**Project Overview: Bias-Variance Trade-off & Auto Dataset Analysis**

**Question 1 – Flexibility vs. Inflexibility in Learning Methods**

Analyzed four scenarios to determine when flexible or inflexible statistical learning methods are preferable:

* **(a) Large n, small p**:  
  Flexible methods perform well due to sufficient data to model complex patterns without overfitting.
* **(b) Small n, large p**:  
  Inflexible models generalize better; flexible models tend to overfit due to high dimensionality and limited observations.
* **(c) Highly non-linear relationships**:  
  Flexible models adapt better to non-linearity, reducing underfitting.
* **(d) High error variance**:  
  Inflexible models help avoid fitting noise, improving generalization.

**Question 2 – Bias-Variance Decomposition and Visualization**

**(a) Simulation of Error Components**

Simulated five curves over increasing model flexibility using R:

* Bias² (↓), Variance (↑)
* Training Error (↓)
* Test Error (U-shaped)
* Bayes Error (Constant)

Visualized using ggplot2, capturing key trade-offs in model complexity.

**(b) Explanation of Curve Shapes**

* **Bias²**: High in simple models, drops with flexibility.
* **Variance**: Low in simple models, increases as models overfit.
* **Training Error**: Decreases rapidly with flexibility.
* **Test Error**: Initially falls, then rises due to overfitting.
* **Bayes Error**: Flat, representing irreducible noise.

**Question 3 – Auto Dataset Analysis**

**(a) Predictor Types**

Identified:

* **Quantitative:** mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin
* **Qualitative:** name

**(b) Ranges of Quantitative Predictors**

Used range() to determine min/max for each quantitative feature.

**(c) Central Tendency & Dispersion**

Computed mean and standard deviation for each quantitative predictor using sapply().

**(d) Subset Analysis**

Removed rows 10–85; recomputed range, mean, and SD for comparison. Changes were minor but measurable in some predictors (e.g., mpg, weight).

**(e) Graphical Exploration**

Used ggpairs() to create a scatterplot matrix for quantitative variables. Strong correlations found:

* mpg vs weight, displacement, horsepower: strong negative
* mpg vs year: positive

**(f) Predictive Indicators for mpg**

Scatterplots showed strong predictive relationships between mpg and:

* **Negative:** weight, horsepower, displacement, cylinders
* **Positive:** year, acceleration (moderate)

These findings support their inclusion in regression modeling for fuel efficiency.

**Tools and Environment**

* **Language:** R
* **Libraries:** ggplot2, reshape2, GGally
* **Data:** Cleaned Auto.data (missing values removed)