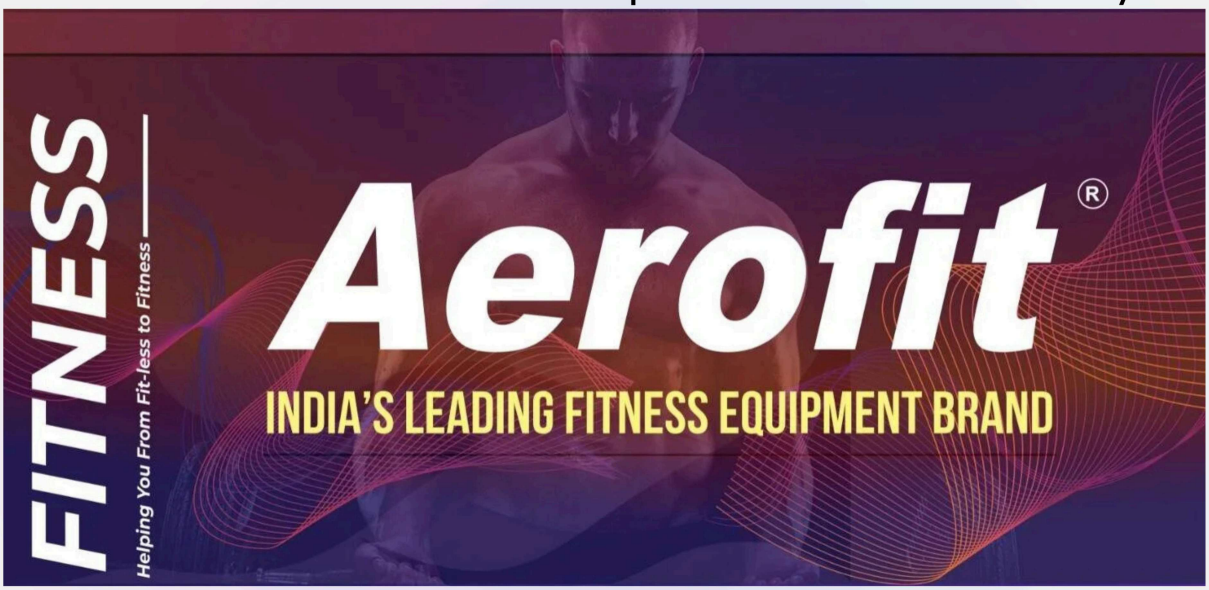
Business Case: Aerofit - Descriptive Statistics & Probability



Compiled by Samarth Kolge.

In [41]:
```
# Import libraries
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [42]:
```
#Load Dataset
df = pd.read_csv("aerofit_treadmill.csv")
```

In [12]:
```
len(df)
```

Out[12]:
```
180
```

In [7]:
```
df.shape
```

Out[7]:
```
(180, 9)
```

In [14]:
```
df.dtypes
```

Out[14]:
```
Product          object
Age               int64
Gender           object
Education         int64
MaritalStatus    object
Usage             int64
Fitness           int64
Income            int64
Miles             int64
dtype: object
```

In [15]:
```
df.head()
```

Out[15]:

	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

In [16]:
```
df.tail()
```

Out[16]:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
175	KP781	40	Male	21	Single	6	5	83416	200
176	KP781	42	Male	18	Single	5	4	89641	200
177	KP781	45	Male	16	Single	5	5	90886	160
178	KP781	47	Male	18	Partnered	4	5	104581	120
179	KP781	48	Male	18	Partnered	4	5	95508	180

In [18]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Product        180 non-null    object
 1   Age            180 non-null    int64 
 2   Gender         180 non-null    object
 3   Education      180 non-null    int64 
 4   MaritalStatus  180 non-null    object
 5   Usage          180 non-null    int64 
 6   Fitness        180 non-null    int64 
 7   Income         180 non-null    int64 
 8   Miles          180 non-null    int64 
dtypes: int64(6), object(3)
memory usage: 12.8+ KB
```

In [23]:
```
df.isnull().sum()
```

Out[23]:
```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
Miles            0
dtype: int64
```

In [39]:
```
df["Product"].nunique()
```

Out[39]:
```
3
```

In [40]:
```
df["Product"].value_counts()
```

Out[40]:
```
KP281    80
KP481    60
KP781    40
Name: Product, dtype: int64
```

There are no missing values in the data.

There are three unique products in the dataset.

KP281 is the most frequently purchased product.

In [54]:
```
df.describe()
```

Out[54]:

	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

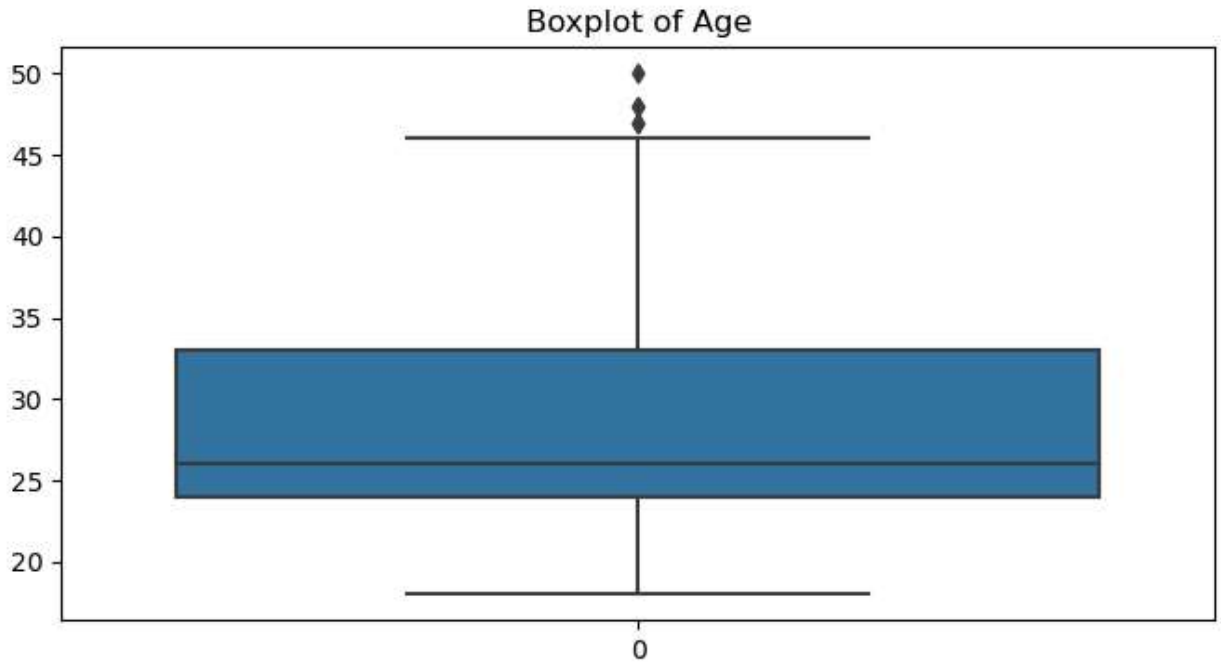
The minimum and maximum ages of the individuals are 18 and 50, respectively, with a mean age of 28.79. Additionally, 75% of individuals are aged 33 or below.

Most individuals have completed 16 years of education, with 75% having 16 years or fewer.

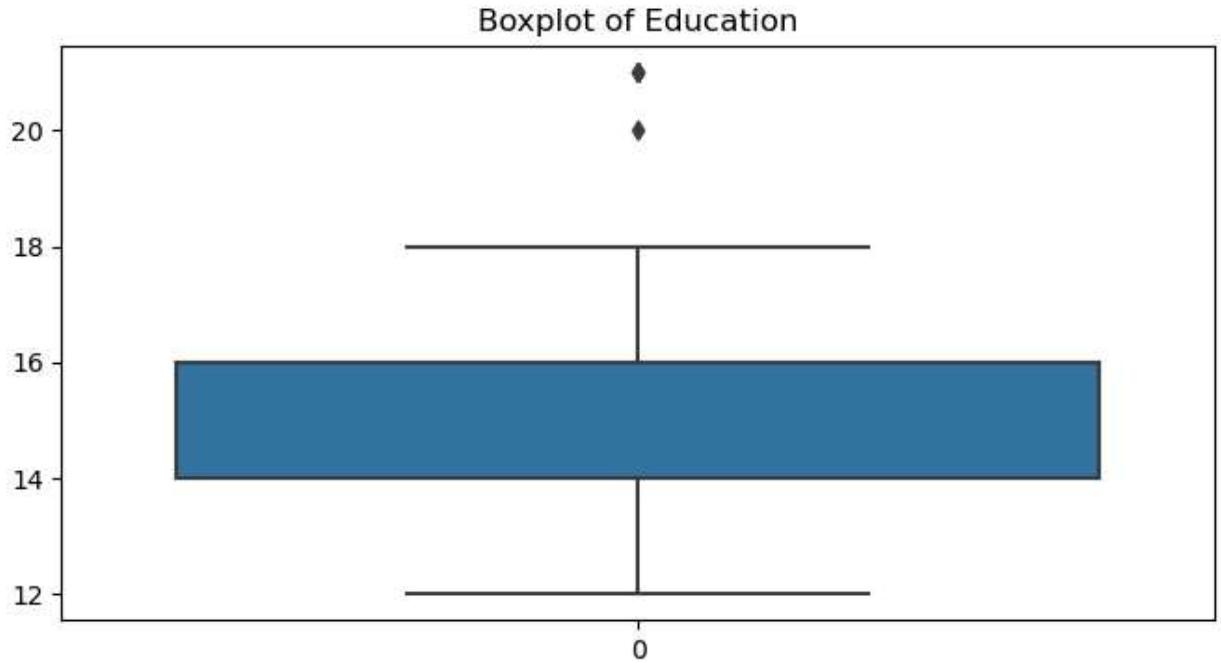
There is a noticeable difference between the mean and median values of both Income and Miles. This difference clearly indicates the presence of outliers.

Let's confirm this using a boxplot.

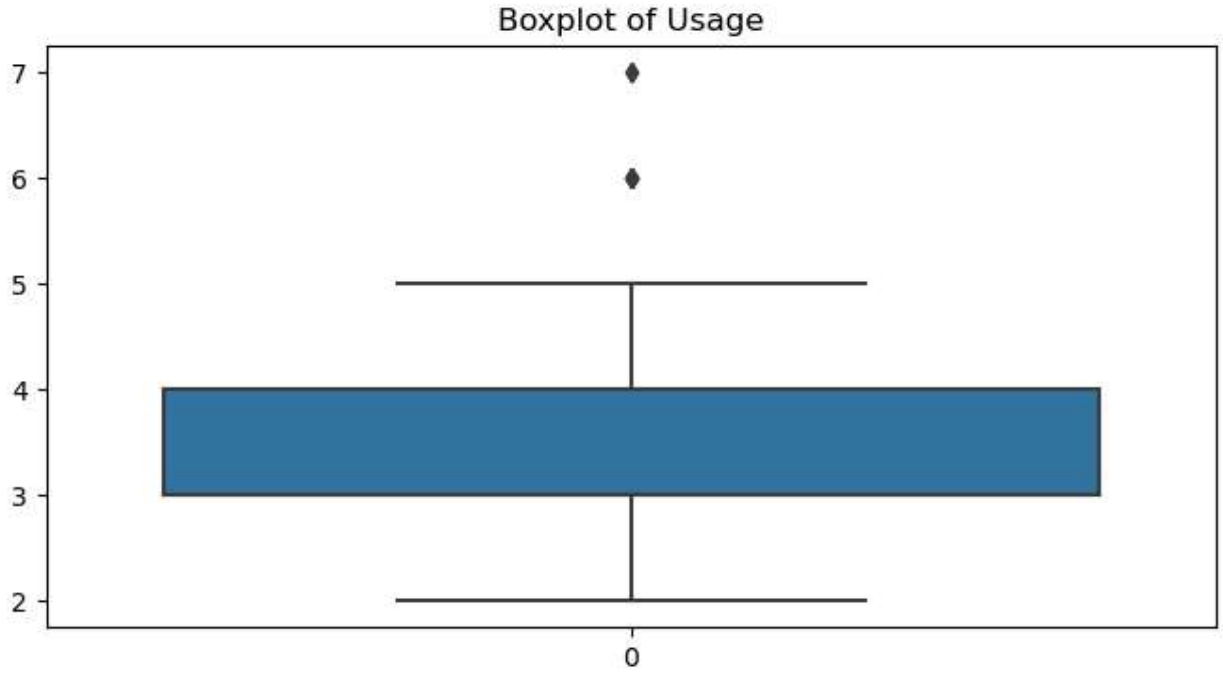
In [48]:
```
# Plot boxplots to find outliers for numerical columns
# 1. Boxplot for Age
plt.figure(figsize=(8, 4))
sns.boxplot(df['Age'])
plt.title('Boxplot of Age')
plt.show()
```



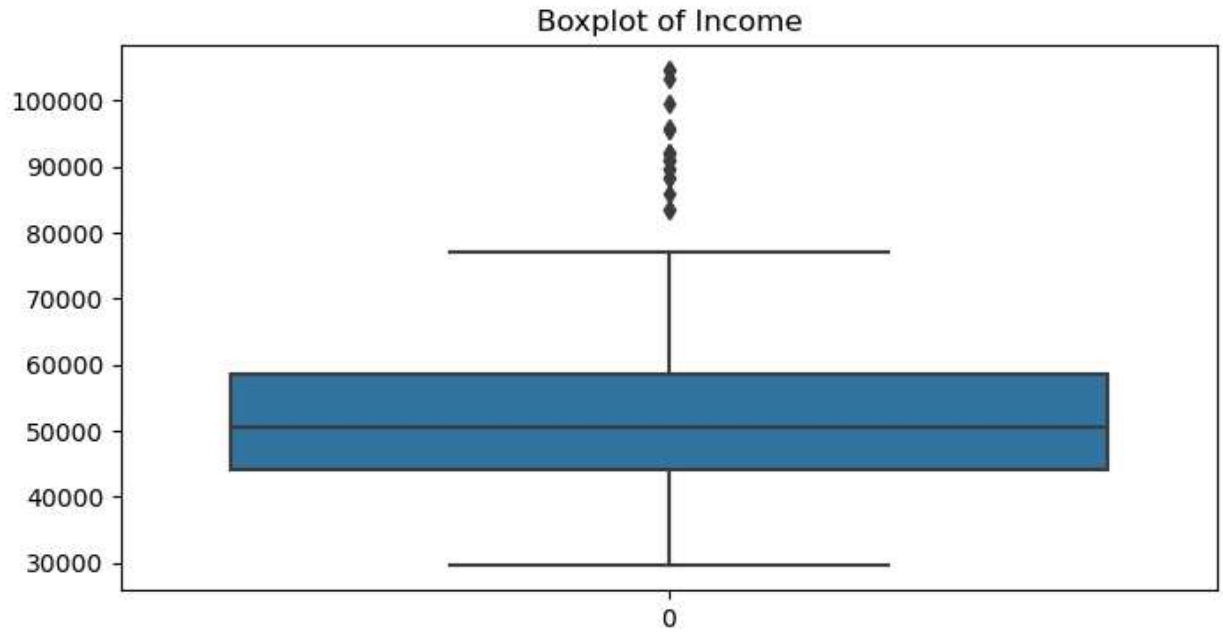
In [49]:
```
# 2. Boxplot for Education
plt.figure(figsize=(8, 4))
sns.boxplot(df['Education'])
plt.title('Boxplot of Education')
plt.show()
```



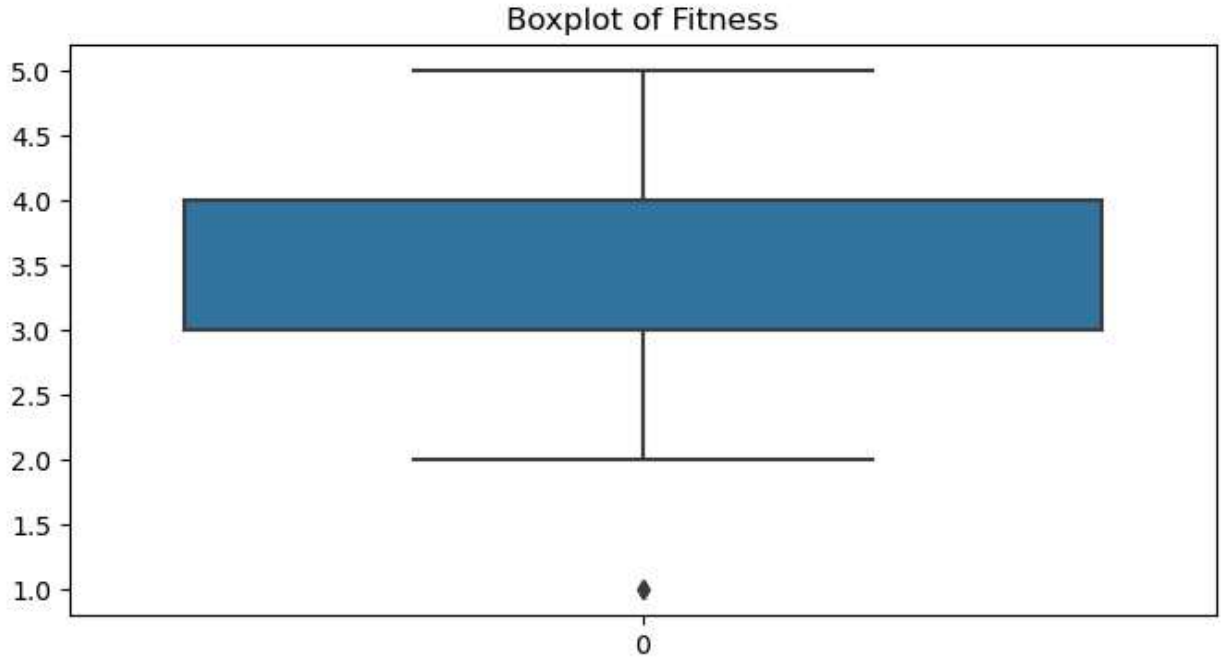
In [50]:
```
# 3. Boxplot for Usage
plt.figure(figsize=(8, 4))
sns.boxplot(df['Usage'])
plt.title('Boxplot of Usage')
plt.show()
```



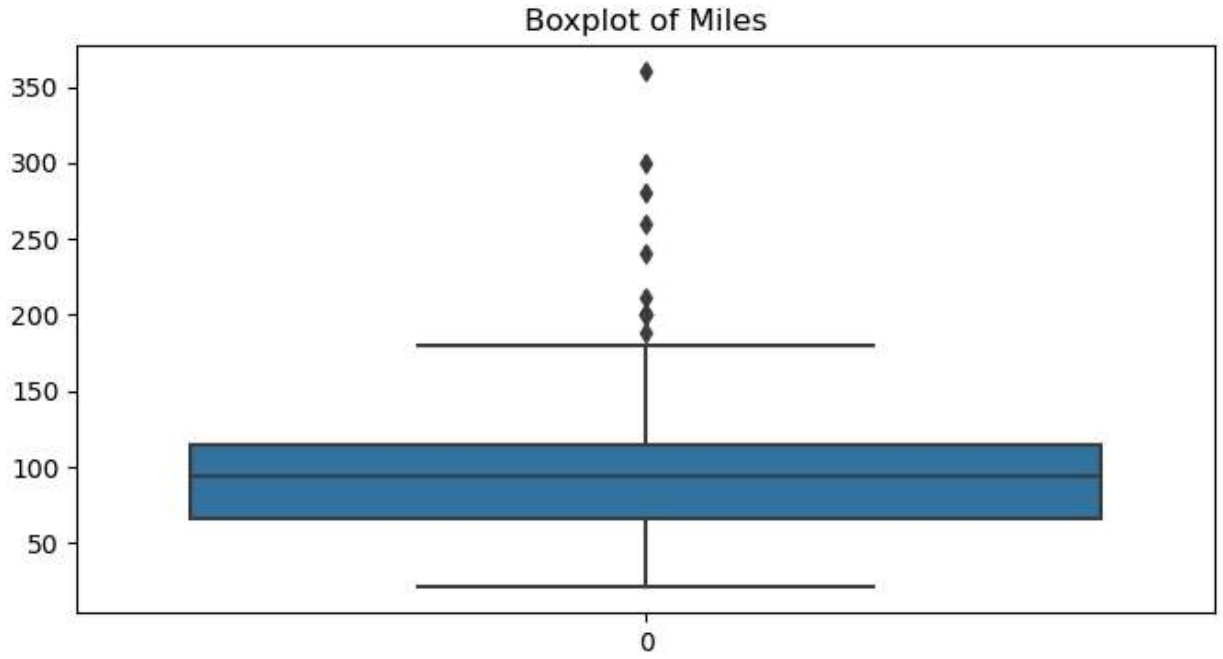
In [51]:
```
# 4. Boxplot for Income
plt.figure(figsize=(8, 4))
sns.boxplot(df['Income'])
plt.title('Boxplot of Income')
plt.show()
```



In [52]:
```
# 5. Boxplot for Fitness
plt.figure(figsize=(8, 4))
sns.boxplot(df['Fitness'])
plt.title('Boxplot of Fitness')
plt.show()
```



In [53]:
```
# 6. Boxplot for Miles
plt.figure(figsize=(8, 4))
sns.boxplot(df['Miles'])
plt.title('Boxplot of Miles')
plt.show()
```



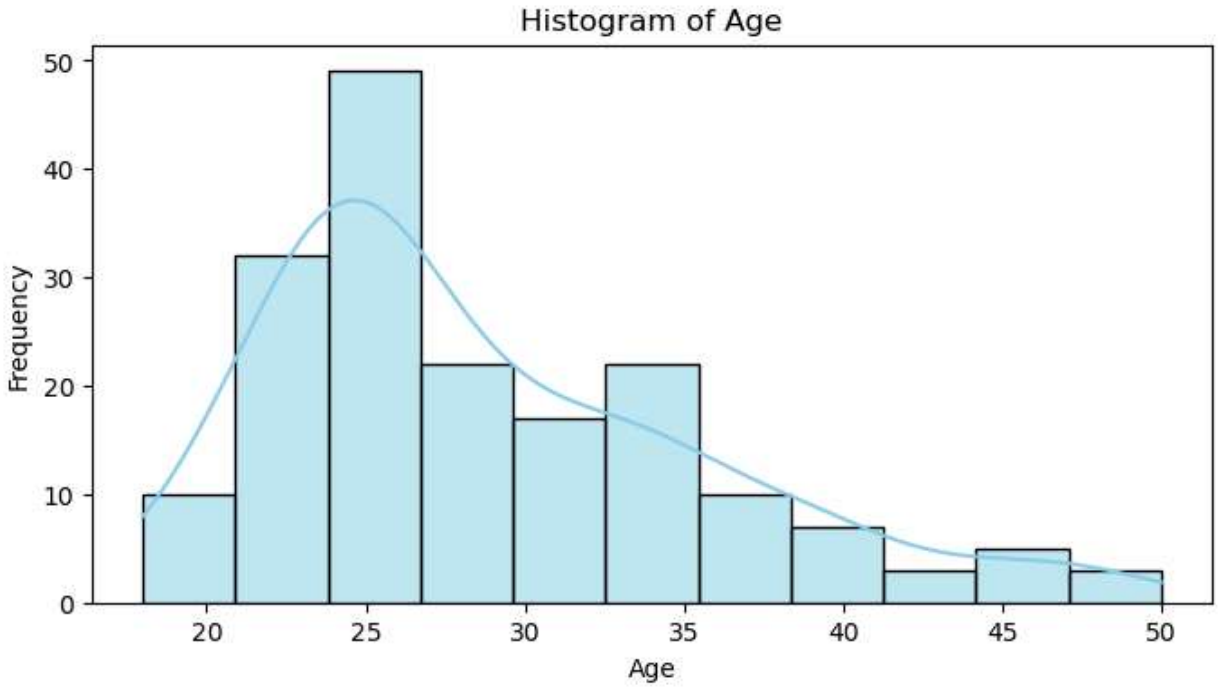
The boxplots of Income and Miles clearly show the presence of outliers.

In [62]:
```
# Display the count of unique values for categorical columns
categorical_cols = ['Product', 'Gender', 'MaritalStatus']

for col in categorical_cols:
    print(f'\nValue counts for {col}:')
    print(df[col].value_counts())
```

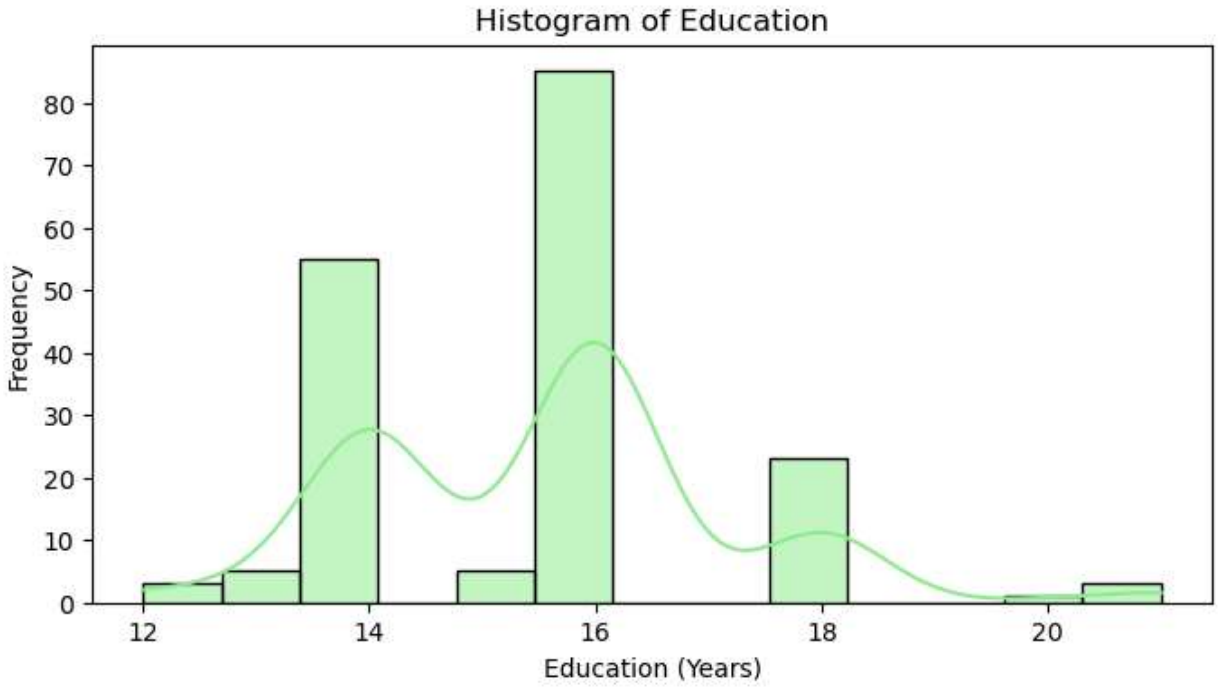
```
Value counts for Product:
KP281    80
KP481    60
KP781    40
Name: Product, dtype: int64

Value counts for Gender:
Male      104
Female     76
Name: Gender, dtype: int64

Value counts for MaritalStatus:
Partnered    107
Single        73
Name: MaritalStatus, dtype: int64
```

The most purchased treadmill product is KP281, with 80 units sold, followed by KP481 with 60 units, and KP781 with 40 units, indicating that the entry-level model (KP281) is the most popular. In terms of gender distribution, males (104) make up the majority of customers compared to females (76). Additionally, most customers are partnered (107) rather than single (73), suggesting that partnered individuals are more likely to purchase fitness equipment. This insight can help Aerofit target their marketing efforts toward younger, partnered males for the KP281 model.

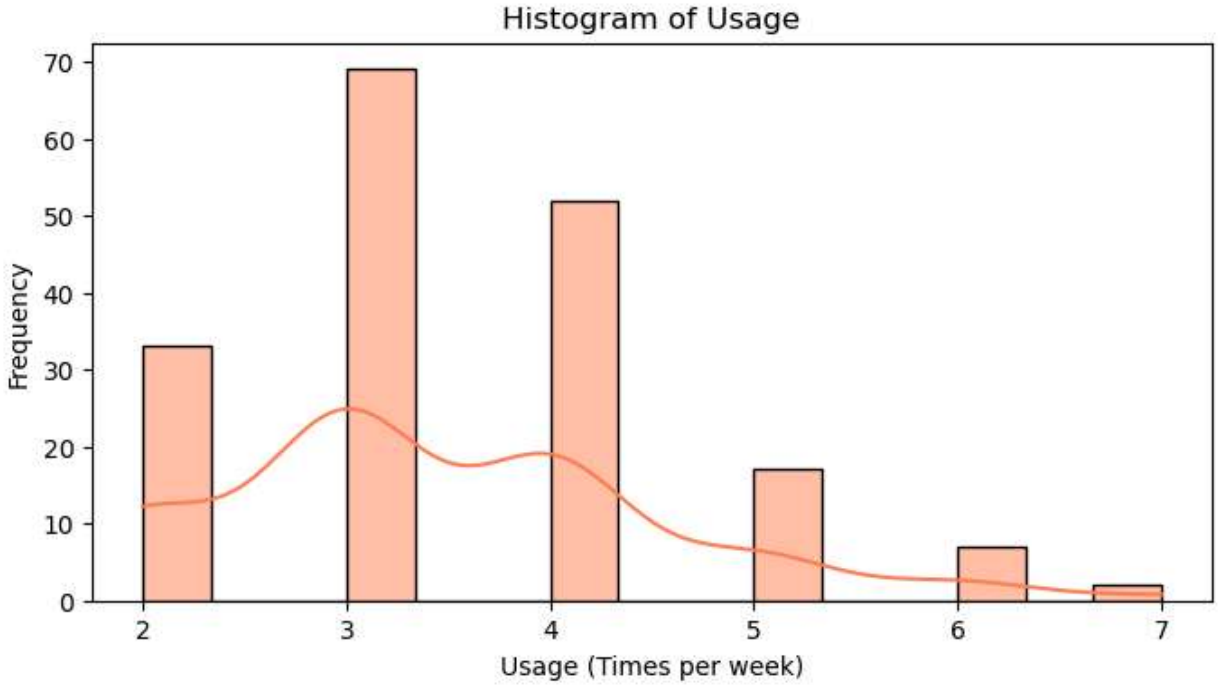
In [68]:
```
# Plot histograms for numerical columns
# 1. Histogram for Age
plt.figure(figsize=(8, 4))
sns.histplot(df['Age'], kde=True, color='skyblue')
plt.title('Histogram of Age')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



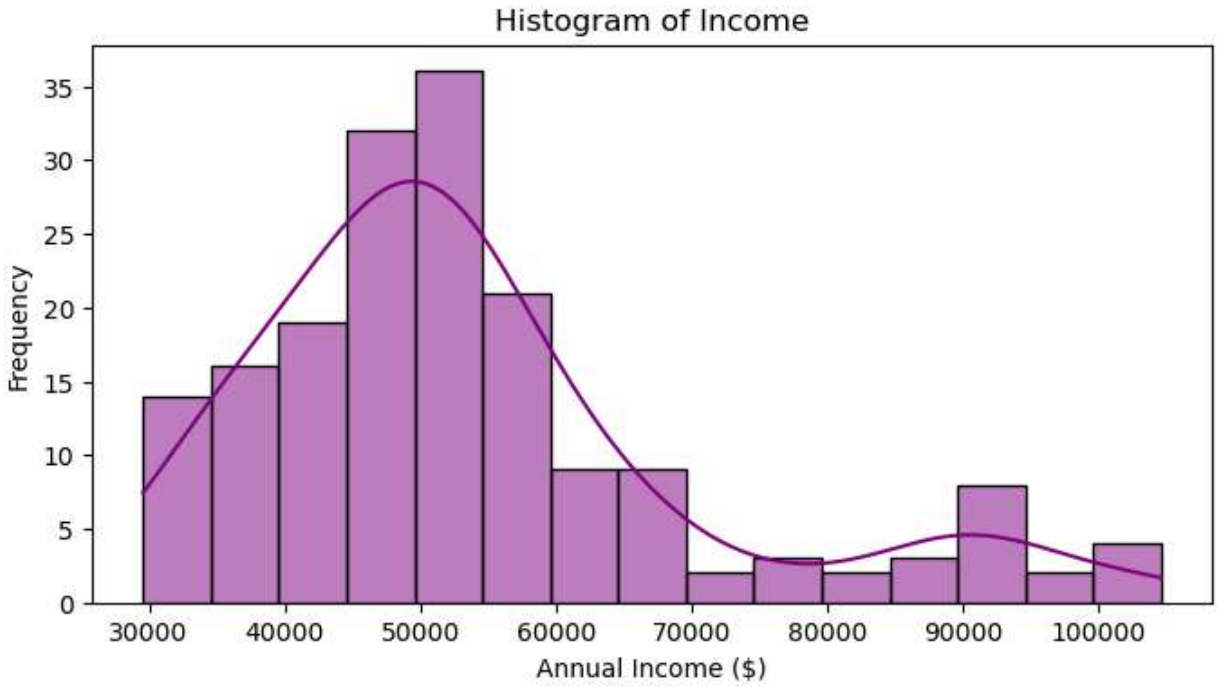
In [69]:
```
# 2. Histogram for Education
plt.figure(figsize=(8, 4))
sns.histplot(df['Education'], kde=True, color='lightgreen')
plt.title('Histogram of Education')
plt.xlabel('Education (Years)')
plt.ylabel('Frequency')
plt.show()
```



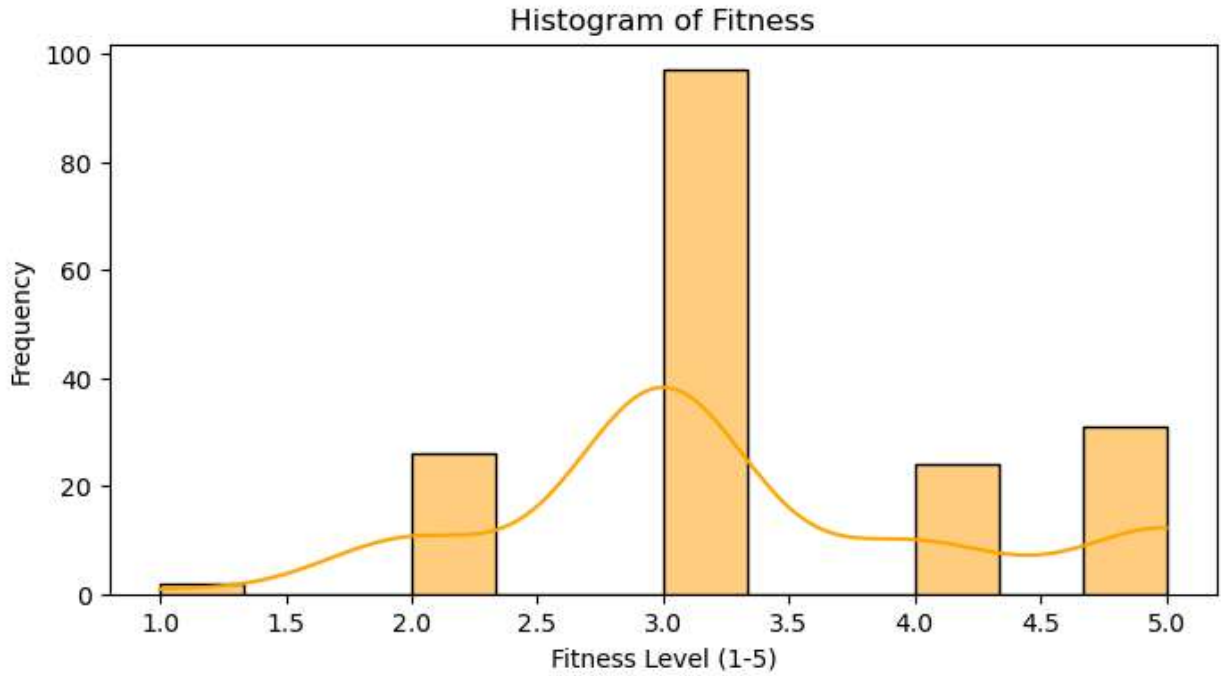
In [70]:
```
# 3. Histogram for Usage
plt.figure(figsize=(8, 4))
sns.histplot(df['Usage'], kde=True, color='coral')
plt.title('Histogram of Usage')
plt.xlabel('Usage (Times per week)')
plt.ylabel('Frequency')
plt.show()
```



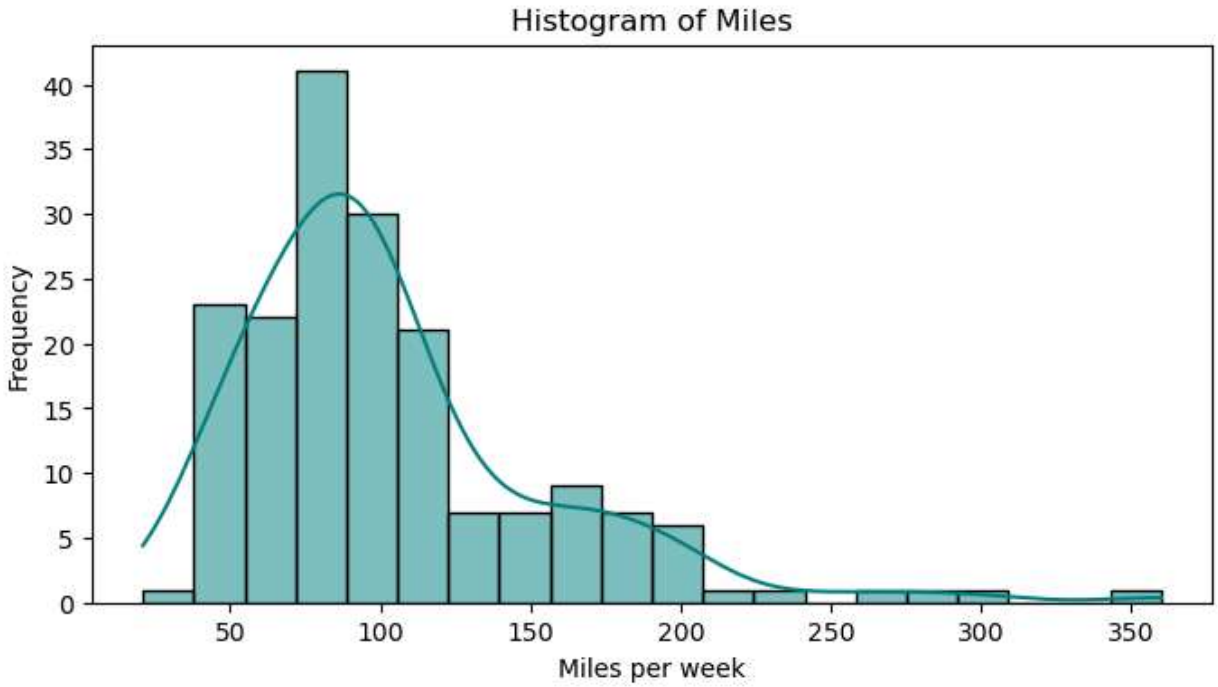
In [72]:
```
# 4. Histogram for Income
plt.figure(figsize=(8, 4))
sns.histplot(df['Income'], kde=True, color='purple')
plt.title('Histogram of Income')
plt.xlabel('Annual Income ($)')
plt.ylabel('Frequency')
plt.show()
```



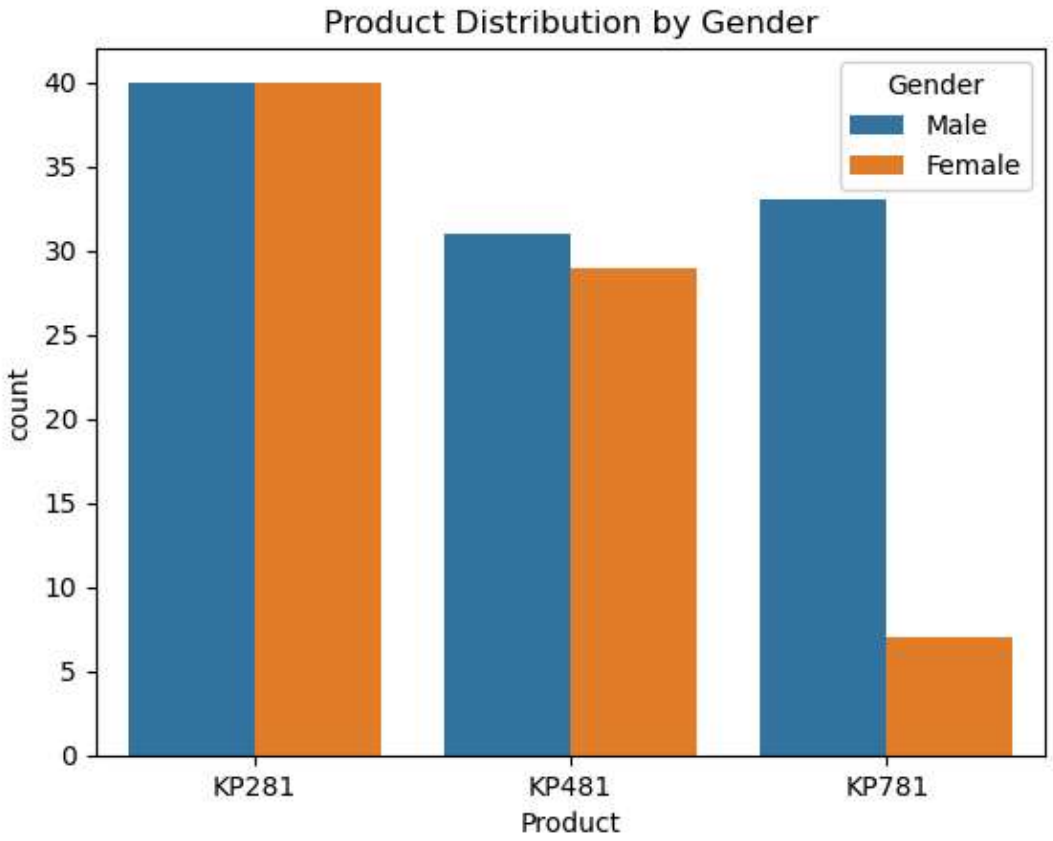
In [73]:
```
# 5. Histogram for Fitness
plt.figure(figsize=(8, 4))
sns.histplot(df['Fitness'], kde=True, color='orange')
plt.title('Histogram of Fitness')
plt.xlabel('Fitness Level (1-5)')
plt.ylabel('Frequency')
plt.show()
```



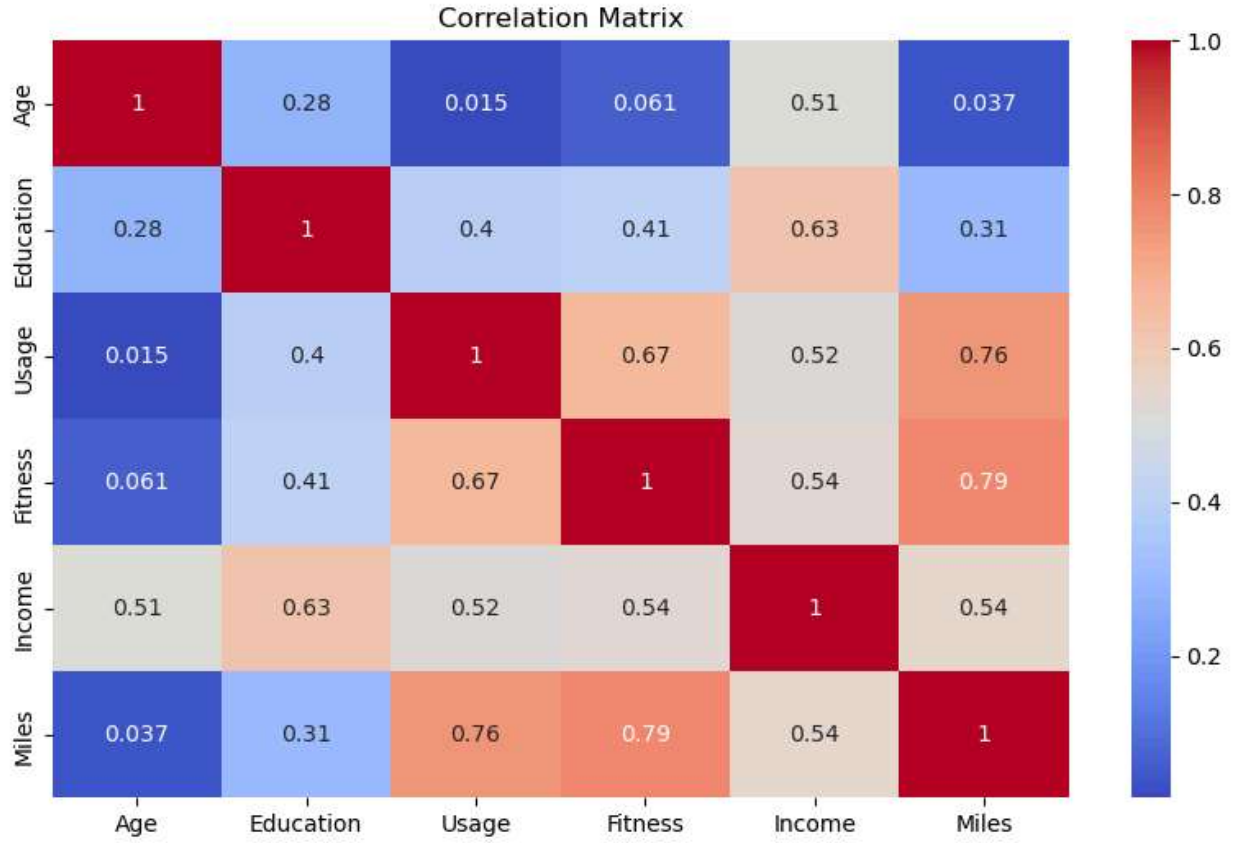
In [74]:
```
# 6. Histogram for Miles
plt.figure(figsize=(8, 4))
sns.histplot(df['Miles'], kde=True, color='teal')
plt.title('Histogram of Miles')
plt.xlabel('Miles per week')
plt.ylabel('Frequency')
plt.show()
```



In [65]:
```
# Compare product distribution by gender
sns.countplot(x='Product', hue='Gender', data=df)
plt.title('Product Distribution by Gender')
plt.show()
```



In [76]:
```
# Create a heatmap to see correlations between numerical features
corr = df.select_dtypes(include='number').corr()
plt.figure(figsize=(10, 6))
sns.heatmap(corr, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```



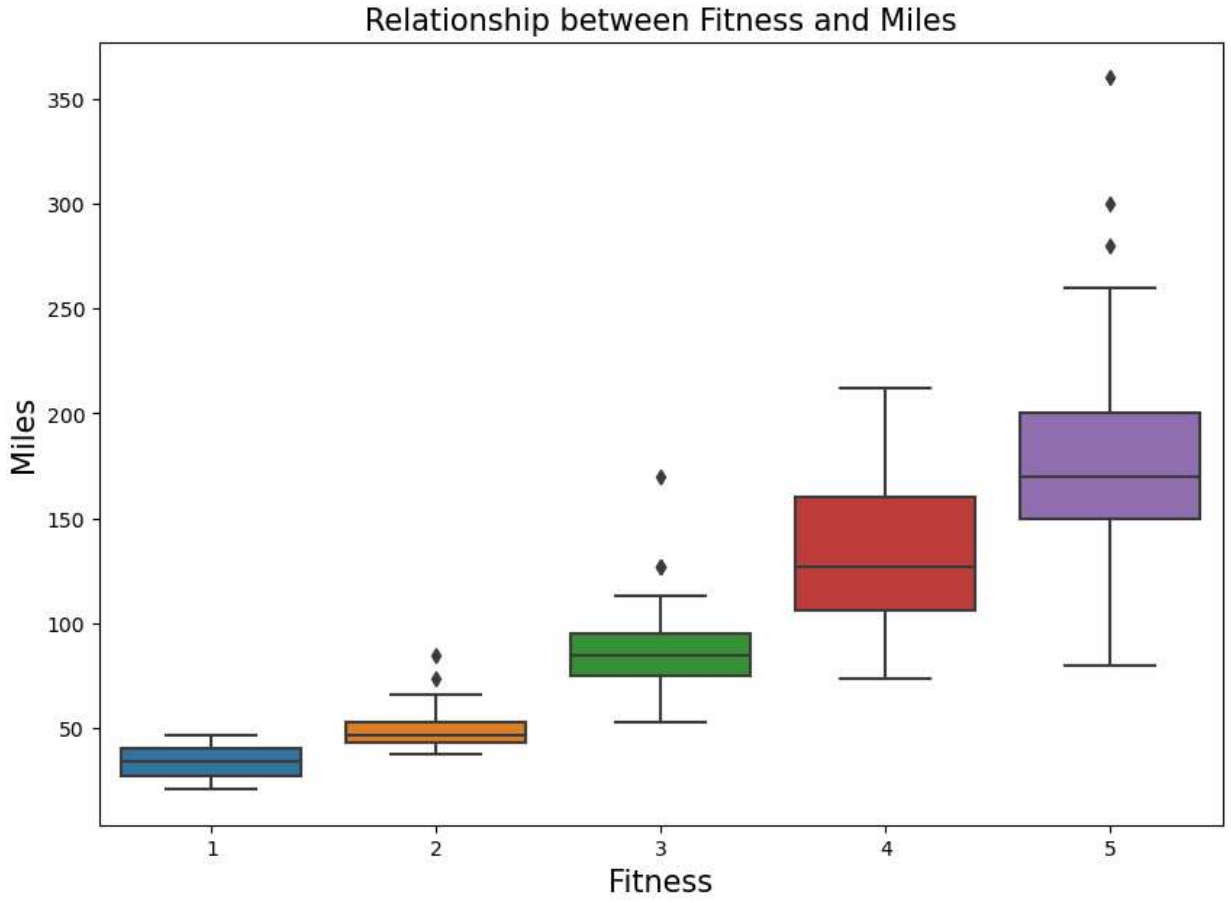
In [90]:
```
# Calculate the conditional probability of gender given the product
product_gender = pd.crosstab(df['Product'], df['Gender'], normalize='index')
print(product_gender)
```

```
Gender     Female      Male
Product
KP281    0.500000  0.500000
KP481    0.483333  0.516667
KP781    0.175000  0.825000
```

In [91]:
```
# Get the average characteristics of customers for each product
profile = df.groupby('Product')[['Age', 'Income', 'Usage', 'Fitness', 'Miles']].mean(
print(profile)
```

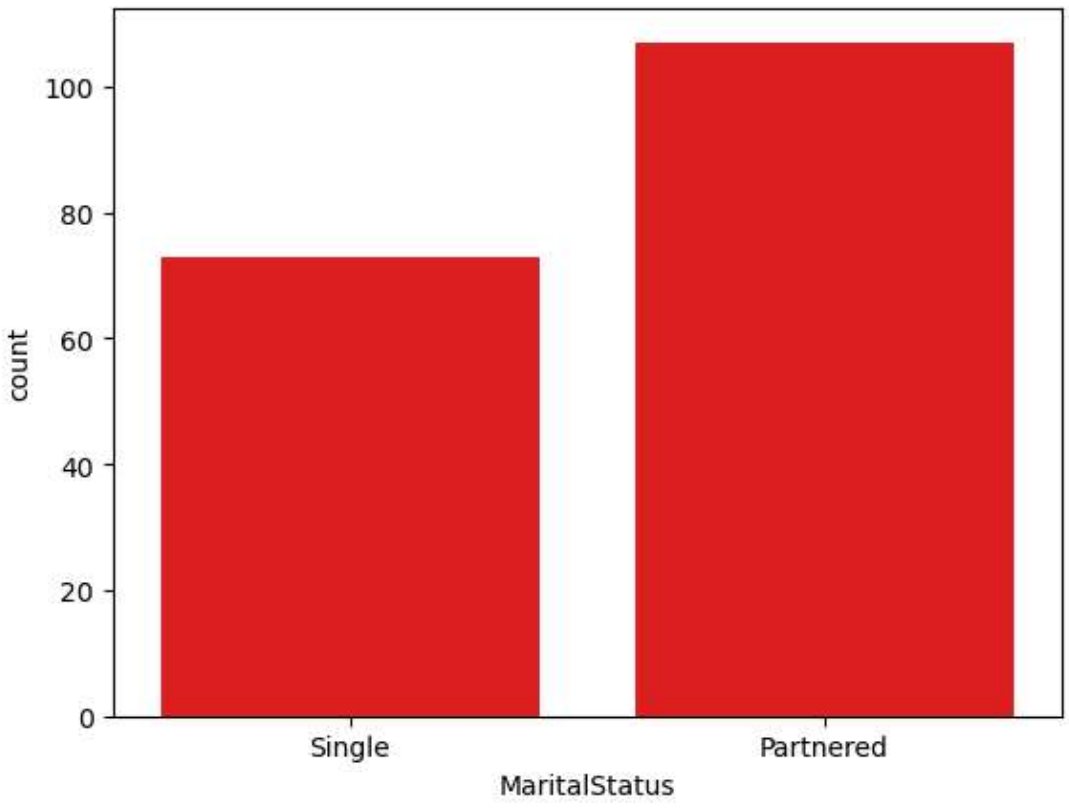
```
             Age     Income     Usage  Fitness       Miles
Product
KP281    28.55  46418.025  3.087500   2.9625   82.787500
KP481    28.90  48973.650  3.066667   2.9000   87.933333
KP781    29.10  75441.575  4.775000   4.6250  166.900000
```

In [84]:
```
fig = plt.figure(figsize =(10, 7))
sns.boxplot(data=df, x='Fitness', y='Miles')
plt.xlabel('Fitness', fontsize=15)
plt.ylabel('Miles', fontsize=15)
plt.title('Relationship between Fitness and Miles', fontsize=15)
plt.show()
```

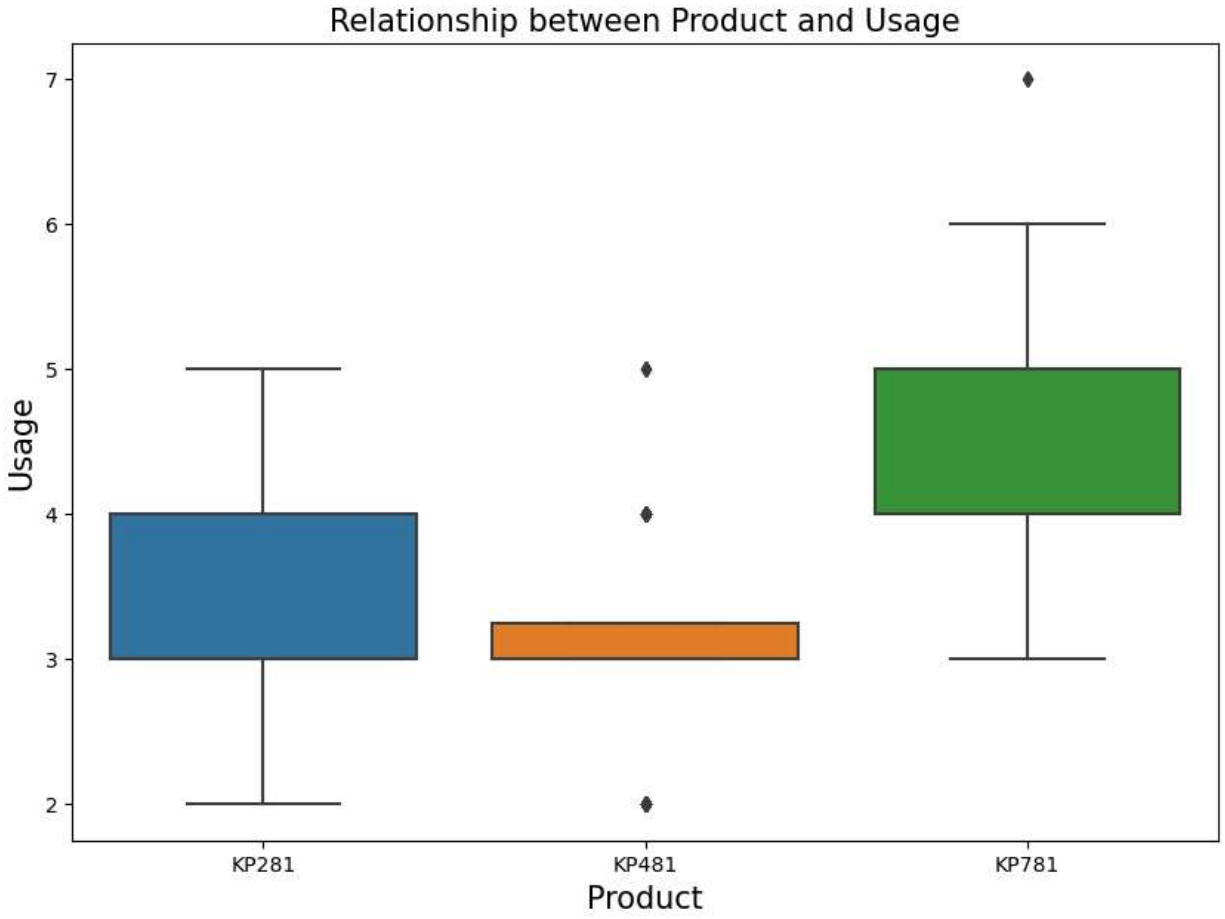


As it is clearly seen from the chart that the person who walks more comparitively much fitter than the person who walks less

In [85]:
```
sns.countplot(data=df, x='MaritalStatus',color='red')
plt.show()
```



In [87]:
```
fig = plt.figure(figsize =(10, 7))
sns.boxplot(df,x='Product',y='Usage')
plt.xlabel('Product',fontsize=15)
plt.ylabel('Usage',fontsize=15)
plt.title('Relationship between Product and Usage', fontsize=15)
plt.show()
```



The product KP481 have very less usage so there may be a reason that the product quality is not not so good.

#Insights:

1. The majority of customers belong to the **25-30 age group**, making it the most significant customer segment. However, there is also potential to attract more customers from the **40-50 age group**.
2. Most customers are from the **partnered group**, and within both the **Partnered** and **Single** categories, the **KP281 treadmill** is the most popular, followed by **KP481**. This trend could be due to **KP281's lower price**, while **KP781**, being the most expensive, has fewer buyers.
3. Customers with **higher incomes** tend to rate themselves as more fit, often giving a **fitness score of 5**.
4. The majority of customers rate their fitness level between **3 and 3.5**, indicating a moderate fitness level among users.
5. **Partnered individuals** generally have **higher incomes** compared to **single individuals**.
6. **KP481** shows relatively **low usage**, which might indicate **quality issues** or **lack of satisfaction** among customers.

#Recommendations:

1. Focus marketing efforts on the **25-30 age group** as they make up the majority of customers, while also exploring ways to attract the **40-50 age group**.
2. Continue promoting **KP281** as the most popular and affordable model, but also highlight the **value and features of KP481** to improve its usage.
3. Target **higher-income customers** with premium models like **KP781**, emphasizing their advanced features and suitability for fitness enthusiasts.
4. Address potential **quality issues with KP481** by gathering customer feedback and making necessary improvements to enhance satisfaction.
5. Create tailored campaigns for **partnered individuals**, as they are more likely to make purchases and tend to have **higher incomes**.
6. Consider offering **fitness improvement plans** for customers with moderate fitness levels (3-3.5) to encourage treadmill usage and build brand loyalty.