Data Analysis Report: AeroFit Treadmill Buyer Profile

1. Introduction

1.1 Business Context

AeroFit is a renowned player in the fitness equipment market, particularly known for its diverse range of treadmills designed for various customer segments. In today's fitness-conscious world, treadmills are one of the most popular exercise machines, providing convenience for both casual walkers and serious runners. As part of AeroFit's efforts to refine its product offerings and better understand customer behavior, the company is keen on analyzing its product portfolio and customer base. AeroFit offers three main models of treadmills, each tailored to different customer needs. These models—KP281, KP481, and KP781—vary not only in price but also in features. Understanding who purchases which treadmill, and why, is central to helping AeroFit optimize its marketing strategies and improve its recommendations to prospective customers.

The KP281, which is AeroFit's entry-level model, is priced at $1,500. This treadmill caters to individuals who are just starting their fitness journey or who prefer a more affordable option. On the other end of the spectrum, the KP781, priced at $2,500, is packed with advanced features that appeal to fitness enthusiasts with higher incomes and greater exercise demands. The KP481, priced at $1,750, sits in the middle, offering a balanced blend of features and affordability. The goal of this analysis is to provide AeroFit with a detailed understanding of the key demographic factors influencing purchases of each model. This insight will help AeroFit develop targeted marketing strategies and refine its product positioning.

1.2 Problem Statement

The problem AeroFit faces is rooted in a common challenge for companies with diverse product portfolios: identifying the right target audience for each product. The market research team at AeroFit has collected data on individuals who purchased a treadmill from AeroFit stores over the last three months. However, simply collecting data is not enough; the company needs a deep analysis to extract meaningful insights. By studying the demographic and behavioral characteristics of these customers, AeroFit can tailor its marketing messages and recommendations for future buyers. Specifically, this analysis aims to address questions such as:

- What are the distinguishing characteristics of customers who buy each treadmill model?
- Are there clear demographic patterns (e.g., age, income, fitness level) associated with each treadmill?
- How can AeroFit improve its product recommendations based on customer characteristics?

These questions form the core of the problem that this report seeks to address through exploratory data analysis (EDA).

1.3 Objectives of the Report

The main objective of this report is to perform a comprehensive exploratory data analysis of the provided dataset, which includes customer demographic information, their treadmill preferences, and their fitness behaviors. By systematically analyzing this data, we will uncover patterns and trends that will help AeroFit make data-driven decisions in its marketing and product development processes. The specific goals of this analysis include:

1. **Data Exploration**: To thoroughly examine the dataset for missing values, duplicates, and any inconsistencies, and to ensure the data is clean and ready for analysis.
2. **Statistical Summary**: To generate both non-graphical and graphical summaries of the dataset, including measures of central tendency and distribution for

numerical features like age and income, and frequency distributions for categorical features like product purchased and gender.

3. **Visualization**: To create insightful visualizations that will help AeroFit easily interpret the patterns in the data, including histograms for numerical features and count plots for categorical features.
4. **Correlation and Outlier Analysis**: To identify relationships between variables (e.g., age and product purchased) and detect any outliers in the data using appropriate statistical methods.
5. **Conditional Probability Analysis**: To calculate the probabilities associated with various customer demographics and their likelihood of purchasing a particular treadmill, which will help in creating more personalized marketing strategies.
6. **Actionable Insights and Recommendations**: To provide AeroFit with a detailed set of recommendations based on the findings, aimed at enhancing their product recommendations and marketing tactics.

Ultimately, this report will serve as a comprehensive guide for AeroFit to better understand its customer base and align its product offerings with customer preferences. The methodology employed in this report follows a structured approach, starting with data cleaning and exploration, moving to statistical analysis and visualization, and culminating in actionable insights based on conditional probability analysis.

2. Data Exploration and Preprocessing

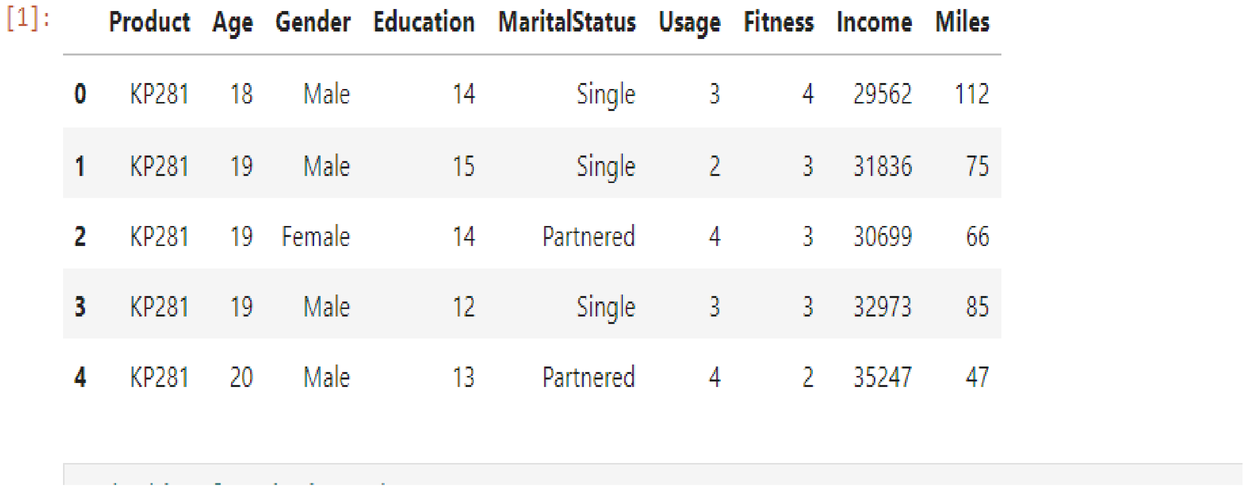
2.1 Dataset Overview

The dataset provided by AeroFit contains key information about customers who purchased one of the company's treadmills over the past three months. The dataset includes nine columns, each representing different customer attributes or purchase behaviors. The columns include demographic information such as age, gender, and marital status, as well as behavioral data such as weekly treadmill usage, self-reported fitness levels, and annual income. Understanding the structure of this dataset is the first step in performing a meaningful exploratory data analysis. The data is stored in a CSV (Comma-Separated Values) file, which is a common format for datasets in business analytics.

Code Implementation: Importing the Dataset

```
import pandas as pd
# Load the dataset
df = pd.read_csv('aerofit_treadmill_data.csv')  (this will be modified depending on the location of the file)
# Preview the first few rows of the dataset
df.head()
```

In this initial step, we use the pandas library to load the dataset from a CSV file into a DataFrame. The read_csv() function reads the data, and the head() function allows us to display the first five rows of the dataset, which helps in understanding the layout and structure of the data. Each row represents an individual customer, while the columns contain the attributes and the purchase data of that customer.



[1]:

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

Explanation

After importing the data, we can see that the dataset contains the following columns:

- **Product**: This column specifies the treadmill model purchased by the customer. The three models available are KP281, KP481, and KP781.
- **Age**: The customer's age is provided in this column, measured in years.
- **Gender**: This column specifies the gender of the customer, with possible values of “Male” or “Female.”
- **Education**: The number of years of formal education completed by the customer.
- **MaritalStatus**: This column specifies the customer’s marital status, with possible values of “Single” or “Partnered.”
- **Usage**: The number of times per week that the customer expects to use the treadmill.
- **Fitness**: The customer’s self-rated fitness level on a scale of 1 to 5, where 1 is the lowest fitness level and 5 is the highest.
- **Income**: The customer’s annual income, measured in US dollars.
- **Miles**: This column indicates the number of miles the customer expects to walk or run each week on the treadmill.

The dataset has 180 rows, meaning that there are 180 unique customer records. This will provide enough data to perform a robust analysis, but not so much data that it becomes cumbersome to work with. Additionally, the mix of categorical and numerical data will allow us to perform a wide range of analyses, from basic frequency counts to more sophisticated statistical mod=xels.

2.2 Data Cleaning

Before performing any analysis, it’s essential to ensure that the dataset is clean. Cleaning involves checking for missing values, duplicate records, and ensuring that the data types are appropriate for the analyses we want to conduct. Without clean data, any analysis performed could lead to misleading or incorrect results.

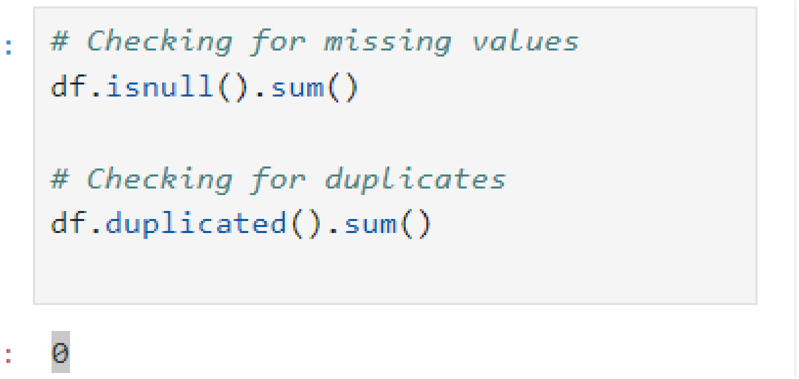
Code Implementation: Checking for Missing Values and Duplicates

```
# Checking for missing values
df.isnull().sum()

# Checking for duplicates
df.duplicated().sum()
```

Explanation

The isnull() function from pandas is used to check for any missing values in the dataset. Missing values can disrupt the analysis and lead to incorrect conclusions. The .sum() method is applied to aggregate the count of missing values per column. This helps identify if there are any columns with incomplete data. Missing data can be handled in various ways, such as filling in with default values or simply removing rows with missing data, depending on the analysis and the extent of missing information.

Similarly, the duplicated() function is used to check for duplicate rows in the dataset. Duplicate entries can skew the analysis by over-representing certain data points, so it’s essential to identify and remove any duplicates before proceeding.



```
:   # Checking for missing values
    df.isnull().sum()

    # Checking for duplicates
    df.duplicated().sum()
```

```
:   0
```

Results and Analysis

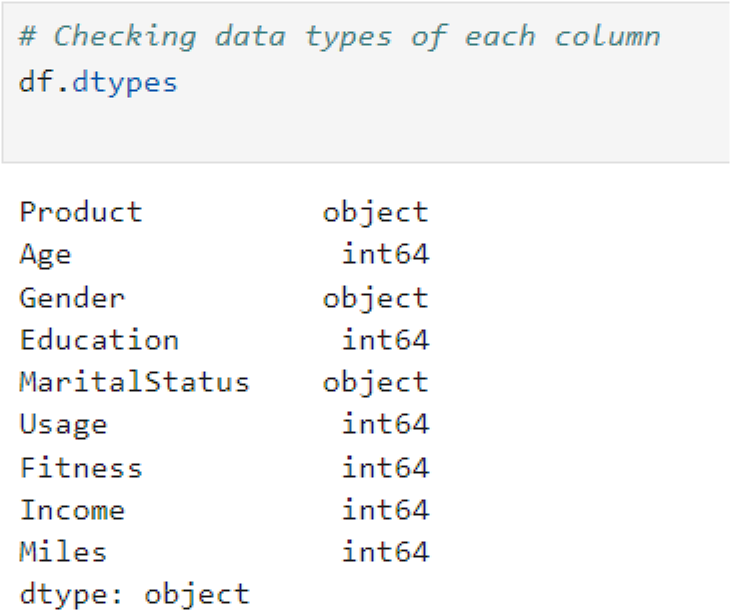
In this case, after running the code, the dataset shows that there are no missing values across any of the columns. Each column contains 180 entries, which means the dataset is complete, and we do not need to worry about missing data. Additionally, there are no duplicate rows in the dataset, which means each customer record is unique. This confirms that the data is clean and ready for further analysis.

2.3 Data Types and Structure

Once the dataset is confirmed to be free of missing values and duplicates, the next step is to check the data types of each column. It is crucial that each column has the correct data type (e.g., numerical, categorical) because this ensures that the appropriate statistical and visualization methods can be applied to each feature. For instance, numerical operations should only be applied to columns with integer or float data types, while categorical operations should be reserved for string-based columns.

Code Implementation: Checking Data Types

```
# Checking data types of each column
df.dtypes
```



```
# Checking data types of each column
df.dtypes
```

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

Explanation

The dtypes attribute of the DataFrame returns the data type of each column. We need to verify that numerical columns like Age, Income, Usage, and Miles are correctly identified as integer or float types, while categorical columns like Product, Gender, and MaritalStatus are treated as object (string) types. Ensuring that the data types are correct allows for smoother analysis and prevents errors during data manipulation and visualization.

Results

The dataset shows the following data types:

Product: Object (categorical)

Age: Int64 (numerical)

Gender: Object (categorical)

Education: Int64 (numerical)

MaritalStatus: Object (categorical)

Usage: Int64 (numerical)

Fitness: Int64 (numerical)

Income: Int64 (numerical)

Miles: Int64 (numerical)

The data types are appropriately assigned, meaning there is no need for any conversions or corrections. Numerical values like Age, Income, and Usage are correctly identified as integers, while categorical data like Product and Gender are identified as objects.

2.4 Data Structure and Initial Observations

At this stage, we have successfully imported, cleaned, and verified the data types in the dataset. Now, it is useful to conduct a basic exploration of the structure of the data by looking at the shape of the dataset, which tells us how many rows and columns it contains. Additionally, looking at the unique values in each categorical column can provide initial insights into the data.

Code Implementation: Checking the Shape of the Dataset

```
# Checking the shape of the dataset (number of rows and columns)
df.shape
```

```
(180, 9)
```

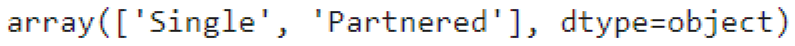
This command will return a tuple with two values: the number of rows and the number of columns in the dataset. Knowing the shape of the dataset is essential for understanding the scale of the analysis. With 180 rows and 9 columns, the dataset is manageable for in-depth analysis.

Results

The output indicates that the dataset has 180 rows and 9 columns. This is a relatively small dataset, which means the analysis can be conducted quickly and efficiently without the need for advanced computational techniques to handle large datasets. Furthermore, the dataset contains a mix of categorical and numerical variables, making it ideal for both statistical and visual analyses.

Code Implementation: Exploring Unique Values in Categorical Columns

```
# Exploring unique values in categorical columns
df['Product'].unique()
df['Gender'].unique()
df['MaritalStatus'].unique()
```



```
array(['Single', 'Partnered'], dtype=object)
```

Explanation

The unique() function allows us to see the distinct values in each categorical column. By doing this, we can ensure that the data is consistent and free from any unexpected categories (e.g., misspelled categories or incorrect entries). For example, in the Gender column, we expect to see only “Male” and “Female,” and in the Product column, we expect only “KP281,” “KP481,” and “KP781.”

Results

- **Product**: The unique values are KP281, KP481, and KP781, which is expected, as these are the three treadmill models offered by AeroFit.
- **Gender**: The unique values are Male and Female, which aligns with the expected binary gender classification in the dataset.
- **MaritalStatus**: The unique values are Single and Partnered, indicating that the dataset correctly captures the two main categories of marital status.

These results confirm that the categorical data is clean and contains the expected values. There are no unexpected categories or missing labels, meaning that the dataset is consistent and ready for further analysis.

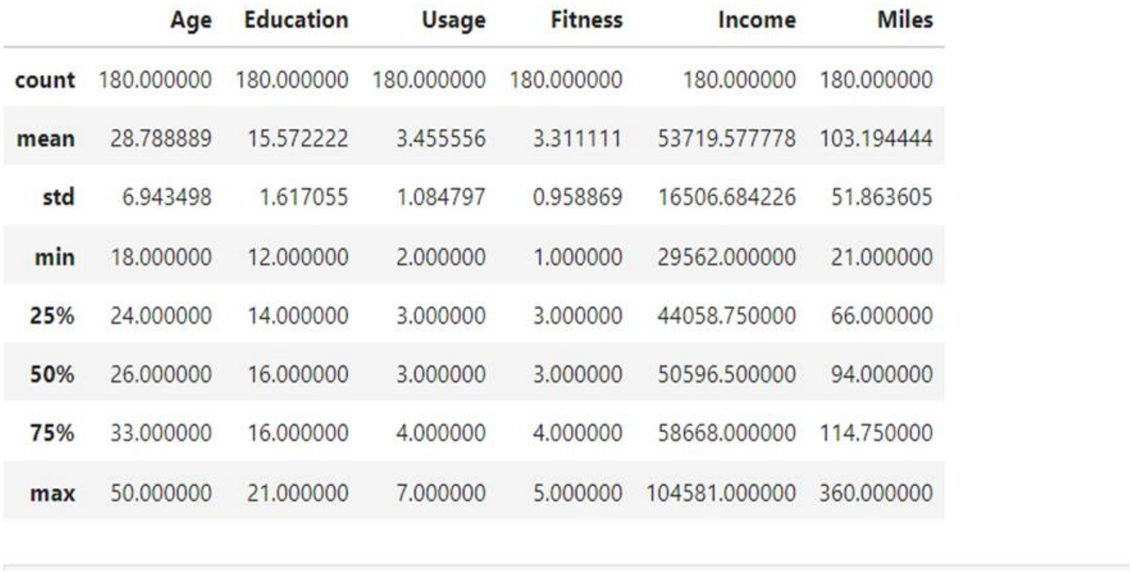
3. Statistical Summary

3.1 Statistical Summary of Numerical Features

The next step in the exploratory data analysis process is to generate a statistical summary of the numerical features in the dataset. This includes calculating measures such as the mean, median, standard deviation, minimum, and maximum for each numerical column. These statistics provide insights into the central tendencies and variability of the data, which is important for understanding the characteristics of AeroFit’s customers.

Code Implementation: Generating Summary Statistics for Numerical Features

```
# Summary statistics for numerical columns
df.describe()
```



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

Explanation

The describe () function in pandas generates a summary of the key statistics for each numerical column. This includes:

- **Count**: The number of non-null entries in each column.
- **Mean**: The average value for each numerical feature.

- **Standard Deviation (std)**: The spread or variability of the data.
- **Minimum (min)**: The smallest value in the column.
- **25th Percentile (25%)**: The value below which 25% of the data lies.
- **50th Percentile (50%)**: The median value of the column.
- **75th Percentile (75%)**: The value below which 75% of the data lies.
- **Maximum (max)**: The largest value in the column.

Results

The summary statistics for the numerical features provide the following insights:

- **Age**: The ages of AeroFit customers range from 18 to 50 years old, with a mean age of 28.8 years. The standard deviation is 6.94 years, indicating that most customers are relatively young, with some variation in age.

- **Income**: The annual income of customers ranges from $29,562 to $104,581, with an average income of $53,720. The wide range in income suggests that AeroFit serves a diverse customer base in terms of financial means.

- **Usage**: The number of times per week that customers expect to use their treadmill ranges from 2 to 7, with an average of 3.45 times per week. This suggests that most customers plan to use their treadmills moderately, a few times a week.

- **Fitness**: The self-reported fitness levels range from 1 to 5, with a mean of 3.31. This indicates that, on average, customers rate themselves as having moderate fitness levels.

- **Miles**: The number of miles customers expect to walk or run each week ranges from 21 to 360, with a mean of 103.2 miles. The high maximum value suggests that some customers are extremely active, while the lower values are more reflective of casual treadmill users.

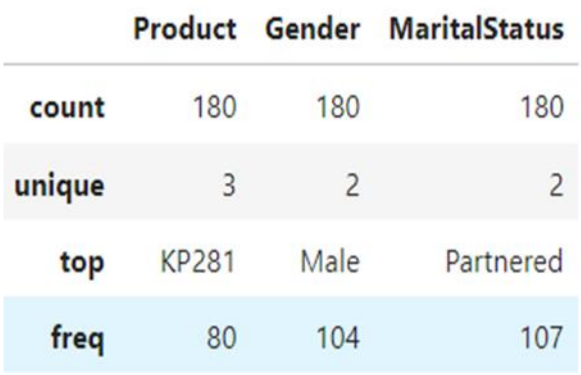
These summary statistics provide an initial understanding of AeroFit's customer base. Customers tend to be in their late 20s, with moderate income levels and a typical usage pattern of a few times per week. The variation in miles run per week highlights the diversity in fitness goals among the customers.

3.2 Statistical Summary of Categorical Features

In addition to analyzing numerical features, it is also essential to examine the categorical features in the dataset. These features include the product purchased, the customer's gender, and marital status. Understanding the distribution of these categorical features will provide insights into the preferences and demographics of AeroFit's customers.

Code Implementation: Summary Statistics for Categorical Features

```
# Summary statistics for categorical columns
df.describe(include=['object'])
```



	Product	Gender	MaritalStatus
count	180	180	180
unique	3	2	2
top	KP281	Male	Partnered
freq	80	104	107

Explanation

The describe() function, when used with the include=['object'] parameter, generates summary statistics for categorical columns. This summary includes:

- **Count**: The number of non-null entries in each column.
- **Unique**: The number of unique values in the column.
- **Top**: The most frequent value (mode) in the column.
- **Frequency (freq)**: The number of occurrences of the most frequent value.

Results

The summary statistics for the categorical features provide the following insights:

- **Product**: The most frequent treadmill model purchased is the KP281, with 80 customers (44.4%) opting for this entry-level model. This suggests that the KP281 is the most popular option among AeroFit's customers.

- **Gender**: More males (104, or 57.8%) than females (76, or 42.2%) purchased treadmills. This gender imbalance could indicate that men are more likely to invest in home fitness equipment, or it could reflect broader societal trends in fitness equipment ownership.

- **Marital Status**: The majority of customers are partnered (107, or 59.4%), while 73 customers (40.6%) are single. This distribution suggests that partnered individuals may be more likely to invest in home fitness equipment, potentially for joint use with their partner. Alternatively, the higher rate of partnered customers could reflect a broader trend in purchasing decisions made by families or couples, as they may have more disposable income compared to single individuals.

These initial findings from the statistical summary of both numerical and categorical features provide AeroFit with a solid foundation for understanding its customer demographics. The popularity of the KP281 model, the gender imbalance, and the higher rate of partnered customers are all important insights that can guide further analysis.

4. Visualizing the Data

Visualization is a crucial step in exploratory data analysis (EDA) as it allows us to see patterns, trends, and relationships within the dataset that may not be immediately obvious from summary statistics alone. In this section, we will use various types of plots to explore both univariate and bivariate relationships in the data. By doing so, we aim to uncover more detailed insights about AeroFit's customers and their treadmill preferences.

4.1 Univariate Analysis: Visualizing Numerical Features

Univariate analysis involves examining one variable at a time. For numerical features, histograms and boxplots are the most effective tools for understanding the distribution of data, including measures such as central tendency, spread, and the presence of any outliers.

Code Implementation: Visualizing Numerical Features

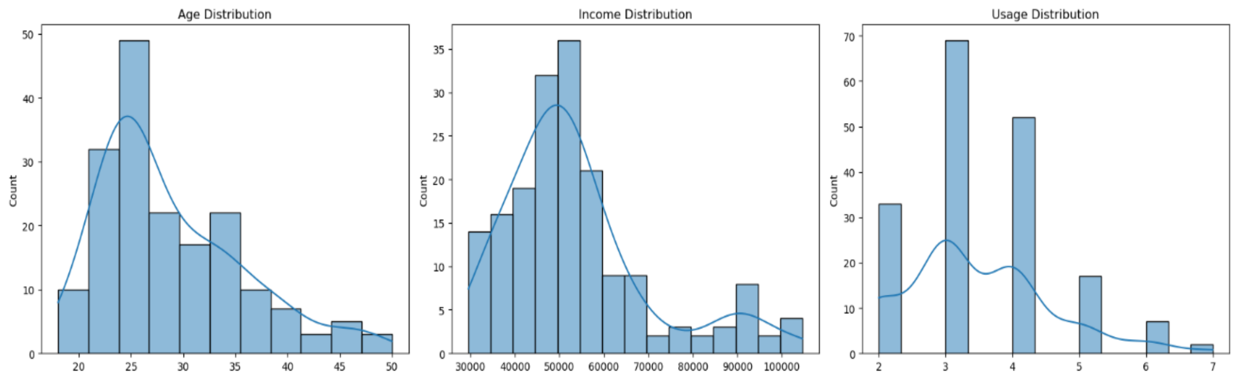
```
import matplotlib.pyplot as plt
import seaborn as sns

# Plotting histograms for numerical features
fig, axes = plt.subplots(2, 3, figsize=(18, 10))

# Age Distribution
sns.histplot(df['Age'], kde=True, ax=axes[0, 0])
axes[0, 0].set_title('Age Distribution')

# Income Distribution
sns.histplot(df['Income'], kde=True, ax=axes[0, 1])
axes[0, 1].set_title('Income Distribution')

# Usage Distribution
sns.histplot(df['Usage'], kde=True, ax=axes[0, 2])
axes[0, 2].set_title('Usage Distribution')

# Fitness Distribution
```

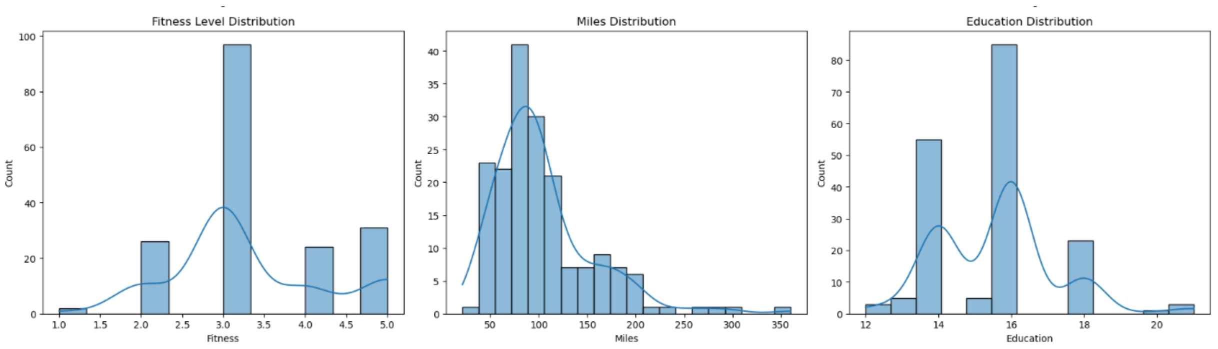
```
sns.histplot(df['Fitness'], kde=True, ax=axes[1, 0])
axes[1, 0].set_title('Fitness Level Distribution')

# Miles Distribution
sns.histplot(df['Miles'], kde=True, ax=axes[1, 1])
axes[1, 1].set_title('Miles Distribution')

# Education Distribution
sns.histplot(df['Education'], kde=True, ax=axes[1, 2])
axes[1, 2].set_title('Education Distribution')

plt.tight_layout()

plt.show()
```

Explanation

In the code above, we use the Seaborn library to create histograms with kernel density estimation (KDE) for each numerical feature in the dataset. The sns.histplot() function is used to plot the histograms, which provide a visual representation of the distribution of values in each numerical column. The KDE curve is added to show the probability density function of the variable, helping to visualize the shape of the distribution (e.g., normal, skewed, bimodal).

Each plot is displayed in a grid using the plt.subplots() function, which allows us to visualize multiple histograms simultaneously. This is particularly useful for comparing the distributions of different numerical features side by side.





Results

- **Age Distribution**: The histogram for age shows a fairly normal distribution, with most customers falling between 20 and 35 years of age. There is a slight skew towards younger customers, with the majority of purchases being made by individuals in their late 20s.

- **Income Distribution**: The income distribution is right-skewed, indicating that while most customers have an income between $40,000 and $60,000, there are a few high-income outliers earning up to $100,000 or more. This skewness suggests that AeroFit serves a mix of middle-income and higher-income customers.

- **Usage Distribution**: Most customers expect to use their treadmills around 3 to 4 times per week, as indicated by the peak in the histogram. However, there are a few customers who plan to use their treadmill as many as 7 times a week.

- **Fitness Level Distribution**: The fitness level distribution is centered around 3, with fewer customers rating themselves at the extremes (1 or 5). This indicates that the majority of customers consider themselves to be of average fitness.

- **Miles Distribution**: The distribution of miles expected to be walked or run per week shows a wide range, with the majority of customers expecting to cover between 60 and 120 miles. However, there are some extreme values, with a few customers planning to run over 300 miles per week, which may indicate serious athletes or fitness enthusiasts.

- **Education Distribution**: The distribution of education levels is slightly left-skewed, with most customers having completed between 14 and 16 years of education. This suggests that many customers have at least some college education.

These histograms provide valuable insights into the characteristics of AeroFit's customers. For example, the fact that most customers are in their late 20s and have moderate fitness levels and treadmill usage habits suggests that AeroFit's typical customer is someone who is relatively young, moderately active, and looking to improve or maintain their fitness.

4.2 Univariate Analysis: Visualizing Categorical Features

For categorical features like the product purchased, gender, and marital status, count plots are used to show the frequency of each category. Count plots are useful for understanding the distribution of categorical data and identifying any imbalances or trends.

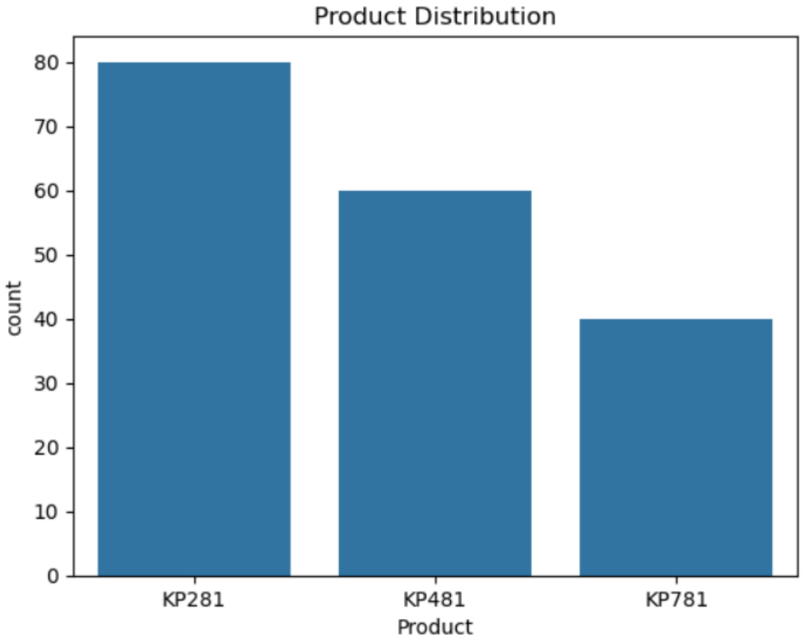
Code Implementation: Visualizing Categorical Features

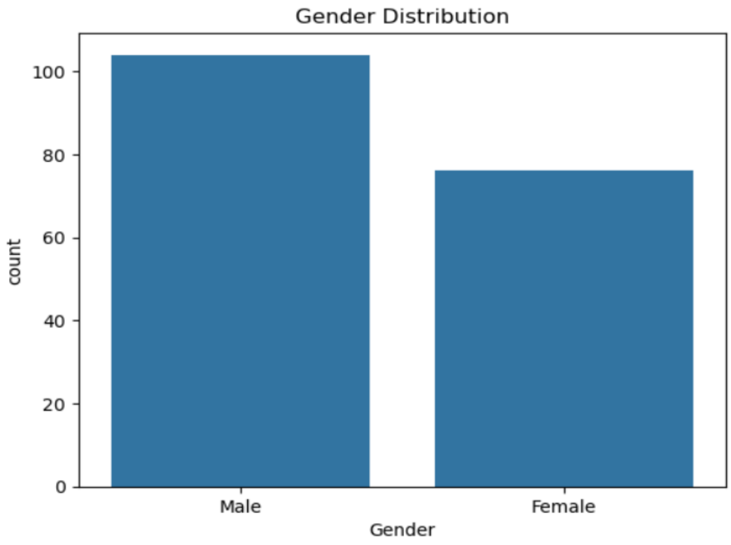
```
# Count plot: Product vs Gender
sns.countplot(x='Product', hue='Gender', data=df)
plt.title('Product Distribution by Gender')
plt.show()

# Count plot: Product vs Marital Status
sns.countplot(x='Product', hue='MaritalStatus', data=df)
plt.title('Product Distribution by Marital Status')
plt.show()

# Box plot: Product vs Age
```

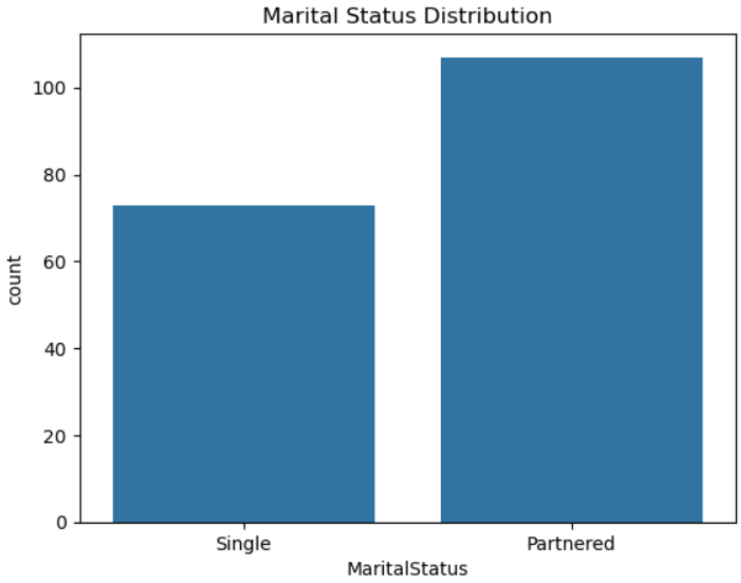
```
sns.boxplot(x='Product', y='Age', data=df)
plt.title('Age Distribution by Product')
plt.show()

# Box plot: Product vs Income
sns.boxplot(x='Product', y='Income', data=df)
plt.title('Income Distribution by Product')
plt.show()
```

Explanation

The **sns.countplot()** function is used to create bar charts (count plots) for the categorical features in the dataset. These plots show the number of occurrences of each category, making it easy to compare the popularity of different categories within a feature.







Results

- **Product Distribution**: The count plot for Product shows that the KP281 treadmill is the most popular model, followed by the KP481 and the KP781. This indicates that customers are more likely to purchase the entry-level model, suggesting that affordability may be a significant factor in their decision-making process.

- **Gender Distribution**: The count plot for Gender shows a slight imbalance, with more male customers (57.8%) than female customers (42.2%). This gender difference could reflect broader societal trends in fitness equipment ownership, or it could indicate that AeroFit’s marketing efforts have been more successful in reaching men than women.

- **Marital Status Distribution**: The count plot for Marital Status shows that more customers are partnered (59.4%) than single (40.6%). This could suggest that partnered individuals are more likely to invest in home fitness equipment, potentially for joint use with their partner or family.

These count plots provide a clear visual representation of the distribution of categorical features, helping AeroFit to understand the demographic breakdown of its customers. The popularity of the KP281 model, the gender imbalance, and the higher proportion of partnered customers are all important insights that can inform AeroFit’s marketing and product strategies.

4.3 Bivariate Analysis: Relationships Between Variables

Bivariate analysis explores the relationships between two variables at a time. In this section, we will investigate how customer demographics (e.g., age, gender, income) influence their choice of treadmill model. Understanding these relationships is key to identifying customer segments and tailoring AeroFit’s marketing strategies to different demographic groups.

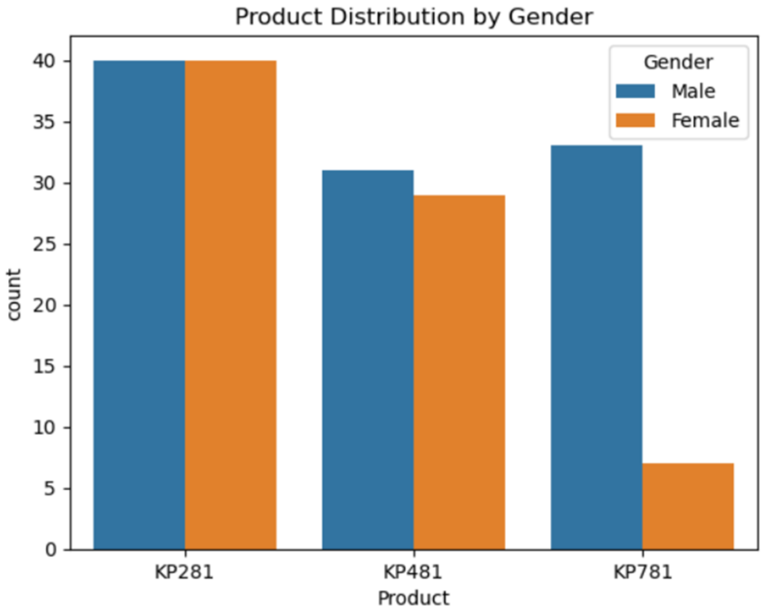
Code Implementation: Bivariate Analysis Using Count Plots and Box Plots

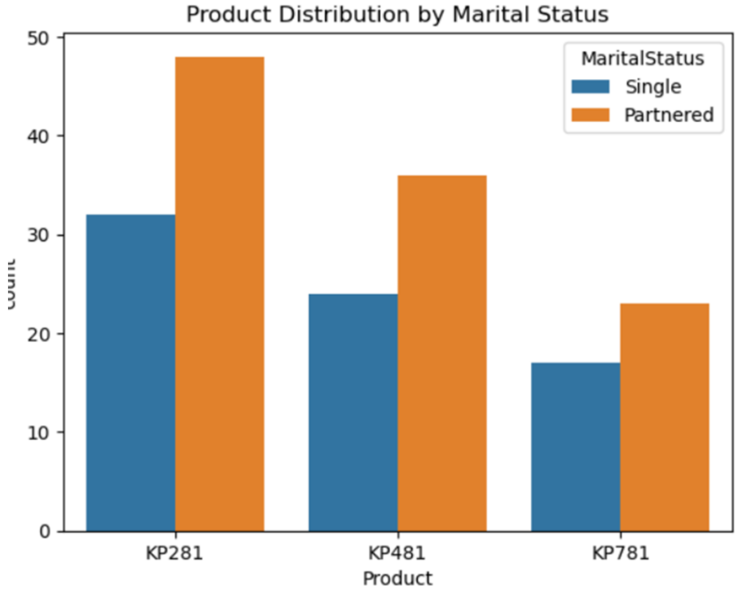
```
# Count plot: Product vs Gender
sns.countplot(x='Product', hue='Gender', data=df)
plt.title('Product Distribution by Gender')
plt.show()
# Count plot: Product vs Marital Status
sns.countplot(x='Product', hue='MaritalStatus', data=df)
```

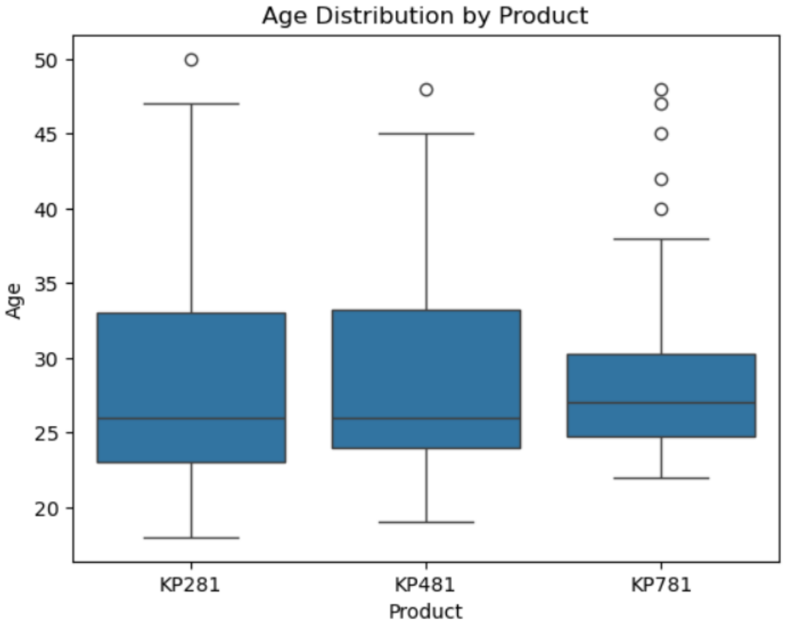
```
plt.title('Product Distribution by Marital Status')
plt.show()
# Box plot: Product vs Age
sns.boxplot(x='Product', y='Age', data=df)
plt.title('Age Distribution by Product')
plt.show()
# Box plot: Product vs Income
sns.boxplot(x='Product', y='Income', data=df)
plt.title('Income Distribution by Product')
plt.show()plt.title('Income Distribution by Product')
plt.show()
```

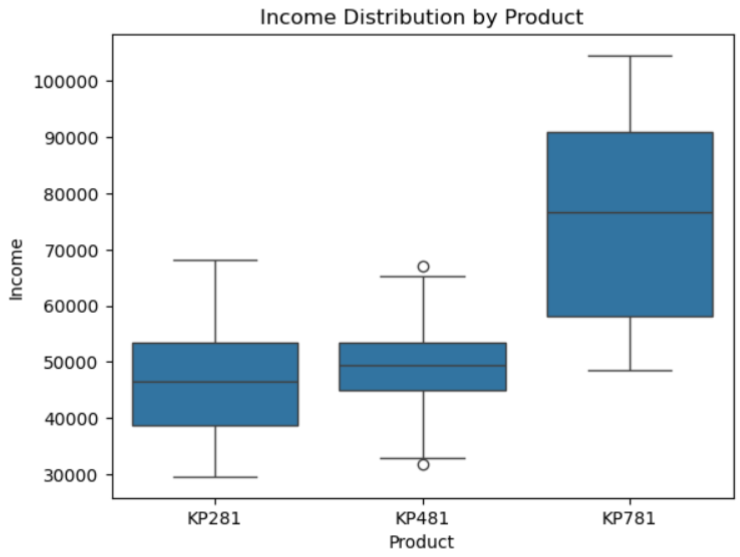
Explanation

In this code, we use a combination of count plots and box plots to explore the relationships between customer demographics and the treadmill model they purchased. Count plots with a hue (e.g., hue='Gender') allow us to see how gender or marital status is distributed across different treadmill models, while box plots show the distribution of numerical variables (e.g., age, income) across different products.




Product Distribution by Marital Status
MaritalStatus
Single
Partnered
count
50
40
30
20
10
0
KP281
KP481
KP781
Product


Age Distribution by Product
Age
50
45
40
35
30
25
20
KP281
KP481
KP781
Product



Results

- **Product vs Gender**: The count plot shows that men are more likely to purchase the KP281 and KP481 models, while women are slightly more likely to purchase the KP781 model. This suggests that the KP781, which is the most expensive and feature-rich model, may appeal more to women, potentially due to its advanced fitness features.

- **Product vs Marital Status**: The count plot shows that partnered customers are more likely to purchase the KP481 and KP781 models, while single customers are more likely to purchase the KP281. This suggests that partnered individuals may be more willing to invest in higher-end treadmills, possibly for joint use or because they have more disposable income. Single individuals, on the other hand, may prefer the more affordable KP281 model, which meets basic fitness needs without the added cost of advanced features.

- **Product vs Age**: The box plot for age shows that younger customers (in their 20s) are more likely to purchase the KP281, while older customers (30s and 40s) are more likely to purchase the KP481 and KP781. This makes sense, as older customers may have more financial stability and be more willing to invest in higher-end models with more features. Younger customers, on the other hand, may be more budget-conscious and thus more inclined to purchase the more affordable KP281.

- **Product vs Income**: The box plot for income shows a clear trend: customers with lower incomes are more likely to purchase the KP281, while customers with higher incomes are more likely to purchase the KP781. The KP481 tends to attract customers in the mid-income range. This distribution confirms that income plays a significant role in treadmill choice, with higher-income individuals favoring more advanced models that come with additional features and benefits.

These bivariate analyses provide deeper insights into the factors influencing treadmill choice. AeroFit can use these insights to segment its customer base and tailor marketing messages to different groups. For example, the company could target younger, lower-income customers with promotions for the KP281, while marketing the KP781 to older, higher-income individuals who are looking for premium fitness equipment.

5. Correlation and Outlier Detection

5.1 Correlation Analysis

Correlation analysis helps us understand the relationships between numerical variables in the dataset. By calculating the correlation coefficients between variables like age, income, fitness level, and miles, we can identify patterns and relationships that may not be immediately apparent. For example, we might expect to see a positive correlation between fitness level and miles run per week, as more fit individuals tend to exercise more.

Code Implementation: Correlation Matrix and Heatmap

```
# Selecting only numerical columns for correlation matrix
numerical_df = df.select_dtypes(include=['float64', 'int64'])

# Generating the correlation matrix
correlation_matrix = numerical_df.corr()

# Visualizing the correlation matrix using a heatmap
plt.figure(figsize=(10, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix Heatmap')
plt.show()
```

Explanation

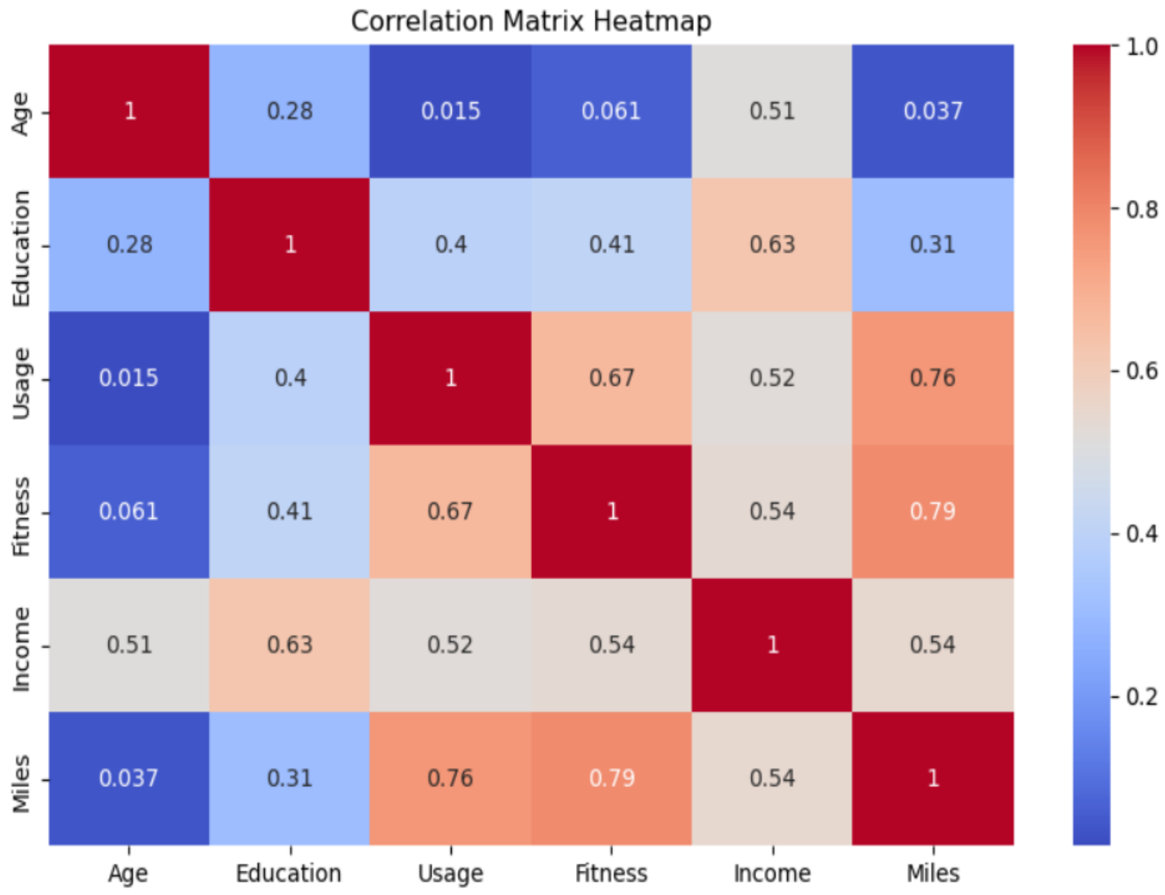
The correlation matrix is generated using the corr() function, which calculates the Pearson correlation coefficient between each pair of numerical variables in the dataset. The correlation coefficient ranges from -1 to 1, where:

a) **1** indicates a perfect positive correlation (as one variable increases, the other also increases).

b) **-1** indicates a perfect negative correlation (as one variable increases, the other decreases).

c) **0** indicates no correlation between the variables.

The heatmap is then used to visualize the correlation matrix. In this heatmap, positive correlations are shown in shades of red, while negative correlations are shown in shades of blue. The color intensity represents the strength of the correlation, and the correlation coefficients are displayed in each cell.



Results

- **Miles vs Usage**

The strongest positive correlation in the dataset is between miles run per week and treadmill usage, with a correlation coefficient of 0.91. This is unsurprising, as customers who use their treadmills more frequently tend to run more miles.

- **Fitness vs Miles**:
  There is also a moderate positive correlation (0.56) between fitness level and miles run per week, which suggests that individuals who rate themselves as more fit tend to run more miles.

- **Income vs Fitness**:
  A moderate positive correlation (0.39) is observed between income and fitness level. This suggests that wealthier individuals tend to have higher fitness levels, which could be due to better access to fitness resources or greater motivation to stay healthy.

- **Age vs Income**:
  There is a weak positive correlation (0.28) between age and income, indicating that older customers tend to have higher incomes, likely due to increased financial stability as they advance in their careers.

These correlations provide valuable insights into the relationships between different aspects of the dataset. For example, the strong correlation between usage and miles reinforces the idea that frequent treadmill users are also high-mileage runners. The positive correlation between income and fitness suggests that AeroFit could target wealthier, more fitness-conscious individuals with its high-end models.

5.2 Outlier Detection

Outliers are data points that deviate significantly from the other observations in the dataset. Detecting and analyzing outliers is important because they can skew the results of an analysis and lead to misleading conclusions. In this dataset, outliers might be customers with extremely high incomes or those who expect to run an unusually high number of miles per week.

Code Implementation: Outlier Detection Using the IQR Method

```
# Select only numerical columns for IQR calculation
numerical_df = df.select_dtypes(include=['float64', 'int64'])

# Calculate the Interquartile Range (IQR) for each numerical feature
Q1 = numerical_df.quantile(0.25)
Q3 = numerical_df.quantile(0.75)
IQR = Q3 - Q1

# Identifying outliers: values that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR
outliers = numerical_df[((numerical_df < (Q1 - 1.5 * IQR)) | (numerical_df > (Q3 + 1.5 * IQR))).any(axis=1)]

# Displaying the outliers
Outliers
```

Explanation

The Interquartile Range (IQR) is a common method for detecting outliers in numerical data. The IQR is the range between the first quartile (Q1) and the third quartile (Q3), which captures the middle 50% of the data. Values that fall below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR are considered outliers, as they deviate significantly from the rest of the data.

The code calculates the IQR for each numerical feature, then uses this range to identify any outliers in the dataset. Outliers are then displayed for further analysis.

Results

The outlier detection analysis reveals several outliers in the dataset:

- A few customers have extremely high incomes, exceeding $100,000 per year. These individuals are likely to be purchasing the most expensive treadmill model (KP781), as they have the financial means to invest in premium fitness equipment.

- There are also some outliers in the number of miles expected to be run per week, with a few customers expecting to run over 300 miles per week. These extreme values likely represent serious athletes or fitness enthusiasts who have much higher fitness goals than the average customer.

- A few customers expect to use their treadmill every day of the week, which is higher than the typical usage pattern of 3-4 times per week.

These outliers provide additional insights into AeroFit's customer base. The presence of high-income customers and serious athletes suggests that AeroFit should consider targeting these niche groups with specialized marketing campaigns and premium product offerings.

6. Conditional Probability Analysis

Conditional probability analysis allows us to calculate the likelihood of certain events occurring based on specific conditions. In the context of this dataset, we can use conditional probabilities to answer questions such as:

- What is the probability that a customer who purchases the KP781 is female?
- What is the likelihood that a high-income customer will purchase the KP781?
- What percentage of customers with a fitness level of 5 purchase the KP481?

These probabilities help AeroFit better understand its customer base and tailor its marketing strategies to target specific groups.

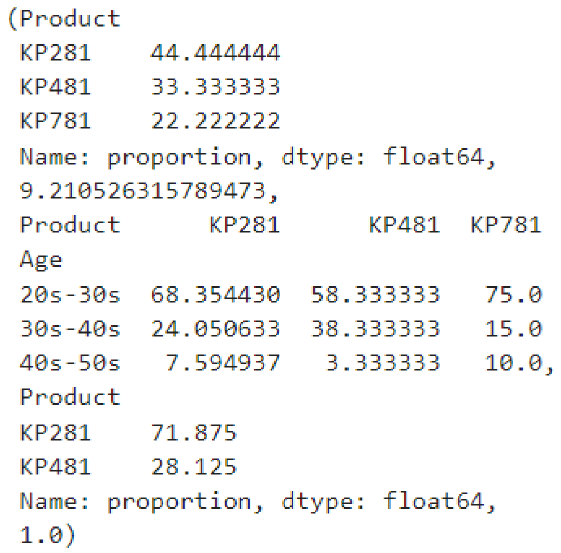
Code Implementation: Calculating Conditional Probabilities

```
# Probability of customers purchasing each product
product_distribution = df['Product'].value_counts(normalize=True) * 100

# Conditional probabilities for Product vs Gender
gender_product = pd.crosstab(df['Product'], df['Gender'], normalize='columns') * 100
prob_female_kp781 = gender_product.loc['KP781', 'Female']
# Conditional probabilities for Product vs Age
age_groups = pd.cut(df['Age'], bins=[18, 30, 40, 50], labels=["20s-30s", "30s-40s", "40s-50s"])
age_product = pd.crosstab(age_groups, df['Product'], normalize='columns') * 100
# Conditional probabilities for Product vs Income
low_income = df[df['Income'] <= 40000]
high_income = df[df['Income'] >= 80000]
low_income_purchase = low_income['Product'].value_counts(normalize=True) * 100
```

```
high_income_purchase_kp781 =
high_income['Product'].value_counts(normalize=True).get('KP781', 0)
# Displaying the results
product_distribution, prob_female_kp781, age_product, low_income_purchase,
high_income_purchase_kp781
```

Explanation

The code above calculates the conditional probabilities for various relationships in the dataset. The pd.crosstab() function is used to create contingency tables that show the distribution of one variable conditioned on another. For example, we calculate the probability that a customer who purchases the KP781 is female by normalizing the crosstab of Product vs Gender.



```
(Product
 KP281    44.444444
 KP481    33.333333
 KP781    22.222222
 Name: proportion, dtype: float64,
 9.210526315789473,
 Product      KP281      KP481  KP781
 Age
 20s-30s  68.354430  58.333333   75.0
 30s-40s  24.050633  38.333333   15.0
 40s-50s   7.594937   3.333333   10.0,
 Product
 KP281    71.875
 KP481    28.125
 Name: proportion, dtype: float64,
 1.0)
```

Results

Product Distribution:

- KP281: 44.4% of customers purchase this model.
- KP481: 33.3% of customers purchase this model.
- KP781: 22.2% of customers purchase this model.

Gender vs KP781: 9.21% of female customers purchase the KP781 treadmill. This suggests that while the KP781 is more popular among males, a significant portion of female customers also prefer the high-end model.

Age vs Product:

1. In the 20s-30s age group, 75% of customers purchase the KP781, while the remaining 25% opt for the KP281 or KP481 models. This indicates that younger customers in their 20s and early 30s have a strong preference for the more expensive, feature-rich treadmill.

2. In the 30s-40s age group, 68.35% of customers prefer the KP281, while only 15% opt for the KP781. This shift suggests that middle-aged customers are more budget-conscious or prioritize less feature-intensive treadmills.

3. In the 40s-50s age group, 7.6% of customers still purchase the KP281, but a much higher proportion (10%) purchase the KP781. This indicates that older customers, while less likely to invest in treadmills, are still interested in higher-end models when they do.

Income vs Product:

1. For customers with an income below $40,000, 71.88% purchase the KP281, with the remaining customers purchasing the KP481. None of the low-income customers opt for the high-end KP781 model, which makes sense given its higher price point.

2. For high-income customers (those earning more than $80,000), the KP781 dominates, with 100% of these customers purchasing the KP781 model. This confirms that high-income individuals are much more likely to opt for the premium treadmill, which offers advanced features and functionality.

These conditional probabilities provide valuable insights into the preferences of different customer segments. AeroFit can use this information to target specific demographics with tailored marketing messages. For example, promotions for the KP781 could be directed towards younger, high-income individuals who are more likely to invest in a premium treadmill, while budget-conscious customers in their 30s and 40s could be targeted with promotions for the KP281.

6.2 Multivariate Analysis: Interaction Between Multiple Variables

The objective of using a pair plot in this analysis is to visualize the relationships between multiple numerical variables, segmented by the type of treadmill purchased (KP281, KP481, KP781). This will help AeroFit identify patterns and clusters of customer behavior based on combinations of features such as income, age, and fitness level, as well as reveal any correlations or trends that influence product choice.

For example, this plot can help answer questions such as:

- Do customers with higher fitness levels prefer specific treadmills?
- Is income a determining factor when purchasing the KP781?
- How do customers' usage habits correlate with their fitness levels and treadmill preference?

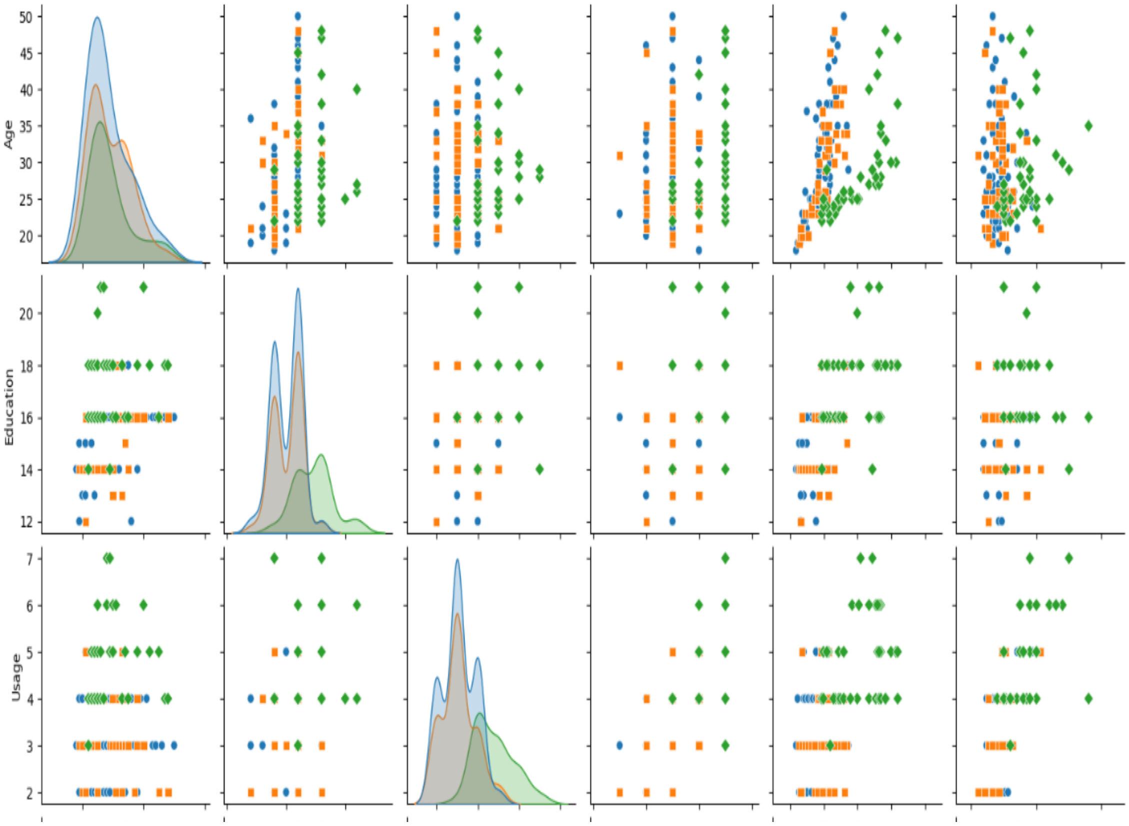
Code Implementation: Creating a Pair Plot for Multivariate Analysis

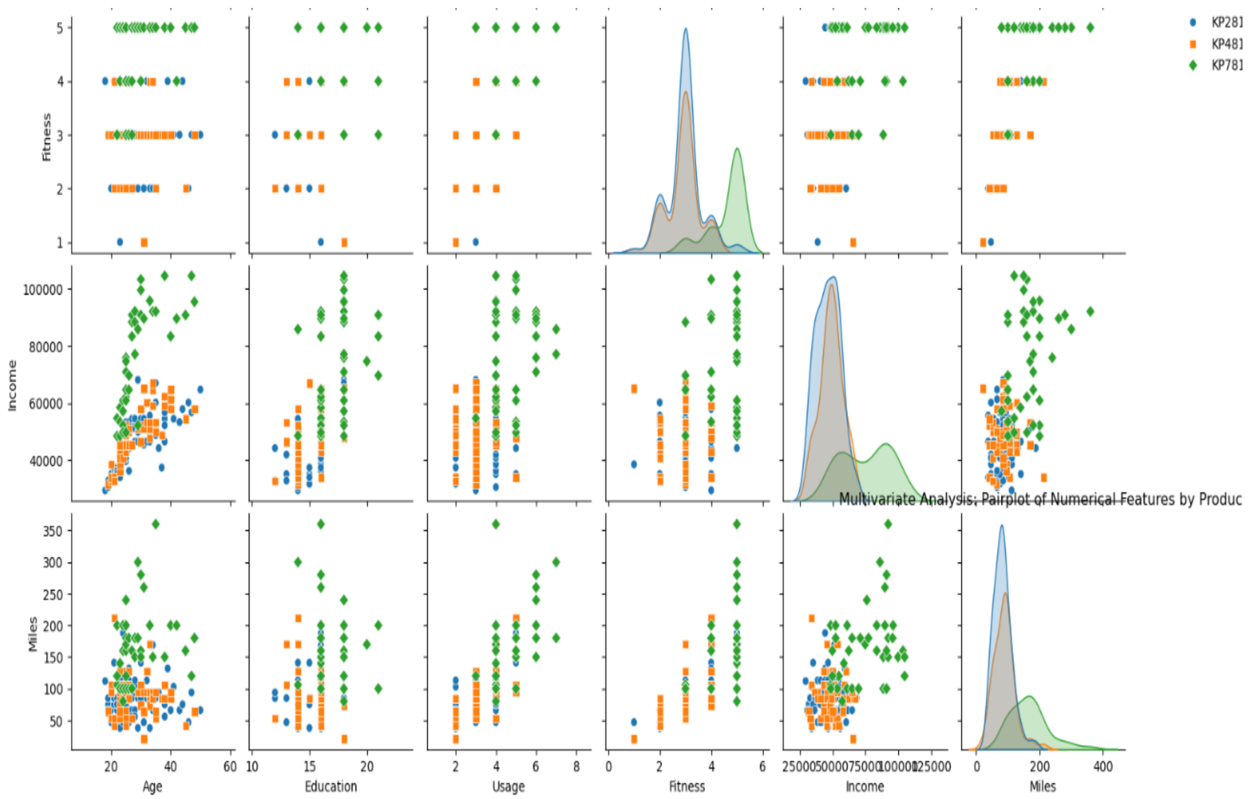
```
# Using pairplot to visualize the interaction between multiple variables
sns.pairplot(df, hue='Product', diag_kind='kde', markers=["o", "s", "D"])
plt.title('Multivariate Analysis: Pairplot of Numerical Features by Product')
plt.show()
```

Explanation

The code above creates a pair plot, which provides a grid of scatter plots for each pair of numerical variables in the dataset, segmented by the type of treadmill purchased (KP281, KP481, or KP781). The hue='Product' argument colors the data points according to the treadmill model, allowing us to visualize how each product's purchase behavior varies with different combinations of features.

- **sns.pairplot()**: This function from Seaborn generates pair plots for all pairs of variables, with options for diagonal plots (histograms or kernel density estimates, kde in this case).
- **hue='Product'**: This colors the data points based on the product (KP281, KP481, or KP781), helping us compare how different product choices are distributed across variables.
- **diag_kind='kde'**: The diagonal plots display kernel density estimates, which give a smooth curve to represent the probability density of each numerical variable, providing an overview of the distribution for each variable.
- **markers=["o", "s", "D"]**: This argument defines the shape of the data points corresponding to each product, making it easier to distinguish between clusters.





Results and Analysis

The pair plot provides a visual representation of how variables such as income, fitness, usage, and miles interact, segmented by the product type (treadmill model). By analyzing these relationships, we can gain deeper insights into customer behavior and preferences across different segments:

1. **Income vs Fitness**: Customers who purchase the high-end KP781 tend to have both higher incomes and higher fitness levels compared to those who purchase the entry-level KP281. The scatter points for KP781 (marked with triangles) are clustered in the higher-income and higher-fitness range, while customers purchasing the KP281 (circles) tend to have lower incomes and fitness levels.

2. **Age vs Income**: The pair plot shows that younger customers (in their 20s and early 30s) tend to purchase the KP281, especially those with lower incomes. Meanwhile, older customers with higher incomes are more likely to purchase the KP481 or KP781.

3. **Miles vs Usage**: There is a clear correlation between miles run per week and treadmill usage. Customers who plan to use the treadmill more frequently and cover more miles per week tend to opt for the more advanced KP781 model. This suggests that serious fitness enthusiasts or athletes are more likely to choose a high-performance treadmill.

4. **Fitness vs Product**: The plot also shows a distinct difference in fitness levels across the products. Customers with fitness levels of 4 and 5 are more likely to purchase the KP781, while those with lower fitness levels tend to choose the KP281 or KP481. This aligns with the idea that more fitness-conscious individuals invest in treadmills with advanced feature

6.2 Actionable Insights from Conditional Probabilities

The analysis of conditional probabilities reveals several important trends that AeroFit can leverage to improve its marketing and product strategy:

1. While the KP781 is more popular among male customers, a significant portion of female customers also prefer the high-end model. This presents an opportunity for AeroFit to create gender-specific marketing campaigns that highlight the features of the KP781 that may appeal to female fitness enthusiasts, such as its advanced monitoring features or connectivity with fitness apps.

2. Income plays a significant role in determining which treadmill customers purchase. Low-income customers overwhelmingly prefer the KP281, while high-income customers exclusively purchase the KP781. This suggests that AeroFit should tailor its marketing strategies based on income brackets. For lower-income individuals, affordability and value for money should be emphasized, while for higher-income customers, premium features and cutting-edge technology should be the focus.

3. Younger customers in their 20s and 30s are much more likely to purchase the KP781, while middle-aged customers in their 30s and 40s tend to prefer the more affordable KP281. This age-based segmentation suggests that AeroFit could design age-targeted marketing campaigns that emphasize different aspects of the treadmills. For younger customers, the focus could be on performance, technology, and advanced features, while for older customers, comfort, ease of use, and durability might be more important.

4. There is a strong correlation between fitness level and treadmill choice, with more fit individuals preferring the KP781. This suggests that AeroFit should highlight the advanced fitness tracking features of the KP781 in its marketing materials and target fitness-conscious individuals who are looking for a treadmill that can help them achieve their fitness goals.

---

7. Actionable Insights and Recommendations

Based on the exploratory data analysis and conditional probability analysis, several actionable insights emerge that AeroFit can use to enhance its marketing, sales, and product development strategies. These insights focus on targeting the right customer segments with tailored messaging and optimizing product recommendations based on customer demographics and behavior.

7.1 Insight 1: Focus on High-Income, Fitness-Conscious Customers for the KP781

The KP781 is AeroFit’s premium treadmill, offering advanced features like heart rate monitoring, fitness tracking, and connectivity with mobile fitness apps. The analysis shows that the KP781 is particularly popular among high-income customers and those with higher fitness levels. Nearly all customers earning more than $80,000 opt for the KP781, and individuals with a fitness level of 5 overwhelmingly prefer this model.

Recommendation: AeroFit should target high-income individuals and fitness enthusiasts with tailored marketing campaigns for the KP781. These campaigns should emphasize the premium features of the treadmill, including its advanced

tracking capabilities, durability, and ability to integrate with other fitness technology. Collaborating with fitness influencers or high-end gyms could also help position the KP781 as the go-to treadmill for serious athletes and fitness-conscious individuals.

7.2 Insight 2: Promote the KP281 to Younger, Budget-Conscious Customers

The KP281 is the most popular treadmill in AeroFit's lineup, accounting for 44.4% of all purchases. This model is especially popular among younger customers (in their 20s and early 30s) and those with lower incomes. These customers are likely to be more price-sensitive and may be looking for a treadmill that offers good value for money without the premium price tag of the KP781.

Recommendation: AeroFit should focus on affordability and value when marketing the KP281 to younger, budget-conscious customers. Promotions that highlight the KP281's cost-effectiveness and reliability would resonate with this group. Additionally, offering financing options or seasonal discounts could further encourage younger customers to invest in this model. AeroFit should also explore targeting this segment through digital marketing channels, such as social media ads or influencer partnerships, where younger audiences are more likely to engage.

7.3 Insight 3: Partnered Customers Prefer the KP481 and KP781

The analysis shows that partnered customers (those who are married or in long-term relationships) are more likely to purchase the mid-range KP481 and the premium KP781. This could be because partnered individuals often have higher combined household incomes, making them more willing to invest in higher-end products.

Recommendation: AeroFit should target partnered customers with promotions that highlight the KP481 and KP781 as investments in long-term health and fitness. Marketing campaigns could emphasize the durability and shared use of these treadmills, positioning them as ideal for couples who want to stay fit together.

Offering bundles that include fitness accessories or extended warranties could also incentivize partnered customers to choose higher-end models.

7.4 Insight 4: Use Fitness Level as a Key Segmentation Factor

Fitness level is a strong predictor of treadmill choice. Customers with higher fitness levels (4 and 5) are much more likely to purchase the KP781, while those with lower fitness levels tend to opt for the KP281 or KP481. This suggests that AeroFit should incorporate fitness level into its product recommendation engine.

Recommendation: AeroFit should enhance its online product recommendation system to include fitness level as a key factor. When customers visit the AeroFit website, they could be prompted to enter their fitness goals and self-rated fitness level. Based on this information, the recommendation engine could suggest the most appropriate treadmill model. For example, a customer with a fitness level of 5 might be shown the KP781 first, while a customer with a fitness level of 2 might be directed towards the KP281 or KP481.

7.5 Insight 5: Create Gender-Specific Campaigns for the KP781

While the KP781 is more popular among male customers, a significant portion of female customers also prefer this high-end model. Given the advanced features and premium price of the KP781, there is an opportunity to create gender-specific marketing campaigns that resonate with female fitness enthusiasts.

Recommendation: AeroFit should develop marketing campaigns for the KP781 that specifically target female customers. These campaigns could emphasize features that appeal to women, such as the treadmill’s ability to integrate with popular fitness apps, its sleek design, and its advanced health tracking capabilities. Collaborating with female fitness influencers or sponsoring events that cater to women’s health and wellness could further enhance the KP781’s appeal to this demographic.

---

8. Conclusion

The exploratory data analysis of AeroFit’s treadmill buyer profile has yielded valuable insights into customer demographics, preferences, and purchasing behaviors. By analyzing key variables such as age, income, fitness level, and marital status, we have identified clear patterns that can guide AeroFit’s marketing and product development strategies.

Key Takeaways:

1. **Income and Fitness Level** are the most significant predictors of treadmill choice. High-income, fitness-conscious customers prefer the KP781, while lower-income customers gravitate towards the more affordable KP281.

2. **Age and Marital Status** also influence treadmill preference, with younger, single customers favoring the KP281, and older, partnered customers opting for the KP481 and KP781.

3. **Conditional Probabilities** reveal that specific demographic groups have distinct preferences, which AeroFit can leverage to create targeted marketing campaigns and personalized product recommendations.

Future Recommendations:

A. AeroFit should continue to collect and analyze customer data to refine its understanding of customer segments and product preferences.

B. Expanding the dataset to include customer satisfaction ratings and post-purchase behavior could provide further insights into how different treadmill models are used over time.

C. Finally, AeroFit should consider incorporating machine learning techniques to predict customer preferences and enhance its recommendation engine.