Data Collection and Preprocessing Phase

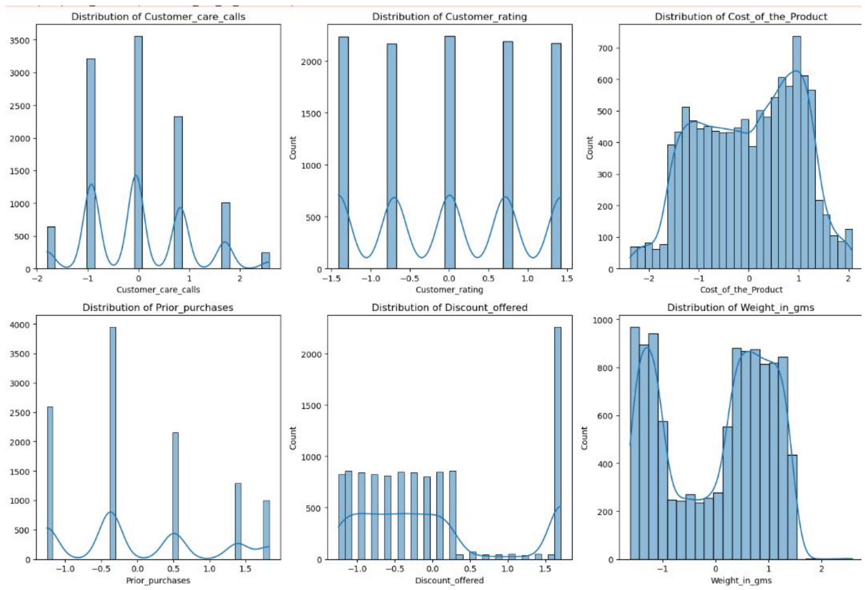
Date	15 June 2024
Team ID	SWTID1720196555
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	6 Marks

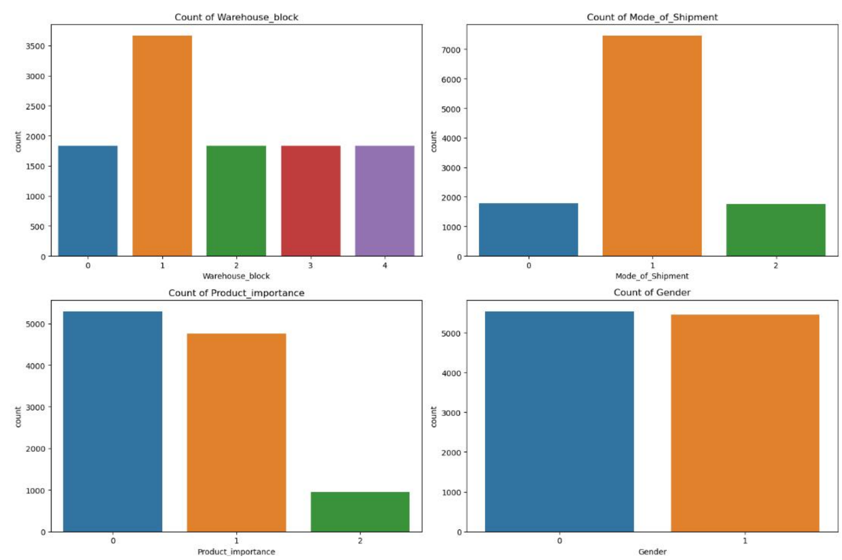
Data Exploration and Preprocessing Template

Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

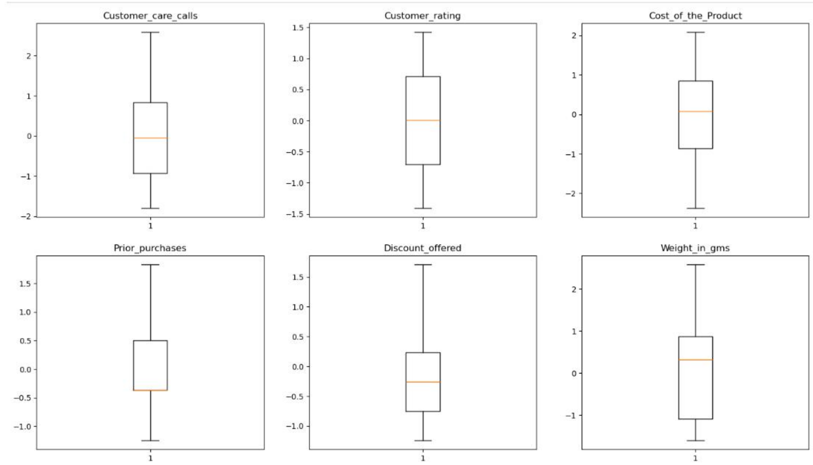
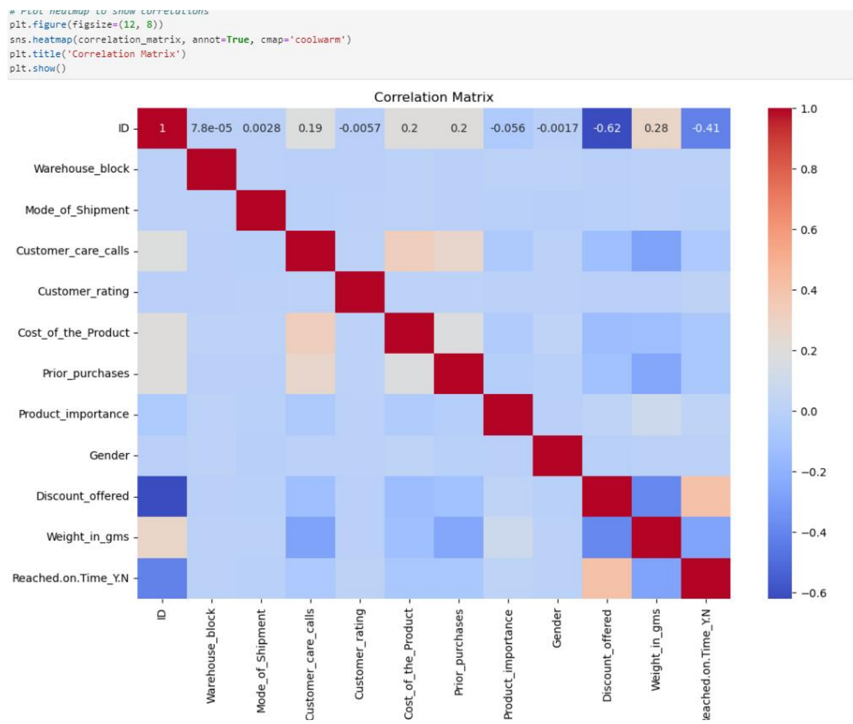
Section	Description
Data Overview	(table below)

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gende
count	10999.00000	10999.000000	10999.000000	1.099900e+04	1.099900e+04	1.099900e+04	1.099900e+04	10999.000000	10999.00000
mean	5500.00000	1.833167	0.998454	2.325624e-16	-5.975562e-17	-1.343694e-16	-7.235275e-17	0.604600	0.49586
std	3175.28214	1.343823	0.567099	1.000045e+00	1.000045e+00	1.000045e+00	1.000045e+00	0.641464	0.50000
min	1.00000	0.000000	0.000000	-1.799887e+00	-1.408199e+00	-2.376077e+00	-1.250497e+00	0.000000	0.00000
25%	2750.50000	1.000000	1.000000	-9.237994e-01	-7.007551e-01	-8.571765e-01	-3.708746e-01	0.000000	0.00000
50%	5500.00000	1.000000	1.000000	-4.771132e-02	6.689172e-03	7.913188e-02	-3.708746e-01	1.000000	0.00000
75%	8249.50000	3.000000	1.000000	8.283768e-01	7.141334e-01	8.489855e-01	5.087480e-01	1.000000	1.00000
max	10999.00000	4.000000	2.000000	2.580553e+00	1.421578e+00	2.076590e+00	1.828182e+00	2.000000	1.00000

<table>
<tr>
<td>Univariate Analysis</td>
<td>


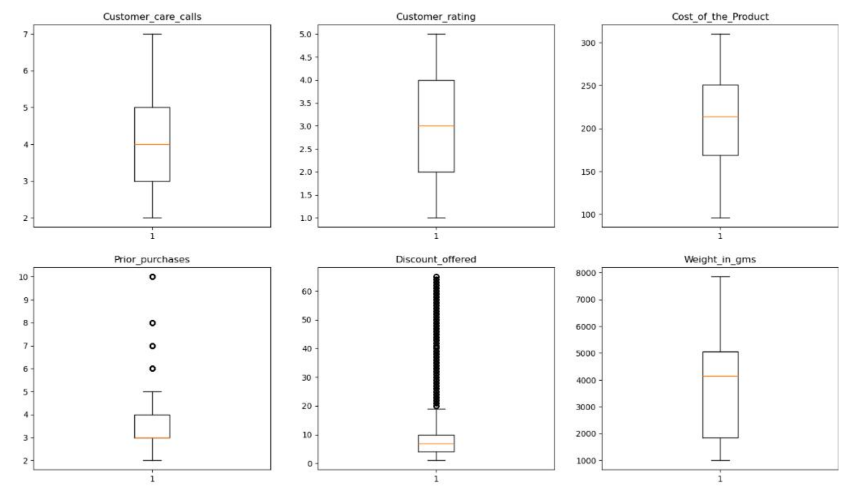


</td>
</tr>
</table>

Bivariate Analysis	
Multivariate Analysis	

<table>
<tr>
<td>Outliers and Anomalies</td>
<td>

```python
import matplotlib.pyplot as plt

c = 0

plt.figure(figsize=(18, 10))

for i in data.drop(columns=[
    'Warehouse_block', 'Mode_of_Shipment', 'Product_importance', 'Gender', 'Reached.on.Time_Y.N', 'ID'
]).columns:

    if str(data[i].dtype) == 'object':
        continue

    plt.subplot(2, 3, c + 1)

    plt.boxplot(data[i])

    plt.title(i)

    c += 1

plt.show()
```



</td>
</tr>
<tr>
<td colspan="2">Data Preprocessing Code Screenshots</td>
</tr>
<tr>
<td>Loading Data</td>
<td>

Code to load the dataset into the preferred environment (e.g., Python, R).

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

data = pd.read_csv('Train.csv')
```

</td>
</tr>
</table>

Handling Missing Data	``` [5]: data.isnull().any() [5]: ID False Warehouse_block False Mode_of_Shipment False Customer_care_calls False Customer_rating False Cost_of_the_Product False Prior_purchases False Product_importance False Gender False Discount_offered False Weight_in_gms False Reached.on.Time_Y.N False dtype: bool [6]: data.isnull().sum() [6]: ID 0 Warehouse_block 0 Mode_of_Shipment 0 Customer_care_calls 0 Customer_rating 0 Cost_of_the_Product 0 Prior_purchases 0 Product_importance 0 Gender 0 Discount_offered 0 Weight_in_gms 0 Reached.on.Time_Y.N 0 dtype: int64 ```
Data Transformation	``` unique_values = { 'Warehouse_block': data['Warehouse_block'].unique(), 'Mode_of_Shipment': data['Mode_of_Shipment'].unique(), 'Product_importance': data['Product_importance'].unique(), 'Gender': data['Gender'].unique() } unique_values {'Warehouse_block': array(['D', 'F', 'A', 'B', 'C'], dtype=object), 'Mode_of_Shipment': array(['Flight', 'Ship', 'Road'], dtype=object), 'Product_importance': array(['low', 'medium', 'high'], dtype=object), 'Gender': array(['F', 'M'], dtype=object)} ```

	`label_map={}` `for i in data.columns:` `    if str(data[i].dtype) == 'object':` `        temp={}` `        cats=data[i].unique()` `        for index in range(len(cats)):` `            temp[cats[index]]=index` `        label_map[i]=temp` `        #labeling` `        data[i]=data[i].map(temp)` `label_map` `{'Warehouse_block': {'D': 0, 'F': 1, 'A': 2, 'B': 3, 'C': 4},` ` 'Mode_of_Shipment': {'Flight': 0, 'Ship': 1, 'Road': 2},` ` 'Product_importance': {'low': 0, 'medium': 1, 'high': 2},` ` 'Gender': {'F': 0, 'M': 1}}`
Feature Engineering	Attached the codes in final submissions
Save Processed Data	-