Choosing the safest place to live

Analysis of COIVD-19 infection rates in the New York neighborhood

Gaitis Kasims

2nd of June 2, 2020

Contents

Introduction

Background

COVID pandemic started in China beginning of 2020 and since then quickly spread to all over the world. Even if a number of daily new cases have stabilized and is not growing exponentially any more, situation in the world is still critical. Also in US number of cases reported daily is not getting down and situation remains critical. The most affected in the US is New York where the pandemic has hit the most. Choosing a best place to live in New York City was linked more to the neighborhood attractions and places, which now might be linked to the pandemic outbreak.

Problem

Different neighborhoods of the New York City are affected at a different scale. In this report we will research the link between the neighborhood data and the pandemic situation. Specifically, we will look at the linking between availability of restaurants, attractions and medical centers to the spread of COVID-19.

Interest

Not only citizens would be interested to understand how their neighborhood is linked to pandemic situation, but it would be important also to predict hot spots in other cities as well.

Data acquisition and cleaning

Data sources

Data on COVID 19 infections per New York City ZIP code in CSV format can be found here. This includes data about COVID cases, death numbers and a population by a ZIP code. Additionally, data from Foursquare API is used to enhance the available data set with the venu information. Finally, zip code geojson information was retrieved from here.

Data cleaning

Data from the Foursquare was combined with the COVID case CSV data. There were however several problems with that, which required preprocessing before data could be used.

Firstly, neighborhood naming of COVID cases was not matching with the Foursquare information an geojson information. It was identified that to get the geocoding information, only first part of neighborhood naming from CSV file could be used.

Secondly, having data per ZIP code showed to have not enough data for most of ZIP codes, for that it was decided to group the data by neighborhood, rather than looking at each ZIP code level.

Finally ZIP code information was acquired in int format and need to be converted into string format.

Feature selection

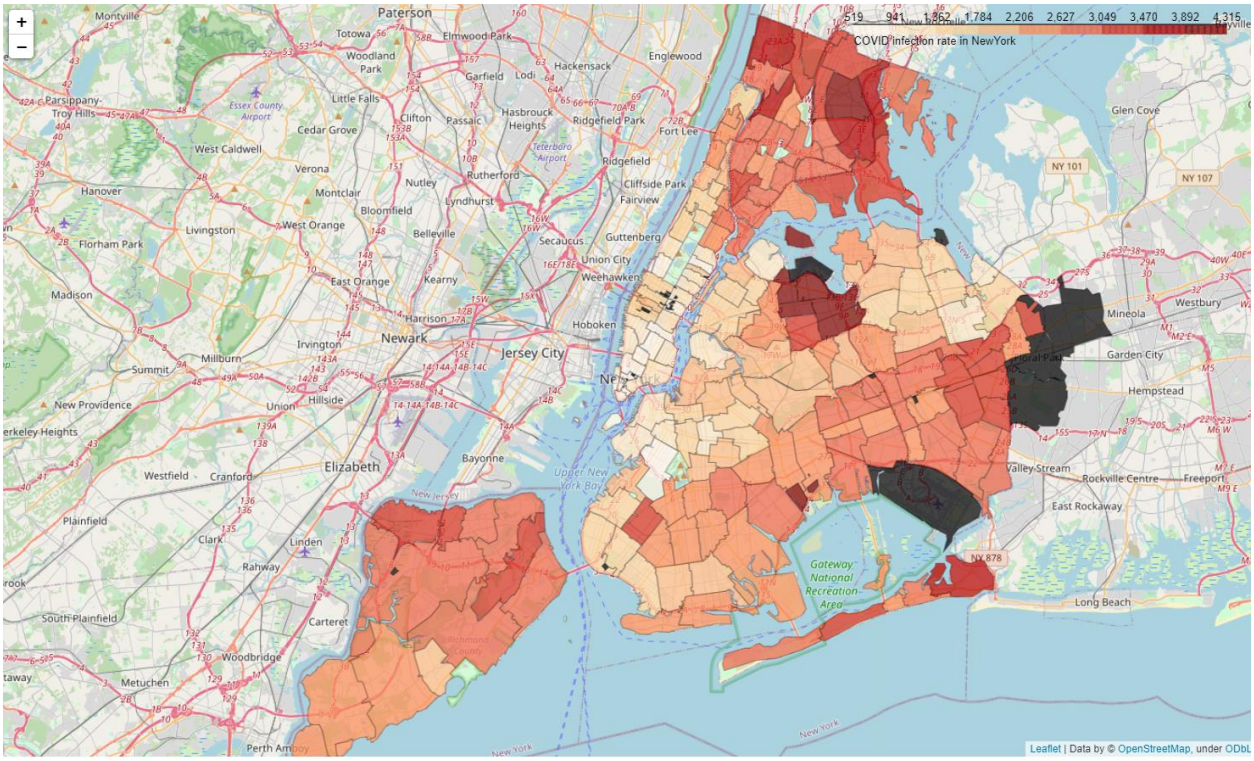
It was decided to group the Foursquare data on master categories Food, Arts & Entertainment and Medical Center. This allows looking at higher level data before deciding to drill down potentially.

Also infection rate for COVID-19 was taken, rather than a death rate, as it is not clear if the death is getting categorized as COVID-19 related using same criteria in different medical centers.

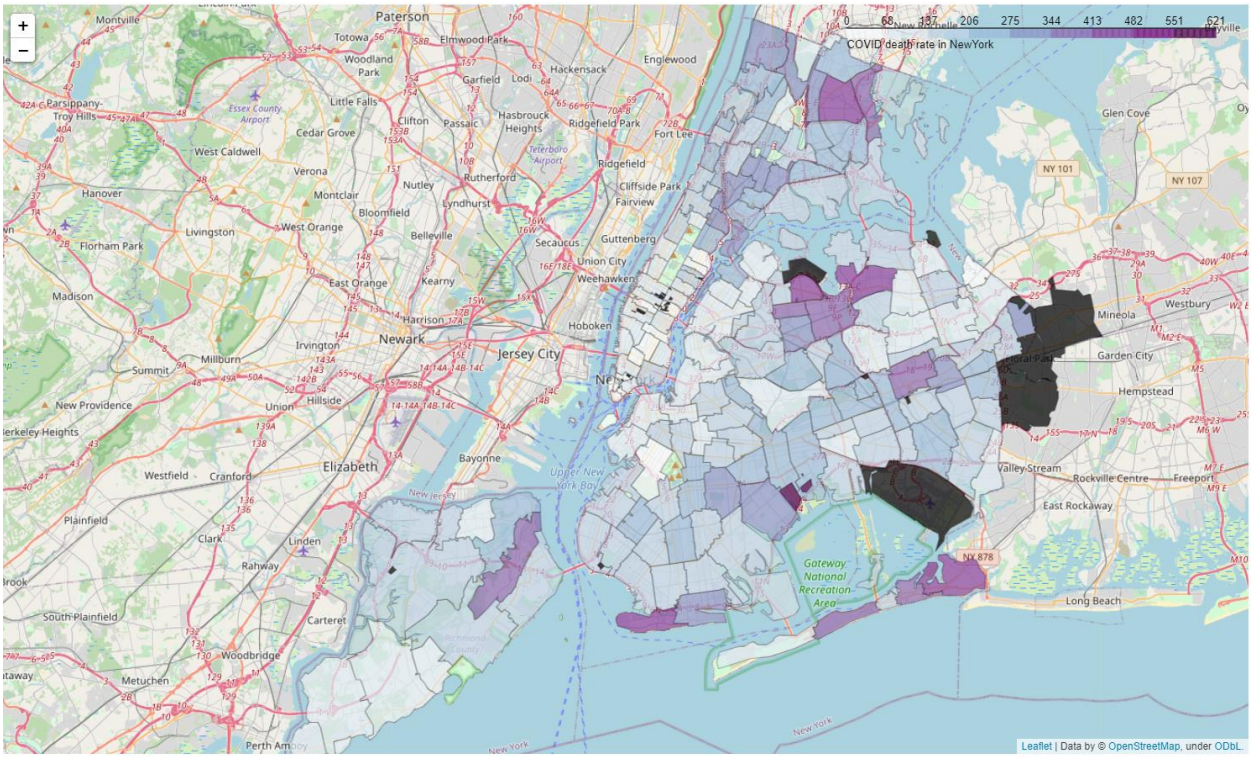
Exploratory Data Analysis

To continue with the analysis, the first question to answer is if there are differences in the infection rate in the different neighborhoods. For that a folium map with chropleth is used for visualization.

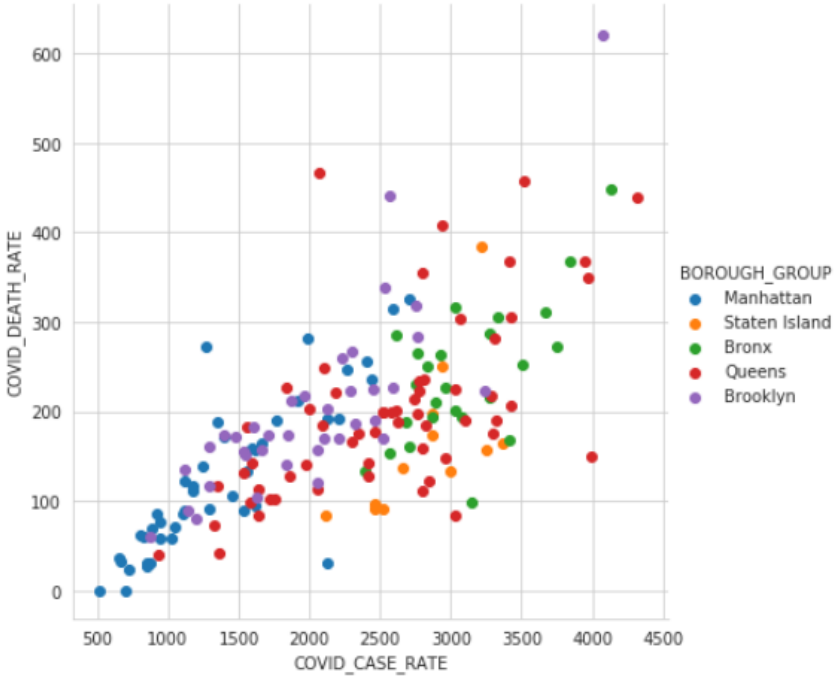
First we look at the COVID infection rates visualization. As we can see there are big differences between the different neighborhoods for infection rates, ranging from 519 to 4315.



Also the death rates vary significantly, from 0 to 621.

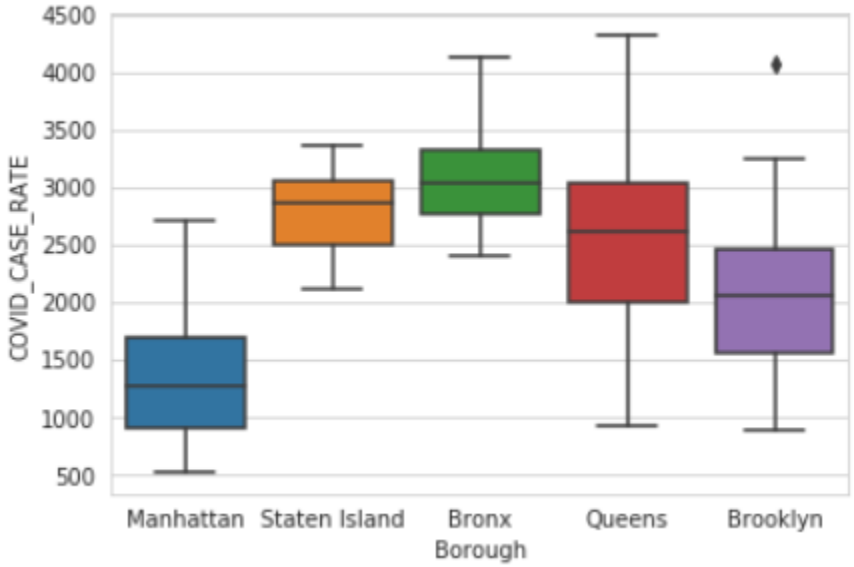


Now we can visualize cross data between infection rate and death rate in different boroughs. As there are too many neighborhoods to visualize in this format, we need to stop at boroughs.



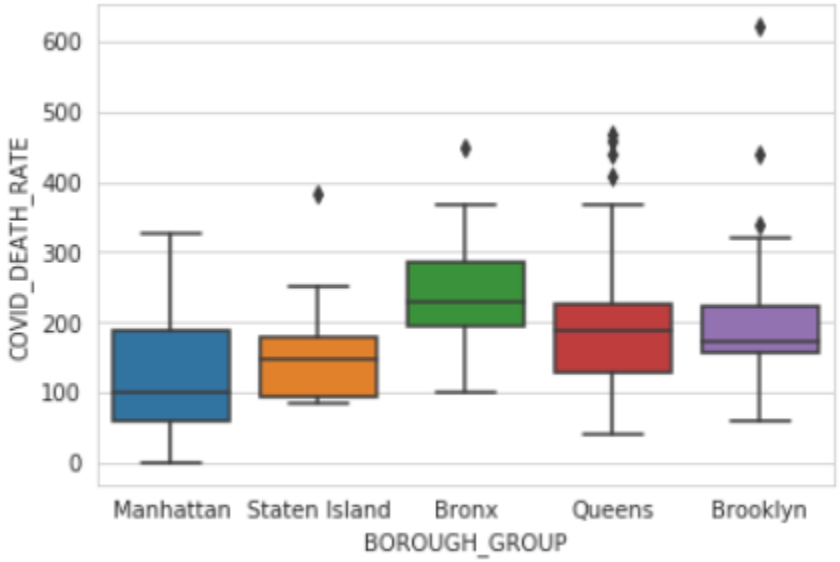
Boroughs having the data points at upper left corner are the worst ones, having relatively much higher death rate outcome.

Looking at more details for COVID cases can use box plots.



We see that, the highest rates are in Bronx and Staten Island, where the box is concentrated at the higher values.

Also the death rates are similar with Bronx as a leader and Mahattan as a best.

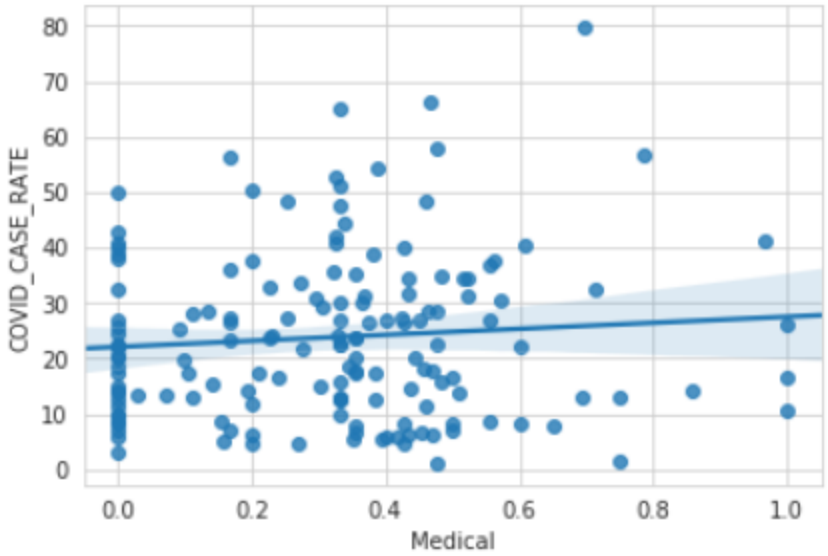


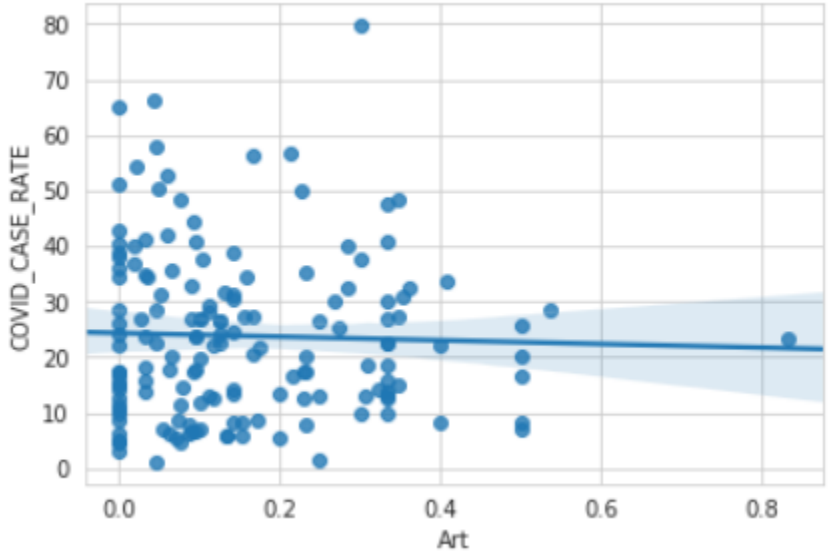
Looking at this data we see that safest is to live in the Manhattan. Let's analyze why this is the case

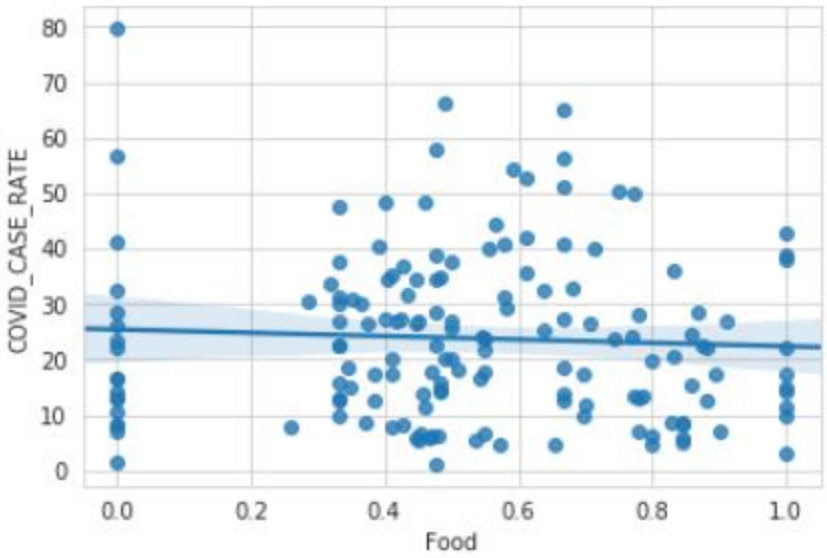
Regression

We will look at the different neighborhood, analyzing how much venues are available there from type Food, Arts & Entertainment or Medical Center.

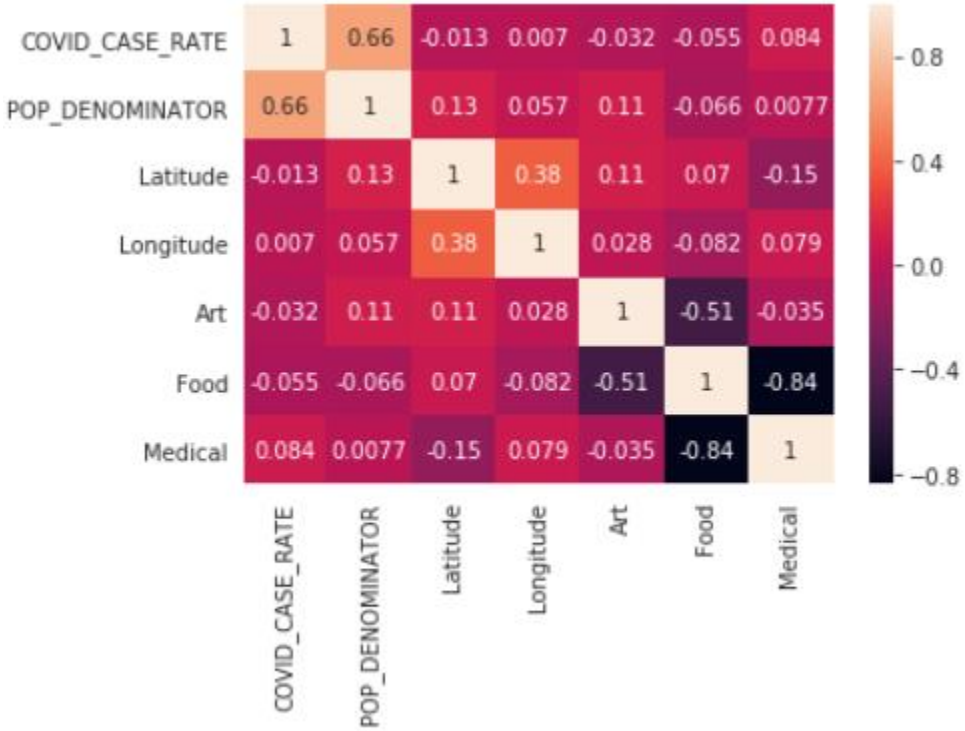
First we can check each of the features separately.


COVID_CASE_RATE
Medical
0
10
20
30
40
50
60
70
80
0.0
0.2
0.4
0.6
0.8
1.0


COVID_CASE_RATE
Art
0
10
20
30
40
50
60
70
80
0.0
0.2
0.4
0.6
0.8



None of these is showing a correlation, so most probably also multiple regression will not give good results as indicated also by correlation matrix.



Only features are correlating with each other.

Linear regression as expected shows very low R2 value

```
LinearRegression(copy_X=True, fit_intercept=True, n_jobs=None,
         normalize=False)
Mean Absolute Error: 11.785929590247573
Mean Squared Error: 218.7551994831311
Root Mean Squared Error: 14.790375231316178
R2 score: 0.007980795135463903
```

Also Multiple factor regression comes only slightly better with R2 of 0.03502016273397435

Conclusions

From the analysis it is clear that the best place to live in New York is in Manhattan, both from infection as well as from the death rates. Against expectation there was no link identified between places where people are gathering or availability of medical centers and the infection rates.

Future work

Next step here would be to analyze at the finer grained level the different venue category availability and compare the data from New York with similar neighborhoods in the other cities.