Applied Data Science Capstone

Rita Lin

04/27/2020

1. Introduction

Coronavirus disease 2019 (COVID-19) is an infectious disease caused by the severe acute respiratory syndrome coronavirus (SARS-CoV-2) that outbroke in December 2019. As of April 27th, 2020, more than 3 million people around the world were infected with coronavirus, among which more than 1 million people are from the United States. So far, the total number of death toll around the world has exceeded 200,000. Therefore, it is of importance for the public to understand what factors are related to the number of COVID-19 cases, so that effective measures can be taken to reduce the risk of exposure to coronavirus.

The targeted audience of this work is the general public who are interested in learning the relationship between daily activities and COVID-19 infections. This work aims to answer if different *venue types* in the neighborhoods is related to the number of COVID-19 cases. New York City, the epicenter of the coronavirus outbreak in the US, will be studied as an example. Besides *venue types* in in New York City neighborhoods, *population* and *population density* will also be explored as factors for COVID-19 infections on the county level across New York State.

2. Data

The data used for this work includes the following:

- County information in New York State, scraped from Wikipedia. *Population*, *population density* and *FIPS code* data will be extracted from this source.
- COVID-19 statistics for New York State, obtained from the New York State government website.
- COVID-19 statistics for New York City, obtained from NYC Department of Health and Mental Hygiene.
- County and zip code coordinate data, obtained from github open sources.
- Venue information in New York City, obtained using Foursquare API.