**Project 1: The Impact of COVID-19 on the US by State**

*By Mohamed Aamer, Kevin Jeong, Tracey Geneau, Ana Torres*

**Abstract**

This research project aims to conduct an investigatory data analysis of COVID-19 cases across the United States from 2020 to 2022. By looking at data taken from Data.gov, this project will seek to answer several questions about the impact of COVID-19 on the US by state as well as aggregating the data to observe impact by region: the Northeast, the Midwest, the South, and the West. Subsequently, the data would then be presented in a visually comprehensive manner. In doing so, the project posits the following questions to answer:

* What were the total number of COVID-19 cases and COVID-19 deaths by region over the 2020-2022 period?
* How did each region experience the spread of COVID-19 over time?
* What was the rate of death caused by COVID-19 for each state? Which 5 states experienced the most death caused by COVID-19 as a % of total cases? Which 5 states experienced the least death caused by COVID-19 as a % of total cases?
* Can we predict a given state’s COVID-19 spread in 2023 following a scatter plot’s regression analysis of new COVID-19 cases reported in that state between 2020-2022?

The dataset used was pulled from the US government’s data.gov website which offers data sets in several formats, including CSV: <https://catalog.data.gov/dataset/united-states-covid-19-cases-and-deaths-by-state-over-time>

**Tasks**

* Data cleaning/munging
* Aggregating COVID-19 case data (total number of cases, total number of deaths) for each state into a list for each given region as understood by the US Census Bureau (<https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf>)
* Presenting the regional data in easily understandable data visualizations (bar charts, line graph showing each region’s case numbers over time)
* Calculating the rate of death caused by COVID-19 as a proportion of total COVID-19 cases for each state in the data set
* Presenting the 5 states with the highest fatalities and the 5 states with the lowest fatalities as bar charts
* Selecting a specific state and creating a scatter plot reflecting the number of new COVID-19 cases reported in the data set over time
* Building a regression analysis model of that scatter plot