ORDER PRIORITY CLASSIFICATION MODEL

A project report submitted to ICT Academy of Kerala

in partial fulfillment of the requirements

for the certification of

CERTIFIED SPECIALIST

IN

DATA SCIENCE & ANALYTICS

submitted by

Team 12

Members

Muhammed Yaseen

Hashir Haris

S. Sai Krishna

Fathima Shareena

ICT ACADEMY OF KERALA

THIRUVANANTHAPURAM, KERALA, INDIA

Feb 2022

List of Figures

List of Abbreviations

1. SMOTE: Synthetic Minority Oversampling Technique

2. EDA: Exploratory Data Analysis

3. IQR: Inter Quartile Range

4. SVC: Support Vector Classifier

5. KNN: K-Nearest Neighbor

6. CV: Cross Validation

Table of Contents

Abstract

A superstore is a very large supermarket, often selling household goods, clothes, and electrical goods, as well as food. Superstores typically charge anywhere from 15 to 45 percent less than their smaller counterparts. But the situations of covid-19 pandemic caused a negative impact on superstore with online shopping rising to the needs. we'll be using global super store data spread across geographical areas for this project containing product category, their sales, profit, discount, quantity, mode of shipment, order priority etc. With so provided data we'll be looking forward to preprocessing the data and conduct exploratory data analysis discovering patterns and trends like customer analysis and product analysis. Finally creating ML models classifying order priority of customers helping to make better business decisions and fine tuning it to achieve maximum accuracy.

1. Problem Definition

1.1 Overview

In this project, a four step procedure is followed leading to the model classifying Order Priority using global superstore dataset. First the data is acquired, collected and understood. This data undergoes preliminary analysis which includes univariate and bivariate analysis. In the third stage, data undergoes pre-processing taking care of missing and erroneous values in data set. This stage is finally followed by final stage where data is divided into training and test set and various ML models are built and result are evaluated

1.2 Problem Statement

The problem is defined to create a Machine Learning Classification Model prioritizing orders to be High, Medium, Critical and low for processing of order resulting in customer satisfaction and better business growth.

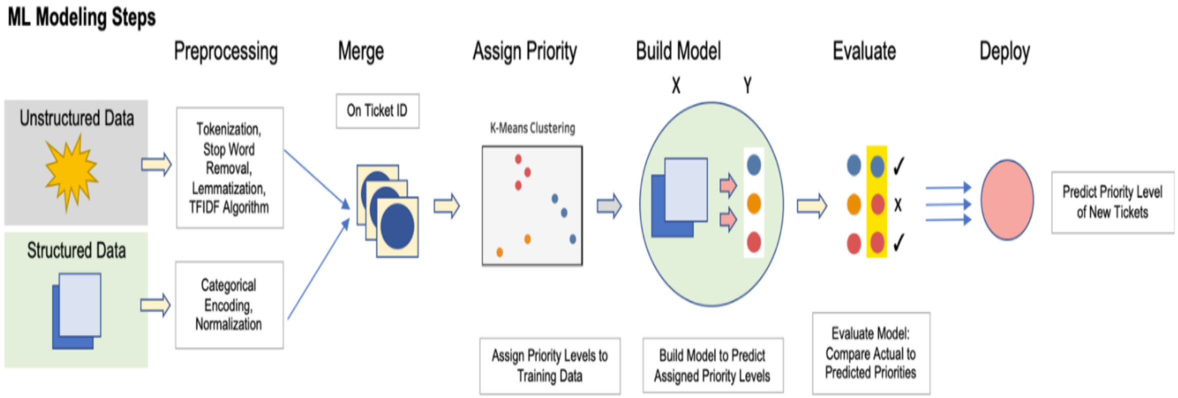
.

2. Introduction

In the modern world, due to the wake of pandemic, Shopping online is currently the need of the hour ensuring safety, incredible offers with well and good shipping. With covid-19 measures becoming lenient, ever growing future demands needs to be addressed by a good order priority processing to be more competent in the market. Order priority is a sequence of actions in business to fulfill a customer purchase. Its usually categorized as tiers. The dataset we are using is built with various dependent and independent variables in the forms of product attributes, data gathered within geographical areas and by the means of customer. The data will be thereafter visualized discovering trends and patterns and refined in order to get accurate classification model and gather interesting results that shed light on our knowledge with respect to the task's data. As good sales are the life of every organization, prioritization of order ensures it. In this project we will be using supervised classification model like Random forest classification model and decision tree classification model to get accurate results.

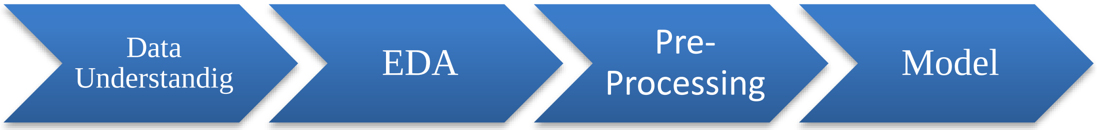
3. Literature Survey

This project is based on literature on ‘A learned approach on priority setting and classification’ published by Chenoa, a solution partner with unqork the industry pioneer no code enterprise application platform. The literature discusses on assignment of priority to business tasks and issues taking the case of IT Support ticketing system. The assignment priority is based on assessment of business impact. The literature focuses on discussing on building IT Ticket priority assignment system with an structured and unstructured data followed by categorical encoding, Normalization and final leading to model assigning priority levels such as low, medium, high etc using various clustering techniques. Challenges faced by them while prioritizing is also discussed by them like predicting wrong priority level even if observations are similar within each cluster and finally concluding by necessitating customer satisfaction in todays world leading to downstream improvements in key metrics such as NPS and help provide quantifiable justification for investing in AI.


ML Modeling Steps
Preprocessing
Merge
Assign Priority
Build Model
Evaluate
Deploy
Unstructured Data
Structured Data
Tokenization, Stop Word Removal, Lemmatization, TFIDF Algorithm
Categorical Encoding, Normalization
On Ticket ID
K-Means Clustering
X
Y
✓
X
✓
Predict Priority Level of New Tickets
Assign Priority Levels to Training Data
Build Model to Predict Assigned Priority Levels
Evaluate Model: Compare Actual to Predicted Priorities

4. Methodology

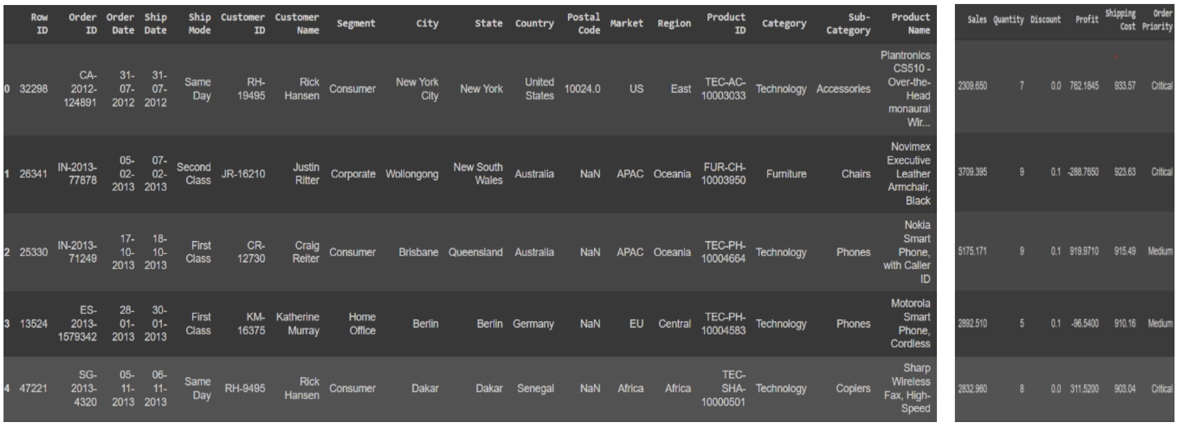
The steps to be followed in this work right from data set preparation to obtaining result are represented:



4.1 Data Understanding

Global superstore data is a customer centric data set, which has the data of all the orders placed through different market and vendors starting from year 2011 till 2015,hailing from 147 different countries. The dataset which can be found at https://www.kaggle.com/apoorvaappz/global-super-store-dataset consists of 51k rows and 24 columns of data on purchases made around the world by the segments, consumers, corporate and home office.

The data set looks like as shown in fig on using head() function on the dataset variable.

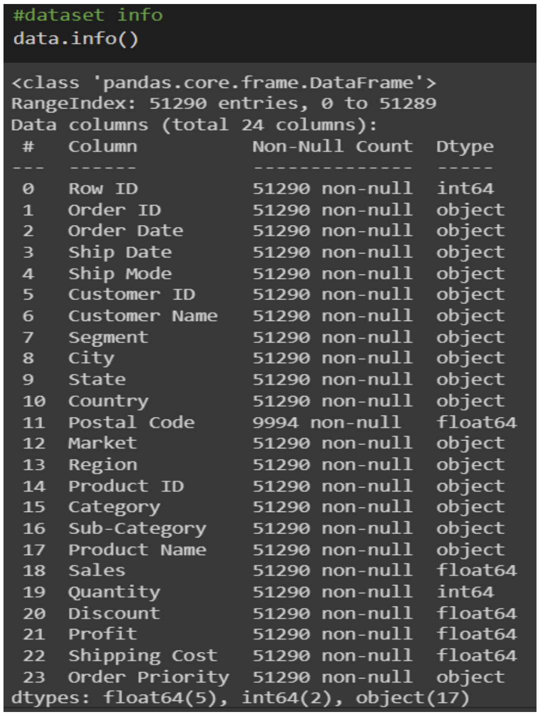


	Row ID	Order ID	Order Date	Ship Date	Ship Mode	Customer ID	Customer Name	Segment	City	State	Country	Postal Code	Market	Region	Product ID	Category	Sub-Category	Product Name
0	32298	CA-2012-124891	31-07-2012	31-07-2012	Same Day	RH-19495	Rick Hansen	Consumer	New York City	New York	United States	10024.0	US	East	TEC-AC-10003033	Technology	Accessories	Plantronics CS510 - Over-the-Head monaural Wir...
1	26341	IN-2013-77878	05-02-2013	07-02-2013	Second Class	JR-16210	Justin Ritter	Corporate	Wollongong	New South Wales	Australia	NaN	APAC	Oceania	FUR-CH-10003950	Furniture	Chairs	Novimex Executive Leather Armchair, Black
2	25330	IN-2013-71249	17-10-2013	18-10-2013	First Class	CR-12730	Craig Reiter	Consumer	Brisbane	Queensland	Australia	NaN	APAC	Oceania	TEC-PH-10004664	Technology	Phones	Nokia Smart Phone, with Caller ID
3	13524	ES-2013-1579342	28-01-2013	30-01-2013	First Class	KM-16375	Katherine Murray	Home Office	Berlin	Berlin	Germany	NaN	EU	Central	TEC-PH-10004583	Technology	Phones	Motorola Smart Phone, Cordless
4	47221	SG-2013-4320	05-11-2013	06-11-2013	Same Day	RH-9495	Rick Hansen	Consumer	Dakar	Dakar	Senegal	NaN	Africa	Africa	TEC-SHA-10000501	Technology	Copiers	Sharp Wireless Fax, High-Speed

Sales	Quantity	Discount	Profit	Shipping Cost	Order Priority
2309.650	7	0.0	762.1845	933.57	Critical
3709.395	9	0.1	-288.7650	923.63	Critical
5175.171	9	0.1	919.9710	915.49	Medium
2892.510	5	0.1	-96.5400	910.16	Medium
2832.960	8	0.0	311.5200	903.04	Critical

Fig 1: Raw dataset

The data set consists of various data types from integer to float to object as shown in Fig

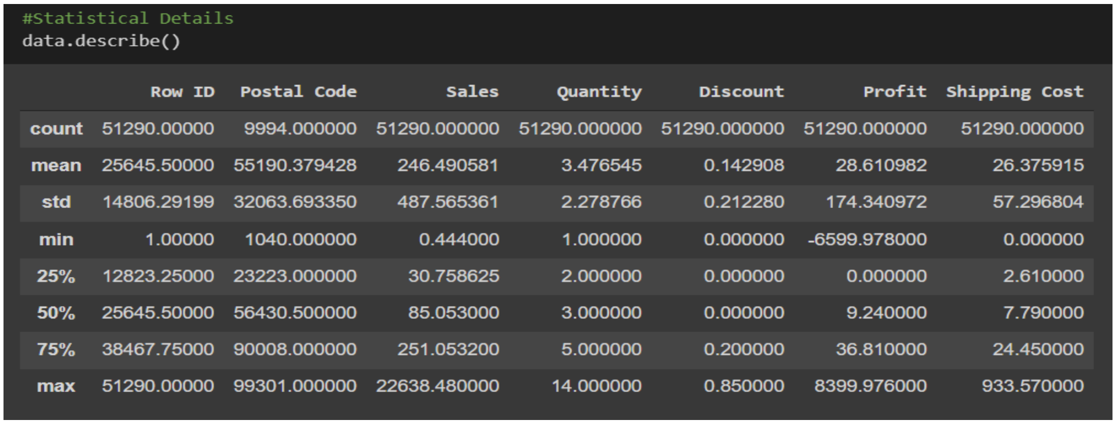


```
#dataset info
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 51290 entries, 0 to 51289
Data columns (total 24 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Row ID          51290 non-null  int64
 1   Order ID        51290 non-null  object
 2   Order Date      51290 non-null  object
 3   Ship Date       51290 non-null  object
 4   Ship Mode       51290 non-null  object
 5   Customer ID     51290 non-null  object
 6   Customer Name   51290 non-null  object
 7   Segment         51290 non-null  object
 8   City            51290 non-null  object
 9   State           51290 non-null  object
 10  Country         51290 non-null  object
 11  Postal Code     9994 non-null   float64
 12  Market          51290 non-null  object
 13  Region          51290 non-null  object
 14  Product ID      51290 non-null  object
 15  Category        51290 non-null  object
 16  Sub-Category    51290 non-null  object
 17  Product Name    51290 non-null  object
 18  Sales           51290 non-null  float64
 19  Quantity        51290 non-null  int64
 20  Discount        51290 non-null  float64
 21  Profit          51290 non-null  float64
 22  Shipping Cost   51290 non-null  float64
 23  Order Priority  51290 non-null  object
dtypes: float64(5), int64(2), object(17)
```

Fig 2: Info of columns

Various factors important by statistical means like mean, standard deviation, median, count of values and maximum value etc. are shown in Fig for numerical variables of our dataset.

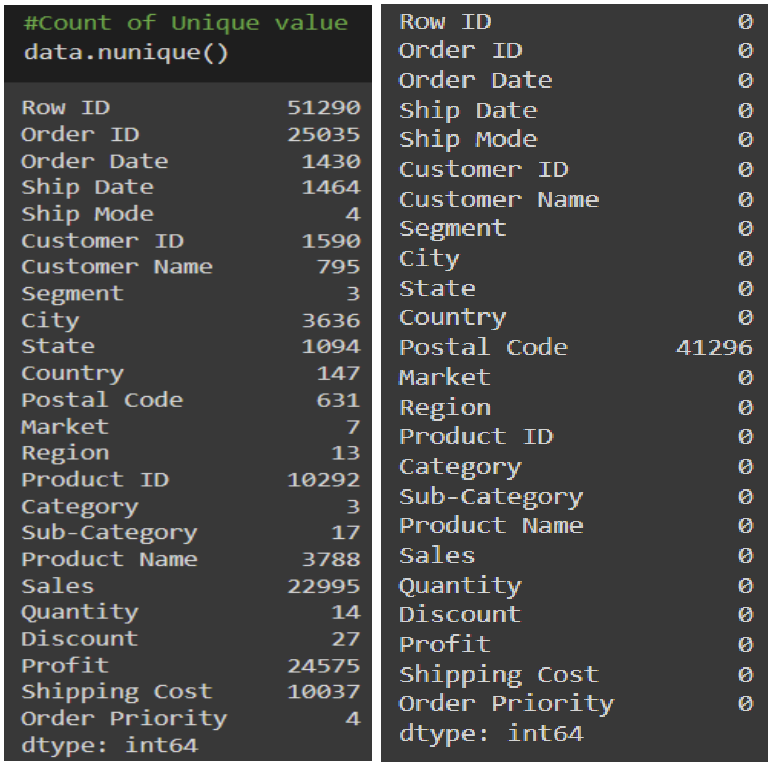


```
#Statistical Details
data.describe()
```

	Row ID	Postal Code	Sales	Quantity	Discount	Profit	Shipping Cost
count	51290.00000	9994.000000	51290.000000	51290.000000	51290.000000	51290.000000	51290.000000
mean	25645.50000	55190.379428	246.490581	3.476545	0.142908	28.610982	26.375915
std	14806.29199	32063.693350	487.565361	2.278766	0.212280	174.340972	57.296804
min	1.00000	1040.000000	0.444000	1.000000	0.000000	-6599.978000	0.000000
25%	12823.25000	23223.000000	30.758625	2.000000	0.000000	0.000000	2.610000
50%	25645.50000	56430.500000	85.053000	3.000000	0.000000	9.240000	7.790000
75%	38467.75000	90008.000000	251.053200	5.000000	0.200000	36.810000	24.450000
max	51290.00000	99301.000000	22638.480000	14.000000	0.850000	8399.976000	933.570000

Fig 3: Description of data

Null values is found in postal code column which had to be dealt during pre-processing stage. Too much of unique value counts is found in columns of data set.



```
#Count of Unique value
data.nunique()

Row ID              51290
Order ID            25035
Order Date           1430
Ship Date            1464
Ship Mode               4
Customer ID          1590
Customer Name         795
Segment                 3
City                 3636
State                1094
Country               147
Postal Code           631
Market                  7
Region                 13
Product ID          10292
Category                3
Sub-Category           17
Product Name         3788
Sales               22995
Quantity               14
Discount               27
Profit              24575
Shipping Cost       10037
Order Priority          4
dtype: int64
```

```
Row ID                0
Order ID              0
Order Date            0
Ship Date             0
Ship Mode             0
Customer ID           0
Customer Name         0
Segment               0
City                  0
State                 0
Country               0
Postal Code       41296
Market                0
Region                0
Product ID            0
Category              0
Sub-Category          0
Product Name          0
Sales                 0
Quantity              0
Discount              0
Profit                0
Shipping Cost         0
Order Priority        0
dtype: int64
```

Fig 4: Unique values and number of null values in data set

skew function is processed to find out the skewness of distribution of non-categorical variables. Negative skewness is found out for Row id and Postal code while all other showing positive skewness.

4.2 Exploratory Data Analysis

EDA is generally an approach to analyze data using visualization to discover trends and patterns within variable and between variables. We have done Univariate analysis and Bivariate analysis under EDA

4.2.1 Univariate Analysis

In univariate analysis, only single variable is analyzed describing data and patterns within itself. The following are the Univariate analysis done from the dataset:

- Distribution plot of Sales: Right Skewed distribution has been observed.
- Distribution plot of Quantity: Right skewed distribution has been observed.

- Distribution plot of profit: from the plot distribution is found to be Right skewed
- Distribution plot of Discount: Right Skewed skewed distribution is observed

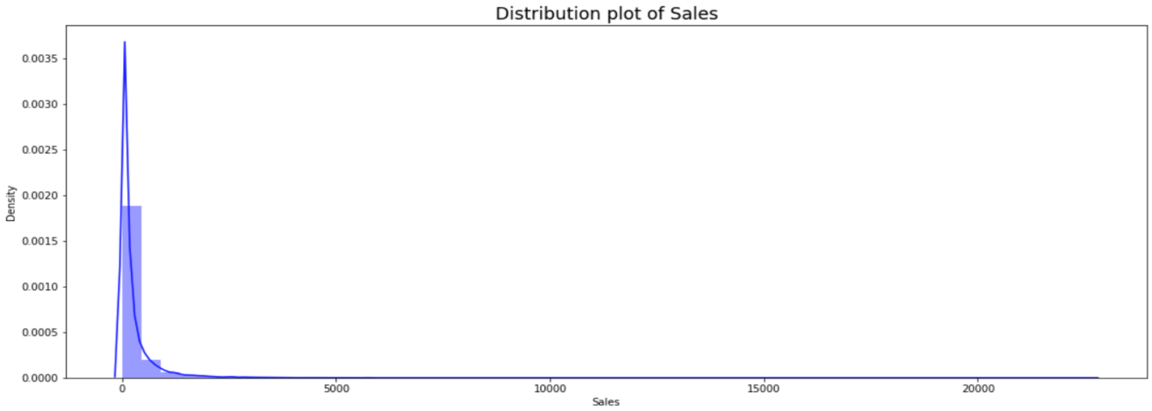


Fig 5: Distribution plot of sales

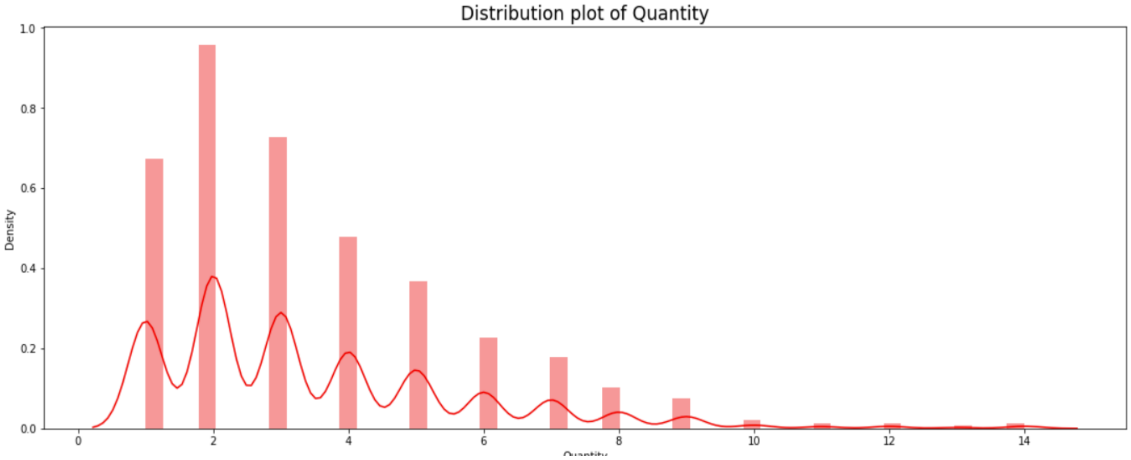


Fig 6: Distribution plot of Quantity

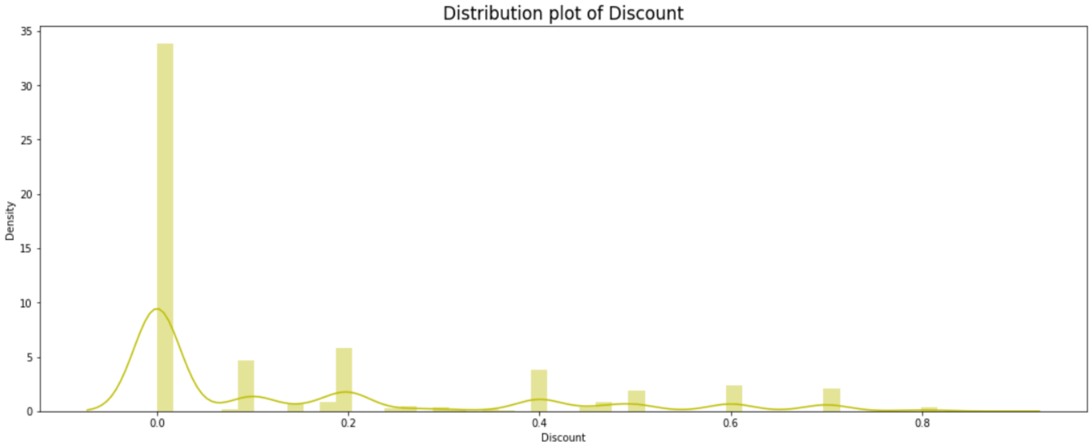


Fig 7: Distribution plot of Discount

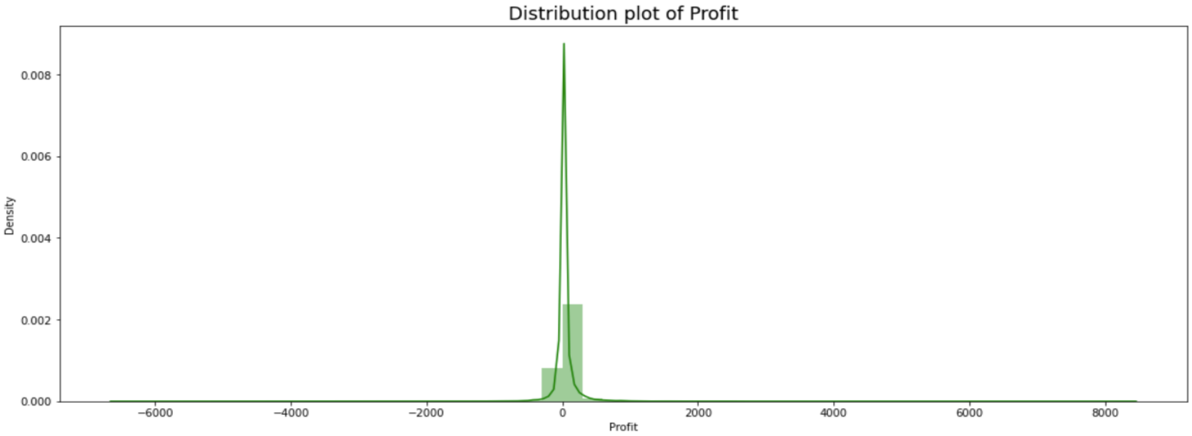


Fig 8: Distribution plot of Profit

- Count plot and pie chart of Segment distribution: Consumer tops in the Segment (51.7% and count=29518) with home office being least (18.2% and count=9343).

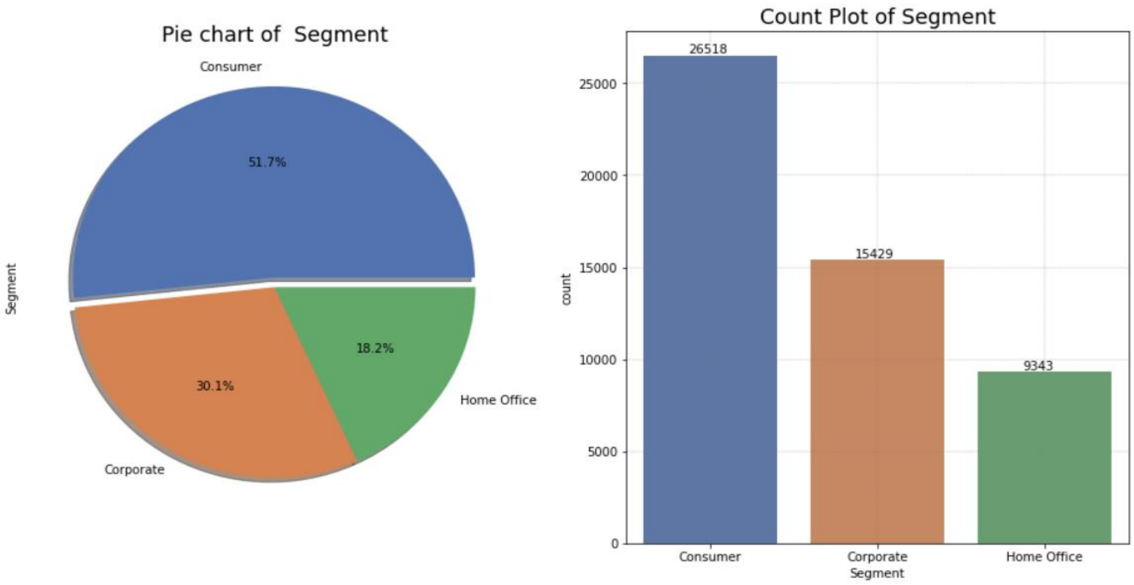


Fig 9: Count plot and Distibution plot of segment

- Count plot and pie chart of Category distribution: Office supplies tops in the segment (61% and count=31273) with furniture being least (19.3% and count=9676).

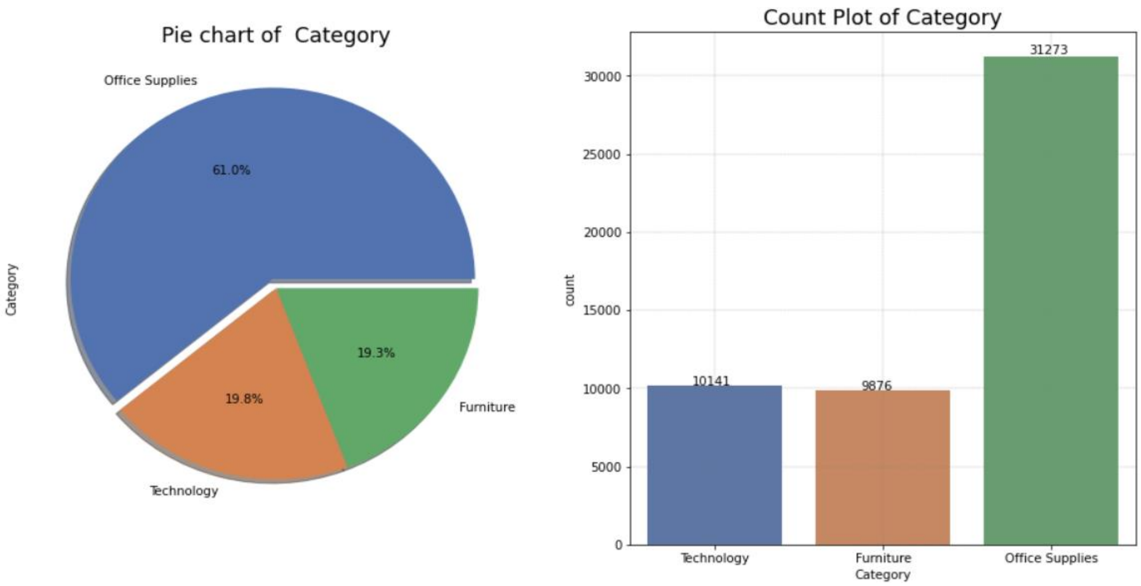


Fig 10: Count plot and pie chart of Category distribution

- Count plot and pie chart of Region distribution: Central tops in the segment (21.7%b and count=11117) with Canada being least (0.7% and count=384).

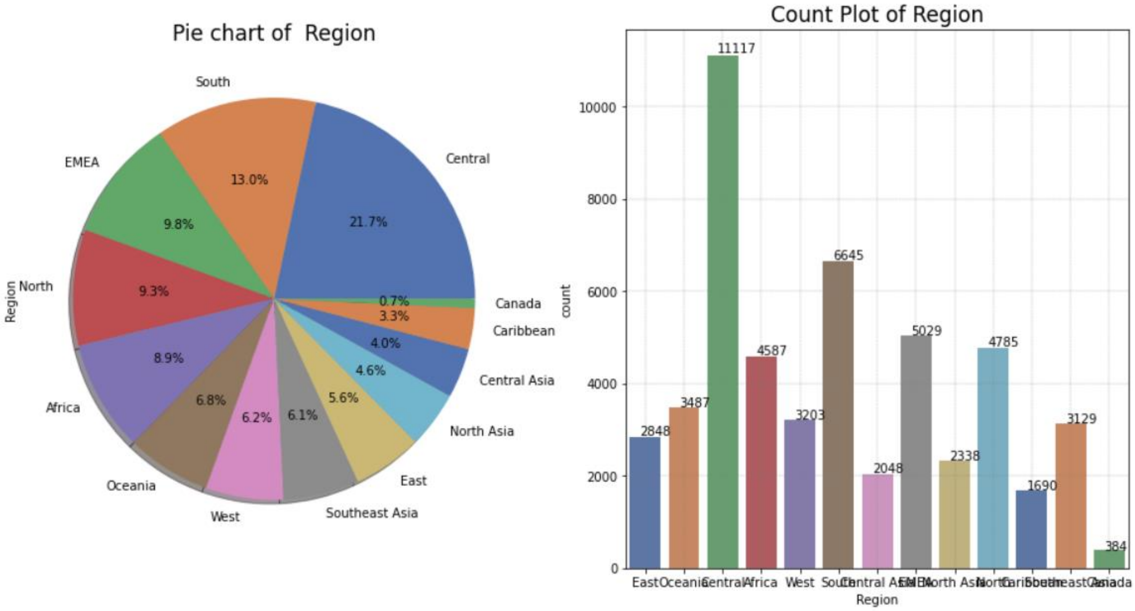


Fig 11: Count plot and pie chart of Region distribution

- <u>Count plot of Sub-Category Distribution</u>: Blinders tops in sub-category distribution with count of 6152 and tables with least count of 861

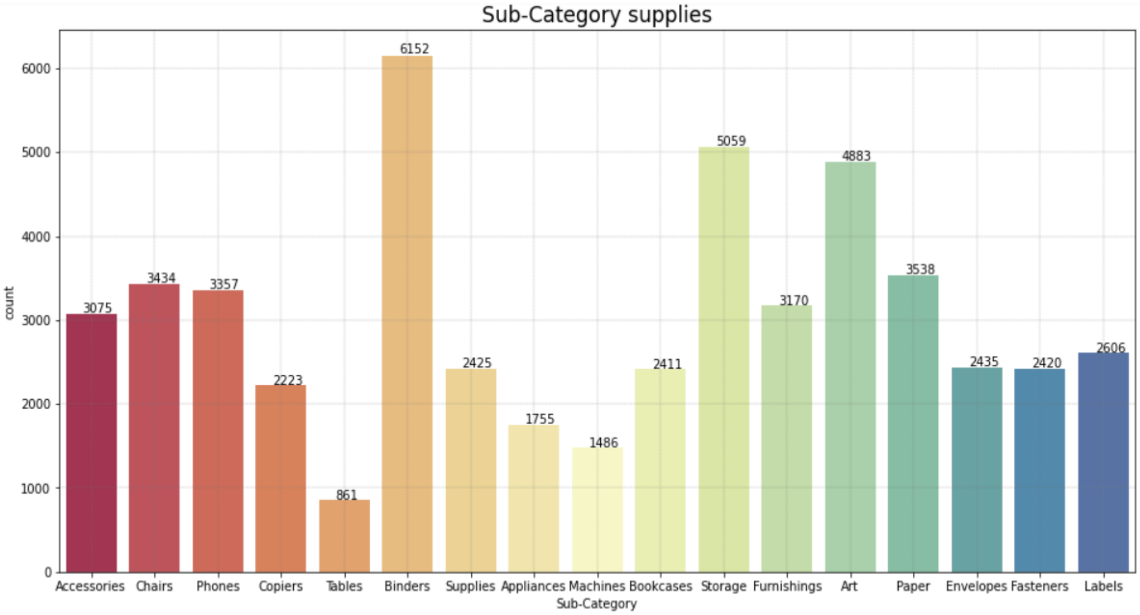


Fig 12: Count plot of Sub-Category Distribution

- Count plot and Pie Chart of Ship Mode: standard class tops in Ship Mode Distribution (60% and count=30775) and least is same day (5.3% and count=2701)

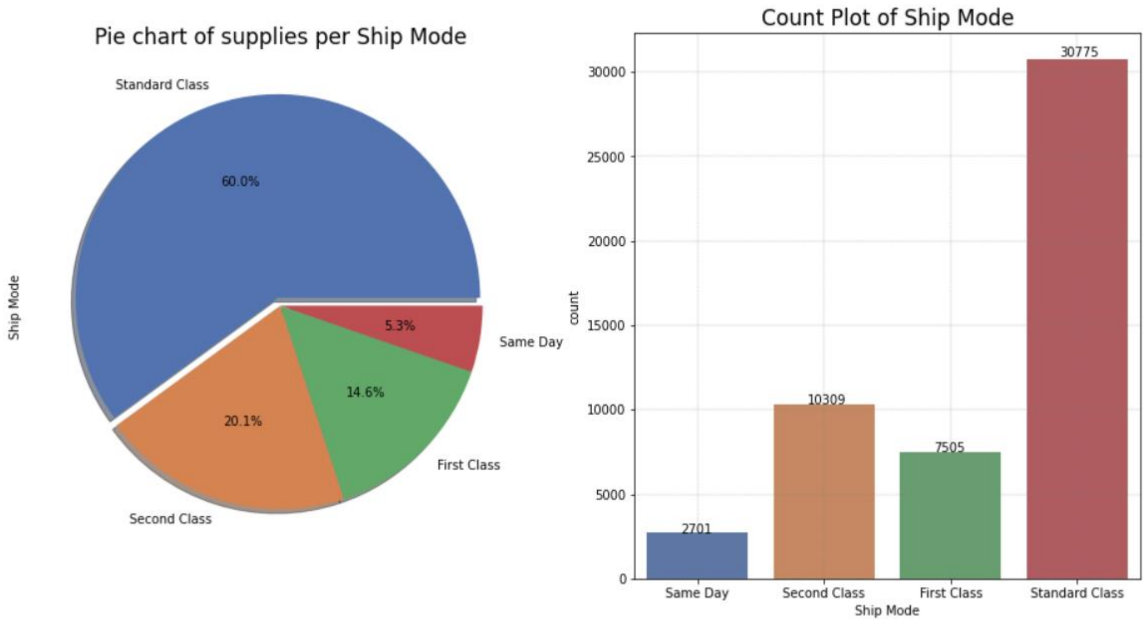


Fig 13: Count plot and Pie Chart of Ship Mode

- Box plot of non-categorical variables like Sales, Quantity, Discount and Shipping cost: Outliers are detected in the plot which have to be dealt during preprocessing

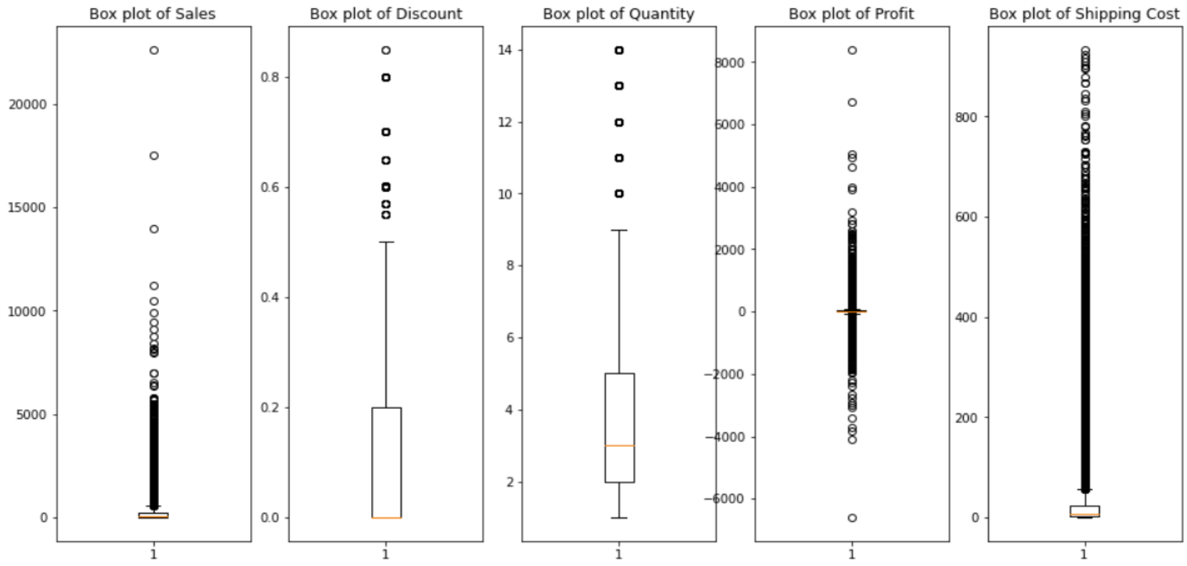


Fig 14: Box plot of Sales, Quantity, Discount and Shipping cost

4.2.2 Bivariate Analysis

Bivariate Analysis is used to determine relationship between 2 variables. The following are the Bivariate Analysis done from Dataset:

- ➢ <u>Line plot of Discount vs Profit</u>: from the plot pattern of as the discount increases profit decreases is observed.

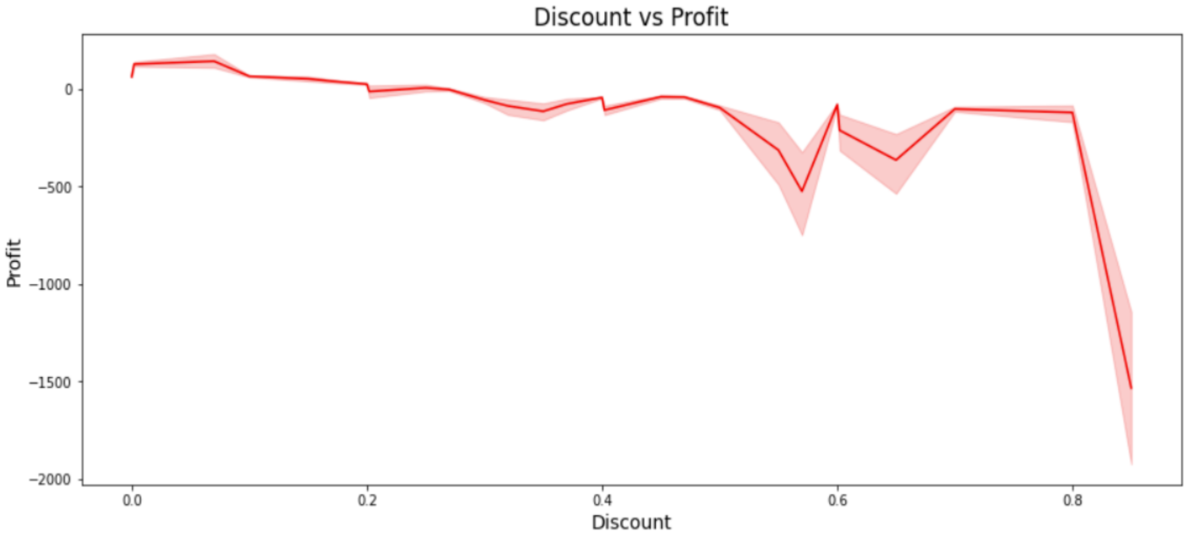


Fig 15: Line plot of Discount vs Profit

- ➢ <u>Line plot of Quantity vs Profit</u>: Both Quantity and Sales increases w.r.t each other has been observed

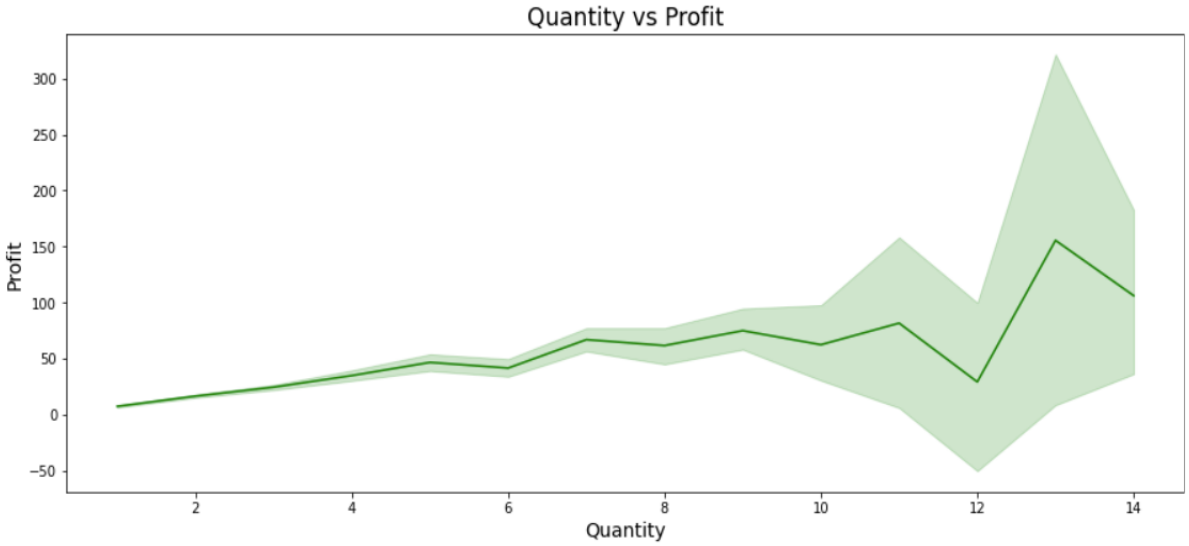


Fig 16: Line plot of Quantity vs Profit

- Scatter plot of Discount vs Sales: Datapoint are to much scattered resulting in no correlation observation

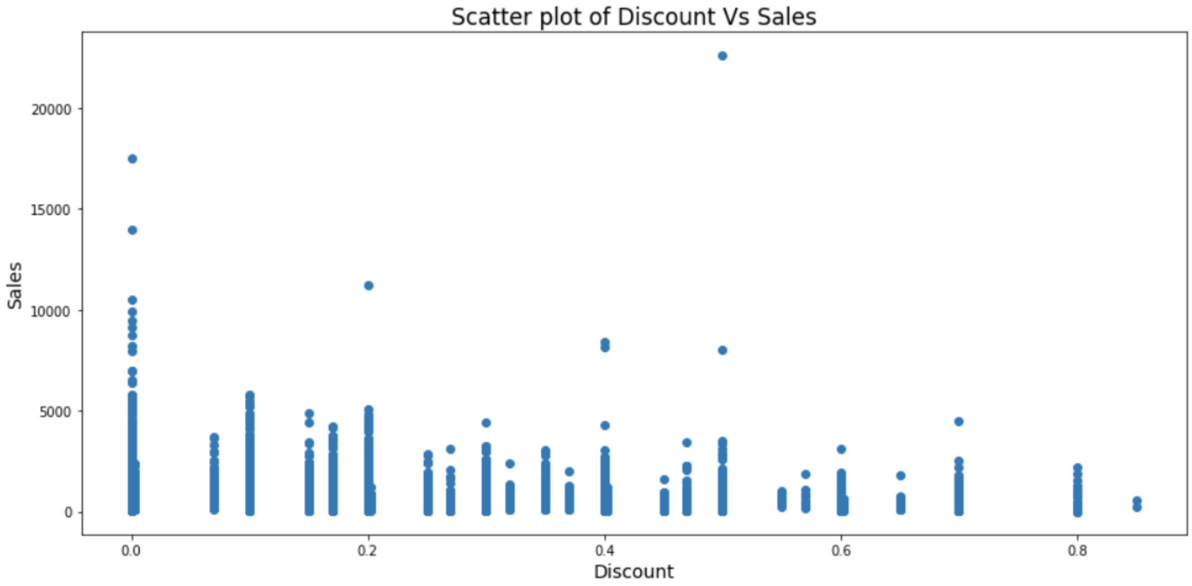


Fig 17: Scatter plot of Discount vs Sales

- Scatter plot of Profit vs Sales: Positive correlation has been observed the two variables resulting in their relationship moving in same direction.

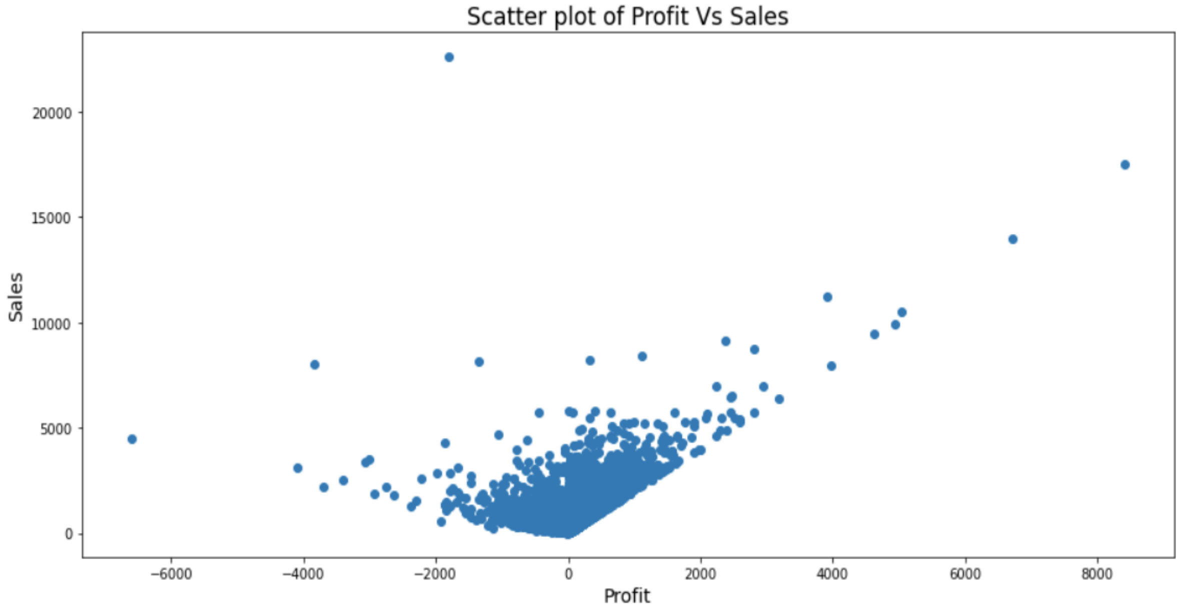


Fig 18: Scatter plot of Profit vs Sales

- Bar plot of Ship Mode wise Total Sales: it has been observed that standard class top the ship mode in terms of total sales with same day being the least.

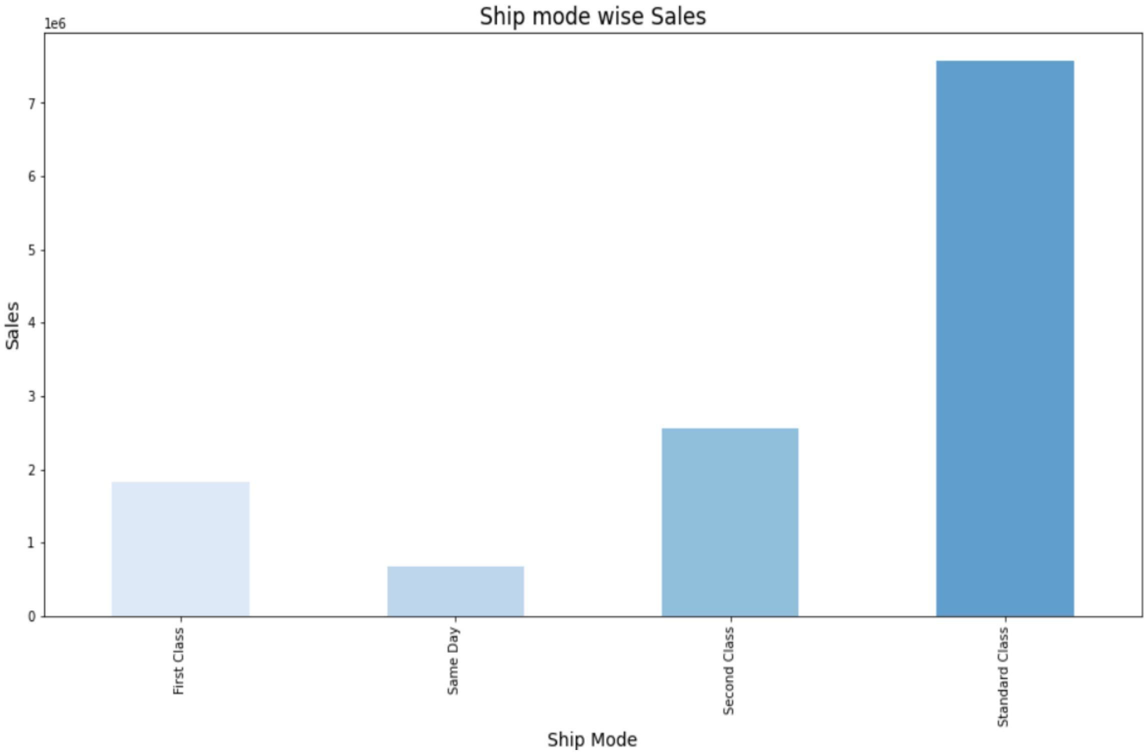


Fig 19: Bar plot of Ship Mode wise Total Sales

- Bar plot of Sub-Category Distribution within Category:

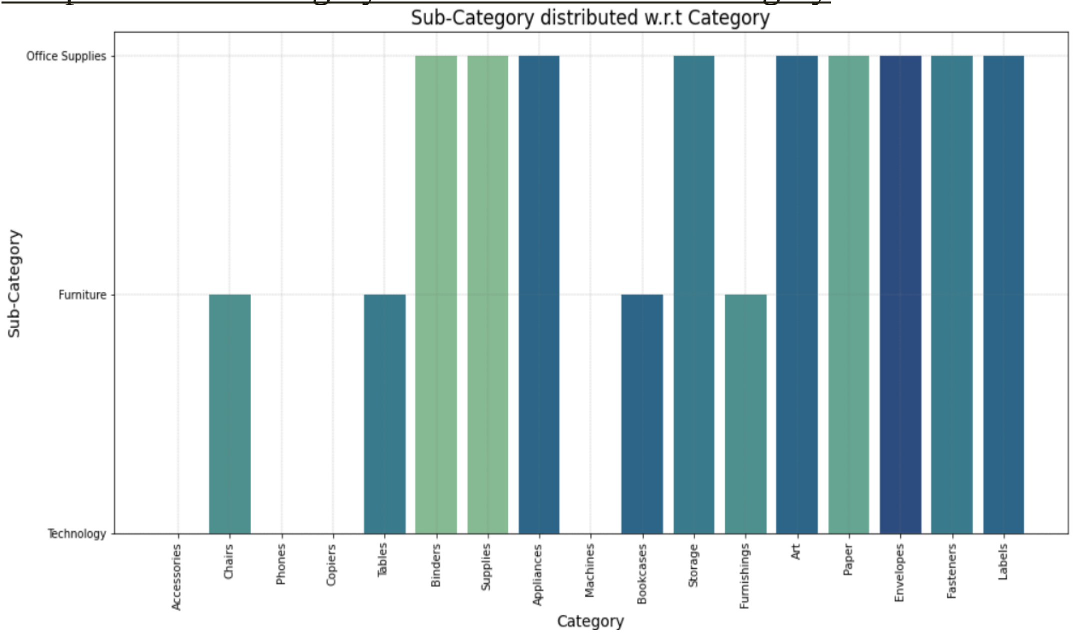


Fig 20: Bar plot of Sub-Category Distribution within Category

- Bar plot Sub-Category wise Total Sales and Profit: total highest sales and profit has been observed in Phones and copiers in Sub-Category respectively whereas least sales and loss has been observed in labels and tables respectively.

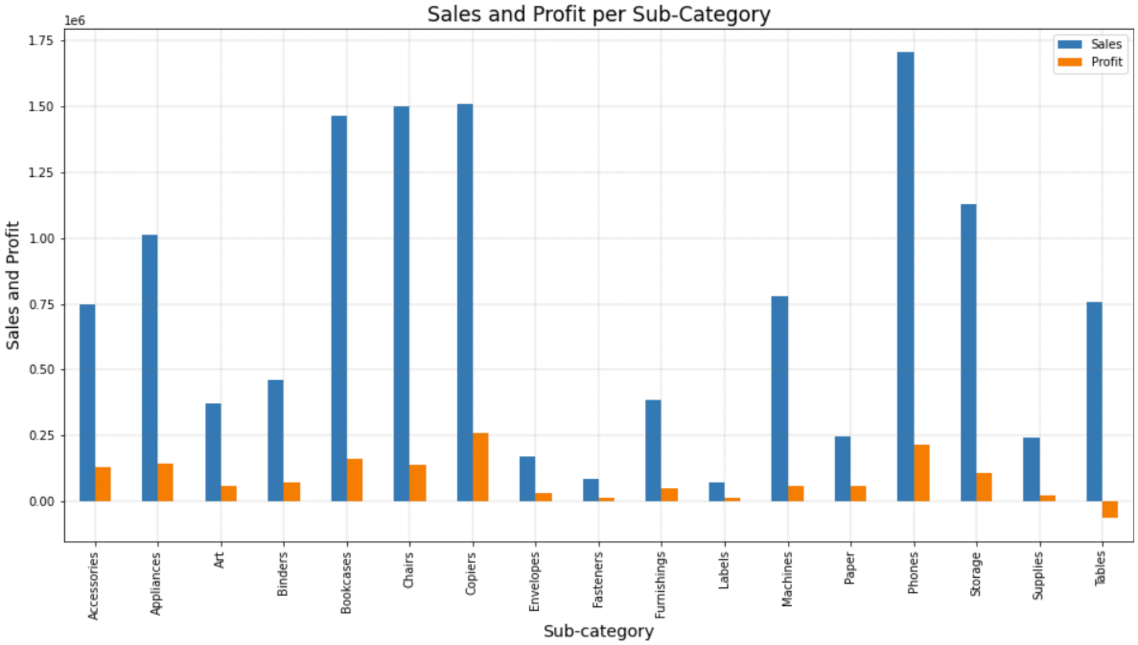


Fig 21: Bar plot Sub-Category wise Total Sales and Profit

- Bar plot Sub-Category Wise Total Profit w.r.t Ship Mode: Highest profit is observed in copiers Sub-Category with standard class Ship mode and loss is observed in tables with only profit when shipped same day.

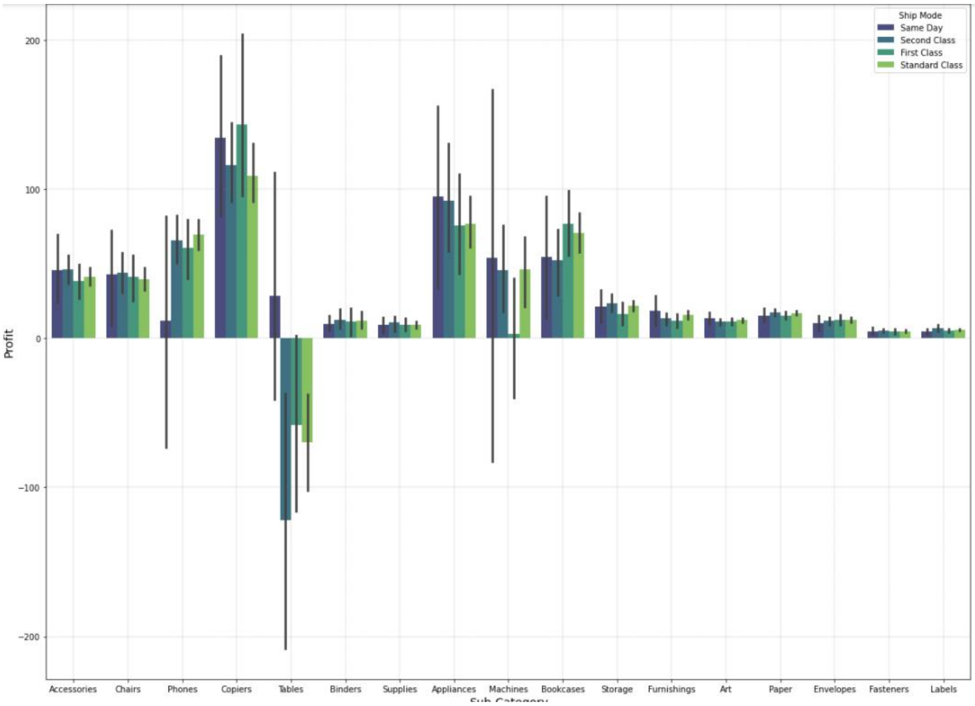


Fig 22: Bar plot Sub-Category Wise Total Profit w.r.t Ship Mode

- <u>Bar plot Quantity sold per Sub-Category</u>: highest quantity sold is blinders and least is found out to be tables

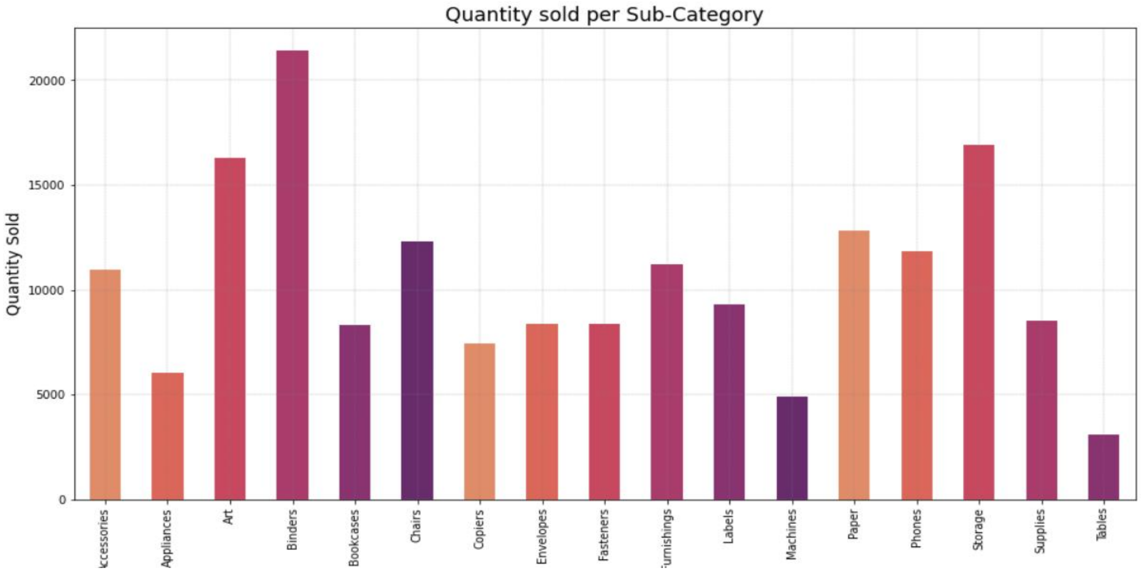
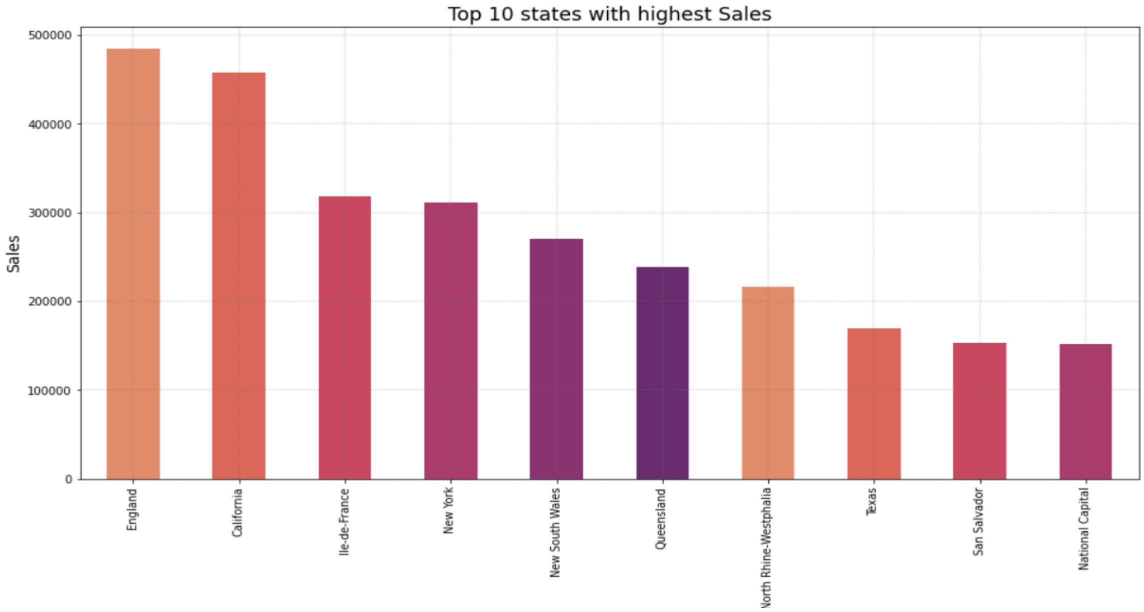
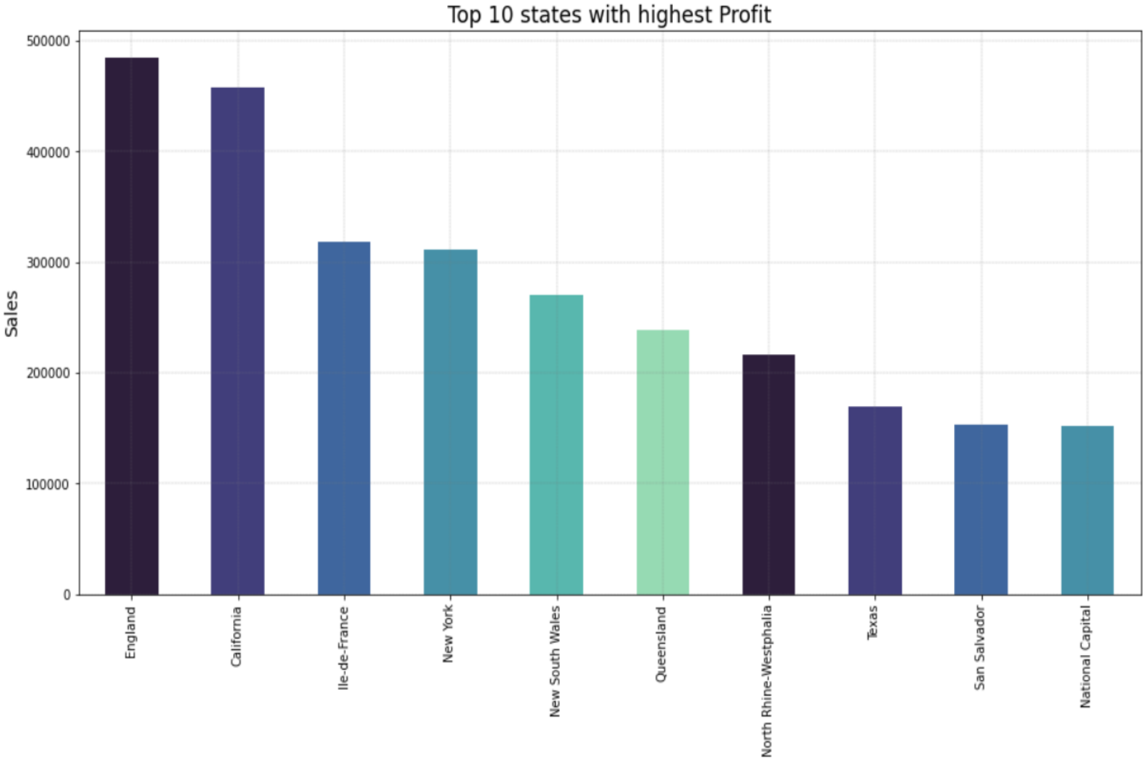


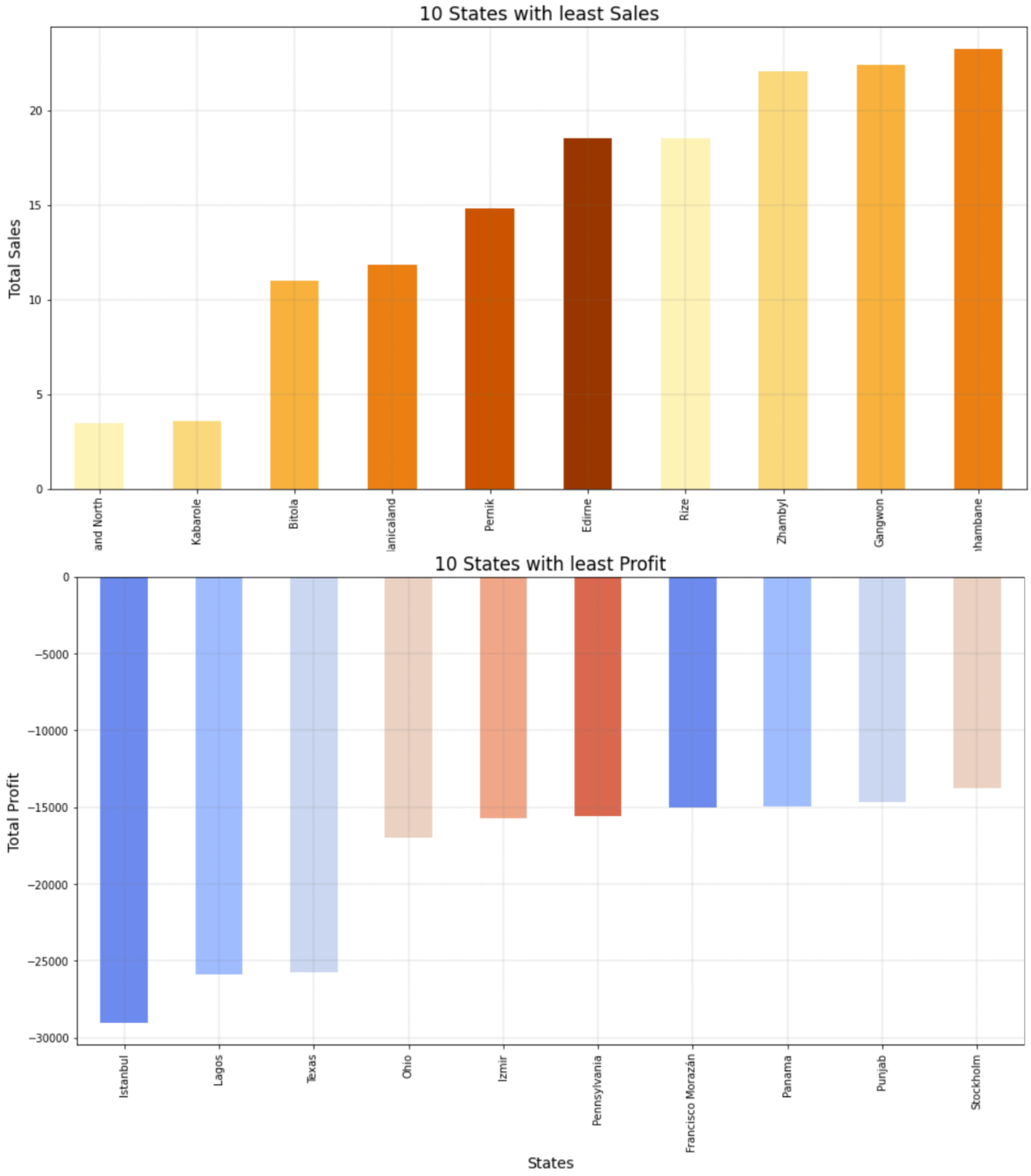
Fig 23: Bar plot Quantity sold per Sub-Category

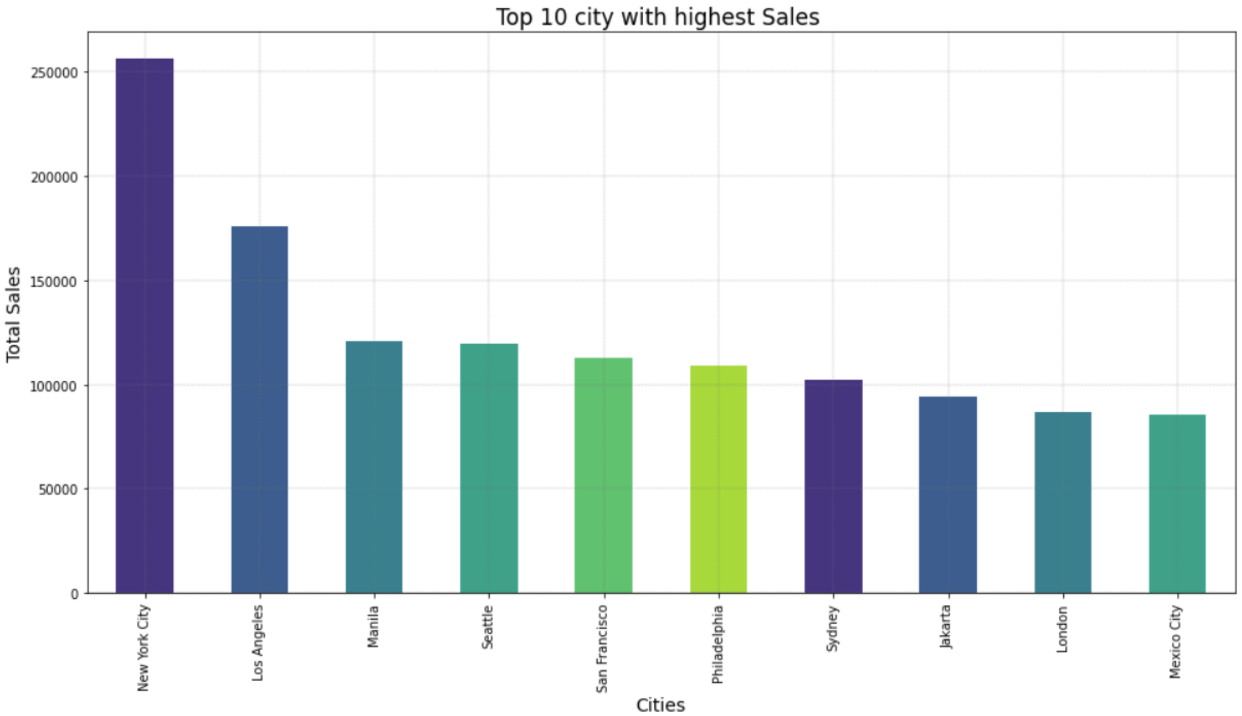
- <u>Bar plot of Total Sale analysis w.r.t State and City</u>:
  - Highest total sales and profit is observed in State of England
  - Least total sales observed is observed in State of Matabeleland North

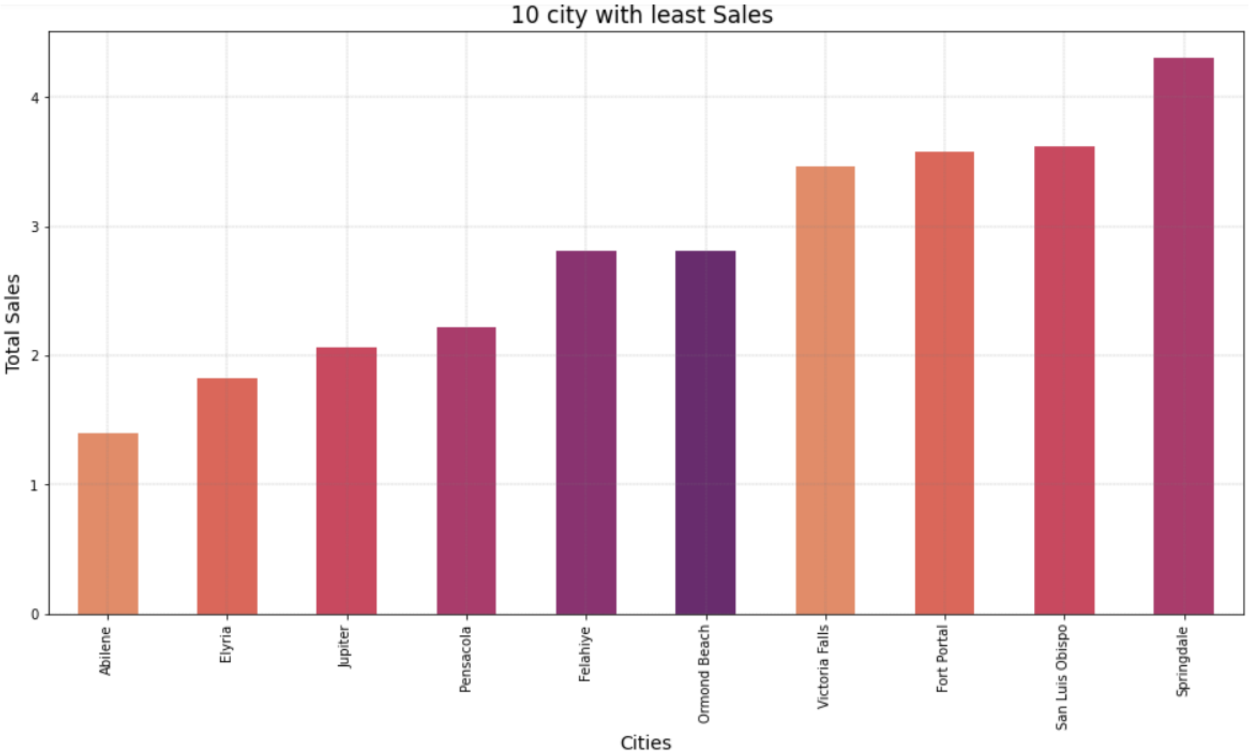
- Huge loss is observed in state of Istanbul
- City with highest total sales and profit is found out to be New York
- City with least sales is found out to be Abilene and huge loss is observed in Lagos






10 States with least Sales
Total Sales
0
5
10
15
20
and North
Kabarole
Bitola
lanicaland
Pernik
Edirne
Rize
Zhambyl
Gangwon
ıhambane
10 States with least Profit
Total Profit
0
−5000
−10000
−15000
−20000
−25000
−30000
Istanbul
Lagos
Texas
Ohio
Izmir
Pennsylvania
Francisco Morazán
Panama
Punjab
Stockholm
States


Top 10 city with highest Sales
Total Sales
250000
200000
150000
100000
50000
0
New York City
Los Angeles
Manila
Seattle
San Francisco
Philadelphia
Sydney
Jakarta
London
Mexico City
Cities


10 city with least Sales
Total Sales
4
3
2
1
0
Abilene
Elyria
Jupiter
Pensacola
Felahiye
Ormond Beach
Victoria Falls
Fort Portal
San Luis Obispo
Springdale
Cities

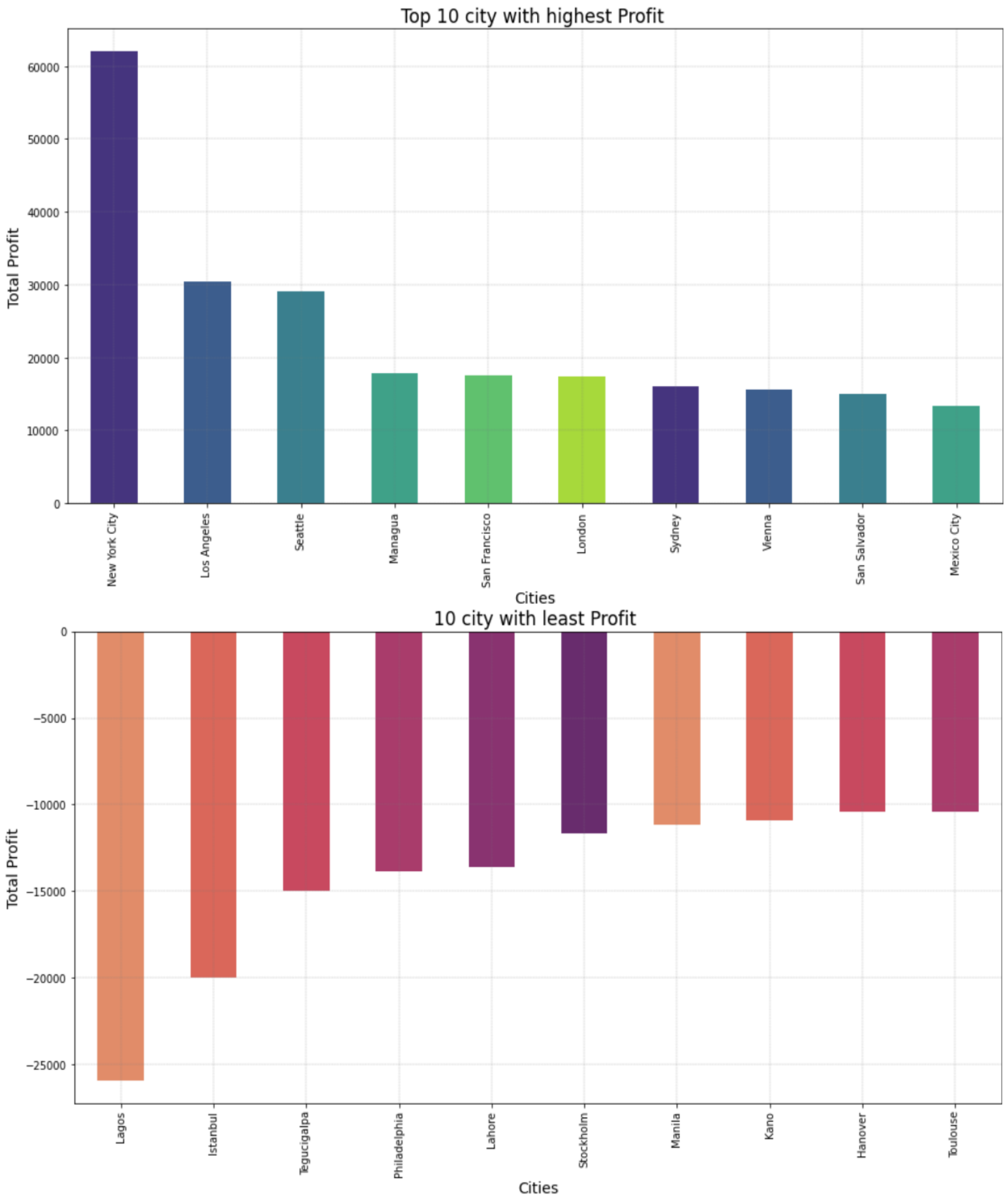


Fig 24: Bar plot of Total Sale analysis w.r.t State and City

- <u>Bar plot of Segment Wise Total Sales and Profit</u>: most sales and profit in Consumer Segment and least in Home office Segment is observed

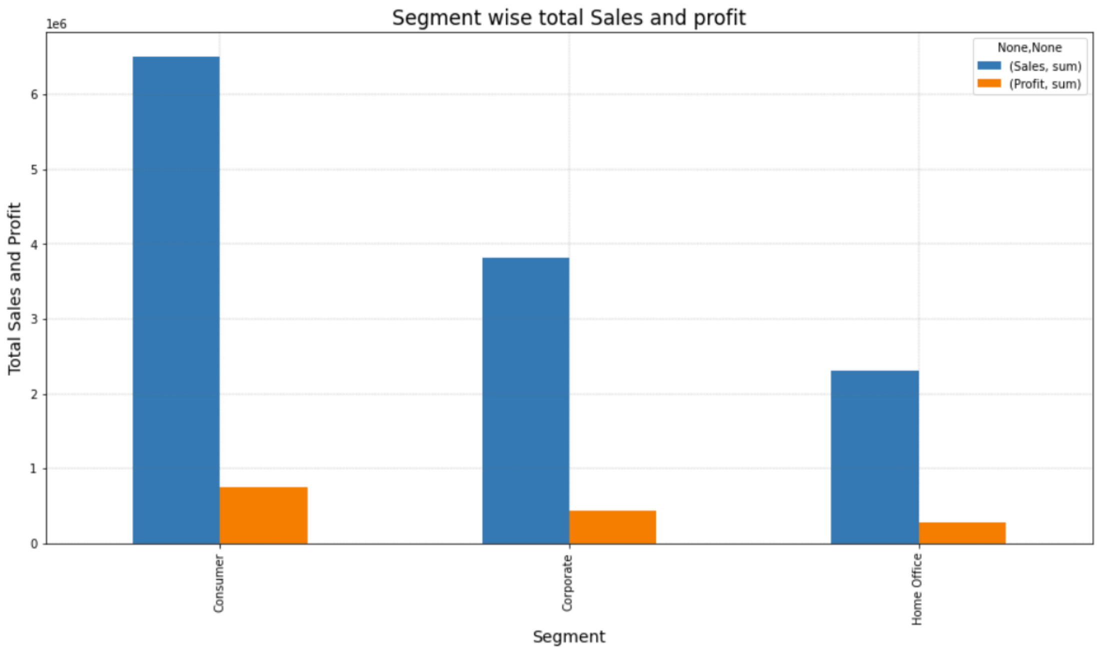


Fig 25: Bar plot of Segment Wise Total Sales and Profit

- Category Wise total Sales and Profit: most sales and profit is found in Technology category and least sales in Office supplies and least profit in furniture has been found.

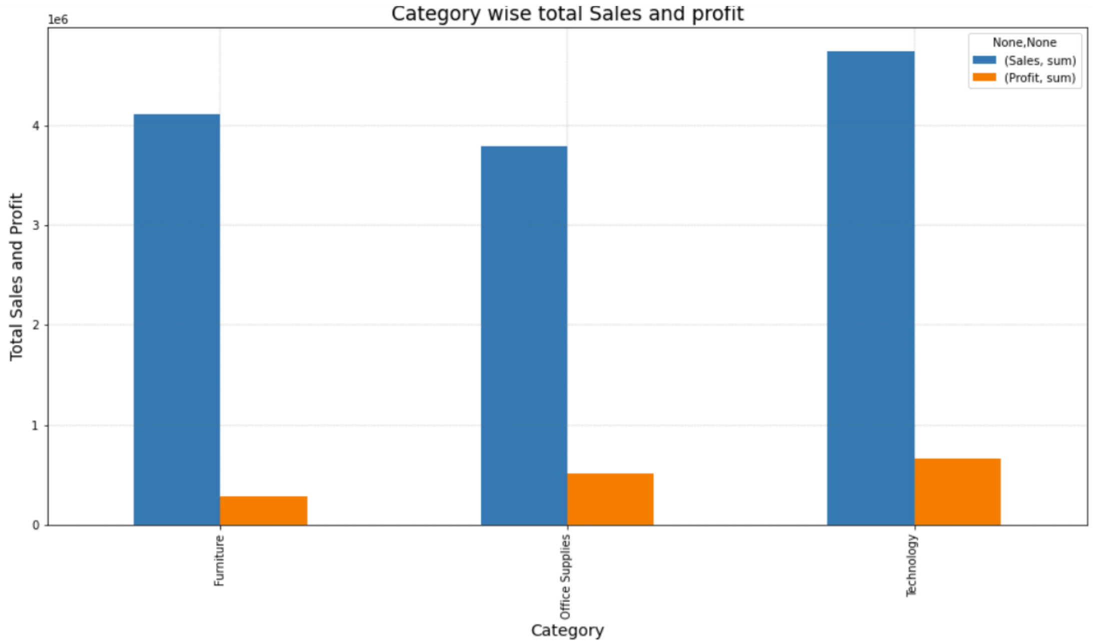


Fig 26: Bar plot ofCategory Wise total Sales and Profit

- Bar plot of Category wise total cost: Added a new column cost by subtracting sum of profit and shipping cost from sales and plotted. It has been observed that highest total cost is that of Technology and least of Office supplies.

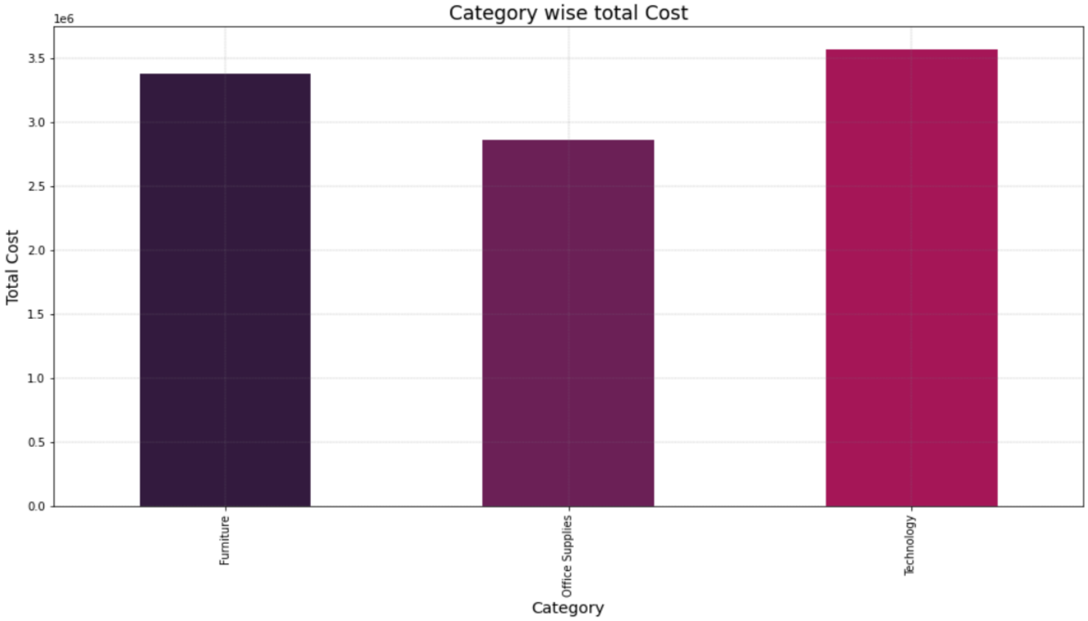


Fig 26: Bar plot of Category wise total cost

➢ <u>Bar plot of Sub-Category wise Total Cost</u>: Highest total cost is observed for tables in Corporate Segment and least for labels equal in all segment.

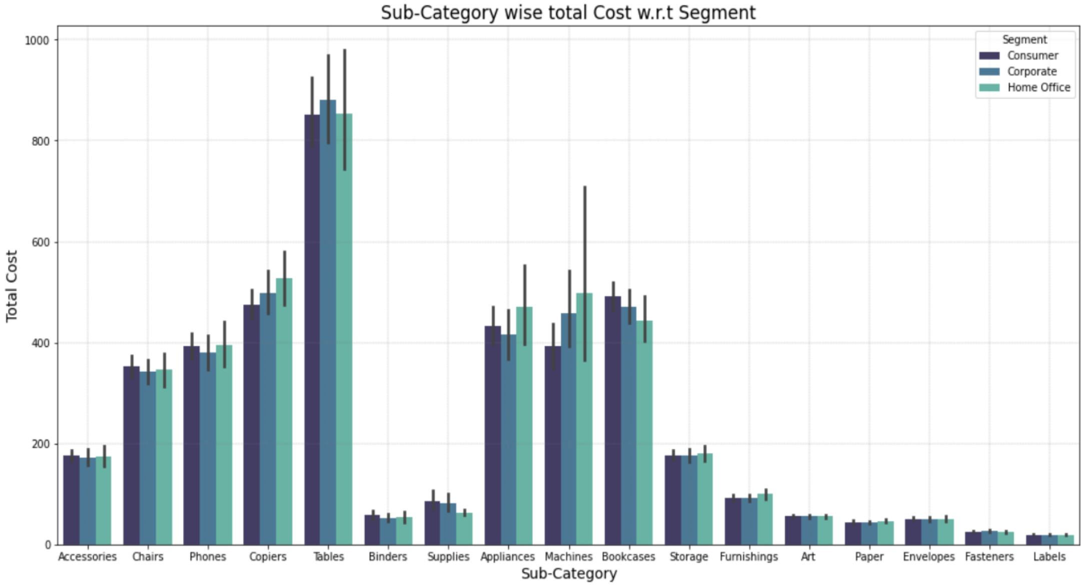


Fig 27: Bar plot of Category wise total cost

➢ <u>Pie Chart of Ship Mode wise total Sales, Discount and Profit</u>:

- Highest total sales discount and profit is observed in standard class with 59.9% , 59.7% and 60.7% of total respectively

- Least total of sales, discount and profit is found in same day with 5.3%, 5.3% and 5.2% of total respectively

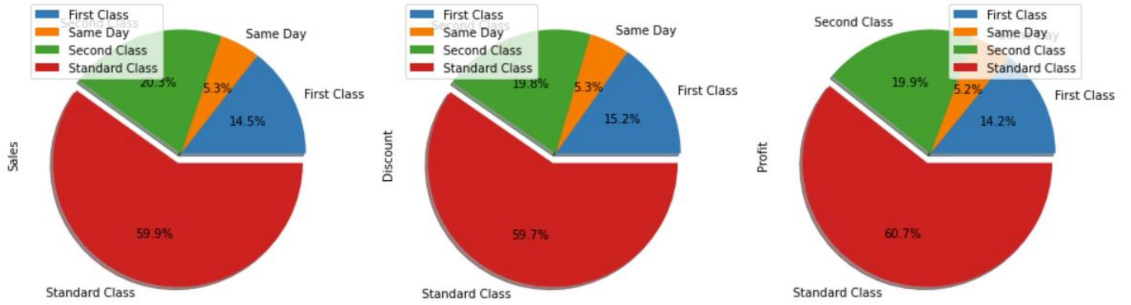


Fig 28: Pie Chart of Ship Mode wise total Sales, Discount and Profit

➢ <u>Pie Chart of Region wise total Sales, Discount and Profit</u>:

- Highest total sales discount and profit in Central with 22% , 21.15% and 21.2% of total respectively
- Least total of sales, discount in Central Asia with 6.0% and 1.9% of total respectively and least profit in EMEA with 3.0 % of total.

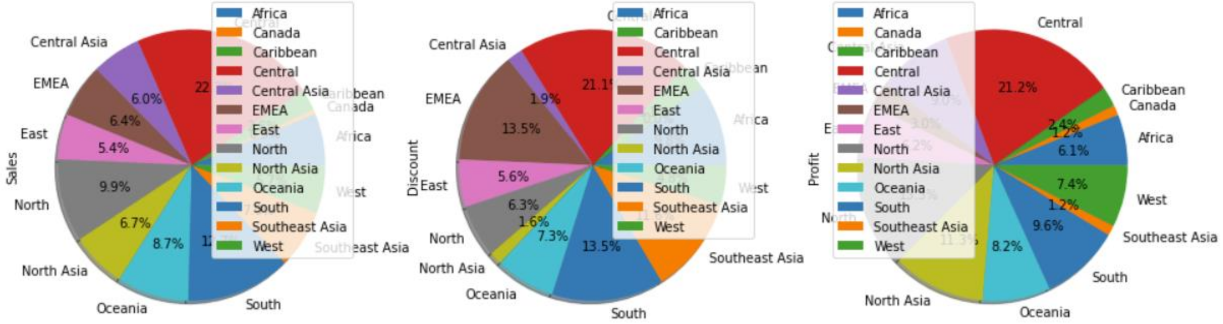


Fig 29: Pie Chart of Ship Mode wise total Sales, Discount and Profit

➢ <u>Heatmap showing correlation between non-categorical variables</u>:

- strong positive correlation is found between shipping cost and sales
- Strong negative correlation is found between profit and discount

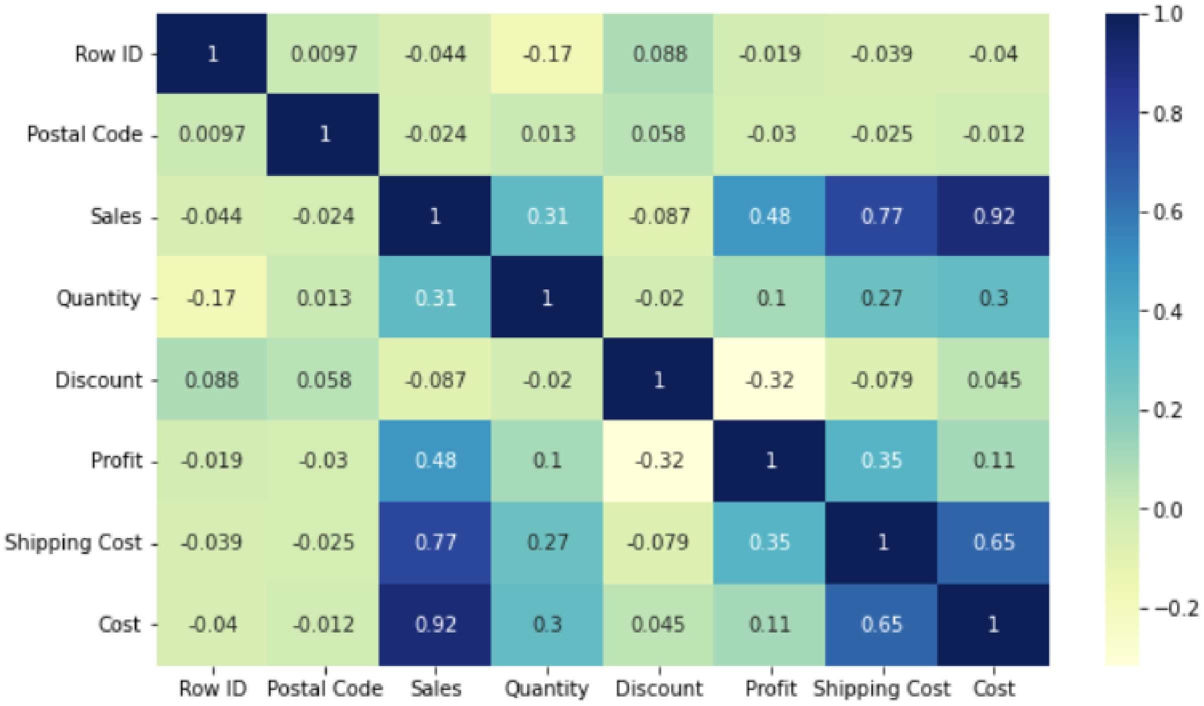


Fig 30: Heatmap showing correlation between non-categorical variables

4.3 Pre-Processing

Before data can be used as input for machine learning algorithms, it often must be cleaned, formatted, and restructured. After processing the missing entries, there are some qualities about certain features that must be adjusted. This preprocessing can help tremendously with the outcome and predictive power of model. The following are some of the preprocessing done in this project.

4.3.1 Splitting the data

Before training the model the data set is split into features and target. Thus initiating more pre-processing on features of data set. So our target here is Order Priority on which classification model has to be built.

4.3.2 Feature Engineering

It’s the process of producing new features with goal of simplifying and speeding up data transformations and thereby enhancing model accuracy. Following are some feature engineering done in this project.

- Feature Creation: New feature ‘Cost’ is created during the process of EDA. It is extracted from difference between Sales and sum of Profit and Shipping Cost. Thus raw cost of Product Segment, Category etc wise is analyzed.
- Feature Extraction: As a part of it after converting Order Date and Ship Date into time datetime data type features like day of week, month and week of year are extracted from both respectively. In day of week feature 0 represents Monday to 6 representing Sunday. Month is represented in integers ranging from 1-12 to corresponding month in order respectively. Week of year represent the week ordinal of the year.

4.3.3 Outlier Handling

Outliers are observation in data set that are too spread out from rest of data set. They may indicate experimental error or heavy skewness in the data. Following steps are followed during outlier handling.

- Visual Detection of outlier with Box plots: Outlier are detected in non-categorical variables like Sales, Discount, Quantity, Profit, Shipping Cost and Cost.
- Determination of outliers using Inter Quartile Range (IQR):
  Steps followed during the process are:
  - Calculation of 1st,2nd and 3rd Quartiles.
  - Computing the Inter Quartile range, IQR=Q3-Q1.
  - Compute:
    lower limit=Q1- (1.5 x IQR) and upper limit=Q2-(1.5 x IQR)
  - Locating index values of data points i.e outliers which falls below lower and upper limits.
- Replacement of outliers using mean/median: Distribution of features is plotted using histogram resulting in observation of normal distribution in Profit plot while Right Skewed distribution in

remaining features plot. Thus Profit outliers are replaced using mean of the feature while Sales, Discount, Quantity, Shipping Cost and Cost with median of each feature respectively.

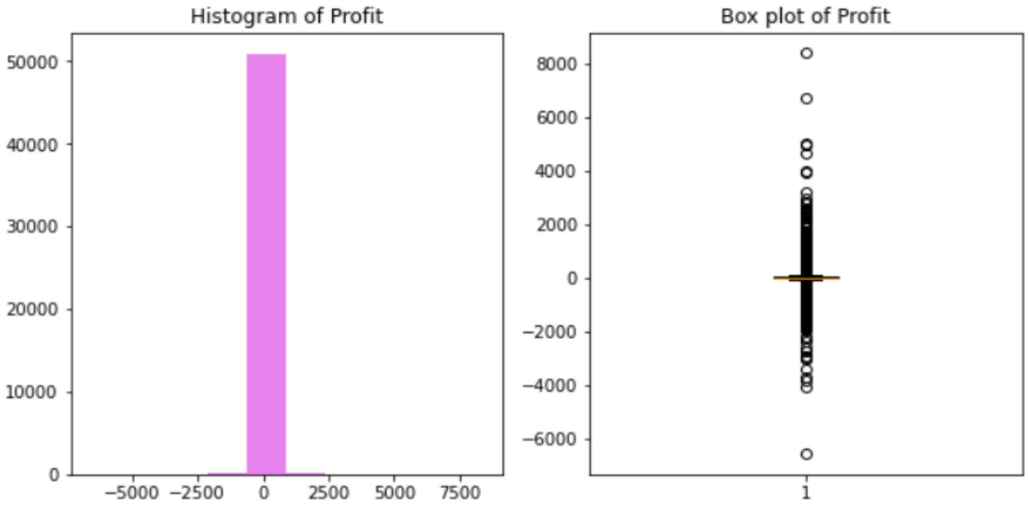


Fig 31: Outlier detection using box plot

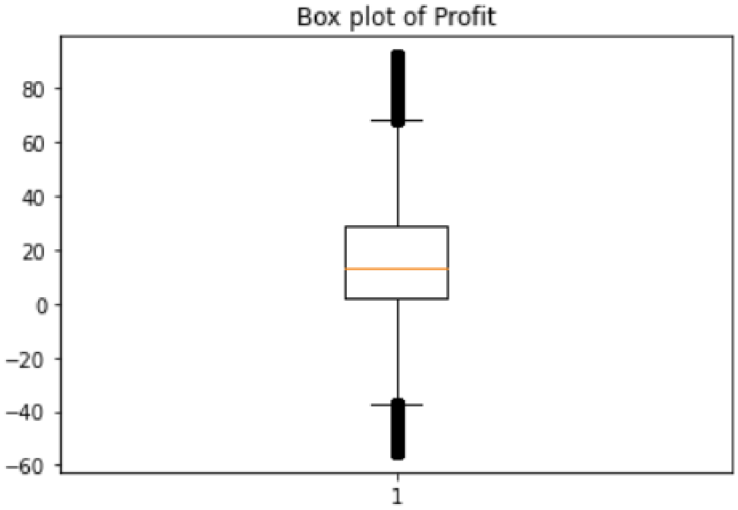


Fig 32: Box plot after outlier handling

4.3.4 Feature Reduction

It's the process of reducing the number of features resulting in faster computation in model. Features like Row ID, Order ID, Customer ID, Customer Name, Product id are removed as its not required by the model. As features are already extracted from Order Date and Ship Date, They are removed.

About 80% of missing values is detected in feature Postal Code which impossible for replacement and thereby removing it.

4.3.4 Encoding

Machine Learning models work with numbers only so categorical variables are to be encoded to numerical values Following are some of the encoding done in this project.

- One Hot Encoding: Here Binary values like 0 and 1 are assigned to features like Ship Mode, Segment and Category as their value counts are low.
- Label Encoding: Here values like 0, 1, 2… are assigned to each unique labels in features like City, Country, State, Market, Region, Sub-Category and Product Name.

4.3.5 Feature Scaling

As a part of feature scaling **standardization** is done on features set to arrive at distribution of 0 mean and 1 standard deviation. This is done so the range of features doesn't differ.

4.3.6 Oversampling

From observing the target, unequal distributions of class labels are observed. This bias in training dataset may influence the ML algorithm and may result in ignoring minority classes. So we have used **SMOTE** (Synthetic Minority

Oversampling Technique) creating new minority examples based on KNN better resembling the number of examples in majority class.

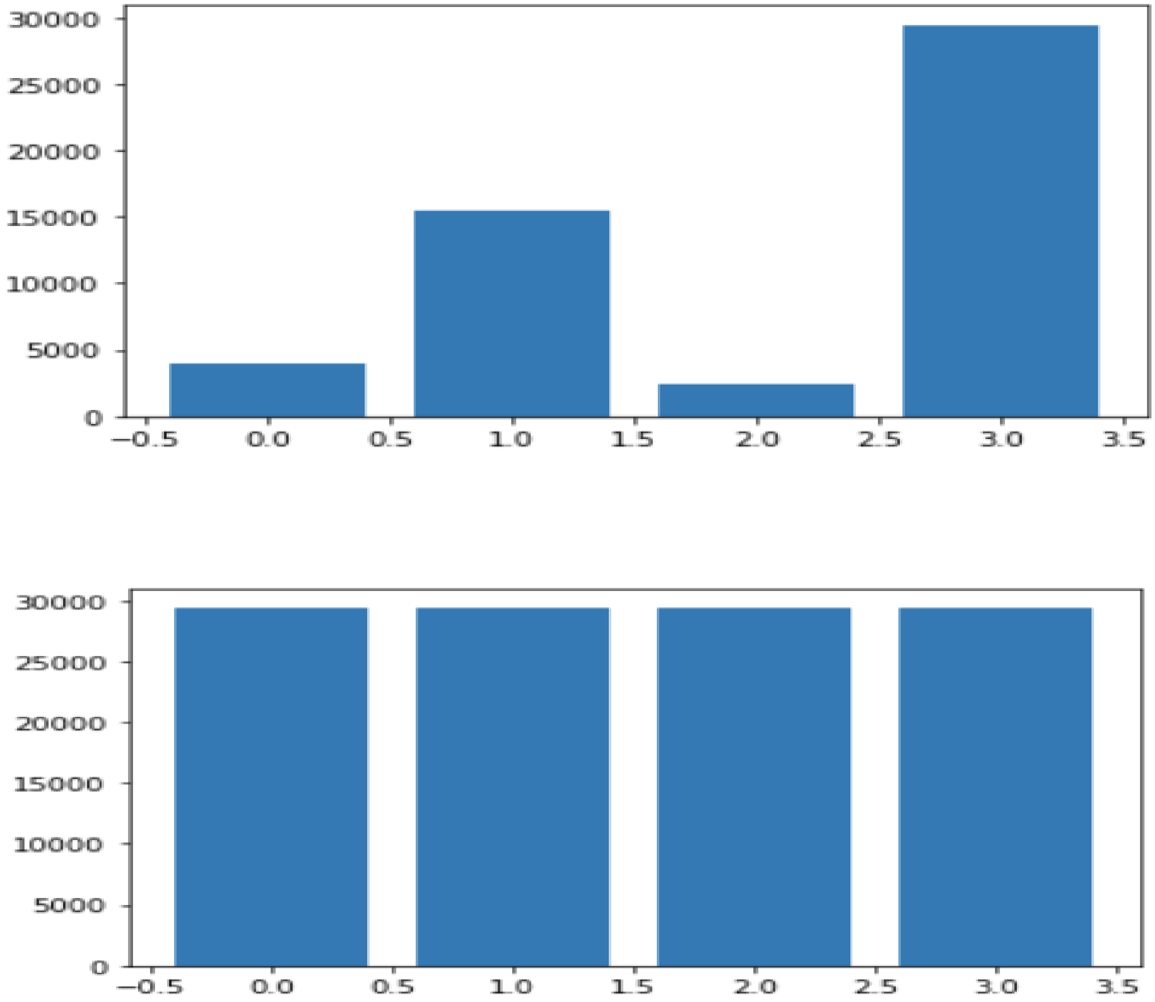


Fig 33: Before and after Oversampling

4.4 Model

The following are the procedure in developing best fit and accurate model.

4.4.1 Splitting the data into train and test set

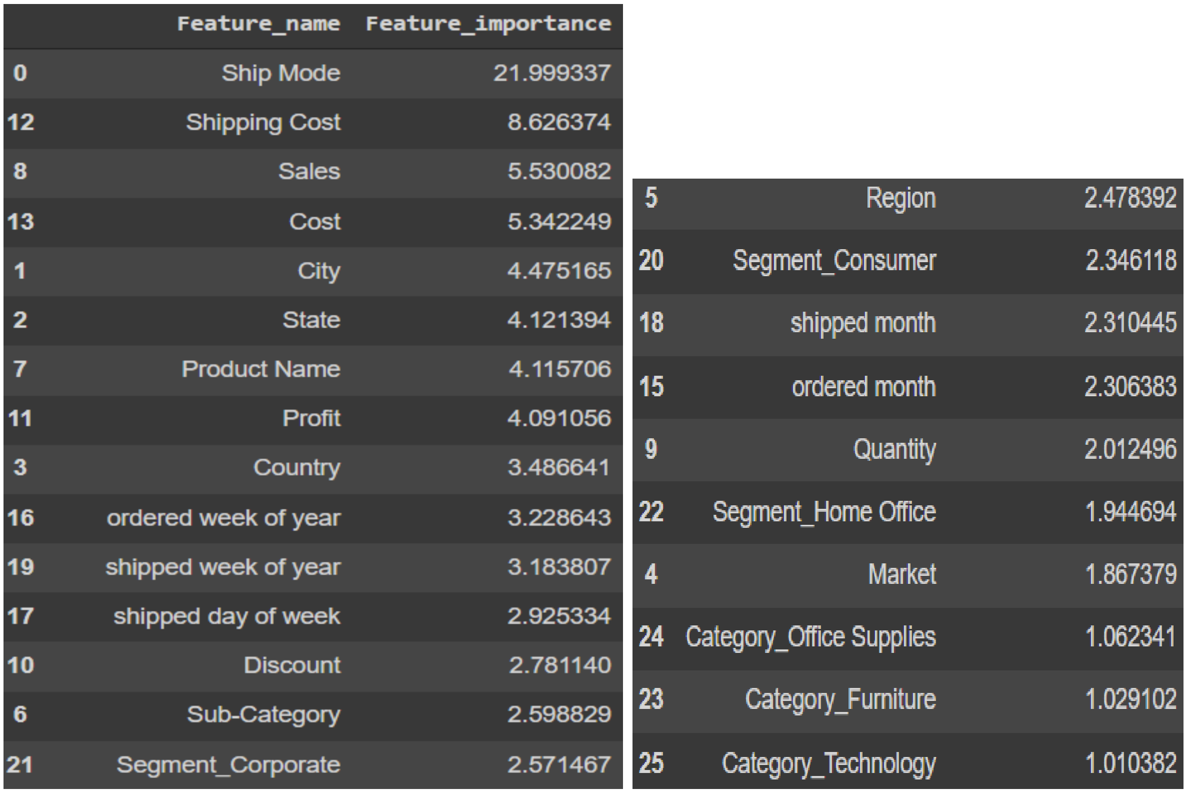
Using train_test_split attribute from sklearn library, features and target is split into train and test sets. The train size is set as 0.3.

4.4.2 Modeling

Different classification models like Logistic Regression, SVC, K-Nearest Neighbor, Decision tree and Random Forest classifier are tried and best among which is Random Forest Classifier giving best score is proceeded. Accuracy on train set is found out to be 100% and on test set of 82.29%.

4.4.3 Model based Feature Importance

Data Frame consisting of features and their Feature importance is constructed and Highest feature importance among which is ‘Ship Mode’ and lowest is ‘Category_Furniture’.



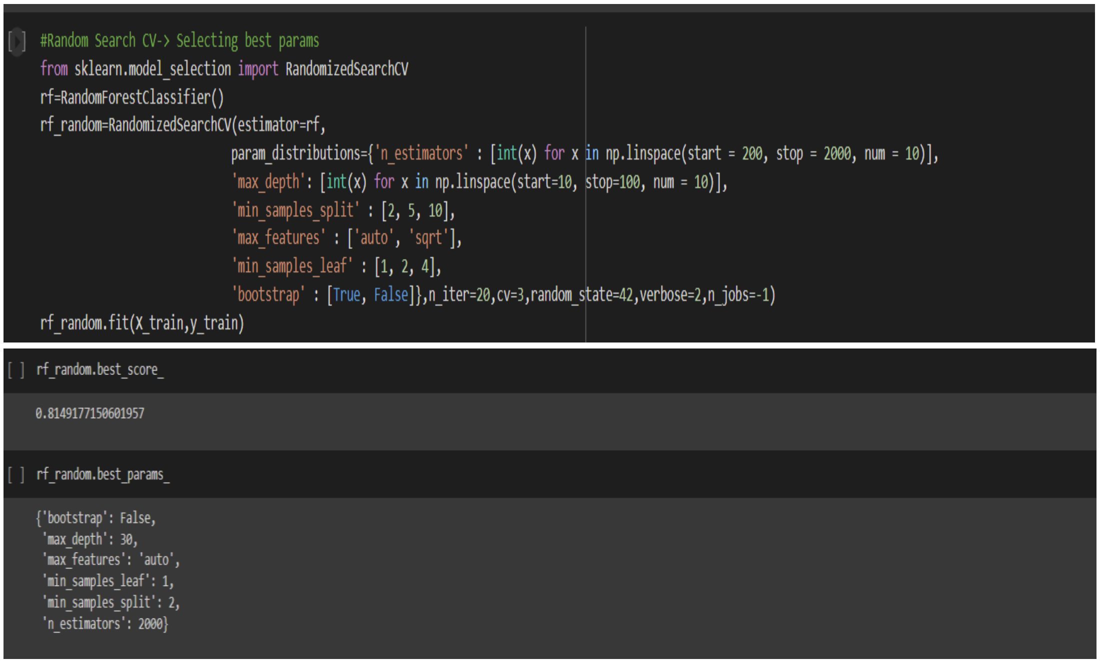
	Feature_name	Feature_importance
0	Ship Mode	21.999337
12	Shipping Cost	8.626374
8	Sales	5.530082
13	Cost	5.342249
1	City	4.475165
2	State	4.121394
7	Product Name	4.115706
11	Profit	4.091056
3	Country	3.486641
16	ordered week of year	3.228643
19	shipped week of year	3.183807
17	shipped day of week	2.925334
10	Discount	2.781140
6	Sub-Category	2.598829
21	Segment_Corporate	2.571467
5	Region	2.478392
20	Segment_Consumer	2.346118
18	shipped month	2.310445
15	ordered month	2.306383
9	Quantity	2.012496
22	Segment_Home Office	1.944694
4	Market	1.867379
24	Category_Office Supplies	1.062341
23	Category_Furniture	1.029102
25	Category_Technology	1.010382

Fig 34: Feature Importance

4.4.4 Fine tuning of Hyperparameter

This consist of 2 steps:

- **Random Search CV:** we will be defining the library required for Random search followed by defining all the parameters we want to test on our model. By specifying a parameter called ‘n_iter’ i.e taking n_iter=20, any random 20 combination will be tried and train it. After this score is checked and best parameter is determined.



```
#Random Search CV-> Selecting best params
from sklearn.model_selection import RandomizedSearchCV
rf=RandomForestClassifier()
rf_random=RandomizedSearchCV(estimator=rf,
                             param_distributions={'n_estimators' : [int(x) for x in np.linspace(start = 200, stop = 2000, num = 10)],
                             'max_depth': [int(x) for x in np.linspace(start=10, stop=100, num = 10)],
                             'min_samples_split' : [2, 5, 10],
                             'max_features' : ['auto', 'sqrt'],
                             'min_samples_leaf' : [1, 2, 4],
                             'bootstrap' : [True, False]},n_iter=20,cv=3,random_state=42,verbose=2,n_jobs=-1)
rf_random.fit(X_train,y_train)
```

```
rf_random.best_score_
```

```
0.8149177150601957
```

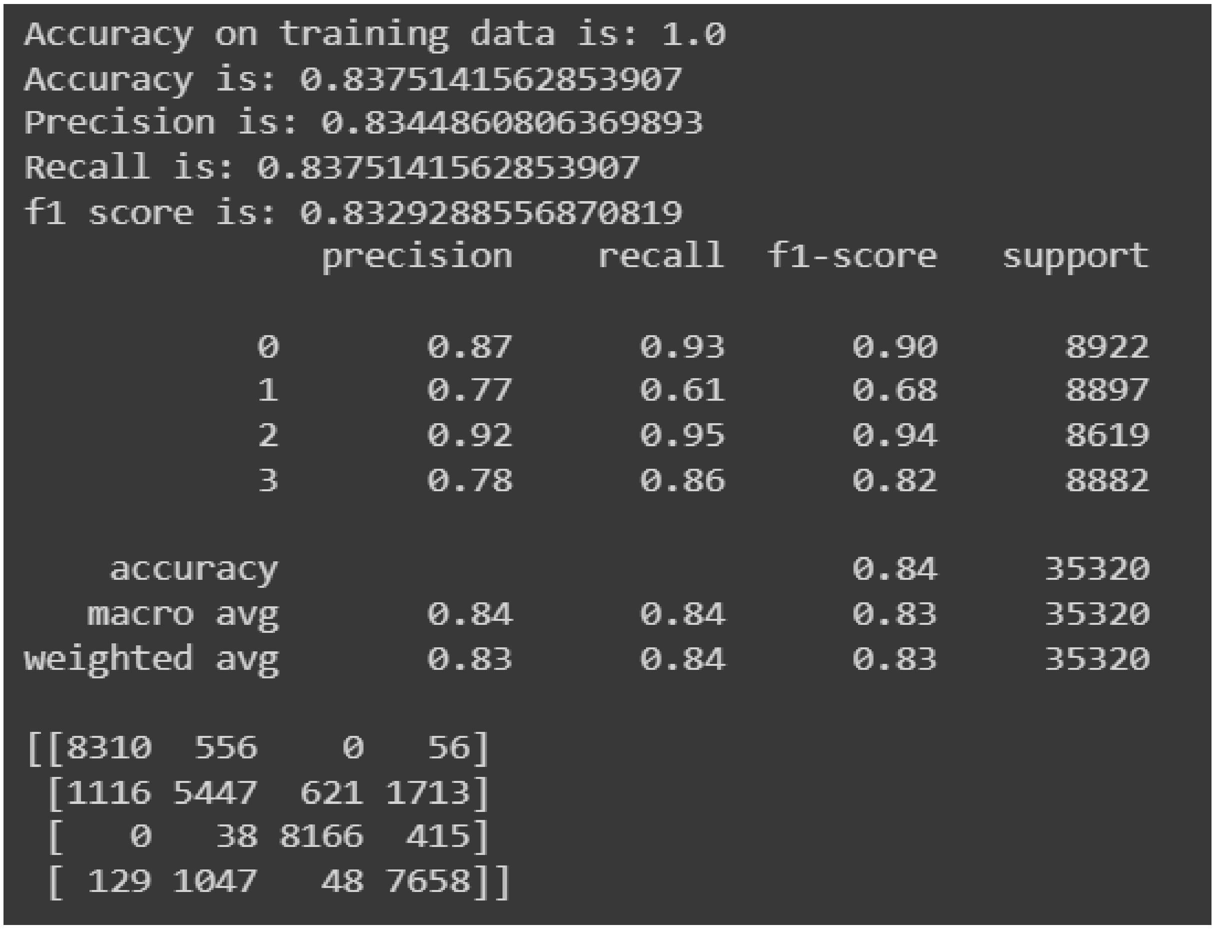
```
rf_random.best_params_
```

```
{'bootstrap': False,
 'max_depth': 30,
 'max_features': 'auto',
 'min_samples_leaf': 1,
 'min_samples_split': 2,
 'n_estimators': 2000}
```

Fig 35: Random search CV

- Final Model after tuning: From the best parameter obtained from random search new RF model is recreated, train it and best possible accuracy of 83.75% is obtained on test data.

7. Result



```
Accuracy on training data is: 1.0
Accuracy is: 0.8375141562853907
Precision is: 0.8344860806369893
Recall is: 0.8375141562853907
f1 score is: 0.8329288556870819
              precision    recall  f1-score   support

           0       0.87      0.93      0.90      8922
           1       0.77      0.61      0.68      8897
           2       0.92      0.95      0.94      8619
           3       0.78      0.86      0.82      8882

    accuracy                           0.84     35320
   macro avg       0.84      0.84      0.83     35320
weighted avg       0.83      0.84      0.83     35320

[[8310  556    0   56]
 [1116 5447  621 1713]
 [   0   38 8166  415]
 [ 129 1047   48 7658]]
```

Fig 36: Final Model Report

From our Final Model we have accuracy score of 83.75% and f1 scores are 0.90, 0.68, 0.94 and 0.82 for Critical, High, Low and Medium respectively.

8. Conclusion

Explored and completed various EDA on dataset such as product analysis, geography centric sales etc discovering trends and patterns. Done some pre-processing on data handling outliers i.e too much spread out, feature engineered some features useful to the model, encoded categorical features and finally created a model classifying order priority as Critical, High, Low and Medium from various characteristic details of features in data set. This type of models can be embedded in retail customer service applications for greater customer satisfaction and be more competent in the market

References

1. Global Superstore data https://www.kaggle.com/apoorvaappz/global-super-store-dataset

2. Literature on 'learned approach to priority setting and classification' by Chenoainc: https://www.chenoainc.com/a-learned-approach-to-priority-setting-classification/

3. Hyper parameter tuning with grid search and random search https://towardsdatascience.com/hyperparameter-tuning-with-grid-search-and-random-search-6e1b5e175144

4. Balancing Multiclass imbalanced classification https://machinelearningmastery.com/multi-class-imbalanced-classification/#:~:text=SMOTE%20Oversampling%20for%20Multi-Class%20Classification%20Oversampling%20refers%20to,the%20number%20of%20examples%20in%20the%20majority%20classes.

5. Imblearn Technique: https://analyticsindiamag.com/what-is-imblearn-technique-everything-to-know-for-class-imbalance-issues-in-machine-learning/

6. Feature Scaling using standardization https://www.analyticsvidhya.com/blog/2020/04/feature-scaling- machine-learning-normalization-standardization/