Exploratory Data Analysis in R

Cardio Fitness Project

Tobiloba Adaramati

July 10, 2020

Contents

1. Project Objective

The objective of this report is to explore the Cardio Fitness dataset in R and extract basic observations about the data.

This exploratory report will consist of the following:

- Importing the data in R
- Understanding the structure of the dataset
- Basic summary of data and graphical exploration
- Identify differences between customers of each product

- Explore relationships between the different attributes of customers
- Coming up a customer profile (characteristics of a customer) of the different products
- Perform uni-variate and bivariate analyses
- Generate a set of insights and recommendations that will help the company in targeting new customers

2. Assumptions

The data is about the customers of the treadmill product(s) of a retail store called Cardio Good Fitness. The dataset used is representative of the population data.

3. Exploratory Data Analysis Steps

3.1 Environmental Set up and Import

3.1.1 Install Necessary Packages

```
## Install and load packages useful for the analysis

library(dplyr) # To manipulate the data
library(rpivotTable)
library(ggplot2) # To create plots
library(psych) # multivariate analysis
library(corrplot) # To plot correlation plot between numerical variables
library(gridExtra) # To plot multiple ggplot graphs in a grid
library(psych) # multivariate analysis
library(knitr) # Necessary to generate source codes from a .Rmd File
```

3.1.2 Set Working Directory

Here we indicate the directory the data is stored in

```
## Set working directory

setwd("C:/Users/bdiam/OneDrive/Desktop/Uni. of Texas/Introduction to R for
Analytics/Datasets used in course")
```

3.1.3 Import the file to use for the analysis

```
## Import the file to use for the analysis
cardio = read.csv("CardioGoodFitness.csv", header = TRUE
```

3.1.4 Global Options Settings

```
## 1.4 Change settings to turn off scientific notation
options(scipen=999) ## This is changed in the global options settings
```

3.2 Variable Identification

To get familiar with the cardio fitness data, the following functions would be used to get an overview

- dim(): this gives us the dimension of the dataset provided. Knowing the data dimension gives us an idea of the size of the data.
- head(): this shows the first 6 rows(observations) of the dataset in tabular form.
- tail(): this shows the last 6 rows(observations) of the dataset. Knowing what the dataset looks like at the end rows also helps us ensure the data is consistent.
- str(): this shows us the structure of the dataset. It helps us determine the datatypes of the features and identify if there are datatype mismatches, so that we can convert them where necessary.
- summary(): this provides statistical summaries of the dataset. This function is important as we can quickly get the 5 number statistical summaries (mean, median, quartiles, min, frequencies/counts, max values etc.).
- View(): helps to look at the entire dataset at a glance

3.2.1 Insights from Variable Identification

```
# Identify the variables of the data
## View some portion of the data at a time for better understanding

names(cardio) #This shows the names of the columns
## [1] "Product"       "Age"            "Gender"         "Education"
## [5] "MaritalStatus" "Usage"          "Fitness"        "Income"
## [9] "Miles"
```

- The dataset has 9 distinct variables

Insight(s) from dim() function

```
dim(cardio) #This returns the dimension of the dataset

## [1] 180   9
```

- This shows that the dataset has 180 observations (rows) and 9 columns or attributes

Insight(s) from head() function

```
head(cardio) #This shows the first 6 observations in the data

##   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
```

- Product contains the model no. of the treadmill
- Age shows the number of years of the customer
- Gender shows the sex of the customer
- Education contains the number of years of education the customer has attained

- Marital Status contains information of a customer’s relationship
- Usage contains the avg. number of times the customer wants to use the treadmill every week
- Fitness contains self-rated fitness score of the customer (5 - very fit, 1 - very unfit)
- Income shows the customer income earned
- Miles contains the distance the customer is expected to run on the treadmill

Insight(s) from tail() function

```
tail(cardio) #This shows the last 6 observations in the data

##     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
```

- Output executed shows the last part of the data is consistent with the upper part

Insight(s) from str() function

```
str(cardio) # This shows the structure of the dataset

## '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 ...
```

- There are 3 character variables and 6 numerical variables

Insight(s) from summary() function

```
summary (cardio) #This provides statistical summary for each column in the dataset

##    Product               Age            Gender             Education    
##  Length:180         Min.   :18.00   Length:180         Min.   :12.00  
##  Class :character   1st Qu.:24.00   Class :character   1st Qu.:14.00  
##  Mode  :character   Median :26.00   Mode  :character   Median :16.00  
##                     Mean   :28.79                      Mean   :15.57  
##                     3rd Qu.:33.00                      3rd Qu.:16.00  
##                     Max.   :50.00                      Max.   :21.00  
##  MaritalStatus          Usage          Fitness          Income      
##  Length:180         Min.   :2.000   Min.   :1.000   Min.   : 29562  
##  Class :character   1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 44059  
##  Mode  :character   Median :3.000   Median :3.000   Median : 50597  
##                     Mean   :3.456   Mean   :3.311   Mean   : 53720  
##                      3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 58668
##                      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
```

- This shows the classes of the dataset
- There are no missing values in the dataset
- The average age of the customers is 28.8 years
- Altogether, the customers have spent an average of 15.6 years attaining education
- The customers indicate that they would use the fitness instrument an average of 3 times a week
- The income of the customers ranges from 29,000 to 104,000 with a mean earning of 53,700
- Customers are expected to run an average of 103 miles on the treadmill
- Based on the context of the data, the variables – product, gender, marital status, usage, and fitness - will be changed to factors so they can form categories to differentiate the customers.

```
## Change the class of some variables
cardio$Fitness = as.factor(cardio$Fitness)
cardio$Product = as.factor(cardio$Product)
cardio$Gender = as.factor(cardio$Gender)
cardio$MaritalStatus = as.factor(cardio$MaritalStatus)
cardio$Usage = as.factor(cardio$Usage)
```

```
## Check the structure again

str(cardio)
## 'data.frame':    180 obs. of  9 variables:
##  $ Product      : Factor w/ 3 levels "TM195","TM498",..: 1 1 1 1 1 1 1 1 1 1 ...
##  $ Age          : int  18 19 19 19 20 20 21 21 21 21 ...
##  $ Gender       : Factor w/ 2 levels "Female","Male": 2 2 1 2 2 1 1 2 2 1 ...
##  $ Education    : int  14 15 14 12 13 14 14 13 15 15 ...
##  $ MaritalStatus: Factor w/ 2 levels "Partnered","Single": 2 2 1 2 1 1 1 2 2 1 ...
##  $ Usage        : Factor w/ 6 levels "2","3","4","5",..: 2 1 3 2 3 2 2 2 4 1 ...
##  $ Fitness      : Factor w/ 5 levels "1","2","3","4",..: 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 ...
```

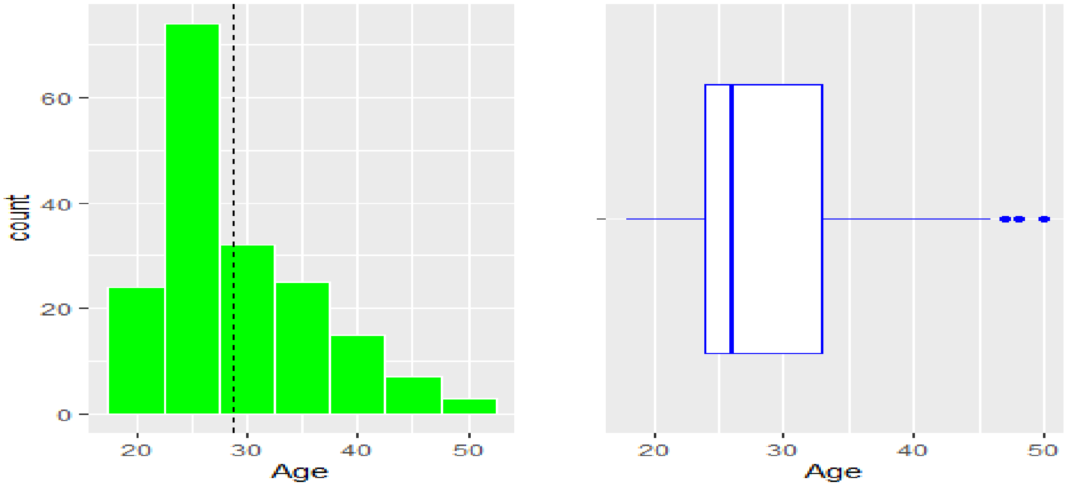
The data is now in the correct format required for the analysis

```r
## View the entire dataset at a glance

View(cardio)
```

3.3 Univariate Analysis

```r
## Explore the numerical variables one after the other
### Create a function to plot histogram and box plot for all numerical variable
plot_histogram_n_boxplot = function(variable, variableNameString, binw)
  {  h = ggplot(data = cardio, aes(x= variable))+
    labs(x = variableNameString,y ='count')+
    geom_histogram(fill = 'green',col = 'white',binwidth = binw)+
    geom_vline(aes(xintercept=mean(variable)),
                color="black", linetype="dashed", size=0.5)
  b = ggplot(data = cardio, aes('',variable))+
    geom_boxplot(outlier.colour = 'blue',col = 'blue',outlier.shape = 19)+
    labs(x = '',y = variableNameString)+ coord_flip()
grid.arrange(h,b,ncol = 2)
}
```

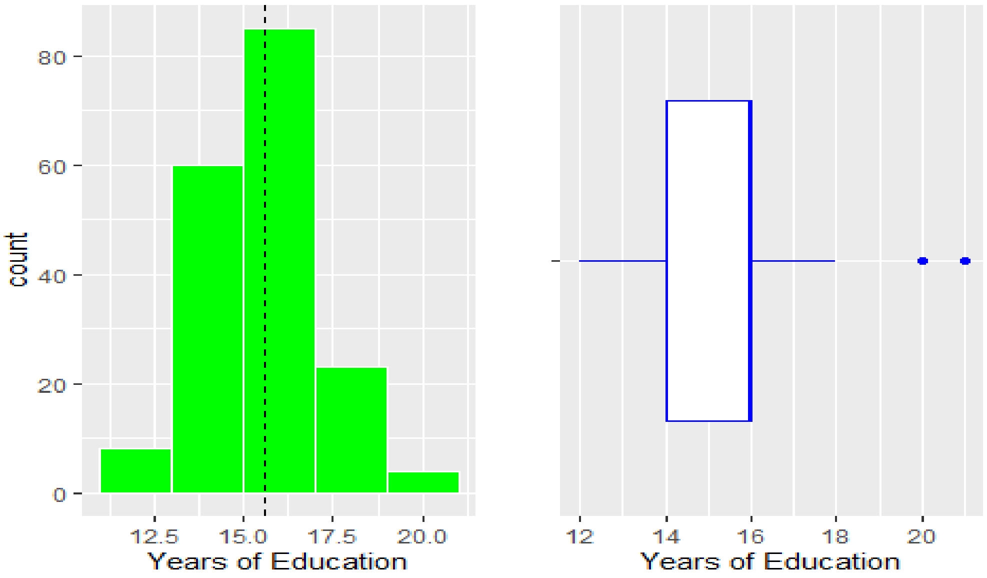
```r
## Histogram and box plots for all numerical variables

plot_histogram_n_boxplot(cardio$Age, 'Age', 5) #Histogram and Box plot for customer age
```



Observations on Age

- The distribution of age is skewed to the right
- There are many more customers between the age of 25 and 40
- The customers are either above or below 26 years

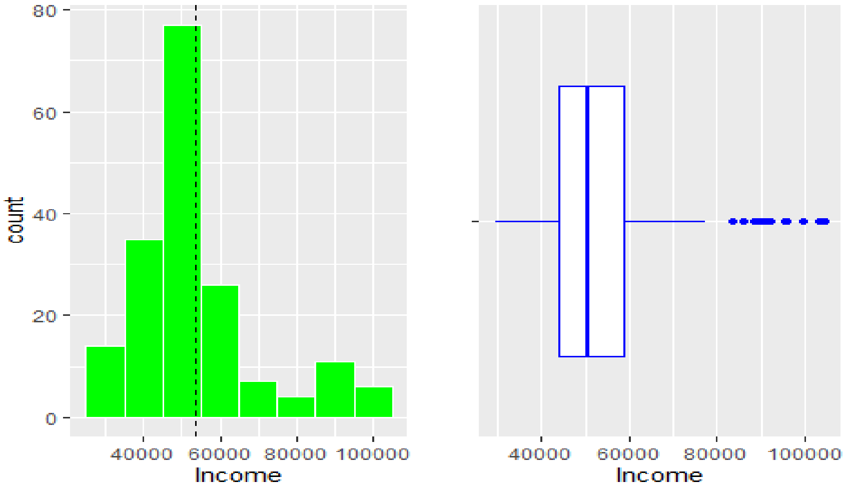
```r
plot_histogram_n_boxplot(cardio$Education, 'Years of Education', 2) #Histogram and Box plot for customer education
```



Observations on Education

- The distribution of education is symmetrical in nature (uniform) i.e. the mean, median and mode are similar

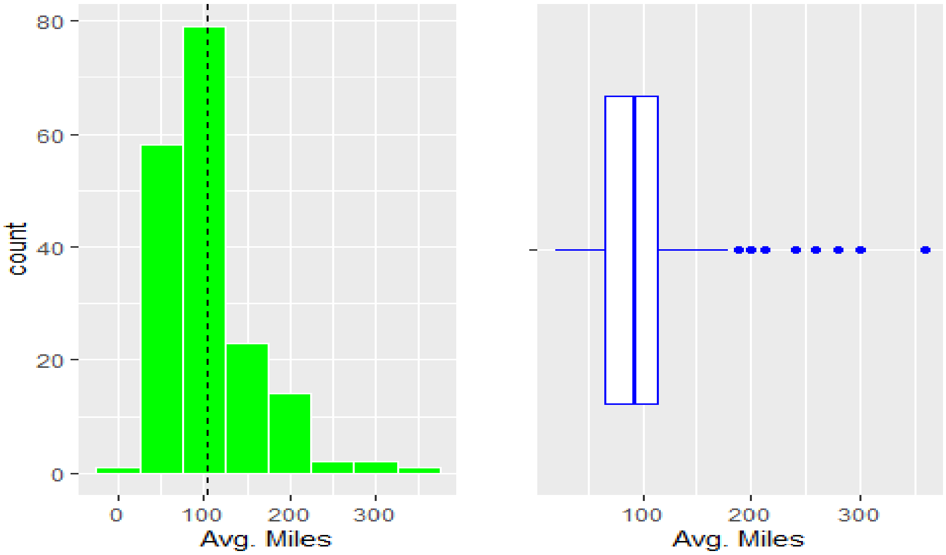
```
plot_histogram_n_boxplot(cardio$Income, 'Income', 10000) #Histogram and Box plot for
customer income
```



Observations on income

- The distribution of age is skewed to the right
- There are outliers in this variable. We have observations where some customers earn twice as much as the average income.
- Majority of the income earned is close between 29,500 and 53,000

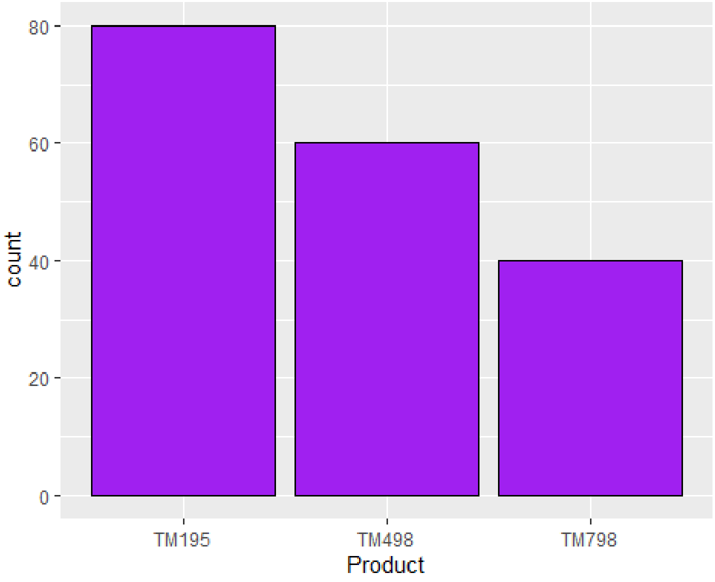
```
plot_histogram_n_boxplot(cardio$Miles, 'Avg. Miles', 50) #Histogram and Box plot for
miles
```



Observations on expected miles to run

- The distribution of age is skewed to the right
- There are outliers in this variable. We have observations where some customers plan to run around 3 times more than the average miles
- The customers expect to run an average of 103 miles after purchasing the treadmill

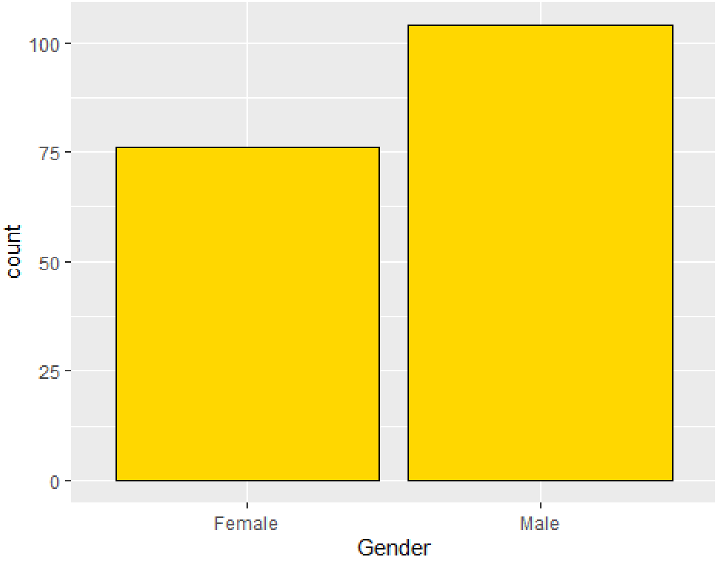
```
## Explore the categorical variables one after the other

### Bar graph for products
ggplot(cardio, aes(x = Product)) +
  geom_bar(fill = c("purple"), color="black")
```



Observations on treadmills

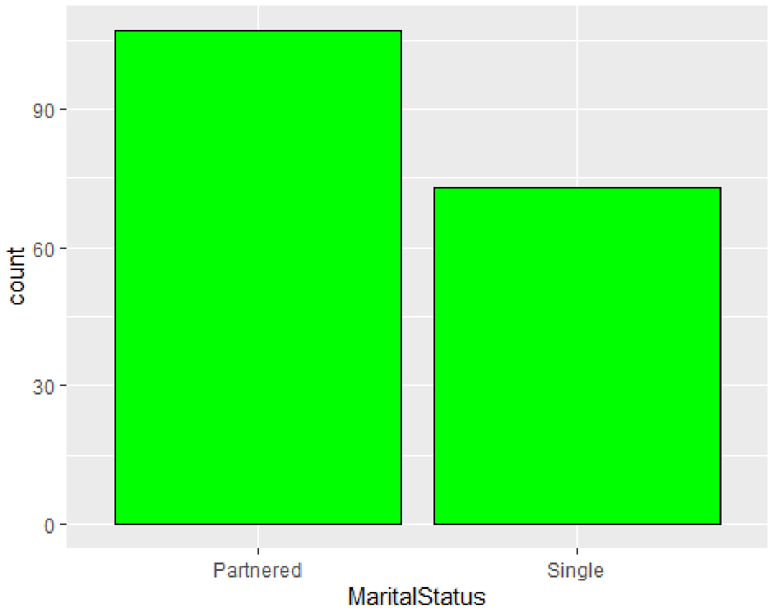
- There is a higher demand for TM195 than the other products

```
### Bar graph for gender
ggplot(cardio, aes(x = Gender)) +
  geom_bar(fill = c("gold"), color="black")
```



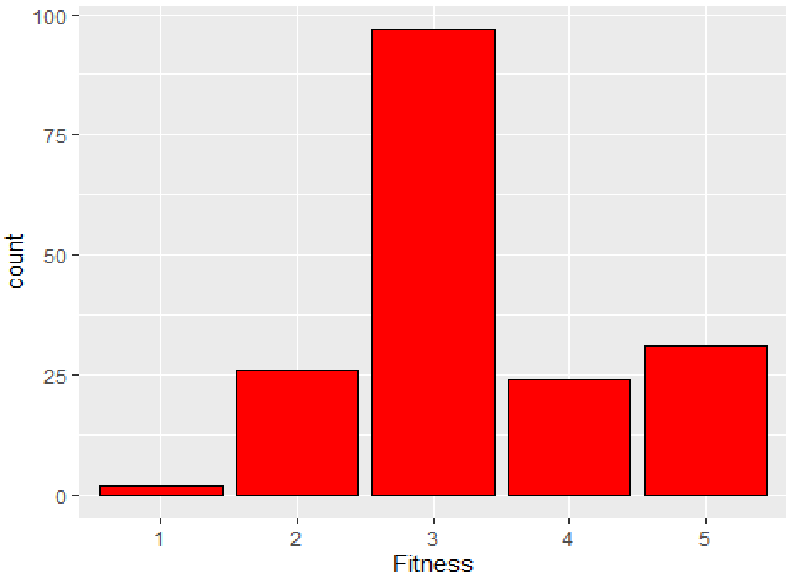
- The business has more male than female customers

```
### Bar graph for marital status
ggplot(cardio, aes(x = MaritalStatus)) +
  geom_bar(fill = c("green"), color="black")
```



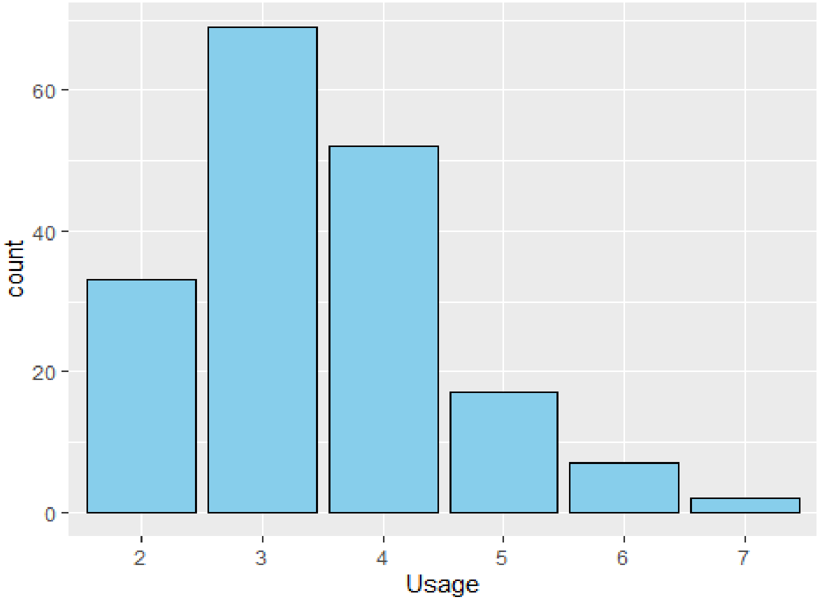
- A larger proportion of the customers have relationship partners
- The company can look into offering some couple centered services

```
### Bar graph for fitness
ggplot(cardio, aes(x = Fitness)) +
  geom_bar(fill = c("red"), color="black")
```



- On the scale of 1- vey unfit and 5 – very fit, a large proportion of the customers say they are fit.

```r
### Bar graph for usage
ggplot(cardio, aes(x = Usage)) +
  geom_bar(fill = c("sky blue"), color="black")
```

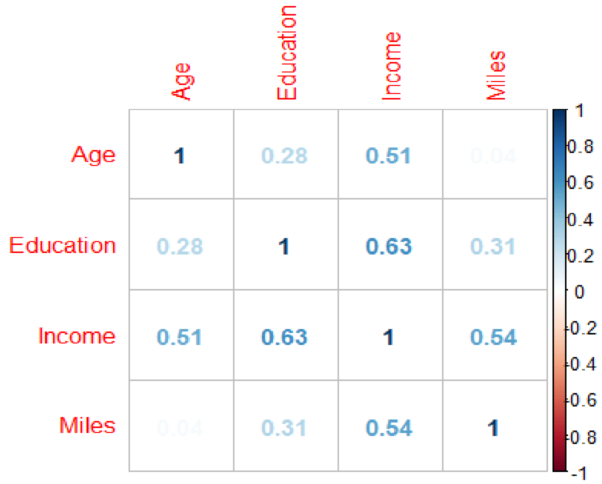


- Most of the customers plan to use the treadmill about 2 to 4 days a week

3.4 Bi-Variate Analysis

In this section, we plot bivariate charts between variables to understand their relationship with each other.

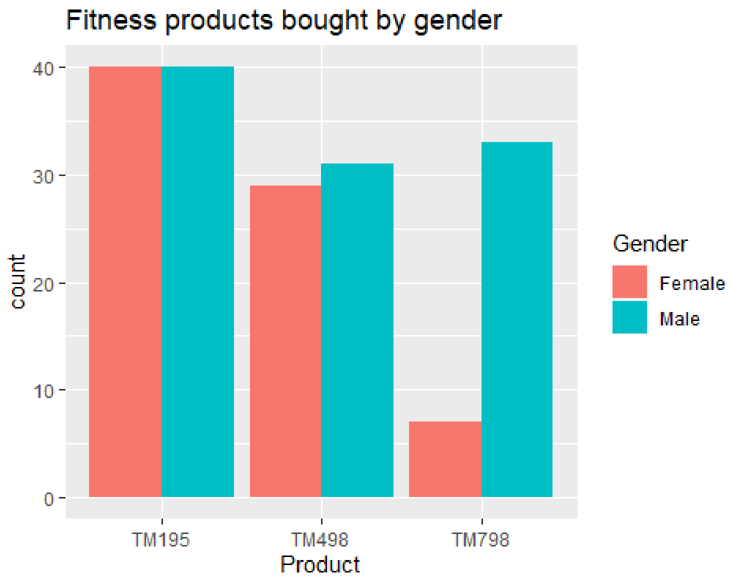
```r
## A. Correlation
## Check for correlation among numerical variables
num_vars = sapply(cardio, is.numeric) #Numeric variables in the cardio data
corrplot(cor(cardio[,num_vars]), method = 'number')
```



Observation from correlation data

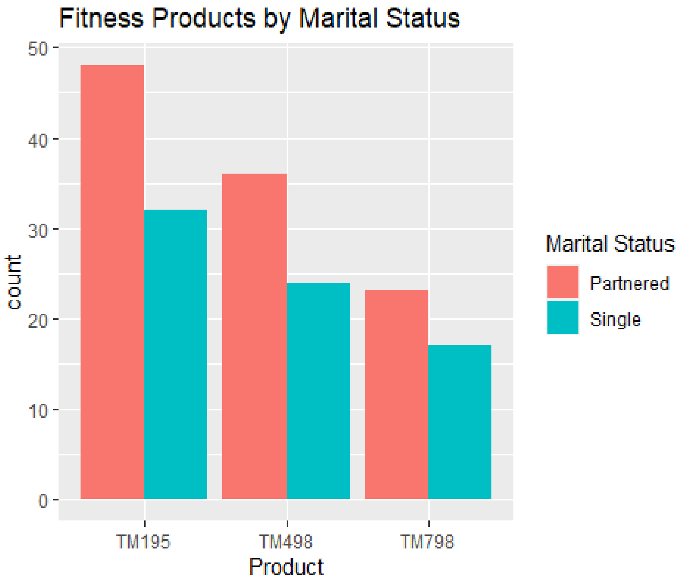
- As expected, income shows high correlation with education.
- All the numerical variables are positively correlated
- It is important to note that correlation does not imply causation.

```
## B. Bar graphs
## identify differences between customers of each product

## Gender and Product (Grouped Bar Graph)
ggplot(cardio, aes(x = Product, fill = Gender)) +
    labs(fill = "Gender", #Legend titles
       x = "Product",
       title = "Fitness products bought by gender") +
    geom_bar(position = position_dodge(preserve = "single"))
```



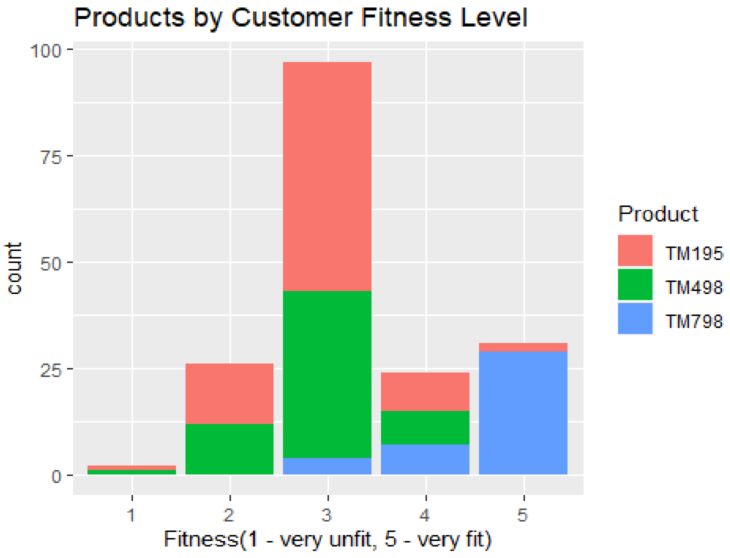
- Both gender prefer the TM195 and TM498 treadmill.
- A lot more men buy the TM798 treadmill than women. There is a high potential of making male customers the target population for TM798

```r
## Marital Status and product (Grouped Bar Chart)
ggplot(cardio, aes(x = Product, fill = MaritalStatus)) +
  labs(fill = "Marital Status", #Legend titles
       x = "Product",
       title = "Fitness Products by Marital Status") +

  geom_bar(position = position_dodge(preserve = "single"))
```



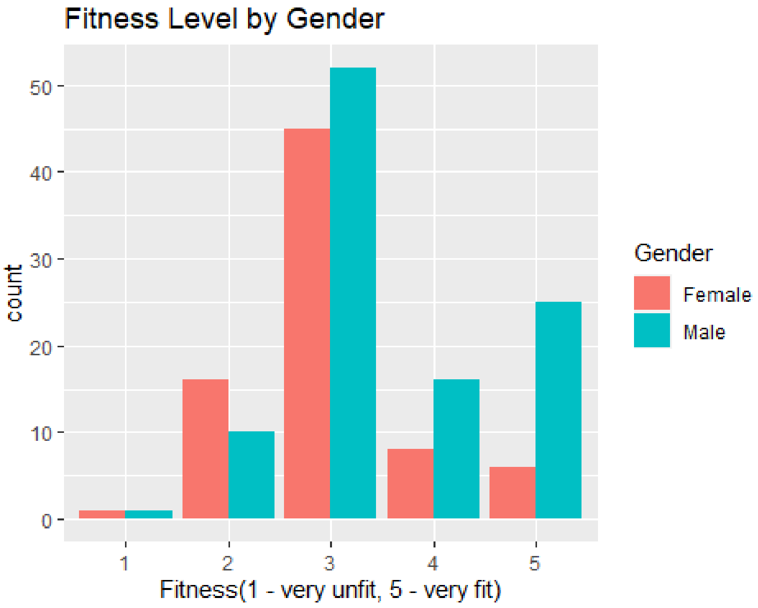
- This result is consistent with the high proportion of partnered customers patronizing the company.

```r
## Fitness Level and product (Stacked Bar Chart)
ggplot(cardio, aes(x = Fitness, fill = Product)) +
  labs(fill = "Product", #Legend titles
       x = "Fitness(1 - very unfit, 5 - very fit)",
       title = "Products by Customer Fitness Level") +
    geom_bar(position = "stack")
```



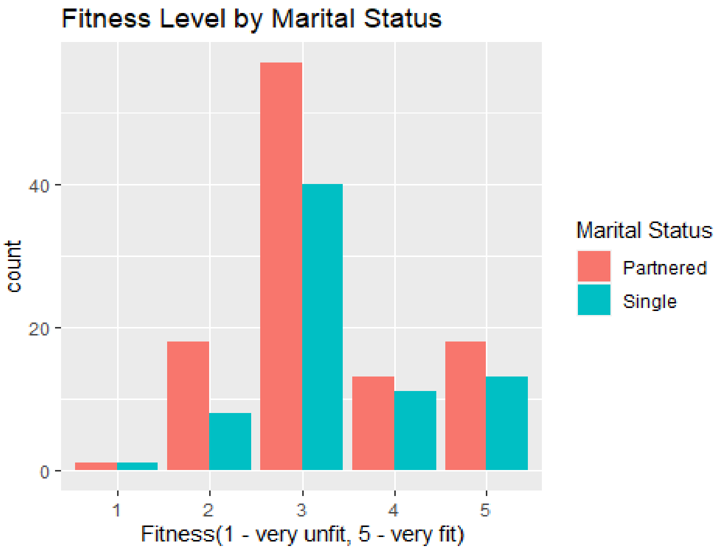
- Many customers that prefer TM798 treadmills consider themselves very fit.

```
## Fitness level and gender (Grouped Bar Chart)
ggplot(cardio, aes(x = Fitness, fill = Gender)) +
  labs(fill = "Gender", #Legend titles
       x = "Fitness(1 - very unfit, 5 - very fit)",
       title = "Fitness Level by Gender") +
  geom_bar(position = position_dodge(preserve = "single"))
```



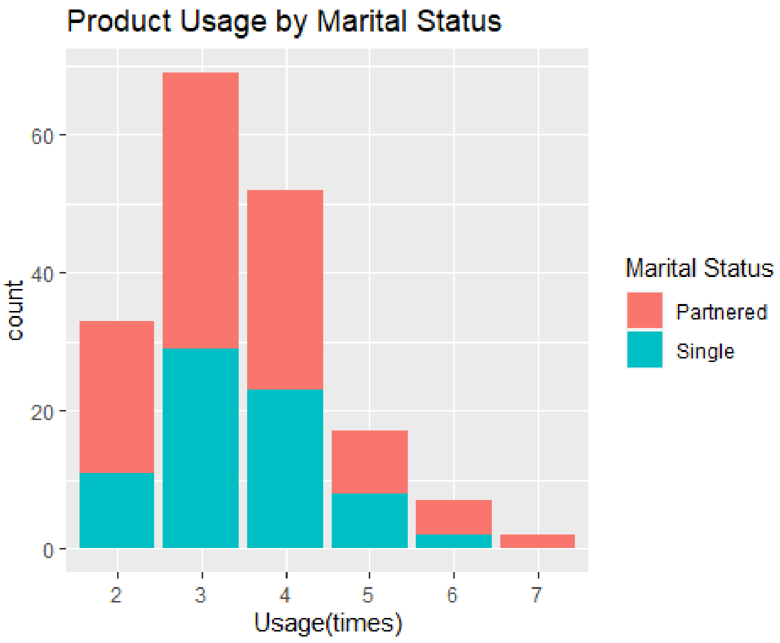
- The result indicates that men are more fit than women. This could also be as a result of the higher proportion of male customers.

```
## Fitness level and marital status (Grouped Bar Chart)
ggplot(cardio, aes(x = Fitness, fill = MaritalStatus)) +
  labs(fill = "Marital Status", #Legend titles
       x = "Fitness(1 - very unfit, 5 - very fit)",
       title = "Fitness Level by Marital Status") +
  geom_bar(position = position_dodge(preserve = "single"))
```



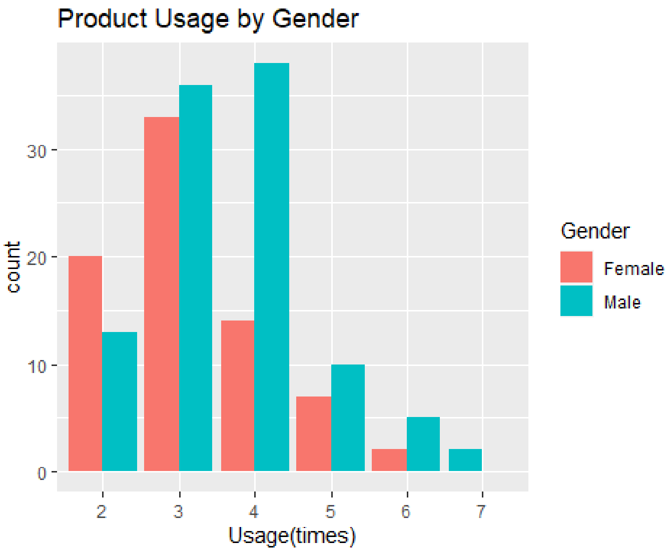
- Customers with partners consider themselves more fit than the singles
- This result is consisted with the large proportion of partnered customers than singles.

```
## Usage and marital status (Stacked Bar Chart)
ggplot(cardio, aes(x = Usage, fill = MaritalStatus)) +
  labs(fill = "Marital Status", #Legend titles
       x = "Usage(times)",
       title = "Product Usage by Marital Status") +
  geom_bar(position = "stack")
```



- Customers with partners are likely to use the treadmills more times than singles
- This result is consisted with the large proportion of partnered customers than singles.

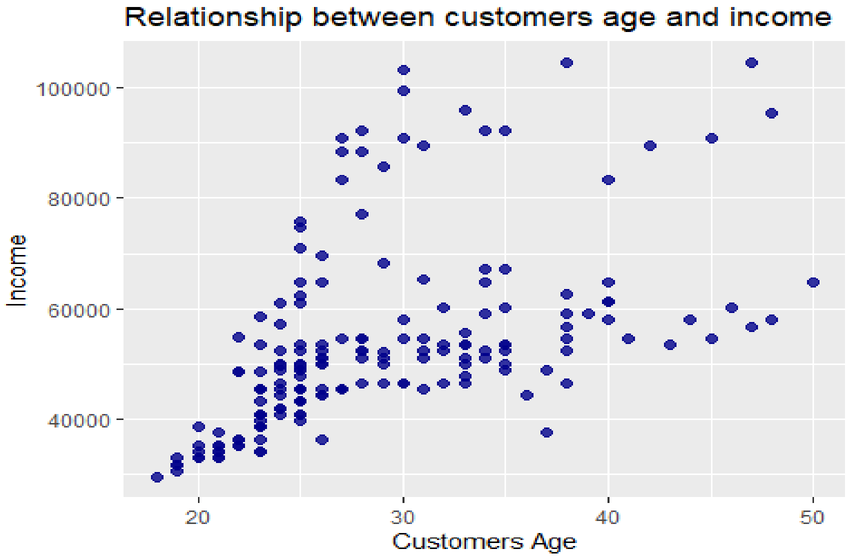
```r
## Usage and gender (Grouped Bar Chart)
ggplot(cardio, aes(x = Usage, fill = Gender)) +
  labs(fill = "Gender", #Legend titles
       x = "Usage(times)",
       title = "Product Usage by Gender") +
  geom_bar(position = position_dodge(preserve = "single"))
```



- The use of the treadmill 3 times a week seems to be a similar plan among both gender

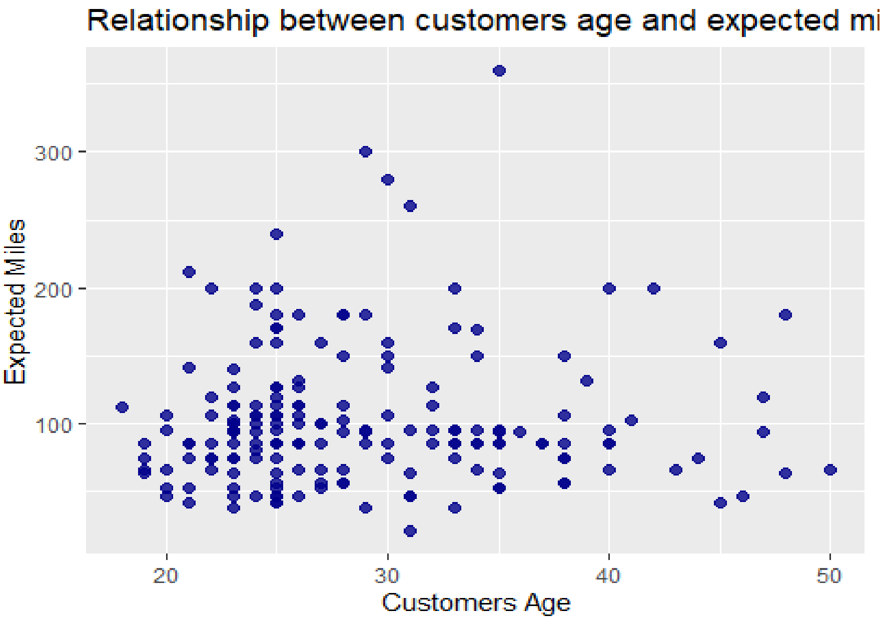
- However, a larger proportion of male customers plan to use the treadmills an average of 4 to 7 times a week.
- An interesting thing of note is that only male customers plan to use their treadmill every day of the week.

```
## C. Scatter Plots
## Check the relationship between the numerical variables

## Age and Income
ggplot(cardio,aes(x = Age,y = Income)) +
  geom_point(color="darkblue", size = 2, alpha=.8) +
  labs(x = "Customers Age",
       y = "Income",
       title = "Relationship between customers age and income")
```



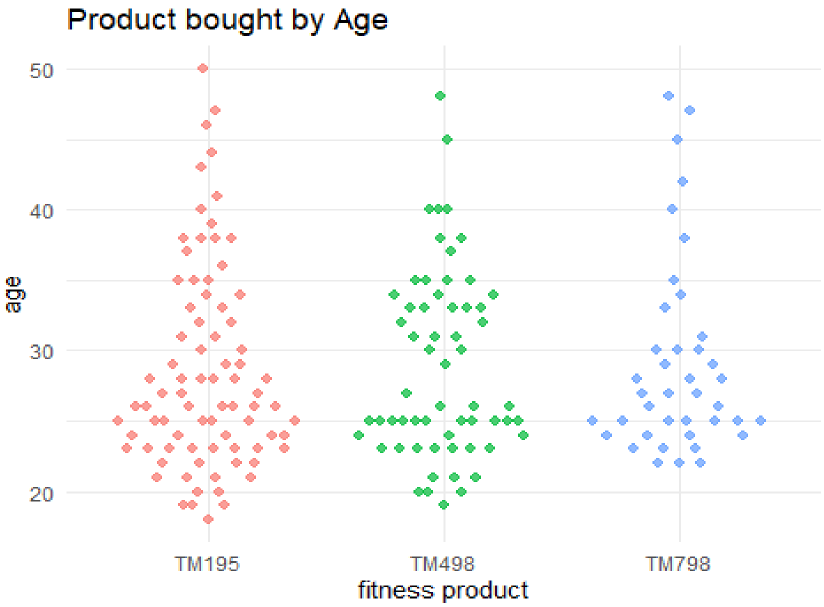
- A large percentage of the customer base earn between 40,000 to 70,000 in income
- Those who earn more than 80,000 range between 28 to 48 years' old
- Majority of the customers in the company are clustered within 23 to 35 years of age.

```
## Miles and Age
ggplot(cardio,aes(x = Age,y = Miles)) +
  geom_point(color="darkblue", size = 2, alpha=.8) +
  labs(x = "Customers Age",
       y = "Expected Miles",
       title = "Relationship between customers age and expected miles")
```



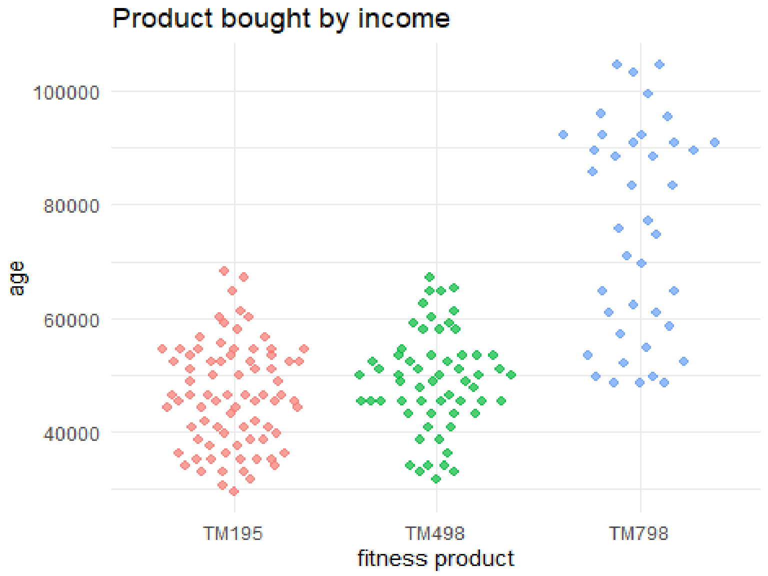
- Both young and old customers are expected to run between 50 to 100 miles after purchasing the treadmill

```
## D. Beeswarm Plots
## Check the relationship between the categorical and numerical variables

## Age and Fitness Product
library(ggbeeswarm)
library(scales)
ggplot(cardio, aes(x = Product,
                   color = Product,
                   y = Age)) +
  geom_quasirandom(alpha = 0.7,
                   size = 1.9) +
  labs(title = "Product bought by Age",
       x = "fitness product",
       y = "age") +
  theme_minimal() +
  theme(legend.position = "none")
```



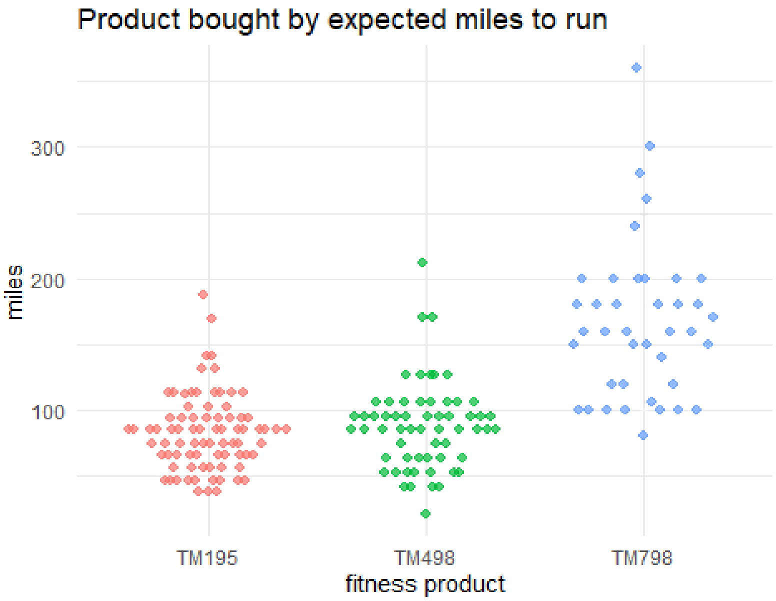
- The customers that prefer TM195 treadmills are much more spread out i.e. both young old prefer this brand.
- More customers within the age of 20 to 35 years purchase TM498 treadmills.
- More customers within the age of 23 to 30 years purchase TM798 treadmills

```
## Income and fitness product
library(ggbeeswarm)
library(scales)
ggplot(cardio, aes(x = Product,
                   color = Product,
                   y = Income)) +
  geom_quasirandom(alpha = 0.7,
                   size = 1.9) +
  labs(title = "Product bought by income",
       x = "fitness product",
       y = "age") +
  theme_minimal() +
  theme(legend.position = "none")
```



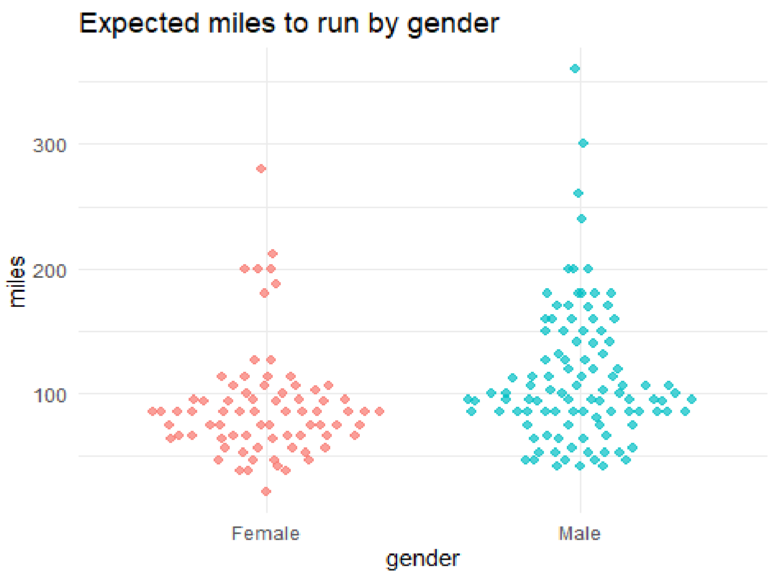
- The customers who earn between 29,000 to 60,000 prefer TM195 and TM498 treadmills
- We can also say from the results that TM195 treadmills are more affordable than the rest.
- TM798 treadmills are the most expensive of the products considered in this analysis.

```r
## Expected miles to run and fitness product
library(ggbeeswarm)
library(scales)
ggplot(cardio, aes(x = Product,
                   color = Product,
                   y = Miles)) +
  geom_quasirandom(alpha = 0.7,
                   size = 1.9) +
  labs(title = "Product bought by expected miles to run",
       x = "fitness product",
       y = "miles") +
  theme_minimal() +
  theme(legend.position = "none")
```



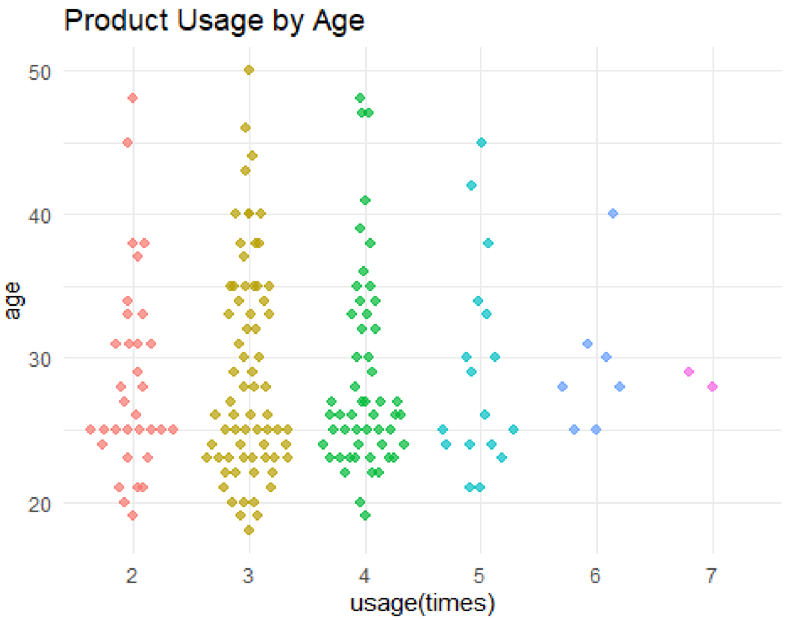
- Customers that purchase TM195 and TM498 are expected to run between 50 to 100 miles
- Based on the data, we can conclude that those who buy TM798 treadmills exercise more than those who buy TM195 and TM498.

```
##Expected miles to run and gender
library(ggbeeswarm)
library(scales)
ggplot(cardio, aes(x = Gender,
                   color = Gender,
                   y = Miles)) +
  geom_quasirandom(alpha = 0.7,
                   size = 1.9) +
  labs(title = "Expected miles to run by gender",
       x = "gender",
       y = "miles") +
  theme_minimal() +
  theme(legend.position = "none")
```



- More female customers are expected to run less than 160 miles
- Male customers are expected to run 50 miles more than females

```r
## Age and Usage
library(ggbeeswarm)
library(scales)
ggplot(cardio, aes(x = Usage,
                   color = Usage,
                   y = Age)) +
  geom_quasirandom(alpha = 0.7,
                   size = 1.9) +
  labs(title = "Product Usage by Age",
       x = "usage(times)",
       y = "age") +
  theme_minimal() +
  theme(legend.position = "none")
```



- Those who plan to use the products within 6 to 7 times a week are within 25 to 30 years old
- Older people from 35 and above plan to use the treadmills 2 to 4 times a week

4. Conclusion and Recommendation

4.1 Conclusion

We analyzed a dataset of 180 customers that patronize Cardio Good Fitness online shop. The main feature of interest in the data is treadmills bought from the company. From a personal and health perspective, regular aerobic exercise, such as a treadmill workout regimen, improves blood circulation in the body and helps to lower blood pressure by strengthening the heart. This makes treadmills a viable product of interest to those interested in fitness. Thus we identified some of the characteristics of present customers in other to attract more.

We have been able to conclude that

1. TM195 treadmills are more popular among people of all age groups and gender
2. In line with intuition, people with partners are more likely to be interested in buying treadmills
3. More male customers prefer the T798 treadmill
4. The TM195 treadmill is much more affordable and should be offered to people of all ages and income bracket
5. TM798 treadmills should be considered as premium treadmills based on the customer base

4.2 Recommendation to Business

- As expected, people with partners are more likely to focus on their fitness than single people. This is in line with the intuitive understanding of people wanting to give off a better image when in relationships. So, the company should consider couple centered marketing.
- The TM195 treadmill is much more in demand by both male and female. It is important to identify the reasons for the high demand in case the factors discovered could be considered for other treadmill products.

5. Appendix: Source Code