Code ▾

Project - Cardio Good Fitness

Juan Esteban Venegas

1. Project Objectives

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

The sample data is representative of the population. There are different customer profiles for different product type in the data set.

3. Exploratory Data Analysis – Step by step approach

3.1 Environment Set up and Data Import

3.1.1 Call necessary Packages and Invoke Libraries

Hide

```
library(ggplot2)
library(data.table)
library(scales)
library(spelling)
library(corrplot)
options(scipen = 999)
```

3.1.2 Set up working Directory

Hide

```
setwd('C:/Users/Juan Esteban Venegas/Desktop
      /Machine Learning Learning/Greatlearning/Project - Cardio Good Fitness')
```

```
Error in setwd("C:/Users/Juan Esteban Venegas/Desktop\n      /Machine Learning Learning/Greatlearning/Project - Cardio Good
Fitness") :
  cannot change working directory
```

3.1.3 Import and Read the Dataset

Hide

```
CardioGoodFitness <- fread(paste0(getwd(),'/info/CardioGoodFitness.csv'))
```

3.2 Variable Identification

3.2.1 Variable Identification – Inferences

Hide

```
names(CardioGoodFitness)
```

```
[1] "Product"       "Age"           "Gender"        "Education"     "MaritalStatus" "Usage"         "Fitness"
[8] "Income"        "Miles"
```

There are 8 columns in the data set. The main column to group information by in order to tie a customer profile to a specific product should be the ‘Product’ column.

Hide

```
str(CardioGoodFitness)
```

```
Classes ‘data.table’ and 'data.frame':  180 obs. of  9 variables:
 $ Product      : chr  "TM195" "TM195" "TM195" "TM195" ...
 $ Age          : int  18 19 19 19 20 20 21 21 21 21 ...
 $ Gender       : chr  "Male" "Male" "Female" "Male" ...
 $ Education    : int  14 15 14 12 13 14 14 13 15 15 ...
 $ MaritalStatus: chr  "Single" "Single" "Partnered" "Single" ...
 $ 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 ...
 - attr(*, ".internal.selfref")=<externalptr>
```

There are three character variables which will be transformed to factor variables. Since there is no information on product pricing or characteristics, an assumption will be made regarding the product. This document will assume the number in the name of the product to be an indicator of the product range and consequently it’s price.

Hide

```
CardioGoodFitness[,`:=` (Product = factor(Product, levels = c('TM195', 'TM498', 'TM798')), Gender = factor(Gender), MaritalS
tatus = factor(MaritalStatus))]
```

Now that the variables are in the desired format it is important to take a look to the data.

Hide

```
head(CardioGoodFitness,4)
```

```
   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
```

Hide

```
tail(CardioGoodFitness,4)
```

```
   Product Age Gender Education MaritalStatus Usage Fitness Income Miles
1:   TM798  42   Male        18        Single     5       4  89641   200
2:   TM798  45   Male        16        Single     5       5  90886   160
3:   TM798  47   Male        18     Partnered     4       5 104581   120
4:   TM798  48   Male        18     Partnered     4       5  95508   180
```

Finally, we look at the summary of the data set. From this, it can be observed that:

- There are no missing values.
- None of the continues variables share the same median and mean but Fitness and Usage have closer values when comparing their own mean against the median than the other variables which might be slightly skewed.
- There are more Males than Females in the sample 57.8%.
- There are more Partnered customers than Singles in the sample 59.4%.

Hide

```
summary(CardioGoodFitness)
```

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

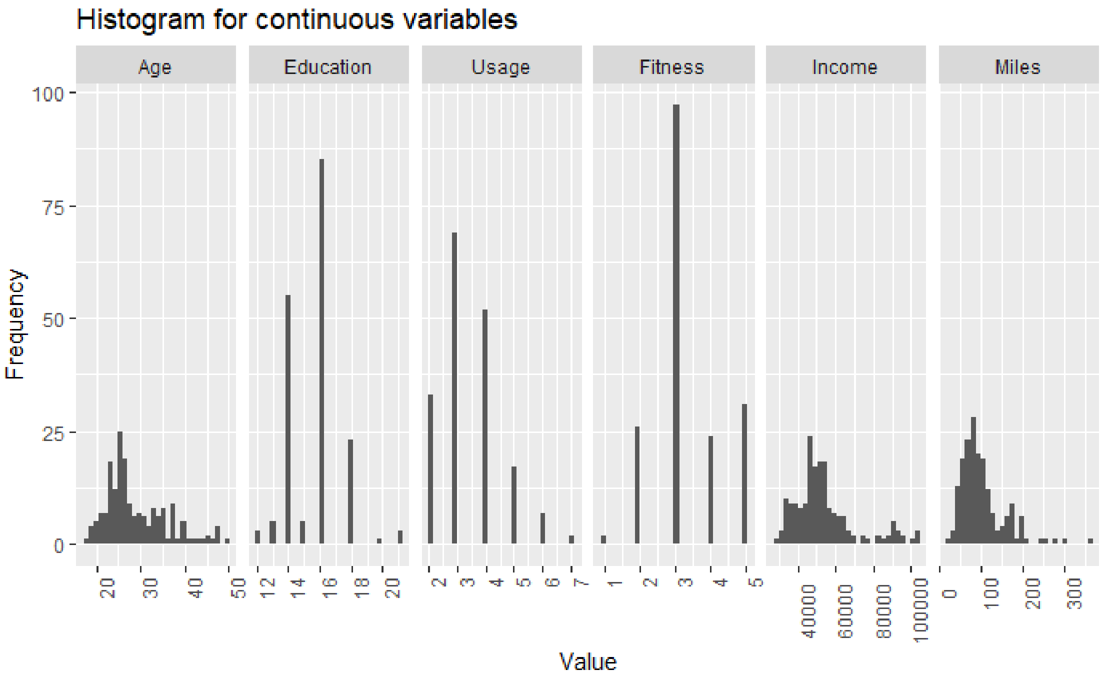
3.3 Univariate Analysis

Hide

```
Univariate <- CardioGoodFitness
Univariate <- melt(Univariate, id.vars = 'Product')
```

Hide

```
ggplot(Univariate[!(variable %in% c('MaritalStatus', 'Gender'))], aes(x = as.numeric(value))) + geom_histogram() + facet_gri
d(.~variable, scales = 'free') + labs(title = 'Histogram for continuous variables', x = 'Value', y = 'Frequency') + theme(ax
is.text.x = element_text(angle = 90))
```

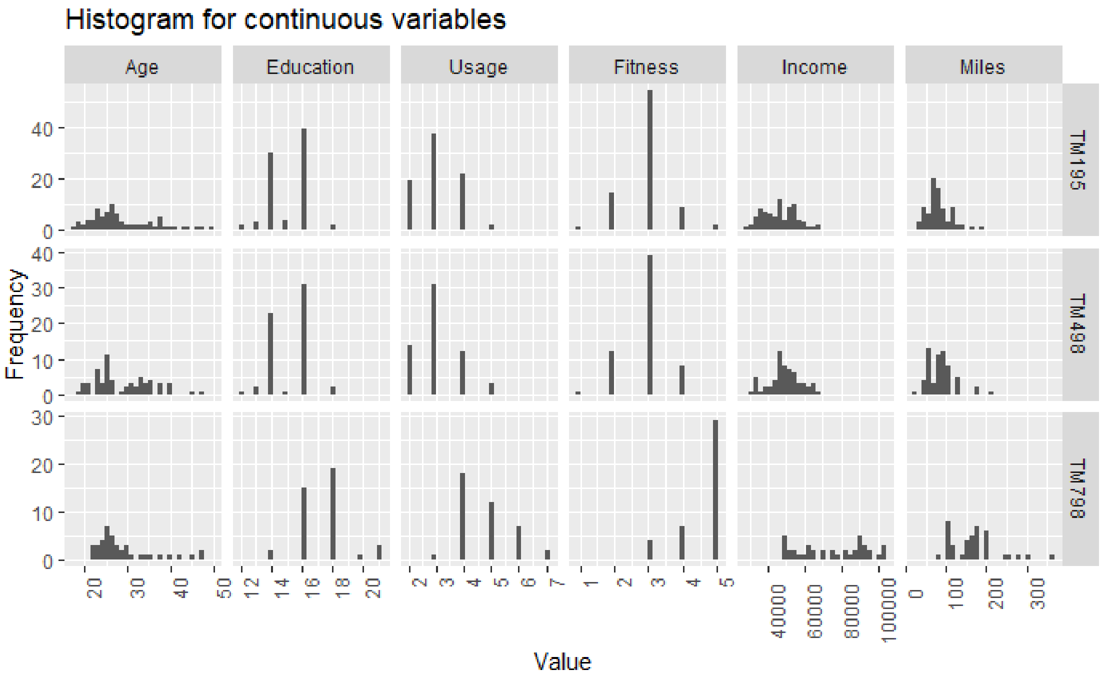


- Age: Average age is 28.79 and median age is 26. This difference is due to the presence of customers with higher age (outliers) which drive the average age towards the right. However, as it can be seen by the median, most of the sample belongs to a lower age group.
- Education: The opposite to age happens here. There is a high number of observations with 14 years of education which is making the mean (15.57) lower than the median (16)
- Usage: There is not a great difference between the mean (3.46) and the median (3.46). The fact that the mean is higher is due to observations where the usage ranges from 5 and above which represent 14.4% of the total sample.
- Fitness: Something similar to usage happens here. However, the difference between the mean (3.31) and the median (3) is lower since the max value is 5
- Income: There is quite a larger gap in income when looking at mean (53719.58) and median (50596.5). This might be interesting to analyze in more detail since it will play a key factor in determining customers segments and how they relate to the product they use. Specially since the range of observations is so wide [29562, 104581]

- Miles: Similar to Income, there is a wide range of observations for this variable wide [21, 360]. It will be necessary to map this variable against product to identify if there is a relationship between these two.

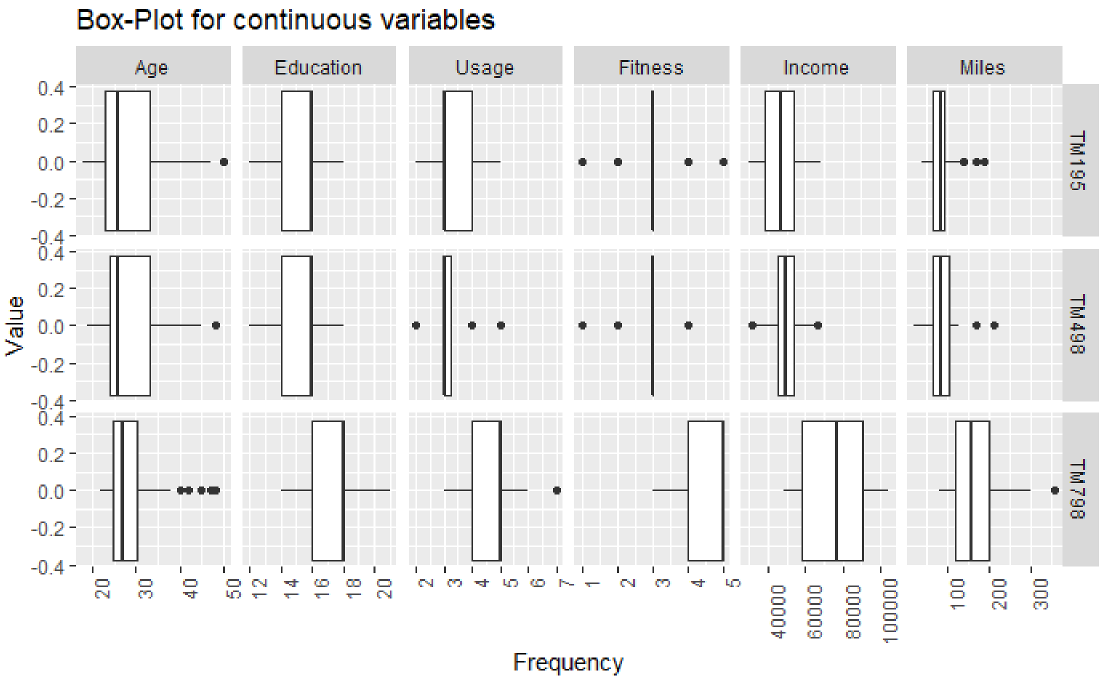
Hide

```
ggplot(Univariate[!(variable %in% c('MaritalStatus', 'Gender'))], aes(x = as.numeric(value))) + geom_histogram() + facet_grid(Product~variable, scales = 'free') + labs(title = 'Histogram for continuous variables', x = 'Value', y = 'Frequency') + theme(axis.text.x = element_text(angle = 90))
```



Hide

```
ggplot(Univariate[!(variable %in% c('MaritalStatus', 'Gender'))], aes(y = as.numeric(value))) + geom_boxplot() + facet_grid(Product~variable, scales = 'free') + coord_flip() + labs(title = 'Box-Plot for continuous variables', x = 'Value', y = 'Frequency') + theme(axis.text.x = element_text(angle = 90))
```



It is interesting to see that there are more similarities between product TM195 AND TM498 when looking at histogram plot and box plot for all the variables by product and that product TM798 seems to have users with a different set of behaviors.

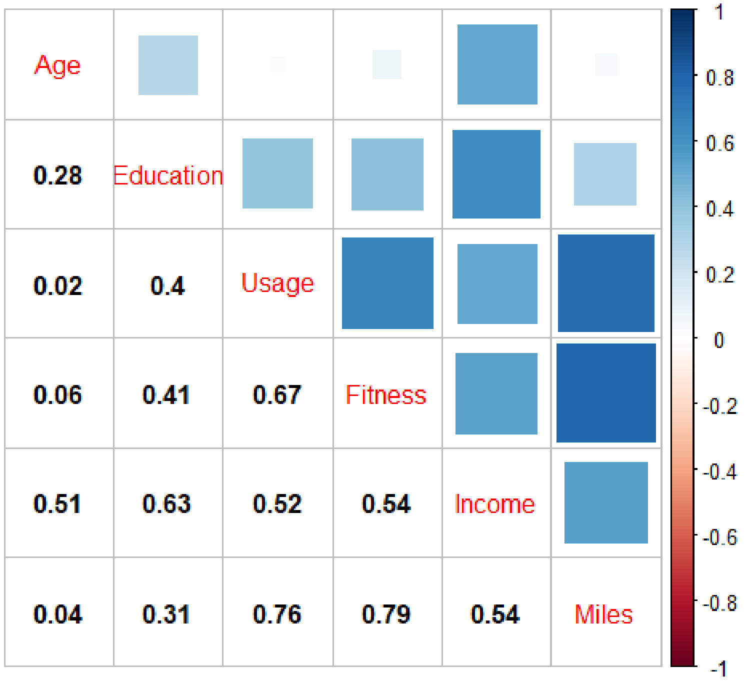
- Age: When looking at age by product, it is clear that there are outliers for the three of them. However, there are more outliers for product TM798. It is also interesting to see how the age range is narrower for this product. This could potentially mean that there could be two types of users for product TM798 regarding age and perhaps it would be interesting to identify if age group of 40+ is being well served by the current product offer.
- Education: Both product TM195 AND TM498 have similar values for education. Product TM798 on the other hand has users who have a higher education.
- Fitness, Income and Miles: Similar to education, product users TM948 have higher scores in all these variables.

Overall, it is clear that there is a difference for users of product TM798 when comparing them to the rest of the sample. For the other two however, the customer profile is not so clear yet.

3.4 Bi-Variate Analysis

Hide

```
corrplot.mixed(cor(CardioGoodFitness[, -c('Product', 'MaritalStatus', 'Gender')]), upper ='square', lower.col = 'black')
```



When looking at the correlation matrix for continuous variables it can be observed that there is a strong positive correlation between users perception on their fitness (Fitness) and the expected miles to run which in turn also explains why there is a strong correlation between Usage and miles and usage and fitness.

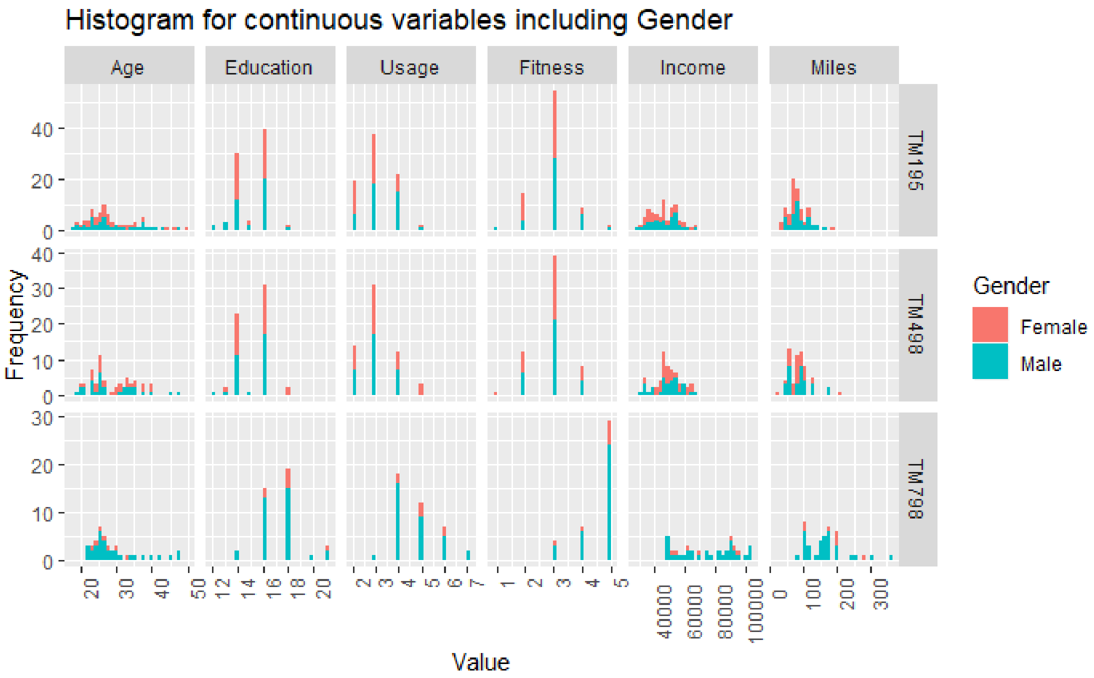
Age also has a positive correlation when paired with Income and Education which means that the higher the age of the user the higher their income and education levels tend to be.

Hide

```
Multivariate <- CardioGoodFitness
Multivariate <- melt(CardioGoodFitness, measure.vars = c('Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles'))
```

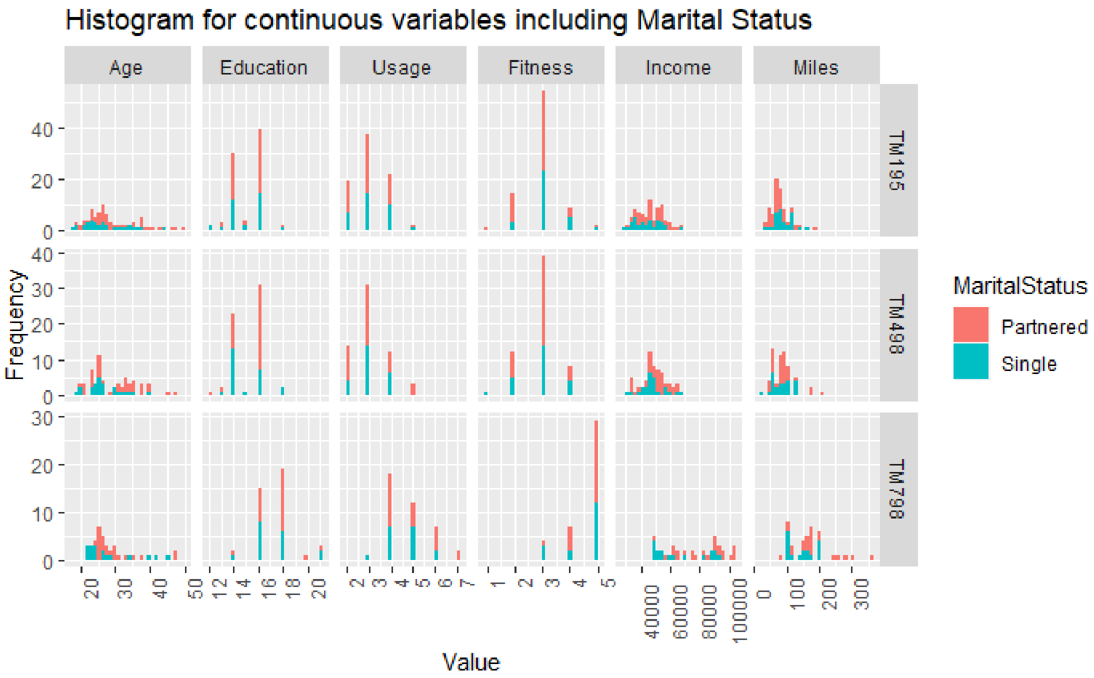
Hide

```
ggplot(Multivariate, aes(x = as.numeric(value), fill = Gender)) + geom_histogram() + facet_grid(Product~variable, scales =
'free') + labs(title = 'Histogram for continuous variables including Gender', x = 'Value', y = 'Frequency') + theme(axis.tex
t.x = element_text(angle = 90))
```



Hide

```
ggplot(Multivariate, aes(x = as.numeric(value), fill = MaritalStatus)) + geom_histogram() + facet_grid(Product~variable, sca
les = 'free') + labs(title = 'Histogram for continuous variables including Marital Status', x = 'Value', y = 'Frequency') +
 theme(axis.text.x = element_text(angle = 90))
```



3.5 Missing Value Identification

There are no missing values in the sample data.

3.6 Outlier Identification

Even though there are some outliers, they will not be normalized since that is out of the scope of this assignment.

3.7 Variable Transformation / Feature Creation

Two new tables were created in order to be able to plot all the variables at once.

Conclusion

There are several user characteristics that can be observed from the data and several customer groups could be created from this. However, it is important to have a better understanding of the characteristics of the products being sold in order to be able to define possible opportunities (Example, identify if the benefits of the products are being understood correctly, if the targeted audience it was thought for is the real audience for this product, if there are opportunities to upsell, downsell or crossell any user into another product category or even if there is a new product that should be released in order to reach an untapped market.)

From the information provided, My most important conclusions would be:

- Intention of use for product TM195 is higher than for product TM498 and the income of the customer moves in a broader range. Product TM195 should be sold as a durable an affordable product for users who would be using it frequently but for shorter periods of times per use.
- For product TM498 it seems like usage and millage is low and the income is very similar when compared to the customers in this sub group. This seems like a product who can be a little more expensive than product TM195 and should have some additional benefits that would appeal to a normal person who wants to work out once in a while in a non competitive or professional way. I would include impulsive

customers who buy gym equipment and rarely use it here and because of that I would think that the way to market this product has to be in a way where additional characteristics or novelties of the product are highlighted.

- Customers interested in this product are mainly male with the intention of using it frequently and for prolonged periods of time. In addition to these, customers in this product category seem to have a higher income and also a higher education when compared to customers in the other two product categories which means that people who are interested in this product are probably more informed and care more about the products characteristics and technical specifications.