E-Commerce and Retail B2B Case Study

Rinkle Das
Priyadharshni Velayutham
Aditi Rani

PROBLEM STATEMENT:

- ***Schuster, a multinational retail company specializing in sports goods and accessories, faces challenges in managing its credit arrangements with the vendors. A significant number of vendors fail to adhere to agreed-upon credit terms, resulting in delayed payments. While Schuster imposes heavy late payment fees, this approach is not sustainable and strains long-term business relationships. Also, the company allocates resources to employees who manually follow up with vendors to ensure timely payments, leading to non-value-added activities, and financial inefficiencies. To address these issues, Schuster aims to analyze its vendors' payment behavior and develop a predictive model to forecast the likelihood of late payments for open invoices to proactively manage vendor relationships, and optimize resource allocation for improved financial outcomes.***

OBJECTIVES:

- *Schuster would like to better understand the customers' payment behaviour based on their past payment patterns (customer segmentation).*
- *It wants to use this information so that collectors can prioritise their work in following up with customers beforehand to get the payments on time.*
- *Proactively predict the likelihood of late payments for open invoices.*
- *Gain insights into the underlying factors influencing vendor payment behavior.*
- *Implement strategies to mitigate the risk of late payments and optimize the collections process.*

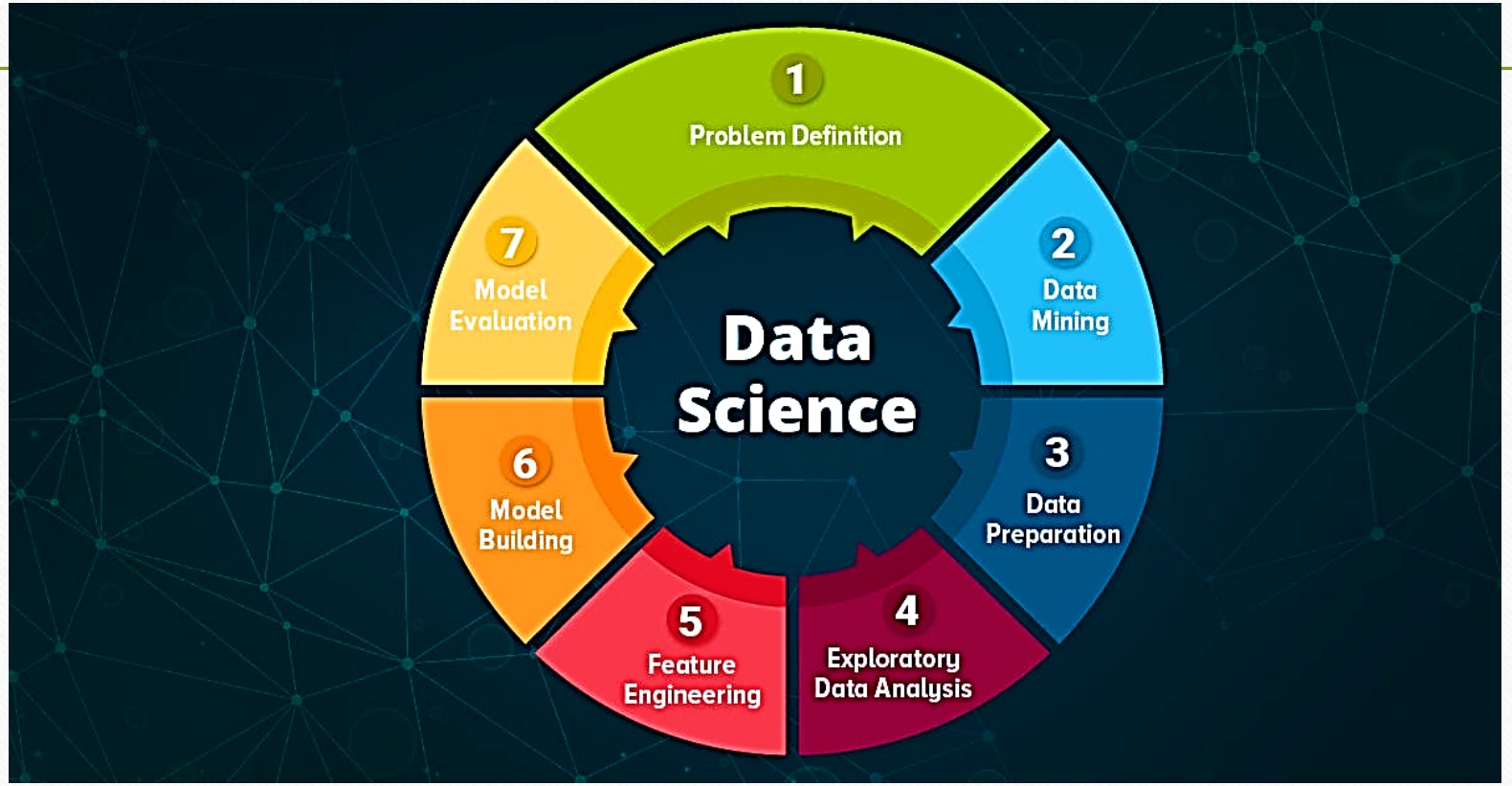
GOAL:

- *The goal is to build a model with the primary objective of identifying important predictor attributes that will help the business understand indicators of late payment.*

FILES USED:

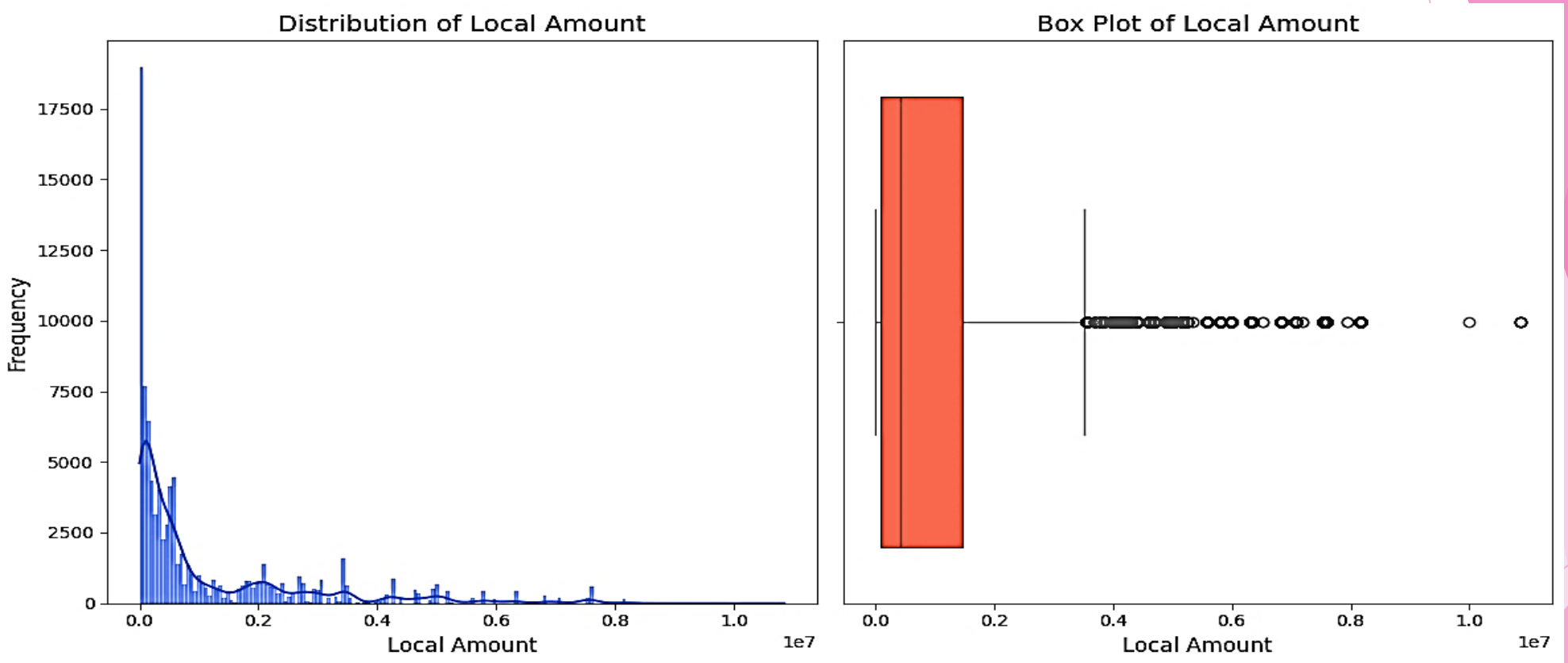
- *Received_Payments_Data.csv*
- *Open_Invoice_data.csv*
- *Data_Dictionary.xlsx*

Exploratory Data Analysis



UNIVARIATE ANALYSIS ON NUMERICAL COLUMNS:

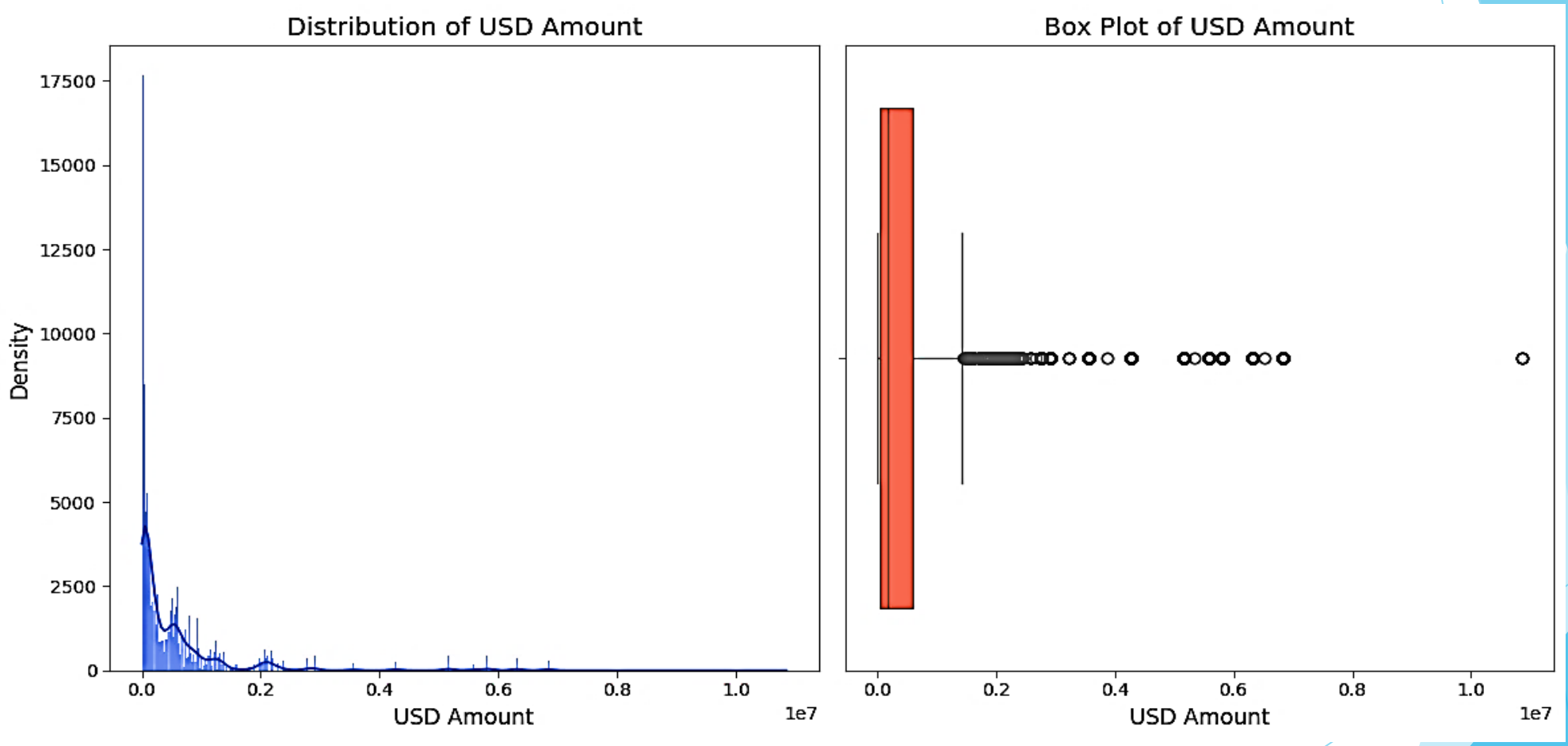
```
#Plotting the distribution of 'Local Amount'
```



There are a few very large "Local Amount" values that are significantly higher than the majority of the data. Also, removing 'Local Amount' as it lacks a single currency and 'USD Amount' is available.

UNIVARIATE ANALYSIS ON NUMERICAL COLUMNS:

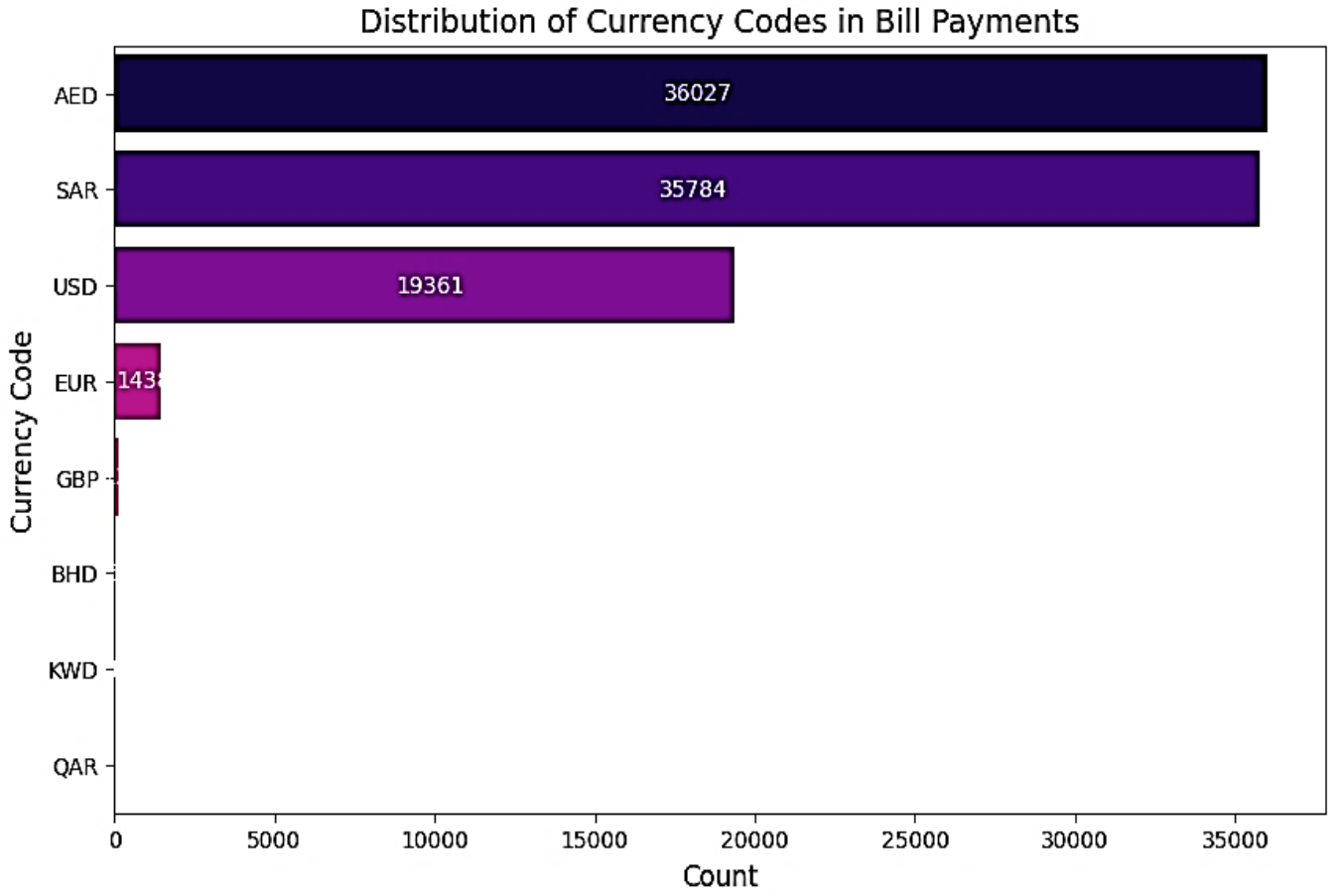
- `#Visualizing the distribution of 'USD Amount'`



Both the histogram and box plot consistently show a heavily skewed distribution with a significant number of outliers on the high end.

UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS:

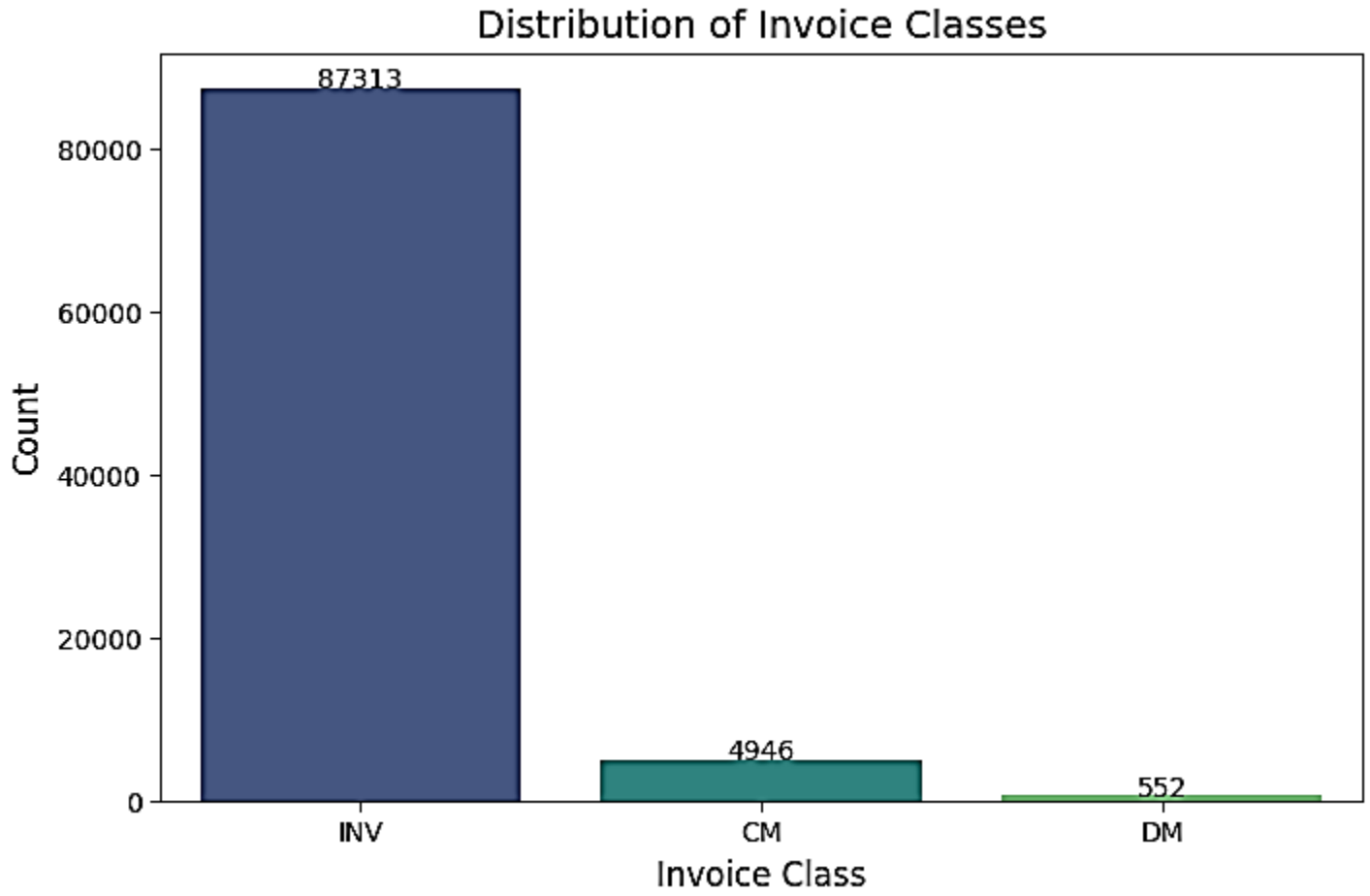
- `#Horizontal bar chart for 'CURRENCY_CODE' column`



Most transactions are processed in USD, SAR, and AED, with minimal use of other currencies.

UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS:

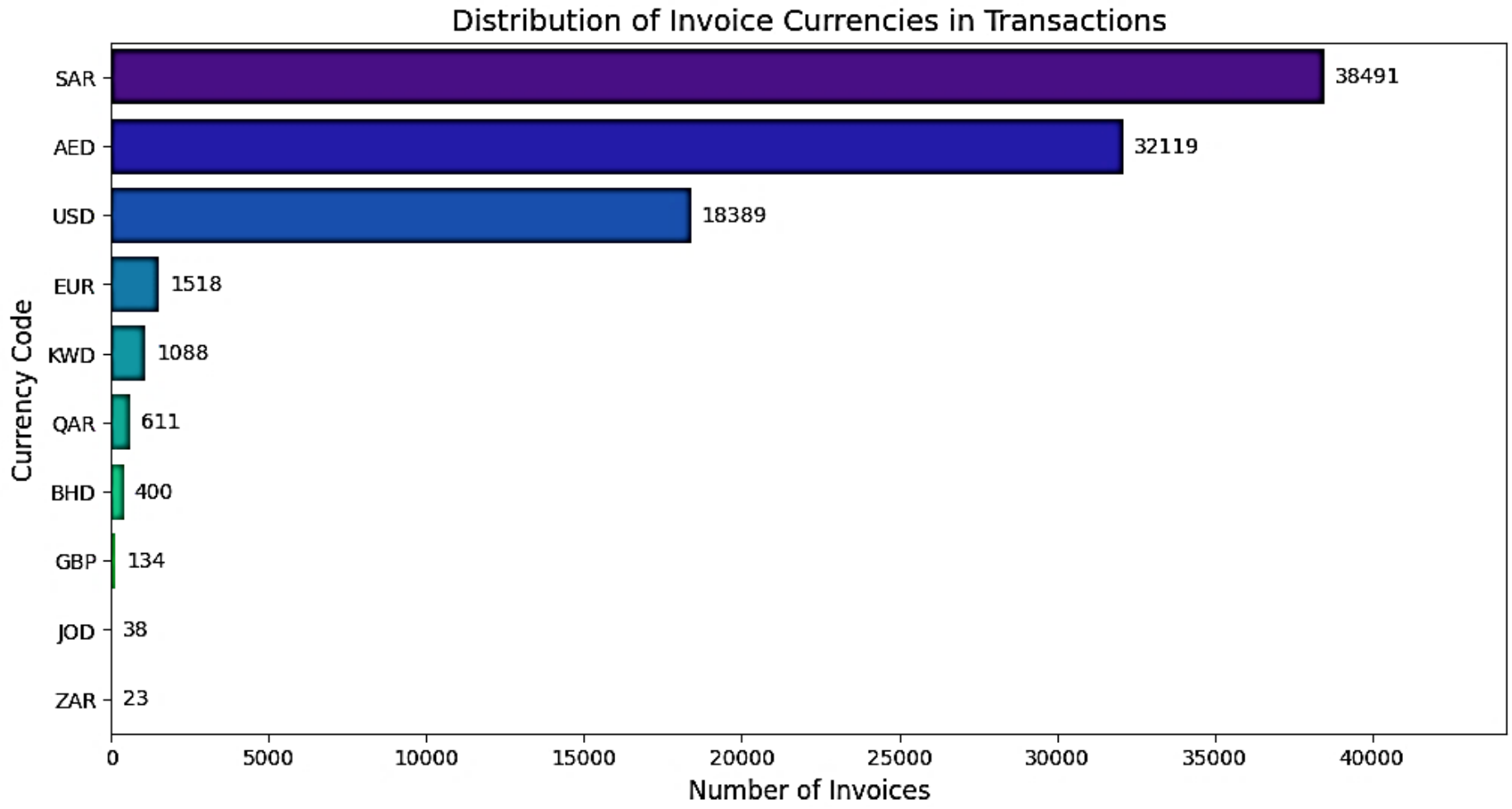
- *#Visualizing the INVOICE_CLASS distribution*



'INV' has the highest number of bills in the INVOICE_CLASS column.

UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS:

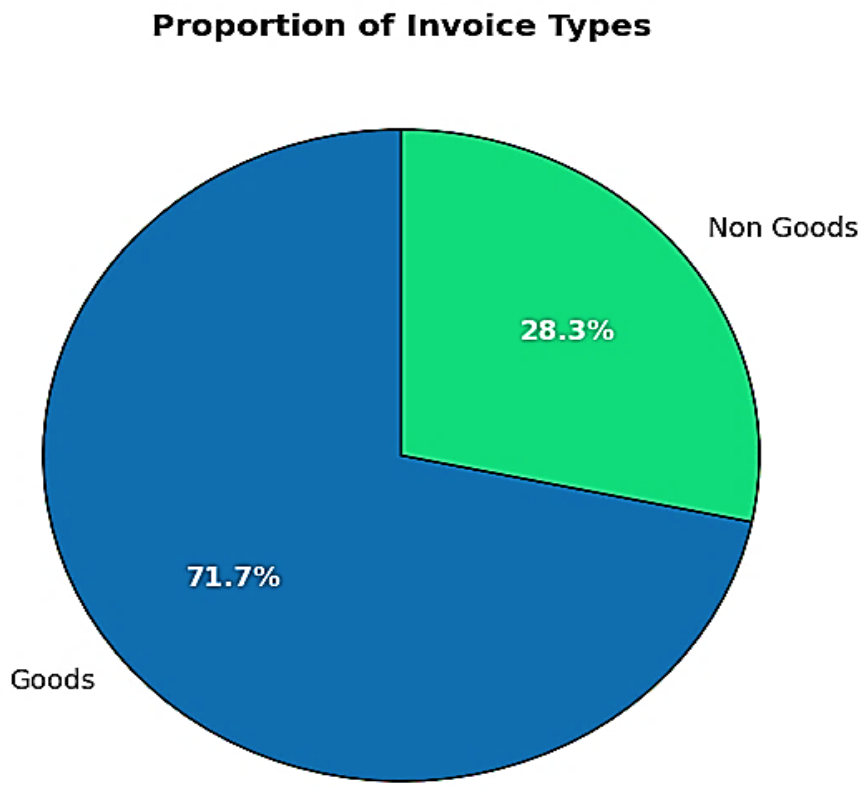
- *#Plotting Invoice Currency Distribution*



`Most invoices were generated in SAR, AED, and USD currencies.`

UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS:

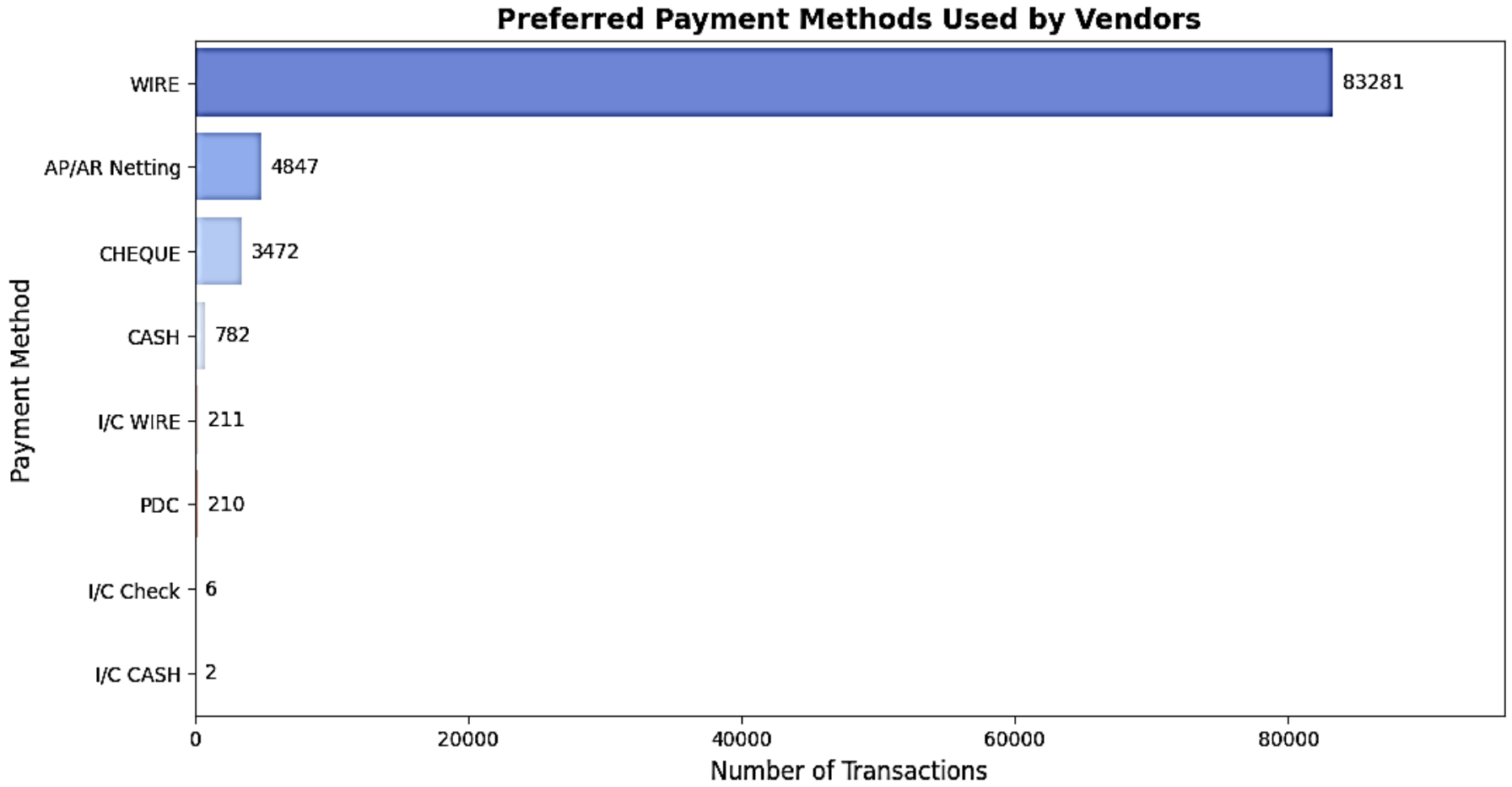
- `# Visualizing the distribution of INVOICE_TYPE`



The majority (71.7%) of invoices are for Goods and [28.3%] for Non-Goods.

UNIVARIATE ANALYSIS ON CATEGORICAL COLUMNS:

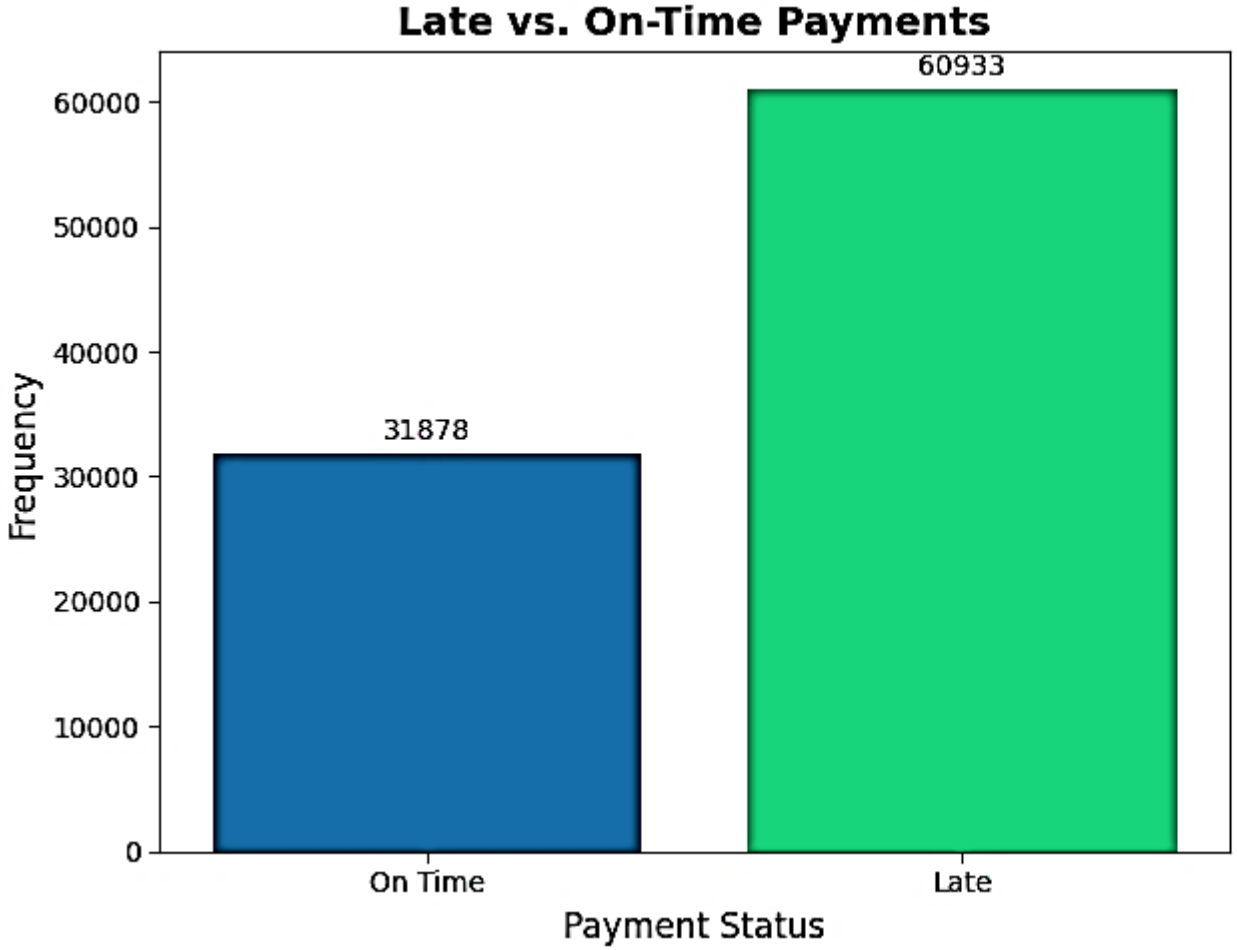
- *#Visualizing the distribution of RECEIPT_METHOD*

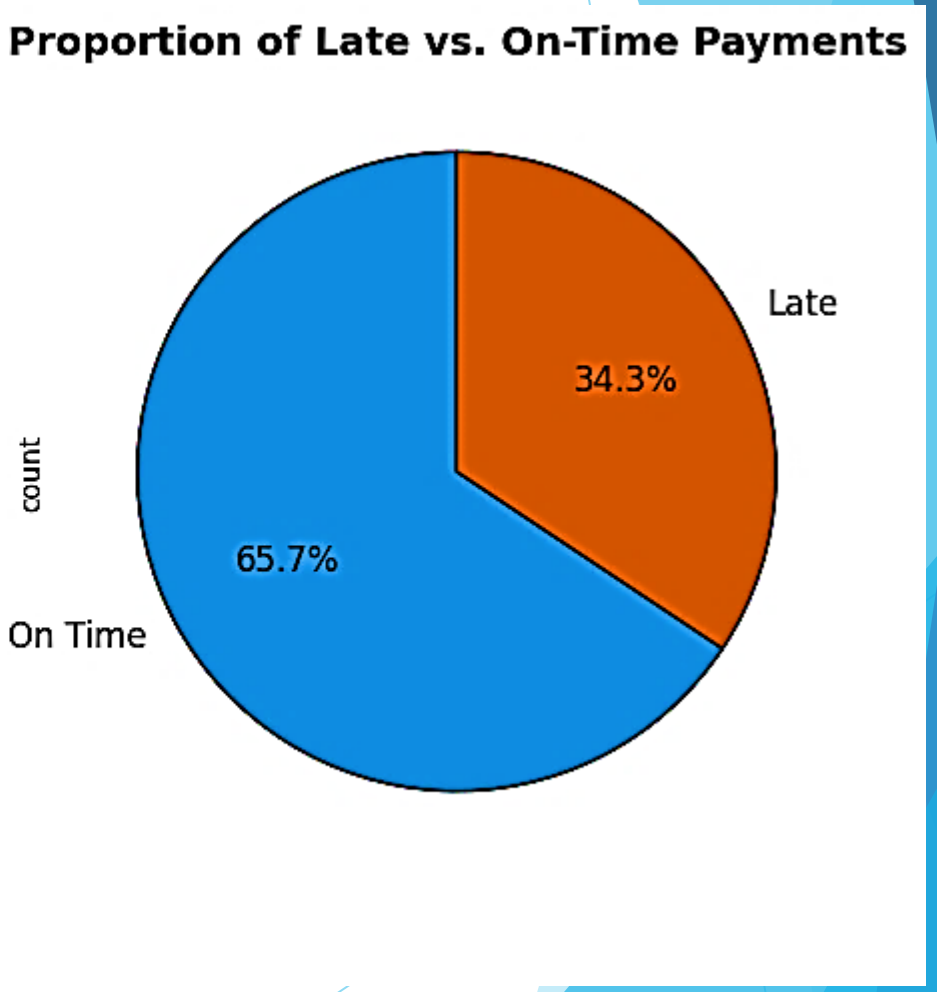


WIRE is the most preferred payment method for bill payments.

Checking for Data Imbalance

- #Visualizing the distribution of RECEIPT_METHOD

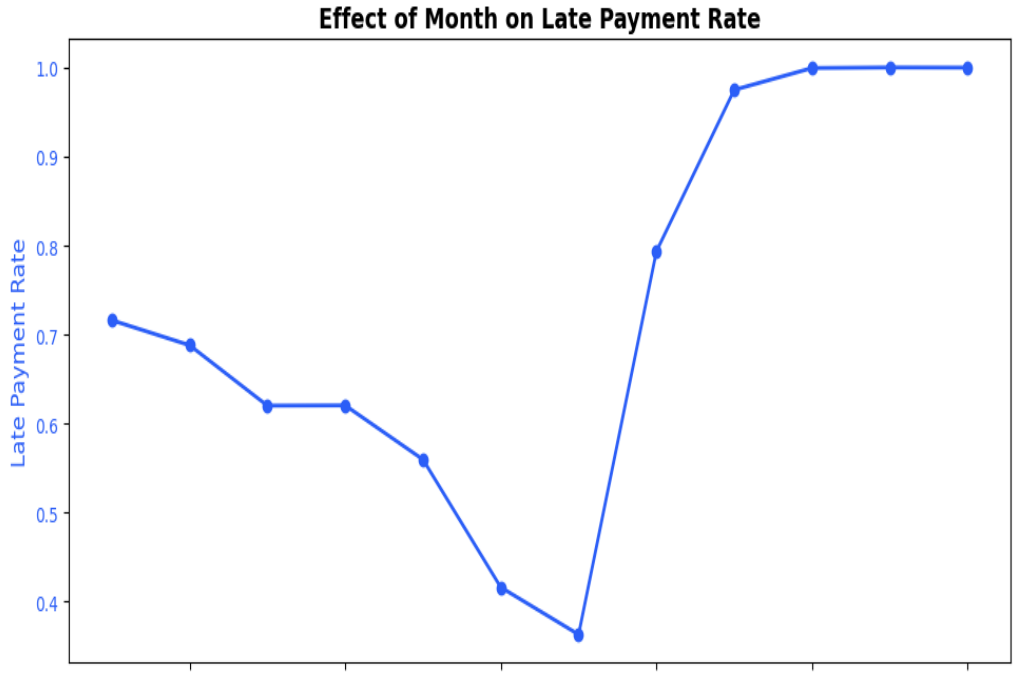


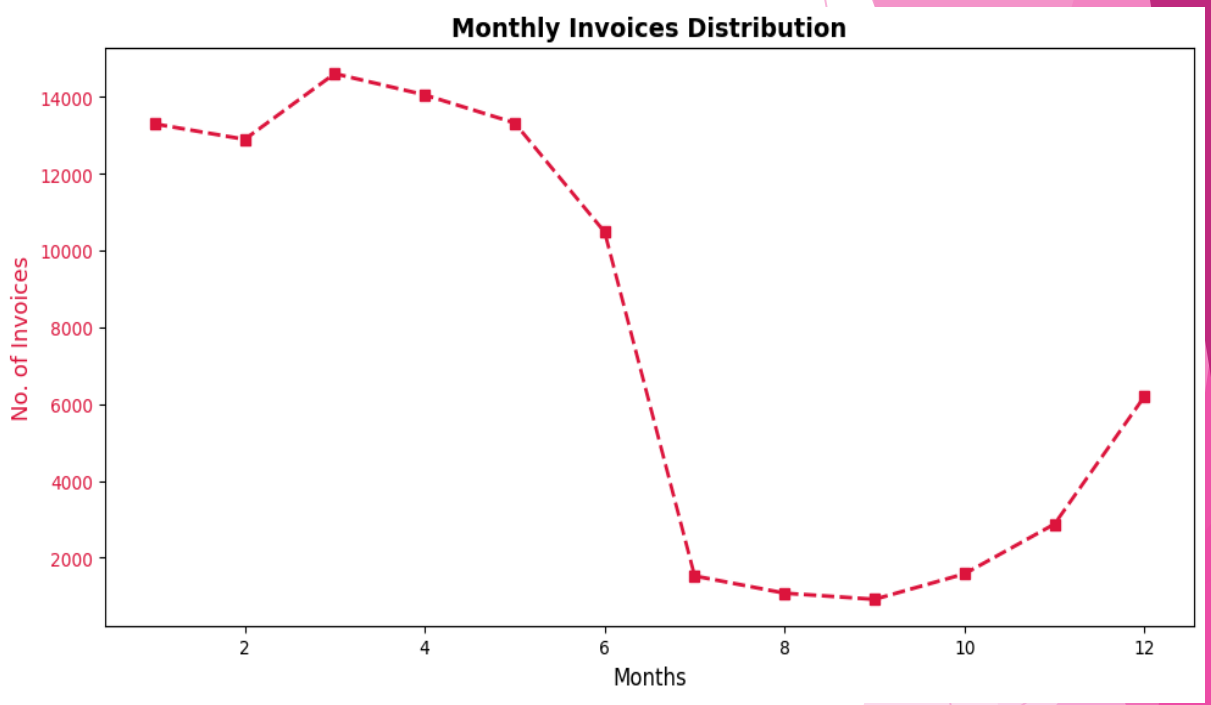


The target variable shows a balanced distribution.

BIVARIATE ANALYSIS

- #Visualizing the influence of month on late payment rate





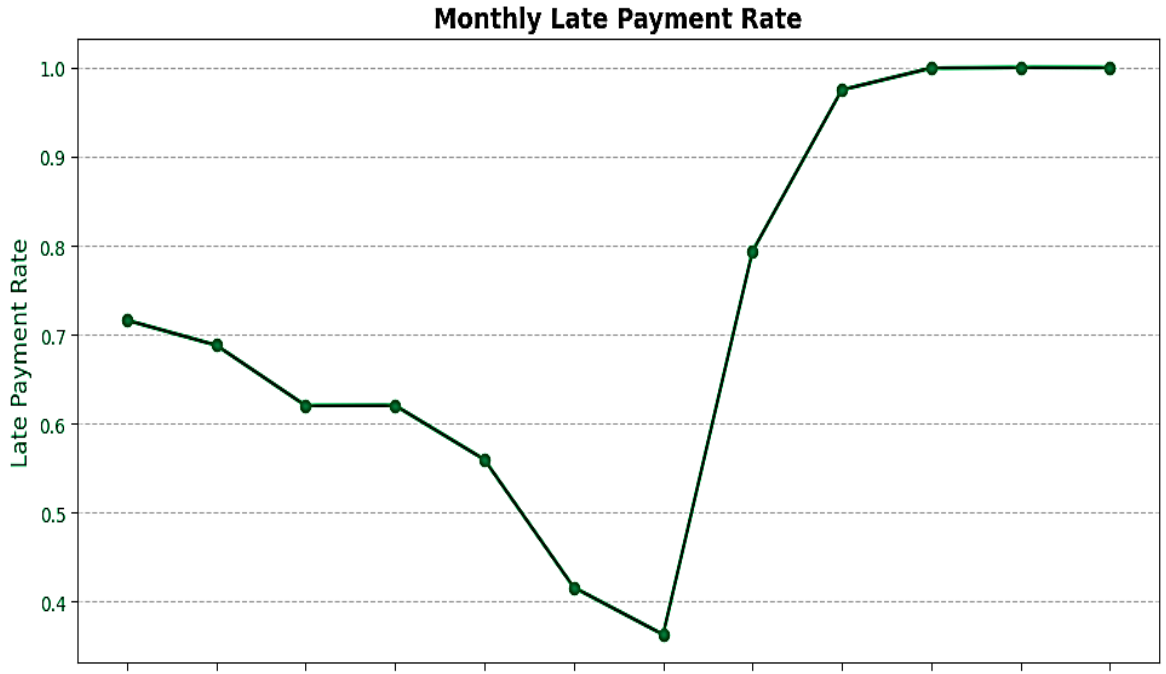
March records the highest number of invoices, yet the late payment rate remains relatively low.

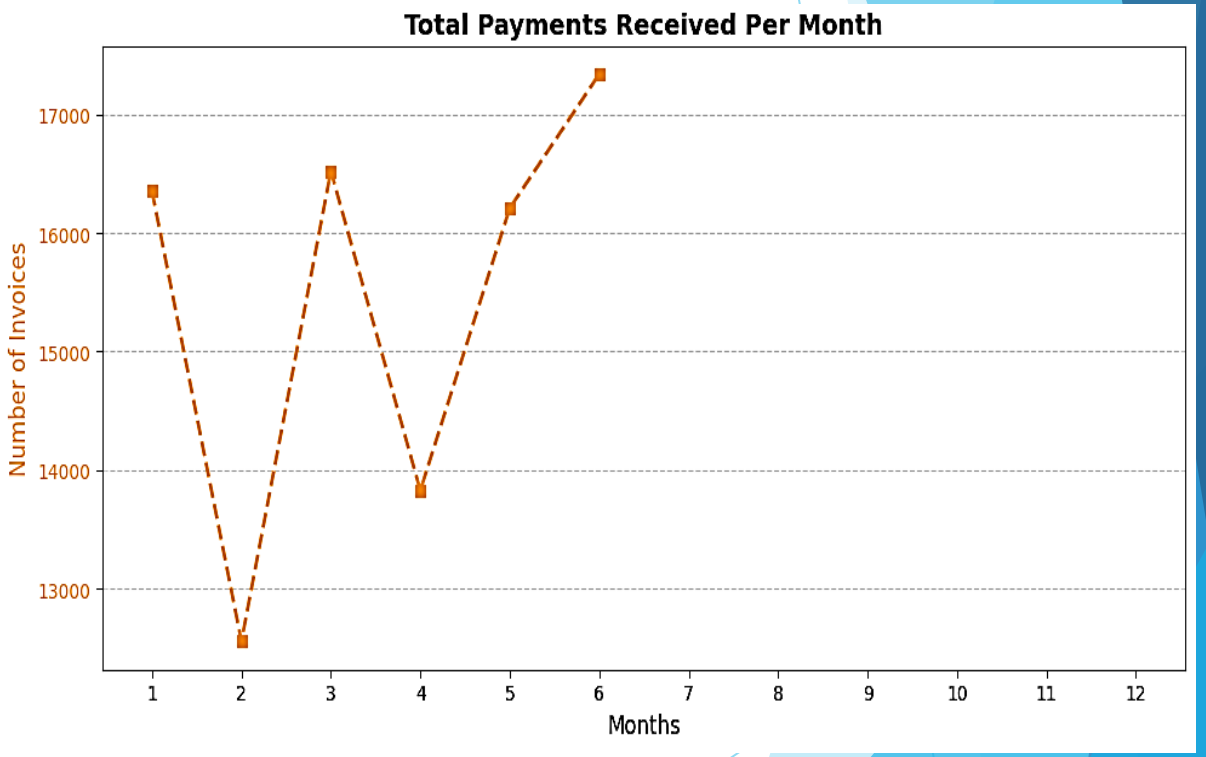
July has the lowest late payment rate, likely due to a significantly lower invoice volume.

In the latter half of the year, late payments increase sharply from July onward, despite a decline in invoice count compared to the first half.

BIVARIATE ANALYSIS

- #Visualizing the impact of due month on late payment rate

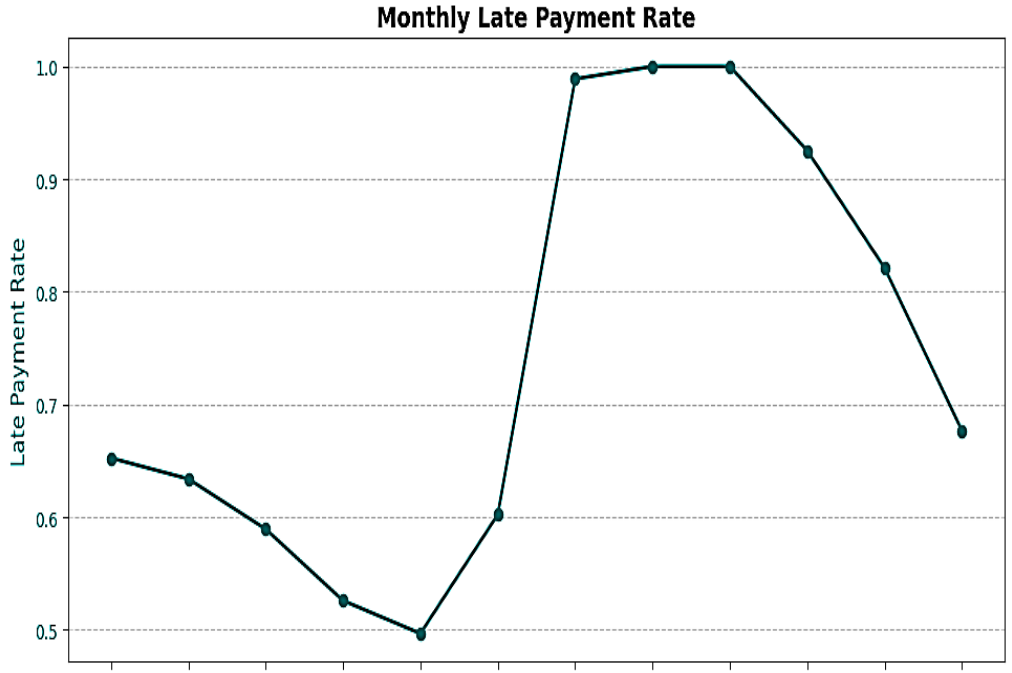


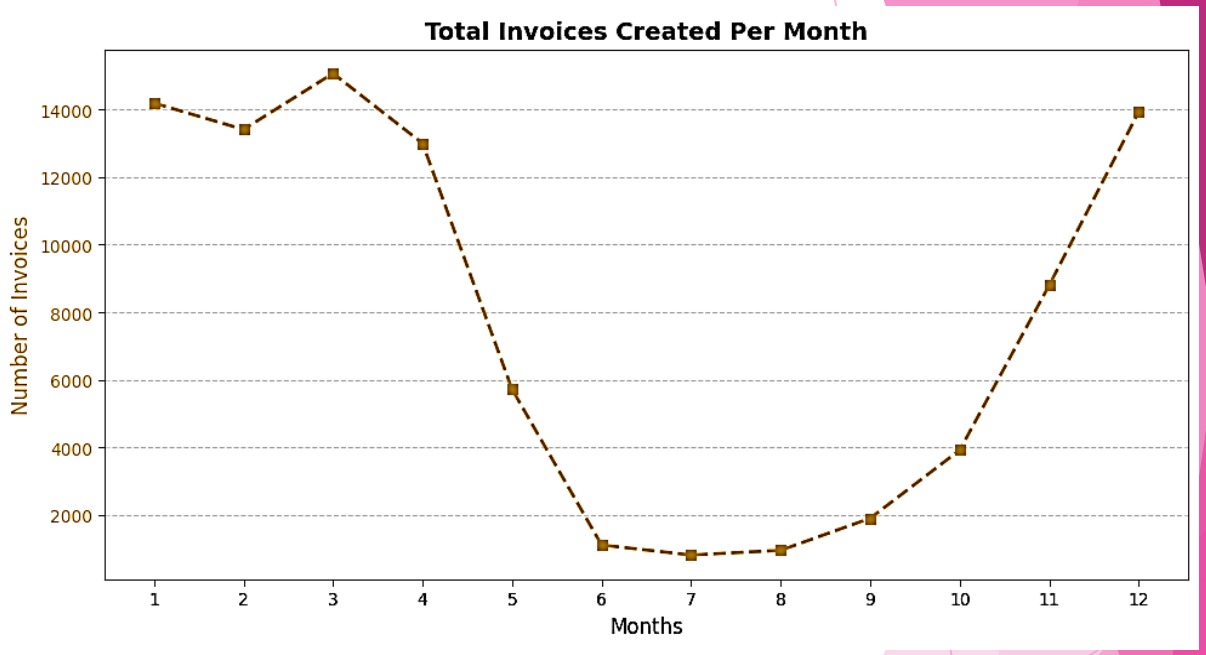


From the 7th month onward, no payments were processed for due invoices.

BIVARIATE ANALYSIS

- #Analyzing the impact of due months on late payment rate





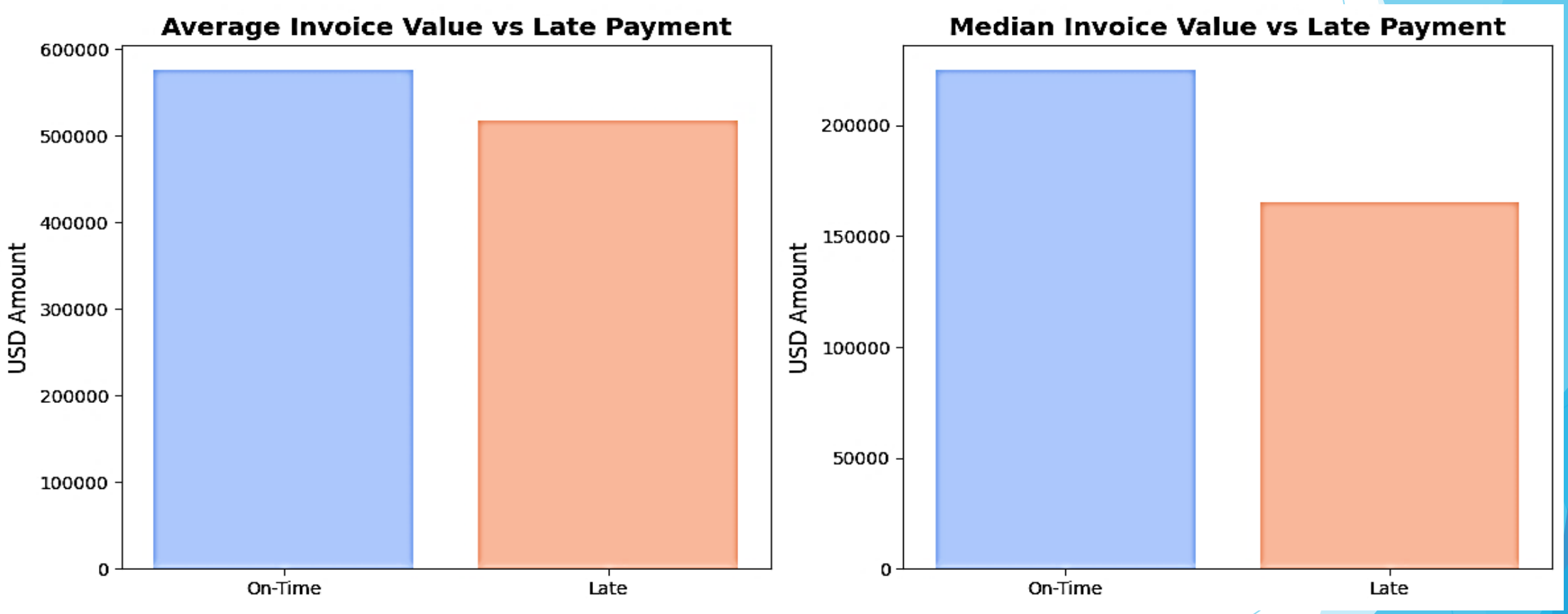
The late payment rate declines steadily from January (Month 1) to May (Month 5), indicating timely payments during this period.

A sharp increase in late payments is observed from July to September (Months 7, 8, and 9), suggesting a potential seasonal or financial trend affecting vendor payments.

Late payment rates remain consistently high beyond September, indicating that

BIVARIATE ANALYSIS

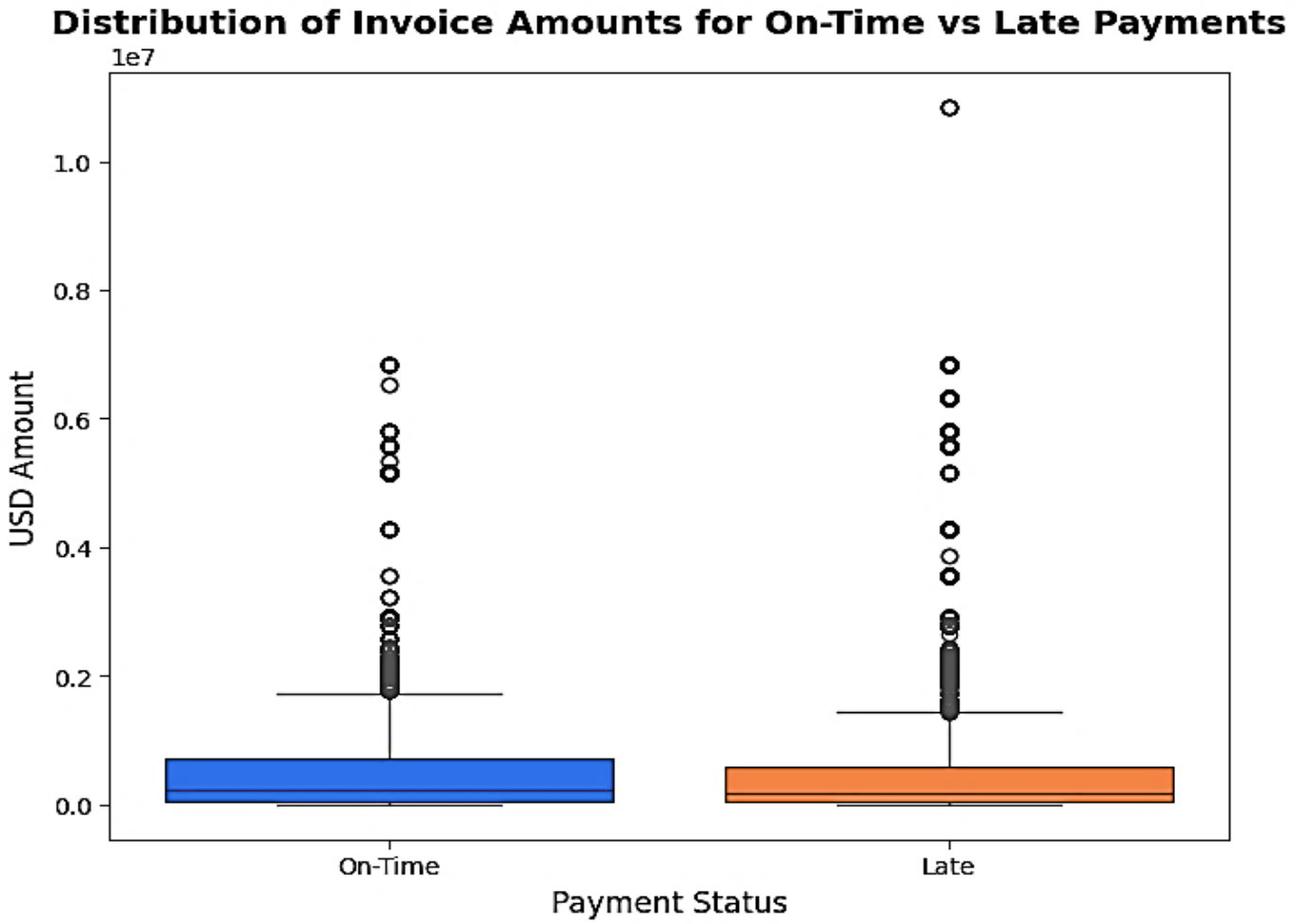
- *#Visualizing the difference between mean and median invoice value with respect to late payment*



On-time payments have a higher average and median invoice value compared t late payments, suggesting that larger transactions are more likely to be paid on

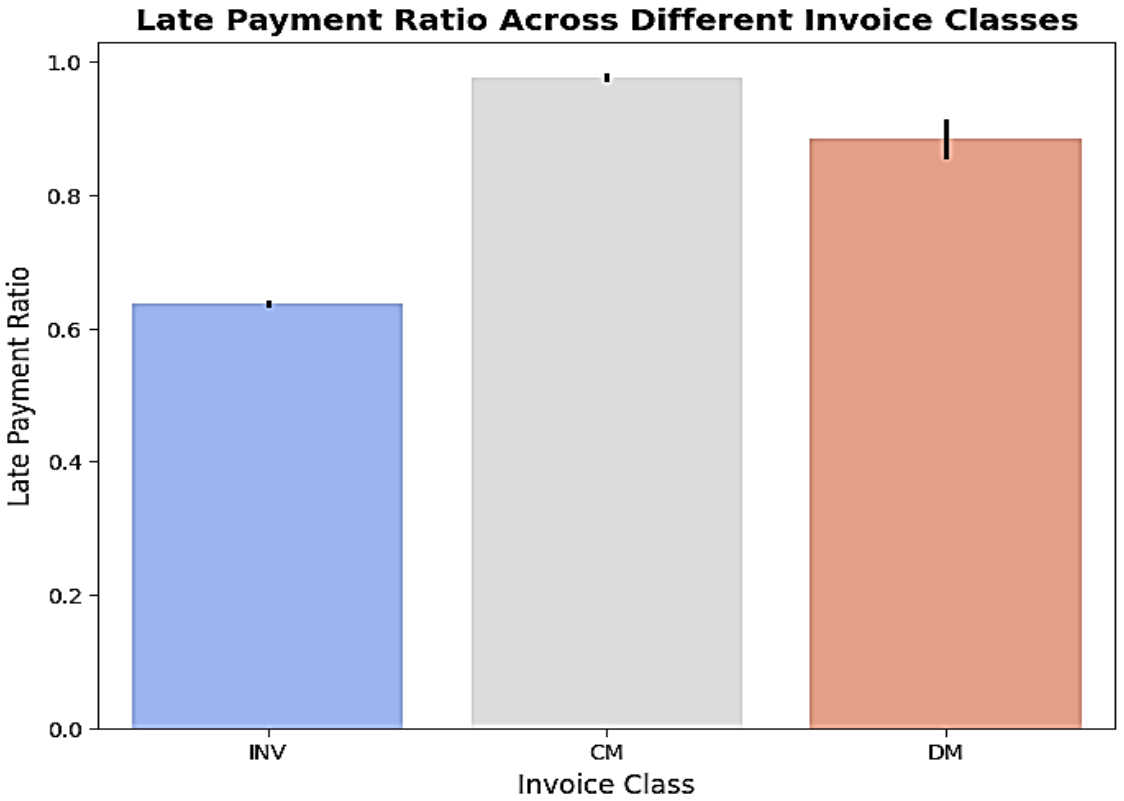
BIVARIATE ANALYSIS

- *#Visualizing USD Amount Comparison between On-Time and Late Payments*



BIVARIATE ANALYSIS

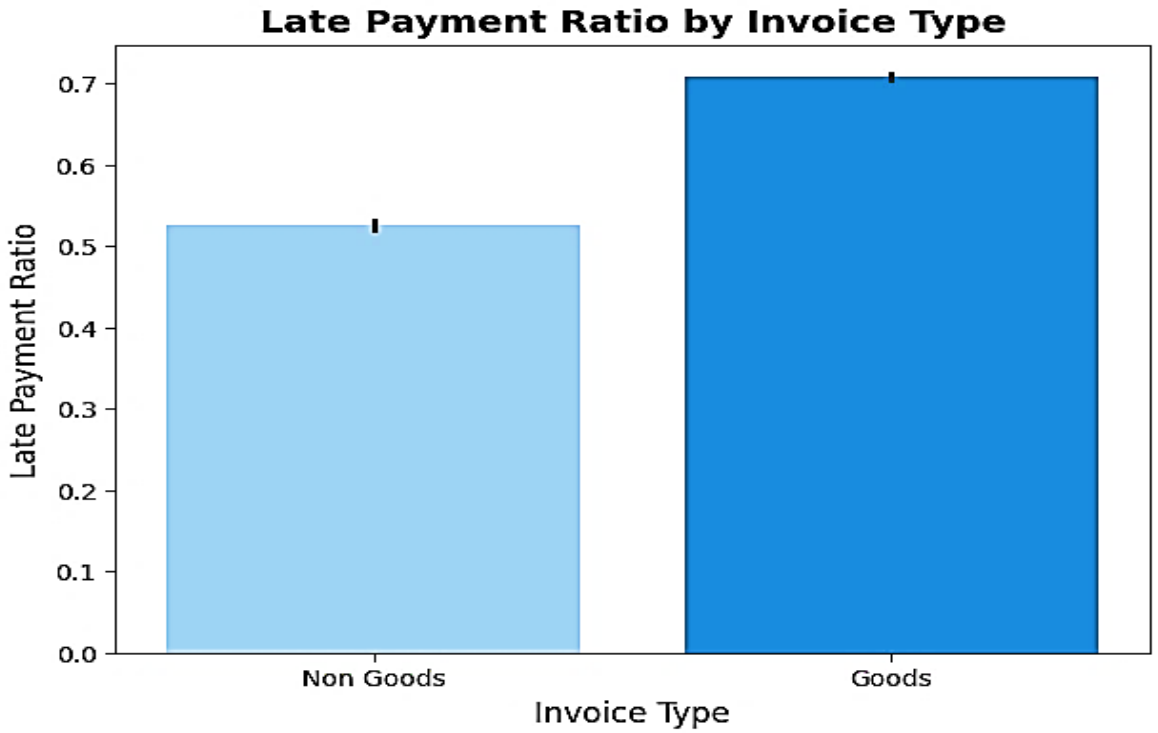
- `#Visualizing the influence of month on late payment rate`



The CM invoice class has the highest late payment ratio, indicating a higher tendency for delayed payments. The DM invoice class experiences frequent late payments, though slightly lower than CM. The INV invoice class has the lowest late payment ratio, suggesting that invoices under this category are more likely to be paid on time.

BIVARIATE ANALYSIS

- *#Analyzing the late payment ratio across different INVOICE_TYPE*

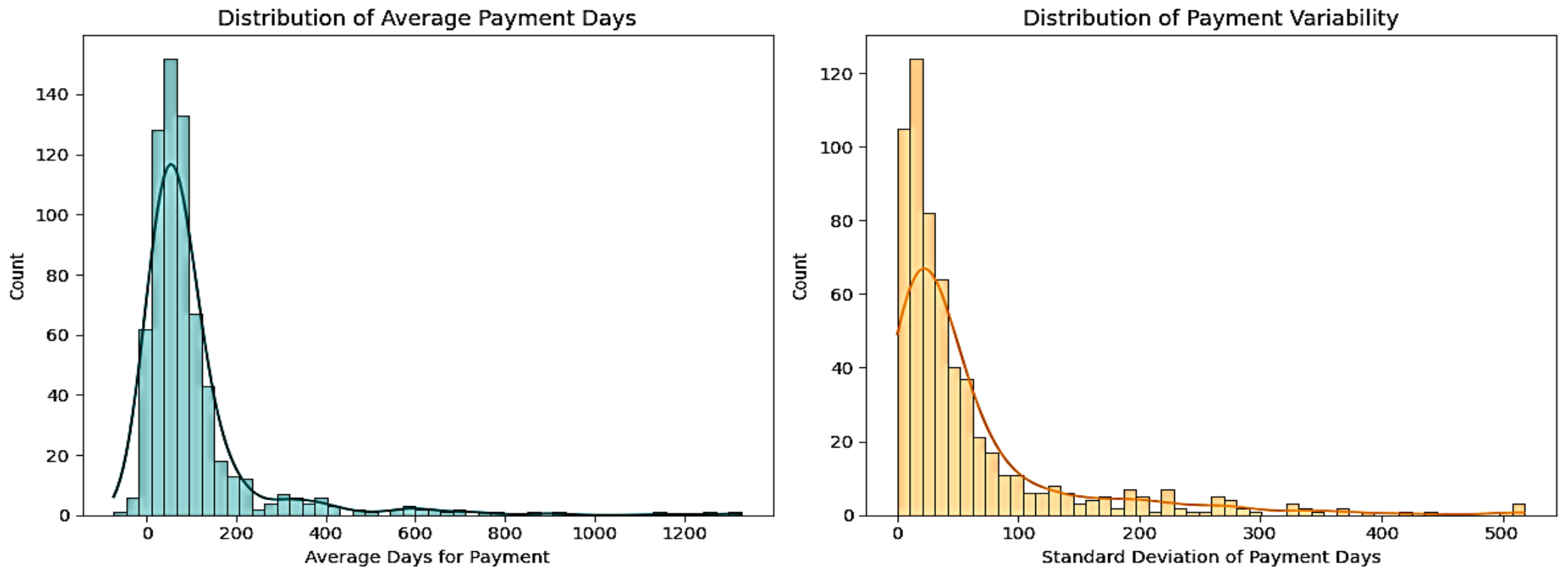


The Goods invoice type has a higher late payment ratio compared to Non-Goods.
The Non-Goods invoices show a lower likelihood of late payments, suggesting possible differences in credit policies or urgency of payment between the two types.
Goods-related invoices may require stricter follow-up strategies to improve timely payments.

CUSTOMER SEGMENTATION

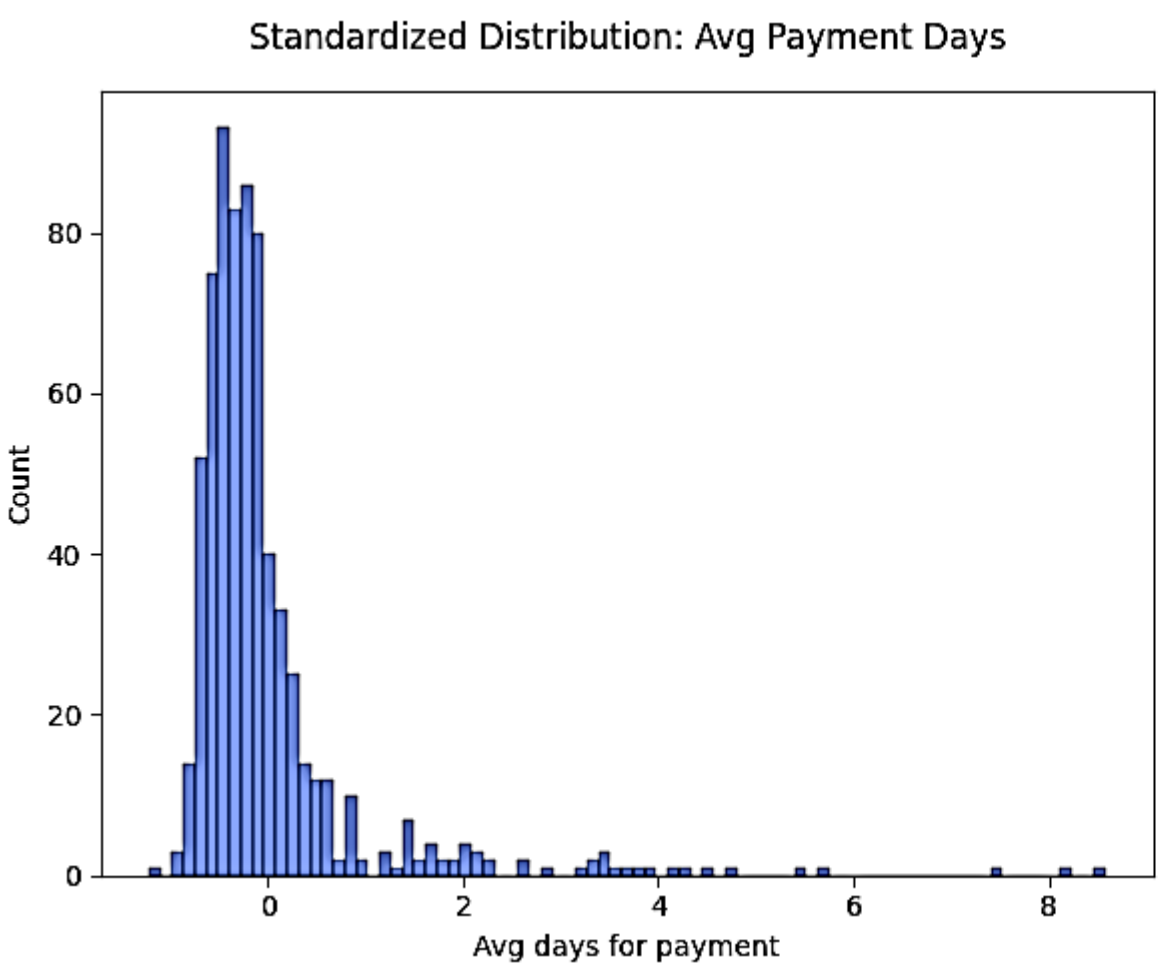
- *Incorporating customer-level attributes as independent variables can enhance the model's predictive power. Customer segmentation can be achieved by analyzing two key factors: the average payment delay (in days) and the variability (standard deviation) in payment behavior. By applying clustering techniques to these variables, we can identify distinct customer groups, which can then serve as valuable input features for the machine learning model.*

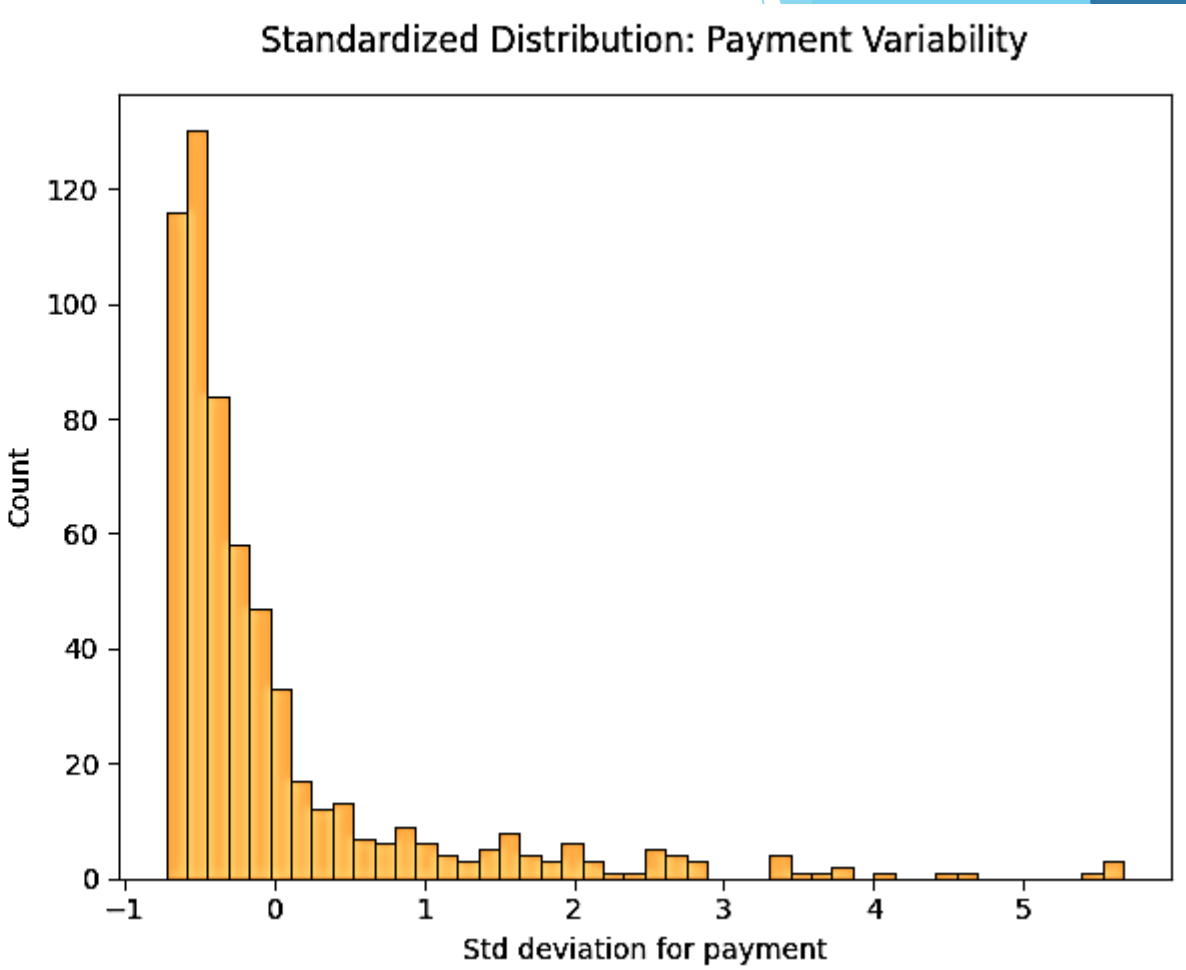
- *#Visualizing the distribution of customer payment behavior*



Both average payment days and standard deviation of payment days is highly right-skewed, indicating that while most customers tend to make payments within a shorter time frame, a few significantly delay their payments. The high variability in payment behavior suggests inconsistent payment patterns, which could be critical for risk analysis and customer segmentation.

- #Revisualizing the standardized data to check distribution

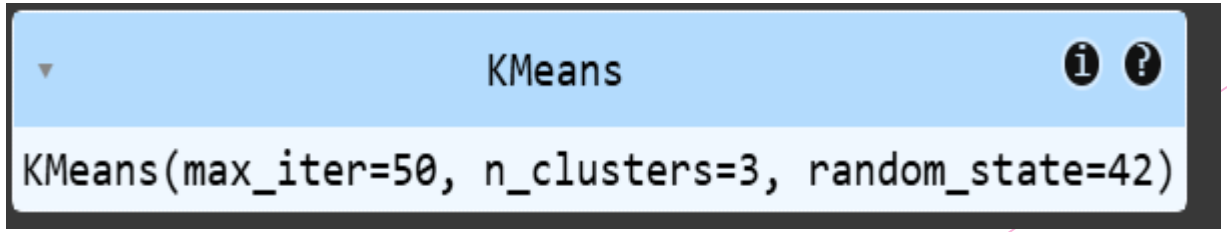




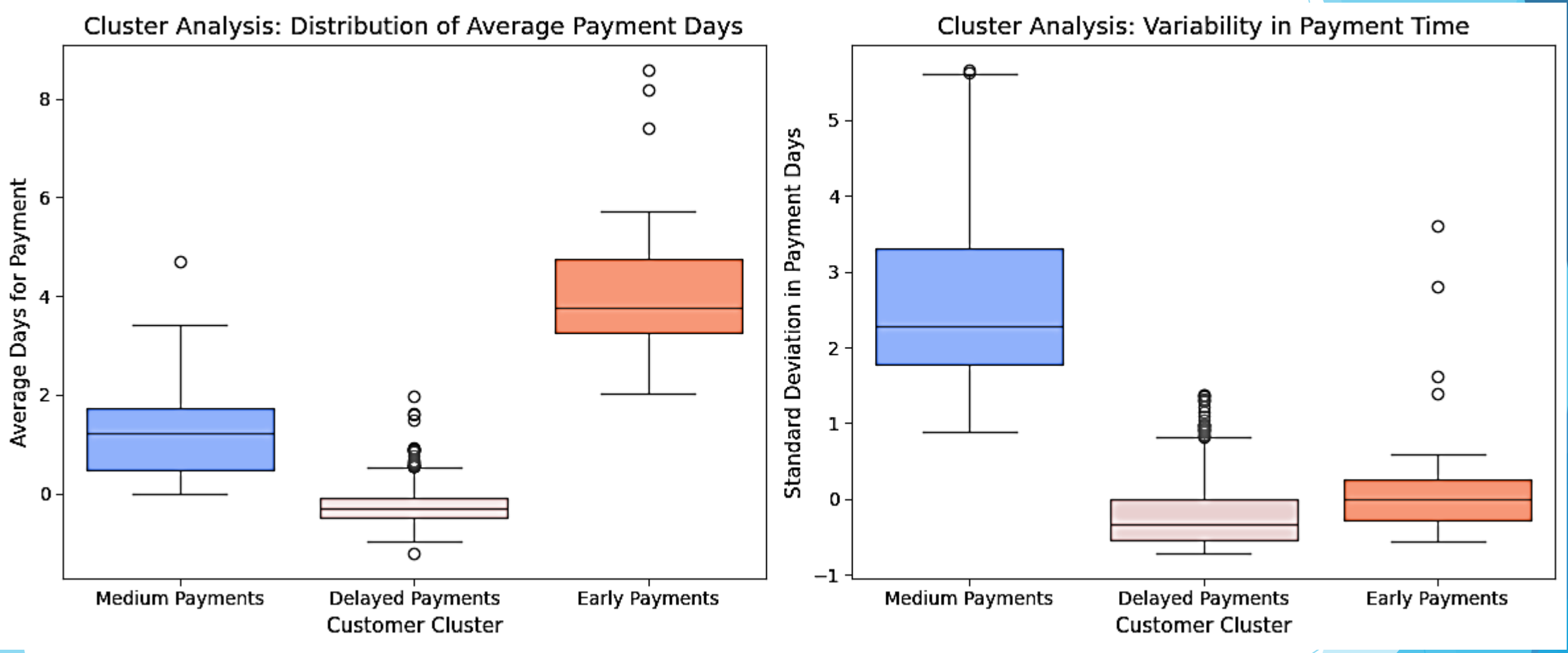
The data has been effectively scaled, ensuring all values fall within a normalized range. Most customers demonstrate lower average payment days with minimal variability, indicating a consistent payment behavior across transactions.

CLUSTERING MODEL - K-MEANS CLUSTERING

- *#Silhouette analysis for different cluster sizes to Identify the optimal number of clusters*
- *cluster_range = [2, 3, 4, 5, 6, 7, 8]*
- *For n_clusters=2, the silhouette score is 0.7557759850933141*
- *For n_clusters=3, the silhouette score is 0.7491797445652462*
- *For n_clusters=4, the silhouette score is 0.6097388985555463*
- *For n_clusters=5, the silhouette score is 0.6173540681032771*
- *For n_clusters=6, the silhouette score is 0.3980238443004184*
- *For n_clusters=7, the silhouette score is 0.4012628375918799*
- *For n_clusters=8, the silhouette score is 0.4145784973897661 5*
- *The silhouette analysis indicates that 3 clusters provide a well-balanced segmentation, ensuring distinct group separation while maintaining cluster cohesion. Selecting K=3 is optimal number of cluster for K-Means clustering algorithm.*



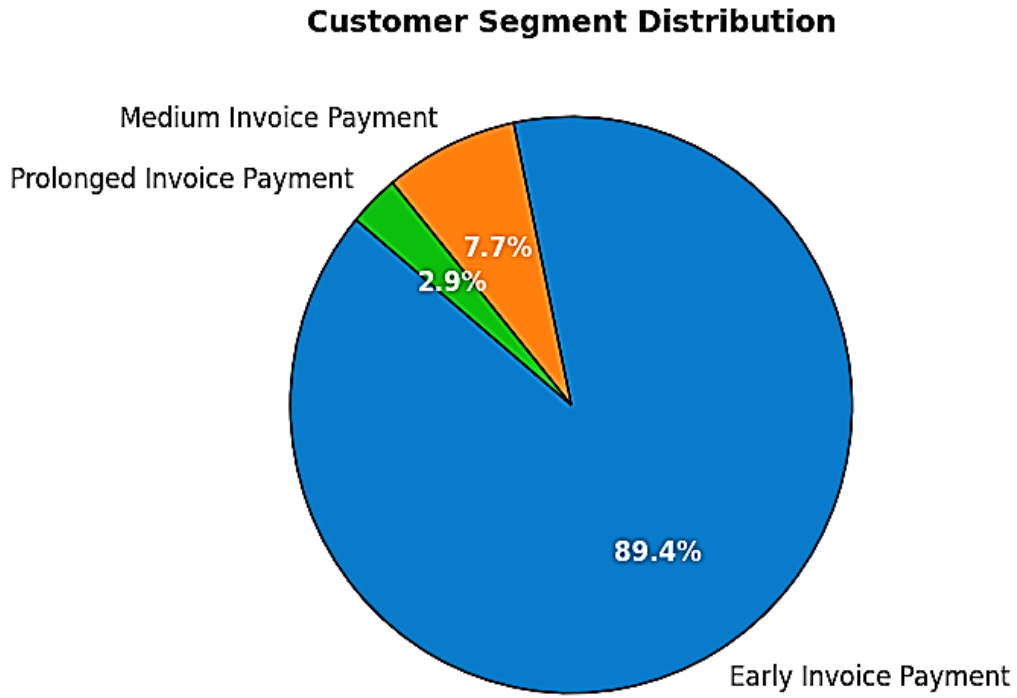
- *#Plotting box plot based on clusters*



Cluster 0 (Medium Payments): Customers have a balanced payment time with som variability.

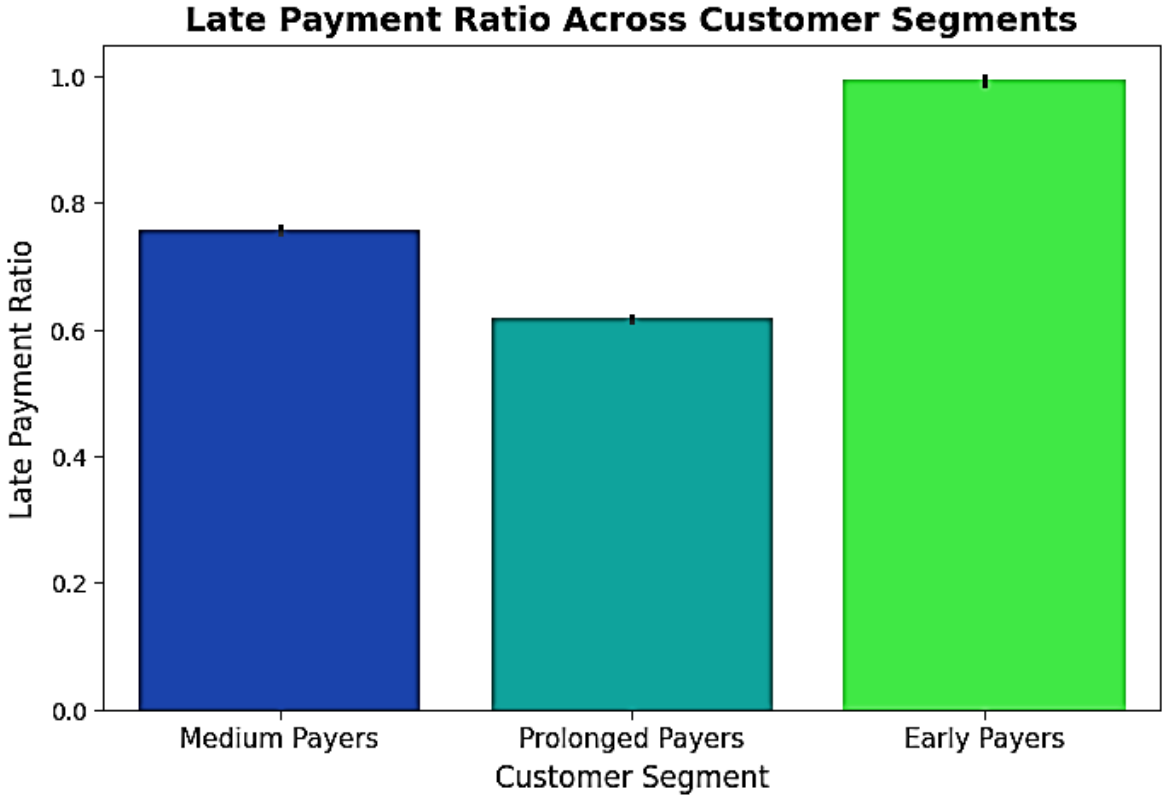
Cluster 1 (Prolonged Payments): Customers take the longest to pay with moderate variability.

- *#Visualizing customer segment distribution*



- *88.4% of customers fall into the Early Invoice Payment category, indicating a strong tendency toward timely payments.*
- *7.7% of customers belong to the Medium Invoice Payment segment, meaning they show moderate payment delays.*
- *Only 2.9% of customers exhibit Prolonged Invoice Payment behavior, highlighting a very small portion that requires intervention. The data suggests that most*

- `# Visualizing the late payment ratio across customer clusters`



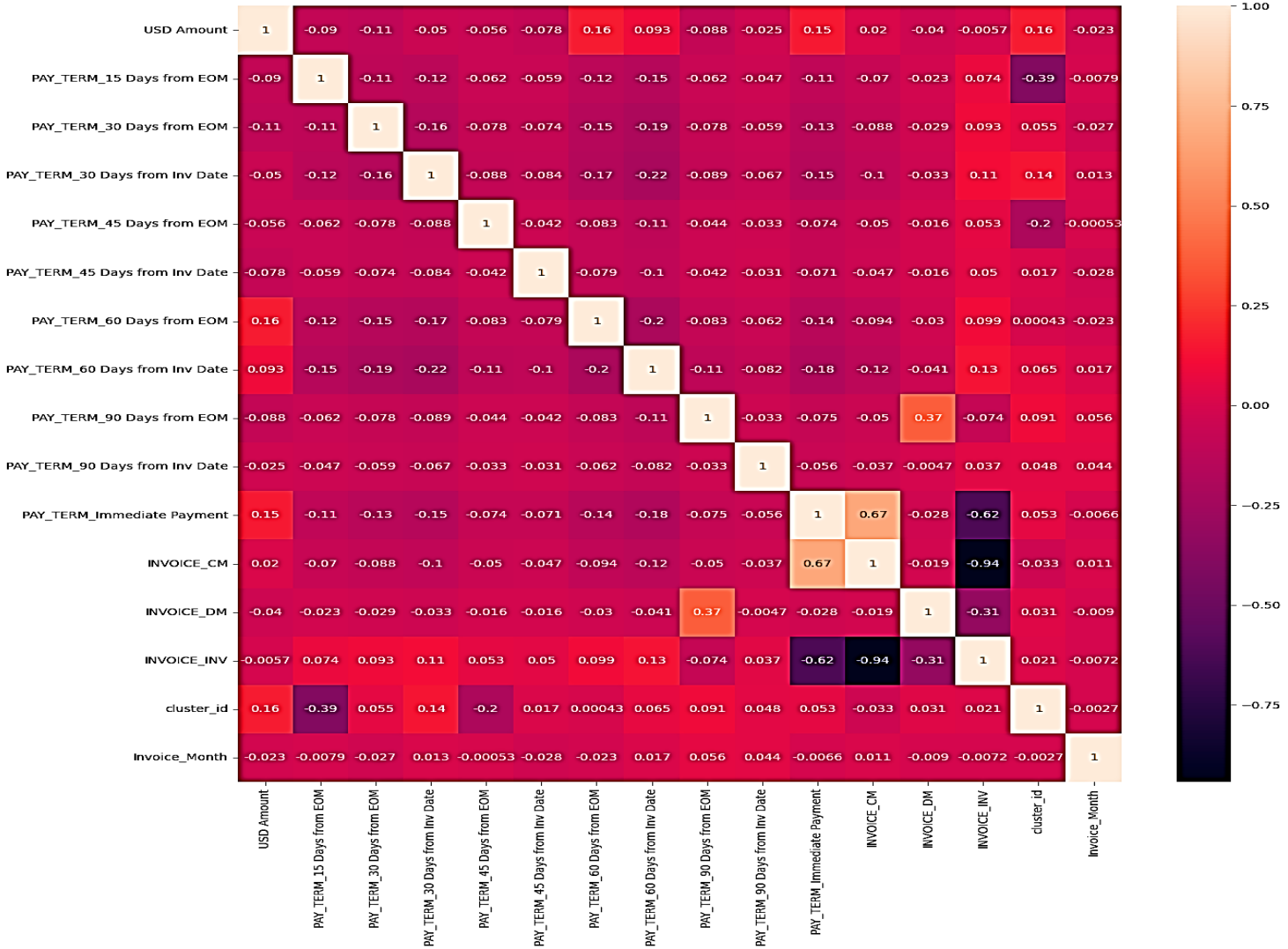
- ***Cluster 1 has the highest late payment ratio, indicating frequent delays.***
- ***Cluster 0 shows a moderate late payment ratio, with mixed payment behavior.***
- ***Cluster 2 has the lowest late payment ratio, representing the timeliest payers.***

LOGISTIC REGRESSION MODEL

- *#Splitting Data into Training and Testing Sets:*
  - *from sklearn.model_selection import train_test_split*
  - *X_train,X_test,y_train,y_test =train_test_split(X,y,train_size=0.7,test_size=0.3,random_state=42)*
  - *X_train.head()*

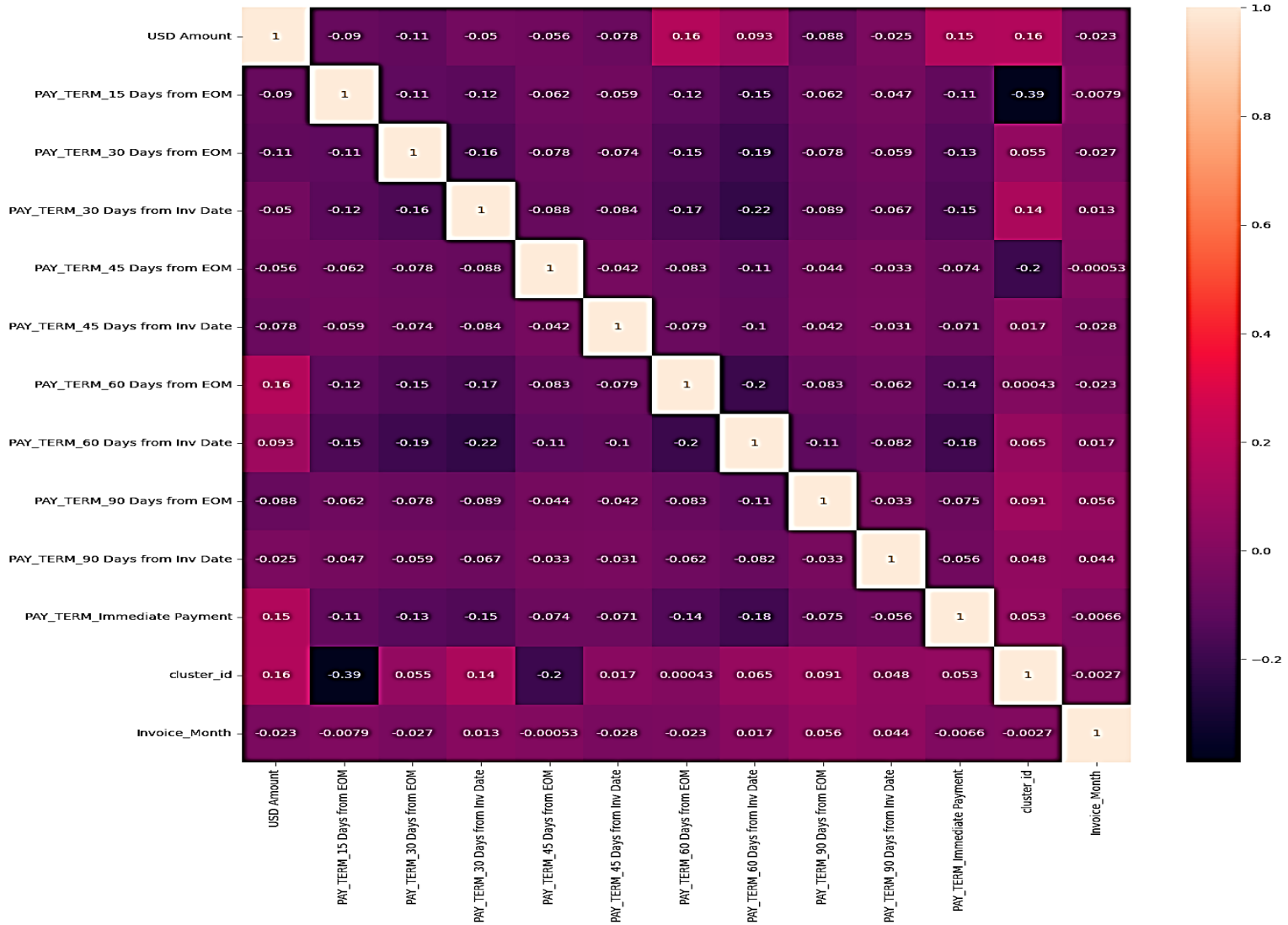
- *#Feature Scaling on Numerical columns:*
  - *scaler = StandardScaler()*
  - *X_train['USD Amount'] = scaler.fit_transform(X_train[['USD Amount']])*

Heatmap for X_train



Dropping Highly Correlated Features: CM & INV, INV & Immediate Payment, DM & 90 Days from EOM

#Replotting Heatmap for X_train to Verify Multicollinearity Reduction



#Checking the P-value and VIF

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:	-30146.
Date:	Tue, 04 Feb 2025	Deviance:	60292.
Time:	06:18:51	Pearson chi2:	6.29e+04
No. Iterations:	7	Pseudo R-squ. (CS):	0.3018
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
const	0.9566	0.052	18.559	0.000	0.856	1.058
USD Amount	-0.0342	0.012	-2.858	0.004	-0.058	-0.011
PAY_TERM_15 Days from EOM	2.4735	0.107	23.119	0.000	2.264	2.683
PAY_TERM_30 Days from EOM	-2.3400	0.053	-44.426	0.000	-2.443	-2.237
PAY_TERM_30 Days from Inv Date	0.2516	0.052	4.877	0.000	0.150	0.353
PAY_TERM_45 Days from EOM	0.3050	0.070	4.382	0.000	0.169	0.441
PAY_TERM_45 Days from Inv Date	-0.3133	0.063	-4.991	0.000	-0.436	-0.190
PAY_TERM_60 Days from EOM	-2.1698	0.052	-41.409	0.000	-2.272	-2.067
PAY_TERM_60 Days from Inv Date	-0.1936	0.049	-3.915	0.000	-0.291	-0.097
PAY_TERM_90 Days from EOM	-0.5079	0.061	-8.326	0.000	-0.627	-0.388
PAY_TERM_90 Days from Inv Date	-1.0336	0.069	-15.044	0.000	-1.168	-0.899
PAY_TERM_Immediate Payment	3.0961	0.104	29.705	0.000	2.892	3.300
cluster_id	-0.4368	0.026	-16.931	0.000	-0.487	-0.386
Invoice_Month	0.0952	0.003	37.574	0.000	0.090	0.100

- *#Checking the P-value and VIF*

	Features	VIF
11	cluster_id	4.08
12	Invoice_Month	2.66
7	PAY_TERM_60 Days from Inv Date	2.01
3	PAY_TERM_30 Days from Inv Date	1.85
2	PAY_TERM_30 Days from EOM	1.58
6	PAY_TERM_60 Days from EOM	1.57
10	PAY_TERM_Immediate Payment	1.54
8	PAY_TERM_90 Days from EOM	1.29
5	PAY_TERM_45 Days from Inv Date	1.17
9	PAY_TERM_90 Days from Inv Date	1.15
1	PAY_TERM_15 Days from EOM	1.14
0	USD Amount	1.11
4	PAY_TERM_45 Days from EOM	1.09

Confusion Matrix and Acurracy score of the model

```
#Confusion Matrix
```

```
array([[12815,  9534],
       [ 4453, 38165]])
```

The model achieves an accuracy of 78.3% on the training dataset

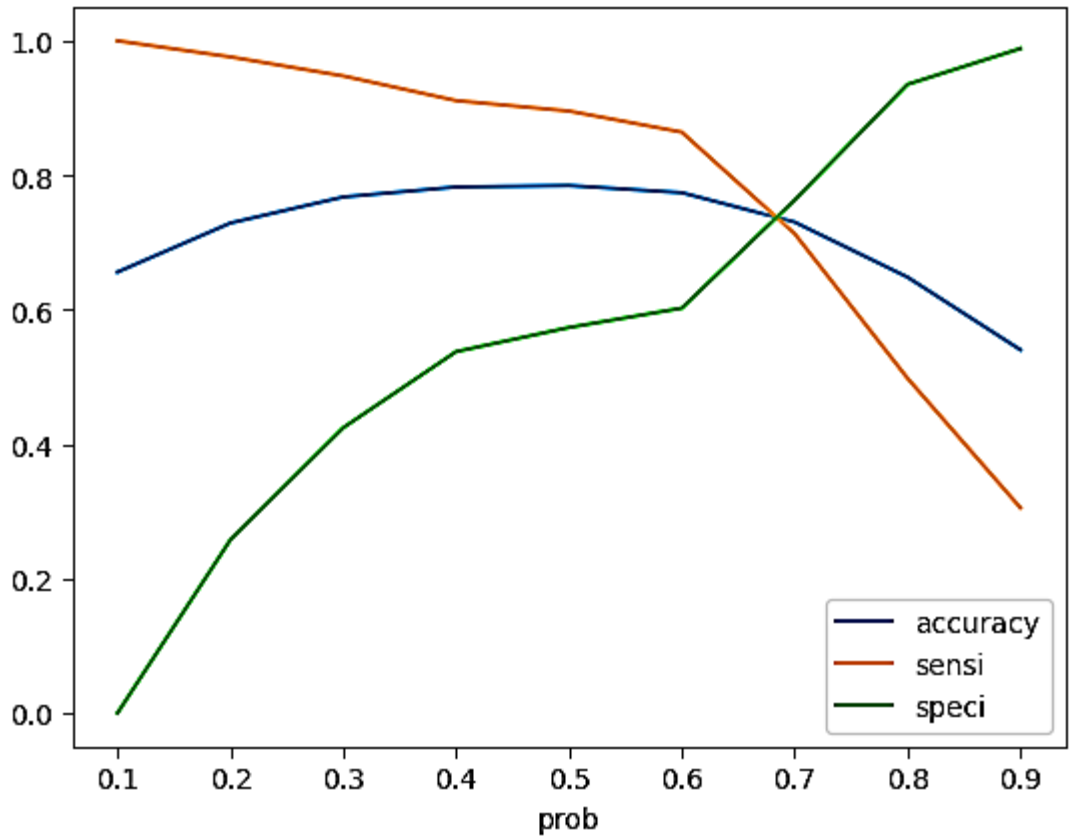
```
#True Positives, True Negatives, False Positives, and False Negatives
```

True Positive = confusion[1,1] = 38165
True Negative = confusion[0,0] = 12815
False Positive = confusion[0,1] = 9534
False Negatiove = confusion[1,0] = 4453

 # Sensitivity(Recall), Specificity, False Positive Rate(FPR), Positive Predictive Value(Precision), and Negative Predictive Value(NPV) of Training Dataset

- ***SENSITIVITY – 89.55%***
- ***SPECIFICITY - 57.34%***
- ***FALSE POSITIVE RATE (FPR) – 42.65%***
- ***POSITIVE PREDICTIVE VALUE (PRECISION) – 80%***
- ***NEGATIVE PREDICTIVE VALUE (NPV) – 74.21%***

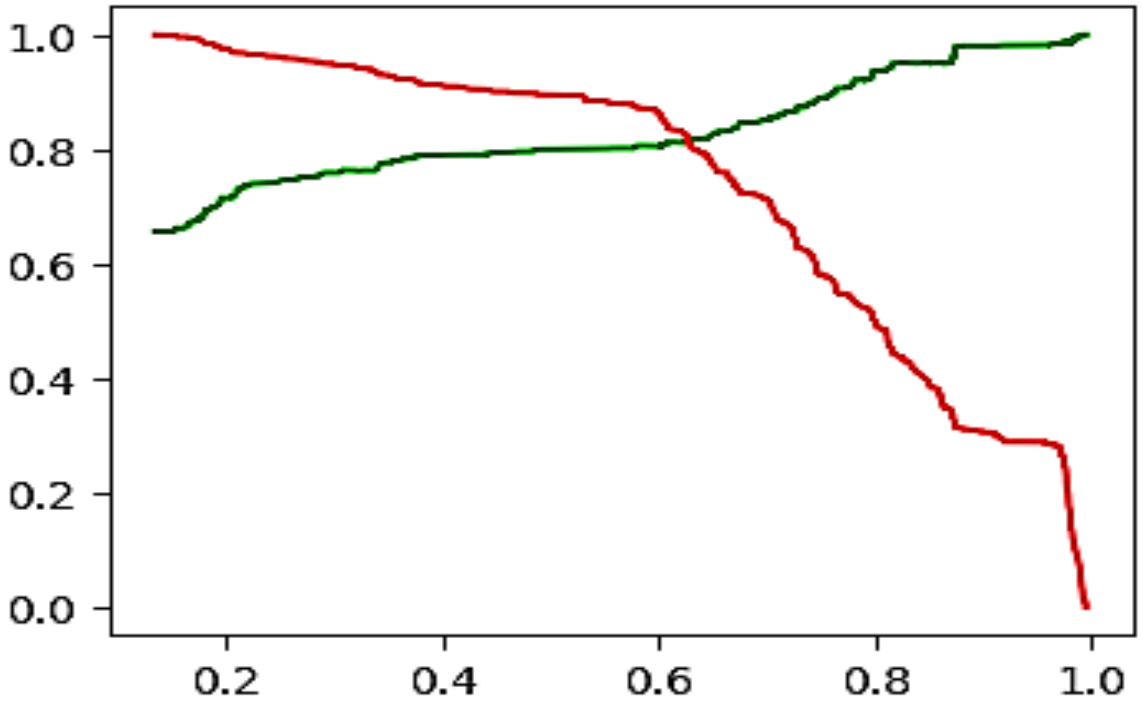
```
#Visualizing accuracy, sensitivity, and specificity
across probability thresholds
```



Based on the precision-recall curve, a threshold of 0.67 is identified as the optimal cutoff probability for classification.

 # Sensitivity(Recall), Accuracy, Positive Predictive Value(Precision) on Train Data

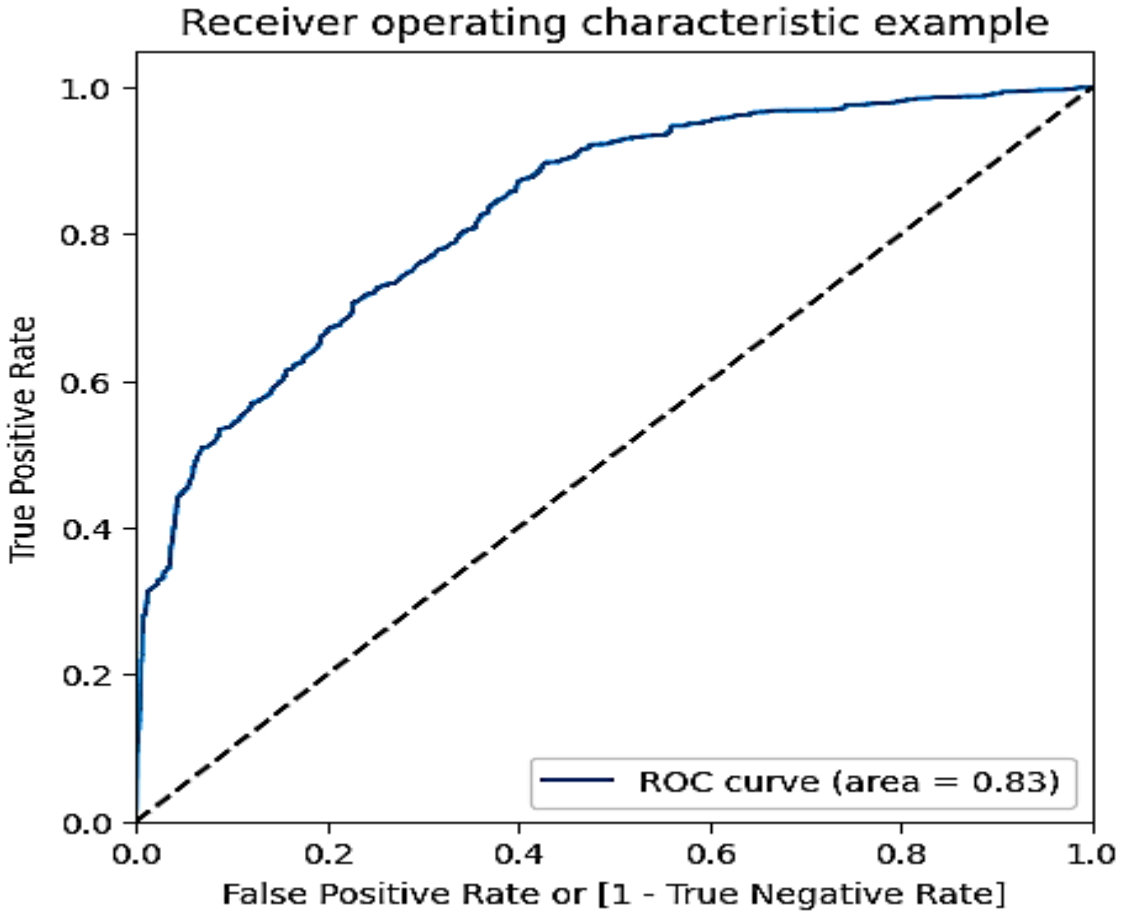
- ***RECALL – 86.41%***
- ***ACCURACY–77.4%***
- ***PRECISION – 80.55%***



Based on the Precision-Recall trade-off, the optimal cutoff was identified between 0.6 and 0.7. Therefore, 0.6 is selected as the final threshold for predictions.

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

#ROC CURVE



AUC = 0.83, indicating a strong model performance in distinguishing between classes.

asses.

•Model performs consistently with train and test accuracy both around 77.5%

```
# Sensitivity(Recall), Accuracy, Positive Predictive
Value(Precision) on Test Data
```

- ***RECALL – 86.41%***
- ***ACCURACY–77.58%***
- ***PRECISION – 80.77%***

Model performs consistently with train and test accuracy both around 77.5%

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

 # Accuracy and classification report on Train Data

	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

Accuracy is : 0.9580864131020365

 # Mean score and Standard Deviation on Train Data

```
Mean score: 0.9520624266036315
Standard deviation: 0.0016240232566768651
```

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

Accuracy and classification report on Test Data

	precision	recall	f1-score	support
0	0.91	0.86	0.88	9529
1	0.93	0.96	0.94	18315
accuracy			0.92	27844
macro avg	0.92	0.91	0.91	27844
weighted avg	0.92	0.92	0.92	27844

Accuracy is : 0.9218503088636669

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

 # Grid Search for hyperparameter tuning to optimize Random Forest model

```
Best hyperparameters: {'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100}
Best f1 score: 0.9393552098689479
              precision    recall  f1-score   support

           0       0.96      0.91      0.94     22349
           1       0.95      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
```

The classification report indicates that the F1-score for training (0.96) and test (0.92) is high, confirming a well-performing model. Proceeding with this as the final model for predictions.

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

- # ANALYZING FEATURE IMPORTANCE IN THE RANDOM FOREST MODel

```
Feature ranking:
1. USD Amount (0.490)
2. Invoice_Month (0.130)
3. PAY_TERM_30 Days from EOM (0.111)
4. PAY_TERM_60 Days from EOM (0.109)
5. PAY_TERM_Immediate Payment (0.044)
6. PAY_TERM_15 Days from EOM (0.028)
7. cluster_id (0.027)
8. PAY_TERM_60 Days from Inv Date (0.014)
9. PAY_TERM_30 Days from Inv Date (0.012)
10. PAY_TERM_90 Days from Inv Date (0.007)
11. PAY_TERM_90 Days from EOM (0.006)
12. INVOICE_INV (0.005)
13. INVOICE_CM (0.005)
14. PAY_TERM_45 Days from EOM (0.005)
15. PAY_TERM_45 Days from Inv Date (0.004)
16. INVOICE_DM (0.001)
```

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

- ***CONCLUSIONS AND RECOOMENDATIONS***
- ***Based on the clustering analysis, several key insights were derived, leading to the following recommendations for improving payment collection efficiency:***
- ***Stricter Policies for Credit Note Payments Among different invoice types, Credit Note Payments exhibit the highest delay rates compared to Debit Notes and standard Invoices. To mitigate this issue, the company should consider implementing stricter payment collection policies specifically for credit note transactions.***
- ***Tighter Payment Terms for Goods-Related Invoices Invoices related to goods transactions experience significantly higher payment delays compared to non-goods invoices. Introducing more stringent payment terms for these transactions may help reduce late payments.***

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

- ***CONCLUSIONS AND RECOOMENDATIONS***
- ***Addressing Delays in Lower-Value Transactions The company could explore penalty-based strategies, where the penalty percentage increases for smaller bill amounts. However, this should be a last resort to ensure customer relationships are maintained.***
- ***Focused Attention on Prolonged Payment Customers Customers were segmented into three clusters based on payment behavior:***
- ***Cluster 0: Medium payment duration***
- ***Cluster 1: Prolonged payment duration***
- ***Cluster 2: Early payment duration***
- ***Customers in Cluster 1 (prolonged payment duration) exhibited the highest delay rates. Therefore, priority should be given to monitoring and engaging with these customers to encourage timely payments.***

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

RANDOM FOREST MODEL

- ***CONCLUSIONS AND RECOOMENDATIONS***
- ***Prioritizing High-Risk Accounts Companies with the highest probability of delayed payments and the greatest number of overdue transactions should be given top priority in collection efforts. Proactive follow-ups and tailored strategies for these accounts can help minimize financial risks.***
- ***By implementing these targeted strategies, the company can enhance its payment collection efficiency and reduce overall delays in receivables.***

•AUC = 0.83, indicating a strong model performance in distinguishing between classes.

Thank you !