In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
df=pd.read_csv("https://d2beiqkhq929f0.cloudfront.net/public_assets/assets/000/001/
df.head()
```

Out[2]:

	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

What Does ‘Good’ Look Like

Checking the Structure & Characteristics of the Dataset

In [3]:
```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 180 entries, 0 to 179
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype 
---  ------         --------------  ----- 
 0   Product        180 non-null    object
 1   Age            180 non-null    int64 
 2   Gender         180 non-null    object
 3   Education      180 non-null    int64 
 4   MaritalStatus  180 non-null    object
 5   Usage          180 non-null    int64 
 6   Fitness        180 non-null    int64 
 7   Income         180 non-null    int64 
 8   Miles          180 non-null    int64 
dtypes: int64(6), object(3)
memory usage: 12.8+ KB
```

In [4]:
```
df.shape
```

Out[4]:
```
(180, 9)
```

In [5]:
```
df.isnull().sum()
```

Out[5]:

	0
Product	0
Age	0
Gender	0
Education	0
MaritalStatus	0
Usage	0
Fitness	0
Income	0
Miles	0

dtype: int64

In [6]:
```
df.describe()
```

Out[6]:

	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

In [7]:
```
df.dtypes
```

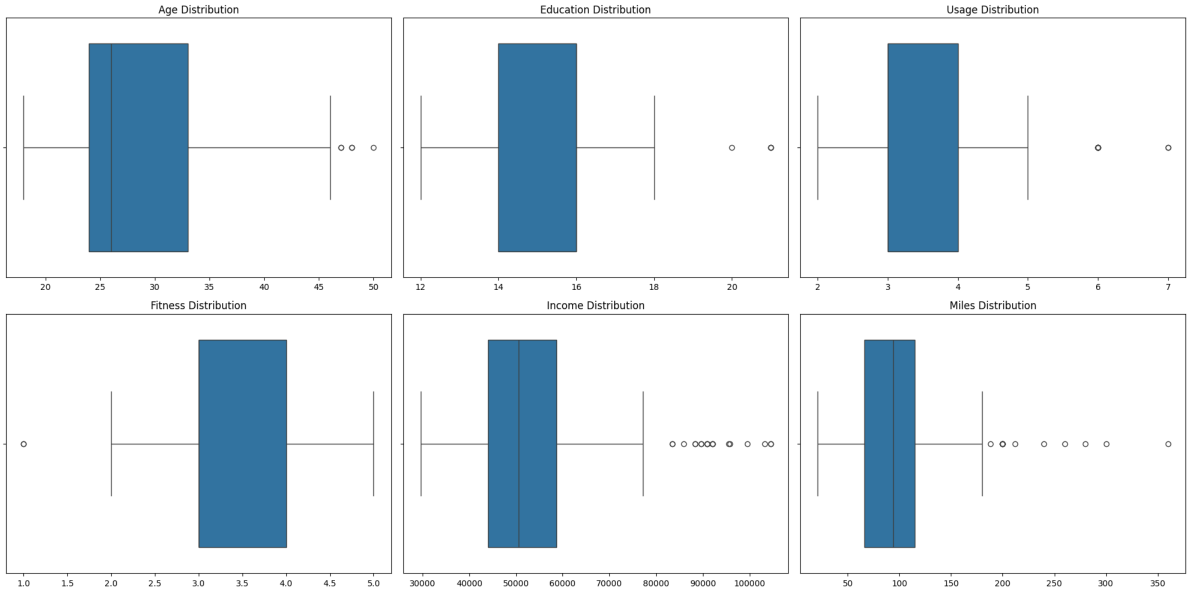
Out[7]:

	0
Product	object
Age	int64
Gender	object
Education	int64
MaritalStatus	object
Usage	int64
Fitness	int64
Income	int64
Miles	int64

dtype: object

Detect Outliers

In [8]:
```
lst=["Age","Education","Usage","Fitness","Income","Miles"]
plt.figure(figsize=(20,10))
for i in range(len(lst)):
    plt.subplot(2,3,i+1)
    sns.boxplot(x=df[lst[i]])
    plt.title(f"{lst[i]} Distribution")
    plt.xlabel("")
plt.tight_layout()
plt.show()
```



In [9]:
```
df1=df.copy()
df1.drop(columns=["Gender","Product","MaritalStatus"],inplace=True)
```

Clipping the Data

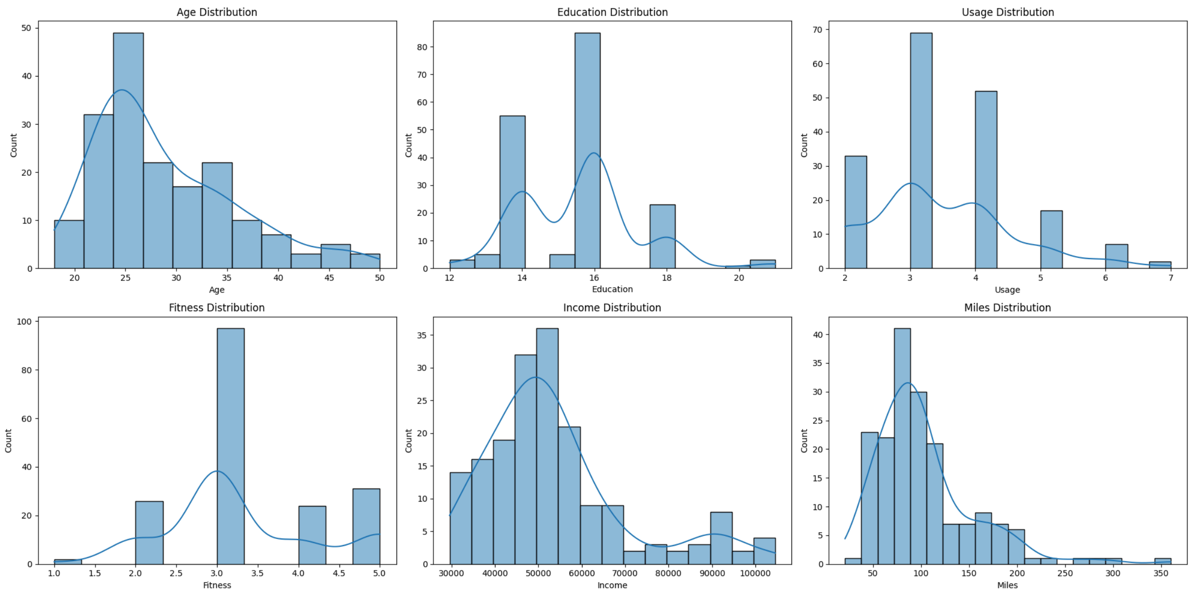
In [10]:
```
lower_percentiles = df1.quantile(0.05)
upper_percentiles = df1.quantile(0.95)
df_clipped = df1.apply(lambda x: np.clip(x, lower_percentiles[x.name], upper_percer
print(df_clipped)
```

```
       Age  Education  Usage  Fitness    Income  Miles
0    20.00         14   3.00        4  34053.15    112
1    20.00         15   2.00        3  34053.15     75
2    20.00         14   4.00        3  34053.15     66
3    20.00         14   3.00        3  34053.15     85
4    20.00         14   4.00        2  35247.00     47
..     ...        ...    ...      ...       ...    ...
175  40.00         18   5.05        5  83416.00    200
176  42.00         18   5.00        4  89641.00    200
177  43.05         16   5.00        5  90886.00    160
178  43.05         18   4.00        5  90948.25    120
179  43.05         18   4.00        5  90948.25    180

[180 rows x 6 columns]
```

In [11]:
```
lst=["Age","Education","Usage","Fitness","Income","Miles"]
plt.figure(figsize=(20,10))
for i in range(len(lst)):
    plt.subplot(2,3,i+1)
```

```
    sns.histplot(x=lst[i],kde=True,data=df)
    plt.title(f"{lst[i]} Distribution")
    plt.xlabel(f"{lst[i]}")
    plt.ylabel("Count")
plt.tight_layout()
plt.show()
```

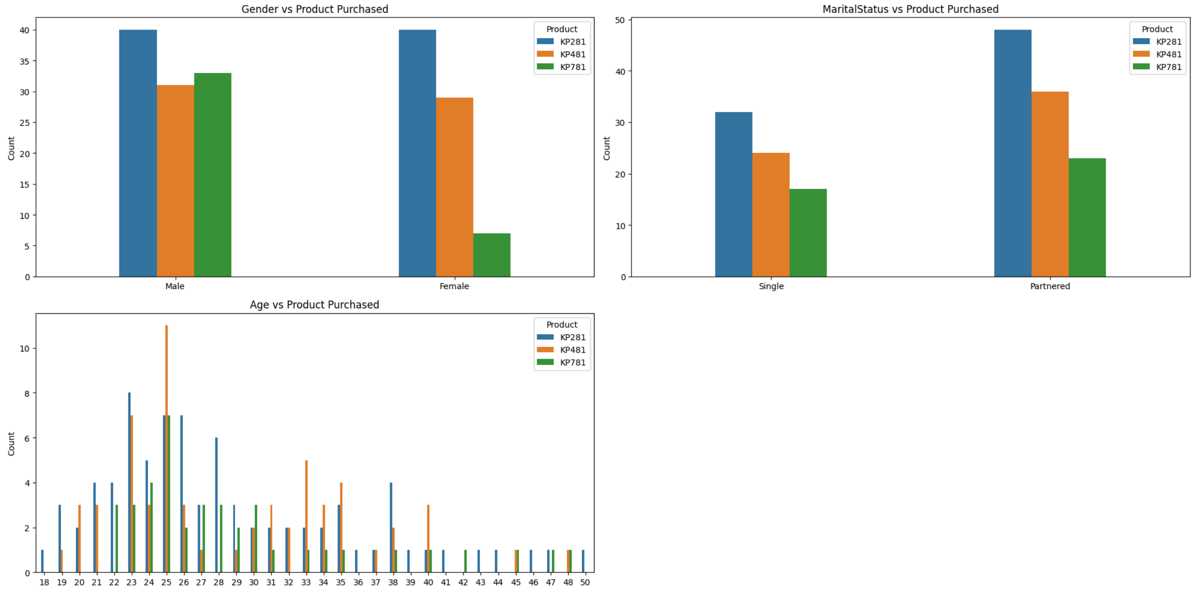


Insights -

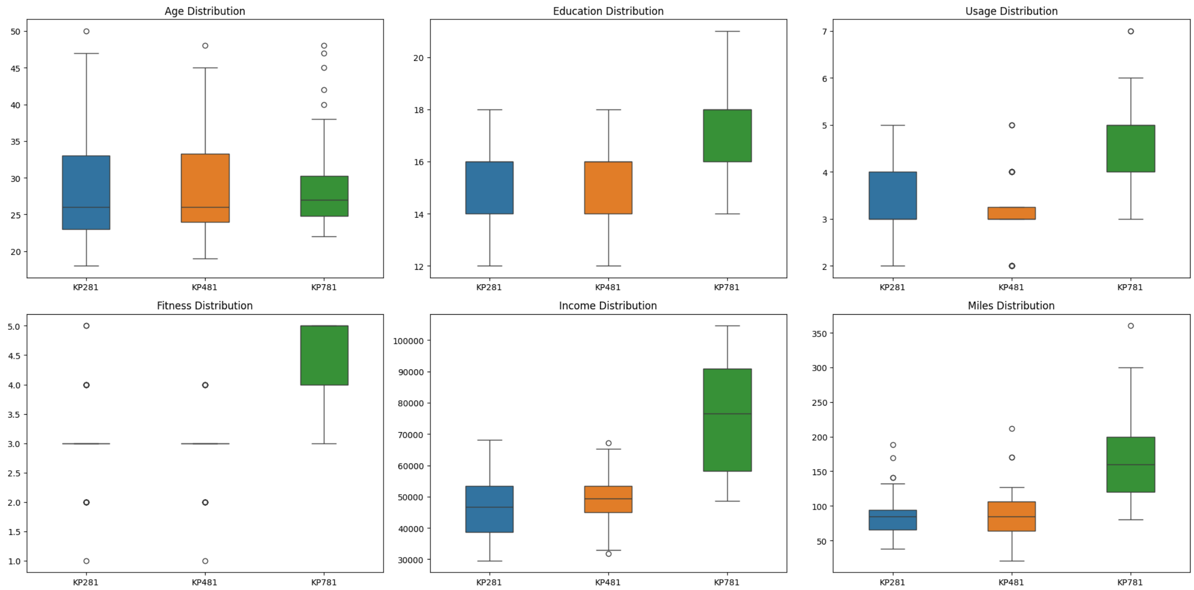
- **Age**: The majority of customers are young adults (20-30 years).
- **Education**: Customers typically have 14 to 16 years of education.
- **Usage**: Most customers use the product 2 to 4 times.
- **Fitness**: A fitness level of 3 is common among customers.
- **Income**: The majority of customers have moderate incomes ($30,000-$70,000).
- **Miles**: Most customers travel relatively short distances (50-150 miles).

Checking Features (MaritalStatus, Gender, Age) effect on the product purchase

In [12]:
```
lst=["Gender","MaritalStatus","Age"]
plt.figure(figsize=(20,10))
for i in range(len(lst)):
    plt.subplot(2,2,i+1)
    sns.countplot(x=lst[i],hue="Product",data=df,width=0.4)
    plt.title(f"{lst[i]} vs Product Purchased")
    plt.xlabel("")
    plt.ylabel("Count")
plt.tight_layout()
plt.show()
```



In [13]:
```
lst=["Age","Education","Usage","Fitness","Income","Miles"]
plt.figure(figsize=(20,10))
for i in range(len(lst)):
    plt.subplot(2,3,i+1)
    sns.boxplot(y=df[lst[i]],x=df["Product"],hue=df["Product"],width=0.4)
    plt.title(f"{lst[i]} Distribution")
    plt.xlabel("")
    plt.ylabel("")
plt.tight_layout()
plt.show()
```



Marginal Probability

In [14]:
```
(round(df["Product"].value_counts(normalize=True),4)*100).apply(lambda x: '{:.2f}%'
```

Out[14]:

	proportion
Product	
KP281	44.44%
KP481	33.33%
KP781	22.22%

dtype: object

Insights -

- **KP281** is the most used product, with **44.44%** of customers choosing it
- **KP481** is the second most used product, with **33.33%** of customers preferring it.
- **KP781** is the least used product, with **22.22%** of customers opting for it.

Probability that the customer buys a product based on Gender,MaritalStatus and Age

In [15]:
```
crosstab=round(pd.crosstab(index=df["Product"],columns=df["Gender"],normalize=True)
crosstab_percentage = crosstab.applymap(lambda x: '{:.2%}'.format(x))
crosstab_percentage
```

Out[15]:

Gender	Female	Male
Product		
KP281	22.22%	22.22%
KP481	16.11%	17.22%
KP781	3.89%	18.33%

Insights -

- The gender distribution for product preferences reveals that KP281 is equally popular among both females and males, each holding 22.22%.
- For KP481, males show a slightly higher preference (17.22%) compared to females (16.11%).
- KP281 is universally popular, KP781 is particularly favored by males.

In [16]:
```
crosstab=round(pd.crosstab(index=df["Product"],columns=df["MaritalStatus"],normaliz
crosstab_percentage = crosstab.applymap(lambda x: '{:.2%}'.format(x))
crosstab_percentage
```

Out[16]:

MaritalStatus	Partnered	Single
Product		
KP281	26.67%	17.78%
KP481	20.00%	13.33%
KP781	12.78%	9.44%

Insights -

- Partnered individuals prefer all three products more than single individuals.
- KP281 is the most popular among both groups, with 26.67% of partnered individuals and 17.78% of single individuals choosing it.
- KP481 follows, with 20.00% of partnered and 13.33% of single individuals.
- KP781 is less popular overall but still preferred more by partnered individuals (12.78% compared to 9.44% of singles).

In [17]:
```
labels = labels = ['18-30', '31-40', '41-50']
df1=pd.qcut(df['Age'], q=3, labels=labels)
crosstab=round(pd.crosstab(index=df["Product"],columns=df1,normalize=True),4)
crosstab_percentage = crosstab.applymap(lambda x: '{:.2%}'.format(x))
crosstab_percentage
```

Out[17]:

Age	18-30	31-40	41-50
Product			
KP281	18.89%	11.67%	13.89%
KP481	15.56%	3.89%	13.89%
KP781	9.44%	7.22%	5.56%

Insights -

- **KP281** is most popular among the 18-30 age group (18.89%) but also maintains a consistent preference across the 31-40 (11.67%) and 41-50 (13.89%) age groups.
- **KP481** has a strong preference in the 18-30 (15.56%) and 41-50 (13.89%) age groups, but significantly drops in popularity among the 31-40 age group (3.89%).
- **KP781** is less popular overall, with its highest preference in the 18-30 age group (9.44%), followed by a decline in the 31-40 (7.22%) and 41-50 (5.56%) age groups.

Conditional Probability that the customer buys a product given that Gender is Male or Female

In [18]:
```
lst = ["KP281", "KP481", "KP781"]
crosstab = pd.crosstab(df['Gender'], df['Product'], normalize='index')
for gender in ['Male','Female']:
    print(f"\nProbabilities for {gender} customers:")
    print("")
```

```
    for product in lst:
        prob = crosstab.loc[gender, product]
        bold_prob = f"\033[1m{prob:.2%}\033[0m"
        print(f"Probability that a {gender.lower()} customer purchases {product}: {
```

```
Probabilities for Male customers:

Probability that a male customer purchases KP281: 38.46%
Probability that a male customer purchases KP481: 29.81%
Probability that a male customer purchases KP781: 31.73%

Probabilities for Female customers:

Probability that a female customer purchases KP281: 52.63%
Probability that a female customer purchases KP481: 38.16%
Probability that a female customer purchases KP781: 9.21%
```

Conditional Probability that the customer buys a product given that MaritalStatus is Single or Partnered

In [19]:

```
lst = ["KP281", "KP481", "KP781"]
crosstab = pd.crosstab(df['MaritalStatus'], df['Product'], normalize='index')
for i in ['Single','Partnered']:
    print(f"\nProbabilities for {i} customers:")
    print("")
    for product in lst:
        prob = crosstab.loc[i, product]
        bold_prob = f"\033[1m{prob:.2%}\033[0m"
        print(f"Probability that a {i.lower()} customer purchases {product}: {bold_
```

```
Probabilities for Single customers:

Probability that a single customer purchases KP281: 43.84%
Probability that a single customer purchases KP481: 32.88%
Probability that a single customer purchases KP781: 23.29%

Probabilities for Partnered customers:

Probability that a partnered customer purchases KP281: 44.86%
Probability that a partnered customer purchases KP481: 33.64%
Probability that a partnered customer purchases KP781: 21.50%
```

Correlation Among Different Factors

In [20]:

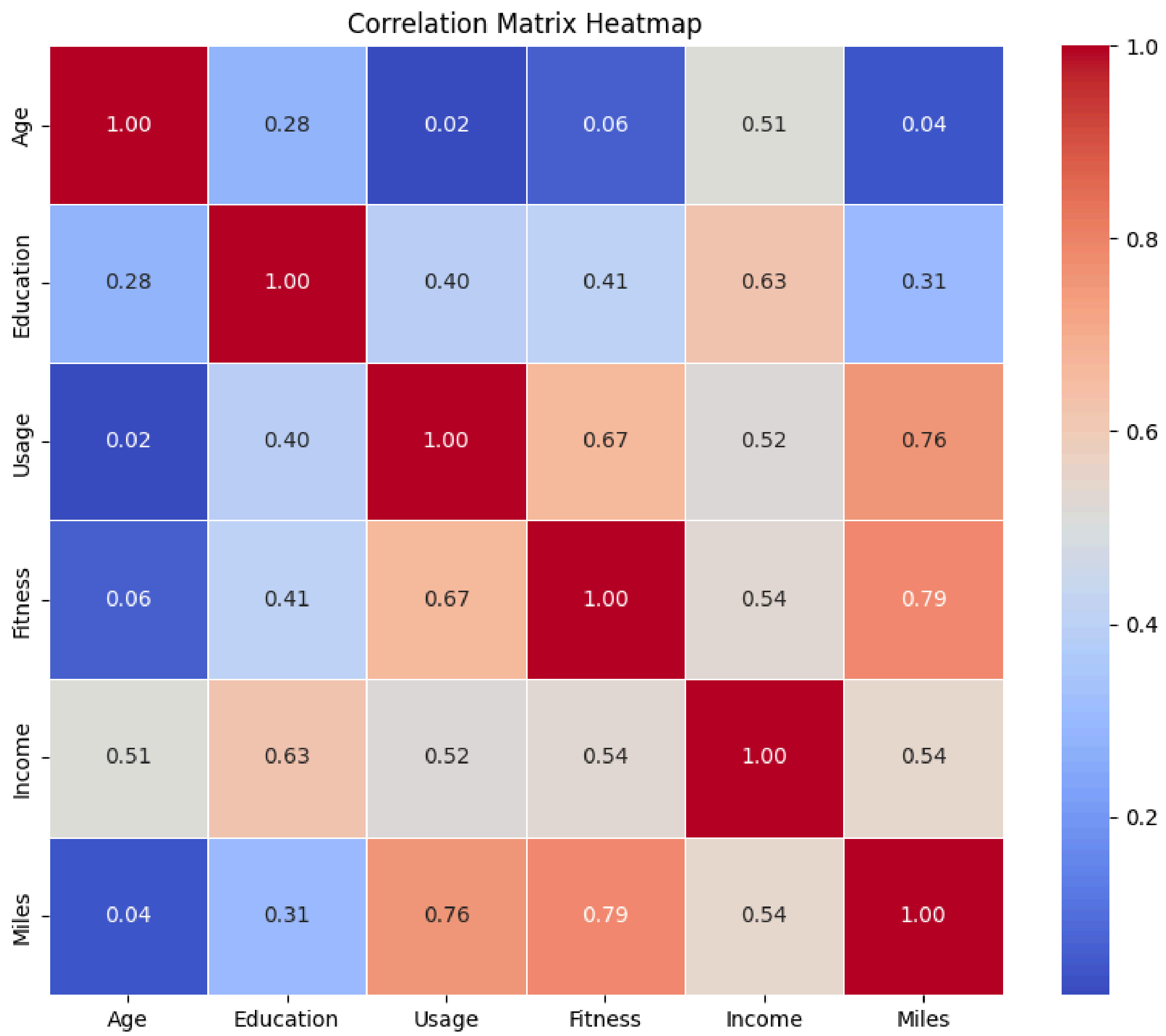
```
numeric_df = df.select_dtypes(include='number')
correl=numeric_df.corr()
correl
```

Out[20]:

	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

In [21]:
```
plt.figure(figsize=(10, 8))
sns.heatmap(correl, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)
plt.title('Correlation Matrix Heatmap')
plt.show()
```



Insights -

- **Strong Positive Correlations -**
- Miles and Fitness have a high correlation (0.79), indicating that as fitness levels increase, the distance traveled (miles) also tends to increase.

- Income and Education are also strongly correlated (0.63), suggesting that higher education levels are associated with higher income.
- **Usage and Miles -**
- Usage is highly correlated with Miles (0.76), implying that higher product usage often results in more miles traveled.

- **Fitness and Usage -**
- There is a notable correlation between Fitness and Usage (0.67), indicating that those who are more fit tend to use the product more frequently.

- **Moderate Correlations -**
- Income shows a moderate correlation with Usage (0.52) and Fitness (0.54), suggesting that higher income individuals tend to use the product more and have higher fitness levels.
- **Weak Correlations -**
- Age shows generally weak correlations with the other variables, indicating that age has a minimal direct impact on the other factors like education, usage, and fitness.

Customer Profiling

Product - KP281

- **Age -** Customers are typically aged around 28, with a range mostly between 25 to 35 years old.

- **Income -** Less than 50,000, generally between 30,000 and 50,000.

- **Fitness -** Fitness level is generally under 3.

- **Miles -** Customers typically travel less than 90 miles, with a range mostly between 50 to 150 miles.

- **Usage -** Product usage is around 3-4 times.
- **Education -** Less than 16 years of education, generally between 13 to 16 years.
- **Marital Status -** KP281 is the most popular among both groups.
- **Gender -** KP281 is equally popular among both females and males.

Product - KP481

- **Age -** Customers are typically aged around 25, with a range mostly between 22 to 30 years old.

- **Income -** Between 50,000 and 70,000.

- **Fitness -** Fitness level is generally around 3.

- **Miles -** Customers typically travel around 60 miles, with a range mostly between 50 to 100 miles.

- **Usage -** Product usage is around 3 times.
- **Education -** Around 16 years of education, generally between 14 to 17 years.
- **Marital Status -** KP481 is the most popular among both groups
- **Gender -** For KP481, males show a slightly higher preference compared to females.

Product - KP781

- **Age -** Customers are typically aged around 32, with a range mostly between 28 to 40 years old.
- **Income -** Between 70,000 and 100,000.
- **Fitness -** Fitness level is above 5.
- **Miles -** Customers typically travel around 200 miles, with a range mostly between 150 to 350 miles
- **Usage -** Product usage is around 4-5 times.
- **Education -** More than 16 years of education, generally between 15 to 20 years.
- **Marital Status -** KP781 is less popular overall but still preferred more by partnered individuals.
- **Gender -** KP781 is particularly favored by males.

Recommendations -

- For **KP281**, marketing efforts should highlight aspects that appeal to female customers, such as usability, moderate fitness, and travel needs, while also targeting younger males.
- For **KP481**, emphasize the product’s versatility and balanced features that appeal to a broad demographic, focusing on professionals and balanced lifestyle attributes.
- For **KP781**, create marketing campaigns that focus on high-income, health-conscious males, and address the barriers that may be limiting its appeal to female customers.
- Highlighting unique features and benefits that cater to high fitness and travel requirements could attract a more diverse customer base.

In [ ]:

```

```