

✓ 1.Importing Libraries

```
import numpy as np
import pandas as pd
```

✓ 2. Data Ingestion

2.1 Location of data defined as dp

```
##Stores the location of the file in the variable dp
dp = "/content/drive/MyDrive/shopping_trends/shopping_trends_with_processdate.csv"
```

✓ 3.Reading Shopping Trend Data

```
#loading the file
shopping_trends_with_processdate = pd.read_csv(dp)
```

✓ 4.Data Preview


4.1 Quickly check how big your dataset is

```
shopping_trends_with_processdate.shape
```

```
(3900, 19)
```

4.2 Displays the first few columns and rows of the table

```
display(shopping_trends_with_processdate.head())
```



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating
0	1	55	Male	Blouse	Clothing	53	Kentucky	L	Gray	Winter	3.
1	2	19	Male	Sweater	Clothing	64	Maine	L	Maroon	Winter	3.
2	3	50	Male	Jeans	Clothing	73	Massachusetts	S	Maroon	Spring	3.
3	4	21	Male	Sandals	Footwear	90	Rhode Island	M	Maroon	Spring	3.
4	5	45	Male	Blouse	Clothing	49	Oregon	M	Turquoise	Spring	2.

4.3 Displays the last 5 rows of the table

```
display(shopping_trends_with_processdate.tail())
```



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating
3895	3896	40	Female	Hoodie	Clothing	28	Virginia	L	Turquoise	Summer	4
3896	3897	52	Female	Backpack	Accessories	49	Iowa	L	White	Spring	4
3897	3898	46	Female	Belt	Accessories	33	New Jersey	L	Green	Spring	2
3898	3899	44	Female	Shoes	Footwear	77	Minnesota	S	Brown	Summer	3
3899	3900	52	Female	Handbag	Accessories	81	California	M	Beige	Spring	3

4.4 Randomly select 5 rows from the dataset

```
display(shopping_trends_with_processdate.sample(5))
```



	Customer ID	Age	Gender	Item Purchased	Category	Purchase Amount (USD)	Location	Size	Color	Season	Review Rating
1188	1189	68	Male	Blouse	Clothing	29	Mississippi	M	Magenta	Summer	
3153	3154	56	Female	Socks	Clothing	88	Alabama	L	Red	Fall	
3007	3008	23	Female	Sunglasses	Accessories	60	New York	L	Lavender	Winter	
15	16	64	Male	Skirt	Clothing	81	Rhode Island	M	Teal	Winter	
1200	1201	27	Male	Coat	Outerwear	22	Connecticut	XL	Black	Winter	

5. Basic Info and Structure

5.1 Checking summary of the DataFrame's structure

```
shopping_trends_with_processdate.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 (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 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 processdate 3900 non-null object
dtypes: float64(1), int64(4), object(14)
memory usage: 579.0+ KB



```

5.2 Checking summary of statistics for the numerical columns in the DataFrame.

```
shopping_trends_with_processdate.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

5.3 Checking column names in the dataset

```
shopping_trends_with_processdate.columns
```

```

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

```

5.4 Checking if the data types are correct for each column

```
shopping_trends_with_processdate.dtypes
```



0

Customer ID	int64
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
processdate	object

dtype: object

5.5 Exploring the uniqueness of my data, counts how many distinct values there are in each column

```
shopping_trends_with_processdate.nunique()
```



	0
Customer ID	3900
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
processdate	3900

dtype: int64

✓ 6.Data Cleaning

6.1 Checking for Missing values/Nulls

```
shopping_trends_with_processdate.isnull().sum()
```



		0
<hr/>		
	Customer ID	0
	Age	0
	Gender	0
	Item Purchased	0
	Category	0
	Purchase Amount (USD)	0
	Location	0
	Size	0
	Color	0
Observation	Season	0