```python
1 import pandas as pd
2 import matplotlib as mlt
```

```python
1 df = pd.read_csv("customer_shopping_behavior_project_pending.csv")
```

```python
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   Customer ID           3900 non-null   int64
 1   Age                   3900 non-null   int64
 2   Gender                3900 non-null   object
 3   Item Purchased        3900 non-null   object
 4   Category              3900 non-null   object
 5   Purchase Amount (USD)  3900 non-null  int64
 6   Location              3900 non-null   object
 7   Size                  3900 non-null   object
 8   Color                 3900 non-null   object
 9   Season                3900 non-null   object
 10  Review Rating         3863 non-null   float64
 11  Subscription Status   3900 non-null   object
 12  Shipping Type         3900 non-null   object
 13  Discount Applied      3900 non-null   object
 14  Promo Code Used       3900 non-null   object
 15  Previous Purchases    3900 non-null   int64
 16  Payment Method        3900 non-null   object
 17  Frequency of Purchases 3900 non-null  object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

```python
1 df.describe(include='all')
```

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases
count	3900.000000	3900.000000	3900	3900	3900	3900.000000	3900	3900	3900	3900	3863.000000	3900	3900	3900	3900	3900.000000
unique	NaN	NaN	2	25	4	NaN	50	4	25	4	NaN	2	6	2	2	NaN
top	NaN	NaN	Male	Blouse	Clothing	NaN	Montana	M	Olive	Spring	NaN	No	Free Shipping	No	No	NaN
freq	NaN	NaN	2652	171	1737	NaN	96	1755	177	999	NaN	2847	675	2223	2223	NaN
mean	1950.500000	44.068462	NaN	NaN	NaN	59.764359	NaN	NaN	NaN	NaN	3.750065	NaN	NaN	NaN	NaN	25.351538
std	1125.977353	15.207589	NaN	NaN	NaN	23.685392	NaN	NaN	NaN	NaN	0.716983	NaN	NaN	NaN	NaN	14.447125
min	1.000000	18.000000	NaN	NaN	NaN	20.000000	NaN	NaN	NaN	NaN	2.500000	NaN	NaN	NaN	NaN	1.000000
25%	975.750000	31.000000	NaN	NaN	NaN	39.000000	NaN	NaN	NaN	NaN	3.100000	NaN	NaN	NaN	NaN	13.000000
50%	1950.500000	44.000000	NaN	NaN	NaN	60.000000	NaN	NaN	NaN	NaN	3.800000	NaN	NaN	NaN	NaN	25.000000
75%	2925.250000	57.000000	NaN	NaN	NaN	81.000000	NaN	NaN	NaN	NaN	4.400000	NaN	NaN	NaN	NaN	38.000000
max	3900.000000	70.000000	NaN	NaN	NaN	100.000000	NaN	NaN	NaN	NaN	5.000000	NaN	NaN	NaN	NaN	50.000000

```
1 df.head()
```

	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating	Subscription Status	Shipping Type	Discount Applied	Promo Code Used	Previous Purchases	Payment Method	Freque Purcha
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	Yes	Yes	14	Venmo	Fortnig
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	Yes	Yes	2	Cash	Fortnig
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	Yes	Yes	23	Credit Card	Wee
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	Yes	Yes	49	PayPal	Wee
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	Yes	Yes	31	PayPal	Annu

```
1 df.isnull().sum()
```

	0
Customer ID	0
Age	0
Gender	0
Item Purchased	0
Category	0
Purchase Amount (USD)	0
Location	0
Size	0
Color	0
Season	0
Review Rating	37
Subscription Status	0
Shipping Type	0
Discount Applied	0
Promo Code Used	0
Previous Purchases	0
Payment Method	0
Frequency of Purchases	0

dtype: int64

```
1 df['Review Rating'] = df.groupby('Category')['Review Rating'].transform(lambda x: x.fillna(x.median()))
```

```
1 df.isnull().sum()
```

	0
Customer ID	0
Age	0
Gender	0
Item Purchased	0
Category	0
Purchase Amount (USD)	0
Location	0
Size	0
Color	0
Season	0
Review Rating	0
Subscription Status	0
Shipping Type	0
Discount Applied	0
Promo Code Used	0
Previous Purchases	0
Payment Method	0
Frequency of Purchases	0

dtype: int64

```
1 df.columns = df.columns.str.lower()
2 df.columns = df.columns.str.replace(' ','_')
3 df = df.rename(columns={'purchase_amount_(usd)':'purchase_amount'})
```

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 18 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   customer_id     3900 non-null   int64
 1   age             3900 non-null   int64
 2   gender          3900 non-null   object
 3   item_purchased  3900 non-null   object
 4   category        3900 non-null   object
 5   purchase_amount 3900 non-null   int64
 6   location        3900 non-null   object
 7   size            3900 non-null   object
 8   color           3900 non-null   object
 9   season          3900 non-null   object
```

```
 10  review_rating           3900 non-null    float64
 11  subscription_status     3900 non-null    object
 12  shipping_type           3900 non-null    object
 13  discount_applied        3900 non-null    object
 14  promo_code_used         3900 non-null    object
 15  previous_purchases      3900 non-null    int64
 16  payment_method          3900 non-null    object
 17  frequency_of_purchases  3900 non-null    object
dtypes: float64(1), int64(4), object(13)
memory usage: 548.6+ KB
```

```
1 df.columns
```

```
Index(['customer_id', 'age', 'gender', 'item_purchased', 'category',
       'purchase_amount', 'location', 'size', 'color', 'season',
       'review_rating', 'subscription_status', 'shipping_type',
       'discount_applied', 'promo_code_used', 'previous_purchases',
       'payment_method', 'frequency_of_purchases'],
      dtype='object')
```

```
1 df['age']
```

	age
0	55
1	19
2	50
3	21
4	45
...	...
3895	40
3896	52
3897	46
3898	44
3899	52

3900 rows × 1 columns

dtype: int64

```
1 labels = ['Young Adult', 'Adult', 'Middle-aged', 'Senior']
2 df['age_group'] = pd.qcut(df['age'], q=4, labels = labels)
```

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3900 entries, 0 to 3899
Data columns (total 19 columns):
 #   Column              Non-Null Count  Dtype
```

```
 ---  ------                --------------  -----
 0    customer_id           3900 non-null   int64
 1    age                   3900 non-null   int64
 2    gender                3900 non-null   object
 3    item_purchased        3900 non-null   object
 4    category              3900 non-null   object
 5    purchase_amount       3900 non-null   int64
 6    location              3900 non-null   object
 7    size                  3900 non-null   object
 8    color                 3900 non-null   object
 9    season                3900 non-null   object
 10   review_rating         3900 non-null   float64
 11   subscription_status   3900 non-null   object
 12   shipping_type         3900 non-null   object
 13   discount_applied      3900 non-null   object
 14   promo_code_used       3900 non-null   object
 15   previous_purchases    3900 non-null   int64
 16   payment_method        3900 non-null   object
 17   frequency_of_purchases 3900 non-null  object
 18   age_group             3900 non-null   category
dtypes: category(1), float64(1), int64(4), object(13)
memory usage: 552.6+ KB
```

```
1 df
```

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	review_rating	subscription_status	shipping_type	dis
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.1	Yes	Express	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.1	Yes	Express	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.1	Yes	Free Shipping	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.5	Yes	Next Day Air	
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.7	Yes	Free Shipping	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4.2	No	2-Day Shipping	
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4.5	No	Store Pickup	
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2.9	No	Standard	
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3.8	No	Express	
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3.1	No	Store Pickup	

3900 rows × 19 columns

```
1 df[['age','age_group']].head(10)
```

	age	age_group
0	55	Middle-aged
1	19	Young Adult
2	50	Middle-aged
3	21	Young Adult
4	45	Middle-aged
5	46	Middle-aged
6	63	Senior
7	27	Young Adult
8	26	Young Adult
9	57	Middle-aged

```
1 frequency_mapping = {
2     'Fortnightly': 14,
3     'Weekly': 7,
4     'Monthly': 30,
5     'Quarterly': 90,
6     'Bi-Weekly': 14,
7     'Annually': 365,
8     'Every 3 Months': 90
9 }
10
11 df['purchase_frequency_days'] = df['frequency_of_purchases'].map(frequency_mapping)
```

```
1 df[['purchase_frequency_days','frequency_of_purchases']].head(10)
```

	purchase_frequency_days	frequency_of_purchases
0	14	Fortnightly
1	14	Fortnightly
2	7	Weekly
3	7	Weekly
4	365	Annually
5	7	Weekly
6	90	Quarterly
7	7	Weekly
8	365	Annually
9	90	Quarterly

```
1 df[['discount_applied','promo_code_used']].head(10)
```

	discount_applied	promo_code_used
0	Yes	Yes
1	Yes	Yes
2	Yes	Yes
3	Yes	Yes
4	Yes	Yes
5	Yes	Yes
6	Yes	Yes
7	Yes	Yes
8	Yes	Yes
9	Yes	Yes

```
1 f = df.drop('promo_code_used', axis=1)
```

```
1 df.columns
```

```
Index(['customer_id', 'age', 'gender', 'item_purchased', 'category',
       'purchase_amount', 'location', 'size', 'color', 'season',
       'review_rating', 'subscription_status', 'shipping_type',
       'discount_applied', 'promo_code_used', 'previous_purchases',
       'payment_method', 'frequency_of_purchases', 'age_group',
       'purchase_frequency_days'],
      dtype='object')
```

```
1 Start coding or generate with AI.
```

```
 1 # To extract day of the week and month, we need a datetime column.
 2 # Assuming the dataset does not have a specific date column, we can create a dummy date column
 3 # using the existing 'customer_id' or simply create a sequence of dates.
 4 # Let's create a sequence of dates for demonstration purposes.
 5
 6 # First, ensure pandas is imported (already done in the notebook)
 7 # import pandas as pd
 8
 9 # Create a date range with the same number of entries as the DataFrame
10 date_range = pd.date_range(start='2023-01-01', periods=len(df), freq='D')
11
12 # Add a 'purchase_date' column to the DataFrame
13 df['purchase_date'] = date_range
14
15 # Extract the day of the week and month
16 df['day_of_week'] = df['purchase_date'].dt.day_name()
17 df['month'] = df['purchase_date'].dt.month_name()
18
19 # Display the updated DataFrame with the new columns
20 display(df.head())
```

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	...	discount_applied	promo_code_used	previous_purchases
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	...	Yes	Yes	14
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	...	Yes	Yes	2
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	...	Yes	Yes	23
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	...	Yes	Yes	49
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	...	Yes	Yes	31

5 rows × 23 columns

```
1 display(df)
```

	customer_id	age	gender	item_purchased	category	purchase_amount	location	size	color	season	...	discount_applied	promo_code_used	previous_purch
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	...	Yes	Yes	
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	...	Yes	Yes	
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	...	Yes	Yes	
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	...	Yes	Yes	