In [1]:
```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
```

In [2]:
```python
df = pd.read_csv(r'C:\Users\suryawaa\OneDrive - TomTom\2022\Scaler\Aerofit_treadmill\aerofit_treadmill.csv')
```

In [3]:
```python
df.head(10)
```

Out[3]:

	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
5	KP281	20	Female	14	Partnered	3	3	32973	66
6	KP281	21	Female	14	Partnered	3	3	35247	75
7	KP281	21	Male	13	Single	3	3	32973	85
8	KP281	21	Male	15	Single	5	4	35247	141
9	KP281	21	Female	15	Partnered	2	3	37521	85

In [4]:
```python
df.shape # 180 rows with 9 columns
```

Out[4]:
```
(180, 9)
```

In [5]:
```python
df.info() # 3 out of 9 are object type and 6 out of 9 are of integer values.
```

```
<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 [6]:
```python
df.isnull().sum() # No Null values. Dataset is clear to perform further analysis.
```

Out[6]:
```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
Miles            0
dtype: int64
```

In [7]:
```
df.describe().round(2) # statistical summary
```
Out[7]:

	Age	Education	Usage	Fitness	Income	Miles
count	180.00	180.00	180.00	180.00	180.00	180.00
mean	28.79	15.57	3.46	3.31	53719.58	103.19
std	6.94	1.62	1.08	0.96	16506.68	51.86
min	18.00	12.00	2.00	1.00	29562.00	21.00
25%	24.00	14.00	3.00	3.00	44058.75	66.00
50%	26.00	16.00	3.00	3.00	50596.50	94.00
75%	33.00	16.00	4.00	4.00	58668.00	114.75
max	50.00	21.00	7.00	5.00	104581.00	360.00

In [8]:
```
df.describe(include=object)
```
Out[8]:

	Product	Gender	MaritalStatus
count	180	180	180
unique	3	2	2
top	KP281	Male	Partnered
freq	80	104	107

In [9]:
```
df.nunique() # count of unique values present in each columns
```
Out[9]:
```
Product           3
Age              32
Gender            2
Education         8
MaritalStatus     2
Usage             6
Fitness           5
Income           62
Miles            37
dtype: int64
```

In [10]:
```
bins = [29000, 50000, 70000, 110000]
categories = ['Low', 'Medium', 'High']
df['Income_category'] = pd.cut(df['Income'], bins=bins, labels=categories)
```

In [11]:
```
df.head()
```
Out[11]:

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

In [12]:
```
bins = [0, 25, 35, 100]
categories = ['<25', '25-35', '35+']
df['Age_Group'] = pd.cut(df['Age'], bins=bins, labels=categories)
```

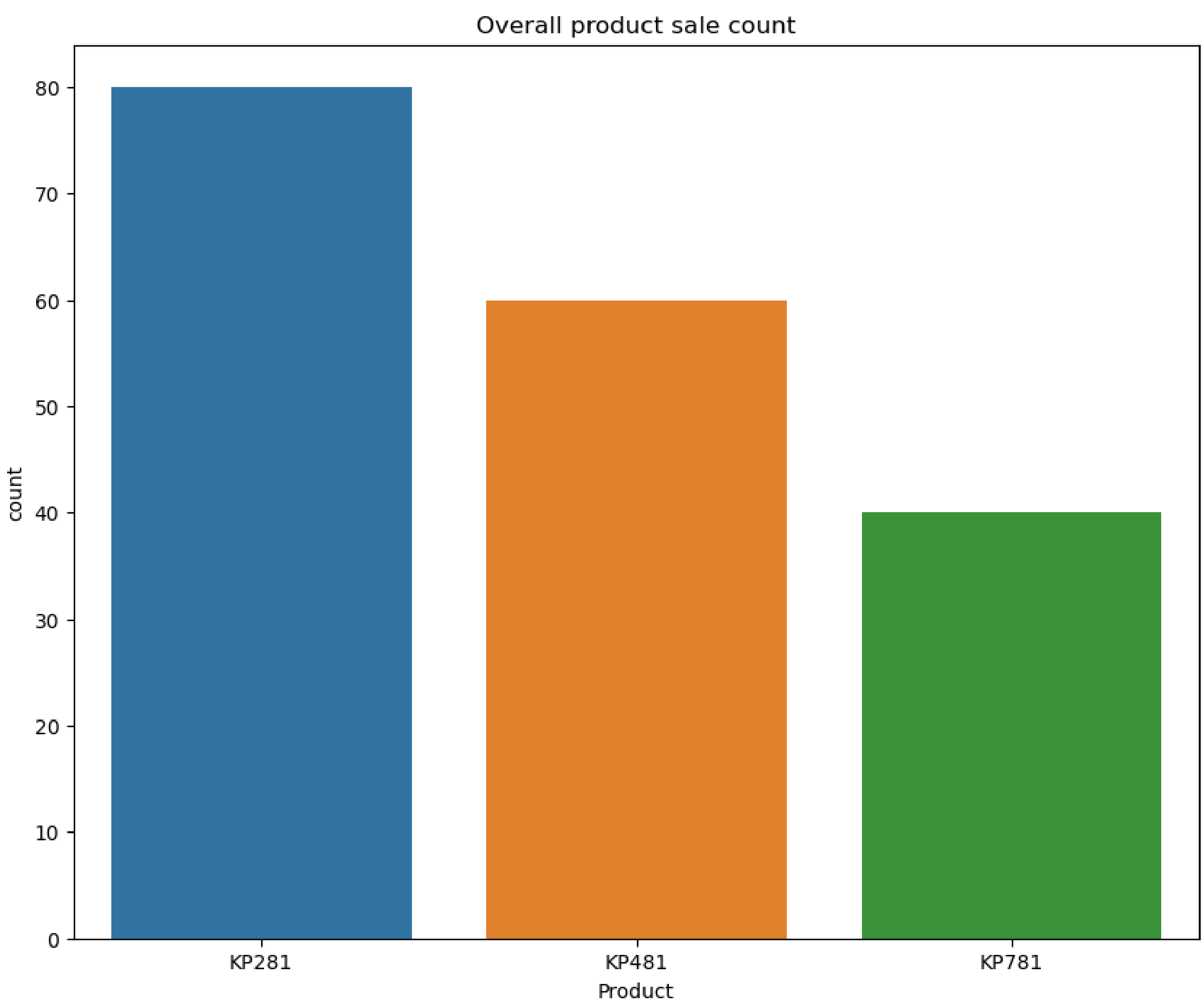
In [13]:
```
df.head()
```
Out[13]:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	Income_category	Age_Group
0	KP281	18	Male	14	Single	3	4	29562	112	Low	<25
1	KP281	19	Male	15	Single	2	3	31836	75	Low	<25
2	KP281	19	Female	14	Partnered	4	3	30699	66	Low	<25
3	KP281	19	Male	12	Single	3	3	32973	85	Low	<25
4	KP281	20	Male	13	Partnered	4	2	35247	47	Low	<25

Product type sale count

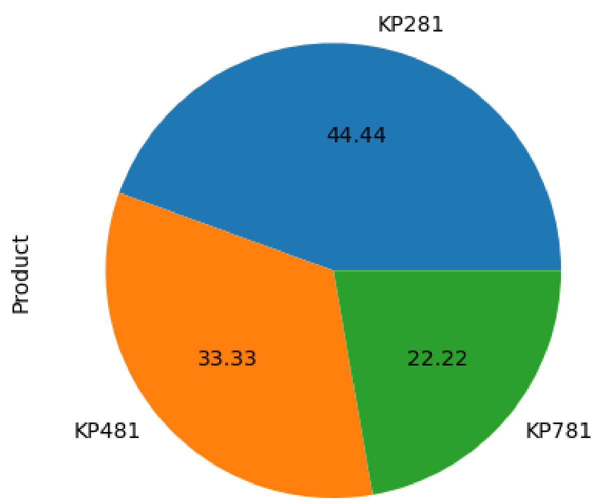
In [14]:
```
df['Product'].value_counts()
```
Out[14]:
```
KP281    80
KP481    60
KP781    40
Name: Product, dtype: int64
```

In [15]:
```
plt.figure(figsize=(10,8))
sns.countplot(data=df,x='Product')
plt.title("Overall product sale count")
plt.show()
```



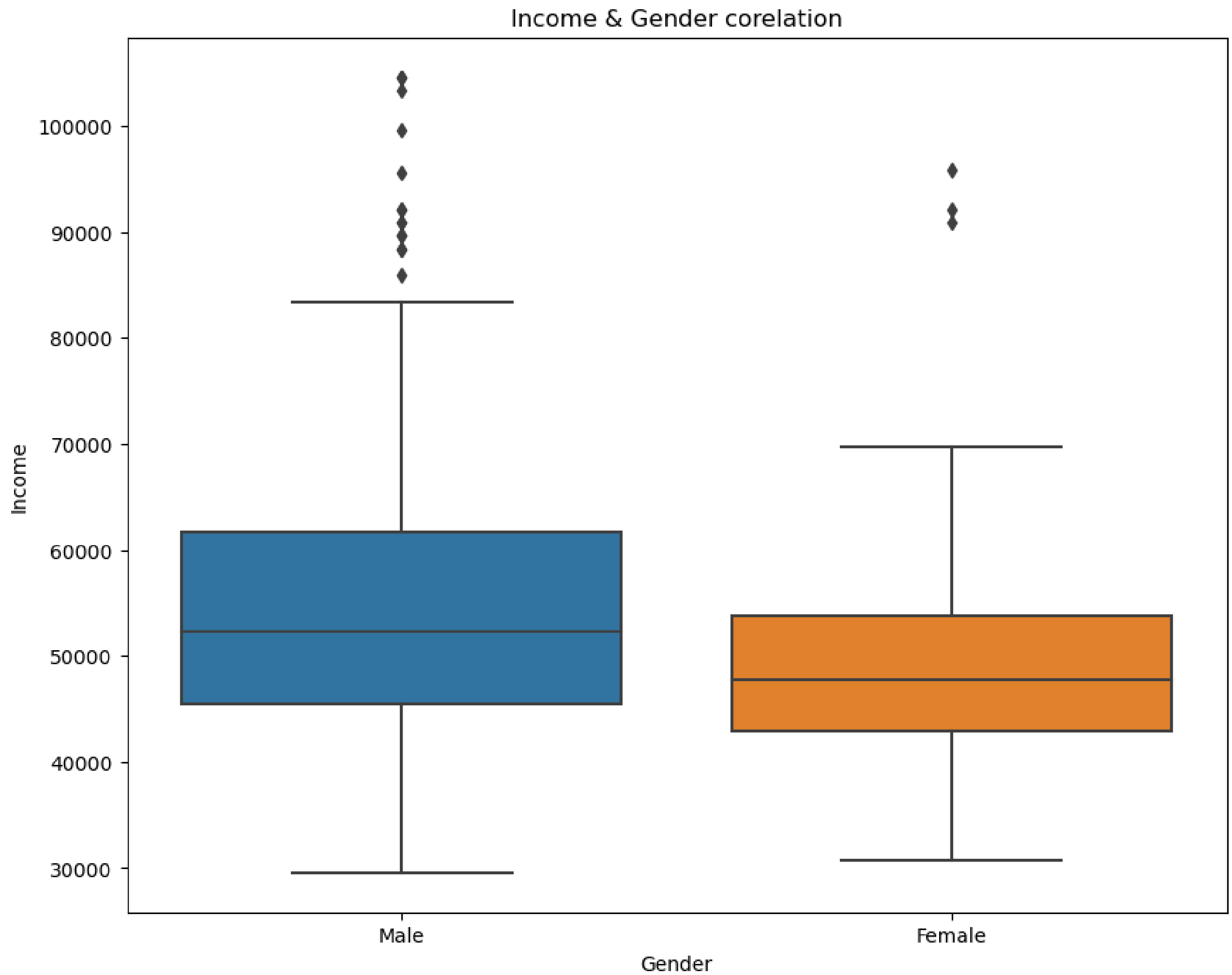
Relative Product type sale distribution overall

In [16]:
```
df['Product'].value_counts().plot(kind='pie',autopct="%.2f")
plt.show()
```



Genderwise,incomewise distribution of Customers

In [17]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x='Gender', y='Income', data=df)
plt.title("Income & Gender corelation")
plt.show()
```



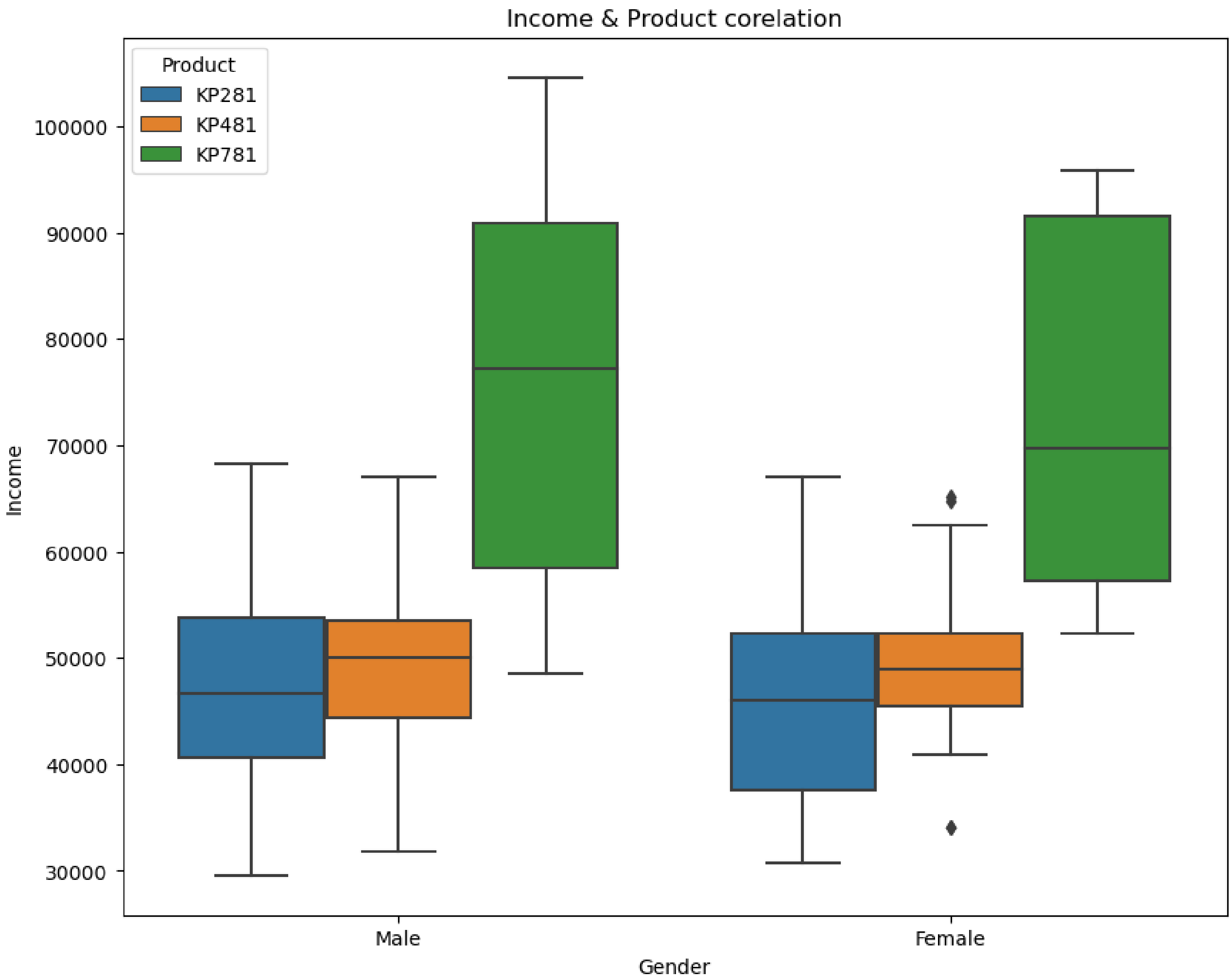
In [46]:
```
df.groupby(['Gender', 'Product'])['Income'].mean().unstack().round(2)
```

Out[46]:

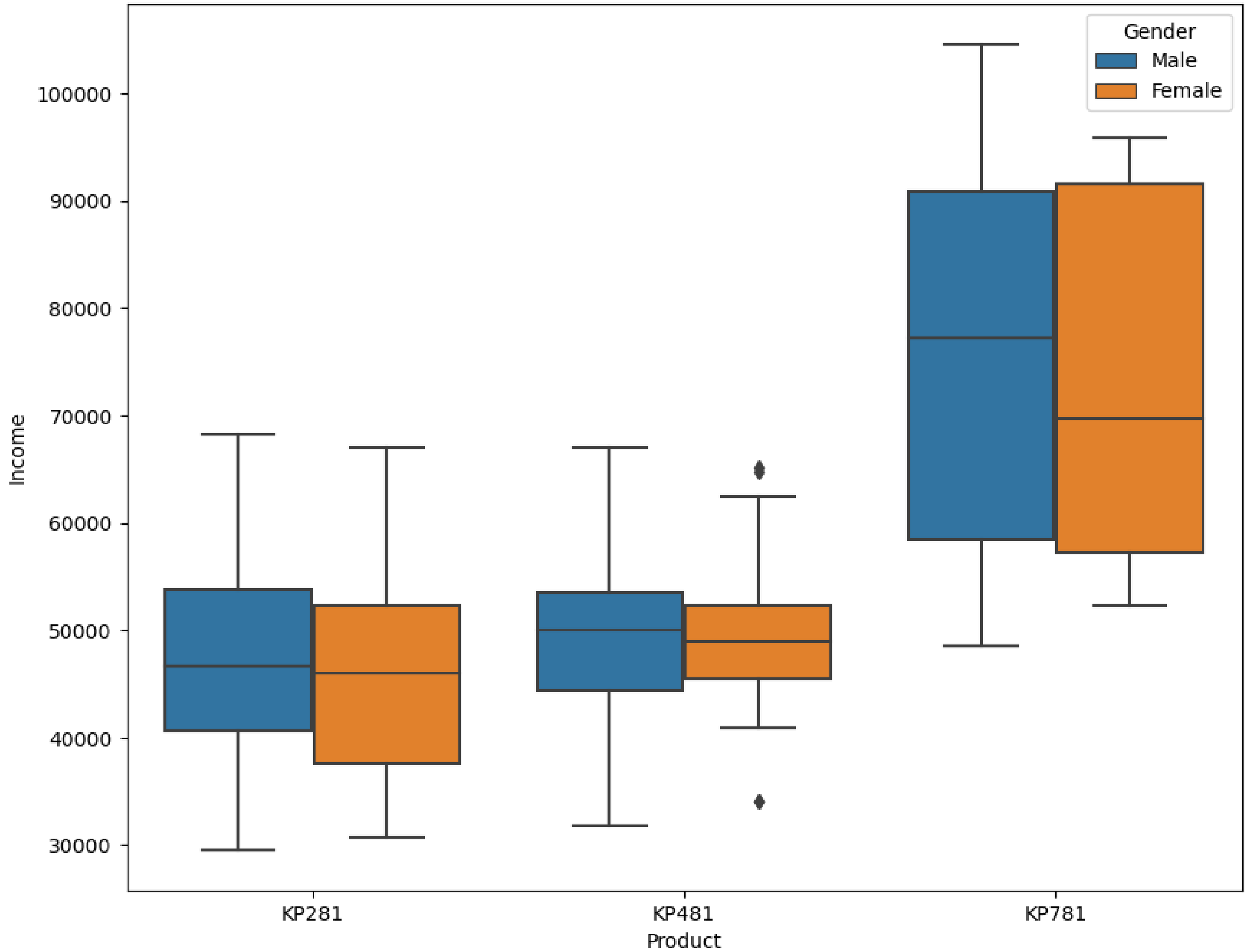
Product Gender	KP281	KP481	KP781
Female	46020.08	49336.45	73633.86
Male	46815.98	48634.26	75825.03

* Males have more income than the females
* Lesser the income more likely they will not opt for high-end product

In [18]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="Gender",y='Income',hue='Product',data=df)
plt.title("Income & Product corelation")
plt.show()
```

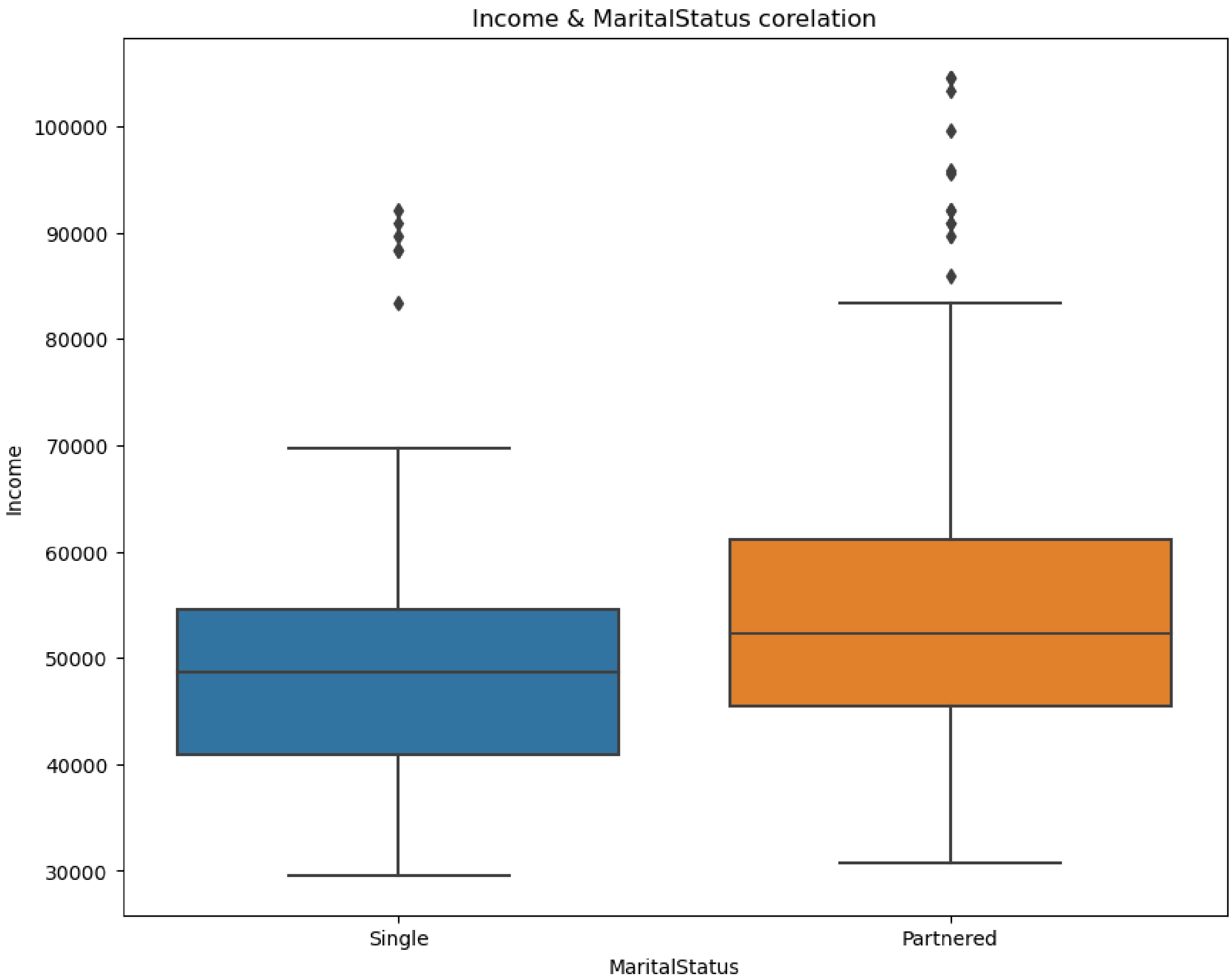


In [19]:
```
plt.figure(figsize=(10,8))
sns.boxplot (y='Income', x='Product', hue='Gender', data=df)
plt.show()
```



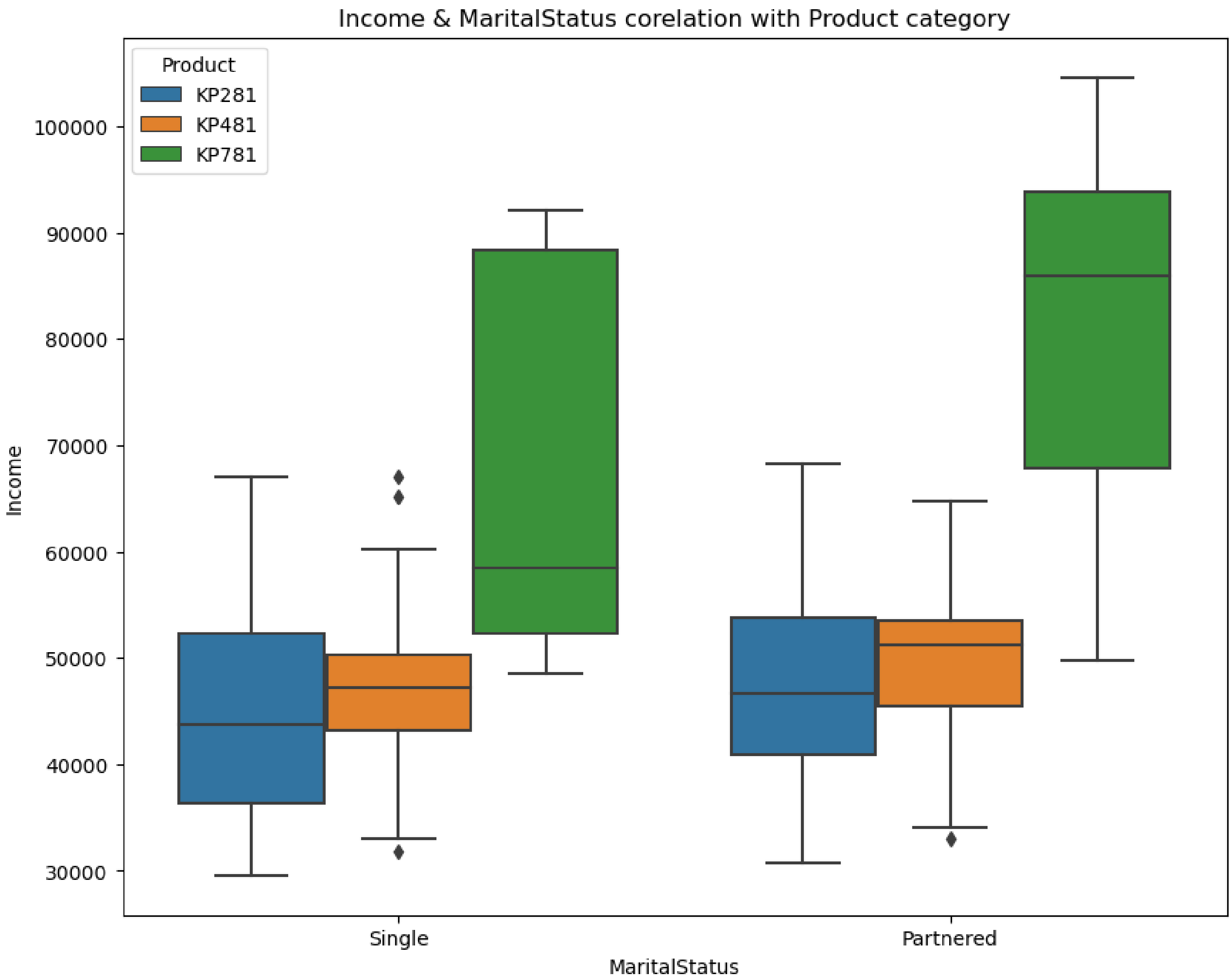
* Irrespective of gender, people with more income buy high end product

In [20]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x='MaritalStatus', y='Income', data=df)
plt.title("Income & MaritalStatus corelation")
plt.show()
```



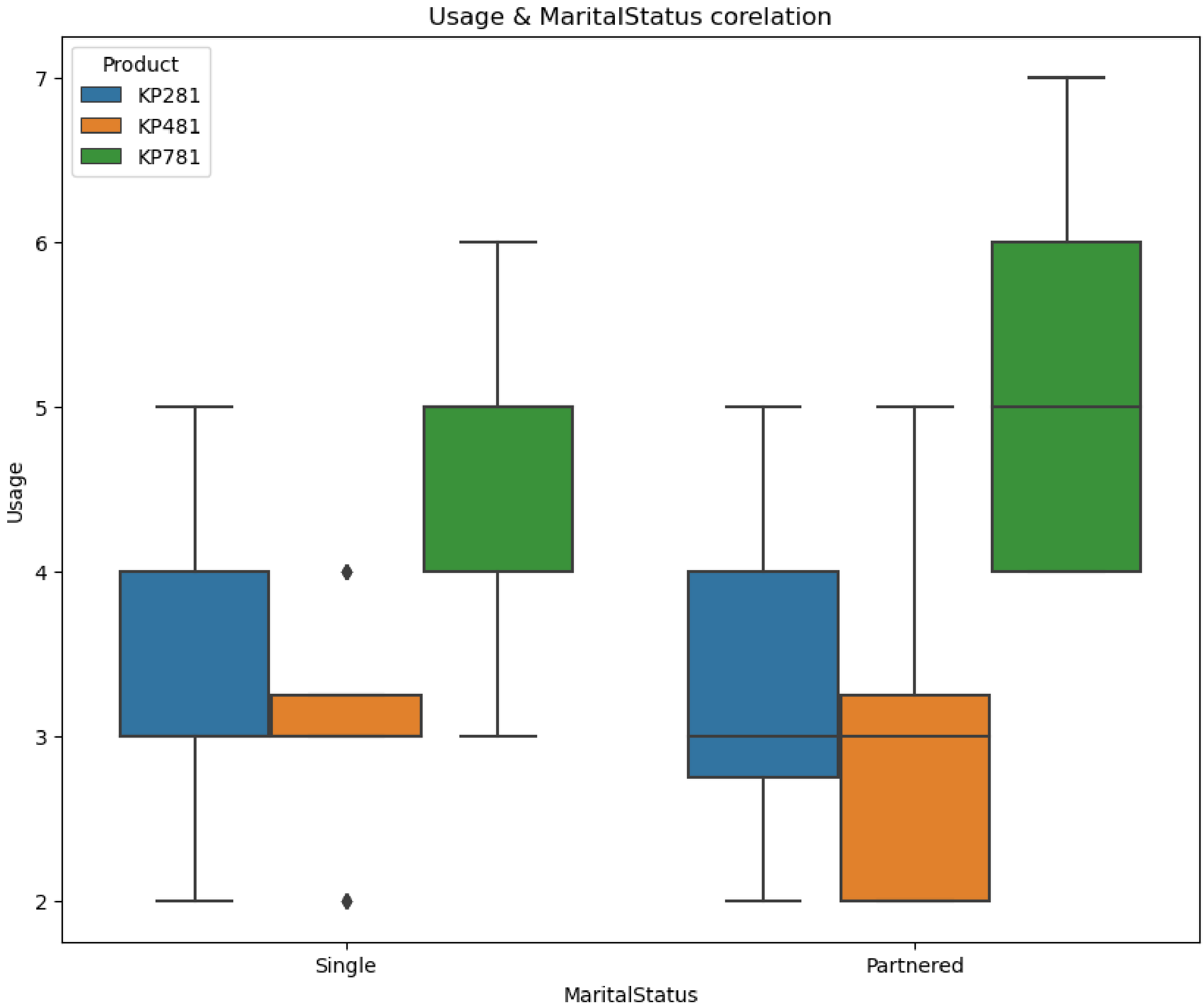
* Married people have slightly higher range of income than single

In [21]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="MaritalStatus",y='Income',hue='Product',data=df)
plt.title("Income & MaritalStatus corelation with Product category")
plt.show()
```



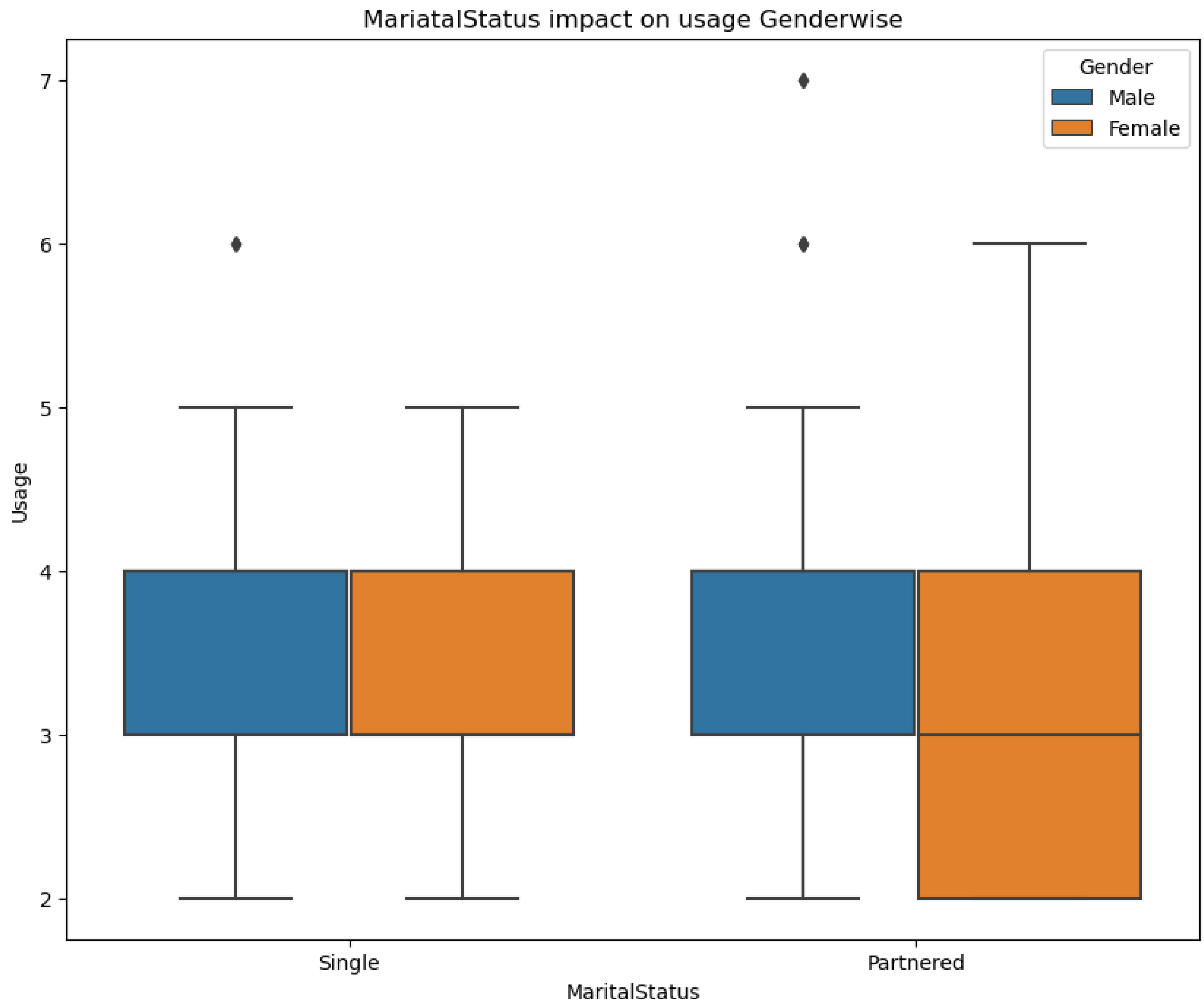
* Single customers with salary range more than 52K opt are more likely to opt for high end product
* while married customer tend to buy high end products when their income range is above 70K
* This difference may be due to the expenses of partnered person being higher than that of single person

In [22]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="MaritalStatus",y="Usage",hue='Product',data=df)
plt.title("Usage & MaritalStatus corelation")
plt.show()
```



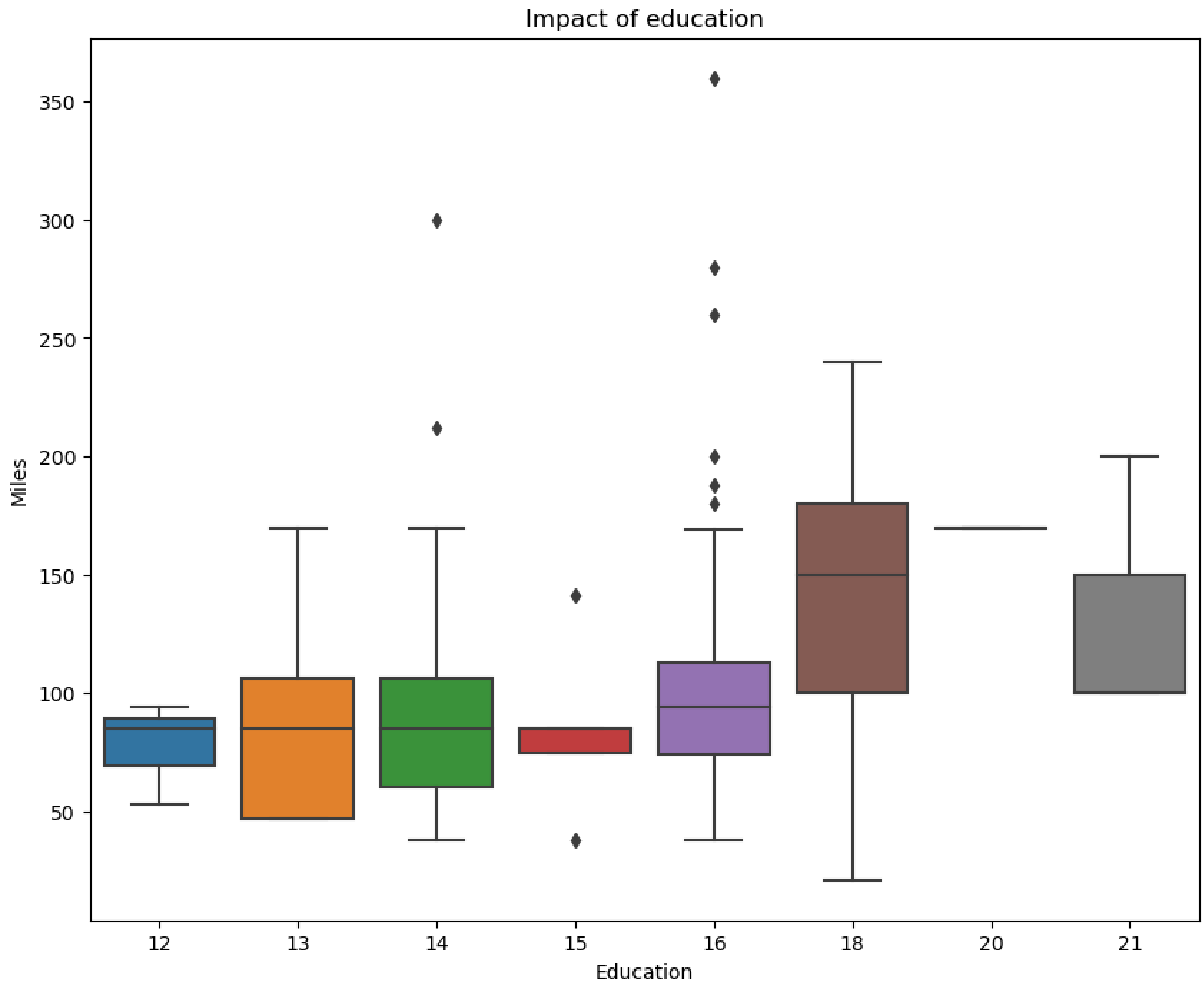
* Usage of highend product is higher in both single and partnered but comparatively more in partnered than single

In [23]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="MaritalStatus",y="Usage",hue='Gender',data=df)
plt.title("MariatalStatus impact on usage Genderwise")
plt.show()
```



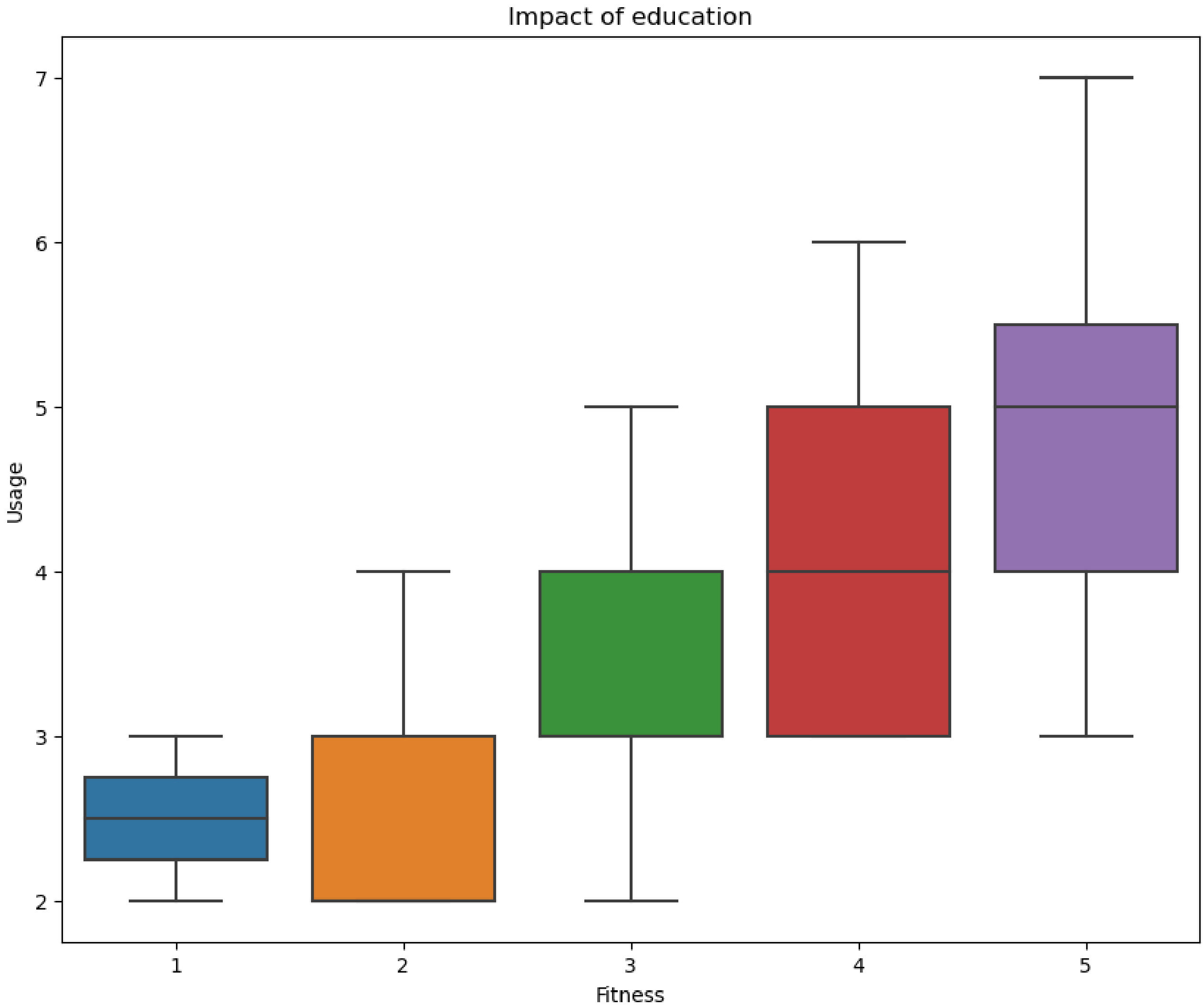
* Usage of products in males is quite maintained, but it drops in females if they are partnered

In [24]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="Education",y='Miles',data=df)
plt.title("Impact of education")
plt.show()
```



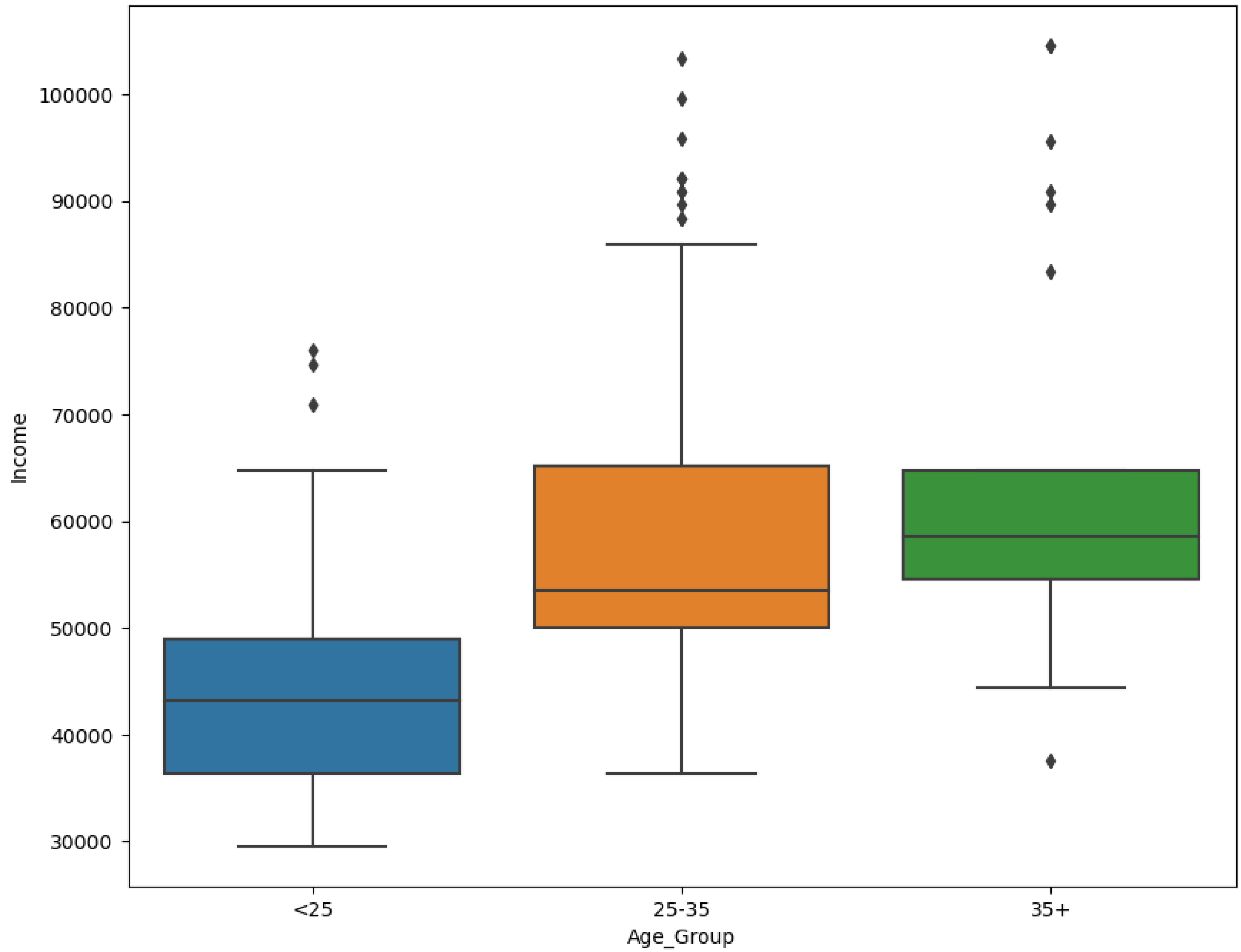

```
* The more educated people then more the miles they achieve. This may be due to educated people tend to
realize the importance of fitness
```

In [25]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="Fitness",y='Usage',data=df)
plt.title("Impact of education")
plt.show()
```



* As obvious the more the usage, more the fitness level

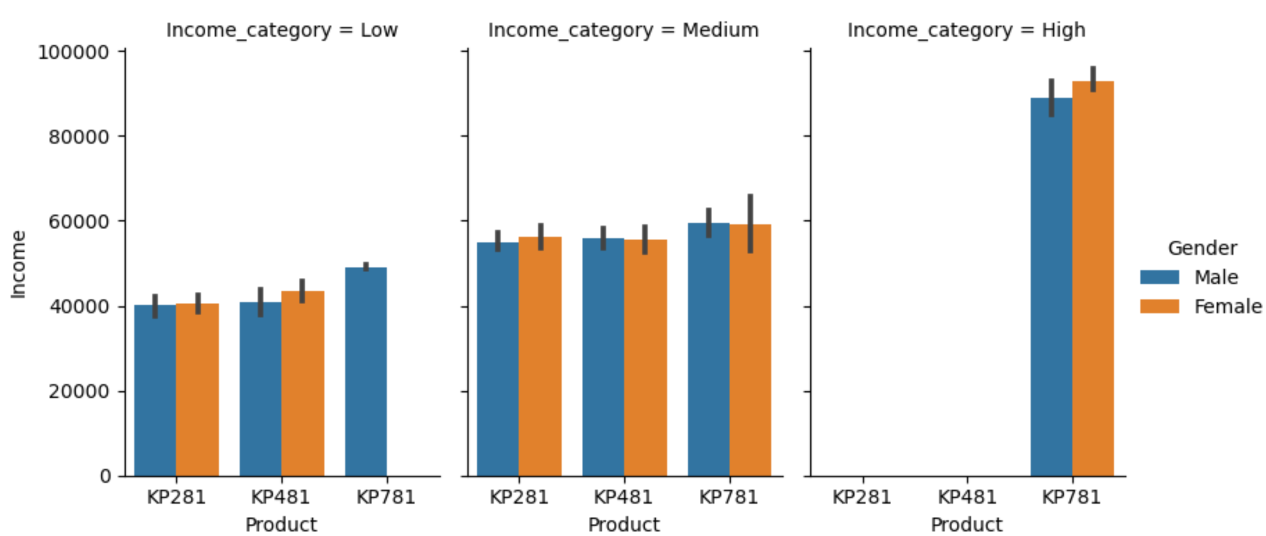
In [26]:
```
plt.figure(figsize=(10,8))
sns.boxplot(x="Age_Group",y='Income',data=df)
plt.show()
```



* Customer age more than 25 years is more likely to have higher income

In [27]:
```
plt.figure(figsize=(10,8))
sns.catplot(x="Product", y="Income", hue="Gender", col="Income_category",data=df, kind="bar", height=4, aspec
plt.show()
```

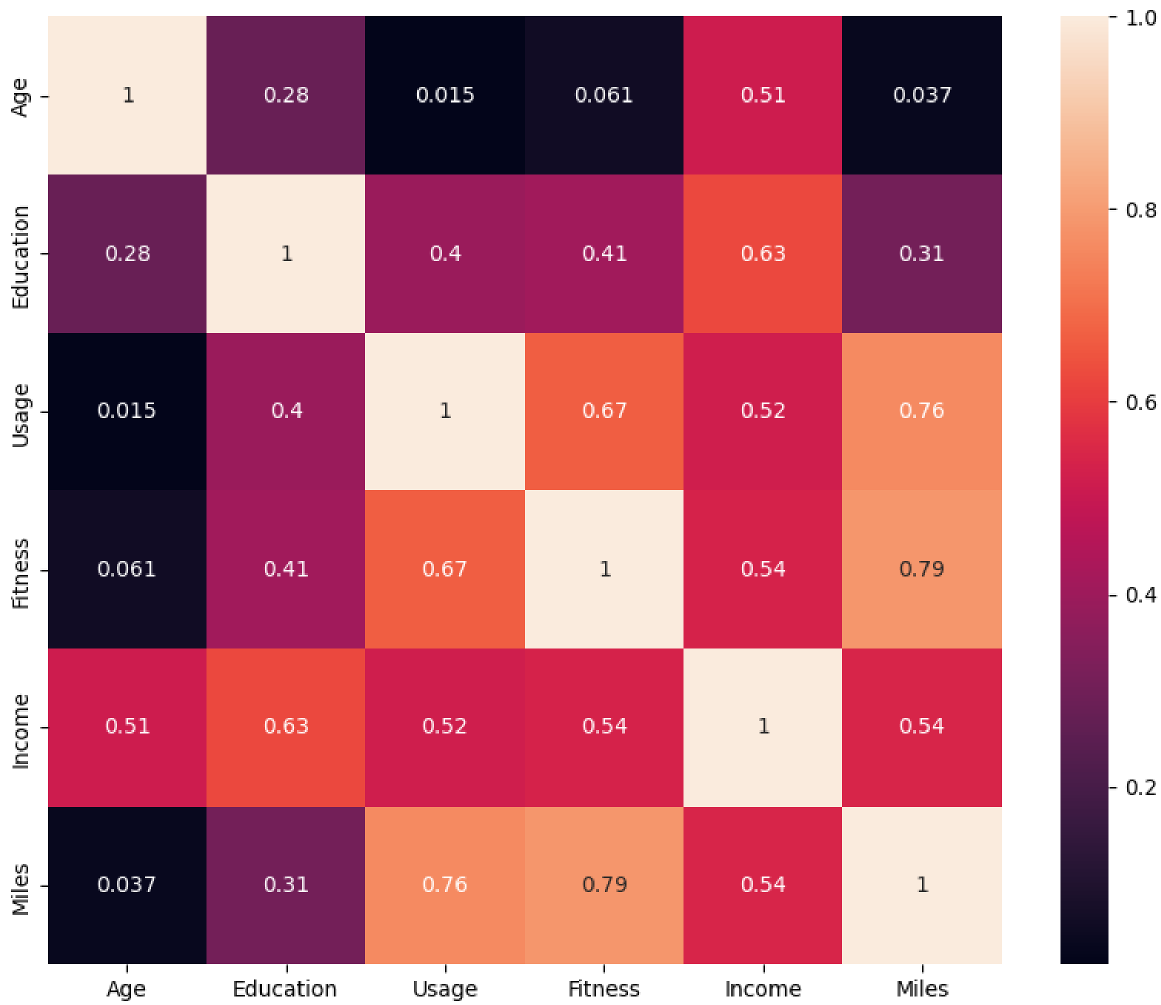
<Figure size 1000x800 with 0 Axes>



```
* Customer with low income category is more over having similar ratio of male and female opting for Product KP281 & KP481, however no females from this category opts for the high end product but males do
* Medium income group has overall same purchases for all 3 products
* High income category people only goes for high end product
```

Corelation Analysis with heatmap

In [28]:
```python
plt.figure(figsize=(10,8))
sns.heatmap(df.corr(),annot=True)
plt.show()
```




```
* Fitness and miles are highly corelated
* Usage and miles are strongly corelated
* Fitness and usage are also strongly corelated
* Age and Income is also somewhat corelated
* Fitness, usage and Age are not corelated as all age group has usage of the product up to some extent
```

In [29]:
```python
pd.crosstab(index=df['Gender'],columns= df['Product'])
```

Out[29]:

Product	KP281	KP481	KP781
Gender			
Female	40	29	7
Male	40	31	33

In [30]:
```
pd.crosstab(index=df['Gender'],columns= df['Product'],margins=True)
```

Out[30]:

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

Probabilites of Product purchase based on Gender

In [31]:
```
pd.crosstab(index=df['Gender'], columns=df['Product'], margins=True, normalize=True).round(4)
```

Out[31]:

Product Gender	KP281	KP481	KP781	All
Female	0.2222	0.1611	0.0389	0.4222
Male	0.2222	0.1722	0.1833	0.5778
All	0.4444	0.3333	0.2222	1.0000

```
* For Product KP281 probability of customer purchasing product is 0.22 in both genders and in total 0.44
* For Product KP481 probability of customer purchasing product is 0.16 in females and 0.17 in males, which
is very close to each others and in total 0.33
* For Product KP781 probability of customer purchasing product is 0.18 in males and very less as 0.03 in
females. In total it is 0.22
```

Probabilites of Product purchase based on Age_groups

In [32]:
```
pd.crosstab(index=df['Age_Group'], columns=df['Product'], margins=True, normalize=True).round(4)
```

Out[32]:

Product Age_Group	KP281	KP481	KP781	All
<25	0.1889	0.1556	0.0944	0.4389
25-35	0.1778	0.1333	0.0944	0.4056
35+	0.0778	0.0444	0.0333	0.1556
All	0.4444	0.3333	0.2222	1.0000

```
* For Product KP281 probability of customer purchasing product from group age <25 years is 0.18, group age
25-35 is 0.17, group age 35+ is 0.07 and in total for all ages is 0.44
* For Product KP481 probability of customer purchasing product from group age <25 years is 0.15, group age
25-35 is 0.13, group age 35+ is 0.04 and in total for all ages is 0.33
* For Product KP781 probability of customer purchasing product from group age <25 years is 0.09, group age
25-35 is 0.09, group age 35+ is 0.03 and in total for all ages is 0.22
```

Probabilites of Product purchase based on Income categories

In [33]:
```
pd.crosstab(index=df['Income_category'], columns=df['Product'], margins=True, normalize=True).round(4)
```

Out[33]:

Product Income_category	KP281	KP481	KP781	All
Low	0.2667	0.1667	0.0278	0.4611
Medium	0.1778	0.1667	0.0667	0.4111
High	0.0000	0.0000	0.1278	0.1278
All	0.4444	0.3333	0.2222	1.0000

```
* For Product KP281 probability of customer purchasing product from low income category is 0.26, medium
income category is 0.17, high income category is 0.0 and in total for is 0.44
```

* For Product KP481 probability of customer purchasing product from low income category is 0.16, medium income category is 0.16, high income category is 0.0 and in total for is 0.33
* For Product KP781 probability of customer purchasing product from low income category is 0.02, medium income category is 0.06, high income category is 0.12 and in total for is 0.22

Probabilities of the particular Gender bying the given products:

In [35]:
```python
pd.crosstab(index=df['Gender'], columns=df['Product'], margins=True, normalize='columns')
```

Out[35]:

Product	KP281	KP481	KP781	All
Gender				
Female	0.5	0.483333	0.175	0.422222
Male	0.5	0.516667	0.825	0.577778

* Probability of the contribution of males & females in buying KP281 is equal i.e. 0.5
* Probability of the contribution of males buying KP481 is 0.51 and in females it is 0.48 which can be considered nearly equal
* Probability of the contribution of males buying KP481 is 0.82 and in females it is 0.17 which greatly vary in the high end product. Males seems to buy high end product more likely than women.

Probabilities of the particular Income category bying the given products:

In [37]:
```python
pd.crosstab(index=df['Income_category'], columns=df['Product'], margins=True, normalize='columns')
```

Out[37]:

Product	KP281	KP481	KP781	All
Income_category				
Low	0.6	0.5	0.125	0.461111
Medium	0.4	0.5	0.300	0.411111
High	0.0	0.0	0.575	0.127778

* Probability of low income category opting to buy KP281 is 0.6, for medium income category it is 0.4 and people with high income range never opt for the product KP281 which is 0.0
* Probability of low income category opting to buy KP481 is 0.5, for medium income category also it is 0.5 and for the product KP481 also people with high income range do not opt to buy hence it is 0.0
* Probability of low income category opting to buy KP781 which is high end product is 0.12, for medium income category it is 0.3 and with high income range it is 0.57 which is greater value than other two income categories

Probabilities of the particular Age group bying the given products:

In [39]:
```python
pd.crosstab(index=df['Age_Group'], columns=df['Product'], margins=True, normalize='columns')
```

Out[39]:

Product	KP281	KP481	KP781	All
Age_Group				
<25	0.425	0.466667	0.425	0.438889
25-35	0.400	0.400000	0.425	0.405556
35+	0.175	0.133333	0.150	0.155556

* Probability of customer buying product KP281 whose age is below 25 is 0.45, whose age is between 25-35 is 0.4 & customer with age more than 35 is 0.17
* Probability of customer buying product KP481 whose age is below 25 is 0.46, whose age is between 25-35 is 0.4 & customer with age more than 35 is 0.13
* Probability of customer buying product KP281 whose age is below 25 is 0.42, whose age is between 25-35 is 0.42 & customer with age more than 35 is 0.15

Distribution of Gender over products:

In [40]:
```
pd.crosstab(index=df['Gender'], columns=df['Product'], normalize='index')*100
```

Out[40]:

Product	KP281	KP481	KP781
Gender			
Female	52.631579	38.157895	9.210526
Male	38.461538	29.807692	31.730769

```
* Portion of Females buying KP281 is 52.63% , KP481 is 38.15% & KP781 is 9.2%  out of total female population
* Portion of males buying KP281 is 38.46% , KP481 is 29.80% & KP781 is 31.7%  out of total male population
```

Distribution of Income category over products:

In [43]:
```
pd.crosstab(index=df['Income_category'], columns=df['Product'], normalize='index').round(4)*100
```

Out[43]:

Product	KP281	KP481	KP781
Income_category			
Low	57.83	36.14	6.02
Medium	43.24	40.54	16.22
High	0.00	0.00	100.00

```
* Portion of low income customer buying KP281 is 57.83% , KP481 is 36.14% & KP781 is 6.02%
* Portion of medium income customer buying KP281 is 43.24% , KP481 is 40.54% & KP781 is 16.22%
* Portion of high income customer buying KP281 is 0.0% , KP481 is 0.0% & KP781 is 100.00% .i.e. customer with high income range always goes for the high-end product
```

Distribution of age group over products:

In [44]:
```
pd.crosstab(index=df['Age_Group'], columns=df['Product'], normalize='index').round(4)*100
```

Out[44]:

Product	KP281	KP481	KP781
Age_Group			
<25	43.04	35.44	21.52
25-35	43.84	32.88	23.29
35+	50.00	28.57	21.43

```
* Portion of customer having age less than 25 years and buying KP281 is 43.04% , KP481 is 35.44% & KP781 is 21.52%
* Portion of customer having age between 25 to 35 years and buying KP281 is 43.84% , KP481 is 32.88% & KP781 is 23.29%
* Portion of customer having age more than 35 years and buying KP281 is 50.00% , KP481 is 28.57% & KP781 is 21.43%
```

Buisness Insights:

Observation

```
* Males have more income than the females
* Lesser the income more likely they will not opt for high-end product
* Irrespective of gender, people with more income buy high-end product
* Married people have slightly higher range of income than single
* Single customers with salary range more than 52K opt are more likely to opt for high-end product, while married customer tend to buy high end products when their income range is above 70K
* This difference may be due to the expenses of partnered person being higher than that of single person
* Usage of high-end product is higher in both single and partnered but comparatively more in partnered than single
* Usage of products in males is quite maintained, but it drops in females if they are partnered
* The more educated people then more the miles they achieve. This may be due to educated people tend to realize the importance of fitness
* As obvious the more the usage, more the fitness level
* Customer age more than 25 years is more likely to have higher income
```

* Customer with low-income category is more over having similar ratio of male and female opting for Product KP281 & KP481, however no females from this category opts for the high-end product but males do.
* Medium income group has overall same purchases for all 3 products
* High income category people only go for high-end product
* Fitness and miles are highly corelated
* Usage and miles are strongly corelated
* Fitness and usage are also strongly corelated
* Age and Income is also somewhat corelated
* Fitness, usage and Age are not corelated as all age group has usage of the product up to some extent

Probabilties and statistics on age group,income range and gender of customer with respect to the products

* For Product KP281 probability of customer purchasing product is 0.22 in both genders and in total 0.44
* For Product KP481 probability of customer purchasing product is 0.16 in females and 0.17 in males, which is very close to each other’s and in total 0.33
* For Product KP781 probability of customer purchasing product is 0.18 in males and very less as 0.03 in females. In total it is 0.22
* For Product KP281 probability of customer purchasing product from group age <25 years is 0.18, group age 25-35 is 0.17, group age 35+ is 0.07 and in total for all ages is 0.44
* For Product KP481 probability of customer purchasing product from group age <25 years is 0.15, group age 25-35 is 0.13, group age 35+ is 0.04 and in total for all ages is 0.33
* For Product KP781 probability of customer purchasing product from group age <25 years is 0.09, group age 25-35 is 0.09, group age 35+ is 0.03 and in total for all ages is 0.22
* For Product KP281 probability of customer purchasing product from low-income category is 0.26, medium income category is 0.17, high income category is 0.0 and in total for is 0.44
* For Product KP481 probability of customer purchasing product from low-income category is 0.16, medium income category is 0.16, high income category is 0.0 and in total for is 0.33
* For Product KP781 probability of customer purchasing product from low-income category is 0.02, medium income category is 0.06, high income category is 0.12 and in total for is 0.22
* Probability of the contribution of males & females in buying KP281 is equal i.e., 0.5
* Probability of the contribution of males buying KP481 is 0.51 and in females it is 0.48 which can be considered nearly equal
* Probability of the contribution of males buying KP481 is 0.82 and in females it is 0.17 which greatly vary in the high-end product. Males seems to buy high-end product more likely than women
* Probability of low-income category opting to buy KP281 is 0.6, for medium income category it is 0.4 and people with high income range never opt for the product KP281 which is 0.0
* Probability of low-income category opting to buy KP481 is 0.5, for medium income category also it is 0.5 and for the product KP481 also people with high income range do not opt to buy hence it is 0.0
* Probability of low-income category opting to buy KP781 which is high-end product is 0.12, for medium income category it is 0.3 and with high income range it is 0.57 which is greater value than other two income categories
* Probability of customer buying product KP281 whose age is below 25 is 0.45, whose age is between 25-35 is 0.4 & customer with age more than 35 is 0.17
* Probability of customer buying product KP481 whose age is below 25 is 0.46, whose age is between 25-35 is 0.4 & customer with age more than 35 is 0.13
* Probability of customer buying product KP281 whose age is below 25 is 0.42, whose age is between 25-35 is 0.42 & customer with age more than 35 is 0.15
* Portion of Females buying KP281 is 52.63%, KP481 is 38.15% & KP781 is 9.2% out of total female population
* Portion of males buying KP281 is 38.46%, KP481 is 29.80% & KP781 is 31.7% out of total male population
* Portion of low-income customer buying KP281 is 57.83%, KP481 is 36.14% & KP781 is 6.02%
* Portion of medium income customer buying KP281 is 43.24%, KP481 is 40.54% & KP781 is 16.22%
* Portion of high-income customer buying KP281 is 0.0%, KP481 is 0.0% & KP781 is 100.00%. i.e., customer with high income range always goes for the high-end product
* Portion of customer having age less than 25 years and buying KP281 is 43.04%, KP481 is 35.44% & KP781 is 21.52%
* Portion of customer having age between 25 to 35 years and buying KP281 is 43.84%, KP481 is 32.88% & KP781 is 23.29%
* Portion of customer having age more than 35 years and buying KP281 is 50.00%, KP481 is 28.57% & KP781 is 21.43%

Business Recommendations:

* Target high-income customers: The data shows that high-income customers are more likely to buy high-end products, Therefore, it would be wise to target high-income individuals for the high-end product (KP781) to maximize sales.

* Consider gender-specific marketing strategies: Males seem to be more likely to purchase the high-end product (KP781), so the company could consider marketing this product specifically to males. On the other hand, females are more likely to purchase KP481, so the company could consider marketing this product specifically to females.

* Consider income-specific marketing strategies: Customers with low income are more likely to purchase KP281, while customers with medium income are more likely to purchase KP481. High-income customers are more likely to purchase the high-end product (KP781). The company should consider targeting these income groups with specific marketing strategies. Although low-income individuals are less likely to purchase high-end products, they still represent a significant portion of the market for the other two products. To increase sales for these individuals, you could consider offering discounts or promotions to make the products more affordable.

* Focus on customers under 25: Customers under the age of 25 are more likely to purchase KP281 and KP481. The company could consider targeting this demographic with marketing campaigns to increase sales.The data shows that different age groups have different probabilities of purchasing each product. Therefore, you could create targeted marketing campaigns for each age group to increase the likelihood of sales.

* Emphasize fitness benefits: The data shows a strong correlation between usage, miles achieved, and fitness levels. The company could consider emphasizing the fitness benefits of the product in their marketing campaigns to appeal to health-conscious customers.

* Consider partnering with fitness influencers: Partnering with fitness influencers could help increase brand awareness and reach the target demographic of health-conscious customers. The data shows a strong correlation between product usage and fitness levels. Therefore, you could consider partnering with fitness brands or influencers to promote the products and appeal to individuals who prioritize their fitness.

* Offer discounts or promotions: Customers with lower incomes are less likely to purchase high-end products, so the company could consider offering discounts or promotions to make the high-end product more accessible to these customers.Since medium-income individuals have a similar purchase probability for all three products, you could consider offering bundle deals to encourage them to purchase all three products together.

* Market to both genders equally for product KP281: The data shows that both males and females are equally likely to purchase product KP281. Therefore, it would be wise to market to both genders equally to maximize sales.
* Focus on marketing to males for the high-end product (KP781): The data shows that males are more likely to purchase the high-end product (KP781) than females. Therefore, you could consider focusing your marketing efforts on males to increase sales of this product.

* Conduct further research on the usage and preferences of partnered females: The data shows a drop in usage of products among partnered females and a lower probability of purchasing high-end products. Further research could be conducted to understand the reasons behind this and identify opportunities for increasing sales to this demographic.