Data Collection and Preprocessing Phase

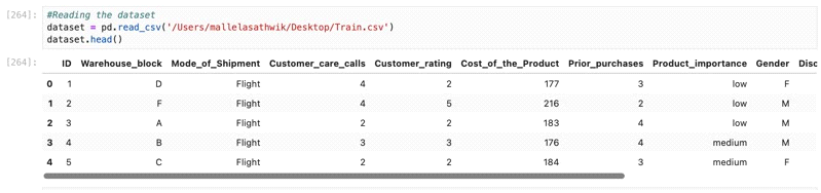
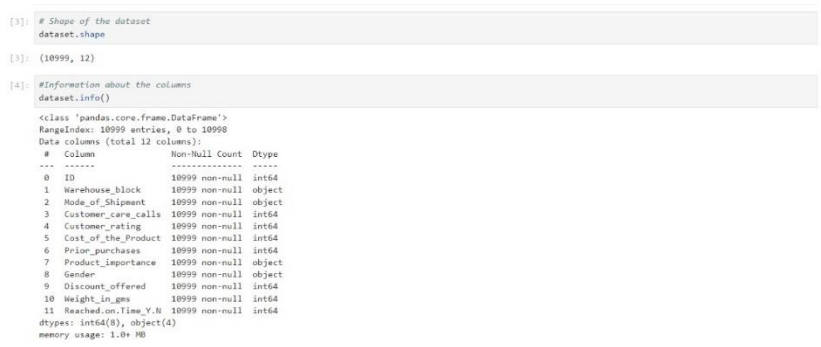
Date	04 JUNE 2024
Team ID	SWTID1720260935
Project Title	**Ecommerce Shipping Prediction Using Machine Learning**
Maximum Marks	6 Marks

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): • Real-time carrier data (shipping rates, transit times) • Weather data (location-based, impacting delivery times) • Holiday calendars (potential delays)
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)
Bivariate Analysis	We expect a positive correlation, meaning locations further away (higher distance) will tend to have longer delivery times. This helps identify factors influencing delivery times.
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	
Handling Missing 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	Disc
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	

```
[3]: # Shape of the dataset
     dataset.shape
```

[3]: (10999, 12)

```
[4]: #Information about the columns
     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
dtypes: int64(8), object(4)
memory usage: 1.0+ MB
```

Data Transformation	[30]: ``` #Splitting data into training and testing data X_train, X_test, y_train, y_test = train_test_split(X,y,test_size=0.2, random_state=42) #Scaling the data ms = MinMaxScaler() X_train = ms.fit_transform(X_train) X_test = ms.fit_transform(X_test) ``` ``` svm_model = svm.SVC(gamma='auto',C=5,kernel='rbf') svm_model.fit(X_train,y_train) y_pred = svm_model.predict(X_test) print(classification_report(y_test,y_pred)) ```
Feature Engineering	[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 Disc 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
Save Processed Data	``` 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 0 D Flight 4 2 177 3 F 1 F Flight 4 5 216 2 M 2 A Flight 2 2 183 4 M 3 B Flight 3 3 176 4 M 4 C Flight 2 2 184 3 F