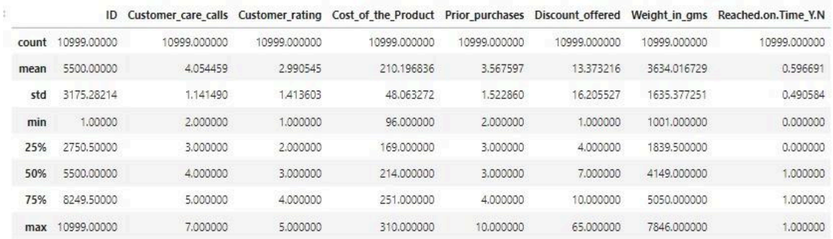
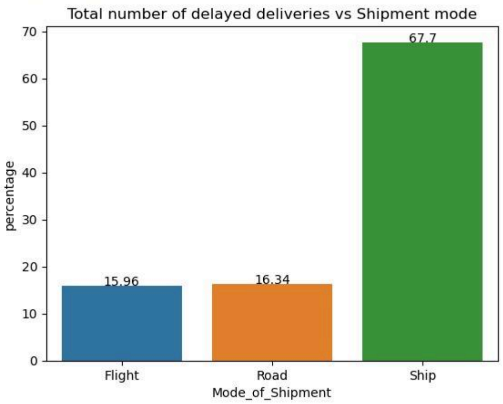
Data Collection and Preprocessing Phase

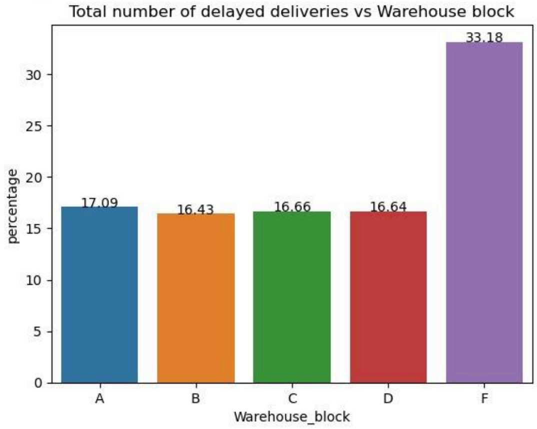
Date	11 JULY 2024
Team ID	SWTID1720116037
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	6 Marks

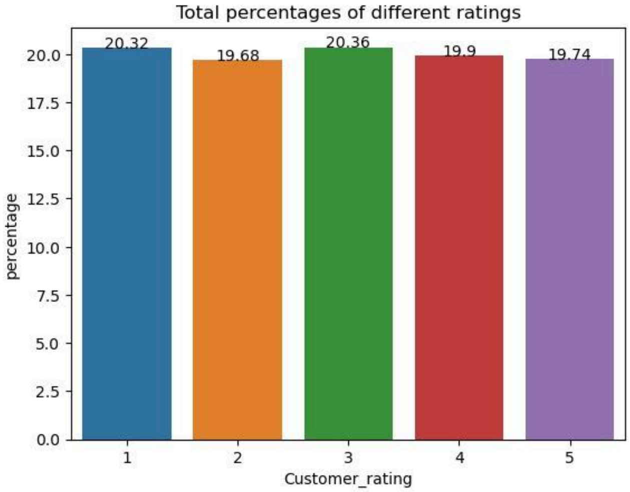
Data Exploration and Preprocessing Template

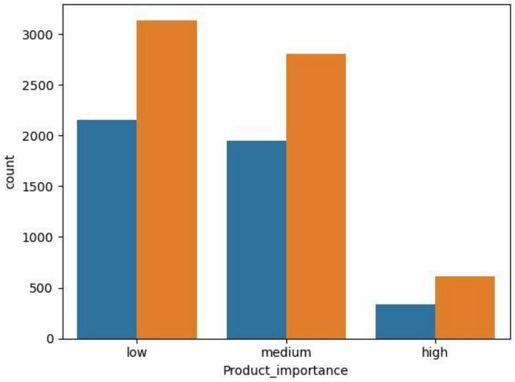
Identifies data sources, assesses quality issues like missing values and duplicates, and implements resolution plans to ensure accurate and reliable analysis.

<table>
<tr><th>Section</th><th>Description</th></tr>
<tr><td>Data Overview</td><td>

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


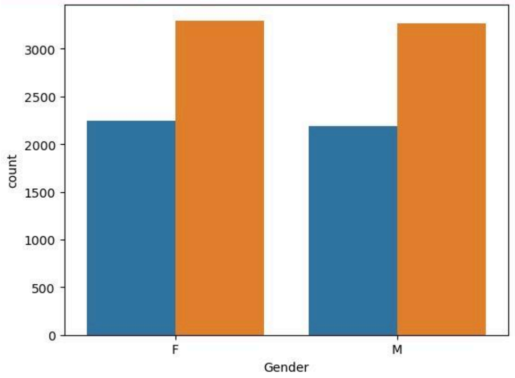
</td></tr>
</table>

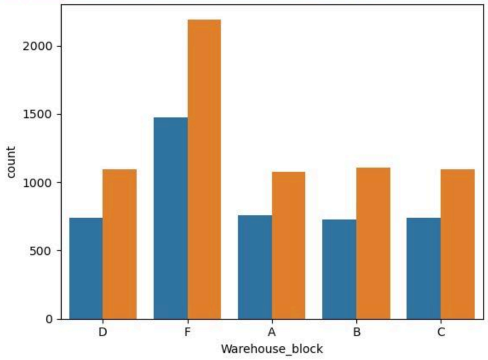
<table>
<tr>
<td></td>
<td>




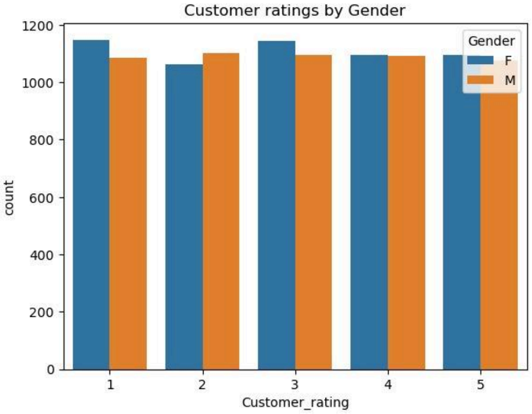
</td>
</tr>
<tr>
<td>Bivariate Analysis</td>
<td>


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

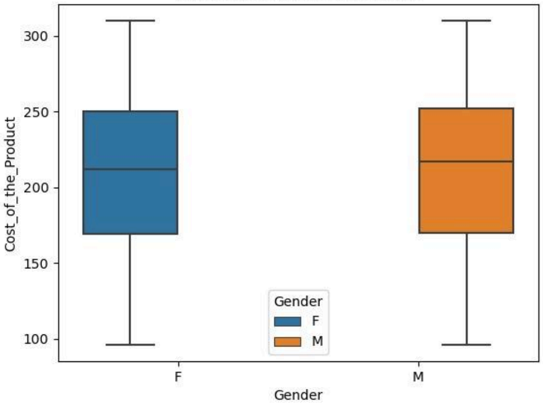


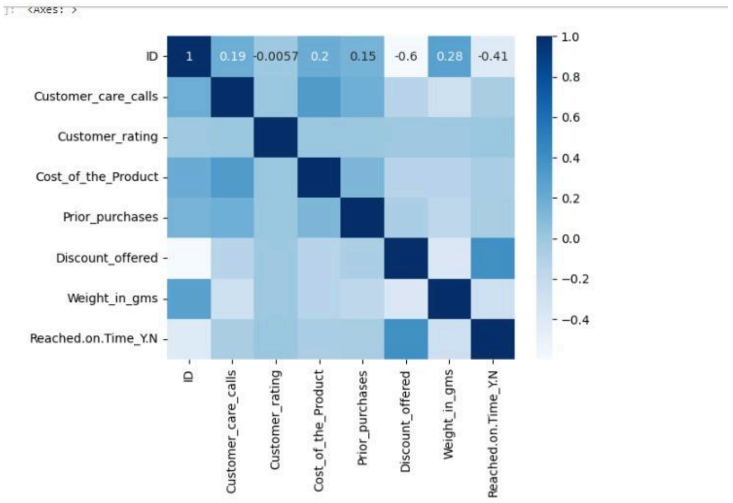
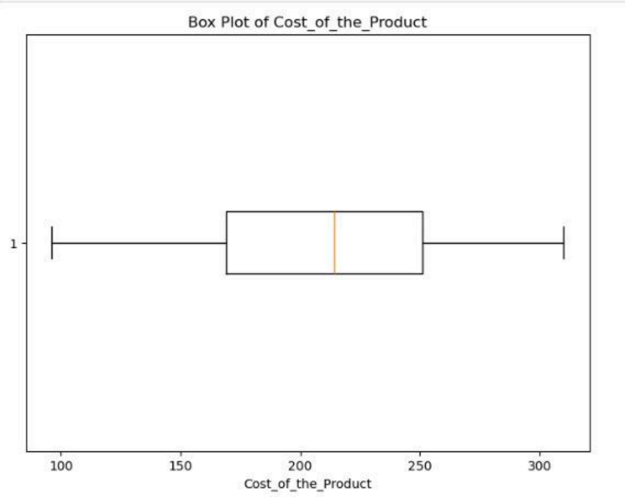
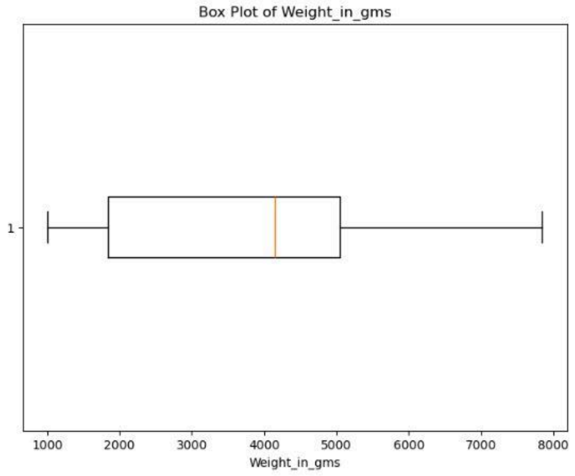


Customer ratings by Gender



Amount spent by each Gender



Multivariate Analysis	
Outliers and Anomalies	 
Data Preprocessing Code Screenshots	

Loading Data	import pandas as pd dataset = pd.read_csv('train.csv') dataset
Handling Missing Data	dataset.isnull().sum()

Data Transformation	# Encode categorical variables
	le = LabelEncoder()
	dataset['Warehouse_block'] = le.fit_transform(dataset['Warehouse_block'])
	dataset['Mode_of_Shipment'] = le.fit_transform(dataset['Mode_of_Shipment'])
	dataset['Product_importance'] = le.fit_transform(dataset['Product_importance'])
	dataset['Gender'] = le.fit_transform(dataset['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']
	dataset[columns_to_scale] = scaler.fit_transform(dataset[columns_to_scale])

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

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