KUDAKWASHE NYIKADZINO

Project - Cardio Good Fitness

PGP-DSBA

Table of Contents

1 Project Objective.............................................................................................................................2

2 Assumptions....................................................................................................................................2

3 Exploratory Data Analysis – Step by step approach .......................................................................2

3.1 Environment Set up and Data Import.....................................................................................2

3.1.1 Install necessary Packages and Invoke Libraries.............................................................2

3.1.2 Set up working Directory ................................................................................................2

3.1.3 Import and Read the Dataset..........................................................................................2

3.2 Variable Identification.............................................................................................................2

3.2.1 Variable Identification – Inferences................................................................................3

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

3.4 Bi-Variate Analysis...................................................................................................................9

3.5 Missing Value Identification....................................................................................................19

3.6 Outlier Identification...............................................................................................................19

3.7 Variable Transformation / Feature Creation ..........................................................................19

4 Conclusion.......................................................................................................................................20

5 Appendix A – Source Code..............................................................................................................21

5 Appendix B – Statistical Summaries…………………………………………………………………………………………….29

1 Project Objective

The objective of the report is to identify the characteristics of the different customers Cardio Good Fitness’s three treadmill models using 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

* All values in the dataset are capture correctly
* The self-ratings given by customers are an accurate representation
* The dataset reflects a good representation of the customers

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

Please refer to Appendix A for source code.

3.1.2 Set up working Directory

Please refer to Appendix A for Source Code.

3.1.3 Import and Read the Dataset

Please refer to Appendix A for Source Code.

3.2 Variable Identification

3.2.1 Functions used

dim()

The dim() function gives the dimensions of the data set (“CardioGoodFitness”). The data set has 180 observations (rows) and 9 variables (columns).

head() and tail()

The head() function gives the 6 observations of the data set by default and it also provides the variables or columns that make up the data set. The tail() functions gives the last 6 observations of the data set by default. Having a view of the first or last ‘n’ rows gives an idea of the data types contained in the data set. The functions also give the variable names. The use of both functions will also confirm the dimensions of the data set.

str()

The str() function gives the structure of the data set where the data type for each variable is stated. Factor variables like Product, it gives the number of factor levels or Products. It also states the data structure in this case it’s a data frame. It also confirms the dimensions of the data set.

summary()

The summary() function gives a summary of each variable and the summary depends on the data type of the variable. For example, factor variables like Gender will have the number of factors or levels in this case Female and Male. For numeric variables like Age, it gives the five number summary of these variables which includes minimum, Q1,Q2,Q3 and maximum.

anyNA()

The anyNA() function helps identify the presence of missing values.

sum()

the function is used to add the total number of NAs in the dataset.

table()

The table function is being used to generate a simple table summary for the variable(s) include in the arguments.

prop.table()

The table function is being used to generate a simple table summary with proportions for the variable(s) include in the arguments.

option()

This function is being used to set the number of decimal places for the numbers as well as removing scientific notation.

plot\_grid() and par()

These functions is being used to plot multiple graphs on one area, with plot grid being used for ggplot graphs and par() for the base function graphs.

3.2.2 Variable Identification – Inferences

The variable names are specified in the correct syntax, and there is no need to use the “make.names()” function to put them in the correct form. Based on the dimensions, all 9 variables have a concise and descriptive name. The naming approach is consistent with all names starting with a capital letter.

From using the head() and tail() functions, there are no formatting issues that appear. There are also no headers or footers included in the data set.

All the data types are correctly specified that is “Product”, “Gender” and “MaritalStatus” are factor variables and the rest are integer variables.

There are no missing values in the data set.

3.3 Univariate Analysis

**Univariate Categorical Variable Observations**

There are three categorical variables under consideration namely, Product, Gender and Marital Status. Gender and Marital Status each have two levels, and these relate to the customer. Product has three levels which relate to the models sold by Cardio Good Fitness.

**Product**

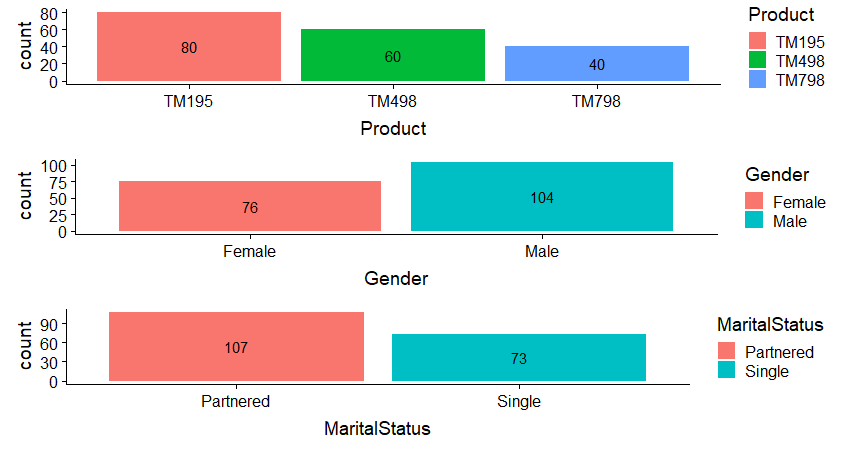
There are three model types of treadmills under consideration sold by Cardio Good Fitness (shown below). The most sales come from the TM195 model with 80 sales which is about 44% of the total treadmills sold. The TM498 model has 60 sales which is about 33% of total sales and the TM798 with the least sales at 40 which is about 22% of total sales. Cardio Good Fitness’ customers have purchased twice as many TM195 treadmills compared to the TM798 treadmills. There is a difference of 20 treadmills sold between the TM195 and the TM498 and between the TM498 and the TM798.

**Gender**

Most of the treadmill customers for three models are male at 104 in total compared to 76 female customers (shown below). Male clients make up approximately 58% of the total customers.

**Marital Status**

Most of the customers who purchased these three treadmill models are partnered. The total number of partnered customers is 107 which is about 59% of customers and single customer making up the remaining 41% which is 73 customers in total.

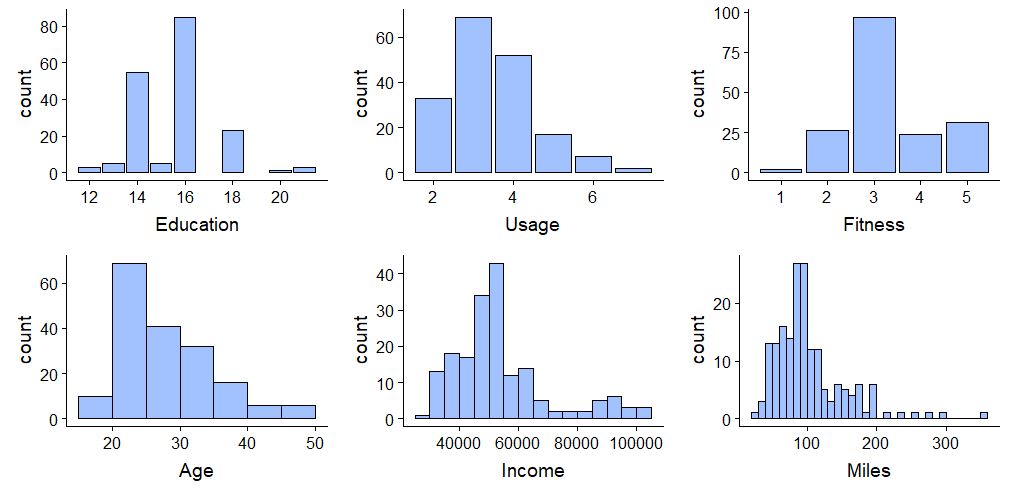


**Univariate Numeric Variable Observations**

There a total of six numeric variables and all share the same data type which is integer. The following information relates to these six variables.

* Age – relates to the customer and it is in years
* Education – relates to the customer and it is in years
* Usage – average number of times the customer intends to use the treadmill
* Fitness – self rating of fitness ranging from 1 – very unfit to 5 – very fit
* Income – relates to the customer
* Miles – distance completed on treadmills by customers

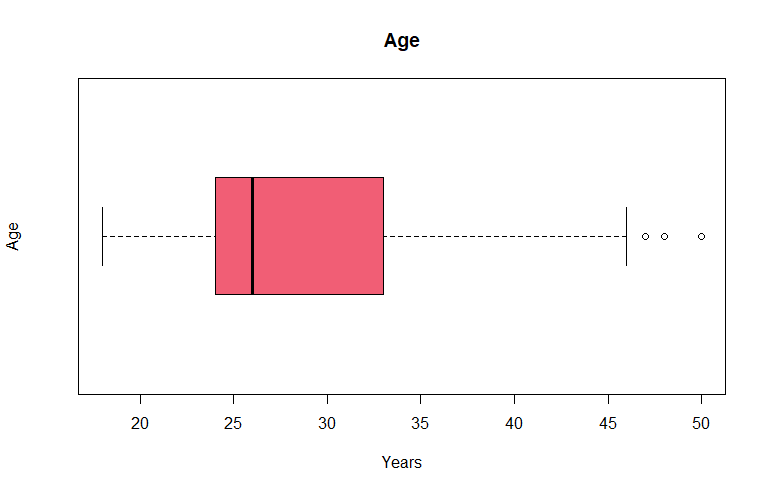
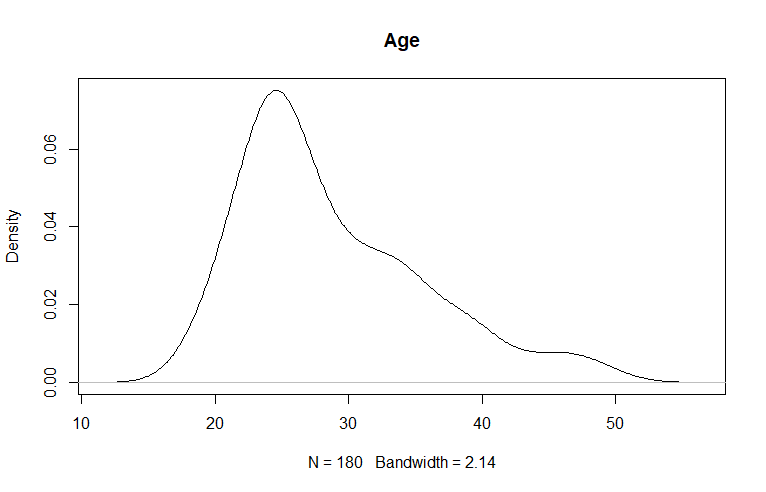
The graphs below show the pattern or distribution of values for each variable across all customers. Bar graphs and histograms are used to show these patterns. Appendix B shows all summary statistics referred to in the following pages.



**Age**

The average age for the customers is close to 29 years and the median age is 26 years and most of the customers are in the 20 – 25 years age group. With the mean, median and mode for Age all different from each other this indicates the age of the customers is skewed. The histogram confirms this skew which shows Age is not distributed symmetrically around the average age. The distribution of Age is right skewed which means the age group where most of the customers are in is less than the average age as well as the median age. The boxplot below confirms this as the median is less than the middle of the box in the plot. The right skew is confirmed by the longer whisker on the right in the box plot. There are outliers that exist in the customers’ ages. The older customers are the outliers in the group of customers for the products.

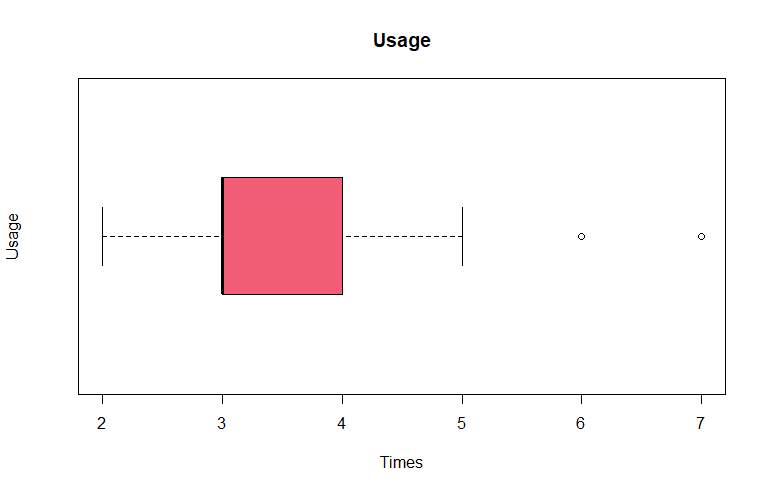
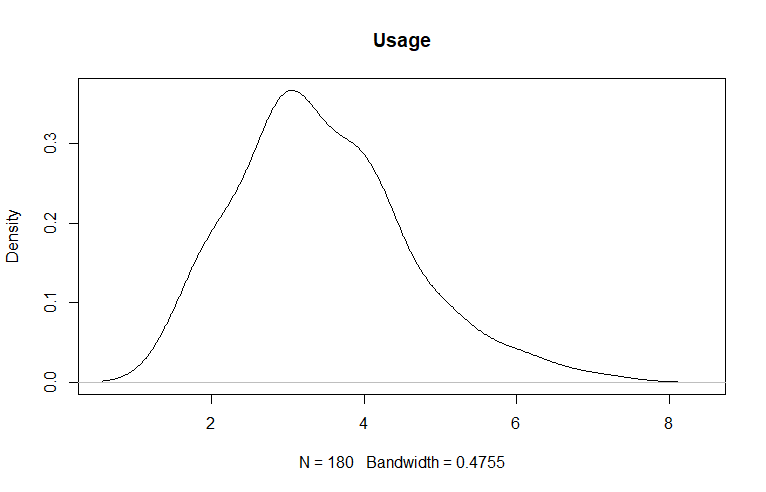
The difference between the youngest customer and the oldest customer is 32 years as indicated by the range. The youngest customer being 18 years old and the oldest being 50 years old. The interquartile range for the age of the customers is 9 years with Q1 being 24 years and Q3 being 33 years. This represents 50% of customers based on their ages arranged in ascending order. The standard deviation is about 7 years, which the age of a customer is on average 7 years away from the average age of all customers.

**Usage**

Customers intent to use the treadmills an average of 3.5 times a week. Most of them intent to use the treadmills 3 times a week. The customers intent to use the treadmills at least twice a week and those using the treadmills the most intending to use them 7 times a week. The median and the mode are the same at 3 times a week and the mean slightly higher at 3.5 times a week which means the distribution of usage is not symmetrical around the mean. Usage is right skewed as a result although the whiskers of the box plot are the same size. The existence of outliers which relates to customers intending to use the treadmills for 6 or 7 times a week increases the mean resulting in the right skew.

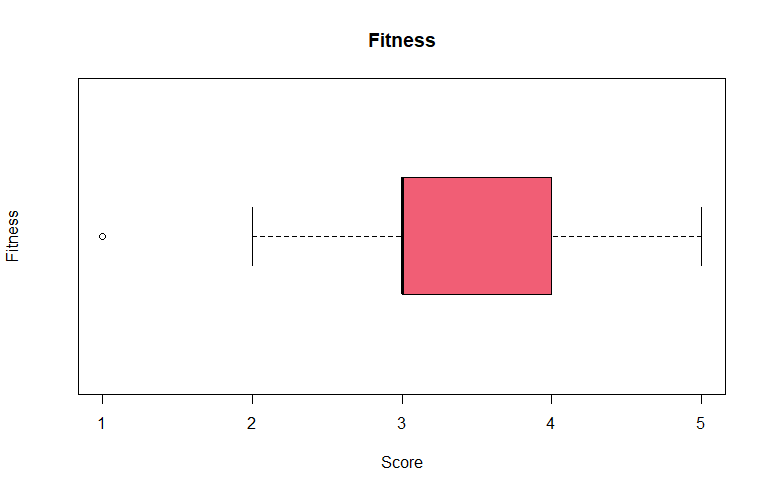
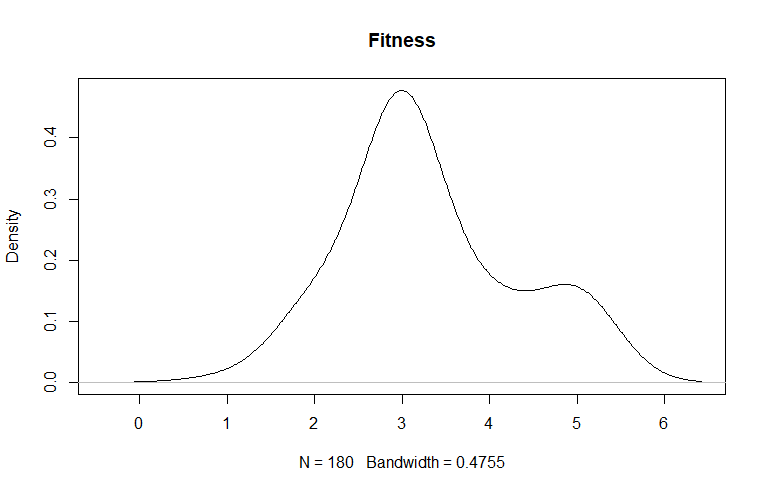
The standard deviation of usage is around 1 which indicates the usage is not widespread around the average usage of 3.5 times. The density plot also shows that the distribution is not very different from a normal distribution.

**Fitness**

Most customers believe that they fall somewhere in the middle of being very fit and very unfit. About 54% of customers believe they a 3 on a 1 to 5 rating of fitness. This means the average fitness at about 3.3 is very close to the median and mode which are at 3. The bar graph indicates that the distribution of fitness is not very far from a normal distribution. Skewness is hardly discernible from observing the bar graph and density plot alone. From the boxplot, the median is below the mean which indicates there is right skewness present. There is an outlier on the lower end of the fitness rating which is likely countering the right skewness which makes it not very apparent from observing the bar graph and density plot.

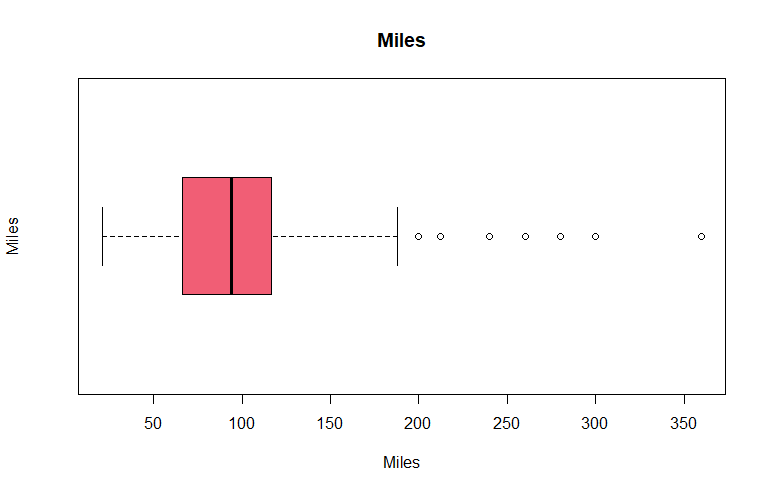
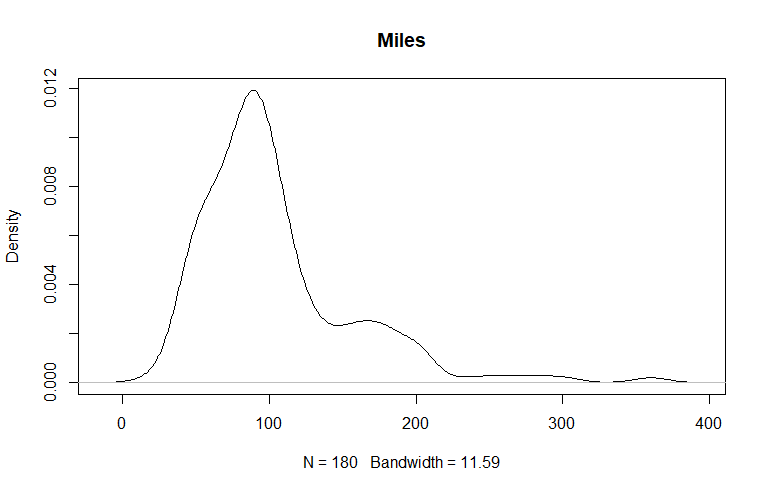
The distribution of fitness scores is mainly around the 3 score which means the variance and standard distribution are low at 0.919 and 0.959 respectively.

**Miles**

On average, customers believe they run 103 miles a week on the treadmills. Most customers run between 80 and 100 miles a week. The mean is higher than the median which is 94 miles, and this indicates the existence of a skewed distribution. The histogram confirms the large right skewness. The density plot confirms the large skewness. The miles customers run are widely spread with the minimum being 21 miles and the highest being 360 miles. The histogram also seems to indicate the presence of outliers on the upper side of miles. The boxplot confirms the skewness with the right whisker being longer than the left. There are also outliers which confirms what the histogram had indicated. The existence of these outliers some of them seemingly quite far from the upper quartile increase the mean which is why its larger than the median.

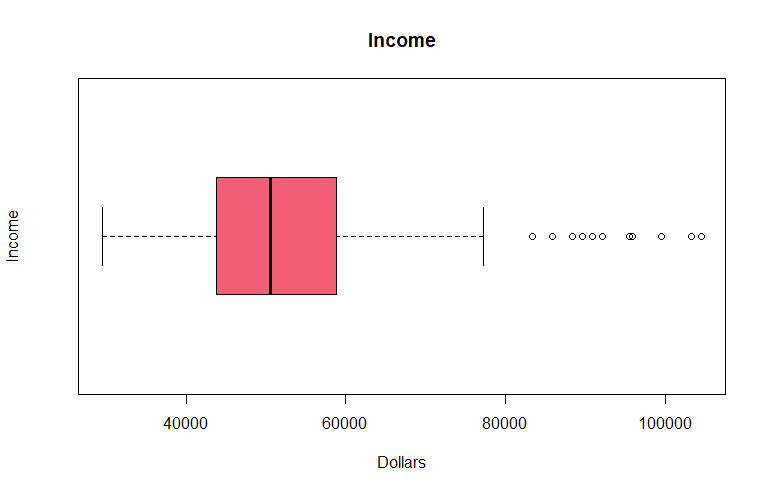
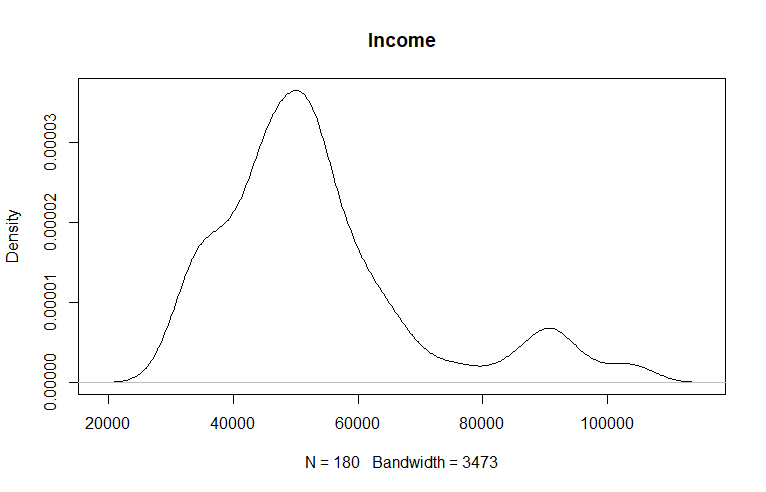
Since the distribution of miles is widespread, the variance and standard deviation also tend to be higher with the outliers increasing the variation.

**Income**

From the histogram, most of the customers fall in the $50 000 - $55 000 income bracket with the average income being $53 720. The median income is lower at $50 596 which indicates skewness. The histogram confirms the skewness which is large bit not as large as the skewness of miles.

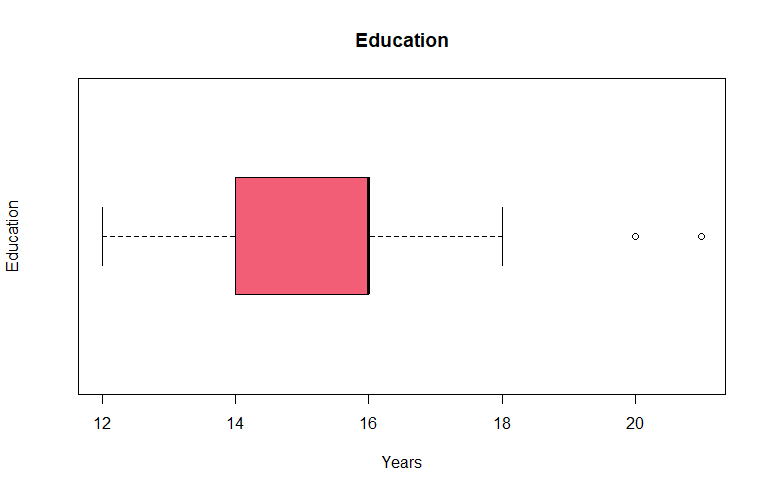
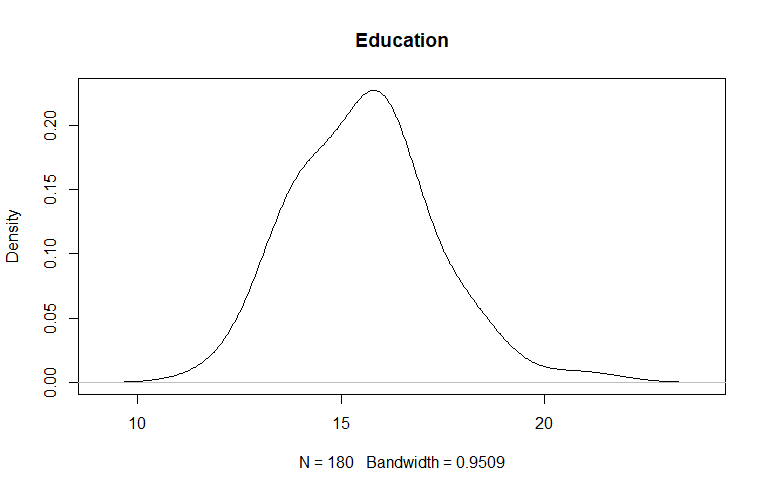
The difference in income between the lowest and highest income of the customers is $75 019 which could indicate a large variance and standard deviation. The distribution is also widespread on the upper side and does not resemble a normal or symmetrical distribution. The existence of outliers highlighted by the boxplot contributes to the right skew. The right whisker of the boxplot is larger than the left whisker also confirming the skew.

**Education**

Most customers have 14, 16 or 18 years of education with the majority having 16 year of education. There are none with 17 or 19 years of education with a few in total having more than 18 years of education. The average years of education is about 15.6 years and the median and mode are 16 years. The distribution of education tends to have a few peaks at 14, 16 and 18 years with few or none of the customers having odd numbered years of education such as 13, 15, 17 or 19 years. The existence of these multiple peaks makes it difficult to determine how the distribution is skewed without making an adjustment. After the adjustment, there is a positive skew that exists as shown in the density plot below. The boxplot whiskers are the same size but the existence of outlier above the upper quartile leads to the skewness.

Since the years of education are closely distributed around 16 years, the standard deviation and variance are not large. The coefficient of variation is the lowest amongst all the variables which means it’s not a widely dispersed compared to the other variables.

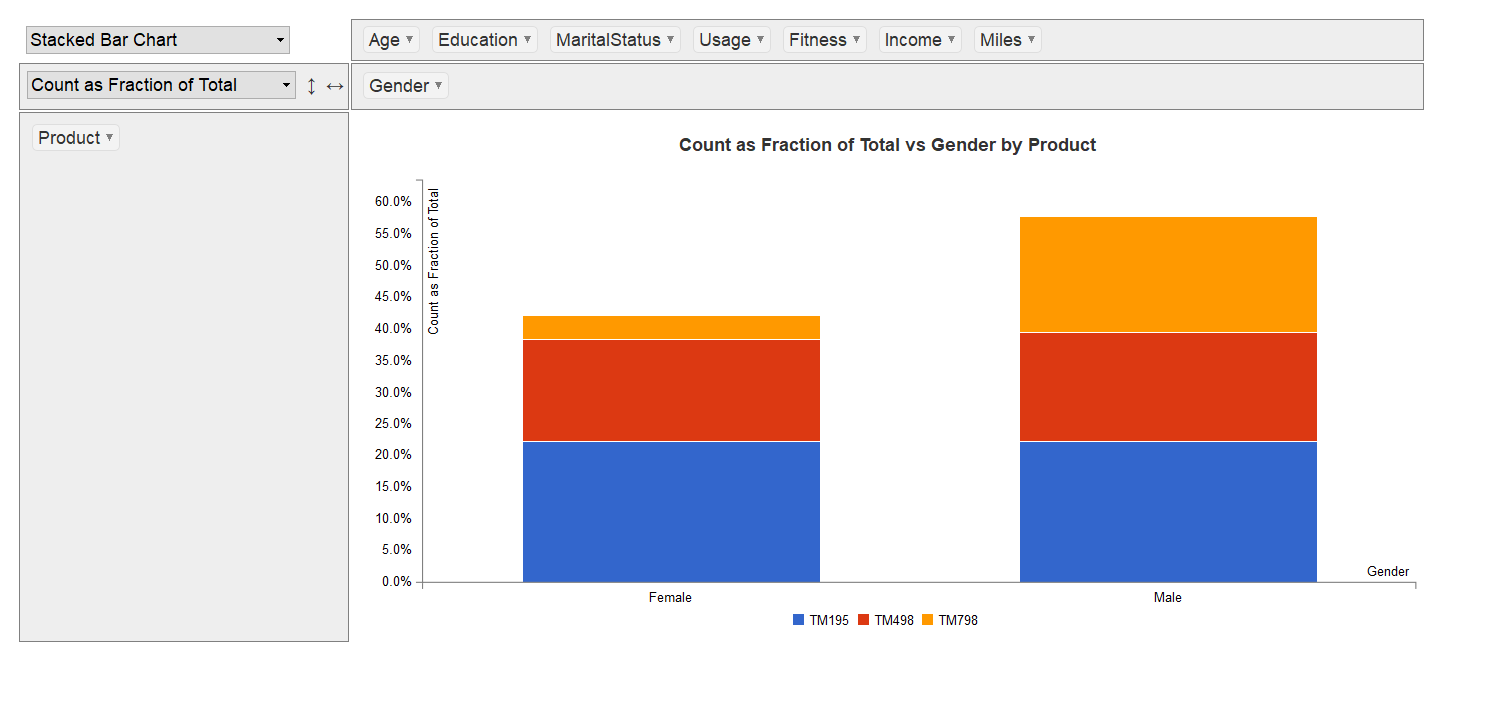
 

3.4 Bi-Variate Analysis

The first part of the bi-variate analysis looks at the focus of the project which is the characteristics of customers and how they relate to the different treadmill. Please refer to Appendix B for statistical summaries used in this section.

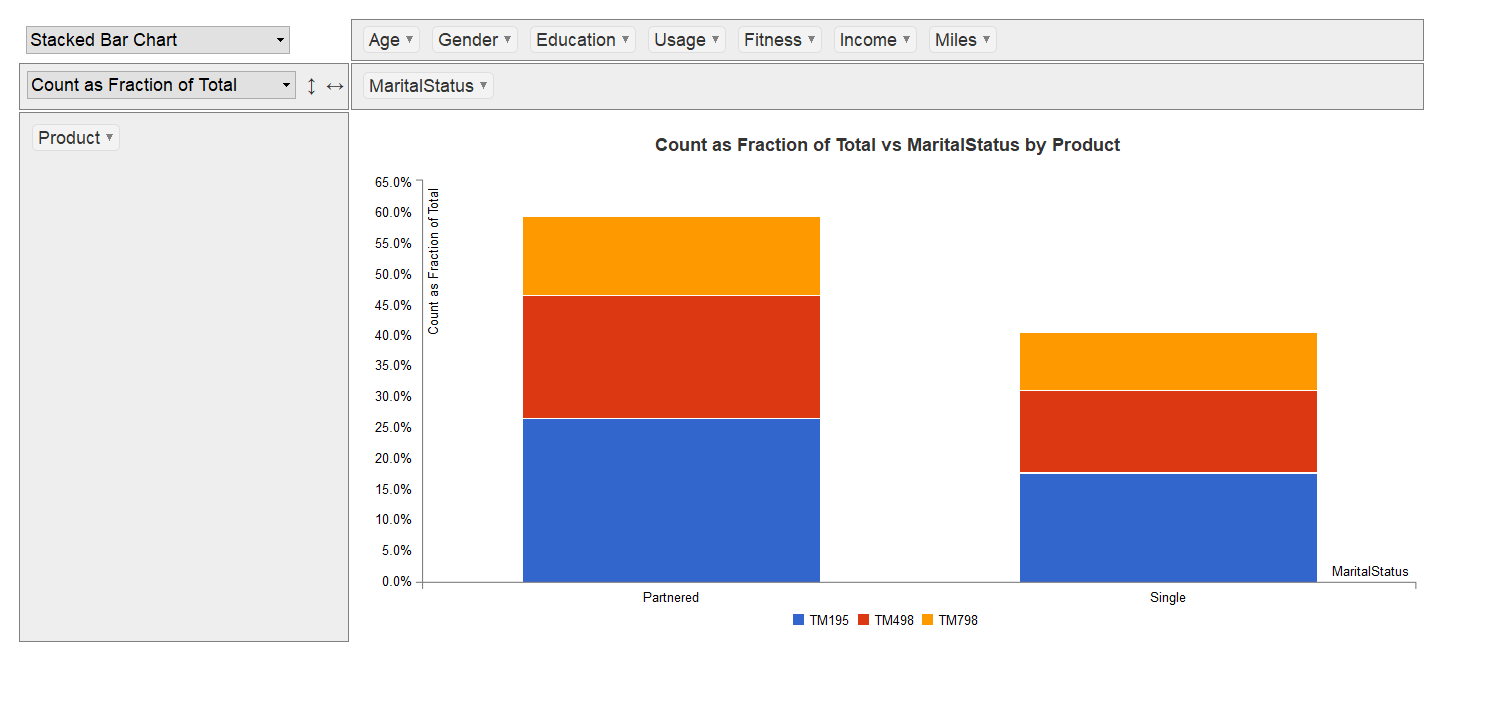
**Product vs Gender**

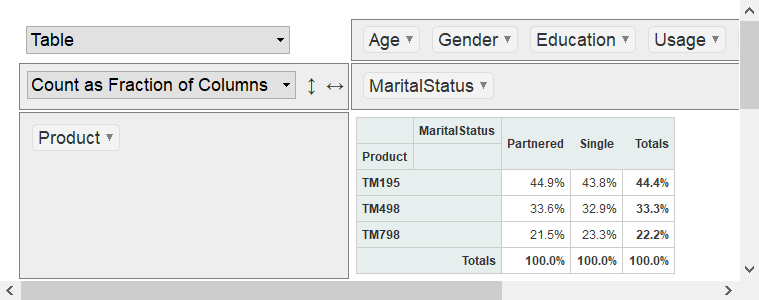
The number of customers using treadmill model TM195 across gender is the same at 40 customers. A similar pattern across gender exists for customer using the TM498 model with 29 female and 31 customers. There is distinction in the type of customers using the TM798 model across gender. Male customers dominate the purchase of this model compared to females with figures of 33 and 7 respectively.



**Product vs Marital Status**

Partnered customers purchase most of the treadmills across all 3 models compared to single customers. The proportion of each model of treadmills purchased relative to the total purchases of each group of customers is similar. For example, partnered customers for purchased the TM195 model make up 44.9% of purchases made by partnered customers. At the same time, single customers who purchased the same model make up 43.8% of all purchases made by single clients. The table below helps in showing this pattern across the two customer groups.



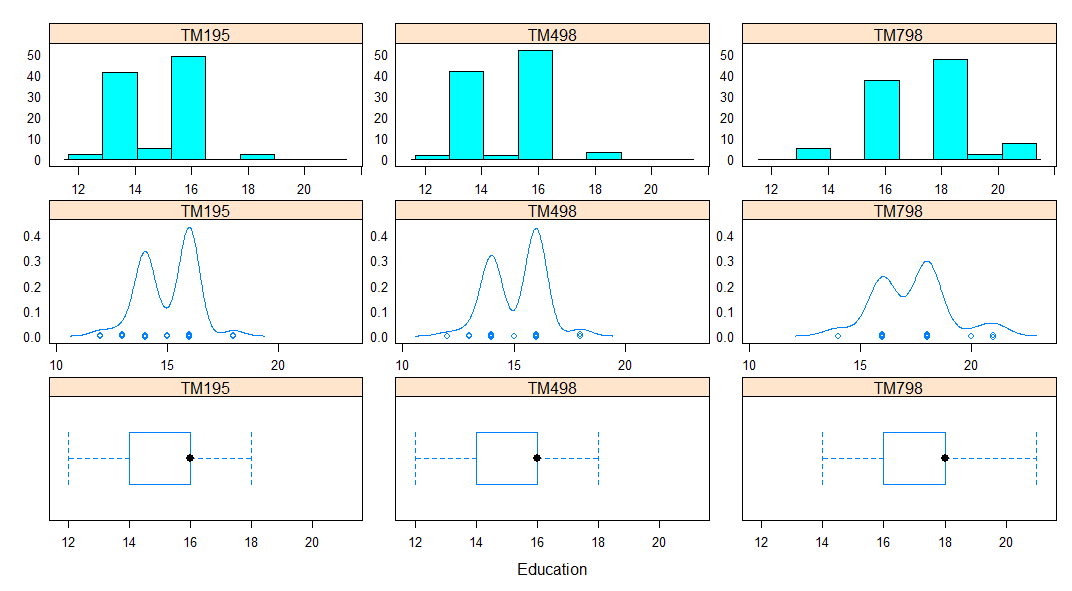


**Product vs Education**

The distribution of the number of years of education between customers of the TM195 and TM498 models is similar as shown by the histogram and density function below. The TM798 models show a different pattern relative to the other products. The degree of skewness is not very visible for the TM195 and TM498 models, but it appears similar. For the TM798 model, there is an indication of years of education being right skewed with the right whisker of the model being longer than the left. Across all three models, there are no outliers present.

The average years of education for customers of the first two models is about 15 years and for the TM798 model customers it about 17 years. The median ages for the first two models are 16 years and that for the TM798 customers is 18 years. These values are also similar to the mode for all three models. There are customers with more than 20 years of education for the TM798 model where the customers of the other two models have a maximum of 18 years of education. The years of education for the TM798 model customers are a bit more dispersed relative to customers of the other two models. The standard deviations of the TM195 and TM498 models is 1.216 and 1.223 respectively. This is less than the standard deviation of the TH798 model which is 1.639. The coefficient of variation also confirms the wider dispersion of years of education of the TM798 model relative to the other two models.

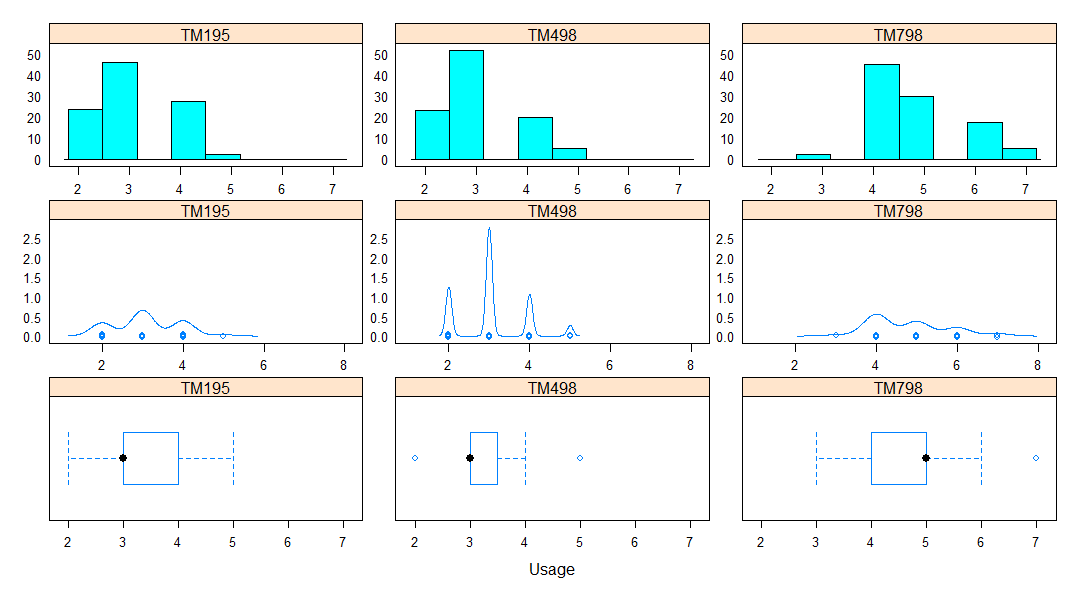
From looking at the education characteristics of customers of the three products, customers of the TM798 model tend to more educated relative to customers of the other two models.



**Product vs Usage**

The histograms of the TM195 and TM498 models show similar pattern of frequency distribution of usage with a the TM798 model showing a different pattern compared to the two models. The density plots seem to be similar for the TM195 and TM798 models. Skewness across all three models is positive or right skewed. The skewness indicating that most usage of the treadmills tends to be on the lower end for each of the model’s distribution of usage. The boxplots show the existence of outliers for the TM498 and TM798 models. The outliers of the TM498 model are on either side of the lower and upper quartiles. The boxplot of the TM498 model does not have a lower whisker indicating the lower quartile is the second value after the lowest value which also an outlier.

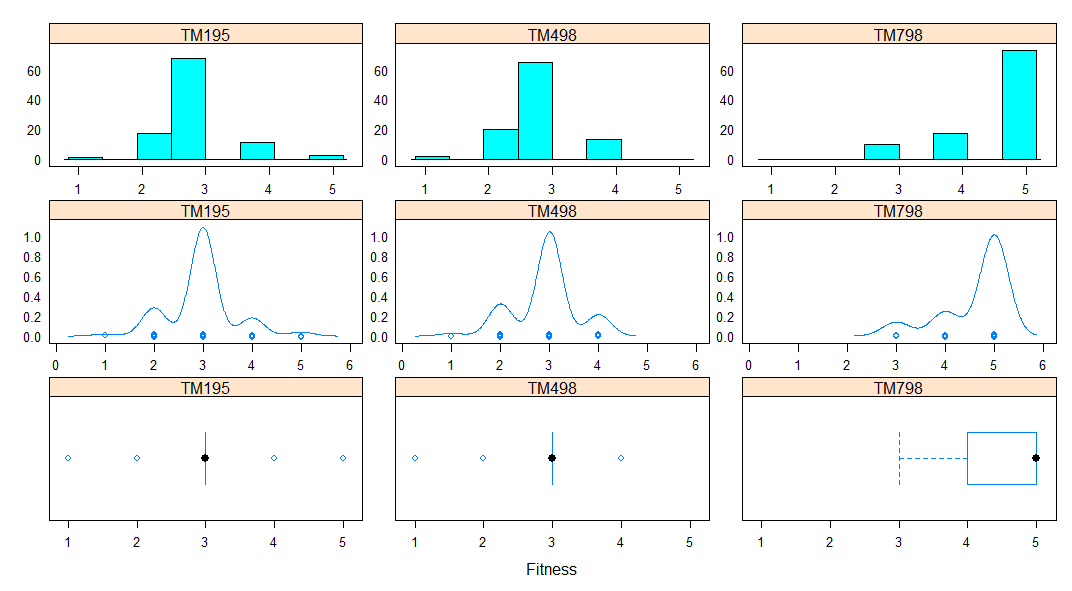
On average, customers of the TM195 and TM498 models intent on using their treadmills about 3 times a week and those of the TM798 model almost 5 times a week. The values for the median are almost like the mean values of the models. The standard deviation of the TM798 model is larger than the other models. However, its coefficient of variation is the least amongst all the models meaning its dispersion of usage is not as wide as suggested by standard deviation alone.



**Product vs Fitness**

The fitness level histograms for the TM195 and TM498 models are similar and are closer to a normal distribution because of the symmetry they show. The TM798 model does not show a symmetrical distribution with a pronounced negative of left skew. The density plot also clearly reveals this skew. The TM195 and TM498 models show boxplots without the box and whiskers. This means the most of the customers reported their fitness levels around the same value for these models. They also have outliers that are on either side of the boxplot contributing to the symmetry. The TM798 model shows customers that rated themselves highly on fitness with the upper quartile equal to the maximum fitness value and no upper whisker.

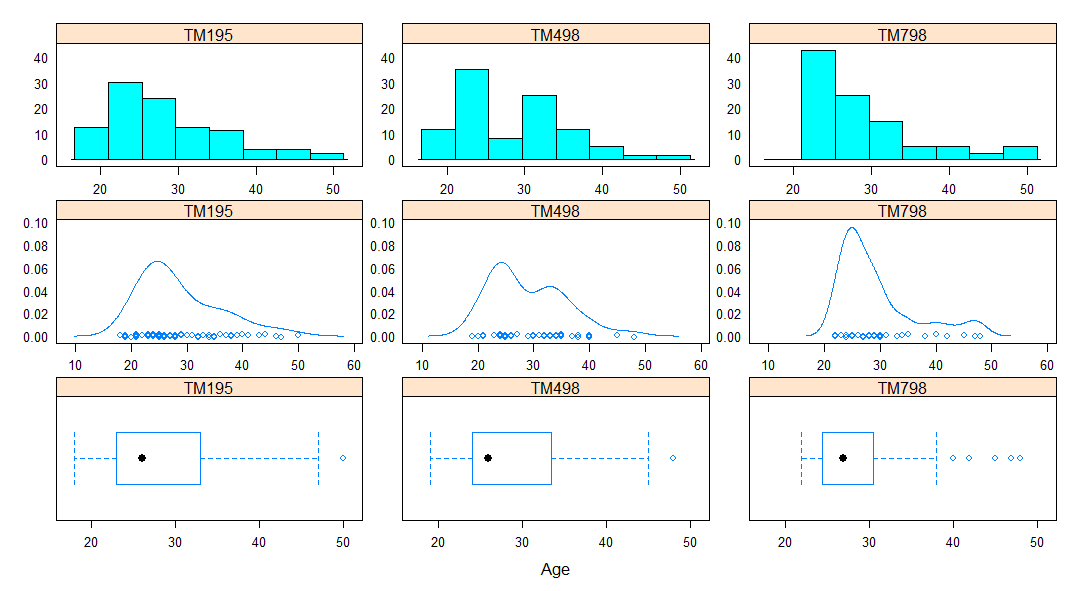
On average customers for the first two models rated themselves 3 on a five-point fitness scale with 1 being very unfit and 5 being very fit. Most customers for the TM798 model rated themselves as very fit with the lowest rating for this model being a 3. The standard deviation across all three products is almost similar. The coefficient of variation for the TM798 model is the lowest amongst all three models showing fitness for the customers is not as widely dispersed as the other two models.



**Product vs Age**

The density plots show that age of the customers across all three models is mainly on the younger side relative to all treadmill customers. Therefore, all models show right skewness. The right whiskers of all three boxplots are longer than the left whiskers. There are outliers existing across all three products though the TM798 models has a larger number of outliers compared to the other two. The existence of these outliers distorts the shape of the distribution which results an asymmetric distribution.

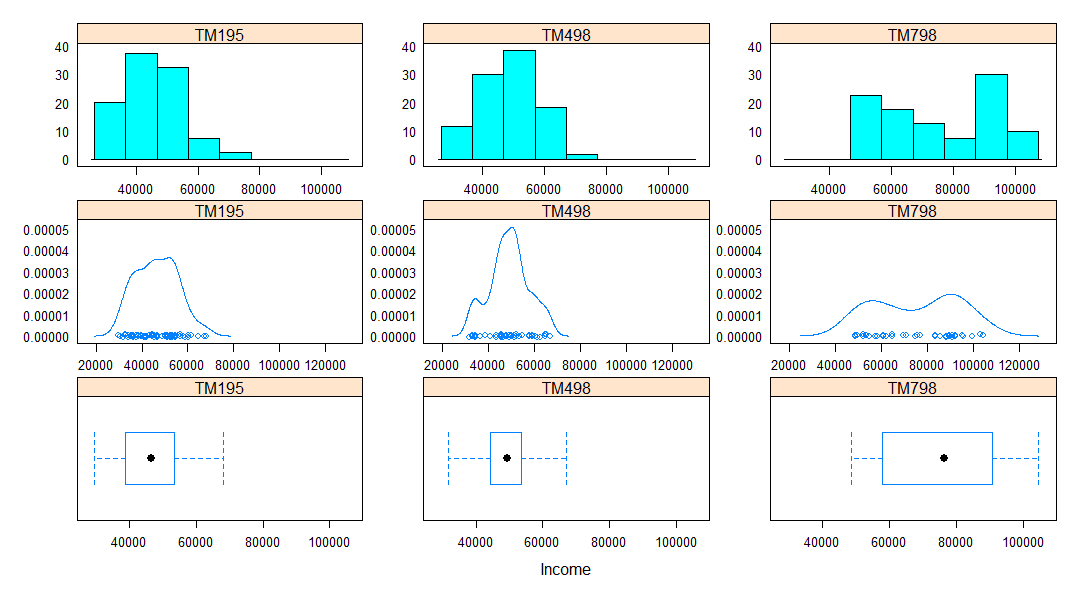
The average age of customers across all three products is around 29 years. Therefore, knowing the age of the customer alone may not give insight into the model a customer may purchase when considering the mean age alone. The median age for the TM195 and TM498 models is 26 years and that of the TM798 is 27 years. Customers of the TM195 model have a wider spread of their age relative to their average compared to the other two models. The coefficient of variation also confirms this wider dispersion.



**Product vs Income**

The distribution of income for customers purchasing the TM195 and TM498 models is closer to a normal distribution compared to those of the TM798 model. The histogram shows that customers of the TM798 model earn more income compared to those of the other two models. Across all three models, there are no customers that are outliers. The width of the boxplot for the TM798 model indicates a wider spread of income range for these customers. Observing the graphs alone does not seem to indicate the degree of skewness.

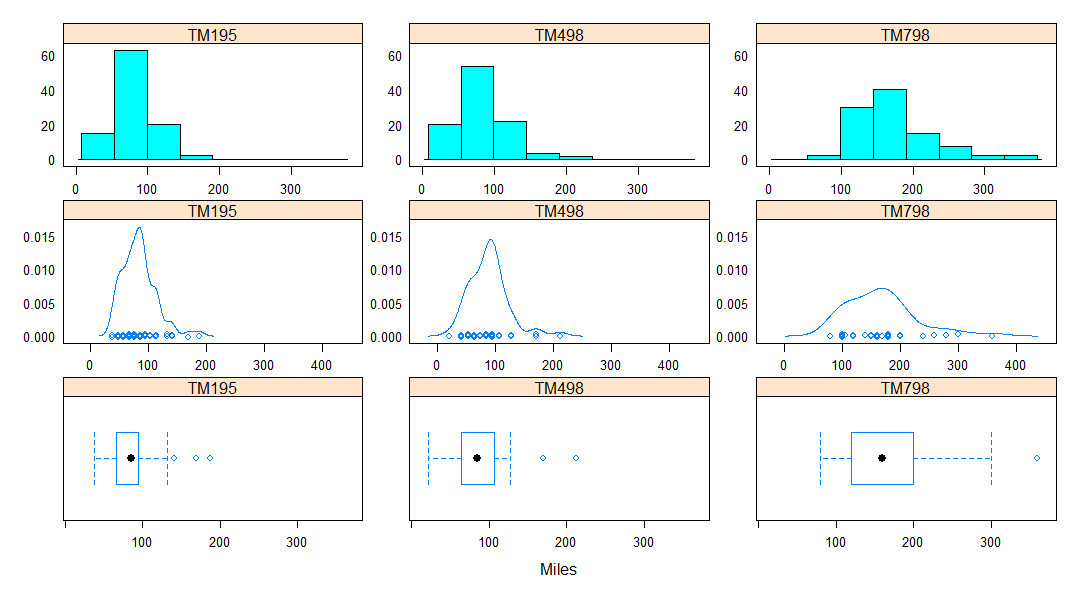
The average incomes across all the three models are all different with the TM195 average at $46 418, the TM498 at $48 974 and the highest average for the TM798 model at $75 567. There is a larger spread between the minimum and maximum incomes as indicated by the range for customers of the TM798 model. The standard deviation for this model is also the highest across all three models. The coefficient of variation also confirms this wide dispersion compared to the other two models.



**Product vs Miles**

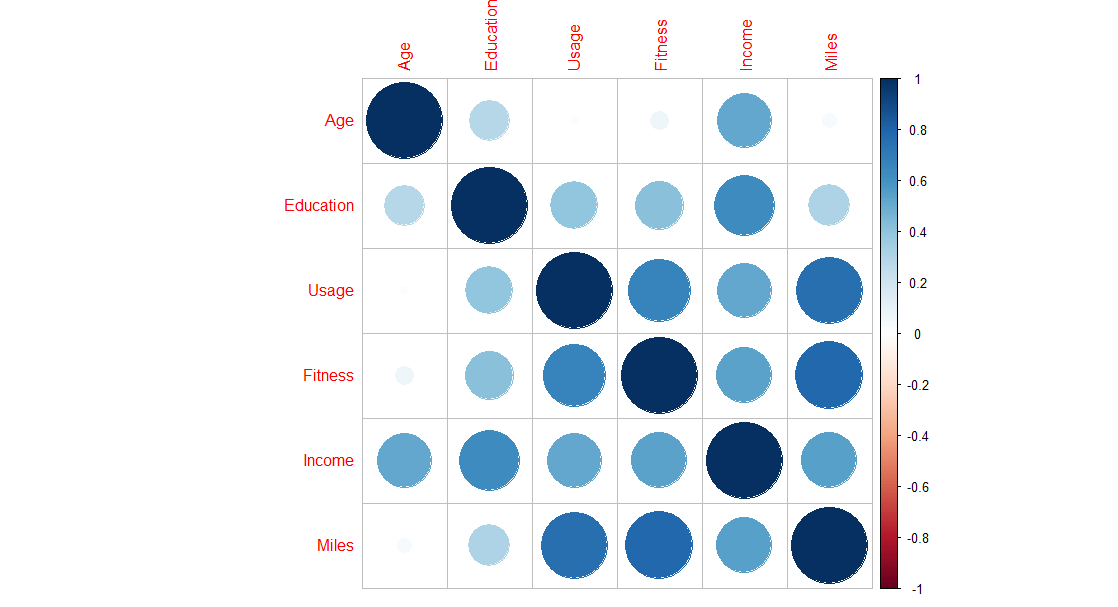
There is similar distribution of miles on the treadmill for customers of the TM195 and TM498 models. This is different from that of the TM798 model. All three models show a degree of positive skewness meaning customers of each model are bunched on the lower end of miles on the treadmill for each respective model. There is an existence of outliers across all three models which distorts the symmetry of the distributions.

The average number of miles on the treadmill for customers of the TM798 model is significantly higher than that of the other two models. It is 166.9 miles compared to the TM195 model at 82.79 miles and the TM498 at 87.93 miles. The other factor that could potentially influence this are the usage and fitness levels which were higher for the TM798 model compared to the other two models. The standard deviation and range of miles for the TM798 model is the largest suggestion a wider spread of miles that customers spend on the machine. However, considering the coefficient of variation, its is not as widely dispersed as these measures would suggest. The coefficient of variation is in the middle relative to the other two models.



**Correlation between numeric variables**

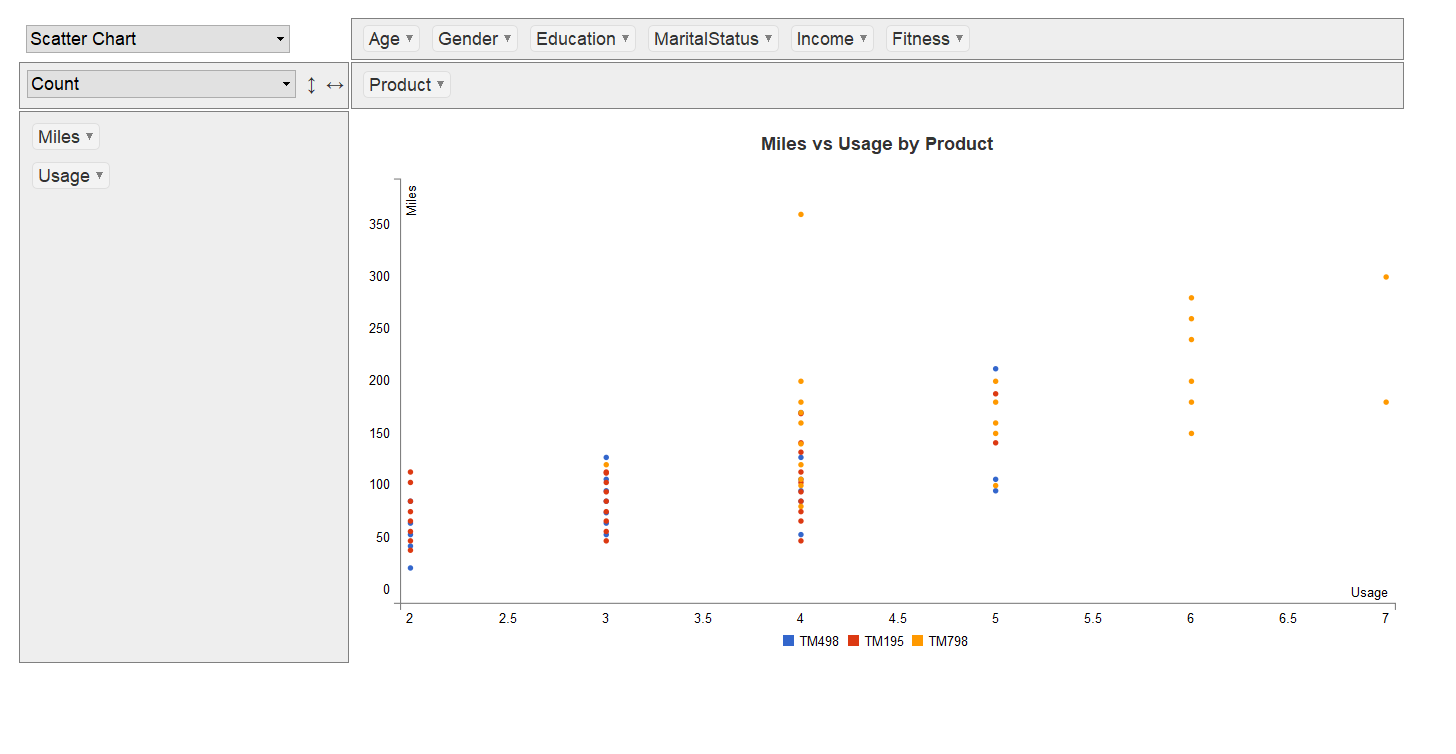
The correlation between the other variables is valuable in identifying the characteristics that tend to go together when describing the customers of each of the models. For example, one can expect to see a high correlation between usage and miles which can be used to identify characteristics of customers of a given model.



All numeric variable have a positive correlation to each. There is strong correlation between Usage and Miles, Miles and Fitness, Usage and Fitness as well as Income and Education. These strong correlations can provide an additional level of depth in characterising customers for these treadmills.

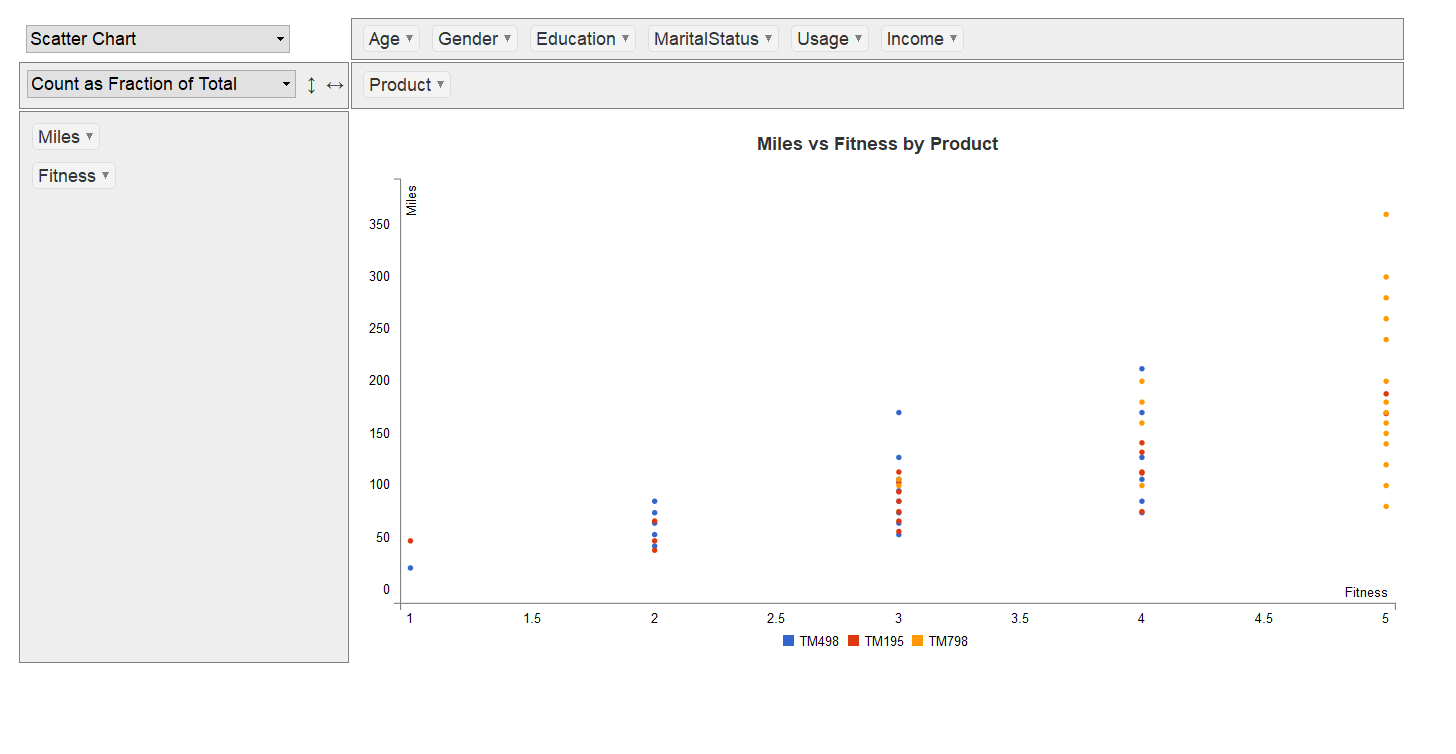
**Miles vs Usage by Product**

Customers that tend to have higher usage and run the most miles on the treadmills are mostly using the TM798 model. This model is not being used by customers on the lower end of both miles and usage.



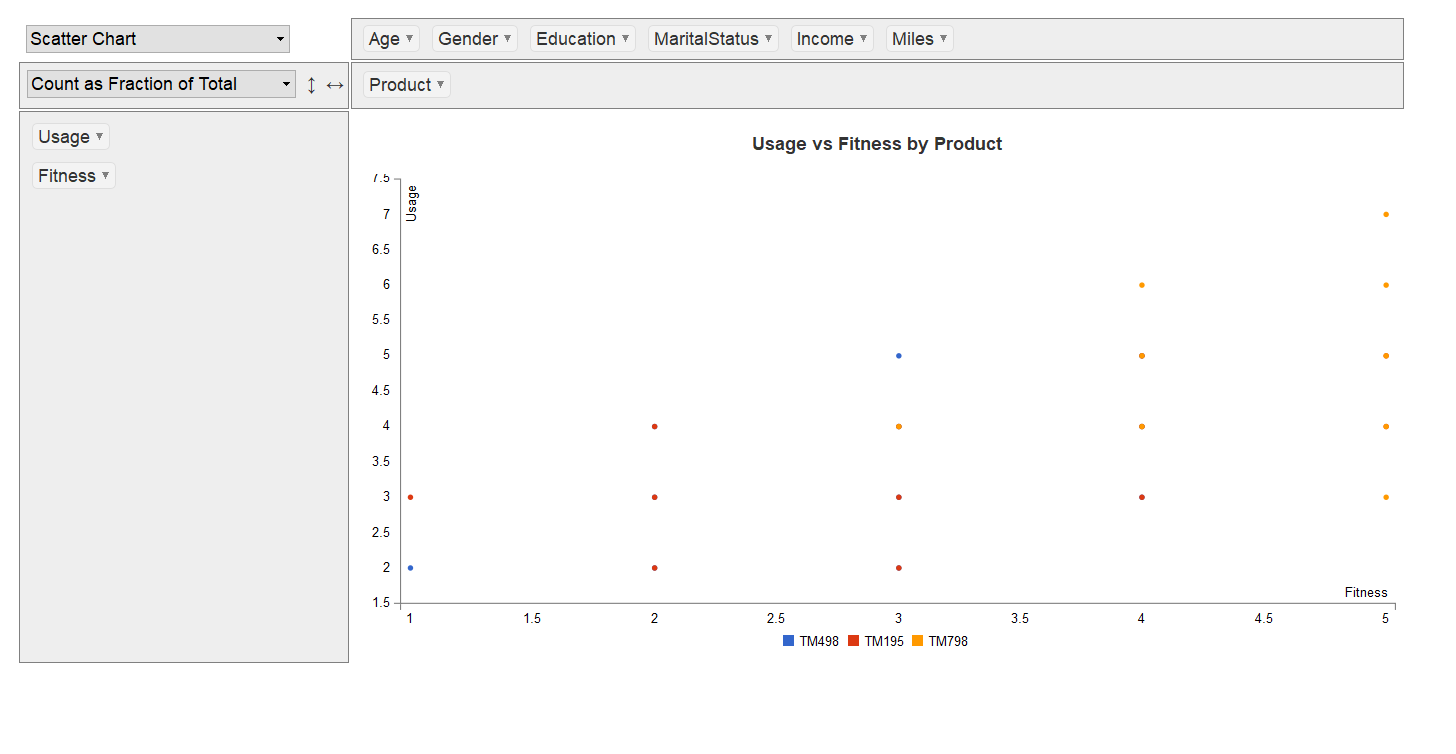
**Miles vs Fitness by Product**

A similar pattern to the above also exists for customers with high fitness levels running more miles on treadmills. These customers are using the TM798 model and none of the customers are with low fitness and running the least miles are using this model.



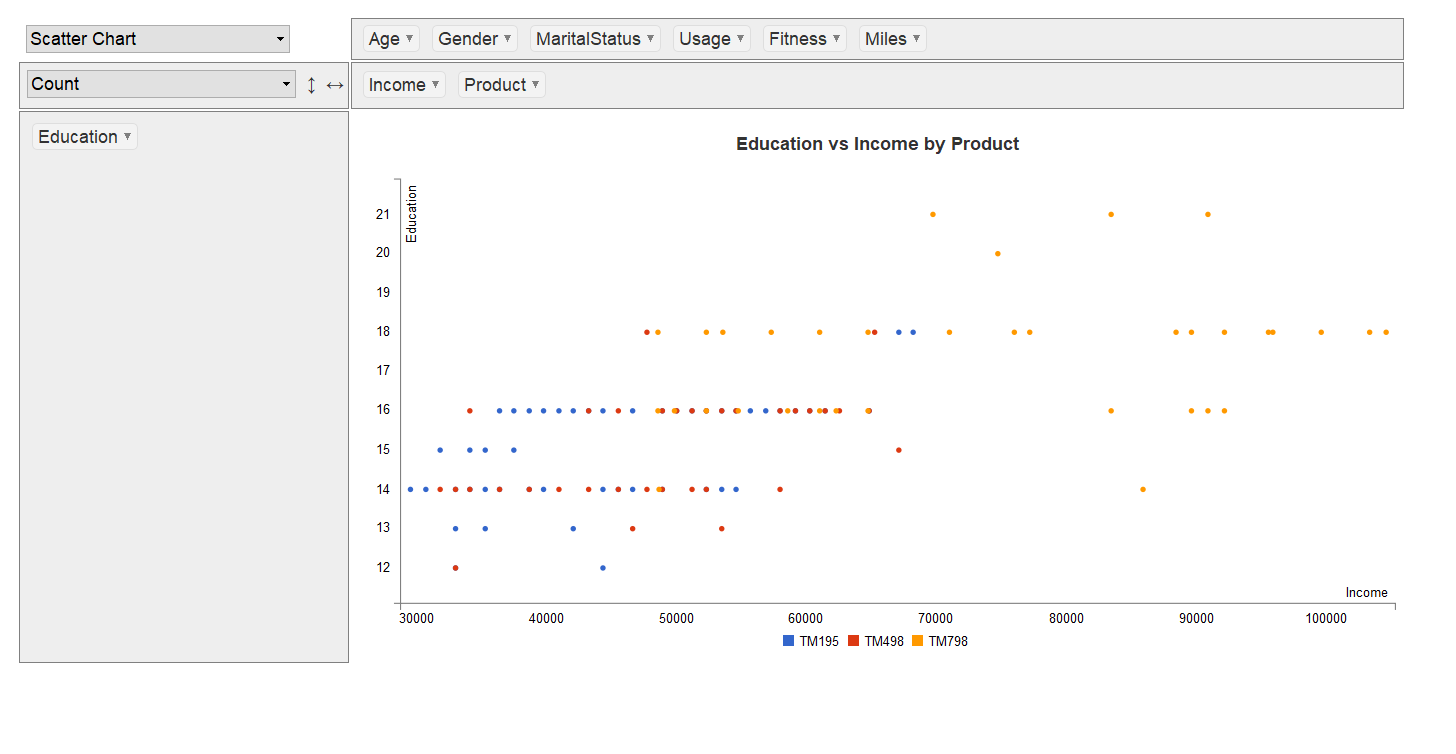
**Usage vs Fitness by Product**

Also, the high usage and high fitness level customers are using the TM798 model and none of the low usage and low fitness level customers are using this model.



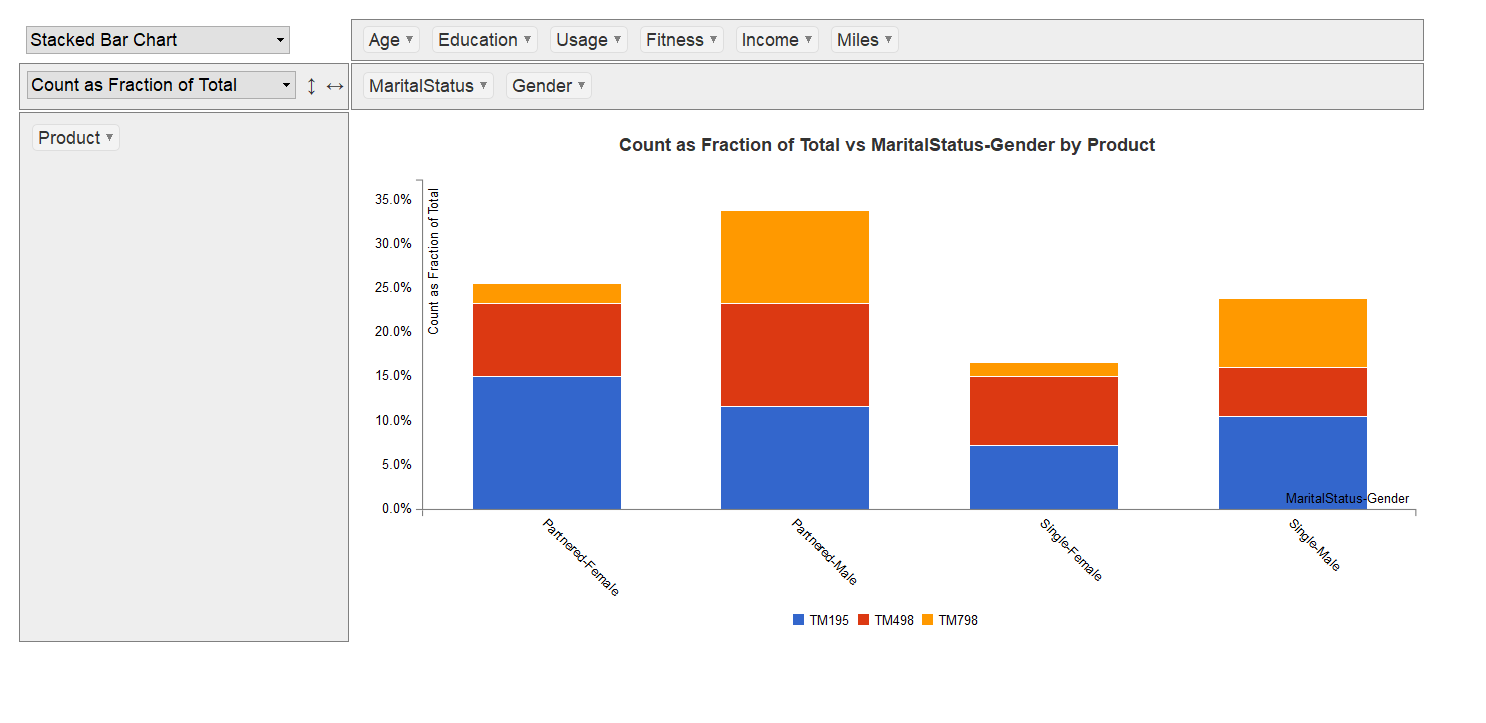
**Education vs Income by Product**

Customers with high incomes and higher levels of education are mainly using the TM798 model. Most of the customers of the other two models are mainly using the other two models.



**Categorical variables – Marital Status vs Gender by Product**

This view show that most of the customers for the TM195 model are female with partners. Most of the customers for the TM498 are male with partners and those of the TM798 model are also male with partners. The largest customer group is males with partners and the smallest customer group is single females across all products.



3.5 Missing Value Identification

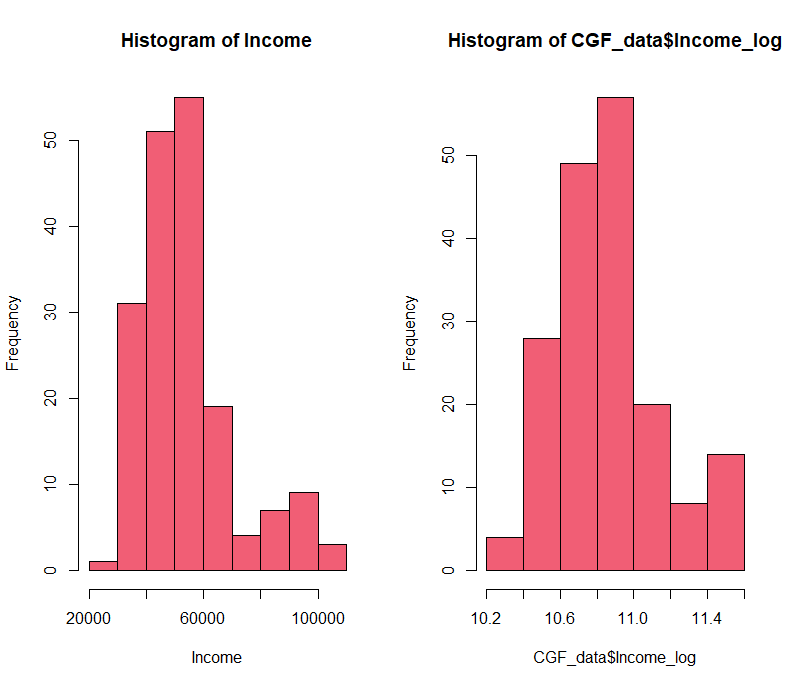
There were no missing values identified in the dataset.

3.6 Outlier Identification

There were outliers identified for some variables in the dataset. No further treatment was taken as they were all considered genuine observations.

3.7 Variable Transformation / Feature Creation

There is need to transform skewed variables to assist in the reduction of the skewness for example income. The graph below shows before and after log transformation. Skewness before transformation was



I would consider creating a new variable which is miles/usage to have an idea of much customers are running for any given use. This would assist in providing customers with information of the product that suit their needs depending on how much they use it.

4 Conclusion

Most of Cardio Good Fitness’ customers are male for all models with a fairly equal distribution of both male and female customers for models TM195 and TM498. A distinctly larger proportion of customers using model TM798 are male.

Most of the customers are partnered with the least number single. The proportion of customers for each model across marital status i.e. partnered vs single is similar. From these observations, partnered male customers make up most of the customers and single female customers the least.

There is not much difference between the education levels of customers of the TM195 and TM498 models. The TM798 model customers tend to have a distinct higher number of years compared to customers of the other two models. There is also a robust positive relationship between the number of education years and income. Customers purchasing the TM798 model in addition to their higher levels of education also tend to have higher incomes.

Customers of all three models have an average age of about 29 years. The age profile of customers across all three models is relatively similar. Not much conclusion or differentiation of customers of each model can be made by looking at their ages alone.

There is strong positive relationship between, Usage, Fitness and Miles. There is a clear distinction between customers purchasing the TM798 model being more fit, with high usage and running more miles compared to customers of the other product.

Finally, looking at all the characteristics, it easier to differentiate customers using the TM798 model as most of their characteristics show a clear distinction compared to customers of the other two models. It is not easy to distinguish customers of these two models as most of the characteristics are similar. In a nutshell, the TM798 model customers can characterised as well educated, earning a higher income and very fit.

5 Appendix A – Source Code

*######------Exploratory Data Analysis------######*

*######------Cardio Good Fitness Project-----#####*

*##Environment Set up and data set import*

*#Required libraries*

*library(readr)*

*library(dplyr)*

*library(corrplot)*

*library(data.table)*

*library(ggplot2)*

*library(cowplot)*

*library(pastecs)*

*library(lattice)*

*library(rpivotTable)*

*library(formattable)*

*library(moments)*

*library(latticeExtra) #Combines laa=ttice graphs into one*

*#Set working directory*

*setwd("C:/Users/kudaakwashe/Documents/Study/PGPDSBA/Introduction\_To\_Analytics/Project\_1")*

*getwd() #check if directory is correctly set up*

*#Import data set*

*CGF\_data = read.csv("CardioGoodFitness.csv")*

*attach(CGF\_data)*

*#get a picture of the data set*

*dim(CGF\_data)*

*head(CGF\_data)*

*tail(CGF\_data)*

*str(CGF\_data)*

*summary(CGF\_data)*

*#checking for missing values*

*anyNA(CGF\_data)*

*sum(is.na(CGF\_data))*

*##Univariate analysis*

*#factor variable analysis*

*table(Product)*

*table(Gender)*

*table(MaritalStatus)*

*prop.table(table(Product))*

*prop.table(table(Gender))*

*prop.table(table(Fitness))*

*table(Usage)*

*#Summary of statistical measures*

*options(digits = 2)*

*options(scipen = 999)*

*ProductB = (ggplot(CGF\_data,aes(Product))*

*+ geom\_bar(aes(fill=Product))*

*+ geom\_text(aes(label=..count..),stat="count",position=position\_stack(0.5)))*

*GenderB = (ggplot(CGF\_data,aes(Gender))*

*+ geom\_bar(aes(fill=Gender))*

*+ geom\_text(aes(label=..count..),stat="count",position=position\_stack(0.5)))*

*MaritalStatusB = (ggplot(CGF\_data,aes(MaritalStatus))*

*+ geom\_bar(aes(fill=MaritalStatus))*

*+ geom\_text(aes(label=..count..),stat="count",position=position\_stack(0.5)))*

*#Plot categorical graphs onto same area*

*plot\_grid(ProductB,GenderB,MaritalStatusB,ncol = 1)*

*#Integer variable graphs*

*EducationB = (ggplot(CGF\_data,aes(x=Education))*

*+ geom\_bar(aes(y=..count..),color="black", fill="#A1C1FF"))*

*UsageB = (ggplot(CGF\_data,aes(x=Usage))*

*+ geom\_bar(aes(y=..count..),color="black", fill="#A1C1FF"))*

*FitnessB = (ggplot(CGF\_data,aes(x=Fitness))*

*+ geom\_bar(aes(y=..count..),color="black", fill="#A1C1FF"))*

*AgeH = (ggplot(CGF\_data,aes(x=Age))*

*+ geom\_histogram(aes(y=..count..), breaks=seq(15,50, by=5),color="black", fill="#A1C1FF"))*

*IncomeH = (ggplot(CGF\_data,aes(x=Income))*

*+ geom\_histogram(aes(y=..count..),breaks=seq(25000,105000, by=5000),color="black", fill="#A1C1FF"))*

*MilesH = (ggplot(CGF\_data,aes(x=Miles))*

*+ geom\_histogram(aes(y=..count..),breaks=seq(20,360, by=10),color="black", fill="#A1C1FF"))*

*#Plot graphs onto same area*

*plot\_grid(EducationB,UsageB,FitnessB,AgeH,IncomeH,MilesH,ncol = 3)*

*#Box and Density Plots*

*AgeBx = (boxplot(Age, main = "Age", xlab = "Years", ylab = "Age",*

*col = "#F15E75", horizontal = TRUE))*

*AgeD = plot(density(Age), main = "Age")*

*UsageBx = (boxplot(Usage, main = "Usage", xlab = "Times", ylab = "Usage",*

*col = "#F15E75", horizontal = TRUE))*

*UsageD = plot(density(Usage, adjust = 2), main = "Usage")*

*FitnessBx = (boxplot(Fitness, main = "Fitness", xlab = "Score", ylab = "Fitness",*

*col = "#F15E75", horizontal = TRUE))*

*FitnessD = plot(density(Fitness, adjust = 2), main = "Fitness")*

*MilesBx = (boxplot(Miles, main = "Miles", xlab = "Miles", ylab = "Miles",*

*col = "#F15E75", horizontal = TRUE))*

*MilesD = plot(density(Miles), main = "Miles")*

*IncomeBx = (boxplot(Income, main = "Income", xlab = "Dollars", ylab = "Income",*

*col = "#F15E75", horizontal = TRUE))*

*IncomeD = plot(density(Income), main = "Income")*

*EducationBx = (boxplot(Education, main = "Education", xlab = "Years", ylab = "Education",*

*col = "#F15E75", horizontal = TRUE))*

*EducationD = plot(density(Education, adjust = 2), main = "Education")*

*#Pivoting data and generating graphs*

*rpivotTable(CGF\_data)*

*#Grouping Numeric variables*

*CGF\_Numeric = CGF\_data %>% select\_if(is.numeric)*

*#Generating Statistical summaries*

*stat.desc(CGF\_Numeric, basic =TRUE)*

*formattable(stat.desc(CGF\_Numeric, basic =TRUE))*

*#Group by Product for Numeric Variables*

*P\_TM195 = (CGF\_data[which(Product=="TM195"),])%>%select\_if(is.numeric)*

*P\_TM498 = (CGF\_data[which(Product=="TM498"),])%>%select\_if(is.numeric)*

*P\_TM798 = (CGF\_data[which(Product=="TM798"),])%>%select\_if(is.numeric)*

*Stats\_TM195 = (setnames(P\_TM195, old = c("Age", "Education", "Usage", "Fitness", "Income", "Miles"),*

*new = c("Age\_TM195", "Education\_TM195", "Usage\_TM195", "Fitness\_TM195", "Income\_TM195", "Miles\_TM195")))*

*formattable(stat.desc(Stats\_TM195))*

*Stats\_TM498 = (setnames(P\_TM498, old = c("Age", "Education", "Usage", "Fitness", "Income", "Miles"),*

*new = c("Age\_TM498", "Education\_TM498", "Usage\_TM498", "Fitness\_TM498", "Income\_TM498", "Miles\_TM498")))*

*formattable(stat.desc(Stats\_TM498))*

*Stats\_TM798 = (setnames(P\_TM798, old = c("Age", "Education", "Usage", "Fitness", "Income", "Miles"),*

*new = c("Age\_TM798", "Education\_TM798", "Usage\_TM798", "Fitness\_TM798", "Income\_TM798", "Miles\_TM798")))*

*formattable(stat.desc(Stats\_TM798))*

*#Determining skewness*

*skewness(P\_TM195$Miles\_TM195)*

*skewness(P\_TM498$Miles\_TM498)*

*skewness(P\_TM798$Miles\_TM798)*

*###BIVARIATE ANALYSIS*

*#Bivariate analysis of categorical variables*

*table(Product,Gender)*

*prop.table(table(Product,Gender))*

*table(Product,MaritalStatus)*

*prop.table(table(Product,MaritalStatus),2)*

*by(CGF\_data, INDICES = Product, FUN = summary)*

*#Generating graphs for Product vs other variable analysis*

*Pro\_Ed\_H = histogram(~Education | factor(Product), data = CGF\_data)*

*Pro\_Ed\_B = bwplot(~Education | factor(Product), data = CGF\_data)*

*Pro\_Ed\_D = densityplot(~Education | factor(Product), data = CGF\_data)*

*Pro\_Us\_H = histogram(~Usage | factor(Product), data = CGF\_data)*

*Pro\_Us\_B = bwplot(~Usage | factor(Product), data = CGF\_data)*

*Pro\_Us\_D = densityplot(~Usage | factor(Product), data = CGF\_data)*

*Pro\_Fi\_H = histogram(~Fitness | factor(Product), data = CGF\_data)*

*Pro\_Fi\_B = bwplot(~Fitness | factor(Product), data = CGF\_data)*

*Pro\_Fi\_D = densityplot(~Fitness | factor(Product), data = CGF\_data)*

*Pro\_Ag\_H = histogram(~Age | factor(Product), data = CGF\_data)*

*Pro\_Ag\_B = bwplot(~Age | factor(Product), data = CGF\_data)*

*Pro\_Ag\_D = densityplot(~Age | factor(Product), data = CGF\_data)*

*Pro\_In\_H = histogram(~Income | factor(Product), data = CGF\_data)*

*Pro\_In\_B = bwplot(~Income | factor(Product), data = CGF\_data)*

*Pro\_In\_D = densityplot(~Income | factor(Product), data = CGF\_data)*

*Pro\_Mi\_H = histogram(~Miles | factor(Product), data = CGF\_data)*

*Pro\_Mi\_B = bwplot(~Miles | factor(Product), data = CGF\_data)*

*Pro\_Mi\_D = densityplot(~Miles | factor(Product), data = CGF\_data)*

*#Grouping the different graphs together*

*c(Pro\_Ed\_B,Pro\_Ed\_D,Pro\_Ed\_H,layout=c(3,3))*

*c(Pro\_Us\_B,Pro\_Us\_D,Pro\_Us\_H,layout=c(3,3))*

*c(Pro\_Fi\_B,Pro\_Fi\_D,Pro\_Fi\_H,layout=c(3,3))*

*c(Pro\_Ag\_B,Pro\_Ag\_D,Pro\_Ag\_H,layout=c(3,3))*

*c(Pro\_In\_B,Pro\_In\_D,Pro\_In\_H,layout=c(3,3))*

*c(Pro\_Mi\_B,Pro\_Mi\_D,Pro\_Mi\_H,layout=c(3,3))*

*#Bivariate Anallysis of numeric variables*

*corrplot(cor(CGF\_Numeric))*

*#Scatterplot Matrix*

*splom(CGF\_Numeric)*

*CGF\_cor = round(cor(CGF\_Numeric),2)*

*CGF\_cor*

*#Income log transformation*

*CGF\_data$Income\_log = log(Income)*

*#comparing skewness change as a result of transformation*

*par(mfrow=c(1,2))*

*hist(Income, col="#F15E75")*

*hist(CGF\_data$Income\_log, col="#F15E75")*

*par(mfrow=c(1,1))*

*#Adding a new feature*

*CGF\_data$Miles\_Per\_Usage = (Miles/Usage)*

*CGF\_data*

*detach(CGF\_data)*

**ANNEXURE B**

