Aerofit_case_Aakash

January 8, 2024

0.1 About Aerofit

Aerofit is a leading brand in the field of fitness equipment. 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.

0.2 Objective

Create comprehensive customer profiles for each AeroFit treadmill product through descriptive analytics. Develop two-way contingency tables and analyze conditional and marginal probabilities to discern customer characteristics, facilitating improved product recommendations and informed business decisions.

0.3 Product Portfolio

- The KP281 is an entry-level treadmill that sells for USD 1,500.
- The KP481 is for mid-level runners that sell for USD 1,750.
- The KP781 treadmill is having advanced features that sell for USD 2,500.

0.4 Features of the dataset:

- Product: Product Purchased KP281, KP481, or KP781
- Age: In years
- Gender: Male/Female
- Education: in years
- MaritalStatus: single or partnered
- Usage: 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 poor shape and 5 is the excellent shape.
- Miles: average number of miles the customer expects to walk/run each week

1 Exploratory Data Analysis

[39]:
```
# import all important libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```python
from scipy.stats import norm
```

[2]:
```python
#importing the data set
data_path="https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/
↪125/original/aerofit_treadmill.csv"
df=pd.read_csv(data_path)
df
```

[2]:
```
    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]
```

[3]:
```python
df.shape
```

[3]: (180, 9)

[4]:
```python
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
```

1.0.1 Insights

- From the above analysis, it is clear that, data has total of 9 features with mixed alpha numeric data. Also we can see that there is no missing data in the columns.
- The data type of all the columns are matching with the data present in them.

1.1 Statistical Summary

[5]:
```
# statisctical summary of object type columns
df.describe(include = 'object')
```

[5]:
```
       Product Gender MaritalStatus
count      180    180           180
unique       3      2             2
top      KP281   Male     Partnered
freq        80    104           107
```

[6]:
```
# statisctical summary of numerical data type columns

df.describe()
```

[6]:
```
              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
```

1.1.1 Insights

1. Age - The age range of customers spans from 18 to 50 year, with an average age of 29 years.

2. Education - Customer education levels vary between 12 and 21 years, with an average education duration of 16 years.

3. Usage - Customers intend to utilize the product anywhere from 2 to 7 times per week, with an average usage frequency of 3 times per week.

4. Fitness - On average, customers have rated their fitness at 3 on a 5-point scale, reflecting a moderate level of fitness.

5. Income - The annual income of customers falls within the range of USD 30,000 to USD 100,000, with an average income of approximately USD 54,000.

6. Miles - Customers' weekly running goals range from 21 to 360 miles, with an average target of 103 miles per week.

1.2 Duplicate Detection

[7]:
```
df.duplicated().value_counts()
```

[7]:
```
False    180
dtype: int64
```

1.2.1 Insights

- There are no duplicate entries in the dataset

1.3 Sanity Check for columns

[8]:
```
# checking the unique values for columns
for i in df.columns:
    print('Unique Values in',i,'column are :-')
    print(df[i].unique())
    print('-'*70)
```

```
Unique Values in Product column are :-
['KP281' 'KP481' 'KP781']
----------------------------------------------------------------------
Unique Values in Age column are :-
[18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41
 43 44 46 47 50 45 48 42]
```

```
----------------------------------------------------------------------
Unique Values in Gender column are :-
['Male' 'Female']
----------------------------------------------------------------------
Unique Values in Education column are :-
[14 15 12 13 16 18 20 21]
----------------------------------------------------------------------
Unique Values in MaritalStatus column are :-
['Single' 'Partnered']
----------------------------------------------------------------------
Unique Values in Usage column are :-
[3 2 4 5 6 7]
----------------------------------------------------------------------
Unique Values in Fitness column are :-
[4 3 2 1 5]
----------------------------------------------------------------------
Unique Values in Income column are :-
[ 29562  31836  30699  32973  35247  37521  36384  38658  40932  34110
  39795  42069  44343  45480  46617  48891  53439  43206  52302  51165
  50028  54576  68220  55713  60261  67083  56850  59124  61398  57987
  64809  47754  65220  62535  48658  54781  48556  58516  53536  61006
  57271  52291  49801  62251  64741  70966  75946  74701  69721  83416
  88396  90886  92131  77191  52290  85906 103336  99601  89641  95866
 104581  95508]
----------------------------------------------------------------------
Unique Values in Miles column are :-
[112  75  66  85  47 141 103  94 113  38 188  56 132 169  64  53 106  95
 212  42 127  74 170  21 120 200 140 100  80 160 180 240 150 300 280 260
 360]
----------------------------------------------------------------------
```

[9]:
```python
# checking the number of unique values for columns
for i in df.columns:
  print('Number of Unique Values in',i,'column are :-')
  print(df[i].nunique())
  print('-'*70)
```

```
Number of Unique Values in Product column are :-
3
----------------------------------------------------------------------
Number of Unique Values in Age column are :-
32
----------------------------------------------------------------------
Number of Unique Values in Gender column are :-
2
----------------------------------------------------------------------
Number of Unique Values in Education column are :-
8
```

```
--------------------------------------------------------------------------
Number of Unique Values in MaritalStatus column are :-
2
--------------------------------------------------------------------------
Number of Unique Values in Usage column are :-
6
--------------------------------------------------------------------------
Number of Unique Values in Fitness column are :-
5
--------------------------------------------------------------------------
Number of Unique Values in Income column are :-
62
--------------------------------------------------------------------------
Number of Unique Values in Miles column are :-
37
--------------------------------------------------------------------------
```

1.3.1 Insights

- The dataset does not contain any abnormal values.

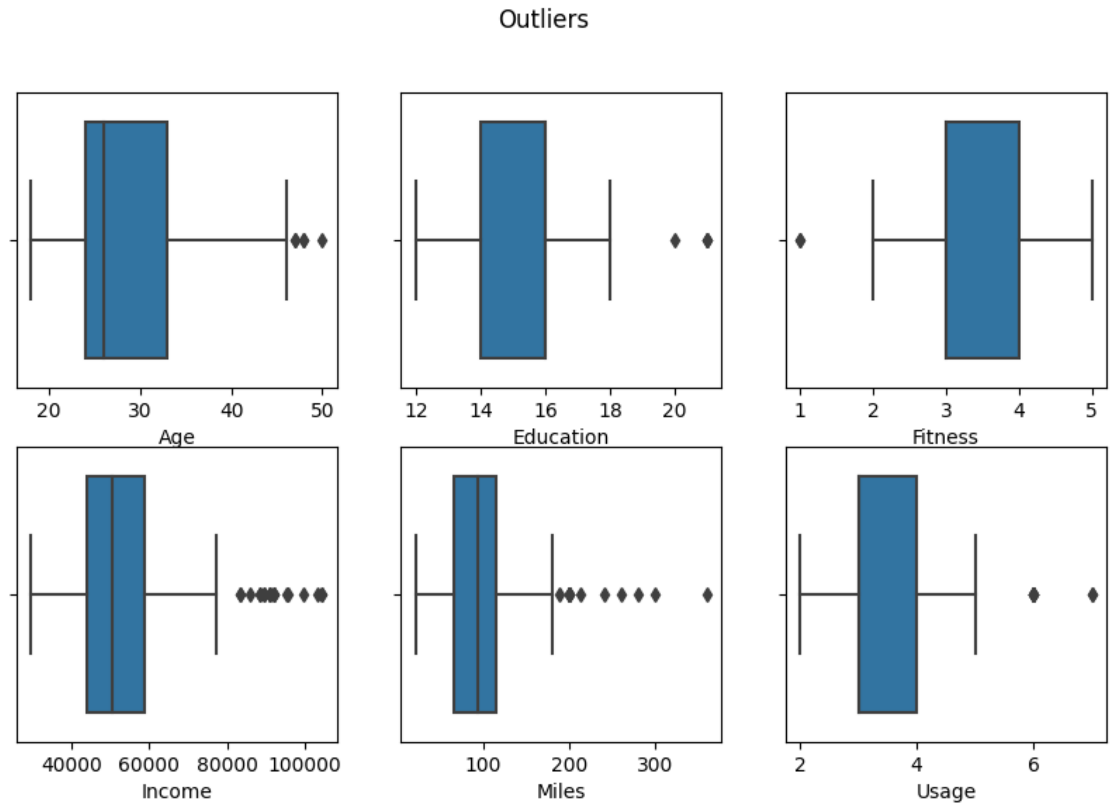
2 Detecting Outliers

Visual Analysis:

2.1 Finding outliers using Boxplot

[10]:
```
fig,ax=plt.subplots(2,3,figsize=(10,6))
fig.suptitle("Outliers")

plt.subplot(2,3,1)
sns.boxplot(data=df,x="Age")

plt.subplot(2,3,2)
sns.boxplot(data=df,x="Education")

plt.subplot(2,3,3)
sns.boxplot(data=df,x="Fitness")

plt.subplot(2,3,4)
sns.boxplot(data=df,x="Income")

plt.subplot(2,3,5)
sns.boxplot(data=df,x="Miles")

plt.subplot(2,3,6)
sns.boxplot(data=df,x="Usage")
```

```
plt.show()
```



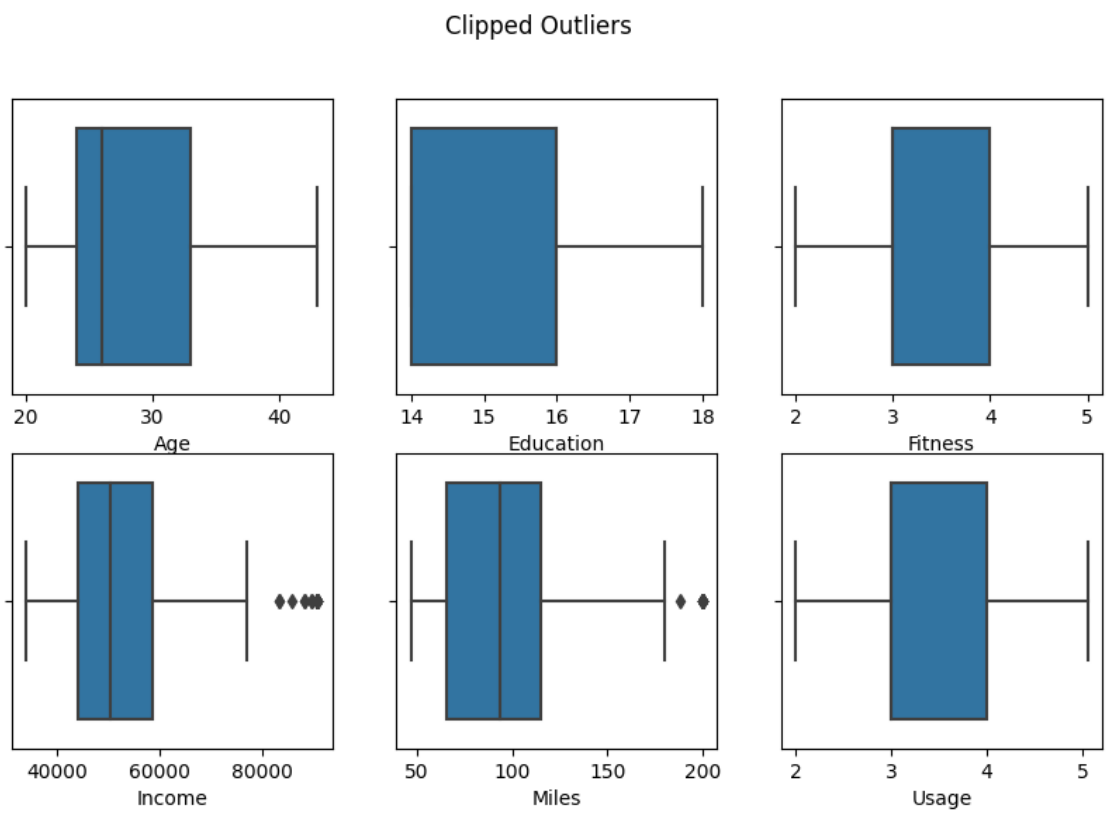
2.1.1 Insights:

Based on the graphical representation, it is evident that both Income and Miles exhibit a substantial number of outliers. In contrast, the remaining variables display only a minor presence of outliers.

2.2 Removing/clipping the data between the 5 percentile and 95 percentile

[11]:
```
# Clipping the data between the 5th and 95th percentiles
clipped_age = np.clip(df['Age'], np.percentile(df['Age'], 5), np.
 ↪percentile(df['Age'], 95))
clipped_education = np.clip(df['Education'], np.percentile(df['Education'], 5),␣
 ↪np.percentile(df['Education'], 95))
clipped_income = np.clip(df['Income'], np.percentile(df['Income'], 5), np.
 ↪percentile(df['Income'], 95))
clipped_usage = np.clip(df['Usage'], np.percentile(df['Usage'], 5), np.
 ↪percentile(df['Usage'], 95))
clipped_miles = np.clip(df['Miles'], np.percentile(df['Miles'], 5), np.
 ↪percentile(df['Miles'], 95))
```

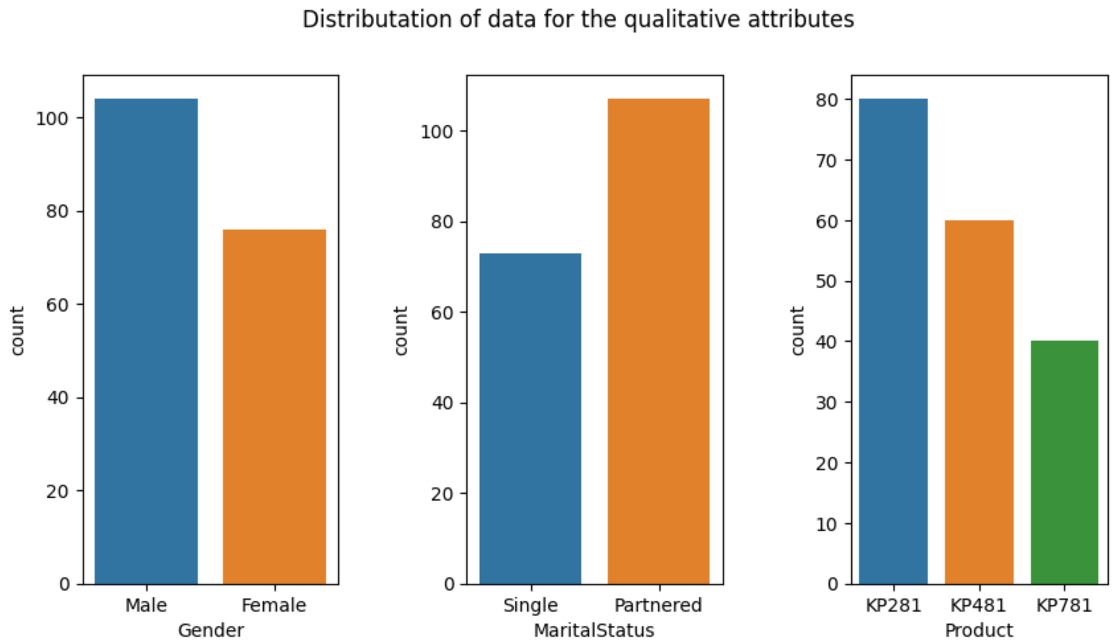
```
clipped_fitness = np.clip(df['Fitness'], np.percentile(df['Fitness'], 5), np.
    ↪percentile(df['Fitness'], 95))

fig,ax=plt.subplots(2,3,figsize=(10,6))
fig.suptitle("Clipped Outliers")


plt.subplot(2,3,1)
sns.boxplot(data=df,x=clipped_age)

plt.subplot(2,3,2)
sns.boxplot(data=df,x=clipped_education)

plt.subplot(2,3,3)
sns.boxplot(data=df,x=clipped_fitness)

plt.subplot(2,3,4)
sns.boxplot(data=df,x=clipped_income)

plt.subplot(2,3,5)
sns.boxplot(data=df,x=clipped_miles)

plt.subplot(2,3,6)
sns.boxplot(data=df,x=clipped_usage)
plt.show()
```



3 Checking if features like marital status, Gender, and age have any effect on the product purchased.

3.1 Univariate Analysis:

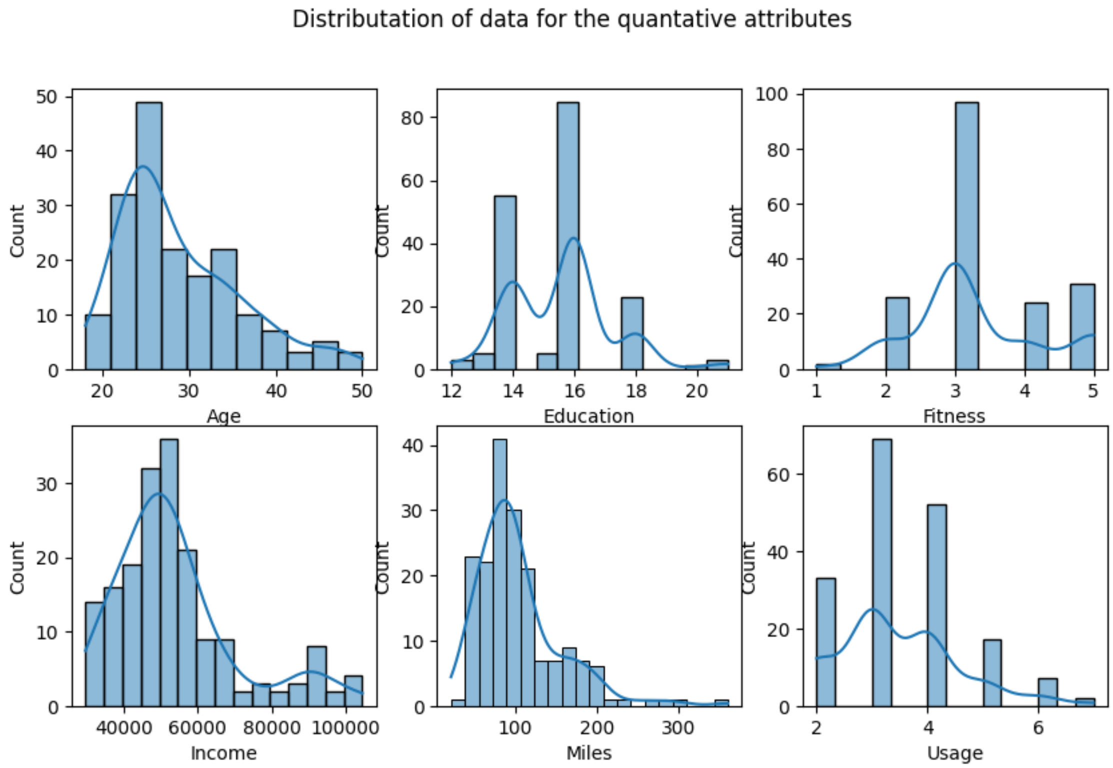
[13]:
```
fig, ax = plt.subplots(1, 3, figsize=(10, 5))
fig.suptitle("Distributation of data for the qualitative attributes")

plt.subplot(1, 3, 1)
sns.countplot(data=df, x="Gender")

plt.subplot(1, 3, 2)
sns.countplot(data=df, x="MaritalStatus")

plt.subplot(1, 3, 3)
sns.countplot(data=df, x="Product")

plt.subplots_adjust(wspace=0.5)
plt.show()
```



3.1.1 Insights:

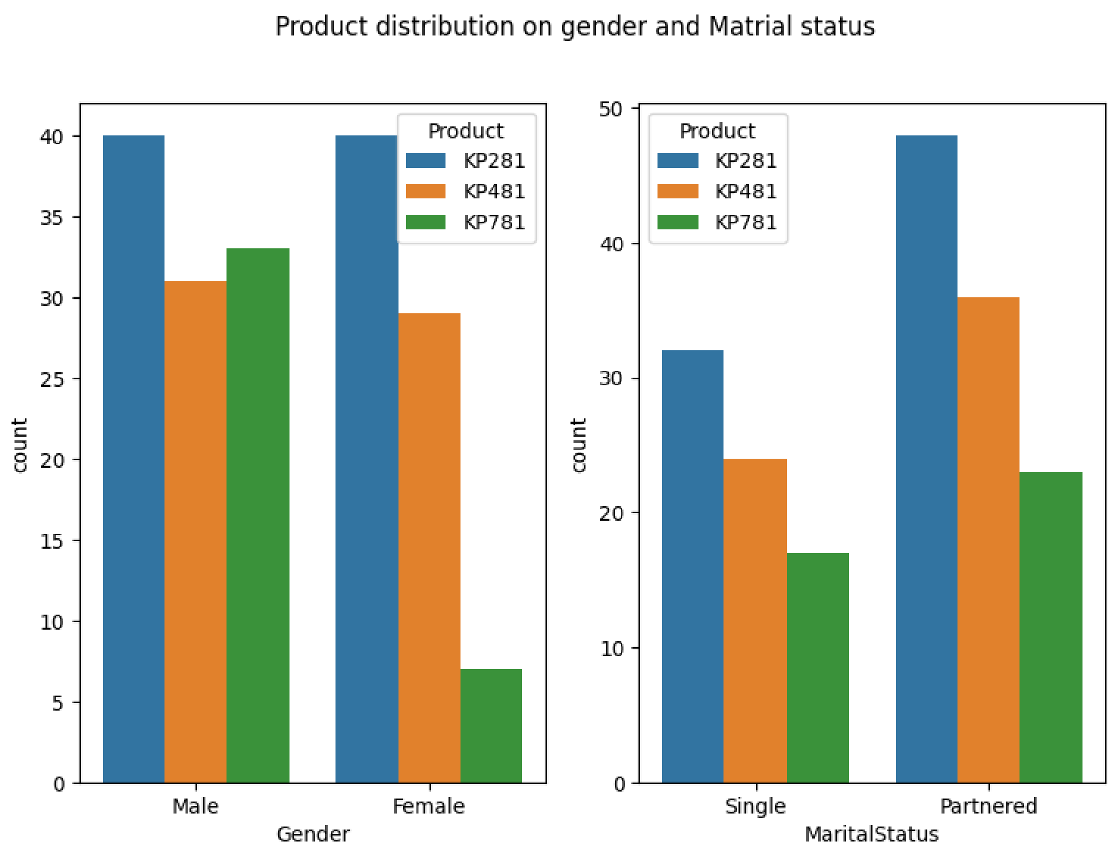
- In the given data, there appears to be a higher number of male customers compared to female customers. Additionally, it seems that partnered customers are more prevalent. Furthermore, it is evident that the product KP281 is the most frequently purchased by customers.

[48]:
```
#Distributation of data for the quantative attributes
fig,ax=plt.subplots(2,3,figsize=(10,6))
fig.suptitle("Distributation of data for the quantative attributes")
plt.subplot(2,3,1)
sns.histplot(data=df,x="Age",kde=True)
plt.subplot(2,3,2)
sns.histplot(data=df,x="Education",kde=True)
plt.subplot(2,3,3)
sns.histplot(data=df,x="Fitness",kde=True)
plt.subplot(2,3,4)
sns.histplot(data=df,x="Income",kde=True)
plt.subplot(2,3,5)
sns.histplot(data=df,x="Miles",kde=True)
plt.subplot(2,3,6)
sns.histplot(data=df,x="Usage",kde=True)
plt.show()
```



3.2 Bivariate Analysis

[15]:
```
#Product distribution on gender and Matrial status
fig,ax=plt.subplots(1,2,figsize=(9,6))
fig.suptitle("Product distribution on gender and Matrial status")

plt.subplot(1,2,1)
sns.countplot(data=df,x="Gender",hue="Product")

plt.subplot(1,2,2)
sns.countplot(data=df,x="MaritalStatus",hue="Product")

plt.show()
```

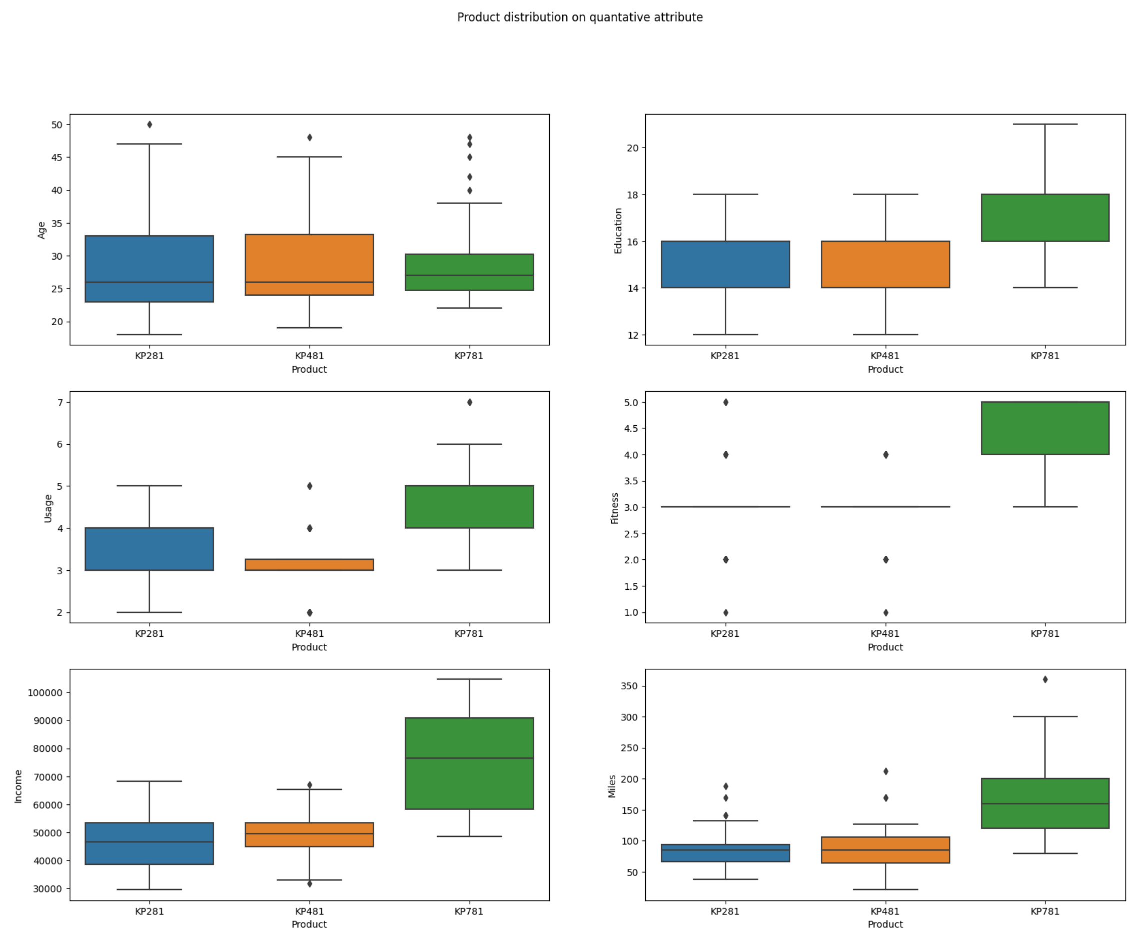


3.2.1 Insights: While both males and females do use KP281, KP781 is predominantly utilized by males. The usage of KP781 among males is notably higher compared to its relatively limited usage among females.

[16]:

```
#Product distribution on quantative attribute
fig,ax=plt.subplots(3,2,figsize=(20,15))
fig.suptitle("Product distribution on quantative attribute")

plt.subplot(3,2,1)
sns.boxplot(data=df,x="Product",y="Age")

plt.subplot(3,2,2)
sns.boxplot(data=df,x="Product",y="Education")

plt.subplot(3,2,3)
sns.boxplot(data=df,x="Product",y="Usage")

plt.subplot(3,2,4)
sns.boxplot(data=df,x="Product",y="Fitness")
```

```
plt.subplot(3,2,5)
sns.boxplot(data=df,x="Product",y="Income")

plt.subplot(3,2,6)
sns.boxplot(data=df,x="Product",y="Miles")
plt.show()
```



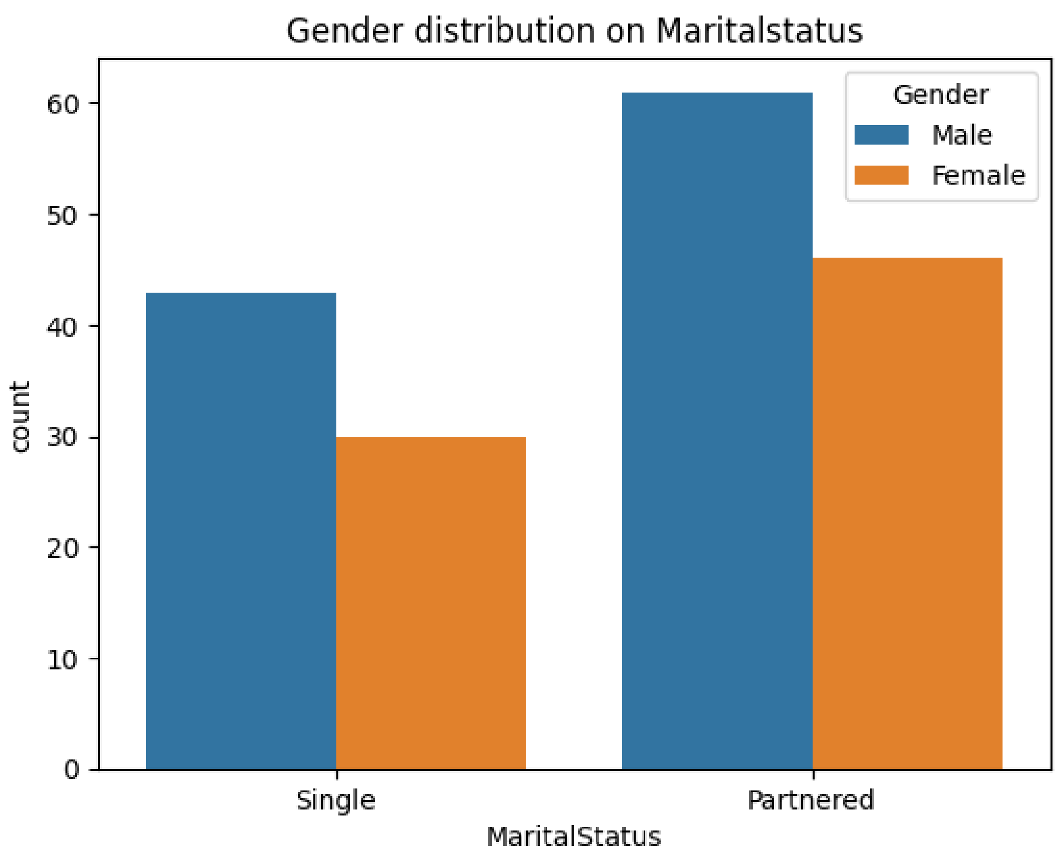
3.2.2 Insights:

- Product vs Age: Both KP281 and KP481 products appear to be popular among customers aged between 22 to 33 years old. On the other hand, KP781 seems to be favored by customers in the 22 to 28 age group, and interestingly, it gains popularity among customers over 40 years old.
- Product vs Education: Customers who predominantly purchase KP281 and KP481 products tend to have a maximum education level of 16 years. In contrast, those who have pursued higher education, up to 18 years or more, seem to prefer KP781.
- Product vs Usage: It appears that customers who intend to use the treadmill more frequently,

specifically greater than four times a week, are more inclined to purchase the KP781 product. On the other hand, customers with different usage patterns are more likely to opt for KP281 or KP481.

- Product vs Fitness: Customers who are opting for the KP781 product may be considered to be in better physical fitness compared to those choosing KP281 and KP481. This assumption suggests that KP781 might cater to a more fitness-conscious or health-oriented customer base.
- Product vs Income: Higher-income customers favor KP781, middle-income customers prefer KP281, and slightly higher middle-income customers opt for KP481, highlighting income's role in product selection.
- Product vs Miles: KP781 offers the highest mileage range, indicating it's ideal for intense workouts, while KP281 and KP481 are better suited for moderate exercise, helping customers match their fitness goals with the right treadmill.

[17]:
```
sns.countplot(data=df,x="MaritalStatus",hue="Gender")
plt.title("Gender distribution on Maritalstatus")
plt.show()
```



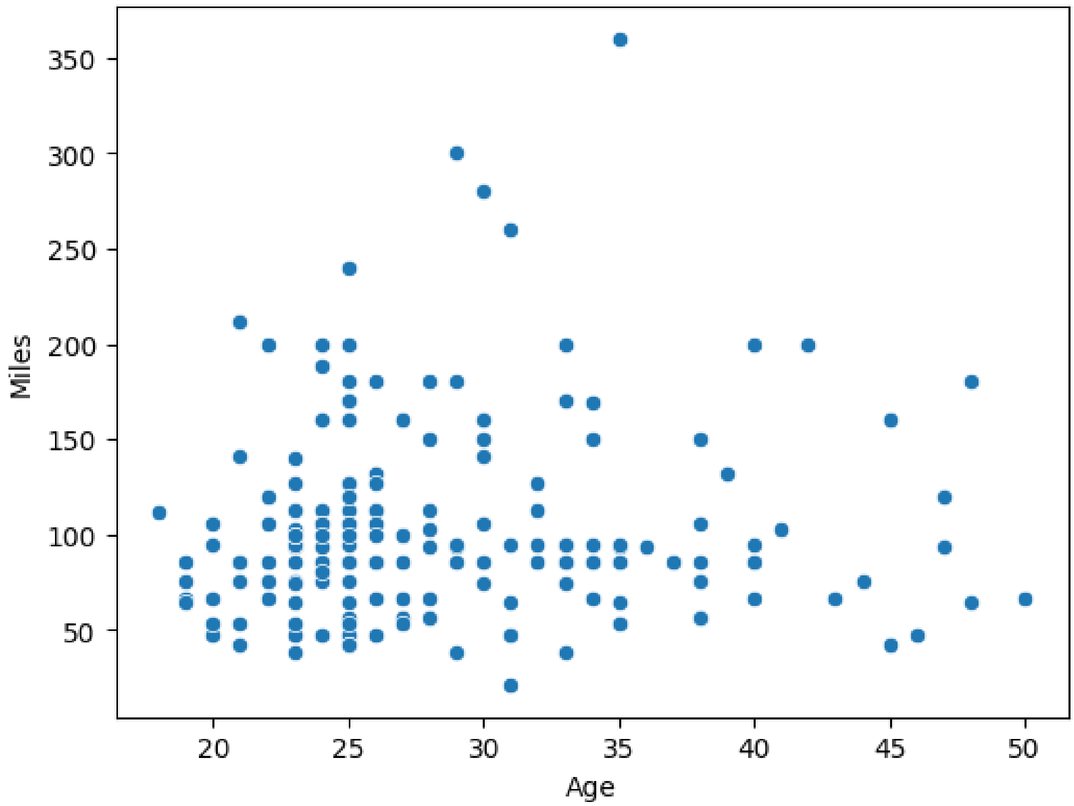
[18]:
```
df.head(3)
```

[18]:

	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

[19]:

```
sns.scatterplot(data=df,x="Age",y="Miles")
plt.show()
```



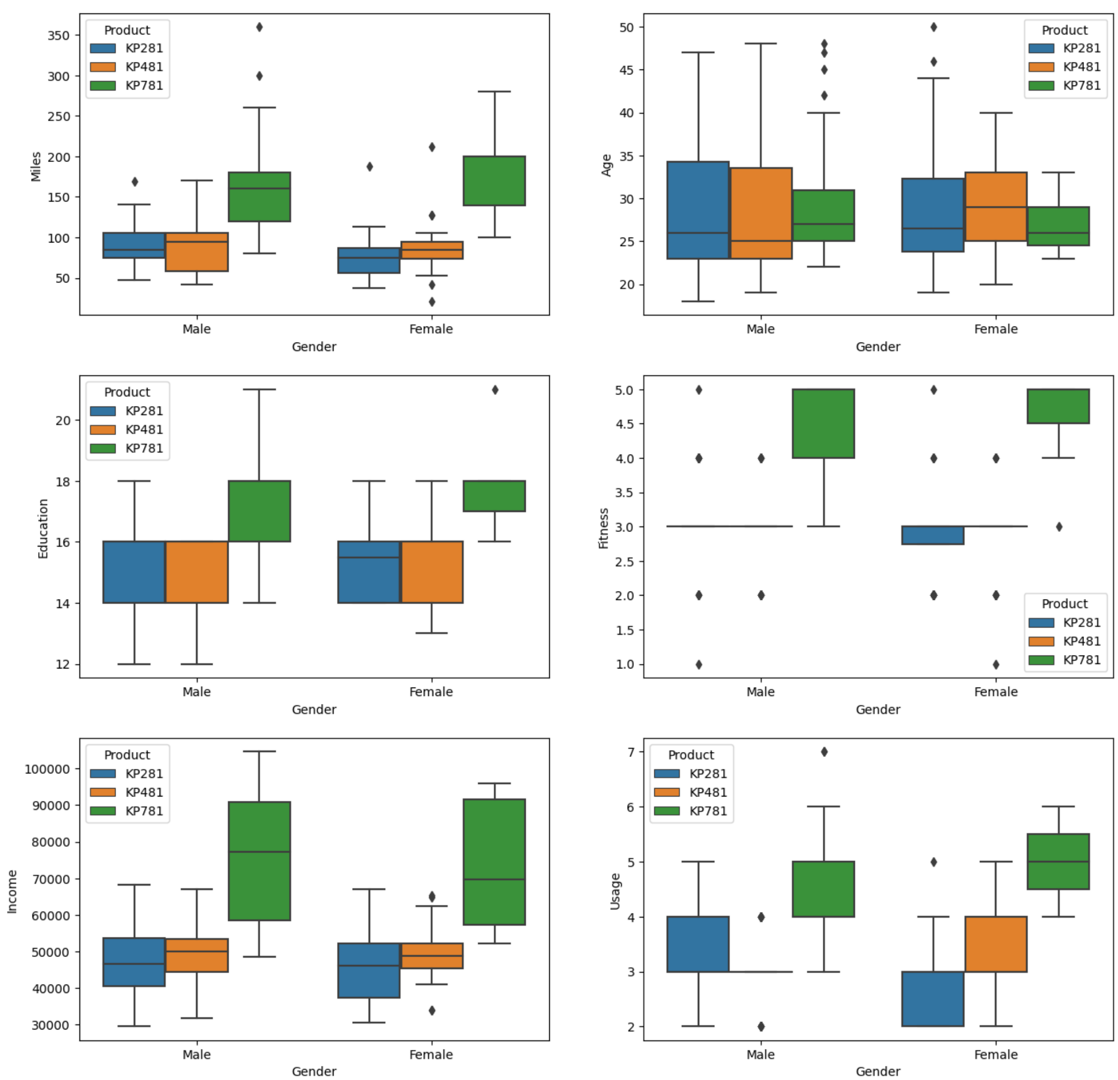
3.3 Multivariate Analysis

[20]:

```
fig,ax=plt.subplots(3,2,figsize=(15,15))
fig.suptitle("Product and Gender distribution on Quantitive attribute")

plt.subplot(3,2,1)
sns.boxplot(data=df,x="Gender",y="Miles",hue="Product")

plt.subplot(3,2,2)
sns.boxplot(data=df,x="Gender",y="Age",hue="Product")

plt.subplot(3,2,3)
sns.boxplot(data=df,x="Gender",y="Education",hue="Product")
```

```python
plt.subplot(3,2,4)
sns.boxplot(data=df,x="Gender",y="Fitness",hue="Product")

plt.subplot(3,2,5)
sns.boxplot(data=df,x="Gender",y="Income",hue="Product")

plt.subplot(3,2,6)
sns.boxplot(data=df,x="Gender",y="Usage",hue="Product")

plt.show()
```

Product and Gender distribution on Quantitive attribute



4 Representing the Probability

4.1 Adding new columns for better analysis

- Creating New Column and Categorizing values in Age , Education, Income and Miles to different classes for better visualization.

4.1.1 Age Column

- Categorizing the values in age column in 4 different buckets:

1. Young Adult: from 18 - 25
2. Adults: from 26 - 35
3. Middle Aged Adults: 36-45
4. Elder :46 and above

4.1.2 Education Column

- Categorizing the values in education column in 3 different buckets:

1. Primary Education: upto 12
2. Secondary Education: 13 to 15
3. Higher Education: 16 and above

4.1.3 Income Column

- Categorizing the values in Income column in 4 different buckets:

1. Low Income - Upto 40,000
2. Moderate Income - 40,000 to 60,000
3. High Income - 60,000 to 80,000
4. Very High Income - Above 80,000

4.1.4 Miles column

- Categorizing the values in miles column in 4 different buckets:

1. Light Activity - Upto 50 miles
2. Moderate Activity - 51 to 100 miles
3. Active Lifestyle - 101 to 200 miles
4. Fitness Enthusiast - Above 200 miles

[30]:
```
#binning the age values into categories
bin_range1 = [17,25,35,45,float('inf')]
bin_labels1 = ['Young Adults', 'Adults', 'Middle Aged Adults', 'Elder']

df['age_group'] = pd.cut(df['Age'],bins = bin_range1,labels = bin_labels1)

#binning the education values into categories
```

```
bin_range2 = [0,12,15,float('inf')]
bin_labels2 = ['Primary Education', 'Secondary Education', 'Higher Education']

df['edu_group'] = pd.cut(df['Education'],bins = bin_range2,labels = bin_labels2)

#binning the income values into categories
bin_range3 = [0,40000,60000,80000,float('inf')]
bin_labels3 = ['Low Income','Moderate Income','High Income','Very High Income']

df['income_group'] = pd.cut(df['Income'],bins = bin_range3,labels = bin_labels3)

#binning the miles values into categories
bin_range4 = [0,50,100,200,float('inf')]
bin_labels4 = ['Light Activity', 'Moderate Activity', 'Active Lifestyle',␣
↪'Fitness Enthusiast ']

df['miles_group'] = pd.cut(df['Miles'],bins = bin_range4,labels = bin_labels4)
```

[31]:
```
df.head()
```

[31]:

```
  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

   Miles     age_group            edu_group income_group         miles_group
0    112  Young Adults  Secondary Education   Low Income    Active Lifestyle
1     75  Young Adults  Secondary Education   Low Income  Moderate Activity
2     66  Young Adults  Secondary Education   Low Income  Moderate Activity
3     85  Young Adults    Primary Education   Low Income  Moderate Activity
4     47  Young Adults  Secondary Education   Low Income     Light Activity
```

4.2 Probability of product purchase w.r.t. gender

[26]:
```
pd.crosstab(index =df['Product'],columns = df['Gender'],margins =␣
↪True,normalize = True ).round(2)
```

[26]:

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

4.2.1 Insights

1. The Probability of a treadmill being purchased by a female is 42%.
   - The conditional probability of purchasing the treadmill model given that the customer is female is:
     - For Treadmill model KP281 - 22%
     - For Treadmill model KP481 - 16%
     - For Treadmill model KP781 - 4%
2. The Probability of a treadmill being purchased by a male is 58%.
   - The conditional probability of purchasing the treadmill model given that the customer is male is -
     - For Treadmill model KP281 - 22%
     - For Treadmill model KP481 - 17%
     - For Treadmill model KP781 - 18%

4.3 Probability of product purchase w.r.t. Age

[32]:
```
pd.crosstab(index =df['Product'],columns = df['age_group'],margins =␣
 ↪True,normalize = True ).round(2)
```

[32]:

age_group / Product	Young Adults	Adults	Middle Aged Adults	Elder	All
KP281	0.19	0.18	0.06	0.02	0.44
KP481	0.16	0.13	0.04	0.01	0.33
KP781	0.09	0.09	0.02	0.01	0.22
All	0.44	0.41	0.12	0.03	1.00

4.3.1 Insights

1. The Probability of a treadmill being purchased by a Young Adult(18-25) is 44%.
   - The conditional probability of purchasing the treadmill model given that the customer is Young Adult is
     - For Treadmill model KP281 - 19%
     - For Treadmill model KP481 - 16%
     - For Treadmill model KP781 - 9%
2. The Probability of a treadmill being purchased by a Adult(26-35) is 41%.
   - The conditional probability of purchasing the treadmill model given that the customer is Adult is -
     - For Treadmill model KP281 - 18%
     - For Treadmill model KP481 - 13%

- For Treadmill model KP781 - 9%

3. The Probability of a treadmill being purchased by a Middle Aged(36-45) is 12%.
4. The Probability of a treadmill being purchased by a Elder(Above 45) is only 3%.

4.4 Probability of product purchase w.r.t. Education level

[33]:
```
pd.crosstab(index =df['Product'],columns = df['edu_group'],margins =␣
↪True,normalize = True ).round(2)
```

[33]:

edu_group Product	Primary Education	Secondary Education	Higher Education	All
KP281	0.01	0.21	0.23	0.44
KP481	0.01	0.14	0.18	0.33
KP781	0.00	0.01	0.21	0.22
All	0.02	0.36	0.62	1.00

4.4.1 Insights

1. The Probability of a treadmill being purchased by a customer with Higher Education(Above 15 Years) is 62%.
   - The conditional probability of purchasing the treadmill model given that the customer has Higher Education is
     - For Treadmill model KP281 - 23%
     - For Treadmill model KP481 - 18%
     - For Treadmill model KP781 - 21%
2. The Probability of a treadmill being purchased by a customer with Secondary Education(13-15 yrs) is 36%.
   - The conditional probability of purchasing the treadmill model given that the customer has Secondary Education is -
     - For Treadmill model KP281 - 21%
     - For Treadmill model KP481 - 14%
     - For Treadmill model KP781 - 1%
3. The Probability of a treadmill being purchased by a customer with Primary Education(0 to 12 yrs) is only 2%.

4.5 Probability of product purchase w.r.t. Income

[34]:
```
pd.crosstab(index =df['Product'],columns = df['income_group'],margins =␣
↪True,normalize = True ).round(2)
```

[34]:

income_group	Low Income	Moderate Income	High Income	Very High Income	All
Product					
KP281	0.13	0.28	0.03	0.00	0.44
KP481	0.05	0.24	0.04	0.00	0.33
KP781	0.00	0.06	0.06	0.11	0.22
All	0.18	0.59	0.13	0.11	1.00

4.5.1 Insights

1. The Probability of a treadmill being purchased by a customer with Low Income(<40k) is 18%.
   - The conditional probability of purchasing the treadmill model given that the customer has Low Income is-
     - For Treadmill model KP281 - 13%
     - For Treadmill model KP481 - 5%
     - For Treadmill model KP781 - 0%
2. The Probability of a treadmill being purchased by a customer with Moderate Income(40k - 60k) is 59%.
   - The conditional probability of purchasing the treadmill model given that the customer has Moderate Income is -
     - For Treadmill model KP281 - 28%
     - For Treadmill model KP481 - 24%
     - For Treadmill model KP781 - 6%
3. The Probability of a treadmill being purchased by a customer with High Income(60k - 80k) is 13%
   - The conditional probability of purchasing the treadmill model given that the customer has High Income is -
     - For Treadmill model KP281 - 3%
     - For Treadmill model KP481 - 4%
     - For Treadmill model KP781 - 6%
4. The Probability of a treadmill being purchased by a customer with Very High Income(>80k) is 11%
   - The conditional probability of purchasing the treadmill model given that the customer has High Income is -
     - For Treadmill model KP281 - 0%
     - For Treadmill model KP481 - 0%
     - For Treadmill model KP781 - 11%

4.6 Probability of product purchase w.r.t. Marital Status

[35]:
```
pd.crosstab(index =df['Product'],columns = df['MaritalStatus'],margins =␣
↪True,normalize = True ).round(2)
```

[35]:
```
MaritalStatus  Partnered  Single   All
Product
KP281               0.27    0.18  0.44
KP481               0.20    0.13  0.33
KP781               0.13    0.09  0.22
All                 0.59    0.41  1.00
```

4.6.1 Insights

1. The Probability of a treadmill being purchased by a Married Customer is 59%.
   - The conditional probability of purchasing the treadmill model given that the customer is Married is
     - For Treadmill model KP281 - 27%
     - For Treadmill model KP481 - 20%
     - For Treadmill model KP781 - 13%
2. The Probability of a treadmill being purchased by a Unmarried Customer is 41%.
   - The conditional probability of purchasing the treadmill model given that the customer is Unmarried is -
     - For Treadmill model KP281 - 18%
     - For Treadmill model KP481 - 13%
     - For Treadmill model KP781 - 9%

4.7 Probability of product purchase w.r.t. Weekly Usage

[36]:
```
pd.crosstab(index =df['Product'],columns = df['Usage'],margins = True,normalize␣
↪= True ).round(2)
```

[36]:
```
Usage       2     3     4     5     6     7   All
Product
KP281    0.11  0.21  0.12  0.01  0.00  0.00  0.44
KP481    0.08  0.17  0.07  0.02  0.00  0.00  0.33
KP781    0.00  0.01  0.10  0.07  0.04  0.01  0.22
All      0.18  0.38  0.29  0.09  0.04  0.01  1.00
```

4.7.1 Insights

1. The Probability of a treadmill being purchased by a customer with Usage 3 per week is 38%.

- The conditional probability of purchasing the treadmill model given that the customer has Usage 3 per week is -
  - For Treadmill model KP281 - 21%
  - For Treadmill model KP481 - 17%
  - For Treadmill model KP781 - 1%

2. The Probability of a treadmill being purchased by a customer with Usage 4 per week is 29%.
   - The conditional probability of purchasing the treadmill model given that the customer has Usage 4 per week is -
     - For Treadmill model KP281 - 12%
     - For Treadmill model KP481 - 7%
     - For Treadmill model KP781 - 10%
3. The Probability of a treadmill being purchased by a customer with Usage 2 per week is 18%
   - The conditional probability of purchasing the treadmill model given that the customer has Usage 2 per week is -
     - For Treadmill model KP281 - 11%
     - For Treadmill model KP481 - 8%
     - For Treadmill model KP781 - 0%

4.8 Probability of product purchase w.r.t. Customer Fitness

[37]:
```
pd.crosstab(index =df['Product'],columns = df['Fitness'],margins =␣
 ↪True,normalize = True ).round(2)
```

[37]:

Fitness	1	2	3	4	5	All
Product						
KP281	0.01	0.08	0.30	0.05	0.01	0.44
KP481	0.01	0.07	0.22	0.04	0.00	0.33
KP781	0.00	0.00	0.02	0.04	0.16	0.22
All	0.01	0.14	0.54	0.13	0.17	1.00

4.8.1 Insights

1. The Probability of a treadmill being purchased by a customer with **‘Average(3) Fitness is 54%.
   - The conditional probability of purchasing the treadmill model given that the customer has Average Fitness‘** is -
     - For Treadmill model KP281 - 30%
     - For Treadmill model KP481 - 22%
     - For Treadmill model KP781 - 2%
2. The Probability of a treadmill being purchased by a customer with Fitness of 2,4,5 is almost 15%.
3. The Probability of a treadmill being purchased by a customer with very low(1) Fitness is only 1%.

4.9 Probability of product purchase w.r.t. weekly mileage

[38]:
```
pd.crosstab(index =df['Product'],columns = df['miles_group'],margins =␣
↪True,normalize = True ).round(2)
```

[38]:
```
miles_group  Light Activity  Moderate Activity  Active Lifestyle  \
Product
KP281                  0.07               0.28              0.10
KP481                  0.03               0.22              0.08
KP781                  0.00               0.04              0.15
All                    0.09               0.54              0.33

miles_group  Fitness Enthusiast   All
Product
KP281                      0.00  0.44
KP481                      0.01  0.33
KP781                      0.03  0.22
All                        0.03  1.00
```

4.9.1 Insights

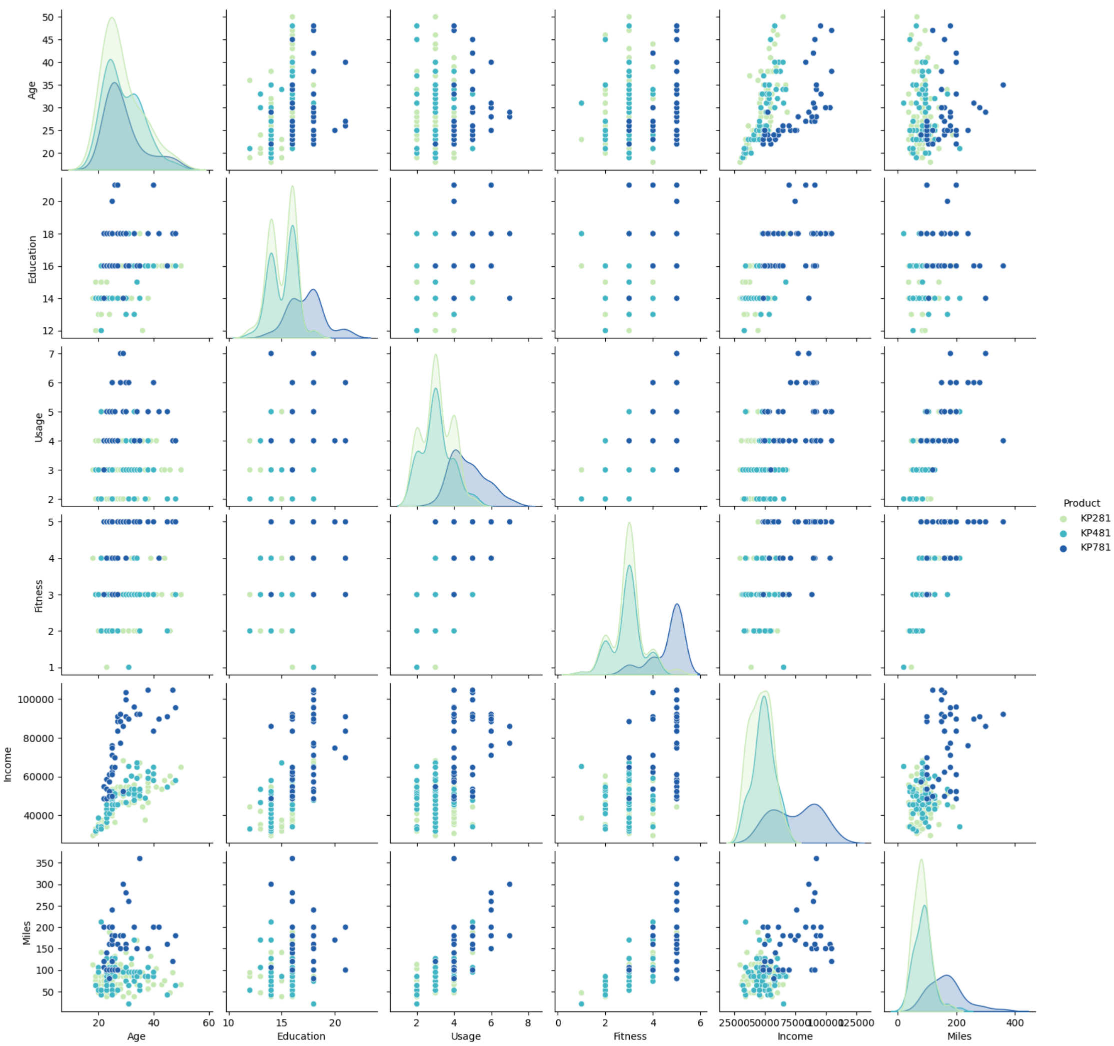
1. The Probability of a treadmill being purchased by a customer with lifestyle of Light Activity(0 to 50 miles/week) is 9%.
    - The conditional probability of purchasing the treadmill model given that the customer has Light Activity Lifestyle is -
        - For Treadmill model KP281 - 7%
        - For Treadmill model KP481 - 3%
        - For Treadmill model KP781 - 0%
2. The Probability of a treadmill being purchased by a customer with lifestyle of Moderate Activity(51 to 100 miles/week) is 54%.
    - The conditional probability of purchasing the treadmill model given that the customer with lifestyle of Moderate Activity is -
        - For Treadmill model KP281 - 28%
        - For Treadmill model KP481 - 22%
        - For Treadmill model KP781 - 4%
3. The Probability of a treadmill being purchased by a customer has Active Lifestyle(100 to 200 miles/week) is 33%.
    - The conditional probability of purchasing the treadmill model given that the customer has Active Lifestyle is -
        - For Treadmill model KP281 - 10%
        - For Treadmill model KP481 - 8%
        - For Treadmill model KP781 - 15%

4. The Probability of a treadmill being purchased by a customer who is Fitness Enthusiast(>200 miles/week) is 3% only

5 Checking the correlation among different factors

5.1 PairPlot

[40]:
```
sns.pairplot(df, hue ='Product', palette= 'YlGnBu')
plt.show()
```



5.1.1 Insights

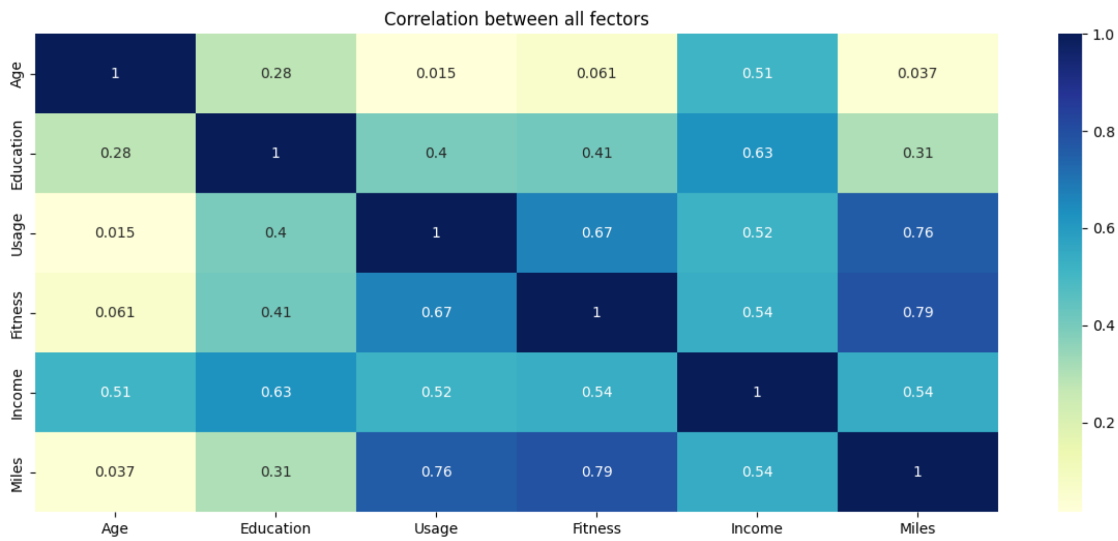
- From the pair plot we can see Age and Income are positively correlated and heatmap also suggests a strong correlation betwwen them.
- Eductaion and Income are highly correlated as its obvious. Eductation also has significatnt

correlation between Fitness rating and Usage of the treadmill.

- Usage is highly correlated with Fitness and Miles as more the usage more the fitness and mileage.

5.2 Heatmap

[47]:
```
import warnings
# Filtering out FutureWarnings
warnings.simplefilter(action='ignore', category=FutureWarning)

plt.figure(figsize=(15,6))
sns.heatmap(df.corr(),cmap="YlGnBu",annot=True)

plt.title("Correlation between all fectors")
plt.show()
```



5.2.1 Insights:

- Usage and Fitness Connection: Usage and fitness level exhibit strong positive correlations (0.76 and 0.67, respectively). This implies that individuals who use fitness equipment more frequently tend to have higher fitness levels.
- Income Influence: Income has notable associations with both education (0.63) and miles covered (0.54). Customers with higher incomes may have pursued more education and might prefer treadmills that offer longer mileage.
- Age's Limited Influence: Age shows relatively weak correlations with other variables, indicating that age alone may not strongly influence factors like income, fitness, or usage patterns.
- Education's Role: Education correlates positively with income (0.63) and, to a lesser extent,

with fitness and usage (0.41 and 0.4, respectively). This suggests that individuals with higher education levels may earn more and engage in fitness activities.

6 Customer Profiling

6.1 Based on above analysis

- Probability of purchase of KP281 = 44%
- Probability of purchase of KP481 = 33%
- Probability of purchase of KP781 = 22%

6.2 Customer Profile for KP281 Treadmill:

```
- Age of customer mainly between 18 to 35 years  with few between 35 to 50 years
- Education level of customer 13 years and above
- Annual Income of customer below USD 60,000
- Weekly Usage - 2 to 4 times
- Fitness Scale - 2 to 4
- Weekly Running Mileage - 50 to 100 miles
```

6.3 Customer Profile for KP481 Treadmill:

```
- Age of customer mainly between 18 to 35 years  with few between 35 to 50 years
- Education level of customer 13 years and above
- Annual Income of customer between USD 40,000 to USD 80,000
- Weekly Usage - 2 to 4 times
- Fitness Scale - 2 to 4
- Weekly Running Mileage - 50 to 200 miles
```

6.4 Customer Profile for KP781 Treadmill:

```
- Gender - Male
- Age of customer between 18 to 35 years
- Education level of customer 15 years and above
- Annual Income of customer USD 80,000 and above
- Weekly Usage - 4 to 7 times
- Fitness Scale - 3 to 5
- Weekly Running Mileage - 100 miles and above
```

7 Recommendations:

7.1 Targeted Marketing:

Given the insights regarding product preferences among different demographics (such as gender, income, and age), consider tailoring your marketing strategies. For instance, focus marketing efforts for KP281 towards females and lower-income customers, while emphasizing KP781 for higher-income and possibly male customers.

7.2 Product Development:

Use the data on product preferences and conditional probabilities to guide product development. If KP281 is popular among certain groups, consider enhancing its features or affordability for wider appeal. For KP781, explore ways to cater to higher-income customers' fitness needs.

7.3 Pricing Strategies:

Based on the correlations between income and product choices, you might adjust pricing strategies to align with customer income levels. Offering different pricing tiers or financing options could attract a broader customer base.

7.4 Education and Engagement:

Leverage the correlation between education and product preferences. Consider educational content or engagement strategies targeted at customers with higher education levels, potentially focusing on the benefits of specific products.

7.5 Customer Segmentation:

Use the provided data to create customer segments and design personalized marketing campaigns or product bundles for each segment. This can enhance customer engagement and increase sales.

7.6 Inventory Management:

Ensure that you have appropriate inventory levels for each product based on their popularity among different demographics. This can help optimize stock management and reduce carrying costs.