Jai Mehrotra-Varma
12/02/22

Homework 3

I collected two sources of data - both Kaggle datasets. One dataset was composed of hospitalized COVID-19 positive patients, while the other dataset was composed of hospitalized COVID-19 negative patients. The COVID+ dataset had 566,602 unique patients, while the COVID- dataset had 1176 unique patients. I am currently searching for a larger COVID- dataset that would allow the comparison to be more accurate; however, if one is not available, I will attempt to do more in-depth analyses to account for the vast difference in size between the two datasets. Both datasets provide demographic and hospital-related information about each patient (ex: their age, gender, comorbidities, lab values, etc.). Since my research question is to determine whether hospitalized COVID+ patients have a higher mortality rate than hospitalized COVID- patients, I have filtered the data to just include 1) their age, 2) their gender, 3) their hypertension status, 4) their diabetes status, 5) their COVID status (which I eventually added manually to the pandas dataframe, and 6) their mortality outcome. The link to the COVID+ dataset is: (https://www.kaggle.com/datasets/tanmoyx/covid19-patient-precondition-dataset?select=covid.csv),
and the link to the COVID- dataset is:
(https://www.kaggle.com/datasets/saurabhshahane/in-hospital-mortality-prediction).

Sample Data:

Both of my datasets were downloaded from Kaggle as CSV files. The original CSV files were much larger (COVID+ dataset had 566,602 unique patients, while the COVID- dataset had 1176 unique patients), so for the sample data, I created two new CSV files selecting 1000 patients from each dataset and saved them to the data folder. They give a good sample overview of what the data looks like (the full datasets can be downloaded from the links above).

Github Repository Link:
https://github.com/biony1209/DSCI-510-Final-Project