# APPLIED DATA SCIENCE CAPSTONE

A STUDY OF COVID-19

INTRODUCTION

- Coronavirus disease 2019 (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus (SARS-CoV-2) that outbroke in December 2019.

- As of April 27th, 2020
  - > 3 million around the world were infected
  - > 1 million Americans were infected

- It is important for the public to understand what factors are related to the number of COVID-19 cases, so that effective measures can be taken to reduce the risk of exposure to coronavirus.

- Factors that are explored in this project;
  - Population
  - Population density
  - Venue types in the neighborhood
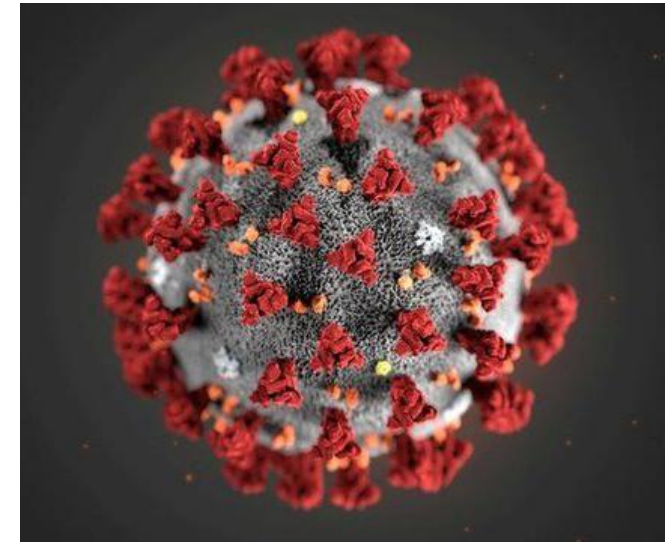
DATA ACQUISITION

- Data Sources:
    - Wikipedia: https://en.wikipedia.org/wiki/List_of_counties_in_New_York
    - New York State government website: https://health.data.ny.gov/browse?tags=covid-19
    - NYC Department of Health and Mental Hygiene GitHub: https://github.com/nychealth/coronavirus-data
    - GitHub open sources
    - Foursquare API
- Methods:
    - BeautifulSoup
    - Requests.get

DATA EXPLORATION
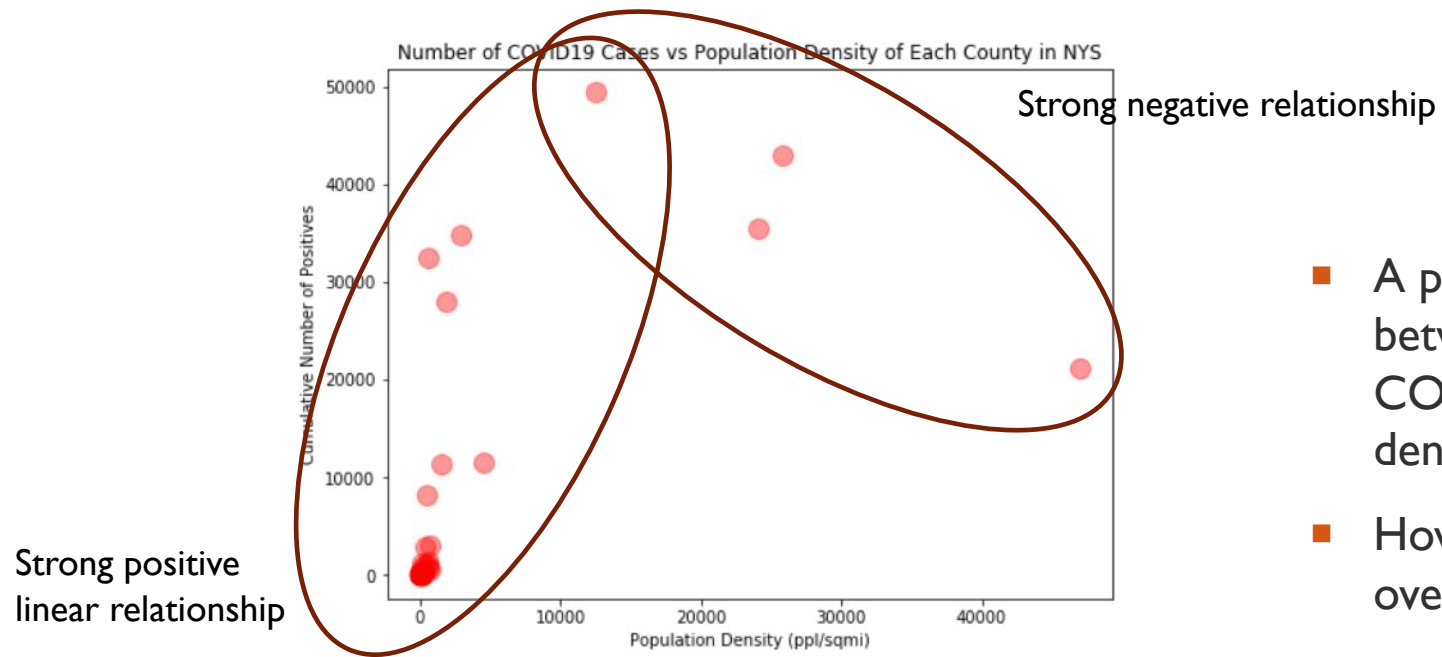
Strong positive linear relationship

Strong negative relationship

Figure 1. Number of COVID-19 Cases vs Population Density of Each County in NYS

- A positive relationship can be observed between the total number of positive COVID-19 cases and the population density of each county.

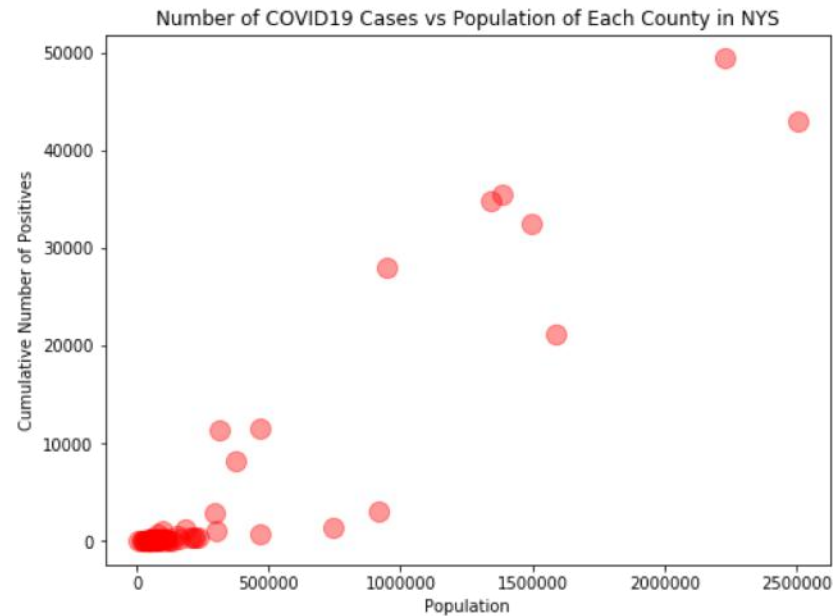- However, data shows weak linearity overall.

DATA EXPLORATION

Figure 2. Number of COVID-19 Cases vs Population of Each County in NYS

- A positive relationship can be observed between the total number of positive COVID-19 cases and the total population of each county.

- Data shows relatively strong linearity.
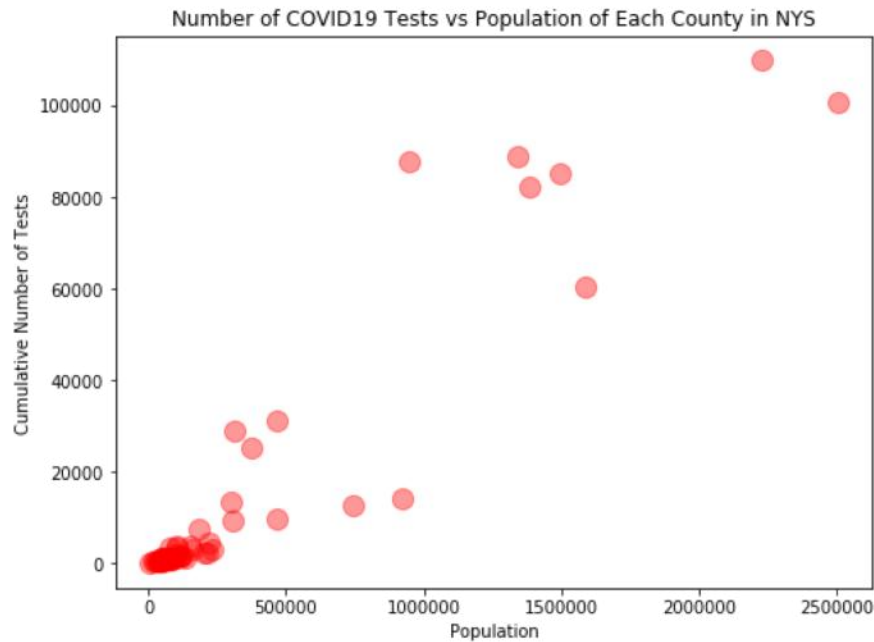
DATA EXPLORATION

Figure 3. Number of COVID-19 Tests vs Population of Each County in NYS

- Similar to the previous plot, a positive relationship can be observed between the total number of positive COVID-19 cases and the total population of each county.

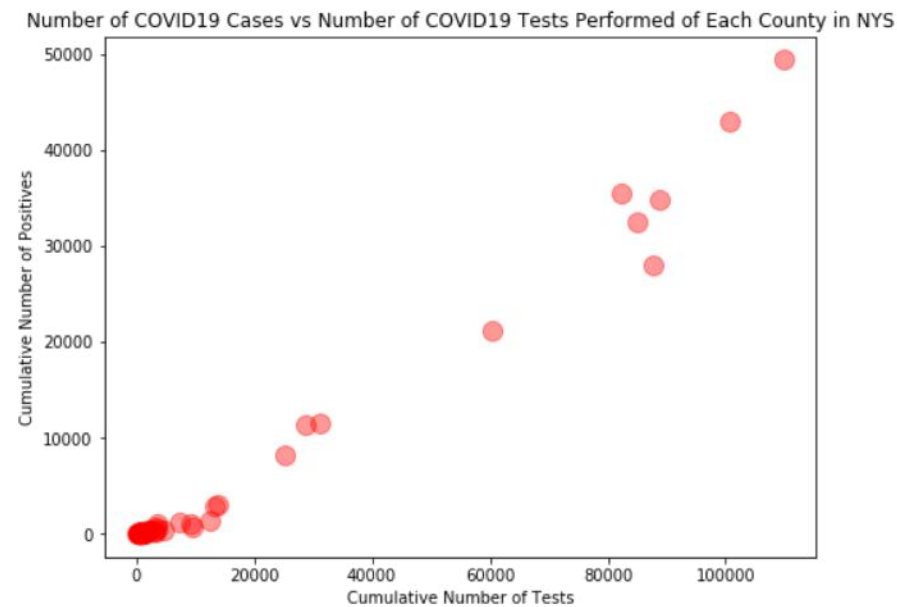- Data shows relatively strong linearity.

DATA EXPLORATION

Figure 3. Number of COVID-19 Tests vs
Population of Each County in NYS

- Data shows strong positive linear relationship.
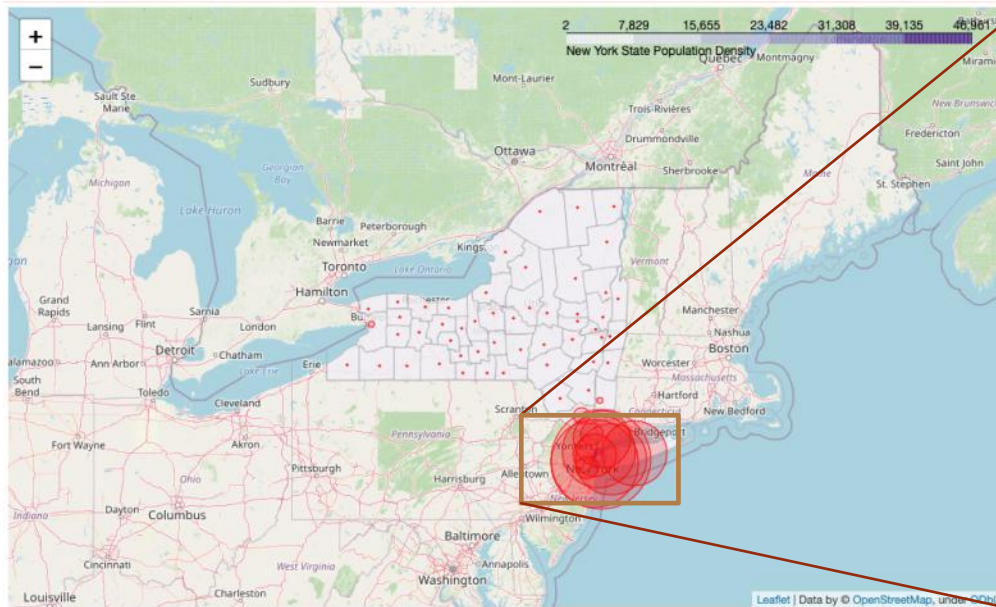
COVID-19 CASE MAPPING

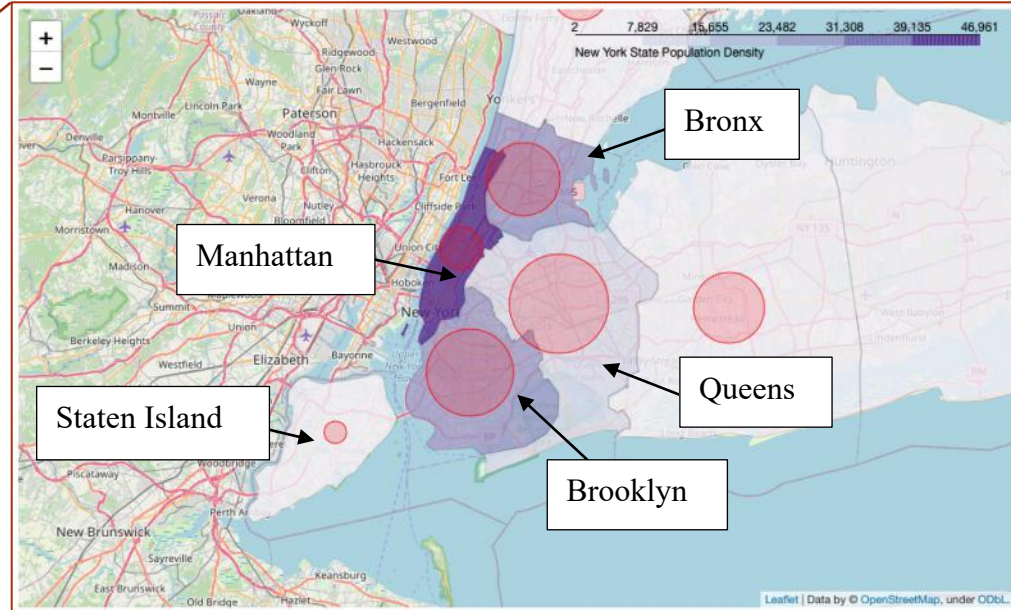Figure 5. COVID-19 Case Mapping (New York State)

Figure 6. COVID-19 Case Mapping (New York City)

- The majority of the COVID-19 cases in New York State is in or close to New York City.

- Queens, Kings (Brooklyn) and Bronx have highest numbers of COVID-19 cases in New York.

- New York county (Manhattan), which has the highest population density in New York, surprisingly has a significantly lower number of COVID-19 cases comparing to the proximal counties.
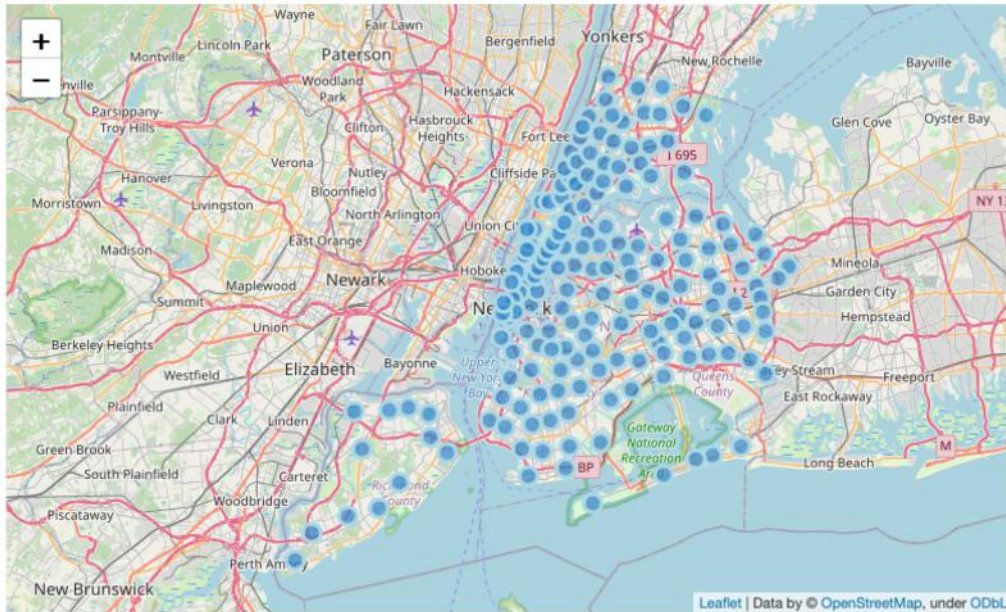
NEW YORK CITY NEIGHBORHOOD CLUSTERING

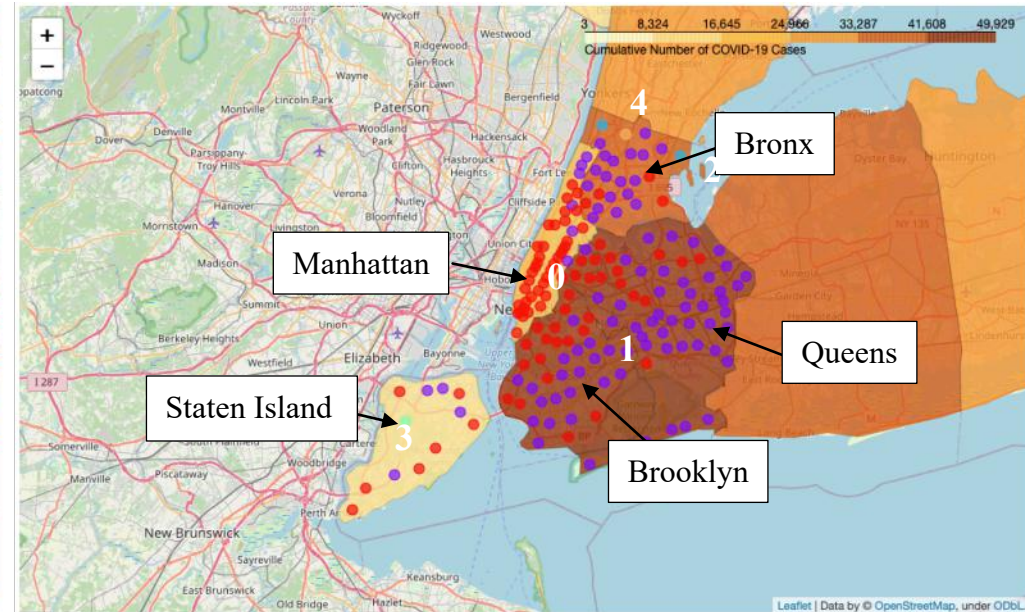Figure 7. Locations of Zip Codes in New York City

Figure 5. Venue Clusters and COVID-19 Case Distribution

- Cluster 0: most frequent venues are **coffee shops, bakeries and bars**. Dominating in areas with less COVID-19 cases

- Cluster 1: most frequent venues are **restaurants, supermarkets and bus stations**. Dominating in areas with more COVID-19 cases

DISCUSSION

- Number of COVID-19 cases and number of COVID-19 tests have a strong positive linear relationship, suggesting the importance of getting enough COVID-19 tests.

- The relationship between number of COVID-19 cases and population density is displays a weak linearity. In counties with high population density, the number of COVID-19 cases is even inversely proportional to population density. This is possibly due to the strict stay at home order.

- Areas with restaurants, supermarkets and bus stations dominating have more COVID-19 cases (Bronx, Brooklyn and Queens).

  - Large flows of people

  - People staying in confined spaces for a relatively long time

- Areas with coffee shops, bakeries and bars dominating have significantly smaller number of COVID-19 cases (Manhattan and Staten Island).

  - Short stay and less frequent visits

CONCLUSION

To avoid the spread of coronavirus:

- Increase number of testing.

- Avoid dining in restaurants and shopping in supermarkets in person.

- Stay at home as much as you can.