



Business Case: Aerofit- Data Exploration and Visualization

Business Case: AeroFit - Data Exploration and Visualisation



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.

- ✓ Perform descriptive analytics to create a customer profile for each AeroFit treadmill product by developing appropriate tables and charts.
- ✓ For each AeroFit treadmill product, construct two-way contingency tables and compute all conditional and marginal probabilities along with their insights/impact on the business.

1. Defining Problem Statement and Analysing basic metrics. Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary.

The business case envisages analysing of the given AeroFit dataset and using different methods like visual and non-visual analysis formulate insights which will help AeroFit in decision making. The business case leverages Python's robust data analytics and visualization capabilities to extract valuable insights from the data set, purchasing patterns, and product performance metrics. By harnessing Python libraries such as Pandas, NumPy, Matplotlib, and Seaborn, the case aims to gain a comprehensive understanding of user preferences, market trends, and market dynamics. Through data-driven analysis and visualization techniques, the case tries to optimize content recommendations to cater to diverse customer segments. This data-centric approach empowers AeroFit to make informed decisions, drive customer retention and growth, and maintain a leading position in the ever-evolving fitness equipment industry landscape.

The data set is about three types of treadmills.

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

The data set have the following columns:

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

The data set is downloaded as 'aerofit_treadmill.csv' and saved as dataframe named 'df'.

```
[3] !wget https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/125/original/aerofit_treadmill.csv
--2024-05-12 07:29:04-- https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/125/original/aerofit_treadmill.csv
Resolving d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)... 108.157.172.10, 108.157.172.183, 108.157.172.173, ...
Connecting to d2beiqkhq929f0.cloudfront.net (d2beiqkhq929f0.cloudfront.net)|108.157.172.10|:443... connected.
HTTP request sent, awaiting response... 200 OK
Length: 7279 (7.1K) [text/plain]
Saving to: 'aerofit_treadmill.csv'

aerofit_treadmill.c 100%[=====>] 7.11K --KB/s in 0s

2024-05-12 07:29:04 (39.1 MB/s) - 'aerofit_treadmill.csv' saved [7279/7279]
```

```
[2] import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
```

```
[4] df=pd.read_csv('aerofit_treadmill.csv')
```

The data sample is observed by `df.head()`

```
[6] 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 data about the products is divided into 9 columns and there are 180 rows in the dataset.

Shape of the dataframe : `df.shape` showed

```
[8] df.shape
```

```
(180, 9)
```

The basic information about dataframe. `df.info()`

```
[9] 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
```

Data type of 3 of the 9 columns are object type, other 6 columns being int64.

Detecting missing values by `isna()`.

```
[11] df.isna().sum()

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

It shows there are no null values or missing values in the dataset.

2. Non-Graphical Analysis: Value counts and unique attributes.

The dataset shows there are 80 purchases for product KP281, 60 for KP481 and 40 for KP 781

```
▶ df['Product'].value_counts()

➡ Product
KP281    80
KP481    60
KP781    40
Name: count, dtype: int64
```

The customers are of age 18 to 50 years.

```
[7] print(df['Age'].min(),df['Age'].max())

➡ 18 50
```


The data contains customers of income of about 30,000 \$ per annum to one lakh \$ per annum.

```
[15] print(df['Income'].min(),df['Income'].max())

➡ 29562 104581
```

The data shows that among the 180 customers, 104 are male and 76 are female.


```
[16] df['Gender'].value_counts()
```



```
Gender
Male      104
Female     76
Name: count, dtype: int64
```

The data shows that among the 180 customers, 107 are partnered and 73 are single.

```
[17] df['MaritalStatus'].value_counts()
```




```
MaritalStatus
Partnered    107
Single        73
Name: count, dtype: int64
```

The age and salary are divided into bins for easiness of finding correlations.



```
[60] bins = [15, 21, 26, 31, 36, 41,46,51]
      labels = ['15-20','20-25','25-30','30-35','35-40','40-45','45-50']
      df['AgeBin'] = pd.cut(df['Age'], bins=bins, labels=labels)
      ibins = [25000, 40000, 55000, 70000, 85000,105000]
      ilabels = ['25k-40k','40k-55k','55k-70k','70k-85k','85k-105k']
      df['IncomeBin'] = pd.cut(df['Income'], bins=ibins, labels=ilabels)
```

The resultant dataframe is:

```
[61] df.head()
```



	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	AgeBin	IncomeBin
0	KP281	18	Male	14	Single	3	4	29562	112	15-20	25k-40k
1	KP281	19	Male	15	Single	2	3	31836	75	15-20	25k-40k
2	KP281	19	Female	14	Partnered	4	3	30699	66	15-20	25k-40k
3	KP281	19	Male	12	Single	3	3	32973	85	15-20	25k-40k
4	KP281	20	Male	13	Partnered	4	2	35247	47	15-20	25k-40k

The Correlation between gender and age in the three varieties of products is found out by crosstab.

```
[62] pd.crosstab(df['Product'],[df['Gender'],df['AgeBin']])
```

Gender	Female							Male						
AgeBin	15-20	20-25	25-30	30-35	35-40	40-45	45-50	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Product														
KP281	4	16	9	6	2	2	1	6	15	7	4	6	1	1
KP481	2	12	5	6	4	0	0	5	12	2	8	2	1	1
KP781	0	4	2	1	0	0	0	0	15	10	2	2	2	2

The data shows in both men and women people of age 20-30 are most probable to buy the Aerofit products.

The correlation between age and income levels are described using crosstab.

```
[ ] pd.crosstab(df['AgeBin'],df['IncomeBin'])
```

	IncomeBin	25k-40k	40k-55k	55k-70k	70k-85k	85k-105k
	AgeBin					
	15-20	17	0	0	0	0
	20-25	14	49	8	3	0
	25-30	0	21	3	2	9
	30-35	0	17	7	0	3
	35-40	1	5	8	1	1
	40-45	0	2	2	0	2
	45-50	0	0	3	0	2

The greatest number of products are bought by income level 40-50 thousand dollars per annum level and in 20-35 years.

The data is divided into three dataframes df2,df4 and df7 separating the data of products KP281,KP481 and KP781 respectively.

```
[14] df2=df[df['Product']=='KP281'].reset_index(drop=True)
df2.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

```
[12] df4=df[df['Product']=='KP481'].reset_index(drop=True)
df4.head()
```



	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP481	19	Male	14	Single	3	3	31836	64
1	KP481	20	Male	14	Single	2	3	32973	53
2	KP481	20	Female	14	Partnered	3	3	34110	106
3	KP481	20	Male	14	Single	3	3	38658	95
4	KP481	21	Female	14	Partnered	5	4	34110	212



```
[13] df7=df[df['Product']=='KP781'].reset_index(drop=True)
df7.head()
```



	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP781	22	Male	14	Single	4	3	48658	106
1	KP781	22	Male	16	Single	3	5	54781	120
2	KP781	22	Male	18	Single	4	5	48556	200
3	KP781	23	Male	16	Single	4	5	58516	140
4	KP781	23	Female	18	Single	5	4	53536	100



The product KP281 appears to be the most popular followed by KP 481 and then KP781.

Customers of Product KP281

- Average age

The average age of customers of KP281 is found to be 28.55 years. Median is 26, denoting the distribution to be unsymmetrical.

```
[24] df2['Age'].mean()
```

```
28.55
```

```
[31] df2['Age'].median()
```

```
26.0
```

- Average Income

The average income of customers of KP281 is found to be 46,000 \$ per annum.

```
[27] df2['Income'].mean()
```

```
46418.025
```


- Gender-wise distribution

The customer pool of product KP 281 found to be consisting of equal number from both genders.

```
[19] df2['Gender'].value_counts()
```

```
Gender
Male      40
Female    40
Name: count, dtype: int64
```

- Average Education years

Average education level of KP 281 customer is 15 years.

```
df2['Education'].mean()
```

```
15.0375
```

- Fitness level distribution

```
df2['Fitness'].value_counts()
```

```
Fitness
3      54
2      14
4       9
5       2
1       1
Name: count, dtype: int64
```

The value counts per fitness level shows that the customers are predominantly who finds themselves in the 2 and 3 fitness level i.e., in medium fitness level.

Customers of Product KP481

- Average age

The average age of customers of KP481 is found to be 28.9 years. Median is 26, denoting the distribution to be unsymmetrical.

```
df4['Age'].mean()
```

```
28.9
```

```
[9] df4['Age'].median()
```

```
26.0
```

- Average Income

The average income of customers of KP481 is found to be about 49,000 \$ per annum.

```
[11] df4['Income'].mean()
```

```
↔ 48973.65
```

- Gender-wise distribution

The customer pool of product KP 481 found to be consisting of equal number from both genders, with a slight upper hand of male customers.

```
[8] df4['Gender'].value_counts()
```

```
↔ Gender
Male      31
Female    29
Name: count, dtype: int64
```

- Average Education years

Average education level of KP 481 customer is 15 years

```
[6] df4['Education'].mean()
```

```
↔ 15.116666666666667
```

- Fitness level distribution

```
[5] df4['Fitness'].value_counts()
```

```
↔ Fitness
3      39
2      12
4       8
1       1
Name: count, dtype: int64
```

The value counts per fitness level shows that the customers are predominantly who finds themselves in the 2 and 3 fitness level i.e., in medium fitness level.

Customers of Product KP781

- Average age

The average age of customers of KP781 is found to be 29 years. Median is 27, denoting the distribution to be unsymmetrical.

```
df7['Age'].mean()
29.1

[19] df7['Age'].median()
27.0
```

- Average Income

The average income of customers of KP781 is found to be about 75,000 \$ per annum. Compared to other models it shows a significant difference in the metrics. Shows that people at higher level in the economic ladder are the customers of this variant.

```
df7['Income'].mean()
75441.575
```

- Gender-wise distribution

The customer pool of product KP 781 found to be consisting of mostly male customers. More than 80% of customers are men.

```
df7['Gender'].value_counts()
Gender
Male      33
Female     7
Name: count, dtype: int64
```

- Average Education years

Average education level of KP 781 customer is 17 years. Comparing to other two variants, it is more.

```
[21] df7['Education'].mean()
17.325
```

- Fitness level distribution

```
[13] df7['Fitness'].value_counts()
```

```

Fitness
5      29
4       7
3       4
Name: count, dtype: int64

```

The value counts per fitness level shows that the customers are predominantly who finds themselves in the 5th fitness level i.e., in excellent shape. There are no one in the customer pool who self-rated a poor shape. The data shows the behaviour of the customer pool for this variant to be keener in maintaining fitness.

Probability Distribution of Customers

- The probability of consumers of different variety of Aerofit product to be male or female is analysed using crosstab.

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

```

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

```

For a customer of KP281 and KP481 to be male or female is equal, but given a person is customer of KP781, there is 82.5% probability that the consumer is male.

- Similarly, the probability distribution of customers choosing the different products depending on their gender is described in:

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

```

Gender  Female  Male
Product
KP281    0.526316  0.384615
KP481    0.381579  0.298077
KP781    0.092105  0.317308

```

Here, given a person is female, the probability for them to choose KP281 is 52.6%, and choosing KP481 is 38.1% and choosing KP781 is only 9.2%. We can figure out that KP281 is most popular among women and KP781 is the least popular.

Among men, probabilities for choosing KP281 and KP781 are 38.5% and 31.7%. And for KP481 it is 29.8%. It describes that all the products have similar popularity among men.

- The probabilities of choosing the products across different fitness levels has been found out among men and women separately.

```
[ ] pd.crosstab(df[df['Gender']=='Female']['Product'],df[df['Gender']=='Female']['Fitness'],margins=True,normalize='index')
```

Fitness	1	2	3	4	5
Product					
KP281	0.000000	0.250000	0.650000	0.075000	0.025000
KP481	0.034483	0.206897	0.620690	0.137931	0.000000
KP781	0.000000	0.000000	0.142857	0.142857	0.714286
All	0.013158	0.210526	0.592105	0.105263	0.078947

```
[ ] pd.crosstab(df[df['Gender']=='Male']['Product'],df[df['Gender']=='Male']['Fitness'],margins=True,normalize='index')
```

Fitness	1	2	3	4	5
Product					
KP281	0.025000	0.100000	0.700000	0.150000	0.025000
KP481	0.000000	0.193548	0.677419	0.129032	0.000000
KP781	0.000000	0.000000	0.090909	0.181818	0.727273
All	0.009615	0.096154	0.500000	0.153846	0.240385

In both cases, for KP281 and KP481, the most probable fitness level is found to be 3 i.e., the chance that a customer of KP281 to belong to fitness level 3 is 70% in women and 65% in men. And for KP481 this is 67.7% and 62%. Interestingly, for the product KP781, the most probable fitness level is 5 with 71.4% in women and 72.7% in men.

- The probability of customer being in different age bins is analysed across gender.

```
[15] pd.crosstab(df2['Gender'],[df['AgeBin']],margins=True,normalize='index')
```

AgeBin	15-20	20-25	25-30	30-35	35-40	40-45	45-50
Gender							
Female	0.100	0.4000	0.225	0.150	0.05	0.0500	0.025
Male	0.150	0.3750	0.175	0.100	0.15	0.0250	0.025
All	0.125	0.3875	0.200	0.125	0.10	0.0375	0.025

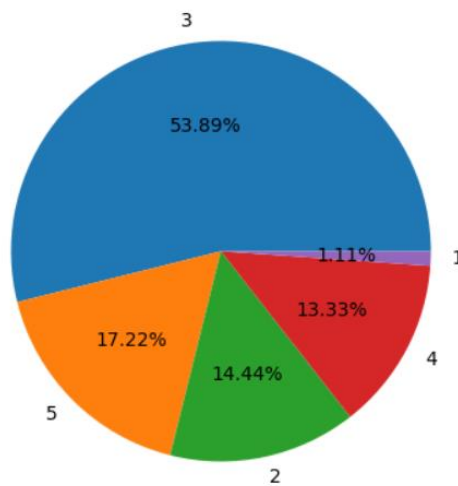
The probability of a male customer being in 20-25 years of age is 37.5% and is the greatest. Similarly, for a female customer to be in 20-25 age bin is 40% being the greatest.

3. Visual Analysis - Univariate & Bivariate

- i. For continuous variable(s): Distplot, countplot, histogram for univariate analysis

The distribution of customers across different fitness levels are described by using pie chart.

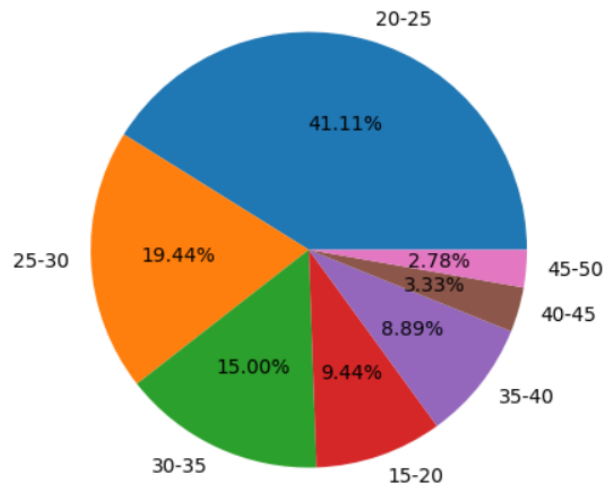
```
[18] counts=df['Fitness'].value_counts()
plt.figure(figsize = (7,7))
plt.pie(counts,
        labels=counts.index,
        startangle=0,
        autopct = '%.2f%%')
plt.show()
```



The major part of customers are rating themselves as 3 out of five in fitness.

Similarly, the distribution of customers across different agebins are described by using pie chart.

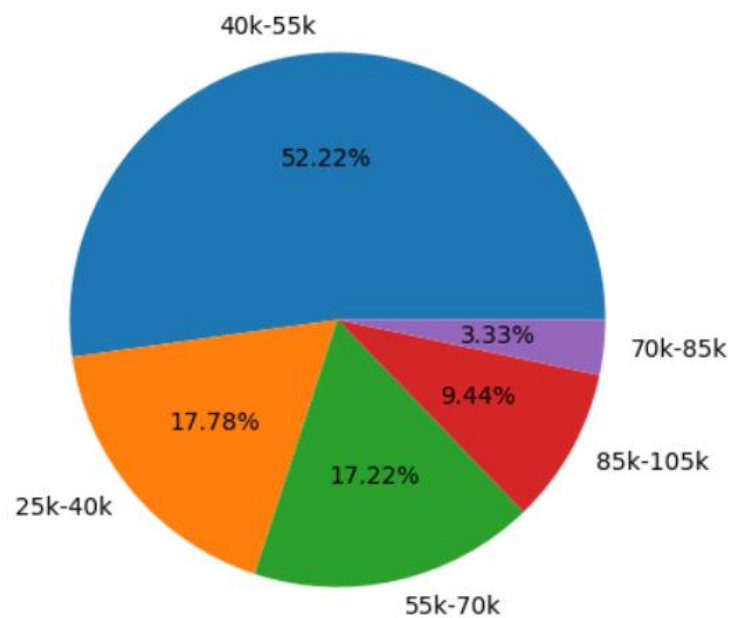
```
counts=df['AgeBin'].value_counts()
plt.figure(figsize = (5,5))
plt.pie(counts,
        labels=counts.index,
        startangle=0,
        autopct = '%.2f%%')
plt.show()
```



The figure shows that 41% of customers are of age 20-25 and thus being the target customers. People of ages 25-30 constitute around 20% of the customers.

Similarly, the distribution of customers across income levels are also depicted.

```
[24] counts=df['IncomeBin'].value_counts()
plt.figure(figsize = (5,5))
plt.pie(counts,
        labels=counts.index,
        startangle=0,
        autopct = '%.2f%%')
plt.show()
```

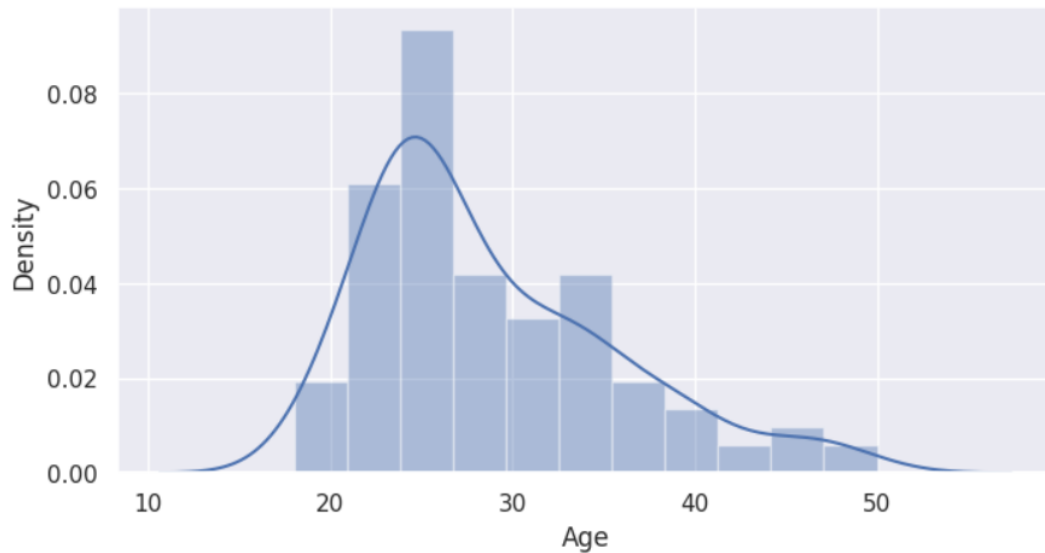


It shows that the people with annual income in the range \$40000-55000 are the target customers constituting 52% of the purchases of Aerofit products.

Distribution of customers across age are depicted using distplot.

```
[ ] sns.set(rc={"figure.figsize": (8, 4)})  
sns.distplot(df['Age'])
```

```
sns.distplot(df['Age'])  
<Axes: xlabel='Age', ylabel='Density'>
```

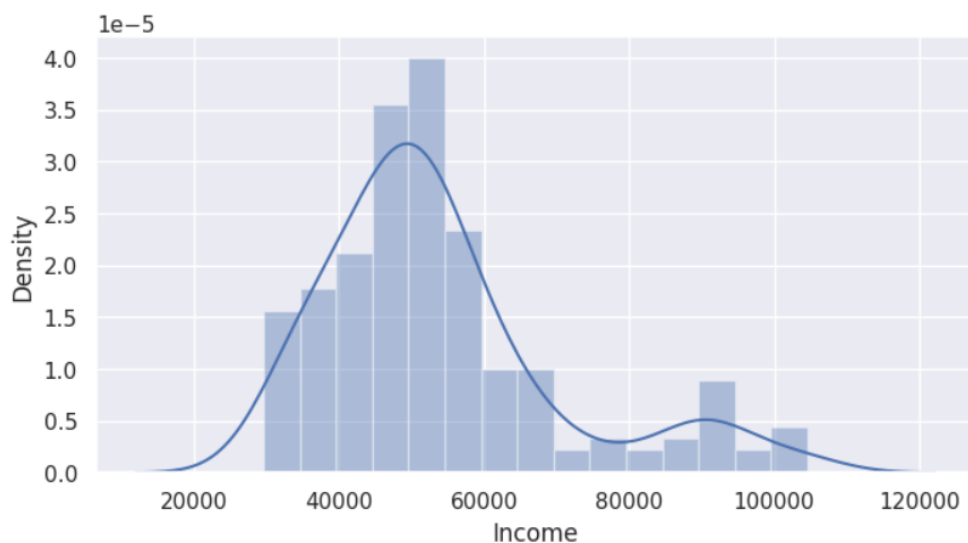


The number of customers peaks at about 25 years of age and then gradually decreases.

Similarly, the distribution of customers across income levels:

```
▶ sns.set(rc={"figure.figsize": (8, 4)})  
sns.distplot(df['Income'])
```

```
sns.distplot(df['Income'])  
<Axes: xlabel='Income', ylabel='Density'>
```

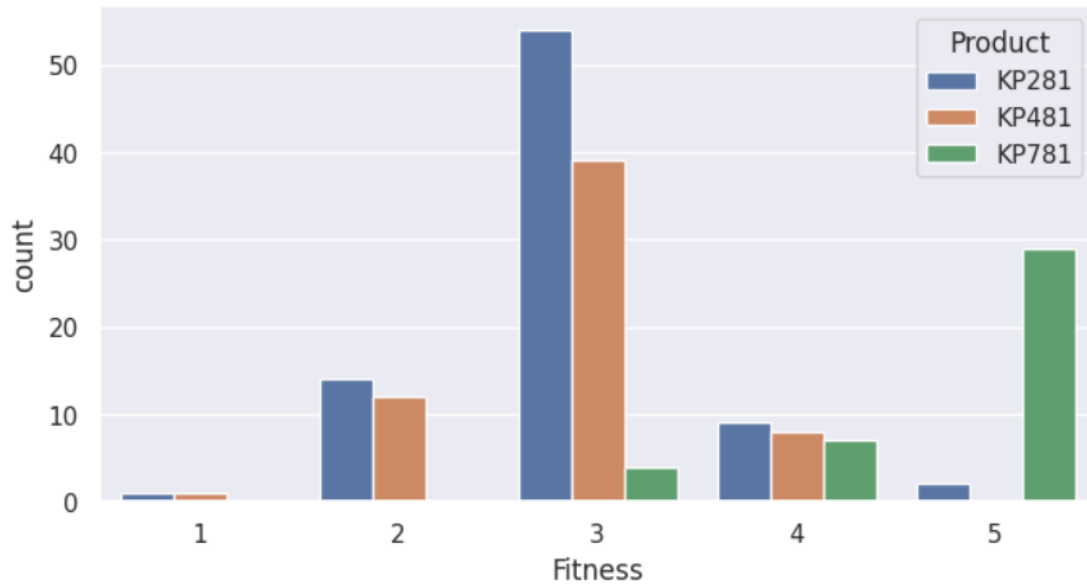


The peak of number of customers are about 50000\$ per annum income level.

Distribution of count of customers of the three varieties of products across fitness levels.

```
[28] sns.set(rc={"figure.figsize": (8, 4)})
      sns.countplot(data=df,x='Fitness',hue='Product')
```

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

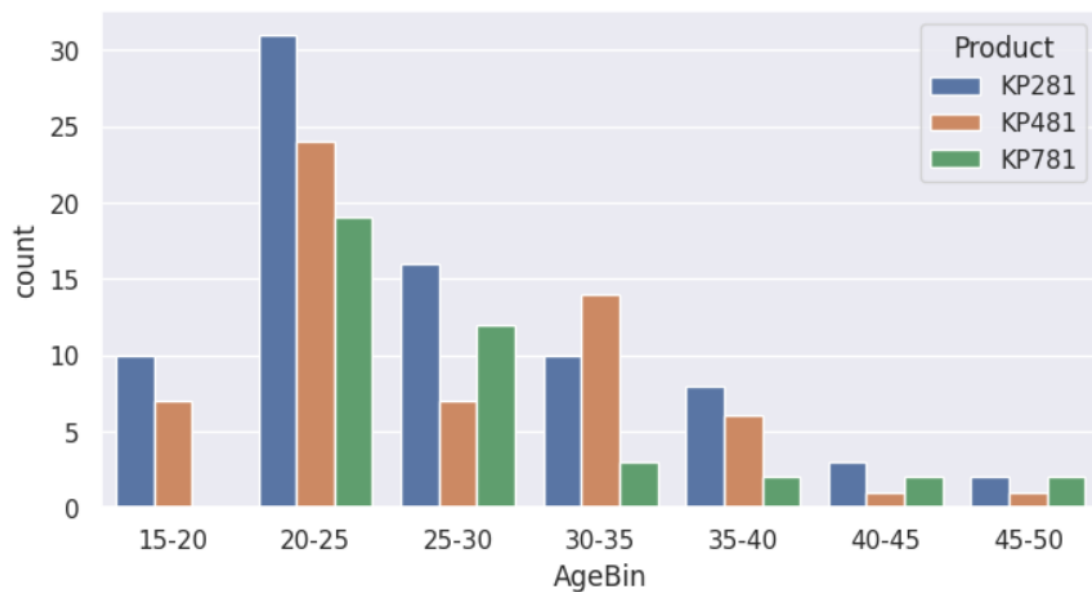


Here, even though the largest customer pool of KP 281 and KP 481 lies in the 3rd fitness level, the customers of KP 781 majorly constitute of 5 fitness level.

When analysed across age groups, the three varieties shown the following trend:

```
[ ] sns.set(rc={"figure.figsize": (8, 4)})
     sns.countplot(data=df,x='AgeBin',hue='Product')
```

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

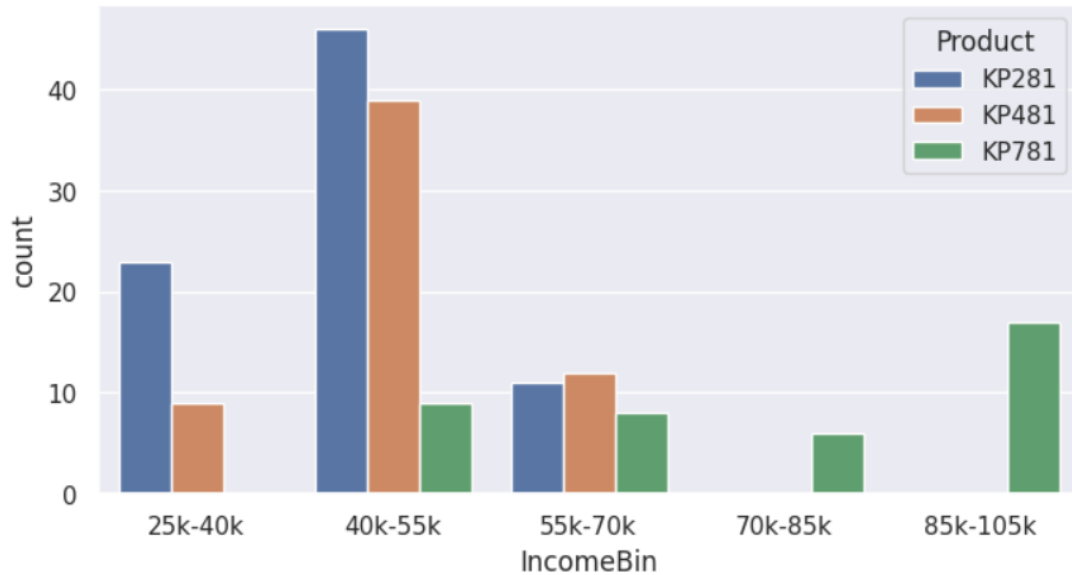


The major share of customers is confined to 20-25 age group. KP451 outplays other models in the 30-35 age bin.

The distribution of customers of the three models across different income levels:

```
[ ] sns.set(rc={"figure.figsize": (8, 4)})  
sns.countplot(data=df, x='IncomeBin', hue='Product')
```

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

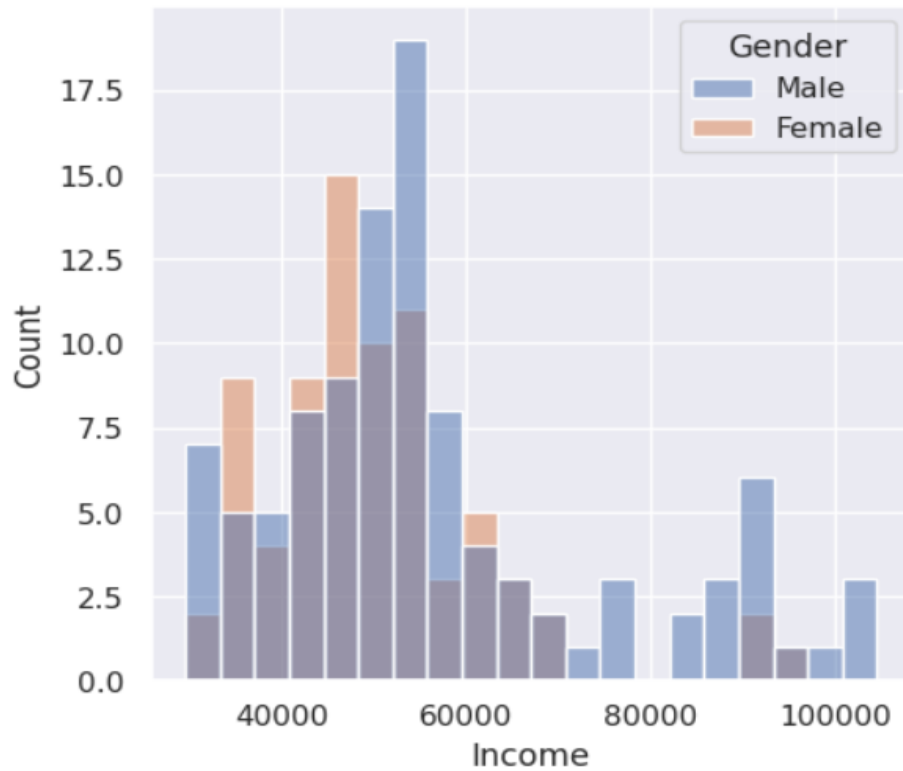


Largest share of customers in number is confined to 40000-55000\$ per annum range. But considering the higher price of the model KP781, the higher bins of >70000\$ per annum also plays an important range in the revenue to the company.

Distribution of customers across different income levels are described in the histogram:

```
[39] plt.figure(figsize = (5,5))  
     sns.histplot(data=df,x='Income',hue='Gender',bins=20)
```

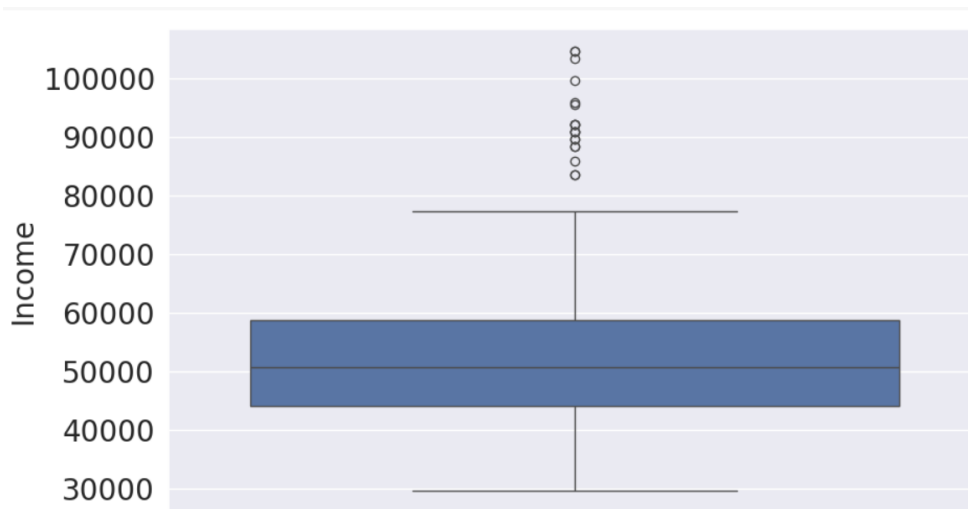
<Axes: xlabel='Income', ylabel='Count'>



ii. For categorical variable(s): Boxplot

The data of customers of different income levels is plotted as a box plot.

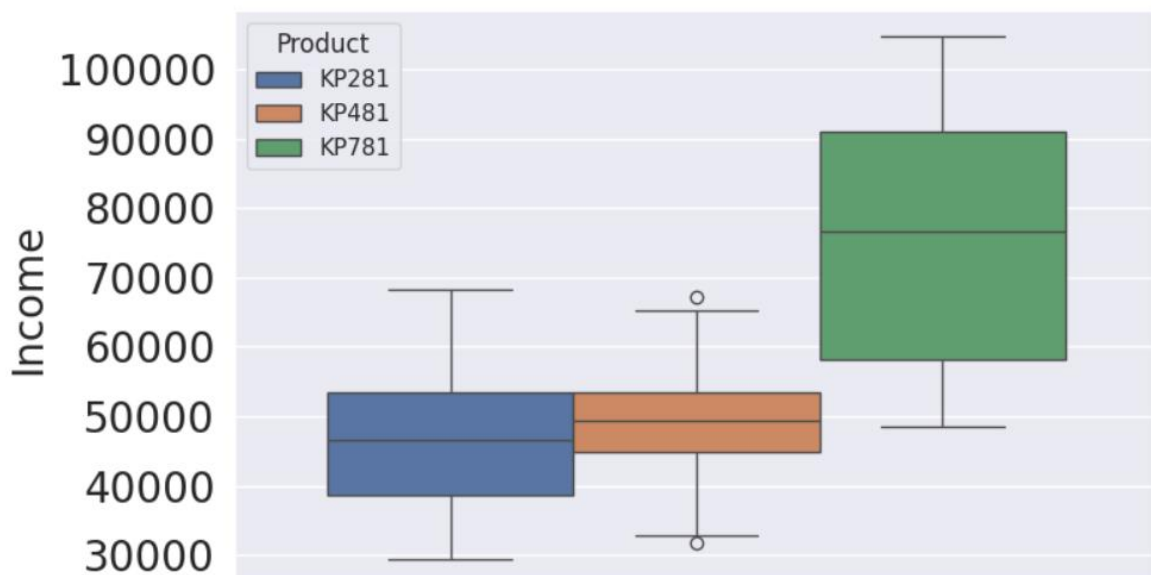
```
plt.figure(figsize=(10,6))  
sns.boxplot(data=df,y='Income')  
plt.yticks(fontsize=20)  
plt.ylabel('Income', fontsize=20)  
plt.show()
```



Here we can observe that there are some outliers.

The income levels of customers of different products:

```
[16] plt.figure(figsize=(10,6))
sns.boxplot(data=df,hue='Product',y='Income')
plt.yticks(fontsize=20)
plt.ylabel('Income', fontsize=20)
```

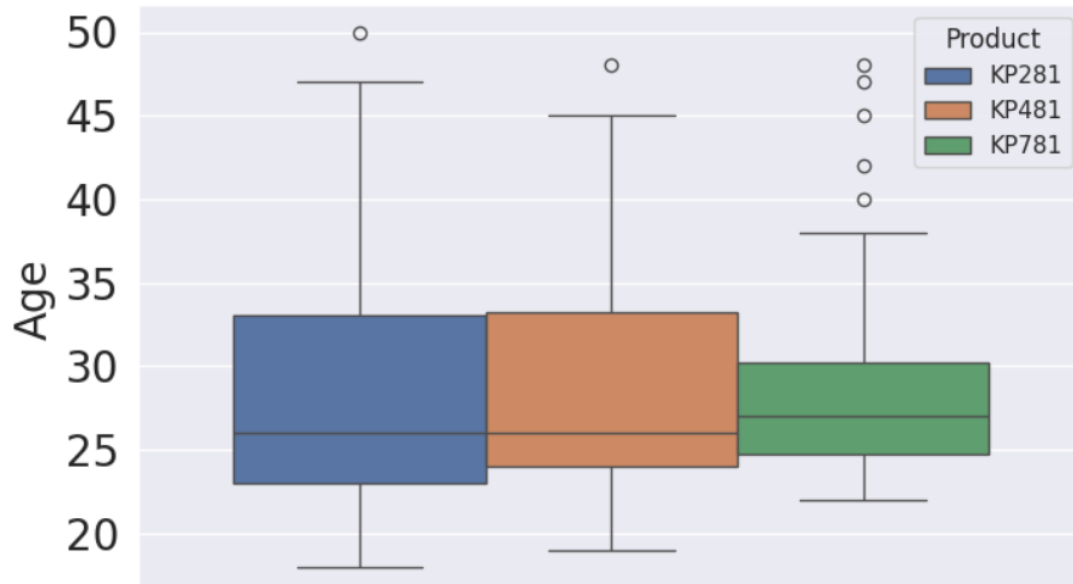


The 50 percentiles of KP281 are around \$40000 to \$50000 where that of KP481 is between \$45000 to \$50000 and that of KP 781 is between \$60000 to \$90000.

The distribution of customers of the three models across their age is depicted in the following box plot:

```
[43] plt.figure(figsize=(8,5))
      sns.boxplot(data=df,hue='Product',y='Age')
      plt.yticks(fontsize=20)
      plt.ylabel('Age', fontsize=20)
```

↗ Text(0, 0.5, 'Age')



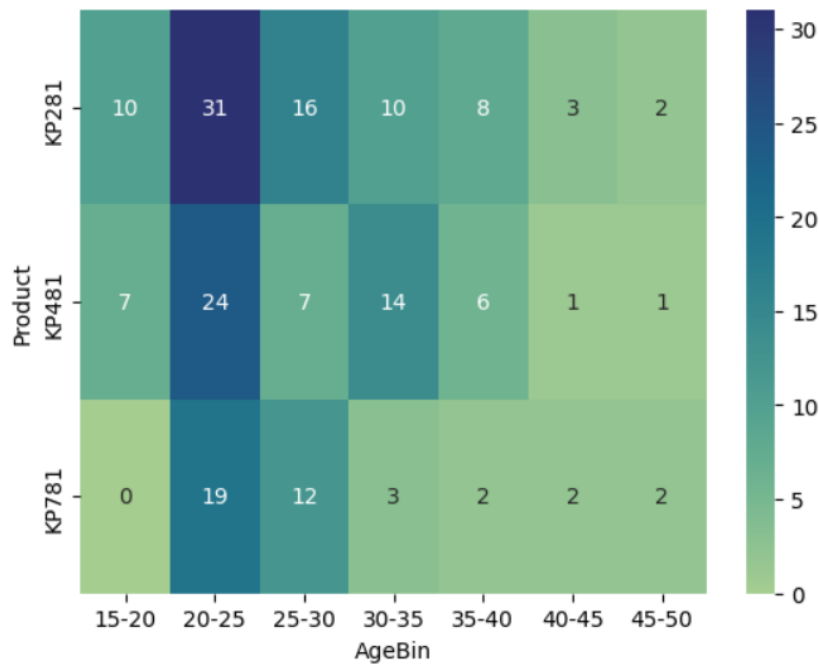
The ages of major share of customers of models KP281 and KP481 lies in the 25-35 years range. In case of KP781, the ages of customers are confined to 25 to 30 years range.

iii) For correlation: Heatmaps, Pairplots

The correlation between age and purchase of three different products is depicted in the heat map.

```
[ ] sns.heatmap(pd.crosstab(df['Product'],df['AgeBin']),annot=True,cmap='crest')
```

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

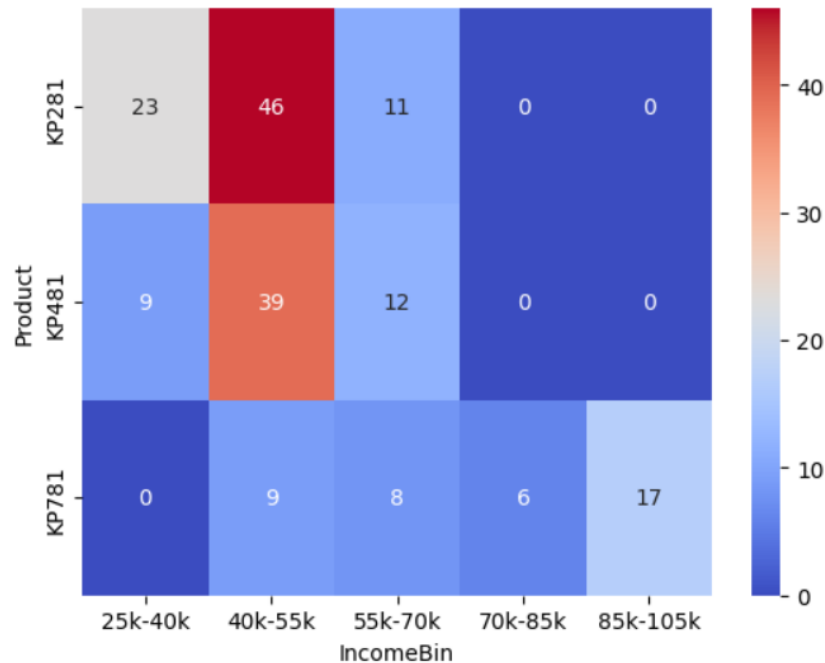


The greatest number of purchases are of KP281 by customers of age group 20-25 years. All the variants have largest number of sales in this age group. The sales of KP281 and KP781 to age group 25-30 follows, but for KP481, the 35-40 age group is second in number of sales.

The correlation between income and purchase of three different products is depicted in the heat map.

```
[8] sns.heatmap(pd.crosstab(df['Product'],df['IncomeBin']),annot=True,cmap='coolwarm')
```

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

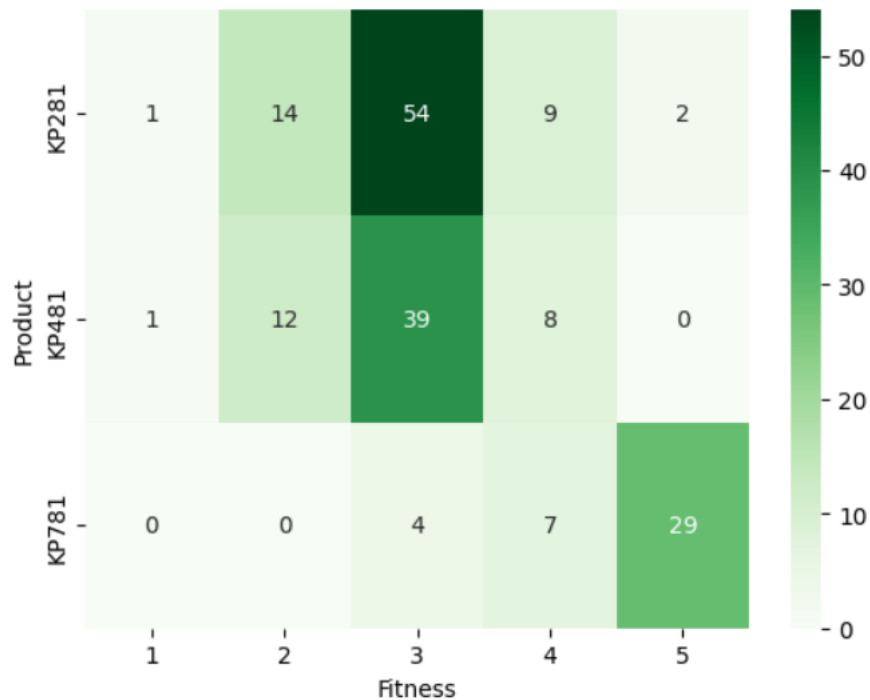


The greatest number of purchases are of KP281 by customers of income group \$40000-\$55000. All the variants have largest number of sales in this income group except KP781. The sales KP781 is maximum to income group \$85000-\$105000.

The distribution of customers across five fitness levels is obtained as follows:

```
[12] sns.heatmap(pd.crosstab(df['Product'],df['Fitness']),annot=True,cmap='Greens')
```

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

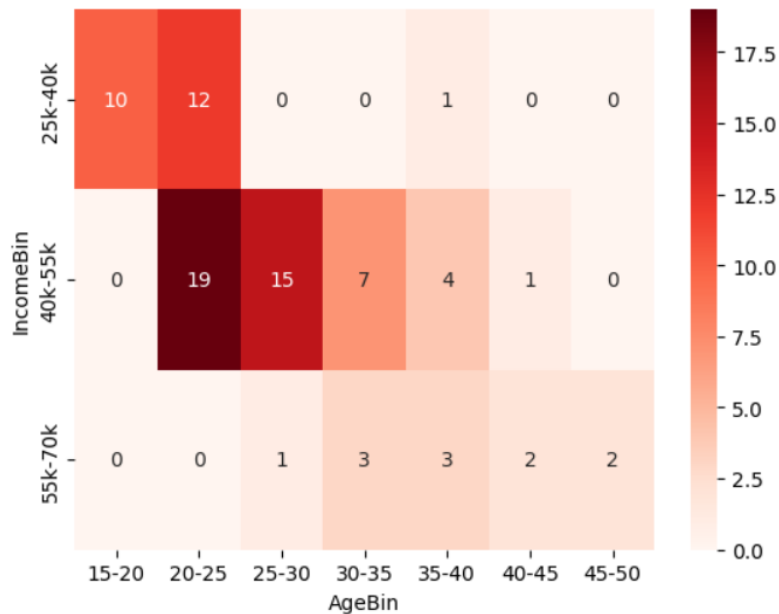


The greatest number of purchases is by KP281 for fitness level 3. KP481 is also having highest number of purchases at fitness level three. Unlike these two the sale of KP781 is highest among people of fitness level 5.

The distribution of customers of KP 281 along age and income are shown below. (df2 is the dataframe which contains the data corresponding to product KP281 alone)

```
[41] sns.heatmap(pd.crosstab(df2['IncomeBin'],df2['AgeBin']),annot=True,cmap='Reds')
```

```
<Axes: xlabel='AgeBin', ylabel='IncomeBin'>
```

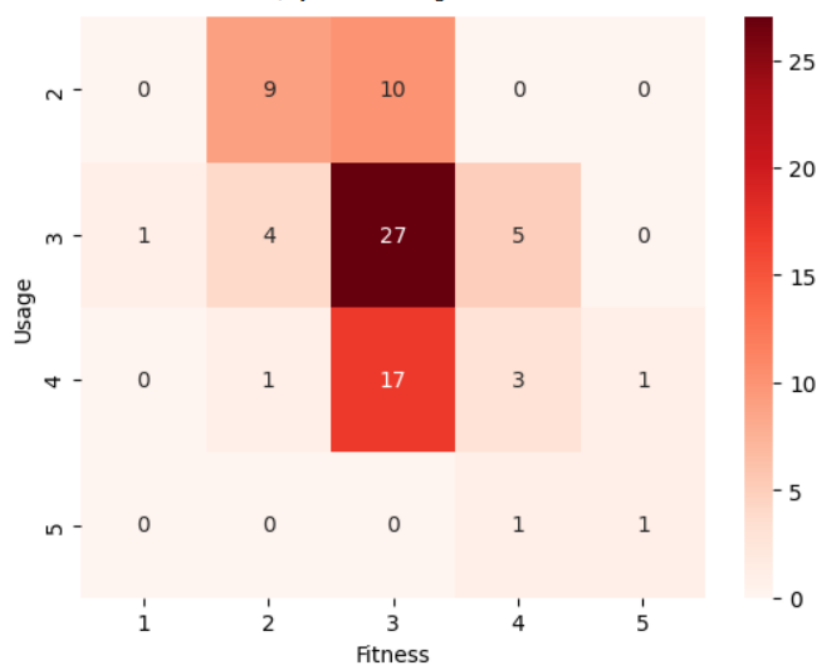


Here, it can be seen that customers are concentrated on the age of 20-25 years and income level \$40000-\$55000 per annum.

The distribution of customers of KP281 across fitness level and expected usage is shown below:

```
sns.heatmap(pd.crosstab(df2['Usage'],df2['Fitness']),annot=True,cmap='Reds')
```

```
<Axes: xlabel='Fitness', ylabel='Usage'>
```

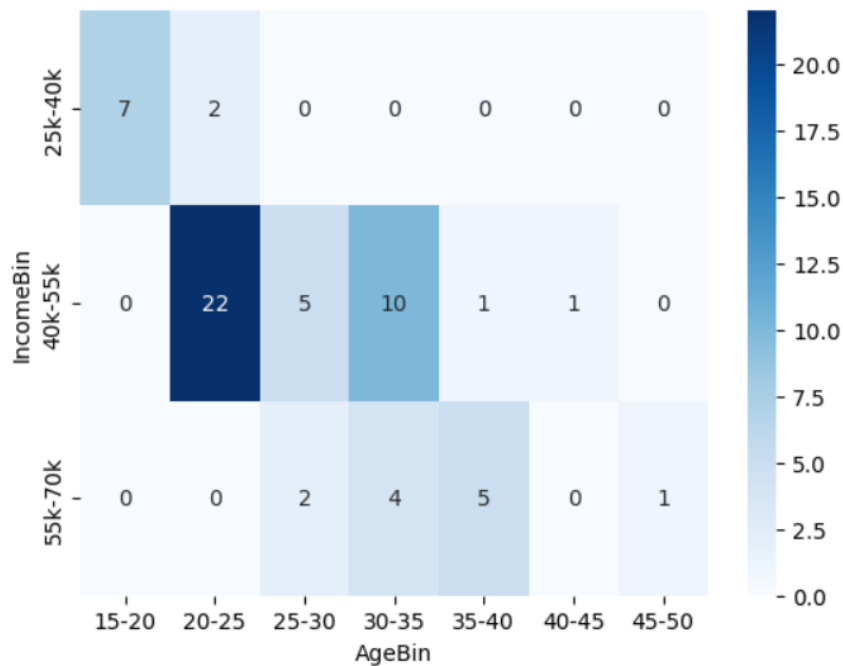


It can be inferred that fitness level 3 is the most popular among customers of KP281. i.e., people of medium fitness are the prime customers of this model. Also, the popular usage is 3 times a week i.e., moderate usage.

The distribution of customers of KP 481 along age and income are shown below. (df4 is the dataframe which contains the data corresponding to product KP481 alone)

```
[39] sns.heatmap(pd.crosstab(df4['IncomeBin'],df4['AgeBin']),annot=True,cmap='Blues')
```

```
<Axes: xlabel='AgeBin', ylabel='IncomeBin'>
```

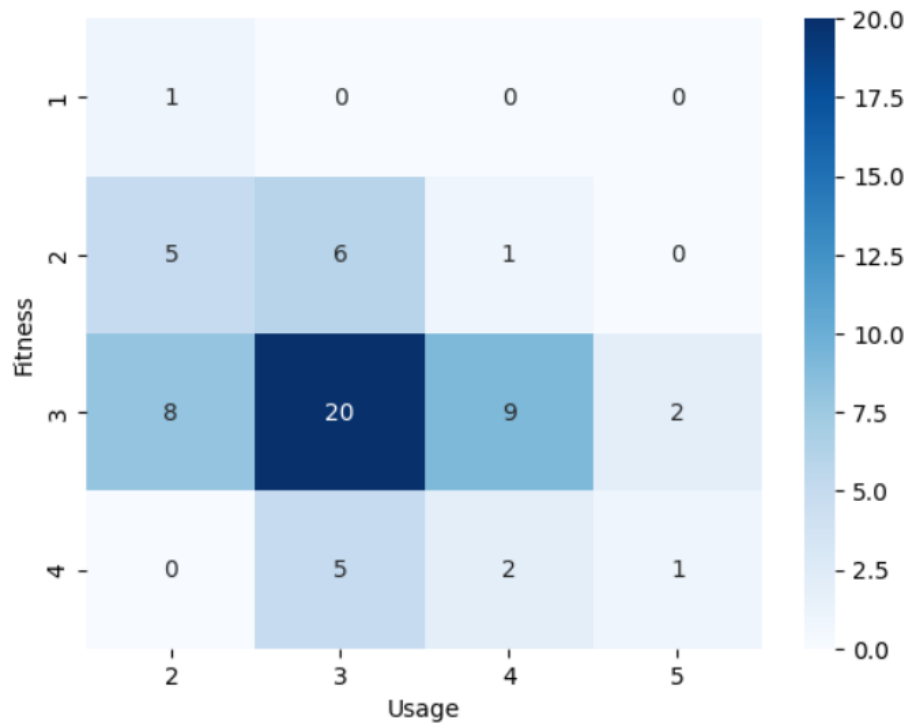


Here, it can be seen that customers are concentrated on the age of 20-25 years and income level \$40000-\$55000 per annum.

The distribution of customers of KP481 across fitness level and expected usage is shown below:

```
[40] sns.heatmap(pd.crosstab(df4['Fitness'],df4['Usage']),annot=True,cmap='Blues')
```

```
<Axes: xlabel='Usage', ylabel='Fitness'>
```

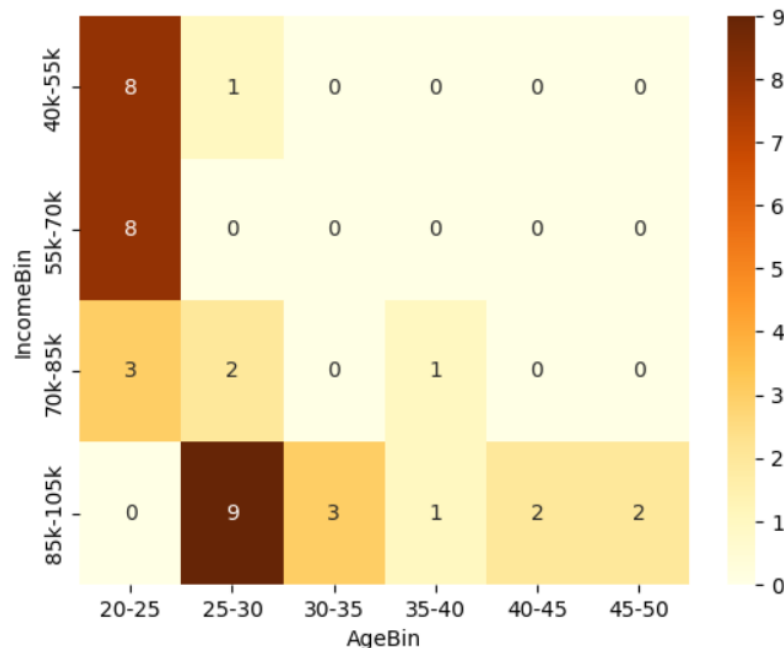


It can be inferred that fitness level 3 is the most popular among customers of KP481. i.e., people of medium fitness are the prime customers of this model. Also, the popular usage is 3 times a week i.e., moderate usage.

The distribution of customers of KP 781 along age and income are shown below. (df7 is the dataframe which contains the data corresponding to product KP781 alone)

```
[37] sns.heatmap(pd.crosstab(df7['IncomeBin'],df7['AgeBin']),annot=True,cmap='YlOrBr')
```

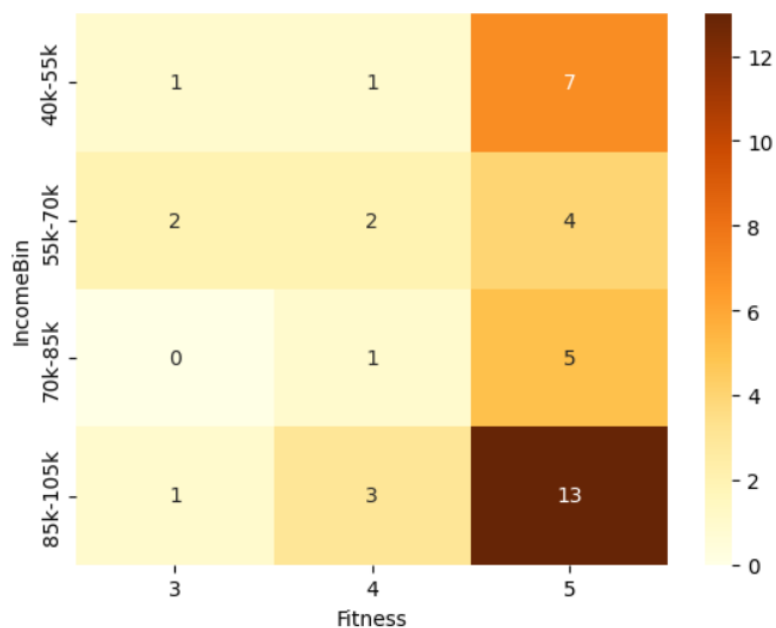
```
<Axes: xlabel='AgeBin', ylabel='IncomeBin'>
```



Here, unlike the other two models, it can be seen that customers are concentrated on the age of 20-30 years and the income levels doesn't show much of a concentration pattern. For further analysis, the distribution of customers of KP781 across income levels and fitness is plotted.

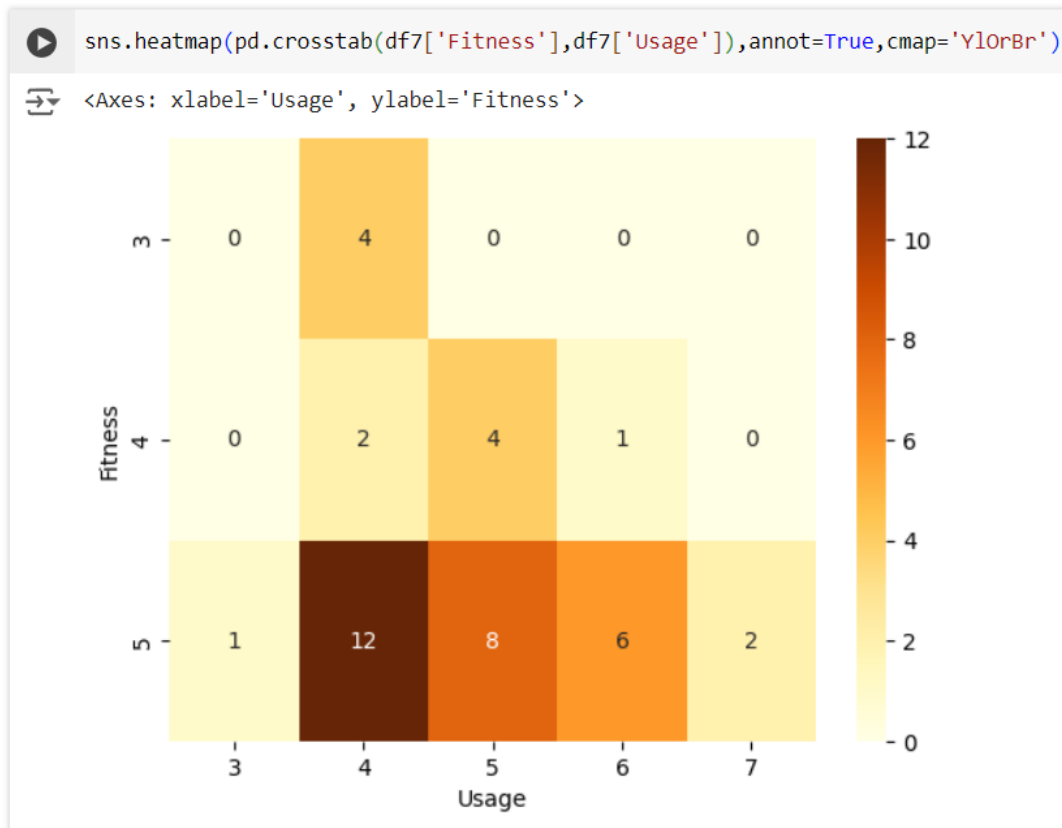
```
[43] sns.heatmap(pd.crosstab(df7['IncomeBin'],df7['Fitness']),annot=True,cmap='YlOrBr')
```

```
<Axes: xlabel='Fitness', ylabel='IncomeBin'>
```



The distribution indicates that the customers are largely concentrated in the fitness level 5, and the purchases from \$40000-\$55000 level is from this group of customers.

The distribution of customers of KP781 across fitness level and expected usage is shown below:



It can be inferred that fitness level 3 is the minimum among customers of KP781. i.e., people of medium to excellent fitness are the prime customers of this model. Also, the popular usage is 4 times a week and more. i.e., high usage.

4. Missing Value & Outlier detection

Detecting missing values by `isna()`.

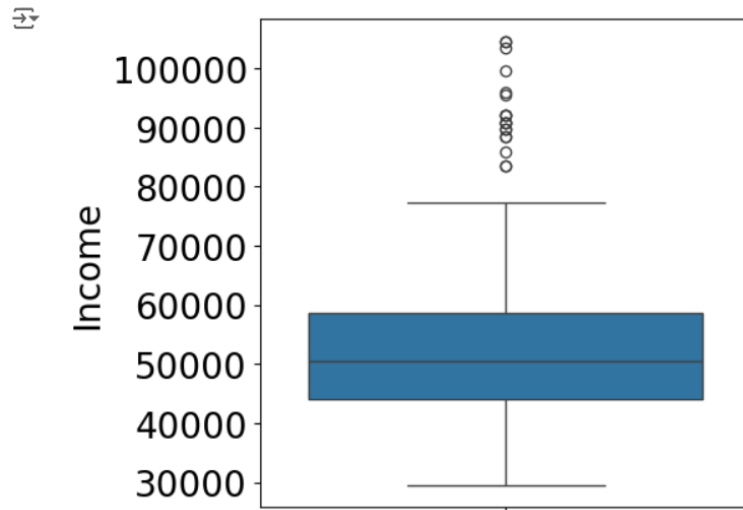
```
[11] df.isna().sum()
```

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

It shows there are no null values or missing values in the dataset.

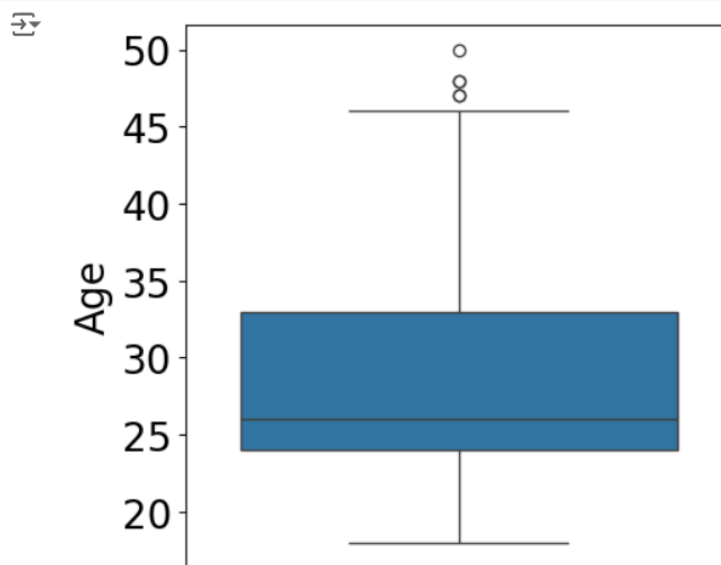
In the box plot drawn using the income of customers, outliers have been spotted.

```
[44] plt.figure(figsize=(5,5))  
     sns.boxplot(data=df,y='Income')  
     plt.yticks(fontsize=20)  
     plt.ylabel('Income', fontsize=20)  
     plt.show()
```



Similarly, in the boxplot drawn using the age of customers, outliers have been spotted.

```
plt.figure(figsize=(5,5))  
sns.boxplot(data=df,y='Age')  
plt.yticks(fontsize=20)  
plt.ylabel('Age', fontsize=20)  
plt.show()
```



5. Business Insights based on Non-Graphical and Visual Analysis

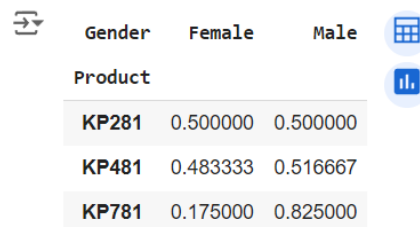
i. Comments on the range of attributes

- The data contains 180 entries corresponding to different customers of the three variety of products.
- There are 8 different attributes corresponding to each customer.
- Among them 80 records correspond to treadmill KP281, 60 records are about customers of treadmill KP481 and 40 are of customers of treadmill KP781.
- Among 180 customers 107 are men and 73 are women.
- The customers of KP281 and KP481 comprise of men and women equally where the customers of KP781 are majorly men.
- The customers age varies from 18 to 50.
- The customers of KP281 and KP481 bought that expecting moderate usage but KP781 are bought majorly expecting high usage.
- The customers of KP281 and KP481 are generally of moderate fitness where the customers of KP781 are from high fitness group.
- The income levels of KP281 and KP481 are moderate. Where in case of KP781, there are two types of income level concentration.

ii. Comments on the distribution of the variables and relationship between them

- The probability distribution of the customers of each model being male or female is observed.

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

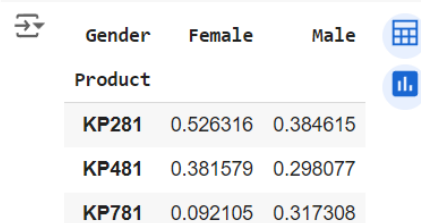


Gender	Female	Male
Product		
KP281	0.500000	0.500000
KP481	0.483333	0.516667
KP781	0.175000	0.825000

The distribution shows an even distribution of customers of KP281 and KP481 across both genders and in case of KP781, the customers are 82.5% men.

- The probability distribution of customers choosing the different products depending on their gender is described in:

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



Gender	Female	Male
Product		
KP281	0.526316	0.384615
KP481	0.381579	0.298077
KP781	0.092105	0.317308

Here, given a person is female, the probability for them to choose KP281 is 52.6%, and choosing KP481 is 38.1% and choosing KP781 is only 9.2%. We can figure out that KP281 is most popular among women and KP781 is the least popular.

Among men, probabilities for choosing KP281 and KP781 are 38.5% and 31.7%. And for KP481 it is 29.8%. It describes that all the products have similar popularity among men.

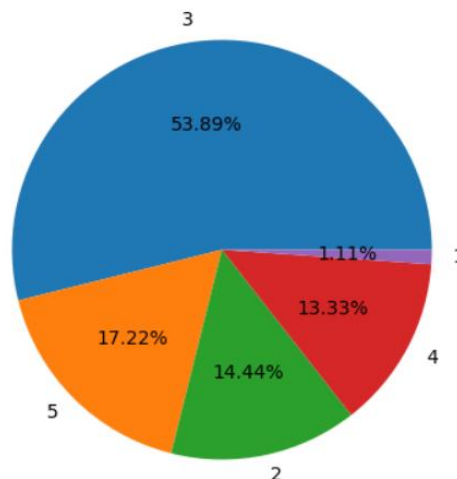
- The correlation between age and income levels are described using crosstab.

```
[ ] pd.crosstab(df['AgeBin'],df['IncomeBin'])
```

IncomeBin	25k-40k	40k-55k	55k-70k	70k-85k	85k-105k
AgeBin					
15-20	17	0	0	0	0
20-25	14	49	8	3	0
25-30	0	21	3	2	9
30-35	0	17	7	0	3
35-40	1	5	8	1	1
40-45	0	2	2	0	2
45-50	0	0	3	0	2

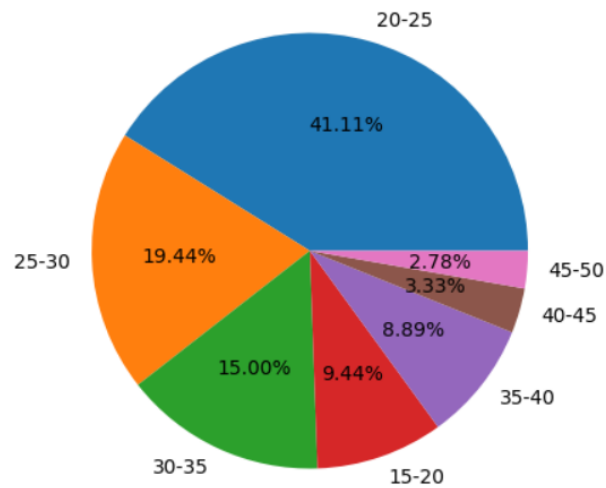
The greatest number of products are bought by income level 40-50 thousand dollars per annum level and in 20-35 years.

- The distribution customers over different fitness levels:



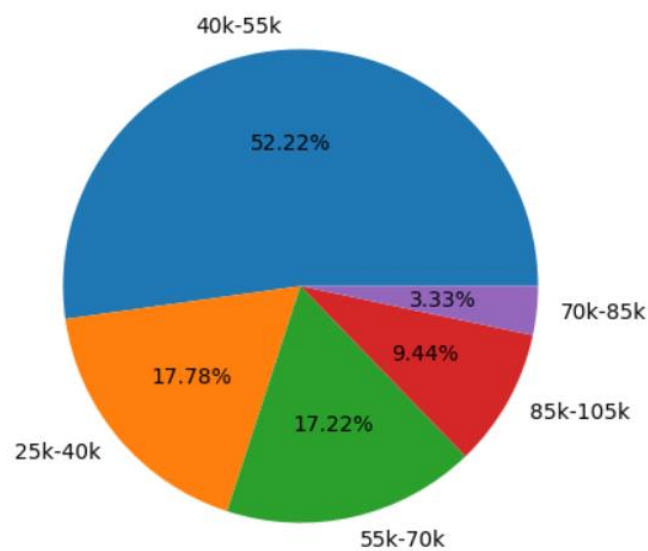
The major part of customers rates themselves as moderate fitness (fitness level-3). An ample share of customers are from fitness level 5, which are mostly customers of KP781 as seen from other metrics.

- The distribution of customers in different age groups:



The major part of customers is from 20-25 age group. The ages of 20-35 cover 75% of all the customers.

- The customers are distributed over their income levels in this chart.

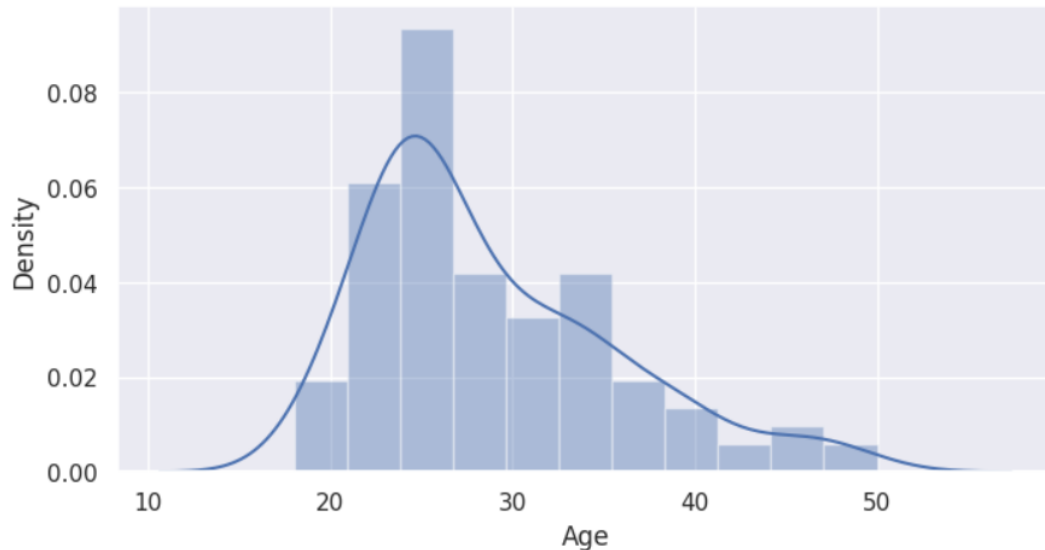


The income levels \$25000-\$70000 covers over 87% of all the consumers.

iii. Comments for each univariate and bivariate plot

- The dist-plot of number of customers across ages of customers show a varying trend which peaks around mid-20s and then starting to decrease

```
sns.distplot(df['Age'])  
<Axes: xlabel='Age', ylabel='Density'>
```

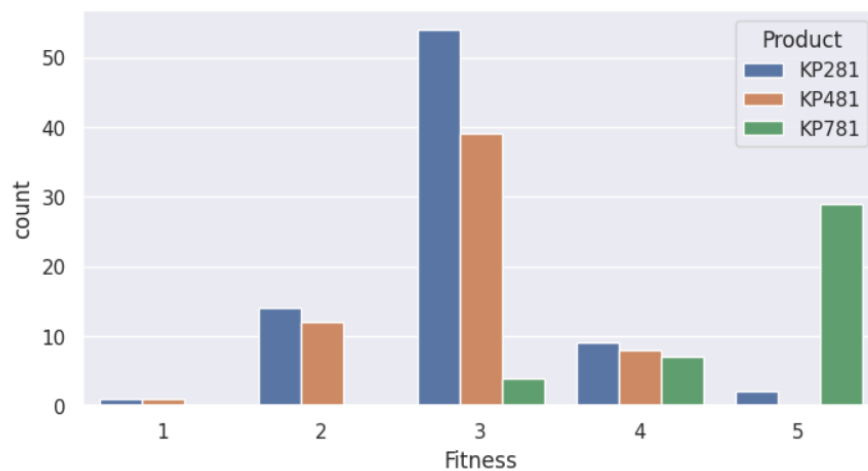


The number of customers peaks at about 25 years of age and then gradually decreases. It shows that the customers to target is fitness-oriented people of age 20s and 30s.

- The number of customers of the three models are distributed over different fitness levels in the following count plot.

```
[28] sns.set(rc={"figure.figsize": (8, 4)})  
sns.countplot(data=df, x='Fitness', hue='Product')
```

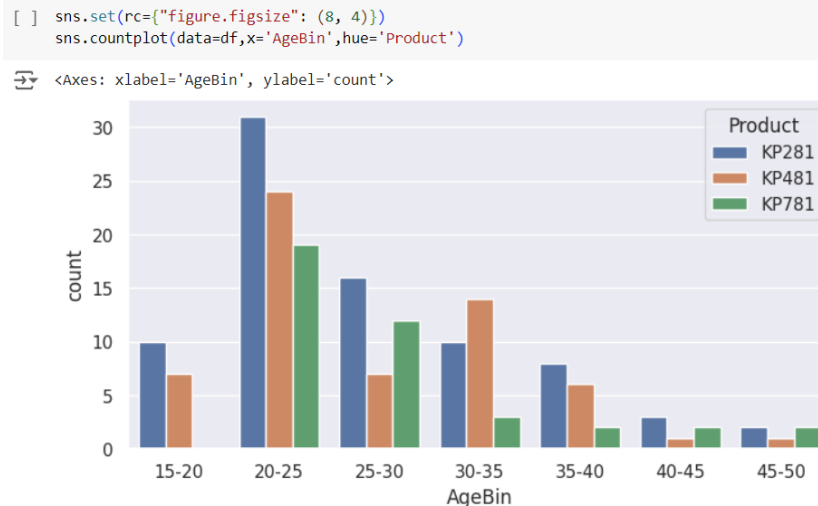
<Axes: xlabel='Fitness', ylabel='count'>



The largest number of customers is concentrated at fitness level 3 i.e., moderate fitness. This peak in number of customers is provided by KP281 and KP781

majorly. It can also be noticed that there is a smaller peak in the number of customers in the fitness level 5, provided by the model KP781. Considering that KP781 is a high end and more costly model, this peak is to be given ample importance.

- When analysed across age groups, the three varieties shown the following trend:



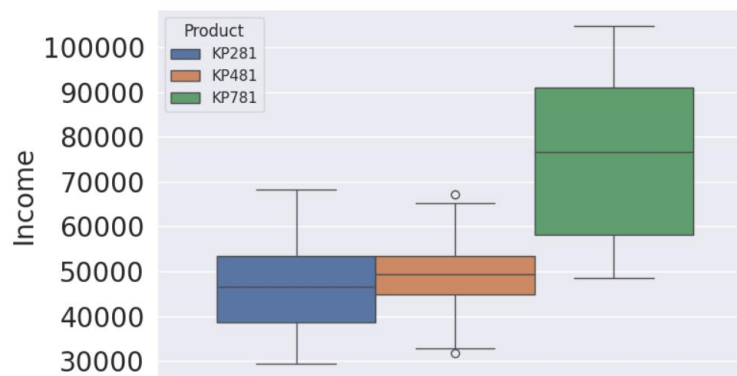
The major share of customers is confined to 20-25 age group. Which clearly defines the target customers of the Aerofit treadmills. KP451 outplays other models in the 30-35 age bin.

- The distribution of customers of the three models across different income levels:



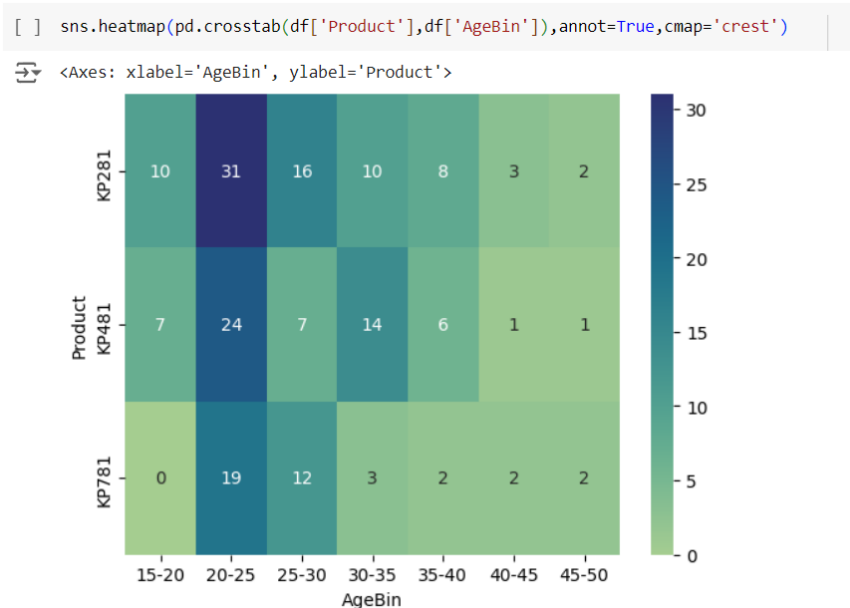
Largest share of customers in number is confined to 40000-55000\$ per annum range. But considering the higher price of the model KP781, the higher bins of >70000\$ per annum also plays an important range in the revenue to the company.

- The income levels plotted in the box plot:



The 50 percentiles of KP281 are around \$40000 to \$50000 where that of KP481 is between \$45000 to \$50000 and that of KP 781 is between \$60000 to \$90000. The plot clearly marks the target customers of each type of products. The target customers of KP781 are exclusive from that of KP281 and KP481 in terms of annual income.

- The correlation between age and purchase of three different products is depicted in the heat map.

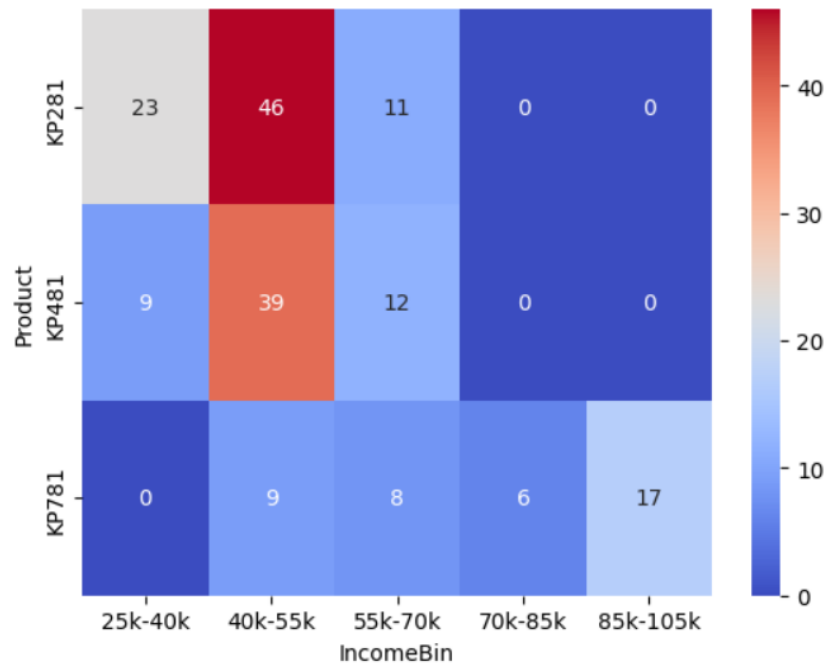


This plot allows us to confine the customer profile with respect to age of the customer. The greatest number of purchases are of KP281 by customers of age group 20-25 years. All the variants have largest number of sales in this age group. The sales of KP281 and KP781 to age group 25-30 follows, but for KP481, the 35-40 age group is second in number of sales.

- The correlation between income and purchase of three different products is depicted in the heat map.

```
[8] sns.heatmap(pd.crosstab(df['Product'],df['IncomeBin']),annot=True,cmap='coolwarm')
```

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



This plot allows us to confine the customer profile with respect to income of the customer. The greatest number of purchases are of KP281 by customers of income group \$40000-\$55000. All the variants have largest number of sales in this income group except KP781. The sales KP781 is maximum to income group \$85000-\$105000, which is an important group of customers considering the higher price of the model.

- Apart from these, the analysis shows that the target customers of KP281 and KP481 are concentrated in moderate-fitness level and moderate-income range. Also, they majorly expect the product to be put to moderate usage. While in the case of KP781, the target customers are concentrated in men of excellent fitness level, who expect the model to be used 4-7 days a week on average.

6. Recommendations.

- The customer profile of KP281 shows that the product is used by both men and women of age 18 to 30 and of income level \$25000 to \$55000. The major share of customers confined to moderate-fitness level. The usage of this model is found to be moderate. There are ways to optimise the product for this usage and promote the product among its target audience.
- The customer profile of KP481 shows that the product is used by both men and women of age 20 to 25 and 30 to 35 and of income level \$40000 to \$55000. The major share of customers confined to moderate-fitness level. The usage of this model is found to be moderate. There are ways to optimise the product for this usage and promote the product among its target audience.

- Another interesting trend is in the case KP781. Unlike other two models of treadmill, KP781 is a more sophisticated and costly product. From the analysis it is found that the customers of KP781 is concentrated on high fitness level. Even though large part of the customers of this model is from high income group, there are also a considerable number of lower-income customers who are good in shape and irrespective of the income level, they expect it to be used frequently. The expected milage and usage shows that this product is exclusively expected to be of high usage. The target customers are thus identified and promotions shall be in such a way to attract the possible customers.
- The model KP781 is also special in the case of customer distribution across gender. Unlike other two models, 82.5% of the customers of this product are men. Considering the cost of this product, and the fact that it contributes to the sales in an equal share of other two variants, the target customers are comparatively concentrated and well defined. Promotional activities for this model can be thus very focused and much cost-effective.