American International University-Bangladesh

Mid Project

Submitted to:

NAME: TOHEDUL ISLAM

Subject: INTRODUCTION TO DATA SCIENCE

SECTION: A

Submitted by:

Name	ID
Sayeem Bin Ezaz	20-43770-2
Ridita Zaman Adikta	20-43679-2
Maliha Mehjabin	20-42508-1

Date of Submission: 11.11.2023

Short Summary of Dataset:

The dataset contains information related to factors contributing to heart attacks and consists of 1319 samples, each comprising nine fields. The fields include age, gender, heart rate, systolic and diastolic blood pressure, blood sugar, CK-MB, and Test-Troponin. Gender is represented as 0 for Female and 1 for Male, while CK-MB and Test-Troponin may indicate enzyme or troponin measurements, respectively. The dataset's output field, labeled "Class," categorizes the presence of heart attacks into "Negative" for absence and "Positive" for presence. It's worth noting that there are missing values, particularly in the gender and CK-MB fields, and some unusual data values that would require further data cleaning and preprocessing for meaningful analysis.

Data Import and View:

Code: MidData<- read.csv("V:/MidProject.csv",header=TRUE ,sep=",")

MidData

```
> MidData
    age gender impluse pressurehight pressurelow glucose    class
1    64   male      66           160          83     160 negative
2    21   male      94            98          46     296 positive
3    55   male      64          -160          77     270 negative
4    64   male      70           120          55     270 positive
5    55   male      64           112          65     300 negative
6    58 female      61           112          58      87 negative
7    32 female      40           179          68     102 negative
8    63   male      60           214          82      87 positive
9    44 female      60            NA          81     135 negative
10   67             61           160          95     100 negative
11   NA female      60           166          90     102 negative
12   63 female      60           150          10     198 negative
13   64   male      60           199           5      92 positive
14   54 female      94           122          67      97 negative
15   47   male      76           120          70     319 negative
16   61   male      81            NA          66     134 positive
17   86 female      73           114          68      87 positive
18   45 female      70           100          68      96 negative
19   37 female      72           107          86     274 negative
20   45   male      60           109          65      89 positive
21   60   male      92           151          78     301 negative
22   48   male     135            98          60     100 positive
23   52   male      76           109          85     227 positive
24   30   male      63           110          68     107 positive
25   NA   male      63           320          63     269 positive
26   72   male      64           106          68     111 positive
27   42   male      65           150          68     101 negative
28   72 female      64           325          60      95 negative
29   47 female      66           134          57     279 positive
30   63   male      66           135          55     166 negative
31   54   male     125           131          82      95 positive
32   35   male      62           137          61     321 negative
33   68   male      61           121          49      98 positive
34   56 female      60           145          62     105 negative
35   50   male      61           136          70     136 positive
36   64   male      58           156          76      82 positive
37   NA   male      60           166          82     117 negative
38   64   male      65           155          75     107 negative
```

Description: Data is read into a variable called MidData from a CSV file.

Unwanted Sign or Invalid value corrected from Pressurehight:

Code: MidData$pressurehight<-gsub("-","",MidData$pressurehight)

MidData

```
> MidData$pressurehight<-gsub("-","",MidData$pressurehight)
> MidData
    age gender impluse pressurehight pressurelow glucose    class
1    64   male      66           160          83     160 negative
2    21   male      94            98          46     296 positive
3    55   male      64           160          77     270 negative
4    64   male      70           120          55     270 positive
5    55   male      64           112          65     300 negative
6    58 female      61           112          58      87 negative
7    32 female      40           179          68     102 negative
8    63   male      60           214          82      87 positive
9    44 female      60          <NA>          81     135 negative
10   67             61           160          95     100 negative
11   NA female      60           166          90     102 negative
12   63 female      60           150          10     198 negative
13   64   male      60           199           5      92 positive
14   54 female      94           122          67      97 negative
15   47   male      76           120          70     319 negative
16   61   male      81          <NA>          66     134 positive
17   86 female      73           114          68      87 positive
18   45 female      70           100          68      96 negative
19   37 female      72           107          86     274 negative
20   45   male      60           109          65      89 positive
21   60   male      92           151          78     301 negative
22   48   male     135            98          60     100 positive
23   52   male      76           109          85     227 positive
24   30   male      63           110          68     107 positive
25   NA   male      63           320          63     269 positive
26   72   male      64           106          68     111 positive
27   42   male      65           150          68     101 negative
28   72 female      64           325          60      95 negative
29   47 female      66           134          57     279 positive
30   63   male      66           135          55     166 negative
31   54   male     125           131          82      95 positive
32   35   male      62           137          61     321 negative
33   68   male      61           121          49      98 positive
34   56 female      60           145          62     105 negative
```

Description: In this dataset gsub() function is used to corrected or remove unwanted signs. There was a "-" sign at preassureheight column. Which may create problem to analysis the data or during work with data.

Replace blank values with null values:

Code:

```
MidData[MidData == ""] <- NA
```

MidData

```
> MidData[MidData == ""] <- NA
> MidData
    age gender impluse pressurehight pressurelow glucose    class
1    64   male      66           160          83     160 negative
2    21   male      94            98          46     296 positive
3    55   male      64           160          77     270 negative
4    64   male      70           120          55     270 positive
5    55   male      64           112          65     300 negative
6    58 female      61           112          58      87 negative
7    32 female      40           179          68     102 negative
8    63   male      60           214          82      87 positive
9    44 female      60          <NA>          81     135 negative
10   67   <NA>      61           160          95     100 negative
11   NA female      60           166          90     102 negative
12   63 female      60           150          10     198 negative
13   64   male      60           199           5      92 positive
14   54 female      94           122          67      97 negative
15   47   male      76           120          70     319 negative
16   61   male      81          <NA>          66     134 positive
17   86 female      73           114          68      87 positive
18   45 female      70           100          68      96 negative
19   37 female      72           107          86     274 negative
20   45   male      60           109          65      89 positive
21   60   male      92           151          78     301 negative
22   48   male     135            98          60     100 positive
23   52   male      76           109          85     227 positive
24   30   male      63           110          68     107 positive
25   NA   male      63           320          63     269 positive
26   72   male      64           106          68     111 positive
27   42   male      65           150          68     101 negative
28   72 female      64           325          60      95 negative
29   47 female      66           134          57     279 positive
30   63   male      66           135          55     166 negative
31   54   male     125           131          82      95 positive
32   35   male      62           137          61     321 negative
```

Description: Removing blank spaces with null value is important. Cause sometime blank space takes place as not null value.

Summary of Dataset:

Code:

MidData$pressurehight<-as.integer(MidData$pressurehight)

summary(MidData)

```
> MidData$pressurehight<-as.integer(MidData$pressurehight)
> summary(MidData)
      age            gender             impluse        pressurehight    pressurelow       glucose          class          
 Min.   : 19.00   Length:150         Min.   :  40.00   Min.   : 85.0   Min.   : 5.00   Min.   : 66.00   Length:150        
 1st Qu.: 46.00   Class :character   1st Qu.:  62.00   1st Qu.:110.8   1st Qu.:60.25   1st Qu.: 97.25   Class :character  
 Median : 56.00   Mode  :character   Median :  74.00   Median :122.5   Median :69.00   Median :116.00   Mode  :character  
 Mean   : 56.14                      Mean   :  81.98   Mean   :129.2   Mean   :68.95   Mean   :148.65                     
 3rd Qu.: 64.00                      3rd Qu.:  83.00   3rd Qu.:140.0   3rd Qu.:80.00   3rd Qu.:179.25                     
 Max.   :155.00                      Max.   :1111.00   Max.   :325.0   Max.   :95.00   Max.   :392.00                     
 NA's   :5                                             NA's   :2
```

Description: This uses the summary() function to create a summary of the MidData data. For removing unwanted sign data type of preassurehight column changed. So we use as.integer function again to change the data type & get the proper summary of data.

Display Dataset:

Code: str(MidData)

```
> str(MidData)
'data.frame':   150 obs. of  7 variables:
 $ age           : int  64 21 55 64 55 58 32 63 44 67 ...
 $ gender        : chr  "male" "male" "male" "male" ...
 $ impluse       : int  66 94 64 70 64 61 40 60 60 61 ...
 $ pressurehight: int  160 98 160 120 112 112 179 214 NA 160 ...
 $ pressurelow   : int  83 46 77 55 65 58 68 82 81 95 ...
 $ glucose       : int  160 296 270 270 300 87 102 87 135 100 ...
 $ class         : chr  "negative" "positive" "negative" "positive" ...
```

Description: R object structures are shown with "str." The data's contents are displayed using the string "str". An alternate function to show the output summary is str(MidData), particularly in cases where the data set is large.

Missing Value Detection (Numeric Value):

Code: is.na(MidData)

```
> is.na(MidData)
          age gender impluse pressurehight pressurelow glucose class
  [1,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [2,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [3,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [4,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [5,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [6,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [7,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [8,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
  [9,] FALSE  FALSE   FALSE           TRUE       FALSE   FALSE FALSE
 [10,] FALSE   TRUE   FALSE          FALSE       FALSE   FALSE FALSE
 [11,]  TRUE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [12,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [13,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [14,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [15,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [16,] FALSE  FALSE   FALSE           TRUE       FALSE   FALSE FALSE
 [17,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [18,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [19,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [20,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [21,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [22,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [23,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [24,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [25,]  TRUE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [26,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [27,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [28,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [29,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [30,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [31,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [32,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [33,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [34,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [35,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
 [36,] FALSE  FALSE   FALSE          FALSE       FALSE   FALSE FALSE
```

Description: A data frame or vector's missing values can be found using the is.na() function. With each element being TRUE if the associated element in the data frame or vector is missing and FALSE otherwise, it yields a logical vector with the same length as the input data.

Missing Value Count in Each Column

Code: colSums(is.na(MidData))

```
> colSums(is.na(MidData))
          age        gender       impluse pressurehight   pressurelow       glucose         class
            5             3             0             2             0             0             0
```

Description: The number of null elements in each column has been determined using the colSums(is.na()) function.

Specific Missing Value row Number

Code: which(is.na(MidData$age))

```
> which(is.na(MidData$age))
[1]  11  25  37  74 122
>
```

Description: The which() function is used for getting exact number of row is missing. We used this for age to see the missing values.

Standard Deviation of all numeric values:

Code: age<-MidData$age

sd(age)

impluse<-MidData$impluse

sd(impluse)

pressurehight<-MidData$pressurehight

sd(pressurehight)

pressurelow<-MidData$pressurelow

sd(pressurelow)

glucose<-MidData$glucose

sd(glucose)

```
> age<-MidData$age
> sd(age)
[1] 17.1392
> impluse<-MidData$impluse
> sd(impluse )
[1] 14.78093
> pressurehight<-MidData$pressurehight
> sd(pressurehight)
[1] 32.58258
> pressurelow<-MidData$pressurelow
> sd(pressurelow)
[1] 13.61182
> glucose<-MidData$glucose
> sd(glucose)
[1] 73.58597
```

Description: The sd() function computes the standard deviation of the column values. The sd() function in R is a built-in function that computes the standard deviation, which is a measure of the amount of variation or dispersion in a set of values. This step is done after all the data cleaning process.

Handling invalid data/outliers in the data set:

Removing Missing Value:

Code: RemovedMidData<-na.omit(MidData)

RemovedMidData

```
> RemovedMidData<-na.omit(MidData)
> RemovedMidData
    age gender impluse pressurehight pressurelow glucose    class
1    64   male      66            160          83     160 negative
2    21   male      94             98          46     296 positive
3    55   male      64            160          77     270 negative
4    64   male      70            120          55     270 positive
5    55   male      64            112          65     300 negative
6    58 female      61            112          58      87 negative
7    32 female      40            179          68     102 negative
8    63   male      60            214          82      87 positive
12   63 female      60            150          10     198 negative
13   64   male      60            199           5      92 positive
14   54 female      94            122          67      97 negative
15   47   male      76            120          70     319 negative
17   86 female      73            114          68      87 positive
18   45 female      70            100          68      96 negative
19   37 female      72            107          86     274 negative
20   45   male      60            109          65      89 positive
21   60   male      92            151          78     301 negative
22   48   male     135             98          60     100 positive
23   52   male      76            109          85     227 positive
24   30   male      63            110          68     107 positive
26   72   male      64            106          68     111 positive
27   42   male      65            150          68     101 negative
28   72 female      64            325          60      95 negative
29   47 female      66            134          57     279 positive
30   63   male      66            135          55     166 negative
31   54   male     125            131          82      95 positive
32   35   male      62            137          61     321 negative
33   68   male      61            121          49      98 positive
34   56 female      60            145          62     105 negative
35   50   male      61            136          70     136 positive
36   64   male      58            156          76      82 positive
38   64   male      65            155          75     107 negative
```

Description: The na.omit() function is used to remove rows from the data MidData that have missing values. The data that has missing values removed is then stored in a new data frame called RemovedMidData.

Removing outliers from Impulse Column:

Code:

MidData$impluse[MidData$impluse > 190] <- NA

MidData

```
> MidData$impluse[MidData$impluse > 190] <- NA
> MidData
    age gender impluse pressurehight pressurelow glucose    class
1    64   male      66           160          83     160 negative
2    21   male      94            98          46     296 positive
3    55   male      64           160          77     270 negative
4    64   male      70           120          55     270 positive
5    55   male      64           112          65     300 negative
6    58 female      61           112          58      87 negative
7    32 female      NA           179          68     102 negative
8    63   male      60           214          82      87 positive
9    44 female      60            NA          81     135 negative
10   67   <NA>      61           160          95     100 negative
11   NA female      60           166          90     102 negative
12   63 female      60           150          10     198 negative
13   64   male      60           199           5      92 positive
14   54 female      94           122          67      97 negative
15   47   male      76           120          70     319 negative
16   61   male      81            NA          66     134 positive
17   86 female      73           114          68      87 positive
18   45 female      70           100          68      96 negative
19   37 female      72           107          86     274 negative
20   45   male      60           109          65      89 positive
21   60   male      92           151          78     301 negative
22   48   male      NA            98          60     100 positive
23   52   male      76           109          85     227 positive
24   30   male      63           110          68     107 positive
25   NA   male      63           320          63     269 positive
26   72   male      64           106          68     111 positive
27   42   male      65           150          68     101 negative
28   72 female      64           325          60      95 negative
29   47 female      66           134          57     279 positive
30   63   male      66           135          55     166 negative
31   54   male      NA           131          82      95 positive
32   35   male      62           137          61     321 negative
33   68   male      61           121          49      98 positive
34   56 female      60           145          62     105 negative
```

Description:

The maximum range of human Impulse is 190 beasts per minutes. From the MidData dataset the noisy impulses of human's are replaced with null value which are out of the range.

Removing missing values of Gender:

Code:

```
MidData <- MidData[!is.na(MidData$gender),]
MidData
```

```
> MidData <- MidData[!is.na(MidData$gender),]
> MidData
          age gender impluse pressurehight pressurelow glucose    class
1    64.00000   male      66           160          83     160 negative
2    21.00000   male      94            98          46     296 positive
3    55.00000   male      64           160          77     270 negative
4    64.00000   male      70           120          55     270 positive
5    55.00000   male      64           112          65     300 negative
6    58.00000 female      61           112          58      87 negative
7    32.00000 female      40           179          68     102 negative
8    63.00000   male      60           214          82      87 positive
9    44.00000 female      60           135          81     135 negative
11   56.13793 female      60           166          90     102 negative
12   63.00000 female      60           150          10     198 negative
13   64.00000   male      60           199           5      92 positive
14   54.00000 female      94           122          67      97 negative
15   47.00000   male      76           120          70     319 negative
16   61.00000   male      81           135          66     134 positive
17   86.00000 female      73           114          68      87 positive
18   45.00000 female      70           100          68      96 negative
19   37.00000 female      72           107          86     274 negative
20   45.00000   male      60           109          65      89 positive
21   60.00000   male      92           151          78     301 negative
22   48.00000   male     135            98          60     100 positive
23   52.00000   male      76           109          85     227 positive
24   30.00000   male      63           110          68     107 positive
25   56.13793   male      63           320          63     269 positive
26   72.00000   male      64           106          68     111 positive
27   42.00000   male      65           150          68     101 negative
28   72.00000 female      64           325          60      95 negative
29   47.00000 female      66           134          57     279 positive
30   63.00000   male      66           135          55     166 negative
31   54.00000   male     125           131          82      95 positive
32   35.00000   male      62           137          61     321 negative
33   68.00000   male      61           121          49      98 positive
34   56.00000 female      60           145          62     105 negative
35   50.00000   male      61           136          70     136 positive
36   64.00000   male      58           156          76      82 positive
37   56.13793   male      60           166          82     117 negative
38   64.00000   male      65           155          75     107 negative
40   34.00000   male      96           105          75     136 positive
41   44.00000   male      94            91          52     208 negative
```

Description:!is.na() function is used for removing missing values of specific column. The missing gender column may impact the nature of our dataset, so we removed gender from our MidData.

Replace Missing values With Mean, Median, Mode

Replace missing values with mean value:

Code: Mean_age<-mean(MidData$age,na.rm = TRUE)
Mean_age
MidData$age[is.na(MidData$age)]<-Mean_age
MidData

```
> Mean_age<-mean(MidData$age,na.rm = TRUE)
> Mean_age
[1] 56.09313
> MidData$age[is.na(MidData$age)]<-Mean_age
> MidData
          age gender impluse pressurehight pressurelow glucose    class
1    64.00000   male      66            160          83     160 negative
2    21.00000   male      94             98          46     296 positive
3    55.00000   male      64            160          77     270 negative
4    64.00000   male      70            120          55     270 positive
5    55.00000   male      64            112          65     300 negative
6    58.00000 female      61            112          58      87 negative
7    32.00000 female      40            179          68     102 negative
8    63.00000   male      60            214          82      87 positive
9    44.00000 female      60             NA          81     135 negative
11   56.13793 female      60            166          90     102 negative
12   63.00000 female      60            150          10     198 negative
13   64.00000   male      60            199           5      92 positive
14   54.00000 female      94            122          67      97 negative
15   47.00000   male      76            120          70     319 negative
16   61.00000   male      81             NA          66     134 positive
17   86.00000 female      73            114          68      87 positive
18   45.00000 female      70            100          68      96 negative
19   37.00000 female      72            107          86     274 negative
20   45.00000   male      60            109          65      89 positive
21   60.00000   male      92            151          78     301 negative
22   48.00000   male     135             98          60     100 positive
23   52.00000   male      76            109          85     227 positive
24   30.00000   male      63            110          68     107 positive
25   56.13793   male      63            320          63     269 positive
26   72.00000   male      64            106          68     111 positive
27   42.00000   male      65            150          68     101 negative
28   72.00000 female      64            325          60      95 negative
29   47.00000 female      66            134          57     279 positive
30   63.00000   male      66            135          55     166 negative
31   54.00000   male     125            131          82      95 positive
32   35.00000   male      62            137          61     321 negative
33   68.00000   male      61            121          49      98 positive
34   56.00000 female      60            145          62     105 negative
```

Description: Calculated the mean value of the "age" column first, and then replaced the missing values with the calculated mean value. We don't want to consider missing values so we used na.rm=TRUE function

Replace missing values of Impulse using median method

Code: Median_impulse<-median(MidData$impluse,na.rm = TRUE)
Median_impulse
MidData$impluse[is.na(MidData$impluse)]<-Median_impulse
MidData

```
> Median_impulse<-median(MidData$impluse,na.rm = TRUE)
> Median_impulse
[1] 74
> MidData$impluse[is.na(MidData$impluse)]<-Median_impulse
> MidData
          age gender impluse pressurehight pressurelow glucose    class
1    64.00000   male      66           160          83     160 negative
2    21.00000   male      94            98          46     296 positive
3    55.00000   male      64           160          77     270 negative
4    64.00000   male      70           120          55     270 positive
5    55.00000   male      64           112          65     300 negative
6    58.00000 female      61           112          58      87 negative
7    32.00000 female      40           179          68     102 negative
8    63.00000   male      60           214          82      87 positive
9    44.00000 female      60            NA          81     135 negative
10   67.00000   <NA>      61           160          95     100 negative
11   56.13793 female      60           166          90     102 negative
12   63.00000 female      60           150          10     198 negative
13   64.00000   male      60           199           5      92 positive
14   54.00000 female      94           122          67      97 negative
15   47.00000   male      76           120          70     319 negative
16   61.00000   male      81            NA          66     134 positive
17   86.00000 female      73           114          68      87 positive
18   45.00000 female      70           100          68      96 negative
19   37.00000 female      72           107          86     274 negative
20   45.00000   male      60           109          65      89 positive
21   60.00000   male      92           151          78     301 negative
22   48.00000   male     135            98          60     100 positive
23   52.00000   male      76           109          85     227 positive
24   30.00000   male      63           110          68     107 positive
25   56.13793   male      63           320          63     269 positive
26   72.00000   male      64           106          68     111 positive
27   42.00000   male      65           150          68     101 negative
28   72.00000 female      64           325          60      95 negative
29   47.00000 female      66           134          57     279 positive
30   63.00000   male      66           135          55     166 negative
31   54.00000   male     125           131          82      95 positive
32   35.00000   male      62           137          61     321 negative
33   68.00000   male      61           121          49      98 positive
```

Description: Calculated the median value of the "impulse" column first, and then replaced the missing values with the calculated median value. We don't want to consider missing values so we used na.rm=TRUE function

Replace missing values of Pressurehight using mode method

Code:

```
custom_mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
Mode_pressurehight <-custom_mode(MidData$pressurehight)
Mode_pressurehight
Midata$pressurehight[is.na(MidData$pressurehight)]<-Mode_pressurehight
MidData
```

```
> Mode_pressurehight <-custom_mode(MidData$pressurehight)
> Mode_pressurehight
[1] 135
> MidData$pressurehight[is.na(MidData$pressurehight)]<-Mode_pressurehight
> MidData
          age gender impluse pressurehight pressurelow glucose    class
1    64.00000   male      66           160          83     160 negative
2    21.00000   male      94            98          46     296 positive
3    55.00000   male      64           160          77     270 negative
4    64.00000   male      70           120          55     270 positive
5    55.00000   male      64           112          65     300 negative
6    58.00000 female      61           112          58      87 negative
7    32.00000 female      40           179          68     102 negative
8    63.00000   male      60           214          82      87 positive
9    44.00000 female      60           135          81     135 negative
10   67.00000   <NA>      61           160          95     100 negative
11   56.13793 female      60           166          90     102 negative
12   63.00000 female      60           150          10     198 negative
13   64.00000   male      60           199           5      92 positive
14   54.00000 female      94           122          67      97 negative
15   47.00000   male      76           120          70     319 negative
16   61.00000   male      81           135          66     134 positive
17   86.00000 female      73           114          68      87 positive
18   45.00000 female      70           100          68      96 negative
19   37.00000 female      72           107          86     274 negative
20   45.00000   male      60           109          65      89 positive
21   60.00000   male      92           151          78     301 negative
22   48.00000   male     135            98          60     100 positive
23   52.00000   male      76           109          85     227 positive
24   30.00000   male      63           110          68     107 positive
25   56.13793   male      63           320          63     269 positive
26   72.00000   male      64           106          68     111 positive
27   42.00000   male      65           150          68     101 negative
28   72.00000 female      64           325          60      95 negative
29   47.00000 female      66           134          57     279 positive
30   63.00000   male      66           135          55     166 negative
31   54.00000   male     125           131          82      95 positive
```

Description: Calculated the mode value of the "pressurehight " column first, and then replaced the missing values with the calculated mode value. We don't get mode function by default so we used Custom_mode function to calculate mode & replace it.

Data Annotation:

Annotating Gender Column:

Code:

MidData$gender<-factor(MidData$gender,levels = c("male","female"),labels = c(1,2))

MidData

```
> MidData$gender<-factor(MidData$gender,levels = c("male","female"),labels = c(1,2))
> MidData
          age gender impluse pressurehight pressurelow glucose    class
1    64.00000      1      66           160          83     160 negative
2    21.00000      1      94            98          46     296 positive
3    55.00000      1      64           160          77     270 negative
4    64.00000      1      70           120          55     270 positive
5    55.00000      1      64           112          65     300 negative
6    58.00000      2      61           112          58      87 negative
7    32.00000      2      40           179          68     102 negative
8    63.00000      1      60           214          82      87 positive
9    44.00000      2      60           135          81     135 negative
11   56.13793      2      60           166          90     102 negative
12   63.00000      2      60           150          10     198 negative
13   64.00000      1      60           199           5      92 positive
14   54.00000      2      94           122          67      97 negative
15   47.00000      1      76           120          70     319 negative
16   61.00000      1      81           135          66     134 positive
17   86.00000      2      73           114          68      87 positive
18   45.00000      2      70           100          68      96 negative
19   37.00000      2      72           107          86     274 negative
20   45.00000      1      60           109          65      89 positive
21   60.00000      1      92           151          78     301 negative
22   48.00000      1     135            98          60     100 positive
23   52.00000      1      76           109          85     227 positive
24   30.00000      1      63           110          68     107 positive
25   56.13793      1      63           320          63     269 positive
26   72.00000      1      64           106          68     111 positive
27   42.00000      1      65           150          68     101 negative
28   72.00000      2      64           325          60      95 negative
29   47.00000      2      66           134          57     279 positive
30   63.00000      1      66           135          55     166 negative
31   54.00000      1     125           131          82      95 positive
32   35.00000      1      62           137          61     321 negative
33   68.00000      1      61           121          49      98 positive
34   56.00000      2      60           145          62     105 negative
35   50.00000      1      61           136          70     136 positive
```

Description: This method is more helpful to represent data in more meaningful ways. Here for Gender column has 1 & 2 values, which considered to represent “male” & “female” respectably. We replaced “male” value with 1 & “female” value with 2.

Annotating Class Column:

Code:

```
MidData$class<-factor(MidData$class,levels = c("positive","negative"),labels = c(1,2))

MidData
```

```
> MidData$class<-factor(MidData$class,levels = c("positive","negative"),labels = c(1,2))
> MidData
          age gender impluse pressurehight pressurelow glucose class
1    64.00000      1      66           160          83     160     2
2    21.00000      1      94            98          46     296     1
3    55.00000      1      64           160          77     270     2
4    64.00000      1      70           120          55     270     1
5    55.00000      1      64           112          65     300     2
6    58.00000      2      61           112          58      87     2
7    32.00000      2      40           179          68     102     2
8    63.00000      1      60           214          82      87     1
9    44.00000      2      60           135          81     135     2
11   56.13793      2      60           166          90     102     2
12   63.00000      2      60           150          10     198     2
13   64.00000      1      60           199           5      92     1
14   54.00000      2      94           122          67      97     2
15   47.00000      1      76           120          70     319     2
16   61.00000      1      81           135          66     134     1
17   86.00000      2      73           114          68      87     1
18   45.00000      2      70           100          68      96     2
19   37.00000      2      72           107          86     274     2
20   45.00000      1      60           109          65      89     1
21   60.00000      1      92           151          78     301     2
22   48.00000      1     135            98          60     100     1
23   52.00000      1      76           109          85     227     1
24   30.00000      1      63           110          68     107     1
25   56.13793      1      63           320          63     269     1
26   72.00000      1      64           106          68     111     1
27   42.00000      1      65           150          68     101     2
28   72.00000      2      64           325          60      95     2
29   47.00000      2      66           134          57     279     1
30   63.00000      1      66           135          55     166     2
31   54.00000      1     125           131          82      95     1
32   35.00000      1      62           137          61     321     2
33   68.00000      1      61           121          49      98     1
34   56.00000      2      60           145          62     105     2
35   50.00000      1      61           136          70     136     1
```

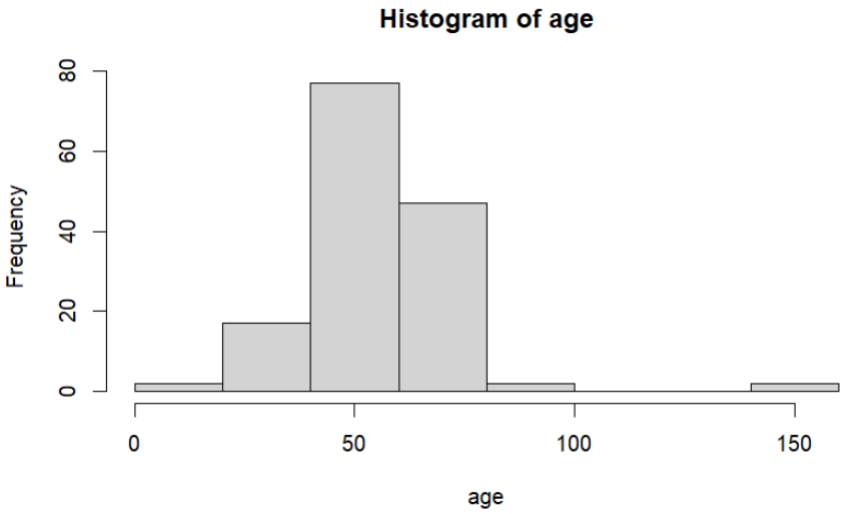
Description: This method is more helpful to represent data in more meaningful ways. Here for Class column has 1 & 2 values, which considered to represent “positive” & “negative” respectably. We replaced “positive” value with 1 & “negative” value with 2.

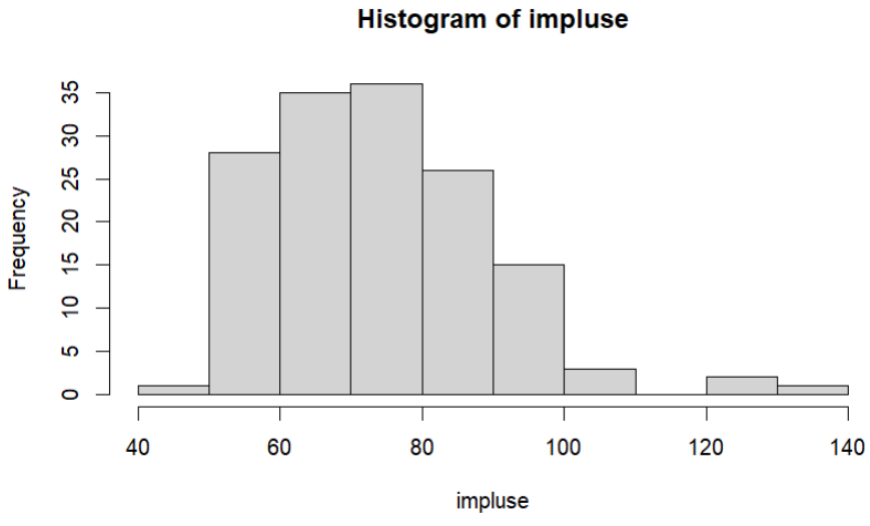
Histogram & BoxPlot:

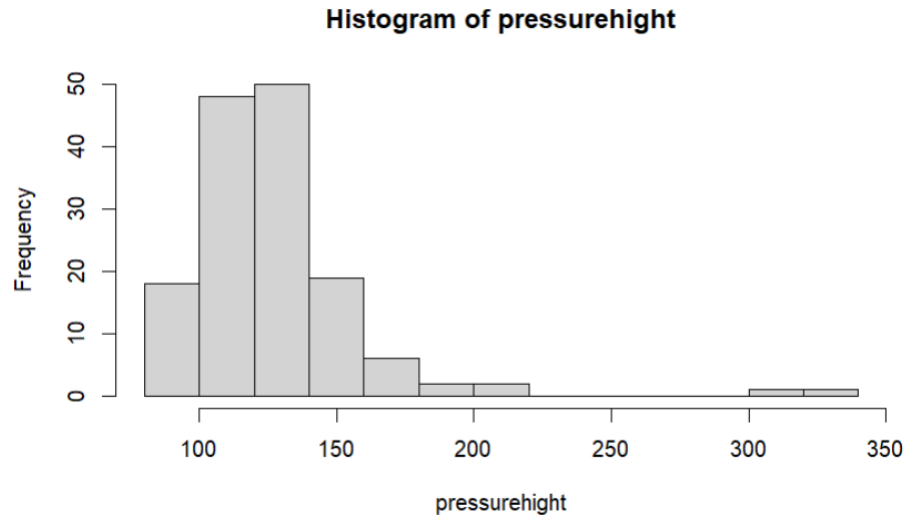
Historgeam for all numeric columns:

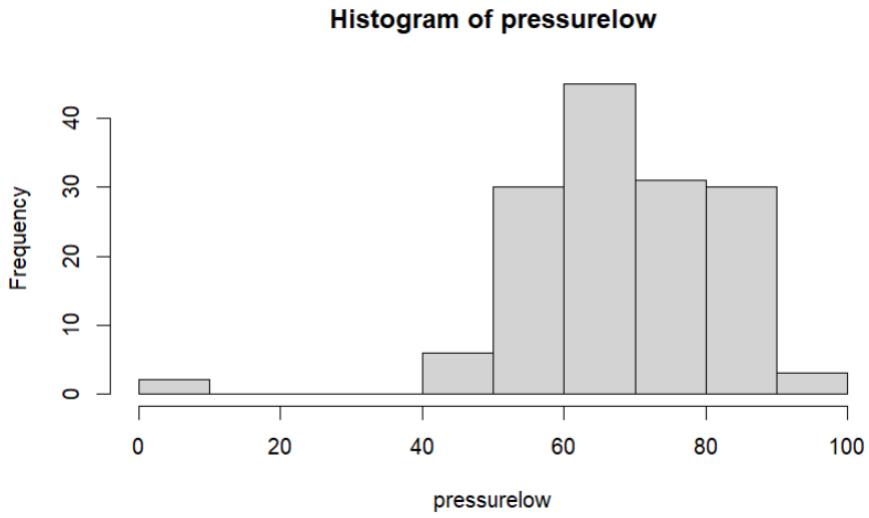
Code:

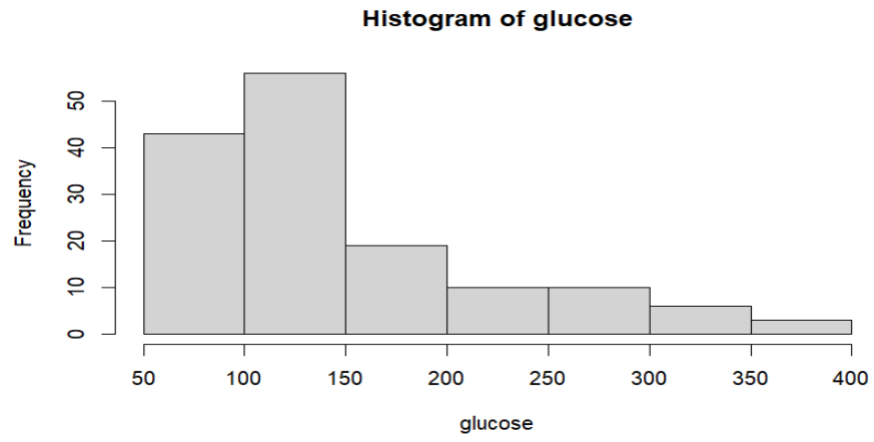
```
hist(age)
hist(impluse)
hist(pressurehight)
hist(pressurelow)
hist(gluecose)
```









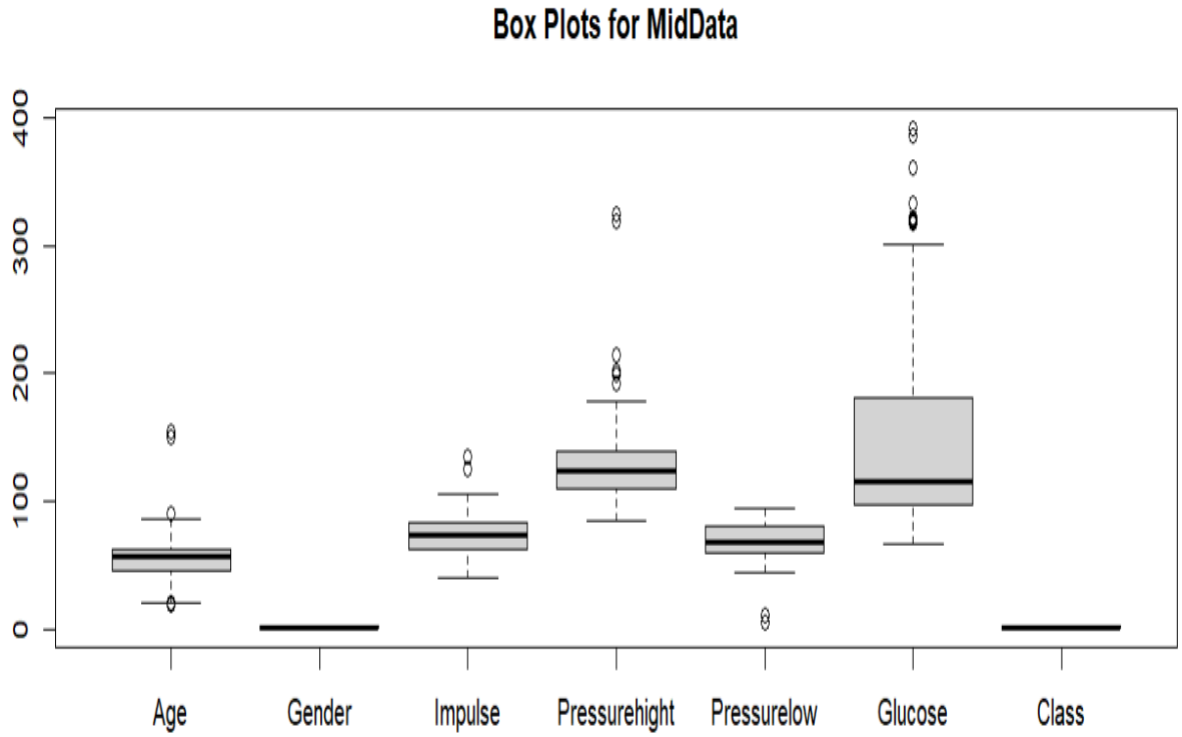


Description: The code hist() generates a histogram of the values in the columns, allowing us to see how the data is distributed.

Boxplot of MidData:

Code:

```
boxplot(MidData$age, MidData$gender ,MidData$impluse, MidData$pressurehight ,
    MidData$pressurelow ,MidData$glucose , MidData$class ,
    main="Box Plots for MidData",
    names=c("Age", "Gender","Impulse","Pressurehight","Pressurelow","Glucose","Class"))
```



Description: This technic is used for identify data points that are fall outside of the range of the most of the values. **boxplot** function is used for plotting boxes, **main** is used for providing name, **names=** is used for providing name each boxes.