About Aerofit

Aerofit is a leading brand in the fitness equipment industry, dedicated to promoting health and wellness through high-quality products. With a diverse range of machines, including treadmills, exercise bikes, and comprehensive gym equipment, Aerofit caters to the fitness needs of all individuals, from beginners to seasoned athletes. Our fitness accessories further enhance the workout experience, ensuring that everyone can achieve their fitness goals with ease and efficiency. At Aerofit, we are committed to innovation, durability, and customer satisfaction, making us a trusted name in fitness worldwide.

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, in order to provide better recommendations to new customers. To achieve this, the team plans to investigate whether there are differences across the products with respect to customer characteristics. This involves analyzing customer demographics, purchase behavior, and preferences.

Objectives:

Customer Profiling: Perform descriptive analytics to create a detailed customer profile for each AeroFit treadmill product (KP281, KP481, and KP781). Develop appropriate tables and charts to illustrate these profiles.

Contingency Analysis: Construct two-way contingency tables for each AeroFit treadmill product to explore the relationship between different customer characteristics and treadmill purchases.

Probability Computation: Compute all conditional and marginal probabilities from the contingency tables. Analyze these probabilities to derive insights on customer behavior and preferences.

Business Insights: Provide insights on how these customer profiles and probabilities can impact business strategies, such as targeted marketing, product recommendations, and inventory management.

Dataset:

The company has collected data on individuals who purchased a treadmill from AeroFit stores during the prior three months. The dataset includes the following variables:

Product Purchased: KP281, KP481, or KP781

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

KP281: Entry-level treadmill that sells for $1,500
KP481: Mid-level treadmill that sells for $1,750
KP781: Advanced treadmill with additional features that sells for $2,500

Content:

1. **Import the Dataset and Perform Initial Analysis**
2. **Detect Outliers**
3. **Analyze the Effect of Features on Product Purchased**
4. **Represent Marginal Probabilities**
5. **Correlation Analysis**
6. **Customer Profiling and Probabilities**
7. **Insights and Recommendations**

1. Import the Dataset and Perform Initial Analysis:

Input:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from matplotlib.ticker import FuncFormatter

# Load the dataset
Aerofit_raw = pd.read_csv('/content/aerofit_treadmill.csv')
# Display first few rows
Aerofit_raw.sample(10)
```

Output:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
52	KP281	29	Female	16	Partnered	4	3	50028	94
1	KP281	19	Male	15	Single	2	3	31836	75
7	KP281	21	Male	13	Single	3	3	32973	85
84	KP481	21	Female	14	Partnered	5	4	34110	212
176	KP781	42	Male	18	Single	5	4	89641	200
164	KP781	28	Male	18	Single	6	5	88396	150
163	KP781	28	Male	18	Partnered	7	5	77191	180
135	KP481	40	Female	16	Partnered	3	3	61398	85
50	KP281	29	Male	18	Partnered	3	3	68220	85
113	KP481	30	Female	14	Single	3	3	57987	74

Input:

```python
# Display the shape of the Aerofit_raw DataFrame
print(f"The shape of the Aerofit_raw DataFrame is:
{Aerofit_raw.shape}")
```

Output:

The shape of the Aerofit_raw DataFrame is: (180, 9)

Input:

```python
# Display data types and missing values
print(Aerofit_raw.info())
```

Output:

```
<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
None
```

Input:

```
# Summary statistics
Aerofit_raw.describe(include="all")
```

Output:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
count	180	180.000000	180	180.000000	180	180.000000	180.000000	180.000000	180.000000
unique	3	NaN	2	NaN	2	NaN	NaN	NaN	NaN
top	KP281	NaN	Male	NaN	Partnered	NaN	NaN	NaN	NaN
freq	80	NaN	104	NaN	107	NaN	NaN	NaN	NaN
mean	NaN	28.788889	NaN	15.572222	NaN	3.455556	3.311111	53719.577778	103.194444
std	NaN	6.943498	NaN	1.617055	NaN	1.084797	0.958869	16506.684226	51.863605
min	NaN	18.000000	NaN	12.000000	NaN	2.000000	1.000000	29562.000000	21.000000
25%	NaN	24.000000	NaN	14.000000	NaN	3.000000	3.000000	44058.750000	66.000000
50%	NaN	26.000000	NaN	16.000000	NaN	3.000000	3.000000	50596.500000	94.000000
75%	NaN	33.000000	NaN	16.000000	NaN	4.000000	4.000000	58668.000000	114.750000
max	NaN	50.000000	NaN	21.000000	NaN	7.000000	5.000000	104581.000000	360.000000

Input:

```
# Display data types and missing values
missing_values = Aerofit_raw.isnull().sum()
missing_values
```

Output:

```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
Miles            0
dtype: int64
```

Insights:

- ❖ **Product Pricing and Segmentation:**
  - ➢ KP281 ($1,500) is an entry-level treadmill and the most popular product, indicating a preference for affordability.
  - ➢ KP481 ($1,750) caters to mid-level runners.
  - ➢ KP781 ($2,500) is the most expensive, offering advanced features, which suggests it is targeted at serious or professional users.
- ❖ **Age Distribution:**
  - ➢ Minimum age: 18
  - ➢ Maximum age: 50
  - ➢ Mean age: 28.79
  - ➢ 75% of the customers are 33 years old or younger, indicating a younger customer base.
- ❖ **Gender Distribution:**
  - ➢ Out of 180 data points, 104 are male (57.78%), and 76 are female (42.22%). This suggests a male-dominated customer base.
- ❖ **Income and Miles Variability:**
  - ➢ The high standard deviation in income and miles suggests a wide range of income levels and expected treadmill usage among customers, potentially indicating diverse customer segments.

2. Detect Outliers

Input:

```python
import matplotlib.pyplot as plt
import seaborn as sns

# List of numerical columns
numerical_cols = ['Age', 'Education', 'Usage', 'Income', 'Fitness', 'Miles']

# Create subplots
fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(10, 8))

# Flatten axes for easy iteration
axes = axes.flatten()

# Generate a boxplot for each numerical column
for i, col in enumerate(numerical_cols):
    sns.boxplot(x=Aerofit_raw[col], ax=axes[i])
    axes[i].set_title(f'Boxplot for {col}')

# Adjust layout
plt.tight_layout()

# Show plot
plt.show()

# Checking difference between mean and median
print(Aerofit_raw[numerical_cols].mean() -
Aerofit_raw[numerical_cols].median())
```

Output:
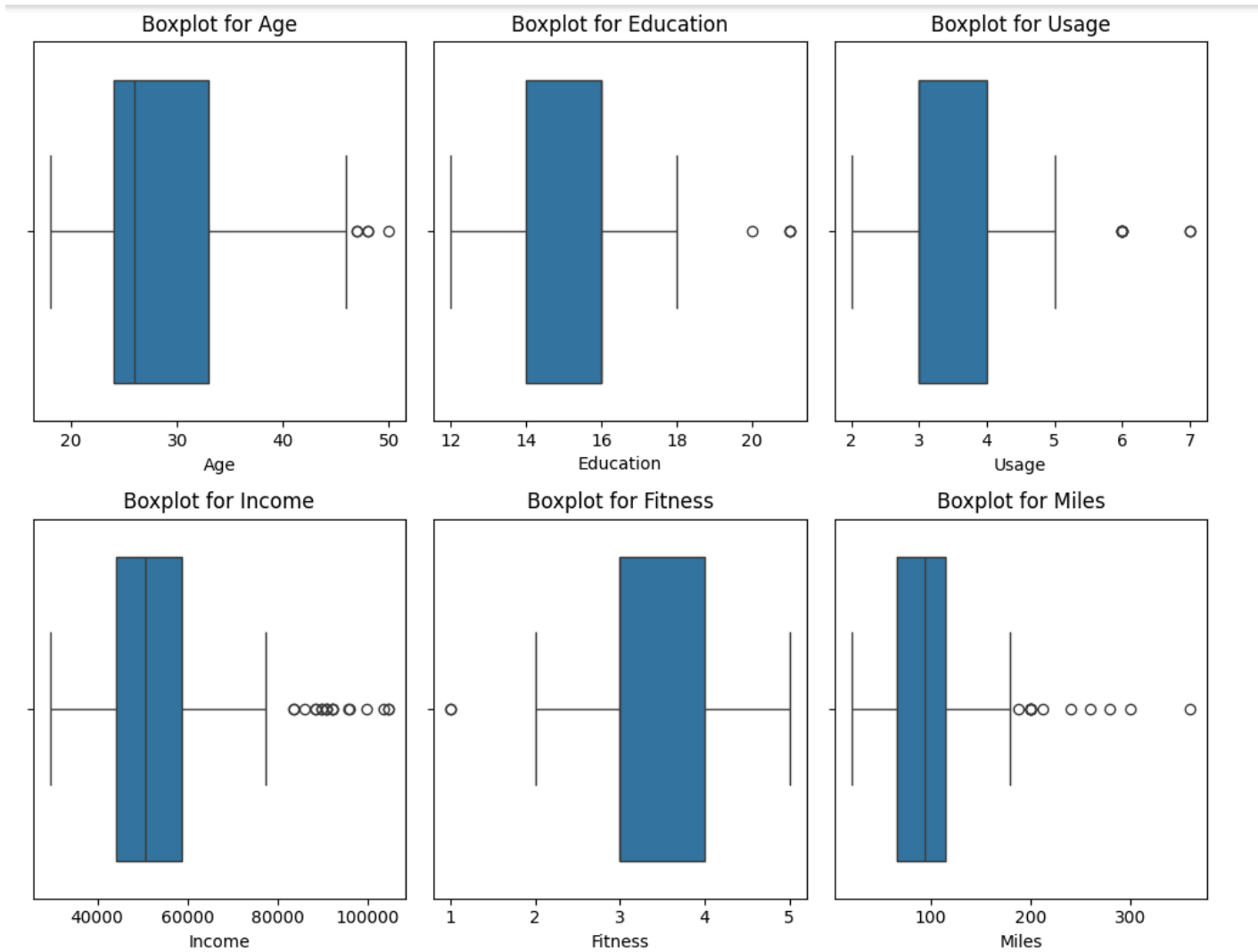

```
Age                2.788889
Education         -0.427778
Usage              0.455556
Income          3123.077778
Fitness            0.311111
Miles              9.194444
dtype: float64
```

Insights:

❖ **Outliers:**
  ➢ Age, Education, and Usage: The presence of few outliers suggests a relatively normal distribution for these variables, which means they are consistent and not heavily skewed.
  ➢ Income and Miles: A higher number of outliers indicates significant variability and potential extreme values. This can skew the data and affect the analysis.
❖ **Difference Between Mean and Median:**
  ➢ Age (2.788889): The mean is slightly higher than the median, suggesting a slight right skew. This means there are a few older individuals pulling the mean higher.
  ➢ Education (-0.427778): The mean is slightly lower than the median, suggesting a slight left skew. This indicates a few individuals with fewer years of education are pulling the mean lower.
  ➢ Usage (0.455556): The mean is slightly higher than the median, indicating a minor right skew. A few individuals plan to use the treadmill significantly more than the average.
  ➢ Income (3123.077778): The substantial difference between mean and median indicates a strong right skew. A few high-income individuals are pulling the mean up significantly.
  ➢ Fitness (0.311111): The small difference suggests a near-normal distribution for fitness ratings.
  ➢ Miles (9.1944444): A notable difference indicates a right skew, with a few individuals expecting to use the treadmill significantly more.

Recommendations:

❖ **Addressing Outliers:**
  ➢ Income and Miles: Perform a detailed analysis to identify and understand the nature of these outliers. Consider using robust statistical methods, such as trimming or winsorizing the data, to reduce the impact of outliers.
  ➢ For marketing, segment these outliers separately and develop tailored strategies to cater to their specific needs.
❖ **Data Transformation:**
  ➢ For highly skewed data like income and miles, consider data transformation techniques (e.g., log transformation) to normalize the distribution. This can make statistical analyses more reliable and insights more accurate.
❖ **Targeted Marketing and Product Development:**
  ➢ Age: Given the slight right skew, consider creating products or marketing strategies that appeal to slightly older demographics within your current age range.
  ➢ Education and Usage: With education being slightly left-skewed, focus on making your products more accessible to individuals with lower educational backgrounds. For usage, highlight features that cater to both casual and frequent users.
  ➢ Implement financing options and flexible payment plans to make higher-end products more accessible.

3. Analyze the Effect of Features on Product Purchased

Input:

```python
# Create a figure with 1 row and 2 columns
fig, axes = plt.subplots(nrows=1, ncols=2, figsize=(10, 6))

# First plot: Product Purchased by Marital Status
sns.countplot(x='Product', hue='MaritalStatus', data=Aerofit_raw,
ax=axes[0])
axes[0].set_title('Product Purchased by Marital Status')

# Second plot: Age Distribution by Product Purchased
sns.histplot(Aerofit_raw, x='Age', hue='Product', multiple='stack',
ax=axes[1])
axes[1].set_title('Age Distribution by Product Purchased')

# Adjust layout to prevent overlap
plt.tight_layout()

# Show plot
plt.show()
```
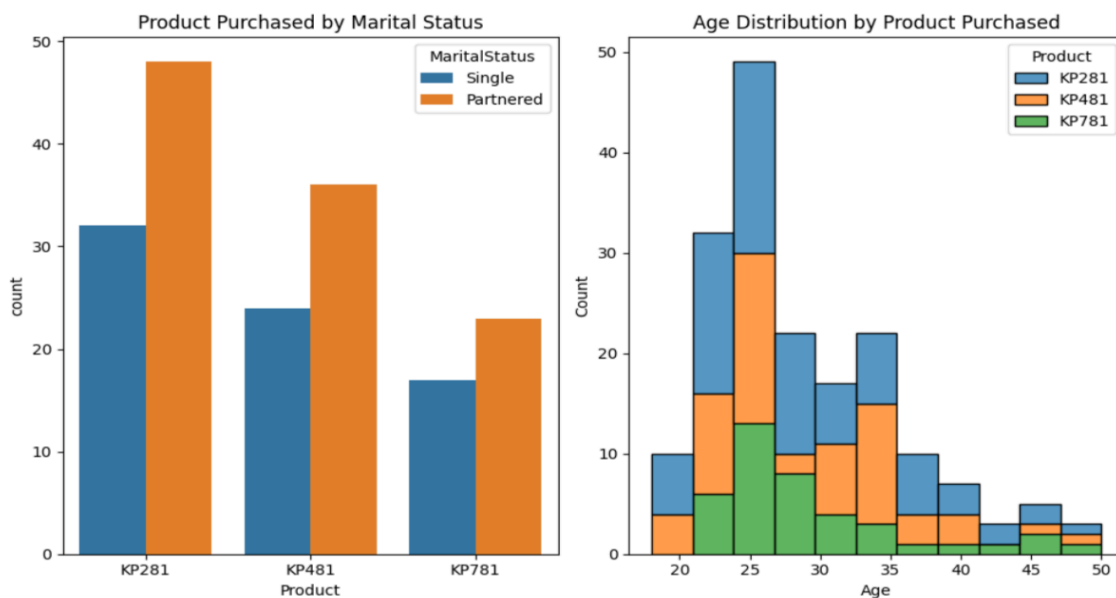
Output:

Input:

```python
# Defining the attributes
Attributes = ['Age', 'Education', 'Usage', 'Income', 'Fitness', 'Miles']
sns.set(style="whitegrid")

# Creating Subplots - 2 rows and 3 columns, adjusting the figure size for
better layout
fig, axes = plt.subplots(nrows=2, ncols=3, figsize=(10, 7))

# Adjust the layout to have more space between plots
fig.subplots_adjust(hspace=0.4, wspace=0.3, top=0.9)

count = 0

# Custom color palette for the products
custom_palette = {'KP281': 'blue', 'KP481': 'orange', 'KP781': 'green'}

# Function to format the y-axis labels for income
def format_income(value, tick_number):
    return f'{int(value / 1000)}k'

# Iterating through Subplots to create Boxplots
for i in range(2):
    for j in range(3):
        sns.boxplot(data=Aerofit_raw, x='Product', y=Attributes[count],
ax=axes[i, j], palette=custom_palette)
        axes[i, j].set_title(f'Product vs {Attributes[count]}', pad=8,
fontsize=13)
        # Format the y-axis for the 'Income' plot
        if Attributes[count] == 'Income':
            axes[i,
j].yaxis.set_major_formatter(FuncFormatter(format_income))
        count += 1

# Show the plot
plt.tight_layout()
plt.show()
```
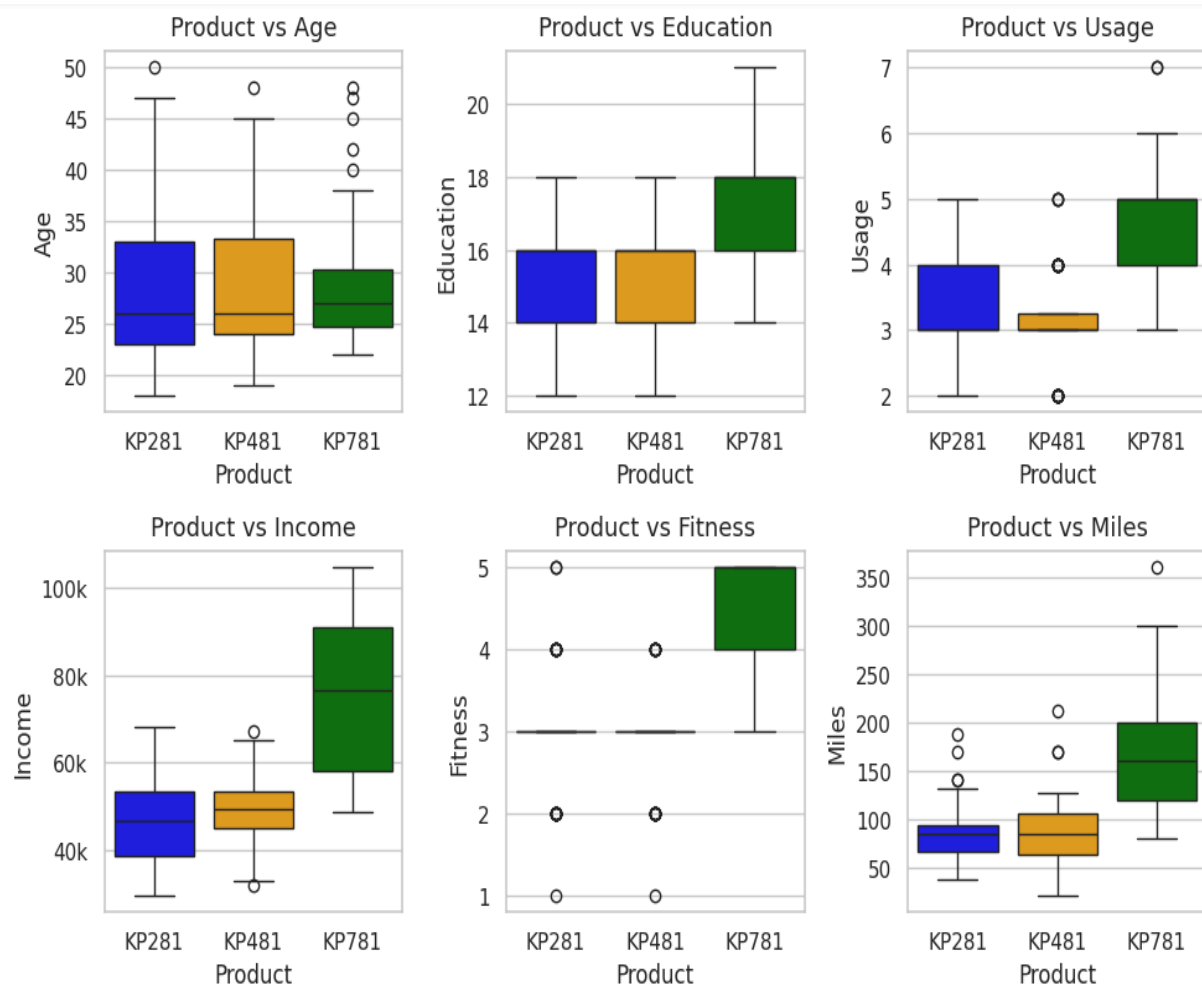
Output:

Insights:

❖ **Product vs Age**:
  - ➤ KP281 & KP481: These products have the same median age, indicating that customers purchasing these treadmills are typically of similar age.
  - ➤ KP781: Customers aged between 25-30 are more likely to purchase the KP781, suggesting this age group values the advanced features and higher performance of the KP781.

❖ **Product vs Education:**
  - ➤ KP781: Customers with education levels greater than 16 years are more likely to purchase this advanced treadmill, indicating a correlation between higher education and preference for advanced features.
  - ➤ KP281 & KP481: Customers with education levels less than 16 years have an equal likelihood of purchasing either of these products, suggesting that basic to mid-level features meet their needs.

❖ **Product vs Usage:**
  - ➤ KP781: Customers planning to use the treadmill more than 4 times a week are more likely to choose this product, indicating it is preferred by frequent users who likely seek durability and advanced functionality.
  - ➤ KP281 & KP481: Customers with lower usage plans are more likely to purchase these treadmills, suggesting that these models are sufficient for less frequent users.

❖ **Product vs Fitness:**
  - ➤ KP781: Customers with fitness levels rated 3 or higher are more likely to purchase this advanced treadmill, indicating that more fit individuals prefer treadmills with advanced features to support their fitness routines.

❖ **Product vs Income:**
  - ➤ KP781: Customers with an income of $60,000 or higher are more likely to purchase the KP781, indicating that higher-income individuals are willing to invest in more expensive, feature-rich treadmills.

❖ **Product vs Miles:**
  - ➤ KP781: Customers who expect to walk or run more than 120 miles per week are more likely to buy this product, suggesting that high mileage users prefer the advanced capabilities and durability of the KP781.

Recommendations:

❖ **Target Marketing by Age:**
  ➢ KP281 & KP481: Focus on marketing to a broad age range since the median age is similar. Emphasize the value and essential features of these treadmills.
  ➢ KP781: Develop targeted marketing campaigns for customers aged 25-30, highlighting the advanced features and performance benefits that appeal to this age group.

❖ **Education-Based Strategies:**
  ➢ KP781: Create content and advertising that appeals to customers with higher education levels, emphasizing the advanced technology and features.
  ➢ KP281 & KP481: Market these models as accessible and reliable options for customers with less than 16 years of education.

❖ **Usage-Focused Promotions:**
  ➢ KP781: Promote to frequent users, highlighting the treadmill's durability, advanced features, and suitability for regular, intense use.
  ➢ KP281 & KP481: Emphasize affordability and essential features for casual users who plan to use the treadmill less frequently.

❖ **Fitness Level Segmentation:**
  ➢ KP781: Target fitness enthusiasts with fitness levels of 3 or higher, using testimonials and success stories from highly fit individuals.
  ➢ KP281 & KP481: Highlight ease of use and essential benefits for those at lower fitness levels.

❖ **Income-Based Marketing:**
  ➢ KP781: Focus on high-income customers with tailored messaging that emphasizes the premium quality and advanced features of the treadmill.
  ➢ Offer financing options or promotions to make the KP781 more accessible to a broader audience.

❖ **High Mileage User Campaigns:**
  ➢ KP781: Create marketing campaigns targeting high mileage users, emphasizing the treadmill's robustness and advanced features that support extensive use.
  ➢ KP281 & KP481: Position these models as cost-effective solutions for moderate mileage users.

4. Represent Marginal Probabilities:

Input:

```
# Create a crosstab to count the occurrences of each product,
including margins
product_counts = pd.crosstab(index=Aerofit_raw['Product'],
columns='count', margins=True)
product_counts
```

Output:

col_0	count	All
Product		
KP281	80	80
KP481	60	60
KP781	40	40
All	180	180

Input:

```
# Calculate the probabilities by dividing the counts by the total sum of
counts (excluding the margin)
product_probs = product_counts / product_counts.loc['All', 'count']
print(product_probs)
```

Output:

```
col_0       count       All
Product
KP281    0.444444  0.444444
KP481    0.333333  0.333333
KP781    0.222222  0.222222
All      1.000000  1.000000
```

Insights:

- ❖ **Product Distribution:**
  - ➢ KP281: This entry-level treadmill accounts for 44.44% of the total products, making it the most popular choice among customers.
  - ➢ KP481: The mid-level treadmill makes up 33.33% of the total products, indicating a significant portion of customers prefer mid-range options.
  - ➢ KP781: The advanced treadmill represents 22.22% of the total products, showing that while it's the least common choice, there is still a notable demand for high-end features.
- ❖ **Overall Probabilities:**
  - ➢ The probabilities indicate the relative popularity of each product. KP281 has the highest probability, followed by KP481 and KP781.

Recommendations:
- ❖ **Inventory Management:**
  - ➢ KP281: Given its high popularity, ensure ample inventory to meet demand. Consider bulk purchasing discounts or promotions to maintain its strong sales.
  - ➢ KP481: Maintain a steady stock but monitor closely for any fluctuations in demand. It's crucial to balance inventory without overstocking.
  - ➢ KP781: While it's the least popular, it still has a significant customer base. Keep sufficient stock to cater to high-end customers but avoid overstocking.
- ❖ **Marketing Strategies:**
  - ➢ KP281: Focus on highlighting its affordability and value for money. Use testimonials and success stories from satisfied customers to attract new buyers.
  - ➢ KP481: Emphasize its suitability for regular runners who need more than basic features but are not ready to invest in high-end products. Position it as the best value for mid-range needs.
  - ➢ KP781: Market as a premium product with advanced features. Target high-income customers and fitness enthusiasts who are willing to invest in top-of-the-line equipment. Use targeted advertising on platforms frequented by these demographics.
- ❖ **Customer Segmentation:**
  - ➢ KP281: Target budget-conscious individuals, beginners, and casual users. Use digital marketing strategies like social media ads and email campaigns.
  - ➢ KP481: Focus on intermediate users who are consistent in their fitness routines but not necessarily professionals. Consider partnerships with fitness influencers to reach this audience.
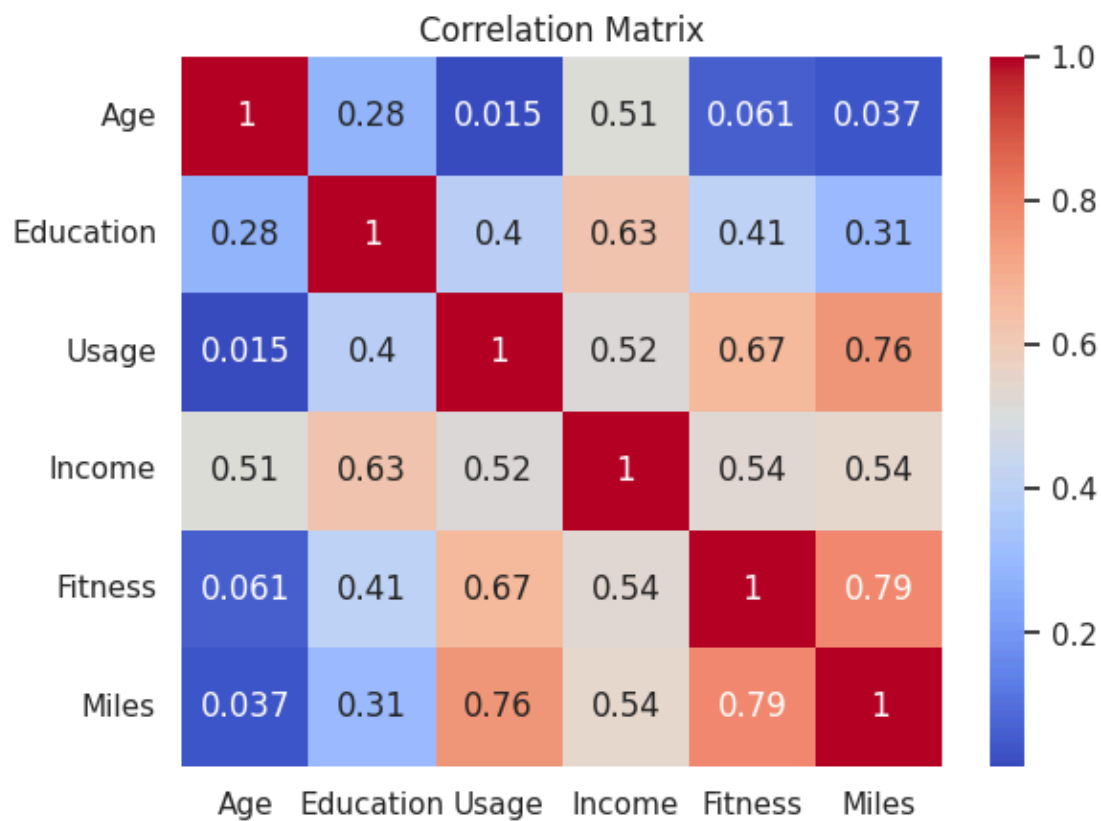
5. Correlation Analysis

Input:

```
# Creating the Correlation Matrix: HeatMap
corr_matrix = Aerofit_raw[numerical_cols].corr()
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```
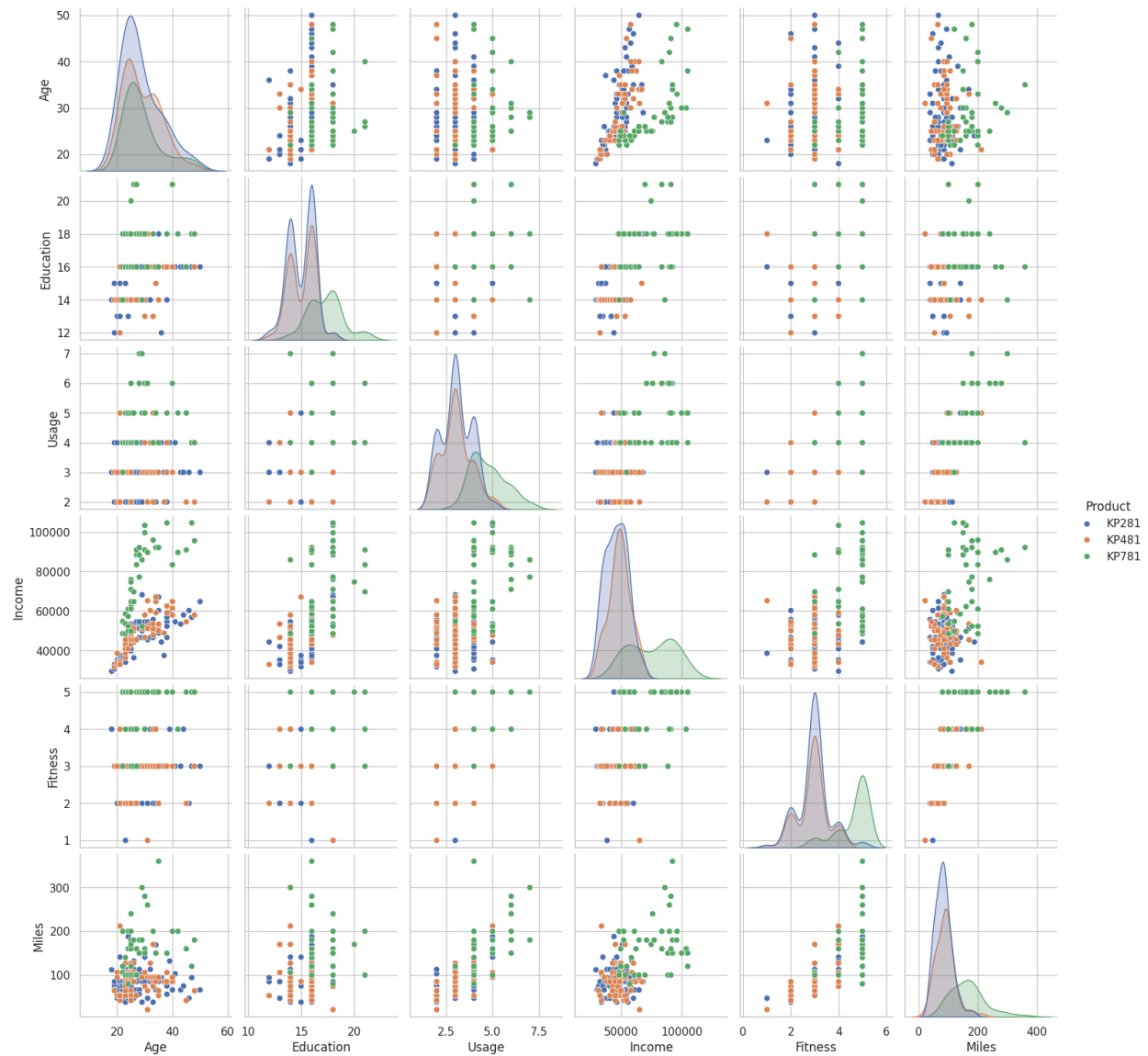
Output:

Input:

```python
# Creating pairplots based on products
sns.pairplot(Aerofit_raw, hue='Product', vars=numerical_cols)
plt.show()
```

Output:

6. Customer Profiling and Probabilities

Input:

```python
# Categorize Users:
profile = Aerofit_raw.groupby('Product').mean(numeric_only=True)
print(profile)
```

Output:

	Age	Education	Usage	Fitness	Income	Miles
Product						
KP281	28.55	15.037500	3.087500	2.9625	46418.025	82.787500
KP481	28.90	15.116667	3.066667	2.9000	48973.650	87.933333
KP781	29.10	17.325000	4.775000	4.6250	75441.575	166.900000

Input:

```python
# Conditional Probabilities:
prob_male_kp781 = len(Aerofit_raw[(Aerofit_raw['Gender'] == 'Male') &
(Aerofit_raw['Product'] == 'KP781')]) /
len(Aerofit_raw[Aerofit_raw['Gender'] == 'Male'])
print(f"Probability of a male customer buying a KP781 treadmill:
{prob_male_kp781:.2f}")
```

Output:

```
Probability of a male customer buying a KP781 treadmill: 0.32
```

Insights:

- ❖ **Product KP281:**
  - ➢ **Age:** Customers have an average age of 28.55, suggesting it appeals to younger adults.
  - ➢ **Education:** Average education level is 15.04 years, indicating customers typically have some college education.
  - ➢ **Usage:** Average planned usage is 3.09 times per week, showing moderate use.
  - ➢ **Fitness:** Average fitness level is 2.96, indicating customers consider themselves moderately fit.
  - ➢ **Income:** Average income is $46,418, suggesting it appeals to budget-conscious customers.
  - ➢ **Miles:** Average planned miles per week is 82.79, indicating moderate usage expectations.
- ❖ **Product KP481:**
  - ➢ **Age:** Customers have an average age of 28.90, slightly older than KP281 customers.
  - ➢ **Education:** Average education level is 15.12 years, similar to KP281.
  - ➢ **Usage:** Average planned usage is 3.07 times per week, similar to KP281.
  - ➢ **Fitness:** Average fitness level is 2.90, comparable to KP281.
  - ➢ Income: Average income is $48,973, slightly higher than KP281 customers.
  - ➢ Miles: Average planned miles per week is 87.93, slightly higher than KP281.
- ❖ **Product KP781:**
  - ➢ **Age:** Customers have an average age of 29.10, the oldest among the three groups.
  - ➢ **Education:** Average education level is 17.33 years, indicating a higher level of education.
  - ➢ **Usage:** Average planned usage is 4.78 times per week, showing high usage.
  - ➢ **Fitness:** Average fitness level is 4.63, indicating high self-rated fitness.
  - ➢ **Income:** Average income is $75,442, significantly higher than the other two groups.
  - ➢ **Miles:** Average planned miles per week is 166.90, indicating very high usage expectations.
- ❖ **Conditional Probability:**
  - ➢ The probability of a male customer buying a KP781 treadmill is 0.32, indicating that about 32% of male customers prefer the KP781 model.

7. Overall Insights, Recommendations and Conclusion:

Insights:

- ❖ **Product Pricing and Segmentation:**
    - ➢ KP281 ($1,500) is the most popular, suggesting a preference for affordability.
    - ➢ KP481 ($1,750) targets mid-level runners.
    - ➢ KP781 ($2,500) is the most expensive and appeals to serious or professional users.
- ❖ **Customer Demographics:**
    - ➢ Age: Mean age is 28.79, with a younger customer base.
    - ➢ Gender: 57.78% male and 42.22% female, indicating a male-dominated customer base.
    - ➢ Income and Miles: Wide variability suggests diverse customer segments.
- ❖ **Product vs Customer Characteristics:**
    - ➢ KP781 is preferred by customers aged 25-30, with higher education levels, higher income, higher fitness levels, and expecting to use the treadmill extensively.
- ❖ **Distribution of Products:**
    - ➢ KP281: 44.44%
    - ➢ KP481: 33.33%
    - ➢ KP781: 22.22%

Recommandations:
- ❖ **Addressing Outliers**
    - ➢ Income and Miles: Perform detailed analysis and consider robust statistical methods (like trimming or winsorizing) to handle outliers.
- ❖ **Targeted Marketing and Product Development**
    - ➢ Age: Develop products and marketing strategies that appeal to slightly older demographics within the current age range.
    - ➢ Education and Usage: Make products more accessible and highlight features that cater to both casual and frequent users.
    - ➢ Income: Offer financing options to make higher-end products more accessible.
- ❖ **Customer Segmentation and Marketing Strategies**
    - ➢ **KP281 & KP481:**
        - ■ Market to a broad age range and emphasize affordability and essential features.
        - ■ Highlight suitability for moderate usage and appeal to budget-conscious individuals.

- ➢ **KP781:**

    - ■ Develop targeted campaigns for customers aged 25-30.
    - ■ Focus on high-income customers and fitness enthusiasts.
    - ■ Promote the treadmill's durability and advanced features for frequent, high-mileage users.
    - ■ Use testimonials and success stories from highly fit individuals.
- ➢ **Inventory Management**
    - ■ Ensure sufficient stock for KP281 due to its popularity.
    - ■ Maintain steady stock for KP481 and KP781, avoiding overstocking.
- ➢ **Product Positioning**
    - ■ **KP281:**
        - ● Emphasize affordability and value for money.
        - ● Use testimonials to attract new buyers.
    - ■ **KP481:**
        - ● Position as the best value for mid-range needs.
        - ● Highlight suitability for regular runners.
    - ■ **KP781:**
        - ● Market as a premium product with advanced features.
        - ● Target high-income customers and fitness enthusiasts.

Conclusion

Implementing these recommendations should help you effectively target different customer segments and maximize the appeal of each product. By focusing on customer insights and refining your marketing strategies accordingly, you can optimize sales and customer satisfaction across the entire product range.