Schuster

E-Commerce & Retail B2B Case Study

Submitted By

Pragati Saxena

Problem Statement :

Schuster is a multinational retail company dealing in sports goods and accessories. They are facing problems with vendors respect credit terms and some of them tend to make payments late. Schuster would wants try to understand its customers' payment behaviour and predict the likelihood of late payments against open invoices.

Objective:

- Analyse the customer transactions data to find different payment patten
- Segregate the customers based on their previous payment patterns
- Predict the likelihood of delayed payment against open invoices from the customers
- Insights from the Developed Model

Approach:

1. Reading the Data and cleaning the data.
   a. Feature RECEIPT_DOC_NO had about 0.03% of null values. They were dropped.
2. Derived the target variable ‘INTIME’ from the difference between 'DUE_DATE’ and RECEIPT_DATE’.
3. Since feature ‘PAYMENT_TERM’ was in string format, we derived this feature from the difference between 'DUE_DATE' and 'INVOICE_CREATION_DATE’
4. We saw that the ‘Received_Payments_Data’ had 65% of the data points as delayed and the rest of the transactions were done Ontime.
5. Performed EDA to derive insights from the features. The highlights of our analysis are described in the Business Insights section.

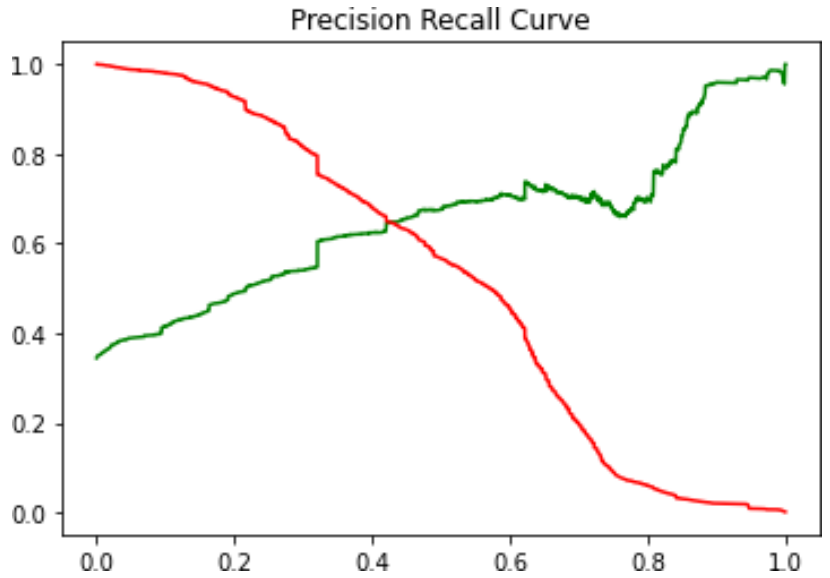
6. We performed clustering using Kmeans
7. The output of the KMeans clustering was used as an input to the Logistic Regression performed in the following steps.
    a) Sum of Squared Distances was used to determine cluster size.
    b) Silhoutte analysis was performed to determine the cluster size.

8. Logistic Regression was performed to derive the probabilities
    a) MinMaxScaler was use to scale the data
    b) Correlation matirix was created to visualize correlations between features.
    c) RFE was used to for feature selection.
    d) Statsmodels summary and VIF were used to eliminate the insignificant features.

9. We trained the model on 70% of the data and calculate the
   a) Accuracy score was calculated ~75% on the train data.
   b) Confusion matrix was created.
   c) Precision recall curve was created to arrive at an optimal probability cutoff. ~0.42

10. Prediction was done on the test set and below metrics were seen

1. Accuracy Score : 0.76
2. Precision : 0.65
3. RECALL : 0.34



11. The Open_Invoice data was used to predict the probabilities of the late payment. We have made recommendations for Shuster at the end of this presentation.

Business Insights

Inferences from Univariate analysis:

1. RECEIPT_METHOD : We can see that 'WIRE' transfer followed by 'AP/AR Netting' is the highest receipt method

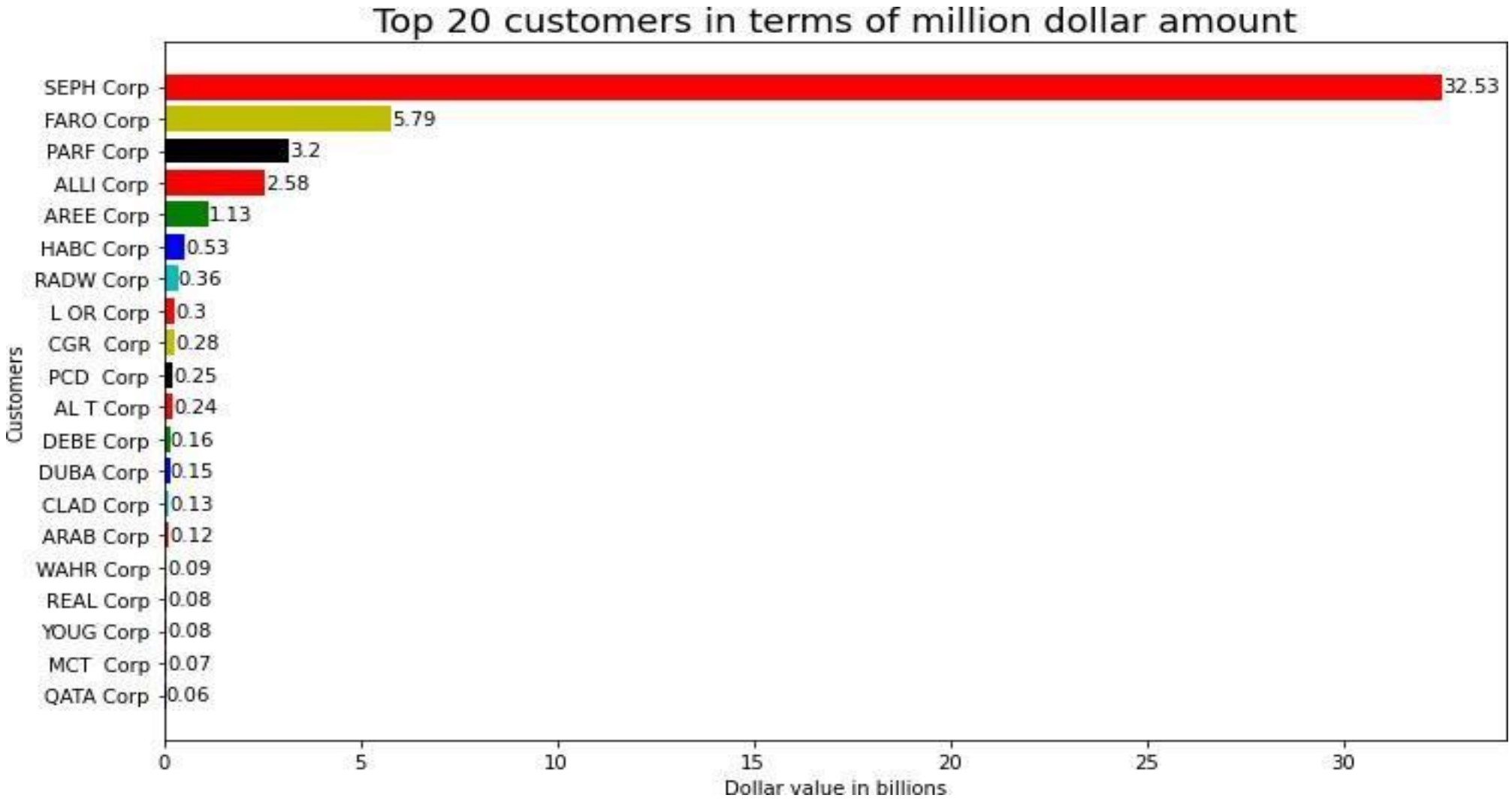
2. CURRENCY_CODE : We can see that 'AED' followed by 'SAR' and 'USD' is the most frequently used currency

3. INVOICE_CLASS : We can see the 'INV' (Invoice) followed by (CM) Credit Memo or Credit Note and then by (DM) Debit Memo or Debit Note

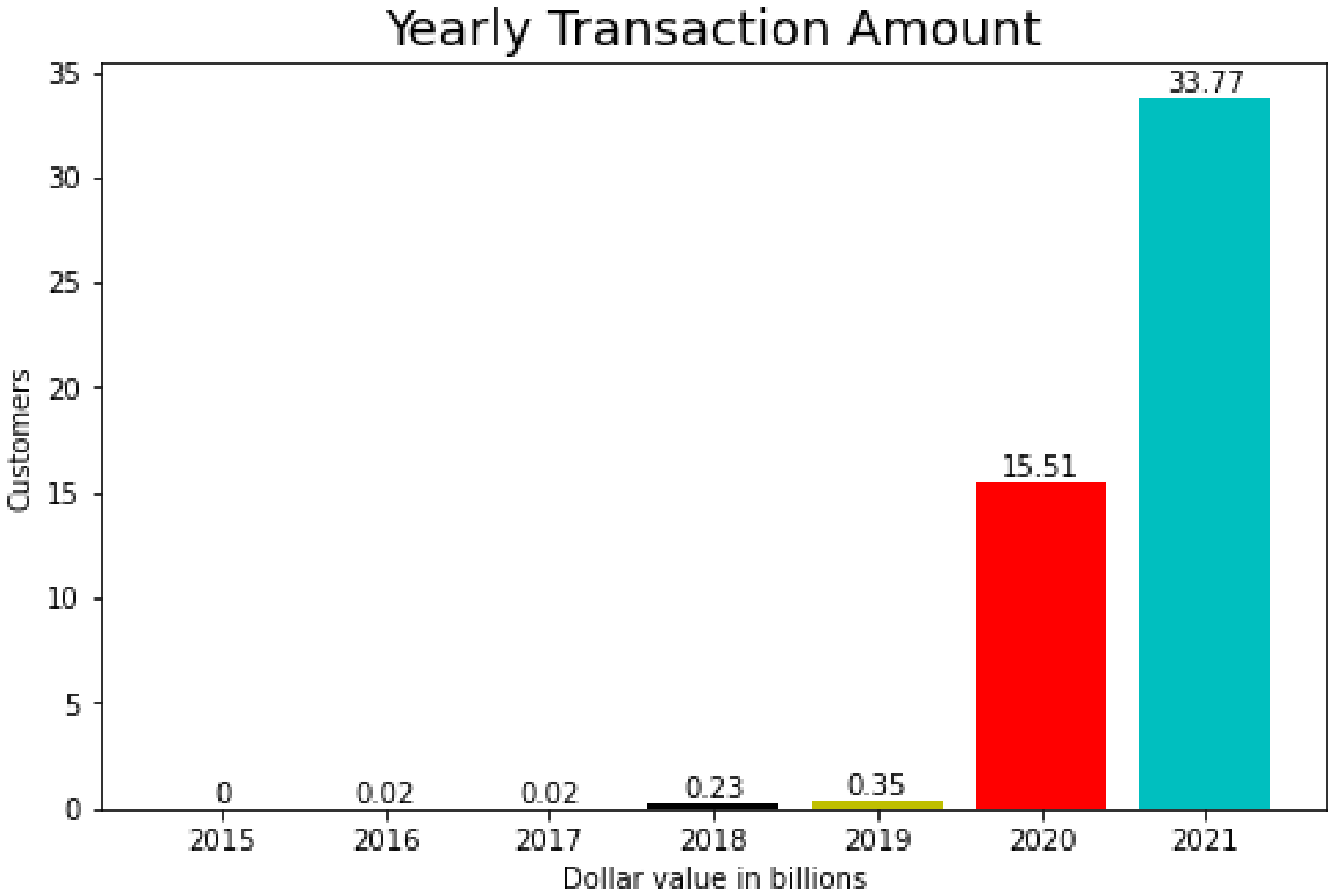
4. INVOICE_TYPE : Goods are the most invoiced type over Non Goods.

5. INTIME: 34.35 % of the payment happen on time and more than 65.65 % of payments that are delayed

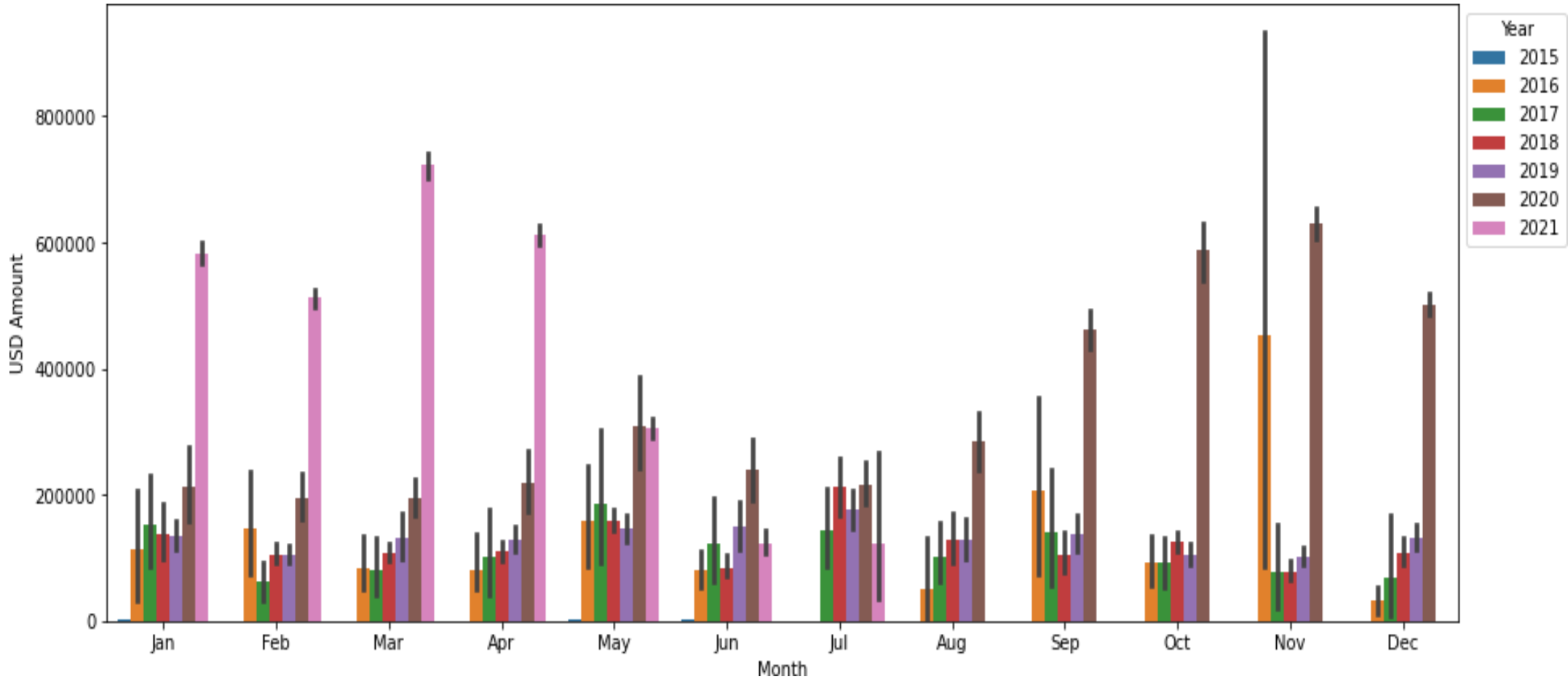
Customer Data in terms of Dollar Amount



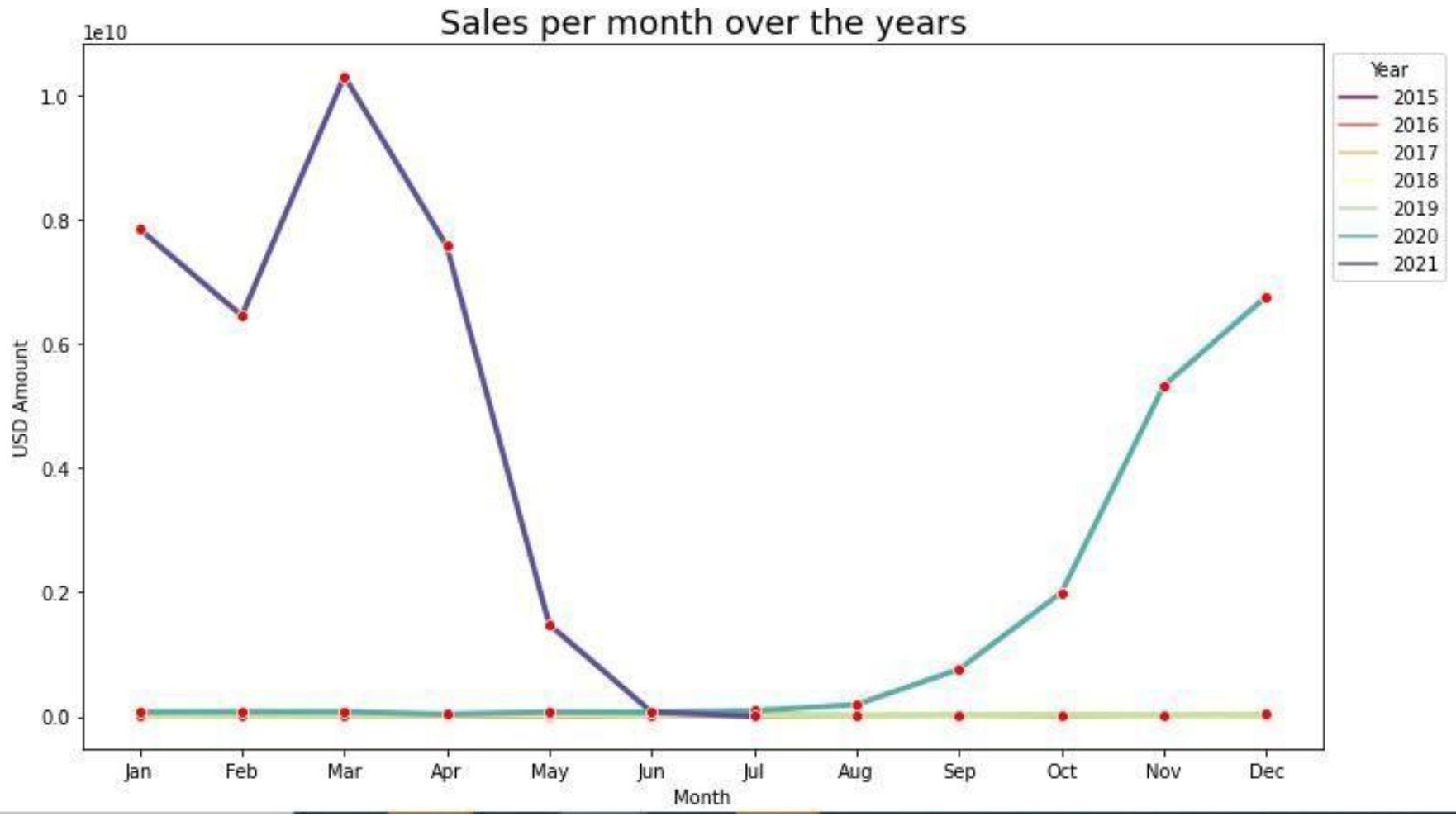
Visualizing the total yearly transactions by year for the company



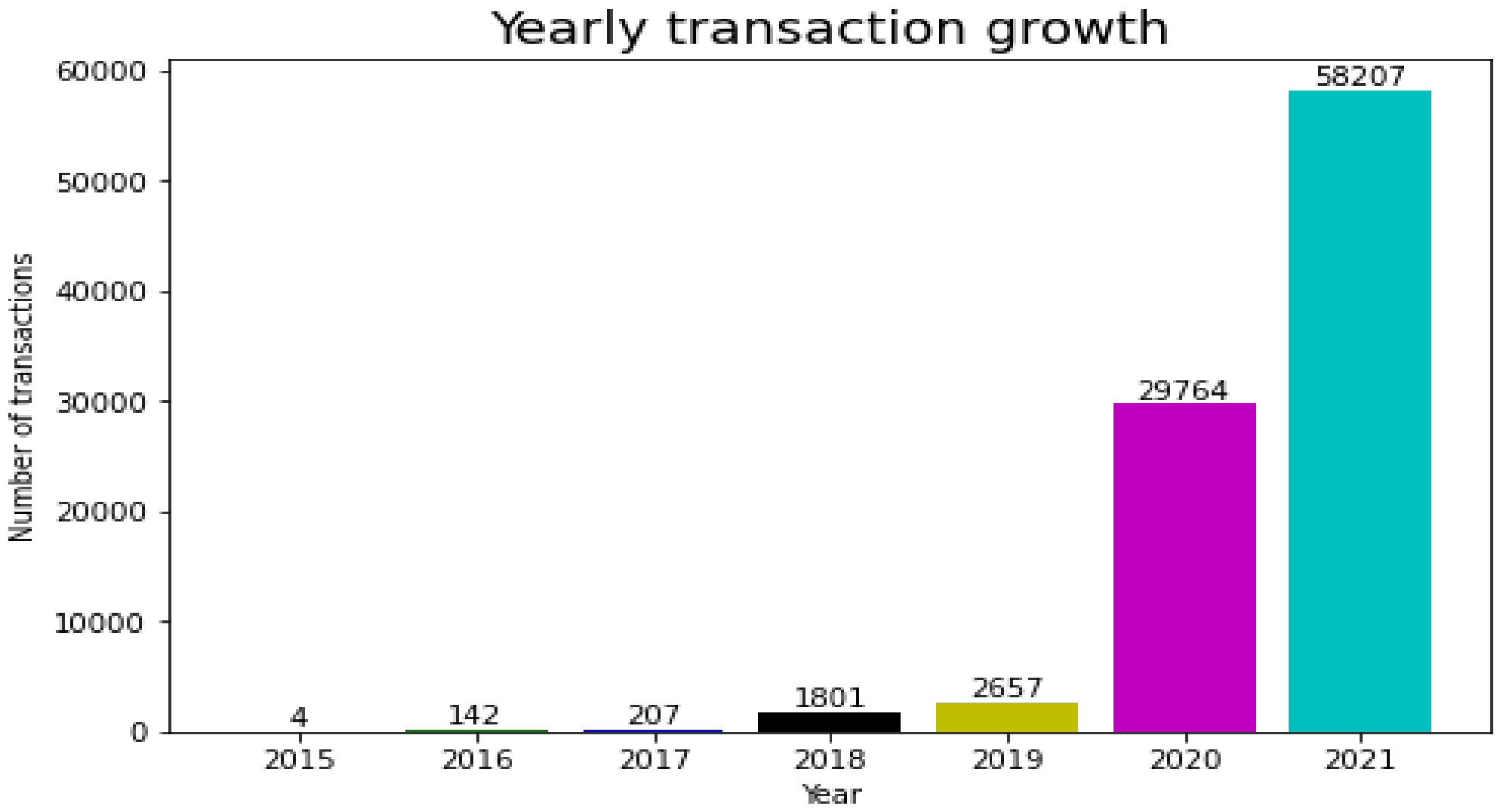
Visualizing the monthly transaction by year for the company



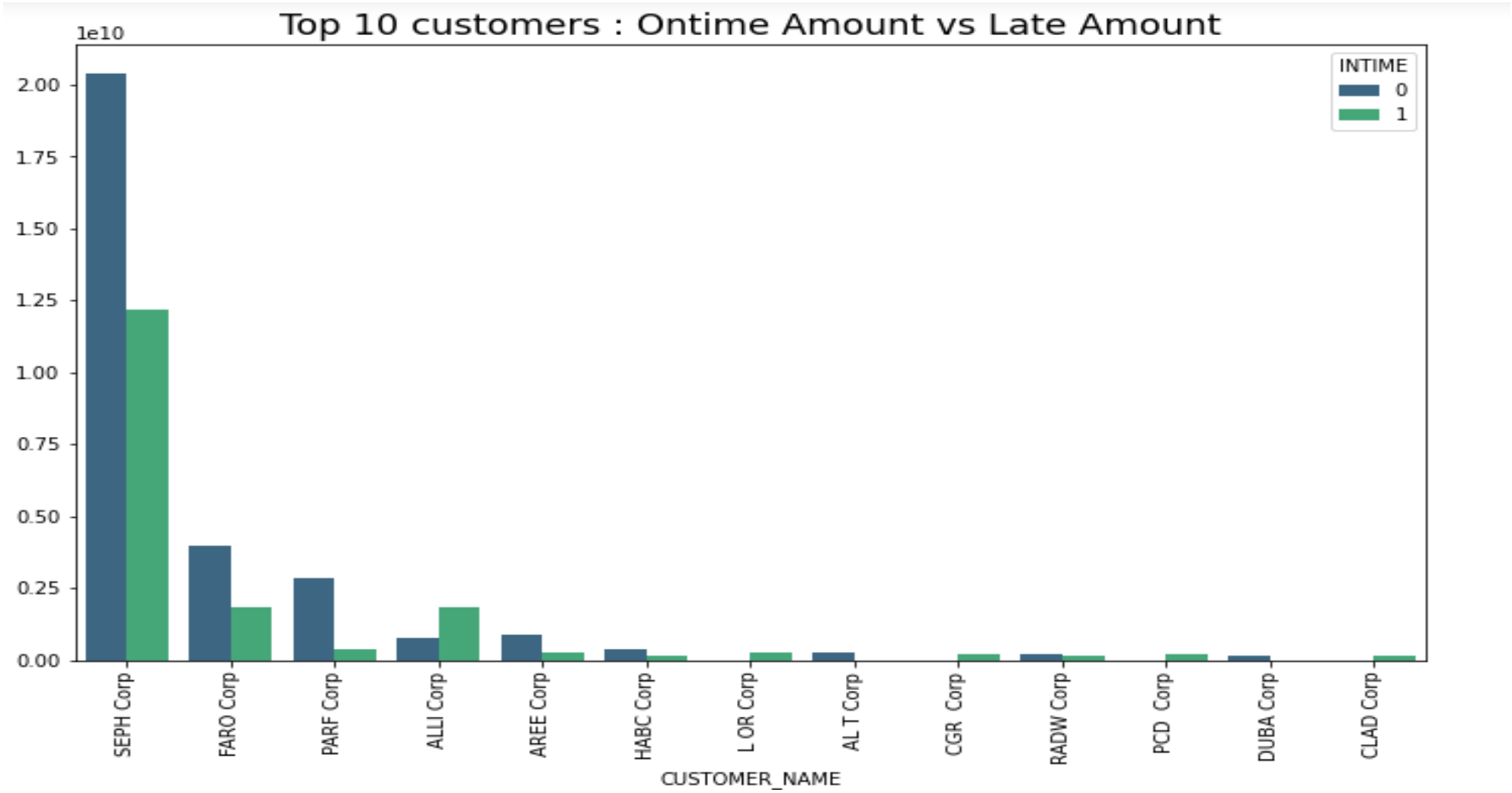
Monthly sales across the years



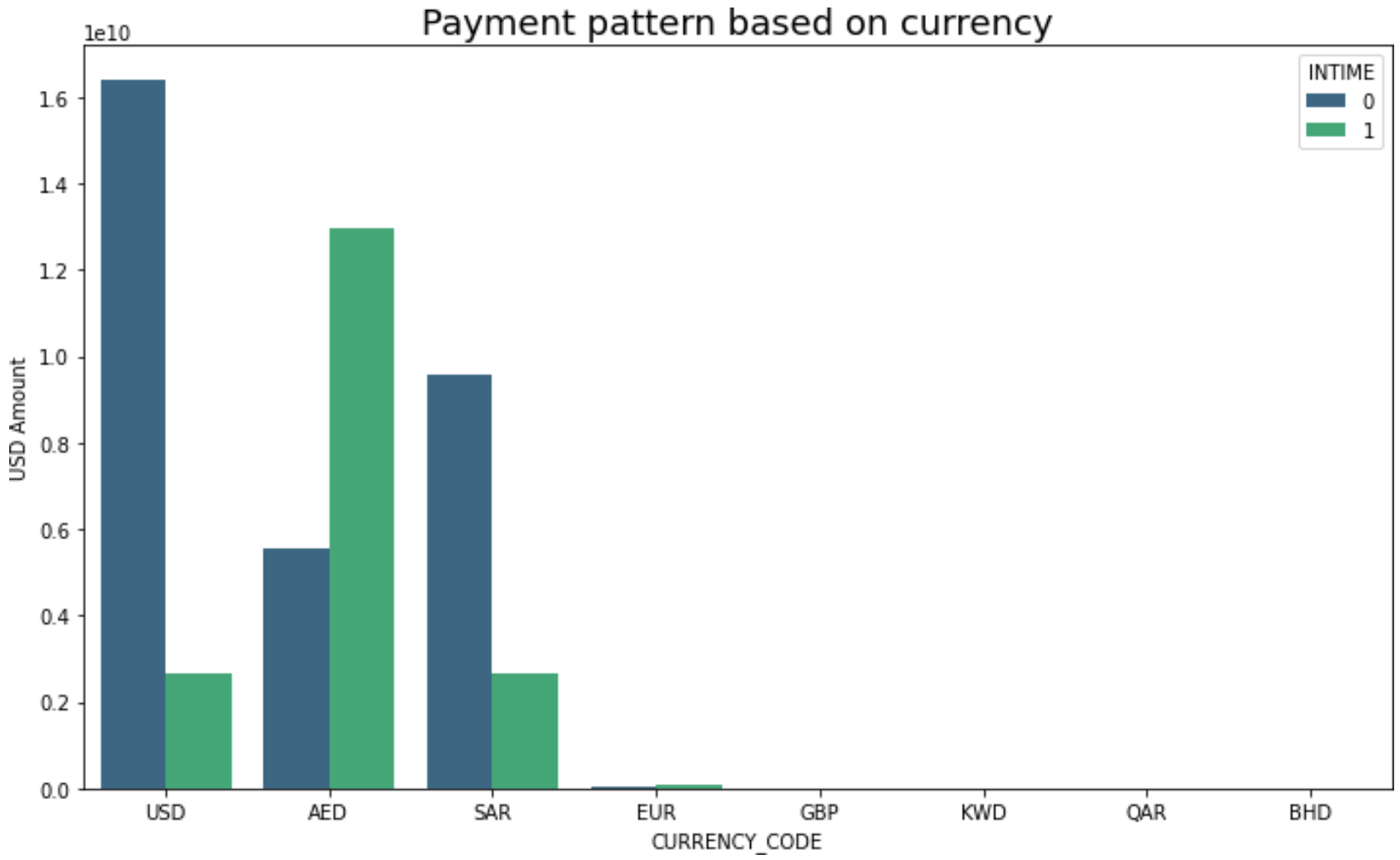
Increase in yearly transaction count.



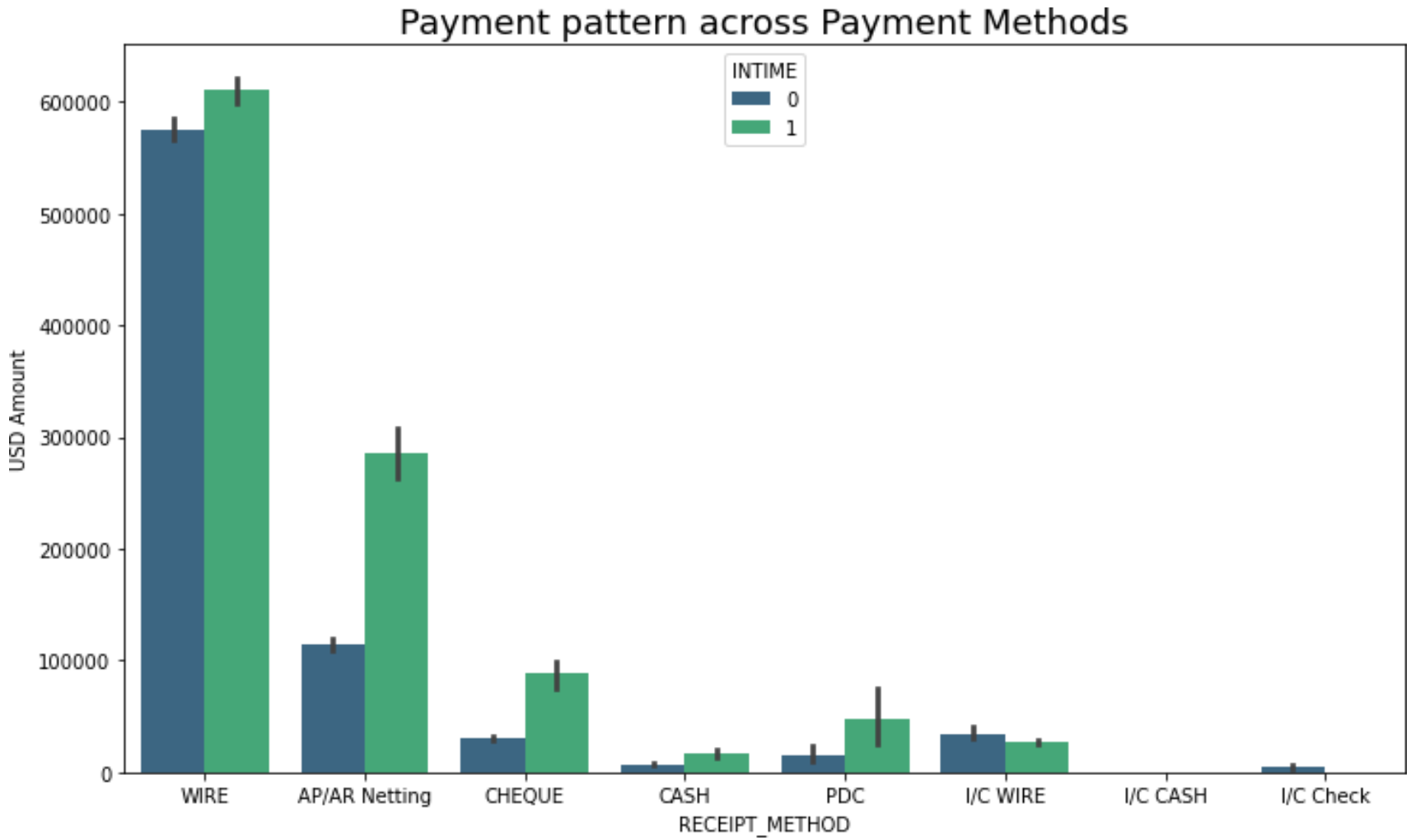
Visualizing the Ontime vs Late payment for top 20 customer.



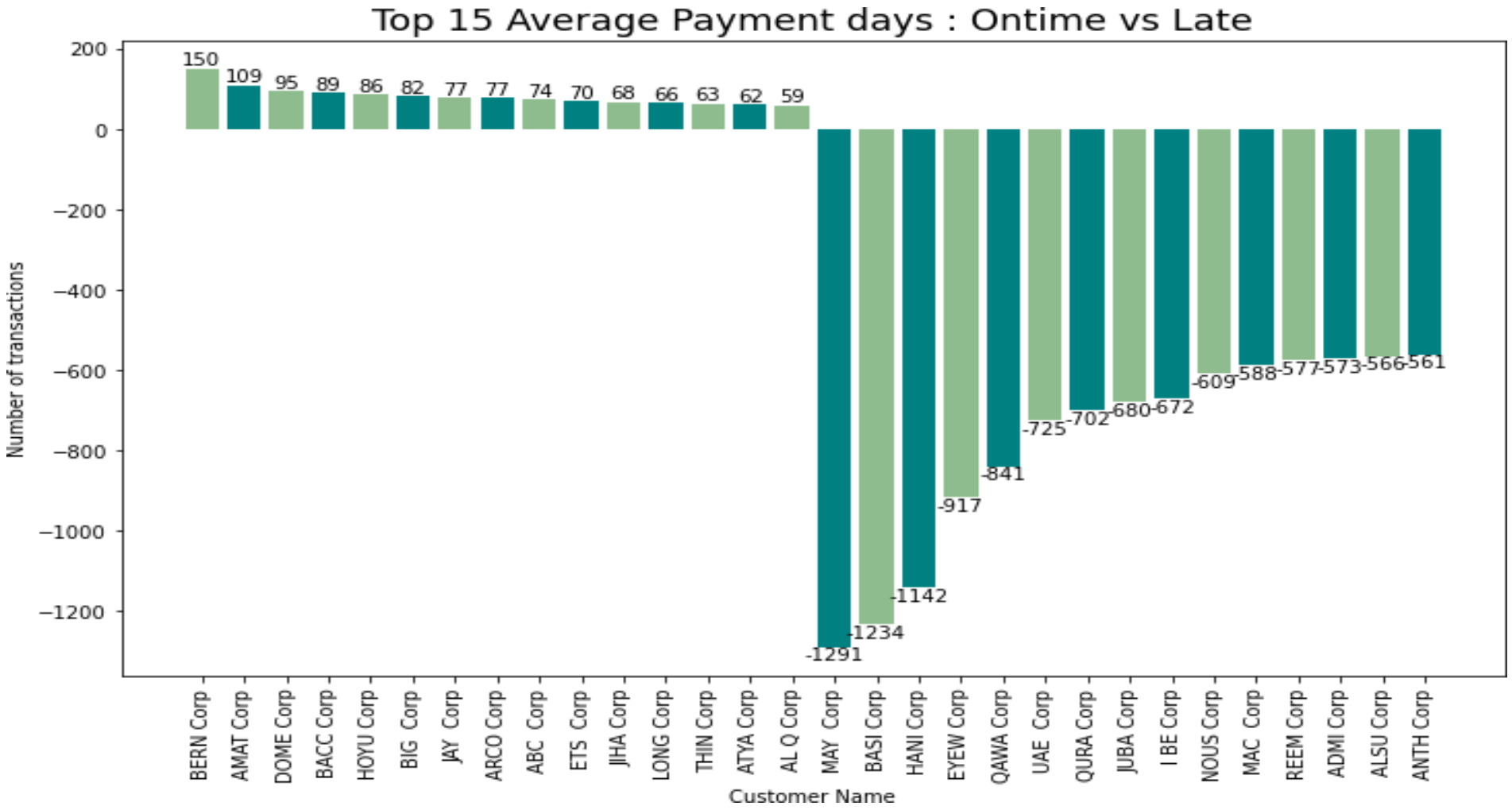
Visualizing the payment pattern base on currency code.



Ontime vs Late payment based on payment method.

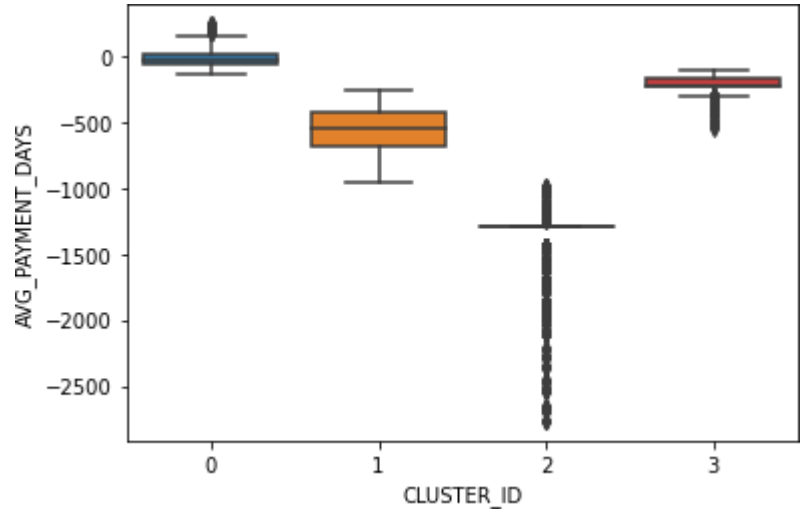


Visualing the top 15 Customers who have the best payment history vs customers with bad payment history



Recommendations for Schuster

From our clustering analysis we can make the following inference



i. Customers in cluster 2 has the highest average payment days, so that might be the ones that make a late payement
ii. Customers in cluster 1 has the second highest average payment days followed by cluster
iii. Customers in cluster 0 have the lowest average payment days and are more likely to make payements on time.

12. After analysing the predictions from logistic regression and clustering and combining the data. We have priotized the transaction that need to be focused to avoid late payment

The below dataframes contain the details of the Customers and Transactions that we need to focus on in the order mentioned below (Excel sheet with details are output in the code)

Late_hybrid_Priority_1

Late_hybrid_Priority_2

Late_hybrid_Priority_3

Late_hybrid_Priority_4

13. Based on the number of transactions that were predicted as Ontime or Late we have made a list of customers who have the highest transactions for each prediction.

We should focus on the customers with the highest transactions that are predicated late, this ensures that we can get a large volume of payments to be done on time. (Excel sheet with details are output in the code)

Late_Customers

Ontime_Customers