```
# Defining Problem Statement and Analysing basic metrics (10 Points)
# Observations on shape of data, data types of all the attributes, conversion of categorical
# Non-Graphical Analysis: Value counts and unique attributes (10 Points)
# Visual Analysis - Univariate & Bivariate (30 Points)
# For continuous variable(s): Distplot, countplot, histogram for univariate analysis (10 Poin
# For categorical variable(s): Boxplot (10 Points)
# For correlation: Heatmaps, Pairplots(10 Points)
# Missing Value & Outlier Detection (10 Points)
# Business Insights based on Non-Graphical and Visual Analysis (10 Points)
# Comments on the range of attributes
# Comments on the distribution of the variables and relationship between them
# Comments for each univariate and bivariate plot
# Recommendations (10 Points) - Actionable items for business. No technical jargon. No compli


# Import the dataset and do usual data analysis steps like checking the structure & character
# Detect Outliers (using boxplot, “describe” method by checking the difference between mean a
# Check if features like marital status, age have any effect on the product purchased (using
# Representing the marginal probability like - what percent of customers have purchased KP281
# Check correlation among different factors using heat maps or pair plots.
# With all the above steps you can answer questions like: What is the probability of a male c
# Customer Profiling - Categorization of users.
#·Probability-·marginal,·conditional·probability.
# Some recommendations and actionable insights, based on the inferences.


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



aerofit_df = pd.read_csv('https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/


aerofit_df.head(5)
```

	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

```
aerofit_df.shape
```

```
(180, 9)
```

As we can see, the dataframe consists of 9 columns and 180 records.

The name of Columns are:

```
aerofit_df.columns
```

```
Index(['Product', 'Age', 'Gender', 'Education', 'MaritalStatus', 'Usage',
       'Fitness', 'Income', 'Miles'],
      dtype='object')
```

```
aerofit_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
```

```
aerofit_df.describe()
```

	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

```
aerofit_df.isnull().sum()
```

```
Product          0
Age              0
Gender           0
Education        0
MaritalStatus    0
Usage            0
Fitness          0
Income           0
Miles            0
dtype: int64
```

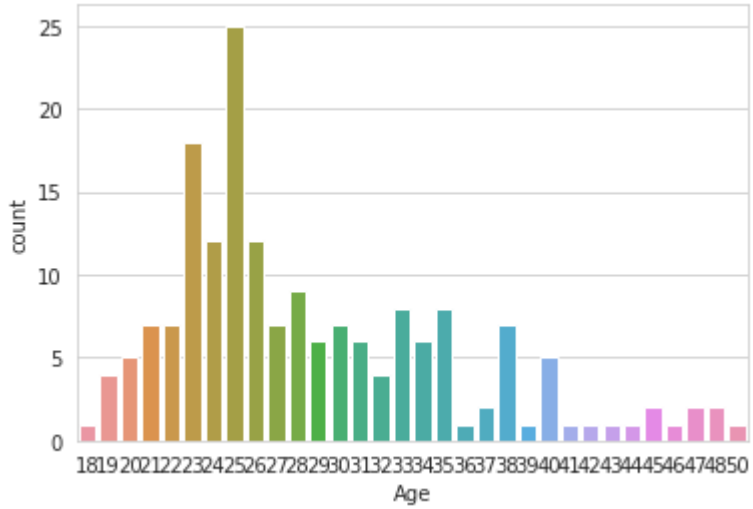
1. There are no missing values in the data.
2. There are 3 unique products in the dataset.
3. KP281 is the most frequent product i.e KP281, KP481, KP781.
4. Minimum & Maximum age of the person is 18 & 50, mean is 28.79 and 75% of persons have age less than or equal to 33.
5. Most of the people are having 16 years of education i.e. 75% of persons are having education <= 16 years.

Uni-Variate Analysis

What is the most popular/targeted AGE DEMOGRAPHIC who purchases the Products?

```
sns.countplot(x=aerofit_df['Age'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a33951d0>
```

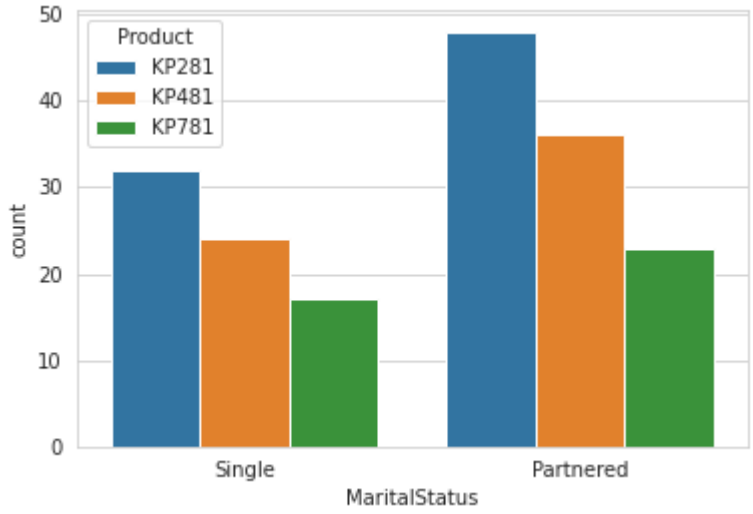


As we can see from above graph the mode as well as Countplot is observing data that most of Demographic Age lies between 22-25 years. So the target audience in this case should be of the age 22-25 years

How the Product Purchasing is Distributed among Peoples?

```
sns.countplot(x=aerofit_df['MaritalStatus'],hue=aerofit_df['Product'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a3330f90>
```



As we can see from above plot, the product is purchased by Married Couples in which most of the cases the Male has purchased the product as compared to Female.

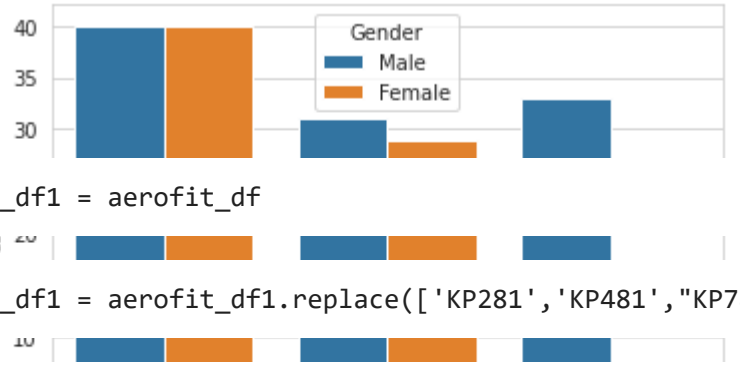
And in Single demographic too we can see that the Male seems to purchase the product more as compare to Female.

So if in future a product is launched, the target of the product should be Couples rather than single persons.

Which Gender prefers what kind of Products?

```
sns.countplot(x=aerofit_df['Product'],hue=aerofit_df['Gender'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a325a810>
```




```
aerofit_df1 = aerofit_df
```

```
aerofit_df1 = aerofit_df1.replace(['KP281','KP481',"KP781"],['281','481','781'])
```

As we can see the Product 'KP281' is more purchased than 'KP481' and 'KP781'.

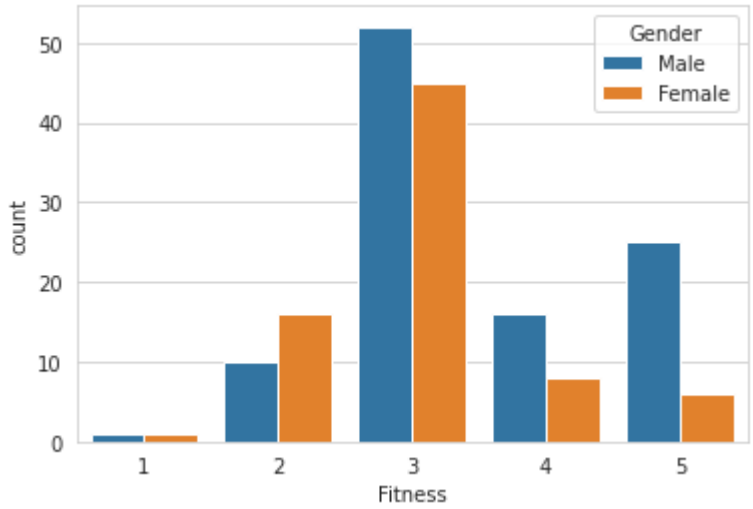
As it seems that KP281 had started early.

Intresting fact about above plot is that while the ratio of Purchasing of Product between Male and Female is higher for Male, but Product KP281 is likely purchased by both the Genders.

What is the distribution of Fitness Levels among the Genders?

```
sns.countplot(x=aerofit_df['Fitness'],hue=aerofit_df['Gender'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a3160d50>
```

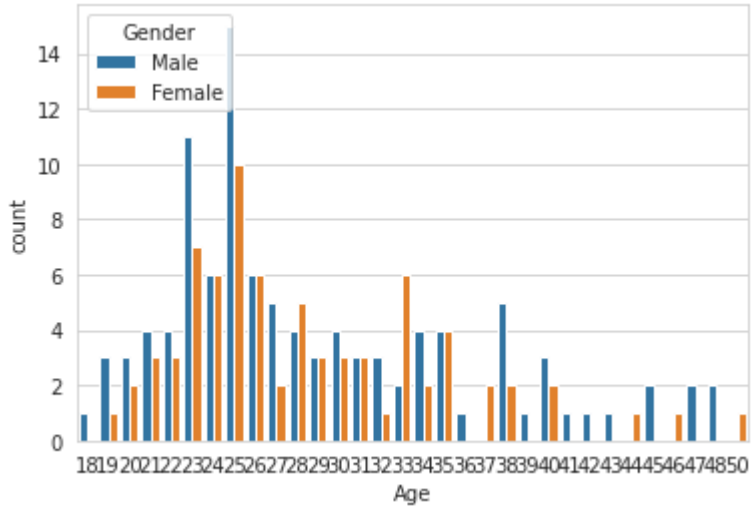


From above plot we can state that the Fitness Levels of Male has Higher than Fitness Level of Female.

Which Gender is more likely to buy the product according to there Ages?

```
sns.countplot(x=aerofit_df['Age'],hue=aerofit_df['Gender'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a3106110>
```



From above Age Demographic plot we can conclude that if we focus on Perma-boosted Age group, Almost few ages is dominated by Male Gender but similarly few age group are equally likely.

What is the MEAN/MEDIAN calculated for each columns?

[ ] ↳ *2 cells hidden*

Individual Product Descriptive Statistics

Product **KP281**

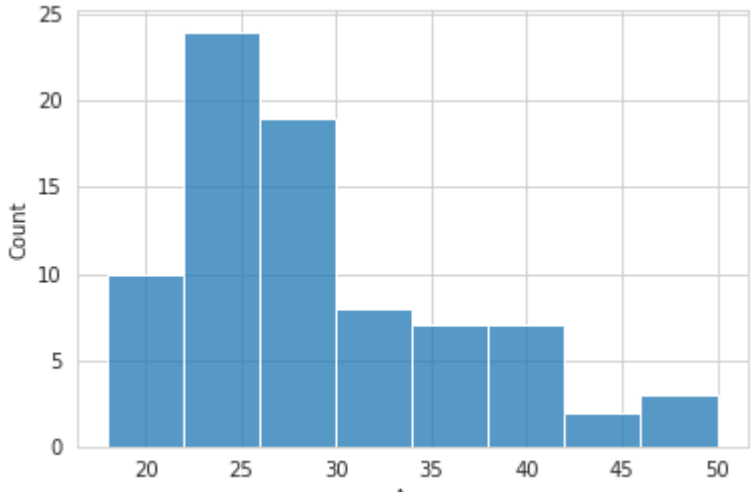
```
KP281 = aerofit_df.loc[aerofit_df['Product']=='KP281']
```

```
KP281.mean()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Dropping
  """Entry point for launching an IPython kernel.
Age             28.5500
Education       15.0375
Usage            3.0875
Fitness          2.9625
Income       46418.0250
Miles           82.7875
dtype: float64
```

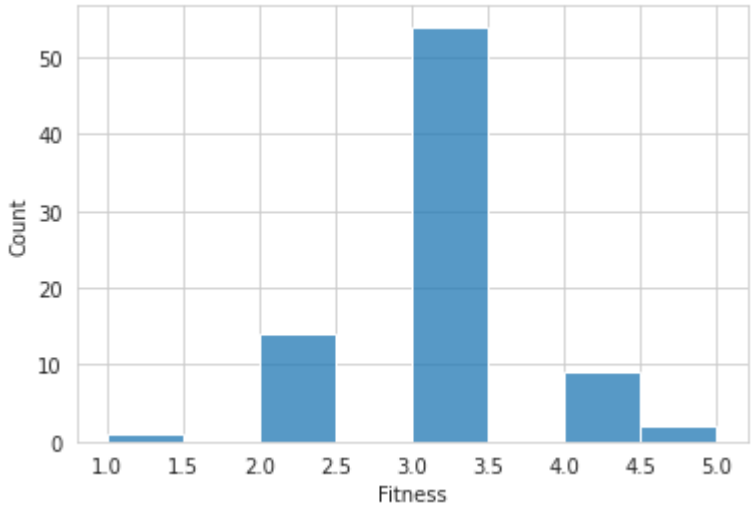
```
sns.histplot(data=KP281,x='Age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a40bb910>
```




```
sns.histplot(data=KP281,x='Fitness')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a37cd550>
```



Product **KP481**

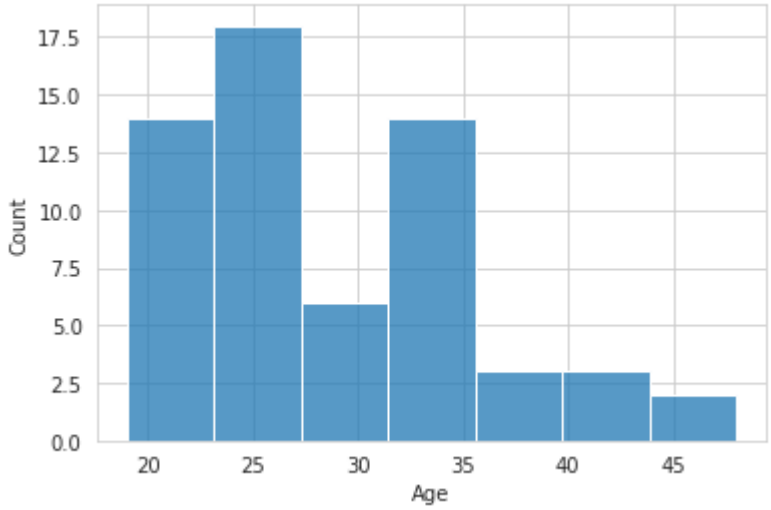
```
KP481 = aerofit_df.loc[aerofit_df['Product']=='KP481']
```

```
KP481.mean()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Dropping
  """Entry point for launching an IPython kernel.
Age             28.900000
Education       15.116667
Usage            3.066667
Fitness          2.900000
Income       48973.650000
Miles           87.933333
dtype: float64
```

```
sns.histplot(data=KP481,x='Age')
```

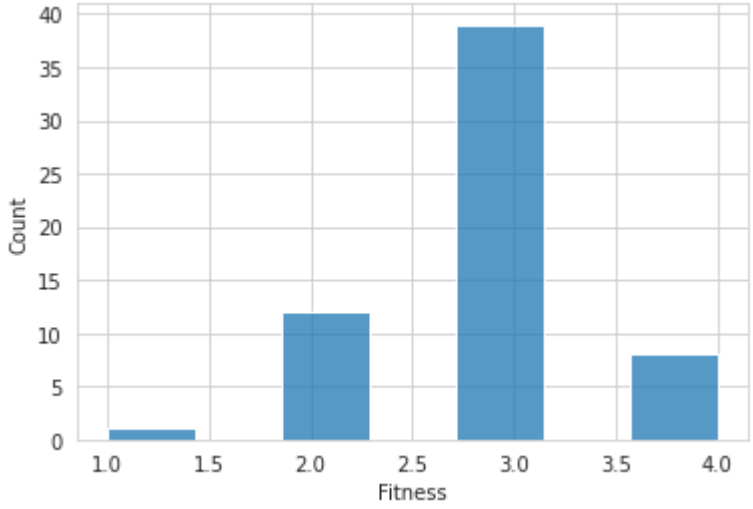
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a3b5ccd0>
```



For product **KP481**, the Age Group targeted are 28-29 years old.

```
sns.histplot(data=KP481,x='Fitness')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a372b410>
```



Product **KP781**

```
KP781 = aerofit_df.loc[aerofit_df['Product']=='KP781']
```

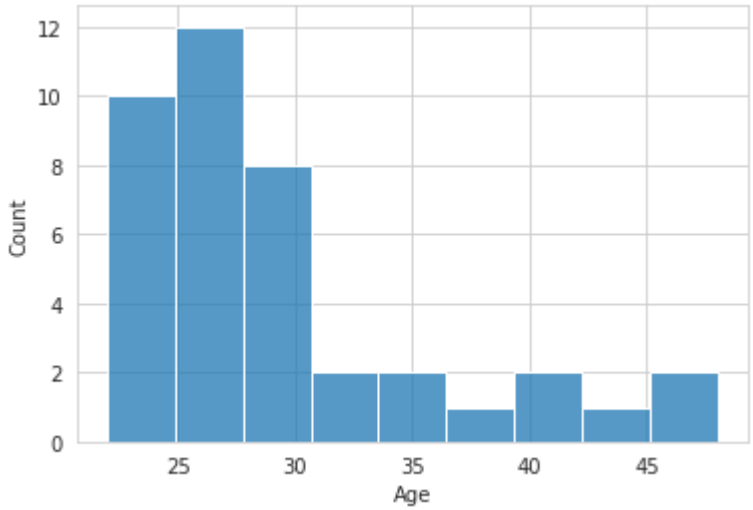
```
KP781.mean()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: FutureWarning: Dropping
  """Entry point for launching an IPython kernel.
```

```
Age             29.100
Education       17.325
Usage            4.775
Fitness          4.625
Income       75441.575
Miles          166.900
dtype: float64
```

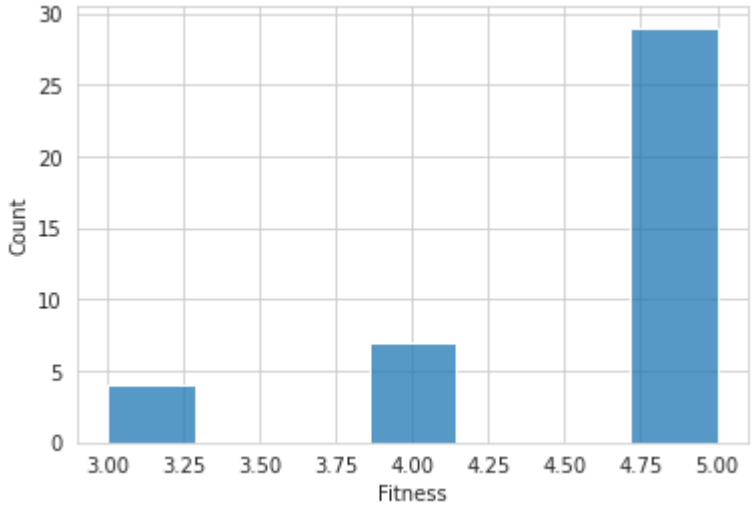
```
sns.histplot(data=KP781,x='Age')
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a40bb050>
```




```
sns.histplot(data=KP781,x='Fitness')
```

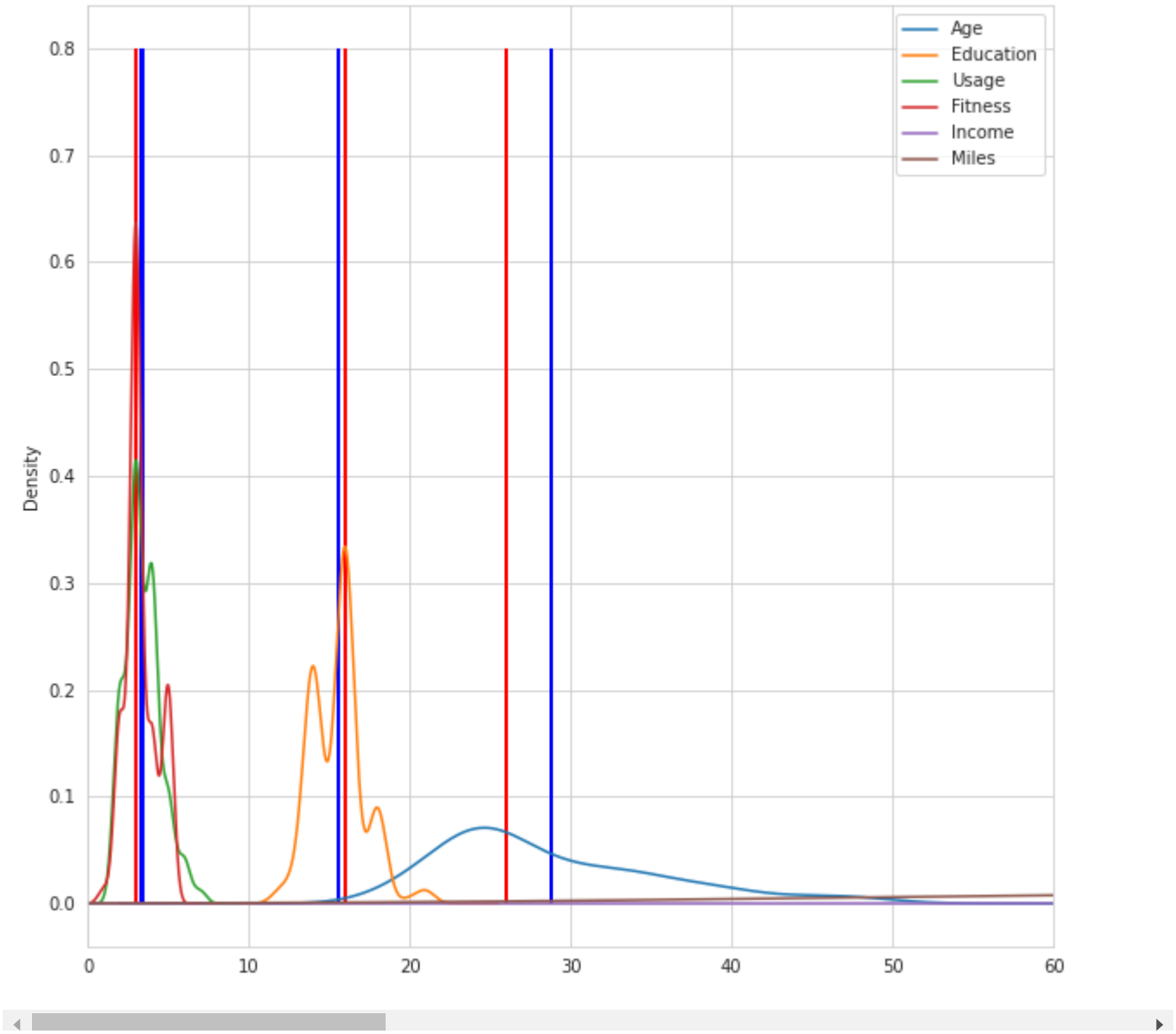
```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a4045c90>
```




```
aerofit_df.plot(kind='density',figsize = (10,10),xlim=(0,60))
#Black line at Mean
plt.vlines(aerofit_df.mean(),ymin=0,ymax=0.8,linewidth=2.0,color='blue');
# Red line at median
plt.vlines(aerofit_df.median(),ymin=0, ymax=0.8, linewidth=2.0,color="red");
```

```
plt.show()
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:3: FutureWarning: Dropping
  This is separate from the ipykernel package so we can avoid doing imports until
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:5: FutureWarning: Dropping
  """
```



From above plot, we can see that "Usage" and "Education" has a Normal distribution but the "Age" follows a Skewed distribution.

```
aerofit_df.mode()
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	25	Male	16	Partnered	3	3	45480	85

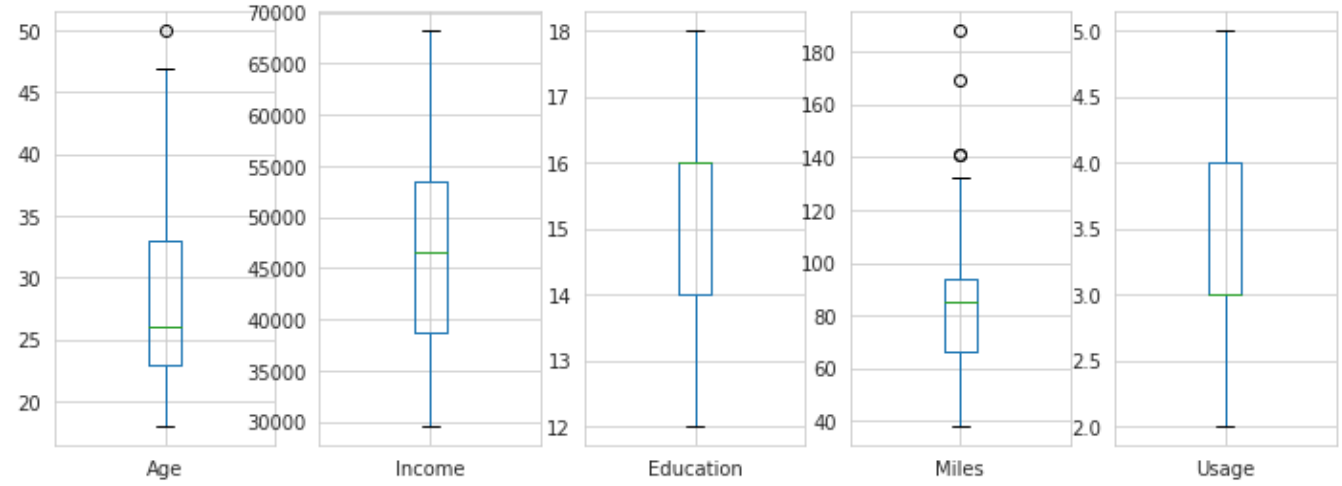
```
fig, axis = plt.subplots(nrows=3, ncols=5, figsize=(12, 10))
```

```
fig.subplots_adjust(top=1.2)
# For Treadmill KP281
KP281.boxplot(column='Age',ax=axis[0,0])
KP281.boxplot(column='Income',ax=axis[0,1])
KP281.boxplot(column='Education',ax=axis[0,2])
KP281.boxplot(column='Miles',ax=axis[0,3])
KP281.boxplot(column='Usage',ax=axis[0,4])


# For Treadmill KP481
KP481.boxplot(column='Age',ax=axis[1,0])
KP481.boxplot(column='Income',ax=axis[1,1])
KP481.boxplot(column='Education',ax=axis[1,2])
KP481.boxplot(column='Miles',ax=axis[1,3])
KP481.boxplot(column='Usage',ax=axis[1,4])

#For Treadmill KP781
KP781.boxplot(column='Age',ax=axis[2,0])
KP781.boxplot(column='Income',ax=axis[2,1])
KP781.boxplot(column='Education',ax=axis[2,2])
KP781.boxplot(column='Miles',ax=axis[2,3])
KP781.boxplot(column='Usage',ax=axis[2,4])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f11a3a23b90>
```



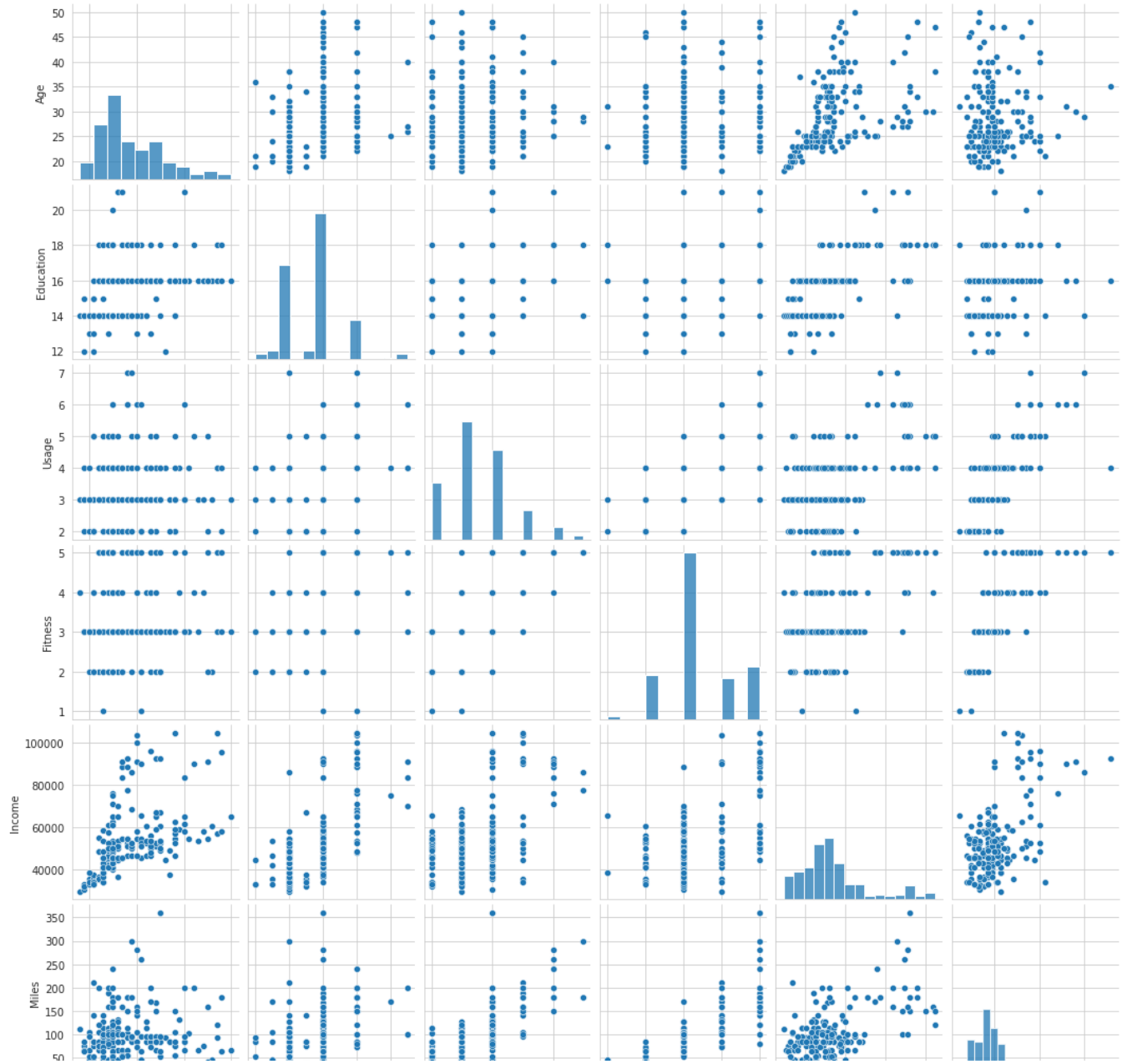
From above Box-Plots we can conclude that:

1. Age, Education and Usage are having very few outliers.
2. While Income and Miles are having more outliers.

Check correlation among different factors using pair plots.

```
sns.pairplot(aerofit_df)
```

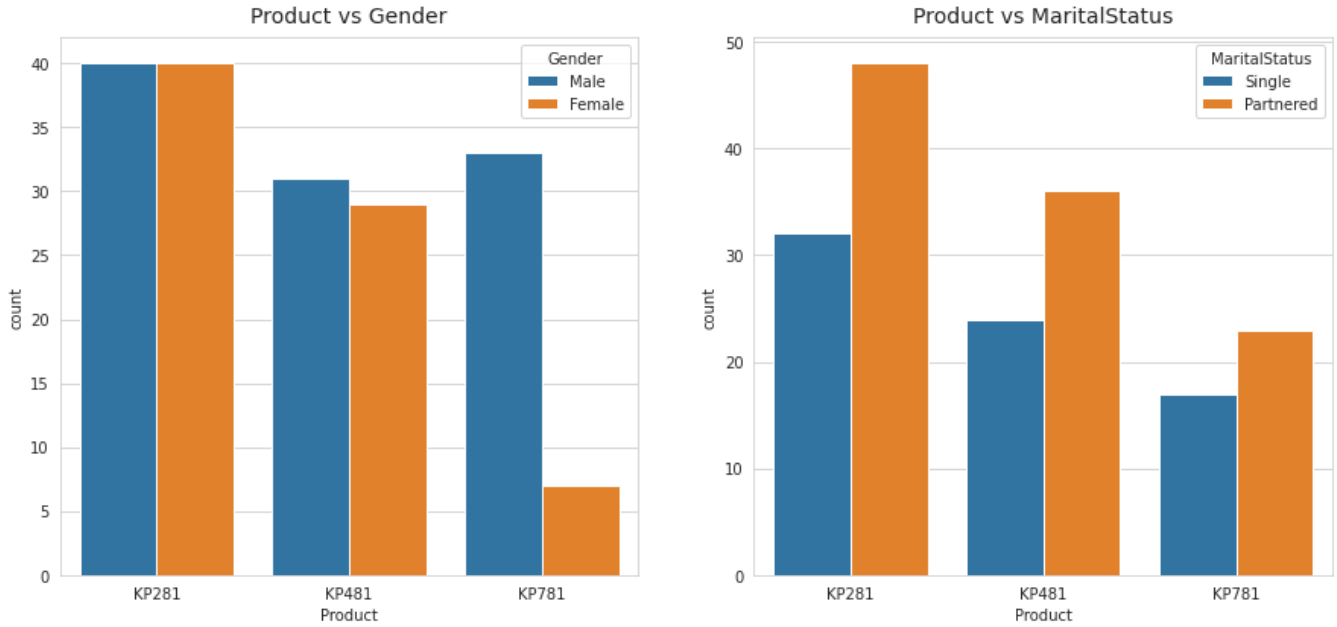
```
<seaborn.axisgrid.PairGrid at 0x7f11a3911950>
```



Bi-Variate Analysis

If features like marital status, age have any effect on the product purchased?

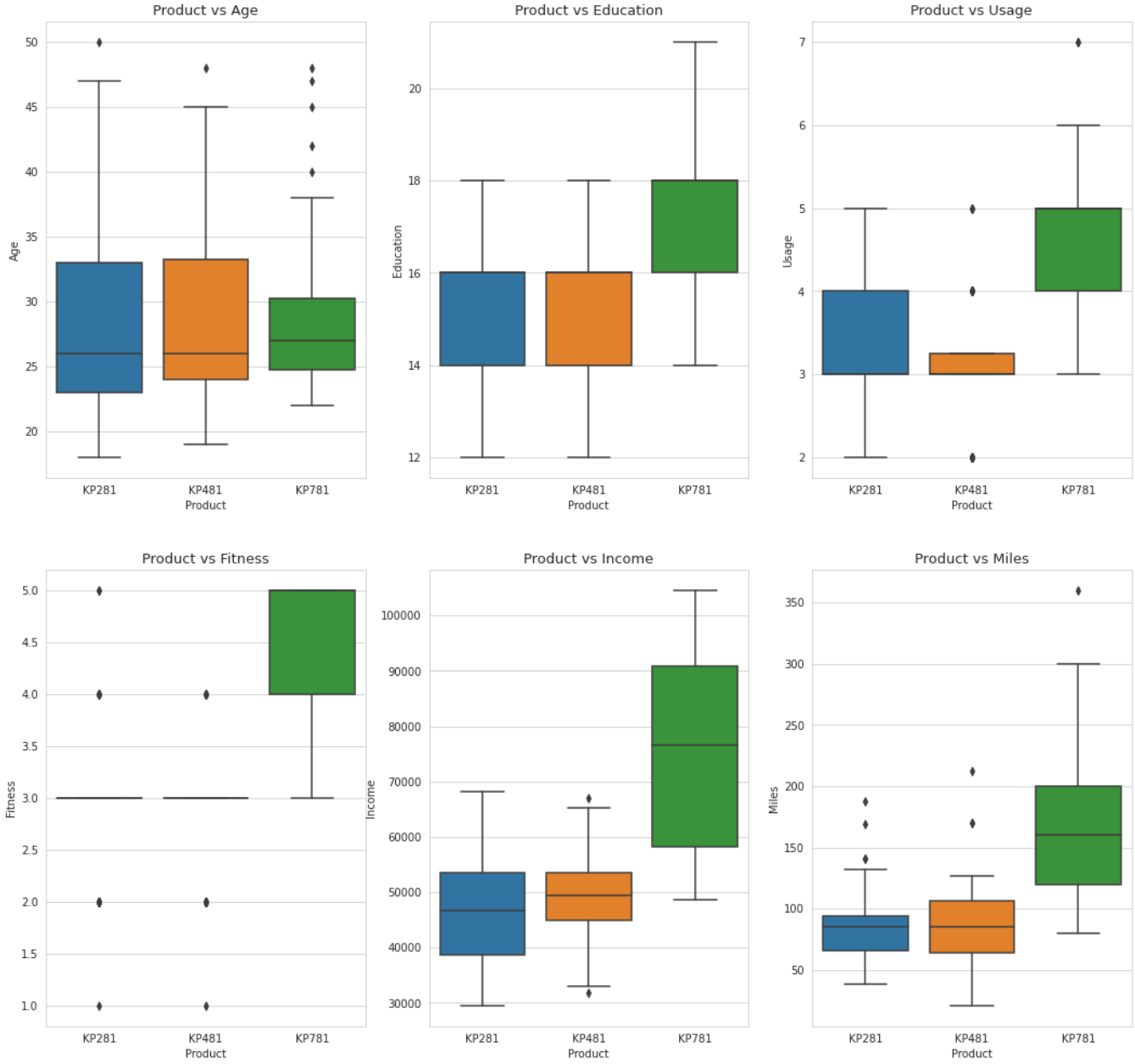
```
sns.set_style(style='whitegrid')
fig, axs = plt.subplots(nrows=1, ncols=2, figsize=(15, 6.5))
sns.countplot(data=aerofit_df, x='Product', hue='Gender', ax=axs[0])
sns.countplot(data=aerofit_df, x='Product', hue='MaritalStatus', ax=axs[1])
axs[0].set_title("Product vs Gender", pad=10, fontsize=14)
axs[1].set_title("Product vs MaritalStatus", pad=10, fontsize=14)
plt.show()
```



From above Plots we can observe that,

1. Product vs Gender count, Equal number of males and females have purchased KP281 product and Almost same for the product KP481.
2. Most of the Male customers have purchased the KP781 product.
3. Product vs Marital Status, the Customers who are Partnered are most likely to purchase the product.

```
bins = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
fig, axs = plt.subplots(nrows=2, ncols=3, figsize=(18, 12))
fig.subplots_adjust(top=1.2)
count = 0
for i in range(2):
    for j in range(3):
        sns.boxplot(data=aerofit_df, x='Product', y=bins[count], ax=axs[i,j])
        axs[i,j].set_title(f"Product vs {bins[count]}",fontsize=13)
        count += 1
```



Product vs Age:

1. Customers purchasing products KP281 & KP481 are having same Age median value.
2. Customers whose age is in between 25-30, are more likely to buy KP781 product

Product vs Education:

1. Customers whose Education is greater than 16, have more chances to purchase the KP781 product. While the customers with Education less than 16 have equal chances of purchasing other products i.e KP281 or KP481.

Product vs Usage:

1. Customers who are planning to use the treadmill greater than 4 times a week, are more likely to purchase the KP781 product.
2. While the other customers are likely to purchasing KP281 or KP481.

Product vs Fitness:

1. The more the customer is fit (fitness >= 3), higher the chances of the customer to purchase the KP781 product.

Product vs Income:

1. Higher the Income of the customer (Income >= 60000), higher the chances of the customer to purchase the KP781 product.

Product vs Miles:

1. If the customer expects to walk/run greater than 120 Miles per week, it is more likely that the customer will buy KP781 product.

What percent of customers have purchased KP281, KP481 or KP781 in a table?

```
aerofit_df3 = aerofit_df[['Product', 'Gender', 'MaritalStatus']].melt()
(aerofit_df3.groupby(['variable', 'value'])[['value']].count() / len(aerofit_df))*100
```

value

Observations regarding Gender:

1. Male had purchased the treadmills more as compared to Female.

Observation regarding Marital Status:

1. Partnered / Couples is observed purchase around 60% of the products.

Observations regarding Products:

1. According to the series released of Treadmills, the rate is decreasing as KP281 has more purchased stake around 44.44%.

Probability- marginal, conditional probability.

What is the probability of a male customer buying a KP781 treadmill?

```
# prob_of_male_buying_KP781
prod_purchased = aerofit_df[aerofit_df['Product']=='KP781']
```

```
prod_purchased.shape
```

```
(40, 9)
```

```
by_male = prod_purchased[prod_purchased['Gender']=='Male']
```

```
by_male.shape
```

```
(33, 9)
```

```
#So the probability that Male will purchase the KP781 treadmill is:
(by_male.shape[0]/prod_purchased.shape[0])*100
```

```
82.5
```

Its 82% chances that if a KP781 treadmill is bought, it is purchased by a Male

```
purchase_by = aerofit_df[aerofit_df['Gender']=='Male']
```

```
purchase_by.shape
```

```
(104, 9)
```

```
# If a Male walks in the store, the probability that it will purchase a KP781 is:
(prod_purchased.shape[0]/purchase_by.shape[0])*100
```

```
38.46153846153847
```

The probability of a Male walks in and buys a KP781 is 38%.

```
# prob_of_male_buying_KP781
prod_purchased_281 = aerofit_df[aerofit_df['Product']=='KP281']
```

```
purchase_by_281 = aerofit_df[aerofit_df['Gender']=='Male']
```

```
(prod_purchased_281.shape[0]/purchase_by_281.shape[0])*100
```

```
76.92307692307693
```

Probability of Male buying the KP281 product is near 77%.

```
aerofit_df.head(5)
```

1 to 5 of 5 entries Filter

index	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

Show 25 per page

Like what you see? Visit the data table notebook to learn more about interactive tables.

Actionable Insights:

1. As it seems that the higher versions are more preferable to higher age groups, so its insight might be the price tag of the product. It may not be affordable for low age groups. And as the series of treadmills are going more and more features are updating in treadmills, thats why the basic version i.e KP281 must be cheap as compare to other versions, hence its sales is higher. If the company brings a mini-version of treadmill that have updated features and might

have price tag similar to KP281, then it might be best-seller as more people can buy that product.

2. If we see the data of Person who is Partnered vs Person who is Single, it is observed that Partnered seems to be more active in Purchasing Advanced Models of Treadmills and there quantity activeness in other things are also well-grouped. If near future, company comes up with some new ideas for Couples then it might be more beneficial as they would be there first and only target audience and might ends-up with a great response.