COVID

A Study of Covid in New York City

Intro
Descriptive Statistics
Inferential Statistics

Since the start of the year, Corvid-19 has spread across the country, resulting in many cases of infections, hospitalizations, and deaths. New York (especially New York City) emerged as an early hot spot, and there are trends among those who are infected.

A Capstone Project for Springboard by Emily Rice

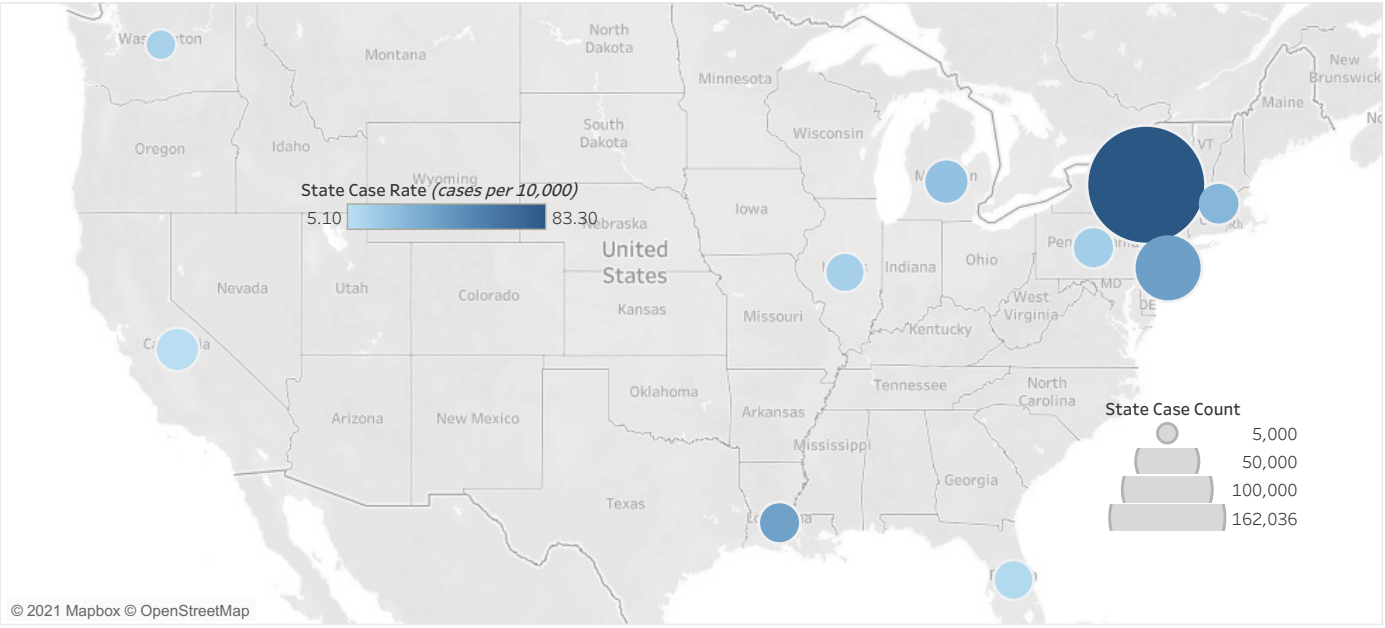
Sources:

github.com/nychealth/coronavirus-data
github.com/BuzzFeedNews/2020-05-covid-city-zip-codes
github.com/nytimes/covid-19-data
2015-2019 American Community Survey 5-Year Estimates

COVID

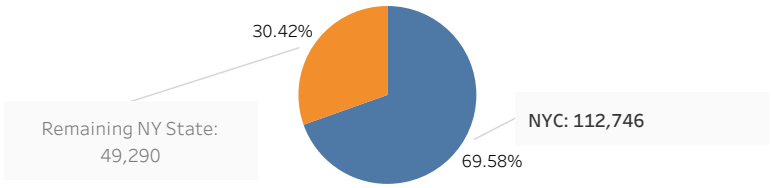
COVID-19 in the United States on April 9, 2020

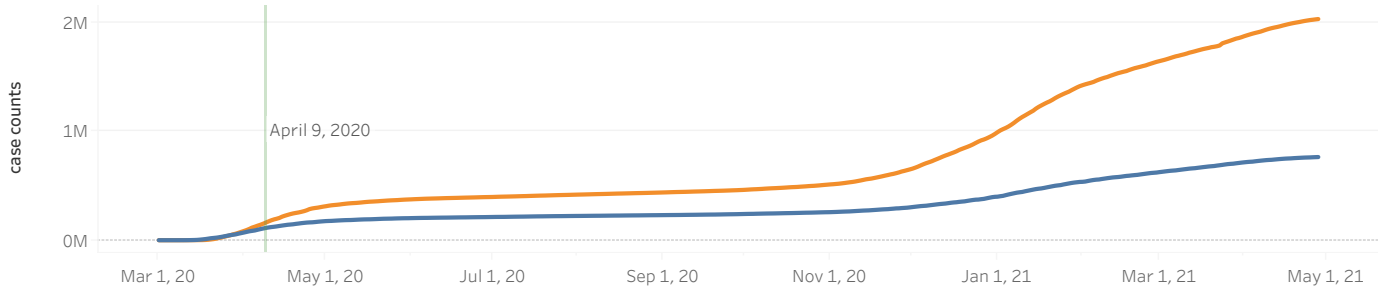
At the start of the pandemic, most of the cases were focused in a handful of hotspots, New York being the main one



COVID

At the start of the pandemic in New York, most of the cases were located in NYC

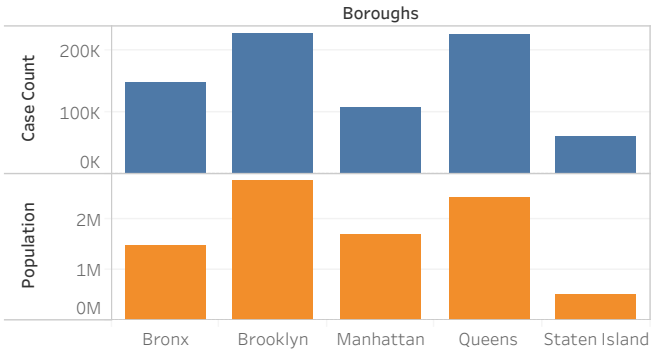




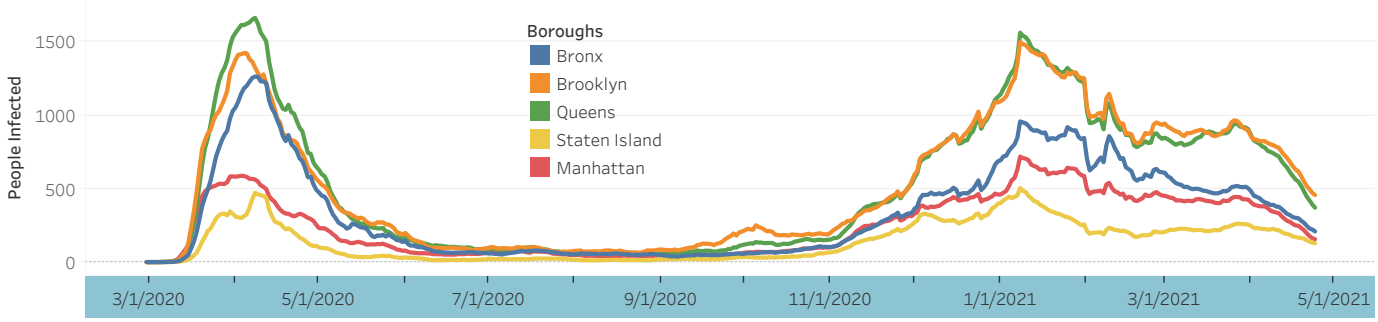
COVID

Throughout 2020 and 2021, NYC experienced two waves of COVID-19. In NYC, case count was high and roughly corresponded to the populations of the five boroughs.

case count and population



cases per day *(based on 7 day avg)*

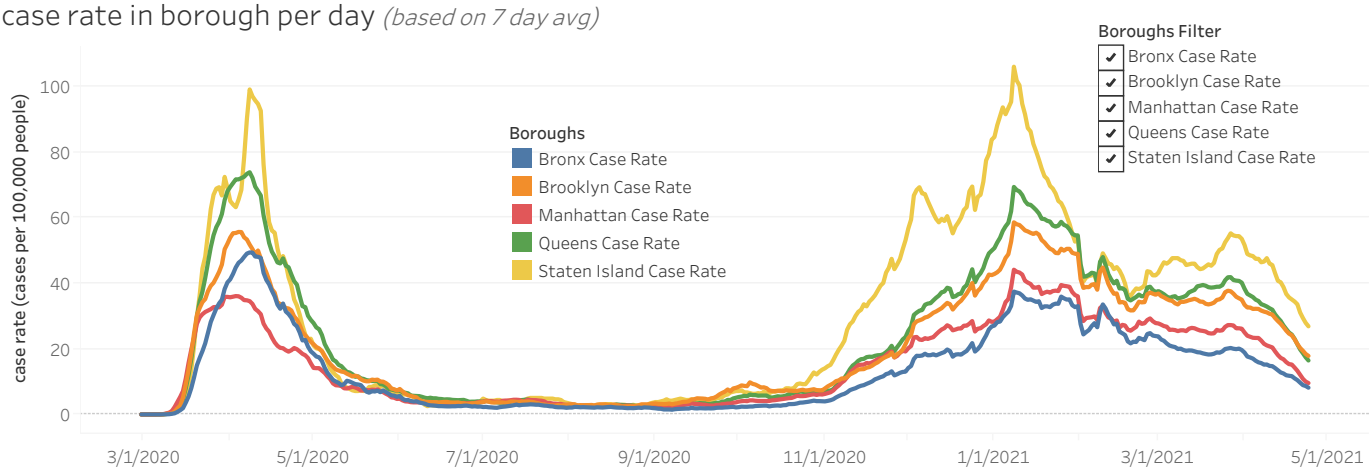


COVID

Case rate was high by US standards. It is interesting to note, however, that although Staten Island had the fewest case counts per day, it consistently had the highest case rate per day, which could be because of the many service jobs or because Staten Island is the only large Repubican area in NYC, as many Republican areas ignored COVID restrictions more than other areas. Staten Island's high case rate will appear in the following scatter plots.

(www.nytimes.com/2020/11/11/nyregion/staten-island-second-wave.html)

case rate in borough per day *(based on 7 day avg)*



COVID

4 Factors of Coronavirus Spread

The 4 demographic variables hypothesized to be related to case rate are:

1. Mean Income
2. Median Age
3. Percent Service Jobs
4. Percent Non-White

These are taken from the Social Vulnerability Index, an index of factors related to a population's ability to handle public health emergencies.

Observed correlations of these 4 demographic variables with case rate will be used to predict the coefficients of a corresponding multi-linear regression.

COVID

Mean income:

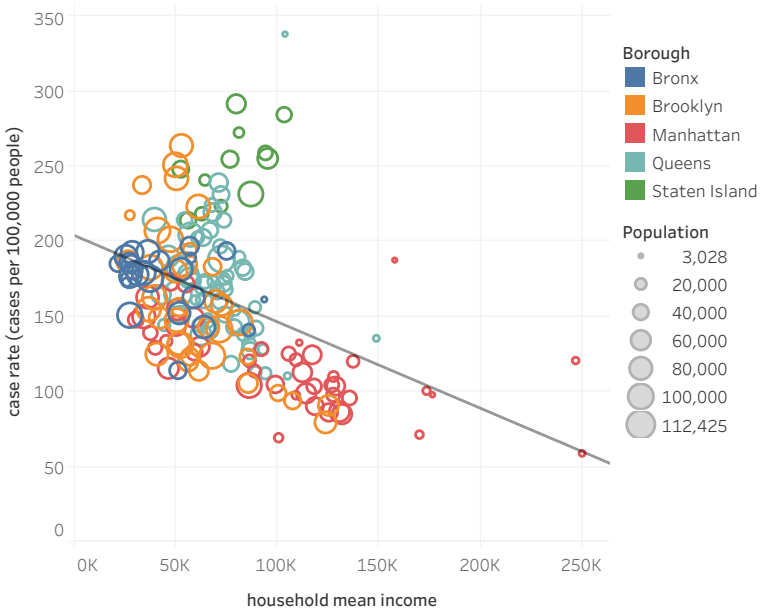
negative coefficient

Poorer areas have less access to high quality health care and usually have a greater population of people with serious health conditions. (www.frontiersin.org/articles/10.3389/fsoc.2020.00047/full)

Household income is negatively correlated

household mean income (by zip code)

(using avg NYC case rates from 8/2020 to 5/2021)

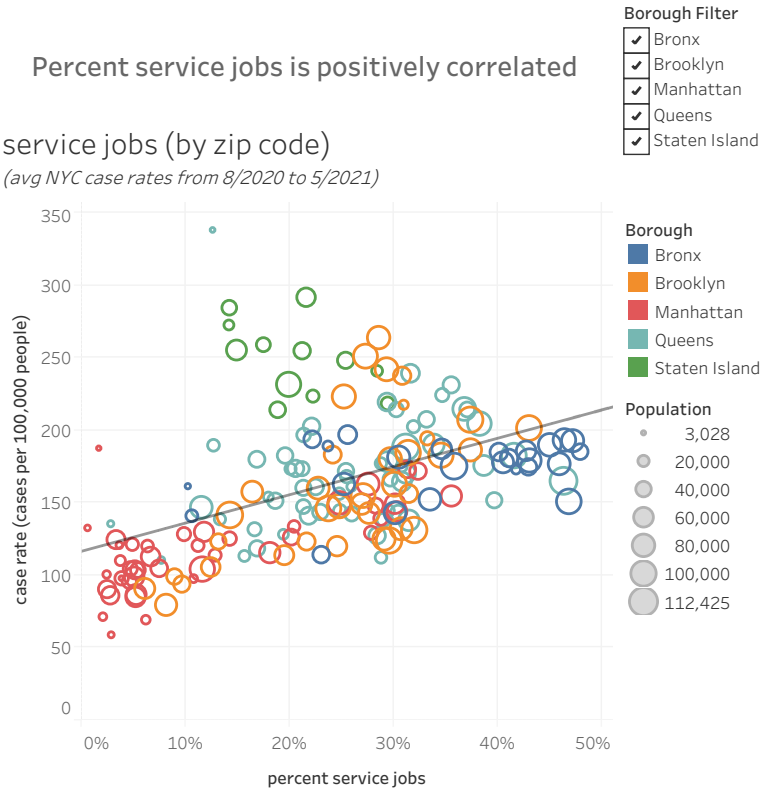


COVID

% Service jobs:

positive coefficient

People with service jobs do not have the luxury of working from home. Service jobs often offer no paid sick leave, making working while ill likely. Many service jobs also make social distancing difficult and expose service workers to the public. (www.nbc-news.com/health/health-news/these-are-most-dangerous-jobs-you-can-have-age-coronavirus-n1201496)



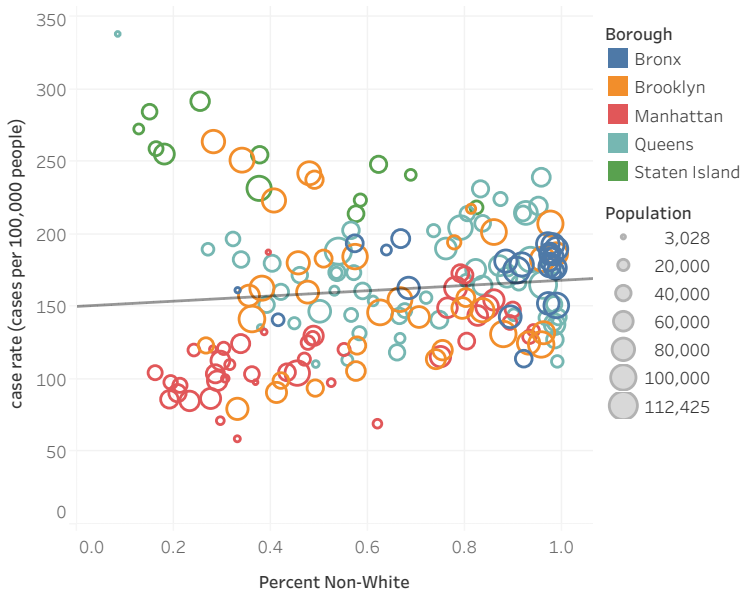
COVID

% Non-white:

positive coefficient

Besides issues correlated with income and percent service jobs (discussed later), sometimes there is racial bias in medical treatment or non-white people are less likely to trust medical professionals. Additionally, redlining of blacks has led to less home ownership and living in subpar neighborhoods, sometimes polluted, and they also tend to be overrepresented in densely populated areas. (www.brookings.edu/blog/fixgov/2020/04/09/why-are-blacks-dying-at-higher-rates-from-covid-19)

Percent non-white is slightly positively correlated



COVID

% Median age:

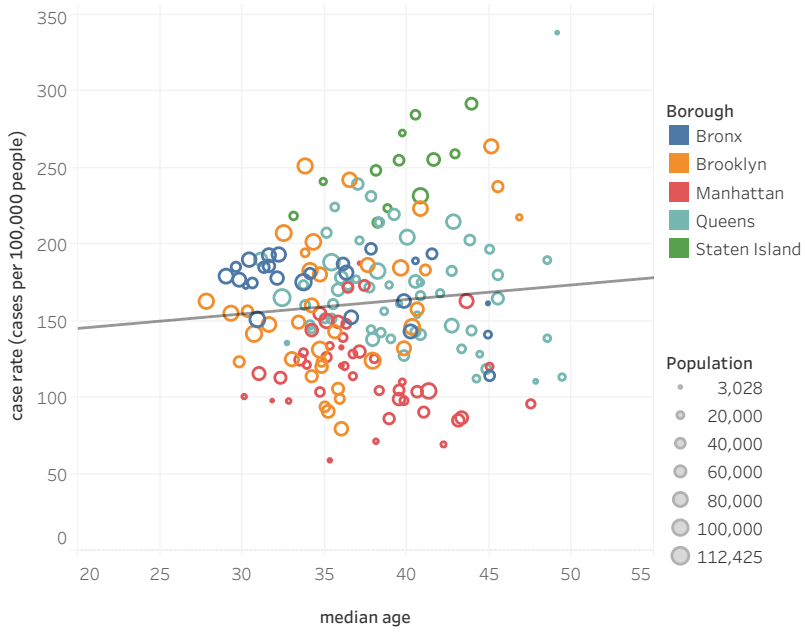
positive coefficient

Older populations are more susceptible to coronavirus in both infection and severity.

Median age is positively correlated

median age (by zip code)

(using avg NYC case rates from 8/2020 to 5/2021)



COVID

REGRESSION MODEL RESULTS

all variables significant ($p < 0.05$)

+ coefficients: median age, % service jobs

- coefficients: mean household income, % non-white (explained later)

regression results

(using standardized variables)

Variables	Coefficients
intercept	163.64
mean household income	-20.89
median age	12.87
percent non-white	-42.76
percent service jobs	42.80

context

Variables	Mean	Standard Dev
household mean income	60,440.00	27,3
median age	36.30	
percent non-white	0.69	
percent service jobs		

R squared = 0.69

dependent variable:

cases per 100,000 people (by zip code)

Each coefficient represents the change in the case rate per change in the number of standard deviations of the independent variable (value minus mean divided by std dev). An R squared value of 0.69 means the regression explains 69% of variation of case rate by zip code.

* only used 73 records (out of 177) where zip code population was above 50,000 (reducing the effect of a zip code with a pop. of 3000 having the same impact as on..

COVID

Why is % non-white's coefficient negative, or associated with a decrease in case rate?

1. May be the correct result, although it is commonly believed that being non-white is positively correlated with contracting COVID.

2. Perhaps there is less testing where there is a higher percentage of non-white people. If fewer non-whites are tested, fewer cases will be reported. The calculated correlation between percent of color and test rate is -0.53 (as calculated by the author).

3. Percent non-white groups all minority groups into one, which may be inappropriate because of socioeconomic and cultural differences.

4. While all the variables are significant, there is great overlap between them. For example, being non-white is correlated with lower incomes and service jobs.