CARDIO FITNESS PROJECT

Karthik Paranthaman

MODEL REPORT

1. PROJECT OVERVIEW

The main aim of this project is to explore the data set provided by the client "Cardio Good Fitness" to find business insights and improve sales.

The R-Studio Platform is used to analyze the dataset using descriptive statistical tools such as Mean, Median, IQR, Standard Deviation and Variances.

The outputs are visualized using varies graphical methodologies to interpret with an aid of plots like Histogram, Bar Plot, Box Plot, Density Plot, and etc.

2. ASSUMPTIONS

I feel the variable fitness level is very subjective to customer's opinion and the mindset he or she was at during data collection period. Therefore, it is advisable to do analysis on Fitness Level and Miles Ran to confirm a linearity before including the Fitness variable into further analysis. It was assumed that all information provided by the customers are accurate. This will be confirmed using Sanity Check methodologies.

3. EXPLORATORY DATA ANALYSIS

In this Chapter we will be discussing about the data exploration approaches to find business insights. The exploratory process consists of the following stages.

1. Environment Setup and Data Import
2. Variable Identification
3. Univariate Analysis
4. Bi-Variate Analysis
5. Missing Value Identification
6. Variable Transformation
7. Feature Creation & Exploration

3.1 Environment Setup and Data Import

3.1.1 WORKING DIRECTORY

The working directory has been set. The dataset is called using the name "Cardio_Data" in R-Studio (Refer to APPENDIX -1 – for R Code).

3.1.2 LIBRARY PACKAGES

The library packages used in this project are listed as follows:

1. readr
2. ggplot2
3. rpivotTable
4. dplyr
5. esquisse

3.2 Missing Value Identification

The Cardio_Data dataset contains 9 Variables and 180 observations.

Sanity check was done to make sure there are no missing values in the dataset by using the following code :

```
anyNA(Cardio_Data) # Outputs if any Missing values were spotted
```

In addition to that "head" & "tail" commands are used to call top & bottom rows to identify for any formatting issue in the dataset given.

All identified character classes are converted into "factor" data type using "as.factor…" command.

Furthermore, output from "summary (Cardio_Data)" command also approves the Sanity Check from missing values and data classes.

3.3 UNIVARIATE ANALYSIS

5 Number Summary analysis is done for each variable and the output is given below.

```
##   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
```

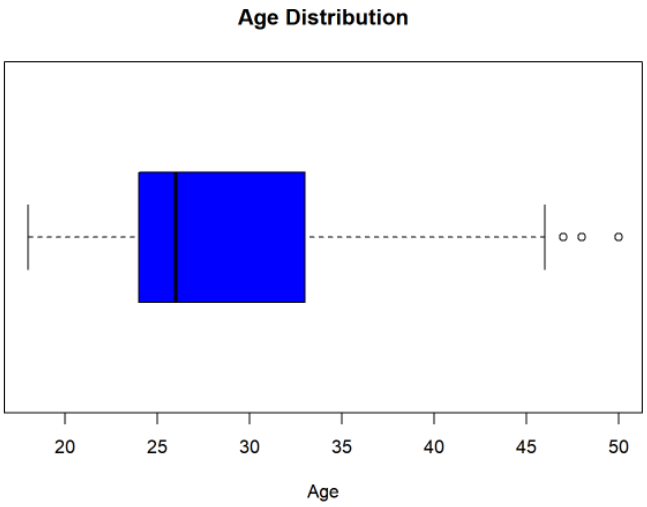
We can say that observations in Age, Usage, Fitness, Income and Miles are distributed as "Positively Skewed" manner while observations from Education variable is "Negatively Skewed".

Majority of the customers are Male and Married. TM195 seems to be the most popular product among the TM series by capturing a prime portion of 44.4% of the overall sales. The upcoming Chapters will be discussing the detailed analysis of each variable.

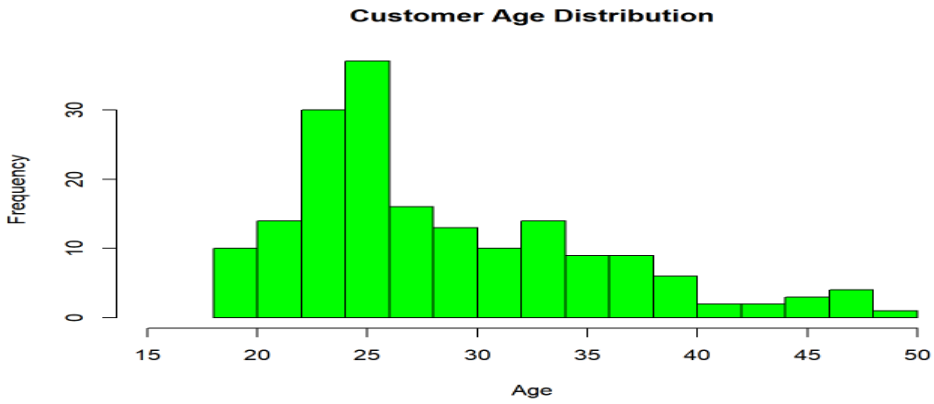
AGE:

Firstly, Box Plot tool is used on Age. Refer to the Appendix-1 for the code. The youngest customer being 18 and eldest is at 50. The diagram below tells us there are 3 outliers. However, repeated outliers can be identified using R code, mentioned in the Appendix -1.

The outliers are 47, 50,48,47,48. In other words any value which is identified in Age column more than 46.5 (i.e.: Maximum IQR).

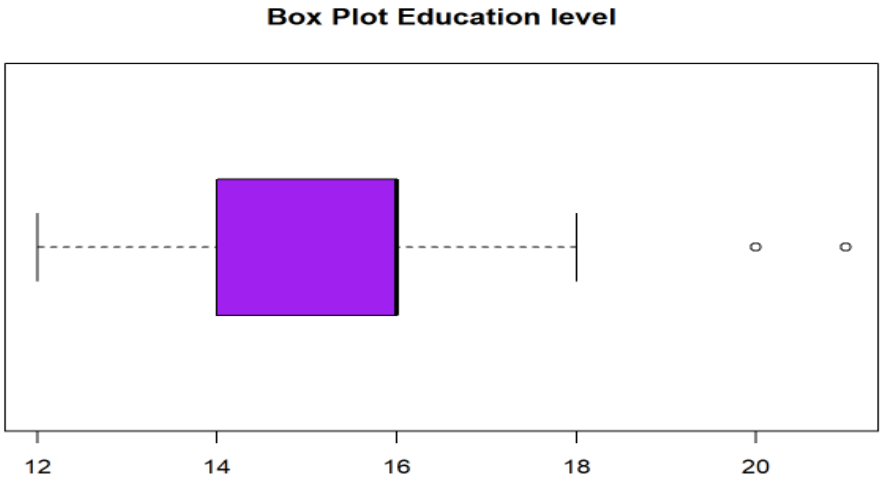


The customer's age distribution is viewed through histogram plot which is strongly skewed to the right. Most of the customers are in their mid-20s.

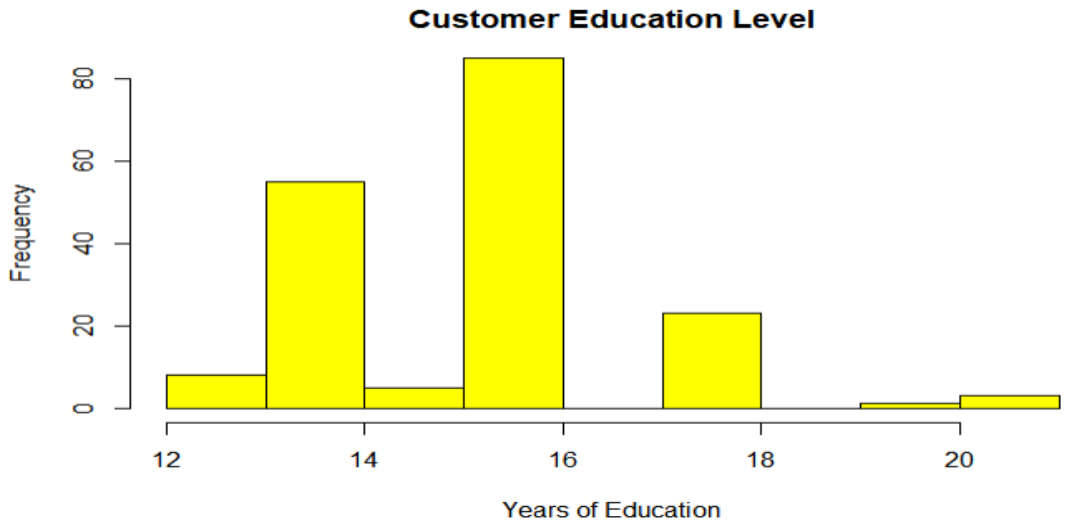


EDUCATION:

Analysis on the Educational background of the customer is done by plotting the observations in box plot.



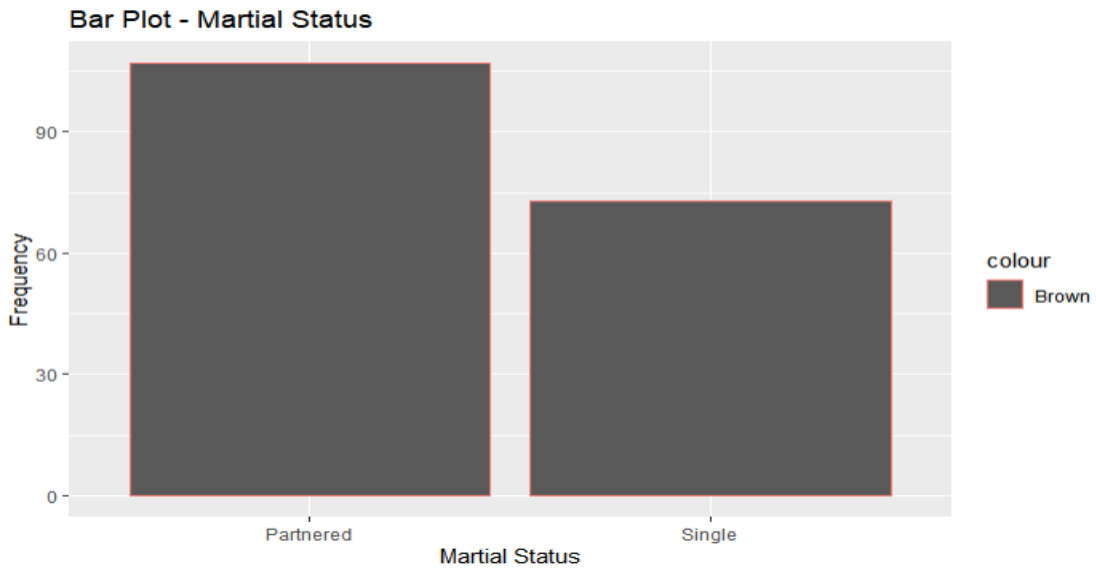
The minimum years of continued education by the customer is for 12 years while most number of continuous education for about 21 years. It was found that there are 4 positive outliers which are 20,21,21,21.



The histogram shows, that dataset contains high number of customers have followed education for 16 years while the overall average years of education found to be 15 years.

MARITIAL STATUS

The univariate analysis on marital status was done using a “qplot” tool and found that married couples are very health conscious and this can be further analysed in depth in upcoming chapters by comparing information on their "Fitness Level”. Married customers dominate about 60% of the collected dataset.

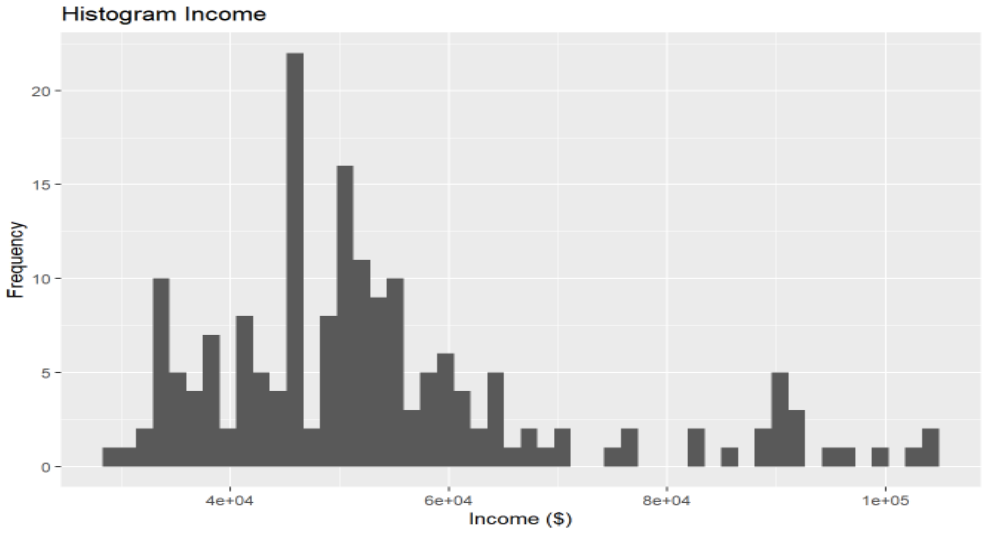


INCOME:

It was found that the lowest salary is approximately US$ 29,500 and the highest being around US$ 104,500 during the box plot analysis on customer Income. This information would be an useful indicator when it comes to pricing on TM series product. In order to understand this variable better we shall identify the outliers on income data observations with an aid outlier calculations in R.

There are 19 positive outliers found in this dataset.

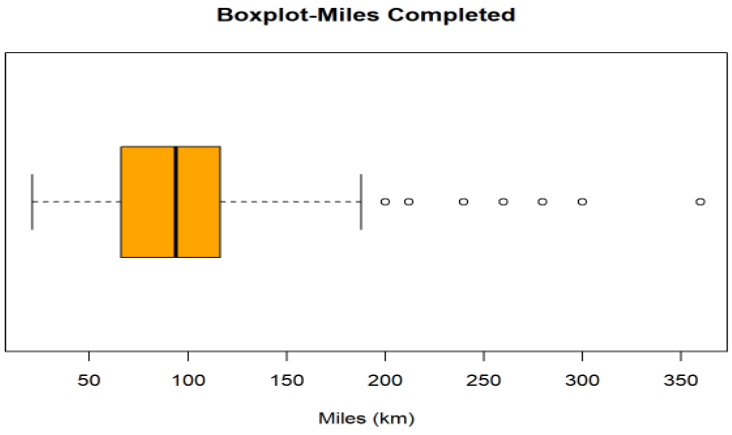
I e: 83416, 88396, 90886, 92131, 88396, 85906, 90886, 103336, 99601, 89641, 95866, 92131, 92131, 104581, 83416, 89641, 90886, 104581 and 95508.



The histogram on Customer Income is positively skewed. The average salary among the customers is around US$ 53,700 and median is around US$ 50,590.

MILES:

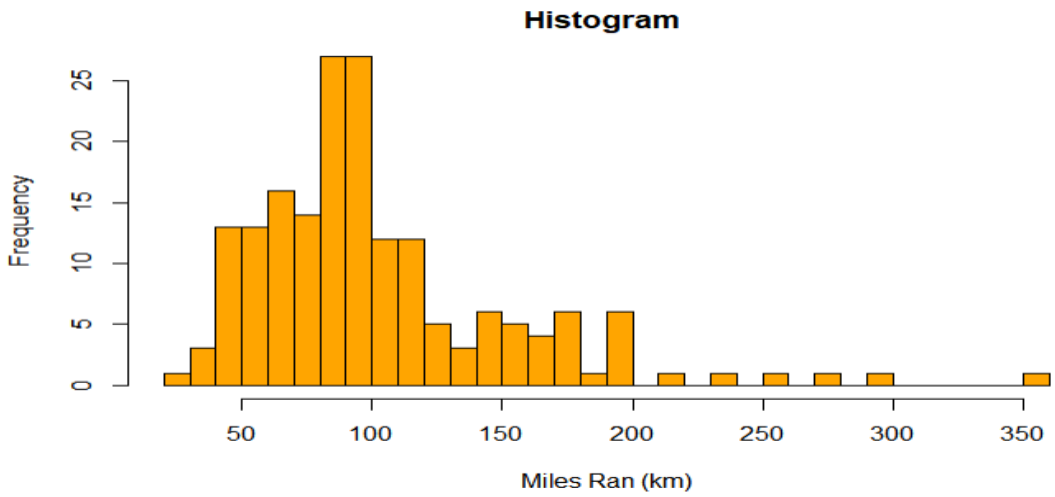
In this section of the Chapter we will be discussing on Univariate Analysis on the Miles ran by each customer using TM series products. The Box Plot Output is:



Box plot output shows some outliers. The calculation for outliers R code is shown in APPENDIX – 1.

The number of outliers are 13.

Outliers: 188,212,200,200,200,240,300,280,260,200,360,200,200

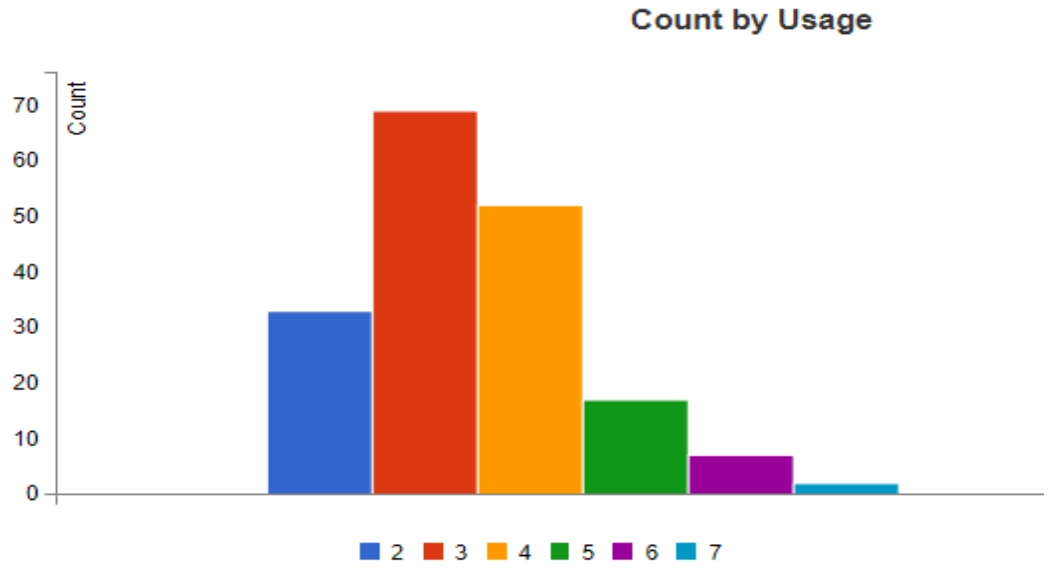


From above histogram we can spot that mean (103.2 km) of the distribution is greater than the median (94) . Hence, distribution is skewed to the right.

USAGE:

There are 9 outliers can be identified with Usage of the threadmills per week from the given dataset.

The outliers are : 6, 6, 6, 7, 6, 7, 6, 6, 6



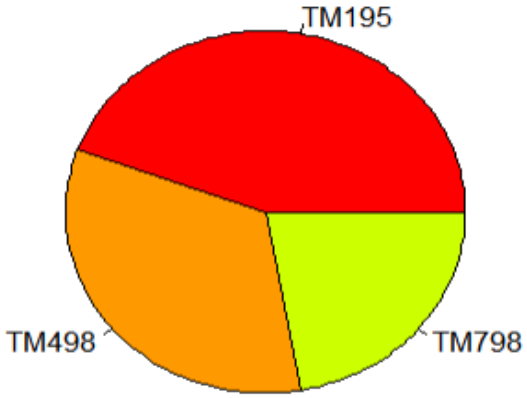
As per the histogram most of the candidates workout 3 times a week. It is interesting to see some of the candidates who work out all 7 days a week.

PRODUCT:

There are 3 types of TM series product was sold among this 180 customers. The sales number is tabulated below.

PRODUCT	NUMBER OF UNITS	%
TM 195	80	44.4
TM 498	60	33.3
TM 798	40	22.2

The most units were sold on TM195 series and the second most is the TM 498 product. The pie- chart gives a graphical representation on the units sold on TM series items.



3.4 Variable Transformation / Feature Creation

Since there are wide range of observations on customer income, I have divided into 4 category of classes based on their salary.

Salary range from,

$20,000 to $40,000 = Low Class

$40,000 to $60,000 = Low mid Class

$60,000 to $80,000 = Upper Mid Class

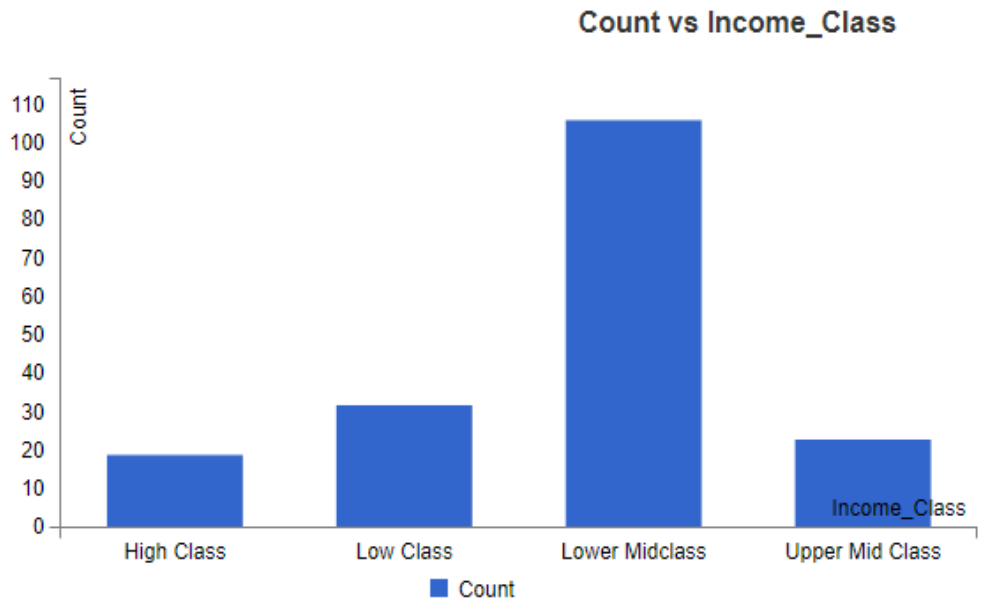
$80,000 and above = High Class

"mutate" code is used to do this transformation. Refer the APPENDIX – 1.

The above transformation is saved as new data frame called "Income_Class" contains of 180 observations with 10 variables. The top 10 entries of the dataset is displayed below.

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	Income_Class
1	TM195	18	Male	14	Single	3	4	29562	112	Low Class
2	TM195	19	Male	15	Single	2	3	31836	75	Low Class
3	TM195	19	Female	14	Partnered	4	3	30699	66	Low Class
4	TM195	19	Male	12	Single	3	3	32973	85	Low Class
5	TM195	20	Male	13	Partnered	4	2	35247	47	Low Class
6	TM195	20	Female	14	Partnered	3	3	32973	66	Low Class
7	TM195	21	Female	14	Partnered	3	3	35247	75	Low Class
8	TM195	21	Male	13	Single	3	3	32973	85	Low Class
9	TM195	21	Male	15	Single	5	4	35247	141	Low Class
10	TM195	21	Female	15	Partnered	2	3	37521	85	Low Class

The histogram below shows the frequency of occurrence among 4 classes that have been created based on customer income range.

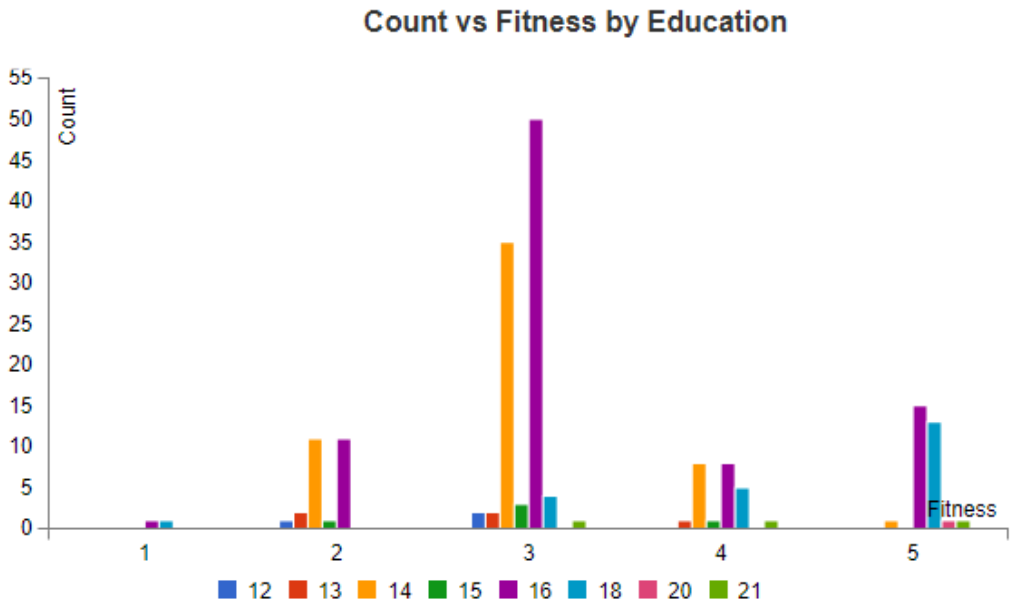


Most customers fall in Lower mid class financial category and High Class customer who earn more than US$ 80,000 being the lowest.

3.5 Bi-Variant Analysis

EDUCATION Vs FITNESS

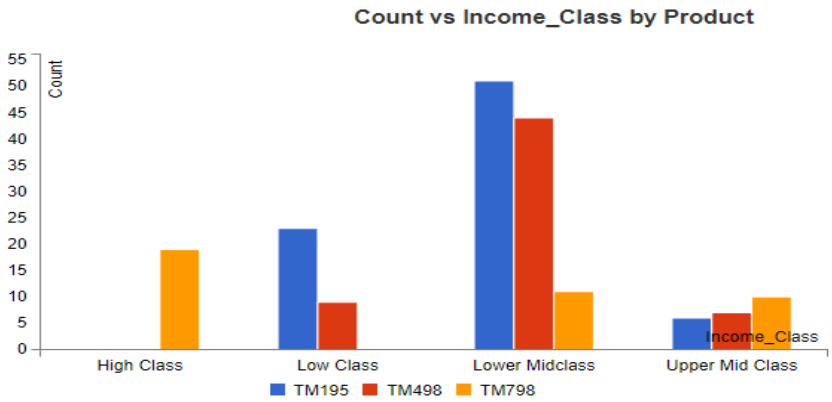
I have decided to build a relationship between Education and Fitness level of the customers to analyze the trend in purchasing. The following bar plot indicates the aforementioned relationship.



People at a range of 14 to 16 years of Education dominates above and below the average fitness level of 3.111. Strongest customers can be spotted with the education level of more than 18 years of experience.

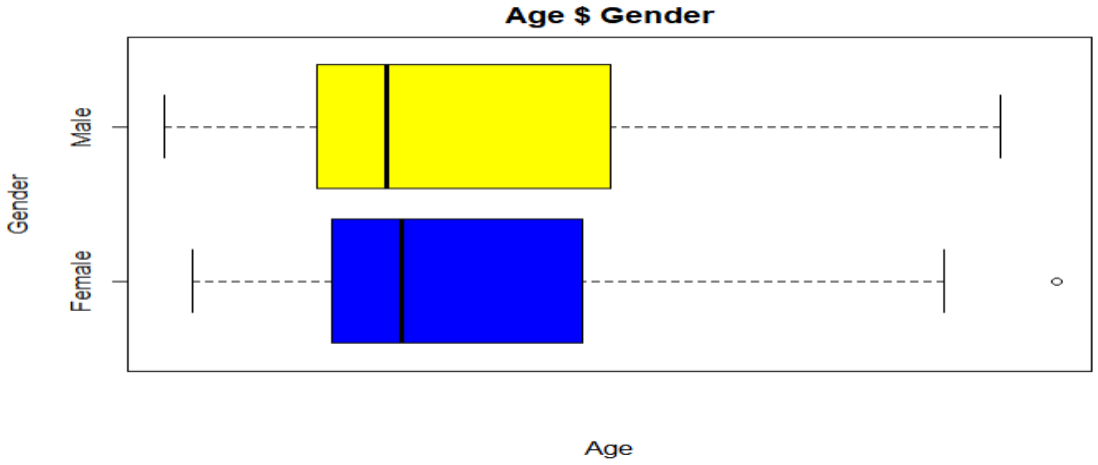
On average people more than 16 years of educational background have realized the importance of having a healthy lifestyle and showing a steady maintenance on their fitness lifestyle.

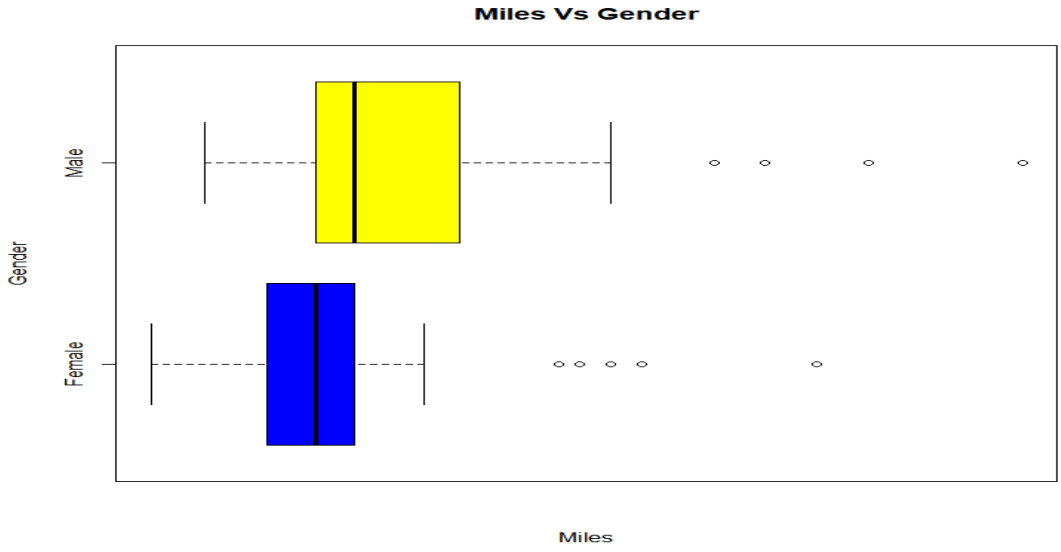
INCOME Vs FITNESS

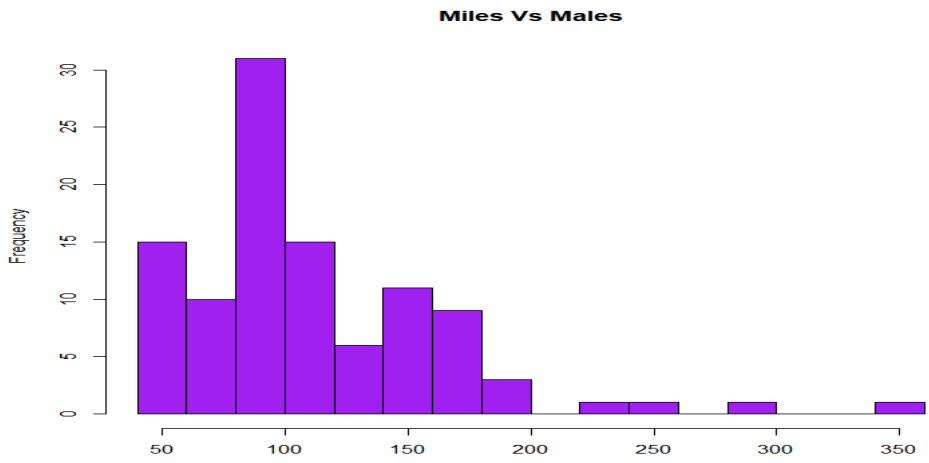


The above bar chart gives the clear insights of sales of TM series product and the financial background of the customer. People at Lower Mid Class are interested on TM 195. High class people are very much interested in TM 798 product.

GENDER Vs MILES







In this section, box plot comparisson is done on miles run between Male and Female customers. In addition to that, It was identified that most men are eleder to women in the dataset given. Larger

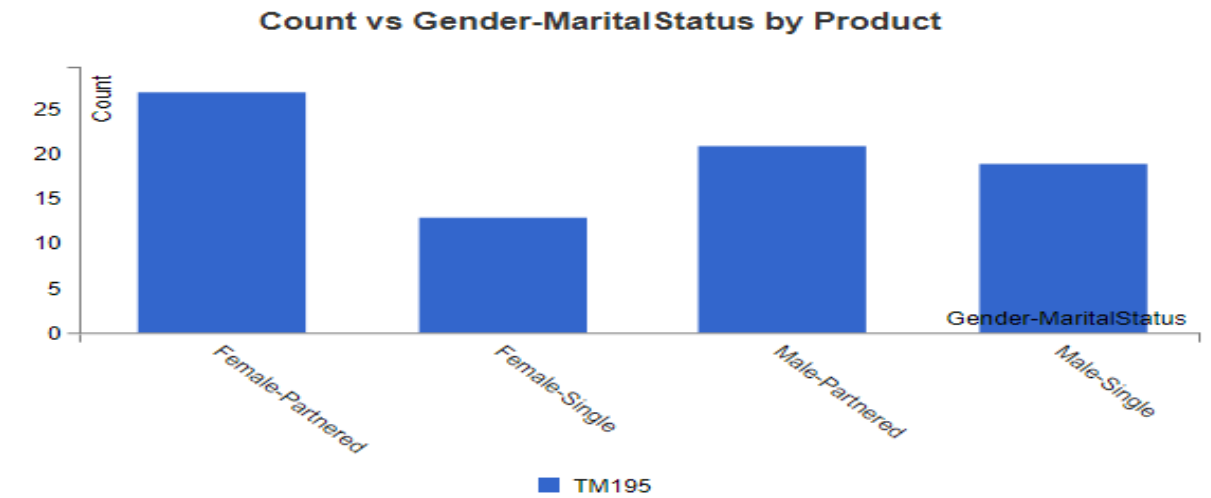
peaks can be obseved around 100 miles and the overall mileage ran by men is more positvely skewed in nature.

Therefore, it can be taken into account during business decision that most Males have reached high miles for the given period . Therefore the main target customer group based on fitness and mileage should be Males.

TM 195

The “filter” command is used to isolate the customer information on most popular product TM 195 and a separate dataset is created called TM195. The “head” command have generated top ten observation from the dataset.

	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
7	TM195	21	Female	14	Partnered	3	3	35247	75
8	TM195	21	Male	13	Single	3	3	32973	85
9	TM195	21	Male	15	Single	5	4	35247	141
10	TM195	21	Female	15	Partnered	2	3	37521	85



Bi

Variate analysis on TM 195 between Customer's Age and Martial Status done through histogram plot. Married women have shown very high interest on maintaining their figure with the popular TM series TM 195.

4.0 CONCLUSION

Based on the analysis from previous chapters it can said that management must focus on maintaining a upper trend on sales of popular product like TM195 by designing their marketing strategies not only centered through financial aspect of the customer also reaching out to people who are well educated and especially, towards married women.

Furthermore, to improve sales on the TM 798 product, business offers could be redesigned in such a way to target high class customers.

5.0 APPENDIX

```
#Installing Library Package to Read & Write .csv files

install.packages("readr")

library(readr)

#Installing Library Package to work with plots

install.packages("ggplot2")

library(ggplot2)

#Installing Library Packages to work with Table

library(rpivotTable)

#Installing Library Packages to Work with data manipulation

library(dplyr)


#=============================================================================

# SETTING UP WORKING DIRECTORY & IMPORTING DATA

```{r message=FALSE, warning=FALSE}


setwd("C:/Users/user/Desktop/Data Science/Week 2/PROJECT REPORT/Cardio_Project_Report")


getwd()

Cardio_Data = read_csv("CardioGoodFitness.csv") # Calling the Data Set .csv file
```
```

```
#=============================================================================

#SANITY CHECK
str(Cardio_Data)

# DATA TYPE CONVERSION
Cardio_Data$Product=as.factor(Cardio_Data$Product)
Cardio_Data$Gender=as.factor(Cardio_Data$Gender)
Cardio_Data$MaritalStatus=as.factor(Cardio_Data$MaritalStatus)
summary(Cardio_Data)

# MISSING VALUE IDENTIFICATION
anyNA(Cardio_Data) # Outputs if any Missing values were spotted
sum(is.na(Cardio_Data)) # Number of Missing Values

# ANALYSING FORMATTING ISSUES
head(Cardio_Data,10)
tail(Cardio_Data,10)

#=============================================================================
#UNIVARIATE ANALYSIS

#AGE
range(Cardio_Data$Age)
boxplot(Cardio_Data$Age,horizontal = TRUE,col="Blue",main = "Age Distribution", xlab= "Age")

# OUTLIER CALCULATION
mean(Cardio_Data$Age)
```

```
median(Cardio_Data$Age)
Age_IQR = IQR(Cardio_Data$Age)
Age_IQR

#OUTLIERS - Q3
Q3 = quantile(Cardio_Data$Age,0.75)
Q3 = unname(Q3)
Q3

#OUTLIERs - Q1
Q1 = unname(quantile(Cardio_Data$Age,0.25))
Q1

Max_Age_IQR = Q3+1.5*Age_IQR # Positive Outlier Calculation
Max_Age_IQR

Min_Age_IQR = Q1-1.5*Age_IQR # Negative Outlier Calculation
Min_Age_IQR

# No Of OUTLIERS
sum(Cardio_Data$Age > Max_Age_IQR)

# Print all Outliers
Cardio_Data$Age[Cardio_Data$Age > Max_Age_IQR]
Cardio_Data$Age[Cardio_Data$Age < Min_Age_IQR]

# Histogram
hist(Cardio_Data$Age,col="Green",xlim=c(15,50),
    main="Customer Age Distribution",xlab="Age",breaks = 15)
```

```
#================================================================================
# EDUCATION
mean(Cardio_Data$Education)
median(Cardio_Data$Education)
Education_IQR = IQR(Cardio_Data$Education)
Education_IQR

#7. OUTLIERS
Q3 = quantile(Cardio_Data$Education,0.75)
Q3 = unname(Q3)
Q3

#6. Q1
Q1 = unname(quantile(Cardio_Data$Education,0.25))
Q1
Max_Education_IQR = Q3+1.5*Education_IQR
Max_Education_IQR

Min_Education_IQR = Q1-1.5*Education_IQR
Min_Education_IQR

#8. No Of OUTLIERS
sum(Cardio_Data$Education > Max_Education_IQR)

#9. Print all Outliers
Cardio_Data$Education[Cardio_Data$Education > Max_Education_IQR]
Cardio_Data$Education[Cardio_Data$Education < Min_Education_IQR]
```

```
# BOX PLOTS & HISTOGRAMS
range(Cardio_Data$Education)
hist(Cardio_Data$Education,col= "yellow",main = "Customer Education Level"
    ,xlab="Years of Education",xlim=c(10,22))

boxplot(Cardio_Data$Education,horizontal=TRUE,col="Purple",
       main = "Box Plot Education level",Xlab = "Years")

#================================================================================
# MARTIAL STATUS
qplot(MaritalStatus, data=Cardio_Data,col="Brown", main="Bar Plot - Martial Status",
     xlab="Martial Status",ylab="Frequency")
#================================================================================
# INCOME

range(Cardio_Data$Income)

Income_IQR =IQR(Cardio_Data$Income) # IQR Calculation
Income_IQR

Q3 = unname(quantile(Cardio_Data$Income,0.75)) # The Third Quatile Calculation
Q3

Q1 = unname(quantile(Cardio_Data$Income,0.25)) # The First Quantile Calculation
Q1

Max_Income_IQR = Q3 + 1.5*Income_IQR
Min_Income_IQR = Q1 - 1.5*Income_IQR
```

```
sum(Cardio_Data$Income > Max_Income_IQR)#Number of Positive Outliers
sum(Cardio_Data$Income < Min_Income_IQR)#Number of Negative Outliers

Cardio_Data$Income[Cardio_Data$Income > Max_Income_IQR]#Print all Income outliers

qplot(Income, data = Cardio_Data,main = "Histogram Income",
     xlab = "Income ($)",ylab="Frequency",bins = 50) # Histogram plot

boxplot(Cardio_Data$Income, horizontal = TRUE , col="Brown",
       main = "Boxplot - Income", xlab="Income") # BOX PLOT

mean(Cardio_Data$Income)
median(Cardio_Data$Income)

#================================================================================
# MILES
boxplot(Cardio_Data$Miles,horizontal = TRUE,col="Orange" ,main= "Boxplot-Miles Completed",
       xlab="Miles (km)") # BOX PLOT

Miles_IQR = IQR(Cardio_Data$Miles)
Miles_IQR

Q3 = unname(quantile(Cardio_Data$Miles,0.75))
Q3

Q1 =unname(quantile(Cardio_Data$Miles,0.25))
Q1

Max_Miles_IQR = Q3 + 1.5*Miles_IQR
```

```
Min_Miles_IQR = Q1 - 1.5*Miles_IQR



sum(Cardio_Data$Miles > Max_Miles_IQR) # Number of Outliers

Cardio_Data$Miles[Cardio_Data$Miles > Max_Miles_IQR] # Outliers on the Right

Cardio_Data$Miles[Cardio_Data$Miles < Min_Miles_IQR] # Outliers on the Left



hist(Cardio_Data$Miles, horizontal = TRUE,col = "orange",main = "Histogram ",

    xlab = "Miles Ran (km)" ,breaks = 15)

#=============================================================================

# USAGE

#=============================================================================

# PRODUCT

Cardio_Product_Table=table(Cardio_Data$Product)

pie(Cardio_Product_Table,col=rainbow(10))

#=============================================================================

# VARIABLE TRANSFORMATION / FEATURE CREATION

#=============================================================================

Income_Class = mutate(Cardio_Data,Income_Class =

                ifelse(Cardio_Data$Income > 20000 & Cardio_Data$Income < 40000,

                    "Low Class",ifelse(Cardio_Data$Income > 40000 & Cardio_Data$Income
<60000,"Lower Midclass",

                                ifelse(Cardio_Data$Income>60000 & Cardio_Data$Income<80000,

                                    "Upper Mid Class","High Class" ))))

Income_Class

# NOTE :  $20,000 between $40,000 = Low Class

#        $40,000 between $60,000 = Low Mid Class
```

```
#       $60,000 between $80,000 = Upper Mid Class
#       $80,000 and above           = High Class
#=============================================================================
# BI VARIATE ANALYSIS
#=============================================================================

rpivotTable(Cardio_Data) # This function is used to do Bivariate Analysis on "Cardio_Data" Data Set.
rpviotTable(Income_Class) # This function is used to do Bivariate Analysis on "Income_Class" Data Set.

# FILTERING POPULAR PRODUCTS
TM195=filter(Cardio_Data,Product =="TM195")
Eduaction= Cardio_Data[order(Cardio_Data$Education),]
View(TM195)
View(head(TM195,10))
rpivotTable(TM195)

# Box Plot comparison on
boxplot(Age~Gender,data=Cardio_Data,horizontal=TRUE,col=c("Blue","Yellow"),
      main = "Age $ Gender",Xlab="ABC",xaxt="n")
boxplot(Miles~Gender,data=Cardio_Data,horizontal=TRUE,col=c("Blue","Yellow"),
      main = "Miles Vs Gender",Xlab="ABC",xaxt="n")

hist(Cardio_Data$Miles[Cardio_Data$Gender=="Male"],col="Purple",main="Miles Vs Males",Xlab="Km
run by Males",breaks=15)


sum(Cardio_Data$Miles > Max_Miles_IQR) # Number of Outliers

Cardio_Data$Miles[Cardio_Data$Miles > Max_Miles_IQR] # Outliers on the Right
```

```
Cardio_Data$Miles[Cardio_Data$Miles < Min_Miles_IQR] # Outliers on the Left
```