Data Collection and Preprocessing Phase

Date	15 March 2024
Team ID	SWTID1720097765
Project Title	Ecommerce Shipping Prediction Using Machine Learning
Maximum Marks	6 Marks

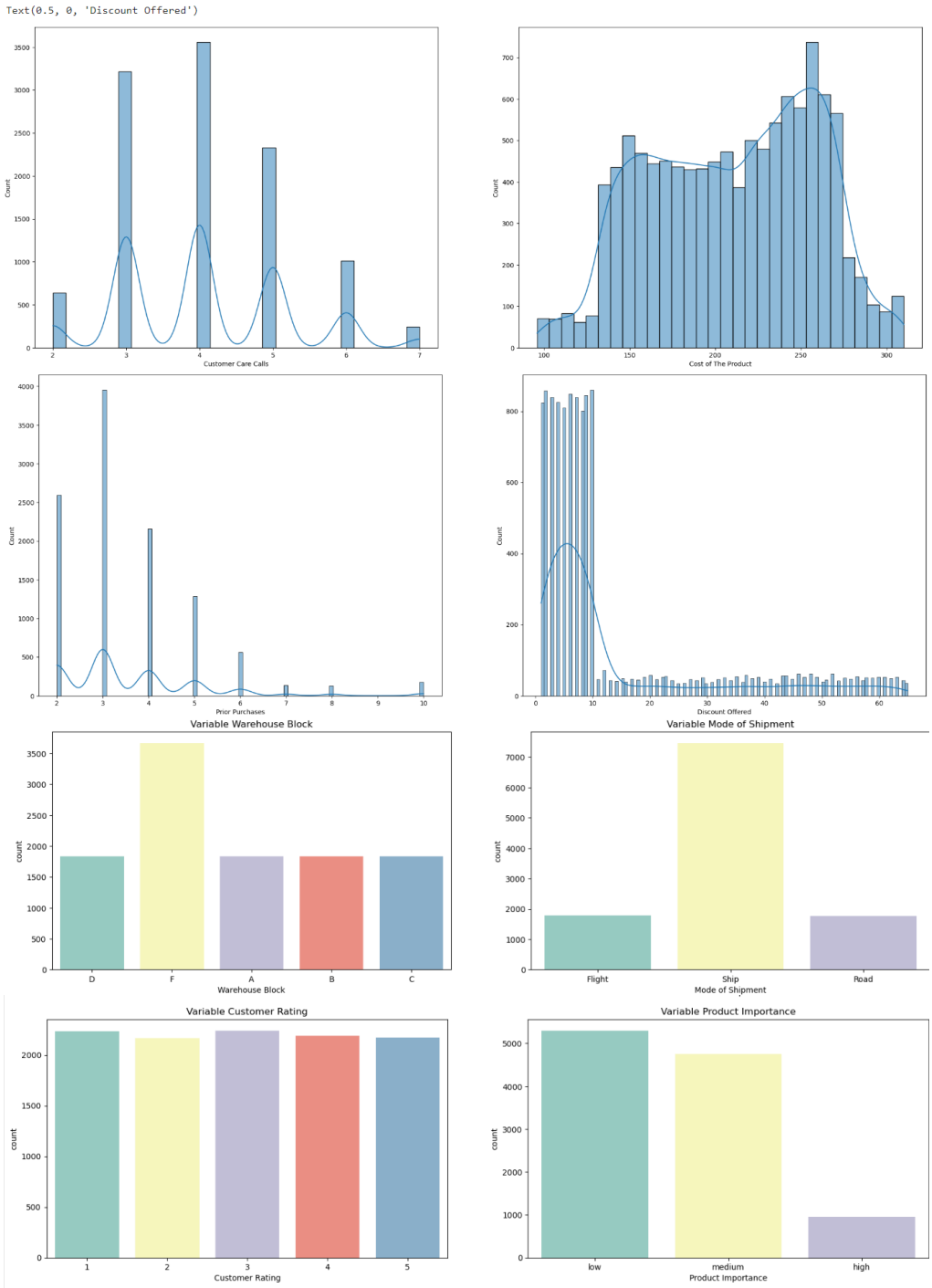
Data Exploration and Preprocessing Template

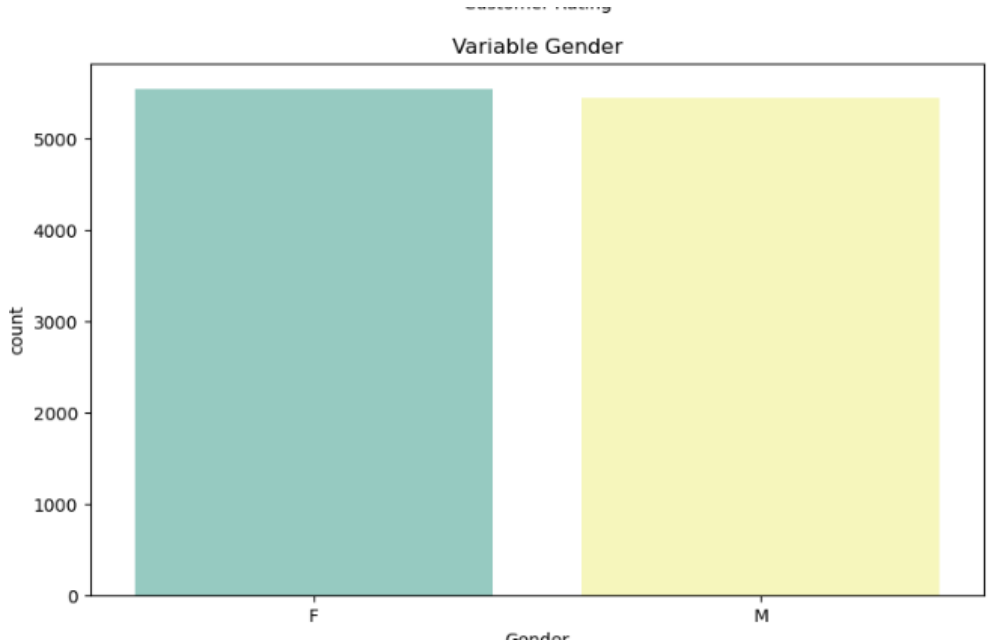
Dataset variables will be statistically analyzed to identify patterns and outliers, with Python employedforpreprocessingtaskslikenormalizationandfeatureengineering. Data cleaning will address missing values and outliers, ensuring quality for subsequent analysis and modeling, and forming a strong foundation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

Section	Description

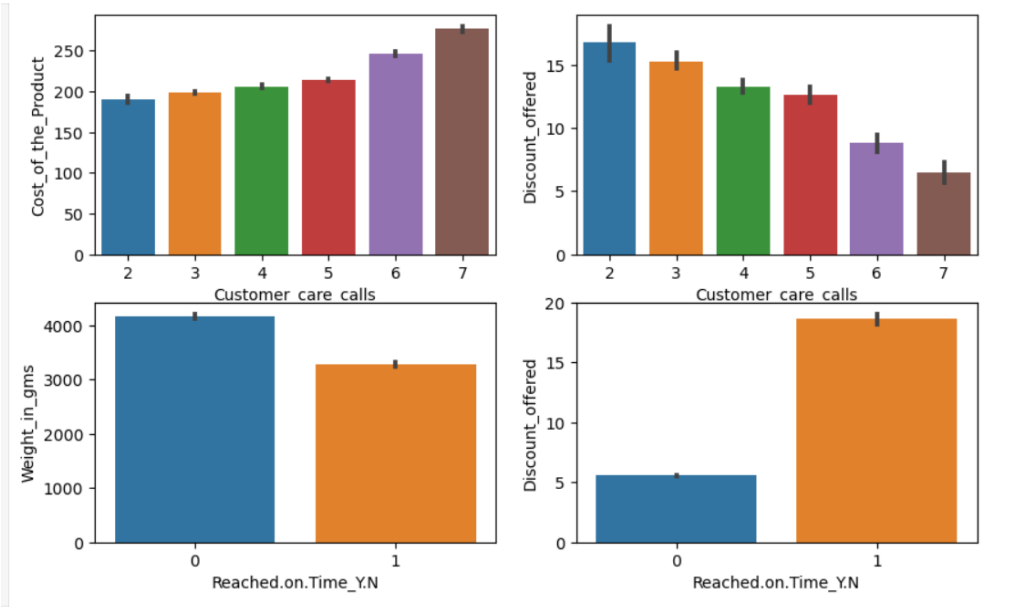
Data Overview	Dimension:
	`[4]: print(data.shape)`
	`(10999, 12)`
	Descriptive analysis:
	`data.describe()`

	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

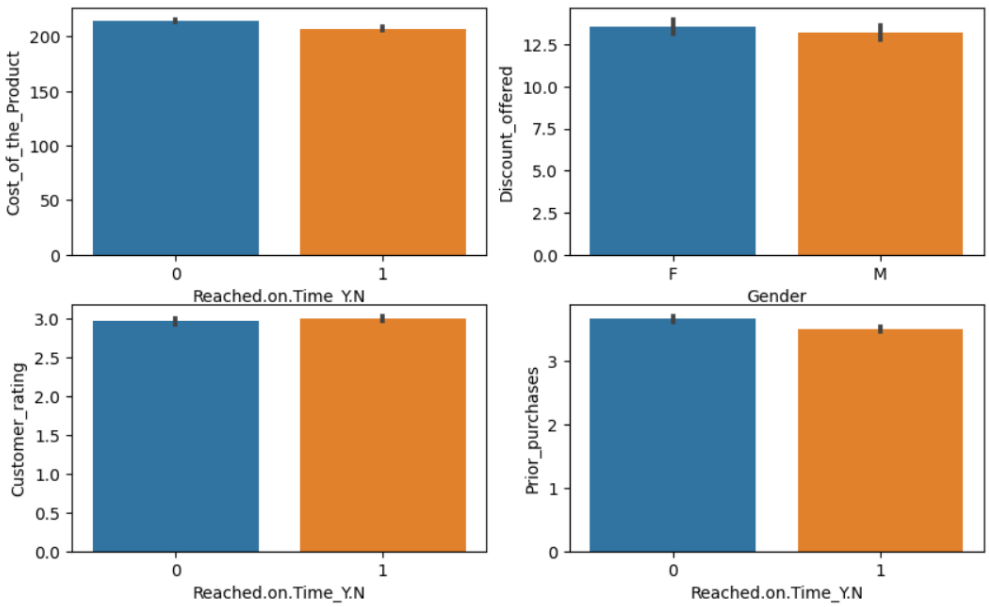
Univariate Analysis	Text(0.5, 0, 'Discount Offered') 

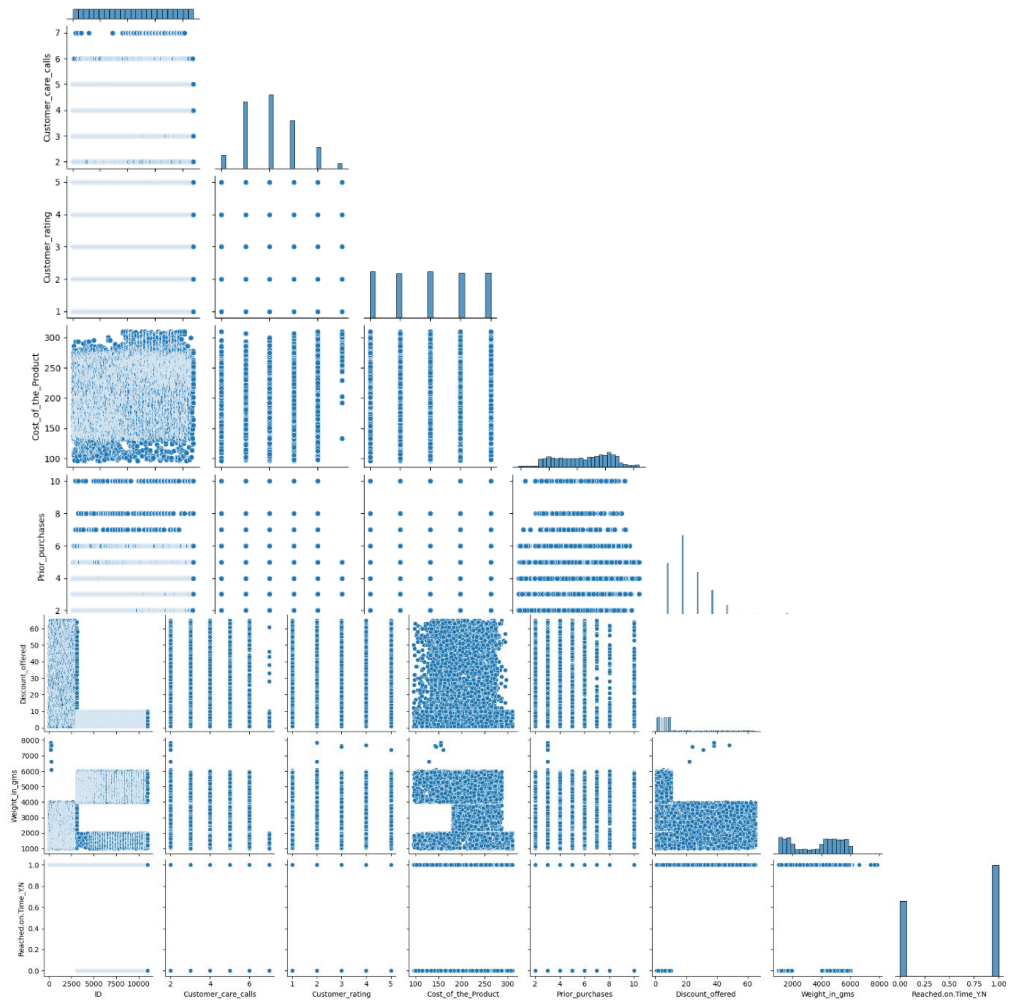


Bivariate Analysis



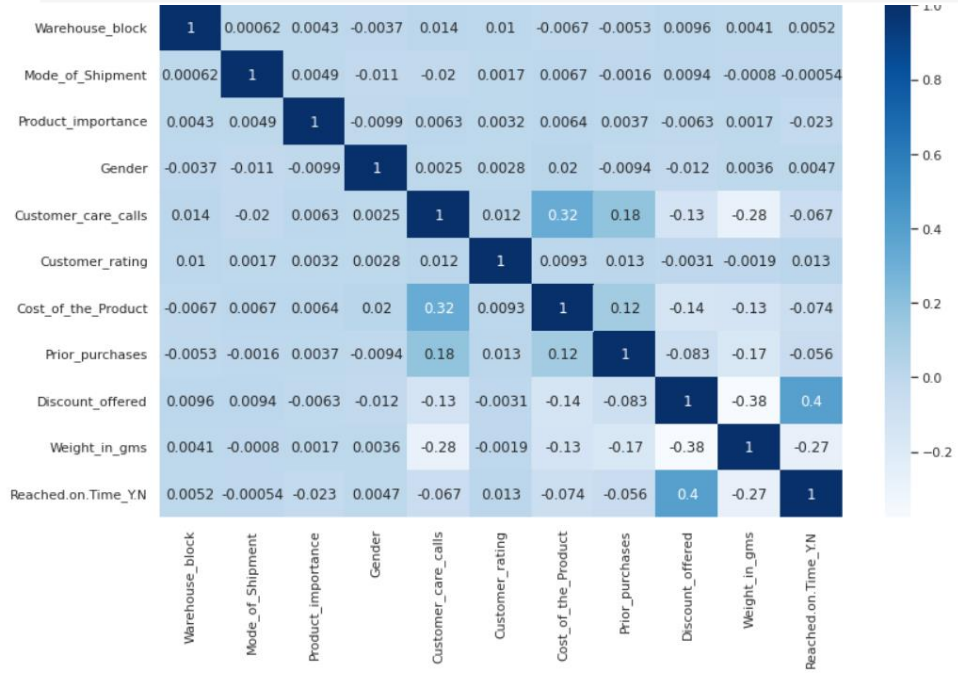
0


Cost_of_the_Product
Reached.on.Time_Y.N
Discount_offered
Gender
Customer_rating
Reached.on.Time_Y.N
Prior_purchases
Reached.on.Time_Y.N


Customer_care_calls
Customer_rating
Cost_of_the_Product
Prior_purchases
Discount_offered
Weight_in_gms
Reached.on.Time_Y.N
ID
Customer_care_calls
Customer_rating
Cost_of_the_Product
Prior_purchases
Discount_offered
Weight_in_gms
Reached.on.Time_Y.N

Multivariate Analysis	Multivariate analysis [16]: `hm=data.corr(numeric_only=True)` `hm` [16]:

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance
ID	1.000000	0.000070	-0.002459	0.188998	-0.005722	0.196791	0.145369	0.029081
Warehouse_block	0.000070	1.000000	0.000617	0.014496	0.010169	-0.006679	-0.005262	0.004260
Mode_of_Shipment	-0.002459	0.000617	1.000000	-0.020164	0.001679	0.006681	-0.001640	0.004911
Customer_care_calls	0.188998	0.014496	-0.020164	1.000000	0.012209	0.323182	0.180771	0.006273
Customer_rating	-0.005722	0.010169	0.001679	0.012209	1.000000	0.009270	0.013179	0.003157
Cost_of_the_Product	0.196791	-0.006679	0.006681	0.323182	0.009270	1.000000	0.123676	0.006366
Prior_purchases	0.145369	-0.005262	-0.001640	0.180771	0.013179	0.123676	1.000000	0.003662
Product_importance	0.029081	0.004260	0.004911	0.006273	0.003157	0.006366	0.003662	1.000000
Gender	-0.001695	-0.003700	-0.011288	0.002545	0.002775	0.019759	-0.009395	-0.009865
Discount_offered	-0.598278	0.009569	0.009364	-0.130750	-0.003124	-0.138312	-0.082769	-0.006251
Weight_in_gms	0.278312	0.004086	-0.000797	-0.276615	-0.001897	-0.132604	-0.168213	0.001652
Reached.on.Time_Y.N	-0.411822	0.005214	-0.000535	-0.067126	0.013119	-0.073587	-0.055515	-0.023483



Outliers and Anomalies	Identification and treatment of outliers.

Data Preprocessing Code Screenshots

Loading Data

```
data=pd.read_csv("train.csv")
data
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	Discount_
0	1	D	Flight	4	2	177	3	low	F	
1	2	F	Flight	4	5	216	2	low	M	
2	3	A	Flight	2	2	183	4	low	M	
3	4	B	Flight	3	3	176	4	medium	M	
4	5	C	Flight	2	2	184	3	medium	F	
...	...	...	...	...	...	...	...	...	...	
10994	10995	A	Ship	4	1	252	5	medium	F	
10995	10996	B	Ship	4	1	232	5	medium	F	
10996	10997	C	Ship	5	4	242	5	low	F	
10997	10998	F	Ship	5	2	223	6	medium	M	
10998	10999	D	Ship	2	5	155	5	low	F	

10999 rows × 12 columns

Handling Missing Data

```
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 Transformation	**Encoding**
	`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['Reached.on.Time_Y.N']=le.fit_transform(data['Reached.on.Time_Y.N'])`
	Scaling
	`sc=StandardScaler()`
	`names=x.columns`
	`x=sc.fit_transform(x)`
	`x=pd.DataFrame(x,columns=names)`

Feature Engineering	
Save Processed Data	-