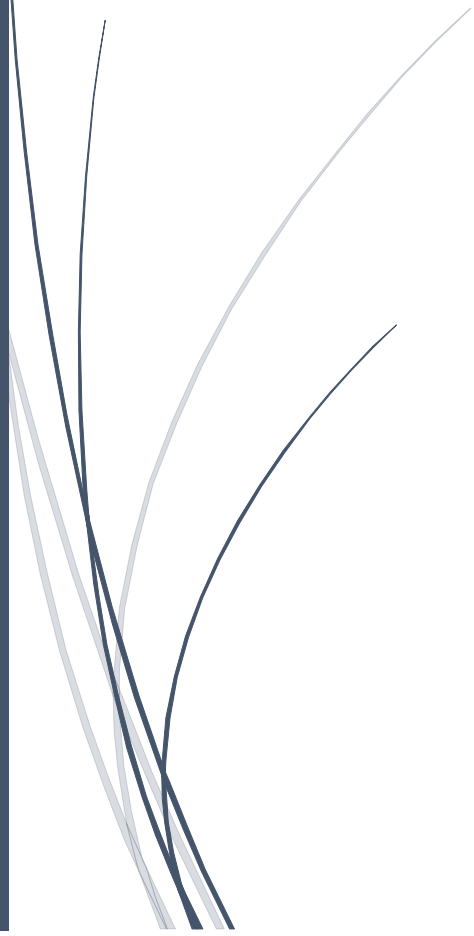
3/14/2020

Project in R - Cardio Good Fitness

Data Analysis



Tahmid Bari

Table of Contents

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 consist of the following:

- Importing the dataset in R
- Understanding the structure of dataset
- Graphical exploration
- Descriptive statistics
- Insights from the dataset

2. Assumptions

After analysing the data, we can say that this is to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. We can decide to investigate whether there are differences across the product lines with respect to customer characteristics. Therefore, it has been decided to collect data on individuals who purchased a treadmill at a CardioGoodFitness retail store during the prior three months. The data are stored in the CardioGoodFitness.csv file.

It has been identified from the dataset that the following customer variables to study: (**product purchased**), **TM195, TM498, or TM798**; gender; age, (**in years**); education, (**in years**); (**relationship status**), single or partnered; annual household (**income**); average number of times the customer plans to (**use the treadmill each week**); average (**number of miles**) the customer expects to walk/ run each week; and self-rated fitness on an 1-to-5 scale, where 1 is poor shape and 5 is excellent shape.

The 180 observations of the dataset relate to 180 unique customers of the treadmill products. The characteristics in the dataset are linked to the fitness level and treadmill usage characteristics of the customers. It can also be assumed that the data provide is accurate as per the survey/ data collected by the company. Below data dictionary is considered for the 9 variables in the dataset:

Sl. No.	Dimension	Detail Description
1	Product	Model of treadmill product (**TM195 / TM498 / TM798**)
2	Age	Age of the customer (**Years**)
3	Gender	Gender of the customer (**Male & Female**)
4	Education	Education of the customer (**Years**)
5	Marital Status	Marital status of the customer (**Single & Partnered/ Married**)
6	Usage	Weekly average number of times the customer plans to use the treadmill (**No. of times per Week**)
7	Fitness	Weekly average number of miles the customer expects to walk/run on the treadmill (**Miles per Week**). 5 being the "very fit" and 1 being "very unfit"
8	Income	Annual income of the customer (**Assumingly in US$**)
9	Miles	Total distance covered on the treadmill (**Miles**)

3. Exploratory Data Analysis – Step by step approach

A Typical Data exploration activity consists of the following steps:

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

Use this section to install necessary packages and invoke associated libraries. Having all the packages at the same places increases code readability.

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.

For example: **Cardio <- read.csv("CardioGoodFitness.csv")**

Please refer Appendix A for Source Code.

3.2 Variable Identification

Following R functions used during the analysis:

- dim (): See dimensions (# of rows/ # of columns) of the data frame.
- names (): See Feature names of the dataset.
- str (): Display internal structure of an R object, to identify classes of the features.

3.2.1 Variable Identification inferences

a. summary (<data frame>): Provides summary of the dataset.

str(Cardio): Provides the structure of the object "Cardio". Here it states as data.frame for the object as it has variables which are of "Factor" and " int" data types.summary (Cardio): gives summary all the 9 variables like the frequency of each variable:

No. of Observations	No. of Variables Dimension
180	9

No. of Females	No. of Males
76	104

Marital Status (Partnered)	Marital Status (Single)
107	73

```
> 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
```

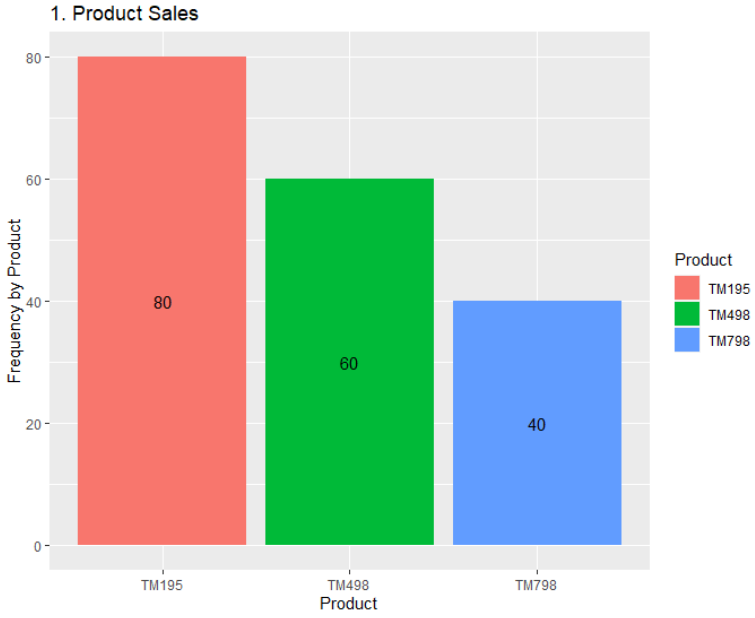
b. colSums(is.na()): Check missing values. There are no missing values in the data set.

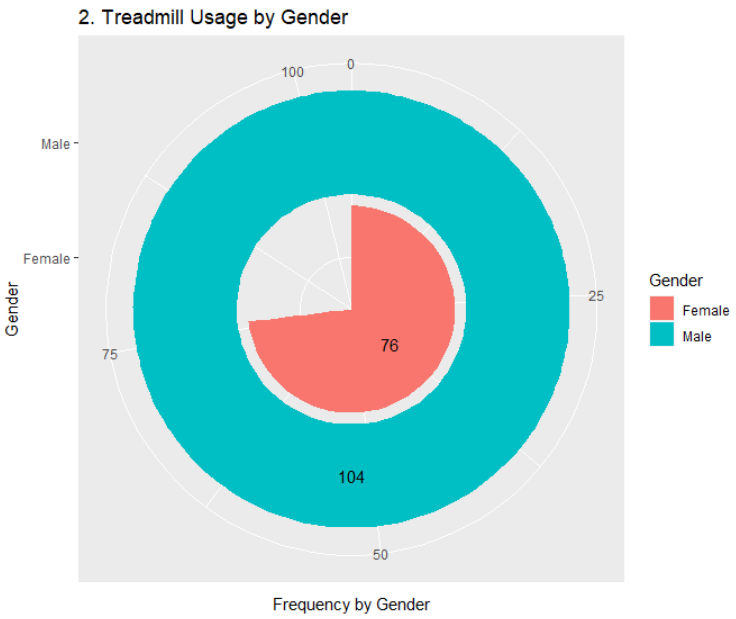
3.3 Univariate Analysis

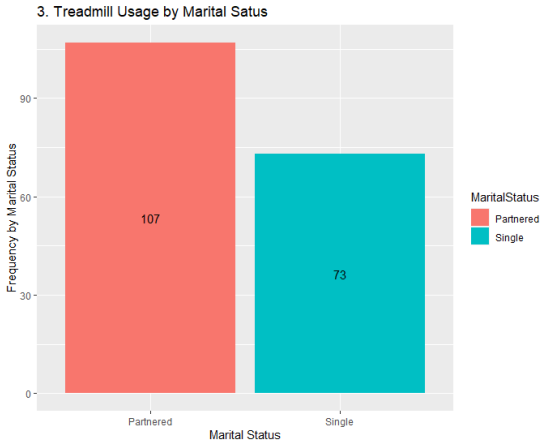
Univariate Analysis can be done on the Categorical Variables and Numeric Variables.

The Categorical Variables are: **Product, Gender and Marital Status**:

Using the following histogram (Bar Chart) to represent the distribution by the three categorical variables using Bar Plot and Pie Chart:







Observations from Categorical Variables Analysis:

(Graph 1) - Product Sales: The count by each product type (i.e. treadmill) is provided. TM195 has the highest count (80), followed by TM498 (60) and TM798 (40). This gives us an understanding that product TM195 is most popular among consumers. The reason might be that it is more affordable. We will see that later on.

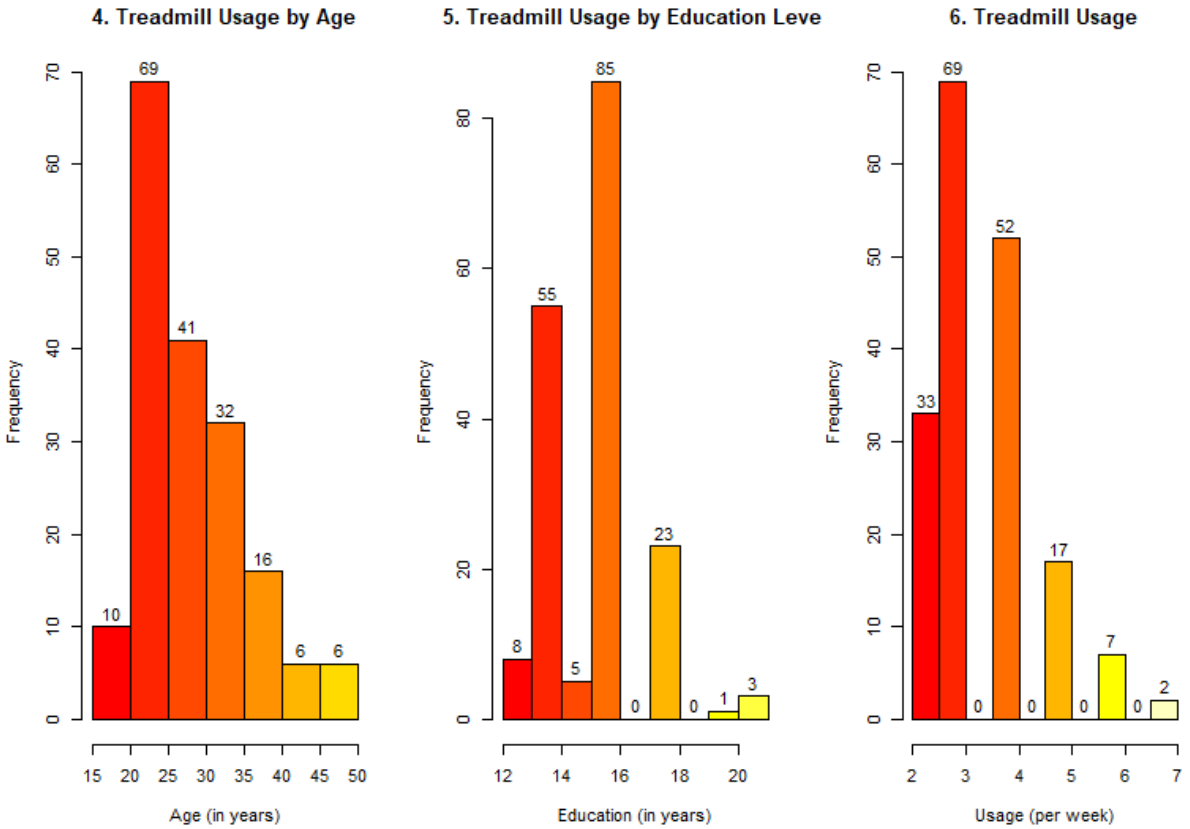
(Graph 2) - Treadmill Usage by Gender: The graph clearly shows that No. of Males (104) clearly outnumbers the No. of Females (76), with a suggestion that Men are more health conscious.

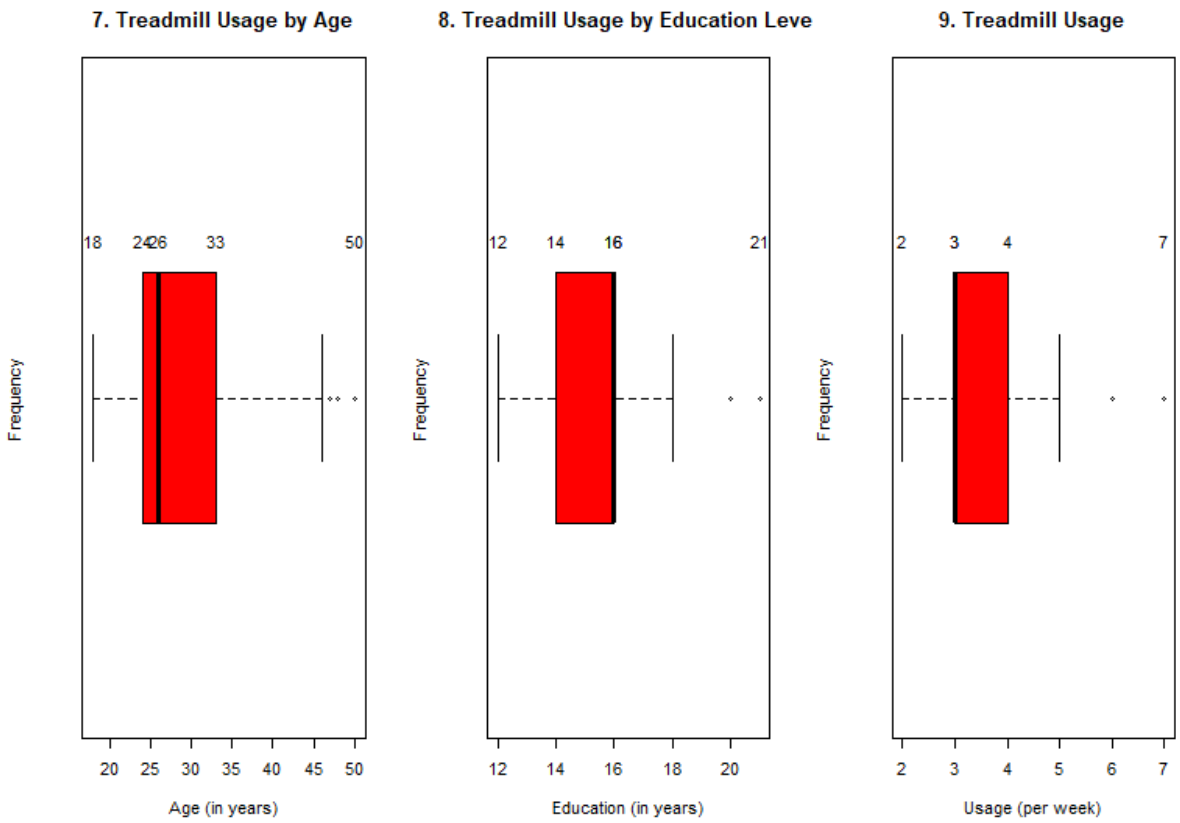
(Graph 3) - Treadmill Usage by Marital Status: Partnered (107) customers is more than single (73) customers.

Univariate Analysis can also be done based on the Integer (Numeric Variables).

The Numeric Variables are: **Age, Education, Usage, Fitness, Income and Miles.**

Using Histogram and Boxplot to represent the data and to identify the Outliers.





Observations from above Numeric Variables Analysis:

(Graph 4 and 7) - Treadmill Usage by Age:

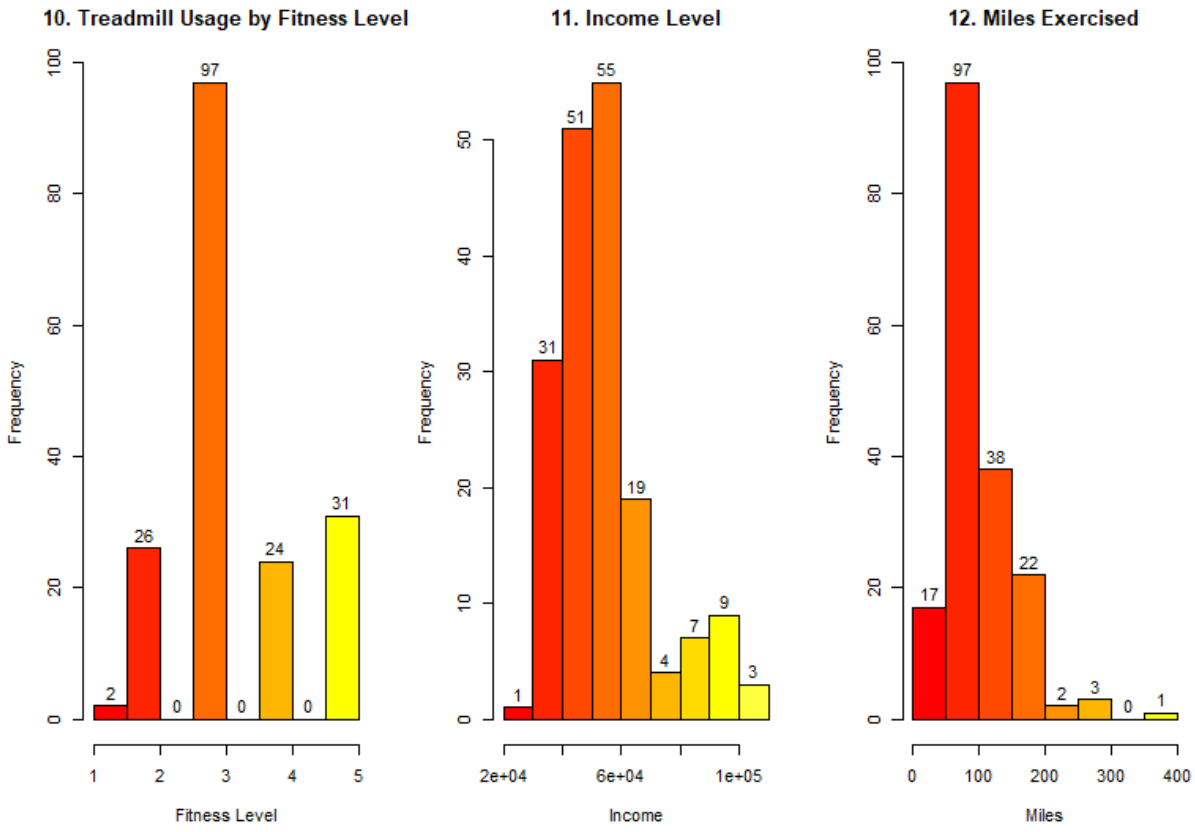
From the Boxplot, we can identify that the minimum age of the customer who used any of the products is 18 years and the maximum age is 50 years which is also an outlier. We can also identify from the Histogram that customers between age 20-25 are the ones who mostly used the treadmill for fitness with a count of 69 customers.

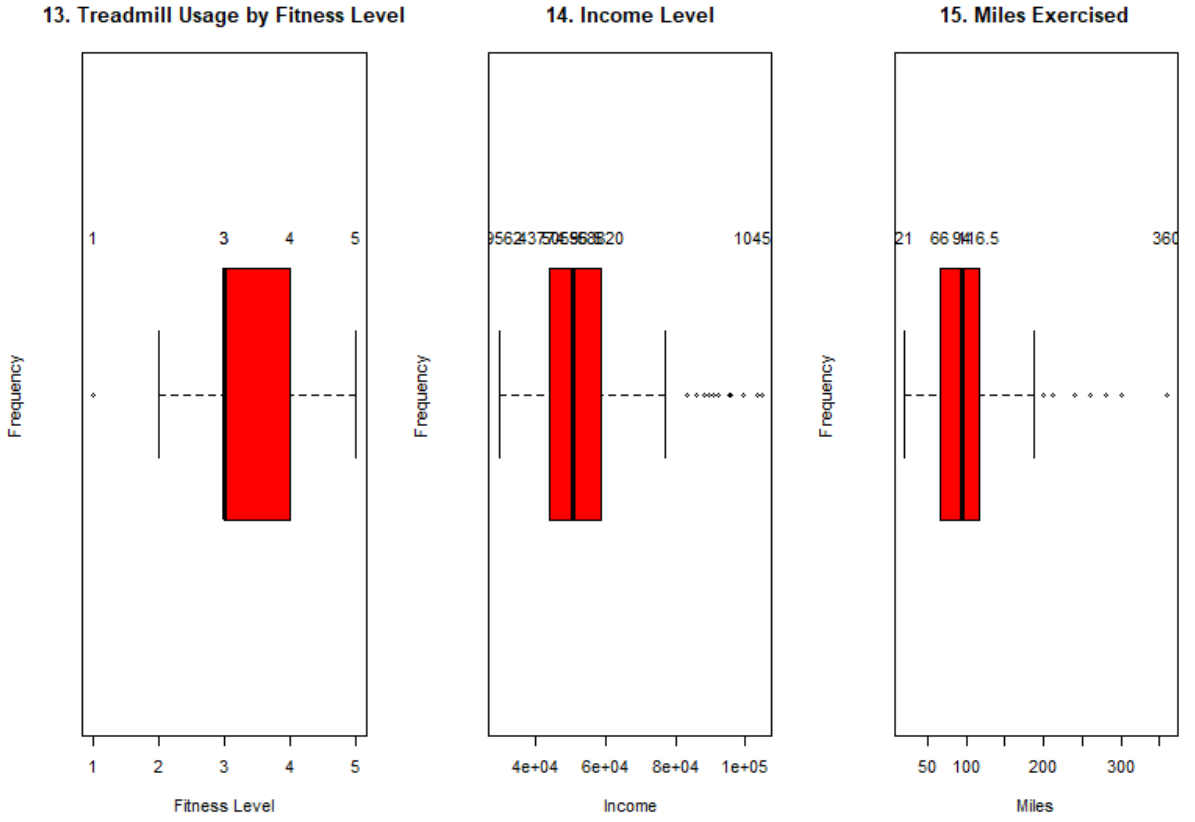
(Graph 5 and 8) - Treadmill Usage by Education Level:

From the Boxplot, we can identify that the minimum education level of the customers is 12 years while the maximum is 21 years. Customers with 15 years of average education are the ones who mostly uses the treadmill.

(Graph 6 and 9) - Treadmill Usage:

From the Boxplot, we identify that customers used the treadmill for a minimum of 2 times per week with maximum of 7 times a week. There are very few customers who exercised more than 6 times a week who can be termed as outliers in this dataset.


10. Treadmill Usage by Fitness Level
Frequency
Fitness Level
11. Income Level
Frequency
Income
12. Miles Exercised
Frequency
Miles


13. Treadmill Usage by Fitness Level
Frequency
Fitness Level
14. Income Level
Frequency
Income
15. Miles Exercised
Frequency
Miles

Observations from above Numeric Variables Analysis:

(Graph 10 and 13) - Treadmill Usage by Fitness Level:

The minimum number of miles the customer covers on the treadmill in a week is 1 and the maximum number of times the customer covers on the treadmill in a week is 5. The average of all the customers is 3.3 miles. The Boxplot indicates an outlier on the lower side, with fitness level 1.

(Graph 11 and 14) - Income Level:

From the Boxplot we can identify that the minimum income of the customers is 29562$, while the maximum is 104581$. There are some outliers with very high-income level above 70000$. The average income is 53720$ and they form the bulk of the treadmill buyers.

(Graph 12 and 15) - Miles Exercised:

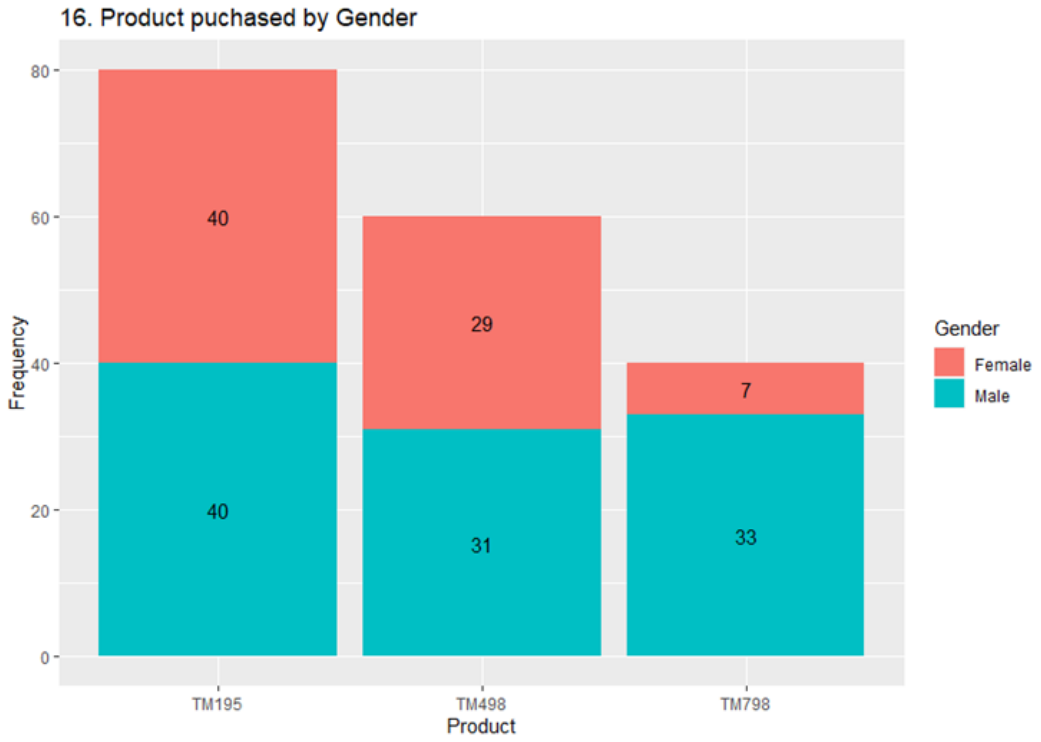
The minimum total miles exercised by the customers is 21 miles and the maximum total miles covered by the customer is 360 miles. The average of all the customers is 103.2 miles. The Boxplot indicates outliers beyond 170 miles.

3.4 Bi-Variate Analysis

Bi-variate analysis is used to represent the relationship between two variables and helps users to identify how changes in one variable affects the other variable. The below combinations are possible:

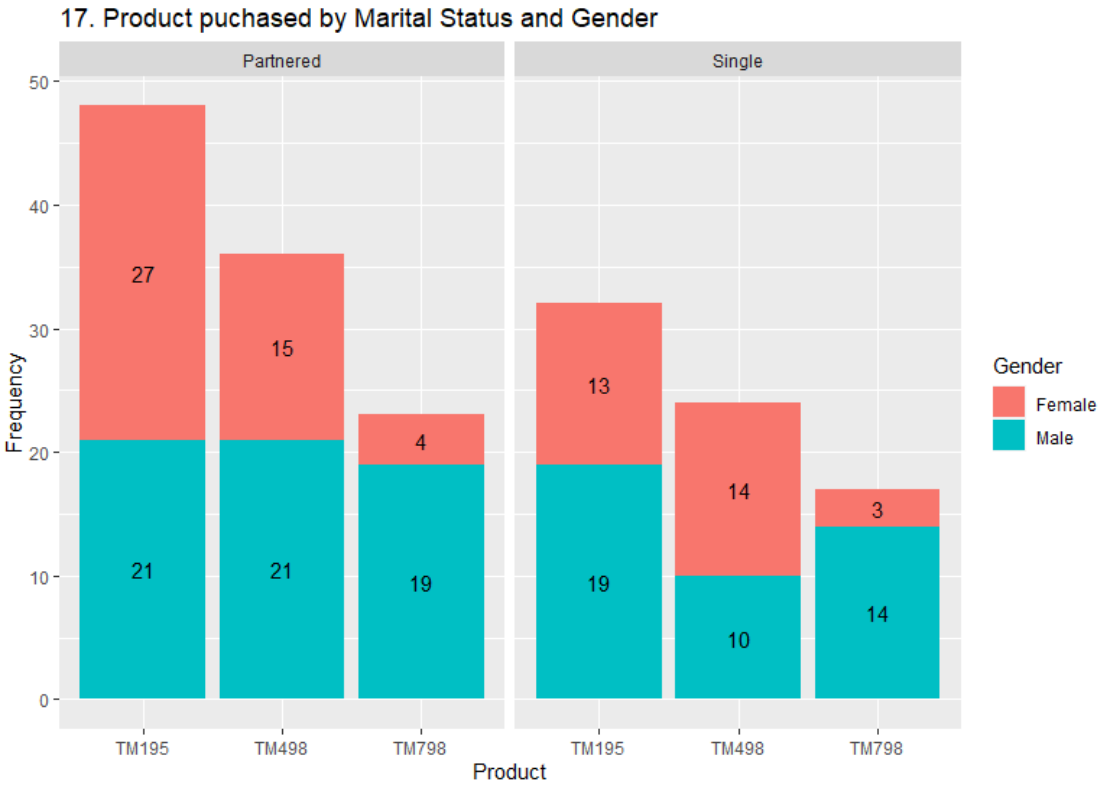
1. Categorical Vs Categorical
2. Categorical Vs Numerical
3. Numerical Vs Numerical

Categorical Vs Categorical: To identify the relationship between two categorical variables, we can use a Stacker Column Chart.



(Graph 16) - Product Purchased by Gender Observations:

Product TM195 is most and equally preferred by both Male and Female, followed by TM498 and TM798. However, all the three products are equally like by men while females like TM195 the most and TM798 the least. This brings up an assumption that the model/color for TM195 might be more unisex and since this information is not available, this extends the scope of analysis further.

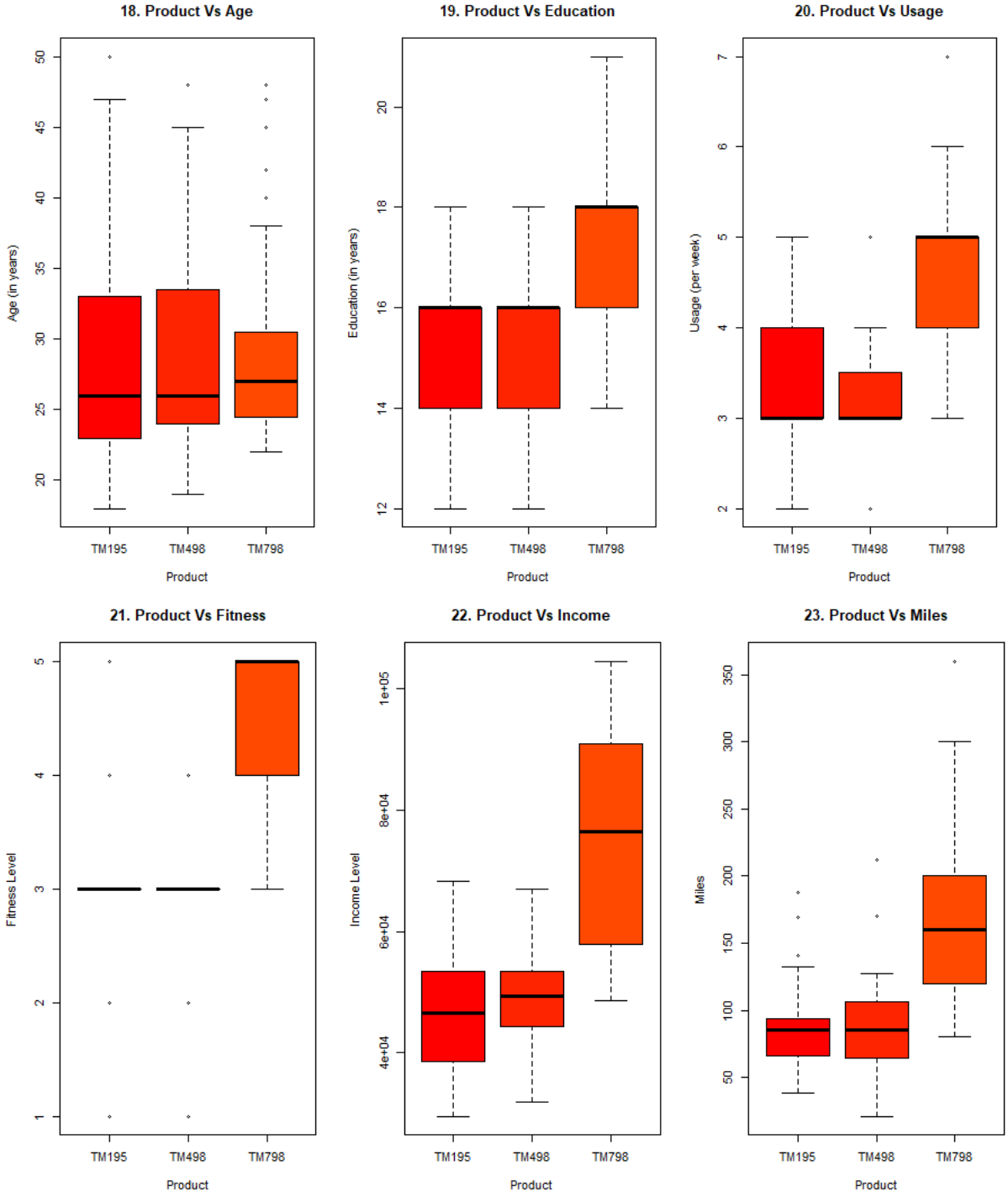


(Graph 17) - Product Purchased by Marital Status and Gender Observations:

We observe that both partnered and single customers prefer product TM195. However, partnered Males have equal liking towards all the product variances.

Categorical Vs Numerical:

To interpret relation between categorical and numerical variables, we can draw box plots for each of the categorical variables



(Graph 18) - Product Vs Age: TM195 is popular among all age groups.

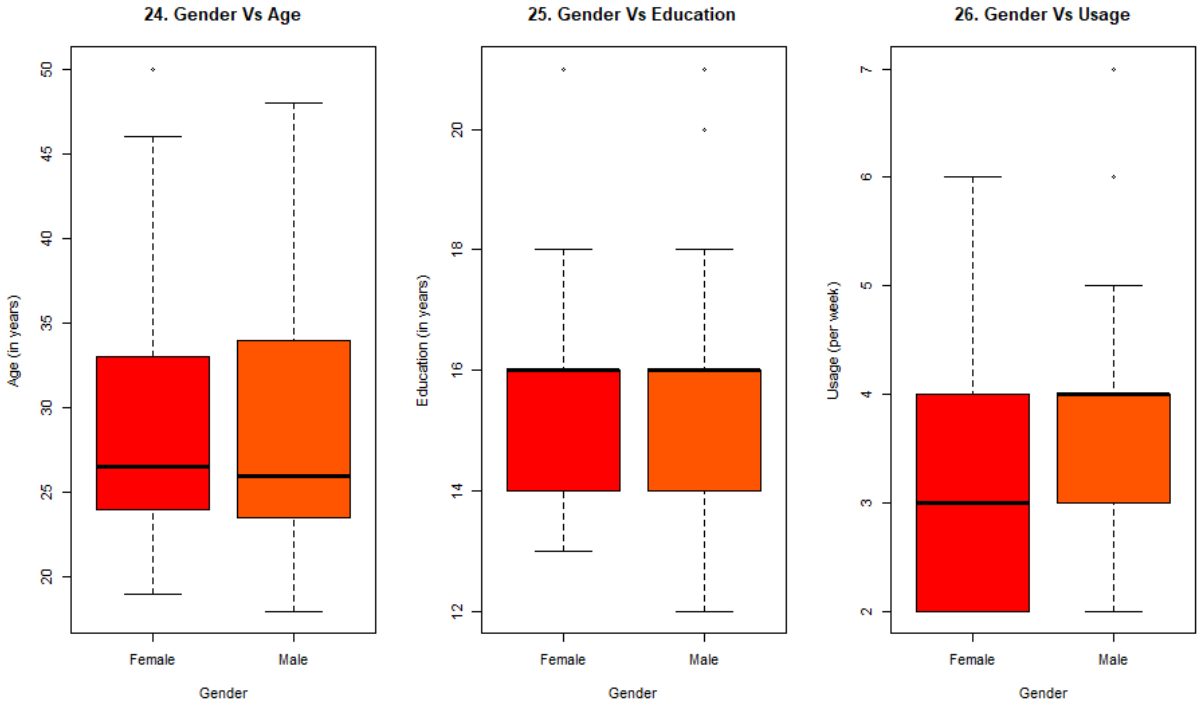
(Graph 19) - Product Vs Education: TM798 is more popular among people with higher education level.

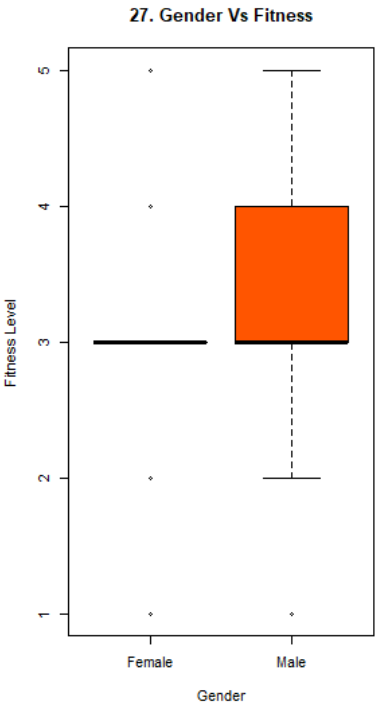
(Graph 20) - Product Vs Usage: Usage of TM798 is higher than others

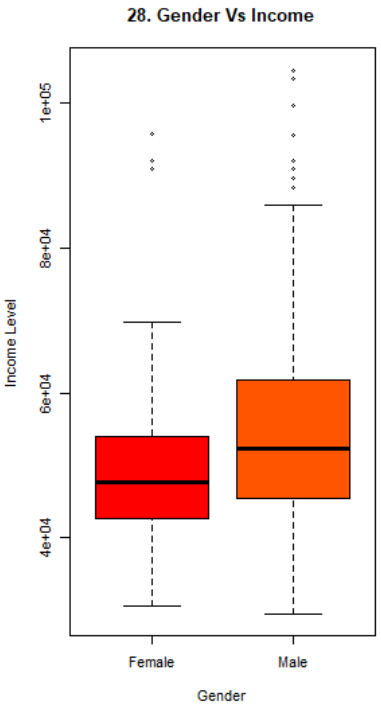
(Graph 21) - Product Vs Fitness: TM195 and TM498 is used by customers with fitness level of 3, whereas TM798 is used by customers with fitness level greater than 4.

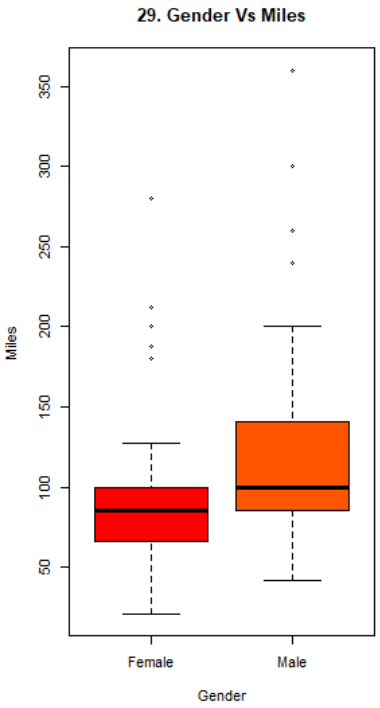
(Graph 22) - Product Vs Income: TM195 and TM498 are popular among low- and medium-income group, whereas TM798 is popular among medium to high income group.

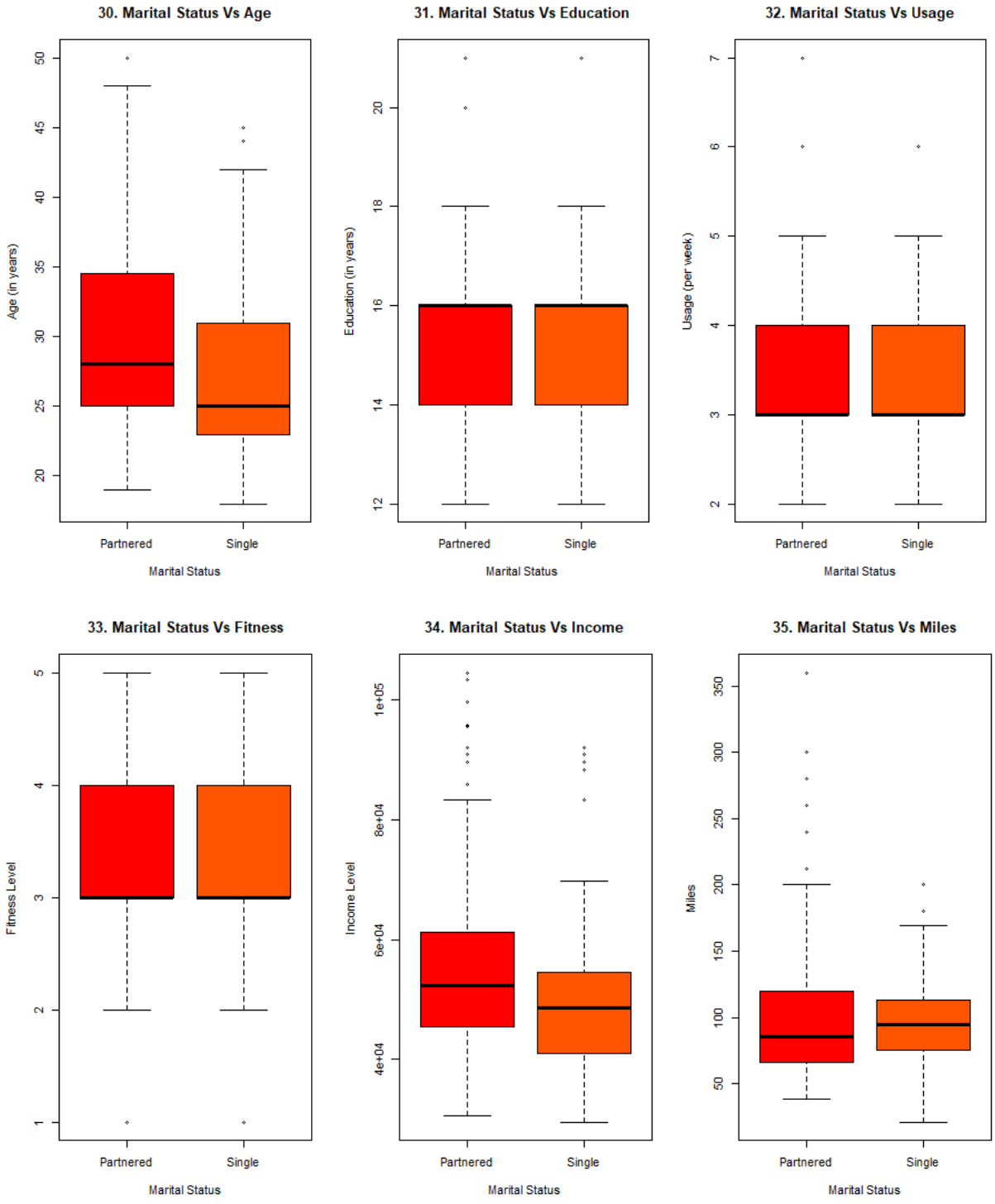
(Graph 23) - Product Vs Miles: Customers using TM798 has covered more miles. This also supports the previous observation that customers using TM798 have higher fitness level.











(Graph 30) - Marital Status Vs Age: Lesser number of Singles have purchased the treadmills.

(Graph 31) - Marital Status Vs Education: No proper relationship can be established.

(Graph 32) - Marital Status Vs Usage: Partnered have outliers and more usage.

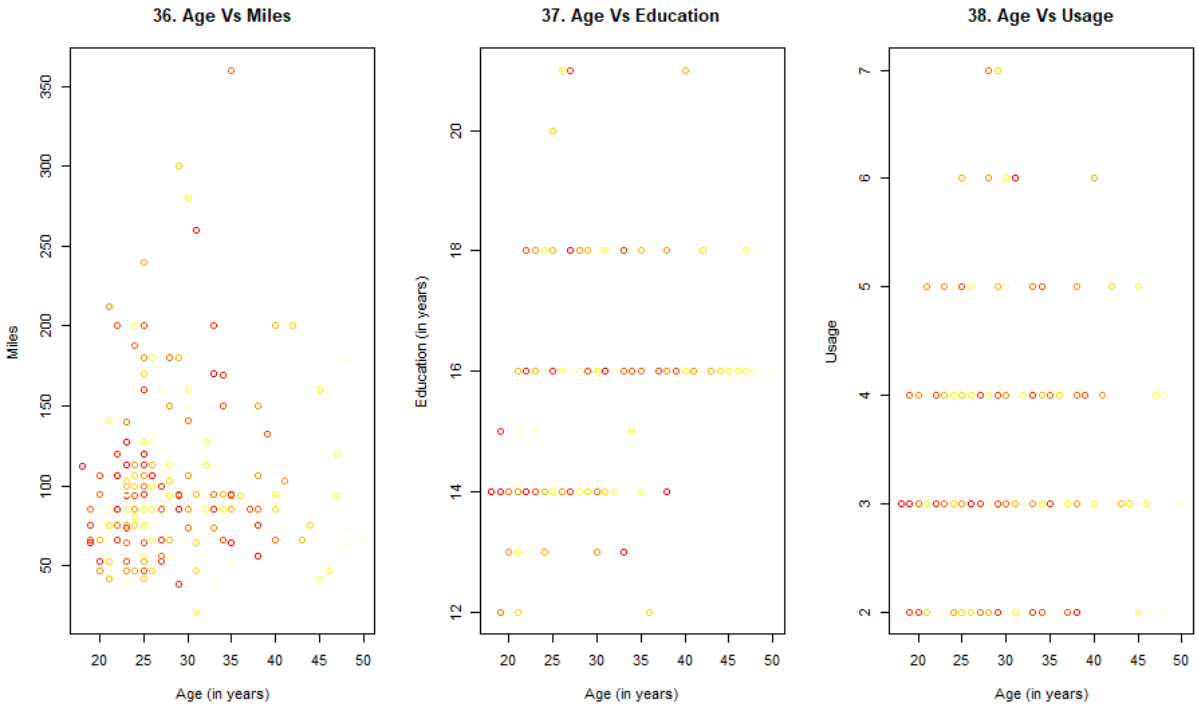
(Graph 33) - Marital Status Vs Fitness: No proper relationship can be established.

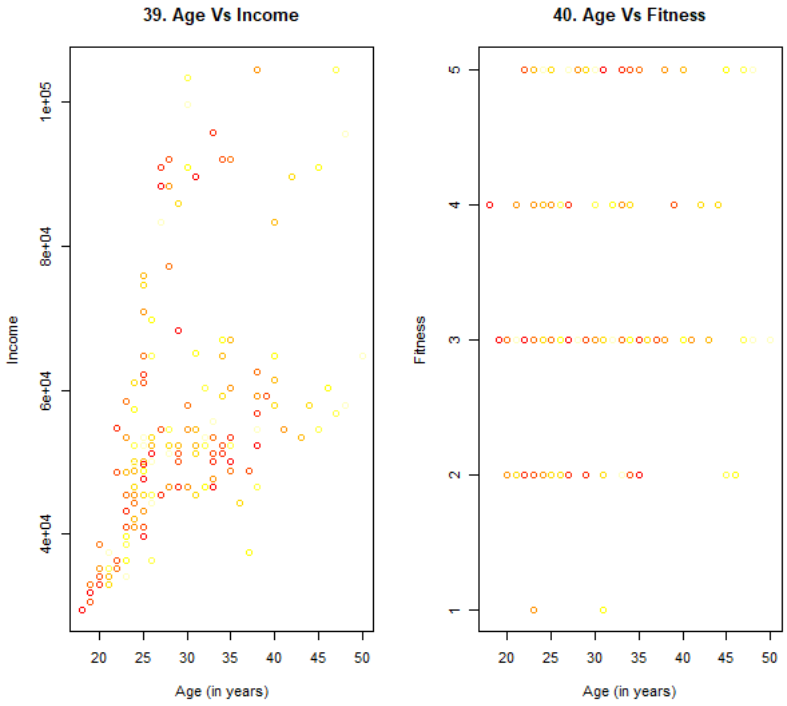
(Graph 34) - Marital Status Vs Income: Partnered persons have more income level than singles. This might because the household income will be higher.

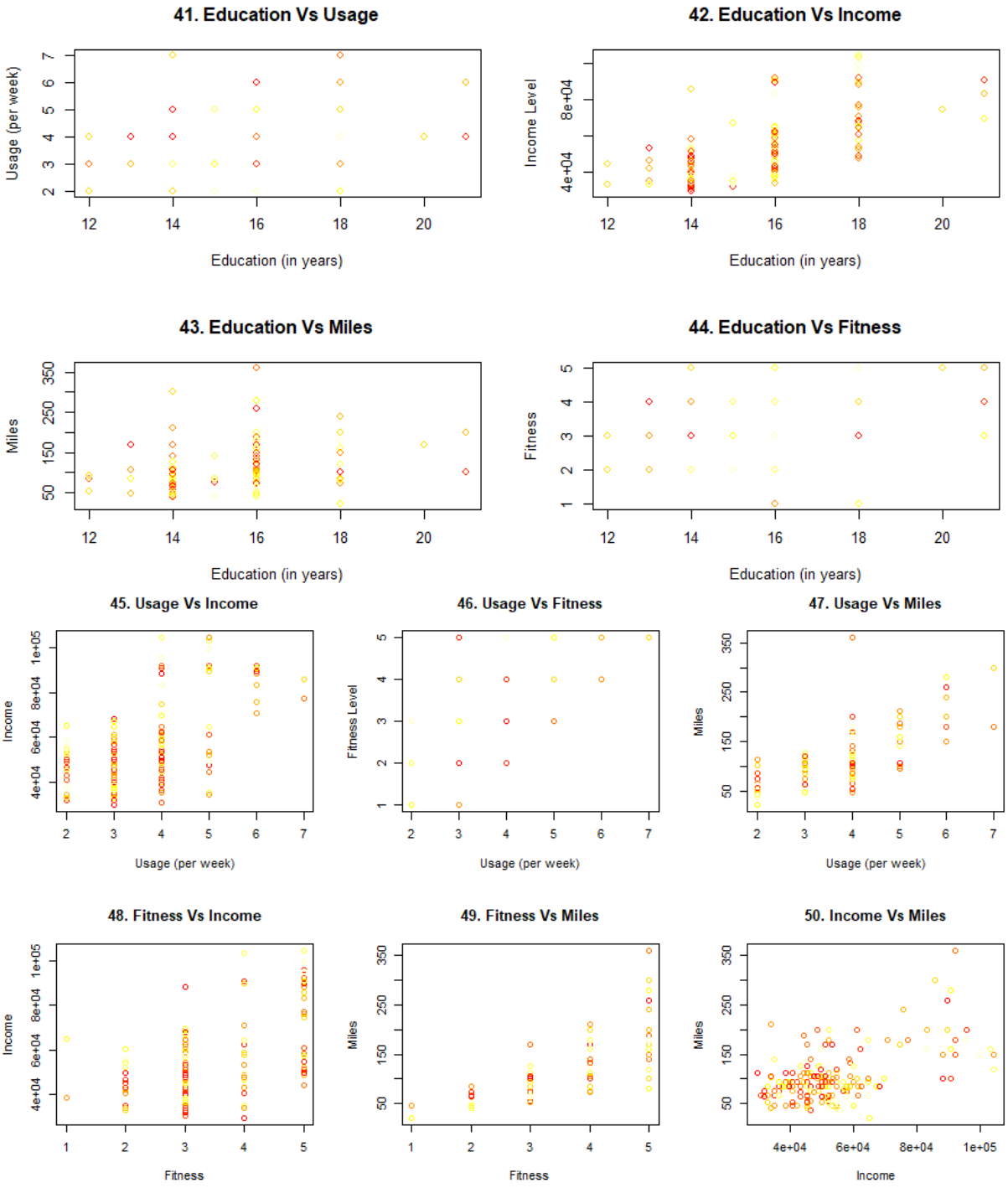
(Graph 35) - Marital Status Vs Miles: Partnered customers burn more miles.

Numerical Vs Numerical:

For bi-variate analysis between two Numerical variables, we should use at scatter plot. The pattern of scatter plot indicates the relationship between variables. The relationships can be linear or non-linear.







(Graph 36) - Age Vs Miles: The data is scattered all over with no proper correlation. This can be grouped by Age ranges for better distribution.

(Graph 37) - Age Vs Education: No proper relationship can be established. This can be grouped by Education in years range for better distribution.

(Graph 38) - Age Vs Usage: Customers from all age groups used the treadmill 2,3,4,5 times and week. We could also see that customers above the age of 35 used it 6,7 times a week also, being the outliers.

(Graph 39) - Age Vs Income: This relationship is directly proportional which means with increase in Age, the Income Level also increases.

(Graph 40) - Age Vs Fitness: No proper relationship can be established.

(Graph 41) - Education Vs Usage: No proper relationship can be established.
(Graph 42) - Education Vs Income: Higher the education, higher the income.
(Graph 43) - Education Vs Miles: Education level of 16 years covered more miles**.**
(Graph 44) - Education Vs Fitness: No proper relationship can be established.
(Graph 45) - Usage Vs Income: Higher usage is observed for high income level. Customers across all income levels used the treadmill on an average of 4 times a week.
(Graph 46) - Usage Vs Fitness: Usage is more for higher fitness level customers.
Graph 47) - Usage Vs Miles: The miles covered is more with increased usage. This is a positive relationship.
(Graph 48) - Fitness Vs Income: Positive relationship is observed. High fitness level for high income group.
(Graph 49) - Fitness Vs Miles: Positive relationship is observed. Customers with High fitness level tend to cover more miles.
(Graph 50) - Income Vs Miles: No proper relationship can be established.

3.5 Missing Value Identification

Missing data is part of any real-world data analysis. It can crop up in unexpected places, making analyses challenging to understand. It can lead to wrong prediction or classification.

In this project, we see that there are no missing values for the variables:

```
> colSums(is.na(cardio))
     Product          Age       Gender    Education MaritalStatus
           0            0            0            0            0
       Usage      Fitness       Income        Miles
           0            0            0            0
```

3.6 Outlier Identification

In statistics, an outlier is defined as an observation which stands far away from the most of other observations. Often an outlier is present due to the measurements error or due to presence of any unnatural data. Therefore, one of the most important tasks in data analysis is to identify and (if is necessary) to remove the outliers.
Outliers can drastically change the results of the data analysis and statistical modelling. It impacts the following:

- It increases the error rate and decreases the power of statistical analysis
- If the outliers are non-randomly distributed, they can decrease/increase normality
- They can give a wrong impression of the data set.

Most commonly used method to detect outliers is through graphical visualization. We have used various visualization methods, like **Boxplot, Histogram, and Scatter Plot** in the above graphs to derive inferences about the data set.

3.7 Variable Transformation / Feature Creation

With variable transformation, we can change the scale of the variables to give more meaningful understanding of the dimension. This might help us in establishing a linear relationship between variables. We can use Scatter Plot to establish any relation between the transformed variables.

We have created dummy variables to create meaningful observations:
For example, Income can be categorized into High Income, Medium Income and Low Income based on its value ranges.
Below are the new variables that are created:

1. Age_Classification (Age)
    - Teenagers: < 20 years
    - Middle Aged Adults: > 35 years
    - Young Adults: 20 to 35 years
2. Usage_Classification (Usage)
    - Amateurs: < 3 times a week
    - Pro: > 5 times a week
    - Regular: 3 to 4 times a week
3. Income_Classification (Income)
    - Low Income < 40000$
    - High Income> 70000$
    - Moderate Income: 40001 to 69999$
4. Fitness_Classification (Fitness)
    - Very Unfit: 1
    - Fit: 2-4
    - Very Fit: 5
5. Miles_Classification (Miles)
    - Low Usage: < 60 Miles
    - Medium Usage: 61 to 129 Miles
    - High Usage: > 130 Miles

A new data frame has been created with the additional 5 attributes.

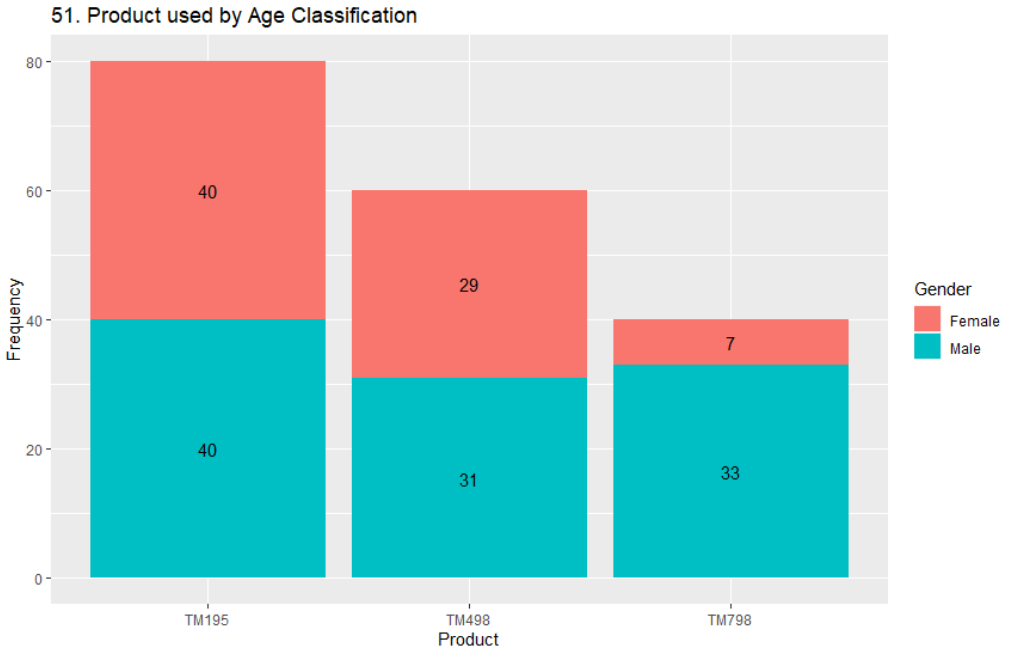
We have created a subset for each of the Product Types to do detailed analysis for each product.

We split the entire data set into 3 subsets:

- Product1: TM195
- Product2: TM498
- Product3: TM798

Refer to the Appendix A for the Source Code.

Data Exploration by New Variables:

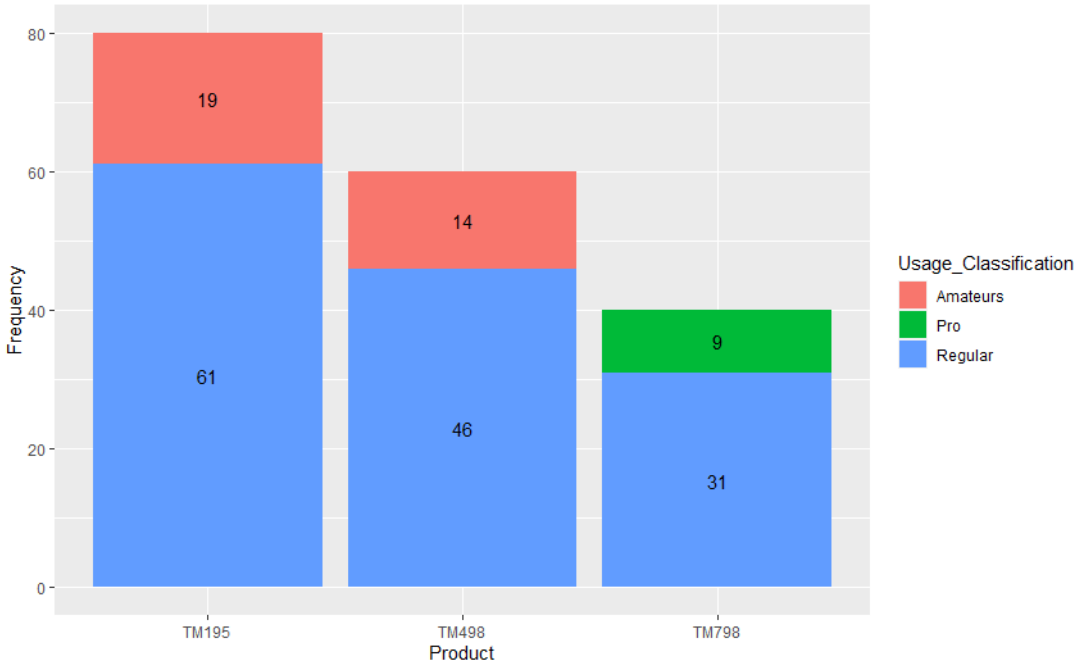


	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	Age_Classification	Usage_Classification	Fitness_Classification	Income_Classification	Miles_Classification
1	TM195	18	Male	14	Single	3	4	29562	112	Teenagers	Regular	Fit	Low Income	Medium Usage
2	TM195	19	Male	15	Single	2	3	31836	75	Teenagers	Amateurs	Fit	Low Income	Medium Usage
3	TM195	19	Female	14	Partnered	4	3	30699	66	Teenagers	Regular	Fit	Low Income	Medium Usage
4	TM195	19	Male	12	Single	3	3	32973	85	Teenagers	Regular	Fit	Low Income	Medium Usage
5	TM195	20	Male	13	Partnered	4	2	35247	47	Young Adults	Regular	Fit	Low Income	Low Usage
6	TM195	20	Female	14	Partnered	3	3	32973	66	Young Adults	Regular	Fit	Low Income	Medium Usage
7	TM195	21	Female	14	Partnered	3	3	35247	75	Young Adults	Regular	Fit	Low Income	Medium Usage
8	TM195	21	Male	13	Single	3	3	32973	85	Young Adults	Regular	Fit	Low Income	Medium Usage
9	TM195	21	Male	15	Single	5	4	35247	141	Young Adults	Regular	Fit	Low Income	High Usage
10	TM195	21	Female	15	Partnered	2	3	37521	85	Young Adults	Amateurs	Fit	Low Income	Medium Usage
11	TM195	22	Male	14	Single	3	3	36384	85	Young Adults	Regular	Fit	Low Income	Medium Usage
12	TM195	22	Female	14	Partnered	3	2	35247	66	Young Adults	Regular	Fit	Low Income	Medium Usage
13	TM195	22	Female	16	Single	4	3	36384	75	Young Adults	Regular	Fit	Low Income	Medium Usage
14	TM195	22	Female	14	Single	3	3	35247	75	Young Adults	Regular	Fit	Low Income	Medium Usage
15	TM195	23	Male	16	Partnered	3	1	38658	47	Young Adults	Regular	Very Unfit	Low Income	Low Usage
16	TM195	23	Male	16	Partnered	3	3	40932	75	Young Adults	Regular	Fit	Moderate Income	Medium Usage

Observations from above graph:

- Most of the treadmill customers had age between 20-34 years across all product types (TM195, TM498 & TM798)
- The average customer age across product types (TM195, TM498 & TM798) is approximately 28.9 years
- TM498 had users across all age groups

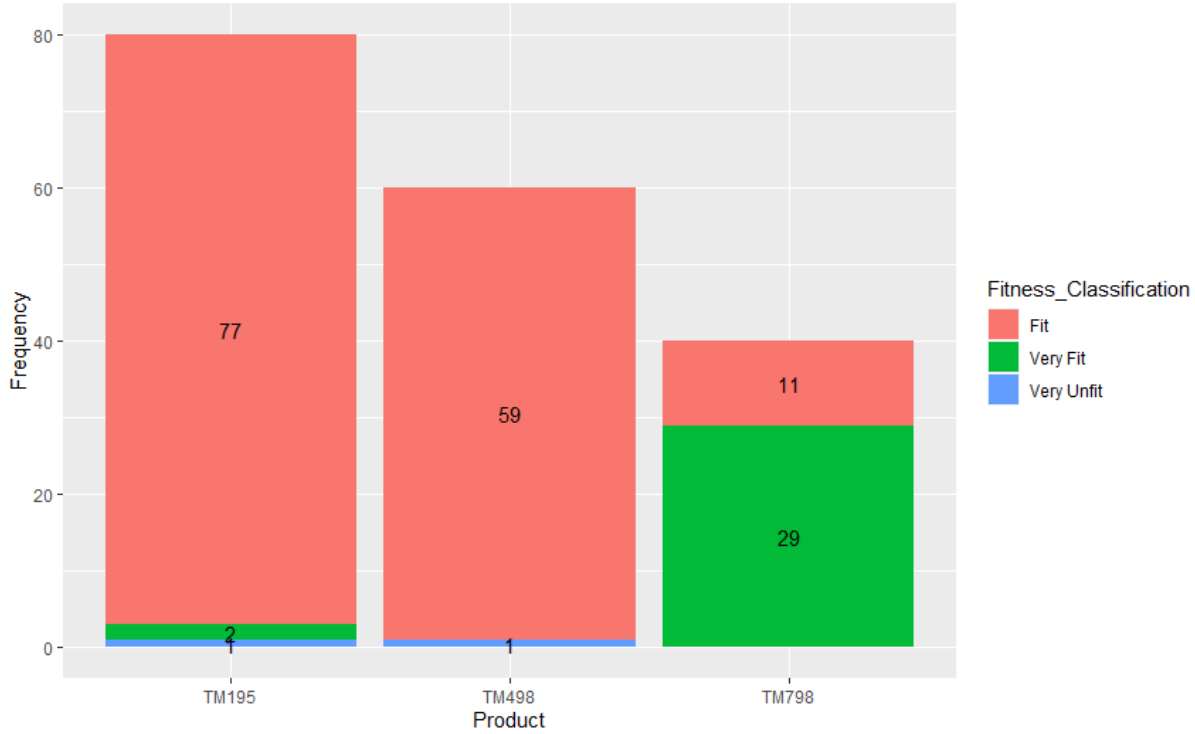
52. Product used by Usage Classification

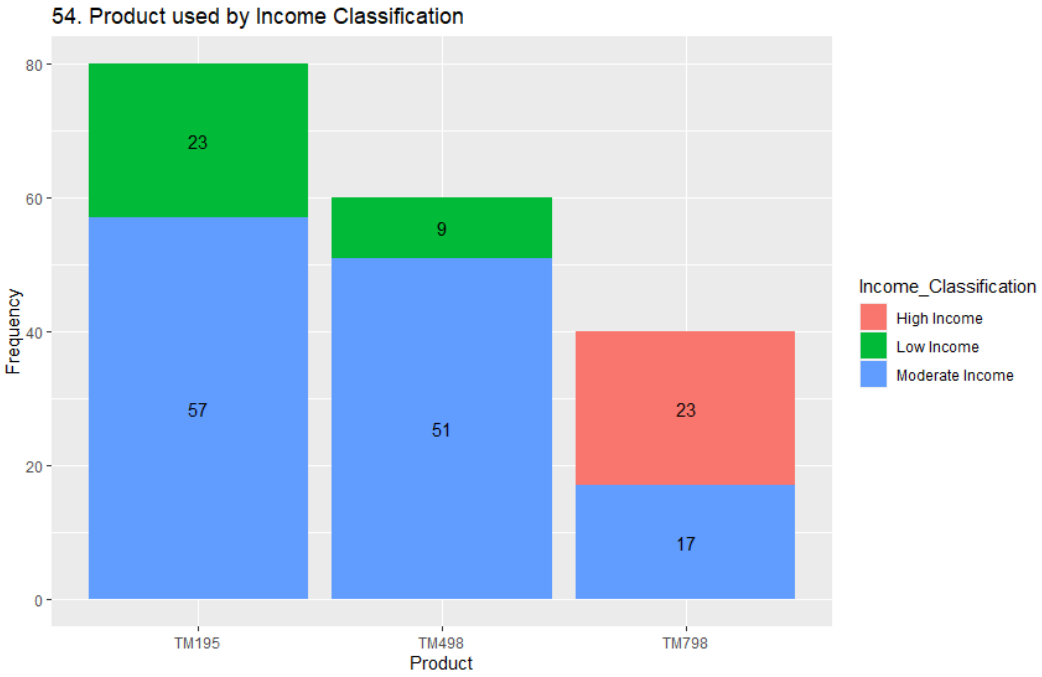


Observations from Above Graph:

- Majority of TM195 & TM498 customers use the treadmill 3 times a week on an average
- TM798 customers use the treadmill 4 times or more a week on an average.

53. Product used by Fitness Classification





Observations from Above Graph:

- Majority of TM195 & TM498 customers have moderate income level.
- TM798 is bought by high income level customers. We might derive that TM798 is the costliest model.

4. Conclusion

We have seen from the aforementioned graphical data presentation and thus we can conclude:

- The data collected is for 3 specific Products – TM195, TM498 & TM798. TM195 being most popular among all Age groups and TM798 being brought by customers with high Income level and between Age group of 20-35 years.
- The data collected has a greater number of Males compared to Females. Males: Females:: 104:76
- It is observed that Male customers have slightly more income than Female customers.
- Male customers are more fitness concerned and exercise 7 times a week.
- Also, males run/walk more miles than females.
- Partnered (107) have shown more usage of the Treadmill compared to Singles (73). This information can be used for targeted marketing among partnered customers.
- TM195 seems to be the entry level variant and can be used by most of the customer groups (by Age, Usage, Fitness and Income Level). However, TM798 seems to be high end costlier model which is used by mostly fitness conscious customers with high income and higher level of education.
- Customers with high education level have higher household income.
- Overall conclusion: TM195 is a great product for beginner level and can be targeted among all segment of customers for sales. TM798 can be targeted among highly educated customers with high income level. The usage of the treadmill still needs some more concrete data for proper relationship.

5. Appendix A – Source Code

```
#=====================================================================#
#Data Analysis  –   Cardio Good Fitness
#Developer     -   Tahmid Bari
#Date          -   March 14, 2020
#=====================================================================#

## Loading linrary ##
library("readr")
library(ggplot2)
library(UsingR)
library(rgl)
library(scatterplot3d)
## scatterplot3d(Age, Education, Usage, main="3D Scatterplot", color="steelblue") ##

## Setting working directory ##
setwd("C:/Users/Tahmid Bari/Desktop/Great_Learning/R_Project/Cardio_Good_Fitness")
getwd()

## Import dataset | Reading a file ##
Cardio <- read.csv("CardioGoodFitness.csv")
attach(Cardio)

#Total Number of Observations and Variables #
dim(Cardio)
str(Cardio)

## Names of the Variables ##
names(Cardio)

## Check for Missing Values ##
colSums(is.na(Cardio))

## Summary of the 'Cardio' dataset ##
summary(Cardio)

## Check frequencies by each variable ##
table(Product)
table(Gender)
table(MaritalStatus)
table(Age)
table(Education)
table(Usage)
table(Fitness)
table(Income)

## *** Univariate Analysis *** ##
# By Product #
ggplot(Cardio, aes(x = Product, fill = Product)) +   theme_gray() +  geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(y = "Frequency by Product",
       title = "1. Product Sales")
```

```
# By Gender #
ggplot(Cardio, aes(x = Gender, fill = Gender)) +
  theme_gray() +
  geom_bar() + coord_polar("y", start = 0)+  theme_gray() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(y = "Frequency by Gender",
       title = "2. Treadmill Usage by Gender")

# By Marital Status #
ggplot(Cardio, aes(x = MaritalStatus, fill = MaritalStatus)) +
  theme_gray() +
  geom_bar() + #coord_polar("y", start = 0)+ theme_void() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(x = "Marital Status" ,
       y = "Frequency by Marital Status",
       title = "3. Treadmill Usage by Marital Satus")

## Create Partitions in the Display Panel ##

par(mfrow=c(1,3))

## Histogram plots to see the data by range ##
# By Age (in years) #
hist(Age,labels = TRUE, main='4. Treadmill Usage by Age',xlab = "Age (in years)",ylab = "Frequency",col =
heat.colors(10))

# By Education Level (in years) #
hist(Education,labels = TRUE,main='5. Treadmill Usage by Education Level',xlab = "Education (in years)",ylab =
"Frequency",col  = heat.colors(10))

# By Usage per week
hist(Usage,labels = TRUE,main='6. Treadmill Usage',xlab = "Usage (per week)",ylab = "Frequency",col =
heat.colors(10))

# By Age (in years) #
boxplot(Age,labels = TRUE,main='7. Treadmill Usage by Age',xlab = "Age (in years)",ylab = "Frequency",col =
heat.colors(10),horizontal = TRUE)
text(x=fivenum(Age), labels =fivenum(Age), y=1.25)

# By Education Level (in years) #
boxplot(Education,labels = TRUE,main='8. Treadmill Usage by Education Level',xlab = "Education (in
years)",ylab = "Frequency",col = heat.colors(10),horizontal = TRUE)
text(x=fivenum(Education), labels =fivenum(Education), y=1.25)

# By Usage per week #
boxplot(Usage,labels = TRUE,main='9. Treadmill Usage',xlab = "Usage (per week)",ylab = "Frequency",col =
heat.colors(10),horizontal = TRUE)
text(x=fivenum(Usage), labels =fivenum(Usage), y=1.25)

# By Fitness Level #
hist(Fitness,labels = TRUE,main='10. Treadmill Usage by Fitness Level',xlab = "Fitness Level",ylab =
"Frequency",col = heat.colors(10))
```

```
# By Income Level #
hist(Income,labels = TRUE,main='11. Income Level',xlab = "Income", ylab = "Frequency",col = heat.colors(10))

# By Miles #
hist(Miles,labels = TRUE,main='12. Miles Exercised',xlab = "Miles",ylab = "Frequency",col = heat.colors(10))


# Boxplots to identify any outliers in the data #

# By Fitness Level #
boxplot(Fitness,labels = TRUE,main='13. Treadmill Usage by Fitness Level',xlab = "Fitness Level",ylab =
"Frequency",col = heat.colors(10),horizontal = TRUE)
text(x=fivenum(Fitness), labels =fivenum(Fitness), y=1.25)

# By Income Level #
boxplot(Income,labels = TRUE,main='14. Income Level',xlab = "Income",ylab = "Frequency",col =
heat.colors(10),horizontal = TRUE)
text(x=fivenum(Income), labels =fivenum(Income), y=1.25)

# By Miles #
boxplot(Miles,labels = TRUE,main='15. Miles Exercised',xlab = "Miles",ylab = "Frequency",col =
heat.colors(10),horizontal = TRUE)
text(x=fivenum(Miles), labels =fivenum(Miles), y=1.25)

########## Bi-Variate Analysis #########
#### Categorical vs Categorical ####

## 16. Between Product and Gender ##
ggplot(Cardio, aes(x = Product, fill = Gender)) +
 theme_gray() +
 geom_bar() +
 geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
 labs(x = "Product",
     y = "Frequency",
     title = "16. Product puchased by Gender")

## 17. Between Product, Gender and Marital Status ##
ggplot(Cardio, aes(x = Product, fill = Gender)) +
 theme_gray() +
 facet_wrap(~ MaritalStatus) +
 geom_bar() +

 geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
 labs(x = "Product",
     y = "Frequency",
     title = "17. Product puchased by Marital Status and Gender")

## *** Bi-Variate Analysis *** ##

# Reset the earlier partition command #
dev.off()
```

```
# Set the new partiton window #
par(mfrow=c(2,3))

## Categorical vs Numerical ##

## BY PRODUCT ##
# 18. Product Vs Age #
plot(Product,Age, main='18. Product Vs Age',xlab = "Product",ylab = "Age (in years)",col = heat.colors(10))

# 19. Product Vs Education #
plot(Product,Education, main='19. Product Vs Education',xlab = "Product", ylab = "Education (in years)",col =
heat.colors(10))

# 20. Product Vs Usage #
plot(Product,Usage, main='20. Product Vs Usage',xlab = "Product", ylab = "Usage (per week)",col =
heat.colors(10))

# 21. Product Vs Fitness #
plot(Product,Fitness, main='21. Product Vs Fitness',xlab = "Product", ylab = "Fitness Level",col =
heat.colors(10))

# 22. Product Vs Income #
plot(Product,Income, main='22. Product Vs Income',xlab = "Product", ylab = "Income Level",col =
heat.colors(10))

# 23. Product Vs Miles #
plot(Product,Miles, main='23. Product Vs Miles',xlab = "Product", ylab = "Miles",col = heat.colors(10))

## BY GENDER ##
# 24. Gender Vs Age #
plot(Gender,Age, main='24. Gender Vs Age',xlab = "Gender", ylab = "Age (in years)",col = heat.colors(5))

# 25. Gender Vs Education #
plot(Gender,Education, main='25. Gender Vs Education',xlab = "Gender", ylab = "Education (in years)",col =
heat.colors(5))

# 26. Gender Vs Usage #
plot(Gender,Usage, main='26. Gender Vs Usage',xlab = "Gender", ylab = "Usage (per week)",col =
heat.colors(5))

# 27. Gender Vs Fitness #
plot(Gender,Fitness, main='27. Gender Vs Fitness',xlab = "Gender", ylab = "Fitness Level",col = heat.colors(5))

# 28. Gender Vs Income #
plot(Gender,Income, main='28. Gender Vs Income',xlab = "Gender", ylab = "Income Level",col = heat.colors(5))

# 29. Gender Vs Miles #
plot(Gender,Miles, main='29. Gender Vs Miles',xlab = "Gender", ylab = "Miles",col = heat.colors(5))


## BY marital Status ##
# 30. Marital Status Vs Age #
```

```
plot(MaritalStatus,Age, main='30. Marital Status Vs Age',xlab = "Marital Status", ylab = "Age (in years)",col = heat.colors(5))

# 31. Marital Status Vs Education #
plot(MaritalStatus,Education, main='31. Marital Status Vs Education',xlab = "Marital Status", ylab = "Education (in years)",col = heat.colors(5))

# 32. Marital Status Vs Usage #
plot(MaritalStatus,Usage, main='32. Marital Status Vs Usage',xlab = "Marital Status", ylab = "Usage (per week)",col = heat.colors(5))

# 33. Marital Status Vs Fitness #
plot(MaritalStatus,Fitness, main='33. Marital Status Vs Fitness',xlab = "Marital Status", ylab = "Fitness Level",col = heat.colors(5))

# 34. Marital Status Vs Income #
plot(MaritalStatus,Income, main='34. Marital Status Vs Income',xlab = "Marital Status", ylab = "Income Level",col = heat.colors(5))

# 35. Marital Status Vs Miles #
plot(MaritalStatus,Miles, main='35. Marital Status Vs Miles',xlab = "Marital Status", ylab = "Miles",col = heat.colors(5))

# Reset the earlier partition command #
dev.off()

# Set the new partiton window #
par(mfrow=c(2,3))

## Numerical vs Numerical ##
# 36. Age Vs Miles #
plot(Age,Miles, main='36. Age Vs Miles',xlab = "Age (in years)", ylab = "Miles",col = heat.colors(10))

# 37. Age Vs Education #
plot(Age,Education, main='37. Age Vs Education',xlab = "Age (in years)", ylab = "Education (in years)",col = heat.colors(10))

# 38. Age Vs Usage #
plot(Age,Usage, main='38. Age Vs Usage',xlab = "Age (in years)", ylab = "Usage",col = heat.colors(10))

# 39. Age Vs Income #
plot(Age,Income, main='39. Age Vs Income',xlab = "Age (in years)", ylab = "Income",col = heat.colors(10))

# 40. Age Vs Fitness #
plot(Age,Fitness, main='40. Age Vs Fitness',xlab = "Age (in years)", ylab = "Fitness",col = heat.colors(10))


# Reset the earlier partition command #
dev.off()

# Set the new partiton window #
par(mfrow=c(2,2))
```

```
# 41. Education Vs Usage #
plot(Education,Usage, main='41. Education Vs Usage',xlab = "Education (in years)", ylab = "Usage (per
week)",col = heat.colors(10))

# 42. Education Vs Income #
plot(Education,Income, main='42. Education Vs Income',xlab = "Education (in years)", ylab = "Income Level",col
= heat.colors(10))

# 43. Education Vs Miles #
plot(Education,Miles, main='43. Education Vs Miles',xlab = "Education (in years)", ylab = "Miles",col =
heat.colors(10))

# 44. Education Vs Fitness #
plot(Education,Fitness, main='44. Education Vs Fitness',xlab = "Education (in years)", ylab = "Fitness",col =
heat.colors(10))

# Reset the earlier partition command #
dev.off()

# Set the new partiton window #
par(mfrow=c(2,3))

# 45. Usage Vs Income #
plot(Usage,Income, main='45. Usage Vs Income',xlab = "Usage (per week)", ylab = "Income",col =
heat.colors(10))
# 46. Usage Vs Fitness #
plot(Usage,Fitness, main='46. Usage Vs Fitness',xlab = "Usage (per week)", ylab = "Fitness Level",col =
heat.colors(10))

# 47. Usage Vs Miles #
plot(Usage,Miles, main='47. Usage Vs Miles',xlab = "Usage (per week)", ylab = "Miles",col = heat.colors(10))

# 48. Fitness Vs Income #
plot(Fitness,Income, main='48. Fitness Vs Income',xlab = "Fitness", ylab = "Income",col = heat.colors(10))

# 49. Fitness Vs Miles #
plot(Fitness,Miles, main='49. Fitness Vs Miles',xlab = "Fitness", ylab = "Miles",col = heat.colors(10))

# 50. Income Vs Miles #
plot(Income,Miles, main='50. Income Vs Miles',xlab = "Income", ylab = "Miles",col = heat.colors(10))


## 3.7 Variable Transformation: Provide meaningful definitions to few Numeric Variables for better
distribution ##

Age_Classification <- ifelse(Age<20, "Teenagers",
                      ifelse(Age>35, "Middle Aged Adults",
                          "Young Adults"))

Usage_Classification <- ifelse(Usage<3, "Amateurs",
                        ifelse(Usage>5, "Pro",
                            "Regular"))
```

```
Fitness_Classification <- ifelse(Fitness == 5, "Very Fit",
                          ifelse( Fitness <=1, "Very Unfit",
                                  "Fit"))


Income_Classification <- ifelse(Income<40000, "Low Income",
                         ifelse(Income>70000, "High Income",
                                "Moderate Income"))

Miles_Classification <- ifelse(Miles<60,  "Low Usage",
                        ifelse(Miles>130, "High Usage",
                               "Medium Usage"))

# Create a new data set with meaningful variables
Cardio_desc =
cbind(Cardio,Age_Classification,Usage_Classification,Fitness_Classification,Income_Classification,Miles_Classifi
cation)

# View the new data set created
View (Cardio_desc)

## Create 3 subsets of the new dataset ##
product1 <- Cardio_desc[which(Cardio_desc$Product == "TM195"),]
product2 <- Cardio_desc[which(Cardio_desc$Product == "TM498"),]
product3 <- Cardio_desc[which(Cardio_desc$Product == "TM798"),]

product_by_age_class = sqldf("select Product,Age_Classification, avg(Age) Avg_Age, count(Product)
Product_Count
                        from Cardio_desc group by Product,Age_Classification")
View(product_by_age_class)

# Average Age of Customers
product_by_age_class = sqldf("select avg(Age) Avg_Age from Cardio_desc ")

## Graphical representation by Age group Classification##
ggplot(Cardio_desc, aes(x = Product, fill = Gender)) +
  theme_gray() +
  geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(x = "Product" ,
       y = "Frequency",
       title = "51. Product used by Age Classification")


## Graphical representation by Usage Classification##
ggplot(Cardio_desc, aes(x = Product, fill = Usage_Classification)) +
  theme_gray() +
  geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(x = "Product" ,
       y = "Frequency",
       title = "52. Product used by Usage Classification")
```

```
## Graphical representation by Fitness Classification##
ggplot(Cardio_desc, aes(x = Product, fill = Fitness_Classification)) +
  theme_gray() +
  geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(x = "Product" ,
       y = "Frequency",
       title = "53. Product used by Fitness Classification")

## Graphical representation by Income Classification##
ggplot(Cardio_desc, aes(x = Product, fill = Income_Classification)) +
  theme_gray() +
  geom_bar() +
  geom_text(aes(label=..count..),stat="count",position=position_stack(0.5)) +
  labs(x = "Product" ,
       y = "Frequency",
       title = "54. Product used by Income Classification")
```