

COVID-19 Spread Assessment in New York Counties

Background

Corona virus disease (COVID-19) is an infectious disease caused by a new virus. The disease causes respiratory illness (like the flu) with symptoms such as a cough, fever, and in more severe cases, leads to pneumonia and breathing difficulties. This disease spreads through contact with an infected person when they cough or sneeze. It also spreads when a person touches a surface or object that has the virus on it, then touches their eyes, nose, or mouth.

New York is the most populous city in the United States of America (USA). It is also deemed as the cultural, financial, and media capital of the world. The state of New York consists of 62 counties, with each county representing an administrative and political subdivision of the state.

With the spread of corona virus around the world, a massive surge in the number of cases is also being reported in the state of New York. There are different sorts of places and venues situated and operational in the NY state with various eateries, cafes, shopping malls, night life, and transportation stations.

Motivation

Since the outburst of COVID-19 pandemic, many researches have been performed which shows how this virus spreads, how long this virus survives on different surfaces, how to keep safe from the virus, etc. The best possible way to keep safe from the virus is to implement social distancing and maintain hygiene and cleanliness. It is imperative that the virus catches from person to person, so it is bound to spread among the people socializing and continuing with their daily dine out, transportation, and night life activities.

Current information available on this topic also shows that older population (people aged over 60) are more likely to die from this disease, while younger population has a better chance of recovery. However, no study has come forward as per my knowledge which analyses the lifestyle with regards to choice of restaurants, travel, and nightlife preferences of people impacted by the disease.

The motivation of this project is to determine the extent to which each sort of venue has contributed to the newly reported cases and deaths in the New York.

Project Scope

1. This project aims to find how night life and different travel and transportation venues across counties is linked to the spread of COVID-19 in the state of New York.
2. In addition, it also aims to analyse the extent to which different kinds of food joints are linked to the number of deaths in the New York.

Significance of this Project

A large number of COVID-19 data are being collected around the world in these times. So, it is important to make sense of this data at deeper levels so to enable people to make better lifestyle choices. This kind of study aims at promoting healthy choices when hopefully the world comes out of this pandemic. For example, in future people may prefer to avoid certain kinds of restaurants which do not serve immunity-boosting foods.

Data

Two datasets are used for this project:

1. COVID-19 data
2. 2020 population data of New York state counties

COVID-19 Data:

The COVID-19 data obtained for this project is obtained from the New York Times Github source:

<https://raw.githubusercontent.com/nytimes/covid-19-data/master/us-counties.csv>

This data has the following attributes:

Table 1: Variables in the COVID-19 dataset

| Variable | Description | Type |
|---------------|---|---------|
| date | Refers to the date of reported cases and deaths in the respective United States county. | Object |
| county | the name of County reporting the cases and deaths. | Object |
| state | the state in which the county is present. | Object |
| fips | The Federal Information Processing Standard Publication 6-4 (FIPS 6-4) is the five-digit Federal Information Processing Standards code which uniquely identifies counties in the United States. | Float64 |
| cases | the reported number of COVID-19 cases in the respective county on a certain date. | Object |
| deaths | the reported number of deaths due to COVID-19 in the respective county on a certain date. | Object |

This data will be used along with the coordinates of each county to explore the most popular venues in each of these counties. An analysis will be performed on the resulting venues and county data to determine how certain travel and nightlife venues impact the number of cases. Also, an analysis will be performed to determine how the number of deaths in a county is linked to the kind of most popular eateries and restaurants.

2020 New York Population Data

Another dataset that is used in this project is 2020 population data of New York. This dataset is obtained from <https://worldpopulationreview.com/us-counties/ny/>.

This dataset has the following attributes:

Table 2: Variables in the New York Population dataset

| Variable | Description | Type |
|--------------------------|-------------------------------------|---------|
| Name | New York county name | Object |
| 2020 Population | Population of county in 2020 | Int64 |
| Growth since 2010 | The growth in population since 2010 | Float64 |

Extract, Transform, and Load (ETL)

The COVID-19 dataset is read as a pandas dataframe. The cumulative count of patients since the first patient was diagnosed for the latest available date is used in this project. The dataframe is further filtered to contain only counties of New York. After this transformation, the data contained the count of cases and deaths in 58 counties of New York on a particular date.

When sorted, it was found that the following counties had the most number of cases as reported on 9th April 2020.

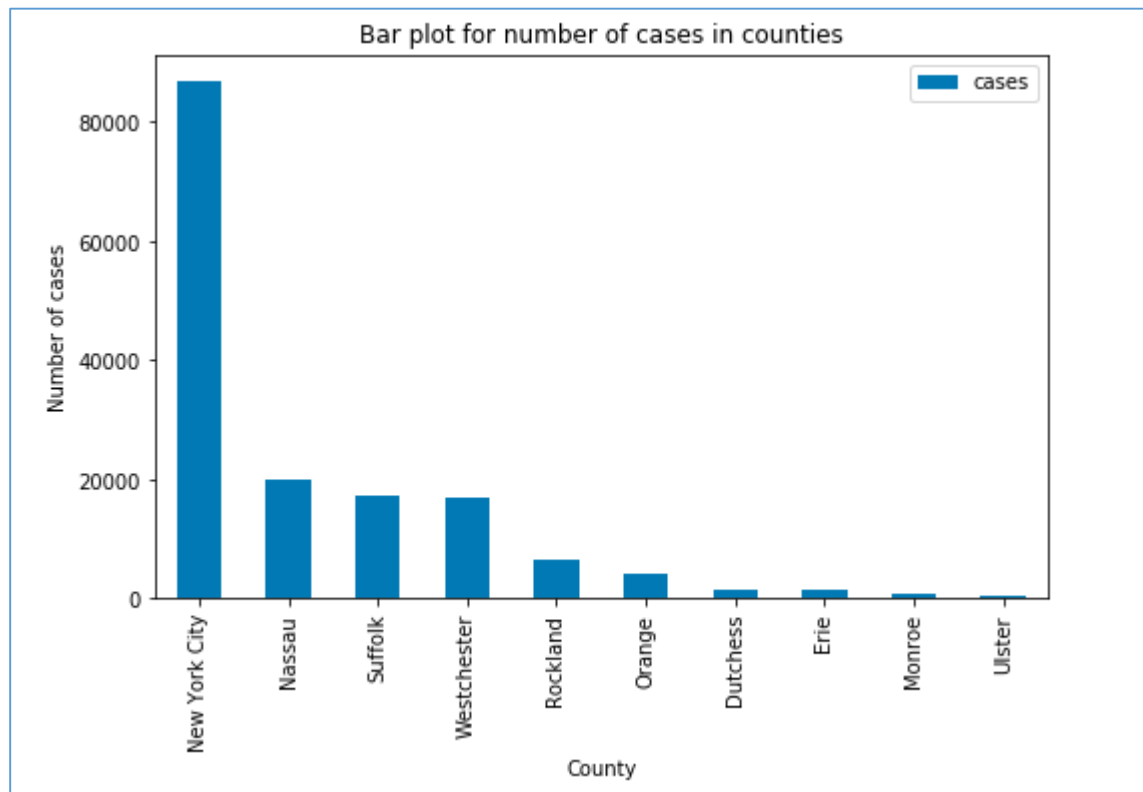


Fig. 1: Counties with the most cases on 9th Apr 2020

It can be seen from the above figure, that the highest number of cases were reported in the New York city county in the state of New York.

To find out what percentage of the population was impacted by the virus, the COVID-19 dataset was merged with the population dataset. For this join, the count name column in population dataframe was stripped of the word "County" using the *rstrip* function to match it with the county names in the COVID-19 dataframe.

In addition, the population dataframe county names had some spelling mistakes which would have resulted in missing values after joining. These values were modified manually using the replace function of pandas dataframe.

The percentage of cases and deaths in each county was calculated in the joined dataframe as follows:

$$\text{Percentage of Deaths} = \frac{\text{Number of deaths}}{\text{Number of cases}} * 100$$

$$\text{Percentage of Cases} = \frac{\text{Number of cases}}{\text{2020 Population of county}} * 100$$

Methodology

After transforming the above datasets into a single dataframe with required columns, the FourSquare endpoint **search** was used to formulate two URLs to access venues in the category of food, and travel and night life. The following common parameters were used for both URLs:

Table 3: Common parameters in both URLs

| | |
|----------------------------|----------------------------|
| categoryIdFood | '4d4b7105d754a06374d81259' |
| categoryIdTravel | '4d4b7105d754a06379d81259' |
| categoryIdNightLife | '4d4b7105d754a06376d81259' |
| LIMIT | 50 |
| Radius | 750 |
| Version | '20200412' |
| Near | <County name, NY> |
| Intent | 'browse' |

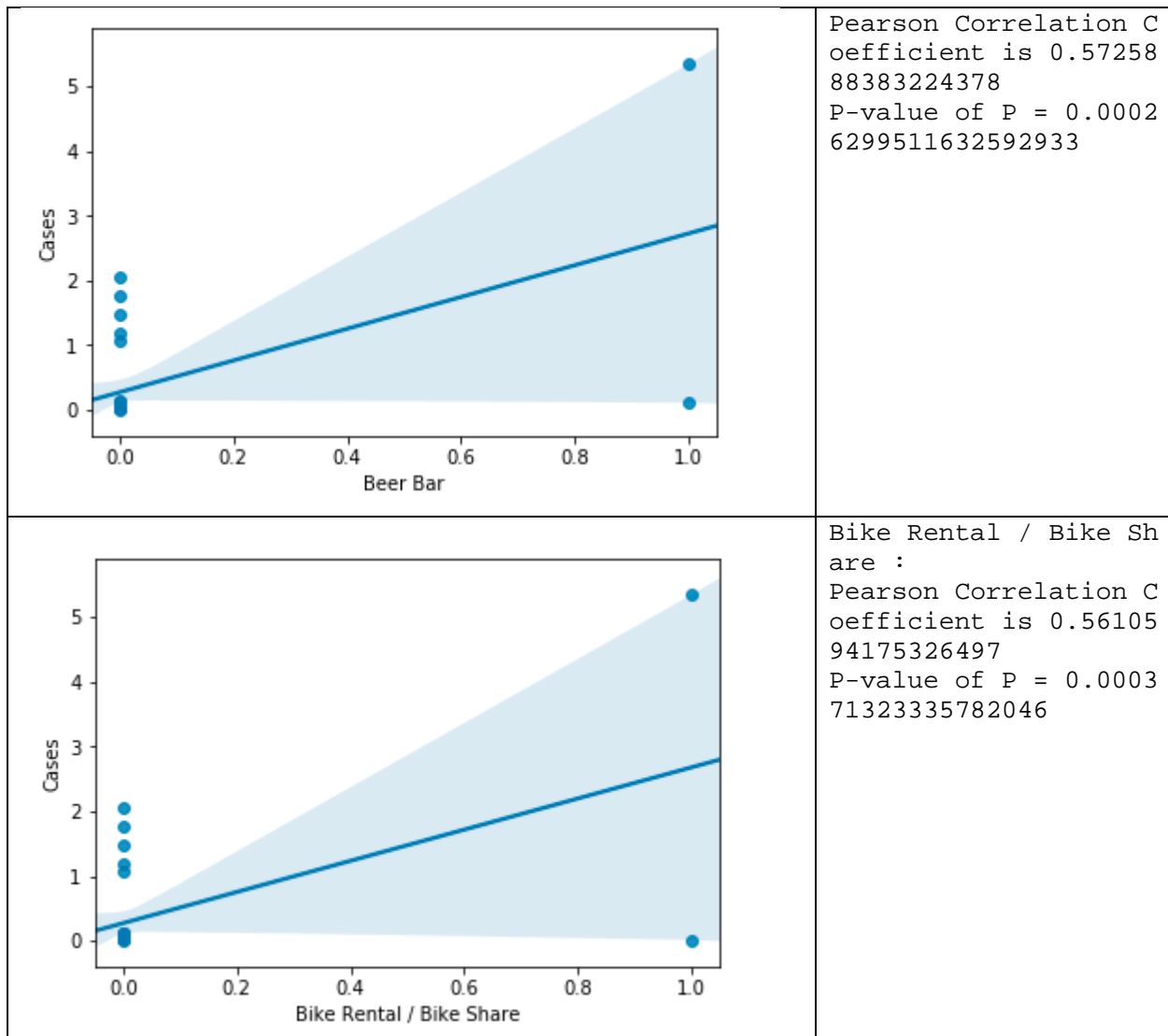
The category ID was changed in both URLs as follows:

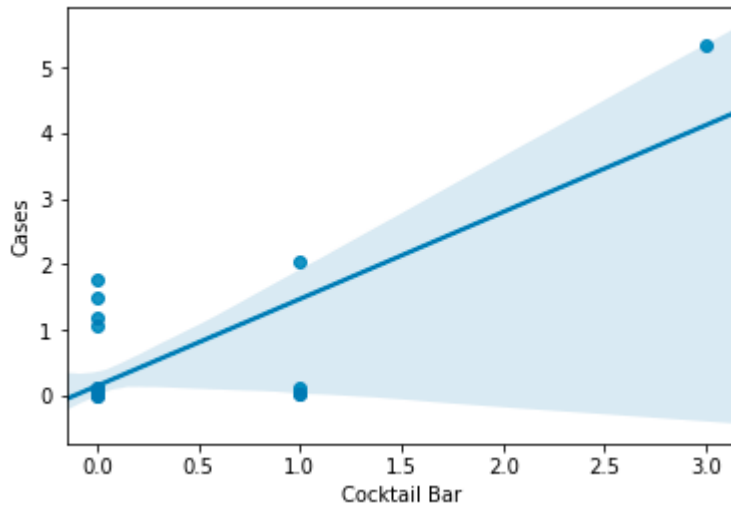
- URL1 to study the effect of food venues in a county on the number of deaths contained categoryIdFood.
- URL2 to study the effect of travel and night life venues in a county on the number of cases contained categoryIdTravel and categoryIdNightLife.

The resulting JSON format was studied and all venues were extracted in a dataframe which was then grouped for each county to contain the sum of venue categories in each county. The final transformed dataframe is shown below:

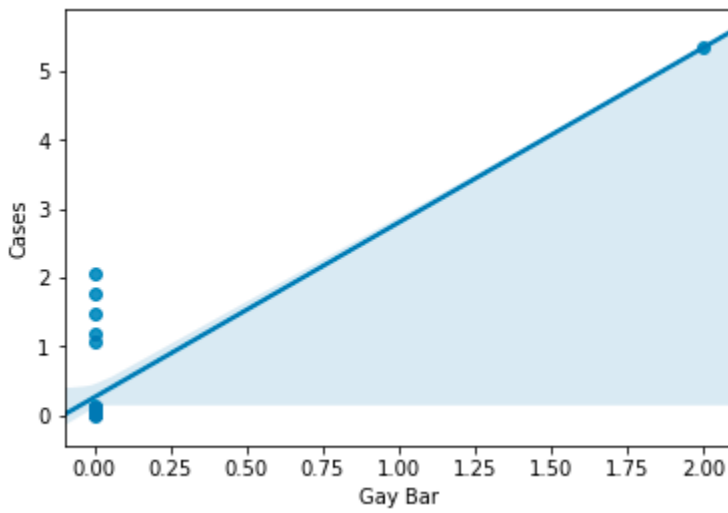
| | County | Cases | Deaths | American Restaurant | Argentinian Restaurant | Asian Restaurant | BBQ Joint | Bagel Shop | Bakery | Bar |
|---|-------------|----------|----------|---------------------|------------------------|------------------|-----------|------------|--------|-----|
| 0 | Albany | 0.123406 | 3.166227 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | Allegany | 0.047383 | 0.000000 | 2 | 0 | 0 | 0 | 0 | 1 | 1 |
| 2 | Cattaraugus | 0.022124 | 0.000000 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | Chautauqua | 0.014069 | 5.555556 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | Clinton | 0.049569 | 0.000000 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |

Results – Correlation of Number of Certain Night Life and Travel Spots with Number of COVID-19 cases

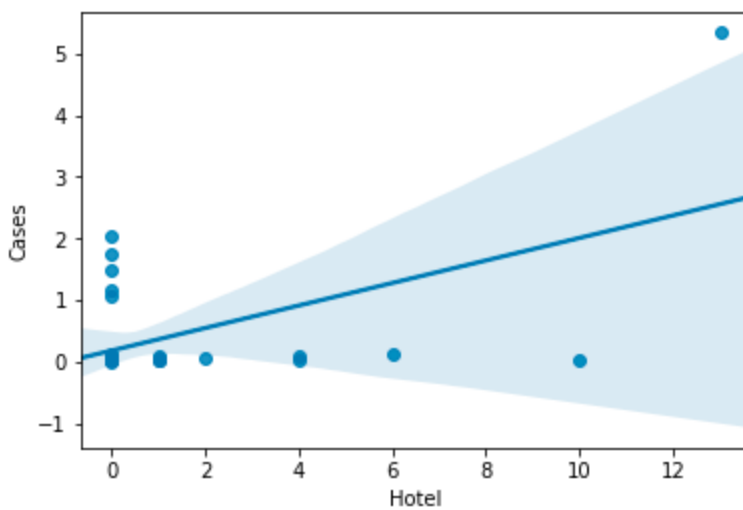




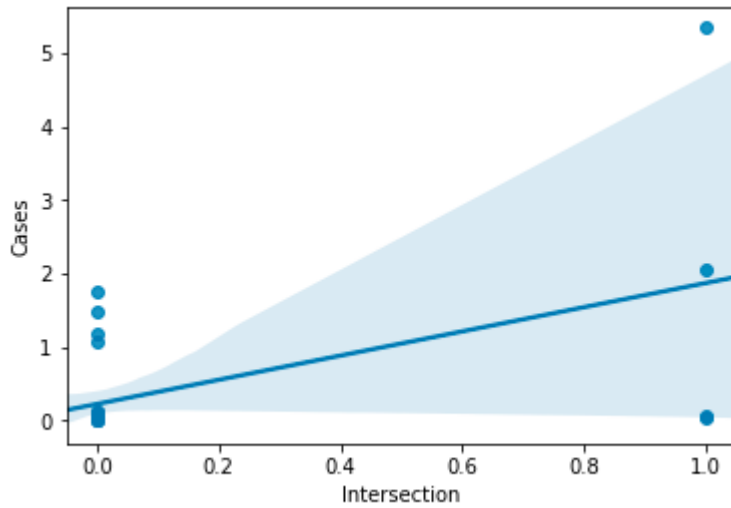
Pearson Correlation Coefficient is 0.7668694895736455
P-value of P = 4.918730847898263e-08



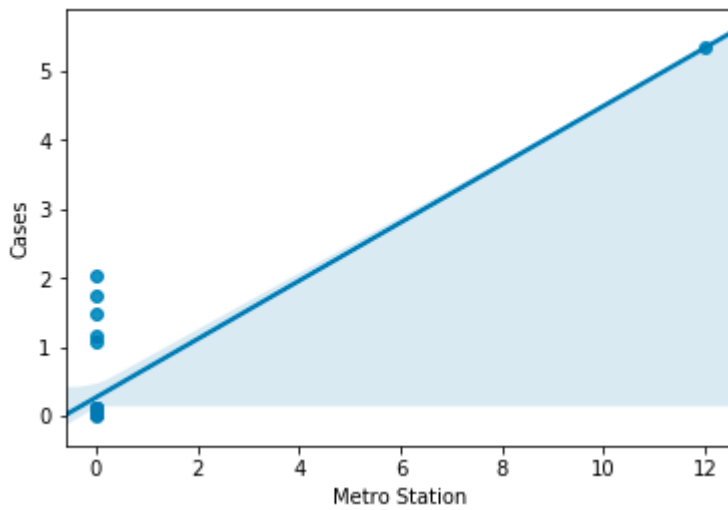
Pearson Correlation Coefficient is 0.8493565924946066
P-value of P = 5.805961946688885e-11



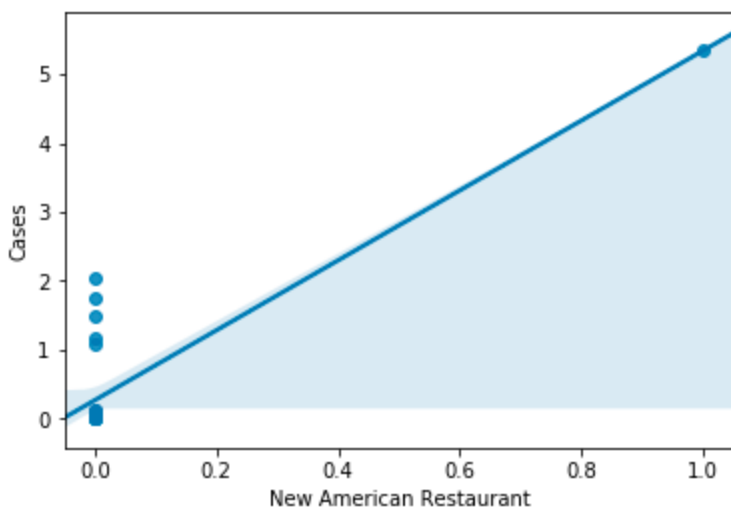
Pearson Correlation Coefficient is 0.5300533454996675
P-value of P = 0.000884399254056661



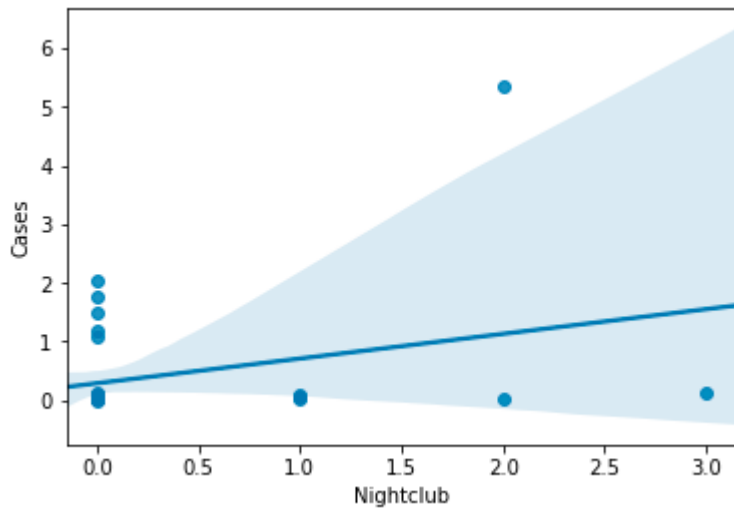
Pearson Correlation Coefficient is 0.5267787046525637
P-value of P = 0.0009646689746400993



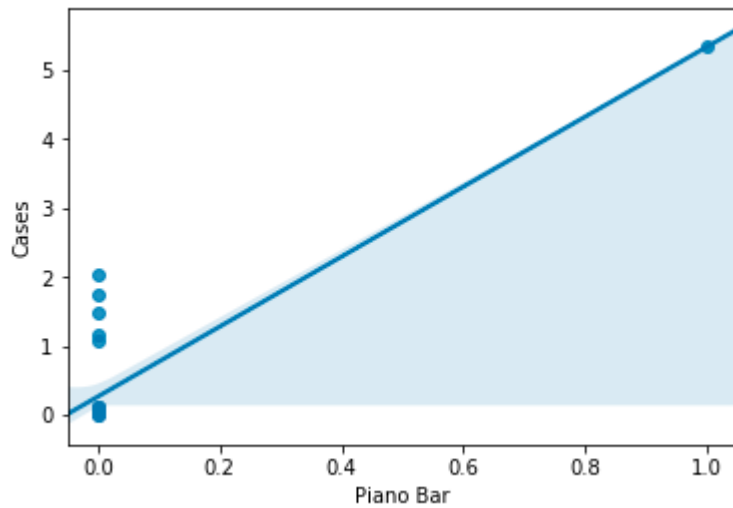
Pearson Correlation Coefficient is 0.8493565924946063
P-value of P = 5.8059619466891044e-11



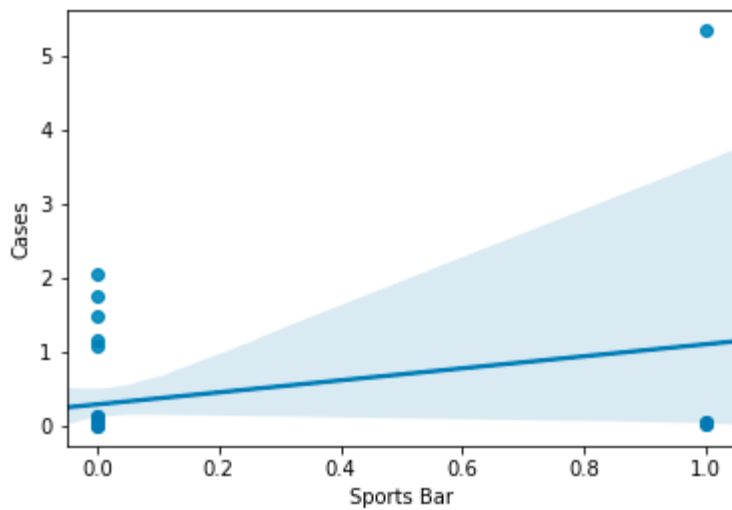
Pearson Correlation Coefficient is 0.8493565924946066
P-value of P = 5.805961946688885e-11



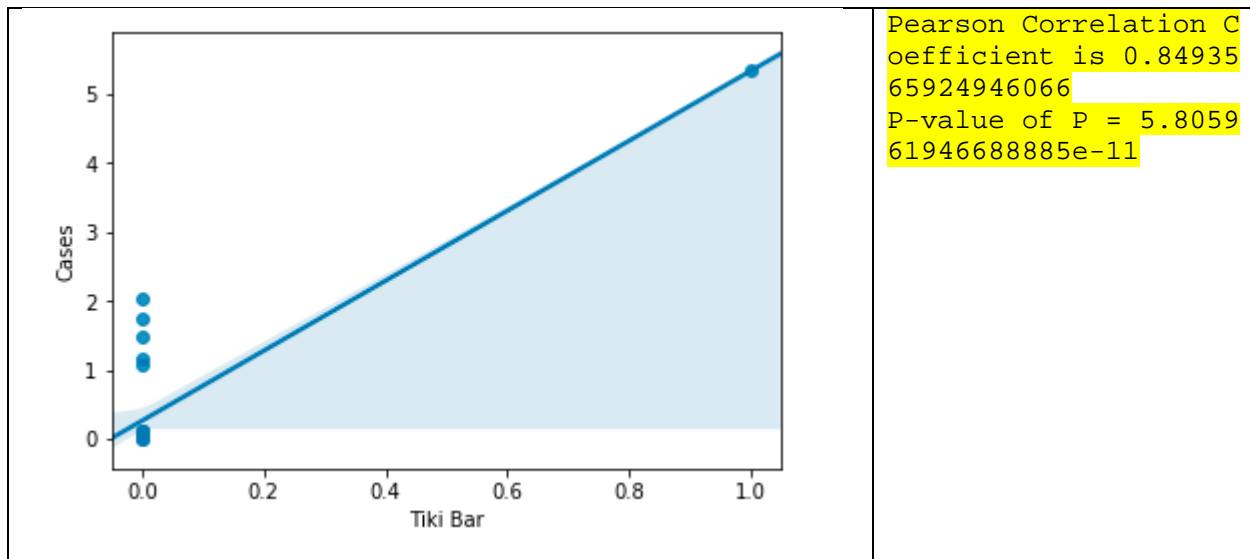
Pearson Correlation Coefficient is 0.2965150286139827
P-value of P = 0.07908600371571615



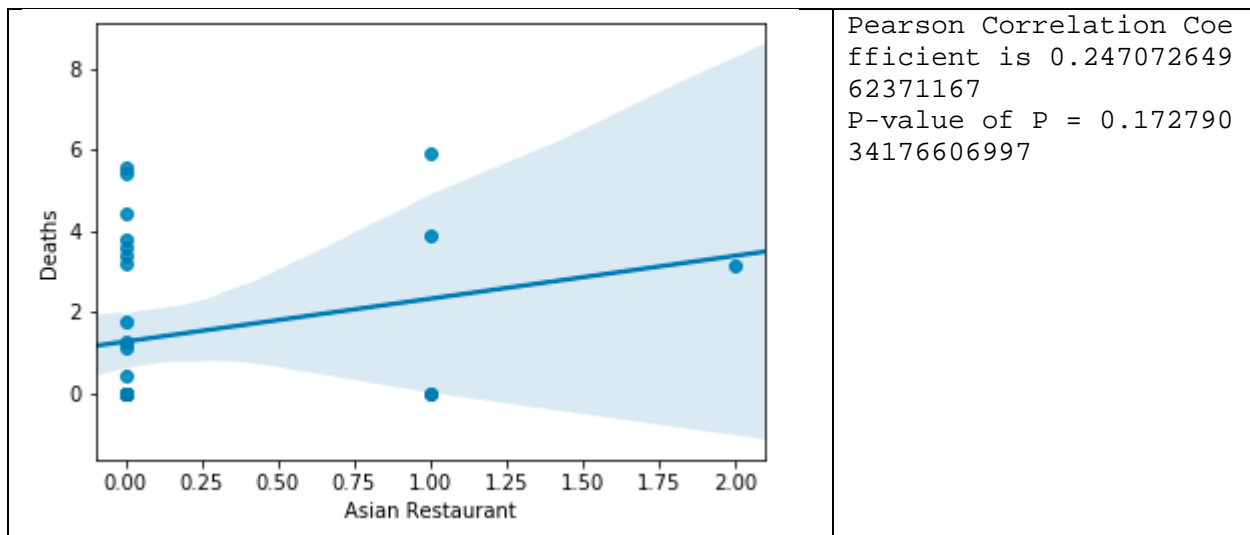
Pearson Correlation Coefficient is 0.8493565924946066
P-value of P = 5.805961946688885e-11

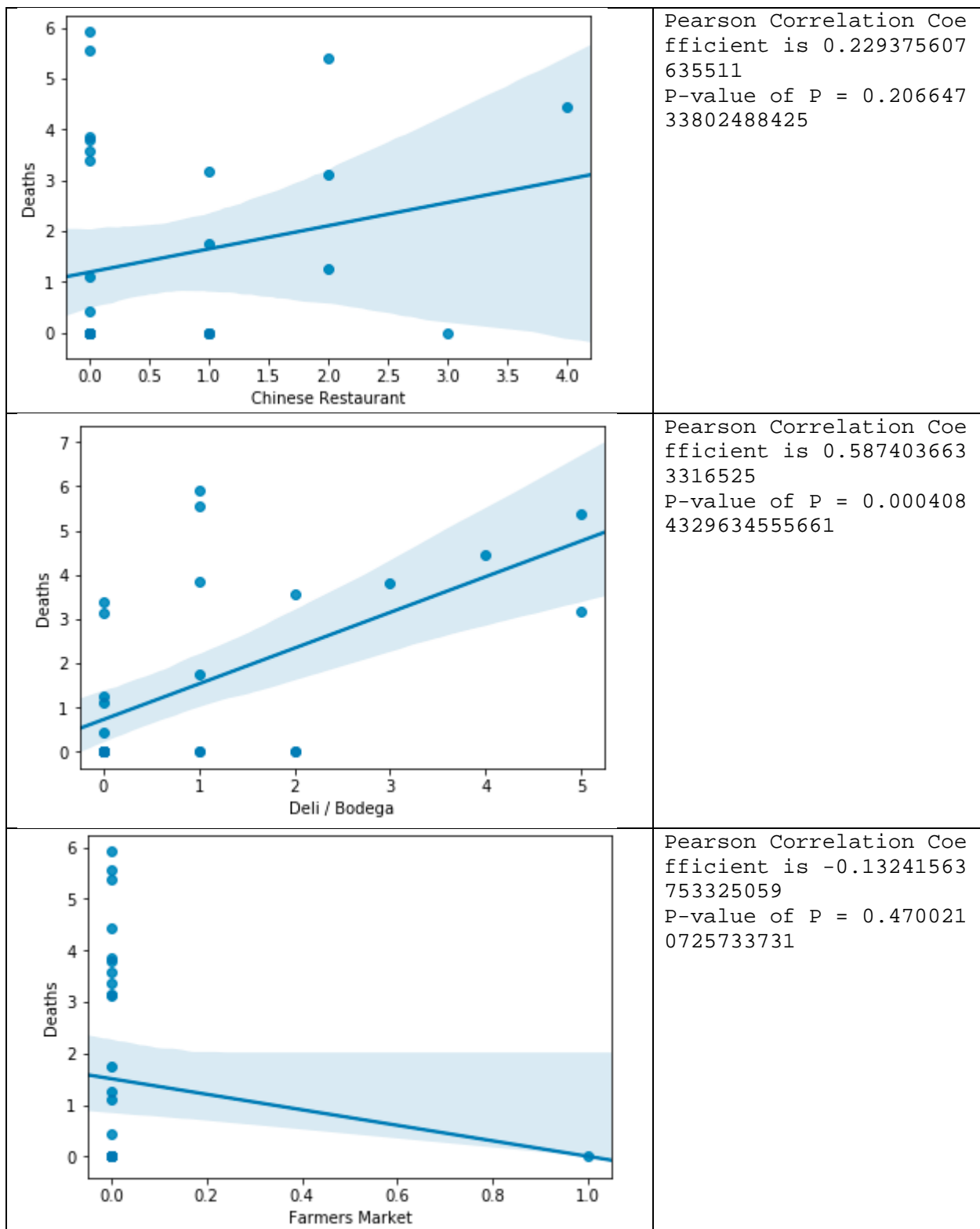


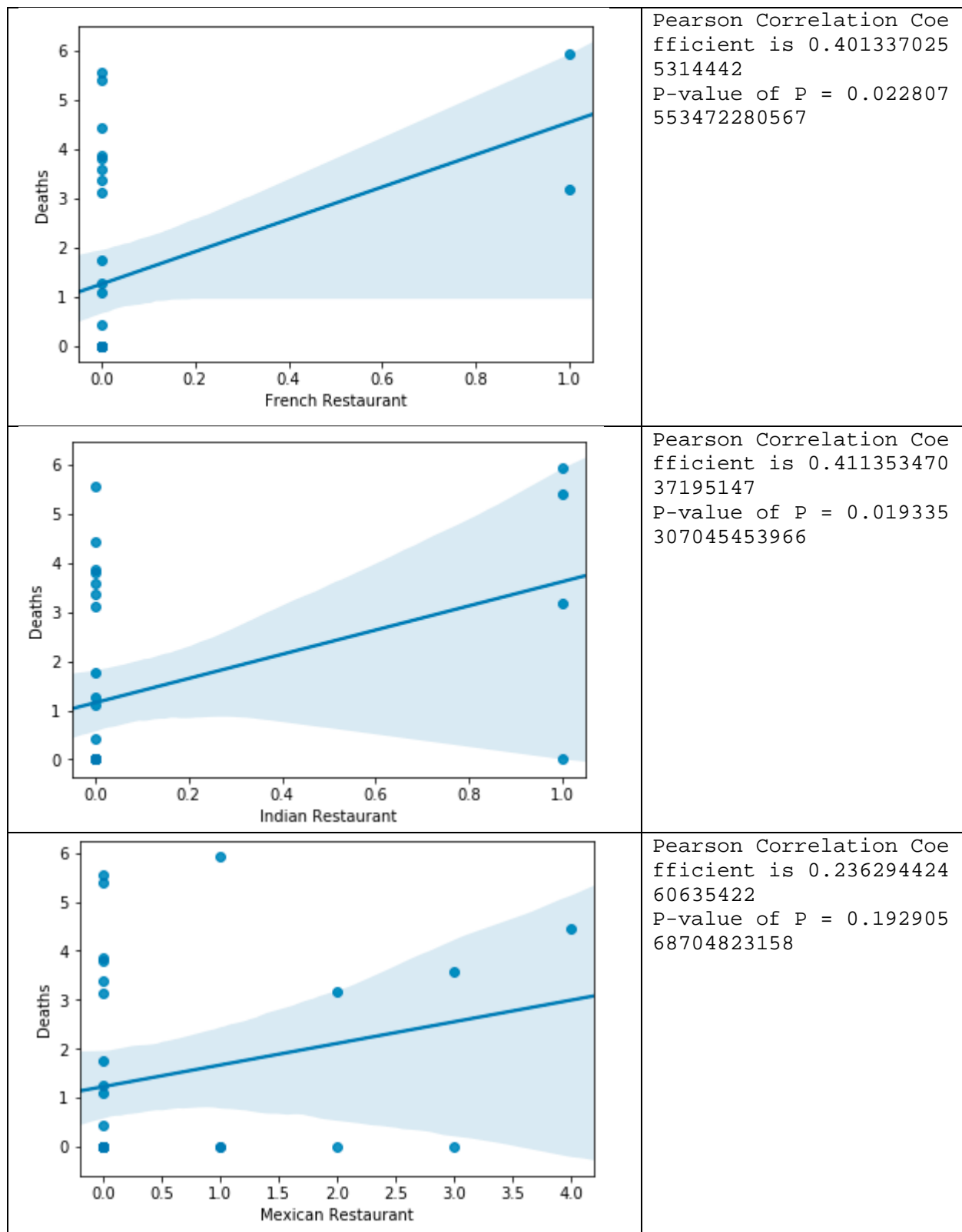
Pearson Correlation Coefficient is 0.28630187327331
P-value of P = 0.09048785946974881

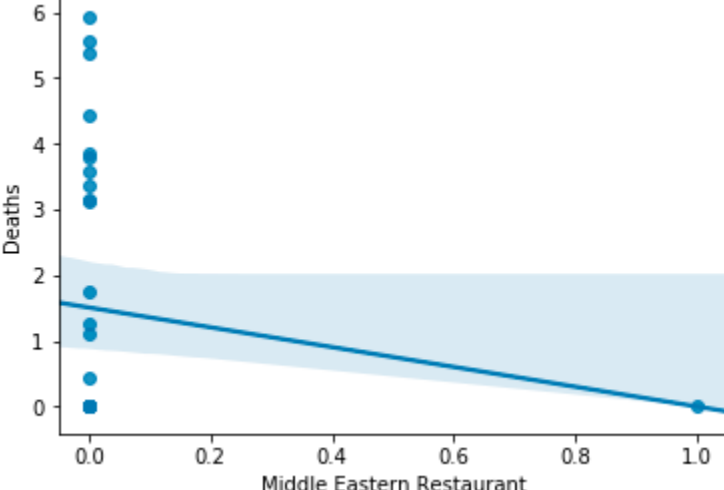
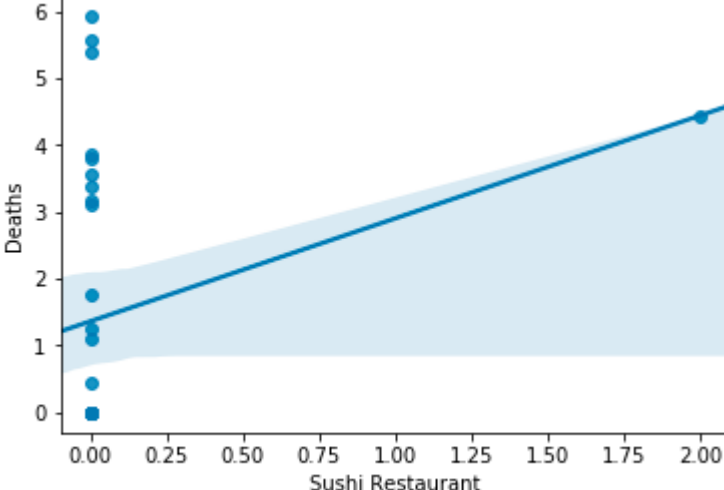
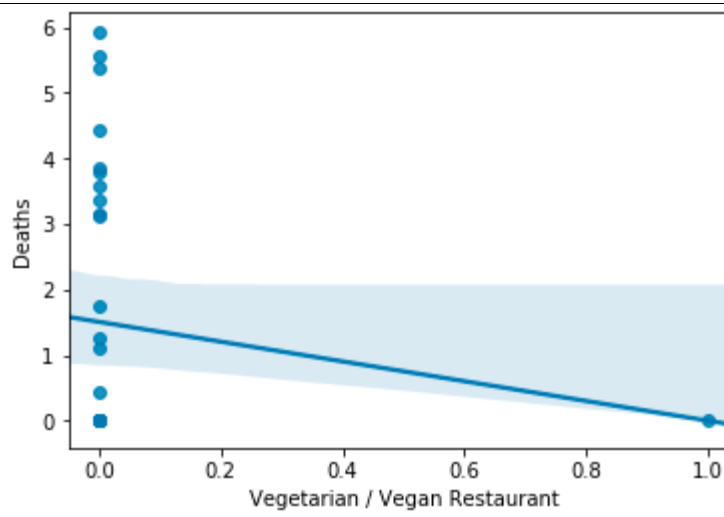


Results – Correlation of Number of Certain Food Venues with Number of COVID-19 deaths







| | |
|---|--|
|  | <p>Pearson Correlation Coefficient is -0.13241563753325059 P-value of P = 0.4700210725733731</p> |
|  | <p>Pearson Correlation Coefficient is 0.27005063473357344 P-value of P = 0.13496918694909094</p> |
|  | <p>Pearson Correlation Coefficient is -0.13241563753325056 P-value of P = 0.4700210725733731</p> |

Observation

The highlighted cells in first results table show a strong correlation of number of various travel and night life spots with number of COVID-19 cases. However, no results can be drawn from the 2nd analysis where the number of certain food venues were studied for their linkage to number of deaths due to coronavirus.

The numbers of Cocktail bar, Gay Bar, Metro Station, Piano Bar, and Tiki Bar were found to be positively correlated with the number of COVID-19 cases in the respective county according to the obtained Pearson coefficient and p-values.

Recommendation

A multi-linear regression model can be built for the travel and night life spots venue categories. Also, Principal Components Analysis (PCA) can be used here to perform dimensionality reduction to study the most related venues with the number of cases and deaths.

Conclusion

In this project, number and abundance of various kinds of venues in different counties of New York was studied in relationship to the number of cases and deaths due to COVID-19.

The exploratory data analysis technique of correlation was performed for this research and the resulting relationships were visualized by Scatter plots, and determined by Pearson coefficients and p-values.

Some popular night life venues and Metro Station transportation venues were observed to be the cause of a plethora of cases in the state of New York.