**Introduction to Data Science**

**Midterm Project**

**Section: A**

Hadiur Rahman Nabil [20-42095-1]

Khan Nushrat Sultana Netu [20-43191-1]

# Dataset Details

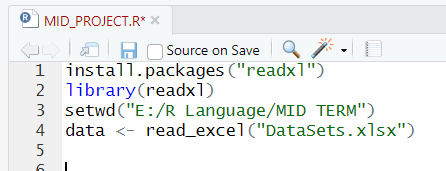
The "Heart Disease Classification" dataset consists of information from 150 patients. This dataset encompasses five attributes: "Age," "Gender," "Impulse," "PressureHigh," "PressureLow," "Glucose," and the target attribute labeled "Class."

The dataset contains a combination of both numerical and categorical data. This diversity in data types reflects the richness of patient information, ranging from age (numerical), gender (categorical), pulse rate (numerical), high blood pressure (numerical), and low blood pressure (numerical), to glucose levels (numerical).

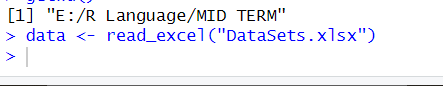
Furthermore, it is worth mentioning that this dataset exhibits missing values, implying that some data points within the dataset are incomplete. There are Outliers in an attribute. Handling these missing values is a critical part of data preprocessing.

# Load the Dataset

R script:



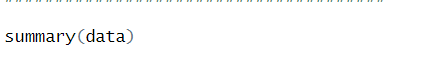
Consol:



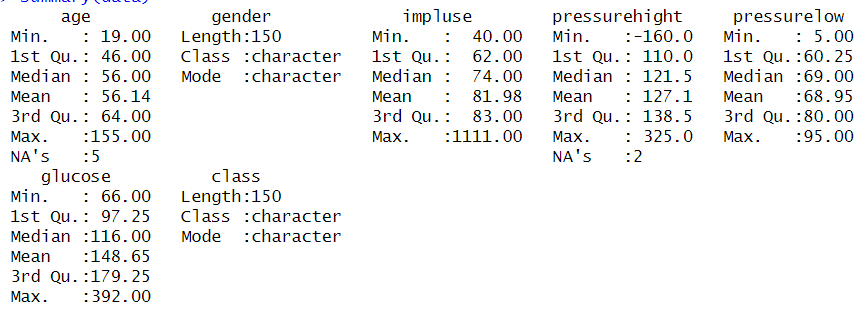
To load the XLSX file it is need to install readxl packages and then use the *setwd()* function to set the directory where the datasets are located. Using *data <- read\_excel("DataSets.xlsx")*  load the dataset into R Studio.

# Identify Missing Values

R Script:



Consol:



To get the details of the dataset there is a function named *summary()* that shows the missing value, the mean value, the median value, Min and Max value of each attribute that consists of numeric data. For the categorical value, this function shows the length of instances and type.

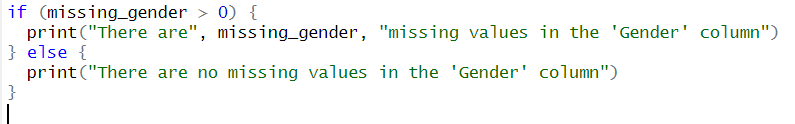
# Replace Categorical value with MODE value

R Script:



Using this *as.character()* the attribute ensures all the values in the gender attribute are character type. Then *max()* function ensures that the values occur for the maximum time. Then the second row select the missing value that was identified and replace that with the mode value.

R Script:



Consol:



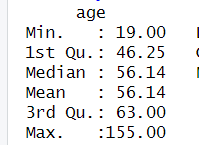
To verify if any missing value exists in the gender column, here it is used IF/ELSE.

# Using the MEAN value to replace the missing value

R Script:



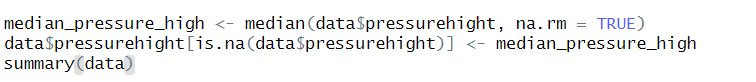
Consol:



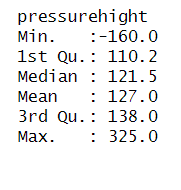
For the age attribute, *mean()* is used to replace all the missing values within this attribute. During this time all the NOT AVAILABLE value is ignored using *na.rm= TRUE.*

# Using the Median value to replace the missing value

R Script:



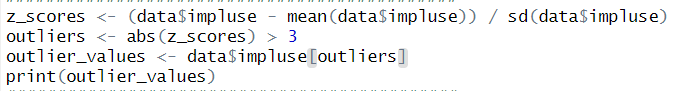
Consol:



In the pressurehight attribute, there are two missing values which are replaced with the median value. During this time all the missing fields are ignored to calculate.

# Detecting Outliers

R Script:



Consol:



To detect outliers we used the z-score in impluse attribute in the dataset, which shows how many standard deviations (sd) a reflection is above or below the mean of the dataset. The z-score is calculated by subtracting the mean of the dataset from the observation, and then dividing the result by the standard deviation of the dataset:

*z = (x - mean) / standard deviation*

Any observation with a z-score greater than 3 or less than -3 is considered an outlier.

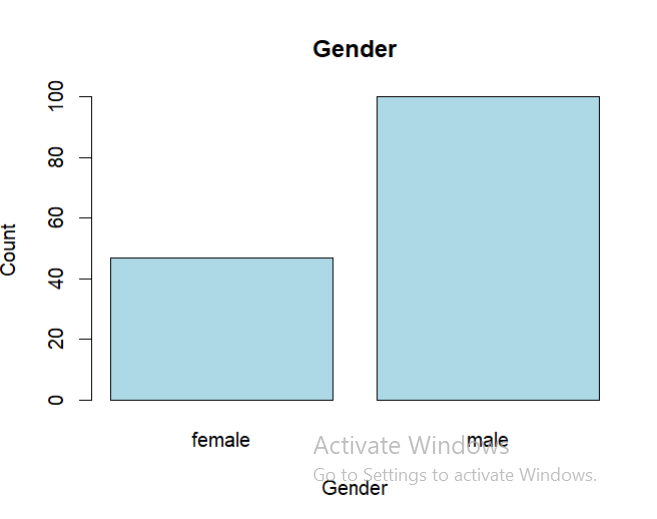
Here we have used the *abs()* function which takes the absolute value of the z-scores.

# Using BarPlot

R Script:



Consol:



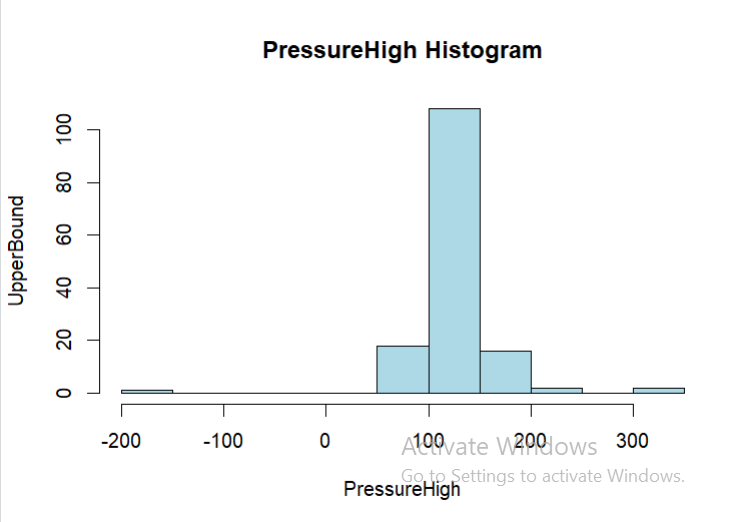
For categorical data we use BarPlot.

# Using Histogram

R Script:



Consol:

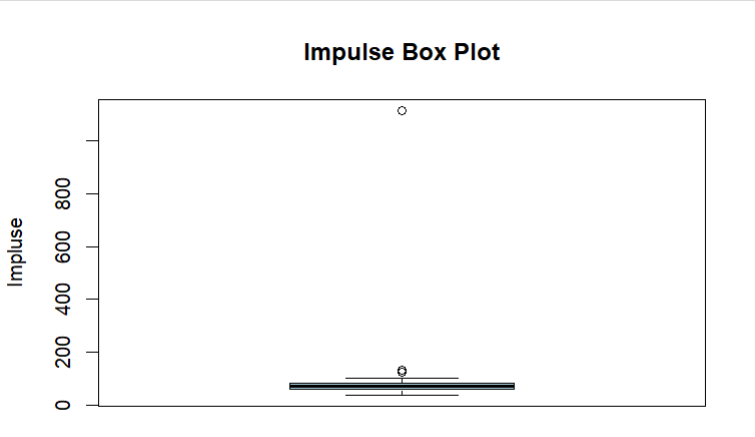


# Using BoxPlot

R Script:



Consol:



For outliers we use boxplot.