```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import warnings as ws
```

The dataset's comprehensive feature set aims to define target audience profiles for each InsuaMart product using tables and charts. Construct new contingency tables for each InsuaMart product to analyze differences in customer characteristics. Compute conditional and marginal probabilities to uncover insights that impact business recommendations.

```
df1=pd.read_csv('./insuaMart.csv')
df1.shape # (rows, columns)
df1.head()
```

```
(7000, 10)
```

	Product_Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4.00	29562
1	KP281	19	Male	15	Single	2	3.00	31836
2	KP281	19	Female	14	Partnered	4	3.00	30699
3	KP281	19	Male	12	Single	3	3.00	32973
4	KP281	20	Male	13	Partnered	4	2.00	35247

```
df1.info()# Number of rows (RP entries)\n\nNumber of columns (RP attributes)\n\n
```

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

<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
dtype: int64, object
memory usage: 12.8 KB
```

```
df1.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	103194.444444
std	NaN	6.943498	NaN	1.617055	NaN	1.084797	0.958869	16506.684226	51804.029655
min	NaN	18.000000	NaN	12.000000	NaN	2.000000	1.000000	29562.000000	21080.000000
25%	NaN	23.000000	NaN	14.000000	NaN	2.000000	3.000000	44058.750000	70966.750000
50%	NaN	26.000000	NaN	16.000000	NaN	3.000000	3.000000	50596.500000	95508.000000
75%	NaN	33.250000	NaN	16.000000	NaN	4.000000	4.000000	58668.000000	115562.500000
max	NaN	50.000000	NaN	21.000000	NaN	7.000000	5.000000	104581.000000	360741.000000

Based on the observations provided:

- **No Missing Values:** The dataset contains no missing values, ensuring completeness in the data.
- **Unique Products:** There are 3 unique products in the dataset, with "KP281" being the most frequently occurring.

1. **Age Distribution:**
   - Minimum age is 18, maximum age is 50.
   - Mean age is approximately 28.79.
   - 25% of individuals are aged 23 or younger, indicating a relatively young customer base.
2. **Education Levels:**
   - Most individuals (75%) have 16 years of education or less.
3. **Gender Distribution:**
   - Out of 180 entries, 104 individuals are male and the rest are female.
4. **Income and Miles:**
   - Standard deviation for income and miles is very high, suggesting the presence of outliers in these variables. This could skew any average calculations and indicate a wide range of income and exercise miles among customers.

These insights provide a snapshot of the demographic and behavioral characteristics of InsuaMart's customers, which can guide further analysis and business strategy. Addressing outliers in income and miles might be crucial for more accurate targeting and customer segmentation efforts.

```
df1["Product1 unique:\
",df1["Product1 unique:\
",df1["Product1 unique:\
",df1.nunique()
```

```
Product1 unique:
Product1 unique:
Product1 unique:
Product    3
Age       21
Gender     2
Education 13
MaritalStatus 2
Usage      6
Fitness    5
Income   62
Miles    37
dtype: int64
```

```
fig, axes = plt.subplots(nrows=3, ncols=3, figsize=(14, 10))
fig.suptitle('Distribution of Variables', fontsize=16)

sns.boxplot(data=df, y='Product', hue='Product', ax=axes[0,0])
sns.boxplot(data=df, y='Age', hue='Product', ax=axes[0,1])
sns.boxplot(data=df, y='Gender', hue='Product', ax=axes[0,2])
sns.boxplot(data=df, y='Education', hue='Product', ax=axes[1,0])
sns.boxplot(data=df, y='MaritalStatus', hue='Product', ax=axes[1,1])
sns.boxplot(data=df, y='Usage', hue='Product', ax=axes[1,2])
sns.boxplot(data=df, y='Fitness', hue='Product', ax=axes[2,0])
sns.boxplot(data=df, y='Income', hue='Product', ax=axes[2,1])
sns.boxplot(data=df, y='Miles', hue='Product', ax=axes[2,2])

plt.show()
```

Based on the observations:

1. **Outliers**
   - Age, Education, and Usage have relatively few outliers based on boxplots.
   - Income and Miles exhibit more outliers, indicating significant variability in these metrics among customers.
2. **Distribution of Qualitative Attributes**
   - **Product:** The dataset includes three products in the dataset, each presumably representing different types of insurance or related products.
   - **Gender:** The dataset includes two genders, likely male and female.
   - **MaritalStatus:** This attribute likely indicates categories such as married, single, Divorced, etc., though specific details on distribution aren't provided.

Understanding the distribution of these qualitative attributes helps in segmenting and profiling customers more effectively. Further analysis could explore how these attributes relate to each other and to customer behaviors and preferences, aiding in targeted marketing and product recommendations.

```
fig, axes = plt.subplots(nrows=1, ncols=3, figsize=(14, 5))
sns.countplot(data=df, x='Product', order=['KP281', 'KP281', 'KP281'], ax=axes[0])
sns.countplot(data=df, x='Gender', order=['Male', 'Female'], ax=axes[1])
sns.countplot(data=df, x='MaritalStatus', order=['Single', 'Partnered'], ax=axes[2])

axes[0].set_title('Product - counts')
axes[1].set_title('Gender - counts')
axes[2].set_title('MaritalStatus - counts')

plt.show()
```

Based on the observations:

1. **Product Frequency**
   - KP281 is the most frequently occurring product among the dataset.
2. **Gender Distribution**
   - There are more males than females in the dataset.
3. **Marital Status**
   - Those who are married (partnered or in a relationship) comprise a larger portion of the dataset compared to single individuals.

Normalized counts for each variable (excluding these are proportions or percentages within the dataset) would provide precise figures for these observations, helping to understand the relative distribution of each category within the data. These insights are foundational for understanding customer demographics and can guide targeted marketing strategies and product recommendations accordingly.

```
df1 = df[['Product', 'Gender', 'MaritalStatus']]
df1.groupby(['Gender', 'MaritalStatus'])['Product'].value_counts()
```

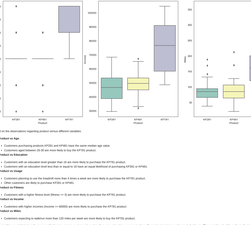
```
Gender  MaritalStatus  Product
Female  Partnered     KP281    34
        Single        KP281    34
Male    Partnered     KP281    73
        Single        KP281    39
        Single        KP281     2
Name: count, dtype: int64
```

Based on the observations provided:

1. **Product Distribution**
   - All 34 of customers have purchased the KP281 product.
   - 34 of customers have purchased the KP281 product.
   - 73 of customers have purchased the KP281 product.
2. **Gender Distribution**
   - 68 of customers are female.
   - 112 of customers are male.
3. **Marital Status Distribution**
   - 34 of customers are partnered (presumably married or in a relationship).
   - 73 of customers are partnered (presumably married or in a relationship).

These observations highlight the relationship between product purchases and demographic factors such as gender and marital status. They indicate potential preferences or behaviors among different customer segments, which could inform targeted marketing strategies and product recommendations.

```
df1 = df[['Product', 'Usage', 'Fitness', 'Income', 'Miles']]
df1.groupby('Product')
fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14, 10))
fig.suptitle('Product vs')
for i in range(5):
    ax = axes[i]
    for j, col in enumerate(['Usage', 'Fitness', 'Income', 'Miles']):
        sns.boxplot(data=df, x='Product', y=col, hue='Product', ax=ax[j], palette='bwr')
        ax[j].set_title('Product vs {}'.format(col), fontsize=14)
        ax[j].set_ylabel('')
        ax[j].set_xlabel('')
plt.show()
```



Based on the observations regarding product versus different variables:

1. **Product vs Age**
   - Customers purchasing products KP281 and KP281 have the same median age value.
   - Customers purchasing Product 3 are more likely to be younger than the KP281 product.
2. **Product vs Education**
   - Customers with an education level greater than 16 are more likely to purchase the KP281 product.
   - Customers with education level less than or equal to 16 have an equal likelihood of purchasing KP281 or KP281.
3. **Product vs Usage**
   - Customers planning to use the treadmill more than 4 times a week are more likely to purchase the KP281 product.
   - Other customers are likely to purchase KP281 or KP281.
4. **Product vs Fitness**
   - Customers with higher fitness levels (fitness >= 3) are more likely to purchase the KP281 product.
5. **Product vs Income**
   - Customers with higher incomes (income >= 60000) are more likely to purchase the KP281 product.
6. **Product vs Miles**
   - Customers expecting to walk/run more than 120 miles per week are more likely to buy the KP281 product.

These insights reveal distinct preferences and behaviors associated with different products based on their age, education, usage intentions, fitness level, income, and expected mileage. They provide valuable guidance for targeting specific customer segments with appropriate marketing strategies and product recommendations.

```
df1 = df[['Product', 'Usage', 'Fitness', 'Income', 'Miles']]
df1.groupby('Product')
fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14, 10))
fig.suptitle('Product vs')
for i in range(5):
    ax = axes[i]
    for j, col in enumerate(['Usage', 'Fitness', 'Income', 'Miles']):
        sns.boxplot(data=df, x='Gender', y=col, hue='Product', ax=ax[j], palette='bwr')
        ax[j].set_title('Product vs {}'.format(col), fontsize=14)
        ax[j].set_ylabel('')
        ax[j].set_xlabel('')
plt.show()
```

Female planning to use treadmill 3-4 times a week, are more likely to buy KP281 product.

```
df1.groupby('Product')
fig, axes = plt.subplots(nrows=1, ncols=5, figsize=(14, 10))
fig.suptitle('Product vs')
for i in range(5):
    ax = axes[i]
    for j, col in enumerate(['Usage', 'Fitness', 'Income', 'Miles']):
        sns.boxplot(data=df, x='Gender', y=col, hue='Product', ax=ax[j], palette='bwr')
        ax[j].set_title('Product vs {}'.format(col), fontsize=14)
        ax[j].set_ylabel('')
        ax[j].set_xlabel('')
plt.show()
```

```
Product
KP281  KP281
KP281  KP281
KP281  KP281
```