consumer-behaviour

July 2, 2024

1 Abstract & Introduction

1.0.1 Context

- The Consumer Behaviour and Shopping Habits The dataset provides comprehensive insights into consumers' preferences, tendencies, and patterns during their shopping experiences.
- This dataset captures a diversity of customer attributes including age, gender, purchase history, preferred payment methods, frequency of purchases, and more.
- The dataset is valuable for businesses aiming to align their strategies with customer needs and preferences.
- Understanding customer preferences and trends is critical for businesses to tailor the Analysects, marketing strategies, and overall customer experience. Analyzing this data can help business makes informed decisions, optimise product offerings, and enhance customer satisfaction.
- Additionally, data on the type of items purchased, shopping frequency, preferred shopping seasons, and interactions with promotional offers is included. With a collection of 3900 recored, this dataset serves as a foundation for business looking to apply data-driven insights for better decision-making and customer-centeric strategies.

1.1 Mount Google Drive

[ ]:
```
from google.colab import drive
drive.mount('/content/drive/')
```

```
Drive already mounted at /content/drive/; to attempt to forcibly remount, call
drive.mount("/content/drive/", force_remount=True).
```

2 Import Libraries

[ ]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
pd.options.display.float_format = '{:,.2f}'.format
```

3 Analysing Data

```
[ ]: behaviour_df = pd.read_csv('/content/drive/MyDrive/data/
      ↪shopping_behavior_updated.csv', index_col='Customer ID')
```

3.1 Initial Inspection

```
[ ]: behaviour_df.head()
```

```
[ ]:              Age Gender Item Purchased  Category  Purchase Amount (USD)  \
Customer ID
1                 55   Male         Blouse  Clothing                     53
2                 19   Male        Sweater  Clothing                     64
3                 50   Male          Jeans  Clothing                     73
4                 21   Male        Sandals  Footwear                     90
5                 45   Male         Blouse  Clothing                     49

                    Location Size      Color  Season  Review Rating  \
Customer ID
1                       Kentucky    L       Gray  Winter           3.10
2                          Maine    L     Maroon  Winter           3.10
3                  Massachusetts    S     Maroon  Spring           3.10
4                   Rhode Island    M     Maroon  Spring           3.50
5                         Oregon    M  Turquoise  Spring           2.70

               Subscription Status  Shipping Type Discount Applied  \
Customer ID
1                              Yes        Express              Yes
2                              Yes        Express              Yes
3                              Yes  Free Shipping              Yes
4                              Yes   Next Day Air              Yes
5                              Yes  Free Shipping              Yes

               Promo Code Used  Previous Purchases Payment Method  \
Customer ID
1                          Yes                  14          Venmo
2                          Yes                   2           Cash
3                          Yes                  23    Credit Card
4                          Yes                  49         PayPal
5                          Yes                  31         PayPal

               Frequency of Purchases
Customer ID
1                         Fortnightly
2                         Fortnightly
3                              Weekly
4                              Weekly
```

```
5                         Annually
```

[ ]:
```
behaviour_df.tail()
```

[ ]:
```
             Age  Gender Item Purchased     Category  Purchase Amount (USD)  \
Customer ID
3896          40  Female         Hoodie     Clothing                     28
3897          52  Female       Backpack  Accessories                     49
3898          46  Female           Belt  Accessories                     33
3899          44  Female          Shoes     Footwear                     77
3900          52  Female        Handbag  Accessories                     81

               Location Size      Color  Season  Review Rating  \
Customer ID
3896            Virginia    L  Turquoise  Summer           4.20
3897                Iowa    L      White  Spring           4.50
3898          New Jersey    L      Green  Spring           2.90
3899           Minnesota    S      Brown  Summer           3.80
3900          California    M      Beige  Spring           3.10

            Subscription Status   Shipping Type Discount Applied  \
Customer ID
3896                         No  2-Day Shipping              No
3897                         No    Store Pickup              No
3898                         No        Standard              No
3899                         No         Express              No
3900                         No    Store Pickup              No

            Promo Code Used  Previous Purchases Payment Method  \
Customer ID
3896                     No                  32          Venmo
3897                     No                  41  Bank Transfer
3898                     No                  24          Venmo
3899                     No                  24          Venmo
3900                     No                  33          Venmo

            Frequency of Purchases
Customer ID
3896                        Weekly
3897                     Bi-Weekly
3898                     Quarterly
3899                        Weekly
3900                     Quarterly
```

[ ]:
```
behaviour_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
```

```
Index: 3900 entries, 1 to 3900
Data columns (total 17 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   Age                     3900 non-null   int64
 1   Gender                  3900 non-null   object
 2   Item Purchased          3900 non-null   object
 3   Category                3900 non-null   object
 4   Purchase Amount (USD)   3900 non-null   int64
 5   Location                3900 non-null   object
 6   Size                    3900 non-null   object
 7   Color                   3900 non-null   object
 8   Season                  3900 non-null   object
 9   Review Rating           3900 non-null   float64
 10  Subscription Status     3900 non-null   object
 11  Shipping Type           3900 non-null   object
 12  Discount Applied        3900 non-null   object
 13  Promo Code Used         3900 non-null   object
 14  Previous Purchases      3900 non-null   int64
 15  Payment Method          3900 non-null   object
 16  Frequency of Purchases  3900 non-null   object
dtypes: float64(1), int64(3), object(13)
memory usage: 548.4+ KB
```

[ ]:
```
behaviour_df.dtypes
```

[ ]:
```
Age                        int64
Gender                    object
Item Purchased            object
Category                  object
Purchase Amount (USD)      int64
Location                  object
Size                      object
Color                     object
Season                    object
Review Rating            float64
Subscription Status       object
Shipping Type             object
Discount Applied          object
Promo Code Used           object
Previous Purchases         int64
Payment Method            object
Frequency of Purchases    object
dtype: object
```

3.2 Analysing the variables in the Dataset

[ ]:
```
behaviour_df.shape
```

[ ]:
```
(3900, 17)
```

- There are 3900 observations (records) and 17 variables (features) in our dataset.

[ ]:
```
behaviour_df.columns
```

[ ]:
```
Index(['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',
       'Frequency of Purchases'],
      dtype='object')
```

Variables Dictionary: * Customer ID - Unique identifier for each customer. * Age - Age of the customer. * Gender - Gender of the customer (Male/Female). * Item Purchased - The item purchased by the customer. * Category - Category of the item purchased. * Purchase Amount (USD) - The amount of the purchase in USD. * Location - Location where the purchase was made. * Size - Size of the purchased item. * Color - Color of the purchased item. * Season - Season during which the purchase was made. * Review Rating - Rating given by the customer for the purchased item. * Subscription Status - Indicates if the customer has a subscription (Yes/No). * Shipping Type - Type of shipping chosen by the customer. * Discount Applied - Indicates if a discount was applied to the purchase (Yes/No). * Promo Code Used - Indicates if a promo code was used for the purchase (Yes/No). * Previous Purchases - Number of previous purchases made by the customer. * Payment Method - Customer's most preferred payment method. * Frequency of Purchases - Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

3.3 Statistic Summary

[ ]:
```
behaviour_df.describe().T
```

[ ]:

	count	mean	std	min	25%	50%	75%	max
Age	3,900.00	44.07	15.21	18.00	31.00	44.00	57.00	70.00
Purchase Amount (USD)	3,900.00	59.76	23.69	20.00	39.00	60.00	81.00	100.00
Review Rating	3,900.00	3.75	0.72	2.50	3.10	3.70	4.40	5.00
Previous Purchases	3,900.00	25.35	14.45	1.00	13.00	25.00	38.00	50.00

From the Statistic summary above, we can infer the findings: > * The Generation spread out from 18-year-old to 70-year-old consumers. > * The Fashion Hierarchy most consumers choose is between Mass Market and Mid-end. > * The Review rating is just around the average, which should be considered about other components such as price, quality, service, shipping types, promotion events, etc. > * The number of customers coming back shows a positive perspective.

[ ]:
```
behaviour_df.describe(include='all').T
```

[ ]:

	count	unique	top	freq	mean	std
Age	3,900.00	NaN	NaN	NaN	44.07	15.21
Gender	3900	2	Male	2652	NaN	NaN
Item Purchased	3900	25	Blouse	171	NaN	NaN
Category	3900	4	Clothing	1737	NaN	NaN
Purchase Amount (USD)	3,900.00	NaN	NaN	NaN	59.76	23.69
Location	3900	50	Montana	96	NaN	NaN
Size	3900	4	M	1755	NaN	NaN
Color	3900	25	Olive	177	NaN	NaN
Season	3900	4	Spring	999	NaN	NaN
Review Rating	3,900.00	NaN	NaN	NaN	3.75	0.72
Subscription Status	3900	2	No	2847	NaN	NaN
Shipping Type	3900	6	Free Shipping	675	NaN	NaN
Discount Applied	3900	2	No	2223	NaN	NaN
Promo Code Used	3900	2	No	2223	NaN	NaN
Previous Purchases	3,900.00	NaN	NaN	NaN	25.35	14.45
Payment Method	3900	6	PayPal	677	NaN	NaN
Frequency of Purchases	3900	7	Every 3 Months	584	NaN	NaN

	min	25%	50%	75%	max
Age	18.00	31.00	44.00	57.00	70.00
Gender	NaN	NaN	NaN	NaN	NaN
Item Purchased	NaN	NaN	NaN	NaN	NaN
Category	NaN	NaN	NaN	NaN	NaN
Purchase Amount (USD)	20.00	39.00	60.00	81.00	100.00
Location	NaN	NaN	NaN	NaN	NaN
Size	NaN	NaN	NaN	NaN	NaN
Color	NaN	NaN	NaN	NaN	NaN
Season	NaN	NaN	NaN	NaN	NaN
Review Rating	2.50	3.10	3.70	4.40	5.00
Subscription Status	NaN	NaN	NaN	NaN	NaN
Shipping Type	NaN	NaN	NaN	NaN	NaN
Discount Applied	NaN	NaN	NaN	NaN	NaN
Promo Code Used	NaN	NaN	NaN	NaN	NaN
Previous Purchases	1.00	13.00	25.00	38.00	50.00
Payment Method	NaN	NaN	NaN	NaN	NaN
Frequency of Purchases	NaN	NaN	NaN	NaN	NaN

3.4 Calculating the null/missing values

[ ]:
```
behaviour_df.isnull().sum()
```

[ ]:
```
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
```

- There is **no NULL/missing values**

3.5 Check for duplication in dataset

[ ]:
```
behaviour_df.duplicated().sum()
```

[ ]:
```
0
```

- There is **no duplicated value** existing in the Dataset

3.6 Check for the number of unique values in each variable

[ ]:
```
behaviour_df.nunique()
```

[ ]:
```
Age                       53
Gender                     2
Item Purchased            25
Category                   4
Purchase Amount (USD)     81
Location                  50
Size                       4
Color                     25
Season                     4
Review Rating             26
Subscription Status        2
Shipping Type              6
Discount Applied           2
Promo Code Used            2
Previous Purchases        50
Payment Method             6
Frequency of Purchases     7
dtype: int64
```

```
[ ]: behaviour_df['Age'].unique()
```

```
[ ]: array([55, 19, 50, 21, 45, 46, 63, 27, 26, 57, 53, 30, 61, 65, 64, 25, 52,
            66, 31, 56, 18, 38, 54, 33, 36, 35, 29, 70, 69, 67, 20, 39, 42, 68,
            49, 59, 47, 40, 41, 48, 22, 24, 44, 37, 58, 32, 62, 51, 28, 43, 34,
            23, 60])
```

```
[ ]: behaviour_df['Item Purchased'].unique()
```

```
[ ]: array(['Blouse', 'Sweater', 'Jeans', 'Sandals', 'Sneakers', 'Shirt',
           'Shorts', 'Coat', 'Handbag', 'Shoes', 'Dress', 'Skirt',
           'Sunglasses', 'Pants', 'Jacket', 'Hoodie', 'Jewelry', 'T-shirt',
           'Scarf', 'Hat', 'Socks', 'Backpack', 'Belt', 'Boots', 'Gloves'],
          dtype=object)
```

```
[ ]: behaviour_df['Category'].unique()
```

```
[ ]: array(['Clothing', 'Footwear', 'Outerwear', 'Accessories'], dtype=object)
```

```
[ ]: behaviour_df['Location'].unique()
```

```
[ ]: array(['Kentucky', 'Maine', 'Massachusetts', 'Rhode Island', 'Oregon',
           'Wyoming', 'Montana', 'Louisiana', 'West Virginia', 'Missouri',
           'Arkansas', 'Hawaii', 'Delaware', 'New Hampshire', 'New York',
           'Alabama', 'Mississippi', 'North Carolina', 'California',
           'Oklahoma', 'Florida', 'Texas', 'Nevada', 'Kansas', 'Colorado',
           'North Dakota', 'Illinois', 'Indiana', 'Arizona', 'Alaska',
           'Tennessee', 'Ohio', 'New Jersey', 'Maryland', 'Vermont',
           'New Mexico', 'South Carolina', 'Idaho', 'Pennsylvania',
           'Connecticut', 'Utah', 'Virginia', 'Georgia', 'Nebraska', 'Iowa',
           'South Dakota', 'Minnesota', 'Washington', 'Wisconsin', 'Michigan'],
          dtype=object)
```

```
[ ]: behaviour_df['Size'].unique()
```

```
[ ]: array(['L', 'S', 'M', 'XL'], dtype=object)
```

```
[ ]: behaviour_df['Color'].unique()
```

```
[ ]: array(['Gray', 'Maroon', 'Turquoise', 'White', 'Charcoal', 'Silver',
           'Pink', 'Purple', 'Olive', 'Gold', 'Violet', 'Teal', 'Lavender',
           'Black', 'Green', 'Peach', 'Red', 'Cyan', 'Brown', 'Beige',
           'Orange', 'Indigo', 'Yellow', 'Magenta', 'Blue'], dtype=object)
```

```
[ ]: behaviour_df['Previous Purchases'].unique()
```

[ ]:
```
array([14,  2, 23, 49, 31, 19,  8,  4, 26, 10, 37, 34, 44, 36, 17, 46, 50,
       22, 32, 40, 16, 13,  7, 41, 45, 38, 48, 18, 15, 25, 39, 35, 29, 21,
       43,  3,  5, 24, 42, 47, 28, 20, 33,  1,  9, 12, 27, 11, 30,  6])
```

[ ]:
```
behaviour_df['Payment Method'].unique()
```

[ ]:
```
array(['Venmo', 'Cash', 'Credit Card', 'PayPal', 'Bank Transfer',
       'Debit Card'], dtype=object)
```

[ ]:
```
behaviour_df['Frequency of Purchases'].unique()
```

[ ]:
```
array(['Fortnightly', 'Weekly', 'Annually', 'Quarterly', 'Bi-Weekly',
       'Monthly', 'Every 3 Months'], dtype=object)
```

4 Exploratory Data Analysis

The EDA step would be much easier because there is no duplicate, missing or null value as the analysing data above.

4.1 Separate Numerical and Categorical variables for easy analysis

[ ]:
```
num_var = behaviour_df.select_dtypes(include=['int64', 'float64'])
cat_var = behaviour_df.select_dtypes(include=['object'])

print(f'Numerical Variables: {num_var.columns}')
print()
print(f'Categorical Variables: {cat_var.columns}')
```

```
Numerical Variables: Index(['Age', 'Purchase Amount (USD)', 'Review Rating',
'Previous Purchases'], dtype='object')

Categorical Variables: Index(['Gender', 'Item Purchased', 'Category',
'Location', 'Size', 'Color',
       'Season', 'Subscription Status', 'Shipping Type', 'Discount Applied',
       'Promo Code Used', 'Payment Method', 'Frequency of Purchases'],
      dtype='object')
```

4.2 Univariate Analysis

[ ]:
```
# Counting the amount of the Item Purchased
behaviour_df['Item Purchased'].value_counts()
```

[ ]:
```
Item Purchased
Blouse        171
Jewelry       171
Pants         171
Shirt         169
```

```
Dress          166
Sweater        164
Jacket         163
Belt           161
Sunglasses     161
Coat           161
Sandals        160
Socks          159
Skirt          158
Shorts         157
Scarf          157
Hat            154
Handbag        153
Hoodie         151
Shoes          150
T-shirt        147
Sneakers       145
Boots          144
Backpack       143
Gloves         140
Jeans          124
Name: count, dtype: int64
```

[ ]:
```
# Counting the amount of different Categories
behaviour_df['Category'].value_counts()
```

[ ]:
```
Category
Clothing       1737
Accessories    1240
Footwear        599
Outerwear       324
Name: count, dtype: int64
```

[ ]:
```
# Counting the amount of different Locations
behaviour_df['Location'].value_counts()
```

[ ]:
```
Location
Montana           96
California        95
Idaho             93
Illinois          92
Alabama           89
Minnesota         88
Nebraska          87
New York          87
Nevada            87
Maryland          86
```

```
Delaware          86
Vermont           85
Louisiana         84
North Dakota      83
Missouri          81
West Virginia     81
New Mexico        81
Mississippi       80
Indiana           79
Georgia           79
Kentucky          79
Arkansas          79
North Carolina    78
Connecticut       78
Virginia          77
Ohio              77
Tennessee         77
Texas             77
Maine             77
South Carolina    76
Colorado          75
Oklahoma          75
Wisconsin         75
Oregon            74
Pennsylvania      74
Washington        73
Michigan          73
Alaska            72
Massachusetts     72
Wyoming           71
Utah              71
New Hampshire     71
South Dakota      70
Iowa              69
Florida           68
New Jersey        67
Hawaii            65
Arizona           65
Kansas            63
Rhode Island      63
Name: count, dtype: int64
```

[ ]:
```
# Counting the Gender
behaviour_df['Gender'].value_counts()
```

[ ]:
```
Gender
Male      2652
```

```
Female    1248
Name: count, dtype: int64
```

[ ]:
```
# Counting the Frequency of Purchases
behaviour_df['Frequency of Purchases'].value_counts()
```

[ ]:
```
Frequency of Purchases
Every 3 Months    584
Annually          572
Quarterly         563
Monthly           553
Bi-Weekly         547
Fortnightly       542
Weekly            539
Name: count, dtype: int64
```

[ ]:
```
# Counting the number of different colours
behaviour_df['Color'].value_counts()
```

[ ]:
```
Color
Olive        177
Yellow       174
Silver       173
Teal         172
Green        169
Black        167
Cyan         166
Violet       166
Gray         159
Maroon       158
Orange       154
Charcoal     153
Pink         153
Magenta      152
Blue         152
Purple       151
Peach        149
Red          148
Beige        147
Indigo       147
Lavender     147
Turquoise    145
White        142
Brown        141
Gold         138
Name: count, dtype: int64
```

[ ]:
```
# Counting the number of Subscription Status
behaviour_df['Subscription Status'].value_counts()
```

[ ]:
```
Subscription Status
No     2847
Yes    1053
Name: count, dtype: int64
```

[ ]:
```
# Counting the number of Discount Applied
behaviour_df['Discount Applied'].value_counts()
```

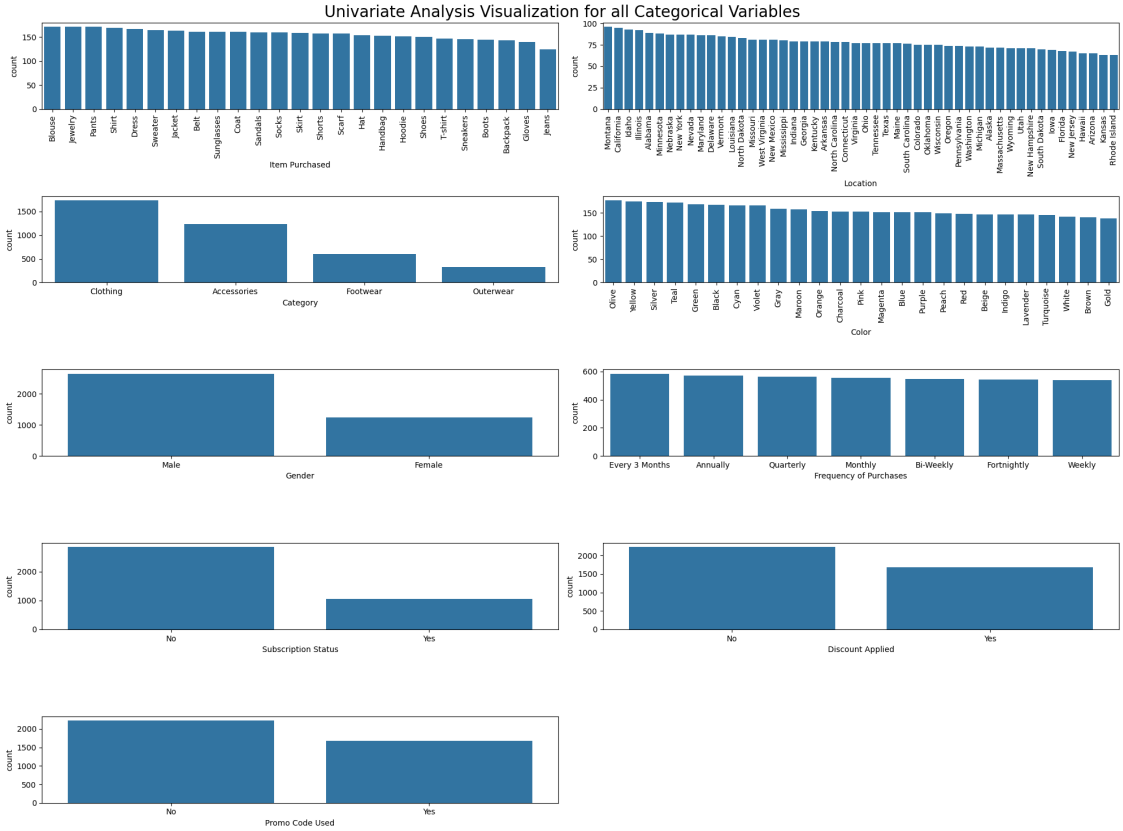
[ ]:
```
Discount Applied
No     2223
Yes    1677
Name: count, dtype: int64
```

[ ]:
```
# Counting the number of Promo Code Used
behaviour_df['Promo Code Used'].value_counts()
```

[ ]:
```
Promo Code Used
No     2223
Yes    1677
Name: count, dtype: int64
```

[ ]:
```
# Visualizing all variables
fig, axes = plt.subplots(nrows=5, ncols=2, figsize=(20, 15))
fig.suptitle('Univariate Analysis Visualization for all Categorical Variables',␣
 ↪fontsize=20);

sns.countplot(ax = axes[0,0], data=behaviour_df, x = 'Item Purchased', order =␣
 ↪behaviour_df['Item Purchased'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=90);
sns.countplot(ax = axes[0,1], data=behaviour_df, x = 'Location', order =␣
 ↪behaviour_df['Location'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=90);
sns.countplot(ax = axes[1,0], data=behaviour_df, x = 'Category', order =␣
 ↪behaviour_df['Category'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=0);
sns.countplot(ax = axes[1,1], data=behaviour_df, x = 'Color', order =␣
 ↪behaviour_df['Color'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=90);
sns.countplot(ax = axes[2,0], data=behaviour_df, x = 'Gender', order =␣
 ↪behaviour_df['Gender'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=0);
sns.countplot(ax = axes[2,1], data=behaviour_df, x = 'Frequency of Purchases',␣
 ↪order = behaviour_df['Frequency of Purchases'].value_counts().index).
 ↪tick_params(axis='x', rotation=0);
```

```
sns.countplot(ax = axes[3,0], data=behaviour_df, x = 'Subscription Status',␣
 ↪order = behaviour_df['Subscription Status'].value_counts().index).
 ↪tick_params(axis='x', rotation=0);
sns.countplot(ax = axes[3,1], data=behaviour_df, x = 'Discount Applied', order␣
 ↪= behaviour_df['Discount Applied'].value_counts().index).
 ↪tick_params(axis='x', rotation=0);
sns.countplot(ax = axes[4,0], data=behaviour_df, x = 'Promo Code Used', order =␣
 ↪behaviour_df['Promo Code Used'].value_counts().index).tick_params(axis='x',␣
 ↪rotation=0);

axes.flat[-1].set_visible(False);
plt.tight_layout();
```



From the count plots, we can see the observations: * 68% of consumers are male. However, according to plot fg No.1, blouses and jewellery, which are the most the number of items purchased, dominating approximately 8.8%, those items belong to females. * Unsurprisingly, clothing is the most prioritized with approximately 44.5%. * Montana is the city that has the most number of consumers, followed by California, Idaho, and Illinois. * Olive, yellow, and silver are the most popular choices in colour. * The frequency of purchases plot shows that consumers purchase the items quite consistently. * The last 3 count plots present the consumers who seem to be not affected by the promotion program.

4.3 Bivariate Analysis

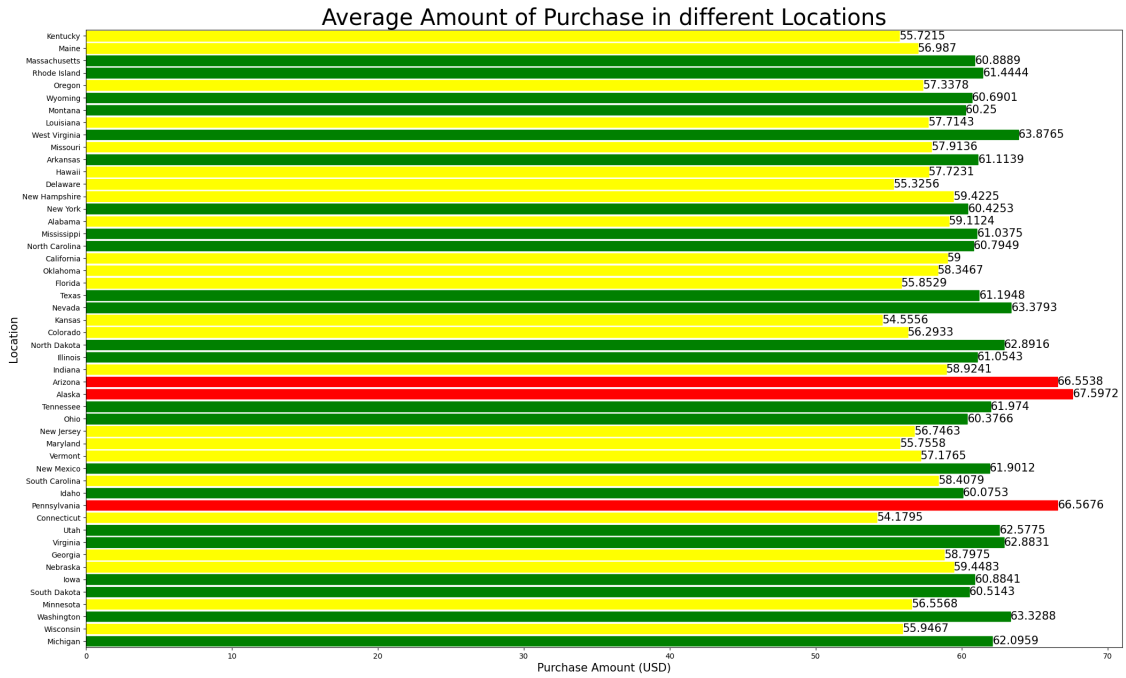
[ ]:
```
# Grouping Location with Purchase Amount
behaviour_df.groupby(['Location'])['Purchase Amount (USD)'].agg(['sum',␣
↪'count', 'mean']).sort_values(by='sum', ascending=False)
```

[ ]:

Location	sum	count	mean
Montana	5784	96	60.25
Illinois	5617	92	61.05
California	5605	95	59.00
Idaho	5587	93	60.08
Nevada	5514	87	63.38
Alabama	5261	89	59.11
New York	5257	87	60.43
North Dakota	5220	83	62.89
West Virginia	5174	81	63.88
Nebraska	5172	87	59.45
New Mexico	5014	81	61.90
Minnesota	4977	88	56.56
Pennsylvania	4926	74	66.57
Mississippi	4883	80	61.04
Alaska	4867	72	67.60
Vermont	4860	85	57.18
Louisiana	4848	84	57.71
Virginia	4842	77	62.88
Arkansas	4828	79	61.11
Maryland	4795	86	55.76
Tennessee	4772	77	61.97
Delaware	4758	86	55.33
North Carolina	4742	78	60.79
Texas	4712	77	61.19
Missouri	4691	81	57.91
Indiana	4655	79	58.92
Ohio	4649	77	60.38
Georgia	4645	79	58.80
Washington	4623	73	63.33
Michigan	4533	73	62.10
Utah	4443	71	62.58
South Carolina	4439	76	58.41
Kentucky	4402	79	55.72
Maine	4388	77	56.99
Massachusetts	4384	72	60.89
Oklahoma	4376	75	58.35
Arizona	4326	65	66.55
Wyoming	4309	71	60.69
Oregon	4243	74	57.34

```
South Dakota    4236     70 60.51
Connecticut     4226     78 54.18
Colorado        4222     75 56.29
New Hampshire   4219     71 59.42
Iowa            4201     69 60.88
Wisconsin       4196     75 55.95
Rhode Island    3871     63 61.44
New Jersey      3802     67 56.75
Florida         3798     68 55.85
Hawaii          3752     65 57.72
Kansas          3437     63 54.56
```

[ ]:
```python
# Visualizing the Average Amount of Purchases in different locations
# Creating a bar plot for visualization
plt.figure(figsize=(25,15))
plt.title('Average Amount of Purchase in different Locations', fontsize=30)
barplot_location = sns.barplot(data=behaviour_df, x='Purchase Amount (USD)',␣
 ↪y='Location', errorbar=None);
# Labelling the value of the number of purchases
barplot_location.bar_label(barplot_location.containers[0], fontsize = 15);
# Customize the bar plot
for bar in barplot_location.patches:
    if bar.get_width() > 65:
        bar.set_color('red')
    elif 60 < bar.get_width() < 65:
        bar.set_color('green')
    else:
        bar.set_color('yellow')
barplot_location.set_xlabel('Purchase Amount (USD)', fontsize=15)
barplot_location.set_ylabel('Location', fontsize=15)
plt.show()
```



- The cities that have spent more than 65 thousand dollars are Arizona, Alaska, and Pennsylvania.

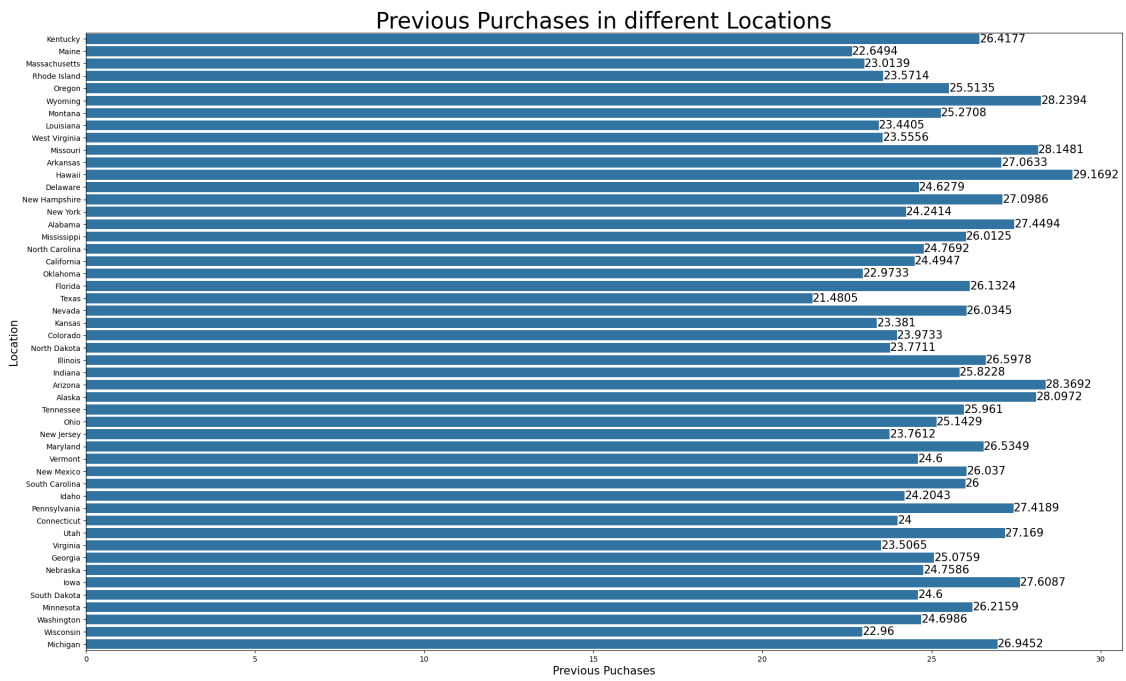
[ ]:
```
# Grouping Location with Previous Purchase
behaviour_df.groupby(['Location'])['Previous Purchases'].agg(['sum', 'count',␣
↪'mean']).sort_values(by='sum', ascending=False)
```

[ ]:

Location	sum	count	mean
Illinois	2447	92	26.60
Alabama	2443	89	27.45
Montana	2426	96	25.27
California	2327	95	24.49
Minnesota	2307	88	26.22
Maryland	2282	86	26.53
Missouri	2280	81	28.15
Nevada	2265	87	26.03
Idaho	2251	93	24.20
Nebraska	2154	87	24.76
Arkansas	2138	79	27.06
Delaware	2118	86	24.63
New Mexico	2109	81	26.04
New York	2109	87	24.24
Vermont	2091	85	24.60
Kentucky	2087	79	26.42
Mississippi	2081	80	26.01

```
Indiana          2040     79 25.82
Pennsylvania     2029     74 27.42
Alaska           2023     72 28.10
Wyoming          2005     71 28.24
Tennessee        1999     77 25.96
Georgia          1981     79 25.08
South Carolina   1976     76 26.00
North Dakota     1973     83 23.77
Louisiana        1969     84 23.44
Michigan         1967     73 26.95
Ohio             1936     77 25.14
North Carolina   1932     78 24.77
Utah             1929     71 27.17
New Hampshire    1924     71 27.10
West Virginia    1908     81 23.56
Iowa             1905     69 27.61
Hawaii           1896     65 29.17
Oregon           1888     74 25.51
Connecticut      1872     78 24.00
Arizona          1844     65 28.37
Virginia         1810     77 23.51
Washington       1803     73 24.70
Colorado         1798     75 23.97
Florida          1777     68 26.13
Maine            1744     77 22.65
Oklahoma         1723     75 22.97
South Dakota     1722     70 24.60
Wisconsin        1722     75 22.96
Massachusetts    1657     72 23.01
Texas            1654     77 21.48
New Jersey       1592     67 23.76
Rhode Island     1485     63 23.57
Kansas           1473     63 23.38
```

[ ]:
```python
# Visualizing the Previous Purchase in different Locations
# Creating a bar plot for visualization
plt.figure(figsize=(25,15))
plt.title('Previous Purchases in different Locations', fontsize=30)
barplot_location = sns.barplot(data=behaviour_df, x='Previous Purchases',␣
 ↪y='Location', errorbar=None);
# Labelling the value of the number of purchases
barplot_location.bar_label(barplot_location.containers[0], fontsize = 15);
# Customize the bar plot
barplot_location.set_xlabel('Previous Puchases', fontsize=15)
barplot_location.set_ylabel('Location', fontsize=15)
plt.show()
```



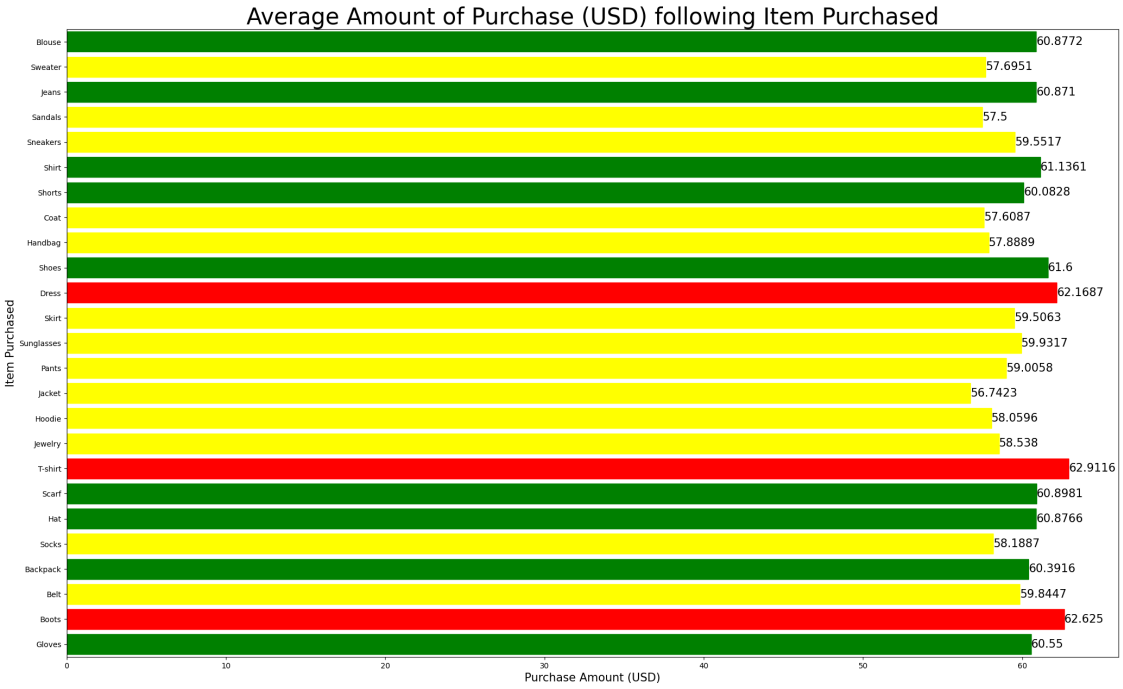
[ ]:
```
# Grouping Purchase Amount by Item Purchased
behaviour_df.groupby(['Item Purchased'])['Purchase Amount (USD)'].agg(['sum',␣
↪'count', 'mean']).sort_values(by='sum', ascending=False)
```

[ ]:

Item Purchased	sum	count	mean
Blouse	10410	171	60.88
Shirt	10332	169	61.14
Dress	10320	166	62.17
Pants	10090	171	59.01
Jewelry	10010	171	58.54
Sunglasses	9649	161	59.93
Belt	9635	161	59.84
Scarf	9561	157	60.90
Sweater	9462	164	57.70
Shorts	9433	157	60.08
Skirt	9402	158	59.51
Hat	9375	154	60.88
Coat	9275	161	57.61
Socks	9252	159	58.19
Jacket	9249	163	56.74
T-shirt	9248	147	62.91
Shoes	9240	150	61.60
Sandals	9200	160	57.50
Boots	9018	144	62.62

```
Handbag          8857    153 57.89
Hoodie           8767    151 58.06
Backpack         8636    143 60.39
Sneakers         8635    145 59.55
Gloves           8477    140 60.55
Jeans            7548    124 60.87
```

[ ]:
```
# Visualizing the Average Amount of Purchases following Item Purchased
# Creating a bar plot for visualization
plt.figure(figsize=(25,15))
plt.title('Average Amount of Purchase (USD) following Item Purchased',␣
 ↪fontsize=30)
barplot_item = sns.barplot(data=behaviour_df, x='Purchase Amount (USD)',␣
 ↪y='Item Purchased', errorbar=None);
# Labelling the value of the number of purchases
barplot_item.bar_label(barplot_item.containers[0], fontsize = 15);
# Customize the bar plot
for bar in barplot_item.patches:
    if bar.get_width() > 62:
        bar.set_color('red')
    elif 60 < bar.get_width() < 62:
        bar.set_color('green')
    else:
        bar.set_color('yellow')
barplot_item.set_xlabel('Purchase Amount (USD)', fontsize=15)
barplot_item.set_ylabel('Item Purchased', fontsize=15)
plt.show();
```



- The amount of sold items such as T-shirts, Dress, and Boots are not more impressive than others, but they brought the most income.

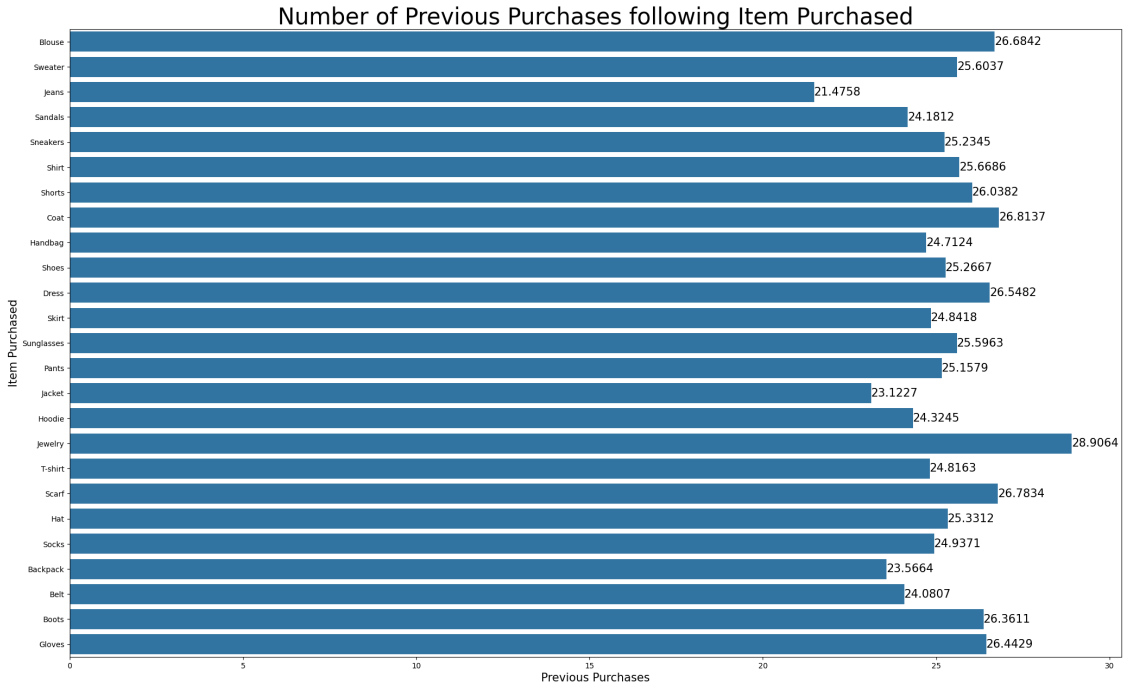
[ ]:
```
# Grouping number of items previously purchased by type of item purchased
behaviour_df.groupby(['Item Purchased'])['Previous Purchases'].agg(['sum',␣
↪'count', 'mean']).sort_values(by='sum', ascending=False)
```

[ ]:

Item Purchased	sum	count	mean
Jewelry	4943	171	28.91
Blouse	4563	171	26.68
Dress	4407	166	26.55
Shirt	4338	169	25.67
Coat	4317	161	26.81
Pants	4302	171	25.16
Scarf	4205	157	26.78
Sweater	4199	164	25.60
Sunglasses	4121	161	25.60
Shorts	4088	157	26.04
Socks	3965	159	24.94
Skirt	3925	158	24.84
Hat	3901	154	25.33
Belt	3877	161	24.08
Sandals	3869	160	24.18
Boots	3796	144	26.36
Shoes	3790	150	25.27
Handbag	3781	153	24.71
Jacket	3769	163	23.12
Gloves	3702	140	26.44
Hoodie	3673	151	24.32
Sneakers	3659	145	25.23
T-shirt	3648	147	24.82
Backpack	3370	143	23.57
Jeans	2663	124	21.48

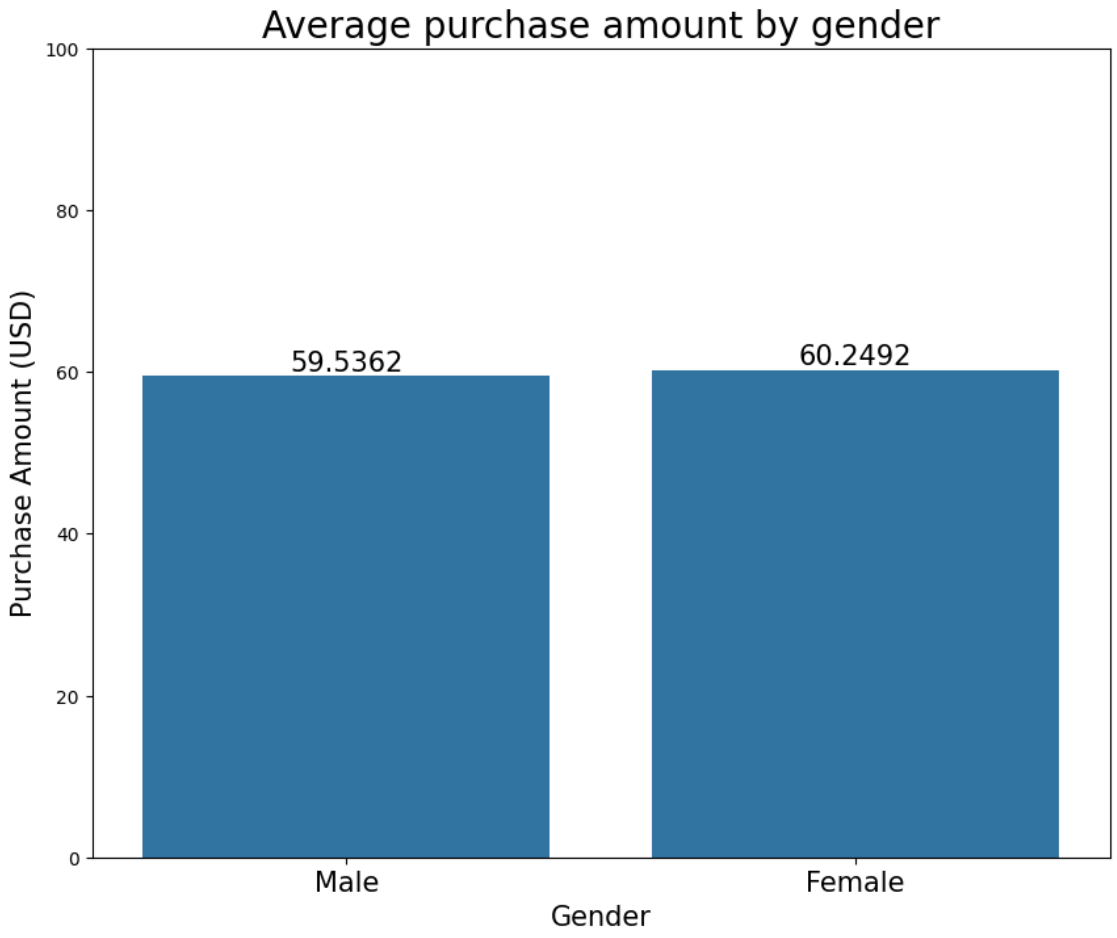
[ ]:
```
# Visualizing the Average Amount of Previous Purchases following Item Purchased
# Creating a bar plot for visualization
plt.figure(figsize=(25,15))
plt.title('Number of Previous Purchases following Item Purchased', fontsize=30)
barplot_item = sns.barplot(data=behaviour_df, x='Previous Purchases', y='Item␣
↪Purchased', errorbar=None);
# Labelling the value of the number of purchases
barplot_item.bar_label(barplot_item.containers[0], fontsize = 15);
# Customize the bar plot
barplot_item.set_xlabel('Previous Purchases', fontsize=15)
```

```
barplot_item.set_ylabel('Item Purchased', fontsize=15)
plt.show();
```



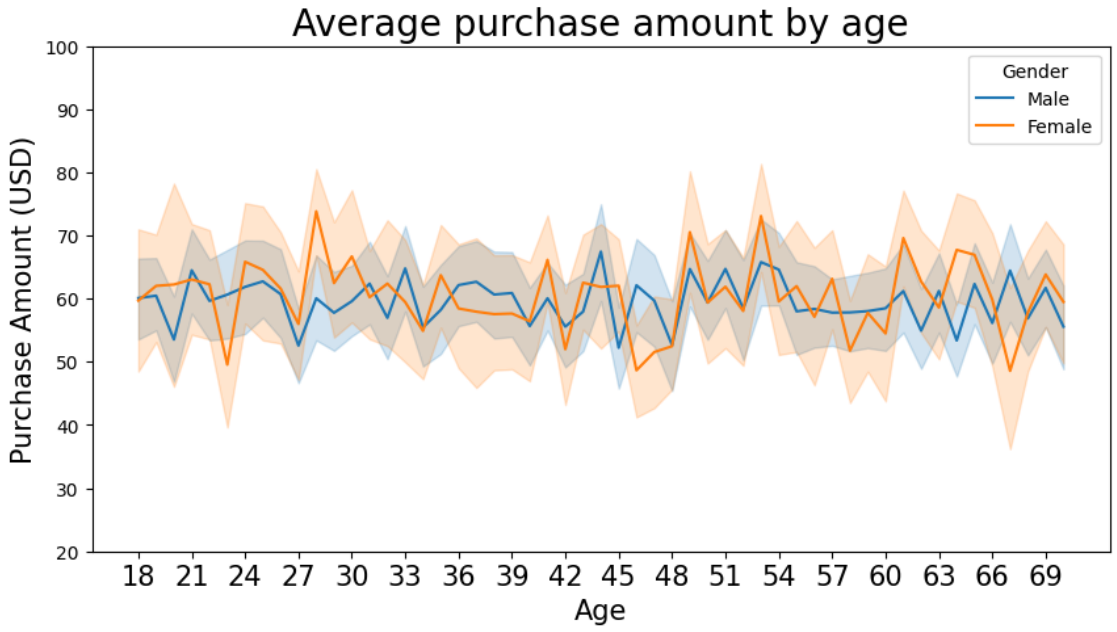
[ ]:
```
plt.figure(figsize=(10,8))
plt.title("Average purchase amount by gender", fontsize = 20)
plt.ylim([0,100])
plot_gender = sns.barplot(x = behaviour_df.Gender, y =behaviour_df['Purchase␣
↪Amount (USD)'], errorbar = None)
plot_gender.bar_label(plot_gender.containers[0], fontsize = 15)
plt.xticks(fontsize = 15)
plt.xlabel("Gender", fontsize = 15)
plt.ylabel("Purchase Amount (USD)", fontsize = 15)
```

[ ]:
```
Text(0, 0.5, 'Purchase Amount (USD)')
```



The number of male customers is more than females, but females spend money slightly more than males as the chart shows.

[ ]:
```
plt.figure(figsize=(10,5))
plt.title("Average purchase amount by age", fontsize = 20)
plt.ylim([20,100])
plot10 = sns.lineplot(x = behaviour_df['Age'], y = behaviour_df['Purchase␣
 ↪Amount (USD)'], hue = behaviour_df['Gender'])
plot10.set_xticks(range(18,70,3))
plt.xticks(fontsize = 15)
plt.xlabel("Age", fontsize = 15)
plt.ylabel("Purchase Amount (USD)", fontsize = 15);
```



[ ]:
```
# Grouping other fields including 'Season','Subscription Status','Discount Applied','Promo Code Used' by Purchase Amount
behaviour_df.groupby(['Season','Subscription Status','Discount Applied','Promo Code Used'])['Purchase Amount (USD)'].agg(['sum', 'count', 'mean']).sort_values(by='sum', ascending=False)
```

[ ]:

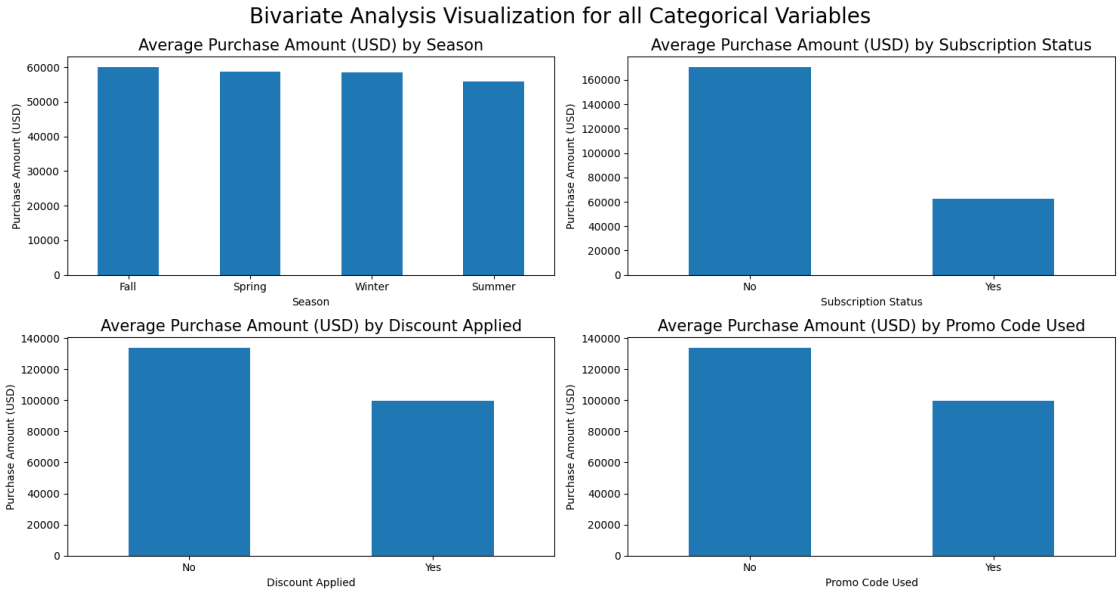
Season	Subscription Status	Discount Applied	Promo Code Used	sum	count	mean
Fall	No	No	No	35566	578	61.53
Winter	No	No	No	33299	554	60.11
Spring	No	No	No	33079	559	59.18
Summer	No	No	No	31726	532	
Fall	Yes	Yes	Yes	16363	264	
Spring	Yes	Yes	Yes	15850	270	
Winter	Yes	Yes	Yes	15379	254	
Summer	Yes	Yes	Yes	15053	265	
Winter	No	Yes	Yes	9929	163	
Spring	No	Yes	Yes	9750	170	
Summer	No	Yes	Yes	8998	158	
Fall	No	Yes	Yes	8089	133	

```
Summer No                  No                  No                  59.64
Fall   Yes                 Yes                 Yes                 61.98
Spring Yes                 Yes                 Yes                 58.70
Winter Yes                 Yes                 Yes                 60.55
Summer Yes                 Yes                 Yes                 56.80
Winter No                  Yes                 Yes                 60.91
Spring No                  Yes                 Yes                 57.35
Summer No                  Yes                 Yes                 56.95
Fall   No                  Yes                 Yes                 60.82
```

[ ]:
```
# Grouping other fields including 'Season','Subscription Status','Discount␣
 ↪Applied','Promo Code Used' by Previous Purchase
behaviour_df.groupby(['Season','Subscription Status','Discount Applied','Promo␣
 ↪Code Used'])['Previous Purchases'].agg(['sum', 'count', 'mean']).
 ↪sort_values(by='sum', ascending=False)
```

[ ]:
```
                                                                    sum  count  \
Season Subscription Status Discount Applied Promo Code Used
Winter No                  No               No                  14405    554
Fall   No                  No               No                  14080    578
Spring No                  No               No                  13793    559
Summer No                  No               No                  13422    532
Spring Yes                 Yes              Yes                  7042    270
Summer Yes                 Yes              Yes                  6951    265
Fall   Yes                 Yes              Yes                  6843    264
Winter Yes                 Yes              Yes                  6631    254
Spring No                  Yes              Yes                  4256    170
Winter No                  Yes              Yes                  4133    163
Summer No                  Yes              Yes                  3892    158
Fall   No                  Yes              Yes                  3423    133

                                                                   mean
Season Subscription Status Discount Applied Promo Code Used
Winter No                  No               No                  26.00
Fall   No                  No               No                  24.36
Spring No                  No               No                  24.67
Summer No                  No               No                  25.23
Spring Yes                 Yes              Yes                 26.08
Summer Yes                 Yes              Yes                 26.23
Fall   Yes                 Yes              Yes                 25.92
Winter Yes                 Yes              Yes                 26.11
Spring No                  Yes              Yes                 25.04
Winter No                  Yes              Yes                 25.36
Summer No                  Yes              Yes                 24.63
Fall   No                  Yes              Yes                 25.74
```

[ ]:
```
fig, axes = plt.subplots(nrows=2, ncols=2, figsize=(15, 8))
fig.suptitle('Bivariate Analysis Visualization for all Categorical Variables',␣
 ↪fontsize=20);
behaviour_df.groupby('Season')['Purchase Amount (USD)'].sum().
 ↪sort_values(ascending=False).plot(kind='bar', ax=axes[0,0]).
 ↪tick_params(axis='x', rotation=0)

axes[0,0].set_title('Average Purchase Amount (USD) by Season',fontsize = 15)
axes[0,0].set_xlabel('Season', fontsize = 10)
axes[0,0].set_ylabel('Purchase Amount (USD)', fontsize = 10)

behaviour_df.groupby('Subscription Status')['Purchase Amount (USD)'].sum().
 ↪sort_values(ascending=False).plot(kind='bar', ax=axes[0,1]).
 ↪tick_params(axis='x', rotation=0)
axes[0,1].set_title('Average Purchase Amount (USD) by Subscription Status',␣
 ↪fontsize = 15)
axes[0,1].set_xlabel('Subscription Status', fontsize = 10)
axes[0,1].set_ylabel('Purchase Amount (USD)', fontsize = 10)

behaviour_df.groupby('Discount Applied')['Purchase Amount (USD)'].sum().
 ↪sort_values(ascending=False).plot(kind='bar', ax=axes[1,0]).
 ↪tick_params(axis='x', rotation=0)
axes[1,0].set_title('Average Purchase Amount (USD) by Discount Applied',␣
 ↪fontsize = 15)
axes[1,0].set_xlabel('Discount Applied', fontsize = 10)
axes[1,0].set_ylabel('Purchase Amount (USD)', fontsize = 10)

behaviour_df.groupby('Promo Code Used')['Purchase Amount (USD)'].sum().
 ↪sort_values(ascending=False).plot(kind='bar', ax=axes[1,1]).
 ↪tick_params(axis='x', rotation=0)
axes[1,1].set_title('Average Purchase Amount (USD) by Promo Code Used',␣
 ↪fontsize = 15)
axes[1,1].set_xlabel('Promo Code Used', fontsize = 10)
axes[1,1].set_ylabel('Purchase Amount (USD)', fontsize =10)

plt.tight_layout();
```



CONCLUSION: > Through EDA, we got useful insights: * The amount of male customers is more than females but the females spent more money than males, according to the data above, most items sold are for females and the purchase amount is even slightly more. * Locations that do not have an impressive amount of consumer counting but spend the most amount of money for items. What factors lead to that happening? The factors can be climate, geographical, economic, cultural styles, etc. * We can tell the yellow and green colours are so appreciated by people who go shopping in the fall. * In a consumer age range that goes around 30 years old and between 50 and 60 years old, the average amount of purchase peaks highest. * Looks like discounts and promo codes are used less than subscriptions because the subscription will keep in touch with customers via email or message.