Aerofit Descriptive Statistics and Probability

Introduction

About Aerofit

Aerofit, a dynamic player in the fitness industry, traces its origins to M/s. Sachdev Sports Co, established in 1928 by Ram Ratan Sachdev. From its modest beginnings in Hyderabad, India, the company evolved into a leading sports equipment supplier across Andhra Pradesh and Telangana. Recognizing the growing need for fitness solutions, M/s. Sachdev Overseas emerged to import quality fitness equipment under the "Aerofit" brand, ensuring affordability and post-sales excellence.

Aerofit provides a product range including machines such as treadmills, exercise bikes, gym equipment, and fitness accessories to cater to the needs of all categories of people.

Objective

The market research team at AeroFit wants to identify the characteristics of the target audience for each type of treadmill offered by the company, to provide a better recommendation of the treadmills to the new customers. The team decides to investigate whether there are differences across the product with respect to customer characteristics. Perform descriptive analytics **to create a customer profile** for each AeroFit treadmill product by developing appropriate tables and charts.For each AeroFit treadmill product, construct **two-way contingency tables** and compute all **conditional and marginal probabilities** along with their insights/impact on the business.

Features of the Dataset:

Product Purchased	:	KP281, KP481, or KP781
Age	:	In years
Gender	:	Male/Female
Education	:	In years
Marital Status	:	Single or partnered
Usage	:	The average number of times the customer plans to use the treadmill each week.
Income	:	Annual income (in $)
Fitness	:	Self-rated fitness on a 1-to-5 scale, where 1 is the poor shape and 5 is the excellent shape.
Miles	:	The average number of miles the customer expects to walk/run each week

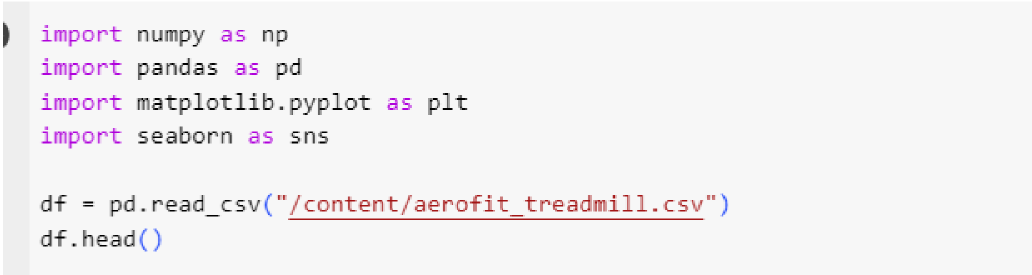
Product Portfolio:

The KP281 is an entry-level treadmill that sells for $1,500.

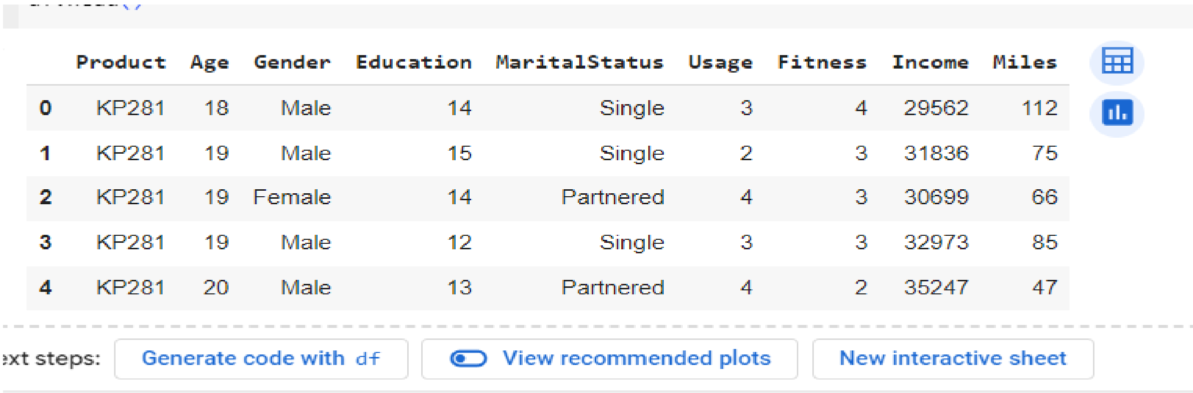
The KP481 is for mid-level runners that sell for $1,750.

The KP781 treadmill is having advanced features that sell for $2,500.

Importing Libraries:

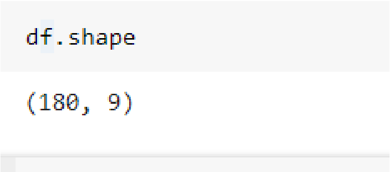


```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

df = pd.read_csv("/content/aerofit_treadmill.csv")
df.head()
```



	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4	29562	112
1	KP281	19	Male	15	Single	2	3	31836	75
2	KP281	19	Female	14	Partnered	4	3	30699	66
3	KP281	19	Male	12	Single	3	3	32973	85
4	KP281	20	Male	13	Partnered	4	2	35247	47

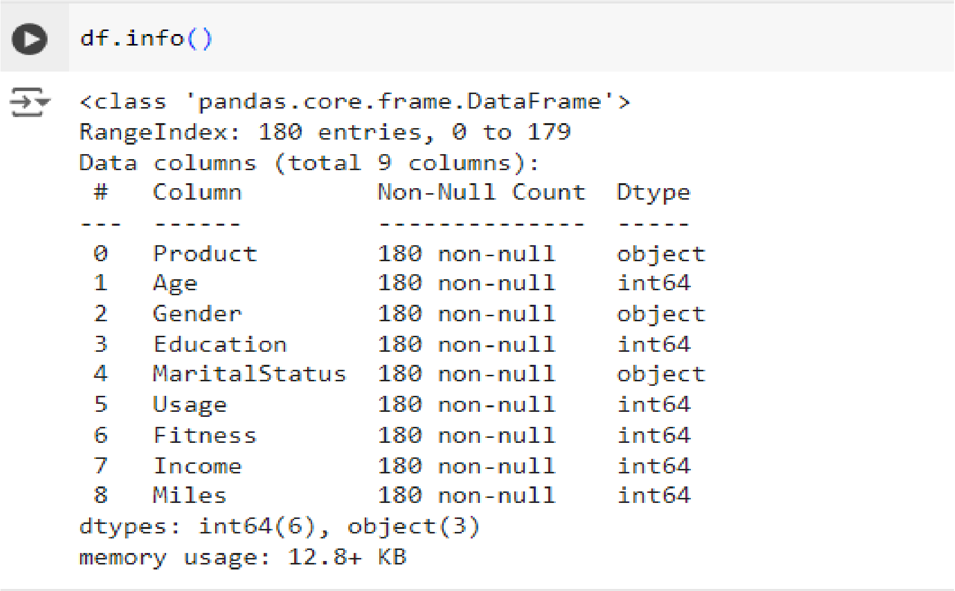
Finding the DataFrame Dimension



```
df.shape
```

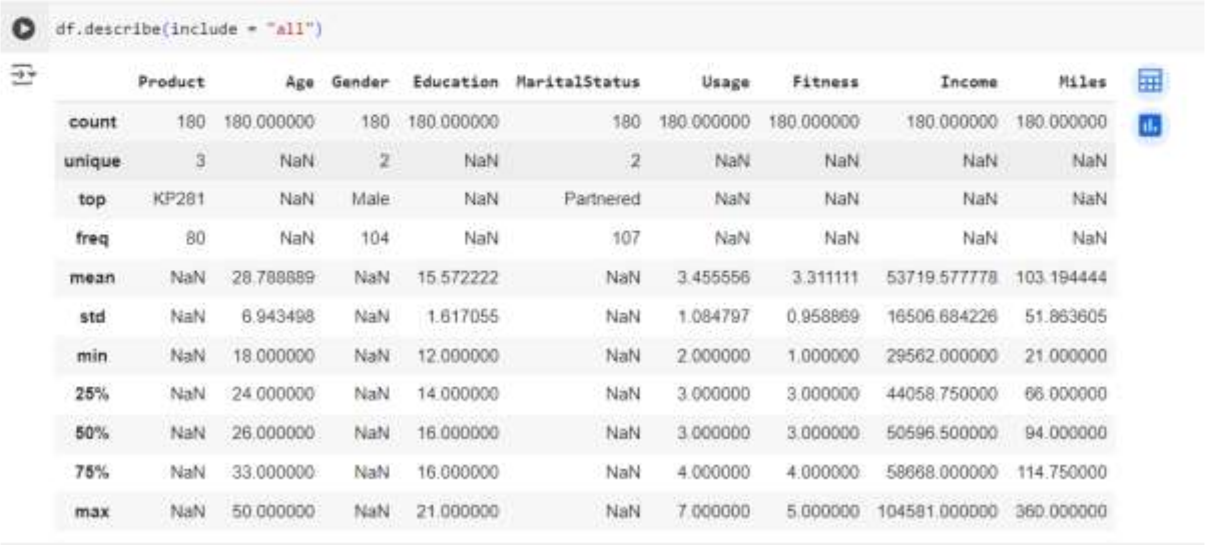
```
(180, 9)
```

Exploration of Data



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Product        180 non-null    object
 1   Age            180 non-null    int64 
 2   Gender         180 non-null    object
 3   Education      180 non-null    int64 
 4   MaritalStatus  180 non-null    object
 5   Usage          180 non-null    int64 
 6   Fitness        180 non-null    int64 
 7   Income         180 non-null    int64 
 8   Miles          180 non-null    int64 
dtypes: int64(6), object(3)
memory usage: 12.8+ KB
```



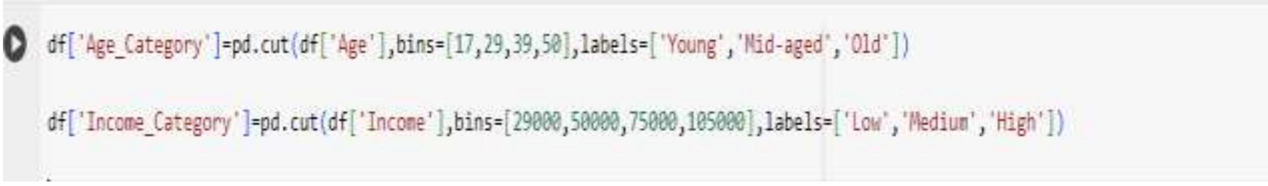
```
df.describe(include = "all")
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
count	180	180.000000	180	180.000000	180	180.000000	180.000000	180.000000	180.000000
unique	3	NaN	2	NaN	2	NaN	NaN	NaN	NaN
top	KP281	NaN	Male	NaN	Partnered	NaN	NaN	NaN	NaN
freq	80	NaN	104	NaN	107	NaN	NaN	NaN	NaN
mean	NaN	28.788889	NaN	15.572222	NaN	3.455556	3.311111	53719.577778	103.194444
std	NaN	6.943498	NaN	1.617055	NaN	1.084797	0.958869	16506.684226	51.863605
min	NaN	18.000000	NaN	12.000000	NaN	2.000000	1.000000	29562.000000	21.000000
25%	NaN	24.000000	NaN	14.000000	NaN	3.000000	3.000000	44058.750000	66.000000
50%	NaN	26.000000	NaN	16.000000	NaN	3.000000	3.000000	50596.500000	94.000000
75%	NaN	33.000000	NaN	16.000000	NaN	4.000000	4.000000	58668.000000	114.750000
max	NaN	50.000000	NaN	21.000000	NaN	7.000000	5.000000	104581.000000	360.000000

Insights:

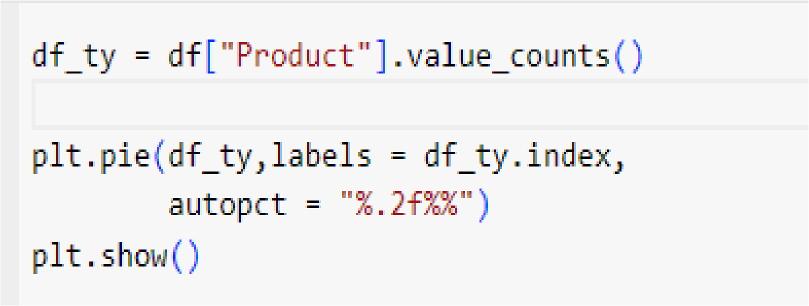
This dataset comprises 180 rows and 9 columns, including 6 integer columns and 3 category columns. There are no null or missing values across any of the columns. Among the categorical columns, there are three unique products, with KP281 being the most prevalent. The age of individuals in the dataset spans from 18 to 50, with an average age of 28.78. Notably, there is a noticeable gender imbalance, with a higher representation of males than females. Furthermore, the standard deviations for the "Income" and "Miles" variables are notably high, suggesting the possible presence of outliers in these data points

Adding Columns for better analysis:

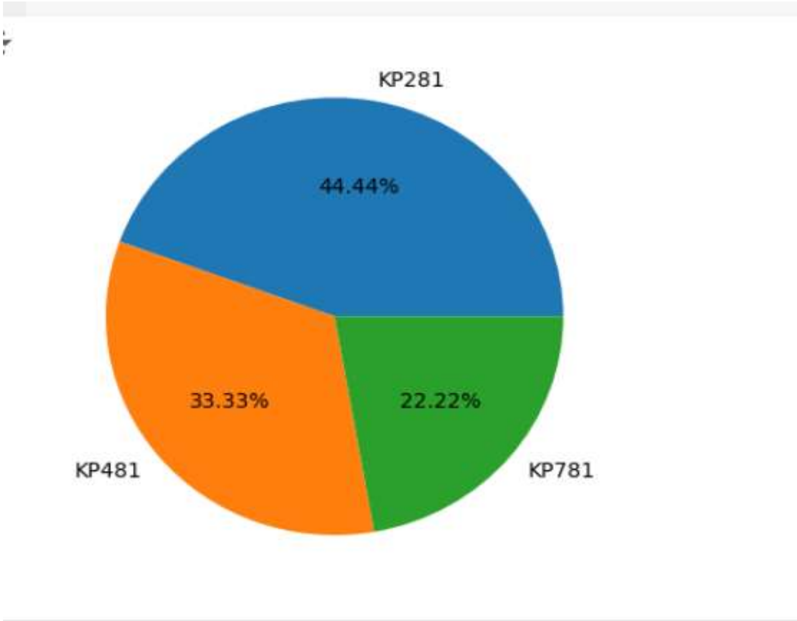


```
df['Age_Category']=pd.cut(df['Age'],bins=[17,29,39,50],labels=['Young','Mid-aged','Old'])

df['Income_Category']=pd.cut(df['Income'],bins=[29000,50000,75000,105000],labels=['Low','Medium','High'])
```

Univariate Analysis

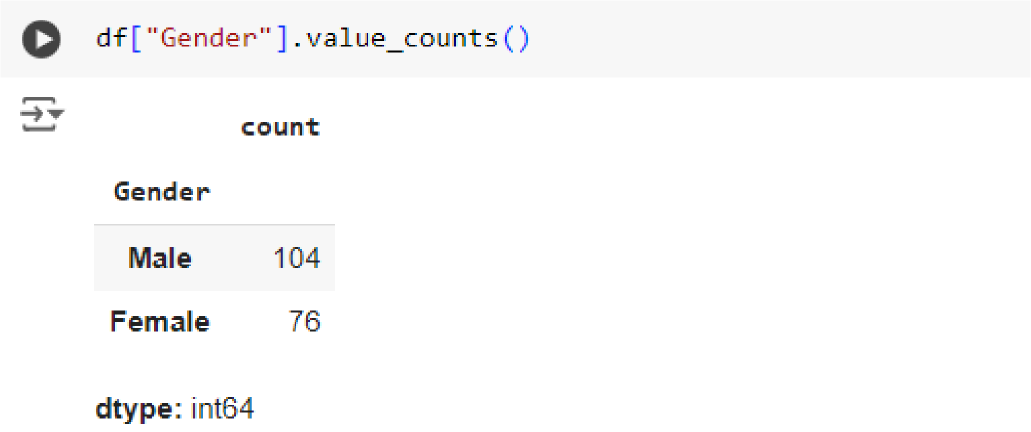


```
df_ty = df["Product"].value_counts()

plt.pie(df_ty,labels = df_ty.index,
        autopct = "%.2f%%")
plt.show()
```



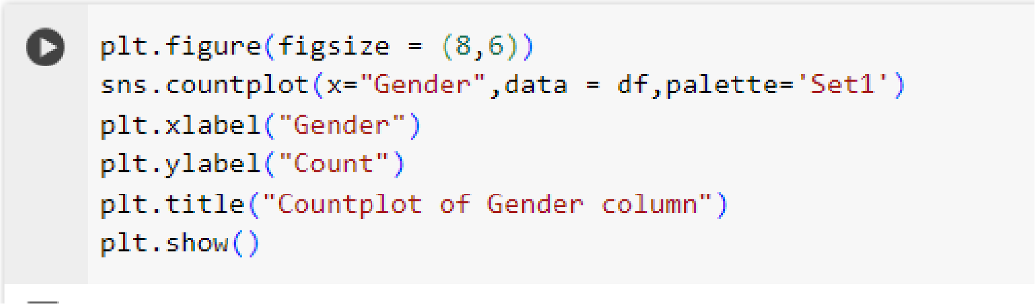
Insights:

Among users, 44.4% express a preference for the KP281 treadmill, while 33.3% favor the KP481 treadmill, and only 22.2% opt for the KP781 treadmill

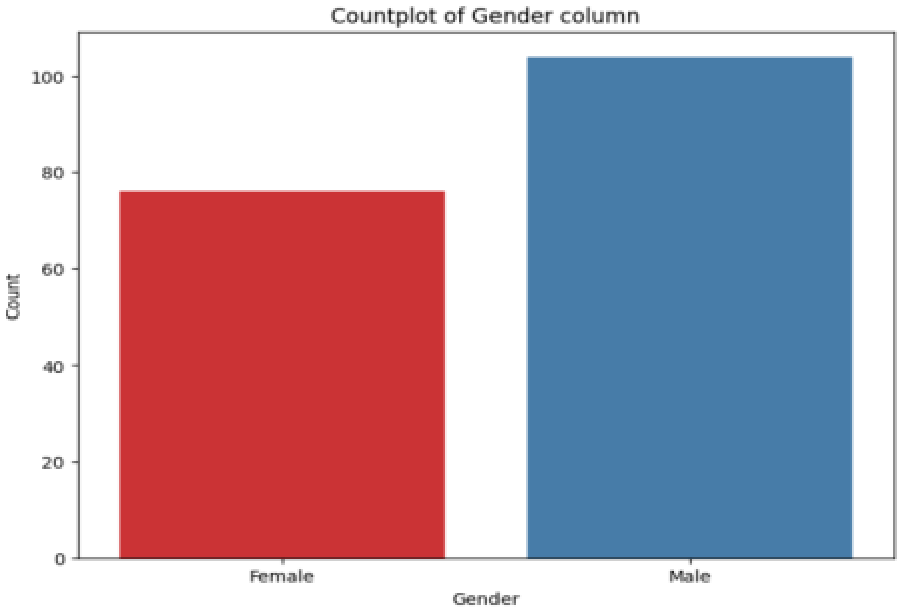


```
df["Gender"].value_counts()
```

```
          count
Gender
  Male      104
Female       76

dtype: int64
```



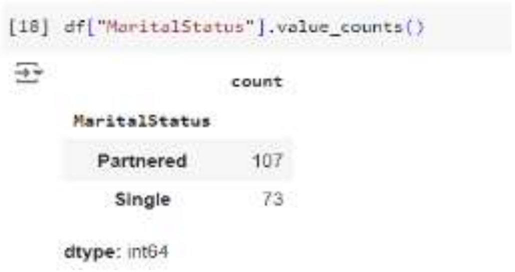
```
plt.figure(figsize = (8,6))
sns.countplot(x="Gender",data = df,palette='Set1')
plt.xlabel("Gender")
plt.ylabel("Count")
plt.title("Countplot of Gender column")
plt.show()
```



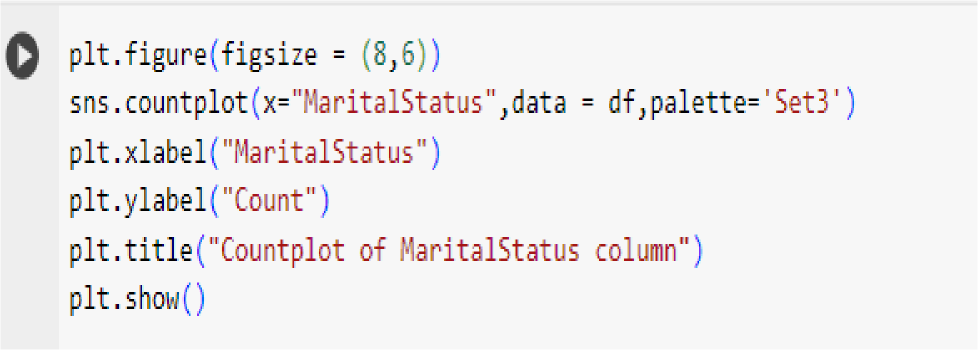
Insights: By observing the above data males are more in number comparing to the females because in general view lot of exercises are done by males. Males are 104 in number and Females are 76 in number.

Recommendations: Aerofit needs to focus on females more because of high in number they need to bring new machineries which can make exercises more effective for men. They need to encourage the females by giving more discounts on female machineries and equipment.

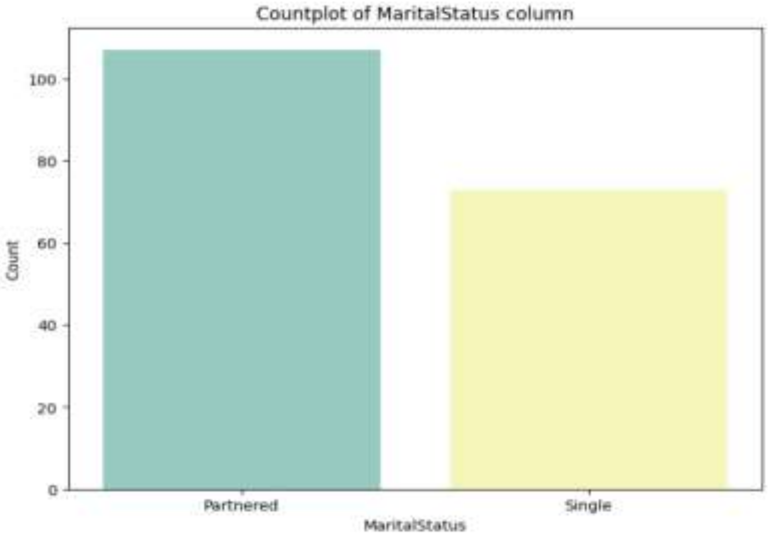
Analysis on Marital Status:



```
[18] df["MaritalStatus"].value_counts()

                 count
MaritalStatus
   Partnered       107
      Single        73

dtype: int64
```

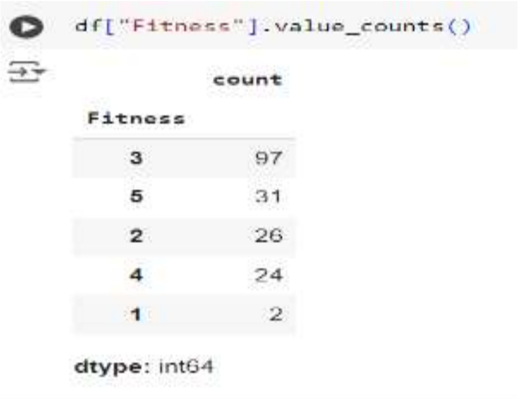


```
plt.figure(figsize = (8,6))
sns.countplot(x="MaritalStatus",data = df,palette='Set3')
plt.xlabel("MaritalStatus")
plt.ylabel("Count")
plt.title("Countplot of MaritalStatus column")
plt.show()
```



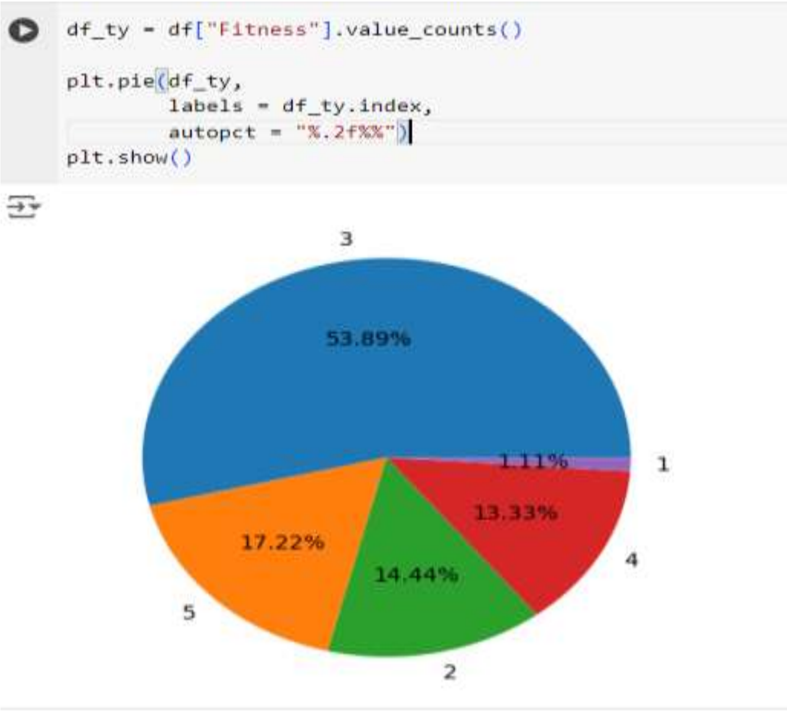
Insights: By observing the above data partnered the people who are married are more in number with 107 numbers and Single are with less number when comparing to the partnered with 73 in number

Analysis on Fitness rated Scale:



```
df["Fitness"].value_counts()
```

```
             count
Fitness
   3           97
   5           31
   2           26
   4           24
   1            2

dtype: int64
```

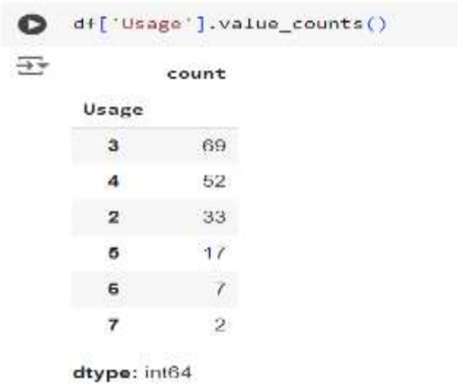


```
df_ty = df["Fitness"].value_counts()

plt.pie(df_ty,
        labels = df_ty.index,
        autopct = "%.2f%%")
plt.show()
```

Insights: By observing the above data three rating has highest number of users with 53.89% users, followed by five rating with 17.22% users, two rating with 14.44%,four rating with 13.33% and least rating is 1.11% is one.

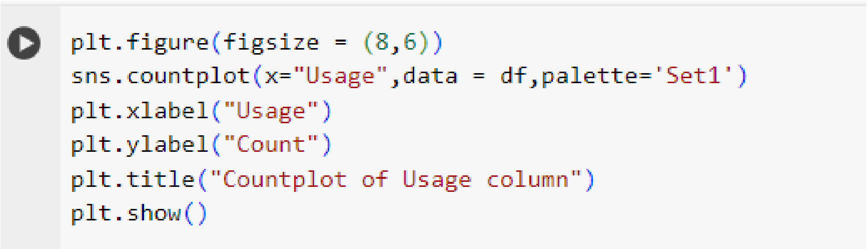
Recommendations: Aerofit needs to make the more efficient production on three rated equipment. Because most of the users belong to the middle scale range only few users are in the high range. It also need to motivate the other users to increase their rating of by conducting the workshops on importance of gym and body fitness which plays a vital role in everyone's life.

Analysis on Usage:

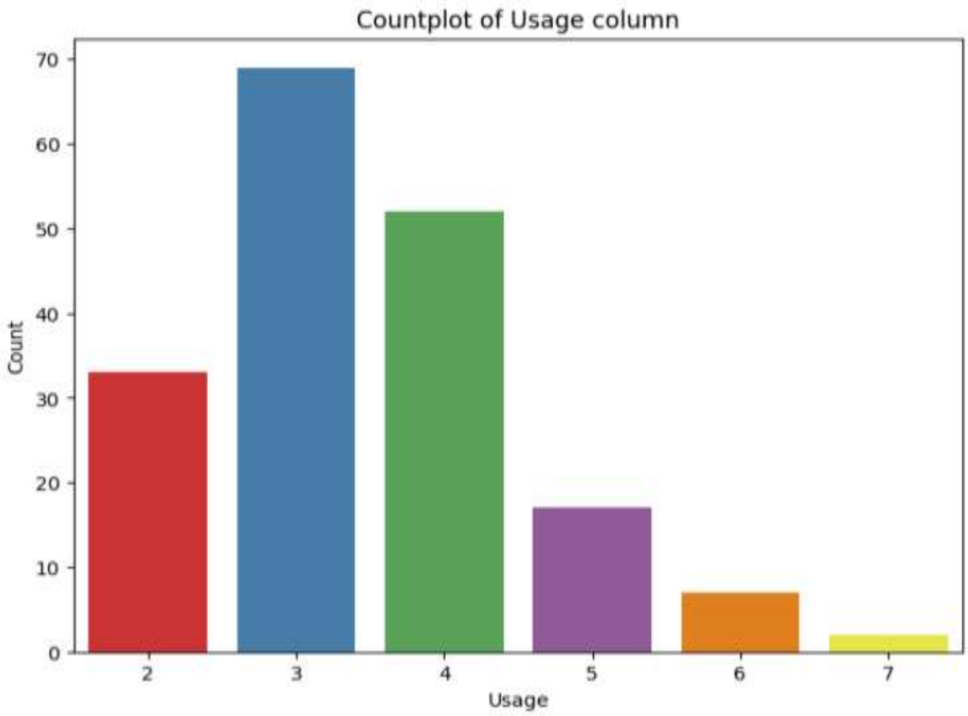


```python
df['Usage'].value_counts()
```

```
       count
Usage
  3       69
  4       52
  2       33
  5       17
  6        7
  7        2

dtype: int64
```

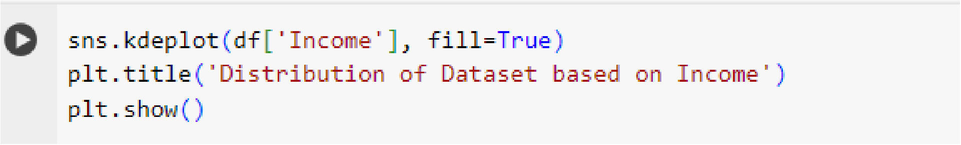


```python
plt.figure(figsize = (8,6))
sns.countplot(x="Usage",data = df,palette='Set1')
plt.xlabel("Usage")
plt.ylabel("Count")
plt.title("Countplot of Usage column")
plt.show()
```

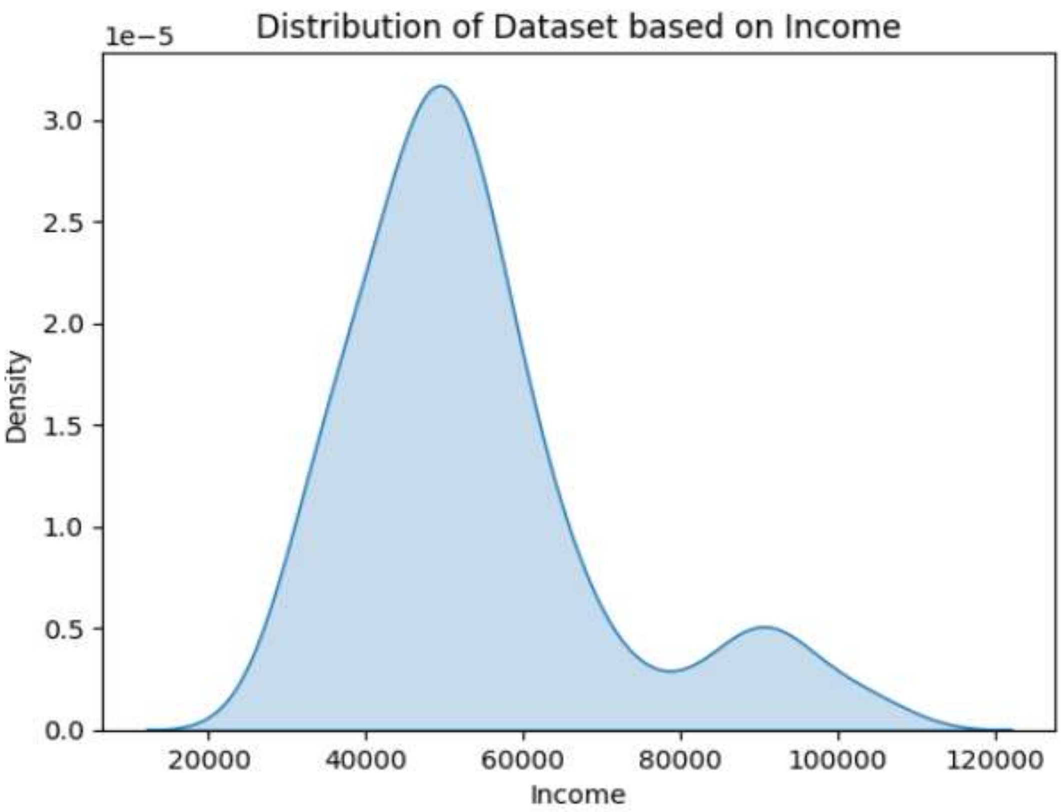


Insights: By observing the above data we can say that more customers prefer 3 times treadmill per week on average is 69 number of customers prefer this and followed by 52 number of customers with 4 times per week these are mostly used by customers, few customers prefer 2 times per week with 33 number of customers , 5 times with 17 customers , 6 times with 7 customers, 7 times with 2 number of customers as we see that number of times increases number of customers decreasing.

Analysis on income:



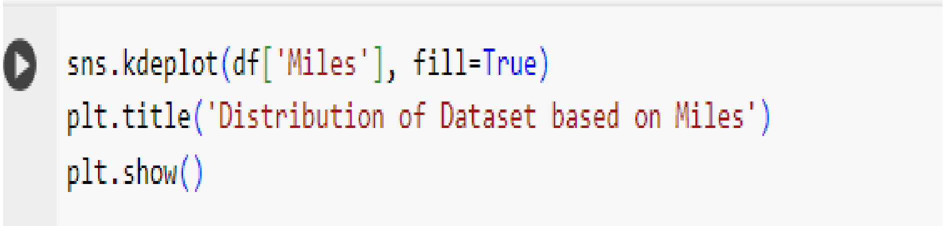
```
sns.kdeplot(df['Income'], fill=True)
plt.title('Distribution of Dataset based on Income')
plt.show()
```



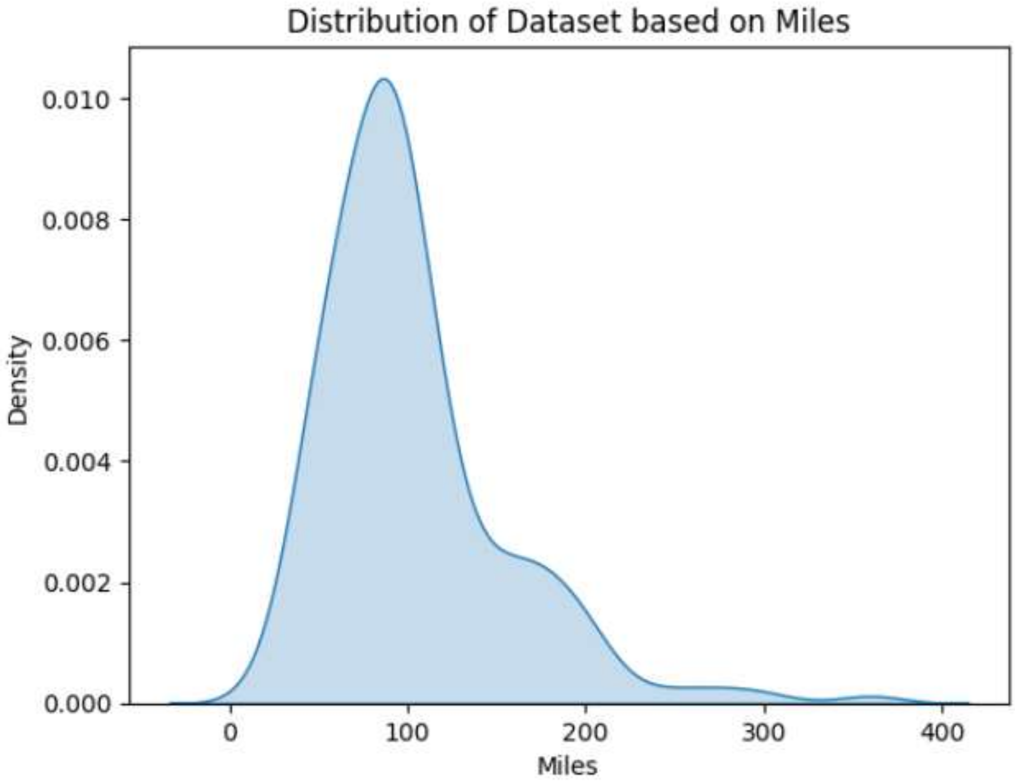
Insights: By observing the above data we can say that customers with income range between 40000 and 65000 are more in number comparing to the other customers, few customers have income range between 80000 and 100000 , very few customers belong to the category where income is above 100000.

Recommendations: Aerofit needs to focus more on customers whose income range between 40000 and 65000 most of the customers belong to this category the prices for the gym equipment should be normal range if the price of equipment is very heigh users cannot afford due to their middle income so price should be in normal range where everyone can afford it. This will increase the sales of gym equipment.

Analysis on Miles:



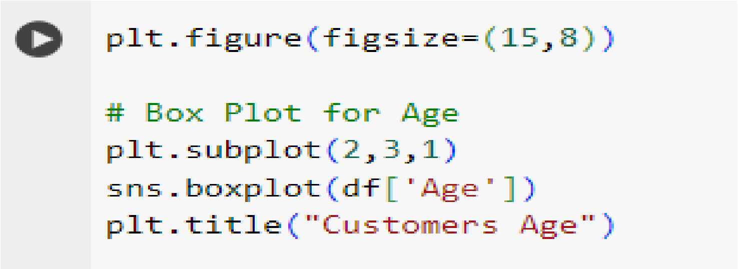
```
sns.kdeplot(df['Miles'], fill=True)
plt.title('Distribution of Dataset based on Miles')
plt.show()
```



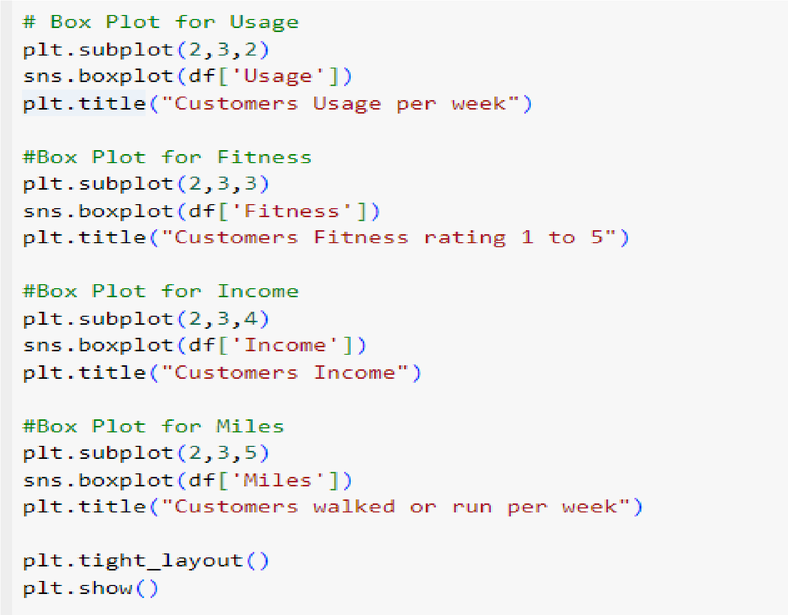
Insights: By observing the above data we can say that more number of customers prefer in between range of 50 to 120 miles per week after 150 the number of customers starts decreasing when it reach 400 miles per week very few run or walk 400 miles per week

Recommendations: Aerofit needs to focus on treadmills average usage because customers prefer according the average of their usage. So Aerofit should include some more advantages when the customers use their treadmills. It need to give all the timings and calories to be displayed after the activity with approximate values like heartbeat, calories burned, number of hours worked. These all need be displayed on the treadmills these will attract more and more customers adding some special features in Aerofit equipment results in increase in sales.

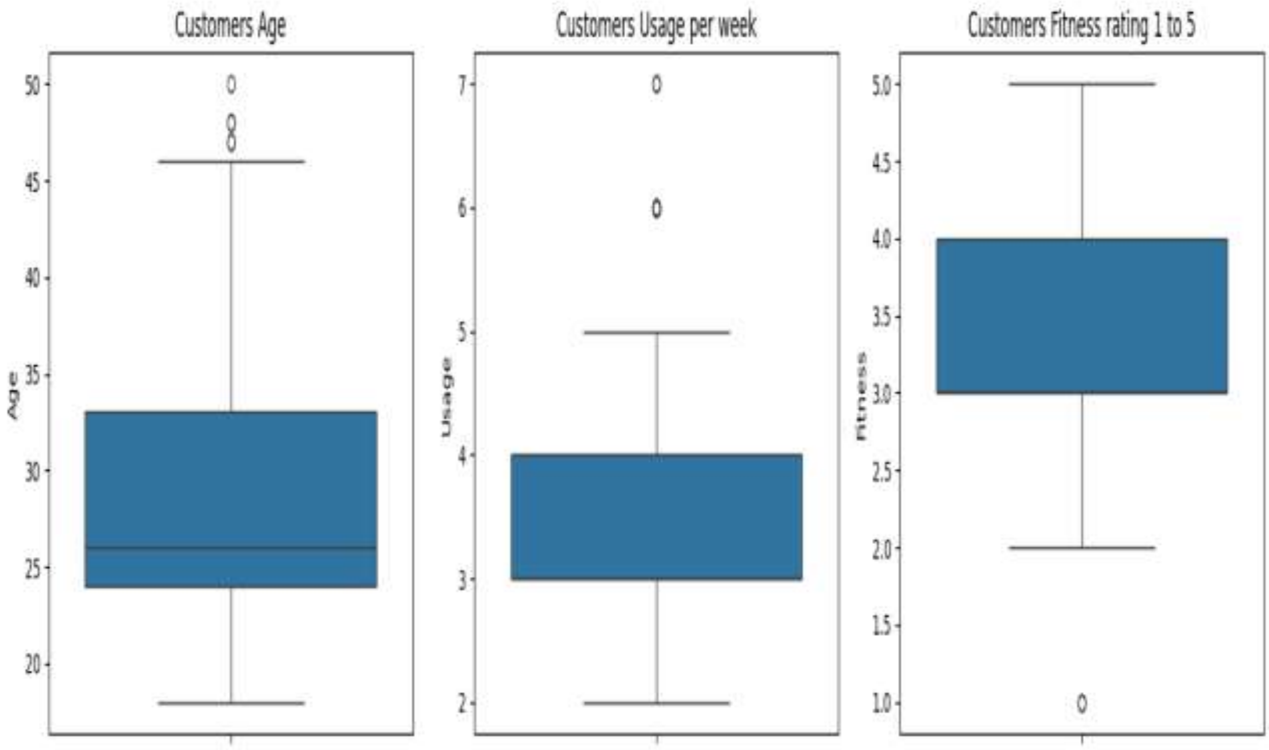
Detecting the outliers:

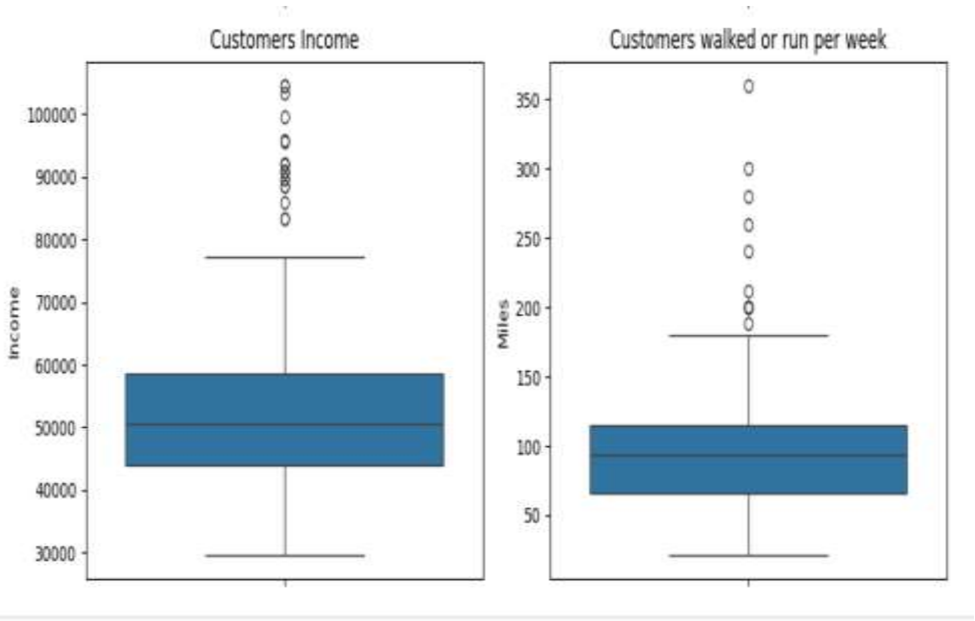


```
plt.figure(figsize=(15,8))

# Box Plot for Age
plt.subplot(2,3,1)
sns.boxplot(df['Age'])
plt.title("Customers Age")
```



```
# Box Plot for Usage
plt.subplot(2,3,2)
sns.boxplot(df['Usage'])
plt.title("Customers Usage per week")

#Box Plot for Fitness
plt.subplot(2,3,3)
sns.boxplot(df['Fitness'])
plt.title("Customers Fitness rating 1 to 5")

#Box Plot for Income
plt.subplot(2,3,4)
sns.boxplot(df['Income'])
plt.title("Customers Income")

#Box Plot for Miles
plt.subplot(2,3,5)
sns.boxplot(df['Miles'])
plt.title("Customers walked or run per week")

plt.tight_layout()
plt.show()
```



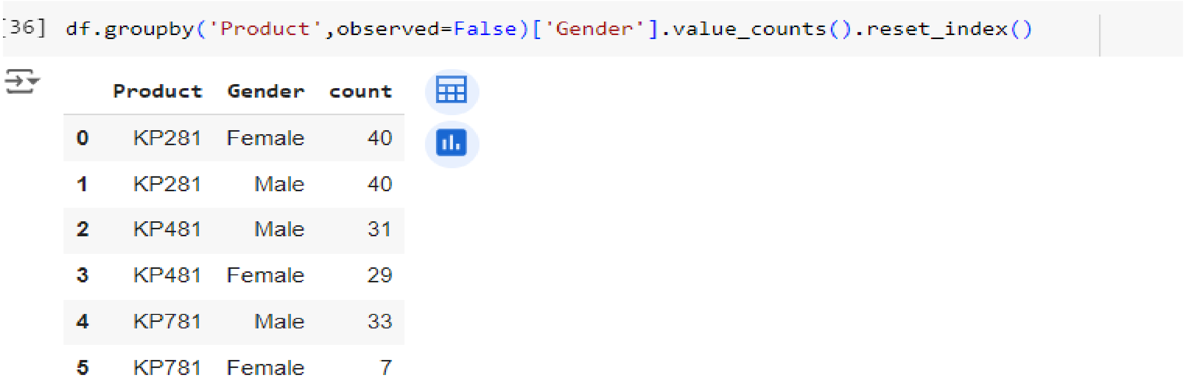


Insights: By observing the above data we can say that there are more number of outliers in the income and miles column for this mean is not a good approach for performing a analysis, median will give some approximate results due to high number of outliers. Age column and usage column contains few outliers which are manageable no outliers are present in the fitness column because rating scale has 1 to 5 values.

Recommendations: id when we take age boxplot median age is 26 which more number customers who are in that age preferring the Aerofit equipment. Median in income column is 50000 customers of Aerofit belong to major this categories so they should focus more on this area. Production of gym equipment and changes in gym equipment needs to be done according to the these analysis which can results in increasing the sales of the equipment.

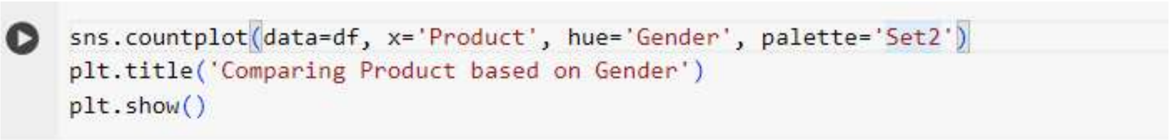
Bivariate Analysis :

Gender v/s Product

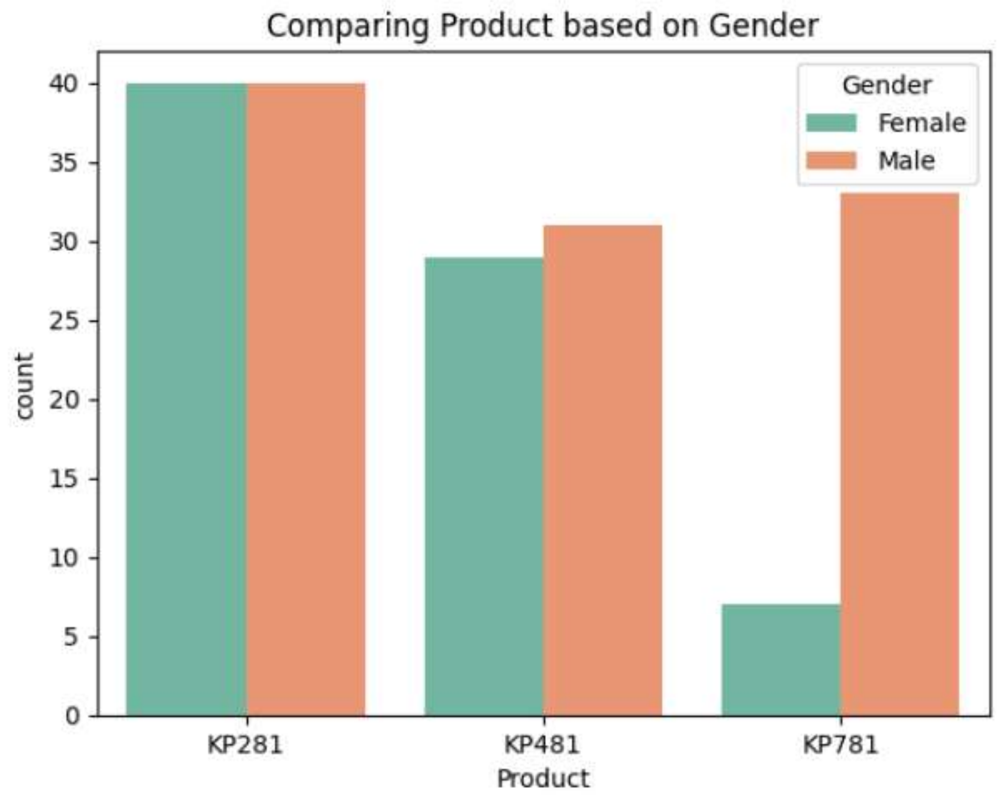


```
36] df.groupby('Product',observed=False)['Gender'].value_counts().reset_index()
```

	Product	Gender	count
0	KP281	Female	40
1	KP281	Male	40
2	KP481	Male	31
3	KP481	Female	29
4	KP781	Male	33
5	KP781	Female	7



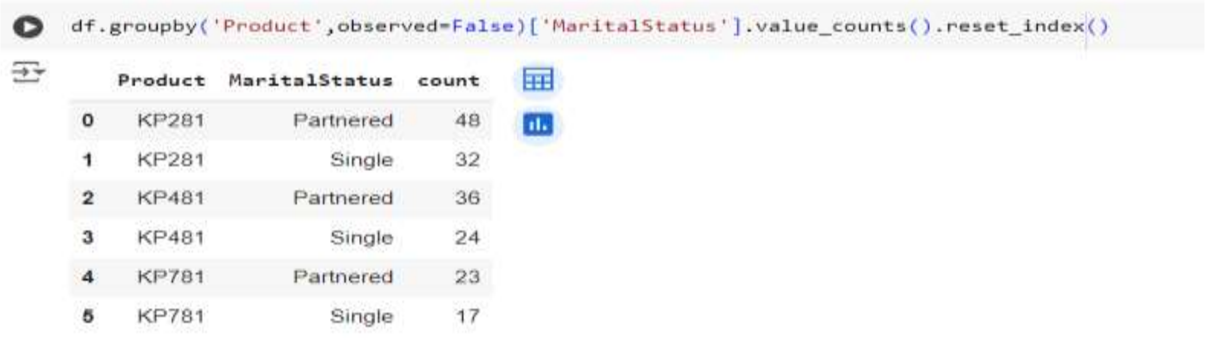
```
sns.countplot(data=df, x='Product', hue='Gender', palette='Set2')
plt.title('Comparing Product based on Gender')
plt.show()
```



Insights: The countplot indicates a balanced distribution between males and females for the KP281 and KP481 products, with both genders showing nearly equal preferences. However, there is a notable difference for the KP781 product, where a significantly larger number of males have KP781.

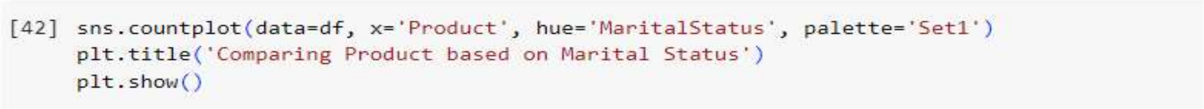
Recommendations: Aerofit gym equipment KP781 treadmill is having advanced features that sell for $2,500 is more used by mail customers because professional gym trainers and other customers who are more serious on their body fitness participating in the gym events and bodybuilding events customers will use these KP781 treadmill more due to its advance features . Aerofit need to provide more discounts for this KP781 treadmill which will results in increasing the sales. When it comes to KP281 treadmill which is KP281 is an entry-level treadmill that sells for $1,500. Males and Females customers are equally shared.

MaritalStatus v/s Product

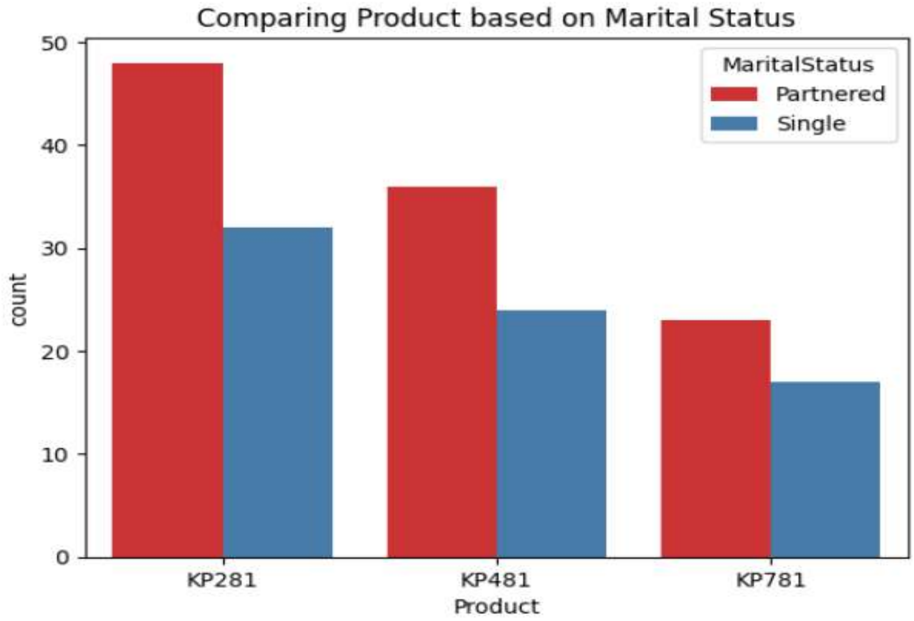


```
df.groupby('Product',observed=False)['MaritalStatus'].value_counts().reset_index()
```

	Product	MaritalStatus	count
0	KP281	Partnered	48
1	KP281	Single	32
2	KP481	Partnered	36
3	KP481	Single	24
4	KP781	Partnered	23
5	KP781	Single	17

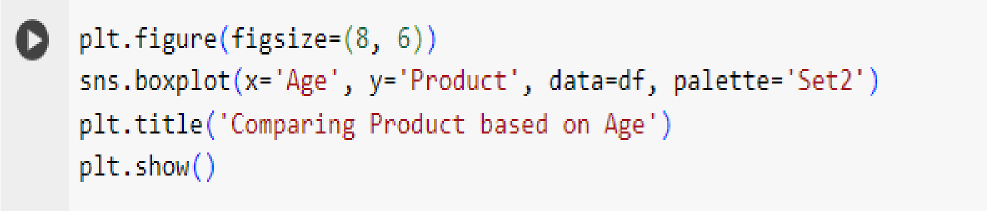


```
[42] sns.countplot(data=df, x='Product', hue='MaritalStatus', palette='Set1')
     plt.title('Comparing Product based on Marital Status')
     plt.show()
```

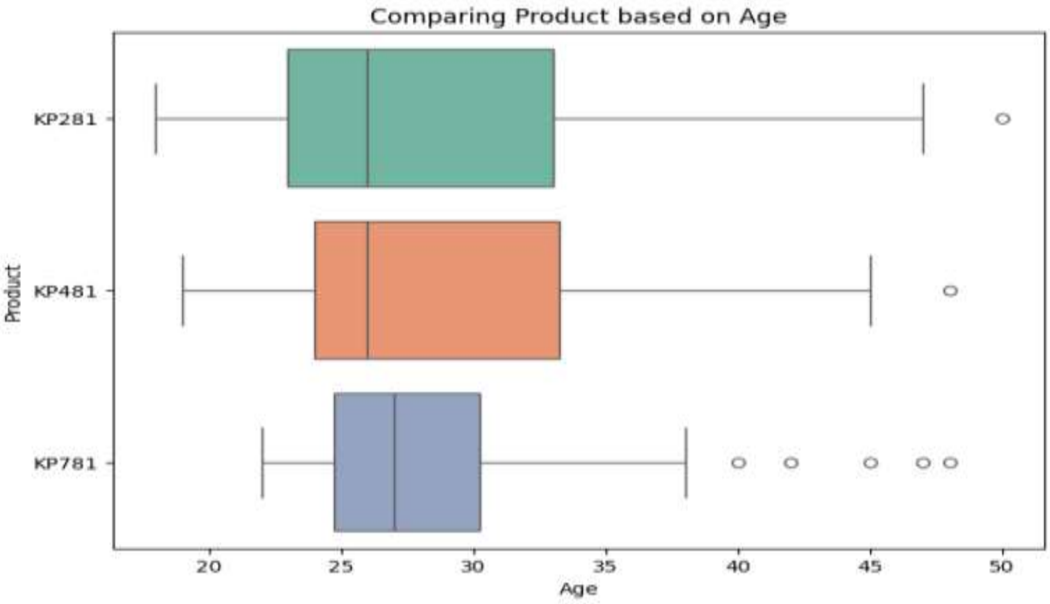


Insights: Customers in a partnered relationship are more inclined to purchase treadmill models KP281, KP481, and KP781 compared to those who are single

Age v/s Product

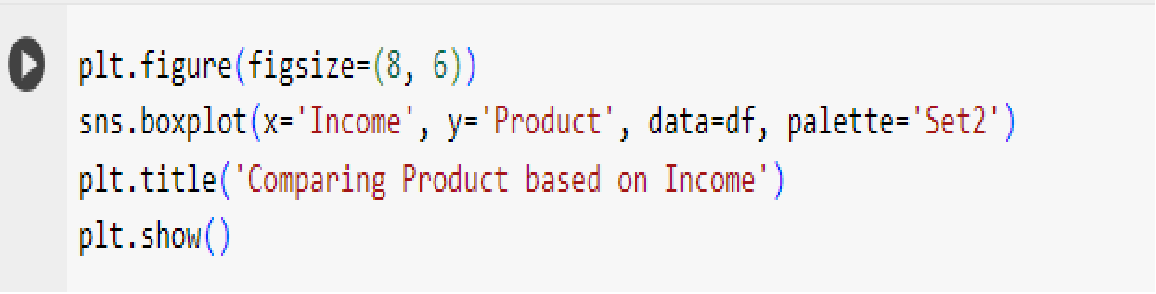


```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Age', y='Product', data=df, palette='Set2')
plt.title('Comparing Product based on Age')
plt.show()
```

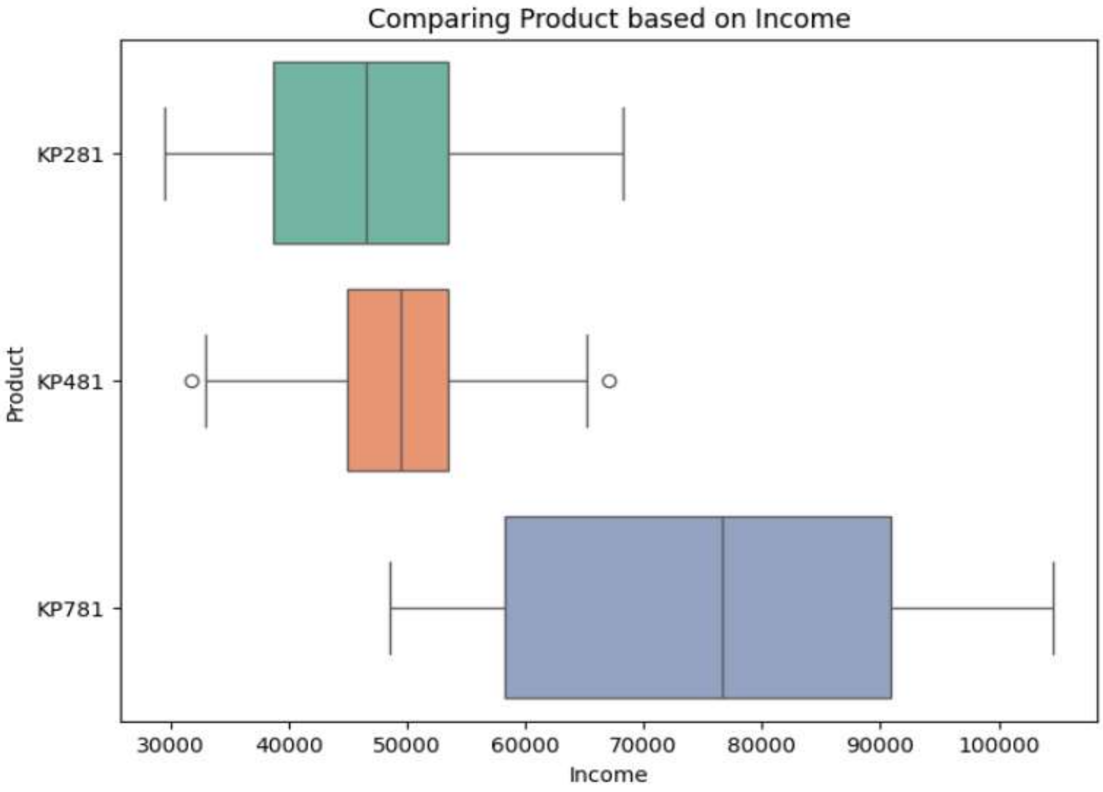


Insights: The KP281 treadmill is owned across all age groups, whereas the KP781 treadmill is predominantly favored by young and middle-aged individuals.

Income v/s Product

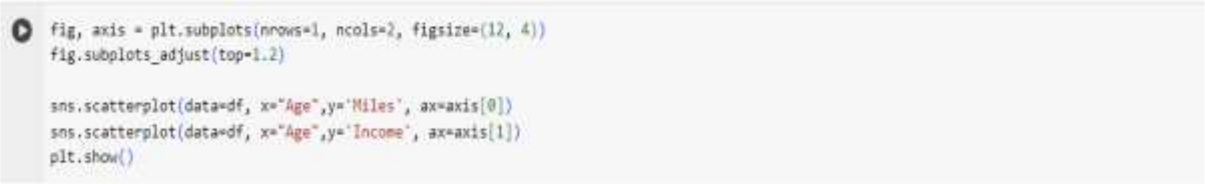


```
plt.figure(figsize=(8, 6))
sns.boxplot(x='Income', y='Product', data=df, palette='Set2')
plt.title('Comparing Product based on Income')
plt.show()
```

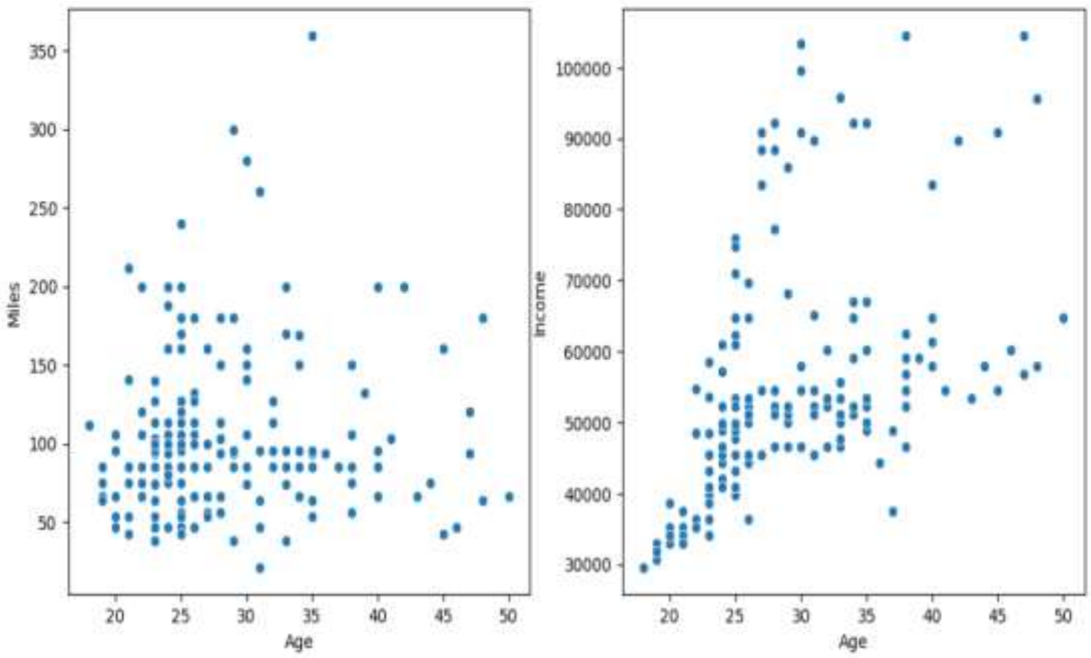


Insights: The ownership of KP781 is primarily associated with individuals earning an income higher than 50,000, whereas KP281 and KP481 are predominantly owned by those with incomes below 60,000.

Age v/s Miles and Income:



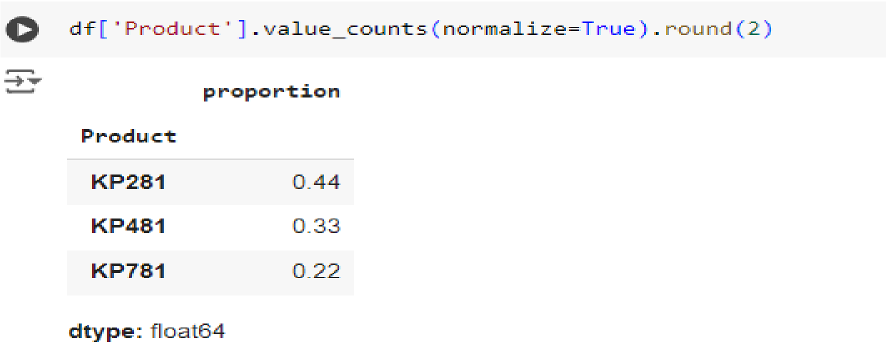
```
fig, axis = plt.subplots(nrows=1, ncols=2, figsize=(12, 4))
fig.subplots_adjust(top=1.2)

sns.scatterplot(data=df, x="Age",y='Miles', ax=axis[0])
sns.scatterplot(data=df, x="Age",y='Income', ax=axis[1])
plt.show()
```



Insights: Age exhibits a direct proportionality with income and an inverse or linear relationship with miles

Representing the Probability

Find the marginal probability

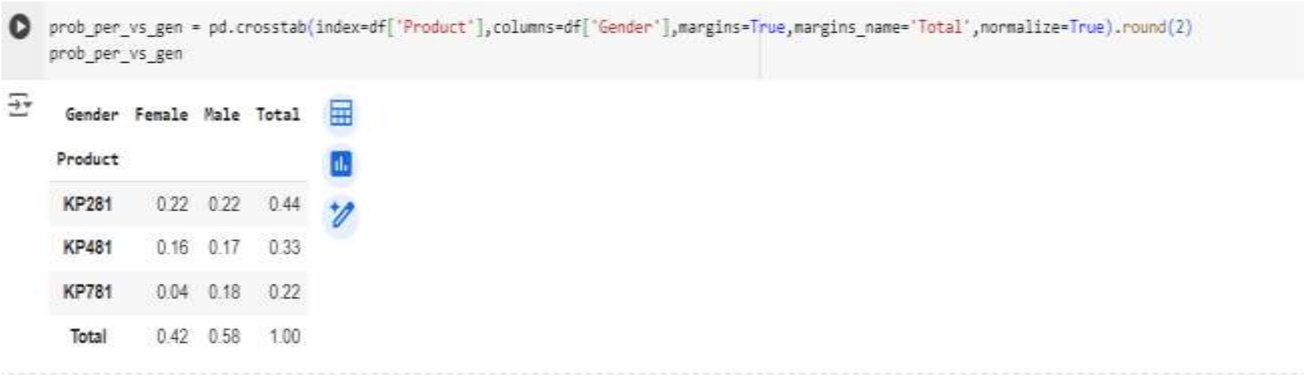


```
df['Product'].value_counts(normalize=True).round(2)
```

```
          proportion
Product
 KP281          0.44
 KP481          0.33
 KP781          0.22

dtype: float64
```

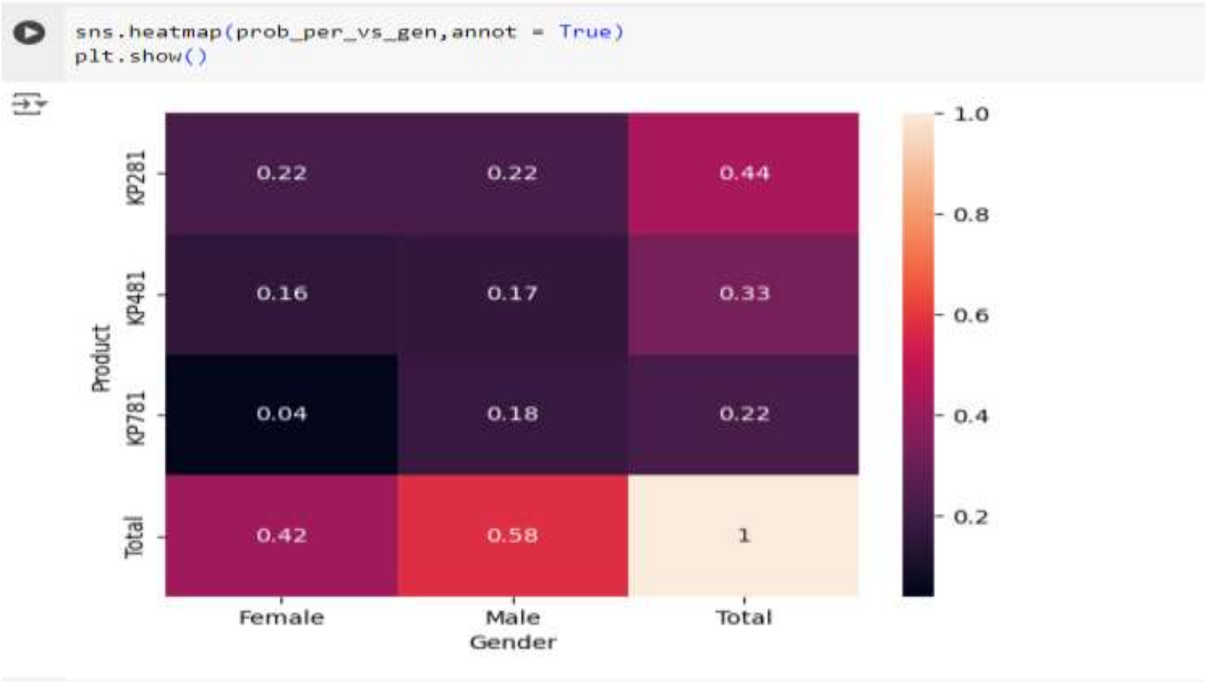
Insights: The probability of the user purchasing KP281 is at its peak, reaching 0.44, while the likelihood of acquiring KP781 is comparatively lower, hovering around 0.22.

1) Finding out the probability. Which treadmill does a person purchase according to the Gender



```
prob_per_vs_gen = pd.crosstab(index=df['Product'],columns=df['Gender'],margins=True,margins_name='Total',normalize=True).round(2)
prob_per_vs_gen
```

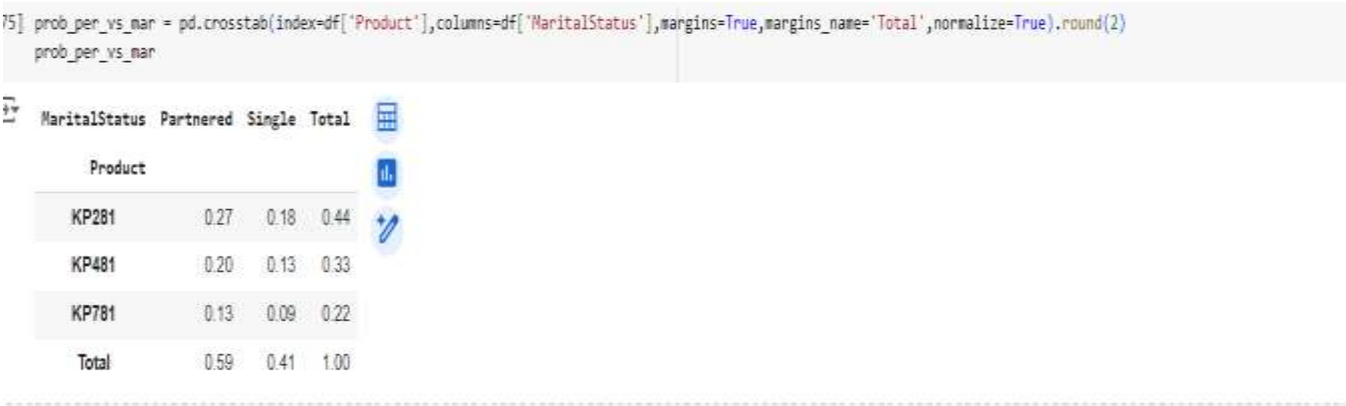
```
Gender   Female  Male  Total
Product
KP281      0.22  0.22   0.44
KP481      0.16  0.17   0.33
KP781      0.04  0.18   0.22
Total      0.42  0.58   1.00
```



```
sns.heatmap(prob_per_vs_gen,annot = True)
plt.show()
```

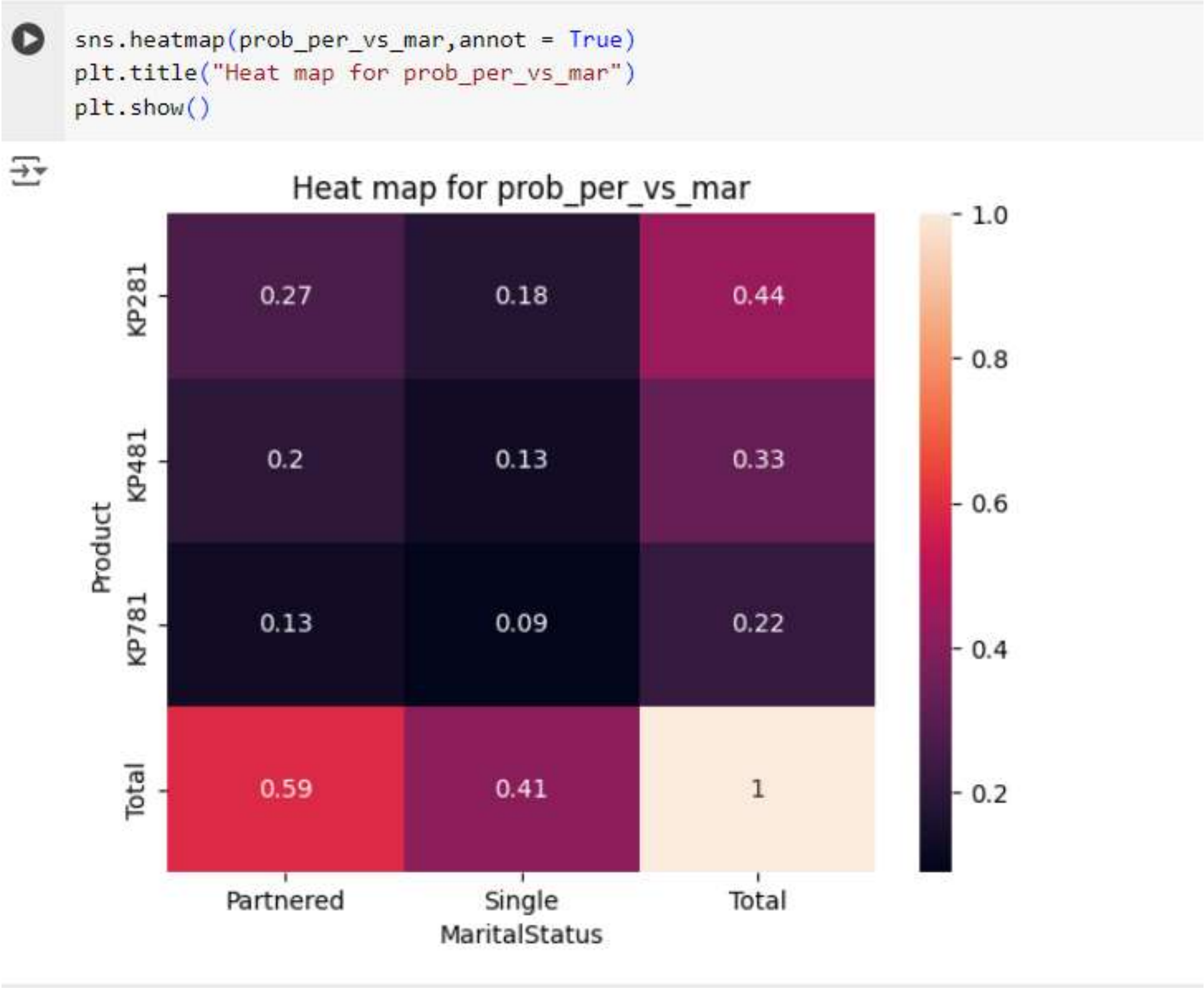
Insights: In the table above, we observe the probabilities associated with gender and the corresponding likelihood of users purchasing one of the three models. Additionally, our analysis encompasses both conditional and marginal probabilities.

2) Finding out the probability. Which treadmill does a person purchase according to the Marital Status



```
75] prob_per_vs_mar = pd.crosstab(index=df['Product'],columns=df['MaritalStatus'],margins=True,margins_name='Total',normalize=True).round(2)
    prob_per_vs_mar
```

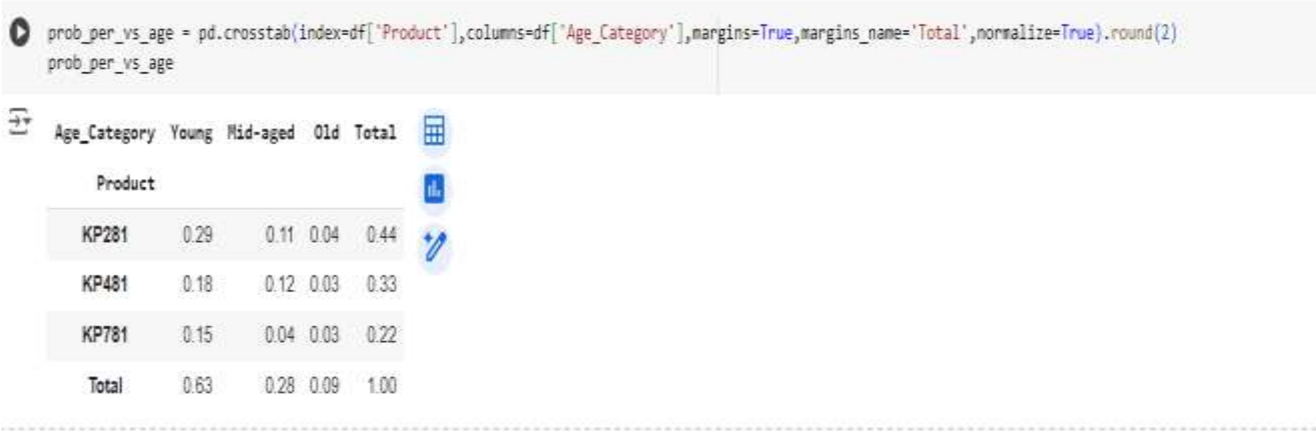
```
MaritalStatus  Partnered  Single  Total
Product
KP281               0.27    0.18   0.44
KP481               0.20    0.13   0.33
KP781               0.13    0.09   0.22
Total               0.59    0.41   1.00
```



```
sns.heatmap(prob_per_vs_mar,annot = True)
plt.title("Heat map for prob_per_vs_mar")
plt.show()
```

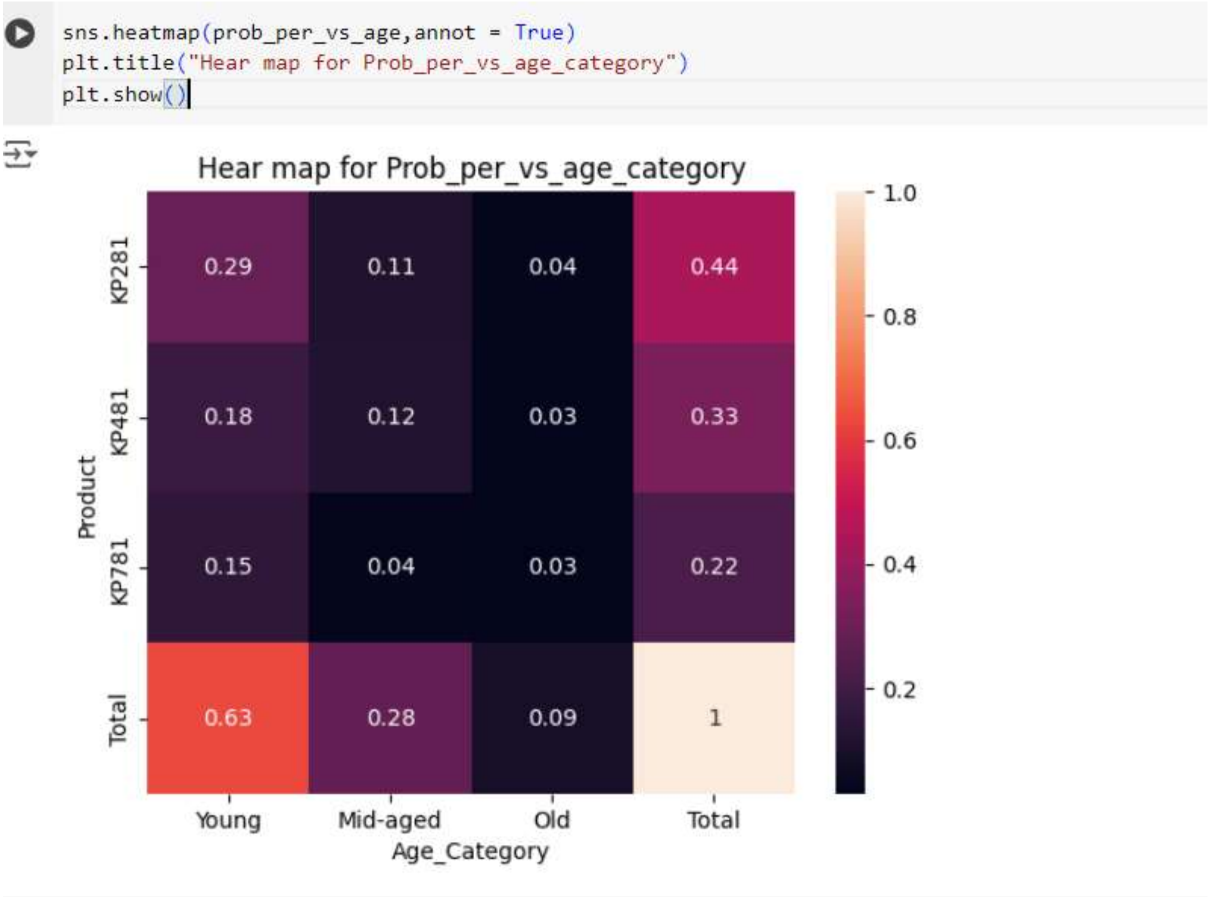
Insights: In the provided table, the probabilities are delineated based on marital status, indicating the likelihood of the user opting for a specific model among the three available choices.

3) Finding out the probability. Which treadmill a person purchase according to the Age Group



```
prob_per_vs_age = pd.crosstab(index=df['Product'],columns=df['Age_Category'],margins=True,margins_name='Total',normalize=True).round(2)
prob_per_vs_age
```

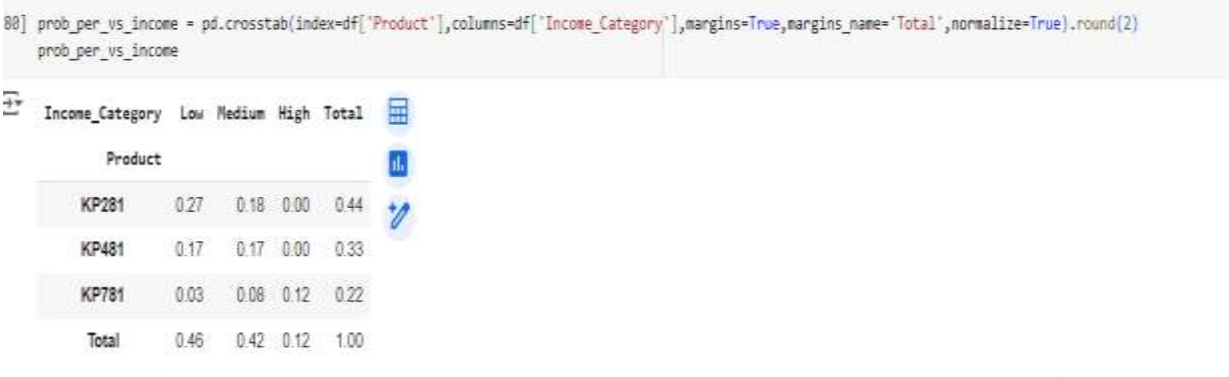
Age_Category	Young	Mid-aged	Old	Total
Product				
KP281	0.29	0.11	0.04	0.44
KP481	0.18	0.12	0.03	0.33
KP781	0.15	0.04	0.03	0.22
Total	0.63	0.28	0.09	1.00



```
sns.heatmap(prob_per_vs_age,annot = True)
plt.title("Hear map for Prob_per_vs_age_category")
plt.show()
```

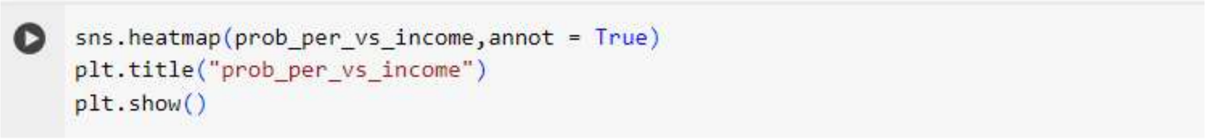
Insights: In the table above, the probabilities are delineated according to the user's Income Group, indicating the likelihood of selecting a particular model from the three available options.

4) Finding out the probability. Which treadmill does a person purchase according to the Income Group

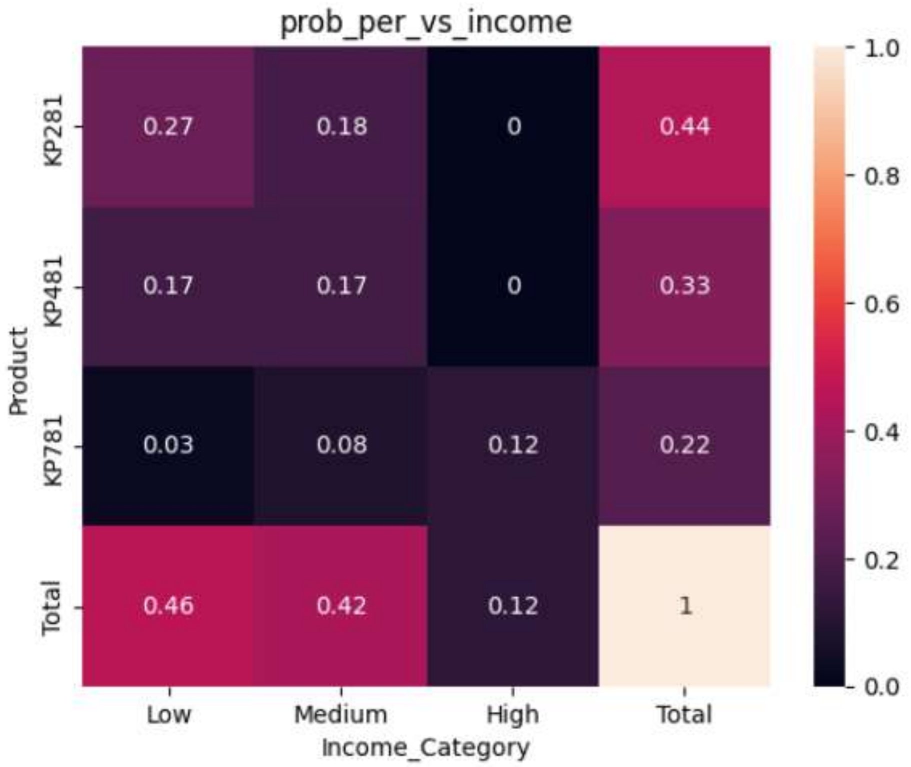


```
80] prob_per_vs_income = pd.crosstab(index=df['Product'],columns=df['Income_Category'],margins=True,margins_name='Total',normalize=True).round(2)
    prob_per_vs_income
```

Income_Category	Low	Medium	High	Total
Product				
KP281	0.27	0.18	0.00	0.44
KP481	0.17	0.17	0.00	0.33
KP781	0.03	0.08	0.12	0.22
Total	0.46	0.42	0.12	1.00

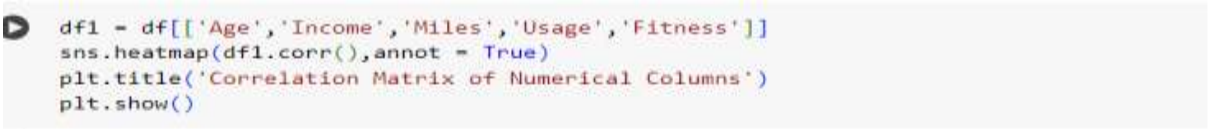


```
sns.heatmap(prob_per_vs_income,annot = True)
plt.title("prob_per_vs_income")
plt.show()
```

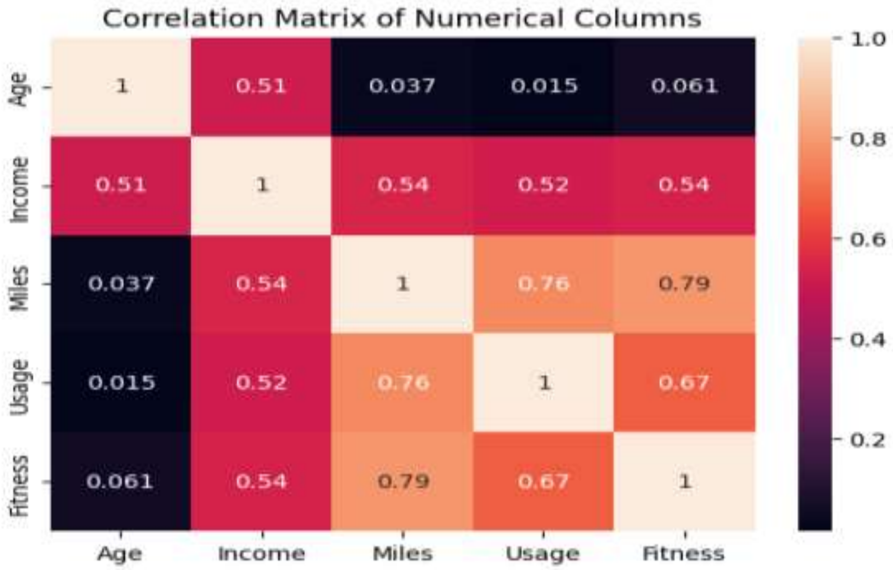


Insights: In the table above, the probabilities are delineated according to the user's Income Group, indicating the likelihood of selecting a particular model from the three available options.

Checking correlation



```
df1 = df[['Age','Income','Miles','Usage','Fitness']]
sns.heatmap(df1.corr(),annot = True)
plt.title('Correlation Matrix of Numerical Columns')
plt.show()
```



Insights: There exists a positive correlation between age and income, signifying that as age increases, income tends to rise as well. Similarly, there is a positive correlation between income and miles, illustrating that as income increases, the number of miles also tends to increase. While there is a relationship between age and miles, the behaviour is somewhat unpredictable. Notably, an increase in usage is associated with a corresponding increase in miles.

Insights:

KP281

- Gender: Both
- Marital Status: Both, preferred Partnered
- Age: 18-28
- Fitness Level: 3
- Income range: 29000 - 50000
- Usage: 3 times a week
- Education: Less than 16 years
- Miles: 70-90 Miles per week

KP481

- Gender: Both, preferred Males
- Marital Status: Both, preferred Partnered
- Age: 20-30
- Fitness Level: 3
- Income range: 30000 - 60000
- Usage: 3 times a week
- Education: Less than 16 years
- Miles: 80-120 Miles per week (Mid-runners)

KP781

- Gender: Males
- Marital Status: Both, preferred Partnered
- Age: 20-30
- Fitness Level: 4-5
- Income range: Above 60000
- Usage: 4 times a week
- Education: Above 16 years
- Miles: Above 120 Miles per week

Recommendations

Insights from the data analysis indicate that the KP281 treadmill is the most popular choice among users, followed by the KP481 and KP781 models. To leverage this information for marketing and sales strategies, the following recommendations are proposed:

1. **KP281 Emphasis:**
    - Emphasize the affordability of the KP281 treadmill.
    - Highlight key features appealing to beginners.
    - Introduce special offers to attract budget-conscious customers.
    - Engage with online fitness communities to promote its entry-level appeal.

1. **KP481 Targeting Mid-Level Runners:**
    - Focus marketing efforts on mid-level runners.
    - Highlight competitive pricing and tailored fitness benefits.
    - Utilize various channels for targeted outreach.
2. **KP781 Advanced Features:**
    - Showcase advanced features of KP781 justifying its higher price.
    - Launch targeted campaigns to raise awareness and interest.
3. **Female Customer Engagement:**
    - Create advertisements emphasizing fitness benefits for women.
    - Showcase female-friendly features of Aerofit treadmills.
    - Offer incentives and discounts for the KP781 to increase female purchases.
4. **Engaging Older Customers:**
    - Provide personalized assistance for customers aged 40-50.
    - Ensure guidance and support to maintain an active lifestyle.
5. **Affordability for Low and Middle-Income Groups:**
    - Introduce tailored discounts and incentives for low and middle-income customers.
    - By implementing these strategies, Aerofit can effectively target specific customer segments, maximize sales, and enhance overall customer satisfaction.