E-COMMERCE & RETAIL B2B CASE STUDY

PRESENTED BY : MANALI POTE

PROBLEM IDENTIFICATION

1) A sports retail company Schuster dealing in B2B transactions often deals with vendors on a credit basis, who might or might not respect the stipulated deadline for payment.
2) Vendors delaying their payments result in financial lag and loss which becomes detrimental to smooth business operations.
3) Additionally, company employees are set up chasing around for collecting payments for a long period of time resulting in no value-added activities and wasteful resource expenditure.

BUSINESS OBJECTIVE

1) Schuster would like to better understand the customers' payment behavior based on their past payment patterns (customer segmentation).
2) Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
3) It wants to use this information so that collectors can priorities their work in following up with customers beforehand to get the payments on time.

READING AND UNDERSTANDING DATA

1) Load the data.
2) Check the datatype.
3) Check and handle NA values and missing values.
4) Dropping unnecessary columns.

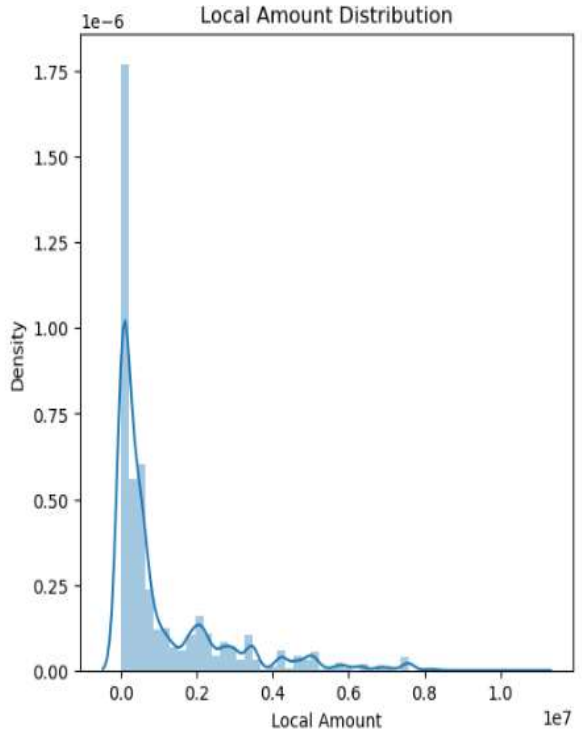
EXPLORATORY DATA ANALYSIS

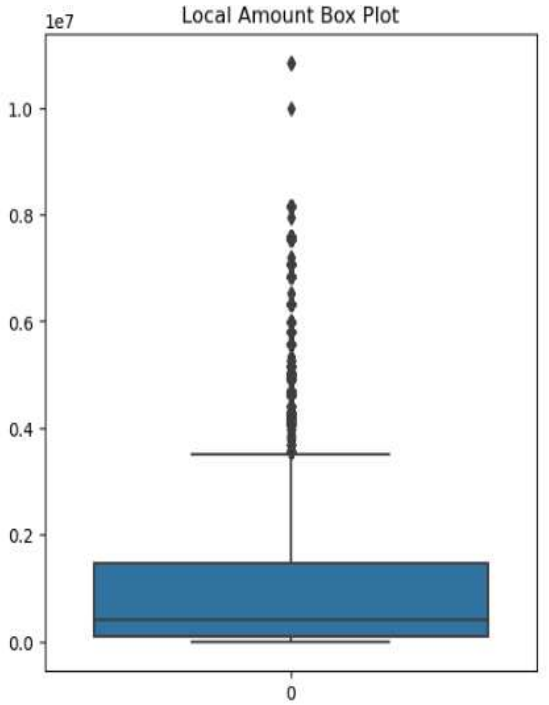
Univariate Analysis:

Customer Number : No Changes required

RECEIPT_DOC_NO : No Changes required

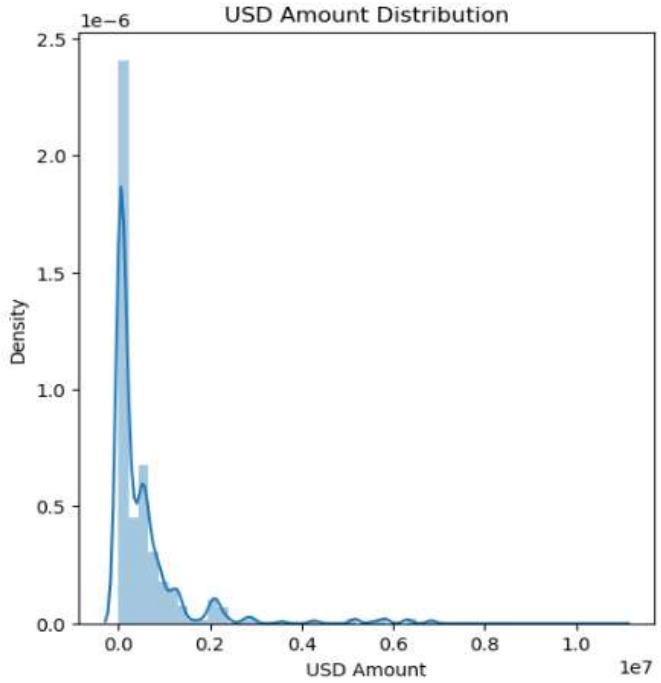
Local Amount : Dropping the 'Local Amount' column as it does not have a single currency value, and we already have 'USD Amount' column for bill amount.

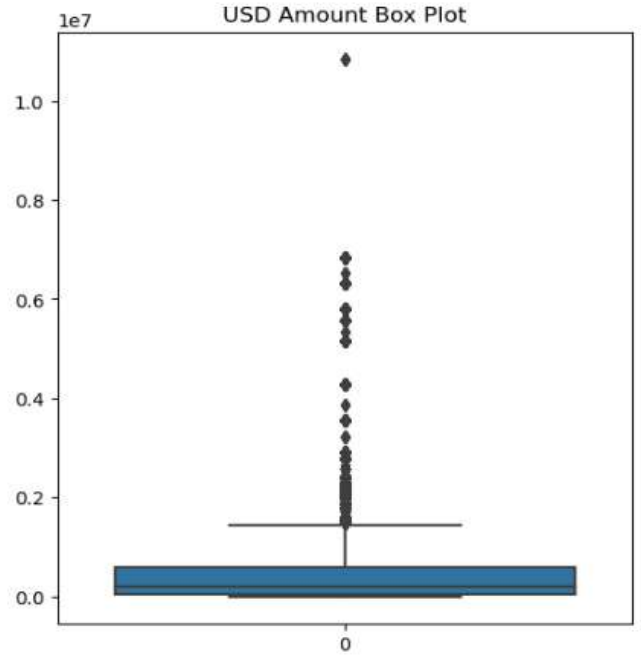




UNIVARIATE ANALYSIS

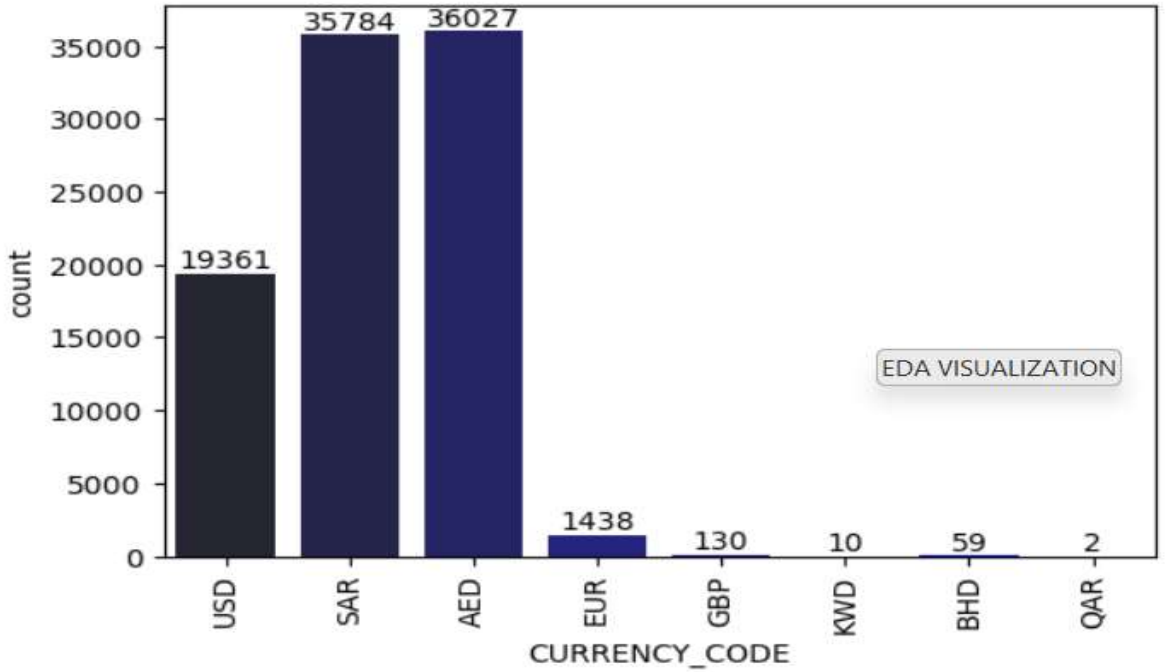
USD value: Since the USD value is distinct for each transaction and there are no outliers in the data that require adjustment, it can be taken into consideration.





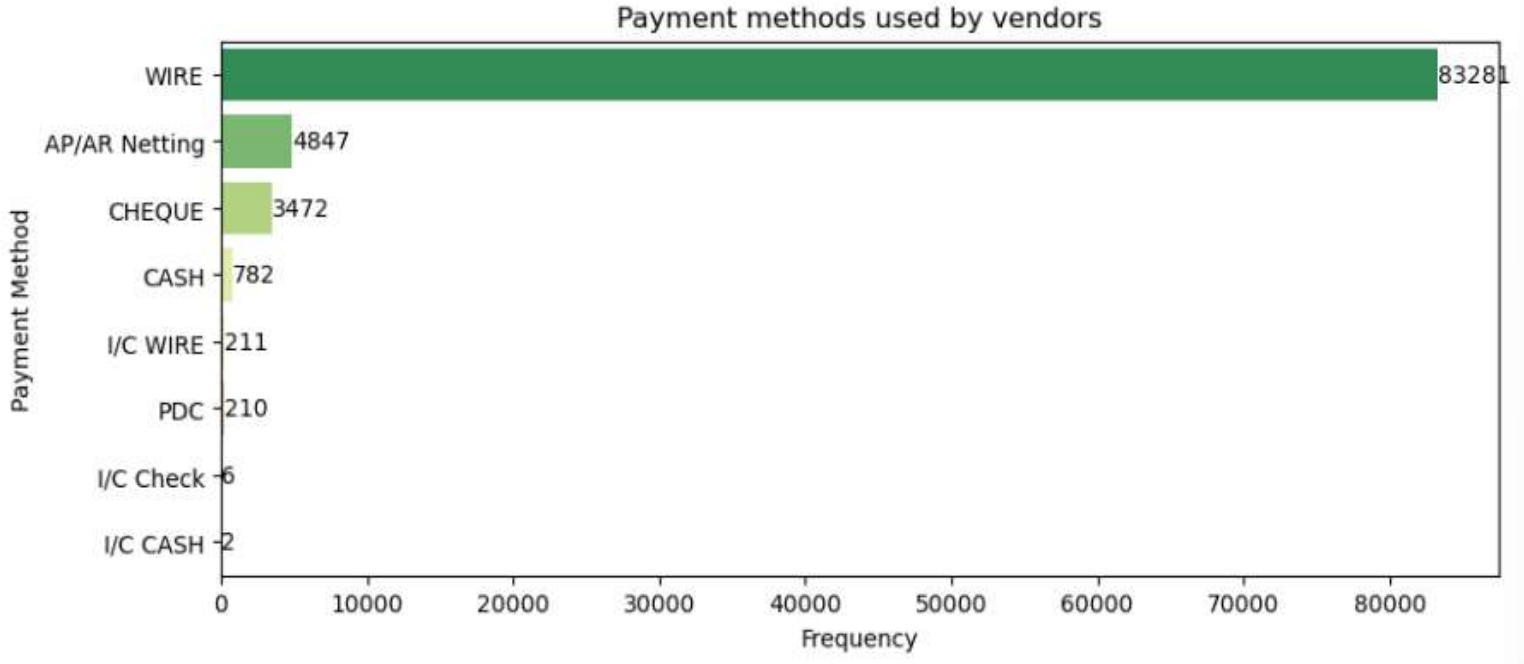
UNIVARIATE ANALYSIS

CURRENCY_CODE : Currency used for bill payments are mostly USD, SAR or AED.



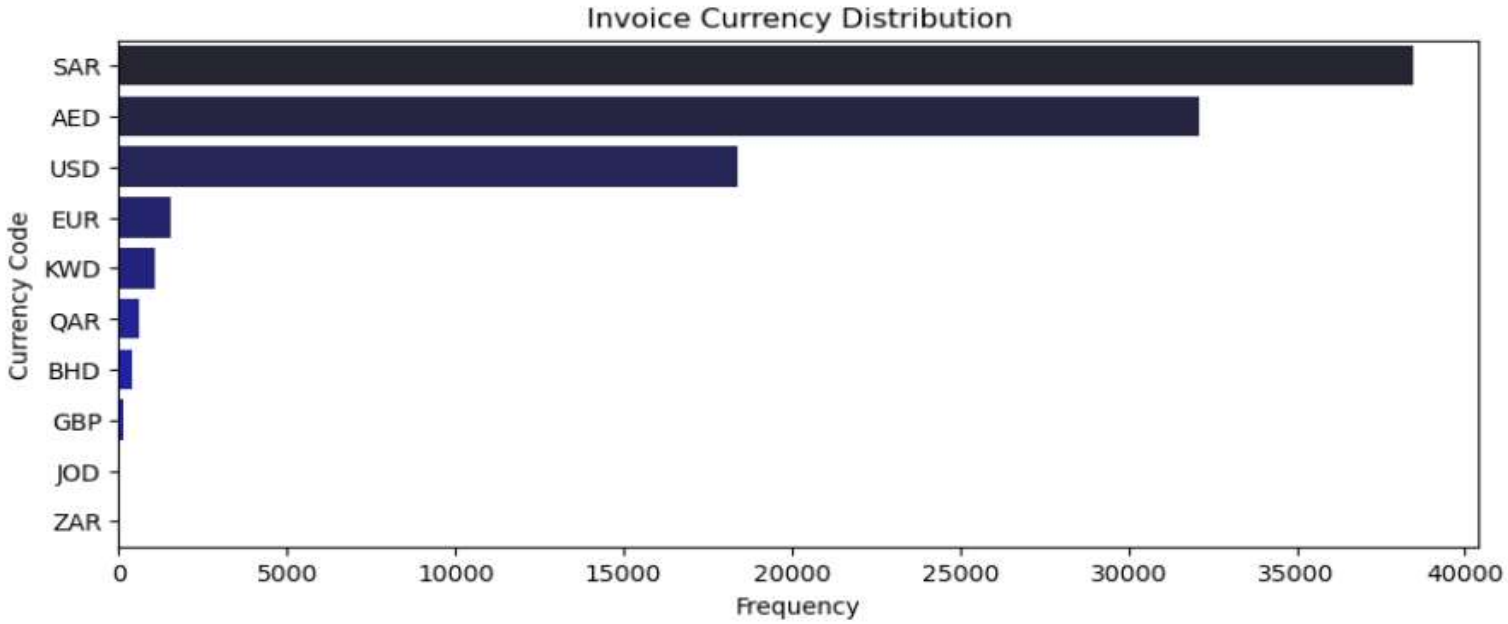
UNIVARIATE ANALYSIS

RECEIPT_METHOD: The most preferred payment method for bill payment is WIRE.



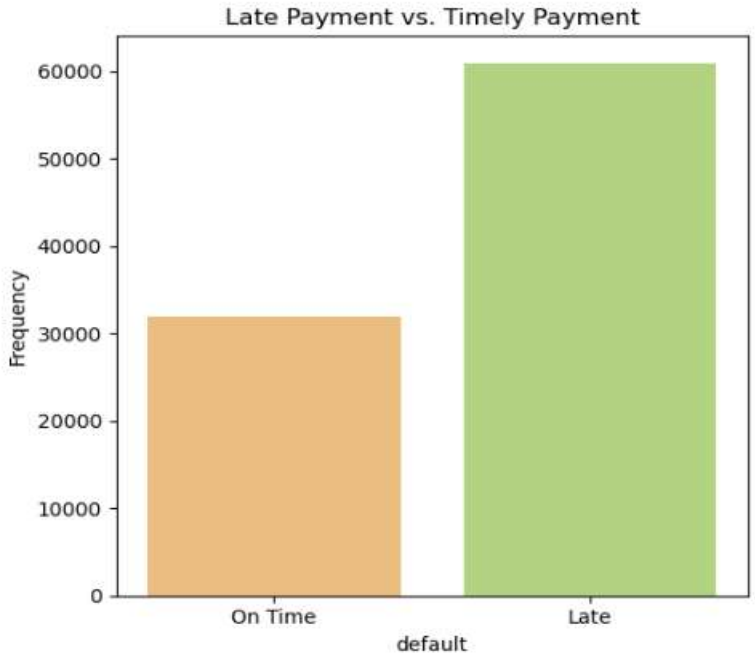
UNIVARIATE ANALYSIS

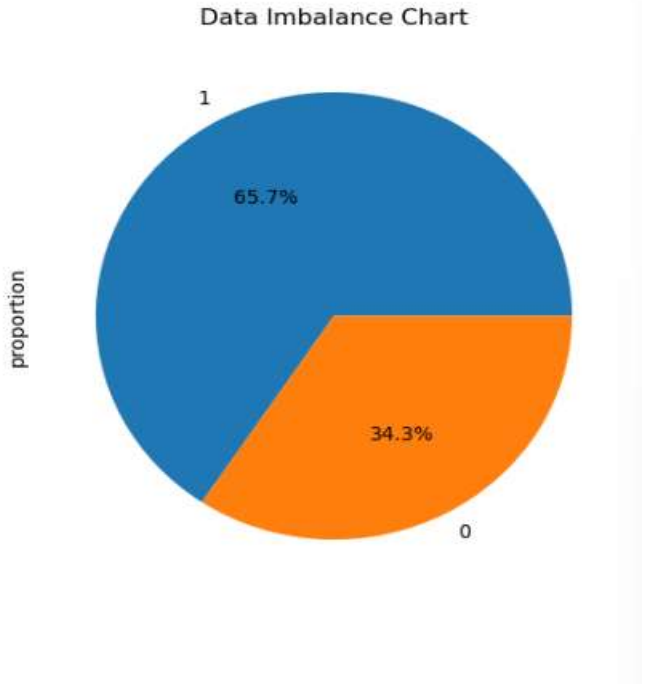
INVOICE_CURRENCY_CODE : Most number of invoices were generated in SAR, AED and USD currency.



Checking Data Imbalance between on Time Payment & Late Payment

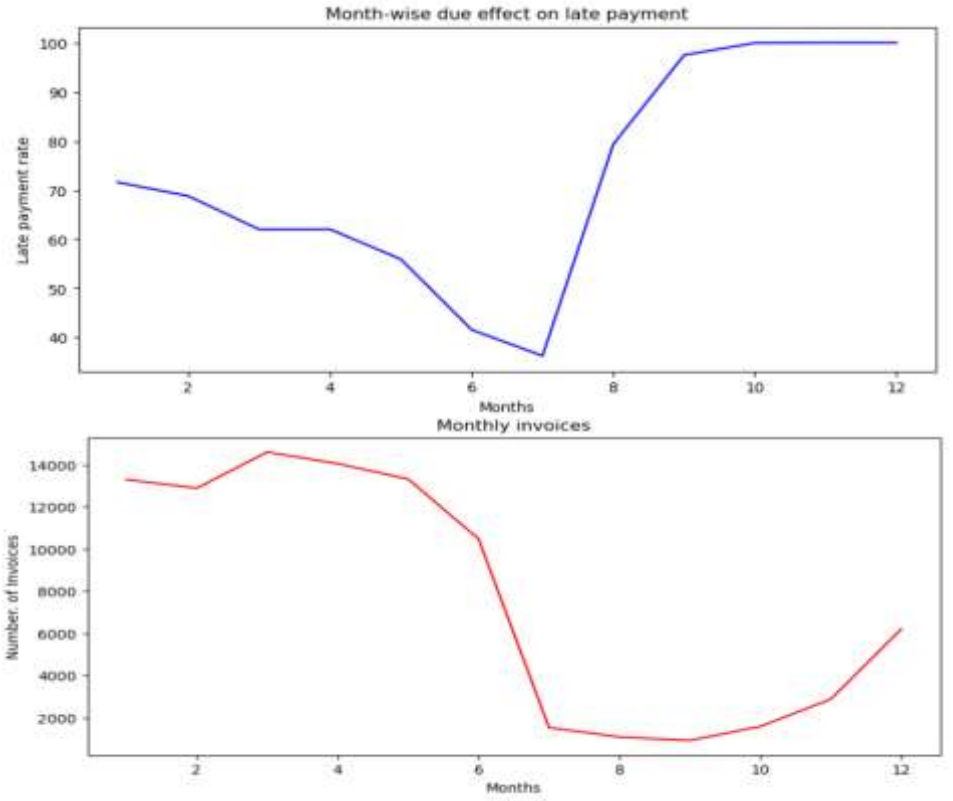
There is a good distribution of data in the target variable.





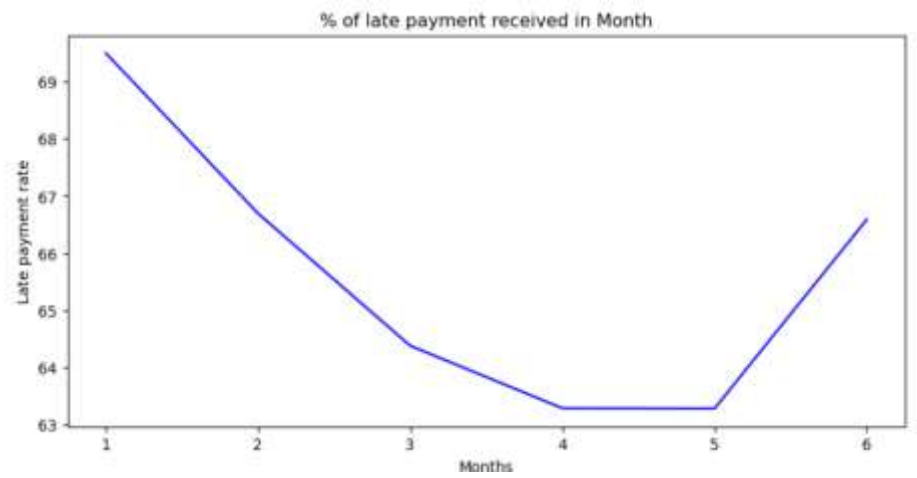
Bi-variate Analysis
1) On the basis of Due month

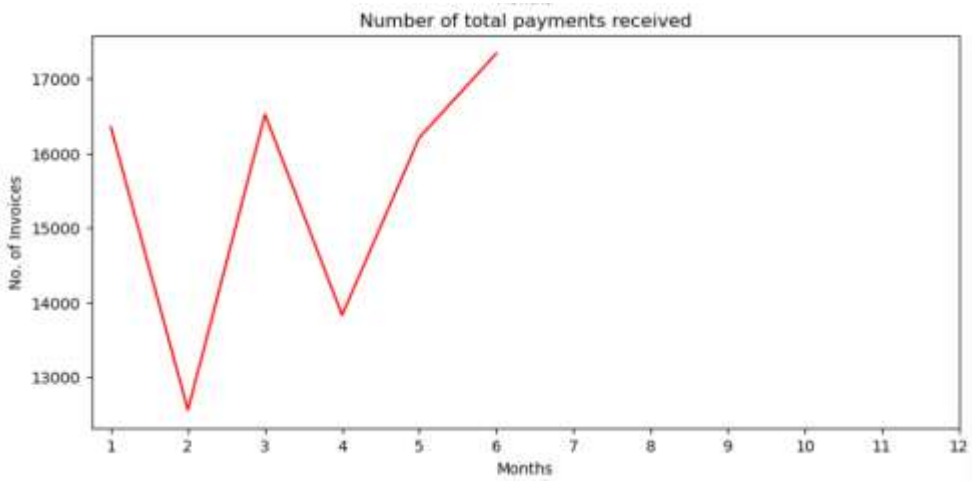
- In the 3rd month, the number of invoices is the highest and the late payment rate is relatively lower compared to other months with a high volume of invoices.
- The late payment rate in Month 7 is very low, likely due to the fact that the number of invoices is also low.
- The late payment rate rises sharply in the 2nd half of the year starting in the 7th month. In comparison to the first half of the year, there are less invoices.



2) On the basis of Receipt_date

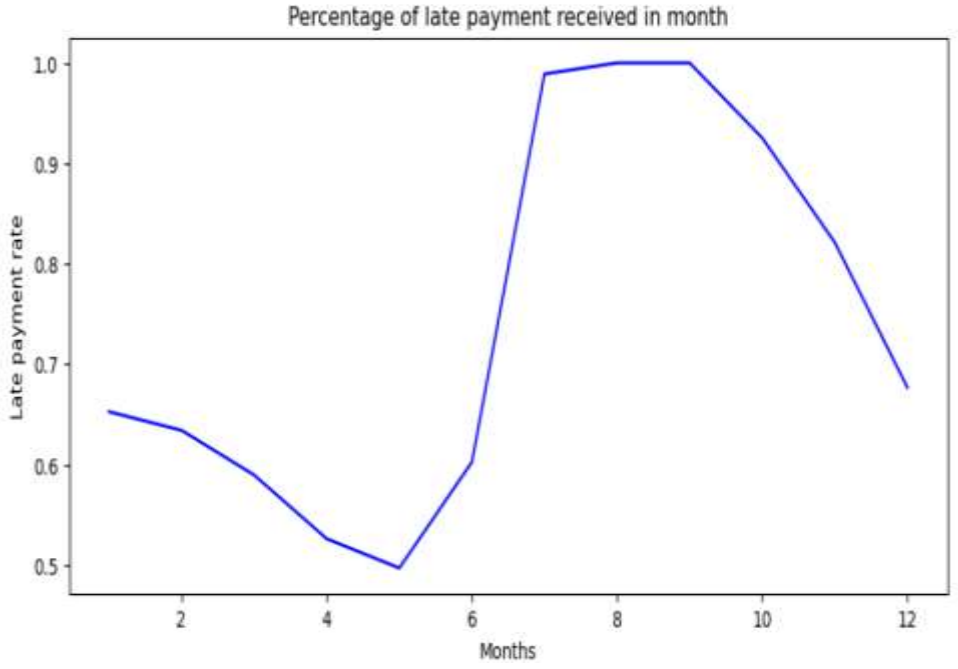
No payment received against any invoice from 7th month onwards.

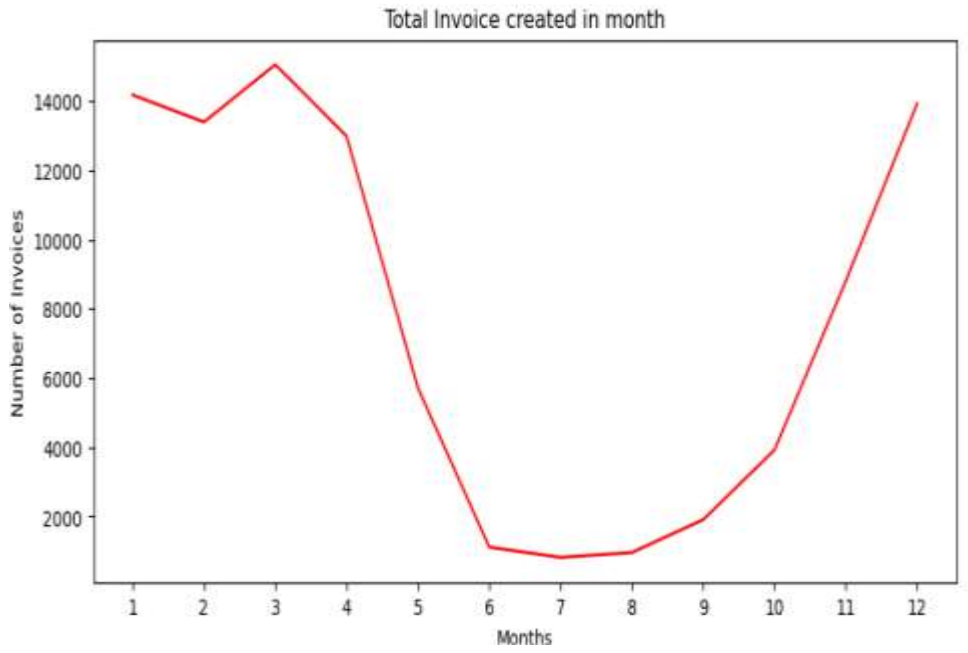




3) On the basis of Invoice Creation month

- Late payment rate is decreases from 1st to 5th month.
- For the months 7, 8 and 9 the late payment rate is very high.



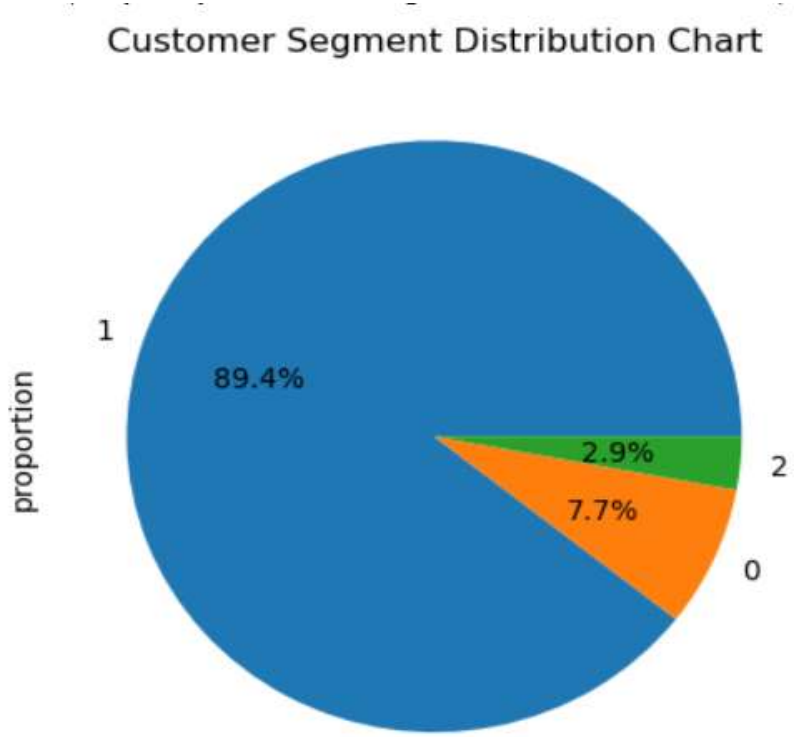


PIE CHART FOR CUSTOMER SEGMENT DISTRIBUTION

Customer Segmentation

- '1' Cluster -- Prolonged Invoice Payment
- '2' Cluster -- Early Invoice Payment
- '0' Cluster -- Medium Invoice Payment

Here we can see that early customers comprise of 89.4% of customers medium and prolonged payers are 10.6% in total.



STEPS FOR MODEL BUILDING

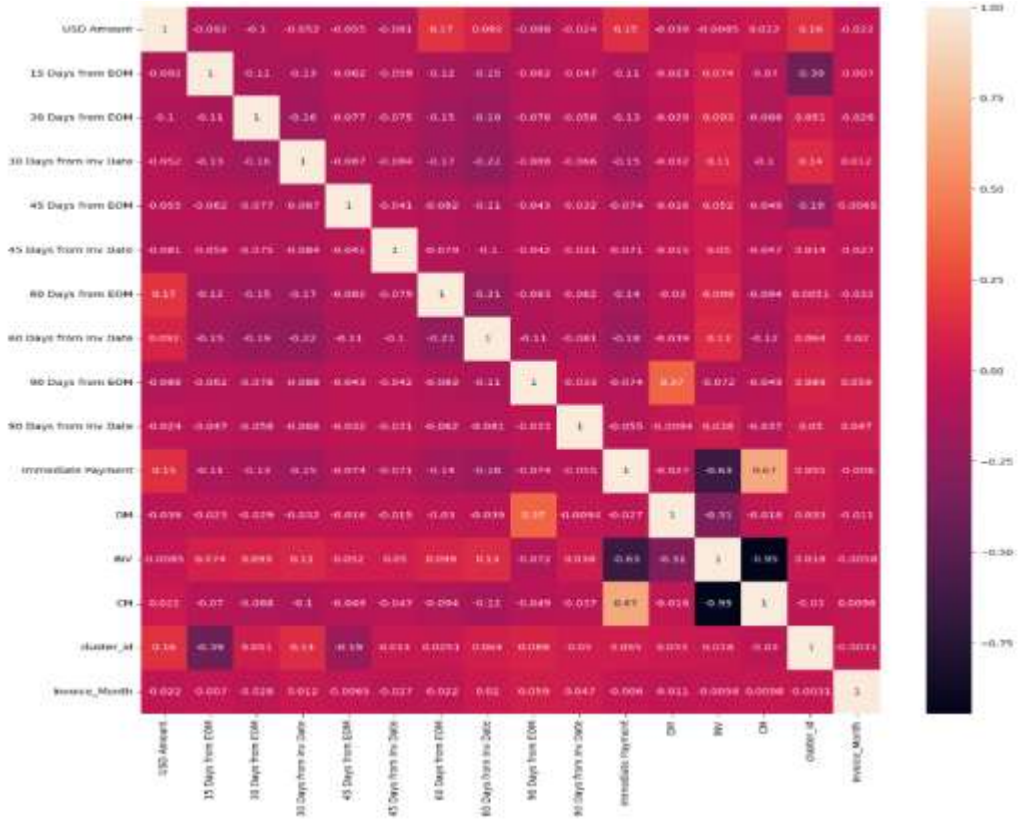
1) Data Preparation

2) Train and Test Split - 70:30 split

3) Feature Scaling

4) Plotting Heatmap for Correlation Matrix

CM & INV, INV & Immediate Payment, DM & 90days from EOM has high multicollinearity, hence dropping these columns.



MODEL BUILDING – LOGISTIC REGRESSION

Since the ‘p-value’ and ‘VIF’ fall within an acceptable range, this model can be used.

	Features	VIF
11	cluster_id	4.02
12	Invoice_Month	2.67
7	60 Days from Inv Date	2.01
3	30 Days from Inv Date	1.85
2	30 Days from EOM	1.58
6	60 Days from EOM	1.58
10	Immediate Payment	1.54
8	90 Days from EOM	1.29
5	45 Days from Inv Date	1.17
1	15 Days from EOM	1.15
9	90 Days from Inv Date	1.15
0	USD Amount	1.11
4	45 Days from EOM	1.09

Generalized Linear Model Regression Results

Dep. Variable:	default	**No. Observations:**	64967
Model:	GLM	**Df Residuals:**	64953
Model Family:	Binomial	**Df Model:**	13
Link Function:	Logit	**Scale:**	1.0000
Method:	IRLS	**Log-Likelihood:**	-30149.
Date:	Wed, 06 Nov 2024	**Deviance:**	60298.
Time:	14:48:36	**Pearson chi2:**	6.31e+04
No. Iterations:	7	**Pseudo R-squ. (CS):**	0.3021
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9276	0.051	18.107	0.000	0.827	1.028
USD Amount	-0.0450	0.012	-3.769	0.000	-0.068	-0.022
15 Days from EOM	2.5535	0.109	23.465	0.000	2.340	2.767
30 Days from EOM	-2.2725	0.052	-43.388	0.000	-2.375	-2.170
30 Days from Inv Date	0.2621	0.052	5.087	0.000	0.161	0.363
45 Days from EOM	0.3054	0.069	4.399	0.000	0.169	0.442
45 Days from Inv Date	-0.2915	0.063	-4.655	0.000	-0.414	-0.169
60 Days from EOM	-2.1775	0.052	-41.559	0.000	-2.280	-2.075
60 Days from Inv Date	-0.1743	0.049	-3.528	0.000	-0.271	-0.077
90 Days from EOM	-0.4685	0.061	-7.631	0.000	-0.589	-0.348
90 Days from Inv Date	-0.9799	0.069	-14.177	0.000	-1.115	-0.844
Immediate Payment	3.1450	0.105	29.863	0.000	2.939	3.351
cluster_id	-0.4302	0.026	-16.855	0.000	-0.480	-0.380
Invoice_Month	0.0954	0.003	37.651	0.000	0.090	0.100

MODEL BUILDING – LOGISTIC REGRESSION

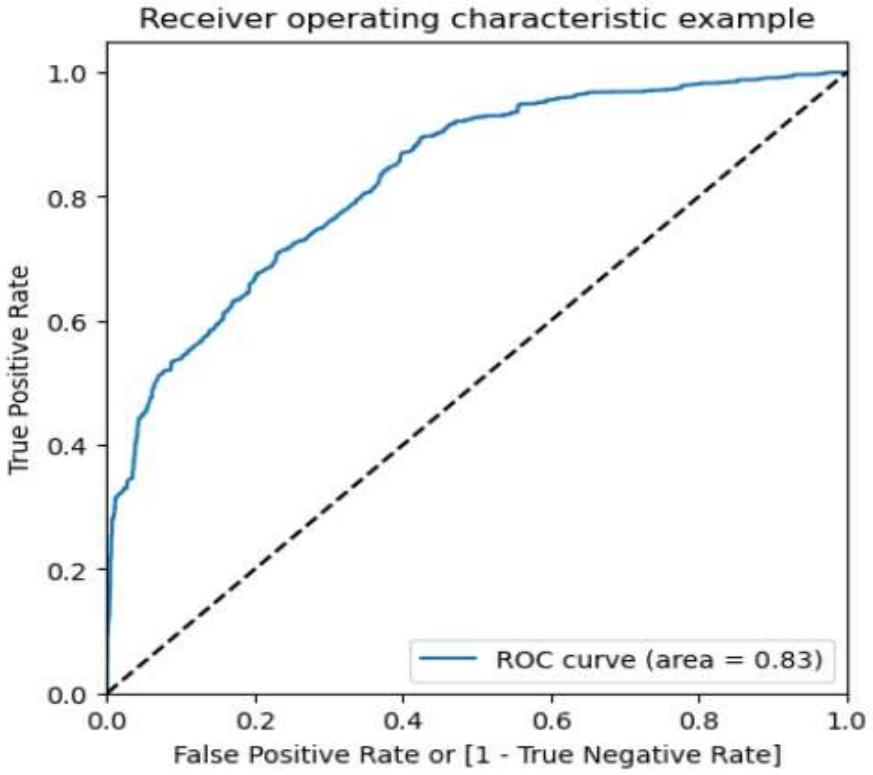
First Model:

Accuracy is 0.7709

Precision is 0.8049

Recall is 0.8585

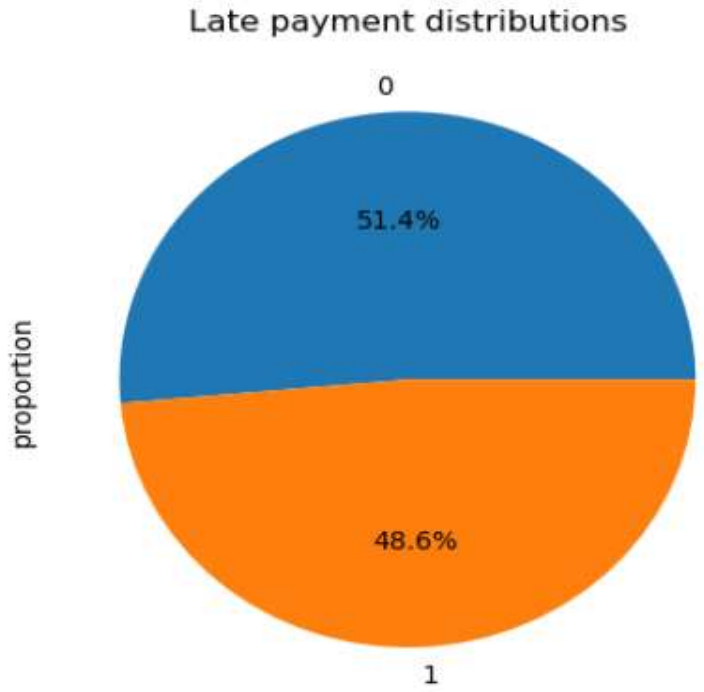
- AUC = 0.83 which shows the model is good.
- With this model our train and test accuracy is almost same around 77.1%



MODEL BUILDING – RANDOM FOREST (CLASSIFICATION MODEL)

Second Model :

	precision	recall	f1-score	support
0	0.97	0.90	0.94	22376
1	0.95	0.99	0.97	42591
accuracy			0.96	64967
macro avg	0.96	0.94	0.95	64967
weighted avg	0.96	0.96	0.96	64967



RECOMMENDATIONS:

1) Our clustering technique allows us to draw the following conclusions.

2) Compared to debit note or invoice type invoice classes, credit note payments have the highest delay rate; therefore, firm policies regarding payment collection should be more stringent with regard to these invoice classes.

3) Goods type invoices had significantly greater payment delay rates than non-goods types and hence can be subjected to stricter payment policies.

4) It is advised to concentrate more on lesser value payments because they make up the majority of transactions and are also more likely to have late payments. Depending on the billing amount, the business may impose penalties; the lower the bill, the higher the percentage of late payment penalties. This must be the last option, of course.

RECOMMENDATIONS:

5) Three categories—0, 1, and 2—which stand for medium, prolonged, and early payment durations, respectively—were created by clustering customer segments. Cluster 1 consumers should receive special attention because it was discovered that their delay rates were noticeably higher than those of early and middle days of payment.

6) Due to their high likelihood rates, the companies listed above with the highest probability and total and delayed payment counts should be given priority and more attention.