AMERICAN INTERNATIONAL UNIVERSITY-BANGLADESH

Faculty of Science and Technology

Project Cover Page

Assignment Title:	MIDTERM PROJECT		
Assignment No:	01	Date of Submission:	18 March 2024
Course Title:	INTRODUCTION TO DATA SCIENCE		
Course Code:	CSC4180	Section:	C
Semester:	Spring 2023-24	Course Teacher:	TOHEDUL ISLAM

Declaration and Statement of Authorship:

1. I/we hold a copy of this Assignment/Case-Study, which can be produced if the original is lost/damaged.
2. This Assignment/Case-Study is my/our original work and no part of it has been copied from any other student's work or from any other source except where due acknowledgement is made.
3. No part of this Assignment/Case-Study has been written for me/us by any other person except where such collaborationhas been authorized by the concerned teacher and is clearly acknowledged in the assignment.
4. I/we have not previously submitted or currently submitting this work for any other course/unit.
5. This work may be reproduced, communicated, compared and archived for the purpose of detecting plagiarism.
6. I/we give permission for a copy of my/our marked work to be retained by the Faculty for review and comparison, including review by external examiners.
7. I/we understand thatPlagiarism is the presentation of the work, idea or creation of another person as though it is your own. It is a formofcheatingandisaveryseriousacademicoffencethatmayleadtoexpulsionfromtheUniversity. Plagiarized material can be drawn from, and presented in, written, graphic and visual form, including electronic data, and oral presentations. Plagiarism occurs when the origin of them arterial used is not appropriately cited.
8. I/we also understand that enabling plagiarism is the act of assisting or allowing another person to plagiarize or to copy my/our work.

* *Student(s) must complete all details except the faculty use part.*
** Please submit all assignments to your course teacher or the office of the concerned teacher.

Group Name/No.: 01

No	Name	ID	Program	Signature
1	KHONDOKER MD. SABIT HASAN	21-45306-2	BSc [CSE]	
2	MIRZA SADMAN MEHRAB	21-45001-2	BSc [CSE]	
3			Choose an item.	
4			Choose an item.	
5			Choose an item.	
6			Choose an item.	
7			Choose an item.	
8			Choose an item.	
9			Choose an item.	
10			Choose an item.	

Faculty use only		
FACULTYCOMMENTS	**Marks Obtained**	
	Total Marks	

Data Exploration

Dataset Description:

The data we will use has been modified from the Maternal Health Risk dataset, which was originally collected from the UC Irvine Machine Learning Repository. The dataset was obtained from a research paper titled "Review and Analysis of Risk Factors of Maternal Health in Remote Areas Using the Internet of Things (IoT)" by Marzia Ahmed, M.A. Kashem, Mostafijur Rahman, and S. Khatun, published in 2020. The data was collected from various hospitals, community clinics, and maternal health care institutions in rural areas of Bangladesh, using an IoT-based risk monitoring system. The dataset is relatively recent, having been published in 2020, and can be considered reliable as it was directly obtained from medical institutions. However, there are some missing and invalid values, as well as some outliers. The dataset consists of 9 attributes, including class attributes. Age, Systolic Blood Pressure (SystolicBP), Diastolic Blood Pressure (DiastolicBP), Blood Sugar (BS), Body Temperature (BodyTemp), and HeartRate are numeric attributes, while Infection and RiskLevel are categorical attributes. Smoking is also a categorical attribute, but with numeric labels. Our class attribute is "RiskLevel," as it is what we are trying to predict. The attributes are classified as follows:

Variable Name	Role	Type
Age	Feature	Numeric
Infection	Feature	Categorical
Smoking	Feature	Categorical with Numeric label
SystolicBP	Feature	Numeric
DiastolicBP	Feature	Numeric
BS	Feature	Numeric
BodyTemp	Feature	Numeric
HeartRate	Feature	Numeric
RiskLevel	Target	Categorical

Data Preparation

Visualize the dataset:

<u>Importing necessary library:</u>

```
library(naniar)
library(ggplot2)
```

```
Console   Terminal   Background Jobs
R 4.3.2 · ~/
> library(naniar)
> library(ggplot2)
```

Importing the dataset and viewing it:

```
mydata_unaltered <- read.csv("D:/AIUB/Spring-24/Data Science/Midterm/Project/Dataset_midterm_Section(
View(mydata_unaltered)
```

```
> mydata_unaltered <- read.csv("D:/AIUB/Spring-24/Data Science/Midterm/Project/Dataset_midterm_Section(C).cs
v", header = TRUE, sep = ",")
> View(mydata_unaltered)
```

Filter

	Age	Infection	Smoking	SystolicBP	DiastolicBP	BS	BodyTemp	HeartRate	RiskLevel
1	25	yes	1	130	80	15.00	98	86	high risk
2	35	yes	1	140	90	13.00	98	70	high risk
3	29	yes	1	90	70	8.00	100	80	high risk
4	30	yes	1	140	85	7.00	98	70	high risk
5	35	no	3	120	60	6.10	98	76	low risk
6	23	yes	1	140	80	7.01	98	70	high risk
7	23		2	130	70	7.01	98	78	mid risk
8	NA	yes	1	85	60	11.00	102	86	high risk
9	32	marginal	2	120	90	6.90	98	70	mid risk
10	42	yes	1	130	80	18.00	98	70	high risk
11	23	no	3	90	60	7.01	98	76	low risk
12	19	marginal	2	120	80	7.00	98	70	mid risk
13	25	no	3	110	89	7.01	98	77	low risk
14	20	marginal	NA	120	75	7.01	100	70	mid risk
15	48	marginal	2	120	80	11.00	98	88	mid risk
16	15	no	3	120	NA	7.01	98	70	low risk

Showing 1 to 16 of 200 entries, 9 total columns

Numeric missing values are represented as "NA", while categorical ones are left blank. To standardize the representation of missing values, "NA" is added to the blank spaces of categorical attributes.

```
mydata_unaltered[mydata_unaltered == ''] <- NA
```

```
> mydata_unaltered[mydata_unaltered == ''] <- NA
```

Printing the number of missing values and the rows with missing values:

```
sum(is.na(mydata_unaltered))
mydata_unaltered[rowSums(is.na(mydata_unaltered))>0,]
```

```
> sum(is.na(mydata_unaltered))
[1] 20
> mydata_unaltered[rowSums(is.na(mydata_unaltered))>0,]
    Age Infection Smoking SystolicBP DiastolicBP    BS BodyTemp HeartRate RiskLevel
7    23      <NA>       2        130          70  7.01       98        78  mid risk
8    NA       yes       1         85          60 11.00      102        86 high risk
14   20  marginal      NA        120          75  7.01      100        70  mid risk
16   15        no       3        120          NA  7.01       98        70  low risk
25   NA        no       3        120          80  7.50       98        76  low risk
27   19      <NA>       3        120          75  7.20       98        66  low risk
34   21        no      NA        120          80  7.10       98        77  low risk
39   45        no       3        120          NA  6.10       98        66  low risk
40   NA        no       3        100          70  6.10       98        66  low risk
59   23      <NA>       3         90          60  6.40       98        76  low risk
61   15        no      NA        120          80  7.20       98        70  low risk
65   NA  marginal       2        120          60  6.10       98        76  mid risk
69   20      <NA>       2        110          60  7.00      100        70  mid risk
79   35  marginal       2        120          NA  6.90       98        78  mid risk
101  NA  marginal       2        120          90  6.80       98        66  mid risk
103  48       yes      NA        140          NA 15.00       98        90 high risk
107  50      <NA>       1        140          90 15.00       98        90 high risk
153  17      <NA>       1        110          75 12.00      101        76 high risk
180  21      <NA>       3         75          50  6.10       98        70  low risk
```

Viewing the overall summary of the dataset to get to know with the dataset:

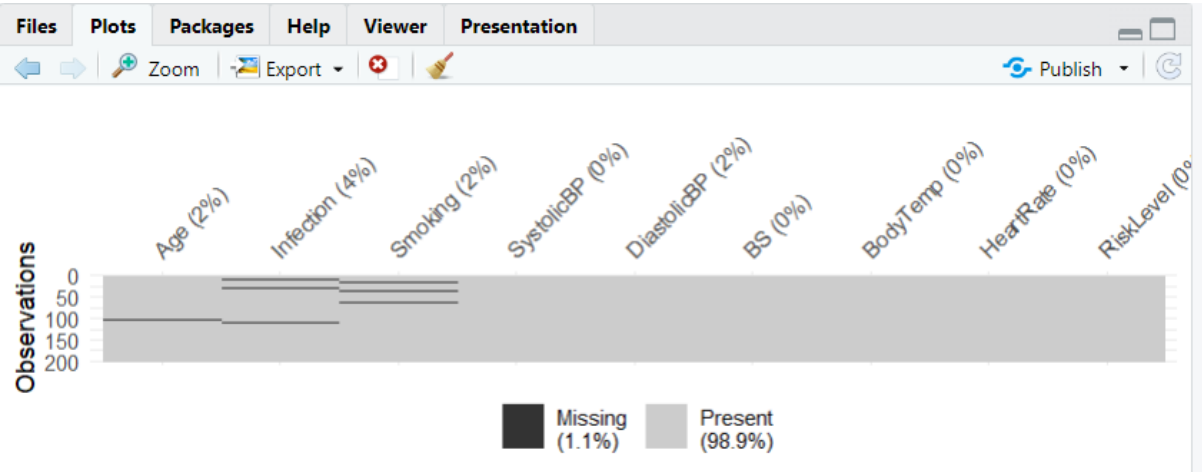
```
summary(mydata_unaltered)
```

```
> summary(mydata_unaltered)
      Age          Infection            Smoking        SystolicBP     DiastolicBP           BS
 Min.   : 10.00   Length:200         Min.   :1.000   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000
 1st Qu.: 21.00   Class :character   1st Qu.:1.000   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.875
 Median : 25.00   Mode  :character   Median :2.000   Median :120.0   Median : 80.00   Median : 7.150
 Mean   : 31.97                      Mean   :2.077   Mean   :114.8   Mean   : 78.32   Mean   : 8.831
 3rd Qu.: 40.00                      3rd Qu.:3.000   3rd Qu.:130.0   3rd Qu.: 90.00   3rd Qu.: 8.000
 Max.   :170.00                      Max.   :3.000   Max.   :160.0   Max.   :100.00   Max.   :19.000
 NA's   :5                           NA's   :4                       NA's   :4
    BodyTemp         HeartRate      RiskLevel
 Min.   :-160.00   Min.   :60.00   Length:200
 1st Qu.:  98.00   1st Qu.:70.00   Class :character
 Median :  98.00   Median :76.00   Mode  :character
 Mean   :  95.94   Mean   :74.89
 3rd Qu.:  98.00   3rd Qu.:80.00
 Max.   : 103.00   Max.   :90.00
```

There may be some outliers/invalid values present in the Age and BodyTemp attributes, as the Age has a max value of 170 and BodyTemp has a min value of -160. Additionally, missing values are present in the dataset.

Viewing the missing values on a graph:

```
vis_miss(mydata_unaltered)
```



We used the vis_miss() function from the Naniar library to visualize the missing data in the dataset. The function revealed that the dataset has an overall missingness of 1.1%. Specifically, Age has a missingness of 2% (5 NAs), Infection has a missingness of 4% (7 NAs), Smoking has a missingness of 2% (4 NAs), and DiastolicBP has a missingness of 2% (4 NAs). All other attributes in the dataset have no missing values.

Displaying the frequency counts of categorical attributes:

```
table(mydata_unaltered$Infection, exclude = NULL)
```

```
> table(mydata_unaltered$Infection, exclude = NULL)

marginal       no      yes    yesss      yoo     <NA>
      52       77       61        1        2        7
```

There are invalid values in the Infection attribute, including 'yesss' and 'yoo'. In addition, there are seven missing values represented as NA.

```
table(mydata_unaltered$RiskLevel, exclude = NULL)
```

```
> table(mydata_unaltered$RiskLevel, exclude = NULL)

high risk  low risk  mid risk
       65        81        54
```

There are no invalid values in the RiskLevel attribute.

Transform the label of the Smoking attribute to the categorical value as it is more sensible-

```
mydata_unaltered$Smoking <- factor(mydata_unaltered$Smoking, levels = c(1,2,3), labels = c("yes"
```

```
> mydata_unaltered$Smoking <- factor(mydata_unaltered$Smoking, levels = c(1,2,3), labels = c("yes","so
metimes","no"), exclude = NA)
```

Converting the Smoking attribute label to a categorical value is more appropriate.

Now, displaying the frequency counts of Smoking categorical attribute-

```
table(mydata_unaltered$Smoking, exclude = NULL)
```

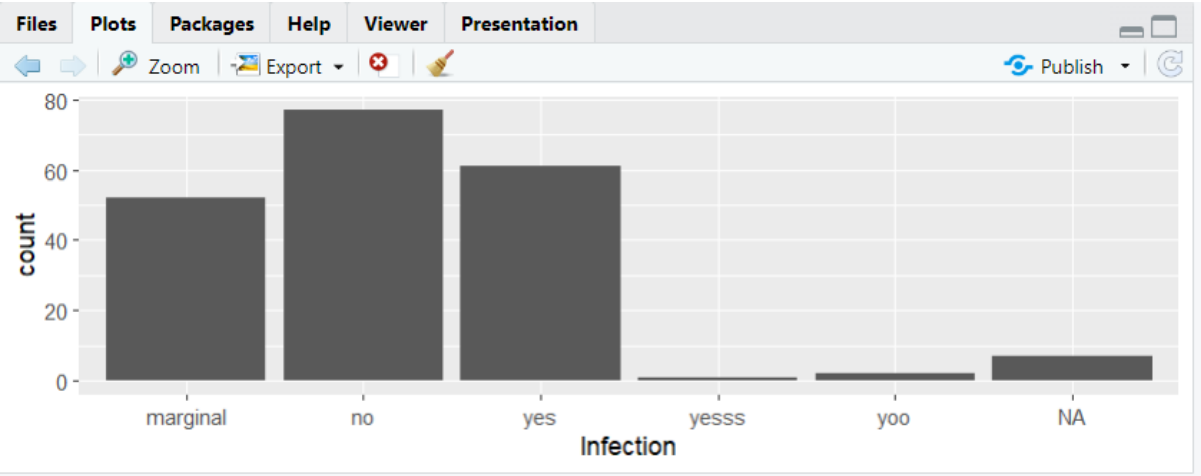
```
> table(mydata_unaltered$Smoking, exclude = NULL)

      yes sometimes        no      <NA>
       64        53        79         4
```

No invalid values are present in the Smoking attribute, but four missing values (NA) exist.

Plotting the categorical attributes on the graph:

```
ggplot(mydata_unaltered, aes(x=Infection)) + geom_bar()
```

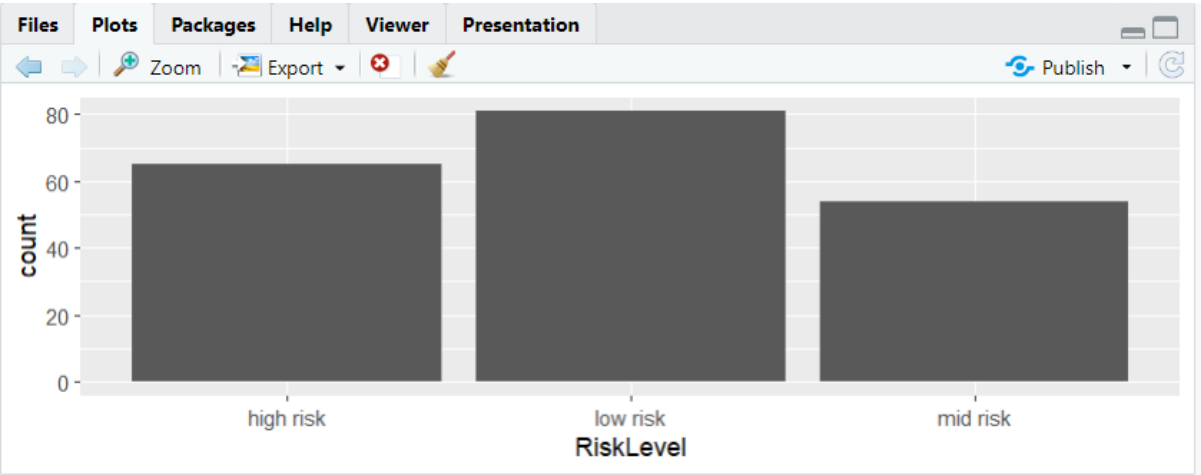
```
> ggplot(mydata_unaltered, aes(x=Infection)) + geom_bar()
```



There are invalid values in the Infection attribute, including 'yesss' and 'yoo'. In addition, some missing values are represented as NA.

```
ggplot(mydata_unaltered, aes(x=RiskLevel)) + geom_bar()
```

```
> ggplot(mydata_unaltered, aes(x=RiskLevel)) + geom_bar()
```



There are no invalid values in the RiskLevel attribute.

```
ggplot(mydata_unaltered, aes(x=Smoking)) + geom_bar()
```

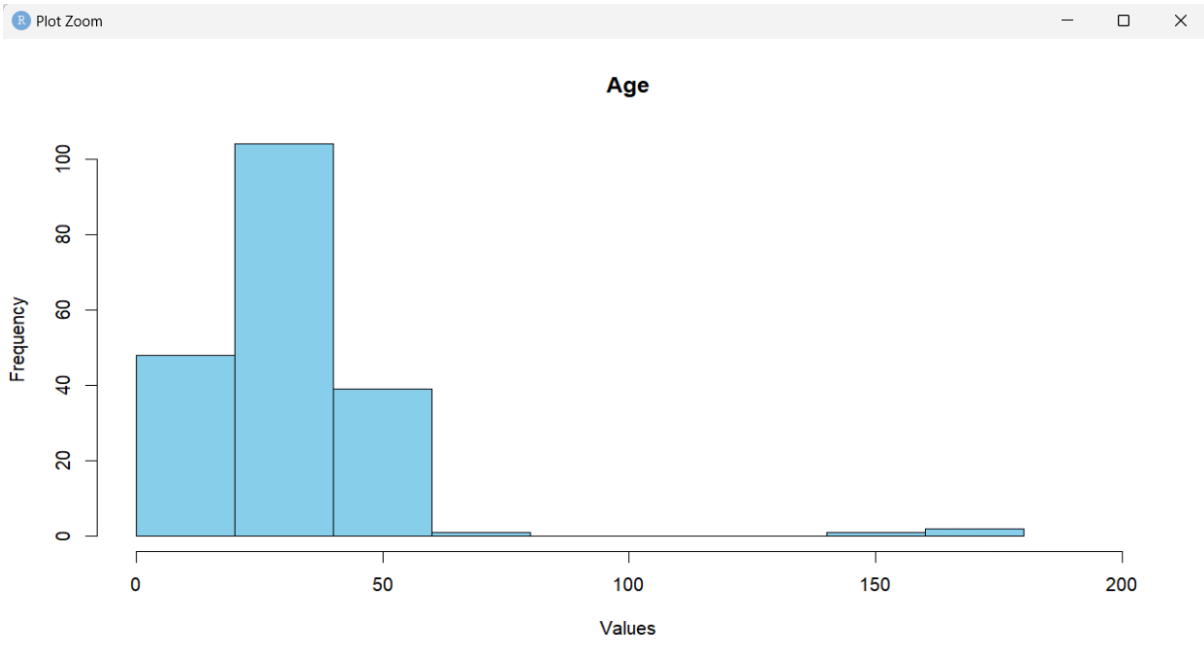
```
> ggplot(mydata_unaltered, aes(x=Smoking)) + geom_bar()
```

No invalid values are present in the Smoking attribute, but some missing values (NA) exist.

<u>Plotting the numeric attributes on the graph:</u>

```
hist(mydata_unaltered$Age, main = "Age", xlab = "Values", ylab = "Frequency", col = "skyblue", b
```

```
> hist(mydata_unaltered$Age, main = "Age", xlab = "Values", ylab = "Frequency", col = "skyblue", borde
r = "black", xlim = c(0,200))
```

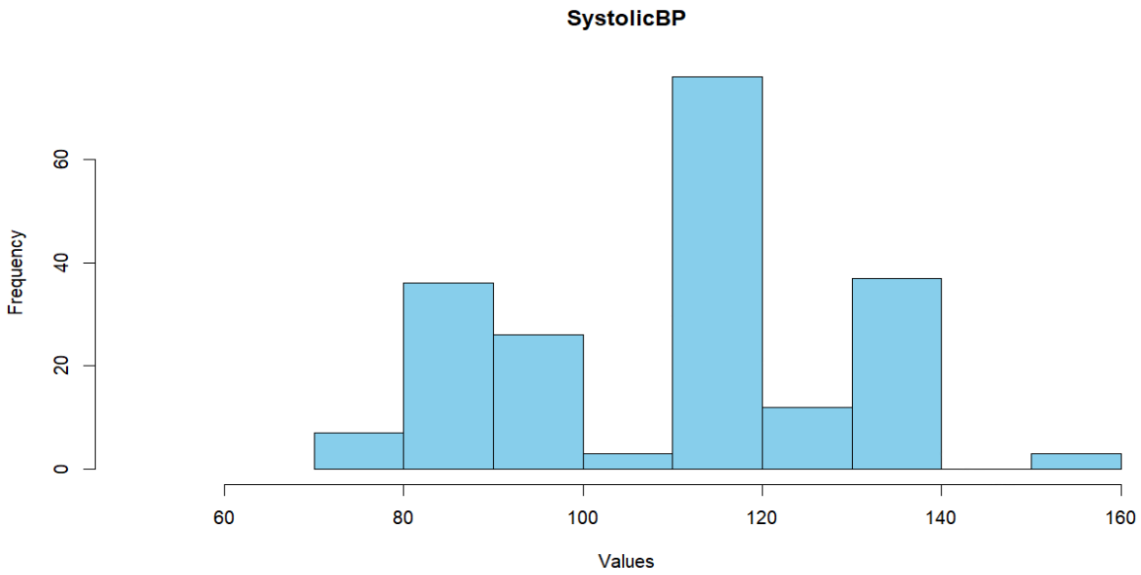


There appear to be some outliers in the Age attribute.

```
hist(mydata_unaltered$SystolicBP, main = "SystolicBP", xlab = "Values", ylab = "Frequency", col
```

```
> hist(mydata_unaltered$SystolicBP, main = "SystolicBP", xlab = "Values", ylab = "Frequency", col = "s
kyblue", border = "black", xlim = c(50,160))
```

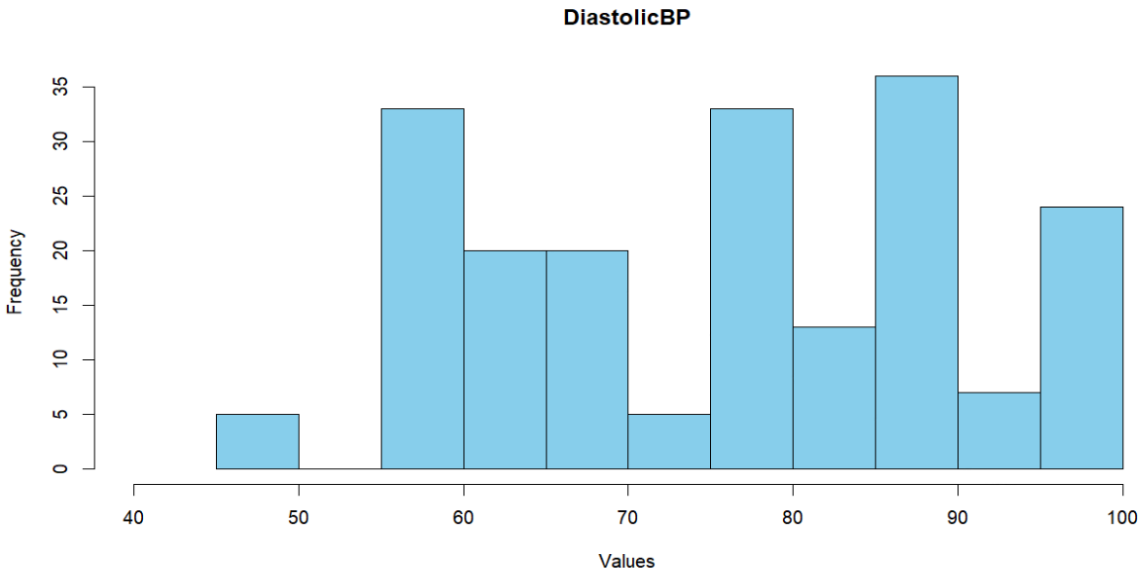
Plot Zoom




```
hist(mydata_unaltered$DiastolicBP, main = "DiastolicBP", xlab = "Values", ylab = "Frequency", co
```

```
> hist(mydata_unaltered$DiastolicBP, main = "DiastolicBP", xlab = "Values", ylab = "Frequency", col =
"skyblue", border = "black", xlim = c(40,100))
```

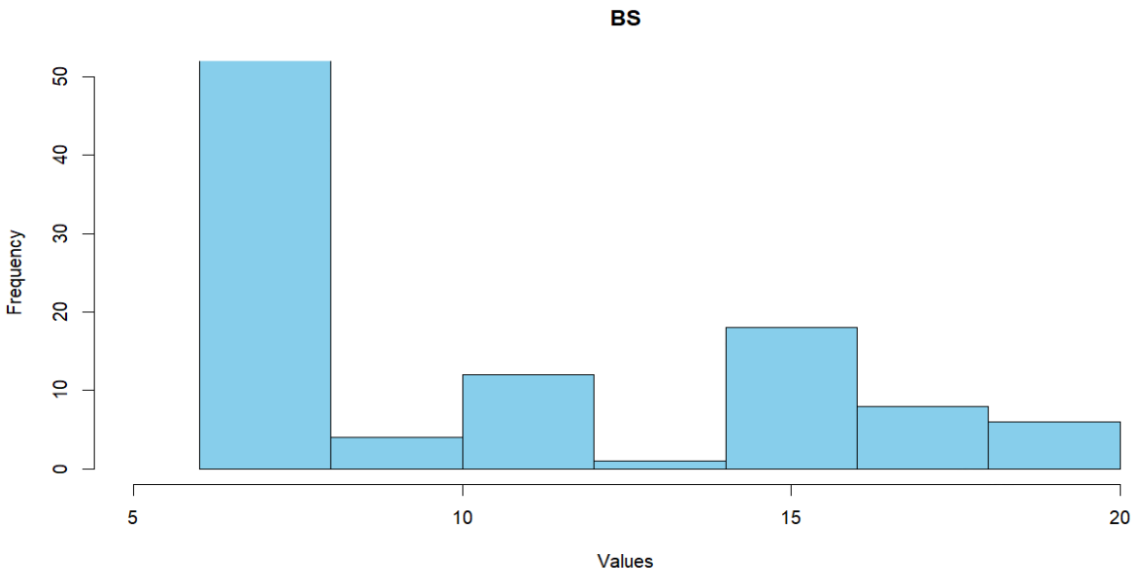
Plot Zoom



```
hist(mydata_unaltered$BS, main = "BS", xlab = "Values", ylab = "Frequency", col = "skyblue", bor
```

```
> hist(mydata_unaltered$BS, main = "BS", xlab = "Values", ylab = "Frequency", col = "skyblue", border
= "black", xlim = c(5,20), ylim = c(0,50))
```

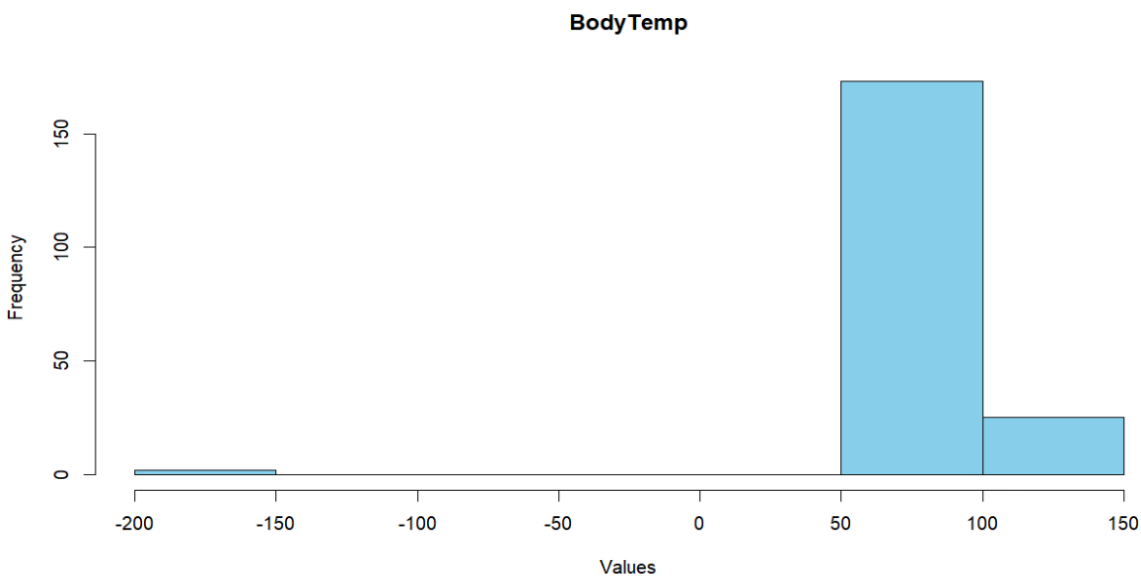
Plot Zoom




```
hist(mydata_unaltered$BodyTemp, main = "BodyTemp", xlab = "Values", ylab = "Frequency", col = "S
```

```
> hist(mydata_unaltered$BodyTemp, main = "BodyTemp", xlab = "Values", ylab = "Frequency", col = "skybl
ue", border = "black")
```

Plot Zoom

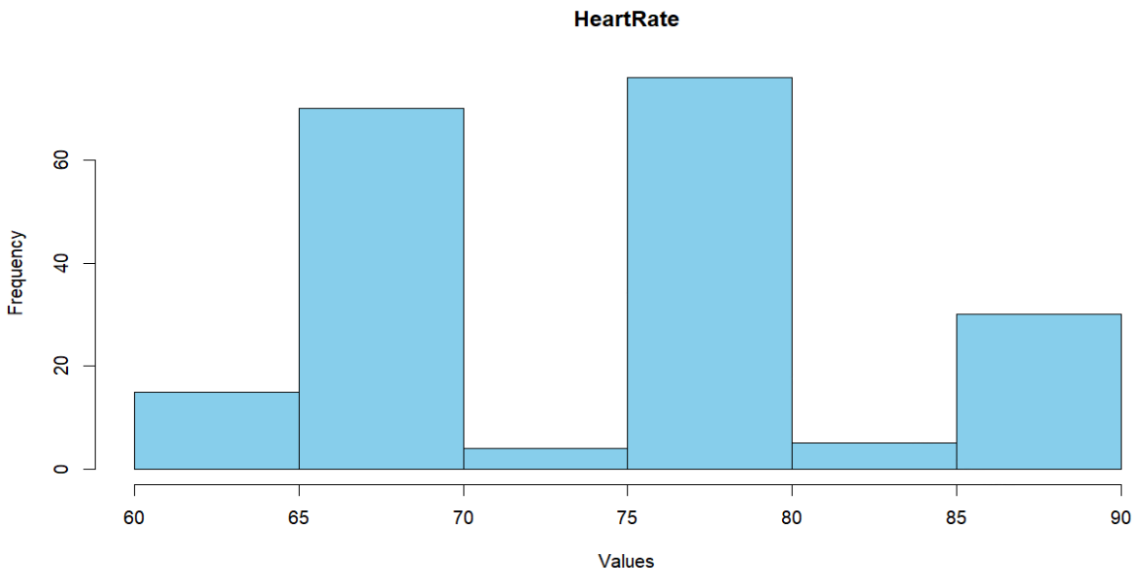


There are negative values present in BodyTemp, which may be considered outliers or invalid values.

```
hist(mydata_unaltered$HeartRate, main = "HeartRate", xlab = "Values", ylab = "Frequency", col =
```

```
> hist(mydata_unaltered$HeartRate, main = "HeartRate", xlab = "Values", ylab = "Frequency", col = "sky
blue", border = "black")
```

Plot Zoom

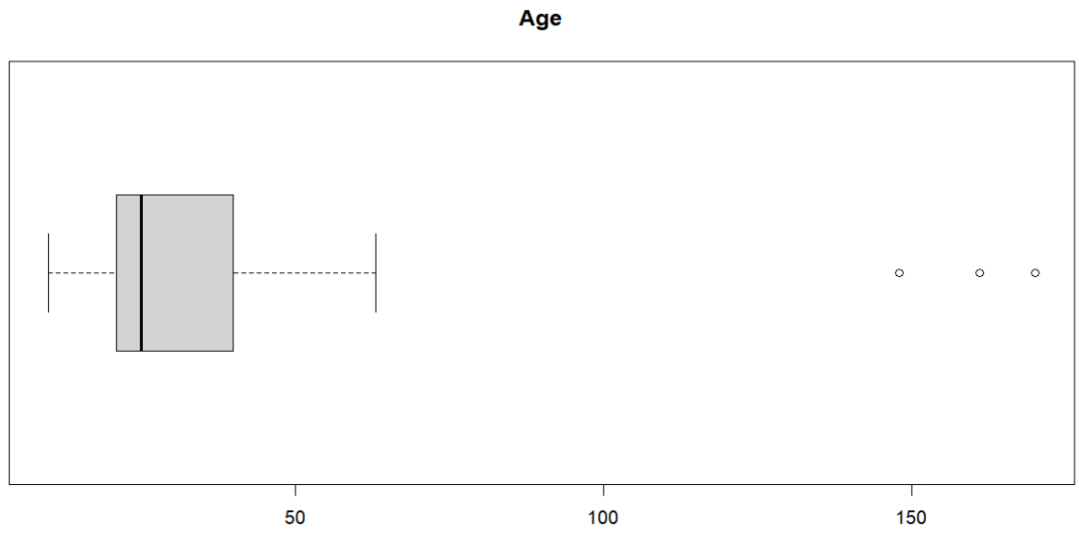


Now, plotting the boxplot to make sure there are any outliers present in the numeric attributes:

```
boxplot(mydata_unaltered$Age, main = "Age", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$Age, main = "Age", horizontal = TRUE)
```

Plot Zoom

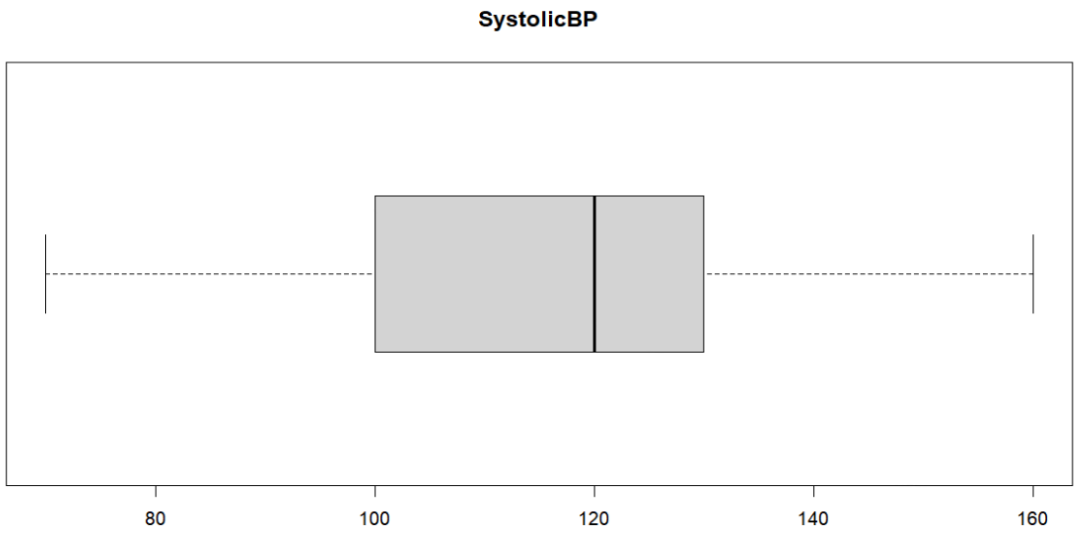


Here, we can see some outliers present in the Age attribute.

```
boxplot(mydata_unaltered$SystolicBP, main = "SystolicBP", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$SystolicBP, main = "SystolicBP", horizontal = TRUE)
```

Plot Zoom

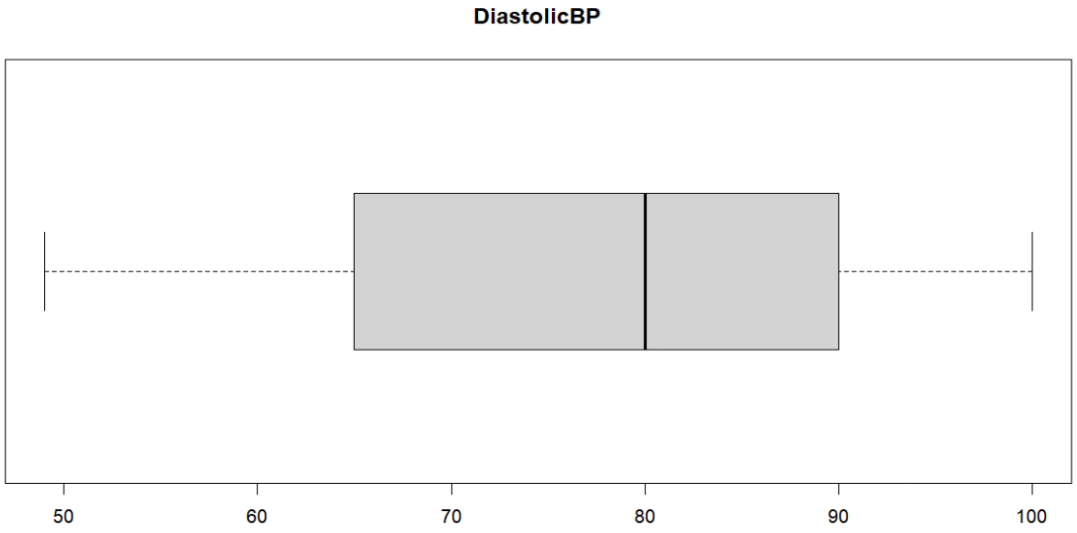


The SystolicBP attribute does not contain any outliers.

```
boxplot(mydata_unaltered$DiastolicBP, main = "DiastolicBP", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$DiastolicBP, main = "DiastolicBP", horizontal = TRUE)
```

Plot Zoom

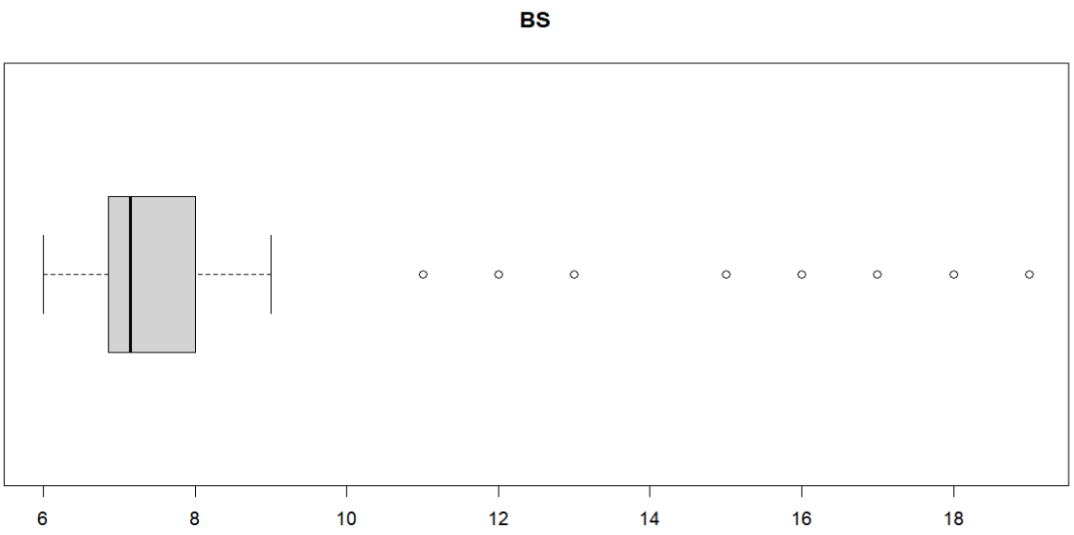


Also, the DiastolicBP attribute does not contain any outliers.

```r
boxplot(mydata_unaltered$BS, main = "BS", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$BS, main = "BS", horizontal = TRUE)
```

Plot Zoom




```r
summary(mydata_unaltered$BS)
```

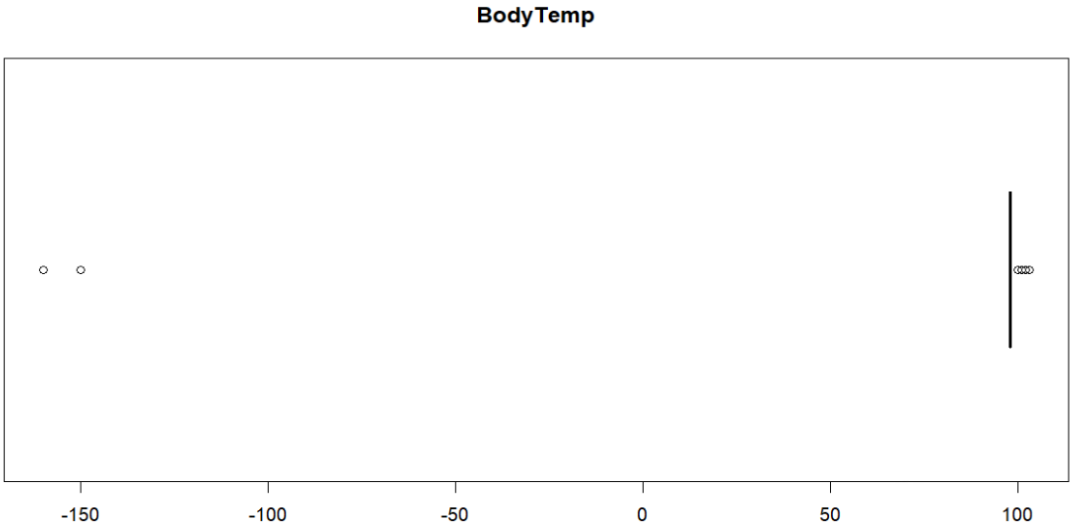
```
> summary(mydata_unaltered$BS)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  6.000   6.875   7.150   8.831   8.000  19.000 
```

The boxplot shows that there are some outliers present in the BS attribute. However, we can ignore them since the maximum value is 19 and the minimum value is 6, which are within the valid range of Blood Sugar.

```
boxplot(mydata_unaltered$BodyTemp, main = "BodyTemp", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$BodyTemp, main = "BodyTemp", horizontal = TRUE)
```

Plot Zoom

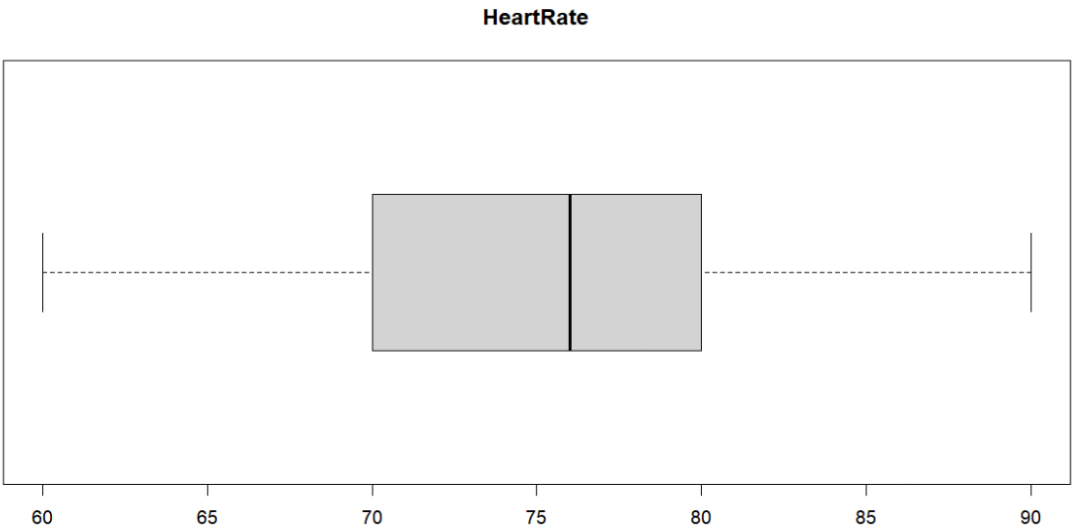


There are negative values present in BodyTemp, which may be considered outliers or invalid values.

```
boxplot(mydata_unaltered$HeartRate, main = "HeartRate", horizontal = TRUE)
```

```
> boxplot(mydata_unaltered$HeartRate, main = "HeartRate", horizontal = TRUE)
```

Plot Zoom



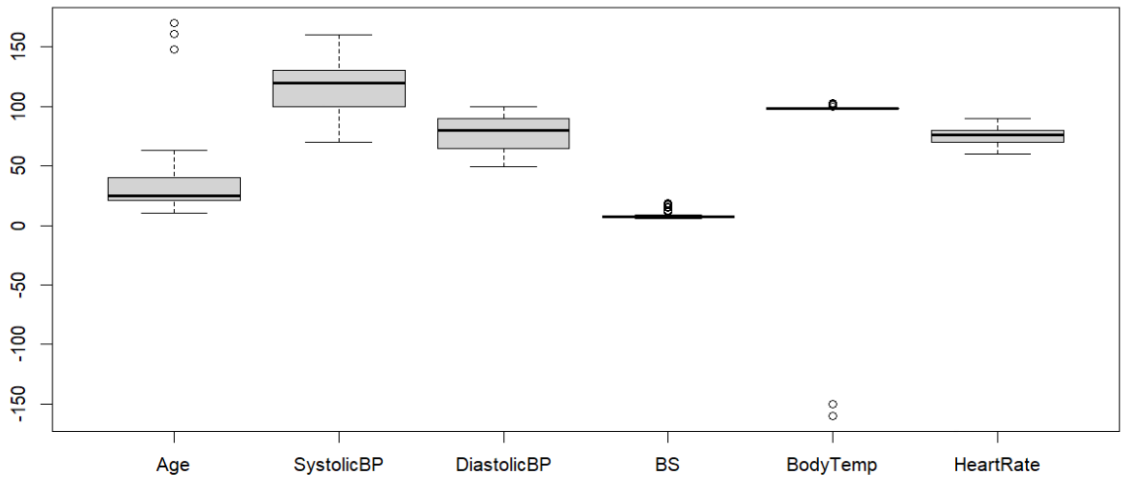
The HeartRate attribute does not contain any outliers.

Let's create a graph that displays boxplots for all the numeric attributes together-

```
mydata_boxplot <- mydata_unaltered[,c(1,4,5,6,7,8)]
boxplot(mydata_boxplot)
```

```
> mydata_boxplot <- mydata_unaltered[,c(1,4,5,6,7,8)]
> boxplot(mydata_boxplot)
```

Plot Zoom



Prepare the dataset:

<u>Dealing with the outliers/invalid values:</u>

For the Age attribute-

```
mydata_wo <- mydata_unaltered
qnt_Age <- quantile(mydata_wo$Age, probs = c(.25, .75), na.rm = T)
caps_Age <- quantile(mydata_wo$Age, probs = c(.05, .95), na.rm = T)
H_Age <- 1.5 * IQR(mydata_wo$Age, na.rm = T)
mydata_wo$Age[mydata_wo$Age < (qnt_Age[1] - H_Age)] <- caps_Age[1]
mydata_wo$Age[mydata_wo$Age > (qnt_Age[2] + H_Age)] <- caps_Age[2]
```

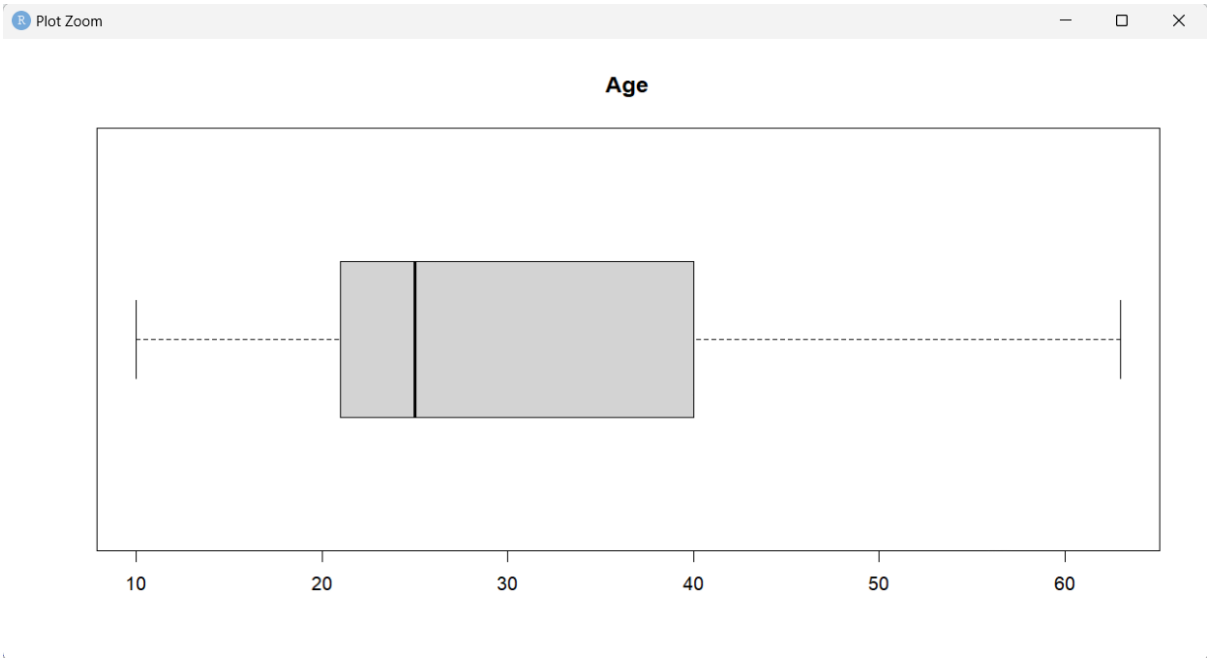
```
> mydata_wo <- mydata_unaltered
> qnt_Age <- quantile(mydata_wo$Age, probs = c(.25, .75), na.rm = T)
> caps_Age <- quantile(mydata_wo$Age, probs = c(.05, .95), na.rm = T)
> H_Age <- 1.5 * IQR(mydata_wo$Age, na.rm = T)
> mydata_wo$Age[mydata_wo$Age < (qnt_Age[1] - H_Age)] <- caps_Age[1]
> mydata_wo$Age[mydata_wo$Age > (qnt_Age[2] + H_Age)] <- caps_Age[2]
```

We have implemented a capping method to handle the outliers present in the 'Age' attribute. For values outside the T-shape whiskers of the boxplot, we have replaced those below the lower limit with the value of the 5th percentile and those above the upper limit with the value of the 95th percentile.

Here is the boxplot of the Age attribute after treating the outliers.

```
boxplot(mydata_wo$Age, main = "Age", horizontal = TRUE)
```

```
> boxplot(mydata_wo$Age, main = "Age", horizontal = TRUE)
```



For the BodyTemp attribute-

Since the max value here is 103, there are no problems with the positive values. However, the histogram (see above) indicates the presence of negative values in the data. This is problematic and needs to be addressed. To handle this issue, we would replace the negative values with the median value of the data.

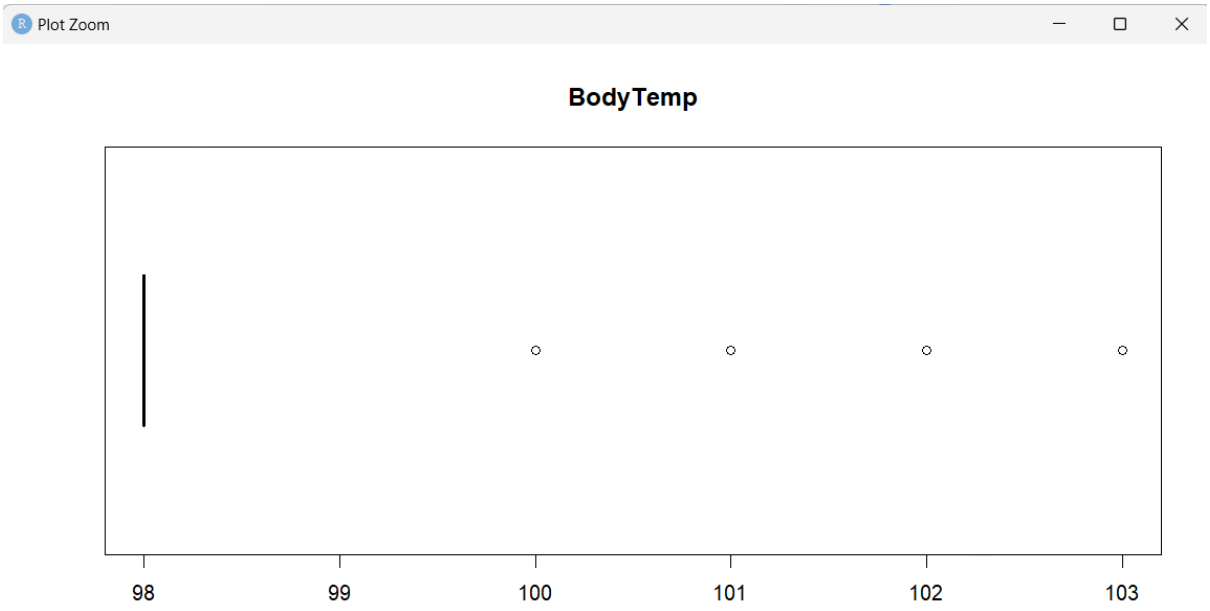
```
median_BodyTemp <- median(mydata_unaltered$BodyTemp, na.rm = T)
mydata_wo$BodyTemp[mydata_wo$BodyTemp < 0] <- median_BodyTemp
```

```
> median_BodyTemp <- median(mydata_unaltered$BodyTemp, na.rm = T)
> mydata_wo$BodyTemp[mydata_wo$BodyTemp < 0] <- median_BodyTemp
```

Here is the boxplot of the BodyTemp attribute after treating the outliers.

```
boxplot(mydata_wo$BodyTemp, main = "BodyTemp", horizontal = T)
```

```
> boxplot(mydata_wo$BodyTemp, main = "BodyTemp", horizontal = T)
```




```
summary(mydata_wo$BodyTemp)
```

```
> summary(mydata_wo$BodyTemp)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  98.00   98.00   98.00   98.47   98.00  103.00 
```

Although the boxplot shows some outliers, we can safely ignore them as the minimum and maximum values fall within the valid range of BodyTemp (98 to 103).

For the Infection attribute-

Some invalid values exist in the 'Infection' attribute, specifically 'yesss' and 'yoo'.

```
mydata_wo$Infection[mydata_wo$Infection %in% c("yesss", "yoo")] <- "yes"
table(mydata_wo$Infection, exclude = NULL)
```

```
> mydata_wo$Infection[mydata_wo$Infection %in% c("yesss", "yoo")] <- "yes"
> table(mydata_wo$Infection, exclude = NULL)

marginal       no      yes     <NA> 
      52       77       64        7 
```

Considering ‘yesss’ and ‘yoo’ were misspelled, we replaced their values with ‘yes’.

Dealing with the missing values:

```
mydata_wom_1 <- mydata_wo
mydata_wom_1 <- na.omit(mydata_wom_1)
```

```
> mydata_wom_1 <- mydata_wo
> mydata_wom_1 <- na.omit(mydata_wom_1)
```

To handle missing values in a dataset, we have the option of removing all the instances with NA values using na.omit(). However, this may not be convenient as those instances could have been necessary for the analysis. Instead, we can replace the missing values with the mean, median, or mode of the corresponding column/attribute. This approach will ensure that the dataset remains complete while minimizing the impact of missing values on the analysis.

For the Age attribute-

There are 5 NA (missing) values present in the Age attribute.

```
mydata_wom$Age[is.na(mydata_wom$Age)] <- median(mydata_wom$Age, na.rm = T)
```

```
> mydata_wom$Age[is.na(mydata_wom$Age)] <- median(mydata_wom$Age, na.rm = T)
```

Since the Age attribute is continuous, missing values are replaced with the median instead of the mean, which would return a floating value.

For the DiastolicBP attribute-

There are 4 NA (missing) values present in the DiastolicBP attribute.

```
mydata_wom$DiastolicBP[is.na(mydata_wom$DiastolicBP)] <- median(mydata_wom$DiastolicBP, na.rm =
```

```
> mydata_wom$DiastolicBP[is.na(mydata_wom$DiastolicBP)] <- median(mydata_wom$DiastolicBP, na.rm = T)
```

Since the DiastolicBP attribute is continuous, missing values are replaced with the median instead of the mean, which would return a floating value.

For the Infection attribute-

There are 7 NA (missing) values present in the Infection attribute.

```
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
mydata_wom$Infection[is.na(mydata_wom$Infection)] <- getmode(mydata_wom$Infection)
```

```
> mydata_wom$Infection[is.na(mydata_wom$Infection)] <- getmode(mydata_wom$Infection)
```

For the Infection attribute, missing values are replaced with the mode since it is categorical.

For the Smoking attribute-

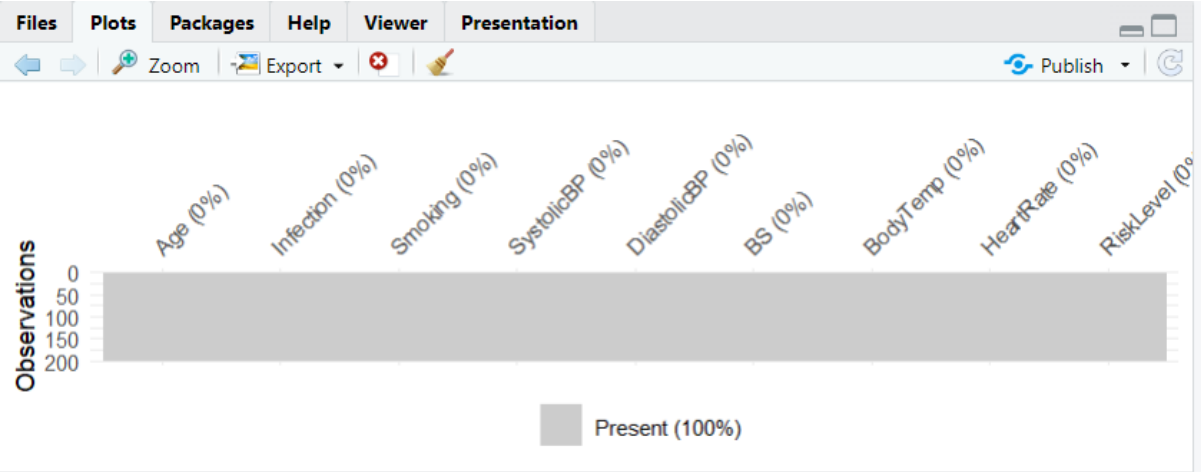
There are 4 NA (missing) values present in the Infection attribute.

```
mydata_wom$Smoking[is.na(mydata_wom$Smoking)] <- getmode(mydata_wom$Smoking)
```

```
> mydata_wom$Smoking[is.na(mydata_wom$Smoking)] <- getmode(mydata_wom$Smoking)
```

For the Smoking attribute, missing values are replaced with the mode since it is categorical.

After dealing with all the missing values, here is the missingness graph-



The dataset has no missing values, indicated by a present value of 100%.

<u>Let's view the overall summary of the final dataset</u>:

```
> summary(mydata_final)
      Age           Infection        Smoking      SystolicBP     DiastolicBP           BS
 Min.   :10.00   marginal:52   yes       :64   Min.   : 70.0   Min.   : 49.00   Min.   : 6.000
 1st Qu.:21.00   no       :84   sometimes:53   1st Qu.:100.0   1st Qu.: 65.00   1st Qu.: 6.875
 Median :25.00   yes      :64   no        :83   Median :120.0   Median : 80.00   Median : 7.150
 Mean   :30.25                                  Mean   :114.8   Mean   : 78.35   Mean   : 8.831
 3rd Qu.:39.25                                  3rd Qu.:130.0   3rd Qu.: 90.00   3rd Qu.: 8.000
 Max.   :63.00                                  Max.   :160.0   Max.   :100.00   Max.   :19.000
    BodyTemp        HeartRate         RiskLevel
 Min.   : 98.00   Min.   :60.00   high risk:65
 1st Qu.: 98.00   1st Qu.:70.00   low risk :81
 Median : 98.00   Median :76.00   mid risk :54
 Mean   : 98.47   Mean   :74.89
 3rd Qu.: 98.00   3rd Qu.:80.00
 Max.   :103.00   Max.   :90.00
```

In conclusion, we can say that this dataset is complete and ready for future work with no missing, invalid, or outlier values.