A logo of a university of windsor

Description automatically generated

**University of Windsor**

Lab 1

COMP 8157 | Advanced Database Topics – Summer 2025

Names:

Aleena Ali Azeem - 110190830

Simranpreet Kaur - 110189426

Priyadharshan Reddy S - 110191285

**Part 1: Data Exploration (12 marks)**

1. Import the **Heart Attack** datasets **(2 marks)**.

A close up of a text

AI-generated content may be incorrect.

This line loads the dataset so we can start analyzing it in R

1. Summarize that Heart Attack dataset and explain the output **(4 marks).**

A close up of a logo

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

This provides a quick **statistical summary** of each column in the data frame df.

For each column in df, it gives:

* **Numerical columns**:

Minimum (Min.)

1st Quartile (1st Qu.)

Median

Mean

3rd Quartile (3rd Qu.)

Maximum (Max.)

* **Categorical columns (factors or characters)**:

A count of the number of observations in each category/level.

* This tells R to display a **summary report** of the dataset so you can quickly understand the distribution and type of data in each column.

3.Show the structure and dimension of the dataset and explain it **(2 marks)**.

A screenshot of a computer code

AI-generated content may be incorrect.



**str(df)**

* The str() function gives you a **snapshot of your dataset**.
* It tells you:

What kind of object it is (usually a data.frame)

How many **rows (data points)** and **columns (variables)** it has

The **names of the columns**

The **type of data** in each column (like numbers, text, or categories)

A **sneak peek** at a few actual values in each column

**dim(df)**

* The dim() function shows you **how big the dataset is**.
* It returns two numbers:
  1. The number of **rows** - how many data entries you have
  2. The number of **columns** - how many features or variables are recorded for each entry

4.Show the first 8 rows and the last 5 rows of the dataset **(2 marks)**.

A screenshot of a computer screen

AI-generated content may be incorrect.

* head() displays the **first few rows** of the dataset.
* By default, it shows the first **6** rows, but here we've specified 8, so it shows the **first 8 rows** of df.
* tail() shows the **last few rows** of the dataset.
* By default, it also shows 6 rows, but here we're asking for **5 rows** - so it displays the **last 5 rows** of df.

5. Show the column names of the Heart Attack dataset **(2 marks)**.   


colnames() stands for **column names**.

It returns a **character vector** containing the list of **names of all the columns** in the data frame df.

**Part 2: Data Pre-Processing (28 marks)**

6.What is the class variable in the Heart Attack dataset? What does it indicate **(4 marks)**.?   
  
It tells the output based on variables such as cholesterol etc.   
The output column is used as the class variable. It contains values like 0 and 1, where:

0 means the person did not have a heart attack

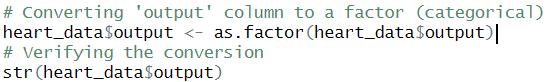
1 means the person had a heart attack



7. What is the datatype of the class variable **(4 marks)**?   


This code checks if the class variable is stored as a number or as a category. If it is numeric (like integer or double).  


8. Change the class type of the class variable of Heart Attack dataset to factor. Show the output after the conversion **(4 marks)**.

So, the factor is applied to classify the levels of the column and how many each level is repeated.   


The output column is changed to a factor. This means it will be treated as a category, not a number. This is important when doing classification tasks, as it clearly separates the two groups: heart attack and no heart attack.



9. Find the sum of the missing values in Heart Attack dataset **(4 marks)**.

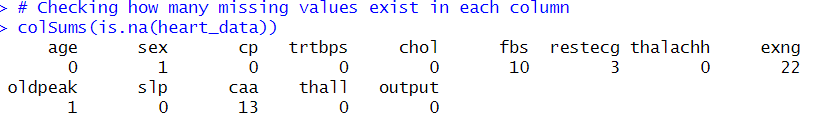
  
 This line checks how many cells in the dataset are empty or have missing values. It's important to find missing data before starting any analysis or building models.



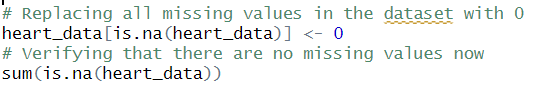
10. Find which columns contain missing values in the dataset. What is the total missing values for each column **(4 marks)**?



**Explanation:**  
This code shows how many missing values are in each column. It helps to know which specific variables need fixing or attention before moving forward with data analysis.



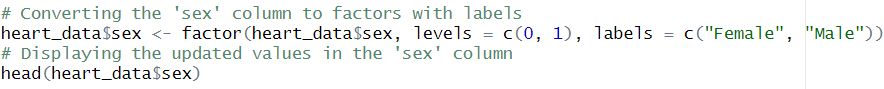
11. Replace the missing values in the Heart Attack by 0. Check what if the missing values was replaced successfully **(4 marks)**.

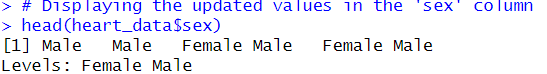


All missing values in the dataset are replaced with 0. After this, the code checks again to make sure there are no more missing values. This step helps clean the data so that it won’t cause errors during analysis.



12. Rename the sex attribute from (0 and 1) to (Male and Female). Show the conversion output of the specific attribute **(4 marks)**.

  
 In the original dataset, the sex column used numbers (1 for Male, 0 for Female). These numbers are changed to text labels ("Male" and "Female") to make the data easier to read and understand when doing analysis or creating charts.



Part 3: Clustering (14 marks)

13. Apply clustering techniques. Explain how many clusters there are and why (7 marks).

First, we use the numerical data and then we use the scale the data. Then we apply the elbow method.

A screenshot of a computer program

AI-generated content may be incorrect.

A graph with numbers and lines

AI-generated content may be incorrect.

After which we will make clusters. Using the elbow method, we use K = 2. As the decrease becomes less after that.

A screenshot of a computer code

AI-generated content may be incorrect.

A diagram of a cluster of numbers

AI-generated content may be incorrect.

14. Which two variables have strong relationships with each other and why (7 marks)

A screenshot of a computer screen

AI-generated content may be incorrect.

Above numerical representation of the variables in our data and their correlation. There are negative and positive correlations.

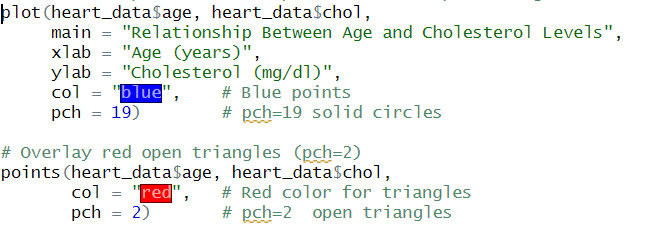
The two variables with the strongest correlation in the Heart Attack dataset are oldpeak and slp, showing a strong negative correlation of -0.58. This indicates that as ST depression (oldpeak) increases, the slope of the peak exercise ST segment (slp) tends to decrease. Since both features relate to the heart's response during exercise, this relationship is medically meaningful and highlights how changes in one measurement often reflect changes in the other.

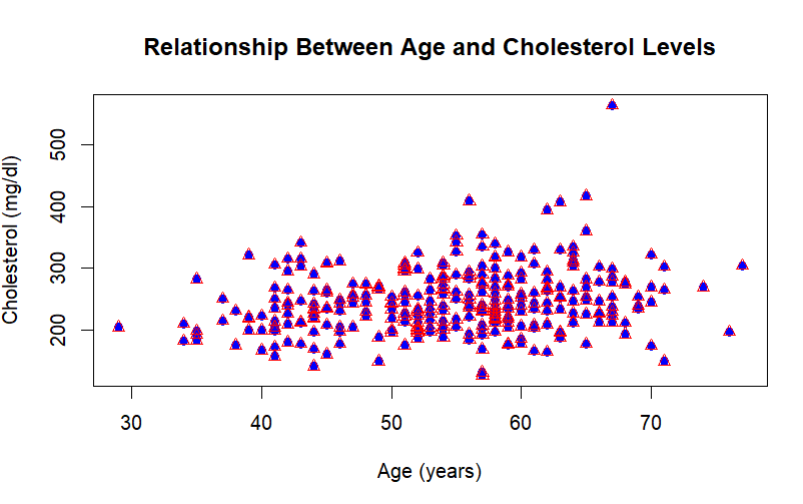
Part 4: Data Visualization (60 marks)

15. Create a scatter plot. The plot should show the relationship between the cholesterol and the age attributes (10 marks).

a. Add labels, title, and color to the plot. The color should be blue.

b. Add open red triangles to the plot.



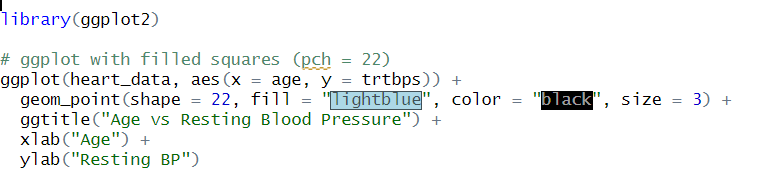


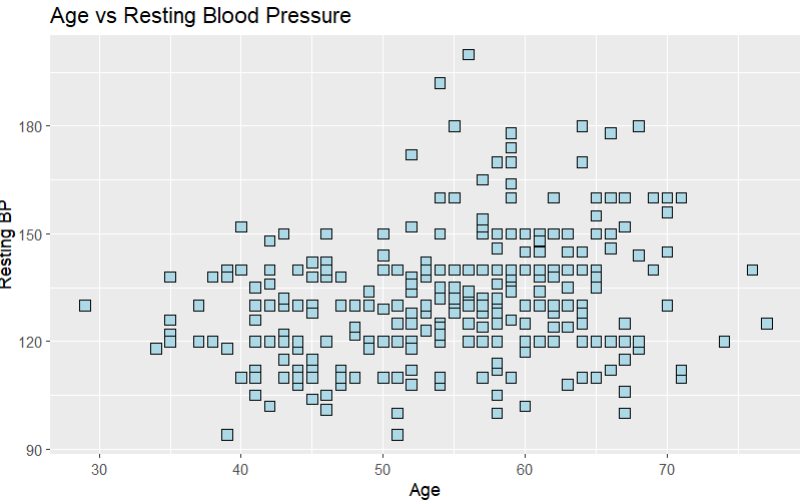
This code creates a scatter plot showing the relationship between age and cholesterol levels. The first layer plots solid blue circles, and the second layer overlays open red triangles at the same points. Labels and titles are added to describe the axes

16. Use the ggplot function to plot any two variables (10 marks).

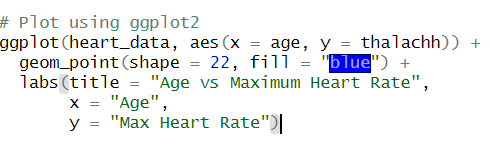
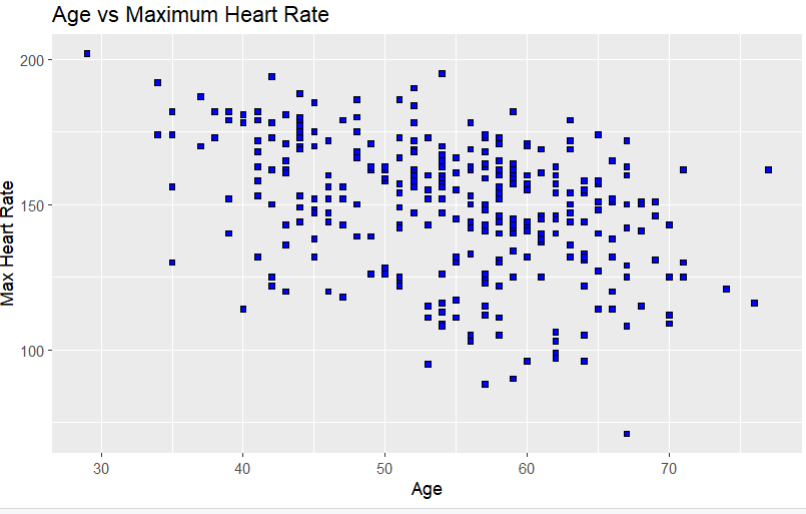
a. The points shape should be filled square.

Example 1:



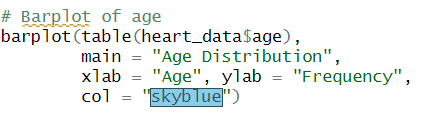


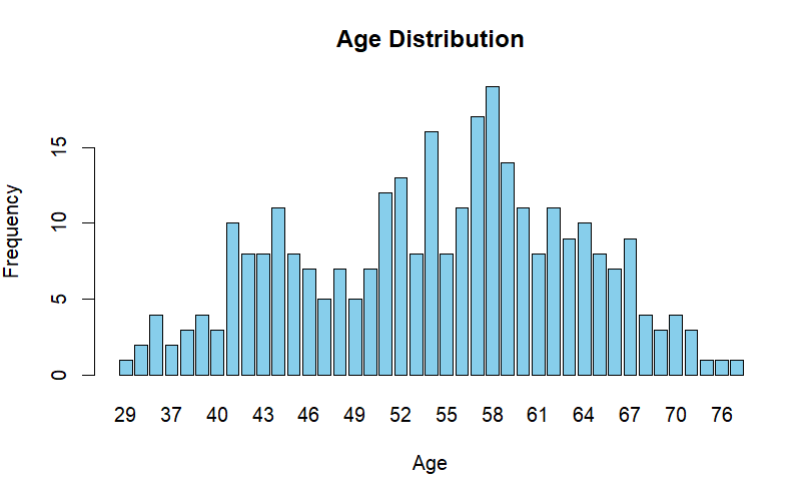
This chart shows how age relates to Resting Blood Pressure achieved . The square filled points help distinguish this plot visually

Question 16: Example: 2   
  
  
  


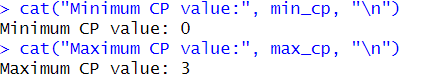
This chart shows how age relates to maximum heart rate achieved (thalachh). The square filled points help distinguish this plot visually.

17. barplot the ‘age’ variable of the Heart Attack dataset (10 marks):   
  
a. Add labels, title, and color to the plot.

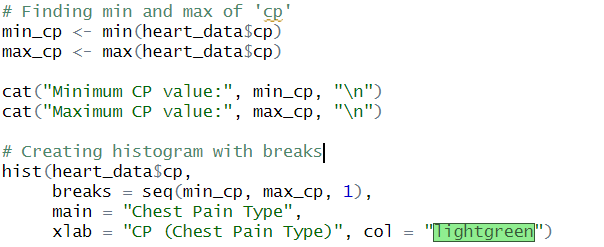


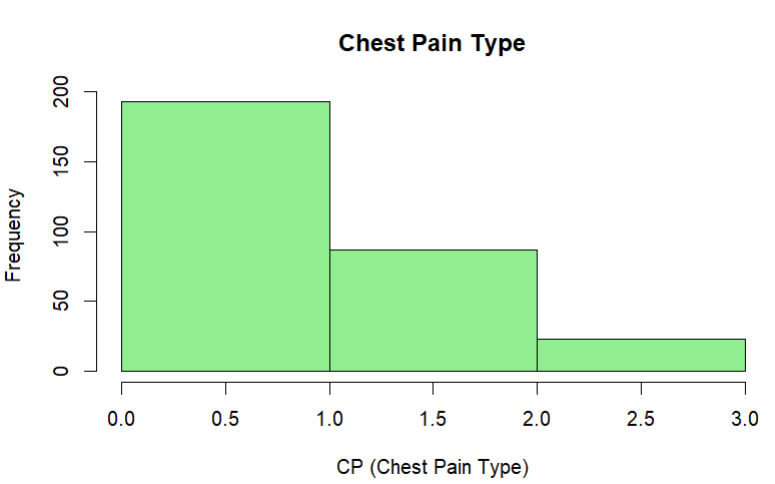


This shows the age distribution in our data. The highest frequency is of people aged between 58 to 61.  
  
18. Create a histogram of the ‘cp’ attribute (10 marks):

1. Find the minimum and maximum of the attribute.  
     
   
2. Add a break function and use the seq(x, y, z) function.

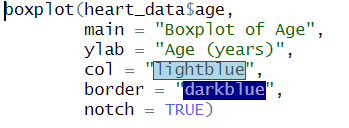
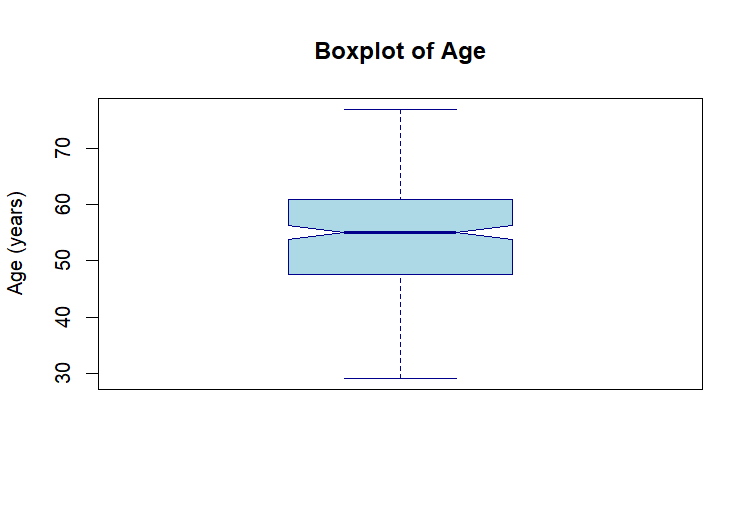
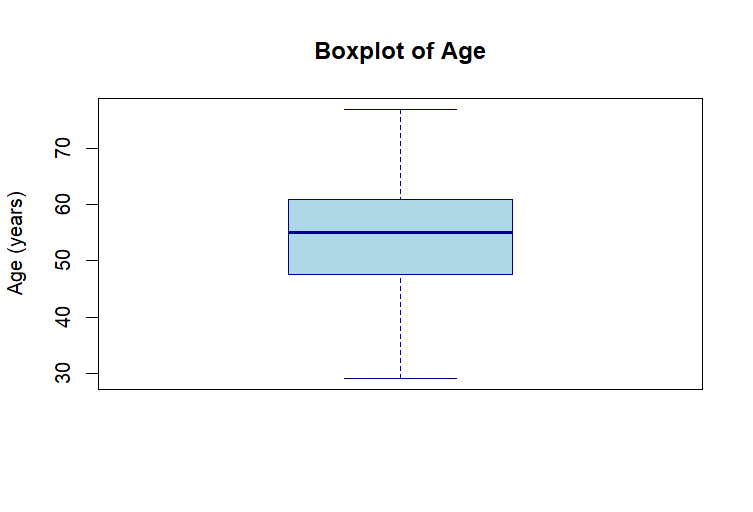
c. Add labels, title = (Chest Pain type), and color to the plot.





It controls how the histogram bins the data. Chest pain is usually **categorical (0, 1, 2, 3)** so seq(min, max, 1) ensures each type gets its own bin. The histogram of the cp (chest pain type) attribute reveals how chest pain categories are distributed among patients in the dataset. Each bar represents one of the four chest pain types: typical angina (0), atypical angina (1), non-anginal pain (2), and asymptomatic (3). From the plot, we can observe which chest pain type is most prevalent.

19. Boxplot the ‘age’ attribute and explain the output (10 marks).

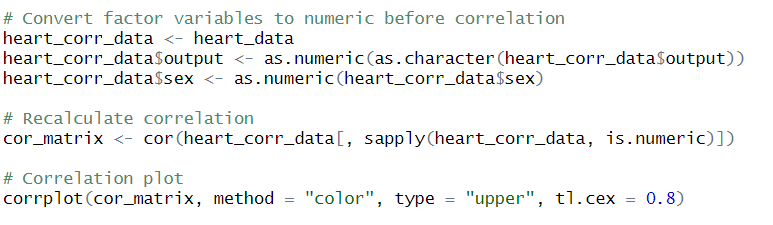
  
  
Output with Notch:  
  
  
  
Output without Notch:  
  


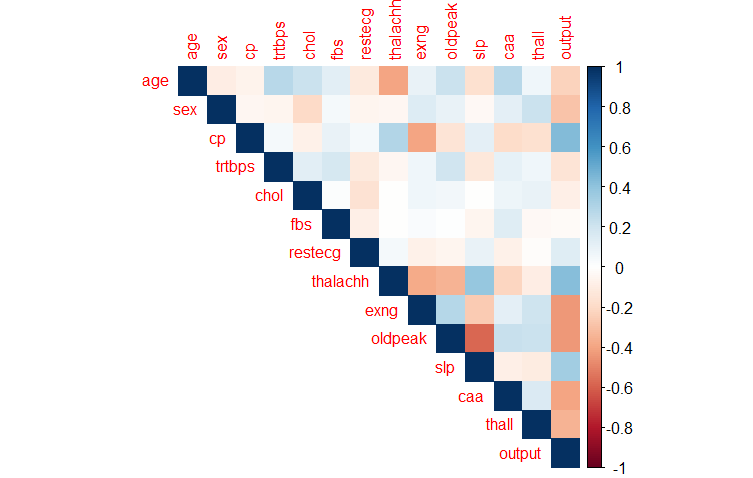
**Explanation:**

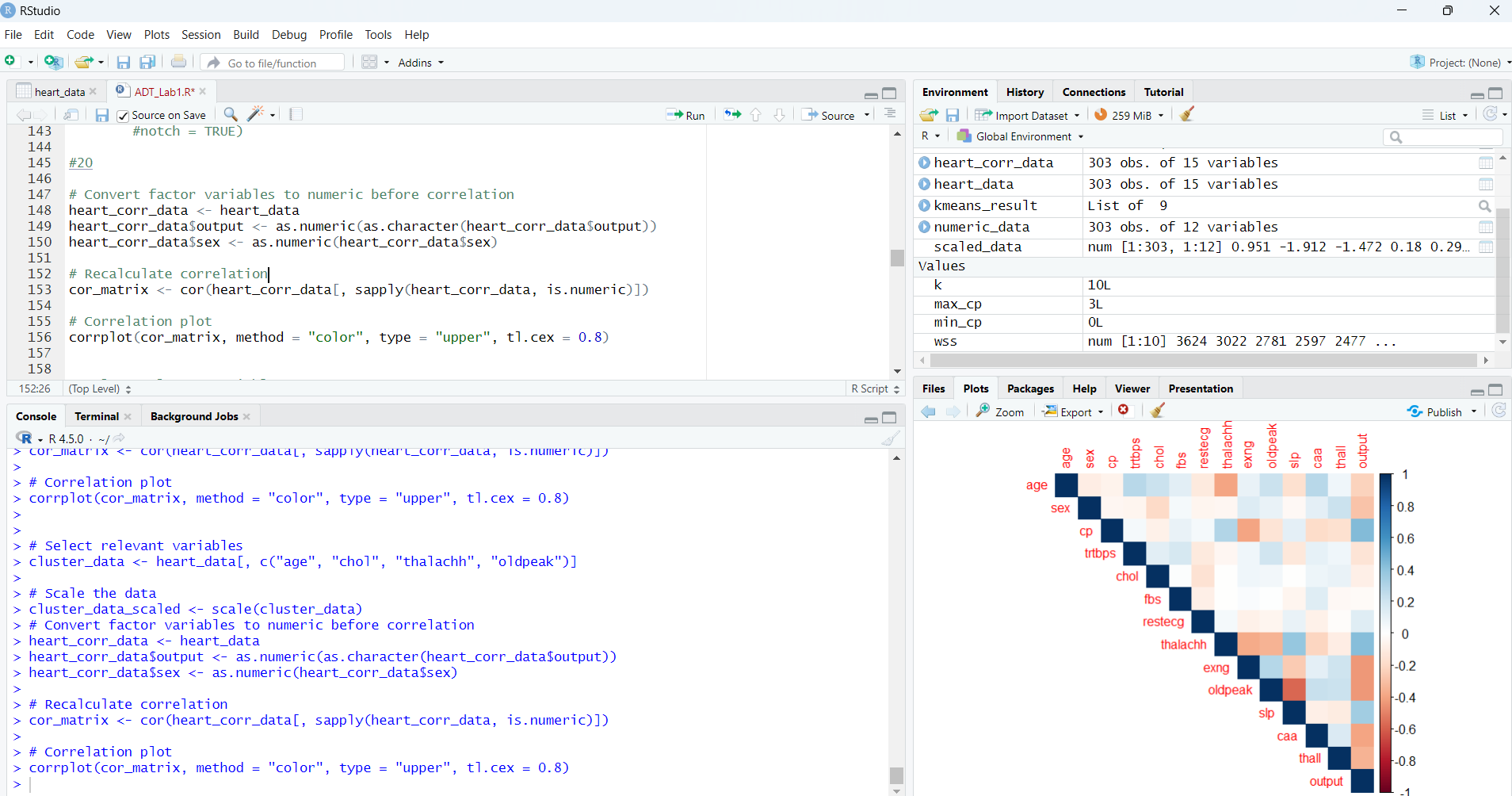
This boxplot shows how patient ages are spread out.

* The **box** represents the middle 50% of ages.
* The **line** inside shows the **median age**.
* **Whiskers** extend to typical min and max values.
* **Dots outside** are **outliers -** patients with unusually high or low ages.

It gives a quick view of age distribution and highlights any unusual values.

20. Create a correlation plot of the whole dataset variables and explain the output. Do not forget to convert some of the variable’s datatype (10 marks).  


Output Image:  
  


  
**Explanation:**

**This correlation plot makes it easy to spot patterns ,one** can quickly see which variables tend to increase or decrease together (**positive or negative correlation**). It’s a powerful visual tool for identifying **strong relationships**, uncovering **hidden trends**, and deciding which variables might be important for **predictive modeling or deeper analysis**.

A document with text and images

AI-generated content may be incorrect.