business-case-aerofit

October 26, 2024

#Business Case: Aerofit - Descriptive Statistics & Probability

```python
[34]: import pandas as pd
      import numpy as np
      import matplotlib.pyplot as plt
      import seaborn as sbn
```

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

```python
[35]: df=pd.read_csv('/content/drive/MyDrive/Data sets/aerofit_treadmill.csv')
```

```python
[36]: df
```

```
[36]:      Product  Age  Gender  Education MaritalStatus  Usage  Fitness  Income  \
      0      KP281   18    Male         14        Single      3        4   29562
      1      KP281   19    Male         15        Single      2        3   31836
      2      KP281   19  Female         14     Partnered      4        3   30699
      3      KP281   19    Male         12        Single      3        3   32973
      4      KP281   20    Male         13     Partnered      4        2   35247
      ..       ...  ...     ...        ...           ...    ...      ...     ...
      175    KP781   40    Male         21        Single      6        5   83416
      176    KP781   42    Male         18        Single      5        4   89641
      177    KP781   45    Male         16        Single      5        5   90886
      178    KP781   47    Male         18     Partnered      4        5  104581
      179    KP781   48    Male         18     Partnered      4        5   95508

           Miles
      0      112
      1       75
      2       66
      3       85
      4       47
      ..     ...
      175    200
      176    200
      177    160
```

```
178    120
179    180

[180 rows x 9 columns]
```

[37]: `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
```

2 2.Detect Outliers (using boxplot, "describe" method by checking the difference between mean and median)

[38]: `df.isna().sum()`

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

[39]: `df.describe()`

```
[39]:              Age    Education        Usage     Fitness         Income  \
      count  180.000000  180.000000  180.000000  180.000000     180.000000
      mean    28.788889   15.572222    3.455556    3.311111   53719.577778
      std      6.943498    1.617055    1.084797    0.958869   16506.684226
```

```
min       18.000000   12.000000   2.000000   1.000000    29562.000000
25%       24.000000   14.000000   3.000000   3.000000    44058.750000
50%       26.000000   16.000000   3.000000   3.000000    50596.500000
75%       33.000000   16.000000   4.000000   4.000000    58668.000000
max       50.000000   21.000000   7.000000   5.000000   104581.000000

              Miles
count    180.000000
mean     103.194444
std       51.863605
min       21.000000
25%       66.000000
50%       94.000000
75%      114.750000
max      360.000000
```

[40]:
```python
# calculate diff between mean & median
df.describe().loc['mean']-df.describe().loc['50%']
```

[40]:
```
Age                2.788889
Education         -0.427778
Usage              0.455556
Fitness            0.311111
Income          3123.077778
Miles              9.194444
dtype: float64
```

KeyInsights: diffrence between Mean & Median: 1.Age 2.78 years 2.Income $ 3123 3.Miles 9.19

[41]:
```python
df.head()
```

[41]:

	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

[42]:
```python
df.shape
```

[42]:
```
(180, 9)
```

[43]:
```python
pd.crosstab(index=df['Gender'],columns='count')
```

[43]:
```
col_0    count
Gender
Female      76
```

```
Male        104
```

Key insights:

1. Total male customers are 104 2.Total Female customers are 76

Actionable Insights:

1.Male customers are more inclined towards fitness 2.There is good scope of increasing female customers. 3.create marketing strategies to attract females as there population is almost 50%. 4.Create fitness awarness progrmames for females.

```
[44]: pd.crosstab(index=df['Age'],columns='count')
```

```
[44]: col_0  count
      Age
      18         1
      19         4
      20         5
      21         7
      22         7
      23        18
      24        12
      25        25
      26        12
      27         7
      28         9
      29         6
      30         7
      31         6
      32         4
      33         8
      34         6
      35         8
      36         1
      37         2
      38         7
      39         1
      40         5
      41         1
      42         1
      43         1
      44         1
      45         2
      46         1
      47         2
      48         2
      50         1
```

Key Insights:

1. Avg age is around 25 years 2.Most customers are in the range of age 20-40

Actionable Insights:

1.Focus on people above 25 years age , as there disposable income is high.

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

```
[45]: # create box plot of Income vs Education

sbn.boxplot(x="Education",y="Income",data=df)
sbn.set(style="whitegrid")
```
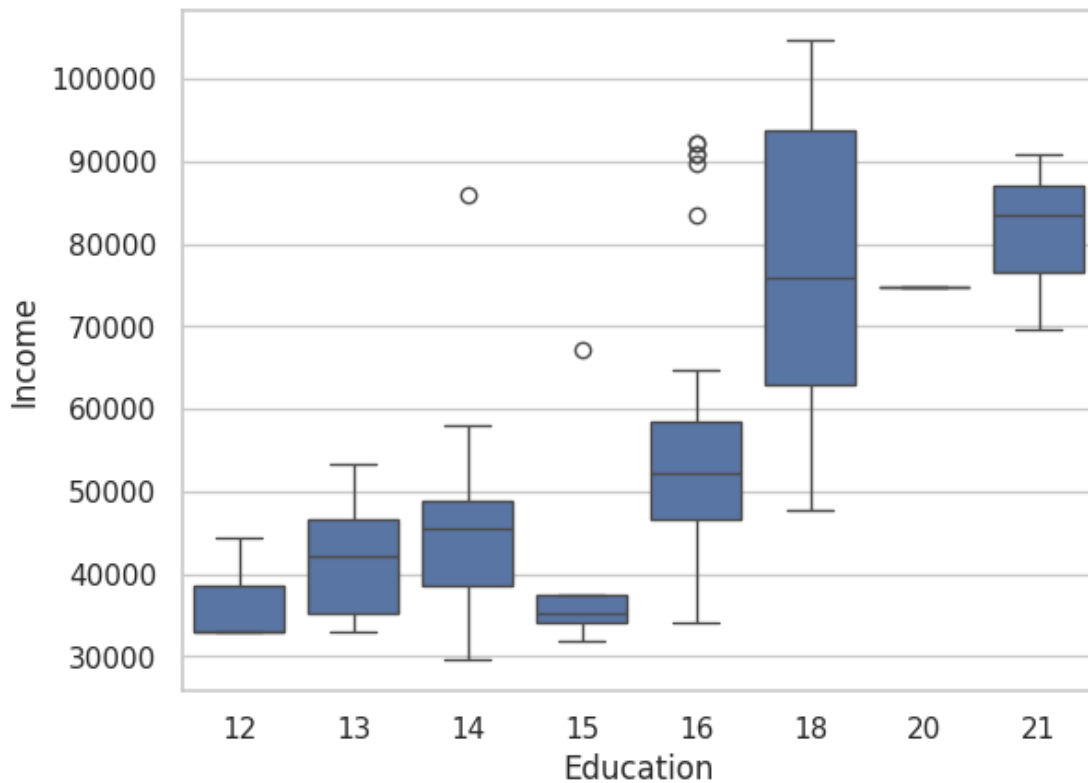

Key Insights:

1. As education level increases the Income level also increases.

Actionable Insights:

1.Higher Education leads to Higher Income 2.Focus on Educated people to sell product KP781. As The KP781 treadmill is having advanced features that sell for $2,500. 3.Helps in Increase in Revenue & profits.

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.

```
[46]: # Box plot representing Relationship bweteen Education & Miles.

sbn.boxplot(x="Education",y="Miles",data=df)
sbn.set(style="whitegrid")
```
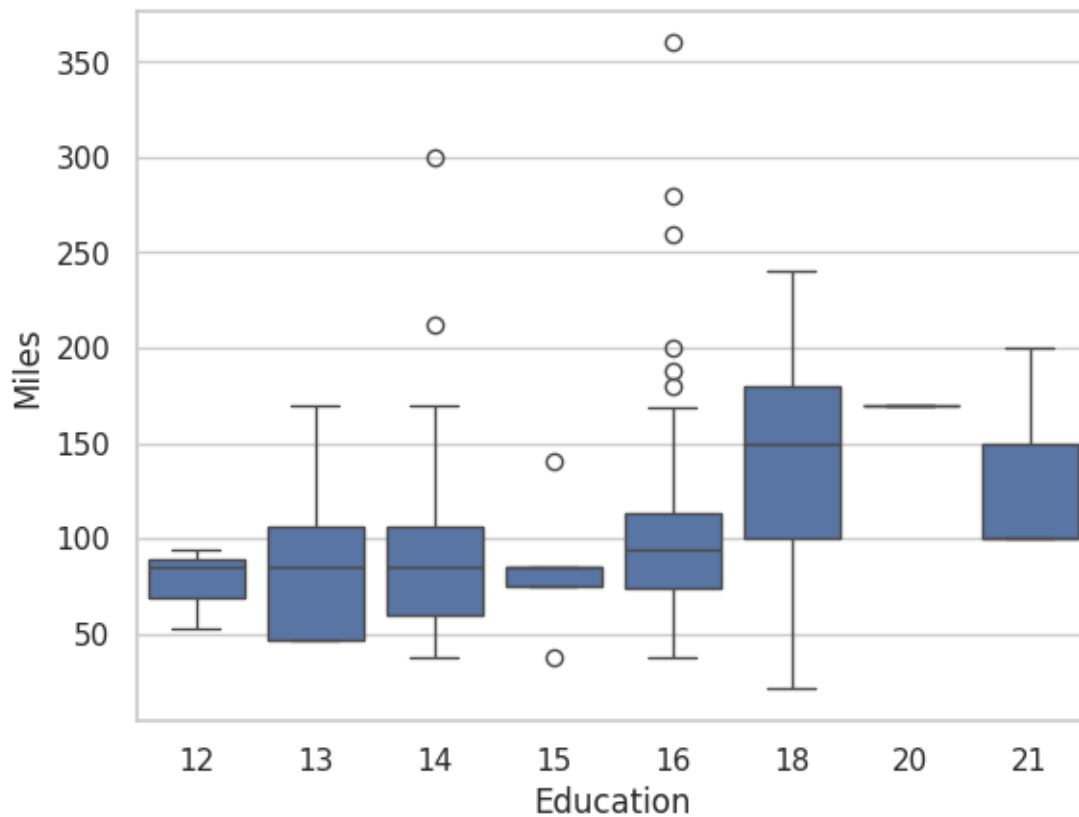

Key Insights:

1. It is observed from the above boxplot, that people upto 16 levels of education are having less miles of walking/running. 2.People with 18 levels of education are having more miles, i.e they are walking/running more.

Actionable Insights:

1. Focus on Gradute, post gradute customers as they are health concious comparing to under-grads.

```
[47]:  # Box plot representing   relationship between variables Education & Fitness.

       sbn.boxplot(x="Education",y="Fitness",data=df)
       sbn.set(style="whitegrid")
```
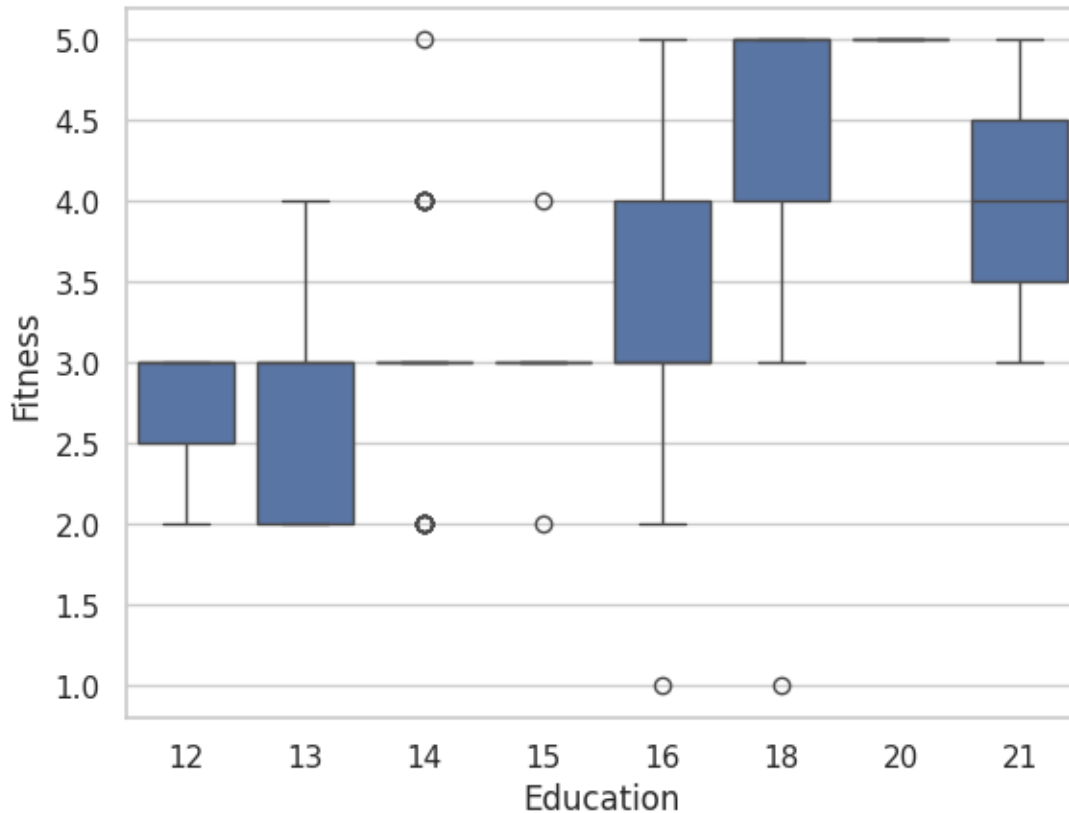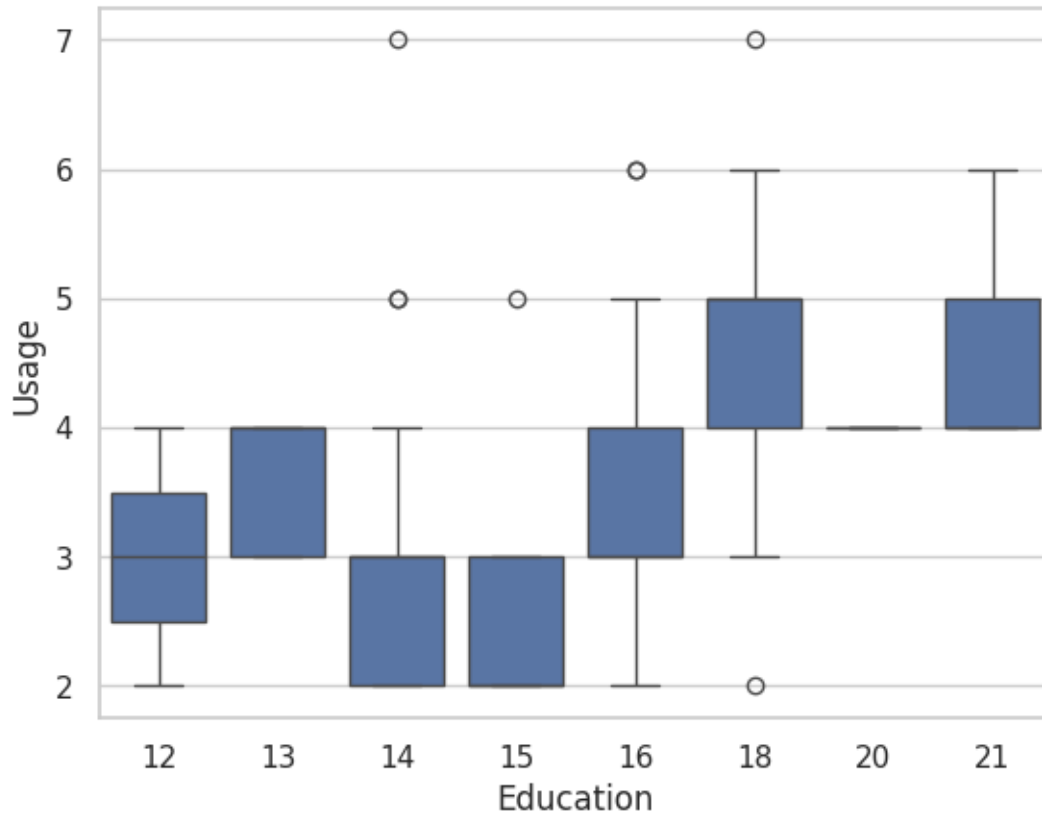

Key Insights:

1. It is observed from the above boxplot, that people with education level 16 and above,are
   having good fitness.

Actionable Insights:

1. Focus on Gradute, post gradute customers as they are health concious comparing to under-
   grads. 2.People with Higher education level, are having more disposable income. 3.Premium
   model KP781 can be sold to customers with higher education level.

```
[48]:  # Box plot representing Education vs usage.
       sbn.boxplot(x="Education",y="Usage",data=df)
       sbn.set(style="whitegrid")
```
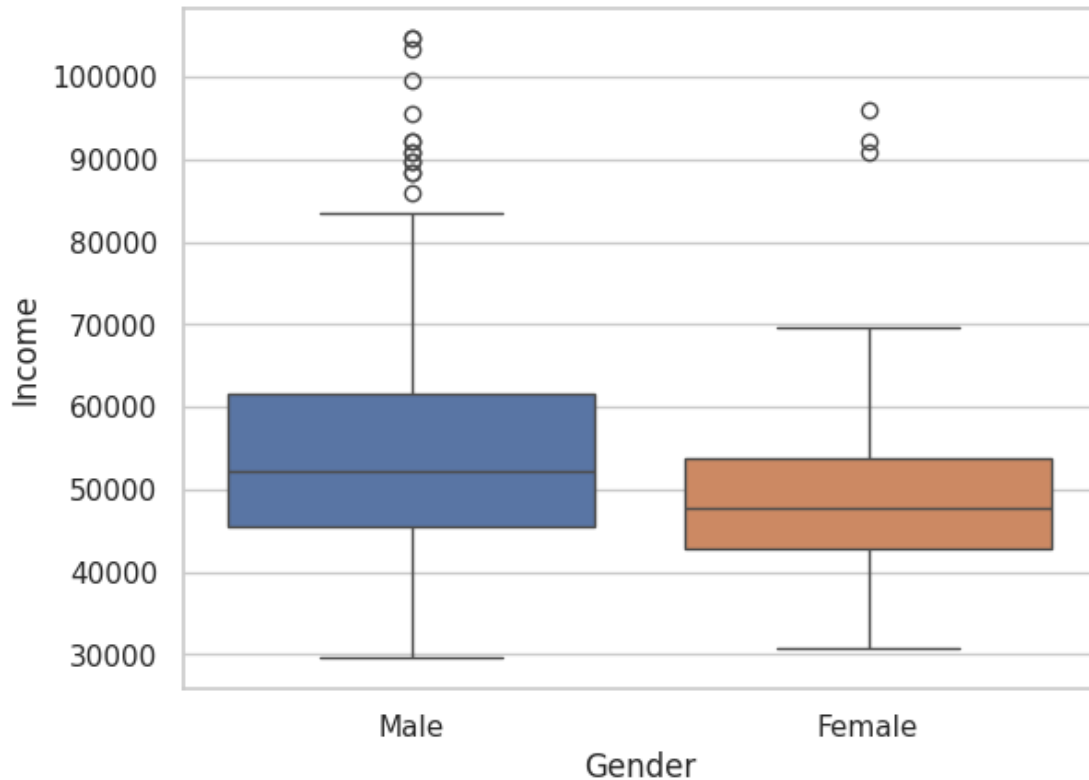
Key Insights:

1. It is evident from the above boxplot, that customers with higher education are having higher usage,

Actionable Insights:

1. Focus more on high usage cutomers to up sell new mdels with premium features.

```
[49]: # box plot showing Impact of Gender on Income.

sbn.boxplot(x="Gender",y="Income",hue='Gender',data=df)
sbn.set(style="whitegrid")
```
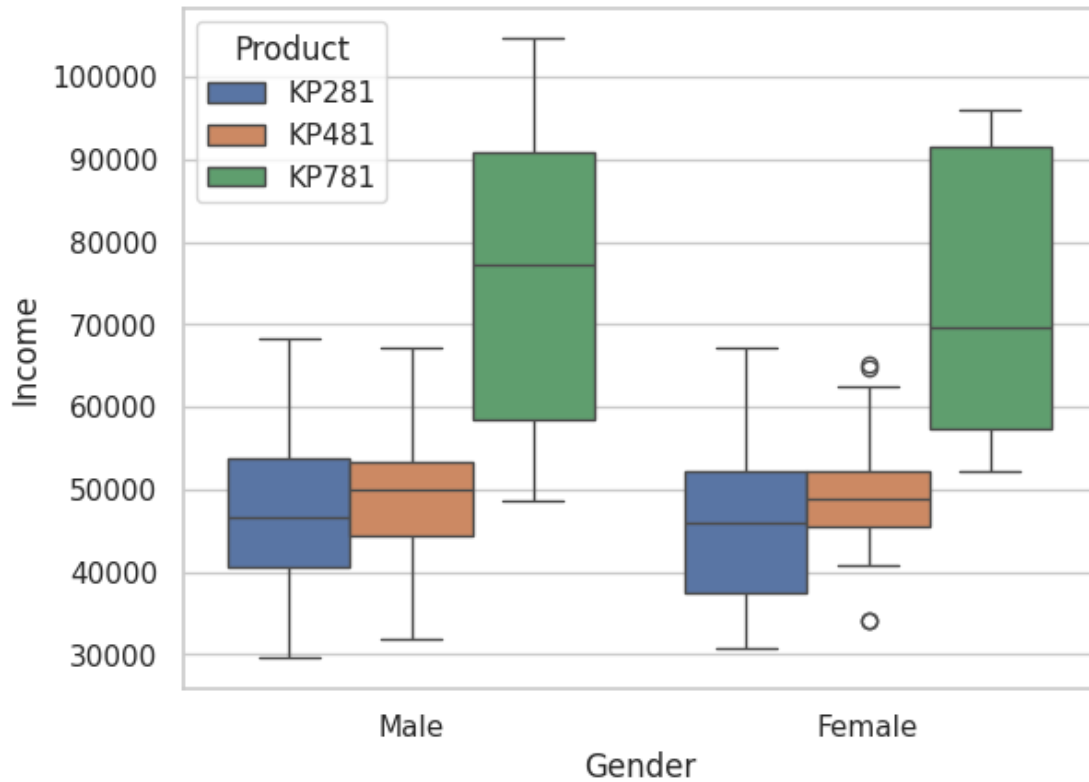
Key Insights:

1. Males are getting more income comparing to Females. 2.There is NO Gender Equality in Income

Actionable Insights:

1. Focus more on Male customers as they have more disposable income comapring to Female customers. 2.Male customers are the bread earners in most households. 3.Target marketing for Females is needed.

```
[50]: # Box plot showing relationship with respect to Gender,Income,product.
      sbn.boxplot(x="Gender",y="Income", hue='Product',data=df)
      sbn.set(style="whitegrid")
```
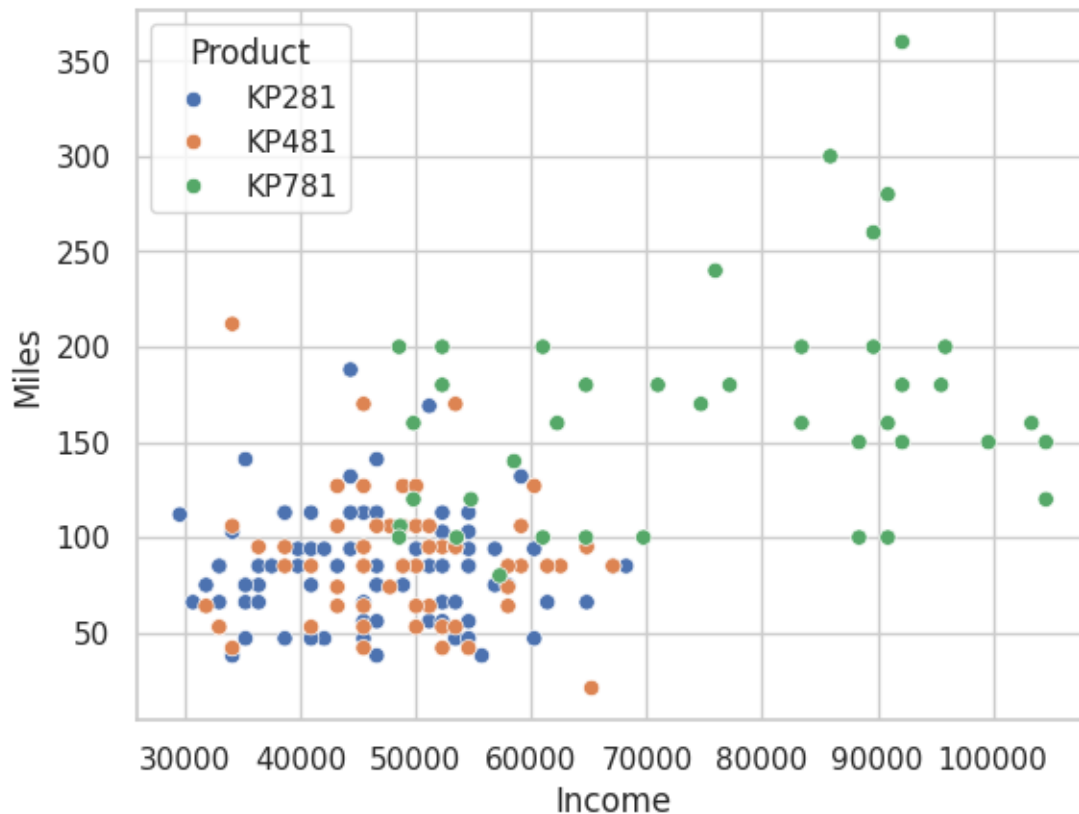
Key Insights:

1.People having higher Income, are buying KP 781 irrespective of Gender.

Actionable Insights:

1.Irrespective of gender, focus on high income & Highly educated people to sell premium models like KP781.

```
[51]: # Scatter plot representing products, income & miles.
      sbn.scatterplot(x="Income",y="Miles",hue='Product',data=df)
      sbn.set(style="whitegrid")
```

NOTE: Using scatter plot as both variables are continues

key Insights:

Customers with low income are having lesser miles comparing to High income customers.

1.People with low income are buying KP281,KP481 2.Peple with high Income are buying KP781 3.people with high Income are more health cautious & Running more Miles 4.People with annual Income of more than $ 70,000 are tend to buy higher model i.e KP781

Actionable Insights:

1.Target marketing is recommended to move people from 1st quadrent to 3rd quadent. 2.make strategies to push low variant product customers KP281,KP481 to Higher Model KP 781 3.It is clear that customers with High Education are having More income and they are having more miles of walk/running with high usage of treadmills. 4.Futher customer segementation is needed for focus target marketing.

```
[52]:  # Realtionship between Income & Product.

       sbn.boxplot(x="Product",y="Income",hue='Product',data=df)
       sbn.set(style="whitegrid")
```
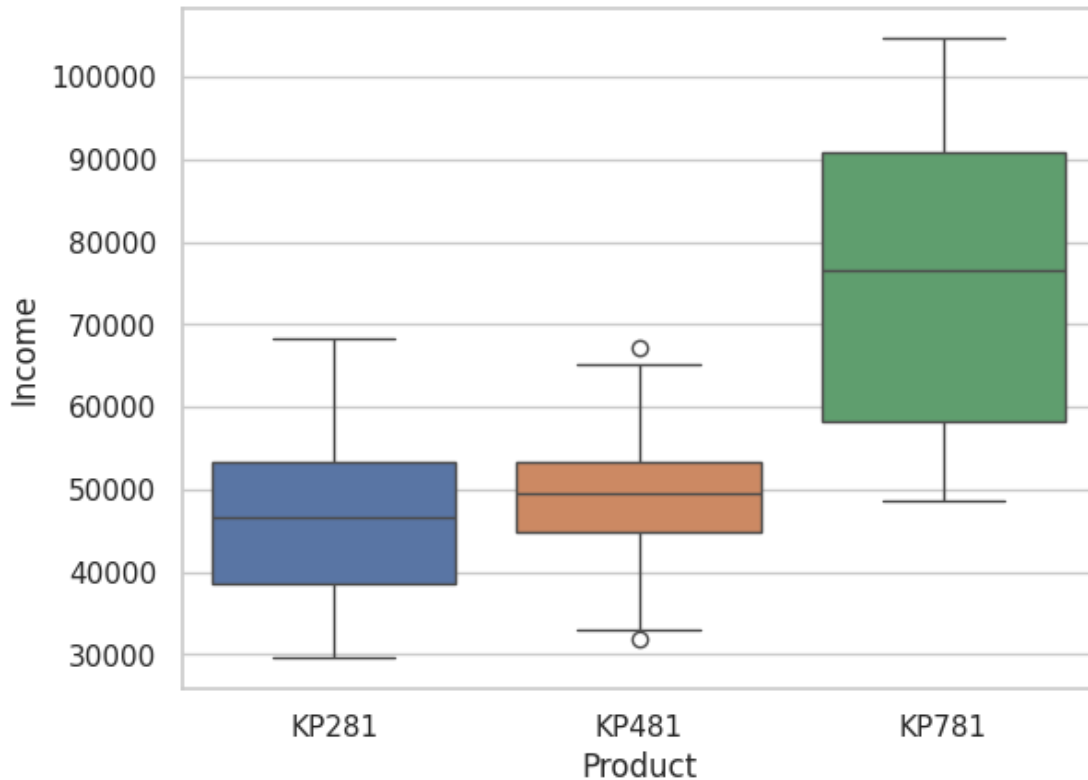
key Insights:

1.There is a strong realtionship between income & product. 2.Low income people prefer low models KP281 & KP481 3.High Income people prefer premium model KP781.

Actionable Insights:

1.Target marketing is needed to move KP481 user to KP781. 2.KP781 users can be upgraded with new premium models with new exclusive features.

```
[53]: #Check if features like marital status, age have any effect on the product␣
      ↪purchased (using countplot, histplots, boxplots etc)
      pd.
        ↪crosstab(index=df['MaritalStatus'],columns=[df['Product'],df['Age']],margins=True)
```

```
[53]: Product        KP281                               … KP781                    \
      Age            18 19 20 21 22 23 24 25 26 27  …    33 34 35 38 40 42 45
      MaritalStatus                                     …
      Partnered       0  1  2  2  1  4  2  5  5  2  …     1  0  1  1  0  0  0
      Single          1  2  0  2  3  4  3  2  2  1  …     0  1  0  0  1  1  1
      All             1  3  2  4  4  8  5  7  7  3  …     1  1  1  1  1  1  1

      Product            All
```

```
Age          47 48
MaritalStatus
Partnered     1  1  107
Single        0  0   73
All           1  1  180
```

```
[3 rows x 69 columns]
```

Key Insights:

1.Partnered/Married customers are 107 2.Single customers are 73 3.Total Customers (107+73)=180

Actionable Insights: Customer Segmentation on the basis of MaritalStatus

1. Around 60% customers are Partnered i.e, 59.44% (107/180*100)
2. Focus more on married prople to increase sales 3.Married people are having more disposable
   income 4.Married people are more health catious than single.

```
[54]: # histplot for marital status,age
      sbn.histplot(x="Age",hue='MaritalStatus',data=df)
      sbn.set(style="whitegrid")
```
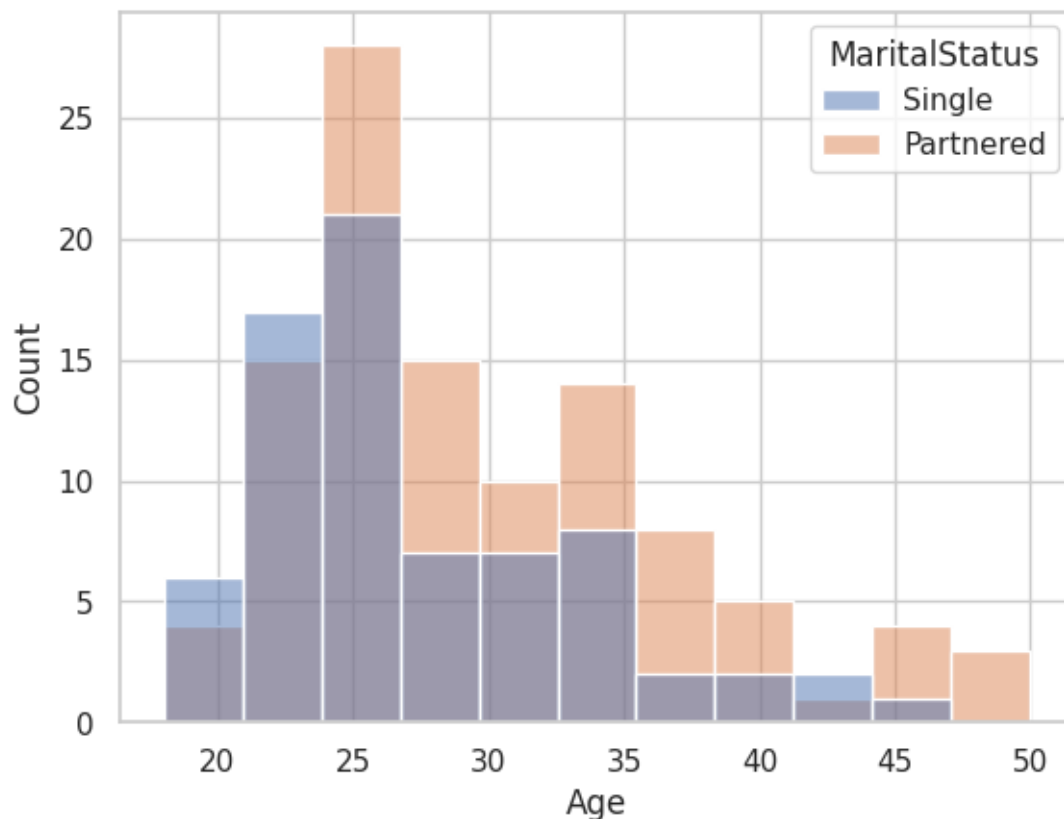

key Insights:

1.Most customers are of the age 25 2.Married i.e Partnered customers are buying more tredmills than single customers.
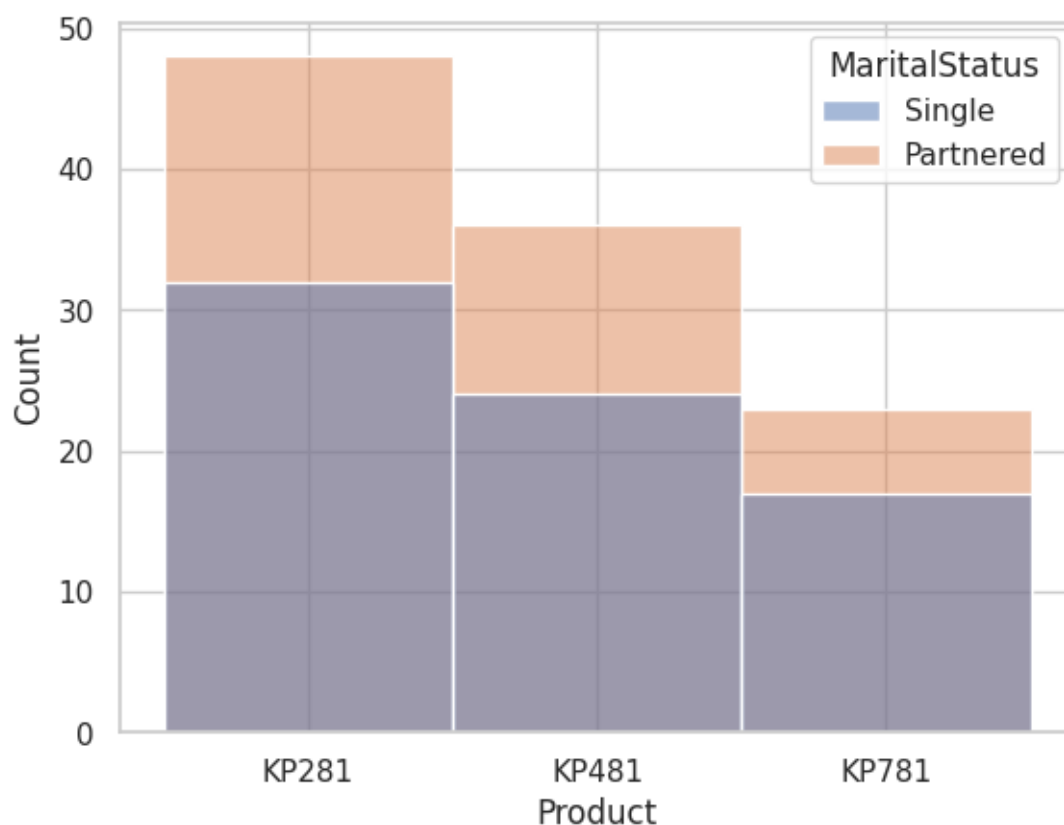
Actionable Insights:

1.Focus more on customer age group 25-35 2.Age 25-35 customers tend to buy more tredmills of premium quality & features.

Key Insights: 1.Partnered 59.44% 2.Single 40.55%

Actionable Insights:

1.Most customers are married people 2.Unique Special offers for Partnered & Single customers needed 3.KP781 premium model should be focused on Partnered customers

```
[55]:  # Hist plot for product,maritalstaus, age
       sbn.histplot(x="Product",hue='MaritalStatus',data=df)
       sbn.set(style="whitegrid")
```


key Insights:

1.KP281 is the most sold product 2.Most customers are from Partnered Maritalstatus.

Actionable Insights:

1. Give more attarctive offers for KP481 single customers & try to move them to KP781.

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

```
[56]: # what is the probality of female buying KP281 when all the products are␣
       ↪available
       pd.
        ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize=True)*100
```

```
[56]: Product       KP281       KP481       KP781          All
       Gender
       Female    22.222222   16.111111    3.888889    42.222222
       Male      22.222222   17.222222   18.333333    57.777778
       All       44.444444   33.333333   22.222222   100.000000
```

**Key Insights:** (40/76) * (76/180) * 100 = 22.222

NOTE: p(A^B) = P(A/B) * P(B) P(KP281 ^ female) = P(KP281/female) * P(Female)

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

6 Multi-variate Analysis using Heatmap correaltion

```
[57]: # Create new DataFrame df1.
       # Select only numerical features for correlation calculation
       df1 = df.select_dtypes(include=['number'])

       # Calculate correlation matrix on numerical features only
       sbn.heatmap(df1.corr(), annot=True)
       sbn.set(style="whitegrid")
```
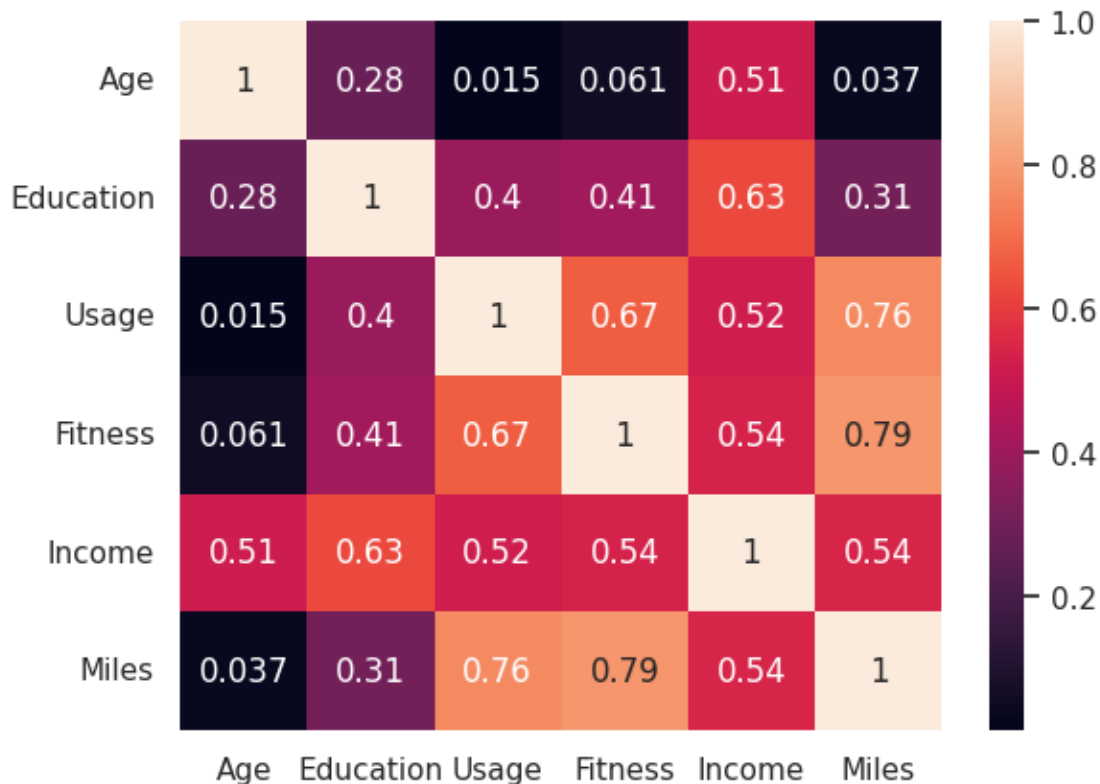
Key insights:

1.There is a strong positive correlation between people with High Education & High Income (0.63)
2.There is a strong positive correlation between people with High Usage & More Fitness (0.67)
3.There is a strong positive correlation between people with more miles & More Fitness (0.79)

Actionable Insights:

1.To increase sales, more focus is needed for customers with High education & High Income.

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

```
[58]: # What is the probability of a male customer buying a KP781 treadmill?

      pd.
      ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize='columns')*100
```

```
[58]: Product  KP281       KP481  KP781         All
      Gender
      Female    50.0  48.333333   17.5  42.222222
      Male      50.0  51.666667   82.5  57.777778
```

key insights:

KP781 bought by;

1.Female $7/40 = 17.5\%$ 2.Males $33/40 = 82.5\%$

Actionable Insights:

1.Males are buying more KP781 model i. 82.5% of over all quantity sold i.e 40. 2.Hence Focus on Males with hihger education& Income to sell premium model KP781.

8   7.Customer Profiling - Categorization of users.

```
[59]: # calculating Count of Partnered, single percentage
      pd.crosstab(index=df['MaritalStatus'],columns='count',normalize=True)*100
```

```
[59]: col_0               count
      MaritalStatus
      Partnered       59.444444
      Single          40.555556
```

Key Insights: 1.Partnered cusomters are majority with 59.44%, followed by single 40.55%

Actionable Insights:

1.Partnered customers are tend to buy premium models KP781 2.there is scope to upsell single user with more benefits.

9   8.Probability- marginal, conditional probability.

```
[60]: pd.crosstab(index=df['Gender'],columns=df['Product'],margins=True)
```

```
[60]: Product  KP281  KP481  KP781  All
      Gender
      Female      40     29      7   76
      Male        40     31     33  104
      All         80     60     40  180
```

```
[61]: # Assuming; If the manufacture wants to produce model KP 781, 1000 units how␣
      ↪much will be purchased by females & males respectively ?
      pd.
      ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize='columns')*100
```

```
[61]: Product  KP281      KP481  KP781        All
      Gender
      Female    50.0  48.333333   17.5  42.222222
      Male      50.0  51.666667   82.5  57.777778
```

Insights:

10 Using conditional probablity

The current ratio of KP 781 female is 17.5% and male is 82.5% Hence it can be concluded that;

female ==> $1000 \cdot 17.5\% = 175$ units Male ==> $1000 \cdot 82.5\% = 825$ units

Total ==> 175+825=1000 units

```
[62]: # Assuming; there is a Exhibiton on fitness products, and we are expecting 2000␣
      ↪females to visit the exhibition, Find how many females will buy KP481?
      pd.
      ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize='index')*100
```

```
[62]: Product        KP281        KP481        KP781
      Gender
      Female    52.631579   38.157895    9.210526
      Male      38.461538   29.807692   31.730769
      All       44.444444   33.333333   22.222222
```

Key Insights:

1. Only females are visiting the Exhibition.
2. Only product sold in exhibition is Model KP481

Calculation:

KP481 Only females:

Only females 29/76*2000 = 763

OR 2000 * 38.15% = 763

Actionable Insights:

1. Aerofit company must keep ready at least 763 units of KP481 model, to be sold in female exhibition.

```
[63]: # Assumption: Arrange a meeting for 800 males & females....How many Chairs to␣
      ↪keep on Left side & right side.
      #Method 1
      pd.
      ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize='columns')*100
```

```
[63]: Product  KP281        KP481  KP781          All
      Gender
      Female    50.0   48.333333   17.5    42.222222
      Male      50.0   51.666667   82.5    57.777778
```

Using Marginal probablity:

Total Females ==> $76/180 \cdot 800 = 338$ Total Males ==> $104/180 \cdot 800 = 462$

```
[64]: # Method 2 to get the percentages directly.( use normalize=True)
      pd.
        ↪crosstab(index=df['Gender'],columns=df['Product'],margins=True,normalize=True)*100
```

```
[64]: Product       KP281      KP481      KP781         All
      Gender
      Female    22.222222  16.111111   3.888889   42.222222
      Male      22.222222  17.222222  18.333333   57.777778
      All       44.444444  33.333333  22.222222  100.000000
```

Key Insights:

All Female : 42.222% All Male : 57.777%

Actionable Insights:

1.Arrange 338 chairs for females 2.Arrange 462 chairs for males

Females ==> 800*42.222% = 338 Males ==> 800*57.777% = 462

11   9. Some recommendations and actionable insights, based on the inferences.

12   Recommendations:

1.As female custoers are less comparing to males,There is good scope of increasing female customers. 2.create marketing strategies to attract females as there population is almost 50%. 3.Create fitness awarness progrmames for females. 4.Target marketing is recommended to move people from 1st quadrent to 3rd quadent. 5.make strategies to push low variant product customers KP281,KP481 to Higher Model KP 781 6.Focus on people above 25 years age , as there disposable income is high. 7.Focus on Educated people to sell product KP781. As The KP781 treadmill is having advanced features that sell for \$2,500. 8.Focus on Gradute, post gradute customers as they are health concious comparing to undergrads. 9.Target marketing is needed,for high usage cutomers to up sell new mdels with premium features. 10.Special Target Marketing, Target marketing for Females is needed. 11.Target marketing is needed to move KP481 user to KP781. 12.KP781 users can be upgraded with new premium models with new exclusive features. 13.There is a strong positive correlation between people with High Education & High Income (0.63) 14.Males are buying more KP781 model i. 82.5% of over all quantity sold i.e 40.Hence Focus on Males with hihger education& Income to sell premium model KP781. 15.Most customers are married people. Focus on couples for more sales. 16.Focus more on customer age group 25-35 17.Age 25-35 customers tend to buy more tredmills of premium quality & features.

13   Conclusion:

14   Further Analysis is recommended,Futher customer segementation is needed for focus target marketing to Increase in Revenue & profits.