ABSTRACT

"Cardio good fitness" data set is used to generate insights about the data set using R. The main aim is to come up with a customer profile (characteristics of a customer) of the different products. Based on the data we have generated a set of insights and recommendations that will help the company in targeting new customers. This exploration report will consist of importing the dataset in r, understanding the structure of the data set, graphical exploration and representation, descriptive statistics, insights gathered from the dataset. Exploratory data set is used to analyze a dataset and understand its key characteristics, understand the pattern in the data and form hypothesis, identify outliers, missing data, incorrect data.

INDEX

1. PROJECT OBJECTIVES

The objective of the 'Cardio Good Fitness' project 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 the data set
- Graphical exploration and representation
- Descriptive statistics
- Insights gathered from the dataset

2. ASSUMPTIONS

The assumptions made in this report are as follows:

- Dataset has been cleaned; no errors in entries
- Data obtained is not biased or skewed
- Data are normally distributed
- Relationship between independent and dependent variables are linear
- Variables are multivariate normal
- Data has little or no multicollinearity in data

3. EXPLORATORY DATA ANALYSIS

The data exploratory activity consists of the following steps:

3.1 Environment Set up and Data Import
3.2 Variable Identification
3.3 Univariate Analysis
3.4 Bi-Variate Analysis
3.5 Missing Value Treatment
3.6 Outlier Treatment
3.7 Variable Transformation / Feature Creation
3.8 Feature Exploration

3.1 DATA ANALYSIS

3.1.1 Install necessary Packages and invoke Libraries

Packages installed and libraries invoked are:

- rpivotTable
- lattice

3.1.2 Set up working directory

Setting a working directory on starting of the R session makes importing and exporting data and code files easier. The working directory is the location or folder on the PC where the data, codes, etc. related to theproject is being stored.

3.1.3 Import and Read the Dataset

The dataset "CardioGoodFitness" is in the .csv format. Hence, the command 'read.csv' is used for importingthe file. The command 'header = TRUE' is to include the header of the columns from the dataset.

3.2 Variable Identification

Six functions were used in the identification of variables in R. They are 'dim', 'head', 'tail', 'names', 'str', 'summary'.

The 'dim' function was used to identify the dimension of a matrix, array or data frame. In this case it returns the number of rows and columns in the dataset.

The 'head' and 'tail' function was used to view the top and bottom rows of the data set respectively. This is to ensure that there are no formatting issues such as headers or footers in the data set.

The 'names' function was used to get the names of an object. In this case, it returns the headers of the columns from the dataset.

The 'str' function was used to check the data types and structure of the dataset. This shows which variablesare taken as factors or numeric variables.

The 'summary' function was used to produce result summaries of the results of various model fitting functions. In this case, it shows the summary of the various variables.

3.2.1 Variable Identification – Inferences

The 'dim' function returned the number of rows and columns of the dataset, which is 180 and 9 respectively. This confirms that the data set has 180 observations and 9 variables.

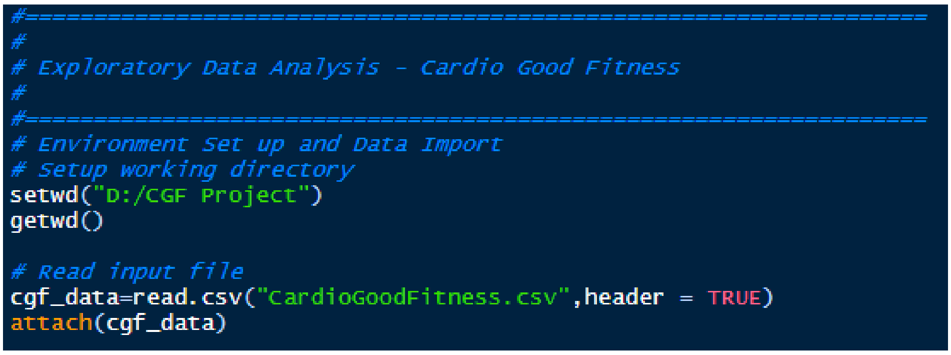
The 'head' and 'tail' function returned the top and bottom few rows of the dataset respectively with the headers/variables included. From this, the dataset has no formatting issues such as headers or footers.

The 'names' function returned the variable names of the dataset. There are 9 variables namely, "Product", "Age", "Gender", "Education", "MaritalStatus", "Usage", "Fitness", "Income" and "Miles".

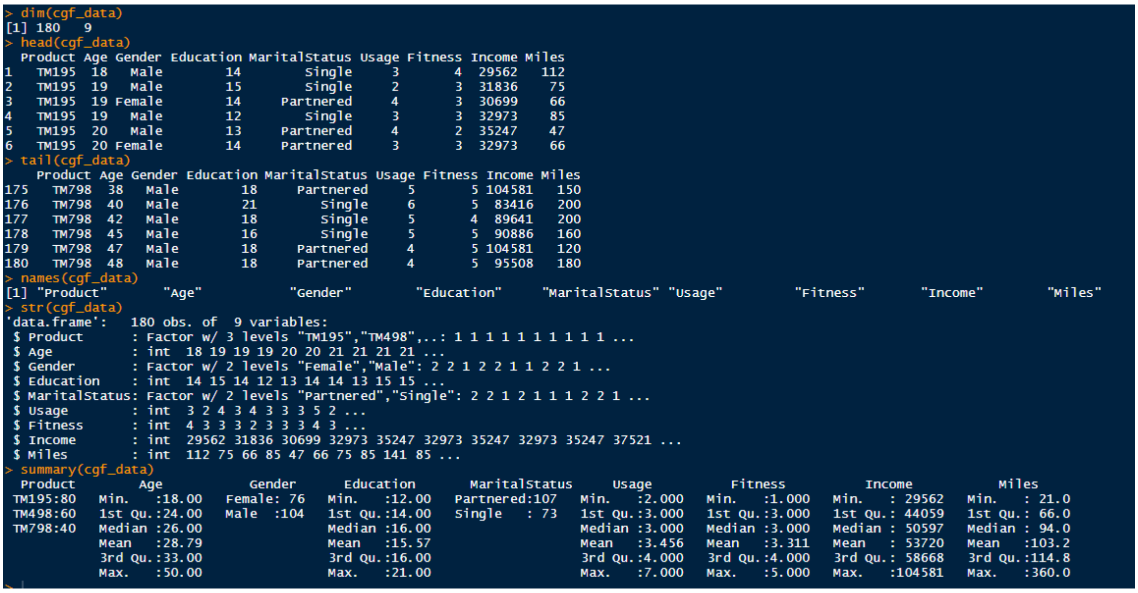
The 'str' function returned information about the data types and structure of the dataset. From this, there are 3 variables that is a 'Factor' type with different levels and are termed as categorical variables. They are the "Product", "Gender", and "Marital Status". The remaining 6 variables are termed as numeric variables with the 'int' type, and they are, "Age", "Education", "Usage", "Fitness", "Income" and "Miles". Thus we are able to identify our variables from the dataset.

The 'summary' function returned a summary of the variables. For those categorical variables, it returnedthe different levels and how many observations. While the numeric variables, it returned the statistical information of the variables such as minimum, 1st quartile, median, mean, 3rd quartile and maximum.

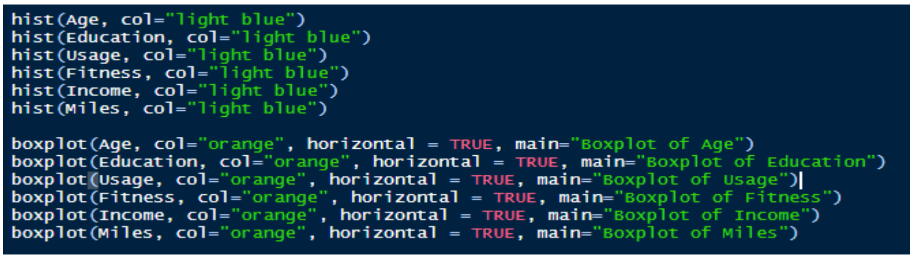
4. Appendix – Source Code



```
#======================================================================
#
# Exploratory Data Analysis - Cardio Good Fitness
#
#======================================================================
# Environment Set up and Data Import
# Setup working directory
setwd("D:/CGF Project")
getwd()

# Read input file
cgf_data=read.csv("CardioGoodFitness.csv",header = TRUE)
attach(cgf_data)
```



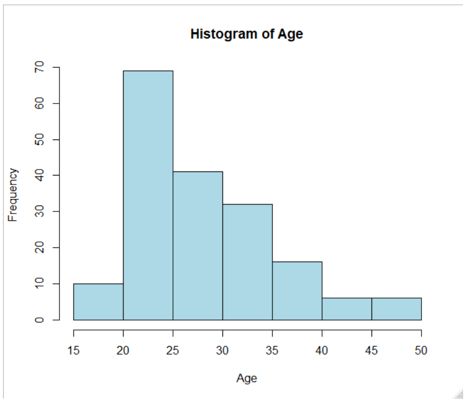
```
> dim(cgf_data)
[1] 180   9
> head(cgf_data)
  Product Age Gender Education MaritalStatus Usage Fitness Income Miles
1   TM195  18   Male        14        Single     3       4  29562   112
2   TM195  19   Male        15        Single     2       3  31836    75
3   TM195  19 Female        14     Partnered     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
> tail(cgf_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
> names(cgf_data)
[1] "Product"       "Age"           "Gender"        "Education"     "MaritalStatus" "Usage"         "Fitness"       "Income"        "Miles"
> str(cgf_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 ...
> summary(cgf_data)
  Product        Age           Gender       Education       MaritalStatus      Usage          Fitness          Income           Miles
 TM195:80   Min.   :18.00   Female: 76   Min.   :12.00   Partnered:107   Min.   :2.000   Min.   :1.000   Min.   : 29562   Min.   : 21.0
 TM498:60   1st Qu.:24.00   Male  :104   1st Qu.:14.00   Single   : 73   1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 44059   1st Qu.: 66.0
 TM798:40   Median :26.00                Median :16.00                   Median :3.000   Median :3.000   Median : 50597   Median : 94.0
            Mean   :28.79                Mean   :15.57                   Mean   :3.456   Mean   :3.311   Mean   : 53720   Mean   :103.2
            3rd Qu.:33.00                3rd Qu.:16.00                   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 58668   3rd Qu.:114.8
            Max.   :50.00                Max.   :21.00                   Max.   :7.000   Max.   :5.000   Max.   :104581   Max.   :360.0
>
```

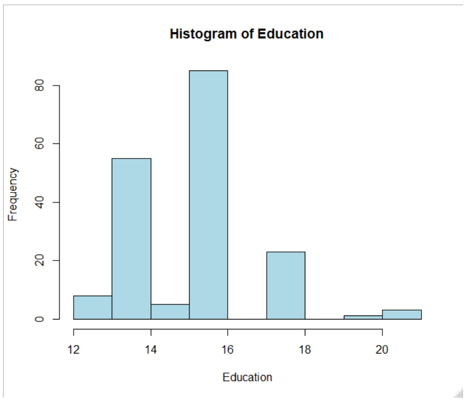


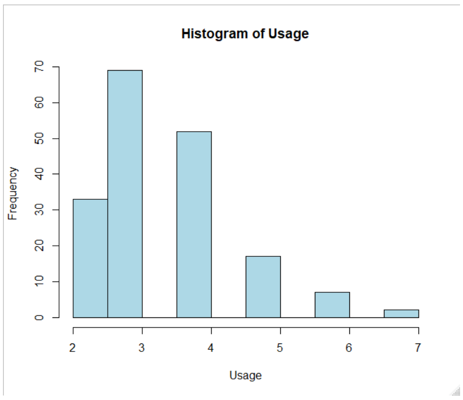
```
hist(Age, col="light blue")
hist(Education, col="light blue")
hist(Usage, col="light blue")
hist(Fitness, col="light blue")
hist(Income, col="light blue")
hist(Miles, col="light blue")

boxplot(Age, col="orange", horizontal = TRUE, main="Boxplot of Age")
boxplot(Education, col="orange", horizontal = TRUE, main="Boxplot of Education")
boxplot(Usage, col="orange", horizontal = TRUE, main="Boxplot of Usage")
boxplot(Fitness, col="orange", horizontal = TRUE, main="Boxplot of Fitness")
boxplot(Income, col="orange", horizontal = TRUE, main="Boxplot of Income")
boxplot(Miles, col="orange", horizontal = TRUE, main="Boxplot of Miles")
```

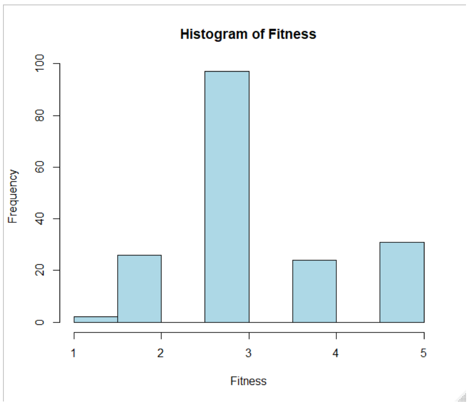
4.1 Univariate Analysis

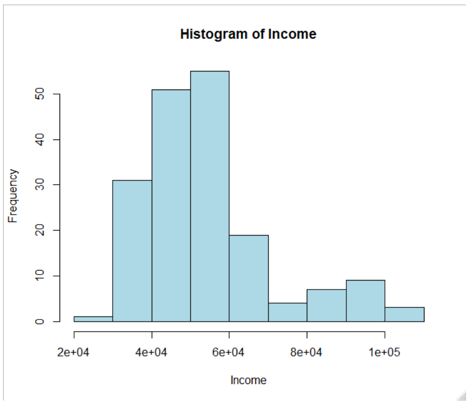
This section explores the individual variables mainly the numeric variables. Using the functions ‘hist’ and ‘boxplot’, the various histograms and boxplot can be generated for individual numeric variables and analysis and inference can be performed.

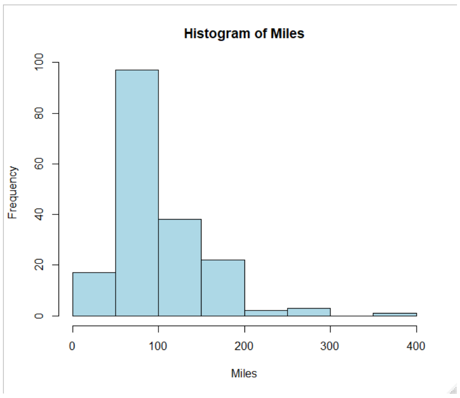




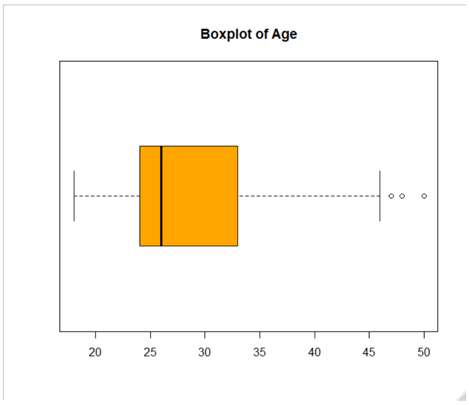


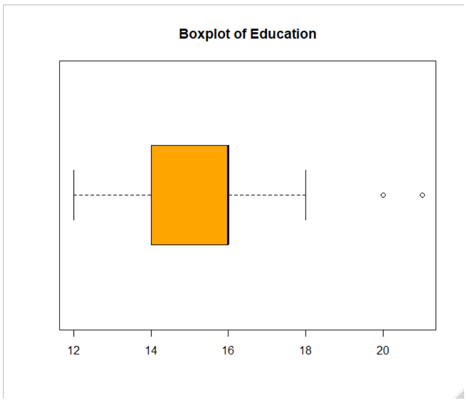


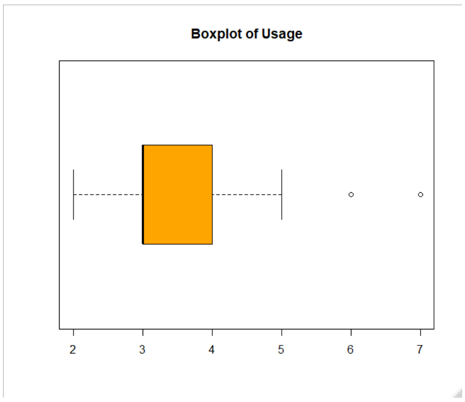




From the histograms, it can be observed that highest frequency of customers are within the age range of 20 to 25, with number of education years of 16, average number of times the customer wants to use the treadmill every week is 3, self-rated fitness level of 3, with income level between 50000 and 60000 and miles covered is between 50 to 100.

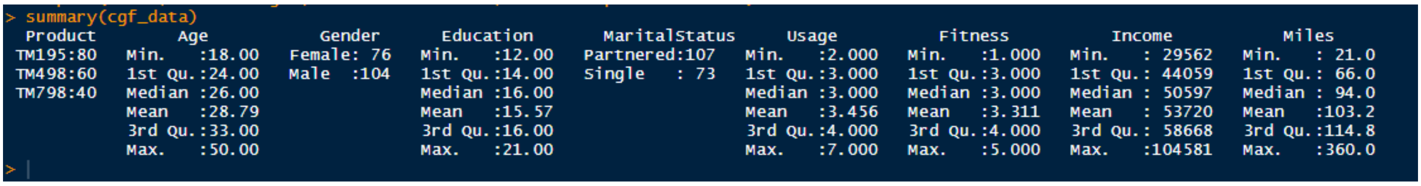






The boxplot shows the minimum, 1st quartile, median, 3rd quartile and maximum values of each individual numeric variables. From the boxplots, it can be observed that more customer’s age falls after the median of 26, their education lesser than the median of 16 years, while the usage of more than the median 3 times a week, self-rated more than the median of 3, with income slightly above the median of 50597 and lesser miles covered below the median of 94.

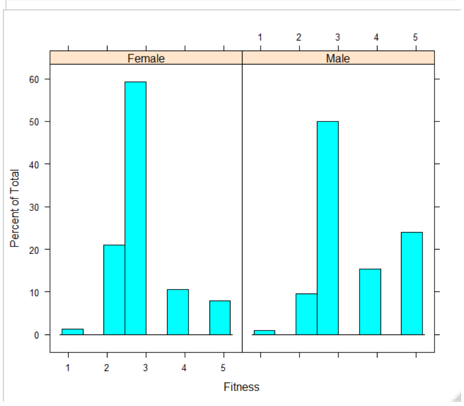
Below is the statisical summary of the variables.

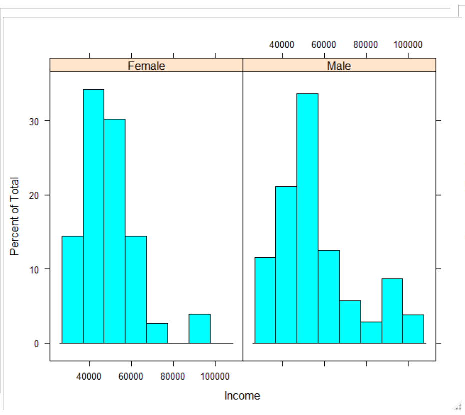


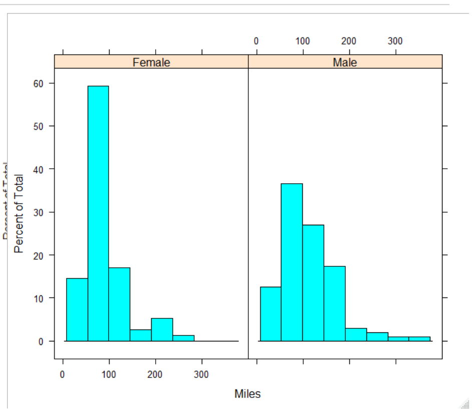
```
> summary(cgf_data)
 Product       Age           Gender       Education       MaritalStatus      Usage          Fitness          Income           Miles      
 TM195:80   Min.   :18.00   Female: 76   Min.   :12.00   Partnered:107   Min.   :2.000   Min.   :1.000   Min.   : 29562   Min.   : 21.0  
 TM498:60   1st Qu.:24.00   Male  :104   1st Qu.:14.00   Single   : 73   1st Qu.:3.000   1st Qu.:3.000   1st Qu.: 44059   1st Qu.: 66.0  
 TM798:40   Median :26.00                Median :16.00                   Median :3.000   Median :3.000   Median : 50597   Median : 94.0  
            Mean   :28.79                Mean   :15.57                   Mean   :3.456   Mean   :3.311   Mean   : 53720   Mean   :103.2  
            3rd Qu.:33.00                3rd Qu.:16.00                   3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.: 58668   3rd Qu.:114.8  
            Max.   :50.00                Max.   :21.00                   Max.   :7.000   Max.   :5.000   Max.   :104581   Max.   :360.0  
> |
```

However, the boxplot also shows outliers represented as circles in the plots. These will be further discussed in the later section on outlier identification.

A further analysis can be further segregated by the 3 categorical variables, namely by “Product”, “Gender” and “MaritalStatus”. The variable “Gender” shall be taken an example and analyzed in this project. The library ‘lattice’ command can be invoked to generate the following plots:





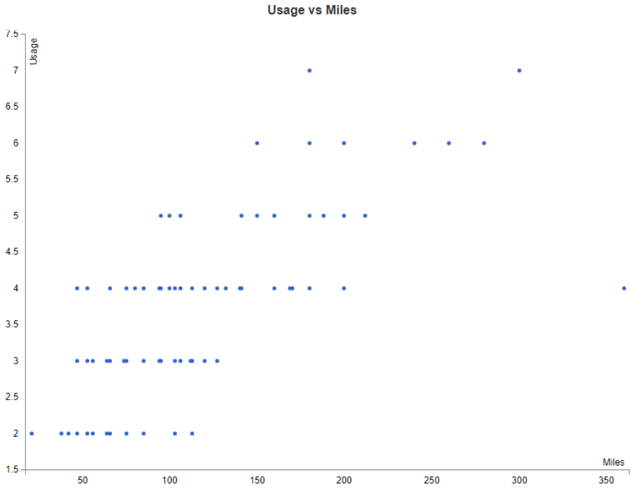


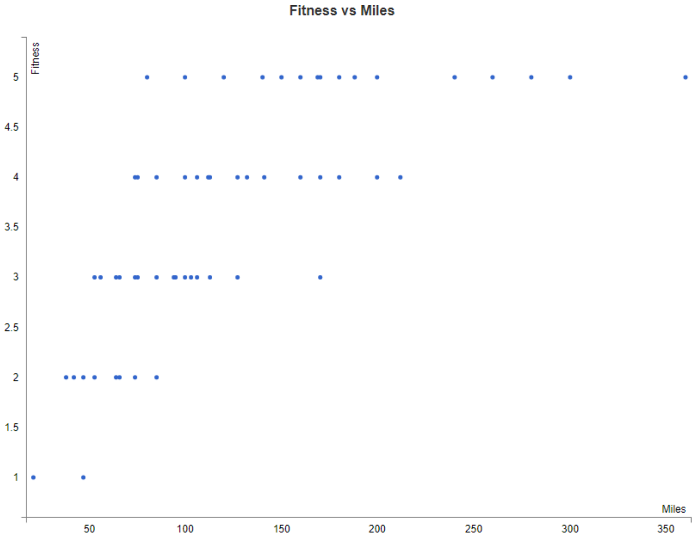
From the histograms, it shows the various information of the age distribution, education, usage of treadmils, fitness rating, income and miles covered between the different genders. Overall, there is not much of a difference betweenthe different genders as in terms of frequency, both genders are similar.

4.2Bi-Variate Analysis

Bivariate analysis is the simultaneous analysis of two variables (attributes). It explores the concept of relationship between two variables, whether there is a correlation and the strength of this correlation, or whether there are differences between two variables and the significance of these differences.

A scatter plot can be plotted to view the relationship between the two variables. The library ‘rpivotTable’ commandcan be invoked to draw the scatter plot. The correlation can be calculated using the ‘cor’ command. Below are someexamples of variables and their scatter plots





The correlation between Usage and Miles are 76% and between Fitness and Miles are 79% thus these variables arequite correlated with each other. A linear model can be created between the variables and the equations are as follows:

- Miles Covered = -22.22 + 36.294 * Usage; (R^2 = 0.5739)
- Miles Covered = -37.519 + 42.497 * Fitness; (R^2 = 0.6152)
- Miles Covered = -56.743 + 27.206 * Fitness + 20.215 * Usage; (R^2 = 0.713)

By observing the coefficient of determination, R^2 value, the equation: Miles Covered = -56.743 + 27.206 * Fitness + 20.215 * Usage is the more accurate equation to use in predicting the miles covered by the customer.

4.3 Missing Value Identification

To identify whether there are missing values in the dataset, the command 'anyNA' can be used. However, in this dataset it returns the value FALSE. This means that there are no missing values in this dataset.

4.4 Outlier Identification

An outlier is a data point that differs significantly from other observations. As seen from the boxplots in section 3.3, the outliers can be identified by the circles beyond the maximum and before minimum of the boxplots. Outliers can be determined the formulas: if it's lesser than Q1 – 1.5*IQR or more than Q3 + 1.5*IQR, where Q1 is the 1st quartile,Q3 is the 3rd quartile and IQR is the interquartile range.

4.5Variable Transformation / Feature Creation

While the information in the dataset are comprehensive enough and a linear equation with a relatively high coefficient of determination. More information / variables could be obtained such as heart rate; number of steps, time spent in hours, calories burnt/intensity of exercise, sleeping patterns, etc. to have a better understanding of thedata and also better predict the customer's behavior and patterns, and advise customers to help optimize their fitness regime. In turn, this could be used to help maximize the sales and profit of Cardio Good Fitness in their sales of treadmill products to the customers.

5 Conclusion

In conclusion, the dataset provided a comprehensive understanding about the customers behavior and consumption patterns. There is a strong correlation between miles covered, their fitness ratings and usage of the treadmill which can be predicted with a linear model of relatively high coefficient of determination, Miles Covered =
-56.743 + 27.206 * Fitness + 20.215 * Usage. This would be a useful model to help present and explain the results to the customers and advise customers to help optimize their fitness regime. However, more variables could be included such as heart rate, number of steps, time spent in hours, calories burnt/intensity of exercise, sleeping patterns, etc. to further understand and generate a more accurate model to predict the customers behavior and consumption patterns. In turn, Cardio Good Fitness can utilize these data to help maximize their sales and profits from the sales of their treadmill products.

6 References

www.kaggle.com