# Project 3: Data Wrangling

BYU STAT 250

## Introduction

In this project you will investigate the effect of income and population density on COVID-19 case counts in the United States during 2021. You will use web scraping and data cleaning techniques to collect and merge data from multiple sources. The goal is to explore how these factors relate to COVID-19 case numbers and to present your findings using data visualizations and a written analysis.

## Data Sources

You will work with three primary data sources:

1. **COVID-19 Data:** A CSV file `https://drbob-richardson.github.io/Stat250_W2025/data/covid.csv` containing COVID-19 case counts (and, if available, population data) for U.S. states in 2021.
2. **Landsize Data:** Data on state size scraped from the following URL: `https://statesymbolsusa.org/symbol-official-item/national-us/uncategorized/states-size`
3. **Income Data:** U.S. states' median household income data available on Wikipedia at: `https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_income`

## Project Tasks

### Task 1: Data Import and Web Scraping

- **COVID Data:** Read in the provided `covid.csv` file.
- **Income Data:** Use the `rvest` package in R to scrape the table from the Wikipedia page that contains median household income data.
- **Landsize Data:** Scrape the state landsize information from the landsize URL.
- **Inspection:** Examine each dataset to understand its structure and identify common keys (e.g., state names).

## Task 2: Data Cleaning and Merging Landsize with COVID Data

- **Standardize Keys:** Clean and format the state names in all datasets so they match consistently.
- **Merge Datasets:** Create a dataset that combines state landsize and COVID-19 case counts for all 50 states.
- **Discussion:** Explain why simply adding columns with a `bind_cols` command does not work in this scenario. Discuss which join functions (e.g., `left_join`, `inner_join`) are appropriate for merging these datasets and why.

## Task 3: Analyzing Population Density and COVID-19 Cases

- **Calculate Density:** Compute the population density for each state as:

$$\text{Population Density} = \frac{\text{Population}}{\text{Landsize}}$$

- **Visualization:** Create a scatter plot showing COVID-19 case numbers versus population density.
- **Interpretation:** Discuss any trends or relationships you observe between population density and COVID-19 case counts.

## Task 4: Merging Income Data with COVID and Landsize Data

- **Combine Datasets:** Merge the income data (scraped from Wikipedia) with the dataset from Task 2. Ensure that the state names are aligned correctly.
- **Final Dataset:** Create a comprehensive dataset that includes COVID-19 case counts, state landsize, population density, and median household income for the 50 states.
- **Challenges:** Describe any challenges or issues you encountered during the merging process and how you addressed them.

## Task 5: Exploring the Relationship Between Income and COVID-19 Cases

- **Visualization:** Create a scatter plot showing COVID-19 case counts against median household income.
- **Analysis:** Analyze and interpret the plot. What does the relationship (or lack thereof) suggest about the effect of income on COVID-19 case counts?

## Submission Guidelines

- Submit a PDF report generated from your final analysis.
- Ensure that your code is displayed and is well-commented and organized.
- Your final report should include:
  - All code used for data scraping, cleaning, merging, and visualization.
  - The generated plots.
  - A written summary of your findings and reflections on the project.
- Collaboration is allowed; you may work individually or in groups, but you must submit your own analysis and code.

## Additional Notes

- Pay close attention to details in web scraping; sometimes tables require significant cleaning before they are usable.
- Consider potential pitfalls when merging datasets from different sources (e.g., inconsistent state names, missing values).
- Use appropriate visualization techniques to effectively communicate your findings.

Good luck with your project!