

Midterm project

# Introduction to Data Science[C]

**Submitted By: Group 8**

**Members Name:**

1. **MD. RAIYAN AL SULTAN**

**Id: 19-41639-3**

1. **MEDHA CHOWDHURY**

**Id: 20-41930-1**

**Section: C**

**Submitted to: TOHEDUL ISLAM**

**Data Set name: Heart Attack Analysis & Prediction Dataset**

# About the Data Set:

In this modified data set we have 12 attributes or columns and 150 instances or rows. The names of attributes are Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS,

RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST\_Slope, HeartDisease.

This data set will give us a proper overview of heart condition and prediction of heart

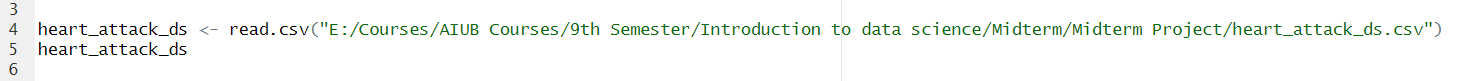
disease of a person of different age and gender.

GitHub link of our project: <https://github.com/NabilRaiyan/Heart-attack-dataset-analysis->

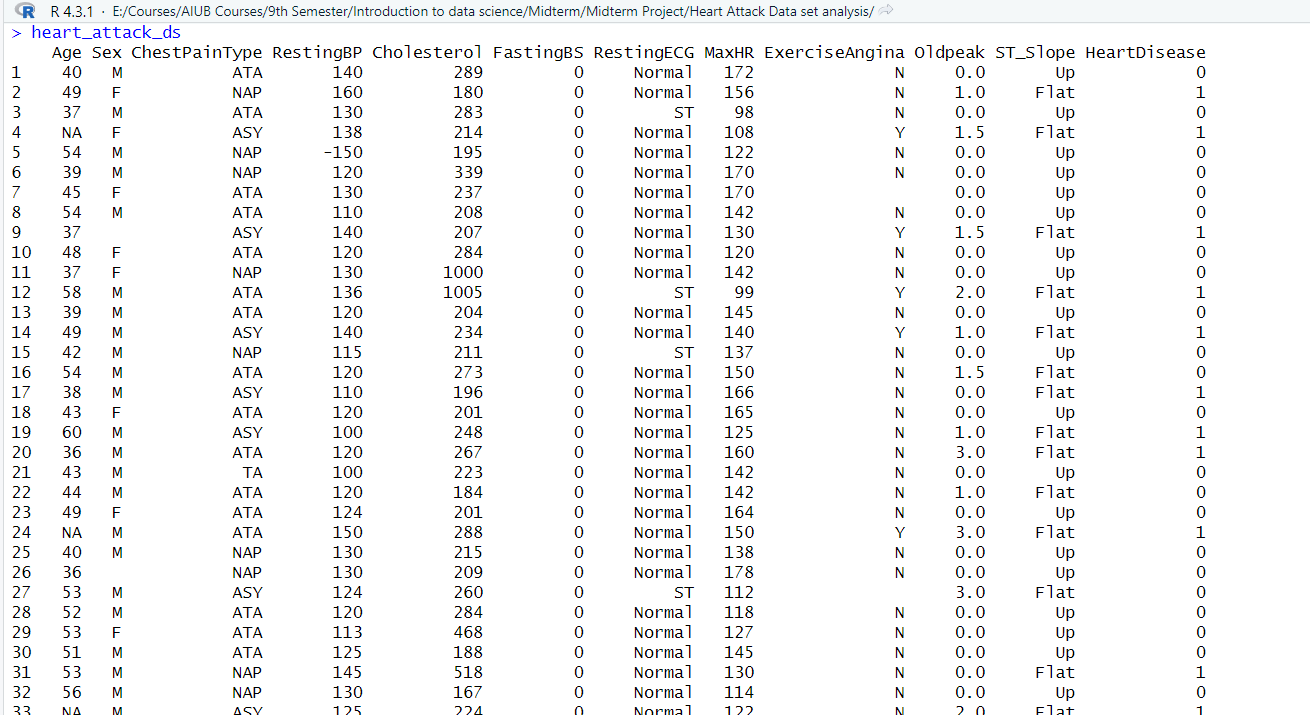
Importing the data set: Importing csv format of the data set in R studio.

Code: In this code we import our modified heart attack analysis data set which is a csv

file.



Output:

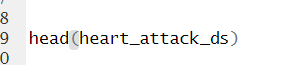


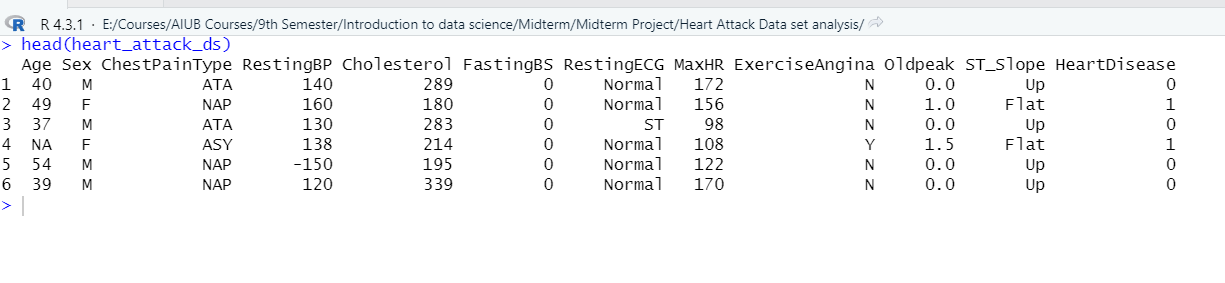
Inspecting the data set: Using head(), tail(), summary(), str(), dim() functions to inspect

the data set.

First few rows:

Code: In this code we inspect first few rows of our data set using head() function

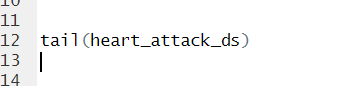


Output:

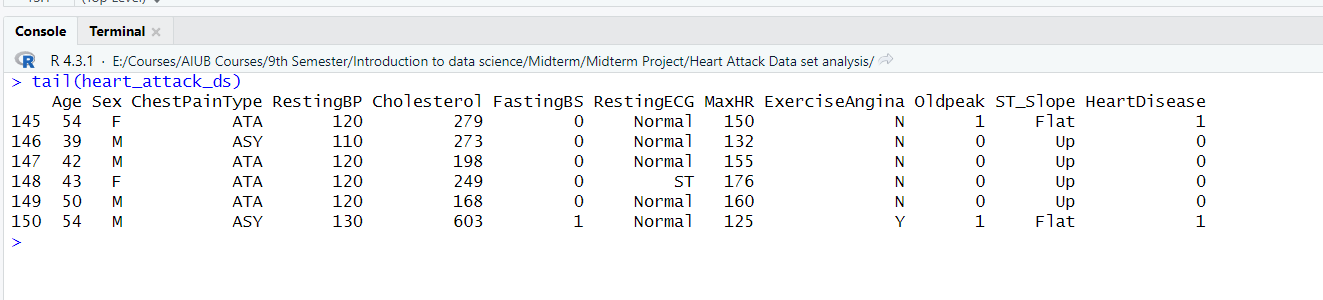
Last few rows: In this code we inspect last few rows of our data set using tail()

function

Code:

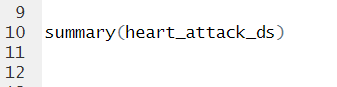


Output:

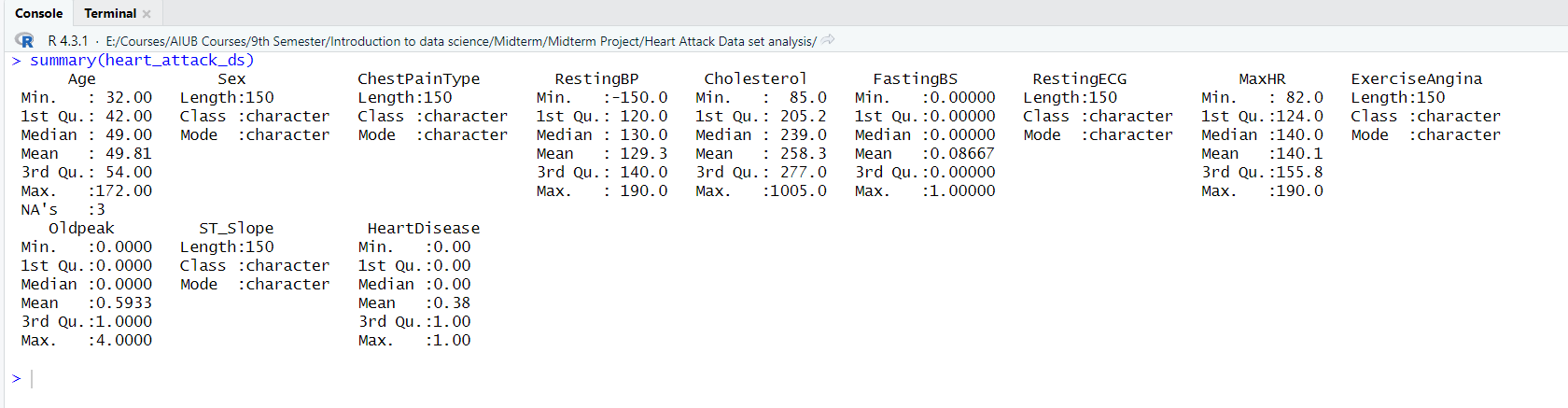


Summary of the whole data set:

Code: summary() function will give us the total summary of out data set

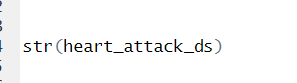


Output:

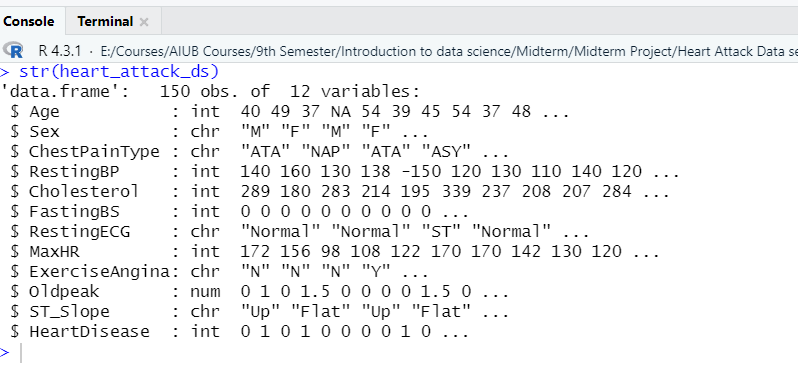


Inspecting data types and structure of the data set:

Code:

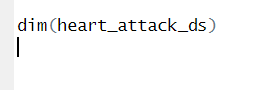


Output:

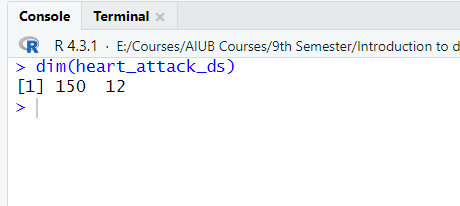


Inspecting the dimension of the data set:

Code: The output will show us the number of columns and rows in our data set

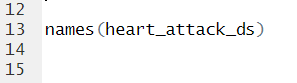


Output:

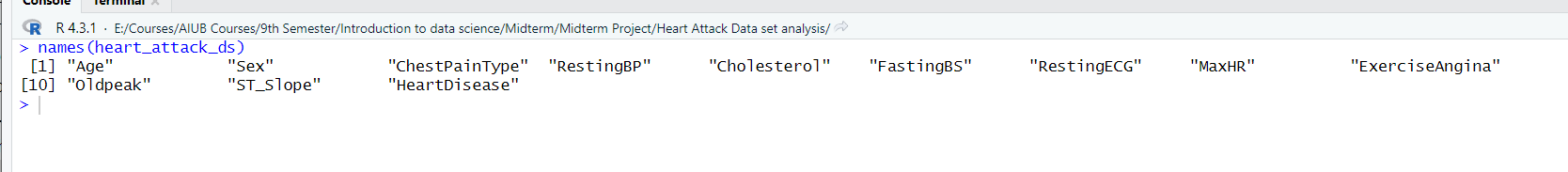


Inspecting the attributes names of the data set:

Code:

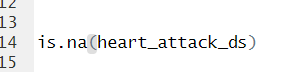


Output:

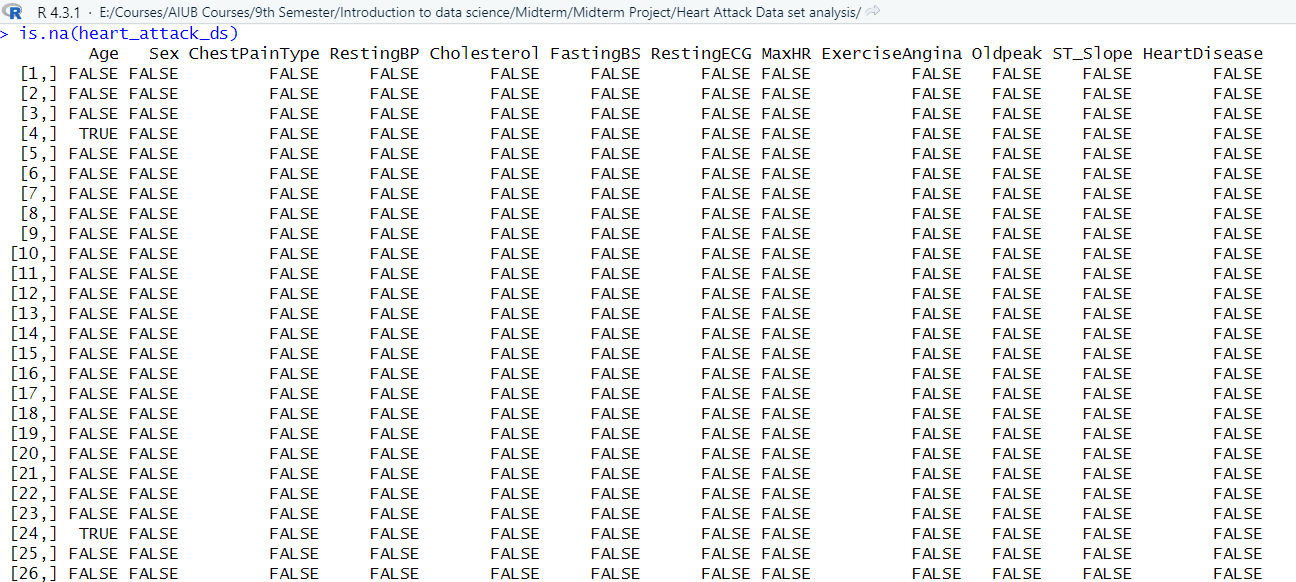


Quick view to check if there is any null value in the data set:

Code:



Output:

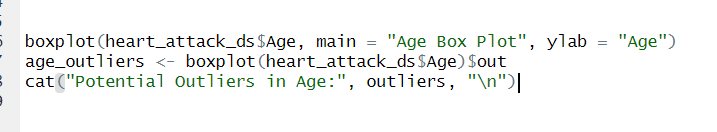


Recovering Missing values from Data Set:

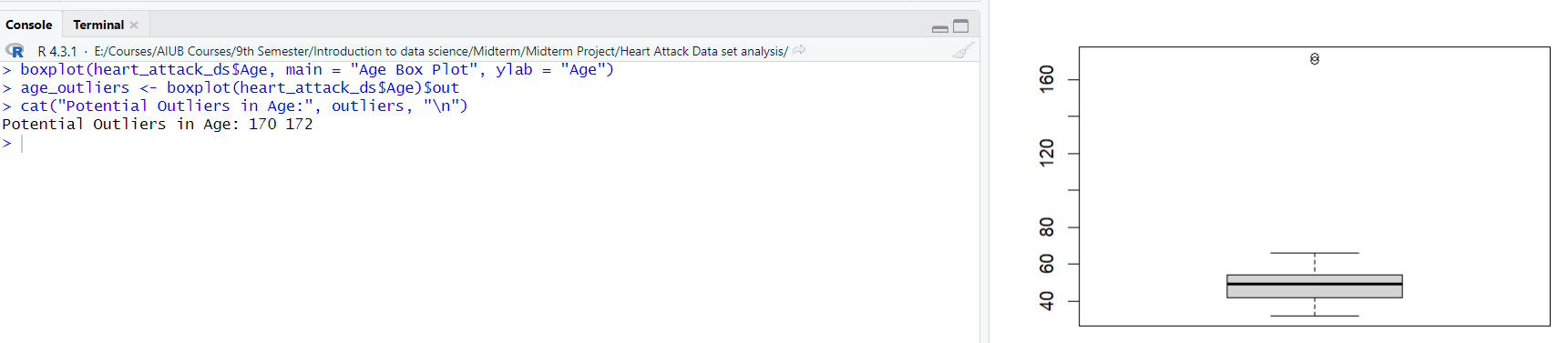
**Age Column:**

Finding Outliers and recovering missing values in Age Column:

Code: In this code we will plot a boxplot to see if there is any outliers in Age column and print the outliers value

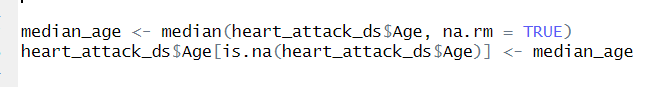


Output:



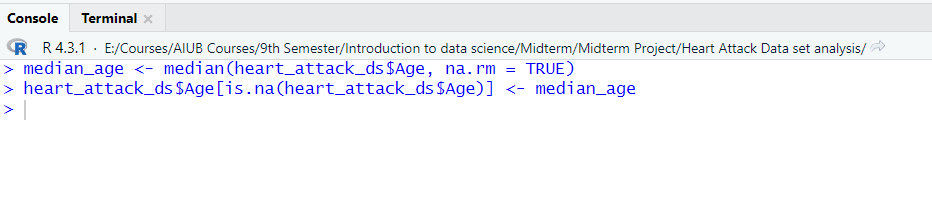
Recovering missing value of Age column by median:

Code:



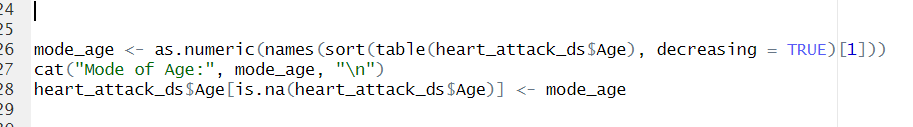
Output:

Code:

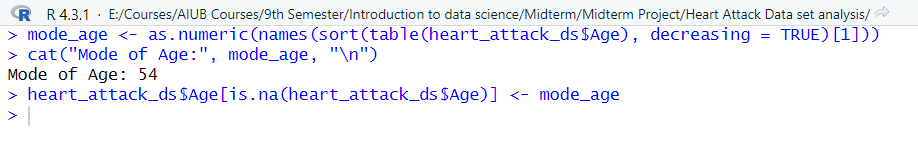


Recovering missing value of Age column by mode:

Code:

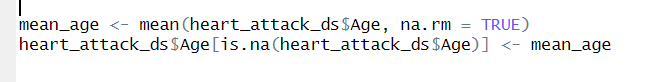


Output:

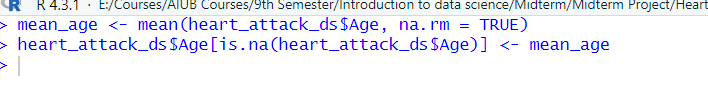


Recovering missing value of Age column by mean:

Code:

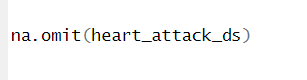


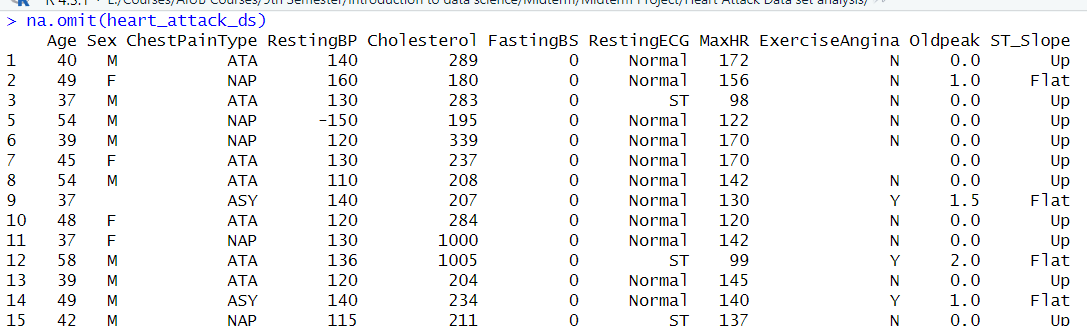
Output:



Removing missing values row from the Age column:

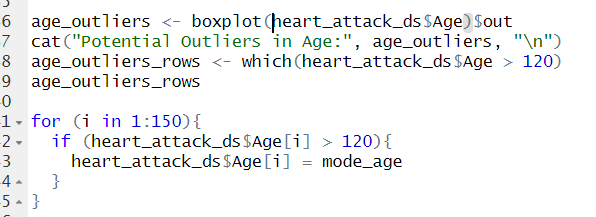
Code:



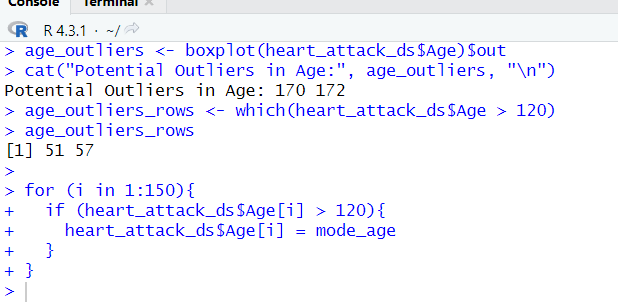
Output:

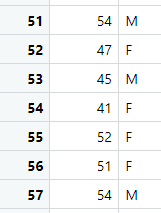
Replacing outliers of the Age column using mode value:

Code:



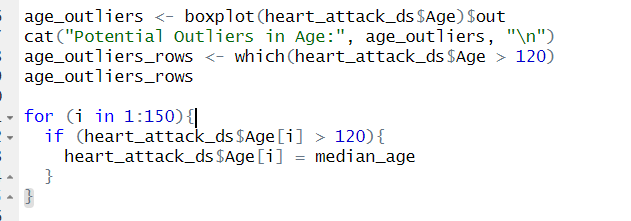
Output:



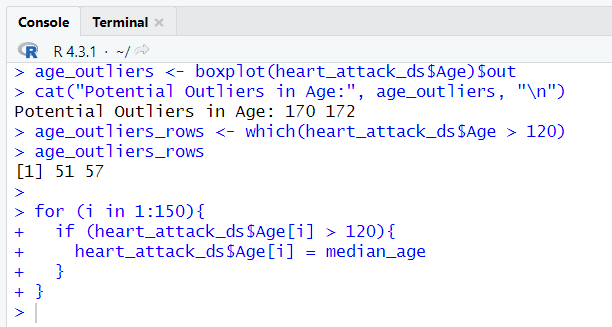


Replacing outliers of the Age column using median value:

Code:



Output:

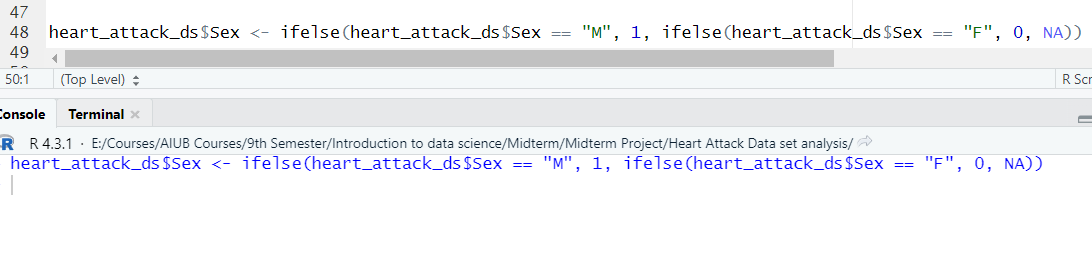


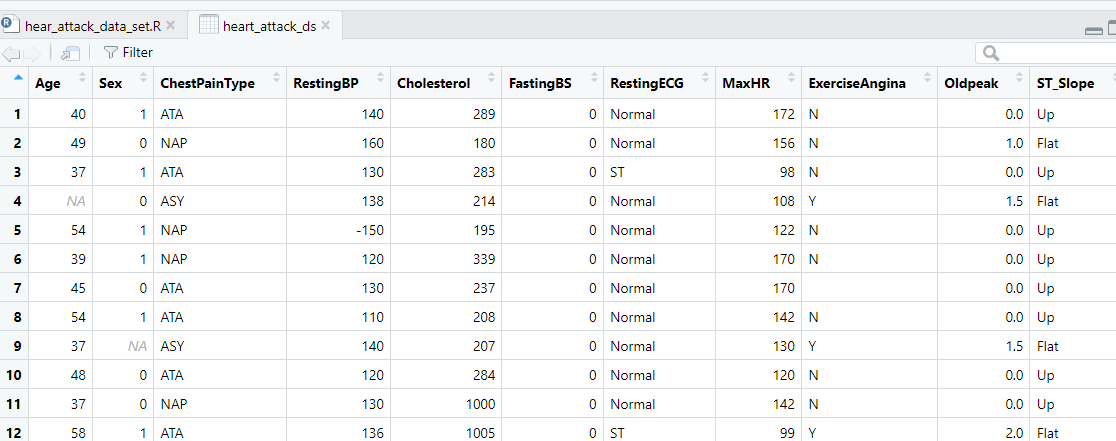
**Sex Column:**

Transforming Sex column, finding and recovering missing values:

Code: In this code we transform the Sex column value from char to int to explore that

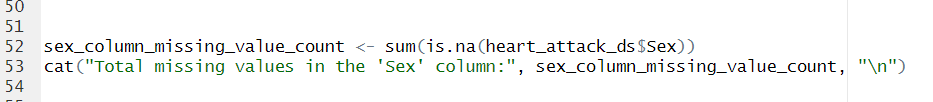
column. So, if there is M then 1, if there is F then 0. (M = Male, F = Female).

****

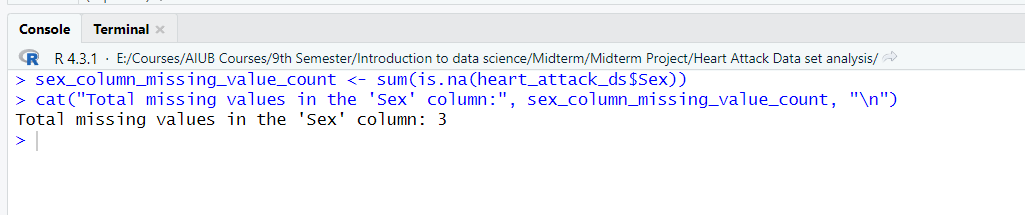
Output:

Finding total missing values in Sex column:

Code:

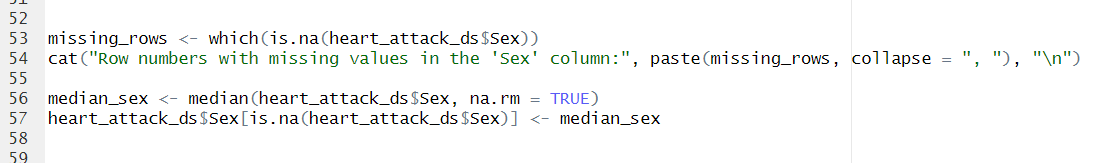


Output:

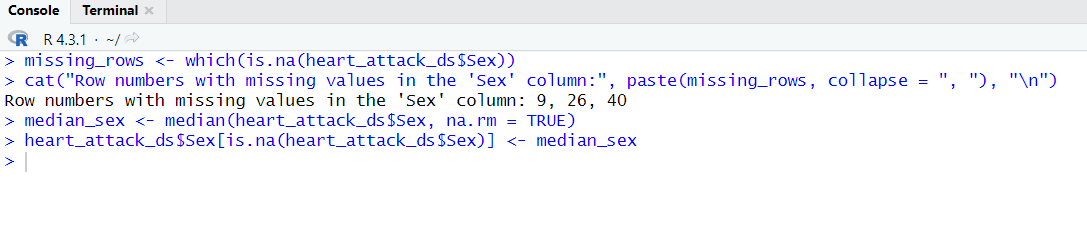


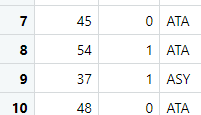
Recovering missing values in Sex column using median:

Code:



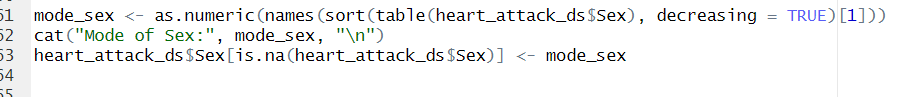
Output:



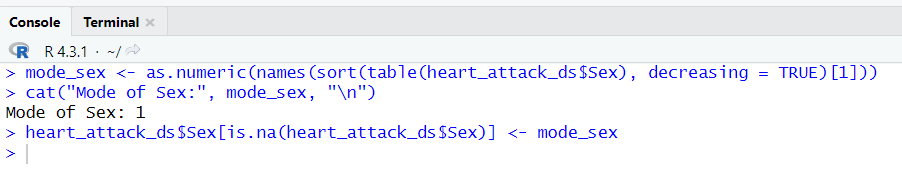


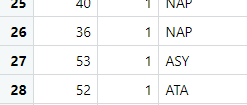
Recovering missing values in Sex column using mode:

Code:



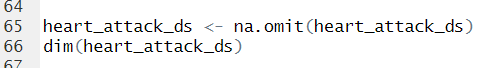
Output:



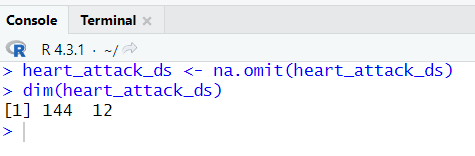


Removing missing value’s rows:

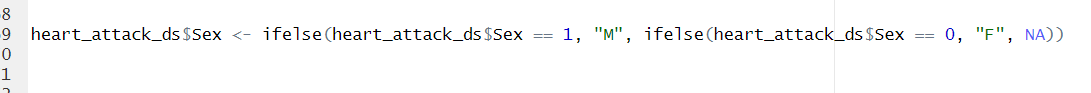
Code:



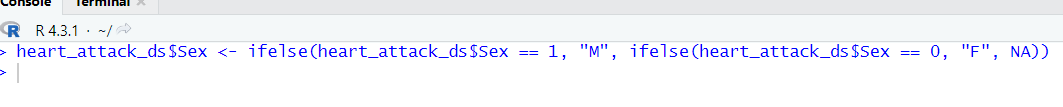
Output:

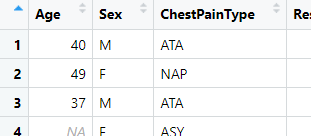


Transforming Sex column to its original value after recovering missing values:

Code: If 1 than Male (“M”) and if 0 than Female (“F”)

Output:

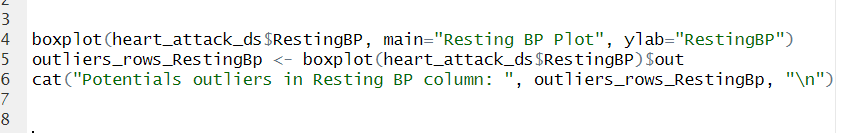




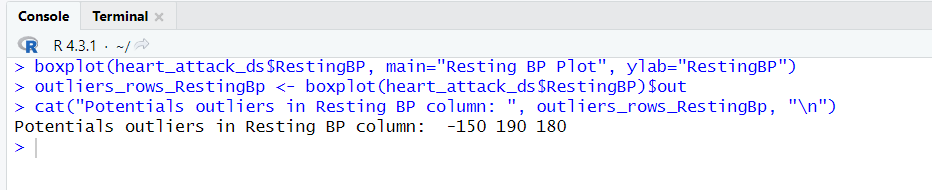
**Resting BP Column:**

Finding outliers in Resting BP column:

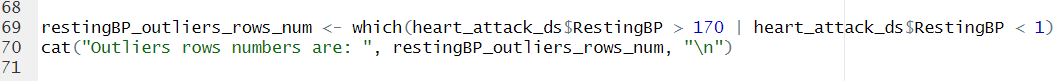
Code:



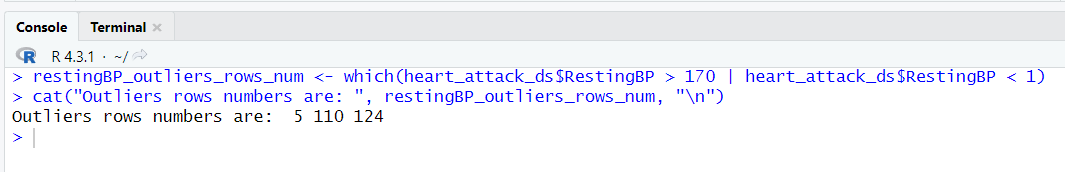
Output:



Finding outliers rows numbers:

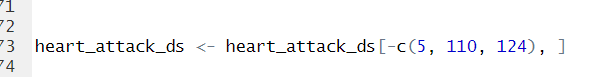
Code: In this code we find the rows number of missing values using which() function

Output:

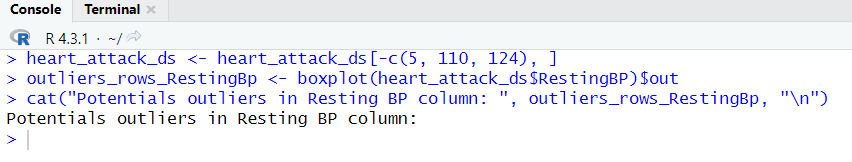


Removing outliers’ rows from the data set:

Code:

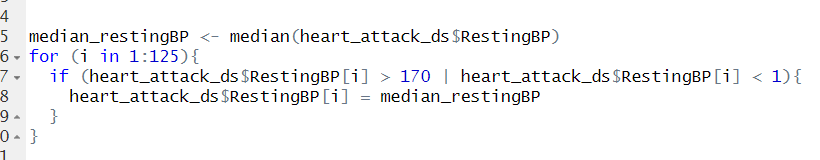


Output:

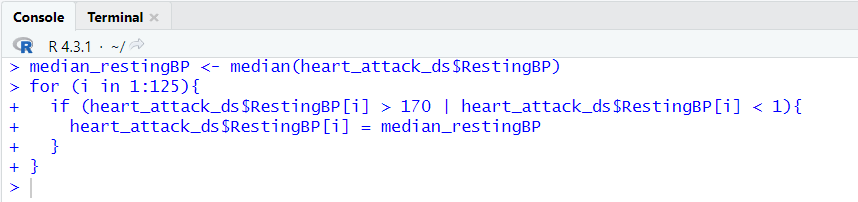


Finding median value of Resting BP column and replacing outliers with median value:

Code:



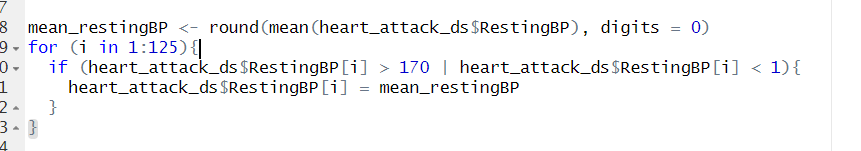
Output:



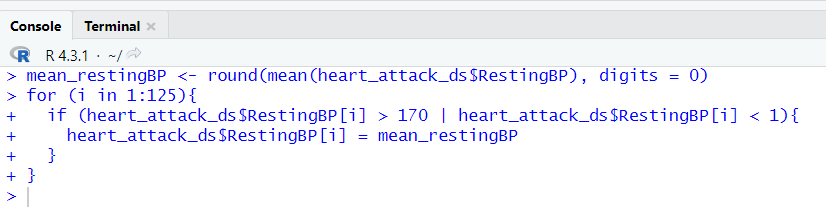


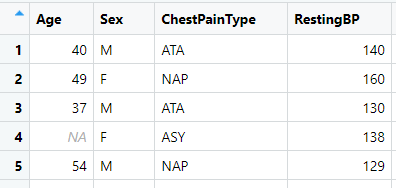
Finding mean value and replace outliers value of Resting BP column with it:

Code:



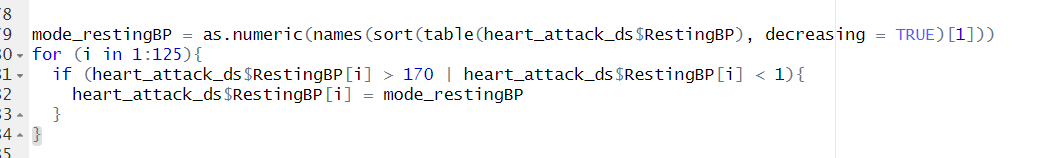
Output:



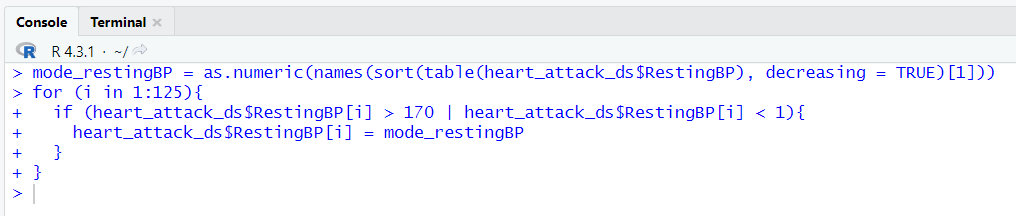


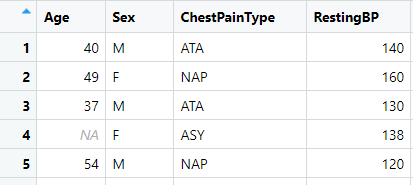
Finding mode value and replace outliers value of Resting BP column with it:

Code:



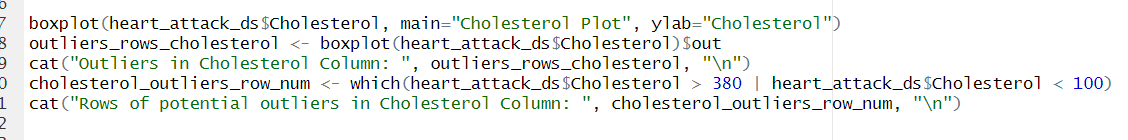
Output:

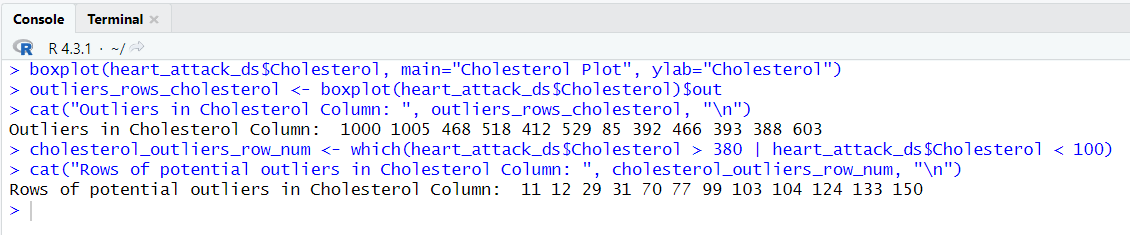


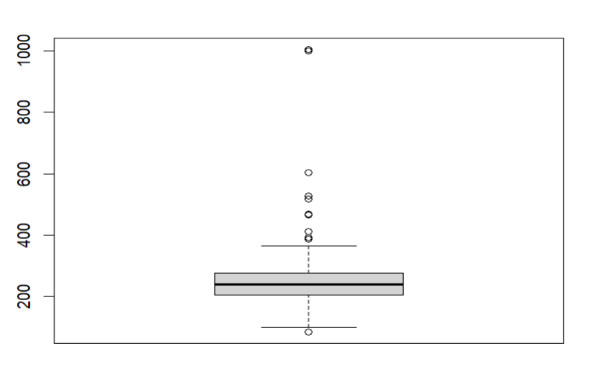


**Cholesterol Column:**

Finding outliers in Cholesterol Column:

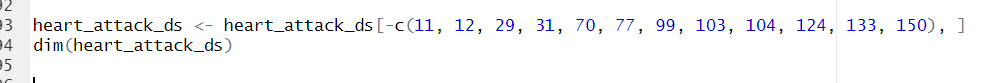
Code:

Output:

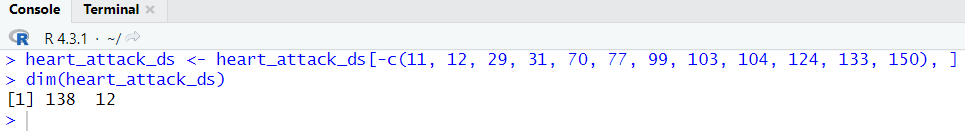


Removing outliers’ rows from the data set:

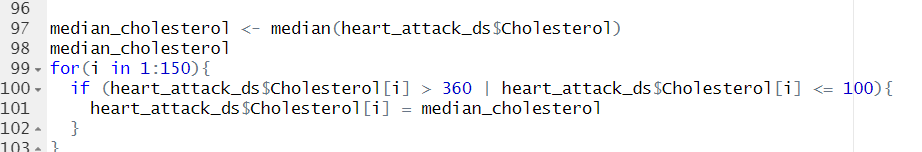
Code:

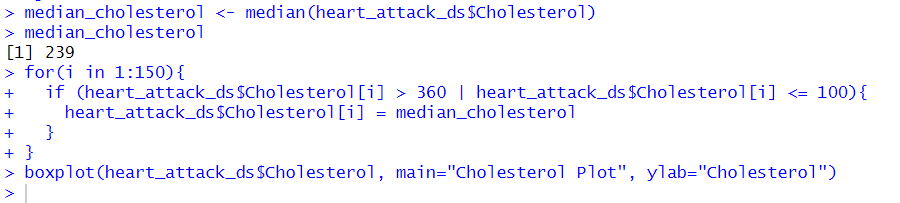


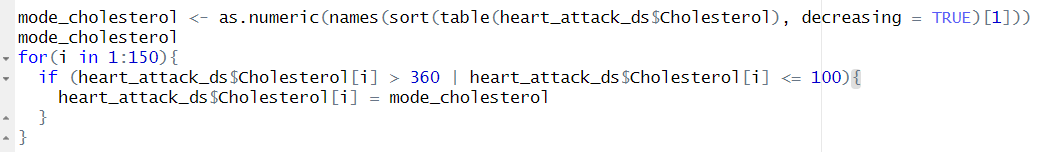
Output:

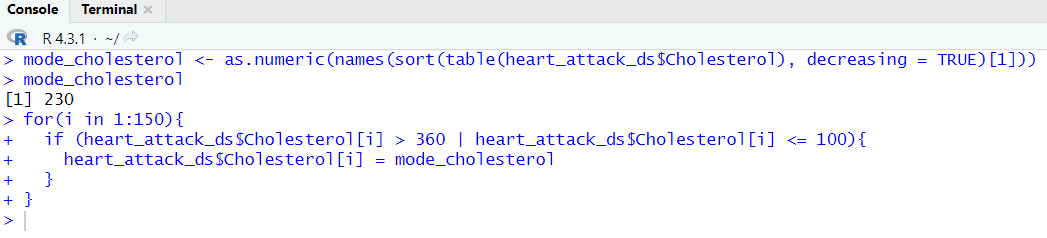


Replacing outliers’ value with median, mode value:

Code for median:

Output:

Code for mode:

Output:

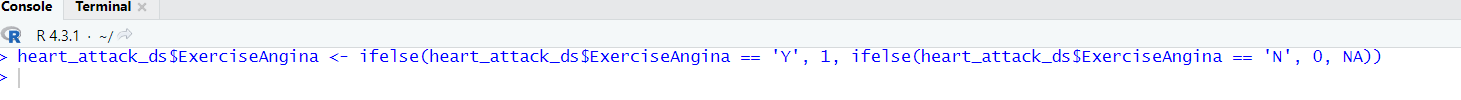
**Exercise Angina Column:**

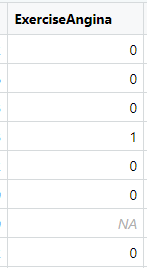
Transforming Exercise Angina column with 0 and 1:

Code: If ‘Y’ then 1 if ‘N’ then 0

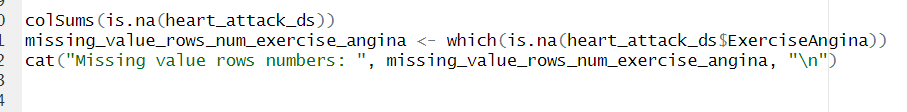


Output:

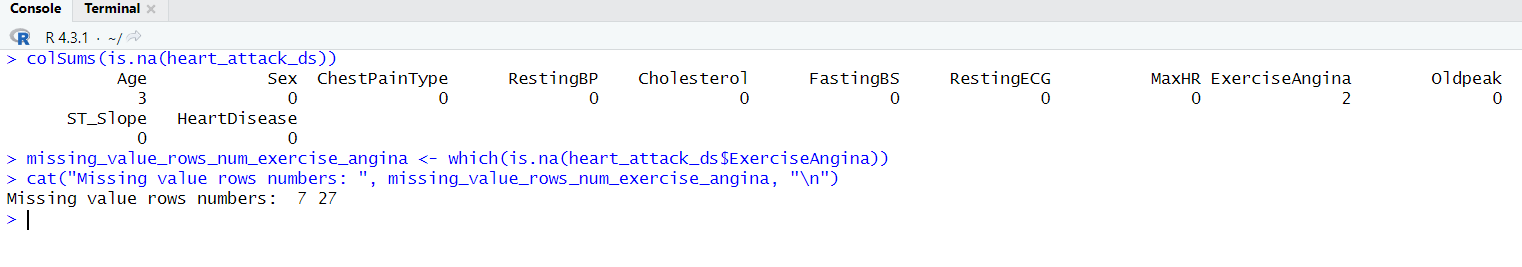




Checking if there are any missing values in Exercise Angina and finding row numbers:

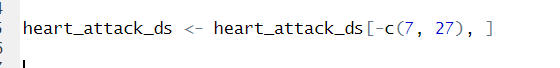
Code:

Output:

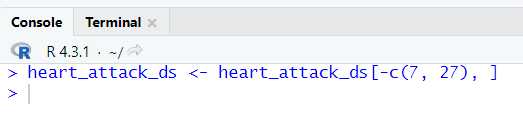


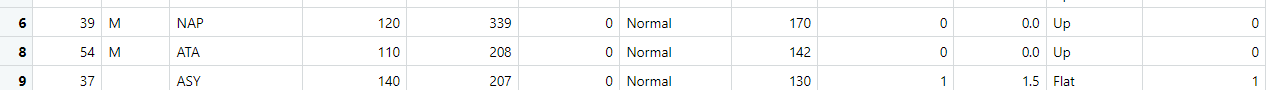
Removing missing values row:

Code:



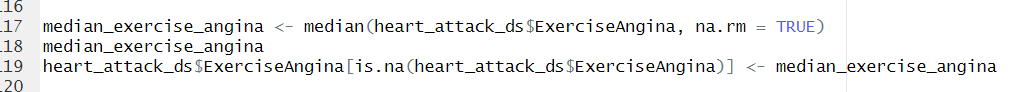
Output:



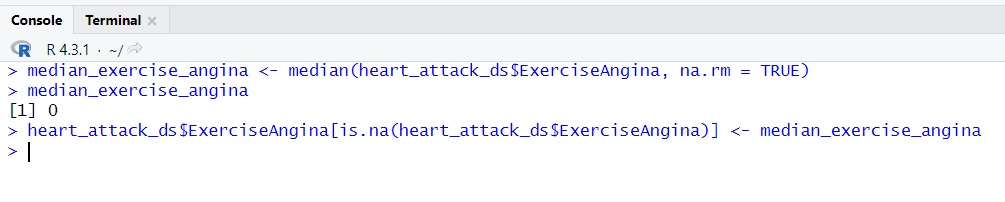


Recovering missing value using median:

Code:

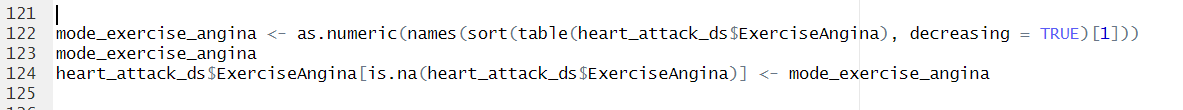


Output:

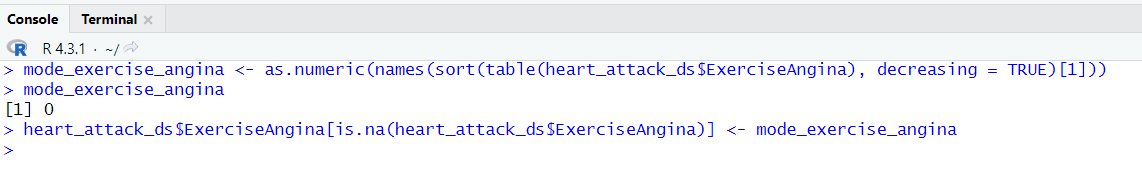


Recovering missing value using mode:

Code:



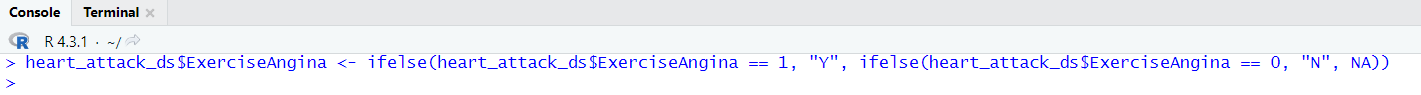
Output:



Transforming back to the original values of Exercise Angina Column:

Code:

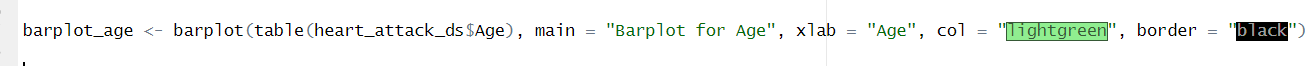
Output:



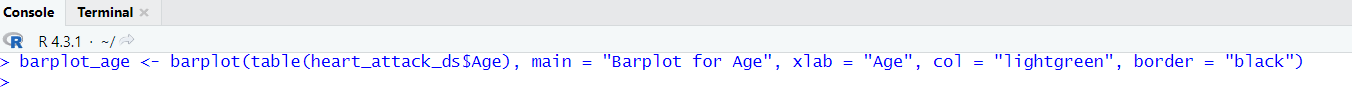
**Data Visualization:**

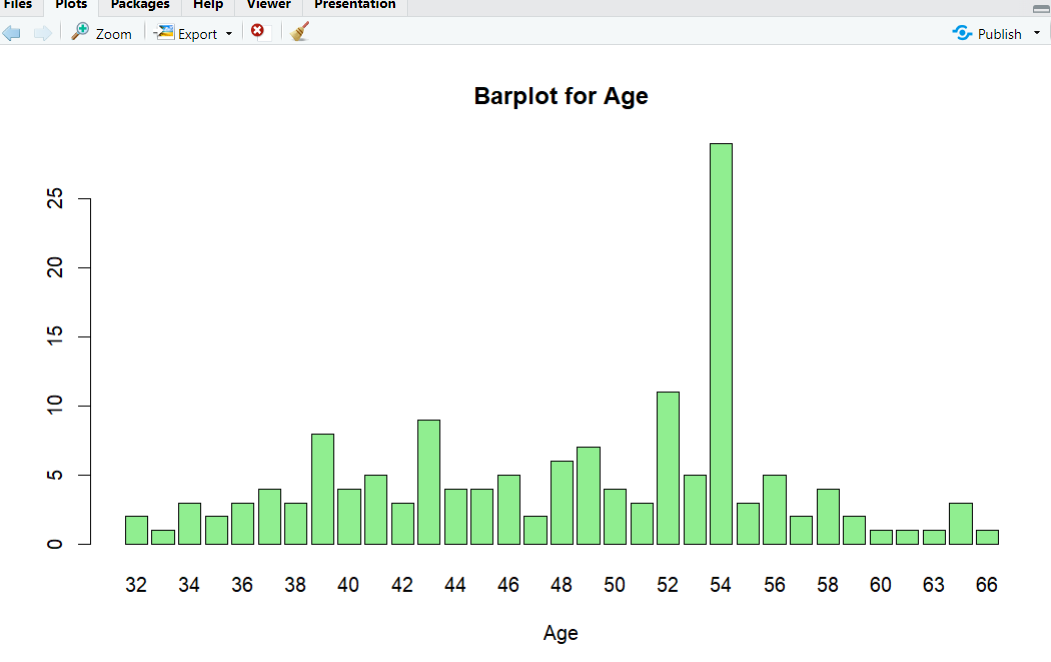
**Age Column:**

Code: In this code we plot a bar plot for Age column

****

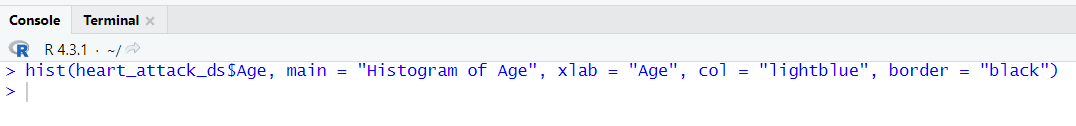
Output:

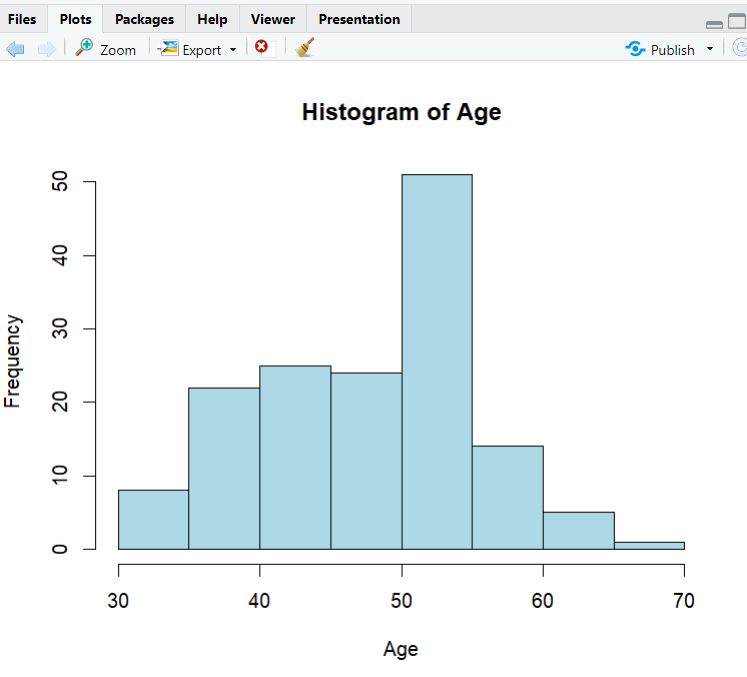




Code: In this code we plot a histogram for Age column

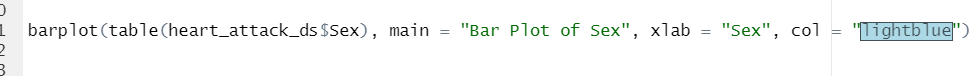
Output:



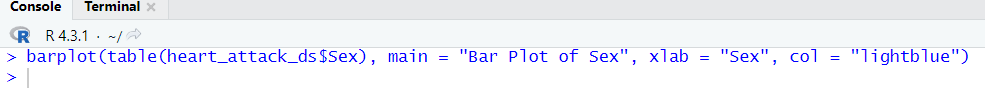


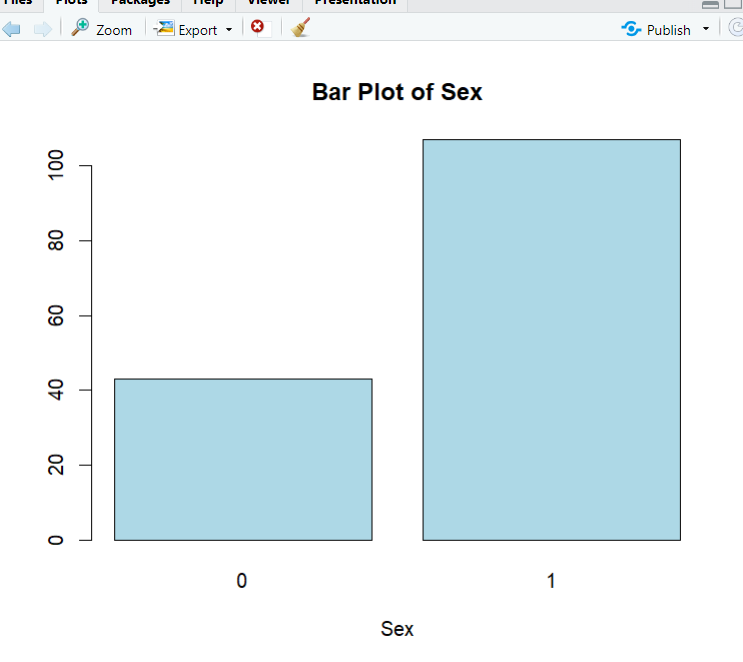
**Sex Column:**

Code: In this code we plot a bar plot for Sex column

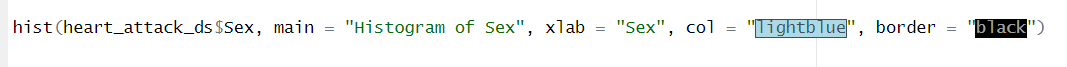


Output:

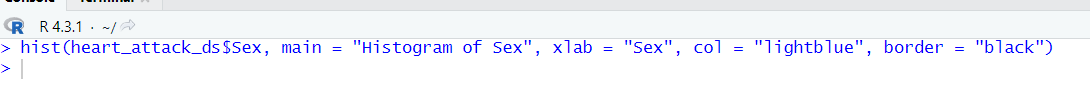


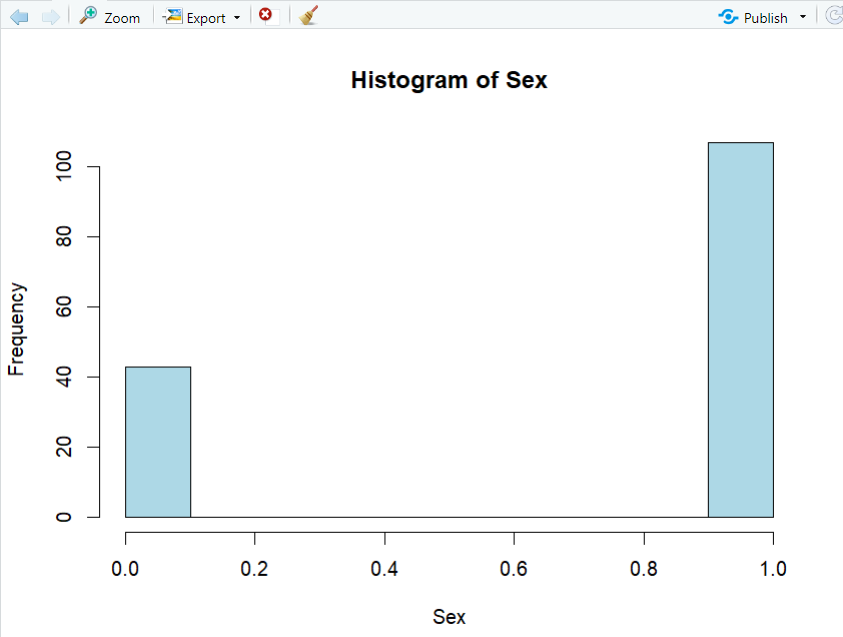


Code: In this code we plot a histogram for Sex column



Output:



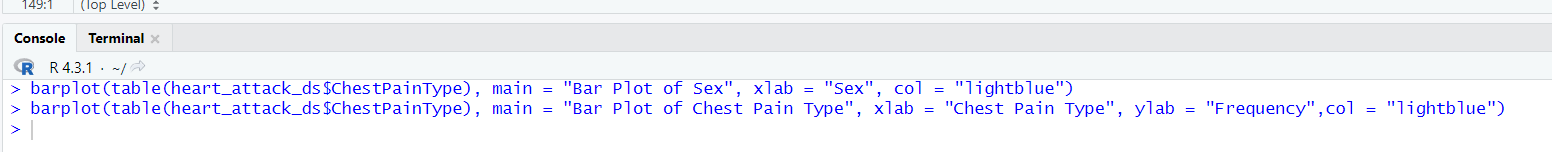


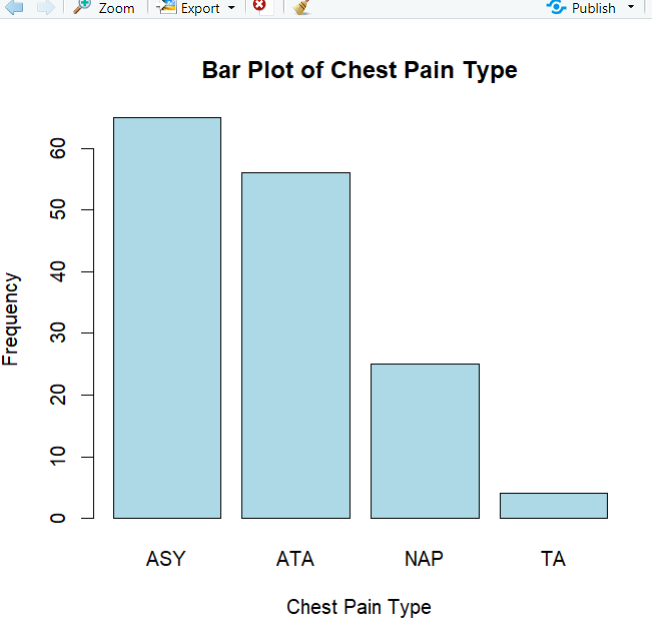
**Chest Pain Type Column:**

Code: In this code we plot a bar plot for Chest Pain Typecolumn



Output:



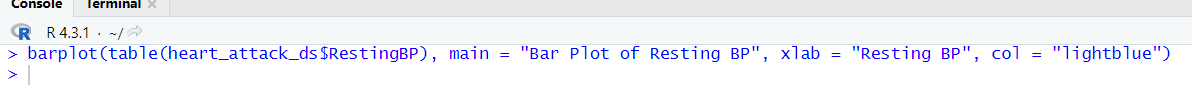


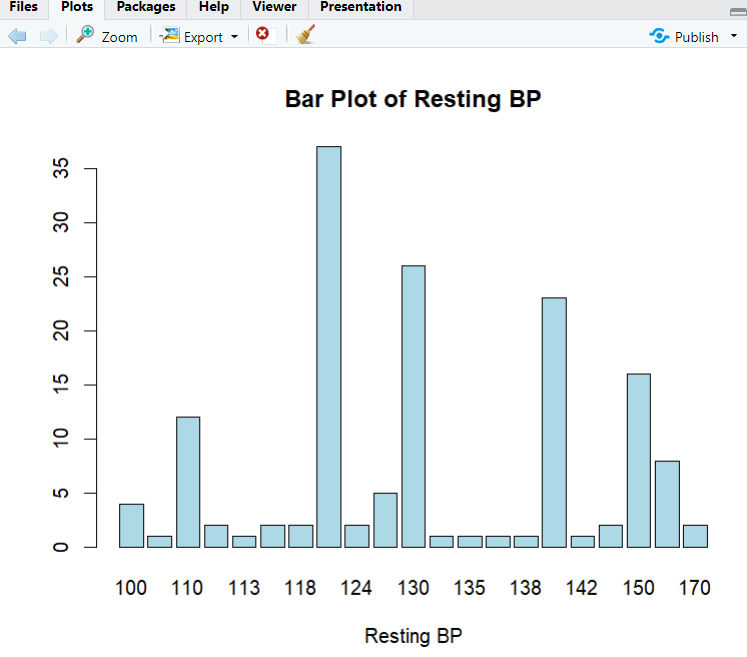
**Resting BP Column:**

Code: In this code we plot a bar plot for Resting BPcolumn



Output:

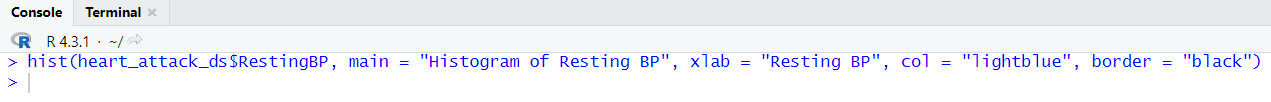


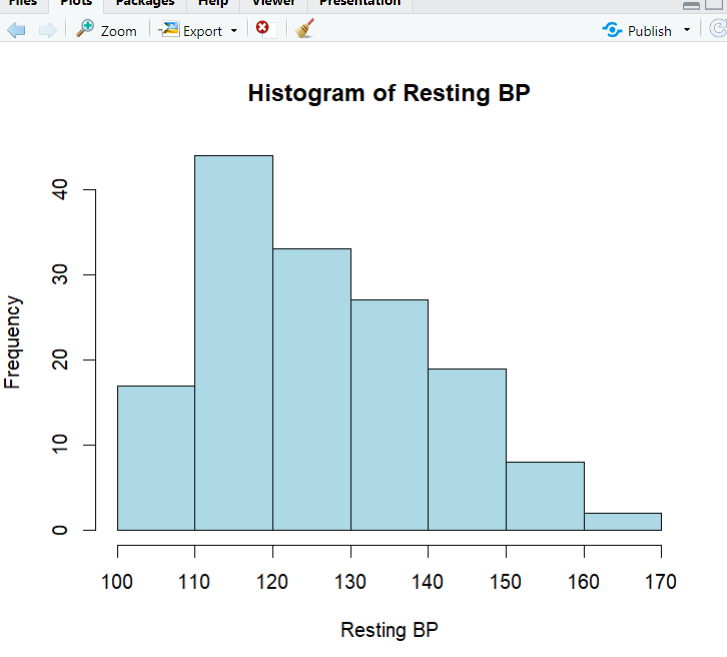


Code: In this code we plot a histogram for Resting BPcolumn



Output:

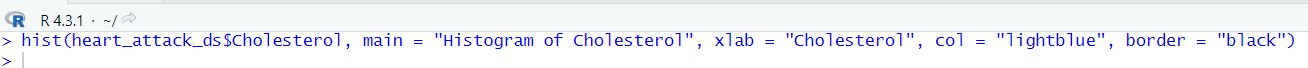


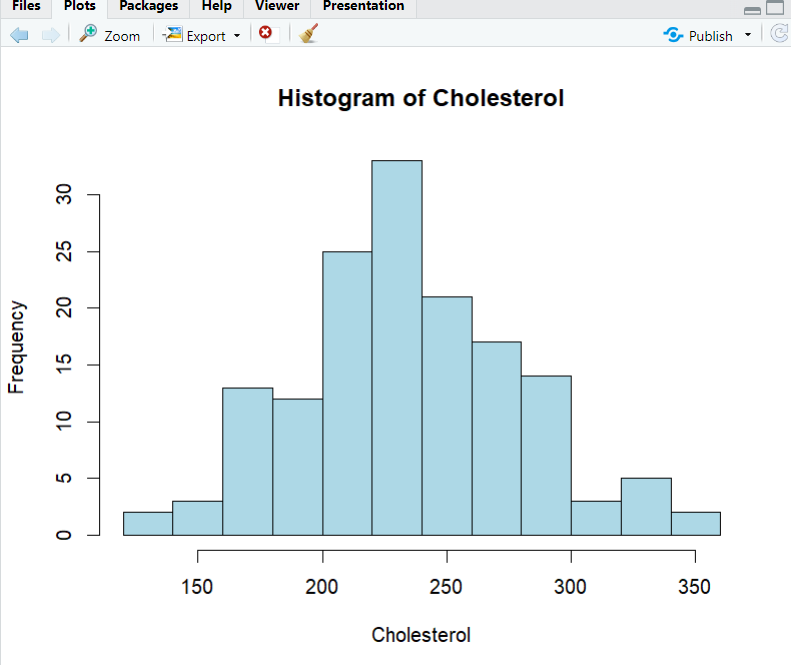


**Cholesterol Column:**

Code: In this code we plot a histogram for cholesterolcolumn

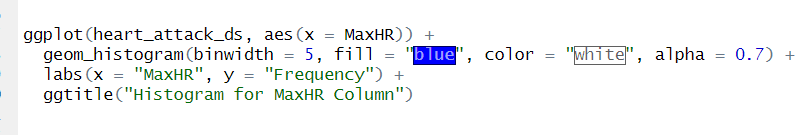
Output:



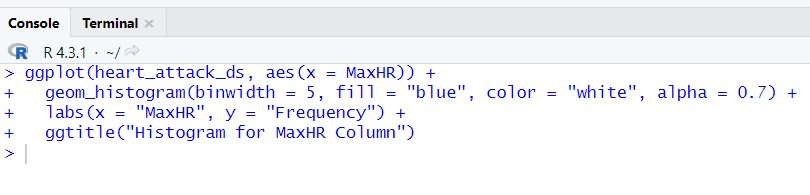


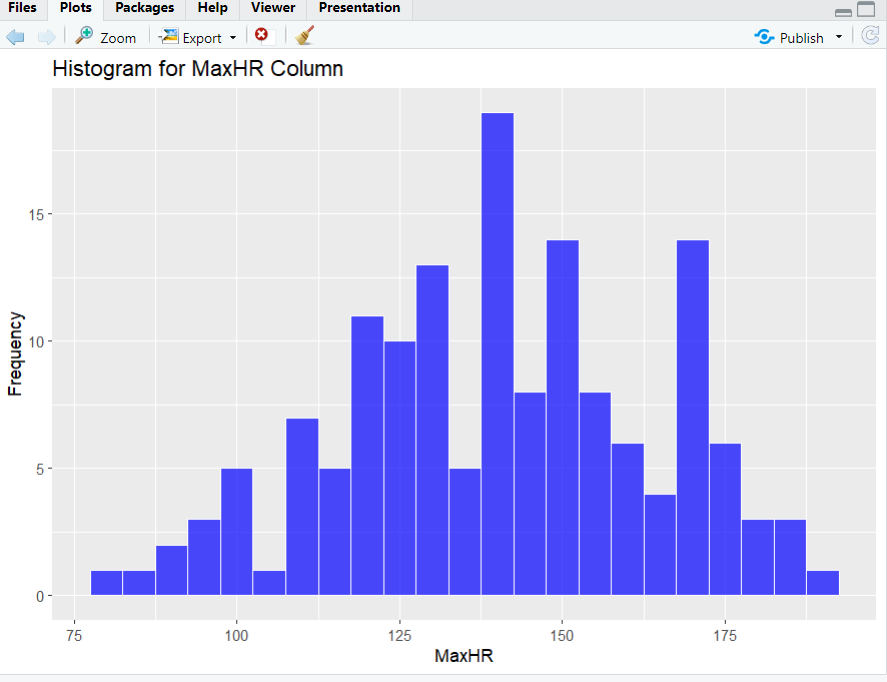
**Max Heart Rate Column:**

Code: In this code we plot a histogram for Max Heart Ratecolumn



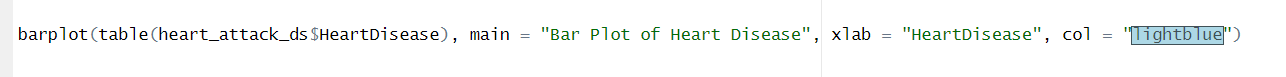
Output:



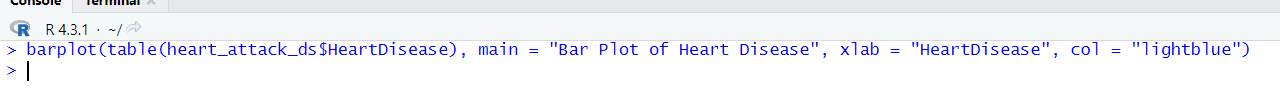


**Heart Disease:**

Code: In this code we plot a bar plot for Heart Diseasecolumn

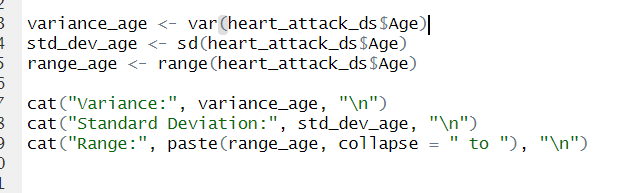


Output:

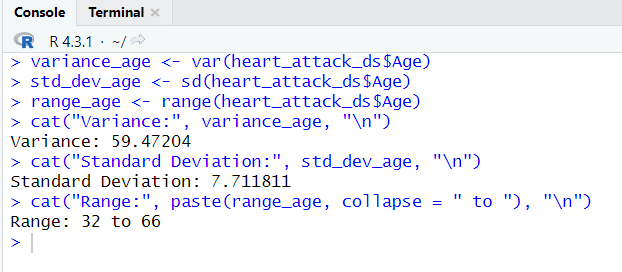


**Calculating measure of Spread for Age Column:**

Code:

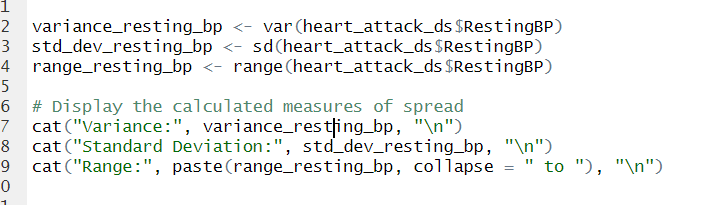


Output:

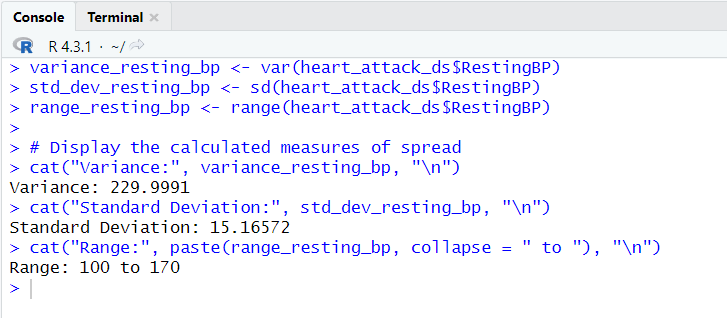


**Calculating measure of Spread for Resting BP Column:**

Code:

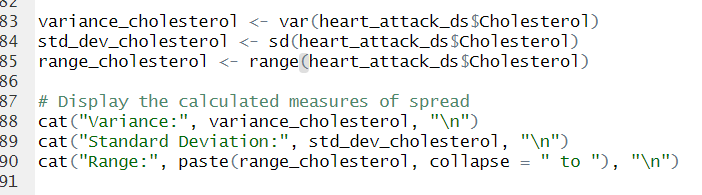


Output:

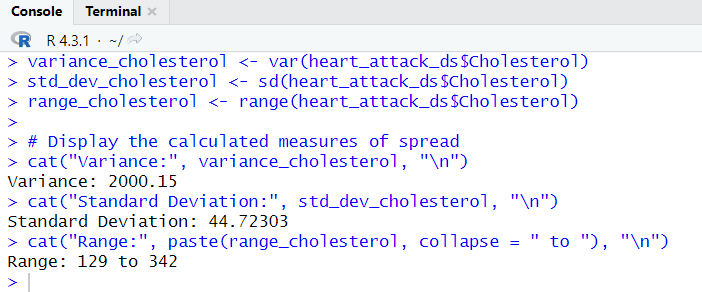


**Calculating measure of Spread for Cholesterol Column:**

Code:

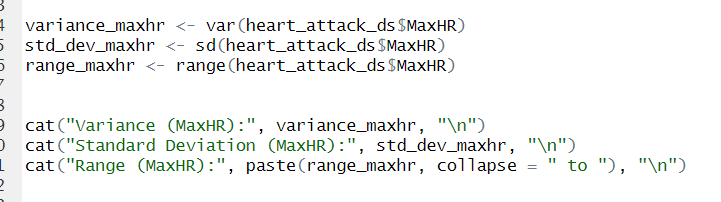


Output:



**Calculating measure of Spread for MaxHR Column:**

Code:



Output:

