# Project Overview

Using data from 3,900 customer transactions, this project examines how people shop across different product categories. It highlights key patterns in spending, customer groups, product interests, and subscription usage to help drive better business strategies.

# Dataset Summary

- **Data Size:** 3,900 rows × 18 columns
- **Includes:**
  - Customer details (Age, Gender, Location, Subscription Status)
  - Transaction info (Items, Categories, Amount, Season, Size, Color)
  - Shopping behavior (Discounts, Promo Codes, Past Purchases, Frequency, Ratings, Shipping)
- **Missing Values:** 37 entries missing in the Review Rating field

# Python-Based Exploratory Data Analysis

We started the workflow with data preparation and cleaning in Python:

- **Data Loading:** The dataset was imported using *pandas*.
- **Initial Exploration:** Performed `df.info()` and basic summary checks to understand the dataset's structure and key statistics

| 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 55 | Male | Blouse | Clothing | 53 | Kentucky | L | Gray | Winter | 3.1 | Yes | Express | Yes | Yes | 14 | Venmo |
| 2 | 19 | Male | Sweater | Clothing | 64 | Maine | L | Maroon | Winter | 3.1 | Yes | Express | Yes | Yes | 2 | Cash |
| 3 | 50 | Male | Jeans | Clothing | 73 | Massachusetts | S | Maroon | Spring | 3.1 | Yes | Free Shipping | Yes | Yes | 23 | Credit Card |
| 4 | 21 | Male | Sandals | Footwear | 90 | Rhode Island | M | Maroon | Spring | 3.5 | Yes | Next Day Air | Yes | Yes | 49 | PayPal |
| 5 | 45 | Male | Blouse | Clothing | 49 | Oregon | M | Turquoise | Spring | 2.7 | Yes | Free Shipping | Yes | Yes | 31 | PayPal |

[9]:
```python
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
```

- Handled missing values by assigning *Review Rating* values using the median per product category.

```
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
```

```
customer_id            0
age                    0
gender                 0
item_purchased         0
category               0
purchase_amount        0
location               0
size                   0
color                  0
season                 0
review_rating          0
subscription_status    0
shipping_type          0
discount_applied       0
promo_code             0
previous_purchases     0
payment_method         0
buy_freq               0
dtype: int64
```

```python
# filling missing values with median but grouping by category for more accurate results

df['review_rating']=df.groupby(['category'])['review_rating'].transform(lambda a:a.fillna(a.median()))
```

- standardized column names to snake_case.

```python
# Change the column names to snake_case so they're easier to read and document

df.columns=df.columns.str.lower().str.replace(' ','_')
df=df.rename(columns={'purchase_amount_(usd)':'purchase_amount','promo_code_used':'promo_code','frequency_of_purchases':'buy_freq'})
```

- Added new columns : age_group and numeric buy_frequency.

```python
# Created a new column called age_group to categorise different age of shoppers

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

```python
df[['age','age_group']]
```

| | age | age_group |
|---|---|---|
| **0** | 55 | Middle Aged |
| **1** | 19 | Young |
| **2** | 50 | Middle Aged |
| **3** | 21 | Young |
| **4** | 45 | Middle Aged |

```python
# Converted column buy_freq/ frequency of purchase into numeric for better analysis

df['buy_freq']=df['buy_freq'].map({'Fortnightly': 14,'Weekly': 7,'Monthly': 30,'Quarterly': 90,'Bi-Weekly': 14,'Annually': 365,'Every 3 Months': 90})
```

```python
df.head()
```

| on | size | color | season | review_rating | subscription_status | shipping_type | discount_applied | promo_code | previous_purchases | payment_method | buy_freq | age_group |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ky | L | Gray | Winter | 3.1 | Yes | Express | Yes | Yes | 14 | Venmo | 14 | Middle Aged |
| ne | L | Maroon | Winter | 3.1 | Yes | Express | Yes | Yes | 2 | Cash | 14 | Young |
| tts | S | Maroon | Spring | 3.1 | Yes | Free Shipping | Yes | Yes | 23 | Credit Card | 7 | Middle Aged |
| nd | M | Maroon | Spring | 3.5 | Yes | Next Day Air | Yes | Yes | 49 | PayPal | 7 | Young |
| on | M | Turquoise | Spring | 2.7 | Yes | Free Shipping | Yes | Yes | 31 | PayPal | 365 | Middle Aged |

- Removed the repeated discount_applied column.

- Loaded the cleaned dataset into MySQL using a Python–database connection.

```python
from sqlalchemy import create_engine

# MySQL connection
username = "root"
password = "admin"
host = "localhost"
port = "3306"
database = "customer_behavior"

engine = create_engine(f"mysql+pymysql://{username}:{password}@{host}:{port}/{database}")

# Write DataFrame to MySQL
table_name = "customer"
df.to_sql(table_name, engine, if_exists="replace", index=False)
```

# Data Analysis using SQL

Performed deeper analysis in MySQL to answer business questions:

- Customer segments generating the highest revenue

  Gender , Location, Age group

| Gender | Revenue |
|---|---|
| Male | 157890 |
| Female | 75191 |

| location | revenue |
|---|---|
| Montana | 5784 |
| Illinois | 5617 |
| California | 5605 |
| Idaho | 5587 |
| Nevada | 5514 |

| age_group | Revenue |
|---|---|
| Young | 62143 |
| Middle Aged | 59197 |
| Adult | 55978 |
| Senior | 55763 |

- Total revenue by Categories and top 5 selling items

| category | Revenue |
|---|---|
| Clothing | 104264 |
| Accessories | 74200 |
| Footwear | 36093 |
| Outerwear | 18524 |

| item_purchased | count |
|---|---|
| Blouse | 171 |
| Pants | 171 |
| Jewelry | 171 |
| Shirt | 169 |
| Dress | 166 |

- Do discount or promo codes actually drive a higher revenue and what seasons drive the highest sales

| Discount_or_promo_code | Total_customers | AVG_revenue | revenue |
|---|---|---|---|
| No | 2223 | 60.1305 | 133670 |
| Yes | 1677 | 59.2791 | 99411 |

| season | AVG_revenue | revenue |
|---|---|---|
| Fall | 61.5569 | 60018 |
| Spring | 58.7377 | 58679 |
| Winter | 60.3574 | 58607 |
| Summer | 58.4052 | 55777 |

- Relation between subscription status and repeat purchase

| subscription_status | Frequency_of_purchase |
|---|---|
| No | 253088 |
| Yes | 94531 |

- Shipping methods and payment modes

| payment_method | revenue |
|---|---|
| Credit Card | 40310 |
| PayPal | 40109 |
| Cash | 40002 |
| Debit Card | 38742 |
| Venmo | 37374 |
| Bank Transfer | 36544 |

| shipping_type | Total_customers | AVG_purchase_amount | revenue |
|---|---|---|---|
| 2-Day Shipping | 627 | 60.7337 | 38080 |
| Express | 646 | 60.4752 | 39067 |
| Free Shipping | 675 | 60.4104 | 40777 |
| Store Pickup | 650 | 59.8938 | 38931 |
| Next Day Air | 648 | 58.6312 | 37993 |
| Standard | 654 | 58.4602 | 38233 |

- top 5 products having the highest percentage of sales with discount applied

| item_purchased | percentage_of_sales |
|---|---|
| Hat | 50.0 |
| Sneakers | 49.7 |
| Coat | 49.1 |
| Sweater | 48.2 |
| Pants | 47.4 |

- Do review ratings affect customer purchases?

| review_rating | purchase_frequency |
|---|---|
| 3.4 | 16693 |
| 2.9 | 16555 |
| 4.2 | 16201 |
| 4 | 16122 |
| 4.9 | 15811 |

# Dashboard Creation in PowerBI