Objective :- Import the dataset and understand its basic structure, data types, size, and missing values.

```
from google.colab import files
files.upload()
```

Choose Files aerofit_data...e study.csv
aerofit_data.csv - auroft case study.csv(text/csv) - 7458 bytes, last modified: 2/3/2026 - 100% done

```
Saving aerofit_data.csv - auroft case study.csv to aerofit_data.csv - auroft case study.csv
{'aerofit_data.csv - auroft case study.csv':
b'Product,Age,Gender,Education,MaritalStatus,Usage,Fitness,Income,Miles\r\nKP281,18,Male,14,Single,3,4,29562,112\r\nKP281,19
```

```
df = pd.read_csv("aerofit_data.csv - auroft case study.csv")
df.head()
```

	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

Next steps: Generate code with df New interactive sheet

```
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. Required Libraries

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

```
sns.set(style="whitegrid")
```

1- **Data Understanding & Basic Checks**

1.1 - Convert Categorical Columns

```
cat_cols = ['Product', 'Gender', 'MaritalStatus']

for col in cat_cols:
    df[col] = df[col].astype('category')


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    category
 1   Age            180 non-null    int64
 2   Gender         180 non-null    category
 3   Education      180 non-null    int64
 4   MaritalStatus  180 non-null    category
 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: category(3), int64(6)
memory usage: 9.5 KB
```

Insights - Product, Gender, and MaritalStatus were converted to categorical data type.

1.2 **Check Missing Values**

```
df.isnull().sum()
```

	0
Product	0
Age	0
Gender	0
Education	0
MaritalStatus	0
Usage	0
Fitness	0
Income	0
Miles	0

dtype: int64

Insight:

No missing values were found in the dataset.

Hence, no data cleaning was required for missing values.

1.3 **Statistical Summary (Describe)**

```
df.describe()
```

	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

Insights :- Age of customers ranges from young adults to older individuals.

Income and Miles show a wide range, indicating different customer segments.

Difference between mean and median suggests slight skewness in some variables.

1.4 **Value Counts**

```
df['Product'].value_counts()
```

Product	count
KP281	80
KP481	60
KP781	40

dtype: int64

```
df['Gender'].value_counts()
```

Gender	count
Male	104
Female	76

dtype: int64

```
df['MaritalStatus'].value_counts()
```

MaritalStatus	count
Partnered	107
Single	73

dtype: int64

Insights :- KP281 has the highest number of purchases, indicating it is the most popular entry-level product.

Male customers are slightly more than female customers.

Partnered customers form a significant portion of buyers.

1.5 **Unique Values Check**

```
df.nunique()
```

	0
Product	3
Age	32
Gender	2
Education	8
MaritalStatus	2
Usage	6
Fitness	5
Income	62
Miles	37

dtype: int64

Insight:

The dataset has limited unique categories, making it suitable for categorical analysis.

Product variable has three unique treadmill mode

O2. utlier Detection & Treatment

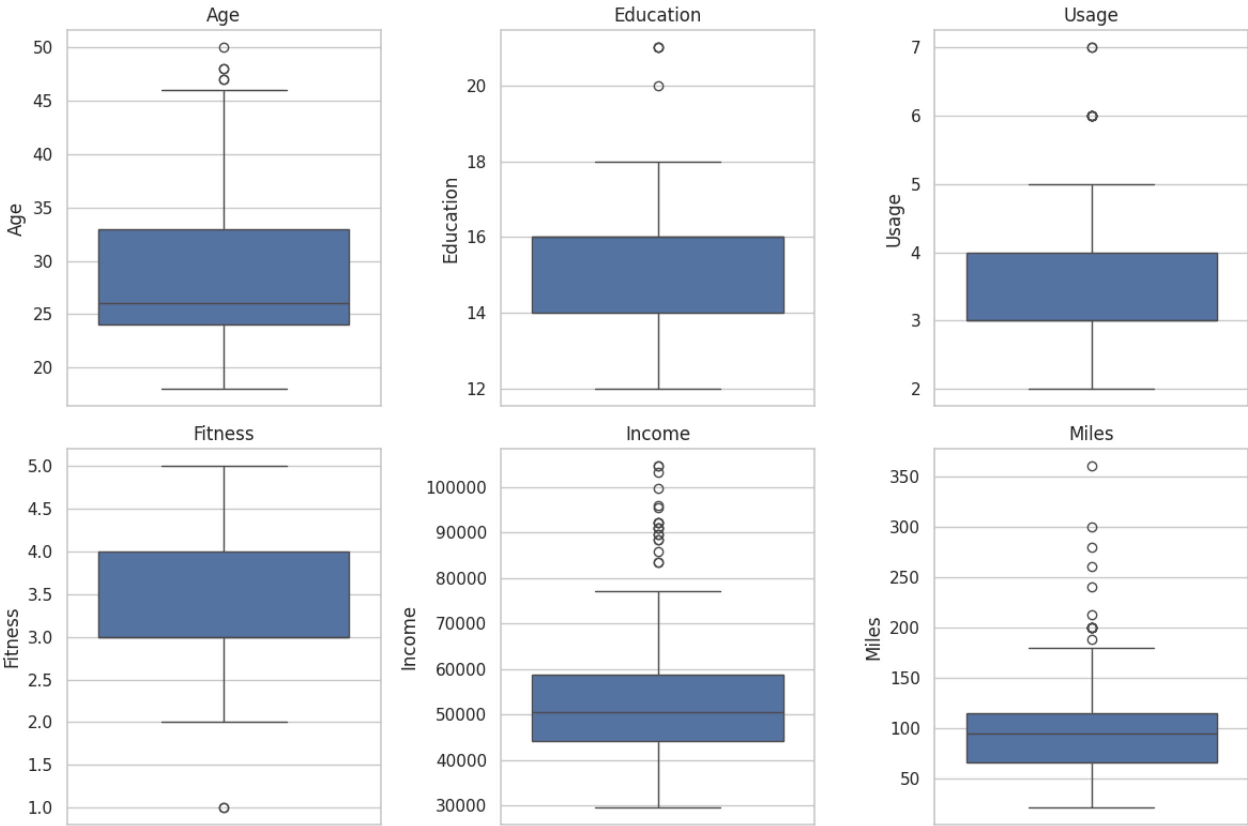
```
2.1 Identify Continuous Variables
```

```
cont_cols = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
cont_cols
```

```
['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
```

2.2 Boxplots Before Outlier Treatment

```
plt.figure(figsize=(12, 8))

for i, col in enumerate(cont_cols):
    plt.subplot(2, 3, i+1)
    sns.boxplot(y=df[col])
    plt.title(col)

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



**Insight **:-

Income and Miles show visible outliers on the higher end.

Usage and Fitness have relatively fewer outliers.

Presence of outliers indicates diverse customer behavior, especially for premium products.

2.3 **Outlier Detection via Describe (Mean vs Median)**

```
df[cont_cols].describe()
```

	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

Insight:

For variables like Income and Miles, mean is higher than median, indicating right skewness.

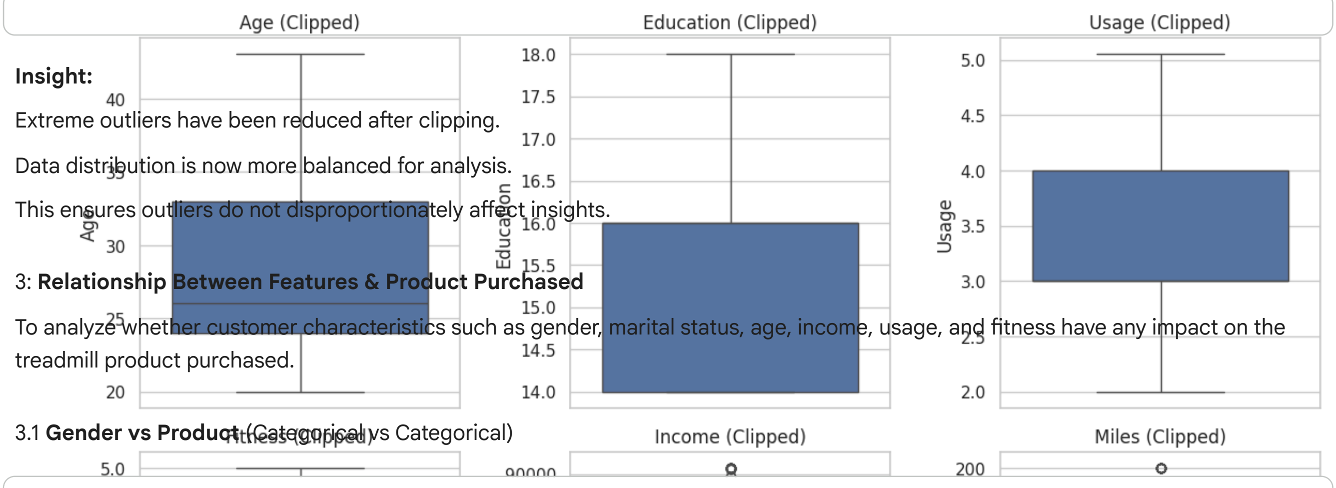
This skewness suggests the presence of high-value outliers.

2.4 **Outlier** Treatment Clipping (5th & 95th Percentile)

```
df_clipped = df.copy()

for col in cont_cols:
    lower = df[col].quantile(0.05)
    upper = df[col].quantile(0.95)
    df_clipped[col] = np.clip(df[col], lower, upper)
```

2.5 **Boxplots After Clipping**

```
plt.figure(figsize=(12, 8))

for i, col in enumerate(cont_cols):
    plt.subplot(2, 3, i+1)
    sns.boxplot(y=df_clipped[col])
    plt.title(col + " (Clipped)")

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



Insight:

Extreme outliers have been reduced after clipping.

Data distribution is now more balanced for analysis.

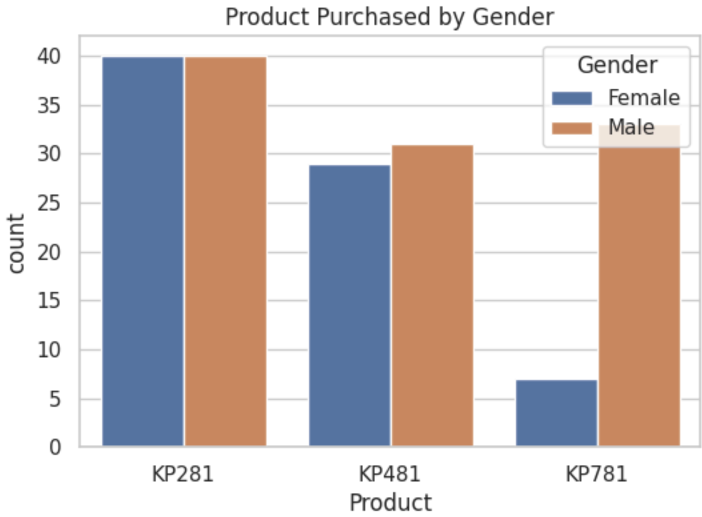
This ensures outliers do not disproportionately affect insights.

3: **Relationship Between Features & Product Purchased**

To analyze whether customer characteristics such as gender, marital status, age, income, usage, and fitness have any impact on the treadmill product purchased.

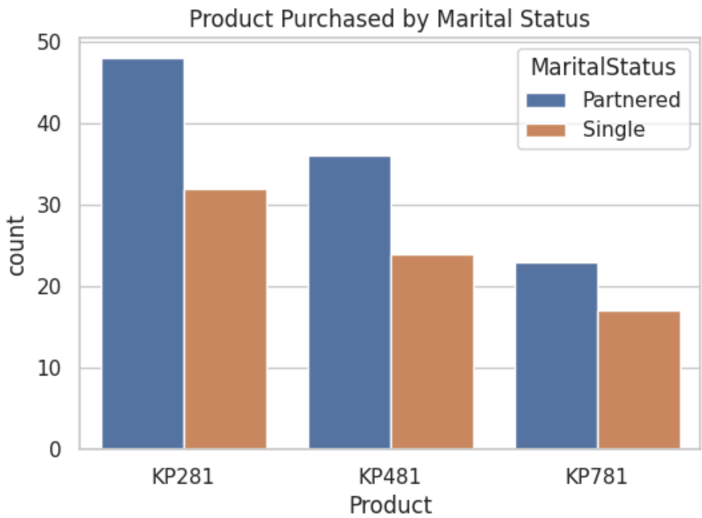
3.1 **Gender vs Product**(Categorical vs Categorical)

```
plt.figure(figsize=(6,4))
sns.countplot(data=df_clipped, x='Product', hue='Gender')
plt.title('Product Purchased by Gender')
plt.show()
```



3.2 **Marital Status vs Product**(Categorical vs Categorical))

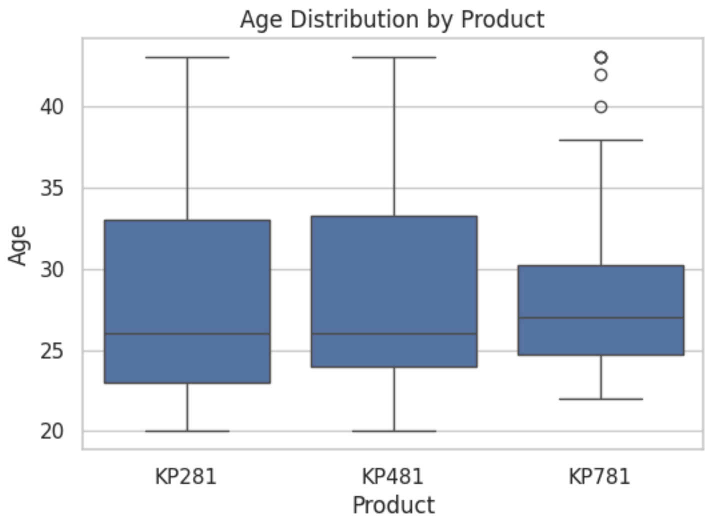
```
plt.figure(figsize=(6,4))
sns.countplot(data=df_clipped, x='Product', hue='MaritalStatus')
plt.title('Product Purchased by Marital Status')
plt.show()
```



INSIGHT:- Partnered customers show a higher preference for KP481 and KP781 compared to single customers. Single customers are more inclined towards KP281, which is an entry-level treadmill. This suggests that partnered customers, possibly with more stable income, tend to purchase mid to premium products.

3.3 Age vs Product(Continuous vs Categorical – Boxplot)

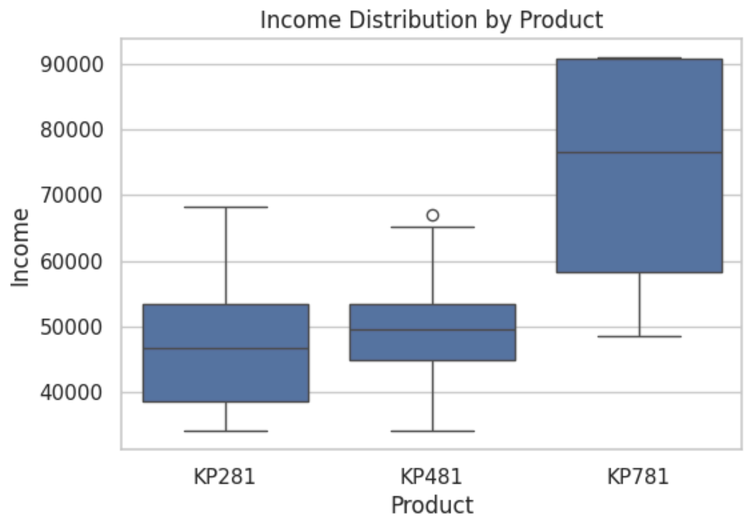
```
plt.figure(figsize=(6,4))
sns.boxplot(data=df_clipped, x='Product', y='Age')
plt.title('Age Distribution by Product')
plt.show()
```



Insights :- Customers purchasing KP281 are generally younger in age. KP481 buyers fall in a middle age range, while KP781 buyers tend to be relatively older. This indicates that age plays a role in treadmill selection, with more experienced users preferring advanced models.

3.4 Income vs Product (Continuous vs Categorical Boxplot)

```
plt.figure(figsize=(6,4))
sns.boxplot(data=df_clipped, x='Product', y='Income')
plt.title('Income Distribution by Product')
plt.show()
```

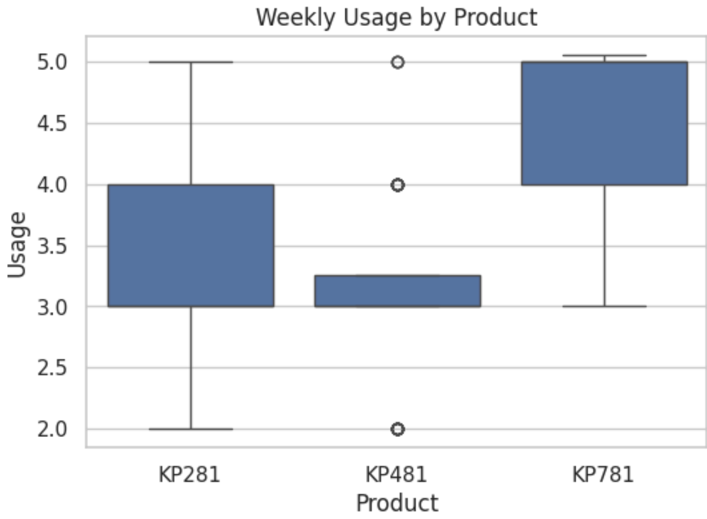


Insights :- There is a clear increase in income levels from KP281 to KP781 customers. KP281 buyers belong to relatively lower income groups, while KP781 buyers fall in significantly higher income brackets. Income is a strong determinant in the purchase of premium treadmills.

3.5 Usage vs Product (Continuous vs Categorical Boxplot)

```
plt.figure(figsize=(6,4))
sns.boxplot(data=df_clipped, x='Product', y='Usage')
plt.title('Weekly Usage by Product')
```

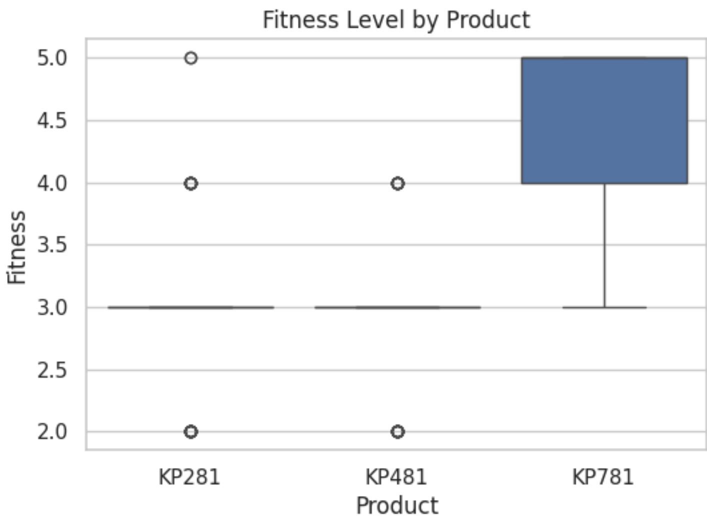
```
plt.show()
```



Insights :- KP781 customers plan to use the treadmill more frequently each week compared to KP281 and KP481 customers. KP281 users show lower to moderate weekly usage, indicating casual fitness routines. Higher planned usage aligns with the advanced features of premium treadmills.

3.6 Fitness vs Product (Continuous vs Categorical – Boxplot)

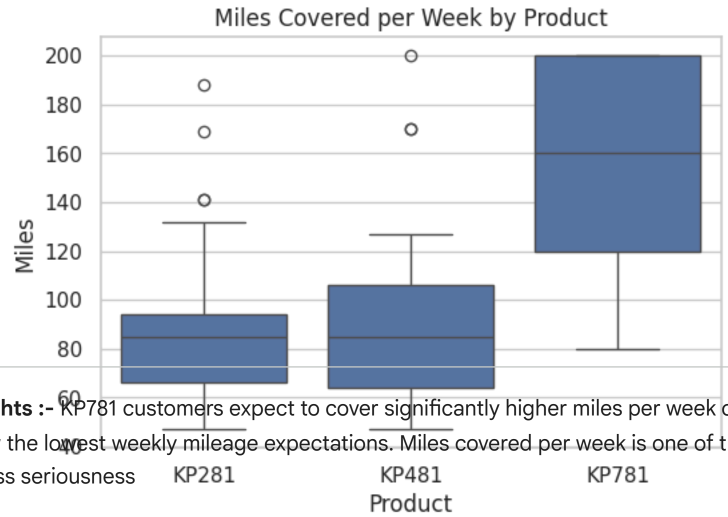
```
plt.figure(figsize=(6,4))
sns.boxplot(data=df_clipped, x='Product', y='Fitness')
plt.title('Fitness Level by Product')
plt.show()
```



Insights :- Customers purchasing KP781 report higher self-rated fitness levels, mostly in the range of 4 to 5. KP281 customers predominantly fall within fitness levels 2 to 3. Fitness level plays a significant role in determining the type of treadmill purchased.

3.7 Miles vs Product (Continuous vs Categorical Boxplot)

```
plt.figure(figsize=(6,4))
sns.boxplot(data=df_clipped, x='Product', y='Miles')
plt.title('Miles Covered per Week by Product')
plt.show()
```



Insights :- KP781 customers expect to cover significantly higher miles per week compared to KP281 and KP481 customers. KP281 users show the lowest weekly mileage expectations. Miles covered per week is one of the strongest indicators of treadmill selection and fitness seriousness

4 Probability Analysis (Marginal & Conditional Probability)

4.1 Marginal Probability - (Overall probability of each product being purchased)

```
product_prob = pd.crosstab(df_clipped['Product'], columns='Count', normalize=True)
product_prob
```

col_0	Count
Product	
KP281	0.444444
KP481	0.333333
KP781	0.222222

Next steps: Generate code with product_prob New interactive sheet

Insight - The highest proportion of customers have purchased the KP281 treadmill, making it the most commonly sold product. KP481 occupies a moderate share, while KP781 has the lowest share, indicating its premium positioning. This shows that entry-level treadmills cater to a larger mass market, while advanced models target niche segments.

4.2 Probability of Product Purchase by Gender

P(Product | Gender)

```
gender_product_prob = pd.crosstab(
    df_clipped['Gender'],
    df_clipped['Product'],
    normalize='index'
)
gender_product_prob
```

Product	KP281	KP481	KP781
Gender			
Female	0.526316	0.381579	0.092105
Male	0.384615	0.298077	0.317308

Next steps: Generate code with gender_product_prob New interactive sheet

Insight

Given that the customer is male, the probability of purchasing KP781 is higher compared to female customers. Female customers show a stronger preference towards KP281 and KP481. Gender influences product choice, especially for premium treadmills.

4.3 Probability of Product Purchase by Marital Status

P(Product | MaritalStatus)

```
marital_product_prob = pd.crosstab(
    df_clipped['MaritalStatus'],
    df_clipped['Product'],
    normalize='index'
)
marital_product_prob
```

MaritalStatus	KP281	KP481	KP781
Partnered	0.448598	0.336449	0.214953
Single	0.438356	0.328767	0.232877

Next steps: Generate code with marital_product_prob | New interactive sheet

Insight

Partnered customers have a higher probability of purchasing KP481 and KP781. Single customers show a higher likelihood of buying KP281. This suggests that lifestyle and financial stability may impact treadmill selection.

4.4 Probability of Product Purchase by Fitness Level

P(Product | Fitness)

```
fitness_product_prob = pd.crosstab(
    df_clipped['Fitness'],
    df_clipped['Product'],
    normalize='index'
)
fitness_product_prob
```

Fitness	KP281	KP481	KP781
2	0.535714	0.464286	0.000000
3	0.556701	0.402062	0.041237
4	0.375000	0.333333	0.291667
5	0.064516	0.000000	0.935484

Next steps: Generate code with fitness_product_prob | New interactive sheet

Insight

Customers with higher fitness levels (4–5) have a much higher probability of purchasing KP781. Customers with lower fitness levels tend to purchase KP281. Fitness level is a strong predictor of advanced treadmill adoption.

4.6 Probability by Usage

```
usage_product_prob = pd.crosstab(
    df_clipped['Usage'],
    df_clipped['Product'],
    normalize='index'
)
usage_product_prob
```

Usage	KP281	KP481	KP781
2.00	0.575758	0.424242	0.000000
3.00	0.536232	0.449275	0.014493
4.00	0.423077	0.230769	0.346154
5.00	0.117647	0.176471	0.705882
5.05	0.000000	0.000000	1.000000

Next steps: Generate code with usage_product_prob New interactive sheet

Insight

Customers with higher planned weekly usage have a greater probability of purchasing KP781. Lower usage customers mostly prefer KP281.

5 : Correlation Analysis (Heatmap) To understand the relationship between different numerical variables and identify which factors move together and influence treadmill usage and purchase behavior.

5.1 Select Numerical Columns

```
num_cols = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
num_cols
```

```
['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
```

5.2 Correlation Matrix

```
corr_matrix = df_clipped[num_cols].corr()
corr_matrix
```

	Age	Education	Usage	Fitness	Income	Miles
Age	1.000000	0.301971	0.015394	0.057361	0.514362	0.029636
Education	0.301971	1.000000	0.413600	0.441082	0.628597	0.377294
Usage	0.015394	0.413600	1.000000	0.661978	0.481608	0.771030
Fitness	0.057361	0.441082	0.661978	1.000000	0.546998	0.826307
Income	0.514362	0.628597	0.481608	0.546998	1.000000	0.537297
Miles	0.029636	0.377294	0.771030	0.826307	0.537297	1.000000

Next steps: Generate code with corr_matrix New interactive sheet

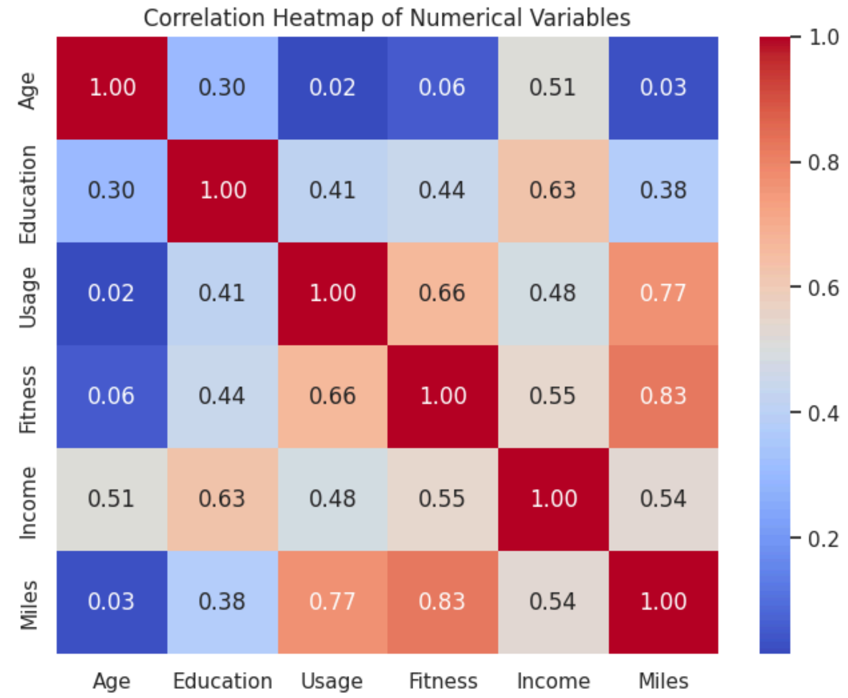
Insight

Correlation values range between -1 and +1.

Positive values indicate variables increase together.

5.3 Heatmap Visualization

```
plt.figure(figsize=(8,6))
sns.heatmap(
    corr_matrix,
    annot=True,
    cmap='coolwarm',
    fmt='.2f'
)
plt.title('Correlation Heatmap of Numerical Variables')
plt.show()
```



**Key Insights **

Usage and Miles show a strong positive correlation, indicating that customers who use the treadmill more frequently also tend to cover more miles.

Fitness has a moderate positive correlation with both Usage and Miles, suggesting fitter customers use the treadmill more intensively.

Income shows a weak to moderate correlation with Usage and Fitness, indicating that while income influences product choice, it does not directly determine usage intensity.

Age has a weak correlation with most variables, suggesting age alone is not a strong driver of treadmill usage behavior.

6. Customer Profiling & Recommendations :- To create detailed customer profiles for each treadmill product and provide actionable business recommendations based on the analysis.

CUSTOMER PROFILING

KP281 :- Entry-Level Treadmill ($1,500)

- Age: Mostly younger customers
- Gender: Both, slightly male-dominated
- Marital Status: Mostly single
- Income: Lower income group
- Fitness Level: 2–3 (beginner to moderate)
- Usage: 2–3 times per week
- Miles: Low weekly mileage

Interpretation

KP281 is preferred by beginners or casual fitness users who are price-sensitive and just starting their fitness journey.

KP481 - Mid-Level Treadmill ($1,750)

Age: Middle-aged group

Gender: Balanced male and female

Marital Status: Largely partnered

Income: Medium income group

Fitness Level: 3–4

Usage: 3–4 times per week

Miles: Moderate weekly mileage

Interpretation

KP481 attracts regular fitness users who want better features but are not ready to invest in premium treadmills.

KP781 - Advanced Treadmill ($2,500)

Age: Older and more experienced users

Gender: Predominantly male

Marital Status: Mostly partnered

Income: High income group

Fitness Level: 4–5 (advanced)

Usage: 4–7 times per week

Miles: High weekly mileage

Interpretation

KP781 is purchased by serious fitness enthusiasts and runners who value advanced features and performance.

BUSINESS RECOMMENDATIONS

1. Product Recommendation Strategy

Recommend KP281 to first-time buyers and budget-conscious customers.

Suggest KP481 to customers with moderate fitness levels looking to upgrade.

Promote KP781 to high-income customers with high fitness and usage levels.

2. Targeted Marketing

Run beginner-focused campaigns for KP281.

Position KP481 as a "value-for-money upgrade."

Market KP781 as a performance-oriented treadmill for serious users.

3. Upselling Opportunities

Encourage KP281 users to upgrade to KP481 after consistent usage.

Offer KP781 trials or demos to KP481 users with increasing usage.

4. Personalised Sales Guidance

Use customer attributes (age, income, fitness) at store level to suggest the most suitable treadmill.

This improves customer satisfaction and reduces return rates.

Close