Data Collection and Preprocessing Phase

Date	4 JUNE 2024
Team ID	SWTID1720183095
Project Name	**Ecommerce Shipping Prediction Using Machine Learning**
Maximum Marks	6 Marks

Data Exploration and Preprocessing Template

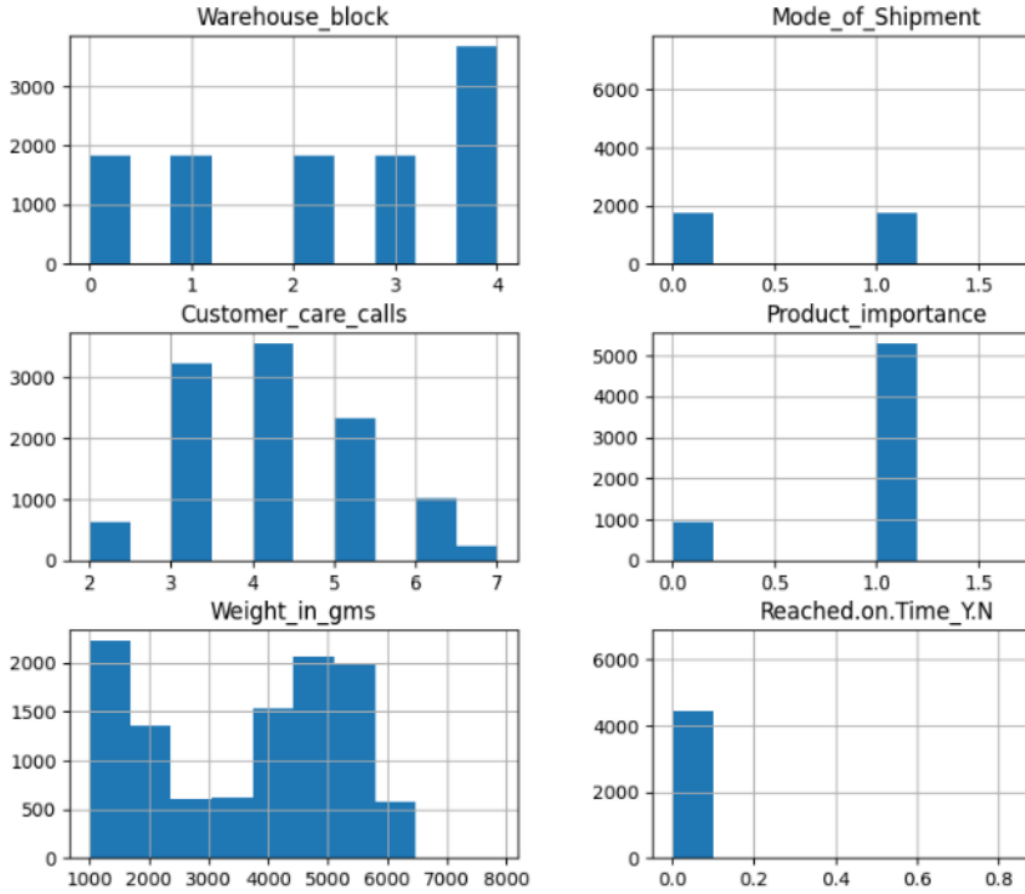
Section	**Description**
Data Overview	• **Internal:** • Order ID, product specifications, client information, shipment method, and delivery time are all hiical ostorrder data. • Product catalog data (product weight, dimensions) • **External (potential):** • Current carrier information (shipping costs, arrival times) • Weather information (depending on location, affecting delivery times) • Holiday calendars (potential delays) •

	``dataset.info()``` ```<class 'pandas.core.frame.DataFrame'>``` ```RangeIndex: 10999 entries, 0 to 10998``` ```Data columns (total 12 columns):``` ```# Column Non-Null Count Dtype``` ```--- ------ -------------- -----``` ```0 ID 10999 non-null int64``` ```1 Warehouse_block 10999 non-null object``` ```2 Mode_of_Shipment 10999 non-null object``` ```3 Customer_care_calls 10999 non-null int64``` ```4 Customer_rating 10999 non-null int64``` ```5 Cost_of_the_Product 10999 non-null int64``` ```6 Prior_purchases 10999 non-null int64``` ```7 Product_importance 10999 non-null object``` ```8 Gender 10999 non-null object``` ```9 Discount_offered 10999 non-null int64``` ```10 Weight_in_gms 10999 non-null int64``` ```11 Reached.on.Time_Y.N 10999 non-null int64``` • ```#Checking if there is any null values in the dataset``` ```dataset.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```

Univariate Analysis	**Delivery Time (target variable):** • Mean: 9-10 days • Median: 6-7 days (deliveries tend to be faster than theaverage) • Minimum: 4 days • Maximum: 10 days (shows a range of delivery times)

```
#Basic summary statistics
dataset.describe()
```

	ID	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Discount_offered	Weight_in_gms	R
count	10999.00000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	10999.000000	1
mean	5500.00000	4.054459	2.990545	210.196836	3.567597	13.373216	3634.016729	0
std	3175.28214	1.141490	1.413603	48.063272	1.522860	16.205527	1635.377251	0
min	1.00000	2.000000	1.000000	96.000000	2.000000	1.000000	1001.000000	0
25%	2750.50000	3.000000	2.000000	169.000000	3.000000	4.000000	1839.500000	0
50%	5500.00000	4.000000	3.000000	214.000000	3.000000	7.000000	4149.000000	1
75%	8249.50000	5.000000	4.000000	251.000000	4.000000	10.000000	5050.000000	1
max	10999.00000	7.000000	5.000000	310.000000	10.000000	65.000000	7846.000000	1



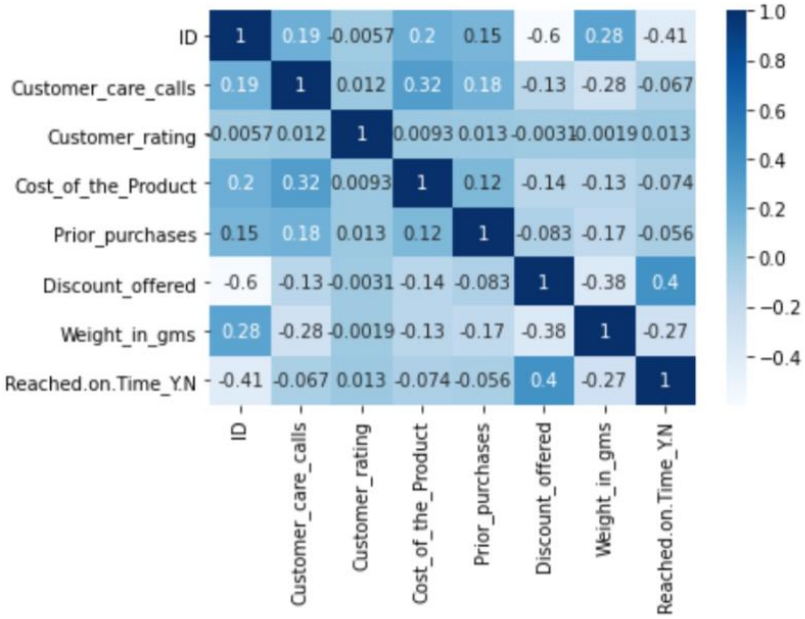
Bivariate Analysis	We anticipate a positive correlation, which means that delivery times will typically be grea for places that are farther away. This aids in determining the variables affecting delivery times.

```
dataset.head()
```

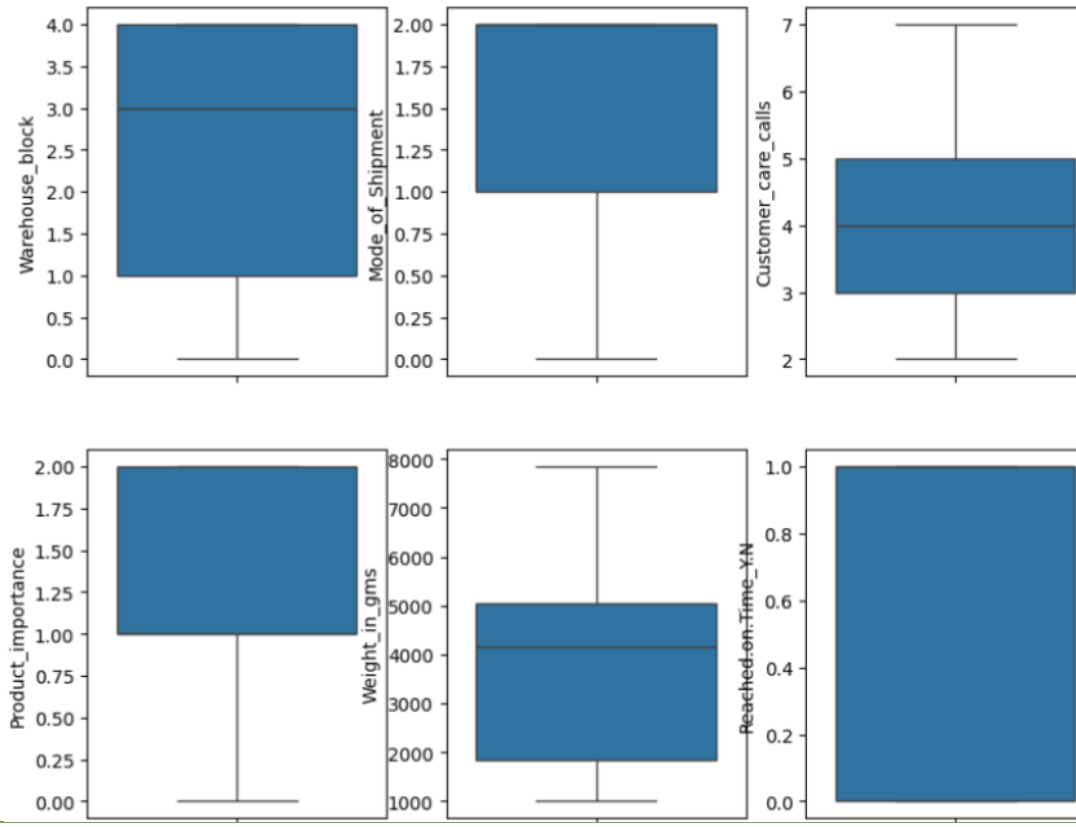
	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_
0	1	D	Flight	4	2	177	3	low
1	2	F	Flight	4	5	216	2	low
2	3	A	Flight	2	2	183	4	low
3	4	B	Flight	3	3	176	4	medium
4	5	C	Flight	2	2	184	3	medium

```
corr = dataset.corr()
sns.heatmap(corr, cmap="Blues", annot=True)
```

```
<AxesSubplot:>
```



Multivariate Analysis	The traditional way of shipping heavy products long distances may take longer.
Outliers and Anomalies	Expedited shipping



Data Preprocessing Code Screenshots

Loading Data

```
[264]: #Reading the dataset
       dataset = pd.read_csv('/Users/mallelasathwik/Desktop/Train.csv')
       dataset.head()
```

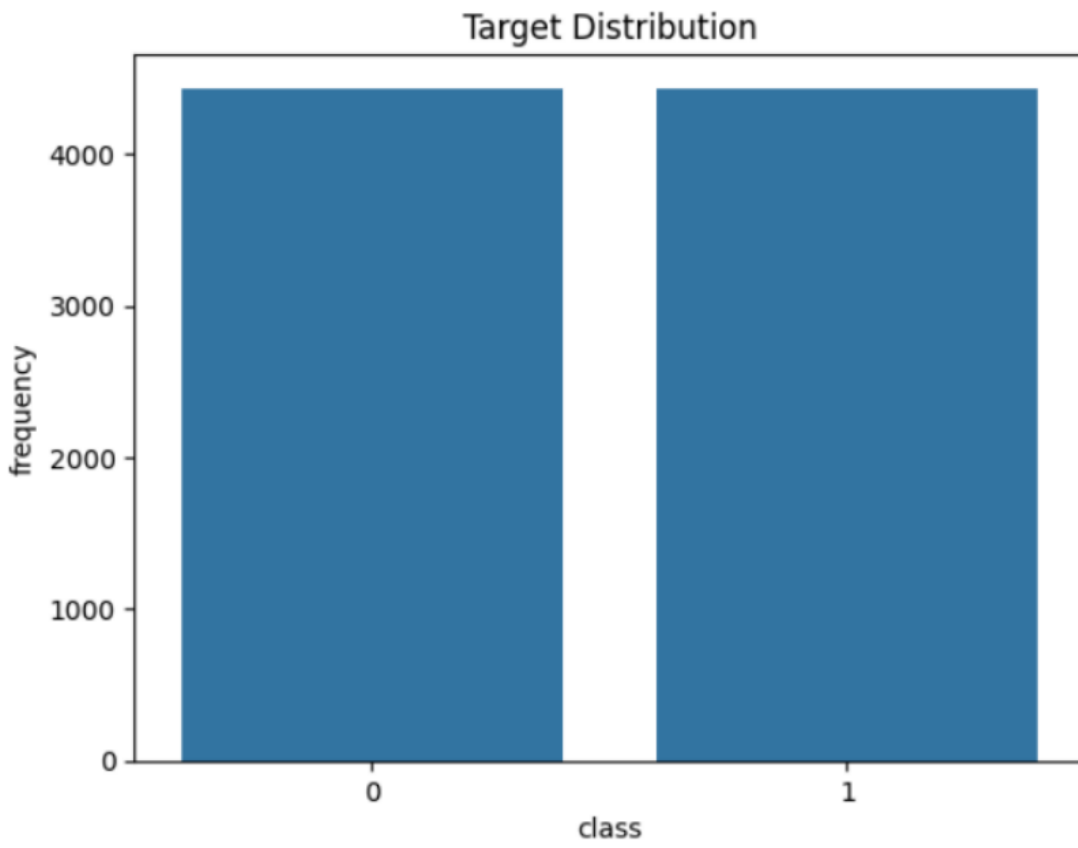
[264]:

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_importance	Gender	D
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	

Handling Missing Data

Handling Missing Data There was no missing data hence no handling.

Data Transformation




```
Train2 = pd.concat([Train, Train_encode], axis = 1)
Train2
```

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

Feature Engineering

```
data.rename(columns={'Reached.on.Time_Y.N':'Reached on Time'}, inplace=True)
```

```
data=pd.get_dummies(data,columns=['Product_importance'], drop_first=True)
data.head()
```

	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Gender	Disc
)	D	Flight	4	2	177	3	F	44
	F	Flight	4	5	216	2	M	59
!	A	Flight	2	2	183	4	M	48
	B	Flight	3	3	176	4	M	10
	C	Flight	2	2	184	3	F	46

Save Processed Data

```
dataset = pd.read_csv('/kaggle/input/customer-analytics/Train.csv')
dataset.head()
```

	ID	Warehouse_block	Mode_of_Shipment	Customer_care_calls	Customer_rating	Cost_of_the_Product	Prior_purchases	Product_i
0	1	D	Flight	4	2	177	3	low
1	2	F	Flight	4	5	216	2	low
2	3	A	Flight	2	2	183	4	low
3	4	B	Flight	3	3	176	4	medium
4	5	C	Flight	2	2	184	3	medium