2. The data

To test our hypothesis, we will use geolocation data (based on zip codes), venue data (obtained through the Foursquare API), and epidemiological data for COVID-19. For model development, we will use the state of New York. The state of New York has several characteristics that make it a desirable target for our preliminary investigation. First, New York is the epicenter of COVID infections in the United States, alone corresponding to more than a quarter of all infections[1]. This led to a massive testing effort, with some of the most comprehensive data found in the country[10]. Second, in addition to New York City, New York state has a huge upstate area with numerous counties, allowing us some of the largest datasets possible. Third, New York state has a comprehensive and well-polished database system, with detailed populational[11], geographical[12], and epidemiological[13] data that can be obtained through the Socrata API (https://dev.socrata.com). Therefore, the data we will need can be easily imported and manipulated in the program.

Epidemiological data and populational data will be used in exploratory analysis to verify some of the premises of our hypothesis. In particular, we want to verify if: there is substantial variability between counties in the number of infected, even when corrected by population; there is substantial variability between counties in the number of tested, even when corrected by population; and if there is substantial variability between counties in the rate of infected (positive cases/tested cases). Substantial variability between counties is a prerequisite for this model. We will also verify correlation factor between these parameters, to understand how population density can affect the spread of COVID-19. If population correlates too strongly with COVID-19 infection, it will eclipse and render our model irrelevant.

If the assumptions for our model are verified, we will obtain geolocation data using zip codes as proxy. While NY city has detailed information for neighborhoods, including geolocation information, the rest of the state does not have such granular data. Therefore, we will use zip codes to obtain latitude-longitude pairs for each zipcode, verify data integrity, and devise a strategy to ensure maximum area coverage for each of these points while minimizing overlapping. Once these "query" areas are obtained, we will employ the Foursquare API (https://developer.foursquare.com) to obtain the types of venues found in each area, quantify them, and then obtain the relative presence of each venue type per zip code and per county. The relative presence of each venue will then be used as independent variables (feature set), while either the rate of infection (continuous variable) or a risk classification (categorical variable) will be used as dependent variables (target set). Models will be trained and tested using NY data and evaluated based on accuracy and recall metrics. Finally, if a model is found that has reliable scores, it will be trained on the whole set of NY data and then tested against data from another state (Illinois[14]) to evaluate how transposable the model is between states.

[10] Source: https://www.politico.com/interactives/2020/coronavirus-testing-by-state-chart-of-new-cases/
[11] Source: https://data.ny.gov/Government-Finance/Annual-Population-Estimates-for-New-York-State-and/krt9-ym2k
[12] Source: https://data.ny.gov/Government-Finance/New-York-State-ZIP-Codes-County-FIPS-Cross-Referen/juva-r6g2
[13] Source: https://health.data.ny.gov/Health/New-York-State-Statewide-COVID-19-Testing/xdss-u53e
[14] Source: https://www.dph.illinois.gov/covid19