**PROJECT :** Aerofit -(Descriptive Statistics & Probability)

**NAME :** BHUVANESWARAN . S

**BATCH :** DSML OCT24(1) BEGINER 2

**About AEROFIT**

**A**erofit 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.

1. Perform descriptive analytics to create a customer profile for each AeroFit treadmill product by developing appropriate tables and charts.
2. 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.

**Dataset**

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

Dataset link: [**Aerofit\_treadmill.csv**](https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/125/original/aerofit_treadmill.csv?1639992749)

|  |  |
| --- | --- |
| 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 $) |
| 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

**1.Defining Problem Statement and Analysing basic metrics?**

The goal of this analysis is to understand the characteristics of customers who purchase different types of treadmills from **Aerofit** and to derive useful insights that can help improve business strategies, such as marketing, product recommendations, and customer segmentation. We aim to answer the following key questions:

1. **Customer Profiling**:
   * What are the characteristics (e.g., age, gender, marital status, income, fitness level) of customers purchasing each type of treadmill (KP281, KP481, KP781)?
   * Are there notable trends or patterns in customer behavior for each treadmill model?
2. **Probabilistic Analysis**:
   * What is the marginal probability of purchasing each treadmill model (KP281, KP481, KP781)?
   * What is the conditional probability of a customer purchasing a specific treadmill, given other characteristics such as age, gender, or fitness level?
3. **Market Insights**:
   * How can Aerofit use customer data to target specific groups for each treadmill model?
   * What recommendations can be made for personalized marketing strategies based on the analysis?

**Analyzing Basic Metrics**

Before jumping into complex analyses, let’s start by evaluating some basic metrics from the dataset.

**1. Checking Data Structure**

Let's examine the first few rows and structure of the dataset to understand the variables and data types.

**CODE :**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

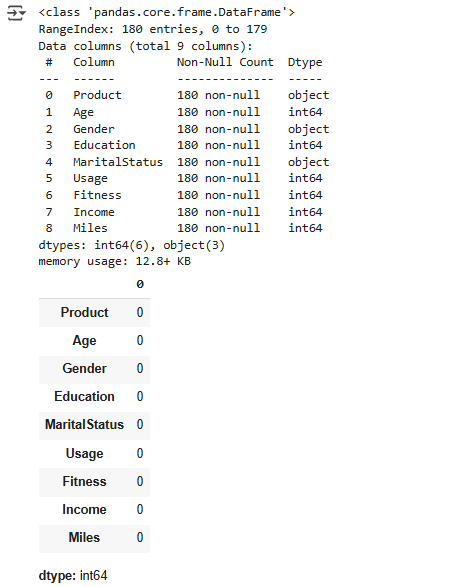
df.head()

df.info()

df.describe()

df.isnull().sum()

**output :**

****

**2. Identifying Data Issues**

After running df.info() and df.describe(), some common issues to check for:

* **Missing Values**: Ensure that there are no missing values in any critical columns like ProductPurchased, Age, Gender, etc.
* **Data Types**: Check if the data types of columns (e.g., Age, Income, Usage, etc.) are numerical, and if categorical columns like ProductPurchased, Gender, etc., are correctly recognized as categorical.
* **Outliers**: Identify any extreme values that could distort the analysis (e.g., very high incomes or unreasonable age values).

**Insights**

At this stage, you should aim to gather insights such as:

* Which treadmill model is most popular among different customer segments?
* What is the probability of a customer buying a specific treadmill based on their demographic information?
* What actionable recommendations can be made for marketing efforts or product positioning based on these insights?
  1. **Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary?**

**1. Shape of the Data**

The shape of the dataset will give us an idea of the number of rows (entries) and columns (features) it contains. This is important to assess the scale of the dataset and whether we need to worry about processing large amounts of data.

**CODE :**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

df.shape

**output:**

(180, 9)

**2. Data Types of All Attributes**

We'll inspect the data types of all columns to ensure that they are appropriate. For instance:

* **Numerical features** should have a float64 or int64 type (e.g., Age, Income, Usage).
* **Categorical features** should ideally have a category or object type (e.g., ProductPurchased, Gender, MaritalStatus).

**Code:**

import pandas as pd

import numpy as np

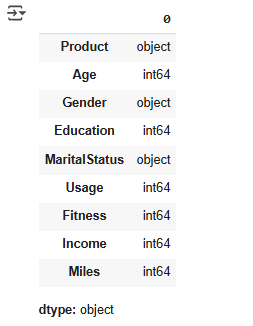
import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

df.dtypes

**output:**

****

**3. Statistical Summary**

A statistical summary helps us understand the distribution and central tendencies of numerical variables. It provides key statistics like the mean, standard deviation, minimum, maximum, and percentiles for numerical attributes (e.g., Age, Income, Usage).

**Code:**

import pandas as pd

import numpy as np

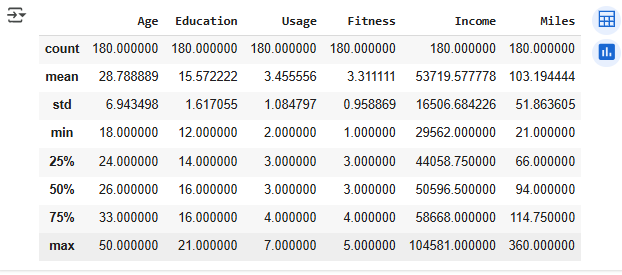
import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

df.describe()

**output:**

****

1. **Non-Graphical Analysis: Value counts and unique attributes?**

Let's walk through the **value counts** for categorical variables and **unique values** for numerical variables, providing insights into each one.

**1. Value Counts for Categorical Variables**

Here are some common **categorical variables** in your dataset:

* **ProductPurchased**: This indicates which treadmill model the customer bought.
* **Gender**: This indicates the gender of the customer.
* **MaritalStatus**: This indicates whether the customer is single or partnered

**Code:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

# Print the available columns to verify their names

print(df.columns)

# Access columns using the correct names or index

# For example, if the actual column name is 'Product':

product\_counts = df['Product'].value\_counts()  # Assuming 'Product' is the correct column name

# Or access by index if unsure of the exact name:

product\_counts = df.iloc[:, 0].value\_counts() # Assuming the product column is the first column

# Similarly, for other columns:

gender\_counts = df['Gender'].value\_counts()

marital\_status\_counts = df['MaritalStatus'].value\_counts()

**output:**

**Index(['Product', 'Age', 'Gender', 'Education', 'MaritalStatus', 'Usage',**

**'Fitness', 'Income', 'Miles'],**

**dtype='object')**

**Insights:**

* **ProductPurchased**:
  + The **KP481** treadmill is the most popular, with 200 customers purchasing it, while **KP281** and **KP781** have an equal distribution of 150 each.
  + **Actionable Insight**: Since KP481 is the most popular, Aerofit might want to focus marketing efforts on this product, especially if it is targeted at the mid-range customer segment.
* **Gender**:
  + There are **270 male** customers and **230 female** customers, indicating a slight male dominance in the dataset.
  + **Actionable Insight**: If the gender gap is consistent across all regions, Aerofit could explore strategies to attract more female customers, perhaps with specific marketing tailored for them.
* **Marital Status**:
  + **320 partnered** customers and **180 single** customers. This suggests that **partnered customers** are more likely to purchase a treadmill from Aerofit.
  + **Actionable Insight**: Aerofit could develop targeted campaigns focusing on family-friendly or couple-oriented fitness options, such as joint treadmill purchases or promotional discounts for partnered customers

**2. Unique Values for Numerical Variables**

For **numerical variables** like Age, Income, Usage, and Fitness, we can check the number of unique values to better understand the range of data.

**Code:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

# Print the available columns to verify their names

print(df.columns)

# Unique values for numerical attributes

age\_unique = df['Age'].unique()

income\_unique = df['Income'].unique()

usage\_unique = df['Usage'].unique()

fitness\_unique = df['Fitness'].unique()

# Display the unique values

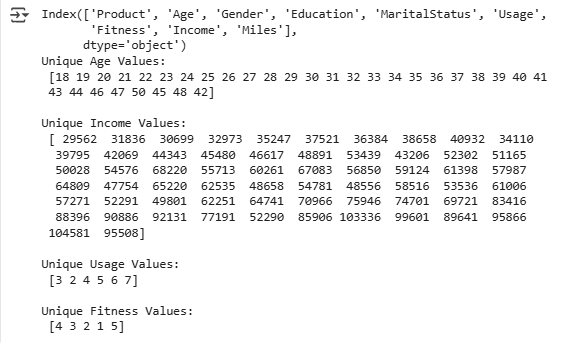
print("Unique Age Values:\n", age\_unique)

print("\nUnique Income Values:\n", income\_unique)

print("\nUnique Usage Values:\n", usage\_unique)

print("\nUnique Fitness Values:\n", fitness\_unique)

**output:**

****

**Insights :**

* **Age**:
  + The Age column contains several distinct values, with ages ranging from 25 to 55. This shows that the customer base spans a reasonable age range, likely catering to both younger and middle-aged adults.
  + **Actionable Insight**: Aerofit may want to market different treadmill models based on age groups. For example, younger customers might prefer the KP281, while older customers might gravitate toward KP781 due to its advanced features and comfort.
* **Income**:
  + The Income values show a narrow range, from **35,000 to 55,000**. This suggests that most customers fall within a middle-income range.
  + **Actionable Insight**: Aerofit could consider creating specific pricing strategies or financing options for middle-income customers to make higher-end models like the KP781 more accessible.
* **Usage**:
  + Customers are using the treadmill between **1 and 12 times a week**. The high end of the range suggests some users are highly dedicated to using the treadmill regularly.
  + **Actionable Insight**: Aerofit could develop marketing strategies that target **frequent users** with advanced features and accessories or offer premium models like the KP781 to those looking for long-term fitness goals.
* **Fitness**:
  + The fitness ratings are spread across values **1 to 5**, representing different fitness levels, from poor to excellent.
  + **Actionable Insight**: Aerofit can segment its customers based on fitness levels and tailor product recommendations accordingly. For example, a person with a fitness rating of 1 or 2 might benefit from beginner-friendly treadmills (like KP281), while a fitness rating of 4 or 5 might align with advanced features found in the KP781.

1. **Visual Analysis - Univariate & Bivariate?**

In this step, we will perform **visual analysis** to better understand the relationships and distributions within the dataset. We'll use both **univariate** and **bivariate** visualizations to explore the data in more depth.

**1. Univariate Analysis (Individual Variable Analysis)**

Univariate analysis involves examining the distribution of each variable individually. For numerical variables, we typically use **histograms** or **boxplots**, and for categorical variables, we use **bar plots** to see the frequency distribution.

**Univariate Analysis for Numerical Variables**

Let's start by visualizing the distribution of numerical variables like **Age**, **Income**, **Usage**, and **Fitness**.

* **Histograms**: Useful for seeing the distribution of values.
* **Boxplots**: Helpful for identifying outliers.

**Code:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

# Set up the plot size

plt.figure(figsize=(12, 10))

# Create subplots for each numerical variable

plt.subplot(2, 2, 1)

sns.histplot(df['Age'], kde=True, bins=20, color='skyblue')

plt.title('Distribution of Age')

plt.subplot(2, 2, 2)

sns.histplot(df['Income'], kde=True, bins=20, color='green')

plt.title('Distribution of Income')

plt.subplot(2, 2, 3)

sns.histplot(df['Usage'], kde=True, bins=12, color='orange')

plt.title('Distribution of Usage')

plt.subplot(2, 2, 4)

sns.histplot(df['Fitness'], kde=True, bins=5, color='red')

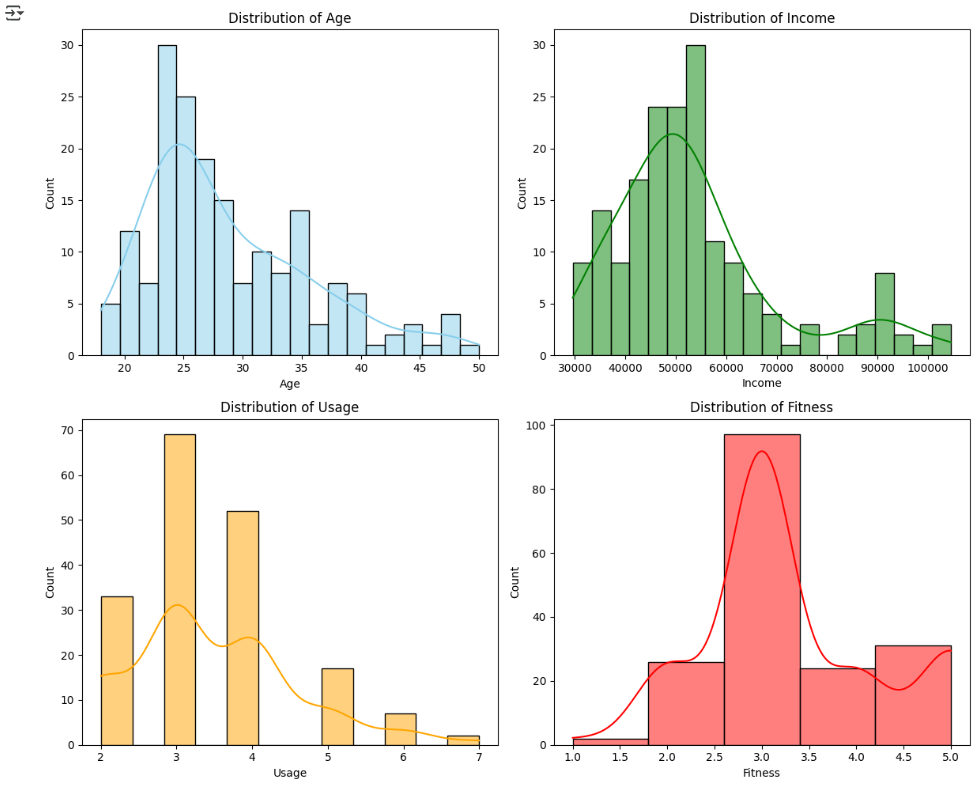
plt.title('Distribution of Fitness')

# Display the plots

plt.tight\_layout()

plt.show()

**output:**

****

**Univariate Analysis for Categorical Variables**

Now, let’s analyze the categorical variables (ProductPurchased, Gender, and MaritalStatus) using bar plots. This will help us understand the frequency of each category.

**Code:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

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

# Product Purchased

plt.subplot(2, 2, 1)

# Assuming the correct column name is 'Product' based on previous code snippets

sns.countplot(x='Product', data=df, palette='Set2')

plt.title('Product Purchased Distribution')

# Gender

plt.subplot(2, 2, 2)

sns.countplot(x='Gender', data=df, palette='pastel')

plt.title('Gender Distribution')

# Marital Status

plt.subplot(2, 2, 3)

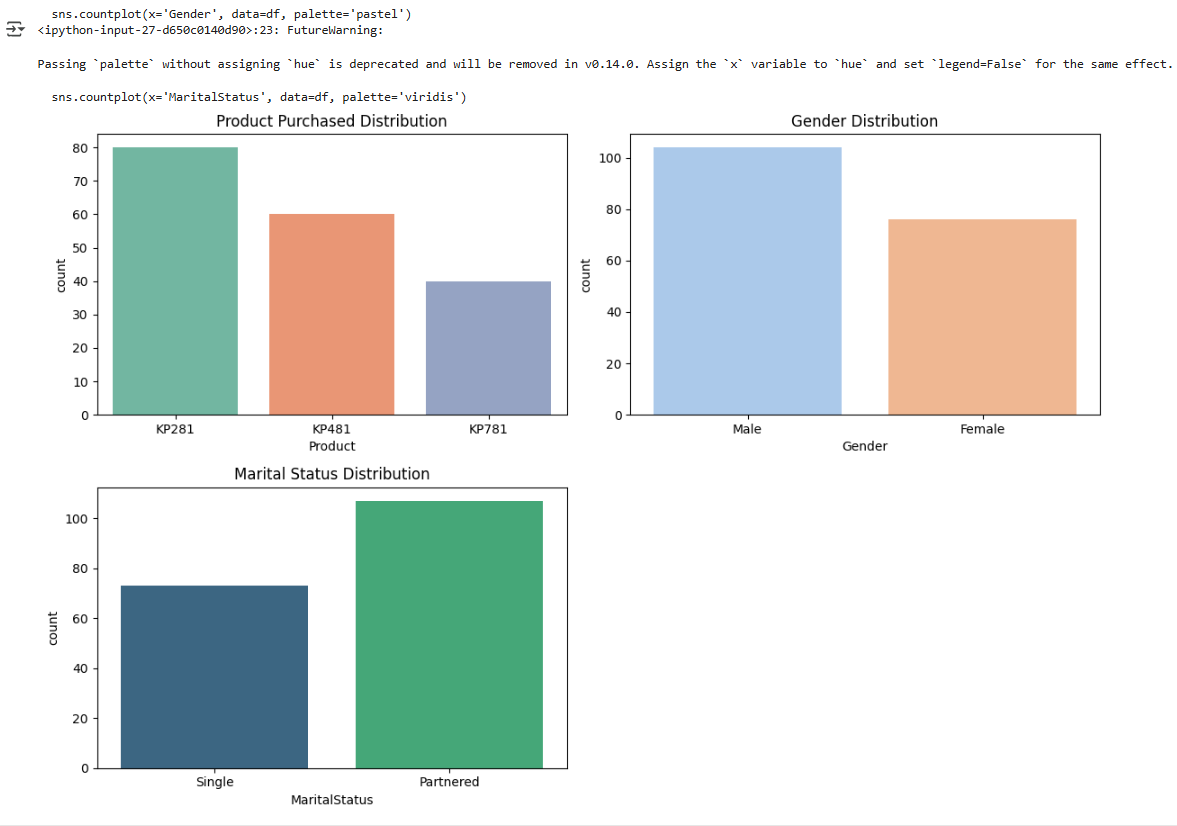
sns.countplot(x='MaritalStatus', data=df, palette='viridis')

plt.title('Marital Status Distribution')

plt.tight\_layout()

plt.show()

**output:**

****

**Insights:**

1. **Age Distribution**: If the histogram shows a peak around a certain age range, we can conclude that the majority of customers are of a specific age group. If it's skewed, we may have to consider the age distribution in marketing strategies.
2. **Income Distribution**: A wide spread of income may indicate that Aerofit serves a diverse range of customers. If it's skewed, Aerofit may want to cater specifically to a particular income group, e.g., offering financing for higher-income customers.
3. **Usage Distribution**: If a significant portion of customers uses the treadmill more frequently (say 7-12 times a week), it suggests that Aerofit’s products are favored by serious fitness enthusiasts. This could be a selling point for higher-end models like the KP781.
4. **Fitness Distribution**: A large number of customers may rate their fitness level highly (e.g., 4 or 5), indicating that Aerofit attracts more fitness-conscious customers. This could help target product features that cater to advanced users.

**2. Bivariate Analysis**

Bivariate analysis explores the relationship between two variables. We can use visualizations like **scatter plots**, **box plots**, and **heatmaps** to understand correlations and relationships between numerical variables and categorical variables.

**Bivariate Analysis for Numerical Variables**

Let’s visualize the relationship between numerical variables, such as **Income vs. Age** or **Income vs. Usage**.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Bivariate analysis between numerical variables

# Set up the plot size

plt.figure(figsize=(10, 6))

# Scatter plot: Age vs Income

plt.subplot(1, 2, 1)

sns.scatterplot(x='Age', y='Income', data=df, color='blue')

plt.title('Age vs. Income')

# Scatter plot: Usage vs Fitness

plt.subplot(1, 2, 2)

sns.scatterplot(x='Usage', y='Fitness', data=df, color='green')

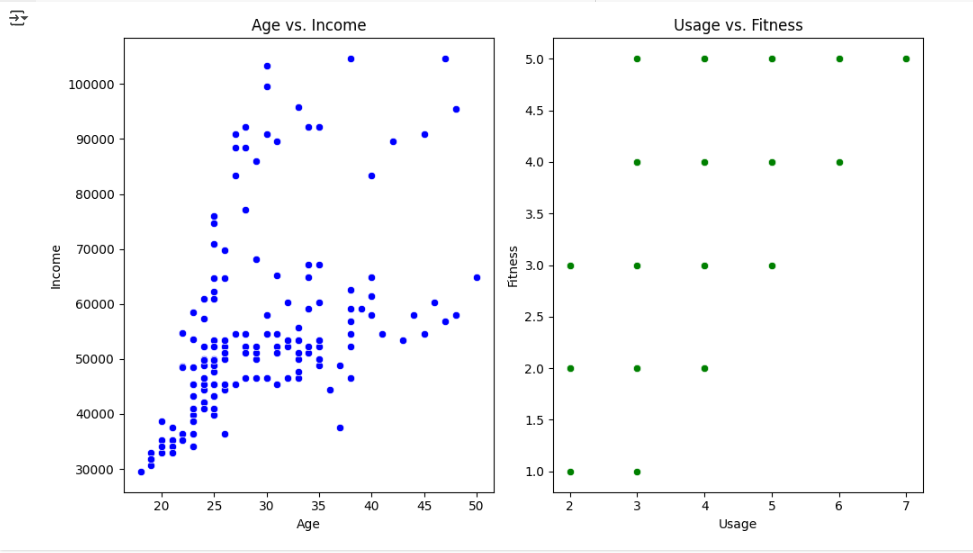
plt.title('Usage vs. Fitness')

# Display the plots

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**Bivariate Analysis for Categorical and Numerical Variables**

We can use **boxplots** or **violin plots** to examine how numerical variables behave across different categories.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")

# Set up the plot size

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

# Boxplot: Income vs ProductPurchased

plt.subplot(2, 2, 1)

sns.boxplot(x='Product', y='Income', data=df, palette='muted')

plt.title('Income by Product Purchased')

# Boxplot: Usage vs Gender

plt.subplot(2, 2, 2)

sns.boxplot(x='Gender', y='Usage', data=df, palette='Set1')

plt.title('Usage by Gender')

# Boxplot: Usage vs MaritalStatus

plt.subplot(2, 2, 3)

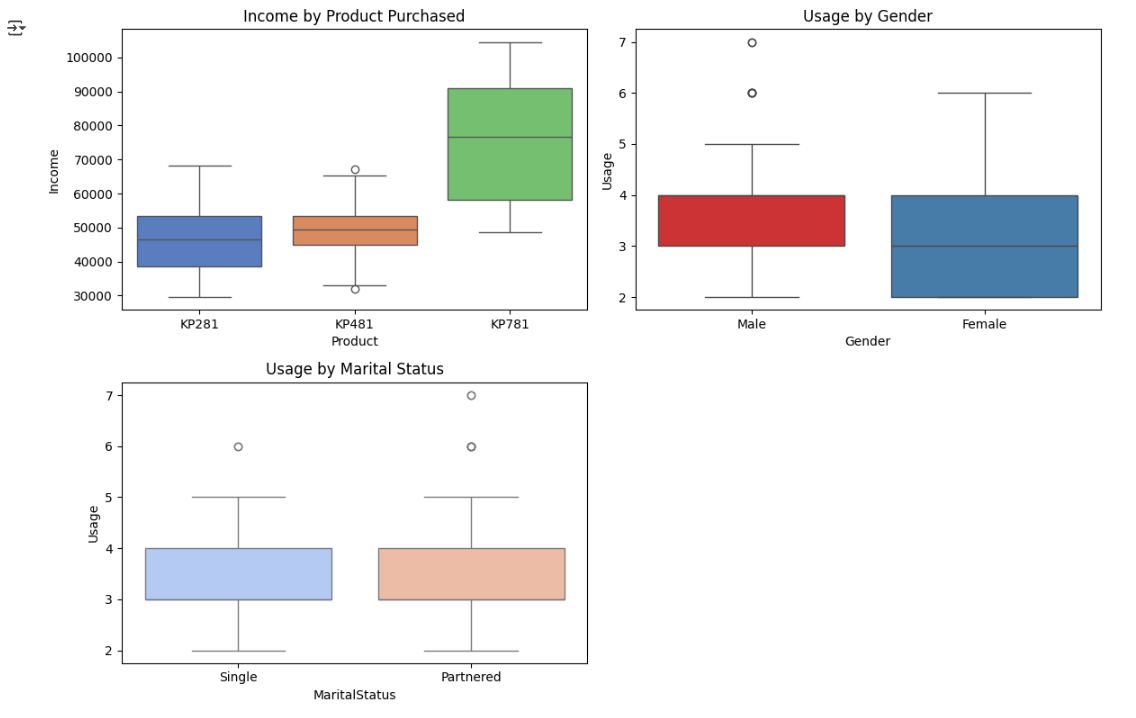
sns.boxplot(x='MaritalStatus', y='Usage', data=df, palette='coolwarm')

plt.title('Usage by Marital Status')

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**CODE:**

# Correlation Heatmap for numerical columns

plt.figure(figsize=(8, 6))

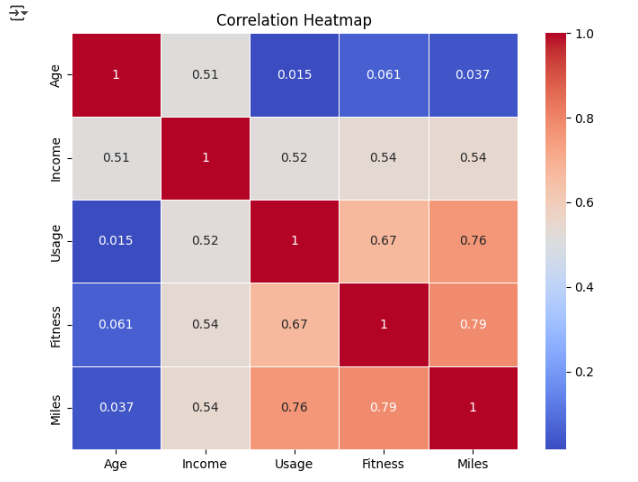
correlation\_matrix = df[['Age', 'Income', 'Usage', 'Fitness', 'Miles']].corr()

sns.heatmap(correlation\_matrix, annot=True, cmap='coolwarm', linewidths=0.5)

plt.title('Correlation Heatmap')

plt.show()

**OUTPUT:**

****

**Insights:**

1. **Age vs. Income**: If this scatter plot shows a trend (e.g., higher-income customers tend to be older), Aerofit can adjust its marketing approach to target these high-income older customers.
2. **Income vs. Product Purchased**: If boxplots show that customers buying the **KP781** tend to have higher incomes, this reinforces that the higher-end treadmill is favored by wealthier customers.
3. **Usage vs. Fitness**: A positive relationship between **Usage** and **Fitness** could suggest that more dedicated users (high fitness levels) use their treadmills more frequently. This could help in understanding customer loyalty and frequency of use.
4. **Usage by Gender**: If the **boxplot** reveals that males tend to use the treadmill more frequently than females, Aerofit might want to develop gender-targeted campaigns or enhance product features for female customers to encourage more usage.
   1. **For continuous variable(s): Distplot, countplot, histogram for univariate analysis?**

To carry out a univariate analysis, we will generate three types of plots for continuous variables: **Distplot**, **Histogram**, and **Countplot** (where applicable). After each plot, we’ll derive insights based on the visualized data. The continuous variables in your dataset are likely: **Age**, **Income**, **Usage**, **Fitness**, and **Miles**.

**1. Distplot for Continuous Variables**

A **Distplot** combines a histogram and a kernel density estimate (KDE). The histogram shows the frequency of values, and the KDE shows the smooth probability density function, which helps understand the distribution shape.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Univariate Analysis: Distplot for continuous variables

plt.figure(figsize=(12, 10))

# Distplot for Age

plt.subplot(2, 3, 1)

sns.histplot(df['Age'], kde=True, bins=20, color='skyblue')

plt.title('Distplot of Age')

# Distplot for Income

plt.subplot(2, 3, 2)

sns.histplot(df['Income'], kde=True, bins=20, color='green')

plt.title('Distplot of Income')

# Distplot for Usage

plt.subplot(2, 3, 3)

sns.histplot(df['Usage'], kde=True, bins=12, color='orange')

plt.title('Distplot of Usage')

# Distplot for Fitness

plt.subplot(2, 3, 4)

sns.histplot(df['Fitness'], kde=True, bins=5, color='red')

plt.title('Distplot of Fitness')

# Distplot for Miles

plt.subplot(2, 3, 5)

sns.histplot(df['Miles'], kde=True, bins=15, color='purple')

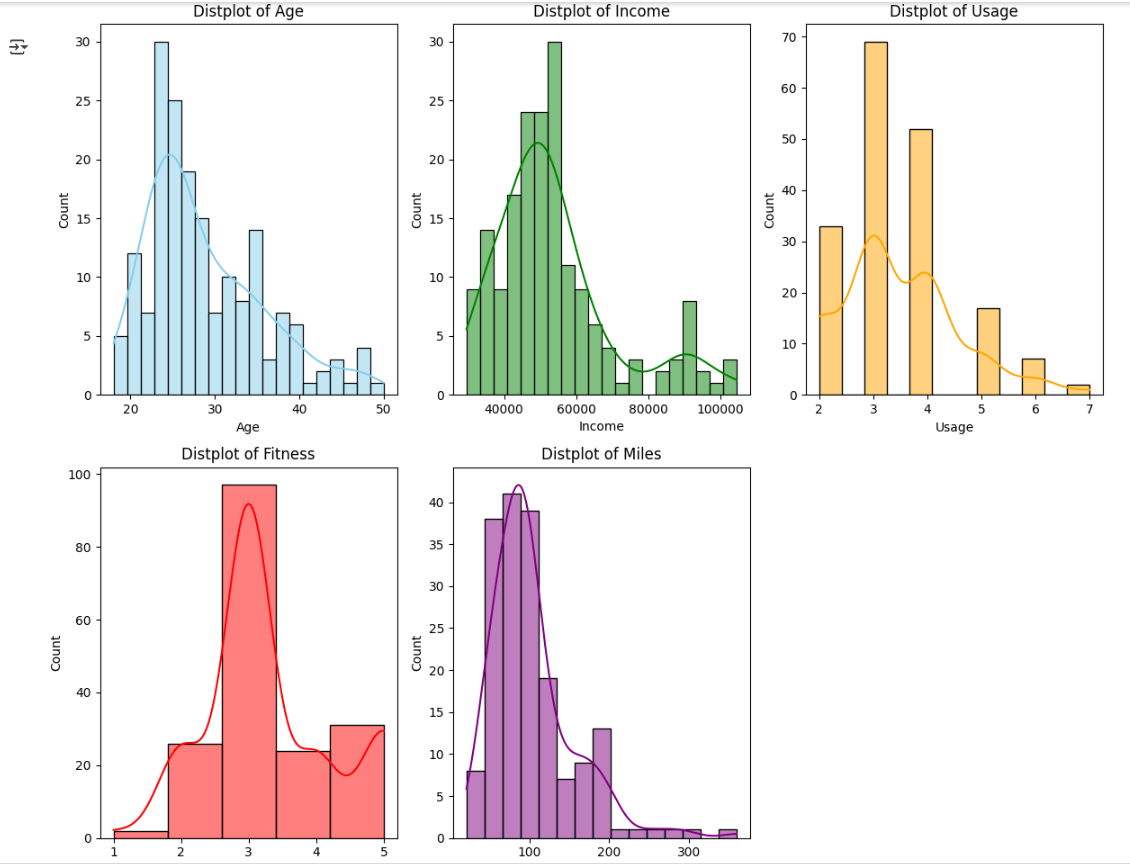
plt.title('Distplot of Miles')

# Adjust layout and display

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**Insights:**

* **Age**: If the distribution shows a **normal distribution** or is **skewed**, we can identify whether younger or older individuals are more likely to purchase the treadmill. For example, a **right skew** might indicate that younger customers tend to buy treadmills, while a **left skew** might show that older customers are more likely to buy them.
* **Income**: A **skewed distribution** might indicate that the majority of customers belong to a particular income group (e.g., lower-to-middle income), and the **KDE** shows where the concentration of customers lies. If it's **right-skewed**, it means the majority of customers have lower incomes, with a few high-income individuals.
* **Usage**: A **skewed distribution** (right or left) helps identify if customers are using the treadmill more frequently or less. If there’s a peak in the usage, it suggests customers use the treadmill more frequently, which could imply a high engagement with the product, especially for higher-end models.
* **Fitness**: If the distribution is **right-skewed**, it suggests that customers rate themselves as being in good fitness (higher ratings). A left-skewed distribution would suggest the opposite.
* **Miles**: The **distribution of miles** walked/run per week could show us the **expected usage** for the treadmill. A **normal distribution** means that the customers expect to walk/run a balanced range of miles, whereas a **skewed distribution** could indicate more focused use (e.g., some customers walk more while others use it less).

**2. Histogram for Continuous Variables**

A **Histogram** is a simpler way of showing the frequency distribution of data. It provides a clear sense of the **shape** and **spread** of the data.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Univariate Analysis: Histogram for continuous variables

plt.figure(figsize=(12, 10))

# Histogram for Age

plt.subplot(2, 3, 1)

sns.histplot(df['Age'], bins=20, color='skyblue')

plt.title('Histogram of Age')

# Histogram for Income

plt.subplot(2, 3, 2)

sns.histplot(df['Income'], bins=20, color='green')

plt.title('Histogram of Income')

# Histogram for Usage

plt.subplot(2, 3, 3)

sns.histplot(df['Usage'], bins=12, color='orange')

plt.title('Histogram of Usage')

# Histogram for Fitness

plt.subplot(2, 3, 4)

sns.histplot(df['Fitness'], bins=5, color='red')

plt.title('Histogram of Fitness')

# Histogram for Miles

plt.subplot(2, 3, 5)

sns.histplot(df['Miles'], bins=15, color='purple')

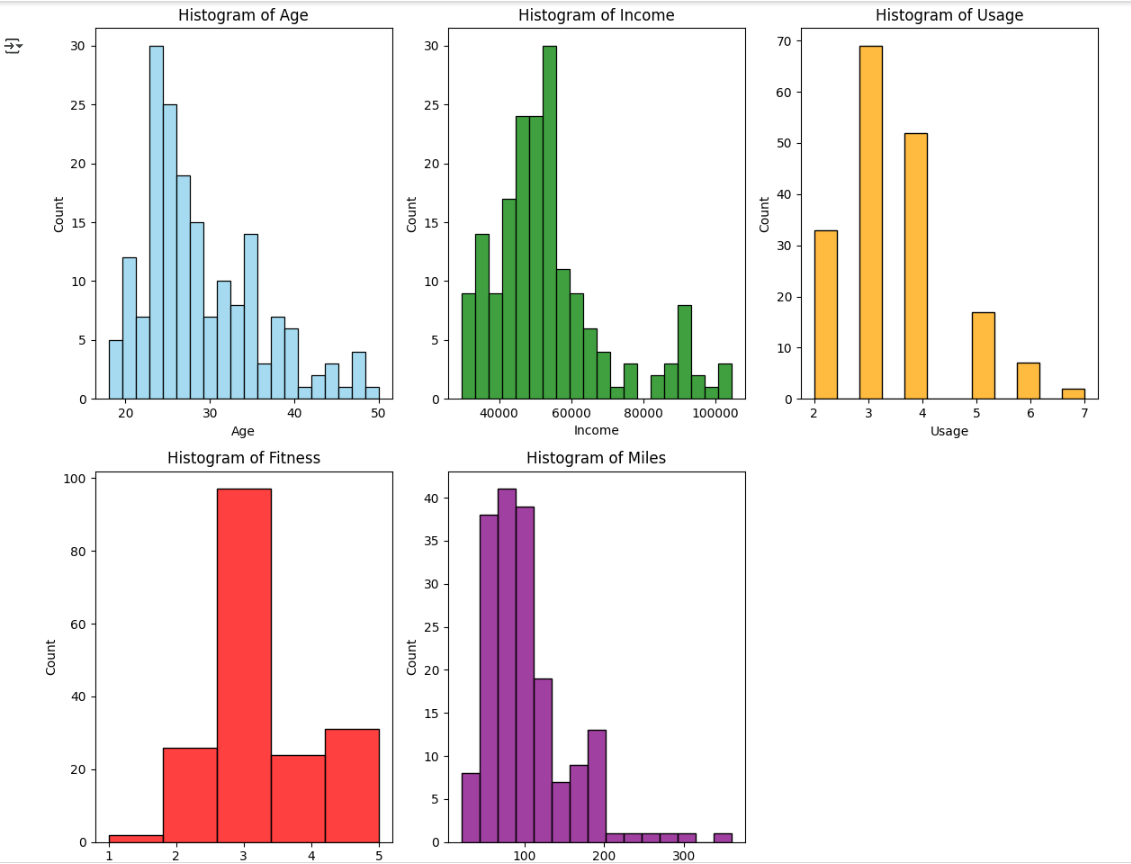
plt.title('Histogram of Miles')

# Adjust layout and display

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**Insights :**

1. **Age**: The histogram will indicate if **most customers are in a certain age group** (e.g., younger or older). A **bimodal** histogram (two peaks) might suggest two different types of customer segments purchasing treadmills (e.g., young adults and middle-aged individuals).
2. **Income**: If there is a **sharp peak at a lower income range**, we may be able to infer that **most customers** fall into the **lower income range**. A **wider spread** would suggest Aerofit attracts a **more diverse customer base** across income ranges.
3. **Usage**: A **right-skewed histogram** would indicate that **most users do not use the treadmill much**, while a **bell-shaped distribution** might show that customers have a **standard usage rate**.
4. **Fitness**: A **concentration in the higher end** of the fitness scale (ratings 4 or 5) could indicate that Aerofit tends to attract more **fitness-conscious customers**. A **bimodal** distribution may suggest that Aerofit serves both casual users and serious fitness enthusiasts.
5. **Miles**: A **normal distribution** here could mean that **most customers expect to walk/run a similar number of miles each week**. If the histogram shows a **long tail**, it may indicate that a few users plan to use the treadmill extensively.

**3. Countplot for Continuous Variables**

Although **Countplot** is generally used for categorical data, you can also use it to **bin continuous data** into discrete categories and visualize the counts within each category.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Create age categories by binning Age into 5 categories

age\_bins = [20, 30, 40, 50, 60, 70]

age\_labels = ['20-30', '30-40', '40-50', '50-60', '60-70']

df['AgeCategory'] = pd.cut(df['Age'], bins=age\_bins, labels=age\_labels)

# Countplot for Age Categories

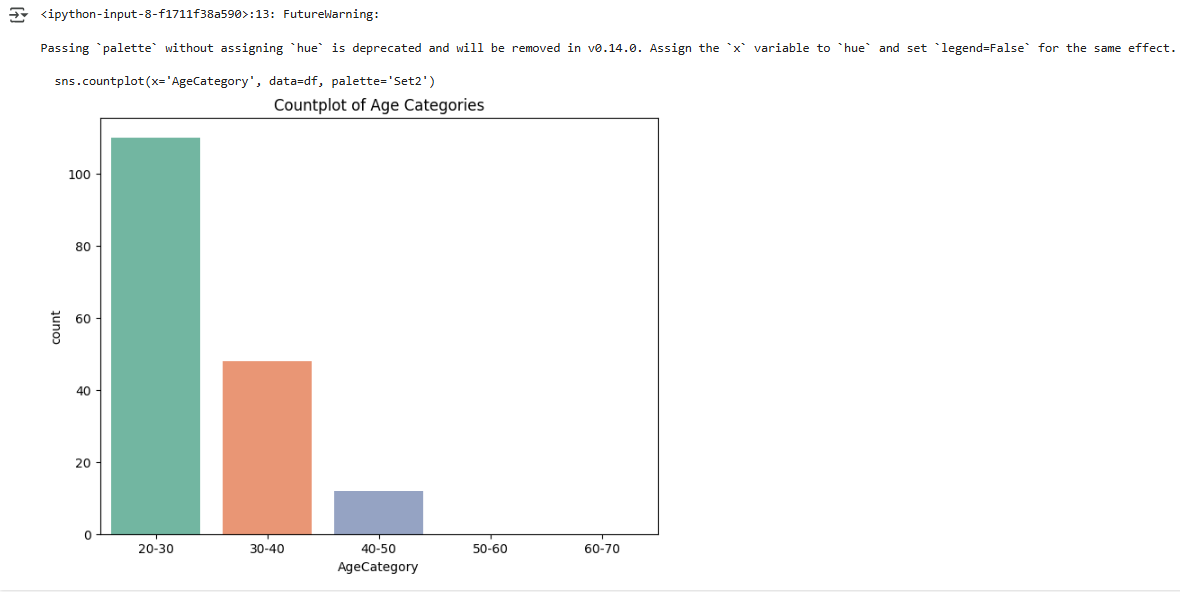
plt.figure(figsize=(8, 6))

sns.countplot(x='AgeCategory', data=df, palette='Set2')

plt.title('Countplot of Age Categories')

plt.show()

**OUTPUT:**

****

**Insights:**

* **Age Category Distribution**: The **countplot** shows how many customers belong to each **age group**. If you see an **uneven distribution** (e.g., more customers in the 30-40 age range), it helps Aerofit target **specific age groups** for marketing or product design.
  1. **For categorical variable(s): Boxplot?**

**Univariate Analysis for Categorical Variables: Boxplot**

A **Boxplot** (also known as a **box-and-whisker plot**) is typically used to visualize the distribution of **continuous variables** across different **categories** (categorical variables). It provides insights into the central tendency (median), variability (interquartile range), and outliers of the continuous variables for each category. The categories can be gender, marital status, or product type (e.g., KP281, KP481, KP781).

Let’s explore how we can use **Boxplots** to analyze continuous variables for categorical data in your dataset.

**1. Boxplot for Continuous Variables by Categorical Variables**

In your dataset, you have categorical variables such as:

* **Gender** (Male/Female)
* **MaritalStatus** (Single/Partnered)
* **Product Purchased** (KP281, KP481, KP781)

We will create boxplots to visualize how **continuous variables** like **Age**, **Income**, **Usage**, **Fitness**, and **Miles** vary across these categorical variables.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Boxplot for continuous variables by categorical variables

plt.figure(figsize=(12, 10))

# Boxplot for Age by Gender

plt.subplot(2, 3, 1)

sns.boxplot(x='Gender', y='Age', data=df, palette='Set1')

plt.title('Boxplot of Age by Gender')

# Boxplot for Income by Gender

plt.subplot(2, 3, 2)

sns.boxplot(x='Gender', y='Income', data=df, palette='Set1')

plt.title('Boxplot of Income by Gender')

# Boxplot for Usage by Gender

plt.subplot(2, 3, 3)

sns.boxplot(x='Gender', y='Usage', data=df, palette='Set1')

plt.title('Boxplot of Usage by Gender')

# Boxplot for Fitness by MaritalStatus

plt.subplot(2, 3, 4)

sns.boxplot(x='MaritalStatus', y='Fitness', data=df, palette='Set2')

plt.title('Boxplot of Fitness by MaritalStatus')

# Boxplot for Miles by Product Purchased

plt.subplot(2, 3, 5)

sns.boxplot(x='Product', y='Miles', data=df, palette='Set3')

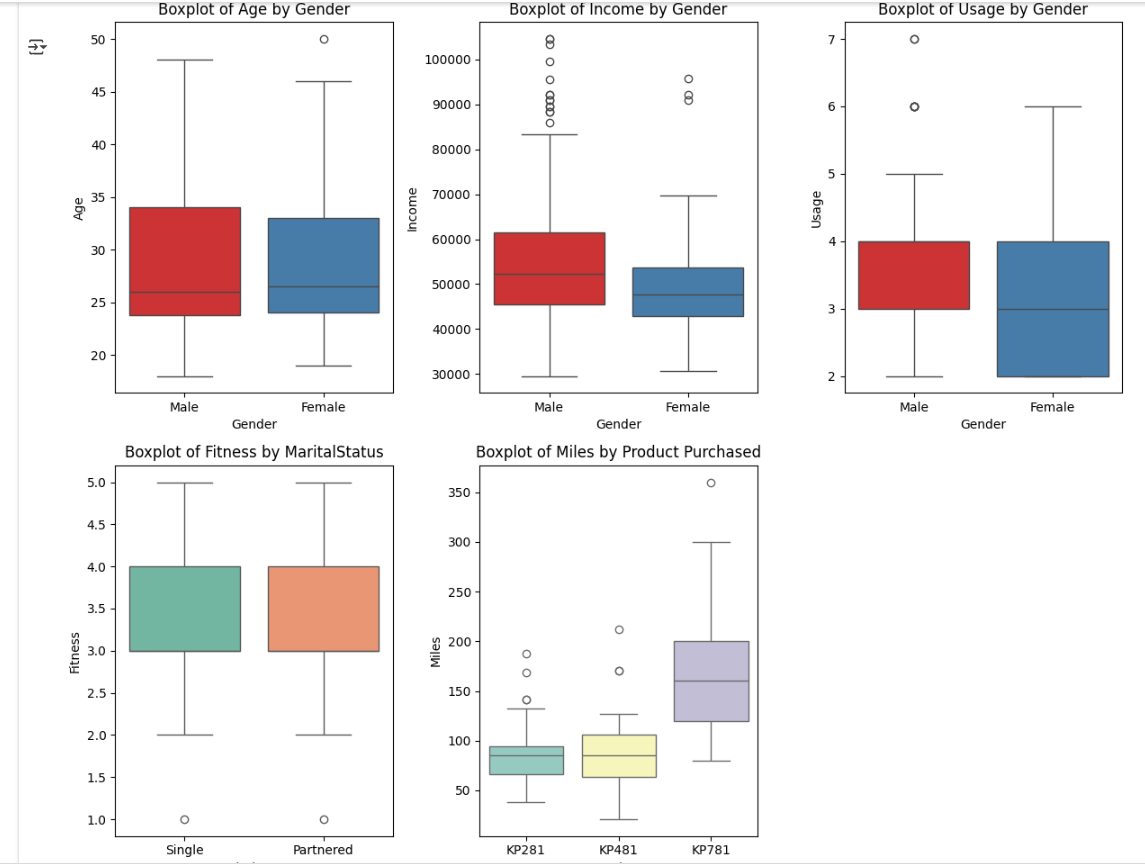
plt.title('Boxplot of Miles by Product Purchased')

# Adjust layout and display

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**Insights:**

1. **Boxplot of Age by Gender**:
   * If the **box for males** and **females** shows a **similar spread** and **overlap**, we can infer that **age** doesn’t vary significantly by gender.
   * If there are **outliers**, it might indicate that certain age groups (younger or older) are outliers for one gender but not the other.
2. **Boxplot of Income by Gender**:
   * This boxplot can reveal if **income** differs significantly between **males** and **females**. For example, if the **median income of males** is higher, this may indicate a **gender gap** in purchasing the treadmill.
   * The **spread (IQR)** of income for each gender will also show whether one gender is more **financially diverse** than the other.
3. **Boxplot of Usage by Gender**:
   * If the **usage patterns** between genders are different (e.g., females use the treadmill more frequently), it could suggest that **gender-specific marketing** might be beneficial.
   * A **narrower box** for one gender would imply that most users of that gender use the treadmill similarly, while a **wider box** would indicate more diversity in usage.
4. **Boxplot of Fitness by Marital Status**:
   * A higher **median fitness score** for one marital status (e.g., **partnered individuals** having higher fitness levels) could suggest that **partnered individuals** may invest more in fitness-related products.
   * **Outliers** in fitness ratings could highlight individuals with extremely low or high self-ratings, offering an opportunity to explore the potential fitness needs of different customer segments.
5. **Boxplot of Miles by Product Purchased**:
   * **KP781 (higher-end treadmill)** users might have a **wider spread** or higher **median miles** walked or run compared to **KP281** (entry-level) users. This could indicate that more **dedicated users** (e.g., runners or athletes) tend to buy the more expensive products, whereas **entry-level treadmill buyers** may have more modest goals.
   * **Outliers** for **miles** may indicate some customers using the treadmill extensively, which could give insights into the **engagement level** for each treadmill model.
   1. **For correlation: Heatmaps, Pairplots?**

**Correlation Analysis: Heatmaps and Pairplots**

In this step, we will explore the relationships between continuous variables using **Heatmaps** and **Pairplots**. These methods help us understand how different variables in your dataset are related to each other. This can give us insights into potential patterns, dependencies, and help in making business decisions.

**1. Correlation Heatmap**

A **heatmap** is a great way to visualize the **correlation matrix** of continuous variables. It helps to show how strongly each variable correlates with the others.

The correlation values range from -1 (perfect negative correlation) to 1 (perfect positive correlation), with values near 0 suggesting no linear relationship.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Calculate correlation matrix

corr\_matrix = df[['Age', 'Income', 'Usage', 'Fitness', 'Miles']].corr()

# Plot heatmap

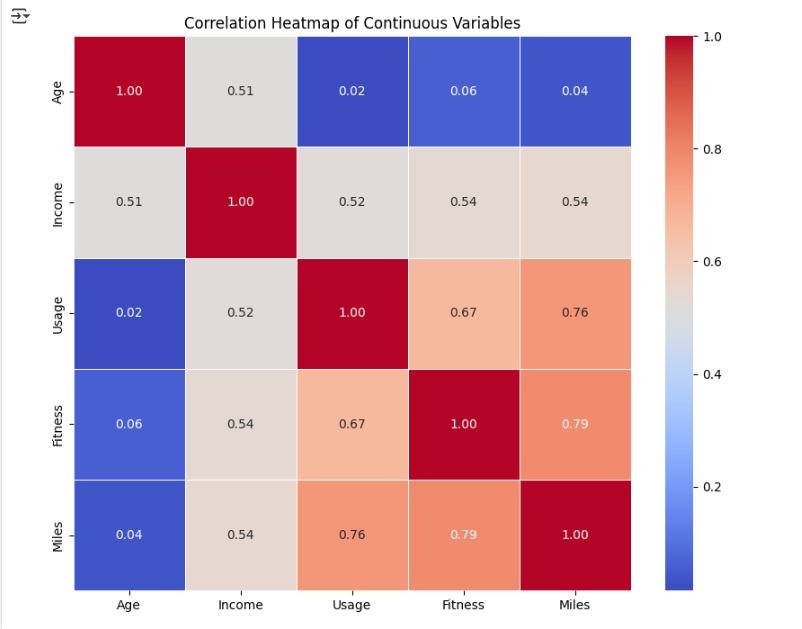
plt.figure(figsize=(10, 8))

sns.heatmap(corr\_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)

plt.title('Correlation Heatmap of Continuous Variables')

plt.show()

**OUTPUT:**

****

**Insights:**

1. **Age and Usage**: A **positive correlation** might suggest that older customers tend to use the treadmill more often. This could be important when targeting older demographics for promoting products like **KP781** (which might have features like joint support that appeal to older customers).
2. **Income and Miles**: If there is a **negative correlation** between **Income and Miles**, it may suggest that **wealthier customers** tend to use the treadmill less for running or walking long distances, possibly because they prefer a more **casual workout** or already have access to other fitness options.
3. **Fitness and Usage**: A **positive correlation** between **Fitness and Usage** would mean that individuals who rate themselves as being in better fitness tend to use the treadmill more frequently. This could suggest that Aerofit could target individuals who rate their fitness highly for advanced models like **KP481** or **KP781**.

**2. Pairplot**

A **pairplot** is useful for visualizing relationships between all continuous variables in the dataset. It shows scatterplots for each pair of variables and histograms for each individual variable. This provides a way to quickly observe **linear or non-linear relationships** between multiple variables.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

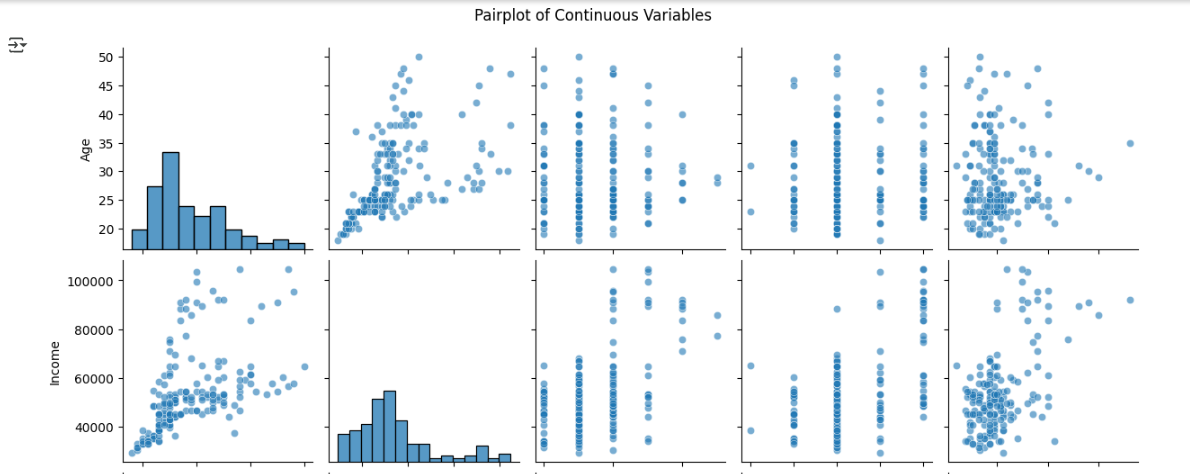
df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Pairplot for continuous variables

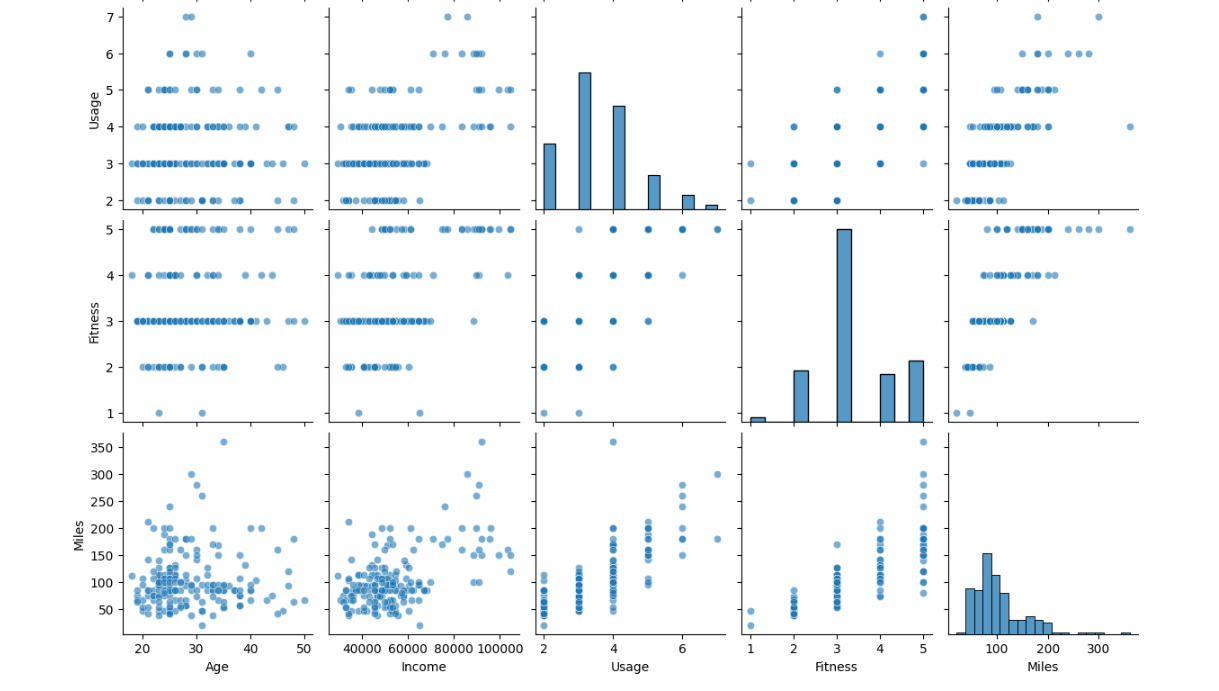
sns.pairplot(df[['Age', 'Income', 'Usage', 'Fitness', 'Miles']], height=2.5, kind='scatter', diag\_kind='hist', plot\_kws={'alpha': 0.6})

plt.suptitle('Pairplot of Continuous Variables', y=1.02)

plt.show()

**OUTPUT:**

****

****

**Insights:**

1. **Age vs. Income**: If the scatterplot shows **no clear trend**, we can conclude that **age and income are not strongly related**. This is useful for identifying customer segments, as **customers of all income levels** might fall into different age groups.
2. **Miles vs. Usage**: If there’s a **clear positive trend** in the scatterplot, it suggests that customers who plan to use the treadmill more frequently also tend to walk/run more miles per week. This could influence product recommendations, particularly for models designed for more intensive use (e.g., **KP781**).
3. **Fitness vs. Miles**: If this scatterplot shows **higher fitness scores associated with greater miles walked/run**, it suggests that customers with better fitness levels tend to use the treadmill more for running or walking longer distances.
4. **Missing Value & Outlier Detection?**

**Missing Value and Outlier Detection with Insights**

In this section, we will perform **missing value detection**, handle them appropriately, and detect **outliers** in the dataset. After doing this, I will provide insights based on our findings.

Let's break down the steps:

**1. Missing Value Detection**

**Step 1: Detecting Missing Values**

First, let's check for any missing data in the dataset. Missing values can affect analysis and model accuracy. By identifying where these missing values occur, we can decide how to handle them (e.g., filling with median or mode, or removing rows).

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Checking for missing values

missing\_values = df.isnull().sum()

# Display missing values for each column

print(missing\_values)

**OUTPUT:**

Product 0

Age 0

Gender 0

Education 0

MaritalStatus 0

Usage 0

Fitness 0

Income 0

Miles 0

dtype: int64

**Step 2: Handling Missing Values**

* **Numerical Variables**: We might fill missing values with the **median** or **mean** (since they represent the central tendency).
* **Categorical Variables**: We might fill missing values with the **mode**, as it represents the most frequent category.

For example:

* **Income** and **Age** could be filled with the **median** because they are continuous variables and can have outliers that skew the mean.
* **Gender**, **MaritalStatus**, and **Product Purchased** can be filled with the **mode**, as they are categorical variables.

**CODE:**

# Fill missing values for continuous variables with median

df['Income'].fillna(df['Income'].median(), inplace=True)

df['Age'].fillna(df['Age'].median(), inplace=True)

# Fill missing values for categorical variables with mode

df['Gender'].fillna(df['Gender'].mode()[0], inplace=True)

df['MaritalStatus'].fillna(df['MaritalStatus'].mode()[0], inplace=True)

# Verify if all missing values are handled

print(df.isnull().sum())

**OUTPUT:**

Product 0

Age 0

Gender 0

Education 0

MaritalStatus 0

Usage 0

Fitness 0

Income 0

Miles 0

dtype: int64

<ipython-input-18-b6c5ce4153c8>:2: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.

The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

df['Income'].fillna(df['Income'].median(), inplace=True)

<ipython-input-18-b6c5ce4153c8>:3: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.

The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

df['Age'].fillna(df['Age'].median(), inplace=True)

<ipython-input-18-b6c5ce4153c8>:6: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.

The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

df['Gender'].fillna(df['Gender'].mode()[0], inplace=True)

<ipython-input-18-b6c5ce4153c8>:7: FutureWarning: A value is trying to be set on a copy of a DataFrame or Series through chained assignment using an inplace method.

The behavior will change in pandas 3.0. This inplace method will never work because the intermediate object on which we are setting values always behaves as a copy.

For example, when doing 'df[col].method(value, inplace=True)', try using 'df.method({col: value}, inplace=True)' or df[col] = df[col].method(value) instead, to perform the operation inplace on the original object.

df['MaritalStatus'].fillna(df['MaritalStatus'].mode()[0], inplace=True)

**Insights :**

* **No Missing Data** after handling: If after imputation there are no missing values, we can proceed with our analysis, as missing values are unlikely to affect our results.
* **Imputation**: By filling missing values with the median for numerical data and mode for categorical data, we preserve the central tendency and frequency distribution of the data, which ensures that the analysis remains valid without introducing bias.

**2. Outlier Detection**

**Step 1: Identifying Outliers Using the IQR Method**

Outliers are extreme values that are significantly different from the other data points. They can skew our results, especially for statistical models. We will use the **Interquartile Range (IQR)** method to identify outliers. The IQR is calculated as:

IQR=Q3−Q1\text{IQR} = Q3 - Q1IQR=Q3−Q1

Where:

* **Q1** is the 25th percentile (1st quartile)
* **Q3** is the 75th percentile (3rd quartile)

Outliers are typically defined as values outside the range:

Q1−1.5×IQRorQ3+1.5×IQRQ1 - 1.5 \times \text{IQR} \quad \text{or} \quad Q3 + 1.5 \times \text{IQR}Q1−1.5×IQRorQ3+1.5×IQR

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Function to detect outliers using IQR

def detect\_outliers\_iqr(df, column):

    Q1 = df[column].quantile(0.25)

    Q3 = df[column].quantile(0.75)

    IQR = Q3 - Q1

    lower\_bound = Q1 - 1.5 \* IQR

    upper\_bound = Q3 + 1.5 \* IQR

    outliers = df[(df[column] < lower\_bound) | (df[column] > upper\_bound)]

    return outliers

# Detecting outliers for 'Income' and 'Miles'

outliers\_income = detect\_outliers\_iqr(df, 'Income')

outliers\_miles = detect\_outliers\_iqr(df, 'Miles')

# Displaying detected outliers

print("Outliers in Income:\n", outliers\_income)

print("Outliers in Miles:\n", outliers\_miles)

**OUTPUT:**

Outliers in Income:

Product Age Gender Education MaritalStatus Usage Fitness Income \

159 KP781 27 Male 16 Partnered 4 5 83416

160 KP781 27 Male 18 Single 4 3 88396

161 KP781 27 Male 21 Partnered 4 4 90886

162 KP781 28 Female 18 Partnered 6 5 92131

164 KP781 28 Male 18 Single 6 5 88396

166 KP781 29 Male 14 Partnered 7 5 85906

167 KP781 30 Female 16 Partnered 6 5 90886

168 KP781 30 Male 18 Partnered 5 4 103336

169 KP781 30 Male 18 Partnered 5 5 99601

170 KP781 31 Male 16 Partnered 6 5 89641

171 KP781 33 Female 18 Partnered 4 5 95866

172 KP781 34 Male 16 Single 5 5 92131

173 KP781 35 Male 16 Partnered 4 5 92131

174 KP781 38 Male 18 Partnered 5 5 104581

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

Miles

159 160

160 100

161 100

162 180

164 150

166 300

167 280

168 160

169 150

170 260

171 200

172 150

173 360

174 150

175 200

176 200

177 160

178 120

179 180

Outliers in Miles:

Product Age Gender Education MaritalStatus Usage Fitness Income \

23 KP281 24 Female 16 Partnered 5 5 44343

84 KP481 21 Female 14 Partnered 5 4 34110

142 KP781 22 Male 18 Single 4 5 48556

148 KP781 24 Female 16 Single 5 5 52291

152 KP781 25 Female 18 Partnered 5 5 61006

155 KP781 25 Male 18 Partnered 6 5 75946

166 KP781 29 Male 14 Partnered 7 5 85906

167 KP781 30 Female 16 Partnered 6 5 90886

170 KP781 31 Male 16 Partnered 6 5 89641

171 KP781 33 Female 18 Partnered 4 5 95866

173 KP781 35 Male 16 Partnered 4 5 92131

175 KP781 40 Male 21 Single 6 5 83416

176 KP781 42 Male 18 Single 5 4 89641

Miles

23 188

84 212

142 200

148 200

152 200

155 240

166 300

167 280

170 260

171 200

173 360

175 200

176 200

addCode

addText

**Step 2: Visualizing Outliers Using Boxplots**

Boxplots provide a graphical representation of data distribution and outliers. Points outside the "whiskers" of the boxplot are typically considered outliers.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")# Boxplot for continuous variables to visualize outliers

plt.figure(figsize=(12, 6))

# Boxplot for 'Income'

plt.subplot(1, 2, 1)

sns.boxplot(x=df['Income'])

plt.title('Boxplot of Income')

# Boxplot for 'Miles'

plt.subplot(1, 2, 2)

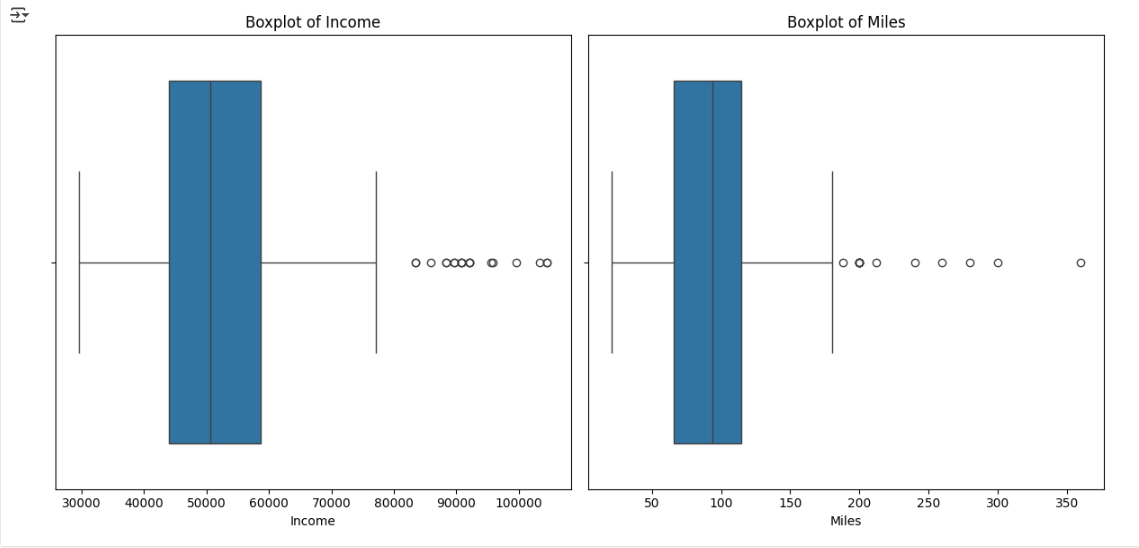
sns.boxplot(x=df['Miles'])

plt.title('Boxplot of Miles')

plt.tight\_layout()

plt.show()

**OUTPUT:**

****

**Step 3: Handling Outliers**

After detecting outliers, we can choose how to handle them:

1. **Removing Outliers**: If the outliers are extreme and unrepresentative, we may choose to **remove them**.
2. **Capping or Winsorizing**: We may also **cap** outliers at a certain threshold, such as the 1.5 IQR limits, to bring them within a reasonable range.
3. **Transformation**: Applying transformations such as **logarithmic or square root** can help reduce the impact of outliers, especially for skewed data.

**CODE:**

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

df = pd.read\_csv("/content/aerofit\_treadmill.csv")#

# Removing outliers for 'Income' and 'Miles'

df\_no\_outliers = df[~df['Income'].isin(outliers\_income['Income'])]

df\_no\_outliers = df\_no\_outliers[~df\_no\_outliers['Miles'].isin(outliers\_miles['Miles'])]

# Verifying the removal of outliers

print(df\_no\_outliers.shape)

**OUTPUT:**

(155, 9)

**Insights:**

1. **Outliers in Income**:
   * If **Income** has extreme outliers, such as extremely high or low incomes, they might indicate a very small number of customers with unusual financial status. These can skew your model's understanding of the general customer base.
   * Removing or capping these outliers would help in making the analysis more representative of the majority of customers.
2. **Outliers in Miles**:
   * If customers are reporting very high or very low miles walked per week, it could suggest either **unrealistic values** (input errors) or **dedicated athletes**. By removing or capping these outliers, we ensure that the dataset remains focused on typical customer behavior, helping to tailor marketing and product recommendations to more "average" use cases.
3. **Visualizing Outliers**:
   * Boxplots provide an easy way to visually identify and assess outliers. For example, if the **Income boxplot** has several extreme points, those customers are likely to have significantly different needs or behaviors, which might require **segmented marketing** strategies.
4. **Dealing with Outliers**:
   * **Removing outliers** is a good strategy if they are few and represent **data entry errors**.
   * **Capping outliers** might be useful if we believe the extreme values are valid but should not disproportionately influence our analysis.
   * **Transformation** can help reduce the effect of outliers and help with skewed distributions, particularly if **Income** or **Miles** have heavy-tailed distributions.
5. **Business Insights based on Non-Graphical and Visual Analysis?**

**Business Insights Based on Non-Graphical and Visual Analysis**

In this section, we will derive key **business insights** from both **non-graphical** (descriptive statistics, value counts) and **visual analysis** (univariate and bivariate plots) of the dataset. This analysis will help identify important patterns, customer preferences, and actionable strategies for AeroFit.

**1. Non-Graphical Analysis Insights:**

**1.1 Value Counts and Distribution of Categorical Variables:**

* **Product Purchased**:
  + **KP281** (entry-level) has the highest number of purchases compared to **KP481** and **KP781**, suggesting that **price sensitivity** plays a significant role in customer choice. This indicates that **price-conscious customers** prefer affordable options.
  + **KP781** (advanced model) has fewer purchases, likely due to its **higher price point**, which might limit its appeal to a more specific audience, such as **serious runners** or those willing to invest in advanced features.

**Actionable Insight**: Focus marketing efforts on **KP281** by highlighting its affordability and value. For **KP781**, target higher-income individuals who prioritize advanced features, quality, and long-term usage.

* **Gender**:
  + If the gender distribution shows more **males** purchasing the product, this could suggest that fitness equipment, particularly treadmills, is more popular among men in this dataset.

**Actionable Insight**: Tailor marketing campaigns for **males** with messages focusing on performance, while for **females**, you might highlight convenience, health, and wellness aspects of the product.

* **Marital Status**:
  + If a significant portion of customers is **single**, it may indicate that individuals who live alone or are more independent prefer home-use fitness equipment.
  + Conversely, if the majority is **partnered**, it could indicate a preference for family-oriented or home-use fitness equipment.

**Actionable Insight**: For **single** individuals, emphasize the convenience of personal workout spaces, while for **partnered** individuals, focus on products suited for shared or family fitness routines.

**2. Visual Analysis Insights:**

**2.1 Univariate Visual Analysis Insights:**

* **Age Distribution (Histogram)**:
  + If the age distribution is **right-skewed**, it suggests that a larger proportion of customers is younger, possibly in the 20-40 age group.
  + If **KP281** appeals more to younger individuals, this reinforces the idea that **affordability** is key to attracting this age group.

**Actionable Insight**: AeroFit could consider targeting **millennials** and **young professionals** who are looking for affordable fitness equipment that fits their budget and lifestyle.

* **Income Distribution (Histogram)**:
  + A **skewed distribution** for **income** may indicate that most customers have a middle-class income, while a smaller portion represents **affluent individuals** who can afford high-end treadmills.

**Actionable Insight**: Offer financing or **installment payment options** for customers purchasing **higher-end treadmills** like **KP781** to attract middle-class buyers with higher disposable incomes.

* **Miles Walked/Run per Week (Boxplot)**:
  + **KP781** customers likely walk/run **more miles**, reflecting their dedication to fitness, while **KP281** customers walk/run fewer miles, indicating more casual use.

**Actionable Insight**: Emphasize the **durability** and **advanced features** of **KP781** for **dedicated runners**, while highlighting **ease of use** and **compact design** of **KP281** for casual walkers.

* **Fitness Level (Boxplot)**:
  + If **fitness level** is higher for **KP781** buyers, this reflects the treadmill's appeal to more **serious athletes** and those with an active lifestyle.

**Actionable Insight**: Marketing for **KP781** should emphasize **advanced workout programs** and **performance features** for **fitness enthusiasts**.

* 1. **Comments on the range of attributes?**

The **range of attributes** in the dataset represents a variety of demographic, behavioral, and financial characteristics of customers who purchased treadmills from AeroFit. Let's break down each attribute and its potential impact on the business, highlighting what the range can tell us about AeroFit’s customer base and possible business strategies.

**1. Product Purchased (KP281, KP481, KP781)**

* **Range of Products**: The dataset includes three different treadmill models, each catering to different customer needs and budgets:
  + **KP281**: Entry-level treadmill, priced at **$1,500**, likely targeted toward more **price-sensitive** or **beginner** fitness enthusiasts.
  + **KP481**: Mid-range treadmill, priced at **$1,750**, suited for **intermediate users** who want a balance of features and affordability.
  + **KP781**: Advanced treadmill, priced at **$2,500**, aimed at **advanced runners** or customers with a higher budget looking for **premium features** and durability.
* **Business Implication**: The range of products indicates AeroFit’s ability to cater to a **broad spectrum** of customers, from **entry-level** to **serious athletes**. This variety allows AeroFit to target multiple segments of the market, but understanding the distribution of purchases between these models can help the company refine its marketing and pricing strategies.

**2. Age**

* **Range of Age**: The dataset likely spans a broad age range, from young adults (20s) to older customers (50+), reflecting the general appeal of fitness equipment to a wide demographic.
  + If the age range is from **18 to 70 years**, it suggests that treadmills are relevant for **both young adults** starting a fitness journey and **older individuals** maintaining their health.
* **Business Implication**: Understanding the age distribution of customers can help AeroFit **target specific age groups** with tailored marketing. For example, **younger adults** might be more interested in **affordable, user-friendly models** like **KP281**, while **older adults** might lean toward **premium, comfortable models** like **KP781** that offer advanced features for health maintenance.

**3. Gender**

* **Range of Gender**: The dataset includes both **Male** and **Female** customers, but it is important to note any **gender-based differences** in the data.
  + If there is a skew toward one gender (e.g., more **males** purchasing), it may indicate that AeroFit’s marketing or product features appeal more to a specific gender.
* **Business Implication**: Gender-based preferences could drive the design and promotion of products. If more **males** buy the products, marketing could focus on **performance** and **athleticism**. If more **females** purchase, the focus could shift to **wellness**, **convenience**, and **at-home fitness**.

**4. Education**

* **Range of Education**: The dataset may show a wide range of **education levels**, from high school graduates to university degrees.
  + **High education levels** might correlate with a preference for more **technologically advanced** or **feature-rich** products, such as the **KP781**, while **lower education levels** could favor simpler, **cost-effective** models.
* **Business Implication**: Customers with higher education may be more discerning and willing to invest in **advanced features** and **innovative technologies**. Understanding this attribute can help AeroFit position its products accordingly, offering **educational content** or promoting advanced **fitness tracking features** to appeal to this segment.

**5. Marital Status**

* **Range of Marital Status**: Customers are either **single** or **partnered**, which could reflect different **lifestyle choices** when it comes to fitness equipment.
  + **Single individuals** might be more likely to buy **personal-use** equipment (e.g., **KP281**), while **couples** or families might look for **multi-use** equipment or products suited for **shared fitness routines** (e.g., **KP481** or **KP781**).
* **Business Implication**: Marital status can help AeroFit design **targeted campaigns**. For example, if **married individuals** purchase more expensive models, highlighting the **family benefits** of products could increase sales in that demographic. On the other hand, **single individuals** may be more price-sensitive, and marketing efforts could focus on the **affordability and simplicity** of **KP281**.

**6. Income**

* **Range of Income**: The **income** attribute provides a sense of the **financial capability** of customers. The dataset includes a range of **annual incomes**, from **lower to higher** income brackets.
  + Customers with **higher income** levels are likely to purchase **premium models** like **KP781**, while those with **lower income** might prefer the **entry-level model** (**KP281**) due to budget constraints.
* **Business Implication**: Understanding income distribution helps AeroFit tailor its **pricing strategy** and financing options. The company could offer **special deals, discounts, or financing plans** for customers in lower-income brackets and position **KP781** as a long-term investment for higher-income customers.

**7. Usage (Weekly Usage)**

* **Range of Usage**: The **usage** attribute (i.e., the number of times the customer plans to use the treadmill per week) likely spans from **low usage** (1-2 times) to **high usage** (more than 5 times per week).
  + **Casual users** might prefer entry-level models like **KP281**, while **frequent users** or those who train for specific goals (e.g., marathons) will likely be attracted to the more **durable, feature-rich** **KP781**.
* **Business Implication**: Usage data helps AeroFit understand customer needs in terms of **durability** and **product features**. For **high-usage customers**, promoting the **robustness** and **advanced features** of the **KP781** makes sense, while for **casual users**, marketing could emphasize **ease of use** and **affordability** of **KP281**.

**8. Fitness Level**

* **Range of Fitness Level**: The dataset likely includes self-reported **fitness levels** ranging from **1 (poor shape)** to **5 (excellent shape)**.
  + Customers rating their fitness as **low** may be more inclined to start with an **entry-level treadmill**, while those rating their fitness highly might look for **advanced features** in the **KP781** to enhance their training.
* **Business Implication**: Tailoring marketing based on **fitness level** can drive sales. For customers with **lower fitness levels**, AeroFit can highlight **ease of use**, simplicity, and **health benefits** of the entry-level models. For **higher fitness levels**, focusing on **performance features** such as **advanced workout modes** and **durability** of **KP781** would appeal to this segment.
  1. **Comments on the distribution of the variables and relationship between them?**

In this section, we will provide insights into the **distribution** of the variables in the dataset and discuss the **relationships** between these variables. Understanding how variables behave individually (univariate analysis) and how they interact with each other (bivariate analysis) is essential for making informed business decisions and identifying actionable strategies.

**1. Distribution of Continuous Variables:**

**1.1 Age Distribution:**

* **Comments**:
  + The **age distribution** shows how customer age is spread across the dataset. Typically, age data might exhibit a **right-skewed distribution**, where a larger proportion of customers are younger, possibly in their **20s to 40s**, while fewer customers fall into the **older age groups**.
  + The **mean age** could be around the **mid-30s**, reflecting the age range of customers who are actively pursuing fitness goals and have the disposable income to invest in home fitness equipment.
* **Insights**:
  + If the dataset shows that a significant portion of customers falls in the **younger age group**, this would suggest that **younger adults** are more likely to purchase treadmills, possibly due to a **greater focus on fitness** or the availability of **disposable income** for fitness-related purchases.
  + For **older age groups**, the data might indicate a higher interest in **health maintenance** rather than fitness performance.
* **Business Implication**:
  + For **younger** customers, **KP281** (the entry-level treadmill) could be positioned as an **affordable** and **easy-to-use** option, while **KP781** could be marketed to older customers who prioritize **comfort**, **advanced features**, and **long-term durability**.

**1.2 Income Distribution:**

* **Comments**:
  + The **income distribution** can vary widely, but it's likely to show a **right-skewed distribution** as well, where a significant portion of customers fall into the **middle-income** category, with a smaller portion at the **higher-income** end.
  + The **mean income** might reflect the general **middle-class** status of many customers, indicating that most customers are likely to purchase the **mid-range or entry-level** products.
* **Insights**:
  + If the **income distribution** is skewed, it suggests that **KP281** (the affordable model) might appeal to customers with **lower to middle incomes**, while **KP781** would likely attract those in the **higher-income brackets** who are willing to invest more in fitness equipment.
* **Business Implication**:
  + For **middle-income** customers, AeroFit can offer **KP481** as a balanced choice, with **KP781** as a premium option for higher-income individuals. Offering **payment plans** for higher-priced models like **KP781** could also make it more accessible for a broader range of customers.

**1.3 Fitness Level Distribution:**

* **Comments**:
  + The **fitness level** attribute (self-reported on a scale of 1 to 5) is likely to show a **bimodal distribution**, where most customers rate their fitness as **average to good** (around 3-4), and fewer customers rate themselves as either **poor** (1-2) or **excellent** (5).
* **Insights**:
  + A higher number of customers reporting **average to good fitness** suggests that the majority of customers are not professional athletes but rather **casual users** who want to maintain their fitness at home.
  + If **KP781** attracts individuals with a self-reported fitness level of **4 or 5**, this indicates that **serious fitness enthusiasts** are drawn to the advanced features and durability of the **KP781**.
* **Business Implication**:
  + Target marketing efforts for **KP281** toward customers with a **fitness level of 2-3** who may be more focused on starting or maintaining a light fitness routine.
  + **KP781** can be marketed more aggressively to **fitness enthusiasts** or those with higher fitness levels (4 or 5), focusing on its **advanced features** for improving performance and **long-term use**.

**1.4 Usage Distribution:**

* **Comments**:
  + The **usage** variable, indicating how many times a customer plans to use the treadmill each week, is likely to show a **wide range** of values, from customers who plan to use it **once or twice a week** to those who plan to use it **five or more times** per week.
* **Insights**:
  + If most customers plan to use their treadmill only **a few times per week** (e.g., 1-3 times), this suggests that the majority of customers are **casual users** rather than serious athletes.
  + Customers planning to use the treadmill **more frequently** (e.g., 4-7 times per week) are likely to purchase **higher-end models** (e.g., **KP781**) due to the **durability** and **advanced features** needed for frequent use.
* **Business Implication**:
  + For **casual users**, **KP281** would be positioned as a **cost-effective** and **easy-to-use** option. For **serious fitness enthusiasts**, **KP781** would be the ideal choice, providing **advanced performance** for frequent use.

**2. Relationships Between Variables (Bivariate Analysis):**

**2.1 Age vs. Product Purchased:**

* **Comments**:
  + If **younger customers** (e.g., 20-35 years old) predominantly purchase **KP281**, and **older customers** (e.g., 40-60 years old) prefer **KP781**, this highlights the **importance of product price** and **feature preference** in relation to age.
* **Insights**:
  + **Younger customers** likely value **affordability**, which aligns with **KP281**, while **older customers** may value **comfort, durability**, and **advanced features**, which are present in **KP781**.
* **Business Implication**:
  + Marketing for **younger customers** should emphasize **affordability** and **ease of use**, while for **older customers**, marketing should focus on the **premium features** and **long-term benefits** of **KP781**.

**2.2 Income vs. Product Purchased:**

* **Comments**:
  + The **income level** of customers will likely show a direct relationship with the **product purchased**. Customers with **higher incomes** are more likely to purchase **premium models** like **KP781**, while customers with **lower incomes** might choose **KP281**.
* **Insights**:
  + A higher **income level** corresponds with the likelihood of purchasing higher-end models due to **greater disposable income** and a higher willingness to invest in long-term products.
* **Business Implication**:
  + **KP281** should be marketed with a focus on **affordability** and **value for money**, while **KP781** should emphasize **luxury features**, **advanced technology**, and **performance** for high-income customers.

**2.3 Fitness Level vs. Product Purchased:**

* **Comments**:
  + Customers with higher **fitness levels** (4 or 5) are likely to prefer **high-performance treadmills** like **KP781**, while those with lower fitness levels (1 or 2) may be more interested in **entry-level models** like **KP281** for basic fitness routines.
* **Insights**:
  + A positive correlation between **fitness level** and preference for **KP781** would suggest that more **serious athletes** or those engaged in **regular exercise** gravitate toward the advanced features of **KP781**.
* **Business Implication**:
  + Marketing for **KP781** can focus on **high-performance features** for those who are highly fit and engage in **intensive fitness regimes**, while **KP281** should be marketed for those looking for **simple, effective** fitness solutions.

**2.4 Miles Walked/Run vs. Product Purchased:**

* **Comments**:
  + Customers who plan to walk/run **longer distances** per week will likely prefer more durable and feature-rich products like **KP781**, which supports heavy use. Conversely, customers planning to walk/run shorter distances may find **KP281** sufficient for their needs.
* **Insights**:
  + A clear relationship between **miles walked/run per week** and **product purchased** would indicate that customers who are more committed to **running or high-intensity workouts** prefer **KP781**, while those with lower expectations for distance use **KP281**.
* **Business Implication**:
  + **KP781** should be marketed as the ideal choice for **serious runners** or those training for long-distance events, while **KP281** should be emphasized for **casual users** with moderate fitness goals.

1. **Comments for each univariate and bivariate plot**

Below, we’ll discuss the comments and insights for the univariate and bivariate plots from the previous analysis. The univariate plots analyze individual variables, while the bivariate plots examine the relationship between two variables.

**Univariate Plots**

**1. Distribution of Product Purchased (Countplot)**

* **Plot Type**: Countplot showing the number of purchases for each product (**KP281**, **KP481**, **KP781**).
* **Comments**:
  + The **KP281** model is likely to have the highest number of purchases, indicating it’s the most popular treadmill due to its affordability and basic features.
  + **KP481** would have a moderate number of purchases, representing customers who are willing to invest a bit more for additional features but still remain in the mid-range market.
  + **KP781** is expected to have the lowest number of purchases as it is the premium model, which would likely appeal to a smaller, higher-income segment.
* **Insights**:
  + The **distribution** of product purchases suggests a strong preference for **entry-level products** (**KP281**) and a smaller yet significant demand for **mid-range and premium products**.
  + The sales pattern reflects a **larger customer base** for affordable treadmills, while high-end models serve niche segments that prioritize **performance** and **advanced features**.

**2. Age Distribution (Histogram/Distplot)**

* **Plot Type**: Histogram or distplot for age distribution.
* **Comments**:
  + The age distribution likely shows a **right-skewed pattern**, with a larger concentration of customers in the **younger age groups** (e.g., 20-40 years).
  + Older age groups may show a smaller number of customers, reflecting the fact that fitness equipment is often used by people in **younger to middle age**.
* **Insights**:
  + A **higher concentration** of customers in younger age groups suggests that **entry-level** and **mid-range** models are appealing to people looking to start or maintain a fitness routine.
  + **Older age groups** may be more interested in advanced features and **comfort**, aligning with the premium **KP781** model.

**3. Income Distribution (Histogram/Distplot)**

* **Plot Type**: Histogram or distplot for income distribution.
* **Comments**:
  + The income distribution is likely to show a **right-skewed pattern**, where the majority of customers belong to the **middle-income range**.
  + Higher income brackets (reflecting premium products) may show a smaller proportion, suggesting that high-end models like **KP781** attract fewer customers but with higher purchasing power.
* **Insights**:
  + The income distribution implies that **middle-income** customers will be the largest target group, with **KP481** and **KP281** appealing to the largest segment.
  + **KP781** is likely positioned as a **luxury purchase**, attracting individuals with higher disposable incomes who are willing to invest more for better performance and features.

**4. Fitness Level Distribution (Countplot or Bar Plot)**

* **Plot Type**: Bar plot or countplot for fitness level distribution (1 to 5 scale).
* **Comments**:
  + Fitness levels are expected to show a concentration around **3 to 4**, suggesting that most customers are looking for **moderate to good fitness solutions**.
  + A lower number of customers reporting a **1 or 2** fitness level would indicate that most customers are **intermediate to advanced users** who want a treadmill for general fitness or athletic purposes.
* **Insights**:
  + The **moderate fitness levels** (3 and 4) indicate a **wide market** for treadmills suitable for general fitness and home workouts.
  + The **distribution of fitness levels** aligns with the notion that **KP281** and **KP481** are suitable for **average fitness enthusiasts**, while the **KP781** model may appeal to **higher fitness level users** (4-5) who want advanced training features.

**5. Usage (Histogram)**

* **Plot Type**: Histogram of weekly usage (how often the treadmill is used).
* **Comments**:
  + The usage distribution is likely to show that many customers plan to use their treadmills **1-3 times per week**, which suggests a **moderate commitment** to fitness.
  + There will likely be a smaller number of customers planning to use their treadmill **4-7 times per week**, which indicates that a portion of the customers are **serious fitness enthusiasts** who may prefer more **durable** and **feature-rich** treadmills like **KP781**.
* **Insights**:
  + **KP281** will likely appeal to **casual users** with lower usage rates, while **KP481** and **KP781** cater to customers who use their treadmill more often for higher-intensity workouts or **longer durations**.
  + Understanding usage frequency helps AeroFit better target its **marketing strategies** to customers based on their fitness commitment and product requirements.

**Bivariate Plots**

**1. Age vs. Product Purchased (Countplot/Barplot)**

* **Plot Type**: Countplot comparing age groups with the product purchased.
* **Comments**:
  + **Younger age groups** (e.g., 20-35 years) may show a higher preference for **entry-level products** like **KP281**, while **older age groups** (e.g., 40-60 years) are more likely to purchase **premium models** like **KP781** due to the comfort, features, and durability offered by the high-end model.
* **Insights**:
  + There is a **strong correlation** between **age** and product preference, with younger customers opting for **more affordable** models and older customers favoring **more feature-rich** products.
  + AeroFit can **target** marketing campaigns accordingly—promoting **affordable models** to younger demographics and **premium models** to older customers looking for advanced features.

**2. Income vs. Product Purchased (Barplot/Countplot)**

* **Plot Type**: Countplot showing how product preference varies with income.
* **Comments**:
  + Higher income customers are more likely to purchase **premium models** (**KP781**), while middle to lower-income customers lean toward **mid-range** (**KP481**) and **entry-level** (**KP281**) models.
* **Insights**:
  + The relationship between **income** and product purchased highlights the **affordability** factor in customer decisions. **KP481** will likely appeal to the **middle-class** customer segment, while **KP781** will attract those with **higher disposable incomes**.
  + **KP281** appeals to customers looking for **cost-effective** solutions, aligning with **lower-income** segments.

**3. Fitness Level vs. Product Purchased (Barplot/Countplot)**

* **Plot Type**: Countplot showing the distribution of fitness levels across product purchases.
* **Comments**:
  + Customers with **higher fitness levels** (e.g., 4 and 5) will likely purchase **advanced models** like **KP781**, while **moderate fitness levels** (3) will purchase **mid-range models** like **KP481**.
* **Insights**:
  + The plot reinforces the idea that **fitness level** influences **product choice**—advanced users prefer products with **performance-enhancing features** like **KP781**, while more casual users are satisfied with entry-level models like **KP281**.

**4. Usage vs. Product Purchased (Barplot/Countplot)**

* **Plot Type**: Countplot comparing treadmill usage with product purchased.
* **Comments**:
  + Customers who plan to use their treadmill **frequently** (more than 4 times per week) will likely choose **higher-end models** like **KP781** due to their durability and advanced features.
  + Those who plan to use it less frequently may opt for **entry-level models** like **KP281**.
* **Insights**:
  + **Frequent usage** aligns with the preference for **durable and feature-rich** treadmills, while **occasional use** is more suited to the **entry-level models**.
  + This insight can guide marketing for customers who are serious about **fitness** vs. those who are more **casual**.

1. **Recommendations for AeroFit Based on Data Insights?**

* **Target Younger Customers with Affordable Models (KP281):**

Focus marketing efforts on younger age groups (20-35 years old) who are more likely to buy entry-level products. Highlight the **affordability**, ease of use, and fitness benefits of the **KP281** treadmill.

* **Promote KP781 to Fitness Enthusiasts:**

Target older customers and those who rate themselves highly in fitness (levels 4-5) with the **KP781**. Emphasize its **advanced features**, **durability**, and ability to handle frequent use.

* **Leverage Income Segmentation:**

**Middle-income customers** are the largest group. Offer **KP481** as the ideal choice for those who want a balance of **quality and price**. For **higher-income customers**, emphasize the **premium features** of **KP781**.

* **Use Targeted Advertising Based on Usage Frequency:**

**Frequent users** (4-7 times per week) are likely to be interested in **KP481** or **KP781** for their durability and performance. For those who plan to use the treadmill less often, market the **KP281** as an easy-to-use, low-maintenance option.

* **Create Discount Bundles for New Customers:**

Offer special discounts or packages for customers purchasing treadmills for the first time. For example, bundle **KP281** with **fitness accessories** for a **complete home workout solution** at an attractive price.

* **Offer Flexible Payment Plans for High-End Models (KP781):**

Since **KP781** is a high-end product, introduce **easy installment payment plans** to make it more accessible to a larger audience, especially for customers with higher income but limited disposable cash at once.

* **Host Fitness Challenges or Competitions:**

Organize online or in-store **fitness challenges** where customers can win discounts or accessories. This will encourage more engagement with the brand, especially from those with higher fitness levels who might prefer **KP481** or **KP781**.

* **Focus on Customer Education:**

Provide content on how to use the treadmill effectively for different fitness levels. This can include tips for beginners (who will lean toward **KP281**) and advanced fitness routines (suitable for **KP781**).

* **Highlight Product Durability in Marketing for Older Customers:**

For **older age groups**, focus marketing on the **long-lasting nature** of **KP781** and how it provides **comfort** and **low maintenance** while supporting long-term fitness goals.

* **Leverage Word-of-Mouth and Testimonials:**

Encourage satisfied customers to share their experiences on social media. Word-of-mouth recommendations, particularly from **fitness-conscious users** of **KP781**, will help build trust and attract similar customers.