INFX 502 COURSE PROJECT: DATA ANALYSIS OF CARDIO GOOD FITNESS PRODUCTS

Ipek Kaya

ULID: C00528687

COURSE: Informatics

Table of Contents

I. Dataset

1. Dataset Description

This project covers a visual analysis of the Cardio Good Fitness dataset which is collected in the CardioGoodFitness.csv file. The Cardio Good Fitness dataset is created by the market research team at AdRirght. The team investigated the 3 different treadmill products offered by Cardio Good Fitness based on customer characteristics. The dataset includes nine different variables.

Numerical variables;

- Age; in years
- Education; in years
- Income; Annual household income ($)
- Usage; the average number of times the customer plans to use the treadmill each week
- Miles; the average number of miles the customer expects to walk/run each week

Categorical variables;

- The first categorical variables are Products which is separated into three groups; TM195, TM498, or TM798.
- The second categorical variable is Gender which is separated into two groups; female and male.
- The third categorical variable is MaritalStatus which is separated into two groups; single and partnered.
- The fourth categorical variable is Fitness which is separated into five groups; categorical variable is marital status which is separated into two groups; 1 is poor shape and 5 is excellent shape.
- The numerical variables in this dataset are age, education, income, usage, and miles.

2. Dataset Loading

Before visualizing the dataset, dataset is saved as a ‘cvc’ filed and it is loaded into R from Microsoft Excel CVS file by using the commend ‘read.csv’.

```
> goodfitness <- read.csv("C:/Users/msiip/Desktop/INFX502/project 2/CardioGoodFitness.csv")
```

3. Variable Identification

To identify variables six following function is used; 'summary', 'dim', 'names', 'tail', 'str', and 'head'. In the project, the 'dim' function is used for identifying the dimension of the matrix, array, or data frame. The 'names' function is used for getting the names of an object. The 'head' and 'tail' functions are used for viewing the top and bottom rows of the dataset. The 'str' function is used for checking dataset variables types and dataset structure. The 'summary' function is used for producing result summaries of the results of various model fitting functions.

4. Data Head

In the figure below the head and tail of the dataset are included. The Head of the dataset is the top six variables in the dataset and the tail dataset is the last six variables of the dataset.

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

Figure 1: Head and Tail of the Dataset

5. Original dataset structure

In this dataset there are 180 rows and 9 columns which is received by 'dim' function. This function reports that the data set has 180 observations and 9 variables.

The first and last 6 rows is received by using 'head' and 'tail' functions. It also provides that in the dataset, there is no formatting issues such as headers or footers.

In the dataset there are 9 variables namely which are received by using 'names' function.

Before visualizing the dataset, the variables types and data structure are checked by using 'str' function. There are three characteristic variables which are 'Product', 'Gender', and 'Maritalstatus'. There are six integers variables which are "Age", "Education", "Usage", "Fitness", "Income" and "Miles" in the dataset. Therefore, the variable types are needed to change and clean.

To receive the statical information of the variables the ‘summary’ function is used. this statistical information of the variables is included minimum, 1st quartile, median, mean, 3rd quartile and maximum.

All this information is included in the Figure 2 below.

```
> dim(goodfitness)
[1] 180   9
> head(goodfitness)
  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
> tail(goodfitness)
    Product Age Gender Education MaritalStatus Usage Fitness Income Miles
175   TM798  38   Male        18     Partnered     5       5 104581   150
176   TM798  40   Male        21        Single     6       5  83416   200
177   TM798  42   Male        18        Single     5       4  89641   200
178   TM798  45   Male        16        Single     5       5  90886   160
179   TM798  47   Male        18     Partnered     4       5 104581   120
180   TM798  48   Male        18     Partnered     4       5  95508   180
> names(goodfitness)
[1] "Product"       "Age"           "Gender"        "Education"     "MaritalStatus" "Usage"         "Fitness"       "Income"
[9] "Miles"
> str(goodfitness)
'data.frame':   180 obs. of  9 variables:
 $ Product      : chr  "TM195" "TM195" "TM195" "TM195" ...
 $ Age          : int  18 19 19 19 20 20 21 21 21 21 ...
 $ Gender       : chr  "Male" "Male" "Female" "Male" ...
 $ Education    : int  14 15 14 12 13 14 14 13 15 15 ...
 $ MaritalStatus: chr  "Single" "Single" "Partnered" "Single" ...
 $ Usage        : int  3 2 4 3 4 3 3 3 5 2 ...
 $ Fitness      : int  4 3 3 3 2 3 3 3 4 3 ...
 $ Income       : int  29562 31836 30699 32973 35247 32973 35247 32973 35247 37521 ...
 $ Miles        : int  112 75 66 85 47 66 75 85 141 85 ...
> summary(goodfitness)
   Product               Age           Gender           Education     MaritalStatus          Usage          Fitness          Income
 Length:180         Min.   :18.00   Length:180         Min.   :12.00   Length:180         Min.   :2.000   Min.   :1.000   Min.   : 29562
 Class :character   1st Qu.:24.00   Class :character   1st Qu.:14.00   Class :character   1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 44059
 Mode  :character   Median :26.00   Mode  :character   Median :16.00   Mode  :character   Median :3.000   Median :3.000   Median : 50597
                    Mean   :28.79                      Mean   :15.57                      Mean   :3.456   Mean   :3.311   Mean   : 53720
                    3rd Qu.:33.00                      3rd Qu.:16.00                      3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 58668
                    Max.   :50.00                      Max.   :21.00                      Max.   :7.000   Max.   :5.000   Max.   :104581
     Miles
 Min.   : 21.0
 1st Qu.: 66.0
 Median : 94.0
 Mean   :103.2
 3rd Qu.:114.8
 Max.   :360.0
```

Figure 2: Original Dataset Structure

6. Updating the dataset structure

In the dataset variable types needed to change. the ‘age’, ‘education’, ‘usage’, ‘income’, and ‘miles’ variable types are changed as numeric variable. The ‘fitness’ variable is changed as a factor variable. After changing variable types, the ‘str’ function is used the received the new variables’ types. Also, the ‘any’ and ‘is.na’ functions are used to ensure whether there is

missing information in the dataset. The ‘is.na’ function print is ‘FALSE’ which means there is no missing part in the dataset. The detail is shown in the Figure 3 below.

```
> goodfitness$Age <- as.numeric(goodfitness$Age)
> goodfitness$Education <- as.numeric(goodfitness$Education)
> goodfitness$Usage <- as.numeric(goodfitness$Usage)
> goodfitness$Fitness <- as.factor(goodfitness$Fitness)
> goodfitness$Income <- as.numeric(goodfitness$Income)
> goodfitness$Miles <- as.numeric(goodfitness$Miles)
> str(goodfitness)
'data.frame':   180 obs. of  9 variables:
 $ Product      : chr  "TM195" "TM195" "TM195" "TM195" ...
 $ Age          : num  18 19 19 19 20 20 21 21 21 21 ...
 $ Gender       : chr  "Male" "Male" "Female" "Male" ...
 $ Education    : num  14 15 14 12 13 14 14 13 15 15 ...
 $ MaritalStatus: chr  "Single" "Single" "Partnered" "Single" ...
 $ Usage        : num  3 2 4 3 4 3 3 3 5 2 ...
 $ Fitness      : Factor w/ 5 levels "1","2","3","4",..: 4 3 3 3 2 3 3 3 4 3 ...
 $ Income       : num  29562 31836 30699 32973 35247 ...
 $ Miles        : num  112 75 66 85 47 66 75 85 141 85 ...
> any(is.na(goodfitness))
[1] FALSE
> summary(goodfitness)
   Product              Age           Gender           Education     MaritalStatus         Usage       Fitness     Income
 Length:180         Min.   :18.00   Length:180         Min.   :12.00   Length:180         Min.   :2.000   1: 2    Min.   : 29562
 Class :character   1st Qu.:24.00   Class :character   1st Qu.:14.00   Class :character   1st Qu.:3.000   2:26    1st Qu.: 44059
 Mode  :character   Median :26.00   Mode  :character   Median :16.00   Mode  :character   Median :3.000   3:97    Median : 50597
                    Mean   :28.79                      Mean   :15.57                      Mean   :3.456   4:24    Mean   : 53720
                    3rd Qu.:33.00                      3rd Qu.:16.00                      3rd Qu.:4.000   5:31    3rd Qu.: 58668
                    Max.   :50.00                      Max.   :21.00                      Max.   :7.000           Max.   :104581
     Miles
 Min.   : 21.0
 1st Qu.: 66.0
 Median : 94.0
 Mean   :103.2
 3rd Qu.:114.8
 Max.   :360.0
```

Figure 3: Updated Dataset Structure

7. Purpose and Expectations

The purpose of this project is to analyze each treadmill product. By analyzing the dataset, the company will understand their customers' features and the correlation between the data variables. In the project, there will be many visual representations based on each variable in the dataset. Each data is compressed with the other; therefore, the company will understand the impact on their treadmill products. The project will be starting to analyze the structure of the dataset. The first step of data analysis will be analyzing the statistical values of the numerical dataset. The purpose of this step is to help the company understand each data with statistical values. Outliers are important for the data, because they can change the report of the analysis. Therefore, outliers and the impact of outliers for each data will be analyzing. Correlation between numerical variables will be analyzing. The purpose of analyzing the correlation is discovering strong and weak correlation between numerical variables. Frequency

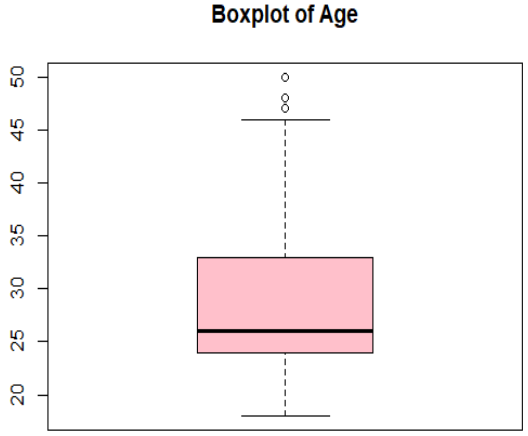
of the variables will be providing the frequency of each data. Numerical variables will be comprising. This part will provide the relationship and impact between variables. The project also included distribution of categorical variables. The goal of this step is discovering the numbers of categorical variables. The expectation of the project is the correlation between Age, Miles and Usage will be strong. Also, the relationship between Income and Education is expected strong relationship. In addition, for TM798 the Usage and Income variables is expected high value.

II. Data Analysis

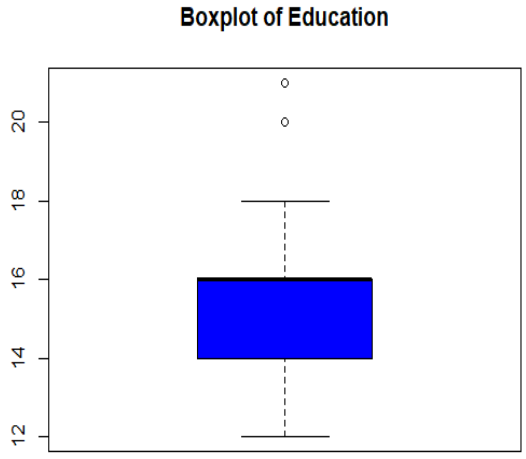
8. Boxplots and Summary Statistics

In this part numerical variables of the dataset will be analyzing by using boxplot. Each numeric variables are summarized separately and included in the table. This section will help to understand each numerical variable better. Based on the boxplots, data in the Income and Miles columns are more consistent in each side of the box.

```
> summary(goodfitness$Age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  18.00   24.00   26.00   28.79   33.00   50.00
> sd(goodfitness$Age)
[1] 6.943498
```

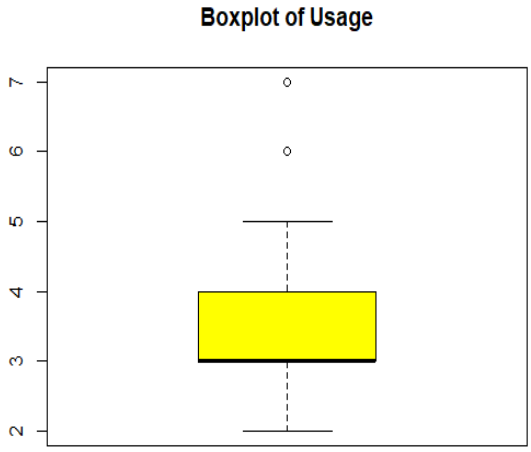



```
> summary(goodfitness$Education)
   Min. 1st Qu.  Median    Mean 3rd Qu.
  12.00   14.00   16.00   15.57   16.00
   Max.
  21.00
> sd(goodfitness$Education)
[1] 1.617055
```



In the Age column, the mean is greater than the median. This provides that the distribution is positively skewed. All the customers' ages are less than 50 years old and older than 18 years old. The range of Age is 32 years old and the standard deviation is high which also shows that data is widely spread out. At least 75% of the customers are 24 years old or older. To compare both sides of the box, from Q1 to Q2 there are not a lot of variabilities and the data is more concentrated. However, from Q2 to Q3 the data is less concentrated. The bottom whisker length is shorter than the top whisker which means the first 25% of the data is more consistent than the last 25% of the data.	In the Education column, the mean is least than the median. This provides that the distribution is negatively skewed. All the customers' education years are less than 21 years and higher than 12 years. The range of Education is 9 years. The standard deviation is low. Therefore, the data is not widely spread out. At least 75% of the customers are educated for 14 years or longer. To compare both sides of the box, from Q1 to Q2 there is a lot of variabilities and the data is less concentrated. However, Q2 and Q3 have the same value. The bottom whisker length is almost the same as the top whisker which means the first 25% of the data and the last 25% of the data has the same concentration.

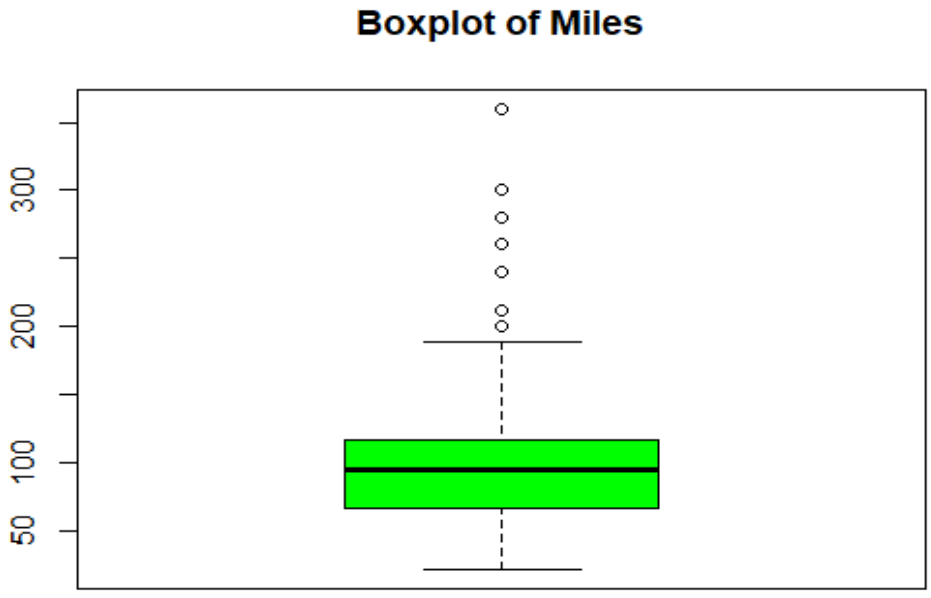
```
> summary(goodfitness$Usage)
Min. 1st Qu. Median Mean 3rd Qu. Max.
2.000 3.000 3.000 3.456 4.000 7.000
> sd(goodfitness$Usage)
[1] 1.084797
```




```
> summary(goodfitness$Income)
Min. 1st Qu. Median Mean 3rd Qu. Max.
29562 44059 50597 53720 58668 104581
> sd(goodfitness$Income)
[1] 16506.68
```

In the Usage column, the mean is greater than the median. This provides that the distribution is positively skewed. All the customers are planning to use the products more than twice a week and less than 7 times a week. The range of Usage is 5 times a week and the standard deviation is high. Therefore, the data is widely spread out. At least 75% of the customers are planning to use the products 4 times or more in a week. To compare both sides of the box, Q1 and Q2 have the same value. However, from Q2 to Q3 the data is less concentrated. The bottom whisker length is almost the same as the top whisker which means the first 25% of the data and the last 25% of the data has the same concentration.	In the Income column, the mean is greater than the median. This provides that the distribution is positively skewed. All the customers' incomes are less than $104581 and higher than $29562. The range of Income is $75,019 and the standard deviation is high which also shows that data is widely spread out. At least 75% of the customers' annual income is $44059 or higher. The variability of both sides of the box is almost the same. So, the data from Q1 to Q2 and from Q2 to Q3 have a similar concentration. The bottom whisker length is shorter than the top whisker which means the first 25% of the data is more consistent than the last 25% of the data.

```
> summary(goodfitness$Miles)
Min. 1st Qu.  Median   Mean   3rd Qu.   Max.
21.0    66.0    94.0   103.2   114.8   360.0
> sd(goodfitness$Miles)
[1] 51.8636
```



In the Miles column, the mean is greatest than the median. This provides that the distribution is positively skewed. All the costumers are planning to walk less than 360 miles and higher than 21 miles each week. The range of the Miles is 339 and the standard deviation is high which also shows that data is widely spread out. At least 75% of the customers are planning to walk 66 miles or higher. To compare both sides of the box, from Q1 to Q2 there is more variability and the data is less concentrated. However, from

Q2 to Q3 the data is more concentrated. The bottom whisker length is shorter than the top whisker which means the first 25% of the data is more consistent than the last 25% of the data.

9. Outliers Analysis

Outliers for each numerical variable are included in the tables below. Basically, outliers are the extreme values of the data. The lower outliers are the lowest data values that are far from the data point and the upper outliers are the highest values that are far from the data point. These outliers have a strong impact on the data statistic because they are extreme values. In the dataset, Income variables have more outliers than any other numerical variables. Education has the lowest outliers in all the numerical variables. All the data have positive outliers except the lower outlier of the Miles variable.

Age Outliers	Usage Outliers	Education Outliers
> loweroutlier [1] 10.5 > upperoutlier [1] 46.5 Outliers: 47 50 48 47 48 Length: 5	> loweroutlier1 [1] 1.5 > upperoutlier1 [1] 5.5 Outliers: 6 6 6 7 6 7 6 6 6 Length: 9	> loweroutlier4 [1] 11 > upperoutlier4 [1] 19 Outliers: 20 21 21 21 Length: 4

Table 1: Outliers of Age, Usage, and Education

Miles Outliers
> loweroutlier3 [1] -7.125
> upperoutlier3 [1] 187.875
Outliers: [1] 188 212 200 200 200 240 300 280 260 200 360 200 200

Length: 13

Table 2: Outliers of Miles

Income Outliers
> loweroutlier2 [1] 22144.88
> upperoutlier2 [1] 80581.88
Outliers: [1] 83416 88396 90886 92131 88396 85906 90886 103336 99601 89641 95866 92131 92131 104581 83416 89641 90886 [18] 104581 95508
Length: 19

Table 3: Outliers of Income

9.1 The impact of outliers on the summary statistic

To investigate the impact of outliers on the numerical data, the data is cleaned from outliers and reported their new statistical value. The new statistic values are under the "out.outlier". In the table below there are both new statistic values that received after deleting outliers and old statistic values of the original data. This statistic report provides the impact of the outliers on the dataset. Belong to the result, the strongest impact of outliers is received on Income and Miles variables.

As is shown the mean value is lower when the outliers are clean in the Age data. The impact of outliers is weak on the Age data.
 	
`> summary(out.outlier)`
`Min. 1st Qu. Median Mean 3rd Qu. Max.`
`18.00 24.00 26.00 28.24 33.00 46.00`
`> summary(goodfitness$Age)`
`Min. 1st Qu. Median Mean 3rd Qu. Max.`
`18.00 24.00 26.00 28.79 33.00 50.00`
	For the Miles data, the impact of outliers is strong. The mean, median, and Q3 values are lower when the outliers are deleted in the Miles data.
	
	`> summary(out.outlier)`
	`Min. 1st Qu. Median Mean 3rd Qu. Max.`
	`21.00 66.00 85.00 93.59 107.50 188.00`
	`> summary(goodfitness$Miles)`
	`Min. 1st Qu. Median Mean 3rd Qu. Max.`
	`21.0 66.0 94.0 103.2 114.8 360.0`

Outliers' impact is stronger on the Income data than other data. When the outliers are deleted in the Income data, all the statistic values are lower expect the minimum.

```
> summary(out.outlier)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29562   43206   48891   49119   54576   77191
> summary(goodfitness$Income)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  29562   44059   50597   53720   58668  104581
```

The impact of outliers is weak on the Usage data because the mean value decreased a little bit when the outliers are cleaned in the Usage data.

```
> summary(out.outlier)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   2.00    3.00    3.00    3.31    4.00    5.00
> summary(goodfitness$Usage)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  2.000   3.000   3.000   3.456   4.000   7.000
```

The impact of the outliers is also weak on the Education data. The mean value is decreased when the outliers are deleted from the education data.

```
> summary(out.outlier)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   14.00   16.00   15.45   16.00   18.00
> summary(goodfitness$Education)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  12.00   14.00   16.00   15.57   16.00   21.00
```

3. Correlation matrix

There are categorical variables in the dataset and categorical variables cannot be used in “cor” function. Therefore, the new dataset is created and only included numerical variables.

```
> goodfitness.SUB <- print(goodfitness[,unlist(lapply(goodfitness, is.numeric))])
```

Now the dataset is ready to use in “cor” function.

```
> fitness.c <- cor(goodfitness.SUB)
> fitness.c
                Age Education      Usage     Income      Miles
Age       1.00000000 0.2804957 0.01506447 0.5134137 0.03661757
Education 0.28049567 1.0000000 0.39515522 0.6258273 0.30728428
Usage     0.01506447 0.3951552 1.00000000 0.5195372 0.75913048
Income    0.51341369 0.6258273 0.51953723 1.0000000 0.54347326
Miles     0.03661757 0.3072843 0.75913048 0.5434733 1.00000000
```

Figure 4: Correlation Matrix

There are no negative correlations in the correlation matrix. To find the max correlations “max” function is used.

```
> max(fitness.c)
[1] 1
> max(fitness.c[fitness.c!=max(fitness.c)])
[1] 0.7591305
```

1 is the highest correlation in the correlation matrix, however, it is between the same data therefore it does not count. So, the second highest correlation is the highest correlation between two different data. The “Max” function is used to find the second and third highest correlation.

The highest correlation is 0.7591305 which is between the Miles and Usage datasets. Therefore, the relationship is strong between Miles and Usage variables. The third highest correlation is 0.6258273 which is between Income and Education variables. These related numerical variables are visualized in the Figure 5 below. The "plot" function is used to visualize the relationship between variables. As it is shown in the plot graph, when the Miles variables increases, the Usage variables also increase and when the Annual income increases, the education years also increase.

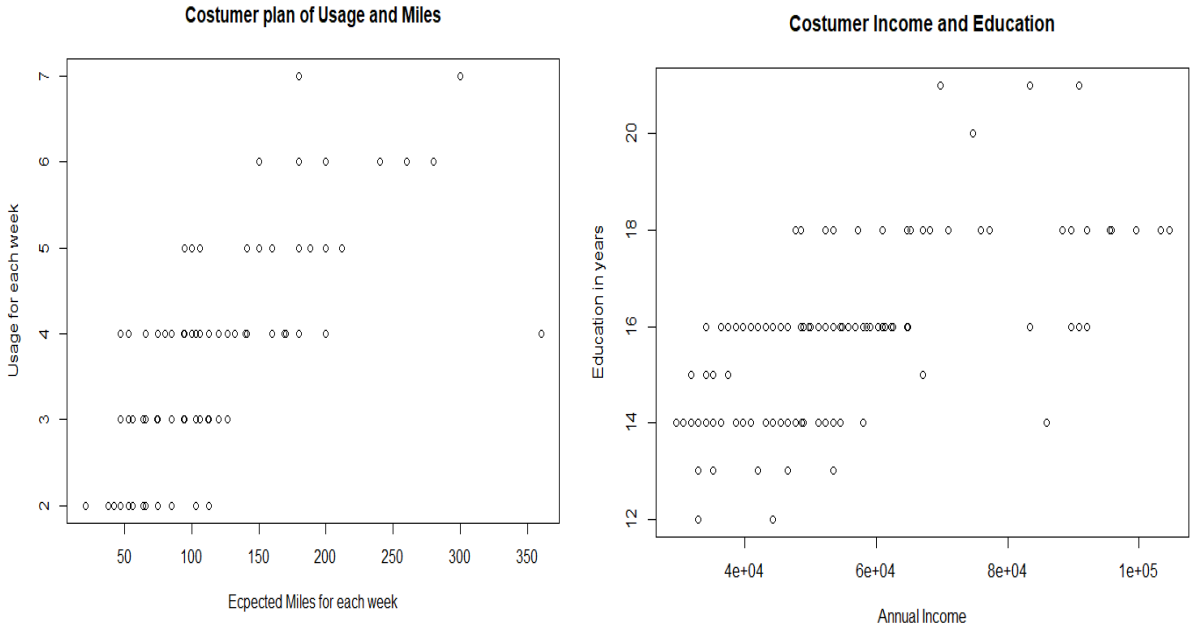


Figure 5: Correlation between Usage and Miles; Education and Income

Correlation can be either strong or weak between variables. These relationships are shown in the correlation matrix below. The relationship between variables is getting stronger when the color gets closer to dark blue and red. The relationship between variables is getting weak when the color gets closer to white. In addition, the correlation is positive on the blue part and it is negative on the red part. The relationship between the same variables of course has the highest correlation value which is 1. As is shown in the correlation matrix the relationship between Miles and Age is very weak because the color is almost white and the correlation value is close to zero. There is also same relationship between Usage and Miles; very weak. The strongest relationship is received between Usage and Miles in the graph.

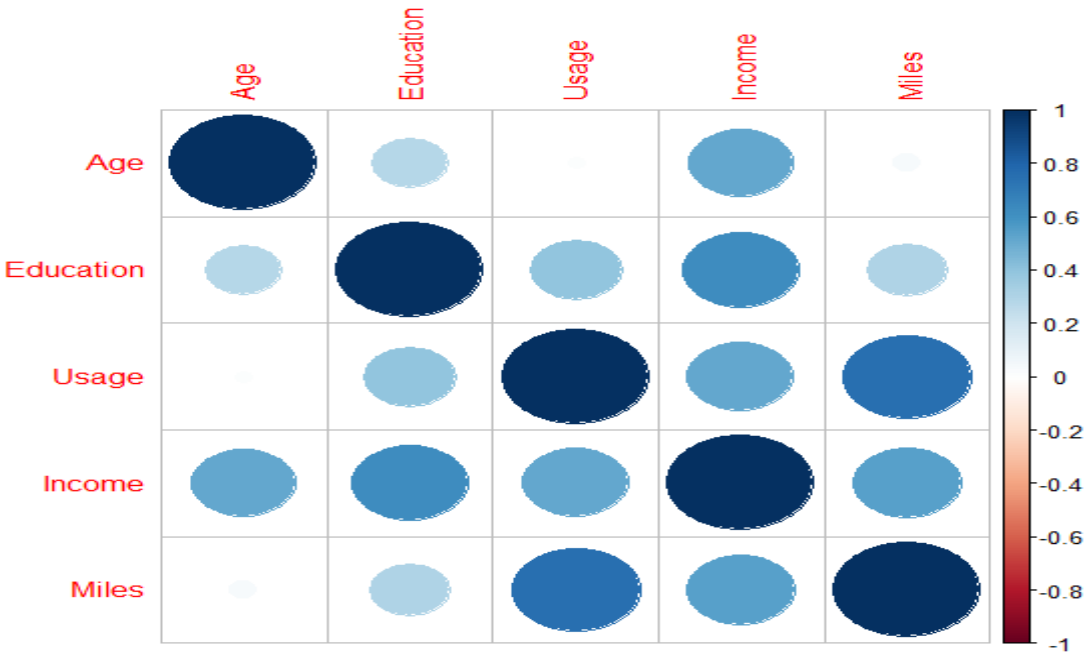


Figure 6: Correlation Matrix of Variables

4. Frequency of Data

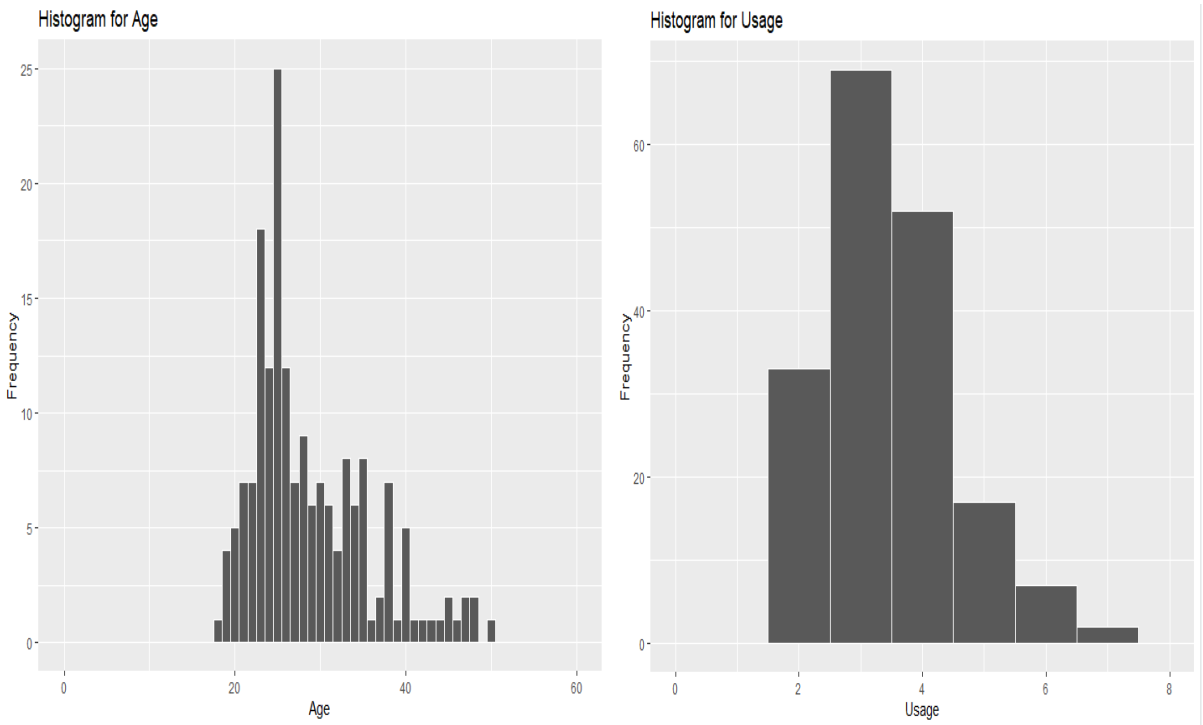


Figure 7:Frequency of Age and Usage

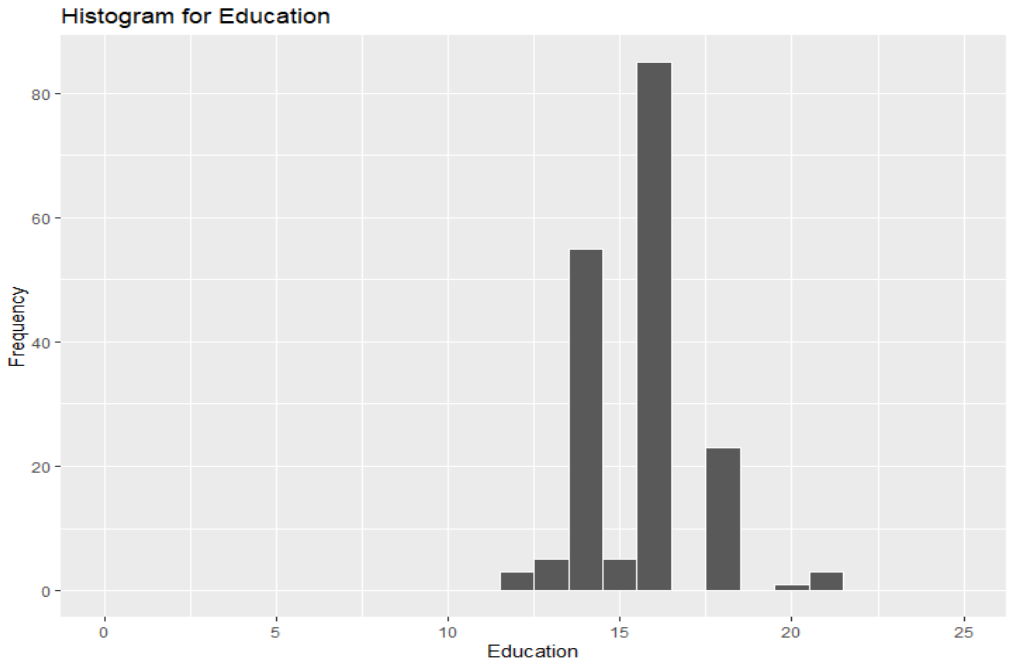


Figure 8:Frequency of Education

The graphs below show the frequency of Miles, Education, Age, and Usage variables.

Histogram for Usage shows most of the customers are planning to use the product 3 times each week, because, the greatest frequency is received on the 3 times. The minority of the customers are planning to use the product 7 times each week, because, the least frequency is received on the 7 times.

The histogram of Education provides most customers have 16 years of education because the highest frequency is received on 16 years of education. The minority of customers have 20 years of education, because, the least frequency is received on 20 years of education.

The histogram for Age shows most customers are 25 years old, because, the highest frequency is received at 25 years old. The minority of customers are 39, 41, 43, 44, 46, and 50 years old. Because the lowest frequency is received at those ages.

5. Correlation between Income and Usage According Miles variables

In this section Income and Usage, numeric variables are visualized based on Miles variables. In the visualization that is Figure 9 below, the Mile variables are categorized into 4 quantile groups; values between the minimum and the first quartile are labeled as Low; between the first and second quartiles are labeled as Normal; between the second and third quartiles are labeled as Medium, and between the third quartile and maximum are labeled as High. The group High has included the high Miles variables and the group Low is included the low Miles variables. So, in the group High covers the customer who is planning to walk the greater miles. The group Low covers the customers who are planning to walk the least miles.

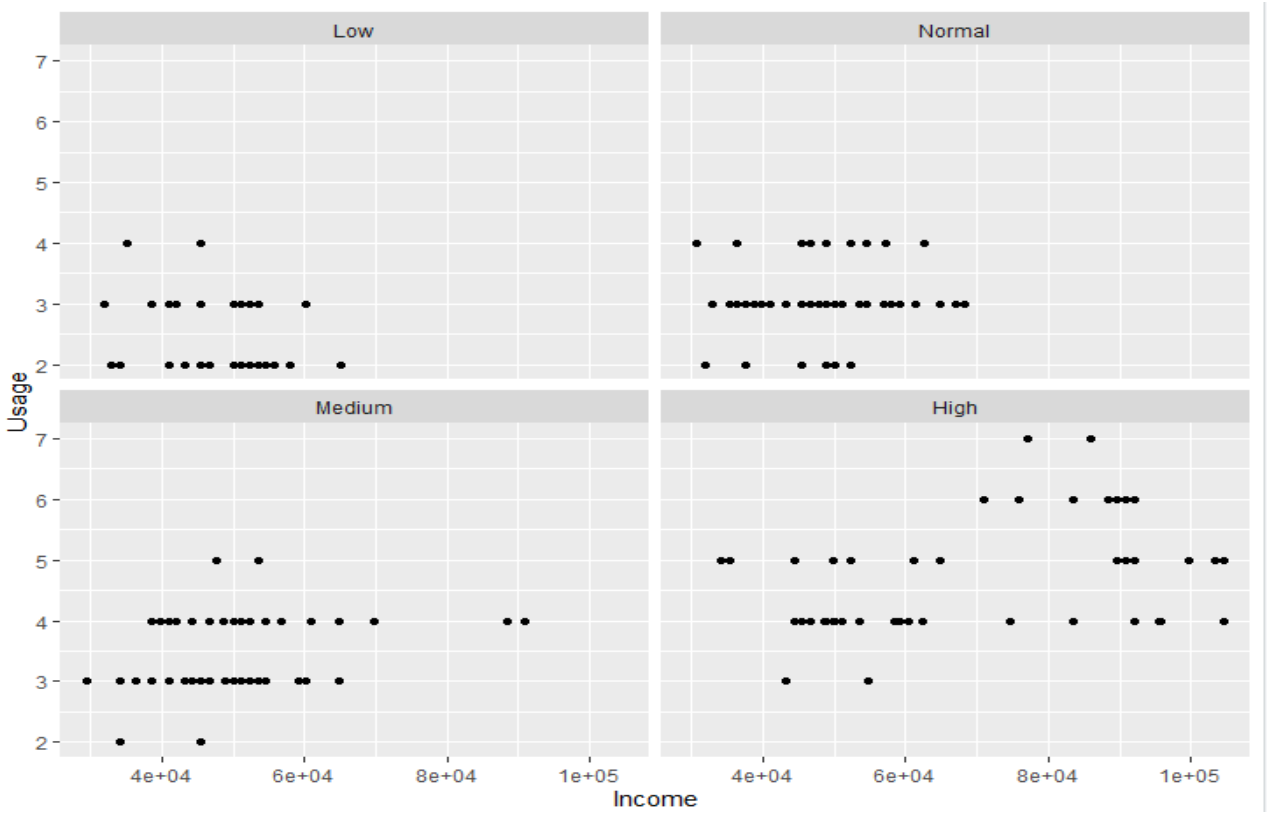


Figure 9: Correlation between Income and Usage According Miles

Group Low;

The average income and usage variables are received in the group low. That means the customers who are planning to walk the least have the lower annual income and uses the products least than other customers. In the group low, the correlation is not widely spread out. Most of the customers are planning to use products 2 times a week which is the lowest usage variable.

Group Normal

The average income and usage variables are higher than the group low. Most of the customers are planning to use the products 3 times a week and the correlation is not widely spread out.

Group Medium;

The average income and usage variables are higher than the group normal. Income variables started to spread out in the medium group. Most of the customers are planning to use the products 3 or 4 times a week.

Group High;

The highest income and usage variables are received in the group high. This proves that customers who are planning to walk high miles are planning to use the products more often. They also have a high income. In addition, the data is widely spread out in the group high.

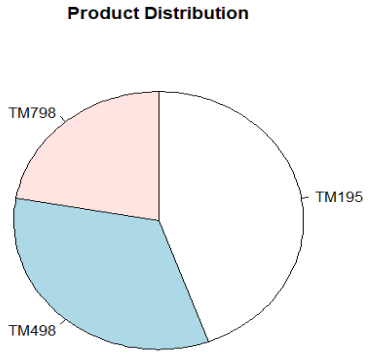
As a result, the relationship between Income, Miles, and Usage variables is directly proportional.

6. Pie Chart and Distribution of Categorical Variables Analysis

Product

As it is mentioned before, there are three products in the dataset. In the pie chart and table on the left, there are the total preferred products by customers. The most popular product is TM195. 80 customers are preferred the TM195. The second most popular product is TM498, and 60 customers are preferred TM498. The last popular product is TM798 and 40 customers are prepared TM 798. 180 products are preferred by customers.

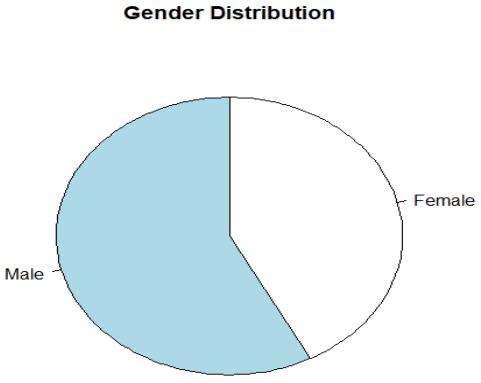
Product	Total
TM195	80
TM498	60
TM798	40
Total	180



Gender

On the left, there are gender table and a pie chart. 80 customers preferred the TM195 half of the 80 customers are male and the other half it is female. So, the total female and male customer numbers are equal for TM195. The female customers are least than male customers for TM498 and TM798. For TM498, the total female customer number is 29 and the total male customer number is 31. For TM798, the female customer number is 7 and the total male customer number is 33. Most of the female customers are preferred TM195 and most of the male customers are preferred TM798.

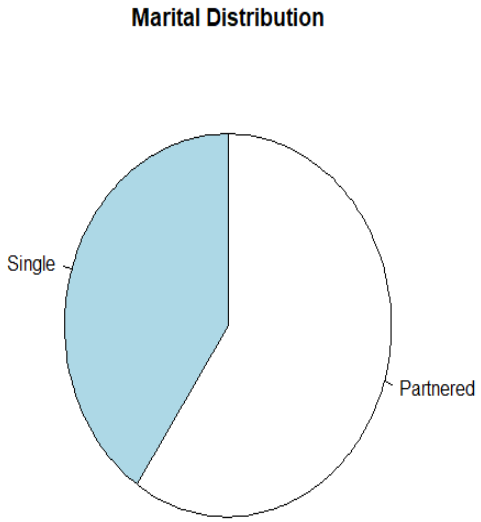
Product	Female	Male
TM195	40	40
TM498	29	31
TM798	7	33
Total	76	104



Marital Status

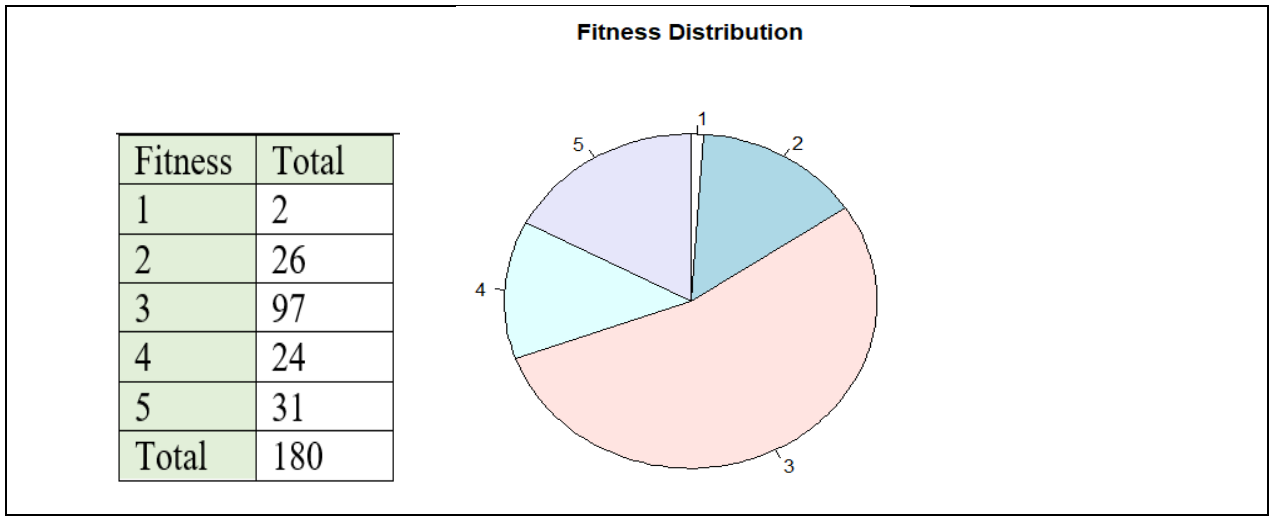
The table and pie chart on the bottom shows the marital status of the customers. 107 of the customers are married and 73 of the customers are single. As a result, customers who are partnered are greatest than costumer who are single.

Marital Status	Total
Single	73
Partnered	107
Total	180



Fitness

On the bottom, the table and pie chart show customer numbers based on their body shape. Customer who has poor body shape is in level 1 and customers who have the fittest body shape are in level 5. There are 2 customers who have a poor body shape and there are 31 customers who have the fittest body shape. Also, the fit body shape number is greatest than any other customer number in the fitness levels. As a result, most of the customers have the good body shape and the least customer has poor body shape in the dataset.



Fitness	Total
1	2
2	26
3	97
4	24
5	31
Total	180

7. Product by Fitness variables

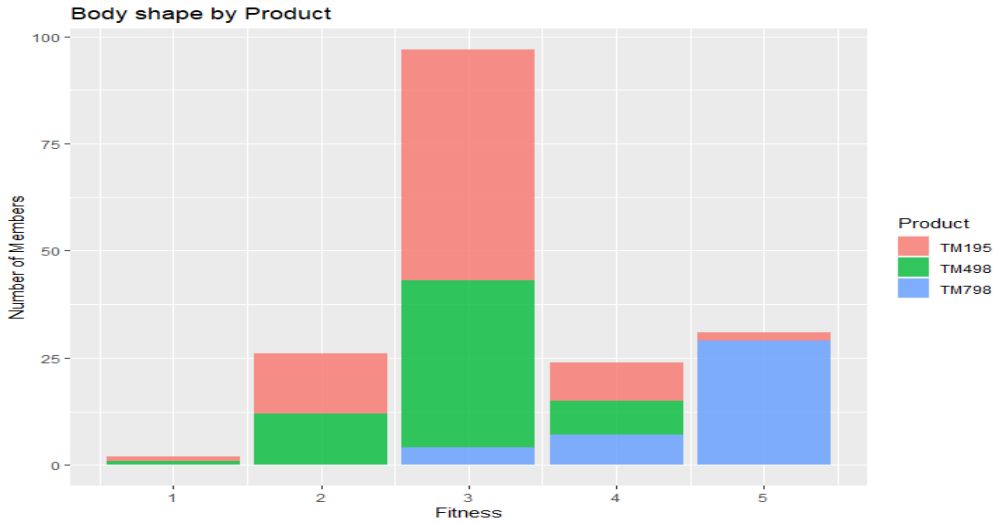


Figure 10: Product by Fitness

The bar graph shows the product distribution into Fitness variables. Each product has a different color in the bar graph and they are visualized based on fitness levels.

Customers who have level 1 body shape:

Most of them are planning to use TM195 and the second preferred product is TM498. None of the customers are planning to use TM798.

Customers who have level 2 body shape:

Most of them are preferred to use TM195 and the second preferred product is TM498. None of the customers are planning to use TM798.

Customers who have level 3 body shape:

Most of them are preferred to use TM195 and the second most preferred product is TM498. The least preferred product is TM798.

Customers who have level 4 body shape:

Distribution of each product is very close to each other, however, most of them are planning to use TM195.

Customers who have level 5 body shape:

Most of them are planning to use TM798. The second most preferred product is TM195 none of them are not planning to use TM498.

8. Product by Gender Variables

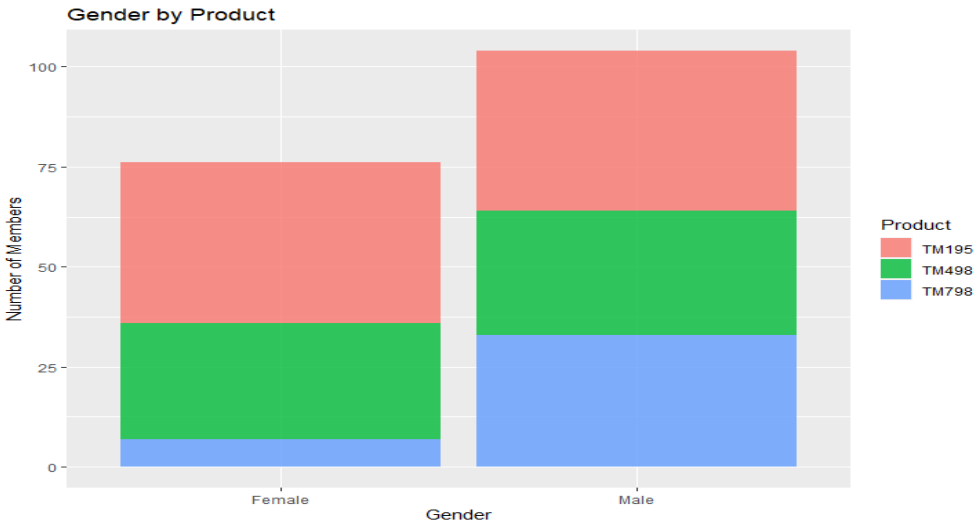


Figure 11: Product by Gender

The bar graph below shows the gender and product distribution. Gender is spread out into two groups; Female and Male. As it is shown male customers are higher than female customers. Most of the female customers preferred to use TM196 and the least female customers are preferred to use TM798. Most of the male customers are planning to use TM195 and the second preferred product by males is TM798. The least preferred product is TM498 among male customers.

9. Product by Marital Status

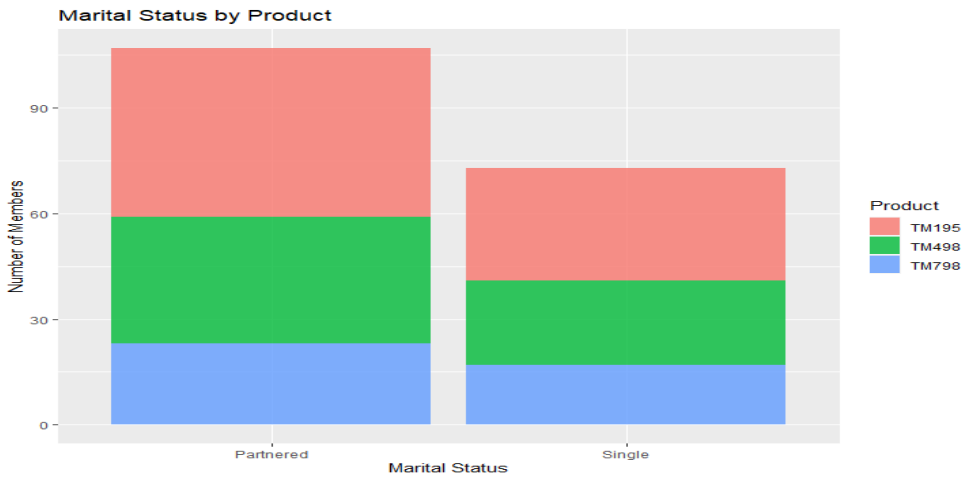


Figure 12: Product by Marital Status

The bar plot below provides product distribution into Marital Status. Marital status is spread out into 2 groups; Partnered and Single. Partnered customers are greater than single customers. Most of the customers who are partnered are preferred to TM195. The second preferred product is TM498 and the least preferred product is TM798 by partnered customers. Most of the single customers are planning to use TM195. The second preferred product is TM498 and the least preferred product is TM798 by single customers.

10. Box plot Analysis for Each Numerical Variables

In this part of the project, each numerical variable will be visualized and analyzed according to product types by using a box plot.

10.1 Income Variables by Product

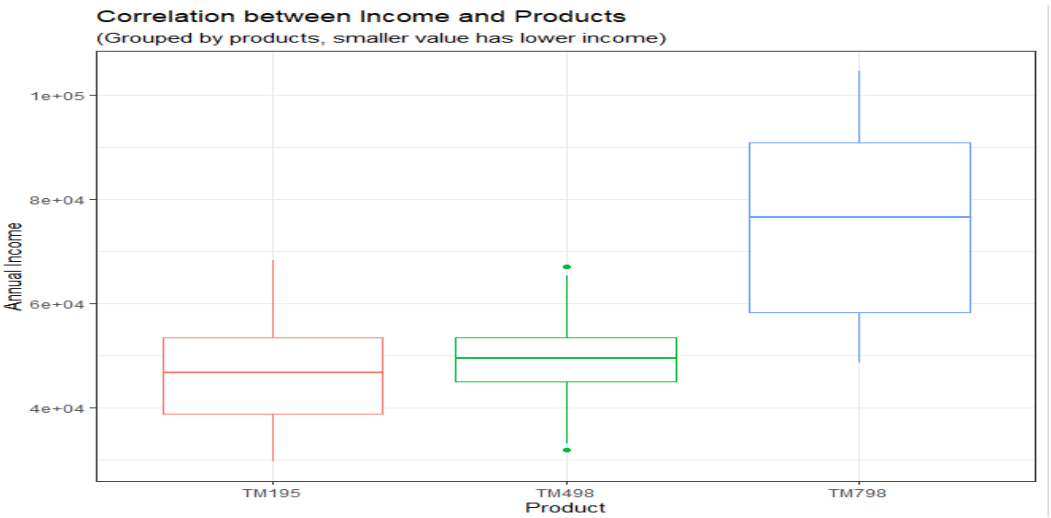


Figure 13: Product and Income

To understand the box plot better the statistical values of Income are shown in the Table 4 below.

Product	Q1	Q2 (median)	Q3	Mean	Minimum mark	Maximum mark	Range	SD
TM798	58,205	76,569	90,886	75,442	48,556	104,581	56,025	18505.84
TM498	44,912	49,460	53,439	48,974	31,836	67,083	35,247	8653.989
TM195	38,658	46,617	53,439	46,418	29,562	68,220	38,658	9075.783

Table 4: Statistical Values of Income

Based on the data distribution the product order is TM798 > TM195 > TM498. In the table, it is received that all the products' median is higher than their mean which shows the distribution is positively skewed for all the products.

Based on the maximum income value the product order is TM798 > TM195 > TM498. Most of the customers who have high annual income purchased TM798 and the customers who have the less income purchased the TM498.

The product order is TM798 > TM498 > TM195 based on the mean and median values. The reason for this is that the income value is greater on the TM798 data and the least income value is on TM498. TM498 and TM195 have the same Q3 value.

10.2 Education Variables by Product

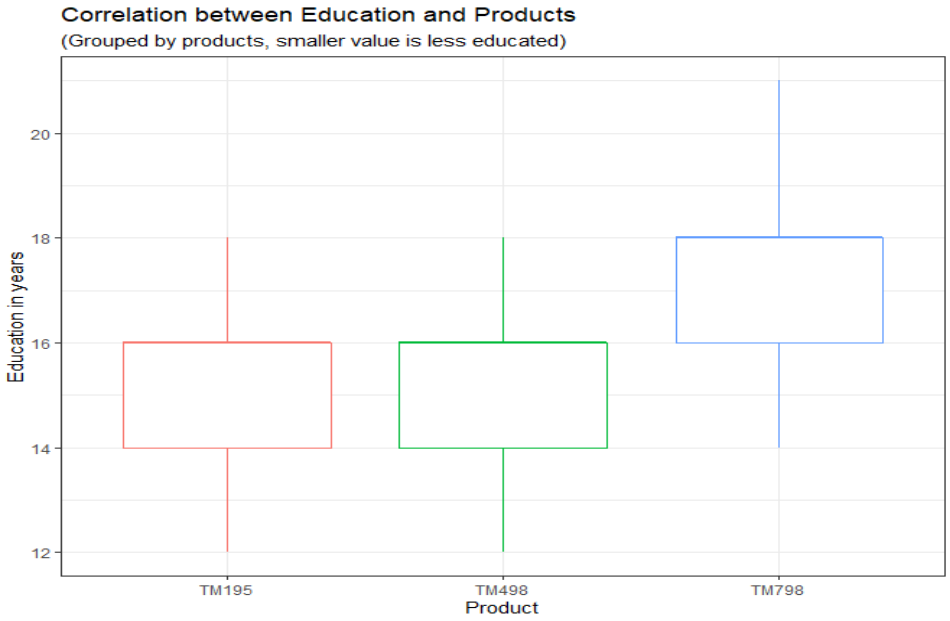


Figure 14: Product and Education

To understand the boxplot better the statistical values of Education are shown in the Table 5 below.

Product	Q1	Q2 (median)	Q3	Mean	Minimum mark	Maximum mark	Range	SD
TM798	16	18	18	17.32	14	21	7	1.6390
TM498	14	16	16	15.12	12	18	6	1.2225
TM195	14	16	16	15.04	12	18	6	1.2163

Table 5: Statistical Values of Education

Based on the data distribution the product order is TM798 > TM195 > TM498. In the table, it is received that all the products' median is higher than their mean which shows the distribution is positively skewed for all the products.

Based on the maximum, minimum, Q1, Q2, and Q3 education values the product order is TM798 > TM195 = TM498.

Most of the customers who have longer education years purchased the TM798 and the customers who have the least income purchased the TM498 and TM195.

Interestingly, for TM498 and TM195, all the statistic values are the same except their mean value. Also, for each product Q1 and Q3 values are the same.

10.3 Miles Variables by Product

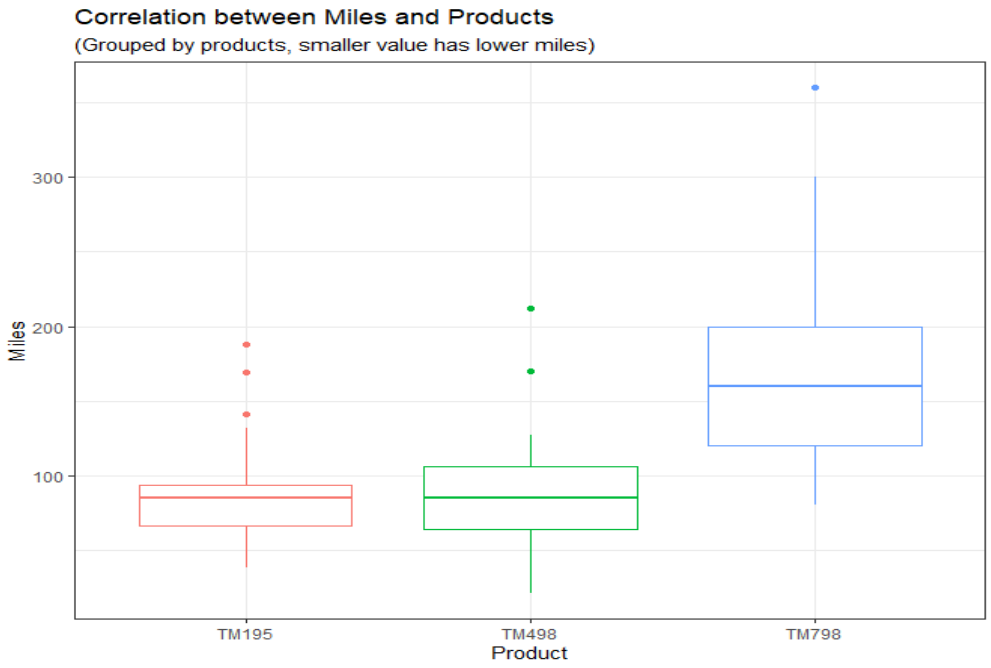


Figure 15: Product and Miles

To understand the boxplot better the statistical values of Miles are shown in the Table 6 below.

Product	Q1	Q2 (median)	Q3	Mean	Minimum mark	Maximum mark	Range	SD
TM798	120	160	200	166.9	80	360	280	60.06654
TM498	64	85	106	87.63	21	212	191	33.26314
TM195	66	85	94	82.79	38	188	150	28.8741

Table 6: Statistical Values of Miles

Based on the data distribution the product order is TM798 > TM498 > TM195. In the table, it is received that median is lower than the mean for TM798 and TM498 which shows the distribution is negatively skewed for these products. However, for TM195 median is higher than the mean which means the distribution is positively skewed for this product.

Based on the maximum and maximum marks the product order is TM798 > TM498 > TM195. The greatest miles value is on TM798 and the least miles value is on TM195.

The Median is the same for TM498 and TM195. TM798 has the greatest median. Therefore, the customer who is planning to use TM798 is also planning to walk longer miles. There are outliers in the box plot; therefore, the data is not distributed normally. Customers who are planning to use TM498 are also planning to walk the least miles.

The data is more concentrated from the median to Q3 from TM195. However, data contribution is similar from Q1 to Q2 and from Q2 to Q3 for TM798 and TM498.

10.4 Income Variables by Product

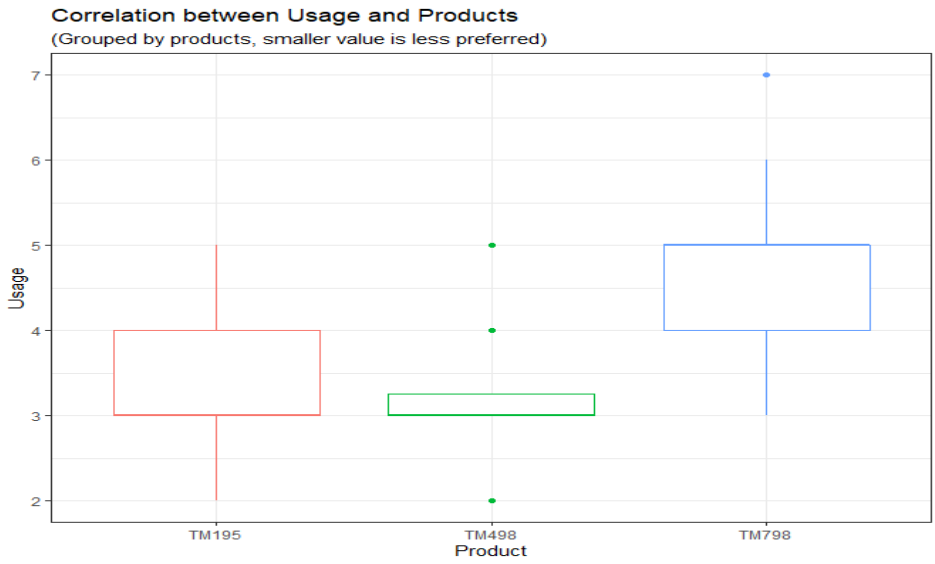


Figure 16: Product and Usage

To understand the boxplot better the statistical values of Usage are shown in the Table 7 below.

Product	Q1	Q2 (median)	Q3	Mean	Minimum mark	Maximum mark	Range	SD
TM798	4	5	5	4.775	3	7	4	0.946992
TM498	3	3	3.067	3.250	2	5	3	0.799717
TM195	3	3	4	3.087	2	5	3	0.782623

Table 7: Statistical Values of Usage

Based on the data distribution the product order is TM798 > TM498 > TM195. In the table, it is received that median is lower than the mean for TM195 and TM498 which shows the distribution is negatively skewed for these products. However, for TM798 median is higher than the mean which means the distribution is positively skewed for this product.

Based on the maximum and maximum mark the product order is TM798 > TM498 = TM195. The greatest usage value is on TM798.

The Median is the same for TM498 and TM195. TM798 has the greatest median. Therefore, the customer number who are planning to use TM798 is higher. There are outliers in the box plot; therefore, the data is not distributed normally

Q1, median, minimum, and maximum values are the same for TM498 and TM195. Q3 and median are the same for TM798.

10.5 Age Variables by Product

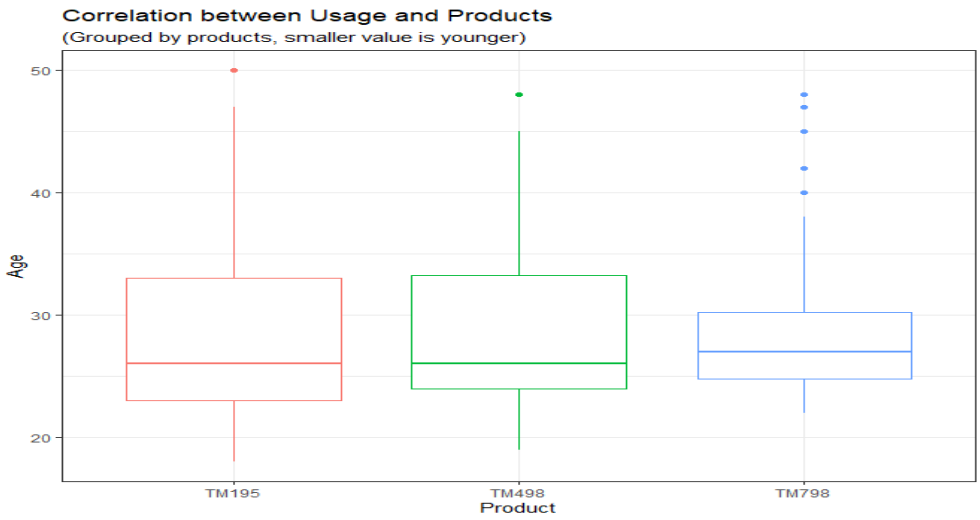


Figure 17: Age and Product

To understand the boxplot better the statistical values of Age are shown in the Table 8 below.

Product	Q1	Q2 (median)	Q3	Mean	Minimum mark	Maximum mark	Range	SD
TM798	24.75	27	30.25	29.10	22	48	26	6.971738
TM498	24	26	33.25	28.90	19	48	29	6.645248
TM195	23	26	33	28.55	18	50	32	7.221452

Table 8: Statistical Values of Age

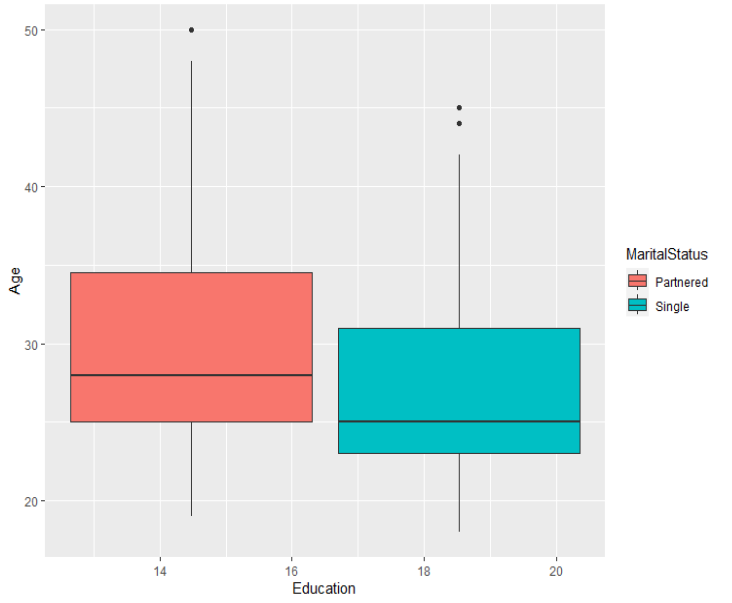
Based on the data distribution the product order is TM195 > TM498 > TM798. In the table, it is received that median is lower than the mean for all the products. That shows the distribution is negatively skewed for these products.

Based on the maximum the product order is TM798 = TM498 < TM195. Therefore, customers who are planning to use TM195 have the oldest.

The Median is the same for TM498 and TM195. TM798 has the greatest median. Therefore, the average age of the customer is higher than other customer average age values. There are outliers in the box plot; therefore, the data is not distributed normally.

The most outliers are received on TM798 and this makes the result unreliable. Data is more concentrated between Q1 on median on TM798 and TM498.

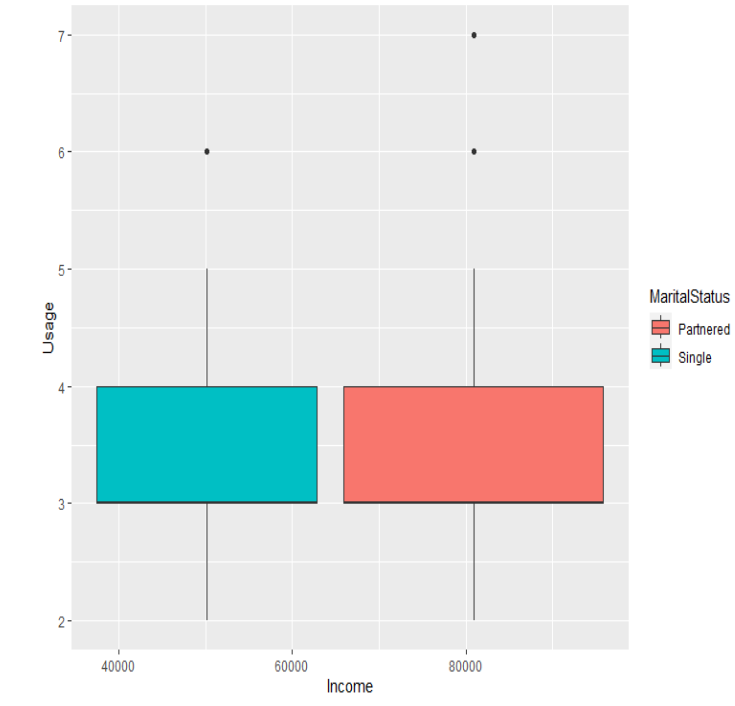
11. Correlation of Numerical Variables on Box Plot



>cor(goodfitness$Education,goodfitness$Age)

[1] 0.2804957

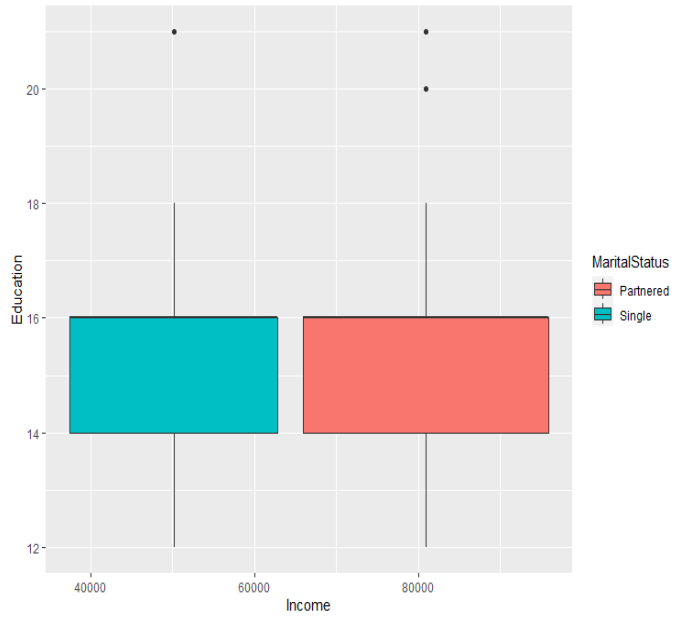
The correlation between Age and Education is weak. The partnered customer has a greater age value in the box plot; however, they have the least education value than the single customer number. The partnered median is greater than the single median. Therefore, the partnered age range is higher than the single customers' age range. There are outliers in the box plot. That means data is not distributed normally. From Q2 to Q3 the data is less concentrated for both single and partnered. The bottom whisker length is shorter than the top whisker which means the first 25% of the data and the last 25% of the data has the same concentration.



>cor(goodfitness$Income,goodfitness$Usage)

[1] 0.5195372

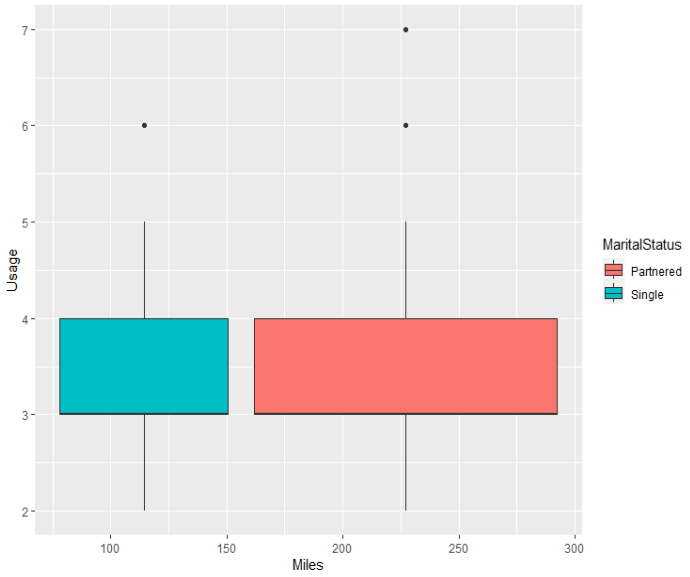
The correlation between Income and Usage is strong. The partnered customers have a higher income than single customers. Single and partnered customers are planning to use the products 3 and 4 times each week. The box plot has outliers; therefore, data is not distributed normally. Their median and Q1 value are equal. The bottom whisker length is almost the same as the top whisker which means the first 25% of the data and the last 25% of the data has the same concentration.



>cor(goodfitness$Education,goodfitness$Income)

[1] 0.6258273

The correlation between Education and Income is very strong. The partnered customers have higher incomes than single customers. Education years are almost the same distribution in both single and partnered customer numbers. The box plot has outliers; therefore, data is not distributed normally. Their median and Q1 value are equal. The bottom whisker length is almost the same as the top whisker which means the first 25% of the data and the last 25% of the data has the same concentration.



> cor(goodfitness$Miles,goodfitness$Usage)

[1] 0.7591305

The correlation between Miles and Usage is very strong. The partnered customers are planning to walk long miles than single customers. Partner customers are planning to walk between 150 and 300 miles. Single customers are planning to walk between 50 to 150 miles. Partnered customers, who are planning to walk longer, are also planning to use products more than single customers who are also planning to walk less. from Q2 to Q3 the data is less concentrated. The bottom whisker length is almost the same as the top whisker which means the first 25% of data and the last 25% of data has the same concentration. The box plot has outliers; therefore, data is not distributed normally. Their median and Q1 value are equal.

12. Heatmap of Variables

12.1 Heatmap of Usage and Fitness Variables

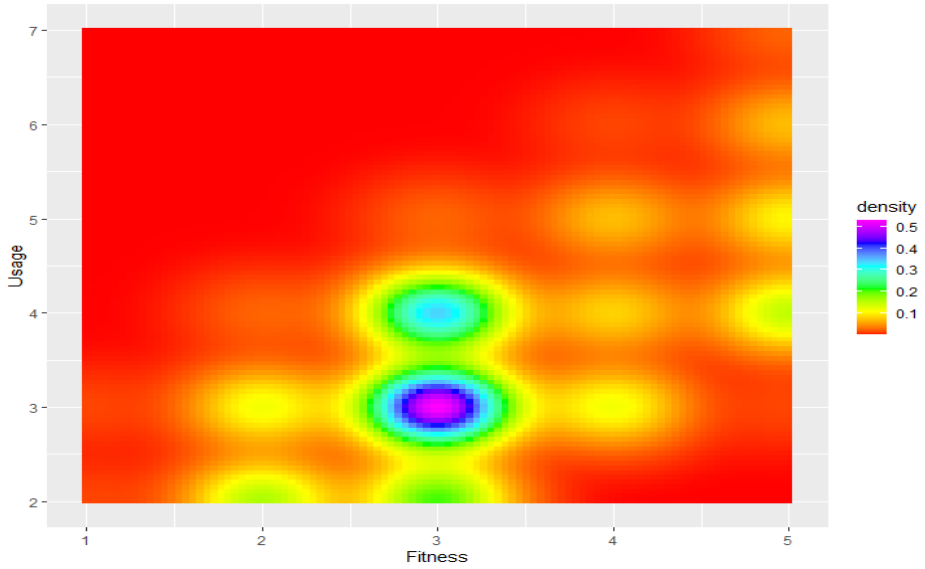


Table 9: Heatmap of Usage and Fitness

Usage variables are analyzed by using “ggplot” function to plot the heatmap. On the map, the density shows the concentration of the data. The density consists of 5 levels. When the level gets closer to 5, the color is turning pink. When the level gets closer to 1, the color is turning red. In the heatmap red areas expresses the least concentration between Fitness and Usage variables. the most concentrated area is received between Usage 3 and Fitness 3 values. That means most of the data is more intensive on 3 values for both variables. In other words, most of the customers have a body shape as level 3 and plan the use the product three times each week. The second concentration area is between the Usage 4 value and Fitness level 3.

12.2 Heatmap of Fitness and Marital Status According Product

Figure 18: Heatmap of Marital Status and Fitness

In the heatmap, Fitness and Marital variables are analyzing according to product. Each product has different color. Customers who are planning to use TM798 have level 5 body shape which is better than rest of the customers' body shape. All the customer who are planning to use TM195 have level 1 body shape which is poor body shape and customers who are planning to use TM498 have level 2 body shape. Interestingly, all the customers who are planning to use TM195 are partnered. Half of the TM798 customers are partnered another half is single. On TM498, there are more customers who are partnered.

13.Correlation Between Miles Variables and Customers' Body Shape

The next visualization is analyzing the Fitness variables based on the Miles variable. As it is mentioned before Fitness is categorized into 5 levels. Level 5 expresses a fit body and group 1 expresses poor body shape. Level 5 has the highest Miles variable which is 360. Customer who has the fittest body walked longer than others. So, it also proves that walking has a positive impact on body shape. Level 1 has the lowest Miles variable. Customers who walked less have poor shape. As a result, the relationship between Miles and Fitness variables is directly proportional. Customer who has fit body shapes walked longer miles and customer who has poor body shape walked the least.

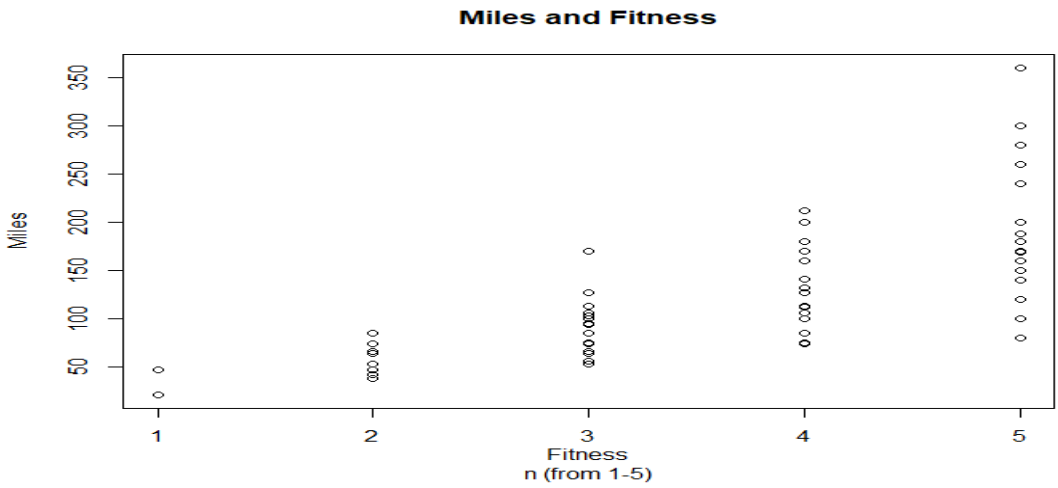


Figure 19: Miles and Fitness

14. Correlation of Miles and Gender According Usage

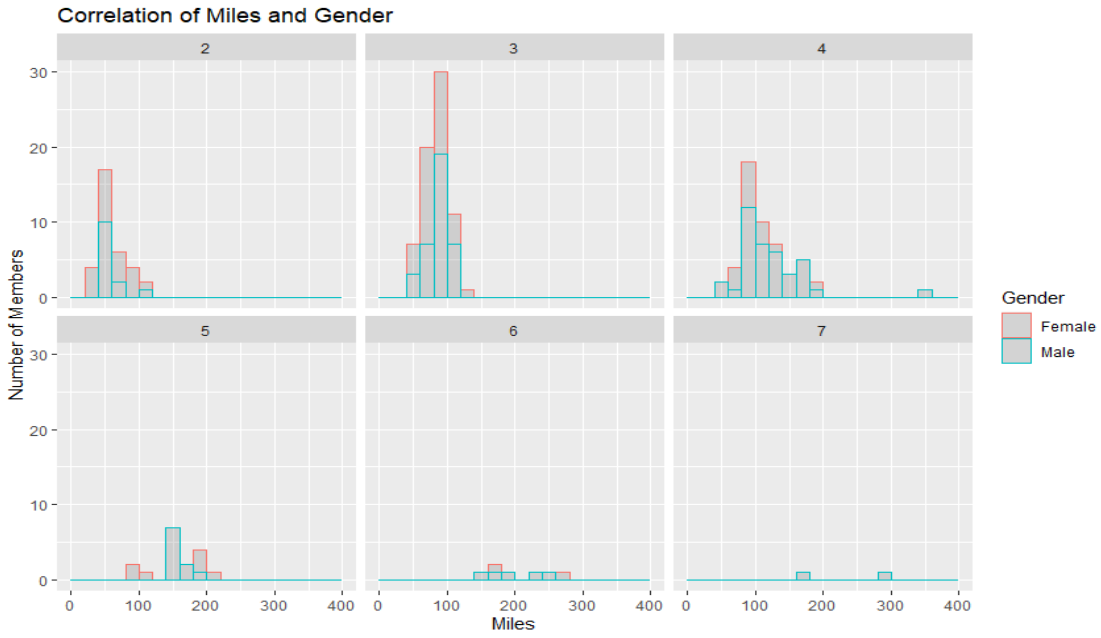


Figure 20: Miles and Usage According Gender

In the bar graphs, gender distribution is analyzed on Miles and Usage variables. Usage data is separated into 6 groups; 2, 3, 4, 5, 6, and 7 times each week. Gender is separated into 2 groups; female and male. In the graph, the Female is red in color and the male is blue in color.

Most of the customers who are planning to use the products 2 times a week are female and they are planning to walk between 20 to 120 miles each week.

Most of the female customers are planning to use the products 3 times a week. Most male customers are planning to use the products 3 times and they are planning to walk between 50 to 100 miles each week.

Most of the customer who is planning to use the products 5 times a week are male and they are planning to walk between 150 to 190 miles each week.

Most of the customer who is planning to use the products 6 times a week are male and they are planning to walk between 150 to 250 miles each week.

All the customer who is planning to use the products 7 times each week is male. None of the female customers are planning to use the product 7 times a week. Those male customers are planning to walk between 160 to 300 miles each week. The customers are widely spread out on in group 7.

15.Chi-Square Test

In this part of the project will cover Chi-square Test which is used to determine if two categorical variables are independent or dependent.

15.1 Marital Status and Product

The relationship between Marital Status and Product matters with the following Hypothesis:

H0: The two categorical variables are independent

H1: The two categorical variables are dependent

```
> cont.table1 = table(goodfitness$MaritalStatus, goodfitness$Product)
> cont.table1

            TM195 TM498 TM798
  Partnered    48    36    23
  Single       32    24    17
> cs.1<- chisq.test(cont.table1)
> cs.1

        Pearson's Chi-squared test

data:  cont.table1
X-squared = 0.080655, df = 2, p-value = 0.9605
```

The Chi-square test for independence is 0.9605 which is higher than the conventionally accepted significance level of < 0.05, therefore, it indicates strong evidence for the null hypothesis. That means the null hypothesis needs to retain and the alternative hypothesis can reject. So, the result is not statistically significant and the variables are independent which is H2.

15.2 Fitness and Product

The relationship between Fitness and Product matters with the following Hypothesis:

H0: The two categorical variables are independent

H1: The two categorical variables are dependent

```
> cont.table2 = table(goodfitness$Fitness, goodfitness$Product)
> cont.table2

   TM195 TM498 TM798
1      1     1     0
2     14    12     0
3     54    39     4
4      9     8     7
5      2     0    29
> cs.2 <- chisq.test(cont.table2)
Warning message:
In chisq.test(cont.table2) : Chi-squared approximation may be incorrect
> cs.2

        Pearson's Chi-squared test

data:  cont.table2
X-squared = 118.78, df = 8, p-value < 2.2e-16
```

The Chi-square test for independence is < 2.2e-16 which is less than the conventionally accepted significance level of < 0.05, therefore, it does not indicate strong evidence for the null hypothesis. That means the null hypothesis can reject. As a result, it is statistically significant and the Fitness and Product variables are dependent which is H1.

15.3 Gender and Product

The relationship between Fitness and Product matters with the following Hypothesis:

H0: The two categorical variables are independent

H1: The two categorical variables are dependent

```
> cont.table3 = table(goodfitness$Gender, goodfitness$Product)
> cont.table3

        TM195 TM498 TM798
  Female    40    29     7
  Male      40    31    33
> cs.3<- chisq.test(cont.table3)
> cs.3

        Pearson's Chi-squared test

data:  cont.table3
X-squared = 12.924, df = 2, p-value = 0.001562
```

The Chi-square test for independence is 0.001562 which is less than the conventionally accepted significance level of < 0.05, therefore, it does not indicate strong evidence for the null hypothesis. That means the null hypothesis can reject. As a result, it is statistically significant and the Gender and Product variables are dependent which is H1. However, Gender is more related to Product than Fitness.

III. Conclusion

Data gives a lot of information that company can increase their sales by following the data result. The company can find which product is most likely to sell based on customer Income, Marital Status, Education level, and Age. Before completing the project, the expectation was before completing the project; receiving a strong correlation between Age, Miles, and Usage variables; receiving a strong correlation between Income and Education; receiving greater Usage and Income value for TM798. The project analysis result;

- Partnered and young people are most likely to buy a treadmill. (Less than 30 years old)
- Most popular product is TM195 which is different than project expectations.
- High-educated people who also have greater income preferer TM798.
- The strongest correlation is between Miles and Usage variables.
- The lowest correlation is between Usage and Age.
- Most of the customers are planning to use a treadmill 3 times each week.
- Male customer number 104 which is higher than female customer number 76.
- Female and male customer numbers are the same for TM195.
- 107 customers are partnered and 73 customers are single.
- Most of the customer's Fitness level is 3.
- All the customers who have level 5 body shapes are preferred to use TM798.
- High miles are received on TM798.
- Younger people are preferring TM798.
- TM498 is popular with people in the income range of 40-60k and with people around 30-35 years old.

IV. Appendix

```
OUTLIERS
quant.age <- quantile(goodfitness$Age)
quant.age
summary(goodfitness$Age)
loweroutlier <- quantile(goodfitness$Age, probs = 0.25, names = FALSE)-1.5*IQR(goodfitness$Age)
loweroutlier
upperoutlier <- quantile(goodfitness$Age, probs = 0.75, names = FALSE)+1.5*IQR(goodfitness$Age)
upperoutlier
v1 <- goodfitness$Age[goodfitness$Age<loweroutlier|goodfitness$Age>upperoutlier]
v1
length(v1)

quant.usage <- quantile(goodfitness$Usage)
quant.usage
summary(goodfitness$Usage)
loweroutlier1 <- quantile(goodfitness$Usage, probs = 0.25, names = FALSE)-1.5*IQR(goodfitness$Usage)
loweroutlier1
upperoutlier1 <- quantile(goodfitness$Usage, probs = 0.75, names = FALSE)+1.5*IQR(goodfitness$Usage)
upperoutlier1
v2 <- goodfitness$Usage[goodfitness$Usage<loweroutlier1|goodfitness$Usage>upperoutlier1]
v2
length(v2)

quant.income <- quantile(goodfitness$Income)
quant.income
summary(goodfitness$Income)
loweroutlier2 <- quantile(goodfitness$Income, probs = 0.25, names = FALSE)-1.5*IQR(goodfitness$Income)
loweroutlier2
upperoutlier2 <- quantile(goodfitness$Income, probs = 0.75, names = FALSE)+1.5*IQR(goodfitness$Income)
upperoutlier2
v3 <- goodfitness$Income[goodfitness$Income<loweroutlier2|goodfitness$Income>upperoutlier2]
v3
length(v3)

quant.miles<- quantile(goodfitness$Miles)
quant.miles
summary(goodfitness$Miles)
loweroutlier3 <- quantile(goodfitness$Miles, probs = 0.25, names = FALSE)-1.5*IQR(goodfitness$Miles)
loweroutlier3
upperoutlier3 <- quantile(goodfitness$Miles, probs = 0.75, names = FALSE)+1.5*IQR(goodfitness$Miles)
upperoutlier3
v4 <- goodfitness$Miles[goodfitness$Miles<loweroutlier3|goodfitness$Miles>upperoutlier3]
v4
length(v4)

quant.education<- quantile(goodfitness$Education)
quant.education
summary(goodfitness$Education)
loweroutlier4 <- quantile(goodfitness$Education, probs = 0.25, names = FALSE)-1.5*IQR(goodfitness$Education)
loweroutlier4
upperoutlier4 <- quantile(goodfitness$Education, probs = 0.75, names = FALSE)+1.5*IQR(goodfitness$Education)
upperoutlier4
v5 <- goodfitness$Education[goodfitness$Education<loweroutlier4|goodfitness$Education>upperoutlier4]
v5
```

```r
goodfitness <- read.csv("C:/Users/msiip/Desktop/INFX502/project 2/CardioGoodFitness.csv", header = TRUE, sep = ",")
goodfitness
dim(goodfitness)
head(goodfitness)
tail(goodfitness)
names(goodfitness)
str(goodfitness)
summary(goodfitness)
goodfitness$Age <- as.numeric(goodfitness$Age)
goodfitness$Education <- as.numeric(goodfitness$Education)
goodfitness$Usage <- as.numeric(goodfitness$Usage)
goodfitness$Fitness <- as.factor(goodfitness$Fitness)
goodfitness$Income <- as.numeric(goodfitness$Income)
goodfitness$Miles <- as.numeric(goodfitness$Miles)
str(goodfitness)
any(is.na(goodfitness))
summary(goodfitness)


install.packages("ggplot2")
library("ggplot2")
library("ggthemes")
install.packages("corrplot")
library(package="corrplot")
install.packages("GGally")
library("GGally")
install.packages("gridExtra")
library("gridExtra")
install.packages("plotly")
library("plotly")
install.packages("AER")
library("AER")


BOXPLOTS AND SUMMARY STATISTIC
boxplot(goodfitness$Age, col="pink", harizontal = TRUE, main="Boxplot of Age")
summary(goodfitness$Age)
sd(goodfitness$Age)
boxplot(goodfitness$Education, col="blue", harizontal = TRUE, main="Boxplot of Education")
summary(goodfitness$Education)
sd(goodfitness$Education)
boxplot(goodfitness$Usage, col="yellow", harizontal = TRUE, main="Boxplot of Usage")
summary(goodfitness$Usage)
sd(goodfitness$Usage)
boxplot(goodfitness$Income, col="red", harizontal = TRUE, main="Boxplot of Income")
summary(goodfitness$Income)
sd(goodfitness$Income)
boxplot(goodfitness$Miles, col="green", harizontal = TRUE, main="Boxplot of Miles")
summary(goodfitness$Miles)
sd(goodfitness$Miles)

OUTLIERS
quant.age <- quantile(goodfitness$Age)
quant.age
```

```
v5
length(v5)

EDUCATION WITHOUT OUTLIERS
education.new <- goodfitness$Education
out.outlier <- education.new[!education.new%in% boxplot(education.new)$out]
out.outlier
summary(out.outlier)
boxplot(out.outlier, col="blue", harizontal = TRUE, main="Boxplot of Education Witout Outliers")

CORRELATION MATRIX
plot(goodfitness$Miles, goodfitness$Usage, main = "Costumer plan of Usage and Miles", xlab = "Ecpected Miles for each week", ylab = "Usage for
plot(goodfitness$Income, goodfitness$Education, main = "Costumer Income and Education", xlab = "Annual Income", ylab = "Education in years")
goodfitness.SUB <- print(goodfitness[,unlist(lapply(goodfitness, is.numeric))])
fitness.c <- cor(goodfitness.SUB)
fitness.c
max(fitness.c)
max(fitness.c[fitness.c!=max(fitness.c)])
corrplot(fitness.c)

NUMERICAL FREQUENCY
qplot(goodfitness$Usage, geom="histogram", binwidth = 1, col=I("white"), main = "Histogram for Usage", xlab
      = "Usage", ylab = "Frequency", xlim=c(0,8)) #Histogram for Usage
qplot(goodfitness$Education, geom="histogram", binwidth = 1, col=I("white"), main = "Histogram for Education", xlab
      = "Education", ylab = "Frequency", xlim=c(0,25))
qplot(goodfitness$Age, geom="histogram", binwidth = 1, col=I("white"), main = "Histogram for Age", xlab
      = "Age", ylab = "Frequency", xlim=c(0,))
qplot(goodfitness$Miles, geom="histogram", binwidth = 1, col=I("white"), main = "Histogram for Miles", xlab
      = "Miles", ylab = "Frequency", xlim=c(0,380))

MILES, INCOME, USAGE CORRELATION
Miles.c1 <- cut(goodfitness$Miles, breaks=quantile(goodfitness$Miles), include.lowest = TRUE, labels =
                c("Low", "Normal", "Medium", "High"), right = FALSE, ordered_result = TRUE)
Miles.c1
goodfitness1 <- cbind(goodfitness, Miles.c1)
goodfitness1
ggplot(data=goodfitness1, mapping=aes(x=Income, y=Usage)) + geom_point() + facet_wrap(facets=~Miles.c1)

GENDER
table(Product,Gender)
pie(table(Gender),main="Gender Distribution", clockwise=TRUE)

PRODUCT
table(Product)
pie(table(Product),main="Product Distribution", clockwise=TRUE)

MARITAL STATUS
table(MaritalStatus)
pie(table(MaritalStatus),main="Marital Distribution", clockwise=TRUE)

FITNESS
table(Fitness)
pie(table(Fitness),main="Fitness Distribution", clockwise=TRUE)
```

```
VARIABLES BY PRODUCT
qplot(Fitness, fill=Product, data=goodfitness, geom ="bar", alpha=I(.8), main="Body shape by Product",xlab="Fitness",ylab="Number of Members")
qplot(Gender, fill=Product, data=goodfitness, geom ="bar", alpha=I(.8), main="Gender by Product",xlab="Gender",ylab="Number of Members")
qplot(MaritalStatus, fill=Product, data=goodfitness, geom ="bar", alpha=I(.8), main="Marital Status by Product",xlab="Marital Status",ylab="Num

CATEGORICAL VARIABLES BY PRODUCT
goodfitness %>% filter(Product %in% goodfitness$Product) %>%
  ggplot(aes(x=Income, y=Product, col=Product)) + guides(col=FALSE) +
  geom_boxplot() +  theme_bw() + coord_flip() +
  labs(x="Annual Income", y="Product",
       title="Correlation between Income and Products", subtitle="(Grouped by products, smaller value has lower income)")

goodfitness %>% filter(Product %in% goodfitness$Product) %>%
  ggplot(aes(x=Education, y=Product, col=Product)) + guides(col=FALSE) +
  geom_boxplot() +  theme_bw() + coord_flip() +
  labs(x="Education in years", y="Product",
       title="Correlation between Education and Products", subtitle="(Grouped by products, smaller value is less educated)")

grid.arrange(d1,d2,ncol=2)

goodfitness %>% filter(Product %in% goodfitness$Product) %>%
  ggplot(aes(x=Miles, y=Product, col=Product)) + guides(col=FALSE) +
  geom_boxplot() +  theme_bw() + coord_flip() +
  labs(x="Miles", y="Product",
       title="Correlation between Miles and Products", subtitle="(Grouped by products, smaller value has lower miles)")

goodfitness %>% filter(Product %in% goodfitness$Product) %>%
  ggplot(aes(x=Usage, y=Product, col=Product)) + guides(col=FALSE) +
  geom_boxplot() +  theme_bw() + coord_flip() +
  labs(x="Usage", y="Product",
       title="Correlation between Usage and Products", subtitle="(Grouped by products, smaller value is less preferred)")

grid.arrange(d3,d4,ncol=2)


goodfitness %>% filter(Product %in% goodfitness$Product) %>%
  ggplot(aes(x=Age, y=Product, col=Product)) + guides(col=FALSE) +
  geom_boxplot() +  theme_bw() + coord_flip() +
  labs(x="Age", y="Product",
       title="Correlation between Usage and Products", subtitle="(Grouped by products, smaller value is younger)")


TABLES OF PRODUCT
TM798 <- subset(goodfitness,goodfitness$Product=="TM798")
summary(TM798$Age)
summary(TM798$Miles)
summary(TM798$Income)
summary(TM798$Education)
summary(TM798$Usage)
sd(TM798$Income)
sd(TM798$Education)
sd(TM798$Miles)
sd(TM798$Usage)
sd(TM798$Age)
```

```
sd(TM798$Miles)
sd(TM798$Usage)
sd(TM798$Age)

TM498 <- subset(goodfitness,goodfitness$Product=="TM498")
summary(TM498$Age)
summary(TM498$Miles)
summary(TM498$Income)
summary(TM498$Education)
summary(TM498$Usage)
sd(TM498$Usage)
sd(TM498$Miles)
sd(TM498$Income)
sd(TM498$Education)
sd(TM498$Age)

TM195 <- subset(goodfitness,goodfitness$Product=="TM195")
summary(TM195$Age)
summary(TM195$Miles)
summary(TM195$Income)
summary(TM195$Education)
summary(TM195$Usage)
sd(TM195$Income)
sd(TM195$Education)
sd(TM195$Usage)
sd(TM195$Miles)
sd(TM195$Age)


BOXPLOT OF NUMERICAL
ggplot(goodfitness, aes(x=Education, y=Age, fill=MaritalStatus)) +
  geom_boxplot()
cor(goodfitness$Education,goodfitness$Age)

ggplot(goodfitness, aes(x=Income, y=Usage, fill=MaritalStatus)) +
  geom_boxplot()
cor(goodfitness$Income,goodfitness$Usage)

ggplot(goodfitness, aes(x=Miles, y=Usage, fill=MaritalStatus)) +
  geom_boxplot()
cor(goodfitness$Miles,goodfitness$Usage)

ggplot(goodfitness, aes(x=Income, y=Education, fill=MaritalStatus)) +
  geom_boxplot()
cor(goodfitness$Education,goodfitness$Income)


HEATMAP OF FITNES AND USAGE
ggplot(data=goodfitness, mapping=aes(x=Fitness, y=Usage)) + stat_density2d(geom = "tile",
                                                                           contour=FALSE, aes(fill = ..density..)) + scale_fill_gradientn(color

FITNESS AND MILES
plot(goodfitness$Fitness, goodfitness$Miles, main = "Miles and Fitness", xlab = "Fitness
n (from 1-5)", ylab = "Miles")
```

```
MILES, GENDER,USAGE
ggplot(goodfitness, aes(Miles, color=Gender)) +
  geom_histogram(breaks=seq(0,400,by=20),alpha=0.2) + facet_wrap(~Usage)+
  labs(x="Miles",y="Number of Members", title=paste("Correlation of Miles and Gender"))


ggpairs(goodfitness)
```