
L8: Exploratory Data Analysis Lab

Complete the following exercises in R. Work in your small group to solve them. For each question, first write the code on the paper, and once you have completed the whole sheet, you can use R to evaluate yourself and see if the code worked as expected. Your end result should be a knitted RMarkdown file, with Word preferred as the output format.

Use the heart dataset in the MSMU library for this lab.

1. Getting to Know the Data

Before doing any analysis, we want to understand what the dataset looks like.

- (a) What code can we use to load the **MSMU** library?
- (b) What code can we use to view the documentation for the **heart** dataset?
- (c) What code can we use to compare the logical structure from the documentation with the physical structure of the dataset in R?
- (d) How many observations and variables are in the dataset?

2. Checking Data Conditions

Next, we want to check whether the dataset has any missing or unusual values.

- (a) What code can we use to determine if any missing values are present in the entire dataset?
- (b) What code can we use to determine how many missing values are present in each column?
- (c) If there is a missing value in the variable **Ca**, how could we calculate the mean of **Ca** while ignoring the missing value?
- (d) Use **sort()** and **unique()** on one quantitative variable. Do you notice anything unusual?

3. Preparing the Data

Some columns may need to be changed or cleaned before they are useful for analysis.

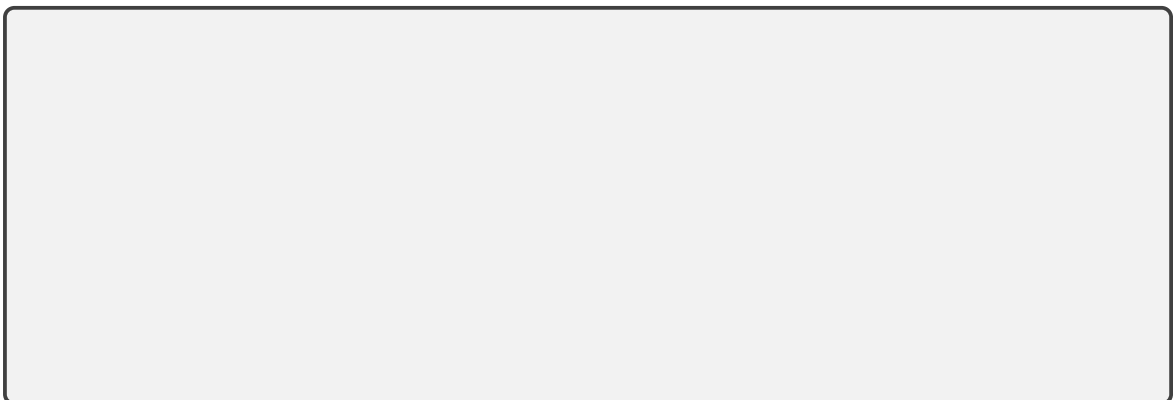
- (a) How could we create a new factor version of the **Sex** variable with the correct level names?
- (b) You find out that any value of **Chol** above 400 should be treated as missing. Create a new column called **Chol_clean** where values above 400 are changed to **NA**.
- (c) Create a new variable using the `cut()` function called **MaxHR_group** that breaks **MaxHR** into the following groups:
 - **low** if less than 120
 - **below average** if 120 to less than 150
 - **above average** if 150 to less than 175
 - **high** if 175 or above



4. Initial Summaries

Now that we have inspected and prepared the data, we want to summarize it.

- (a) Use `summary()` on the full dataset.
- (b) Use `summary()` on the cleaned cholesterol variable.
- (c) Use `table()` to count how many people are in each **MaxHR_group**.
- (d) In 2–3 sentences, explain one thing you learned from these summaries.



5. Creating Basic Visualizations

Create visualizations that help us understand the data.

- (a) Create a histogram of the **MaxHR** variable.
- (b) Create a histogram of the **MaxHR** variable with a density line overlaid on it.
- (c) Create a barplot showing the number of people in each **MaxHR_group**.
- (d) In 2–3 sentences, explain what one of your visualizations helps us see.

6. Final EDA Reflection

Write a short paragraph answering the following questions:

- What is one thing you learned about the **heart** dataset?
- What is one issue or limitation you noticed in the data?
- Why is it important to inspect and prepare data before trying to draw conclusions?

7. Make sure your final turned-in work is a knitted RMarkdown document. Turn in both the **.Rmd** file and the **.docx** file to gain full points.