E-COMMERCE & RETAIL B2B CASE STUDY

Agenda

- Problem Statement
- Objective
- Reading and Understanding the data
- Data Cleaning
- Exploratory Data Analysis
- Clustering
- Data Preparation
- Model Building
- Model Evaluation
- Conclusion

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. The company has some employees who keep chasing vendors to get the payment on time; this procedure nevertheless also results in non-value-added activities, loss of time and financial impact. Schuster would thus try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

Problem Statement

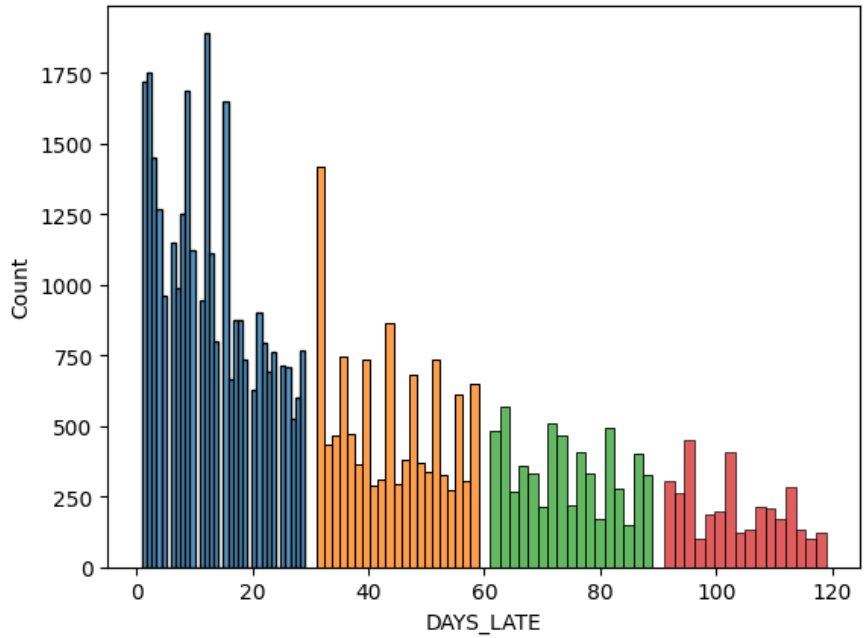
- Schuster would like to better understand the customers' payment behavior based on their past payment patterns(customer segmentation).

- Using historical information, it wants to be able to predict the likelihood of delayed payment against open invoices from its customers.

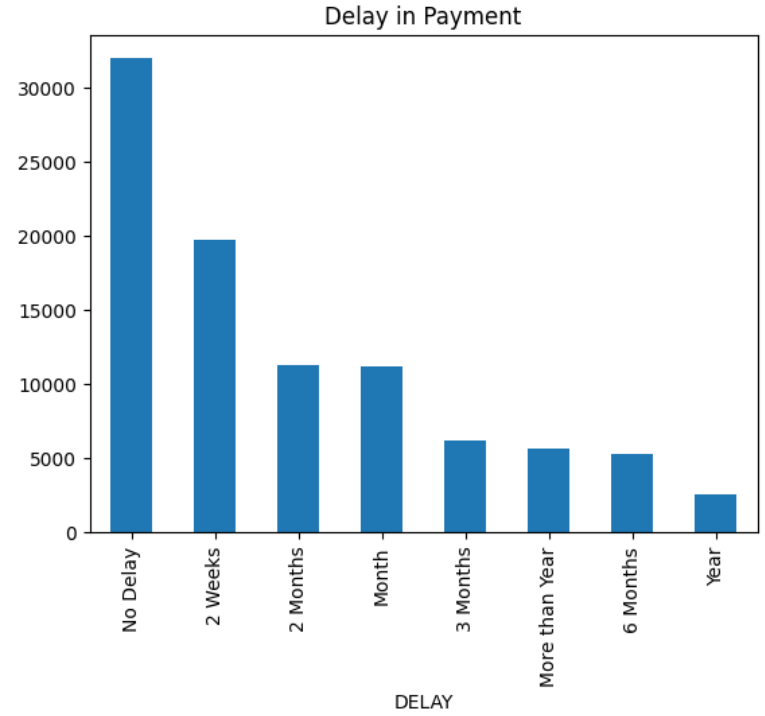
- It wants to use this information so that collectors can prioritize their work in following up with customers before hand to get the payments on time.

EDA

Distribution of payments across due days

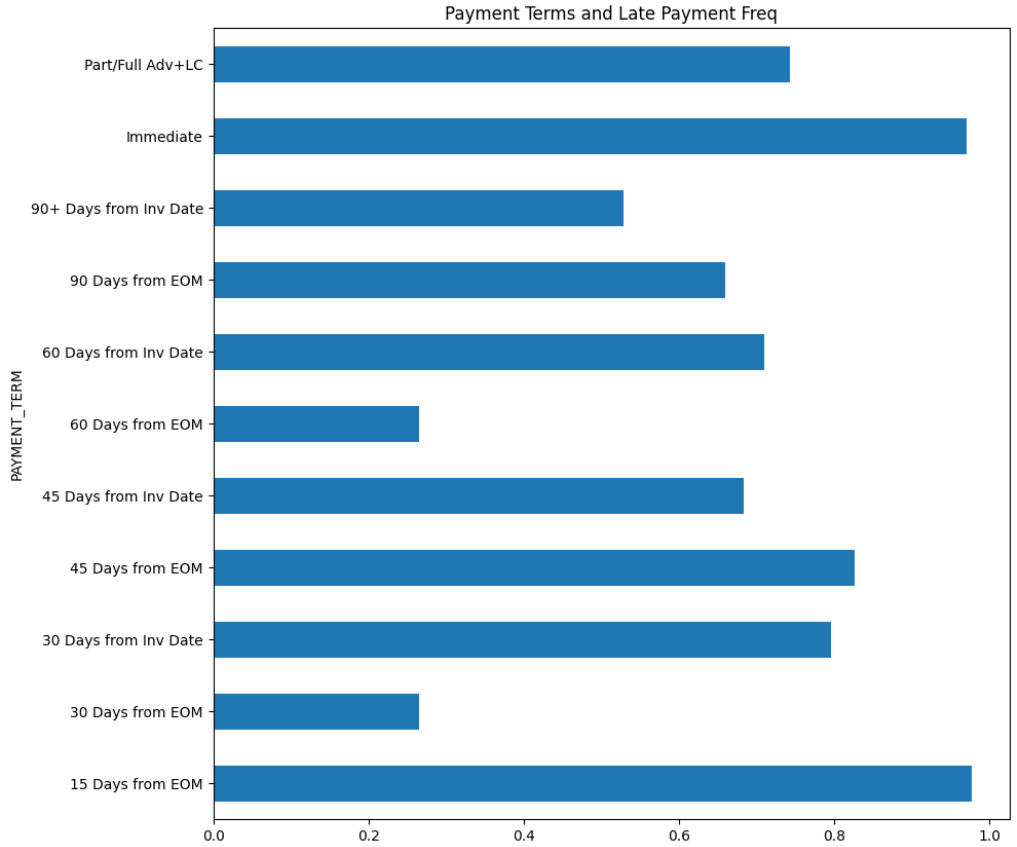


- Delay in payment
- Most of payments are cleared with no delay average
- 2 weeks delayed payments are above the average

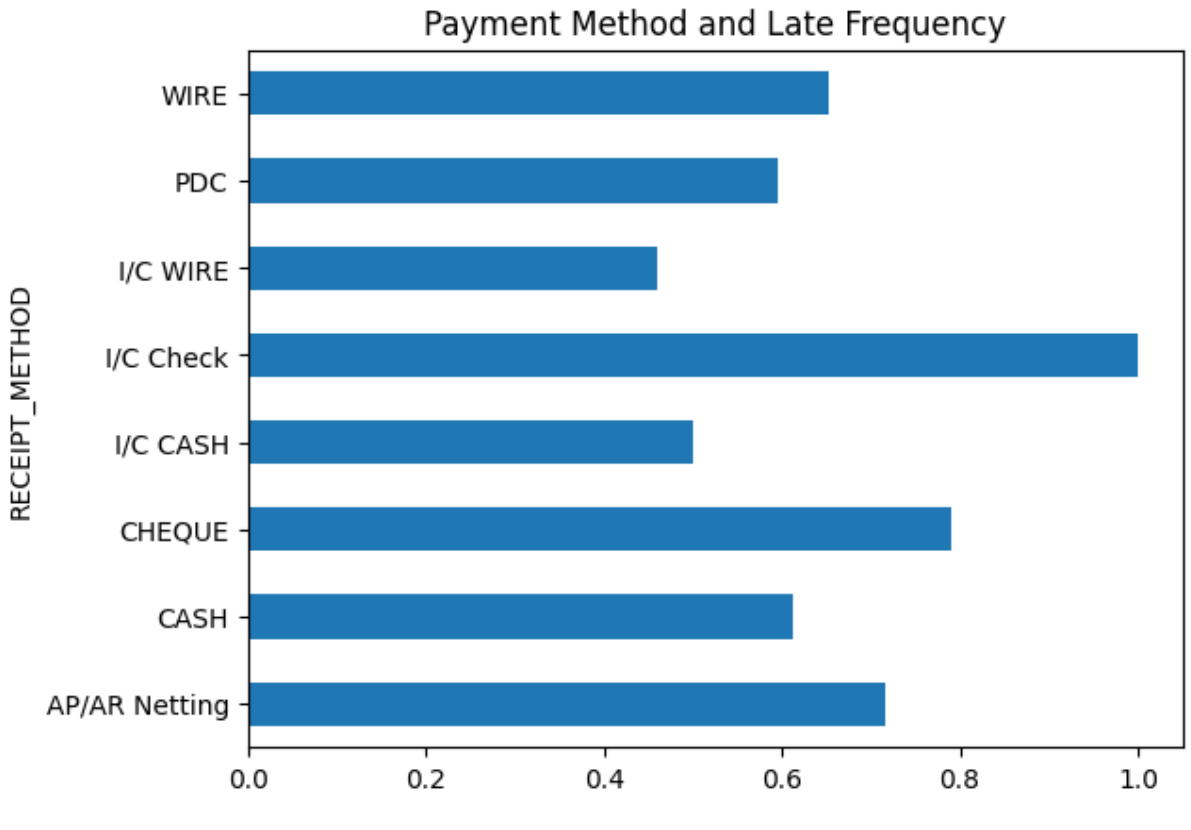


EDA

Late payment frequency by Payment term

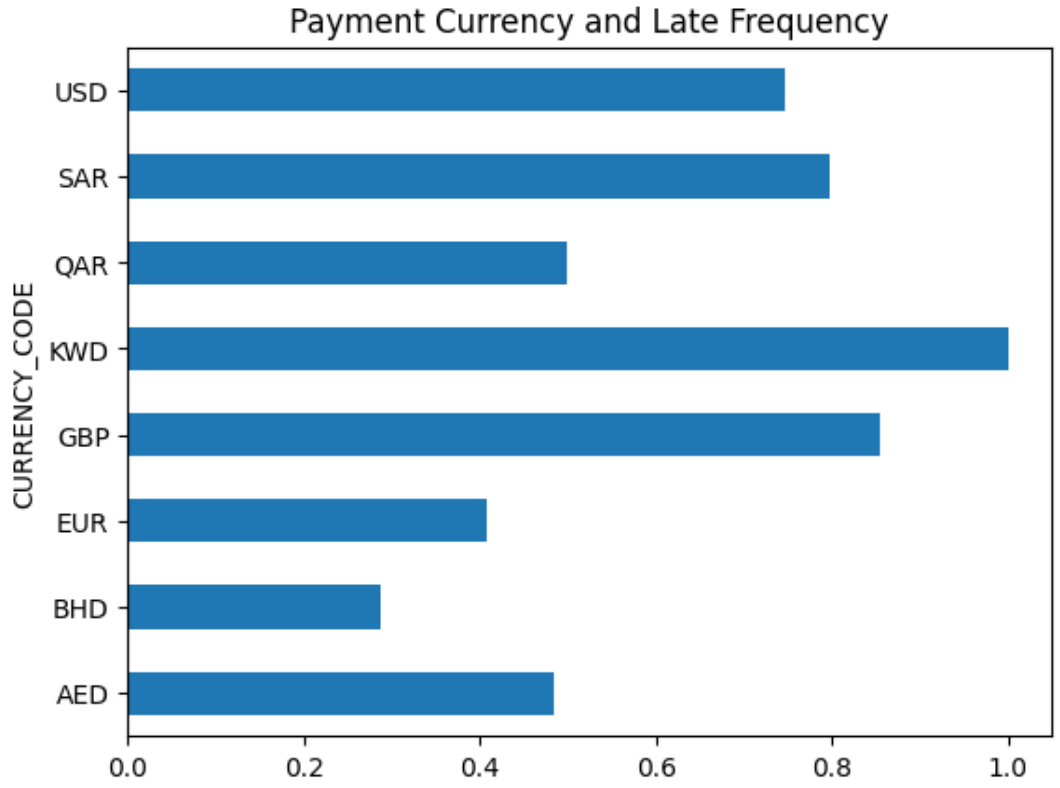


Late payment frequency by Payment method

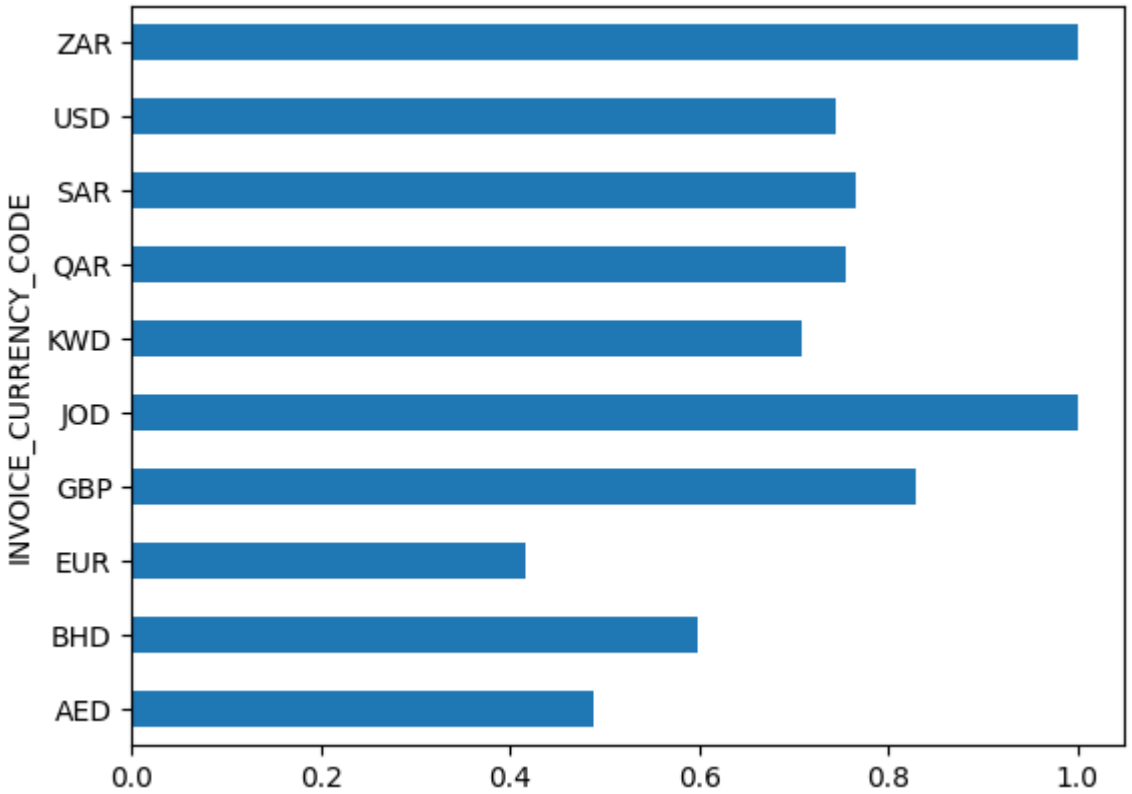


EDA

Late payment frequency by Payment currency

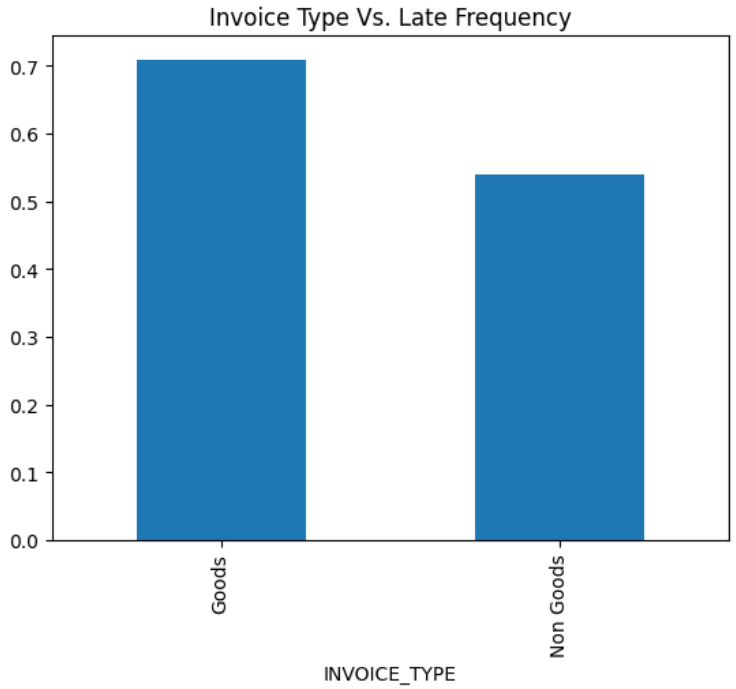


Late payment frequency by Payment invoice currency

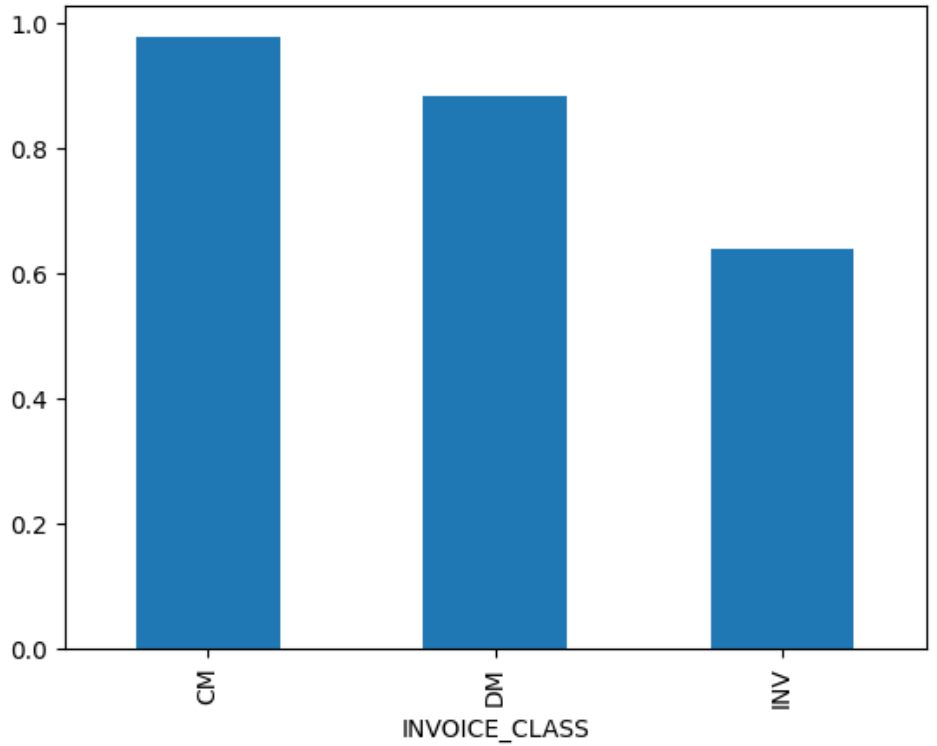


EDA

- Late payment frequency by Invoice class

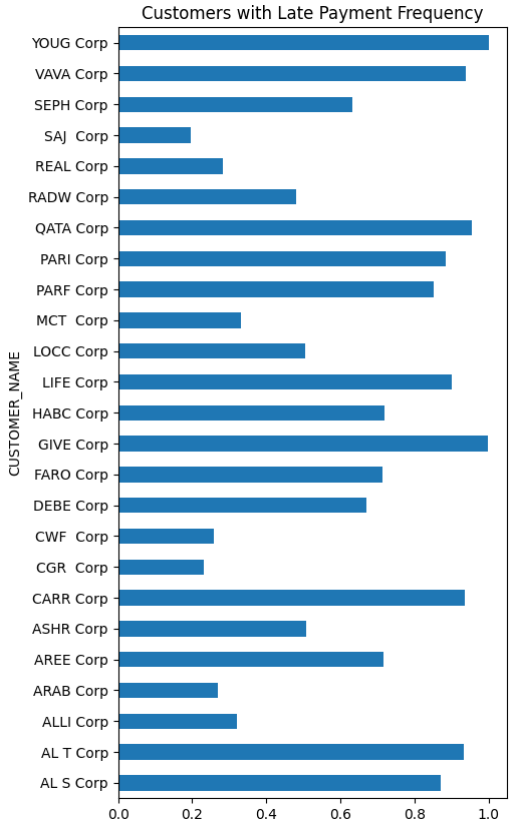


- Late payment frequency by Invoice type

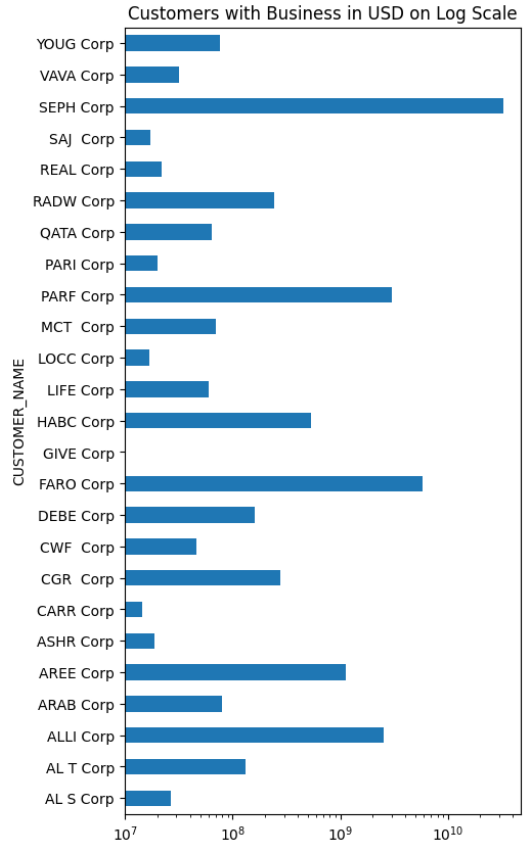


EDA

Late payment frequency by Customers

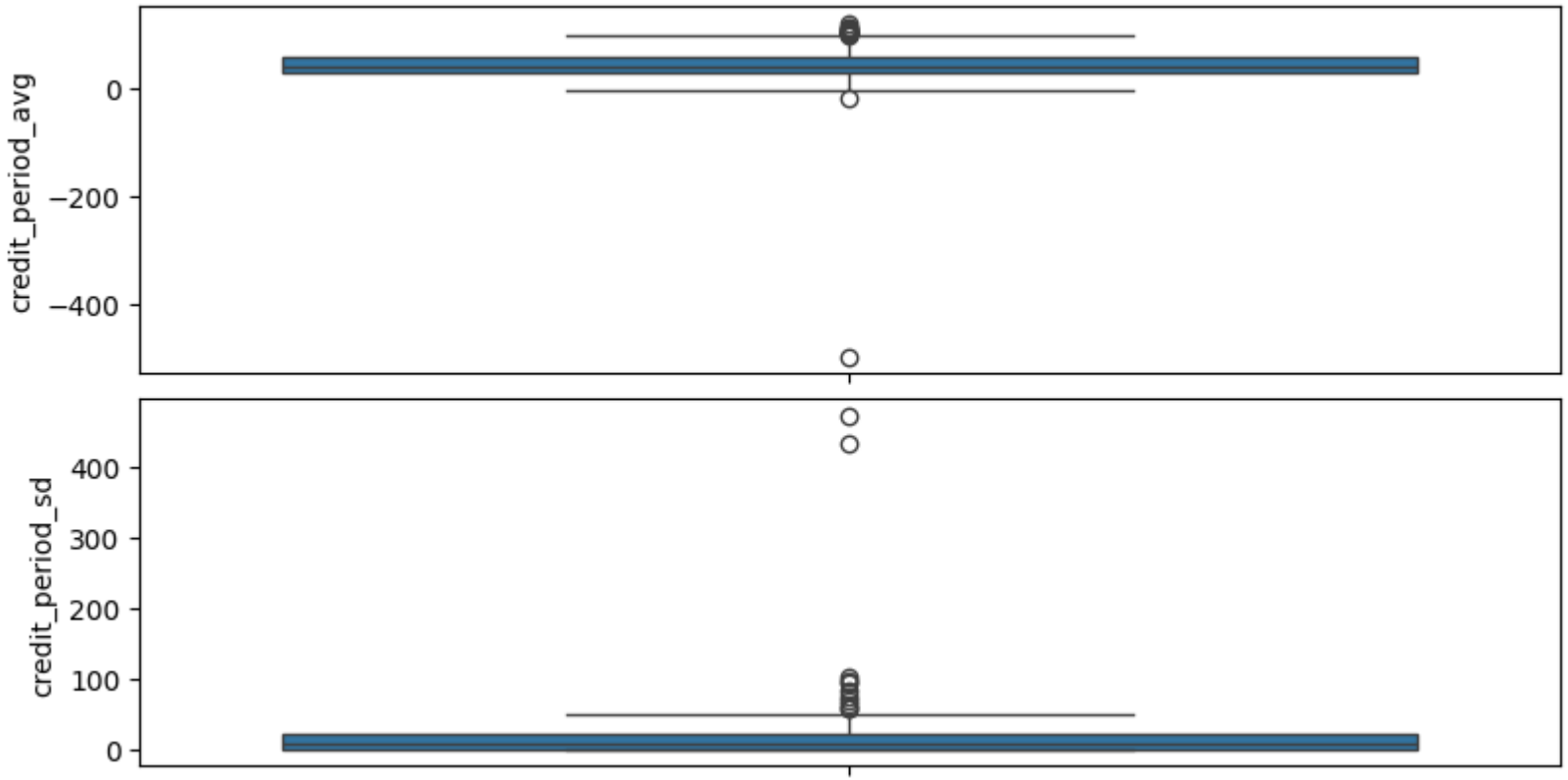


Customer count with Business in USD (log scaled)



EDA

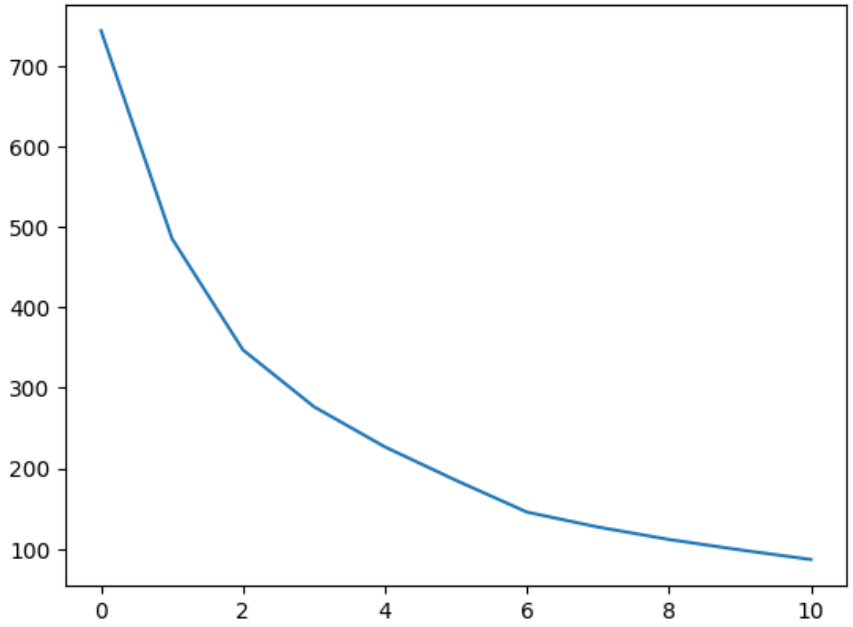
Used IQR(InterQuartileRange) method to remove outliers and scaled the data using Standard scalar method to do clustering



EDA

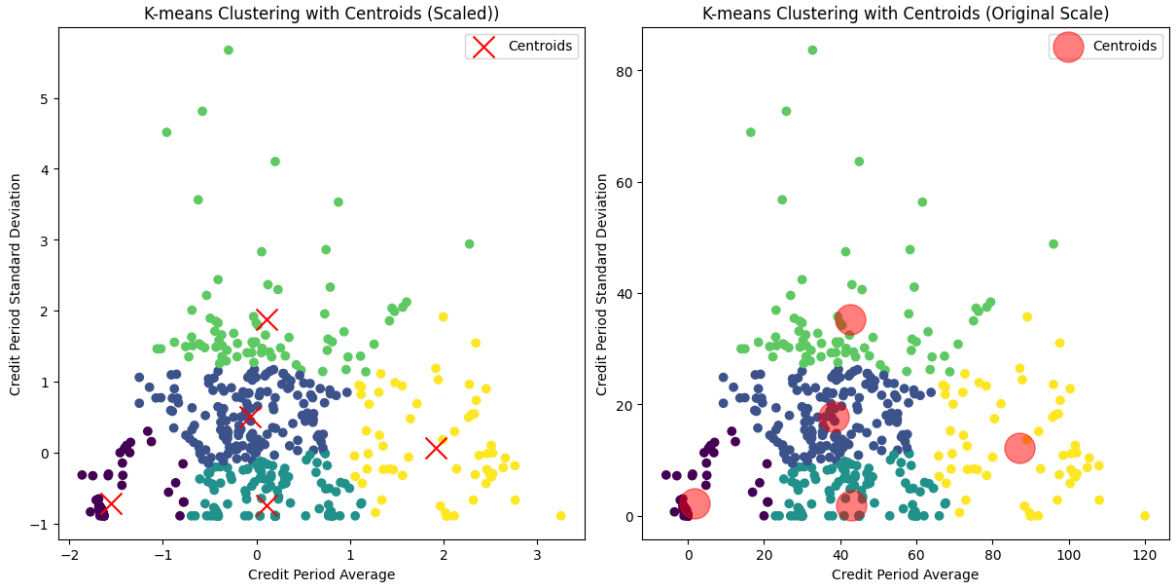
Customer segmentation

- K-Means clustering on scaled data.
- Elbow curve and Silhouette score to determine optimal cluster
- Too many cluster will loose its importance so choosing k=5byanalyzingSilhouettescore



For n_clusters=2, the silhouette score is 0.3927907291076642 7
For n_clusters=3, the silhouette score is 0.396733974500572 4
For n_clusters=4, the silhouette score is 0.463110044261739 1
For n_clusters=5, the silhouette score is 0.440303149831987 9
For n_clusters=6, the silhouette score is 0.444554259055814 5
For n_clusters=7, the silhouette score is 0.451478078171557 44
For n_clusters=8, the silhouette score is 0.478659247805615 4
For n_clusters=9, the silhouette score is 0.488010152972929 7
For n_clusters=10, the silhouette score is 0.489537193483609 1
For n_clusters=11, the silhouette score is 0.508755930049355 8
For n_clusters=12, the silhouette score is 0.528530630081288 8

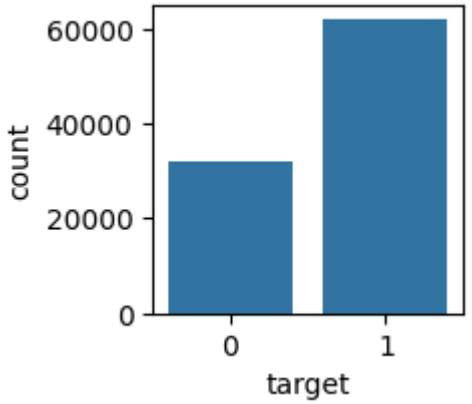
CLUSTERING



- There are clear 5 clusters of customers having different average payment days
- Most of the customers are offered between 20 to 60 days of payment terms on an average
- When credit period is under 20 days on average, the variability in credit period is very less.
- When credit period is more than 60 days on average, there is relatively moderage variability in the offered credit period
- The variability is highest when credit period is between 20 to 60 days

Data Preparation

The dataset is moderately imbalanced with approx. 66% delayed and 34% not delayed.



Feature Selection

- Selected top features which are>0.02, other columns are dropped as they do not contribute much

```
USD Amount 0.498703
credit_period 0.173386
PAYMENT_TERM_30 Days from EOM 0.097347
PAYMENT_TERM_60 Days from EOM 0.076179
INVOICE_CURRENCY_CODE_SAR 0.025932
PAYMENT_TERM_15 Days from EOM 0.016894
PAYMENT_TERM_60 Days from Inv Date 0.016142
INVOICE_CURRENCY_CODE_USD 0.015978
PAYMENT_TERM_Immediate 0.010788
PAYMENT_TERM_Immediate Payment 0.010682
dtype: float64
```

Model Selection

- Random Forest performs much better with highAccuracy, Precision, Recall andF1Score.
- Logistic Regression has better Recall than Random Forest but much lower Accuracy and Precision.

```
clasification report:
              precision    recall  f1-score   support

           0       0.86      0.78      0.82      9588
           1       0.89      0.94      0.91     18594

    accuracy                           0.88     28182
   macro avg       0.88      0.86      0.87     28182
weighted avg       0.88      0.88      0.88     28182


confussion matrix:
[[ 7512  2076]
 [ 1174 17420]]
```

Conclusion and Recommendations

The analysis identifies the top 10 contributors to delayed payments,with the highest impact factors being USD Amount, credit_period ,and specific payment terms such as PAYMENT_TERM_30 Days from EOM and PAYMENT_TERM_60 Days from EOM.This suggests that addressing these factors could significantly reduce payment delays

To improve the payment process, it is recommended that the client consider adopting milestone or staggered invoicing strategies rather than waiting to invoice for the entire order at once.

Additionally, it is important to exercise caution with payment termssuchasPAYMENT_TERM_30Days fromEOMandPAYMENT_TERM_60Days from EOM, as these have been identified as contributors to delayed payments

Thank you