Assingment - 1

(STATISTICS)

NAME:- Suraj Ravi Shiwal

Reg. No. :- **17225760059**

Loading Libraries

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------------------------------------------------- tidyver
se 1.2.1 --
```

```
## v ggplot2 2.2.1     v purrr   0.2.4
## v tibble  1.4.1     v dplyr   0.7.4
## v tidyr   0.7.2     v stringr 1.2.0
## v readr   1.1.1     v forcats 0.2.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------ tidyverse_con
flicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(readxl)

platelet <- read_excel("Dataset 2.xlsx")
platelet<- as.data.frame(platelet)
```

Change necessary columns to factor

```
col_names <- names(platelet)
for (i in col_names) {
  if(length(unique(platelet[,i])) <= 4)
  {
    platelet[,i] <- as.factor(platelet[,i])
  }
}
names(platelet)[2]<-"Study_Group"
```

Binning

```
#Binning Continuos Columns
for (col in 3:ncol(platelet)) {
  column = platelet[,col]
 if(is.numeric(column))
  {
  range=(max(column,na.rm = T)-min(column,na.rm = T))/5
    range=round(range)
    min_val<-min(column,na.rm = T)-range
    max_val=max(column,na.rm = T)
    bin=seq(from=min_val,to=max_val,by=range)
    temp <- cut(column,bin)

    platelet<-cbind(platelet,temp)
    names(platelet)[match("temp",names(platelet))]<-paste0(names(platelet)[col],"_bins")
    }
}
```

Questions:-

Question 1 Construct frequency distribution of all variables according to Group 1 and Group 2

```
freq_dist<- function(clm_to_dist)
{
  dataset <- platelet
  m <<- m+1
  freq_db <- data.frame(cat="a",group="b")
  num_col<-ncol(dataset)
  if(is.numeric(clm_to_dist))
  {

    range=(max(clm_to_dist,na.rm = T)-min(clm_to_dist,na.rm = T))/5
    range=round(range)
    min_val<-min(clm_to_dist,na.rm = T)-range
    max_val=max(clm_to_dist,na.rm = T)
    bin=seq(from=min_val,to=max_val,by=range)
    dataset[,num_col+1] <- cut(clm_to_dist,bin)

     t= dataset %>% group_by(dataset[,(num_col+1)]) %>% summarise(Group_1=sum(Study_Group=="Group 1"),Group_2=sum(Study_Group=="Group 2"))
    names(t)[1] =  names(dataset)[m-1]

    }
  else if(is.factor(clm_to_dist))
   {
     fac_col<-names(dataset)[m-1]
     t = dataset %>% group_by(dataset[,c(fac_col)]) %>% summarise(Group_1 = sum(Study_Group=="Group 1"),Group_2=sum(Study_Group=="Group 2"))
      names(t)[1] <- fac_col
     }

  return(t)
}

m=1
a<-lapply(platelet,freq_dist)
a
```

```
## $Serial
## # A tibble: 6 x 3
##   Serial  Group_1 Group_2
##   <fct>     <int>   <int>
## 1 (-23,1]       1       0
## 2 (1,25]       24       0
## 3 (25,49]      24       0
## 4 (49,73]      11      13
## 5 (73,97]       0      24
## 6 <NA>          0      23
## 
## $Study_Group
## # A tibble: 2 x 3
##   Study_Group Group_1 Group_2
##   <fct>         <int>   <int>
## 1 Group 1          60       0
## 2 Group 2           0      60
## 
## $`Age (yrs)`
## # A tibble: 6 x 3
##   `Age (yrs)` Group_1 Group_2
##   <fct>         <int>   <int>
## 1 (11,21]           0       1
## 2 (21,31]           9       7
## 3 (31,41]          19      18
## 4 (41,51]          16      17
## 5 (51,61]           9       9
## 6 (61,71]           7       8
## 
## $Sex
## # A tibble: 2 x 3
##   Sex    Group_1 Group_2
##   <fct>    <int>   <int>
## 1 Female      30      36
## 2 Male        30      24
## 
## $`Family Income(Rs)`
## # A tibble: 6 x 3
##   `Family Income(Rs)` Group_1 Group_2
##   <fct>                 <int>   <int>
## 1 (1.2e+04,1.5e+04]         6       3
## 2 (1.5e+04,1.8e+04]        27      17
## 3 (1.8e+04,2.1e+04]        18      26
## 4 (2.1e+04,2.4e+04]         4       3
## 5 (2.4e+04,2.7e+04]         4      10
## 6 (2.7e+04,3e+04]           1       1
## 
## $`Duration of Hospitalization`
## # A tibble: 6 x 3
##   `Duration of Hospitalization` Group_1 Group_2
##   <fct>                           <int>   <int>
## 1 (-2,1]                              4       3
## 2 (1,4]                              50      34
## 3 (4,7]                               6      14
## 4 (7,10]                              0       6
## 5 (10,13]                             0       2
## 6 <NA>                                0       1
## 
## $Platelets
## # A tibble: 6 x 3
##   Platelets           Group_1 Group_2
##   <fct>                 <int>   <int>
## 1 (-5.58e+04,6e+03]         3       1
## 2 (6e+03,6.78e+04]         30      23
## 3 (6.78e+04,1.3e+05]       15      25
## 4 (1.3e+05,1.91e+05]        6       8
## 5 (1.91e+05,2.53e+05]       4       1
## 6 (2.53e+05,3.15e+05]       2       2
## 
## $`Systolic blood presure`
## # A tibble: 7 x 3
##   `Systolic blood presure` Group_1 Group_2
##   <fct>                      <int>   <int>
## 1 (64,90]                        1       1
## 2 (90,116]                      10       6
## 3 (116,142]                     25      24
## 4 (142,168]                     14      21
## 5 (168,194]                      4       2
## 6 (194,220]                      1       1
## 7 <NA>                           5       5
##
## $`Diastolic blood presure`
## # A tibble: 7 x 3
##   `Diastolic blood presure` Group_1 Group_2
##   <fct>                       <int>   <int>
## 1 (33,44]                         1       0
## 2 (44,55]                         2       5
## 3 (55,66]                        14      10
## 4 (66,77]                        21      19
## 5 (77,88]                        13      19
## 6 (88,99]                         4       2
## 7 <NA>                            5       5
##
## $BMI
## # A tibble: 7 x 3
##   BMI         Group_1 Group_2
##   <fct>         <int>   <int>
## 1 (14.6,17.6]       2       2
## 2 (17.6,20.6]      12       5
## 3 (20.6,23.6]      20      27
## 4 (23.6,26.6]      15      13
## 5 (26.6,29.6]       7       8
## 6 (29.6,32.6]       3       3
## 7 <NA>              1       2
##
## $`Culture 1`
## # A tibble: 2 x 3
##   `Culture 1` Group_1 Group_2
##   <fct>         <int>   <int>
## 1 Negative         59      54
## 2 Positive          1       6
##
## $`Culture 2`
## # A tibble: 2 x 3
##   `Culture 2` Group_1 Group_2
##   <fct>         <int>   <int>
## 1 Negative         43      42
## 2 Positive         17      18
##
## $`Age (yrs)_bins`
## # A tibble: 6 x 3
##   `Age (yrs)_bins` Group_1 Group_2
##   <fct>              <int>   <int>
## 1 (11,21]                0       1
## 2 (21,31]                9       7
## 3 (31,41]               19      18
## 4 (41,51]               16      17
## 5 (51,61]                9       9
## 6 (61,71]                7       8
##
## $`Family Income(Rs)_bins`
## # A tibble: 6 x 3
##   `Family Income(Rs)_bins` Group_1 Group_2
##   <fct>                      <int>   <int>
## 1 (1.2e+04,1.5e+04]              6       3
## 2 (1.5e+04,1.8e+04]             27      17
## 3 (1.8e+04,2.1e+04]             18      26
## 4 (2.1e+04,2.4e+04]              4       3
## 5 (2.4e+04,2.7e+04]              4      10
## 6 (2.7e+04,3e+04]                1       1
##
## $`Duration of Hospitalization_bins`
## # A tibble: 6 x 3
##   `Duration of Hospitalization_bins` Group_1 Group_2
##   <fct>                                <int>   <int>
## 1 (-2,1]                                   4       3
## 2 (1,4]                                   50      34
## 3 (4,7]                                    6      14
## 4 (7,10]                                   0       6
## 5 (10,13]                                  0       2
## 6 <NA>                                     0       1
```

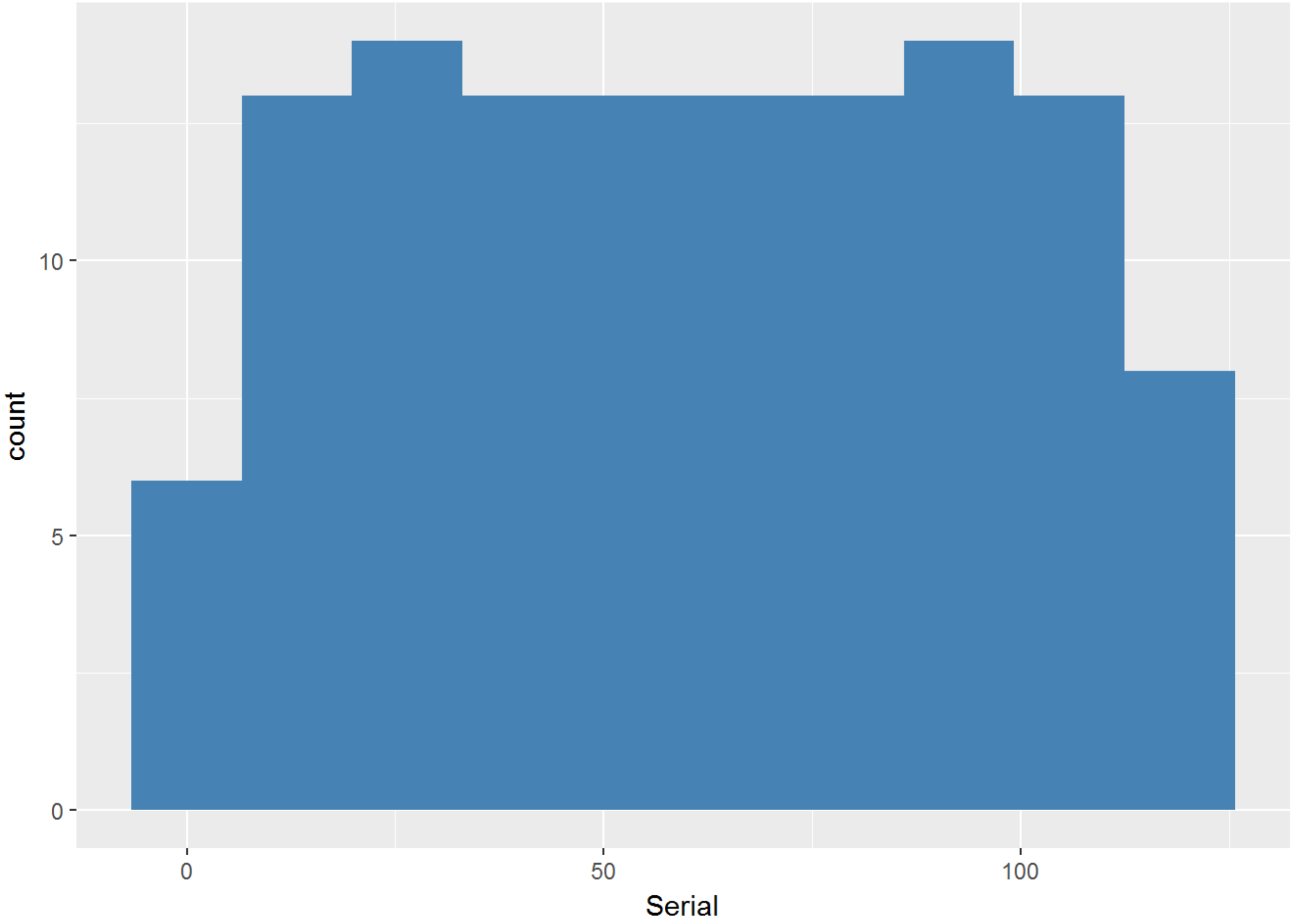
```
##
## $Platelets_bins
## # A tibble: 6 x 3
##   Platelets_bins      Group_1 Group_2
##   <fct>                 <int>   <int>
## 1 (-5.58e+04,6e+03]         3       1
## 2 (6e+03,6.78e+04]         30      23
## 3 (6.78e+04,1.3e+05]       15      25
## 4 (1.3e+05,1.91e+05]        6       8
## 5 (1.91e+05,2.53e+05]       4       1
## 6 (2.53e+05,3.15e+05]       2       2
##
## $`Systolic blood presure_bins`
## # A tibble: 7 x 3
##   `Systolic blood presure_bins` Group_1 Group_2
##   <fct>                           <int>   <int>
## 1 (64,90]                             1       1
## 2 (90,116]                           10       6
## 3 (116,142]                          25      24
## 4 (142,168]                          14      21
## 5 (168,194]                           4       2
## 6 (194,220]                           1       1
## 7 <NA>                                5       5
##
## $`Diastolic blood presure_bins`
## # A tibble: 7 x 3
##   `Diastolic blood presure_bins` Group_1 Group_2
##   <fct>                            <int>   <int>
## 1 (33,44]                              1       0
## 2 (44,55]                              2       5
## 3 (55,66]                             14      10
## 4 (66,77]                             21      19
## 5 (77,88]                             13      19
## 6 (88,99]                              4       2
## 7 <NA>                                 5       5
##
## $BMI_bins
## # A tibble: 7 x 3
##   BMI_bins    Group_1 Group_2
##   <fct>         <int>   <int>
## 1 (14.6,17.6]       2       2
## 2 (17.6,20.6]      12       5
## 3 (20.6,23.6]      20      27
## 4 (23.6,26.6]      15      13
## 5 (26.6,29.6]       7       8
## 6 (29.6,32.6]       3       3
## 7 <NA>              1       2
```

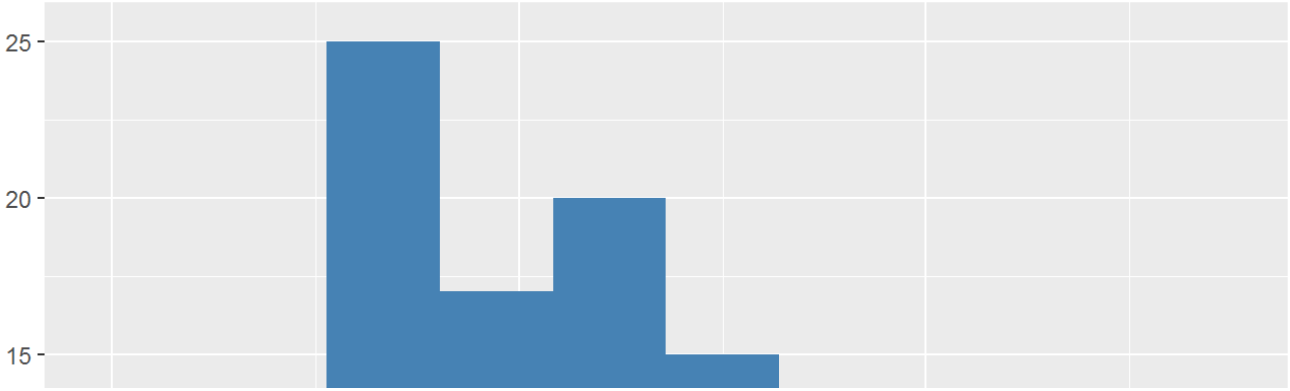
```
View(platelet)
```

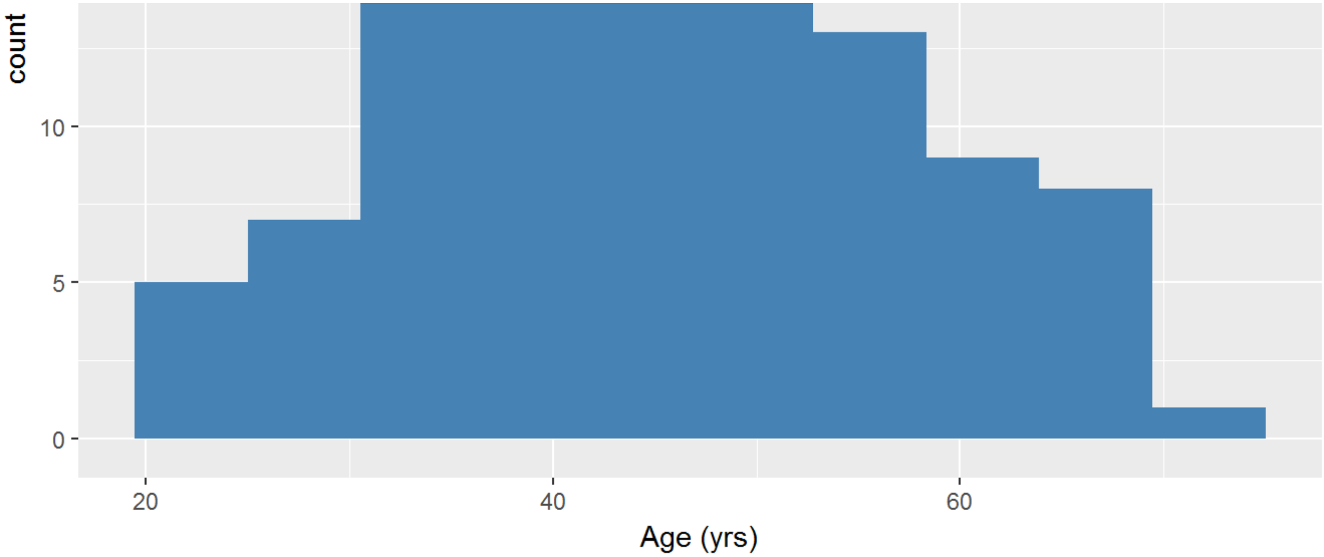
Question 2 Represent the all the given variables below using appropriate graphical presentation

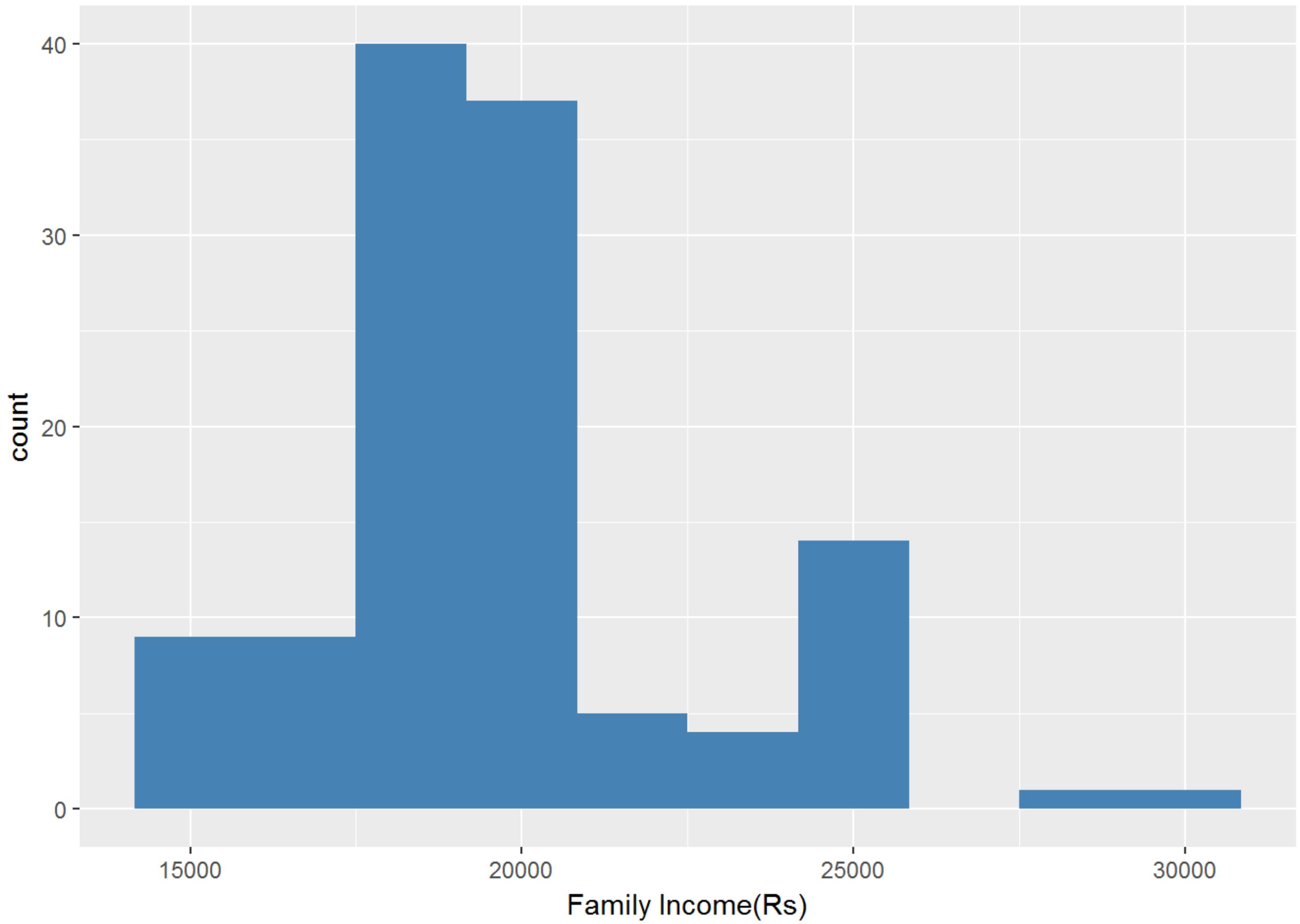
Numerical Column

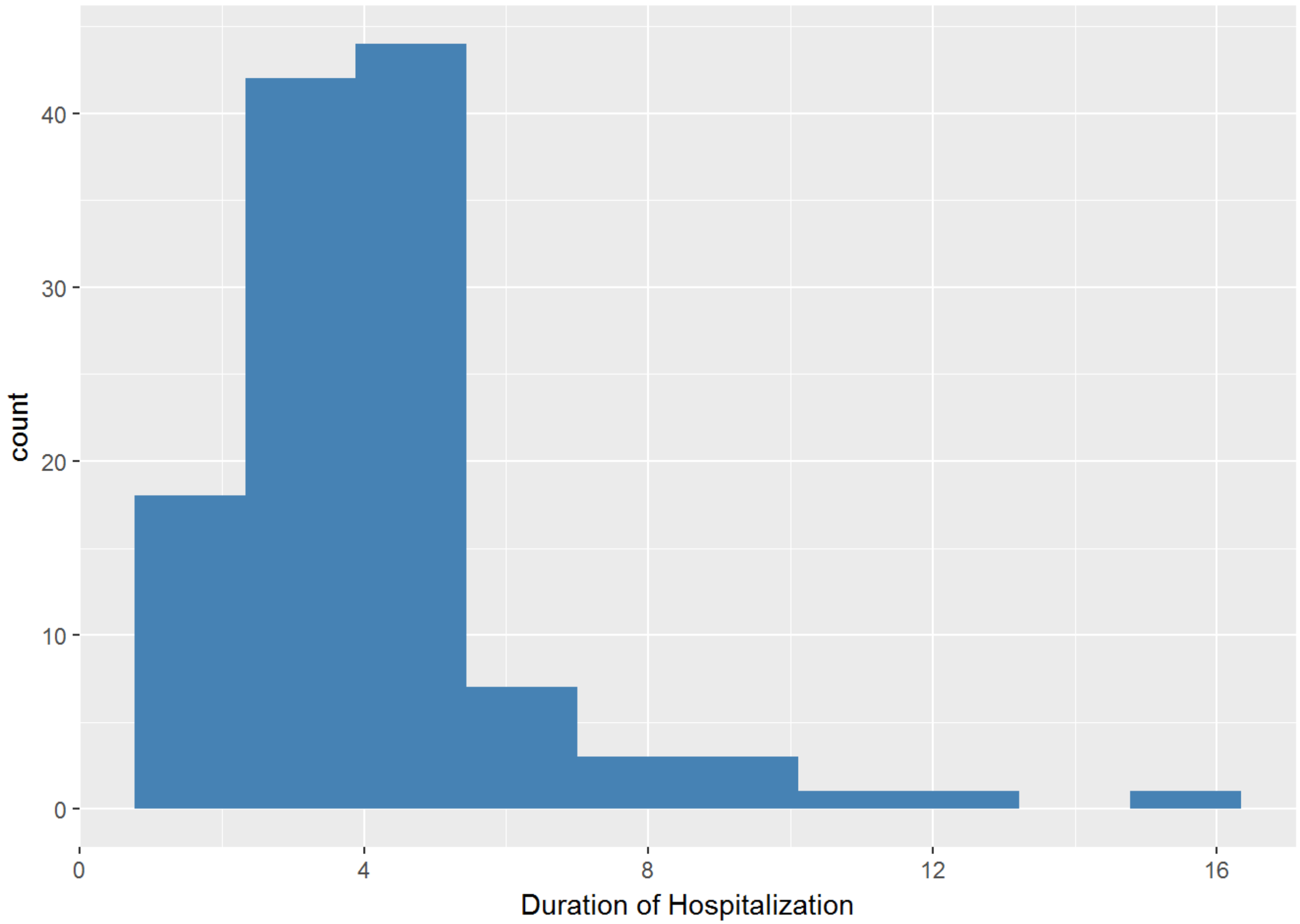
```
num = sapply(platelet, is.numeric)

plot_graph_num = function(col){
  new = data.frame(col = platelet[,col])
  print(new %>% na.omit() %>%
          ggplot(aes(x=col)) + geom_histogram(fill = 'steelblue',bins = 10) + xlab(col))
}

sapply(names(platelet)[num], plot_graph_num)
```

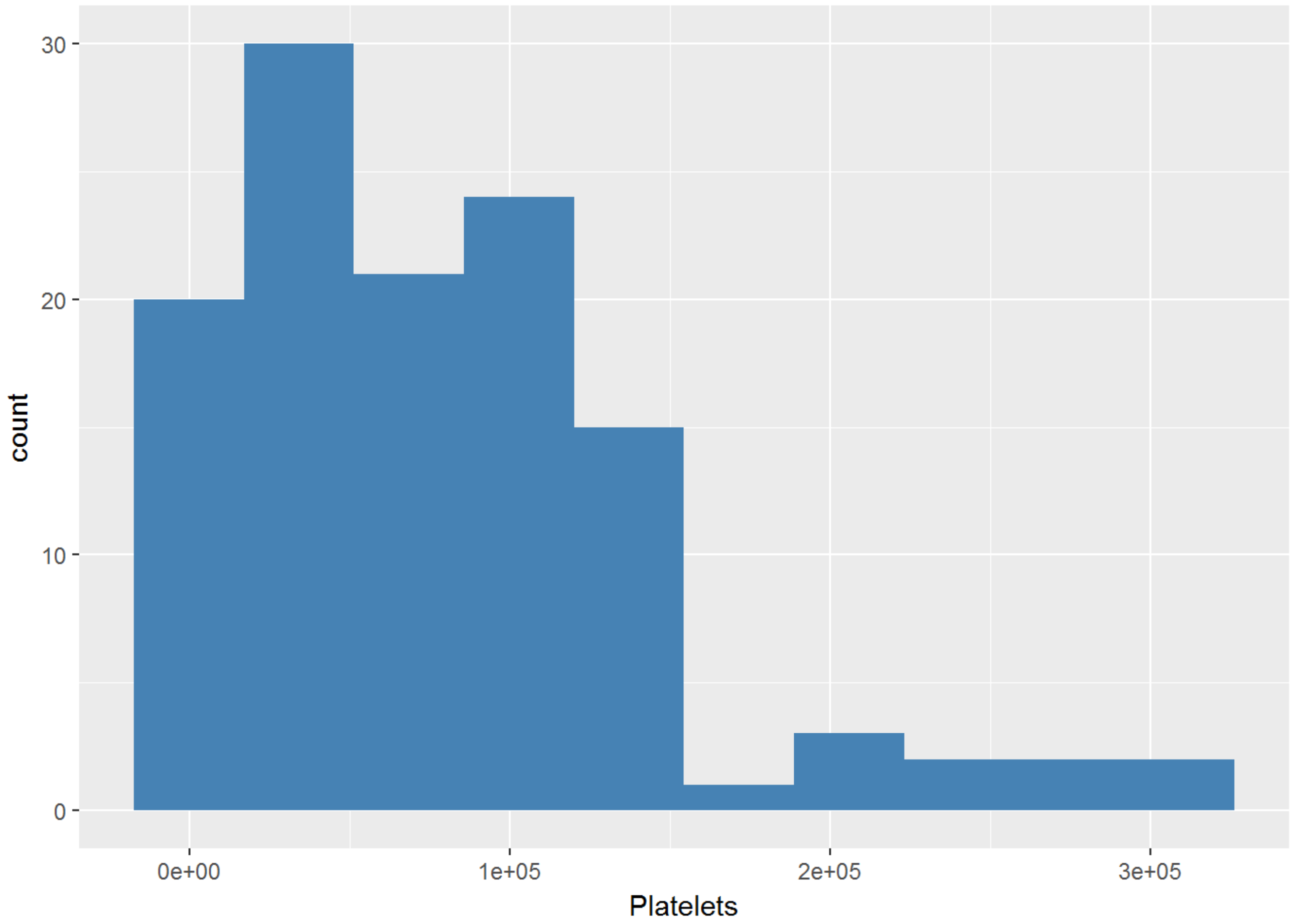

count
10
5
0
0
50
100
Serial

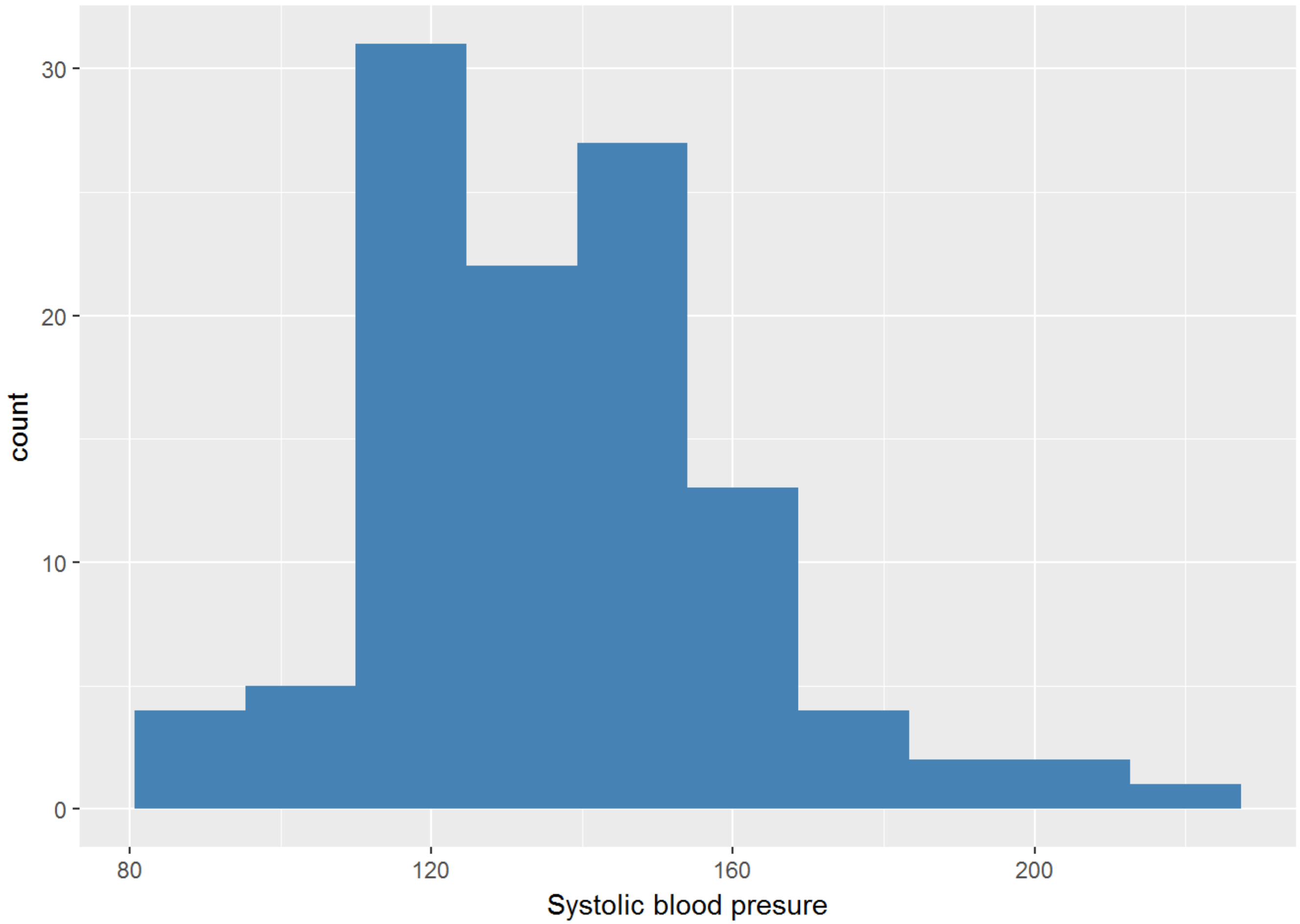

25
20
15

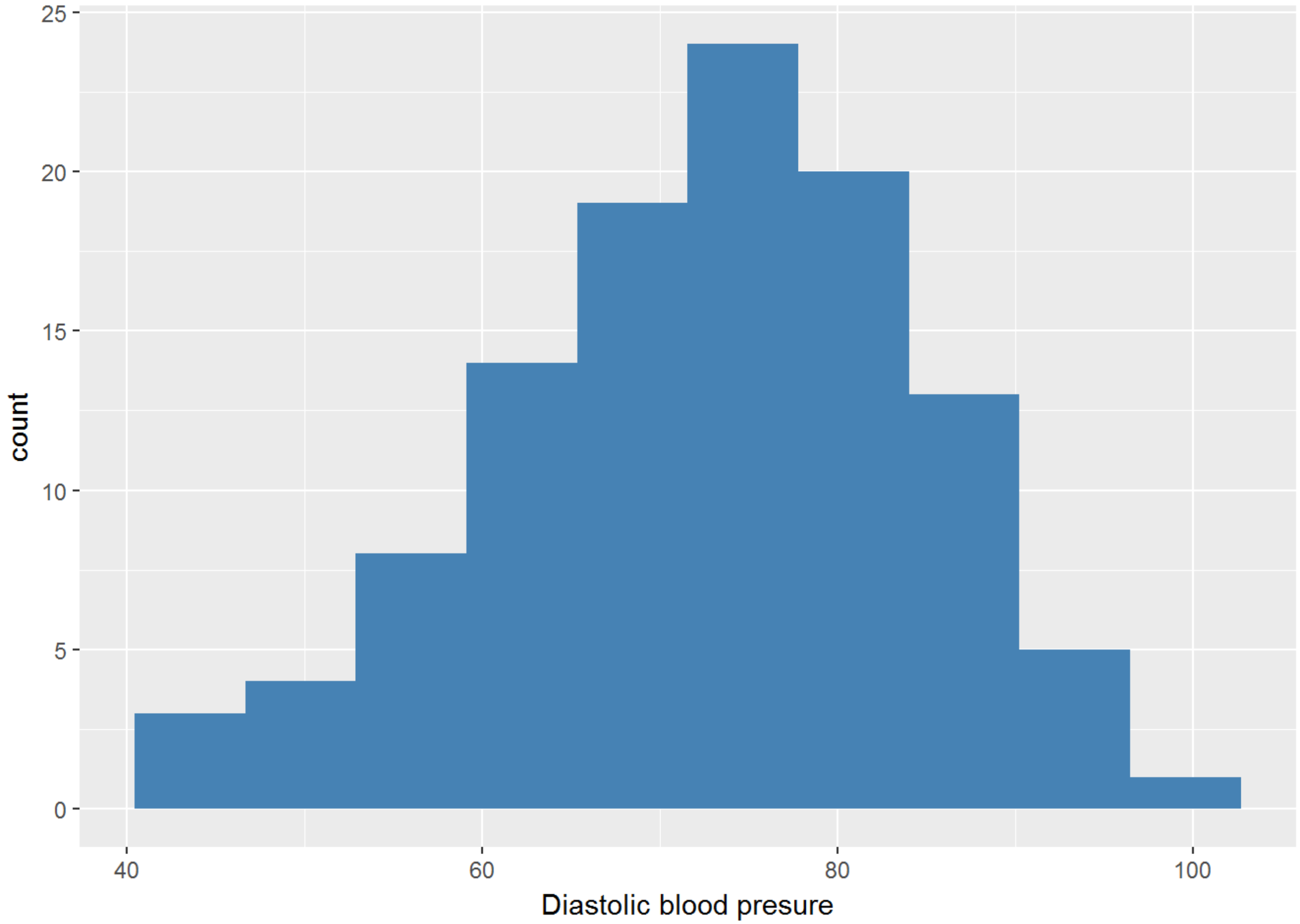

count
10
5
0
20
40
60
Age (yrs)

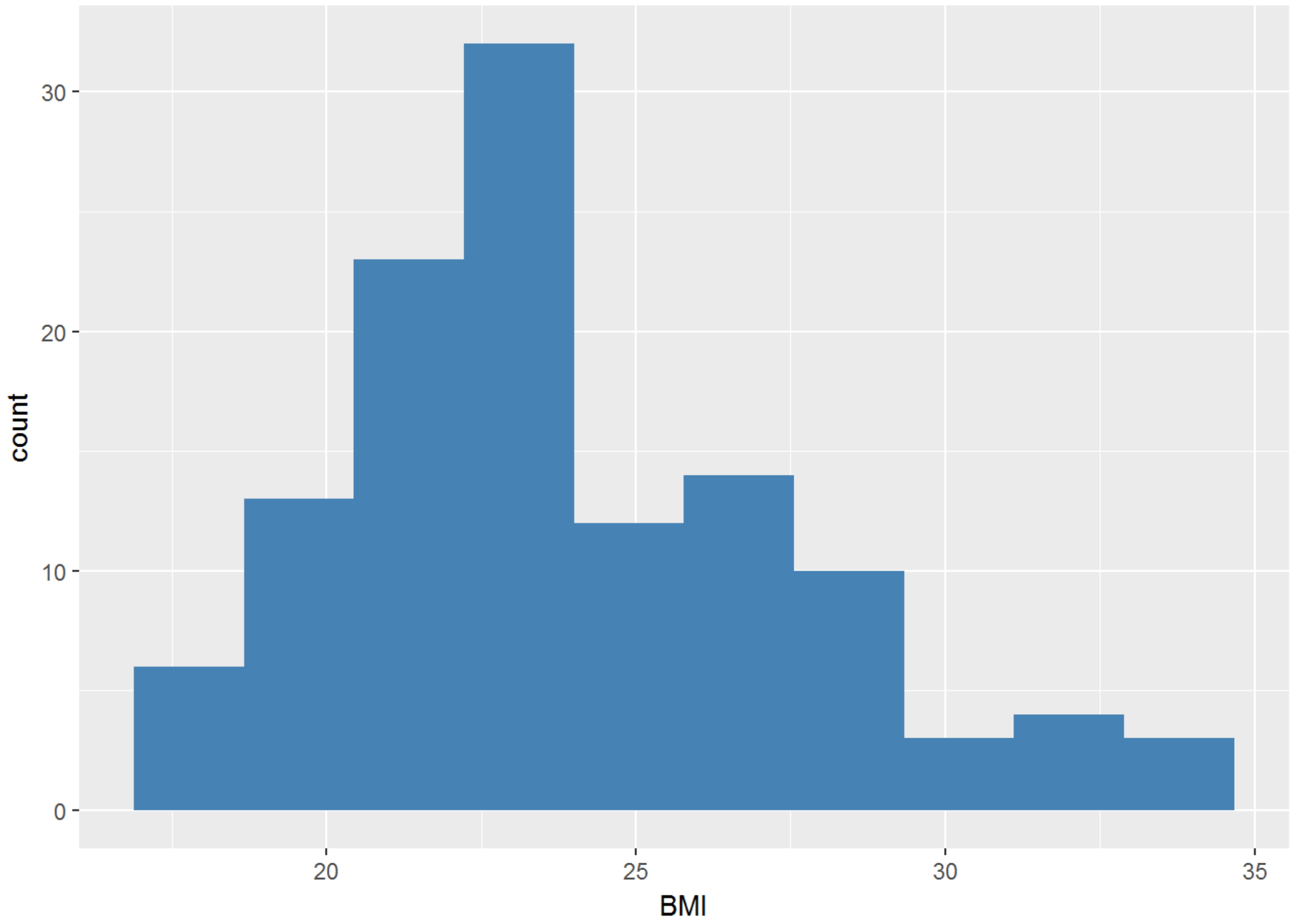

count
40
30
20
10
0
15000
20000
25000
30000
Family Income(Rs)


count
40
30
20
10
0
0
4
8
12
16
Duration of Hospitalization


30
20
10
0
count
0e+00
1e+05
2e+05
3e+05
Platelets


30
20
10
0
count
80
120
160
200
Systolic blood presure

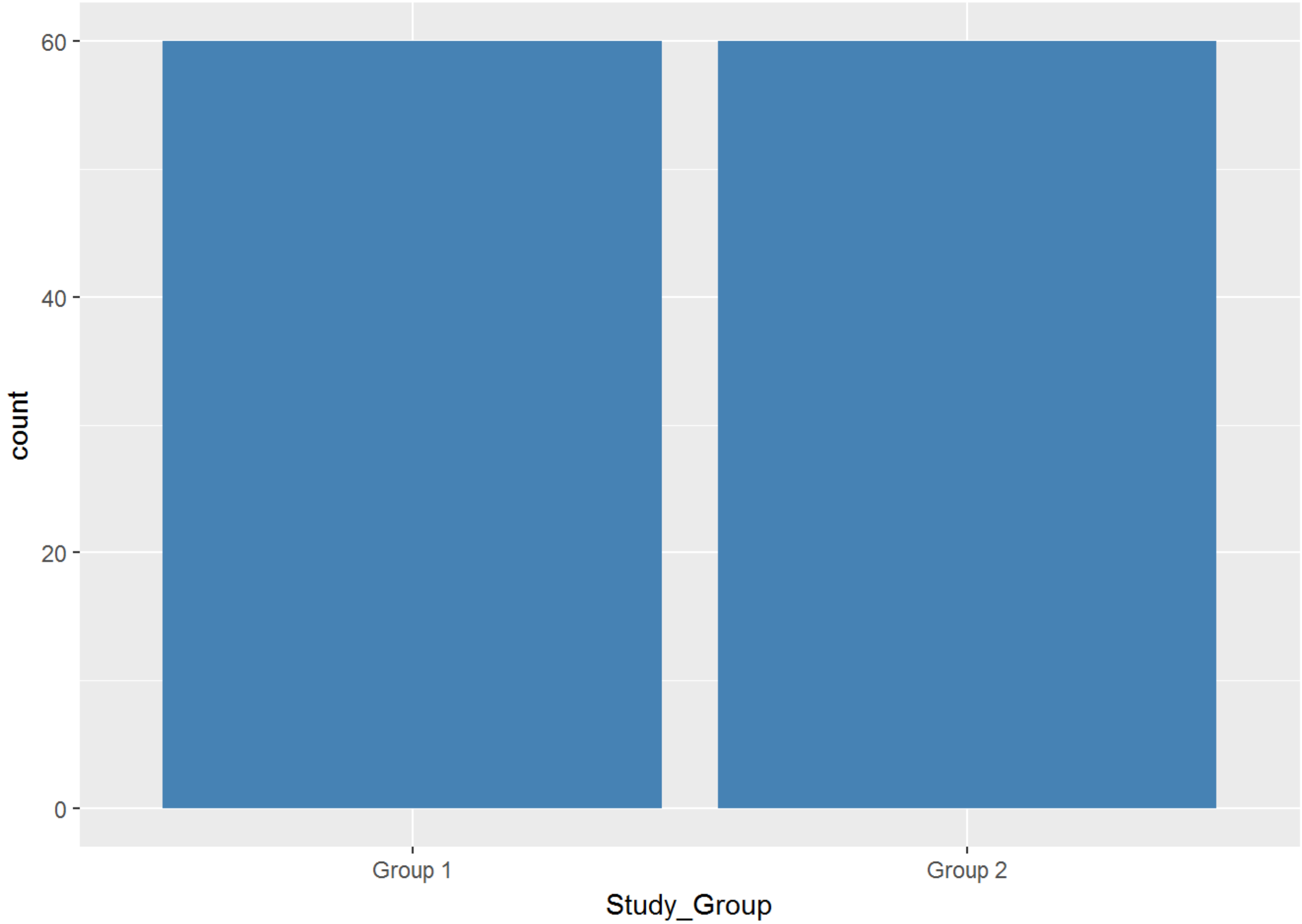

25
20
15
10
5
0
count
40
60
80
100
Diastolic blood presure

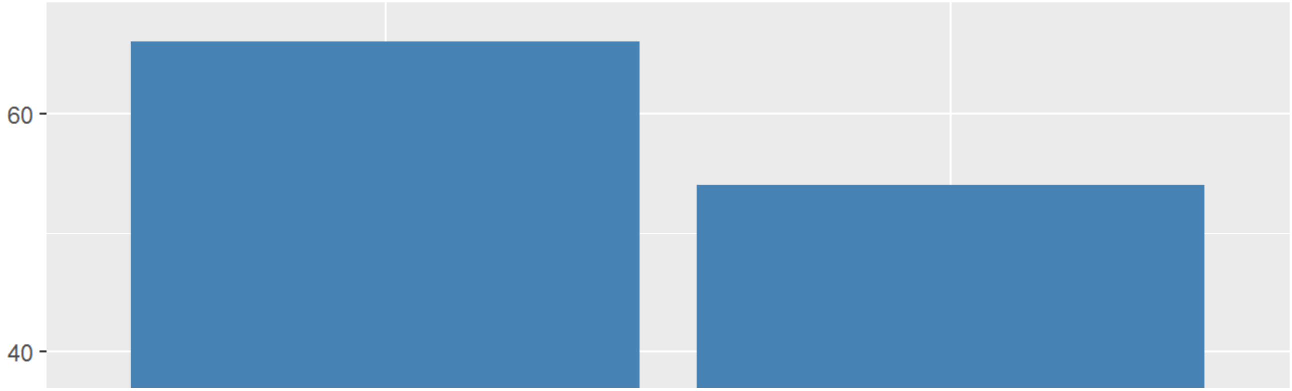


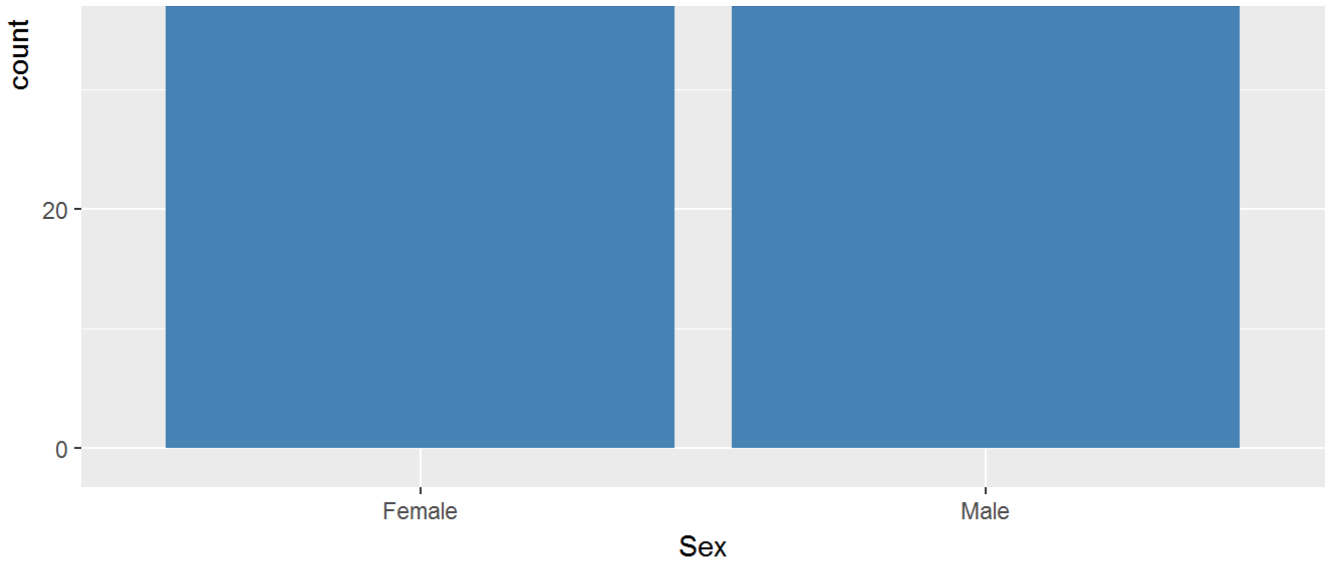
```
##        Serial Age (yrs) Family Income(Rs) Duration of Hospitalization
## data   List,1 List,1    List,1            List,1
## layout ?      ?         ?                 ?
## plot   List,9 List,9    List,9            List,9
##        Platelets Systolic blood presure Diastolic blood presure BMI
## data   List,1    List,1                 List,1                  List,1
## layout ?         ?                      ?                       ?
## plot   List,9    List,9                 List,9                  List,9
```

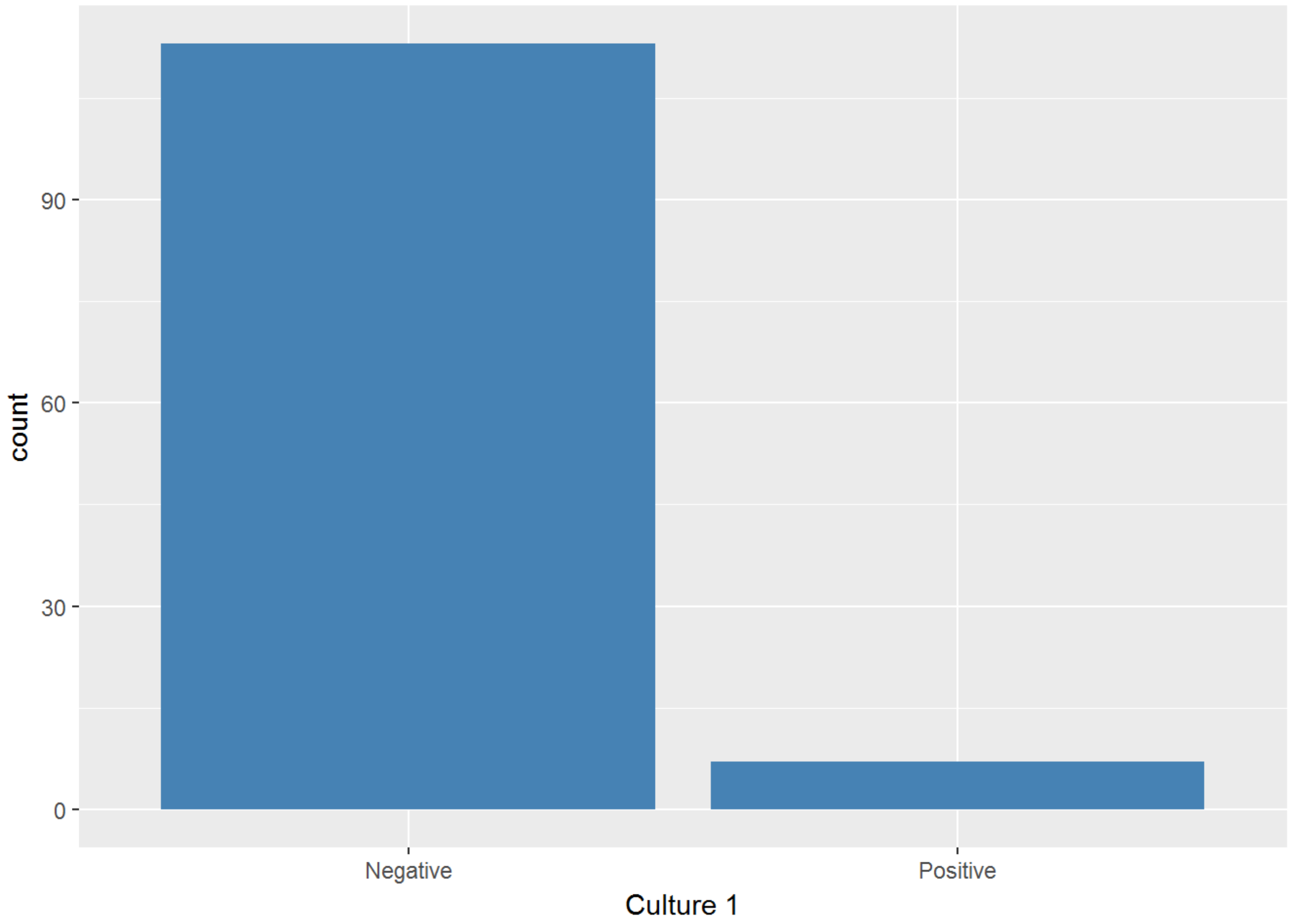
Categorical Column

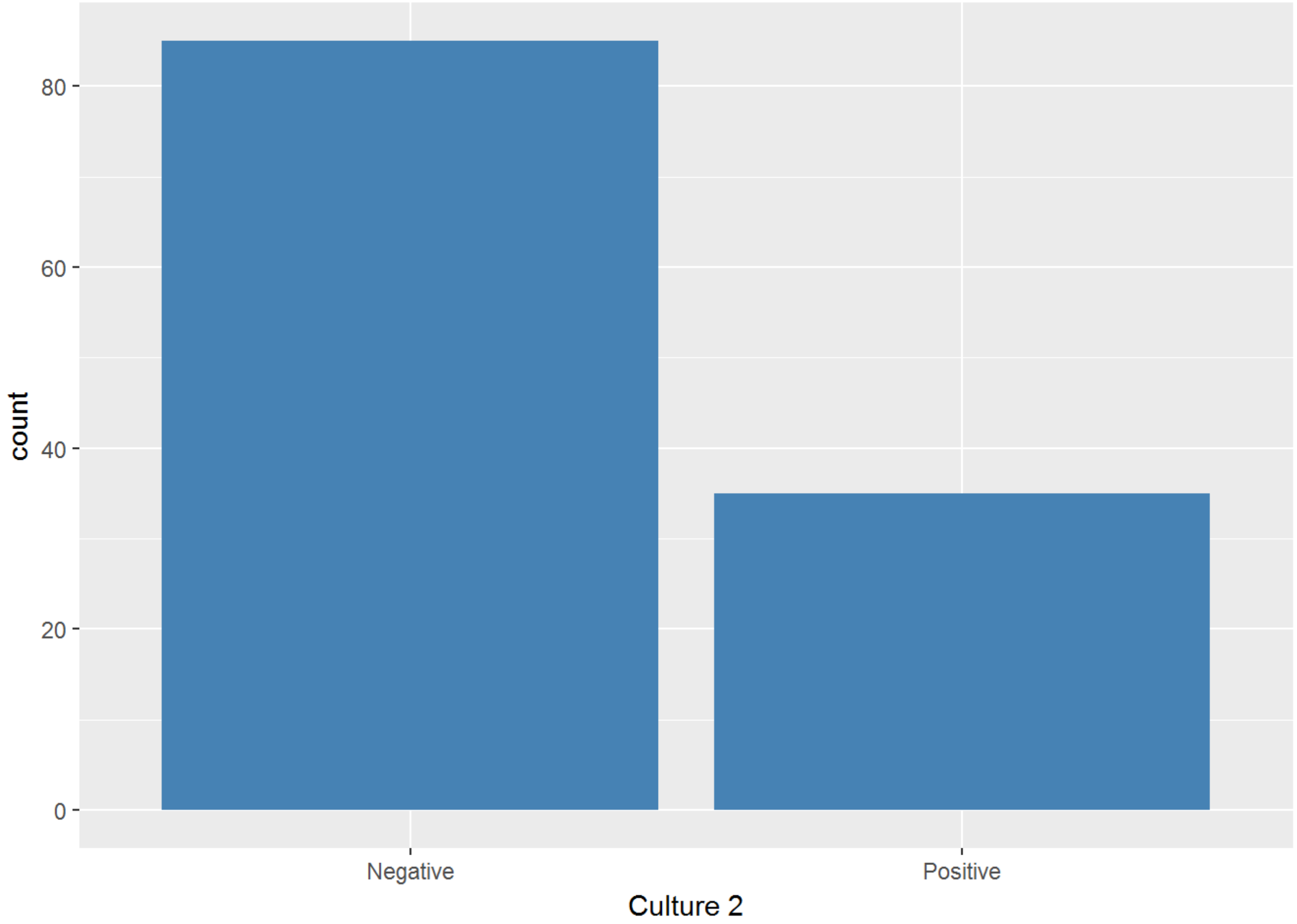
```r
num = sapply(platelet, is.numeric)

gen_plot = function(col){
  new = data.frame(col = platelet[,col])
  print(new %>% na.omit() %>%
            ggplot(aes(x=col)) + geom_bar(fill = 'steelblue') + xlab(col))
}

sapply(names(platelet)[!num], gen_plot)
```

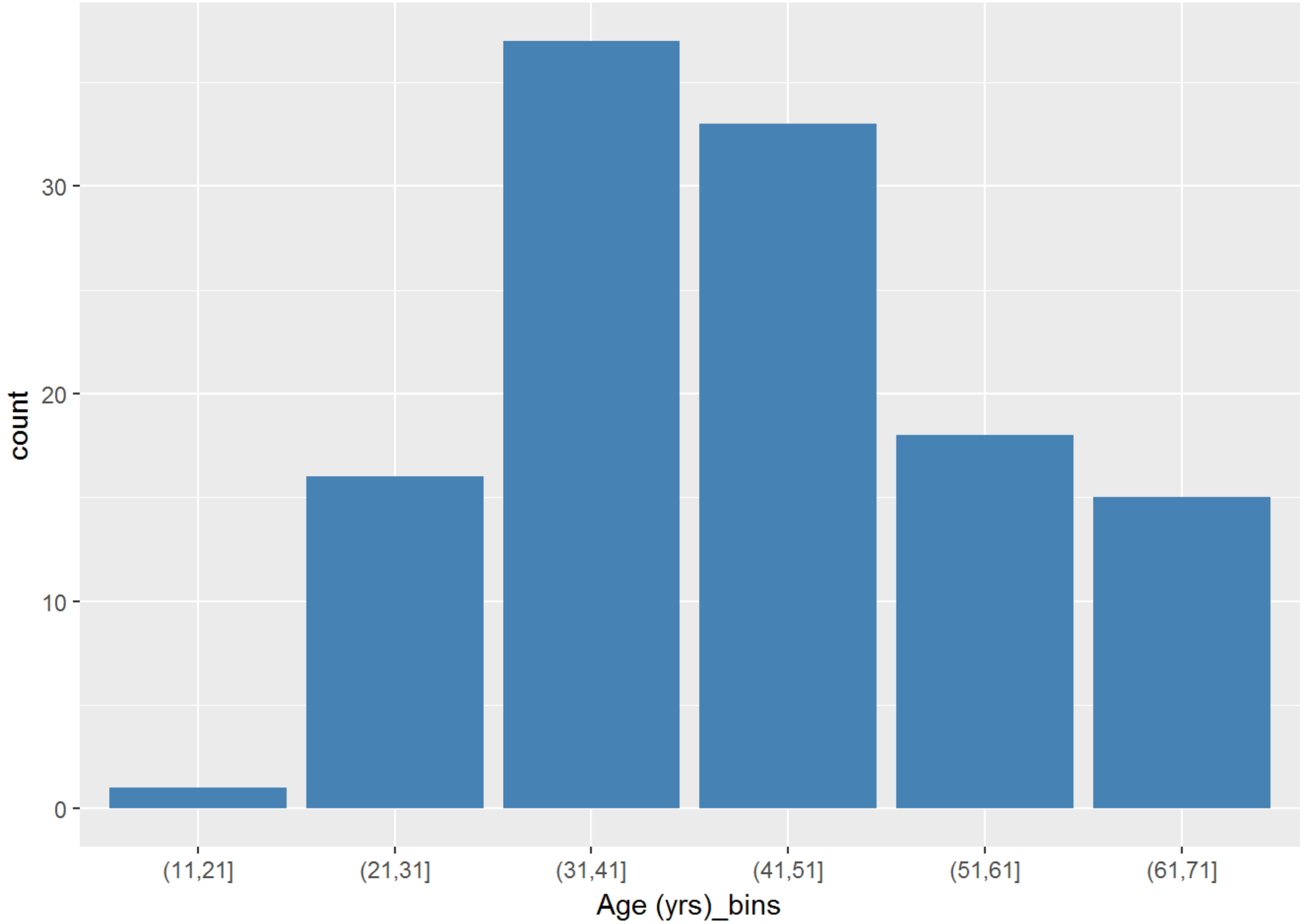

60
40
20
0
count
Group 1
Group 2
Study_Group

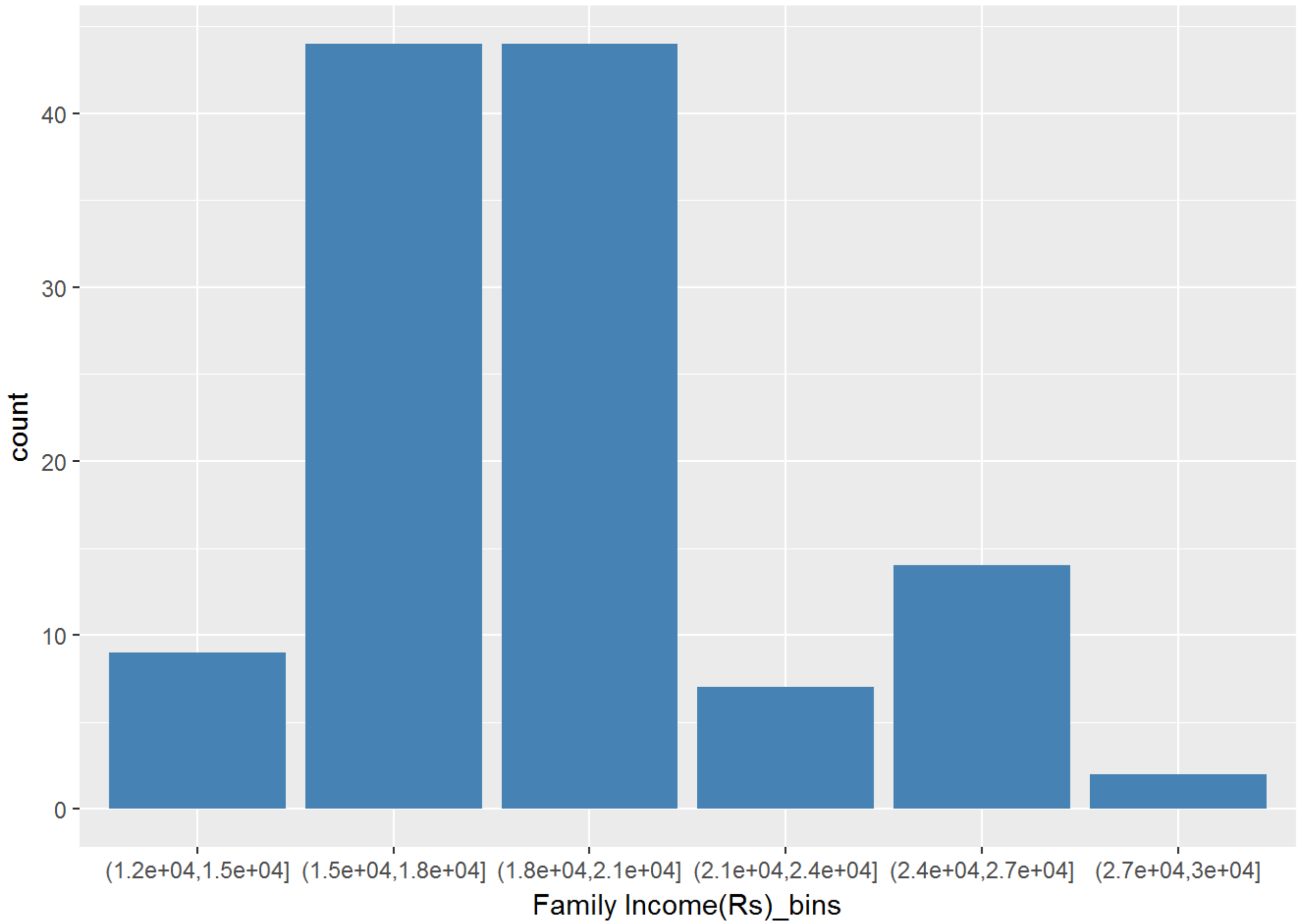

60
40

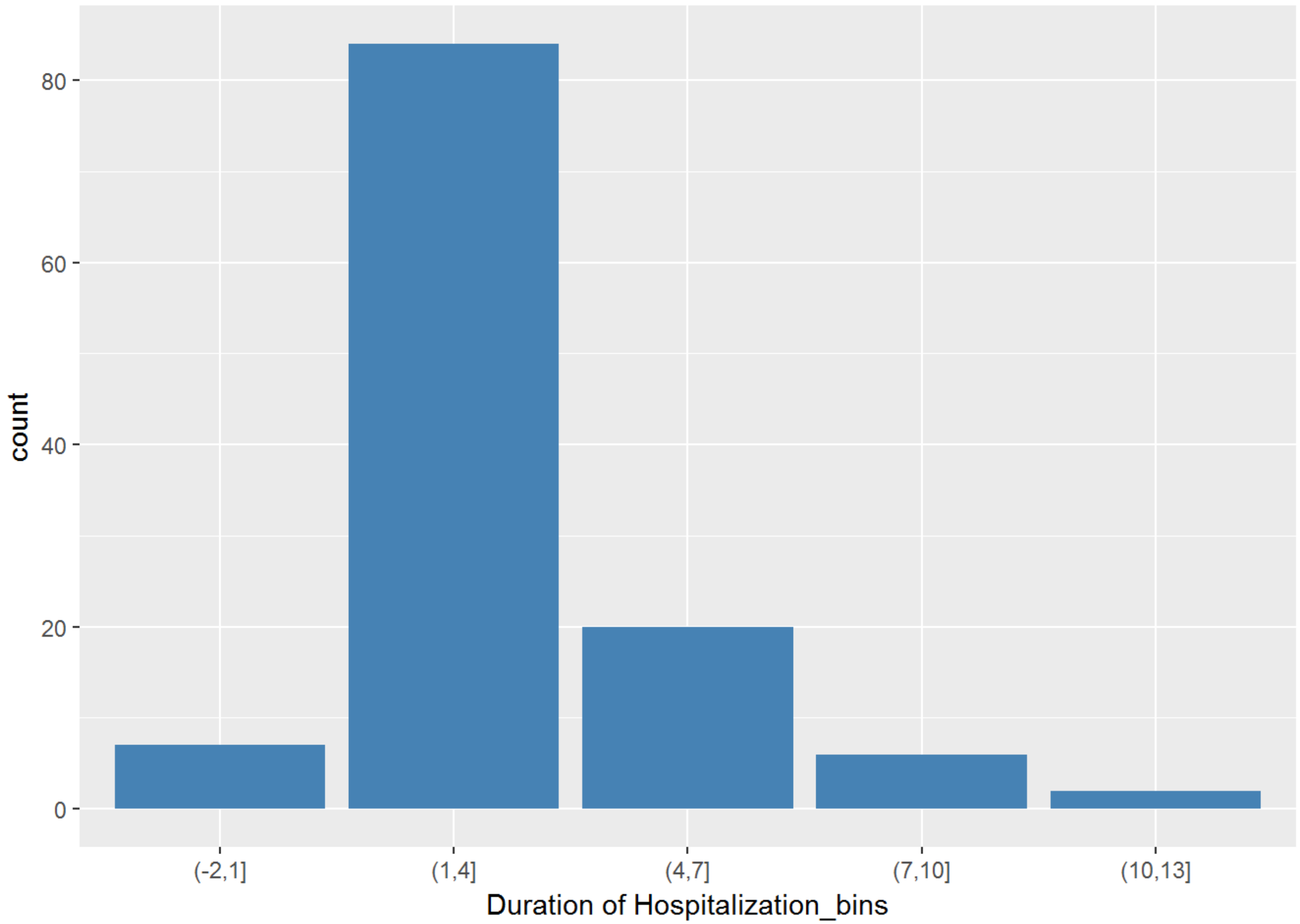

count
20
0
Female
Male
Sex

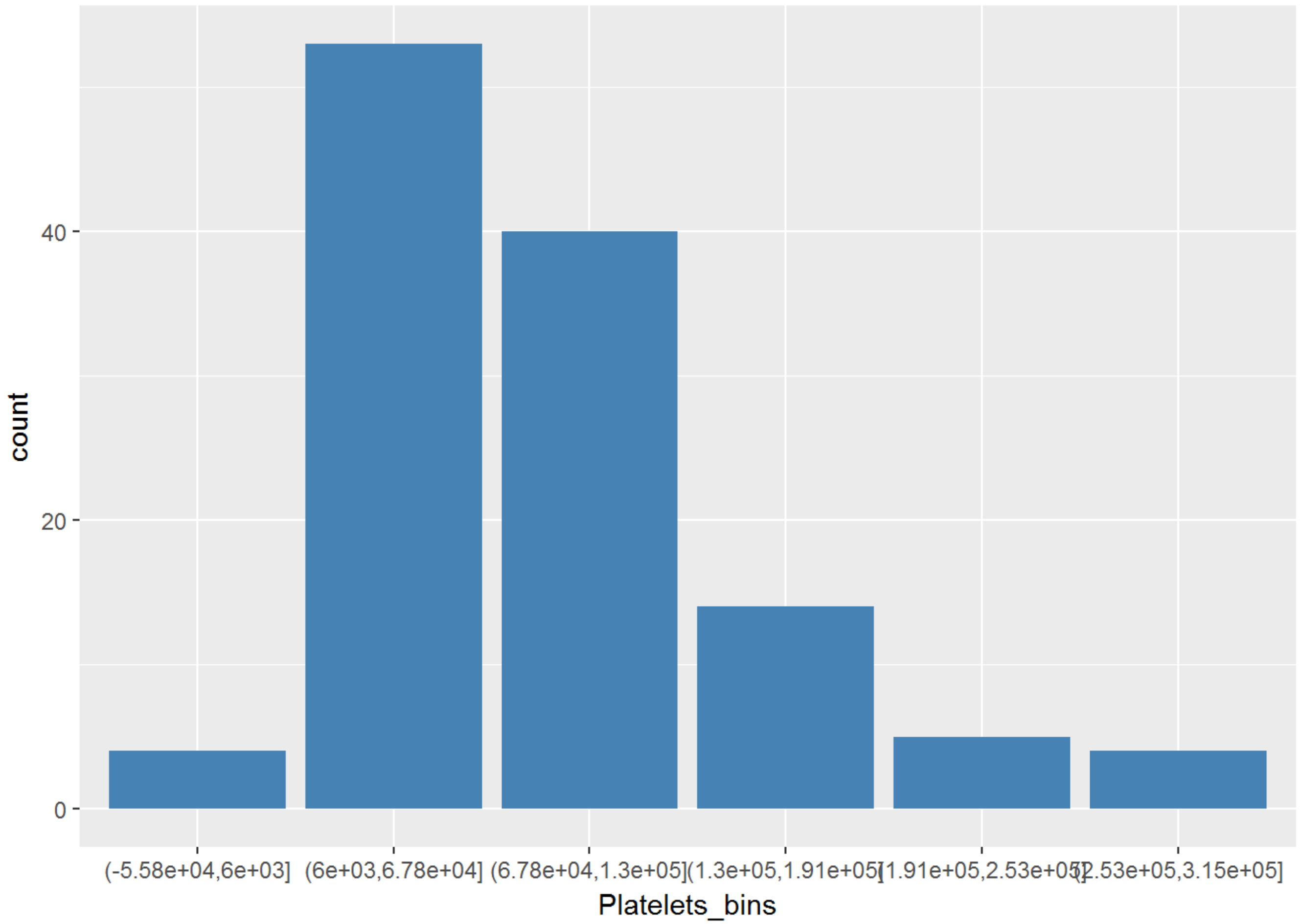

count
90
60
30
0
Negative
Positive
Culture 1

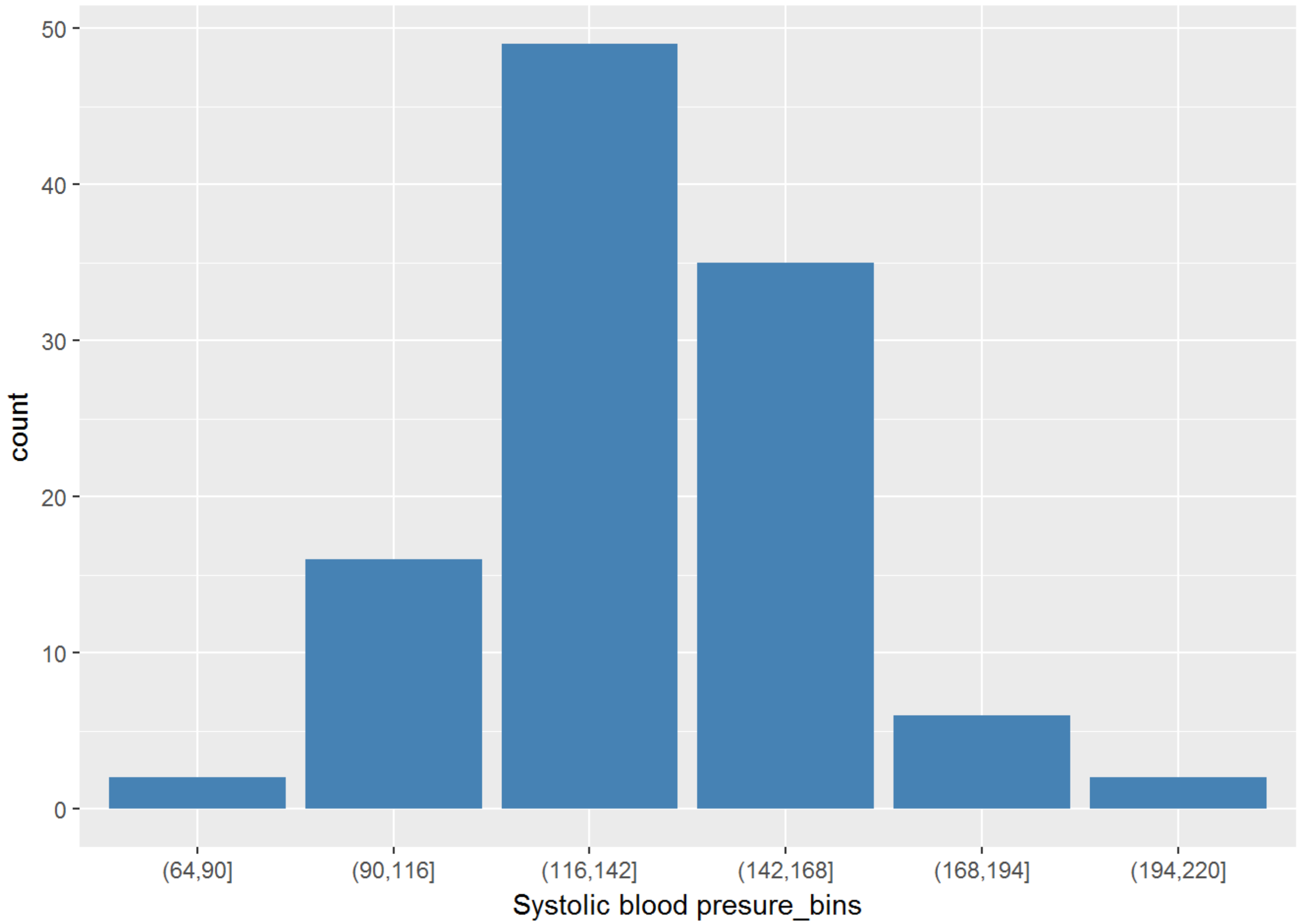

count
80
60
40
20
0
Negative
Positive
Culture 2

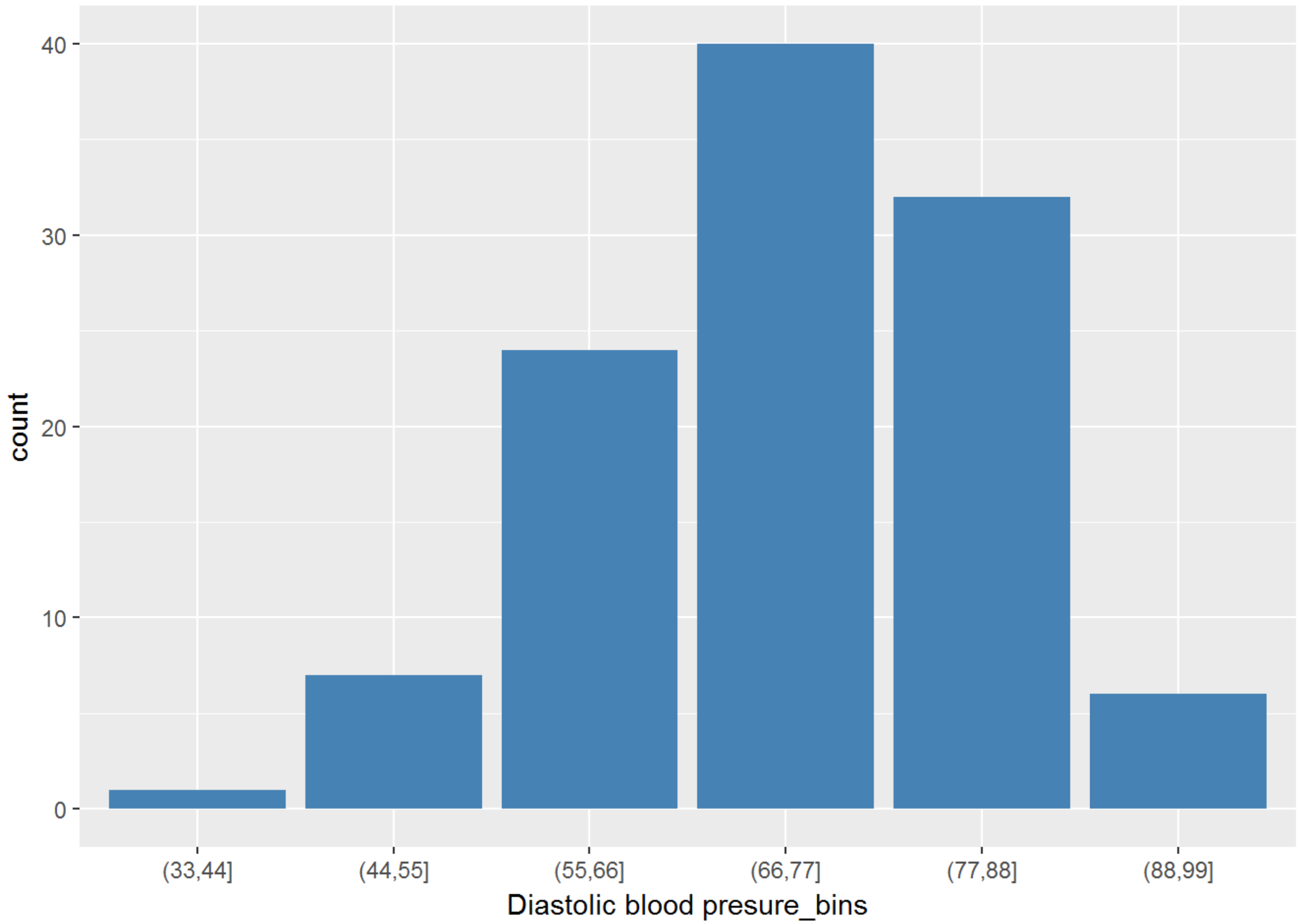

count
30
20
10
0
(11,21]
(21,31]
(31,41]
(41,51]
(51,61]
(61,71]
Age (yrs)_bins

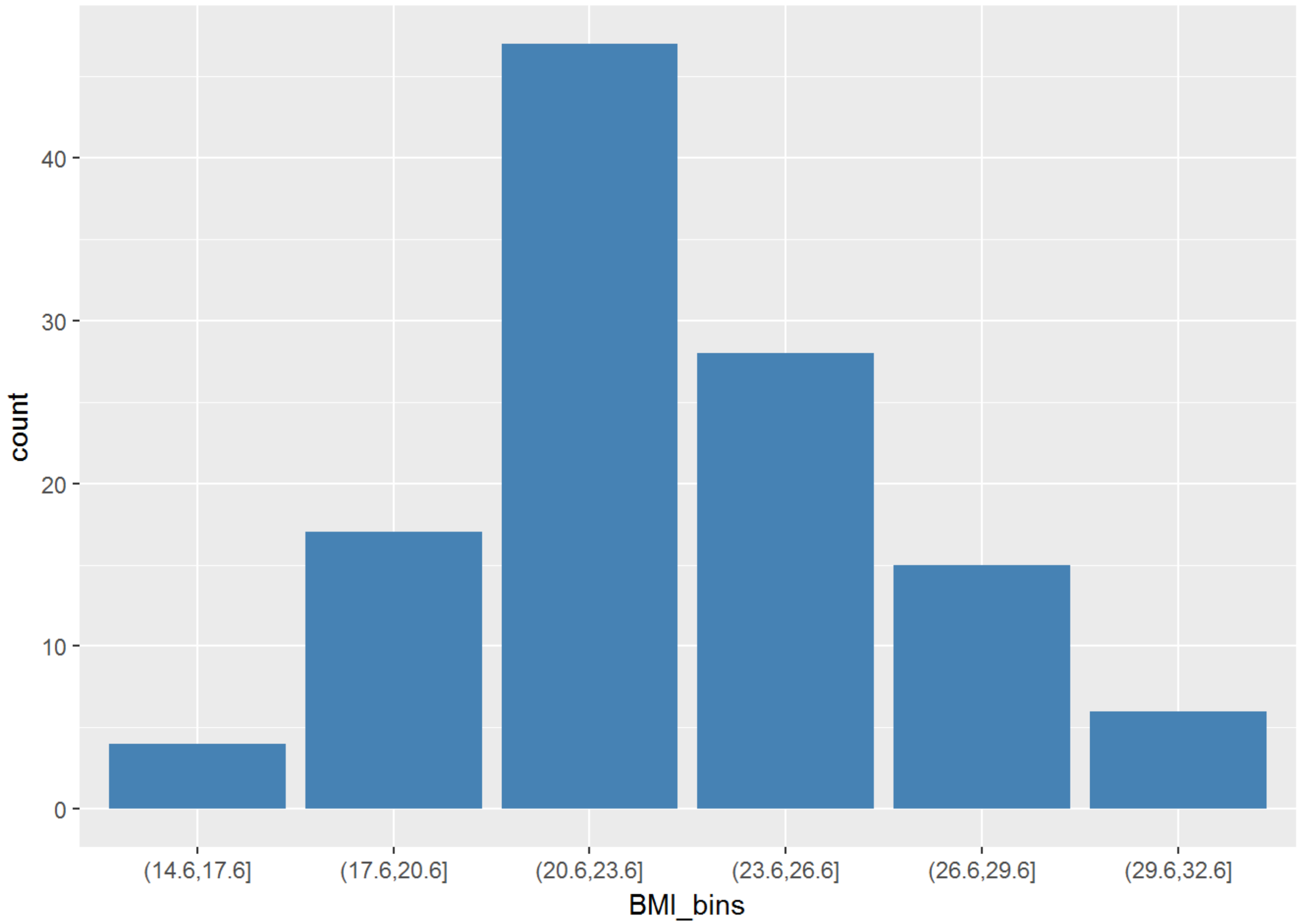

count
0
10
20
30
40
(1.2e+04,1.5e+04]
(1.5e+04,1.8e+04]
(1.8e+04,2.1e+04]
(2.1e+04,2.4e+04]
(2.4e+04,2.7e+04]
(2.7e+04,3e+04]
Family Income(Rs)_bins


count
80
60
40
20
0
(-2,1]
(1,4]
(4,7]
(7,10]
(10,13]
Duration of Hospitalization_bins


count
40
20
0
(-5.58e+04,6e+03]
(6e+03,6.78e+04]
(6.78e+04,1.3e+05]
(1.3e+05,1.91e+05]
(1.91e+05,2.53e+05]
(2.53e+05,3.15e+05]
Platelets_bins


50
40
30
20
10
0
count
(64,90]
(90,116]
(116,142]
(142,168]
(168,194]
(194,220]
Systolic blood presure_bins


count
0
10
20
30
40
(33,44]
(44,55]
(55,66]
(66,77]
(77,88]
(88,99]
Diastolic blood presure_bins



```
##        Study_Group Sex    Culture 1 Culture 2 Age (yrs)_bins
## data   List,1      List,1 List,1    List,1    List,1
## layout ?           ?      ?         ?         ?
## plot   List,9      List,9 List,9    List,9    List,9
##        Family Income(Rs)_bins Duration of Hospitalization_bins
## data   List,1                 List,1
## layout ?                      ?
## plot   List,9                 List,9
##        Platelets_bins Systolic blood presure_bins
## data   List,1         List,1
## layout ?              ?
## plot   List,9         List,9
##        Diastolic blood presure_bins BMI_bins
## data   List,1                       List,1
## layout ?                            ?
## plot   List,9                       List,9
```

Question 3 Construct the cross tables of Age versus Sex, Culture 1 and Culture 2

```
cross_tab <- function(versus=c("Sex","Culture 1","Culture 2"))
{
   cross_tab_list=list()
  for (col in versus) {
    cross_tab_list[[col]]=table(platelet[,"Age (yrs)_bins"],platelet[,col])
  }
  return(cross_tab_list)
}


cross_tab()
```

```
## $Sex
## 
##            Female Male
##   (11,21]       0    1
##   (21,31]       9    7
##   (31,41]      25   12
##   (41,51]      15   18
##   (51,61]      10    8
##   (61,71]       7    8
## 
## $`Culture 1`
```

```
##
##            Negative Positive
##   (11,21]         0        1
##   (21,31]        15        1
##   (31,41]        35        2
##   (41,51]        33        0
##   (51,61]        17        1
##   (61,71]        13        2
##
## $`Culture 2`
##
##            Negative Positive
##   (11,21]         0        1
##   (21,31]        10        6
##   (31,41]        29        8
##   (41,51]        24        9
##   (51,61]        15        3
##   (61,71]         7        8
```

Question 4 Compute the mean and standard deviation of data obtained in the age frequency distribution

```
freq_dist_age <- platelet %>% group_by(`Age (yrs)_bins`) %>% summarise(Total_occurence=n())
mid_vector=c()
for (row in 1:6) {
  bined_age<-freq_dist_age$`Age (yrs)_bins`[row]
 trimed_range<-gsub("\\(|\\]","",bined_age)
 num_min_max<-as.numeric(unlist(strsplit(trimed_range,",")))
 mid = (num_min_max[2]-num_min_max[1])/2
 mid_vector <- append(mid_vector,(mid+num_min_max[1]))
}
freq_dist_age<-cbind(freq_dist_age,mid_vector)

freq_dist_age<-freq_dist_age %>% mutate(F.X=Total_occurence*mid_vector)
mean_Age_Freq_Distribution<-sum(freq_dist_age$F.X)/sum(freq_dist_age$Total_occurence)
```

```
SD_Age_Freq_Didtribution <- sqrt(sum(((freq_dist_age$mid_vector-mean_Age_Freq_Distribution)^2)*freq_dist_age$Tota
l_occurence))/sqrt(sum(freq_dist_age$Total_occurence))

mean_Age_Freq_Distribution
```

```
## [1] 44
```

```
SD_Age_Freq_Didtribution
```

```
## [1] 12.35584
```

Question 5 Construct the cross tables between Culture 1 and Culture 2

```
cross_tab<- platelet %>% group_by(`Culture 1`) %>% summarise(Culture2_Negative=sum(`Culture 2`=="Negative"),Cultu
re2_positive=sum(`Culture 2`=="Positive"))
cross_tab
```

```
## # A tibble: 2 x 3
##   `Culture 1` Culture2_Negative Culture2_positive
##   <fct>                   <int>             <int>
## 1 Negative                   85                28
## 2 Positive                    0                 7
```

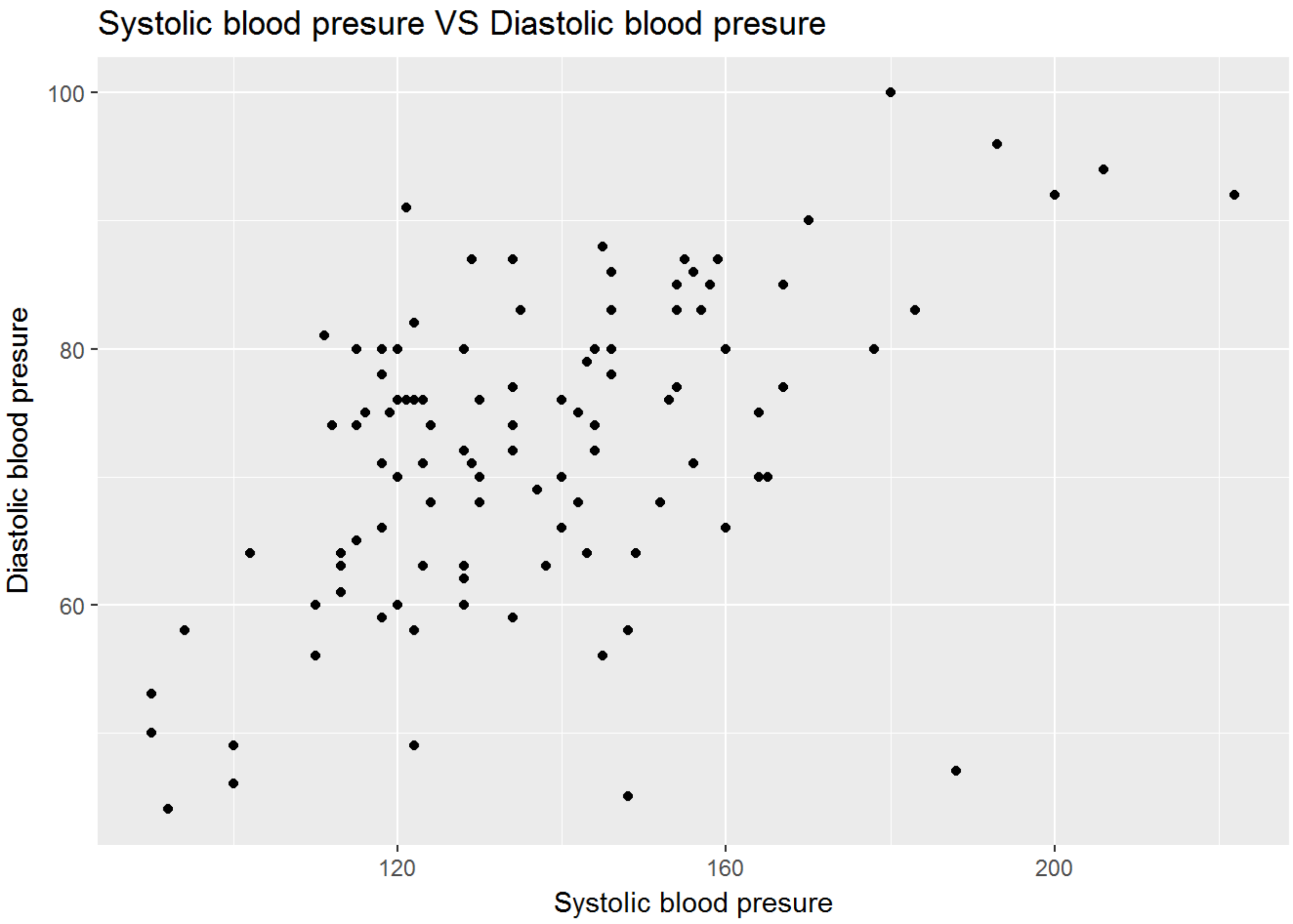
Question 6 Present the summary statistics of all quantitative variables

```r
num_col <- sapply(platelet, is.numeric)
sum_list<- lapply(platelet[,num_col], function(x){if(is.numeric(x)){summary(x)}else{return()}})
sum_list
```

```
## $Serial
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   30.75   60.50   60.50   90.25  120.00
##
## $`Age (yrs)`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   21.00   35.00   43.00   44.25   53.00   71.00
##
## $`Family Income(Rs)`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   15000   18000   20000   19650   20000   30000
##
## $`Duration of Hospitalization`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   3.500   4.008   4.000  15.000
##
## $Platelets
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6000   25000   69500   79850  112000  315000
##
## $`Systolic blood presure`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    90.0   120.5   134.0   137.2   152.5   222.0       9
##
## $`Diastolic blood presure`
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   44.00   64.50   74.00   72.79   80.00  100.00       9
##
## $BMI
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.60   21.49   23.20   23.87   25.91   33.60
```

Question 7 Draw a scattered diagram between Systolic blood pressure and Diastolic blood pressure

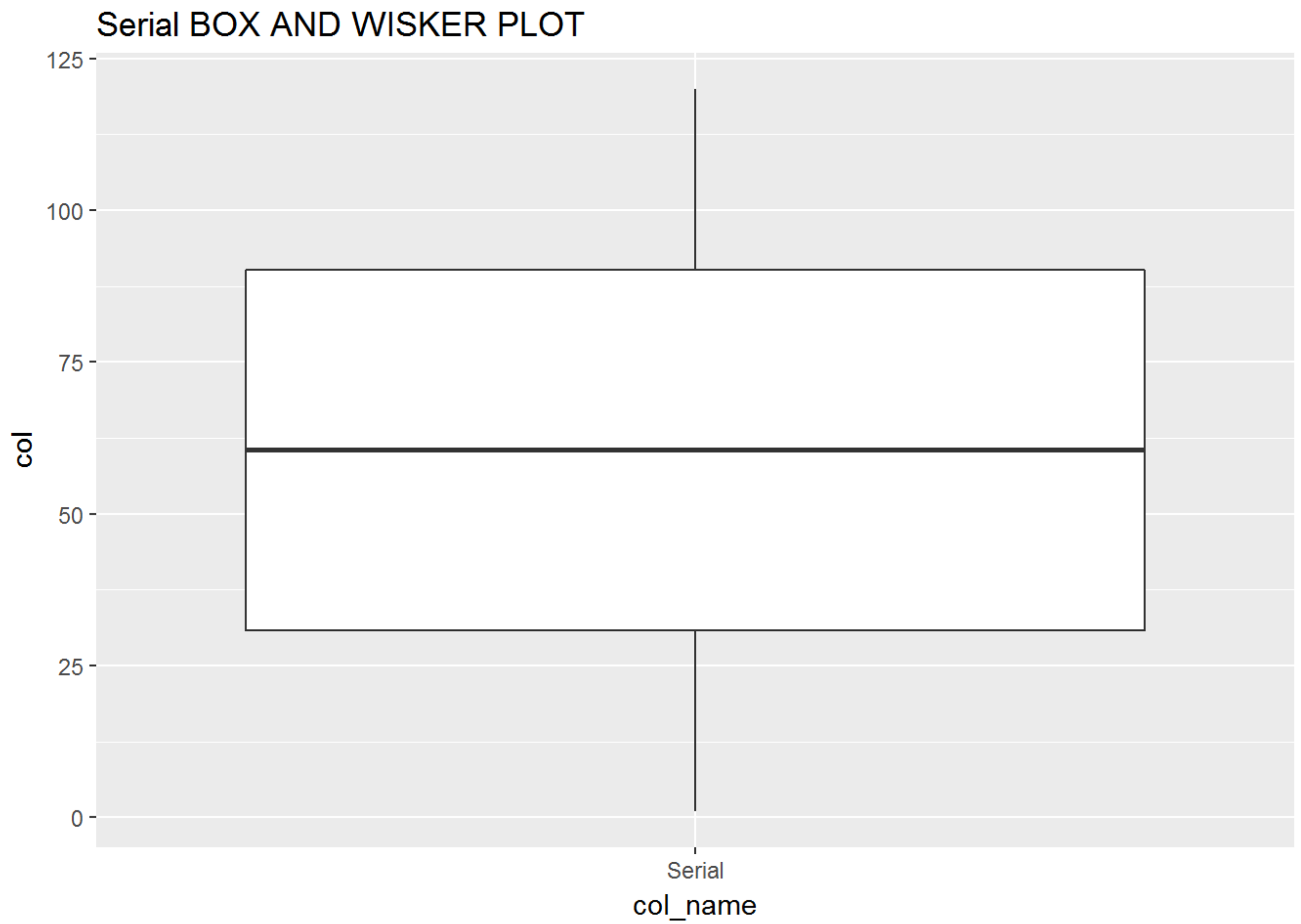
```
ggplot(platelet,aes(x=`Systolic blood presure`,y=`Diastolic blood presure`))+geom_point() +ggtitle("Systolic bloo
d presure VS Diastolic blood presure")
```



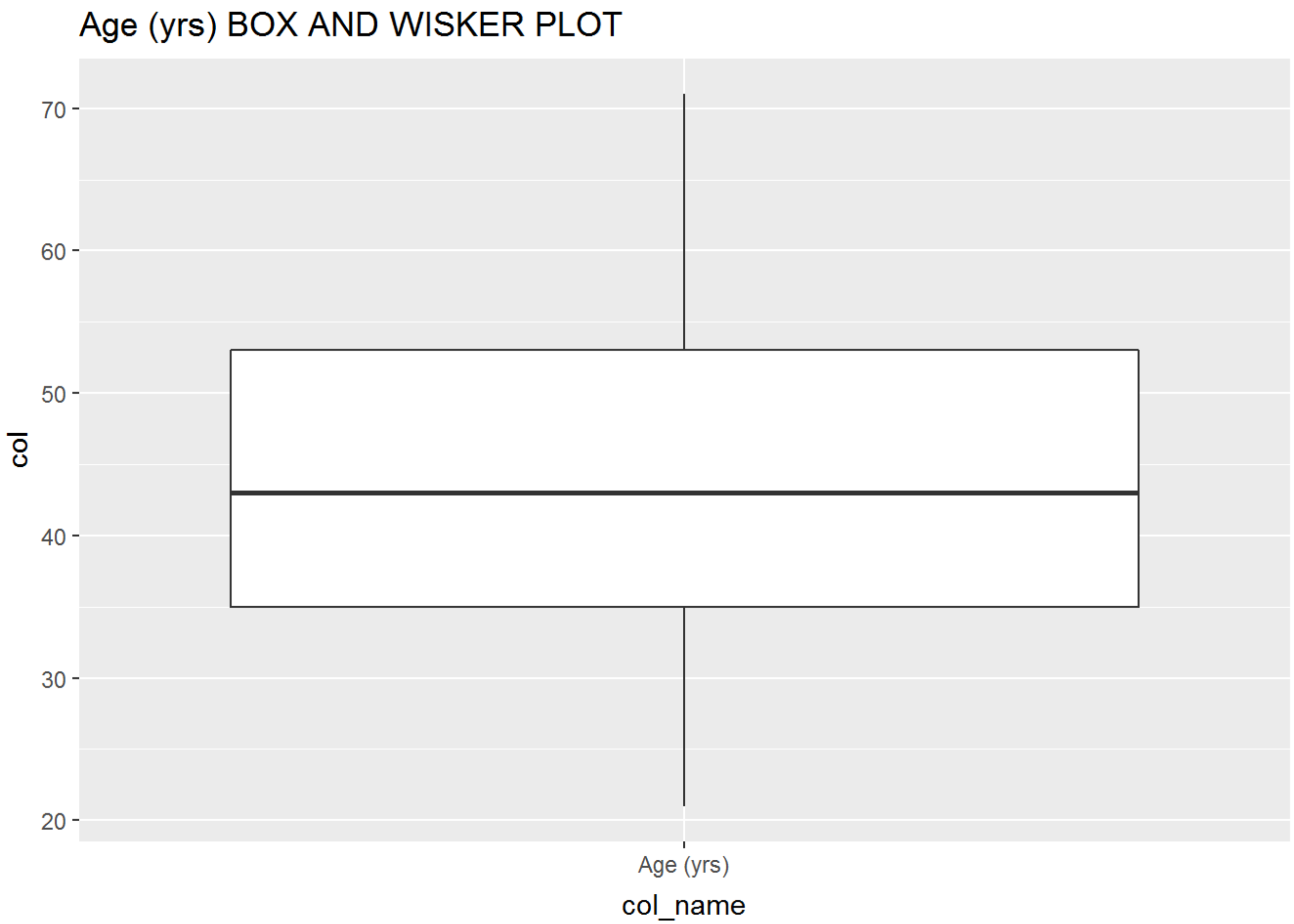
##Question 8 Present all the quantitative data using box-and Whisker plot

```
num_col <- sapply(platelet, is.numeric)

n=0
plot1 <- function(col) {
  n<<-n+1
  if(is.numeric(col))
  {
    col_name <- names(platelet)[n]
     ggplot(platelet,aes(x=col_name,y=col)) + geom_boxplot() +ggtitle(paste0(col_name," BOX AND WISKER PLOT"," "
))
  }
}
a<-sapply(platelet, plot1)
b<-sapply(a, function(x){!is.null(x)})
a[b]
```

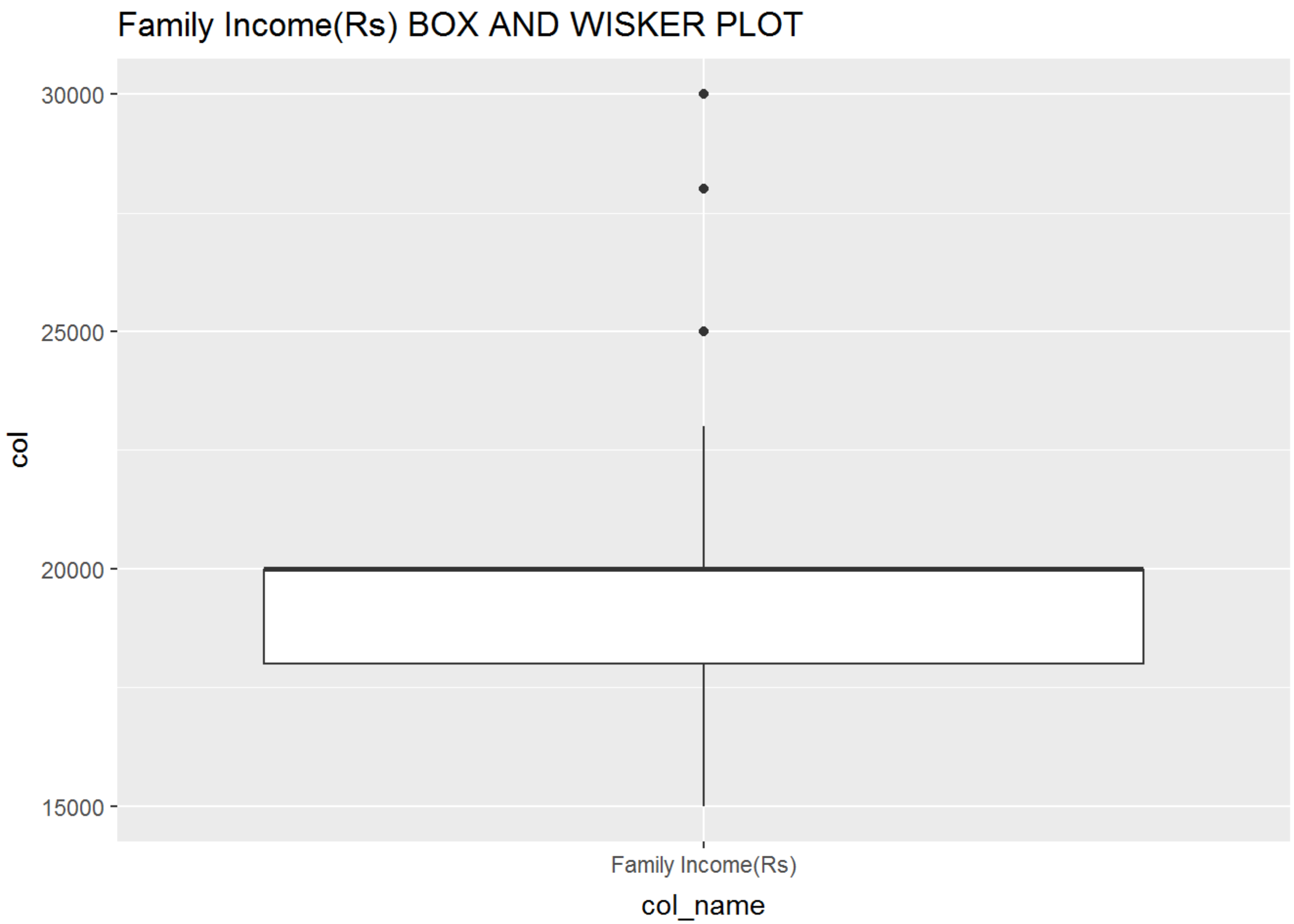
```
## $Serial
```


Serial BOX AND WISKER PLOT
125
100
75
50
25
0
col
Serial
col_name

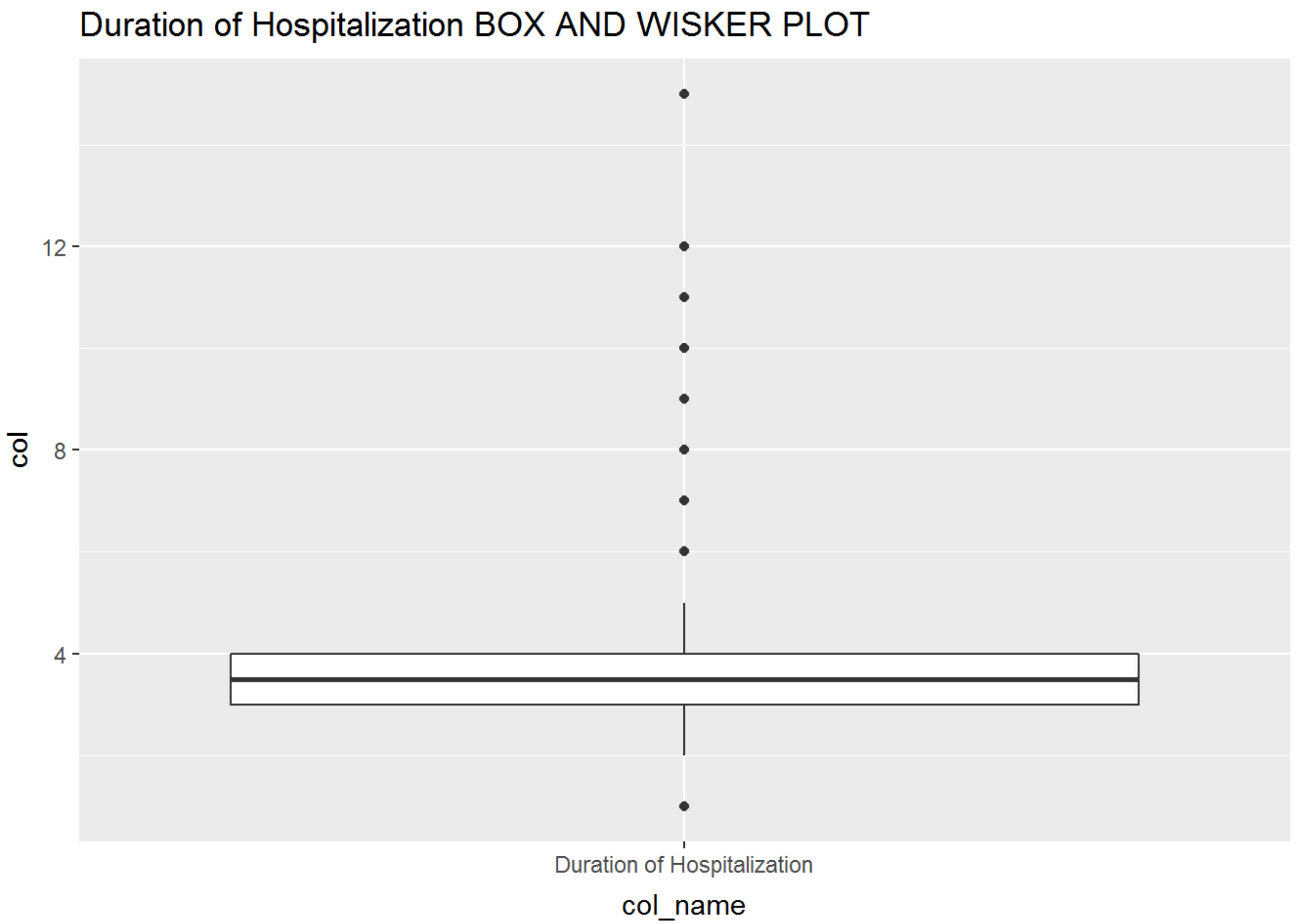
```
##
## $`Age (yrs)`
```



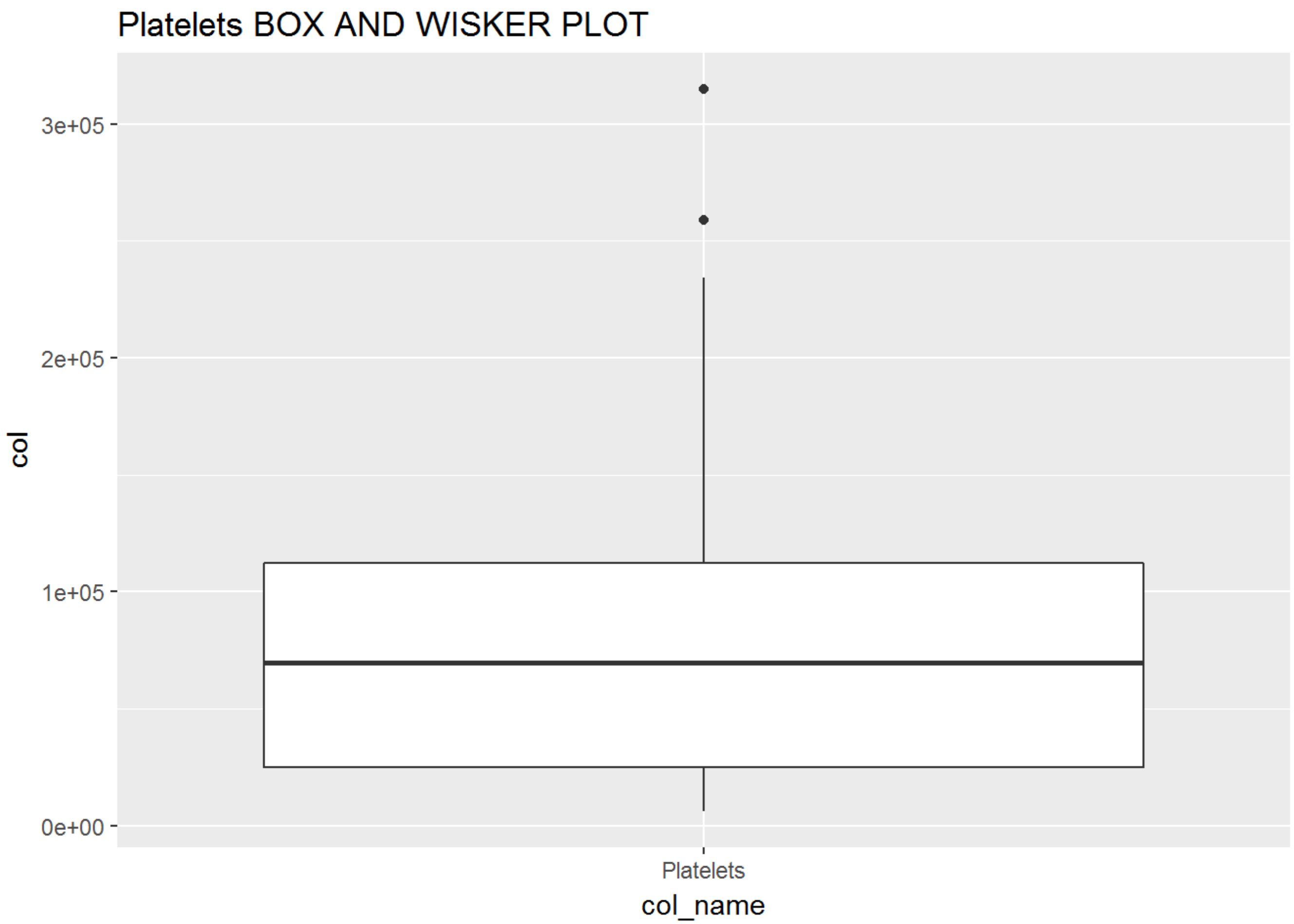
```
##
## $`Family Income(Rs)`
```



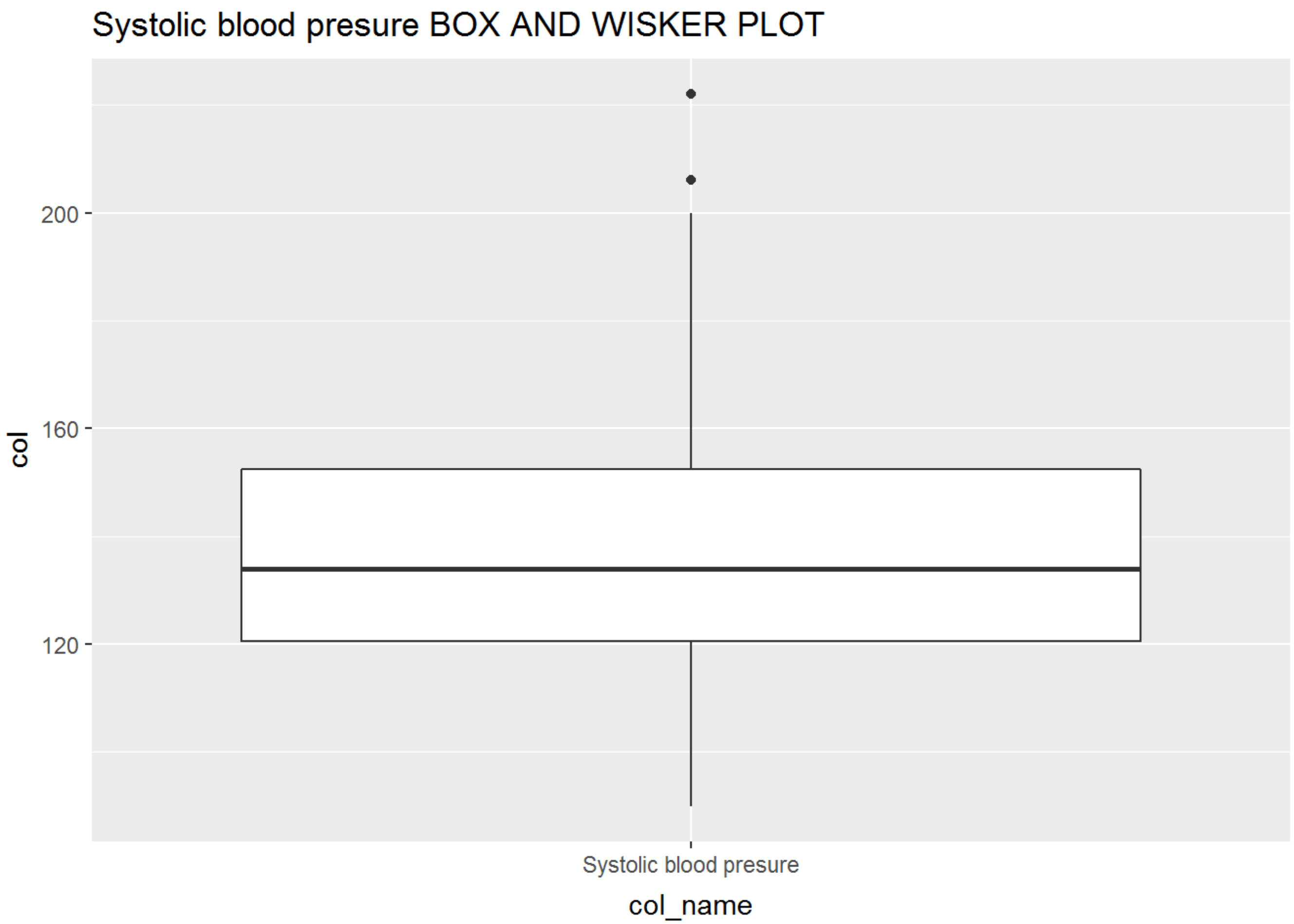
```
##
## $`Duration of Hospitalization`
```


Duration of Hospitalization BOX AND WISKER PLOT
col
12
8
4
Duration of Hospitalization
col_name

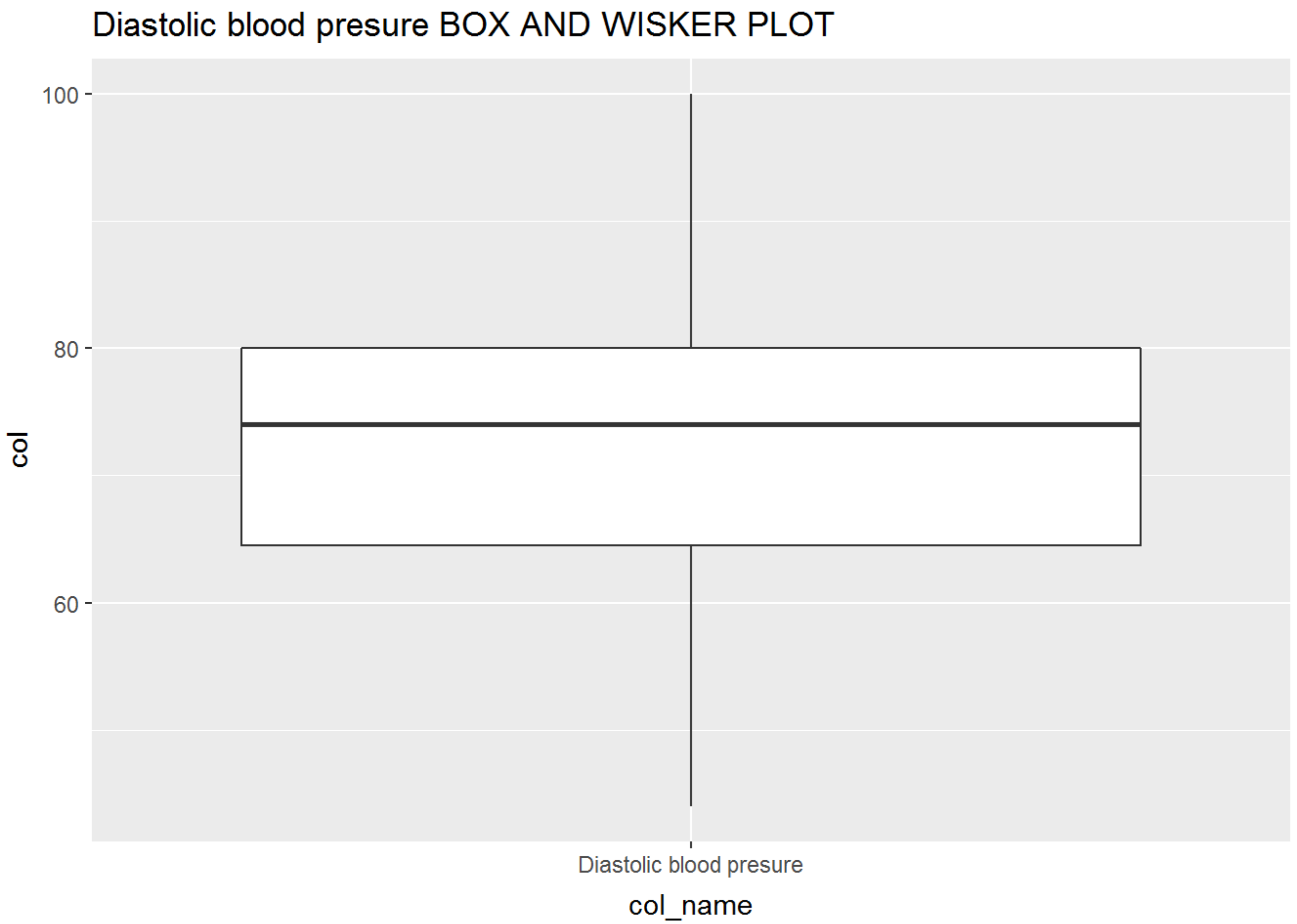
```
##
## $Platelets
```



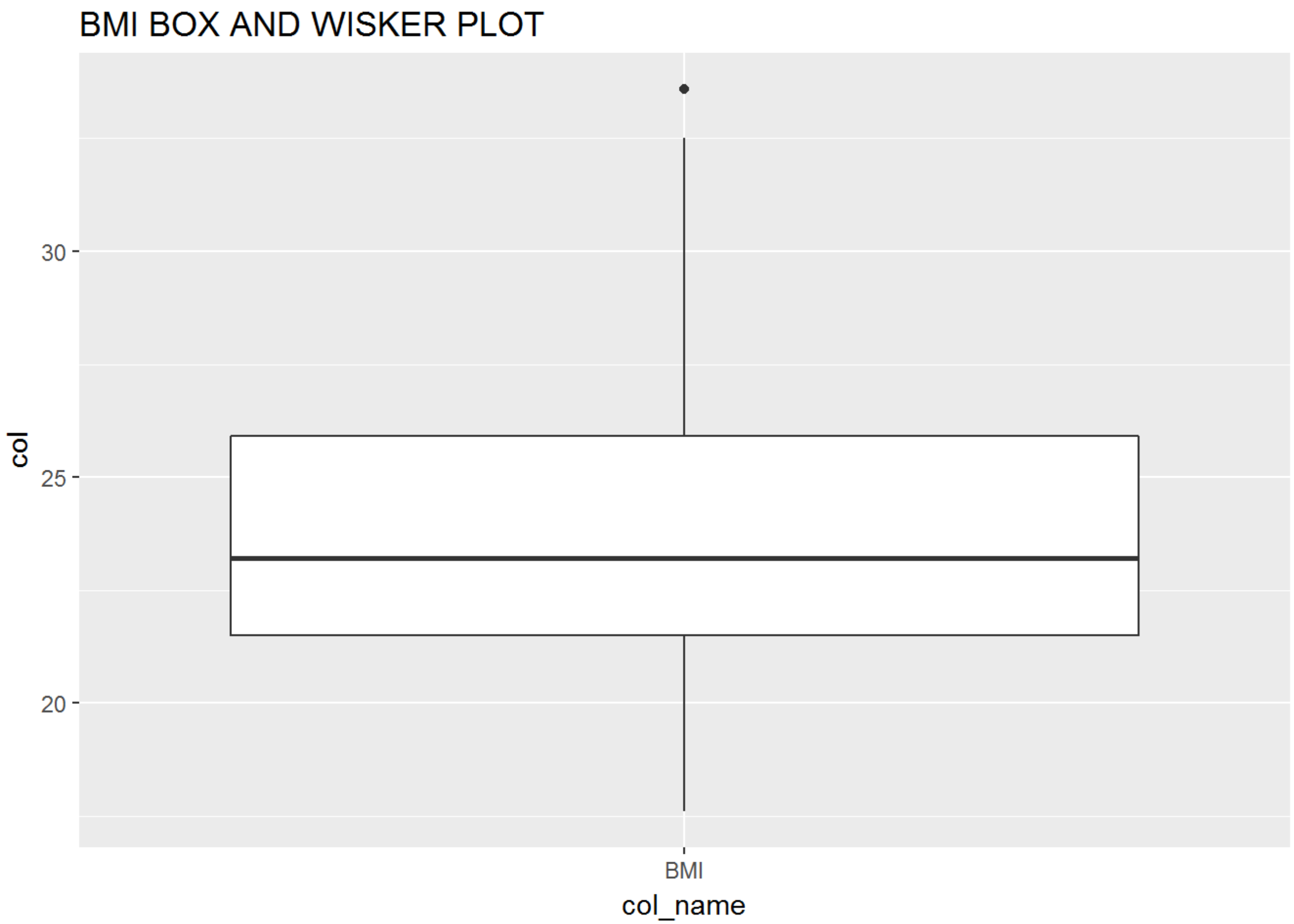
```
##
## $`Systolic blood presure`
```



```
##
## $`Diastolic blood presure`
```


Diastolic blood presure BOX AND WISKER PLOT
100
80
60
col
Diastolic blood presure
col_name

```
##
## $BMI
```



BMI BOX AND WISKER PLOT