Cardio-Fitness-Project.R

daoud

2020-03-11

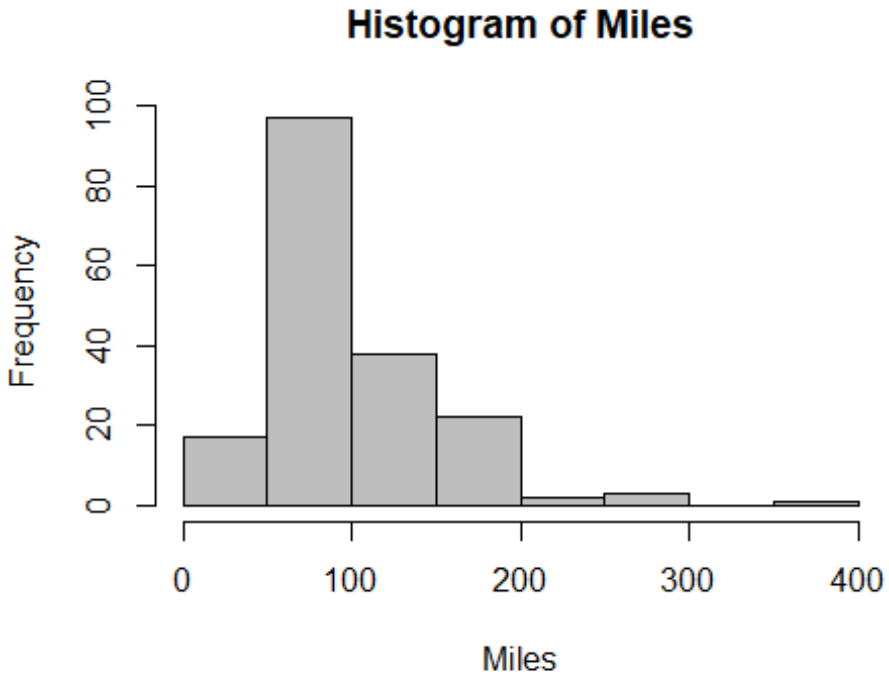
```
## ============================================================
## Cardio  Good  Fitness
## ============================================================

## packages
#install.packages("readr")
#install.packages("visdat")
#install.packages("dplyr")
#library(readr)
#library(visdat)
#library(dplyr)


## Set Working Directory

setwd("C:/Users/daoud/Downloads/PGP DSBA/Introduction to R/week 3")
getwd()

## [1] "C:/Users/daoud/Downloads/PGP DSBA/Introduction to R/week 3"

## read Input Data
goodFitness=read.csv("CardioGoodFitness.csv")

## view and explore Data

names(goodFitness)

## [1] "Product"       "Age"           "Gender"        "Education"
## [5] "MaritalStatus" "Usage"         "Fitness"       "Income"
## [9] "Miles"

dim(goodFitness)

## [1] 180   9

View(goodFitness)
summary(goodFitness)

##   Product        Age           Gender      Education       MaritalStatus
##  TM195:80   Min.   :18.00   Female: 76   Min.   :12.00   Partnered:107
##  TM498:60   1st Qu.:24.00   Male  :104   1st Qu.:14.00   Single   : 73
##  TM798:40   Median :26.00                Median :16.00
##             Mean   :28.79                Mean   :15.57
```

```
##             3rd Qu.:33.00                3rd Qu.:16.00
##             Max.   :50.00                Max.   :21.00
##      Usage          Fitness          Income           Miles
##  Min.   :2.000   Min.   :1.000   Min.   : 29562   Min.   : 21.0
##  1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 44059   1st Qu.: 66.0
##  Median :3.000   Median :3.000   Median : 50597   Median : 94.0
##  Mean   :3.456   Mean   :3.311   Mean   : 53720   Mean   :103.2
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 58668   3rd Qu.:114.8
##  Max.   :7.000   Max.   :5.000   Max.   :104581   Max.   :360.0
```

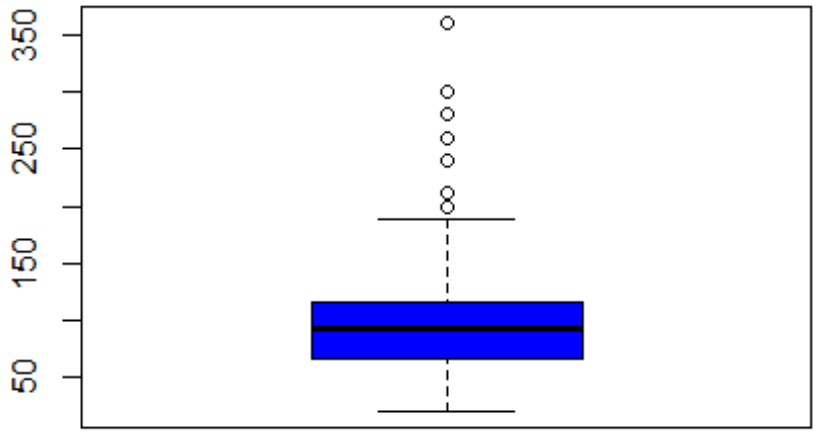
```
str(goodFitness)
```

```
## 'data.frame':    180 obs. of  9 variables:
##  $ Product      : Factor w/ 3 levels "TM195","TM498",..: 1 1 1 1 1 1 1 1 1
1 ...
##  $ Age          : int  18 19 19 19 20 20 21 21 21 21 ...
##  $ Gender       : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 1 1 2 2 1
...
##  $ Education    : int  14 15 14 12 13 14 14 13 15 15 ...
##  $ MaritalStatus: Factor w/ 2 levels "Partnered","Single": 2 2 1 2 1 1 1 2
2 1 ...
##  $ Usage        : int  3 2 4 3 4 3 3 3 5 2 ...
##  $ Fitness      : int  4 3 3 3 2 3 3 3 4 3 ...
##  $ Income       : int  29562 31836 30699 32973 35247 32973 35247 32973
35247 37521 ...
##  $ Miles        : int  112 75 66 85 47 66 75 85 141 85 ...
```

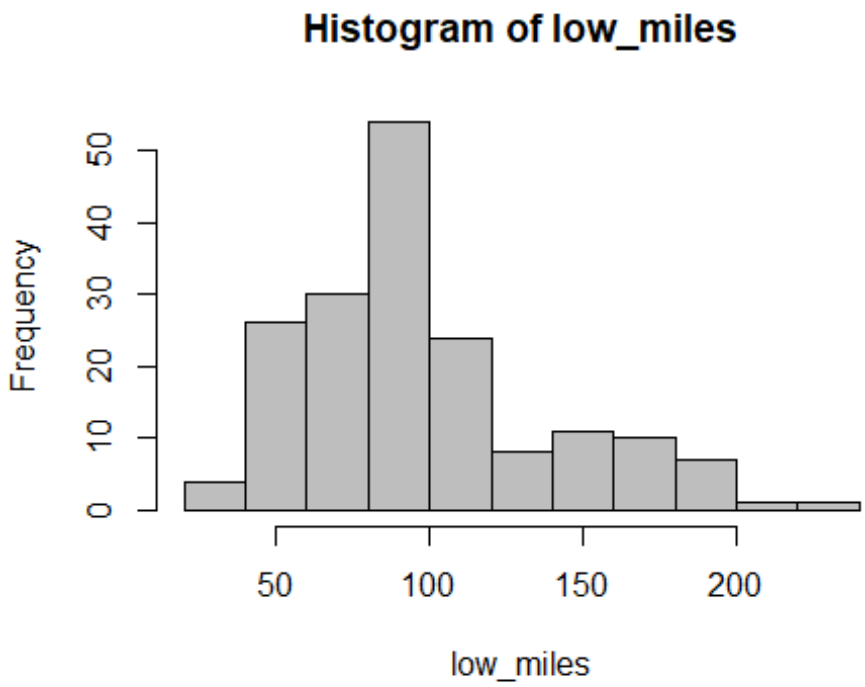
```
## OBSERVATIONS:
# 1. Deppendent variable : Miles
# 2. all indeppendent variable are integer except : Product,Gender and
MaritalStatus .
# 3. No missing values on data

## examine Miles variable :

attach(goodFitness)

hist(Miles,col = 'grey')
```



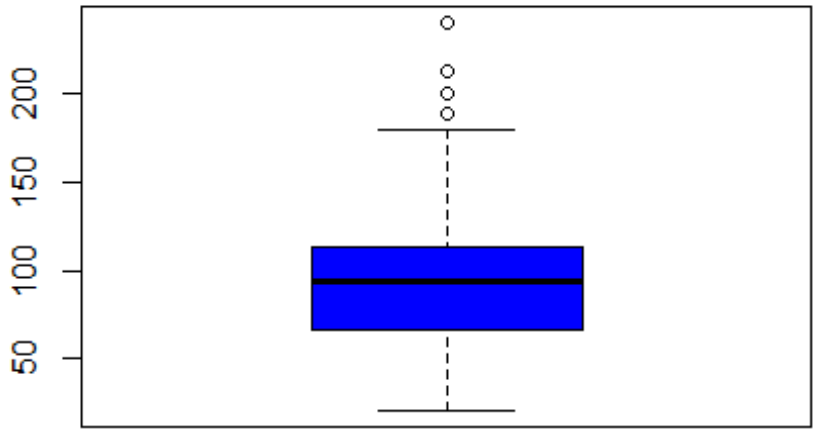

```r
# possiply outlier effacting histogram
boxplot(Miles,col = 'blue')
```



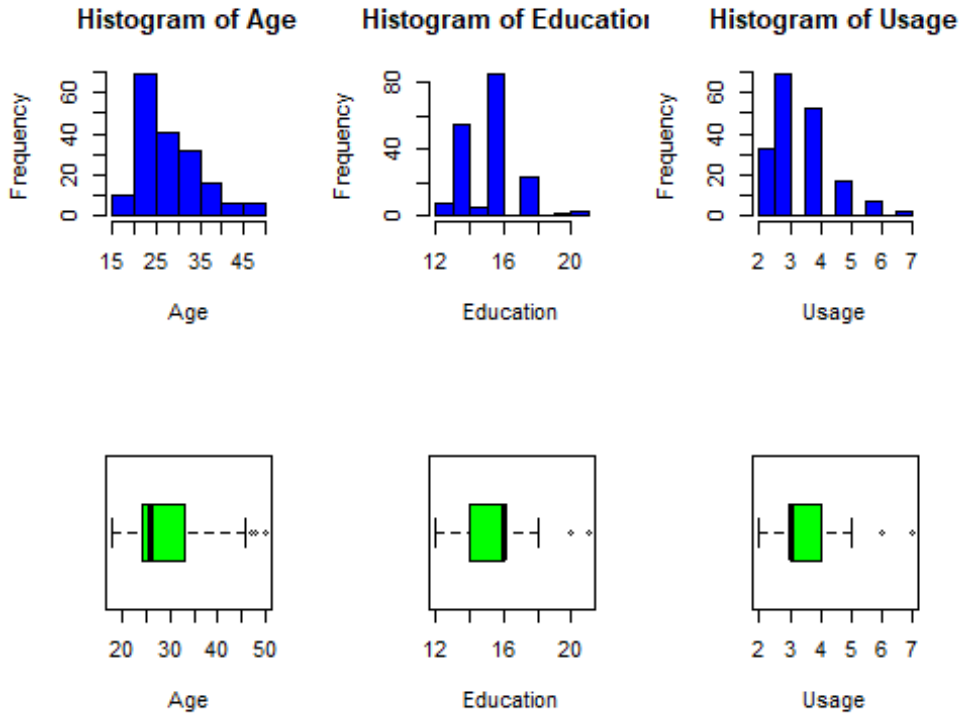
```r
## few outliers showed on the boxplot may effact

## let us examine Miles low then < 250


low_miles=Miles[Miles<=250]

hist(low_miles,col = 'grey')
```



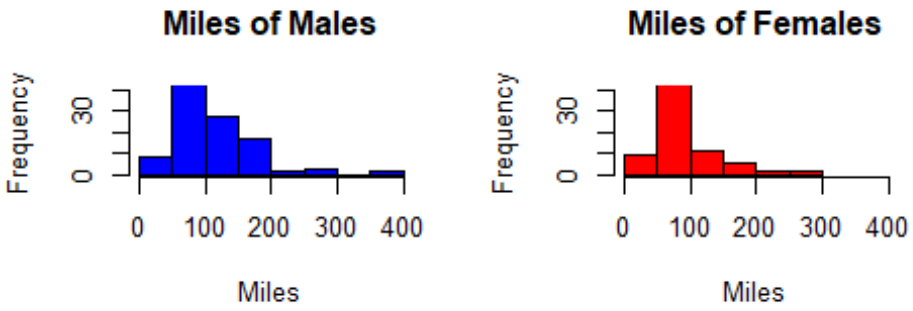

```r
boxplot(low_miles,col = 'blue')
```

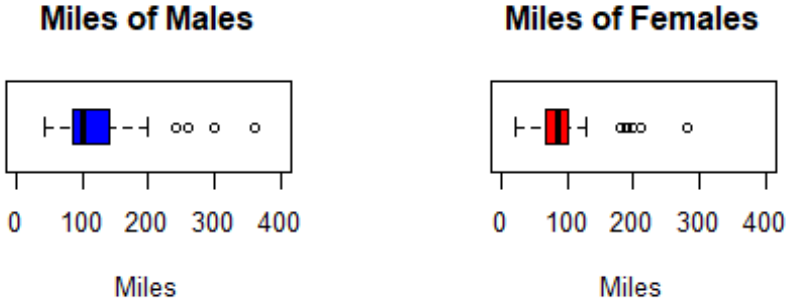



```
## OBSERVATIONS :
# number of OBS reduce from 180 to 176 , there was 4 obs above 250 , it
doesn't seem too much differnt on the plots
# lets examine indeppendent variable with original data

# let us use the original dataset

par(mfrow=c(2,3))
hist(Age,col = "blue",xlab = "Age")
hist(Education,col = "blue",xlab = "Education")
hist(Usage,col = "blue",xlab = "Usage")
boxplot(Age,horizontal = TRUE,col = "green",xlab = "Age")
boxplot(Education,horizontal = TRUE,col = "green",xlab = "Education")
boxplot(Usage,horizontal = TRUE,col = "green",xlab = "Usage")
```



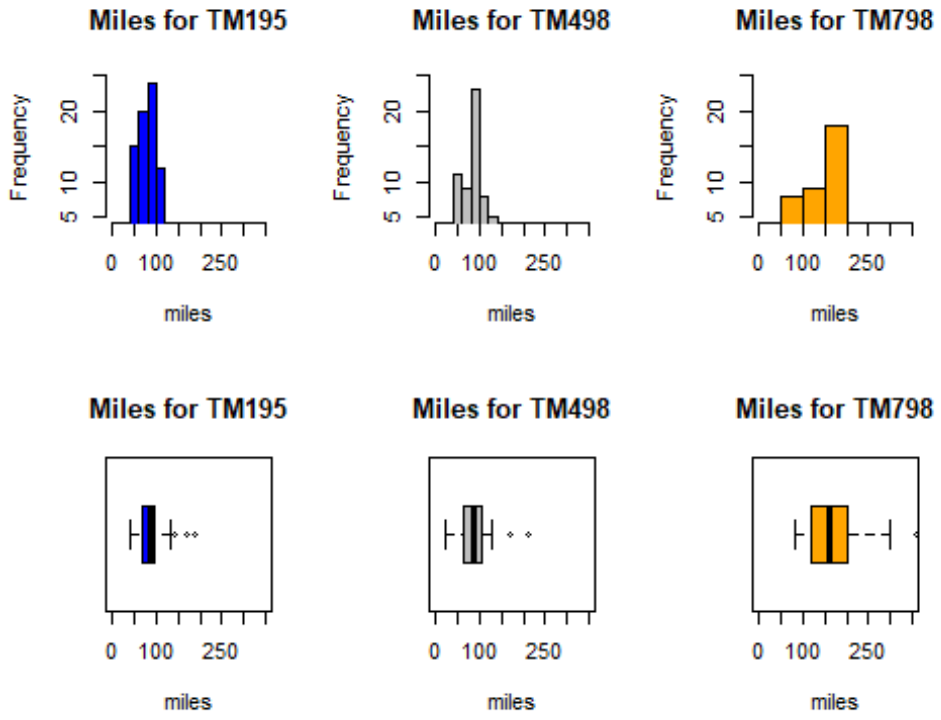

```
## OBSERVATIONS:
# Age histogram shows the Right skew in distribution on average 20 to 25
years, means the majorty early adulthood
# Education histogram shows the symmetric distribution 15 to 16 years, means
the majorty are in Secondary school
# Usage histogram shows the Right skew in distribution on average 3 time a
week

## Miles VS Gender

# note: we use xlim and ylim to present all plot with same XY for easy
comparsion .
par(mfrow=c(2,2))
hist(Miles[Gender=='Male'],main='Miles of Males',xlab='Miles',col =
'blue',xlim = c(0,400) ,ylim = c(0,40))
hist(Miles[Gender=='Female'],main='Miles of Females',xlab='Miles',col =
'red',xlim = c(0,400) ,ylim = c(0,40))
boxplot(Miles[Gender=='Male'],main='Miles of Males',xlab='Miles',horizontal =
TRUE,col = 'blue',ylim = c(0,400))
boxplot(Miles[Gender=='Female'],main='Miles of
Females',xlab='Miles',horizontal = TRUE,col = 'red',ylim = c(0,400))
```






```
summary(Product[Gender=='Male'])

## TM195 TM498 TM798
##    40    31    33

summary(Product[Gender=='Female'])

## TM195 TM498 TM798
##    40    29     7

## OBSERVATIONS:
# 1. we obs from boxplot that Males running longer distance then Fimale.
# 2. we obs from summary that Male use TM798 more then Female by 33 to 7, but
the same quantity for TM195 and TM498.
# 3. maybe TM798 NOT suitable for Female ? or maybe TM798 is very expensive ?

## Miles VS Product

summary(Product)

## TM195 TM498 TM798
##    80    60    40

par(mfrow=c(2,3))
hist(Miles[Product=='TM195'],main='Miles for TM195',xlab='miles',col =
'blue',xlim = c(0,350),ylim = c(5,25))
hist(Miles[Product=='TM498'],main='Miles for TM498',xlab='miles',col =
'grey',xlim = c(0,350),ylim = c(5,25))
```

```
hist(Miles[Product=='TM798'],main='Miles for TM798',xlab='miles',col =
'orange',xlim = c(0,350),ylim = c(5,25))
boxplot(Miles[Product=='TM195'],main='Miles for TM195',xlab='miles',col =
'blue',horizontal = TRUE,ylim = c(0,350))
boxplot(Miles[Product=='TM498'],main='Miles for TM498',xlab='miles',col =
'grey',horizontal = TRUE,ylim = c(0,350))
boxplot(Miles[Product=='TM798'],main='Miles for TM798',xlab='miles',col =
'orange',horizontal = TRUE,ylim = c(0,350))
```

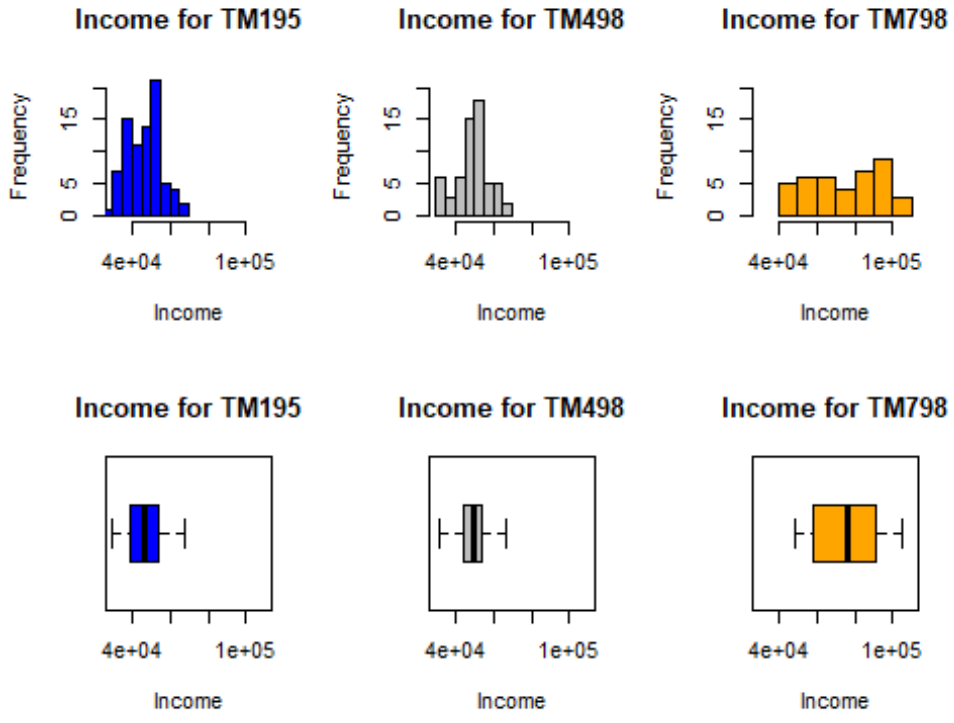



```
# mean Income VS Product VS Gender
mean(Income[Product=='TM195'&Gender=='Female'])

## [1] 46020.07

mean(Income[Product=='TM195'&Gender=='Male'])

## [1] 46815.97

mean(Income[Product=='TM498'&Gender=='Female'])

## [1] 49336.45

mean(Income[Product=='TM498'&Gender=='Male'])

## [1] 48634.26

mean(Income[Product=='TM798'&Gender=='Female'])

## [1] 73633.86
```

```r
mean(Income[Product=='TM798'&Gender=='Male'])
```

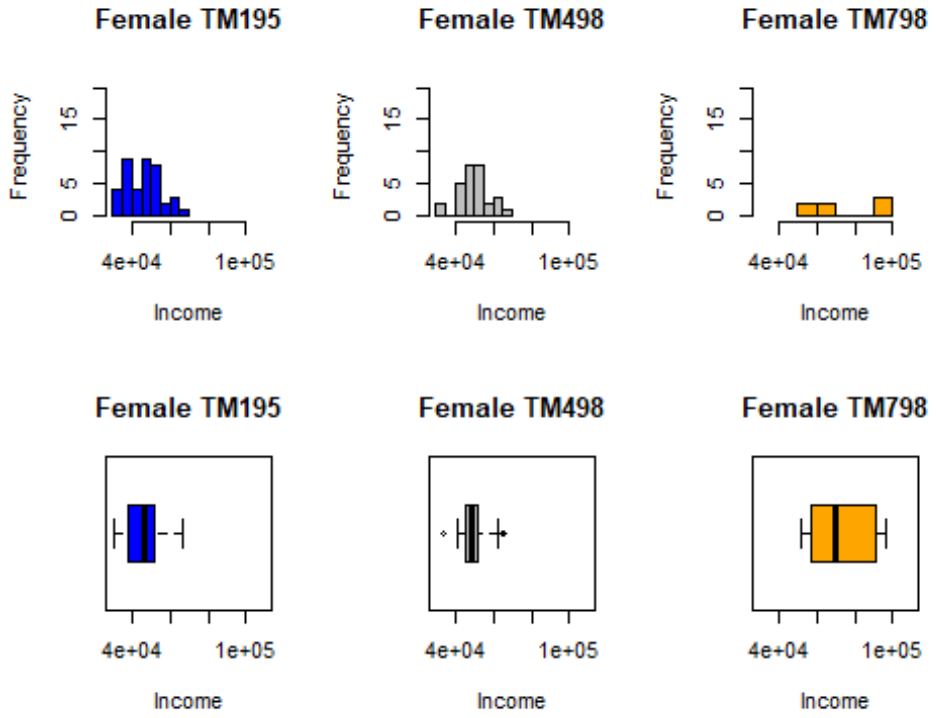
```
## [1] 75825.03
```

```r
# plot for Income VS Product :
par(mfrow=c(2,3))
hist(Income[Product=='TM195'],main='Income for TM195',xlab='Income',col =
'blue',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM498'],main='Income for TM498',xlab='Income',col =
'grey',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM798'],main='Income for TM798',xlab='Income',col =
'orange',xlim = c(30000,110000),ylim = c(0,22))
boxplot(Income[Product=='TM195'],main='Income for TM195',xlab='Income',col =
'blue',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM498'],main='Income for TM498',xlab='Income',col =
'grey',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM798'],main='Income for TM798',xlab='Income',col =
'orange',horizontal = TRUE,ylim=c(30000,110000))
```

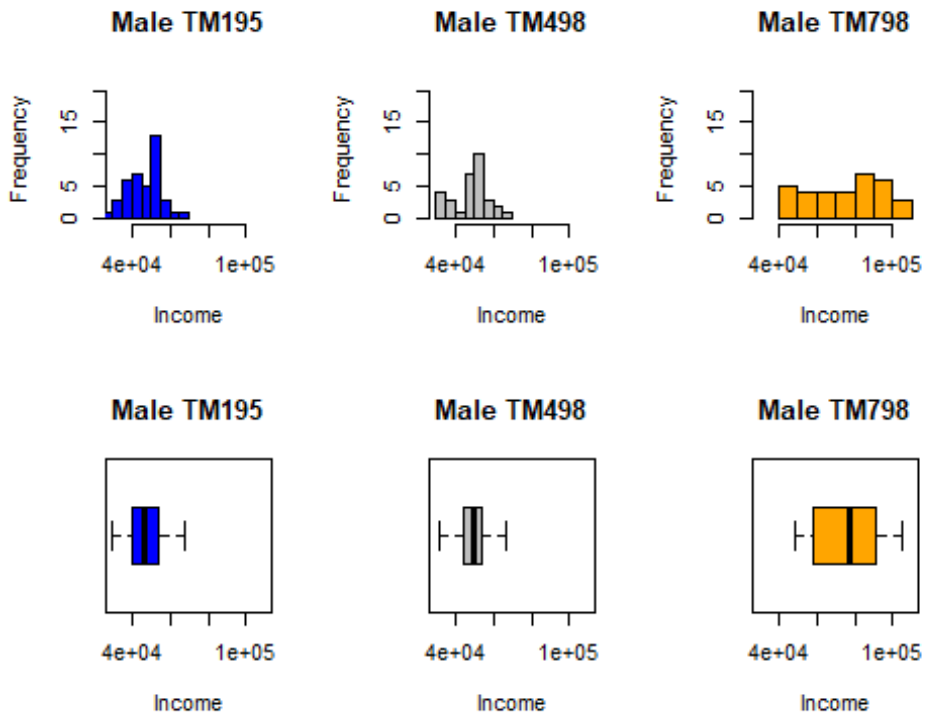



```r
# plot for Incom VS Produt for Female :
par(mfrow=c(2,3))
hist(Income[Product=='TM195'&Gender=='Female'],main='Female
TM195',xlab='Income',col = 'blue',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM498'&Gender=='Female'],main='Female
TM498',xlab='Income',col = 'grey',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM798'&Gender=='Female'],main='Female
TM798',xlab='Income',col = 'orange',xlim = c(30000,110000),ylim = c(0,22))
```

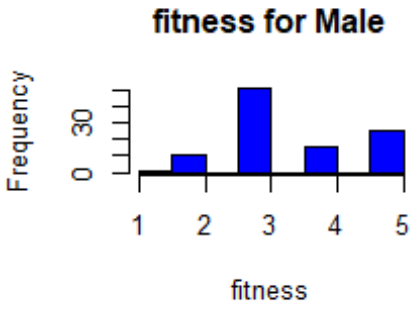
```r
boxplot(Income[Product=='TM195'&Gender=='Female'],main='Female
TM195',xlab='Income',col = 'blue',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM498'&Gender=='Female'],main='Female
TM498',xlab='Income',col = 'grey',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM798'&Gender=='Female'],main='Female
TM798',xlab='Income',col = 'orange',horizontal = TRUE,ylim=c(30000,110000))
```

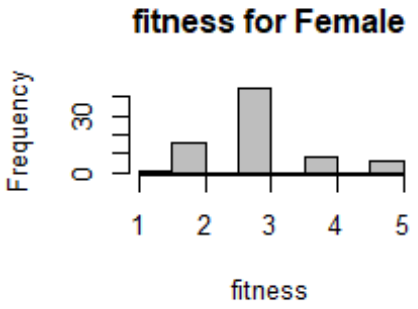


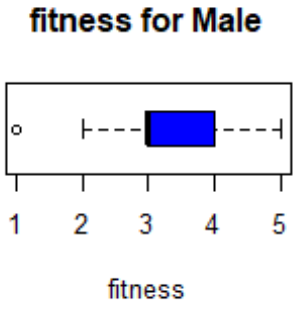

```r
# plot for Incom VS Produt for Male :
par(mfrow=c(2,3))
hist(Income[Product=='TM195'&Gender=='Male'],main='Male
TM195',xlab='Income',col = 'blue',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM498'&Gender=='Male'],main='Male
TM498',xlab='Income',col = 'grey',xlim = c(30000,110000),ylim = c(0,22))
hist(Income[Product=='TM798'&Gender=='Male'],main='Male
TM798',xlab='Income',col = 'orange',xlim = c(30000,110000),ylim = c(0,22))
boxplot(Income[Product=='TM195'&Gender=='Male'],main='Male
TM195',xlab='Income',col = 'blue',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM498'&Gender=='Male'],main='Male
TM498',xlab='Income',col = 'grey',horizontal = TRUE,ylim=c(30000,110000))
boxplot(Income[Product=='TM798'&Gender=='Male'],main='Male
TM798',xlab='Income',col = 'orange',horizontal = TRUE,ylim=c(30000,110000))
```

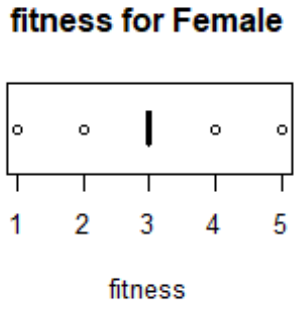



```
## OBSERVATIONS:
# 1. we obs that TM195 is the most Product consumed with 80 of 180 .
# 2. but when we compare Product VS Miles on Boxplot , we obs that biggest
Miles distance of runner are using TM798,
# 3. the Income very approach for Male and Female .so we Drop out the
prospect of TM798 being expensive for Female .

## Fitness VS Gender

par(mfrow=c(2,2))
hist(Fitness[Gender=='Male'],xlab='fitness',main='fitness for Male',col =
'blue')
hist(Fitness[Gender=='Female'],xlab='fitness',main='fitness for Female',col =
'grey')
boxplot(Fitness[Gender=='Male'],xlab='fitness',main='fitness for Male',col =
'blue',horizontal = TRUE)
boxplot(Fitness[Gender=='Female'],xlab='fitness',main='fitness for
Female',col = 'grey',horizontal = TRUE)
```









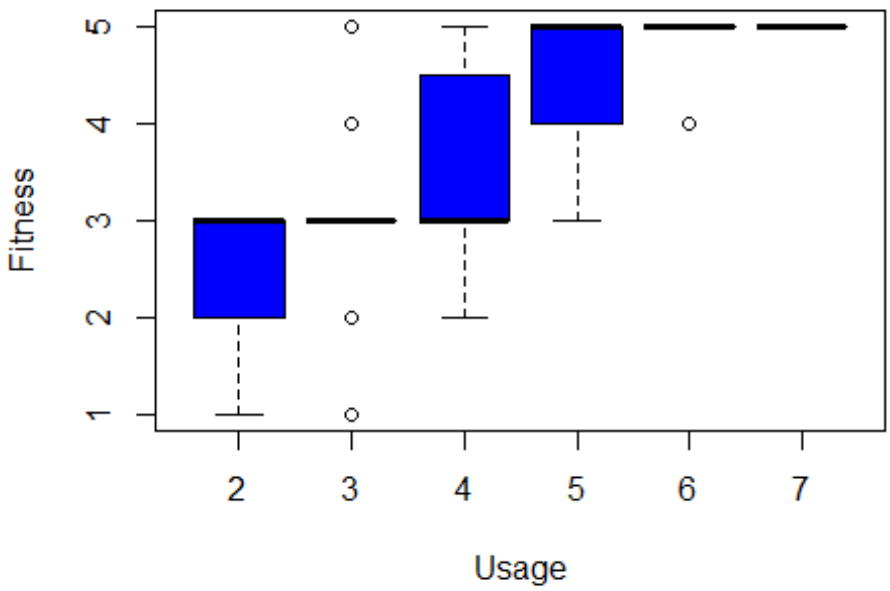

```r
sum(Fitness[Gender=='Female'])
```

```
## [1] 230
```

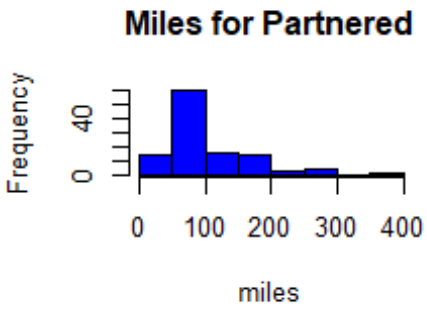
```r
sum(Fitness[Gender=='Male'])
```

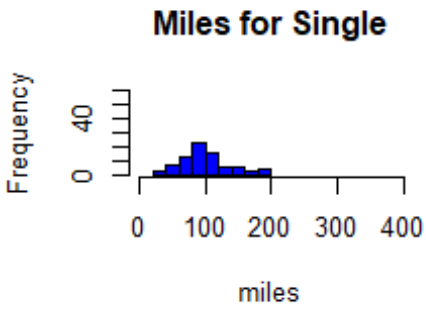
```
## [1] 366
```

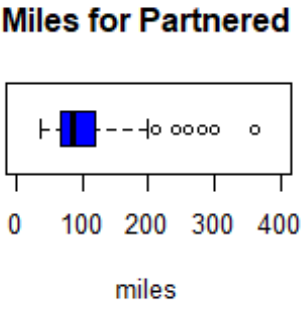
```r
par(mfrow=c(1,1))
boxplot(Fitness~Usage,data = goodFitness ,col='blue')
```

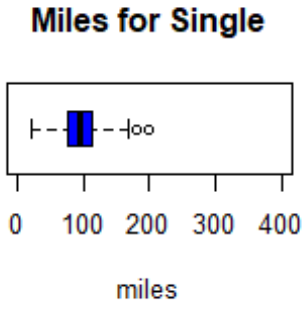



```
## OBSERVATIONS:
# 1. the Normal Fitness 3 is more often, Fitness distribut to Male by 366 but
Female by 230  .
# 2. hight possibility of outlier for Female Fitness , the more Product use
the better you Fitness is.

## Miles VS MaritalStatus:

summary(MaritalStatus)

## Partnered    Single
##       107        73

par(mfrow=c(2,2))
hist(Miles[MaritalStatus=='Partnered'],xlab='miles',main='Miles for
Partnered',col = 'blue',xlim = c(0,400),ylim = c(0,60))
hist(Miles[MaritalStatus=='Single'],xlab='miles',main='Miles for Single',col
= 'blue',xlim = c(0,400),ylim = c(0,60))
boxplot(Miles[MaritalStatus=='Partnered'],xlab='miles',main='Miles for
Partnered',col = 'blue',horizontal = TRUE,ylim = c(0,400))
boxplot(Miles[MaritalStatus=='Single'],xlab='miles',main='Miles for
Single',col = 'blue',horizontal = TRUE,ylim = c(0,400))
```










```
sum(Miles[MaritalStatus=='Partnered' & Gender=='Male' ])
```

```
## [1] 6866
```

```
sum(Miles[MaritalStatus=='Partnered'&Gender=='Female'])
```

```
## [1] 4293
```

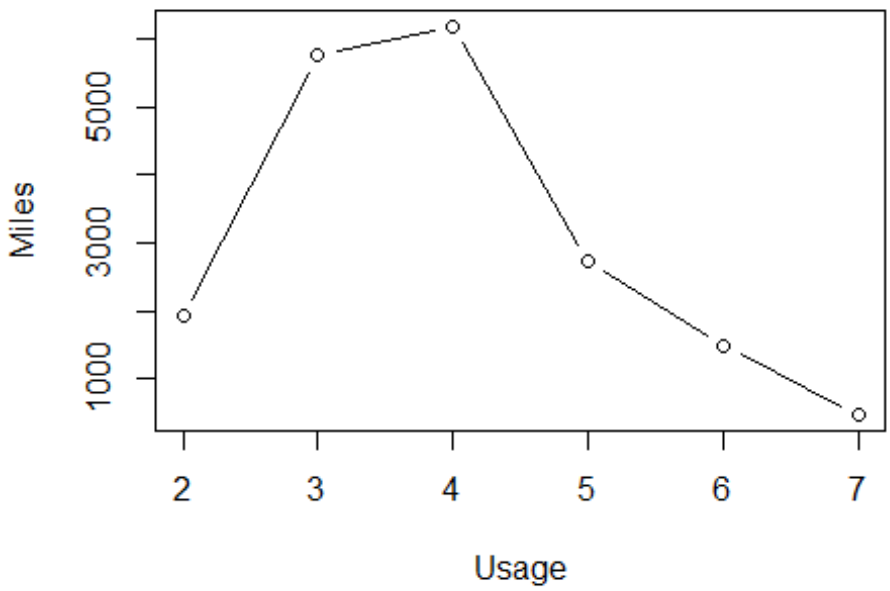
```
sum(Miles[MaritalStatus=='Single' & Gender=='Male' ])
```

```
## [1] 4868
```

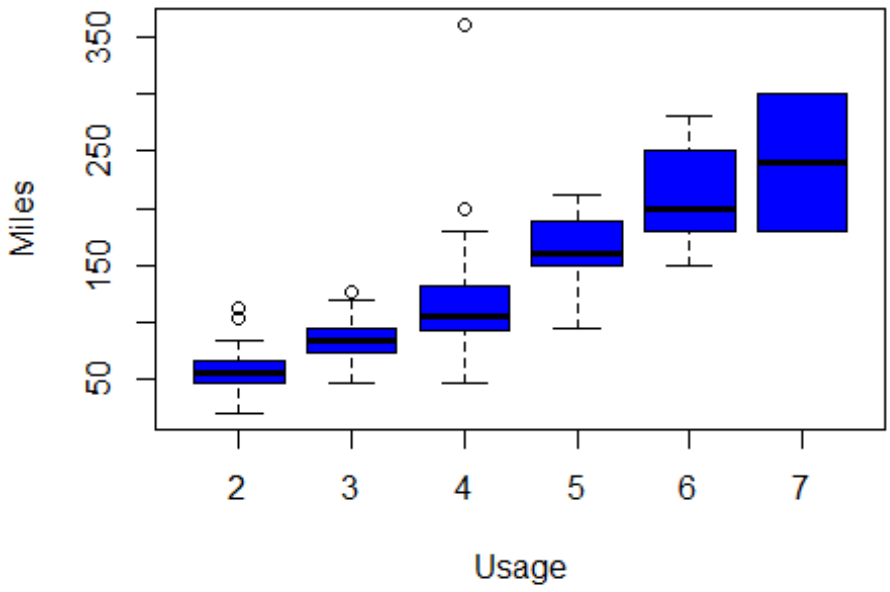
```
sum(Miles[MaritalStatus=='Single'&Gender=='Female'])
```

```
## [1] 2548
```

```
## OBSERVATIONS:
# 1. we obs that Partnered run 11159 more then Single .
# 2. Male with Partnered run more distance then Female.
# 3. Single Male run Double distance then Female, maybe Female need courage
by Partnered ?

## Miles VS Usage:

par(mfrow=c(1,1))
plot(aggregate(Miles~Usage,data=goodFitness, sum), type="b")
```




```
boxplot(Miles~Usage,data = goodFitness ,col='blue')
```




```
summary(Miles[Usage=='4'])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    47.0    94.0   106.0   118.9   132.0   360.0

## OBSERVATIONS:
# 1. we obs from the plot that usage 4 have the biggest running Miles
# 2. clearly there is outlier on Usage 4, we can obs hight different between
3IQR and Outlier, 132 to 360
# 3. from boxplot we obs the more you use Product the more Miles you run .


##
**************************************************************************
**********************
##  CONCLUSIONS:
##
**************************************************************************
**********************
# 1. young people  are using the Product more often
# 2. Female with Partnered run better then Single .
# 3. the more Product Usage the better cardio Fitnees you get and the more
Miles you Run the healthiest you be
# 4. DATA IS NOW READY FOR MODEL BUILDING!!!
```