AEROFIT CASE STUDY

About Aerofit

Aerofit is a leading brand in the field of fitness equipment. Aerofit provides a product range including machines such as treadmills, exercise bikes, gym equipment, and fitness accessories to cater to the needs of all categories of people.

Business Problem

The market research team at AeroFit wants to identify the characteristics of the target audience for each type of treadmill offered by the company, to provide a better recommendation of the treadmills to the new customers. The team decides to investigate whether there are differences across the product with respect to customer characteristics.

The Dataset used

Aerofit_treadmill.csv

Details of columns

Product Purchased:	KP281, KP481, or KP781
Age:	In years
Gender:	Male/Female
Education:	In years
MaritalStatus:	Single or partnered
Usage:	The average number of times the customer plans to use the treadmill each week.
Income:	Annual income (in $)
Fitness:	Self-rated fitness on a 1-to-5 scale, where 1 is the poor shape and 5 is the excellent shape.
Miles:	The average number of miles the customer expects to walk/run each week

Product Portfolio:

- The KP281 is an entry-level treadmill that sells for $1,500.
- The KP481 is for mid-level runners that sell for $1,750.
- The KP781 treadmill is having advanced features that sell for $2,500.

What good looks like?

Importing the dataset and performing usual data analysis steps like checking the structure & characteristics of the dataset.

Here the analysis is performed on **Google Colab Notebook.**

```
import pandas as pd
df = pd.read_csv("aerofit.txt")
df.head()
```

	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

The text file is converted to pandas dataframe for better visualization.

```
[ ] df.shape     #this shows the shape of data frame with number of rows and columns

    (180, 9)

[ ] df.info()   # this gives column details ... here its clear that there are no null values

    <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
```

Inference:

The DataFrame contains 180 rows and 9 columns (output using shape attribute)

Df.info shows the column’s datatype details … . Here few columns are categorical (Product, Gender ,Marital Status) and few are continuous.(Age , Education , Income , fitness, Miles).

It can be clearly inferred from above and the below code that there are no null values in any of the columns ..

```
df.isnull().any()     #to check if there are any null values
```

```
Product          False
Age              False
Gender           False
Education        False
MaritalStatus    False
Usage            False
Fitness          False
Income           False
Miles            False
dtype: bool
```

1. To detect Outliers

```
df1 = df.describe(include = "all")
df1
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
count	180	180.000000	180	180.000000	180	180.000000	180.000000	180.000000	180.000000
unique	3	NaN	2	NaN	2	NaN	NaN	NaN	NaN
top	KP281	NaN	Male	NaN	Partnered	NaN	NaN	NaN	NaN
freq	80	NaN	104	NaN	107	NaN	NaN	NaN	NaN
mean	NaN	28.788889	NaN	15.572222	NaN	3.455556	3.311111	53719.577778	103.194444
std	NaN	6.943498	NaN	1.617055	NaN	1.084797	0.958869	16506.684226	51.863605
min	NaN	18.000000	NaN	12.000000	NaN	2.000000	1.000000	29562.000000	21.000000
25%	NaN	24.000000	NaN	14.000000	NaN	3.000000	3.000000	44058.750000	66.000000
50%	NaN	26.000000	NaN	16.000000	NaN	3.000000	3.000000	50596.500000	94.000000
75%	NaN	33.000000	NaN	16.000000	NaN	4.000000	4.000000	58668.000000	114.750000
max	NaN	50.000000	NaN	21.000000	NaN	7.000000	5.000000	104581.000000	360.000000

✓ 0s completed at 08:50

Inference:

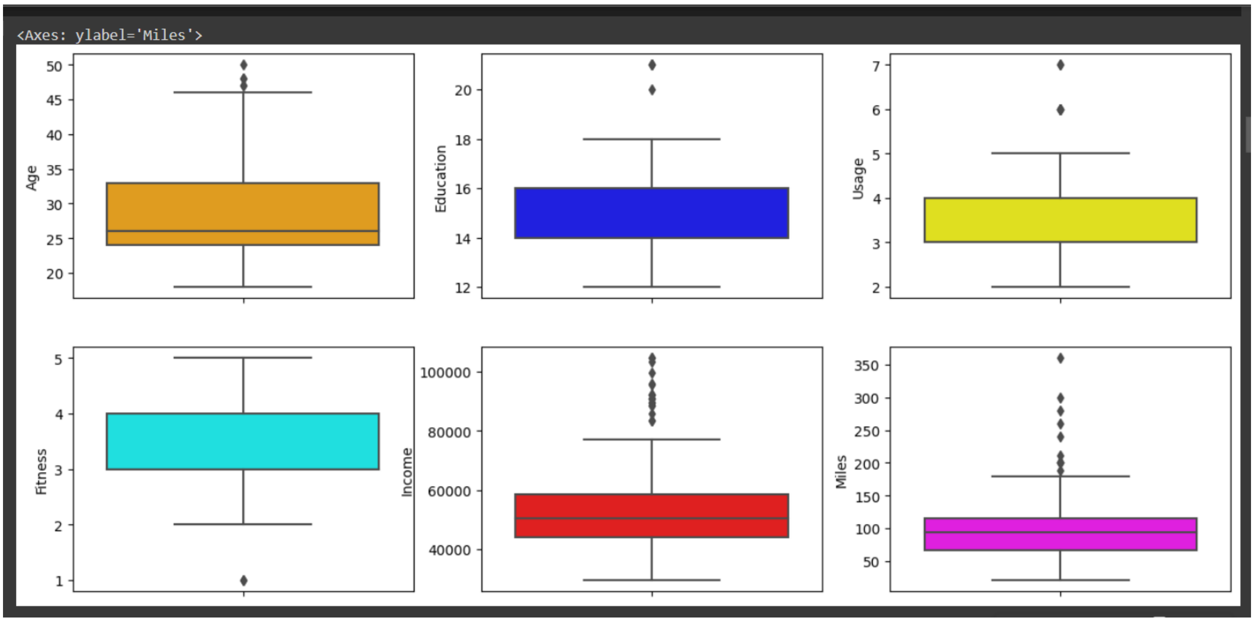
- Using the describe method . The various values are obtained .. Here , the columns Income and miles shows a huge difference in mean and median (50%) . This clearly shows that there are few outliers . Also the std values is considerably high compared to other columns.
- The top product purchased by customers is KP281.
- Also the male customers have purchased the most products compared to females
- The customers' age group is 18-50 . With mean of 28 years .. Also . 75% of the customers are <=33 years .. Surely there are outliers.
- The income range is 29500-104600. 75% of the population has income less than 60000.
- This means there are definitely outliers.

OUTLIERS DETECTION USING BOXPLOTS

```
[32] import matplotlib.pyplot as plt
     import seaborn as sns
     fig, axis = plt.subplots(nrows = 2, ncols = 3 , figsize=(15,7))
     sns.boxplot(data = df , y ="Age" , ax =axis[0,0], color = "orange")
     sns.boxplot(data = df , y ="Education" , ax =axis[0,1], color = "blue")
     sns.boxplot(data = df , y ="Usage" , ax =axis[0,2], color = "yellow")
     sns.boxplot(data = df , y ="Fitness" , ax =axis[1,0], color = "aqua")
     sns.boxplot(data = df , y ="Income" , ax =axis[1,1], color = "red")
     sns.boxplot(data = df , y ="Miles" , ax =axis[1,2], color = "magenta")
```

Various boxplots are plotted using seaborn lib. On the columns(Age , Education , Usage , Fitness , Income , Miles)

Output :



Inference:

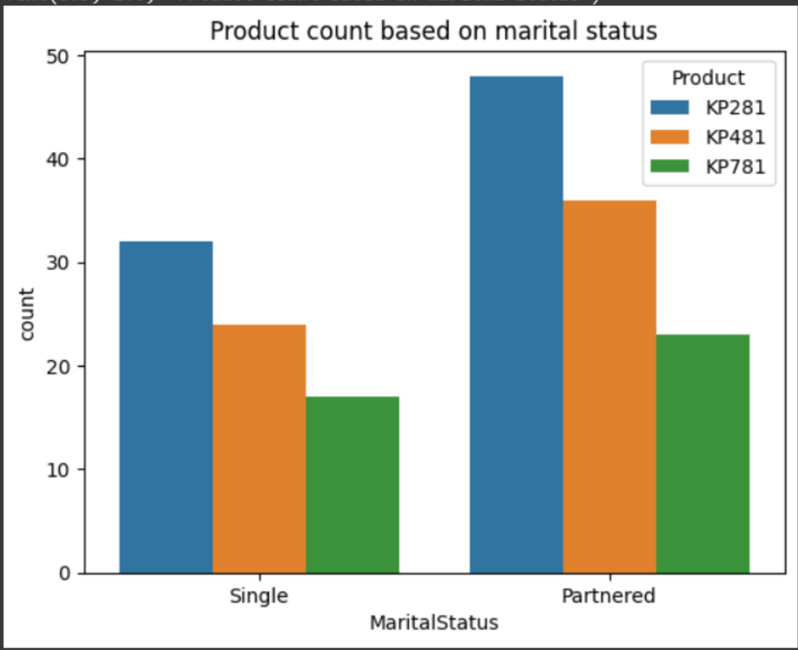
- Here , age , Income and miles have most outliers .
- Average age of customers is nearly 28.. The average income is nearly 50000 and the average miles covered is close to 100.
- The maximum education years is 18 with 2 outlier .
- The maximum age is 45 with few outliers.75%tile of the population is less than 33 years.
- 75%tile of the population earns less than 60000.

1. Check if features like marital status, age have any effect on the product purchased (using countplot, histplots, boxplots etc)

EFFECT OF MARITAL STATUS ON PRODUCT PURCHASED

```
sns.countplot(x = "MaritalStatus", data = df , hue = "Product")
plt.title("Product count based on marital status")  #EFFECT OF MARITAL STATUS ON PRODUCT PURCHASED
```

```
Text(0.5, 1.0, 'Product count based on marital status')
```

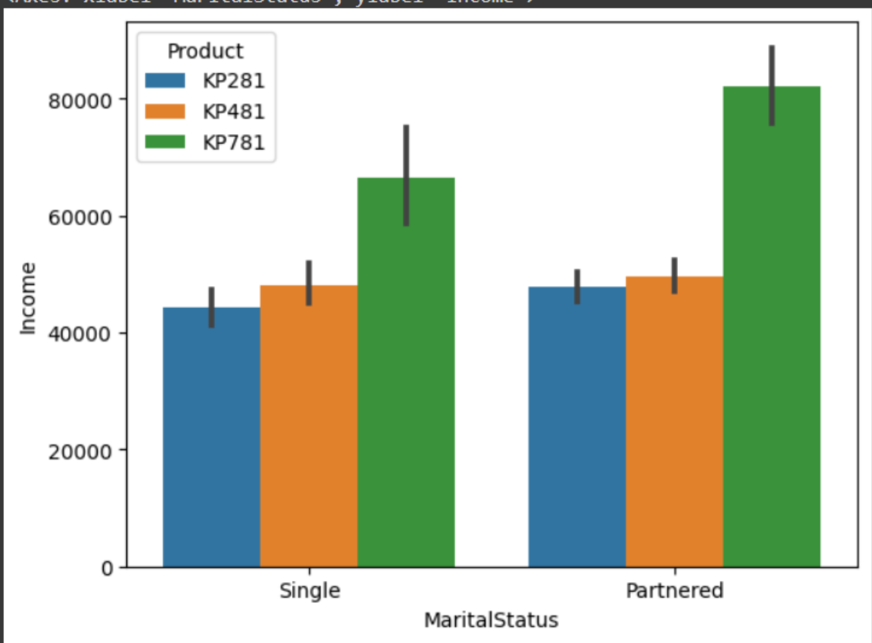


Inference:

- The above plot clearly shows that partnered customers are more health enthusiast than single ones .
- All the products purchased by partnered customers are more than single ones . With the sale of KP281 being the highest.

```
sns.barplot(x = "MaritalStatus", y = "Income",  data = df , hue = "Product")
```

```
<Axes: xlabel='MaritalStatus', ylabel='Income'>
```

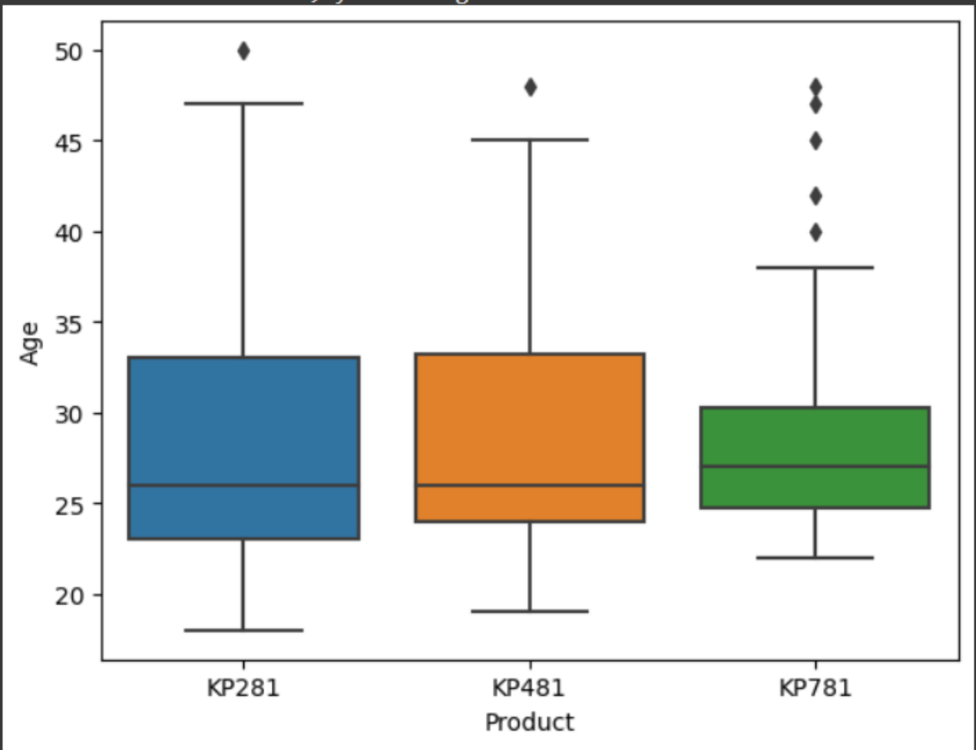


Inference : The reason of more patterned customers can be their income slab .. Which is comparatively higher than single ones .

EFFECT OF AGE ON PRODUCT PURCHASED

```
import seaborn as sns
sns.boxplot( x = "Product",  y = "Age",  data = df)
```

```
<Axes: xlabel='Product', ylabel='Age'>
```

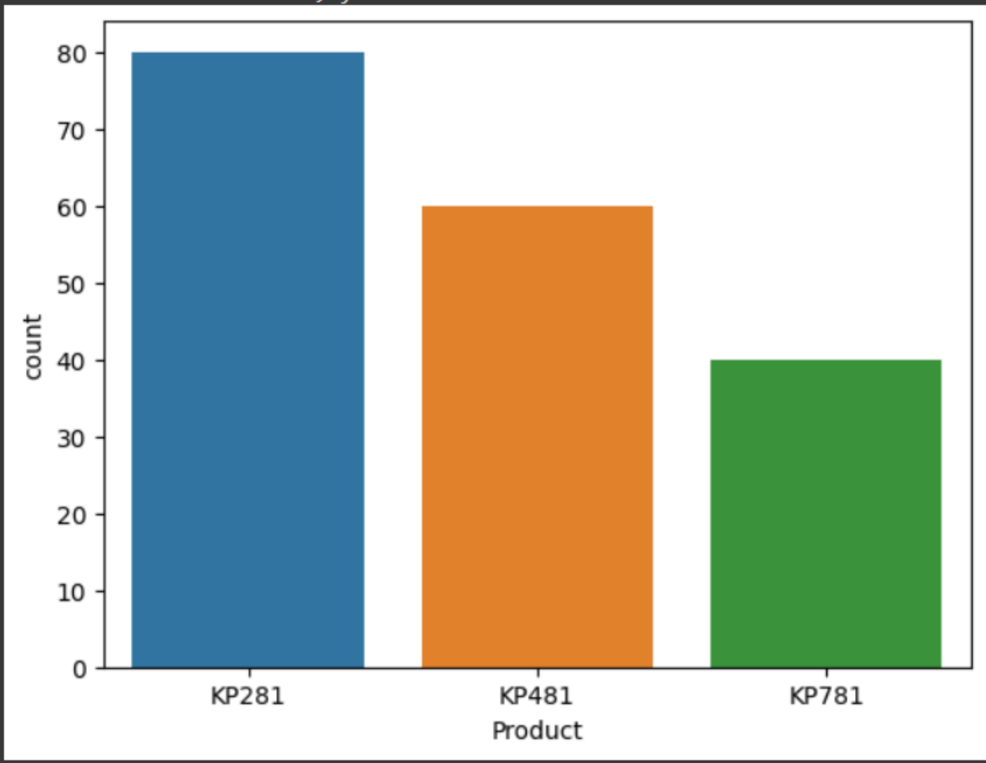


Inference :

- The age group for KP281 is largest … with minimum of 24 and maximum 48. 75 % of KP281 customers are less than 33 years. which is same for KP481.
- Here KP781 has difference age group of customers ranging from 24 – 37 years . With several outliers..

```
sns.countplot(x= "Product" , data = df)
```

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

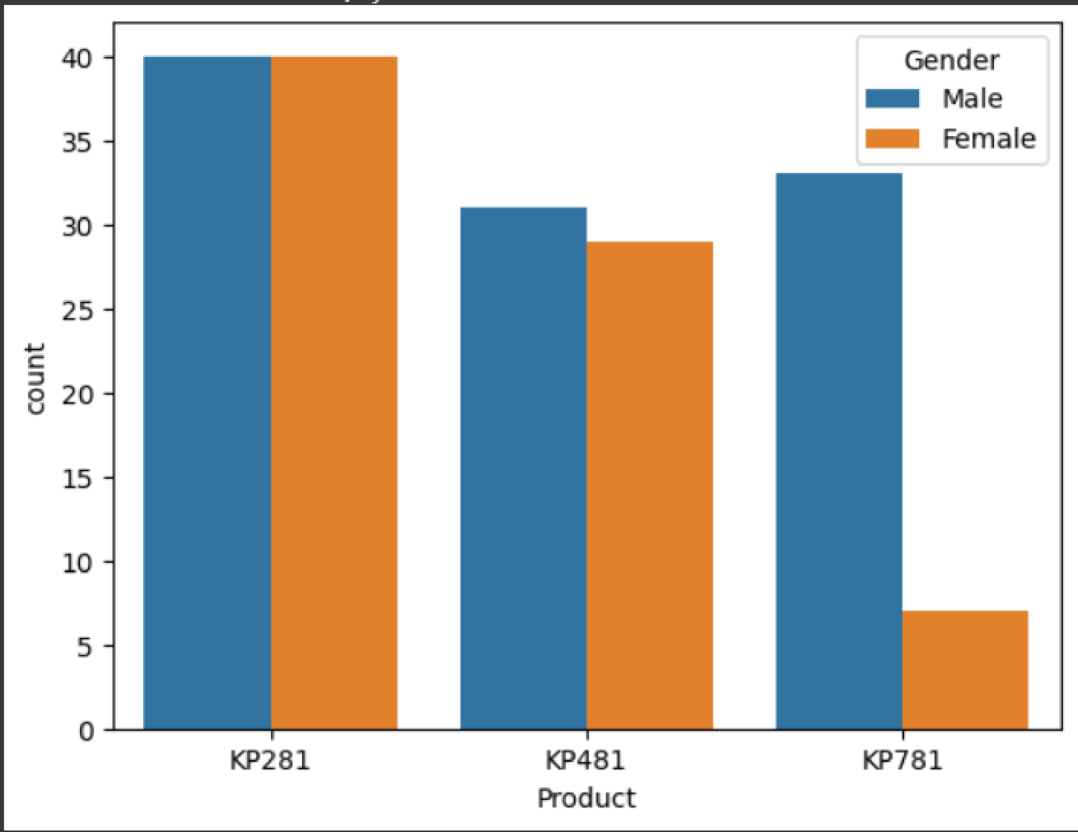


Inference : The above plot indicates that KP781 is less popular product among customers .

EFFECT OF GENDER ON PRODUCT PURCHASED

```python
sns.countplot(x= "Product" , data = df, hue = "Gender")
```

<Axes: xlabel='Product', ylabel='count'>



Inference : KP781 is less popular among the female customers.

Representing the marginal probability like - what percent of customers have purchased KP281, KP481, or KP781 in a table

By using Pandas Crosstab

```python
pd.crosstab(index=df['Product'], columns=df['Gender'], margins = "True")
```

Gender	Female	Male	All
Product			
KP281	40	40	80
KP481	29	31	60
KP781	7	33	40
All	76	104	180

To find the marginal probability . Normalizing the tab

```
pd.crosstab(index=df['Product'], columns=df['Gender'], normalize = "all")
```

Gender Product	Female	Male
KP281	0.222222	0.222222
KP481	0.161111	0.172222
KP781	0.038889	0.183333

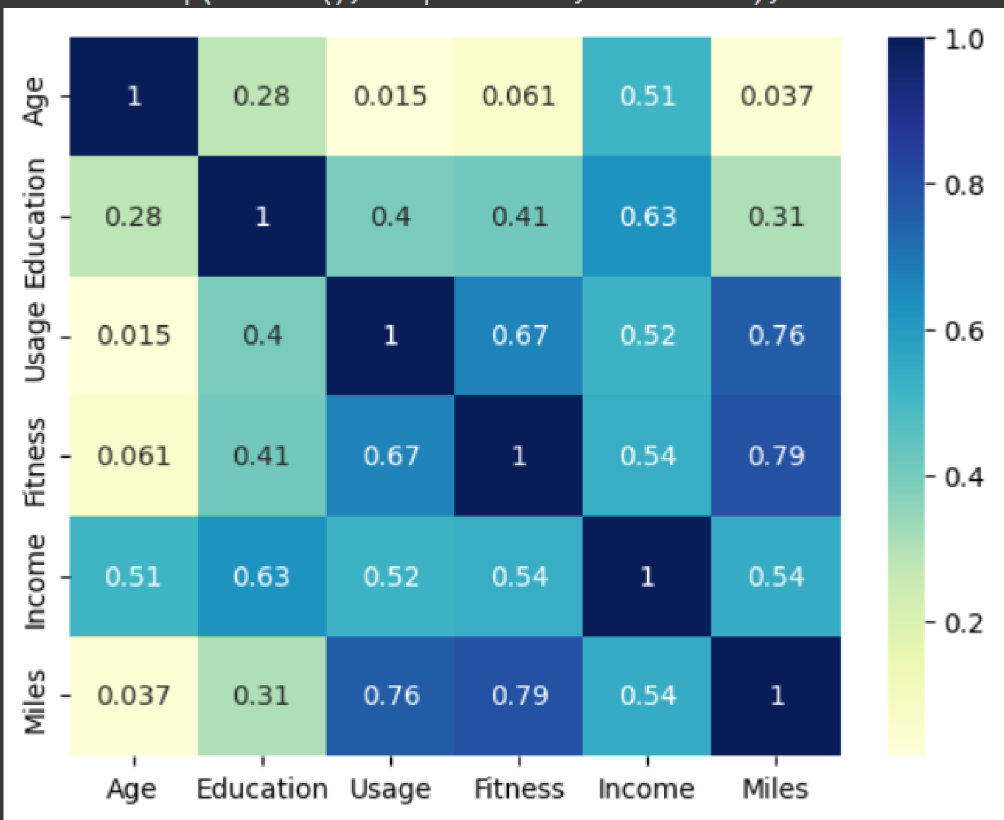
Inference:

- For KP281 the probability if 50/50 for both gender.
- Also , for KP481 I s nearly similar for both the genders
- However for KP781 .. Female customers are less probable to buy the product compare to male customers.
- Overall KP281 has the highest probability of purchase
- The probability of male customers buying KP781 is 0.183

Checking correlation among different factors using heat maps or pair plots.

```
sns.heatmap(df.corr(), cmap="YlGnBu", annot=True);
```

```
<ipython-input-48-6a09541e6a1a>:3: FutureWarning: The default value of numeric_
  sns.heatmap(df.corr(), cmap="YlGnBu", annot=True);
```



Inference :

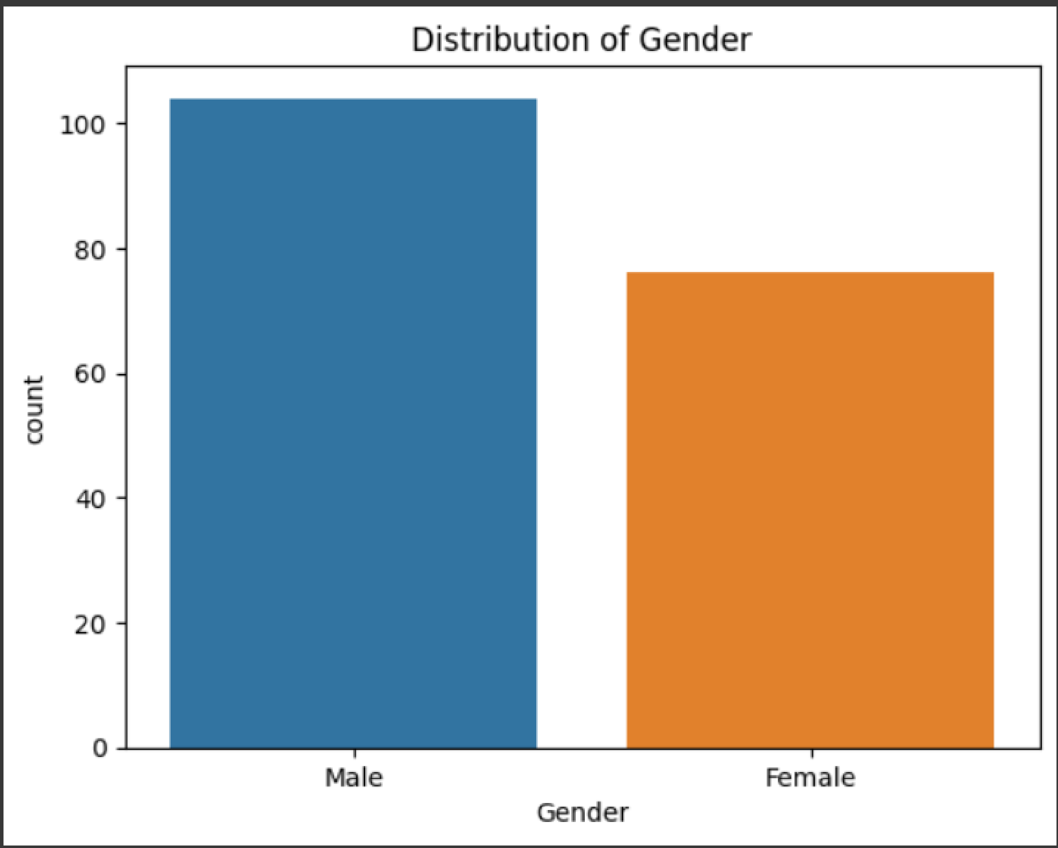
- Fitness and Usage are having strong positive correlation.
- Variables such age and miles , age and Income are having strong negative correlations.
- There are several variables that have no correlation and whose correlation value is near 0.
- Generally speaking, a Pearson correlation coefficient value greater than 0.7 indicates the presence of **multi-collinearity.**

Customer Profiling and Segmentation – An Analytical Approach To Business Strategy

To check the distribution of customer's gender in the dataset

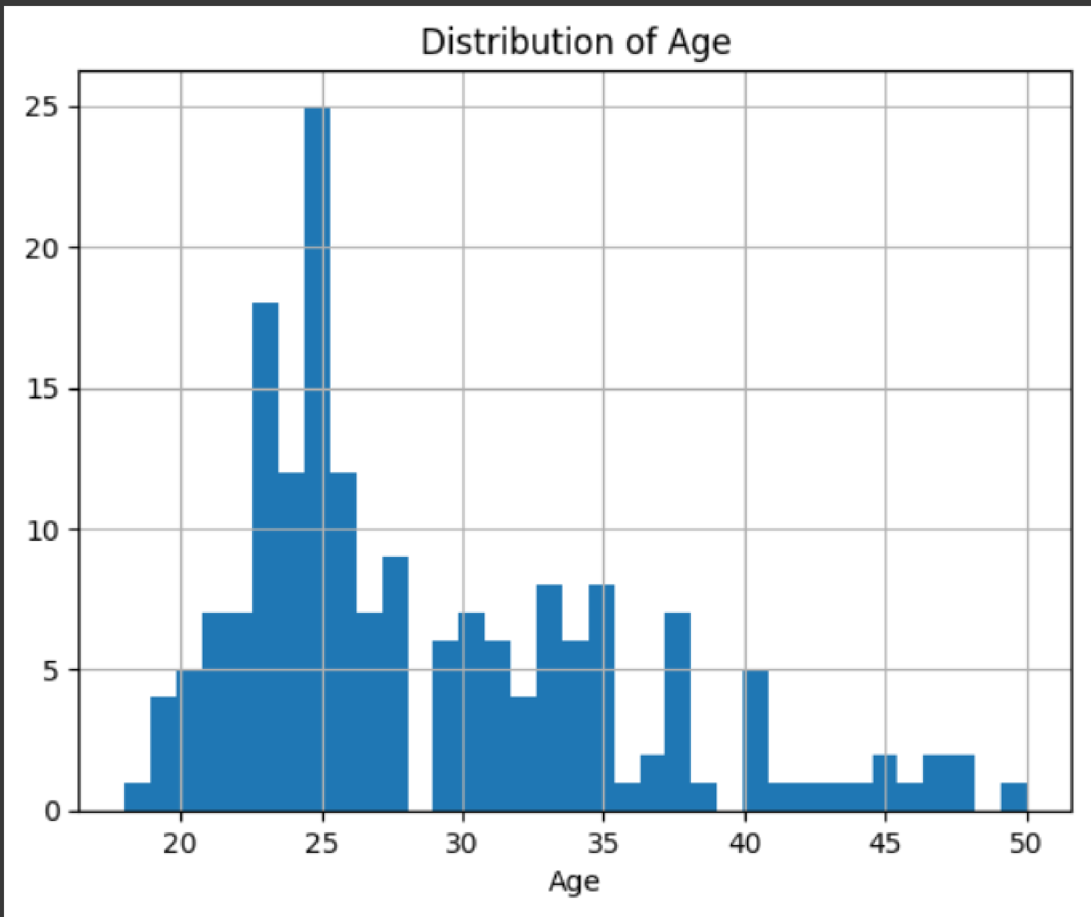
UNIVARIATE ANALYSIS

```
# See the distribution of gender to recognize different distributions
sns.countplot(x='Gender', data=df);
plt.title('Distribution of Gender');
```



To check the distribution of customer's age in the dataset

```
df.hist('Age', bins=35);
plt.title('Distribution of Age');
plt.xlabel('Age');
```



```
[54] from scipy.stats import norm
     df["Age"].describe()
```

```
count    180.000000
mean      28.788889
std        6.943498
min       18.000000
25%       24.000000
50%       26.000000
75%       33.000000
max       50.000000
Name: Age, dtype: float64
```

```
[57] new = norm( 29,7)
```

```
[58] new.cdf(40)
```

```
0.9419584331306725
```

```
[60] new.cdf(18)
```

```
0.05804156686932752
```

```
new.cdf(40)-new.cdf(18)
```

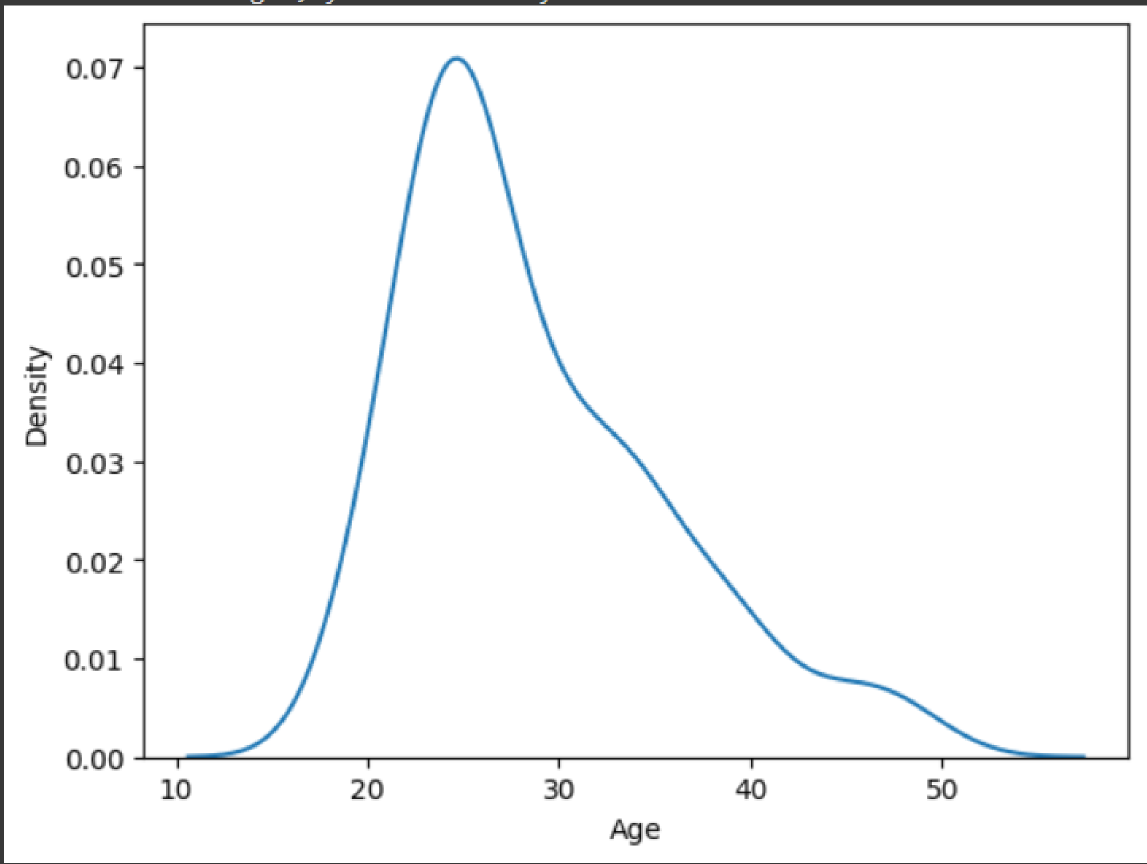
```
0.883916866261345
```

Inference:

Close to 88% of customers are between age group 18-40

```
sns.kdeplot(df["Age"])
```

```
<Axes: xlabel='Age', ylabel='Density'>
```



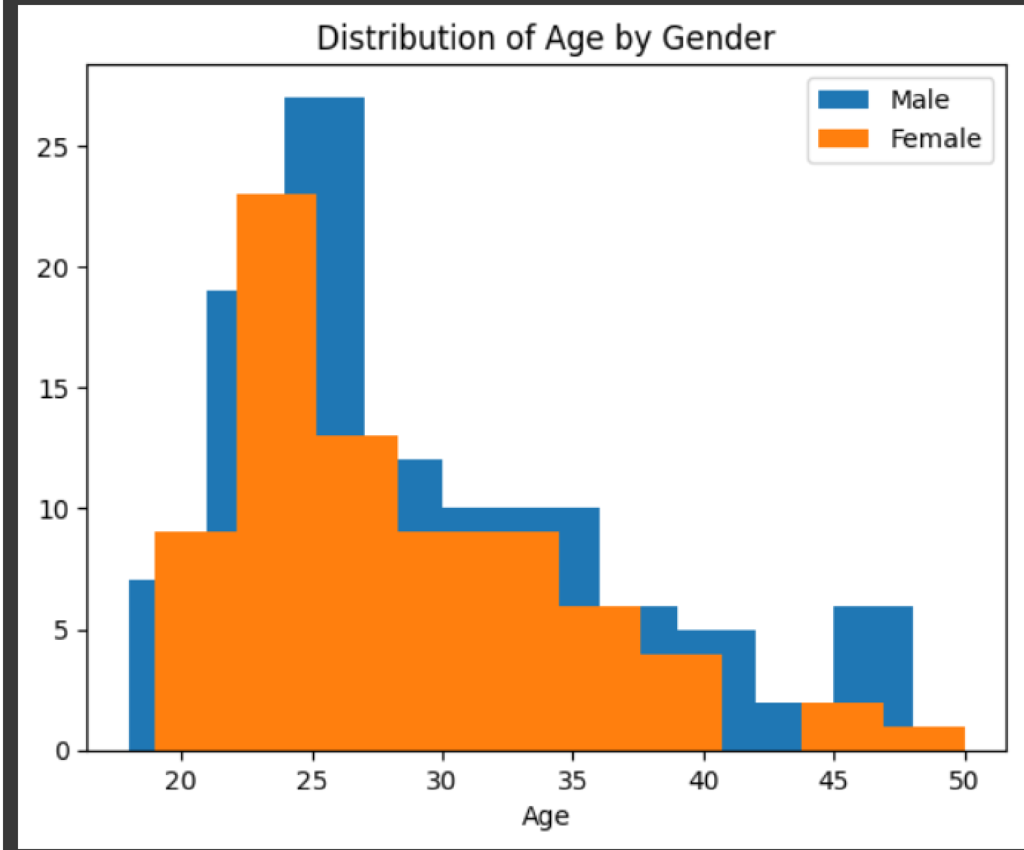
Inference

The ages are mostly between 23 and 35. Recalling the describe() call results this makes sense. The average age was 28.8. There are less older customers, so this distribution is right-skewed because of its longer right tail. This could be because of more aged customers are not comfortable on treadmill as the younger customers.

we can add detail to this by overlaying two histograms, creating one age histogram for each gender.

<u>To check the distribution of customer's age by Gender in the dataset</u>

```
plt.hist('Age', data=df[df['Gender'] == 'Male'],  label='Male');
plt.hist('Age', data=df[df['Gender'] == 'Female'],  label='Female');
plt.title('Distribution of Age by Gender');
plt.xlabel('Age');
plt.legend();
```

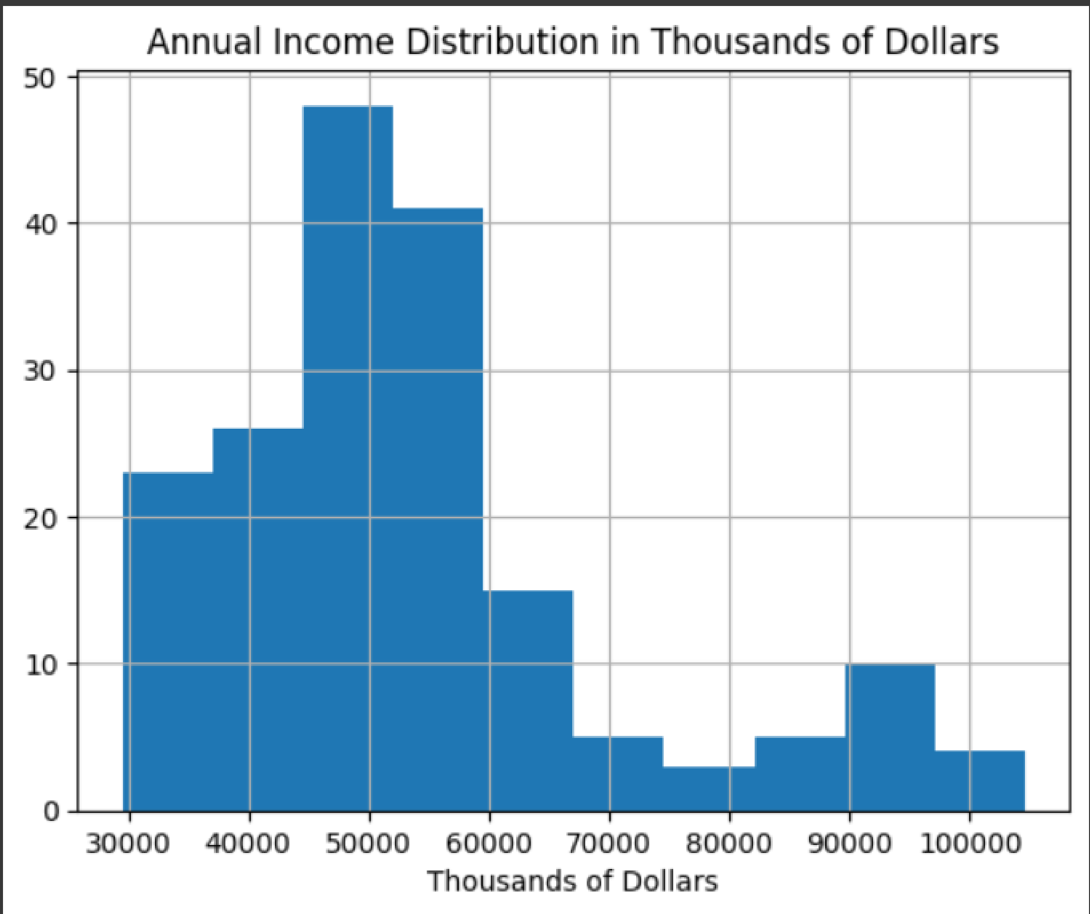


Inference:

```
The women in this data set tended to be younger than the men. You can see
the spike around the age of 22-25 for the women, There are also more
middle-aged women in this data set than men. There are more senior men in
the 45-50 year old bucket.
```

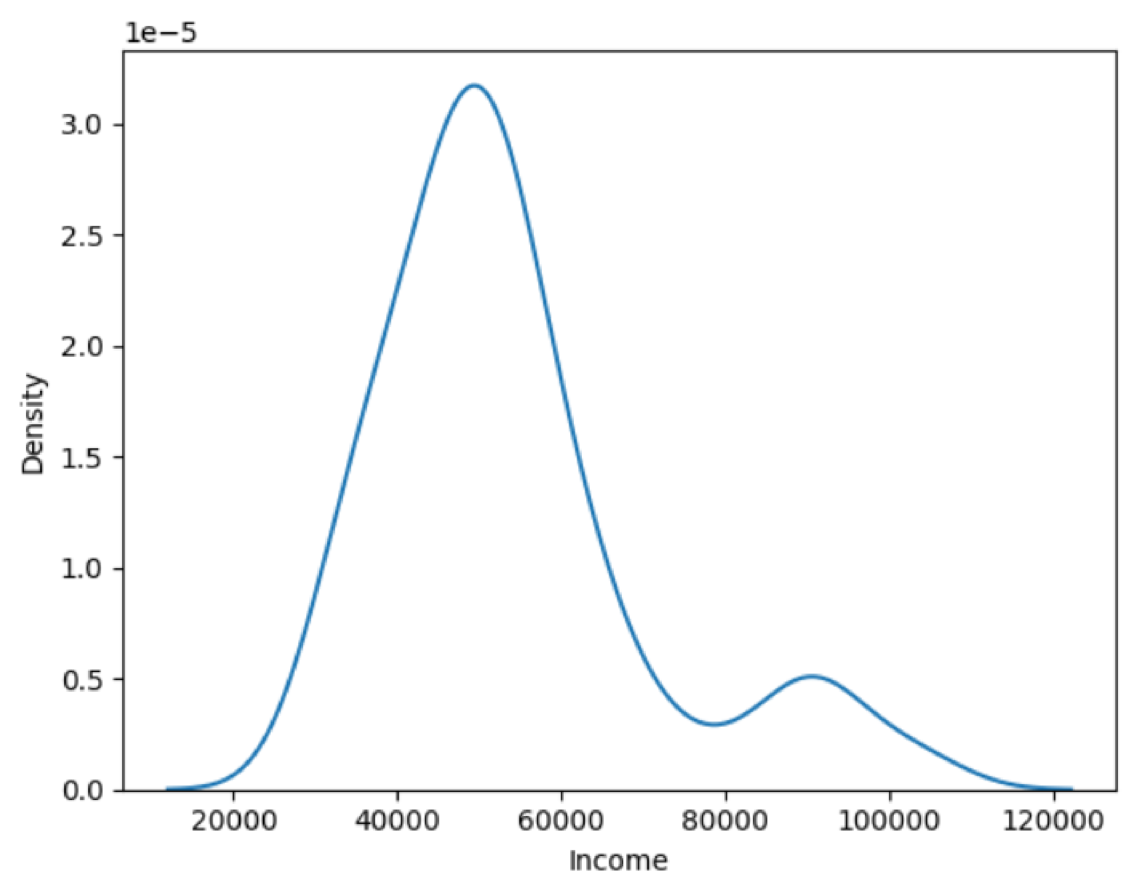
__**To check the distribution of customer's annual income in the dataset**__

```
df.hist('Income');
plt.title('Annual Income Distribution in Thousands of Dollars');
plt.xlabel('Thousands of Dollars');
```



```python
sns.kdeplot(df["Income"])
```

```
<Axes: xlabel='Income', ylabel='Density'>
```



```
[62] from scipy.stats import norm
     df["Income"].describe()
```

```
count       180.000000
mean      53719.577778
std       16506.684226
min       29562.000000
25%       44058.750000
50%       50596.500000
75%       58668.000000
max      104581.000000
Name: Income, dtype: float64
```

```
[63] new = norm(53719,16506.6)
```

```
[64] new.cdf(20000)
```

```
0.020538167665540005
```

```
[65] new.cdf(80000)
```

```
0.9443246120184545
```

```
new.cdf(80000)-new.cdf(20000)
```

```
0.9237864443529146
```

Inference:

Most customer are in the range 50000-60000 . Also the distribution is skewed right.. Most of the customes are below 60000. Close to 92% of the customer are in the income bracket 20000-80000 The plot is skewed on right due to presence of outliers

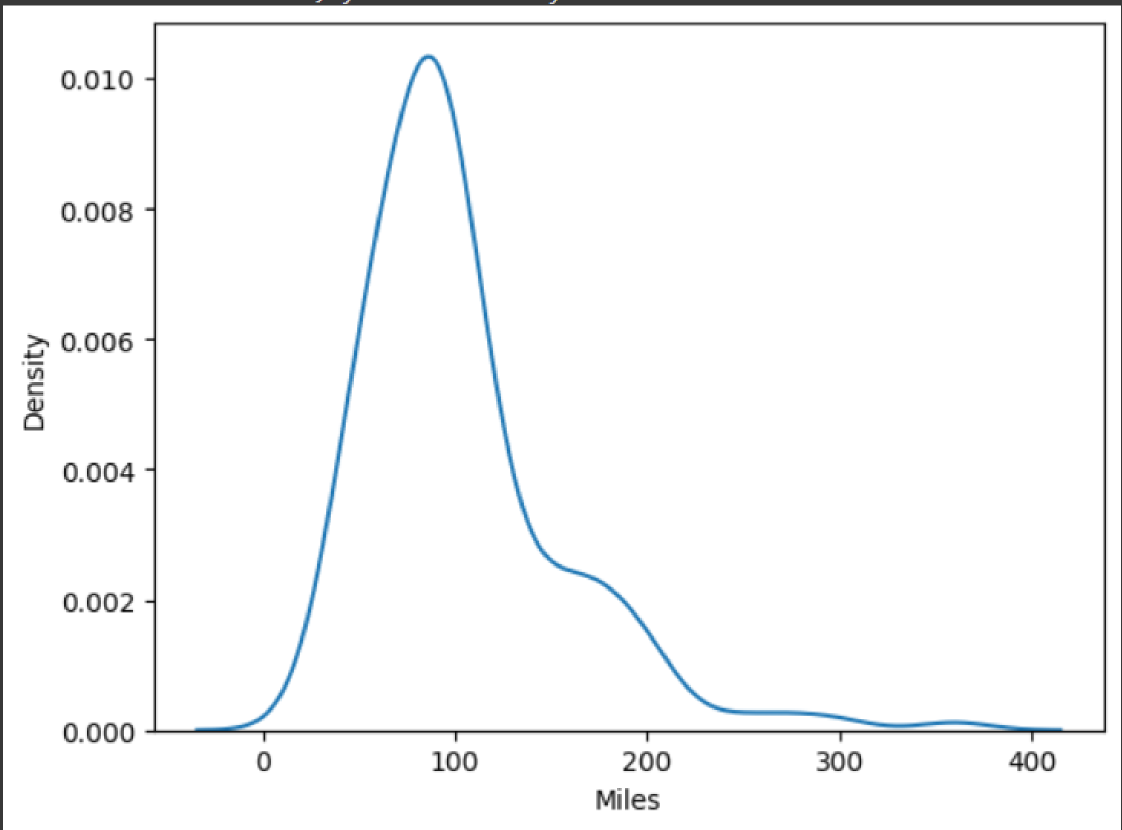
```
[67] from scipy.stats import norm
     df["Miles"].describe()
```

```
count    180.000000
mean     103.194444
std       51.863605
min       21.000000
25%       66.000000
50%       94.000000
75%      114.750000
max      360.000000
Name: Miles, dtype: float64
```

```
[69]
```

```
sns.kdeplot(df["Miles"])
```

```
<Axes: xlabel='Miles', ylabel='Density'>
```



```
new = norm(103,51.8)
```

```
new.cdf(0)
```

```
0.023382795731984762
```

```
new.cdf(200)
```

```
0.9694372722610985
```

```
new.cdf(200)-new.cdf(0)
```

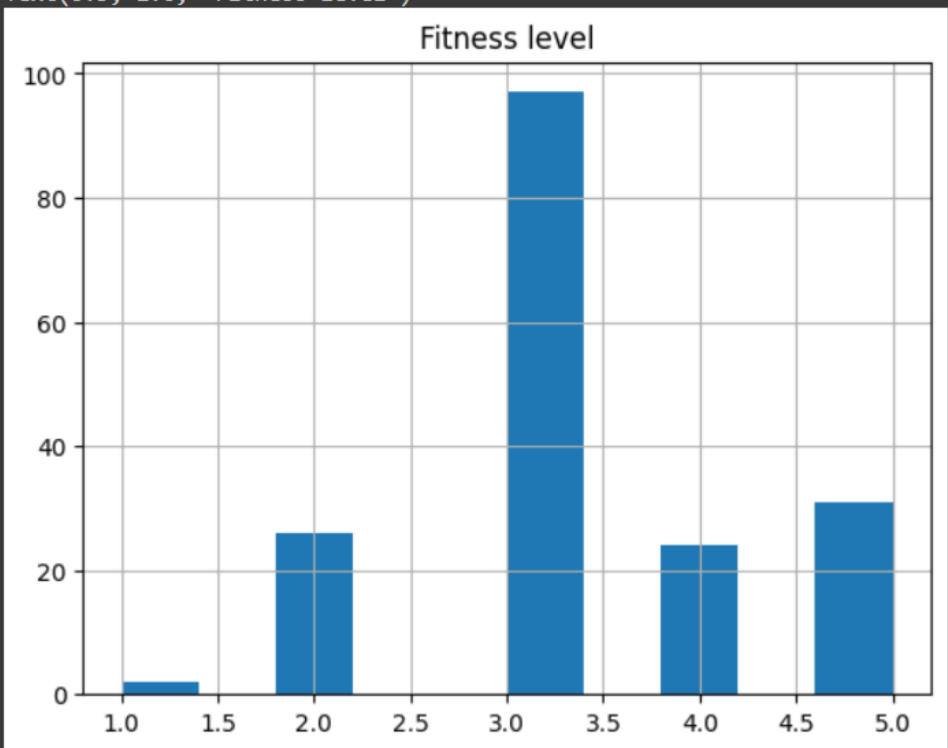
```
0.9460544765291138
```

Nearly 94% of the miles expected per week are 200 miles

FITNESS LEVEL DISTRIBUTION

```
df.hist('Fitness')
plt.title("Fitness level")
```

```
Text(0.5, 1.0, 'Fitness level')
```



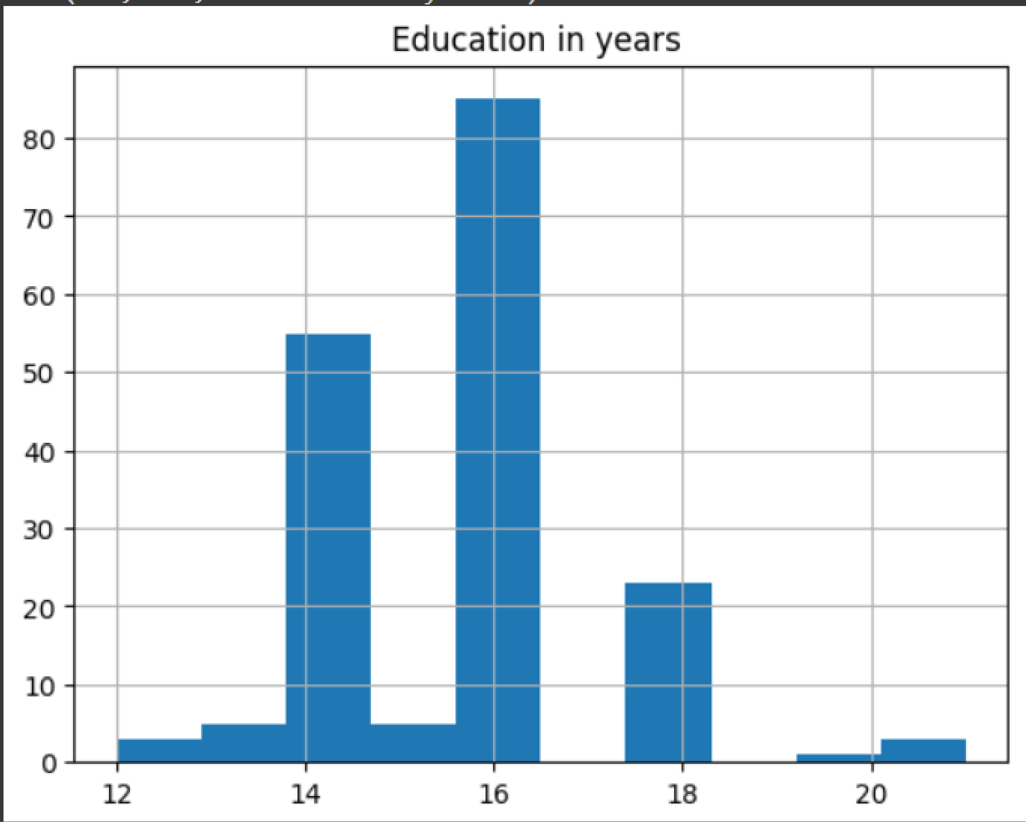
0s completed at 13:05

Inference : Most customer are average in shape as per the rating ..

EDUCATION LEVEL DISTRIBUTION

```
df.hist('Education')
plt.title("Education in years ")
```

```
Text(0.5, 1.0, 'Education in years ')
```

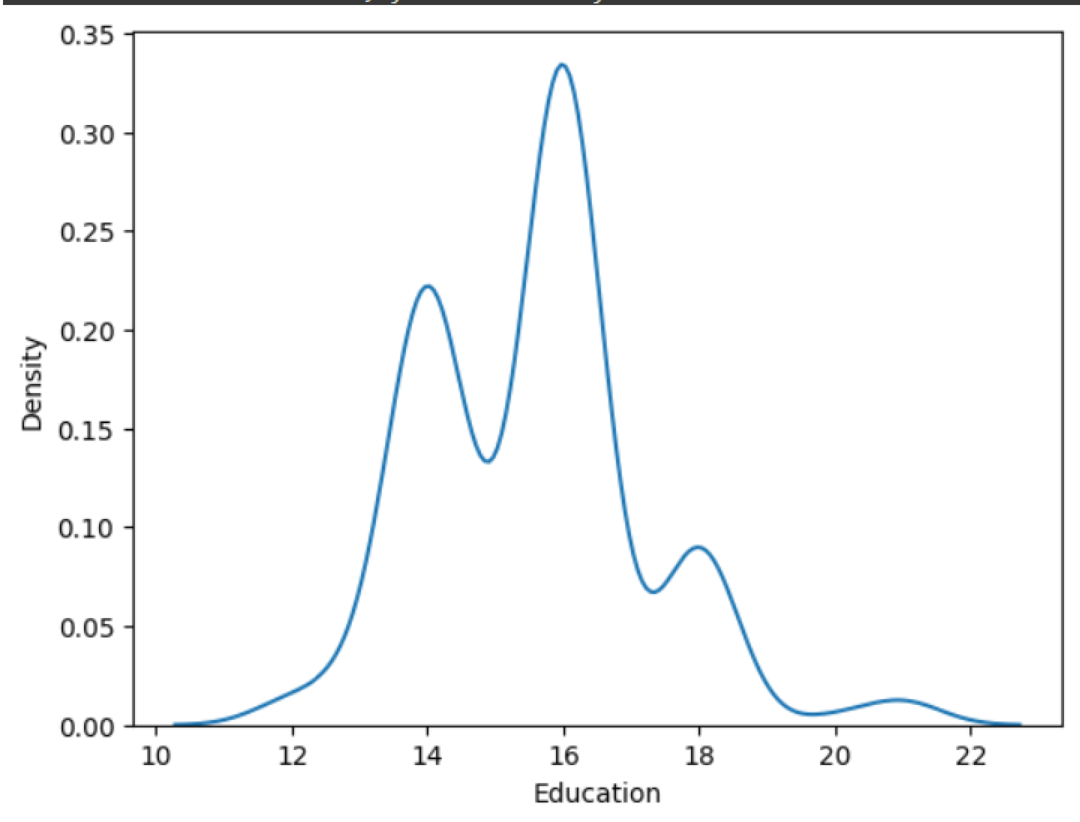


✓ 0s completed at 13:09

Inference : Most of the customer have completed 16 years of education . Possible graduates and above. With possible outliers

```
sns.kdeplot(df["Education"])
```

```
<Axes: xlabel='Education', ylabel='Density'>
```



ACTIONABLE INSIGHTS

- From the above analysis. We can infer that poduct KP281 is the most popular product among men and women , single and partenerd, and is puirchased by both average income people and high income people as it is cheaper and user friendly with not much functions
- Hence it is the most profit giving product.
- Most of the customers with this product are below 33 years ..So mostly this product is popular among youngsters and working people in both genders . More features can be added to the same product to increase sales .
- The product KP781 is the least selling product . It is less popular among the female customers .
- It has advanced features which are not required as per women customers . Hence features can be modified so that it can attract more female customers .

REFERENCE COLAB LINK:

https://colab.research.google.com/drive/1I39dR7cGlZxXNahpZFudY8VcG4R9k3DT?usp=sharing