

# **Customer Shopping Trends**

Niharika Madhadi

Instructor: Mr. Gahangir Hossain

Data Visualization INFO 5709

Term Project

## Introduction

The customer shopping preferences dataset offers insightful information on the behaviour and buying habits of consumers. Companies may customise their goods, marketing plans, and general customer experience by knowing consumer trends and preferences. A wide range of consumer characteristics are included in this dataset, such as age, gender, past purchases, preferred payment methods, and frequency of purchases. By analysing this data, organisations may improve consumer happiness, optimise product offers, and make well-informed decisions. It is a useful tool for businesses looking to match their tactics to the constantly changing demands and tastes of their clients.

This dataset includes a variety of factors related to consumer buying habits, offering crucial data to companies looking to improve their understanding of their clientele. Customer age, gender, purchase quantities, preferred payment methods, frequency of purchases, and feedback ratings are all included in these characteristics. The collection also includes information on the kind of goods purchased, how often people buy, when they like to shop, and if they take advantage of promotions. This 3900-record dataset is a vital resource for companies looking to use data-driven insights to improve decision-making and create customer-focused strategies.

## Dataset

This dataset is collected from Kaggle. It is a synthetic dataset which shows the shopping trends of customers. It has 3900 records and 19 attributes.

Dataset collected from Kaggle:

<https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset>

## Attributes

The dataset has total of 19 attributes. Below is the list of attributes:

- **Customer ID** – It is a unique number given to each customer
- **Age** – It gives age in years
- **Gender** - Gender of the customer (Male/Female)
- **Item Purchased** – It gives product purchased by customers
- **Category** – It gives the category to which item belongs to
- **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** - The total count of transactions concluded by the customer at the store, excluding the ongoing transaction
- **Payment Method** - Customer's most preferred payment method
- **Frequency of Purchases** - Frequency at which the customer makes purchases (e.g., Weekly, Fortnightly, Monthly)

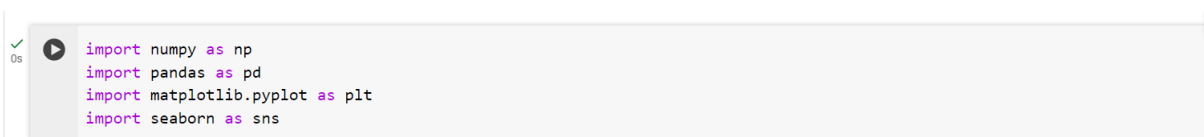
## Tools

Python- For performing the EDA (Exploratory Data Analysis) on the dataset

Tableau- To visualize the dataset for given hypothesis.

## Exploratory Data Analysis

Import all the required packages for performing EDA



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

Load dataset into data frame and get its shape (number of rows and columns), info (gives the datatype and non-null count of each column) and head (gives the first 5 rows of the dataset).

Our dataset has 3900 entries and 19 columns. It has one float64 data type, four int64 data type and 14 object data type.

```
[2] shopping_df=pd.read_csv("/content/shopping_trends.csv")
print(shopping_df.shape)
print(shopping_df.info())
shopping_df.head()
```

```
(3900, 19)
<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 (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                       3900 non-null   float64
11  Subscription Status                 3900 non-null   object
12  Payment Method                     3900 non-null   object
13  Shipping Type                      3900 non-null   object
14  Discount Applied                   3900 non-null   object
15  Promo Code Used                    3900 non-null   object
16  Previous Purchases                 3900 non-null   int64
17  Preferred Payment Method           3900 non-null   object
18  Frequency of Purchases             3900 non-null   object
dtypes: float64(1), int64(4), object(14)
memory usage: 579.0+ KB
None
```

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

Used describe method to get mean, minimum, maximum and quartile values of all numerical columns.

Age column has mean of 44 years, the youngest one has 18 years of age and the oldest one has 70 years of age.

For Purchase Amount the average amount is 59.7 USD, the least amount of purchase is 20 USD, and the highest amount of purchase is 100 USD.

The average review rating is 3.7, the least rating is 2.5 and highest rating is 5.0

```
[3] shopping_df.describe()
```

	Customer ID	Age	Purchase Amount (USD)	Review Rating	Previous Purchases
count	3900.000000	3900.000000	3900.000000	3900.000000	3900.000000
mean	1950.500000	44.068462	59.764359	3.749949	25.351538
std	1125.977353	15.207589	23.685392	0.716223	14.447125
min	1.000000	18.000000	20.000000	2.500000	1.000000
25%	975.750000	31.000000	39.000000	3.100000	13.000000
50%	1950.500000	44.000000	60.000000	3.700000	25.000000
75%	2925.250000	57.000000	81.000000	4.400000	38.000000
max	3900.000000	70.000000	100.000000	5.000000	50.000000

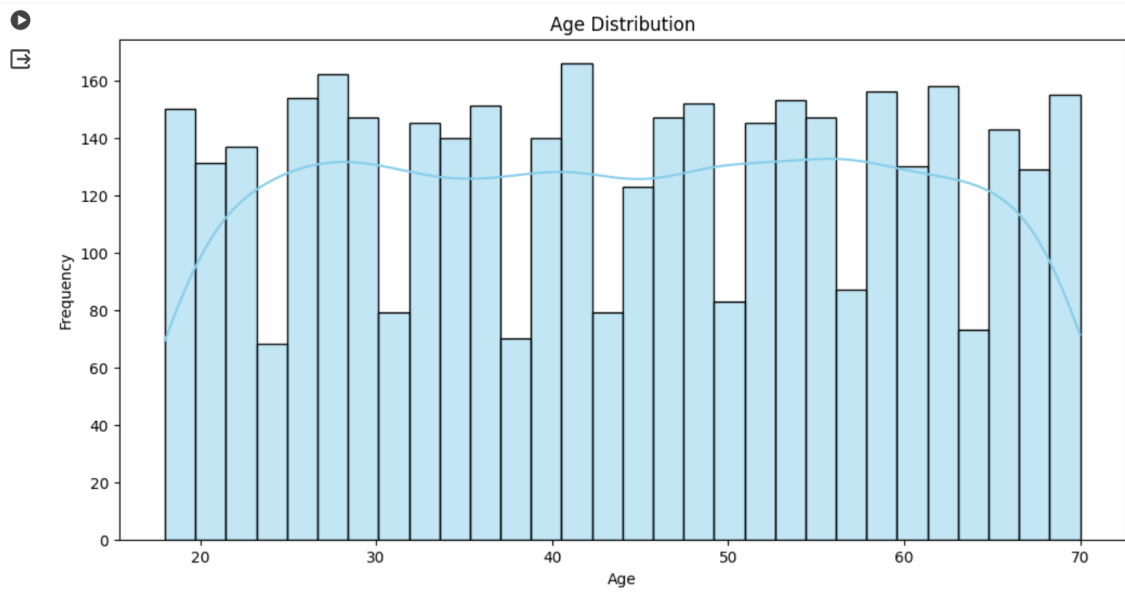
I have checked if the dataset has any null values using the below code. We could see that there are no null values present.

```
shopping_df.isnull().sum()
```

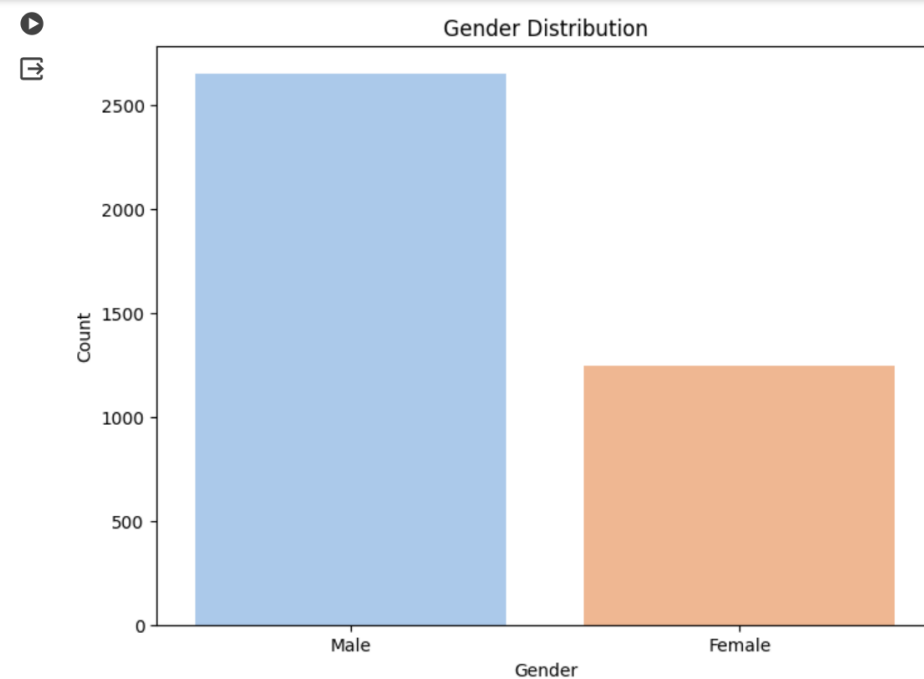
```
Customer ID      0
Age              0
Gender           0
Item Purchased   0
Category         0
Purchase Amount (USD)  0
Location         0
Size            0
Color           0
Season          0
Review Rating    0
Subscription Status  0
Payment Method   0
Shipping Type    0
Discount Applied 0
Promo Code Used  0
Previous Purchases 0
Preferred Payment Method 0
Frequency of Purchases 0
dtype: int64
```

1. What is the distribution of age, gender and location among customers?

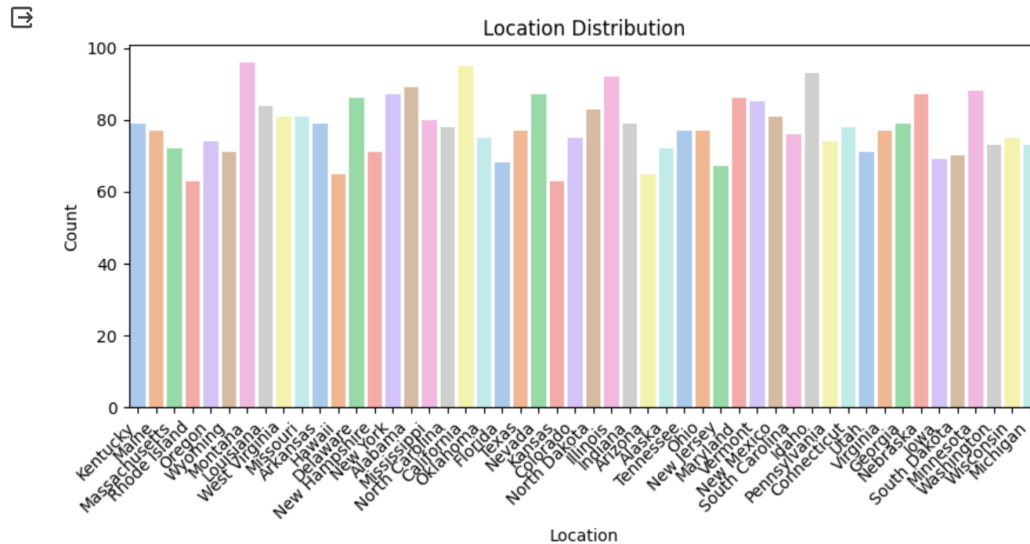
```
# Analyze Age Distribution
plt.figure(figsize=(12, 6))
sns.histplot(shopping_df['Age'], bins=30, kde=True, color='skyblue')
plt.title('Age Distribution')
plt.xlabel('Age')
plt.ylabel('Frequency')
plt.show()
```



```
[ ] # Analyze Gender Distribution
plt.figure(figsize=(8, 6))
sns.countplot(x='Gender', data=shopping_df, palette='pastel')
plt.title('Gender Distribution')
plt.xlabel('Gender')
plt.ylabel('Count')
plt.show()
```



```
# Analyze Location Distribution
plt.figure(figsize=(12, 8))
sns.countplot(x='Location', data=shopping_df, palette='pastel')
plt.title('Location Distribution')
plt.xlabel('Location')
plt.ylabel('Count')
plt.xticks(rotation=45, ha='right') # Rotate x-axis labels for better visibility
plt.show()
```



Result:

We have seen three visualizations for distribution of age, gender and location. From age distribution graph we can see that the frequency is normally distributed and could see from 25-30 years of age the amount of people who buy are little more than others. From gender distribution we could see that Male's count is higher than Female. From location distribution it is mostly similarly distributed having Louisiana with greater count.

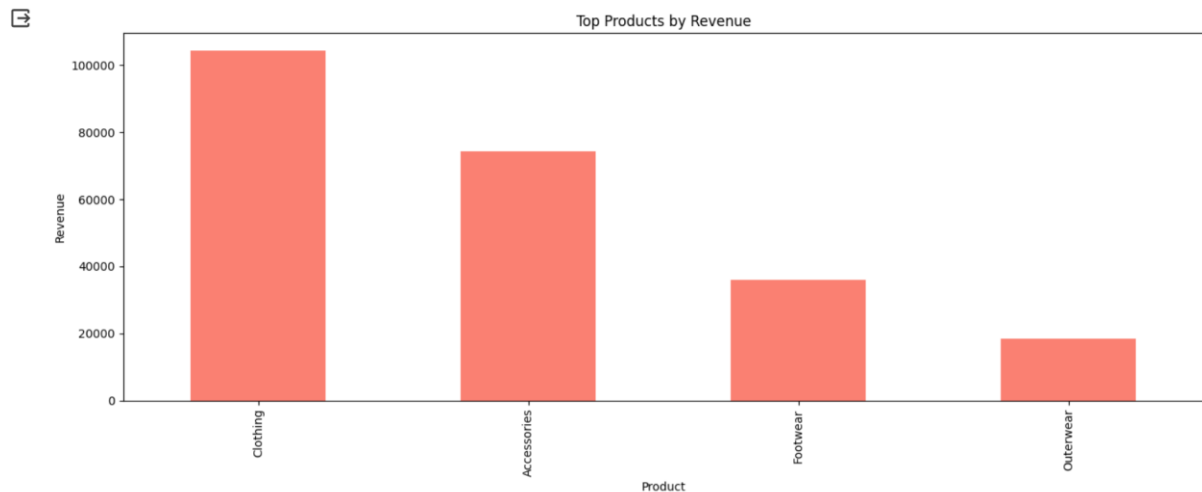
## 2. What are most popular categories in terms of revenue?

```
# Top Categories by Revenue
top_products_by_revenue = shopping_df.groupby('Category')['Purchase Amount (USD)'].sum().sort_values(ascending=False).head(10)

# Plotting
plt.figure(figsize=(14, 6))

# Bar plot for Revenue
top_products_by_revenue.plot(kind='bar', color='salmon')
plt.title('Top Products by Revenue')
plt.xlabel('Product')
plt.ylabel('Revenue')

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



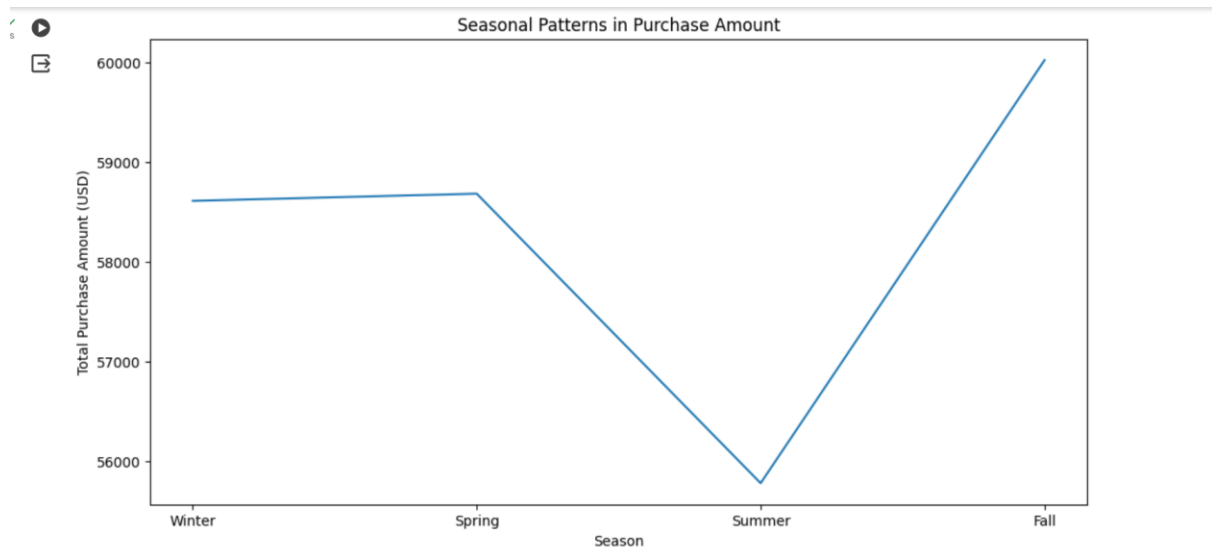
Result:

We can see that Clothing category has highest revenue and Outerwear has lowest revenue generated. We could conclude that clothing and accessories are most bought categories by customers.

### 3. Are there any seasonal patterns in purchasing behaviour?

```
[9] # Analyze Seasonal Patterns
plt.figure(figsize=(12, 6))
sns.lineplot(x='Season', y='Purchase Amount (USD)', data=shopping_df, estimator='sum', errorbar=None)
plt.title('Seasonal Patterns in Purchase Amount')
plt.xlabel('Season')
plt.ylabel('Total Purchase Amount (USD)')
plt.show()
```



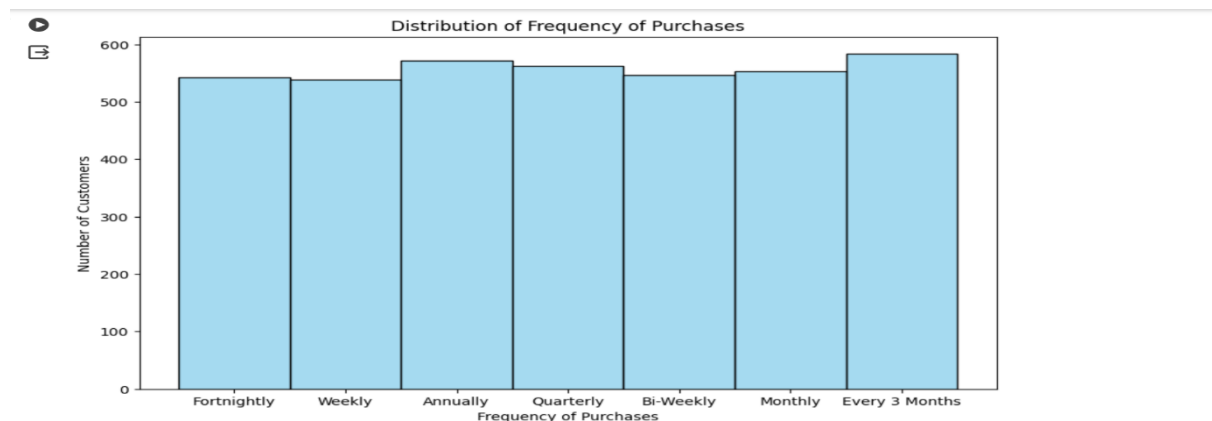


Result:

From above visualization, we can see that Fall season has higher purchase done than others. Winter and Spring season have almost similar purchases and Summer has the lower purchases done.

#### 4. How frequently do customers make purchases?

```
[10] # Analyze Frequency of Purchases
plt.figure(figsize=(10, 6))
sns.histplot(shopping_df['Frequency of Purchases'], bins=20, kde=False, color='skyblue')
plt.title('Distribution of Frequency of Purchases')
plt.xlabel('Frequency of Purchases')
plt.ylabel('Number of Customers')
plt.show()
```

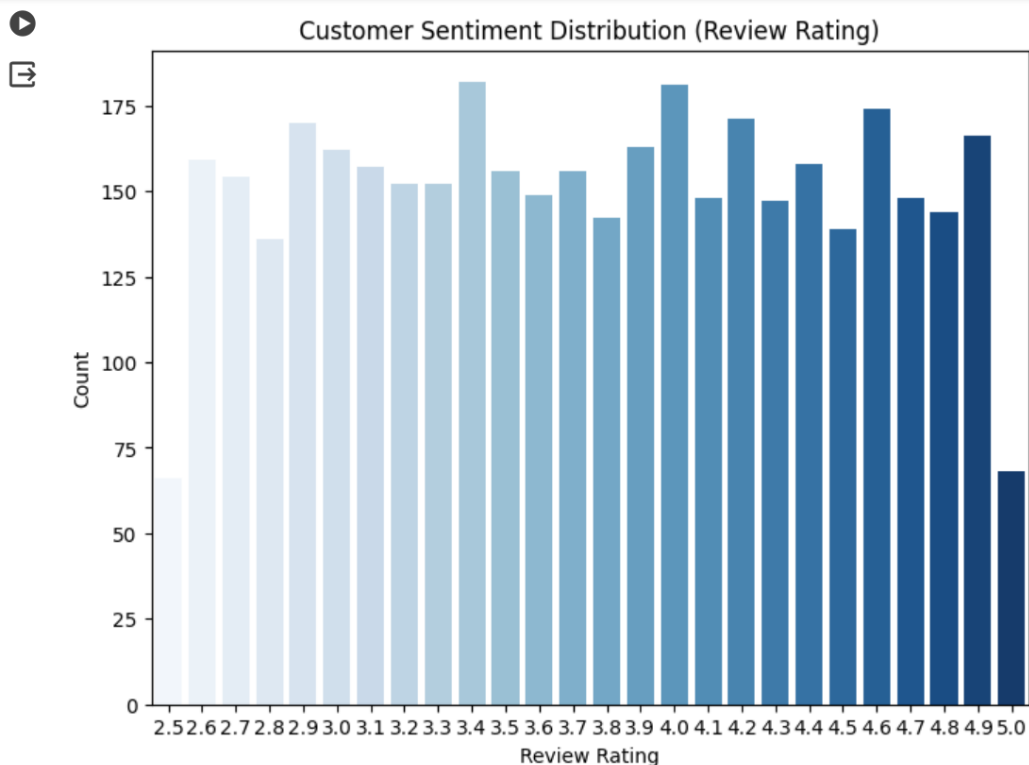


Result: From above histogram the distribution of frequency is similar for most of them. But Every 3 months is the highest among them where customer would make purchases again. Annually also the purchases are made frequently.

## 5. How does customer sentiment correlate with sales?

```
[1] # Bar plot: Customer Sentiment Distribution (Review Rating)
plt.figure(figsize=(10, 6))
sns.countplot(x='Review Rating', data=shopping_df, palette='Blues')
plt.title('Customer Sentiment Distribution (Review Rating)')
plt.xlabel('Review Rating')
plt.ylabel('Count')
plt.show()

# Correlation analysis: Customer Sentiment and Purchase Amount
correlation_purchase = shopping_df['Review Rating'].corr(shopping_df['Purchase Amount (USD)'])
print(f'Correlation between Customer Sentiment and Purchase Amount: {correlation_purchase}')
```



Correlation between Customer Sentiment and Purchase Amount: 0.030775923073914465

### Result:

From above code we could see the correlation between reviews and revenue is positive but very less which means purchase amount increases if they have higher review rating, but the dependency is less. The frequency distribution of reviews are plotted and could see higher count is for 3.4 and 4.0 rating.

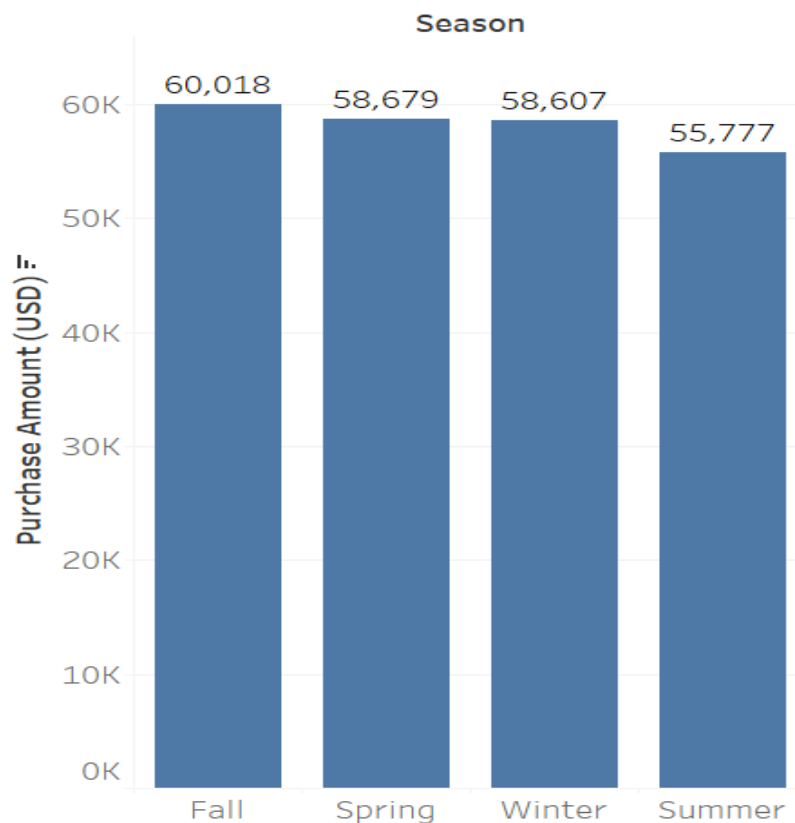
## Hypothesis:

### 1. Seasonal Impact on Purchase behaviour:

To check if there is any significant difference in customer purchasing behaviour across different seasons. I have used three visualizations to check the relationship between seasons and purchase amount.

Bar Chart: In this Season is given in columns and Purchase Amount is given in rows. Then it is sorted in descending order and labelled with purchase amount. So, we can see the sales in each season.

### Seasonal Impact on Purchase Behavior

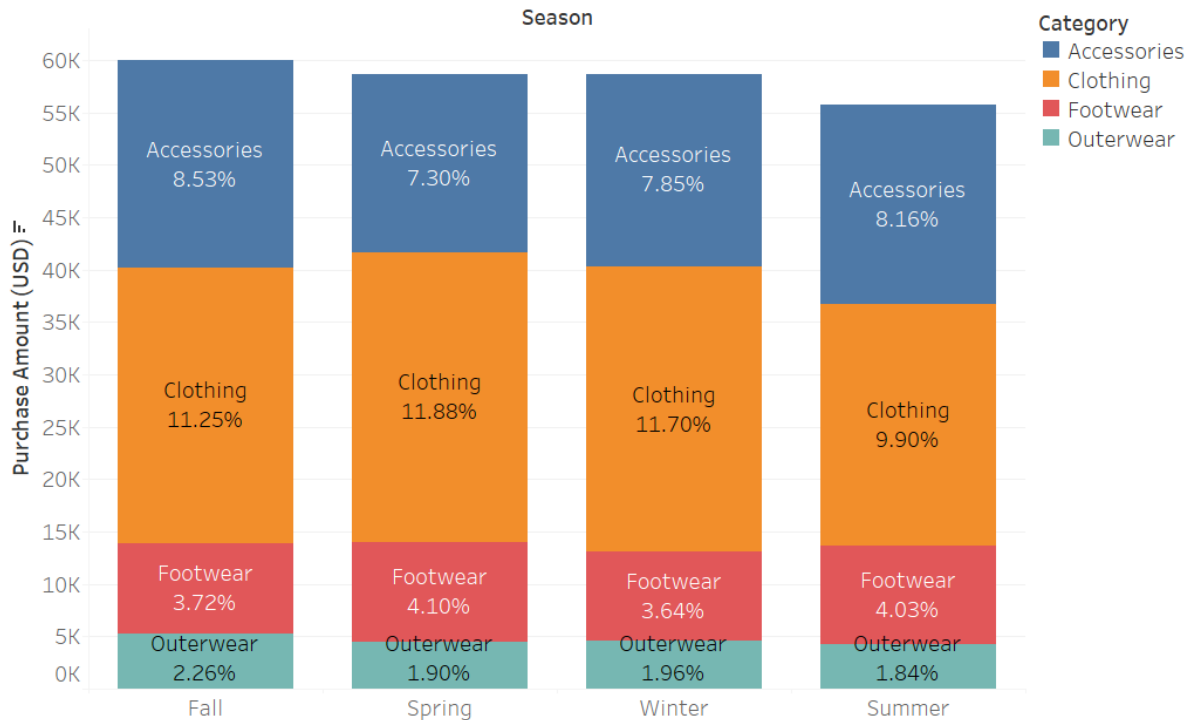


Sum of Purchase Amount (USD) for each Season. The marks are labeled by sum of Purchase Amount (USD).

Result: Fall season has most sales done which is 60,018 USD. Summer has the least sales done which is 55,777 USD. We can see a difference in sales based on season, hence the dataset is having a seasonal impact on purchase behaviour.

Stacked bar chart: In this Season is given in columns and Purchase Amount is given in rows and sorted in descending order. Category is given in Color marks and in label so that in each season we can get breakdown of each category. The percent of total calculation is done for Purchase amount and given in label.

Stacked bar chart

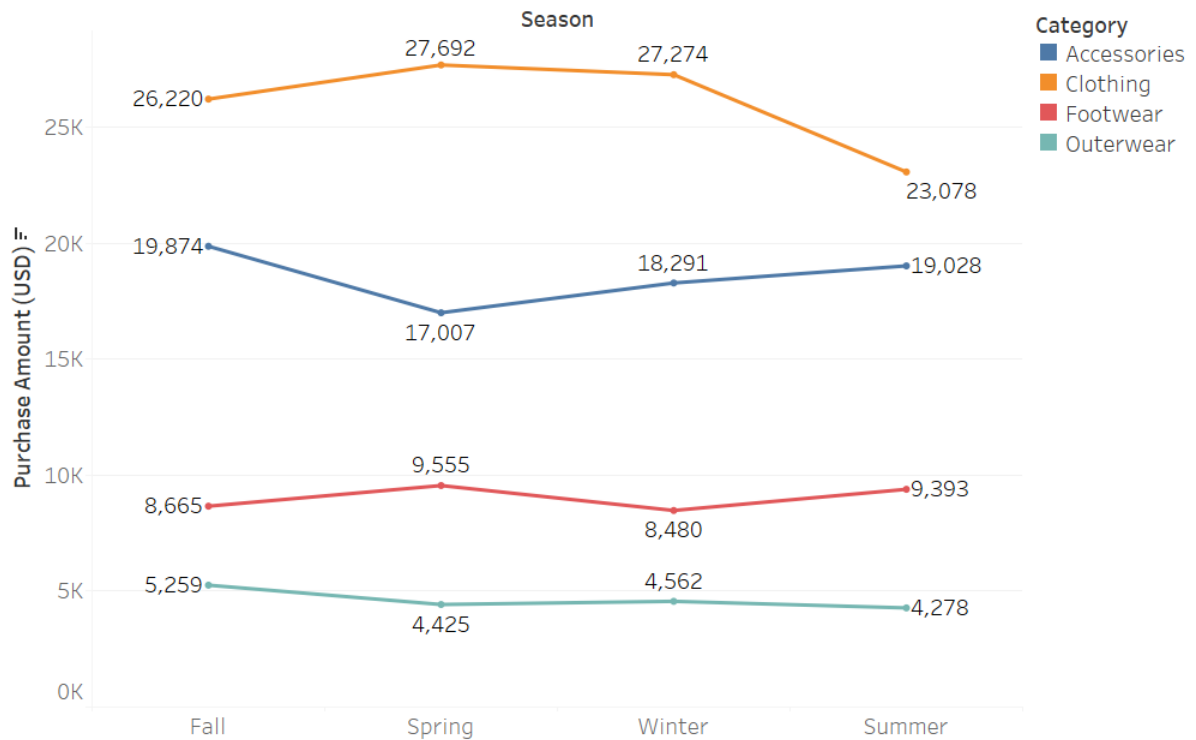


Sum of Purchase Amount (USD) for each Season. Color shows details about Category. The marks are labeled by Category and % of Total Purchase Amount (USD). The view is filtered on Category, which keeps Accessories, Clothing, Footwear and Outerwear.

Result: In all seasons we can see Clothing category has highest percent of total and Outerwear has least percent of total. Comparing with fall season, the accessories and outerwear percent is less for next two seasons, so to improve these they can market and promote those products in next seasons as well. Summer has least percent in terms of all categories, so it is important for business to improve their sales in summer season.

Line Chart: In this Season is given in columns and Purchase Amount is given in rows and the purchase amount is labelled. I have given Category in color marks so that we get separate line charts for each category.

Line chart



The trend of sum of Purchase Amount (USD) for Season. Color shows details about Category. The marks are labeled by sum of Purchase Amount (USD).

Result: For clothing highest sales is done in Spring season which is 27,692 USD. For accessories highest sales is done in Fall season which is 19,874 USD. For footwear highest sales is done in Spring season which is 9,555 USD and for outerwear highest sales is done in Fall season which is 5,259 USD.

### Conclusion:

The proposed hypothesis is to be accepted as we can see each category has varying sales in each season. Thus, the seasonal impact on purchasing behaviour is there and business could tailor their marketing strategies, product offerings and promotions to align with seasonal preferences.

## 2. Effect of shipping type on purchase amount

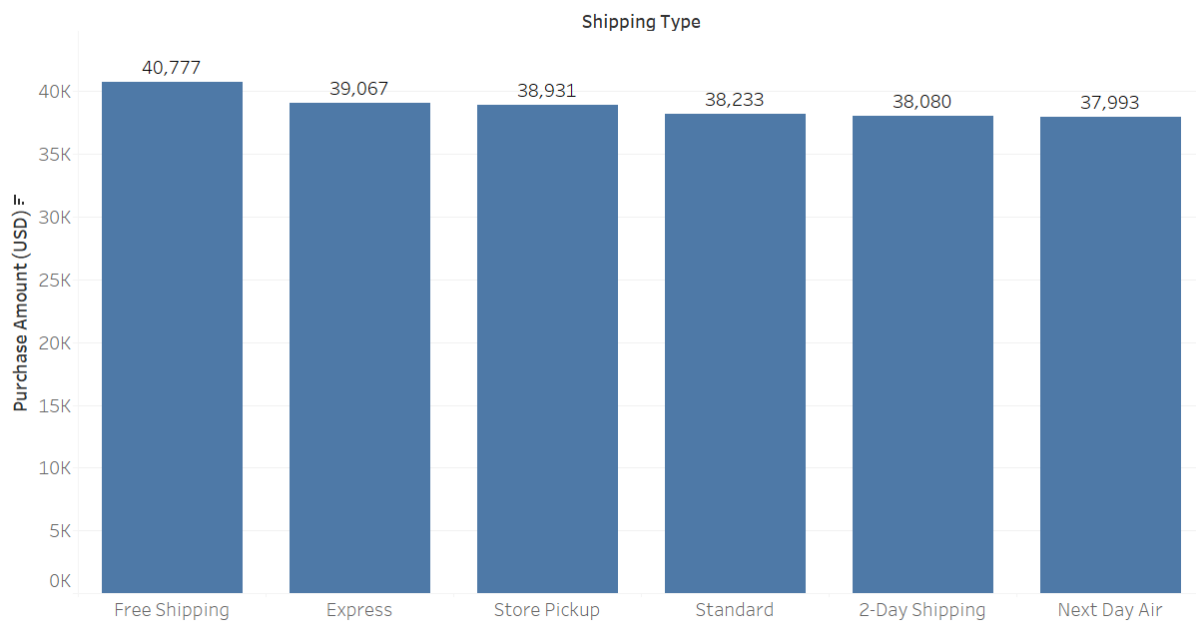
To check if there is any significant difference in customer purchasing behaviour for different shipping types. I have used three visualizations to check the relationship between shipping type and purchase amount.

Bar Chart:

A bar chart can show the average purchase amount for each shipping type, providing a straightforward comparison.

In this Shipping type is given in columns and Purchase Amount is given in rows. Then it is sorted in descending order and labelled with purchase amount. So, we can see the sales in each shipping type.

Bar chart



Sum of Purchase Amount (USD) for each Shipping Type. The marks are labeled by sum of Purchase Amount (USD).

Result: Free shipping has the highest amount of sales which is 40,777 USD. Next day air has the lowest amount of sales which is 37,993 USD. Other types have little difference in their sales. There is no greater difference in sales for different types of shipping, they only differ in few hundreds.

Scatter plot:

A scatter plot can reveal individual purchase amounts for each transaction, and adding trend lines can help identify any patterns or correlations.

Previous purchases are given in columns and purchase amount is given in rows. Shipping type is given in Color marks. Trend lines are also given for each type.

## Scatter plot



Sum of Previous Purchases vs. sum of Purchase Amount (USD). Color shows details about Shipping Type.

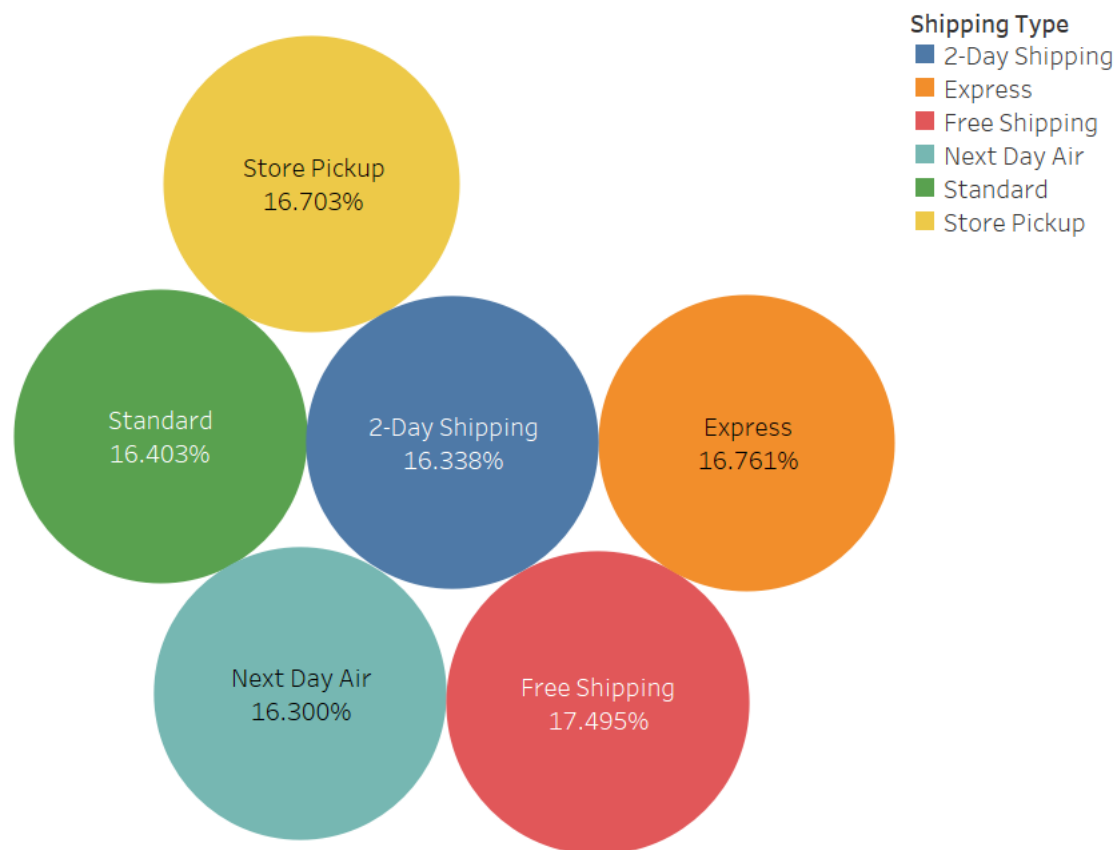
Result: As the difference in sales for different shipping type is not higher, we could see they are all scattered to similar points. Free shipping has highest purchase amount and standard shipping has highest previous purchases. We can conclude that free shipping and standard shipping are the most preferred among others.

Bubble chart:

A bubble chart can be used to compare the average purchase amount and the frequency of purchases for each shipping type. The size of each bubble represents the frequency.

Shipping type is placed in color mark. Purchase amount is given to size mark. Shipping type and percent of total calculation of purchase amount is given in labels.

Bubble chart



Shipping Type and % of Total Purchase Amount (USD). Color shows details about Shipping Type. Size shows sum of Purchase Amount (USD). The marks are labeled by Shipping Type and % of Total Purchase Amount (USD).

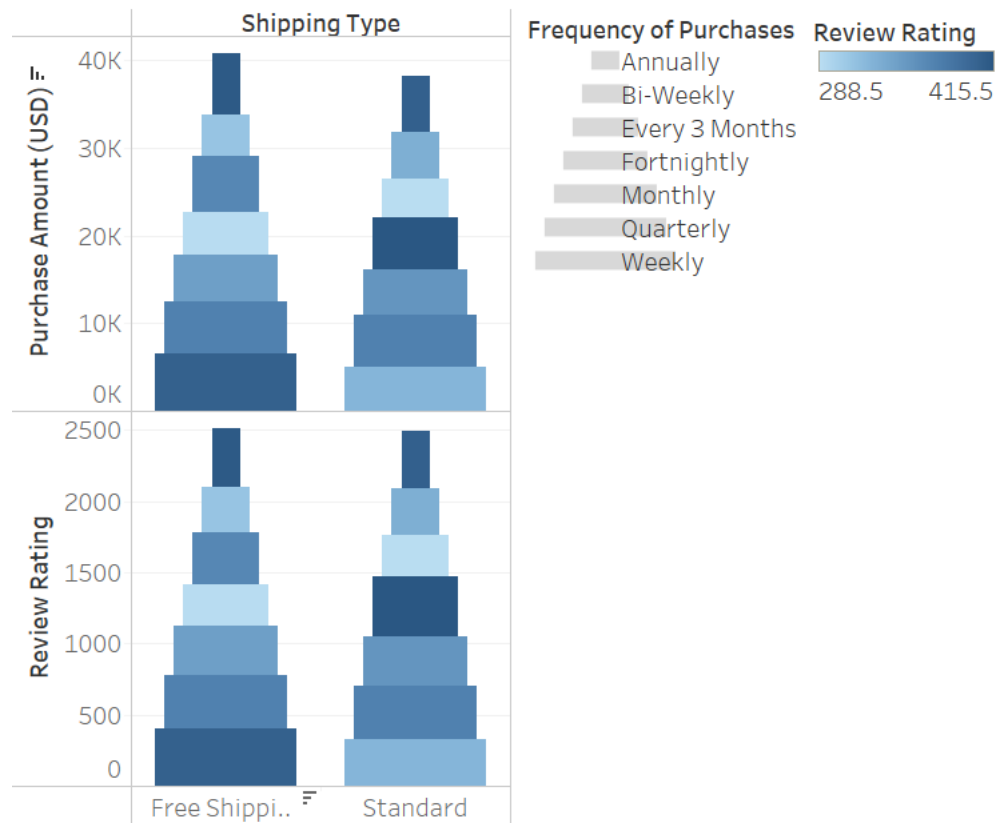
Result: As the difference is less all the bubbles are of similar size, however free shipping has highest percent of total among others which is 17.49% and next day air has least percentage which is 16.3%.



Parallel coordinates:

I have placed Shipping type in columns and Purchase amount and Review rating to rows. Review rating is given to Color marks and Frequency of purchase is given to Size mark, then shipping type is given in filters where I have kept only Free shipping and Standard shipping which are the highest used among others.

## Parallel coordinates



Sum of Purchase Amount (USD) and sum of Review Rating for each Shipping Type. Color shows sum of Review Rating. Size shows details about Frequency of Purchases. The view is filtered on Shipping Type, which keeps Free Shipping and Standard.

Conclusion:

The proposed hypothesis can be rejected as we can see there is not much significant effect on sales based on shipping type. There are free shipping and standard shipping which are preferable over others but that does not effect the sales amount done in the business. So the business can continue to maintain same prices or process for different shipping type.

### 3. Size and color impact on Purchase amount

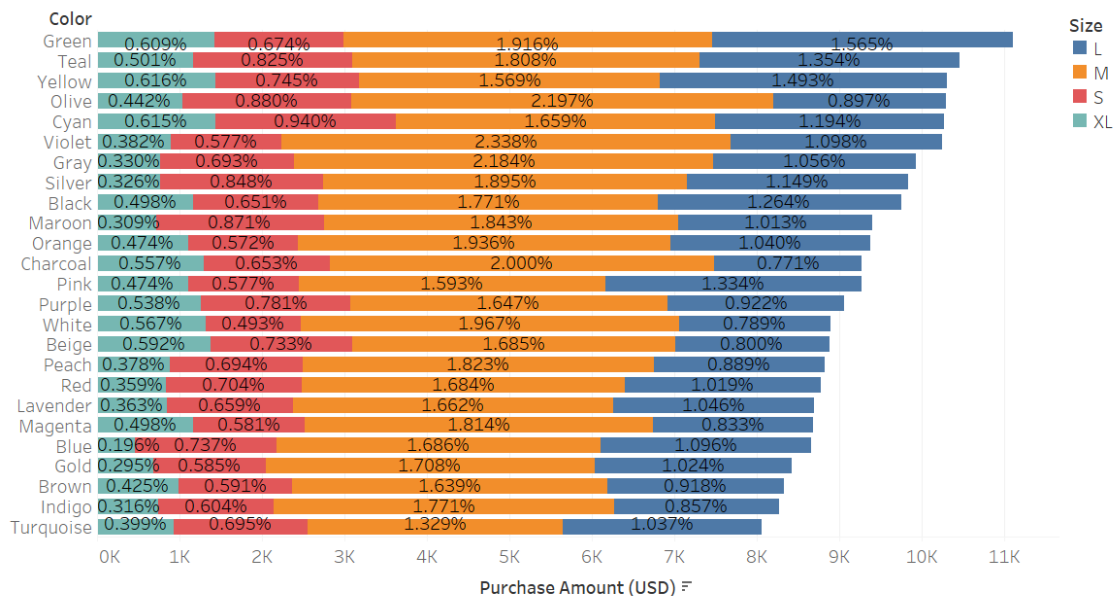
To check if there are any significant preferences in Size and Color that effects the sales. I have used three visualizations to check the relationship between size, color and purchase amount.

Grouped bar chart:

Visualize the distribution of color and size preferences on purchase amount using a grouped bar chart.

I have given purchase amount in columns and color in rows. Size is given in color marks. The purchase amount is calculated in terms of percent of total and labelled.

Grouped bar chart



Sum of Purchase Amount (USD) for each Color. Color shows details about Size. The marks are labeled by % of Total Purchase Amount (USD). The view is filtered on Size, which keeps L, M, S and XL.

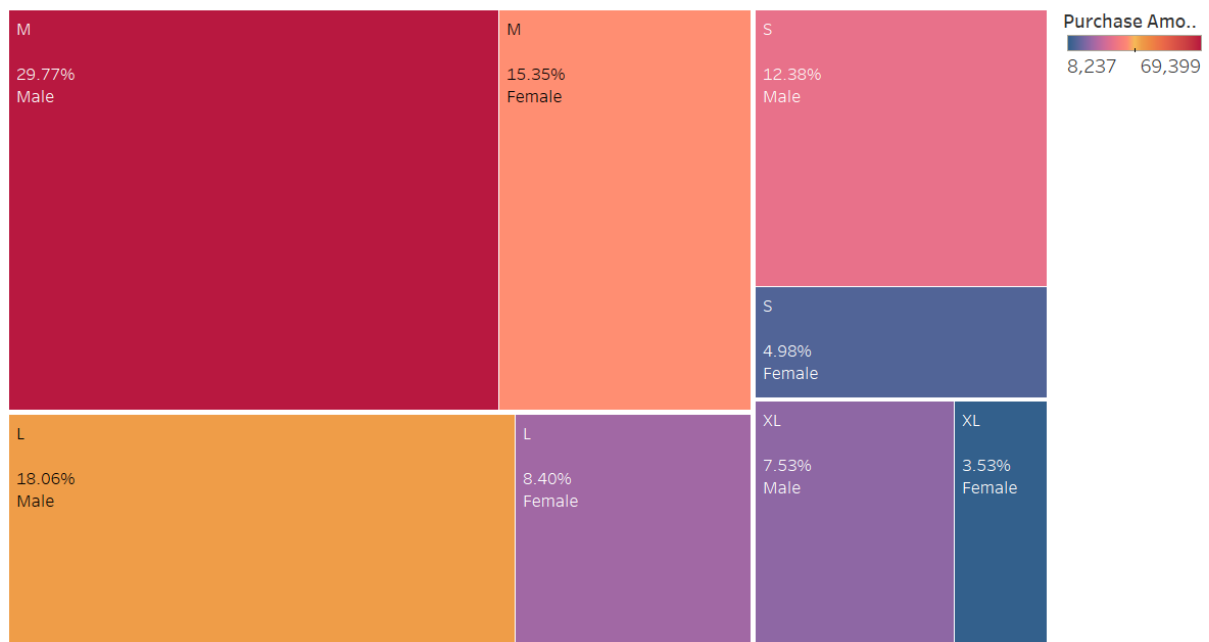
Result: Green color has highest amount of sales done and Turquoise has least amount of sales done. Within each colour's sales we have categorized it on size, and we could see L and M are most bought sizes.

## Tree Map:

Visualize the proportion of color preferences on gender and purchase amount using a tree map.

Purchase amount is given in color, size and label marks. Size and gender are given in label marks.

Tree map



Size, % of Total Purchase Amount (USD) and Gender. Color shows sum of Purchase Amount (USD). Size shows sum of Purchase Amount (USD). The marks are labeled by Size, % of Total Purchase Amount (USD) and Gender.

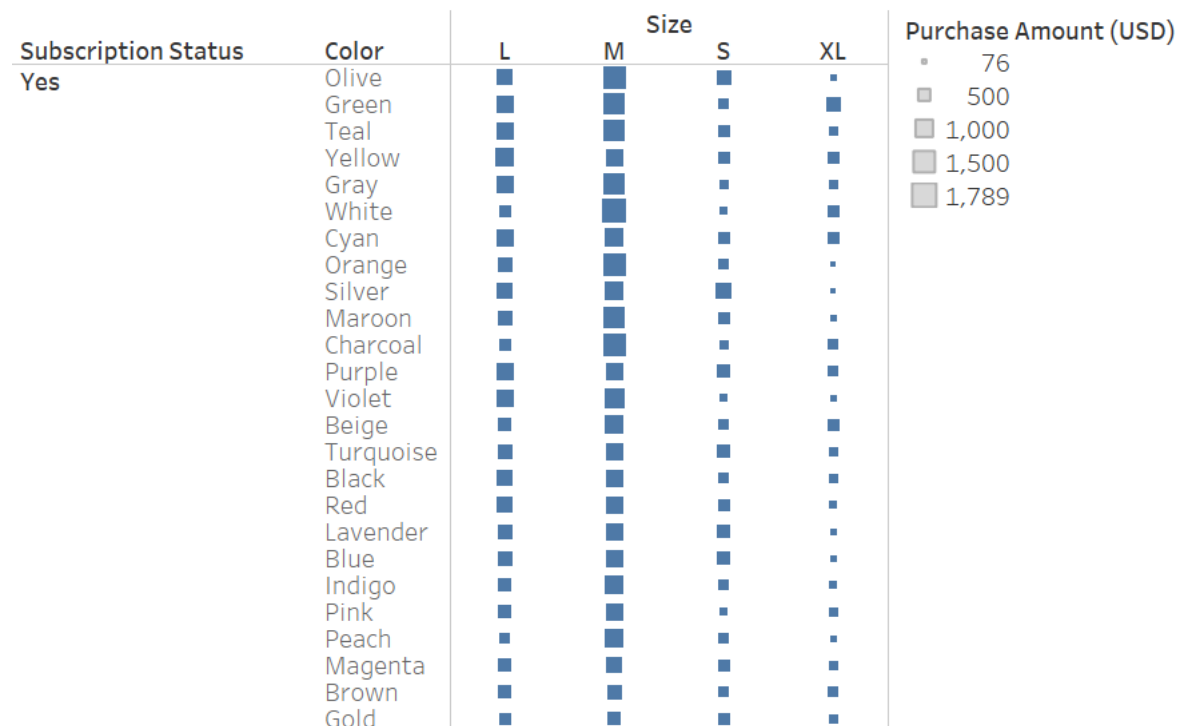
Result: Among males, size M has highest percent of sales which is 29.77% and L has percentage of 18.06%. Among females, size M has highest percent of sales which is 15.35% and L has percentage of 8.40%. In both gender we can see size M has done highest amount of sales done followed by L size.

Heat map:

Explore the relationships between color and size preferences using a heat map.

Size is given in columns. Color and subscription status is given in rows, where subscription status is filtered with Yes. It is sorted in descending order by sum of purchase amount within each color. Purchase amount is given in size mark.

## Heat map



Sum of Purchase Amount (USD) (size) broken down by Size vs. Subscription Status and Color. The view is filtered on Subscription Status, which keeps Yes.

Result: For all the consumers who have subscribed, the purchase amount least is 76 USD and highest is 1,789 USD. We can see M size has highest sales and XL has least amount of sales done.

Conclusion:

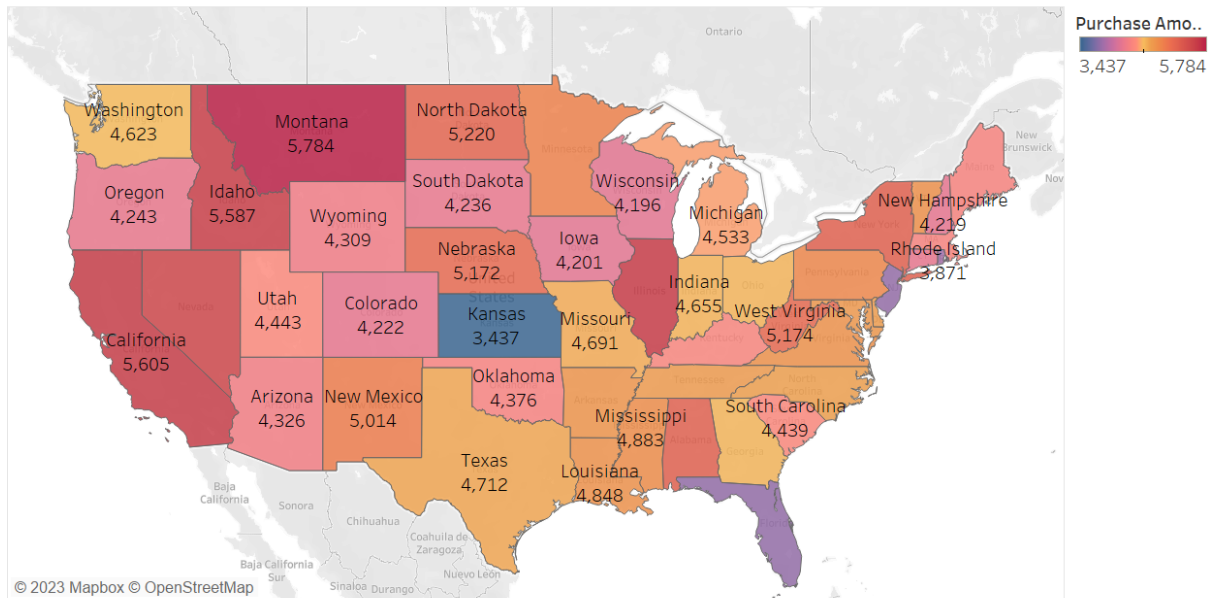
The proposed hypothesis is to be accepted as we can see each size and color has varying sales. Thus, the size and color impact on purchasing behaviour is there and business could tailor their marketing strategies, product offerings and promotions to align with those preferences like the production size has to be increased for the preferred sizes and colours like M size and shades of green color need to be produced more.

## Conclusion

This project analysed the various factors that influence the customer shopping trends. These factors include season, size, color, age, shipping type, purchase amount. This project will help organisations to improve consumer happiness, optimise product offers, and make well-informed decisions.

The map that shows which region has highest sales is given below:

Map

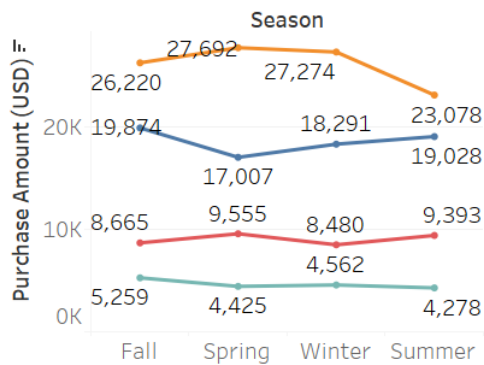


Map based on Longitude (generated) and Latitude (generated). Color shows sum of Purchase Amount (USD). The marks are labeled by Location and sum of Purchase Amount (USD). Details are shown for Location.

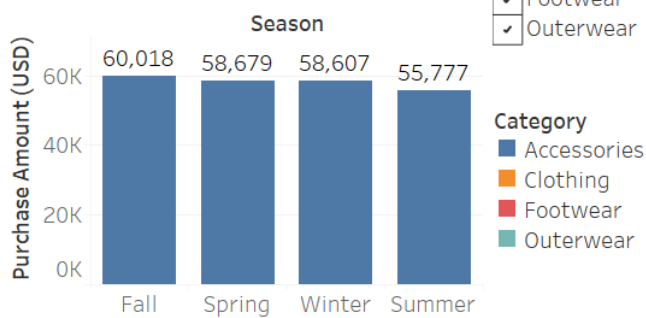
The dashboards of three hypotheses are given below:

## Seasonal Impact on Purchase Behavior

Line chart



Seasonal Impact on Purchase Behavior

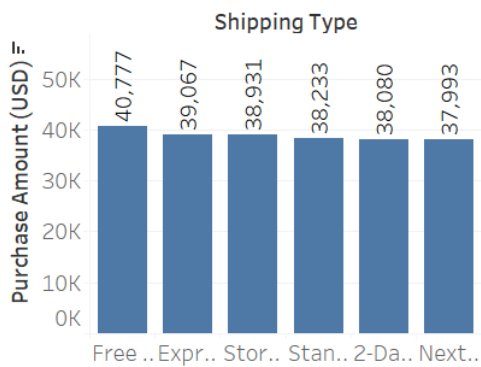


Stacked bar chart

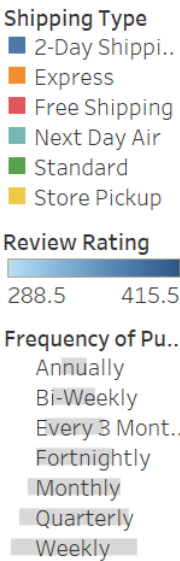
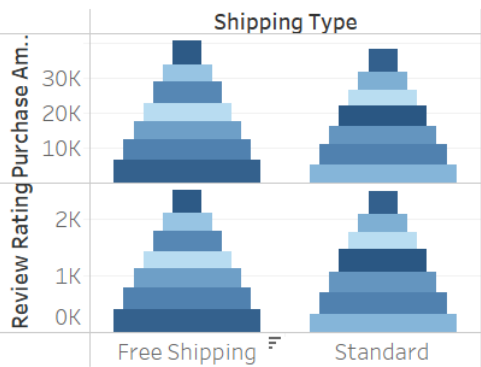


# Effect of shipping type on Sales

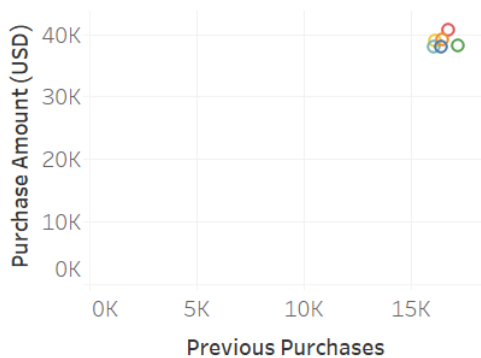
Bar chart



Parallel coordinates



Scatter plot

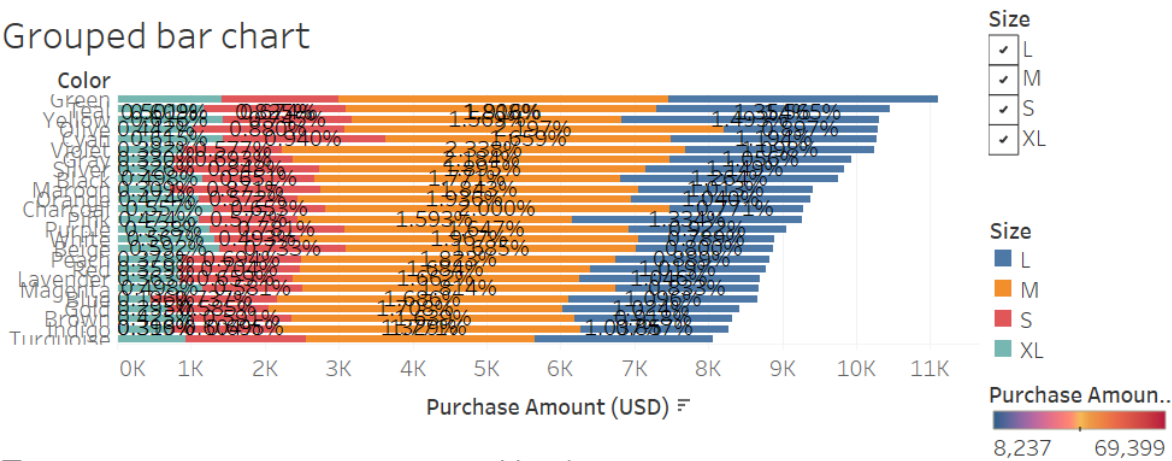


Bubble chart

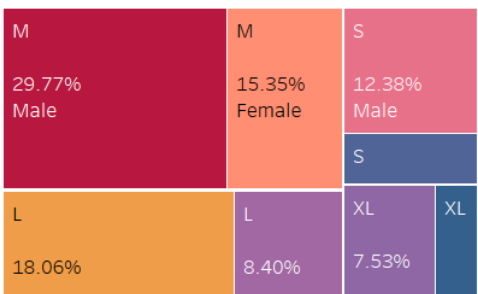


# Size and Color impact on Purchase Amount

Grouped bar chart



Tree map



Heat map



## References

Dataset from Kaggle: <https://www.kaggle.com/datasets/iamsouravbanerjee/customer-shopping-trends-dataset>



