STAT 3355.001 COVID 19 Data Analysis: Project Proposal
Group DCB: Roopali Kallem, Kenan Stredic, Jacqueline Rodriguez

Dataset Information:

Our dataset contains daily Covid-19 reports from both global and local sources. This includes information about the number of cases and deaths from Covid-19 between 2021 and 2022. Along with this, the dataset also includes the exact locations and dates of these cases.

Availability: This dataset is part of the JHU CSSE COVID-19 repository, which is available for free on Github. This site does not require an account to access the information.

Constraints: Our dataset contains 192,754 observations, which exceeds the >5000 requirement, and 29 variables, which meets the >10 requirement.

Background and Motivation:

For this project we will use information from various data sets to gain a better understanding of the situation various countries and the United States faced between the years 2021 and 2022 during the Covid Pandemic. The datasets we chose for this analysis were the 2021 Covid19 US Daily Report, the 2022 Covid19 US Daily Report, and the WHO Covid19 Global Data all from Github. These datasets were chosen because of their clarity and overlapping information. These datasets analyzed the pandemic within the years 2021 and 2022 as well as cover key details of the pandemic such as the number of new and active cases, deaths, and various locations.

Having lived through the COVID-19 Pandemic there was little time to process how the world and the United States were affected. Therefore, in this project we aim to analyze datasets regarding the number of confirmed cases and deaths in the United States, and compare it with that of other countries. This will bring us to a better understanding of the longstanding effects left on the people of the world as well as the state or country at most disadvantage due to being most impacted by the pandemic.

Questions:

- Which countries were most affected by the pandemic with regard to confirmed cases and deaths?
- Between the two years is there any pattern between the number of cases and the number of deaths? (increase in cases in one year results in less deaths the following year??)
- How can this data best be modeled to display active cases, confirmed cases, and deaths?
- How the number of cases increased or decreased as the time progressed(month wise or quarter wise)?
- How the number of cases, recovery and deaths changed within the US from year 2021 to 2022, as all businesses were back up?
- Which state had the highest cases for the month?
- Is there a pattern in the number of cases or recovery rate within the US?
- Is there any notable relation in the number of cases with the countries that are close to each other?(use the long, lat)
- Try to predict the percent increase on an interval, to understand if the spread was ever constant.
- Which state or country stands at the greatest disadvantage due to being most impacted by the pandemic?

Link to the data:

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_daily_reports_us/01-01-2021.csv

https://github.com/CSSEGISandData/COVID-19/blob/master/csse_covid_19_data/csse_covid_19_daily_reports_us/01-01-2022.csv

https://covid19.who.int/WHO-COVID-19-global-data.csv