Project - Cardio Fitness

Jan 2020 Author: Snehal Gawand

Table of Contents

1. Project Objective

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

Importing the dataset in R
Understanding the structure of dataset
Graphical exploration
Descriptive statistics
Insights from the dataset

2. Assumptions

This dataset contains data about customers performing cardio exercises using treadmill. Age, Education, Usage, Fitness, Income and Miles are integers and rest are categorical.

3. Exploratory Data Analysis – Step by step approach

3.1 Environment Set up and Data Import

3.1.1 Install necessary Packages and Invoke Libraries

For performing analysis on ''CardioGoodFitness" dataset, the below mentioned packages are installed and their corresponding libraries are invoked

- readr (For reading the csv file)
- ggplot2. (For plotting the graphs)
- corrplot (To identify relation between columns/variables)

Please refer Appendix A for Source Code.

3.1.2 Set up working Directory

In order to import the data from "CardioGoodFitness" dataset and performing the analysis, the location/folder on the computer where this dataset is stored is set as working directory in R. This location can further used to store the new/updated dataset.

Please refer Appendix A for Source Code.

3.1.3 Import and Read the Dataset

The given dataset is in .csv format. Hence, the command 'read.csv' is used for importing the file.

Please refer Appendix A for Source Code.

3.2 Variable Identification

3.2.1 Variable Identification – Inferences

The below listed functions are used to explore the variables used in the dataset

- dim → to check amount of data present in input dataset
- nrow and ncol → can also be used to check amount of data (no of observations and columns)
- names → list down names of columns used in dataset
- str → view structure of the dataset
- summary → to check the summary of each column of dataset

The result of these functions is described as follows:

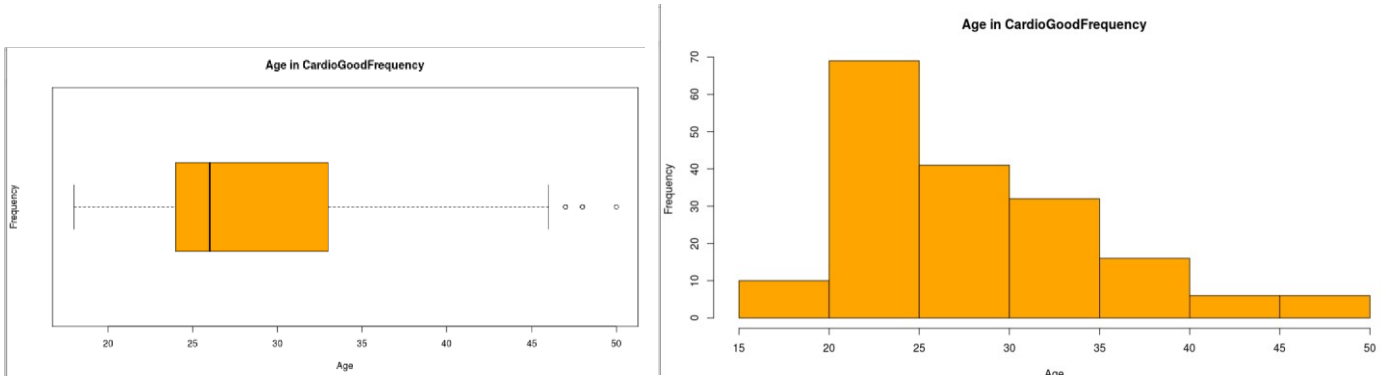
- The dataset consists of data of customers using the three types of treadmill with product codes TM195, TM498 and TM798
- The minimum age of customers is 18 whereas the maximum age is 50
- The treadmills are used by 76 females and 104 males. i.e. 57.77% are males and 42.22% are females
- Out of the total population of customers, 107 are partnered and 73 are single. This indicates, the customers interested in working out using treadmill are partnered
- The dataset is clean with very less outliers in Age, Usage, Fitness, Income and miles columns

Please refer Appendix A for Source Code.

3.3 Univariate Analysis

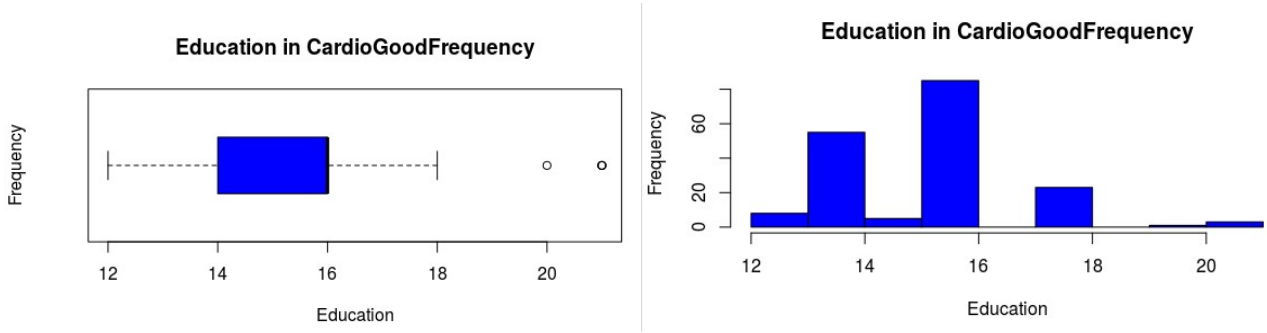
The univariate analysis is performed on the categorical variables to check for the frequency of occurrence in dataset.

- **AGE**



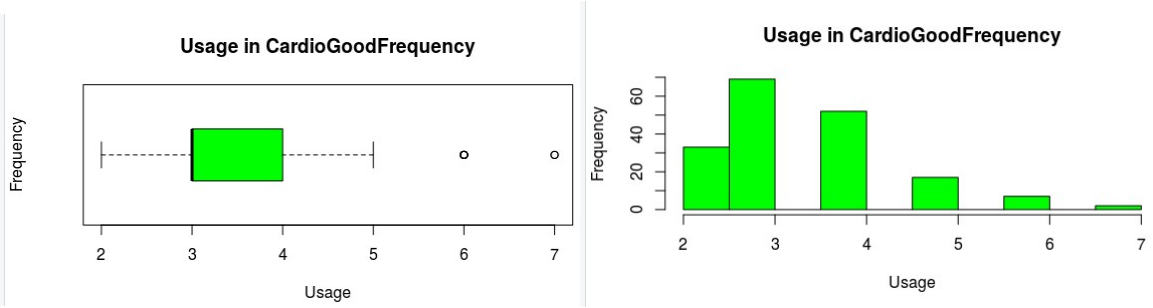
- Graph illustrates approximately 70 customers using treadmill, are between age group 20 to 25.

- **EDUCATION**



Most of the users have Education level of 15 to 16. Very few users have higher education level(19 and above)

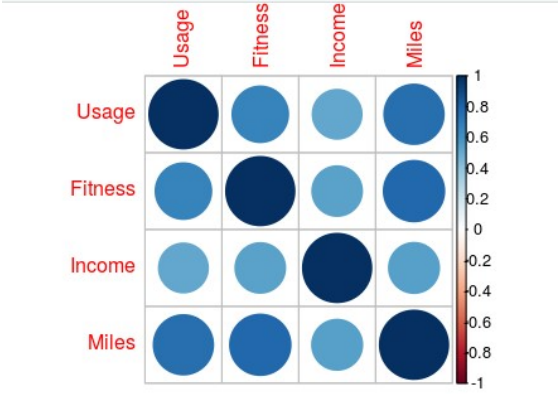
- **USAGE**

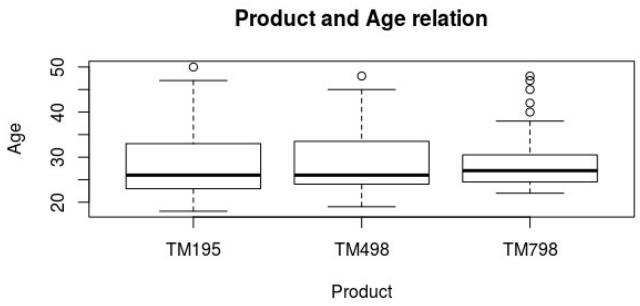


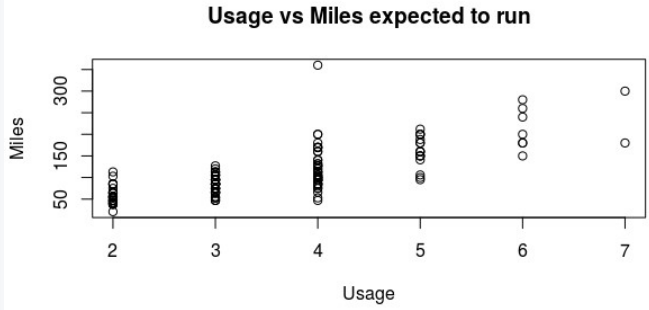
The customers want to use the cardio treadmill at an average 3 times in a week

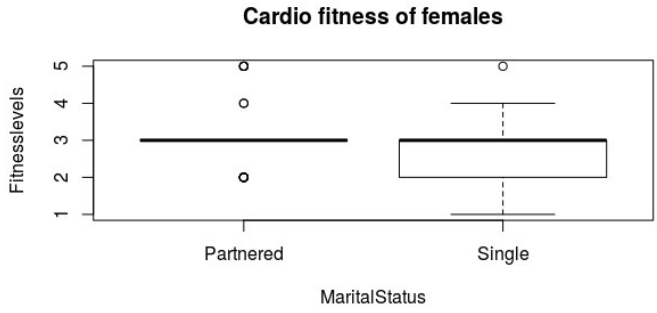
3.4 Bi-Variate Analysis

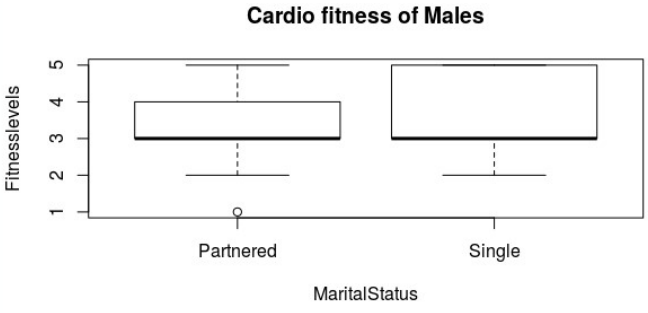
We are using the 'corr' function in R to identify the correlated variables











These graphs illustrates the following:

- The treadmills with product code TM195 and TM498 are widely used by customers between age group 20 to 25 whereas treadmills with product code TM798 are used by customers between age group 25 to 30
- The number of miles expected to run by customer is directly proportional to the usage of treadmill • Partnered females have average fitness score 3, whereas singles have fitness score between 2 and 3. This shows Single females are more inclined towards fitness
- Partnered males have fitness score between 3 and 4 whereas Singles have between 3 and 5. Single males are more fitter than partnered

- The usage of three types of treadmill is almost equally distributed amongst Male customers, whereas 50% females use TM195.

- TM195 is the popular type of treadmill amongst all the customers
- Male customers are more inclined towards fitness as they cover more miles as compared to females

3.5 Missing Value Identification

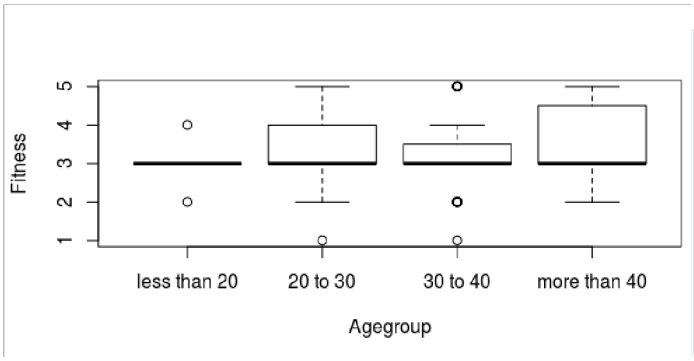
There are no missing values in the dataset. All the columns have complete information about the customers.

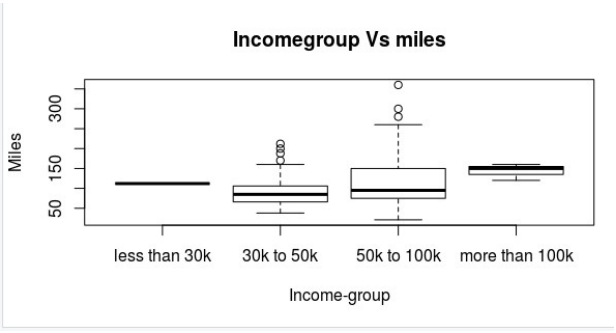
3.6 Outlier Identification

Few outliers in Miles covered by males. Partnered females have almost fitness score of 3, with few outliers having score 2,4,and 5

3.7 Variable Transformation / Feature Creation

Two new variables Age-group and Income group are added to understand the relation between variables in more detailed manner.





- More than 40 age group has high fitness score although the number of users in this age group is less. 20 to 30 age group has max fitness score 4
- Most of the customers are between income group 50k to 100k with avg expectation of 100 miles to run
- Users with income-group 50k to 100k have fitness score between 3 and 5, whereas users with income-group 30k to 50k have fitness score 3

4. Conclusion

- The dataset shows the treadmill is used mostly by Male customers

- The level of fitness is higher as the income of customers is increased. Also the fitness levels and Miles achieved are linear.
- Customers use the treadmill at an average 3 times a week.
- Single male and female users are more fitter than partnered
- Customers within age group 20 and 30 have maximum fitness level of 4.

- As TM195 is the most popular product amongst all customers, the count of this product can be increased to attract new customers.

5. Appendix A – Source Code

```
#########################################################################
############                    SOURCE CODE                  ##################
#########################################################################

#install packages and invoke libraries
install.packages("readr") library("readr")

install.packages("ggplot2") library("ggplot2")

install.packages("corrplot") library("corrplot")

#setup working directory
setwd = ('/Users/snehal/Documents/Projects/Project 1-Week 3') getwd()

#read input CardioGoodFitness.csv dataset cgf_data
<- read.csv("CardioGoodFitness.csv")

#variable Identification
#Check how much data is present in input dataset
dim(cgf_data)   #[1] 180   9     #dataset has 180 observations and 9 variables
nrow(cgf_data) # [1]  180       #180 rows  ncol(cgf_data)  # [1]  9          #9
cols
#total number of records : 180

#List down the variable names in dataset names(cgf_data)
[1] "Product"        "Age"            "Gender"         "Education"      "MaritalStatus"
[6] "Usage"          "Fitness"        "Income"         "Miles"

#View top and bottom rows  head(cgf_data)
tail(cgf_data)
```

```
#check structure 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 ...
```

#check type of data class(Product)
#Product is factor class(Age) #Age is integer class(Gender) #Gender is factor
class(Education) #Education is integer
class(MaritalStatus) #MaritalStatus is factor
class(Usage) #Usage is integer
class(Fitness) #Fitness is integer
class(Income) #Income is integer
class(Miles) #Miles is integer

#summary statistics summary(cgf_data)

```
Product        Age            Gender      Education       MaritalStatus     Usage
 TM195:80   Min.   :18.00   Female: 76   Min.   :12.00   Partnered:107   Min.   :2.000
 TM498:60   1st Qu.:24.00   Male  :104   1st Qu.:14.00   Single   : 73   1st Qu.:3.000
TM798:40   Median :26.00                Median :16.00                   Median :3.000
            Mean   :28.79                Mean   :15.57                   Mean   :3.456
            3rd Qu.:33.00                3rd Qu.:16.00                   3rd Qu.:4.000
            Max.   :50.00                Max.   :21.00                   Max.   :7.000
    Fitness          Income          Miles
 Min.   :1.000   Min.   : 29562   Min.   : 21.0
1st Qu.:3.000   1st Qu.: 44059   1st Qu.: 66.0
 Median :3.000   Median : 50596   Median : 94.0
 Mean   :3.311   Mean   : 53720   Mean   :103.2
 3rd Qu.:4.000   3rd Qu.: 58668   3rd Qu.:114.8
 Max.   :5.000   Max.   :104581   Max.   :360.0
```

```
##############################
#Exploratory data analysis  #
##############################
```

total_users = nrow(cgf_data) total_users

```
[1] 180
```

```
#Percentage of males
no_of_males = nrow(cgf_data[which(cgf_data$Gender == 'Male'),])
cat("Number of male users",no_of_males) Number
of male users 104

cat("Percentage of male users",no_of_males /total_users*100 ,"%") Percentage
of male users 57.77778 %

#Percentage of females
no_of_females = nrow(cgf_data[which(cgf_data$Gender == 'Female'),])
cat("Number of Female users",no_of_females) Number
of Female users 76

cat("Percentage of female users",no_of_females /total_users*100 ,"%") Percentage
of female users 42.22222 %

########################
#       univariate analysis       #
########################

#Age: boxplot(cgf_data$Age,main = "Age in
CardioGoodFrequency",col =
"orange",xlab="Age",ylab="Frequency",horizontal=TRUE)

hist(cgf_data$Age,main = "Age in CardioGoodFrequency",col =
"orange",xlab="Age",ylab="Frequency",horizontal=TRUE)

#Education: boxplot(cgf_data$Education,main = "Education in
CardioGoodFrequency",col =
"blue",xlab="Education",ylab="Frequency",horizontal=TRUE)

hist(cgf_data$Education,main = "Education in CardioGoodFrequency",col =
"blue",xlab="Education",ylab="Frequency",horizontal=TRUE)


#Usage:
boxplot(cgf_data$Usage,main = "Usage in CardioGoodFrequency",col =
"green",xlab="Usage",ylab="Frequency",horizontal=TRUE)

hist(cgf_data$Usage,main = "Usage in CardioGoodFrequency",col =
"green",xlab="Usage",ylab="Frequency",horizontal=TRUE)


#########################
#          Bi variate analysis          #
#########################
install.packages("corrplot") library(corrplot)
corrplot(cor(cgf_data[,6:9]))
```

```
#relationship between product and age
plot(cgf_data$Product,cgf_data$Age,xlab="Product",ylab="Age")


#relationship between Usage and Miles
plot(cgf_data$Usage,cgf_data$Miles,xlab="Usage",ylab="Miles")


#Analysis of cardiofitness for females
cgf_female = cgf_data[which(cgf_data$Gender =="Female"),] summary(cgf_female)

hist(cgf_female$Fitness,col="green") hist(cgf_female$Age,col="blue")
plot(cgf_female$MaritalStatus,cgf_female$Fitness)

#Analysis of cardiofitness for males
cgf_male = cgf_data[which(cgf_data$Gender=="Male"),] summary(cgf_male)
hist(cgf_male$Fitness,col="green") hist(cgf_male$Age,col="blue")
plot(cgf_male$MaritalStatus,cgf_male$Fitness)

#Analysis of cardiofitness for singles
cgf_single = cgf_data[which(MaritalStatus=="Single"),]
summary(cgf_single) hist(cgf_single$Fitness,col="green")
hist(cgf_single$Age,col="blue")
plot(cgf_single$MaritalStatus,cgf_single$Fitness)

#product and gender based usage
plot(cgf_data$Gender,cgf_data$Product)

#gender and miles
plot(cgf_data$Gender,cgf_data$Miles)

 #Missing values
 anyNA(cgf_data)
 [1]  FALSE


#add new variable age-group
cgf_data$Agegroup <- cut(cgf_data$Age, breaks = c(0,20,30,40,50), labels = c("less than 20", "20 to 30","30 to
40","more than 40"))
```

```
#relationship between age-group and fitness
plot(cgf_data$Agegroup,cgf_data$Fitness,xlab="Agegroup",ylab="Fitness")

# add new variable income-group
cgf_data$Incomegroup <- cut(cgf_data$Income, breaks = c(0,30000,50000,100000,200000), labels = c("less than
30k", "30k to 50k","50k to 100k","more than 100k"))

#relationship between incomegroup vs miles
plot(cgf_data$Incomegroup,cgf_data$Miles,main = "Incomegroup vs miles",xlab="Incomegrp",ylab="Miles")

#relationship between incomegroup vs fitness
plot(cgf_data$Incomegroup,cgf_data$Fitness,main = "Incomegroup vs
fitness",xlab="Incomegroup",ylab="Fitness")


write.csv(cgf_data,"derivedcgf_data.csv")

#########################################################################
############                    END OF  CODE                  ##################
#########################################################################
```