AMERICAN INTERNATIONAL UNIVERSITY–BANGLADESH (AIUB)

FACULTY OF SCIENCE AND TECHNOLOGY

MIDTERM ASSIGNMENT

INTRODUCTION TO DATA SCIENCE

Spring 2022-2023

Section: D

Submitted By

AL SHAKIB E ELAHI

ID: 20-43665-2

Department: CSE

Supervised By

Tohedul Islam

Assistant Professor

Department of Computer Science

Date of Submission: **March 12, 2023**

Data set import:

```
library(dplyr)
library(readxl)
dataframe<-read_excel("D:/Codes/Data_Science/Datasets/Dataset_midterm.xlsx")
summary(dataframe)
```

Output:

```
> library(dplyr)
> library(readxl)
> dataframe<-read_excel("D:/Codes/Data_Science/Datasets/Dataset_midterm.xlsx")
> summary(dataframe)
       id              Age          weight(kg)     Delivery_number Delivery_time
 Min.   : 1.00   Min.   :18.00   Min.   : 49.00   Min.   :1.000   Min.   :0.0000
 1st Qu.:20.75   1st Qu.:25.00   1st Qu.: 61.00   1st Qu.:1.000   1st Qu.:0.0000
 Median :40.50   Median :28.00   Median : 63.50   Median :1.500   Median :0.0000
 Mean   :40.50   Mean   :29.68   Mean   : 65.13   Mean   :1.679   Mean   :0.6234
 3rd Qu.:60.25   3rd Qu.:32.00   3rd Qu.: 68.00   3rd Qu.:2.000   3rd Qu.:1.0000
 Max.   :80.00   Max.   :95.00   Max.   :110.00   Max.   :4.000   Max.   :2.0000
                 NA's   :3       NA's   :3        NA's   :2       NA's   :3
    Blood               Heart          Caesarian
 Length:80          Min.   :0.000   Min.   :0.0000
 Class :character   1st Qu.:0.000   1st Qu.:0.0000
 Mode  :character   Median :0.000   Median :1.0000
                    Mean   :0.375   Mean   :0.5641
                    3rd Qu.:1.000   3rd Qu.:1.0000
                    Max.   :1.000   Max.   :1.0000
                                    NA's   :2
```

Data Pre-processing:

Age Section

From summary, it is observed that "Age" section contains 3 null values.

- **Detecting null values' row number:**

```
> which(is.na(dataframe$Age))
[1] 50 62 78
```

- **Measure of Center Tendency:**

```
mn_age<-mean(dataframe$Age, na.rm = TRUE)
mdian_age<-median(dataframe$Age, na.rm = TRUE)
md_tage <- table(dataframe$Age)
md_age<-names(which.max(md_tage))
```

Output:

```
> print(paste("Mean: ",mn_age))
[1] "Mean:  29.6753246753247"
> print(paste("Median: ",mdian_age))
[1] "Median:  28"
> print(paste("Mode: ",md_age))
[1] "Mode:  26"
```

Explanation:

- ❖ na.rm == TRUE is used to skip the null (N/A) values
- ❖ table() provides the number of frequencies for a certain Age
- ❖ which.max() gives the max frequency value and associated Age
- ❖ names() extracts the maximum frequency
- ❖ paste() function mainly combines the string and numeric data type for the output.

Here mean is slightly greater than median and the mode.

- ➢ **Replacing null values with median group by Delivery_number:**

```
dataframe<-dataframe %>%
  group_by(Delivery_number) %>%
  mutate(Age = ifelse(is.na(Age), floor(median(Age, na.rm = TRUE)) , Age)) %>%
  ungroup()

summary(dataframe$Age)
```

Output:

```
> dataframe<-dataframe %>%
+   group_by(Delivery_number) %>%
+   mutate(Age = ifelse(is.na(Age), floor(median(Age, na.rm = TRUE)) , Age)) %>%
+   ungroup()
> summary(dataframe$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   25.00   27.00   29.56   32.00   95.00
```

Explanation:

- ❖ group_by() creates subsets according to each unique value of Delivery_number
- ❖ ungroup() removes the group created by group_by()
- ❖ ifelse() is to allow conditional operations on dataframe
- ❖ mutate() if for modifying existing dataframe information with the median

In my observation, it is obvious that Age is slightly dependent on number of Deliveries.

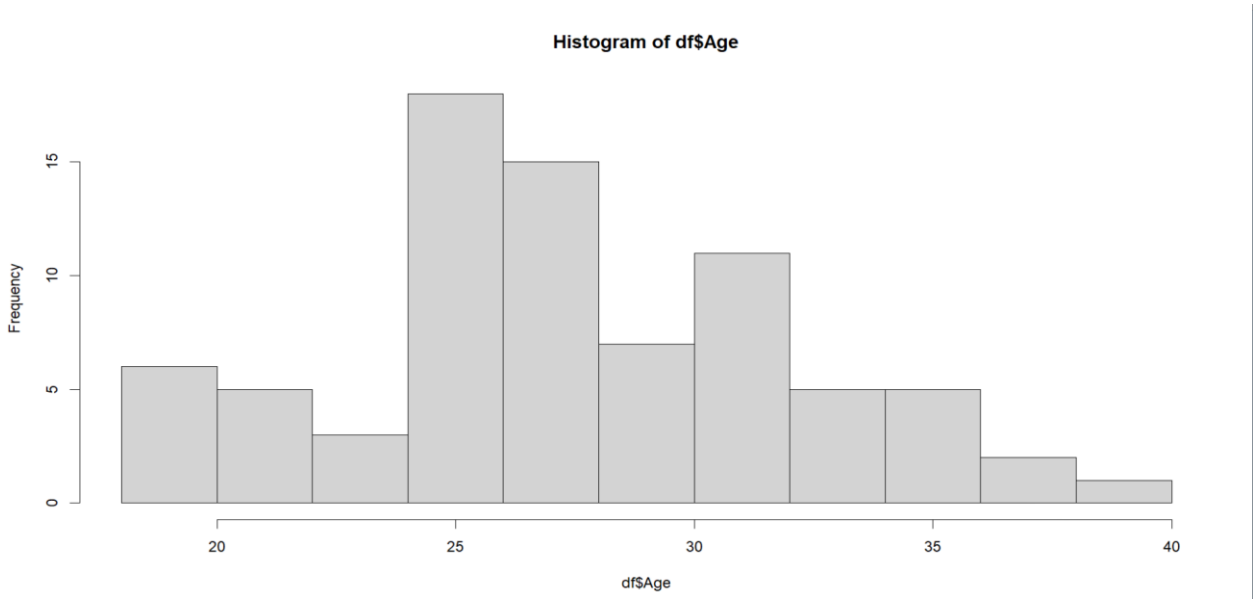
- ➢ **Outliers Detection:**

```
26 sds_age<-sd(dataframe$Age, na.rm = TRUE)
27 sds_age
28 dataframe$zScore <- ((dataframe$Age-mn_age)/sds_age)
29
30 minZ<-min(dataframe$zScore)
31 maxZ<-max(dataframe$zScore)
32 print(paste("Minimum Z-Score: ",minZ))
33 print(paste("Maximum Z-Score: ",maxZ))
34
35 df <- subset(dataframe, dataframe$zScore<=abs(minZ))
36 hist(df$Age)
37 summary(df$Age)
38
39 df$zScore<-NULL
```

Output:

```
> sds_age<-sd(dataframe$Age, na.rm = TRUE)
> sds_age
[1] 11.18752
> dataframe$zScore <- ((dataframe$Age-mn_age)/sds_age)
> minZ<-min(dataframe$zScore)
> maxZ<-max(dataframe$zScore)
> print(paste("Minimum Z-Score: ",minZ))
[1] "Minimum Z-Score:  -1.04360258495259"
> print(paste("Maximum Z-Score: ",maxZ))
[1] "Maximum Z-Score:  5.83906674339434"
> df <- subset(dataframe, dataframe$zScore<=abs(minZ))
> hist(df$Age)
> summary(df$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   25.00   27.00   27.95   32.00   40.00
> df$zScore<-NULL
```

Graph:



Explanation:

First, standard deviation of Age is calculated. Then Z-Score is determined. It is obvious that, for a normal distribution, maximum and minimum value for Z-Score should be almost similar. But in this case, maximum is far away and that's mean, there is a outlier. So with subset() function, a subset of dataframe was taken with threshold of absolute value of the minimum value of the Z-Score. Lastly, to eliminate zScore column, it has been assigned NULL.

Weight Section

From summary, it is observed that "Weight" section also contains 3 null values.

- **Detecting null values' row number:**

```
> which(is.na(df$`weight(kg)`))
[1] 47 50 61
```

- **Measure of Center Tendency:**

```
46  mnw<-mean(df$`weight(kg)`, na.rm = TRUE)
47  mdianw<-median(df$`weight(kg)`, na.rm = TRUE)
48  md_tw <- table(df$`weight(kg)`)
49  mdw<-names(which.max(md_tw))
50
51  mnw;mdianw;mdw
```

Output:

```
> mnw<-mean(df$`weight(kg)`, na.rm = TRUE)
> mdianw<-median(df$`weight(kg)`, na.rm = TRUE)
> md_tw <- table(df$`weight(kg)`)
> mdw<-names(which.max(md_tw))
> mnw;mdianw;mdw
[1] 63.99733
[1] 63
[1] "63"
```

Here mean, median and mode are almost equal.

- **Replacing null values with median:**

```
53 df<-df %>%
54   mutate(`weight(kg)` = ifelse(is.na(`weight(kg)`), median(`weight(kg)`, na.rm = TRUE) , `weight(kg)`))
55
56 summary(df$`weight(kg)`)
```

Output:

```
> summary(df$`weight(kg)`)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  49.00   61.12   63.00   63.96   67.50   82.00
```

Explanation:

- ifelse() is to allow conditional operations on dataframe
- mutate() if for modifying existing dataframe information with the median

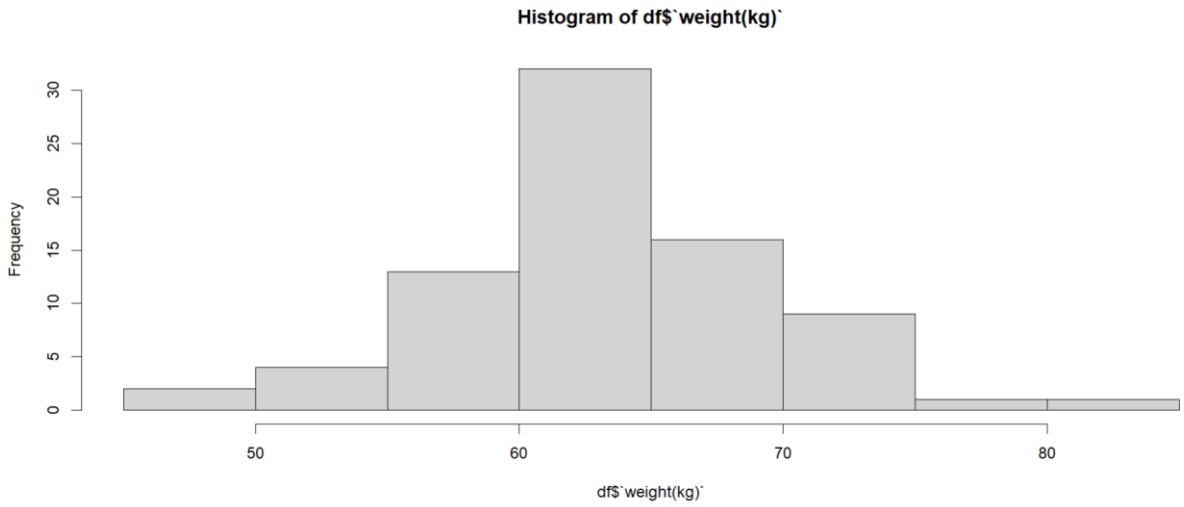
- **Outliers Detection:**

```
65 sdsw<-sd(df$`weight(kg)`, na.rm = TRUE)
66 sdsw
67 df$zScore <- ((df$`weight(kg)`-mnw)/sdsw)
68
69 minZW<-min(df$zScore)
70 maxZW<-max(df$zScore)
71 print(paste("Minimum Z-Score: ",minZW))
72 print(paste("Maximum Z-Score: ",maxZW))
73
74 hist(df$`weight(kg)`)
```

Output:

```
> sdsw
[1] 6.307148
> df$zScore <- ((df$`weight(kg)`-mnw)/sdsw)
> minZW<-min(df$zScore)
> maxZW<-max(df$zScore)
> print(paste("Minimum Z-Score: ",minZW))
[1] "Minimum Z-Score:  -2.37783128560653"
> print(paste("Maximum Z-Score: ",maxZW))
[1] "Maximum Z-Score:  2.85432770432604"
> View(df)
> boxplot(df$`weight(kg)`)
> hist(df$`weight(kg)`)
```

Graph:



Explanation:
Here, difference between the maximum and minimum Z-Score is quite low and so there is no outliers according to my observation.

Delivery Number

From summary, it is observed that “Delivery_number” section also contains 2 null values.

- **Detecting null values’ row number:**

```
> which(is.na(df$Delivery_number))
[1] 24 26
```

- **Measure of Center Tendency:**

```
83  mndn<-mean(df$Delivery_number, na.rm = TRUE)
84  mdiandn<-median(df$Delivery_number, na.rm = TRUE)
85  md_tdn <- table(df$Delivery_number)
86  mddn<-names(which.max(md_tdn))
87
88  mndn;mdiandn;mddn
```

Output:

```
> mndn;mdiandn;mddn
[1] 1.697368
[1] 2
[1] "1"
```

Here mean, median and mode are in a certain range.

- **Replacing null values with mode:**

```
90  df<-df %>%
91    mutate(Delivery_number = ifelse(is.na(Delivery_number), as.integer(mddn) , Delivery_number))
92
93  summary(df$Delivery_number)
94  hist(df$Delivery_number, breaks = c(0,1,2,3,4))
```

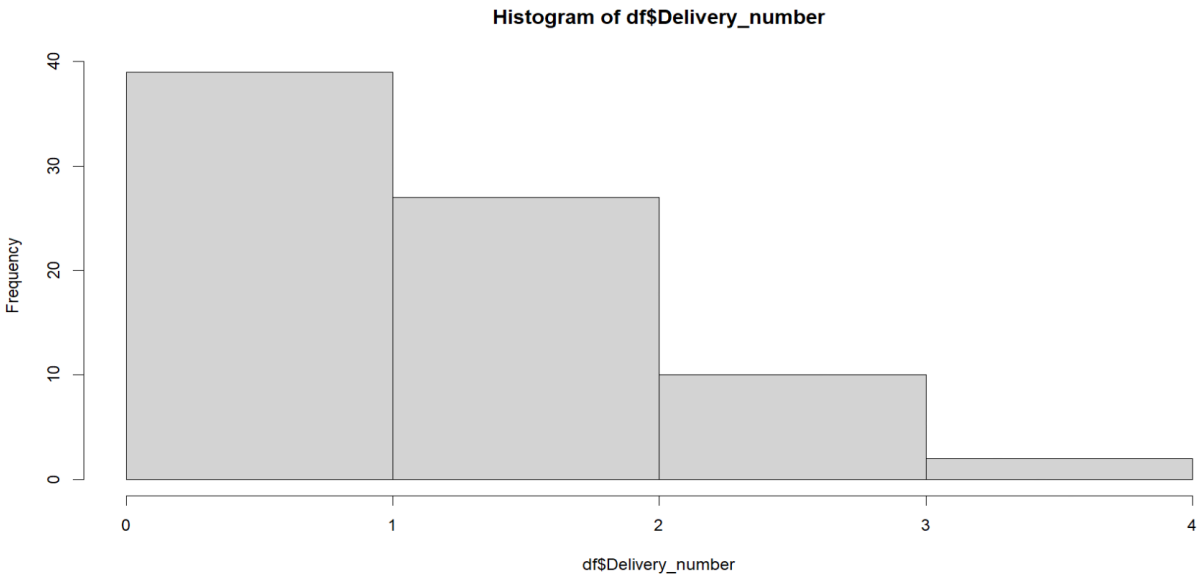
Output:

```
> df<-df %>%
+   mutate(Delivery_number = ifelse(is.na(Delivery_number), as.integer(mddn) , Delivery_number))
> summary(df$Delivery_number)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.500   1.679   2.000   4.000
> hist(df$Delivery_number, breaks = c(0,1,2,3,4))
```

Explanation:

- as.interger() is used to change the data-type of “mddn” from character to integer as Delivery_number is numeric.
- In hist() breaks is used to personalized.

Graph:



Delivery Time

- **Detecting null values' row number:**

```
> which(is.na(df$Delivery_time))
[1] 15 24 27
```

- **Measure of Center Tendency:**

```
> mnt; mdiant;mdt
[1] 0.64
[1] 0
[1] "0"
```

- **Replacing null values with mode:**

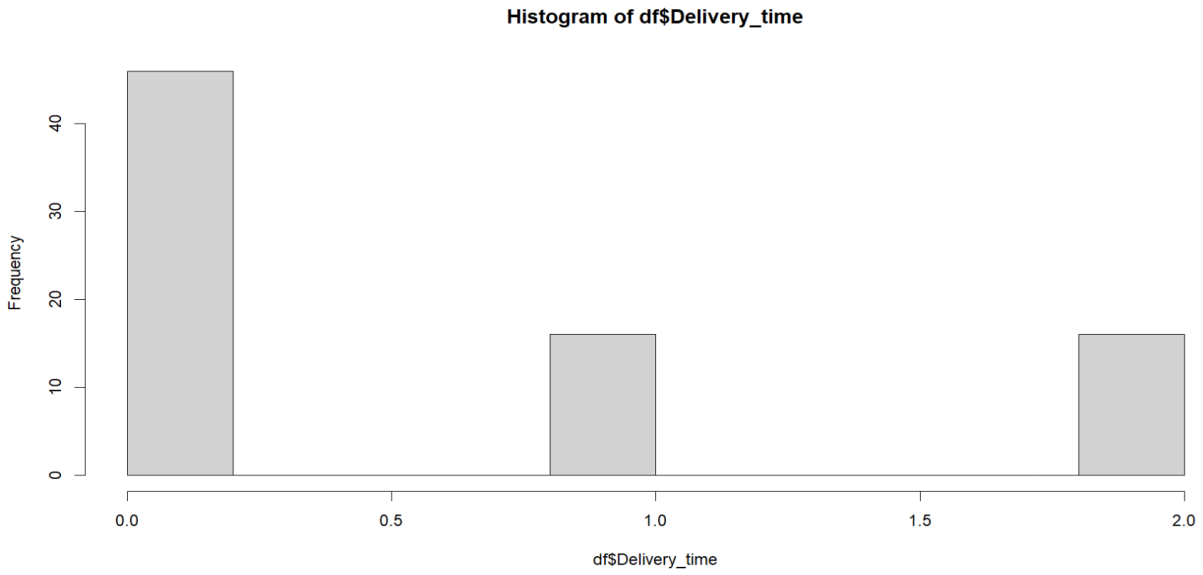
```
df<-df %>%
  mutate(Delivery_time = ifelse(is.na(Delivery_time), as.integer(mdt), Delivery_time))

summary(df$Delivery_time)
```

Output:

```
> summary(df$Delivery_time)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 0.0000  0.0000  0.0000  0.6154  1.0000  2.0000
```

Graph:



Blood

- **Detecting null values' row number:**

```
> which(is.na(df$Blood))
[1]  9 15 70
```

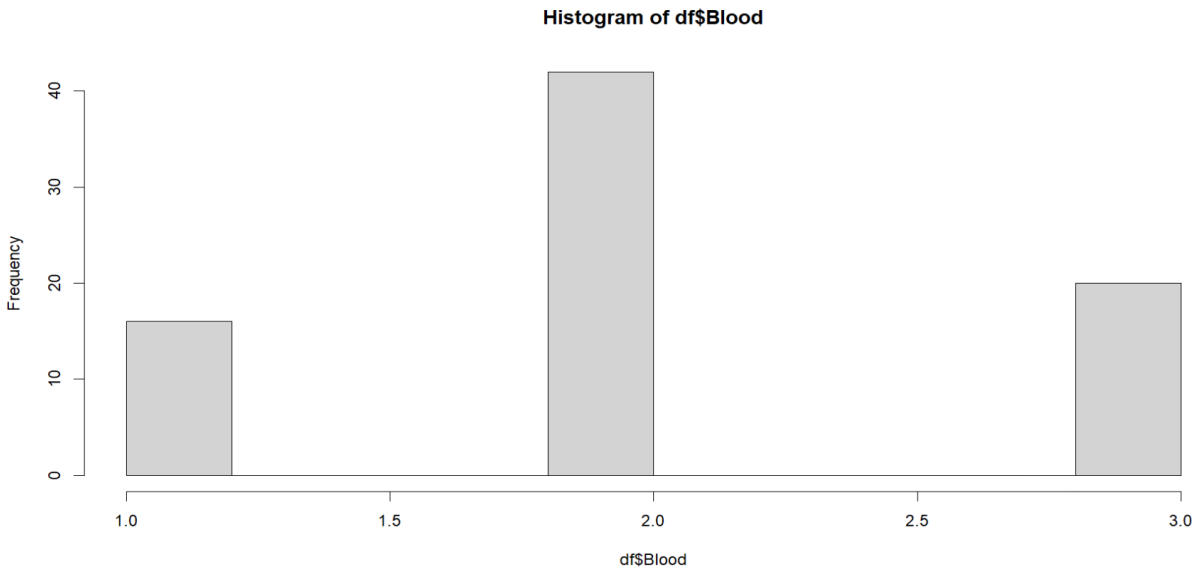
- **Replacing null values with mode:**

```
> md_tb <- table(df$Blood)
> mdb<-names(which.max(md_tb))
> mdb
[1] "normal"
> df<-df %>%
+   mutate(Blood = ifelse(is.na(Blood), mdb , Blood))
```

- **Converting to numeric:**

```
128 df$Blood<-as.numeric(factor(df$Blood,levels = c("low","normal","high"), labels = c(0,1,2)))
129
130 summary(df$Blood)
131 hist(df$Blood)
```

Graph:



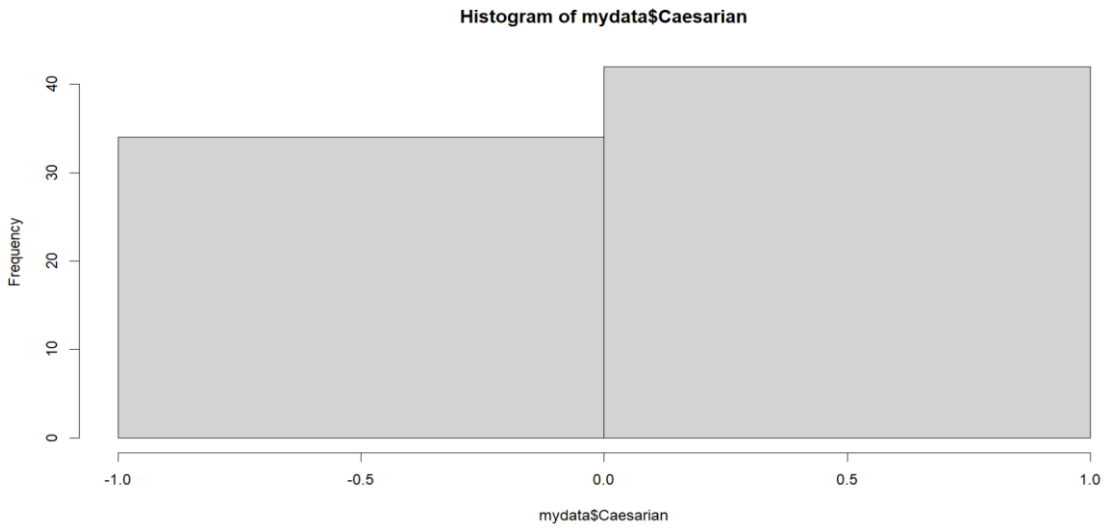
Ceasarian

- **Omitting null values:**

```
137  #as this is totally dependent on medial history
138  mydata<-na.omit(df)
139  summary(mydata)
140  hist(mydata$Caesarian, breaks = c(-1,0,1))
```

Explanation:
In my observation, caesarian is the dependent variable that is going to be predict and it is a sensitive data. So I preferred to omit this instead of replacing.

Graph:



- **Summary:**

```
> summary(mydata)
       id              Age          weight(kg)    Delivery_number Delivery_time        Blood
 Min.   : 1.00   Min.   :18.00   Min.   :49.00   Min.   :1.000   Min.   :0.0000   Min.   :1.000
 1st Qu.:19.75   1st Qu.:25.00   1st Qu.:60.50   1st Qu.:1.000   1st Qu.:0.0000   1st Qu.:2.000
 Median :38.50   Median :27.00   Median :63.00   Median :1.000   Median :0.0000   Median :2.000
 Mean   :39.13   Mean   :27.86   Mean   :63.89   Mean   :1.645   Mean   :0.6184   Mean   :2.039
 3rd Qu.:57.50   3rd Qu.:32.00   3rd Qu.:67.62   3rd Qu.:2.000   3rd Qu.:1.0000   3rd Qu.:2.250
 Max.   :80.00   Max.   :40.00   Max.   :82.00   Max.   :4.000   Max.   :2.0000   Max.   :3.000
     Heart          Caesarian
 Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :1.0000
 Mean   :0.3816   Mean   :0.5526
 3rd Qu.:1.0000   3rd Qu.:1.0000
 Max.   :1.0000   Max.   :1.0000
```