Data Analysis
and
Data Visualization
of
Happy Bank Data

Objective

The objective of this project is to use the data analysis techniques to predict customer attrition. By applying predictive modeling to the customer survey data, we aim to develop a model that can anticipate which customers are more likely to cancel their credit card services. This will enable Happy Bank to proactively address the concerns of those customers and provide better services to turn customer's decisions in the opposite direction.

Introduction

Welcome to our project on customer attrition prediction for Happy Bank's credit card services. In this project we aim to help Happy Bank understand why more and more customers are leaving their credit card services and predict customer attrition based on various attributes such as Customer age, credit limit, dependent count.

Happy Bank is a financial institution that offers a range of credit services, including credit cards to its customers. However recently, the bank has

noticed an increase in the number of customers canceling their credit card accounts. This has raised concerns for the bank's management team as they strive to provide the best possible service to their customers.

To gain insights into the reasons behind this customer attrition, the management team at happy bank conducted a customer survey. They collected information on various attributes such as customer age, credit limit, and other relevant factors. By analyzing the data, the team hopes to uncover patterns and trends that can help them understand why customers are leaving the credit card services.

Methodology

In order to address the challenge of increasing customer attrition in Happy bank credit card services, we will followed a systematic approach. The methodology involved several key steps to gather data, analyze customer attributes, and develop a predictive model. These steps are outlined below:

1. Data collection: The team conducted a comprehensive customer survey to collect data on various attributes such as Customer age, credit limit, dependent count.
2. Display data: This involved reaching out to a sample of data to gather the information.
3. Checking the shape of data: This involved the amount of rows and columns present in the data.
4. Checking percentage of missing values in each column of data: The collected data underwent a thorough cleaning process to remove any (inconsistencies)missing values.

5. Checking for duplicate rows: The collected data underwent a thorough cleaning process to remove any (inconsistencies)duplicate values and will give unique values.
6. MODEL DEVELOPMENT: tasks in these modelling given as below:

1.Checking the distribution of the Customer_Age column.

2.Checking the basic statistics like mean,median and standard deviation of the age column.

3. Plot 2 Boxplots and 2 pie charts of the parameter of your own choice and write your Intitution about it.

4. plot a box-plot of Total_Revolving_Bal and Card_Category by Characterizing with Attrition_Flag.

5. Plot a percentage segment bar graph between education_level and Attrition_flag of the customers.

6. Plot a percentage segment bar graph between Income_Category and Attrition_Flag of the customers.

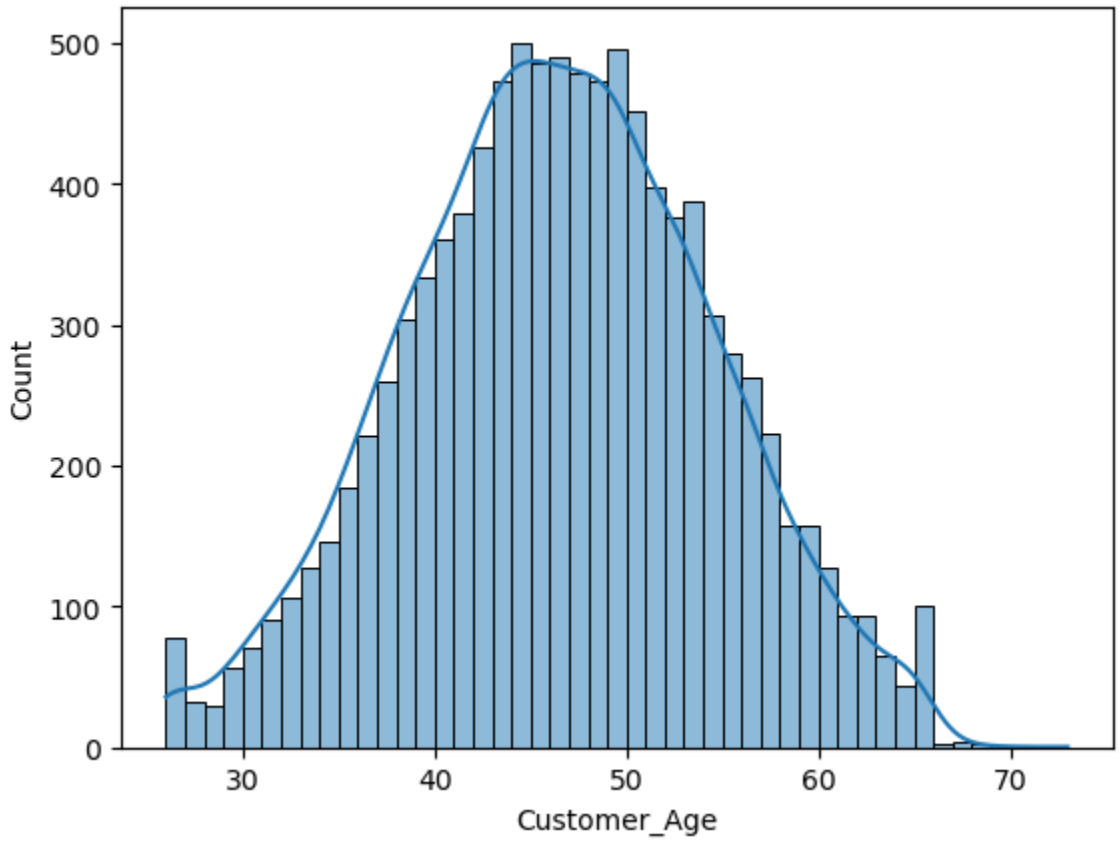
7. Drop CLIENTNUM column.Make a sub data frame which consists of all the numerical columns(i.e.int64,float64) along with the Attrition_Flag column. Plot a clear heatmap to view the correlation using seaborn.

8. Plot a boxplot for the Credit_Limit column and check if it contains any outlier or not.

9. Map the Attrition_Flag values to 0 and 1(i.e. Existing Customer=0 and Attrited Customer=1. Standardize the columns.

FINDING INSIGHTS IN DATA:

1.Checking the distribution of customer's age column:



Analysis of customer age distribution:

- To gain insights into the distribution of customer age in the dataset, a histogram was created to visualize the number of Customers within different age groups. The histogram provides a clear understanding of

the distribution pattern and helps to identify any significant trends or patterns.

- The customer age data was divided into the Several age groups, such as 0-30, 30-40,40-50,50-60,60-70,and so on. The number of Frequency of customers falling into each group was then Plotted the Vertical axis,while ranges were represented on the horizontal axis.

Upon analyzing the histogram, the following observations can be made:

a.)Age group distribution: It provides an overview of the customers within different age ranges.

b.)Central tendency: The histogram allows us to identify the central tendency or the most common age group among the customers. This can be determined by identifying the peak with the highest frequency.

- Example: Here we got the highest frequency of age group distribution between the age group of 40-

c.)Skewness: The shape of the histogram can indicate whether the age distribution is skewed to the left, right or symmetric.

- Example: The shape of histogram indicates the age distribution towards Right.

2. Checking the basic statistics like mean,median and Standard deviation of the age column

count	10127.000000
mean	46.325960
std	8.016814
min	26.000000
25%	41.000000
50%	46.000000
75%	52.000000
max	73.000000
median	46.000000

Analysis of mean,median and standard deviation:

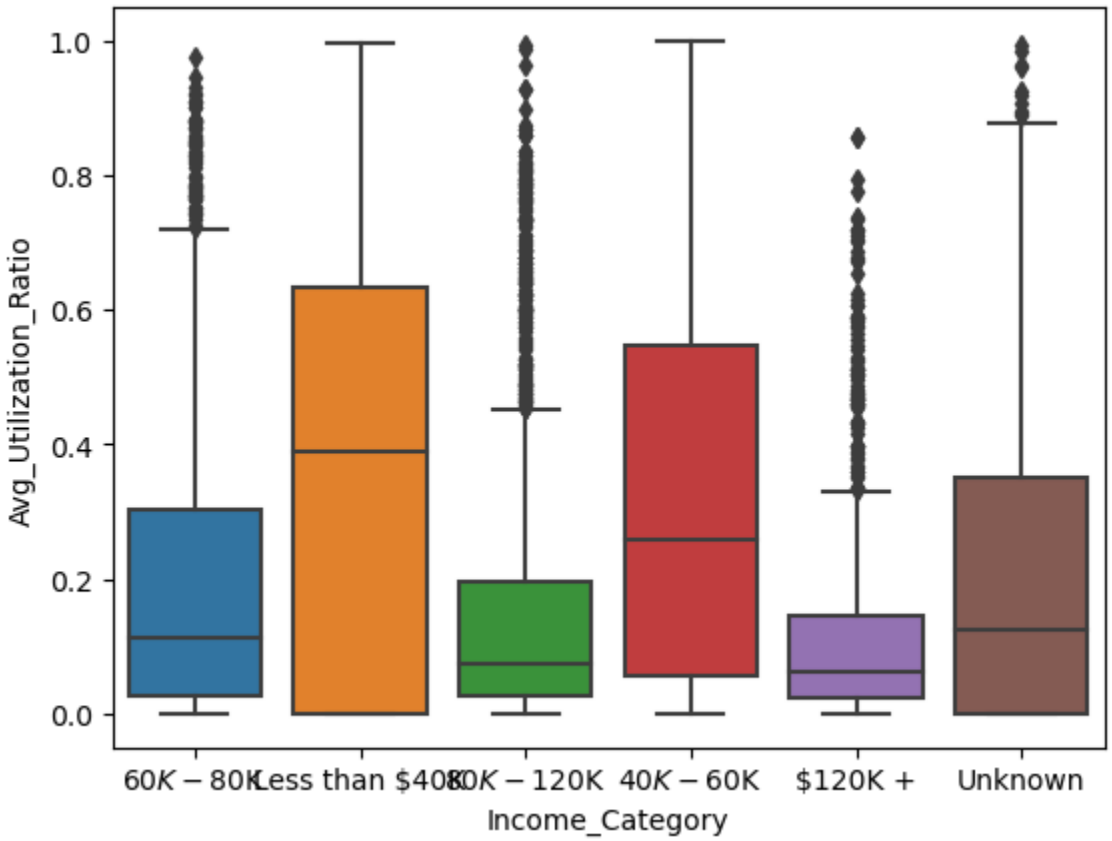
Explanation of **mean,median** and **standard deviation analysis** :

a.) **Mean**: The mean is the average value of a set of numbers. Here Mean helps us to interpret the average value of the age group. Here the value of mean would be 46.4.

b.) **Median**: The median is the middle value in a set of numbers.The median is useful because it is not affected by extreme values.Here the value of median would be 46.

c.)**Standard Deviation**: The standard deviation measures how spread out or varied the numbers are in a dataset. It tells us how much the numbers are spread out and there is more variability. If the standard deviation is low, it means the numbers are closer to the mean and there is less variability. Here the value of standard deviation is 8.016814,. Since this is the average value of standard deviation hence the variability is average too.

3. Plotting of box plots



This is a box plot between **income category** and **average utilization ratio**.It helps us to understand how the numbers are spread out and where the middle value is.

In a box plot,There is a box that represents the middle 50% of the numbers. The line in the middle of the box shows the middle value or median. If the line is closer to the bottom of the box,it means most of the numbers are smaller. If the line is closer to the top,it means most of the numbers are bigger.

Analysis of Box plot between income category and avg_utilisation.

INCOME CATEGORY:

- The box plot for the income category represents the distribution of the data across different income levels. Each box represents a specific income category, and the height of the box indicates the spread of the data within that category.
- You can look at the medians of the boxes to understand the central tendency of the income category. If the medians are higher in certain categories, it suggests that the income tends to be higher in those groups.

AVERAGE UTILIZATION:

- The box plot for average utilization represents the distribution of utilization rates for different groups or individuals. The utilization rate refers to the percentage of credit limit used by customers.
- The height of the box indicates the spread of utilization rates within the dataset. A taller box suggests a wider range of utilization rates, while a shorter box indicates a narrower spread
- outliers , represented by individual points which can provide insights. Outliers suggest exceptional cases present in any particular case.

ANALYSIS :

Case 1: Less than 40K : There Avg_Utilization_Ratio is Less Than 0.8 No Outliers

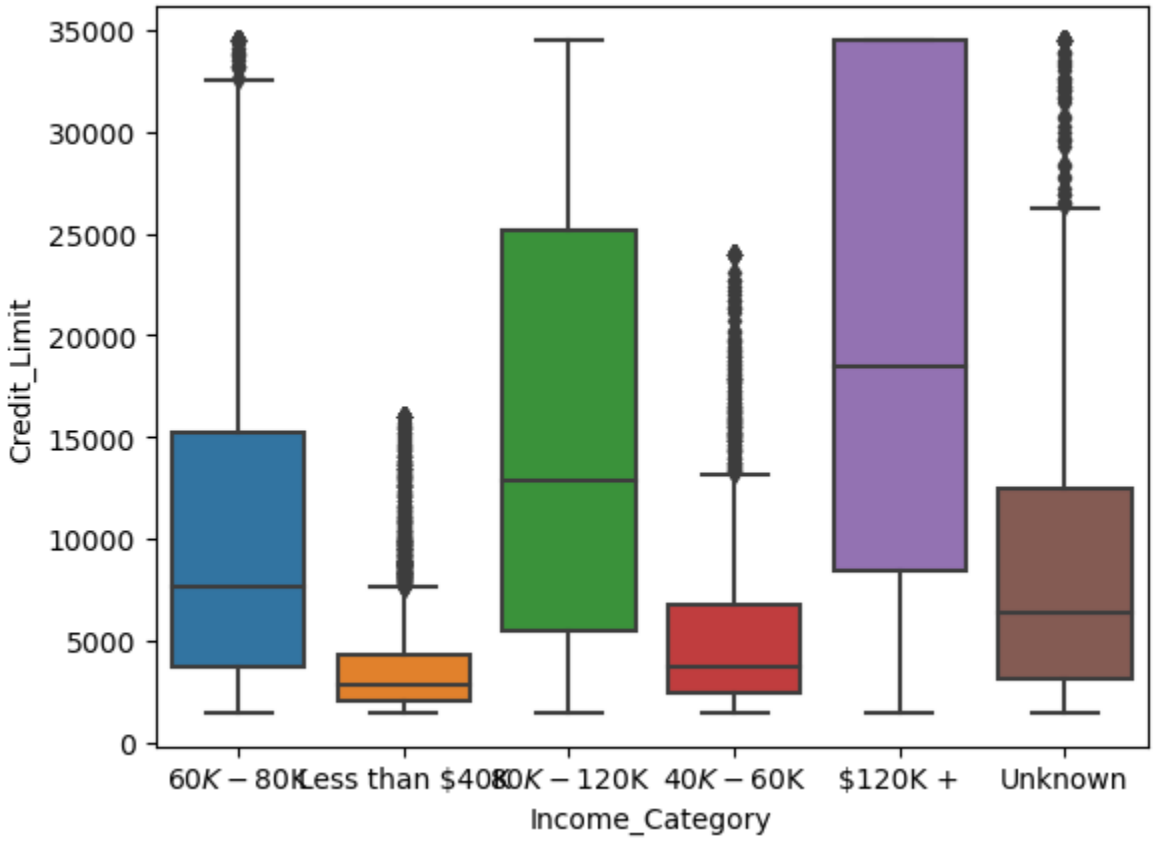
Case 2: 40k-60k $: There Avg_Utilization_Ratio is More Than 0.0 Less Than 0.6 No Outliers

Case 3: 60k-80k $: There Avg_Utilization_Ratio is More Than 0.0 Less Than 0.4 Outliers : More Than 0.6 and till 1.0

Case 4: 80k-120k $: There Avg_Utilization_Ratio is more than 0.0 less than 0.2 Outliers : More Than 0.4 and till 1.0

Case 5: 120K+ $: There Avg_Utilization_Ratio is more than 0.0 less than 0.2 Outliers : More Than 0.2 less than 1.0

Case 6: Unknown : There Avg_Utilization_Ratio is More Than 10K and Less Than 0.4 Outliers : More Than 0.8 and till 1.0



ANALYSIS :

This is a box plot between **income category** and **credit limit**.

Case 1: Less than 40K : There Credit_Limit is Less Than 5K $ Outliers : More Than 5K and Less Than 20K

Case 2: 40k-60k $: There Credit_Limit is More Than 5K$ Less Than 10K $ Outliers : More Than 10K and Less Than 25K

Case 3: 60k-80k $: There Credit_Limit is 15K $ Outliers : More Than 30K and Less Than 35K

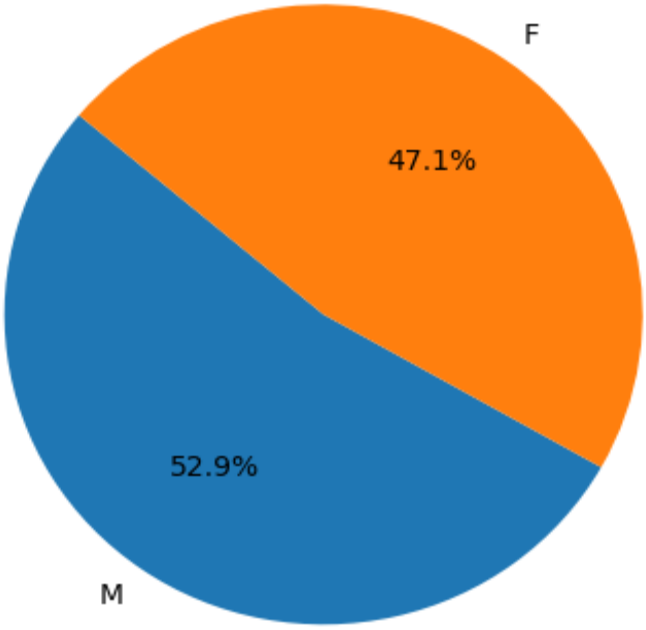
Case 4: 80k-120k $: There Credit_Limit is 25K $ Outliers : No Outliers

Case 5: 120K+ $: There Credit_Limit is 35K $ Outliers : No Outliers

Case 6: Unknown : Their Credit_Limit is More Than 10K and Less Than 15K $ Outliers : More Than 25K and Less Than 35K.

Pie charts:

a.) This pie chart is all about gender.



A **pie chart** is a circular chart divided into slices, where each slice represents a category or group. In this case, we are using a pie chart to represent the **distribution of females** and **males** in a dataset.

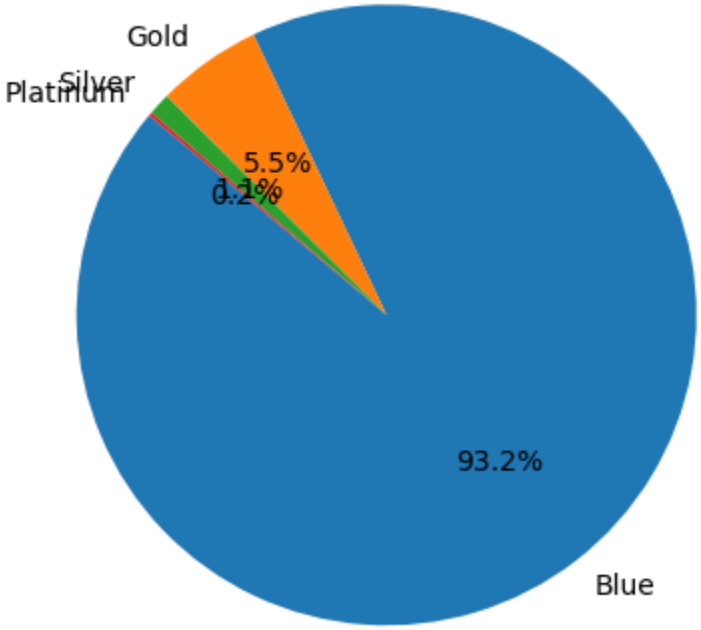
Females: The pie chart will have a slice dedicated to females. The size of the slice will correspond to the proportion or percentage of females in the dataset. The larger the slice, the higher the percentage of females.

Males: Similarly, the pie chart will have a slice dedicated to males. The size of the slice will represent the proportion or percentage of males in the dataset. The larger the slice, the higher the percentage of males.

Analysis:

- In this pie chart data we can easily analyze that the ratio of male is 52.9% and the ratio of female is **47.1%.**
- There are total : Females : 5358 Males : 4769
- By looking at the pie chart, you can quickly determine the relative representation of females and males in the dataset. If the female slice is larger, it suggests that females make up a larger proportion of the group. If the male slice is larger, it indicates a higher percentage of males.
- Pie charts are effective visual tools for comparing proportions and understanding the composition of a dataset.

This pie chart is all about the card category.

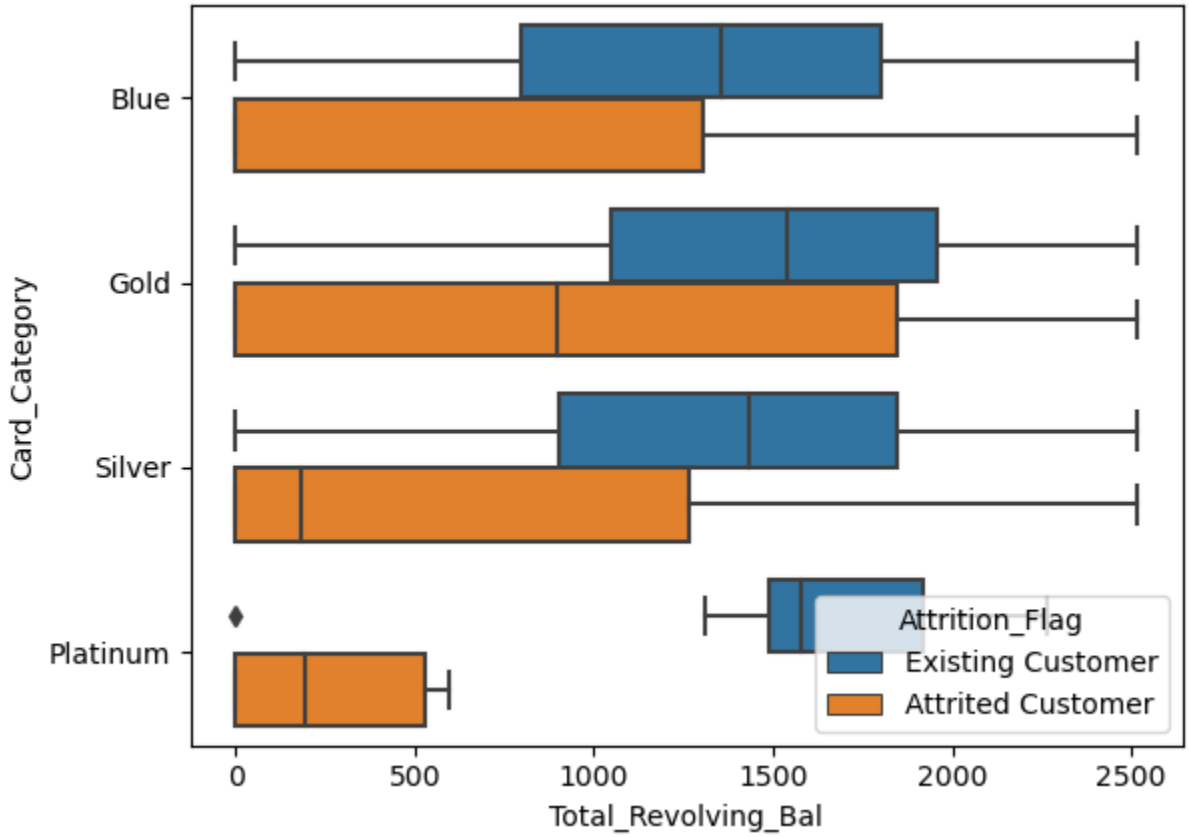


Analysis:

- In this pie chart data we can easily analyze that the ratio of gold is **5.5%** ,ratio of silver is **1.1%,** ratio of platinum is **0.2%** and the ratio of blue is **93.2%.**
- There are total: **Blue : 9436 Silver : 555 Gold : 116 Platinum : 20**

4. Box-plot of Total_Revolving_Bal and Card_Category by characterizing with Attrition_Flag.

This is the box plot between card category and total revolving balance with including the third variable as attrition flag.

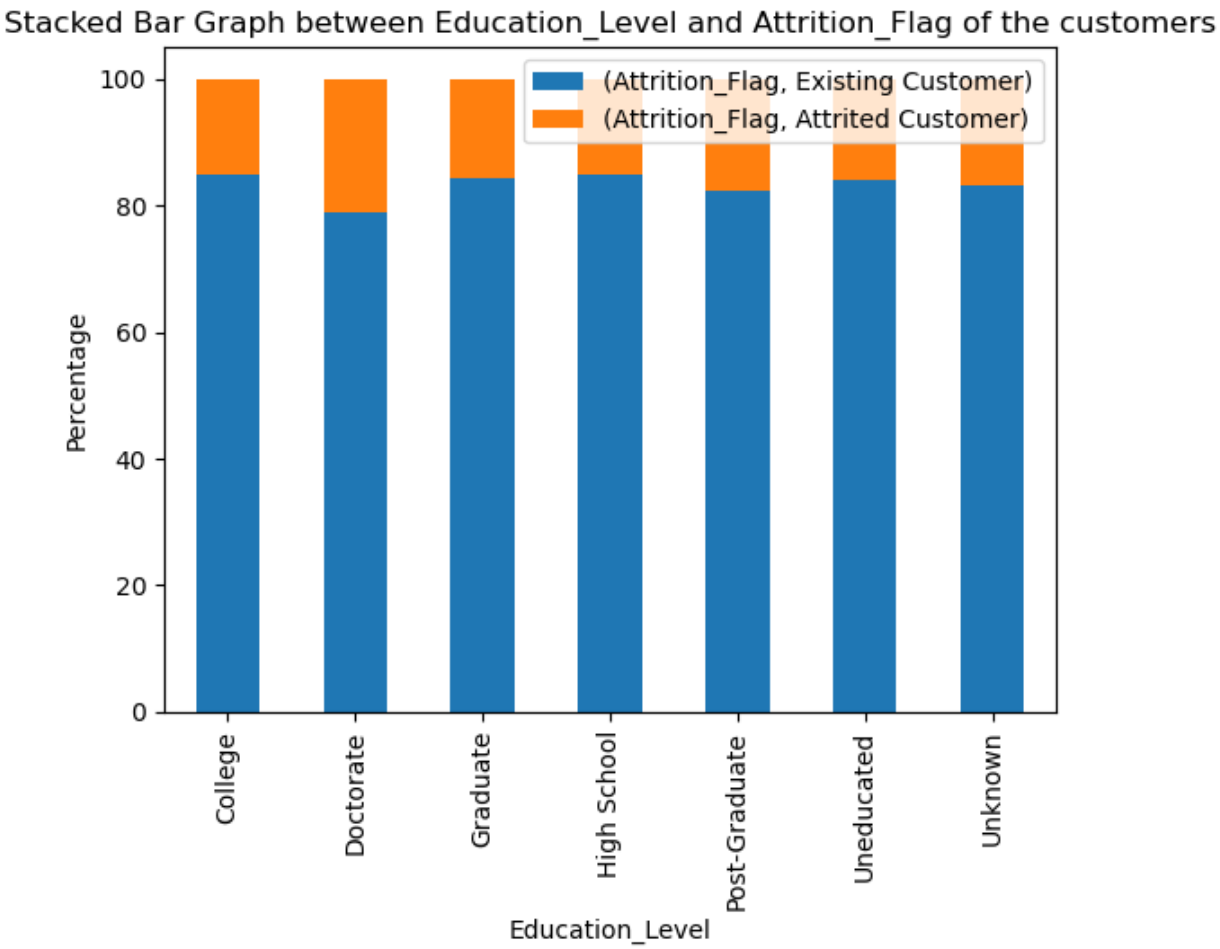


ANALYSIS:

The reason behind using the third variable is to distinguish between existing customer and attrited customer with respect to total revolving balance and card category.

CARD CATEGORY	ATTRITED CUSTOMER REVOLVING BALANCE	EXISTING CUSTOMER REVOLVING BALANCE
Blue	0 to less than 1500	More than 500 and less than 2000
Gold	0 to less than 2000	More than 1000 to 2000
SILVER	0 to less than 1500	More than 500 and less than 2000
PLATINUM	0 to 500	1500 to less than 2000

5. Plot a percentage segment bar graph between Education_Level and Attrition_Flag of the customers.



A stacked bar graph is a visual representation of data that shows multiple categories or groups stacked on top of each other within a bar. Each category is represented by a different color, and the height of the bar indicates the total value of all the categories combined.

Analysis of a stacked bar graph involves several key aspects:

Comparison of Category Sizes: By looking at the heights of the bars, you can compare the total values of each category. The taller the bar, the larger the total value of that category.

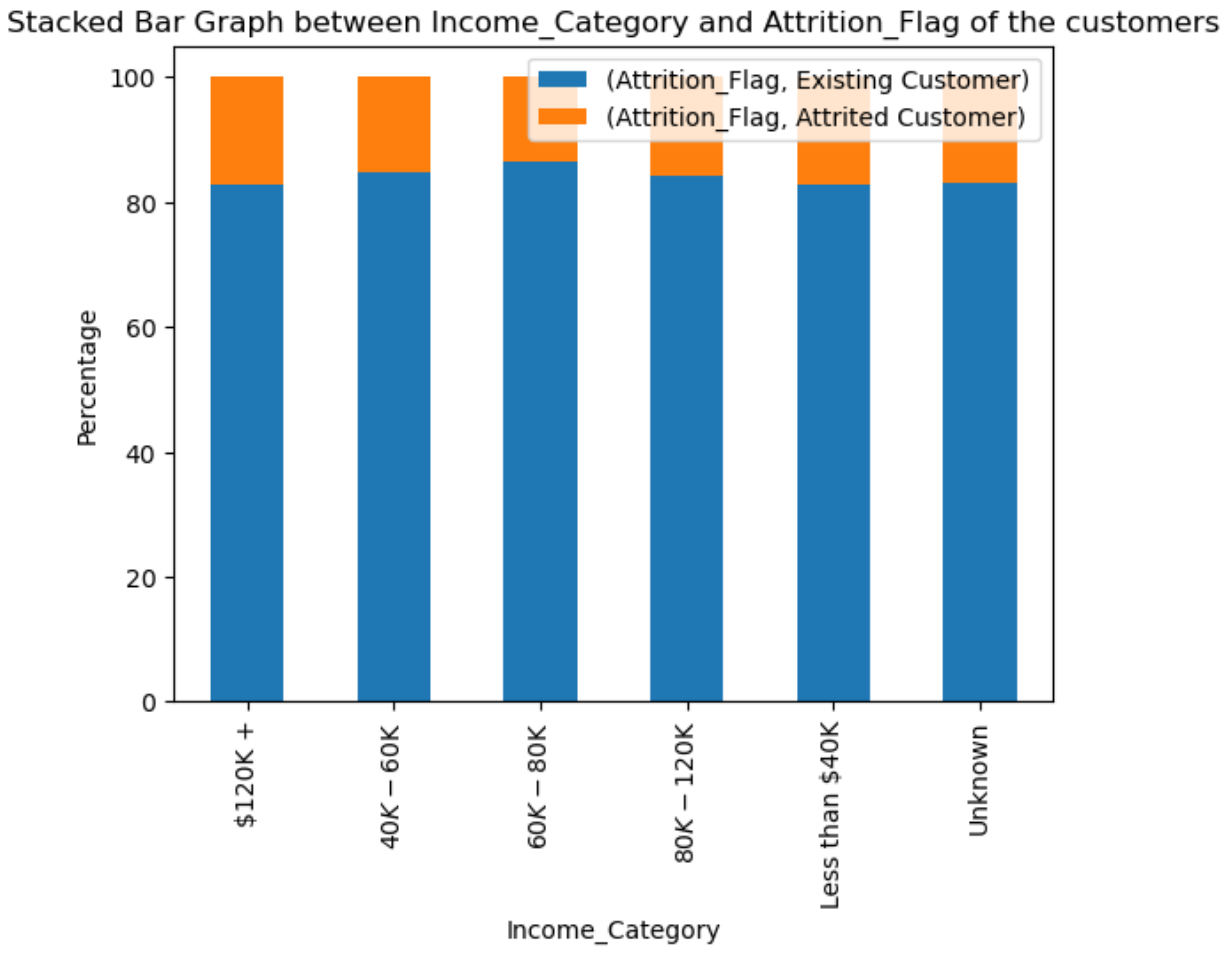
Proportional Breakdown: Within each bar, the individual segments represent the proportions of the different categories. By examining the lengths of these segments, you can understand the relative contribution of each category to the total value.

This is the stacked bar graph between education level and attrition flag of the customer.

EDUCATION LEVEL	PERCENTAGE OF ATTRITED CUSTOMER	PERCENTAGE OF EXISTING CUSTOMER
College	Less than 20%	More than 80% and less than 85%
doctorate	22%	78%
graduate	More than 80% and less than 85%	More than 80% and less than 85%
High school	More than 80% and less than 85%	More than 80% and less than 85%

Post graduate	20%	80%
uneducated	More than 80% and less than 85%	More than 80% and less than 85%
unknown	20%	80%

6. Plot a percentage segment bar graph between Income_Category and Attrition_Flag of the customers.



ANALYSIS:

INCOME CATEGORY	PERCENTAGE OF ATTRITED CUSTOMER	PERCENTAGE OF EXISTING CUSTOMER
120K+ $	22%	78%
40K - 60k $	23%	77%
60K - 80k $	24%	76%
80K - 120k $	21%	79%
Less than 40K $	20%	80%
unknown	20%	80%

7.Drop CLIENTNUM column.

When conducting a drop column analysis, you are examining the impact of removing or dropping a particular column from your dataset.

In this we dropped the CLIENT NUM column.

ANALYSIS:

Hence, table was look like this:

]:

	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_
0	Existing Customer	45	M	3	High School	Married	$60K-80K$	Blue	39	
1	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	
2	Existing Customer	51	M	3	Graduate	Married	$80K-120K$	Blue	36	
3	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	
4	Existing Customer	40	M	3	Uneducated	Married	$60K-80K$	Blue	21	
...	...	...	...	...	...	...	...	...	...	
10122	Existing Customer	50	M	2	Graduate	Single	$40K-60K$	Blue	40	
10123	Attrited Customer	41	M	2	Unknown	Divorced	$40K-60K$	Blue	25	
10124	Attrited Customer	44	F	1	High School	Married	Less than $40K	Blue	36	
10125	Attrited Customer	30	M	2	Graduate	Unknown	$40K-60K$	Blue	36	
10126	Attrited Customer	43	F	2	Graduate	Married	Less than $40K	Silver	25	

]:

Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1
5	1	3	12691.0	777	11914.0	1.335
6	1	2	8256.0	864	7392.0	1.541
4	1	0	3418.0	0	3418.0	2.594
3	4	1	3313.0	2517	796.0	1.405
5	1	0	4716.0	0	4716.0	2.175
...	...	...	...	...	...	...
3	2	3	4003.0	1851	2152.0	0.703
4	2	3	4277.0	2186	2091.0	0.804
5	3	4	5409.0	0	5409.0	0.819
4	3	3	5281.0	0	5281.0	0.535
6	2	4	10388.0	1961	8427.0	0.703

:[19]:

Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
12691.0	777	11914.0	1.335	1144	42	1.625	0.061
8256.0	864	7392.0	1.541	1291	33	3.714	0.105
3418.0	0	3418.0	2.594	1887	20	2.333	0.000
3313.0	2517	796.0	1.405	1171	20	2.333	0.760
4716.0	0	4716.0	2.175	816	28	2.500	0.000
...	...	...	...	...	...	...	...
4003.0	1851	2152.0	0.703	15476	117	0.857	0.462
4277.0	2186	2091.0	0.804	8764	69	0.683	0.511
5409.0	0	5409.0	0.819	10291	60	0.818	0.000
5281.0	0	5281.0	0.535	8395	62	0.722	0.000
10388.0	1961	8427.0	0.703	10294	61	0.649	0.189

Make a sub data frame which consists of all the numerical columns(i.e.int64,float64) along with the Attrition_Flag column.

TASK: Here we did the task of removing all data contained string except the Attrition flag and only we have left with numeric data including attrition flag column.

Hence the table was look like this:

Analysis:

22]:

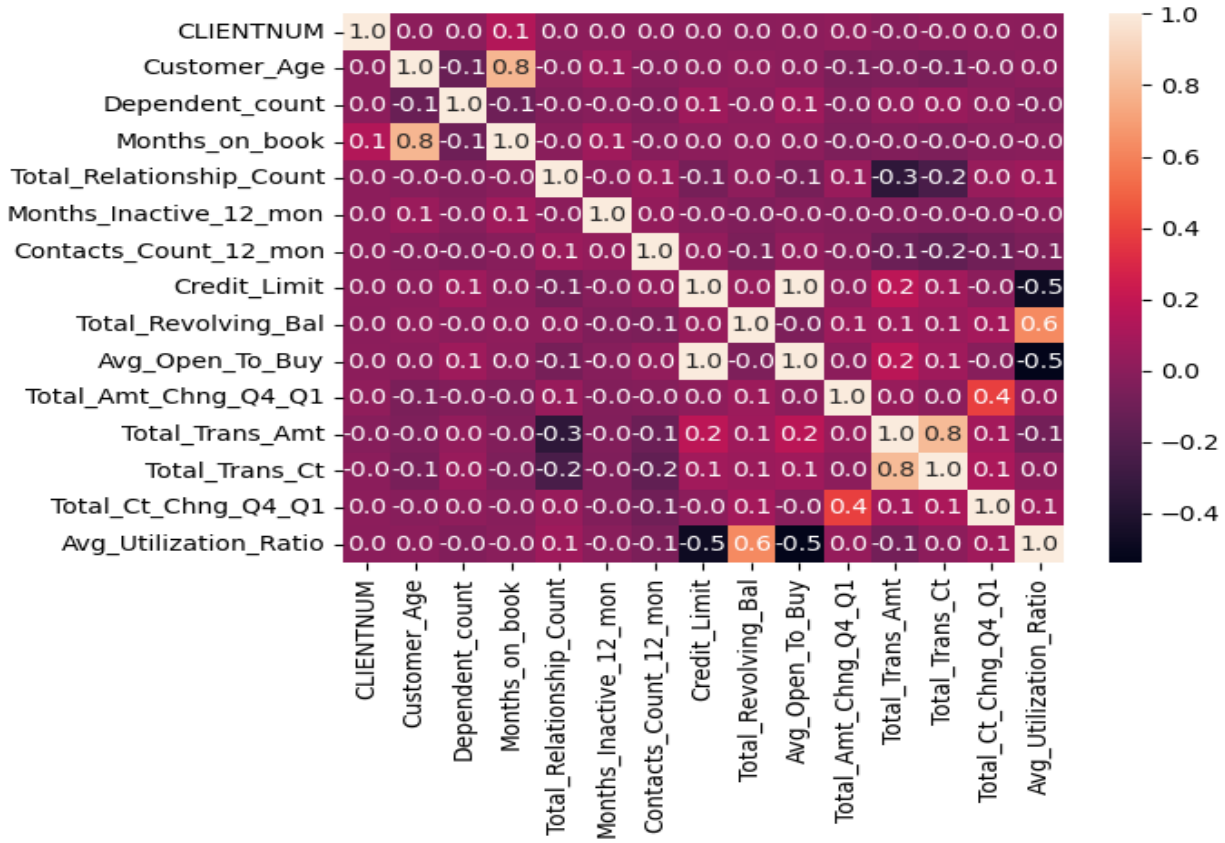
	Attrition_Flag	Customer_Age	Dependent_count	Marital_Status	Months_on_book	Total_Relationship_Count	Months_Inactive_12_mon	Contacts_Cour
0	Existing Customer	45	3	Married	39	5	1	
1	Existing Customer	49	5	Single	44	6	1	
2	Existing Customer	51	3	Married	36	4	1	
3	Existing Customer	40	4	Unknown	34	3	4	
4	Existing Customer	40	3	Married	21	5	1	
...	...	...	...	...	...	...	...	
10122	Existing Customer	50	2	Single	40	3	2	
10123	Attrited Customer	41	2	Divorced	25	4	2	
10124	Attrited Customer	44	1	Married	36	5	3	
10125	Attrited Customer	30	2	Unknown	36	4	3	
10126	Attrited Customer	43	2	Married	25	6	2	

Contacts_Count_12_mon	Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_(
3	12691.0	777	11914.0	1.335	1144	42	1.6
2	8256.0	864	7392.0	1.541	1291	33	3.7
0	3418.0	0	3418.0	2.594	1887	20	2.3
1	3313.0	2517	796.0	1.405	1171	20	2.3
0	4716.0	0	4716.0	2.175	816	28	2.5
...	...	...	...	...	...	...	
3	4003.0	1851	2152.0	0.703	15476	117	0.8
3	4277.0	2186	2091.0	0.804	8764	69	0.6
4	5409.0	0	5409.0	0.819	10291	60	0.8
3	5281.0	0	5281.0	0.535	8395	62	0.7
4	10388.0	1961	8427.0	0.703	10294	61	0.6

Credit_Limit	Total_Revolving_Bal	Avg_Open_To_Buy	Total_Amt_Chng_Q4_Q1	Total_Trans_Amt	Total_Trans_Ct	Total_Ct_Chng_Q4_Q1	Avg_Utilization_Ratio
12691.0	777	11914.0	1.335	1144	42	1.625	0.061
8256.0	864	7392.0	1.541	1291	33	3.714	0.105
3418.0	0	3418.0	2.594	1887	20	2.333	0.000
3313.0	2517	796.0	1.405	1171	20	2.333	0.760
4716.0	0	4716.0	2.175	816	28	2.500	0.000
...	...	...	...	...	...	...	...
4003.0	1851	2152.0	0.703	15476	117	0.857	0.462
4277.0	2186	2091.0	0.804	8764	69	0.683	0.511
5409.0	0	5409.0	0.819	10291	60	0.818	0.000
5281.0	0	5281.0	0.535	8395	62	0.722	0.000
10388.0	1961	8427.0	0.703	10294	61	0.649	0.189

Plot a clear heatmap to view the correlation using seaborn.

Here we are Plotting a heatmap to view the correlation between each column of Dataframe



Analysis:

Analyze the resulting heatmap to understand the correlation relationships between variables. The color intensity and the correlation coefficient values can provide insights into the strength and direction of the correlations. Positive correlations are represented by warmer colors (e.g., red), while negative correlations are represented by cooler

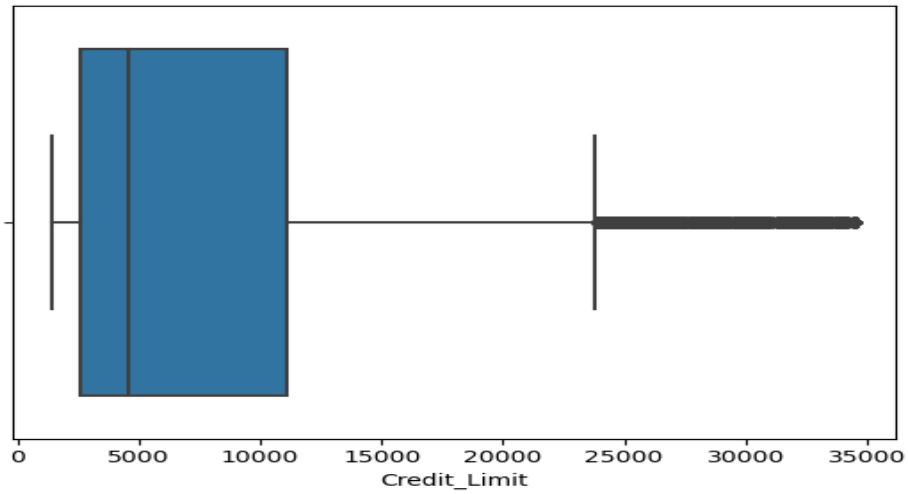
colors (e.g., blue). The stronger the correlation, the darker or more intense the color will be.

By using a heatmap, you can quickly identify variables that are strongly correlated (either positively or negatively) and identify potential patterns or relationships within your dataset. It provides a visual summary of the correlation structure, making it easier to spot trends and make data-driven decisions.

8. Plot a boxplot for the Credit_Limit column and check if it contains any outlier or not

Task:

- Plotting a Box Plot for Credit_Limit column .
- Checking for the outliers present in the Box Plot of Credit_Limit column.



Analysis:

By analyzing the box plot for the credit limit column, we can gain insights into the spread of the data and presence of any outlier in the data.

This analysis helps in understanding the distribution of credit limits and identifying any unexpected or extreme values that may need attention in further analysis.

OUTLIER IS PRESENT.
There is an outlier in Credit_limit which is from less than 20k to 35k.

9. Map the Attrition_Flag values to 0 and 1(i.e. Existing Customer=0 and Attrited Customer=1. Standardize the columns.

Tasks :

- Mapping the values of Attrition_Flag column to Existing Customer=0 and Attrited Customer=1 .
- After mapping the values now , we are standardizing the columns of the Dataframe .

Analysis :

Before mapping the value table looks like this:

8]:

	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_on_book	Total_
0	Existing Customer	45	M	3	High School	Married	$60K - 80K$	Blue	39	
1	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	
2	Existing Customer	51	M	3	Graduate	Married	$80K - 120K$	Blue	36	
3	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	
4	Existing Customer	40	M	3	Uneducated	Married	$60K - 80K$	Blue	21	
...	...	...	...	...	...	...	...	...	...	
10122	Existing Customer	50	M	2	Graduate	Single	$40K - 60K$	Blue	40	
10123	Attrited Customer	41	M	2	Unknown	Divorced	$40K - 60K$	Blue	25	
10124	Attrited Customer	44	F	1	High School	Married	Less than $40K	Blue	36	
10125	Attrited Customer	30	M	2	Graduate	Unknown	$40K - 60K$	Blue	36	
10126	Attrited Customer	43	F	2	Graduate	Married	Less than $40K	Silver	25	

10127 rows × 20 columns

After mapping the values the table looks like this :

ut[53]:

	CLIENTNUM	Attrition_Flag	Customer_Age	Gender	Dependent_count	Education_Level	Marital_Status	Income_Category	Card_Category	Months_o
0	768805383	0	45	M	3	High School	Married	$60K – 80K	Blue	
1	818770008	0	49	F	5	Graduate	Single	Less than $40K	Blue	
2	713982108	0	51	M	3	Graduate	Married	$80K – 120K	Blue	
3	769911858	0	40	F	4	High School	Unknown	Less than $40K	Blue	
4	709106358	0	40	M	3	Uneducated	Married	$60K – 80K	Blue	
...	...	...	...	...	...	...	...	...	...	
10122	772366833	0	50	M	2	Graduate	Single	$40K – 60K	Blue	
10123	710638233	1	41	M	2	Unknown	Divorced	$40K – 60K	Blue	
10124	716506083	1	44	F	1	High School	Married	Less than $40K	Blue	
10125	717406983	1	30	M	2	Graduate	Unknown	$40K – 60K	Blue	
10126	714337233	1	43	F	2	Graduate	Married	Less than $40K	Silver	

10127 rows × 21 columns

Now here we are standardizing the columns of the Dataframe

CLIENTNUM	3.690378e+07
Customer_Age	8.016814e+00
Dependent_count	1.298908e+00
Months_on_book	7.986416e+00
Total_Relationship_Count	1.554408e+00
Months_Inactive_12_mon	1.010622e+00
Contacts_Count_12_mon	1.106225e+00
Credit_Limit	9.088777e+03
Total_Revolving_Bal	8.149873e+02
Avg_Open_To_Buy	9.090685e+03
Total_Amt_Chng_Q4_Q1	2.192068e-01
Total_Trans_Amt	3.397129e+03
Total_Trans_Ct	2.347257e+01
Total_Ct_Chng_Q4_Q1	2.380861e-01
Avg_Utilization_Ratio	2.756915e-01