Data Collection and Preprocessing Phase

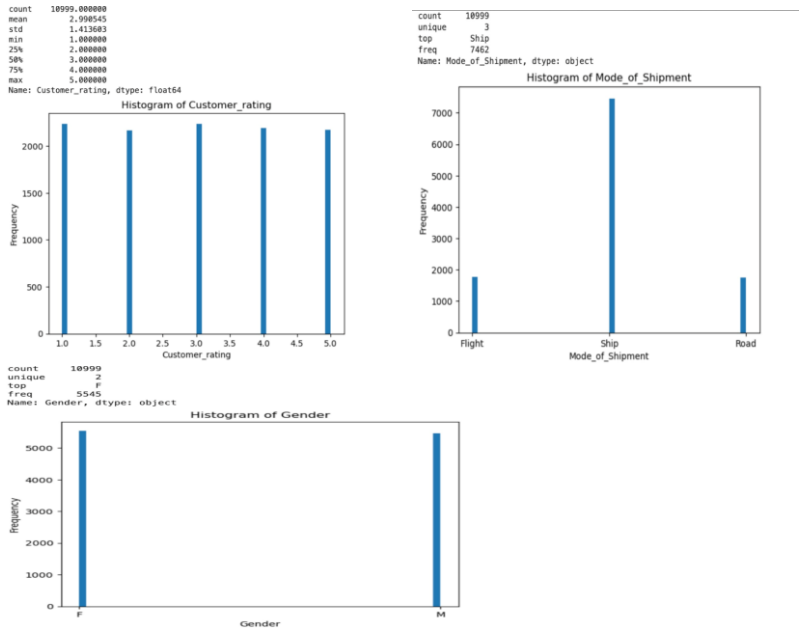
Date	**11 JULY 2024**
Team ID	**SWTID1720115788**
Project Title	**Ecommerce Shipping Prediction Using Machine Learning**
Maximum Marks	**6 Marks**

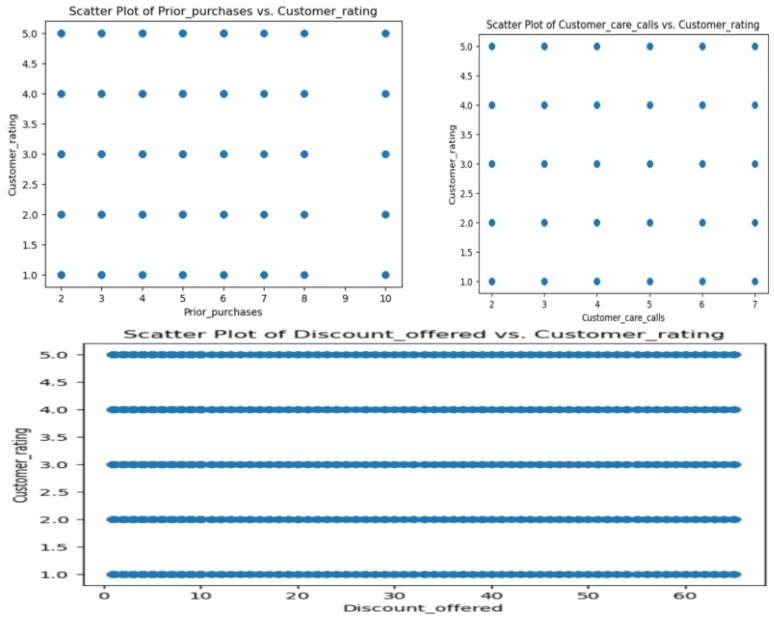
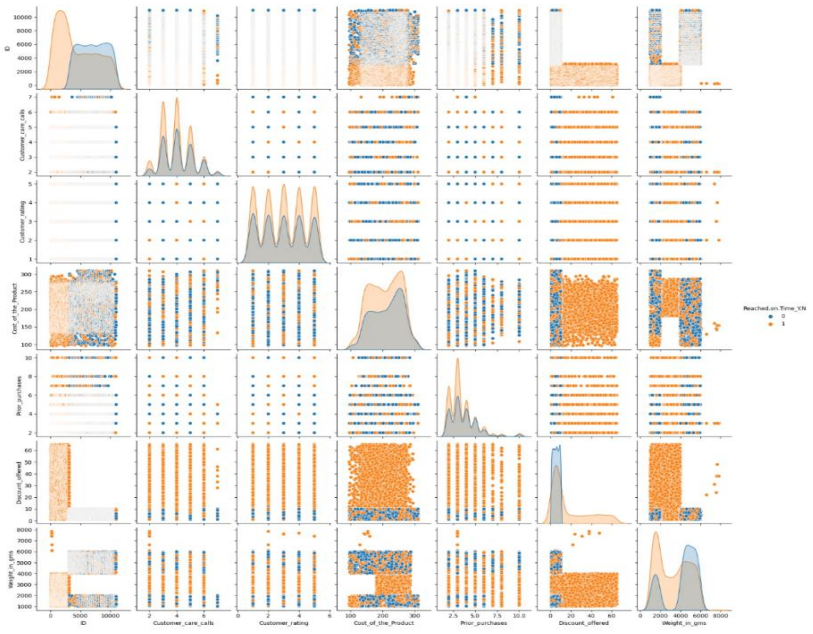
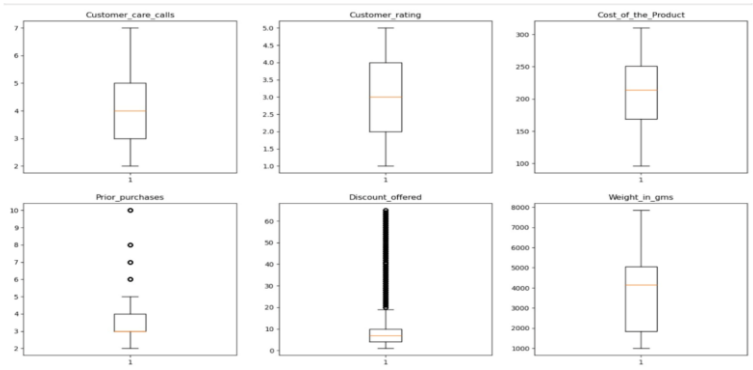
Data Exploration and Preprocessing Template

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	
Univariate Analysis	

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	Reached.on.Time_Y.N
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0.596691
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0.490584
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0.000000
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0.000000
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1.000000
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1.000000
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1.000000



Bivariate Analysis	
Multivariate Analysis	
Outliers and Anomalies	

Data Preprocessing Code Screenshots	
Loading Data	import pandas as pd df = pd.read_csv('/Users/shanmukhanandudu/Downloads/train (3).csv') df
Handling Missing Data	df.isnull().sum()

Data Transformation	# Encode categorical variables
	le = LabelEncoder()
	df['Warehouse_block'] = le.fit_transform(df['Warehouse_block'])
	df['Mode_of_Shipment'] = le.fit_transform(df['Mode_of_Shipment'])
	df['Product_importance'] = le.fit_transform(df['Product_importance'])
	df['Gender'] = le.fit_transform(df['Gender'])
	
	# Scale/normalize features
	scaler = StandardScaler()
	columns_to_scale = ['Customer_care_calls', 'Customer_rating', 'Cost_of_the_Product', 'Prior_purchases', 'Discount_offered', 'Weight_in_gms']
	df[columns_to_scale] = scaler.fit_transform(df[columns_to_scale])

Feature Engineering	import pandas as pd
	
	# create a sample dataframe
	data = {'priority': ['low', 'medium', 'high', 'low', 'medium', 'high']}
	df = pd.DataFrame(data)
	
	# create a new column with the mapped values
	df['priority_code'] = df['priority'].map({'low': 0, 'medium': 1, 'high': 2})
	
	print(df)

Save Processed Data	df.to_csv('my_dataset.csv', index=False)