Cardio Fitness Project

Dosubi Joshua

3/7/2020

Table of Contents

1. Project Objective

The objective of this report is to explore the cardio data set ("CardioGoodFitness") in R and generate insights about the data set. This exploration report will consist of the following:

- Importing the dataset in R
- Understanding the structure of dataset
- Graphical exploration
- Descriptive statistics
- Insights from the dataset
- Recommendations that will help the company in targeting new customers

2. Assumptions

1. There are no missing data in any of the fields of the data set. This assumption prevents us from performing field transformations to handle missing data thereby reducing the complexity of exploration and analysis, especially since this project does not cover treating missing values(*See Model Report*).
2. Outliers in the fields of the data set are Negligible. Thisassumption helps simplifies our exploration and analysis, especially since this project does not cover treating Outliers(*See Model Report*).
3. The Target Customers are at least in the same region as the physical store. This assumption helps us to remove Location from variables to consider especially since Location/address data is not provided in the data.
4. The Data Provided is Sample Data not Population Data. This assumption helps us when making insights as we would be able to apply recurrent patterns on other similar subsets.

3. Exploratory Data Analysis – Step by step approach

```
#=========================================================================
#
# Exploratory Data Analysis - CardioGoodFitness
#
#=========================================================================
```

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

```
# Environment Set up and Data Import

# Invoking Libraries
library(tidyverse) # contains ggplot2,dplyr,forcats,lubridate etc
library(gridExtra) # Needed for plotting multiple ggplot graphs side-by-side
library(corrplot) # Needed to plot correlation plots
library(scales) # For Big Number formating
library(knitr) # Necessary to generate sourcecodes from a .Rmd File
```

3.1.2 Set up working Directory

```
# Setup Working Directory
setwd("C:/Users/USER/Documents/El-PaDJo/R programming language/1-Introduction to R & Statistics/week_3")
```

3.1.3 Import and Read the Dataset

```
# Read Input File
cardio_data = read.csv("CardioGoodFitness.csv")

#making all columns accessible without '$' sign usage
attach(cardio_data)
```

3.1.4 Global options settings and Function Definitions

```
# Global options settings
options(scipen=999)  # turn off scientific notation like 1e+06

# Function to calculate mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}
```

3.2 Variable Identification

In order for us to get familiar with the cardio data, below are the functions we would be using to get overview information:

1. **dim()**: this gives us the dimension of the dataset provided. knowing our dimension gives us an idea of how large the data is helping us discern what analysis methods would suffice.
2. **head()**: this shows the first 6 rows(observations) of the dataset. It is essential for us to get a glimpse of the dataset in a tabular format without revealing the entire dataset if we are to properly analyse the data.
3. **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
4. **str()**: this shows us the structure of the dataset. This function is essential as it helps us to determine if there are datatype mismatches specifically (in a very brief format) so that we handle these ASAP to avoid wrong results from our analysis.
5. **summary()**: this provides statistical summaries of the dataset. This function is important as we can quickly get statistical summaries (mean,median, quartiles, min, frequencies/counts, max values etc.) from which we can make insights before even diving into the data themselves for analysis.

3.2.1 Variable Identification - Insights

Insight(s) from dim():

```
# Variable Identification
# check dimension of dataset
dim(cardio_data)
```

```
## [1] 180   9
```

The dataset is not a 'big' dataset.

Insight(s) from head():

```
#see first 6 rows(observations) of dataset
head(cardio_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 names start with 'TM',
- Age is recorded as integers,
- Gender is recorded as Factors "Male"/"Female",
- Education is recorded as integers,
- Marital Status is recorded as Factors "Single"/"Partnered",
- Usage is recorded as integers,
- Fitness is recorded as integers,
- Income is recorded as integers (without comma seperators, Currency or decima points),
- Miles is recorded as integers.
- No NA fields so far

Insight(s) from tail():

```
#see last 6 rows(observations) of dataset
tail(cardio_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
```

Values in all fields are consistent in each column.

Insight(s) from str():

```
# check structure of dataset
str(cardio_data)
```

```
## '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        : 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 ...
```

- All the columns(variables) have appropriate datatypes except the Fitness Variable. As defined in the problem statement, the fitness variable should be a factor, so we would need to chage it to factor for better analysis,

```
#Change fitness variable type to factor
cardio_data$Fitness = as.factor(Fitness)
```

Insight(s) from summary():

```
# get summary of dataset
summary(cardio_data)
```

```
##   Product        Age           Gender      Education       MaritalStatus
##  TM195:80   Min.   :18.00   Female: 76   Min.   :12.00   Partnered:107
##  TM498:60   1st Qu.:24.00   Male  :104   1st Qu.:14.00   Single   : 73
##  TM798:40   Median :26.00                Median :16.00
##             Mean   :28.79                Mean   :15.57
##             3rd Qu.:33.00                3rd Qu.:16.00
##             Max.   :50.00                Max.   :21.00
##      Usage       Fitness     Income          Miles
##  Min.   :2.000   1: 2    Min.   : 29562   Min.   : 21.0
##  1st Qu.:3.000   2:26    1st Qu.: 44059   1st Qu.: 66.0
##  Median :3.000   3:97    Median : 50597   Median : 94.0
##  Mean   :3.456   4:24    Mean   : 53720   Mean   :103.2
##  3rd Qu.:4.000   5:31    3rd Qu.: 58668   3rd Qu.:114.8
##  Max.   :7.000           Max.   :104581   Max.   :360.0
```

- TM195 is the people's choice of product from all the products available,
- Middle aged (24-33) people want these fitness equipments more than other age groups,
- Males want these fitness products more than females,
- Averagely Educated (14-16) People want these fitness products more than others,
- "Partnered" people want these fitness products more than "Single" people,
- People who indicate that they want to use the treadmills extensively every week tend to patronise these products more than those who indicate that they would use it sparingly every week,

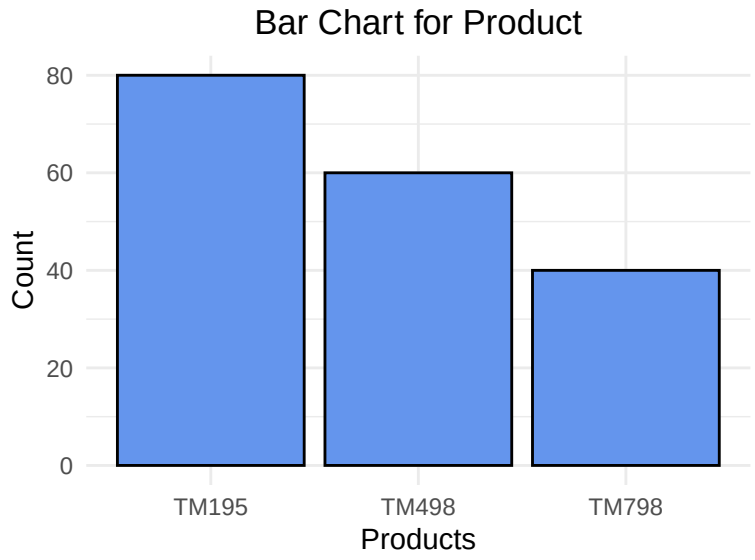
- People who have a fitness conciousness and indicate that they are fit patronise the products,
- People whose incomes are good or better tend to find the need for these equipments more important,
- People who indicate that they need to run average mileage(66-114.8) tend to patronise these products more.

3.3 Univariate Analysis

1. **Observations on Product**:

    - TM195 was bought more than TM498 *as shown on the bar chart*,
    - TM798 is the least purchased *as shown on the bar chart*.

```
# Bar Plot for Product Variable
ggplot(cardio_data, aes(x = Product)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Product", x="Products", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```

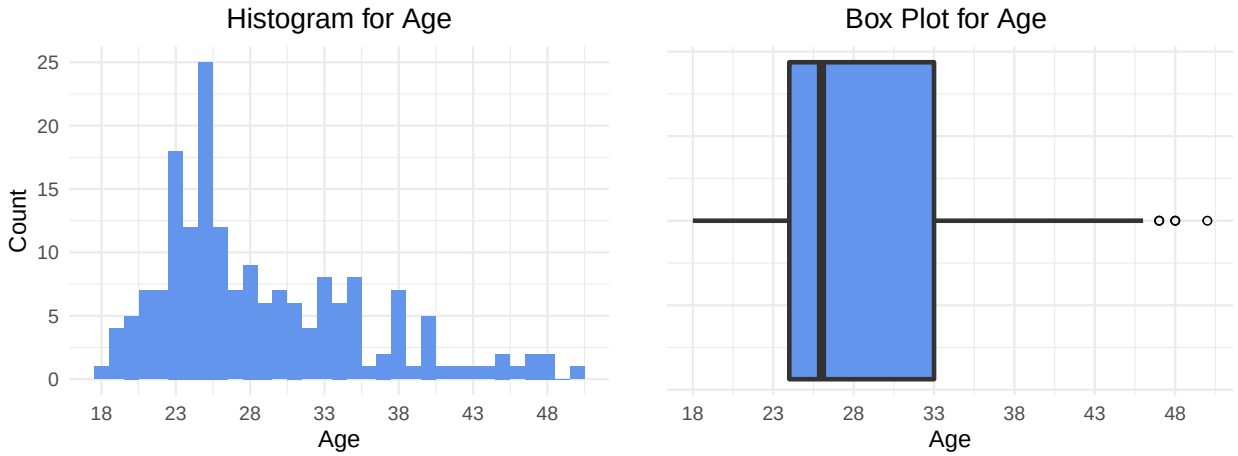


2. **Observations on Age**:

    - Age 25 is fitness age! *as shown on the histogram*,
    - The skewness looks to be a left skewed as the curve looks like that of a standard left skew, *as observed on the histogram*,
    - Some old people still workout! *as shown on the Box plot as outliers.*

```
# Histogram for Age Variable
age_histogram = ggplot(cardio_data, aes(x = Age)) +
  geom_histogram(fill = "cornflowerblue", bins = 33) +
  labs(title="Histogram for Age", x="Age", y="Count") +
  scale_x_continuous(breaks = seq(18, max(Age)+5, by = 5)) +
  scale_y_continuous(breaks = seq(0, 30, by = 5)) +
  theme_minimal() +
```

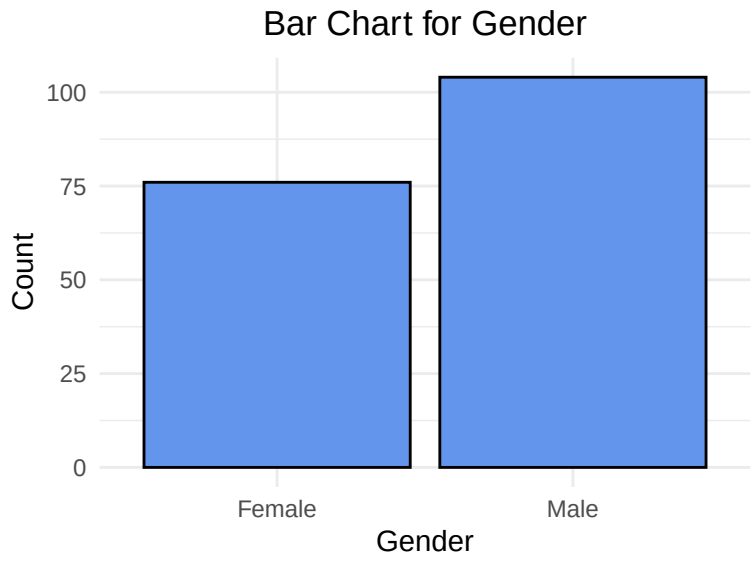
```
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for Age Variable
age_boxplot = ggplot(cardio_data, aes(x = 0, y = Age)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Age ", x = "", y="Age") +
  scale_y_continuous(breaks = seq(18, max(Age)+5, by = 5)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(age_histogram,age_boxplot,nrow=1,ncol= 2)
```



3. **Observations on Gender**:

- The majority of buyers are male *as shown on the bar chart*,
- The females also make purchases, just not as much as the males *as shown on the bar chart.*

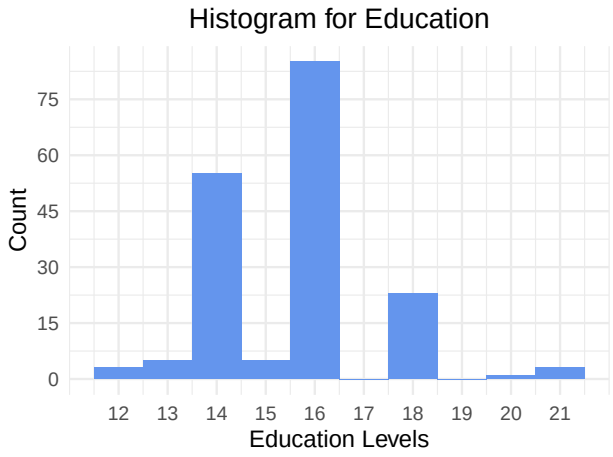
```
# Bar Plot for Gender Variable
ggplot(cardio_data, aes(x = Gender)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Gender", x="Gender", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```

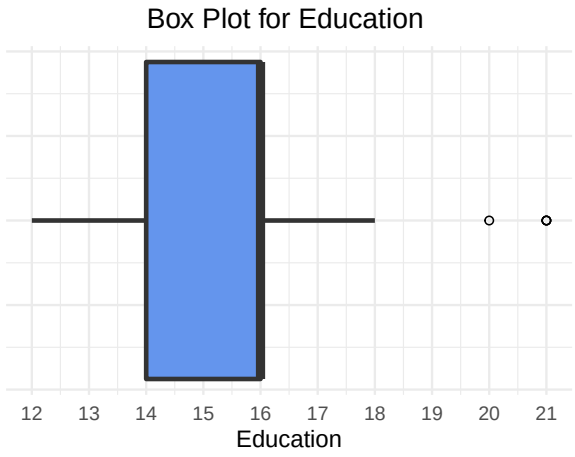


4. **Observations on Eduation**:

   - People with average educational levels buy more products *as shown on the Histogram*,
   - The very educated people seem to have less interest in buying these products *as shown on the Histogram*,
   - The skewness looks to be a left skewed as the curve looks like that of a standard left skew, *as observed on the histogram*,
   - Once in a very rare while, people at the peak of their career educationally tend to buy the products *as shown on the Box plot as outliers.*

```
# Histogram for Education Variable
edu_histogram = ggplot(cardio_data, aes(x = Education)) +
  geom_histogram(fill = "cornflowerblue", bins = 10) +
   labs(title="Histogram for Education", x="Education Levels", y="Count") +
  scale_x_continuous(breaks = seq(min(Education), max(Education), by = 1)) +
  scale_y_continuous(breaks = seq(0, 100, by = 15)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box for Education Variable
edu_boxplot = ggplot(cardio_data, aes(x = 0, y = Education)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Education ", x = "", y="Education") +
  scale_y_continuous(breaks = seq(min(Education), max(Education), by = 1)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(edu_histogram,edu_boxplot,nrow=1,ncol= 2)
```

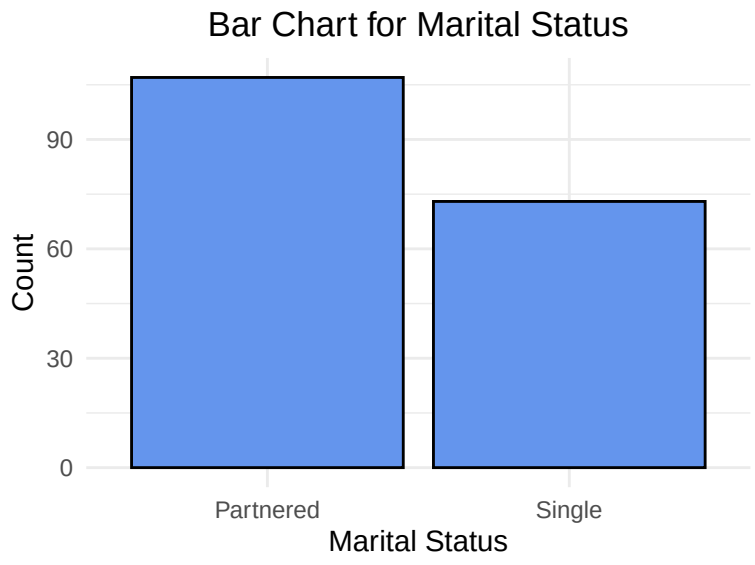




5. **Observations on Marital Status**:

- The majority of buyers are Partneres *as shown on the bar chart*,
- The Singles also make purchases, just not as much as the Partnered *as shown on the bar chart*.

```
# Bar Plot for Marital Status Variable
ggplot(cardio_data, aes(x = MaritalStatus)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Marital Status", x="Marital Status", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```

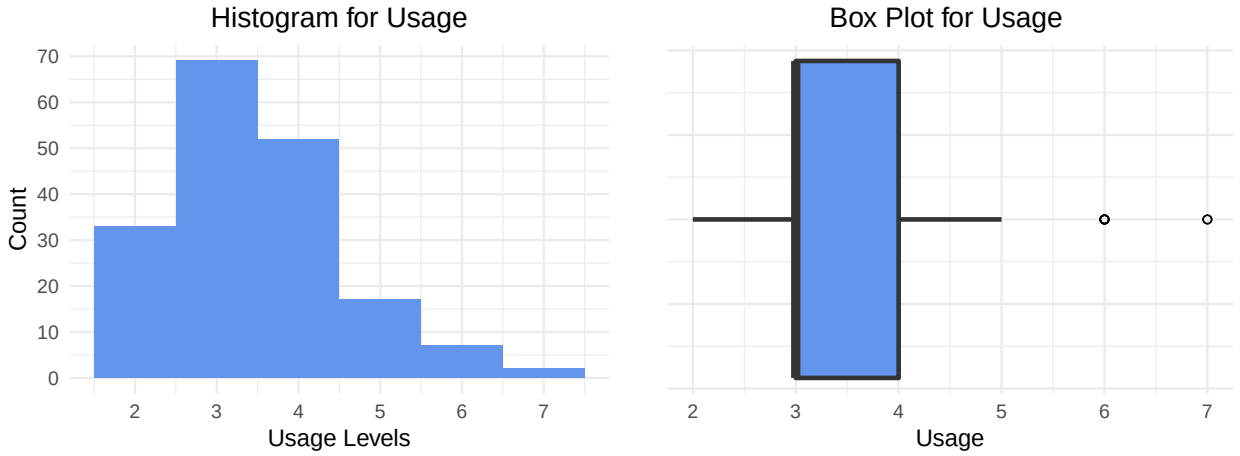


6. **Observations on Usage**:

- People who intend to use these devices sparingly during the week buy more products *as shown on the Histogram*,
- People who claim that the would use the device so often that it breaks dont end up buying the product *as shown on the Histogram*,

- The skewness looks to be a left skewed as the curve looks like that of a standard left skew, *as observed on the histogram*,
- Once in a very rare while, people with high claims tend to buy the products *as shown on the Box plot as outliers*, but they are too few to matter in analysis.

```
# Histogram for Usage Variable
usage_histogram = ggplot(cardio_data, aes(x = Usage)) +
  geom_histogram(fill = "cornflowerblue", bins = 6) +
   labs(title="Histogram for Usage", x="Usage Levels", y="Count") +
  scale_x_continuous(breaks = seq(min(Usage), max(Usage), by = 1)) +
  scale_y_continuous(breaks = seq(0, 80, by = 10)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box for Usage Variable
usage_boxplot = ggplot(cardio_data, aes(x = 0, y = Usage)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Usage ", x = "", y="Usage") +
  scale_y_continuous(breaks = seq(min(Usage), max(Usage), by = 1)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(usage_histogram,usage_boxplot,nrow=1,ncol= 2)
```

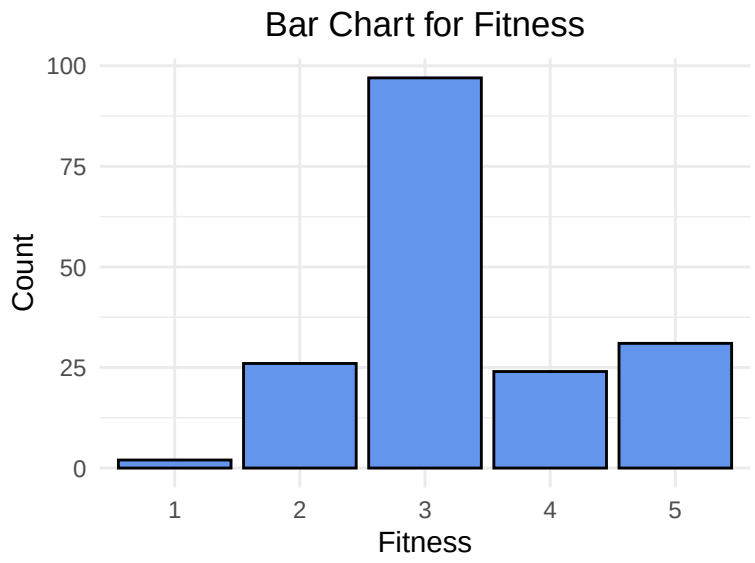


7. **Observations on Fitness**:

- The people who say they are fit boldly buy these products *as shown on the bar chart*,
- Other people who claim to be super fit or who are not very familiar with fitness usually dont buy the products *as shown on the bar chart*.

```
# Bar Plot for Fitness Variable
ggplot(cardio_data, aes(x = factor(Fitness))) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Fitness", x="Fitness", y="Count") +
```

```
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```



8. **Observations on Income**:

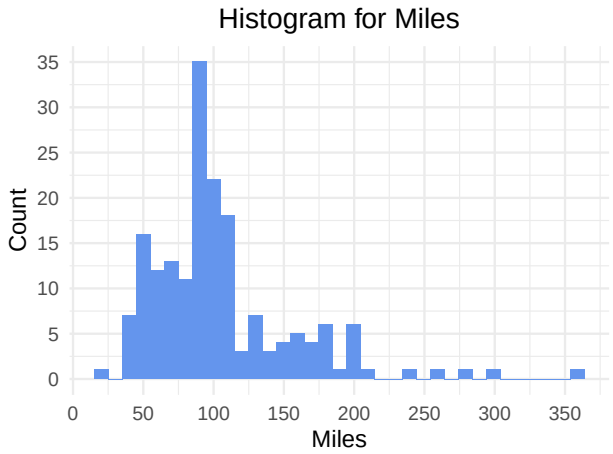
   - People with average incomes in general buy more products *as shown on the Histogram*,
   - The very rich people dont really buy these products *as shown on the Histogram*,
   - Once in a very rare while, very rich people tend to buy the products *as shown on the Box plot as outliers*.

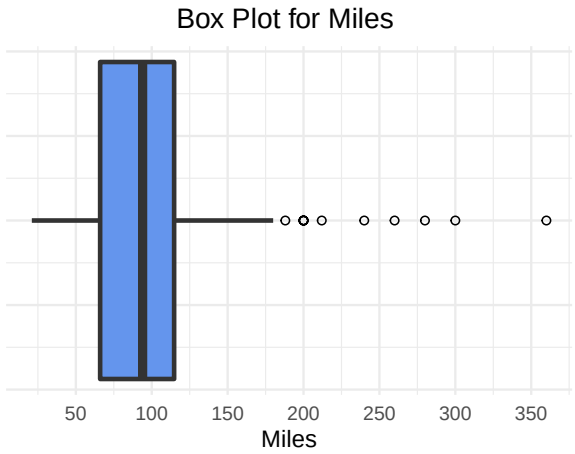
```
# Histogram for Income Variable
income_histogram = ggplot(cardio_data, aes(x = Income)) +
  geom_histogram(fill = "cornflowerblue", bins = 50) +
  labs(title="Histogram for Income", x="Income Levels", y="Count") +
  scale_x_continuous(breaks = seq(20000, 120000, by = 10000), labels = comma) +
  scale_y_continuous(breaks = seq(0, 25, by = 5)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5))

# Box for Income Variable
income_boxplot = ggplot(cardio_data, aes(x = 0, y = Income)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Income ", x = "", y="Income") +
  scale_y_continuous(breaks = seq(20000, 120000, by = 10000), labels = comma) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(income_histogram,income_boxplot,nrow=1,ncol= 2)
```

9. **Observations on Miles**:

- People who intend to use these devices for just a few sufficient miles buy more products *as shown on the Histogram*,
- People who claim that the would use the device so often that it breaks dont end up buying the product *as shown on the Histogram*,
- The skewness looks to be a left skewed as the curve looks like that of a standard left skew, *as observed on the histogram*,
- Once in a very rare while, people with high claims tend to buy the products *as shown on the Box plot as outliers*, but they are too few to matter in analysis.

```
# Histogram for Miles Variable
miles_histogram = ggplot(cardio_data, aes(x = Miles)) +
  geom_histogram(fill = "cornflowerblue", bins = 35) +
  labs(title="Histogram for Miles", x="Miles", y="Count") +
  scale_x_continuous(breaks = seq(0, max(Miles)+10, by = 50)) +
  scale_y_continuous(breaks = seq(0, 40, by = 5))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box for Miles Variable
miles_boxplot = ggplot(cardio_data, aes(x = 0, y = Miles)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Miles ", x = "", y="Miles") +
  scale_y_continuous(breaks = seq(0, max(Miles)+10, by = 50)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(miles_histogram,miles_boxplot,nrow=1,ncol= 2)
```



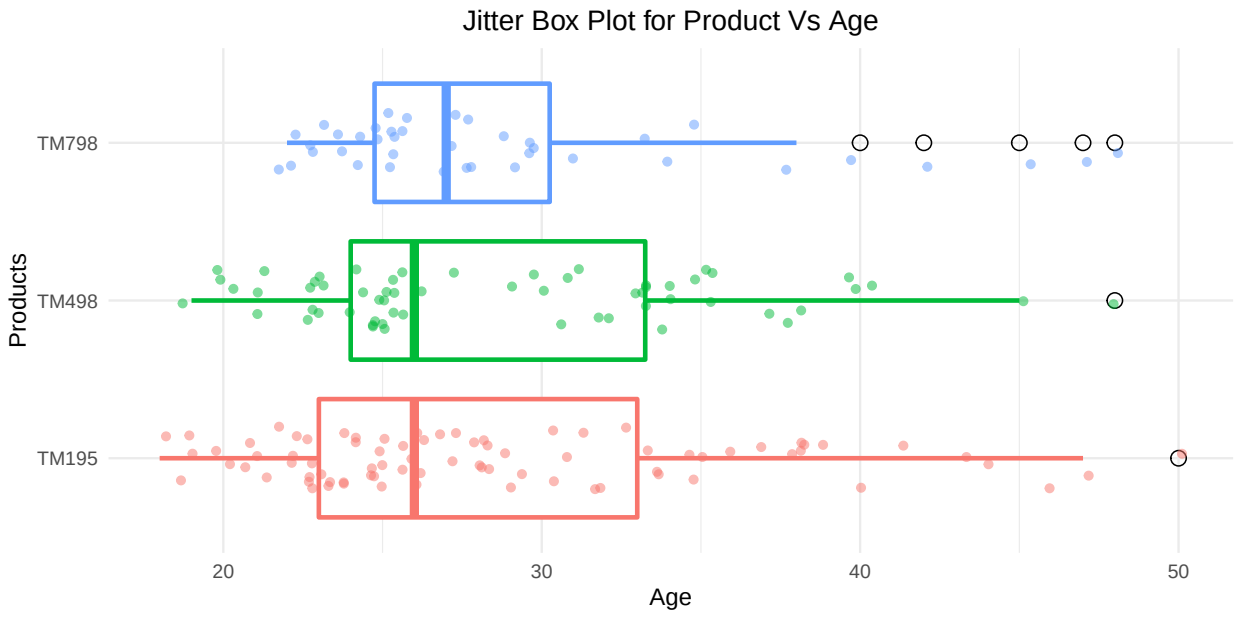


3.4 Bivariate Analysis

In order to get more interpretations, we plot bivariate charts to see the relationships between the variables:

1. PRODUCT VS AGE:
    - Every Age group likes buying TM195 and TM498 but younger people go for TM798,
    - Older people dont like TM798.

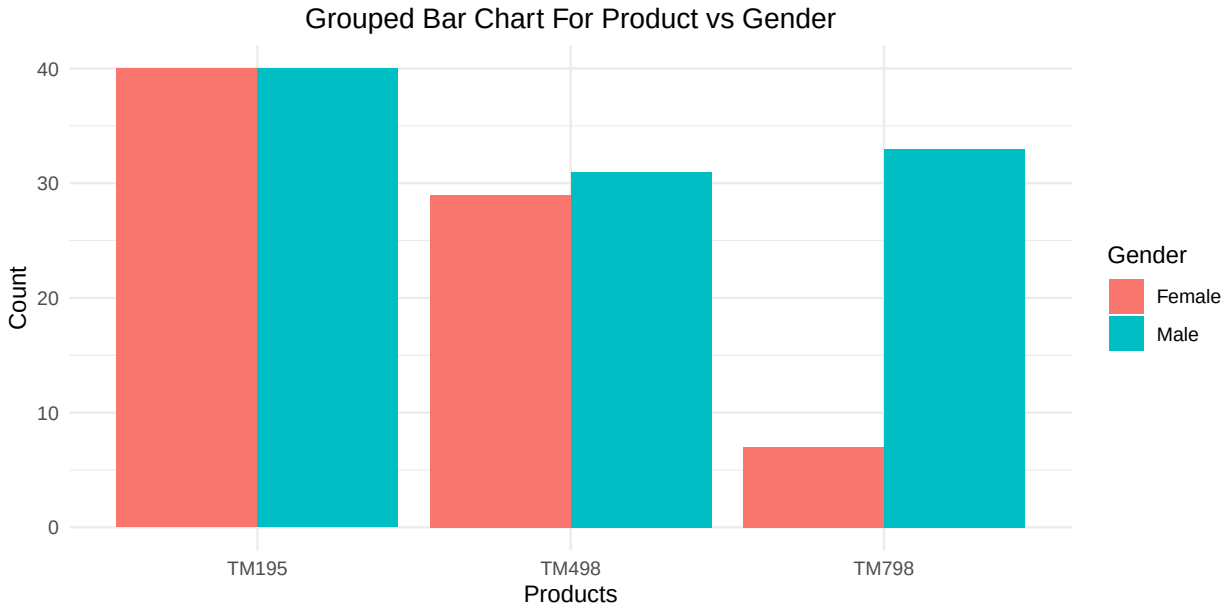
```
# Jitter box Plot for Product vs Age
ggplot(cardio_data, aes(x = Product, y = Age, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Age", x = "Products", y = "Age") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()
```



2. PRODUCT VS GENDER:

- Every Gender likes buying TM195 and TM498 but Males Majorly go for TM798,
- Females dont like TM798,

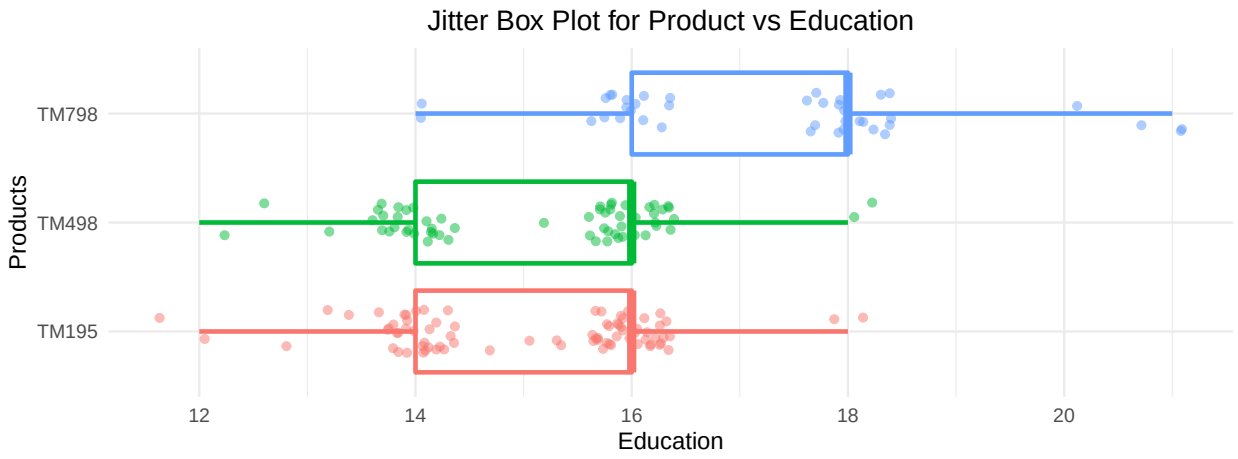
```
# Grouped Bar Chart For Product vs Gender
ggplot(cardio_data, aes(x = Product, fill = Gender)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Gender", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```



3. PRODUCT VS EDUCATION:

- Well Educated people buy 798 but other educational leveled people prefer TM195 and TM498 *as shown on the Jitter box Plot*,
- People with low education levels are not interested in buying any of the products *as shown on the Jitter box Plot*,

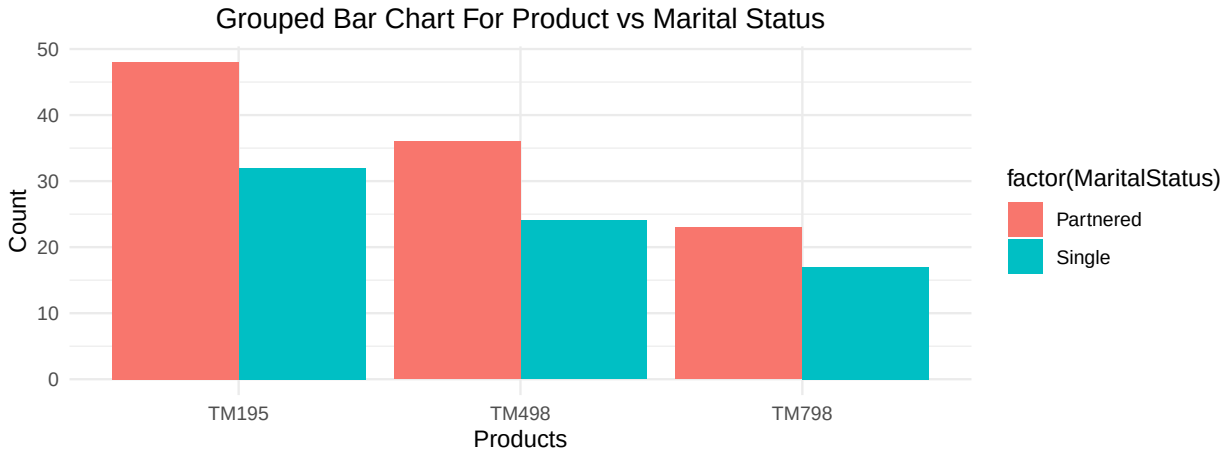
```
# Jitter box Plot for Product vs Education
ggplot(cardio_data, aes(x = Product, y = Education, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product vs Education", x = "Products", y = "Education") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()
```



4. PRODUCT VS MARITAL STATUS:

- Marital status does not really determine the type of product bought *as shown on the Grouped Bar Chart.*

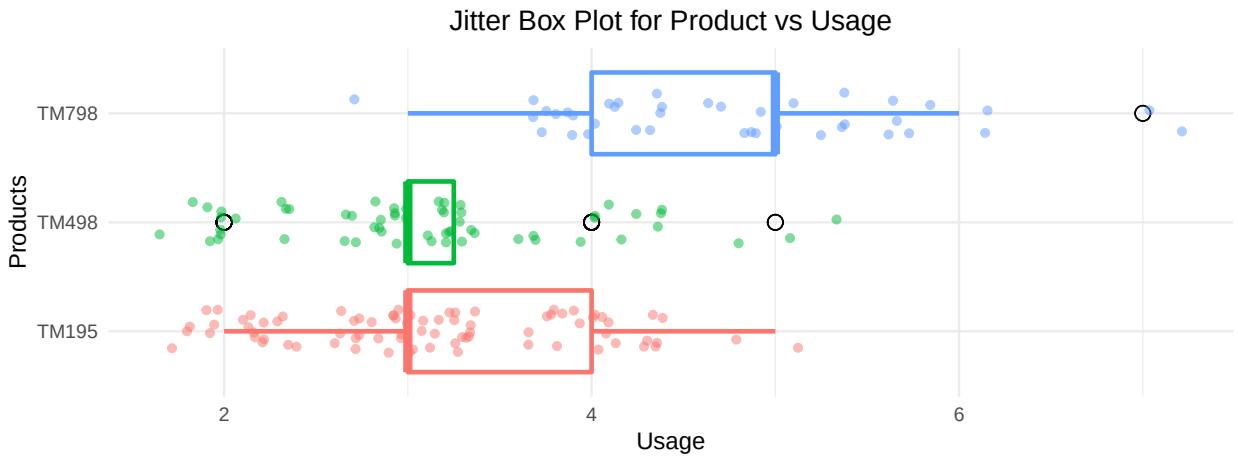
```
# Grouped Bar Chart For Product vs Marital Status
ggplot(cardio_data, aes(x = Product, fill = factor(MaritalStatus))) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Marital Status", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```



5. PRODUCT VS USAGE:

- People who intend an average use per week and a good number of people who indicate high usage buy TM798 *as shown on the Jitter box Plot*,
- People who indicate low usage weekly go for TM195 and TM498 *as shown on the Jitter box Plot*,
- People who indicate low usage weekly prefer TM195 over TM498 *as shown on the Jitter box Plot*,
- the jitter spread of usage for TM498 is quite noisy, we'll need more analysis to see what really happens here *as shown on the Jitter box Plot.*

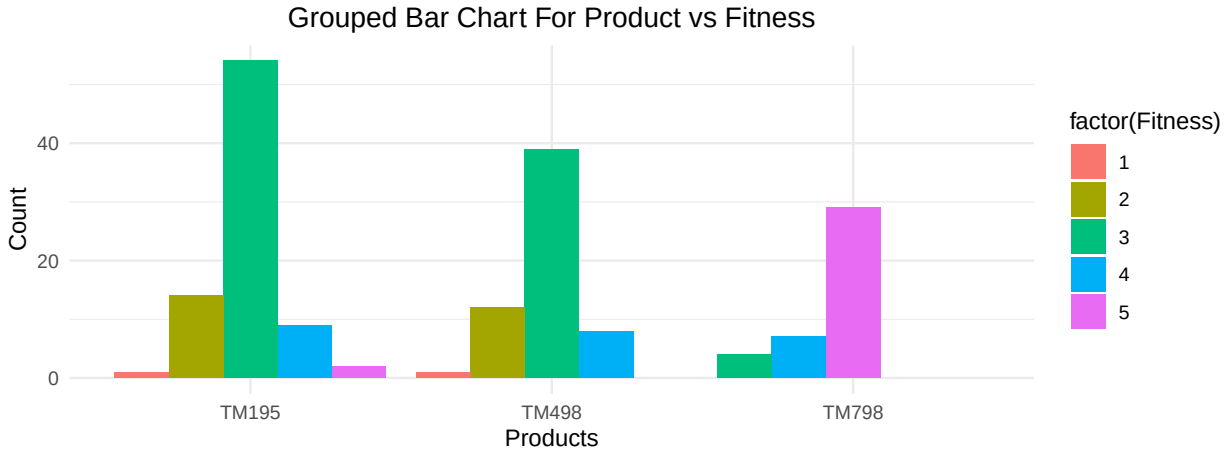
```r
# Grouped Bar Chart For Product vs Usage Status
ggplot(cardio_data, aes(x = Product, y = Usage, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product vs Usage", x = "Products", y = "Usage") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()
```



6. PRODUCT VS FITNESS:

- people who claim to be Very fit buy TM798 *as shown on the Grouped Bar Chart*,
- people who claim to be fit buy TM195 and TM498 *as shown on the Grouped Bar Chart*,
- people who claim to be very unfit dont bother buying any product *as shown on the Grouped Bar Chart*,

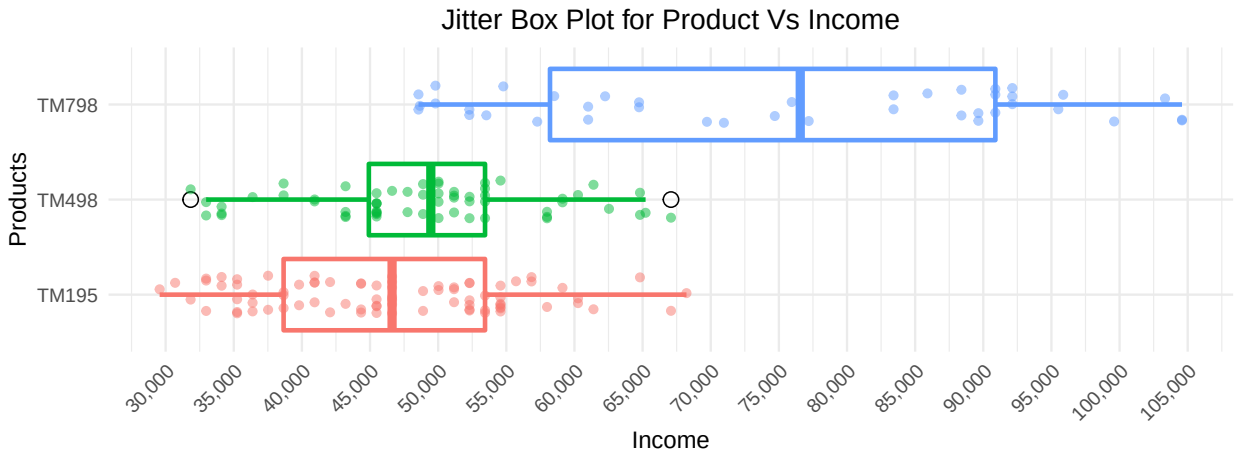
```r
# Grouped Bar Chart For Product vs Marital Status
ggplot(cardio_data, aes(x = Product, fill = factor(Fitness))) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Fitness", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```



7. PRODUCT VS Income:

- The very High Income Earners buy TM798 *as shown on the Jitter box Plot*,
- The Average Income Earners go for TM498 and TM195 *as shown on the Jitter box Plot*,
- Low Income Earners prefer TM195 over TM498 *as shown on the Jitter box Plot*,
- the jitter spread of Income for TM498 is quite noisy, we'll need more analysis to see what really happens here *as shown on the Jitter box Plot*.

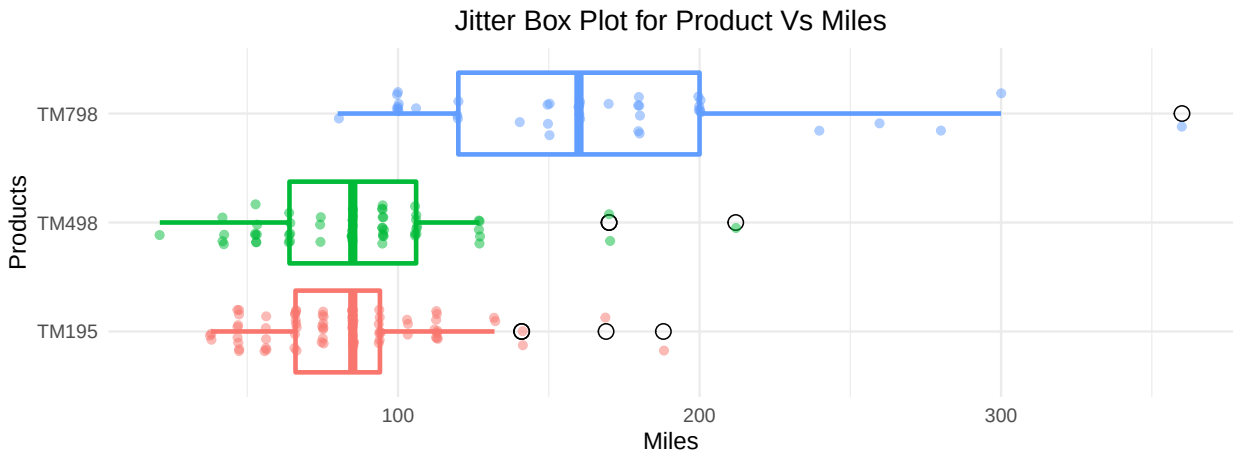
```
# Jitter box Plot for Product vs Income
ggplot(cardio_data, aes(x = Product, y = Income, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Income", x = "Products", y = "Income") +
  scale_y_continuous(breaks = seq(20000, 120000, by = 5000), labels = comma) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```



8. PRODUCT VS Miles:

- People who indicate moderate mileage and a very few ones who indicate high mileage buy TM798 *as shown on the Jitter box Plot*,
- People who indicate low mileage go for TM195 and TM498 *as shown on the Jitter box Plot*,

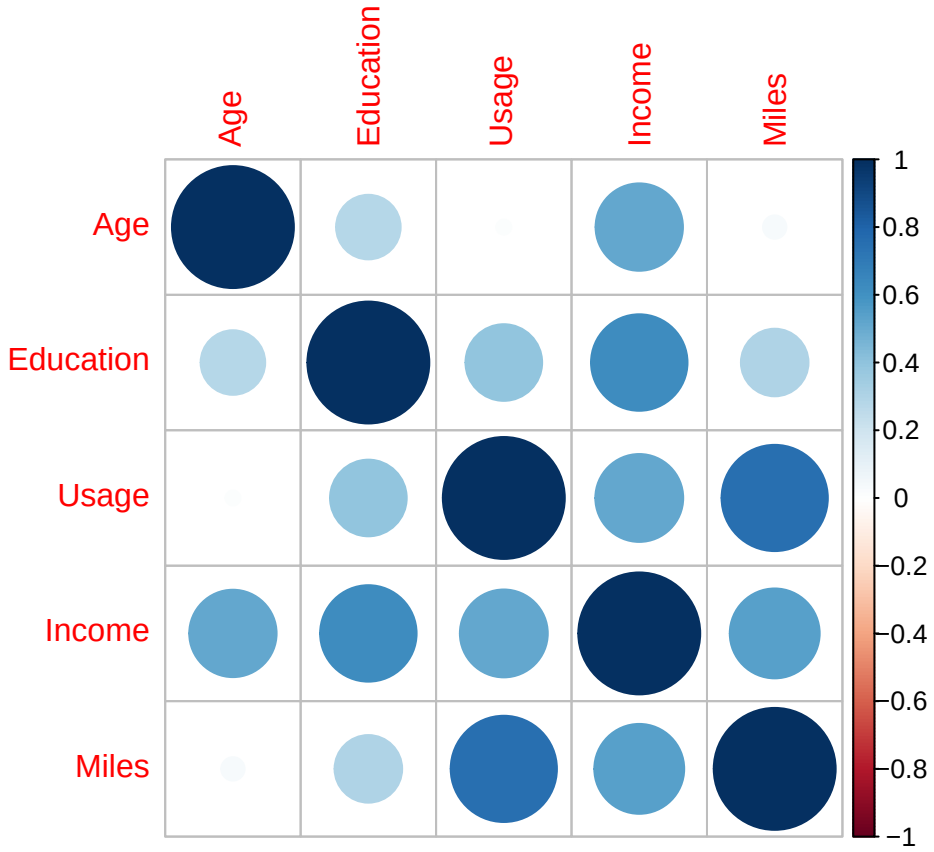
```
# Jitter box Plot for Product vs Age
ggplot(cardio_data, aes(x = Product, y = Miles, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Miles", x = "Products", y = "Miles") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()
```



9. CORRELLATION PLOT for Age, Education, Usage, Income and Miles: As shown on the Subsequent Correllation Plot,

    - AGE and INCOME shows a strong correlation implying that the older a person gets, the higher their income and vice versa,
    - AGE and EDUCATION shows a mild correlation implying that the older a person gets, the higher their Education levels and vice versa. But this doesnt always happen,
    - EDUCATION and USAGE show a strong correlation implying that the higher a person's education level, the higher his intended weekly usage and vice versa,
    - EDUCATION and MILES show a mild correlation implying that the higher a person's education level, the higher his intended mileage and vice versa. But this doesnt always happen,
    - USAGE and INCOME show a strong correlation implying that when a persons income is low, his intended weekly usage is also low vice versa,
    - USAGE and MILES show a very high correlation implying that when the intended weekly usage is low, the mileage is alos low by default and vice versa,
    - INCOME and MILES show a strong correlation implying that the lower the Income, the lower the mileage and vice versa,
    - INCOME seems to be the major Factor to be considered when looking at this correlation data, followed by EDUCATION.

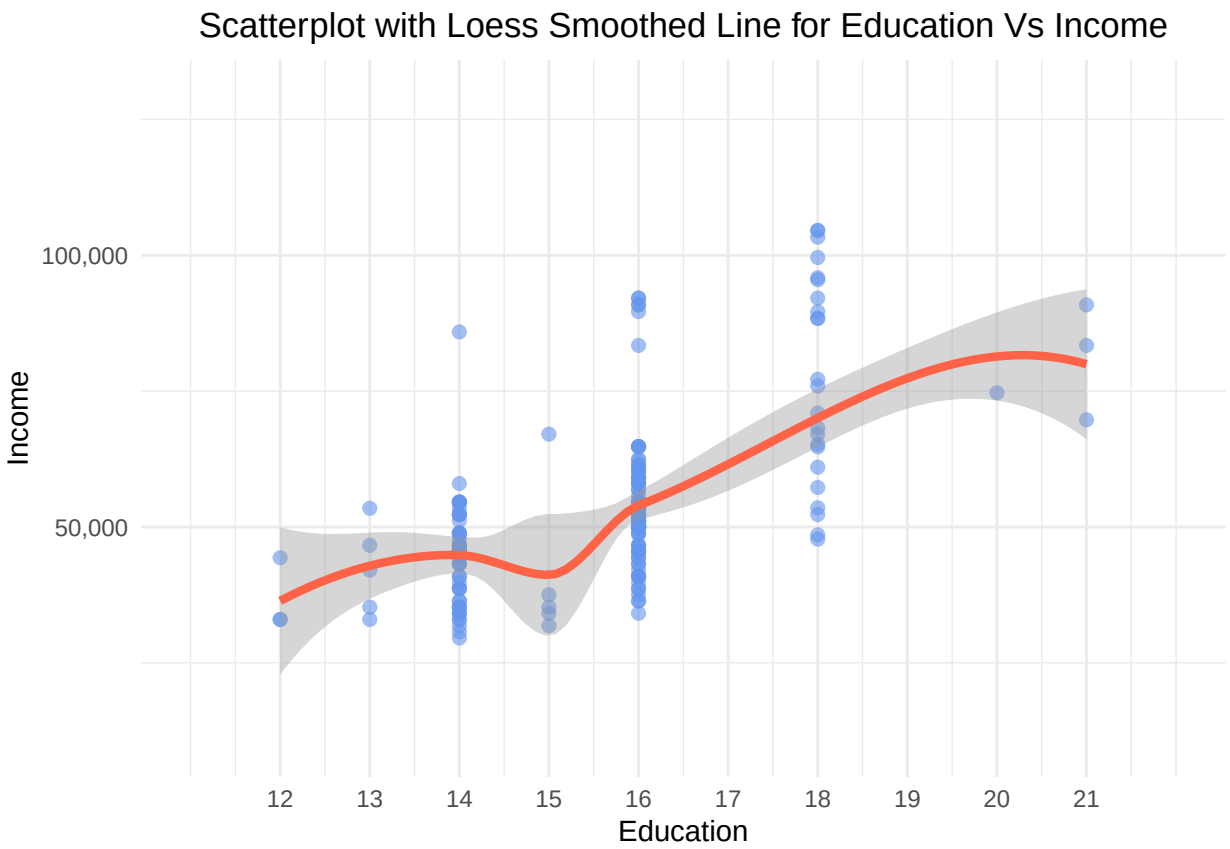
```
#Make a Correllation plot of AGE, Education, Usage, Income and miles
corrplot(cor(cardio_data[,c(2,4,6,8,9)]))
```



10. EDUCATION VS INCOME:

- The Income increases as a person's education level increases,
- When People are at average education level, their incomes sometimes remains stagnant or even deeps as they try to move to the next levels of education,
- At the peak of Education levels, we see that the income doesnt increase anymore, but now starts decreasing as they try to climb higher on the education cadre

```
# scatterplot with loess smoothed line
# and better labeling and color
ggplot(cardio_data, aes(x = Education, y = Income)) +
  geom_point(color="cornflowerblue", size = 2, alpha = .6) +
  geom_smooth(size = 1.5, color = "tomato") +
  scale_y_continuous(labels = comma, limits = c(10000, 130000)) +
  scale_x_continuous(breaks = seq(min(Education), max(Education), by = 1),
                     limits = c(11, 22)) +
  labs(title = "Scatterplot with Loess Smoothed Line for Education Vs Income",
       x = "Education", y = "Income") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```



3.5 Customer Profile

Here, we would like to see the characteristics of the customers that buys each product:

1. TM195:

   - Age group: usually 23 - 33, Averaged at 28.55
   - Gender: either "Female" or "Male" in a 50:50 ratio
   - Education category: usually 14 - 16, , Averaged at 15.04
   - Marital STatus: usually "Partnered"
   - Usage Category: usually 3 - 4, , Averaged at 3.087
   - Fitness Category: Mostly 3
   - Income Category: usually 38658 - 53439, Averaged at 46418
   - Miles category: usually 66 - 94, Averaged at 82.79

   IN A SENTENCE: *"Educated Middle-aged men/women, maritally partnered with Moderate incomes, who see themselves as quite fit and plan to use the product around 3 times a week expecting to run just an adequate number of miles or thereabout on it."*

   IN A PHRASE: *"Preferred By the Everyone"*

2. TM498:

   - Age group: usually 24 - 33.25, Averaged at 28.55
   - Gender: either "Female" or "Male" in a 48:52 ratio
   - Education category: usually 14 - 16, , Averaged at 15.12

- Marital STatus: usually “Partnered”
- Usage Category: usually 3 - 3.250, , Averaged at 3.067
- Fitness Category: Mostly 3
- Income Category: usually 44912 - 53439, Averaged at 48974
- Miles category: usually 64 - 106, Averaged at 87

IN A SENTENCE: *“Educated Middle-aged men(Mostly), maritally partnered with Moderate incomes, who see themselves as quite fit and plan to use the product around 3 times a week expecting to run just adequate number of miles or thereabout on it.”*

IN A PHRASE: *“Preferred by the Average Man”*

3. TM798:

- Age group: usually 24.75 - 30.25, Averaged at 29.10
- Gender: mostly “Male”
- Education category: usually 16 - 18, , Averaged at 17.32
- Marital STatus: usually “Partnered”
- Usage Category: usually 4 - 5, , Averaged at 4.775
- Fitness Category: Mostly 5,
- Income Category: usually 58205 - 90886, Averaged at 75442
- Miles category: usually 120 - 200, Averaged at 166

IN A SENTENCE: *“well-Educated Middle-aged men, maritally partnered with High incomes, who see themselves as very fit and plan to use the product many times a week expecting to run a high number of miles on it”*

IN A PHRASE: *“Preferred by Educated Rich Men”*

3.6 Missing Value Identification

```
+ There are no NA fields in this data set
```

```
# check if NA field exists in the dataset
anyNA(cardio_data)
```

```
## [1] FALSE
```

3.7 Outlier Identification

According to the Box plots in the Univariate Analysis section,

- AGE: there are a few outliers to the high end here but quite negligible,
- EDUCATION: there are a few outliers to the high end here which can really affect the average calculation,
- USAGE: there are a few outliers to the high end here but quite negligible,
- INCOME: there are a lot of outliers to the high end here and it needs to be analysed closely for better results,
- MILES: there are a lot of outliers to the high end here too.

3.8 Variable Transformation / Feature Creation:

Fitness variable was transformed from the default integer datatype to a factor datatype to match the question statement. the as.factor() function was used to accomplish this.

4. Conclusion And Recommendations

4.1 Conclusion:

Based on a sample data of 180 entries for customers of the treadmill product(s) of a retail store called Cardio Good Fitness, we have been able to successfully use R to carry out a successful Preliminary Data Analysis which enatails the Exploration of the dataset, performing various univariate and bivariate analysis, and generation of useful insights and recommendation all in order to help the store target new customers.

We have been able to conclude that the most efficient and fastest method of targeting new customers is by focusing more on TM195 (which has proven so far to be the most favorable product without a doubt) and TM798(whose uniqueness in being a favorite of elite individuals and professionals makes it stand out). Conclusions derived from each data variable are further discussed below:

1. **Age**: middle aged people consist of the largest market share of the products especially TM195 and TM498. However, it would be in the best interest of the store not to try getting older people to buy TM798. Highly educated High income earners can boost the odds greatly.
2. **Gender**: Both gender of people consist of the largest market share of the products especially TM195 and TM498. How ever, it would be in the best interest of the store not to try getting Females to buy TM798.
3. **Education**: well educated people would always go for TM195 and TM498 but very highly educated people would always choose TM798. Having at least a level 16 customer is the benchmark to ensure customer conversion as this variable is one of the strongest factor to consider here.
4. **Marital Status**: both statuses have high impact in improving sales and getting new customers. However, the chance is greater when partnered people are targeted.
5. **Usage**: people who want an average weekly usage consist of the largest market share of the products especially TM195 and TM498. Interestingly, people who want a high weekly usage buy TM798
6. **Fitness**: generally the people that patronise the store are those who are fit, especially for TM195 and TM498. People who are very fit go for TM798. Its best not to engage anyone who is unfit.
7. **Income**: High income earners prefer to buy TM798 over the other products whie low and majorly average income earners go for TM195 and TM498. High income earners seriously ensures customer conversion as this variable is the strongest factor to consider here.
8. **Miles**: People who want to run a very long mileage prefer to buy TM798 over the other products whie the majority of people wanting low and medium mileage go for TM195 and TM498.

4.2 Recommendations:

After a careful exploration of the data provided and an in-depth analysis using Modern Statistical tools, below are the recommendations for "Cardio Good Fitness" Store to help them target New customers:

1. TM198 and TM498 should be given a major campaign targeting middle aged working class people who are fit,
2. TM798 should be uniquely styled and crafted to high taste to favor the elite and the very wealthy,
3. There should be Bonuses for Men on successful acquisition of TM198 and TM468
4. There should be new packages made to encourage partnered people for all products

```
#=====================================================================
#
# T H E - E N D
#
#=====================================================================
```

5. Appendix A – Source Code

```
#Generate the .R file to hold the source code
## 4. Conclusion And Recommendations
purl("project-1.Rmd", documentation = 0)
```

```
## [1] "project-1.R"
```

```
#=====================================================================
#
# Exploratory Data Analysis - CardioGoodFitness
#
#=====================================================================

# Environment Set up and Data Import

# Invoking Libraries
library(tidyverse) # contains ggplot2,dplyr,forcats,lubridate etc
library(gridExtra) # Needed for plotting multiple ggplot graphs side-by-side
library(corrplot) # Needed to plot correlation plots
library(scales) # For Big Number formating
library(knitr) # Necessary to generate sourcecodes from a .Rmd File

# Setup Working Directory
setwd("C:/Users/USER/Documents/El-PaDJo/R programming language/1-Introduction to R & Statistics/week_3")

# Read Input File
cardio_data = read.csv("CardioGoodFitness.csv")

#making all columns accessible without '$' sign usage
attach(cardio_data)

# Global options settings
options(scipen=999)  # turn off scientific notation like 1e+06

# Function to calculate mode
getmode <- function(v) {
  uniqv <- unique(v)
  uniqv[which.max(tabulate(match(v, uniqv)))]
}

# Variable Identification
# check dimension of dataset
dim(cardio_data)

#see first 6 rows(observations) of dataset
head(cardio_data)

#see last 6 rows(observations) of dataset
tail(cardio_data)

# check structure of dataset
```

```r
str(cardio_data)

#Change fitness variable type to factor
cardio_data$Fitness = as.factor(Fitness)

# get summary of dataset
summary(cardio_data)

# Bar Plot for Product Variable
ggplot(cardio_data, aes(x = Product)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Product", x="Products", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Histogram for Age Variable
age_histogram = ggplot(cardio_data, aes(x = Age)) +
  geom_histogram(fill = "cornflowerblue", bins = 33) +
  labs(title="Histogram for Age", x="Age", y="Count") +
  scale_x_continuous(breaks = seq(18, max(Age)+5, by = 5)) +
  scale_y_continuous(breaks = seq(0, 30, by = 5)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box Plot for Age Variable
age_boxplot = ggplot(cardio_data, aes(x = 0, y = Age)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Age ", x = "", y="Age") +
  scale_y_continuous(breaks = seq(18, max(Age)+5, by = 5)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(age_histogram,age_boxplot,nrow=1,ncol= 2)

# Bar Plot for Gender Variable
ggplot(cardio_data, aes(x = Gender)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Gender", x="Gender", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Histogram for Education Variable
edu_histogram = ggplot(cardio_data, aes(x = Education)) +
  geom_histogram(fill = "cornflowerblue", bins = 10) +
   labs(title="Histogram for Education", x="Education Levels", y="Count") +
  scale_x_continuous(breaks = seq(min(Education), max(Education), by = 1)) +
  scale_y_continuous(breaks = seq(0, 100, by = 15)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))
```

```r
# Box for Education Variable
edu_boxplot = ggplot(cardio_data, aes(x = 0, y = Education)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Education ", x = "", y="Education") +
  scale_y_continuous(breaks = seq(min(Education), max(Education), by = 1)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(edu_histogram,edu_boxplot,nrow=1,ncol= 2)

# Bar Plot for Marital Status Variable
ggplot(cardio_data, aes(x = MaritalStatus)) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Marital Status", x="Marital Status", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Histogram for Usage Variable
usage_histogram = ggplot(cardio_data, aes(x = Usage)) +
  geom_histogram(fill = "cornflowerblue", bins = 6) +
   labs(title="Histogram for Usage", x="Usage Levels", y="Count") +
  scale_x_continuous(breaks = seq(min(Usage), max(Usage), by = 1)) +
  scale_y_continuous(breaks = seq(0, 80, by = 10)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box for Usage Variable
usage_boxplot = ggplot(cardio_data, aes(x = 0, y = Usage)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Usage ", x = "", y="Usage") +
  scale_y_continuous(breaks = seq(min(Usage), max(Usage), by = 1)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(usage_histogram,usage_boxplot,nrow=1,ncol= 2)

# Bar Plot for Fitness Variable
ggplot(cardio_data, aes(x = factor(Fitness))) +
  geom_bar(fill = "cornflowerblue", color="black") +
  labs(title="Bar Chart for Fitness", x="Fitness", y="Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Histogram for Income Variable
income_histogram = ggplot(cardio_data, aes(x = Income)) +
  geom_histogram(fill = "cornflowerblue", bins = 50) +
  labs(title="Histogram for Income", x="Income Levels", y="Count") +
```

```
  scale_x_continuous(breaks = seq(20000, 120000, by = 10000), labels = comma) +
  scale_y_continuous(breaks = seq(0, 25, by = 5)) +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(plot.title=element_text(hjust=0.5))

# Box for Income Variable
income_boxplot = ggplot(cardio_data, aes(x = 0, y = Income)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Income ", x = "", y="Income") +
  scale_y_continuous(breaks = seq(20000, 120000, by = 10000), labels = comma) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(income_histogram,income_boxplot,nrow=1,ncol= 2)

# Histogram for Miles Variable
miles_histogram = ggplot(cardio_data, aes(x = Miles)) +
  geom_histogram(fill = "cornflowerblue", bins = 35) +
  labs(title="Histogram for Miles", x="Miles", y="Count") +
  scale_x_continuous(breaks = seq(0, max(Miles)+10, by = 50)) +
  scale_y_continuous(breaks = seq(0, 40, by = 5))+
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Box for Miles Variable
miles_boxplot = ggplot(cardio_data, aes(x = 0, y = Miles)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", fill="cornflowerblue") +
  labs(title = "Box Plot for Miles ", x = "", y="Miles") +
  scale_y_continuous(breaks = seq(0, max(Miles)+10, by = 50)) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(axis.text.y = element_blank()) +
  coord_flip()

# Print charts side-by-side
grid.arrange(miles_histogram,miles_boxplot,nrow=1,ncol= 2)

# Jitter box Plot for Product vs Age
ggplot(cardio_data, aes(x = Product, y = Age, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Age", x = "Products", y = "Age") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()

# Grouped Bar Chart For Product vs Gender
```

```r
ggplot(cardio_data, aes(x = Product, fill = Gender)) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Gender", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))


# Jitter box Plot for Product vs Education
ggplot(cardio_data, aes(x = Product, y = Education, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product vs Education", x = "Products", y = "Education") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()

# Grouped Bar Chart For Product vs Marital Status
ggplot(cardio_data, aes(x = Product, fill = factor(MaritalStatus))) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Marital Status", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Grouped Bar Chart For Product vs Usage Status
ggplot(cardio_data, aes(x = Product, y = Usage, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product vs Usage", x = "Products", y = "Usage") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()

# Grouped Bar Chart For Product vs Marital Status
ggplot(cardio_data, aes(x = Product, fill = factor(Fitness))) +
  geom_bar(position = position_dodge(preserve = "single")) +
  labs(title = "Grouped Bar Chart For Product vs Fitness", x = "Products", y = "Count") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# Jitter box Plot for Product vs Income
ggplot(cardio_data, aes(x = Product, y = Income, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Income", x = "Products", y = "Income") +
  scale_y_continuous(breaks = seq(20000, 120000, by = 5000), labels = comma) +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  coord_flip()
```

```r
# Jitter box Plot for Product vs Age
ggplot(cardio_data, aes(x = Product, y = Miles, color = Product)) +
  geom_boxplot(size=1, outlier.shape = 1, outlier.color = "black", outlier.size  = 3) +
  geom_jitter(alpha = 0.5, width=.2) +
  labs(title = "Jitter Box Plot for Product Vs Miles", x = "Products", y = "Miles") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5)) +
  theme(legend.position = "none") +
  coord_flip()

#Make a Correllation plot of AGE, Education, Usage, Income and miles
corrplot(cor(cardio_data[,c(2,4,6,8,9)]))

# scatterplot with loess smoothed line
# and better labeling and color
ggplot(cardio_data, aes(x = Education, y = Income)) +
  geom_point(color="cornflowerblue", size = 2, alpha = .6) +
  geom_smooth(size = 1.5, color = "tomato") +
  scale_y_continuous(labels = comma, limits = c(10000, 130000)) +
  scale_x_continuous(breaks = seq(min(Education), max(Education), by = 1),
                     limits = c(11, 22)) +
  labs(title = "Scatterplot with Loess Smoothed Line for Education Vs Income",
       x = "Education", y = "Income") +
  theme_minimal() +
  theme(plot.title=element_text(hjust=0.5))

# check if NA field exists in the dataset
anyNA(cardio_data)

#Generate the .R file to hold the source code
#Save all Sourcecodes in a corresponding .R file
purl("project-1.Rmd", documentation = 0)

#=============================================================================
#
# T H E - E N D
#
#=============================================================================

##
##
```