**Exploratory Data Analysis (EDA) (by Xinyi Gui)**

## **1. Introduction**

In this section of the project, I focus on **Exploratory Data Analysis (EDA)** to gain initial insights into the dataset. EDA is a crucial step that helps us understand data distributions, detect anomalies, and formulate hypotheses. Specifically, I use univariate analyses—such as **histograms**, **boxplots**, and **summary statistics**—to explore each variable in isolation.

## **2. Data and Variables**

For demonstration, I used either:

* A **built-in dataset** (e.g., mtcars or iris), or
* A **user-uploaded dataset** via a Shiny app interface.

After loading the data, I identified the **numeric variables** for histogram and boxplot visualization. In this example, let’s assume a numeric variable named **mpg** (miles per gallon) from the mtcars dataset.

## **3. Methods and Workflow**

### **3.1 Data Loading**

1. **User Upload**: A file input allows the user to upload CSV, Excel, JSON, or RDS.
2. **Built-in Datasets**: For quick testing, mtcars or iris can be loaded.

### **3.2 Univariate Analysis**

* **Histogram**: Shows the distribution of a single numeric variable (e.g., mpg).
* **Boxplot**: Highlights the median, quartiles, and potential outliers.
* **Summary**: Provides numeric summaries like mean, median, min, max, and quartiles.

**Explanation**:

1. **UI**:
   * A **selectInput** ("uni\_var") for choosing a numeric variable.
   * **radioButtons** ("uni\_plot\_type") for histogram or boxplot.
   * An **actionButton** ("run\_uni\_eda") to trigger the EDA.
2. **Server**:
   * An **observeEvent** listens for the "run\_uni\_eda" button click.
   * **renderPlot** uses ggplot2 to produce either a histogram or a boxplot, depending on the user’s choice.
   * **renderPrint** calls summary() on the chosen variable to show mean, median, quartiles, etc.

**4. Results and Observations**

1. **Histogram** (bins = 30):
   * Shows that mpg is roughly unimodal, with most values around 15–20.
   * A few higher mpg values (above 30) are present, indicating more fuel-efficient cars.
2. **Boxplot**:
   * Median mpg is around 19.2.
   * Lower quartile is about 15.4, upper quartile about 22.8.
   * Potential outliers appear above 30 mpg.
3. **Summary**:
   * **Min**: 10.4
   * **1st Qu.**: 15.4
   * **Median**: 19.2
   * **Mean**: 20.1
   * **3rd Qu.**: 22.8
   * **Max**: 33.9

This tells us that **mpg** in the dataset spans from about 10 to 34, with a mean of roughly 20.1. Such insights can guide further analysis or modeling decisions (e.g., investigating the characteristics of cars with mpg > 30).

## **5. Key Takeaways**

* **Histogram** and **boxplot** provide quick snapshots of the distribution and outliers.
* **Summary statistics** confirm the central tendency and spread.
* Combining these visuals and numeric summaries gives a clear initial understanding of each variable’s behavior.
* This EDA step is foundational, preparing us for more advanced analyses (like bivariate relationships, correlations, and feature engineering).

**Xinyi Gui ’s Contribution:**

In the **EDA Part**, I implemented univariate analysis to explore individual numeric variables. This approach reveals data distribution, outlier presence, and summary statistics crucial for guiding deeper analyses. The methods described (histogram, boxplot, summary) serve as a **baseline** for more sophisticated EDA techniques, which we will build upon in subsequent parts of the project.