DEFINING PROBLEM STATEMENT & ANALYZING BASIC METRICS.

Aerofit stands as a prominent figure in the realm of fitness gear, offering a comprehensive selection of products tailored to suit various fitness requirements. Their diverse range encompasses essentials like treadmills, exercise bikes, gym apparatus, and an array of fitness accessories. Whether you're an avid gym-goer, a seasoned athlete, or someone just beginning their fitness journey, Aerofit ensures there's something for everyone, catering to a wide spectrum of individuals with varying fitness goals and preferences. With a commitment to quality and innovation, Aerofit remains dedicated to empowering individuals on their path to better health and wellness.

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

In [2]:
```
df = pd.read_csv('aerofit.csv')
```

In [3]:
```
df.head(5)
```

Out[3]:

	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 [4]:
```
# Data types of all attributes and the number of null and non-null values
print("Data types of all the attributes are:\n")
df.info()
```

```
Data types of all the attributes are:

<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 [5]:
```
# Shape of data.
shape = df.shape
print(f'Shape of the data is: {shape}')
```

```
Shape of the data is: (180, 9)
```

In [6]:
```
df['Product'].value_counts()
```

Out[6]:
```
Product
KP281    80
KP481    60
KP781    40
Name: count, dtype: int64
```

In [7]:
```
# Statistical summary for the dataset.
df.describe()
```

Out[7]:

	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

NON GRAPHICAL ANALYSIS: VALUE COUNTS AND UNIQUE ATTRIBUTES.

In [8]:
```
# Number of unique values for each column.
df.nunique()
```

Out[8]:
```
Product           3
Age              32
Gender            2
Education         8
MaritalStatus     2
Usage             6
Fitness           5
Income           62
Miles            37
dtype: int64
```

In [9]:
```
# Most occuring age value.
df['Age'].value_counts().head(10)
```

Out[9]:
```
Age
25    25
23    18
24    12
26    12
28     9
35     8
33     8
30     7
38     7
21     7
Name: count, dtype: int64
```

In [10]:
```
# Most occuring income value.
df['Income'].value_counts().head(10)
```

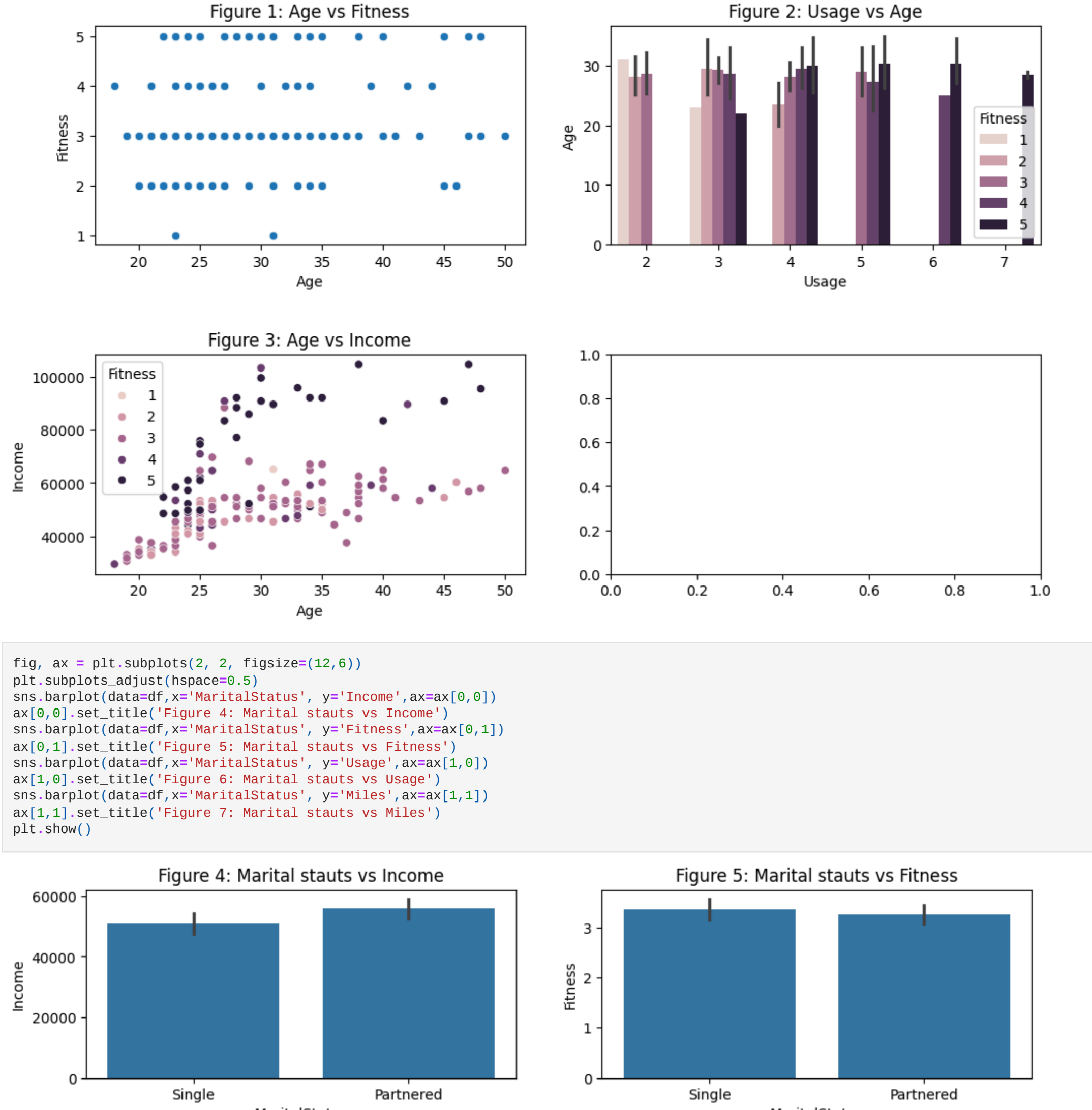
Out[10]:
```
Income
45480    14
52302     9
46617     8
54576     8
53439     8
50028     7
51165     7
40932     6
48891     5
32973     5
Name: count, dtype: int64
```

In [11]:
```
# Most occuring fitness(type of body shape) value.
df['Fitness'].value_counts()
```

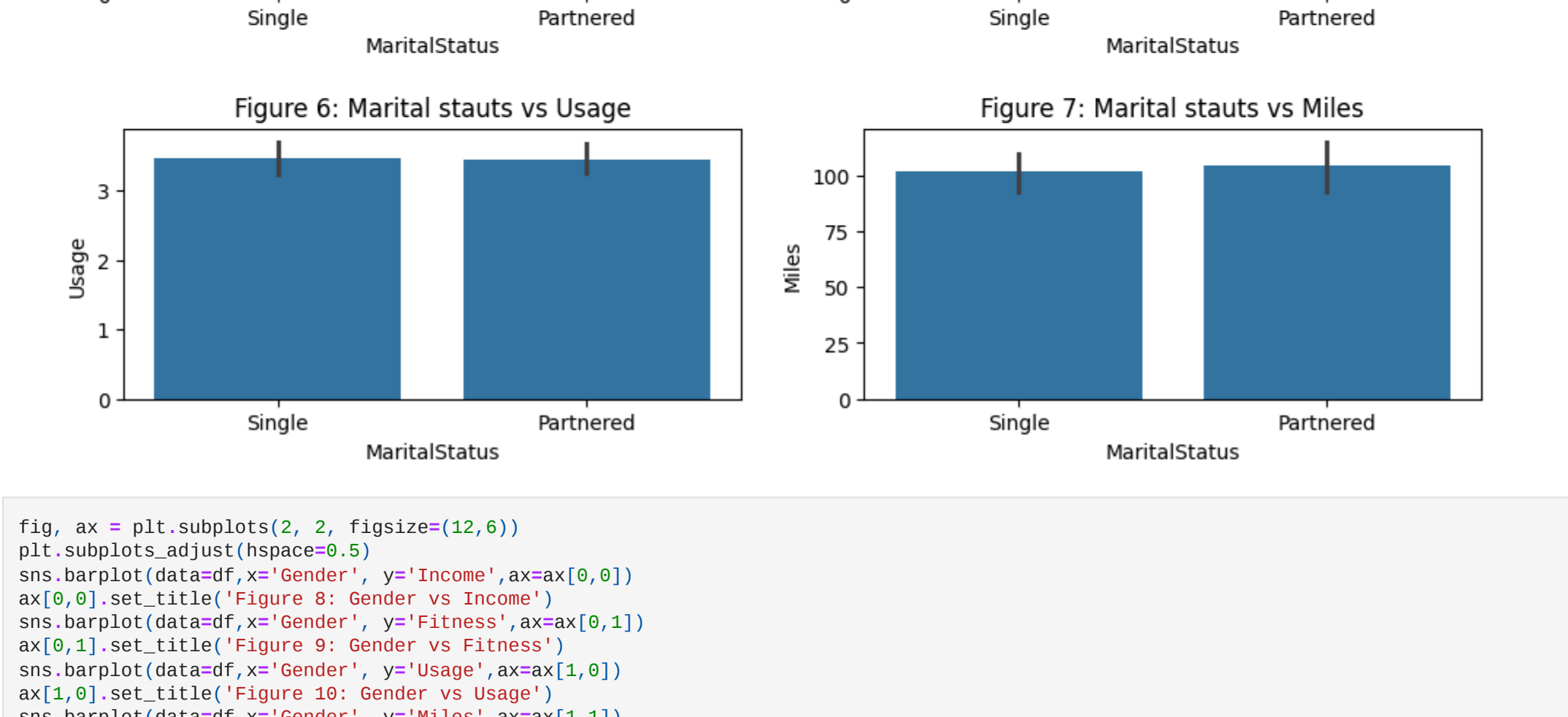
Out[11]:
```
Fitness
3    97
5    31
2    26
4    24
1     2
Name: count, dtype: int64
```

VISUAL ANALYSIS - UNIVARIATE AND BIVARIATE.

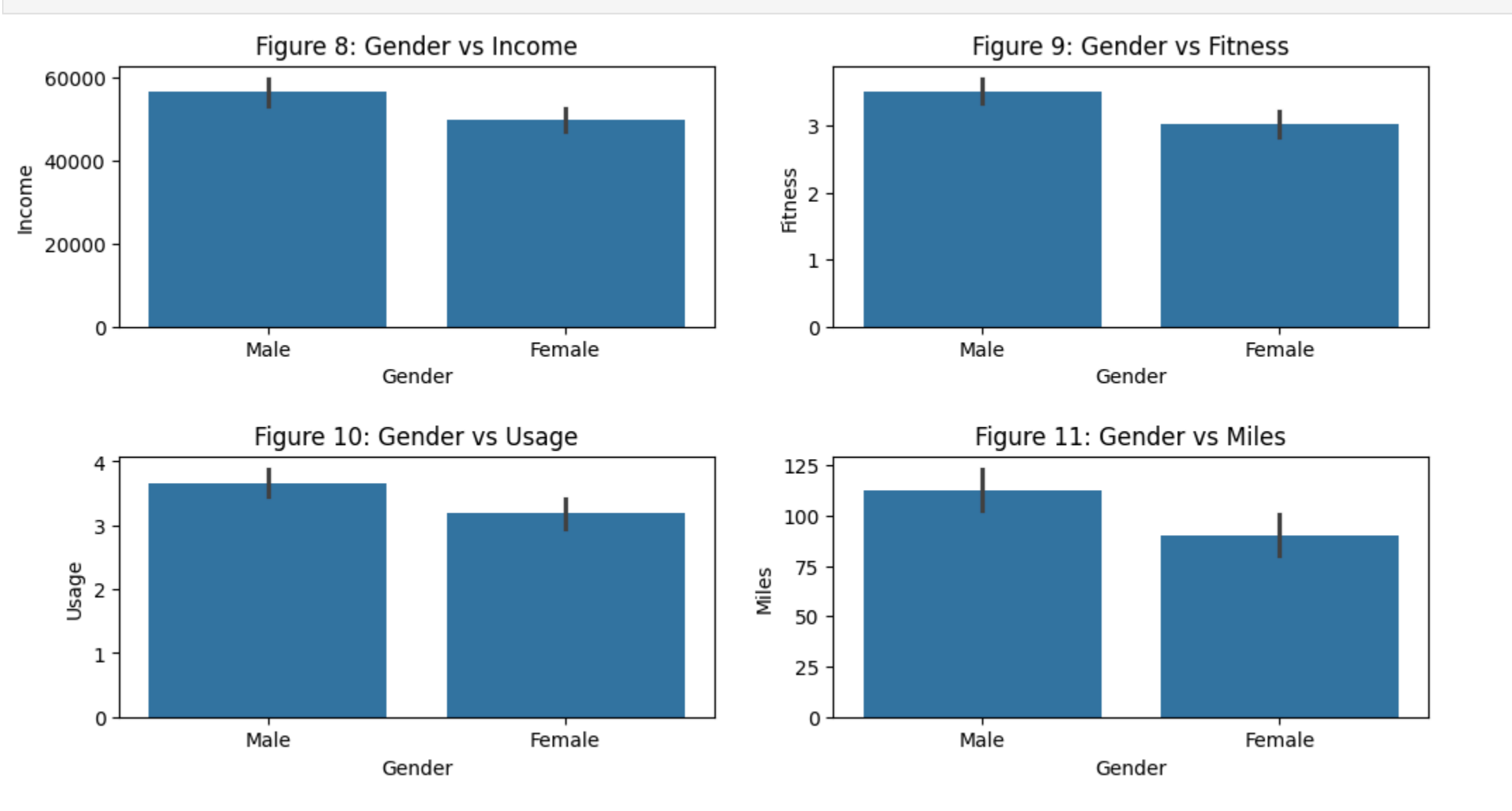
In [12]:
```
fig, ax = plt.subplots(2, 2, figsize=(12,7))
plt.subplots_adjust(hspace=0.5)
sns.scatterplot(data=df,x='Age',y='Fitness',ax=ax[0,0])
ax[0,0].set_title('Figure 1: Age vs Fitness')
sns.barplot(data=df,x='Usage',y='Age',hue='Fitness',ax=ax[0,1])
ax[0,1].set_title('Figure 2: Usage vs Age')
sns.scatterplot(data=df,x='Age',y='Income',hue='Fitness',ax=ax[1,0])
ax[1,0].set_title('Figure 3: Age vs Income')
plt.show()
```



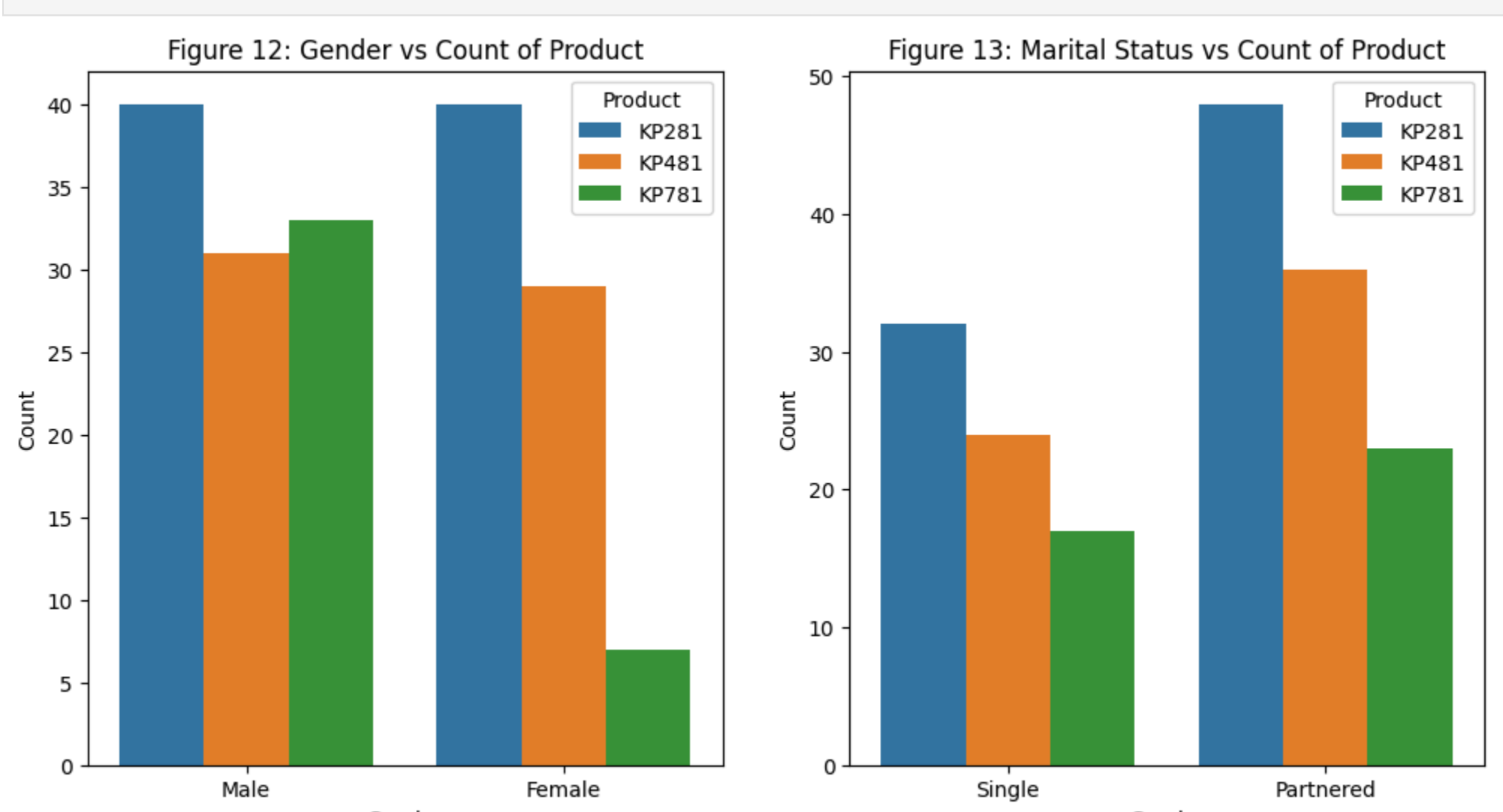
In [13]:
```
fig, ax = plt.subplots(2, 2, figsize=(12,6))
plt.subplots_adjust(hspace=0.5)
sns.barplot(data=df,x='MaritalStatus', y='Income',ax=ax[0,0])
ax[0,0].set_title('Figure 4: Marital stauts vs Income')
sns.barplot(data=df,x='MaritalStatus', y='Fitness',ax=ax[0,1])
ax[0,1].set_title('Figure 5: Marital stauts vs Fitness')
sns.barplot(data=df,x='MaritalStatus', y='Usage',ax=ax[1,0])
ax[1,0].set_title('Figure 6: Marital stauts vs Usage')
sns.barplot(data=df,x='MaritalStatus', y='Miles',ax=ax[1,1])
ax[1,1].set_title('Figure 7: Marital stauts vs Miles')
plt.show()
```



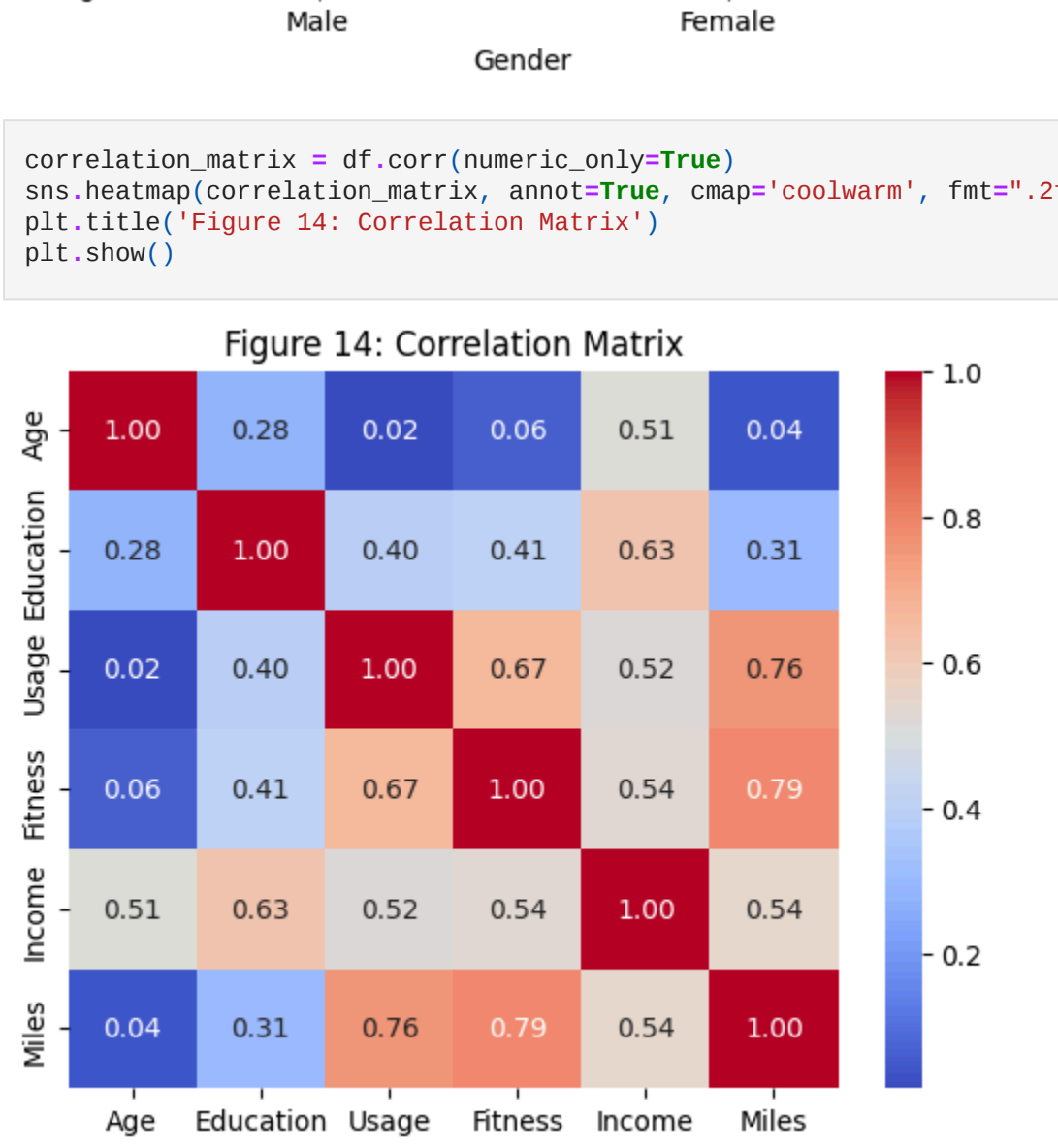
In [14]:
```
fig, ax = plt.subplots(2, 2, figsize=(12,6))
plt.subplots_adjust(hspace=0.5)
sns.barplot(data=df,x='Gender', y='Income',ax=ax[0,0])
ax[0,0].set_title('Figure 8: Gender vs Income')
sns.barplot(data=df,x='Gender', y='Fitness',ax=ax[0,1])
ax[0,1].set_title('Figure 9: Gender vs Fitness')
sns.barplot(data=df,x='Gender', y='Usage',ax=ax[1,0])
ax[1,0].set_title('Figure 10: Gender vs Usage')
sns.barplot(data=df,x='Gender', y='Miles',ax=ax[1,1])
ax[1,1].set_title('Figure 11: Gender vs Miles')
plt.show()
```



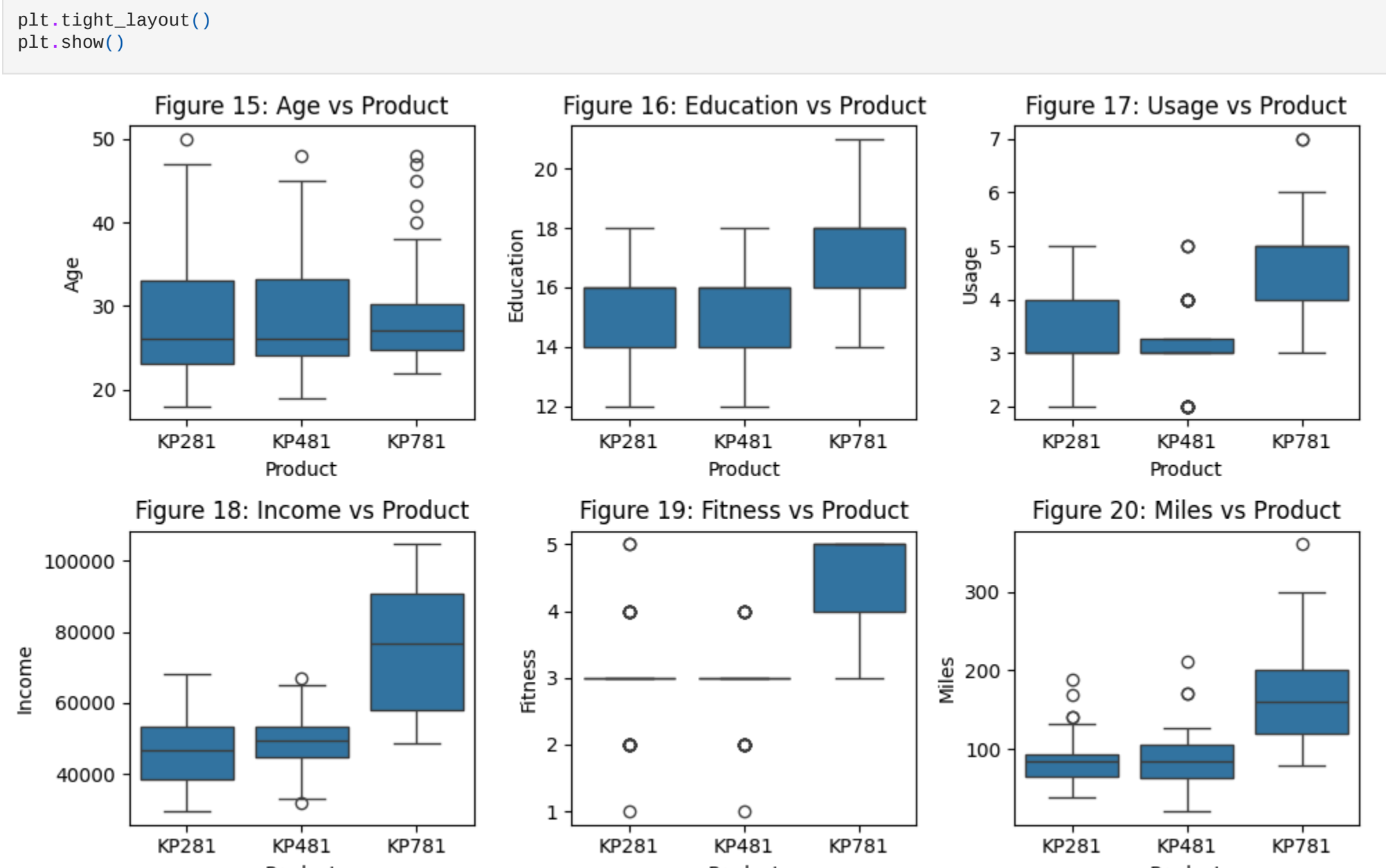
In [15]:
```
fig, ax = plt.subplots(1, 2, figsize=(12,6))
plt.subplots_adjust(hspace=0.5)
sns.countplot(data=df, x='Gender', hue='Product',ax=ax[0])
ax[0].set_xlabel('Gender')
ax[0].set_ylabel('Count')
ax[0].set_title('Figure 12: Gender vs Count of Product')
sns.countplot(data=df, x='MaritalStatus', hue='Product',ax=ax[1])
ax[1].set_xlabel('Gender')
ax[1].set_ylabel('Count')
ax[1].set_title('Figure 13: Marital Status vs Count of Product')
plt.show()
```



In [16]:
```
correlation_matrix = df.corr(numeric_only=True)
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title('Figure 14: Correlation Matrix')
plt.show()
```

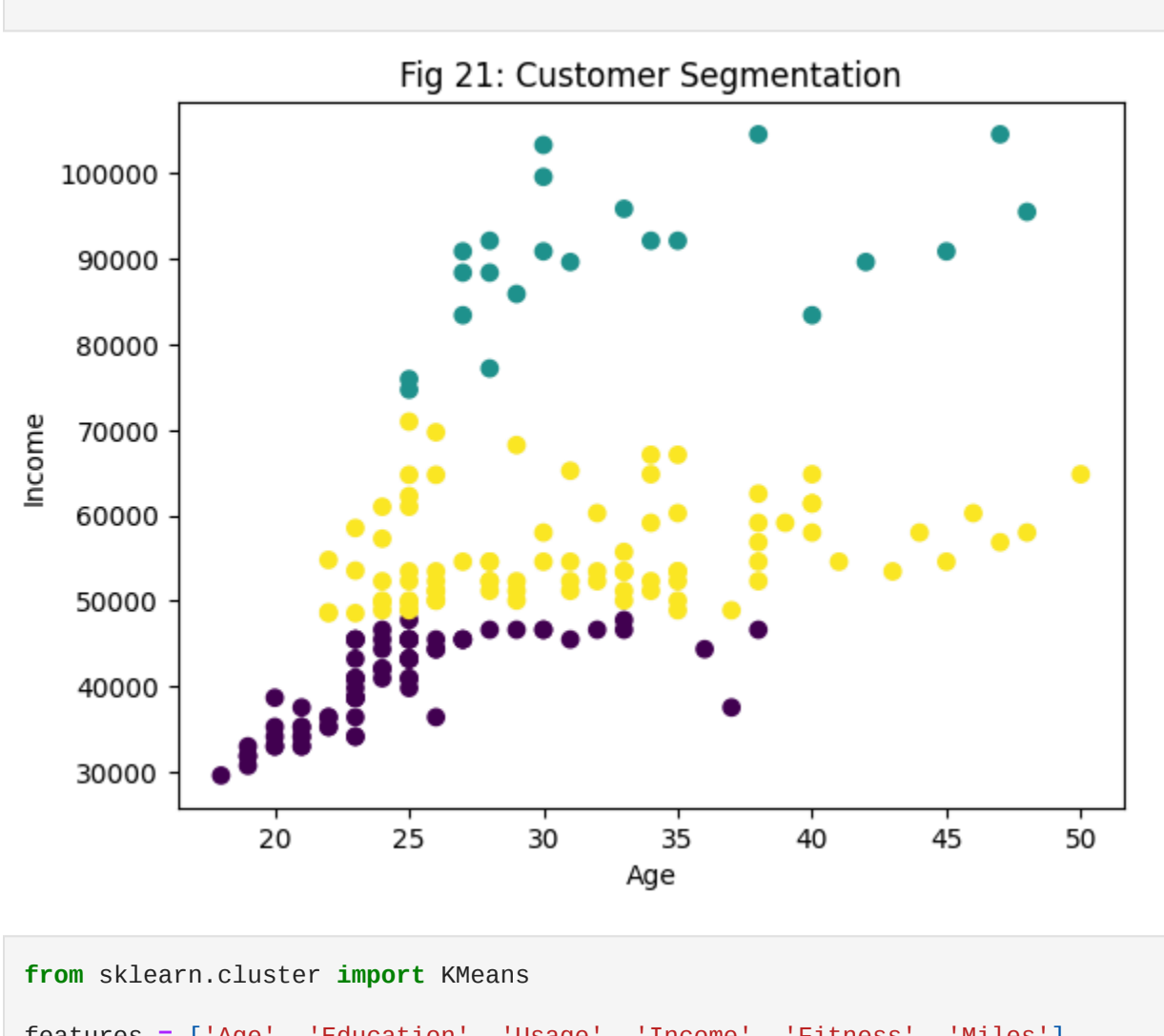


In [17]:
```
plt.figure(figsize=(10, 6))

# Age vs Product Purchased
plt.subplot(2, 3, 1)
sns.boxplot(data=df, x='Product', y='Age')
plt.title('Figure 15: Age vs Product')

# Education vs Product Purchased
plt.subplot(2, 3, 2)
sns.boxplot(data=df, x='Product', y='Education')
plt.title('Figure 16: Education vs Product')

# Usage vs Product Purchased
plt.subplot(2, 3, 3)
sns.boxplot(data=df, x='Product', y='Usage')
plt.title('Figure 17: Usage vs Product')

# Income vs Product Purchased
plt.subplot(2, 3, 4)
sns.boxplot(data=df, x='Product', y='Income')
plt.title('Figure 18: Income vs Product')

# Fitness vs Product Purchased
plt.subplot(2, 3, 5)
sns.boxplot(data=df, x='Product', y='Fitness')
plt.title('Figure 19: Fitness vs Product')

# Miles vs Product Purchased
plt.subplot(2, 3, 6)
sns.boxplot(data=df, x='Product', y='Miles')
plt.title('Figure 20: Miles vs Product')

plt.tight_layout()
plt.show()
```



CUSTOMER PROFILING

In [18]:
```
from sklearn.cluster import KMeans

features = ['Age', 'Education', 'Usage', 'Income', 'Fitness', 'Miles']

# Perform K-means clustering
kmeans = KMeans(n_clusters=3, random_state=42, n_init = 10)
df['Cluster'] = kmeans.fit_predict(df[features])

# Visualize the clusters (using only two features for simplicity)
plt.scatter(df['Age'], df['Income'], c=df['Cluster'], cmap='viridis')
plt.xlabel('Age')
plt.ylabel('Income')
plt.title('Fig 21: Customer Segmentation')
plt.show()
```



In [19]:
```
from sklearn.cluster import KMeans

features = ['Age', 'Education', 'Usage', 'Income', 'Fitness', 'Miles']

# Perform K-means clustering
kmeans = KMeans(n_clusters=3, random_state=42, n_init = 10)
df['Cluster'] = kmeans.fit_predict(df[features])

# Visualize the clusters (using only two features for simplicity)
plt.scatter(df['Product'], df['Income'], c=df['Cluster'], cmap='viridis')
plt.xlabel('Product')
plt.ylabel('Income')
plt.title('Fig 22: Customer Segmentation')
plt.show()
```

MISSING VALUES AND OUTLIER DETECTION.

In [20]:
```
# Missing values.
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 10 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
 9   Cluster        180 non-null    int32
dtypes: int32(1), int64(6), object(3)
memory usage: 13.5+ KB
```

There are no missing values in any column and the data types of all the columns are correct and hence we have a clean dataset by default

PROBABILITY

In [21]:
```
contingency_table = pd.crosstab(index=df['Product'], columns=df['Gender'])

# Display the contingency table
print("Two-Way Contingency Table:")
print(contingency_table)

# Compute conditional probabilities
conditional_probabilities = contingency_table.div(contingency_table.sum(axis=1), axis=0)

# Display conditional probabilities
print("\nConditional Probabilities:")
print(conditional_probabilities)

# Compute marginal probabilities
marginal_probabilities = contingency_table.div(contingency_table.sum().sum())

# Display marginal probabilities
print("\nMarginal Probabilities:")
print(marginal_probabilities)
```

```
Two-Way Contingency Table:
Gender   Female  Male
Product
KP281        40    40
KP481        29    31
KP781         7    33

Conditional Probabilities:
Gender     Female      Male
Product
KP281    0.500000  0.500000
KP481    0.483333  0.516667
KP781    0.175000  0.825000

Marginal Probabilities:
Gender     Female      Male
Product
KP281    0.222222  0.222222
KP481    0.161111  0.172222
KP781    0.038889  0.183333
```

The contingency table tells us that KP281 has equal share of customers from both genders, whereas for KP481 the share of men is infinitesimally higher and for KP781 the share of men is significantly higher.

The conditional probability of women buying KP281, KP481 and KP781 is 0.50, 0.48 and 0.17 respectively and that of men is 0.50, 0.52 and 0.83.

BUSINESS INSIGHTS

RANGE OF ATTRIBUTES

The dataset provided encompasses a diverse range of attributes capturing various facets of customer behavior and demographics. It includes demographic information such as age, gender, education, marital status, and income, offering insights into the socioeconomic backgrounds of customers. Additionally, behavioral attributes such as usage frequency, fitness level, and intended mileage per week provide valuable indicators of customer engagement with treadmill products. The dataset further delves into purchasing patterns through the 'Product Purchased' column, shedding light on customer preferences across different treadmill models. With a comprehensive set of attributes covering demographic, behavioral, and purchasing dimensions, the dataset offers a rich source of information for understanding customer profiles and tailoring marketing strategies to meet their needs and preferences

DISTRIBUTION OF VARIABLES AND RELATIONSHIP BETWEEN THEM

The dataset exhibits a varied distribution of variables, reflecting the diversity of customer characteristics and behaviors. Age, income, and education appear to follow typical distributions observed in demographic data, with age likely exhibiting a relatively normal distribution, while income and education may skew towards higher values due to the presence of outliers or higher-income individuals. Marital status and gender are categorical variables, likely showing a relatively balanced distribution between categories.

Regarding the relationship between variables, several interesting patterns emerge. Income positively correlates with education level, as individuals with higher education often command higher incomes. Similarly, age and income exhibit a positive correlation, with older individuals typically having higher incomes due to career progression. Fitness level correlates positively with age, as younger individuals may prioritize fitness more than older individuals. Usage frequency could correlate positively with income, as individuals with higher disposable incomes may afford more frequent usage of treadmill products. Additionally, marital status is neutral with usage frequency, as married individuals have different lifestyle priorities compared to single individuals but mindset towards fitness is same. Overall, exploring the relationships between these variables can provide valuable insights into customer behavior and preferences, enabling targeted marketing strategies and product offerings.

COMMENTS FOR EACH UNIVARIATE AND BIVARIATE PLOT

Figure 1: Indicates that individuals aged 20-42 show greater involvement in fitness activities, with most falling within the decent to perfect fitness range.

Figure 2: Reveals a correlation between fitness level and intended treadmill usage, where individuals with higher fitness intend to use the treadmill more frequently.

Figure 3: Demonstrates that younger individuals with varying income levels exhibit more active engagement in fitness, while older individuals with higher income levels demonstrate greater dedication to fitness.

Figure 4: Illustrates that married individuals tend to earn slightly more than single individuals.

Figure 5: Suggests that single individuals show a higher inclination towards fitness compared to married individuals.

Figure 6: Indicates consistent usage intentions across marital statuses, with individuals planning similar usage regardless of marital status.

Figure 7: Shows that married individuals commit to walking longer distances per week compared to single individuals.

Figure 8: Highlights a gender disparity in income, with men earning more than women.

Figure 9: Indicates that men tend to be in better physical shape than women.

Figure 10: Suggests that men plan to use the product more frequently than women.

Figure 11: Demonstrates that men tend to walk/run more than women.

Figure 12: Displays a gender preference in product purchases, with the higher-end product (KP781) being more popular among men.

Figure 13: Indicates that partnered individuals contribute more to revenue compared to single individuals.

Figure 14: Shows the correlation matrix depicting relationships between numerical attributes.

Figure 15: Reveals differences in median age across treadmill products, with KP781 attracting slightly older customers.

Figure 16: Indicates that more educated individuals tend to purchase higher-end products.

Figure 17: Suggests a correlation between usage frequency and product preference, with lower-end products favored by those using the treadmill 3-4 times a week, and higher-end products preferred by those using it more frequently.

Figure 18: Indicates that KP781 is preferred by individuals with higher income levels.

Figure 19: Shows a correlation between self-reported fitness level and product preference, with lower-end products preferred by those rating themselves as decently fit, and the higher-end product preferred by those rating themselves as perfectly fit.

Figure 20: Demonstrates a relationship between planned miles and product range, where individuals planning to cover more miles tend to prefer higher-end products.

Figure 21: Illustrates customer segmentation by age and income, indicating an increase in income with age, with individuals aged 27 and above falling into the medium to higher income brackets.

Figure 22: Shows a correlation between income levels and product preference, with KP281 preferred by those earning between 30k - 50k, KP481 preferred by those earning between 50k-70k, and KP781 preferred by those earning higher than 70k.

RECOMMENDATIONS

Based on the insights gleaned from the figures, attributes, and relationships within the dataset, here are some actionable recommendations for Aerofit:

Targeted Marketing Campaigns:

Tailor marketing campaigns to different age groups, with a focus on promoting fitness activities among individuals aged 20-42 who exhibit higher engagement levels. Highlight the benefits of treadmill usage for individuals across different fitness levels, emphasizing the versatility and effectiveness of Aerofit products in catering to varying fitness needs.

Product Development and Pricing:

Consider developing specialized features or packages targeting specific age groups or income brackets to align with their preferences and purchasing power. Offer competitive pricing strategies for different product models based on customer segments' income levels and perceived value, ensuring affordability while maintaining profitability.

Customer Segmentation:

Segment customers based on attributes such as age, income, and fitness level, to better understand their needs and preferences. Customize product offerings, promotions, and support services to cater to the unique requirements of each customer segment, enhancing customer satisfaction and loyalty.

Promotion of Fitness Programs:

Launch fitness programs or challenges targeting individuals with higher fitness levels who intend to use the treadmill more frequently, encouraging continuous engagement and goal achievement. Collaborate with fitness influencers or experts to endorse Aerofit products and promote a healthy lifestyle, leveraging their credibility and reach to attract new customers.

Customer Experience Enhancement:

Offer personalized recommendations or incentives based on customer profiles and purchase history to enhance the overall shopping experience and foster repeat purchases. Provide educational resources or workshops targeting younger, more educated customers interested in fitness, offering valuable insights and tips for achieving their fitness goals effectively.

Market Expansion and Partnership:

Explore partnerships with fitness centers, wellness clubs, or corporate wellness programs to expand Aerofit's reach and tap into new customer segments. Leverage customer data and insights to identify untapped market opportunities and develop targeted strategies for market expansion and penetration. By implementing these actionable insights, Aerofit can strengthen its competitive position, drive customer engagement and satisfaction, and capitalize on emerging market trends to achieve sustainable growth and success in the fitness industry.