COGS 109 Analysis Pipeline

Group: Alexander Schonken, Bingzhe Wang, Jolene Leung

Dataset Description:

- 577 samples, 99 features
- Each sample is from a specific year and state
- Examples of Features: year, state, arts/music GPA, family income, SAT score, etc.
- Link to dataset: https://corgis-edu.github.io/corgis/csv/school_scores/
- Previous Analyses: male vs. female scores, state scores comparison, correlation between art/music GPA and high SAT score

Data Exploration and Visualization:

- Plot the SAT scores in each state across the United States of America
- Plot the GPAs (art, math, etc.) in each state across the United States of America
- Plot GPA vs. SAT score on an xy-coordinate system
- Plot bar graph of states along the x-axis and two bars per state (GPA, SAT)

Research Question:

- Can you predict SAT scores based on different GPAs (Art, Music, Math, Science, etc.) and state?

Analysis Plan:

- Plot out the data in different ways and make sure to separate every plot out by year to account for changes over time.
- We will use cross validation as we train our linear regression model since we will have multiple parameters to figure out.
- This model will help answer our research question by creating a function whereby we can predict SAT scores (output) with GPA scores (input). If the error is low, then it is possible.

Report Plan:

- We will report our results by walking the reader through our initial analysis and plots of data that reveal the insights we used to approach this research question.
- We will report a single model if the scores are highly correlated and the model, after multiple cross validation iterations, holds up in producing a valid result. If not, we will present other models that show the direction of our research question is plausible even if our model is not.