CUSTOMER PAYMENT ANALYTICS

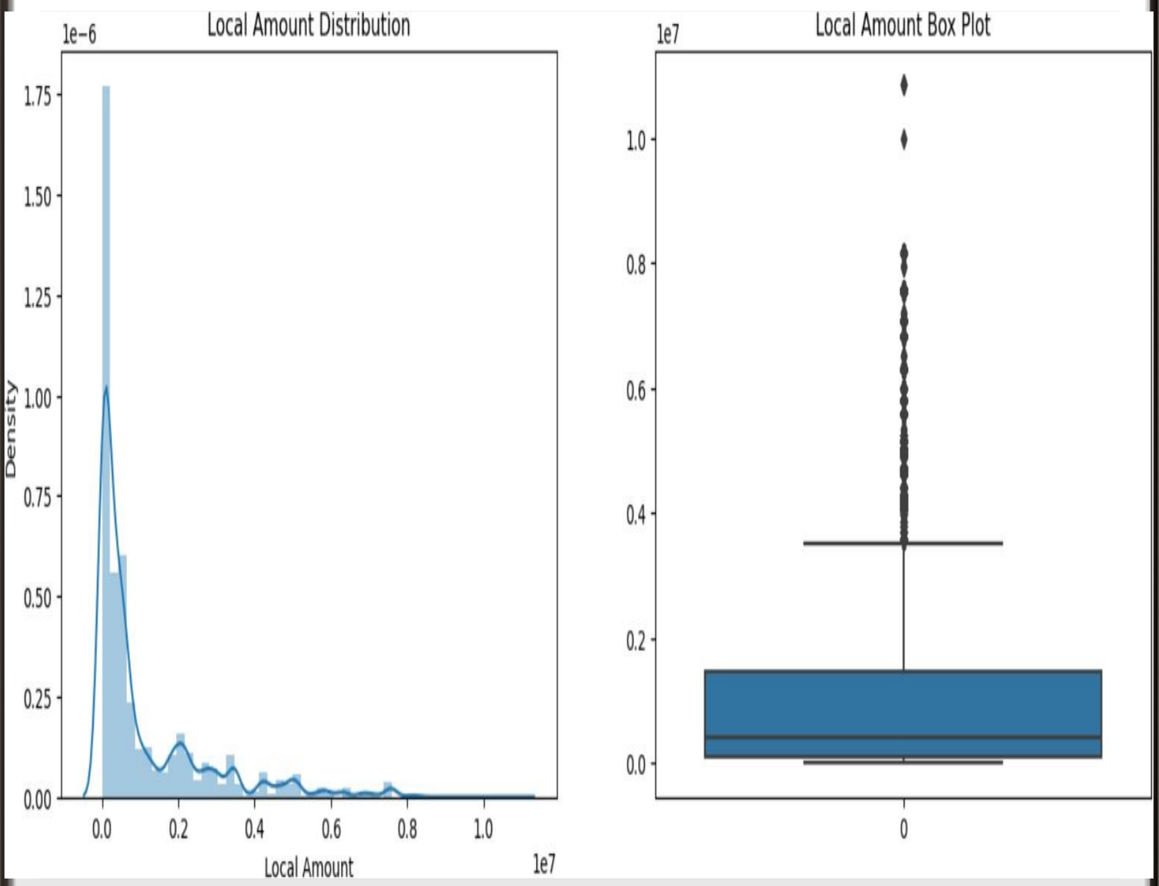
E-COMMERCE AND RETAIL B2B CASE STUDY

PROBLEM STATEMENT

- Schuster is a multinational retail company dealing in sports goods and accessories
- Schuster conducts significant business with hundreds of its vendors, with whom it has credit arrangements.
- Unfortunately, not all vendors respect credit terms and some of them tend to make
- payments late.
- Schuster levies heavy late payment fees, although this procedure is not beneficial to either party in a long-term business relationship.
- Every time a transaction of goods takes place with a vendor, the accounting team
- raises an invoice and shares it with the vendor.
- Schuster would try to understand its customers' payment behavior and predict the likelihood of late payments against open invoices.

BUSINESS OBJECTIVE

- Schuster would like to better understand the customers' payment behavior based on their past payment patterns (customer segmentation).
- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.
- 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.

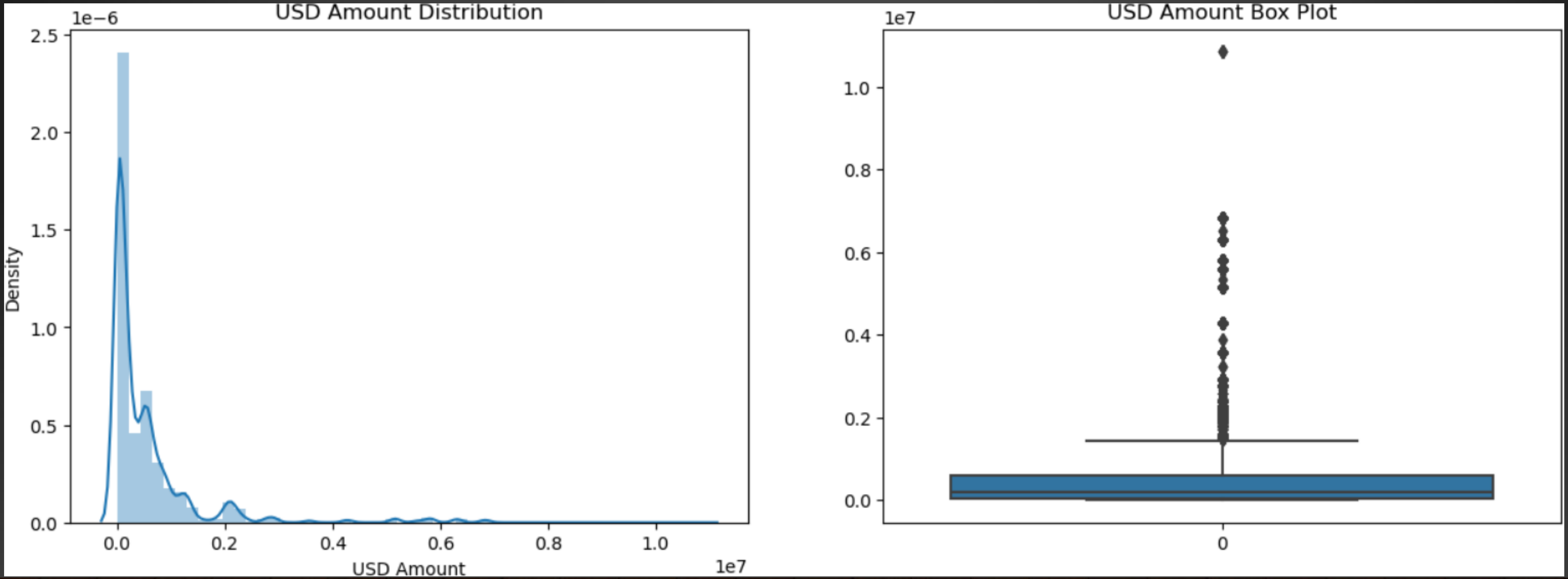


EXPLORATORY DATA ANALYSIS (EDA)

- **Uni-Variate Analysis**
  - **Customer Number :** No Changes required
  - **RECEIPT_DOC_NO :** No Changes required
  - **Local Amount :** Could be dropped as the local currencies are different & amounts will be not matching

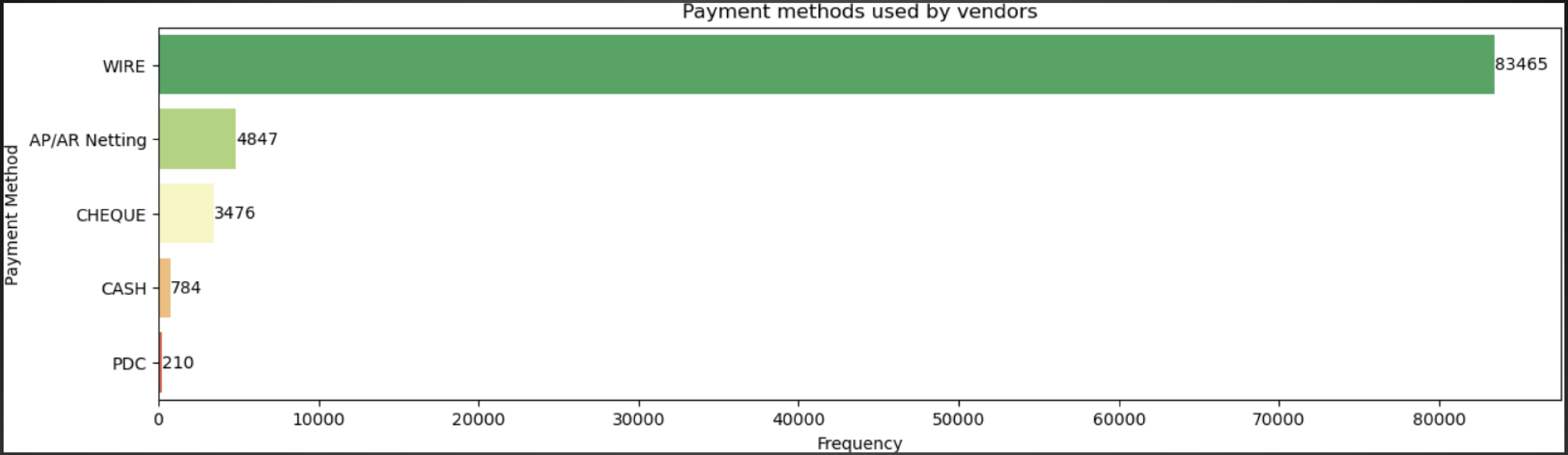
- **Uni-Variate Analysis**

  - USD Amount : USD amount can be considered as this is unique for all the transactions, also the data do not have any outlier which needs to be changed



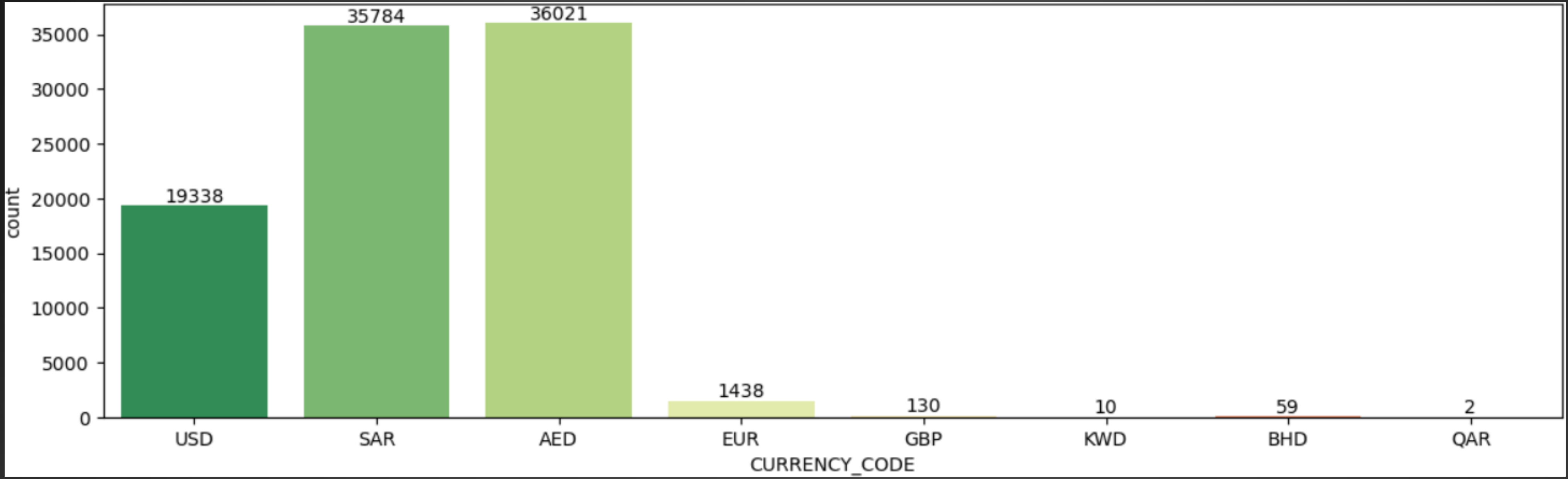
- Uni-Variate Analysis

  - **RECEIPT_METHOD :** The Most preferred method of payment is WIRE



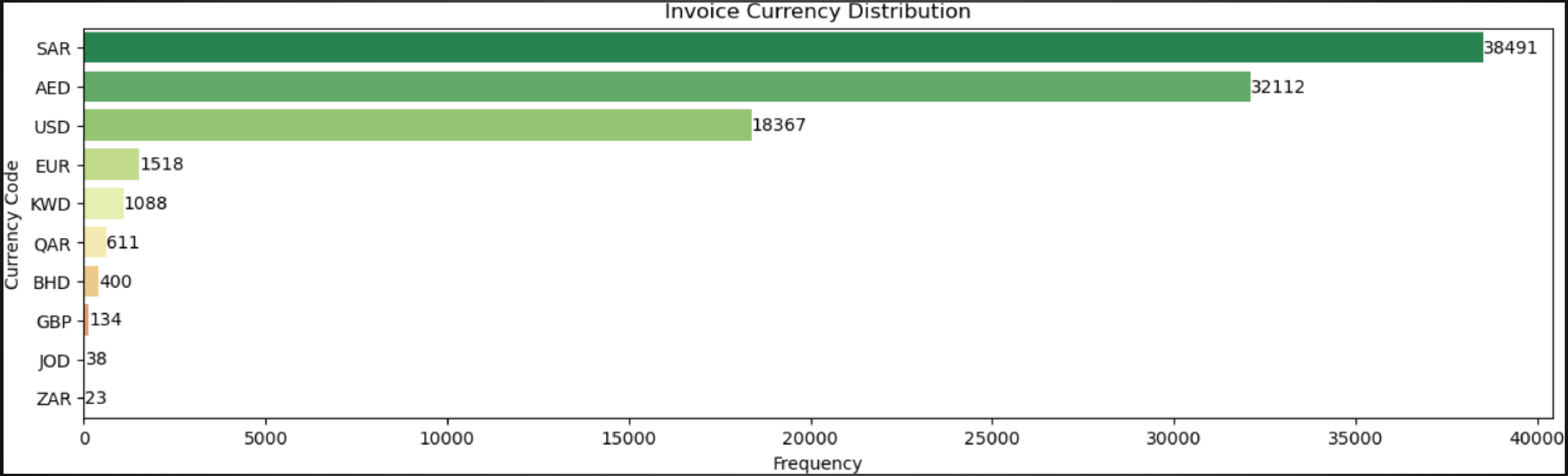
- **Uni-Variate Analysis**

  - **Currency_code :** The Most used currencies are SAR, AED & USD

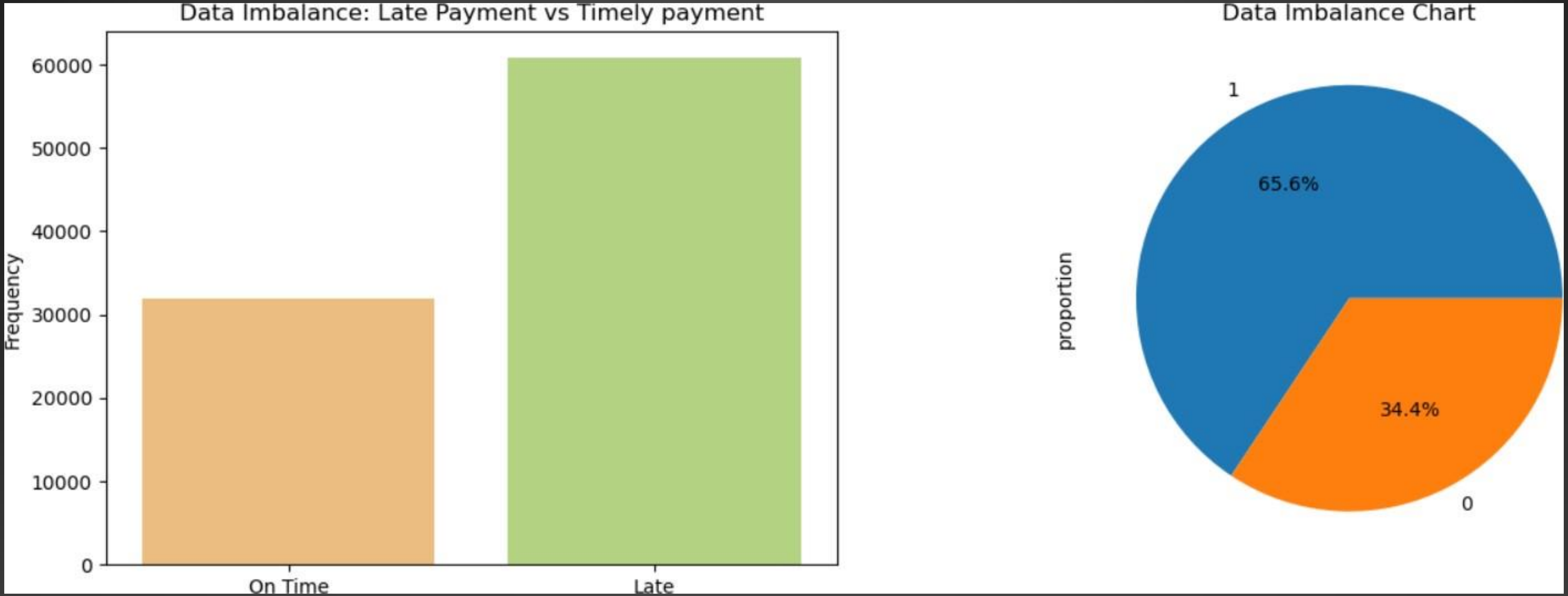


- **Uni-Variate Analysis**

  - **Invoice_Currency :** The Most used currencies are SAR, AED & USD, similar to Payment currencies.

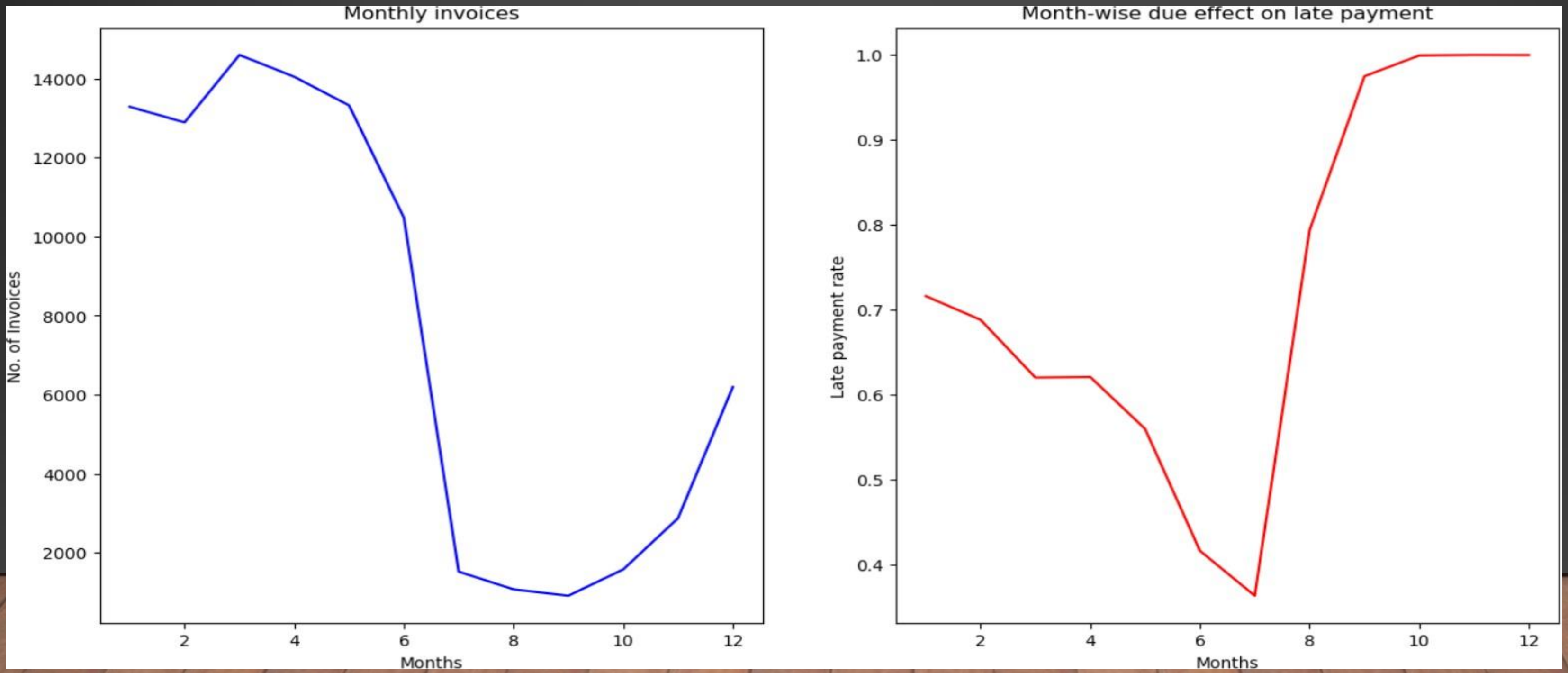


- **Checking Data imbalance between On-Time payment & Late payment**
  - There is no such data imbalance we can go ahead with the available data

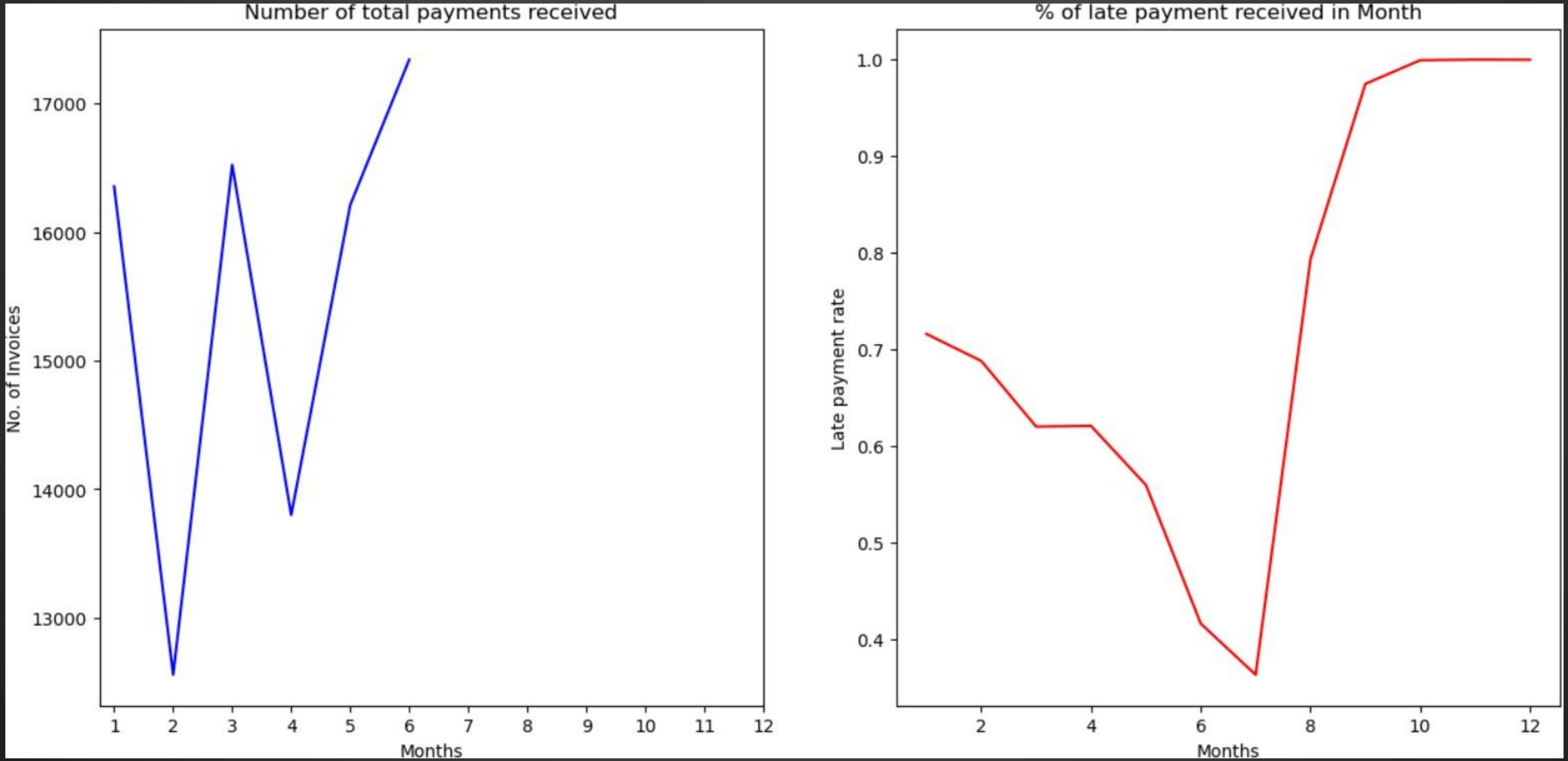


- **Bi-Variate Analysis**
  1. **Basis of Due month :**

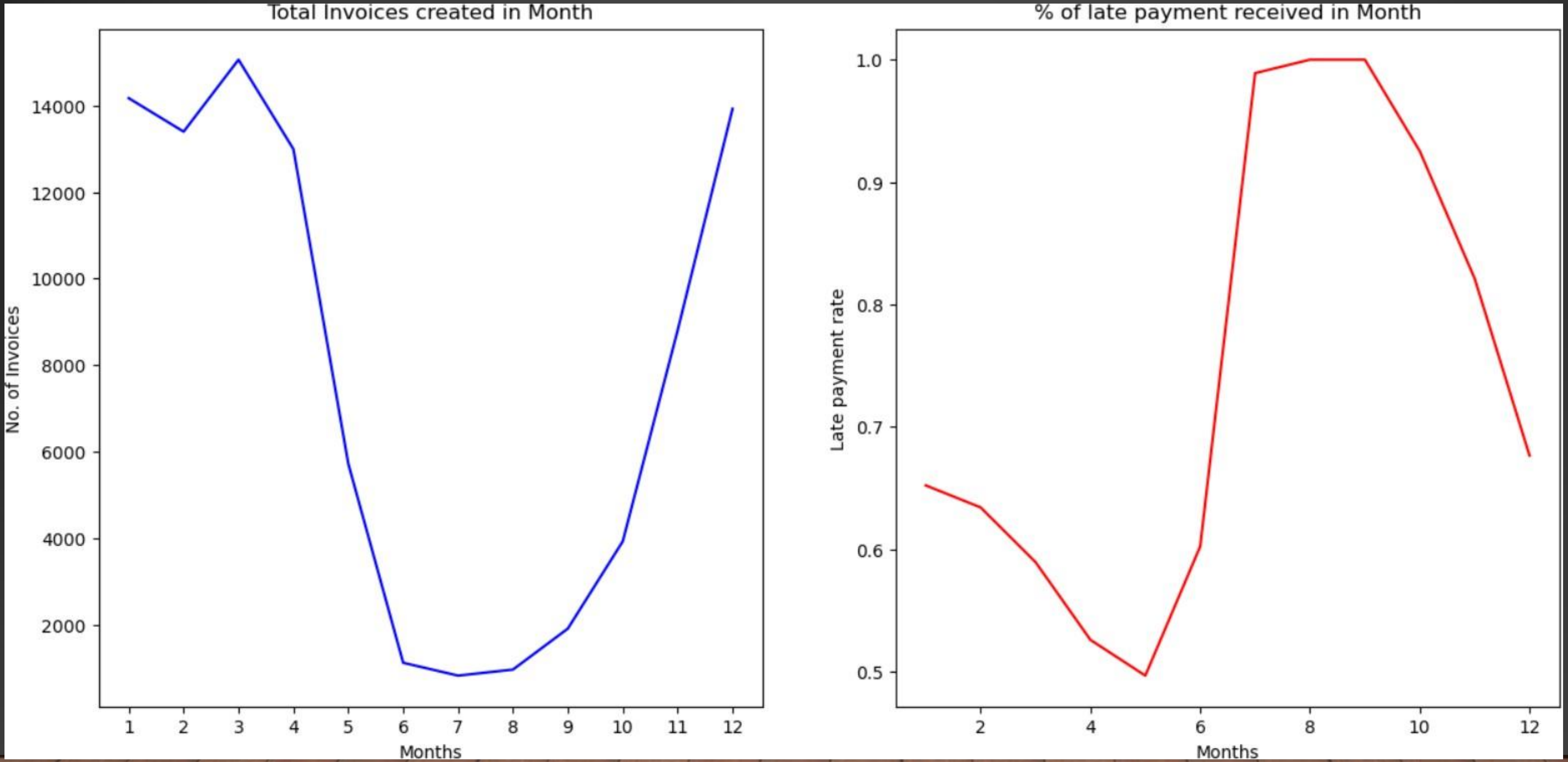
- For the 3rd month, the number of invoices is the highest and late payment rate is comparatively lower than other months with large number of invoices.
- Month 7 has the very low late payment rate, this can be because of the fact that the number of invoices is also low.
- In the 2nd half of the year, the late payment increases steeply from 7th month onwards. The number of invoices are comparatively lower than the first half of the year.



- **Bi-Variate Analysis**
  - **2. Analysis on Basis of Receipt Months :**
    - No payment received against any invoices from 7th month onwards even having due.



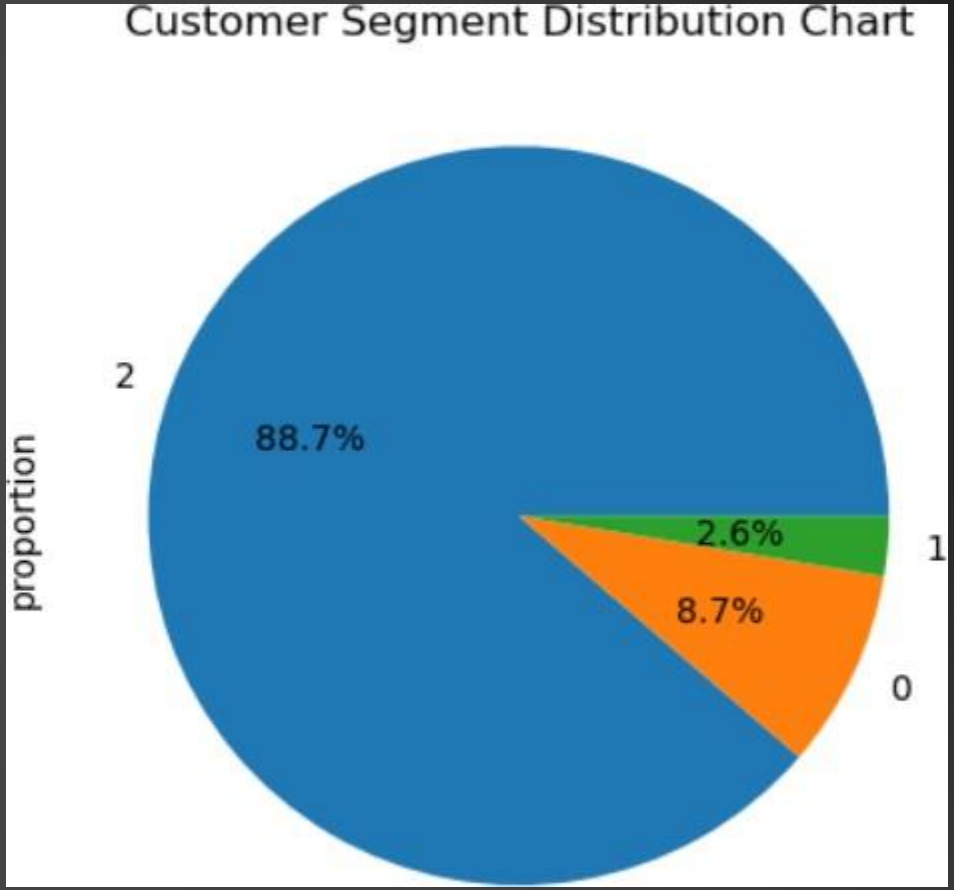
- **Bi-Variate Analysis**
  - **3. Analysis on Basis of Invoice Creation Months** :
    - Late payment rate is decreases from 1st to 5th month.
    - For the months 7, 8 and 9, the late payment rate is very high.



ANALYSIS ON OPEN_INVOICE_DATA" DATASET

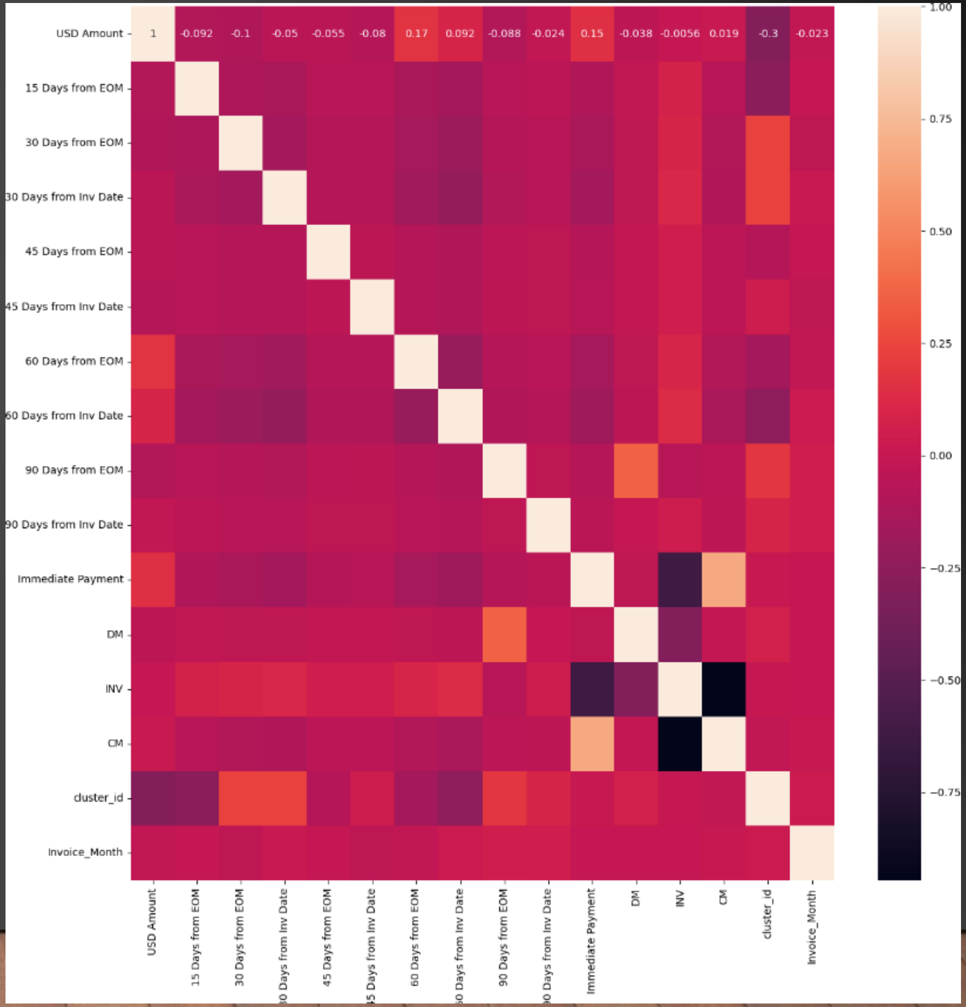
- **Customer Segmentation**
  - '1' Cluster -- Prolonged Invoice Payment
  - '2' Cluster -- Early Invoice Payment
  - '0' Cluster -- Medium Invoice Payment

- We can say that Early payers comprise of 88.7% of customers whereas medium and prolonged payers are 11.3% in total



STEPS FOR MODEL BUILIDNG

- 1. Data Preparation
- 2. Train Test Split - 70:30 Spilt
- 3. Feature Scaling
- 4. Plotting Heatmap for Correlation matrix
  - "CM" & "INV", "INV" & "Immediate Payment", "DM" & "90 days" from "EOM" has high multicollinearity, hence dropping these columns.



MODEL BUILDING - LOGISITIC REGRESSION

Generalized Linear Model Regression Results

Dep. Variable:	Default	No. Observations:	64947
Model:	GLM	Df Residuals:	64933
Model Family:	Binomial	Df Model:	13
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-30170.
Date:	Mon, 06 May 2024	Deviance:	60339.
Time:	10:13:39	Pearson chi2:	6.34e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.3012
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.7495	0.050	15.124	0.000	0.652	0.847
USD Amount	-0.1054	0.012	-8.748	0.000	-0.129	-0.082
15 Days from EOM	2.6146	0.108	24.267	0.000	2.403	2.826
30 Days from EOM	-2.2548	0.052	-42.950	0.000	-2.358	-2.152
30 Days from Inv Date	0.2638	0.052	5.102	0.000	0.162	0.365
45 Days from EOM	0.3968	0.070	5.704	0.000	0.260	0.533
45 Days from Inv Date	-0.3347	0.063	-5.338	0.000	-0.458	-0.212
60 Days from EOM	-2.2158	0.053	-41.704	0.000	-2.320	-2.112
60 Days from Inv Date	-0.2641	0.051	-5.219	0.000	-0.363	-0.165
90 Days from EOM	-0.4898	0.062	-7.953	0.000	-0.611	-0.369
90 Days from Inv Date	-1.0483	0.069	-15.203	0.000	-1.183	-0.913
Immediate Payment	3.0618	0.103	29.634	0.000	2.859	3.264
cluster_id	-0.1355	0.012	-11.123	0.000	-0.159	-0.112
Invoice_Month	0.0978	0.003	38.542	0.000	0.093	0.103

First Model

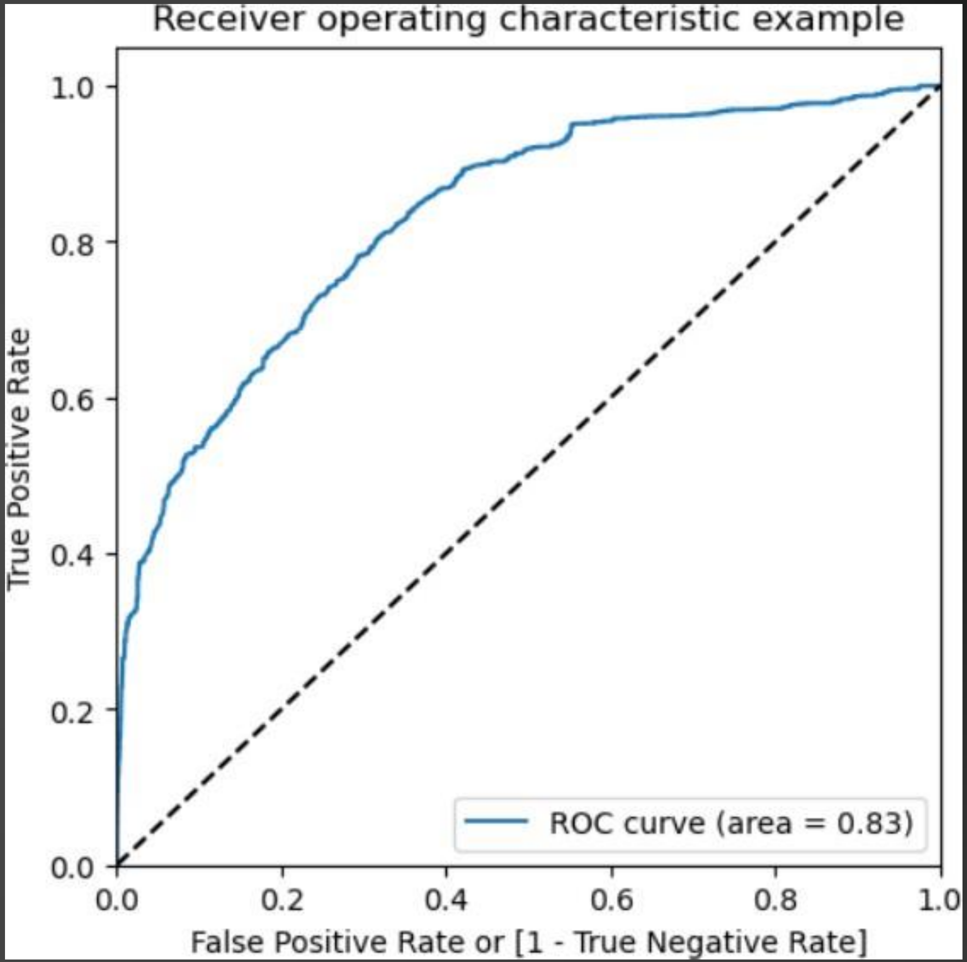
Both the "p-value" and "VIF" are in acceptable range, hence going ahead with this model.

	Features	VIF
12	Invoice_Month	2.67
11	cluster_id	2.60
3	30 Days from Inv Date	1.66
2	30 Days from EOM	1.52
7	60 Days from Inv Date	1.45
10	Immediate Payment	1.36
6	60 Days from EOM	1.31
8	90 Days from EOM	1.25
0	USD Amount	1.20
1	15 Days from EOM	1.14
9	90 Days from Inv Date	1.12
5	45 Days from Inv Date	1.10
4	45 Days from EOM	1.08

MODEL BUILDING - LOGISITIC REGRESSION

First Model

```
Accuracy is 0.7723369858092329
Precision is 0.8089661576557986
Recall is 0.8565089799272215
```



•AUC = 0.83 which shows the model is good.

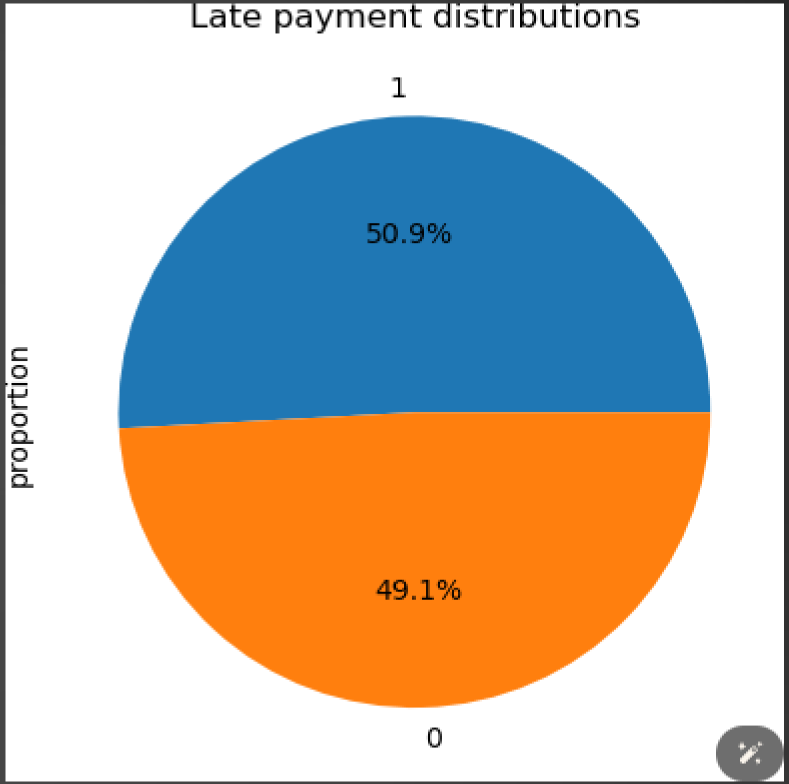
•With this model our train and test accuracy is almost same around 77.2 %

MODEL BUILDING - RANDOM FOREST (CLASSIFICATION MODEL)

SECOND MODEL

	precision	recall	f1-score	support
0	0.97	0.91	0.94	22352
1	0.95	0.98	0.97	42595
accuracy			0.96	64947
macro avg	0.96	0.95	0.95	64947
weighted avg	0.96	0.96	0.96	64947

Accuracy is : 0.9572882504195729



•From the above pie chart, we can observe that 50.9% payments in the open invoice data with AGE value negative(indicating due date not crossed)

RECOMMENDATIONS

- Credit Note Payments observe the greatest delay rate compared to Debit Note or Invoice type invoice classes, hence company policies on payment collection could be made stricter around such invoice classes.
- Goods type invoices had significantly greater payment delay rates than non-goods types and hence can be subjected to stricter payment policies.
- Since lower value payments comprise of the majority of the transactions, also late payments are seen more on lower value payments, it is recommended to focus more on those.
- The company can apply penalties depending on billing amount, the lesser the bill, the greater the percentage of penalty on late payments. Of course this has to be last resort.

RECOMMENDATIONS

- Customer segments were clustered into three categories, viz., 0,1 and 2 which mean medium, prolonged and early payment duration respectively.
- It was found that customers in cluster 1 (prolonged days) had significantly greater delay rates than early and medium days of payment.
- Hence cluster 1 customers should be paid extensive focus
- The above companies with the greatest probability and total & delayed payment counts should be first priority and should be focused on more due to such high probability rates