Covid-19 vs. New York City

Yong Chen

Background

COVID-19 is caused by a coronavirus called SARS-CoV-2. Older adults and people who have severe underlying medical conditions like heart or lung disease or diabetes seem to be at higher risk for developing more severe complications from COVID-19 illness.
In 2020, COVID-19 outbroke severely in the whole world. New York City is one of the most affected areas. Up to date, there are 6,226,409 confirmed cases worldwide; 1,835,300 of them are in the United States, while New York City itself has 203,303 confirmed cases.

Problem

In this project, we will explore insights and exciting information about the correlation between location data and COVID-19 data by zip code in New York City to see how they affect each other.

Data

Data source

data-zsnEP.csv

This dataset includes the Cases, Cases per 100,000, Deaths per 100,000, and Percent of people tested who tested positive of Covid-19 by zip code within New York City.

Source:[NYC Health] https://www1.nyc.gov/site/doh/covid/covid-19-data.page

Foursquare API

Based on the zip code, coordinates of the areas and venues within 500 meters are collected through Foursquare API, including venue name, venue Latitude, venue Longitude, and venue category.

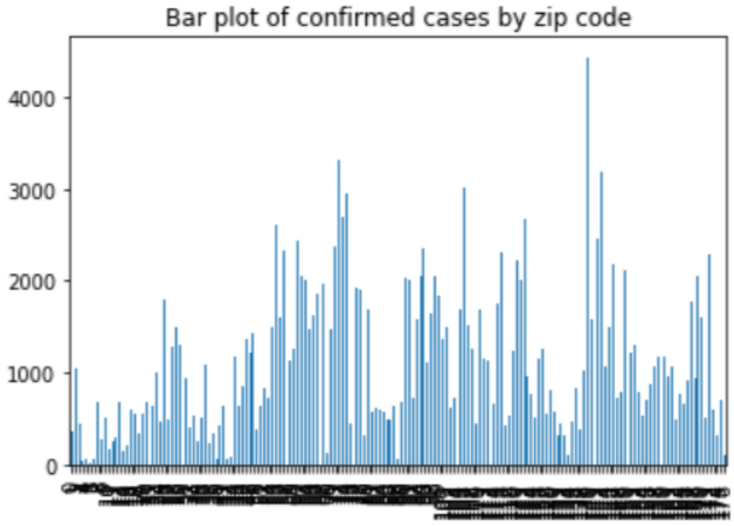
Data cleaning

During the process, the latitude and longitude of zip code “11234” from Geopy were misleading, which caused an empty result from Foursquare. After updating the coordinates into the one on Google, the venue information was successfully updated.

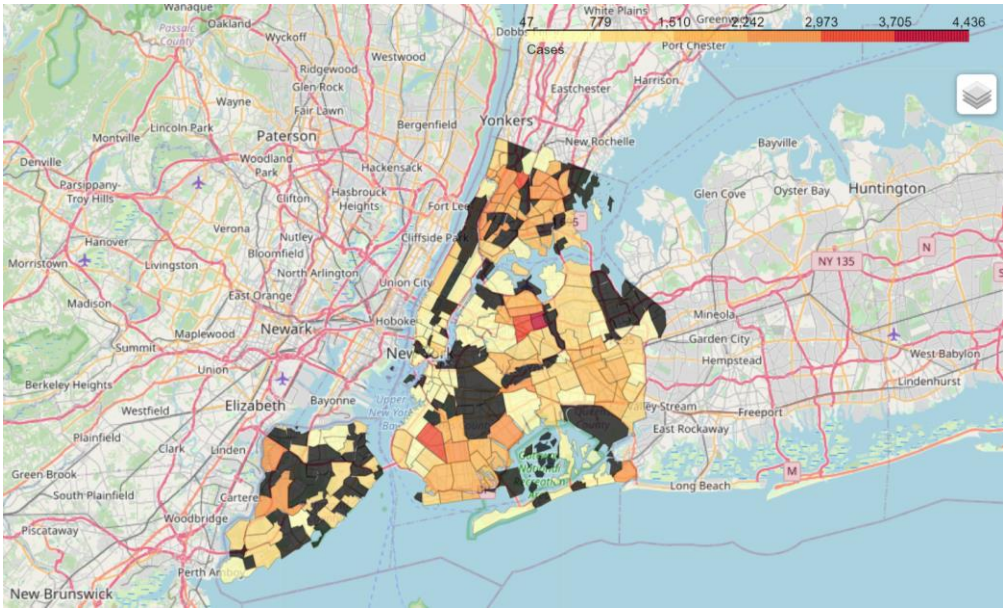
After getting the nearby venue's information with Foursquare, the venue category is our focus. Since it is a categorical variable with hundreds of categories, one-hot-encoding was implemented to flatten the data and make it one zip code per row.

The final data is 177*349, including Zipcode, 344 one hot coding categories, and four columns of COVID case data.

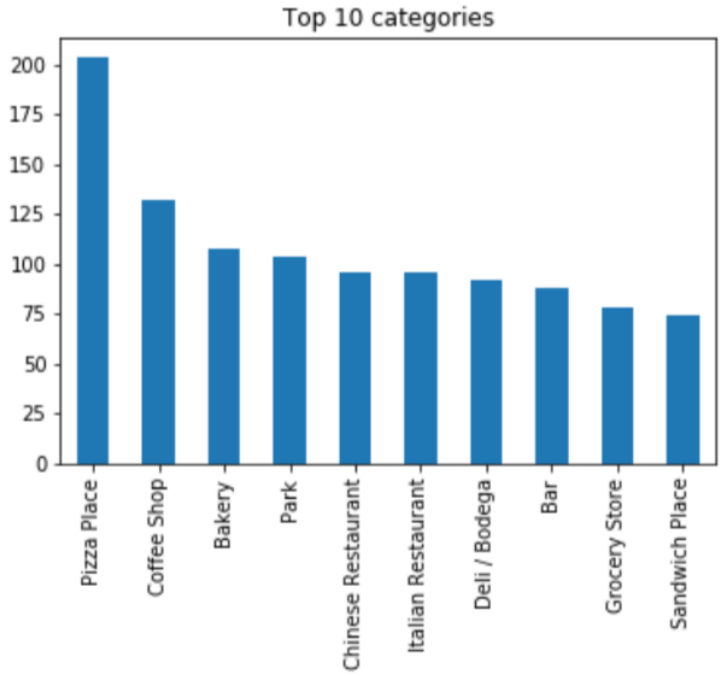
Exploratory data analysis



The above plot is a bar plot that shows how many confirmed cases there are by postal code in New York City. You can notice that the numbers of the confirmed case are sparse. Some neighborhoods have a lot more cases than the other ones, indicating that some unknown factors might be affecting the numbers.



Then, breaking down the zip code, a Choropleth map was created to show the range of cases in each neighborhood; however, some neighbor data were missing. The whole colored area on the map is NYC, while the deeper the color is, the more cases there are. The black areas are where data is missing. We can see there are several spots in the middle which have the most cases. Half of the domain is between 779 and 2242.



Out of all the venues extracted to be the data, here are the top 10 categories nearby NYC zip code centers. The eight out of ten are all food-related; the other two are “park” and “bar”.

Methodology

First of all, Linear Regression is used to prove if there is any linear relationship between the categories and the number of cases. 344 one hot coding categories as predictors, X, while "Cases" as the target variable. However, since the predictors are either 0 or 1, the numbers of "Cases" are much higher. Therefore we would take a log transformation on "Cases" to balance the data.

We are adding Random Forest as a more complicated algorithm to see if any difference.

Since the feature dimension is a lot higher than the number of observations, PCA is applied to reduce dimensions. Then Kmeans is used to cluster the areas.

Results

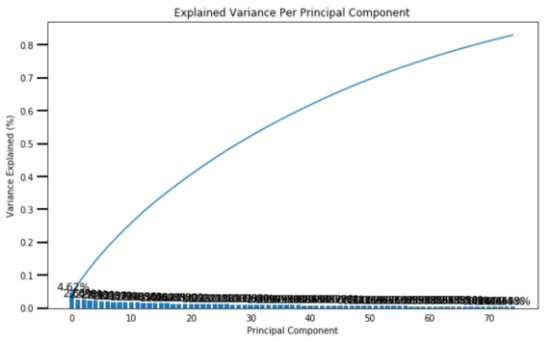
Coeffiencts	Category
1.725480	Art Museum
1.696269	Department Store
1.655098	Cheese Shop
1.641648	Fried Chicken Joint
1.485366	Dumpling Restaurant

Coeffiencts	Category
-1.132608	Mediterranean Restaurant
-1.196123	Community Center
-1.261439	Mattress Store
-1.425798	Border Crossing
-2.392710	Dog Run

The data frame above is the top 10 most essential features generating from linear regression, also known as a category. "Art museum", "Department Store", "Cheese Shop", "Fried Chicken Joint" and "Dumpling Restaurant" have the highest positive effect on the number of cases. On the contrary, areas with “Mediterranean Restaurant”, “Community Center”, “Mattress Store”, “Border Cross” and “Dog Run” tend to have lower cases.

Importance	Category
0.186724	Monument / Landmark
0.063804	Dog Run
0.050711	Park
0.035435	Cycle Studio
0.026387	Border Crossing
0.021089	Italian Restaurant
0.020799	Coffee Shop
0.016144	Pizza Place
0.015416	Pier
0.015192	American Restaurant

The above table is the feature importance of Random Forest. It shows the categories that are the most important to the COVID-19 cases. However, we can notice that these categories have many overlapping with the top 10 categories. Therefore, it can only indicate that the more a category in the area, the more important it is for the algorithm, which makes sense.

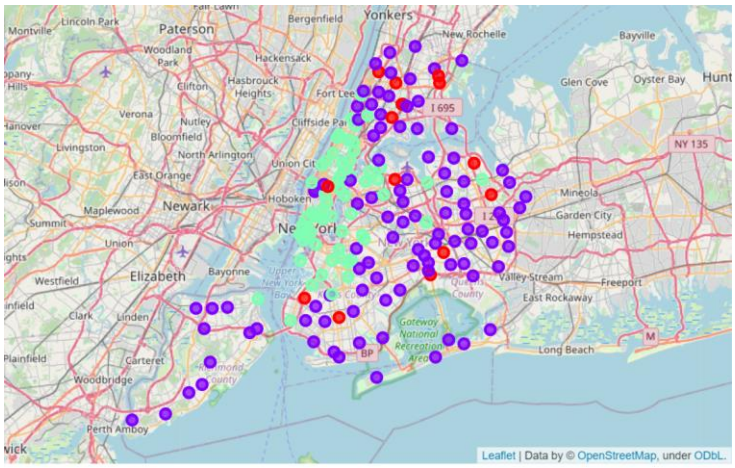


Considering how to choose the components of the Principal Components Analysis, 75 components were used since it is the threshold for 80% explained variance.

After PCA reducing the feature dimension from over 300 to 75, K Means clustering was applied to form three clusters.

label	Count	Mean
0	14	1351.785714
1	101	1268.247525
2	62	762.048387

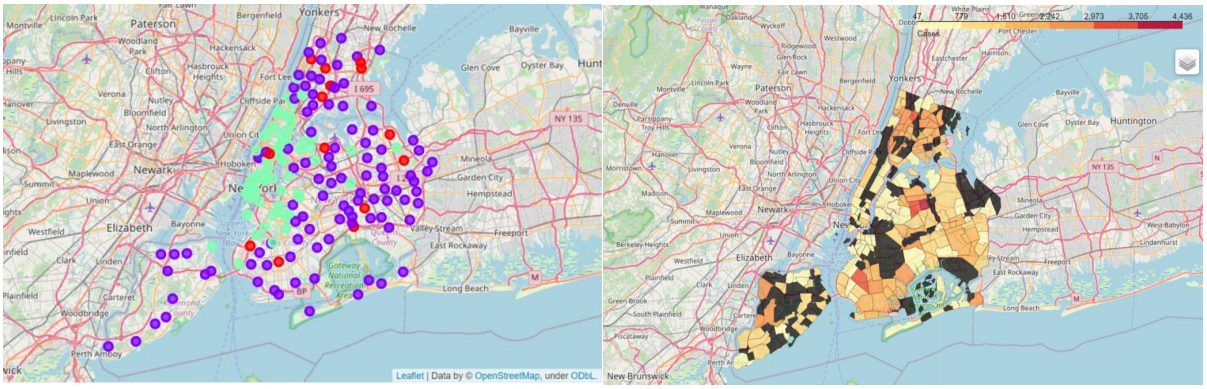
Mean case numbers of each class are shown in the table above. The first class has the highest mean but lowest count, the second class has the highest count, while the third class has the lowest mean.



The map shows how the clusters locate. The purple points are class 1, the greens are class 2 and the red ones are from class 0.

Discussion

The left map shows three different clusters of the zip code areas, while the right map shows the distribution of number of cases of Covid-19. Comparing these two maps, you can easily notice that they are connected. Even though the clustering factors are difficult to elaborate and explain, you can tell the the red points are the areas that have the most cases, the green ones are the areas that have the least cases while the purple ones are moderate affected areas.



Conclusion

Areas with "Art museum", "Department Store", "Cheese Shop", "Fried Chicken Joint" and "Dumpling Restaurant" tend to have higher number of cases. On the contrary, areas with “Mediterranean Restaurant”, “Community Center”, “Mattress Store”, “Border Cross” and “Dog Run” tend to have less cases. The Kmeans model did a good job clustering the areas with three classes corresponding to different levels of COVID-19 cases.