CS4331 Project I Report

Enrique Cervantes, Jacob Zahn

The goal of this project is to gain insight into and attempt to model the county level number of cases and deaths of the COVID-19 outbreak in America. We intend to combine two datasets; 2019 county population estimates from [census.gov](https://www2.census.gov/programs-surveys/popest/datasets/2010-2019/counties/totals/) it’s [Data Sheet](https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2019/co-est2019-alldata.pdf) and information on hospitals in the U.S. from [hifld-geoplatform.opendata.arcgis.com](https://hifld-geoplatform.opendata.arcgis.com/datasets/hospitals). Combining these datasets gives us information on a county level regarding the number of hospital beds, population, hospital helipads, private, government, and nonprofit hospitals. This dataset can then be combined with [county level data](https://github.com/nytimes/covid-19-data) from the New York Time’s GitHub. This data is organized by date and county containing cases and deaths. Combining these datasets creates a time series dataset of cases and deaths in the United States organized by county that includes hospital and population information. State level testing and recovery numbers may be added to further enrich the dataset.

Data Preparation:

The first goal of this phase was to merge and then clean up the data. This is done in combineCountyHospital.py written by Jacob Zahn. This script is only meant to be run if CountyHospitalCombined.csv is not made, as it takes a while to run. The python script interprets the Hospitals.csv file and its variables, to combine them with the population provided by census projections and geo location provided by the Nominatim api of counties in the U.S. Another script is provided to make changes to CountyHospitalCombined.csv, updateCountyHospital.py which reuses the geo location data. This completed our initial merging of the data.

The first step of the cleanup was to add an index field to all the data sets, so we can track each individual column. The next step was to remove misleading values. These values mainly showed up in the Hospitals.csv data set, as we had fields that had -999 as a quantity of beds, population, and staff. The final step of the cleanup was to add a z-value for each numeric attribute in the data set.

Next the python script combineCountyHospitalCovData.py written by Jacob Zahn is used to merge Covid-19 cases and deaths as reported by the New York Times in us-counties.csv with CountyHospitalCombined.csv to create CovCountyHospitalTimeSeries.csv. The shape of the dataset is (counties, attributes+2(D)), where D is the difference between January 20th, 2020 the date with no cases in the U.S. and the lasted date in our copy of us-counties.csv. This is the dataset we will use for the model in project 2.

Data Exploration:

Explore univariate relationships between predictors and the target variable:

Figures



