CREDIT CARD CUSTOMER SEGMENTATION

Abstract

Over the years, credit card issuers competing in similar segments with similar products find it hard to differentiate between customers based on their behaviours.

Segmenting customers primarily by their needs and attitudes while exposing the underlying reasons why customers use their credit cards and taking into account a small number of financial behaviours, helps us in achieving this problem.

Credit card issuers have traditionally targeted consumers by using information about their behaviours and demographics. Behaviours are often based on credit bureau reports on how a person spends and pays over time;
Customer segmentation is the practice of dividing a customer base into groups of individuals that are similar in specific ways. A customer segmentation model allows for the effective allocation of marketing resources and the maximisation of cross and up-selling opportunities.

Armed with enhanced segmentation, card issuers can not only craft better value propositions but also identify groups that are not well served by current offers.

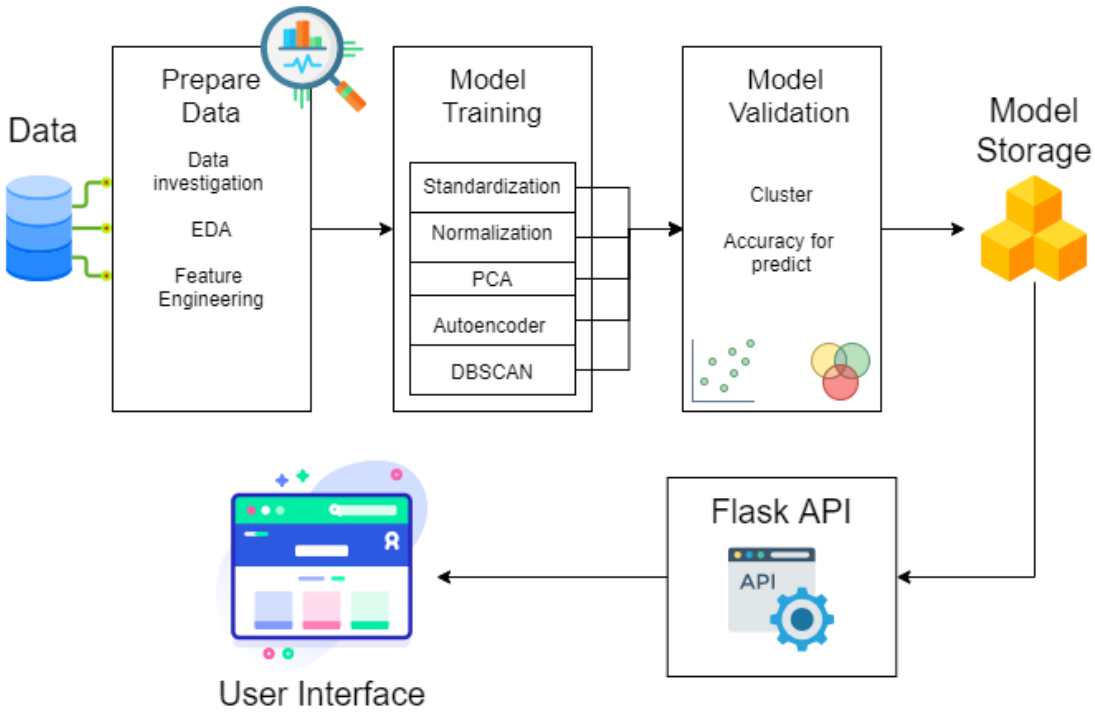
Problem Statement

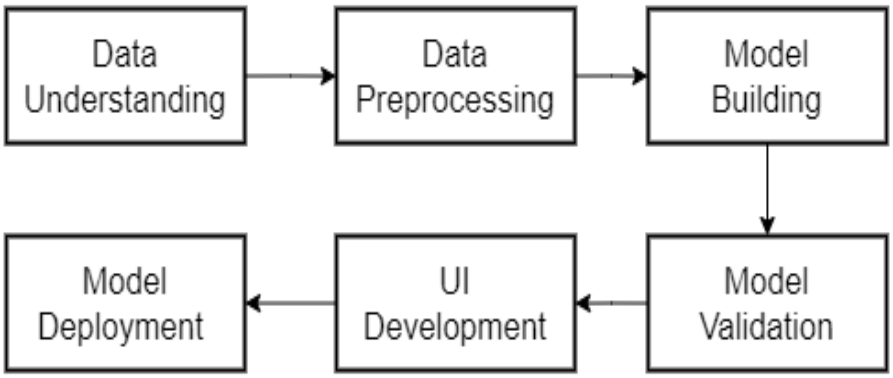
RBL's marketing department collects various customer specific data of the credit card holders. They need a mechanism to segment the customer based on underlying characteristics and form market clusters which will be easy for them to target and provide product ideas to the management. Currently these tasks are performed manually by trusting the judgment of experts in the field. These can lead to human error, biased decision and other factors which may not be helpful to create a customer cluster that actually exist. They want to design a system that would automate this process and help the different stakeholders to make informed business decision.

Objective

1. To develop a solution focused on developing a clustering algorithm based on unsupervised learning.
2. Try various clustering algorithm, and identify the best one for the business scenario and build and fine tune them based on characteristics.
3. Deploy the machine learning model using Flask API and pickle files.
4. Design a UI for entering model inputs and display results accordingly.

5. Architecture of the project





Project Overflow

Steps:

1. Initial Data Understanding:

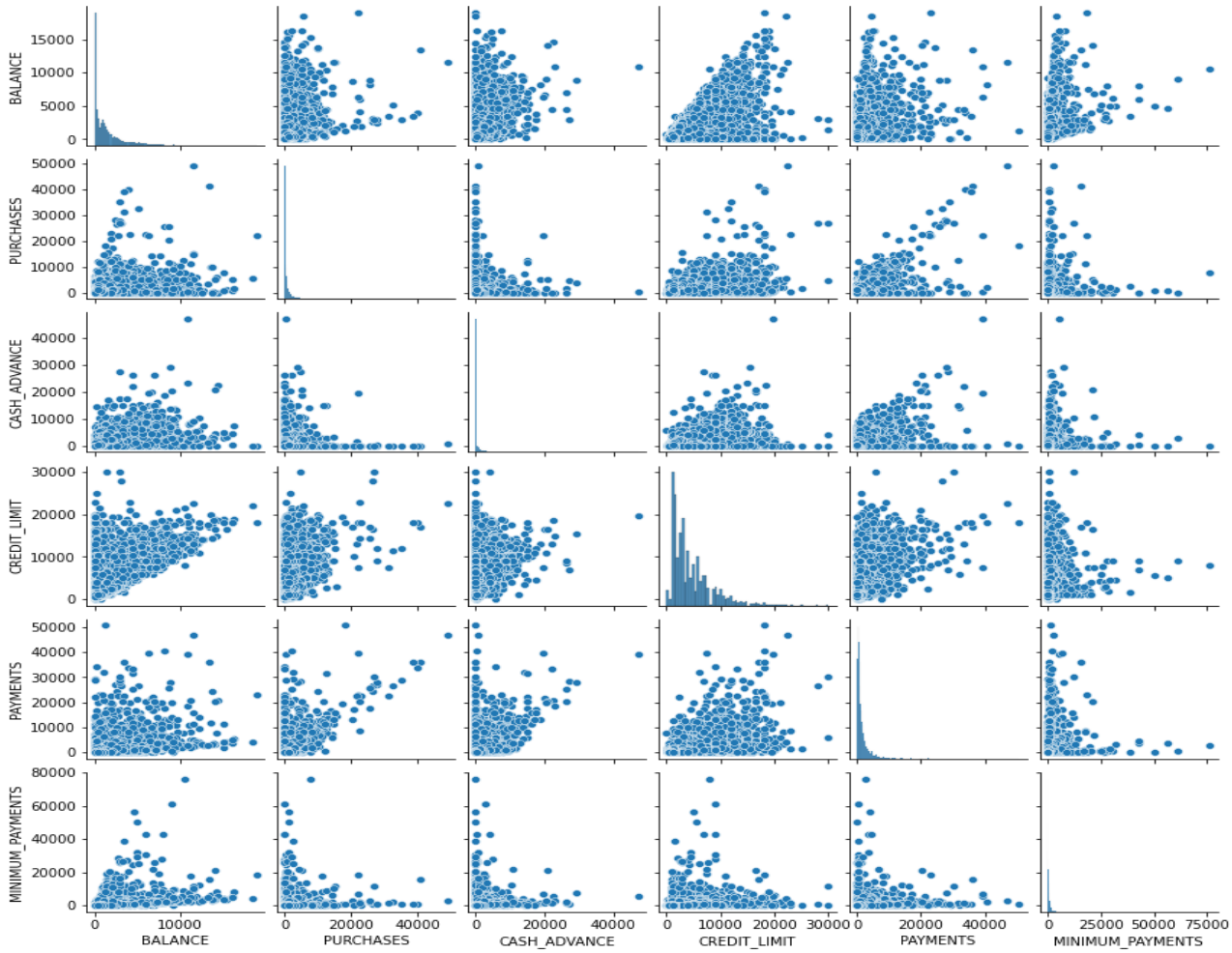
- Load the dataset onto Jupyter IDE and import all necessary libraries that will be needed for analysis.
- Performed initial analysis of the provided dataset, checking the format and type of the various numerical and categorical columns.
- Deduced the functional/business sense of each of those columns, their relevance in predicting the business solution and form initial hypothesis of the most critical features which could be crucial in designing the model.

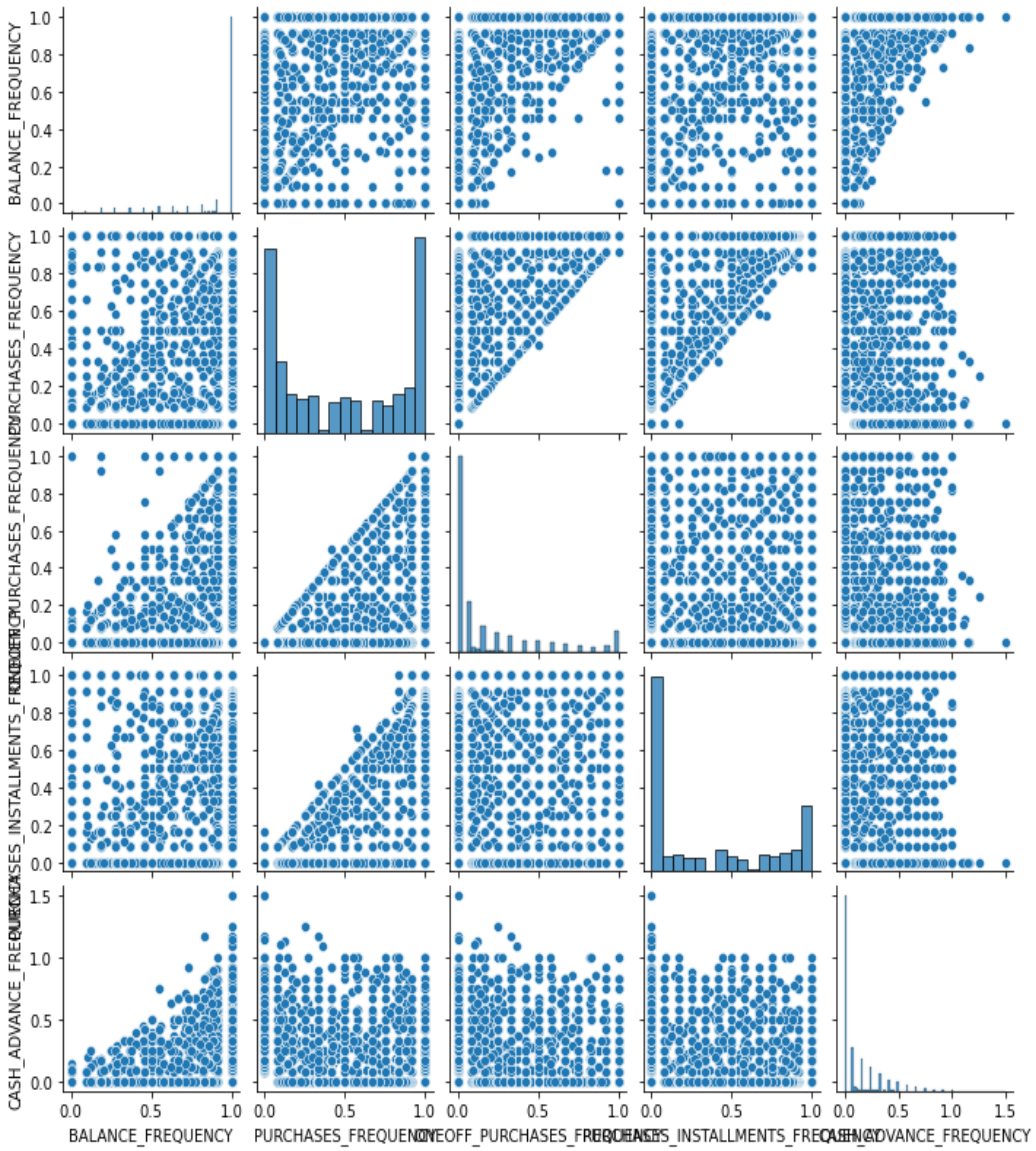
2. Data Preparation:

- Performed initial quality checks on the dataset.
- Checked the columns for missing values and treated those columns with feasible imputation methods.
- Column **CREDIT_LIMIT** with only one row having null value was dropped from the dataset.
- Column **MINIMUM_PAYMENTS** had around 313 missing rows. For some rows it seemed that for null value of MINIMUM_PAYMENTS, PAYMENTS column is also 0 which might suggest that the customer has not made any payment at all. For rows where MINIMUM_PAYMENTS column is Null, we imputed the value 0 considering that the customer has not made any MINIMUM_PAYMENTS.
- In a different approach we also can impute the median of that column to fill missing values.
- Column **CUST_ID** was dropped from the dataset, since no business sense can be derived from it and does not add value to our insights.
- The datatype of the columns was also checked and no discrepancies were found.

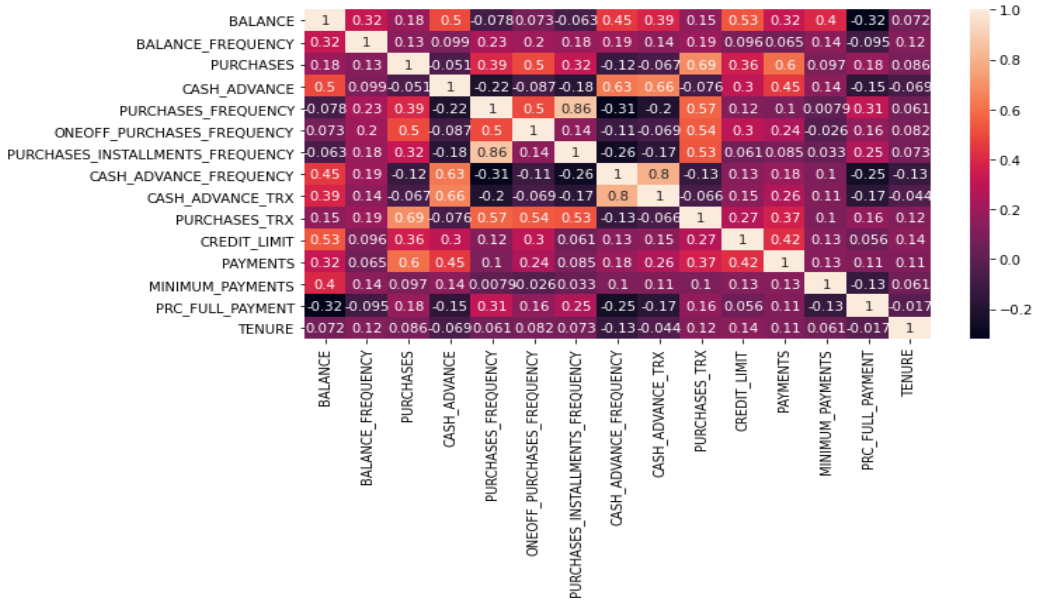
3. Exploratory Data Analysis

- For Exploratory Data analysis, we use the columns BALANCE","PURCHASES","CASH_ADVANCE","CREDIT_LIMIT","PAYMENTS","MINIMUM_PAYMENTS" And created scatter plot to check collinearity. We see that there is no significant visible collinearity.

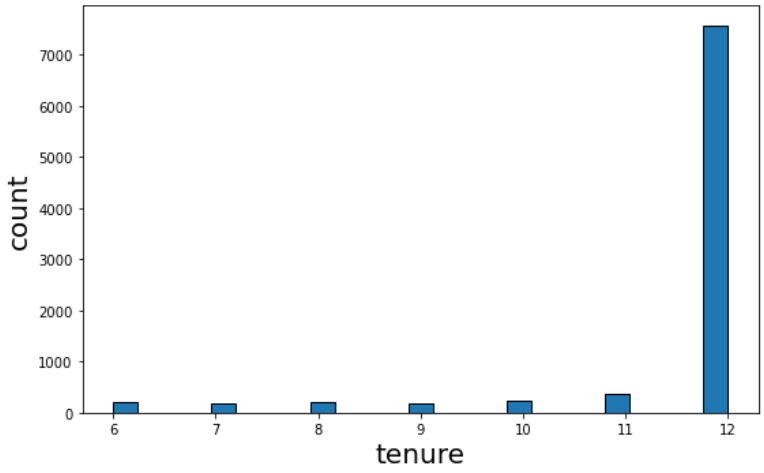

BALANCE
PURCHASES
CASH_ADVANCE
CREDIT_LIMIT
PAYMENTS
MINIMUM_PAYMENTS



- We use the columns BALANCE_FREQUENCY,PURCHASES_FREQUENCY, ONEOFF_PURCHASES_FREQUENCY, PURCHASES_INSTALLMENTS_FREQUENCY, CASH_ADVANCE_FREQUENCY. To check for collinearity, but here too we do not find any such collinearity.
- We further plot the heat map to further investigate.



- There is a correlation between the columns PURCHASE_INSTALLMENT_FREQ & PURCHASES_FREQ which is 0.86. So we may remove any one of these columns in the further processing.
- COLMN WISE ANALYSIS OF THE DATA SET:

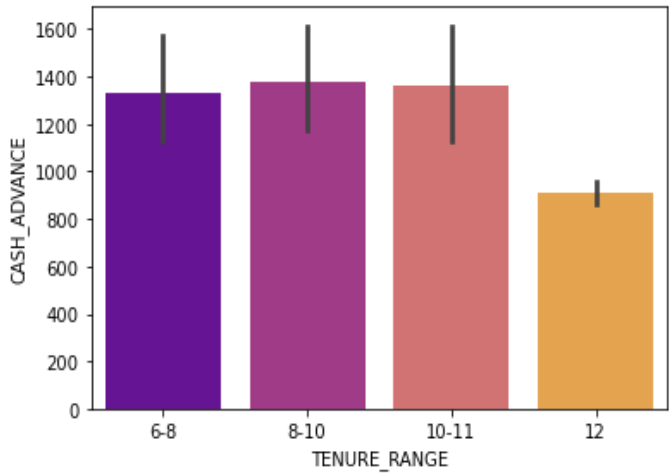


Most of the customers have 12 months as their Tenure. So further analysis of the data set we can perform binning of the column. Where we define the bins as 6-8, 8-10, 10-11, 12 months.

bins=[6,8,10,11,12]

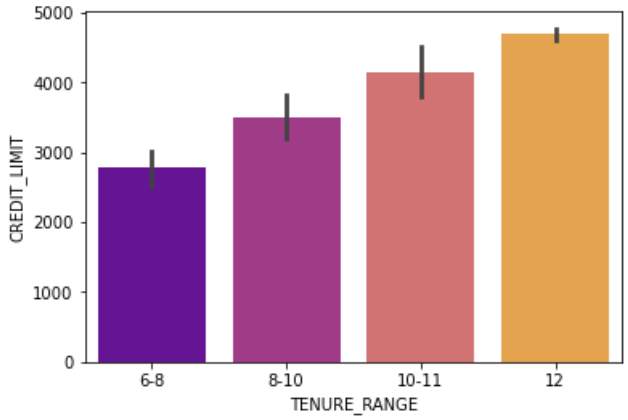
group_names=["6-8","8-10","10-11","12"]

- We Further carry on our analysis with respect to tenure.
- TENURE RANGE V CASH ADVANCE



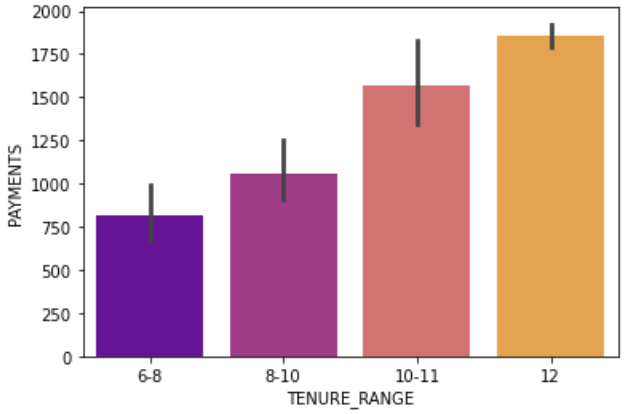
We see that majority of the customers who have taken maximum cash advance are in the tenure of 6-11 months. Even though their number is less.

- TENURE RANGE VS CREDIT LIMIT



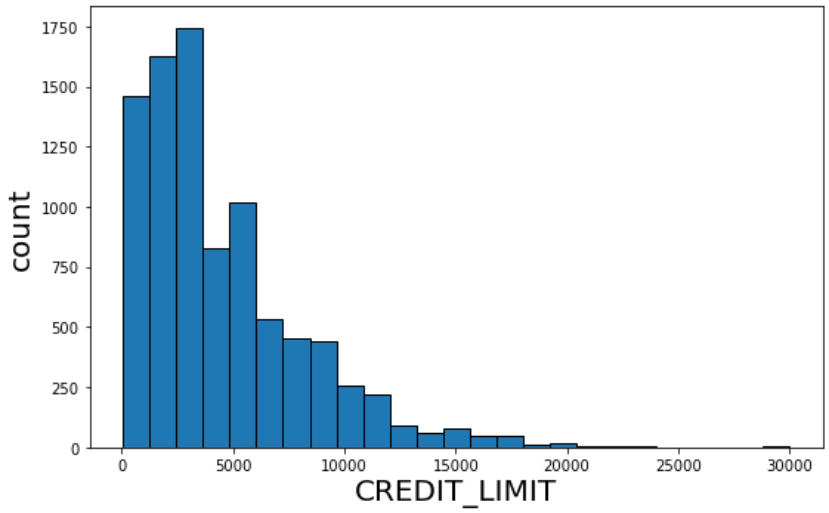
Higher is the tenure, higher is the credit limit.

- TENURE RANGE V PAYMENTS.



Higher the tenure higher the payments made

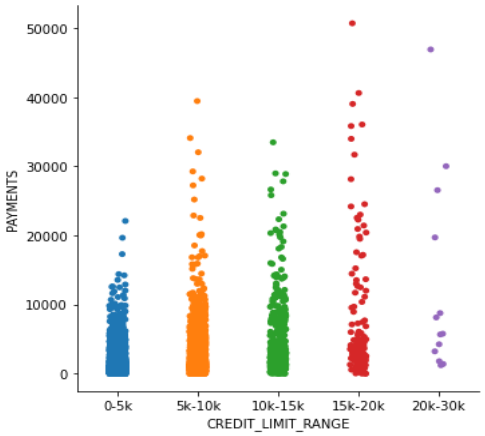
- Insights: Customers with a tenure range of 12 months have Higher spending power, They also pay the credit amount, and take the least cash advance.
- CREDIT_LIMIT VISUALISATIONS



MOST OF THE PEOPLE HAVE CROSSED THE CREDIT LIMIT.

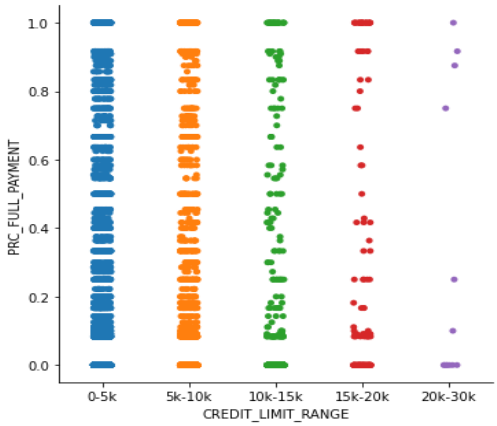
- We see from the above subplot that the majority of the customers have their credit limit belw 5000$ and between 5000 to 20000$. So we perform binning operations on them, to get better insights.
- We assign the bins as
- bins=[0,5000,10000,15000,20000,30000]
- group_names=["0-5k","5k-10k","10k-15k","15k-20k","20k-30k"]

- CREDIT LIMIT VS PAYMENTS



Significant customers from 0-15k have made the payments, upto 10k $ and there is a decrease in 15-20k and the least in 20-30k.

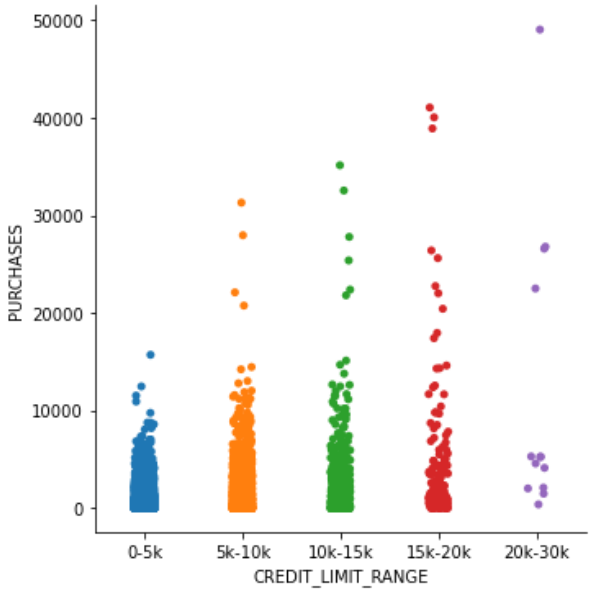
- PRC OF PAYMENT VS CREDIT LIMIT



Customers between 0-10k are paying regularly, compared to other credit ranges.

- People with credit limit bw 0-10k, and tenure of 12 months, are more comfortable in paying.

- CREDIT LIMIT VS PURCHASES

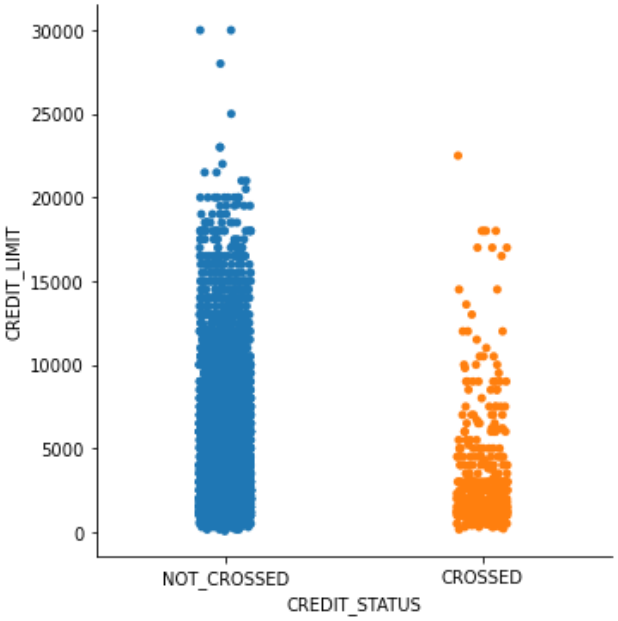


customers between 0-5k & 5k-10k credit have purchases beyond their limit Compared to others. Meaning there are customers who cross their credit limit, which must be taken note.

- Meaning there are customers who cross their credit limit, which must be taken note. customers with credit limit from 0-10k have a tendency to go befond their limit.
- Insights: People with credit limit bw 0-10k, and tenure of 12 months, are more comfortable in paying.
- Meaning there are customers who cross their credit limit, which must be taken note. customers with credit limit from 0-10k have a tendency to go befond their limit.

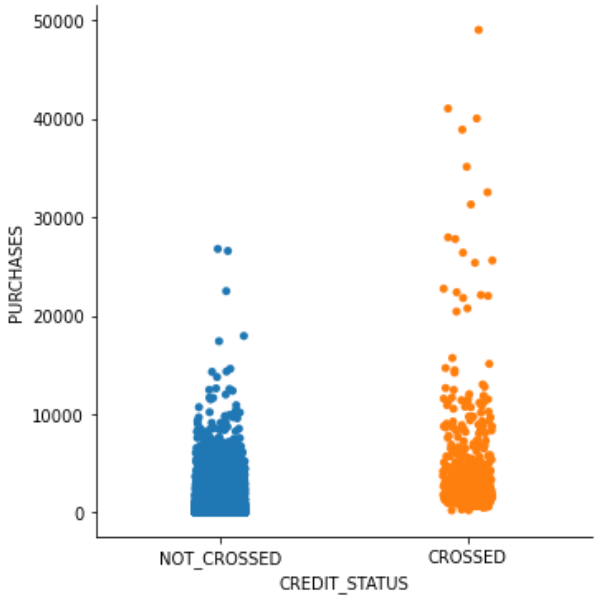
- So now we group the data set further into customers that have crosses the credit limit and not crossed the credit limit.

- Most of the customers have not crossed, but there are less than 1000 customers that have crossed. We analyse this further.
- We create a new column called CREDIT_STATUS to indicate this.
- CREDIT STATUS VS CREDIT LIMIT.



Customers with limit mostly between 0-5k and upto 20k are over purchasing.

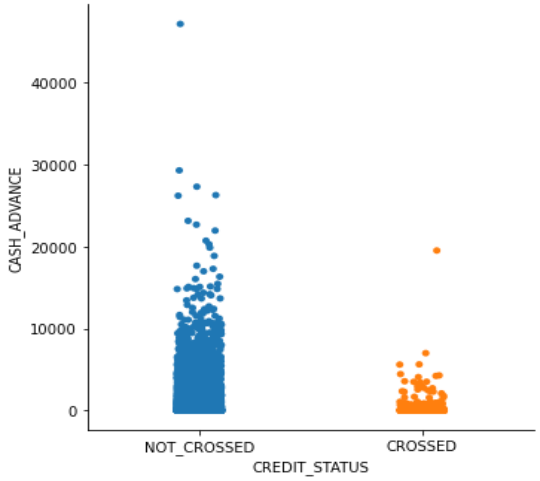
- PURCHASES VS CREDIT STATUS



Crossed customers lie between 0-10k and rise up to 40k.

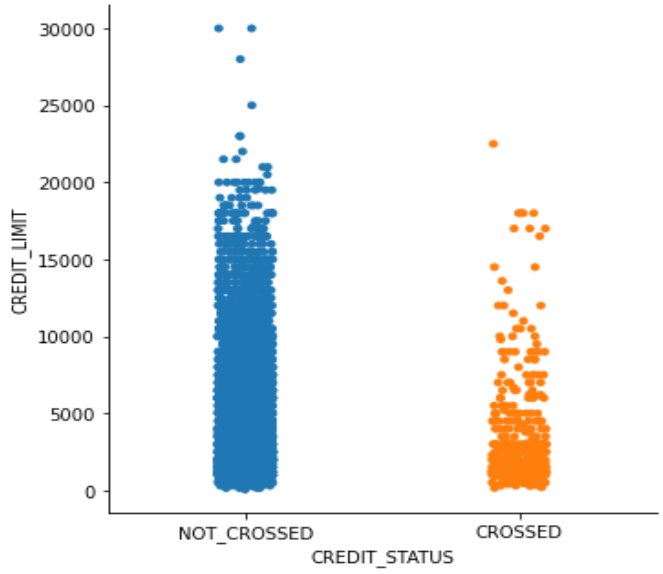
- Look out for customers with purchases from 0-10k

CREDIT STATUS VS CASH ADVANCE



Most people who have cash advance do not cross credit limit.

- Taking cash advance is not a sign of crossing.
- CREDIT LIMI VS CREDIT STATUS.



people with credit limit between 0-5k have mostly crossed.

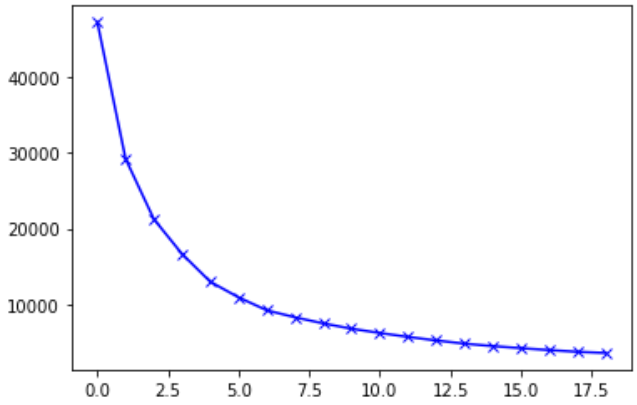
BUSINESS INFERENCE

- Significant people from 0-15k have made the payments, there is significant decrease in 15-20k and the least in 20-30k
- taking cash advance is not a sign of crossing.
- look out for customers with purchases from 0-10k as they tend to go over limit
- people with credit limit between 0-10k, and tenure of 12 months, are more comfortable in paying regularly
- people with credit limit between 0-5k have mostly crossed.

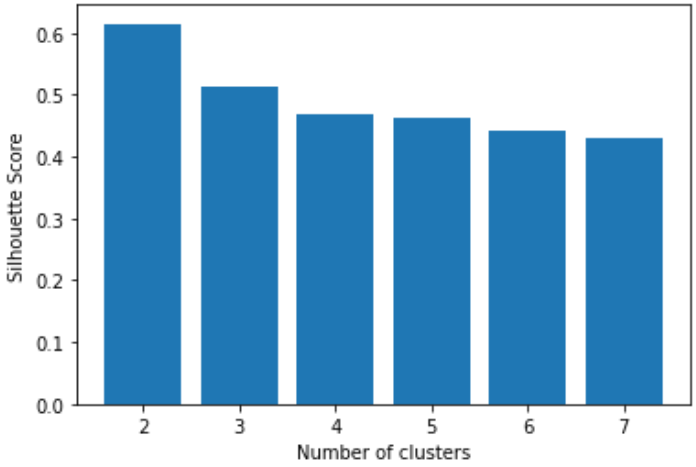
4. Feature Engineering Selection/ Creation:

- Feature Engineering during EDA
- We divided the data set into customers who have crossed their credit limit and those who have not crossed their credit limit, by creating a new column called CREDIT_STATUS.
- The feature engineering and the new features that were created are. "TOT_TRANSACTION"="PURCHASES"]+"CASH_ADVANCE"
- "TOT_TRANSACTION" column was created by adding the PURCHASES AND CASH ADVANCE.
- Total number of transactions column was created by adding the CASH_ADVANCE_TRX & PURCHASES_TRX columns
  TOT_TRX=CASH_ADVANCE_TRX+PURCHASES_TRX

5. ELBOW CURVE

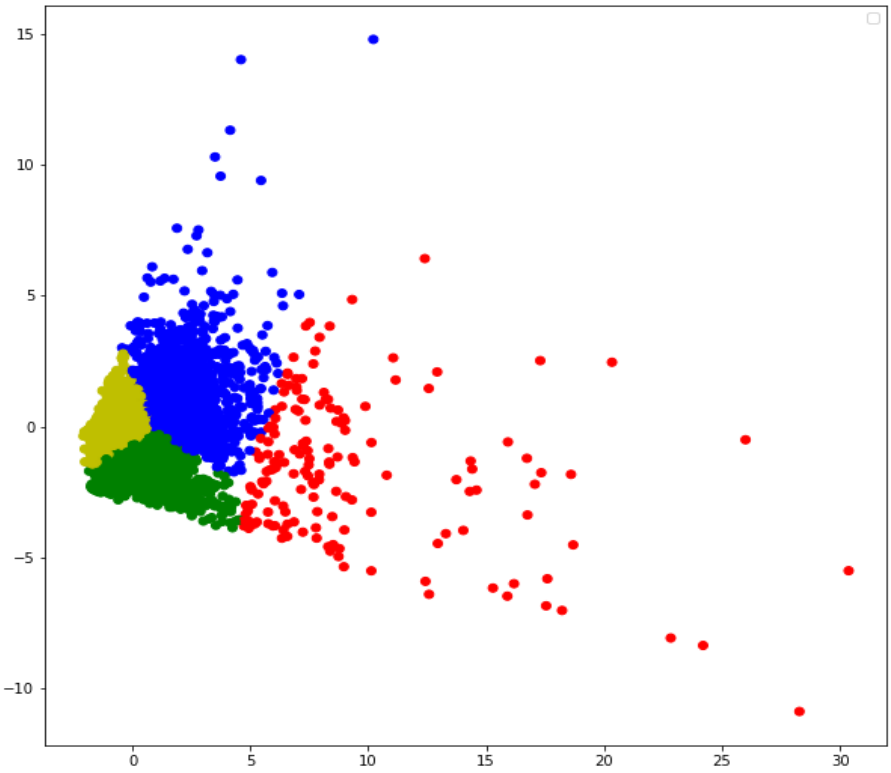


We See that in the in the elbow curve, we the slope of the curve starts changing after 4 or 5, hence we can consider any of these two numbers as optimal number of clusters.



We see that we get the best silhouette score at cluster 2.

- The approach we take to choose the cluster is through silhouette score and Elbow curve, even though the silhouette score is highest for 2, to get better insights, we can go for 4 or 5 clusters.



- We see that we are getting 4 clusters with minimum noise and overlapping. So 4 is the optimum number of clusters.

INSIGHTS AND RECOMMENDATIONS:

Cluster-0 : these are the customers who have a *high credit limit*, and tend to cross their credit limits, but they also make significant payments to their account, they usually spend for one off purchases.
Hence, they are high spenders, and they also maintain good balance compared to other clusters. But we must be careful as they cross their credit limit.

Cluster-1: they are low spenders; they usually maintain low balance and their purchase and spending depends on their balance
Encourage these customers to spend more, through cross selling and offers

Cluster-2 : they are ideal customers, where their total transactions never crosses, the credit limit, and they payment is also upto the mark, they usually spend on both one off purchases and installments,
They must be retained. They are an ideal group of customers, and we must do dome promotional activities so that they spend more

CLUSTER-3 : these are customers, who maintain good balance and their transaction is usually proportion to their balance, they don't cross the credit limit, and they have returned significant portion of their, credit value.

They usually pay their credit amount in much lesser transactions than other customers. Hence, they are also an ideal set of customers, we must reach out to, so that they can do more transactions.

Conclusion of Insights

- **Cluster 0 :** These customers are low spenders, despite having a decent credit limit. But maintain very low balance. They usage of credit card is very minimal.

They must encourage to transact more.

- **Cluster 1 :** This cluster does the maximum transactions with the credit card,they dont cross their limit and also make payments to the bank.They take very less cash avance.

This is a very good cluster of customers, and they must be given offers and incentives, to retain them.

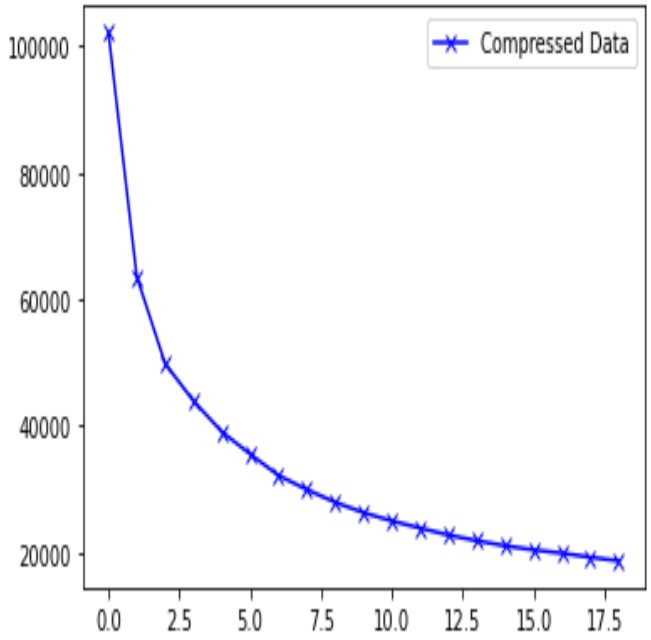
- **Cluster 2 :** These customers mostly take cash advance, and do not use their credit card regularly. They maintain good balance

They must be encouraged to use their card more frequently

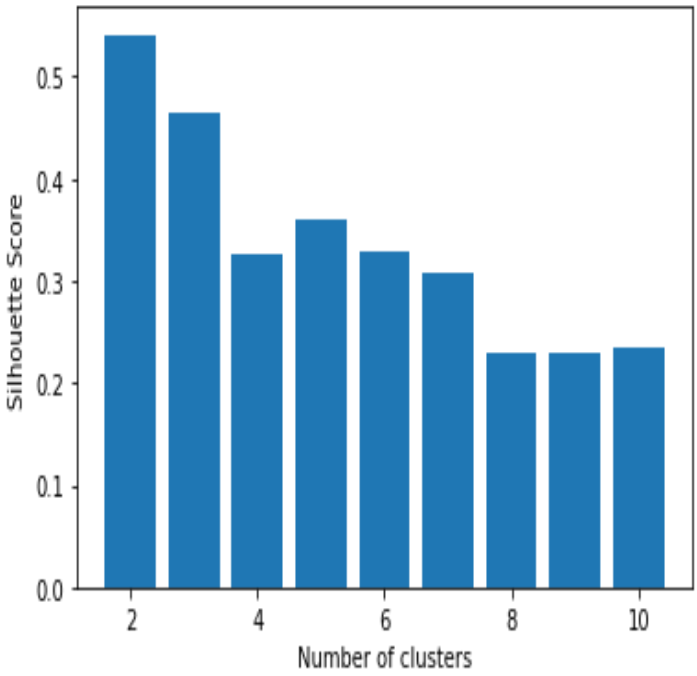
- **Cluster 3 :** These customers maintain very low balance, and all their transactions are proportional to their balance, despite a good credit limit.

They must be encouraged to spend more, and must be targeted using offers and promotions.

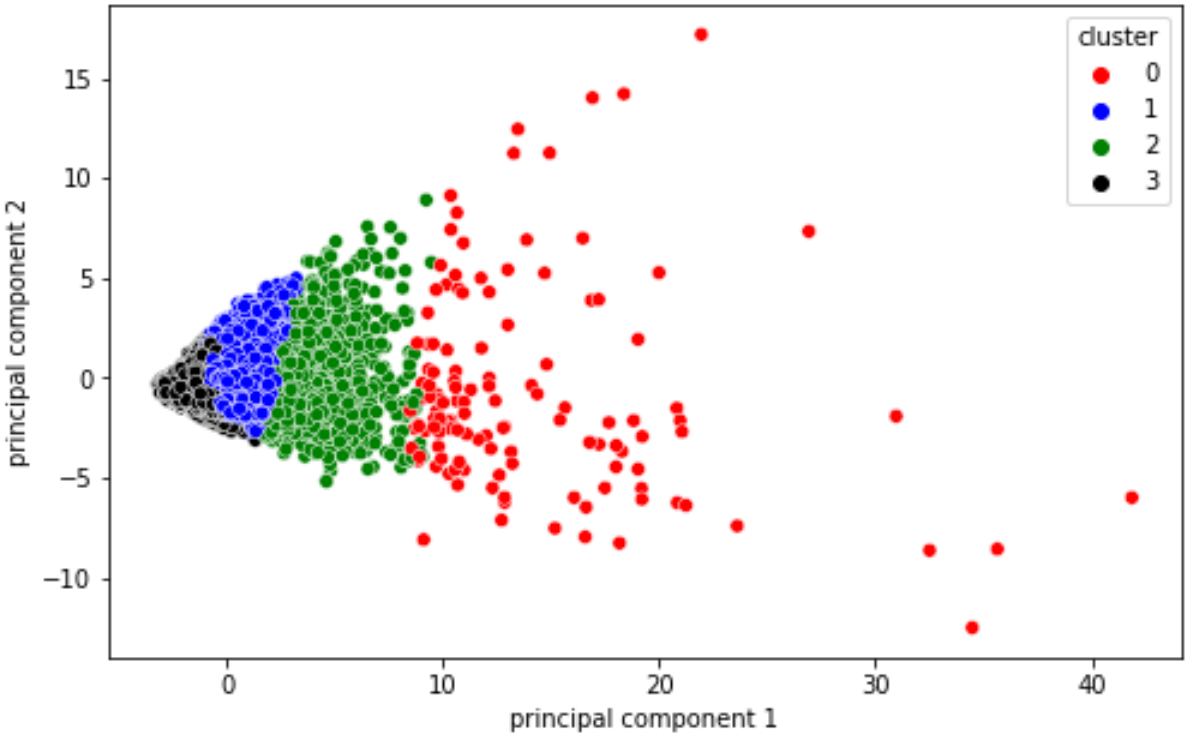
6. K- Means Algorithm with Autoencoder:



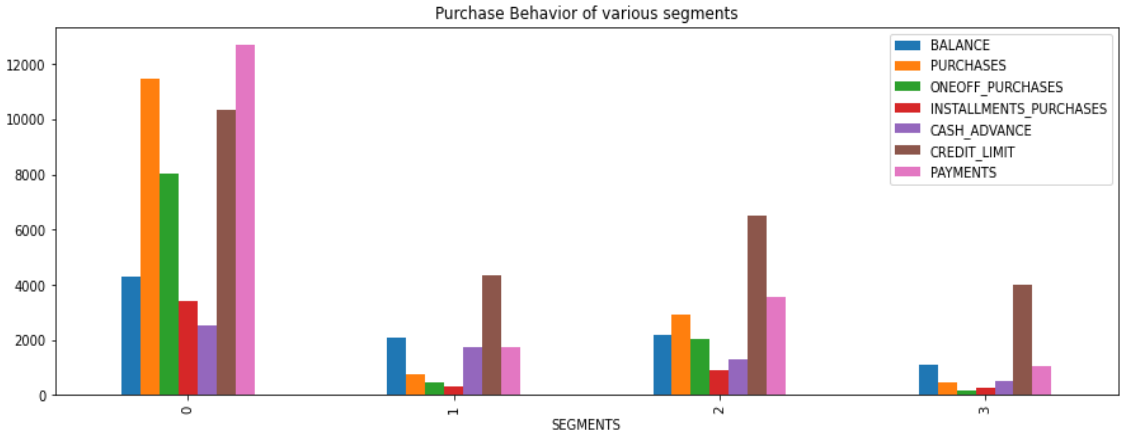
We See that in the in the elbow curve, we the slope of the curve starts changing after 4 or 5, hence we can consider any of these two numbers as optimal number of clusters.



We see that we get the best silhouette score at cluster 2.



- The approach we take to choose the cluster is through silhouette score and Elbow curve, even though the silhouette score is highest for 2, to get better insights, we can go for 4 or 5 clusters.



Conclusion

- **Cluster 0 :** These customers have the highest credit limit compared to others, and they also tend to make transactions, beyond their limit. But, they have also payd all their credit amount as seen from th graph.

They are good customers as they spend more, use credit card regularly for one-off and installment purchases and also pay the credit amount back even though they cross their credit limit. And they do not take much cash advance.

- **Cluster 1 :** These customers do very less transaction with their credit card, and mostly prefer cash advance, but they maintain good balance in their account, despite the cash advance, and also make significant payments. But they do not cross the credit limit.

These customers need to be encouraged to transact more with their credit card,

- **Cluster 2 :** These customers have a high credit limit, and do not spend beyond their credit limit. They also do significant transactions with the credit card. And take very less cash advance.

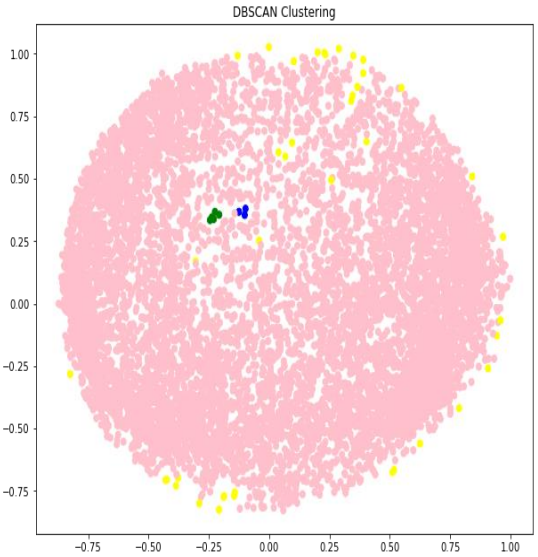
These customers are ideal customers and must be encouraged to transact more, as they do not make full use of their credit limit.

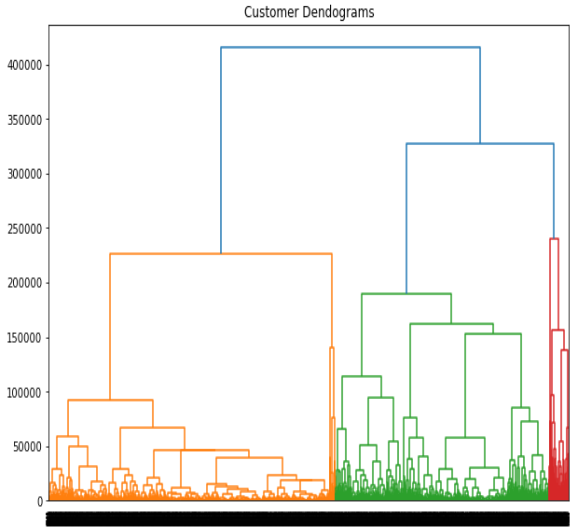
- **Cluster 3 :** These customers are low spenders, despite having a good credit limit. But maintain very low balance. They usage of credit card is very minimal.

they must be encouraged to transact more.

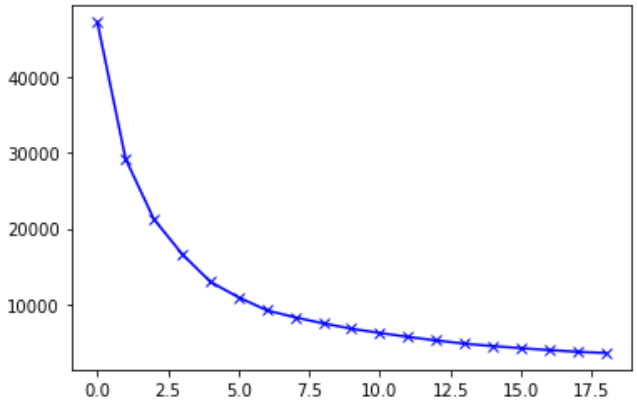
7. DBSCAN Clustering Algorithm:

Density-Based Clustering refers to unsupervised learning methods that identify distinctive groups/clusters in the data, based on the idea that a cluster in data space is a contiguous region of high point density, separated from other such clusters by contiguous regions of low point density. On trying to implement this, we are not able to distinguish between clusters and hence we try to implement Hierarchical clustering as well. We finally decide to develop using K means algorithm which gives us clear clusters.



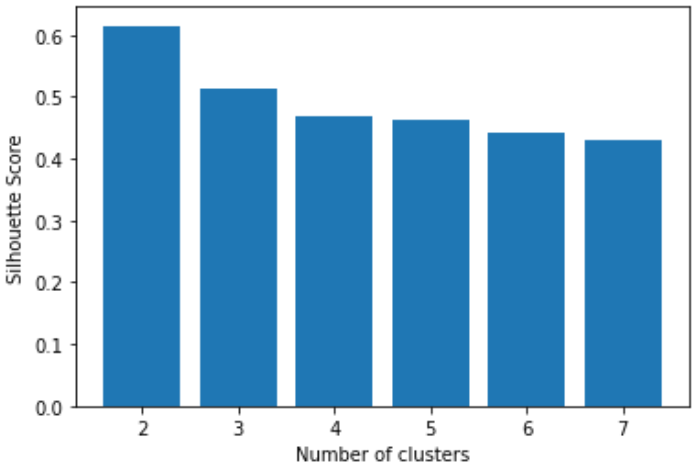


8. Elbow Method:



We See that in the in the elbow curve, we the slope of the curve starts changing after 4 , hence we can consider 4 as optimal number of clusters.

9. Silhouette Score



10. Business Recommendations & Insights:

CLUSTER-0 : THESE ARE THE CUSTOMERS WHO HAVE A *HIGH CREDIT LIMIT*, AND TEND TO CROSS THEIR CREDIT LIMITS, BUT THEY ALSO MAKE SIGNIFICANT PAYMENTS TO THEIR ACCOUNT, THEY USUALLY SPEND FOR ONE OFF PURCHASES.

HENCE THEY ARE HIGH SPENDERS, AND THEY ALSO MAINTAIN GOOD BALANCE COMPARED TO OTHER CLUSTERS. BUT WE MUST BE CAREFUL AS THEY CROSS THEIR CREDIT LIMIT.

CLUSTER-1:THEY ARE LOW SPENDERS, THEY USUSALLY MAINTAIN LOW BALANCE AND THEIR PURCHASE AND SPENDING MDEPENDS ON THEIR BALANCE
ENCOURAGE THESE CUSTOMERS TO SPEND MORE, THROUGH CROSS SELLING AND OFFERS

CLUSTER-2 : THEY ARE IDEAL CUSTOMERS, WHERE THEIR TOTAL TRANSACTIONS NEVER CROSSES, THE CREDIT LIMIT, AND THEY PAYMENT IS ALSO UPTO THE MARK, THEY USUALLY SPEND ON BOTH ONE OFF PURCHASES AND INSTALLMENTS,

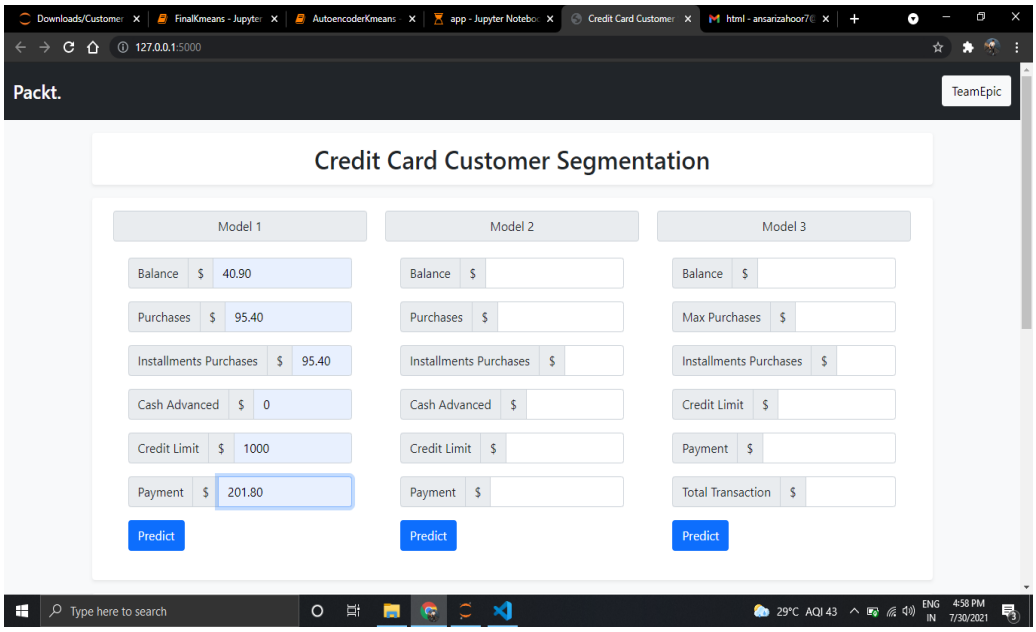
THEY MUST BE RETAINED. THEY ARE AN IDEAL GROUP OF CUSTOMERS, AND WE MUST DO DOME PROMOTIONAL ACTIVITIES SO THAT THEY SPEND MORE

CLUSTER-3 : THESE ARE CUSTOMERS, WHO MAINTAIN GOOD BALANCE AND THEIR TRANSACTION IS USUALLY PROPORTION TO THEIR BALANCE, THEY DONT CROSS THE CREDIT LIMIT, AND THEY HAVE RETURNED SIGNIFICANT PORTION OF THEIR, CREDIT VALUE.

THEY USUALLY PAY THEIR CREDIT AMOUNT IN MUCH LESSER TRANSACTIONS THAN OTHER CUSTOMERS. HENCE THEY ARE ALSO AN IDEAL SET OF CUSTOMERS, WE MUST REACH OUT TO, SO THAT THEY CAN DO MORE TRANSACTIONS.

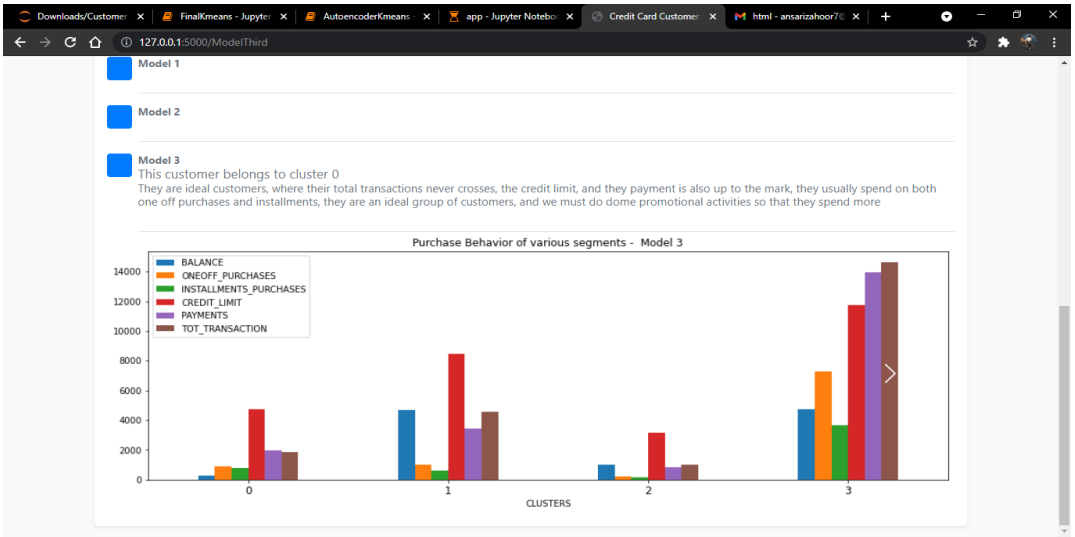
11. UI Development:

- In order to develop an interactive and smooth User Interface (UI) to enable the end user for input his/her own inputs and receive model insights, we have used web development frameworks such as **HTML**, **CSS** and **bootstrap**.
- Bootstrap is a powerful and extremely useful toolkit – a collection of HTML, CSS and JavaScript tools for creating and building web pages and applications.
- We incorporated three model along with user inputs and predict buttons for each of these models that are integrated with the backend using **Flask** Framework.
- Flask is a micro web framework written in python. It is classified as microframework because it doesn't require any particular tools and libraries.



12. Deployment & UI Testing:

- Flask API was used to deploy the machine learning model in backend. It is used to manage HTTP requests and uses API function to get the data and display the result to the end user in front end UI.
- User enters the inputs for all three models, post routing of those values, our machine learning model makes predictions and returns the same.
- As a final result, on inputting the user inputs, a backend call routed via Flask framework helps us in providing relevant insights and recommendations to the end user.


Downloads/Customer
FinalKmeans - Jupyter
AutoencoderKmeans
app - Jupyter Notebo
Credit Card Customer
html - ansarizahoor7
127.0.0.1:5000/ModelThird
Model 1
Model 2
Model 3
This customer belongs to cluster 0
They are ideal customers, where their total transactions never crosses, the credit limit, and they payment is also up to the mark, they usually spend on both one off purchases and installments, they are an ideal group of customers, and we must do dome promotional activities so that they spend more
Purchase Behavior of various segments - Model 3
BALANCE
ONEOFF_PURCHASES
INSTALLMENTS_PURCHASES
CREDIT_LIMIT
PAYMENTS
TOT_TRANSACTION
14000
12000
10000
8000
6000
4000
2000
0
0
1
2
3
CLUSTERS
Type here to search
29°C AQI 43
ENG
IN
5:00 PM
7/30/2021


Downloads/Customer
FinalKmeans - Jupyter
AutoencoderKmeans
app - Jupyter Notebo
Credit Card Customer
html - ansarizahoor7
127.0.0.1:5000/ModelSecond
Predictions
Model 1
Model 2
This customer belongs to cluster 3
These customers are low spenders, despite having a good credit limit. But maintain very low balance. They usage of credit card is very minimal. They must encourage to transact more.
Model 3
Purchase Behavior of various segments - Model 2
BALANCE
PURCHASES
INSTALLMENTS_PURCHASES
CASH_ADVANCE
CREDIT_LIMIT
PAYMENTS
10000
8000
6000
4000
2000
0
Type here to search
29°C AQI 43
ENG
IN
4:59 PM
7/30/2021


Downloads/Customer
FinalKmeans - Jupyter
AutoencoderKmeans
app - Jupyter Notebo
Credit Card Customer
html - ansarizahoor7
127.0.0.1:5000/predict
Predictions
Model 1
This customer belongs to cluster 1
These customers are low spenders, despite having a decent credit limit. But maintain very low balance. They usage of credit card is very minimal.They must encourage to transact more.
Model 2
Model 3
Purchase Behavior of various segments - Model 1
BALANCE
PURCHASES
INSTALLMENTS_PURCHASES
CASH_ADVANCE
CREDIT_LIMIT
PAYMENTS
10000
8000
6000
4000
2000
0
Type here to search
29°C AQI 43
ENG
IN
4:59 PM
7/30/2021

Technology Stack:

1. Jupyter : Open-source IDE used for development
2. Python : Programming language
3. Numpy : Python library for creating arrays
4. Pandas : python library for data manipulation and analysis
5. EDA Libraries : matplotlib, sea-born visualization libraries
6. Flask API : Web Framework for Python for making web apps & managing HTTP requests
7. HTML, CSS : website designing frameworks

Conclusion:

1. We were able to load the dataset onto Jupyter IDE and perform relevant analysis and data manipulation for deriving further insights.
2. Initial Data Quality measures were performed for making the data useful for further analysis.
3. Exploratory Data Analysis was performed in order to obtain visual information on how the data was impacting certain business requirements. Feature correlations were visualized and their distributions were plotted.
4. Post EDA, further data quality improvement measures such as Standardization, Normalization and dimensionality reduction techniques were implemented.
5. Multiple trials of various cluster formation techniques such as DBSCAN clustering, Hierarchical clustering, K means clustering (with/without encoders, with/without PCA) were implemented in order to finalize best possible clustering algorithm to implement the model upon.
6. Cluster Evaluation techniques such as Elbow Score, silhouette method was used to determine the quality of each of these clusters.
7. Cluster visualization helped us determine the critical features and their insights to conclude on business recommendations.
8. UI development using HTML, FLASK API was done in order to create a suitable and interactive UI for determining the clusters based on end user's inputs

Future Scope:

1. The UI web page can be tweaked to introduce more user interactive and automatic features and make it easier for end user to enter inputs and get predictions.
2. Features like CASH_ADVANCE can be visualized more and new features can be derived in order to obtain more valuable insights.
3. More feature engineering in terms of interest rate.
4. Verbosity of insights and recommendations can be changed.