1.4: Data Exploration

Stat 5100: Dr. Bean

1 Why Data Exploration

Data Modeling is a lot like:

In order to avoid disaster, you need to **look** before you **jump**.

Example: Consider four scenarios where we use to create a model that uses values of x to predict values of y. We make the assumption in each case that the data can be modeled as

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \epsilon_i \tag{1}$$

This assumption means that we assume that X and Y share a linear relationship. That is, as X increases, Y will increase proportionally. We will explore this further in Handout 2.1.

Data Explorations BEFORE modeling will help us to detect:

- Skewed distributions
- Outlier points
- Non-linear trends

Often, we can use **variable transformations** to get data that are normal, or at least symmetric, in distribution.

Common Exploratory Plots

- **Boxplots:**: Show the five quartlies of the data (min, 25th percentile, median, 75th percentile, and maximum).
  - Values that are farther than 1.5*IQR (Interquartile Range, which is the 75th percentile minus the 25 percentile) above the 75th percentile or below the 25th percentile are typically plotted as "outlier" points.
  - Great way to quickly summarize the range of values.