Descriptive Analysis of Cardio Good Fitness Treadmill Product Lines

Sushrit Kafle

2024-09-23

Introduction

Cardio Good Fitness, a leading retailer in the fitness equipment industry, aims to better understand its customer base to improve its product marketing strategy and increase sales.The market research team is assigned the task to identify 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.

Dataset Description

The team identifies the following customer variables to study:

- **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.

The data are stored in the CardioGoodFitness.csv file.

Methodology

The following steps were undertaken to analyze the data and draw insights:

- **Data Preprocessing**: Identification of missing value, classification of variables into categorical and numerical for analysis.
- **Data Summary**: A preliminary analysis to understand the central tendencies and spread of numerical variables and frequency counts for categorical variables.
- **Univariate, Bivariate, Multivariate Analysis and Tests**: Performed appropriate tests to see if gender affects the model purchased. Performed tests to see if there any difference in mean age among customers that purchased different models.

- **Visualization**: Histograms and density plots were created to assess the distribution of numerical variables, bar charts and boxplots to compare categorical and numerical variables and multivariate plots to explore the combined effect of several customer attributes on treadmill purchases.

Results and Analysis

In this section, the findings from Cardio Good Fitness dataset are discussed. We ask and answer important questions about the dataset to further understand the data.

About The Dataset:

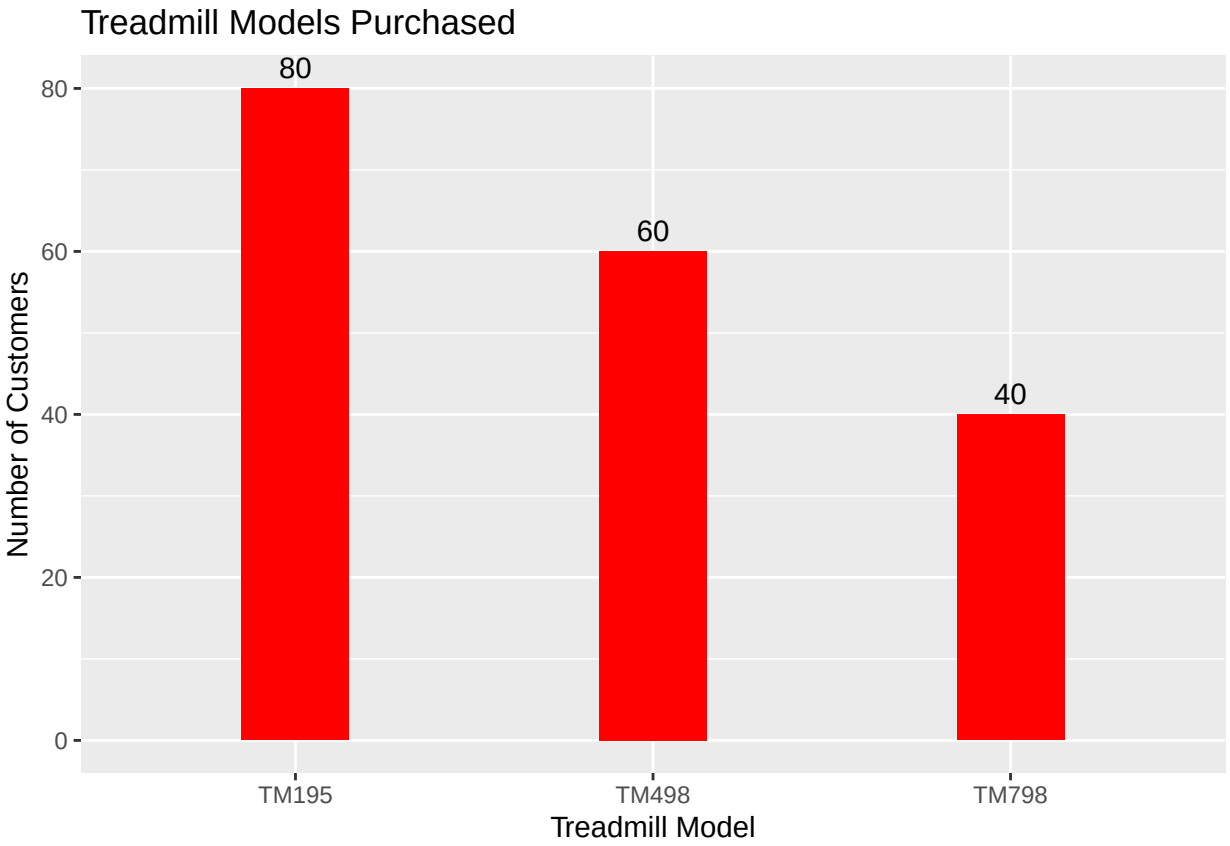
- The dataset has the shape of (180,9),i.e. 180 observations and 9 variables.
- The types of variables include:
  - **Numerical Variable**: Age(Discrete), Education(Discrete), Usage(Discrete), Fitness(Ordinal), Income(Continuous), Miles(Discrete).
  - **Categorical Variable**: Product, Gender, Marital Status.
- There are no missing data in the table.

Table 1: First 5 Rows of CardioGoodFitness Data

Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
TM195	18	Male	14	Single	3	4	29562	112
TM195	19	Male	15	Single	2	3	31836	75
TM195	19	Female	14	Partnered	4	3	30699	66
TM195	19	Male	12	Single	3	3	32973	85
TM195	20	Male	13	Partnered	4	2	35247	47

Univariate Analysis

1. Product



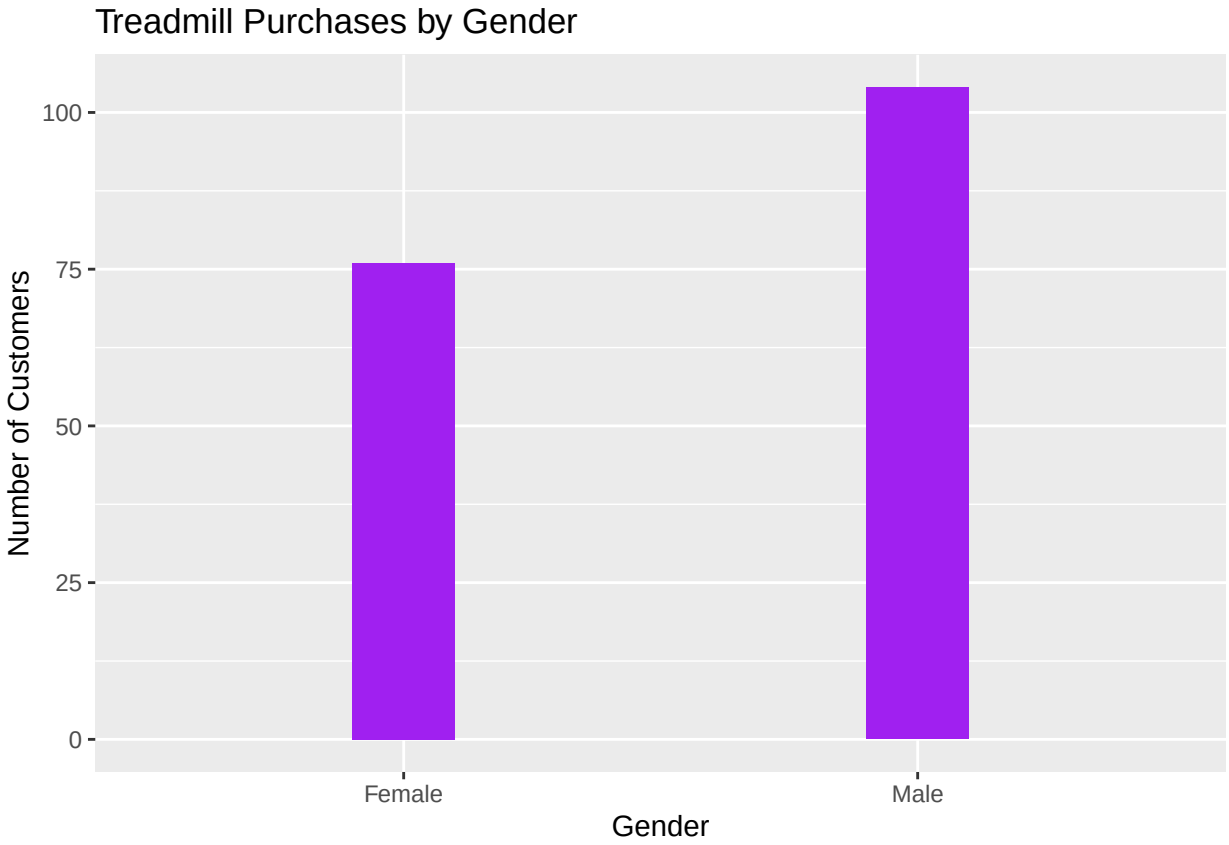
How many models does store have?

->The store has 3 models: “TM195” “TM498” and “TM798”.

Which is most sold Model?

-> TM195 is the most sold model.

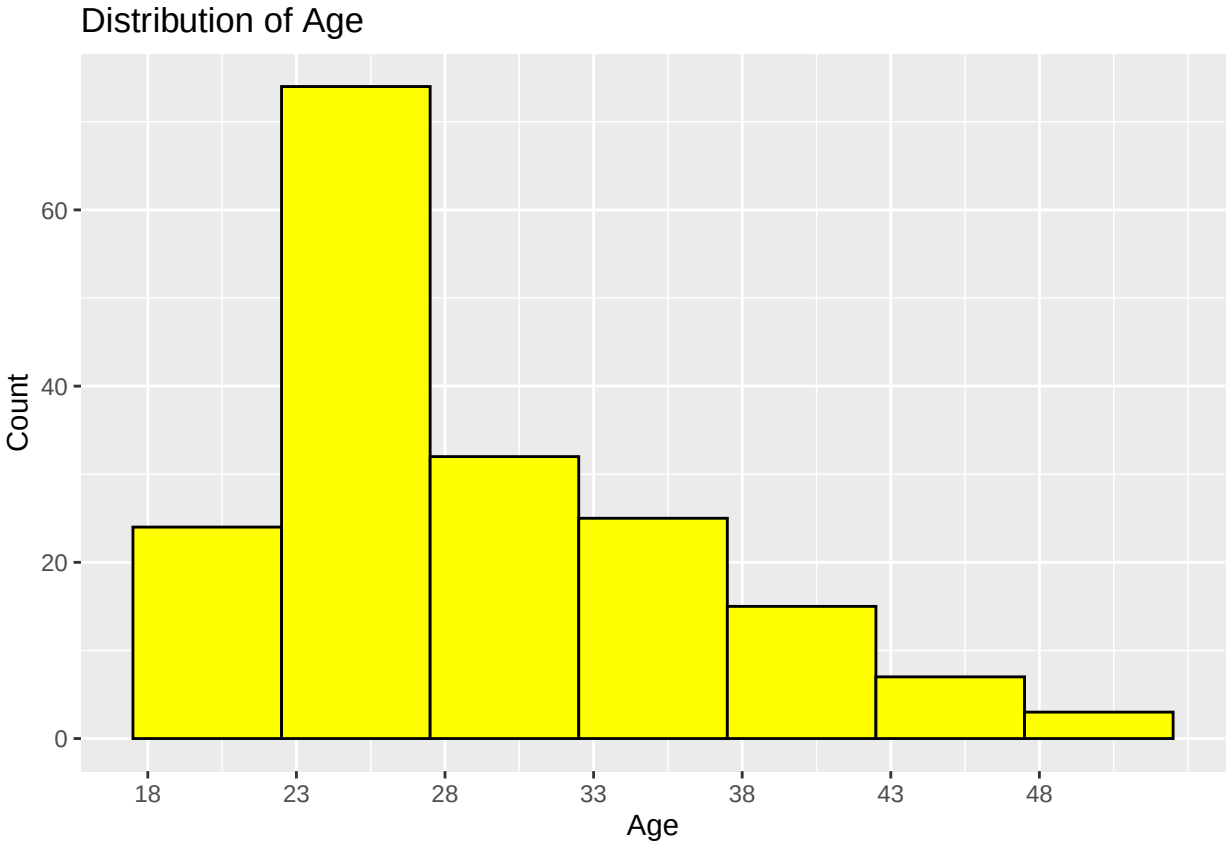
2. Gender



Are Male customers buying treadmill more than female customers?

-> Yes, male customers are buying more treadmills.

3. Age



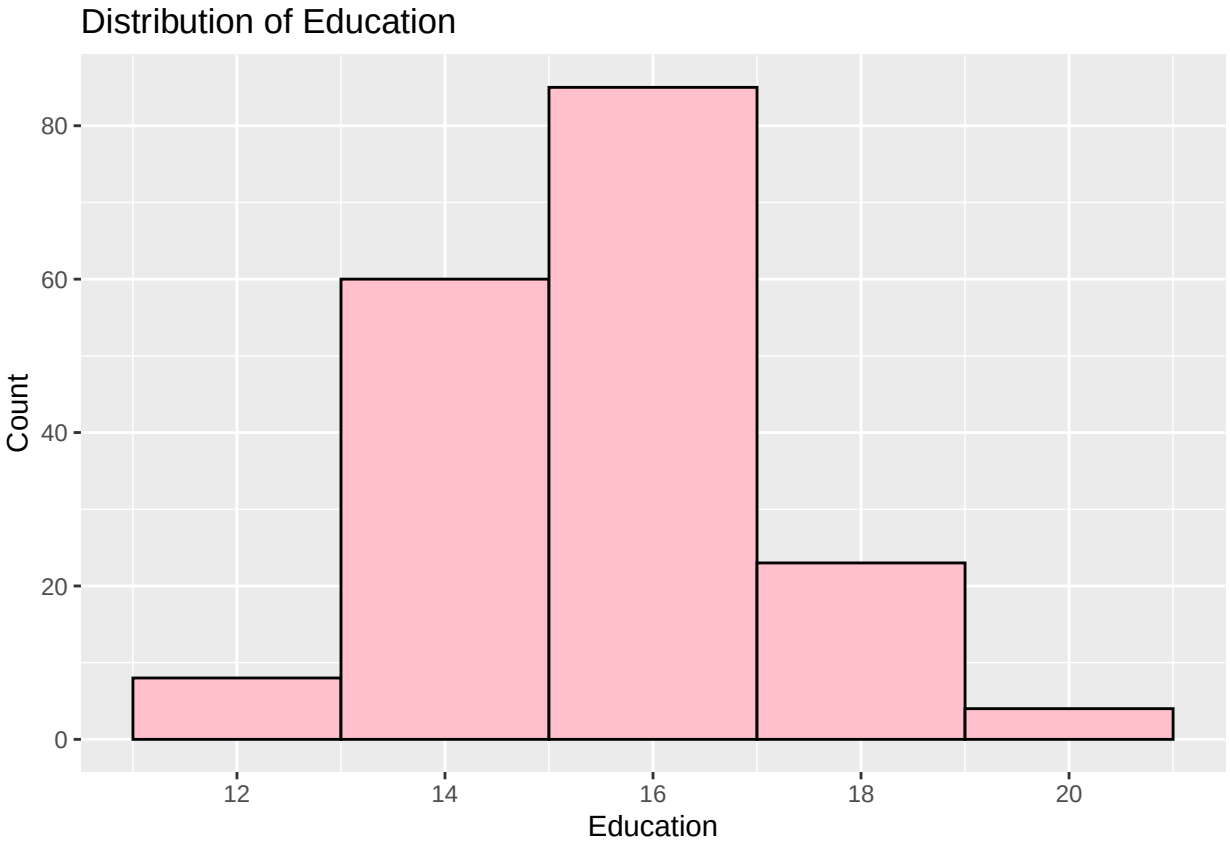
The graph is right skewed with majority between the range of 23-28.

Table 2: Age Summary

Statistic	Value
Min.	18.00000
1st Qu.	24.00000
Median	26.00000
Mean	28.78889
3rd Qu.	33.00000
Max.	50.00000

The age range is (18-50). After age 28, the sales decrease with the increase in age.

4. Education



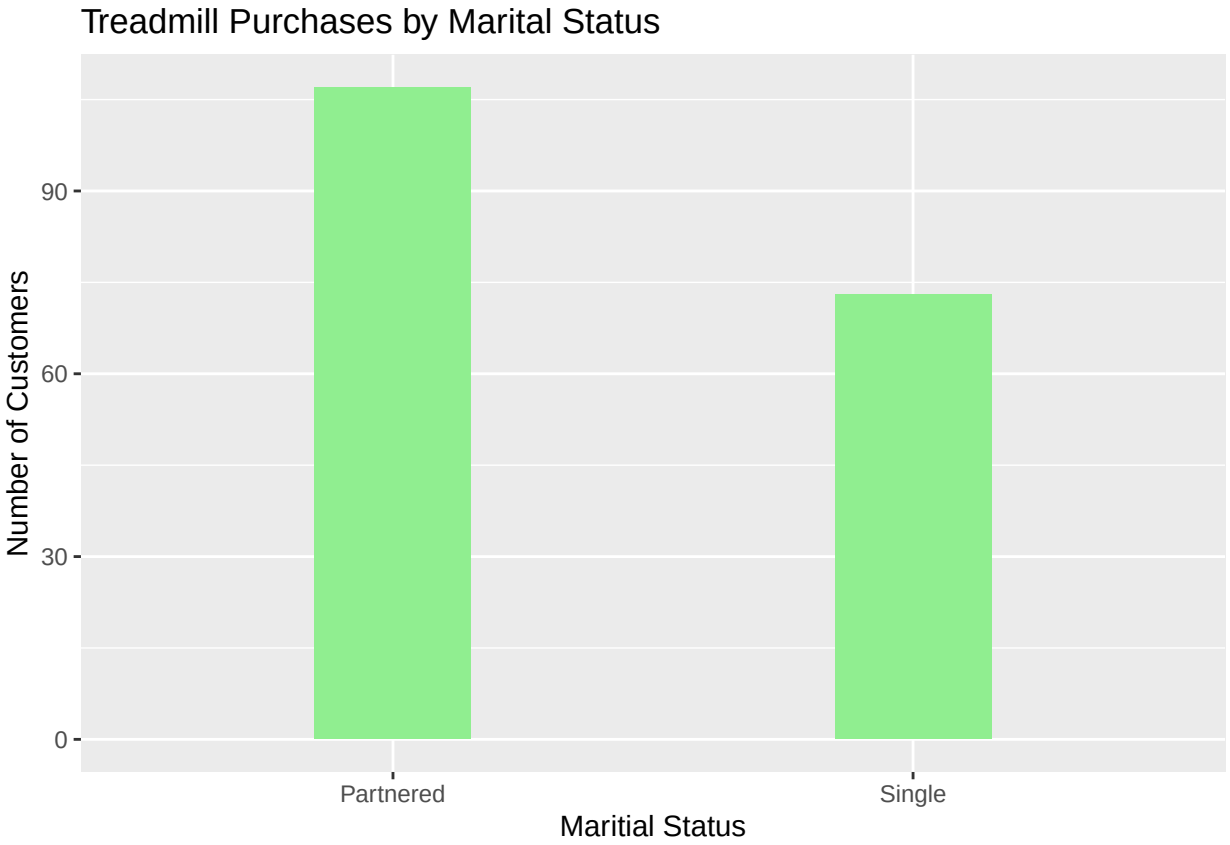
The graph shows that the education level peaks between 14-16 years.

Table 3: Education Summary

Statistic	Value
Min.	12.00000
1st Qu.	14.00000
Median	16.00000
Mean	15.57222
3rd Qu.	16.00000
Max.	21.00000

As per the data, majority of customers are high school graduate (12th). Few have advanced level education i.e more than 16 years which also shows right skewness of the graph.

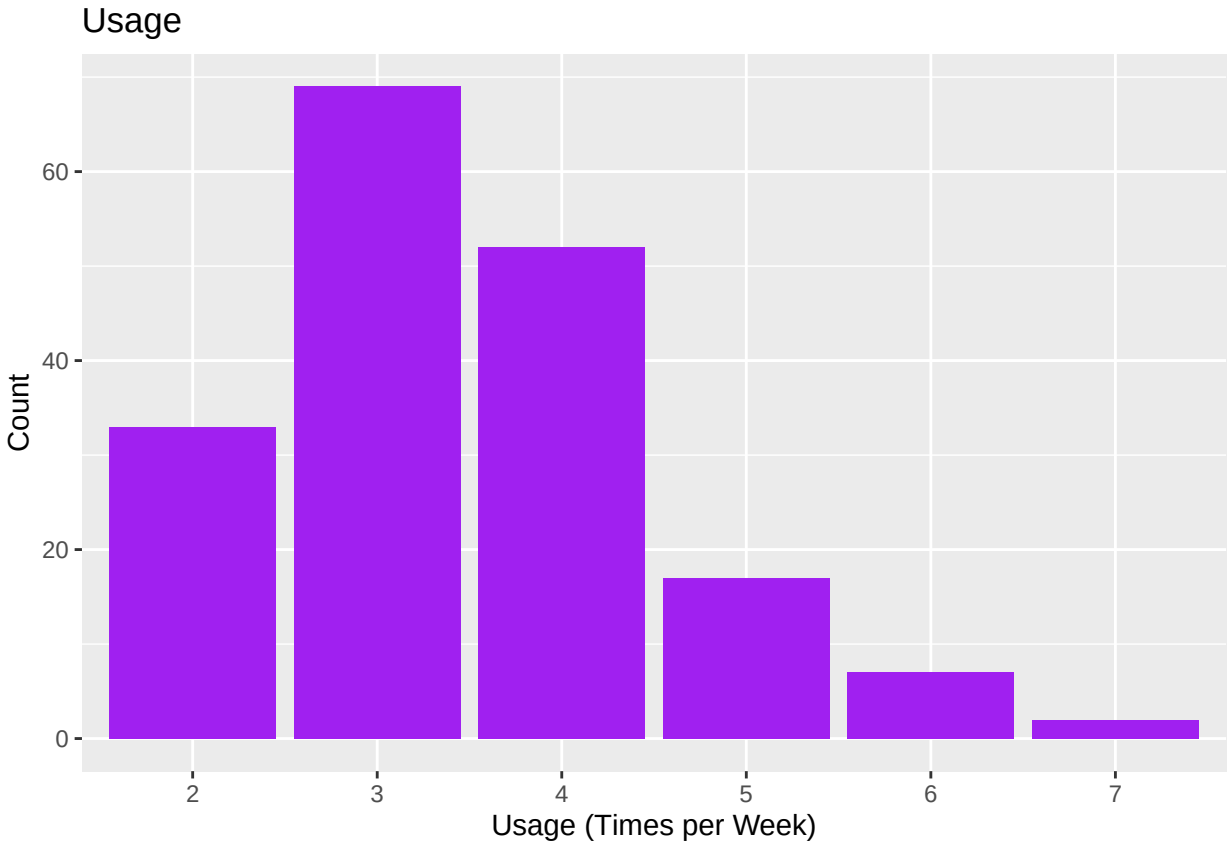
5. Martial Status



Are married customers buying Treadmill more than Single customers?

-> Yes, married people are buying more Treadmills than Single customers.

6. Usage

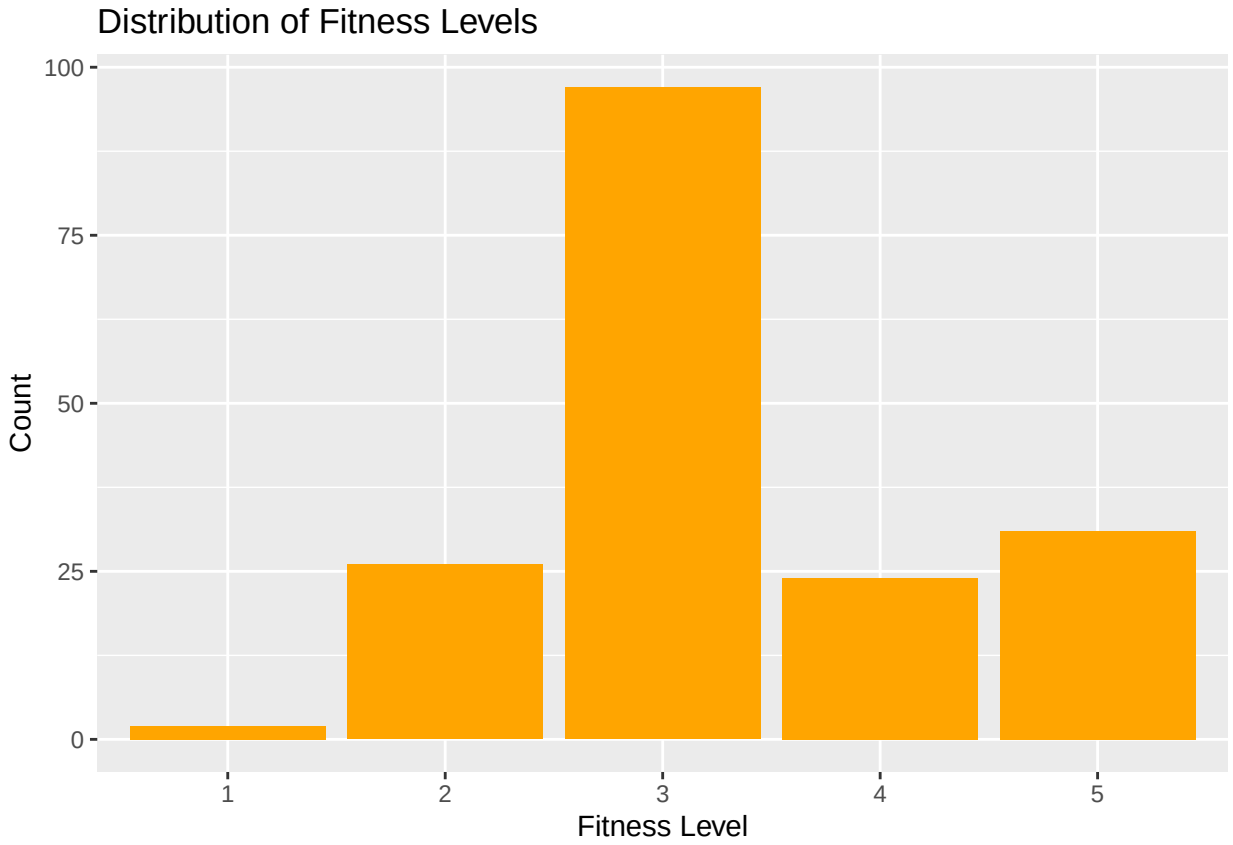


Above graph is right skewed. People on average want to use the treadmill 3 times a week. Few will use it more than 4 times a week, hence making the graph right skewed.

Table 4: Usage Summary

Statistic	Value
Min.	2.000000
1st Qu.	3.000000
Median	3.000000
Mean	3.455556
3rd Qu.	4.000000
Max.	7.000000

7. Fitness



The histogram shows that fitness level 3 is the most common, with close to 100 customers rating themselves at this level.

Table 5: Fitness Summary

Statistic	Value
Min.	1.000000
1st Qu.	3.000000
Median	3.000000
Mean	3.311111
3rd Qu.	4.000000
Max.	5.000000

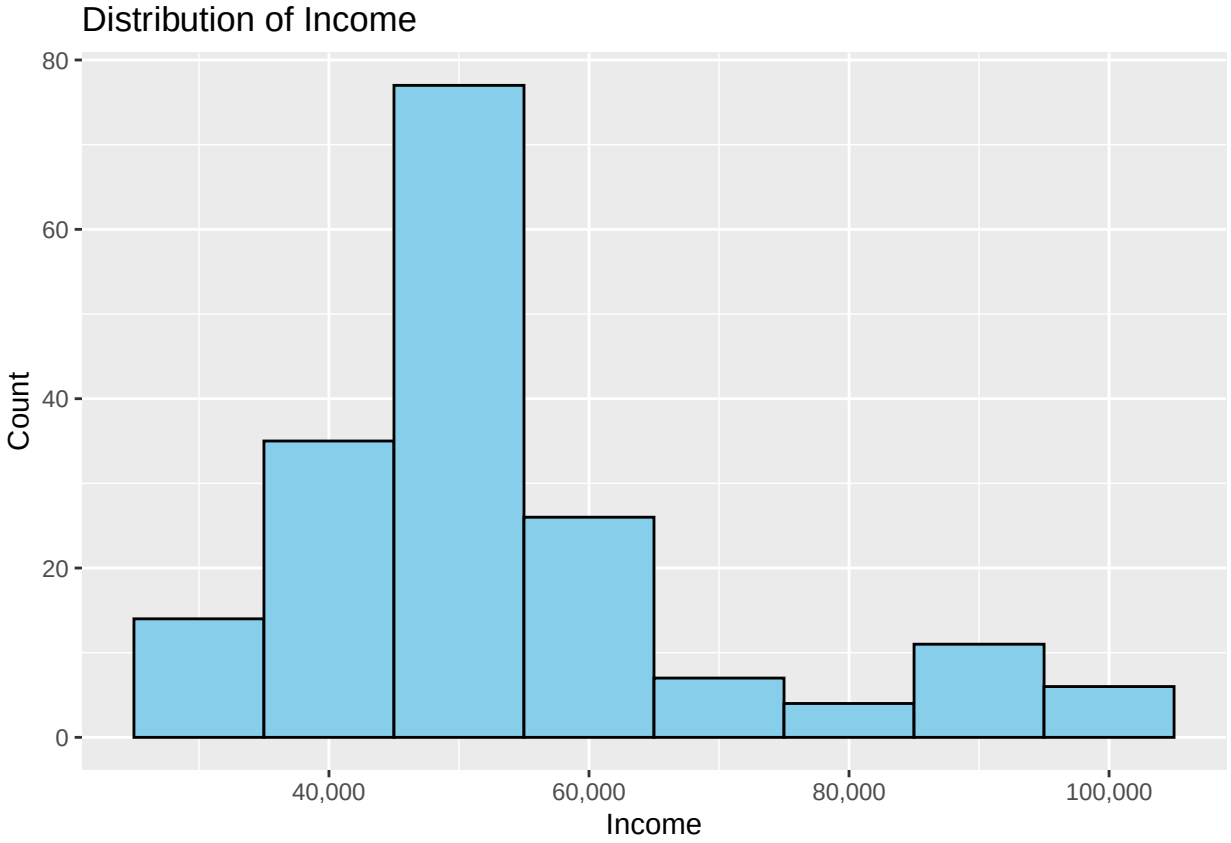
What is the average self-rated fitness of customers buying treadmill?

-> The average self-rate fitness is given as:

Table 6: Average self-rated fitness of customers buying treadmill

Avg_Rating
3.311111

8. Income

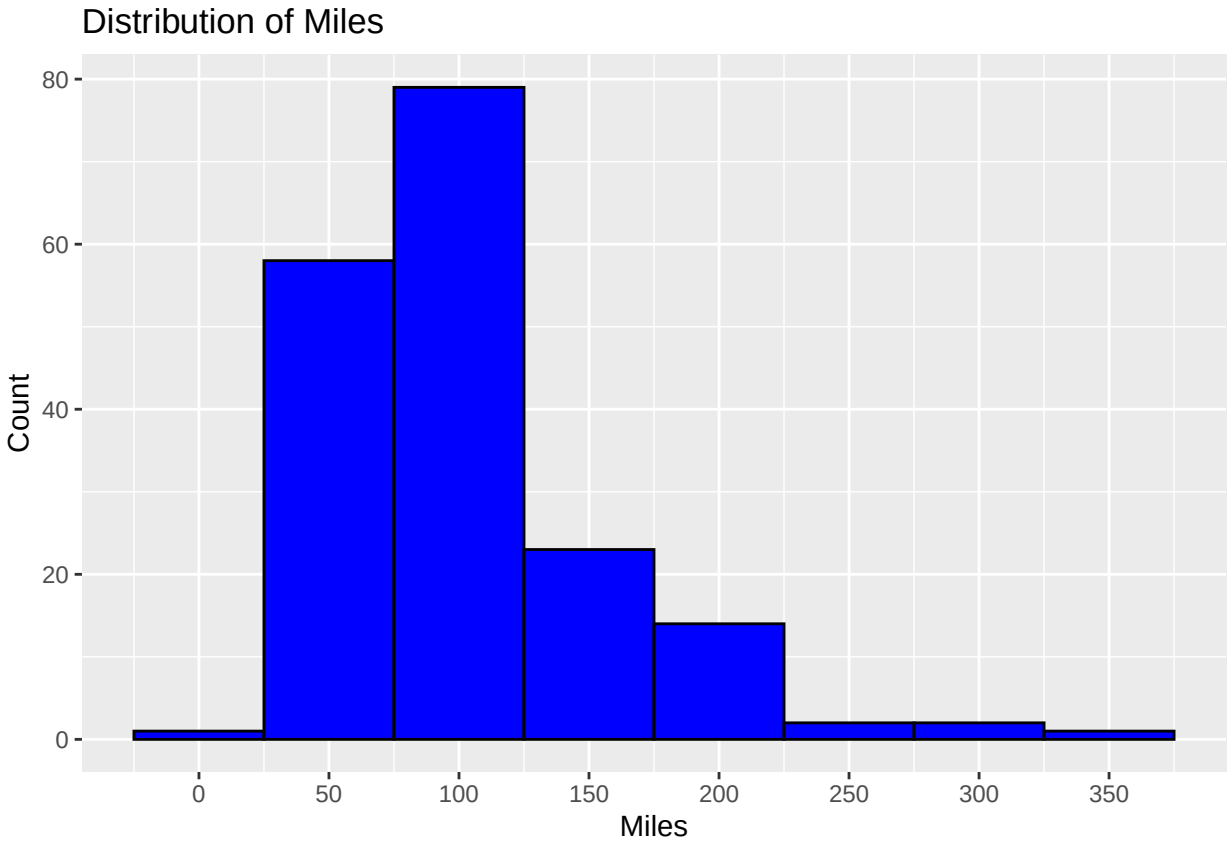


The graph is right skewed. The most common range of income is between 40000 and 60000. Very few customers have higher income which makes the graph right skewed.

Table 7: Income Summary

Statistic	Value
Min.	29562.00
1st Qu.	44058.75
Median	50596.50
Mean	53719.58
3rd Qu.	58668.00
Max.	104581.00

9. Miles



It is a right skewed graph. An average customer expects to run around 100 miles on their treadmill.

Table 8: Miles Summary

Statistic	Value
Min.	21.0000
1st Qu.	66.0000
Median	94.0000
Mean	103.1944
3rd Qu.	114.7500
Max.	360.0000

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

-> According to the data,

Table 9: Average Days and Miles Expected on Treadmill

Avg_Usage	Avg_Miles
3.455556	103.1944

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

-> The averages can be observed as below:

Table 10: Averages of Treadmill Buyers

Avg_Income	Avg_Age	Avg_Education
53719.58	28.78889	15.57222

Bivariate Analysis

1. Age vs Model

Is there any relation between Age and model?

-> Forming hypothesis:

Null: No relation between Age and model

Alternate: Relation between Age and model

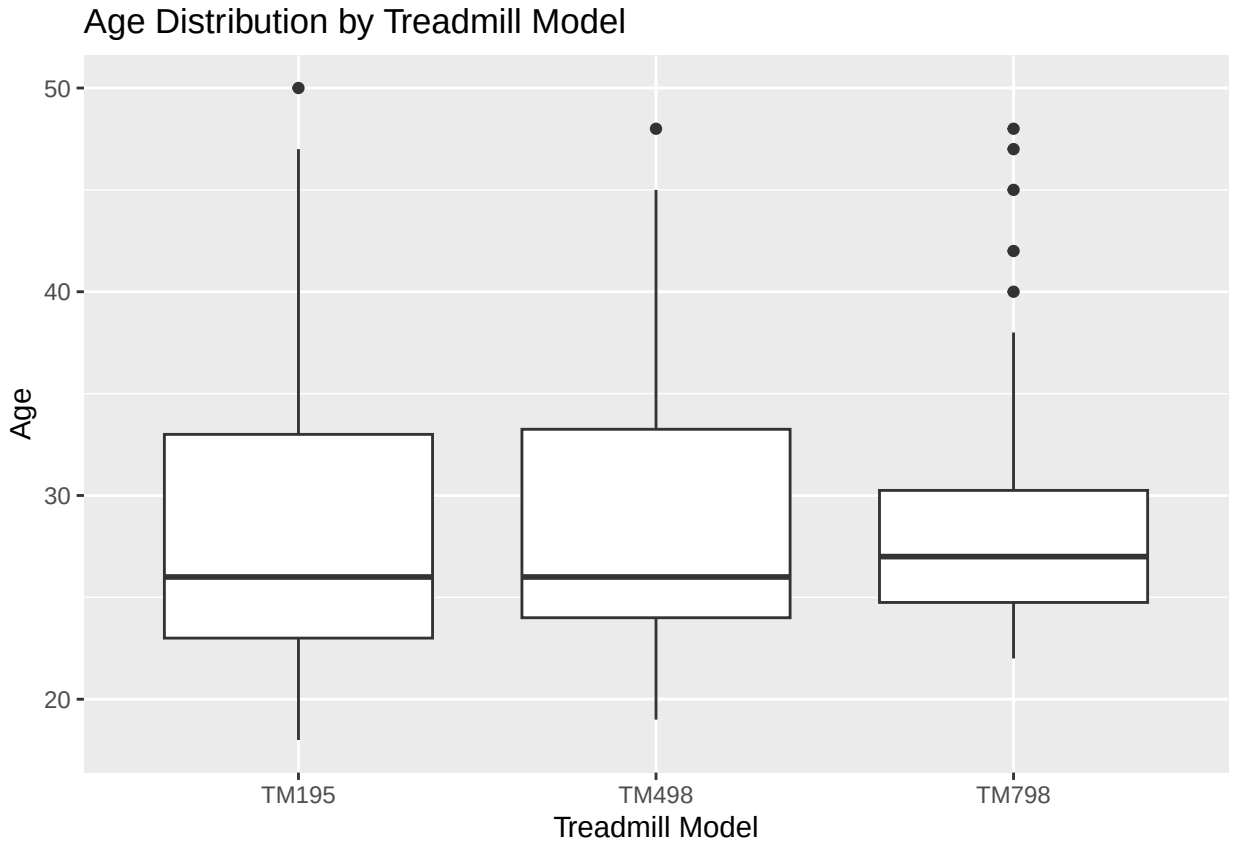
Performing ANOVA test:

Table 11: ANOVA Table: Age vs Product

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product	2	9.177778	4.588889	0.0942179	0.9101301
Residuals	177	8620.800000	48.705085	NA	NA

As p-value > 0.05, we do not have enough evidence to reject the null hypothesis. So we conclude that there is no relation between age and model choice.

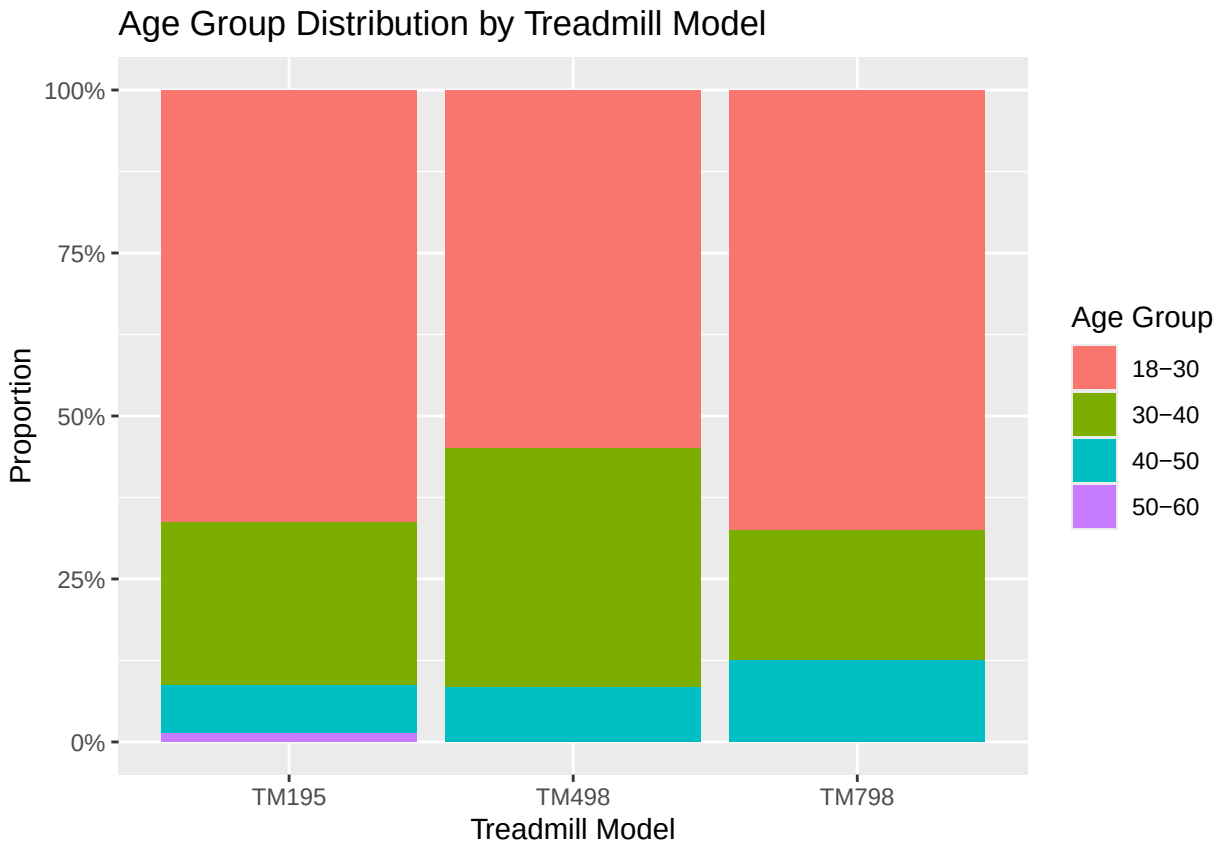
We can visualize the relation with box plot below:



It can be observed that the medians for the different treadmill models are similar. TM195 and TM498 have similar distribution but TM798 have smaller distribution with a few outliers.

Are different age groups buying different models?

-> Visualizing:



As per the figure above, people of age 18-30 are the majority buyers for all the models. While people of age 30-40 are slightly more inclined to buy model TM498.

2. Income vs Model

Is there any relation between Income and model?

-> Forming hypothesis:

Null: There is no relation between Income and Model of Treadmill.

Alternate: There is relation between Income and Model of Treadmill.

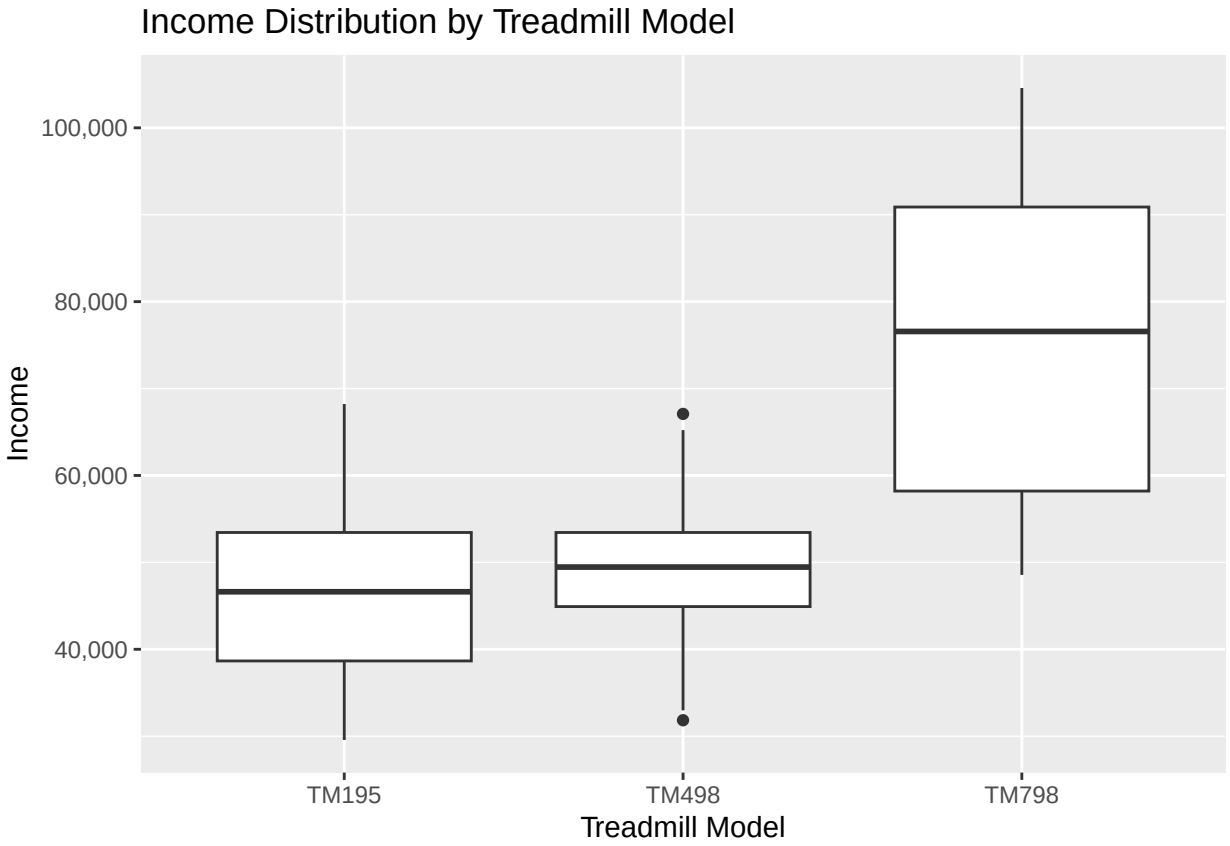
Performing ANOVA test:

Table 12: ANOVA Table: Income ~ Product

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product	2	24490250199	12245125099	89.25904	0
Residuals	177	24281991523	137186393	NA	NA

As, F-value is very high and p-value < 0.05(typical significant value); we do not have enough evidence to accept the null hypothesis. Hence, there is relation between income and model of treadmill.

We can observe the relation below:



TM798 is more popular among higher income while TM195 and TM498 are popular among low income.

3. Fitness vs Model

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

-> Forming hypothesis;

Null: No relation between self-rated fitness and model.

Alternate: There is relation between self-rated fitness and model

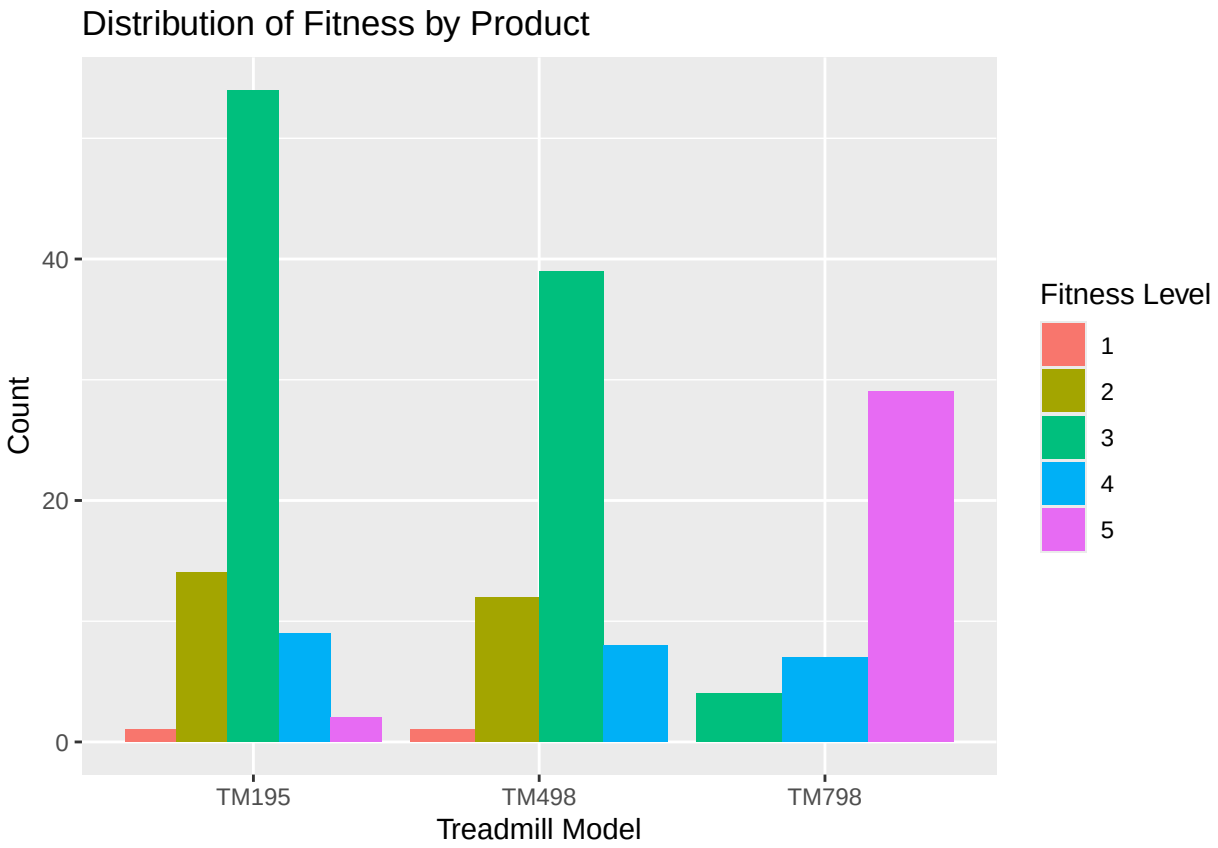
Performing Chi-Square Test for categorical variable:

Table 13: Chi-Square Test Results: Fitness vs Product

	Statistic	DF	P_value
X-squared	118.7768	8	0

For p-value close to 0, we do not have enough evidence to accept null hypothesis at 0.05 level, so we conclude self-rated fitness and model have a relation.

Observing with plot:



For the model TM195 and TM498, people with fitness level 3 are more attracted. However, the model TM798 is bought by more fitness enthusiasts with fitness greater than 3.

4. Education vs Model

Is there any relation between education and model?

-> Forming Hypothesis:

Null: There is no relation between education and model

Alternate: There is relation between education and model.

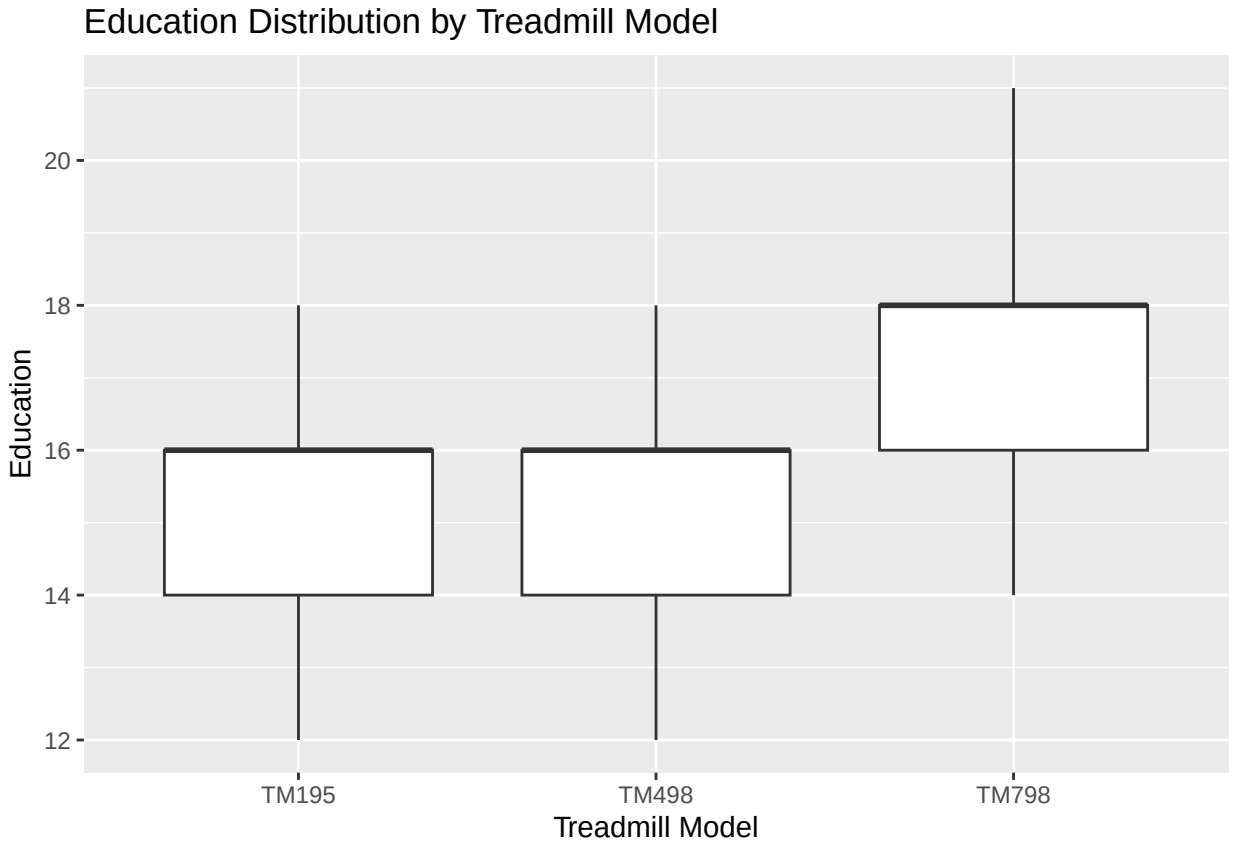
Performing Test:

Table 14: ANOVA Table: Education vs Product

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product	2	158.2153	79.107639	45.19038	0
Residuals	177	309.8458	1.750541	NA	NA

As the p-value is close to 0, we reject the null hypothesis and conclude that we have enough evidence to support the alternative hypothesis, which says that there is relation between the education level and model of the treadmill.

We can further explore the relation with a plot:



It can be observed that people with more years of eduaction choose TM798 model.

5. Gender vs Model

Does gender have any effect on model customer buy?

->Forming Hypothesis:

Null Hypothesis: Gender doesn't have any effect on model customer buy.

Alternate Hypothesis: Gender has effect on the model customer buy.

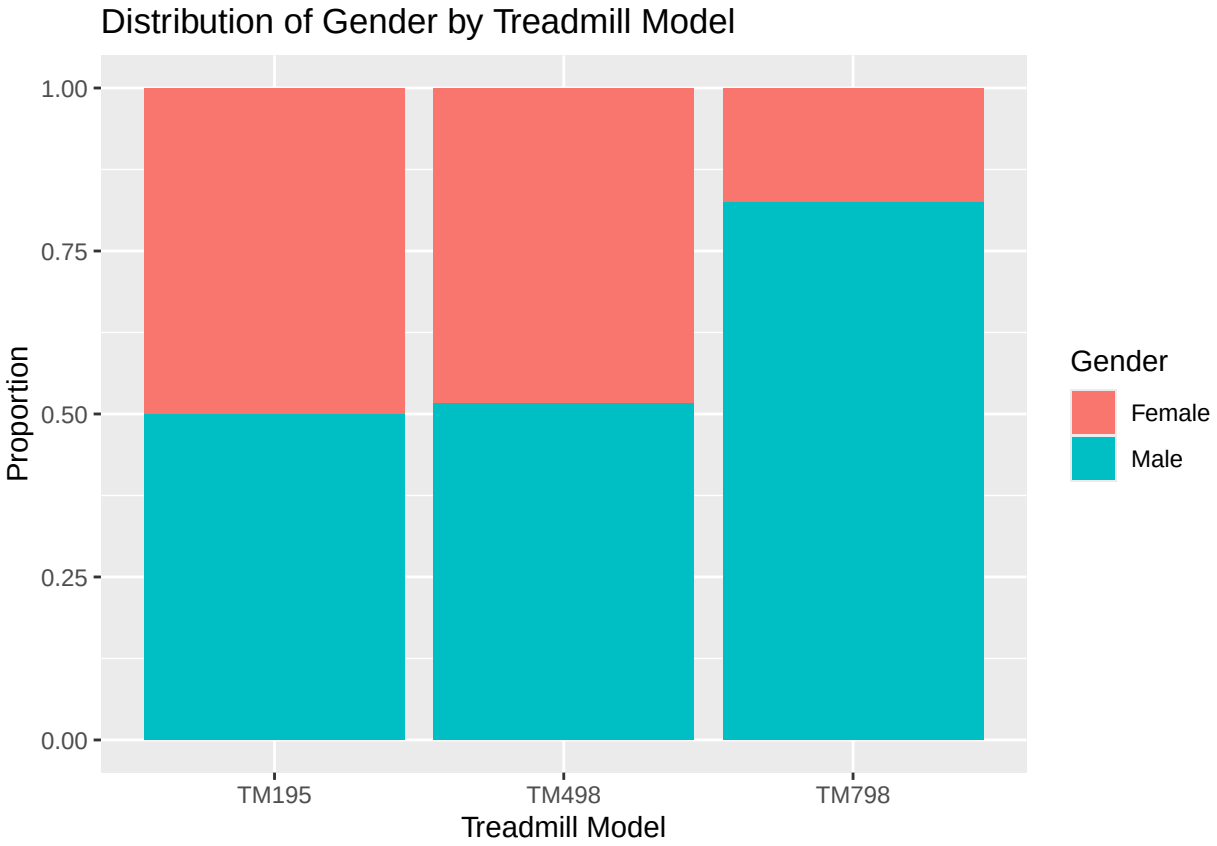
Performing Test:

Table 15: Chi-Square Test: Gender vs Treadmill Model

	Statistic	DF	P_value
X-squared	12.92384	2	0.0015618

As the p-value is close to 0, we reject the null hypothesis and conclude that we have enough evidence to support the alternative hypothesis, which says that there is relation between the gender and model of the treadmill.

Visualizing the relation:



There is equal proportion for male and female for TM195 and TM498 but male are more attracted to the TM798 model than females.

6. Gender vs Model

Does Martial status have any effect model customer buy?

-> Forming hypothesis:

Null Hypothesis: Martial status has no effect on model

Alternate Hypothesis: Martial status has effect on the model customer choose.

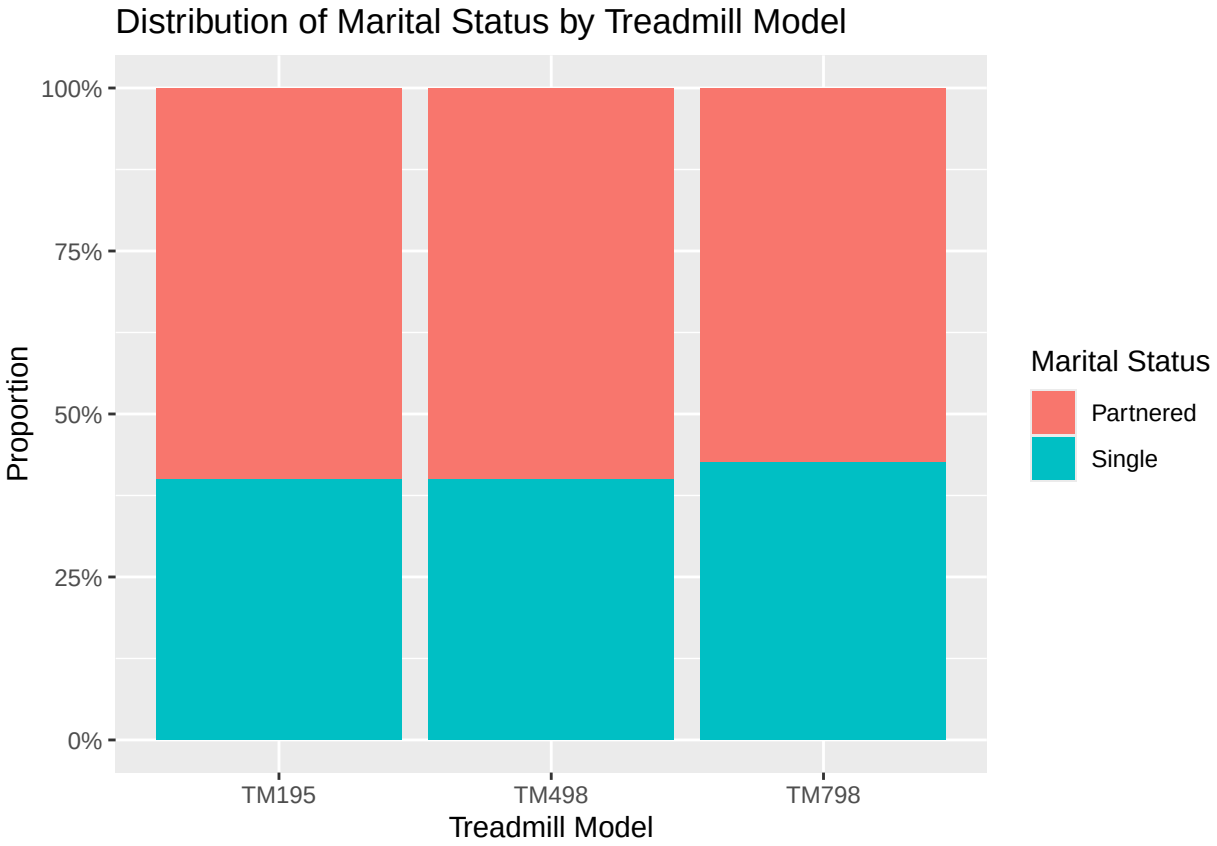
Performing Test:

Table 16: Chi-Square Test: Martial Status vs Treadmill Model

	Statistic	DF	P_value
X-squared	0.0806555	2	0.9604746

As, p-value (0.9604746) > 0.05, we fail to reject the null hypothesis, and conclude that there is not enough evidence to suggest that the martial status has effect on the model customer choose.

Visualizing:



As per the figure, equal proportion of martial status is found in all the models.

7. Usage vs Model

Is there relation between usage and choice of model?

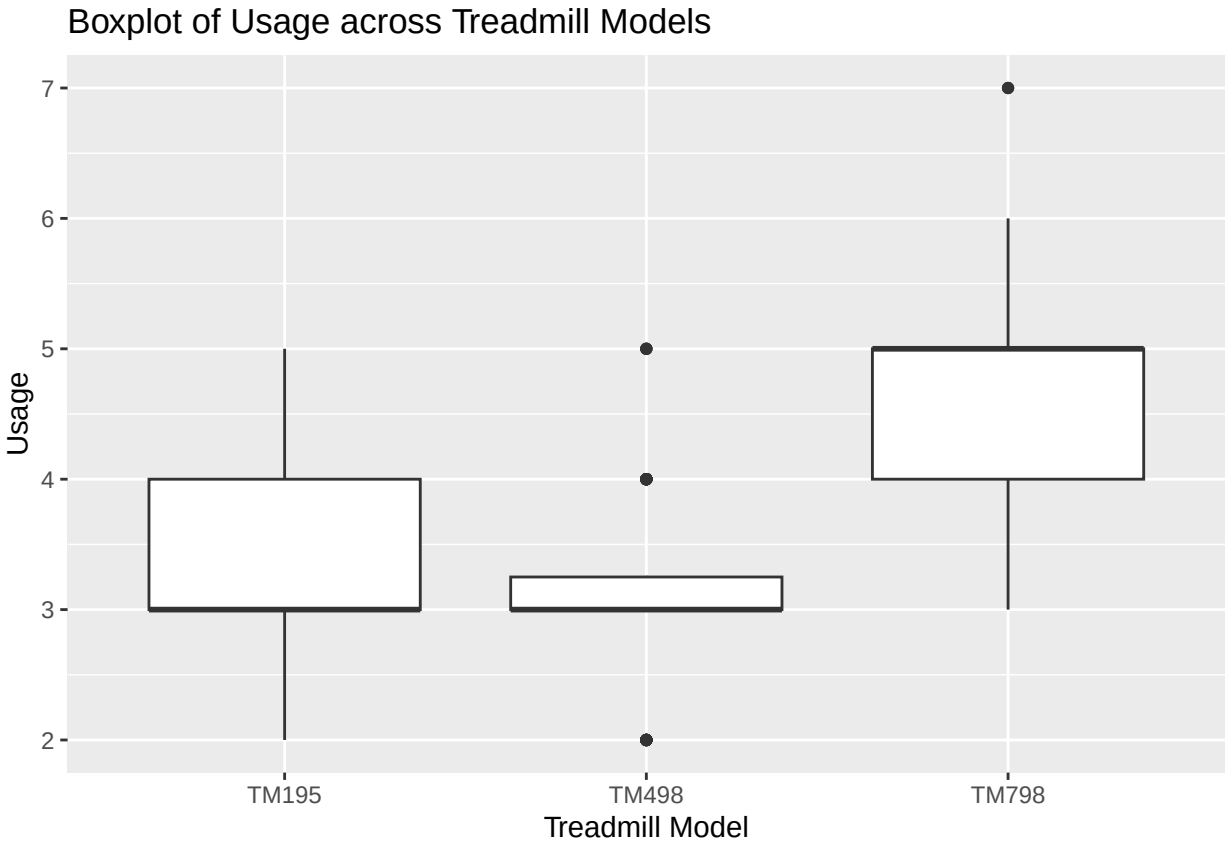
-> Forming hypothesis: Null: There is no relation between usage and choice of model. Alternate: There is relation between usage and the choice of model.

Performing Test:

Table 17: ANOVA Table: Usage vs Product

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product	2	89.54861	44.7743056	65.44447	0
Residuals	177	121.09583	0.6841573	NA	NA

As, p-value is closer to $0 < 0.05$, we do not have enough evidence to reject the null hypothesis, thus we accept the alternate hypothesis. So, accepting the alternate hypothesis, we can say that, there is relation between usage and the choice of model



It can be observed from the box plot that the model TM498 and TM195 have similar average usage but the usage of model TM195 is more varying. The model TM798 however has higher average usage compared to other two models.

8. Miles vs Product

Is there a relation between model and miles?

-> Null Hypothesis: There is no relation between model and miles

Alternate Hypothesis: There is a relation between model and miles ran.

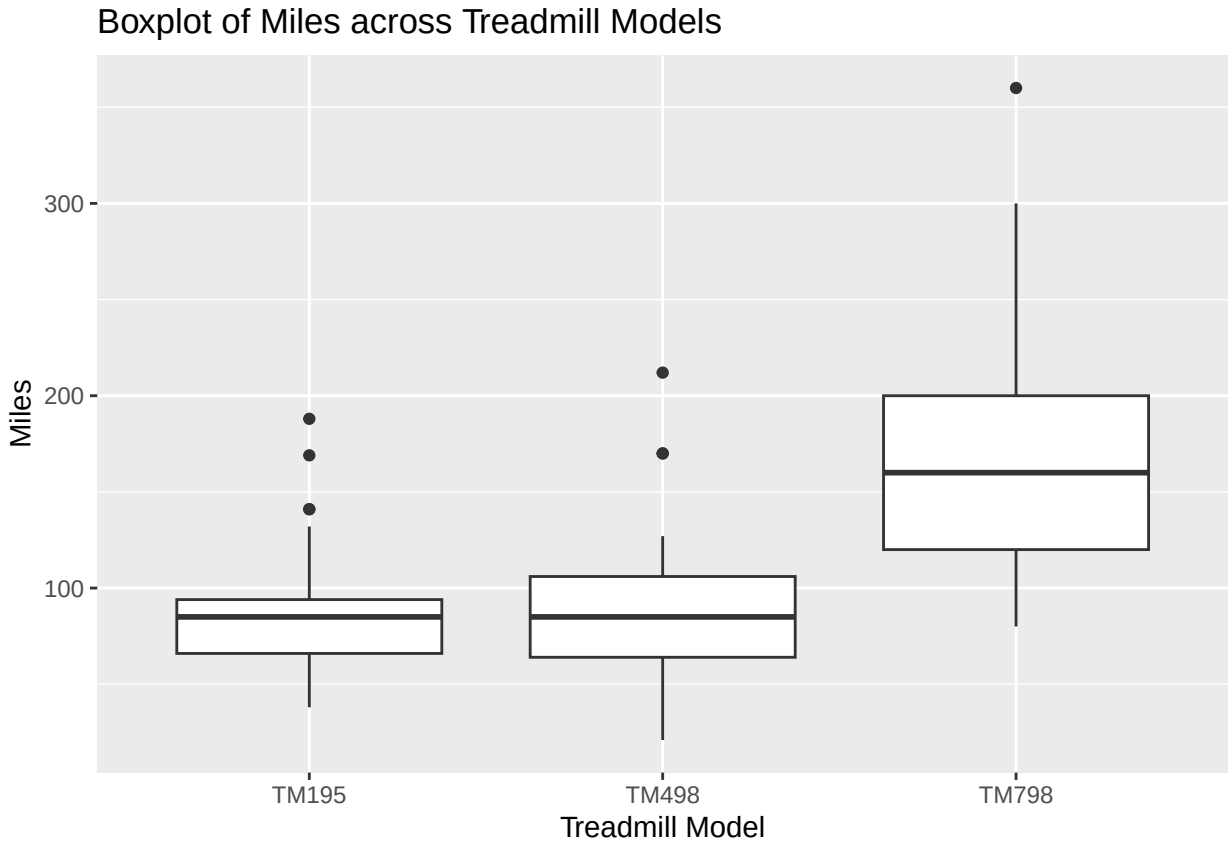
Performing test:

Table 18: Miles vs Product

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Product	2	209625.5	104812.737	68.2418	0
Residuals	177	271854.7	1535.902	NA	NA

As p-value is close to 0, we do not have enough evidence to support the null hypothesis. Thus, we accept the alternate hypothesis that models and the miles run have a realtion.

Visualizing:



It can be observed that people who run longer miles on average choose TM798 model. The remaining two have similar but lower average miles covered with varying distribution.

Multivariate Analysis

Relation between Age, Income and education and model bought?

->Visualizing:

From the above figure, it can be seen that people with higher income and in their middle ages tend to choose TM798 while the other two models have broader distribution over the income and ages.

It can also be noticed that people of ages below 30 tend to buy TM798 model even with slightly below mean income.

- **For Product Recommendation based on Income**

Performing multiple regression:

```
##
## Call:
## lm(formula = Income ~ Age + Fitness + Education + Product, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -24019  -4166   -599   4004  25969
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -6262.16    7690.21  -0.814  0.41658
## Age           1089.11      93.97  11.590  < 2e-16 ***
## Fitness         31.45     948.25   0.033  0.97358
## Education     1429.29     494.83   2.888  0.00436 **
## ProductTM498  2063.25    1405.64   1.468  0.14395
## ProductTM798 25102.74    2537.39   9.893  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 8219 on 174 degrees of freedom
## Multiple R-squared:  0.759,  Adjusted R-squared:  0.752
## F-statistic: 109.6 on 5 and 174 DF,  p-value: < 2.2e-16
```

For Recommendation based on Income for various models:

Choosing model TM195 as baseline,it can be observed that customers who buy the TM498 model tend to have an income $2,063 higher but this is not statistically significant as p-values is 0.14395. However, for model TM798 the income is 25102.74 higher than the baseline and it is highly significant as p is very small. Thus, income matters for product TM798.

How does Model affect the interation between Age and Fitness?

```
##              Df Sum Sq Mean Sq F value   Pr(>F)
## Product       2  81600   40800   41.16 2.09e-15 ***
## Residuals   177 175449     991
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As p-value < 0.05, we can conclude that choice of model is affected by the interaction between Age and Fitness eventhough from above test we figured that there is no relation between Model and Age.

Summary

Univariate Analysis Results:

- The store has 3 models: "TM195" "TM498" and "TM798".
- TM195 is the most sold model.
- Male customers buy more treadmills compared to Female.
- People of age 23-28 are the most popular age group buying treadmill .
- Customers of education 14-16 years are of highest frequency.
- Married people are buying more Treadmills than Single customers.
- People on average want to use the treadmill 3 times a week.
- Fitness level 3 is the most common, with close to 100 customers rating themselves at this level.
- Average customer has income of 53719.58.
- An average customer expects to run around 100 miles on their treadmill.

Bivariate Analysis Results:

- There is no relation between Age and Model chosen. Age distribution is less varying for TM798 model compared to TM195 and TM498 models and has few outliers. People of age 30-40 prefer TM498 model more than other ones. Age 18-30 are the majority buyers for all the models.
- TM798 is more popular among higher income (Average 80000) while TM195 and TM498 are popular among low income (Average 45000).
- Customers who rate their fitness at average of 3 choose TM195 or TM498. People who rate their fitness above average (mostly 5 (maximum)) fitness choose the TM798 model.
- Customers with more years of education choose model TM798 (Average 18 years) while average of 16 choose the other models.
- Proportion of male is significant for TM798 model. Others have equal proportions.
- Partnered and Single are in equal proportion for all the models.
- Customer with higher expectation of usage i.e., around 5 choose the model TM798 model. For average usage around 3 days choose TM498 and for varying usage of 3-4 days they prefer TM195.
- For miles lesser than 100 miles, customer prefer TM195 and for more varying miles around 100, they choose TM498. For miles around 150 customers prefer TM798.

Multivariate Analysis

- Customer with higher income and in their middle age tend to choose TM798.
- Customers of ages below 30 also tend to buy TM798 model even with slightly below mean income.
- Customers who buy the TM498 model tend to have an income $2,063 higher than TM195, but this is not statistically significant. For model TM798 the income is 25102.74 higher than TM195 and it is highly significant.
- The interaction between Age and Fitness significantly affects the choice of treadmill. Even though the individual effect of Age may not be significant, its interaction with Fitness plays a vital role in determining model preference.

Customer Profile Summary based on model

- **TM195**: Younger, cost-conscious customers with average fitness levels and moderate usage expectations.
- **TM498**: Middle-aged customers seeking value for money and moderate performance, interested in maintaining regular fitness routines.
- **TM798**: High-income, highly educated, and fitness-committed customers looking for advanced features and performance, willing to invest in their fitness.

Conclusion

There is distinction between customer profiles between the three models which suggests that each model tends to different customer needs. Income is a major factor in choosing the models. Higher income customers are more likely to buy TM798 which is also corelated with higher education and fitness rating. Otherwise customers choose the more simple treadmills. Customers for TM798 also demostrate higher committment towards fitness with higher usage and miles expectation. TM195 buyers focus more on affordability and medium usage, TM498 buyers want balance between cost and performance while TM798 buyers prioritize advanced features over price.

Recommendation

- Target high-income and more fitness conscious customers for TM798 model.
- Promote TM498 for customers who want a balance between perfomance and affordability and to a middle aged demographich of 30-40 years.
- Promote TM195 to a more younger audience given they probably have lower level of income and mention it's interest over the audince who care for affordibility.
- Married male customers must be focused as they invest on more high end TM798 model. Also there can be efforts to attract more female demogrpachic for TM798 model.
- TM798 model should be advertised as for serious performance for fitness consious runner, TM498 for reliability and affordibity and TM195 for starters.
- As TM195 is the most sold product, sales of TM498 can also be increased with more adverisement as it appeals to similar profiles of customers.