Treadmill Buyer Profile

Introduction

An established brand in the fitness equipment industry is **Aerofit**. Machines including treadmills, exercise cycles, gym equipment and fitness accessories are the products that are offered by Aerofit to meet the needs of different demographics.

The market research team at AeroFit aims to understand customer demographics and their preferences for treadmill models: **KP281**, **KP481**, and **KP781**.

This analysis investigates how various factors such as gender, age, income, fitness level, and marital status influence treadmill purchases.

Data Description

The dataset has the following features:

- **Product Purchased** - KP281, KP481, or KP781
- **Age** - Customer's age in years
- **Gender** - Male/Female
- **Education** - In years
- **Marital Status** - Single/Partnered
- **Usage** - The average number of times the treadmill is used per week by customers
- **Fitness** - Self-rated fitness on a 1-5 scale
- **Income** - Annual income in US dollars
- **Miles** - The average number of miles the customer expects to walk/run each week

Data Exploration

Dataset Preview

Code

```
df.head()
```

The first 5 rows of this dataset are displayed as follows:

Result

	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

Code

```
df.tail()
```

The last 5 rows are displayed as follows:

Result

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
175	KP781	40	Male	21	Single	6	5	83416	200
176	KP781	42	Male	18	Single	5	4	89641	200
177	KP781	45	Male	16	Single	5	5	90886	160
178	KP781	47	Male	18	Partnered	4	5	104581	120
179	KP781	48	Male	18	Partnered	4	5	95508	180

Shape of Dataframe

Code

```
print("Rows =", df.shape[0])
print("Columns =", df.shape[1])
```

- This dataset has **180** rows and **9** columns.

Datatypes of Columns

Code

```
df.dtypes
```

Result

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

- **3** Categorical columns: Product, Gender and Marital Status
- **6** Numerical columns: Age, Education, Usage, Fitness, Income and Miles

Missing and Duplicated Values

Code

```
# for missing values
df.isnull().sum()

# for duplicate values
df[df.duplicated()].count()
```

- There are no missing values in the dataset.
- There are **0** duplicate rows in the dataset.

Statistical Summary

Numerical Features Summary

Code

```
df.describe(include= 'number')
```

Result

	Age	Education	Usage	Fitness	Income	Miles
count	180.000000	180.000000	180.000000	180.000000	180.000000	180.000000
mean	28.788889	15.572222	3.455556	3.311111	53719.577778	103.194444
std	6.943498	1.617055	1.084797	0.958869	16506.684226	51.863605
min	18.000000	12.000000	2.000000	1.000000	29562.000000	21.000000
25%	24.000000	14.000000	3.000000	3.000000	44058.750000	66.000000
50%	26.000000	16.000000	3.000000	3.000000	50596.500000	94.000000
75%	33.000000	16.000000	4.000000	4.000000	58668.000000	114.750000
max	50.000000	21.000000	7.000000	5.000000	104581.000000	360.000000

The summary of numerical features is as follows: **1. Age**

- Individuals are aged from **18-50** with the average of **28.7** years.

2. Education

- Education level ranges from **12-21** years of experience with the average experience of **16**.

3. Usage

- Individuals use treadmill **2-7** time/week.

4. Fitness

- On average, customers have rated their fitness at **3**, on a scale of 1-5.

5. Income

- Annual income ranges from **29K-104K**.

6. Miles

- Customer expects to walk or run **21-360** miles/week, with an average of **94** miles/week.

Categorical Features Summary

Code

```
df.describe(include= ['object'])
```

Result

	Product	Gender	MaritalStatus
count	180	180	180
unique	3	2	2
top	KP281	Male	Partnered
freq	80	104	107

The summary of categorical features is as follows: **Categorical Features Summary 1. Product**

- **3** unique values: KP281, KP481, KP781
- **KP281** is the top buying product, approximatley **44%** of the total sales.

2. Gender

- **2** unique values: Male or Female
- **Male** individuals bought the most treadmills, around **58%**, and **Females** bought **42%** of the treadmills.

3. Marital Status

- **2** unique values: Single or Partnered

- **Partnered** individuals bought the treadmills with the most frequency of approximately **60%**, and **Single** individuals bought **40%** of the treadmills.

Value Counts for all categorical features

Code

```
products = df['Product'].value_counts()
print(products)

print("----------------------------")

gender = df['Gender'].value_counts()
print(gender)

print("----------------------------")

marital_status = df['MaritalStatus'].value_counts()
print(marital_status)
```

Result

```
Product
KP281    80
KP481    60
KP781    40
Name: count, dtype: int64
----------------------------
Gender
Male      104
Female     76
Name: count, dtype: int64
----------------------------
MaritalStatus
Partnered    107
Single        73
Name: count, dtype: int64
```

Insights

1. Product Distribution

- Purchasing quantity for each treadmill model shows that **KP281** is the most popular and purchased treadmill.
- This suggests that it could be the most affordable treadmill among buyers.

2. Gender Distribution

- **Males** are dominant in purchasing the treadmills.
- The count of male and female customers helps determining the gender preference for purchasing a treadmill.

3. Marital Status Distribution

- **Couples/Partnered** individuals purchased the most treadmills.
- If married individuals are purchasing treadmills, it suggests higher preference among families or couples and if single individuals are purchasing, this could indicate that they are using it for personal fitness.

Unique Attributes for all categorical features

Code

```
categorical_cols = df.select_dtypes(include= 'object').columns

for col in categorical_cols:
    total = df[col].nunique()
    unique = df[col].unique()

    print(col)
    print("Total Unique Values =", total)
    print("Unique Values =", unique)
    print("--------------------")
```

Result

```
Product
Total Unique Values = 3
Unique Values = ['KP281' 'KP481' 'KP781']
--------------------
Gender
Total Unique Values = 2
Unique Values = ['Male' 'Female']
--------------------
MaritalStatus
Total Unique Values = 2
Unique Values = ['Single' 'Partnered']
```

Insights

1. Product

- It has **3** unique values.
- [KP281, KP481, KP781]

2. Gender

- It has **2** unique values.
- [Male, Female]

3. MaritalStatus

- It has **2** unique values.
- [Single, Partnered]

Univariate Analysis - Numerical features

Distribution Plot

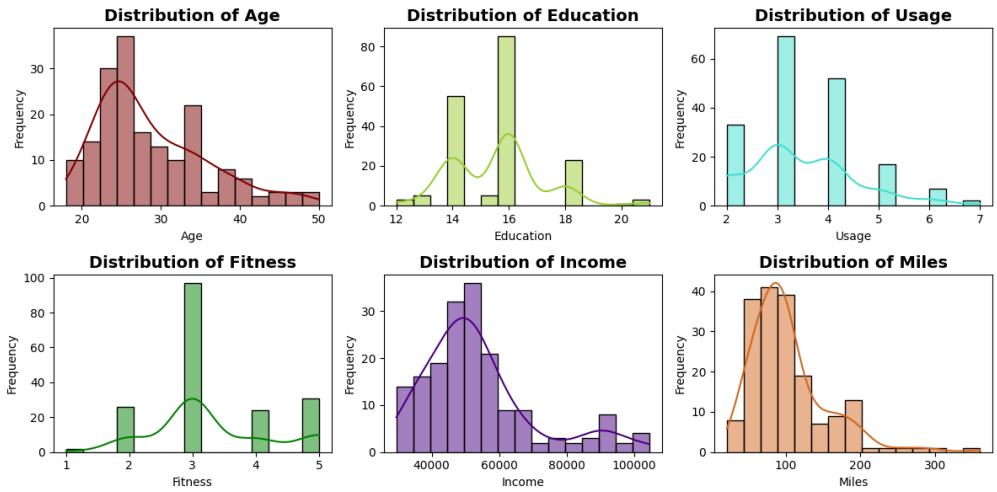
Code

```
plt.figure(figsize= (10, 6))
col_name = df.select_dtypes(include= 'number')

colors = ['maroon', 'yellowgreen', 'turquoise', 'green', 'indigo', 'chocolate']
for row, col in enumerate(col_name):
    plt.subplot(2, 3, row + 1)
    sns.histplot(df[col], kde= True, bins= 15, color= colors[row])
    plt.title(f'Distribution of {col}', fontdict= {'fontsize' : 14, 'fontweight' :
'bold'})
    plt.ylabel('Frequency')
    plt.tight_layout()

plt.show()
```

Result



Insights

1. Age Distribution

- Young adults appear to be the primary treadmill buyers, as the most of the clients are between the age of **20 and 30**.

2. Education Distribution

- The majority of customers have **16** years of education (i.e., Bachelor's degree).
- The two additional peaks on the second are **14** years for college level and **18** years for master's education.

3. Usage Distribution

- Most of the customers use treadmills **3-4** times a week.
- There are a very few customers who use treadmills **6-7** times per week.

4. Fitness Distribution

- Most customers have **3** (average) fitness level.
- Few customers have the lowest fitness level of **1**.

5. Income Distribution

- Majority of the buyers have the annual income between **$40,000** and **$60,000**.
- High income customers above **$80,000** exists, but fewer in number.

6. Miles Distribution

- The majority of the customers run **50-100 miles** per week.
- Some customers run over **200 miles**, but they are a minority.

Count Plot

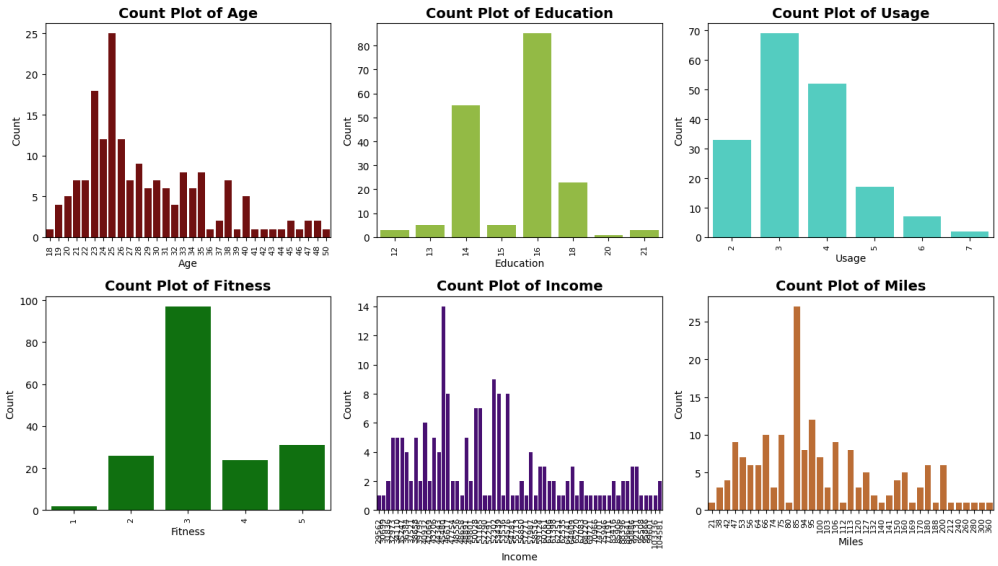
Code

```
plt.figure(figsize= (12, 5))
col_name = df.select_dtypes(include= 'number').columns
colors = ['maroon', 'yellowgreen', 'turquoise', 'green', 'indigo', 'chocolate']

for row, col in enumerate(col_name):
    plt.subplot(2, 3, row + 1)
    sns.countplot(x= df[col], color= colors[row])
    plt.title(f'Count Plot of {col}', fontdict= {'fontsize' : 14, 'fontweight' :
'bold'})
    plt.xticks(rotation= 90, fontsize= 8)
    plt.ylabel('Count')
    plt.tight_layout()

plt.show()
```

Result



Box Plot

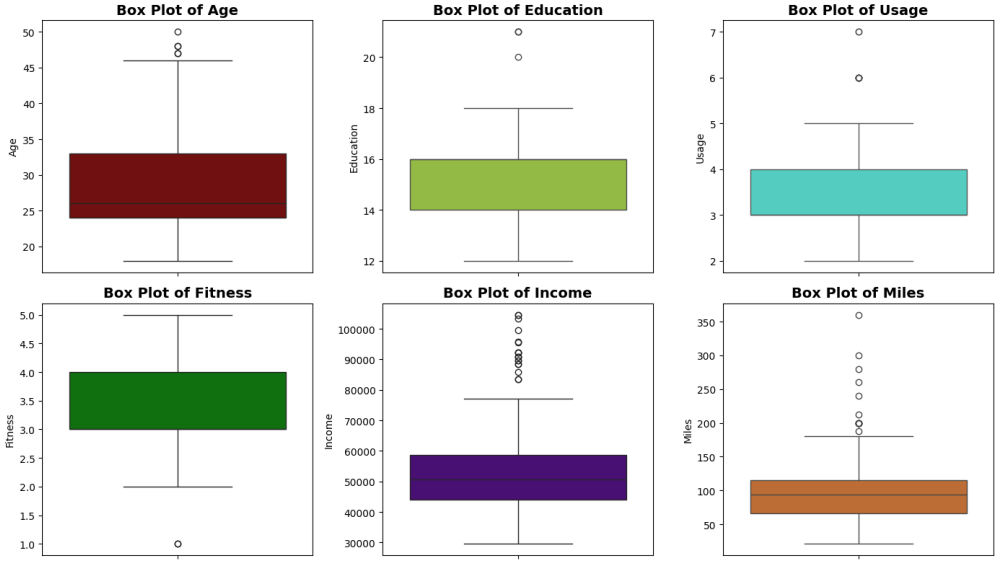
Code

```
plt.figure(figsize= (10, 5))
col_name = df.select_dtypes(include= 'number').columns
colors = ['maroon', 'yellowgreen', 'turquoise', 'green', 'indigo', 'chocolate']

for row, col in enumerate(col_name):
    plt.subplot(2, 3, row + 1)
    sns.boxplot(y= df[col], color= colors[row])
    plt.title(f'Box Plot of {col}', fontdict= {'fontsize' : 14, 'fontweight' :
'bold'})
    plt.tight_layout()

plt.show()
```

Result



Insights

1. Age

- The median age is **26** years.
- 50% of the data falls in the IQR range, from Q1(**24** years) and Q3(**33** years).
- Most of the treadmills are bought by **young adults**.
- There are few outliers above age of **46** years. It indicates that the old buyers are very rare.

2. Education

- The median is **16** years.
- The IQR ranges from **14-16** years, which shows that most of the customers have college and university level education.
- The outliers are above **19** years, which indicates that there are a very customers who have high education level.

3. Usage

- The median is about **3** times/week on average.
- The IQR ranges from **3-4** times/week.
- The outliers exceeds **5** times/week, which shows that the customers are very few.
- Only 1 buyer uses the treadmill **7** times/week, which is quite rare.

4. Fitness

- The median is **3**, which indicates moderate fitness.
- The IQR ranges shows the customer's fitness between **3-4** on a scale of 1-5.
- The outliers are below **1**, which indicates poor fitness.

5. Income

- The median is approximately **$50,000**.
- The IQR ranges between **~($44,000 and $59,000)**.
- There are several outliers which are above **$80,000** and **$100,000**. They are high-income buyers.

6. Miles

- The median is about **90-100** miles/week.
- The IQR ranges from **~(60-120)** miles, which shows that most of the buyers use the treadmill moderately at this rate.

Univariate Analysis - Categorical features

Count Plot

Code

```
fig, axes = plt.subplots(1, 3, figsize= (12, 5))
```

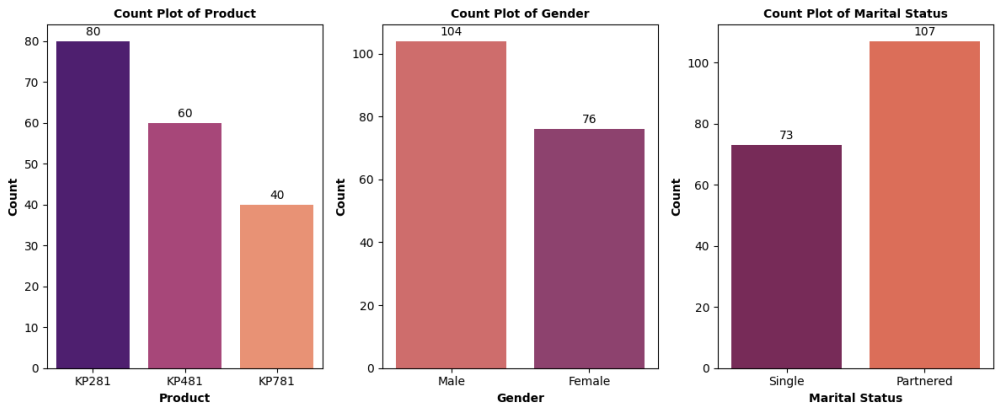
```
ax = sns.countplot(data= df, x= 'Product', ax= axes[0], palette= 'magma')
for container in ax.containers:
    ax.bar_label(container, fontsize=10, padding=3)
axes[0].set_title('Count Plot of Product', fontdict= {'fontsize': 10,
'fontweight': 'bold'})
axes[0].set_xlabel('Product', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
axes[0].set_ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})

ax = sns.countplot(data= df, x= 'Gender', ax= axes[1], palette= 'flare')
for container in ax.containers:
    ax.bar_label(container, fontsize=10, padding=3)
axes[1].set_title('Count Plot of Gender', fontdict= {'fontsize': 10, 'fontweight':
'bold'})
axes[1].set_xlabel('Gender', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
axes[1].set_ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})

ax = sns.countplot(data= df, x= 'MaritalStatus', ax= axes[2], palette= 'rocket')
for container in ax.containers:
    ax.bar_label(container, fontsize=10, padding=3)
axes[2].set_title('Count Plot of Marital Status', fontdict= {'fontsize': 10,
'fontweight': 'bold'})
axes[2].set_xlabel('Marital Status', fontdict= {'fontsize': 10, 'fontweight':
'bold'})
axes[2].set_ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})

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

Result



Insights

1. Product Distribution

- Treadmill model **KP281** is the best-selling treadmill, with the count of **80**.
- Model **KP481** has a count of approximately, **60** purchases.
- Model **KP781** has the least number of buyers with the count of **40** purchases.

2. Gender Distribution

- Males (about **58%**) are the dominant buyers than females (around **42%**)

3. Marital Status Distribution

- Around **60%** of the customers, who bought more treadmills, are **couples** than singles.

Bivariate Analysis

1. Product vs Gender

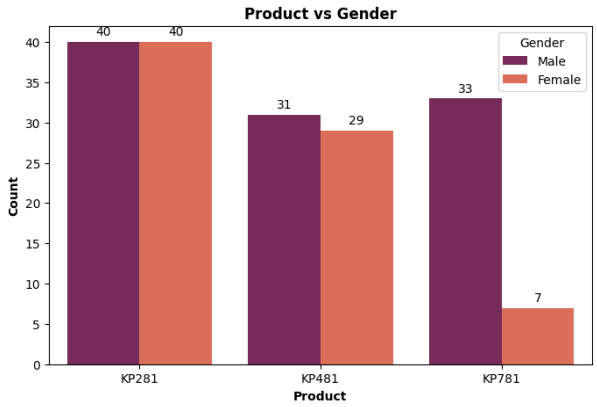
Code

```
fig = plt.figure(figsize= (8, 5))

ax = sns.countplot(data= df, x= 'Product', hue= 'Gender', palette= 'rocket')
for container in ax.containers:
    ax.bar_label(container, fontsize=10, padding=3)

plt.xlabel('Product', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.title('Product vs Gender', fontdict= {'fontsize': 12, 'fontweight': 'bold'})
plt.show()
```

Result



Insights

This shows that **male** individuals purchase more treadmill than females. **1. KP281**

- Both males and females prefer **KP281** equally, with **40** each.
- It may be budget-friendly treadmill.

2. KP481

- Males **(~31)** slightly buy more **KP481** than females **(~29)**.

3. KP781

- As compared to females **(~7)**, around **33** males purchased **KP781**.

- It could be larger, that makes it less attractive to female buyers.

2. Product vs Marital Status

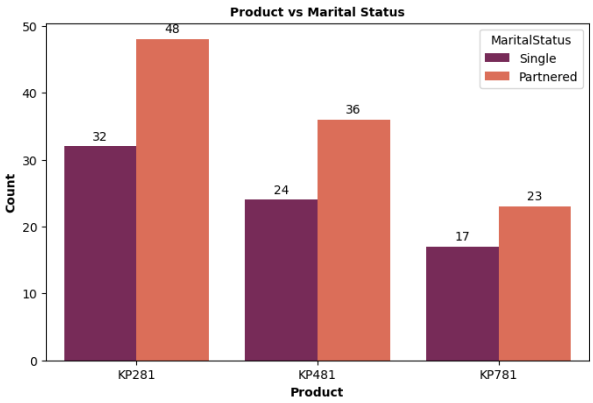
Code

```
fig = plt.figure(figsize= (8, 5))

ax= sns.countplot(data= df, x= 'Product', hue= 'MaritalStatus', palette= 'rocket')
for container in ax.containers:
    ax.bar_label(container, fontsize=10, padding=3)

plt.xlabel('Product', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.title('Product vs Marital Status', fontdict= {'fontsize': 10, 'fontweight':
'bold'})
plt.show()
```

Result



Insights

This shows that Partners buy more treadmills than singles. **1. KP281**

- Married individuals **(~48)** purchase more KP281 than singles **(~32)**.
- It could be due to budget-friendly, or it might be designed for home.

2. KP481

- Partnered individuals **(~36)** buy more treadmills of model KP481 than singles **(~24)**.
- It could be an advanced treadmill for fitness focused couples.

3. KP781

- Partnered individuals **(~23)** buy more KP781 than singles **(~17)**.
- It could be a premium model.
- It could be a high-performance treadmill, that is used by atheletes.

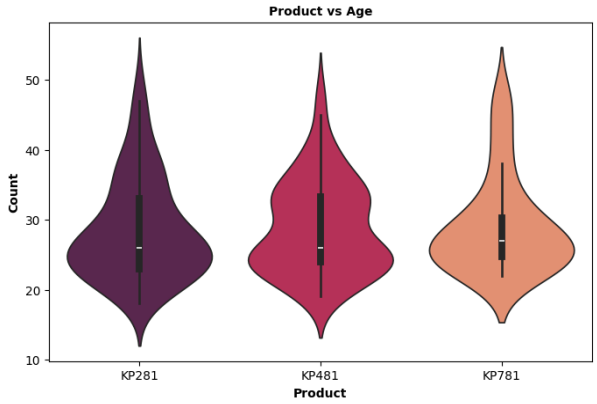
3. Product vs Age

Code

```
fig = plt.figure(figsize= (8, 5))

sns.violinplot(data= df, x= 'Product', y= 'Age', hue= 'Product', palette=
'rocket')
plt.xlabel('Product', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.ylabel('Count', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.title('Product vs Age', fontdict= {'fontsize': 10, 'fontweight': 'bold'})
plt.show()
```

Result



Insights

1. KP281

- The thicker portion of the violin shows that most of the buyers are between **20** and **30**.
- The median age for **KP281** buyers is around **26** years.
- Customers who have age above 40 years also buy KP281, but fewer in number.
- It is popular among younger customers, which suggests that it could be budget-friendly.

2. KP481

- The median age for **KP481** buyers is similar to KP281, around **26** years.
- The number of buyers are more in their 30s and even early 40s
- It could be attracted by professionals.

3. KP781

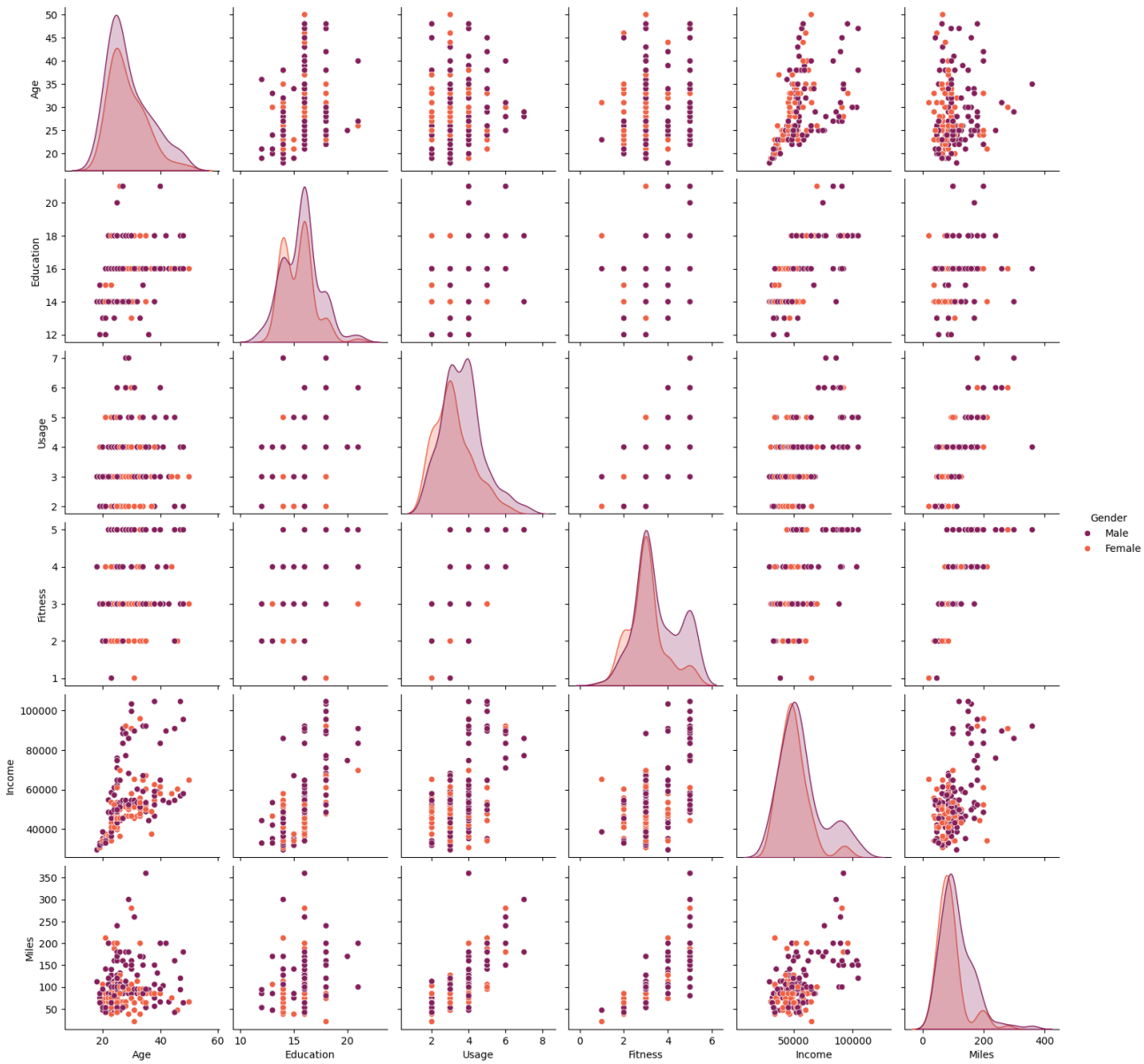
- The median age for **KP781** buyers is around **27** years.
- Compared to KP281 and KP481, there is a distinct presence of senior buyers who are above 40 years.
- There are few younger customers below 25 years who prefer this treadmill.
- This treadmill is preffered by olders.

Multi-Variate Analysis

Code

```
sns.pairplot(data= df, hue= 'Product', palette= 'magma')
```

Result



Insights

1. Age-Income

- Income increases with increase in age.
- Customers below 30 years have low income, and tend to buy KP281.
- Older customers above 40 years tend to have higher income and more likely to buy premium treadmills (KP781).

2. Fitness-Miles

- Customers with higher fitness levels tends to run more miles.

3. Education-Income

- Higher education levels correlate with higher income.
- KP781 is mostly purchased by individuals having higher education lavel.

4. Usage-Fitness Level

- Customers with higher fitness levels 4 and 5 uses treadmills more frequently.
- KP781 buyers are mostly at fitness level 5.

5. Usage-Miles

- Those who use the treadmill more often tend to log more miles.

5. Correlation Analysis

Correlation Matrix

Code

```
corr_matrix = df.select_dtypes(include= ['number']).corr()
print("Correlation Matrix =")
corr_matrix
```

Result

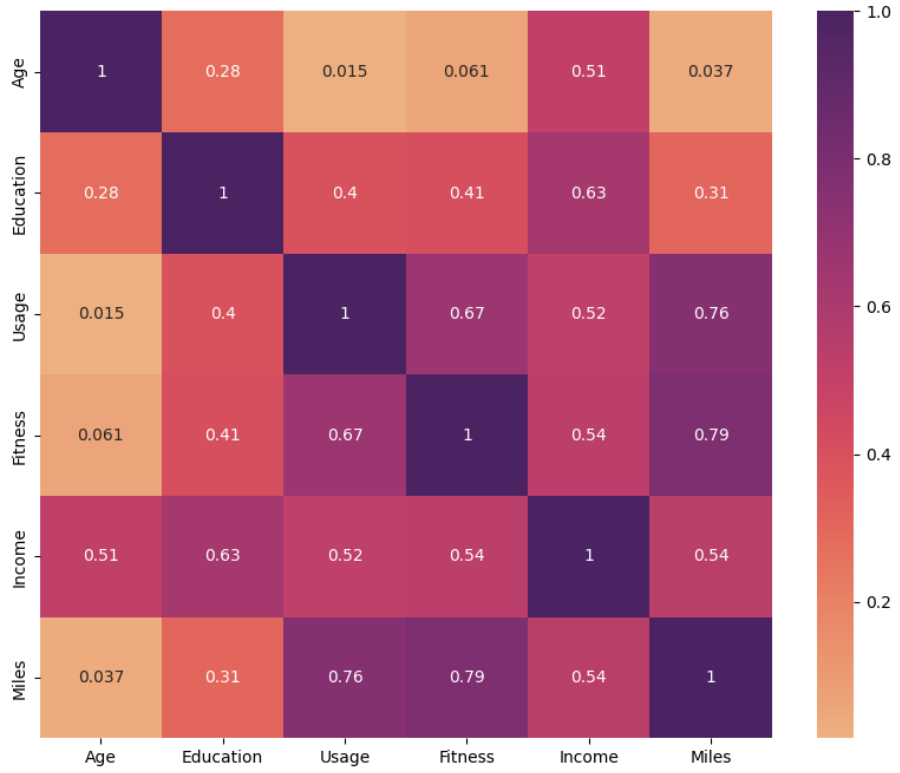
	Age	Education	Usage	Fitness	Income	Miles
Age	1.000000	0.280496	0.015064	0.061105	0.513414	0.036618
Education	0.280496	1.000000	0.395155	0.410581	0.625827	0.307284
Usage	0.015064	0.395155	1.000000	0.668606	0.519537	0.759130
Fitness	0.061105	0.410581	0.668606	1.000000	0.535005	0.785702
Income	0.513414	0.625827	0.519537	0.535005	1.000000	0.543473
Miles	0.036618	0.307284	0.759130	0.785702	0.543473	1.000000

Heatmap

Code

```
plt.figure(figsize= (10, 8))
sns.heatmap(data= corr_matrix, annot= True, cmap= 'flare')
```

Result



Insights

Strong Positive Correlation

1. Age-Income (0.51)

- The correlation between Age and Income is moderate.
- Those customers who are older may have moved into higher-paying roles, which indicates higher income.

2. Education-Income (0.63)

- There is a strong but moderate correlation between Education and Income.
- Those who have higher education level tend to have higher income.

3. Usage-Miles (0.76)

- There is a strong correlation between Usage and Miles.
- This indicates that those who use the treadmill more often tend to log more miles.

4. Fitness-Miles (0.79)

- There is a very strong correlation between Fitness level and Miles run.
- It indicates that the customers who keep up a high level of fitness are more likely to use the treadmill for longer workouts.

Weak Correlation

1. Age-Usage (0.015)

- Almost no correlation between Age and Usage.

- This suggests that there is no difference in treadmill usage between any age group.

2. Miles-Age (0.037)

- There is a very weak correlation between Miles and Age.
- This shows that the number of miles runs by a person is not affected by their age.

3. Fitness-Age (0.061)

- The correlatio between Fitness and Age is 0.061, which is close to 0.
- This suggests that age has no effect on fitness. People of different age groups may have similar fitness levels.

4. Education-Age (0.28)

- There is a weak positive correlation between Education and Age.
- This shows that older customers tend to have slightly higher levels of education.

Outliers Detection

Code

```
column_name = df.select_dtypes(include= 'number').columns

for col in column_name:
    q1 = df[col].quantile(0.25)
    q3 = df[col].quantile(0.75)
    iqr = q3 - q1
    lower = q1 - 1.5 * iqr
    upper = q3 + 1.5 * iqr

    outliers = df[(df[col] < lower) | (df[col] > upper)][col]

    print("\nColumn :", col)
    print(f"Total Outliers : {len(outliers)}")
    print("Outlier Values :", outliers.values)
```

Result

```
Column : Age
Total Outliers : 5
Outlier Values : [47 50 48 47 48]

Column : Education
Total Outliers : 4
Outlier Values : [20 21 21 21]

Column : Usage
Total Outliers : 9
Outlier Values : [6 6 6 7 6 7 6 6 6]

Column : Fitness
Total Outliers : 2
Outlier Values : [1 1]

Column : Income
Total Outliers : 19
Outlier Values : [ 83416  88396  90886  92131  88396  85906  90886 103336  99601  89641
  95866  92131  92131 104581  83416  89641  90886 104581  95508]

Column : Miles
Total Outliers : 13
Outlier Values : [188 212 200 200 200 240 300 280 260 200 360 200 200]
```

Percent of customers who have purchased KP281, KP481 or KP781

Code

```
product_purchase = df['Product'].value_counts(normalize= True) * 100
print("Percentage of customers who purchased treadmill model", product_purchase)
```

Result

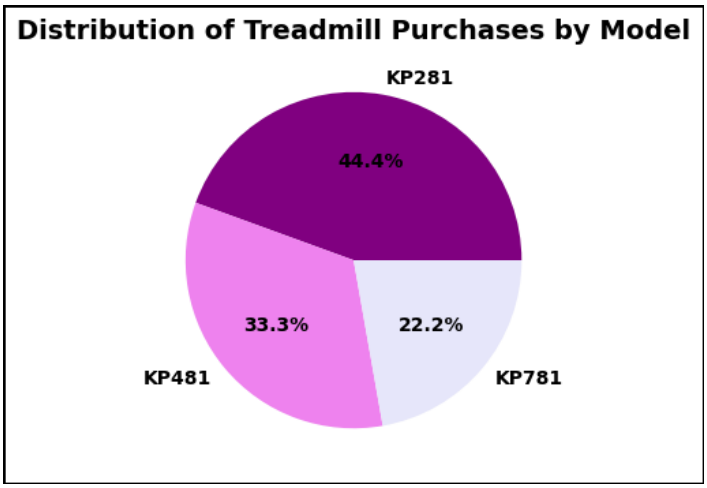
```
Percentage of customers who purchased treadmill model Product
KP281    44.444444
KP481    33.333333
KP781    22.222222
Name: proportion, dtype: float64
```

Code

```
plt.figure(figsize= (4, 4))
plt.pie(product_purchase, labels= product_purchase.index, autopct= '%1.1f%%',
        textprops= {'fontsize':10, 'fontweight':'bold'},
        colors= ['purple', 'violet', 'lavender'])

plt.title('Distribution of Treadmill Purchases by Model', fontdict= {'fontsize' :
14, 'fontweight' : 'bold'})
plt.show()
```

Result



Insights

1. KP281

- As it is the most popular model, around **44.4%** of the customers buy this treadmill.

2. KP481

- Around **33.3%** of the customers purchased, as it is the second most popular model.

3. KP781

- It is the least sold treadmill, with around **22.2%** customers.

Frequency Tables

1. Product – Gender

Code

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

Result

Gender	Female	Male
Product		
KP281	40	40
KP481	29	31
KP781	7	33

Insights

- Percentage of a Male customer purchasing a treadmill = **57.78%**

- Percentage of a Female customer purchasing KP781 = **9.21%**
- Probability of a customer being a Female given that Product is KP281 = **0.50**

2. Product – Age

Code

```
frequency_table_age = pd.crosstab(df['Product'], pd.cut(df['Age'], bins= [0, 19, 30, 40, 50, 100], labels= ["<20", "21-30", "31-40", "41-50", "50+"]))
frequency_table_age
```

Result

Age Product	<20	21-30	31-40	41-50
KP281	4	51	19	6
KP481	1	34	23	2
KP781	0	30	6	4

Insights

- Percentage of customers with Age between 20s and 30s among all customers = **90.56%**

3. Product – Income

Code

```
frequency_table_income = pd.crosstab(df['Product'], pd.cut(df['Income'], bins= 4, labels= ['Low', 'Medium', 'High', 'Very High']))
frequency_table_income
```

Result

Income Product	Low	Medium	High	Very High
KP281	46	32	2	0
KP481	27	32	1	0
KP781	0	16	7	17

Insights

- Percentage of a low-income customer purchasing a treadmill = **46.11%**
- Percentage of a high-income customer purchasing KP781 treadmill = **16.11%**
- Percentage of customer with high-income salary buying treadmill given that Product is KP781 = **72.50%**

4. Product – Fitness

Code

```
frequency_table_fitness = pd.crosstab(df['Product'], df['Fitness'])
frequency_table_fitness
```

Result

Fitness	1	2	3	4	5
Product					
KP281	1	14	54	9	2
KP481	1	12	39	8	0
KP781	0	0	4	7	29

Insights

- Percentage of customers that have fitness level 5 = **17.22%**
- Percentage of a customer with Fitness Level 5 purchasing KP781 treadmill = **72.50%**

5. Product - Marital Status

Code

```
frequency_table_maritalstatus = pd.crosstab(df['Product'], df['MaritalStatus'])
frequency_table_maritalstatus
```

Result

MaritalStatus	Partnered	Single
Product		
KP281	48	32
KP481	36	24
KP781	23	17

Insights

- Percentage of a customers who are partnered using treadmills = **59.44%**

Future Recommendations

KP281

- Promotions for the product **KP281** must remain gender-neutral.

- As this model is purchased more by youngers, so consider student discounts.

KP481

- Since, females interest is also strong with a slight difference of purchasing.
- If this model has features that are more attractive to males, they should emphasize those in their marketing.
- Consider offering bundle deals for couples.
- Focus on performance tracking and custom workouts, as it is attarcted by professionals.

KP781

- The model **KP781** is underperforming.
- It might be more advanced due to strength training or sports training. Consider discounts or feature upgrades for women.
- Advertise to health concious individuals, who are above 40 years.
- Investigate customer perception of KP781. If it is costly, consider for discounts.