Mini Project - Cardio Good Fitness

* Ajay K M

Table of Contents

1 Project Objective ............................................................................................................................. 3

2 Assumptions .................................................................................................................................... 3

3 Exploratory Data Analysis – Step by step approach ....................................................................... 3

3.1 Environment Set up and Data Import ..................................................................................... 4

3.1.1 Install necessary Packages and Invoke Libraries ............................................................. 4

3.1.2 Set up working Directory ................................................................................................ 4

3.1.3 Import and Read the Dataset .......................................................................................... 4

3.2 Variable Identification ............................................................................................................. 4

3.2.1 Variable Identification – Inferences ................................................................................ 4

3.3 Univariate Analysis .................................................................................................................. 4

3.4 Bi-Variate Analysis................................................................................................................... 6

3.5 Missing Value Identification .................................................................................................... 9

3.6 Outlier Identification ............................................................................................................... 9

3.7 Variable Transformation / Feature Creation .......................................................................... 12

4 Conclusion ....................................................................................................................................... 14

5 Appendix A – Source Code ..............................................................................................................15

1. Project Objective :

The objective of the report is to explore the cardio data set (“CardioGoodFitness”) in R and generate insights about the data set. This exploration report will consists of the following:

 Importing the dataset in R

Understanding the structure of dataset

 Graphical exploration

 Descriptive statistics

Insights from the dataset

1. Assumptions :

We have this Usage column in our dataset. I have made this assumption that the usage of it is on daily basis.

1. Exploratory Data Analysis – Step by step approach

1. Environment Set up and Data Import

2. Variable Identification

3. Univariate Analysis

4. Bi-Variate Analysis

5. Missing Value Treatment (Not in scope for our project)

6. Outlier Treatment (Not in scope for our project)

7. Variable Transformation / Feature Creation

8. Feature Exploration

We shall follow these steps in exploring the provided dataset.

Although Steps 5 and 6 are not in scope for this project, a brief about these steps (and other steps as well) is given, as these are important steps for Data Exploration journey.

3.1: Environment Set up and Data Import :

3.1.1 : Install necessary Packages and Invoke Libraries :

install.packages("ggplot2")

library(“ggplot2”)

We are using ggplot2 for making beautiful graphics. ggplot2 lets you use the [grammar of graphics](http://vita.had.co.nz/papers/layered-grammar.pdf) to build layered and customizable plots.

3.1.2 : Set up working Directory

Setting a working directory on starting of the R session makes importing and exporting data files and code files easier. Basically, working directory is the location/ folder on the PC where you have the data, codes etc. related to the project.

Please refer Appendix A for Source Code.

3.1.3 : Import and Read the Dataset

The given dataset is in .csv format. Hence, the command ‘read.csv’ is used for importing the file.

Please refer Appendix A for Source Code.

3.2 : Variable Identification

*3.2.1 :* Variable Identification – Inferences

*Out of 9 variables -Income, Miles are continues variables.*

*Age, education, Marital Status, gender, Usage, Fitness and Product are discrete variables*

*Different functions used :*

*dim – Dimensions. I.e used for* checking the total number of rows and columns.

*Names – Gives the column names in our dataset*

*Str – Basic one line structure is displayed.*

*Summary – Gives summary statistics including the minimum and maximum values, mean, standard deviation, range, and percentiles.*

*This also tells if there are any missing values in our dataset.*

*head – gives first 6 rows of our dataset*

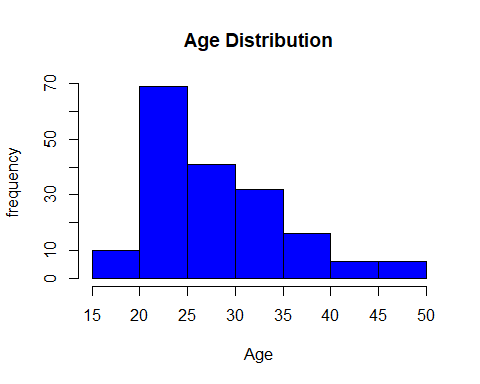
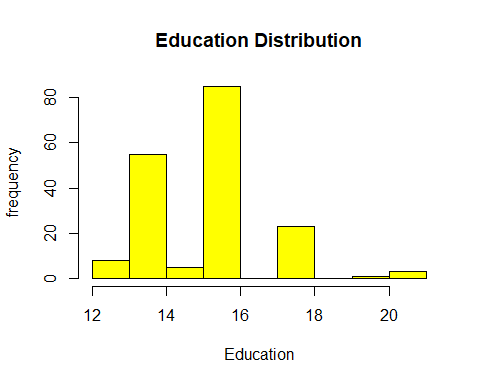
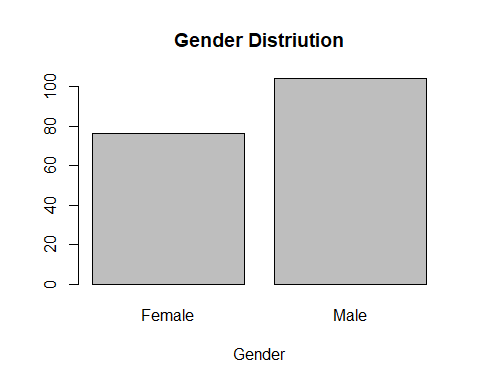
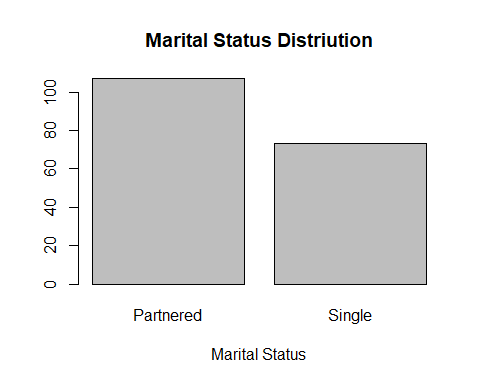
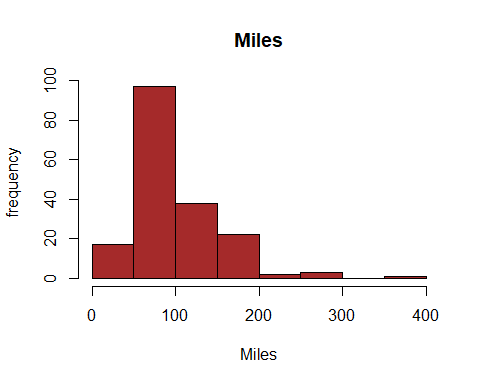
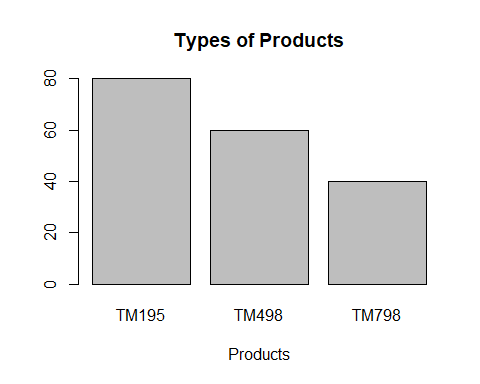
*tail- gives last 6 rows of our dataset.*

*3.3 :* Univariate Analysis

*Lets explore individual variables one by one :*

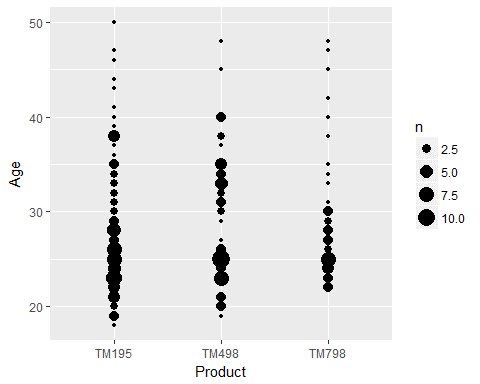
|  |  |
| --- | --- |
| Input Variables | Observations |
| Income | We see more income values lies in 40k-60k range |
| Age | More number of youngsters i.e 20-30 aged people are more in our dataset. |
| Education | We can see Education -14 and 16 are high in our distribution. |
| Miles | From the plot you see a majority of them belong to 50-100 miles group |
| Marital Status | There are more number of partnered people (around 110) than the single ones (around 70). |
| Gender | More no. of Males (around 100) than females (around 70) in our dataset. |
| Products | We have 3 types of products - TM195, TM498 and TM798. Out of which we have more products  of TM195 (80) and less of TM798 (40). Rest of 60 is of TM498. |

Graphs :

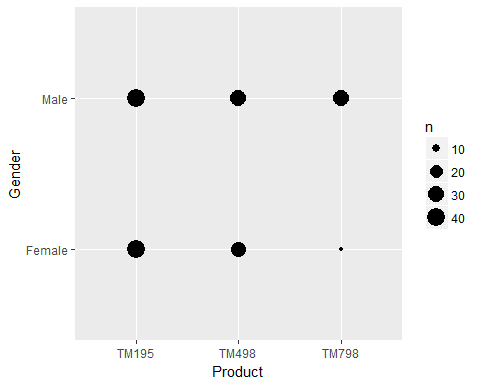
*3.4 :* Bi-Variate Analysis : Lets analyze 2 variables

1. *Product vs Age -*

**

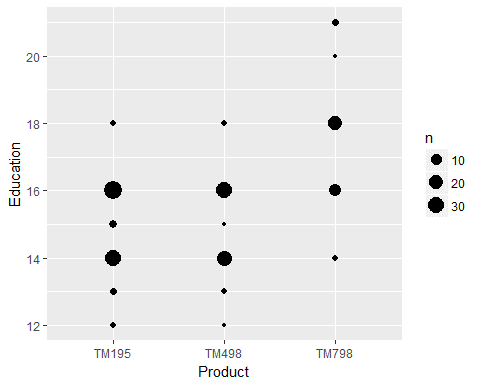
*Observation:*

1. *Out of 3 products, TM195 is used largely.*
2. Many Youngsters (i.e. between 20-30) use TM195.
3. TM798 is used least out of the 3 products.
4. TM798 - Only above 23(approximate) aged people use it.
5. When it comes to TM798, Majority of the users are in age group of 23-30
6. *Product vs Gender –*

**

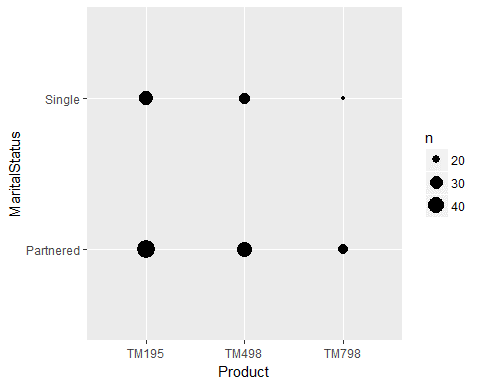
*Observation –*

1. *Very less number of females use TM798*
2. *large number of females use TM195*
3. *When it comes to Males, again TM195 is used by many. But the difference between all 3 products is not very much.*
4. *Product vs Education :*

**

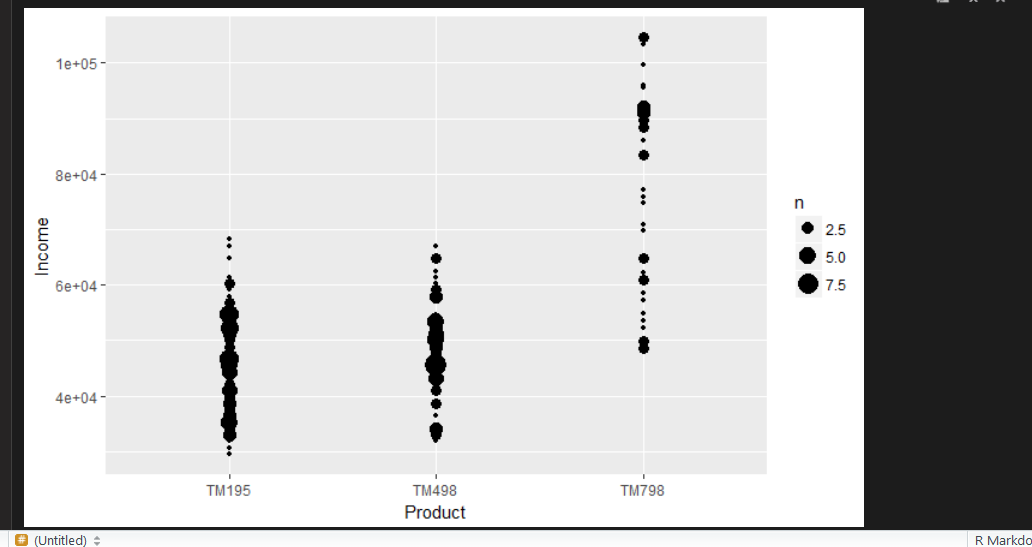
*Observation:*

1. Note that in our dataset, we don’t have information on education - 17 and 19.
2. Out of 3 products, again TM195 is more used.
3. Both TM195 and TM498 has no users after education 18 whereas TM798 has got few users.
4. TM798 usage starts only after education 14.
5. *Product vs Marital :*

**

*Observation :*

1. TM195 : Here we see large number of partnered people using this.
2. TM195 is used little less when it comes to Single people.
3. TM498 : More number of partnered people using it when compared to Single people
4. TM 798 : Very less single people using this product
5. *Product vs Income :*

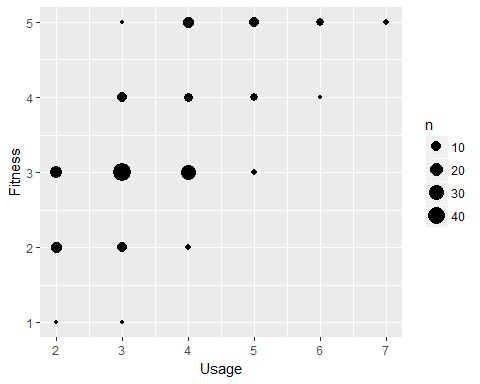
**

*Observation –*

1. *Both TM195 and TM498 has similar observations i.e people having salary between 30k-50k uses it.*
2. *When it comes to TM798 higher salaried people tends to use more. We can even see people with more than 1L using it.*

*We might assume that TM798 is a bit costly when compared to the other 2.*

1. *Usage vs Fitness:*

**

*Observation :*

1. *For getting fitness level 5 all you have do is start using the product thrice or more than that.*
2. *In general we can say the more usage the more fit you are.*

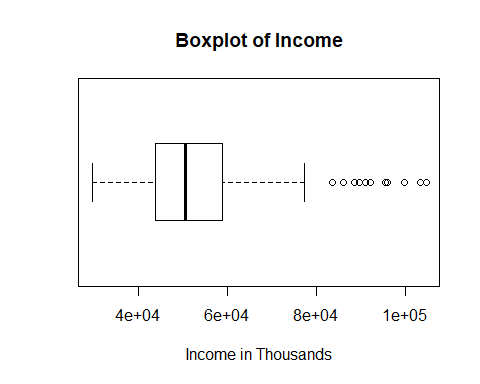
*3.5 :* Missing Value Identification

*There are no Missing values in our dataset.*

3.6 : Outlier Identification:

Though this is not in our scope, I have plotted some boxplots and observed outliers.

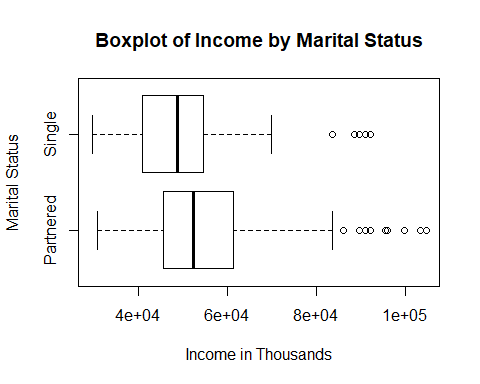
1. Boxplot of only income :



Observation –

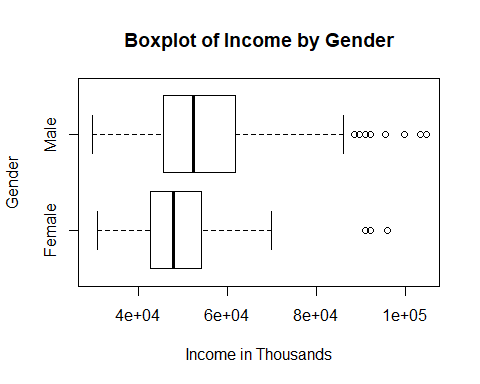
*We have quite a few number of outliers in the higher income range. i.e after 79k*

1. Boxplot of income by marital status –



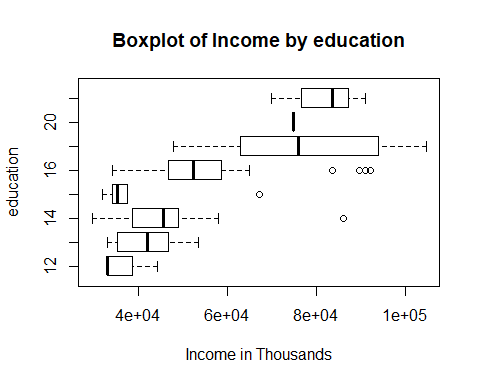
Observation –

1. More number of outliers (higher income) in Partnered than in single.
2. When it comes to higher income, Partnered people are more when compared to single ones.
3. Box plot of Income by Gender –



Observation –

1. More number of outliers (higher income) in Males than in females.
2. When it comes to higher income, males earn more when compared to females.
3. Boxplot of income by education -



Observation –

1. *Clearly we can say outliers (higher income) exists only in education - 14, 15 and 16.*
2. *Education 18 and more earns higher income.*
3. *Note that we do not have any values for education 17 and 19 in our dataset.*

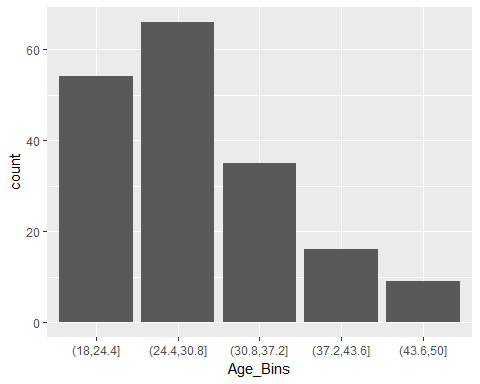
3.7 : Variable Transformation / Feature Creation

*I have created a new column Age\_Bins and used this as binning for age column and have plotted few graphs.*

*# Lets create bins in Age column and do some analysis*

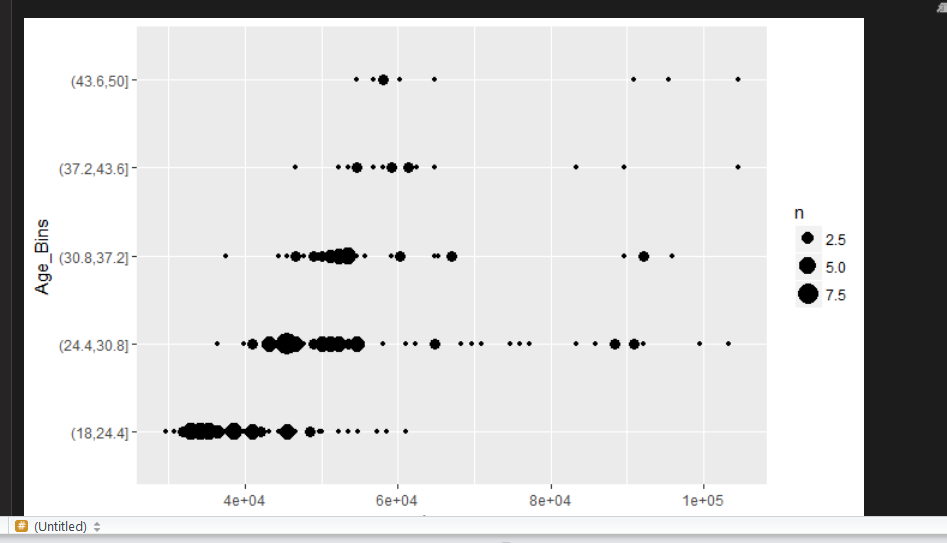
*cardio$Age\_Bins <- cut(Age,5)*

1. *Age count*

**

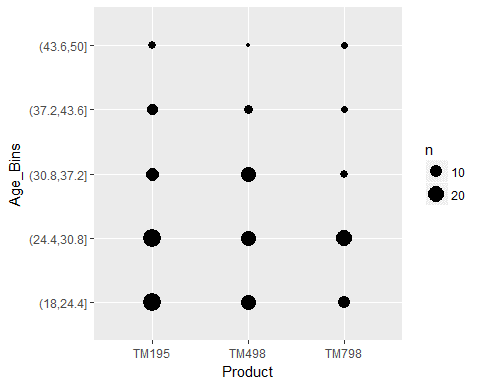
*Observation –*

1. *Here we have Around 54 people in the age range of 18-24.4*
2. *we have Around 65 people in the age range of 24.4-30.8*
3. *we have Around 35 people in the age range of 30.8-37.2*
4. *we have Around 17 people in the age range of 37.2-43.6*
5. *we have Around 9 people in the age range of 43.6-50*
6. Age vs Income –



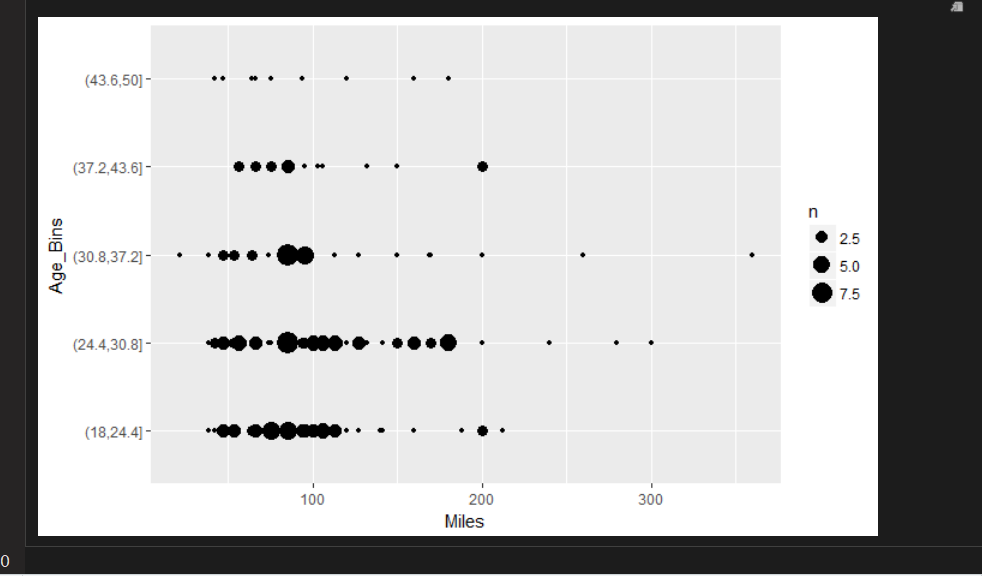
Observation –

1. Most of the users are in the salary range of 25-50k.
2. Few outliers can be seen who earns more than 1L.
3. *Age vs Product –*

**

*Observation –*

1. *TM195 is used majorly out of 3 products.*
2. *Majority of the users of TM195 are in the range of 18-30 Years.*
3. *TM798 has got its users more in the range of 24-30 years.*
4. *When it comes to TM498, Majority of the users are in the range of 18-30 years but after which hardly few users using it.*
5. *Age Vs Miles –*

**

*Observation –*

1. Majority of them are in the range of 40-125 miles
2. We can see 18-30 aged people covering around 85% of our dataset.
3. One outlier aged between 30-37 who has covered more than 300 miles.
4. Conclusion
5. *Out of the 3 products – TM195, TM498 and TM798: TM195 is the one used majorly by the users. We might want to see why the other two is not being used much.*
6. *We found that TM798 is being used more by higher salaried people, which tells that it might be a bit costly. So we might think about reducing the price of it.*
7. *When it comes to age, TM798 is not being used much by > 30 aged people. It has big share only within 18-30 aged people.*
8. *When it comes to education, TM798 is being used by more educated (i.e =>18) people than the other 2 products. Also having information on education level 17 and 19 would have added more value to the analysis.*
9. *TM798 is very least used by females. We might want to see why the product is not used much by females.*
10. Appendix A – Source Code

# Exploratory Data Analysis - Cardio Fitness  
  
#=======================================================================  
  
# ====Environment Set up and Data Import=====   
# Setup Working Directory   
setwd("C:/AJay/AJay/Great Lakes/Fitness")  
getwd()

## [1] "C:/AJay/AJay/Great Lakes/Fitness"

# reading the file   
cardio <- read.csv("CardioGoodFitness.csv")  
attach(cardio)

# ===== Checking the total number of rows and columns =====  
dim(cardio)

## [1] 180 9

# So we have 180 rows and 9 columns in our dataset.  
  
  
# ===== Names of the Columns in the data set =====  
names(cardio)

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

# ===== Find out Class of each Feature, along with internal structure =====  
str(cardio)

## '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 ...

#Out of the 9 Variables - Age, Education, Usage, Fitness, Income and Miles are integer variables. Product, Gender and MaritalStatus are categorical.  
  
# ===== Check if there are any missing values =====  
summary(cardio)

## 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

Observation –

# we have 3 products - TM195,TM498 and TM798  
# Age ranges between 18 andd 50 in our dataset.  
# We have 76 Females and 104 males in our dataset.  
# Education ranges from 12 to 21.  
# Our dataset has 73 singles and 107 partnered people.  
# Usage level is from 2 to 7.  
# Fitness level is from 1 to 5.  
# Income varies from 29562 to 104581.  
# Miles ranges from 21 to 360.

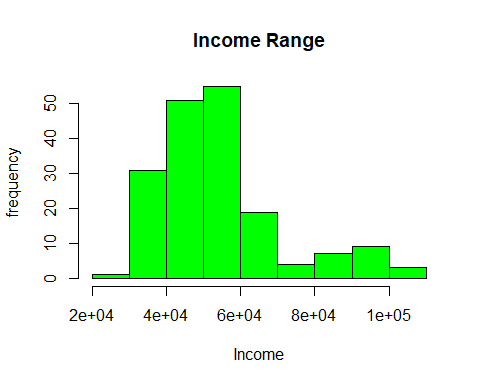
# ===== Display first 6 rows in our data set using head function =====  
head(cardio)

## Product Age Gender Education MaritalStatus Usage Fitness Income Miles  
## 1 TM195 18 Male 14 Single 3 4 29562 112  
## 2 TM195 19 Male 15 Single 2 3 31836 75  
## 3 TM195 19 Female 14 Partnered 4 3 30699 66  
## 4 TM195 19 Male 12 Single 3 3 32973 85  
## 5 TM195 20 Male 13 Partnered 4 2 35247 47  
## 6 TM195 20 Female 14 Partnered 3 3 32973 66

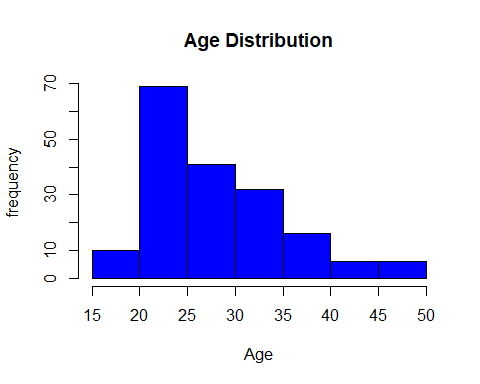
# ===== Display last 6 rows in our data set using tail function =====  
tail(cardio)

## Product Age Gender Education MaritalStatus Usage Fitness Income Miles  
## 175 TM798 38 Male 18 Partnered 5 5 104581 150  
## 176 TM798 40 Male 21 Single 6 5 83416 200  
## 177 TM798 42 Male 18 Single 5 4 89641 200  
## 178 TM798 45 Male 16 Single 5 5 90886 160  
## 179 TM798 47 Male 18 Partnered 4 5 104581 120  
## 180 TM798 48 Male 18 Partnered 4 5 95508 180

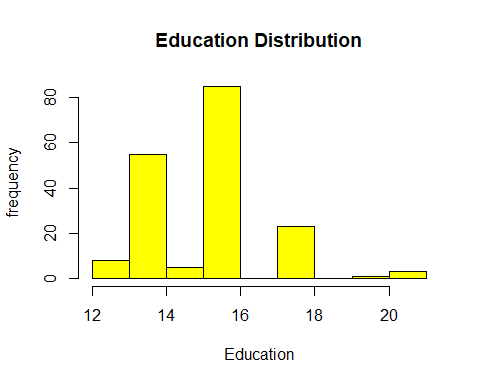
# Lets start with Univariate Analysis, first income column  
hist(Income,main="Income Range",xlab="Income",ylab="frequency",col="green")



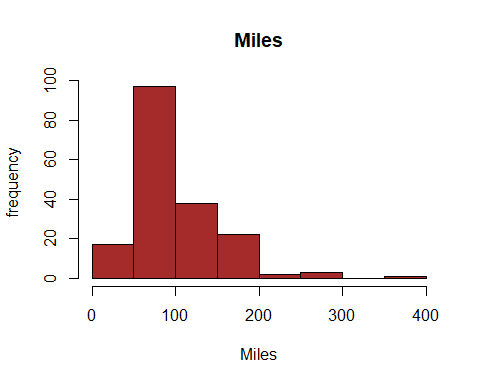
# Observation - We see more income values lies in 40k-60k range.  
  
# lets see the Age distribution  
hist(Age,main="Age Distribution",xlab="Age",ylab="frequency",col="blue")



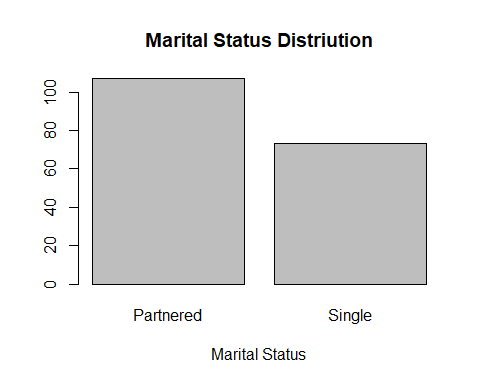
# Observation - 20-30 aged people are more in our dataset.  
  
# Lets see the Education column  
hist(Education,main="Education Distribution",xlab="Education",ylab="frequency",col="yellow")



# Observation - We can see 14 and 16 are high in our distribution  
  
# Lets check the miles column  
hist(Miles,main="Miles",xlab="Miles",ylab="frequency",col="brown")

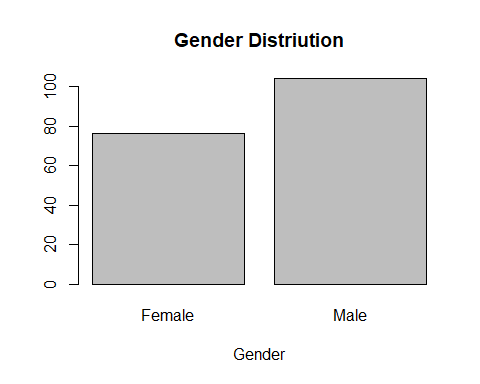


# Observation - from the plot you see a major of them belong to 50-100 miles group.  
  
# lets see the Marital status column   
count <- table(cardio$MaritalStatus)  
barplot(count,main="Marital Status Distriution",xlab= "Marital Status")



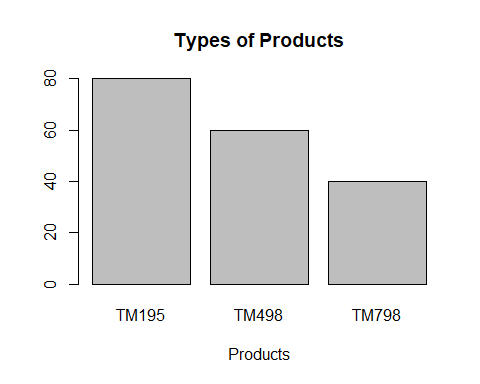
# Observation - there are more number of partnered people than the single ones.

# lets see the Gender column   
count2 <- table(cardio$Gender)  
barplot(count2,main="Gender Distriution",xlab= "Gender")



# Observation - There are more no. of Males than females in our dataset.

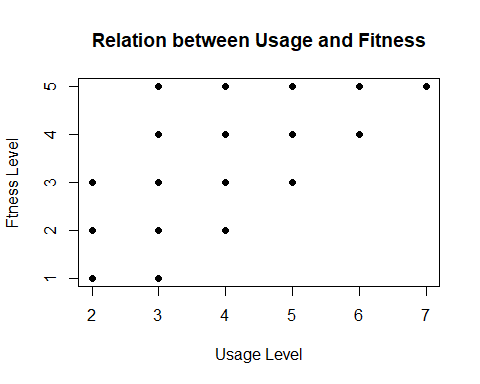
# lets see how many types pf products are there.  
count3 <- table(cardio$Product)  
barplot(count3,main="Types of Products",xlab= "Products")



# Observation - We have 3 types of products - TM195, TM498, TM798. Out of which we have more products of TM195 and less of TM798.

Lets see if we can get somethig out of Usage vs Fitness level.

plot(Usage, Fitness, main="Relation between Usage and Fitness",   
 xlab="Usage Level", ylab="Ftness Level",pch=19)

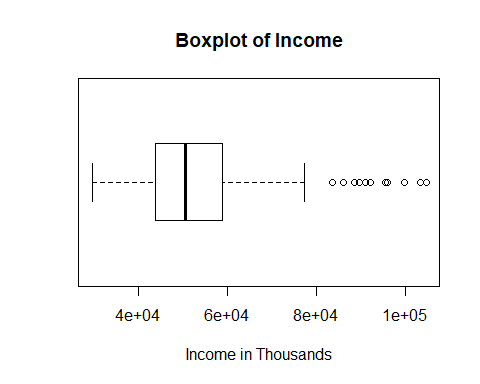


Observation –

# For getting fitness level 5 all you have do is start using the product thrice or more than that.  
# In general we say from the above plot is the more usage the more fit you are.

Box plots for outlier identification –

boxplot(Income,xlab="Income in Thousands",horizontal = TRUE,main="Boxplot of Income")

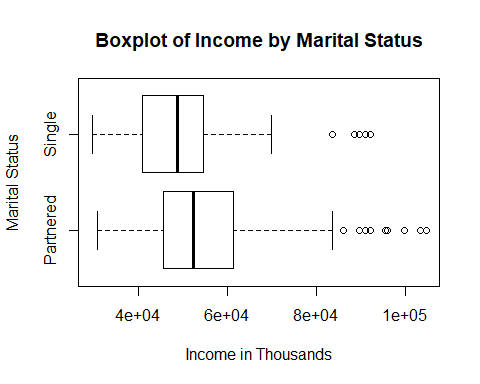


We have quite a few number of outliers in the higher income. i.e after 79k

# Now lets check the outliers again for the income but with other variables like marital status,Gender and education.

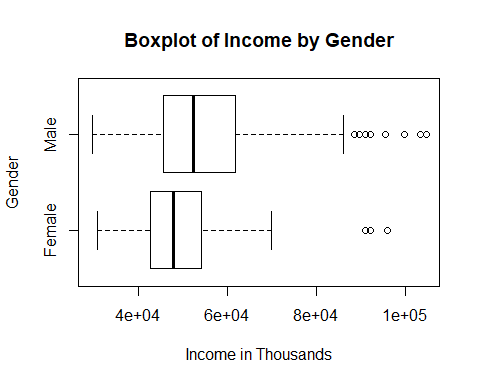
# 1. Martial status

boxplot(Income~MaritalStatus,xlab="Income in Thousands",ylab="Marital Status",horizontal = TRUE,main="Boxplot of Income by Marital Status")



# Observations :   
# 1. More number of outliers (higher income) in Partnered than in single.  
# 2. When it comes to higher income, Partnered people are more when compared to single ones(whci is obvious)

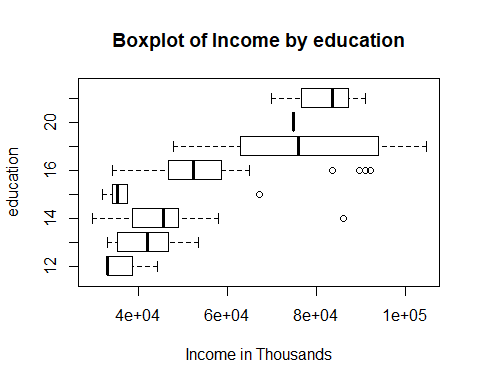
# 2. Gender  
  
boxplot(Income~Gender,xlab="Income in Thousands",ylab="Gender",horizontal = TRUE,main="Boxplot of Income by Gender")



Observations :   
# 1. More number of outliers (higher income) in Males than in females.  
# 2. When it comes to higher income, males earn more when comapred to females

# 3. Education :

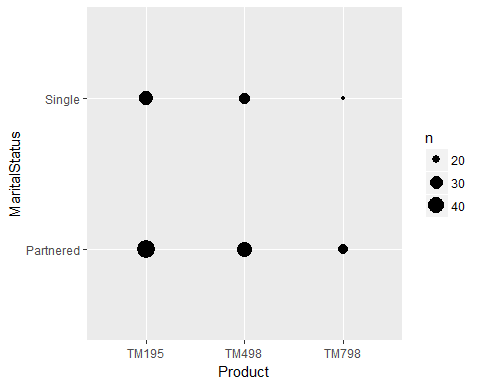
boxplot(Income~Education,xlab="Income in Thousands",ylab="education",horizontal = TRUE,main="Boxplot of Income by education")



# Observations :   
# 1. Clearly we can say outliers (higher income) exists only in education - 14,15 and 16.  
# 2. Education 18 and more earns higher income.

Marital Status Vs Product :

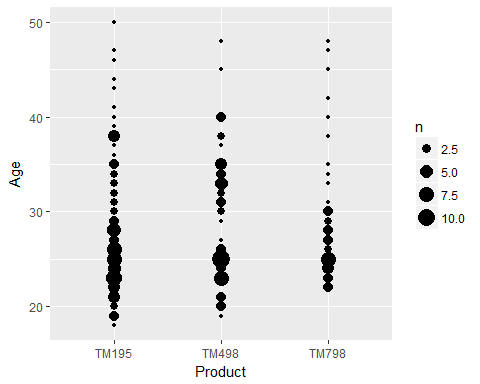
g<-ggplot(cardio,aes(Product,MaritalStatus))  
g + geom\_count()



# Observation -  
# TM195 : Here we see more number large number of partnered people using TM195 where as a little less when it comes to Single people.   
# TM498 : More number of partnered people using it when compared to Single people  
# TM 798 : Very less single people using this product.

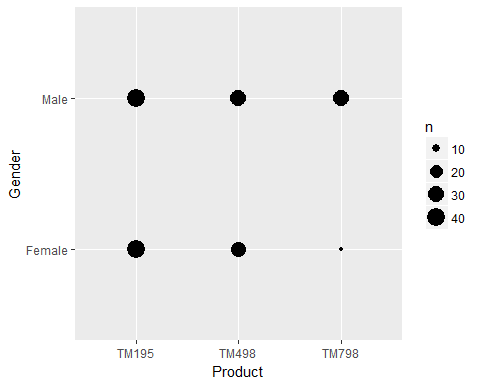
# Lets see the distribution of products vs the Age group

g<-ggplot(cardio,aes(Product,Age))  
g + geom\_count()



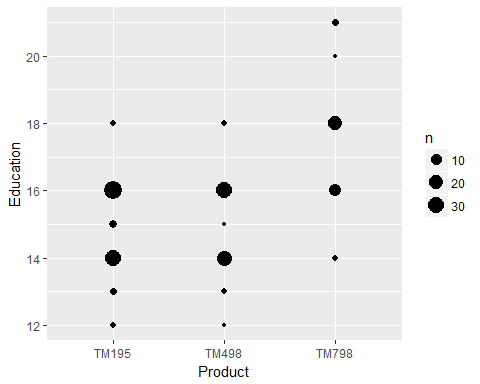
# From the graph below can be said -  
  
# Out of 3 products, TM195 is used largely.  
# Many Youngsters (i.e between 20-30) use TM195  
# TM798 is used least out of the 3 products.  
# TM798 - Only above 23(approximate) aged people use it.   
# When it comes to TM798, Majority of the users are in age group of 23-30.

#Lets see Product vs Gender  
g<-ggplot(cardio,aes(Product,Gender))  
g + geom\_count()



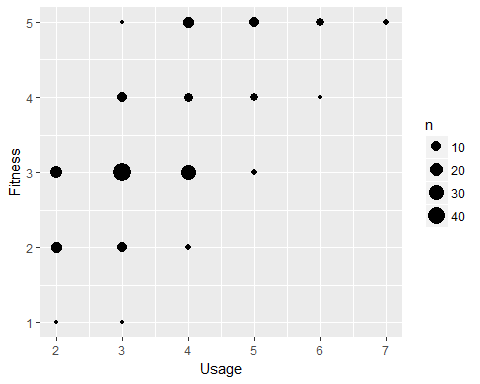
# Observation -   
# Very less number of females use TM798  
# Large number of females use TM195   
# When it comes to Males, again TM195 is used by many. But the difference between all 3 products is not very much.

# Lets see Product vs Eduation level  
g <- ggplot(cardio, aes(Product,Education))  
g + geom\_count()



Observation -   
# Note that in our dataset, we dont have information on education - 17 and 19  
# Out of 3 products, again TM195 is more used.  
# Both TM195 and TM498 has no users after education 18 whereas TM798 has got few users.  
# TM798 usage starts only after education 14.

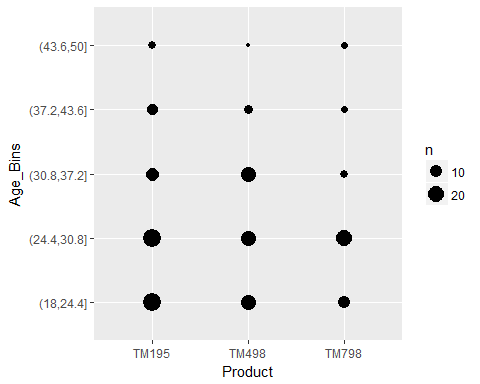
# Lets see the distribution of usage vs fitness  
  
g <- ggplot(cardio, aes(Usage,Fitness))  
g + geom\_count()



# Observation -   
# Users using the product 4 times or more than that has good fitness level.  
# There are more number of people in our dataset who has moderate or average fitness lelvel of 3.

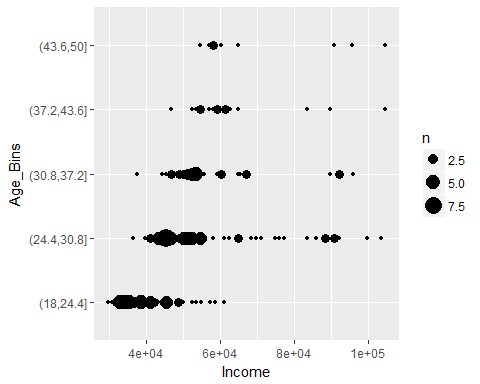
# Lets check the Box plot now for the bins that we have created for Age.

g <- ggplot(cardio, aes(Product,Age\_Bins))  
g + geom\_count()



# Observation -   
  
# TM195 is used majorly out of 3 products.  
# Majority of the users of TM195 are in the range of 18-30 Years.  
# TM798 has got its users more in the range of 24-30 years.  
# When it comes to TM498, Majority of the users are in the range of 18-30 years but then they hardly few users using it.

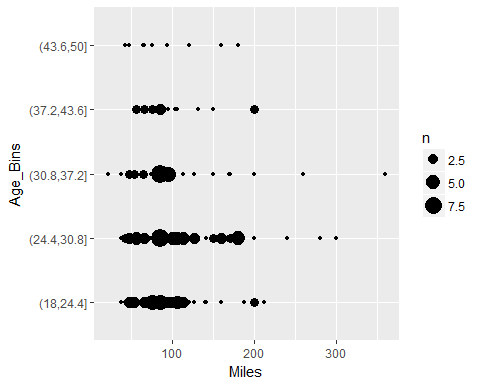
g <- ggplot(cardio, aes(Income,Age\_Bins))  
g + geom\_count()



# obseravtion :   
# Most of the users are in the salary range of 25-50k  
# Few outliers can be seen who earns more than 1L

# Lets see for age bins vs miles

g <- ggplot(cardio, aes(Miles,Age\_Bins))  
g + geom\_count()



# observations :  
# Majority of them are in the range of 40-125 miles   
# we can see 18-30 aged people covering around 90% of our dataset.  
# one outlier aged between 30-37 who has more than 300 miles.

############################################################################

T H E - E N D ############################################################################