

# Existing and Attrite Customer Segmentation using Hybrid Model with Proposed Retention Strategies

## Problem Statement:

The problem of banks or credit card providers is to maintain profitability and good customer relationships. Thus, the key challenge faced by credit card providers is customer attrition. It refers to when a customer decides to close their accounts or reduce their card usage. These behaviors indicate loss to the business and further contribute to full account terminations. The factors that result in low engagement and a decline in customer activities or satisfaction need to be studied to mitigate potential loss. Credit card providers can discover the early signs or underlying factors contributing to customer attrition, such as high revolving balances, limited card benefits, and many more. Segmentation of customers into existing and attrite customers is required to tailor retention strategies to cater to these two segments. Furthermore, segmenting the customers based on their characteristics can help the business identify at-risk and high-value customers so that more resources can be allocated to the segments. Proactive measures are needed to retain them and to prevent future loss. Thus, personalized retention strategies could be made to improve the retention rate and increase card usage.

## Objectives:

1. To identify the characteristics of the attrite and existing customers by conducting customer profile segmentation.
2. To evaluate the effectiveness of customer segmentation using a hybrid model (clustering + decision tree) on segregating new customers into distinct segments.
3. To develop retention strategies for existing and about-to-churn customers by providing personalized intervention to improve retention.

## Scope:

The dataset of this assignment is from Kaggle with a total of 22 variables with 10127 rows. This study will not consider predicting the churning status of a customer, instead, the difference between attrited and existing customers will be analyzed by segmentation. The dataset includes a range of variables that include customer demographic, financial, and behavioral dimensions. Some of the demographic variables such as customer age, gender, and education level; financial and credit usage such as credit limit, and total\_revolving\_balance determine whether a higher utilization ratio or lower credit limit results in churn. The customer engagement and activity variables such as months inactive, total relationship count, and change in transactions help in detecting early signs of attrition. The study will explore the misclassification rate of the predictive models and seek to develop targeted retention strategies based on the segmentation results.

## Link to the dataset:

<https://www.kaggle.com/datasets/thedevastator/predicting-credit-card-customer-attrition-with-m>

## Methodology:

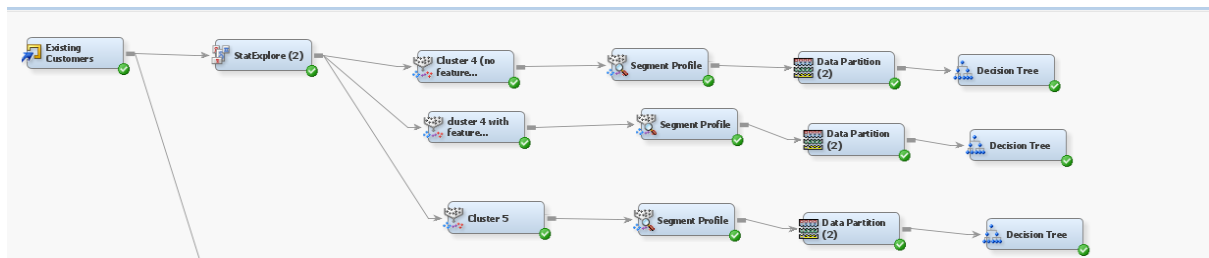
The process starts when the customers are segregated into two distinct groups which are existing customers and attrite customers. The data are loaded into SAS Enterprise Miner. After that, the number of clusters is determined with the clustering algorithm. The customers will be clustered into their representative clusters based on similar characteristics. The statistics of the existing and churned customers are explored using the StatsExplore. A data pipeline is built by connecting the clusters to a data partition (70% testing and 30% validation) and a decision tree. The decision tree is trained on the cluster labels and the features used for clustering. Verification of which clusters give the lowest misclassification rate on the training and validation data is performed. This is to examine the model effectiveness of clustering based on feature splits on the trained model in segmenting new data.

After the appropriate clusters are selected, data preparation and cleaning are performed. Some data preparation steps involved one-hot encoding by converting categorical variables to dummy variables and standardization of variables. Feature selection is also conducted to identify important features and reduce model complexity. Then, the misclassification rate after data cleaning is performed to check if the misclassification rate has improved.

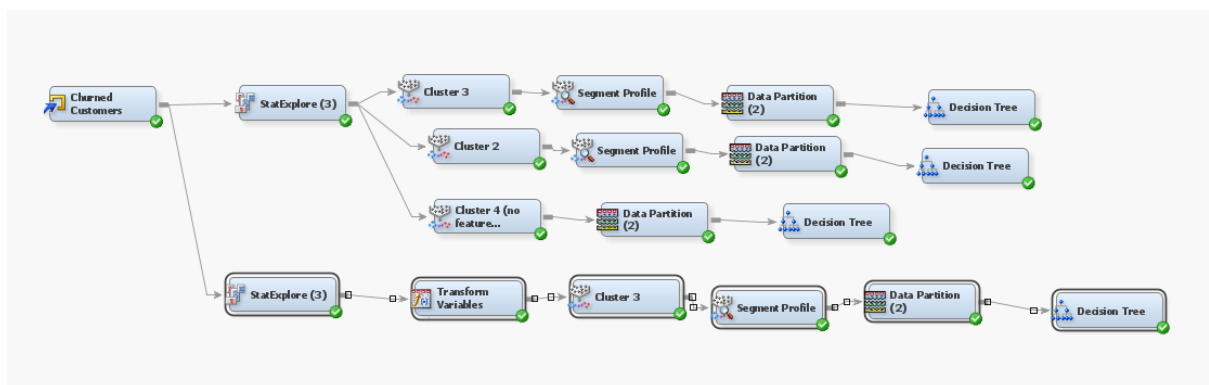
Finally, the cluster profile segmentation examines the key difference between existing customer segments and churned customer segments. Actionable insights on customer retention strategies can be proposed based on the result of the analysis.

# SAS EM Process Flow Diagram

For Existing Customer



For Churned Customer:



# Metadata

Variable Name	Description	Data Types
CLIENTNUM	Unique identifier for each customer.	Integer (Numerical)
Attrition_Flag	Indication of churning	Boolean
Customer_Age	Age of customer.	Integer (Numerical)
Gender	Gender of the customer.	String (Categorical)
Dependent_count	Number of dependents that the customer has.	Integer (Numerical)
Education_Level	Education level of customer.	String (Categorical)
Marital_Status	Marital status of customer.	String (Categorical)
Income_Category	Income category of customer.	String (Categorical)
Card_Category	Type of card held by the customer.	String (Categorical)
Months_on_book	The duration (months) of the customer has been using a credit card since they first opened the account.	Integer (Numerical)
Total_Relationship_Count	Total number of relationships the customer has with the credit card provider.	Integer (Numerical)
Months_Inactive_12_mon	Number of months the customer has been inactive in the last twelve months.	Integer (Numerical)
Contacts_Count_12_mon	Number of contacts the customer has had in the last twelve months.	Integer (Numerical)
Credit_Limit	Credit limit of customer.	Integer (Numerical)
Total_Revolving_Bal	The total amount of credit card debt across all accounts carried from one billing period to the next.	Integer (Numerical)
Avg_Open_To_Buy	Available credit a customer has on their revolving accounts	Integer (Numerical)
Total_Amt_Chng_Q4_Q1	The total amount changed from quarter 4 to quarter 1.	Integer (Numerical)
Total_Trans_Amt	Total transaction amount.	Integer (Numerical)
Total_Trans_Ct	Total transaction count.	Integer (Numerical)
Total_Ct_Chng_Q4_Q1	The difference in the total number of transactions between Q4 and next year's Q1	Integer (Numerical)
Avg_Utilization_Ratio	Average utilization ratio of customers.	Integer (Numerical)
Naive_Bayes_Classifier_Attrition_Flag	Naive Bayes classifier for attrition prediction	Boolean

Name	Use	Report	Role	Level
Attrition_Flag	Default	No	Rejected	Nominal
Avg_Open_To	Default	No	Input	Interval
Avg_Utilization	Default	No	Input	Interval
CLIENTNUM	Default	No	ID	Interval
Card_Category	Default	No	Input	Nominal
Contacts_Count	Default	No	Input	Interval
Credit_Limit	Default	No	Input	Interval
Customer_Age	Default	No	Input	Interval
Dependent_count	Default	No	Input	Interval
Education_Level	Default	No	Input	Nominal
Gender	Default	No	Input	Nominal
Income_Category	Default	No	Input	Nominal
Marital_Status	Default	No	Input	Nominal
Months_Inactive_12_mon	Default	No	Input	Interval
Months_on_book	Default	No	Input	Interval
Naive_Bayes_Classifier	Default	No	Rejected	Interval
Total_Amt_Chng_Q4_Q1	Default	No	Input	Interval
Total_Ct_Chng_Q4_Q1	Default	No	Input	Interval
Total_Relationship_Count	Default	No	Input	Interval
Total_Revolving_Bal	Default	No	Input	Interval
Total_Trans_Amt	Default	No	Input	Interval
Total_Trans_Ct	Default	No	Input	Interval

There are a total of 5 categorical variables, 15 continuous variables, and 2 two variables with Boolean values (1/0). The figure above shows the attrition\_flag and Naïve\_Bayes\_Classifier variables have been omitted (Role: Rejected) since the focus of the study is segmentation. The scale levels such as intervals and nominal variables are depicted in the figure above.

## Data Preparation

### The Customer Profile Overview – Without Differentiation of Churned and Existing Customers

Data Role	Variable Name	Role	Number of Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage
TRAIN	Card_Category	INPUT	4	0	Blue	93.18	Silver	5.48
TRAIN	Education_Level	INPUT	7	0	Graduate	30.89	High School	19.88
TRAIN	Gender	INPUT	2	0	F	52.91	M	47.09
TRAIN	Income_Category	INPUT	6	0	Less than \$40K	35.16	\$40K - \$60K	17.68
TRAIN	Marital_Status	INPUT	4	0	Married	46.28	Single	38.94
TRAIN	Attrition_Flag	TARGET	2	0	Existing Customer	83.93	Attrited Customer	16.07

Based on the output, the categorical variables have no missing values. Thus, imputation is not required.

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Avg_Open_To_Buy	INPUT	7469.14	9090.685	10127	0	3	3474	34516	1.661697	1.798617
Avg_Utilization_Ratio	INPUT	0.274894	0.275691	10127	0	0	0.176	0.999	0.718008	-0.79497
Contacts_Count_12_mon	INPUT	2.455317	1.106225	10127	0	0	2	6	0.011006	0.000863
Credit_Limit	INPUT	8631.954	9088.777	10127	0	1438.3	4549	34516	1.666726	1.808989
Customer_Age	INPUT	46.32596	8.016814	10127	0	26	46	73	-0.03361	-0.28862
Dependent_count	INPUT	2.346203	1.298908	10127	0	0	2	5	-0.02083	-0.68302
Months_Inactive_12_mon	INPUT	2.341167	1.010622	10127	0	0	2	6	0.633061	1.098523
Months_on_book	INPUT	35.92841	7.986416	10127	0	13	36	56	-0.10657	0.4001
Total_Amt_Chng_Q4_Q1	INPUT	0.759941	0.219207	10127	0	0	0.736	3.397	1.732063	9.993501
Total_Ct_Chng_Q4_Q1	INPUT	0.712222	0.238086	10127	0	0	0.702	3.714	2.064031	15.68929
Total_Relationship_Count	INPUT	3.81258	1.554408	10127	0	1	4	6	-0.16245	-1.00613
Total_Revolving_Bal	INPUT	1162.814	814.9873	10127	0	0	1276	2517	-0.14884	-1.14599
Total_Trans_Amt	INPUT	4404.086	3397.129	10127	0	510	3899	18484	2.041003	3.894023
Total_Trans_Ct	INPUT	64.85869	23.47257	10127	0	10	67	139	0.153673	-0.36716

There are no missing values reported in the numerical variables, thus imputation is not needed for the variables. The figure above shows that the total count change from Q4 to Q1 has a

skewness value of 2.06, and the Total Transaction Amount variable has a skewness value of 2.04. Both variables are highly skewed, but transformation using a log algorithm is needed as it might distort the clustering. However, some variables are naturally skewed for instance income, credit limit, or transaction amount. Skewness often reflects the underlying structure of the data though keeping the original scale can preserve its interpretability. However, the transformation of the variables might yield more distinguishable clusters by its distance measure in K-Means Clustering.

Furthermore, all the categorical need to be converted to dummy variables for clustering. This is because K-Means clustering works well with numerical data but struggles with categorical data, especially nominal variables. Thus, converting to dummy variables can allow the distance to be computed properly and treat the categories equally.

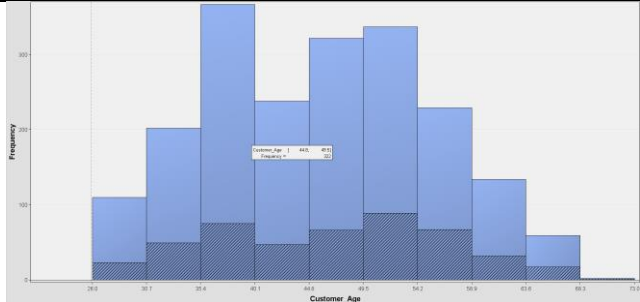
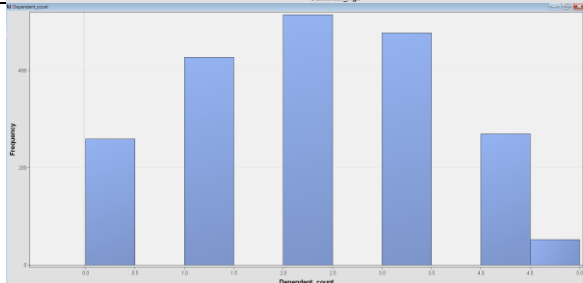
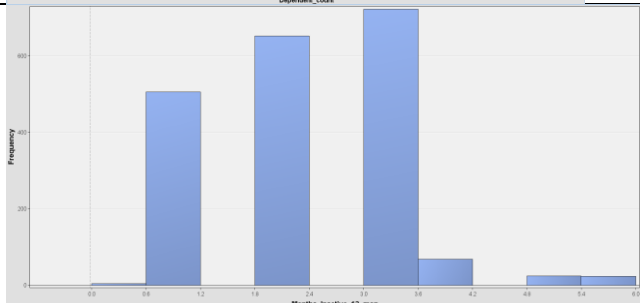
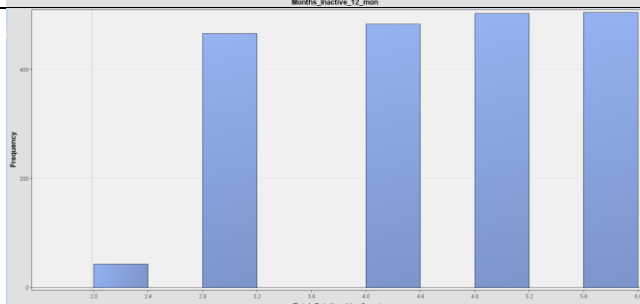
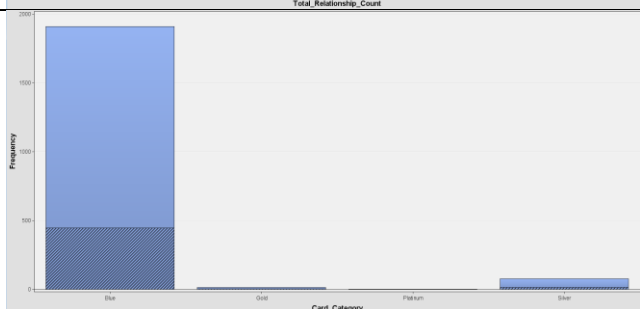
## EDA and Data Preparation:

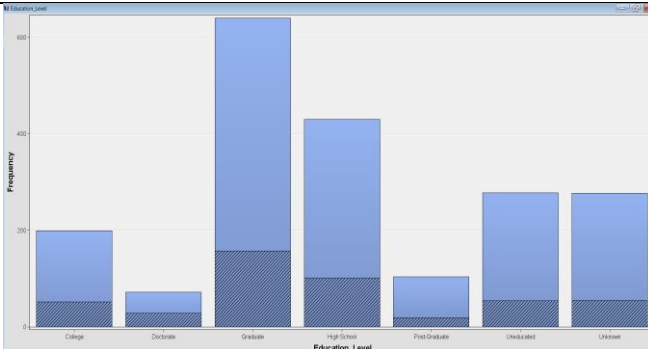
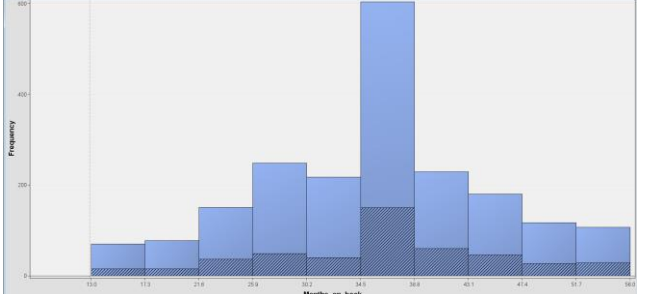
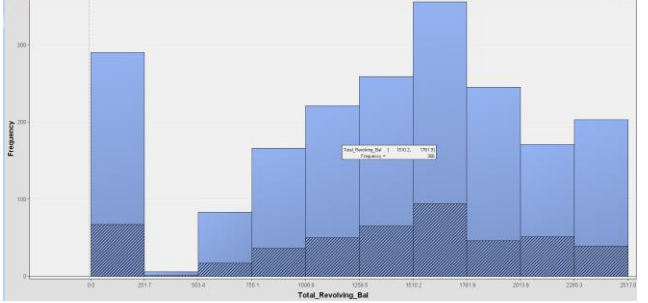
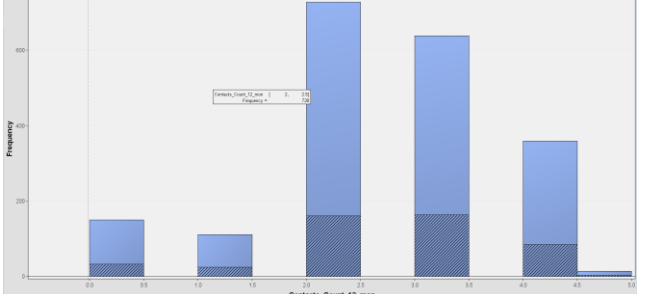

### Existing Customer Profiles

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
Avg_Open_To_Buy	INPUT	7470.273	9087.672	8500	0	15	3469	34516	1.635478	1.707629
Avg_Utilization_Ratio	INPUT	0.296412	0.272568	8500	0	0	0.211	0.994	0.599613	-0.94528
Contacts_Count_12_mon	INPUT	2.356353	1.081436	8500	0	0	2	5	-0.09898	-0.34297
Credit_Limit	INPUT	8726.878	9084.97	8500	0	1438.3	4642	34516	1.642397	1.727134
Customer_Age	INPUT	46.26212	8.081157	8500	0	26	46	73	-0.03023	-0.28881
Dependent_count	INPUT	2.335412	1.303229	8500	0	0	2	5	-0.00449	-0.69199
Months_Inactive_12_mon	INPUT	2.273765	1.016741	8500	0	0	2	6	0.726324	1.17041
Months_on_book	INPUT	35.88059	8.02181	8500	0	13	36	56	-0.10348	0.39133
Total_Amt_Chng_Q4_Q1	INPUT	0.77251	0.217783	8500	0	0.256	0.743	3.397	2.144442	11.69811
Total_Ct_Chng_Q4_Q1	INPUT	0.742434	0.228054	8500	0	0.028	0.721	3.714	2.65745	20.34624
Total_Relationship_Count	INPUT	3.914588	1.528949	8500	0	1	4	6	-0.23839	-0.92975
Total_Revolving_Bal	INPUT	1256.604	757.7454	8500	0	0	1364	2517	-0.34518	-0.82814
Total_Trans_Amt	INPUT	4654.656	3512.773	8500	0	816	4100	18484	1.995948	3.482775
Total_Trans_Ct	INPUT	68.67259	22.91901	8500	0	11	71	139	-0.00364	-0.20165

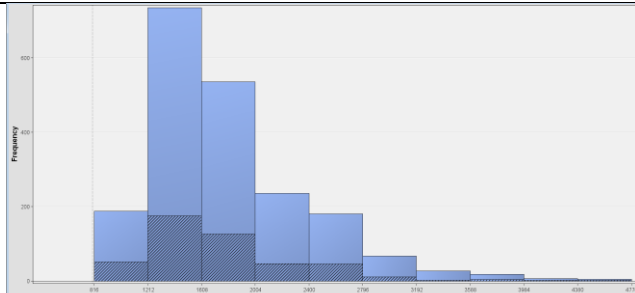
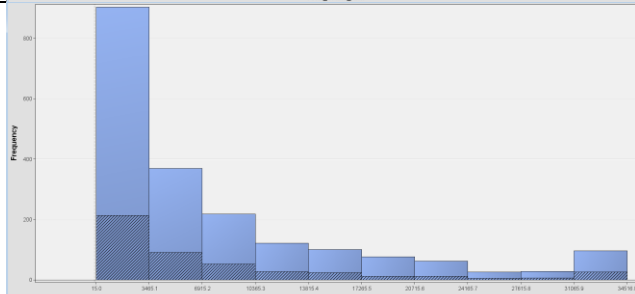
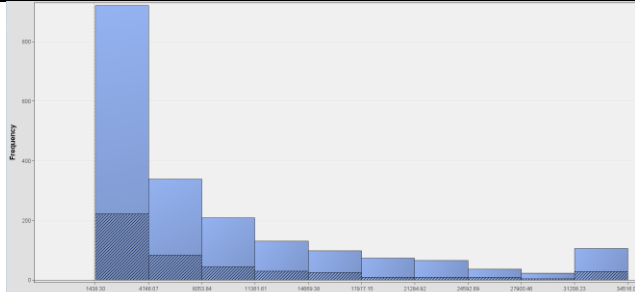
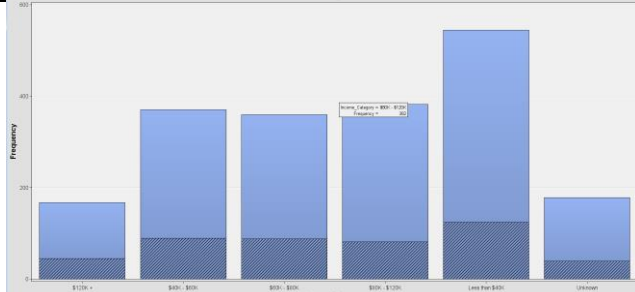
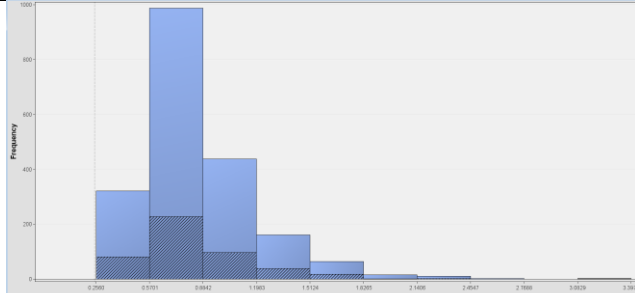
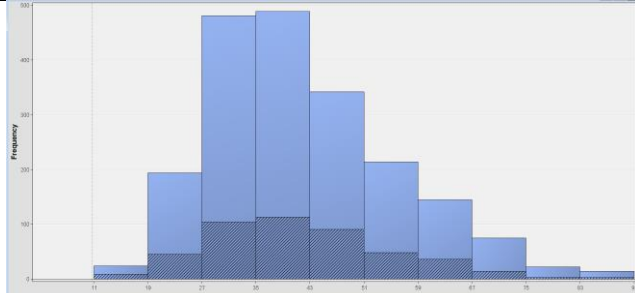
Based on the output above, the Total Transaction Amount Change (Q4 vs Q1) and Total Count Change from (Q4 vs Q1) variables have skewness values of 2.14 and 2.66 respectively. Thus, standardizing the variables is required. A comparison of the misclassification rate between log-transformed variables and non-normalized variables is conducted.

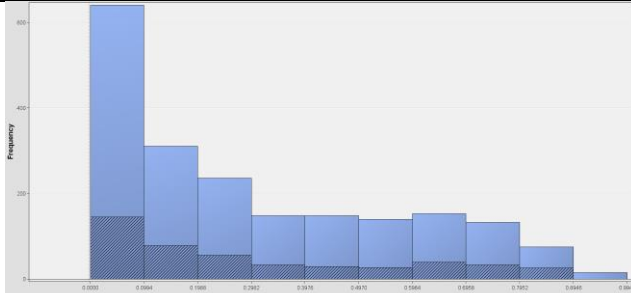
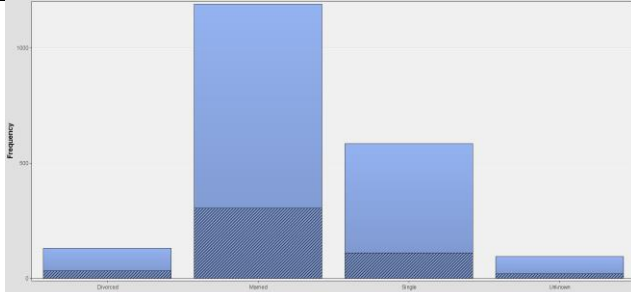
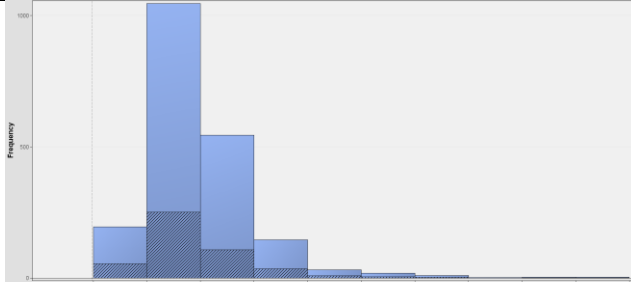
## EDA - Univariate Analysis

Variable Name	Exploration Results (Graphical or statistical)	Interpretations
Customer_Age		From the distribution diagram, the existing customers are mostly from 35 to 40 years old. The mean value is 46.32, the median value is 46 and the skewness value is -0.0336 which is normally distributed
Dependent_count		The dependent count ranges from 0 to 5. Most of the existing customers have 2 dependents. Most existing customer has 2 dependents, and it is normally distributed.
Months_Inactive_12_months		The month_inactive_12_months variable values range from 0 to 6 and it is normally distributed.
Total_Relationship_Count		The total relationship counts variable shows the range from 2 to 6 relationships, which majority of them have 4 to 6 relationships with providers.
Card_Category		Most of the existing customers have blue cards, and very few of them have gold or platinum cards.

Education_level		Most of the existing customers are graduate followed by high schools.
Months_on_book		The months on books have a skewness value of -0.103 which indicates normal distribution. The mean is 35.88 months, median is 36 months. Most of the existing customers have 34.5 to 38.8 months on book.
Total_Revolving_Balance		The total revolving balance of existing customers has a mean value of 1256.60. The median value is 1364. The skewness value of -0.345 indicates it is normally distributed. Most of them have a revolving balance between \$1510.20 to \$1761.90, and some customers have a balance of less than \$251.70.
Contact_Count_12_mon		The contact count variable has a skewness value of -0.09 which indicates normal distribution. Most of the existing customers have 2 to 3 contacts with service providers in a year.
Gender		Most of the existing customers are male, higher than female.



Total_Trans_Amt		<p>The total transaction amount shows that it has a skewness of 1.99. It is moderately skewed to the left. The median value is \$4100. The mean value is \$4654.65.</p>
Avg_Open_To_Buy		<p>The average open-to-buy ratio shows a skewness value of 1.64. The median value is \$3469, and the mean value is \$7470.27. Most of the customers have an average buy ratio of less than \$3465.10.</p>
Avg_Credit_Limit		<p>The average credit limit shows a skewness value of 1.64. The mean value is \$8726.87, and the median value is \$4642. Most of the customers have an average buy ratio of less than \$4746.07.</p>
Income_Category		<p>Most of the existing customer's income is less than \$40K.</p>
Total_Amt_Chng_Q4_Q1		<p>The total amount has a high skewness value of more than 2. The mean value is 0.772, and the median value is 0.743. Most of the customers have a total_amount_change between 0.57 to 0.88.</p>
Total_Trans_Ct		<p>The total transaction change has a skewness value of 0. The mean value is 68.67, and the median value is 71. Most of the customers have a total transaction count between 27 to 43.</p>

Avg_Utilization_Ratio		<p>The average utilization ratio has a skewness value of 0.60. The mean value is 0.29, and the median value is 0.21.</p>
Marital_status		<p>Most of the existing customers are married.</p>
Total_Ct_Chng_Q4_Q1		<p>The total transaction difference (Q4 to Q1) has a high skewness value of more than 2. The mean value is 0.742, and the median value is 0.721. Most of the customers have a total_transaction_change_Q4 to Q1 between 0.40 to 0.76.</p>

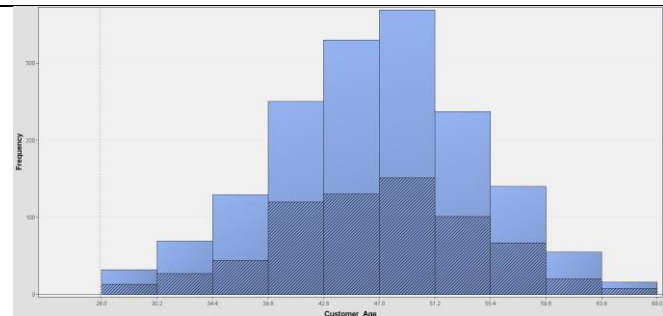

## Churned Customer Profiles

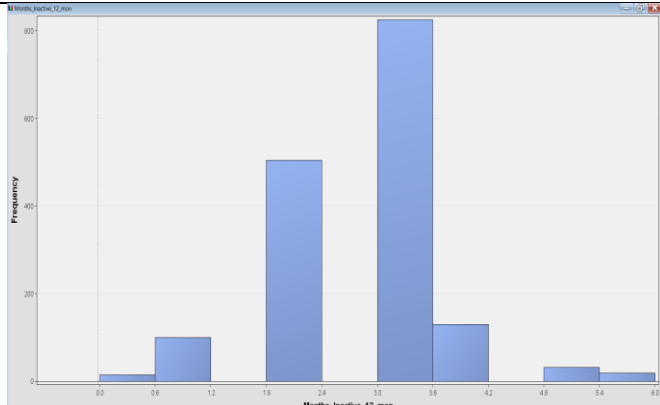
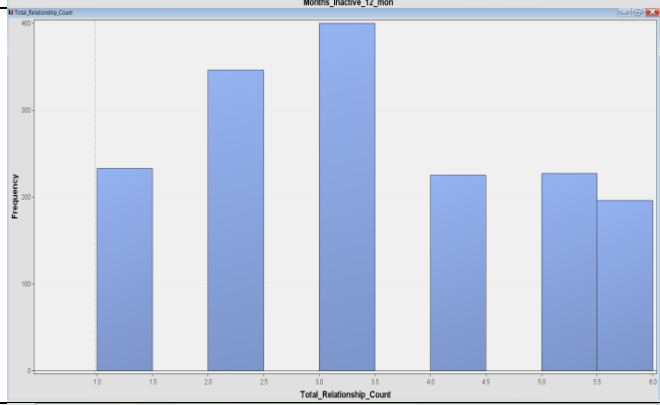
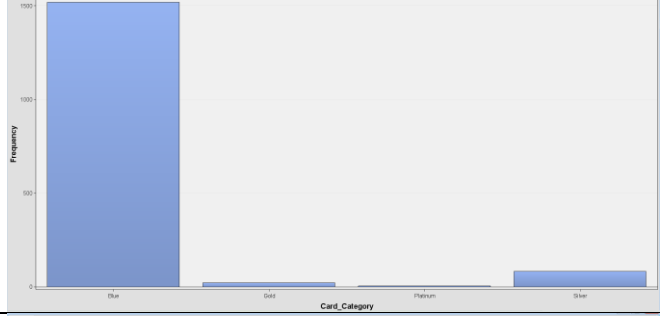
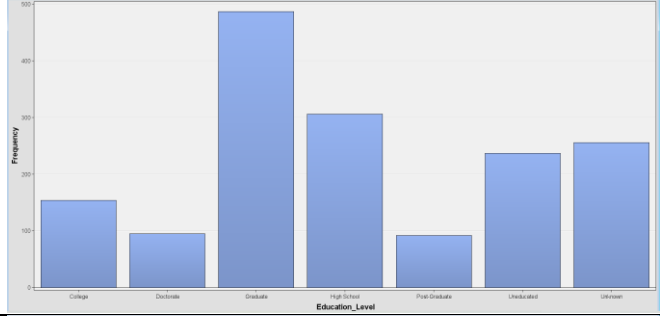
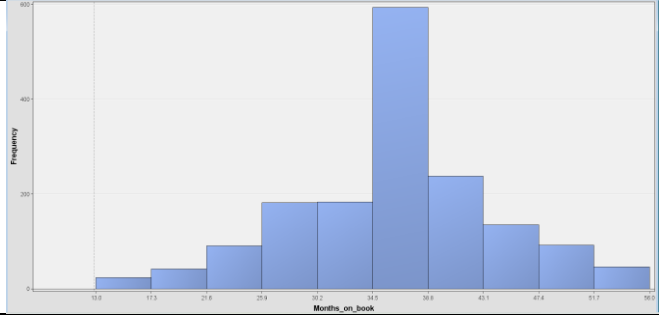
### Churned Customer Statistic

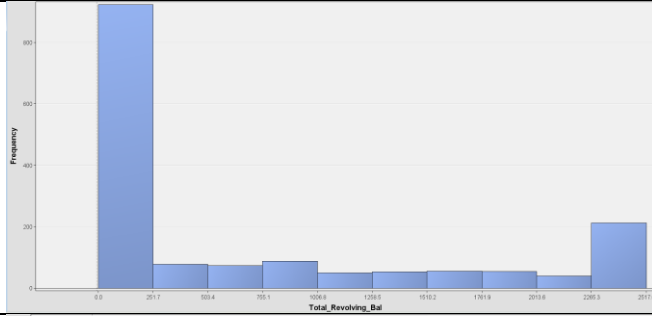
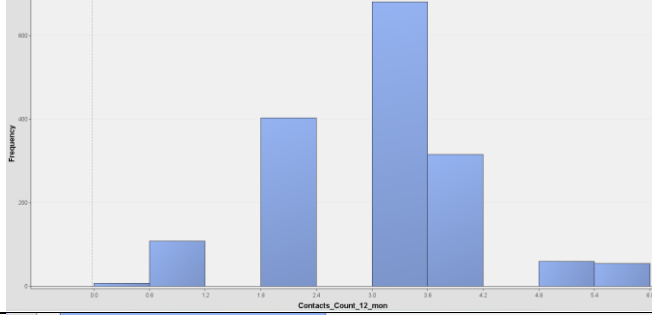
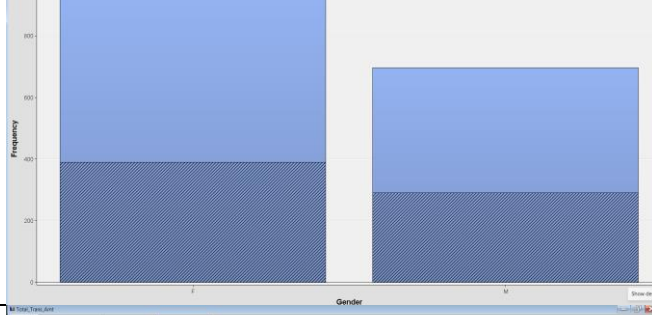
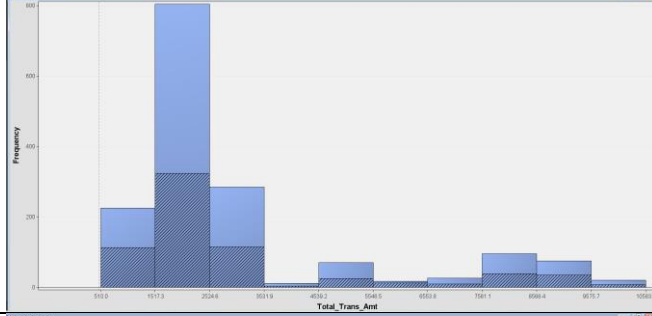
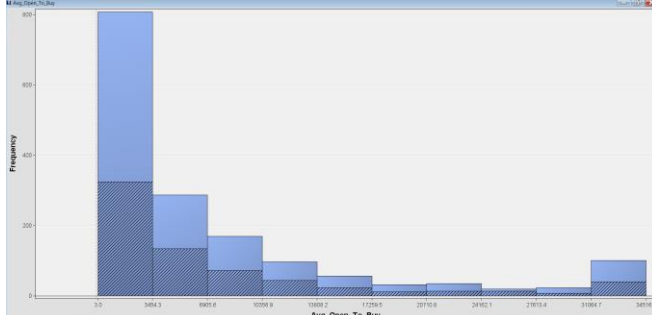
40				Number								
41	Data			of								
42	Role	Variable Name	Role	Levels	Missing	Mode	Mode Percentage	Mode2	Mode2 Percentage			
43												
44	TRAIN	Card_Category	INPUT	4	0	Blue	93.36	Silver	5.04			
45	TRAIN	Education_Level	INPUT	7	0	Graduate	29.93	High School	18.81			
46	TRAIN	Gender	INPUT	2	0	F	57.16	M	42.84			
47	TRAIN	Income_Category	INPUT	6	0	Less than \$40K	37.62	\$40K - \$60K	16.66			
48	TRAIN	Marital_Status	INPUT	4	0	Married	43.58	Single	41.06			
49												
50												
51												
52	Interval Variable Summary Statistics											
53	(maximum 500 observations printed)											
54												
55	Data Role=TRAIN											
56												
57					Standard	Non						
58	Variable	Role		Mean	Deviation	Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
59												
60	Avg_Open_To_Buy	INPUT		7463.216	9109.208	1627	0	3	3488	34516	1.79944	2.280736
61	Avg_Utilization_Ratio	INPUT		0.162475	0.264458	1627	0	0	0	0.999	1.63015	1.423134
62	Contacts_Count_12_mon	INPUT		2.972342	1.090537	1627	0	0	3	6	0.450797	0.672372
63	Credit_Limit	INPUT		8136.039	9095.334	1627	0	1438.3	4178	34516	1.804428	2.294341
64	Customer_Age	INPUT		46.6595	7.665652	1627	0	26	47	68	-0.03975	-0.31268
65	Dependent_count	INPUT		2.402581	1.27501	1627	0	0	2	5	-0.10624	-0.61894
66	Months_Inactive_12_mon	INPUT		2.693301	0.899623	1627	0	0	3	6	0.377828	1.981655
67	Months_on_book	INPUT		36.17824	7.796548	1627	0	13	36	56	-0.11867	0.448711
68	Total_Amt_Chng_Q4_Q1	INPUT		0.694277	0.214924	1627	0	0	0.701	1.492	-0.21522	-0.09221
69	Total_Ct_Chng_Q4_Q1	INPUT		0.554386	0.226854	1627	0	0	0.531	2.5	1.050356	5.306908
70	Total_Relationship_Count	INPUT		3.279656	1.577782	1627	0	1	3	6	0.265179	-1.01283
71	Total_Revolving_Bal	INPUT		672.823	921.3856	1627	0	0	0	2517	1.024055	-0.53447
72	Total_Trans_Amt	INPUT		3095.026	2308.228	1627	0	510	2329	10583	1.685336	1.653971
73	Total_Trans_Ct	INPUT		44.93362	14.56843	1627	0	10	43	94	0.485945	0.570805

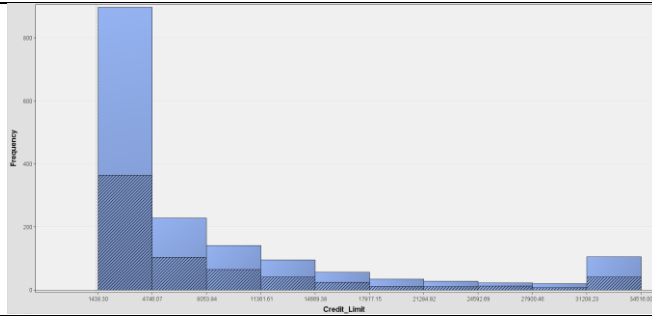
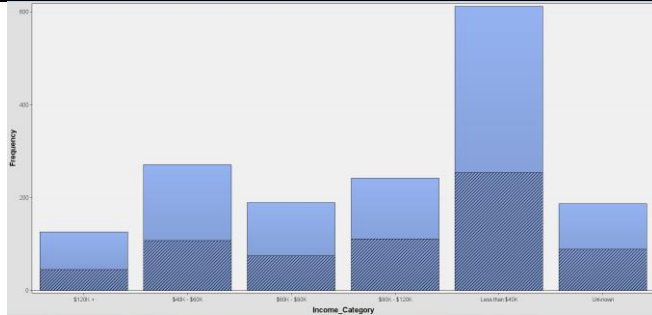
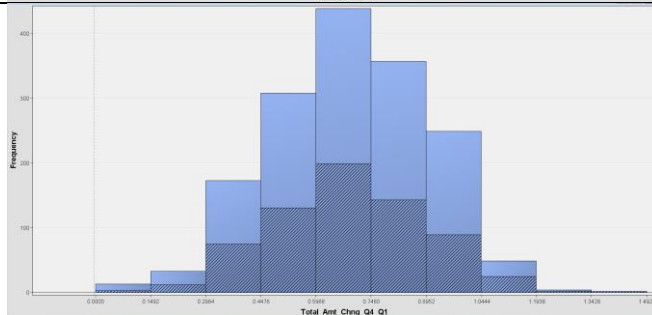
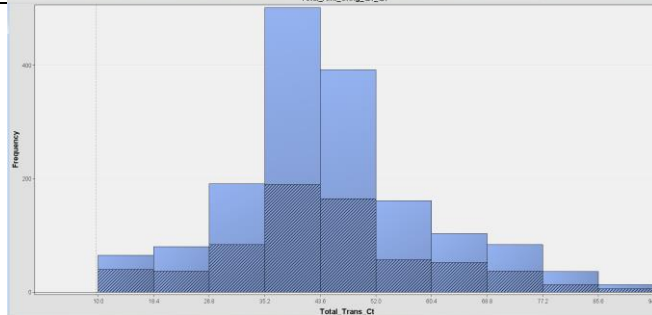
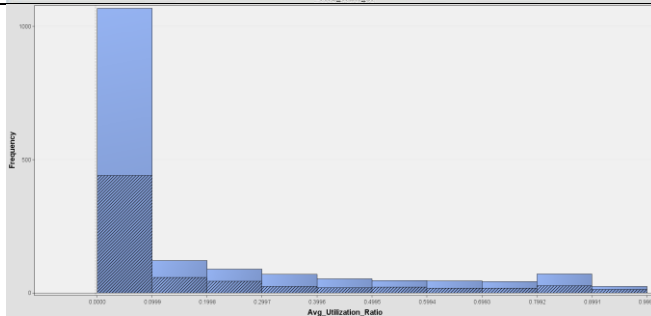
Based on the figure above, the skewness values of the variables such as average open to buy, average utilization ratio, credit limit, and total revolving balance are between 1.60 to 1.80. These variables are moderately skewed. There are total of 1627 churned customers with no missing values.

### EDA -For Churned Customers

Variable Name	Exploration Results (Graphical or statistical)	Interpretations
Customer_Age		From the distribution diagram, the existing customers are mostly from 42 to 51 years old. The mean value is 46.66, the median value is 47 and the skewness value is -0.040 which is normally distributed
Dependent_count		The dependent count ranges from 0 to 5. Most of the churned customers have 3 dependents and it is normally distributed.

Months_Inactive_12_months		The month_inactive variable values range from 0 to 6. Most of the churned customers have 3 or more months of inactive. The skewness value is 0.38, and it is normally distributed.
Total_Relationship_Count		The total relationship counts variable shows the range from 1 to 6 relationships, which majority of them have 3 relationships with providers. It is also normally distributed.
Card_Category		The churned customers have blue cards, and very few of them have gold or platinum cards.
Education_level		Most of the churned customers are graduates followed by high schools.
Months_on_book		The months on books have a skewness value of -0.12 which indicates normal distribution. Most of the churned customers have 34.5 to 38.8 months on the book. The mean and median values are 36 months.

Total_Revolving_Balance		<p>The total revolving balance of existing customers has a mean value of \$672.82. The median value is \$0. The skewness value of 1.02. Many customers have a balance of less than \$251.70.</p>
Contact_Count_12_mon		<p>The contact count variable has a skewness value of 0.45 which indicates normal distribution. Most of the churned customers have 3 to 4 contacts with service providers in a year.</p>
Gender		<p>Most of the existing customers are female, significantly higher than male.</p>
Total_Trans_Amt		<p>The total transaction amount shows that it has a skewness of 1.68. The median value is \$2329. The mean value is \$3095.02.</p>
Avg_Open_To_Buy		<p>The average open-to-buy ratio shows a skewness value of 1.80. The median value is \$3488, and the mean value is \$7463.22. Most of the customers have an average buy ratio of less than \$3454.30.</p>

Credit_Limit		<p>The average credit limit shows a skewness value of 1.80. The mean value is \$8136.04, and the median value is \$4178. Most of the customers have an average buy ratio of less than \$4746.07.</p>
Income_Category		<p>Most of the existing customer's income is less than \$40K.</p>
Total_Amt_Chng_Q4_Q1		<p>The total amount change (Q4 to Q1) has a skewness value of -0.22. The mean value is 0.69, and the median value is 0.71. Most of the customers have a total_amount_change between 0.57 to 0.75.</p>
Total_Trans_Ct		<p>The total transaction change has a skewness value of 0.49. The mean value is 44.94, and the median value is 43. Most of the customers have a total_transaction_change between 35 to 52.</p>
Avg_Utilization_Ratio		<p>The average utilization ratio has a skewness value of 1.63. The mean value is 0.16, and the median value is 0.</p>

Marital_status		Most of the churned customers are married, however, the number of single customers is also significantly higher than that of existing customers.
Total_Ct_Chng_Q4_Q1		The total transaction difference (Q4 to Q1) has a skewness value of 1.05. The mean value is 0.55, and the median value is 0.53. Most of the customers have a total_transaction_change_Q4 to Q1 between 0.25 to 0.75.

## Findings based on EDA

The churned customer has a higher age, and higher months of inactive, the total revolving balance is half the revolving balance of existing customers. Furthermore, the churned customers have more females than existing customers. The total transaction amount is also significantly lower than that of the existing customers. The churned customers also have lower credit limits and average utilization ratios. The total transaction changes and total amount difference (Q4 to Q1) are also lower than that of existing customers.

For the subsequent steps, the number of clusters will be determined by comparison of different cluster numbers that yield the lowest misclassification rate of validation data. Next, the appropriate cluster number is chosen for data preparation – conversion of categorical to dummy variable and standardization for the skewed variables.

## Data Preparation & Model Construction, Optimization, and Validation.

### Step 1: Determination of Number of Clusters

#### Data partition

The data is split into 70% of the training data and 30% of the validation data shown in the figure below to verify the misclassification rate using the Decision Tree.

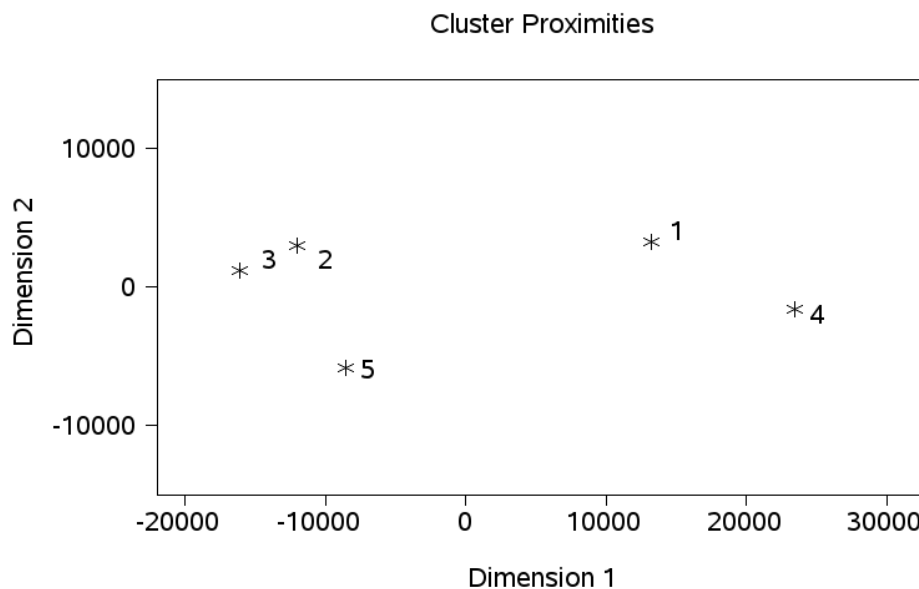
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

Table 1: Misclassification Rates of Training and Validation Datasets for 3, 4 and 5 Clusters

Cluster Number	Misclassification Rate
3 clusters	Training: 0.0735 Validation: 0.087
4 clusters	Training: 0.082 Validation: 0.082
5 clusters	Training: 0.076 Validation: 0.070

The testing of the performance of the clusters in segmenting customers into their respective groups based on cluster features. The focus is to find the model with the lowest misclassification rate to prove that the clustering model can separate customers into meaningful segments. Based on the results above, 5 clusters show the lowest misclassification rate of the training and validation dataset. This indicates that the cluster can assign the customers into distinct groups well for training and validation data. Cluster 3 shows the misclassification rate of validation is higher than that of training. This shows that the decision tree can cluster the segments well for the data it was trained on, but it is struggling to separate the clusters effectively on new data. For cluster 5, we can observe that the training misclassification rate is slightly higher than that of validation. This shows that the decision tree is underfitting. This might be due to the decision tree being too simple, or the training data being noisy.

Number of Clusters: 5

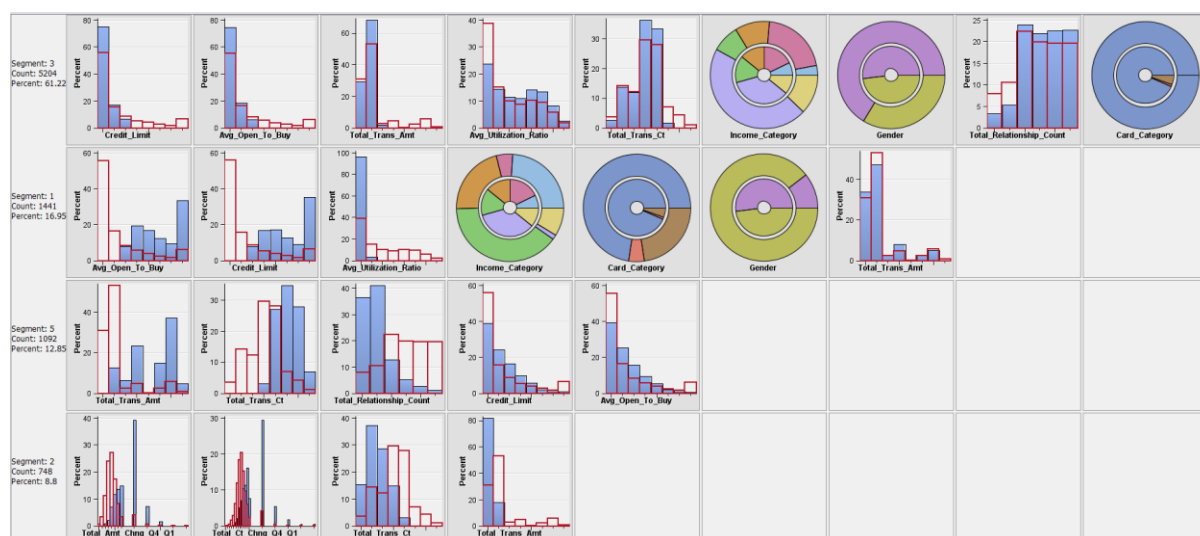


From the output above, the cluster segment 2 and segment 3 are slightly closer to each other. While other cluster segments such as clusters 1, 4, and 5 have clearly defined distances.



27	Frequencies: _SEGMENT_			
28				
29				Percent of
30	Segment	Segment	Frequency	Total
31	Variable	Value	Count	Frequency
32				
33	_SEGMENT_	3	5204	61.2235
34	_SEGMENT_	1	1441	16.9529
35	_SEGMENT_	5	1092	12.8471
36	_SEGMENT_	2	748	8.8000
37	_SEGMENT_	4	15	0.1765

From the figure above, the 5 clusters have an unequal distribution of customers. This is because cluster number 4 only has 15 customers.



The figure above shows clearly that segment 1 has a higher average open-to-buy amount than the overall population. This means that this segment has a higher available credit on their revolving accounts. Furthermore, segment 1 has a lower average utilization ratio than the population, which means they use a lesser amount than their credit limit.

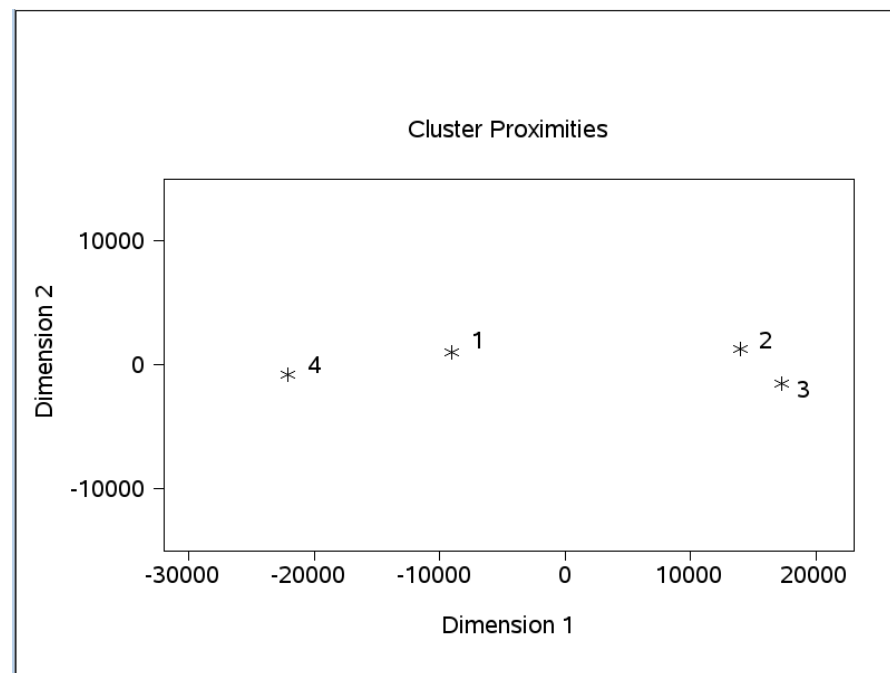
Segment 3 has a lower credit limit, total transaction amount, and average open-to-buy than the overall population. Segment 5 has a higher total transaction amount, and transaction count but a lower total relationship count than the overall population.

For segment 2, the cluster has a higher difference in total transaction count and amount between Q4 to Q1 but lower total transaction amount and count.

34	Variable Importance				
35					
36			Number of	Number of	
37			Splitting	Surrogate	
38	Variable Name	Label	Rules	Rules	Importance
39					
40	Credit_Limit		3	6	1.00000
41	Avg_Open_To_Buy		3	4	0.99939
42	Total_Trans_Ct		0	7	0.97241
43	Total_Trans_Amt		5	4	0.79088
44	Card_Category		1	3	0.73362
45	Avg_Utilization_Ratio		0	4	0.73176
46	Income_Category		1	2	0.71913
47	Total_Ct_Chng_Q4_Q1		3	6	0.51622
48	Total_Amt_Chng_Q4_Q1		2	3	0.46522
49	Customer_Age		0	5	0.44682
50	Total_Relationship_Count		0	3	0.23295
51	Total_Revolving_Bal		0	2	0.16248
52	Months_Inactive_12_mon		0	3	0.13600
53	Gender		0	1	0.09976
54	Marital_Status		0	1	0.08319
55	Months_on_book		0	2	0.06824

In the feature importance, the top 5 variables that contribute to this cluster are credit limit, average open to buy, total transaction count, total transaction amount, and card category which has a feature importance value of between 0.73 to 1.00.

Number of Clusters: 4



From the diagram above, the clusters have a clear distance from each other with no overlapping.

_SEGMENT_	_1	_2	_3	_4
1	0	22960.49	26400.94	13244.67
2	22960.49	0	4378.554	36124.86
3	26400.94	4378.554	0	39386.97
4	13244.67	36124.86	39386.97	0

The distance amongst the clusters is very consistent and is clearly defined.

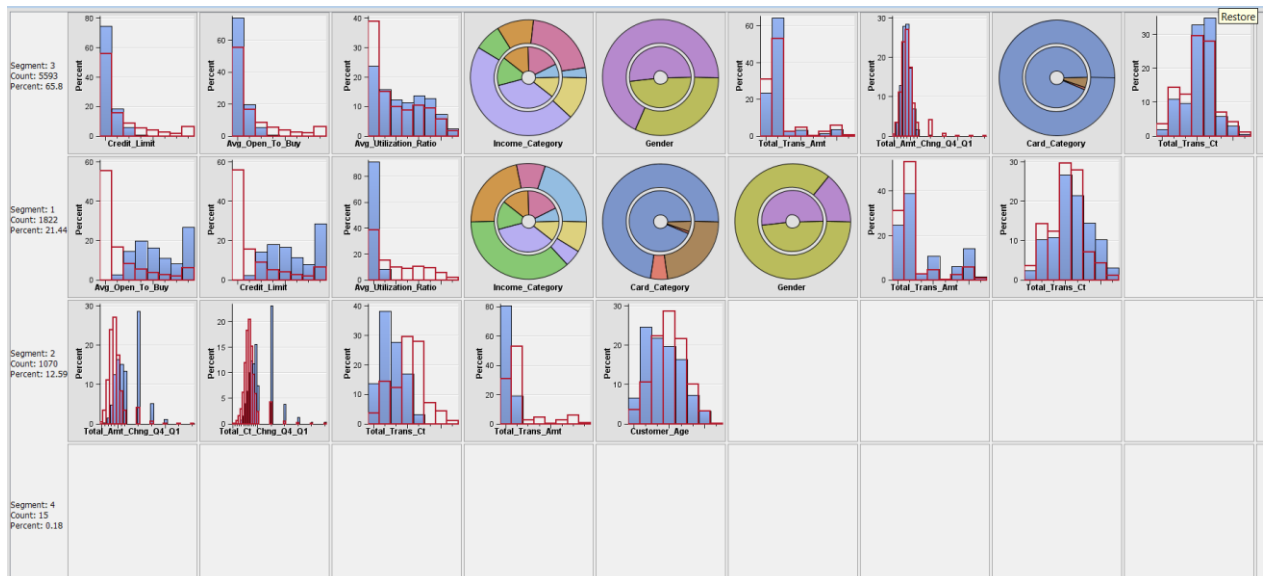
27	Frequencies: _SEGMENT_			
28				
29				
30	Segment	Segment	Frequency	Percent of
31	Variable	Value	Count	Total
32				Frequency
33	_SEGMENT_	3	5593	65.8000
34	_SEGMENT_	1	1822	21.4353
35	_SEGMENT_	2	1070	12.5882
36	_SEGMENT_	4	15	0.1765

The figure above shows that segment number 3 is 65% of the total population, followed by segment number 1 with 12% of the population, segment 2 (12%), and segment 4 only has 15 customers. This is because this segment might be outliers that require further cleaning.

## Variable Importance

Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Avg_Open_To_Buy		3	7	1.00000
Credit_Limit		2	7	0.89940
Total_Trans_Ct		1	9	0.83779
Card_Category		1	3	0.79915
Avg_Utilization_Ratio		0	5	0.79799
Income_Category		1	3	0.79719
Total_Trans_Amt		6	6	0.55359
Total_Ct_Chng_Q4_Q1		2	4	0.53546
Customer_Age		0	8	0.50525
Total_Amt_Chng_Q4_Q1		2	2	0.49693
Total_Relationship_Count		0	4	0.20440
Months_Inactive_12_mon		0	2	0.18961
Gender		1	2	0.18827
Naive_Bayes_Classifier_Attrition		0	3	0.14801
Months_on_book		0	2	0.13608
Total_Revolving_Bal		0	1	0.11283
Dependent_count		0	1	0.06739

The figure above shows the variable importance. The top 5 variables that contribute to this cluster are average open to buy, credit limit, total transaction count, card category, and average utilization ratio. The feature importance values of these variables are between 0.80 to 1.00. The feature importance range of cluster 4 is slightly better than cluster 5.



From the figure above, it is evident that Segment 3 has a lower credit limit and average open-to-buy amount as compared to the general customers. Segment 1 has a higher cluster average open-to-buy and credit limit amount, but a lower average utilization ratio as compared to the overall population. Segment 2 has a higher difference in total transaction amount between Q4 and Q1, and a higher total transaction count (Q4 vs Q1). However, they have lower total transaction amounts and counts compared to the overall population. In addition, Segment 4 has too few data to analyze the differences. We can also discover that categorical variables distribution between the segments are hard to analyze, thus one-hot encoding is needed.

### Insights Derived from Cluster 4 and Cluster 5

Based on all the information, we can conclude that Cluster 4 and Cluster 5 are not very dissimilar to each other. For cluster 5, segments number 2 and 3 are close to each other and can be grouped. It might also be prone to overlap. Furthermore, credit limit and average open-to-buy are the top two features that contribute significantly to these clusters. The subsequent steps are to perform feature selection and data cleaning of clusters 4 and 5 to observe any improvements in the misclassification rate. Alternatively, we can observe that cluster 4 provides consistent misclassification rates on training and validation data. It also provides a clear cluster plot with no overlapping and a distance table that allows interpretability of analysis. Cluster number 4 is more preferred.

### Data Preparation & Cleaning For the Clusters

#### Feature selection of Cluster 4

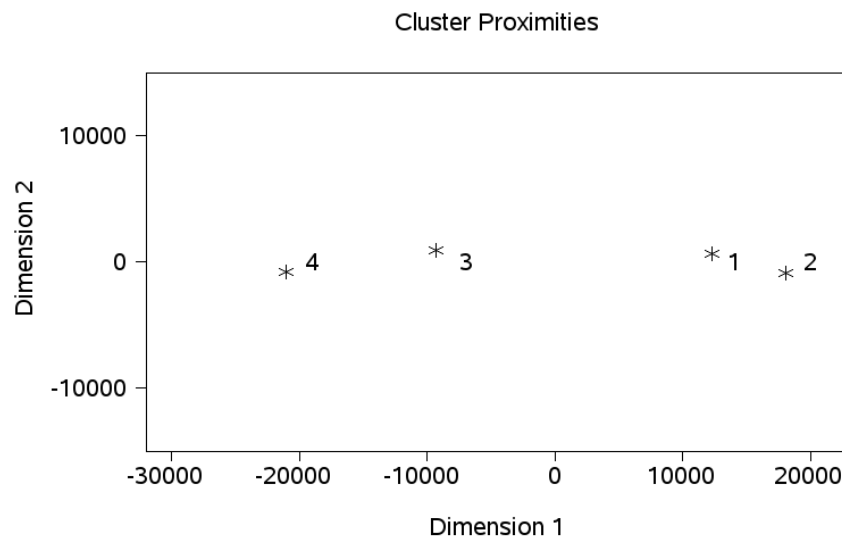
Misclassification Rate for cluster 4 (with feature selection) by removing variables dependent\_count, gender, months\_on\_booking, and total\_revolving as their feature importance values are less than 0.2. The misclassification rate of the training and validation datasets are **0.051** respectively. This shows that the misclassification rate has decreased by 37.80% compared to the previous without feature selections.

#### Converting categorical to dummy variables

Source	Method	Variable Name	Formula	Label
input	Original	Card Category		
input	Original	Education Level		
input	Original	Gender		
input	Original	Income Category		
input	Original	Marital Status		
Output	Computed	TI Card Category1	Dummy	Card Category:Blue
Output	Computed	TI Card Category2	Dummy	Card Category:Gold
Output	Computed	TI Card Category3	Dummy	Card Category:Platinum
Output	Computed	TI Card Category4	Dummy	Card Category:Silver
Output	Computed	TI Education Level1	Dummy	Education Level:College
Output	Computed	TI Education Level2	Dummy	Education Level:Doctorate
Output	Computed	TI Education Level3	Dummy	Education Level:Graduate
Output	Computed	TI Education Level4	Dummy	Education Level:High School
Output	Computed	TI Education Level5	Dummy	Education Level:Post-Graduate
Output	Computed	TI Education Level6	Dummy	Education Level:Uneducated
Output	Computed	TI Education Level7	Dummy	Education Level:Unknown
Output	Computed	TI Gender1	Dummy	Gender:F
Output	Computed	TI Gender2	Dummy	Gender:M
Output	Computed	TI Income Category1	Dummy	Income Category:\$120K +
Output	Computed	TI Income Category2	Dummy	Income Category:\$40K - \$60K
Output	Computed	TI Income Category3	Dummy	Income Category:\$60K - \$80K
Output	Computed	TI Income Category4	Dummy	Income Category:\$80K - \$120K
Output	Computed	TI Income Category5	Dummy	Income Category:Less than \$40K
Output	Computed	TI Income Category6	Dummy	Income Category:Unknown
Output	Computed	TI Marital Status1	Dummy	Marital Status:Divorced
Output	Computed	TI Marital Status2	Dummy	Marital Status:Married
Output	Computed	TI Marital Status3	Dummy	Marital Status:Single
Output	Computed	TI Marital Status4	Dummy	Marital Status:Unknown

All of the categorical variables are converted to dummy variables as shown in the diagram above.

## Checking The Cluster Distance and Distribution After Converting to Categorical Variables



After categorical variables conversion, the cluster now has a clearer distance within the segments.

**Table 2: Misclassification Rates of Training & Validation Datasets of Cluster Number 4**

Misclassification Rate of Training Dataset	0.0049
Misclassification Rate of Validation Dataset	0.0047

The misclassification rate of training and validation datasets has also reduced significantly compared to cluster 4 without dummy variable conversion.

34	Variable Importance				
35					
36					
37			Number of	Number of	
38	Variable Name	Label	Splitting	Surrogate	
39			Rules	Rules	Importance
40	TI_Gender2	Gender:M	0	1	1.00000
41	TI_Gender1	Gender:F	1	0	1.00000
42	TI_Income_Category5	Income_Category:Less than \$40K	1	1	0.90040
43	Credit_Limit		1	3	0.85938
44	Avg_Open_To_Buy		1	3	0.85826
45	TI_Income_Category4	Income_Category:\$80K - \$120K	0	1	0.82193
46	Customer_Age		0	1	0.20317
47	TI_Card_Category2	Card_Category:Gold	1	0	0.19435
48	Avg_Utilization_Ratio		1	1	0.17155
49	TI_Card_Category1	Card_Category:Blue	1	1	0.15098
50	TI_Card_Category4	Card_Category:Silver	0	2	0.15098
51	Total_Revolving_Bal		0	1	0.13752
52	TI_Income_Category2	Income_Category:\$40K - \$60K	0	1	0.08917
53	TI_Income_Category6	Income_Category:Unknown	0	1	0.08917
54	Total_Relationship_Count		0	1	0.07953
55	TI_Card_Category3	Card_Category:Platinum	1	0	0.07764

The top five features now that affect the cluster assignments are Gender (M and F), Income Category (Less than \$40K), Credit Limit, and Average open-to-buy. These features can be further investigated in the analysis later.

## Data Preparation: Standardization of Skewed Variables

Name	Method	Number of Bins	Role	Level
Avg_Open_To	Default	4	Input	Interval
Avg_Utilization	Default	4	Input	Interval
Card_Category	Default	4	Input	Nominal
Contacts_Count	Default	4	Input	Interval
Credit_Limit	Default	4	Input	Interval
Customer_Age	Default	4	Input	Interval
Dependent_count	Default	4	Input	Interval
Education_Level	Default	4	Input	Nominal
Gender	Default	4	Input	Nominal
Income_Category	Default	4	Input	Nominal
Marital_Status	Default	4	Input	Nominal
Months_Inactive	Default	4	Input	Interval
Months_on_book	Default	4	Input	Interval
Total_Amt_Chng	Log	4	Input	Interval
Total_Ct_Chng	Log	4	Input	Interval
Total_Relation	Default	4	Input	Interval
Total_Revolvin	Default	4	Input	Interval
Total_Trans_A	Default	4	Input	Interval
Total_Trans_C	Default	4	Input	Interval

The figure above shows the total amount change (Q4 vs Q1) and the total amount change (Q4 vs Q1) has been standardized using log-transformed.

**Table 3: Cluster 4 (without feature selection) with Log Transform and Categorical to Dummy Variables**

Misclassification Rate of Training Dataset	0.0024
Misclassification Rate of Validation Dataset	0.0020

After standardization, the misclassification rate was further reduced by half compared to the misclassification rate without standardization. The misclassification rate of the training dataset is slightly higher than that of the validation dataset. However, the values are approximate to each other, with only 0.2% of misclassification.

**Table 4: Cluster 4 (feature selection) with Log Transform and Categorical to Dummy Variables**

Misclassification Rate of Training Dataset	0.0028
Misclassification Rate of Validation Dataset	0.0051

After feature selection, the validation misclassification rate is higher than that of the training dataset. Thus, overfitting might occur. It can be observed that the misclassification rates after feature selection become higher. Thus, this feature selection is not feasible.

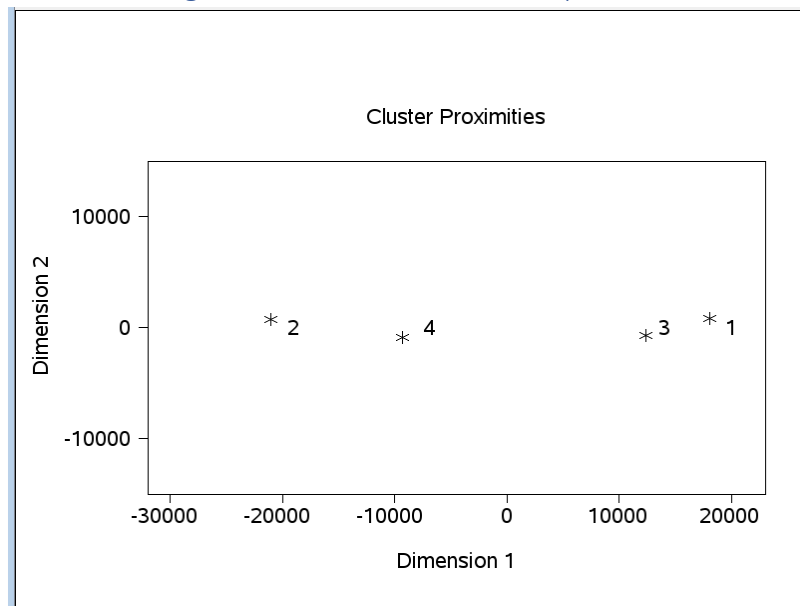
**Table 5: Cluster 5 (without feature selection) with Log Transform and Categorical to Dummy Variables**

Misclassification Rate of Training Dataset	0.0404
Misclassification Rate of Validation Dataset	0.0372

The misclassification rates of training and validation datasets of cluster 5 after data cleaning are significantly higher compared to cluster 4, which is not appropriate for the clustering assignment and is not chosen.

Therefore, we can conclude that cluster 4 (Table 3) without feature selection is selected due to its lower misclassification rate compared to cluster 4 with feature selection. The next step is to observe the visualization of customer segment profiles.

## Customer Segment Profiles of Cluster 4 (without Feature Selection + data cleaned)



The figure above shows that each segment has a clear and distinctive distance.

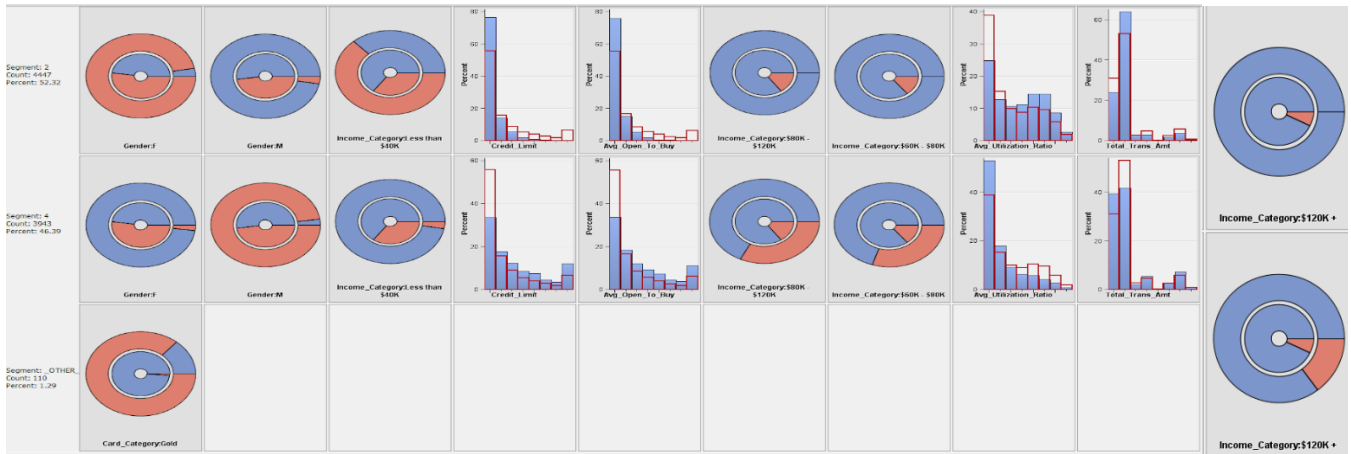
_SEGMENT_	_1	_2	_3	_4
1		0	39107.79	5933.955
2	39107.79		0	27437.87
3	5933.955	33396.17		11838.85
4	27437.87	11838.85	21643.25	0

There is a great distance between each segment.

27	Frequencies: _SEGMENT_			
28				
29				
30	Segment	Segment	Frequency	Percent of
31	Variable	Value	Count	Total
32				Frequency
33	_SEGMENT_	2	4447	52.3176
34	_SEGMENT_	4	3943	46.3882
35	_SEGMENT_	_OTHER_	110	1.2941

The figure above shows that Segment 2 has the highest frequency with a total of 4447 customers, followed by Segment 4 with 3943 customers. Segments 1 and 3 are merged together as others, with a total of 110 customers.

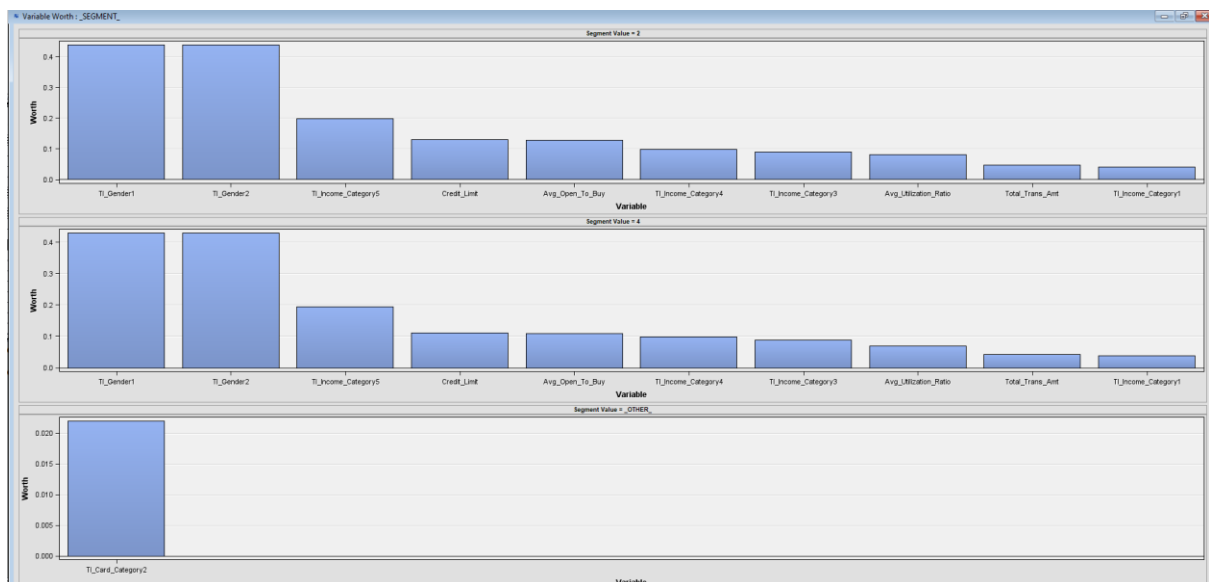




The customer profiles show that segment 2 has an equal amount of male and female representation, the income category of this segment is mainly less than \$40K as it is overrepresented in this segment compared to other income categories. Segment 2 customers have lower credit limits and average open-to-buy compared to the overall population. The average utilization ratio of segment 2 is also slightly higher than the overall population.

For segment 4, the customers are mostly male with females slightly underrepresented in this segment. This segment focuses on the higher income category with income ranges between 60K to 80K, 80K to 120K, and more than 120K. Segment 4 has a higher credit limit with a higher average open-to-buy, which means that this segment has higher credit available and remains unused on average. This segment has a lower average utilization ratio.

The other segments are mainly focused on the customers who hold gold-card which is very unusual for the overall population.



The figure above shows the features importance of each segment. The top variables that contribute to Segment 2 are gender, income category, credit limit, average open-to-buy, average utilization ratio, and total transactions. Segment 4 features importance is same as the Segment 2. While Segment 3 is only looking at the customers who have gold cards.

# Attrite customer

**Table 6: Determine the Number of Clusters of churned customers by Checking Misclassification Rates**

Cluster Number	Misclassification Rate
Cluster 4	Training: 0.097 Validation: 0.130
Cluster 3	Training: 0.016 Validation: 0.024
Cluster 2	Training: 0.013 Validation: 0.025

Based on the table above, cluster number 3 is selected for churned customer segmentation as it produces the lowest misclassification rates of the validation dataset compared to cluster 4 and cluster 2. This indicates that Cluster 3 provides better generalizability in distinguishing between customer segments and the cluster is selected.

However, the result above has shown that all clusters misclassification rate of validation data is higher than training data. This indicates higher misclassification of the predicted cluster assigned compared to the actual cluster of the validation dataset. However, the values are close to 0, thus it is still a meaningful segmentation.

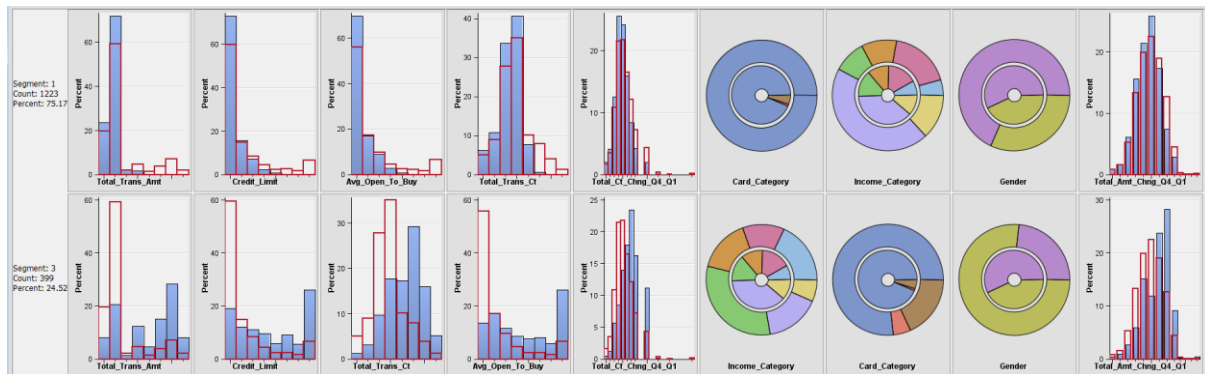
## Customer Segmentation Profile of Cluster Number - 3

Variable Importance				
Variable Name	Label	Number of Splitting Rules	Number of Surrogate Rules	Importance
Total_Amt_Chng_Q4_Q1		0	9	1.00000
Card_Category		1	3	0.97794
Total_Trans_Amt		4	2	0.88070
Total_Trans_Ct		1	5	0.86095
Total_Ct_Chng_Q4_Q1		2	5	0.81171
Customer_Age		0	3	0.78423
Avg_Open_To_Buy		3	7	0.67694
Credit_Limit		2	5	0.66392
Avg_Utilization_Ratio		0	5	0.24142
Contacts_Count_12_mon		0	2	0.15560
Gender		1	1	0.15193
Months_on_book		0	1	0.14895
Total_Relationship_Count		0	1	0.13895
Income_Category		0	1	0.11554
Months_Inactive_12_mon		0	1	0.08741
Dependent_count		0	1	0.08381

The figure above shows the feature importance of Cluster 3 with the top 5 variables being the total amount of change (Q4 vs Q1), Card Category, Total Transaction Amount, Total Transaction Count and Change (Q4 vs Q1).

27	Frequencies: _SEGMENT_			
28				
29				
30	Segment	Segment	Frequency	Percent of
31	Variable	Value	Count	Total
32				Frequency
33	_SEGMENT_	1	1223	75.1690
34	_SEGMENT_	3	399	24.5237
35	_SEGMENT_	2	5	0.3073

Based on the Figure above, segment 1 has the highest percentage of customers followed by segment 3 and segment 2 only has 5 customers.



The figure above shows for cluster 3, segment 1 customers have lower total transaction amounts, credit limit, average open-to-buy, and lower transaction count which might lead to churning. These early signs need to be detected. For segment 1, the customers mostly have higher transaction amounts, total transaction count, higher average open-to-buy, and credit limits. The reason of churning needs to be investigated further, which might be due to card category, income category or gender. The other segments have too few data to analyze the differences.

#### Data Cleaning (convert categorical to dummy) of Churned Customers

**Table 1: Misclassification Rates of Training and Validation Datasets for 3 Clusters**

Number of Clusters	Misclassification rate
3	Training: 0.0035 Validation: 0.0041

Based on the output above, the misclassification rate has significantly decreased by 82.91% from 0.024 to 0.0041 after converting the categorical to a dummy variable.

_SEGMENT_	1	2	3
1		0	29182.67
2		29182.67	0
3		2993.354	26268.36

The distance within the segments is clearly defined and consistent with each other's.

```

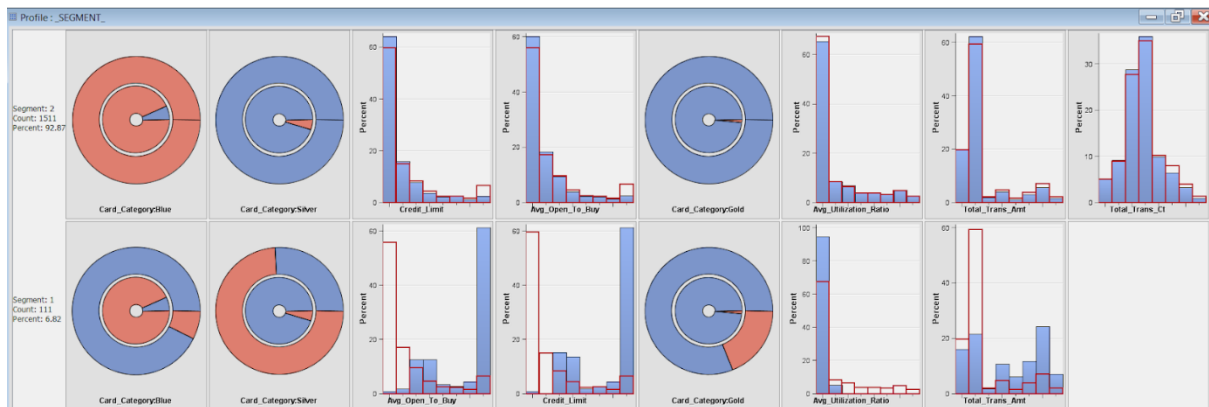
20
27 Frequencies: _SEGMENT_
28
29
30 Segment      Segment      Frequency      Percent of
31 Variable      Value      Count      Total
32
33 _SEGMENT_      2          1511      92.8703
34 _SEGMENT_      1           111      6.8224
35 _SEGMENT_      3           5       0.3073

```

Based on the Figure above, segment 2 has the highest percentage of customers (92.08%) followed by segment 3 and segment 2 only 5 customers.

34	Variable Importance				
35					
36			Number of	Number of	
37			Splitting	Surrogate	
38	Variable Name	Label	Rules	Rules	Importance
39					
40	TI_Card_Category1	Card_Category:Blue	1	0	1.00000
41	TI_Card_Category4	Card_Category:Silver	0	1	0.99351
42	Credit_Limit		0	2	0.98272
43	Avg_Open_To_Buy		1	1	0.98211
44	TI_Card_Category2	Card_Category:Gold	0	1	0.97439
45	Total_Amt_Chng_Q4_Q1		0	1	0.96805
46	Total_Trans_Ct		1	1	0.26981
47	Total_Ct_Chng_Q4_Q1		0	1	0.23249
48	Total_Trans_Amt		0	1	0.23249
49	TI_Card_Category3	Card_Category:Platinum	1	0	0.22112
50	Dependent_count		0	1	0.21991

The features that contribute to this segment are card category (blue, silver and gold), credit limit, average open-to-buy, and total transaction amount change (Q4 vs Q1).



Based on the figure above, segment 2 churned customers have more blue cards compared to the overall population. Not many of the customers in Segment 2 hold silver or gold cards. The credit limit, average open-to-buy, credit utilization ratio, and total transaction are like the overall population. For segment 1, most of the customers have more credit limits and higher average open-to-buy (available credit). Segment 1 has a lower average credit utilization ratio, but a higher total transaction amount compared to segment 2. Furthermore, Segment 1 customers mostly hold a silver card and gold card.

# Discussion of Model Outcomes.

## Retention Strategies of Existing Customers

For the existing customers, the hybrid model can classify the existing clusters into meaningful clusters based on their feature split and distinct segment properties. The misclassification ratio is very low – 0.2%. For existing customers, the credit card providers or the bank can target the lower income category (less than 40K), lower credit limit but a higher utilization ratio. This means they have used the card more with a limited credit limit. Thus, the bank should also adopt budget management tools to gauge their spending and provide some incentive to lower their credit utilization. The bank can also increase their credit usage by boosting their credit limit based on adequate usage to reduce their revolving balance (decrease the burden of debt). This group generally has higher financial risks with might fail to repay the loans. The bank should leverage the interest on balance to generate revenue. However, the banks might face some risk if this segment's financials are not stable.

Segment 4 encompasses high-income males with higher credit limits and an average amount that remains unused (open-to-buy) with a lower utilization ratio. The bank could increase retention of this customer with higher premium cards (gold or platinum) based on their spending habits. More financial benefits could be given for this segment such as cash-back rewards, travel packages or promotions, or high-end financial products. Credit card providers can introduce wealth management or investment options for this segment with low-risk high rewards to further engage them in the business. They are the high-value customers whom the bank can offer upselling and cross-selling to improve revenue. However, the customers do not fully utilize the available credits leads to lower interest payments which restricts revenue. The bank should have tiered reward bonuses or spending bonuses targeted to this segment for cash-back reward in higher transaction count and amount.

The bank can target customers who hold gold cards by offering gold card exclusive benefits or dividends. Upgrading the gold card to a platinum card and exclusive benefits can potentially drive more engagement within this segment. The bank should prevent the customers from these segments from leaving especially if the competitor offers better rewards or services. This is because these customers are more sensitive to benefits than to interest rates.

## Retention Strategies of Churned Customers

The characteristics of churned customers are significantly different from existing customers. This is because the attrite customers are higher-aged, remain inactive for a longer period, and have lower revolving balances and transaction amounts compared to non-churners. They are clustered into three groups with distinct features, the hybrid model has achieved a misclassification rate of 0.4%, which is approximately 0.

One of the target segments of churned customers is mainly holding blue cards, having similar financial behaviours as the overall population. Thus, the bank should craft an incentive to increase their credit spending. For instance, re-engagement activities such as cash-back rewards, credit cards with low interest rates, and many more encourage at-risk customers to stay. Another churned customer segment has a higher credit limit and more available credit but a lower utilization ratio than the overall population. The bank needs to encourage more credit

usage for this segment. Thus, banks should have rewards targeted to them if they spend more. They can utilize their available credit by participating in large purchase offers or low-interest financing. Furthermore, targeted promotions can also be given to silver or gold cardholders.

The banks might have revenue loss due to the customer's dissatisfaction with the services, lack of engagement, or better rewards from competitors. Furthermore, the lower credit utilization and transaction amount lead to the churning of customers. Therefore, the banks can think of having lower interest rates or repayment fees for this segment. Bonus rewards for new card usage can be introduced with a balance transfer offer targeted to the churned customers. The bank should detect the inactive period, low revolving balance, and reduced transactions of customers in real-time by constructing an early prevention or detection system. To increase credit utilization, the bank should have encouraged an increased credit limit program for churners who have good credit behaviour so that they can return as customers. A temporary reduction in interest rates on the revolving balance should be emphasized so that they can consider the benefits before churning.

In addition, the churned customers are mostly women. Thus, the bank should have an engagement program targeted to mothers such as a family reward program. Besides that, the bank should also promote credit card usage based on seasonal trends for the churned customers as their total amount differences (Q4 vs Q1) are lower. In addition, banks can have personalized financial offers, explaining the key benefits of their products to their customers with new updates or special deals in the future (SuperOffice, 2024). The card providers need to have social listening to empathize with their customer's needs and desires so that the customer experience can be enhanced. Defining the high-risk customers is important by providing customer support services, allowing the customer to halt their credit card subscription, and promoting discounted pricing to about-to-churned customers (SuperOffice, 2024). Online marketing or email marketing is crucial to attract customers to stay by offering bundles of special offers to them. Thus, the competitive advantage of the credit card providers or banks needs to be emphasized, and can also consider having longer subscription models instead of month-to-month contracts of credit card subscriptions (SuperOffice, 2024).

## Conclusion

The study has achieved the objectives of performing cluster profile segmentation of existing and churned customers. The existing customers are segmented into 4 clusters while churners have 3 clusters with different characteristics. The cluster segmentation is very effective as the predictive model – Decision Tree manages to classify the training and validation data into distinct clusters. Banks can improve customer retention by re-engaging churned customers through retention campaigns, incentivizing credit utilization, and offering cash-back rewards. For existing customers, identifying income levels and spending behaviors will help maintain loyalty and drive engagement. These targeted strategies can reduce churn rates and boost the overall profitability of credit card providers or banks.

## References

SuperOffice. (2024). *Customer Churn: 12 Strategies to Stop Churn Right Now!* Superoffice.com. <https://www.superoffice.com/blog/reduce-customer-churn/>