Exploratory Data Analysis (EDA)
Report
AeroFit treadmills

This project analyzes customer purchasing trends for AeroFit treadmills using Exploratory Data Analysis (EDA). The objective is to identify the characteristics of the target audience for each treadmill model and provide recommendations to improve product marketing and sales. The analysis explores key demographic attributes such as age, income, fitness level, and marital status. Findings reveal correlations between product preference and various customer traits, highlighting insights that can enhance sales strategies.

Prepared by
Muhammad TAHA Zaman

Prepared for
AeroFit treadmills

Table of Contents:

Abstract:

This project analyzes customer purchasing trends for AeroFit treadmills using **Exploratory Data Analysis (EDA)**. The objective is to identify the characteristics of the target audience for each treadmill model and provide recommendations to improve product marketing and sales. The analysis explores key demographic attributes such as age, income, fitness level, and marital status. Findings reveal correlations between product preferences and various customer traits, offering valuable insights to enhance sales strategies. The goal is to enable targeted recommendations and optimize marketing efforts based on consumer behavior and demographics.

Introduction:

This report investigates customer demographics and purchasing behavior at **AeroFit**. The company offers three treadmill models:

- **KP281**: Entry-level treadmill ($1,500)
- **KP481**: Mid-level treadmill ($1,750)
- **KP781**: High-end treadmill ($2,500)

The primary goal is to analyze **customer attributes and trends** to determine how factors such as **age, income, fitness level, and marital status** influence treadmill selection. This EDA-based study will provide actionable insights for marketing and sales improvements.

The analysis is structured into the following sections:

1. **Data Exploration & Preprocessing**
2. **Statistical & Non-Graphical Analysis**
3. **Graphical Analysis (Univariate, Bivariate, and Multivariate)**
4. **Correlation & Outlier Detection**
5. **Conditional Probability Analysis**
6. **Actionable Insights & Recommendations**

Methodology:

Tools Used:

- **Programming Language**: Python
- **Libraries**: Pandas, NumPy, Seaborn, Matplotlib, SciPy
- **Dataset**: AeroFit Treadmill Sales Data

Steps Involved:

1. **Data Exploration & Preprocessing:**
   - Loaded the dataset and checked for missing values & duplicates.
   - Examined the data structure, column types, and basic statistics.

2. **Statistical Analysis:**

   - Analyzed numerical and categorical variables.
   - Computed summary statistics for age, income, fitness level, and treadmill usage.
   - Age range: 18 - 50 years
   - Education: 12 - 21 years
   - Fitness (1-5 scale): Majority rated 3 or 4
   - Income: $29,562 to $104,581

3. **Non Graphical Analysis:**

   - Gender: 104 Male, 76 Female
   - Marital Status: 107 Partnered, 73 Single
   - Most customers plan to use treadmill 3-4 times/week
   - Analyzed numerical and categorical variables
   - Computed summary statistics for age, income, fitness level, and treadmill usage

4. **Graphical Analysis:**
   - **Univariate Analysis**: Histograms, boxplots, and count plots for numerical and categorical data.
   - **Bivariate Analysis**: Analyzed relationships between product purchase and customer attributes (Gender, Age, Marital Status, and Income).
   - **Multivariate Analysis**: Created pair plots to visualize feature relationships.

5. **Correlation & Outlier Detection:**
   - Generated a correlation heatmap to understand feature relationships.
   - Identified outliers using the IQR method
   - Outliers found in Age (5), Education (4), Usage (9), Fitness (2), Income (19), and Miles (13)

6. **Conditional Probability Analysis:**
   - o Calculated probability distributions for product selection based on gender, income, age, fitness level, and marital status.

   Product Purchase Rates:
   - o KP281: 80/180
   - o KP481: 60/180
   - o KP781: 40/180
   - o Gender: 9.21% of females purchased KP781
   - o Age: 87.78% of customers are aged 20s to 30s
   - o Income: 100% of high-income (>$70k) customers purchased KP781
   - o Fitness: 72% of KP781 buyers have Fitness Level 5
   - o Marital Status: 67% of treadmill users are partnered

7. **Insights & Recommendations:**
   - o Summarized findings and suggested marketing strategies based on data trends.

Graphical Analysis: Univariate Analysis - Numerical features:

Distribution Plot:

These plots were used to visualize the frequency distribution of numerical features such as Age, Income, and Miles. They helped in understanding the shape of the data (normal, skewed) and identifying any visible clusters or gaps.

Distribution of Income

Distribution of Miles
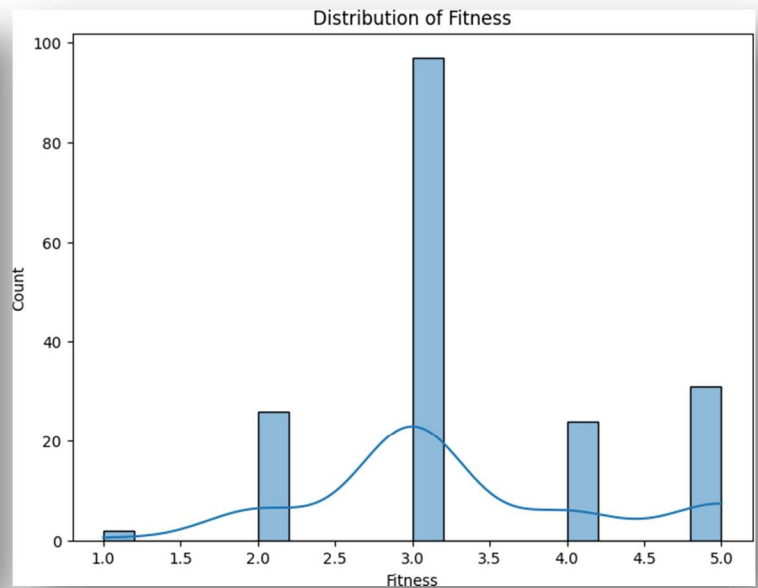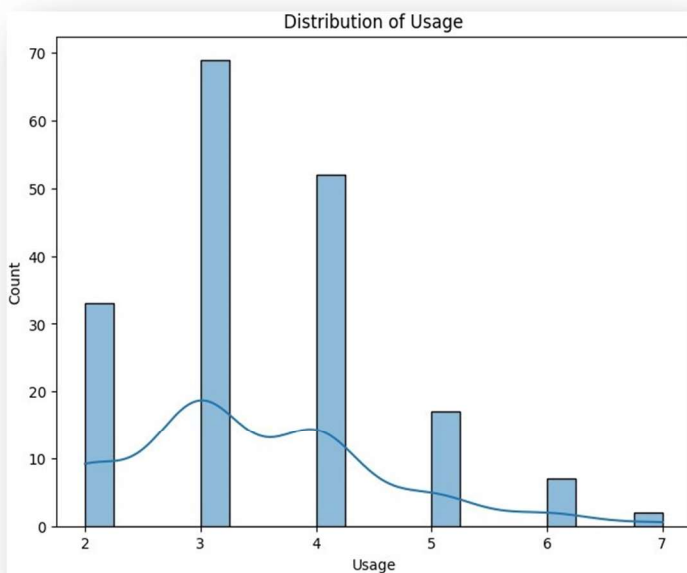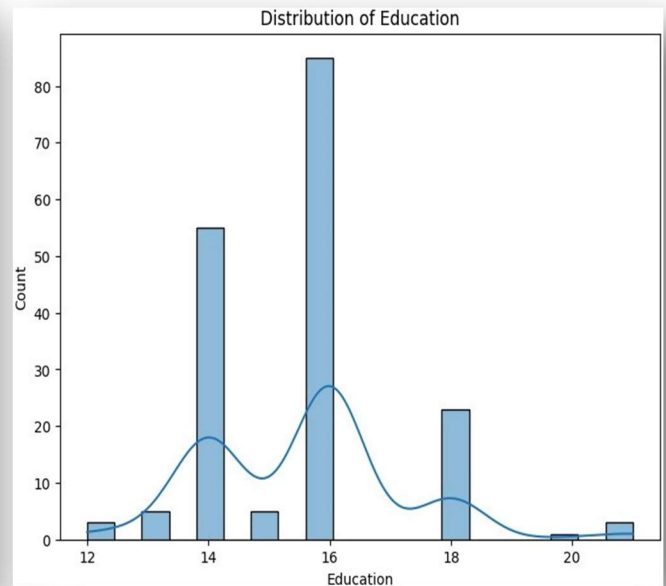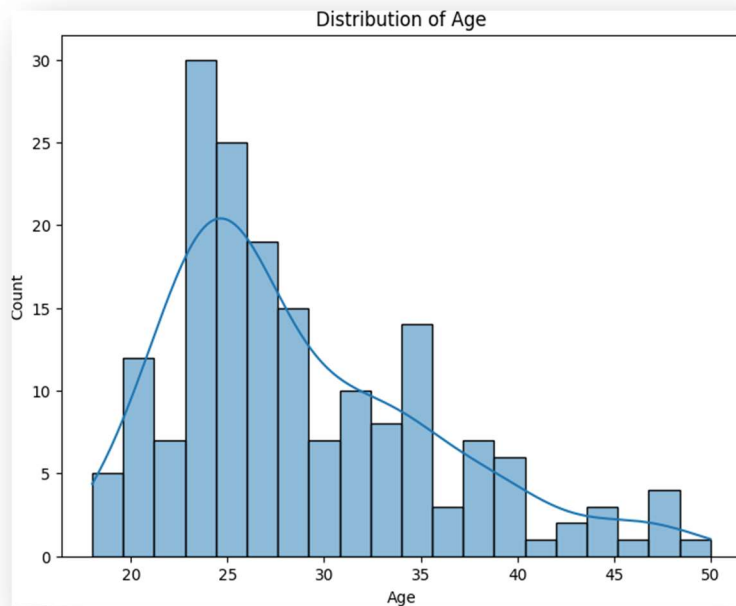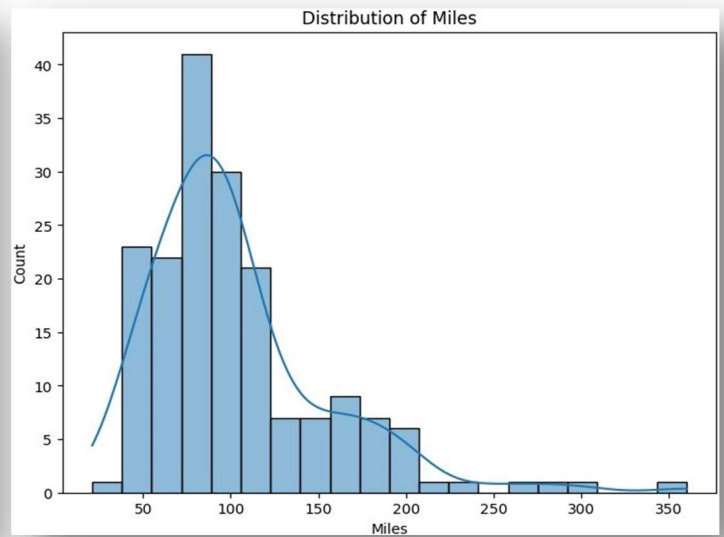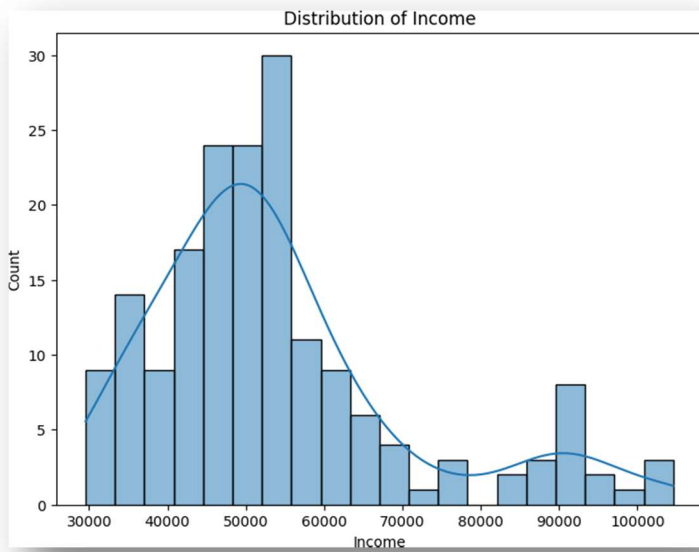
Count Plot:

Count plots were also used for numerical columns that take on discrete values. These plots helped visualize how often each numerical value occurred, making it easier to detect common ratings or usage patterns among customers.
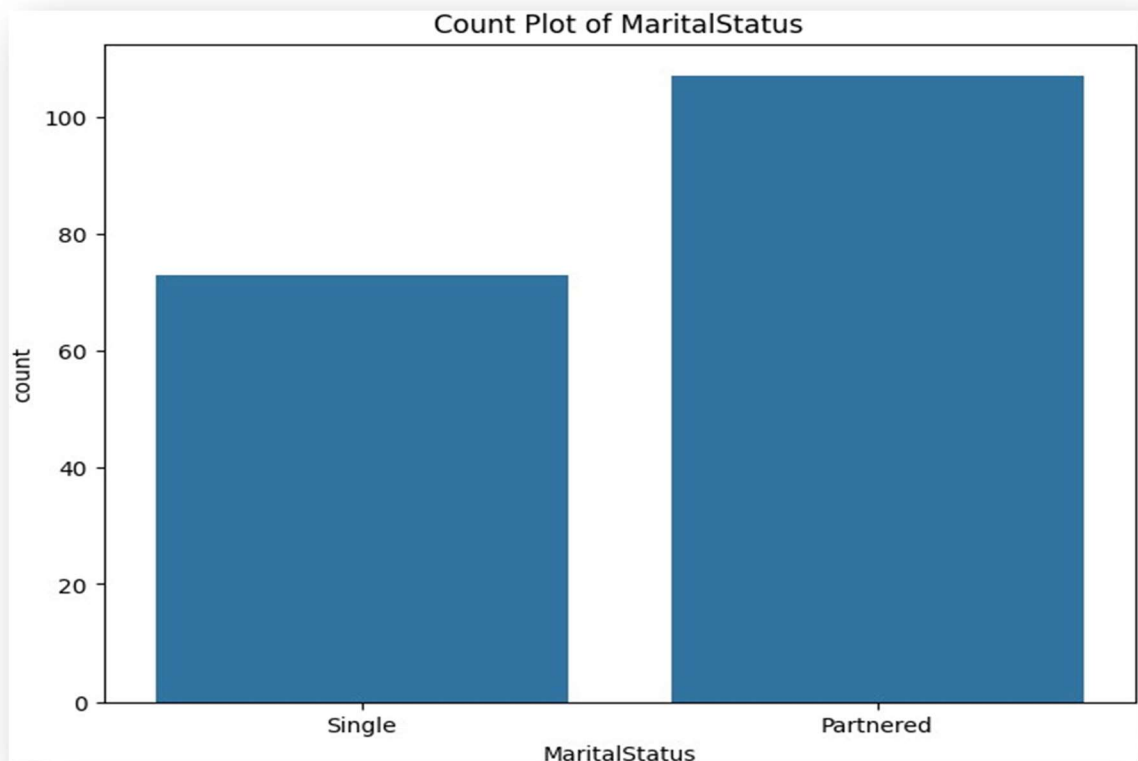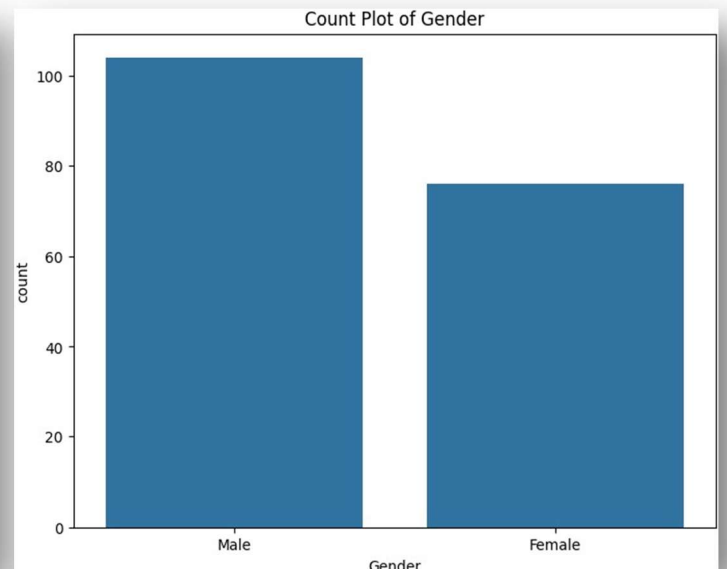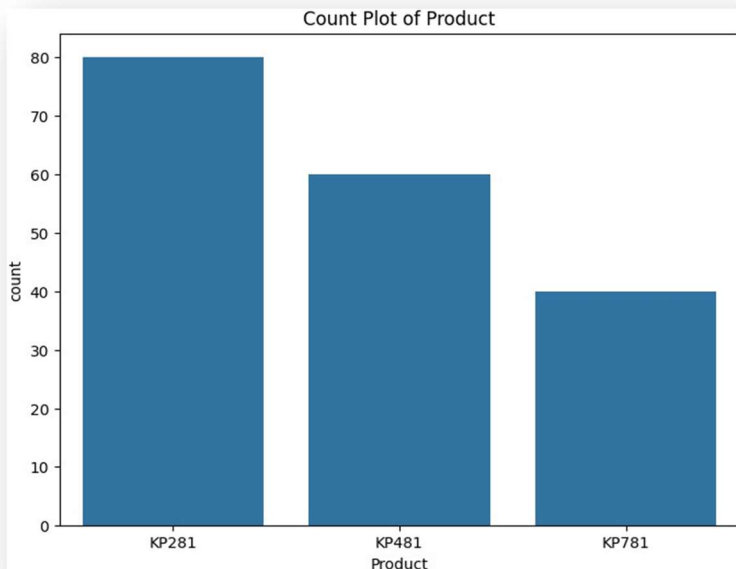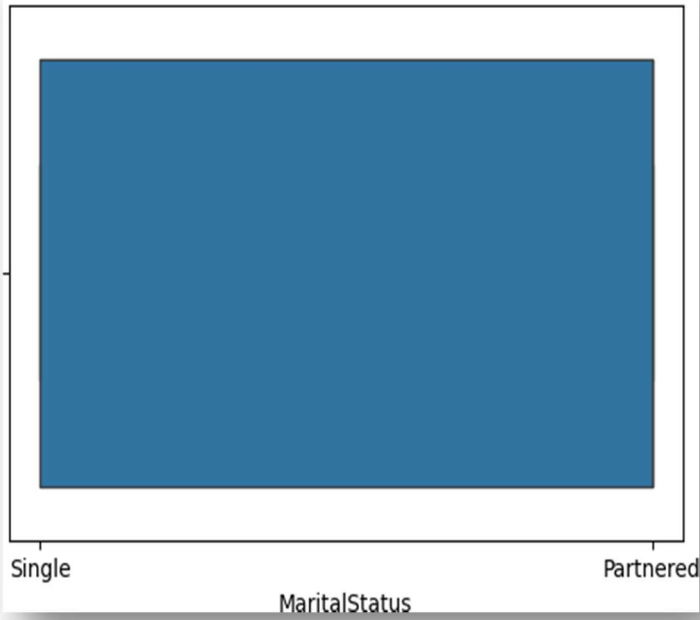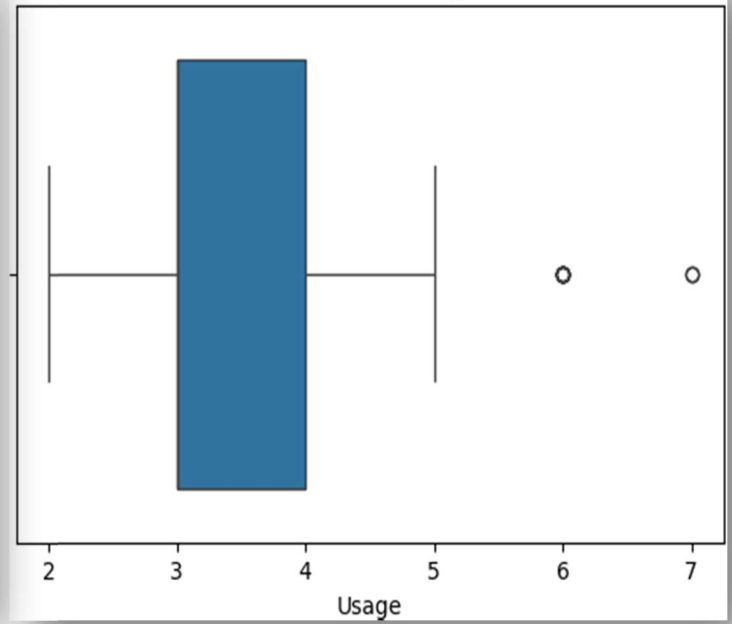
Box Plot:

Box plots allowed us to identify outliers and understand the spread of numerical data. They showed the median, quartiles, and extremes, giving insight into the range and variability of each feature.
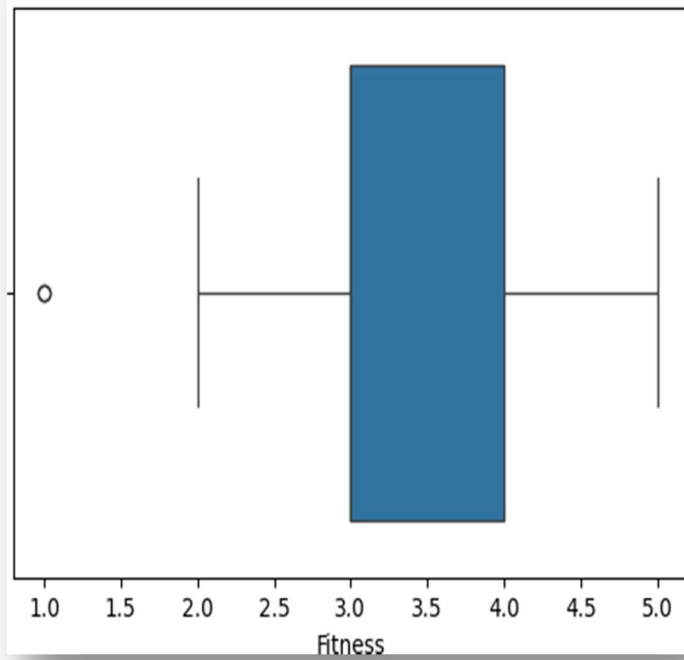

Box Plot of Product

Box Plot of Age

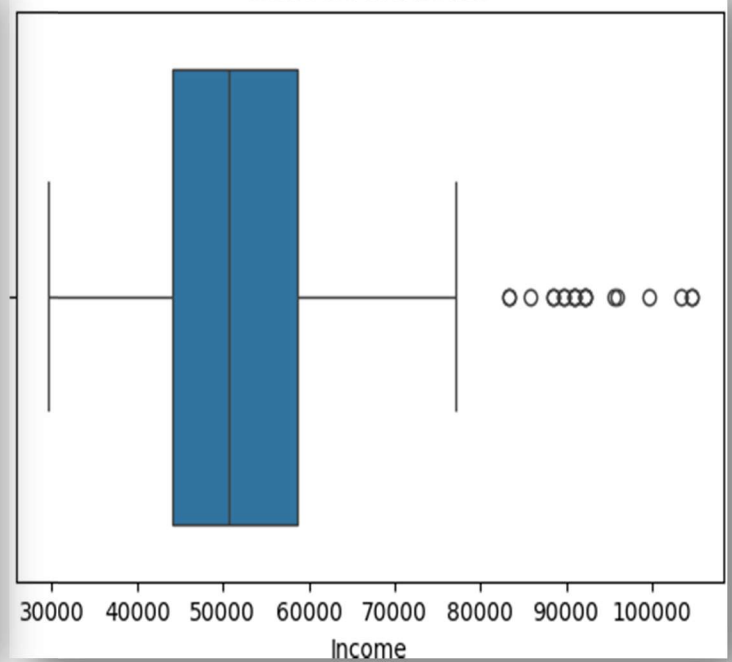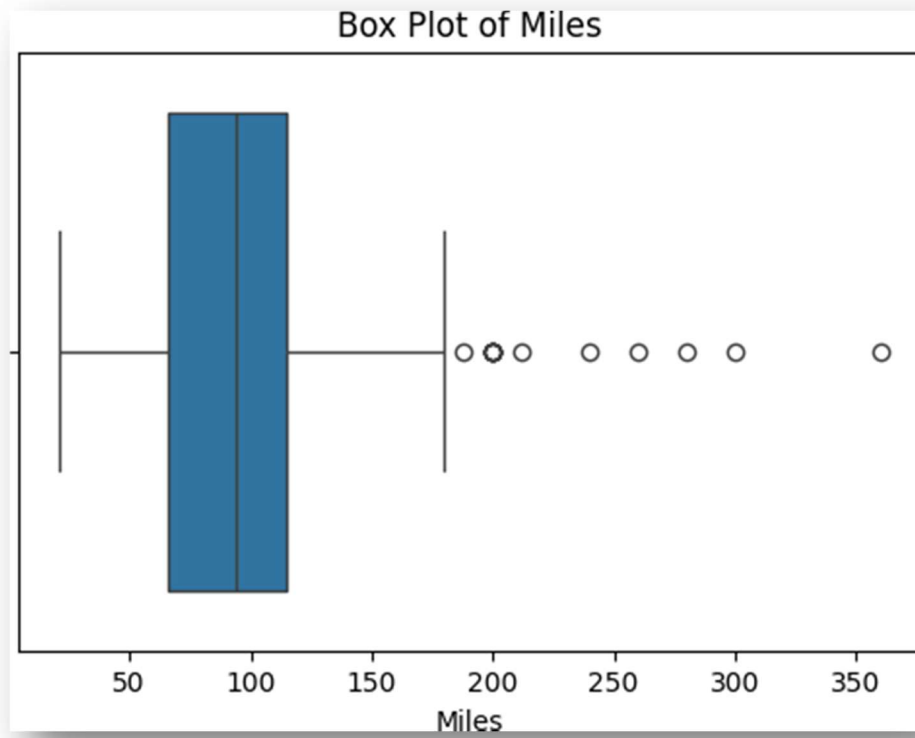Box Plot of Gender

Box Plot of Education
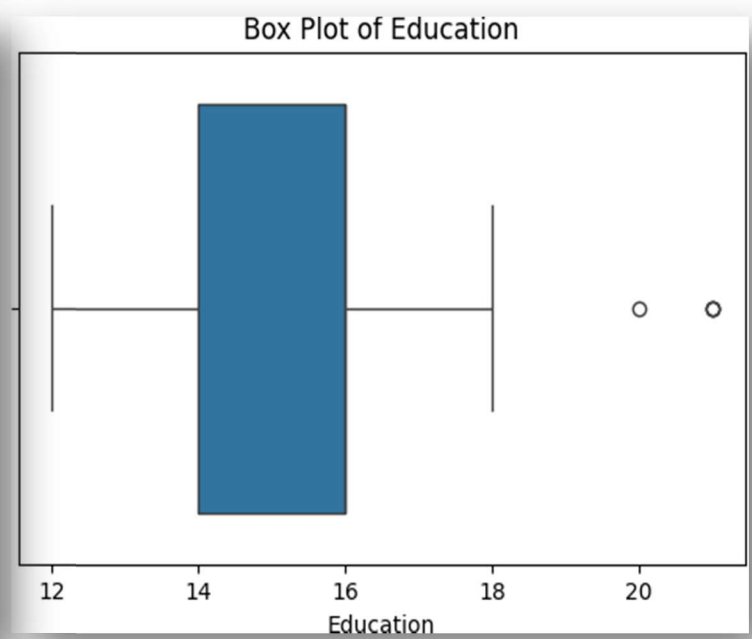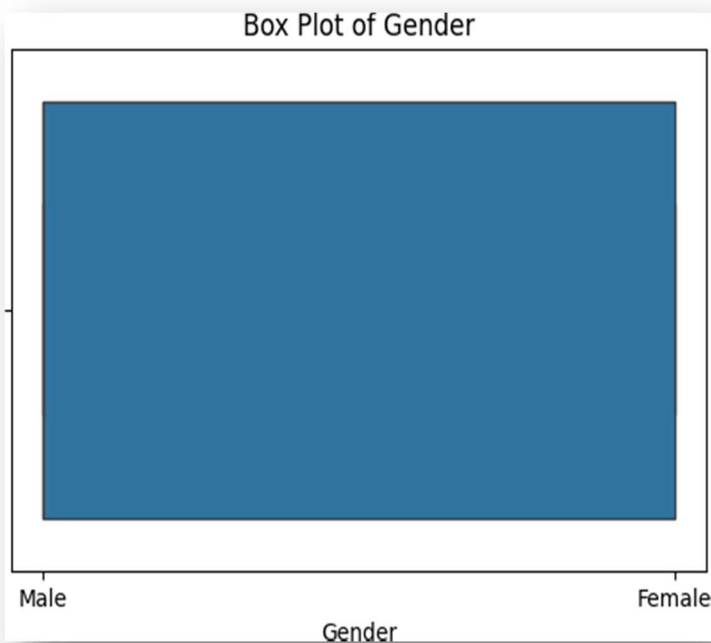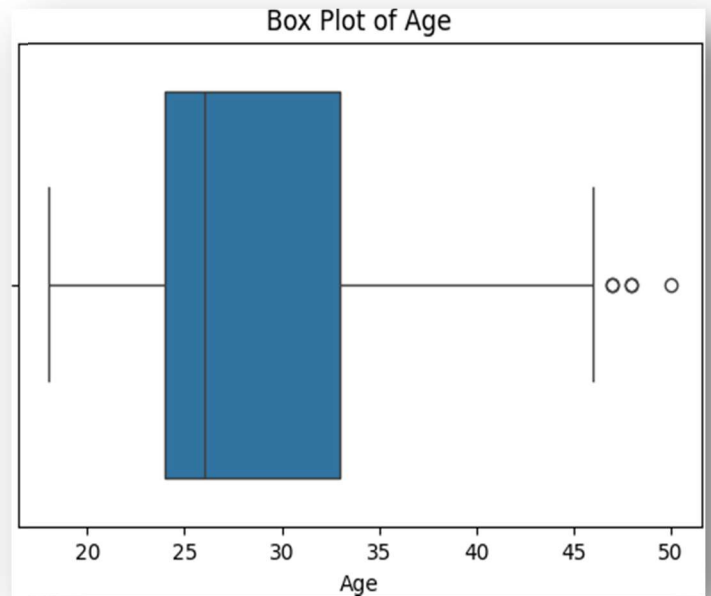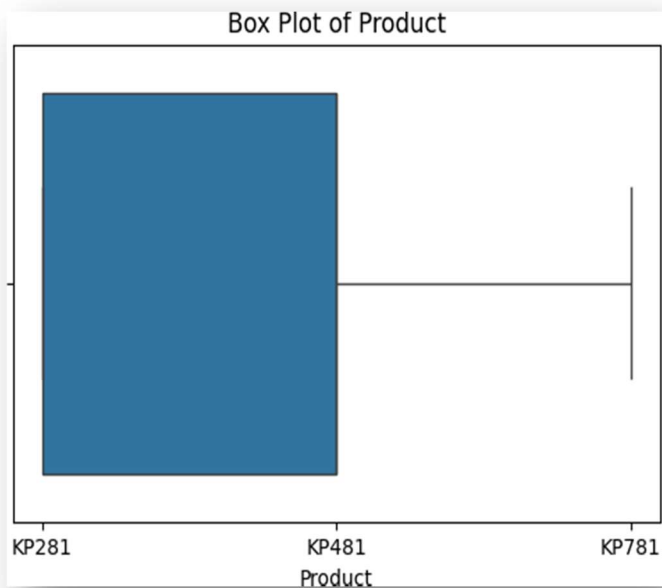
Box Plot of MaritalStatus

Box Plot of Usage

Box Plot of Fitness

Box Plot of Income

Box Plot of Miles

Univariate Analysis - Categorical features:

Count Plot:

Count plots were used to display the frequency of different categories in variables like Gender, Marital Status, and Product. These helped determine the most common values and identify any imbalances in the dataset.


Count Plot of Product

Count Plot of Gender

Count Plot of MaritalStatus
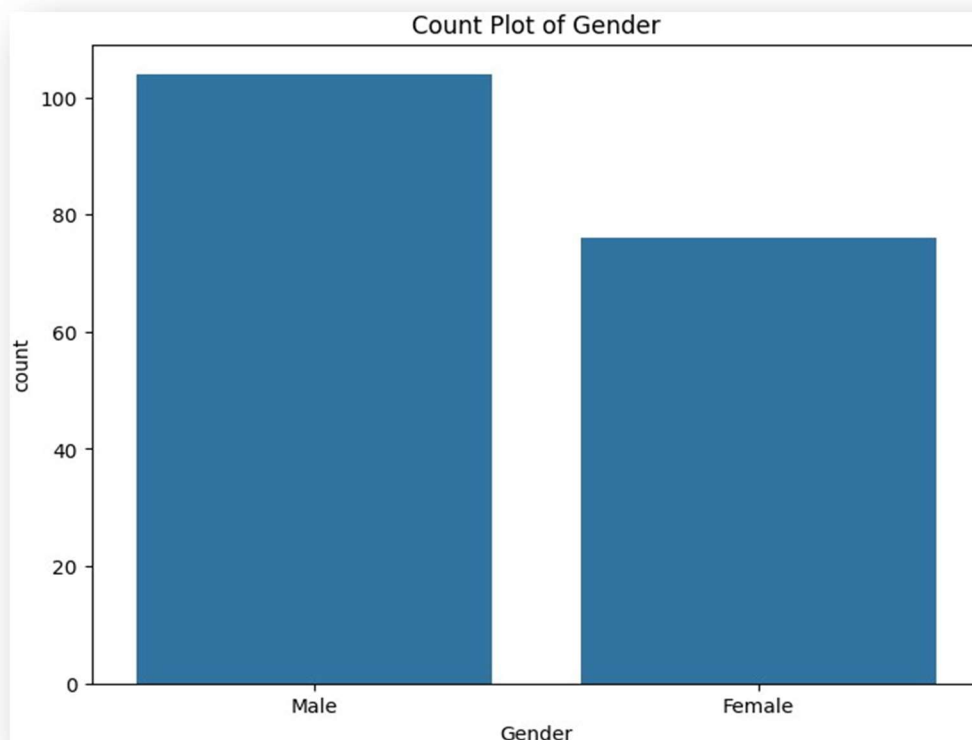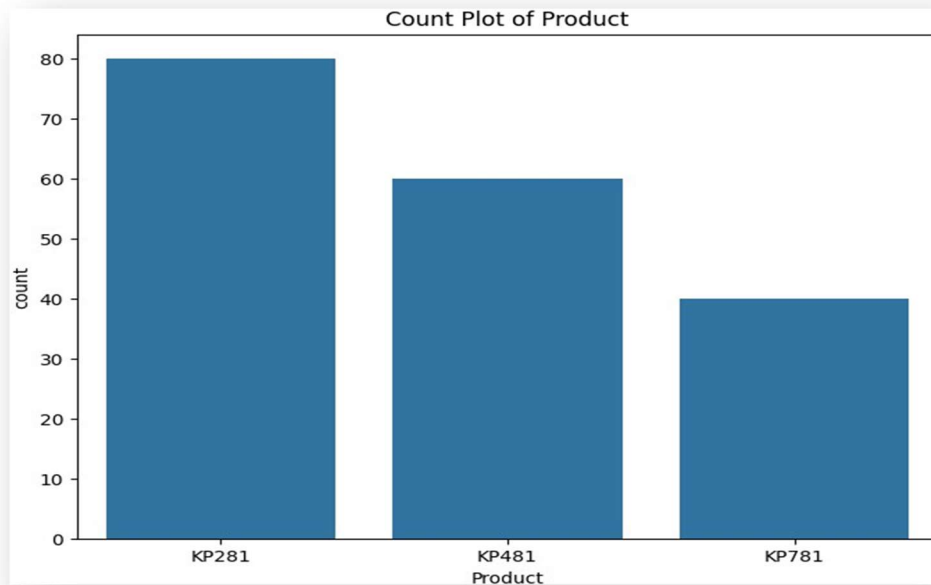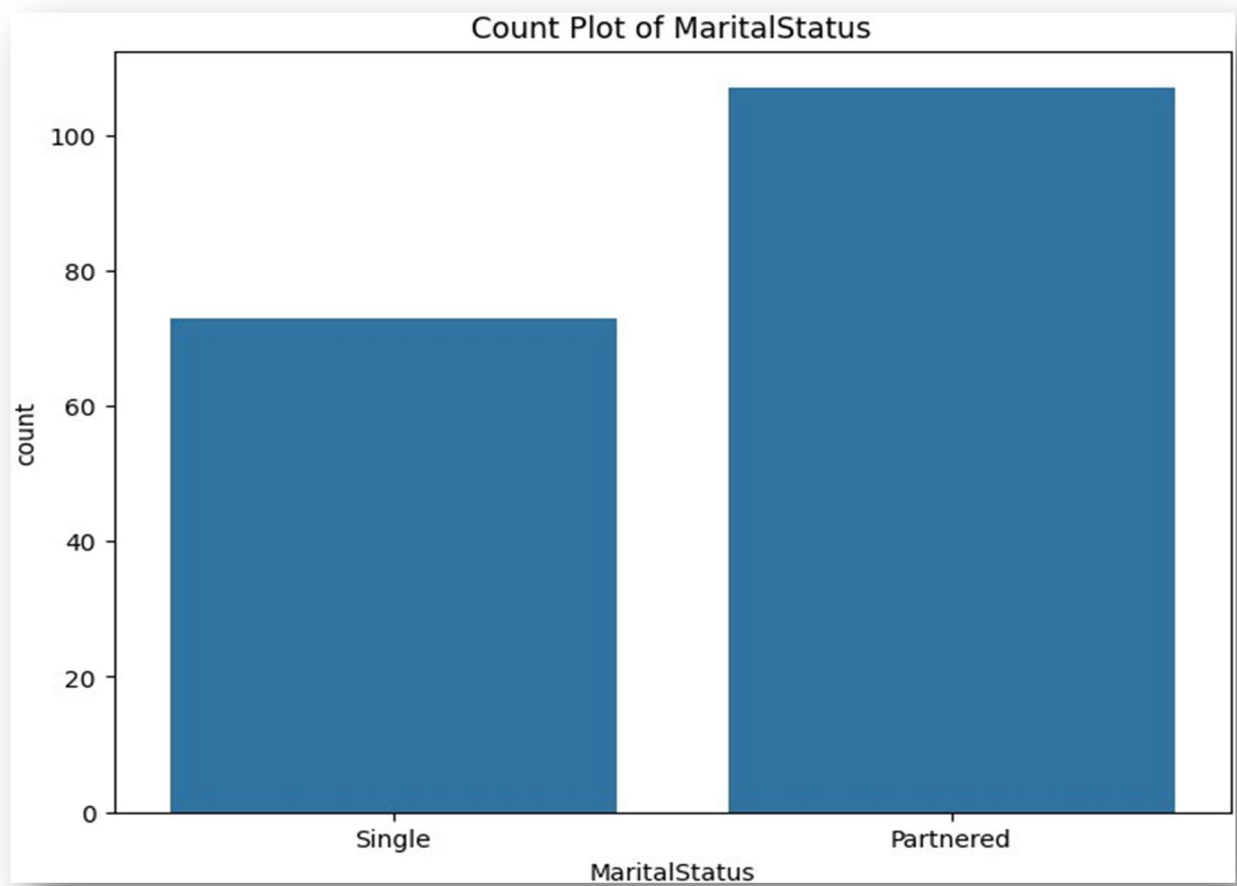
**Bivariate Analysis – Effect of Features on Product Purchased (Product vs Gender, Product vs Marital Status, Product vs Age):**

Product vs Gender / Marital Status / Age:
These plots examined how different features influenced the choice of treadmill. For example, count plots and box plots helped identify if males or partnered individuals preferred certain models, and how age distribution varied across products.

Product vs Age
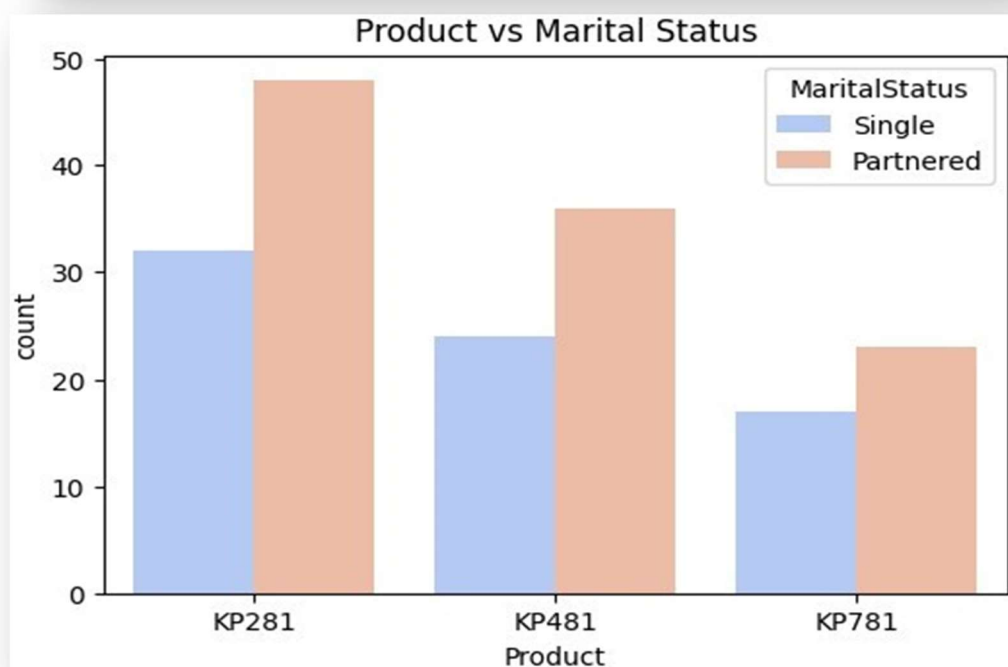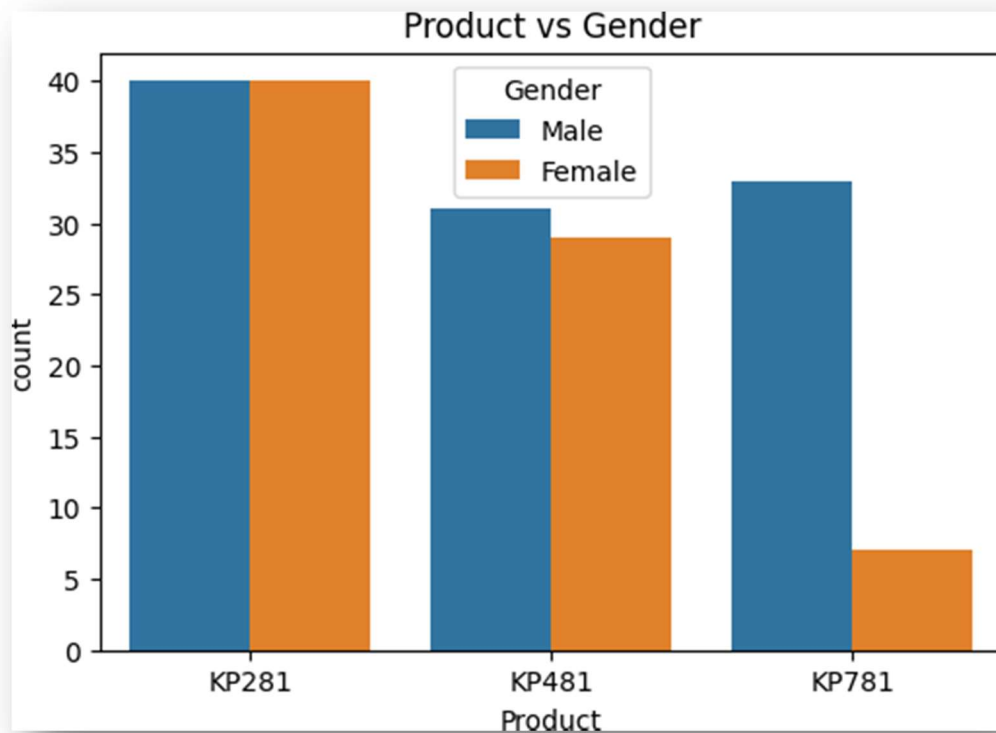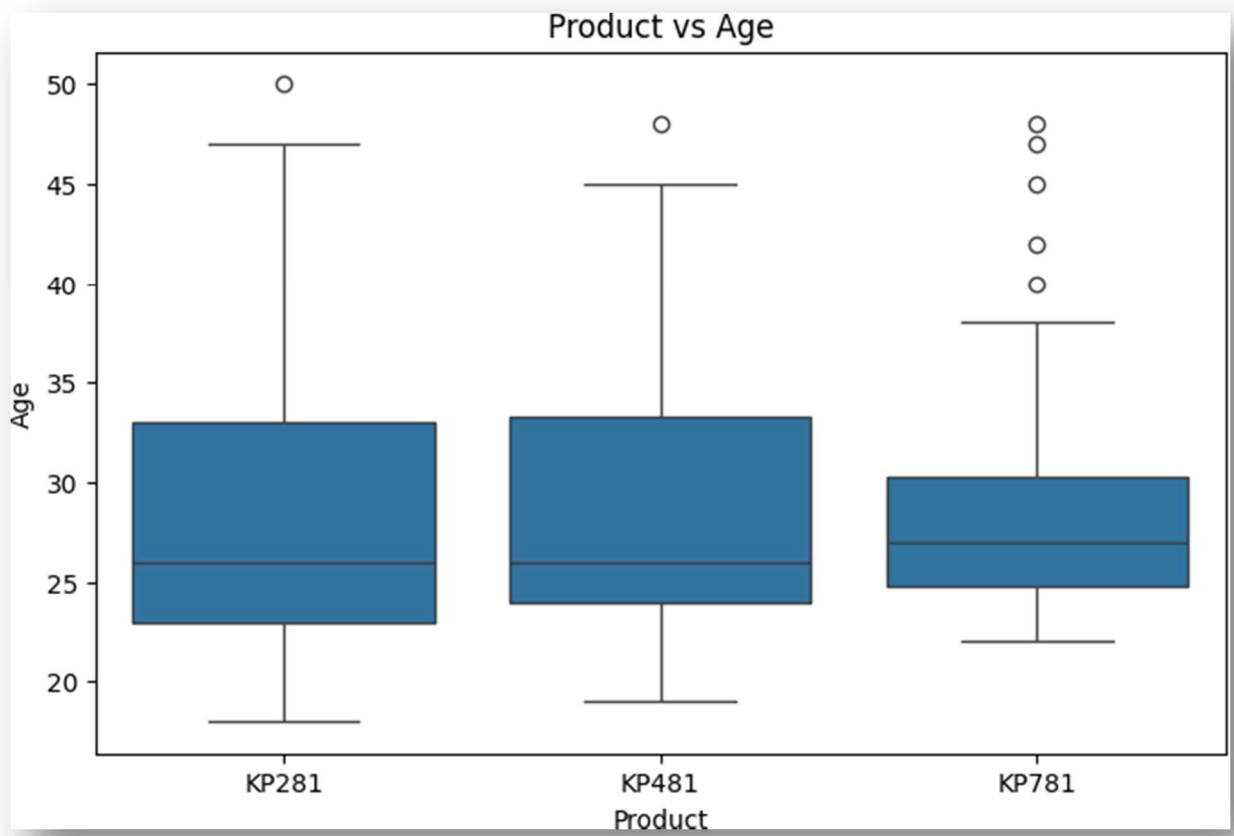
Multivariate Analysis:

Pair Plots:
Pair plots helped visualize the interaction between multiple variables at once, colored by product type. This provided a broad view of how customer traits like age, fitness, and income varied with each treadmill model
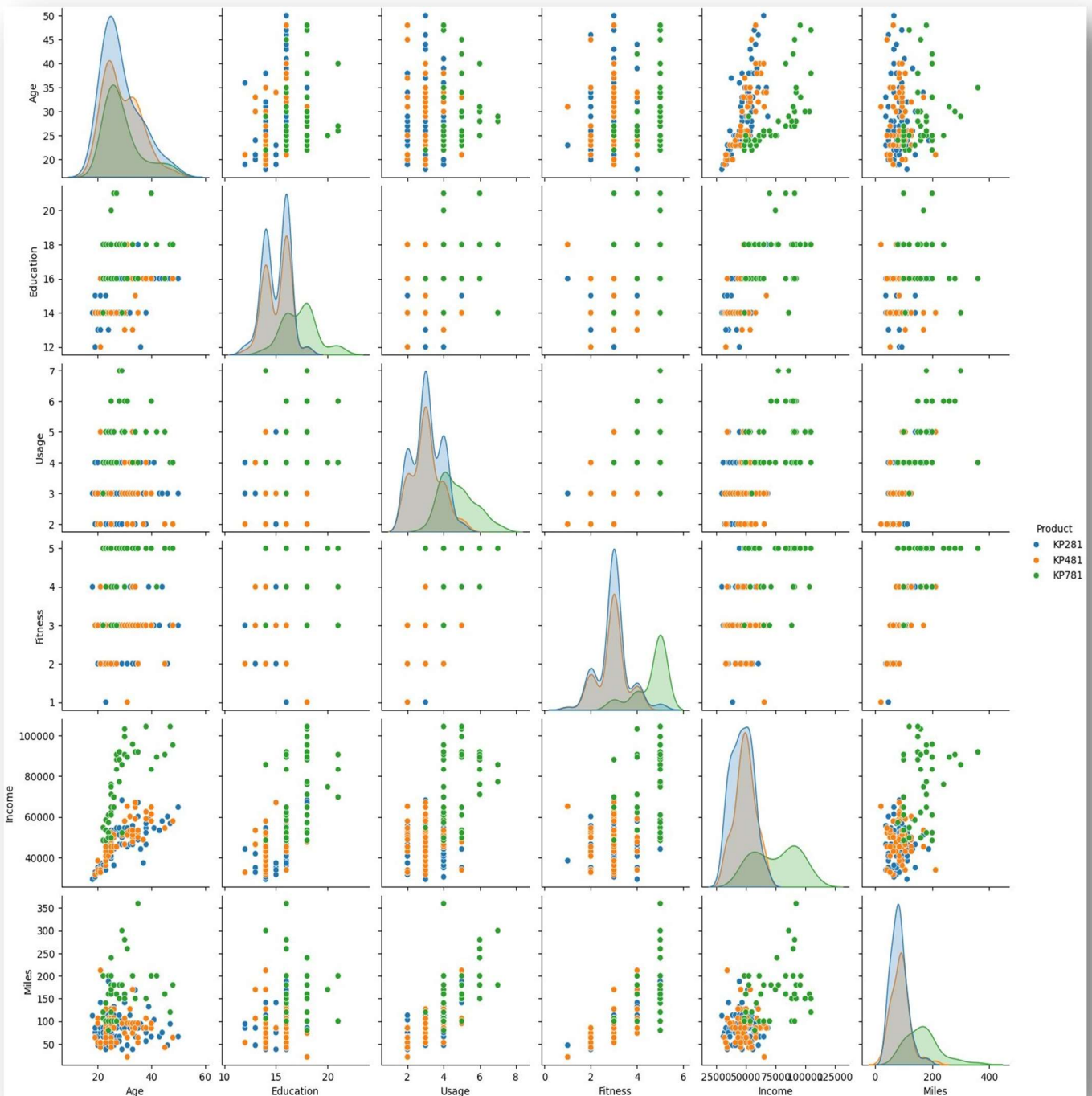
Correlation Analysis:

Heatmap:
A heatmap was used to visualize the correlation matrix of numerical features. It helped identify which variables are strongly or weakly related, such as the positive correlation between treadmill usage and miles run per week.
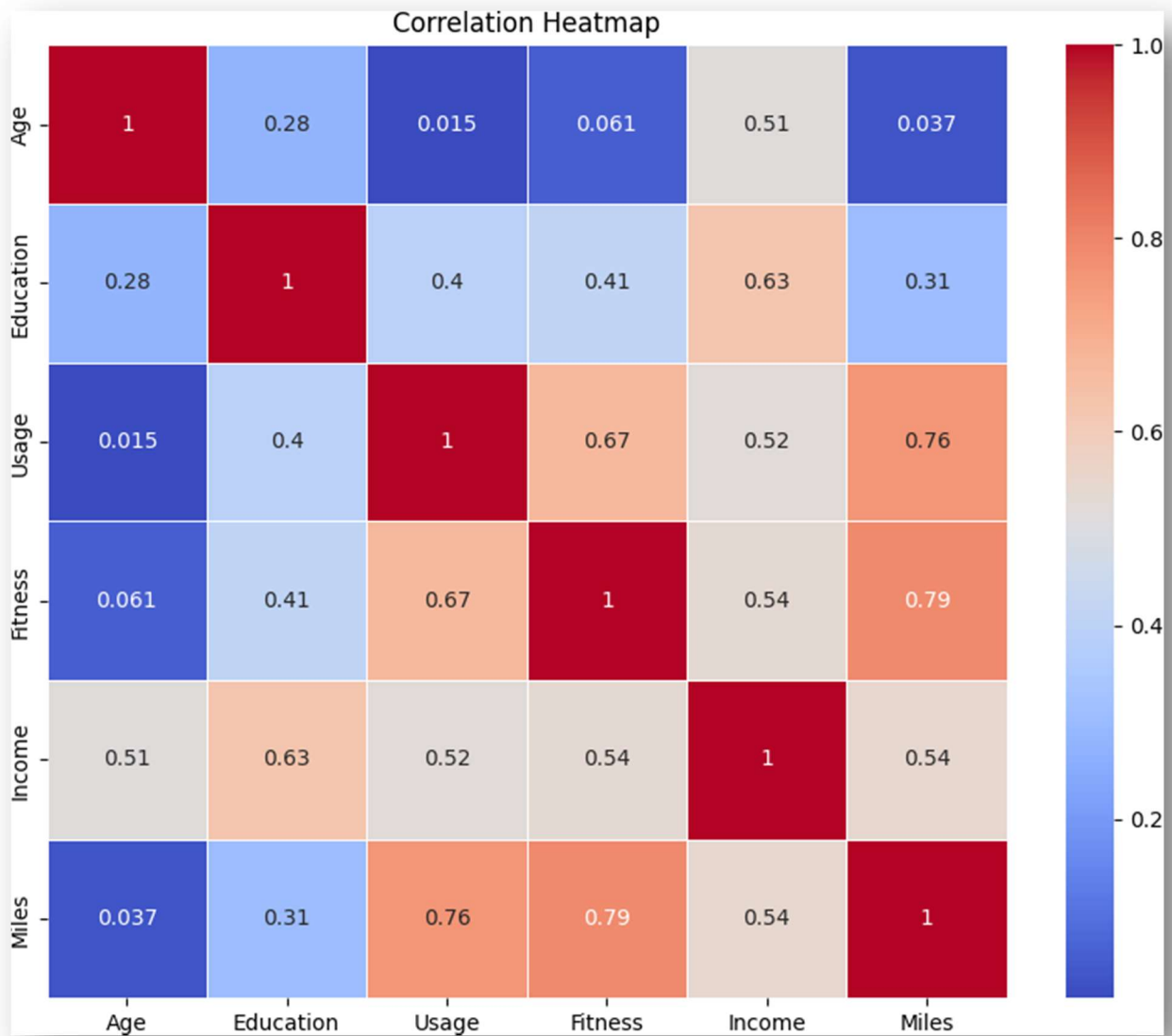
Key Findings & Insights:

1. Customer Demographics & Buying Trends:

- Most buyers fall within the **20s to 30s age range**.
- **KP281 has the highest sales**, while **KP781 lags in demand**.
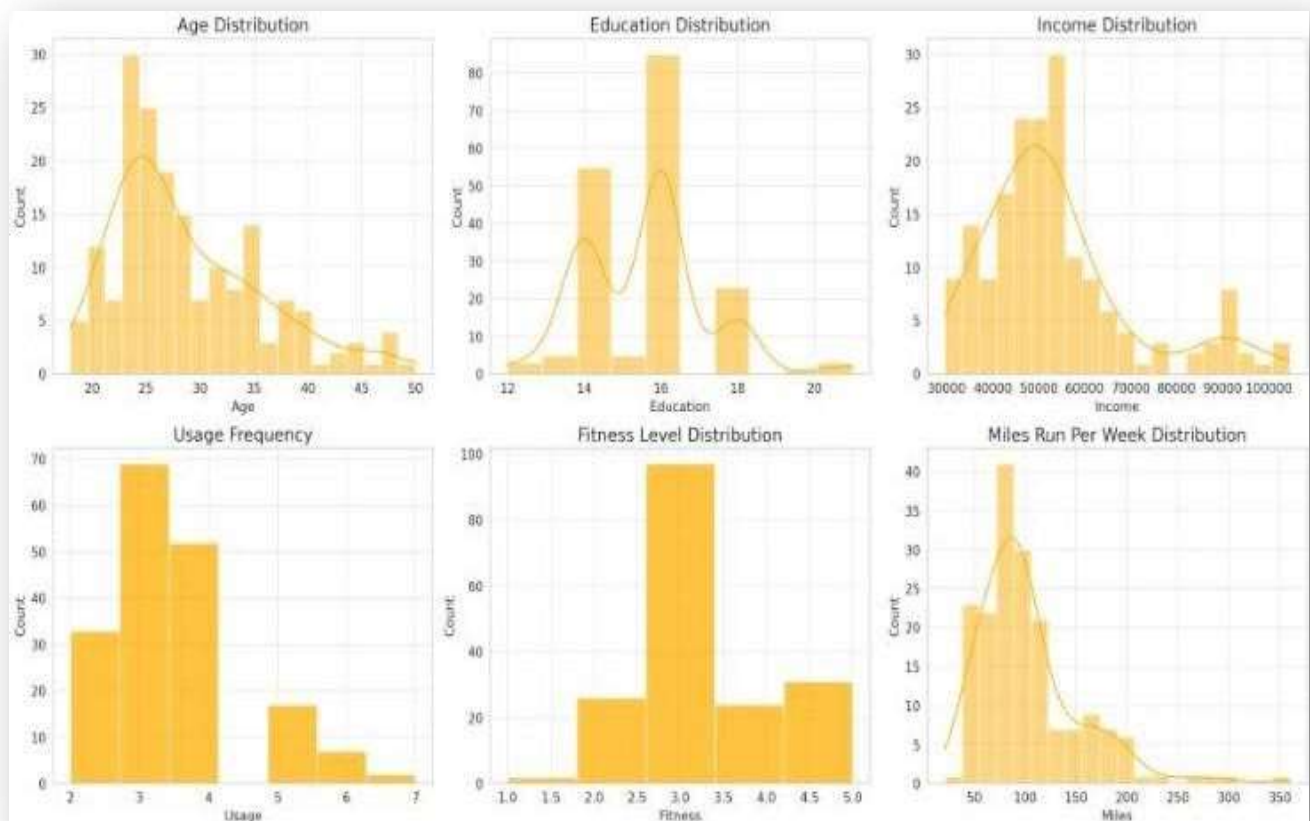- High-income customers (earning $70,000+) do not necessarily buy premium models like **KP781**.

Data Visualization:

2. Product Preferences & Correlation Analysis:

- **Age vs. Product Preference**: Younger customers tend to buy more budget-friendly models.
- **Income vs. Purchase Behavior**: No strong correlation between income and treadmill price selection.
- **Fitness Level vs. KP781 Sales**: **Highly fit customers prefer KP781**, but the overall sales remain low.
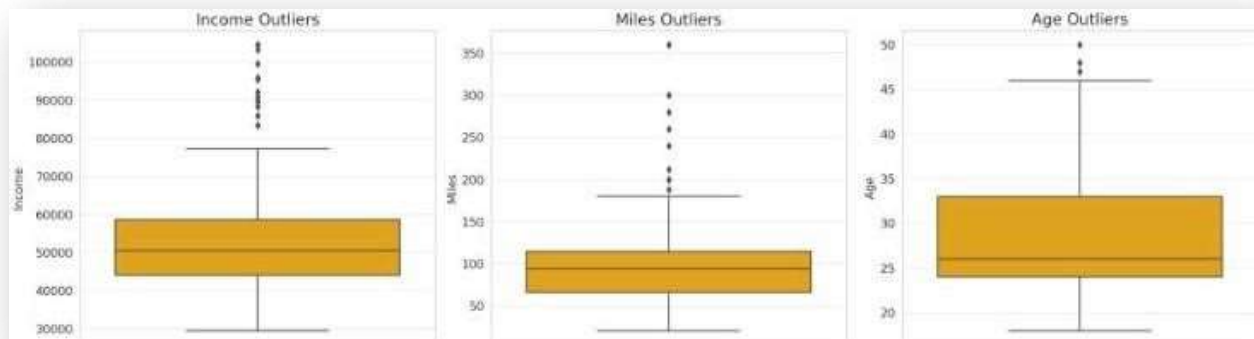
Data Visualization:
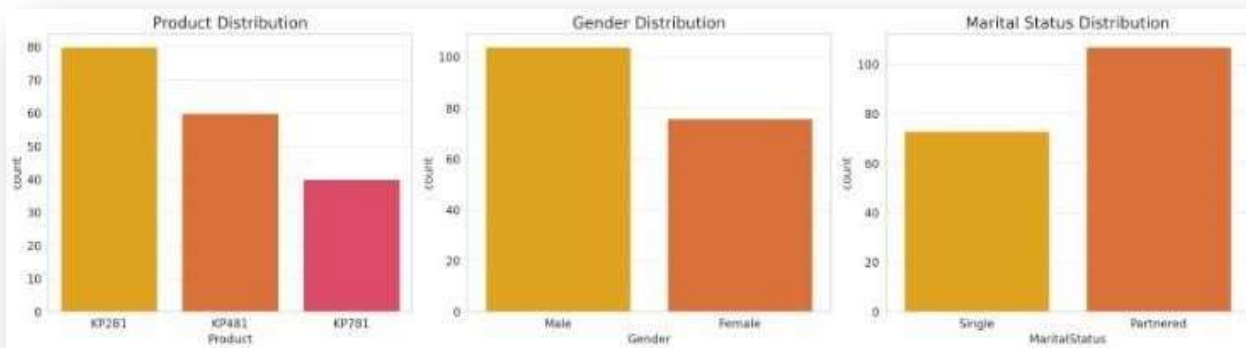
3. Outlier Detection & Data Quality:

- Income and mileage had **significant outliers**, suggesting either **niche customers** or **data entry errors**.
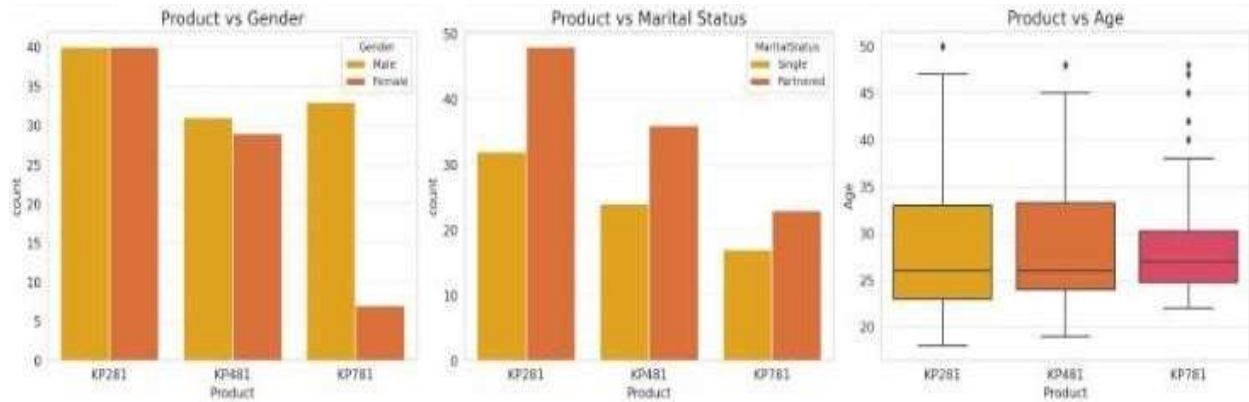- Some **high-salary customers** still opted for entry-level treadmills.

Data Visualization:

4. Gender & Marital Status Trends:

- Males purchase treadmills more frequently than females.
- Partnered customers make up a majority of treadmill buyers.

Data Visualization:

Actionable Recommendations:

1. Targeted Marketing Strategies:

- Focus advertising on **young buyers (20s-30s)** with fitness-related messaging.
- Emphasize KP781's **premium features** to high-income customers.
- Increase engagement among **female buyers** through specialized promotions.

2. Sales Promotions & Pricing Adjustments:

- Introduce **discounts, EMI plans, and bundle deals** for KP281, KP481, and KP781.
- Offer **premium incentives** (e.g., **VIP fitness plans**) to high-income customers to boost KP781 sales.

3. Data Quality & Customer Segmentation:

- Perform **data cleaning** to remove incorrect outliers.
- Use **AI-based recommendations** to provide personalized product suggestions.

4. Product Positioning & Branding:

- Improve **KP781's market appeal** by partnering with **influencers & fitness trainers**.
- Introduce **limited-time offers & loyalty programs** to encourage repeat purchases.

Conclusion:

This project provided an in-depth analysis of AeroFit's treadmill sales data using **EDA techniques**. The study identified trends in customer demographics, income levels, and fitness preferences, leading to actionable recommendations. By **refining marketing strategies, improving product positioning, and optimizing data quality**, AeroFit can enhance customer engagement and drive higher sales. Future analysis can incorporate **predictive modeling** to better forecast customer purchasing behavior.

References:

- AeroFit Treadmill Sales Dataset (aerofit_treadmill_data.xlsx)
- Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
- Exploratory Data Analysis Techniques
- Data Source: AeroFit customer purchase records analyzed through EDA techniques.
- Python Libraries: Pandas, NumPy, Seaborn, Matplotlib
- Exploratory Data Analysis Techniques