Project I Report

Enrique Cervantes, Jacob Zahn

CS4331 Dr. Mengel, Texas Tech University

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. This was done with the “Data Preparation part 1.R” and “Data Preparation part 2.R” scripts written by Enrique Cervantes. The R scripts return the new sets of data New\_CountyHospitalCombined.csv, New\_Hospital\_data.csv, and new\_us\_stats\_covid19\_daily.csv files.

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 further data exploration and the model in project 2.

Data Exploration:

Exploration of frequencies among predictive variables is done in the set of bar graphs with the binned variables to enhance the model. This was done in the “Exploratory Phase.R” written by Enrique Cervantes. The first bar graph tells us about the frequency of beds in hospitals. We can figure out from the data the amount of large hospitals and smaller ones based on the amount of beds available. The second bar graph tells us the frequency of helipads among hospitals. The third graph tells us about the frequency of the size of the population’s among counties.

The exploration of univariate relationships is done with a set of scatterplots done in the “Exploratory Phase.R” written by Enrique Cervantes. The first set of scatter plots look for relationships among beds vs. number of cases/deaths. While beds vs cases show barely any correlation, you can spot a trend with the number of beds and number of deaths. As the number of beds go down the death rate goes up. The second set of scatter plots looks for relationships among population vs. number of cases/deaths. Both plots indeed show a trend where the number of cases/deaths increase as population within the counties go up.

Exploration of multivariate relationships in our data is done in the first plots of dataExploration.py written by Jacob Zahn. The data plotted for this section is a scatter plot of beds versus population (two highly correlated predictors) colored by cases/deaths respectively. These plots are included in the Figures section below. These plots show that as population increases beds tend to increase, but that cases/deaths are not overly correlated with population or beds. Additional multivariate relationships explored were types of hospital versus population colored by cases/deaths respectively. Plots for the cases versions of this graph are also included in the Figures section below. These plots mirror some of the trend between beds and population, as beds are dependently related to hospitals.

A new variable explored from the dataset is total hospitals per ten thousand people, which is a summation of the government, private, and non-profit hospitals in a county divided by the population and then multiplied by ten thousand. Another new variable is hospital beds per ten thousand people, which is beds per county divided by population and then multiplied by ten thousand. Charts of beds per ten thousand people versus cases/deaths respectively are shown in the figures section below. These show a trend where lower numbers of beds per population is sometimes correlated with a higher number of cases.

Figures





