**Predicting the Number of Covid-19 Cases Based On Neighborhood Venues**

Robin Zhao

**1. Introduction**

**1.1 Background**

During 2020, the number of COVID-19 cases has surged across the state of New York. After strict COVID-19 safety guideline enforcements, the infection rate in NYC gradually decreased, and New York city was starting to open up. The reopening of the New York City has four phases, and in around July 2020, New York was entering phase three of the reopening, where outdoor dining, local gyms, shopping malls and other entertainment facilities were reopened. However, with these facilities opening, the COVID-19 infection rate came up simultaneously. In January 2021, New York City’s restaurant’s indoor dining were forced to close again because of the surging virus. There is clearly a positive correlation between the reopening businesses and the virus infection rate. With different infection rates in different neighborhoods of New York City, it would be interesting to examine how different venues in a neighborhood affects that neighborhood’s infection rate as well as if number of venues can effectively predict the neighborhood’s infection rate.

**1.2 Problem**

Different venues in a neighborhood can affect the overall neighborhood’s infection rate. This paper is to examine the relationships between venues and the neighborhood’s infection rate and try to effectively predict the neighborhood’s infection rate based on the venue lists.

**1.3 Interest**

The government would be very interested in an accurate prediction of COVID-19 infection rates in different neighborhoods in order to come up with strategies to combat it. Moreover, individuals would be very interested too because it can give them some precautions in traveling into another possibly highly infected neighborhood.

**2. Data Acquisition**

One dataset is found on NYC Health website in the Percent Positive and Test Rate of Molecular Testing by ZIP code section. This dataset is able to give us the information of the neighborhood name, its ZIP code, and its infection rate. Another dataset is found on opendatasoft in the US Zip code Latitude and Longitude section. This second dataset is able to give us the coordinates for each neighborhood by its ZIP code, and hence we can us Foursquare to acquire the venue lists for each NY neighborhood by its latitude and longitude. Finally, we can conduct our analysis on the venue lists and the infection rates.