Business-Case: Aerofit - Descriptive Statistics & Probability

Submitted by: Archana Bharti

Business Problem

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.

Q1. Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset

In [13]:

```
# Importing necessary libraries

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

In [14]:

```
# Importing the data-set

df = pd.read_csv('D:\\Scaler\\Scaler\\Probability & Stats\\Business Case\\aerofit_treadm
```

In [15]:

```
df.head()
```

Out[15]:

	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

In [24]:

```
df.tail()
```

Out[24]:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
175	KP781	40	Male	21	Single	6	5	83416	200
176	KP781	42	Male	18	Single	5	4	89641	200
177	KP781	45	Male	16	Single	5	5	90886	160
178	KP781	47	Male	18	Partnered	4	5	104581	120
179	KP781	48	Male	18	Partnered	4	5	95508	180

In [25]:

```
df.shape
```

Out[25]:

```
(180, 9)
```

In [26]:

```
df.columns
```

Out[26]:

```
Index(['Product', 'Age', 'Gender', 'Education', 'MaritalStatus', 'Usage',
       'Fitness', 'Income', 'Miles'],
      dtype='object')
```

In [16]:

```
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
```

In [23]:

```
df.isnull().sum()
```

Out[23]:

```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
Miles            0
dtype: int64
```

Observations: No null values found

****

Q2. Detect Outliers (using boxplot, “describe” method by checking the difference between mean and median)

In [17]:

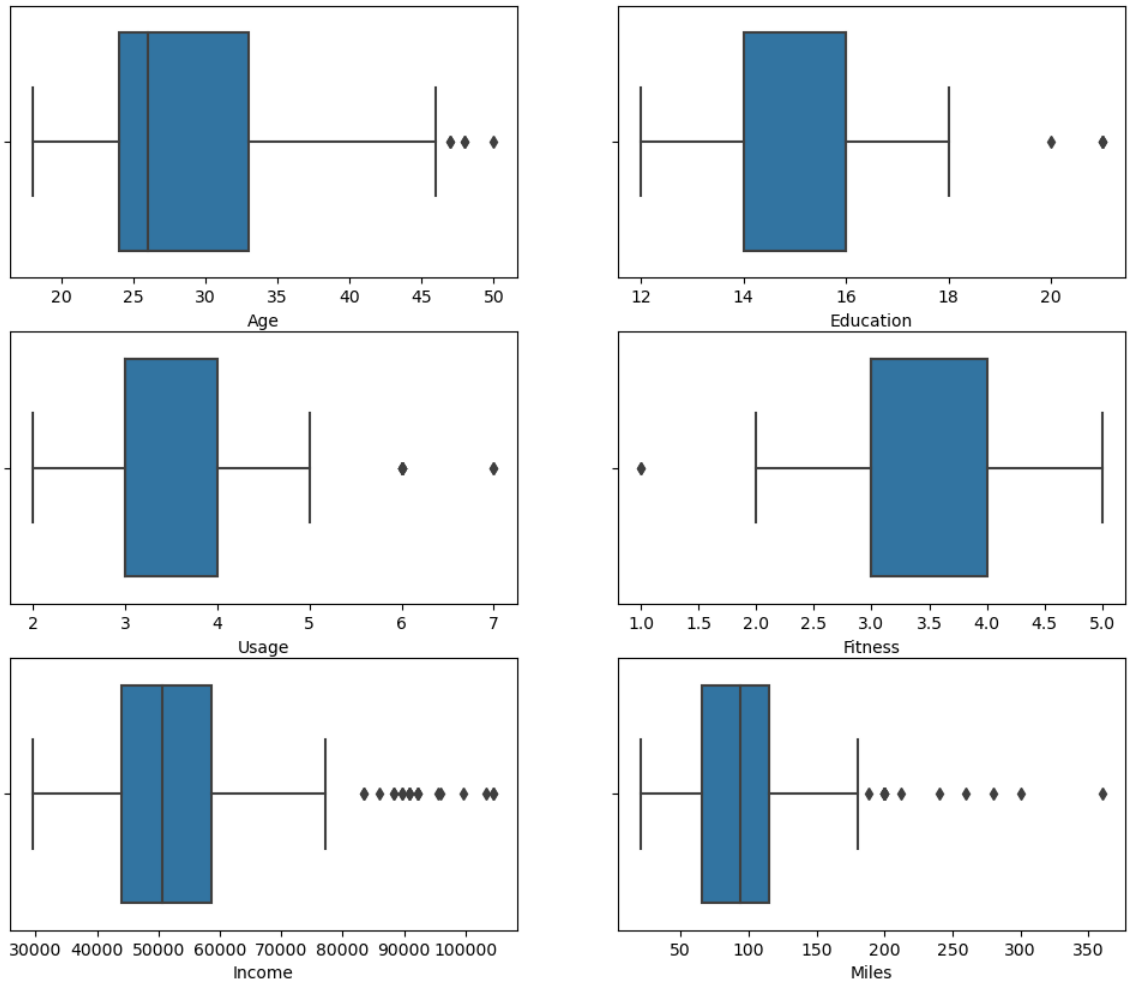
```
df.describe()
```

Out[17]:

	Age	Education	Usage	Fitness	Income	Miles
count	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000
mean	28.788889	15.572222	3.455556	3.311111	53719.577778	103.194444
std	6.943498	1.617055	1.084797	0.958869	16506.684226	51.863605
min	18.000000	12.000000	2.000000	1.000000	29562.000000	21.000000
25%	24.000000	14.000000	3.000000	3.000000	44058.750000	66.000000
50%	26.000000	16.000000	3.000000	3.000000	50596.500000	94.000000
75%	33.000000	16.000000	4.000000	4.000000	58668.000000	114.750000
max	50.000000	21.000000	7.000000	5.000000	104581.000000	360.000000

In [32]:

```
# Boxplot for each numerical column to visualize outliers
fig, axis= plt.subplots(3,2 , figsize=(12,10))
sns.boxplot(data=df,x="Age", orient='h',ax=axis[0,0])
sns.boxplot(data=df,x="Education", orient='h',ax=axis[0,1])
sns.boxplot(data=df,x="Usage", orient='h',ax=axis[1,0])
sns.boxplot(data=df,x="Fitness", orient='h',ax=axis[1,1])
sns.boxplot(data=df,x="Income", orient='h',ax=axis[2,0])
sns.boxplot(data=df,x="Miles", orient='h',ax=axis[2,1])
plt.show()
```



Observations - It is observed from the Boxplot analysis:

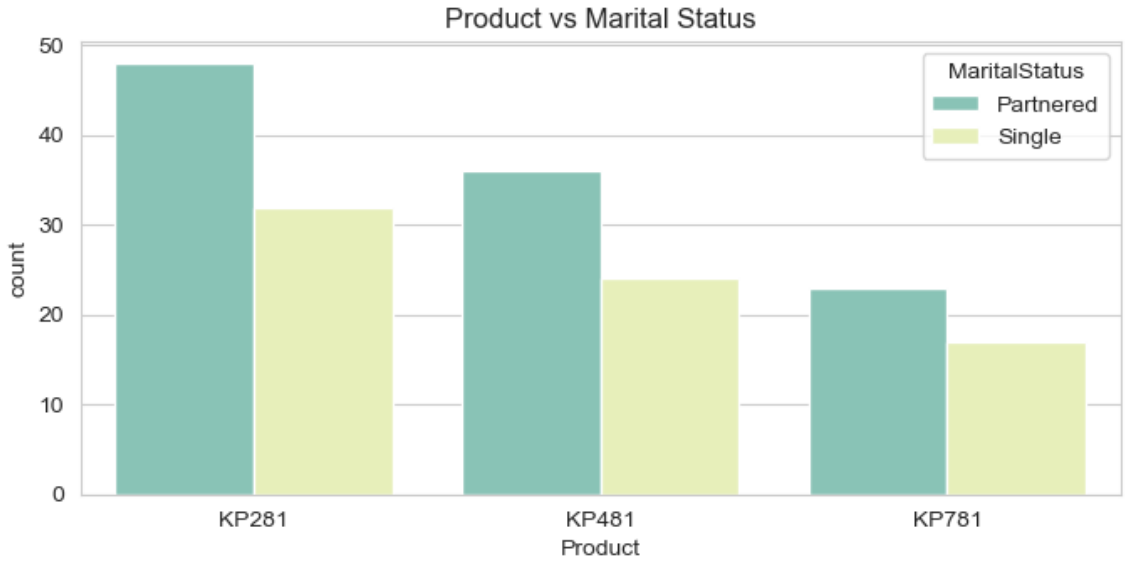
* 'Income' and 'Miles' have more outliers than other parameters

****

Q3. Check if features like marital status, age have any effect on the product purchased (using countplot, histplots, boxplots etc)

In [55]:

```
sns.set_style(style='whitegrid')
fig, axs= plt.subplots(figsize=(8,3.5))
sns.countplot(data=df,x="Product",hue="MaritalStatus",palette=["#7fcdbb","#edf8b1"])
axs.set_title("Product vs Marital Status")
plt.show()
```

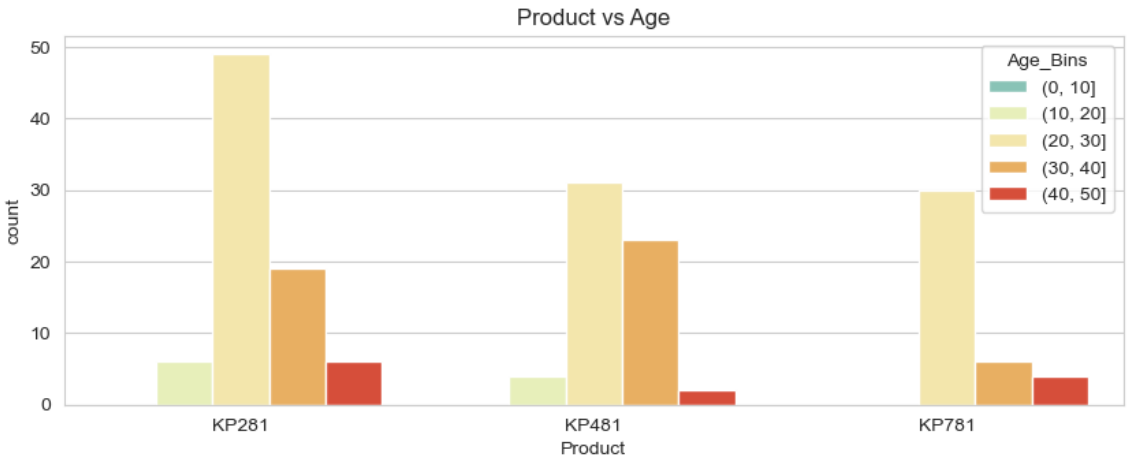


Observations:

* Customers who is "Partnered" is more likely to purchase the product across all the products

In [56]:

```
sns.set_style(style='whitegrid')

fig, axs = plt.subplots(figsize=(10, 3.5))

age_bins = [0, 10, 20, 30, 40, 50]

df['Age_Bins'] = pd.cut(df['Age'], bins=age_bins)

sns.countplot(data=df, x="Product", hue="Age_Bins", palette=["#7fcdbb", "#edf8b1", "#ffe

axs.set_title("Product vs Age")

plt.show()
```



Observations:

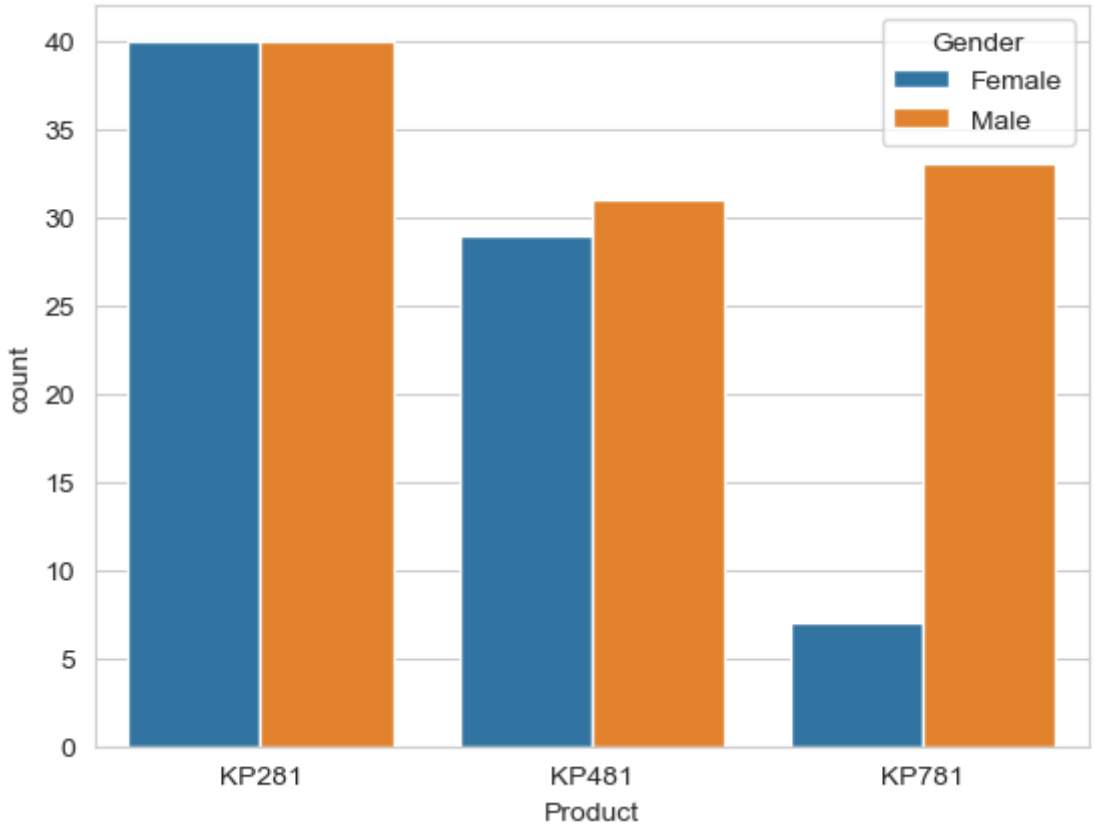
* Age-group of 20-30 is more likely to buy the product across category

In [53]:

```
sns.countplot(x='Product', hue='Gender', data=df)
```

Out[53]:

```
<Axes: xlabel='Product', ylabel='count'>
```



Observations:

* Equal number of Males and Females purchased KP281 product

* Slightly more Males were seen buying KP481 than females

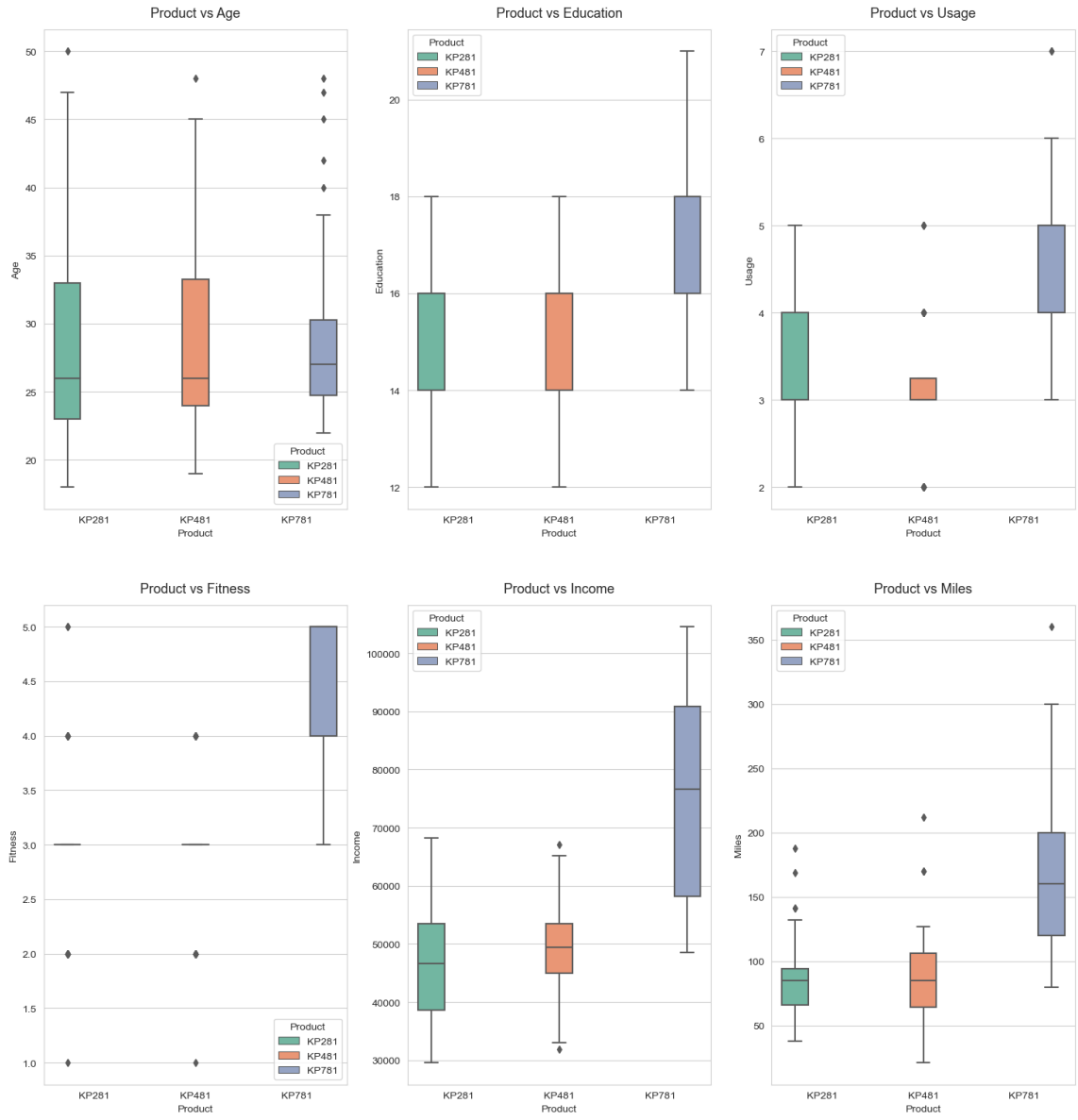
* Most of the male customers purchased KP781 product

Univariate Analysis

Check if the given features have any effect on the product purchased

In [58]:

```
var= ['Age','Education','Usage','Fitness','Income','Miles']
sns.set_style("whitegrid")
fig,axs=plt.subplots(2,3,figsize=(18,12))
fig.subplots_adjust(top=1.3)
count=0
for i in range(2):
    for j in range(3):
        sns.boxplot(data=df,x='Product',y=var[count],ax=axs[i,j],hue='Product',palette="
        axs[i,j].set_title(f"Product vs {var[count]}",pad=12,fontsize=13)
        count+=1
```



Observations:

1. Product vs Age:

*** Customers purchasing products KP281 and KP481 are having same age median value**

*** Customers whose age lies between 25-30 are more likely to buy KP781 product**

2. Product vs Education:

*** Customers whose education is greater than 16, have more chances to purchase the KP781 product**

*** While the customers with education less than 16 have equal chances of purchasing KP281 or KP481**

Product vs Usage:

*** Customers who are planning to use the treadmill greater than 4 times a week are more likely to purchase KP781**

*** While other customers are likely to buy KP281 or KP481**

Product vs Fitness:

*** Customer with with more than 3 rating for "fitness" have the higher chance to purchase KP781**

Product vs Income:

*** The customers with the income > = 60000, have the higer chances to purchase KP781**

Product vs Miles:

****

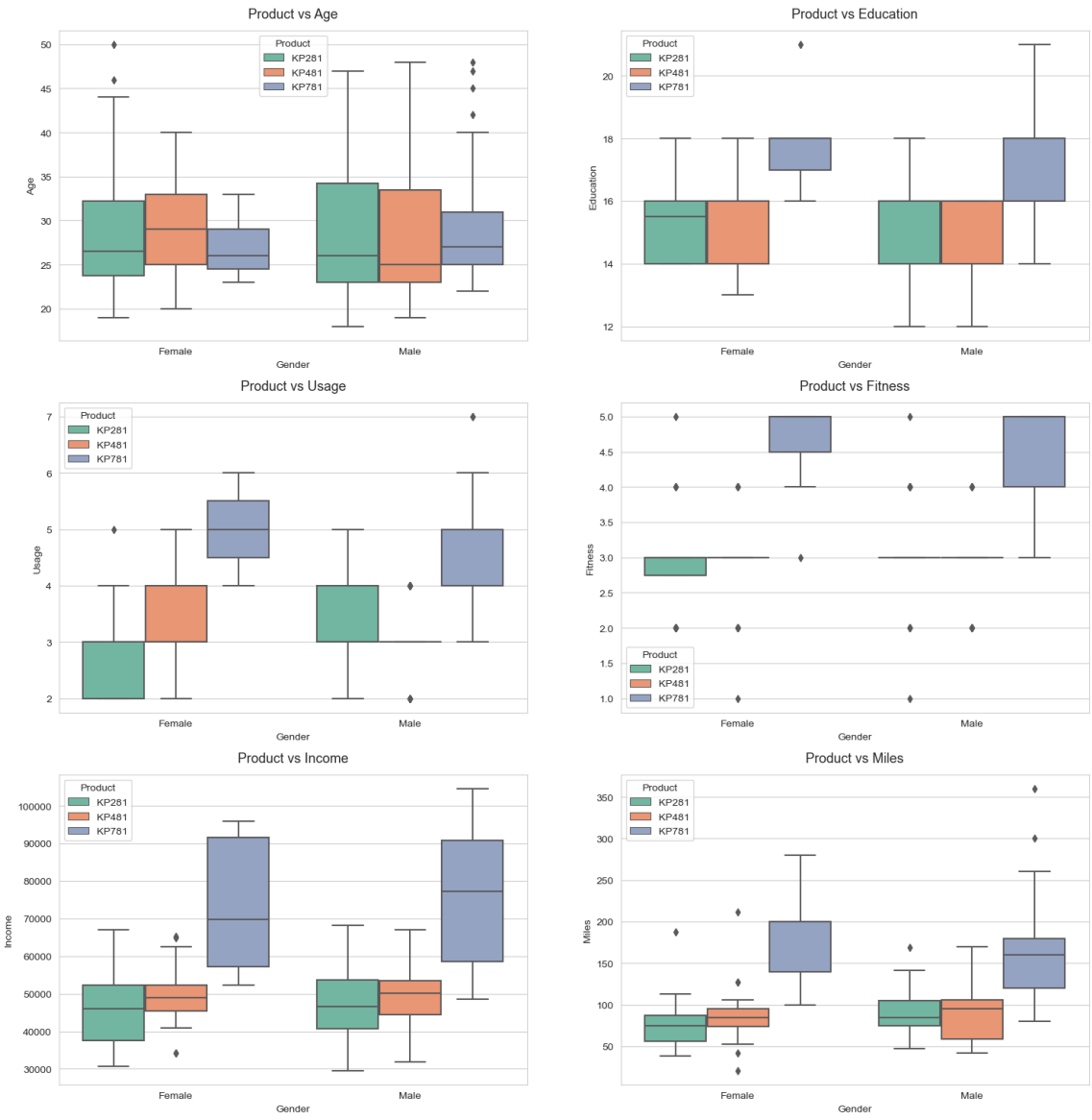
Q7. Customer Profiling - Categorization of users.

Multivariate Analysis

Checking if Gender-wise the given feature have any effect on the product purchased

In [60]:

```
var= ['Age','Education','Usage','Fitness','Income','Miles']
sns.set_style("whitegrid")
fig,axs=plt.subplots(3,2,figsize=(18,12))
fig.subplots_adjust(top=1.3)
count=0
for i in range(3):
    for j in range(2):
        sns.boxplot(data=df,x='Gender',y=var[count],hue='Product',ax=axs[i,j],palette="S
        axs[i,j].set_title(f"Product vs {var[count]}",pad=12,fontsize=13)
        count+=1
```



Observations:

*** In both male and female category, customers whose education is greater than 16, prefer to buy KP781**

*** In both Gender, customers who are planning to use treadmills more than four times, prefer to buy KP781**

*** Females who are planning to use treadmills 3-4 times a week, are more likely to buy KP781**

*** In both Gender, customer whose income is more than 55000, are more likely to buy KP781**

****

Q4. Representing the marginal probability like - what percent of customers have purchased KP281, KP481, or KP781 in a table (can use pandas.crosstab here)

Q8. Probability- marginal, conditional probability

Marginal Probability

In [61]:

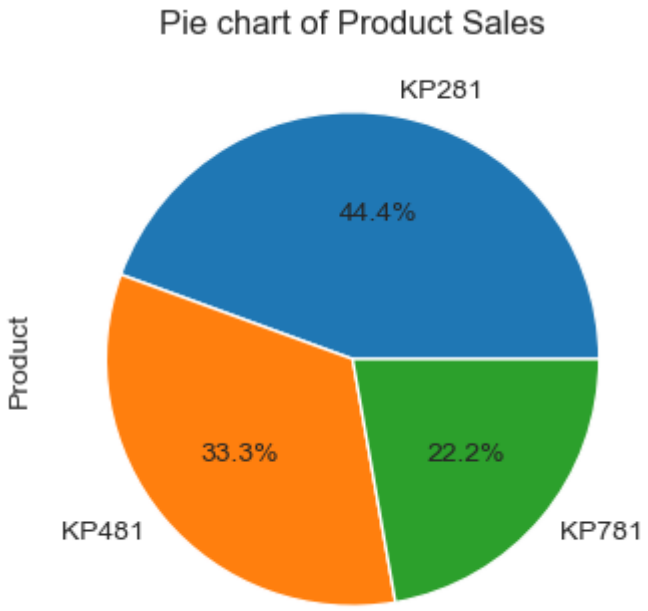
```
pd.concat(
    [
        df.Product.value_counts(),
        df.Product.value_counts(normalize=True)
    ],
    keys=['counts', 'Marginal_Prob'], axis=1, )
```

Out[61]:

	counts	Marginal_Prob
KP281	80	0.444444
KP481	60	0.333333
KP781	40	0.222222

In [65]:

```
plt.figure(figsize=(8,4))
df['Product'].value_counts().plot.pie(autopct='%1.1f%%',figsize=(4,4))
plt.title("Pie chart of Product Sales")
plt.show()
```



Observations:

* KP281 is the most sold product category (with 44% of share) followed by KP481 (with 33% share)

* The least sold product is KP781

Conditional Probability

Probability of each product given gender

In [66]:

```
def p_prod_given_gender(gender, print_marginal=False):
    if gender!= "Female" and gender!= "Male":
        return "Invalid gender value."
    df1= pd.crosstab(index=df['Gender'],columns=[df['Product']])
    p_781= df1['KP781'][gender] / df1.loc[gender].sum()
    p_481= df1['KP481'][gender] / df1.loc[gender].sum()
    p_281= df1['KP281'][gender] / df1.loc[gender].sum()

    if print_marginal:
        print(f"P(Male): {df1.loc['Male'].sum()/len(df):.2f}")
        print(f"P(Female): {df1.loc['Female'].sum()/len(df):.2f}")

    print(f"P(KP781/{gender}):{p_781:.2f}")
    print(f"P(KP481/{gender}):{p_481:.2f}")
    print(f"P(KP281/{gender}):{p_281:.2f}\n")

p_prod_given_gender('Male',True)
p_prod_given_gender('Female')
```

```
P(Male): 0.58
P(Female): 0.42
P(KP781/Male):0.32
P(KP481/Male):0.30
P(KP281/Male):0.38

P(KP781/Female):0.09
P(KP481/Female):0.38
P(KP281/Female):0.53
```

Observations:

*** More male customers are likely to purchase KP781**

*** Both male and female customers are likely to purchase KP281**

****

Q6. With all the above steps you can answer questions like: What is the probability of a male customer buying a KP781 treadmill?

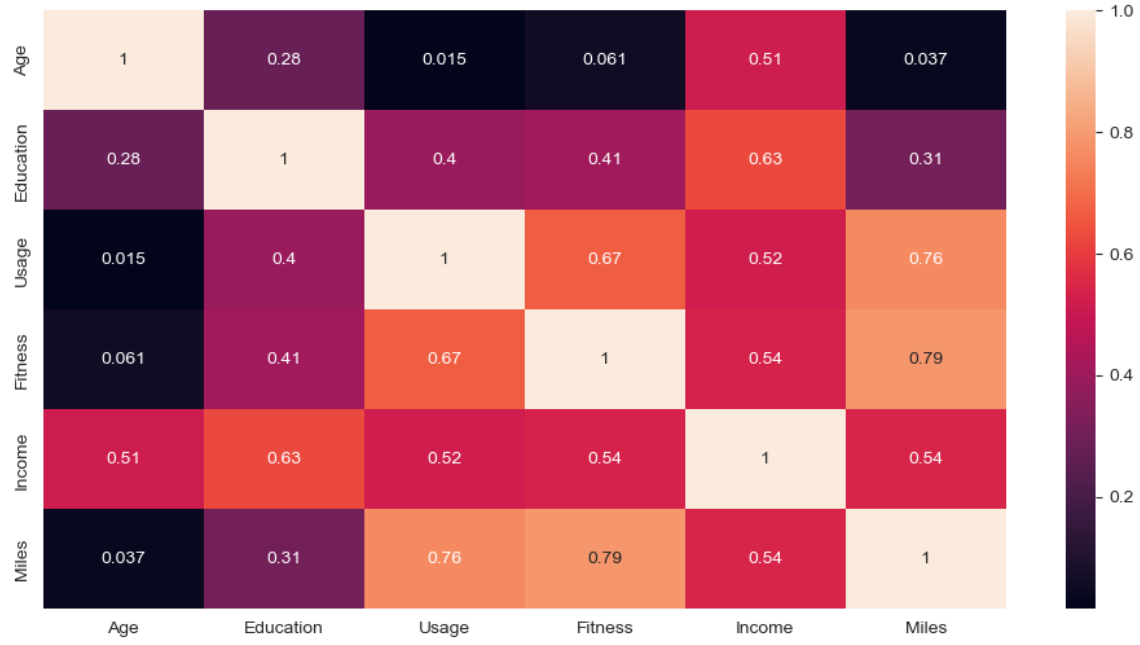
Ans: The probability of male customer are buying a KP781 treadmill is 0.32 (32%)

****

Q5. Check correlation among different factors using heat maps or pair plots.

In [69]:

```
plt.figure(figsize=(12, 6))
sns.heatmap(df.corr(numeric_only=True), annot=True)
plt.show()
```

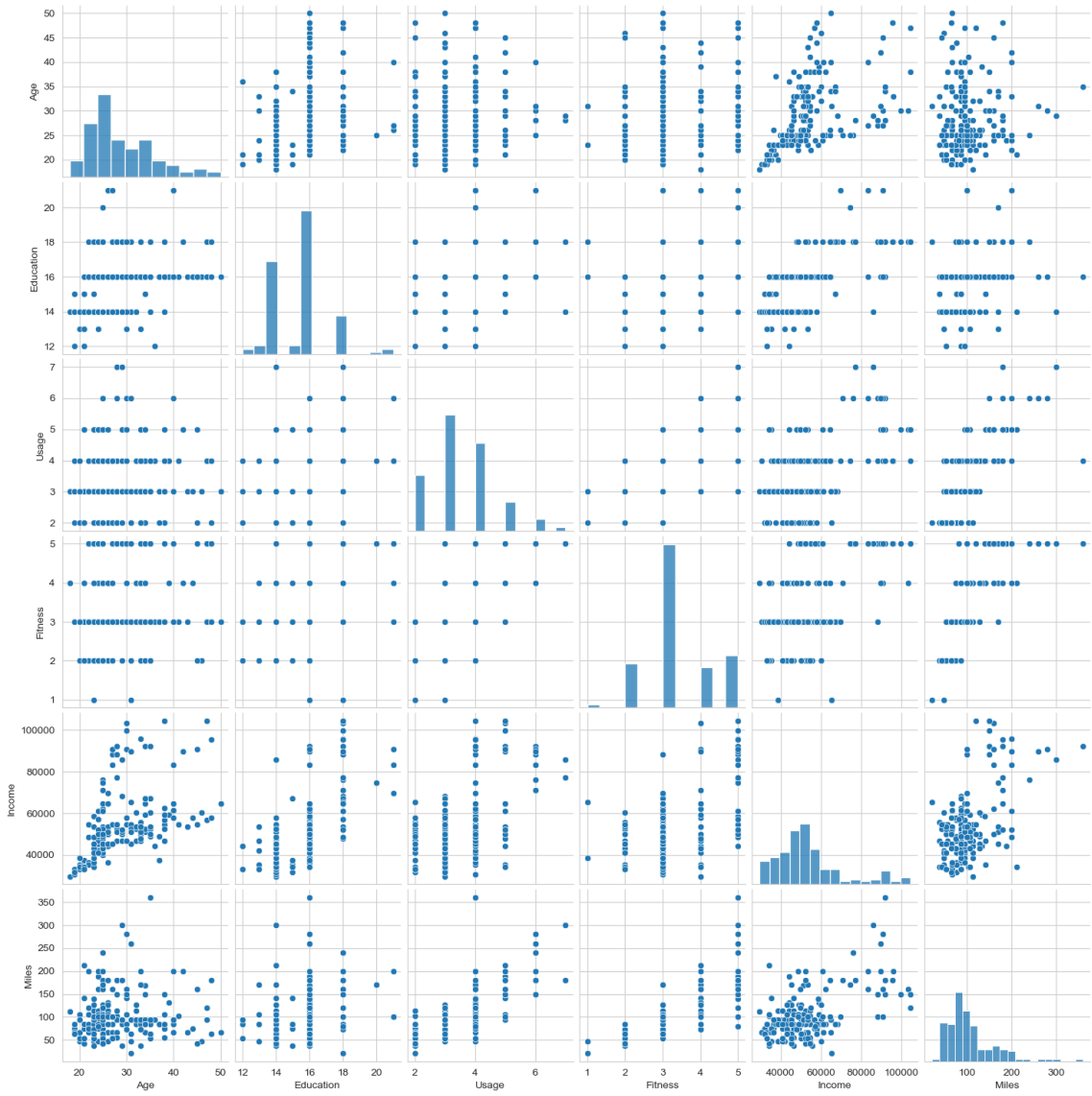


In [70]:

```
sns.pairplot(df)
```

Out[70]:

```
<seaborn.axisgrid.PairGrid at 0x17197b11d20>
```



In [72]:

```
df.corr(numeric_only=True)
```

Out[72]:

	Age	Education	Usage	Fitness	Income	Miles
Age	1.000000	0.280496	0.015064	0.061105	0.513414	0.036618
Education	0.280496	1.000000	0.395155	0.410581	0.625827	0.307284
Usage	0.015064	0.395155	1.000000	0.668606	0.519537	0.759130
Fitness	0.061105	0.410581	0.668606	1.000000	0.535005	0.785702
Income	0.513414	0.625827	0.519537	0.535005	1.000000	0.543473
Miles	0.036618	0.307284	0.759130	0.785702	0.543473	1.000000

Observations:

Each cell in the table represents the correlation coefficient between the corresponding two variables. The correlation coefficient ranges from -1 to 1 and provides insights into the strength and direction of the relationship between the variables.

Here are some insights based on the correlation coefficients:

Age and Education: The correlation coefficient between Age and Education is 0.280496, which indicates a positive correlation. This means that, in general, as age increases, education tends to increase as well. However, the correlation is not very strong, suggesting that age is not a major determinant of education level.

Age and Usage: The correlation coefficient between Age and Usage is very low (0.015064), indicating a weak positive correlation. There is little to no relationship between age and how often the treadmill is used.

Age and Fitness: The correlation coefficient between Age and Fitness is also weak (0.061105), suggesting a minimal positive relationship between age and fitness levels. Age does not play a significant role in determining fitness levels.

Age and Income: The correlation coefficient between Age and Income is 0.513414, which indicates a moderate positive correlation. This suggests that, on average, as age increases, income tends to increase as well. However, other factors may also influence income, and age alone is not a strong predictor of income.

Age and Miles: The correlation coefficient between Age and Miles is very low (0.036618), indicating a weak positive correlation. Age has little impact on the number of miles users run on the treadmill.

Education and Usage: The correlation coefficient between Education and Usage is 0.395155, indicating a moderate positive correlation. Higher education levels are associated with more frequent treadmill usage.

Education and Fitness: The correlation coefficient between Education and Fitness is 0.410581, indicating a moderate positive correlation. Higher education levels tend to be associated with higher fitness levels.

Education and Income: The correlation coefficient between Education and Income is 0.625827, which indicates a relatively strong positive correlation. People with higher education levels tend to have higher incomes.

Education and Miles: The correlation coefficient between Education and Miles is 0.307284, indicating a weak positive correlation. Education levels have a minor impact on the number of miles users run on the treadmill.

Usage and Fitness: The correlation coefficient between Usage and Fitness is 0.668606, indicating a strong positive correlation. Users who use the treadmill more frequently tend to have higher fitness levels.

Usage and Income: The correlation coefficient between Usage and Income is 0.519537, suggesting a moderate positive correlation. People who use the treadmill more frequently tend to have higher incomes, on average.

Usage and Miles: The correlation coefficient between Usage and Miles is 0.759130, indicating a strong positive correlation. Users who use the treadmill more frequently tend to run more miles.

Fitness and Income: The correlation coefficient between Fitness and Income is 0.535005, indicating a moderate positive correlation. Higher fitness levels are associated with higher incomes, on average.

Fitness and Miles: The correlation coefficient between Fitness and Miles is 0.785702, indicating a strong positive correlation. Users with higher fitness levels tend to run more miles on the treadmill.

Income and Miles: The correlation coefficient between Income and Miles is 0.543473, indicating a moderate positive correlation. People with higher incomes tend to run more miles on the treadmill.

****

Q9. Some recommendations and actionable insights, based on the inferences.

INSIGHTS

1.Customers who is "Partnered" is more likely to purchase the product across all the products

2.Equal number of Males and Females purchased KP281 product

3.Most of the male customers purchased KP781 product

4.Customers whose age lies between 25-30 are more likely to buy KP781 product

5.Customers whose education is greater than 16, have more chances to purchase the KP781 product

6.Customers who are planning to use the treadmill greater than 4 times a week are more likely to purchase KP781

7.Customer with with more than 3 rating for "fitness" have the higher chance to purchase KP781

8.The customers with the income > = 60000, have the higer chances to purchase KP781

9.If the customer expects to walk/run more than 120 miles per week, it is more likely that the customers will buy KP781

10.Females who are planning to use treadmills 3-4 times a week, are more likely to buy KP781

11.In both Gender, customer whose income is more than 55000, are more likely to buy KP781

12.KP281 is the most sold product category (with 44% of share) followed by KP481 (with 33% share)

Recommendations:

1.KP781 should be marketed as Premium Model and can be targeted to high-income groups

2.Marketing initiative can be more focused to the age-group of 25-30

3.As the KP781 is a premium product category, it is ideally suited for fitness enthusiasts who have a higher average weekly mileage

4.Aerolift should conduct market research to determine if it can attract customers with income less than 40,000 to expand its customer base

END OF PROJECT