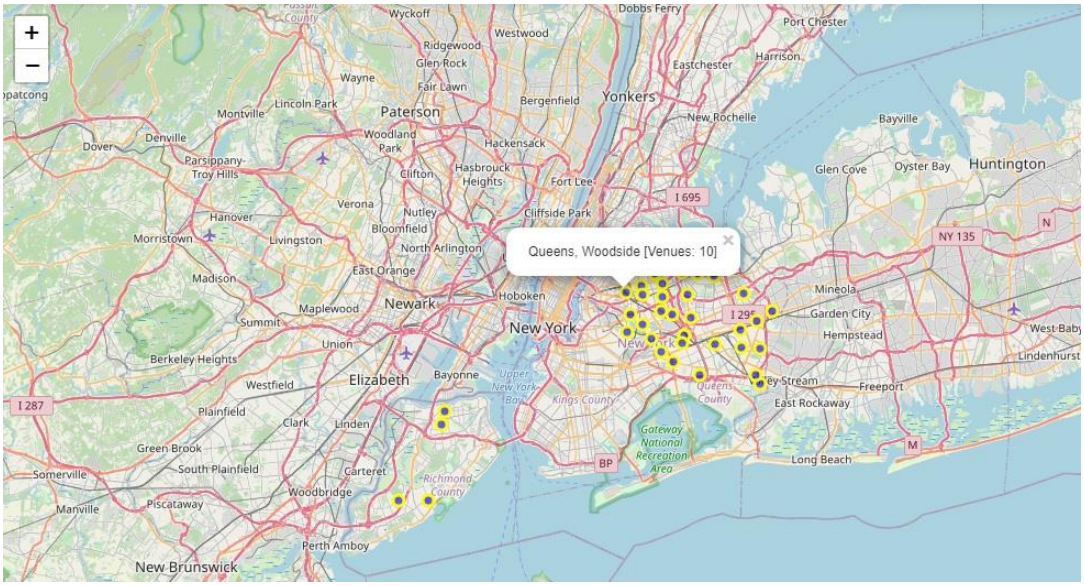
In this step we have collected NY neighborhoods population census data developed by ***NTA***. After review, it is being found that NTA has formalize this data according to their own neighborhoods planning so some of the neighborhood population data was either merged or missed in this dataset. By keeping all such lags in view, we have accepted this dataset for further proceeding. Then we have merged this population data with already developed neighborhood dataset and ***keep only those neighborhoods detail who has population detail***.

Just to have a high glimpse of this data we have put it on following Folium map.



Step 4

In this step we have used ***Foursquare API*** and collect all ***Food and Drink Shops*** detail including Grocery Stores, Supermarket etc. only for shortlisted neighborhood dataset. This venue detail then merged with the neighborhoods dataset to build a comprehensive locations dataset. We have put zero against those neighborhoods whose data was not received using Foursquare API.

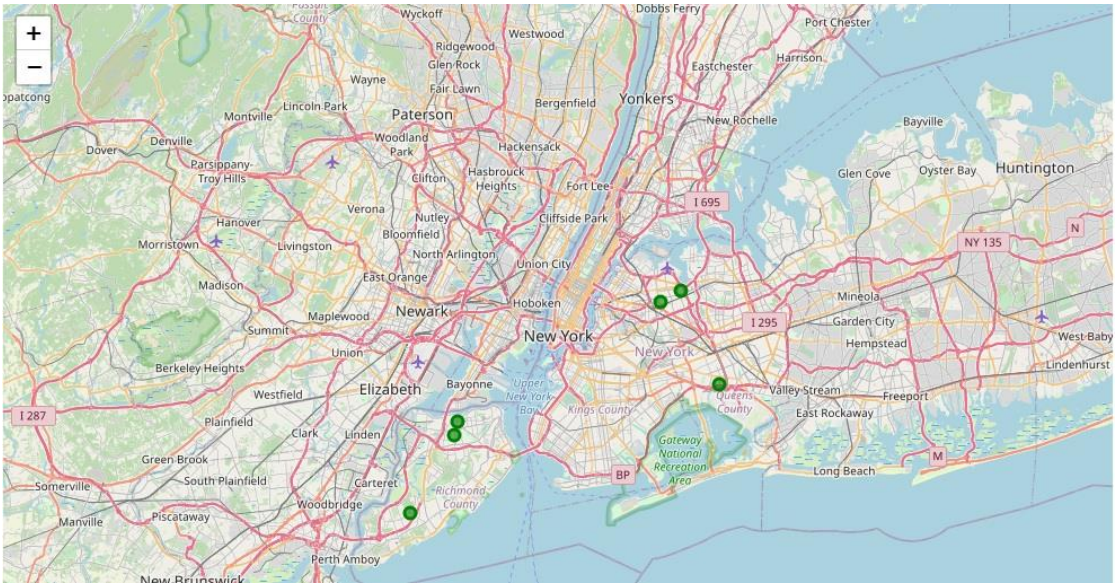


Result:

After processing all above steps we are got a comprehensive dataset of neighborhoods in the boroughs having lowest crime rate along with population and venues detail. Out of this dataset we will propose three optimized locations from each borough to the stakeholder for further review and selection. Proposed six locations are as follows:

1. Queens, Elmhurst
2. Queens, South Ozone Park
3. Queens, North Corona
4. Staten Island, Arden Heights
5. Staten Island, Westerleigh
6. Staten Island, Port Richmond

We have also plot above six locations on map to have a better view of locations.



Discussion:

Purpose of this project was to identify the neighborhoods in lowest crime rate boroughs of New York. The available population dataset and Foursquare API result was not that much satisfactory but being accepted to aid stakeholders in narrowing down the search for optimal location for a new grocery store.

By calculating complaint records, first we have identified boroughs with lowest crime rate and then generated collection of neighborhoods in these boroughs combined with their respective population data. Finally, after analysis and evaluation we have recommended six promising locations to the stakeholder for further review and decision making.

Conclusion:

Final decision on optimal store location will be made by stakeholders based on specific characteristics of neighborhoods and locations in every recommended zone, taking into

consideration additional factors like attractiveness of each location (proximity to park or water), levels of noise / proximity to major roads, real estate availability, prices, social and economic dynamics of every neighborhood etc.

Final Note:

All the above analysis is depended on the adequacy and accuracy of the NTA Population and Foursquare API data. A more comprehensive analysis and future work would need to incorporate data from other external databases.