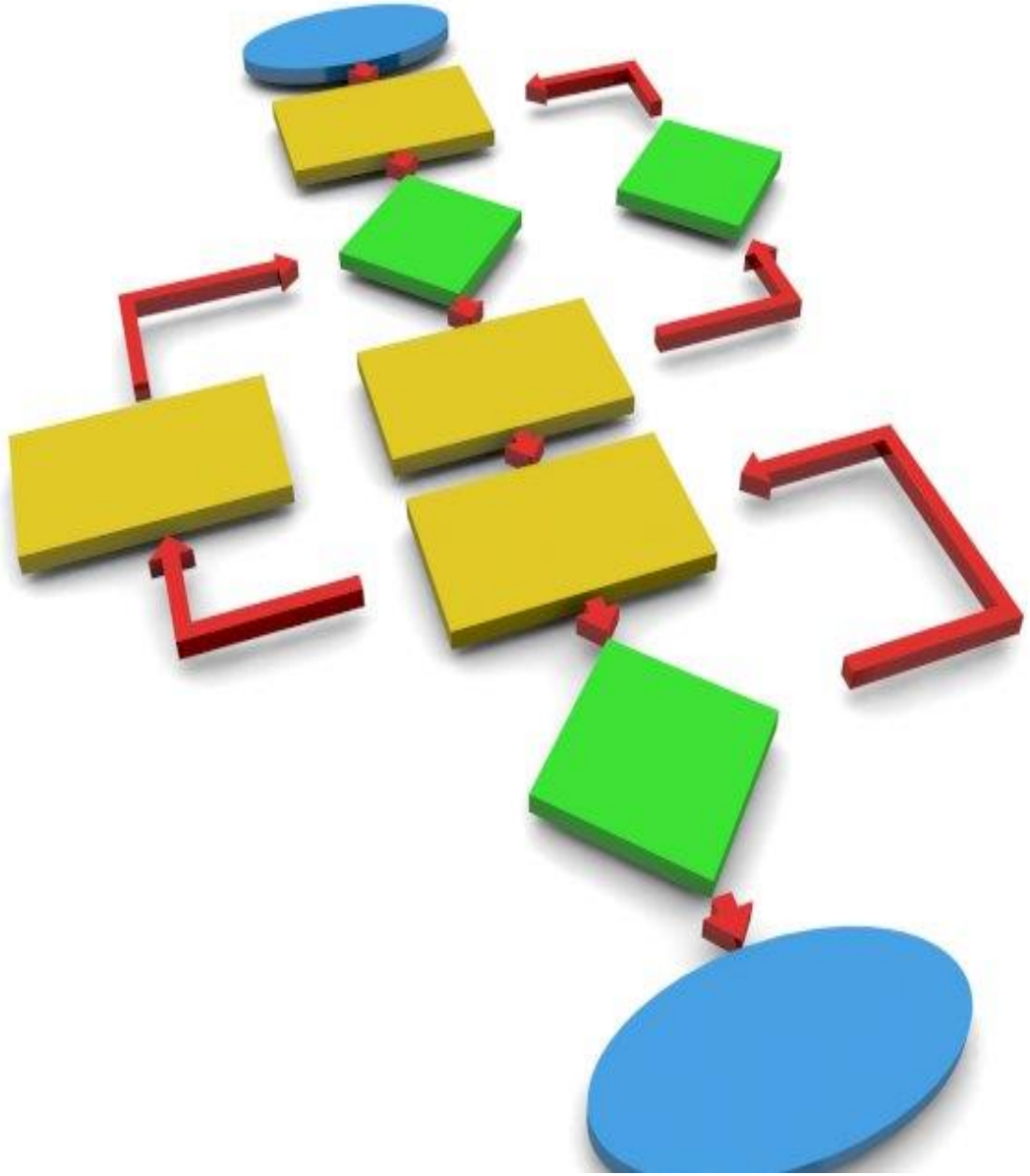
Project Team - Green

Phase-1 Exploratory Data Analysis

Team Members-

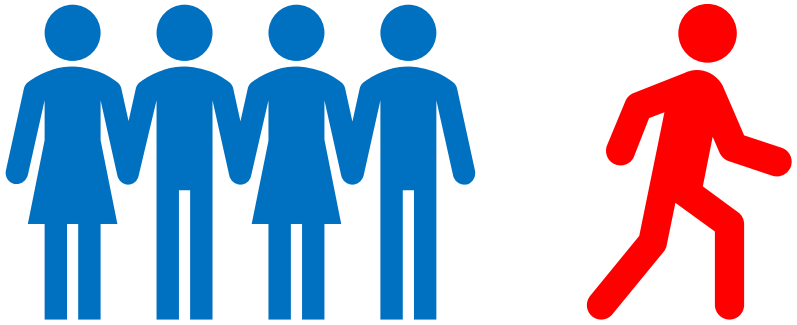
- Yuraja Kadari
- Yash Jivani
- Ameya Kalbande
- Vicky Jadhav



Agenda

- EDA Problem Statement
- Goals of EDA
- Features Selection
- Introduction to Dataset
- Yuraja’s Analysis
- Yash’s Analysis
- Ameya’s Analysis
- Vicky’s Analysis
- Summary

EDA Problem Statement



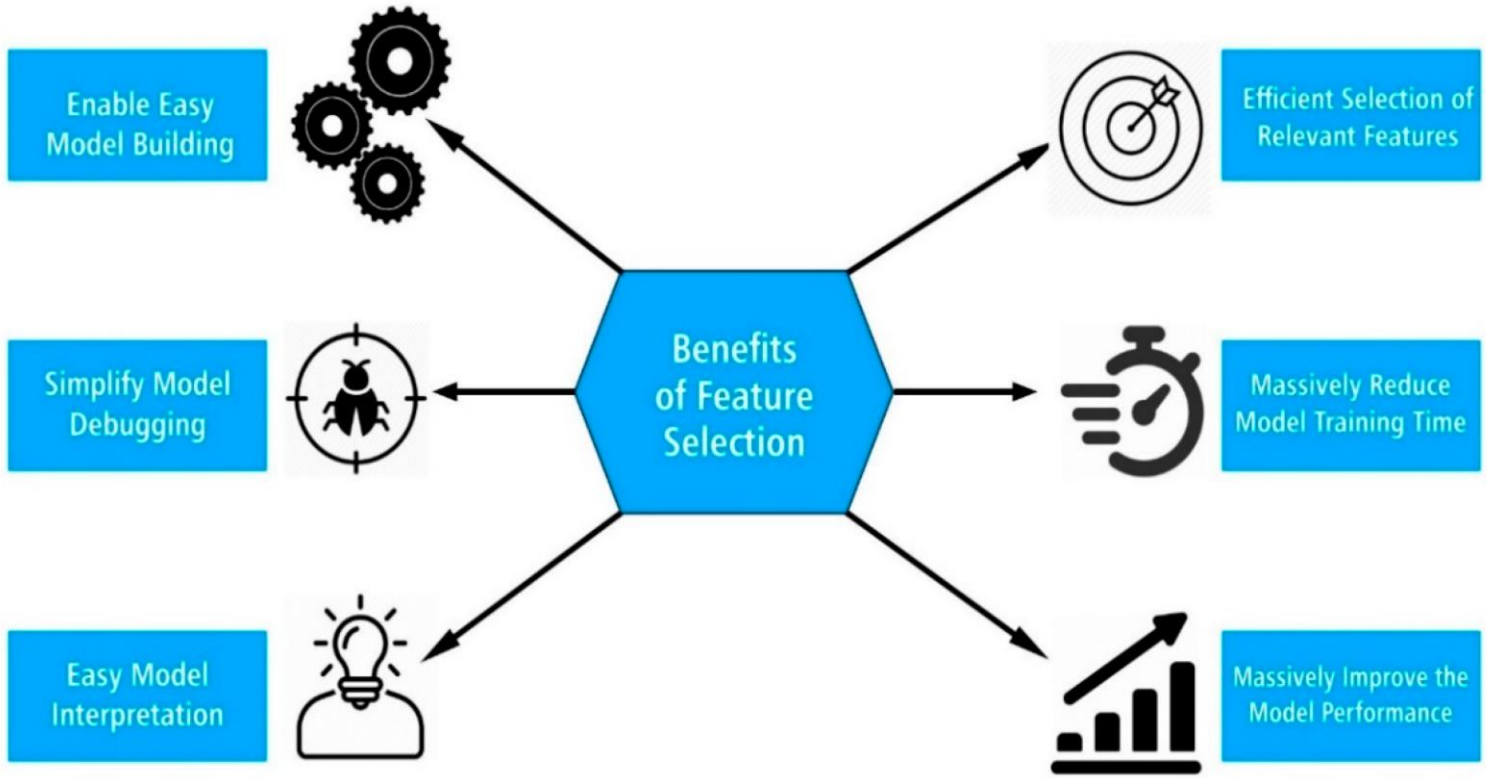
- The CEO of Very Nice Bank is concerned about customer attrition in their credit card services and wants to proactively address the issue by predicting which customers are most likely to cancel their accounts.
- How can we address the issue of customer attrition in the credit card services provided by Very Nice Bank Inc?

Goals of EDA

- The objective is to offer an initial collection of predictor variables and machine learning models that could be employed for predictive modeling.
- This would involve identifying a set of features that may potentially have a significant impact on predicting customer attrition.
- Additionally, exploring various machine learning techniques that could be utilized to generate accurate predictions based on the identified features.
- The end goal is to create a reliable and effective model that can be used to predict which customers are likely to cancel their credit card services with Very Nice Bank.

Features Selection for analysis.

- Jadhav Vicky –Contact Count In 12 Month, Dependent Count
- Jivani Yash – Total Relationship Count, Total Ct Change Q4-Q1
- Kadari Yuraja - Total transaction Count, Total Revolving Balance.
- Kalbande Ameya - Gender, Months Inactive 12 mon.



Introduction to Bank Dataset

- The dataset comprises of 5,998 customers out of a total of 3,100,111 credit card customers, and it includes 21 features.
- One of the features is a flag variable indicating whether a customer has attrited or not, meaning whether they have ceased business with the bank.
- There are 5 categorical variables and 16 numerical variables.

```
Bank_churners<-read_csv("C:/Users/Lenovo/Downloads/BankChurners.csv")
glimpse(Bank_churners)
Bank_churners%>%head()
summary(Bank_churners)
skim(Bank_churners)
```

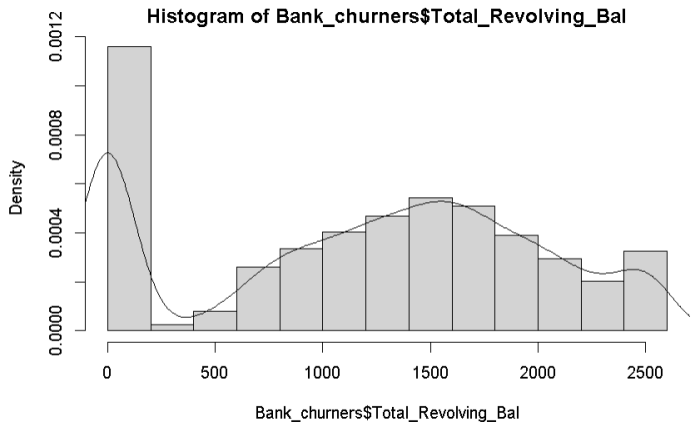
```
Rows: 5,998
Columns: 21
$ CLIENTNUM                 <dbl> 768805383, 818770008, 713982108, 769911858, 709106358, 713061558, 810347208, 818906208, 710930…
$ Attrition_Flag            <chr> "Existing Customer", "Existing Customer", "Existing Customer", "Existing Customer", "Existing …
$ Customer_Age              <dbl> 45, 49, 51, 40, 40, 44, 51, 32, 37, 48, 42, 65, 56, 35, 57, 44, 48, 41, 61, 45, 47, 62, 41, 47…
$ Gender                    <chr> "M", "F", "M", "F", "M", "M", "M", "M", "M", "M", "M", "M", "M", "M", "F", "M", "M", "M", "M",…
$ Dependent_count           <dbl> 3, 5, 3, 4, 3, 2, 4, 0, 3, 2, 5, 1, 1, 3, 2, 4, 4, 3, 1, 2, 1, 0, 3, 4, 2, 3, 1, 1, 3, 4, 3, 2…
$ Education_Level           <chr> "High School", "Graduate", "Graduate", "High School", "Uneducated", "Graduate", "Unknown", "Hi…
$ Marital_Status            <chr> "Married", "Single", "Married", "Unknown", "Married", "Married", "Married", "Unknown", "Single…
$ Income_Category           <chr> "$60K - $80K", "Less than $40K", "$80K - $120K", "Less than $40K", "$60K - $80K", "$40K - $60K…
$ Card_Category             <chr> "Blue", "Blue", "Blue", "Blue", "Blue", "Blue", "Gold", "Silver", "Blue", "Blue", "Blue", "Blu…
$ Months_on_book            <dbl> 39, 44, 36, 34, 21, 36, 46, 27, 36, 36, 31, 54, 36, 30, 48, 37, 36, 34, 56, 37, 42, 49, 33, 36…
$ `Total_Relationship Count` <dbl> 5, 6, 4, 3, 5, 3, 6, 2, 5, 6, 5, 6, 3, 5, 5, 5, 6, 4, 2, 6, 5, 2, 4, 3, 4, 6, 4, 3, 5, 6, 3, 2…
$ Months_Inactive_12_mon    <dbl> 1, 1, 1, 4, 1, 1, 1, 2, 2, 3, 3, 2, 6, 1, 2, 1, 2, 4, 2, 1, 2, 3, 2, 3, 2, 1, 1, 3, 2, 0, 2, 5…
$ Contacts_Count_12_mon     <dbl> 3, 2, 0, 1, 0, 2, 3, 2, 0, 3, 2, 3, 0, 3, 2, 2, 3, 1, 3, 2, 0, 3, 1, 2, 3, 2, 2, 2, 2, 0, 3, 1…
$ Credit_Limit              <dbl> 12691.0, 8256.0, 3418.0, 3313.0, 4716.0, 4010.0, 34516.0, 29081.0, 22352.0, 11656.0, 6748.0, 9…
$ Total_Revolving_Bal       <dbl> 777, 864, 0, 2517, 0, 1247, 2264, 1396, 2517, 1677, 1467, 1587, 0, 1666, 680, 972, 2362, 1291,…
$ Avg_Open_To_Buy           <dbl> 11914.0, 7392.0, 3418.0, 796.0, 4716.0, 2763.0, 32252.0, 27685.0, 19835.0, 9979.0, 5281.0, 750…
$ Total_Amt_Chng_Q4_Q1      <dbl> 1.335, 1.541, 2.594, 1.405, 2.175, 1.376, 1.975, 2.204, 3.355, 1.524, 0.831, 1.433, 3.397, 1.1…
$ Total_Trans_Amt           <dbl> 1144, 1291, 1887, 1171, 816, 1088, 1330, 1538, 1350, 1441, 1201, 1314, 1539, 1311, 1570, 1348,…
$ Total_Trans_Ct            <dbl> 42, 33, 20, 20, 28, 24, 31, 36, 24, 32, 42, 26, 17, 33, 29, 27, 27, 21, 30, 21, 27, 16, 18, 23…
$ Total_Ct_Chng_Q4_Q1       <dbl> 1.625, 3.714, 2.333, 2.333, 2.500, 0.846, 0.722, 0.714, 1.182, 0.882, 0.680, 1.364, 3.250, 2.0…
$ Avg_Utilization_Ratio     <dbl> 0.061, 0.105, 0.000, 0.760, 0.000, 0.311, 0.066, 0.048, 0.113, 0.144, 0.217, 0.174, 0.000, 0.1…
```


YURAJA SOHEL KADARI

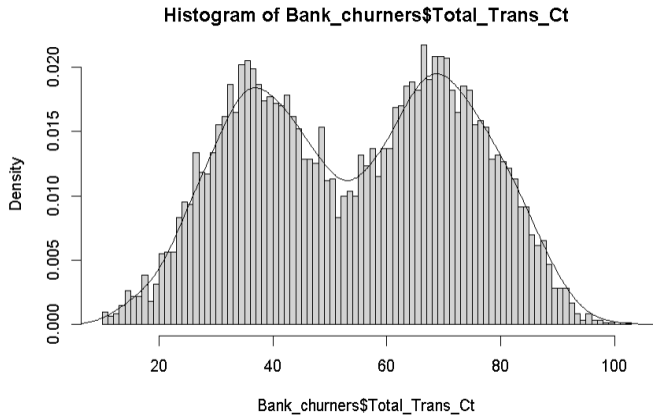
Data Pre-processing

- Upon exploring the dataset, it was found that there were no missing values and all variables had appropriate data types.
- However, some of the column names had blank spaces, which were replaced with underscores using the replace function in R.
- Currently, the variables Total_revolving_bal and Total_transaction_ct have been selected for analysis. Upon checking their skewness, it was observed that they were normally distributed.

```r
hist(Bank_churners$Total_Revolving_Bal,prob=TRUE,breaks=9)
lines(density(Bank_churners$Total_Revolving_Bal))
```




```r
hist(Bank_churners$Total_Trans_Ct,prob=TRUE,breaks=100)
lines(density(Bank_churners$Total_Trans_Ct))
```



	skim_variable <chr>	n_missing <int>	complete_rate <dbl>	mean <dbl>	sd <dbl>	p0 <dbl>	p25 <dbl>
1	CLIENTNUM	0	1	7.397950e+08	3.724549e+07	708083283.0	7.130958e+08
2	Customer_Age	0	1	4.637679e+01	8.518824e+00	26.0	4.000000e+01
3	Dependent_count	0	1	2.277593e+00	1.306390e+00	0.0	1.000000e+00
4	Months_on_book	0	1	3.598166e+01	8.397118e+00	13.0	3.100000e+01
5	Total_Relationship Count	0	1	4.309770e+00	1.270124e+00	1.0	3.000000e+00
6	Months_Inactive_12_mon	0	1	2.299266e+00	9.910233e-01	0.0	2.000000e+00
7	Contacts_Count_12_mon	0	1	2.608369e+00	1.151306e+00	0.0	2.000000e+00
8	Credit_Limit	0	1	8.713911e+03	8.814795e+03	1438.3	2.644000e+03
9	Total_Revolving_Bal	0	1	1.183529e+03	8.035394e+02	0.0	5.805000e+02
10	Avg_Open_To_Buy	0	1	7.530382e+03	8.848616e+03	3.0	1.368500e+03

1-10 of 15 rows | 1-8 of 11 columns

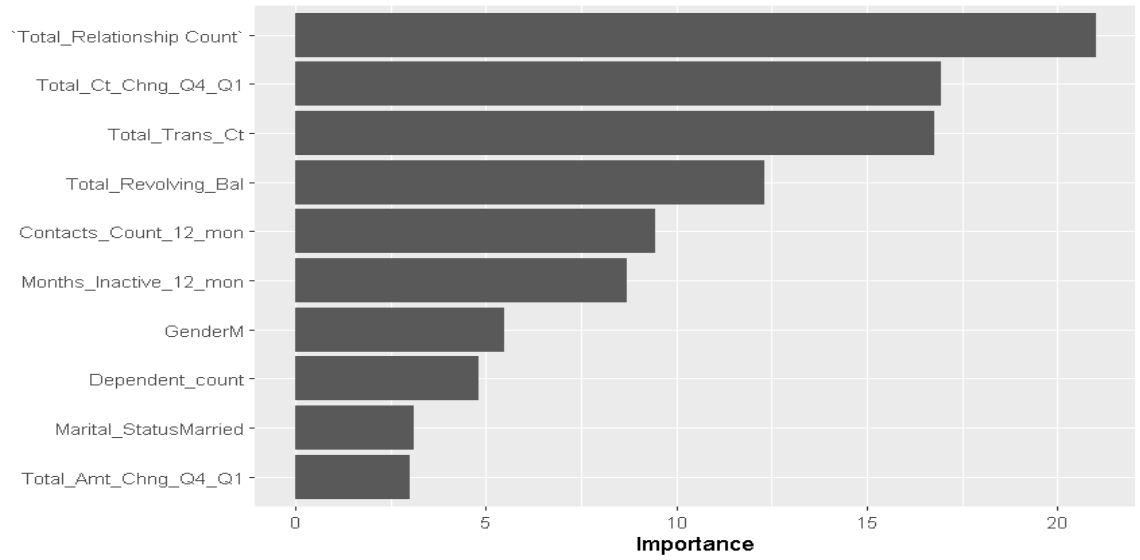
Previous 1 2 Next

YURAJA SOHEL KADARI

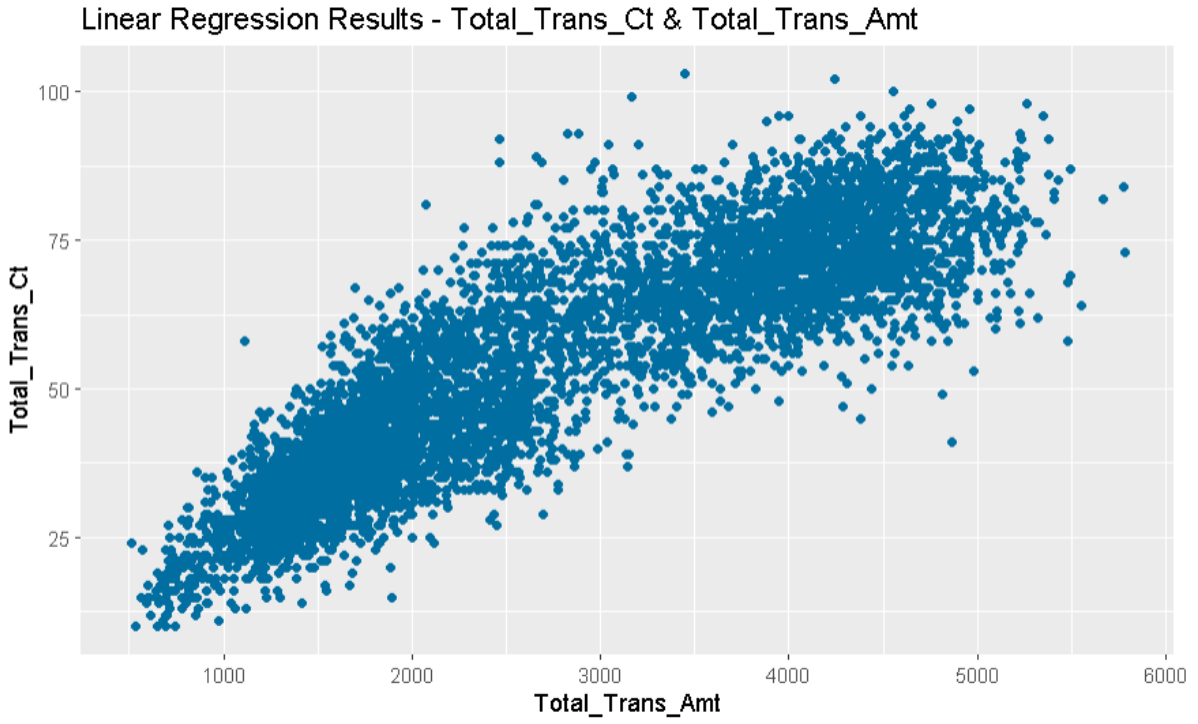
How and Why this features is selected ?

- The feature which I selected is Total transaction Count & Total Revolving Balance.
- So, while selecting the features I have used all the features and created a linear model and through that model Using VIP function it shows the important variables and its level of Importance on our Predictor variable.
- As the upper variables is been selected by other team members so I have selected the given features and there are many reasons for which I have landed on this variables which I have shown in Further slides.

```r
vip(lm_fit)
```



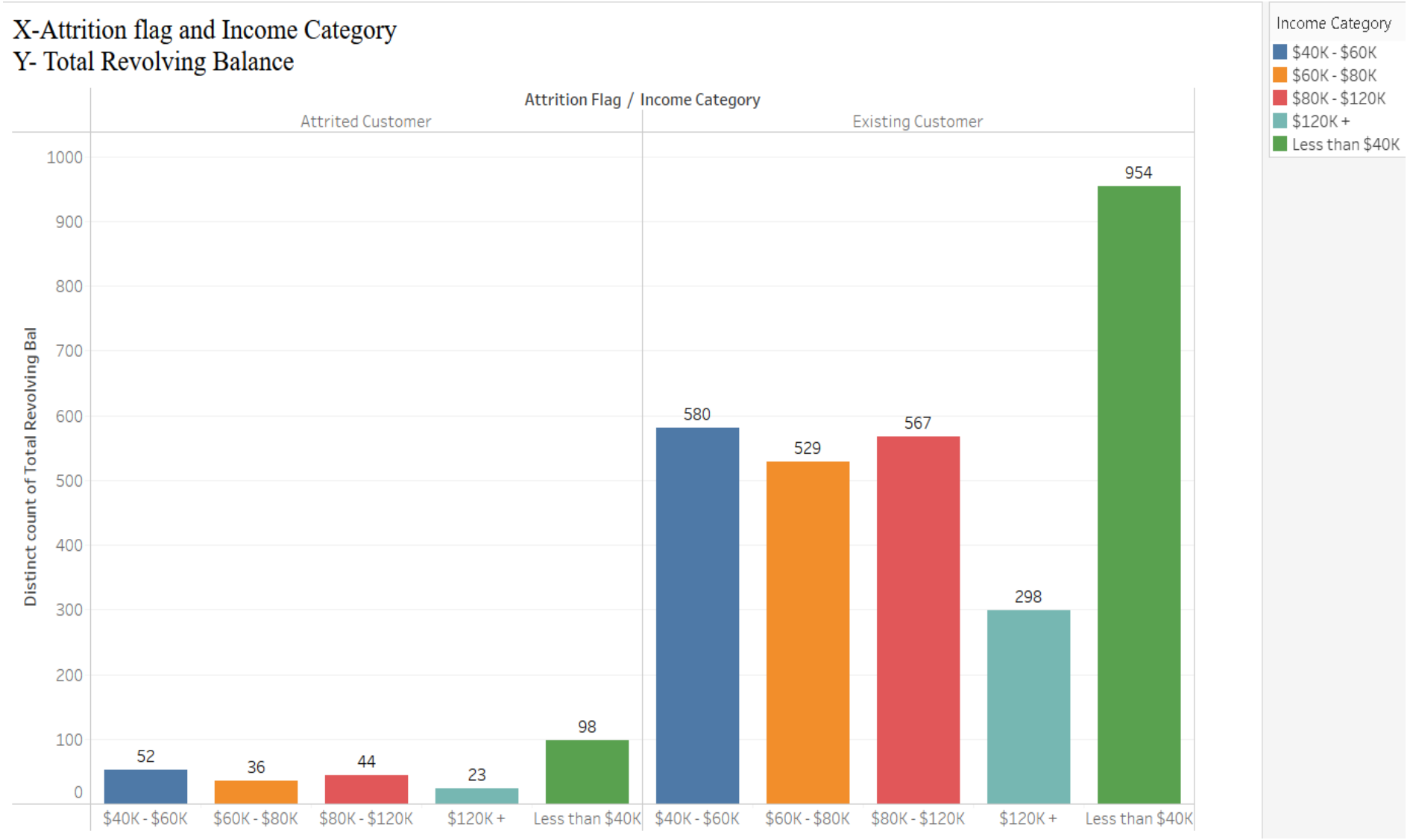

```r
ggplot(data = Bank_churners,
       mapping = aes(x = Total_Trans_Amt, y = Total_Trans_Ct)) +
  geom_point(color = '#006EA1') +
  geom_abline(intercept = 0, slope = 1, color = 'orange') +
  labs(title = 'Linear Regression Results - Total_Trans_Ct & Total_Trans_Amt',
       x = 'Total_Trans_Amt',
       y = 'Total_Trans_Ct')
```



YURAJA SOHEL KADARI

What is Total Revolving Balance and does customers with higher revolving balances are more likely to discontinue using the bank's credit card services?

- Total Revolving Balance refers to the outstanding balance on a credit card account that carries over from one billing period to the next, on which interest is charged.
- The lower your revolving balances, the better your Credit score and a very low balance will keep your financial risks low.
- As you can see in the visualization customers with higher income have lower revolving balance and vice versa and there are a greater number of customer with higher revolving balance have left the services.
- Therefore, we can say that **customers with higher revolving balances are more likely to discontinue using the bank's credit card services.**

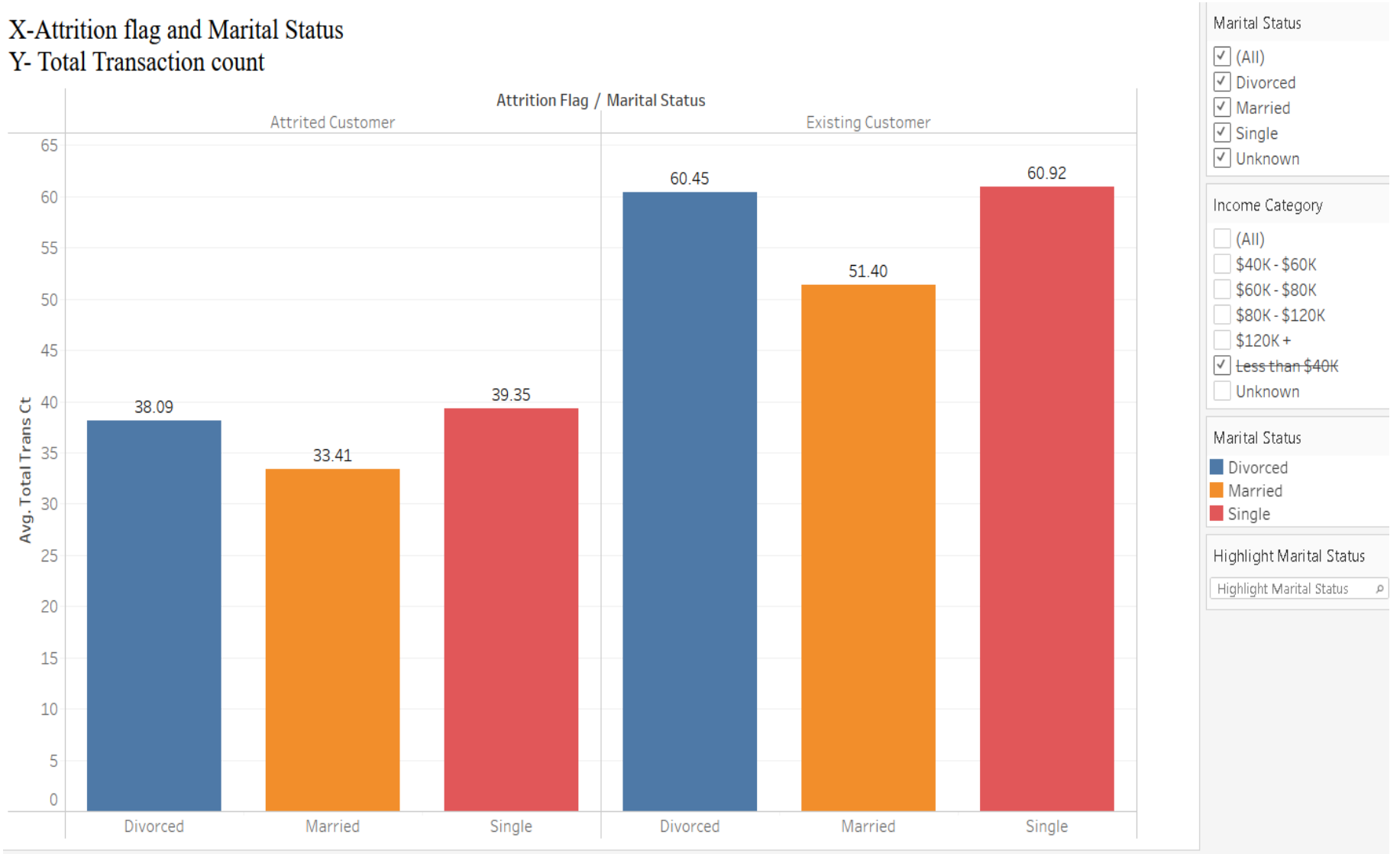


YURAJA SOHEL KADARI

Which Marital status of customer are not leaving our Credit card services on basis of there Transaction count? What is Total Transaction Count?

- Total Transaction Count refers to the total number of transactions made by a customer with their credit card account during a given period.
- In the visualization we can say Married customer are less who is leaving the services.
- We have taken Average of total transaction count.
- Single and Divorced customer are more likely to leave the services based on there transaction count.
- As Total transaction count determines more number of transaction so the customer will stay with the services

X-Attrition flag and Marital Status
Y- Total Transaction count

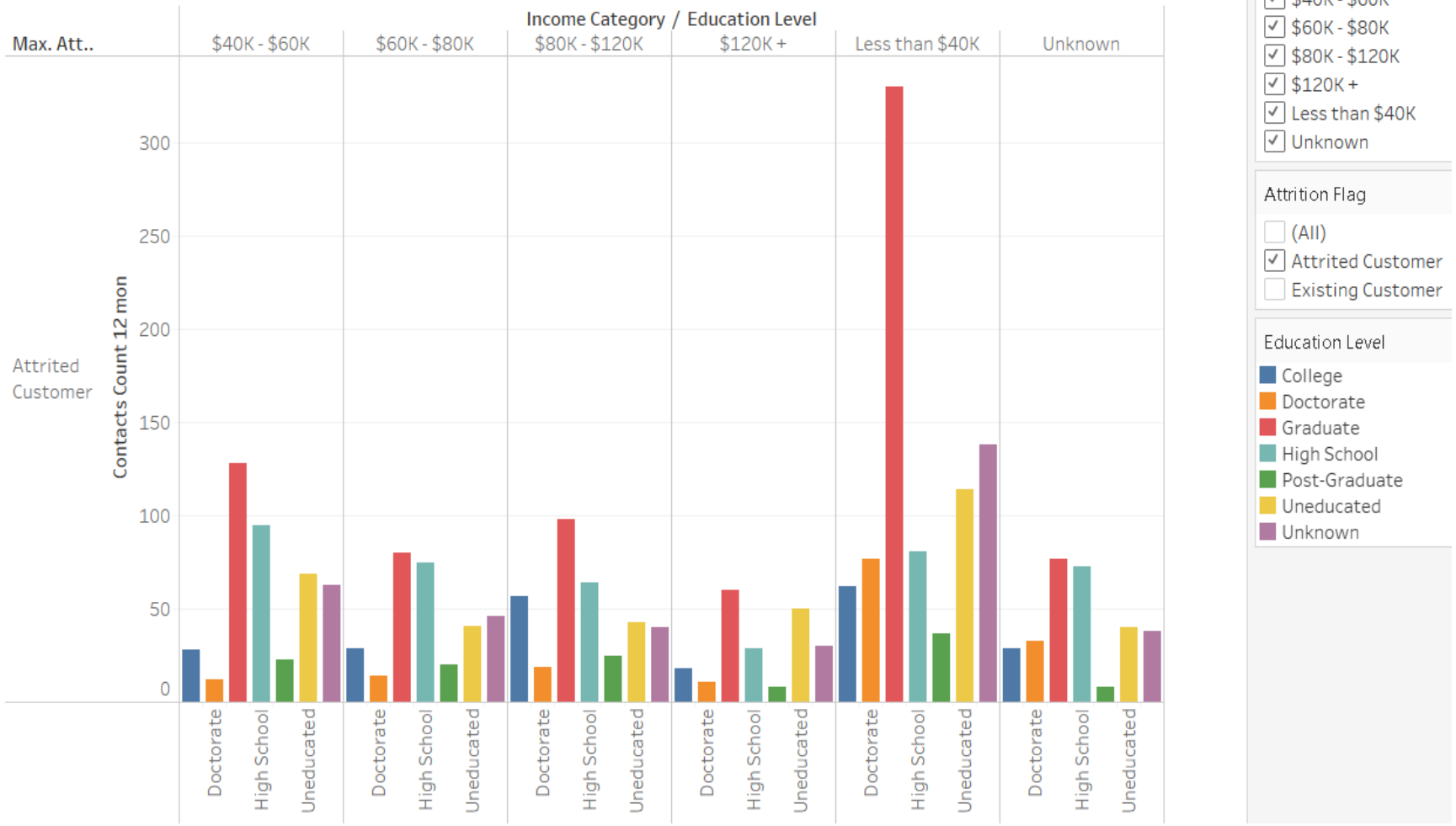


YURAJA SOHEL KADARI

Which salary range customer are more likely to leave the services and what are their Education status ?

- In the given Visualizations you can see Customers who are graduate level and having salary range less than $40K is more like leaves the services.
- We have predicted this by using contact count 12 months.
- Contact count means how many times bank have contacted the customers.
- The outcomes were same using other features like total transaction count.

X-Education Level and Income Category
Y- Contacts count 12 Months and Attrited customer



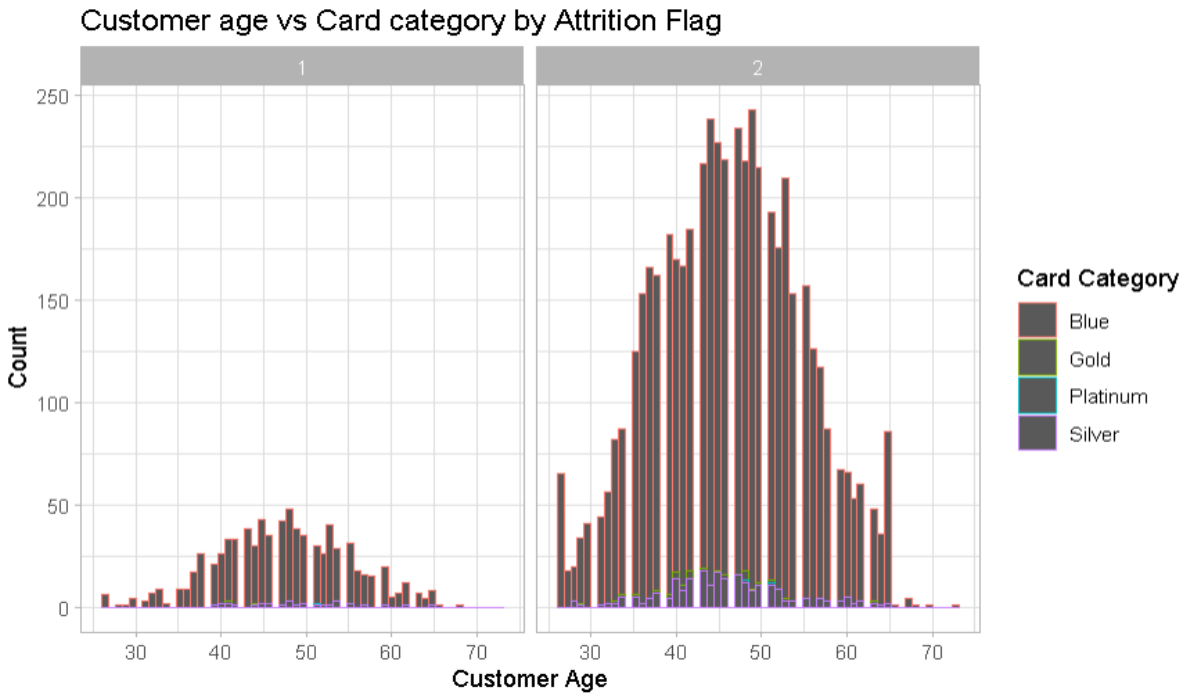
YURAJA SOHEL KADARI

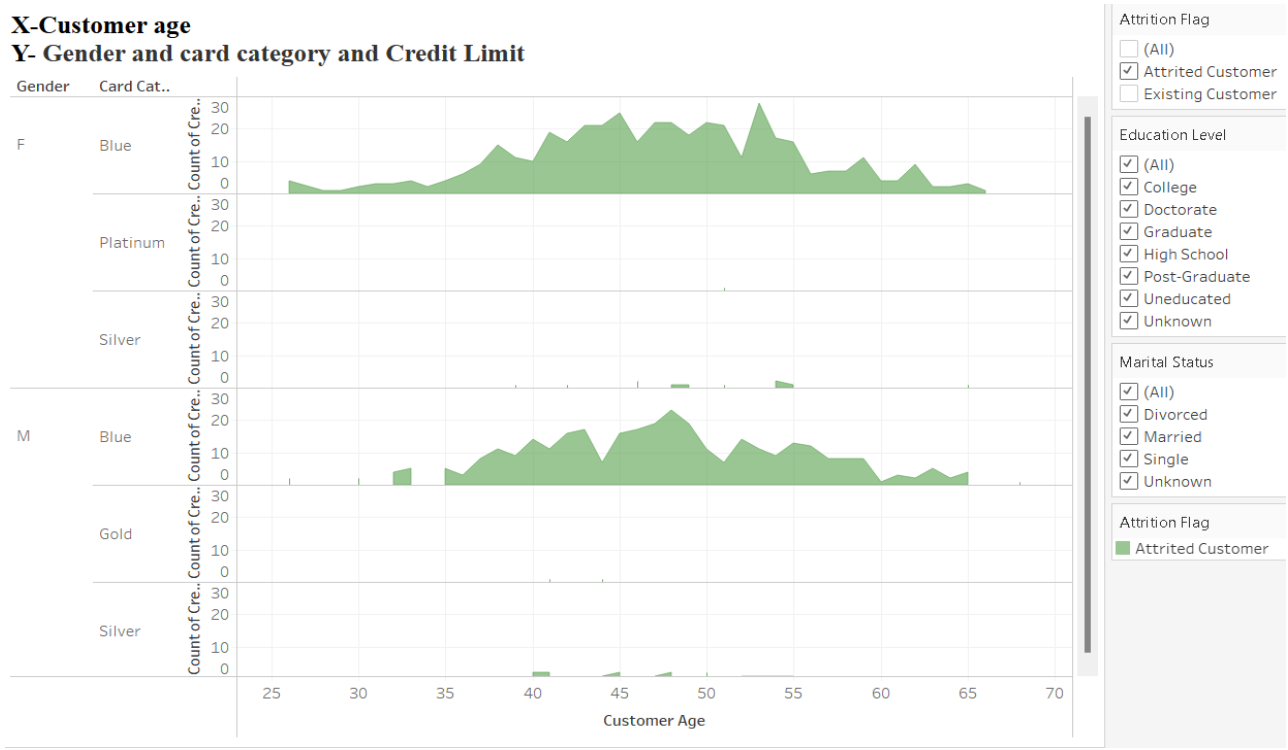
Which age group customers are leaving the services ?

As you can see in the below visualization customer having age group between 45 to 50 having high chances to leave the Services.

The code for getting the visualization is given :

```
ggplot(data = Bank_churners, mapping = aes(x = Customer_Age,color = Card_Category)) +geom_histogram(binwidth = 0.8) +facet_wrap(~ Card_Category) +labs(title ="Customer age vs Card category by Attrition Flag",x = "Customer Age",y = "Count") + scale_color_discrete(name = "Card Category") +theme_light() +facet_wrap(~ Card_Category) +facet_grid(. ~ Attrition_Flag)
```

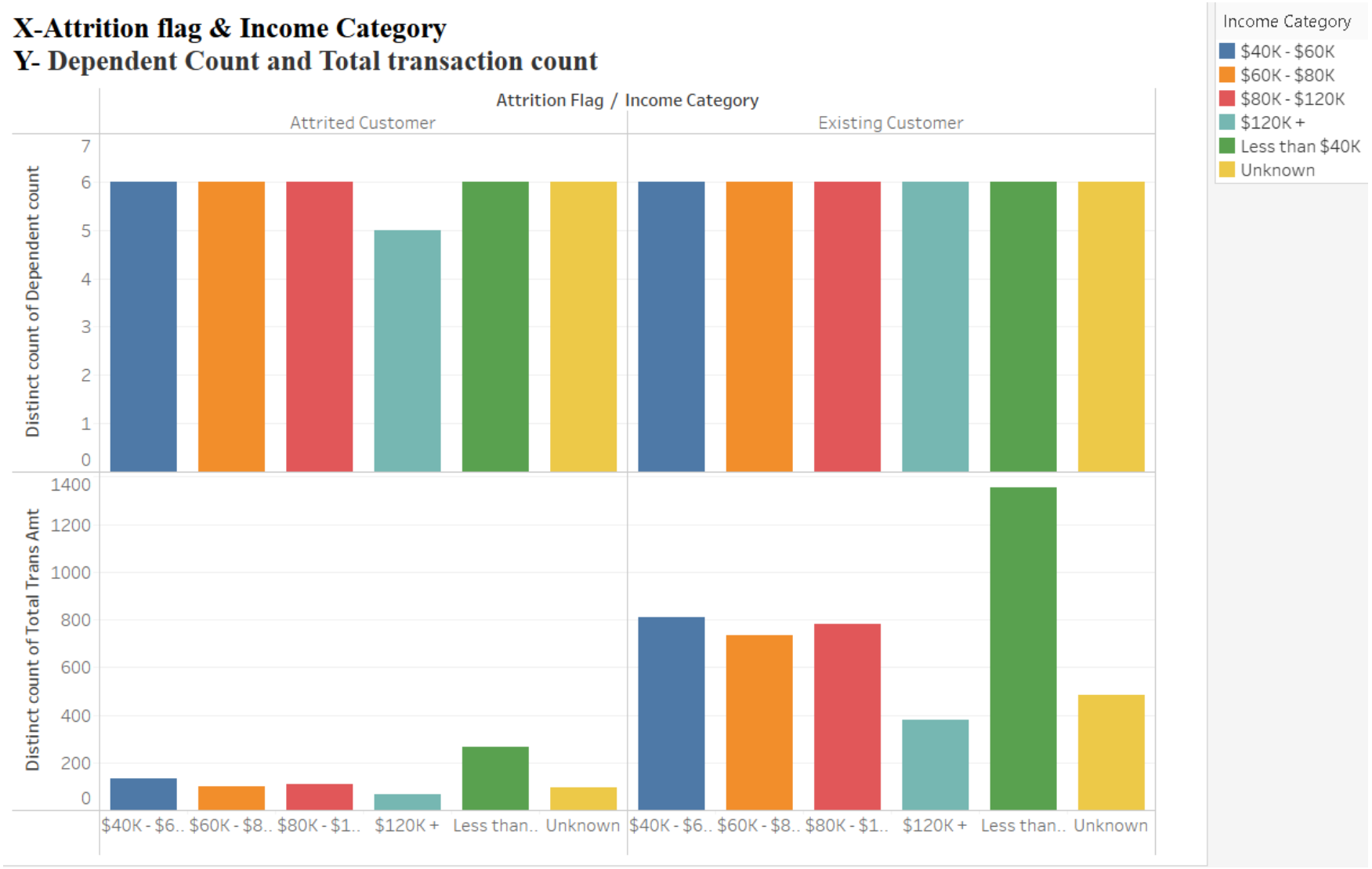




YURAJA SOHEL KADARI

Why not other features is selected ?

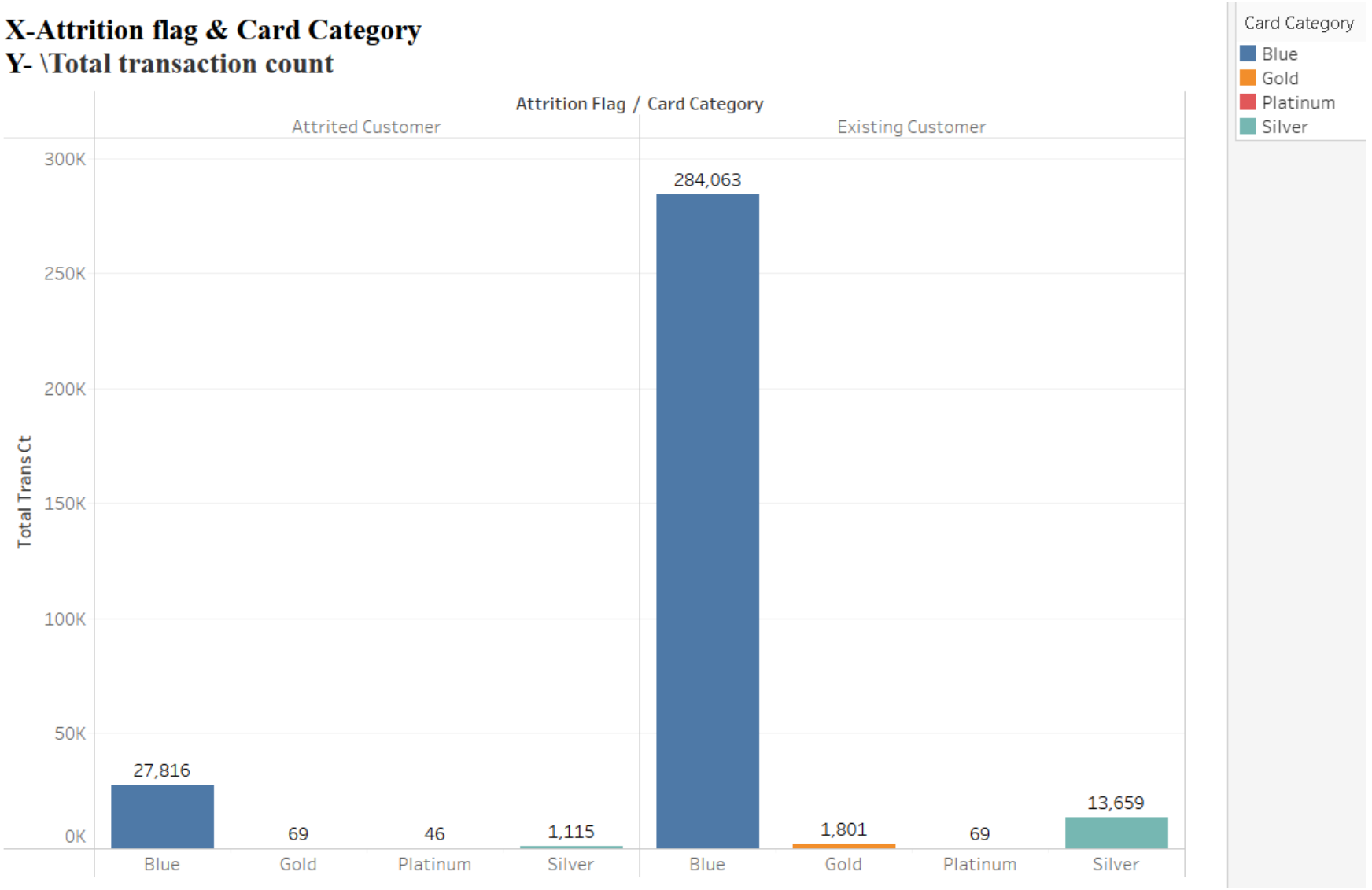
- Here we are taking other variables visualization for visualizing the output.
- So, the variable dependent count is not showing great result on why the customer are leaving the services.
- Whereas the variable Total transaction count Gives you the exact information which customer are leaving the services from which category.
- This is one of the reason why I Didn’t select this features.



YURAJA SOHEL KADARI

Which card category sales should be increased and how ?

- Here you can see platinum card category need more sales over other card category.
- As the customers in platinum and gold card category is drastically less than blue card we need to promote more offers for platinum and gold category members.
- Hence using above analysis we can increased the sales.



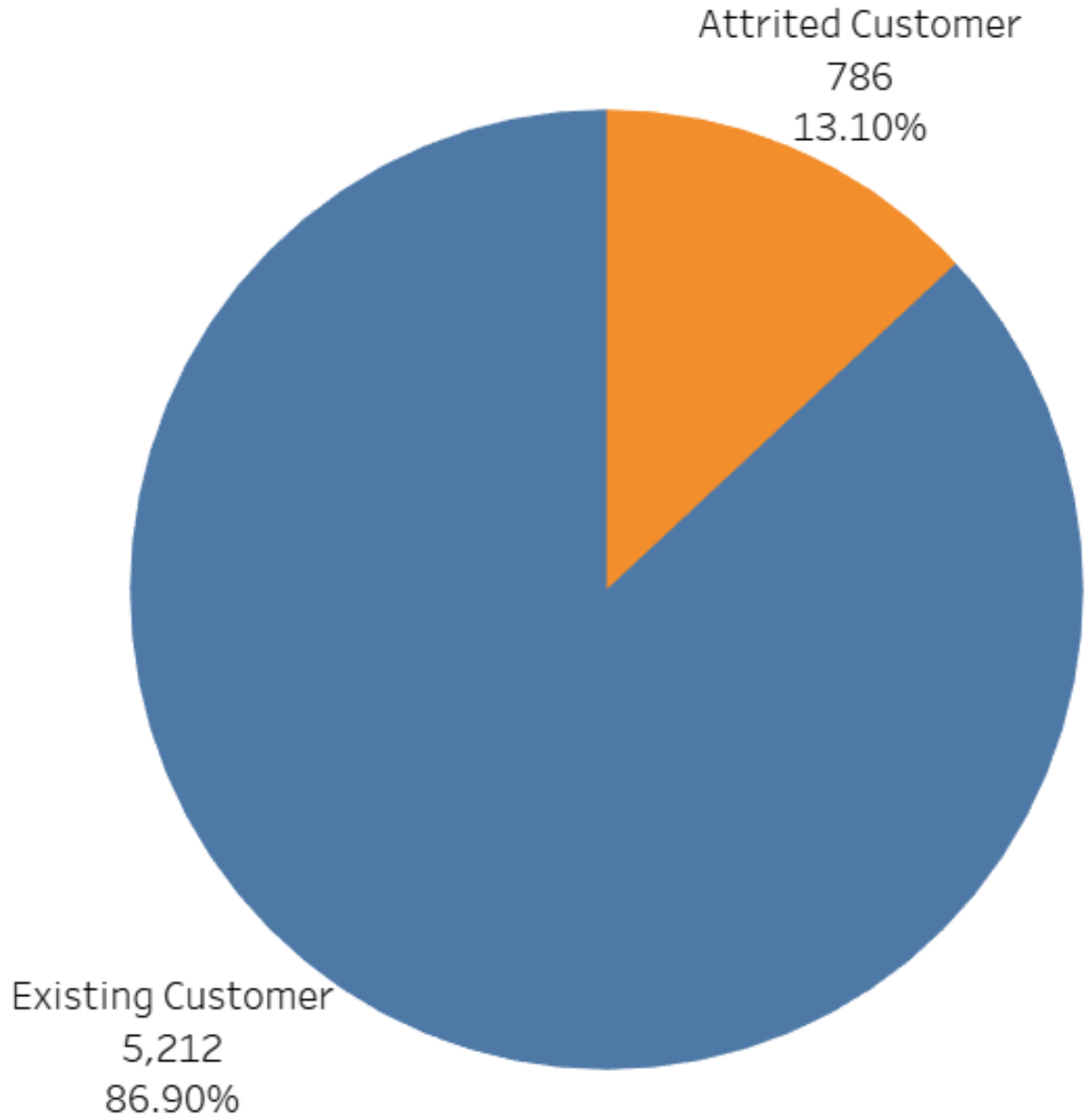
YURAJA SOHEL KADARI

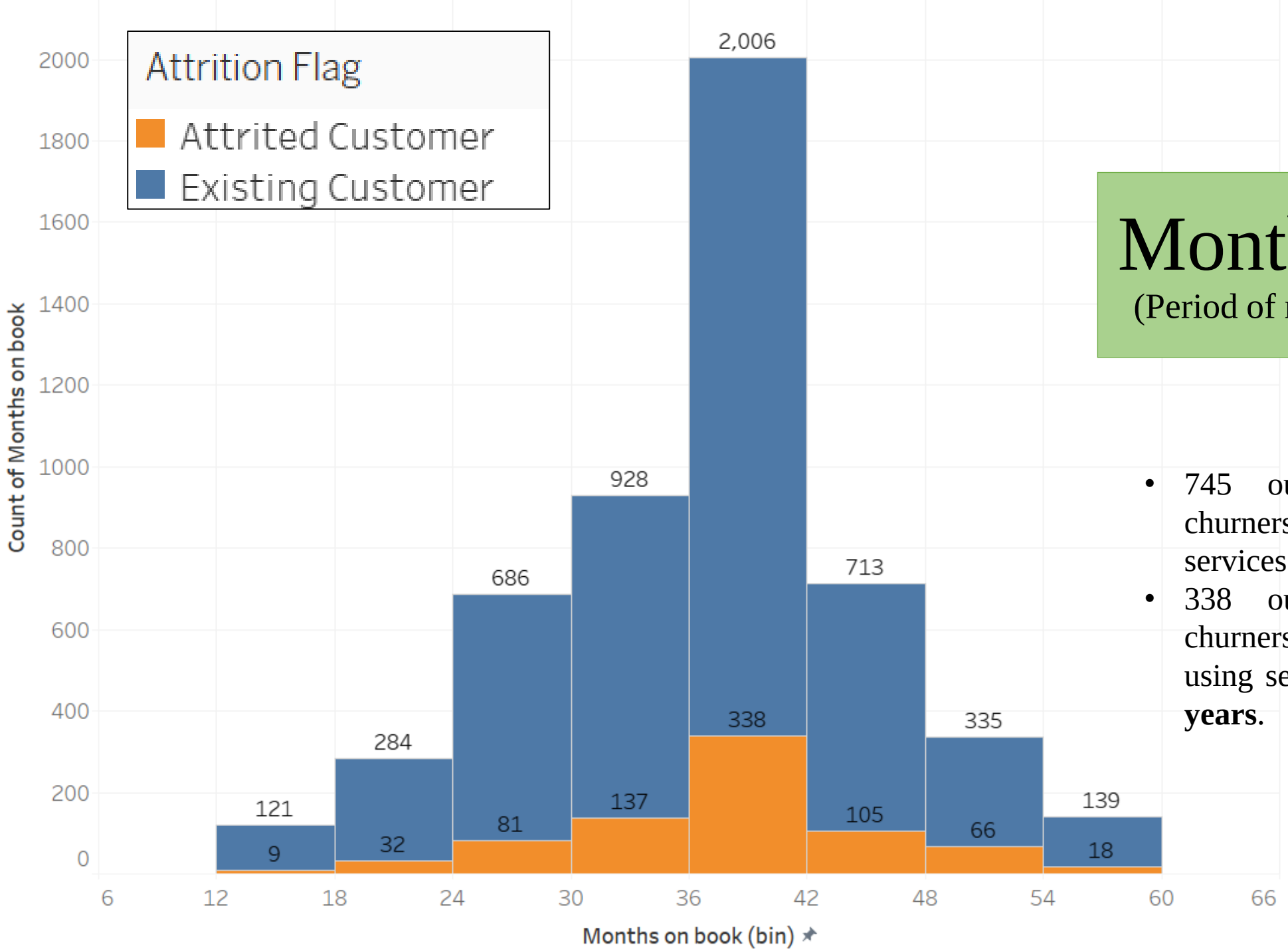
Feedback from Professor

- As professor said : The objective is not only to identify potential customers who are discontinuing the use of credit card services but also to propose a solution to address the issue.
- It is given on slide 14 that we can provide more offers to platinum and gold members so its sales can be increased.
- We can provide more offer to the customer whose age is between 40-55.
- We can provide addition offer to the customer whose salary range is less than $40K.
- Through all the above measures we can solved the issue.

Attrition Count

- Data set contains 86.9% of the customers who are using the bank services.
- It contains 13.1% of the customers who are not using the services of the bank.
- Are there any specific feature/category/sub-category in which the customers are dropping off the bank services?





Months on book

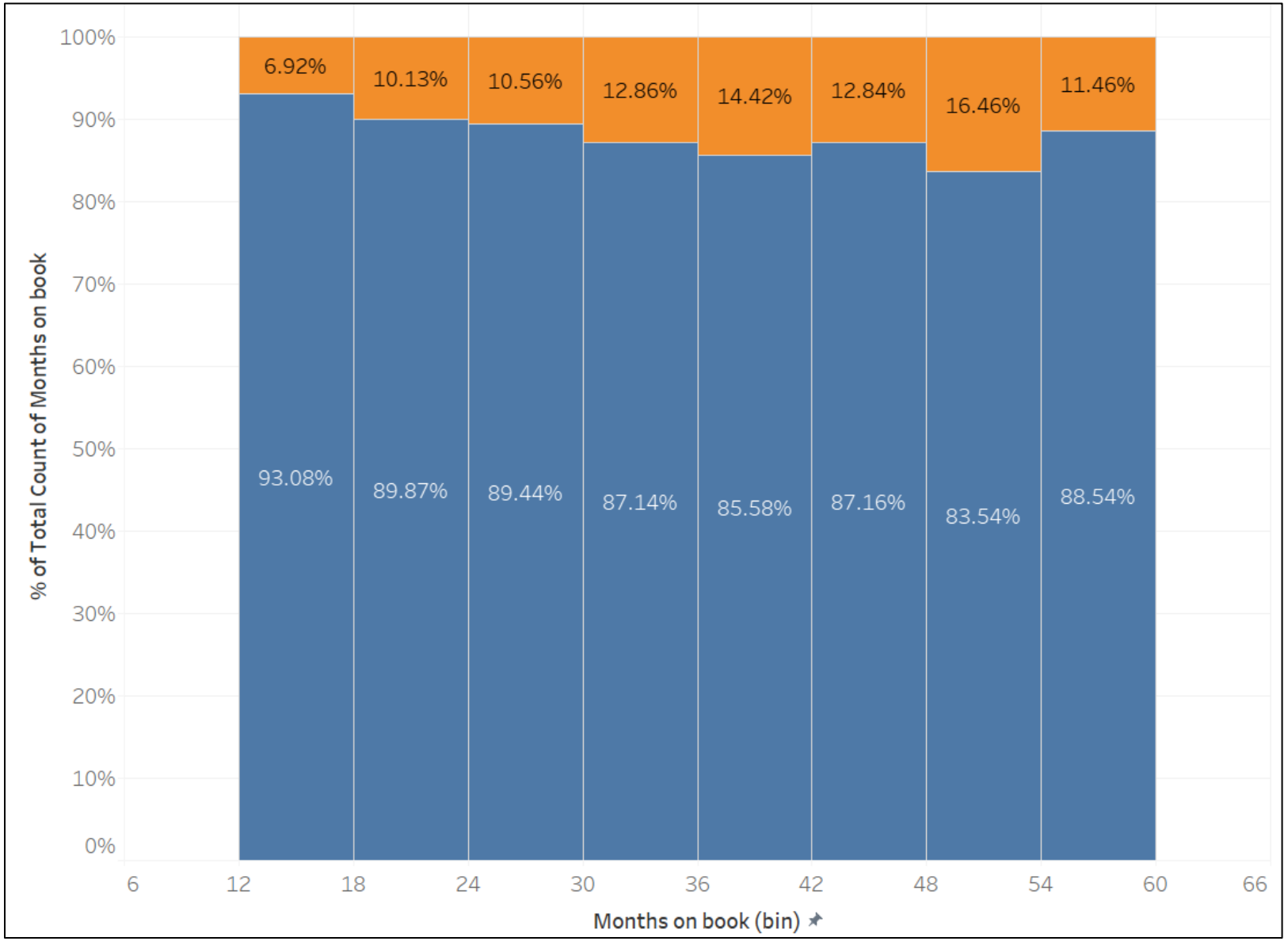
(Period of relationship with bank)

Numerical

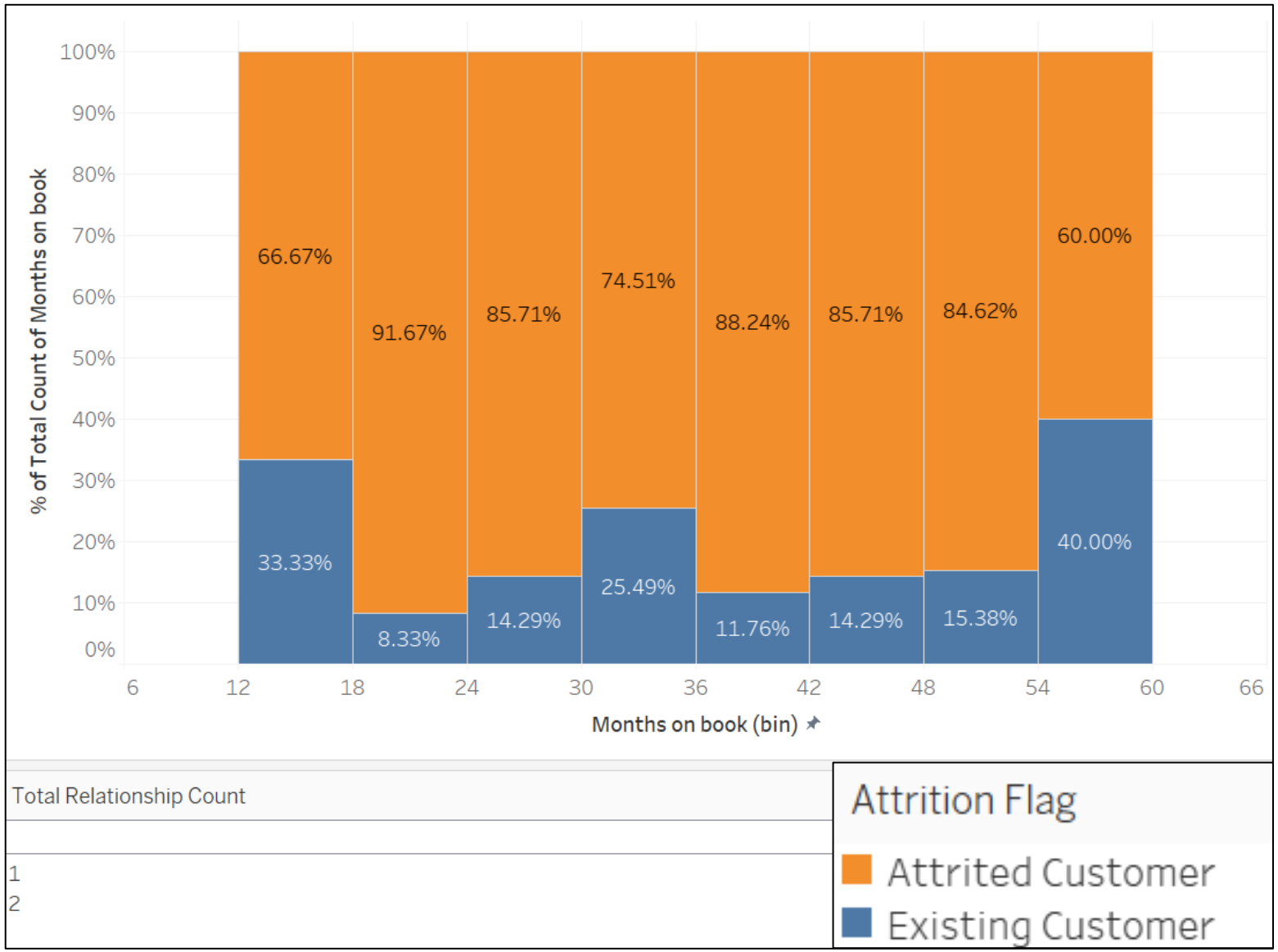
- 745 out of 786 (**95%**) churners are using bank services **more than 2 years**.
- 338 out of 745 (**45%**) churners are the customers using services b/w **3 to 3-1/2 years**.

- Yash Jivani -

Months on book



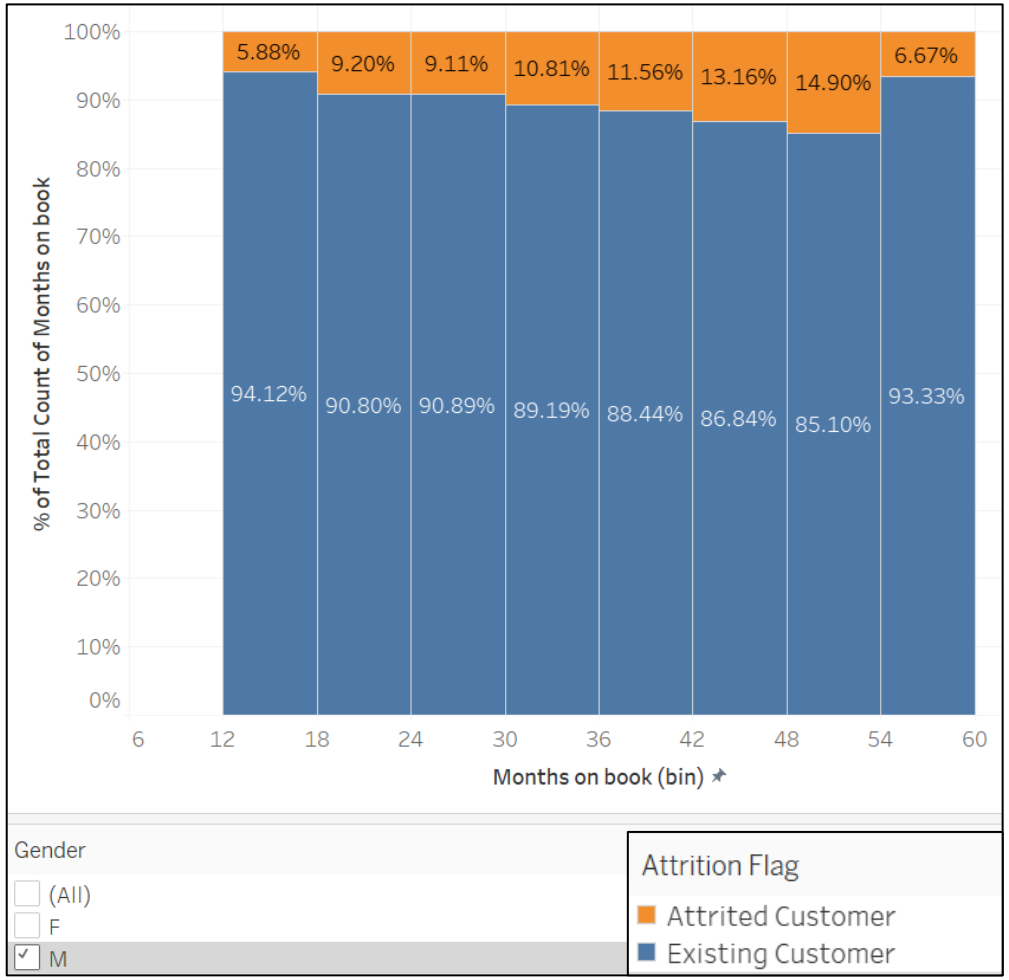
- After normalizing, the bin of **30-36 months** (3 to 3-1/2 yrs), only **14%** of the customers in that period are churners.
- Nearly around **12%** churners in **each** months' bin.

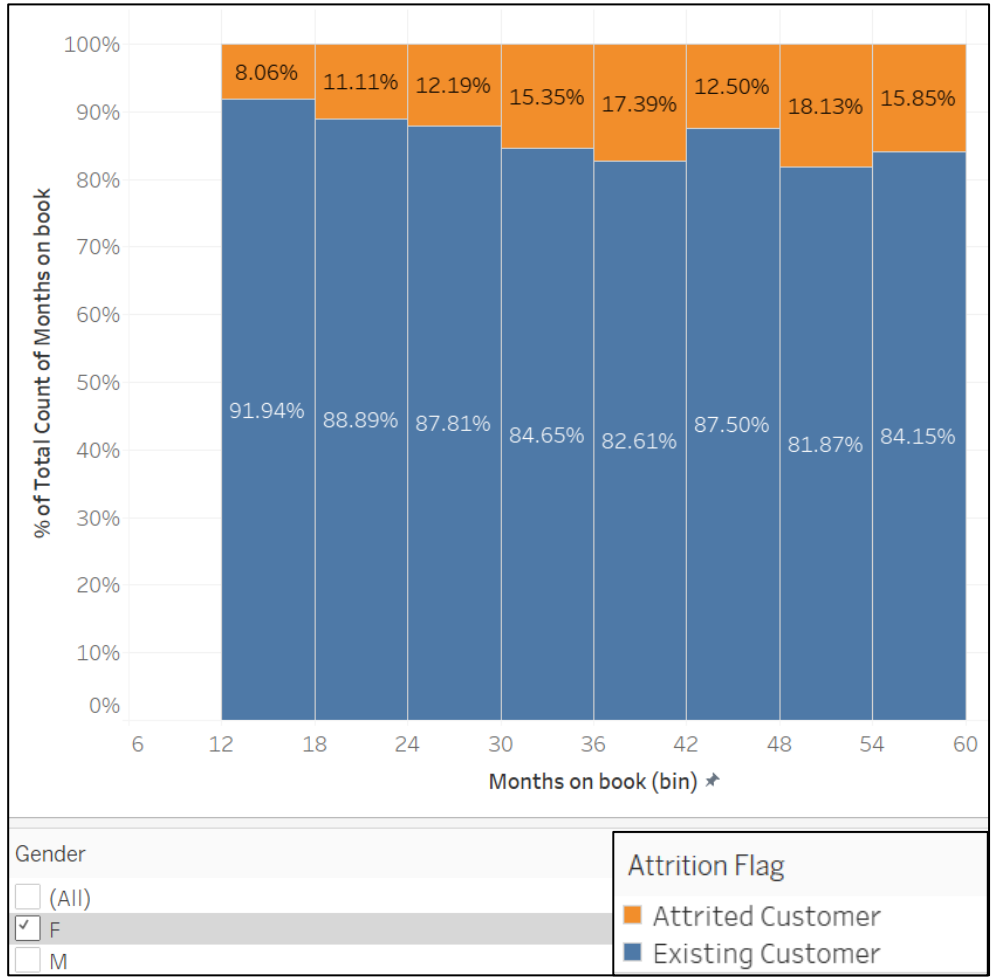


Months on book

After adding filters with Total Relationship Count - 1 & 2, it is clear that customers churn who are almost same.

Months on book





Either Male or Female, customer are churning at almost equal %

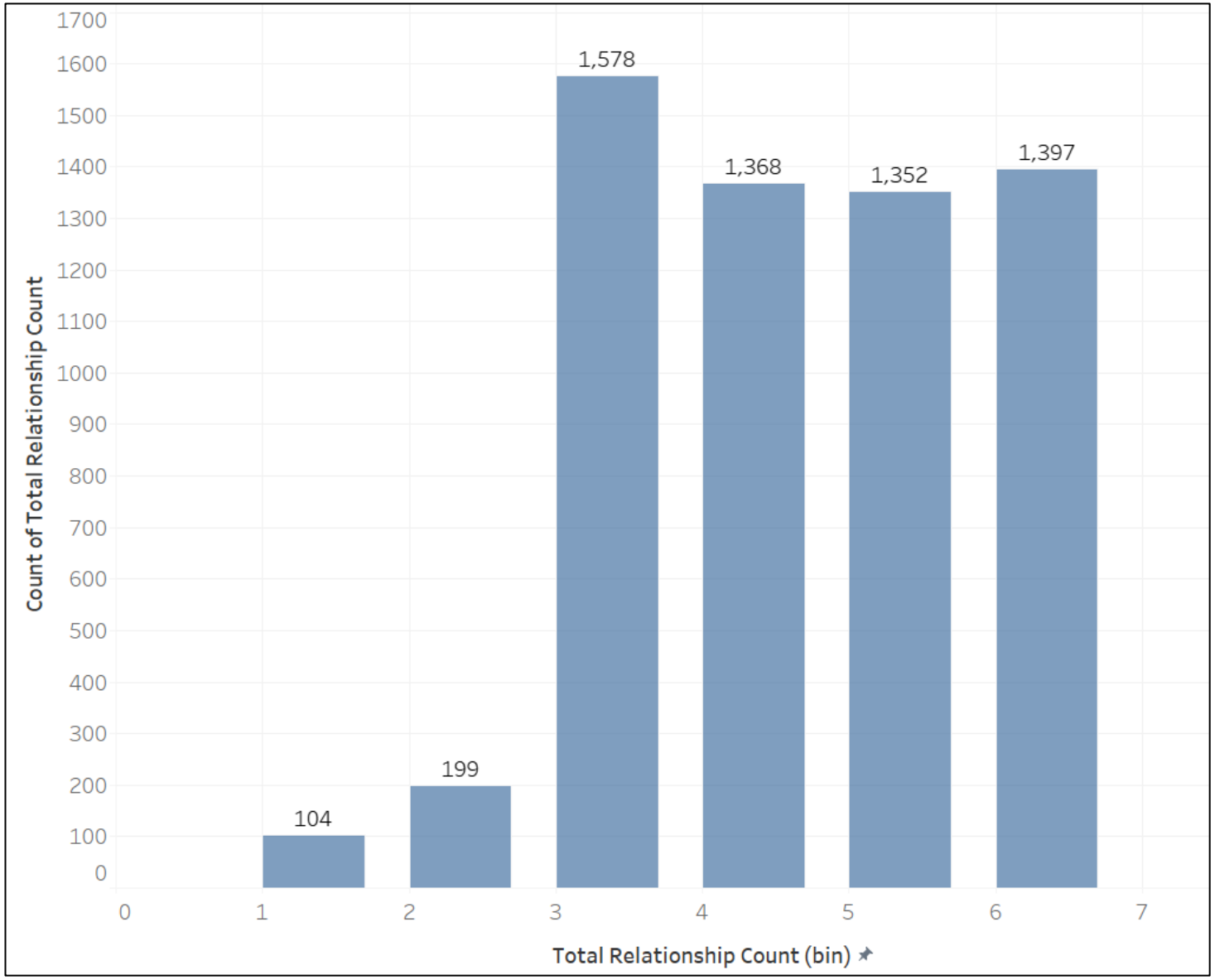
- Yash Jivani -

Total Relationship Count

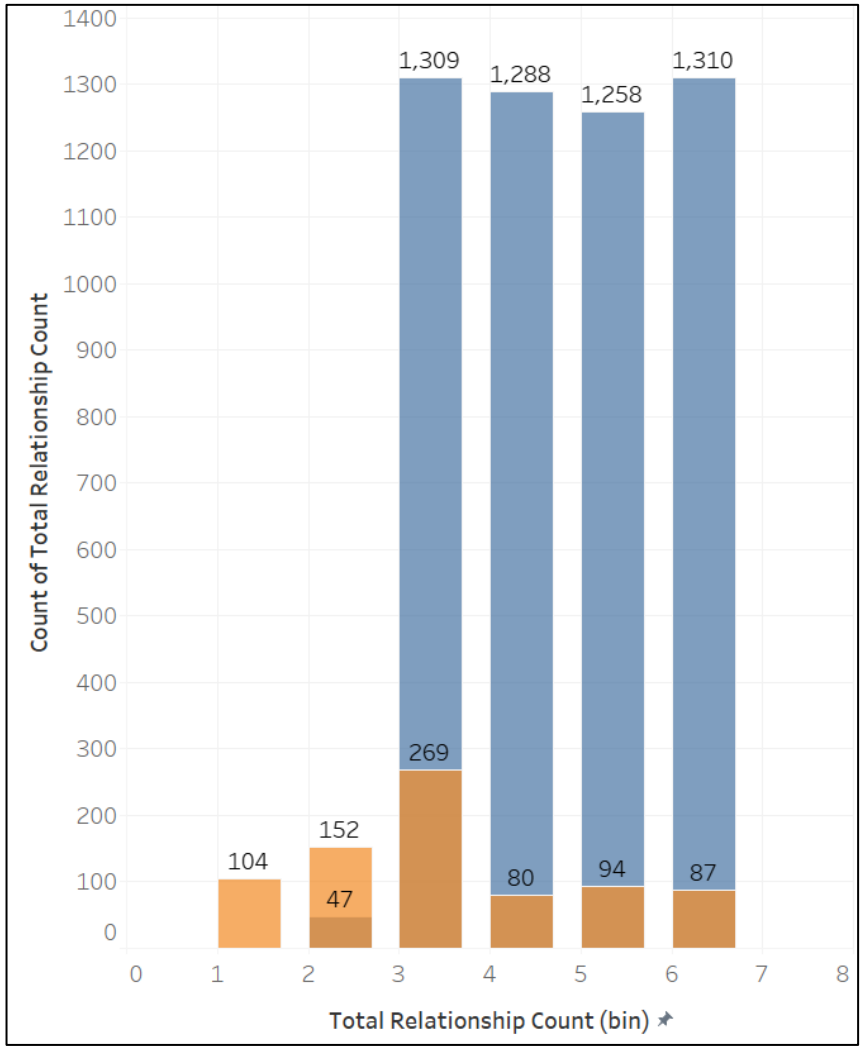
(No. of products/services the customer/s is/are using)

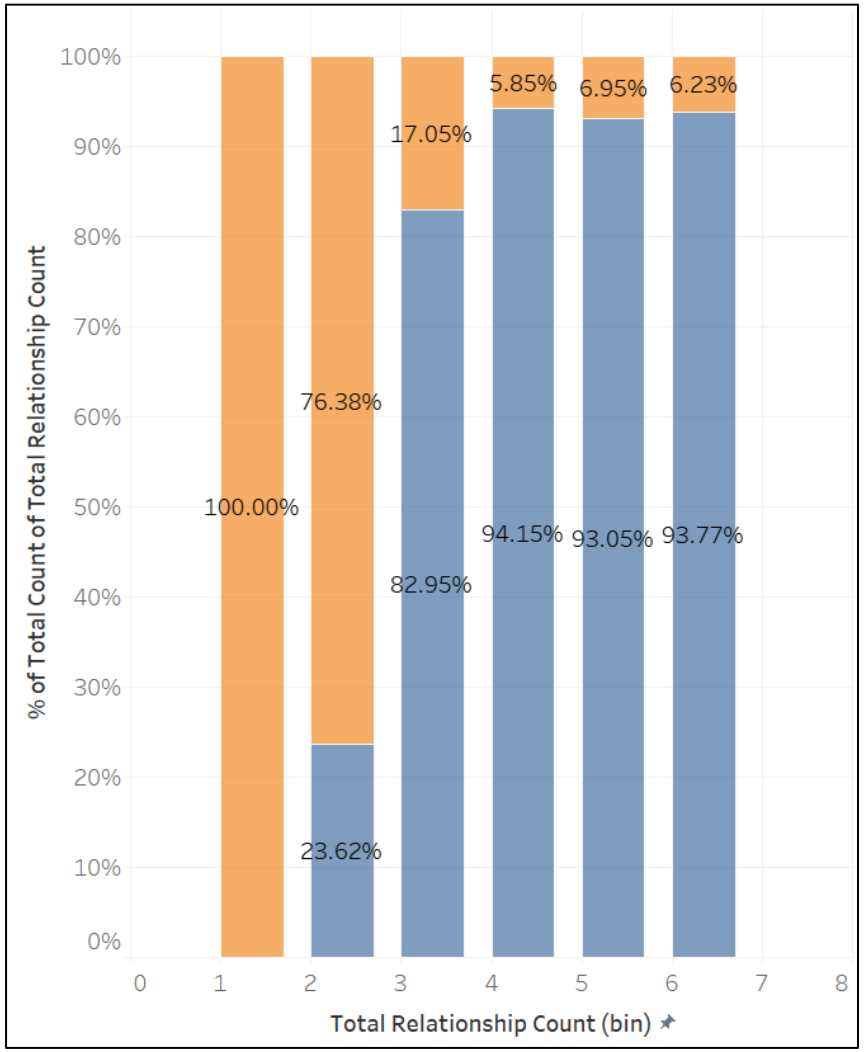
Discrete Categorical

Around **95%** customers use products/services more than 2



Total Relationship Count

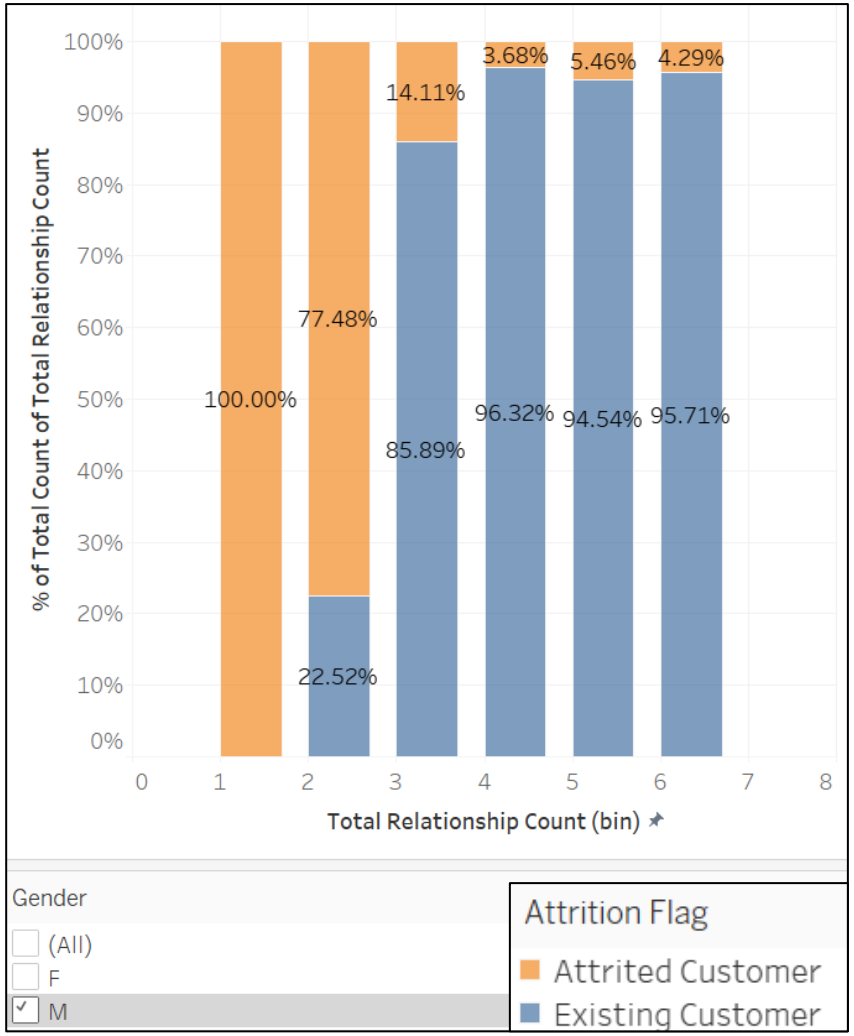


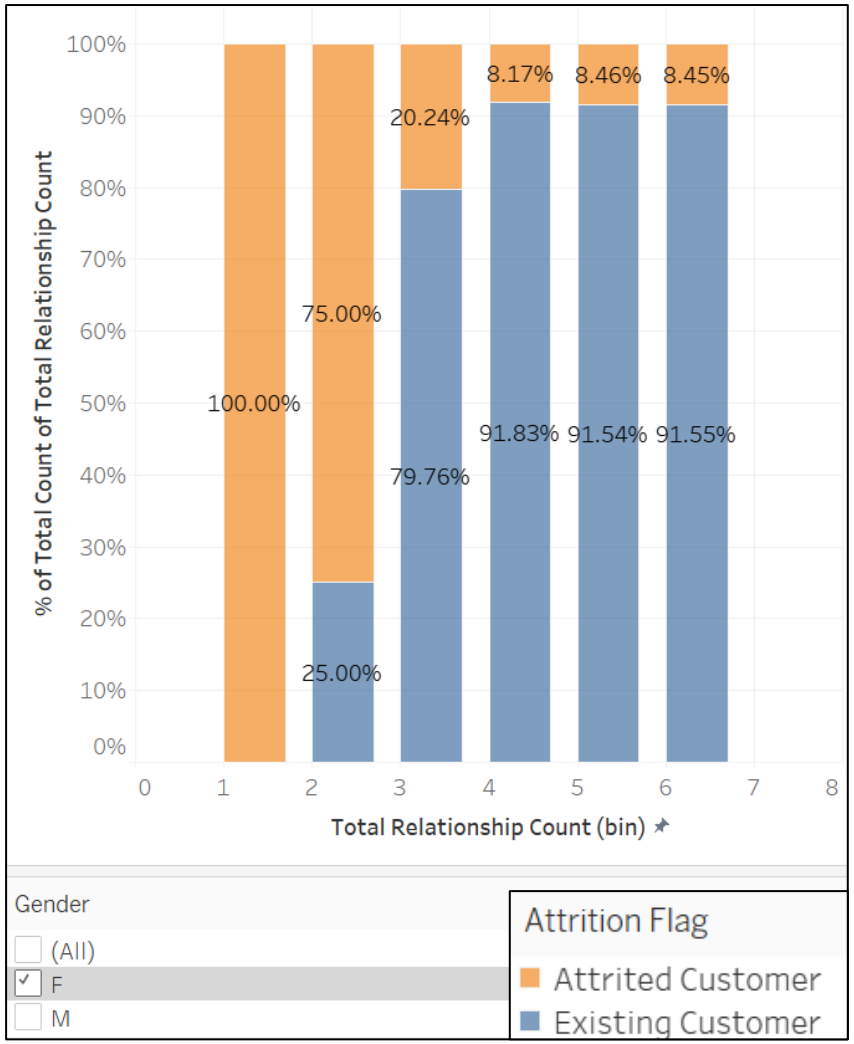


Customers using products/services count-1 or 2 are churning

- Yash Jivani -

Total Relationship Count





Either Male or Female customer using products/services count-1 or 2 are churning

- Yash Jivani -

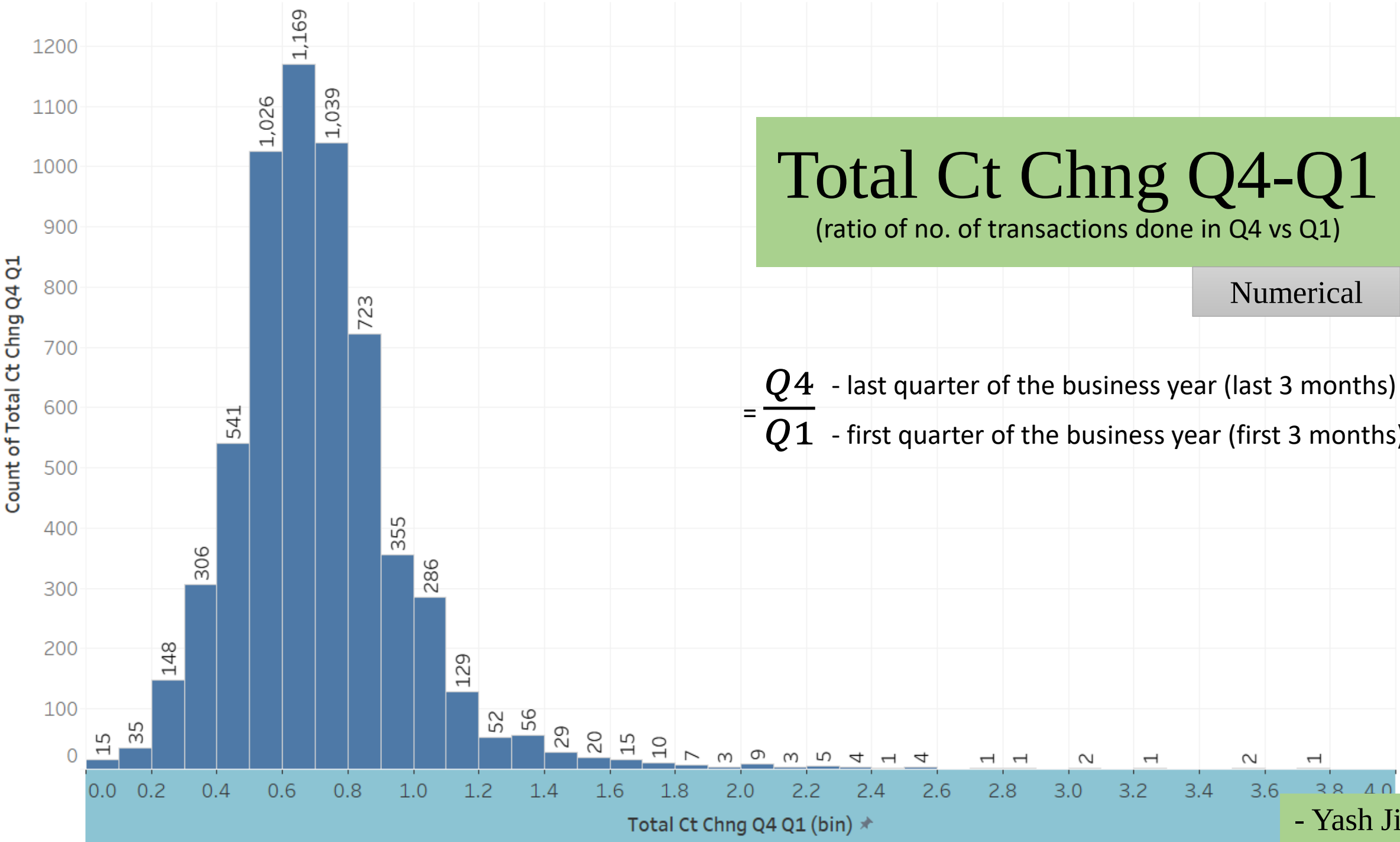
Total Ct Chng Q4-Q1

(ratio of no. of transactions done in Q4 vs Q1)

Numerical

$$= \frac{Q4}{Q1}$$

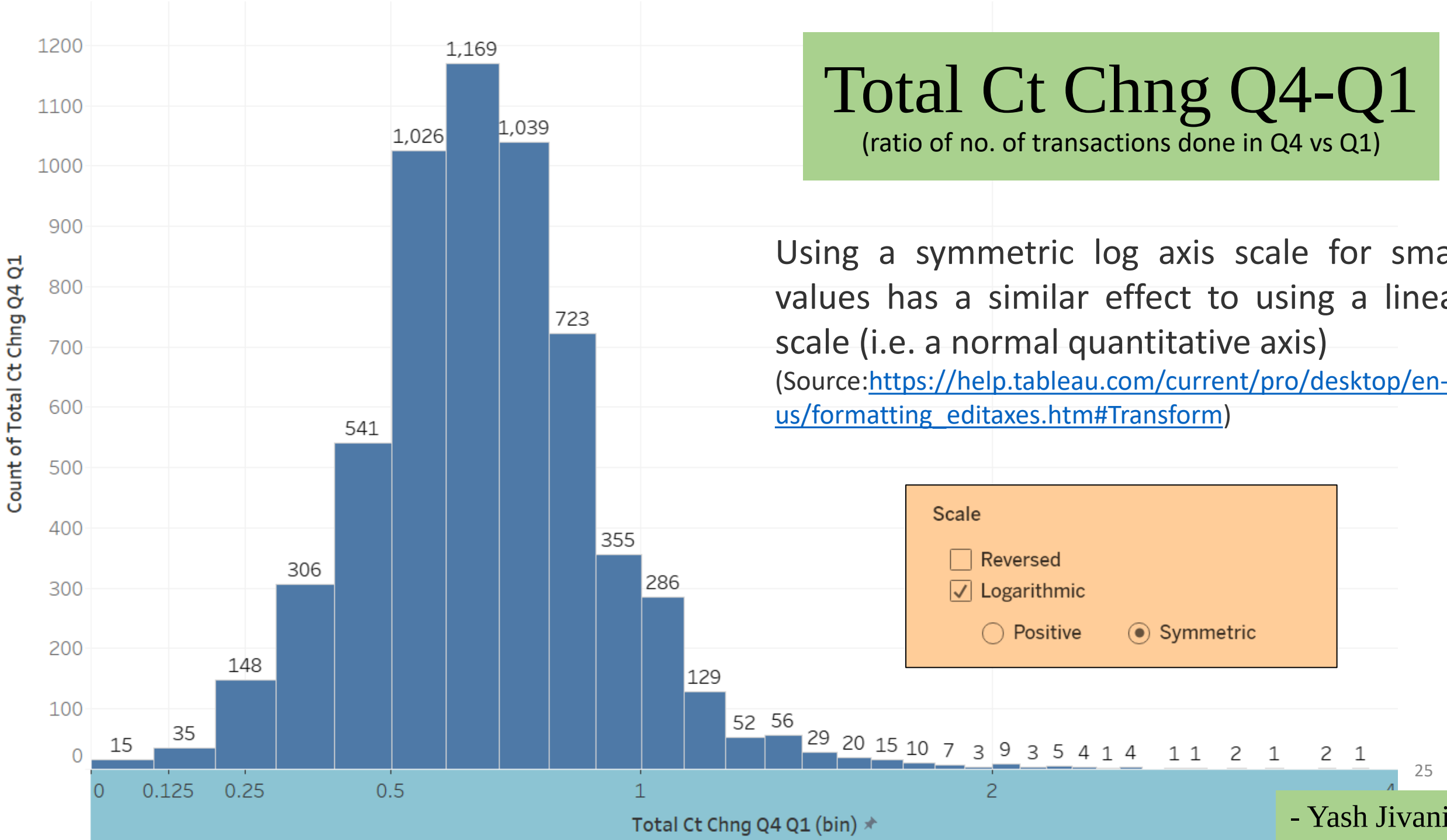
- $Q4$ - last quarter of the business year (last 3 months)
- $Q1$ - first quarter of the business year (first 3 months)



- Yash Jivani -

Total Ct Chng Q4-Q1

(ratio of no. of transactions done in Q4 vs Q1)



Using a symmetric log axis scale for small values has a similar effect to using a linear scale (i.e. a normal quantitative axis)

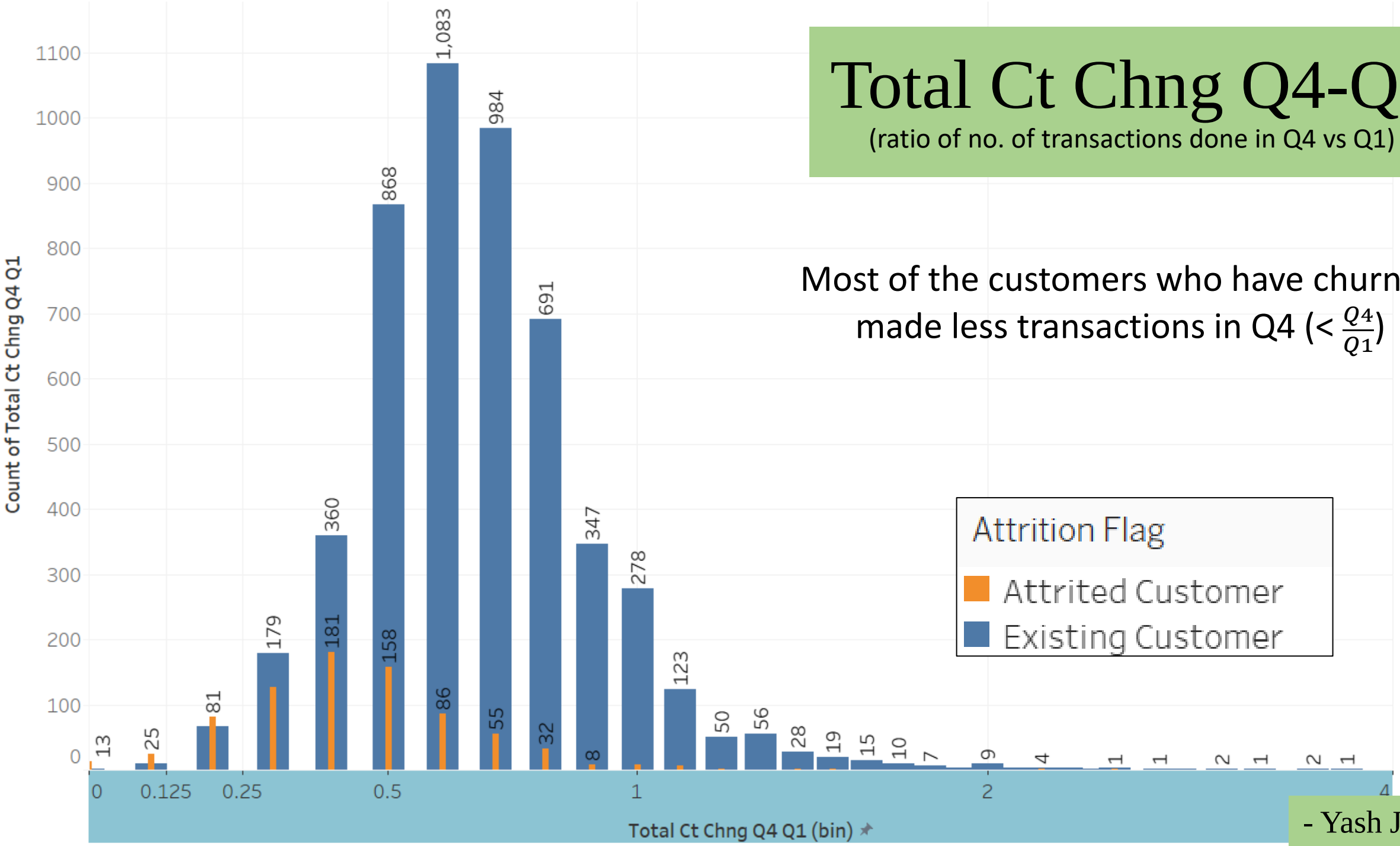
(Source:https://help.tableau.com/current/pro/desktop/en-us/formatting_editaxes.htm#Transform)

- Yash Jivani -

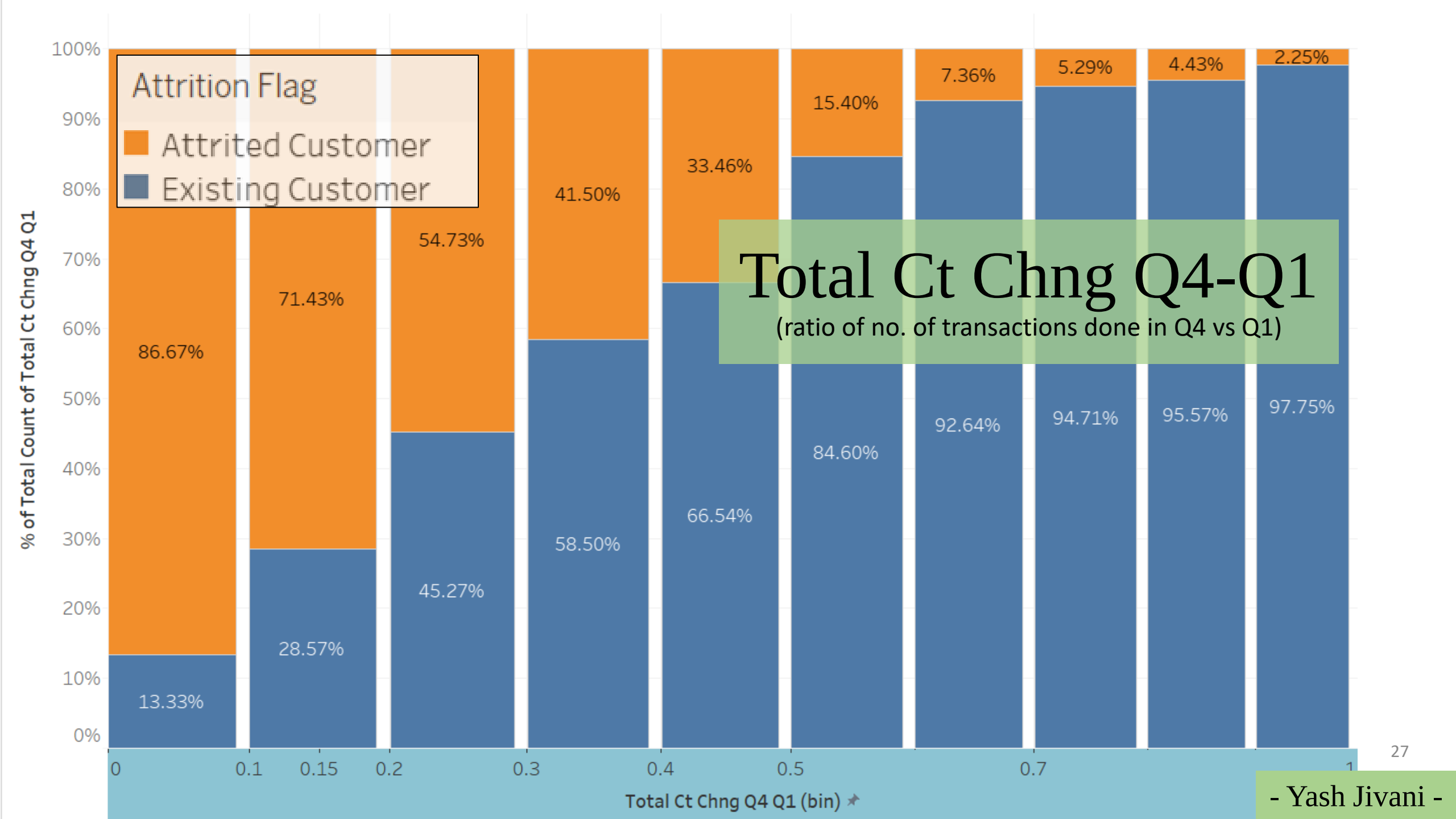
Total Ct Chng Q4-Q1

(ratio of no. of transactions done in Q4 vs Q1)

Most of the customers who have churned made less transactions in Q4 ($< \frac{Q4}{Q1}$)



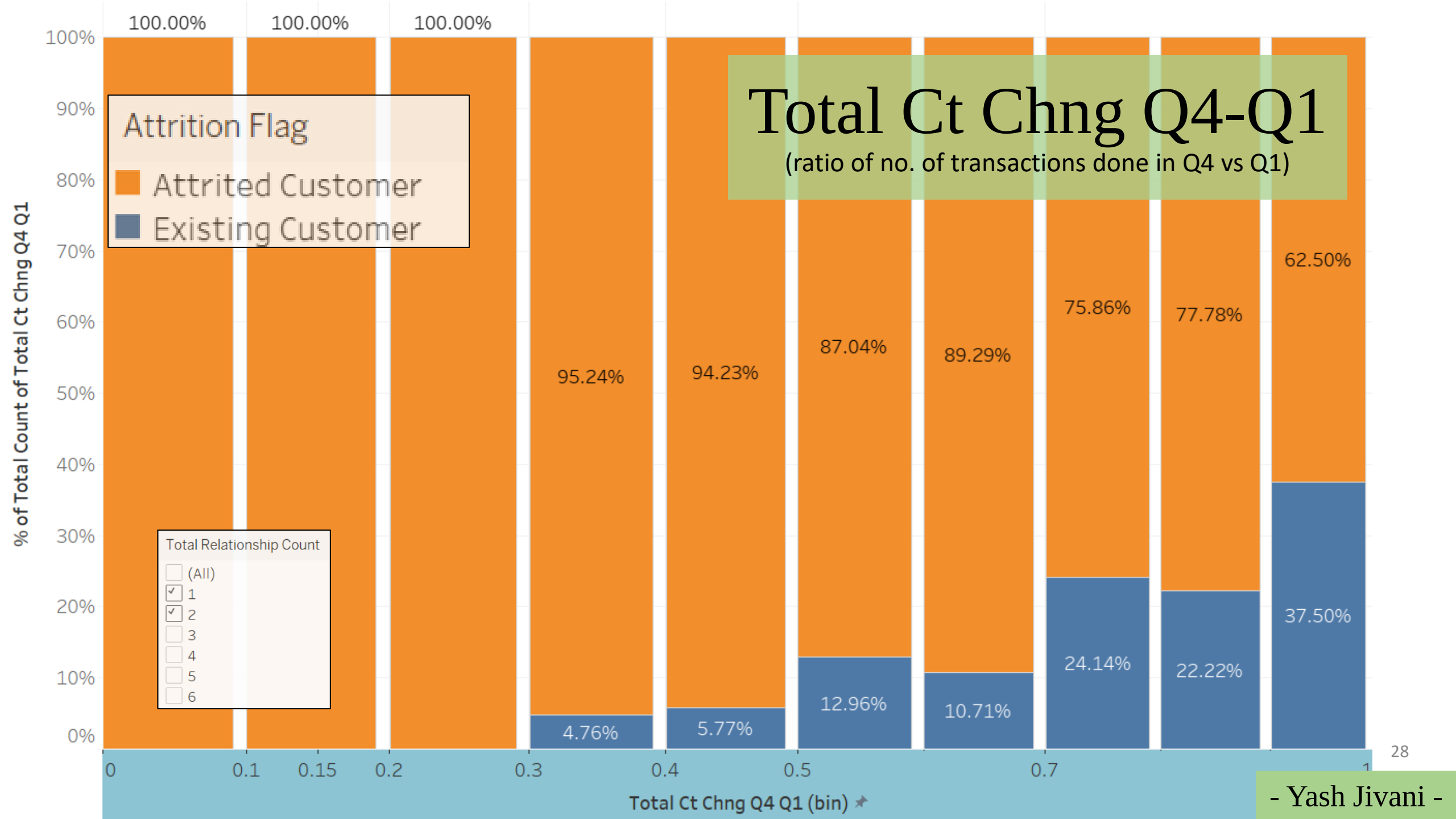
- Yash Jivani -



Total Ct Chng Q4-Q1

(ratio of no. of transactions done in Q4 vs Q1)

- Yash Jivani -


Total Ct Chng Q4-Q1
(ratio of no. of transactions done in Q4 vs Q1)
Attrition Flag
Attrited Customer
Existing Customer
Total Relationship Count
(All)
1
2
3
4
5
6
% of Total Count of Total Ct Chng Q4 Q1
100%
90%
80%
70%
60%
50%
40%
30%
20%
10%
0%
100.00%
100.00%
100.00%
95.24%
94.23%
87.04%
89.29%
75.86%
77.78%
62.50%
4.76%
5.77%
12.96%
10.71%
24.14%
22.22%
37.50%
0
0.1
0.15
0.2
0.3
0.4
0.5
0.7
1
Total Ct Chng Q4 Q1 (bin)
28
- Yash Jivani -

Methodology

1. Imported the dataset into Google Colab.

a) Executed the EDA by importing some fundamental libraries like pandas, matplotlib, seaborn.

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

```
[2] data = pd.read_csv('/content/BankChurners for EDA 2023.csv')
```

```
[4] data.head(5)
```

	CLIENTNUM	Attrition Flag	Customer Age	Gender	Dependent count	Education Level	Marital Status	Income Category	Card Category	Months on book	...	Months Inactive 12 mon	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 Q1
0	768805383	Existing Customer	45	M	3	High School	Married	$60K - $80K	Blue	39	...	1	3	12691.0	777	11914.0	1.335	1144	42	1.625
1	818770008	Existing Customer	49	F	5	Graduate	Single	Less than $40K	Blue	44	...	1	2	8256.0	864	7392.0	1.541	1291	33	3.714
2	713982108	Existing Customer	51	M	3	Graduate	Married	$80K - $120K	Blue	36	...	1	0	3418.0	0	3418.0	2.594	1887	20	2.333
3	769911858	Existing Customer	40	F	4	High School	Unknown	Less than $40K	Blue	34	...	4	1	3313.0	2517	796.0	1.405	1171	20	2.333
4	709106358	Existing Customer	40	M	3	Uneducated	Married	$60K - $80K	Blue	21	...	1	0	4716.0	0	4716.0	2.175	816	28	2.500

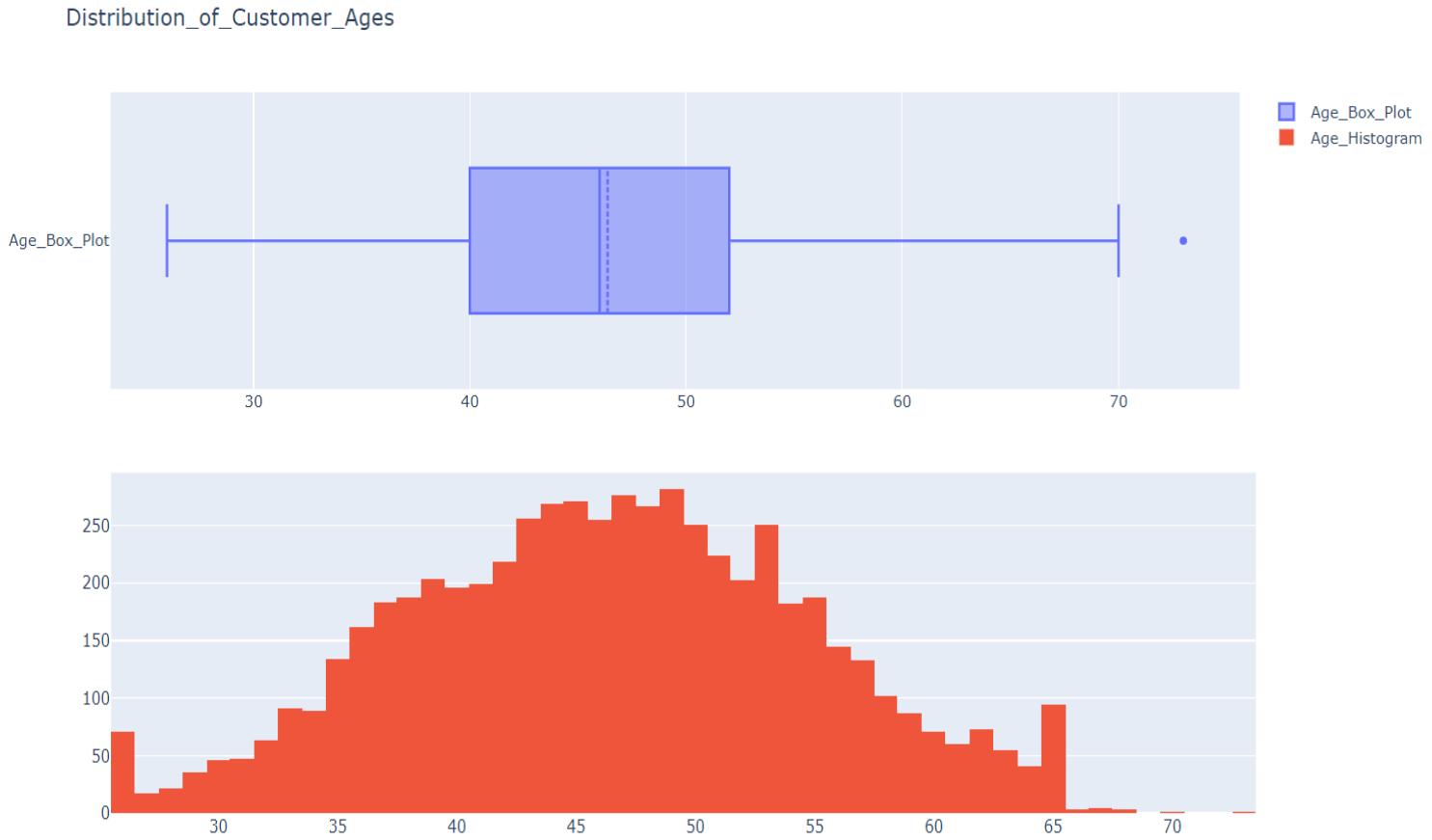
AMEYA

B) Calculated the shape, info., null values of the dataset

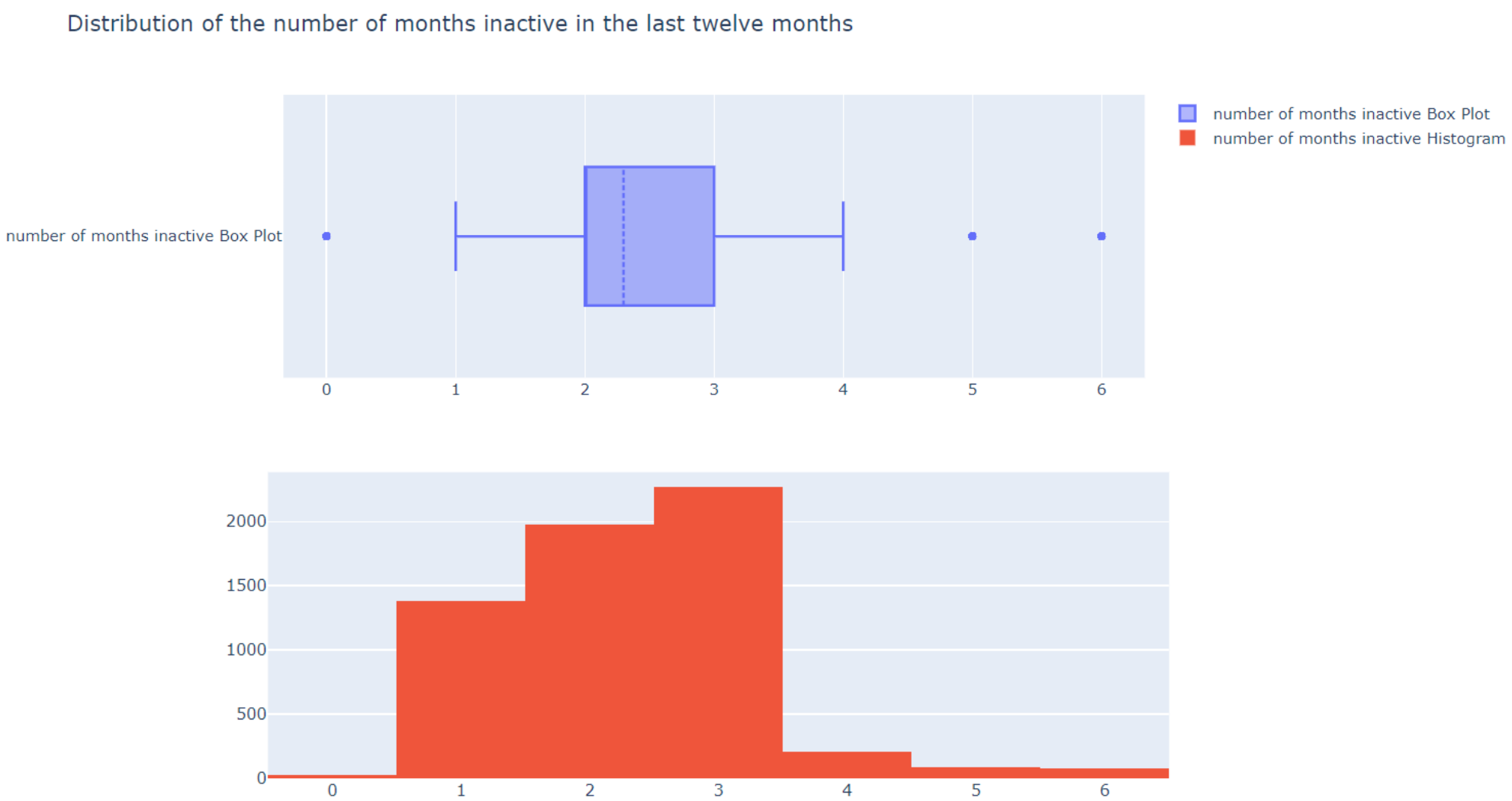
Shape of data: 5998 rows, 21 columns.

There are No NULL values in the dataset.

C) Used Box plot to predict outliers.



AMEYA

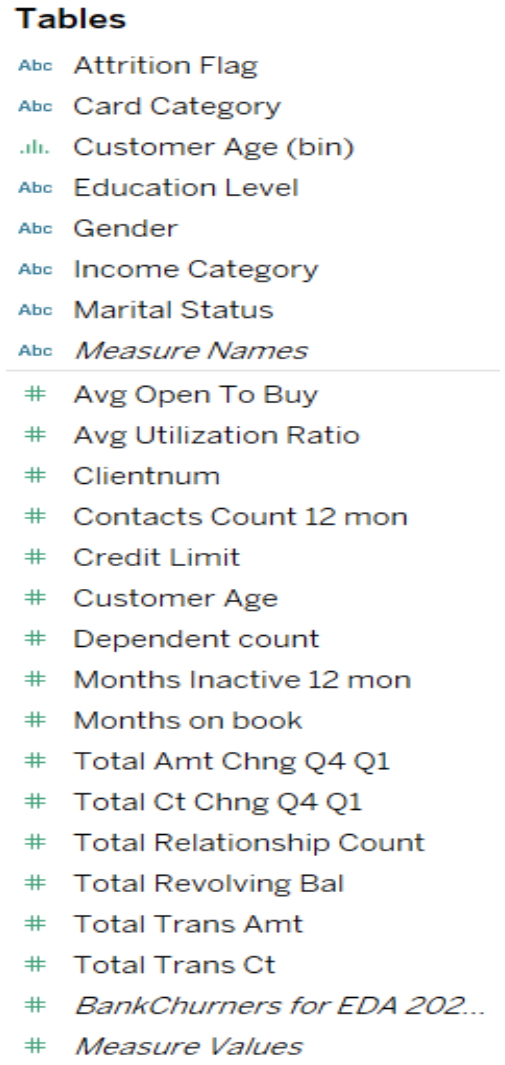


From the graph we can see the customers remain Inactive for 2-3 months.

2. Imported the data to Tableau

Dimension (categorical or qualitative data field that provides context or descriptive information about the data)
In our data, there are **5 Independent variables**

Measures (a quantitative or numerical data field)
There are **16 measure/ dependent variables**

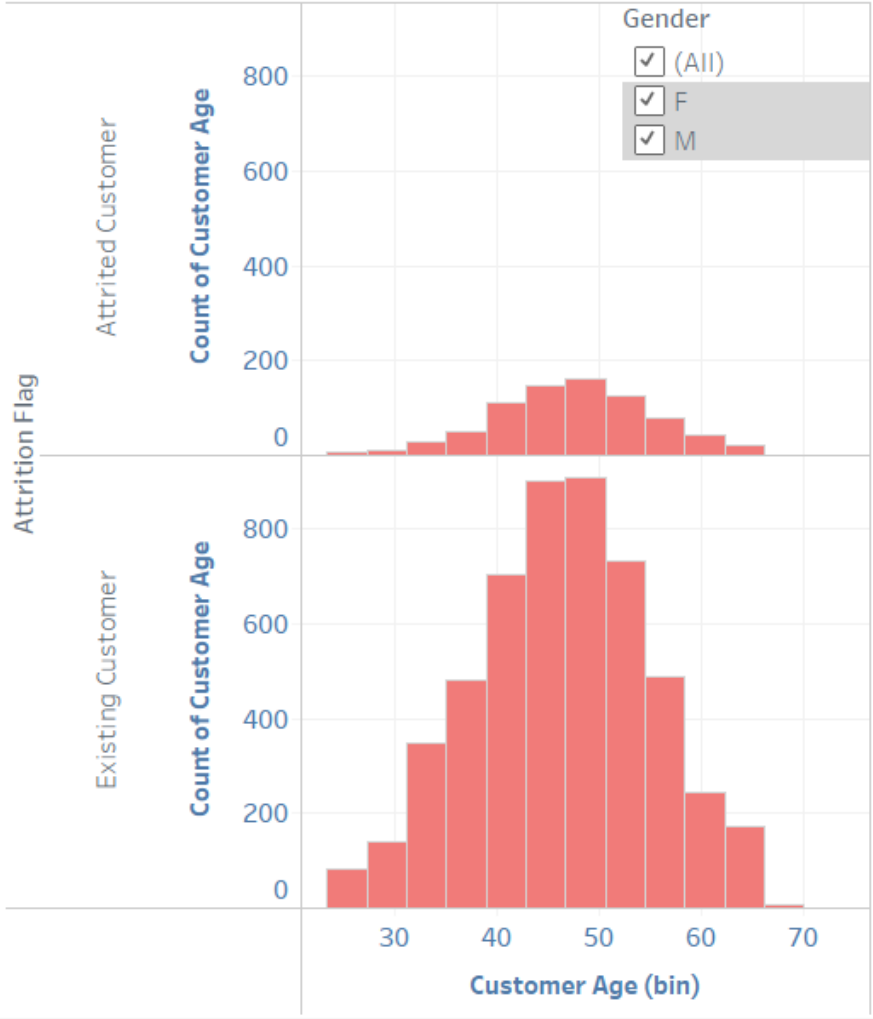


AMEYA

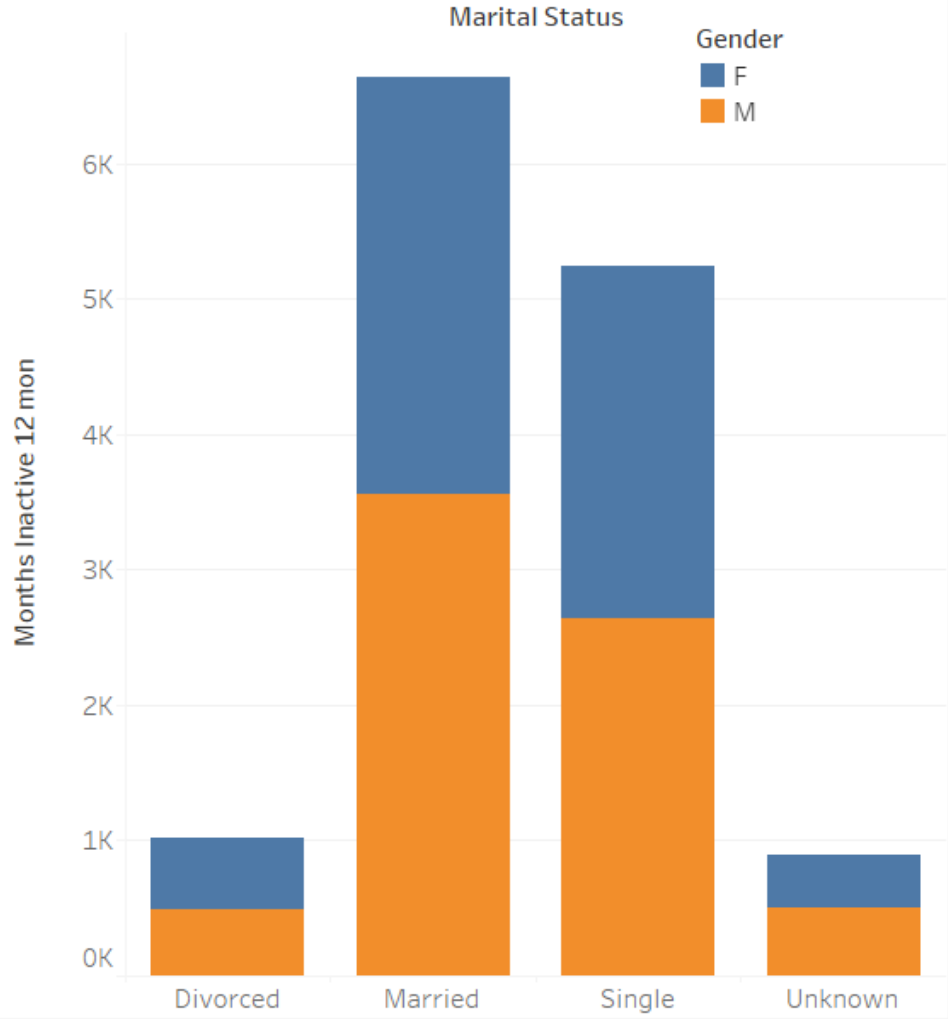
Gender	
F	M
2,883	3,115

Attrition Flag	
Attrited Customer	786
Existing Customer	5,212

Age Vs Attriation



Maritial Status Vs Months Inactive

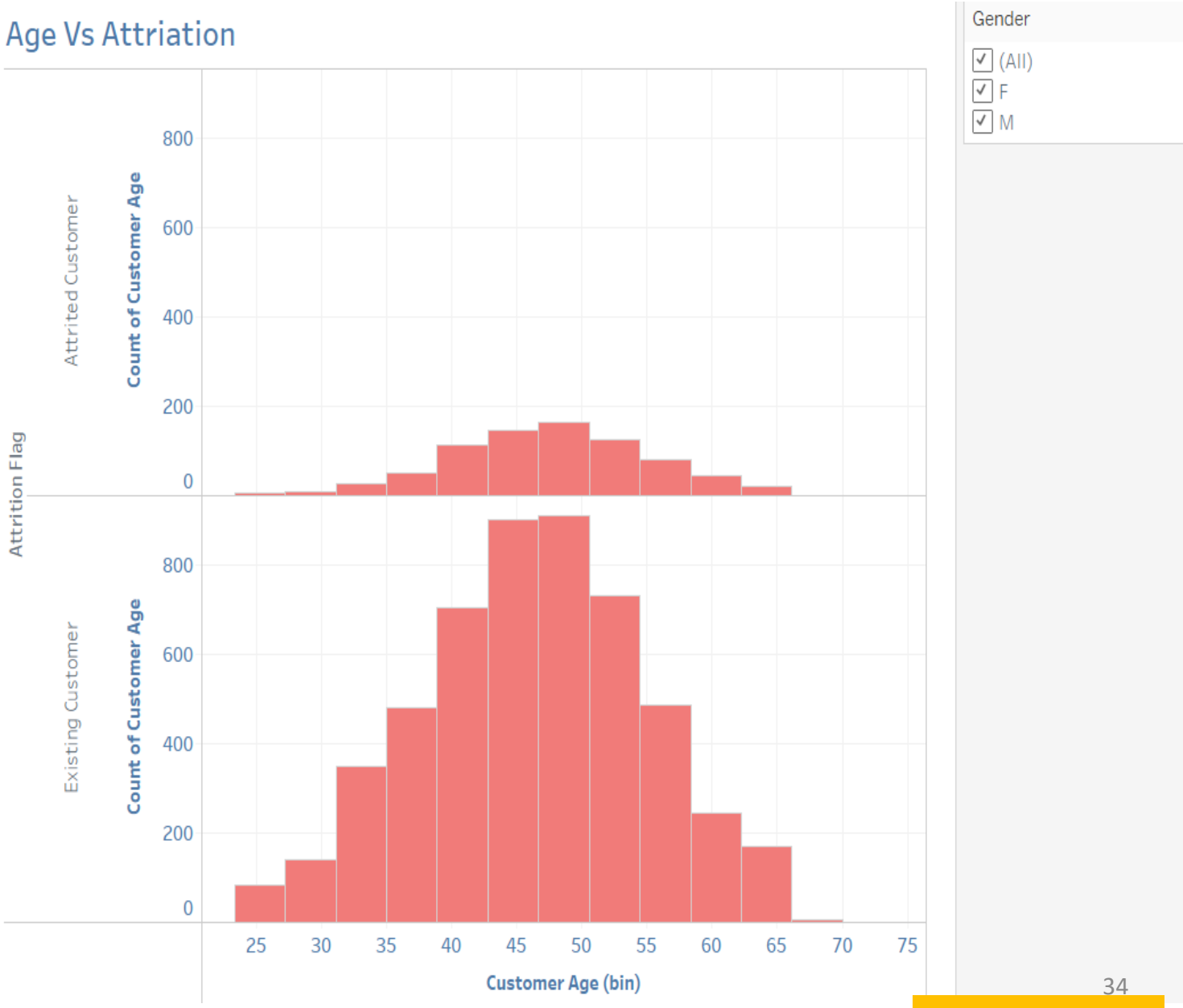


AMEYA

By plotting the customer age on X-axis and Attriation Flag on the Y-axis, and filtering the figure using Gender attribute. We can observe that individuals in range 40-55 has taken a break from the bank service.

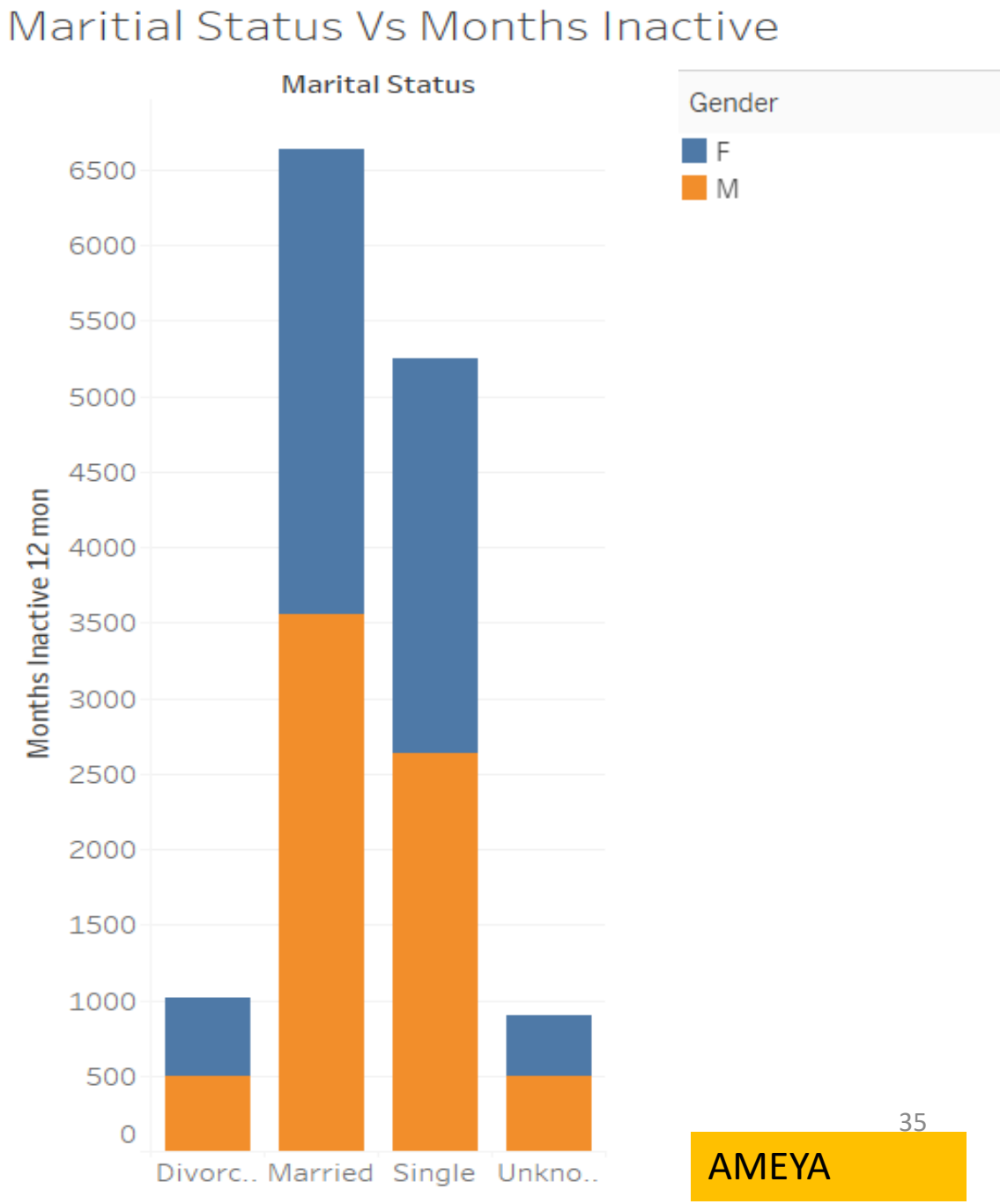
Gender	
F	M
2,883	3,115

Attrition Flag	
Attrited Customer	786
Existing Customer	5,212



AMEYA

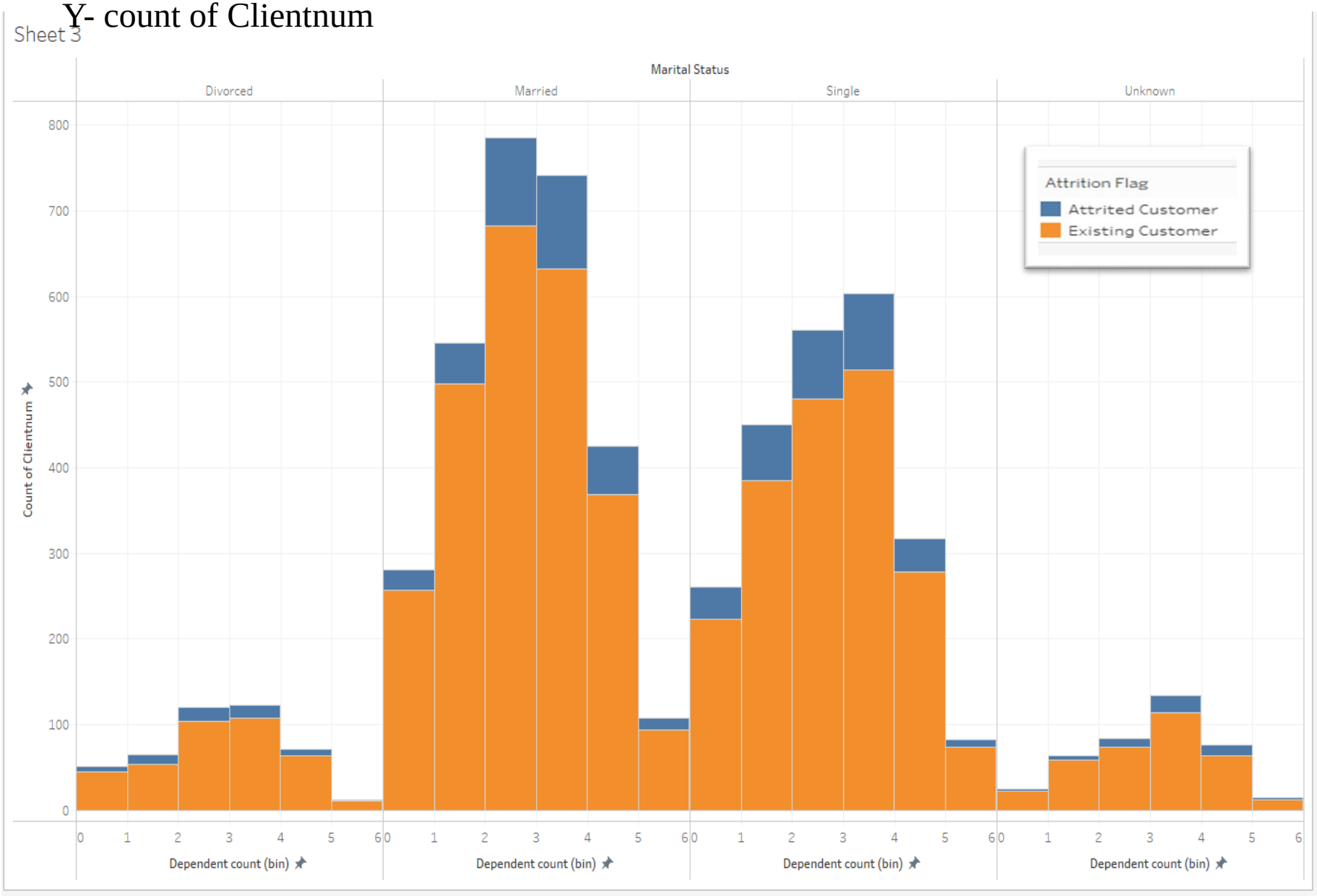
From the second figure which is based on Maritial Status and Months Inactive, and further filtering it using Gender attribute. We can see that married and singles of both the genders showed the similar trend of inactivity in a year.



AMEYA

Dependent Count & Marital Status

X- Marital Stauts & Dependent Count
Y- count of Clientnum



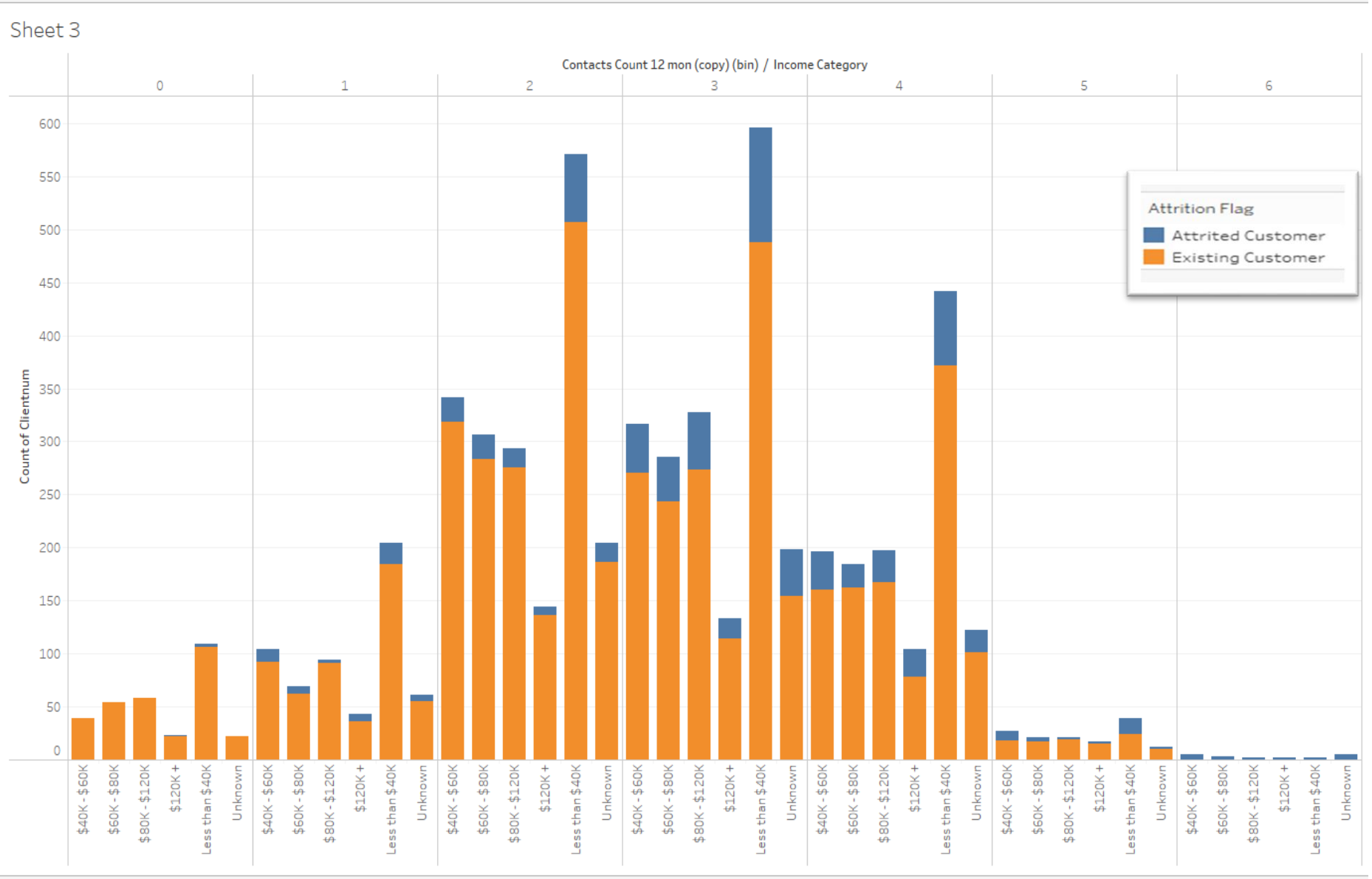
Dependent Count:- dependent count refers to the number of people who are claimed as dependents on a taxpayer's income tax return. They may be children, Elderly, anyone who dependent on taxpayer support.

By filtering the data based on marital status and Dependent Count, we can observe that individuals with more than 2 or 3 dependents are more likely to leave the bank

Vicky

Contacts Count 12 Mon & Income Category

X- Contact count 12 Month and Income category
Y- Count of clientnum

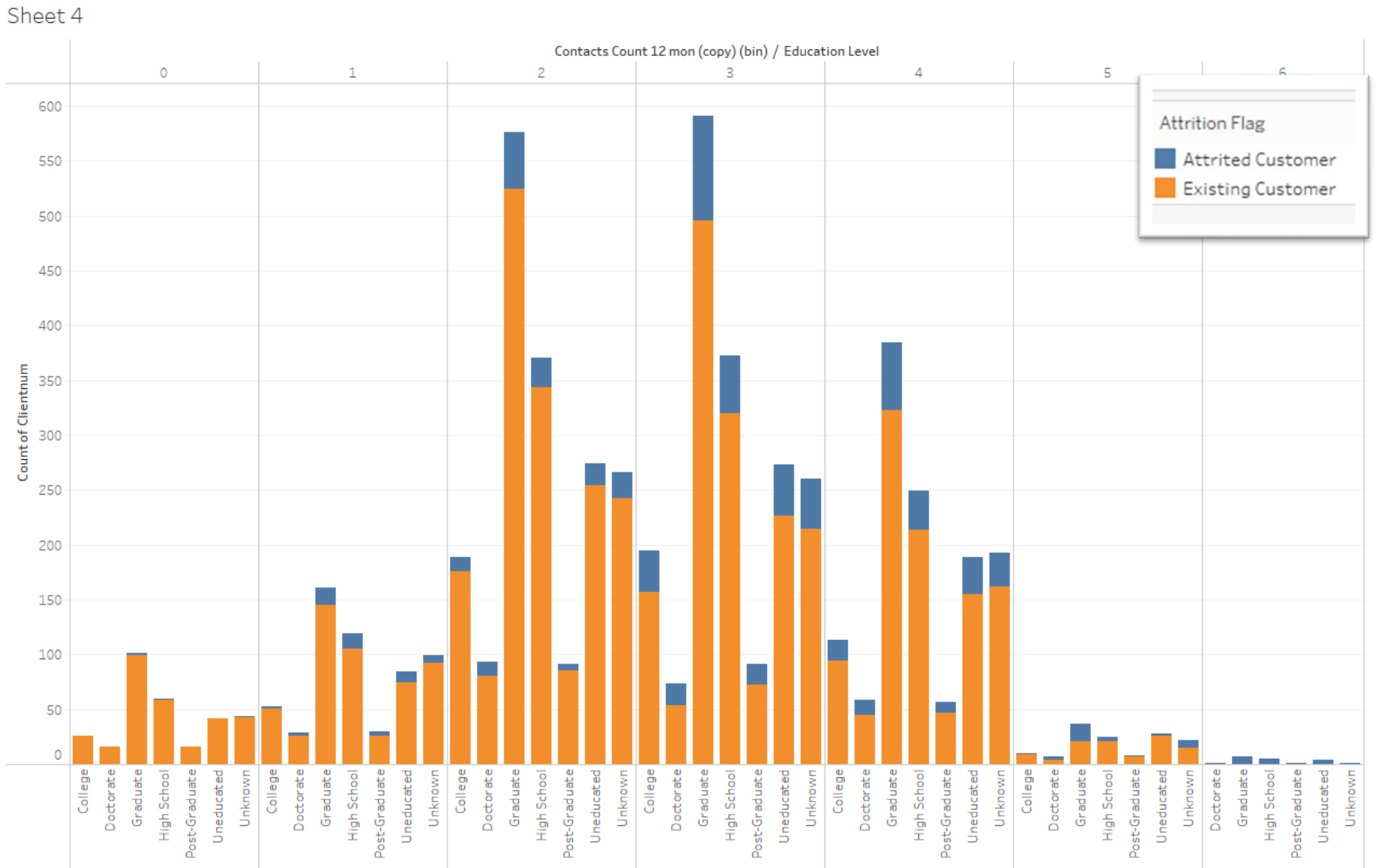


- The Contact Count 12-Month metric refers to the number of times a customer or client has reached out to a business or organization within a 12-month period. This metric is useful in gauging customer engagement and satisfaction levels.
- Upon filtering the database by income category and 12-month contact count, it is observed that customers with an income of less than $40,000 are more likely to leave the bank, while the remaining customers have a relatively similar ratio of leaving the bank or conducting business with it.

Vicky

Contacts Count 12 Mon & Education Level

X- Contact 12 month and Education Level
Y – Count of clientnum

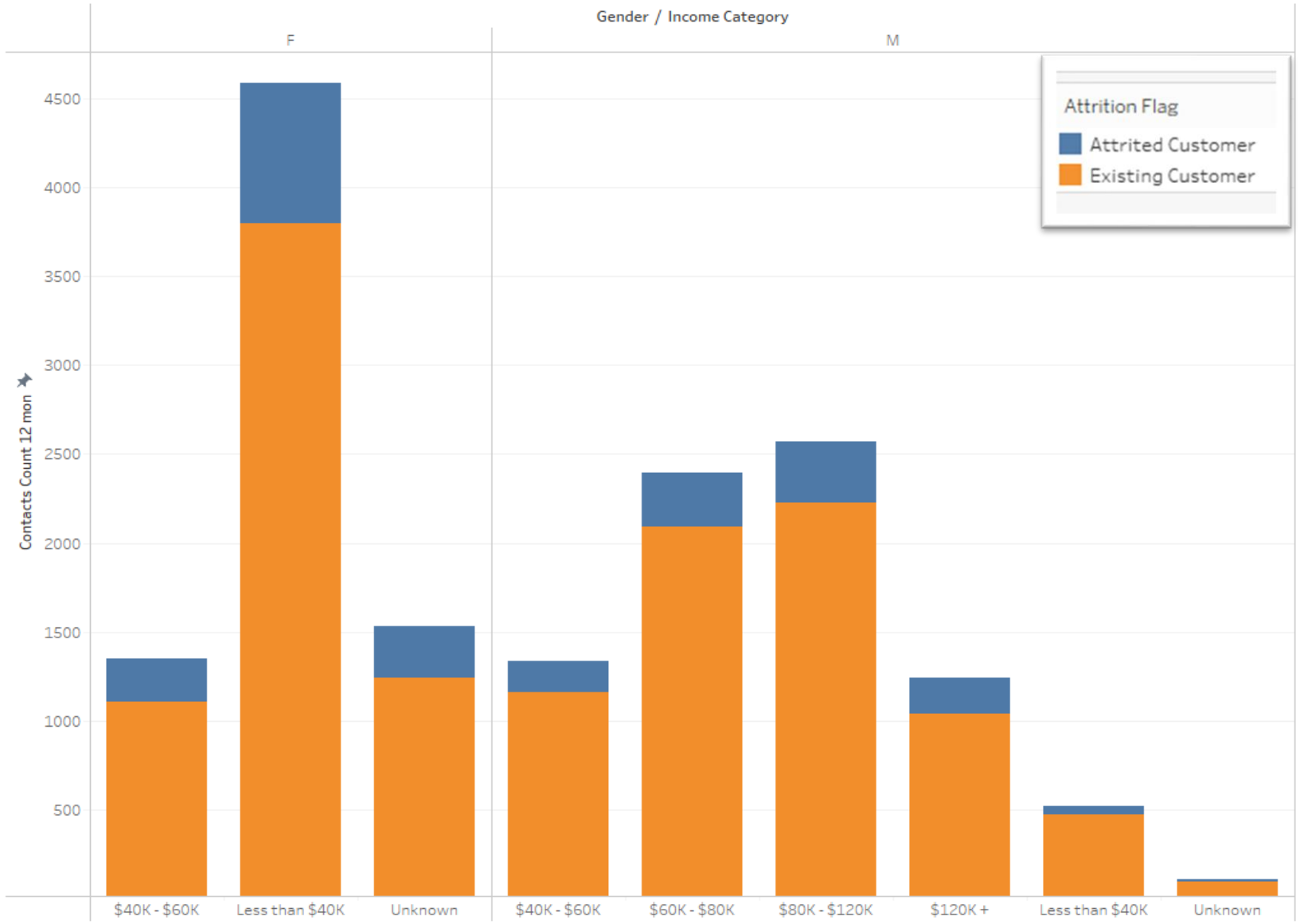


Based on Contact count and Education Level, we can see that Graduate are the one who have been contacted by the bank, Also most of the Graduates stopped doing business with the bank followed by Undergraduates and College students.

Vicky

Gender and Income Category

X-axis :- Gender and Income Category
Y- axis :- Contact count in 12 month

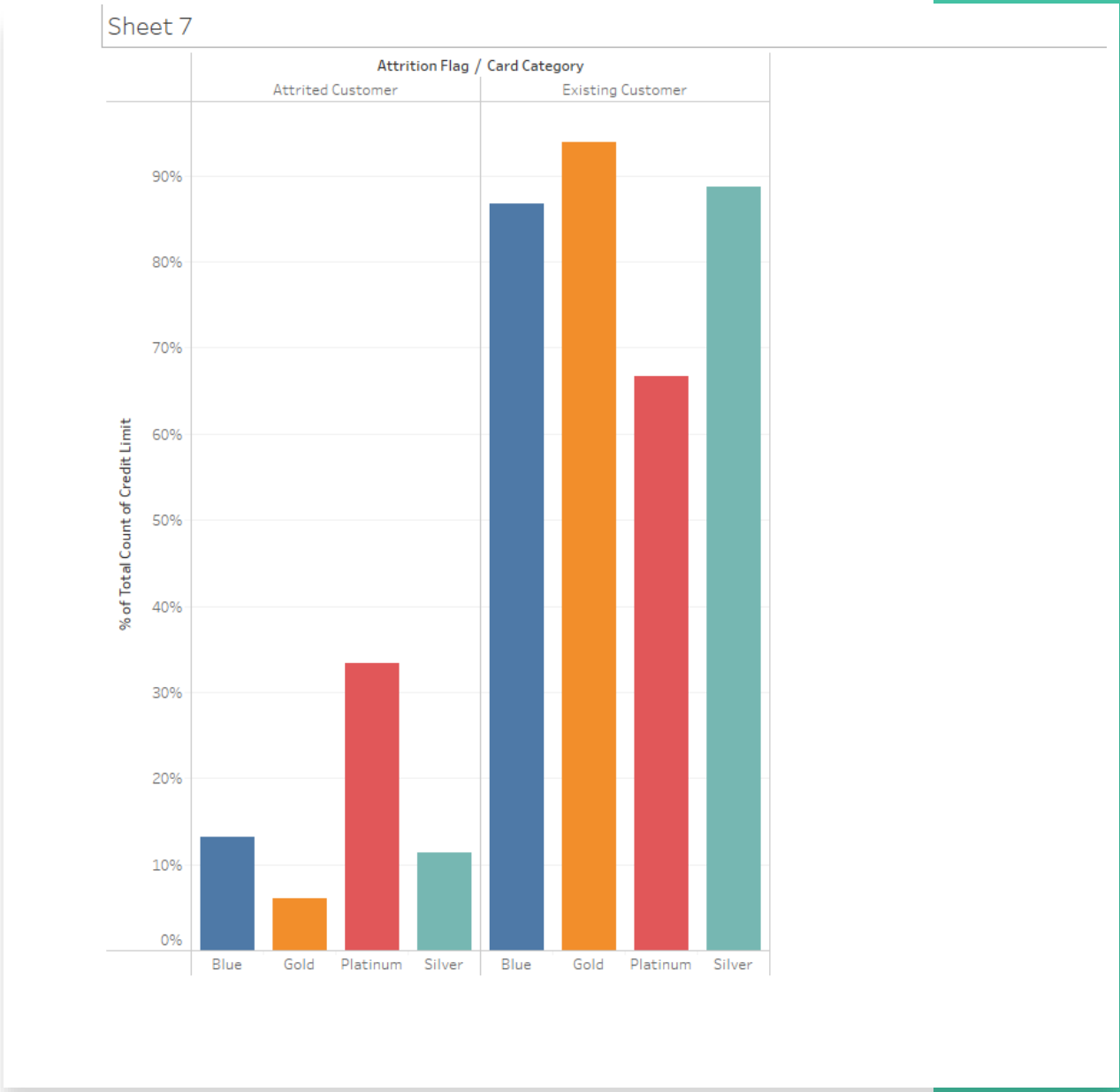


- Based on the visualization, it appears that females with an income below $40K are more likely to leave the bank. Additionally, males in the $60K to $80K and $80K to $120K income categories have also shown a tendency to stop doing business with the bank.

- As a result, the bank should consider focusing more on retaining female customers with lower incomes and males within the $60K to $120K income range.

Vicky

X- Card Category
Y- Total Credit Limit



In this visualization, it appears that the Platinum cardholders are the most likely to leave, followed by Blue, Silver, and Gold. Based on this information, it may be beneficial for the bank to focus more on retaining their Platinum cardholders.

"The count of Card Category and Total Count of Credit Limit is not providing significant insights into the reasons behind customer .I have focused on tracking the number of contacts made within the past 12 months, as well as the number of dependents associated with each account. This data allows for clear visualization of who customers have been in contact with and who may be leaving the bank."

Vicky

Professor Feedback

As per the professor's statement, the goal is not just to detect customers who are ceasing to use credit card services but also to suggest a resolution to tackle the problem.

One possible solution mentioned on slide 14 is to increase sales by providing more offers to platinum and gold members.

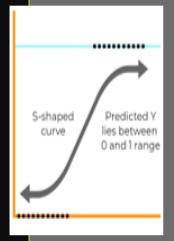
Another approach could be to target customers between the ages of 40-55 by offering them additional incentives.

Vicky

SUMMARY

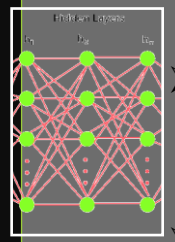
As per the analysis and after study of various Machine Learning (ML) Algorithms, we suggest the following ML models : -

Logistic Regression (binomial)



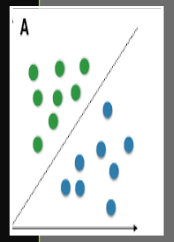
- It is commonly used for many binary classification tasks.
- Logistic regression is a powerful tool for analyzing bank data because it can help banks make informed decisions and customers analysis.

Deep Learning



- It is computer model which learns to perform classification tasks directly from dataset.
- Deep learning models can achieve state-of-the-art accuracy, sometimes exceeding human-level performance.
- Deep learning algorithms can automatically learn features from the input data and make complex decisions based on those features

Naive Bayes



- Naive Bayes is that it is relatively simple and easy to implement. It does not require complex parameter tuning, and it is not prone to overfitting.
- This makes it a popular choice for small and medium-sized banks that may not have the resources to implement more complex machine learning algorithms.

THANK YOU!

Any Questions