The treadmill problem

Sagar Kalauni

2023-08-22

options(repos = "https://cran.rstudio.com/")

**Cardio Good Fitness [The treadmill problem]**

The market research team is assigned the task of identifying the profile of the typical customer for each treadmill product offered by Cardio Good Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a Cardio Good Fitness retail store during the prior three months. The data are stored in the CardioGoodFitness.csv file. The team identifies the following customer variables to study: product purchased, TM195, TM498, or TM798; gender; age in years; education in years; relationship status, single or partnered; annual household income ($); average number of times the customer plans to use the treadmill each week; average number of miles the customer expects to walk/run each week; and self-rated fitness on an 1-to-5 scale, where 1 is poor shape and 5 is excellent shape. Perform descriptive analytics to create a customer profile for each Cardio Good Fitness treadmill product line.

**Dataset Information :**

cardiogoodfitness.csv: The csv contains data related to customers who have purchased different model from Cardio Good Fitness : Product - the model no. of the treadmill Age - in no of years, of the customer Gender - of the customer Education - in no. of years, of the customer Marital Status - of the customer Usage - Avg. # 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

## Objective

Come up with a customer profile (characteristics of a customer) of the different products Based on the data we have to generate a set of insights and recommendations that will help the company in targeting new customers

# ASK

Questions to be answered

1. How many models does store have?
2. Which is most sold Model?
3. Are Male customers buying treadmill more than female customers?
4. What is the Income ,Age , Education of people buying treadmill?
5. How many days and miles customer expect to run on treadmill?
6. What is the self rated fitness of customers buying treadmill?
7. Are married customer buying Treadmill more than Single customers?
8. Is there any relation between Income and model?
9. Is there any relation between Age and model? Is there any relation between self rated fitness and model?
10. Is there any relation between education and model?
11. Does gender has any effect on model customer buy?
12. Does Martial status has any effect model customer buy?
13. Is there different age groups buying different models?
14. Relation between Age, Income and education and model bought?

# Prepare

# Installing the necessary packages for the data-manupulation and data visualization  
install.packages("tidyverse")

## Installing package into 'C:/Users/Dell/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'tidyverse' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Dell\AppData\Local\Temp\RtmpWAgWc9\downloaded\_packages

install.packages('skimr')

## Installing package into 'C:/Users/Dell/AppData/Local/R/win-library/4.3'  
## (as 'lib' is unspecified)

## package 'skimr' successfully unpacked and MD5 sums checked  
##   
## The downloaded binary packages are in  
## C:\Users\Dell\AppData\Local\Temp\RtmpWAgWc9\downloaded\_packages

library(tidyverse)

## Warning: package 'tidyverse' was built under R version 4.3.1

## Warning: package 'ggplot2' was built under R version 4.3.1

## Warning: package 'lubridate' was built under R version 4.3.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.2 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1

## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors

library(skimr)

## Warning: package 'skimr' was built under R version 4.3.1

library(gridExtra)

## Warning: package 'gridExtra' was built under R version 4.3.1

##   
## Attaching package: 'gridExtra'  
##   
## The following object is masked from 'package:dplyr':  
##   
## combine

mydata= read\_csv("C:/Users/Dell/Desktop/Foundation of data Science- STAT 560/Project\_Stat\_560/Data\_project\_stat\_560/CardioGoodFitness.csv")

## Rows: 180 Columns: 9  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (3): Product, Gender, MaritalStatus  
## dbl (6): Age, Education, Usage, Fitness, Income, Miles  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

Having a first look at my dataset

glimpse(mydata)

## Rows: 180  
## Columns: 9  
## $ Product <chr> "TM195", "TM195", "TM195", "TM195", "TM195", "TM195", "T…  
## $ Age <dbl> 18, 19, 19, 19, 20, 20, 21, 21, 21, 21, 22, 22, 22, 22, …  
## $ Gender <chr> "Male", "Male", "Female", "Male", "Male", "Female", "Fem…  
## $ Education <dbl> 14, 15, 14, 12, 13, 14, 14, 13, 15, 15, 14, 14, 16, 14, …  
## $ MaritalStatus <chr> "Single", "Single", "Partnered", "Single", "Partnered", …  
## $ Usage <dbl> 3, 2, 4, 3, 4, 3, 3, 3, 5, 2, 3, 3, 4, 3, 3, 3, 2, 4, 4,…  
## $ Fitness <dbl> 4, 3, 3, 3, 2, 3, 3, 3, 4, 3, 3, 2, 3, 3, 1, 3, 3, 3, 3,…  
## $ Income <dbl> 29562, 31836, 30699, 32973, 35247, 32973, 35247, 32973, …  
## $ Miles <dbl> 112, 75, 66, 85, 47, 66, 75, 85, 141, 85, 85, 66, 75, 75…

**Observation:**

From above, we can clearly see that my data set has total of 180 observation and 9 fields.

Getting more familier with my data

head(mydata) # this code is used to see the head of my dataset

## # A tibble: 6 × 9  
## Product Age Gender Education MaritalStatus Usage Fitness Income Miles  
## <chr> <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 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(mydata) # Similarly this code is used to see the tail of my dataset

## # A tibble: 6 × 9  
## Product Age Gender Education MaritalStatus Usage Fitness Income Miles  
## <chr> <dbl> <chr> <dbl> <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 TM798 38 Male 18 Partnered 5 5 104581 150  
## 2 TM798 40 Male 21 Single 6 5 83416 200  
## 3 TM798 42 Male 18 Single 5 4 89641 200  
## 4 TM798 45 Male 16 Single 5 5 90886 160  
## 5 TM798 47 Male 18 Partnered 4 5 104581 120  
## 6 TM798 48 Male 18 Partnered 4 5 95508 180

View(mydata) # To view how my dataset in general look like

str(mydata)

## spc\_tbl\_ [180 × 9] (S3: spec\_tbl\_df/tbl\_df/tbl/data.frame)  
## $ Product : chr [1:180] "TM195" "TM195" "TM195" "TM195" ...  
## $ Age : num [1:180] 18 19 19 19 20 20 21 21 21 21 ...  
## $ Gender : chr [1:180] "Male" "Male" "Female" "Male" ...  
## $ Education : num [1:180] 14 15 14 12 13 14 14 13 15 15 ...  
## $ MaritalStatus: chr [1:180] "Single" "Single" "Partnered" "Single" ...  
## $ Usage : num [1:180] 3 2 4 3 4 3 3 3 5 2 ...  
## $ Fitness : num [1:180] 4 3 3 3 2 3 3 3 4 3 ...  
## $ Income : num [1:180] 29562 31836 30699 32973 35247 ...  
## $ Miles : num [1:180] 112 75 66 85 47 66 75 85 141 85 ...  
## - attr(\*, "spec")=  
## .. cols(  
## .. Product = col\_character(),  
## .. Age = col\_double(),  
## .. Gender = col\_character(),  
## .. Education = col\_double(),  
## .. MaritalStatus = col\_character(),  
## .. Usage = col\_double(),  
## .. Fitness = col\_double(),  
## .. Income = col\_double(),  
## .. Miles = col\_double()  
## .. )  
## - attr(\*, "problems")=<externalptr>

**Observation:**

From above we can see that variables like Age, Education, Usage, Fitness, Income, Miles are Numerical type But the variables like Product, Gender,MaritalStatus are string type.

#changing it to object dtype to category to save memory  
mydata$Product = as.factor(mydata$Product)  
mydata$Gender = as.factor(mydata$Gender)  
mydata$MaritalStatus = as.factor(mydata$MaritalStatus)

Converting string data to categorical variables in data analysis enhances memory efficiency, computational speed, and visualization while facilitating statistical analysis.

colnames(mydata) # to look at the column name of our dataset.

## [1] "Product" "Age" "Gender" "Education"   
## [5] "MaritalStatus" "Usage" "Fitness" "Income"   
## [9] "Miles"

**Observation:**

In exploring my dataset, I discovered it contains 9 observations: “Product,” “Age,” “Gender,” “Education,” “MaritalStatus,” “Usage,” “Fitness,” “Income,” and “Miles,” with a total of 180 entries.

Upon examining the variables in mydata dataset, we categorize them as follows:

* Product: Categorical data
* Age: Numerical data
* Gender: Categorical data
* Education: Numerical data
* MaritalStatus: Categorical data
* Usage: Numerical data
* Fitness: Categorical data
* Income: Numerical data
* Miles: Numerical data

Checking for the unique data in each column

unique(mydata$Product)

## [1] TM195 TM498 TM798  
## Levels: TM195 TM498 TM798

unique(mydata$Age)

## [1] 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 43  
## [26] 44 46 47 50 45 48 42

unique(mydata$Gender)

## [1] Male Female  
## Levels: Female Male

unique(mydata$Education)

## [1] 14 15 12 13 16 18 20 21

unique(mydata$MaritalStatus)

## [1] Single Partnered  
## Levels: Partnered Single

unique(mydata$Usage)

## [1] 3 2 4 5 6 7

unique(mydata$Fitness)

## [1] 4 3 2 1 5

unique(mydata$Income)

## [1] 29562 31836 30699 32973 35247 37521 36384 38658 40932 34110  
## [11] 39795 42069 44343 45480 46617 48891 53439 43206 52302 51165  
## [21] 50028 54576 68220 55713 60261 67083 56850 59124 61398 57987  
## [31] 64809 47754 65220 62535 48658 54781 48556 58516 53536 61006  
## [41] 57271 52291 49801 62251 64741 70966 75946 74701 69721 83416  
## [51] 88396 90886 92131 77191 52290 85906 103336 99601 89641 95866  
## [61] 104581 95508

unique(mydata$Miles)

## [1] 112 75 66 85 47 141 103 94 113 38 188 56 132 169 64 53 106 95 212  
## [20] 42 127 74 170 21 120 200 140 100 80 160 180 240 150 300 280 260 360

Looking for is there any cell in the dataset has the null value. If the null values are more then 2% of the total data, it is hard to perform calculation

sum(is.na(mydata)== 'TRUE') # this code check for the null value in the each cell of the dataset and if there is null value it will count it as 1

## [1] 0

we will try to make our Analysis process more systematics. Let’s first look the data summary of all three kind of product we have i.e TM195,TM498, TM798 models of the treadmill.

# Process

## TM195 Data Preprocessing:

Customer profile using model TM195

library(skimr)  
mydata\_TM195 = mydata %>% filter(Product == "TM195")  
skim\_without\_charts(mydata\_TM195)

Data summary

|  |  |
| --- | --- |
| Name | mydata\_TM195 |
| Number of rows | 80 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Product | 0 | 1 | FALSE | 1 | TM1: 80, TM4: 0, TM7: 0 |
| Gender | 0 | 1 | FALSE | 2 | Fem: 40, Mal: 40 |
| MaritalStatus | 0 | 1 | FALSE | 2 | Par: 48, Sin: 32 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 28.55 | 7.22 | 18 | 23 | 26 | 33 | 50 |
| Education | 0 | 1 | 15.04 | 1.22 | 12 | 14 | 16 | 16 | 18 |
| Usage | 0 | 1 | 3.09 | 0.78 | 2 | 3 | 3 | 4 | 5 |
| Fitness | 0 | 1 | 2.96 | 0.66 | 1 | 3 | 3 | 3 | 5 |
| Income | 0 | 1 | 46418.03 | 9075.78 | 29562 | 38658 | 46617 | 53439 | 68220 |
| Miles | 0 | 1 | 82.79 | 28.87 | 38 | 66 | 85 | 94 | 188 |

**Observations:**

80 people purchased the TM195 model treadmill. The average age of these customers is 28.6 years, with a median age of 26 years. The data is right-skewed, indicating the presence of older customers. On average, customers have 15 years of education, but the median education level is 16 years. Customers plan to use the treadmill about 3 days per week. The average expected weekly running distance is 82.8 miles, with a median distance of 85 miles. Customers rated their fitness level as an average of 3.09, representing an average fitness level. The average income and median income are around $46.5K.

## TM498 Data Preprocessing:

Customer profile using model TM498

mydata\_TM498 = mydata %>% filter(Product == "TM498")  
skim\_without\_charts(mydata\_TM498)

Data summary

|  |  |
| --- | --- |
| Name | mydata\_TM498 |
| Number of rows | 60 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Product | 0 | 1 | FALSE | 1 | TM4: 60, TM1: 0, TM7: 0 |
| Gender | 0 | 1 | FALSE | 2 | Mal: 31, Fem: 29 |
| MaritalStatus | 0 | 1 | FALSE | 2 | Par: 36, Sin: 24 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 28.90 | 6.65 | 19 | 24.0 | 26.0 | 33.25 | 48 |
| Education | 0 | 1 | 15.12 | 1.22 | 12 | 14.0 | 16.0 | 16.00 | 18 |
| Usage | 0 | 1 | 3.07 | 0.80 | 2 | 3.0 | 3.0 | 3.25 | 5 |
| Fitness | 0 | 1 | 2.90 | 0.63 | 1 | 3.0 | 3.0 | 3.00 | 4 |
| Income | 0 | 1 | 48973.65 | 8653.99 | 31836 | 44911.5 | 49459.5 | 53439.00 | 67083 |
| Miles | 0 | 1 | 87.93 | 33.26 | 21 | 64.0 | 85.0 | 106.00 | 212 |

Observation:

60 customers purchased the TM498 model treadmill. The average age of these customers is 28.9 years, with a median age of 26 years. The data displays a right-skewed distribution, indicating the presence of older customers. On average, customers have 15.1 years of education, while the median education level remains at 16 years. Customers intend to use the treadmill around 3 days per week. The average expected weekly running distance is 87.9 miles, with a median value of 85 miles. These customers have an average income of $48,974, and the median income is (dollor) 49,460.

## TM798 Data Preprocessing:

Customer profile using model TM798

mydata\_TM798 = mydata %>% filter(Product == "TM798")  
skim\_without\_charts(mydata\_TM798)

Data summary

|  |  |
| --- | --- |
| Name | mydata\_TM798 |
| Number of rows | 40 |
| Number of columns | 9 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Column type frequency: |  |
| factor | 3 |
| numeric | 6 |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ |  |
| Group variables | None |

**Variable type: factor**

| skim\_variable | n\_missing | complete\_rate | ordered | n\_unique | top\_counts |
| --- | --- | --- | --- | --- | --- |
| Product | 0 | 1 | FALSE | 1 | TM7: 40, TM1: 0, TM4: 0 |
| Gender | 0 | 1 | FALSE | 2 | Mal: 33, Fem: 7 |
| MaritalStatus | 0 | 1 | FALSE | 2 | Par: 23, Sin: 17 |

**Variable type: numeric**

| skim\_variable | n\_missing | complete\_rate | mean | sd | p0 | p25 | p50 | p75 | p100 |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 0 | 1 | 29.10 | 6.97 | 22 | 24.75 | 27.0 | 30.25 | 48 |
| Education | 0 | 1 | 17.32 | 1.64 | 14 | 16.00 | 18.0 | 18.00 | 21 |
| Usage | 0 | 1 | 4.78 | 0.95 | 3 | 4.00 | 5.0 | 5.00 | 7 |
| Fitness | 0 | 1 | 4.62 | 0.67 | 3 | 4.00 | 5.0 | 5.00 | 5 |
| Income | 0 | 1 | 75441.57 | 18505.84 | 48556 | 58204.75 | 76568.5 | 90886.00 | 104581 |
| Miles | 0 | 1 | 166.90 | 60.07 | 80 | 120.00 | 160.0 | 200.00 | 360 |

**Observation:**

For customers who bought the TM798 model, the average age is 29.1 years, with a median age of 27 years. On average, these customers have 17.3 years of education, and the middle education value is 18 years. Their treadmill usage plan ranges from 4 to 5 days a week. The average expected weekly running distance for these customers is 167 miles, while the median distance is 160 miles. The average income of these customers is $75,400, with a median income of (dollor)76,000. Now since we have the general idea of all three model of the treadmill, so we can start answering the problem questions

# Analyze

1) **How many models does store have?**

-**Observation:** -This clearly shows, Cardio Good Fitness company provides three type of product namely:“TM195”, “TM498”, “TM798”.

unique(mydata$Product) # code to get all the unique models of the treadmill in the product section.

## [1] TM195 TM498 TM798  
## Levels: TM195 TM498 TM798

1. **Which is most sold Model?**

-**Observation** -This indicates that the TM195 model is the highest-selling, chosen by 80 users, whereas the TM798 model is the least popular, selected by only 40 users.

mydata %>% group\_by(Product) %>% summarise(Count=n())

## # A tibble: 3 × 2  
## Product Count  
## <fct> <int>  
## 1 TM195 80  
## 2 TM498 60  
## 3 TM798 40

1. **Are Male customers buying treadmill more than female customers?**

-**Observation:** -Based on the displayed results, it is evident that there is a higher number of male customers purchasing treadmills compared to female customers. Specifically, there are 104 male customers and only 76 female customers.

mydata %>% group\_by(Gender) %>% summarise(count=n())

## # A tibble: 2 × 2  
## Gender count  
## <fct> <int>  
## 1 Female 76  
## 2 Male 104

1. **What is the average Income, Age, Education of people buying treadmill?**

-**Observation:** -According to the output, the average income of treadmill buyers is approximately $53,719.58, their average age is around 28.79, and they have an average education background of 15.57 years.

mean(mydata$Income)

## [1] 53719.58

mean(mydata$Age)

## [1] 28.78889

mean(mydata$Education)

## [1] 15.57222

1. **How many days and miles customer expect to run on treadmill?**

-**Observation:** -So from these data we can say that the customer are expect to use the treadmill for the average of 3.4555 days a week and customer are expect to run on the treadmill for average of 103.1944 miles.

# Calculate the average number of days and miles expected to run on the treadmill  
expected\_days\_miles = mydata %>%  
 summarise(  
 Avg\_Days\_Usage = mean(Usage),  
 Avg\_Miles\_Expected = mean(Miles)  
 )  
  
expected\_days\_miles

## # A tibble: 1 × 2  
## Avg\_Days\_Usage Avg\_Miles\_Expected  
## <dbl> <dbl>  
## 1 3.46 103.

1. **What is the average self-rated fitness of customers buying treadmill?**

-**Observation:** -From the output results, it can be inferred that the average self-rated fitness level of customers purchasing treadmills is 3.311 out of 5.

mean(mydata$Fitness)

## [1] 3.311111

1. **Are married customers buying Treadmill more than Single customers?**

-**Observation:** -The data clearly indicates that married(Partnered) customers are buying more treadmills than single customers.

mydata %>% group\_by(MaritalStatus) %>% summarise(count=n())

## # A tibble: 2 × 2  
## MaritalStatus count  
## <fct> <int>  
## 1 Partnered 107  
## 2 Single 73

1. **Is there any relation between Income and model?**

Hypothesis Test

H0: There is no significant differences in income means among the individual using various treadmill models.

H1 : There is significant differences in income means among the individual using various treadmill models.

mydata %>% group\_by(Product) %>% summarise(count = n(), mean = mean(Income, na.rm = TRUE),sd = sd(Income, na.rm = TRUE))

## # A tibble: 3 × 4  
## Product count mean sd  
## <fct> <int> <dbl> <dbl>  
## 1 TM195 80 46418. 9076.  
## 2 TM498 60 48974. 8654.  
## 3 TM798 40 75442. 18506.

aov(Income~Product, data=mydata)

## Call:  
## aov(formula = Income ~ Product, data = mydata)  
##   
## Terms:  
## Product Residuals  
## Sum of Squares 24490250199 24281991523  
## Deg. of Freedom 2 177  
##   
## Residual standard error: 11712.66  
## Estimated effects may be unbalanced

summary(aov(Income~Product, data=mydata))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Product 2 2.449e+10 1.225e+10 89.26 <2e-16 \*\*\*  
## Residuals 177 2.428e+10 1.372e+08   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Observation:**

Here our p-value is quite small indicating we do have strong evidence to believe that the average amount of income for different group of individual using different product model is different. i,e we reject null hypothesis.

1. **Are different age groups buying different models?**

Hypothesis Test

H0 : There is no significant differences in Average age among the individual using(buying) various treadmill models.

H1 : There is significant differences in Average age among the individual using(buying) various treadmill models.

mydata %>% group\_by(Product) %>% summarise(count = n(), mean = mean(Age, na.rm = TRUE),sd = sd(Age, na.rm = TRUE))

## # A tibble: 3 × 4  
## Product count mean sd  
## <fct> <int> <dbl> <dbl>  
## 1 TM195 80 28.6 7.22  
## 2 TM498 60 28.9 6.65  
## 3 TM798 40 29.1 6.97

aov(Age~Product, data=mydata)

## Call:  
## aov(formula = Age ~ Product, data = mydata)  
##   
## Terms:  
## Product Residuals  
## Sum of Squares 9.178 8620.800  
## Deg. of Freedom 2 177  
##   
## Residual standard error: 6.978903  
## Estimated effects may be unbalanced

summary(aov(Age~Product, data=mydata))

## Df Sum Sq Mean Sq F value Pr(>F)  
## Product 2 9 4.59 0.094 0.91  
## Residuals 177 8621 48.71

**Observation:**

Here our p-value is big enough indicating we do not have the sufficient evidence to believe that the average age of the individual for different group of individual using(buying) different product model is different. i,e we fail to reject null hypothesis.

1. **Is there any relation between self-rated fitness and model?**

Hypothesis Test

H0 : There is no significant difference in the average self-rated fitness across different product models.

H1 : There is a significant difference in the average self-rated fitness across different product models.

aov(Fitness~Product, data=mydata)

## Call:  
## aov(formula = Fitness ~ Product, data = mydata)  
##   
## Terms:  
## Product Residuals  
## Sum of Squares 88.91528 75.66250  
## Deg. of Freedom 2 177  
##   
## Residual standard error: 0.6538132  
## Estimated effects may be unbalanced

summary(aov(Fitness~Product, data=mydata))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Product 2 88.92 44.46 104 <2e-16 \*\*\*  
## Residuals 177 75.66 0.43   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Observation:**

-Here our p-value is quite small indicating we do have strong evidence to believe that There is a significant difference in the average self-rated fitness across different product models. i,e we reject null hypothesis.

1. **Is there any relation between education and model**?

Hypothesis Test

H0 : There is no significant difference in the average years of education level across different product models.

H1 : There is a significant difference in the average years of education level across different product models.

aov(Education~Product, data=mydata)

## Call:  
## aov(formula = Education ~ Product, data = mydata)  
##   
## Terms:  
## Product Residuals  
## Sum of Squares 158.2153 309.8458  
## Deg. of Freedom 2 177  
##   
## Residual standard error: 1.32308  
## Estimated effects may be unbalanced

summary(aov(Education~Product, data=mydata))

## Df Sum Sq Mean Sq F value Pr(>F)   
## Product 2 158.2 79.11 45.19 <2e-16 \*\*\*  
## Residuals 177 309.9 1.75   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

**Observation:**

Here our p-value is quite small indicating we do have strong evidence to believe that there is a significant difference in the average years of education level across different product models. i,e we reject null hypothesis.

1. **Does gender have any effect on model customer buy?**

Hypothesis Test

H0 : Gender and Treadmill Model are independent (no association)

H1 : Gender and Treadmill Model are not independent (significant association)

# First, create a contingency table  
contingency\_table = table(mydata$Gender, mydata$Product)  
contingency\_table

##   
## TM195 TM498 TM798  
## Female 40 29 7  
## Male 40 31 33

# Perform the Chi-Square Test  
chi\_square\_result <- chisq.test(contingency\_table)  
  
# Print the test results  
print(chi\_square\_result)

##   
## Pearson's Chi-squared test  
##   
## data: contingency\_table  
## X-squared = 12.924, df = 2, p-value = 0.001562

**Observation:**

Here our p-value is quite small indicating we do have strong evidence to believe that there is significant association between Gender and Treadmill Model (not independent). i,e we reject null hypothesis.

1. **Does Martial status have any effect model customer buy?**

Hypothesis Test

H0 : Marital Status and Treadmill Model are independent (no association)

H1 : Marital Status and Treadmill Model are not independent (significant association)

# First, create a contingency table  
contingency\_table1 = table(mydata$MaritalStatus, mydata$Product)  
contingency\_table1

##   
## TM195 TM498 TM798  
## Partnered 48 36 23  
## Single 32 24 17

# Perform the Chi-Square Test  
chi\_square\_result = chisq.test(contingency\_table1)  
  
# Print the test results  
print(chi\_square\_result)

##   
## Pearson's Chi-squared test  
##   
## data: contingency\_table1  
## X-squared = 0.080655, df = 2, p-value = 0.9605

**Observation:**

Here our p-value is big enough indicating we do not have the sufficient evidence to believe that There is significant association between Marital Status and Treadmill Model. i,e we fail to reject null hypothesis.

# Share

Data Visualization Let’s try to get more information and inferance from plots and graphs. We will try to make this analysis part also more systematic. So what we gonna do is, we will first do visualization for univarite and then we will go for multi-variate.

-In univeriate also first we will look for numerical variable and then we will look for the categorical Variables.

**Univariate-[Numerical Variable]**

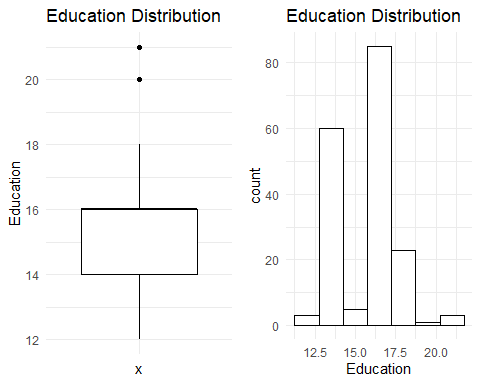
**For Education**

**Observation:**

-The majority of customers have 16 years of education, which can be assumed to be college graduates or bachelor’s degree holders.

-There are a few outliers in the education data, which means there are a few customers with education levels significantly different from the majority, either higher or lower. These outliers represent individuals with educational backgrounds that deviate from the typical customer profile.

library(gridExtra)  
# Box plot  
box\_plot <- ggplot(mydata, aes(x = "", y = Education)) +  
 geom\_boxplot(color = "black", fill = "white") +  
 labs(title = "Education Distribution") +  
 theme\_minimal() +  
 theme(legend.position = "none",  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank())  
  
# Histogram  
histogram\_plot <- ggplot(mydata, aes(x = Education)) +  
 geom\_histogram(binwidth = 1.5, color = "black", fill = "white") +  
 labs(title = "Education Distribution") +  
 theme\_minimal()  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot, histogram\_plot, ncol = 2)

 For Income

**For Income**

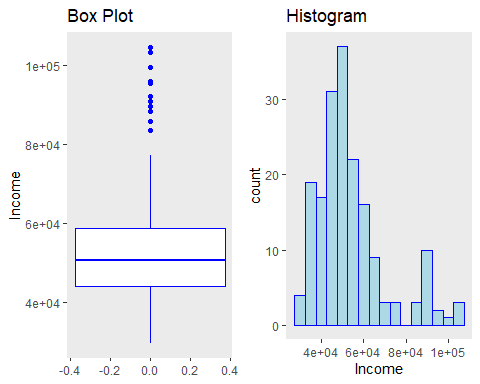
**Observation:**

-The income data is right-skewed, indicating that there are some high-earning customers, while most customers have lower incomes. The median income is $50,000, which is the middle value when all incomes are arranged from lowest to highest. The mean income is (dollor)55,000, which is the average income of all customers. The mode, which represents the most common income value, is (dollor)45,000.

-The majority of customers fall within the lower pay range and earn less than (dollor)70,000.

-There are some outliers in the income data, where a few customers earn beyond (dollor)80,000, which are considerably higher than the majority of incomes.

# Box plot  
box\_plot <- ggplot(mydata, aes(y = Income)) +  
 geom\_boxplot(color = "blue") +  
 labs(title = "Box Plot") +  
 theme(panel.grid = element\_blank())  
  
# Histogram  
histogram\_plot <- ggplot(mydata, aes(x = Income)) +  
 geom\_histogram(binwidth = 5000, color = "blue", fill = "lightblue") +  
 labs(title = "Histogram") +  
 theme(panel.grid = element\_blank())  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot, histogram\_plot, ncol = 2)



**For Age**

**Observation:**

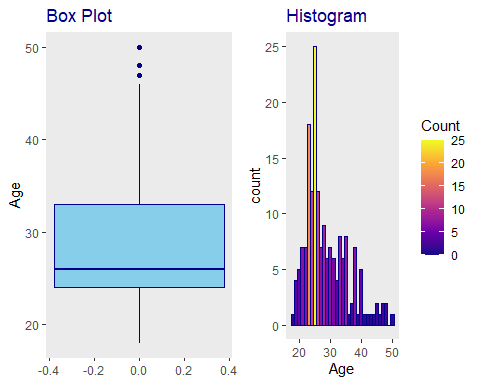
-The age distribution of customers buying the treadmill is right-skewed, which means there are relatively more younger customers.

-The average age of customers is 28, with the middle value (median) being 26, and the most common age (mode) being 25.

-There are very few customers who buy the treadmill after the age of 40 or before the age of 20. The majority of customers fall within the age range of 20 to 40.

# Box plot  
box\_plot <- ggplot(mydata, aes(y = Age)) +  
 geom\_boxplot(fill = "skyblue", color = "darkblue") +  
 labs(title = "Box Plot", fill = "Age") +  
 theme(panel.grid = element\_blank(), plot.title = element\_text(color = "darkblue"))  
  
# Histogram  
histogram\_plot <- ggplot(mydata, aes(x = Age, fill = ..count..)) +  
 geom\_histogram(binwidth = 1, color = "darkblue") +  
 labs(title = "Histogram", fill = "Count") +  
 scale\_fill\_viridis\_c(option = "plasma") +  
 theme(panel.grid = element\_blank(), plot.title = element\_text(color = "darkblue"))  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot, histogram\_plot, ncol = 2)

## Warning: The dot-dot notation (`..count..`) was deprecated in ggplot2 3.4.0.  
## ℹ Please use `after\_stat(count)` instead.  
## This warning is displayed once every 8 hours.  
## Call `lifecycle::last\_lifecycle\_warnings()` to see where this warning was  
## generated.



**For Usage**

**Observation:**

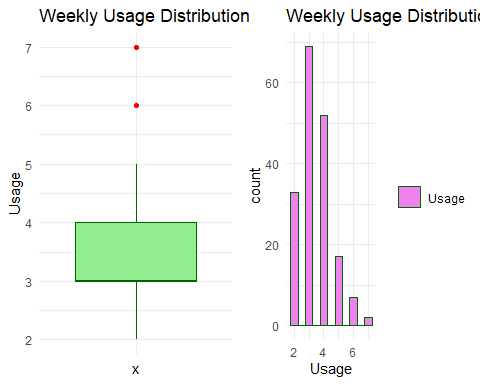
-The majority of customers expect to use the treadmill 3-4 days per week. This is the range where most of the data is concentrated, indicating that it’s the most common usage expectation among customers.

-There are a few outliers in the data where customers are expecting to use the treadmill for 6 or 7 times a week. These outliers represent individuals with higher usage expectations, deviating from the typical usage pattern.

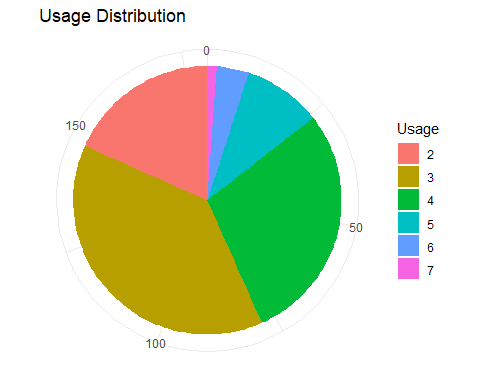
-This plot provides an informative overview of the expected treadmill usage distribution among the customers, highlighting the most common usage pattern and the presence of outliers with higher expectations.

-It shows that the majority of customers (represented by the largest slice of the pie) want to use the treadmill for 3 times a week and fewer customers want to use the treadmill for 4, 5, 6, and 7 times a week,as their slice is getting smaller with 7 times a week being the least preferred option.

# Box plot  
box\_plot <- ggplot(mydata, aes(x = "", y = Usage)) +  
 geom\_boxplot(color = "darkgreen", fill = "lightgreen", outlier.color = "red") +  
 labs(title = "Weekly Usage Distribution", fill = "") +  
 theme\_minimal() +  
 theme(legend.position = "none",  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank())  
  
# Histogram  
histogram\_plot <- ggplot(mydata, aes(x = Usage, fill = "Usage")) +  
 geom\_histogram(binwidth = 0.5, color = "darkgreen") +  
 labs(title = "Weekly Usage Distribution", fill = "") +  
 scale\_fill\_manual(values = c("violet", "purple")) +  
 theme\_minimal()  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot, histogram\_plot, ncol = 2)



# Usage Distribution Plot - Pie Chart  
usage\_plot <- ggplot(mydata, aes(x = "", fill = factor(Usage))) +  
 geom\_bar(width = 1) +  
 labs(title = "Usage Distribution", fill = "Usage", x = NULL, y = NULL) +  
 theme\_minimal() +  
 coord\_polar(theta = "y")  
  
print(usage\_plot)



**For Miles**

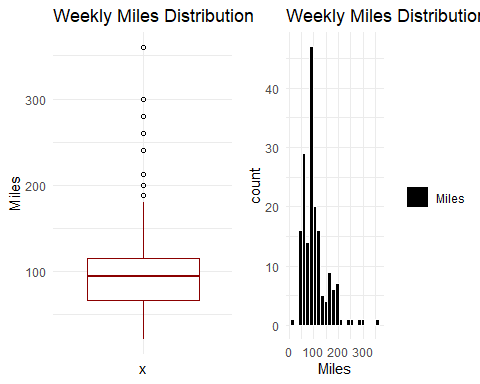
**Observation:**

-The miles data is right-skewed, which means that there are some customers with significantly higher expected miles per week, pulling the distribution towards the right.

-The box plot will show the central tendency (median) and spread of the miles data. The median value (indicated by the line inside the box) is likely to be around 80 miles per week, suggesting that half of the customers expect to run 80 miles or below per week.

-There are some outliers in the data where customers are expecting to run more than 200 miles per week. These outliers represent individuals with much higher expected mileage compared to the majority of the customers.

# Box plot  
box\_plot <- ggplot(mydata, aes(x = "", y = Miles)) +  
 geom\_boxplot(color = "darkred", outlier.color = "black", outlier.shape = 1) +  
 labs(title = "Weekly Miles Distribution", fill = "") +  
 theme\_minimal() +  
 theme(legend.position = "none",  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank())  
  
# Histogram  
histogram\_plot <- ggplot(mydata, aes(x = Miles, fill = "Miles")) +  
 geom\_histogram(binwidth = 15, color = "white") +  
 labs(title = "Weekly Miles Distribution", fill = "") +  
 scale\_fill\_manual(values = c("black", "purple")) +  
 theme\_minimal()  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot, histogram\_plot, ncol = 2)



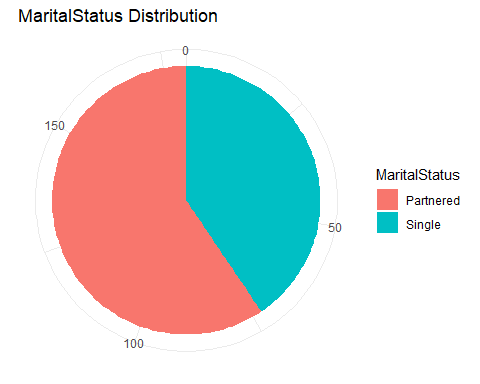
**Univariate- [Categorical Variable]**

**For MaritalStatus**

**Observation:**

-Clearly, the pie chart indicates that partnered customers are purchasing the treadmill more frequently compared to single customers.

# MaritalStatus Distribution Plot - Simple Pie Chart  
MaritalStatus\_plot <- ggplot(mydata, aes(x = "", fill = factor(MaritalStatus))) +  
 geom\_bar(width = 1) +  
 labs(title = "MaritalStatus Distribution", fill = "MaritalStatus", x = NULL, y = NULL) +  
 theme\_minimal() +  
 coord\_polar(theta = "y")  
  
print(MaritalStatus\_plot)

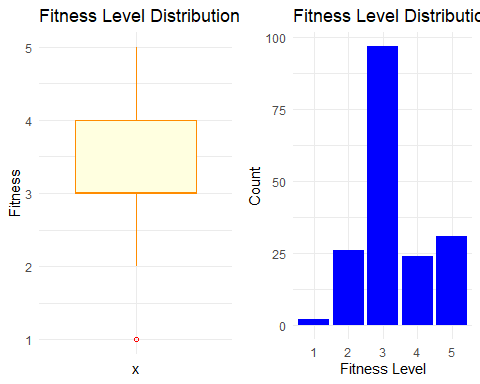


**For Fitness**

**Obaservation**

-Based on the result from above plot, it seems that most of the customers have self-rated their fitness level as 3, which is an average level of fitness.

# Bar plot for Fitness Level Distribution  
FitnessLevel\_plot <- ggplot(mydata, aes(x = factor(Fitness))) +  
 geom\_bar(fill = "blue") +  
 labs(title = "Fitness Level Distribution", x = "Fitness Level", y = "Count") +  
 theme\_minimal()  
  
#print(FitnessLevel\_plot)  
  
box\_plot <- ggplot(mydata, aes(x = "", y = Fitness)) +  
 geom\_boxplot(color = "darkorange", fill = "lightyellow", outlier.color = "red", outlier.shape = 1) +  
 labs(title = "Fitness Level Distribution", fill = "") +  
 theme\_minimal() +  
 theme(legend.position = "none",  
 axis.text.x = element\_blank(),  
 axis.ticks.x = element\_blank())  
  
  
  
# Combine the plots using grid.arrange from gridExtra package  
grid.arrange(box\_plot,FitnessLevel\_plot, ncol = 2)

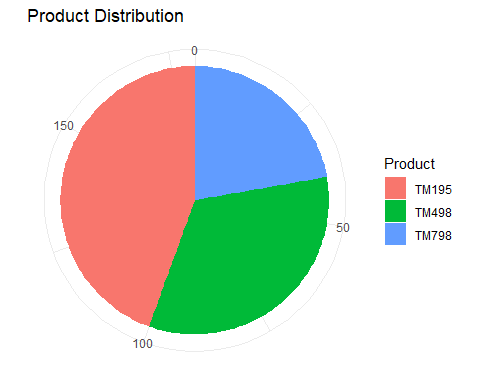


**For Product**

**Observation**

-The pie chart clearly illustrates the distribution of customers among different treadmill models. Model TM195 stands out as the most popular choice shown by the larger segment. On the other hand, model TM798 appears to be the least preferred, as indicated by its much smaller segment.

# Product Distribution Plot - Simple Pie Chart  
Product\_plot <- ggplot(mydata, aes(x = "", fill = factor(Product))) +  
 geom\_bar(width = 1) +  
 labs(title = "Product Distribution", fill = "Product", x = NULL, y = NULL) +  
 theme\_minimal() +  
 coord\_polar(theta = "y")  
  
print(Product\_plot)

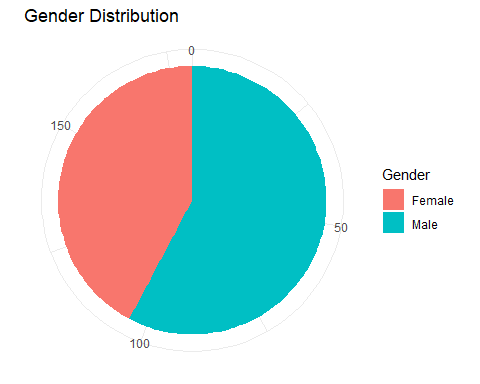


**For Gender**

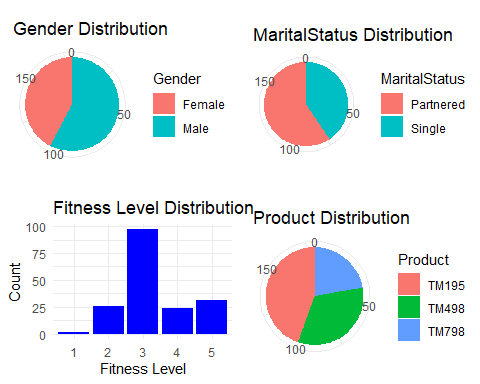
**Observation:**

-We can see that the Male use more treadmill as compared to female.

# Product Distribution Plot - Simple Pie Chart  
Gender\_plot <- ggplot(mydata, aes(x = "", fill = factor(Gender))) +  
 geom\_bar(width = 1) +  
 labs(title = "Gender Distribution", fill = "Gender", x = NULL, y = NULL) +  
 theme\_minimal() +  
 coord\_polar(theta = "y")  
  
print(Gender\_plot)



# Arrange all the plots in one figure  
all\_plots <- grid.arrange(Gender\_plot, MaritalStatus\_plot, FitnessLevel\_plot, Product\_plot, nrow = 2, ncol = 2, widths = c(4, 4))



# Display the combined figure  
print(all\_plots)

## TableGrob (2 x 2) "arrange": 4 grobs  
## z cells name grob  
## 1 1 (1-1,1-1) arrange gtable[layout]  
## 2 2 (1-1,2-2) arrange gtable[layout]  
## 3 3 (2-2,1-1) arrange gtable[layout]  
## 4 4 (2-2,2-2) arrange gtable[layout]

**Multi-variate**

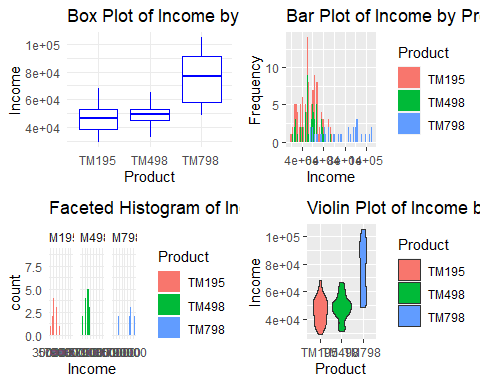
**Plot of Income by Product**

**Observation:**

-The graphs clearly illustrate that individuals with high earning prefer the TM798 treadmill model over others, as it shows a significantly higher median income compared to TM195 and TM498 users. However, it is essential to consider that TM798 users also exhibit greater income variability.

# Box plot of Income by Product  
boxplot\_Income <- ggplot(mydata, aes(x = Product, y = Income)) +  
 geom\_boxplot(color = "blue", outlier.shape = NA) +  
 labs(title = "Box Plot of Income by Product") +  
 theme\_minimal()  
  
# bar plot of Income by Product  
g = ggplot(data=mydata,aes(x=Income,fill=Product))  
barplot\_Income = g + geom\_bar(width = 800) +  
labs(y = "Frequency", x= "Income",  
 title = "Bar Plot of Income by Product")  
  
# Faceted histogram of Income by Product  
histogram\_Income <- ggplot(mydata, aes(x = Income, fill = Product)) +  
 geom\_histogram(binwidth = 800, position = "identity") +  
 facet\_wrap(~Product, ncol = 3) +  
 labs(title = "Faceted Histogram of Income by Product") +  
 theme\_minimal()  
  
  
  
# Violin plot of Income by Product  
violinplot\_Income <- ggplot(mydata, aes(x = Product, y = Income, fill = Product)) +  
 geom\_violin() +  
 labs(title = "Violin Plot of Income by Product")  
  
# Combine all plots into one figure  
grid.arrange(boxplot\_Income, barplot\_Income, histogram\_Income,violinplot\_Income, nrow = 2, ncol = 2, widths = c(4, 4))

## Warning: `position\_stack()` requires non-overlapping x intervals

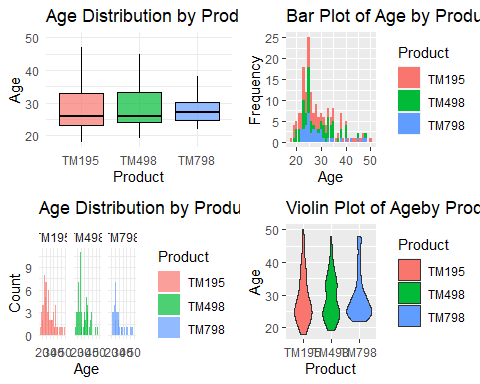


**Plots of Age by Product**

**Observation:**

-Based on the graph, it’s evident that the median age of individuals using different treadmill models is quite similar. However, there is a slight difference in age variability among users of the TM798 model. The majority of TM798 users fall within the age range of 20-30. This suggests that while the average age may not differ significantly across treadmill models, the TM798 attracts a younger demographic, with a concentration of users in their 20s and 30s.

# Box plot of Age by Product  
boxplot\_age <- ggplot(mydata, aes(x = Product, y = Age, fill = Product)) +  
 geom\_boxplot(color = "black", alpha = 0.7, outlier.shape = NA) +  
 labs(title = "Age Distribution by Product", x = "Product", y = "Age") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
# bar plot of Age by Product  
g = ggplot(data=mydata,aes(x=Age,fill=Product))  
barplot\_age = g + geom\_bar() +  
labs(y = "Frequency", x= "Age",  
 title = "Bar Plot of Age by Product")  
  
# Faceted histogram of Age by Product  
histogram\_Age <- ggplot(mydata, aes(x = Age, fill = Product)) +  
 geom\_histogram(binwidth = 1, alpha = 0.7, position = "identity") +  
 facet\_wrap(~Product, ncol = 3) +  
 labs(title = "Age Distribution by Product", x = "Age", y = "Count") +  
 theme\_minimal()  
  
# Violin plot of Age by Product  
violinplot\_age <- ggplot(mydata, aes(x = Product, y = Age, fill = Product)) +  
 geom\_violin() +  
 labs(title = "Violin Plot of Ageby Product")  
  
# Combine all plots into one figure  
grid.arrange(boxplot\_age, barplot\_age, histogram\_Age,violinplot\_age, nrow = 2, ncol = 2, widths = c(4, 4))

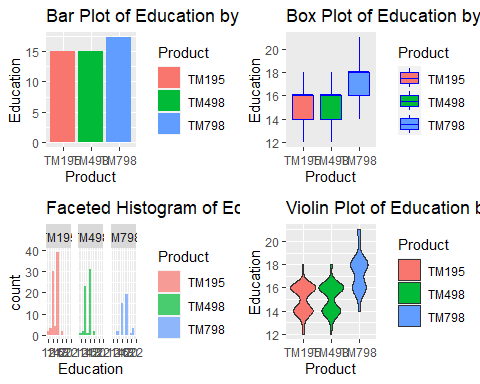


**Plot of Education by Product**

**Interpertation:**

-The provided plot shows that the median years of education for users of TM195 and TM498 treadmill models are very similar and lower compared to the median years of education for users of the TM798 model. This indicates that individuals with higher levels of education tend to favor the TM798 model over the other two options. In other words, the data suggests that more educated individuals show a preference for the TM798 treadmill.

library(gridExtra)  
# Bar plot of Education by Product  
barplot\_education <- ggplot(mydata, aes(x = Product, y = Education, fill = Product)) +  
 geom\_bar(stat = "summary", fun = "mean") +  
 labs(title = "Bar Plot of Education by Product")  
  
# Box plot of Education by Product  
boxplot\_Education <- ggplot(mydata, aes(x = Product, y = Education, fill = Product)) +  
 geom\_boxplot(color = "blue") +  
 labs(title = "Box Plot of Education by Product") +  
 theme(panel.grid = element\_blank())  
  
# Faceted histogram of Education by Product  
histogram\_education <- ggplot(mydata, aes(x = Education, fill = Product)) +  
 geom\_histogram(binwidth = 1, position = "identity", alpha = 0.7) +  
 facet\_wrap(~Product, ncol = 3) +  
 labs(title = "Faceted Histogram of Education by Product")  
  
# Violin plot of Education by Product  
violinplot\_Education <- ggplot(mydata, aes(x = Product, y = Education, fill = Product)) +  
 geom\_violin() +  
 labs(title = "Violin Plot of Education by Product")  
  
# Arrange all plots into one figure  
grid.arrange(barplot\_education, boxplot\_Education, histogram\_education,violinplot\_Education, nrow = 2, ncol = 2, widths = c(4, 4))



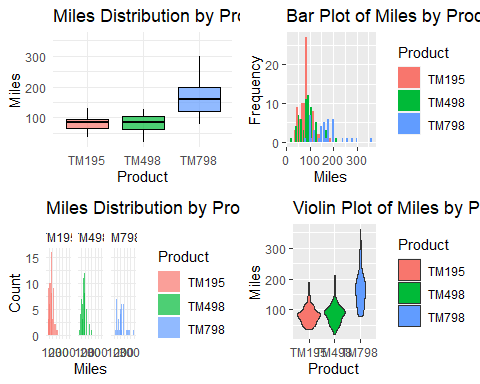
**Miles by Product**

**Observation:**

-Based on the plot, we observe that the median distance expected to be run by individuals using the TM195 and TM498 treadmill models is quite similar, and it is significantly lower than the median distance expected to be run by those using the TM798 model. This implies that users of the TM798 treadmill are generally expected to cover more miles than users of the other two models. In summary, the data suggests that individuals using the TM798 treadmill are likely to achieve greater distances in their workouts compared to users of TM195 and TM498.

# Box plot of Miles by Product  
boxplot\_Miles <- ggplot(mydata, aes(x = Product, y = Miles, fill = Product)) +  
 geom\_boxplot(color = "black", alpha = 0.7, outlier.shape = NA) +  
 labs(title = "Miles Distribution by Product", x = "Product", y = "Miles") +  
 theme\_minimal() +  
 theme(legend.position = "none")  
  
# bar plot of Age by Product  
g = ggplot(data=mydata,aes(x=Miles,fill=Product))  
barplot\_Miles = g + geom\_bar(width = 8) +  
labs(y = "Frequency", x= "Miles",  
 title = "Bar Plot of Miles by Product")  
  
# Faceted histogram of Miles by Product  
histogram\_Miles <- ggplot(mydata, aes(x = Miles, fill = Product)) +  
 geom\_histogram(binwidth = 8, alpha = 0.7, position = "identity") +  
 facet\_wrap(~Product, ncol = 3) +  
 labs(title = "Miles Distribution by Product", x = "Miles", y = "Count") +  
 theme\_minimal()  
  
# Violin plot of Miles by Product  
violinplot\_Miles <- ggplot(mydata, aes(x = Product, y = Miles, fill = Product)) +  
 geom\_violin() +  
 labs(title = "Violin Plot of Miles by Product")  
  
# Combine all plots into one figure  
grid.arrange(boxplot\_Miles, barplot\_Miles, histogram\_Miles,violinplot\_Miles, nrow = 2, ncol = 2, widths = c(4, 4))

## Warning: `position\_stack()` requires non-overlapping x intervals



**Descriptive Analysis:**

Now, below we will do some Descriptive Analysis based on few questions

1. **What are the basic statistics for each variable in the dataset (e.g., mean, median, standard deviation, min, max)?**

summary(mydata)

## 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 Min. :1.000 Min. : 29562 Min. : 21.0   
## 1st Qu.:3.000 1st Qu.:3.000 1st Qu.: 44059 1st Qu.: 66.0   
## Median :3.000 Median :3.000 Median : 50597 Median : 94.0   
## Mean :3.456 Mean :3.311 Mean : 53720 Mean :103.2   
## 3rd Qu.:4.000 3rd Qu.:4.000 3rd Qu.: 58668 3rd Qu.:114.8   
## Max. :7.000 Max. :5.000 Max. :104581 Max. :360.0

1. **How many data points (customers) are there for each treadmill model?**

mydata %>% count(Product)

## # A tibble: 3 × 2  
## Product n  
## <fct> <int>  
## 1 TM195 80  
## 2 TM498 60  
## 3 TM798 40

1. **What is the dimensions of the data? How may observations and how many variables?**

# Find the dimensions of the data  
data\_dimensions <- dim(mydata)  
  
# Extract the number of observations (rows) and variables (columns)  
num\_observations <- data\_dimensions[1]  
num\_variables <- data\_dimensions[2]  
  
# Print the results  
cat("Number of observations (rows):", num\_observations, "\n")

## Number of observations (rows): 180

cat("Number of variables (columns):", num\_variables, "\n")

## Number of variables (columns): 9

1. **What is the range of customer education and what is the median?**

# Find the range of customer education  
education\_range <- range(mydata$Education)  
  
# Calculate the median of customer education  
education\_median <- median(mydata$Education)  
  
# Print the results  
cat("Range of customer education:", education\_range[1], "to", education\_range[2], "\n")

## Range of customer education: 12 to 21

cat("Median of customer education:", education\_median, "\n")

## Median of customer education: 16

1. **What is the average age of customer who purchases TM195?**

# Filter the data to include only rows where the product is TM195  
tm195\_data <- filter(mydata, Product == "TM195")  
  
# Calculate the average age of customers who purchased TM195  
average\_age\_tm195 <- mean(tm195\_data$Age)  
  
# Print the result  
cat("The average age of customers who purchase TM195 is:", average\_age\_tm195, "years.\n")

## The average age of customers who purchase TM195 is: 28.55 years.

1. **Perform appropriate tests to see if gender affects the model purchased**?

Hypothesis Test

H0 : The distribution of the “Product” variable is independent of the “Gender” variable.

H1 : The distribution of the “Product” variable is not independent of the “Gender” variable.

# Create a contingency table between Gender and Product  
contingency\_table <- table(mydata$Gender, mydata$Product)  
  
# Perform the chi-square test  
chi\_sq\_result <- chisq.test(contingency\_table)  
  
# Print the results  
cat("Chi-square test result:\n")

## Chi-square test result:

print(chi\_sq\_result)

##   
## Pearson's Chi-squared test  
##   
## data: contingency\_table  
## X-squared = 12.924, df = 2, p-value = 0.001562

**Observation:**

Here our p-value is quite small indicating we do have strong evidence to believe that The distribution of the “Product” variable is not independent of the “Gender” variable. i,e we reject null hypothesis.

1. **Which model is most popular? Which model attracts people with lower income?**

# Calculate the frequency distribution of the 'Product' variable  
product\_frequency <- table(mydata$Product)  
  
# Find the most popular model  
most\_popular\_model <- names(product\_frequency)[which.max(product\_frequency)]  
  
# Print the most popular model  
cat("The most popular model is for lower income people( for more info, look graph above):", most\_popular\_model, "\n")

## The most popular model is for lower income people( for more info, look graph above): TM195

**Conclusion**

**Important general observations**

* -TM195 model is the highest-selling, chosen by 80 users, whereas the TM798 model is the least popular, selected by only 40 users.
* -There are more male customer then female customer
* -Patnered customer purchase more trademill then single
* -The most popular model for lower income people is TM195
* -Users using the TM798 model treadmill are cover greater miles in their workouts compared to users of TM195 and TM498.
* -More educated individuals perfer treadmill of model TM798
* -Younger individual are attracted more towards model TM798, specifically individual in their 20s or 30s
* -Most of the customer rate their fitness level as 3, which is close to average also
* -The majority of customers fall within the lower pay range and earn less than (dollor)70,000.
* -There are very few customers who buy the treadmill after the age of 40 or before the age of 20. The majority of customers fall within the age range of 20 to 40.

**Act**

**Recommendations:**

* Based on the analysis, it appears that the treadmill model TM195 attracts people with lower income, as there is a concentration of customers with lower income values who have purchased this model. Given this insight, marketing the TM195 as a budget-friendly treadmill could be a strategic approach. By positioning the TM195 as an affordable and value-for-money option, it may appeal to a broader customer base, especially those looking for cost-effective fitness equipment. However, before proceeding with the marketing strategy, it’s essential to conduct a thorough market analysis and consider other factors to ensure the success of the campaign. Customer feedback and preferences should also be taken into account to fine-tune the marketing message for the TM195.
* The analysis suggests that TM798 is preferred by individual with higher income. Positioning it as a high-end luxurious treadmill may appeal to discerning customers seeking top-quality fitness equipment. To ensure success, conduct thorough market analysis, consider customer preferences, and highlight its premium features in the marketing campaign. Emphasize TM798’s unique value proposition and ability to elevate the fitness experience for affluent customers
* To attract more female customers, we could run special promotions on Women’s Day and Mother’s Day, emphasizing the importance of staying fit and active.
* As we have a larger number of younger individuals using our product, it’s important to keep this in mind when creating advertisements.

*The End.*

*Your suggestions are always appreciated.*