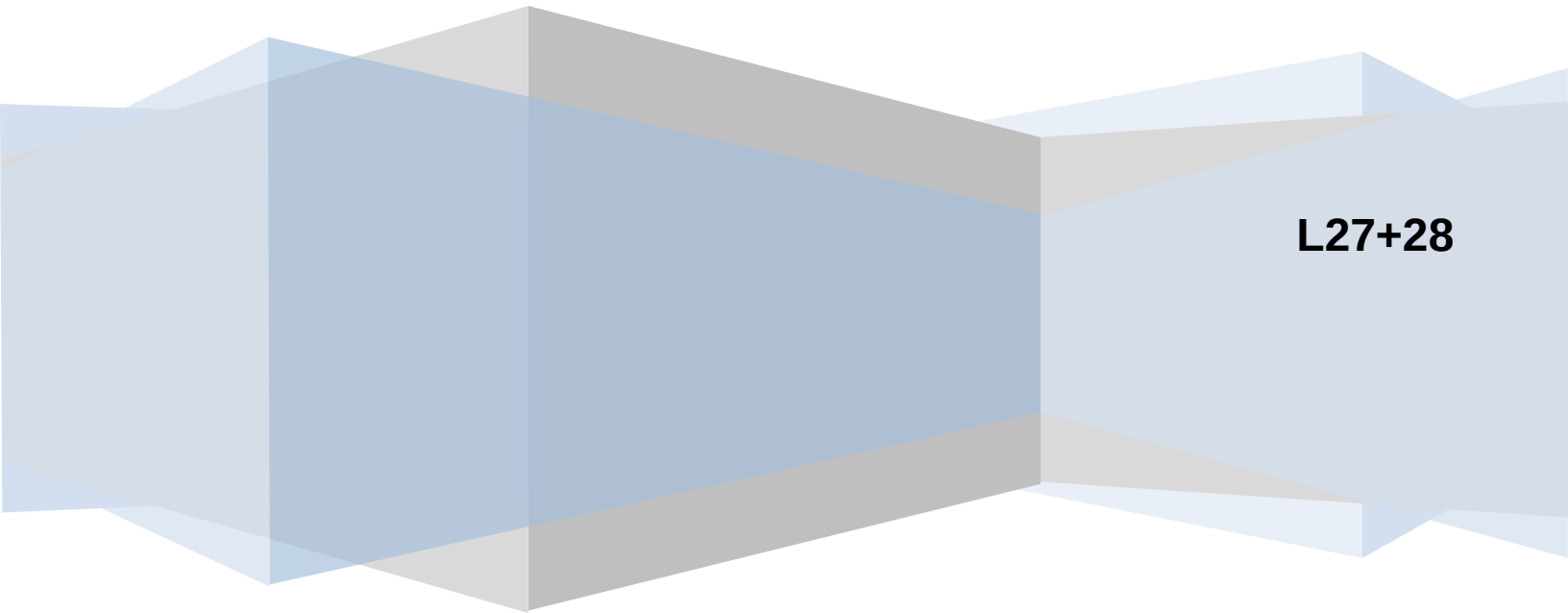
C S Sahil 19BCE2094

Data Visualization

Lab assignment - 2

PRAVAT KUMAR JENA



L27+28

Cardio Good Fitness Case Study - Descriptive Statistics

The market research team at AdRight is assigned the task to identify the profile of the typical customer for each treadmill product offered by CardioGood Fitness. The market research team decides to investigate whether there are differences across the product lines with respect to customer characteristics. The team decides to collect data on individuals who purchased a treadmill at a CardioGoodFitness retail store during the prior three months. The data are stored in the CardioGoodFitness.csv file.

The team identifies the following customer variables to study:

- product purchased, TM195, TM498, or TM798;
- gender;
- age, in years;
- education, in years;
- relationship status, single or partnered;
- annual household income ;
- average number of times the customer plans to use the treadmill each week;
- average number of miles the customer expects to walk/run each week;
- and self-rated fitness on an 1-to-5 scale, where 1 is poor shape and 5 is excellent shape.

Perform descriptive analytics to create a customer profile for each CardioGood Fitness treadmill product line.

Load the necessary packages

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
import warnings
warnings.filterwarnings('ignore') # To supress warnings
sns.set(style="whitegrid") # set the background for the graphs
```

Load the Cardio Dataset

```
mydata = pd.read_csv('CardioGoodFitness-1.csv')
```

Q1: Show few data from begin and end

```
mydata.head()
```

```
  Product  Age  Gender  Education MaritalStatus  Usage  Fitness  Income  Mile
s
0   TM195   18    Male         14        Single      3        4   29562    11
```

```
2
1   TM195   19    Male         15        Single      2        3   31836     7
5
2   TM195   19  Female         14     Partnered      4        3   30699     6
6
3   TM195   19    Male         12        Single      3        3   32973     8
5
4   TM195   20    Male         13     Partnered      4        2   35247     4
7
```

```
mydata.tail()
```

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

     Miles
175    200
176    200
177    160
178    120
179    180
```

Q2: Give a statistical description of all varibales available in the datasets.

```
mydata.describe(include="all")
```

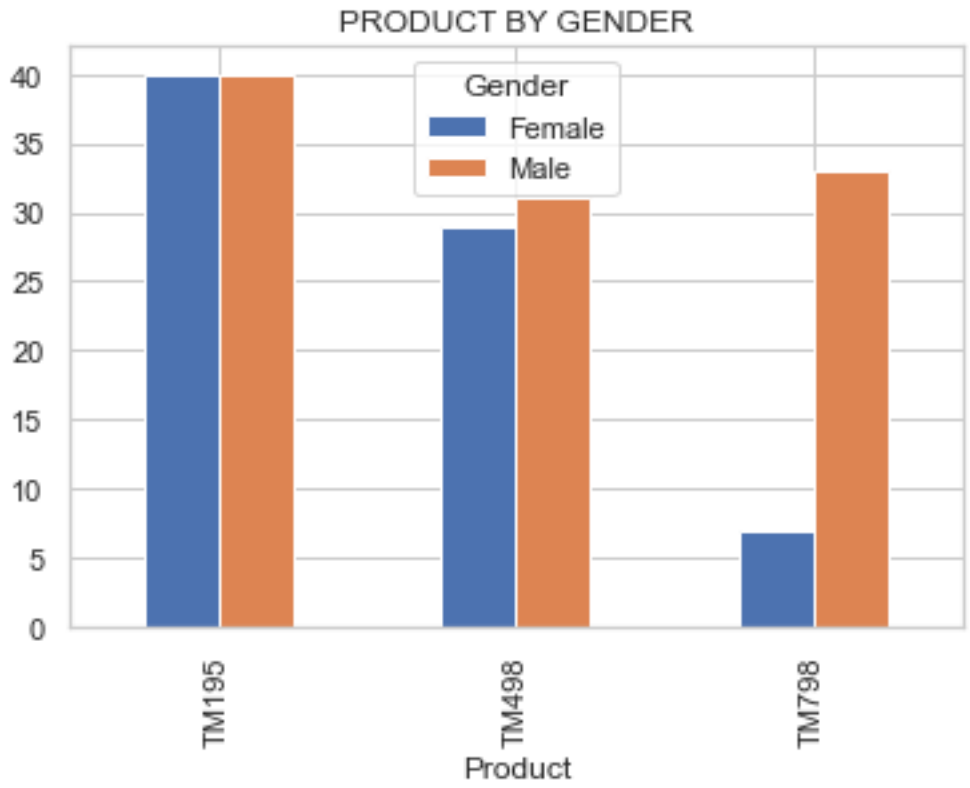
```
       Product         Age Gender   Education MaritalStatus       Usage  \
count      180  180.000000    180  180.000000           180  180.000000
unique       3         NaN      2         NaN             2         NaN
top      TM195         NaN   Male         NaN     Partnered         NaN
freq        80         NaN    104         NaN           107         NaN
mean       NaN   28.788889    NaN   15.572222           NaN    3.455556
std        NaN    6.943498    NaN    1.617055           NaN    1.084797
min        NaN   18.000000    NaN   12.000000           NaN    2.000000
25%        NaN   24.000000    NaN   14.000000           NaN    3.000000
50%        NaN   26.000000    NaN   16.000000           NaN    3.000000
75%        NaN   33.000000    NaN   16.000000           NaN    4.000000
max        NaN   50.000000    NaN   21.000000           NaN    7.000000

           Fitness         Income       Miles
count   180.000000     180.000000  180.000000
unique         NaN            NaN         NaN
top            NaN            NaN         NaN
freq           NaN            NaN         NaN
```

```
mean      3.311111   53719.577778  103.194444
std       0.958869   16506.684226   51.863605
min       1.000000   29562.000000   21.000000
25%       3.000000   44058.750000   66.000000
50%       3.000000   50596.500000   94.000000
75%       4.000000   58668.000000  114.750000
max       5.000000  104581.000000  360.000000
```

Q3: Which product of treadmill has been frequently used by male

```
plt.figure(figsize=(10,10))
prd_gender=pd.crosstab(mydata['Product'],mydata['Gender'] )
print(prd_gender)

ax=prd_gender.plot(kind='bar')

plt.title("PRODUCT BY GENDER")
```

```
Gender   Female  Male
Product
TM195        40    40
TM498        29    31
TM798         7    33

Text(0.5, 1.0, 'PRODUCT BY GENDER')

<Figure size 720x720 with 0 Axes>
```



TM195 Has been most frequently used

Q4: How many objects are there in the datasets

```
mydata.info()
```

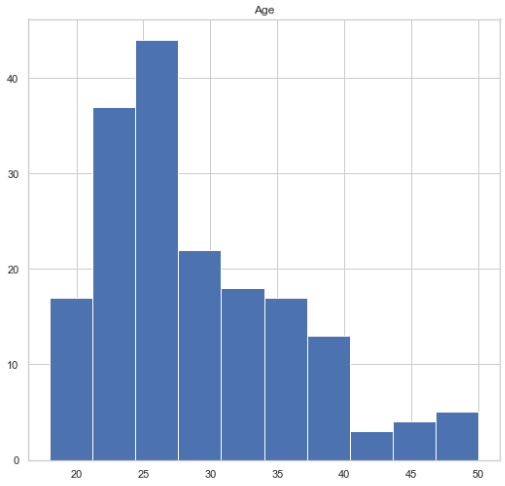
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Product        180 non-null    object
 1   Age            180 non-null    int64 
 2   Gender         180 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
```

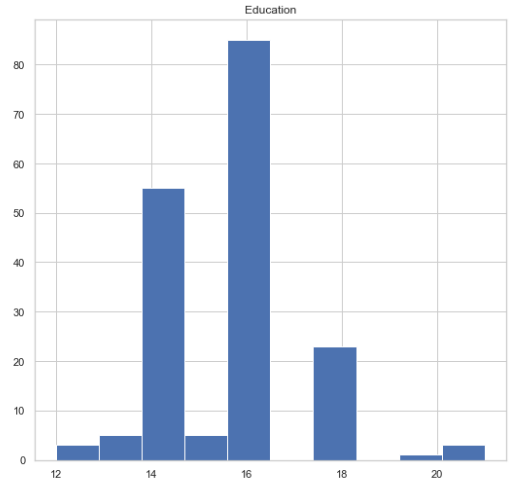
There are 3 Objects in the database

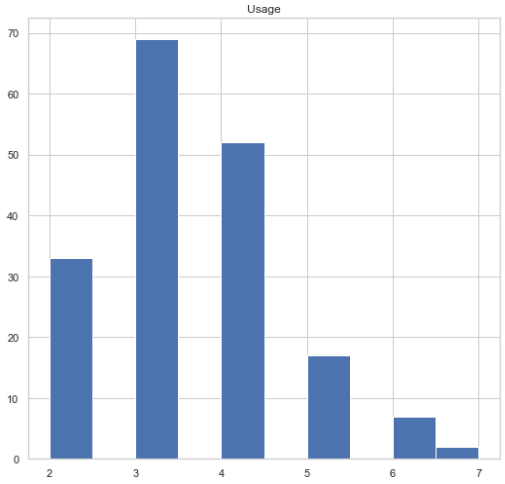
Q5: What your intution says about the numeric attributes such as Age, Income, Miles, and usage are normally distributed? Justified through required graphic

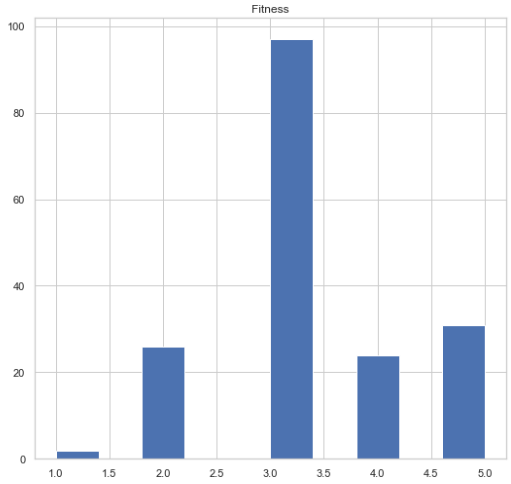
```
mydata.hist(figsize=(20,30))
```

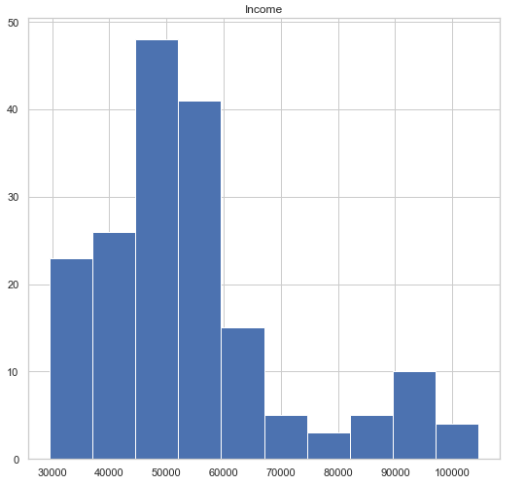
```
array([[<AxesSubplot:title={'center':'Age'}>,
        <AxesSubplot:title={'center':'Education'}>],
       [<AxesSubplot:title={'center':'Usage'}>,
        <AxesSubplot:title={'center':'Fitness'}>],
       [<AxesSubplot:title={'center':'Income'}>,
        <AxesSubplot:title={'center':'Miles'}>]], dtype=object)
```

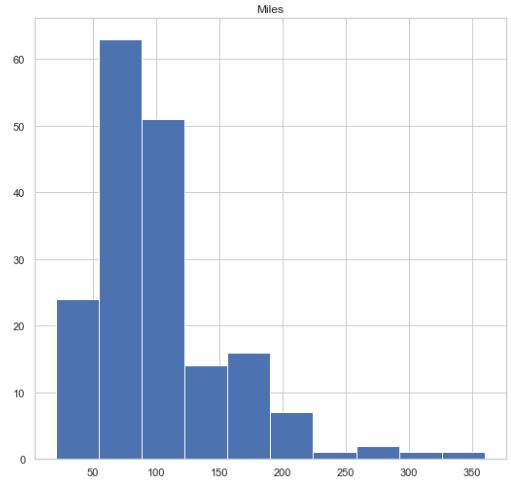












Education & Fitness Aproximately look normally distributed

Q6: Find the outlier if any exists in the variable Age. Hint: calcualte the IQR and use to filter the ourlier

```
data = mydata['Age']
sort_data = np.sort(data)
sort_data
```

```
array([18, 19, 19, 19, 19, 20, 20, 20, 20, 20, 21, 21, 21, 21, 21, 21, 21,
       22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 23, 23, 23, 23, 23,
       23, 23, 23, 23, 23, 23, 23, 23, 24, 24, 24, 24, 24, 24, 24, 24, 24,
       24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25,
       25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26,
       26, 26, 26, 26, 26, 26, 27, 27, 27, 27, 27, 27, 27, 28, 28, 28, 28,
       28, 28, 28, 28, 28, 29, 29, 29, 29, 29, 29, 30, 30, 30, 30, 30, 30,
       30, 31, 31, 31, 31, 31, 31, 32, 32, 32, 32, 33, 33, 33, 33, 33, 33,
       33, 33, 34, 34, 34, 34, 34, 34, 35, 35, 35, 35, 35, 35, 35, 35, 36,
       37, 37, 38, 38, 38, 38, 38, 38, 38, 39, 40, 40, 40, 40, 40, 41, 42,
       43, 44, 45, 45, 46, 47, 47, 48, 48, 50], dtype=int64)
```

```
Q1 = np.percentile(data, 25, interpolation = 'midpoint')
Q2 = np.percentile(data, 50, interpolation = 'midpoint')
Q3 = np.percentile(data, 75, interpolation = 'midpoint')
```

```
IQR = Q3 - Q1
print('Interquartile range is', IQR)
```

```
Interquartile range is 9.0
```

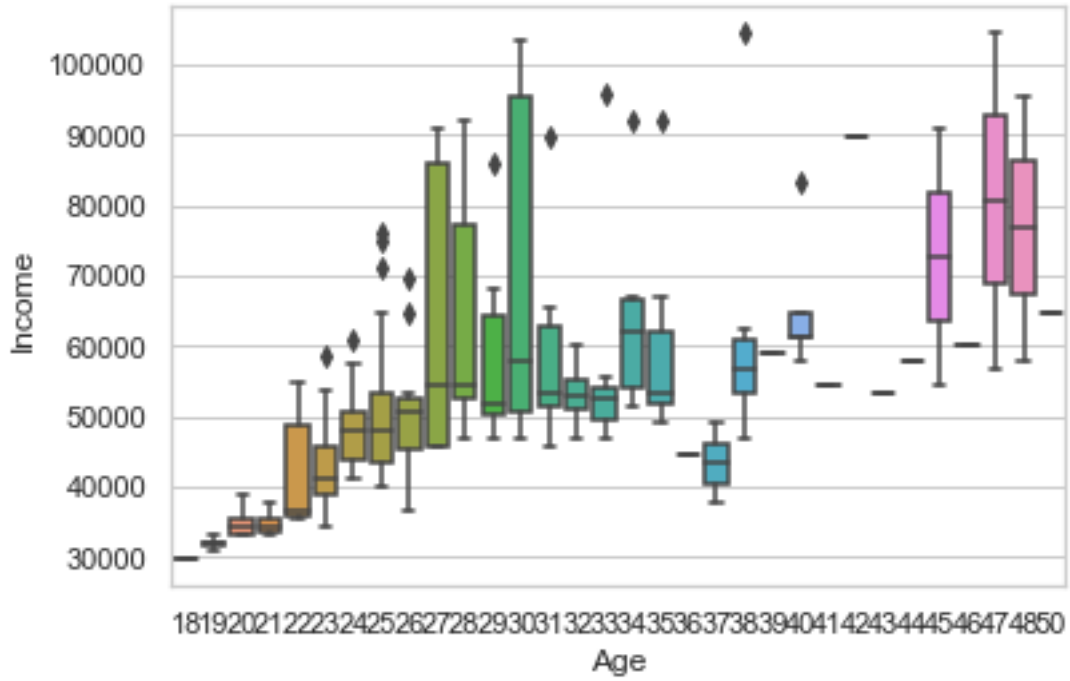
```
low_lim = Q1 - 1.5 * IQR
up_lim = Q3 + 1.5 * IQR
outlier =[]
for x in data:
    if ((x> up_lim) or (x<low_lim)):
         outlier.append(x)
print(' outlier in the dataset is', outlier)
```

```
 outlier in the dataset is [47, 50, 48, 47, 48]
```

Q7: Which plot is required to display five statistics of the variables Income with respect to Age. Display the graphics

```
sns.boxplot(x=mydata['Age'],y=mydata['Income'])
```

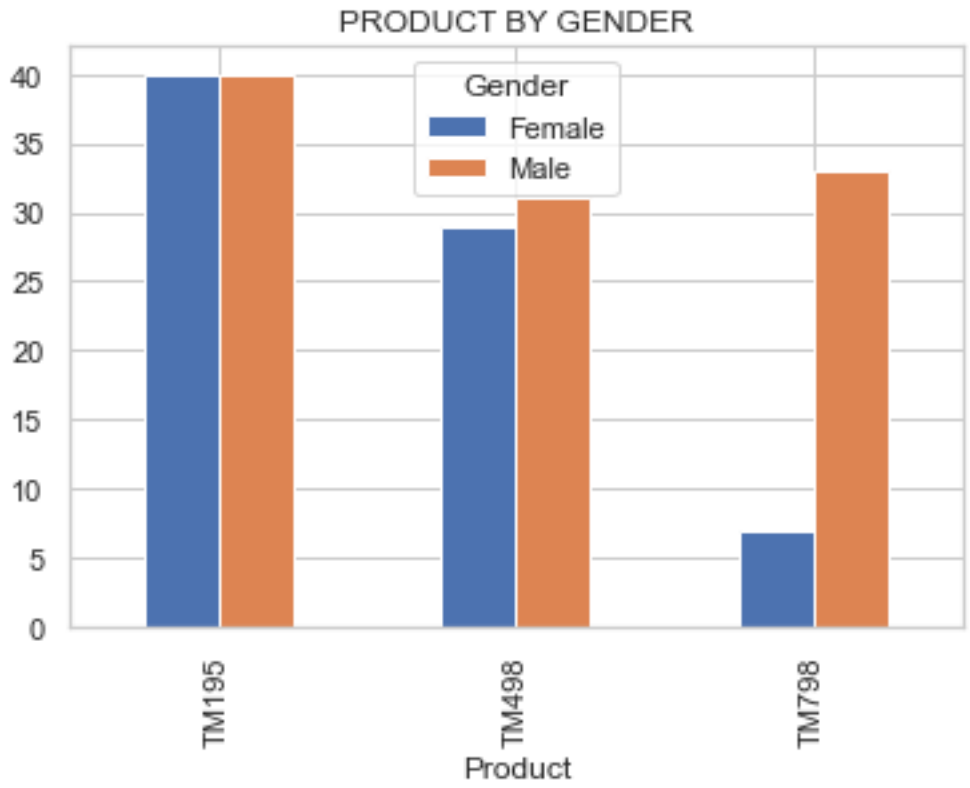
```
<AxesSubplot:xlabel='Age', ylabel='Income'>
```



Q8: How do you compare among the product of treadmill? or Which product is frequently used by gender-wise. Show your result through plot.

```
plt.figure(figsize=(10,10))
prd_gender=pd.crosstab(mydata['Product'],mydata['Gender'] )
print(prd_gender)

ax=prd_gender.plot(kind='bar')

plt.title("PRODUCT BY GENDER")
```

```
Gender   Female  Male
Product
TM195        40    40
TM498        29    31
TM798         7    33

Text(0.5, 1.0, 'PRODUCT BY GENDER')

<Figure size 720x720 with 0 Axes>
```



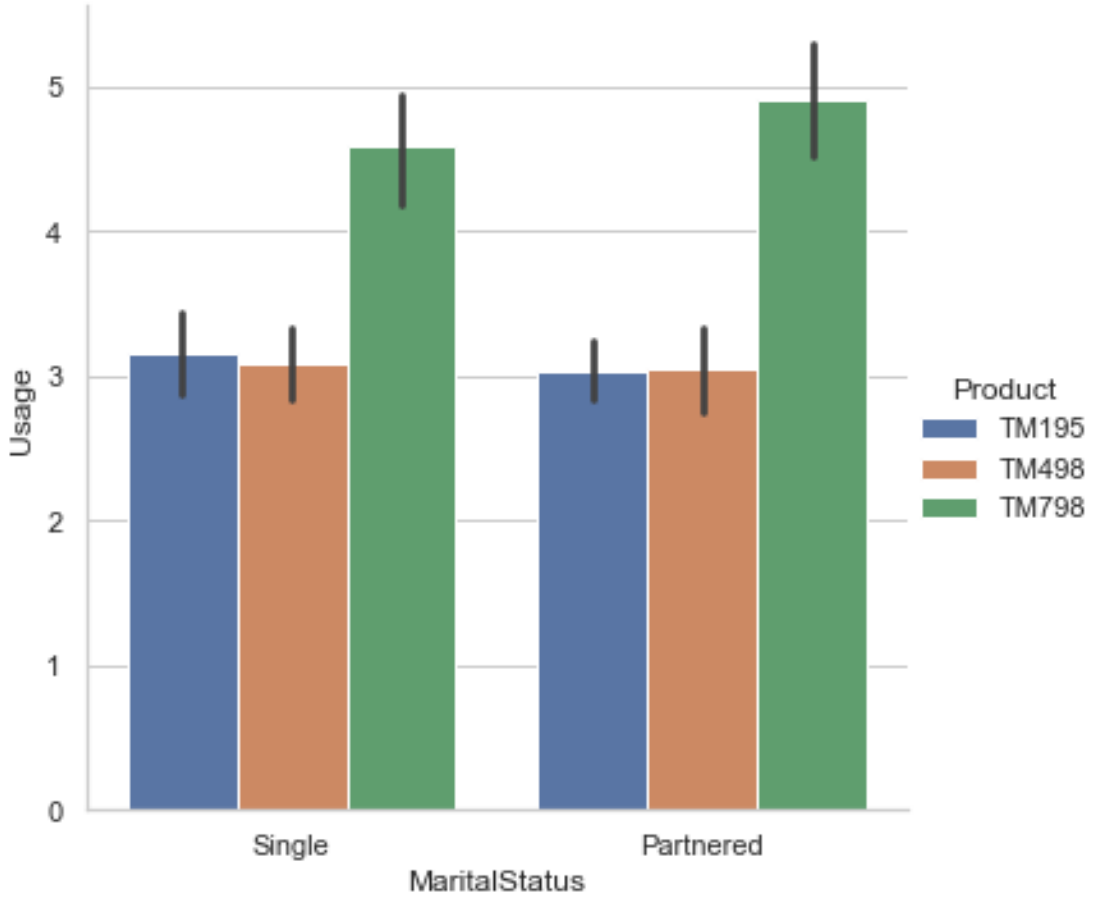

```
##  TM195 model was equally bought my Male and Female
##  Compared to females, male bought TM498 model .
##  TM798 model is popular in Males than in female.
```

Q9: Is maritial status affect the utiliation of the product of the treadmill? If so justify your results through the index matrix form

```
plt.figure(figsize=(12,7))
sns.catplot(x='MaritalStatus', y='Usage',hue='Product' ,kind="bar", data=myda
ta)
```

```
<seaborn.axisgrid.FacetGrid at 0x24f0b2acfa0>
```

```
<Figure size 864x504 with 0 Axes>
```




```
## Partnered Status had more usage for TM798 than Single
## Single has slightly higher usage for TM195
## TM498 Has equal usage for Singe and Partnered
```

Q10: How do you explain the relation between the numeric attributes? Which variables are correlated and quatify the relation?

```
corr_pairs = mydata.corr().unstack()
print( corr_pairs[abs(corr_pairs)>0.5])
```

```
Age        Age          1.000000
           Income       0.513414
Education  Education    1.000000
           Income       0.625827
Usage      Usage        1.000000
           Fitness      0.668606
           Income       0.519537
           Miles        0.759130
Fitness    Usage        0.668606
           Fitness      1.000000
           Income       0.535005
```

```
            Miles        0.785702
Income      Age          0.513414
            Education    0.625827
            Usage        0.519537
            Fitness      0.535005
            Income       1.000000
            Miles        0.543473
Miles       Usage        0.759130
            Fitness      0.785702
            Income       0.543473
            Miles        1.000000
dtype: float64
```

```
##Age and Income has some in significant correlation
##Education and Income has very little correlation
##There is some corelation between Usage and Income
##Fitness and miles are corelated
##TM798 model is correlated to Education, Usage,Fitness, Income and Miles.
##Miles and usage are positively correlated
```

Q11: Develope a model which can predict distance in miles with respect to fitness and usage.

```
from sklearn import linear_model


regr = linear_model.LinearRegression()

y = mydata['Miles']
x = mydata[['Usage','Fitness']]


regr.fit(x,y)

LinearRegression()

regr.coef_

array([20.21486334, 27.20649954])

regr.intercept_

-56.74288178464856

## To use this model just use like predictedDistance = -56.74 + 20.21*Usage +
27.20*Fitness
```