Introduction

The effects of and each area's reaction to the COVID-19 pandemic have been far from homogenous. As cases in the United States surge, it is urgently important that we better understand why this virus affects some populations more than it does others. With the volume of data and timeliness of the subject makes this an apt project for data science. Using available data, we can determine how the popular locations of each area are related to that area's COVID-19 burden.

The problem of not fully understanding how this coronavirus spreads is a pressing matter to local, state, national, and global health departments. Uncovering how hyperlocal (per zip code) differences in popular venue categories are affecting COVID-19 cases can help inform health department guidelines and help predict future hotspots. Because this project only deals with zip codes within Alameda County, it would be of immediate importance to their health department but expanding the scope beyond one county would enhance the statistical power.

Data

I will use two datasets to elucidate the relationship between confirmed COVID-19 cases and popular local venues in Alameda county. COVID-19 rates are hosted, and updated daily, by Alameda Public Health at: https://services3.arcgis.com/1iDJcsklY3l3KIjE/arcgis/rest/services/AC_Rates_Zip_Code/FeatureServer/0/query?where=1%3D1&outFields=*&outSR=4326&f=json. For each of the 53 zip codes in Alameda county, this dataset contains the city, COVID-19 case count, population, and COVID-19 case rate. The provided zip codes can be converted to latitude and longitude by GeoPy for the FourSquare API request. My analysis only requires the zip code and case rate from this dataset.

The other dataset is the result of using the FourSquare API to find recommended venues in each zip code. This requires a "GET" request to the FourSquare API with the "explore" endpoint. I can then pair recommended venues with COVID-19 case rates from the first dataset and analyze how popular venues affect COVID-19 cases.