Learning Goals

- Understand the role of Exploratory Data Analysis (EDA) in the data science life cycle and identify when and why it is necessary.
- Investigate dataset documentation and structure using functions like `help()` and `str()` to determine what the data is and how it is organized.
- Detect and handle special values (e.g., NA, NaN, Inf) and assess data conditions using logical checks and summary functions.
- Prepare data by modifying variable types, filtering missing values, or creating new variables through transformation and recoding.
- Generate initial summaries using the summary() function and visualize quantitative and categorical variables to gain insight into distributions and patterns in the data.

Key Functions

For each of the following functions below write down a brief definition of what it does and a basic example

- **is.na():**
- **unique():**
- **sort():**
- **complete.cases():**
- **as.numeric():**
- **cut():**
- **summary():**
- **table():**
- **hist():**

Key Concepts

1. Why is it important to compare dataset documentation with the actual data?

2. How do special values (NA, NaN, Inf) affect analysis?

3. When would you use a histogram versus a barplot?

4. How might we identify unusual and missing values that are not coded as NA?

Practice Problems

- What line of code can we use to install and load the "MSMU" library?

- Given a vector called "grades", how could you determine if any (and how many) values are missing?

- Given a vector called "grades", how could we create another vector called "grade_categories" (A: 90-100, B: 80-89, C: 70-79, D: 60-69, F: 0-59)?