Cardio Good Fitness Project

The data is about customers of the treadmill product(s) of a retail store called Cardio Good Fitness. It contains data related to customers who have purchased different product from Cardio Good Fitness. We are going to explore the dataset to come up with a customer profile (characteristics of a customer) of the different products. Perform univariate and multivariate analyses. Generate a set of insights and recommendations that will help the company in targeting new customers.

Dataset Informations:

CardioGoodFitness.csv: It contains informations of customers of the treadmill product(s) of a retail store called Cardio Good Fitness. It contains the following columns: Product, Age, Gender, Education, Marital Status, Usage, Fitness, Income, Miles

Objectives:

- Overview of the dataset shape, datatypes - Statistical summary and check for missing values
- Analysis of spread and distribution of every feature in the dataset.
- Analysis of interaction between features, in the dataset using univariate and multivariate analyses method
- Conclude with the key insights/observations

Learning Outcomes:

```
Learning about fundamentals of AIML
```

Domain

```
marketing and manufacturing domain
```

Understanding the structure of the data

1. Import the necessary packages

In [3]:

```python
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
```

2. Read the dataset

In [4]:

```python
# Reading dataset by using read_csv from pandas package
```

```
cgfitness = pd.read_csv("CardioGoodFitness.csv")
```

3. Understand the shape of the dataset

In [7]:

```
# Analysing the data by looking at the first 5 rows of the data using Head funct

cgfitness.head()
```

Out[7]:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	TM195	18	Male	14	Single	3	4	29562	112
1	TM195	19	Male	15	Single	2	3	31836	75
2	TM195	19	Female	14	Partnered	4	3	30699	66
3	TM195	19	Male	12	Single	3	3	32973	85
4	TM195	20	Male	13	Partnered	4	2	35247	47

In [8]:

```
# Analysing the data by looking at the last 5 rows of the data using tail functi

cgfitness.tail()
```

Out[8]:

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
175	TM798	40	Male	21	Single	6	5	83416	200
176	TM798	42	Male	18	Single	5	4	89641	200
177	TM798	45	Male	16	Single	5	5	90886	160
178	TM798	47	Male	18	Partnered	4	5	104581	120
179	TM798	48	Male	18	Partnered	4	5	95508	180

In [9]:

```
# Finding the shape of the dataset using dataframe.shape

cgfitness.shape
```

Out[9]:

```
(180, 9)
```

Observation: There are 180 rows and 9 columns in the cgfitness dataset

4. Check the data types of the columns for the dataset.

In [5]:

```
# We use dataframe.dtypes to get the data types of each column
fitness.dtypes
```

Out[5]:

```
Product          object
Age               int64
Gender           object
Education         int64
MaritalStatus    object
Usage             int64
Fitness           int64
```

```
Income           int64
Miles            int64
dtype: object
```

Observation:

- Age, Education, Usage, Fitness, Income and Miles columns are of integer data types.
- Product, Gender and MaritalStatus columns are of string data type.

5. Check the statistical summary for the dataset.

In [6]:

```
# We use dataframe.describe to get the statistical summary of each column
fitness.describe()
```

Out[6]:

	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

Observation:

- The Mean, Median and Standard Deviation of Ages are 28.78, 26 & 6.94 respectively
- The Mean, Median and Standard Deviation of Educations are 15.57, 16 & 1.617 respectively
- The Mean, Median and Standard Deviation of Usage are 3.45, 3 & 1.08 respectively
- The Mean, Median and Standard Deviation of Fitness are 3.31, 3 & 0.95 respectively
- The Mean, Median and Standard Deviation of Income are 53719.57, 50596.5 & 16506.68 respectively
- The Mean, Median and Standard Deviation of Miles are 103.19, 94 & 51.86 respectively

6. Check for missing values

In [12]:

```
# We use dataframe.isnull().sum() to get the missing values and sum of the total
cgfitness.isnull().sum()
```

Out[12]:

```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
```

```
Miles            0
dtype: int64
```

Observation: There are no missing values in the fitness dataset

Data Visualization - Univariate Data Analysis

Univariate analysis refer to the analysis of a single variable. The main purpose of univariate analysis is to summarize and find patterns in the data. The key point is that there is only one variable involved in the analysis.

Let us take the fitness dataset and work on that for the univariate analysis.

Analysis of spread and distribution of every feature in the dataset

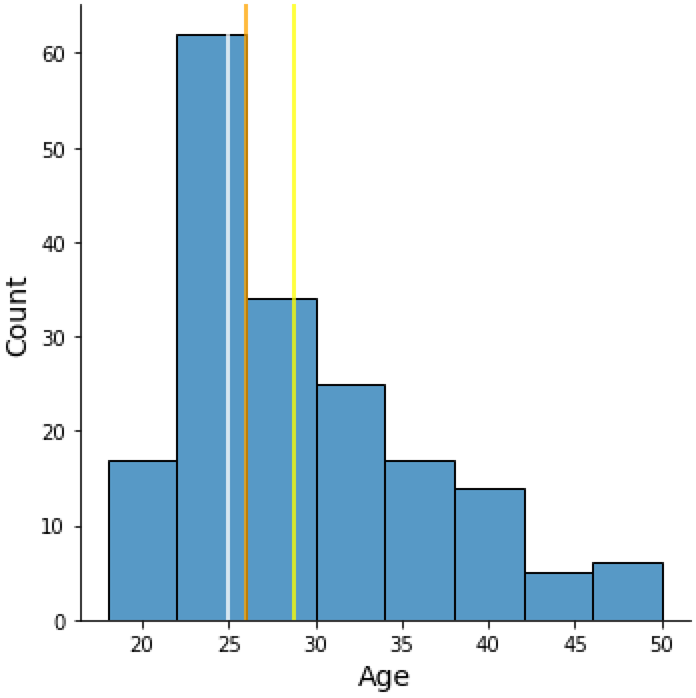
In [19]:

```
# plots a histogram plt using the seaborn package for Age column.
# Using displot since distplot going to be decommissioned in the future

sns.displot(cgfitness,
            x = "Age",
            bins=8,
            height=5)
plt.xlabel("Age", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Age.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Age.median(),
            color='orange')
plt.axvline(x=cgfitness.Age.mode()[0],
            color='white')
```

Out[19]:

```
<matplotlib.lines.Line2D at 0x7fa241d5bf10>
```



Observations:

- In this case we are analysing age column. We can safely say based on visual observation that Age is skewed towards right side most of the count present on the data are less than 30 ages
- In the above histograms we can see that a bulk of the observations lie within the first six classes. The rest of the five classes contain only a very few observations.
- From the above figure we can see that mean is represented by the yellow line and the mode by the white line . The median is represented by the orange line.
- We can see from the above figure that the mode and the median are very close to each other and that the mean is higher than both.
- There are a very few counts that are more than 30 ages. Once we pass the 30 ages point the number of observations drops further.

Now we have an idea of how the data is distributed for age.

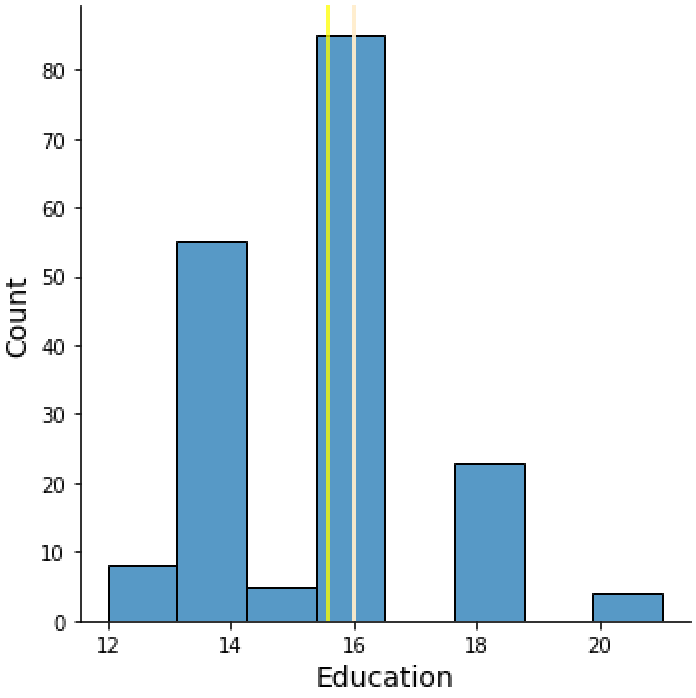
In [20]:

```
# plots a histogram plt using the seaborn package for Education column.

sns.displot(cgfitness,
            x = "Education",
            bins=8,
            height=5)
plt.xlabel("Education", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Education.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Education.median(),
            color='orange')
plt.axvline(x=cgfitness.Education.mode()[0],
            color='white')
```

Out[20]:

```
<matplotlib.lines.Line2D at 0x7fa242091f70>
```



Observations:

- In this case we are analysing Education column. We can safely say based on visual observation that most of the customers have 16 years of education
- In the above histograms we can see that a bulk of the observations lie in the 2nd and 4th classes. The rest of the classes contain only a very few observations.
- From the above figure we can see that mean is represented by the yellow line, the mode by the white line and The median is represented by the orange line.
- We can see from the above figure that the mode and the median are very close to each other and the mean is lesser than both.

Now we have an idea of how the data is distributed for Education.

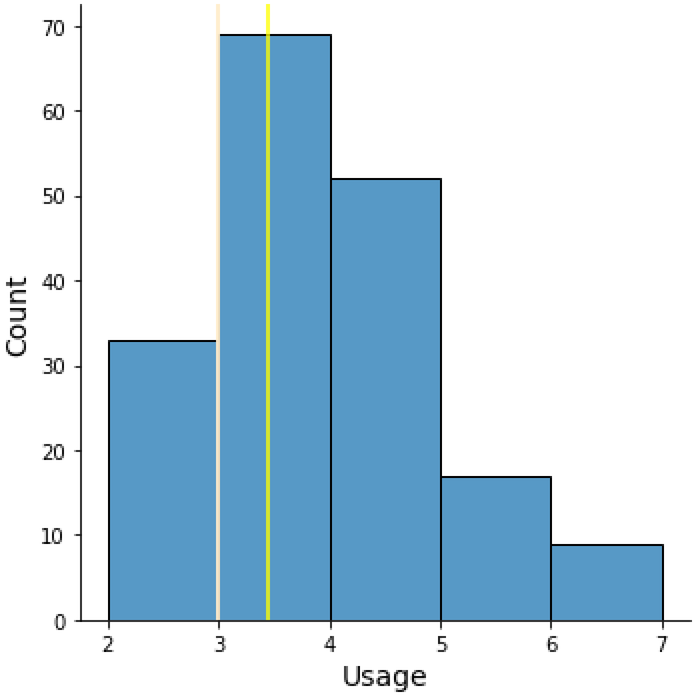
In [23]:

```
# plots a histogram plt using the seaborn package for Usage column.

sns.displot(cgfitness,
            x = "Usage",
            bins=5,
            height=5)
plt.xlabel("Usage", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Usage.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Usage.median(),
            color='orange')
plt.axvline(x=cgfitness.Usage.mode()[0],
            color='white')
```

Out[23]:

```
<matplotlib.lines.Line2D at 0x7fa24255e0a0>
```



Observations:

- In this case we are analysing Usage column. We can safely say based on visual observation that most of the count present on the data are 3-4 days per week

- In the above histograms we can see that a bulk of the observations lie within the first four classes. The rest of the four classes contain only a very few observations.
- From the above figure we can see that mean is represented by the yellow line, the mode by the white line and The median is represented by the orange line.
- We can see from the above figure that the mode and the median are very close to each other and the mean is higher than both.
- There are a very few counts that are more than 4 Usage. Once we pass the 4 Usage point the number of observations drops further.

Now we have an idea of how the data is distributed for Usage.

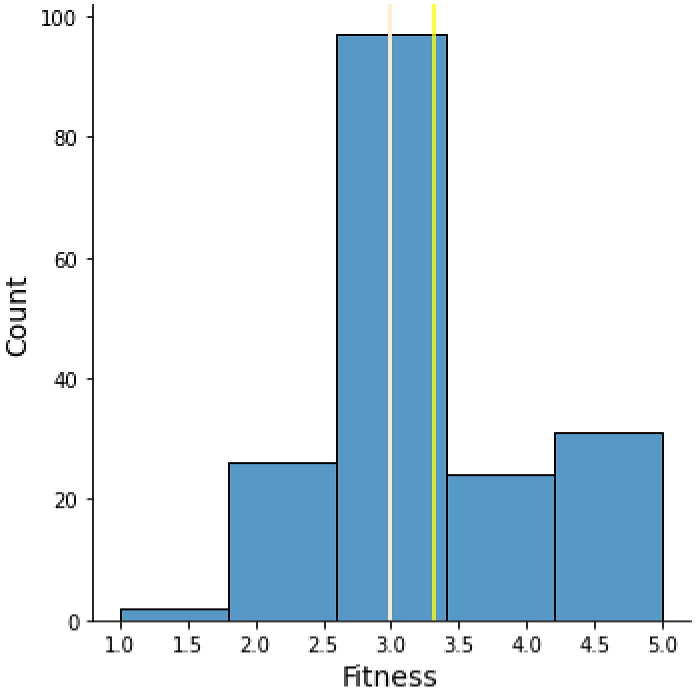
In [25]:

```
# plots a histogram plt using the seaborn package for Fitness column.

sns.displot(cgfitness,
            x = "Fitness",
            bins=5,
            height=5)
plt.xlabel("Fitness", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Fitness.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Fitness.median(),
            color='orange')
plt.axvline(x=cgfitness.Fitness.mode()[0],
            color='white')
```

Out[25]:

```
<matplotlib.lines.Line2D at 0x7fa24255e070>
```



Observations:

- In this case we are analysing Fitness column. We can safely say based on visual observation that most of the count present on the data are 3 rated and other contains very few observations.

- From the above figure we can see that mean is represented by the yellow line, the mode by the white line and The median is represented by the orange line.
- We can see from the above figure that the mode and the median are very close to each other and the mean is higher than both.

Now we have an idea of how the data is distributed for Fitness.

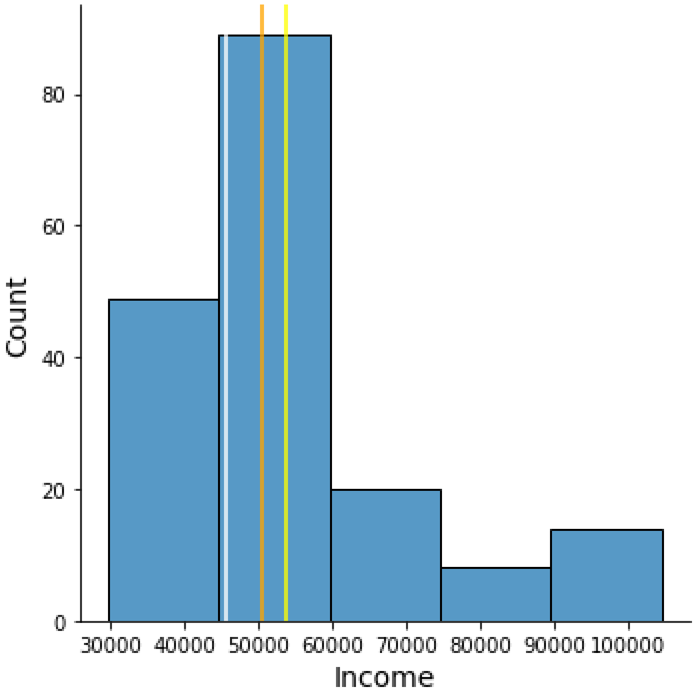
In [30]:

```
# plots a histogram plt using the seaborn package for Income column.

sns.displot(cgfitness,
            x = "Income",
            bins=5,
            height=5)
plt.xlabel("Income", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Income.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Income.median(),
            color='orange')
plt.axvline(x=cgfitness.Income.mode()[0],
            color='white')
```

Out[30]:

```
<matplotlib.lines.Line2D at 0x7fa2440f3f40>
```



Observations:

- In this case we are analysing Income column. We can safely say based on visual observation that most of the count present on the data are less than 60000 Income and Income is skewed towards right
- Most of the customers are in lower pay range and earn less than 60K.

- In the above histograms we can see that a bulk of the observations lie within the first four classes. The rest of the four classes contain only a very few observations.
- From the above figure we can see that mean is represented by the yellow line, the mode by the white line and The median is represented by the orange line.
- We can see from the above figure that the mean and the median are close to each other and the mode is lesser than both.
- There are a very few counts that are more than 60000 income. Once we pass the 60000 Income point the number of observations drops further.

Now we have an idea of how the data is distributed for Income.

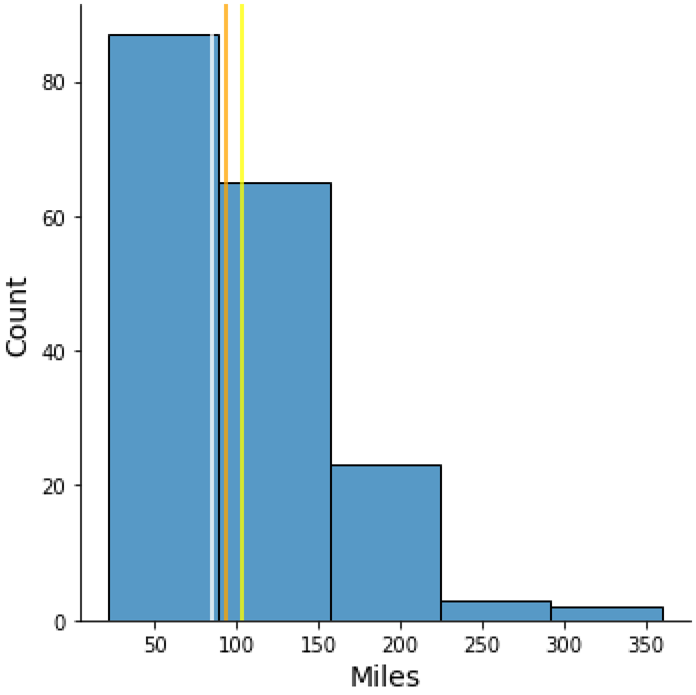
In [32]:

```
# plots a histogram plt using the seaborn package for Miles column.

sns.displot(cgfitness,
            x = "Miles",
            bins=5,
            height=5)
plt.xlabel("Miles", size=14)
plt.ylabel("Count", size=14)
plt.axvline(x=cgfitness.Miles.mean(),
            color='yellow')
plt.axvline(x=cgfitness.Miles.median(),
            color='orange')
plt.axvline(x=cgfitness.Miles.mode()[0],
            color='white')
```

Out[32]:

```
<matplotlib.lines.Line2D at 0x7fa2444ffd30>
```



Observations:

- In this case we are analysing Miles column. We can safely say based on visual observation that most of the count present on the data are less than 200 Miles and Miles is skewed

towards right.
- In the above histograms we can see that a bulk of the observations lie within the first four classes. The rest of the four classes contain only a very few observations.
- From the above figure we can see that mean is represented by the yellow line, the mode by the white line and The median is represented by the orange line.
- We can see from the above figure that the mode and the median are close to each other and the mean is higher than both.
- Customers are running average 80 miles per week.
- There are a very few counts that are more than 200 Miles. Once we pass the 200 Miles point the number of observations drops further.

Now we have an idea of how the data is distributed for Miles.

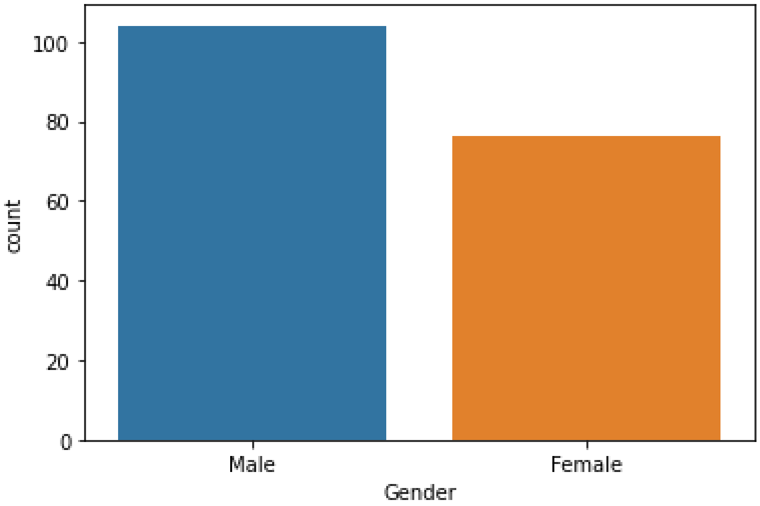
In [41]:

```
# Based on Gender column who is buying more treadmill( Male or Female )

sns.countplot(x="Gender", data=cgfitness, palette="tab10")
```

Out[41]:

```
<AxesSubplot:xlabel='Gender', ylabel='count'>
```



Observation: We can safely say that Male cutomers are more than Female cutomers.

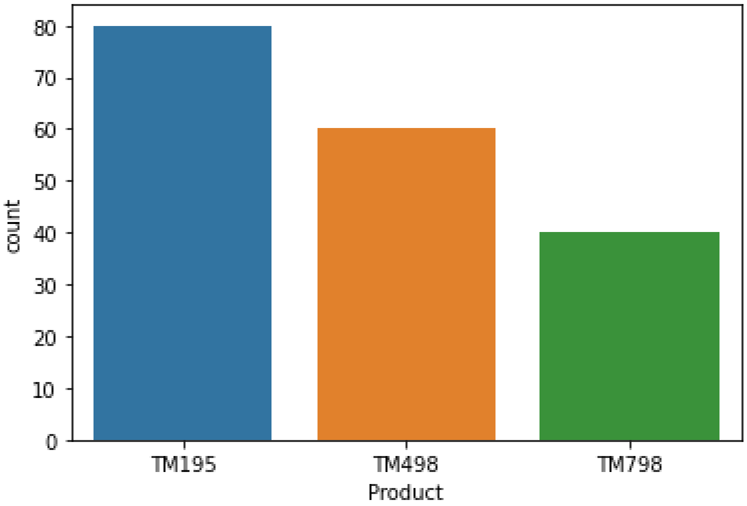
In [43]:

```
# Based on Product which treadmill model is most sold

sns.countplot(x="Product", data=cgfitness, palette="tab10")
```

Out[43]:

```
<AxesSubplot:xlabel='Product', ylabel='count'>
```

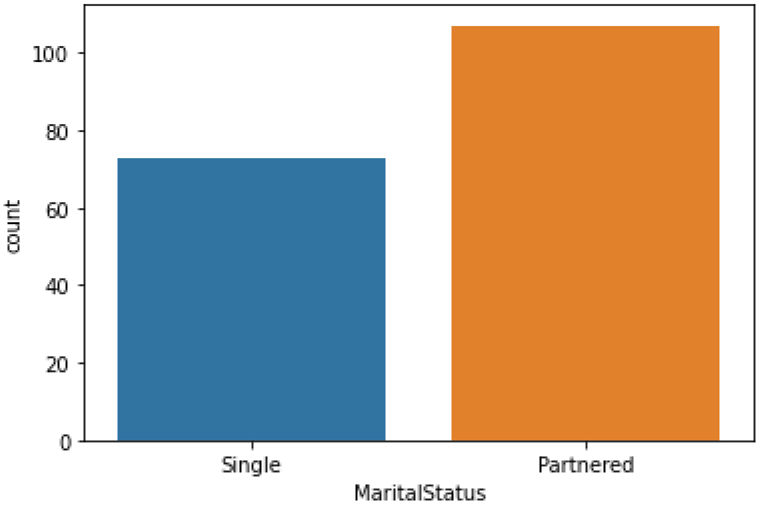


Observation: We can safely say that TM195 treadmill model is most sold model.

In [44]:
```
# Based on MaritalStatus who is buying more treadmill(Partnered or Single)

sns.countplot(x="MaritalStatus", data=cgfitness, palette="tab10")
```

Out[44]:
```
<AxesSubplot:xlabel='MaritalStatus', ylabel='count'>
```



Observation: We can safely say that customers who are Partnered are buying more treadmill.

Multivariate Data Analysis

Multivariate analysis is performed to understand interactions between different fields in the dataset (or) finding interactions between variables more than 2

Example: Pairplot, 3D scatter plot

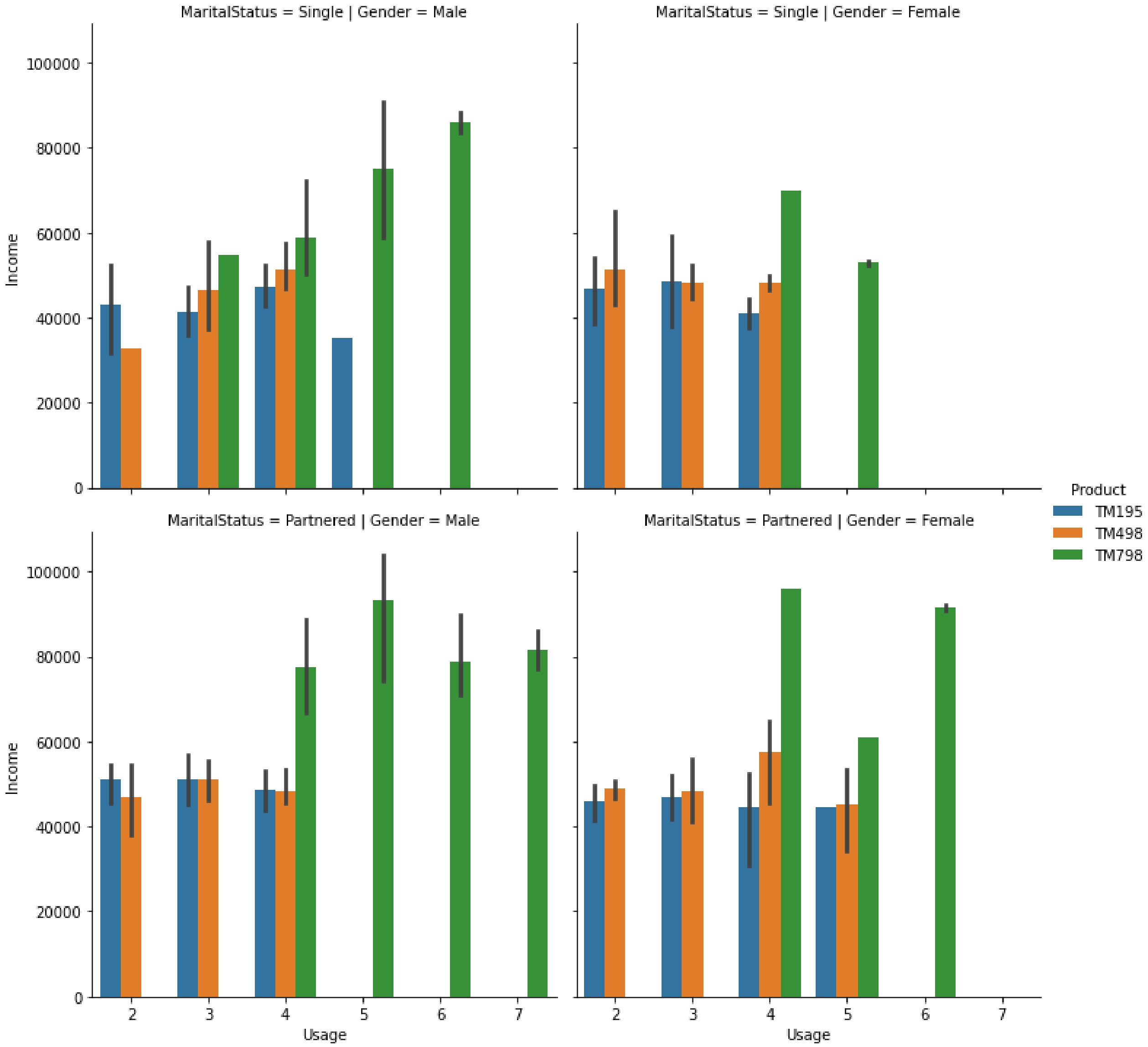
Analysis of interaction between features, in the dataset

In [22]:
```
# plots a cat plot using the catplot seaborn package for cgfitness dataset.
```

```
plt.figure(figsize=(10,10))
sns.catplot(x='Usage', y='Income', row = 'MaritalStatus', col='Gender',hue='Proc
```

Out[22]:

```
<seaborn.axisgrid.FacetGrid at 0x7feeda53e040>
<Figure size 720x720 with 0 Axes>
```



Observations:

- Male customers are more than female customer overall
- Male customer with higher income ,bought TM798 Model and they are using treadmill 4 to 6 days a week
- TM195 and TM498 models customers are in same income range and they are using treadmill 3 to 4 days a week
- Most of the customers who bought the TM798 are male and partnered.
- Single Male customers bought TM195 Model compared to Single Female.
- Partnered Male customers bought TM798 model more than Single Male customers.
- Single Female customers bought TM498 model more than Single male customers.

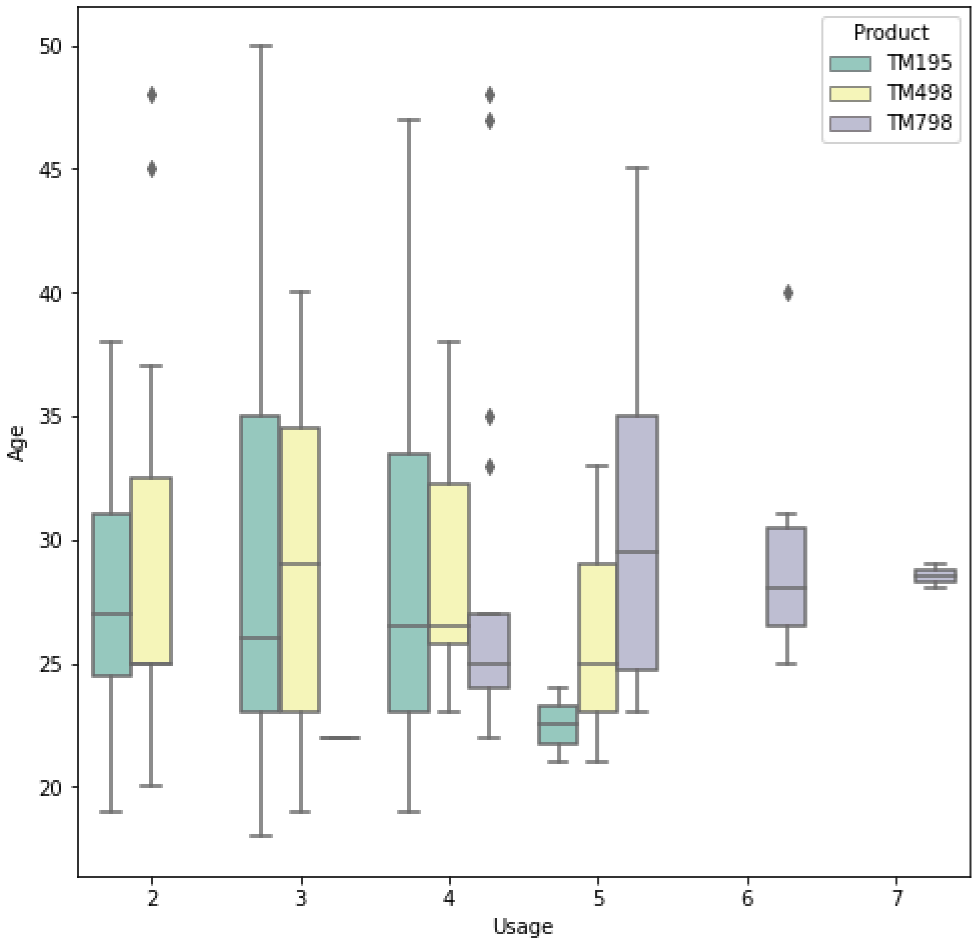
In [20]:

```
# plots a cat plot using the catplot seaborn package for cgfitness dataset.
```

```
plt.figure(figsize=(8,8))
#sns.catplot(x='Usage', y='Age', col='Gender',hue='Product' ,kind="bar", data=cg
#sns.catplot(x="feat_1", y="target", data=train)

sns.boxplot(x="Usage", y="Age", hue="Product",data=cgfitness, palette="Set3")
```

```
Out[20]: <AxesSubplot:xlabel='Usage', ylabel='Age'>
```



Observations:

- Products TM195 and TM498 models are bought by customers with more in range of 23 to 35 Age and usage of 3 days a week
- Product TM798 model is mainly bought by customers whose age falls in range of 24 to 35 Age and usage of 5 days a week. So TM789 model customer are usage is more than other two model customers.

Conclusion

- Customers buying most treadmills are younger and Average Age is 28.
- Customers buying most treadmills are having 16 years of education(Average).
- Customers buying most treadmills are using 3 to 4 days per week.
- 3 rated fitness customers buying most treadmills.
- Most of the customers are in lower pay range and earn less than 60K.
- Most of the customers average running is 80 miles
- More male customers bought Treadmill than Female customers.

- More customers bought TM195 model Treadmill. TM195 model is the most purchased model. TM498 was purchased more than TM798.
- Customers who purchased treadmill are Partnered than Single.
- Male customer with higher income ,bought TM798 Model and they are using treadmill 4 to 6 days a week
- TM195 and TM498 models customers are in same income range and they are using treadmill 3 to 4 days a week
- Single Female customers bought TM498 model than Single male customers.
- Products TM195 and TM498 models are bought by customers with more in range of 23 to 35 Age and usage of 3 days a week

Recommendations:

- From the conclusion TM798 model bought by higher income customers with most usage than other two models. So, this model should be marketed as a pro model for experienced and royal cutomers. So we can reduce publishing ad about this model and increase for other models because mostly higher income customers aren't going to watch ads.
- From the conclusion TM195 & TM498 model treadmills are most sold and bought by same less range income customers. So, it concludes that we should market these models as a mid range Treadmill and Increase the ad with feature comparison of these two models. We should motivate them to use more days per week and show case the extra features in the other high end models, by doing this we can push them to buy higher end model which is TM798.
- Our product is less familiar with 35 plus Age group so we need to focus more on this by giving offers or ads.
- From the conclusion we know that Female customers are lesser than Male customers but we also know that married/partnered Marital Status customers are higher than Single. So, we can focus more on married/partnered Martial status women and attract them with some offers which is useful for married people. We also need to focus more on Single women category with additional ads or benefits.