Explorative Data Analysis (EDA)

Hamid Abdulsalam

Explorative Data Analysis (EDA)

Explorative Data Analysis (EDA)

Exploratory data analysis (EDA) is used by data scientists to analyze, investigate and summarize the main characteristics of data sets, often employing data visualization methods.

EDA is often used to provide a better understanding of data set variables and the relationships between the the variables. It makes it easier to discover patterns in the data set.

Tidyverse

Tidyverse is a collection of essential R packages for data science created by Hadley Wickham.

The following packages are included in the core tidyverse: ggplot2, dplyr, tidyr, readr, purrr , tibble, stringr and forcats.

Data

The **Gym** data contains the information of customers of the treadmill products of a retail store. It's Variables are:

- Product — the model number of the treadmill
- Age — in number of years, of the customer
- Gender — of the customer
- Education — in number of years, of the customer
- Marital Status — of the customer
- Usage — Average number of times the customer wants to use the treadmill every week
- Fitness — Self rated fitness score of the customer (5 — very fit, 1 — very unfit)
- Income — of the customer
- Miles- expected to run

Loading the Data

```
library(tidyverse)
gy<-read.csv("data/gym.csv", header = T,)
glimpse(gy)
```

```
## Rows: 180
## Columns: 9
## $ Product       <chr> "TM195", "TM195", "TM195", "TM195"
## $ Age           <int> 18, 19, 19, 19, 20, 20, 21, 21, 21
## $ Gender        <chr> "Male", "Male", "Female", "Male",
## $ Education     <int> 14, 15, 14, 12, 13, 14, 14, 13, 15
## $ MaritalStatus <chr> "Single", "Single", "Partnered", "
## $ Usage         <int> 3, 2, 4, 3, 4, 3, 3, 3, 5, 2, 3, 3
## $ Fitness       <int> 4, 3, 3, 3, 2, 3, 3, 3, 4, 3, 3, 2
## $ Income        <int> 29562, 31836, 30699, 32973, 35247,
## $ Miles         <int> 112, 75, 66, 85, 47, 66, 75, 85, 1
```

Univariate Analysis of Categorical

The variables Product, Gender and Marital Status are characters, hence they need to be converted to factors (Categorical variable in R).

```
gy$Product<-as.factor(gy$Product)
gy$Gender<-as.factor(gy$Gender)
gy$MaritalStatus<-as.factor(gy$MaritalStatus)
```

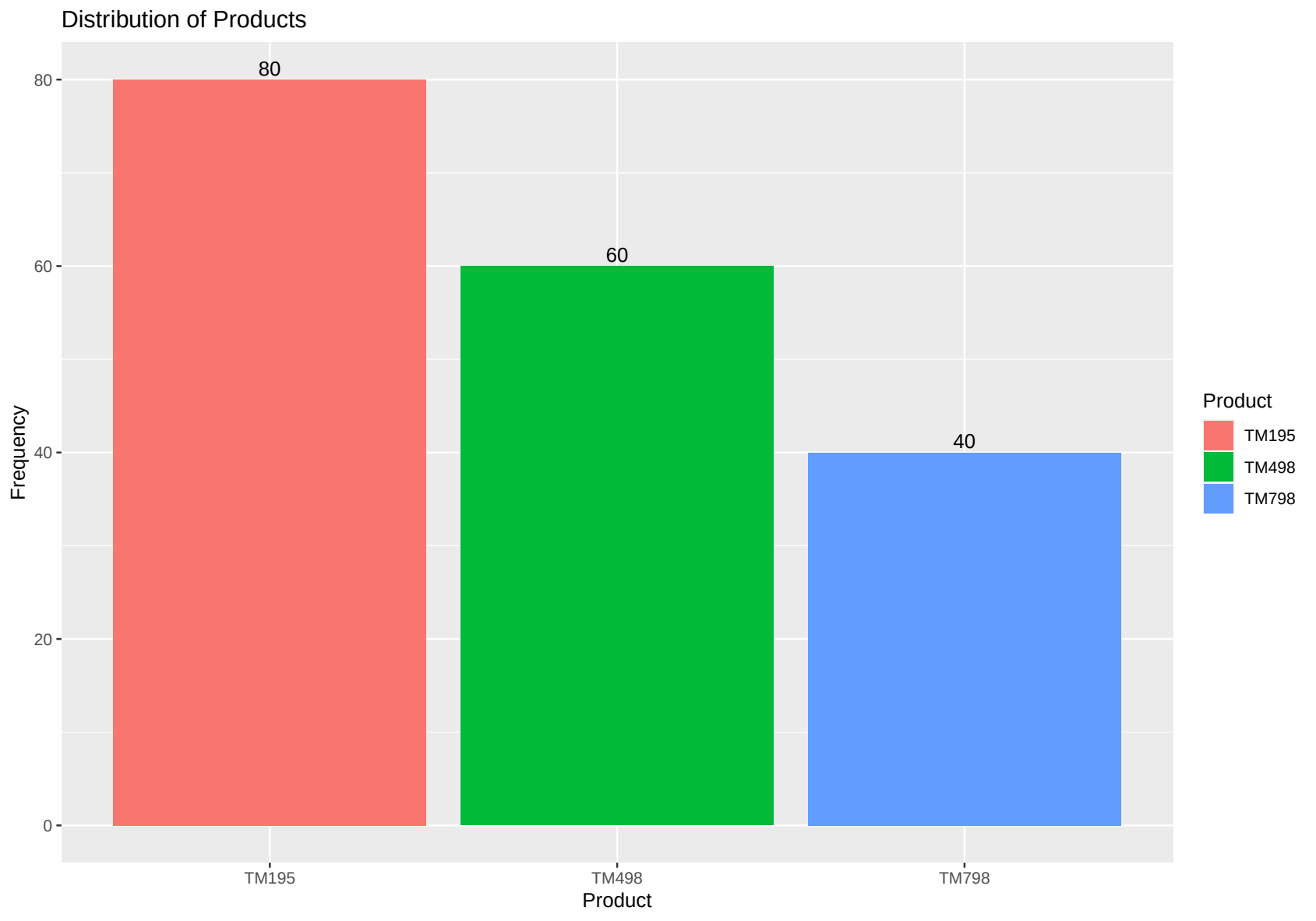
Univariate Analysis

Univariate analysis is defined as analysis carried out on only one ("uni") variable ("variate") to summarize or describe the variable. Let's carry out univariate analysis to understand the overall profiles of customers who purchased at least one treadmill.

Univariate Analysis

```
b1<-ggplot(gy,aes(x = Product, fill=Product))+geom_bar()+
labs(x="Product",y="Frequency",
title="Distribution of Products")+
geom_text(stat='count',aes(label=..count..),vjust=-0.3)
```

Univariate Analysis of Product



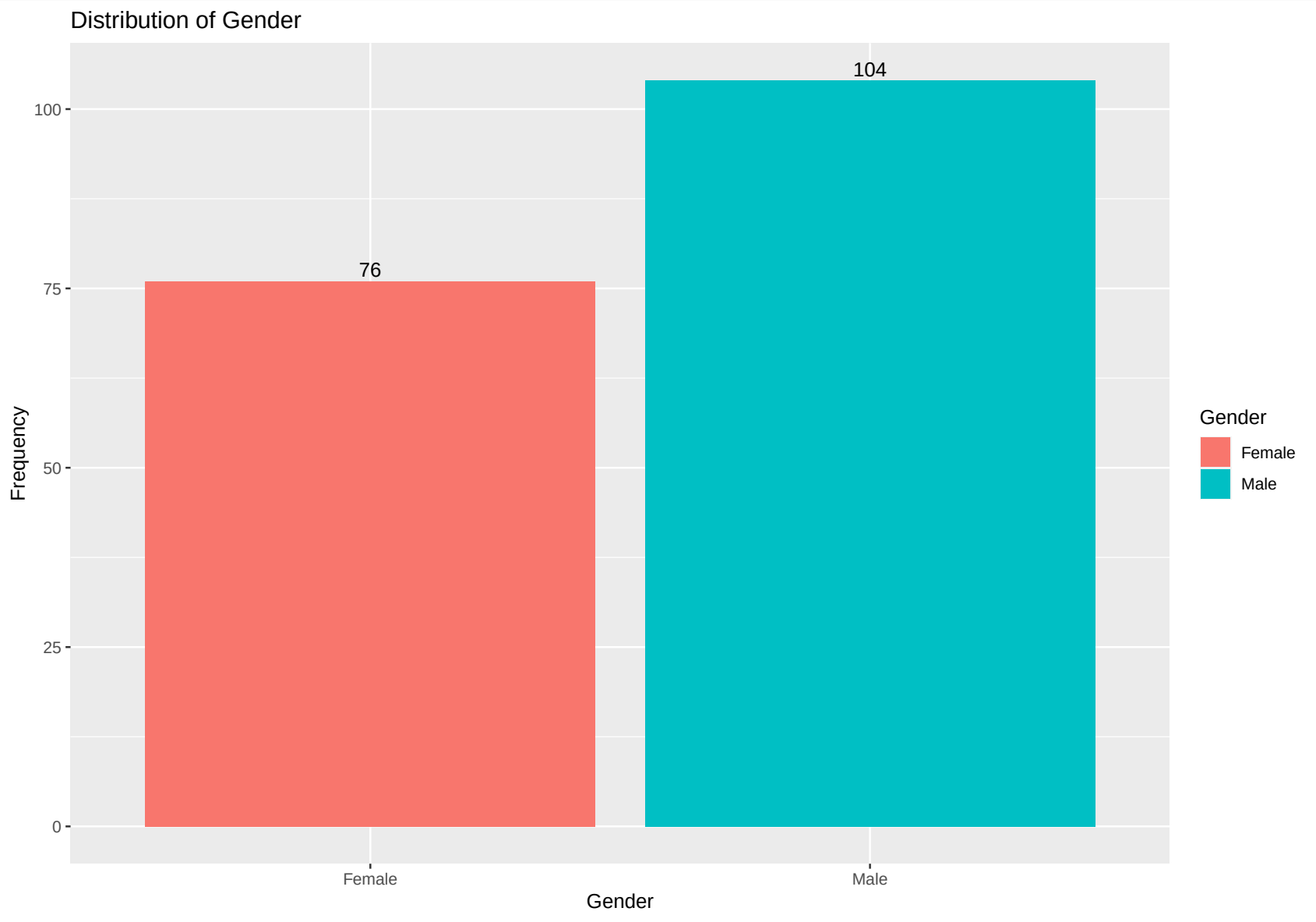
Univariate Analysis of Product

From the Bar Plot of the Product, we can see that TM195 appears to be the most purchased treadmill product with 80 purchases, followed by TM498 with 60 purchases and TM798 is the least popular product with 40 purchases.

Univariate Analysis of Gender

```
b2<-ggplot(gy,aes(x = Gender, fill=Gender))+geom_bar()+
labs(x="Gender",y="Frequency",
title="Distribution of Gender")+
geom_text(stat='count',aes(label=..count..),vjust=-0.3)
```

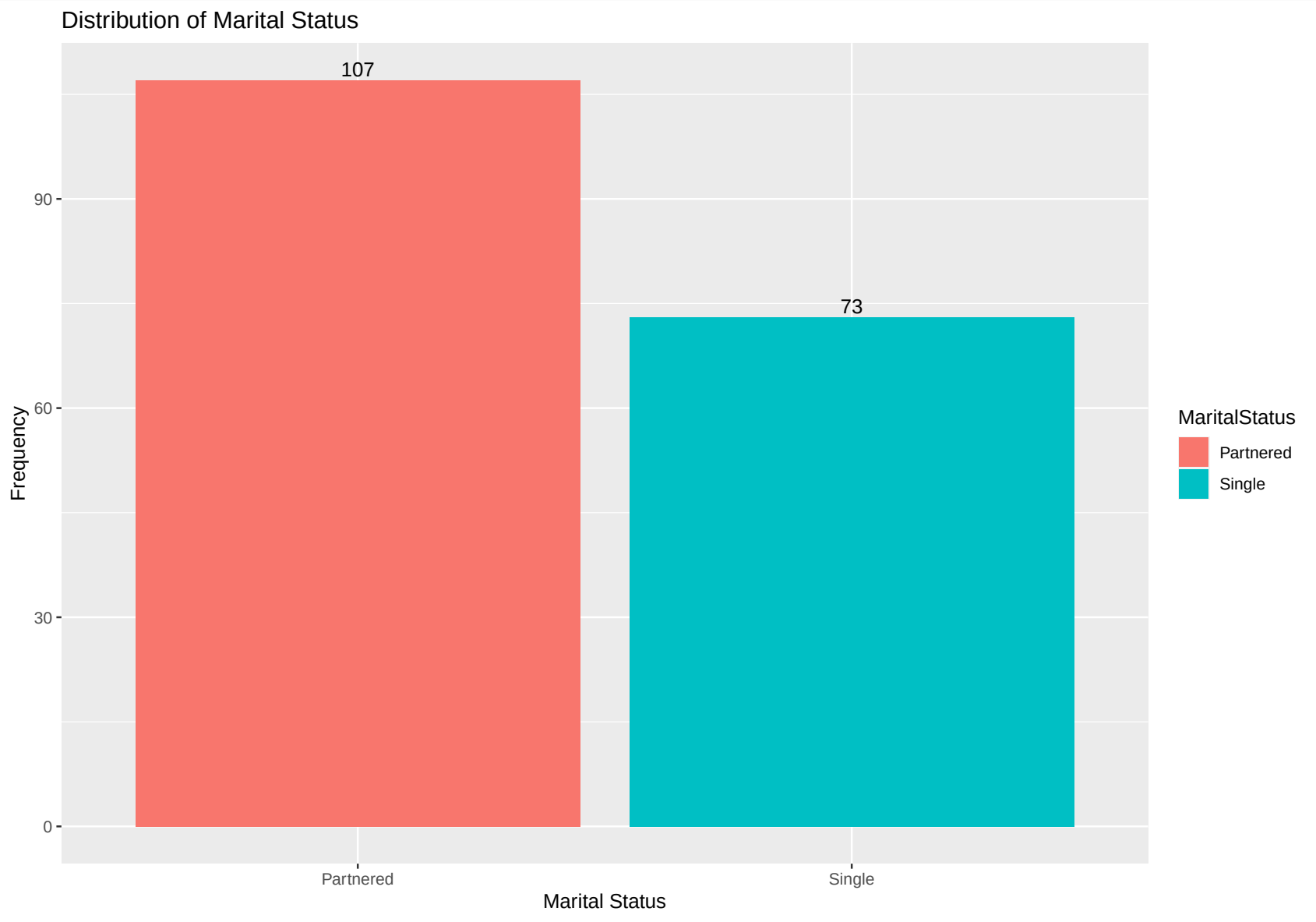
Univariate Analysis of Gender



Univariate Analysis of Marital Status

```
b3<-ggplot(gy,aes(x = MaritalStatus, fill=MaritalStatus))+
geom_bar()+labs(x="Marital Status",y="Frequency",
title="Distribution of Marital Status")+
geom_text(stat='count',aes(label=..count..),vjust=-0.3)
```

Univariate Analysis of Marital Status



Univariate Analysis of Gender and Marital Status

The Distribution of Gender shows that the Treadmills are more popular among males than the females. 76 Females bought a treadmill compared to 104 Males.

The Bar Plot of Marital Status shows that the Treadmill are more popular among partnered customers than single customers. 107 partnered customers bought a treadmill as compared to 73 single customers.

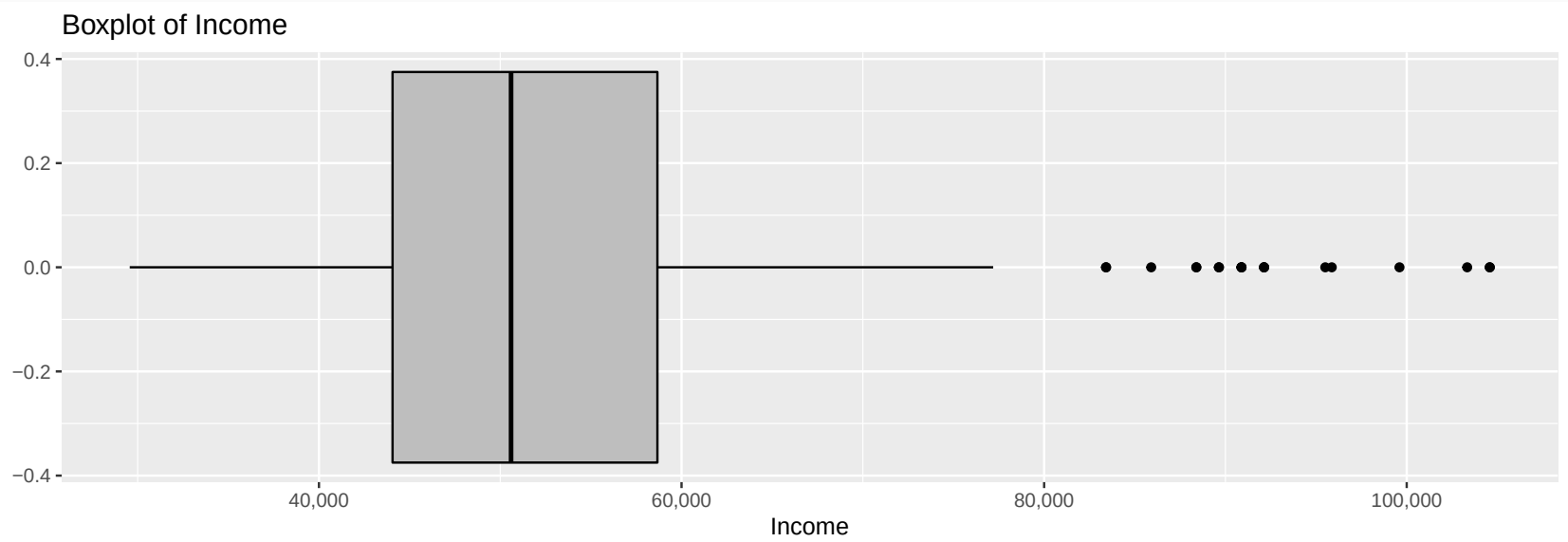
Univariate Analysis of Numeric Variables.

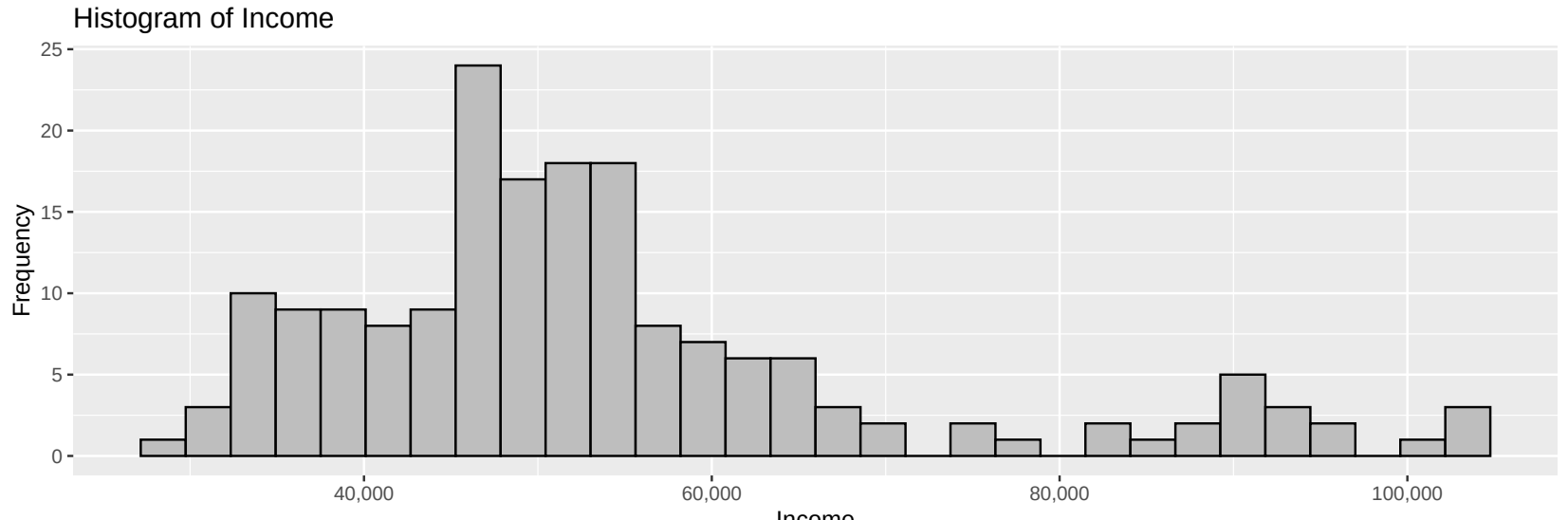
Let's Examine the distribution of Income, Age, Usage, Fitness, Miles and Education

```
library(gridExtra)
library(scales)
p1<-ggplot(gy, aes(x = Income)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Income",
x="Income",y="Frequency")

p2<- ggplot(gy, aes(x = Income)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Income",x="Income")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p2, p1, heights = c(1, 1))
```





Univariate Analysis of Numeric Variables.

The Histogram of Income is Skewed to the right, majority of the buyers' income range between approximately $30,000 and $80,000.

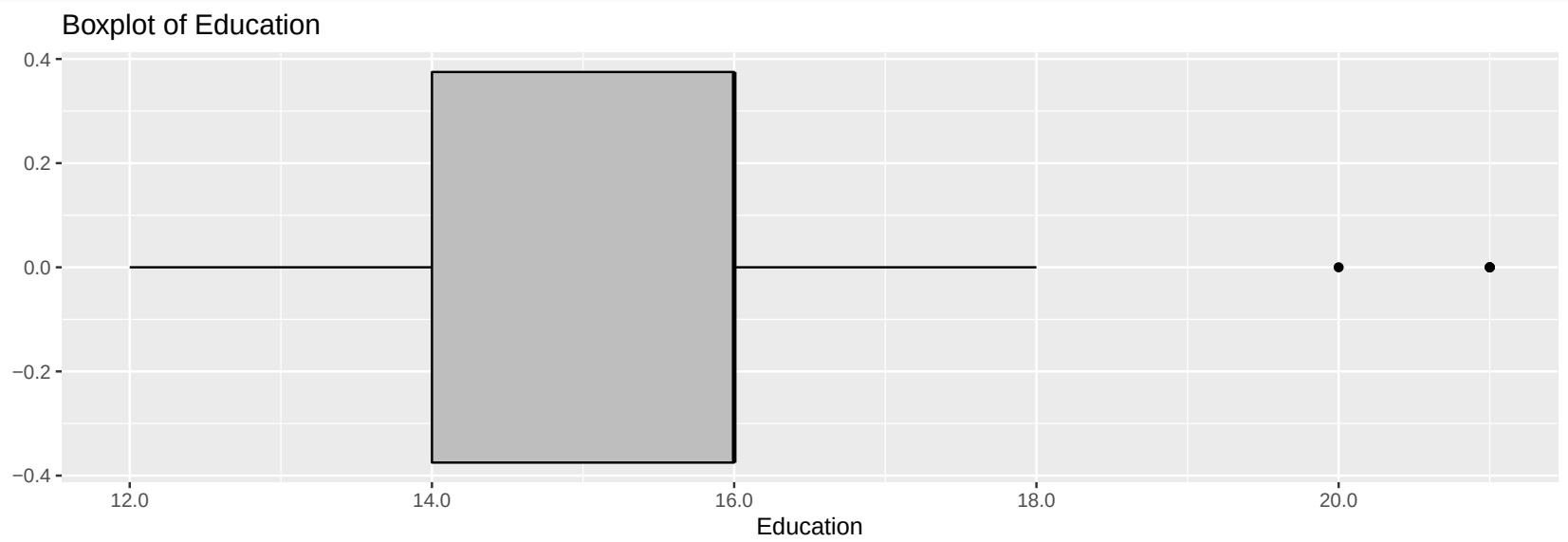
The middle 50% falls between $44000 and $58000 as shown in the Income boxplot

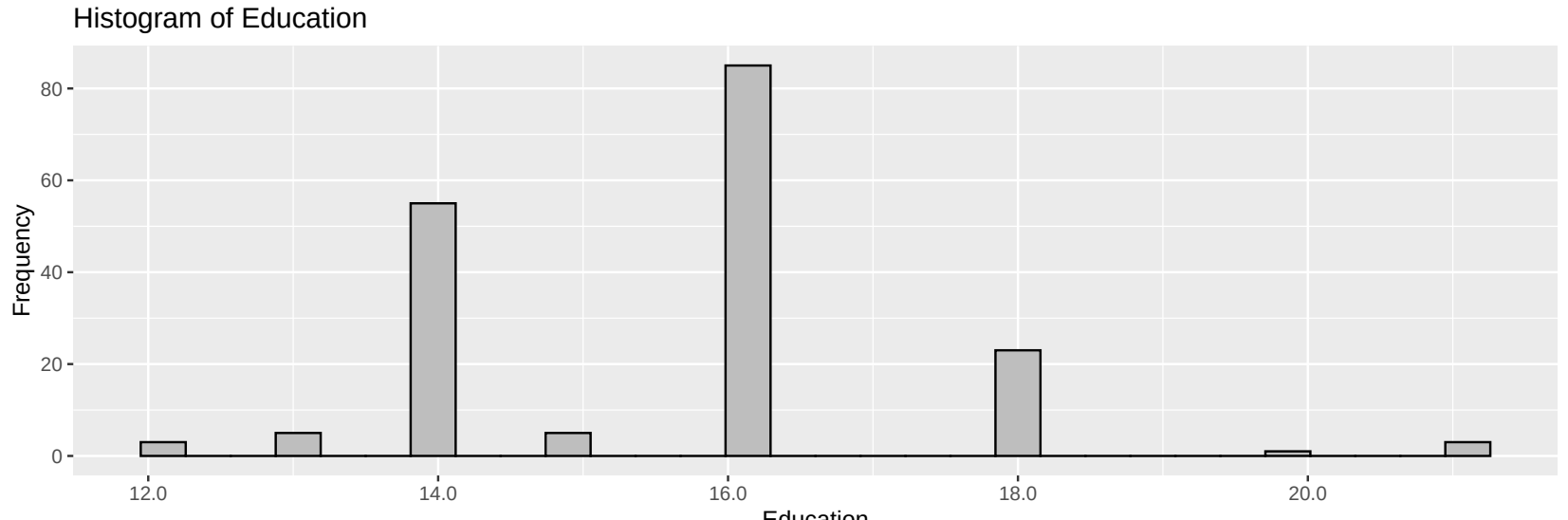
Univariate Analysis of Numeric Variables.

```
p3<-ggplot(gy, aes(x = Education)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Education",
x="Education",y="Frequency")

p4<- ggplot(gy, aes(x = Education)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Education",x="Education")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p4, p3, heights = c(1, 1))
```





Univariate Analysis of Numeric Variables.

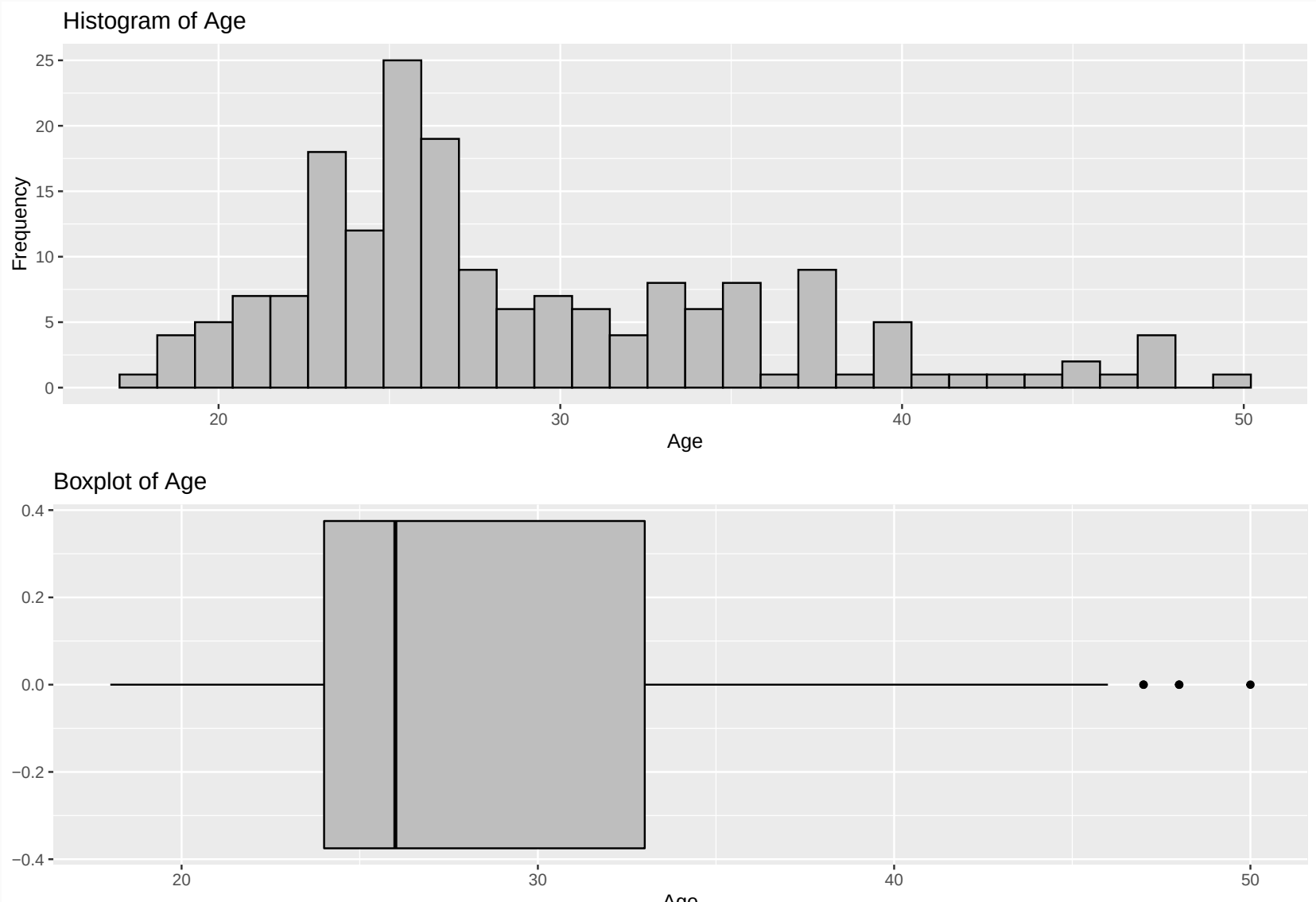
The boxplot for Education shows that the middle 50% of buyers have 14 to 16 years of education and the upper 25% of the buyers have 16 to 18 years of education.

Univariate Analysis of Numeric Variables.

```
p5<-ggplot(gy, aes(x = Age)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Age",
x="Age",y="Frequency")

p6<- ggplot(gy, aes(x = Age)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Age",x="Age")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p5, p6, heights = c(1, 1))
```



Univariate Analysis of Numeric Variables.

The Age boxplot indicates that the middle 50% of the buyers are between 24 and 33 years old. We see that 25% of all buyers fall between 24 and 26 years old

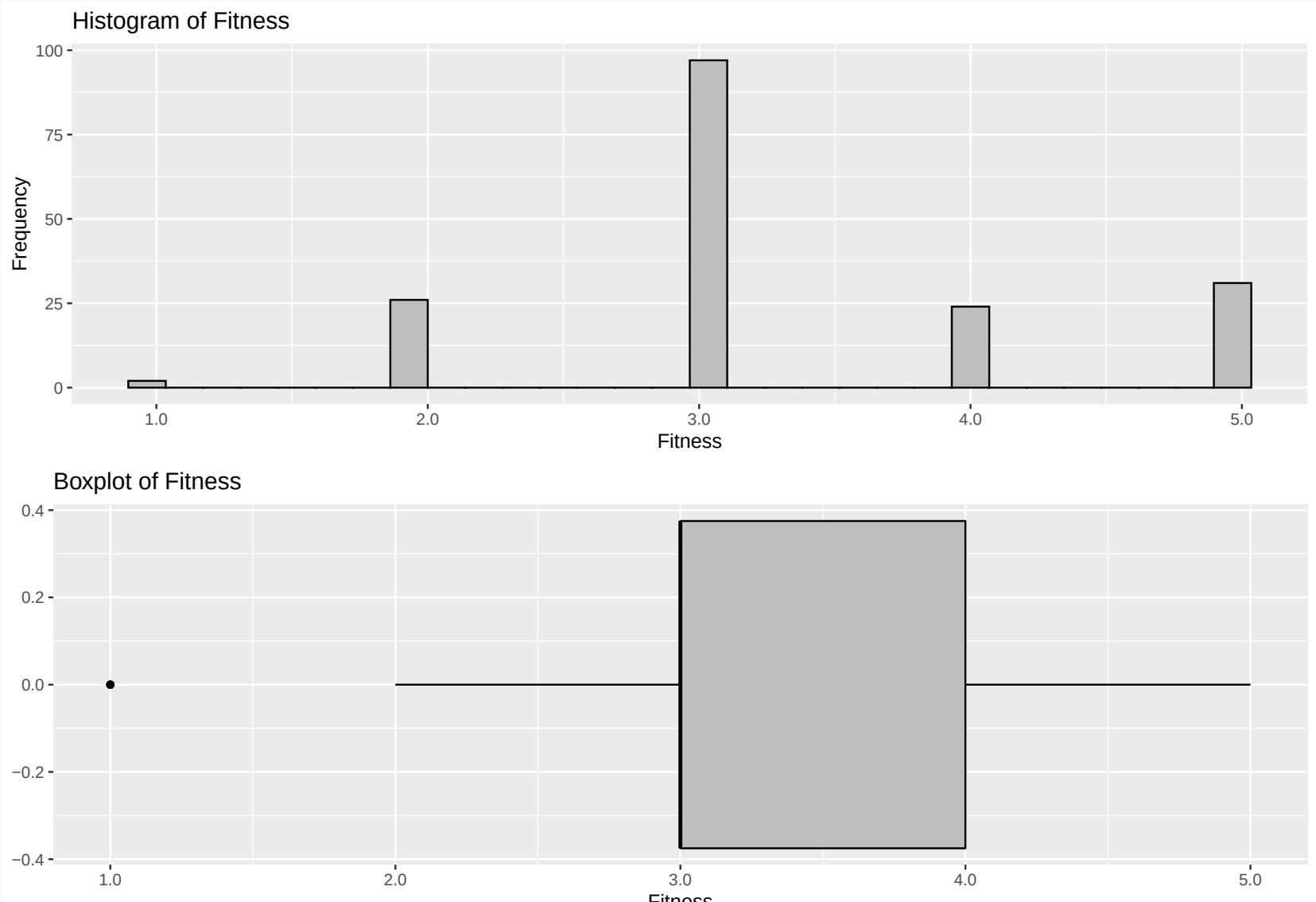
As seen in the Histogram of the Age , the distribution of Age is Skewed to right.

Univariate Analysis of Numeric Variables.

```
p7<-ggplot(gy, aes(x = Fitness)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Fitness",
x="Fitness",y="Frequency")

p8<- ggplot(gy, aes(x = Fitness)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Fitness",x="Fitness")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p7, p8, heights = c(1, 1))
```



Univariate Analysis of Numeric Variables.

The Fitness Histogram shows that three is the mode for Fitness, it also shows that Fitness is skewed to the left.

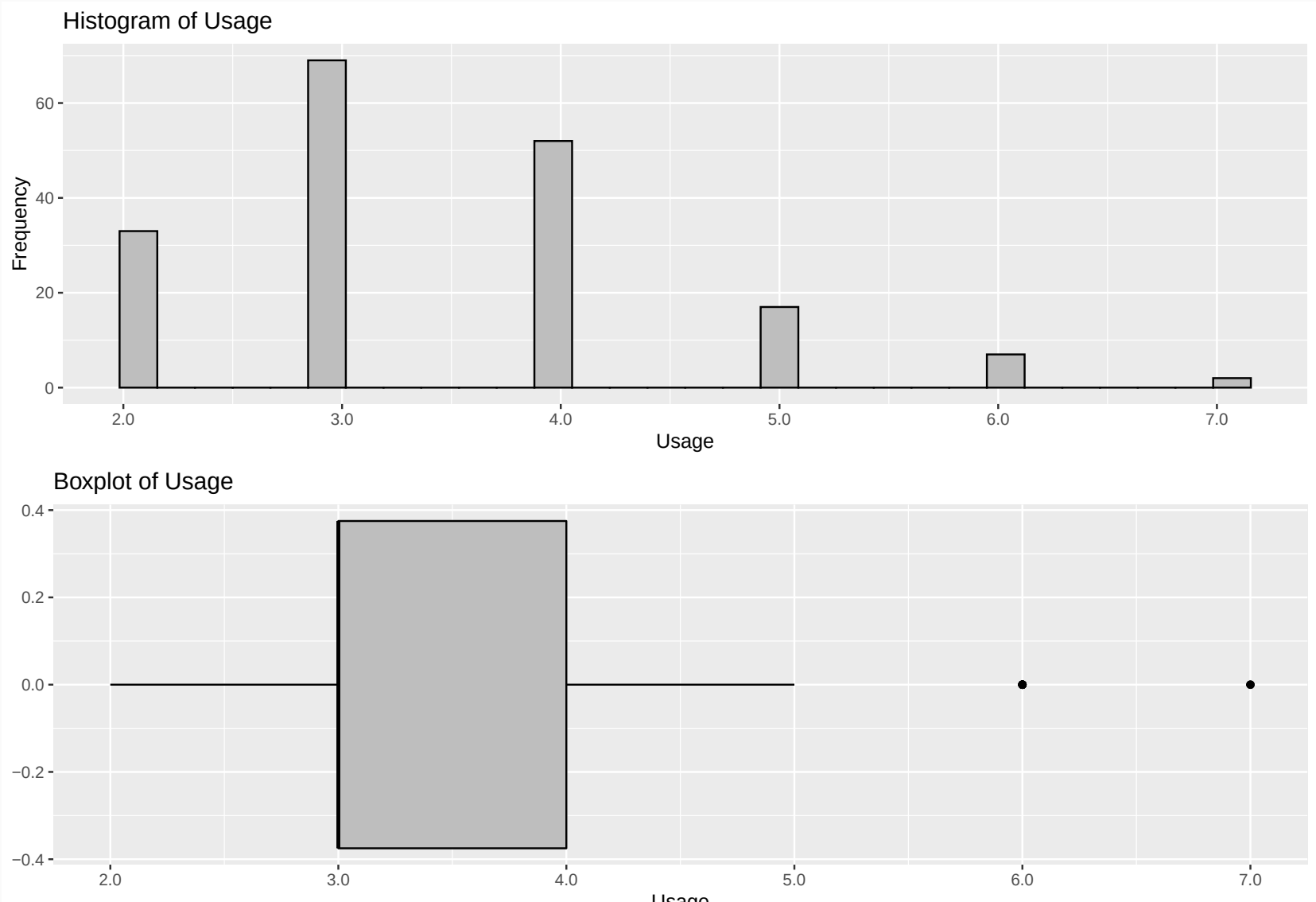
Middle 50% of the buyers rated their fitness levels between 3 and 4, as seen in the Fitness boxplot.

Univariate Analysis of Numeric Variables.

```
p9<-ggplot(gy, aes(x = Usage)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Usage",
x="Usage",y="Frequency")

p10<- ggplot(gy, aes(x = Usage)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Usage", x="Usage")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p9, p10, heights = c(1, 1))
```



Univariate Analysis of Numeric Variables.

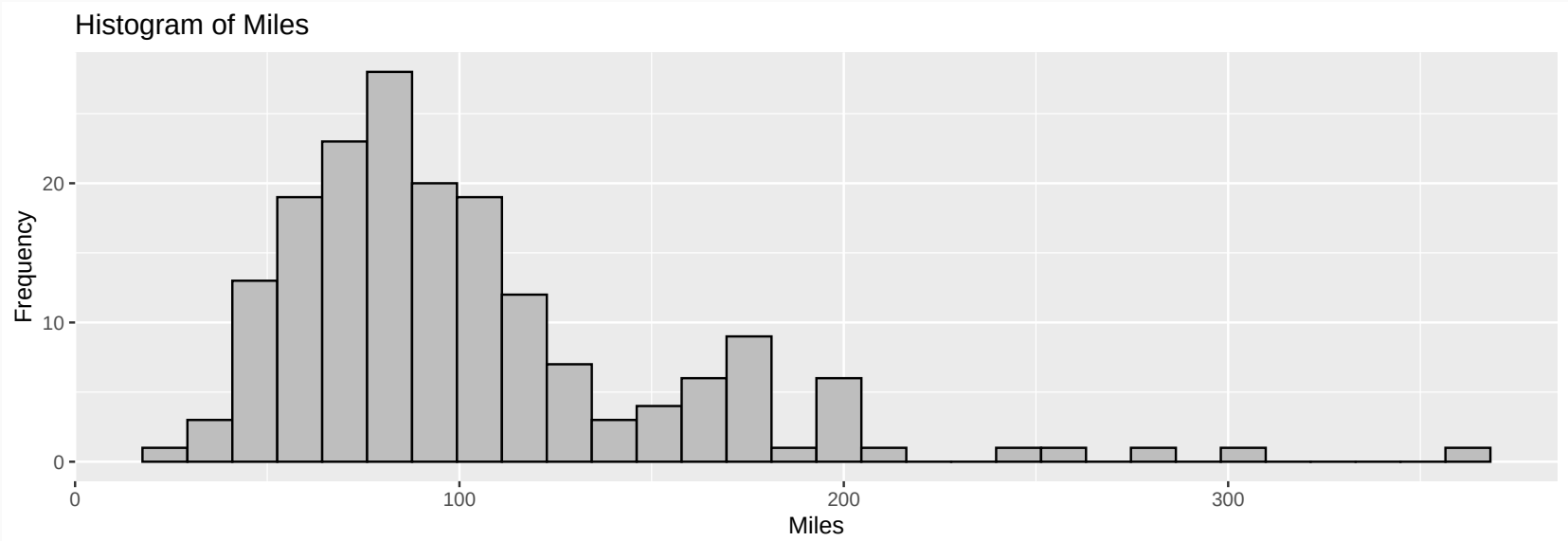
The Histogram of Usage is skewed to the right. Middle 50% of the buyers plan to use the treadmills between 3 to 4 times a week, as shown in the Usage boxplot, with three times a week as the mode shown in the Usage histogram.

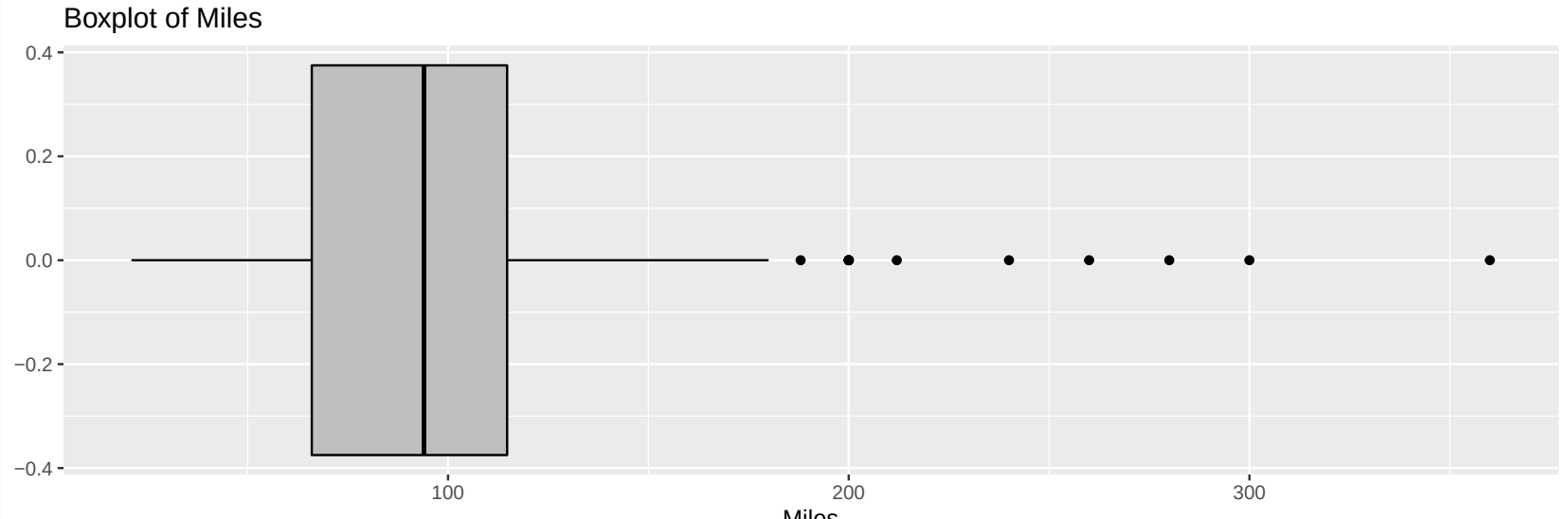
Univariate Analysis of Numeric Variables.

```
p11<-ggplot(gy, aes(x = Miles)) +
geom_histogram(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Histogram of Miles",
x="Miles",y="Frequency")

p12<- ggplot(gy, aes(x = Miles)) +
geom_boxplot(color = "black", fill = "gray")+
scale_x_continuous(labels = comma)+
labs(title = "Boxplot of Miles",x="Miles")
```

Univariate Analysis of Numeric Variables.

```
grid.arrange(p11, p12, heights = c(1, 1))
```





Univariate Analysis of Numeric Variables.

The Miles Histogram is skewed to the right. The boxplot suggests that the middle 50% of them expect to run between 66 miles and 116.5 miles

Bivariate Analysis

Let's conduct further analyses to explore how customer profiles differ across treadmill products TM195, TM498, and TM798.

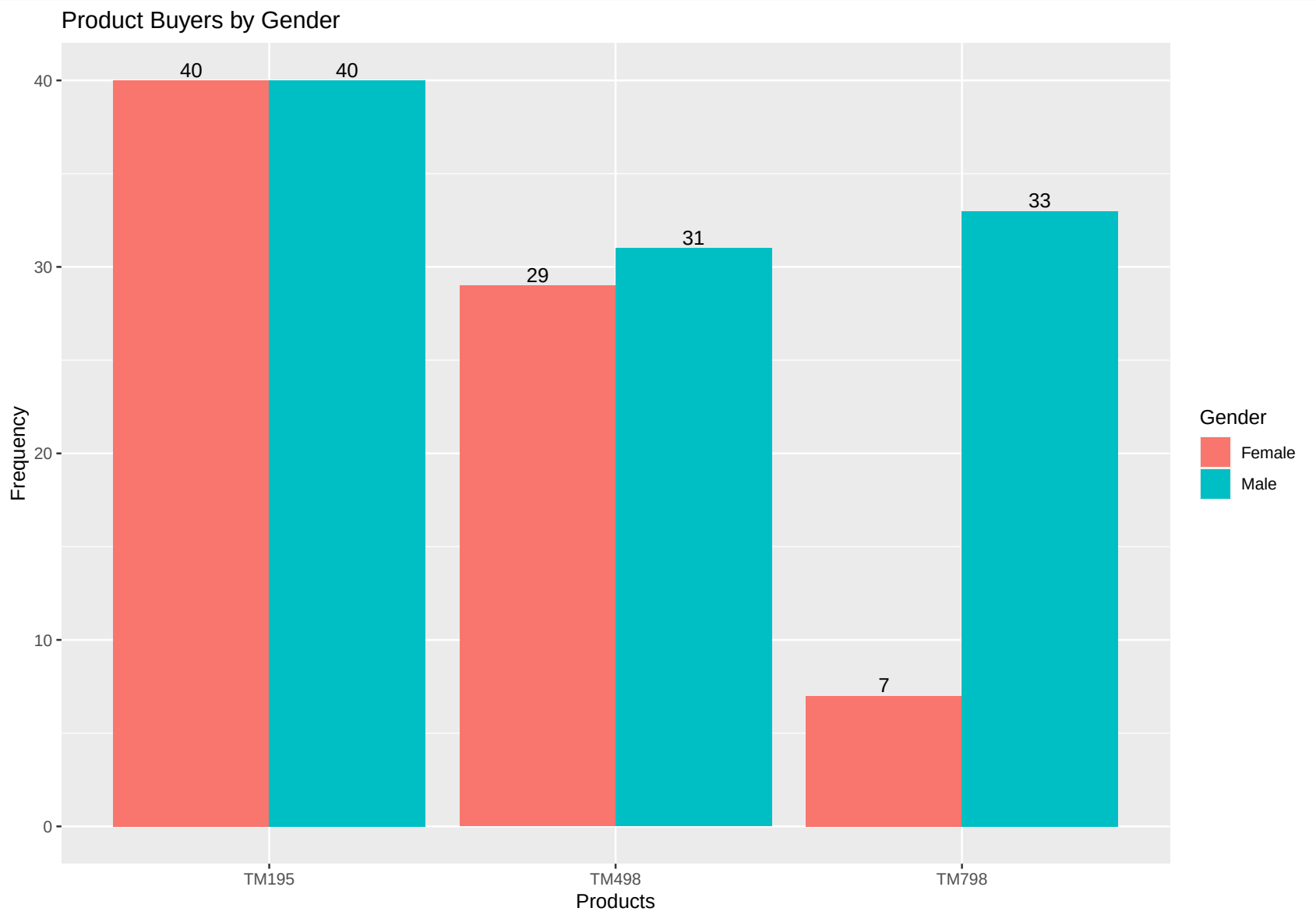
Bivariate Analysis of Product and Gender

```
b4<-ggplot(gy,aes(x =Product, fill=Gender))+
geom_bar(position = "dodge")+
labs(x="Products",y="Frequency",
  title="Product Buyers by Gender")

b4_1<-b4+geom_text(stat='count',aes(label=..count..)
                   ,position=position_dodge(width=0.9),
                   vjust=-0.25)
```

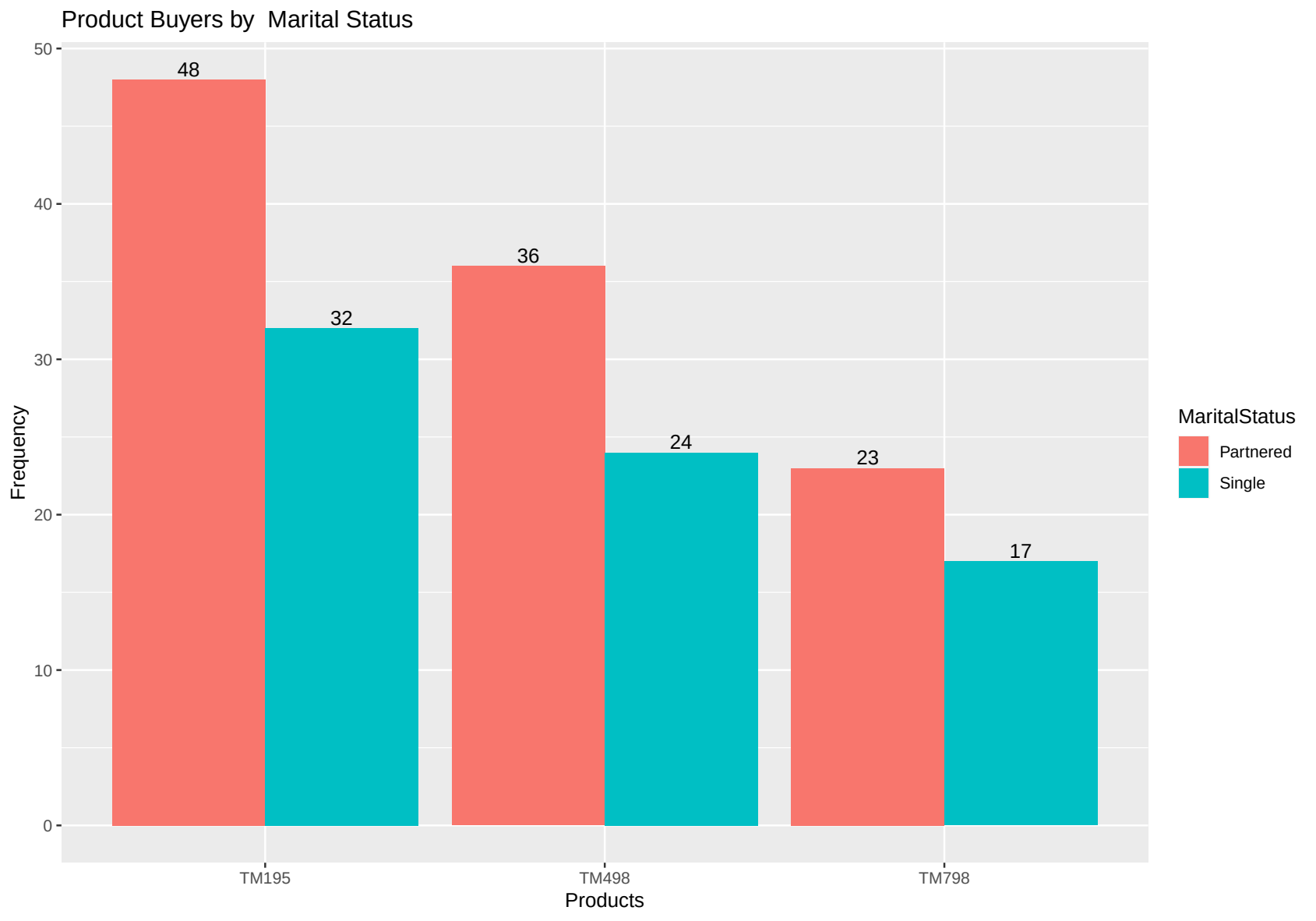
Bivariate Analysis of Product and Marital Status

```
b5<-ggplot(gy,aes(x =Product, fill=MaritalStatus))+
geom_bar(position = "dodge")+
labs(x="Products",y="Frequency",
title="Product Buyers by  Marital Status")

b5_1<- b5+geom_text(stat='count',aes(label=..count..),
                    position=position_dodge(width=0.9),
                    vjust=-0.25)
```

Bivariate Analysis of Product and Gender



Bivariate Analysis of Product and Gender



Bivariate Analysis of Product and Gender

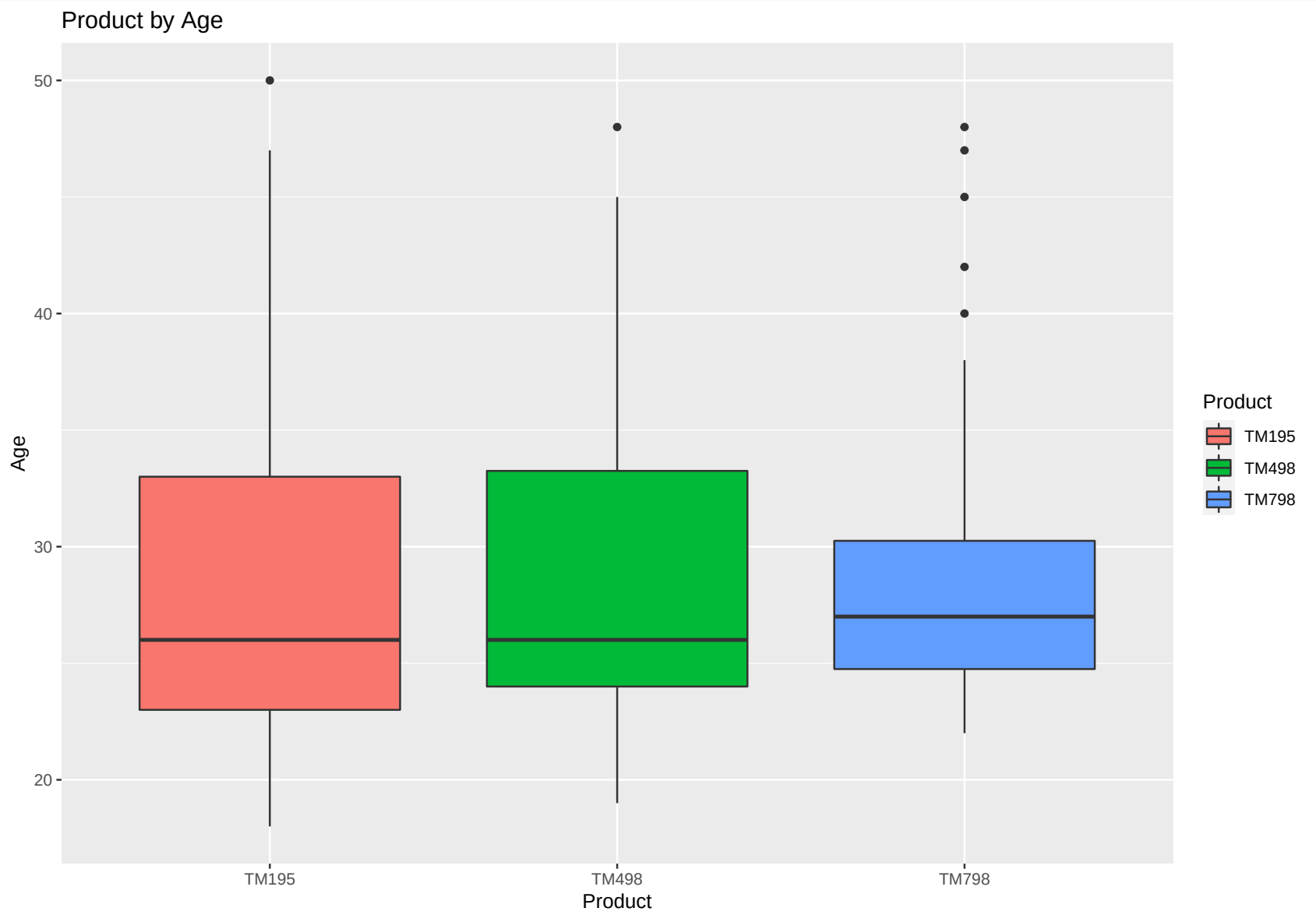
There is a considerable preference for TM798 among males while product TM195 and TM498 are splitted almost evenly among both gender.

For the Product by Marital Status plot, no significant difference in the distribution of the products between partnered and single.

Bivariate Analysis of Product and Age

```
bi1 <- ggplot(gy, aes(x=Product, y=Age, fill=Product)) +
  geom_boxplot()+
  labs(x= "Product", y= "Age",title="Product by Age")
```

Bivariate Analysis of Product and Age



Bivariate Analysis of Product and Age

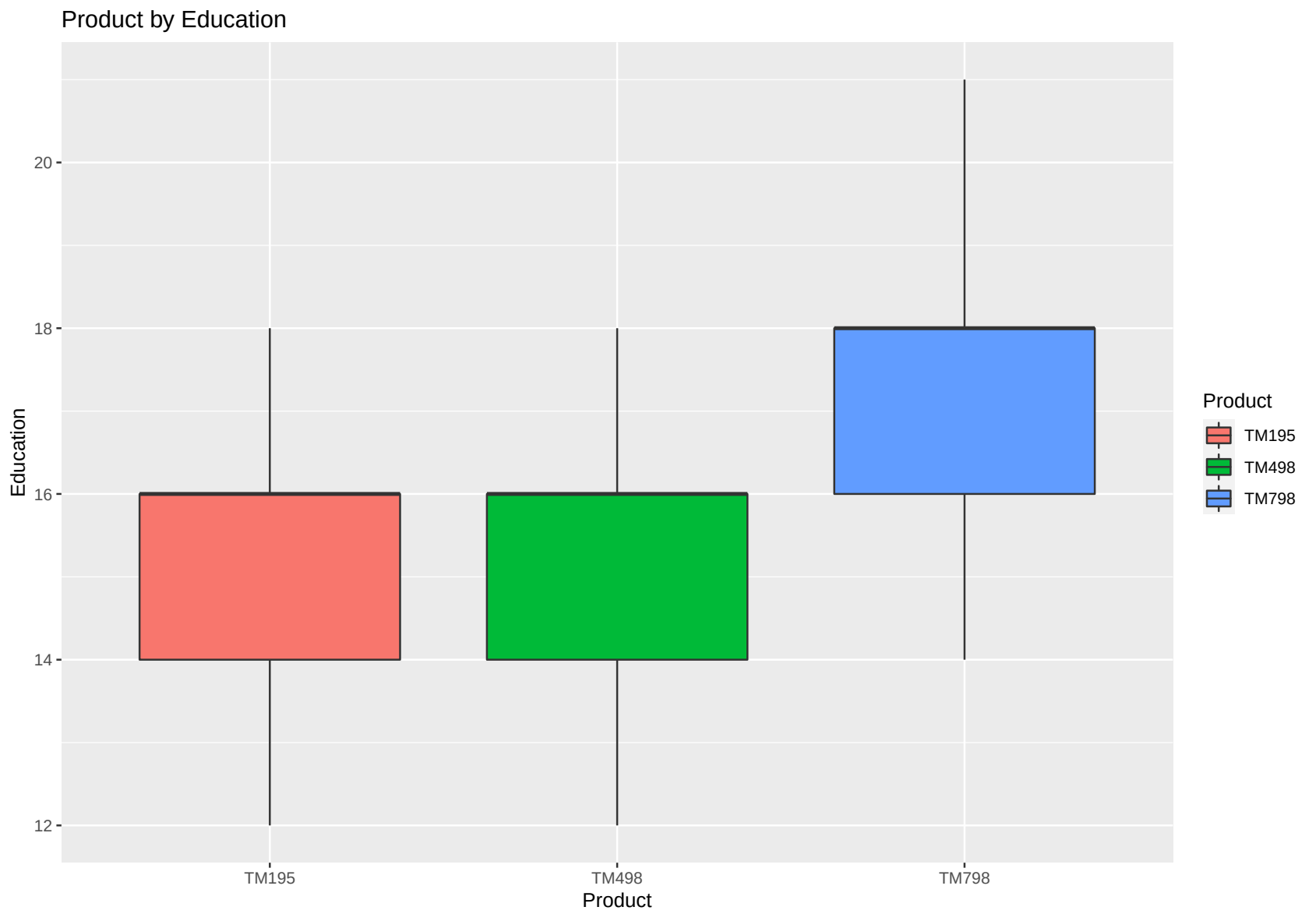
The Age range for TM195 and TM498 buyers are between approximately 18 - 47 years old with the median age as 26 years.

For product TM798 ,the median age is 27 years. On Average, TM798 buyers are older than TM195 and TM498 buyers.

Bivariate Analysis of Product and Education

```
bi2<-ggplot(gy,aes(x=Product,y=Education
           ,fill=Product)) +geom_boxplot()+
  labs(x= "Product", y= "Education",
       title="Product by Education")
```

Bivariate Analysis of Product and Education



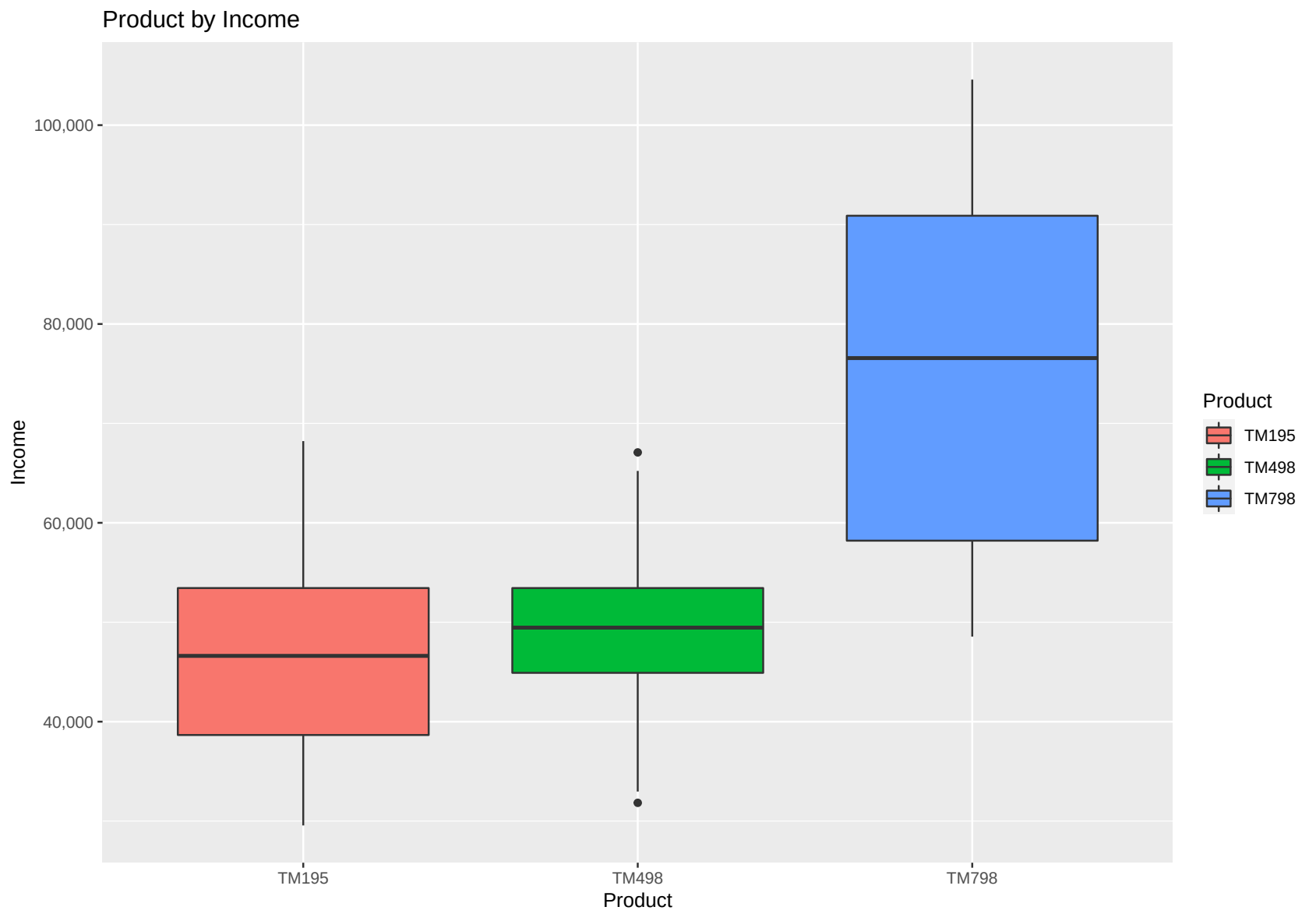
Bivariate Analysis of Product and Education

From the Box Plot, TM798 buyers appear to be highly educated than buyers of TM498 and TM195. Also, the years of education of TM195 and TM498 buyers are between 12 to 18 years while that of TM498 buyers are between 14 to 21 years.

Bivariate Analysis of Product and Income

```
bi3<- ggplot(gy, aes(x=Product, y=Income
    , fill=Product)) + geom_boxplot()+
  scale_y_continuous(labels = comma)+
  labs(x= "Product", y= "Income",
       title="Product by Income")
```

Bivariate Analysis of Product and Income

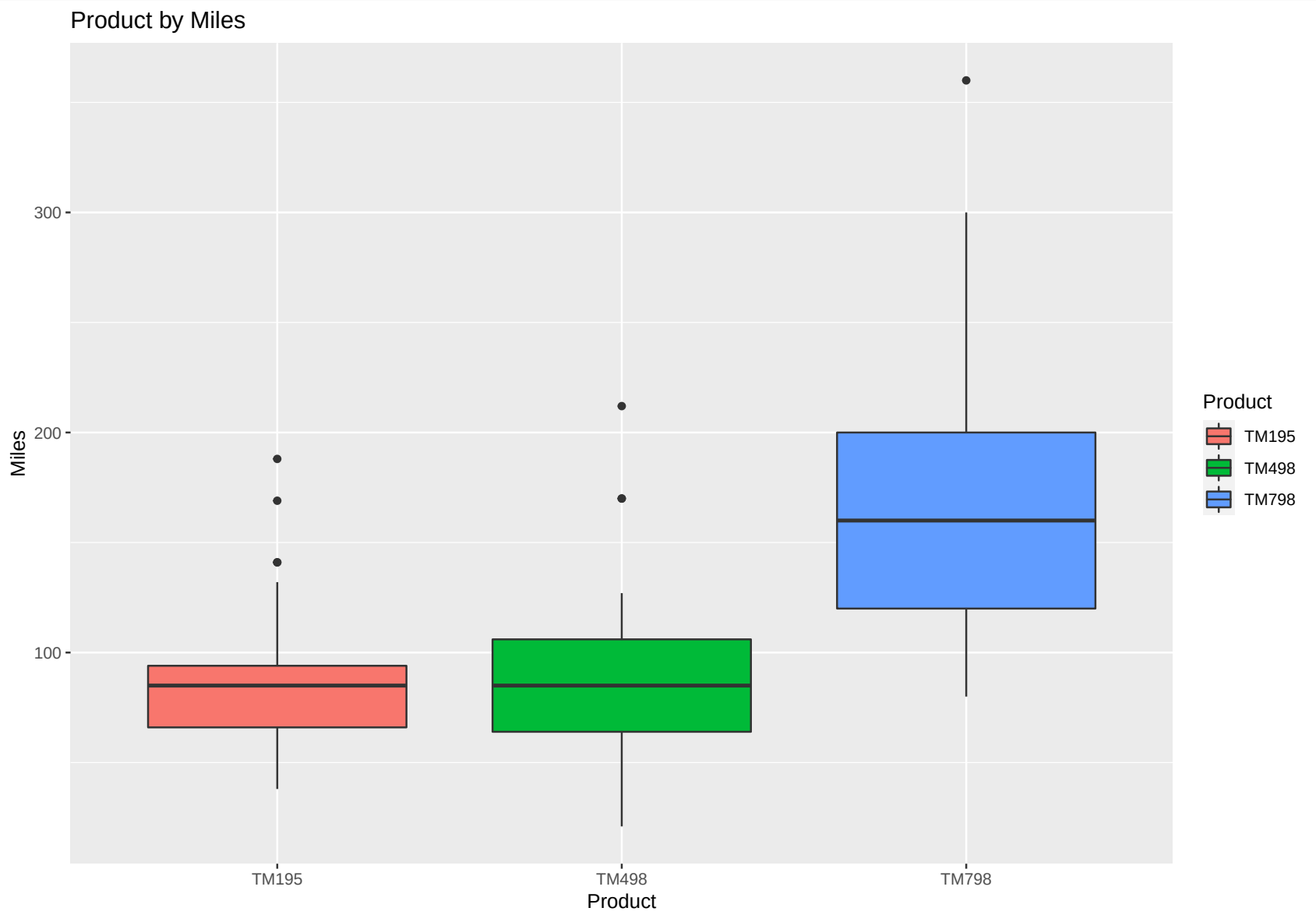


Bivariate Analysis of Product and Income Product TM195

Bivariate Analysis of Product and Miles

```
bi4<- ggplot(gy, aes(x=Product, y=Miles
    , fill=Product)) + geom_boxplot()+
  scale_y_continuous(labels = comma)+
  labs(x= "Product", y= "Miles",
       title="Product by Miles")
```

Bivariate Analysis of Product and Miles



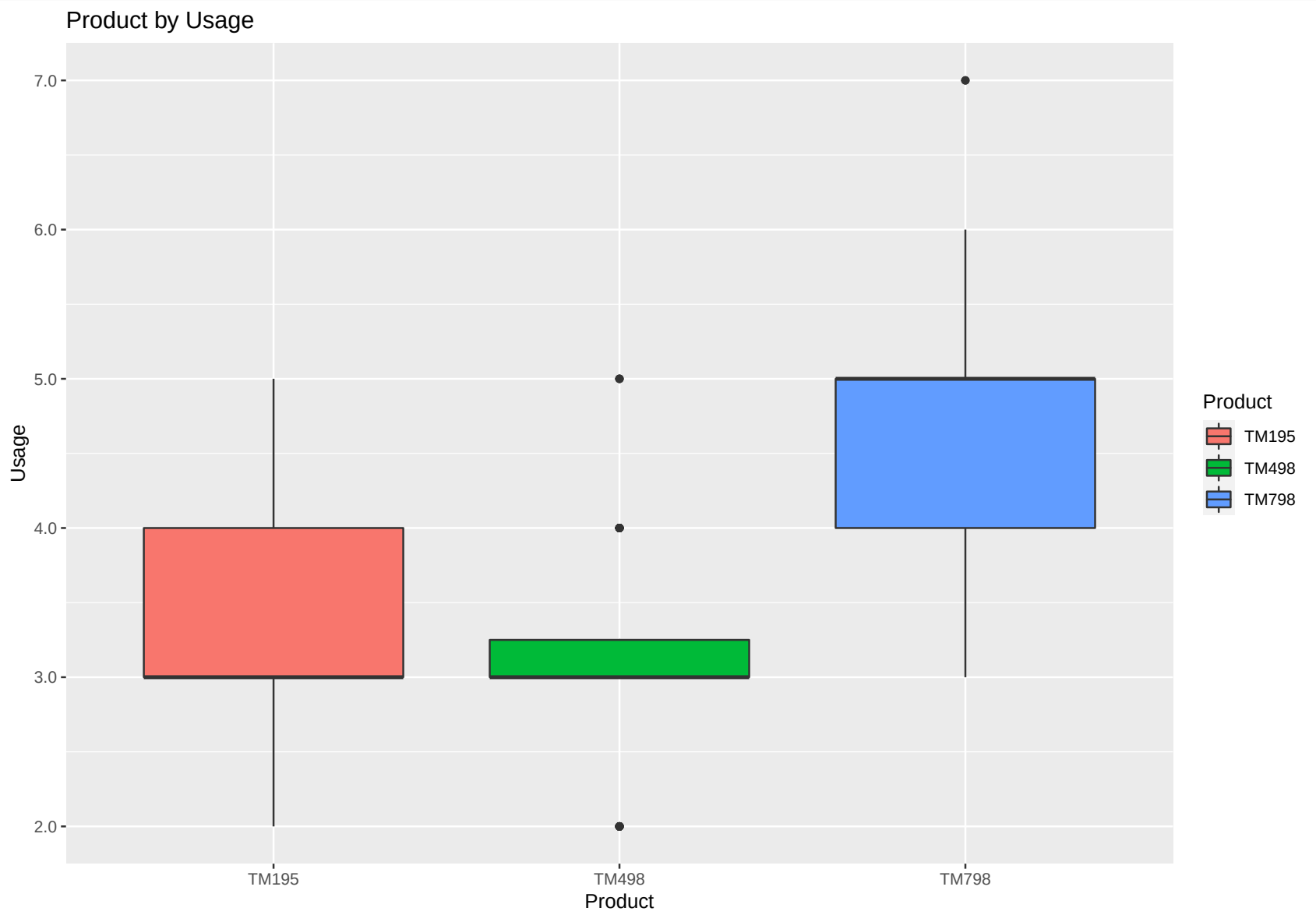
Bivariate Analysis of Product and Miles

TM798 is expected to run more miles than TM195 and TM498. It is expected to cover between 75 miles and 300 miles, with a median of 160 miles. This is two times the median of the expected miles of TM195 and TM498 with median of 80 miles

Bivariate Analysis of Product and Usage

```
bi5<- ggplot(gy, aes(x=Product, y=Usage
    , fill=Product)) + geom_boxplot()+
  scale_y_continuous(labels = comma)+
  labs(x= "Product", y= "Usage",
       title="Product by Usage")
```

Bivariate Analysis of Product and Usage



Bivariate Analysis of Product and Usage

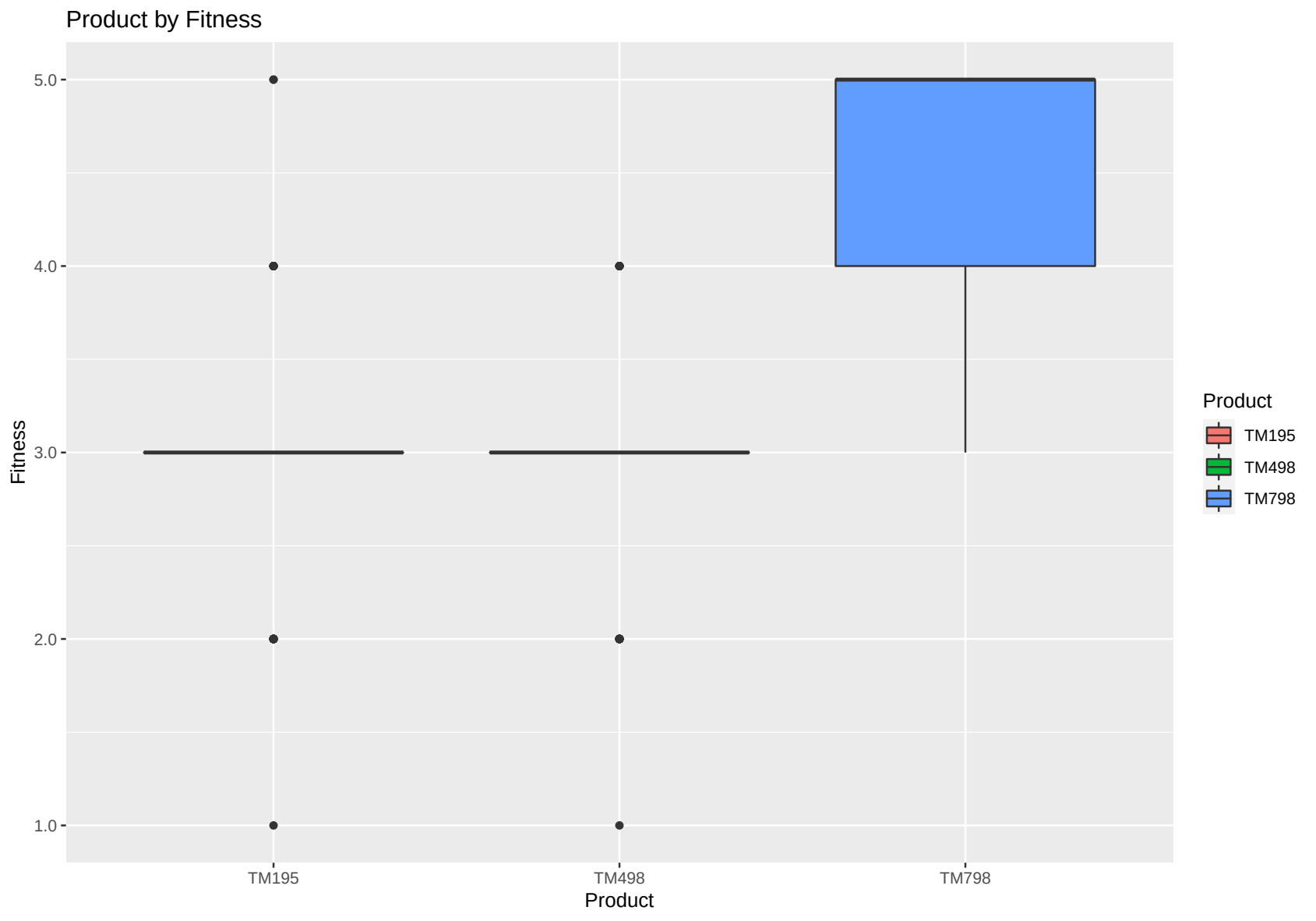
TM798 buyers plan to use their treadmill heavily than buyers of TM195 and TM498. They expect to use their treadmills 3 to 6 times a week, with a median usage of 5 times a week.

TM195 and TM498 buyers plan to use their treadmills between 2 to 5 times and 3 to 4 times respectively with median usages of 3 times a week.

Bivariate Analysis of Product and Fitness

```
bi6<- ggplot(gy, aes(x=Product, y=Fitness
    , fill=Product)) + geom_boxplot()+
  scale_y_continuous(labels = comma)+
  labs(x= "Product", y= "Fitness",
       title="Product by Fitness")
```

Bivariate Analysis of Product and Fitness



Bivariate Analysis of Product and Fitness

TM798 Buyers rate their fitness levels more highly than buyers of TM195 and TM498. TM798 have a median fitness rating of 5 while buyers of TM195 and TM498 have a median fitness level rating of 3.

END