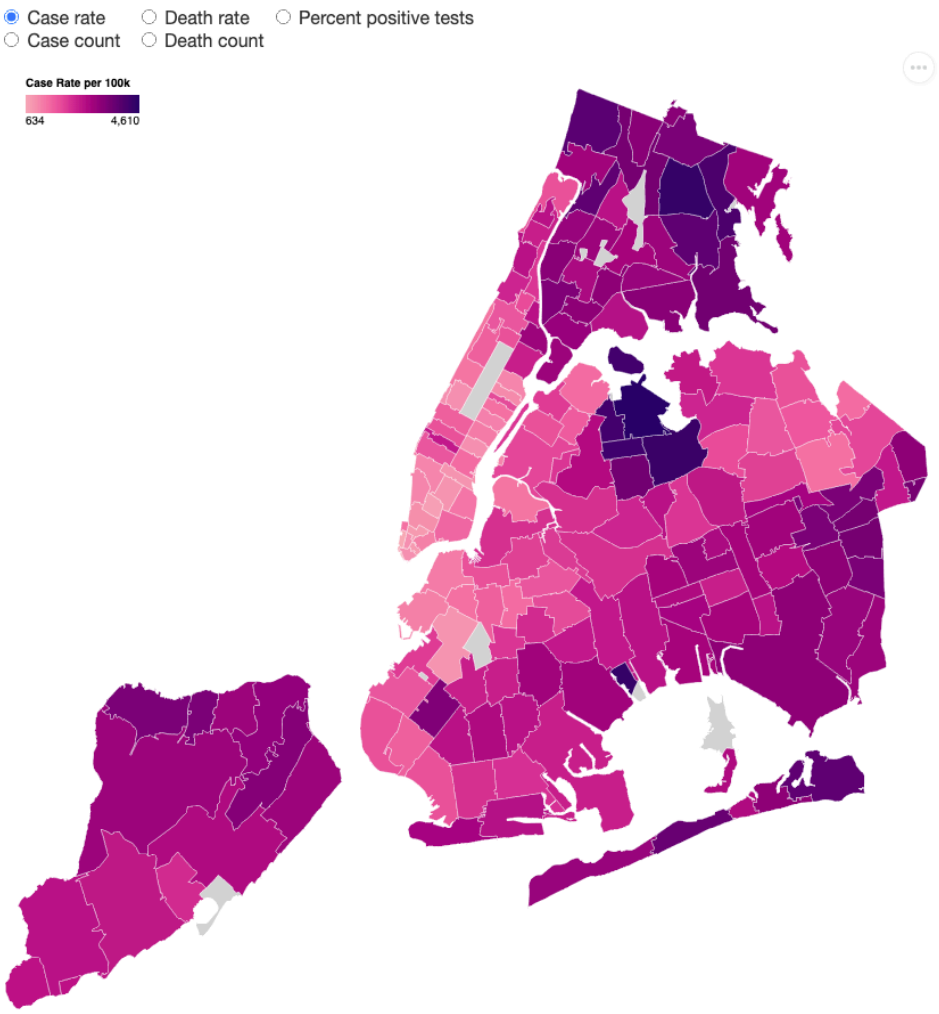
The Composition of Venues in NYC neighborhoods and Their Relationship with COVID-19 Infection rates -- IBM Capstone Coursera

Introduction:

COVID-19 has devastated the world and one of the epicenters of this pandemic is New York City. While many studies and research has been conducted on the spread and infection rates of the virus, I wanted to try an analysis of how the type of venue in a neighborhood is related to infection rates of a neighborhood.

As New York City enters into the later phases of re-opening, I believe it's important to see if there is any correlation between being opened venues (restaurants, bars, coffee shops) and infection rates. The NYC government and the residents/businesses of NYC could benefit from the insights of this analysis through the recommendation of different strategies of reopening, the development of contingency plans, and administration of safety protocols.



Research Problem:

In this project, we will explore the neighborhoods of New York City:

- Whether the type of venues can affect the infection rates of a neighborhood
- Whether the concentration of venues could affect the infection rates of a neighborhood
- Whether we could use this information on recommending specific strategies of reopening or the development of contingency plans for the people and businesses affected by the pandemic.

Data Sources:

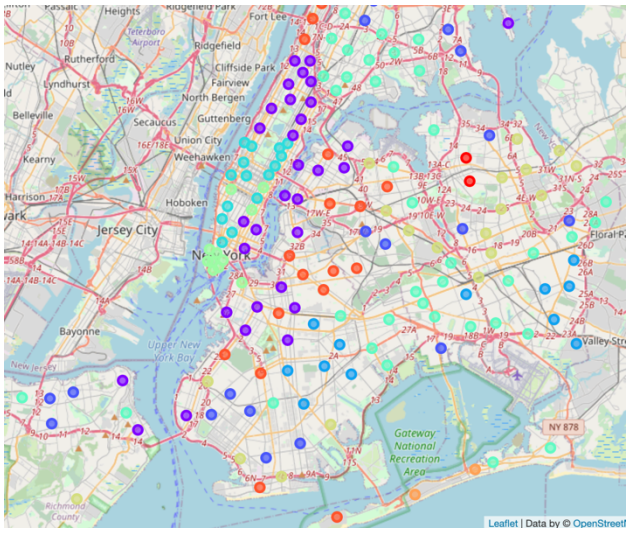
The main data used for this project will be from these three sources:

- The infection rate of New York city neighborhoods. (NYC.gov)
- Official GitHub of NYC Health data sources. (Github NYC Health)
- The venues in each neighborhood. (FourSquare API)

*Note: Understandably infection rates may not be reported at where the infected person caught the virus (i.e. the venues in question); however, they could reveal how venues in proximity where people live could affect infection rates

Methodology:

1. Infection rates by neighborhood will be extracted from NYC Data's website
2. For each neighborhood/zip code, Geocoder will acquire its coordinates
3. For each neighborhood/zip code's coordinates, Foursquare API will acquire all surrounding venues' information, such as "venue type"
4. The composition of types of venue for each neighborhood will be analyzed against the infection rates
5. PCA/Clustering algorithms will be utilized to generate clusters of neighborhoods
6. Venue counts and clusters generated can be used in regressions to see how well they can be used predict infection rates

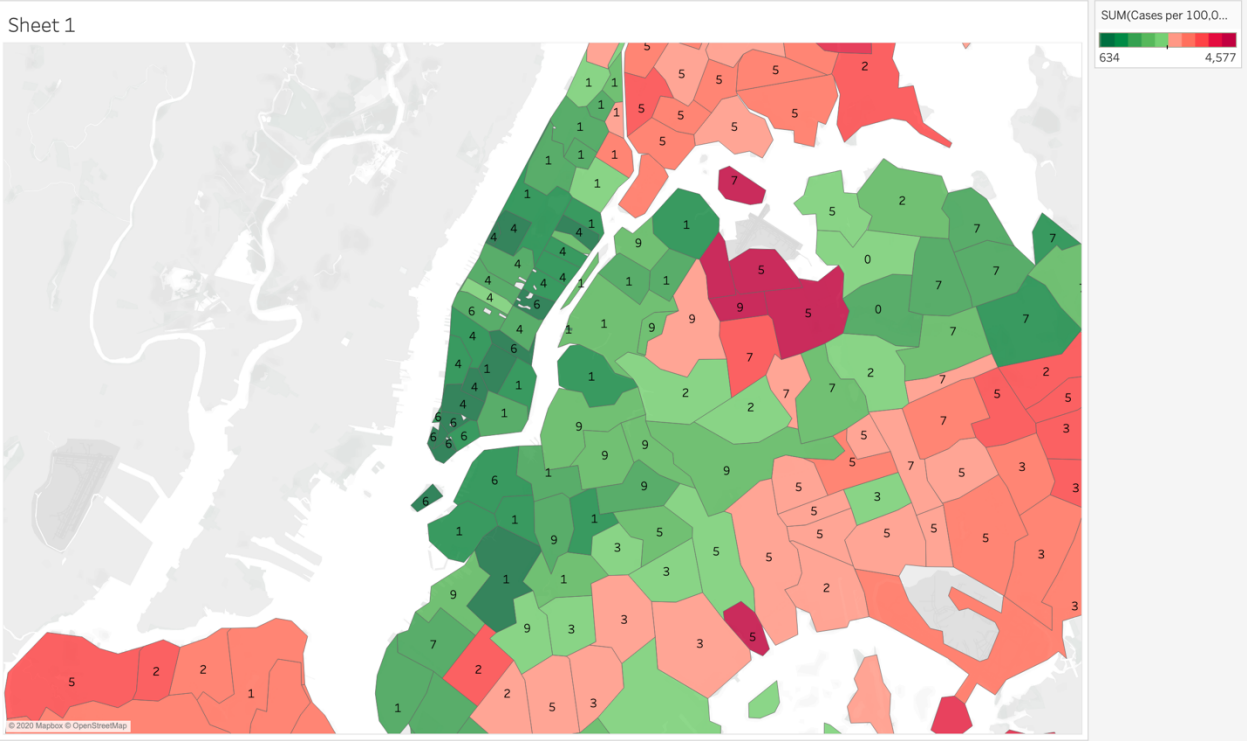


Results:

Using simple linear regression on individual venues did not yield a strong predictive model, resulting in a low R2 of 0.13; however, clustering neighborhoods based on venues yielded a stronger predictive model, resulting in a much higher R2 of 0.578.

It would appear that the combination of certain venues in a group gives more predictive power than individual components of venues. I decided to go further and utilized PCA, which is an algorithm that reduces the dimensionality by utilizing the correlations of certain venues to each other into clusters.

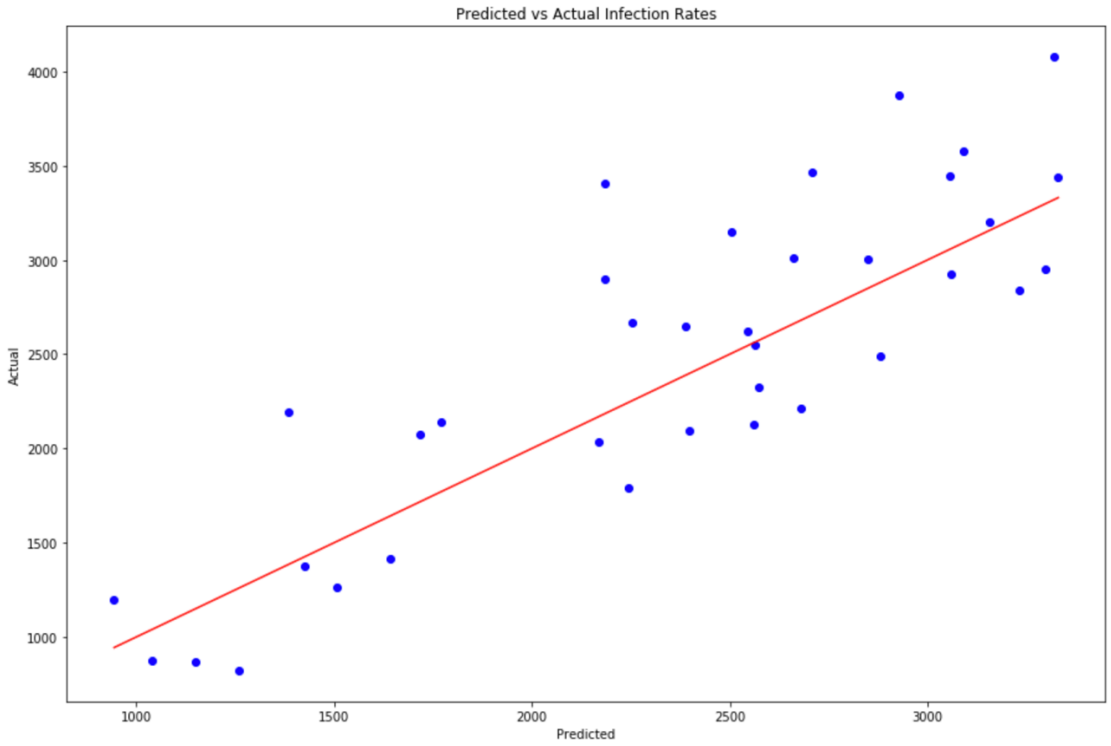
Manhattan Neighborhoods Clusters and Infection Rates:



PCA Results:

Utilizing PCA and linear regressions, I am able to find a moderately strong relationship between certain venues and infection rates of an area. After optimizing for the number of **components (8),** the model yielded an **R2 of 0.70 and a Mean Absolute Error of 384**, which can be interpreted that 70% of the variance of infection rates across neighborhoods could be explained/predicted from the composition of a neighborhood's top venues.

PCA Actual vs. Predicted Infection Rates:



Discussion:

According to my analysis, it appears that neighborhoods that have establishments that are frequented by lower socioeconomic populations, such as **Fast Food Restaurants and Bodegas** have higher rates of infections than neighborhoods that have venues that cater towards higher socioeconomic populations, such as **Cycle Studios and Wine Bars**. Speculating from this observation, it appears that people from poorer socioeconomic backgrounds are more negatively affected by this pandemic.

VS

Conclusion:

By applying this model to any given set of coordinates or neighborhood, we can reasonably predict how drastic the effects of the pandemic could be in terms of infection rates based on the composition of the venues in the area. We can then allocate resources, such as medical supplies and personal to areas in most in need. We also need to be mindful that poorer communities are still at risk and should be closely monitored, as NYC reopens. Perhaps, one course of action would be to place more testing facilities and medical resources in neighborhoods at risk and placing more careful reopening guidelines on those neighborhoods at risk to reduce the spread of the pandemic.