Business Case: Aerofit - Descriptive Statistics & Probability

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.

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. ¶

- **Product Purchased**: KP281, KP481, or KP781
- **Age**: In years
- **Gender**: Male/Female
- **Education**: In years
- **MaritalStatus****: Single or partnered
- **Usage**: The average number of times the customer plans to use the treadmill each week.
- **Income**: Annual income (in dollars)
- **Fitness**: Self-rated fitness on a 1-to-5 scale, where 1 is the poor shape and 5 is the excellent shape.
- **Miles**: The average number of miles the customer expects to walk/run each week

Product Portfolio:

- 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".

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import binom
```

In [2]:
```
original_data = pd.read_csv('aerofit_treadmill.csv')
original_data.head()
```

Out[2]:

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

In [3]:
```
data = original_data.copy(deep=True)
data.head()
```

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

In [4]:
```
data.shape
```

Out[4]: (180, 9)

In [5]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Product        180 non-null    object
 1   Age            180 non-null    int64 
 2   Gender         180 non-null    object
 3   Education      180 non-null    int64 
 4   MaritalStatus  180 non-null    object
 5   Usage          180 non-null    int64 
 6   Fitness        180 non-null    int64 
 7   Income         180 non-null    int64 
 8   Miles          180 non-null    int64 
dtypes: int64(6), object(3)
memory usage: 12.8+ KB
```

In [6]:
```
data.isnull().sum()
```

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

In [7]:
```
data.describe()
```

Out[7]:

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

Business Problem

The market research team at AeroFit wants to identify the characteristics of the target audience for each type of treadmill offered by the company, to provide a better recommendation of the treadmills to the new customers. The team decides to investigate whether there are differences across the product with respect to customer characteristics.

Perform descriptive analytics to create a customer profile for each AeroFit treadmill product by developing appropriate tables and charts.

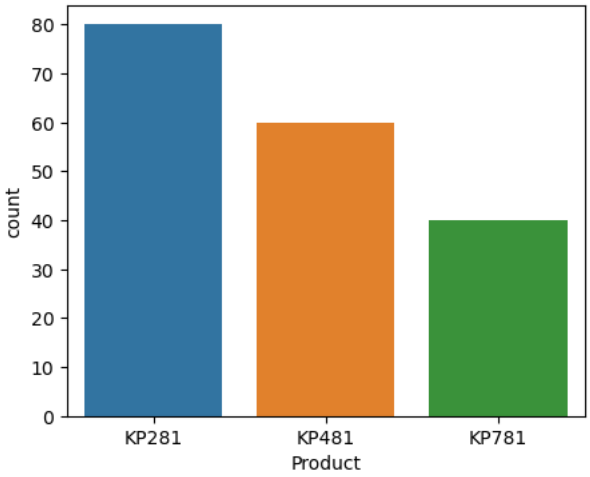
For each AeroFit treadmill product, construct two-way contingency tables and compute all conditional and marginal probabilities along with their insights/impact on the business.

Top Selling Product among 3 products

In [8]:
```
data['Product'].value_counts()
```

Out[8]:
```
Product
KP281    80
KP481    60
KP781    40
Name: count, dtype: int64
```

In [9]:
```
plt.figure(figsize=(5,4))
sns.countplot(data, x="Product")
plt.show()
```



Clearly KP281 is the most Selling Product

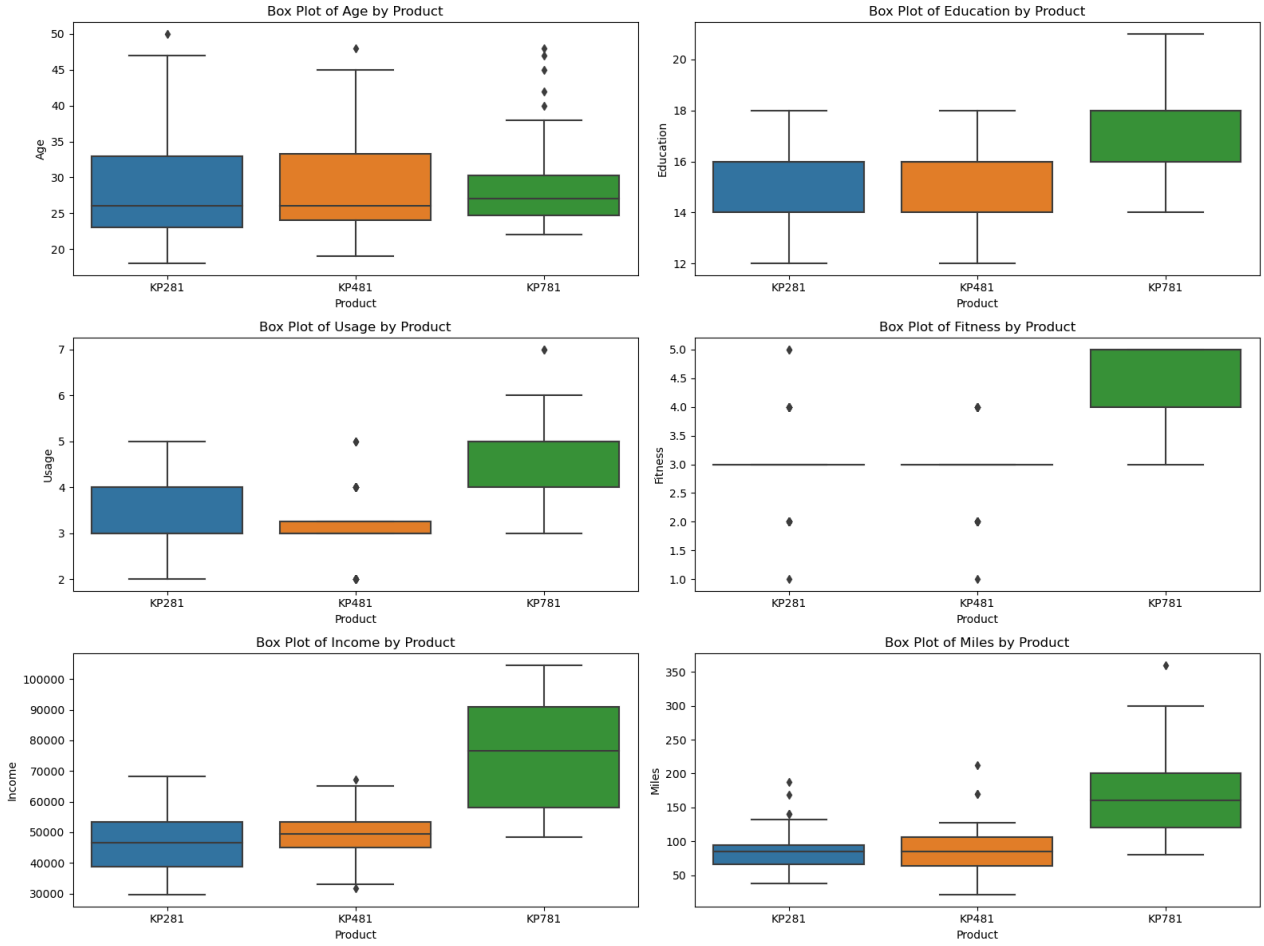
Let's investigate whether there are differences across the product with respect to customer characteristics.

In [10]:
```
data.describe()
```

Out[10]:

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

In [11]:
```
"""Smart Way of Writing code 😊"""

plt.figure(figsize=(16, 12))

integer_columns = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']

for i, column in enumerate(integer_columns, 1):
    plt.subplot(3, 2, i)  # 3 rows, 2 columns of subplots
    sns.boxplot(x='Product', y=column, data=data)
    plt.title(f'Box Plot of {column} by Product')
    plt.xlabel('Product')
    plt.ylabel(column)

plt.tight_layout()
plt.show()
```



Descriptive Analytics Summary for Aerofit Treadmill Products:

KP281:

Age: Average age is 28.55 years, ranging from 18 to 50.
Education: Average education level is 15.04 years (12 to 18 years).
Usage: Customers use the treadmill an average of 3 times per week (range: 1 to 5).
Fitness Level: The average fitness level is 3.09 (on a scale of 1 to 5).
Income: Average annual income is 46,418 (range: 29,562 to 68,220).
Miles: Customers run an average of 82.79 miles per week (range: 38 to 188 miles).

KP481:

Age: Average age is 28.90 years, ranging from 19 to 48.
Education: Average education level is 15.12 years (12 to 18 years).
Usage: Customers use the treadmill an average of 3.07 times per week (range: 1 to 5).
Fitness Level: The average fitness level is 3.07 (on a scale of 1 to 4).
Income: Average annual income is 48,974 (range: 31,836 to 67,083).
Miles: Customers run an average of 87.93 miles per week (range: 21 to 212 miles).

KP781:

Age: Average age is 29.10 years, ranging from 22 to 48.
Education: Average education level is 17.33 years (14 to 21 years).
Usage: Customers use the treadmill an average of 4.78 times per week (range: 2 to 7).

Fitness Level: The average fitness level is 4.78 (on a scale of 3 to 5).
Income: Average annual income is 75,442 (range: 48,556 to 104,581).
Miles: Customers run an average of 166.9 miles per week (range: 80 to 360 miles).

Observations:

- KP781 is used by **Higher-Income**, more **Educated**, and **Fitter** customers, who also run significantly more miles compared to users of KP281 and KP481.
- KP281 and KP481 are similar in terms of usage and fitness level, but KP481 users have slightly higher income and run more miles per week.

In [12]:
```
data[data['Product']=='KP781']['Miles'].mean()
```

Out[12]: 166.9

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

In [13]:
```
data.head()
```

Out[13]:

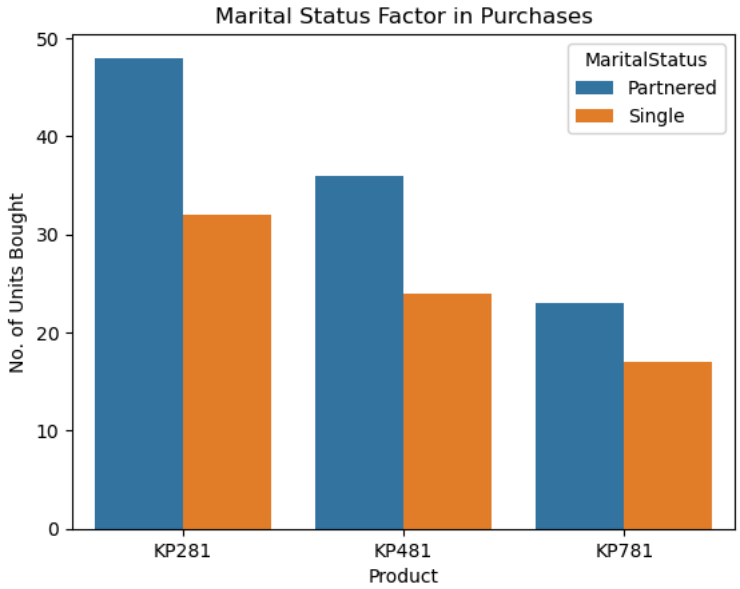
	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4	29562	112
1	KP281	19	Male	15	Single	2	3	31836	75
2	KP281	19	Female	14	Partnered	4	3	30699	66
3	KP281	19	Male	12	Single	3	3	32973	85
4	KP281	20	Male	13	Partnered	4	2	35247	47

In [14]:
```
marital_data = data.groupby(by=['Product', 'MaritalStatus'])['Product'].agg(['count']).reset_index()
marital_data
```

Out[14]:

	Product	MaritalStatus	count
0	KP281	Partnered	48
1	KP281	Single	32
2	KP481	Partnered	36
3	KP481	Single	24
4	KP781	Partnered	23
5	KP781	Single	17

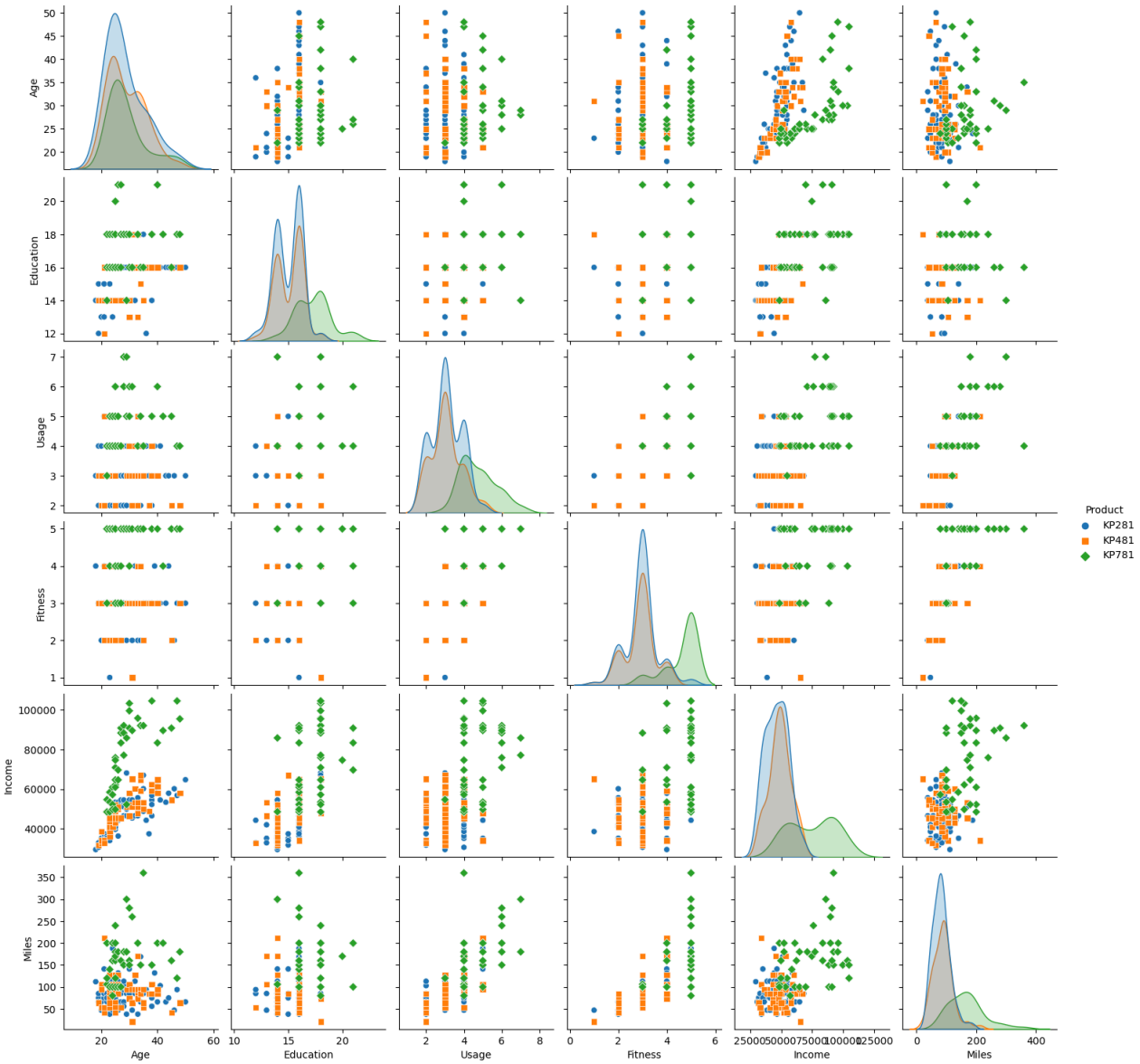
In [15]:
```
sns.barplot(marital_data, x='Product', y='count', hue="MaritalStatus")
plt.ylabel("No. of Units Bought")
plt.title("Marital Status Factor in Purchases")
plt.show()
```



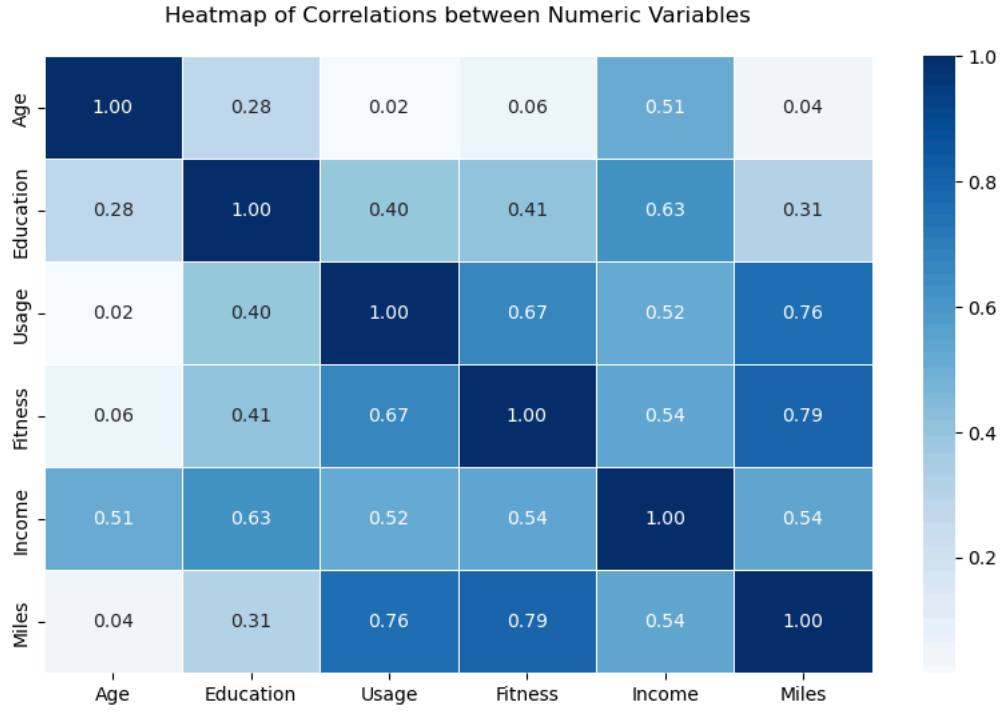
Partnerd customers purchase count is higher it tells that they want to stay fit. So we can target partnered customers to sell other fitness related products. And for single customers we can use different marketing strategy like ads.

In [16]:
```
sns.pairplot(data=data, hue='Product', markers=["o", "s", "D"])
plt.show()
```

```
C:\Users\302sy\anaconda3\Lib\site-packages\seaborn\axisgrid.py:118: UserWarning: The figure layout has changed to tight
  self._figure.tight_layout(*args, **kwargs)
```



In [17]:
```
# Select only the numeric columns for correlation
numeric_data = data[['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']]

# Calculate the correlation matrix
correlation_matrix = numeric_data.corr()

# # Plot the heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='Blues', linewidths=0.5, fmt='.2f')
plt.title('Heatmap of Correlations between Numeric Variables\n')
plt.show()
```



What is the probability of a male customer buying a KP781 treadmill?

In [18]:
```
male_kp781 = data[(data['Product']=='KP781')&(data['Gender']=='Male')].shape[0]

total_male = data[data['Gender']=='Male'].shape[0]
```

In [19]:
```
prob_male_kp781 = male_kp781/total_male
print(f"The Probability of male customers buying a KP781 is {round(prob_male_kp781*100,2)}%")
```

```
The Probability of male customers buying a KP781 is 31.73%
```

Practising more like the above question

Probability of a Customer Being Female if chosen randomly

In [20]:
```
total_cus = len(data['Gender'])
fem_cus = len(data[data['Gender']=='Female'])

prob_cus_female = fem_cus/total_cus # 76/180
print(f"Probability of a Customer Being Female is: {round(prob_cus_female,4)*100}%")
```

```
Probability of a Customer Being Female is: 42.22%
```

Probability of a High Fitness Level (Fitness > 3) if chosen randomly

In [21]:
```
fit_cus = data[data['Fitness'] > 3].shape[0]

prob_fit_cus = fit_cus/total_cus #55/180
print(f"Probability of a Customer Being Fit is: {round(prob_fit_cus,4)*100}%")
```

```
Probability of a Customer Being Fit is: 30.56%
```

Out of all customers, how many both use the KP781 treadmill and have a high fitness level?

In [22]:
```
cus_fit_kp781 = data[(data['Fitness']>3) & (data['Product']=="KP781")].shape[0]

prob_cus_fit_kp781 = cus_fit_kp781/total_cus

print(f"Probability of a Customer randomly chosen is Being Fit and using KP781 is: {round(prob_cus_fit_kp781,4)*100}%")
```

```
Probability of a Customer randomly chosen is Being Fit and using KP781 is: 20.0%
```

Probability of a High Fitness Level (Fitness > 3) using KP781

P(Fit_n_using_KP781 | total_KP781) = P(high_fitness ∩ KP781) / P(total_KP781)

In [23]:
```
prob_fitter_KP781 = (data[(data['Fitness']>3) & (data['Product']=="KP781")].shape[0])/len(data[data["Product"]=="KP781"])

## prob = 36/40 => 0.9
print(f"Probability of a Customer with High Fitness Level (Fitness > 3) using KP781 is: {round(prob_fitter_KP781,4)*100}%")
```

```
Probability of a Customer with High Fitness Level (Fitness > 3) using KP781 is: 90.0%
```

Probability of High Fitness Level for KP781 Users

The probability that a customer using KP781 has a fitness level greater than 3 is high, indicating this model is more popular among fitter individuals.

Insight: KP781 is a popular choice for those who consider themselves highly fit, aligning with its premium features and higher price point.

Actionable Item: Market KP781 with a focus on professional athletes, fitness enthusiasts, or customers aiming for intensive workouts.

In [24]:
```
data.groupby(by=['Product', 'Gender'])['Gender'].agg(['count']).reset_index()
```

Out[24]:

	Product	Gender	count
0	KP281	Female	40
1	KP281	Male	40
2	KP481	Female	29
3	KP481	Male	31
4	KP781	Female	7
5	KP781	Male	33

Use Conditional Probability for few problems

If a customer is Partnered, what is the conditional probability that their Usage is 4 or more times per week

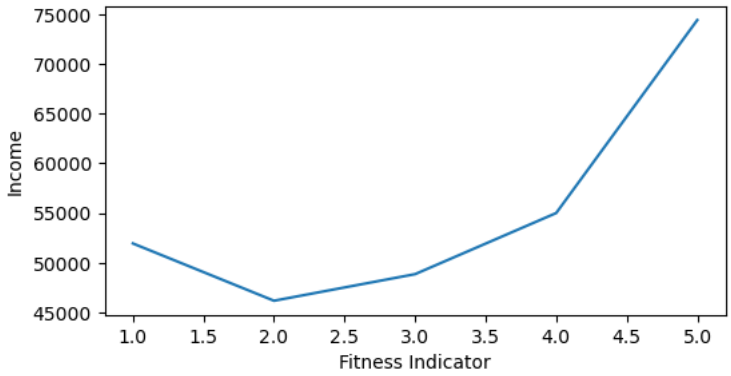
In [25]:
```
no_of_partnered = data[data['MaritalStatus']=='Partnered'].shape[0]

prob_cus_partnered = no_of_partnered/total_cus # 0.5944

# prob(use > 4 | partnered) = prob(use>4 & partnered)/ prob(partnered)

prob_highusage_partnered_intesection = data[(data['MaritalStatus']=="Partnered")&(data['Usage'] >= 4)].shape[0]/total_cus # (

prob_highusage_partnered = prob_highusage_partnered_intesection/prob_cus_partnered #Conditional Probability

print(f"If a customer is Partnered, the conditional probability that their Usage is 4 or more per week is {round(prob_highusa
```

```
If a customer is Partnered, the conditional probability that their Usage is 4 or more per week is 42.06%
```

In [26]:
```
 # If a customer is Single, what is the probability that they run more than 100 miles per week?

prob_single = 1-prob_cus_partnered

# prob(run>100 miles | Single) = prob(sing & miles>100)/ prob(single)

prob_single_and_runner = data[(data['MaritalStatus']=="Single")&(data['Miles']>100)].shape[0]/total_cus

prob_single_runner = prob_single_and_runner/prob_single

print(f"If a customer is Single, the probability that they run more than 100 miles per week is {round(prob_single_runner*100
```

```
If a customer is Single, the probability that they run more than 100 miles per week is 42.47%
```

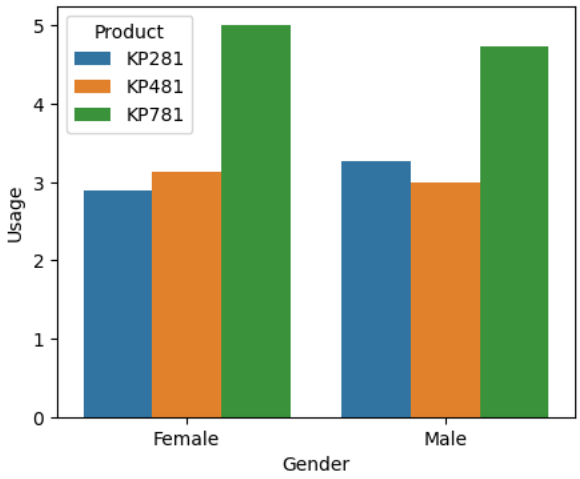
In [27]:
```
# What is the average income of customers based on their fitness levels?

df = data.groupby(by='Fitness')['Income'].agg("mean").reset_index()
# print(df)
plt.figure(figsize=(6,3)) # change it to 5,5 when you want bigger picture
plt.plot(df["Fitness"],df["Income"])
plt.xlabel("Fitness Indicator")
plt.ylabel("Income")
plt.show()
```



Insight: It is Observed that People who earns more or has high average Income and focuses more on their fitness. For that they are happy to spend amount in order to stay fit.

Actionable Item: We can provide them more fitness services and products like Yoga Sessions, Protein powder etc

In [28]:
```
# How does treadmill usage differ between male and female customers?

usage_by_gender = data.groupby(by=['Gender', "Product"])['Usage'].mean().reset_index()
usage_by_gender

plt.figure(figsize=(5,4))
sns.barplot(data=usage_by_gender,x = "Gender", y="Usage", hue="Product")
plt.show()
```

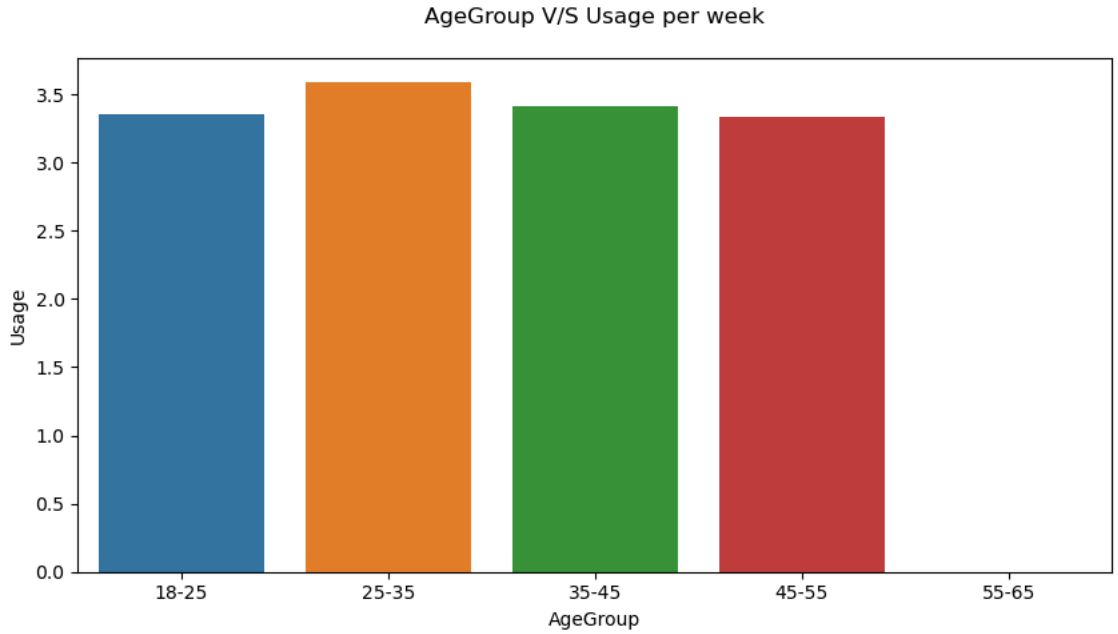


The usage of treadmill is higher in female that too most of them uses KP781 product even though the Product purchase count of KP781 is less but it is used more.

In [29]:
```
# If we randomly select 5 customers, what is the probability that exactly 3 of them have a Fitness level of 4 or higher?
n = 5
k = 3
p = data[data['Fitness'] >= 4].shape[0]/len(data)

binom.pmf(n=n,k=k,p=p)
```

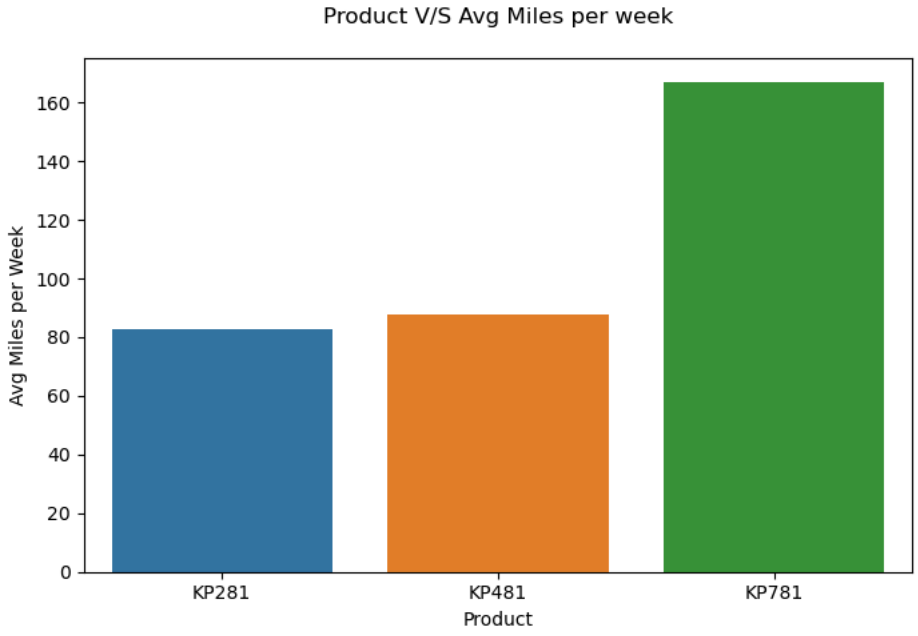
Out[29]:
```
0.1375769157288861
```

In [30]:
```
# Which age group (e.g., 18-25, 26-35, etc.) uses the treadmill the most?
bins = [18, 25, 35, 45, 55, 65]
labels = ["18-25", "25-35", "35-45", "45-55", "55-65"]
data['AgeGroup']=pd.cut(data["Age"], bins=bins, labels = labels)

usage_by_age = data.groupby("AgeGroup")['Usage'].mean().reset_index()

plt.figure(figsize=(10,5))

sns.barplot(data=usage_by_age, x = "AgeGroup", y="Usage")
plt.title("AgeGroup V/S Usage per week\n")
plt.show()
```



Age 25-35 tend to use tredmill more in comparison to other age groups as they are young and wants to stay fit. Followed by 35-45 age group.

In [31]:
```
# How many miles do customers typically run on different treadmill models?

avg_cus_miles = data.groupby("Product")['Miles'].agg(["mean"]).reset_index()
# print(avg_cus_miles)
plt.figure(figsize=(8,5))
sns.barplot(data= avg_cus_miles, x="Product", y="mean")
plt.title("Product V/S Avg Miles per week\n")
plt.ylabel("Avg Miles per Week")
plt.show()
```



Avg Miles per Week on Specific Tread mill KP781 is used more followed by KP481 and KP281

Conclusion

KP781 should be positioned as the premium treadmill for high-income, highly fit customers who seek advanced features and heavy usage.
KP281 can be marketed as a budget-friendly option for younger, less experienced users with lower fitness levels.
KP481 sits between the two, attracting slightly more educated customers, and could be promoted to families or mid-tier fitness users.

In [ ]: