The battle of neighbourhood

1. Introduction/Business Problem

New York City, the most populous city in the United States. With an estimated 2018 population of 8,398,748 distributed over about 302.6 square miles (784 km2), New York is also the most densely populated major city in the United States. As one of the most popular cities around the globe. New York city attract tons of tourists annually. There are a different variety of cuisines for the tourists to choose.

This study is to help some small individual investors who would like to open up pharmacy shop at New York City. We will utilize the data to gather the information to identify which place would be the ideal place for a newly established pharmacy shop. The analysis will also include a bunch of parameters which the investors need to consider for decision-making.

2. Data

In this report, we will be using the following data:

New York City data that contains list Boroughs, Neighborhoods along with their latitude and longitude.

Data source : https://cocl.us/new\_york\_dataset

Description : This data set contains the required information. And we will use this data set to explore various neighborhoods of new york city.

pharmacy shops in each neighborhood of new york city.

Data source : Fousquare API

Description : By using this api we will get all the venues in each neighborhood. We can filter these venues to get only coffee shops.

GeoSpace data

Data source : https://data.cityofnewyork.us/City-Government/Borough-Boundaries/tqmj-j8zm

Description : By using this geo space data we will get the New york Borough boundaries that will help us visualize choropleth map.

3. Methodology

BeautifulSoup to scrape information and parse data from the Wikipedia page, so that I can get an organized table of neighborhoods information of New York City.

In addition, Geopy is used to get the exact geological location of the neighbourhood.

Furthermore, we use Foursqaure API to search venue information of the neighbourhood to gain more understanding of each neighbourhood.

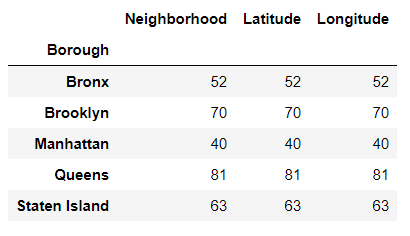
Finally, we use Kmeans to cluster similar neighbourhood to reach our conclusion.

4. Analysis

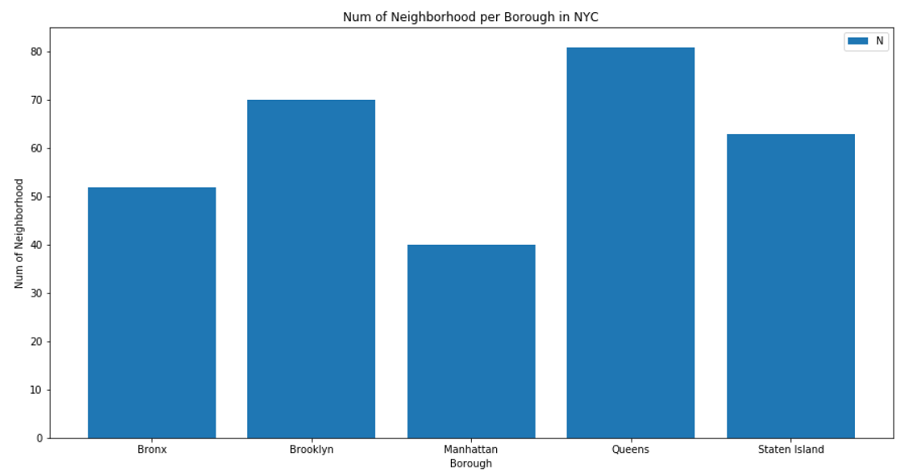
Firstly, we scrape the information from the Wikipedia to get the below dataframe.



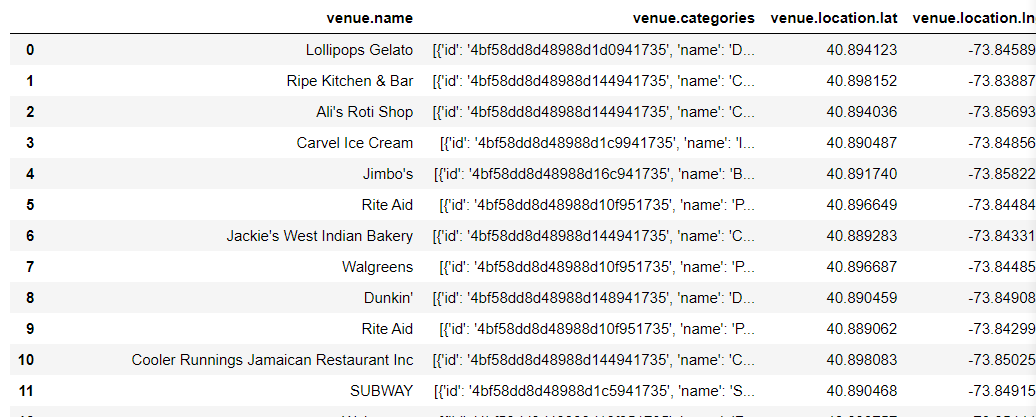
After we calculate the sum of neighbourhoods, we found that Queens has the highest number of neighbourhoods.



Below is a bar graph visual.

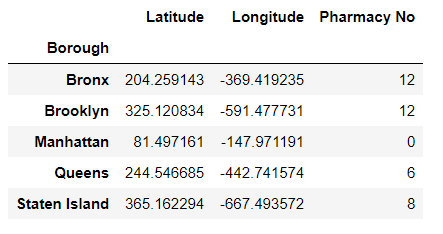


After that, we interact the Foursquare API to gain more geographic data as below.



After cleaning the data for a little bit and grouping data. We can now see how much pharmacy shops are within 1000 m of each neighbourhood.

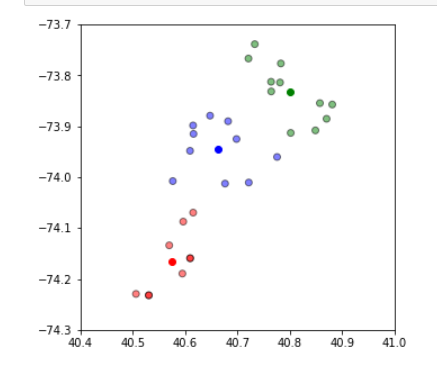
We took a random sample of 30 and found that Bronx has the highest number of pharmacy shops.



We can also distribute the sample on the map.



By using Kmeans clustering, we can cluster the 30 neighbourhood into below scatter plot.



5. Conclusion

In the random sample, we found that Bronx and Brooklyn have the highest number of pharmacy shops. However, in the scatter plot, Bronx and Brooklyn belong to two different clusters. As Brooklyn has a higher number of neighbourhoods, we recommend that investors may prefer to open a pharmacy shop in Brooklyn.