# Description of Health Outcome Socio-Econ Data

This dataset contains a wealth of health-related information and socio-economic data aggregated from multiple sources such as the American Community Survey, clinicaltrials.gov, and cancer.gov, covering a variety of US counties. Your task is to use this collection of data to build an Ordinary Least Squares (OLS) regression model that predicts the target death rate in each county. The model should incorporate variables related to population size, health insurance coverage, educational attainment levels, median incomes, and poverty rates. Additionally, you will need to assess linearity between your model parameters; measure serial independence among errors; test for heteroskedasticity; evaluate normality in the residual distribution; identify any outliers or missing values and determine how categories variables are handled; compare models through implementation with k=10 cross-validation within linear regressions as well as assessing multicollinearity among model parameters. Examine your results by utilizing statistical agreements such as R-squared values and Root Mean Square Error (RMSE) while also interpreting implications uncovered by your analysis based on health outcomes compared to correlates among demographics surrounding those affected most closely by land structure along geographic boundaries throughout the United States

https://www.kaggle.com/datasets/thedevastator/uncovering-trends-in-health-outcomes-and-socioec

## Research Ideas

- Analysis of factors influencing target deathrates in different US counties.
- Prediction of the effects of varying poverty levels on health outcomes in different US counties.
- In-depth analysis of how various socio-economic factors (e.g., median income, educational attainment, etc.) contribute to overall public health outcomes in US counties