E-COMMERCE & RETAIL B2B CASE STUDY

By – Somil Nagar | Soujanya TV | Rahul Biswas

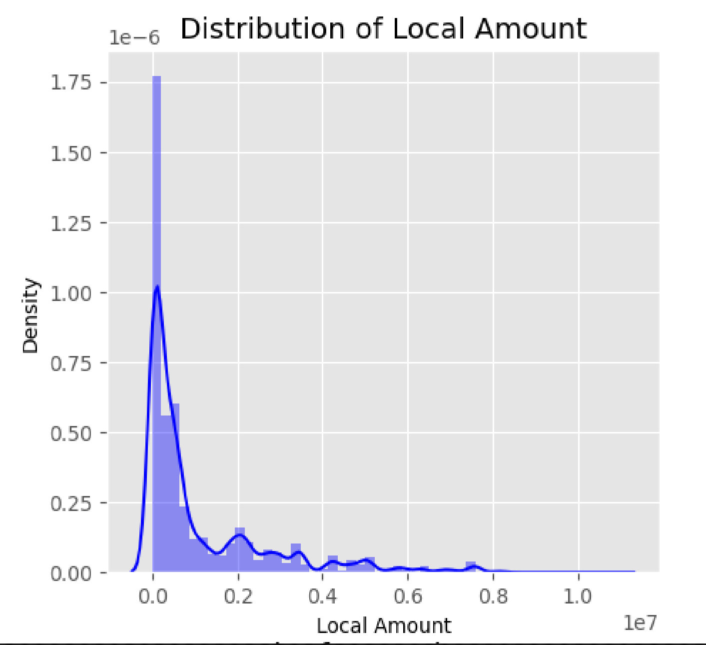
READING AND UNDERSTANDING THE DATA

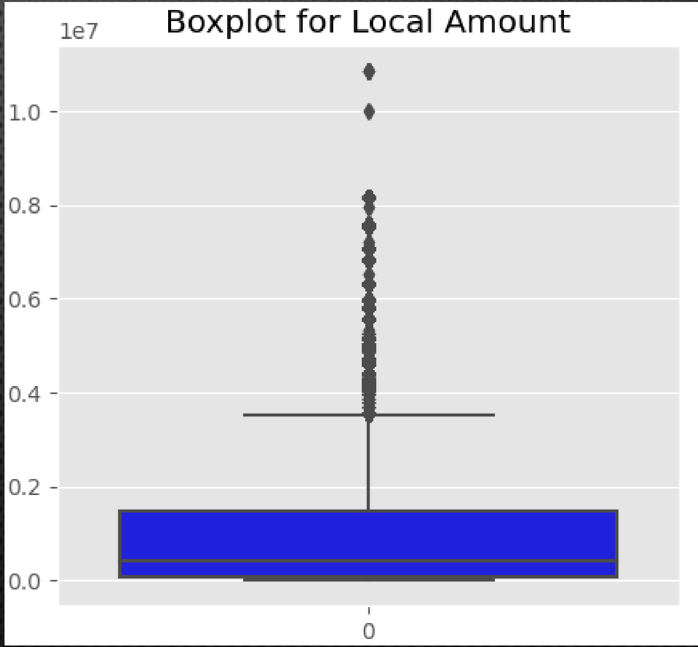
1) Data Type Checks
2) Treating Missing Values
3) Dropping unnecessary columns
4) Outlier detection
5) Handling Outliers
6) Derived Columns

- Numeric columns -[USD_Amount]
- Categorical column -payment Term, Invoice_class.
- Date columns -reciept_date, due_date, Invoice_date
- Created Target Var -LATE_PAY
- Dropped columns like Local_Amount,RECEIPT_DOC_NO,customer_name,class_currency_code,inv_curr_code, reciept_method which were not contributing to our target variable

EDA -Visualization

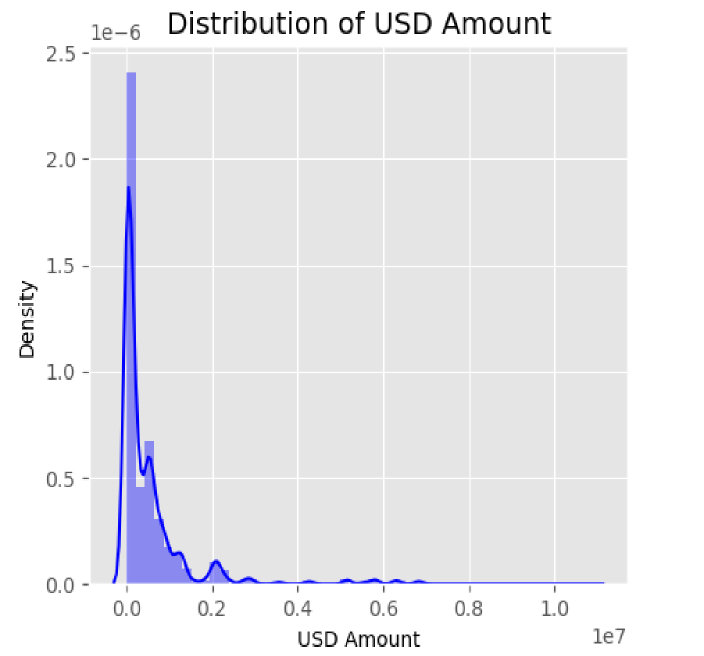
LOCAL AMOUNT

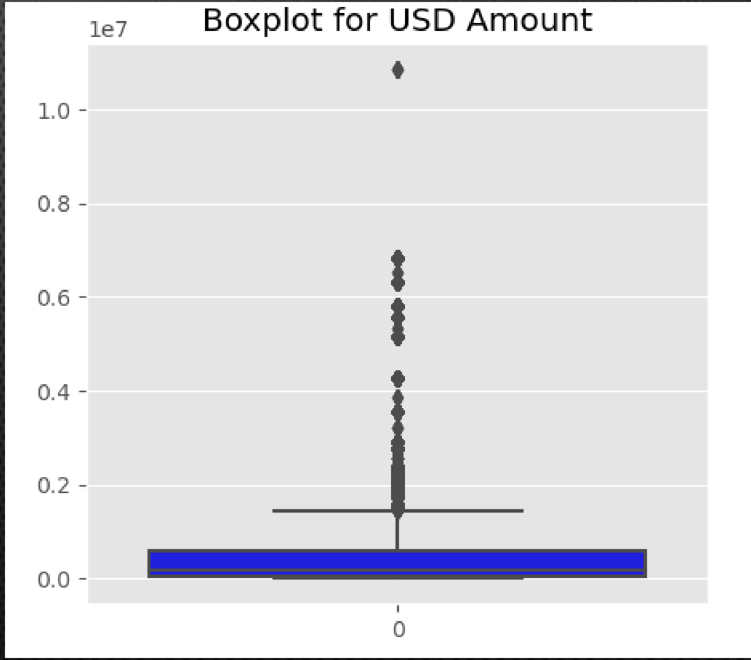




EDA - Visualization

USD AMOUNT





EDA - CATEGORICAL COLUMNS

- Top 10 customers based on frequency

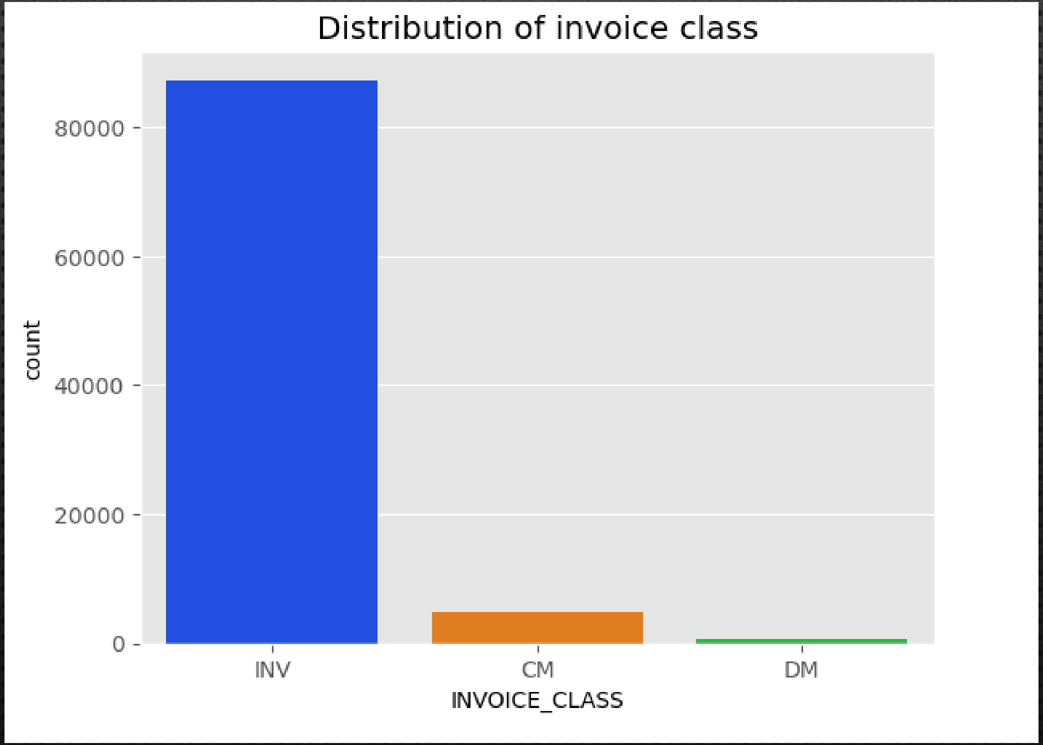
SEPH Corp	23075
FARO Corp	15004
PARF Corp	6624
ALLI Corp	5645
AREE Corp	2224
DEBE Corp	2133
RADW Corp	1647
YOUG Corp	1480
HABC Corp	1402
CARR Corp	952

- Top 10 customers based on amount

CUSTOMER_NAME	
SEPH Corp	32,533,709,059.000
FARO Corp	5,790,071,209.000
PARF Corp	3,200,510,261.000
ALLI Corp	2,580,740,593.000
AREE Corp	1,125,144,489.000
HABC Corp	534,321,619.000
RADW Corp	362,237,576.000
L OR Corp	295,550,941.000
CGR Corp	279,516,184.000
PCD Corp	246,606,985.000

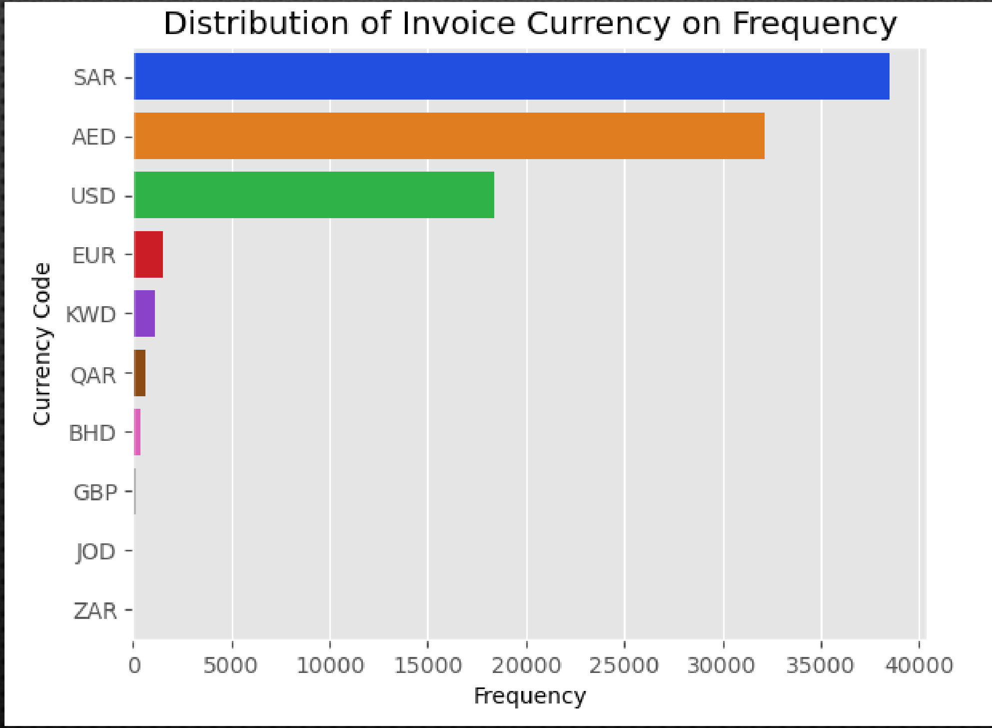
EDA - Categorical Columns

INVOICE CLASS



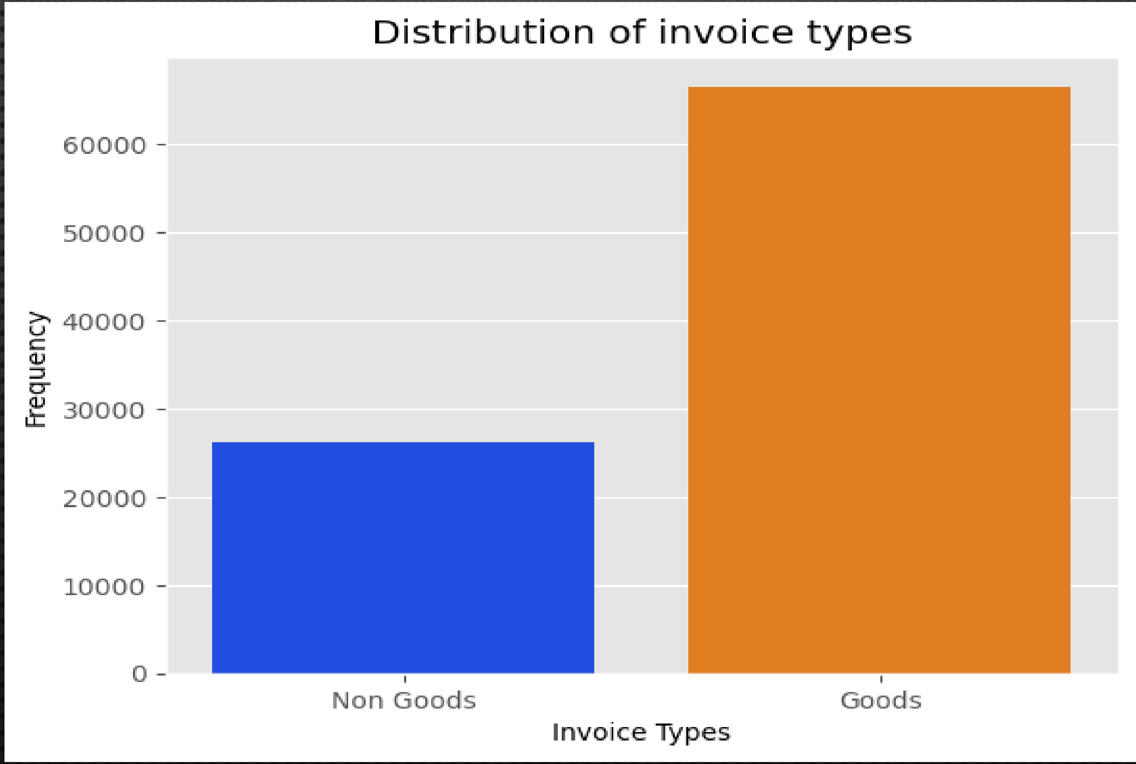
EDA - Categorical Columns

CURRENCY CODES



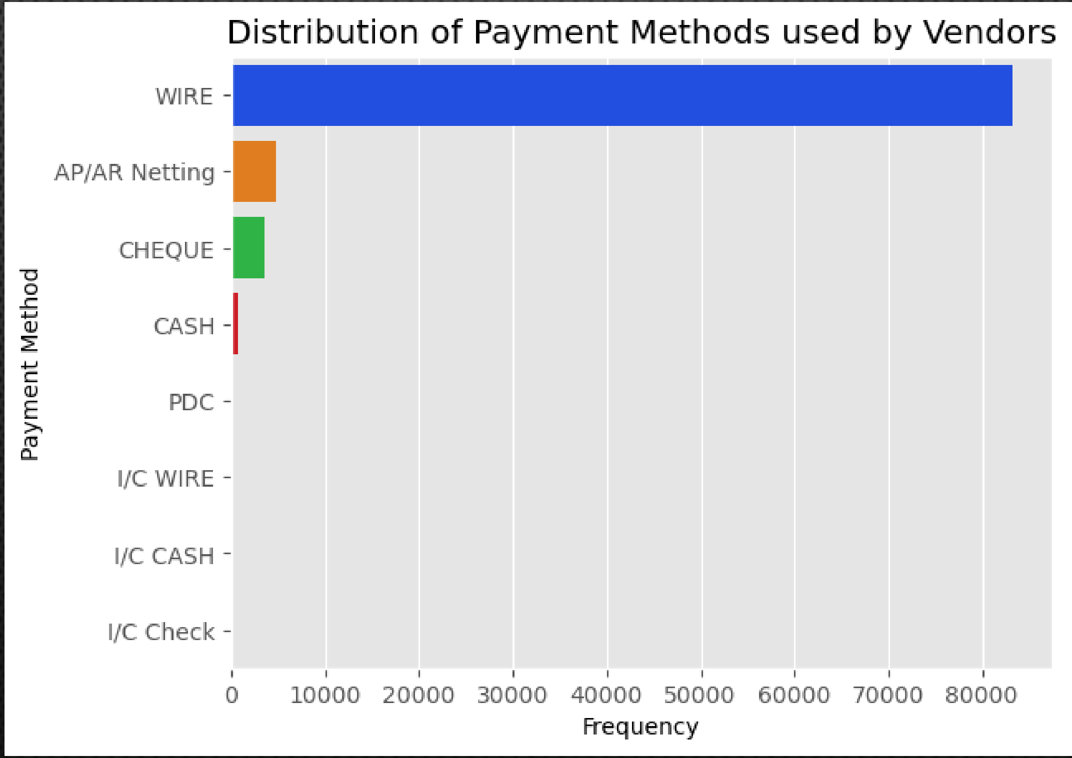
EDA - Categorical Columns

INVOICE TYPE



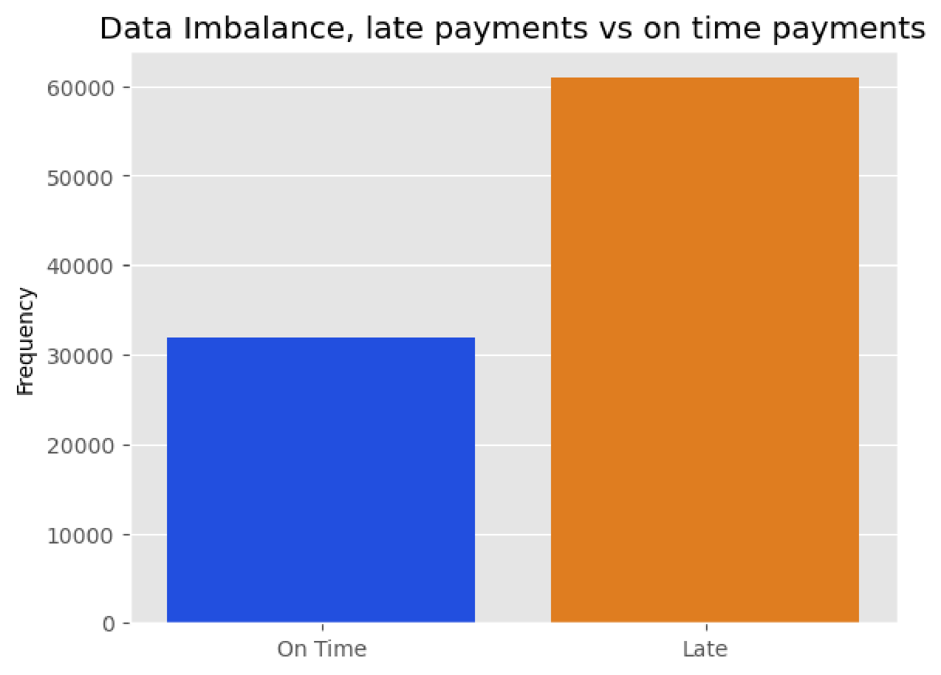
EDA - Categorical Columns

PAYMENT METHODS



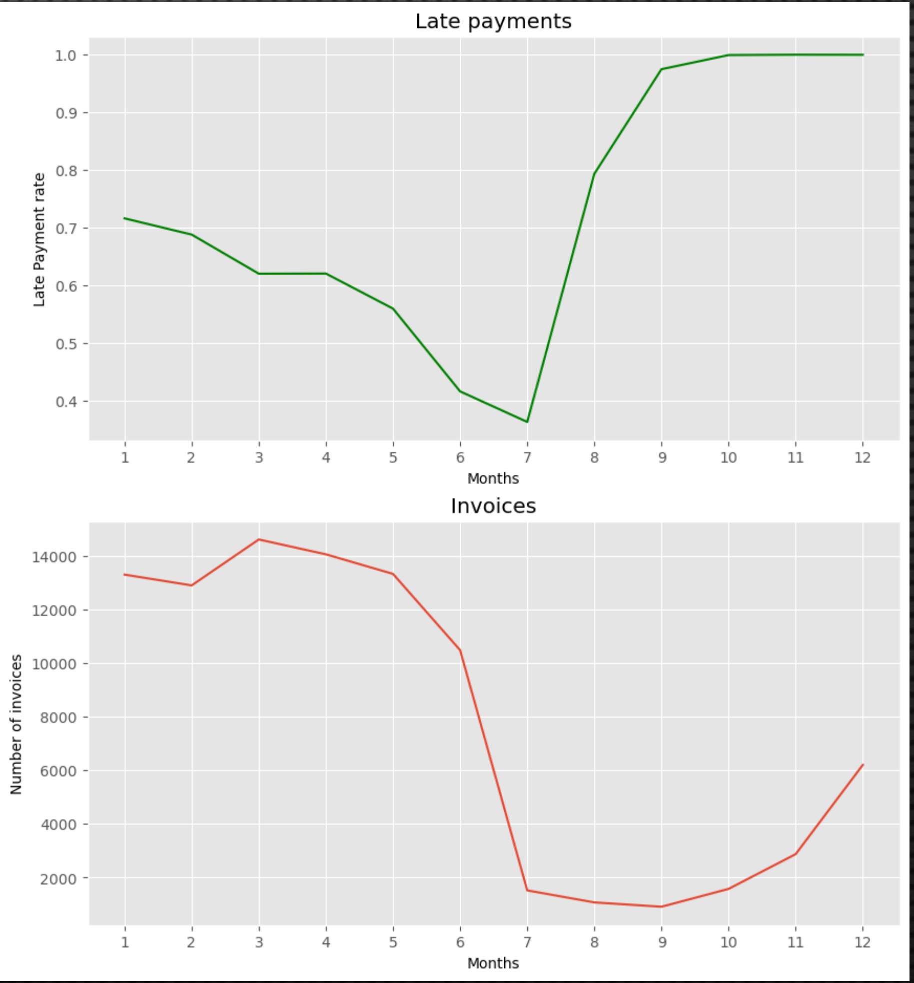
CLASS IMBALANCE

- The class imbalance is not that high , so we can work with the present data itself.
- 65~35 ratio is acceptable.



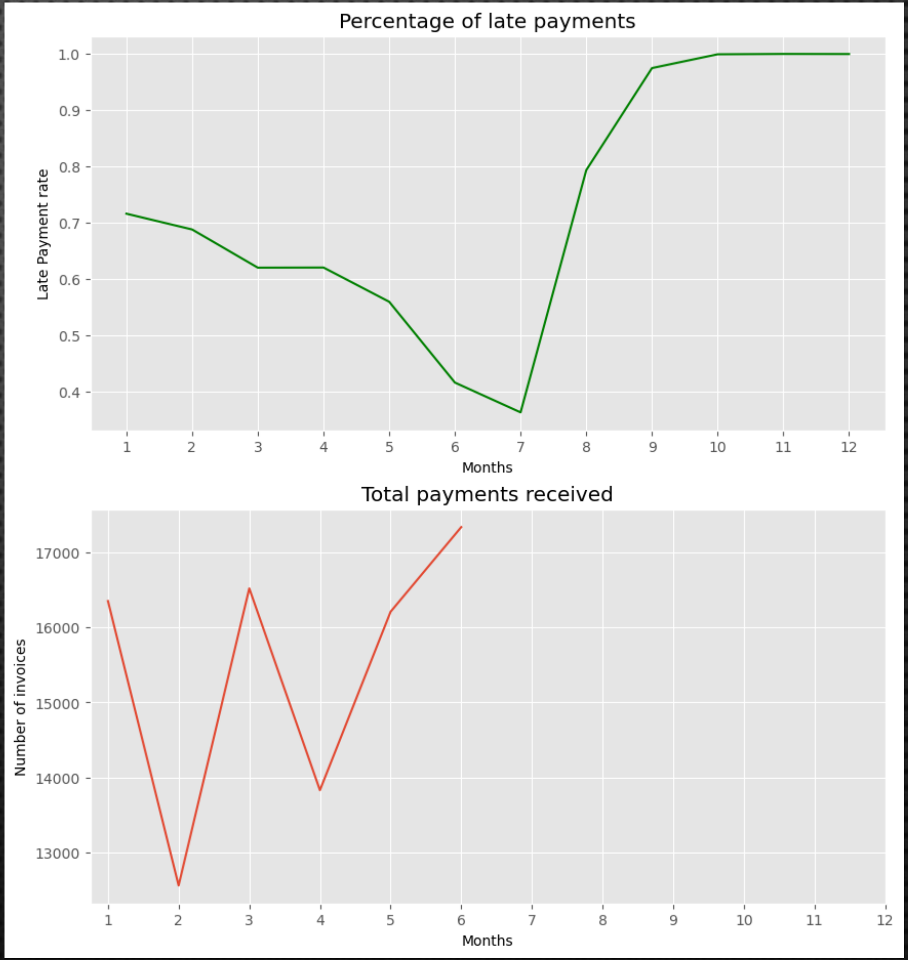
BI-VARIATE ANALYSIS

- Monthly affects on payments and invoices.



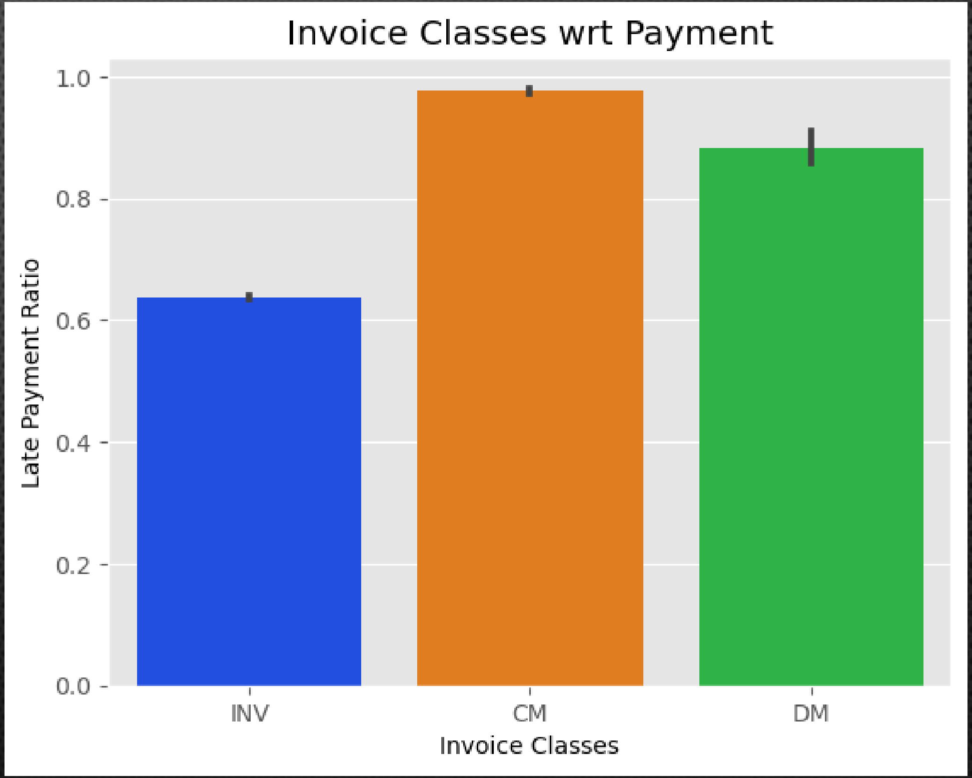
BI-VARIATE ANALYSIS

- A stark effect is noted here, which provides that all payments are received in the first half of the year only



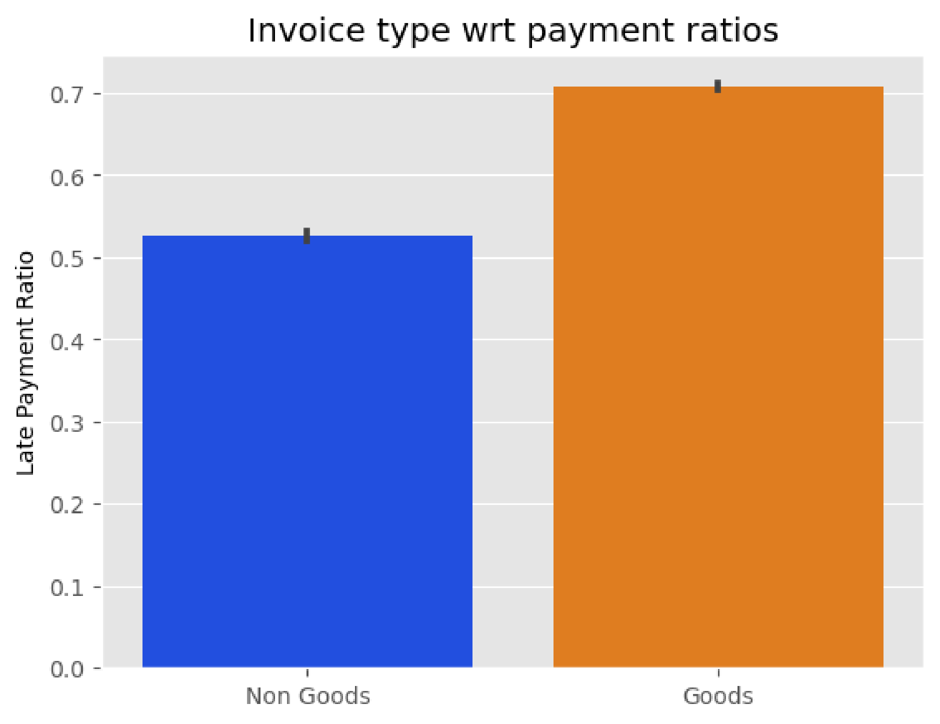
BI-VARIATE ANALYSIS

- It is observed that both credit and debit memo have high late payment ratios, however it is also to be noted that there are only a few invoices with CM and DM Class



BI-VARIATE ANALYSIS

- It is observed that late payment ratio is much higher for goods.



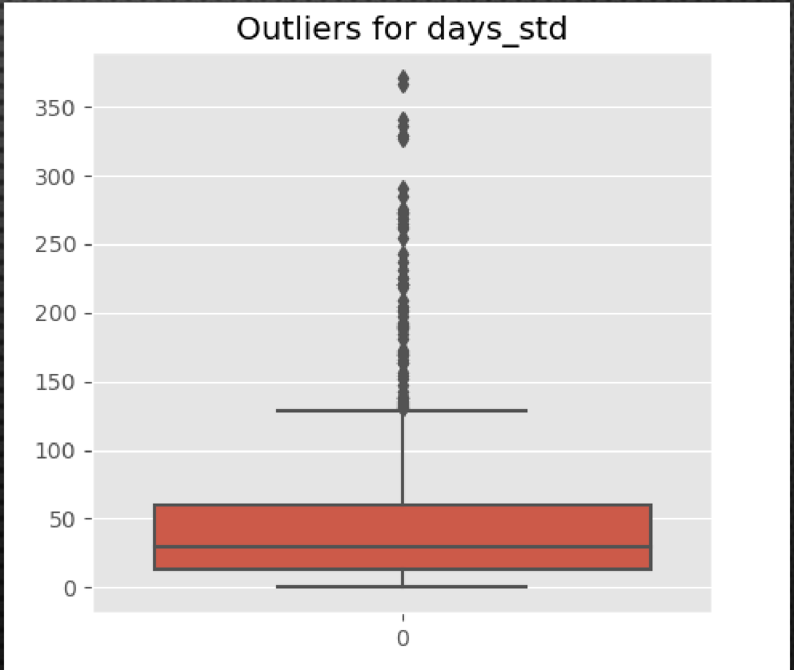
FEATURE ENGINEERING

- Created Dummy variables for ‘Payment _Term’ and ‘Invoice_class’
- First combined similar payment terms and then clubbed every other payment term except top 10.
- Open_Invoice_Data table -removed unnecessary columns and created dummy variables ’

CLUSTERING

Customer Segmentation

- Outlier treatment
- Removed about quantile 0.99.



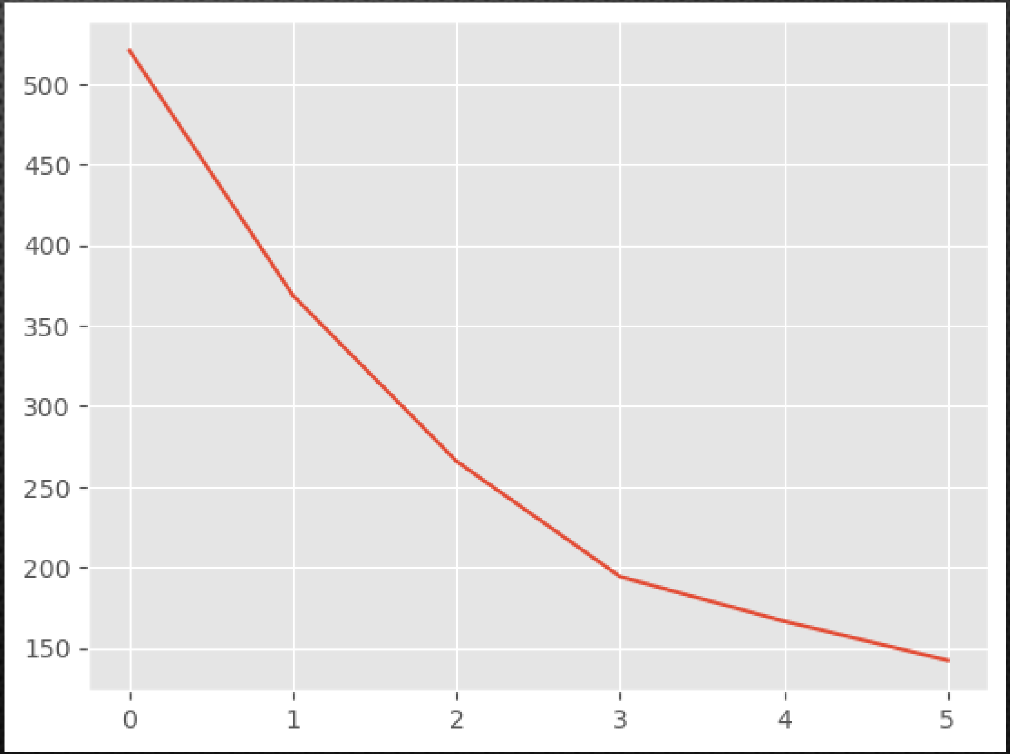
SCALING AND HOPKINS TEST

- We can see negative values for days mean which means that the customer has done immediate payment, while the invoice was created later.
- To maintain this data integrity, it would be advisible to use standardization over normalization for scaling.
- On running hopkins Test, we got a value of 0.91337.
- A value close to 1 tends to indicate the data is highly clustered, random data will tend to result in values around 0.5, and uniformly distributed data will tend to result in values close to 0.

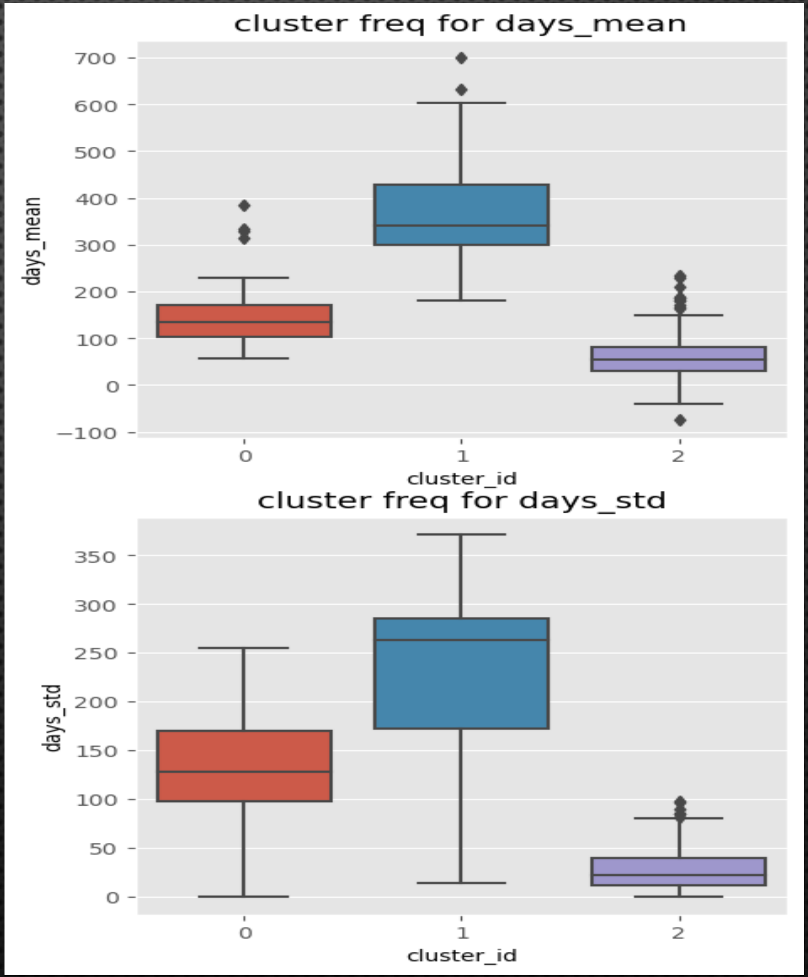
OPTIMAL CLUSTERS

Elbow Method

Optimal cluster at 3.

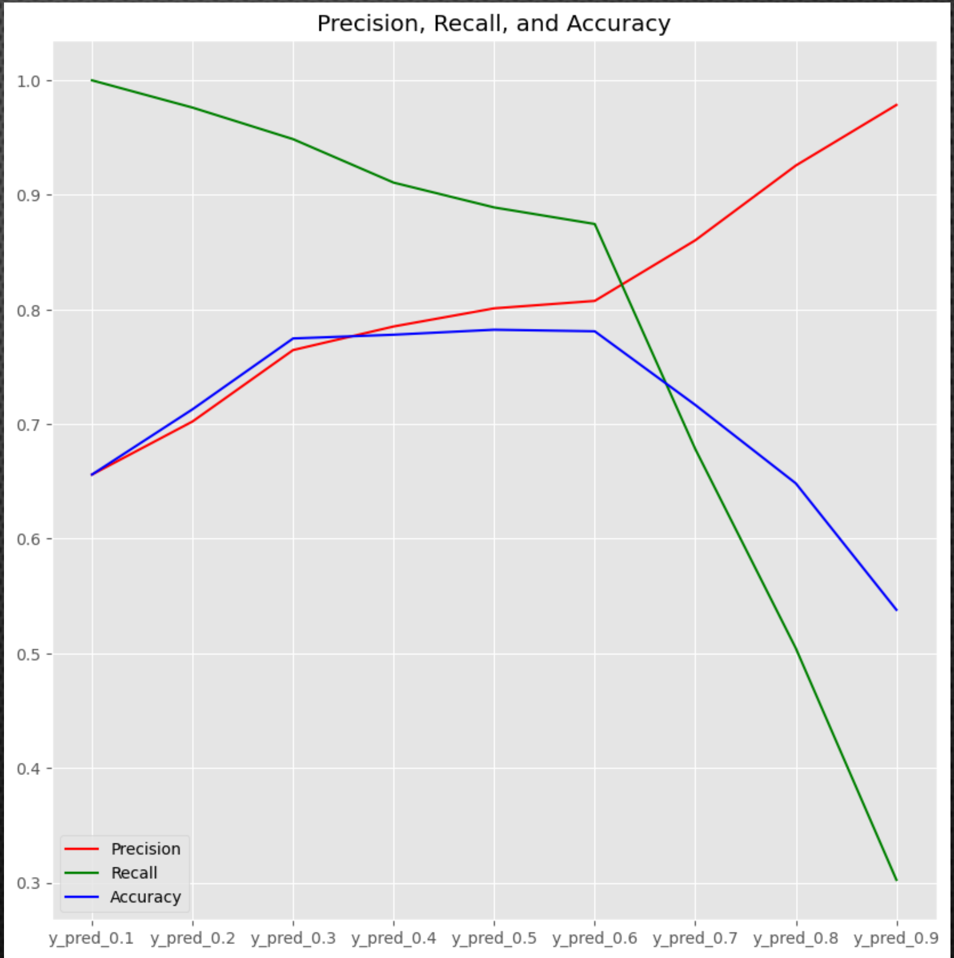


SUMMARY - CLUSTERING



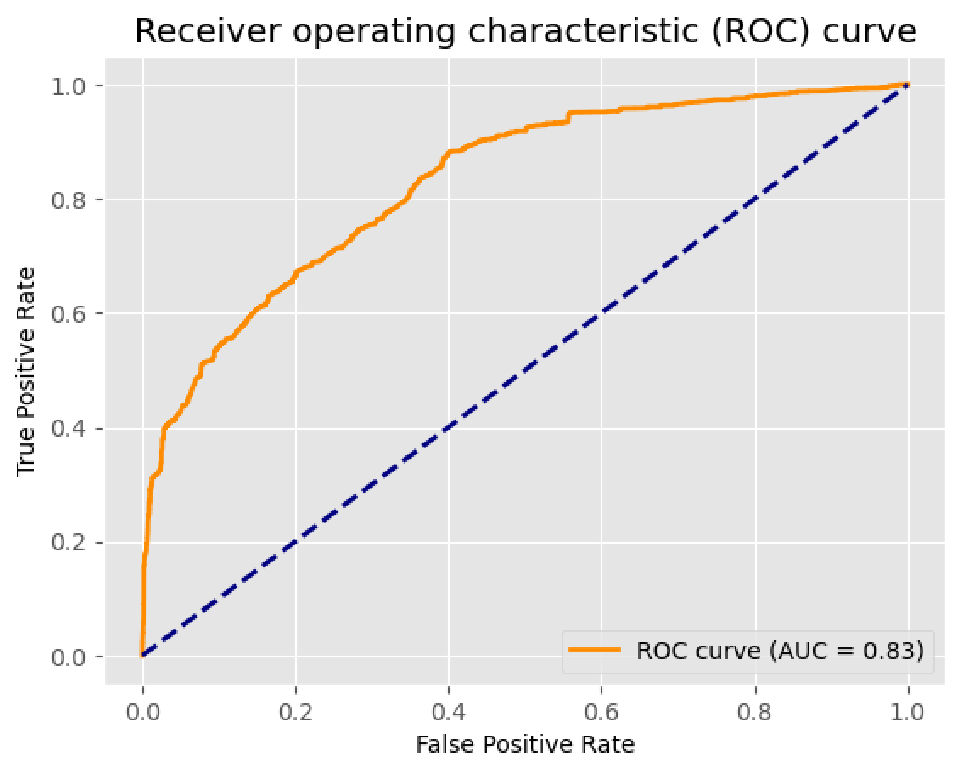
LOGISTIC REGRESSION

- Precision , recall and accuracy curve give us the optimal cutoff at 0.5



LOGISTIC REGRESSION

- The logistic regression algorithm is working, and has an AUC score of 0.83



LOGISTIC REGRESSION

Evaluation metrics on training dataset

	precision	recall	f1-score	support
0	0.73	0.58	0.65	22349
1	0.80	0.89	0.84	42618
accuracy			0.78	64967
macro avg	0.77	0.73	0.74	64967
weighted avg	0.78	0.78	0.78	64967

Evaluation metrics on the test dataset

	precision	recall	f1-score	support
0	0.73	0.58	0.65	9529
1	0.80	0.89	0.84	18315
accuracy			0.78	27844
macro avg	0.77	0.74	0.75	27844
weighted avg	0.78	0.78	0.78	27844

- The data is almost similar and hence we can say that our algorithm is working as expected.

RANDOM FOREST

Basic random forest model is also working well with the following metric scores on training dataset.

- We see that the basic model itself has a high accuracy, recall, and precision. But here we focus on recall of the positive class in both the train and test set. The basic model itself is able to identify 93-94% of all positive instances.

	precision	recall	f1-score	support
0	0.94	0.88	0.91	9529
1	0.94	0.97	0.96	18315
accuracy			0.94	27844
macro avg	0.94	0.93	0.93	27844
weighted avg	0.94	0.94	0.94	27844

Accuracy is : 0.9402384714839822

RANDOM FOREST-HYPERPARAMETER TUNING (TRAIN)

- Grid Search CV

	precision	recall	f1-score	support
0	0.96	0.91	0.94	22349
1	0.96	0.98	0.97	42618
accuracy			0.96	64967
macro avg	0.96	0.95	0.95	64967
weighted avg	0.96	0.96	0.96	64967

RANDOM FOREST (TEST)

- We have found the best model, which is giving great performance for us. All the metrics including accuracy, recall, and precision are great. We will use this model to make predictions on our invoices.

	precision	recall	f1-score	support
0	0.91	0.85	0.88	9529
1	0.93	0.96	0.94	18315
accuracy			0.92	27844
macro avg	0.92	0.90	0.91	27844
weighted avg	0.92	0.92	0.92	27844

RANDOM FOREST- FEATURE IMPORTANCE

- Although feature importance does not show the direction in which the possibility of late payment is affected (whether it will affect it positively or negatively) it does show that features such as Amount, Invoice Month, and Receipt Month affect the possibility of late payment

```
Feature ranking:
1. USD Amount (0.233)
2. Invoice_Month (0.202)
3. Reciept_Month (0.139)
4. 60 Days from EOM (0.110)
5. 30 Days from EOM (0.105)
6. cluster_id (0.057)
7. Immediate Payment (0.044)
8. 15 Days from EOM (0.030)
9. 60 Days from Inv Date (0.017)
10. 30 Days from Inv Date (0.015)
11. 90 Days from EOM (0.012)
12. 90 Days from Inv Date (0.010)
13. 45 Days from EOM (0.007)
14. INV (0.006)
15. 45 Days from Inv Date (0.006)
16. CM (0.006)
17. DM (0.001)
```

PREDICTIONS

- We use both classification models to check performance.
- As both the models were performing well, we need to select the one which is more interpretable.
- Algorithm which helps us define the linear relationship of features with target variable.

	prob_logreg	Prob_rf
Customer_Name		
2H F Corp	0.0802	0.622222
3D D Corp	0.0000	0.252195
6TH Corp	0.0465	0.152873
ABDU Corp	0.0000	0.412946
ABEE Corp	0.4145	0.460000
...	...	...
ZAIN Corp	0.2711	0.730333
ZALL Corp	0.1761	0.300654
ZALZ Corp	0.0006	0.582234
ZINA Corp	0.1955	0.113333
ZUHA Corp	0.1716	0.271866

CONCLUSION

- We have got two models predicting different probability values for a customer to have a late payment.

- It is to be noted that Random forest is performing much better than logistic regression. We can take those into account and make pre-emptive calls to the customers to have them pay their invoice amounts on time. Anyone with a high value in Column Prob_rf- shows that that customer has high probability of making a late payment.

- Since logistic regression shows linear relationship between probability and the features we can use it to find the relation.

THANK YOU