The Battle of the Neighborhoods

Introduction

Background

New York is the most populous city in the United States. It is diverse and is the financial capital of USA. It is a global hub of business and commerce. The city is a major center for banking and finance, retailing, world trade, transportation, tourism, real estate, new media, traditional media, advertising, legal services, accountancy, insurance, theater, fashion, and the arts in the United States.

In 2019, New York had its first case of coronavirus disease 2019 (Covid-19) pandemic and 2020 emerged as the United States major epicenter.

Problem Description

The first case of coronavirus in New York was identified in New York City on February 29. Soon after, the city cautioned against public gatherings, and New Rochelle, New York, closed schools as the virus spread quickly there. With more than 80,000 cases and 4,260 coronavirus deaths, according to the city's website, New York is one of the major epicenters for the coronavirus outbreak in the United States.

It is evident that to survive and curb the various more testing must be done, and this study aids to shows the areas with more cases. it is very important to strategically plan how to distribute testing kits to areas gravely infected.

Why Data?

Without leveraging data to make decisions about locations, the country will end up distributing testing material to locations where they are not needed. Data will provide better answers and better solutions to their task at hand.

Criteria

The analysis will focus on cities in New York and, not on specific boroughs. Narrowing down the areas options derived from analysis allows for either further research to be conducted.

Outcome

The success criteria of the project will be a good recommendation of city in New York with the highest number of Covid 19 cases.

Methodology and Exploratory Data Analysis

The Data Science Workflow for Part 1 & 2 includes the following:

Business Understanding:

Our main goal is to get the areas in New York that are highly infected and require adequate testing

Analytic Approach:

New York city neighbourhood has a total of 5 boroughs and 306 neighborhoods. In this project first part is clustering of Manhattan and Brooklyn. And second part is clustering of Bronx, Queens, and Staten Island. This is done because of the following Exploratory data analysis.

- **Outline the initial data that is required:**
  - District data for New York including names, location data if available, and any other details required.

- **Obtain the Data:**
  - Research and find suitable sources for the district data for New York.
  - Access and explore the data to determine if it can be manipulated for our purposes.

- **Initial Data Wrangling and Cleaning:**
  - Clean the data and convert to a useable form as a data frame.

The Data Science Workflow for parts 3 & 4 includes:
Data Analysis and Location Data:
- Data manipulation and analysis to derive subsets of the initial data.
- Identifying the high traffic areas using data visualisation and statistical analysis.

Visualization:
- Analysis and plotting visualizations.
- Data visualization using various mapping libraries.

Discussion and Conclusions:
- Recommendations and results based on the data analysis.
- Discussion of any limitations and how the results can be used, and any conclusions that can be drawn

We will be analyzing **New York City**

We will be using the below datasets for analysing New York city

Data 1: Neighborhood has a total of 5 boroughs and 306 neighborhoods. To segment the neighborhoods and explore them, we will essentially need a dataset that contains the 5 boroughs and the neighborhoods that exist in each borough as well as the latitude and longitude coordinates of each neighborhood.

This dataset exists for free on the web. Link to the dataset is:
https://www.health.ny.gov/statistics/cancer/registry/appendix/neighborhoods.htm

	Borough	Neighborhood	ZIP Codes
0	Bronx	Central Bronx	10453, 10457, 10460
1	Bronx	Bronx Park and Fordham	10458, 10467, 10468
2	Bronx	High Bridge and Morrisania	10451, 10452, 10456
3	Bronx	Hunts Point and Mott Haven	10454, 10455, 10459, 10474
4	Bronx	Kingsbridge and Riverdale	10463, 10471

Data 2: New York city geographical coordinates data will be utilized as input for the Foursquare API, that will be leveraged to provision venues information for each neighborhood. We will use the Foursquare API to explore neighborhoods in New York City. The below is image of the Foursquare API data.

	Borough	Neighborhood	Latitude	Longitude
0	Bronx	Wakefield	40.894705	-73.847201
1	Bronx	Co-op City	40.874294	-73.829939
2	Bronx	Eastchester	40.887556	-73.827806
3	Bronx	Fieldston	40.895437	-73.905643
4	Bronx	Riverdale	40.890834	-73.912585

Data 3: Second data which will be used is the data from NYC Health.
Website- https://github.com/nychealth/coronavirus-data/blob/master/boro.csv

BOROUGH_GROUP	COVID_CASE_COUNT	COVID_CASE_RATE
The Bronx	40148	2728.8
Brooklyn	46977	1732.54
Manhattan	21862	1164.06
Queens	54558	2183.62
Staten Island	12452	2477.13
Citywide	176086	

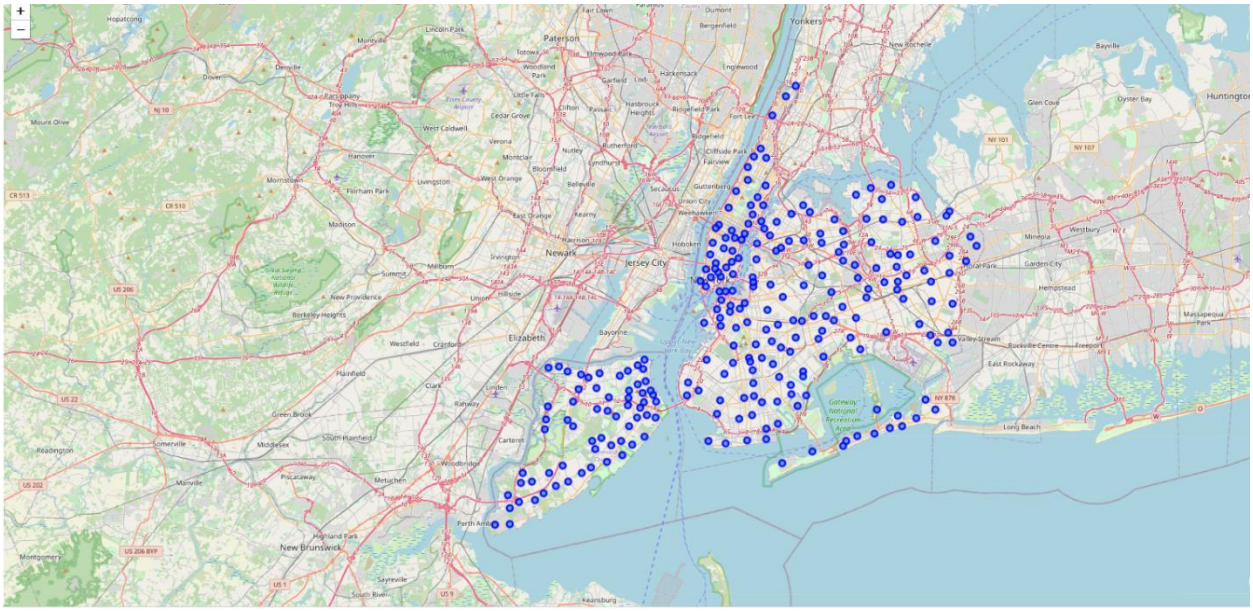
Exploratory Data Analysis:

Data 1- New York city Geographical Coordinates Data.

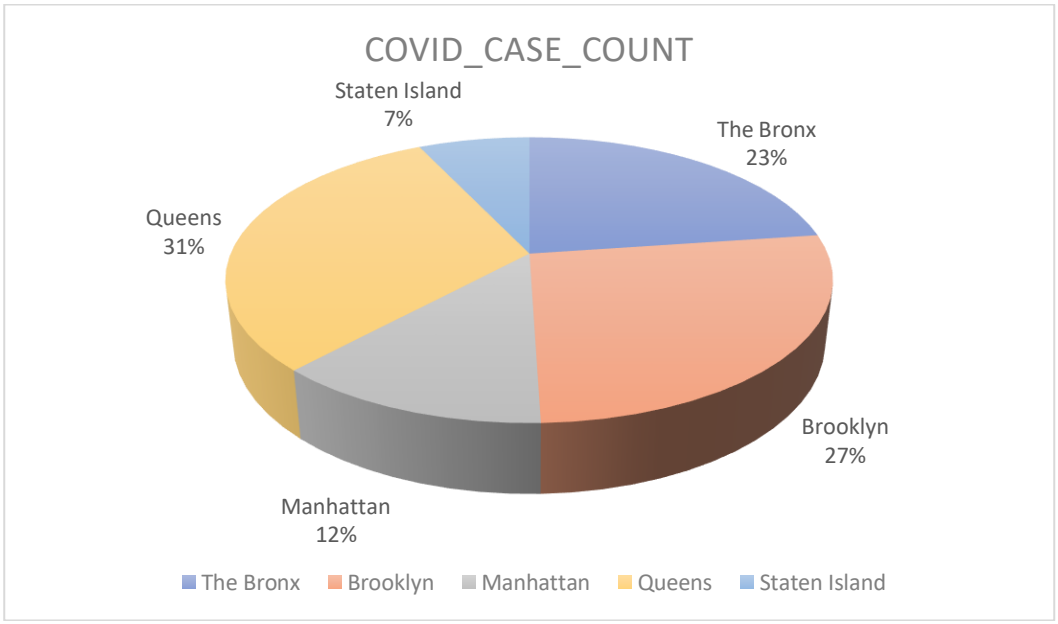
1. In this we load the data and explore data from newyork_data.json file.
2. Transform the data of nested python dictionaries into a pandas data frame.

3. This data frame contains the geographical coordinates of New York city neighborhoods.
4. We used geopy and folium libraries to create a map of New York city with neighborhoods superimposed on top.

New York neighbourhood visualization showing areas affected by Covid 19

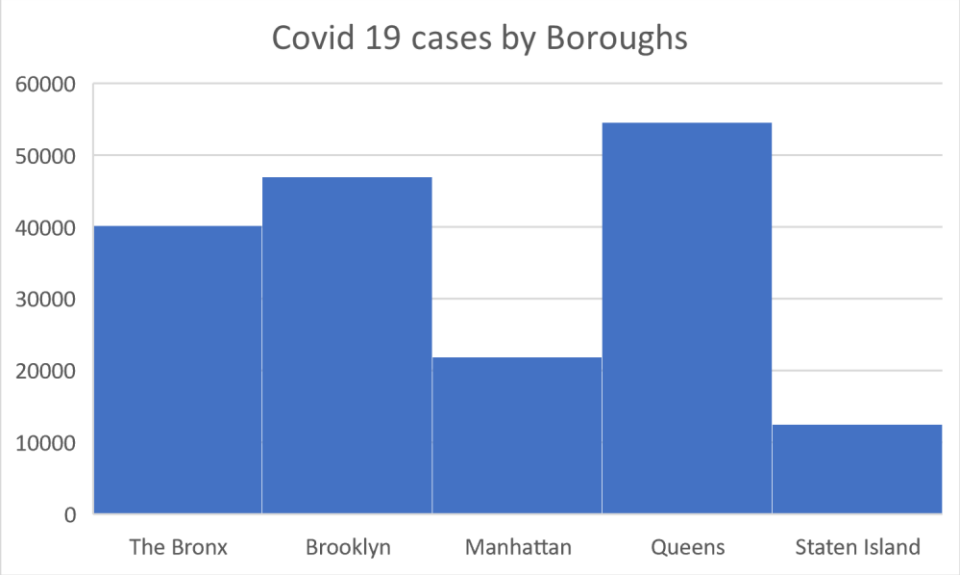


Visualization showing percentage affected by Covid 19



Covid 19 cases by Boroughs

This chart shows the number of positive cases COVID-19 on a daily basis since March 3.



RESULTS

From this venues data we filtered and used only the data that aligns with the Covid report eliminating the Zip codes clustering. As we focussed only on areas with high volumes.

Neighborhood K-Means clustering based on mean occurrence of venue category:

To cluster the neighborhoods into two clusters we used the K-Means clustering Algorithm. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. It uses iterative refinement approach.

DISCUSSION

1. There is scope to increase testing in Queens, Brooklyn and The Bronx
2. There is a huge possibility that rate of exposure in Queens, Brooklyn and The Bronx is high.
3. Very few data have been acquired from neighborhood in Staten Island

CONCLUSION

This analysis is performed on limited data. This may be right or may be wrong. But if good amount of data is available there is scope to come up with better results. If there are lot of infection probably testing will be in high demand. Queens, Brooklyn and The Bronx has high concentration of infected persons. Very competitive market. Bronx, Queens and Staten Island also has had a substantial amount of testing but not as many as required. So, this can be explored.