1. INTRODUCTION

1.1. BACKGROUND

The primary function of the pharmaceutical and medicine manufacturing is to create value by discovering and producing effective medicines, vaccines and services that improve patients' well-being, and can be sold in markets at a profit. As well as increasing shareholder value, this contributes significantly to the quality and protection of life and helps make the world a better place.

In the past 60 years, innovation and technology have driven huge improvements in global health. According to the Human Development Report 2010, growth in life expectancy that took over 300 years to achieve in developed countries has been secured by developing countries in just half a century, thanks largely to innovations in medicine and other public health interventions.

Pharmaceutical manufacturing is a niche and tough business. Nevertheless, like any other business, it has its own time-tested business models that have allowed numerous companies to be successful in this field. With each model, numerous economic decisions are made that may not always be in sync with engineering decisions. In addition, as times change, newer business models are introduced.

The days of big pharmaceutical companies owning many manufacturing plants are slowly ending, and every company, no matter what size, is looking for ways to lessen costs to survive in an increasingly tough and unpredictable economic climate. Therefore, pressure is increasing on all disciplines in pharmaceutical development to optimize their operations. One of the best ways to make sure that your pharmaceuticals factory is a success is to make sure that it has economic advantage in terms of transportation costs, customers and location of the factory.

New York City is the most diverse and populated city in the United States. In New York, you'll find almost every race and language you can think of under the sun, hence the common moniker, "the melting pot of the world." The city is made up of 5 boroughs: Manhattan, Brooklyn, Queens, the Bronx and Staten Island, all of which are grouped together into a single city.

New York City is widely recognized as the global center for the financial services industry. It's also the heartbeat of the American publishing industry (traditional anyway), media, entertainment (along with California), telecommunications, law and advertising industries. The city houses the main headquarters for many global corporations and business entities. The city is not popular of having pharmaceutical manufactures. It is widely known that pharmaceutical manufacturing can be housed indoors in contrast to other chemical plants like petrol manufacturing. Starting a pharmaceutical manufacturing factory in New York can be advantageous for investors as the competition in the city is minimal.

1.2. PROBLEM STATEMENT

Figure 1.2.1: Pharmaceutical products

Pharmaceutical manufacturing companies proof to be very crucial in the survival and the wellbeing of life. In the corona virus pandemic we have come to realize the importance of having large and efficient reliable supply of pharmaceuticals/medication. Multiple medium pharmaceutical factories can be very important in helping to reduce scarcity of medication and pharmaceutical products. This is very crucial especially in densely populated cities such as New York City. These pharmaceutical factories will not only help with the supply of medication and pharmaceutical products, but will also increase job availability in the densely populated city. The objective of this study is to explore neighborhoods in New York City to find the best location to open a small to medium scale pharmaceutical/Medicine manufacturing company.

The company should have a wide variety of employees in terms of cultural diversity to promote a variety of perspectives, increased creativity, global impact, improved performance and boost the brands reputation. Cultural diversity will allow ethical way of medical research by allowing diverse employees to express their different beliefs and views. The company will have a global impact by making sure that cultural diversity allows the delivery of different kinds of medication practiced across the world, this will be advantageous to the company by allowing these diverse employees to come up with new ways of manufacturing and researching medicinal products and have wide variety of medicine for the diverse population of New York city.

1.3. TARGET AUDIENCE

This report will be of most value to the following classes of people:

1. Investors who are looking to open a startup or medium pharmaceutical/medicine production company.
2. Doctors who are looking to open up a practice next to the area that has many pharmacies within a smaller radius.
3. Investors who would like to open pharmacies in an area that does not have many pharmacies which reduces competition.
4. Government officials or private companies looking to open a hospital in an area that will be convenient for patients to obtain wide variety of medication from different kinds of pharmacies.

5. Job seekers who are looking to work in pharmacies, this study will show them areas which have dense distribution of pharmacies for them to go and job hunt at.

2. DATA DESCRIPTION

2.1. NEW YORK CITY'S DEMOGRAPHIC, RACE AND ETHNIC DATA

The data that will be used is New York City's population race and ethnicity data. The demographic data of New York City will also be used to explore New York City. They will be found at the below links:

Data source 1: https://en.wikipedia.org/wiki/New_York_City

Data source 2: https://en.wikipedia.org/wiki/Demographics_of_New_York_City

Web scraping techniques were used to get the New York City's race and ethnicity data of the population and the population's demographic data from Wikipedia. This data will be used to study the population density of neighborhoods and boroughs in New York City, the cultural diversity of the neighborhoods and boroughs will also be studied. This information will motivate the location of the pharmacy by looking at diversity and population density which is related to number of customers to pharmacies and therefore allow pharmacies to buy more often from the pharmaceutical factory.

New York City's population data was scraped from Wikipedia and cleaned to obtain a data frame with borough, county, population, gdp per capita, land per square kilometer and persons per square kilometer.

New York city's population race and ethnicity demographic data was scraped from Wikipedia and cleaned to obtain a data frame that has jurisdiction which is the borough, percentage distribution of white, black or African American, Asian, mixed, Hispanic Latino and other races.

2.2. NEW YORK CITY'S FOURSQUARE API DATA

The second set of data that will be used is the New York City's data containing list of boroughs, neighborhoods along with their latitude and longitude. The data will be obtained from FourSquare API. Four Square API will also be used to obtain all the pharmacies in New York City and their location data.

The second set of data from FourSquare API will be used to study pharmacies locations in New York City to determine the areas with high number of pharmacies. This information will allow the determination of the best pharmaceutical factory location for the small to medium scale factory to ensure the factory has many pharmacies nearby to reduce operating costs such as transportation and have many pharmacy to supply to at a lower cost of factory transport operation.

Neighborhoods of New York City were obtained using FourSquare API. The New York City data was cleaned to get a list of Boroughs, Neighborhoods along with their latitude and longitude.

The venue details of the pharmacies such as count, rating, tip counts for a given pharmacy id were pulled to determine customer support of the pharmacies to get an idea of which pharmacies have higher customer support.

A data frame containing a list of pharmacies in each neighborhood was pulled from foursquare and merged with the boroughs data frame on boroughs to plot them on a map by borough in order to determine the best possible location to locate a pharmaceutical factory.

3. EXPLORATORY DATA ANALYSIS AND DISCUSSION

3.1. Land area of boroughs in New York City

```
[6]: import pandas as pd

import matplotlib.pyplot as plot

# Draw a vertical bar chart

df.plot.bar(x="borough", y=["land_sqkm"], rot=70, title="Area of the Boroughs in New York City in SqKm");

plot.show(block=True);
```

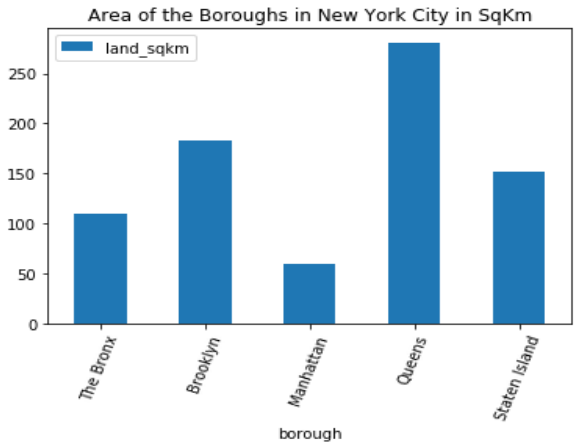

Figure 3.1.1: Area of the boroughs in New York City

Queens has the highest area of all the boroughs in New York City. Brooklyn has the second largest area in the boroughs from New York City. This means that the two best areas to build a pharmaceutical/Medicine manufacturing plant in New York city is at Brooklyn and Queens, these two boroughs have enough area to find the best plot/building to build a pharmaceutical company, there is a high chance of finding large area building or plots in Queens and Brooklyn to build a pharmaceutical/medicine manufacturing company that will allow for feature expansion in the case that manufacturing expansion is required due to high demand of pharmaceuticals/medicine. Using the Area of boroughs it can be seen that Queens and Brooklyn are the two winners.

3.2. Density of boroughs in New York City

```
[6]: import pandas as pd

import matplotlib.pyplot as plot

# Draw a vertical bar chart

df.plot.bar(x="borough", y=["persons_sqkm"], rot=70, title="Density of the Boroughs in New York City");

plot.show(block=True);
```

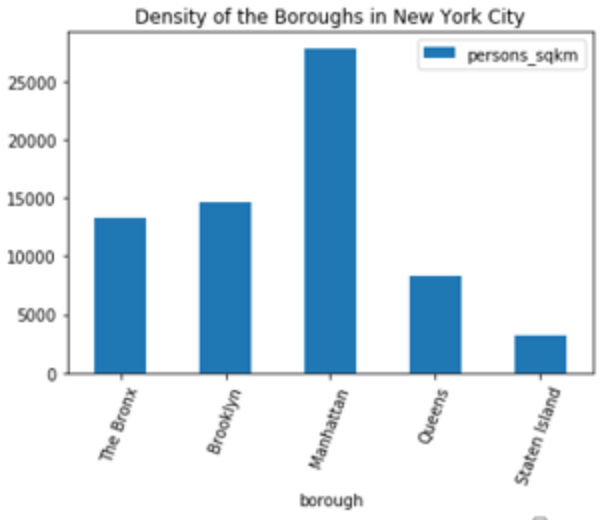

Figure 3.2.1: Density of the boroughs in New York City

The above graph shows that the density of Manhattan is very high compared to all other boroughs in New York City. This is not a good thing for a pharmaceutical factory because these kind of industry requires open areas with no dense population to allow for convenient transportation of goods to the customers. Due to the high density of Manhattan, the borough is removed from the possible location to build a pharmaceutical company. The borough will not cater for possible expansion of the factory in future if required.

3.3. Population amount of the boroughs in New York City

[7]:
```
import pandas as pd

import matplotlib.pyplot as plot

# Draw a vertical bar chart

df.plot.bar(x="borough", y="population", rot=70, title="Population of the Boroughs in New York City");

plot.show(block=True);
```

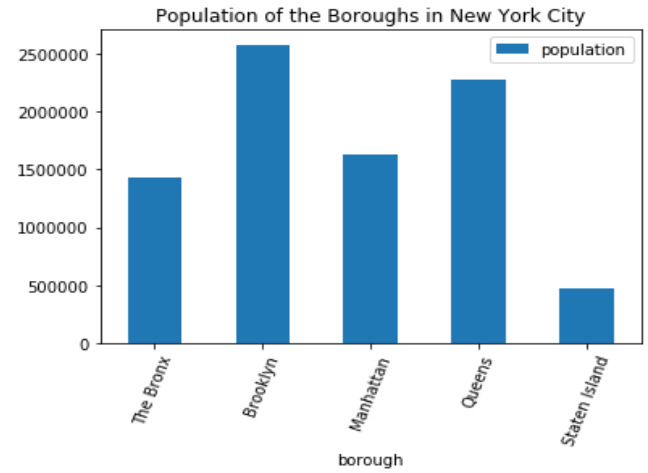

Figure 3.3.1: Population amount of the boroughs in New York City

Brooklyn has the highest population in the boroughs in New York City. Queens has the second highest population in the New York cities' boroughs which is 12% less than Brooklyn's population. The population of Brooklyn is 12% higher than its second lowest runner up which makes it ideal to open a business. There will be enough customers. Here I do not conclude this by population alone as the direct customers to the pharmaceutical/medicine manufacture will not directly be the people but pharmacies. Pharmacies are used for this determinant as they buy in large quantities compared to shops and have a 100% certainty of selling medicine/pharmaceuticals in contrast to the normal shops.

3.4. Simpson's diversity index of the boroughs in New York City

Statistical exploration of the boroughs in New York City was done using Simpson's Diversity Index to determine the most diverse boroughs in New York City to locate the pharmaceutical factory. The following bar graph is the Inverse Simpson's diversity index for the boroughs. The higher the index the more racially diverse is the borough.

[124]:
```
import matplotlib.pyplot as plt
import numpy as np

plt.rcdefaults()
fig, ax = plt.subplots()

# Example data
y_pos = np.arange(len(b_df['Borough']))
boroughs = ('Brooklyn', 'Queens', 'Manhattan', 'Bronx', 'Staten Island')

ax.barh(y_pos, b_df['Inverse_Simpsons_Index'])
ax.set_yticks(y_pos)
ax.set_yticklabels(boroughs)
ax.invert_yaxis()  # labels read top-to-bottom
ax.set_xlabel("Inverse Simpson's Diversity Index")
ax.set_title('Diversity measurement of the boroughs in New York City')

plt.show()
```

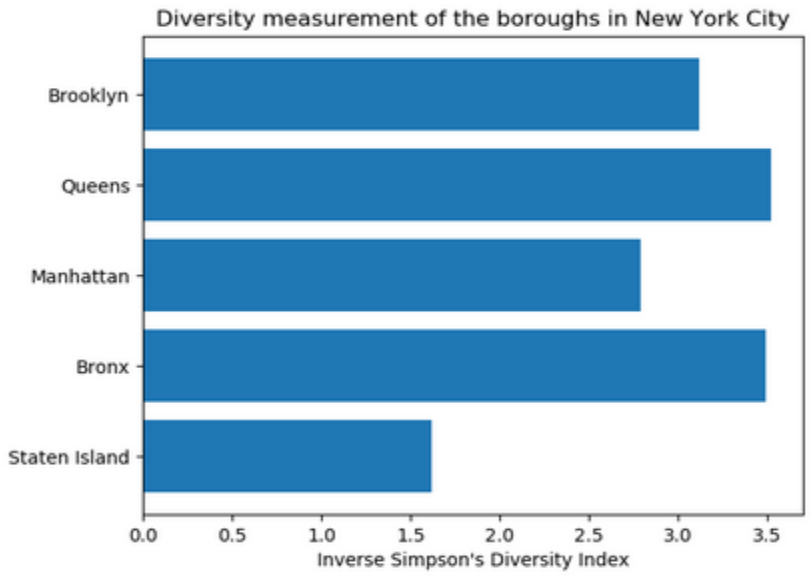

Figure 3.4.1: Diversity measurement of the boroughs in New York City

The top 3 boroughs with higher cultural diversity is Bronx, Queens and Brooklyn in descending order using the Inverse Simpson's diversity index. The highest possible index in this study is 5 since we have 5 races

and the higher the index the more diverse the borough. It can be seen that Bronx and Queens are the ideal boroughs to locate the factory followed by Brooklyn.

3.5. The Number of Neighborhoods in the boroughs of New York City

Figure 3.5.1: Number of neighborhoods in the New York City's boroughs

Foursquare API was used to obtain all neighborhoods and group them by their boroughs in order to be able to plot the bar graph of number of neighborhoods in each borough. It can be seen that Bronx has the highest number of neighborhoods followed by Staten Island and Queens. Staten Island is removed from the possible locations to build a factory, it has the lowest population of all the boroughs as it is discussed in 3.3. Therefore, although Staten Island has the third highest number of neighborhoods, it has the lowest population number which makes it not ideal to open a business in.

If a factory is built in Staten Island, the many pharmacies will not buy stock frequently as compared to Bronx, Queens and Brooklyn as they have higher population numbers. In this category the top three boroughs considered for building a factory are Bronx, Queens and Brooklyn as they are the top 3 boroughs with higher number of neighborhoods when excluding Staten Island. The graph above shows that there is a total of 306 neighborhoods in New York City which might all have differing numbers of pharmacies. The next step is to write a code that pulls all pharmacies in each borough.

3.6. Pharmacies in the New York City's Neighborhoods

The Following code extracts all the pharmacies in all the boroughs of New York City. The data frame obtained has the borough, neighborhood, pharmacy ID and the pharmacy ID.

The Following code extracts all Pharmacies in the New City's Neighborhoods

[23]:
```
# Preparing neighborhood list that contains Pharmacies
column_names=['Borough', 'Neighborhood', 'ID','Name']
pharmacy_ny=pd.DataFrame(columns=column_names)
count=1
for row in new_york_data.values.tolist():
    Borough, Neighborhood, Latitude, Longitude=row
    venues = get_venues(Latitude,Longitude)
    pharmacies=venues[venues['Category']=='Pharmacy']
    #print('(',count,'/',len(new_york_data),')','Pharmacies in '+Neighborhood+', '+Borough+':'+str(len(pharmacies)))
    for Pharmacy_detail in pharmacies.values.tolist():
        id, name , category=Pharmacy_detail
        pharmacy_ny = pharmacy_ny.append({'Borough': Borough,
                                          'Neighborhood': Neighborhood,
                                          'ID': id,
                                          'Name' : name
                                         }, ignore_index=True)
    count+=1
```

The following data frame contains all Pharmacies, their borough, Neighborhood, ID and Pharmacy name

[24]:
```
pharmacy_ny.tail()
```

[24]:

	Borough	Neighborhood	ID	Name
500	Queens	Bayswater	50058548a4b07c19da39affd	Vista Pharmacy & Surgical
501	Staten Island	Fox Hills	4b2aa393f964a5200aa424e3	CVS pharmacy
502	Staten Island	Fox Hills	4ca97b54595cb1f723a8ca14	Rite Aid
503	Staten Island	Fox Hills	4df38b0522718759f82017a1	Millers pharmacy
504	Staten Island	Fox Hills	4c4722f71ddac9289a639c32	Classic Pharmacy, Inc.

[25]:
```
pharmacy_ny.shape
```

[25]:
```
(505, 4)
```

The tail method was used to see the last 5 pharmacies in the data frame. It can be seen that the number of pharmacies in New York City pharmacies data frame is 506. After careful exploration of the data frame, it was found that it contains duplicates. The duplicates were removed and the total number of pharmacies were reduced to 184. This is a very good customer base for a small to medium pharmaceutical factory. The next important thing to do is to find the neighborhood(s) that has higher density of pharmacies. The results of this will be used in conjunction with all the findings of 3.1 – 3.5 above to decide on the area that is best suitable to locate the pharmaceutical factory.

To get the neighborhoods with higher density of the pharmacies, all the pharmacies were grouped by neighborhood. This allowed to sort the neighborhoods by the top neighborhoods which have many pharmacies. The borough of these neighborhoods were found and then compared to the results of the analysis in 3.1 – 3.5 to see if the results are resonating with each other. The graph below shows the top 5 Neighborhoods with high number of pharmacies.

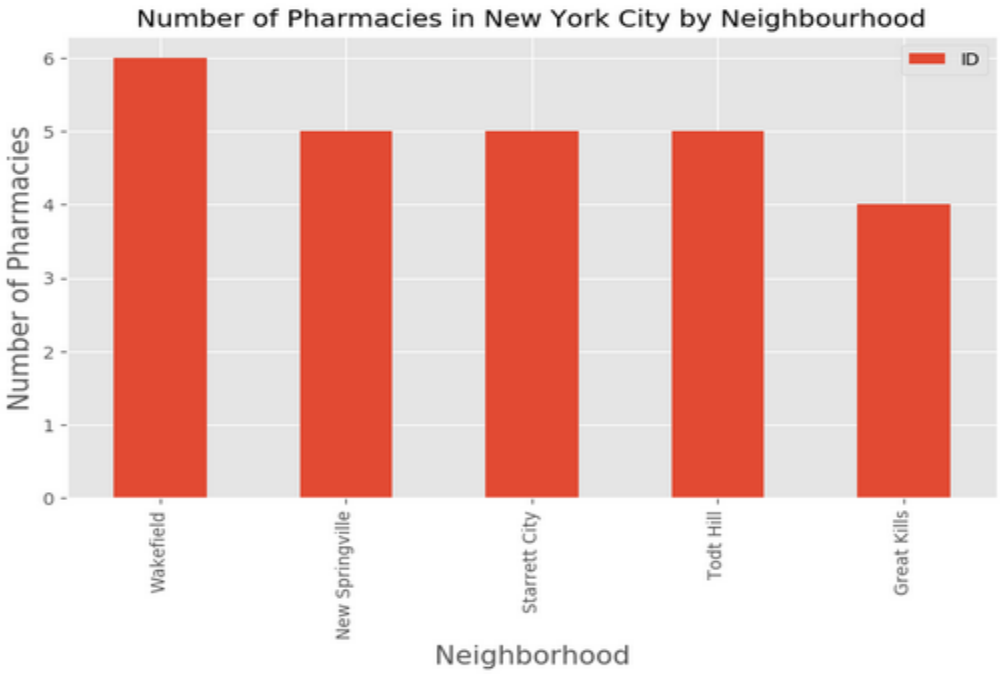

Figure 3.6.1: Number of pharmacies in the New York City's top 5 neighborhoods

It can be seen that the top five neighborhoods from the above graph all have 6 or less number of pharmacies. These neighborhoods are plotted on the map to compare their distribution in the below map. The top 5 neighborhoods contain 26 pharmacies all together which is relatively small in contrast to the total of 184 pharmacies. Due to pulling the pharmacies using the neighborhood, there is possibility of having duplicate pharmacies even though the ID's are not the same. This problem will be solve by writing a code that will extract pharmacy coordinates and these can be used to actually remove the duplicates further as no two pharmacies can be located at the same coordinates. These are foursquare shortfalls.

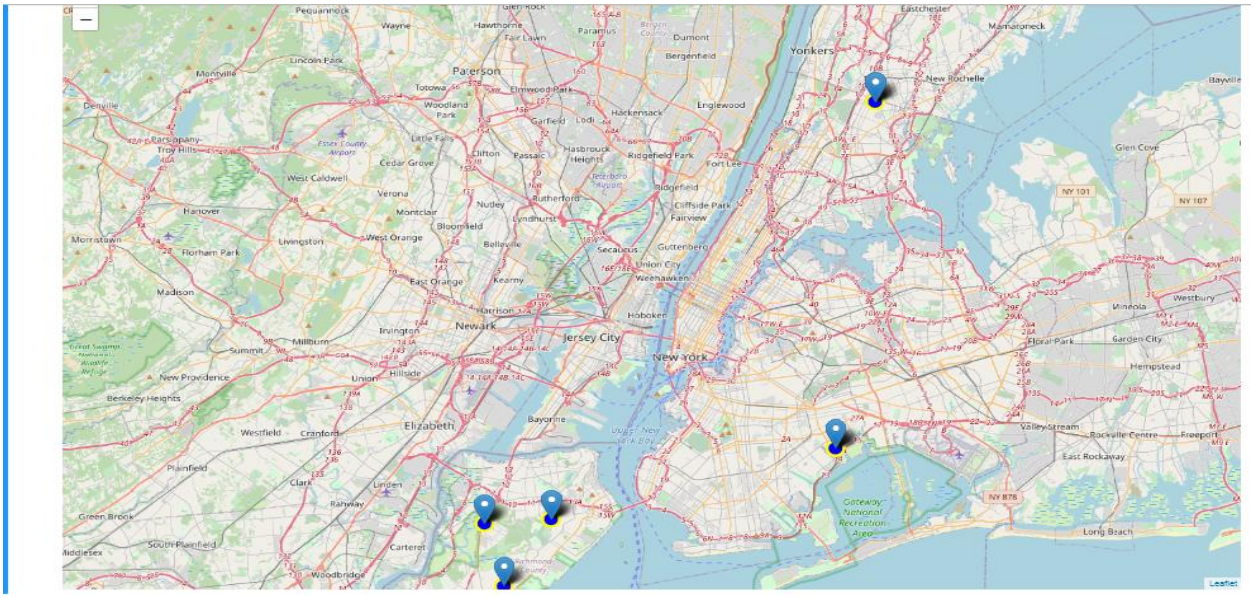

Figure 3.6.2: Map of the top 5 New York City's neighborhoods with many pharmacies

The above map shows that the top 5 neighborhoods in the bar graph above are scattered. Therefore this data cannot be used alone to determine the best location of the pharmaceutical factory. Finding the centroid of these neighborhoods and using it as the location can be misleading. Therefore, clustering of all neighborhoods will be performed using K-means to determine the more clustered region where a pharmaceutical factory will have many customers. It should be noted that a business case will be presented by considering all exploratory data analysis to motivate the final chosen location for the pharmaceutical factory.

3.7. K-means clustering of the pharmacies in New York City

K-means clustering is performed due to the inconclusive results of the top 5 neighborhoods which have many pharmacies. The neighborhoods are too scattered as shown in the map in 3.6 above and therefore cannot be relied on to determine the best dense region of New York. K-means will help in identifying the region which is dense with pharmacies.

The following is the scatter plot to visualize the location of all the pharmacies in New York City as obtained from foursquare.

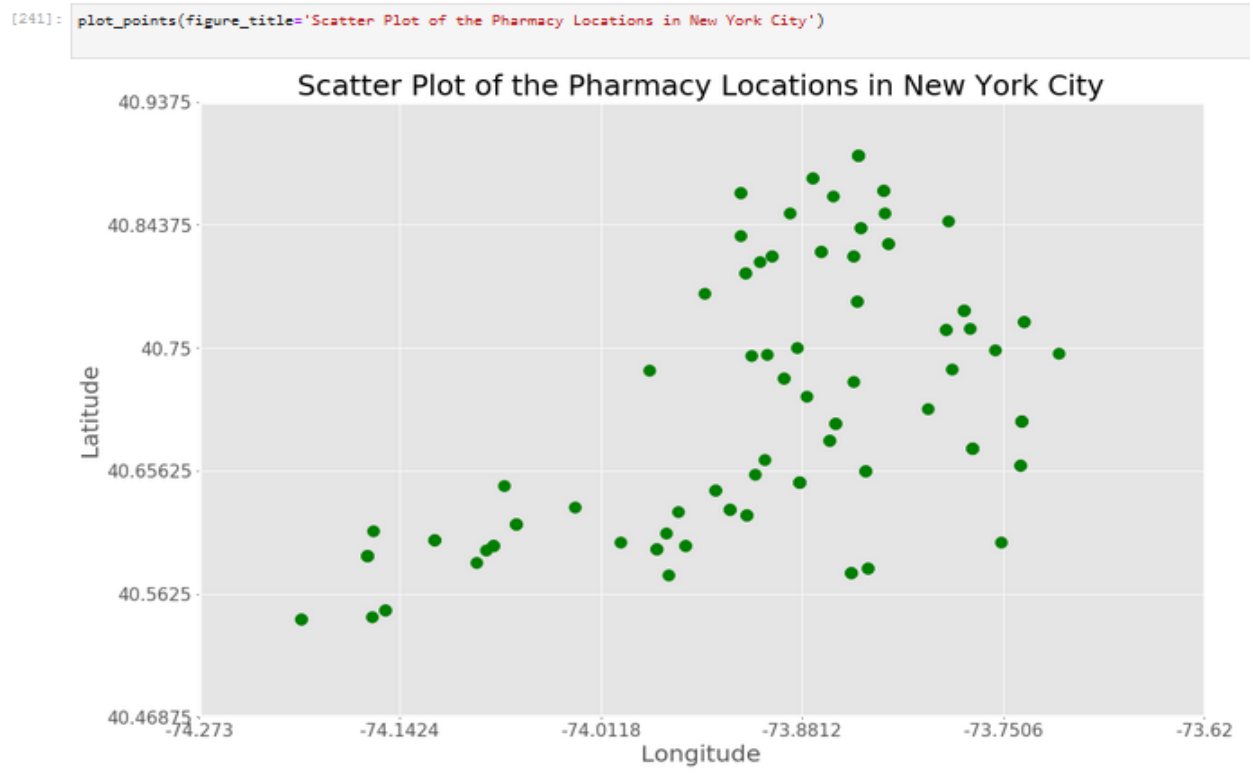

Figure 3.7.1: Scatter plot of the pharmacies in New York City

The above scatter plot will be clustered into three clusters that the pharmaceutical factory can be located in. Three clusters were used to maximize the number of factory's customer base.

A K-means algorithm was implemented to cluster the pharmacies into 3 clusters. The clusters are shown in the below clustered scatter plot visual. The cluster which has the most pharmacies for the factory to supply to is the green cluster followed by the blue cluster and then the red cluster respectively.

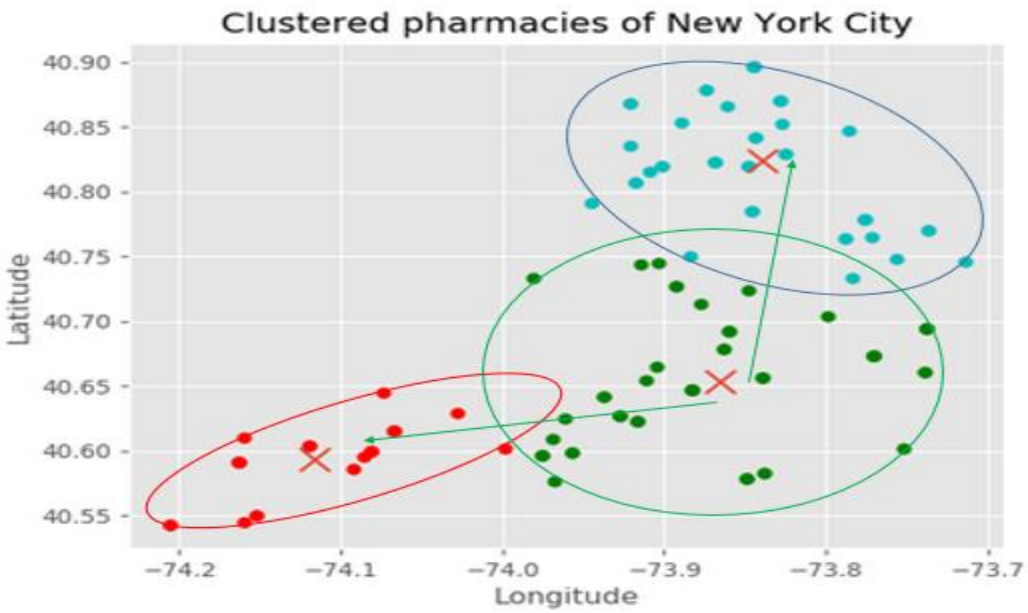

Figure 3.7.2: Clustered pharmacies in New York City using K-means flat clustering

The green cluster has higher customer base and the option to expand into the blue and the red clustered regions. This shows that the green cluster centroid shown by the x in the green cluster circle is the best location to build the factory. The factory will have higher customers base and also a room for expansion shown by the green arrows. This puts it a good advantage than being located in any of the other blue and red clusters. Therefore the best location to build the factory is given by the following coordinates [-73.8654, 40.6537] and this location is in Seaview Ave in Brooklyn borough. The second best location given by the blue cluster is given by the following coordinates [-73.8398, 40.8239] which is in Wetchester Ave in Bronx borough. The third best location is in Yellow Trail in Staten Island borough.

Figure 3.7.3: Map of all the pharmacies and the top 3 best locations of the factory shown by pop-ups

The number 1 best location figure 3.7.3 is shown by the pop up with description which is Seaview in Brooklyn according to the k-means clustering results.

3.8. Comparing the exploratory data analysis results to choose best location

The table below looks at all the exploratory data analysis performed in 3.1 – 3.7 and chooses the top two winning boroughs for each exploratory analysis with the exception of clustering which only the best cluster centroid location is chosen as decided in number 3.7. The count of borough in the "Winning Borough" column is used to get the "Borough Score" column which is then used to motivate the number best location to build a pharmaceutical factory.

	Boroughs					
Study	**The Bronx**	**Brooklyn**	**Manhattan**	**Staten Island**	**Queens**	**Winning Borough**
Land Area	Small	2nd Highest	Small	3rd Highest	Highest	Queens, Brooklyn
Population Density	3rd Highest	2nd Highest	Very High	Low	Low	Manhattan, Brooklyn
Population Number	4th Highest	Highest	3rd Highest	Low	2nd Highest	Brooklyn, Queens
Diversity	High	2nd Highest	3rd Highest	Low	High	Bronx, Queens, Brooklyn
Number of Neighborhoods	Highest	4th Highest	Low	3rd Highest	2nd Highest	Bronx, Queens
Cluster Centroid Location	Blue: Wetchester Ave	Green: Seaview Ave	-	Red: Yellow Trail	-	Brooklyn Seaview Ave
Borough Score	**2**	**5**	**1**	**0**	**4**	**-**

Table 3.8.1: Summary of the exploratory data analysis and the scoring of the locations

Based on the borough scoring it can be seen that the best location for the pharmaceutical factory is Brooklyn followed by Queens and then Bronx. This table will be used to motivate for the best location of the factory as it summarizes the results of the exploratory data analysis.

4. Conclusion

The Table in 3.8 shows that Brooklyn has the highest borough score of 5 which makes it a winner in terms of the best location to build a pharmaceutical factory. An overall view of Brooklyn shows that Brooklyn has the highest population number compared to all the boroughs. This is good because the higher the population the higher the customer base of the pharmacies and the more the pharmacies will require stock supply from the pharmaceutical factory. This point makes Brooklyn the top ideal borough to locate the pharmaceutical factory. The results of K-means clustering showed that the best cluster was the green cluster as shown and discussed in number 3.7. Green cluster is located in Seaview which is in Brooklyn. This cluster is the ideal location of locating a pharmaceutical factory with the room for expanding into the other two clusters as discussed in number 3.7.

The Land area of Brooklyn is the 2nd highest which is a reasonable because we can see that Queens has the highest land area but the lowest population density. It can be seen from table in 3.8 that Brooklyn diversity is the second highest of all the boroughs but looking at the diversity index graph in 3.4 we can see that index of Queens which is the highest is almost equal to that of Brooklyn, they do not differ significantly. This makes Brooklyn a good place as well as it will be able to cater for the diversity needs of the pharmaceutical factory. This exploratory data analysis of New York City shows that the best location to locate the Pharmaceutical factory is in Seaview Brooklyn.

5. Recommendations

The results of the exploratory study show that Brooklyn is the best location for a medium scale pharmaceutical factory. In the case that a factory is opened in Brooklyn, small investors can perform a deeper study of Queens to understand the clusters and the borough statistics. Queens looks like a second best promising location to build a second lowest scale pharmaceutical factory. A second lowest is stated explicitly here because once a medium scale factory is opened in Brooklyn, investors should not attempt to open the same scale or larger scale factory in contrast to Brooklyn factory. This is because the results of this study proof that this factory will have an enormous operating advantage and customer advantage. Queens should be considered for only smaller factories.

The Results of this study can also be used by government or private officials who are looking to build a hospital, doctors who are looking to build a practice and the investors who are looking to open new pharmacies.