Aerofit Business Case

1. Import the dataset and do usual data analysis steps like checking the structure & characteristics of the dataset

```
!wget https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/125/ori
```

```
--2025-02-16 15:16:51--  https://d2beiqkhq929f0.cloudfront.net/public_asset
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)...
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)
HTTP request sent, awaiting response... 200 OK
Length: 7279 (7.1K) [text/plain]
Saving to: ‘aerofit_treadmill.csv’

aerofit_treadmill.c 100%[===================>]   7.11K  --.-KB/s    in 0s

2025-02-16 15:16:51 (2.32 GB/s) - ‘aerofit_treadmill.csv’ saved [7279/7279]
```

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings('ignore')
```

```
df = pd.read_csv('aerofit_treadmill.csv')
```

```
df
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	M
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	

180 rows × 9 columns

```
#Will check the data type of each column in th data
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
```

```
#let's check the number of rows & columns
df.shape
```

```
(180, 9)
```

The data has 180 rows & 9 columns

```
#Check for the missing values and find the number of missing values in each colu
df.isnull().sum()
```

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

dtype: int64

No missing values in any column.

2. Detect Outliers

```
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

Key Observations from Summary Statistics:

Age: Ranges from 18 to 50.

Mean: 28.8 years, Median: 26 years → Slightly right-skewed.

Education (Years of Study): Ranges from 12 to 21 years. Mean: 15.57 years, Median: 16 years → Mostly balanced.

Usage (Times per Week): Ranges from 2 to 7 times. Mean: 3.46 times, Median: 3 times → Most customers use it around 3-4 times per week.

Fitness (Self-Rating 1-5): Mean: 3.31, Median: 3. Most people rate themselves as moderately fit.

Income: Ranges from $29,562 to$104,581. Mean: $53,719, Median:$50,596 → Right-skewed (some high-income customers).

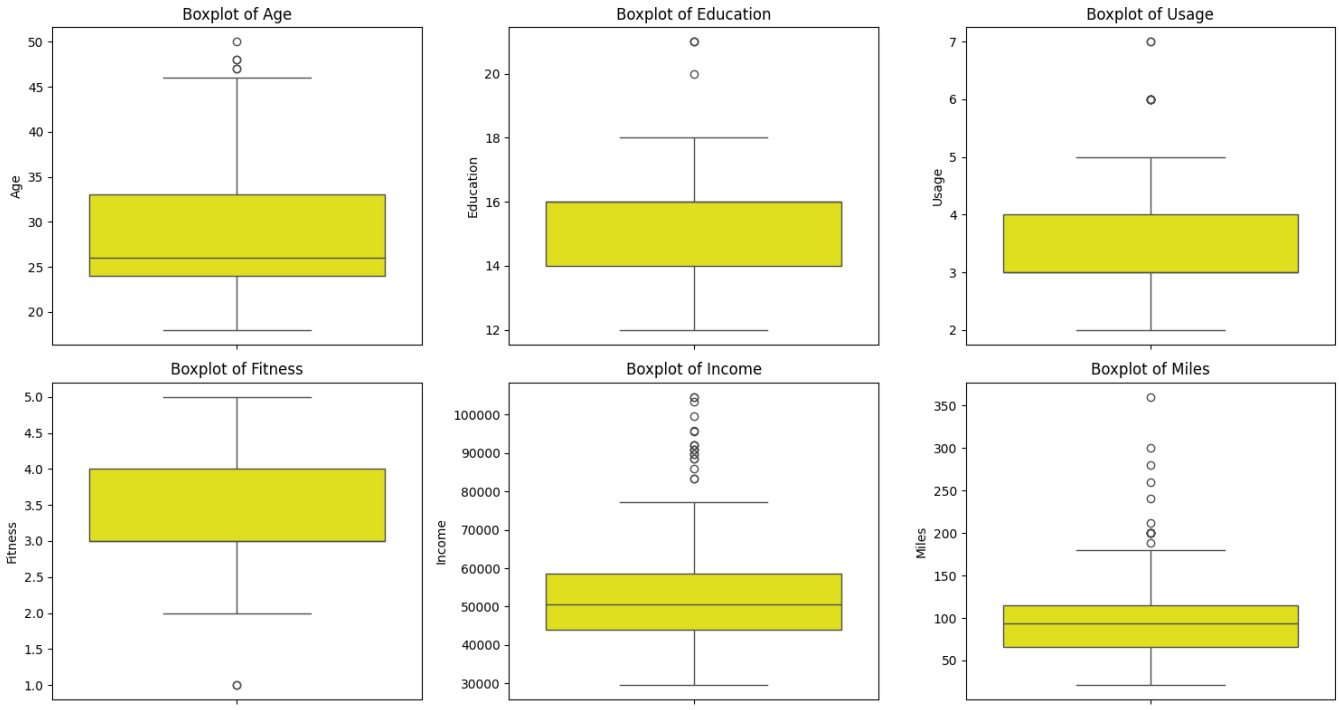
Miles (Expected to Run/Walk per Week): Ranges from 21 to 360 miles. Mean: 103 miles, Median: 94 miles → Some extreme values.

Next, I'll visualize outliers using boxplots.

```
plt.figure(figsize=(15, 8))

# Plot boxplots for numerical columns
numerical_columns = ["Age", "Education", "Usage", "Fitness", "Income", "Miles"]
for i, col in enumerate(numerical_columns, 1):
    plt.subplot(2, 3, i)
    sns.boxplot(y=df[col], color='yellow')
    plt.title(f"Boxplot of {col}")

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



Outlier Observations from Boxplots:

Age, Education, Usage, and Fitness: No significant outliers.

Income: A few customers have significantly high income (above ~$90,000), which may be outliers.

Miles: There are extreme values (above 300 miles), indicating possible outliers.

Remove/clip the data between the 5 percentile and 95 percentile

```
numerical_cols = ["Age", "Education", "Usage", "Fitness", "Income", "Miles"]

# Compute 5th and 95th percentiles
lower_bounds = df[numerical_cols].quantile(0.05)
upper_bounds = df[numerical_cols].quantile(0.95)


df_clipped = df.copy()
# df_clipped[numerical_cols] = np.clip(df[numerical_cols], lower_bounds, upper_

# # Display before and after summary statistics
# df.describe(), df_clipped.describe()
df_clipped[numerical_cols] = df[numerical_cols].apply(
    lambda x: np.clip(x, lower_bounds[x.name], upper_bounds[x.name])
)

# Display summary statistics before and after clipping
df.describe(), df_clipped.describe()
```

```
(              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  ,
              Age   Education       Usage     Fitness        Income  \
count  180.000000  180.000000  180.000000  180.000000    180.000000
mean    28.641389   15.572222    3.396944    3.322222  53477.070000
std      6.446373    1.362017    0.952682    0.937461  15463.662523
min     20.000000   14.000000    2.000000    2.000000  34053.150000
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     43.050000   18.000000    5.050000    5.000000  90948.250000

            Miles
count  180.000000
mean   101.088889
std     43.364286
min     47.000000
25%     66.000000
50%     94.000000
75%    114.750000
max    200.000000  )
```

As you can see now if a value is below the 5th percentile, set it to the 5th percentile value and if a value is above the 95th percentile, set it to the 95th percentile value.

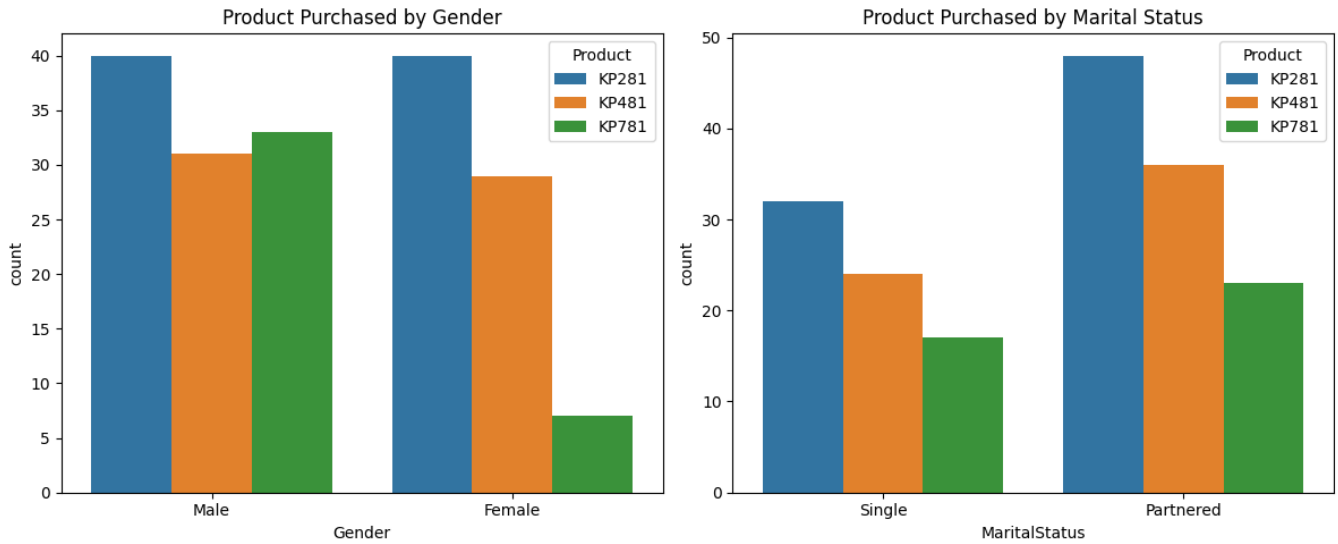
3. Check if features like marital status, Gender, and age have any effect on the product purchased

○ Find if there is any relationship between the categorical variables and the output variable in the data.

```
# Count plot for Gender vs Product
plt.figure(figsize=(12, 5))
plt.subplot(1, 2, 1)
sns.countplot(data=df, x="Gender", hue="Product")
plt.title("Product Purchased by Gender")

# Count plot for Marital Status vs Product
plt.subplot(1, 2, 2)
sns.countplot(data=df, x="MaritalStatus", hue="Product")
plt.title("Product Purchased by Marital Status")

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



The count plots reveal:

Gender vs. Product: The distribution of treadmill purchases varies across genders. We can observe if certain models are preferred more by males or females.

Marital Status vs. Product: The purchase pattern differs between single and partnered individuals.

○ Find if there is any relationship between the continuous variables and the output variable in the data.

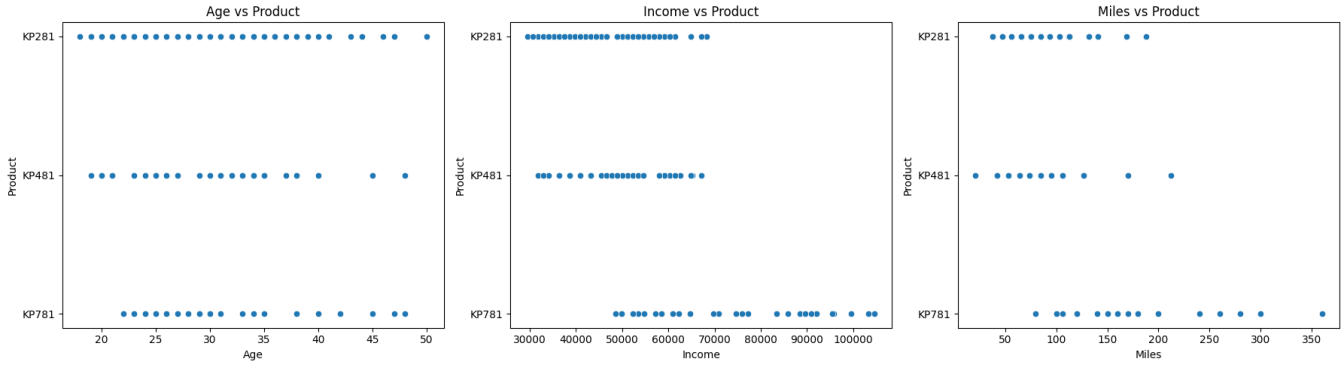
```
# Scatter plots for continuous variables vs Product using scatter plot
plt.figure(figsize=(18, 5))

# Age vs Product
plt.subplot(1, 3, 1)
sns.scatterplot(data=df, x="Age", y="Product")
plt.title("Age vs Product")

# Income vs Product
plt.subplot(1, 3, 2)
sns.scatterplot(data=df, x="Income", y="Product")
plt.title("Income vs Product")

# Miles vs Product
plt.subplot(1, 3, 3)
sns.scatterplot(data=df, x="Miles", y="Product")
plt.title("Miles vs Product")

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



The scatter plots show:

Age vs. Product: Different age groups may have preferences for specific treadmill models.

Income vs. Product: Higher or lower income levels might be associated with specific product choices.

Miles vs. Product: Customers who run more miles per week may favor particular treadmills.

4. Representing the Probability

○ Find the marginal probability (what percent of customers have purchased KP281, KP481, or KP781)

```
product_counts = pd.crosstab(index=df["Product"], columns="Count", normalize=Tr
product_counts
```

col_0	Count
Product	
KP281	44.444444
KP481	33.333333
KP781	22.222222

It shows what percentage of total customers purchased each treadmill model.

Find the probability that the customer buys a product based on each column.

```
prob_by_gender = pd.crosstab(df["Gender"], df["Product"], normalize="index") *
prob_by_marital_status = pd.crosstab(df["MaritalStatus"], df["Product"], normal
prob_by_education = pd.crosstab(df["Education"], df["Product"], normalize="inde
prob_by_gender, prob_by_marital_status, prob_by_education
```

```
(Product      KP281      KP481      KP781
 Gender
 Female   52.631579  38.157895   9.210526
 Male     38.461538  29.807692  31.730769,
 Product            KP281      KP481      KP781
 MaritalStatus
 Partnered      44.859813  33.644860  21.495327
 Single         43.835616  32.876712  23.287671,
 Product        KP281      KP481       KP781
 Education
 12         66.666667  33.333333    0.000000
 13         60.000000  40.000000    0.000000
 14         54.545455  41.818182    3.636364
 15         80.000000  20.000000    0.000000
 16         45.882353  36.470588   17.647059
 18          8.695652   8.695652   82.608696
 20          0.000000   0.000000  100.000000
 21          0.000000   0.000000  100.000000)
```

Gender	KP281 (%)	KP481 (%)	KP781 (%)
Female	52.63	38.16	9.21
Male	38.46	29.81	31.73

Females are more likely to purchase KP281 (52.63%) and less likely to buy KP781.

Males have a more even distribution, but KP781 is more popular among them (31.73%).

Marital Statu	KP281 (%)	KP481 (%)	KP781 (%)
Partnered	44.86	33.64	21.5
Single	43.84	32.88	23.29

Partnered customers slightly prefer KP281 (44.86%) over others.

Single customers have a similar trend but with a higher percentage choosing KP781 (23.29%).

```
prob_by_age = pd.crosstab(df["Age"], df["Product"], normalize="index") * 100
prob_by_income = pd.crosstab(df["Income"], df["Product"], normalize="index") *

# Display results
prob_by_age, prob_by_income
```

```
(Product       KP281      KP481       KP781
Age
18       100.000000   0.000000    0.000000
19        75.000000  25.000000    0.000000
20        40.000000  60.000000    0.000000
21        57.142857  42.857143    0.000000
22        57.142857   0.000000   42.857143
23        44.444444  38.888889   16.666667
24        41.666667  25.000000   33.333333
25        28.000000  44.000000   28.000000
26        58.333333  25.000000   16.666667
27        42.857143  14.285714   42.857143
28        66.666667   0.000000   33.333333
29        50.000000  16.666667   33.333333
30        28.571429  28.571429   42.857143
31        33.333333  50.000000   16.666667
32        50.000000  50.000000    0.000000
33        25.000000  62.500000   12.500000
34        33.333333  50.000000   16.666667
35        37.500000  50.000000   12.500000
36       100.000000   0.000000    0.000000
37        50.000000  50.000000    0.000000
38        57.142857  28.571429   14.285714
39       100.000000   0.000000    0.000000
40        20.000000  60.000000   20.000000
41       100.000000   0.000000    0.000000
42         0.000000   0.000000  100.000000
43       100.000000   0.000000    0.000000
44       100.000000   0.000000    0.000000
45         0.000000  50.000000   50.000000
46       100.000000   0.000000    0.000000
47        50.000000   0.000000   50.000000
48         0.000000  50.000000   50.000000
50       100.000000   0.000000    0.000000,
Product  KP281  KP481  KP781
Income
29562    100.0    0.0    0.0
30699    100.0    0.0    0.0
31836     50.0   50.0    0.0
32973     60.0   40.0    0.0
34110     40.0   60.0    0.0
...        ...    ...    ...
95508      0.0    0.0  100.0
95866      0.0    0.0  100.0
99601      0.0    0.0  100.0
103336     0.0    0.0  100.0
104581     0.0    0.0  100.0

[62 rows x 3 columns])
```

1. people with income in range 29k$to 30.7k$ prefer to buy KP281.
2. People with higher income from range 95k$to 105k$ prefer to buy KP781
3. People with medium income prefer to buy KP281 & KP481

Find the conditional probability that an event occurs given that another event has occurred. (Example: given that a customer is female, what is the probability she'll purchase a KP481)

Conditional Probabilities:

Given that a customer is Female: 52.63% chance of purchasing KP281 38.16% chance of purchasing KP481 9.21% chance of purchasing KP781

Given that a customer is Male: 38.46% chance of purchasing KP281 29.81% chance of purchasing KP481 31.73% chance of purchasing KP781

Given that a customer is Partnered: 44.86% chance of purchasing KP281 33.64% chance of purchasing KP481 21.50% chance of purchasing KP781

Given that a customer is Single: 43.84% chance of purchasing KP281 32.88% chance of purchasing KP481 23.29% chance of purchasing KP781

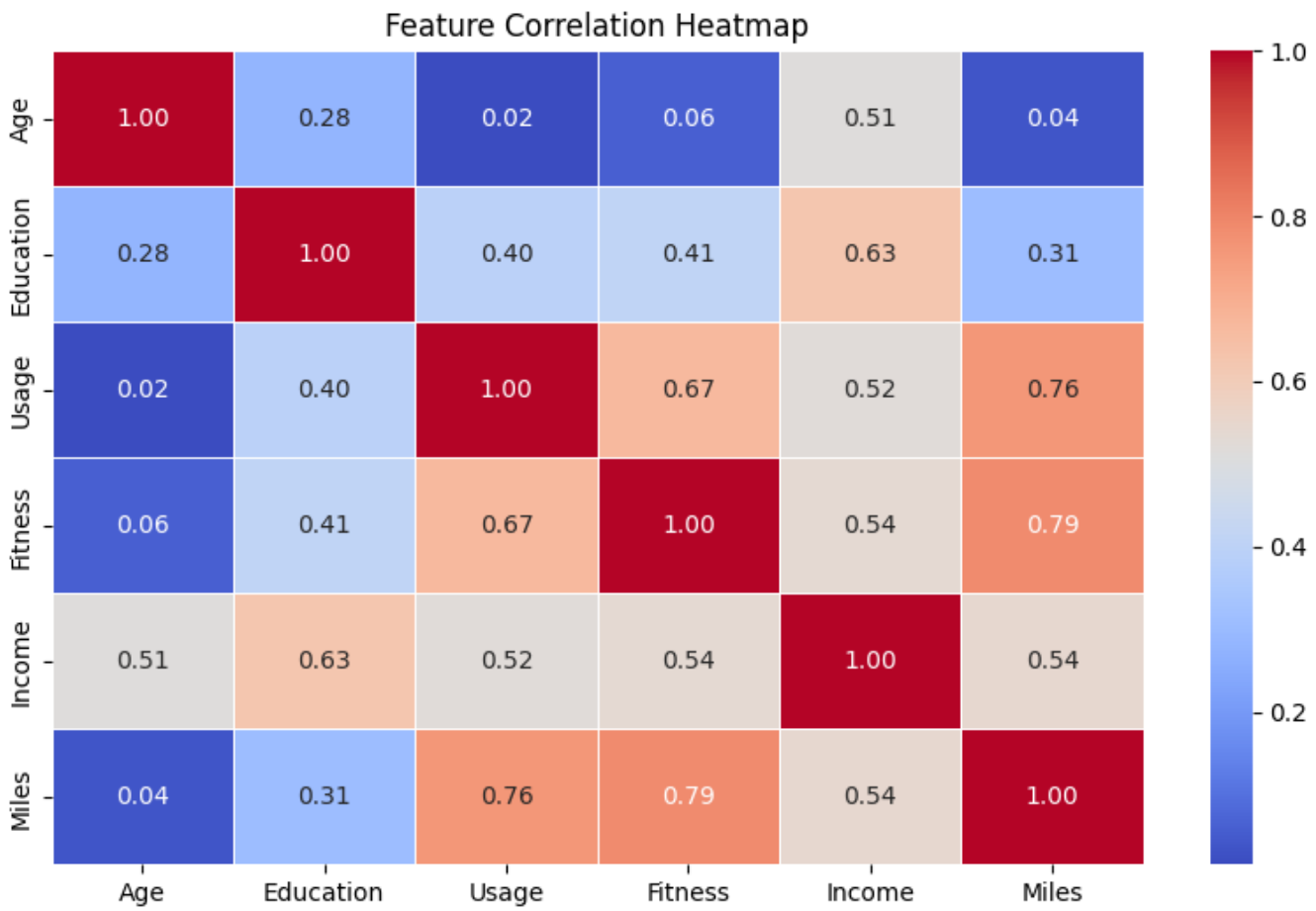
Similar thing is applicable for other columns like age, income. So I am not mentioning it

5. Check the correlation among different factors

```
numerical_df = df.select_dtypes(include=['number'])

# Compute correlation matrix
correlation_matrix = numerical_df.corr()

# Plot heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap="coolwarm", fmt=".2f", linewic
plt.title("Feature Correlation Heatmap")
plt.show()
```



Summary of Correlation Heatmap:

```
1    Strong Positive Correlations:

◦    Usage & Fitness (0.67): Customers who use the treadmill more frequently tend to h
◦    Miles & Fitness (0.79): Higher fitness levels are strongly associated with more m
◦    Miles & Usage (0.76): The more frequently a treadmill is used, the more miles are
◦    Education & Income (0.63): Higher education levels are linked to higher income.

2    Moderate Positive Correlations:

◦    Income & Fitness (0.54): Customers with higher income tend to have a better fitne
◦    Income & Miles (0.54): Higher-income customers tend to run more miles.
◦    Education & Fitness (0.41): More educated individuals tend to have a better fitne

3    Weak or No Correlation:

◦    Age & Miles (0.04): Age has almost no impact on the number of miles run.
◦    Age & Usage (0.02): Age does not significantly affect treadmill usage.
◦    Age & Fitness (0.06): Fitness level does not show a strong relationship with age.
```

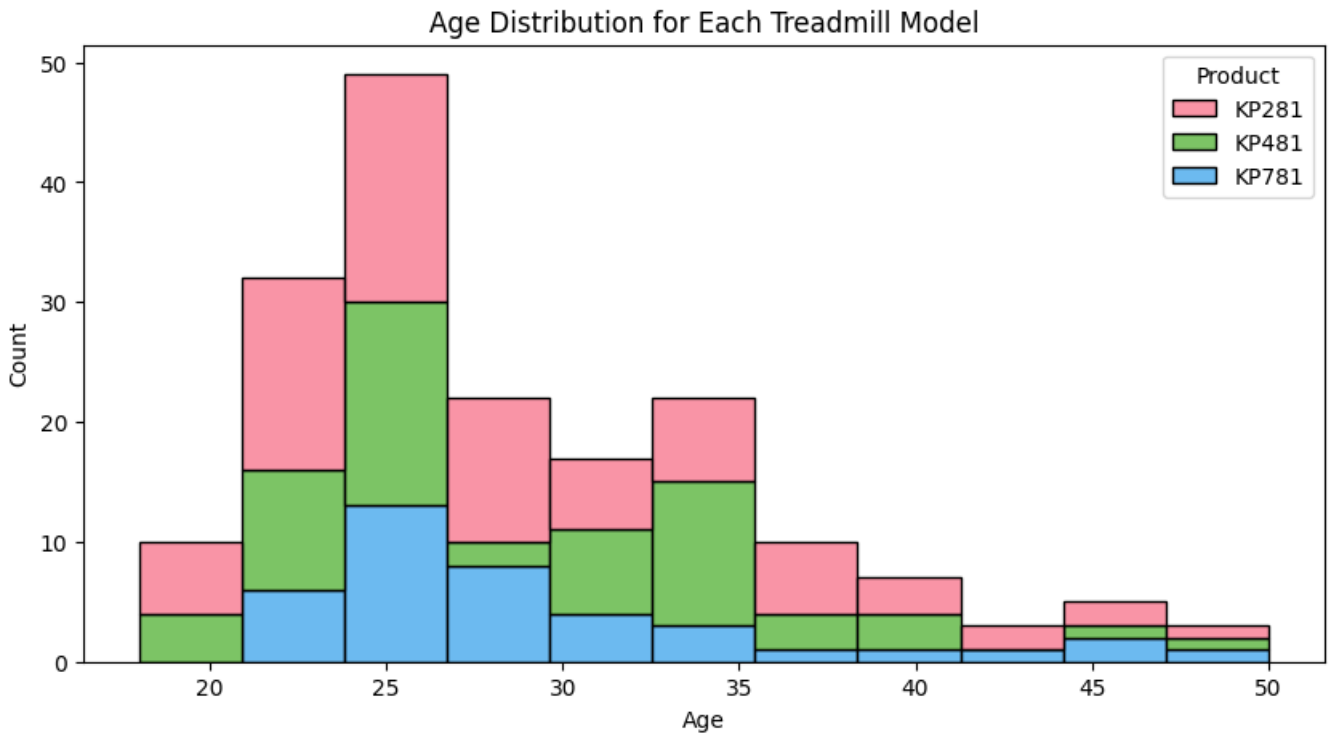
Key Takeaways:

```
•    Fitness, treadmill usage, and miles run are highly correlated. This suggests that
•    Income and education levels influence fitness and treadmill usage. Higher-income,
•    Age has minimal impact on treadmill usage or fitness. This indicates that people
```

6. Customer profiling and recommendation

○ Make customer profilings for each and every product.

```
plt.figure(figsize=(10, 5))
sns.histplot(data=df, x="Age", hue="Product", fill=True,multiple='stack', palet
plt.title("Age Distribution for Each Treadmill Model")
plt.xlabel("Age")
plt.ylabel("Count")
# plt.legend(title="Product")
plt.show()
```



Insights from the Age Distribution Graph

1️⃣ Young Customers (20-30 years) Dominate Sales

- The largest group of buyers falls between 20–30 years old, especially around 25 y
- This suggests younger individuals are the main target audience for these treadmil

2️⃣ KP281 (Pink) is the Most Popular Model

- Across most age groups, KP281 is the most purchased treadmill.
- This could indicate that KP281 is an entry-level or budget-friendly model preferr

3️⃣ KP481 (Green) and KP781 (Blue) are Purchased Less Frequently

- KP481 has a moderate number of buyers across different age groups.
- KP781 (Blue) has the least number of buyers, suggesting it might be a premium tre

4️⃣ Sales Drop Significantly After Age 35

- There is a steep decline in purchases after age 35, showing that older customers
- Marketing efforts should focus on convincing older customers of the benefits of h

Possible Recommendations

✅ Target young adults (20-30 years old) with fitness campaigns and promotions.

✅ KP281 is the most popular → Market it as a "starter treadmill" for beginners.

✅ KP781 has low sales → If it's a premium product, highlight its advanced features more.

✅ Encourage older adults (35+) to purchase by focusing on health benefits.

```
plt.figure(figsize=(10, 5))
sns.scatterplot(data=df, x="Age", y="Income", hue="Product", palette="husl", al
plt.title("Income vs. Age for Each Treadmill Model")
plt.xlabel("Age")
plt.ylabel("Income")
# plt.legend(title="Product")
plt.show()
```

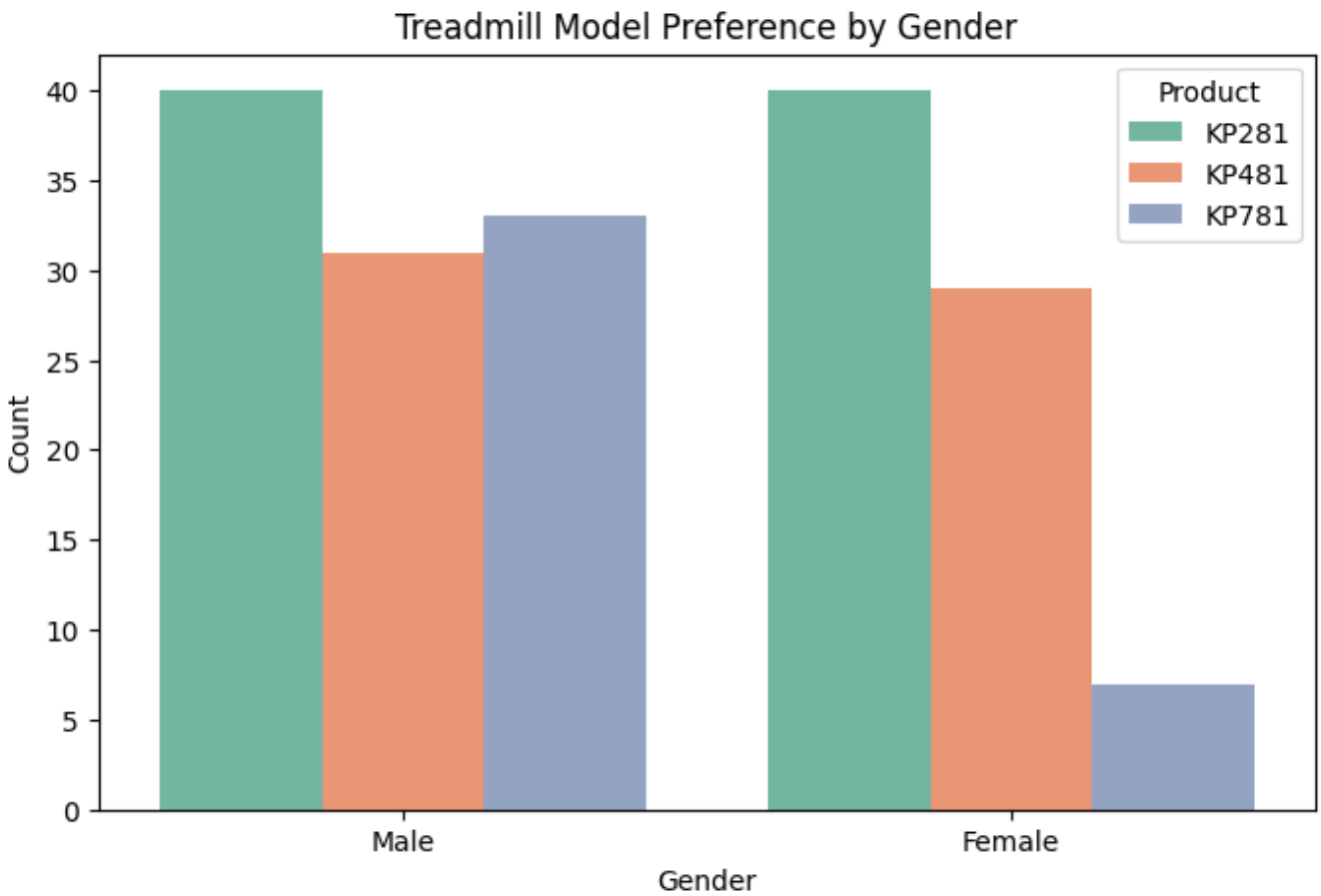
✅ Insights to Look For:

Are higher-income customers buying premium models like KP781?

Are younger customers (20-30) buying budget-friendly models?

Are middle-aged (40-50) individuals investing in mid-range models?

```
plt.figure(figsize=(8, 5))
sns.countplot(data=df, x="Gender", hue="Product", palette="Set2")
plt.title("Treadmill Model Preference by Gender")
plt.xlabel("Gender")
plt.ylabel("Count")
plt.legend(title="Product")
plt.show()
```

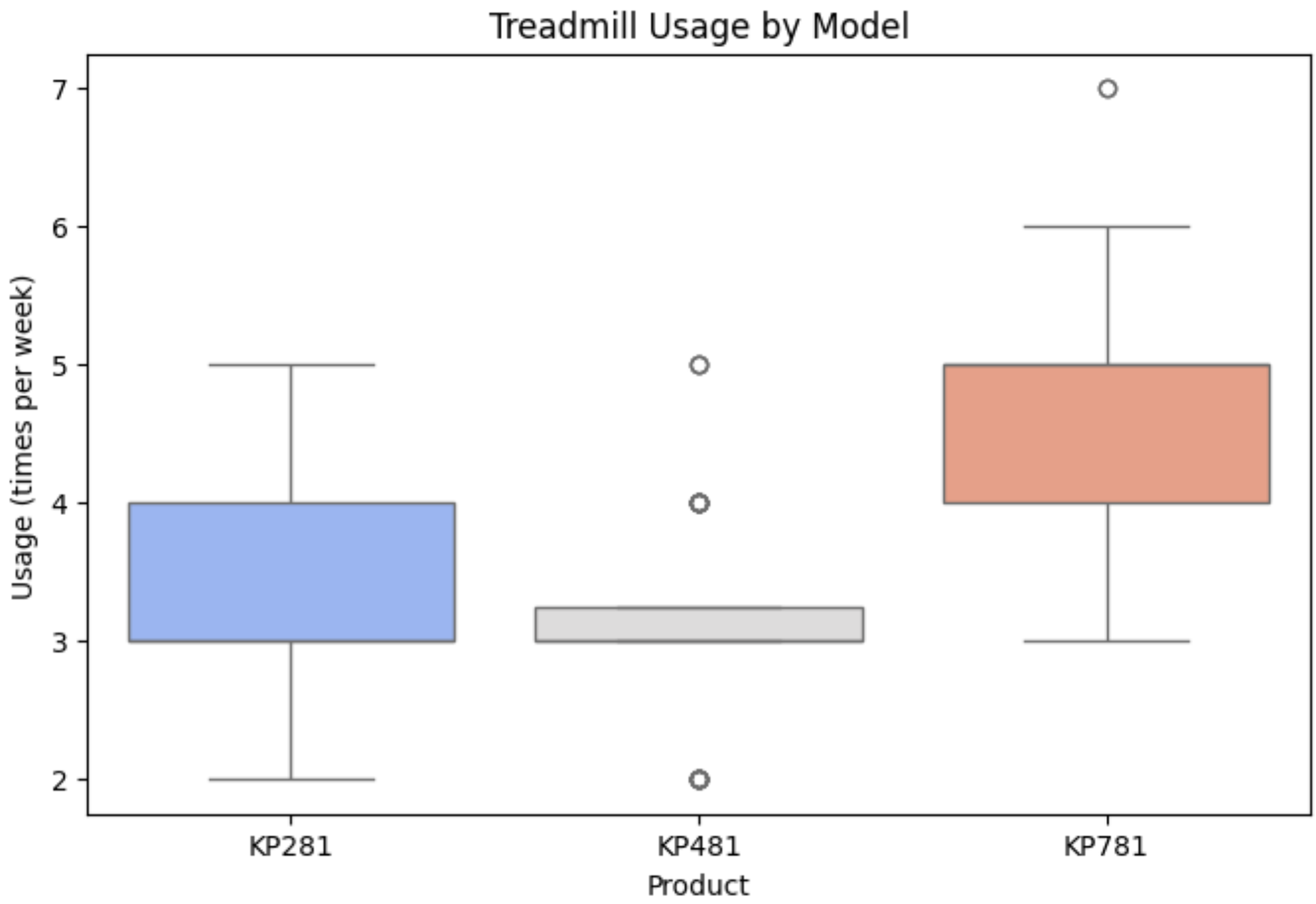


✅ Insights to Look For:

Do men and women prefer different models?

If one model is heavily male/female-dominated, what could be the reason? (E.g., power, durability, or features like incline settings).

```
plt.figure(figsize=(8, 5))
sns.boxplot(data=df, x="Product", y="Usage", palette="coolwarm")
plt.title("Treadmill Usage by Model")
plt.xlabel("Product")
plt.ylabel("Usage (times per week)")
plt.show()
```



✅ Insights to Look For:

If a model has a higher median usage, it may be preferred by gym-goers or fitness enthusiasts.

If a model has low but consistent usage, it might be used for casual workouts at home.

Detailed report -

1️⃣ Target Audience and Demographics

A. Focus on the 20-30 Age Group

✅ The majority of treadmill buyers fall between the ages of 20 to 30, with a peak around 25 years old.

✅ This indicates that younger individuals are the primary consumers, likely due to their focus on fitness, health, and home workout routines.

✅ Recommendation:

Run targeted digital marketing campaigns on platforms like Instagram, TikTok, and YouTube, emphasizing the benefits of treadmills for weight management and overall fitness. Collaborate with fitness influencers who resonate with young audiences.

Offer student discounts or young professional fitness plans to attract first-time buyers.

B. Engage the 35+ Age Group with Health-Focused Messaging

❌ Sales drop significantly after the age of 35, meaning older customers are not engaging as much with Aerofit products.

✅ The 35+ group is often more health-conscious, with increasing concerns about lifestyle diseases (e.g., diabetes, heart conditions).

✅ Recommendation:

Create campaigns emphasizing health benefits, such as reducing cardiovascular risks and improving joint mobility.

Feature real-life testimonials from middle-aged users who have benefited from treadmill workouts.

Partner with health professionals and physiotherapists to promote the benefits of regular treadmill usage for older adults.

2️⃣ Product-Specific Insights and Marketing Strategy

A. KP281 – The Best-Selling Model (Entry-Level Treadmill) ✅ KP281 is the most popular choice across all age groups, especially among young customers.

✅ This suggests it is likely an entry-level or budget-friendly treadmill.

✅ Recommendation:

Position KP281 as the "Perfect First Treadmill" for new fitness enthusiasts.

Offer bundle deals (e.g., KP281 + fitness tracker + workout guide).

Highlight affordable pricing and easy EMI options for first-time buyers.

B. KP481 – A Balanced Choice for Serious Users

✅ KP481 is moderately purchased across different age groups, suggesting that it is a mid-range treadmill preferred by those with a greater fitness commitment.

✅ Recommendation:

Market KP481 to serious fitness enthusiasts who want an upgrade from a basic treadmill.

Focus on durability, enhanced features, and better cushioning for long-term workouts.

Offer comparison content showing why KP481 is a better choice than KP281.

C. KP781 – The Least Purchased Model (Premium Treadmill)

❌ KP781 has the lowest number of buyers, suggesting it is a high-end treadmill with a niche audience.

✅ This could be due to higher pricing or advanced features that appeal to only a small segment.

✅ Recommendation:

Position KP781 as a luxury product targeted at high-income professionals and serious athletes.

Offer premium installation services and personal training sessions as an add-on.

Use high-quality video advertisements showcasing top-tier features and why it's worth the investment.

3️⃣ Gender-Based Marketing Strategy

✅ Data suggests that males and females purchase treadmills at different rates.

✅ Recommendation:

Run separate marketing campaigns for men and women, highlighting benefits that appeal to each gender.

For women, focus on weight loss, home workouts, and convenience.

For men, emphasize muscle endurance, cardio improvement, and performance training.

4️⃣ Income-Based Customer Profiling

✅ Higher-income individuals tend to prefer premium models (KP781 & KP481), while lower-income buyers gravitate toward KP281.

✅ Recommendation:

Introduce "Buy Now, Pay Later" EMI plans for mid-range and premium models.

Offer discounts for fitness clubs, corporate wellness programs, and gym owners.

Promote KP781 as a long-term investment in fitness rather than a one-time purchase.

5️⃣ Sales Strategy and Promotions

📌 Leverage Seasonal Discounts: Offer big discounts during New Year, Black Friday, and fitness months (January, April, September).

📌 Run Referral Programs: Encourage existing customers to refer friends and earn discounts on accessories.

📌 Provide After-Sales Services: Offer free maintenance checks for 6 months to retain customers.

Final Thoughts

🔷 Aerofit should focus on young adults (20-30 years old) for the majority of sales.

🔷 KP281 should be heavily promoted as the best option for beginners.

🔷 Older adults (35+) should be encouraged through health-focused marketing.

🔷 KP781 should be marketed as a luxury treadmill for high-income individuals.

🔷 Flexible pricing, financing options, and strategic campaigns will increase sales across all segments.

End of Case Study