

Seth Hillis
DSCI – 235
3/21/2022
Final Project Proposal

COVID-19 US CASES AND DEATHS

For this project I have found a set of data that contains 20+ CSV files that contains data on COVID-19 such as states and county breakdowns of the cases, but also, as well as data on colleges, prisons, mask use and deaths. I believe that this data set will provide an appropriate challenge for the techniques that we have learned in this class. The data can be found at this link <https://github.com/nytimes/covid-19-data> and is thankfully already collected, but I will need to clean features that are not relevant to my questions and combine the multiple data bases that are provided in the data into one final data frame of the cleaned data.

Based on the data I want to answer the following questions; How do the number of cases in each state compare to the college cases for schools in the same state? How about the prison cases? How did the infection rate of COVID change from 2020, 2021, and 2022, are there patterns in the spread?

To ensure that this project is finished in time my plan is to have a clean set of data saved to a new CSV for each of the data frames that I plan to use for the project clean within a week after proposal approval. From there I plan to spend 4 to 6 weeks graphing, visualizing, and interpreting my data to answer the questions listed above. The final step will be 1 to 2 weeks that I spend formatting and preparing my results for the presentation in a jupyter notebook.