Aerofit Business case
All the codes are written on Jupyter and link to the workbook is (link)
Workbook drive (__________)<- Click here

Q1. Defining Problem Statement and Analysing basic metrics **(10 Points)**

- Observations on shape of data, data types of all the attributes, conversion of categorical attributes to 'category' (If required), statistical summary

Problem Statement

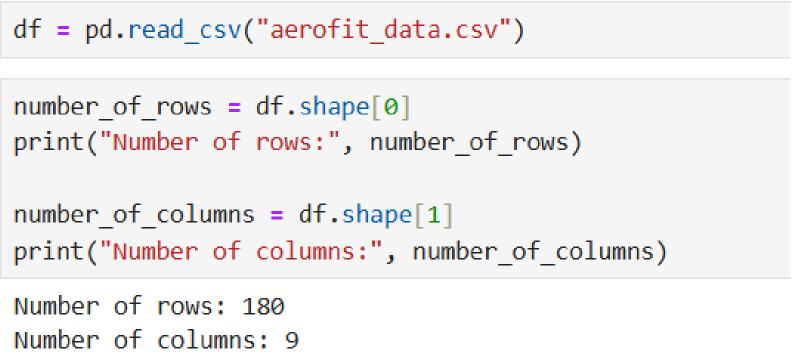
The goal is to analyze customer data related to various products (KP281, KP781) to understand the characteristics and usage patterns of the customers. This includes examining demographic details, usage metrics, fitness levels, and income. The insights derived can help in better understanding the customer segments, their behaviors, and potentially improving marketing strategies and product offerings.

Data Analysis

1. Observations on Shape of Data

First, let's look at the shape of the dataset to understand the number of records and attributes we have:

Code is as:



```
df = pd.read_csv("aerofit_data.csv")
```

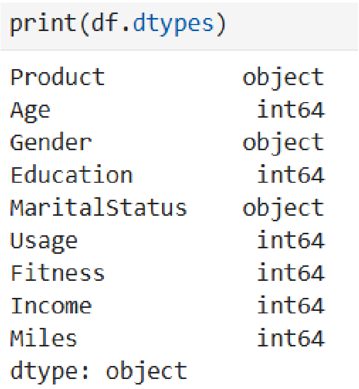
```
number_of_rows = df.shape[0]
print("Number of rows:", number_of_rows)

number_of_columns = df.shape[1]
print("Number of columns:", number_of_columns)
```

```
Number of rows: 180
Number of columns: 9
```

2. Data Types of All Attributes

Next, we will check the data types of each column to identify which attributes are categorical and which are numerical:

Code is as:



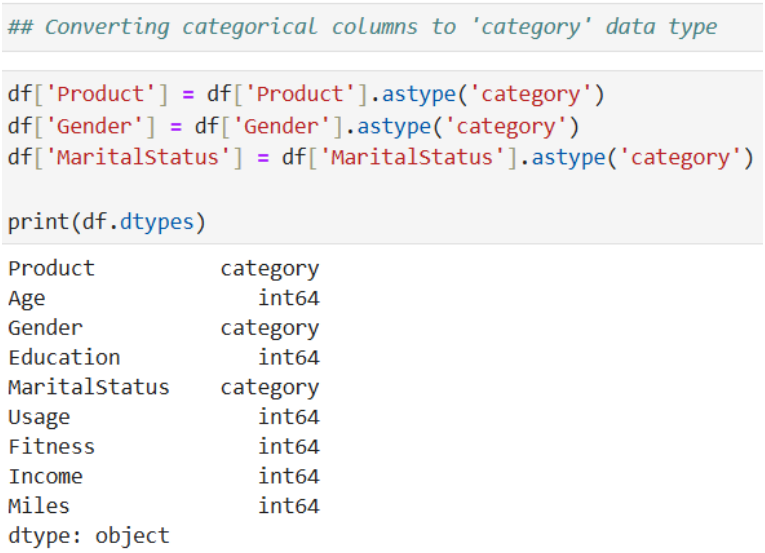
```
print(df.dtypes)
```

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

3. Conversion of Categorical Attributes to 'Category'

Categorical attributes identified (if not already converted) will be converted to the 'category' data type:

Code is as:



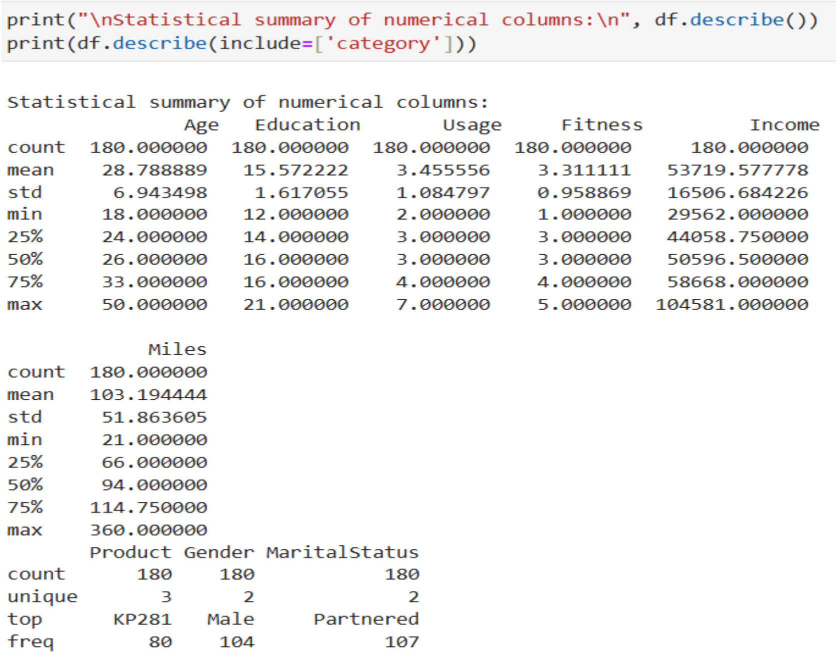
```
## Converting categorical columns to 'category' data type

df['Product'] = df['Product'].astype('category')
df['Gender'] = df['Gender'].astype('category')
df['MaritalStatus'] = df['MaritalStatus'].astype('category')

print(df.dtypes)
```

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

4. Statistical Summary

A statistical summary will be provided to give an overview of the numerical attributes:

code is as:



```
print("\nStatistical summary of numerical columns:\n", df.describe())
print(df.describe(include=['category']))
```

```
Statistical summary of numerical columns:
               Age   Education       Usage     Fitness         Income
count  180.000000  180.000000  180.000000  180.000000     180.000000
mean    28.788889   15.572222    3.455556    3.311111   53719.577778
std      6.943498    1.617055    1.084797    0.958869   16506.684226
min     18.000000   12.000000    2.000000    1.000000   29562.000000
25%     24.000000   14.000000    3.000000    3.000000   44058.750000
50%     26.000000   16.000000    3.000000    3.000000   50596.500000
75%     33.000000   16.000000    4.000000    4.000000   58668.000000
max     50.000000   21.000000    7.000000    5.000000  104581.000000

            Miles
count  180.000000
mean   103.194444
std     51.863605
min     21.000000
25%     66.000000
50%     94.000000
75%    114.750000
max    360.000000
       Product Gender MaritalStatus
count      180    180           180
unique       3      2             2
top      KP281   Male     Partnered
freq        80    104           107
```

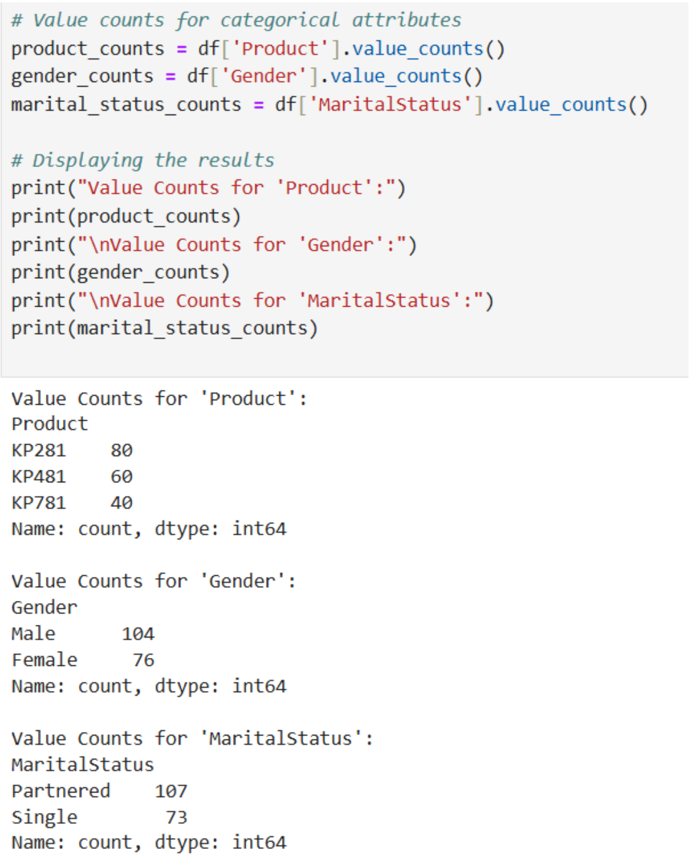
Q2. Non-Graphical Analysis: Value counts and unique attributes **(10 Points)**:

To perform a non-graphical analysis of the dataset, we will focus on obtaining value counts and identifying unique attributes for the categorical columns. This provides insight into the distribution of categories and the diversity within the dataset.

1. Value Counts

We will use the value_counts() method to get the counts of unique values in each categorical column.

Code is as:



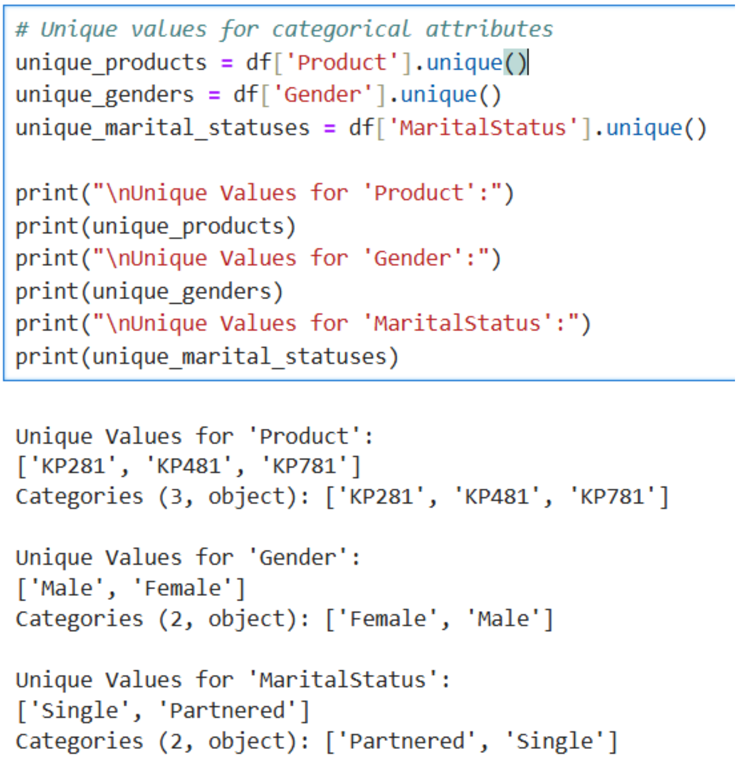
```
# Value counts for categorical attributes
product_counts = df['Product'].value_counts()
gender_counts = df['Gender'].value_counts()
marital_status_counts = df['MaritalStatus'].value_counts()

# Displaying the results
print("Value Counts for 'Product':")
print(product_counts)
print("\nValue Counts for 'Gender':")
print(gender_counts)
print("\nValue Counts for 'MaritalStatus':")
print(marital_status_counts)
```

```
Value Counts for 'Product':
Product
KP281    80
KP481    60
KP781    40
Name: count, dtype: int64

Value Counts for 'Gender':
Gender
Male      104
Female     76
Name: count, dtype: int64

Value Counts for 'MaritalStatus':
MaritalStatus
Partnered    107
Single        73
Name: count, dtype: int64
```

2. Unique Attributes

We will use the unique() method to list the unique values for each categorical column.

Code is as:



```
# Unique values for categorical attributes
unique_products = df['Product'].unique()
unique_genders = df['Gender'].unique()
unique_marital_statuses = df['MaritalStatus'].unique()

print("\nUnique Values for 'Product':")
print(unique_products)
print("\nUnique Values for 'Gender':")
print(unique_genders)
print("\nUnique Values for 'MaritalStatus':")
print(unique_marital_statuses)
```

```
Unique Values for 'Product':
['KP281', 'KP481', 'KP781']
Categories (3, object): ['KP281', 'KP481', 'KP781']

Unique Values for 'Gender':
['Male', 'Female']
Categories (2, object): ['Female', 'Male']

Unique Values for 'MaritalStatus':
['Single', 'Partnered']
Categories (2, object): ['Partnered', 'Single']
```

Q3. Visual Analysis - Univariate & Bivariate **(30 Points)**

1. For continuous variable(s):Distplot, countplot, histogram for univariate analysis (10 Points)

2. For categorical variable(s): Boxplot (10 Points)
3. For correlation: Heatmaps, Pairplots(10 Points)

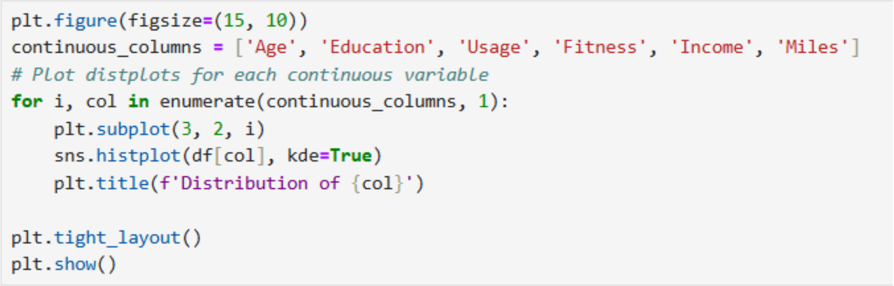
For continuous variable(s):Distplot, countplot, histogram for univariate analysis (10 Points):

To conduct a visual analysis of continuous variables in the dataset, we will use various types of plots.

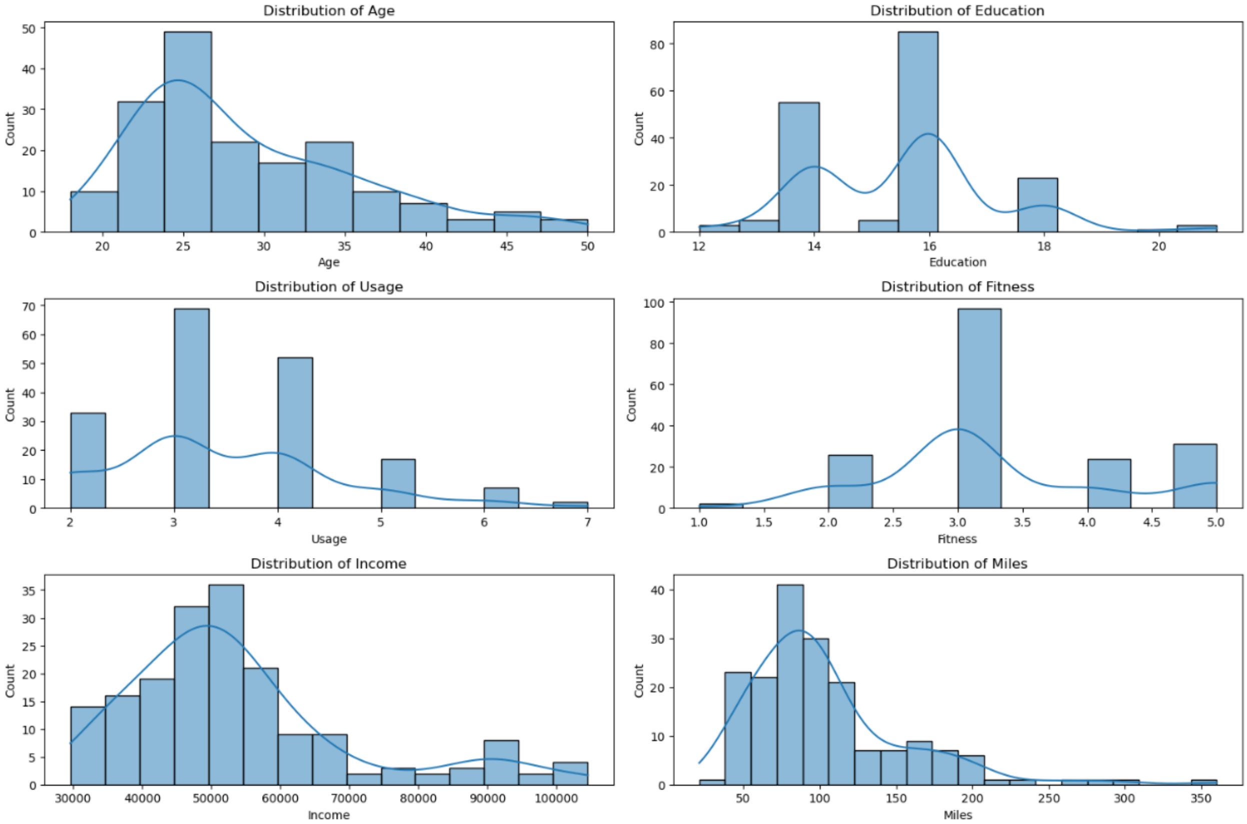
Univariate Analysis

1. **Distplot**: This shows the distribution of a continuous variable.

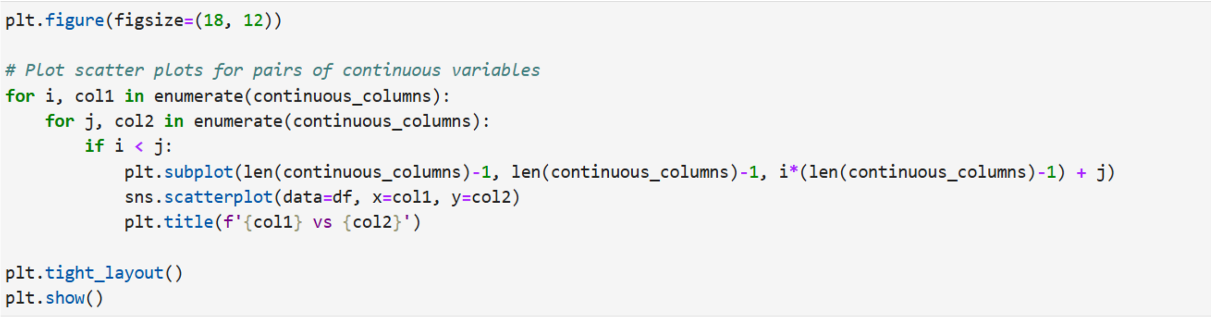
   code is as:



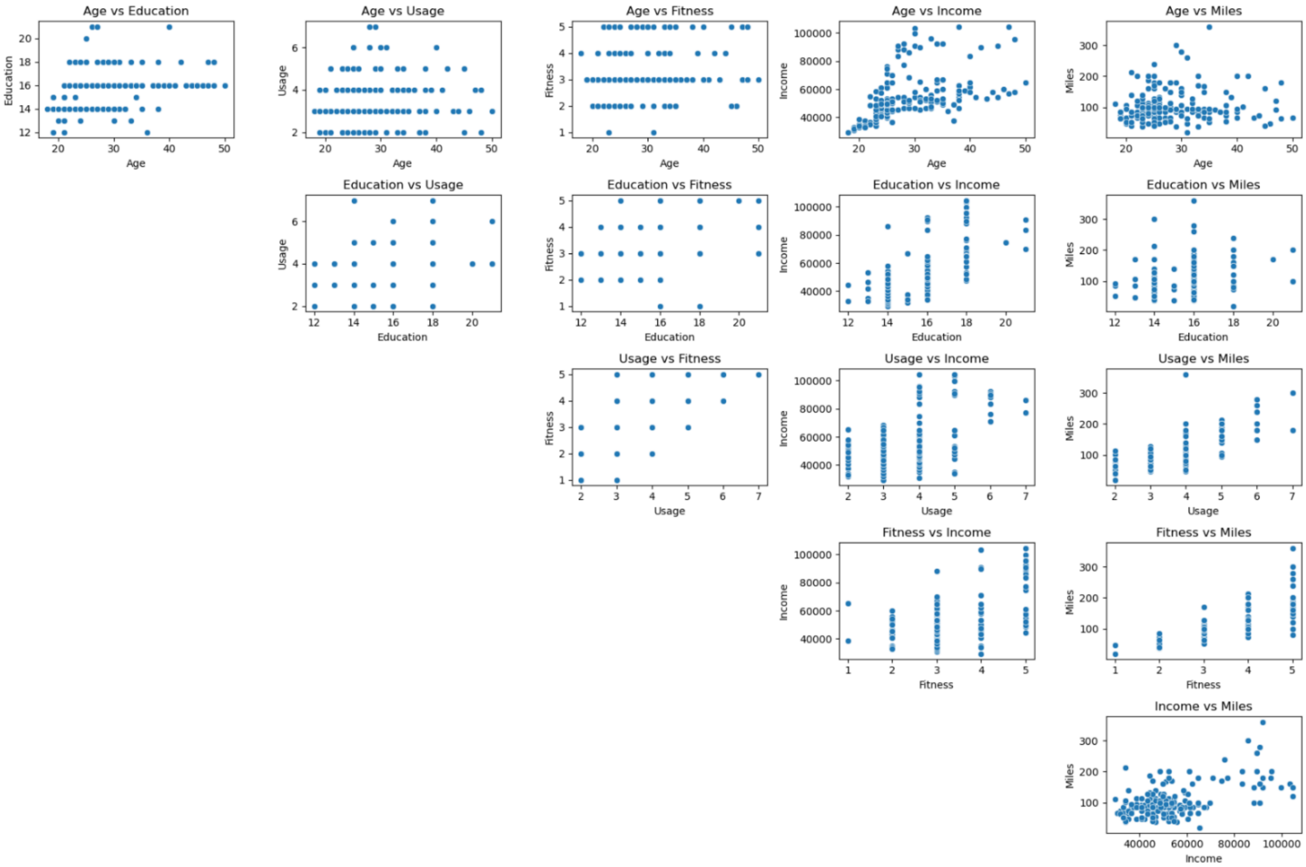
```python
plt.figure(figsize=(15, 10))
continuous_columns = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
# Plot distplots for each continuous variable
for i, col in enumerate(continuous_columns, 1):
    plt.subplot(3, 2, i)
    sns.histplot(df[col], kde=True)
    plt.title(f'Distribution of {col}')

plt.tight_layout()
plt.show()
```



2. **Histogram**: This shows the frequency distribution of a continuous variable.
Code is as:



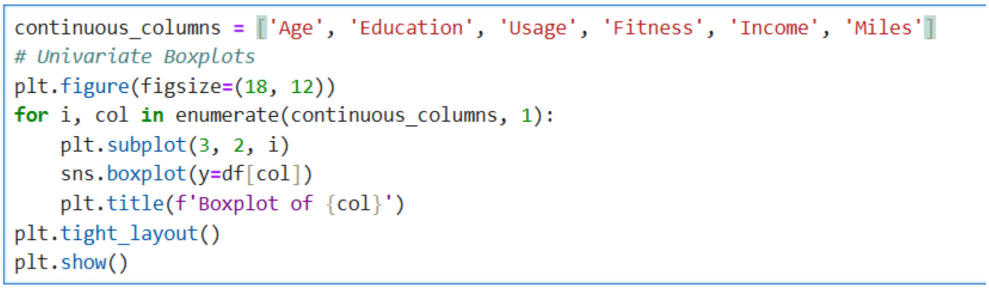
```python
plt.figure(figsize=(18, 12))

# Plot scatter plots for pairs of continuous variables
for i, col1 in enumerate(continuous_columns):
    for j, col2 in enumerate(continuous_columns):
        if i < j:
            plt.subplot(len(continuous_columns)-1, len(continuous_columns)-1, i*(len(continuous_columns)-1) + j)
            sns.scatterplot(data=df, x=col1, y=col2)
            plt.title(f'{col1} vs {col2}')

plt.tight_layout()
plt.show()
```



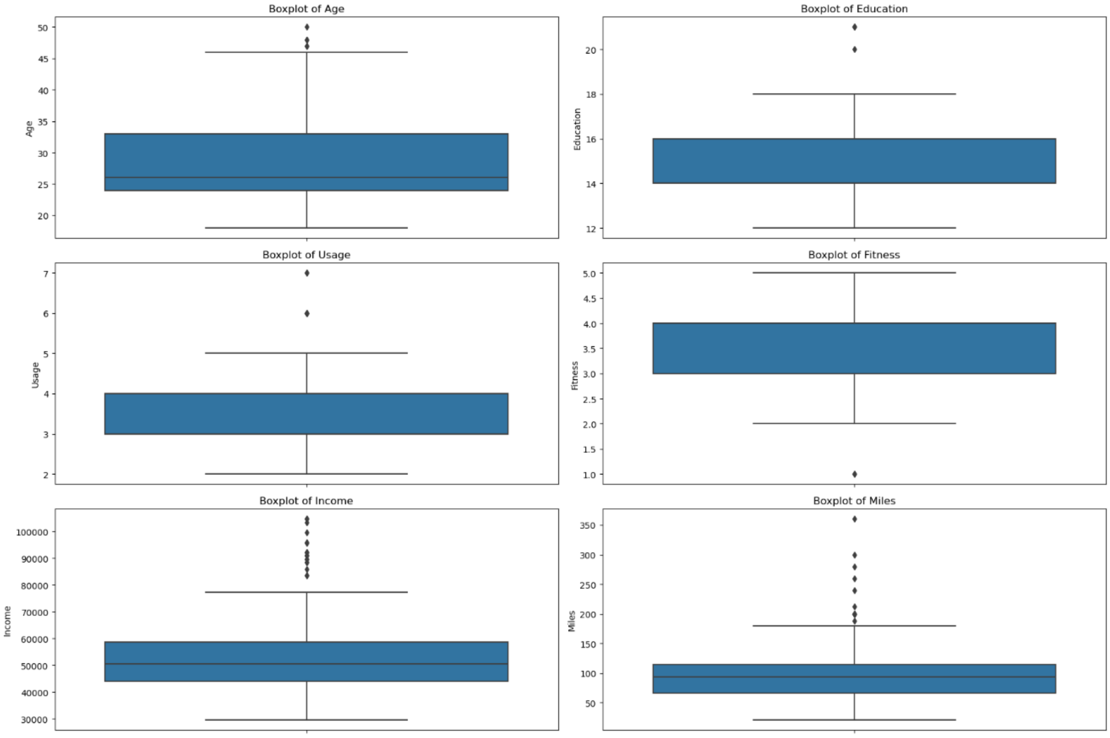
3.2) For categorical variable(s): Boxplot (10 Points)

Here's how we can boxplot on categorical variable:

Code is as:



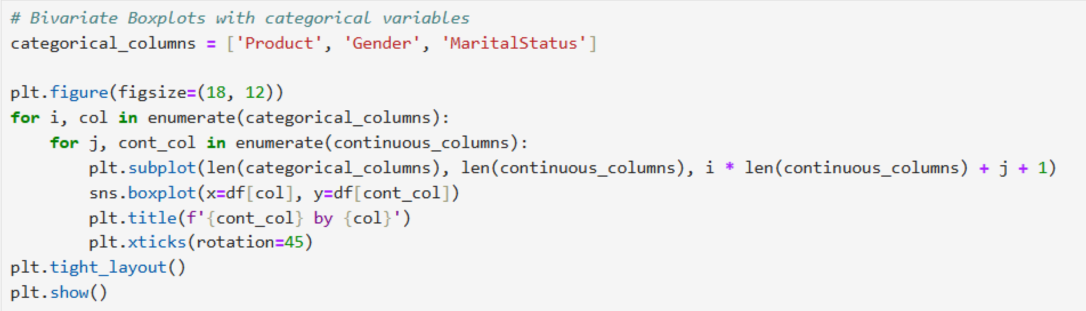
```
continuous_columns = ['Age', 'Education', 'Usage', 'Fitness', 'Income', 'Miles']
# Univariate Boxplots
plt.figure(figsize=(18, 12))
for i, col in enumerate(continuous_columns, 1):
    plt.subplot(3, 2, i)
    sns.boxplot(y=df[col])
    plt.title(f'Boxplot of {col}')
plt.tight_layout()
plt.show()
```



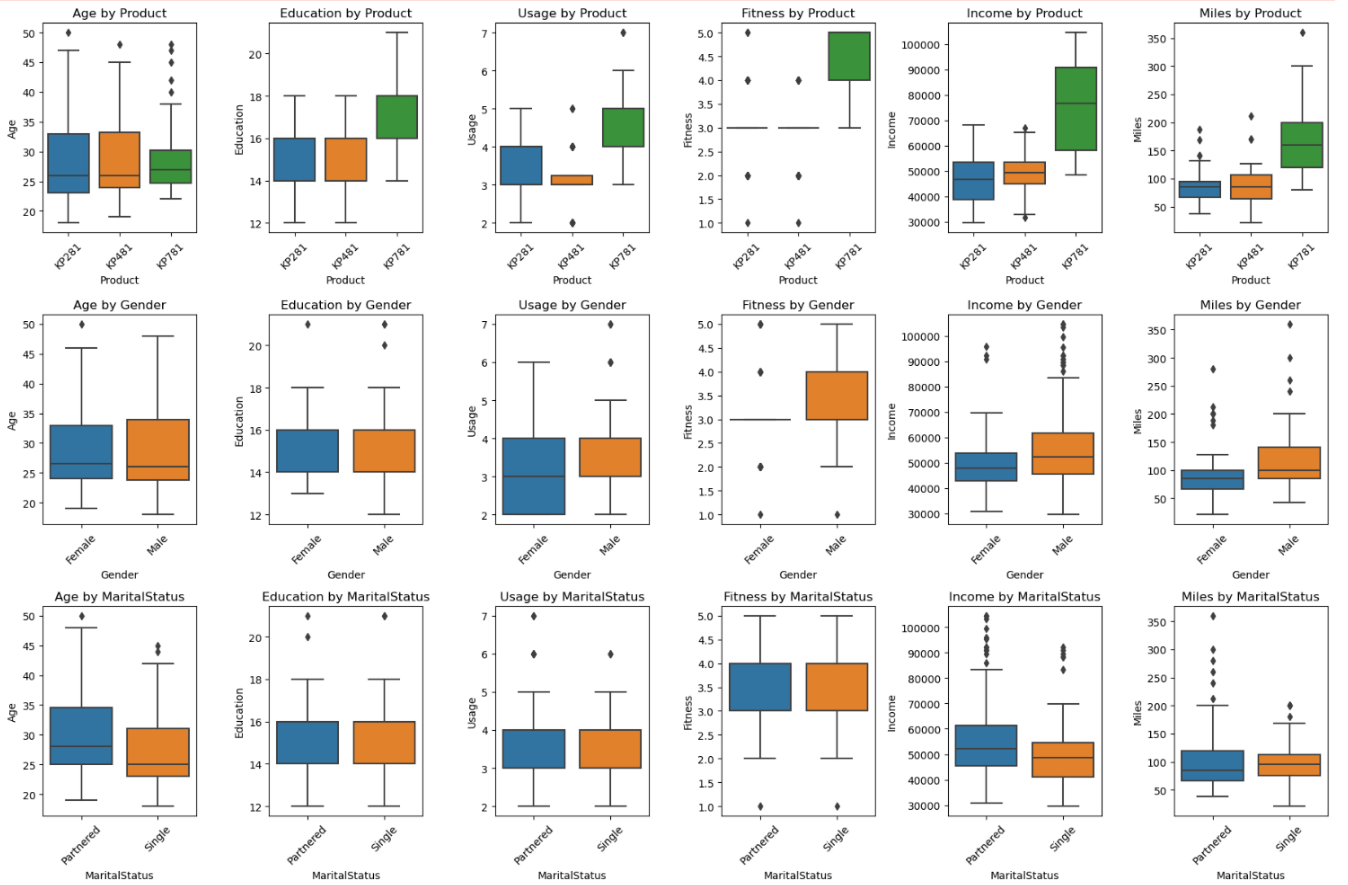
Bivariate Analysis:

Here is how we can do Bivariate Analysis on categorical variables,

Code is as:



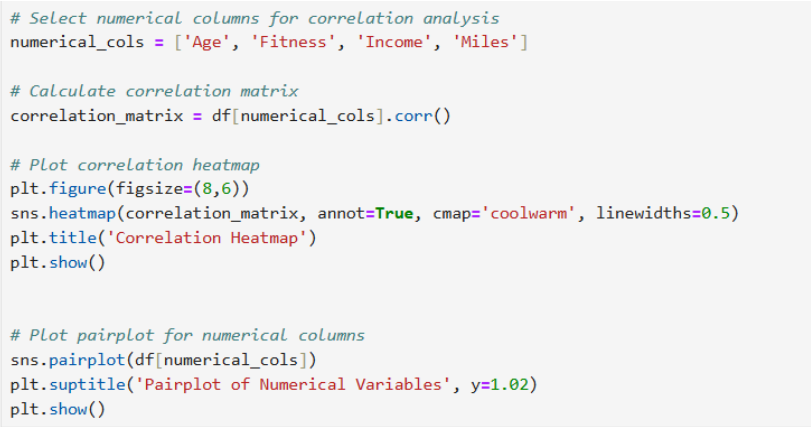
```
# Bivariate Boxplots with categorical variables
categorical_columns = ['Product', 'Gender', 'MaritalStatus']

plt.figure(figsize=(18, 12))
for i, col in enumerate(categorical_columns):
    for j, cont_col in enumerate(continuous_columns):
        plt.subplot(len(categorical_columns), len(continuous_columns), i * len(continuous_columns) + j + 1)
        sns.boxplot(x=df[col], y=df[cont_col])
        plt.title(f'{cont_col} by {col}')
        plt.xticks(rotation=45)
plt.tight_layout()
plt.show()
```



3.3) For correlation: Heatmaps, Pairplots(10 Points):

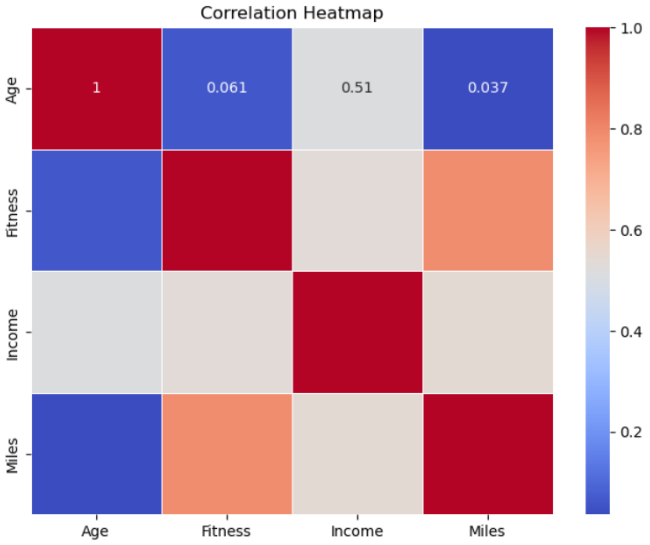
Univariate analysis typically involves examining the distribution and summary statistics of individual variables, while correlation and pairplots are methods used for bivariate analysis to explore relationships between variables

Code is as:

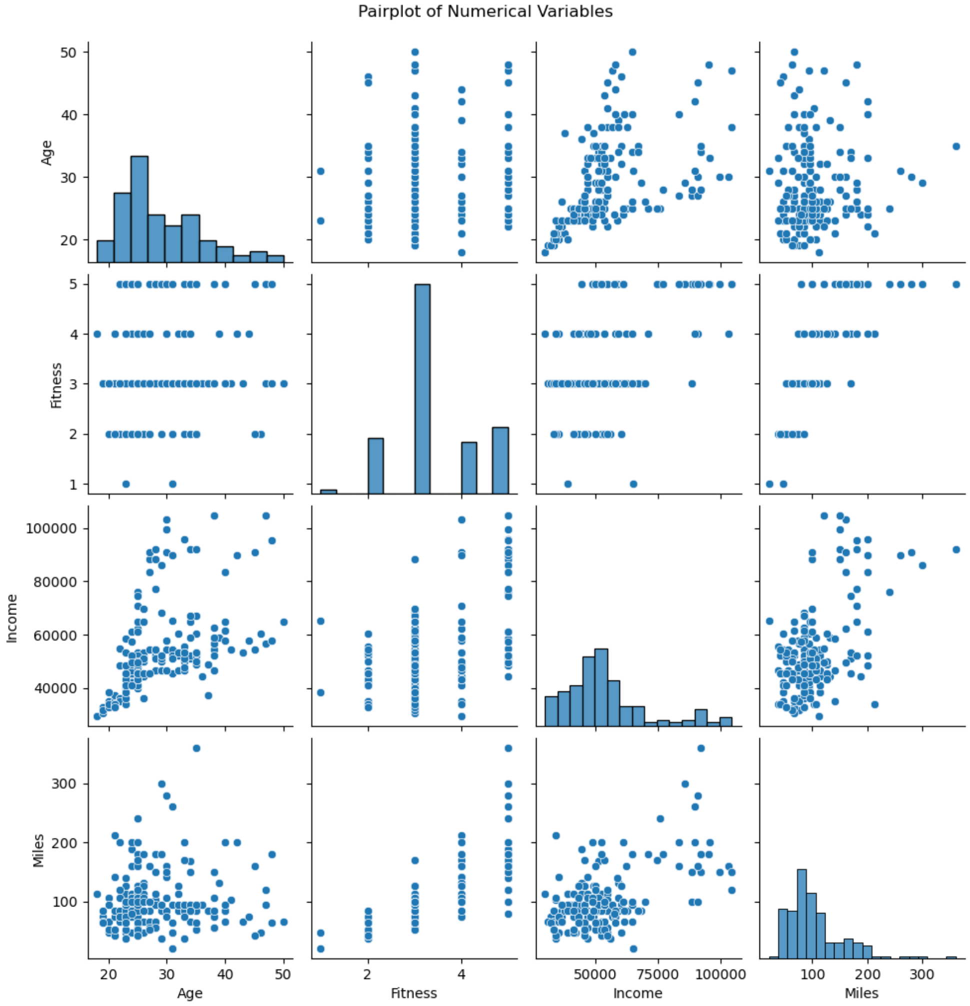


```
# Select numerical columns for correlation analysis
numerical_cols = ['Age', 'Fitness', 'Income', 'Miles']

# Calculate correlation matrix
correlation_matrix = df[numerical_cols].corr()

# Plot correlation heatmap
plt.figure(figsize=(8,6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', linewidths=0.5)
plt.title('Correlation Heatmap')
plt.show()


# Plot pairplot for numerical columns
sns.pairplot(df[numerical_cols])
plt.suptitle('Pairplot of Numerical Variables', y=1.02)
plt.show()
```

Heatmap ->



Pairplot->

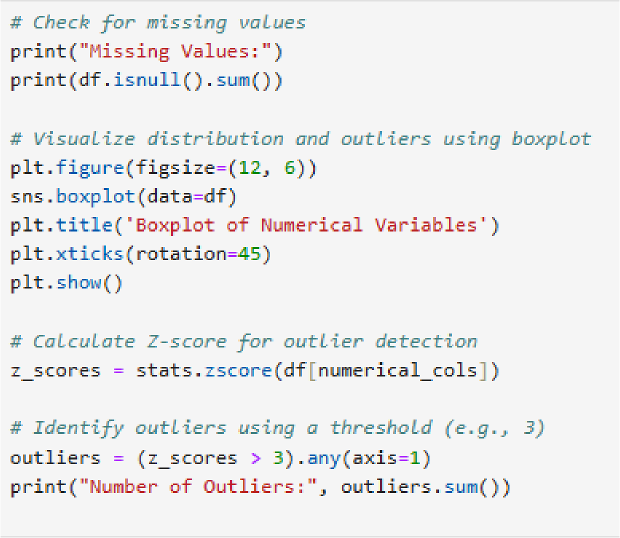


Q4. Missing Value & Outlier Detection:

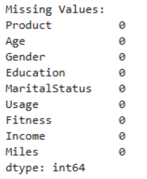
Detecting missing values and outliers is crucial for data preprocessing to ensure the quality and reliability of your analysis.

Missing value detection and outliers can be checked as:

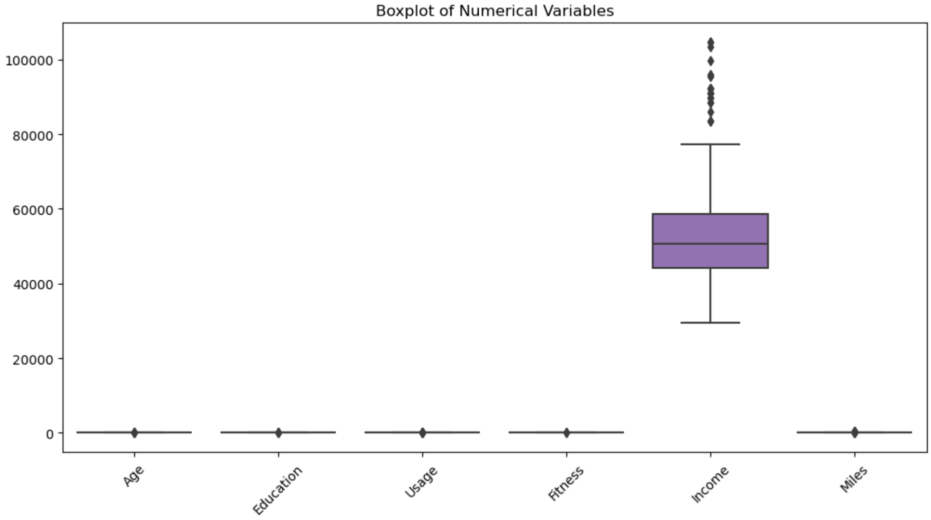
Code is as:



```
# Check for missing values
print("Missing Values:")
print(df.isnull().sum())

# Visualize distribution and outliers using boxplot
plt.figure(figsize=(12, 6))
sns.boxplot(data=df)
plt.title('Boxplot of Numerical Variables')
plt.xticks(rotation=45)
plt.show()

# Calculate Z-score for outlier detection
z_scores = stats.zscore(df[numerical_cols])

# Identify outliers using a threshold (e.g., 3)
outliers = (z_scores > 3).any(axis=1)
print("Number of Outliers:", outliers.sum())
```



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



```
Number of Outliers: 8
```

There is no missing values in nay columns and Total number of outliers are 8.

Q 5. Business Insights based on Non-Graphical and Visual Analysis **(10 Points)**

1. Q5.1 Comments on the range of attributes.

Business Insights from Non-Graphical and Visual Analysis

Non-Graphical Analysis

Descriptive Statistics:

1. **Age:**
   The average age in the dataset provides insights into the target demographic. For example, if the mean age is 30, the product may appeal more to young adults.
2. **Income:**
   High variability in income might suggest the product is bought by both high and low-income groups, indicating a broad market appeal.
3. **Usage and Fitness Levels:**

Average and median usage levels, as well as fitness scores, can indicate how engaged customers are with the product and their general fitness.

4. **Missing Values:**

If certain columns have a high number of missing values, it could indicate areas where data collection needs improvement. For instance, if 'Income' has many missing entries, it could suggest that customers are hesitant to disclose their financial information.

Visual Analysis

Boxplots:

Miles by Product: Significant differences in the median miles across different products can indicate which product models are used more frequently.

Income by Product: If higher-income individuals prefer a specific product, it can guide premium pricing strategies.

Pairplots:

Visualization of pairwise relationships can help in understanding the interactions between different numerical variables. For example, a pairplot showing Income vs. Miles might reveal that higher-income individuals tend to use the product more extensively.

Correlation Heatmap:

A heatmap can visually confirm the strength and direction of relationships between variables, highlighting key factors that influence customer behaviour. For example, a strong positive correlation between Usage and Miles would suggest that more frequent usage leads to higher mileage.

Bar Plots:

1. **Product by Gender:** If one gender predominantly purchases a certain product model, it could guide targeted marketing campaigns.
2. **Product by Marital Status:** Identifying preferences based on marital status can help in customizing marketing messages, such as promoting family-friendly features to partnered individuals.

Histograms:

Histograms of Age, Income, Usage, etc., help in understanding the distribution and central tendencies. For example, a right-skewed income distribution might indicate that a smaller segment of affluent customers significantly influences average income figures.

Scatterplots:

1. **Age vs. Miles:** Scatterplots can reveal trends such as younger users being more active.

2. **Income vs. Fitness:** A scatterplot might show whether higher-income individuals have higher fitness levels, suggesting that fitness features could be marketed as premium benefits.

Q5.2 Comments on the distribution of the variables and relationship between them

Distribution of Variables:

1. **Age**:
   - The age distribution appears to be relatively evenly spread, with no significant skewness observed. This suggests a diverse user base across different age groups.
   - Most users seem to fall within the range of 18 to 48 years, indicating that the product appeals to a wide demographic.
2. **Income**:

- The income distribution exhibits some skewness, with a majority of users having incomes in the mid-range.
- There are outliers with higher incomes, indicating the presence of a segment of more affluent users.

3. **Usage and Fitness**:
    - Usage and fitness levels appear to vary widely across the dataset, with no clear pattern observed from a visual inspection.
    - There may be clusters of users with similar usage and fitness levels, but further analysis is needed to identify any distinct patterns or segments.
4. **Miles**:
    - The distribution of miles appears to be right-skewed, with a majority of users logging lower mileage.
    - There are outliers with significantly higher mileage, suggesting the presence of a segment of highly active users.

Relationships Between Variables:

1. **Age and Income**:
    - There doesn't seem to be a strong linear relationship between age and income based on visual inspection. However, further analysis, such as correlation coefficients, would provide more insights into their relationship.
2. **Age and Usage**:
    - There doesn't appear to be a clear relationship between age and usage levels based on visual inspection. However, it's possible that certain age groups may exhibit higher or lower usage patterns, which could be explored through statistical analysis.
3. **Income and Usage**:
    - There doesn't seem to be a strong linear relationship between income and usage levels. However, there may be certain income brackets that exhibit higher or lower usage, which could be explored further.
4. **Usage and Miles**:
    - There appears to be a positive relationship between usage and miles based on visual inspection. Users with higher usage levels tend to log more miles, suggesting that more active users use the product more frequently.
5. **Fitness and Miles**:
    - There doesn't appear to be a clear relationship between fitness levels and miles logged based on visual
    - inspection. However, it's possible that users with higher fitness levels may log more miles, which could be explored through further analysis.

Q5.3 Comments for each univariate and bivariate plot:

Univariate Plots:

1. **Age Distribution Histogram**:
    - Comments: The histogram of age distribution shows that the majority of users fall between the ages of 18 to 48, with a relatively even spread across the range. This indicates a diverse user base in terms of age.
2. **Income Distribution Histogram**:
    - Comments: The income distribution histogram displays a slightly right-skewed distribution, with most users having incomes in the mid-range. However, there are outliers with higher incomes, suggesting the presence of a segment of more affluent users.
3. **Usage Distribution Histogram**:
    - Comments: The usage distribution histogram indicates variability in usage levels among users, with no clear pattern observed. Further analysis is needed to identify any distinct usage patterns or segments.
4. **Fitness Distribution Histogram**:
    - Comments: The fitness distribution histogram shows variability in fitness levels among users, with no clear pattern observed. Further analysis is needed to understand if there are any relationships between fitness levels and other variables.
5. **Miles Distribution Histogram**:

- Comments: The miles distribution histogram exhibits a right-skewed distribution, with a majority of users logging lower mileage. However, there are outliers with significantly higher mileage, indicating the presence of a segment of highly active users.

Bivariate Plots:

1. **Age vs. Income Scatter Plot**:
   - Comments: The scatter plot of age vs. income does not show a clear relationship between the two variables. However, there may be certain income brackets that correspond to specific age groups, which could be explored further.
2. **Age vs. Usage Scatter Plot**:
   - Comments: The scatter plot of age vs. usage levels does not reveal a clear relationship between the two variables. Further analysis is needed to determine if there are any age-related trends in usage patterns.
3. **Usage vs. Miles Scatter Plot**:
   - Comments: The scatter plot of usage vs. miles logged displays a positive relationship between the two variables. Users with higher usage levels tend to log more miles, suggesting that more active users use the product more frequently.
4. **Fitness vs. Miles Scatter Plot**:
   - Comments: The scatter plot of fitness vs. miles logged does not show a clear relationship between the two variables. However, it's possible that users with higher fitness levels may log more miles, which could be explored further.

Q.6 Recommendations **(10 Points)** - Actionable items for business. No technical jargon. No complications. Simple action items that everyone can understand.

1. **Create Targeted Ads**: Tailor your advertisements to specific age groups within the 18-48 range to resonate better with different segments of your audience.
2. **Introduce Budget-Friendly Options**: Develop entry-level products or subscription plans to attract customers with lower incomes, ensuring affordability for all.
3. **Highlight Premium Features**: Showcase premium product features in marketing campaigns to attract users with higher incomes, emphasizing the value they bring.
4. **Reward Active Users**: Implement a loyalty program or offer discounts to users who consistently engage with your product, encouraging ongoing usage.
5. **Promote Family Deals**: Advertise special discounts or family packages to appeal to partnered individuals or families, making your product more accessible to a wider audience.
6. **Seek Customer Feedback**: Conduct surveys or online polls to gather insights