**Colab NoteBook Link**  
**https://colab.research.google.com/drive/1RvnychR23DtRwB6RkI3oedU7AVTAjABL?usp=sharing**  
  
**PDF Notebook Link  
  
https://drive.google.com/file/d/1PGz15igJixLWJ\_UxR0KDoFmFkz\_tGFDU/view?usp=sharing**

**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.

1. Perform descriptive analytics **to create a customer profile** for each AeroFit treadmill product by developing appropriate tables and charts.
2. 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.

**Dataset**

The company collected the data on individuals who purchased a treadmill from the AeroFit stores during the prior three months. The dataset has the following features:

|  |  |
| --- | --- |
| 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.

**Problem Statement and Solutioning**

**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**

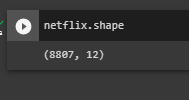
 We basically have to identify the features of the target audience for each type of treadmill (**KP281, KP481, KP781**) offered by the company in order to provide a better recommendation of the treadmills to the new customers. This could increase the company's sales and business.

* Below can be an approach :

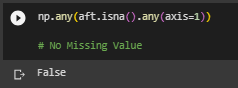
Analyzing that in the last 3 months, what type of customers (based on their gender, marital status, age, income, education, fitness, and usage) have opted for what type of Treadmill.

To infer these insights, we can create two-way contingency tables between treadmill type and customer characteristics mentioned above using pandas crosstab and can calculate marginal probabilities, joint probabilities, and conditional probabilities to suggest the treadmill to the right set of customers.

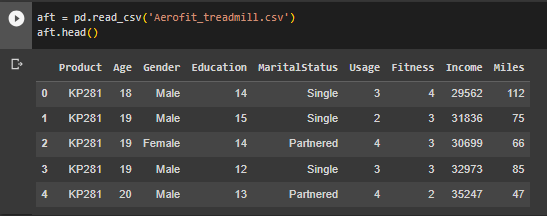
**Basic Metrics**:

* There are 180 records of customer purchases in the dataset.
* There are three types of treadmills in the dataset  (**KP281, KP481, KP781**).
* The Shape of the data is (180,9).  
    
  ****
* There are no-null values present in any column Product, Age, Gender, Education, MaritalStatus, Usage, Fitness, Income, Miles.

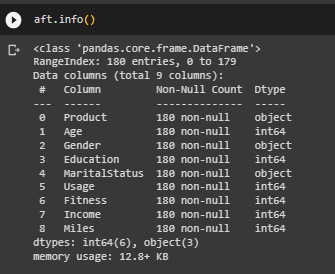
np.any(aft.isna().any(axis=1))

****

* Top 5 rows in the data set to have a glimpse of data.

****

* Column**Product, Gender, and MaritalStatus** are **object**types (since they contain **string**), all other columns Age, Education, Usage, Fitness, Income, and Miles are**Integer types.**

****

* Binned and Categorized **Age, Income, and Education** columns in new Columns **AgeCategory**, **IncomeCategory,**and **EducationCategory respectively**using the below pandas cut code.

aft['AgeCategory'] = pd.cut(aft['Age'], bins = [15,25,35,55], labels = ['Youth','Middle Age Adults','Older Adults'])

aft['AgeCategory'] = aft['AgeCategory'].astype(str)

aft.describe(include=object)

aft['IncomeCategory'] = pd.cut(aft['Income'], bins = [10000,30000,50000,70000,90000,110000], labels = ['Lower Middle','Middle','Upper Middle','Wealthy','Very Wealthy'])

aft['IncomeCategory'] = aft['IncomeCategory'].astype(str)

aft.describe(include=object)

aft['EducationCategory'] = pd.cut(aft['Education'], bins = [0,12,16,18,22], labels = ['Less Educated','Moderately Educated','Highly Educated','Very Highly Educated'])

aft['EducationCategory'] = aft['EducationCategory'].astype(str)

aft.describe(include=object)

The categorization criteria are mentioned below :

# Age

# 15-25 - Youth

# 25-35 - Middle Age Adults

# 35-55 - Older Adults

# Income

# 10k-30k                 - Lower Middle

# 30k-50k                 - Middle

# 50k-70k                 - Upper Middle

# 70k-90k                 - Wealthy

# 90k and 110k            - Very Wealthy

# Education

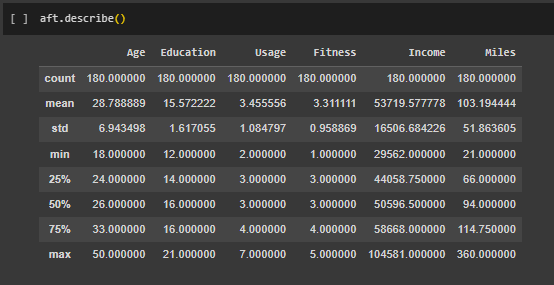
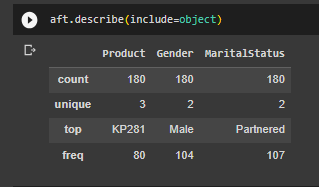
# 0-12                 - Less Educated

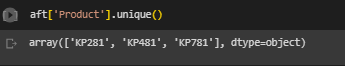
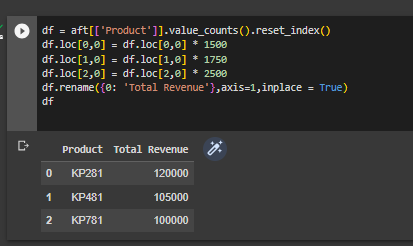
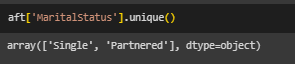
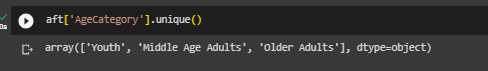
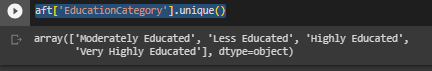
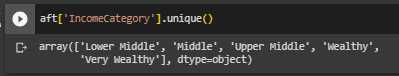
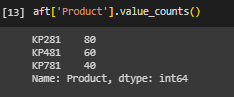
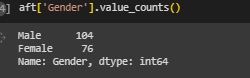
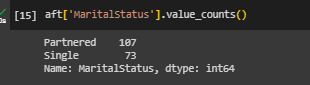
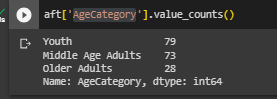
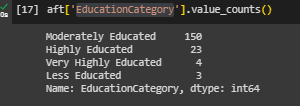
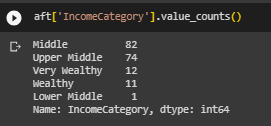
# 12-16                - Moderately Educated

# 16-18                - Highly Educated

# 18-22                - Very Highly Educated

* The statistical summary can be found using **describe**method mentioned below.

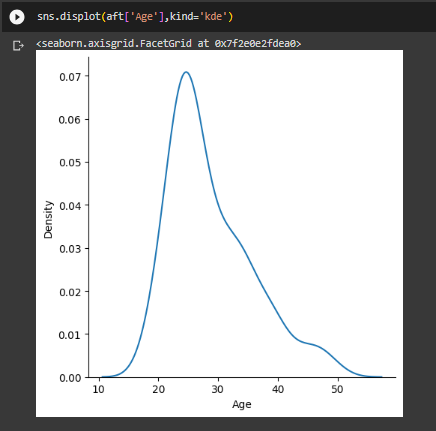
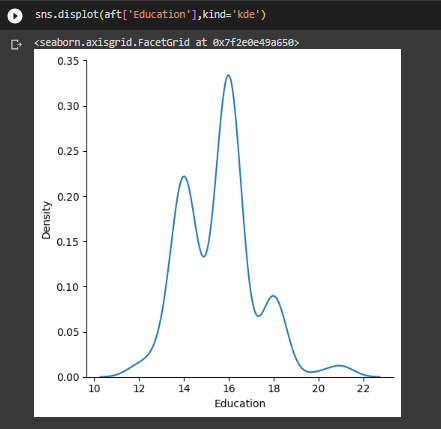
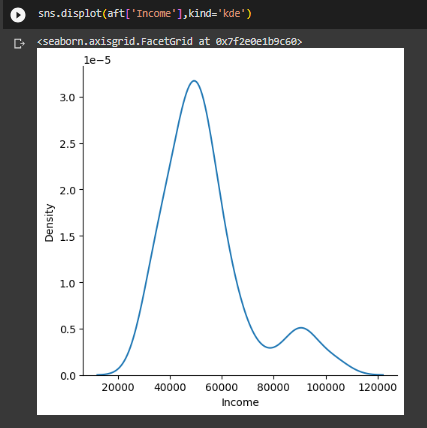
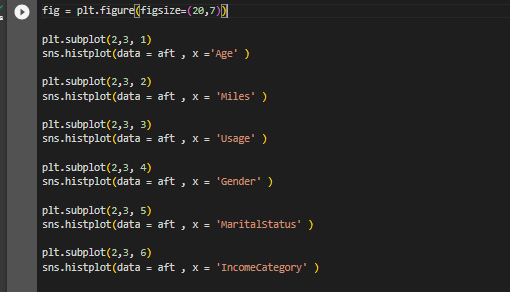
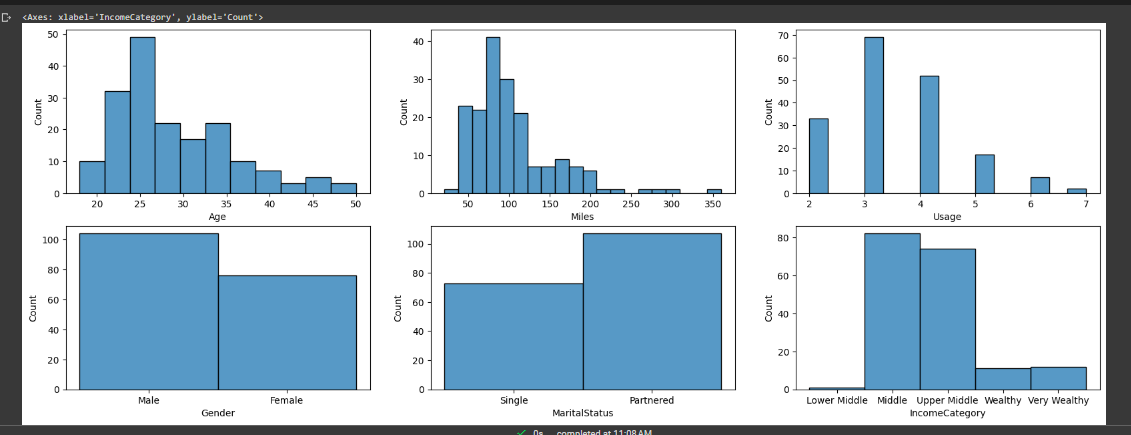
****  
  
****  
  
​​2. **Non-Graphical Analysis: Value counts and unique attributes**

* There are 3 types of treadmills 'KP281', 'KP481', 'KP781'. The total Sales amount for KP281 is **120000 dollars,** for **KP481 is 105000 dollars**, and for **KP781 is 100000 dollars**.  
    
  ****  
    
  ****
* **There are 2 Unique genders ( Male and Female )  
    
  **
* **There are 2 Unique marital status (Single, Partnered)   
    
  **
* **We have added 3 new columns after categorizing Age, Education, and Income as mentioned above.**
* **There are 3 Unique Age Categories.  
    
  **
* **There are 4 Unique Education Categories.**  
    
  ****
* **There are 5 Unique Income Categories.**  
  ****
* **Product Value\_Counts: There are 80 treadmills sold of type KP281, 60 treadmills sold of type KP481, and 40 treadmills sold of type KP481  in the last 3 months.   
    
  **
* **Gender Value\_Counts: There are 104 Males and 76 Females who have purchased the treadmills.  
    
  **
* **MaritalStatus Value\_Counts:  There are 107 Partnered and 73 Singles who have purchased the treadmills.  
    
  **
* **AgeCategory Value\_Counts: Most customers are in the Age Category of Youth (18-25 years) and Middle Age Adults (25-35 years)  
    
  **
* **EducationCategory Value\_Counts: Most customers are moderately Educated (12-16 years)**  
  ****
* **IncomeCategory Value\_Counts: Annual Salary of most customers is between 30K to 70K dollars consisting of Middle and Upper Middle Class.**  
  ****

3.  **Visual Analysis - Univariate, Bivariate after pre-processing of the data**

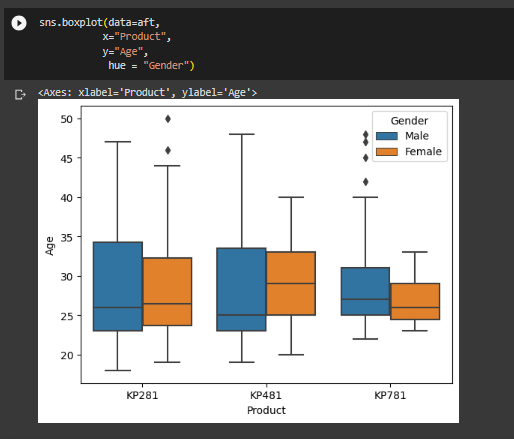
3.1 For continuous variable(s): Distplot, countplot, histplot for univariate analysis

\*Covering **5.1, 5.2. 5.3** as well.

* **Most customers are of age 20-30 years.**
* **The minimum age for customers is 18 years and the max is 50 years.**
* ****
* **Customers with qualification of around 16 years have purchased the most treadmill.**
* **The minimum education qualification for customers is 12 years and the max is 21 years.  
    
  **
* **Customers with an Annual Income of around 50K dollars have purchased the most treadmill.**
* **The minimum Annual Income for customers is around 30k dollars and the max is around 105k dollars.**
* ****
* **Treadmills of type KP281 are sold the most followed by KP481 and KP781.  
    
  **
* **Most customers are of age 20-30 years.**
* **Most customers would walk/run around 50-100 Miles per week.**
* **Most customers plan to use the treadmill for 3 to 4 days.**
* **There are around 11 percent (~20 in number) more male customers than female customers.**
* **There are around 19 percent (~34 in number ) more partnered customers than single customers.**
* **  
  **

3.2 For categorical variable(s): **Boxplot**

Covering **5.1, 5.2. 5.3** as well

* The median age of male customers for Treadmill type KP281 and KP481 is less than for female customers.
* The median age of male customers for Treadmill type KP781 is more than that for female customers.
* ****  
   **Categorical Vs Categorical using CountPlot**
* Treadmill-type KP781 is brought mostly by customers having an annual income of 50k dollars or more.
* Treadmill-type KP481 is brought mostly by customers having an annual income of around 30-70k dollars.
* Treadmill-type KP281 is brought mostly by customers having an annual income of around 30-50k dollars.
* Most of the treadmills are purchased by Youth and Middle Age Adults.
* Moderately Educated customers (12-16 years) buy KP481 and KP281 treadmills more.
* Most of the KP781 treadmills are brought by customers having Top Self Rated Fitness of 5.
* Mostly Males buy KP781 treadmills.
* Partnered customers buy treadmills more than single customers.
* fig = plt.figure(figsize=(17,11))

plt.subplot(2,3, 1)

sns.countplot(data=aft, x="Product", hue = "IncomeCategory")

plt.subplot(2,3, 2)

sns.countplot(data=aft, x="Product", hue = "AgeCategory")

plt.subplot(2,3, 3)

sns.countplot(data=aft, x="Product", hue = "EducationCategory")

plt.subplot(2,3, 4)

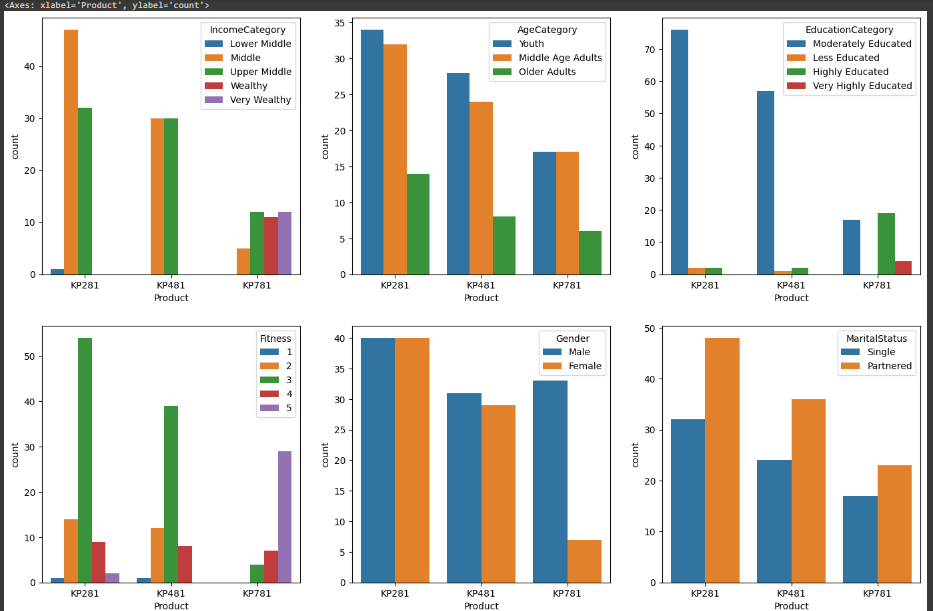
sns.countplot(data=aft, x="Product", hue = "Fitness")

plt.subplot(2,3, 5)

sns.countplot(data=aft, x="Product", hue = "Gender")

plt.subplot(2,3, 6)

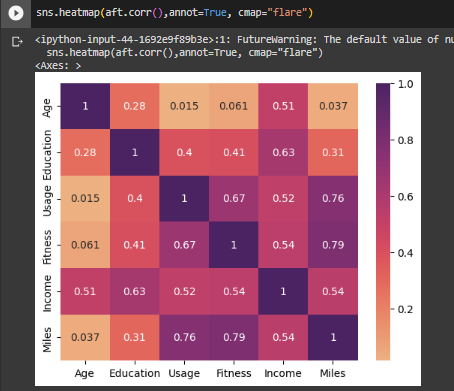
sns.countplot(data=aft, x="Product", hue = "MaritalStatus")

****

3.3 For correlation: Heatmaps, Pairplots

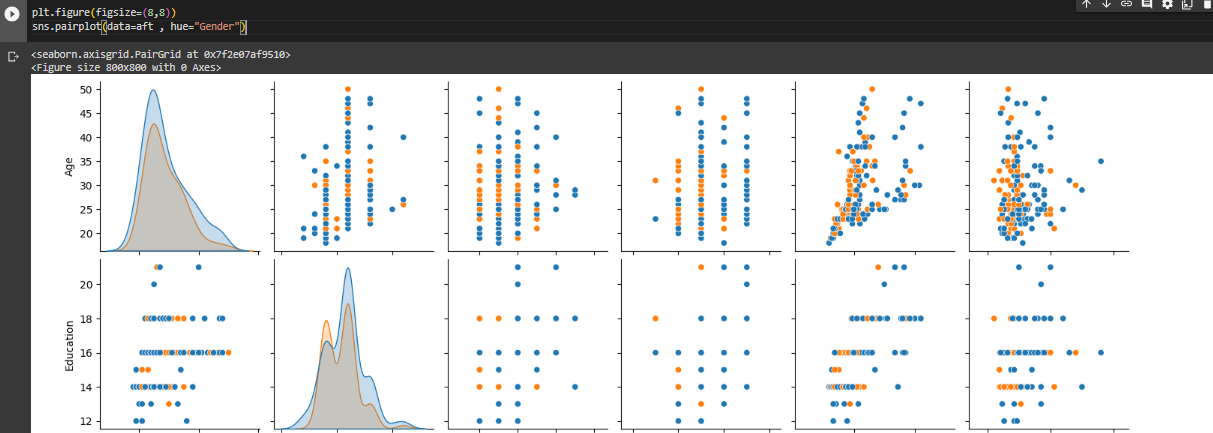
Covering **5.1, 5.2. 5.3** as well  
  
**Correlation using Heatmaps**

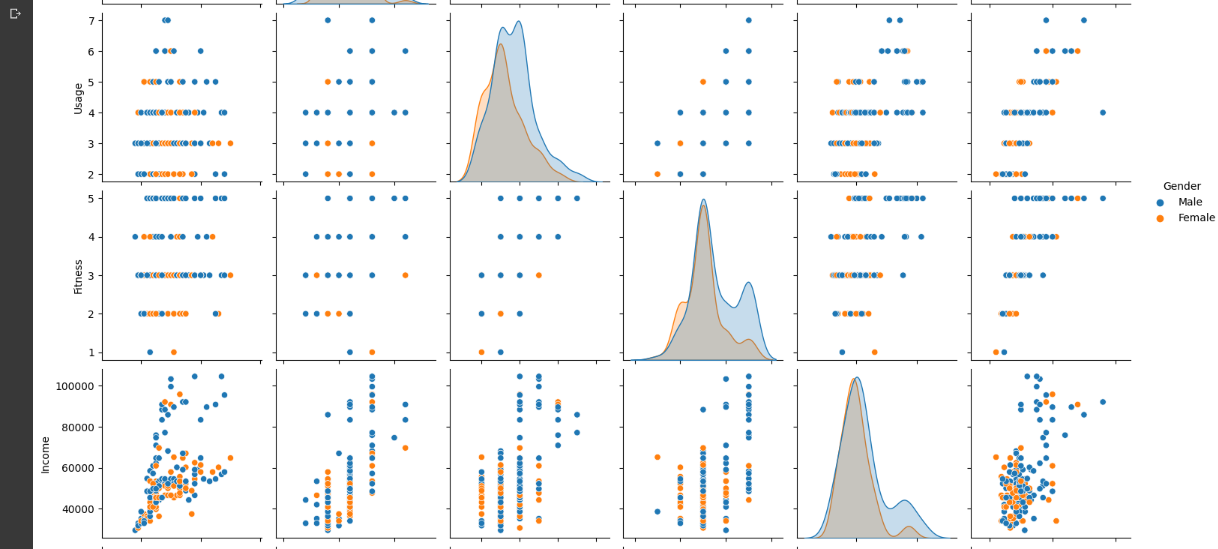
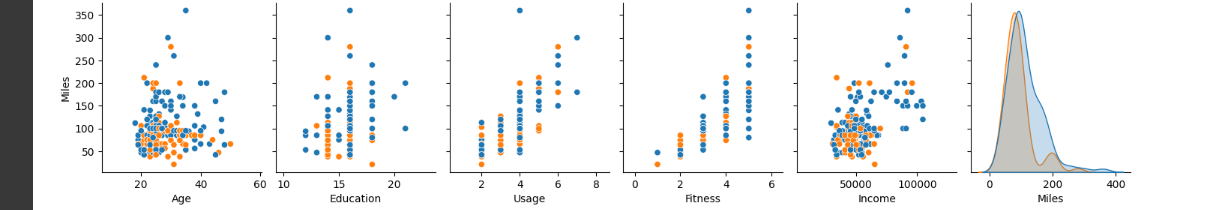
* **The number of Miles**covered by customers is **highly correlated to customer fitness**.
* **The number of Miles**covered by customers is **highly correlated to the customer usage in a week**.
* **Education and Annual Income**are also correlated to some extent.

****

**Correlation using PairPlot**

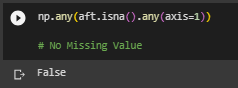
* + **Male customers tend to use the treadmill more than female customers.**
  + **Partnered customers tend to use the treadmill more than single customers.**

****

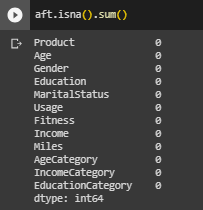
****  
****  
 **4. Missing Value & Outlier check (Treatment optional)**

 There are no-null values present in any column **Product, Age, Gender, Education, MaritalStatus, Usage, Fitness, Income, Miles**.

np.any(aft.isna().any(axis=1))

****

aft.isna().sum()

****

**Outlier using BoxPlot**

fig = plt.figure(figsize=(20,7))

plt.subplot(2,3, 1)

sns.boxplot(data = aft , x ='Age' )

plt.subplot(2,3, 2)

sns.boxplot(data = aft , x = 'Education' )

plt.subplot(2,3, 3)

sns.boxplot(data = aft , x = 'Income' )

plt.subplot(2,3, 4)

sns.boxplot(data = aft , x = 'Miles' )

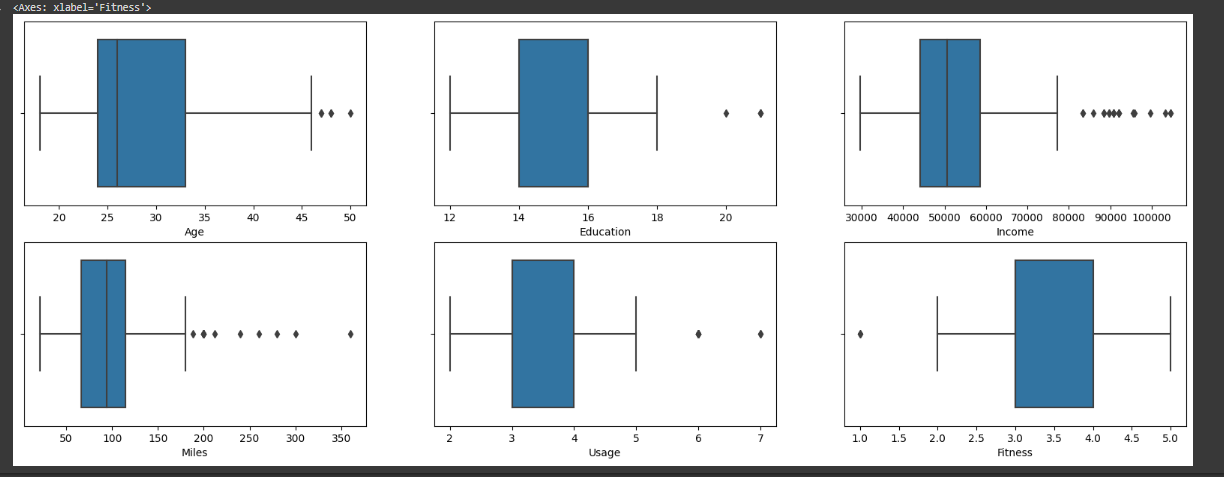
plt.subplot(2,3, 5)

sns.boxplot(data = aft , x = 'Usage' )

plt.subplot(2,3, 6)

sns.boxplot(data = aft , x = 'Fitness' )

* There are very less customers having an age of more than 45 years.
* There are very less customers having education more than 18 years.
* Most customers are having a median annual salary of around 50k dollars.
* Few customers plan to use the treadmill for 6 to 7 days a week.
* Hardly any customer has rated his/her fitness as 1.

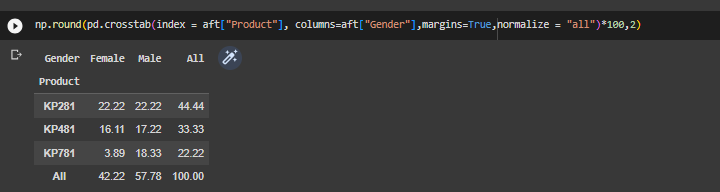
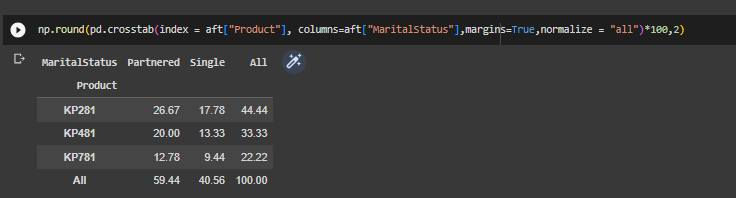
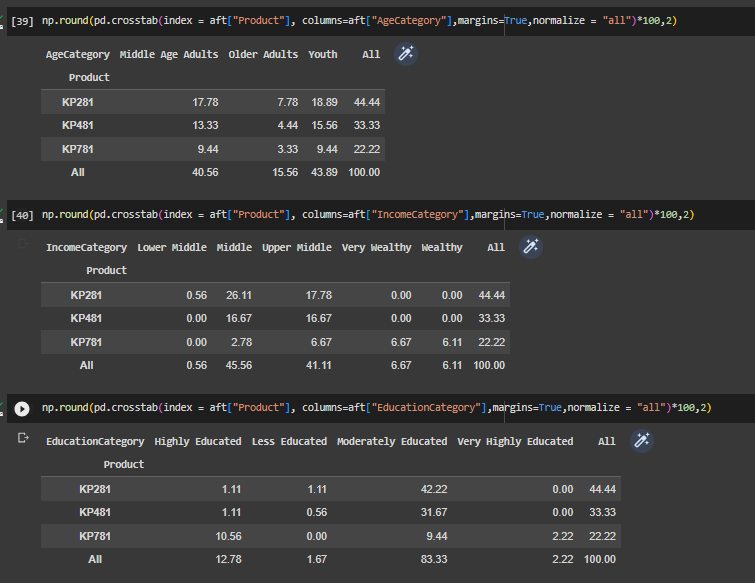
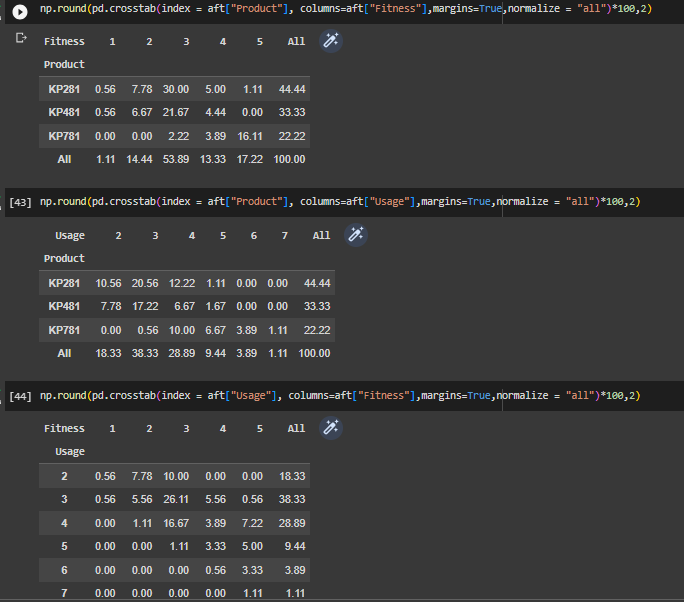
****

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

* Insights based on Non-Graphical Analysis covered in point 2
* Insights based on Visual Analysis covered in point 3.

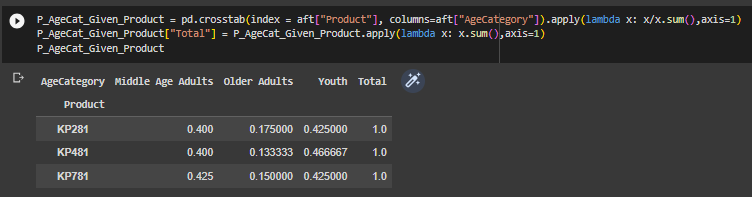
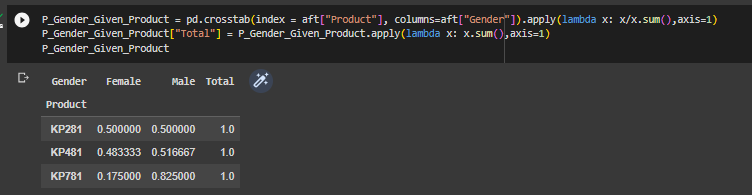
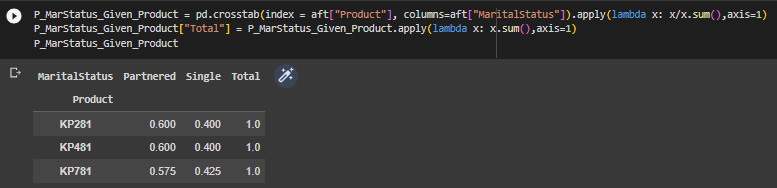
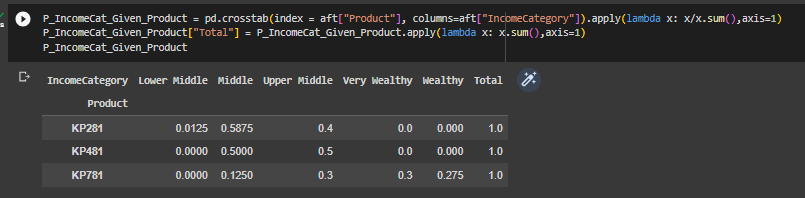
6. **Business Insights**

**MARGINAL AND JOINT PROBABILITIES**

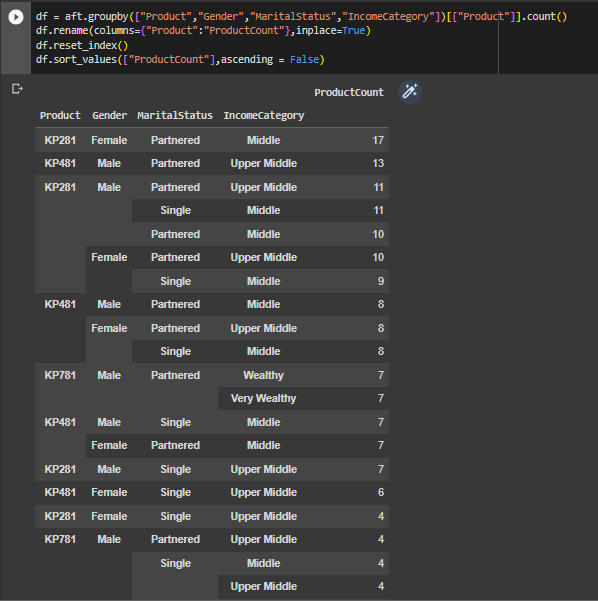
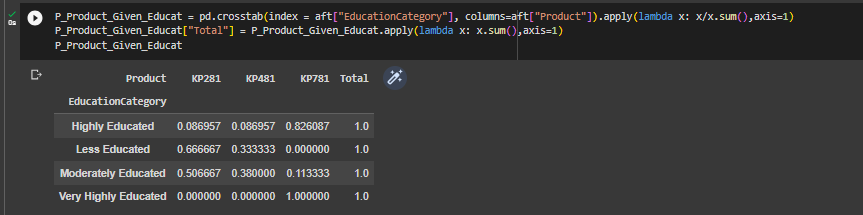
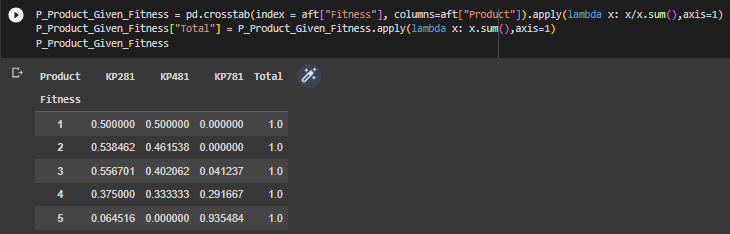
* **57.78% of Males and 42.22% of Females have brought the treadmill.   
    
  **
* **59.44% of Partnered and 40.56% of Single customers have brought the treadmill.   
    
  **
* **43.89% of Youth, 40.56% of Middle Age Adults, and 15.56% of Older Adults have brought the treadmill.**
* **45.56% of the Middle Class and 41.11% of the Upper Middle-Class customers have brought the treadmill.**
* **83.33% of customers are Moderately Educated (12-16 years)**  
  ****
* **53.89% of customers have rated themselves as 3 on the Fitness level and around 17.22% as 5.**
* **38.33% of customers plan to use the treadmill for 3 days a week and  28.89% of customers plan to use the treadmill for 4 days a week**
* **  
    
  CONDITIONAL PROBABILITIES  
    
  Of Customer Characteristics given that they have brought a Treadmill Product**

P(Parameter|Product)

**P(AgeCategory|Product)**

* **There are 42.5% of Youth, 42.5% of Middle Age Adults, and 15% of Older Adults given that they brought KP781.**
* **There are 46.66% of Youth, 40 % of Middle Age Adults, and 13.3% of Older Adults given that they brought KP481.**
* **There are 42.5% of Youth, 40% of Middle Age Adults, and 17.5% of Older Adults given that they brought KP281.   
    
    
    
  P(Gender|Product)**
* **There are 82.5% of Males and 17.5% of Females given that they brought KP781.**
* **There are 51.6% of Males and 48.33% of Females given that they brought KP481.**
* **There are 50% of Males and 50% of Females given that they brought KP281.   
    
    
    
  P(MaritalStatus|Product)**
* **There are 57.5% of Partnered and 42.5% of Single customers given that they brought KP781.**
* **There are 60% of Partnered and 40% of Single customers given that they brought KP481.**
* **There are  60% of Partnered and 40% of Single customers given that they brought KP281.   
    
    
    
  P(IncomeCategory|Product)**
* **There are 27.5% of Wealthy, 30% of Very Wealthy, and 30% of Upper Middle-class customers given they brought KP781**
* **There are 50% of Upper Middle and 50% of Middle-class customers given they brought KP481**
* **There are 58.5% of Middle and 40% of Upper Middle-class customers given they brought KP281    
    
  **

8. **Recommendations**- Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand.

* **Recommendation 1: Male**, **Partnered**, **Wealthy**(**70K to 90K dollars per annum income**), and **Very Wealthy**(**90K and above dollars per annum income**) customers are more likely to buy a Treadmill of product type **KP781**
* **Recommendation 2: Male**, **Partnered, and** **Upper Middle Class**(**50K to 70K dollars per annum income**) customers are more likely to buy a Treadmill of product type **KP481**
* **Recommendation 3**: **Female**, **Partnered, and** **Middle Class**(**30K to 50K dollars per annum income**) customers are more likely to buy a Treadmill of product type **KP281   
    
  **
* **Recommendation 4:**If a customer is**highly educated (18-22 years),** he/she is more likely to buy Treadmill of product type **KP781**  
    
  ****
* **Recommendation 5:**If a customer has rated themselves as 5 on the Fitness chart**,** he/she is more likely to buy Treadmill of product type **KP781**  
    
  ****