Business Case Study: Descriptive Stats and Probability

Python Notebook Link

Submitted by: Abhishek Singh

Business Problem:

- 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.

Pre-Requisites:

1. Required Python Libraries were imported.
2. Dataset was imported to Jupyter notebook, stored in a data frame and preliminary analysis performed on the data to understand the data size, check null values, and derive statistical observations from the data

Preliminary Data Cleansing and Exploration:

1. Data types of the columns was checked. Columns 'Product', 'Gender' and 'MaritalStatus' were converted to 'category' datatype for efficient memory utilization.

```
#Checking the Dataset datatypes
data.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
```

```
#Convert the columns to categorical dtype wherever applicable
data['Product']=data['Product'].astype('category')
data['Gender']=data['Gender'].astype('category')
data['MaritalStatus']=data['MaritalStatus'].astype('category')
```

```
data.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    category
 1   Age            180 non-null    int64   
 2   Gender         180 non-null    category
 3   Education      180 non-null    int64   
 4   MaritalStatus  180 non-null    category
 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: category(3), int64(6)
memory usage: 9.5 KB
```

2. Shape of dataset was checked to consist of 180 Rows x 9 Columns.
3. Checked first 5 entries of dataset to understand the provided data.

```
#Check first 5 record of DF to understand the data provided
data.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

Column Name	Column Description
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

4. Descriptive analysis was performed on the columns to check the mean, median, minimum value, maximum value and 25%, 50% and 70%-ile values and IQR calculated.

```
#Statistical Analysis of the Dataset
data.describe(include="all")
```

	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

5. Significant Variance between the mean and 50% value in column Age, Income and Miles indicate presence of outliers in the provided data.
6. The IQR for the columns are calculated using the 25% and 75%ile values. Using the calculated IQR, we calculate the upper and lower limit of the values to eliminate possible outliers. The final data is tabulated as below:

Column Name	25% value	75% value	IQR	Lower Limit	Upper Limit
Age	24.00	33.00	9.00	10.500	46.500
Education	14.00	16.00	2.00	11.000	19.000
Usage	3.00	4.00	1.00	1.500	5.500
Fitness	3.00	4.00	1.00	1.500	5.500
Income	44058.75	58668.00	14609.25	22144.875	80581.875
Miles	66.00	114.75	48.75	-7.125	187.875

7. Checked the existence of null values on all the columns and the number of unique values on all the columns.

```
#Checking how many null entries we have across all the columns
data.isnull().sum()
```

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

```
#Checking the number of unique values in each columns
data.nunique()
```

```
Product           3
Age              32
Gender            2
Education         8
MaritalStatus     2
Usage             6
Fitness           5
Income           62
Miles            37
dtype: int64
```

8. Value counts was checked for all the columns to check for any irregularities/incorrect values and no such discrepancies in values were found

```
#checking the value counts of each column to check for any inconsistencies
for i in list(data.columns):
    print("Column" ,i,":")
    print(data[i].value_counts())
    print()
```

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

Column Age :
25    25
23    18
24    12
26    12
28     9
35     8
33     8
30     7
38     7
21     7
22     7
27     7
```

9. New columns ‘AgeGroup’ and ‘IncomeGroup’ were created to segregate the entries based on the ‘Age’ and ‘Income’ values.

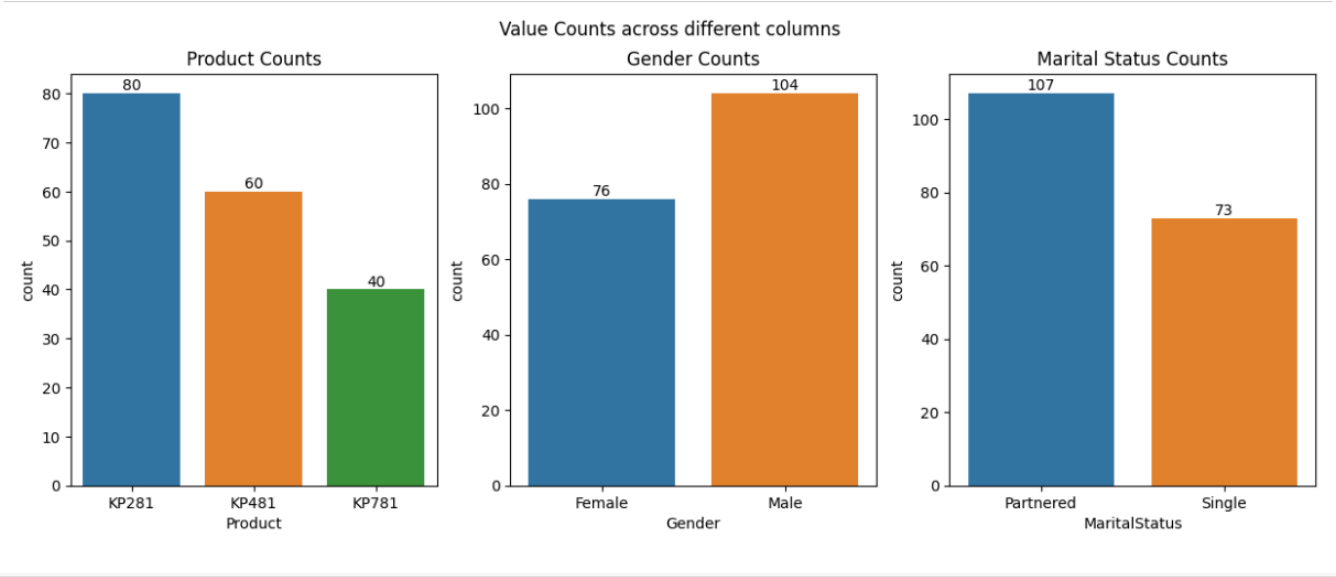
```
#Creating Age brackets to club data
data['AgeGroup']=data['Age'].apply(lambda x : '40-50' if x>=40 else '30-39' if x>=30 else '18-29')
```

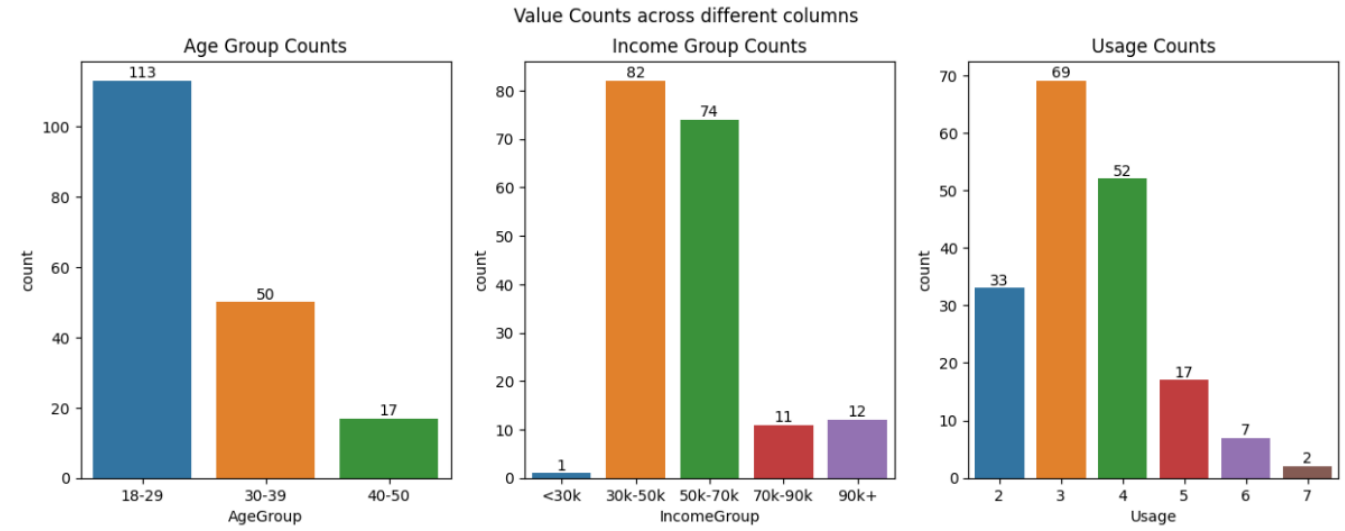
```
#Creating Income brackets to club income data
data['IncomeGroup']=data['Income'].apply(lambda x : '90k+' if x>=90000 else '70k-90k' if x>=70000 \
                                         \ else '50k-70k' if x>=50000 else '30k-50k' if x>=30000 else '<30k')
```

10. Price of the products was added to the provided dataset to generate further insights.

Univariate Analysis:

- Checked the value count distribution of the values across all the columns to understand the frequency of the values occurring in each column.





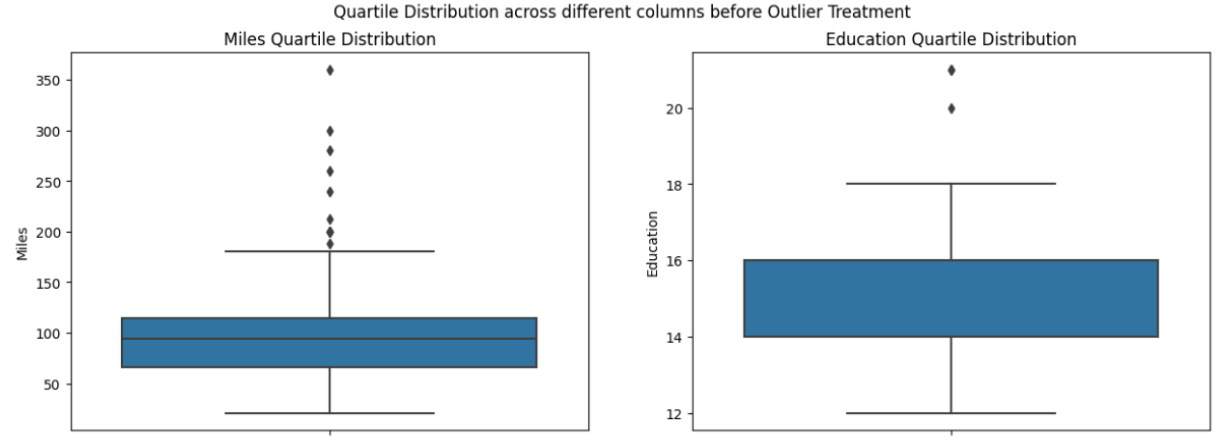
```
#Outlier Treatment. Creating a new column with the outlier values being replaced with the Upper Limit.
#Creating a new column as it will help retain the original values if needed

data.loc[data['Age']>description.loc['Age','Upper Limit'],'NewAge']=description.loc['Age','Upper Limit']
data.loc[data['Income']>description.loc['Income','Upper Limit'],'NewIncome']=description.loc['Income','Upper Limit']
data.loc[data['Miles']>description.loc['Miles','Upper Limit'],'NewMiles']=description.loc['Miles','Upper Limit']
data.loc[data['Education']>description.loc['Education','Upper Limit'],'NewEducation']=description.loc['Education','Upper Limi

#Filling the remaining values with original values
data['NewAge'].fillna(data['Age'],inplace=True)
data['NewMiles'].fillna(data['Miles'],inplace=True)
data['NewIncome'].fillna(data['Income'],inplace=True)
data['NewEducation'].fillna(data['Education'],inplace=True)
```

```
#Dataset after removal of Outliers from Age, Income, Miles and Education
data.head()
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	NewAge	NewIncome	NewMiles	NewEducation
0	KP281	18	Male	14	Single	3	4	29562	112	18.0	29562.0	112.0	14.0
1	KP281	19	Male	15	Single	2	3	31836	75	19.0	31836.0	75.0	15.0
2	KP281	19	Female	14	Partnered	4	3	30699	66	19.0	30699.0	66.0	14.0
3	KP281	19	Male	12	Single	3	3	32973	85	19.0	32973.0	85.0	12.0
4	KP281	20	Male	13	Partnered	4	2	35247	47	20.0	35247.0	47.0	13.0




```
#Outlier Treatment. Creating a new column with the outlier values being replaced with the Upper Limit.
#Creating a new column as it will help retain the original values if needed

data.loc[data['Age']>description.loc['Age','Upper Limit'],'NewAge']=description.loc['Age','Upper Limit']
data.loc[data['Income']>description.loc['Income','Upper Limit'],'NewIncome']=description.loc['Income','Upper Limit']
data.loc[data['Miles']>description.loc['Miles','Upper Limit'],'NewMiles']=description.loc['Miles','Upper Limit']
data.loc[data['Education']>description.loc['Education','Upper Limit'],'NewEducation']=description.loc['Education','Upper Limi

#Filling the remaining values with original values
data['NewAge'].fillna(data['Age'],inplace=True)
data['NewMiles'].fillna(data['Miles'],inplace=True)
data['NewIncome'].fillna(data['Income'],inplace=True)
data['NewEducation'].fillna(data['Education'],inplace=True)
```

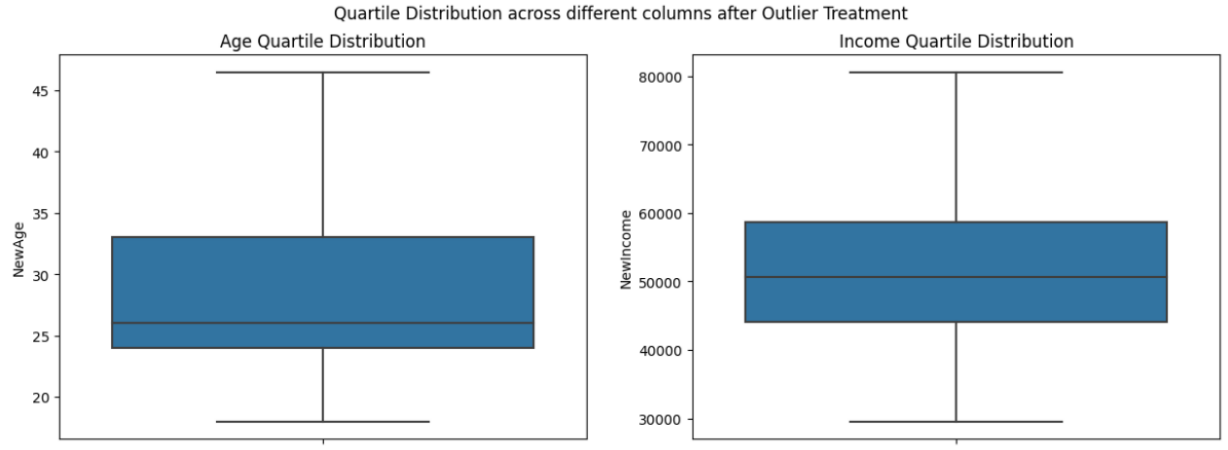
```
#Dataset after removal of Outliers from Age, Income, Miles and Education
data.head()
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles	NewAge	NewIncome	NewMiles	NewEducation
0	KP281	18	Male	14	Single	3	4	29562	112	18.0	29562.0	112.0	14.0
1	KP281	19	Male	15	Single	2	3	31836	75	19.0	31836.0	75.0	15.0
2	KP281	19	Female	14	Partnered	4	3	30699	66	19.0	30699.0	66.0	14.0
3	KP281	19	Male	12	Single	3	3	32973	85	19.0	32973.0	85.0	12.0
4	KP281	20	Male	13	Partnered	4	2	35247	47	20.0	35247.0	47.0	13.0

```
#Quartile Distribution across different columns after Outlier Treatment
plt.figure(figsize=(15,5))
plt.subplot(1,2,1)
a=sns.boxplot(y=data.NewAge)
plt.title('Age Quartile Distribution', fontsize=12)

plt.subplot(1,2,2)
b=sns.boxplot(y=data.NewIncome)
plt.title('Income Quartile Distribution', fontsize=12)

plt.suptitle('Quartile Distribution across different columns after Outlier Treatment')
plt.show()
```

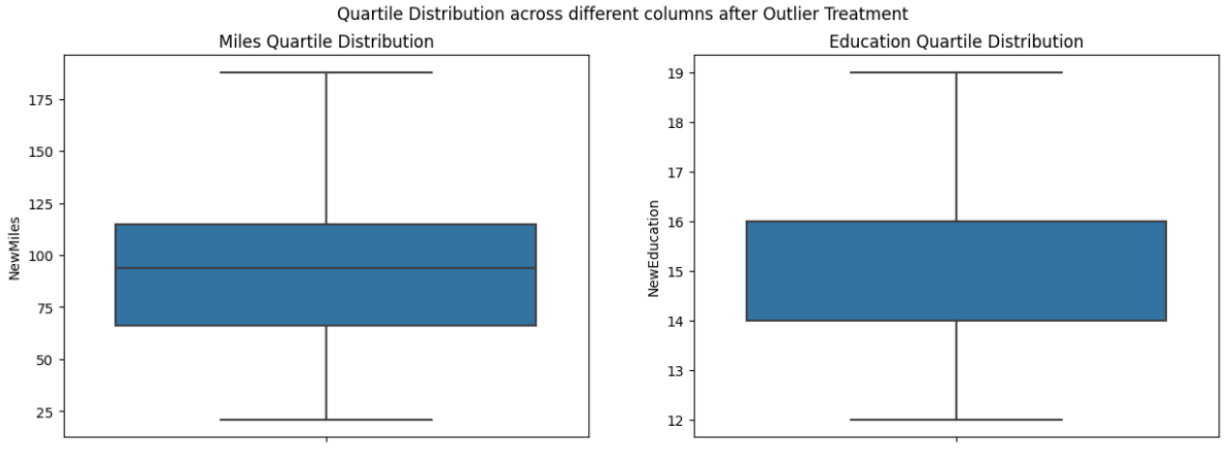



```
#Quartile Distribution across different columns before Outlier Treatment
plt.figure(figsize=(15,5))

plt.subplot(1,2,1)
c=sns.boxplot(y=data.NewMiles)
plt.title('Miles Quartile Distribution', fontsize=12)

plt.subplot(1,2,2)
c=sns.boxplot(y=data.NewEducation)
plt.title('Education Quartile Distribution', fontsize=12)

plt.suptitle('Quartile Distribution across different columns after Outlier Treatment')
plt.show()
```

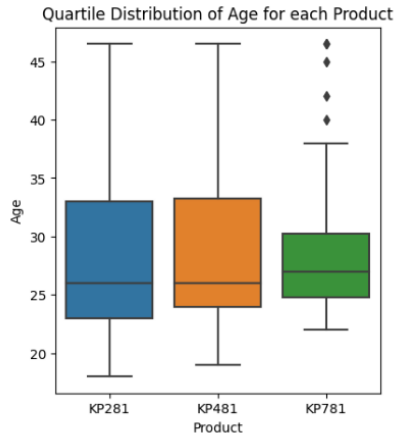


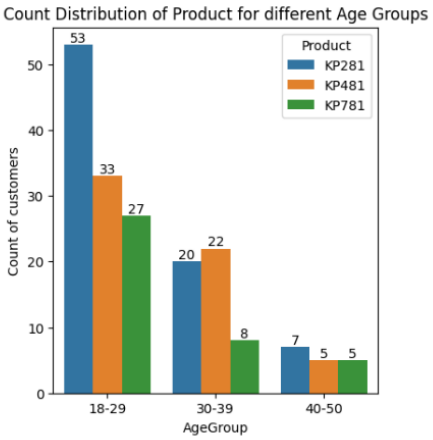
Business Insights from Univariate Analysis:

- Available data suggests that the price of products has the highest impact on overall revenue generated per product, as the highest units have sold from KP281 (least expensive) and witnesses a linear decrease as the product price increases (KP281 units > KP481 units > KP781 units)
- Male: female customer count is 104:76 i.e., on average 36% higher male customers compared to the female customers. This could be due to underlying reasons such as the average/median income across the genders, an overall higher interest in purchase of home fitness equipment in one gender compared to another, the %age of women with independent sources of income to allow them such purchases, or a combination of any of these and would require further analysis to be done at later stage.
- Partnered (Married/Live-in) people have a larger sale volume of ~46% higher than single customers indicating that customers with partners are more likely to get involved in home fitness routine.
- The younger customer base in the age group 18-29 years, with income 30-50k annually and estimated planned used of 3 times a week contribute the most significant %age to the overall revenue, indicating that the company has been able to establish itself firmly in the youth population with a middle class earning, and should not work towards establishing the brand value and confidence in the financial upper segment of users to be able to increase the profit margins significantly. The same can inferred from the median and 25%-ile and 75%-ile distributions of the overall Age, Income and Miles used data provided.

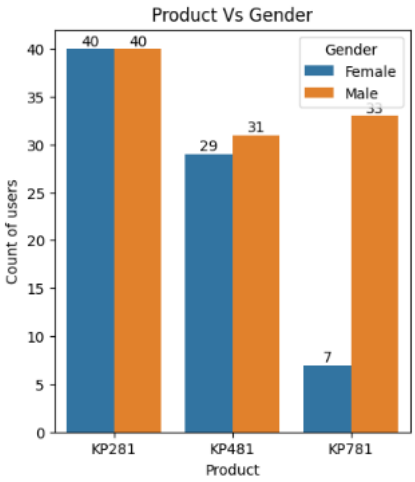
Bi-Variate Analysis:

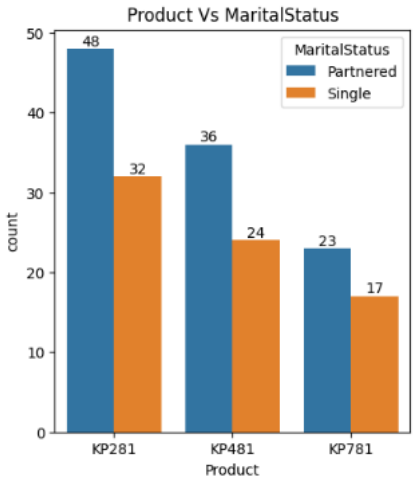
- Bi-variate analysis was performed between the Age and Product data to understand the Quartile Distribution of Age for each Product and Count Distribution of Product for different Age Groups.



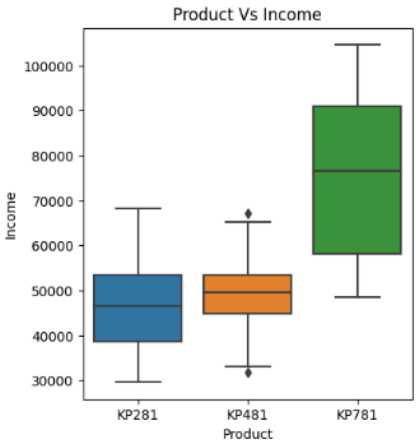


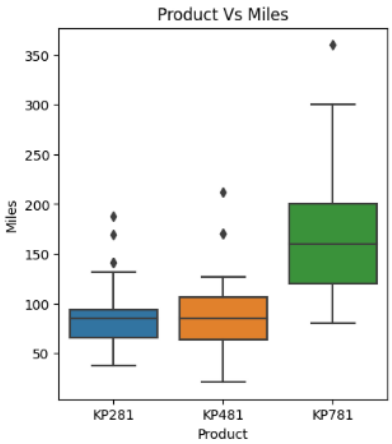
- Analyzed the impact on shopping trends for the available products between Male and Female genders and Marital Status of the customer.



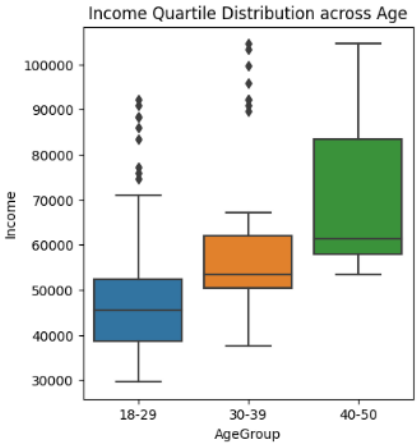


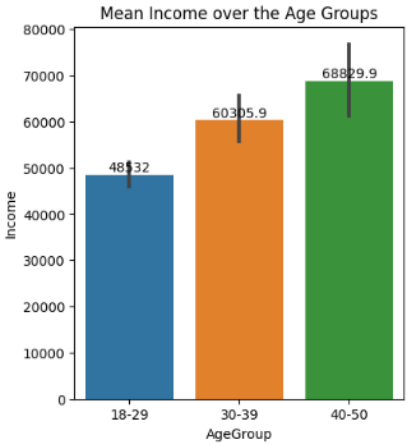
- Analyzed the impact on shopping trends for the available products based on the annual income and estimated miles usage of the customer.

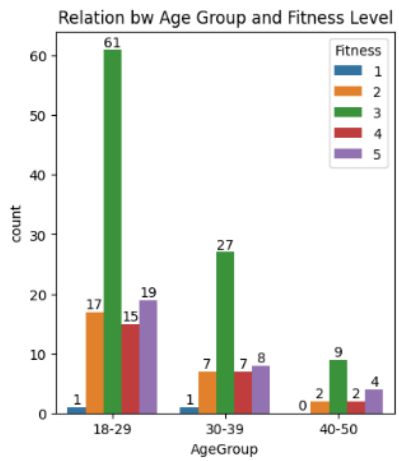


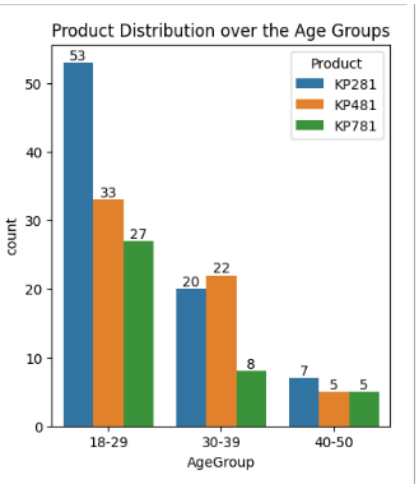


- Bi-variate analysis performed to identify trends in the income, fitness level and product distribution over the age groups.

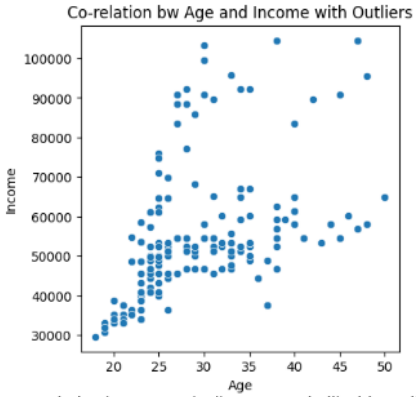


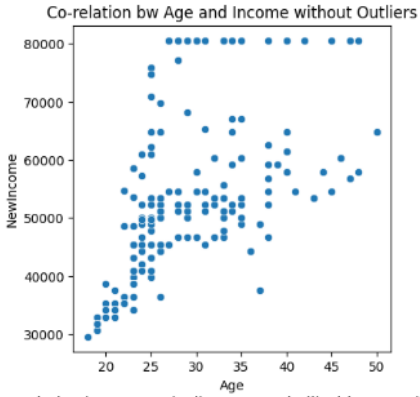


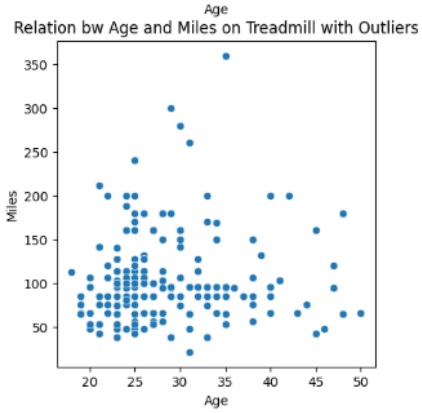


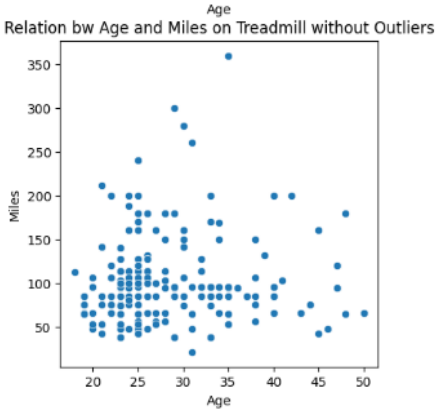


- Analyzed the co-relation between Age and Income and Age and Miles of Usage on Treadmill with and without outliers for a comprehensive understanding of the customer base.

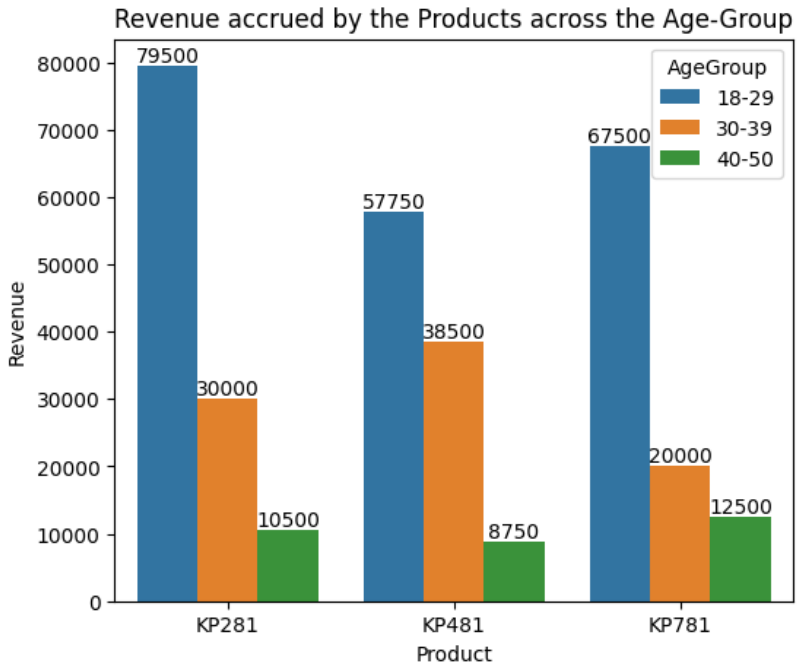








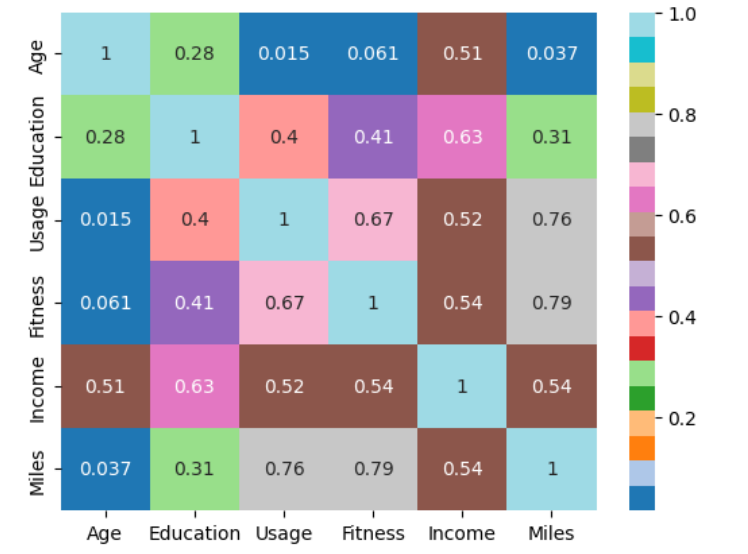
- Analyzed the revenue accrued by the sale of each product across the age group to generate insights about the profit leaders.



- Heatmap was generated to have insights on degree of correlation between multiple combination of columns.

```
sns.heatmap(correlation_matrix,annot=True,cmap='tab20')
```

```
<Axes: >
```



Business Insights from Bi-Variate Analysis:

- Median age of customers across the three available product range increases with the product offering (KP281<KP481<KP781) indicating the customers' preference shifts towards the more expensive and feature rich device with age.
- The youth population within the age bracket 18-29 contributes the highest to the overall sales volume. The sales volume witnesses a linear decrease as the age group increases indicating majority of the customer base is youth population. Furthermore, KP281 is the most selling product in the youth (18-29) and senior adult (40-50) population and KP481 being the highest selling product for people in 30-39 age group. It might be inferred with age as the income increases, the customer preference shifts to accommodate more features than what is offered by the base model and again reverts to base model with age progressing in the 40-50 bracket. This might be due to the additional responsibilities (senior citizen parents, young kids) that the customers might have to take up during this phase.
- Overall, male customers and Partnered customers have a higher purchase rate across the board. This might be indicative of a higher interest of men in fitness and curating offers, promotions and clubbing fitness products/supplements with KP781, for men might have positive impact in transitioning purchase decision to a higher tier product (men might end up purchasing KP781 when they originally came in with a mindset to purchase KP481).
- Purchase decision is directly impacted by the income and miles expected to walk/run on the treadmill. Higher income households by the income and miles expected to walk/run on the treadmill. Higher income households and users with expected higher miles to be registered on the treadmill tend to go for more feature rich product KP781 and the product tier decreases linearly with the Income and Miles usage.
- Overall, with age the median income and miles usage increases. It can be inferred from the previous observations, that customers on the higher age bracket may be persuaded to purchase KP781 considering the income data.
- Customers majorly assess themselves as average on the fitness level across the age groups. This might be a powerful utility to influence purchase decisions and navigate customers towards purchasing KP481 or higher due to the feature and usage limitations of KP281.
- Across the available product line, the maximum revenue is accrued by the youth customers (18-29 years old) which is the largest customer base. It is imperative that company develops features in demand by the older population as they could become the significant profit leaders due to the higher income and potential to purchase the higher tier products.

Customer Profile Generation:

Several Probability matrices were created based on multiple unique combinations of columns and the marginal and conditional probability of the customer purchasing either of the products calculated for a thorough customer profile.

```
#Probability Distribution of Product across gender
pd.crosstab([data.Gender],data.Product,margins=True,normalize=True).round(3)
```

Product	KP281	KP481	KP781	All
Gender				
Female	0.222	0.161	0.039	0.422
Male	0.222	0.172	0.183	0.578
All	0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender and Marital Status
pd.crosstab([data.Gender,data.MaritalStatus],[data.Product],margins=True,normalize=True).round(3)
```

	Product	KP281	KP481	KP781	All
Gender	MaritalStatus				
Female	Partnered	0.150	0.083	0.022	0.256
	Single	0.072	0.078	0.017	0.167
Male	Partnered	0.117	0.117	0.106	0.339
	Single	0.106	0.056	0.078	0.239
All		0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender,Age Group and Marital Status
pd.crosstab([data.Gender,data.MaritalStatus,data.AgeGroup],[data.Product],margins=True,normalize=True).round(3)
```

		Product	KP281	KP481	KP781	All
Gender	MaritalStatus	AgeGroup				
Female	Partnered	18-29	0.117	0.039	0.011	0.167
		30-39	0.022	0.039	0.011	0.072
		40-50	0.011	0.006	0.000	0.017
	Single	18-29	0.039	0.044	0.017	0.100
		30-39	0.028	0.028	0.000	0.056
		40-50	0.006	0.006	0.000	0.011
Male	Partnered	18-29	0.056	0.056	0.067	0.178
		30-39	0.039	0.044	0.028	0.111
		40-50	0.022	0.017	0.011	0.050
	Single	18-29	0.083	0.044	0.056	0.183
		30-39	0.022	0.011	0.006	0.039
		40-50	0.000	0.000	0.017	0.017
All			0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender, Age Group and Income Group
pd.crosstab([data.Gender,data.AgeGroup,data.IncomeGroup],data.Product,margins=True,normalize=True).round(3)
```

		Product	KP281	KP481	KP781	All
Gender	AgeGroup	IncomeGroup				
Female	18-29	30k-50k	0.122	0.067	0.000	0.189
		50k-70k	0.033	0.017	0.022	0.072
		90k+	0.000	0.000	0.006	0.006
	30-39	30k-50k	0.022	0.017	0.000	0.039
		50k-70k	0.028	0.050	0.000	0.078
		90k+	0.000	0.000	0.011	0.011
	40-50	50k-70k	0.017	0.011	0.000	0.028
Male	18-29	30k-50k	0.094	0.083	0.028	0.206
		50k-70k	0.039	0.017	0.044	0.100
		70k-90k	0.000	0.000	0.044	0.044
		90k+	0.000	0.000	0.006	0.006
		<30k	0.006	0.000	0.000	0.006
	30-39	30k-50k	0.022	0.000	0.000	0.022
		50k-70k	0.039	0.056	0.000	0.094
		70k-90k	0.000	0.000	0.006	0.006
		90k+	0.000	0.000	0.028	0.028
	40-50	50k-70k	0.022	0.017	0.000	0.039
		70k-90k	0.000	0.000	0.011	0.011
		90k+	0.000	0.000	0.017	0.017
All			0.444	0.333	0.222	1.000

```
pd.crosstab([data.Gender,data.IncomeGroup],data.Product,margins=True,normalize=True).round(3)
```

	Product	KP281	KP481	KP781	All
Gender	IncomeGroup				
Female	30k-50k	0.144	0.083	0.000	0.228
	50k-70k	0.078	0.078	0.022	0.178
	90k+	0.000	0.000	0.017	0.017
Male	30k-50k	0.117	0.083	0.028	0.228
	50k-70k	0.100	0.089	0.044	0.233
	70k-90k	0.000	0.000	0.061	0.061
	90k+	0.000	0.000	0.050	0.050
	<30k	0.006	0.000	0.000	0.006
All		0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender & Usage
pd.crosstab([data.Gender,data.Usage],data.Product,margins=True,normalize=True).round(3)
```

	Product	KP281	KP481	KP781	All
Gender	Usage				
Female	2	0.072	0.039	0.000	0.111
	3	0.106	0.078	0.000	0.183
	4	0.039	0.028	0.011	0.078
	5	0.006	0.017	0.017	0.039
	6	0.000	0.000	0.011	0.011
Male	2	0.033	0.039	0.000	0.072
	3	0.100	0.094	0.006	0.200
	4	0.083	0.039	0.089	0.211
	5	0.006	0.000	0.050	0.056
	6	0.000	0.000	0.028	0.028
	7	0.000	0.000	0.011	0.011
All		0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender and Age Group
pd.crosstab([data.Gender,data.AgeGroup],data.Product,margins=True,normalize=True).round(3)
```

Gender	AgeGroup	KP281	KP481	KP781	All
Female	18-29	0.156	0.083	0.028	0.267
	30-39	0.050	0.067	0.011	0.128
	40-50	0.017	0.011	0.000	0.028
Male	18-29	0.139	0.100	0.122	0.361
	30-39	0.061	0.056	0.033	0.150
	40-50	0.022	0.017	0.028	0.067
All		0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender & Fitness Level
pd.crosstab([data.Gender,data.Fitness],data.Product,margins=True,normalize=True).round(3)
```

Gender	Fitness	KP281	KP481	KP781	All
Female	1	0.000	0.006	0.000	0.006
	2	0.056	0.033	0.000	0.089
	3	0.144	0.100	0.006	0.250
	4	0.017	0.022	0.006	0.044
	5	0.006	0.000	0.028	0.033
Male	1	0.006	0.000	0.000	0.006
	2	0.022	0.033	0.000	0.056
	3	0.156	0.117	0.017	0.289
	4	0.033	0.022	0.033	0.089
	5	0.006	0.000	0.133	0.139
All		0.444	0.333	0.222	1.000

```
#Probability Distribution of Product across Gender & Education Level
pd.crosstab([data.Gender,data.Education],data.Product,margins=True,normalize=True).round(3)
```

Gender	Education	KP281	KP481	KP781	All
Female	13	0.000	0.006	0.000	0.006
	14	0.100	0.067	0.000	0.167
	15	0.011	0.000	0.000	0.011
	16	0.106	0.078	0.011	0.194
	18	0.006	0.011	0.022	0.039
	21	0.000	0.000	0.006	0.006
Male	12	0.011	0.006	0.000	0.017
	13	0.017	0.006	0.000	0.022
	14	0.067	0.061	0.011	0.139
	15	0.011	0.006	0.000	0.017
	16	0.111	0.094	0.072	0.278
	18	0.006	0.000	0.083	0.089
	20	0.000	0.000	0.006	0.006
	21	0.000	0.000	0.011	0.011
All		0.444	0.333	0.222	1.000

Recommendations:

1. Establishing fitness, well-being and nutrition awareness campaigns targeted towards a specific age group could influence the senior age groups (30-39 & 40-50) with high income and spending potential to purchase high tier product (KP781 & KP481).
2. Introduction of complimentary repair and maintenance service for high tier product (KP781) could influence first time customers to purchase the same, by addressing the maintenance concerns, which would help with higher revenues overall.
3. Since a higher percentage of users are male, setting up fitness classes, providing complimentary diet plans and overall exercise programs targeted towards women could have two-fold returns. First, it could encourage more women to purchase the available fitness equipment and secondly, it could encourage more men influenced by their partners to purchase the product.
4. Creation of a customer reward program to provide additional offers, promotions, services to returning customers with prior purchase history to purchase higher tier products as the high tier products are preferred by users with prior experience in health & fitness and overall exercise regimen.
5. Sales representatives should be trained appropriately to collect relevant information such as gender, age, approx. income, marital status, etc. during their conversations with potential customers as it would help calculate the probability of them purchasing a specific product which would help the sales representative deliver the sales pitch accordingly.