Data Collection and Preprocessing Phase

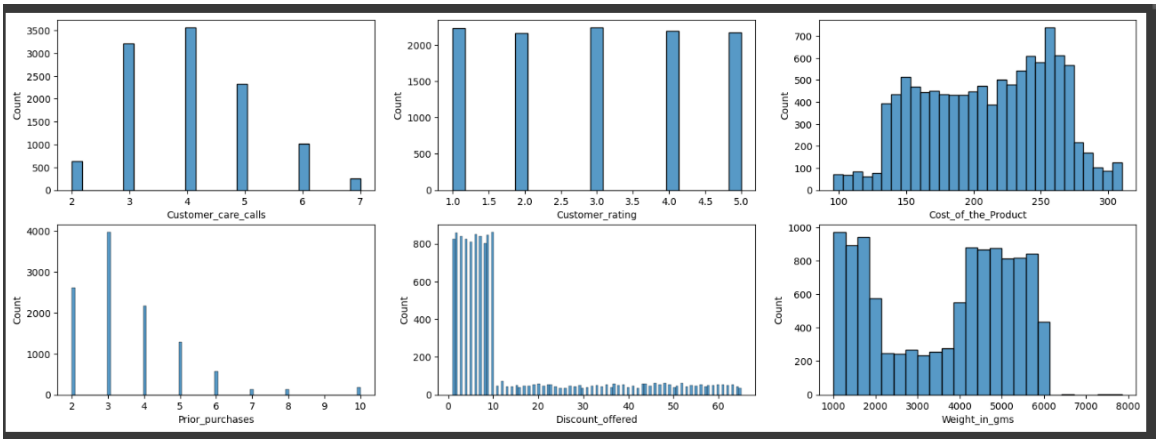
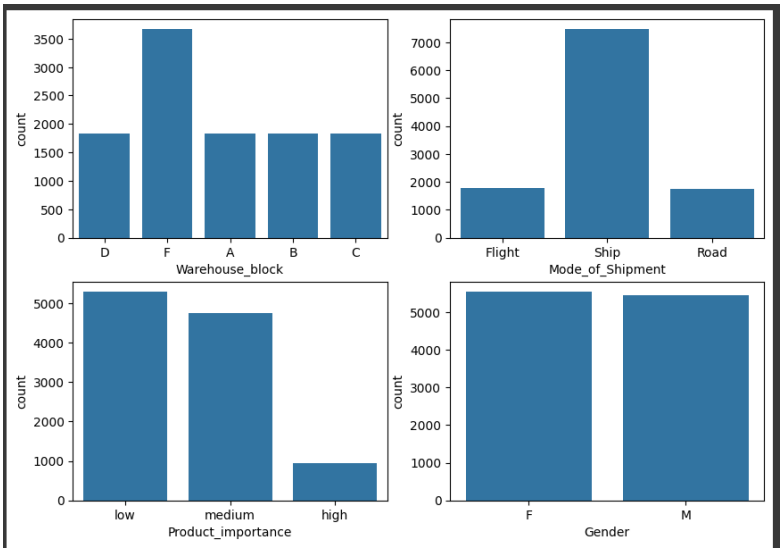
Date	5 July 2024
Team ID	SWTID1720082658
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	6 Marks

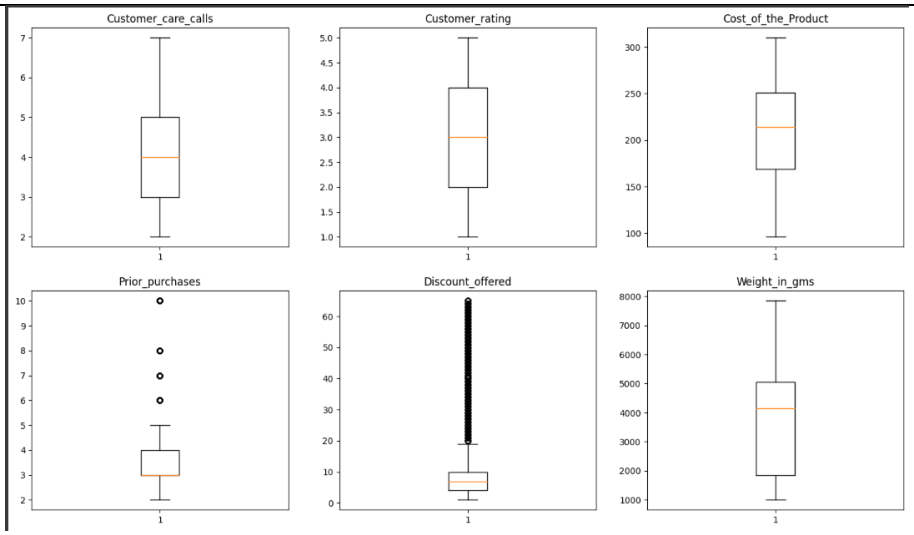
Data Exploration and Preprocessing Template

Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employed for preprocessing tasks like normalization and feature engineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation for insights and predictions.

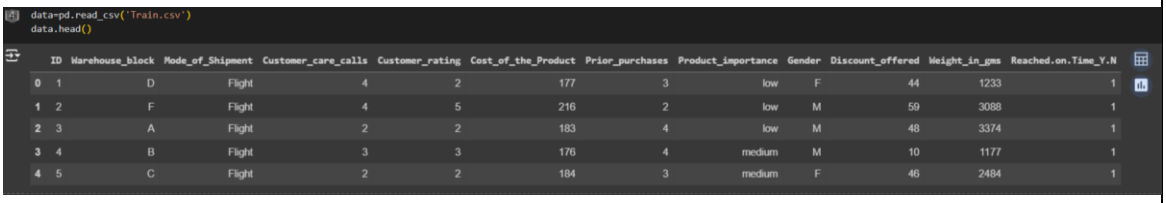
Section	Description
Data Overview	Dimensions : 10999 rows x 12 columns

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	2.333394	1.516865	4.054459	2.990545	210.196836
std	3175.28214	1.490726	0.756894	1.141490	1.413603	48.063272
min	1.00000	0.000000	0.000000	2.000000	1.000000	96.000000
25%	2750.50000	1.000000	1.000000	3.000000	2.000000	169.000000
50%	5500.00000	3.000000	2.000000	4.000000	3.000000	214.000000
75%	8249.50000	4.000000	2.000000	5.000000	4.000000	251.000000
max	10999.00000	4.000000	2.000000	7.000000	5.000000	310.000000

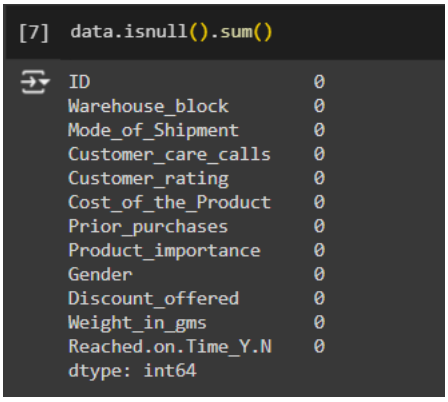
Univariate Analysis	
Bivariate Analysis	
Multivariate Analysis	

Outliers and Anomalies	

Data Preprocessing Code Screenshots

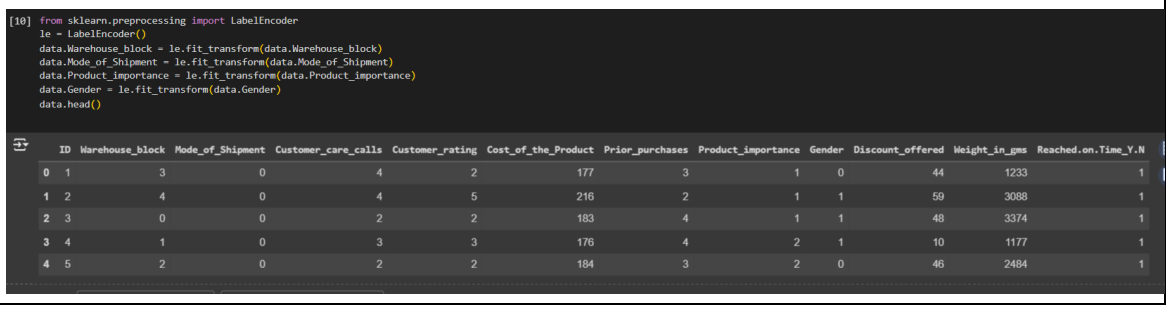
Loading Data	

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
0	1	D	Flight	4	2	177	3	low	F	44	1233	1
1	2	F	Flight	4	5	216	2	low	M	59	3088	1
2	3	A	Flight	2	2	183	4	low	M	48	3374	1
3	4	B	Flight	3	3	176	4	medium	M	10	1177	1
4	5	C	Flight	2	2	184	3	medium	F	46	2484	1

Handling Missing Data	

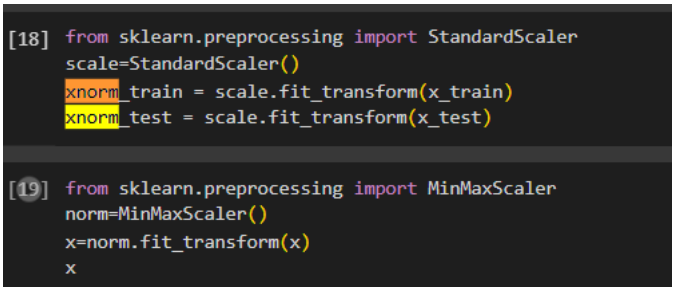
```
[7] data.isnull().sum()

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 Encoding	

```
[10] from sklearn.preprocessing import LabelEncoder
     le = LabelEncoder()
     data.Warehouse_block = le.fit_transform(data.Warehouse_block)
     data.Mode_of_Shipment = le.fit_transform(data.Mode_of_Shipment)
     data.Product_importance = le.fit_transform(data.Product_importance)
     data.Gender = le.fit_transform(data.Gender)
     data.head()
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
0	1	3	0	4	2	177	3	1	0	44	1233	1
1	2	4	0	4	5	216	2	1	1	59	3088	1
2	3	0	0	2	2	183	4	1	1	48	3374	1
3	4	1	0	3	3	176	4	2	1	10	1177	1
4	5	2	0	2	2	184	3	2	0	46	2484	1

Data Transformation	
Feature Engineering	Code is in the final code submitted.
Save Processed Data	-

```
[18] from sklearn.preprocessing import StandardScaler
     scale=StandardScaler()
     xnorm_train = scale.fit_transform(x_train)
     xnorm_test = scale.fit_transform(x_test)

[19] from sklearn.preprocessing import MinMaxScaler
     norm=MinMaxScaler()
     x=norm.fit_transform(x)
     x
```