Business Case:Aerofit

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

About Case Study

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.

Defining Problem Statement

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.

Objective

Create a 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.

```
df=pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/
assets/000/001/125/original/aerofit_treadmill.csv?1639992749')
df
```

```
    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]
```

Dataset Characteristics

Dataset contains following columns

Product Purchased: KP281, KP481 and KP781, are the 3 different types of treadmills that are purchased by customers

- **Age** : In years, age of the customer who purchased
- **Gender**: Gender of the purchased customer
- **Education**: represented in years
- **Marital Status**: Single or partnered
- **Usage**: The average number of times the customer has planned to use the treadmill each week
- **Fitness**: Self rated fitness of the user rated from 1 (as poor shape) to 5 (as excellent shape)
- **Miles**: The average number of miles the customer expects to walk or run each week
- **Income**: Annual income of the user in Dollars $

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

       Product         Age  Gender   Education MaritalStatus
Usage  \
count      180  180.000000     180  180.000000           180
180.000000
unique       3         NaN       2         NaN             2
NaN
```

```
top       KP281          NaN   Male          NaN     Partnered
NaN
freq         80          NaN    104          NaN           107
NaN
mean        NaN    28.788889    NaN    15.572222           NaN
3.455556
std         NaN     6.943498    NaN     1.617055           NaN
1.084797
min         NaN    18.000000    NaN    12.000000           NaN
2.000000
25%         NaN    24.000000    NaN    14.000000           NaN
3.000000
50%         NaN    26.000000    NaN    16.000000           NaN
3.000000
75%         NaN    33.000000    NaN    16.000000           NaN
4.000000
max         NaN    50.000000    NaN    21.000000           NaN
7.000000

           Fitness         Income       Miles
count   180.000000     180.000000  180.000000
unique         NaN            NaN         NaN
top            NaN            NaN         NaN
freq           NaN            NaN         NaN
mean      3.311111   53719.577778  103.194444
std       0.958869   16506.684226   51.863605
min       1.000000   29562.000000   21.000000
25%       3.000000   44058.750000   66.000000
50%       3.000000   50596.500000   94.000000
75%       4.000000   58668.000000  114.750000
max       5.000000  104581.000000  360.000000
```

Observations:

- There are no missing values in the data.
- There are 3 unique products in the dataset.
- KP281 is the most frequent product.
- Minimum & Maximum age of the person is 18 & 50, mean is 28.79 and 75% of persons have age less than or equal to 33.
- Most of the people are having 16 years of education i.e. 75% of persons are having education <= 16 years.
- Out of 180 data points, 104's gender is Male and rest are the female.
- Standard deviation for Income & Miles is very high. These variables might have the outliers in it.

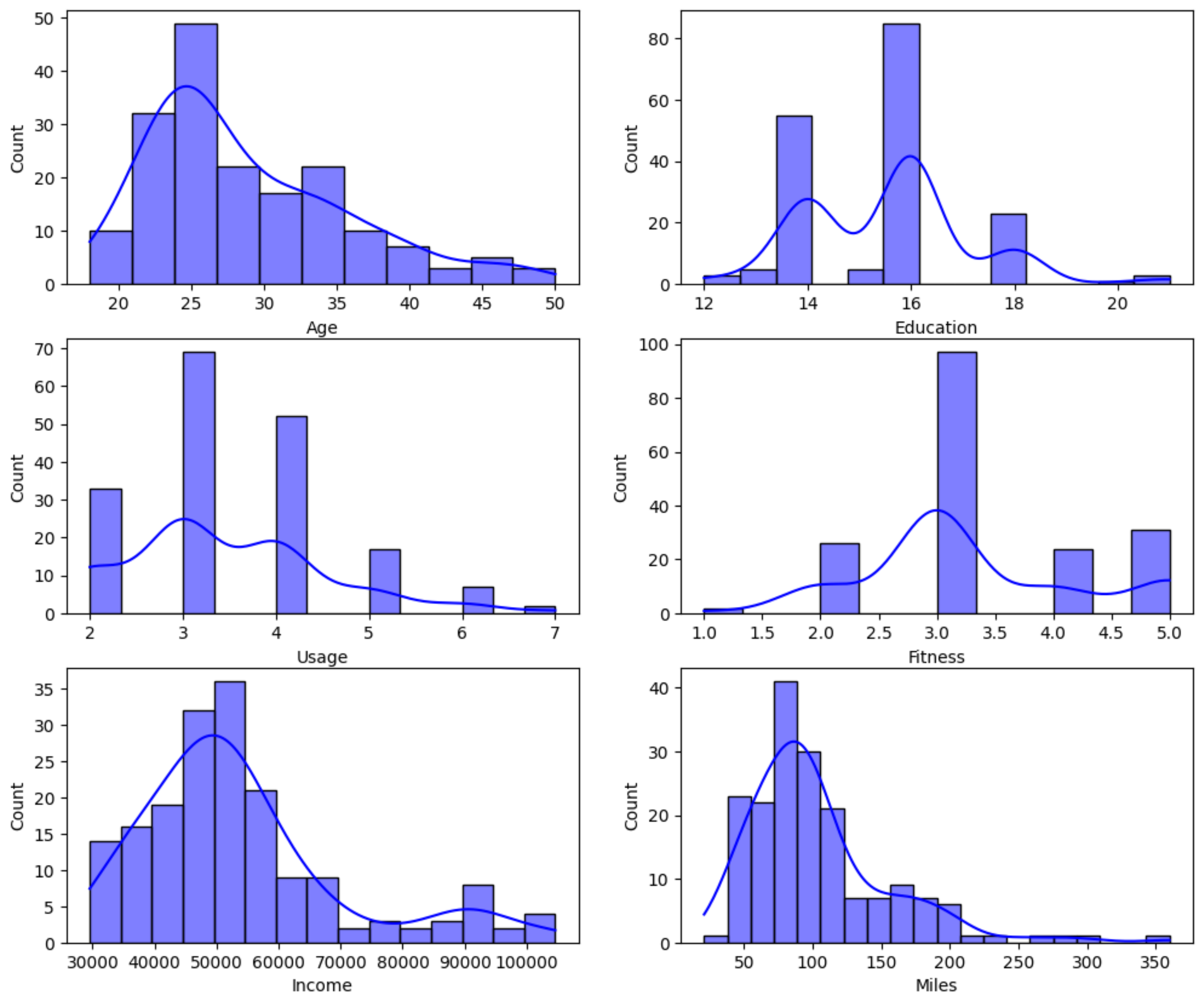
```
#computing number of rows

rows=len(df.axes[0])
rows

180

#computing number of columns
columns=len(df.axes[1])
columns

9

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['Product'].value_counts()

KP281    80
KP481    60
KP781    40
Name: Product, dtype: int64
```

There are 3 unique products available in the dataset.

#**Univariate Analysis** Understanding the distribution of the data for the quantitative attributes:

- Age
- Education
- Usage
- Fitness
- Income
- Miles

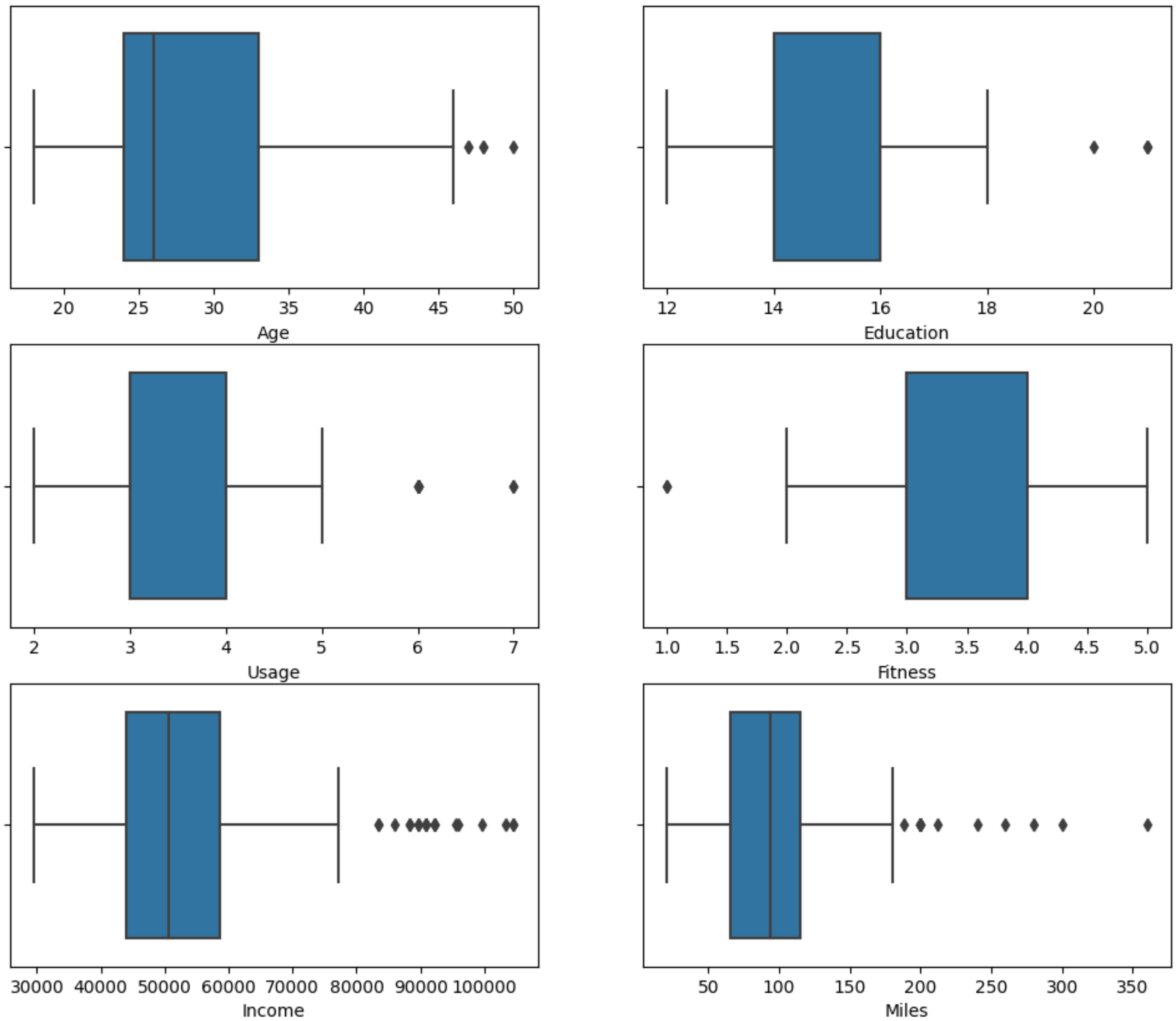
```
fig,axis= plt.subplots(3,2 , figsize=(12,10))
sns.histplot(data=df, x="Age" , kde=True , ax= axis[0,0],color="blue")
sns.histplot(data=df, x="Education" , kde=True , ax=
axis[0,1],color="blue")
sns.histplot(data=df, x="Usage" , kde=True , ax=
axis[1,0],color="blue")
sns.histplot(data=df, x="Fitness" , kde=True , ax=
axis[1,1],color="blue")
sns.histplot(data=df, x="Income" , kde=True , ax=
axis[2,0],color="blue")
sns.histplot(data=df, x="Miles" , kde=True , ax=
axis[2,1],color="blue")
plt.show()
```



Outliers detection using BoxPlots

```
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()
```



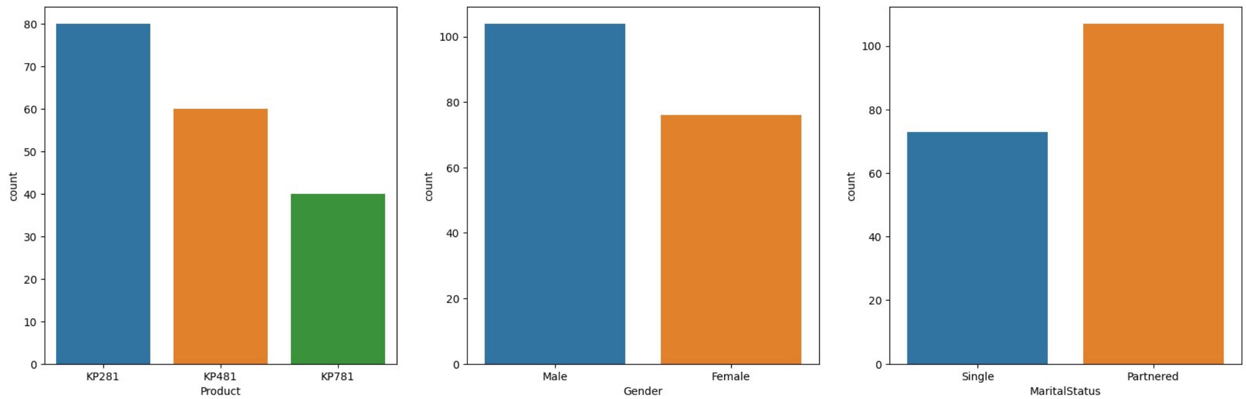
Observation Even from the boxplots it is quite clear that:

- Age, Education and Usage are having very few outliers.
- While Income and Miles are having more outliers.

Understanding the distribution of the data for the qualitative attributes:

- Product
- Gender
- MaritalStatus

```
fig, axs= plt.subplots(1,3 , figsize=(20,6))
sns.countplot(data=df,x='Product',ax=axs[0])
sns.countplot(data=df,x='Gender',ax=axs[1])
sns.countplot(data=df,x='MaritalStatus',ax=axs[2])
plt.show()
```



Obervations

- KP281 is the most frequent product.
- There are more Males in the data than Females.
- More Partnered persons are there in the data.

To be precise - normalized count for each variable is shown below

```
df1= df[["Product","Gender","MaritalStatus"]].melt()
df1.groupby(['variable','value'])[['value']].count()/len(df)
```

variable	value	value
Gender	Female	0.422222
	Male	0.577778
MaritalStatus	Partnered	0.594444
	Single	0.405556
Product	KP281	0.444444
	KP481	0.333333
	KP781	0.222222

Obervations

Product

- 44.44% of the customers have purchased KP2821 product.
- 33.33% of the customers have purchased KP481 product.
- 22.22% of the customers have purchased KP781 product.

Gender

- 57.78% of the customers are Male.

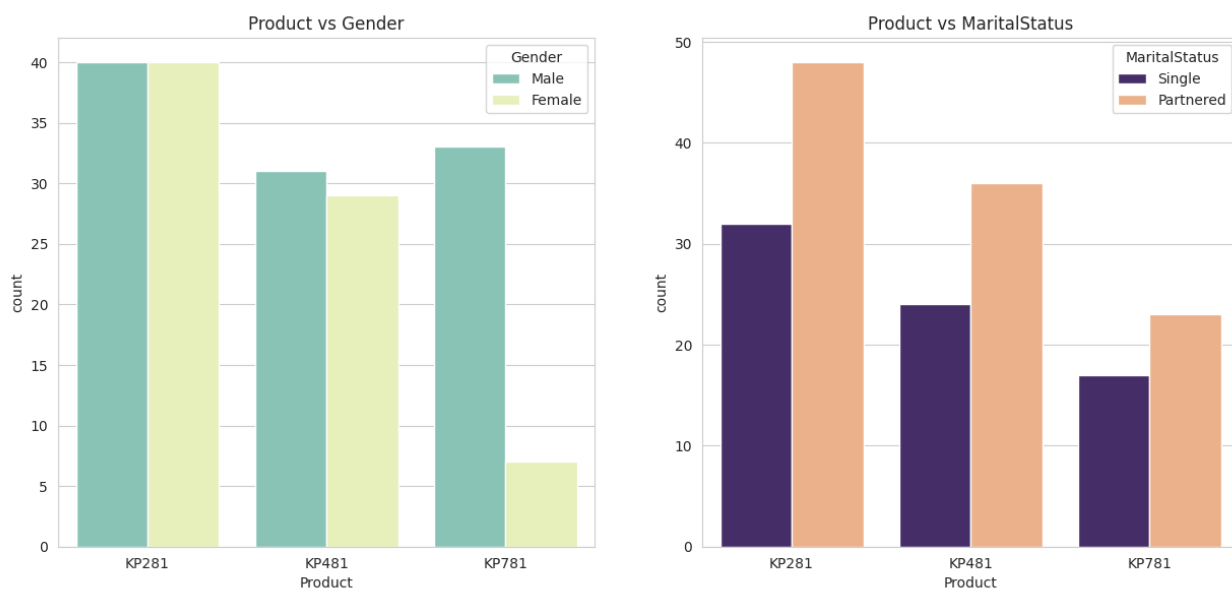
MaritalStatus

- 59.44% of the customers are Partnered.

Bivariate Analysis

Checking if features - Gender or MaritalStatus have any effect on the product purchased.

```
sns.set_style(style='whitegrid')
fig, axs= plt.subplots(1,2,figsize=(15,6.5))
sns.countplot(data=df,x="Product",hue="Gender",palette=["#7fcdbb","#ed
f8b1"],ax=axs[0]) ## #7fcdbb - This color is a shade of greenish-blue
#edf8b1 -  This color is a pale shade of yellow
axs[0].set_title("Product vs Gender")
sns.countplot(data=df,x="Product",hue="MaritalStatus",palette=['#43237
1',"#FAAE7B"],ax=axs[1])
##  #432371 - This color is a deep shade of purple
##  #FAAE7B - This color is a warm shade of light orange or peach
axs[1].set_title("Product vs MaritalStatus")
plt.show()
```



Obervations

Product vs Gender

- Equal number of males and females have purchased KP281 product and Almost same for the product KP481
- Most of the Male customers have purchased the KP781 product.

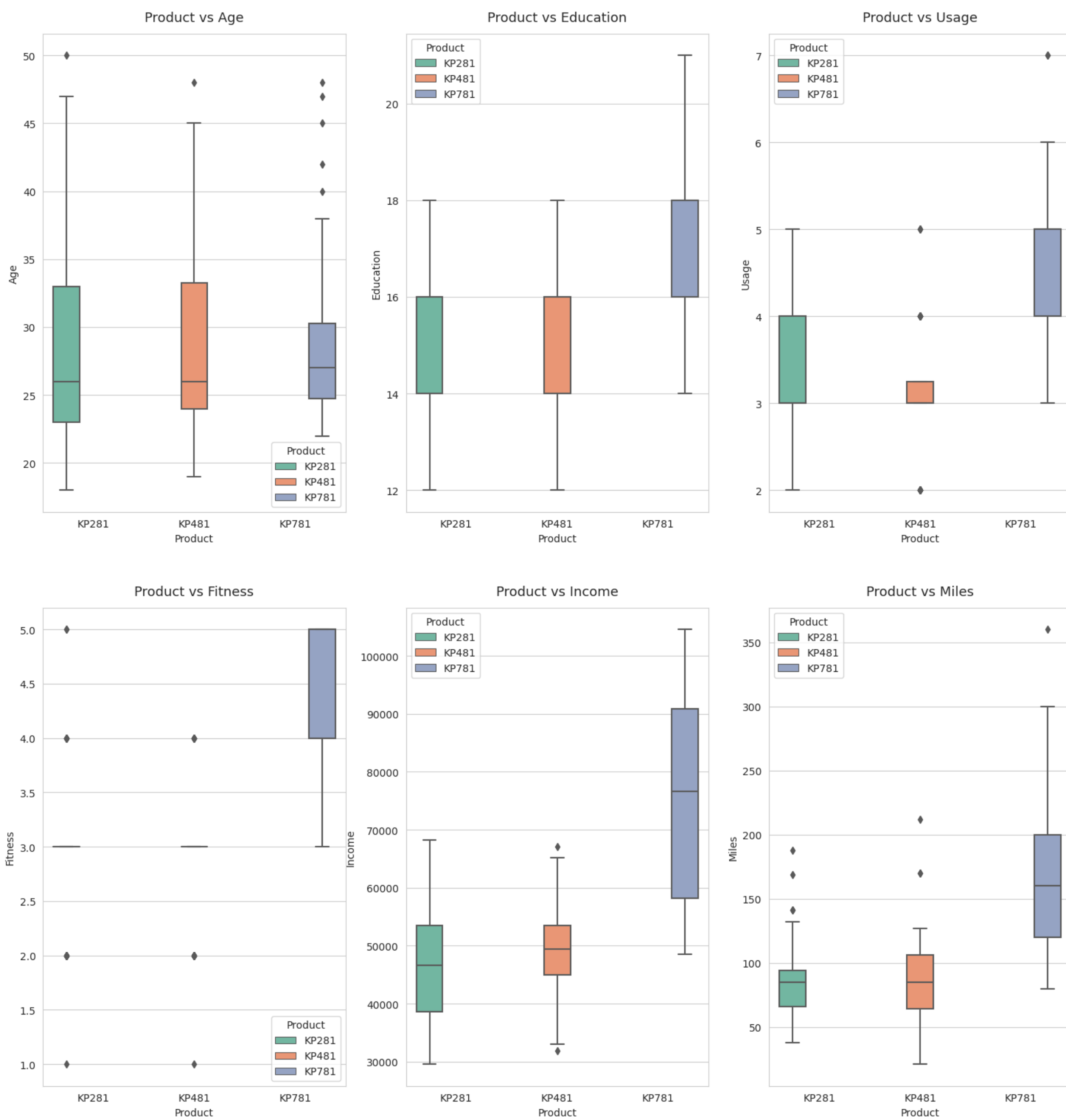
Product vs MaritalStatus

- Customer who is Partnered, is more likely to purchase the product.

Checking if following features have any effect on the product purchased:

- Age
- Education
- Usage
- Fitness
- Income
- Miles

```
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="Set2")
    axs[i,j].set_title(f"Product vs {var[count]}",pad=12,fontsize=13)
    count+=1
```



Observations

Product vs Age

- Customers purchasing products KP281 & KP481 are having same Age median value.
- Customers whose age lies between 25-30, are more likely to buy KP781 product

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 the KP781 product.
- While the other customers are likely to purchasing KP281 or KP481.

Product vs Fitness

- The more the customer is fit (fitness >= 3), higher the chances of the customer to purchase the KP781 product.

Product vs Income

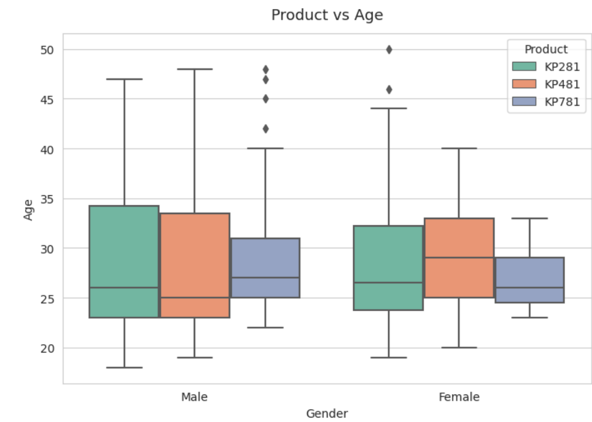
- Higher the Income of the customer (Income >= 60000), higher the chances of the customer to purchase the KP781 product.

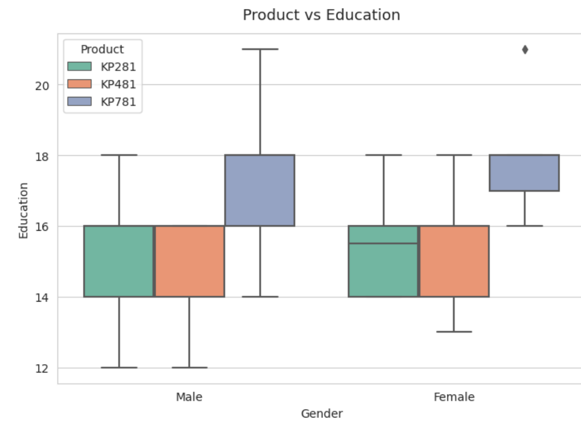
Product vs Miles *If the customer expects to walk/run greater than 120 Miles per week, it is more likely that the customer will buy KP781 product.

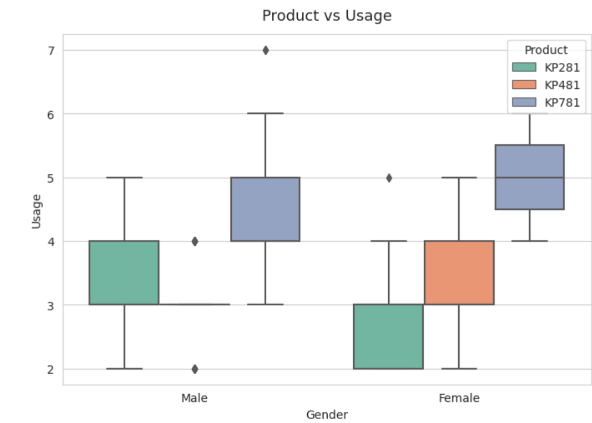
Multivariate Analysis

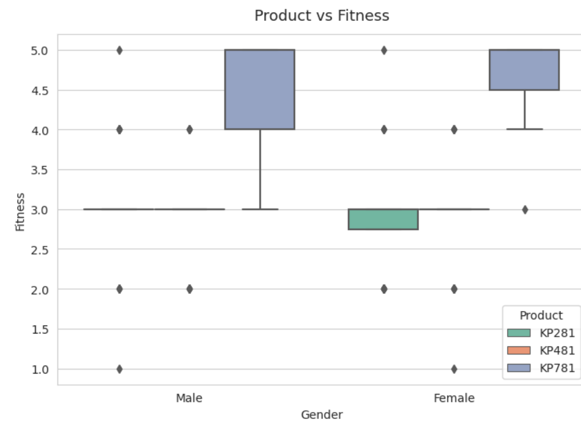
- Age
- Education
- Usage
- Fitness
- Income
- Miles

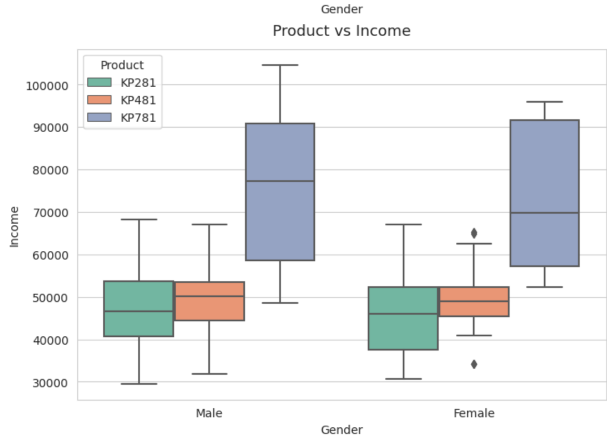
```
var= ['Age','Education','Usage','Fitness','Income','Miles']
sns.set_style("whitegrid")
fig,axs=plt.subplots(3,2,figsize=(18,12))
fig.subplots_adjust(top=1.3)
count=0
for i in range(3):
  for j in range(2):

sns.boxplot(data=df,x='Gender',y=var[count],hue='Product',ax=axs[i,j],
palette="Set2")
    axs[i,j].set_title(f"Product vs {var[count]}",pad=12,fontsize=13)
    count+=1
```

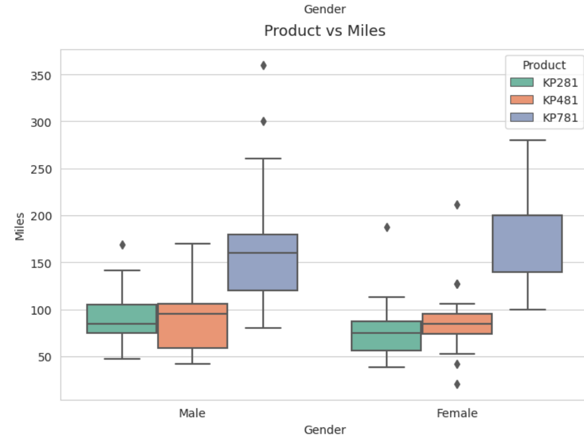












Observations

- In both Gender,customers whose education is greater then 16(Education>=16) prefer to buy KP781 product.
- In both Gender,customers who ue planning to use treadmill more then four times(Usage>=14) prefer to buy KP781 product.
- Females who are planning to use treadmill 3-4 times a week are more likely to buy KP481 product.
- In both Gender,custorner whose income is more than 55000 are more likely to buy KP781 product.

Computing Marginal & Conditional Probabilities

Marginal Probability

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

```
KP281    0.444444
KP481    0.333333
KP781    0.222222
Name: Product, dtype: float64
```

Conditional Probabilities

Probability of each product given gender

```
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
```

Probability of each product given MaritalStatus

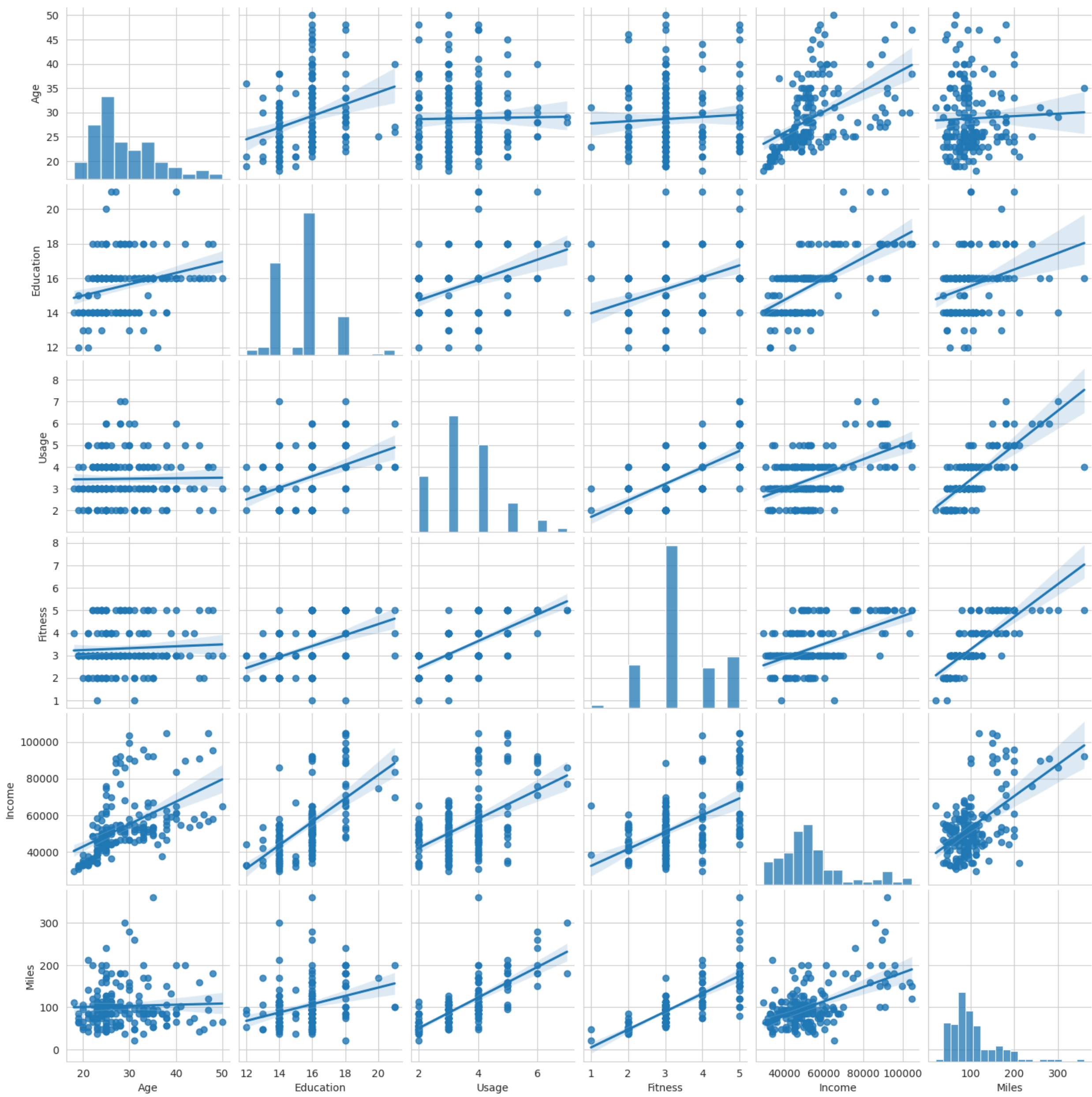
```python
def p_prod_given_MaritalStatus(status, print_marginal=False):
  if status!= "Single" and status!= "Partnered":
    return " invalid MaritalStatus value."
  df1= pd.crosstab(index=df['MaritalStatus'],columns=[df['Product']])
  p_781= df1['KP781'][status] / df1.loc[status].sum()
  p_481= df1['KP481'][status] / df1.loc[status].sum()
  p_281= df1['KP281'][status] / df1.loc[status].sum()
  if print_marginal:
    print(f"P(Single): {df1.loc['Single'].sum()/len(df):.2f}")
    print(f"P(Partnered): {df1.loc['Partnered'].sum()/len(df):.2f}\n")


  print(f"P(KP781/{status}):{p_781:.2f}")
  print(f"P(KP481/{status}):{p_481:.2f}")
  print(f"P(KP281/{status}):{p_281:.2f}\n")
p_prod_given_MaritalStatus('Single',True)
p_prod_given_MaritalStatus('Partnered')
```

```
P(Single): 0.41
P(Partnered): 0.59

P(KP781/Single):0.23
P(KP481/Single):0.33
P(KP281/Single):0.44

P(KP781/Partnered):0.21
P(KP481/Partnered):0.34
P(KP281/Partnered):0.45
```

#Coorelation between measurable quantities

```python
sns.pairplot(data = df, kind = 'reg')
plt.plot()
```

```
[]
```



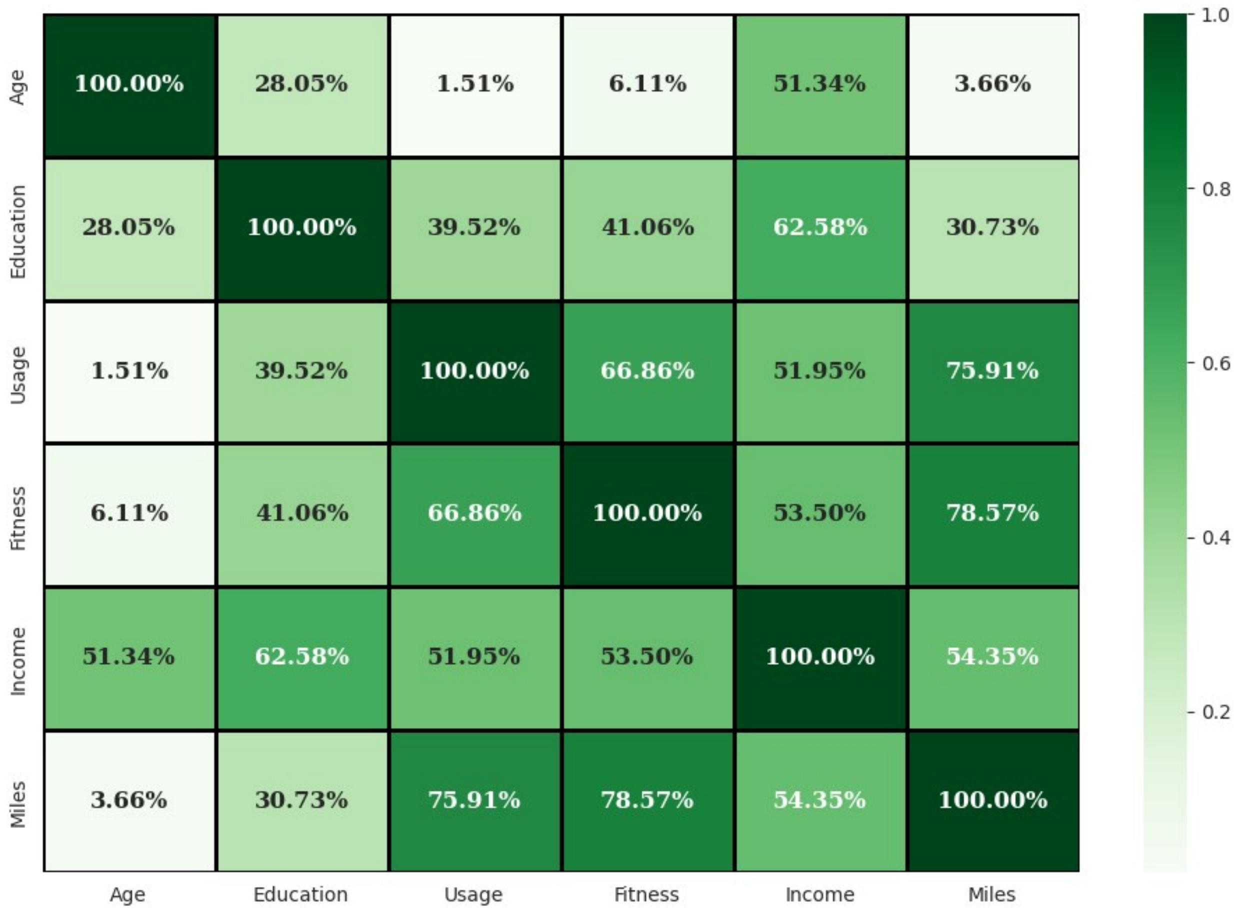

```
df_corr = df.corr()
df_corr

<ipython-input-34-0c96883f2151>:1: FutureWarning: The default value of
numeric_only in DataFrame.corr is deprecated. In a future version, it
will default to False. Select only valid columns or specify the value
of numeric_only to silence this warning.
  df_corr = df.corr()
```

	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
```

```python
plt.figure(figsize = (12, 8))
sns.heatmap(data = df_corr,
            annot = True,
            fmt = '.2%',
            cmap='Greens',
            linewidth = 2,
            linecolor = 'black',
            annot_kws = {'fontsize' : 'large',
                         'fontfamily' : 'serif',
                         'fontweight': 'bold'})
plt.plot()
```

```
[]
```



- The customer with high fitness scale is more likely to run or walk more miles.
- The customer who expects to use the treadmill more times in a week generally expects to walk or run more miles in the week.

- The customer who have a high fitness scale generally uses the treadmill more frequently in a week.

What is the product buying behaviors of both the genders ?

```
print(pd.crosstab(index = df['Product'], columns = df['Gender'],
margins = True))
print()
print('-' * 26)
print()
print("Product-wise normalization : ")
print(np.round(pd.crosstab(index = df['Product'], columns =
df['Gender'], normalize = 'index') * 100, 2))
print()
print('-' * 23)
print()
print("Gender-wise normalization : ")
print(np.round(pd.crosstab(index = df['Product'], columns =
df['Gender'], normalize = 'columns') * 100, 2))
```

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

--------------------------

Product-wise normalization :
Gender   Female   Male
Product
KP281     50.00  50.00
KP481     48.33  51.67
KP781     17.50  82.50

-----------------------

Gender-wise normalization :
Gender   Female   Male
Product
KP281     52.63  38.46
KP481     38.16  29.81
KP781      9.21  31.73
```

Customers who bought KP781, 82.5% of them are males rest are females.

Among all female customers, only 9.21 % buy KP781. Females mostly buy products KP281 or KP481.

Objective:

Customer Profiling for Each Product

Customer profiling based on the 3 product categories provided

KP281

Easily affordable entry level product, which is also the maximum selling product.

KP281 is the most popular product among the entry level customers.

This product is easily afforded by both Male and Female customers.

Average distance covered in this model is around 70 to 90 miles.

Product is used 3 to 4 times a week.

Most of the customer who have purchased the product have rated Average shape as the fitness rating.

Younger to Elder beginner level customers prefer this product.

Single female & Partnered male customers bought this product more than single male customers.

Income range between 39K to 53K have preferred this product.

KP481

This is an Intermediate level Product.

KP481 is the second most popular product among the customers.

Fitness Level of this product users varies from Bad to Average Shape depending on their usage.

Customers Prefer this product mostly to cover more miles than fitness.

Average distance covered in this product is from 70 to 130 miles per week.

More Female customers prefer this product than males.

Probability of Female customer buying KP481 is significantly higher than male.

KP481 product is specifically recommended for Female customers who are intermediate user.

Three different age groups prefer this product - Teen, Adult and middle aged.

Average Income of the customer who buys KP481 is 49K.

Average Usage of this product is 3 days per week.

More Partnered customers prefer this product.

There are slightly more male buyers of the KP481.

The distance travelled on the KP481 treadmill is roughly between 75 - 100 Miles. It is also the 2nd most distance travelled model.

The buyers of KP481 in Single & Partnered, Male & Female are same.

The age range of KP481 treadmill customers is roughly between 24-34 years.

KP781

Due to the High Price & being the advanced type, customer prefers less of this product.

Customers use this product mainly to cover more distance.

Customers who use this product have rated excelled shape as fitness rating.

Customer walk/run average 120 to 200 or more miles per week on his product.

Customers use 4 to 5 times a week at least.

Female Customers who are running average 180 miles (extensive exercise) , are using product KP781, which is higher than Male average using same product.

Probability of Male customer buying Product KP781(31.73%) is way more than female(9.21%).

Probability of a single person buying KP781 is higher than Married customers. So , KP781 is also recommended for people who are single and exercises more.

Middle aged to higher age customers tend to use this model to cover more distance.

Average Income of KP781 buyers are over 75K per annum

Partnered Female bought KP781 treadmill compared to Partnered Male.

Customers who have more experience with previous aerofit products tend to buy this product

This product is preferred by the customer where the correlation between Education and Income is High.

Recommendations:

- KP781 should be marketed as a Premium model and marketing it to high income groups and educational over 20 years market segments could result in more sales.
- Aerofit should conduct market research to determine if it can attracts customers under 40,000 to expand its customer base
- The KP781 is a premium model, so it is idealy suited for sporty people who have a high average weekly milege
- Female who prefer exercising equipments are very low here. Hence, we should run a marketing campaign on to encourage women to exercise more
- KP281 & KP481 treadmills are preferred by the customers whose annual income lies in the range of 39K - 53K Dollars. These models should promoted as budget treadmills.

- As KP781 provides more features and functionalities, the treadmill should be marketed for professionals and athletes.
- KP781 product should be promotted using influencers and other international atheletes.
- Research required for expanding market beyond 50 years of age considering health pros and cons.
- Provide customer support and recommend users to upgrade from lower versions to next level versions after consistent usages.
- KP781 can be recommended for Female customers who exercises extensively along with easy usage guidance since this type is advanced.
- Target the Age group above 40 years to recommend Product KP781.