Quantitative Techniques

Customer Churn Analysis for "ABC" Bank

(a European bank, operating in France, Spain and Germany)

Table of Contents

Table of Contents 3
Table of Figures 4
Table of Figures 5
1. Introduction 6
1.1. Motivation 6
1.2. Objective 6
1.3. Outcomes 6
2. Dataset 7
2.1. Data Source 7
2.2. Data Sample 7
2.3. Dataset 7
3. Variables 8
3.1. Variable definition 8
4. Descriptive Analysis 9
4.1. Data Visualization - Pie Charts 9
4.2. Data Visualization – Bar Graphs 11
4.3. Data Visualization – Line Charts 13
4.4. Data Visualization – Side-by-Side Bar Graphs 14
4.5. Data Visualization – Histograms 18
4.6. Scatter Plot 22
4.7. Numerical Summary of Variables 25
4.8. Box and Whisker Plot 27
5. Probability 29
5.1. Operations with Probability 29
5.2. Probability Distribution 30
6. Inferential Statistics 31
6.1. Population and Parameter 31
6.2. Sample and Statistics 31
6.3. Sampling Distribution of Statistic 31
6.4. 95% Confidence Interval of Parameter 32
6.5. Hypothesis Testing 32
7. Multiple Linear Regression 34
7.1. Regression Coefficients 34
7.2. Model Assumption Validation 35
7.3. Variables Summary Outcomes 36
8. ANOVA 37
9. Conclusion from Analysis 38
10. Contribution of Team Members 39

Table of Figures

Chart 1- Churn Rate 9
Chart 2 – Gender Distribution 9
Chart 3 – Geography Distribution 9
Chart 4 – Credit Card Distribution 10
Chart 5 – Active Member Distribution 10
Chart 6 – Age Distribution 11
Chart 7 – Estimated Salary Distribution 11
Chart 8 – Balance Distribution 11
Chart 9 – Credit Score Distributions 12
Chart 10 – Customer Tenure Distribution 12
Chart 11 – Number of Products Distribution 12
Chart 12 – Tenure of Customers 13
Chart 13 – Age of Customers 13
Chart 14 – Churn Rate vs Age 14
Chart 15 – Churn Rate vs Gender and Country 14
Chart 16 – Churn Rate vs Balance 15
Chart 17 – Churn Rate vs Estimated Salary 15
Chart 18 – Churn Rate vs Credit Score 15
Chart 19 – Churn Rate vs Credit Card 16
Chart 20 – Churn Rate vs Active Member 16
Chart 21 – Churn Rate vs Tenure 16
Chart 22 – Churn Rate vs Number of Products 17
Chart 23 – Age Histogram 18
Chart 24 – Balance Histogram 18
Chart 25 – Salary Histogram 19
Chart 26 – Credit Score Histogram 19
Chart 27 – Tenure Histogram 20
Chart 28 – Number of Products Histogram 20
Chart 29 – Age vs Tenure 22
Chart 30 – Age vs Balance 22
Chart 31 – Age vs Credit Score 23
Chart 32 – Age vs Estimate Salary 23
Chart 33 – Age vs Churn 27
Chart 34 – Age vs Churn per Country 27
Chart 35 – Tenure vs Churn 28
Chart 36 – Normal Probability Plot 35
Chart 37 – Age Residual Plot 35
Chart 38 – Estimated Salary Residual Plot 35
Chart 39 – Balance Residual Plot 36

Table of Figures

TABLE 1 – DATA SAMPLE 7
TABLE 2 – VARIABLES DEFINITION 8
TABLE 3 – CORRELATION MATRIX 24
TABLE 4 – SUMMARY DESCRIPTIVE OF VARIABLES 25
TABLE 5 – AGE GROUP DISTRIBUTION 29
TABLE 6 – NUMBER OF PRODUCTS DISTRIBUTION 29
TABLE 7 – HYPOTHESIS TEST FOR CREDIT SCORE 32
TABLE 8 - HYPOTHESIS TEST FOR AGE 33
TABLE 9 – MULTIPLE LINEAR REGRESSION VARIABLES 34
TABLE 10 – REGRESSION COEFFICIENTS 34
TABLE 11 – MULTIPLE LINEAR REGRESSION OUTCOME 36
TABLE 12 – ANOVA 37

1. Introduction

1.1. Motivation

Banking is one of the industries we can all relate to as we all use banks in our day to day life and probably one of the common services we all used. As a group, when we started to discuss the topics, everyone came up with different options but most of us agreed to pick up this topic as we have all used bank in our lives and all of us have changed our preferred banks for one or other regions. So, we decided to use bank churn analysis as our topic.

In one of the studies related to banking industry, it was derived that cost of one new customer joining the bank is 7 times higher than retaining existing ones. This clearly shows that churning is one of the major issues for a bank and also if bank needs to be profitable and save costs, it had to tackle churning efficiently.

Banking is one of the largest industries around the world with more than 8 trillion dollars market cap. Also, over the decades, newer technologies and transformations has led to more features in banking services and in turn increased the competition. Better services, products, wealth etc. have given customer the power to choose any bank which they want and for that matter even switch the bank whenever they want.

Data, now a days, is considered as an invaluable asset for every industry and banks are no different in using to successfully understand today's ever-changing environment. Data analysis can help a bank to differentiate themselves with peers and take competitive edge through achieving a better understanding of customers.

1.2. Objective

"ABC" bank has new CEO and his aim is to cut the cost of the company and increase profits. To achieve this, he is focusing on a key metrics of the industry "Churn Rate". So, he gathered sample data from the bank and trying to focus on below points:

- Identify and visualize what are the factors which contribute to customer churn.
- Classify customers who are more likely to leave the bank (i.e. close their bank account).

This will make it easier for customer service teams to target specific customers in their efforts to prevent churn.

1.3. Outcomes

"ABC" bank CEO is expecting below mentioned outcomes from this analysis:

- Identify important factors that affects the customer churning from the bank.
- Using the different types of analysis on sample data, try to better understand the customer behavior and help bank to retain the customers.

2. Dataset

2.1. Data Source

The relevant dataset which is being used to carry out the churn analysis is taken from Kaggle platform. The dataset contains 10,000 observations and 14 variables.

URL - https://www.kaggle.com/kmalit/bank-customer-churn-prediction

2.2. Data Sample

Data of the chosen dataset is shown below

Row Number	CustomerId	Surname	CreditScore	Geography	Gender	Age	Tenure	Balance	NumOfProducts	HasCrCard	IsActiveMember	EstimatedSalary	Exited
1	15634602	Hargrave	619	France	Female	42	2	0	1	1	1	101348.88	1
2	15647311	Hill	608	Spain	Female	41	1	83807.86	1	0	1	112542.58	0
3	15619304	Onio	502	France	Female	42	8	159660.8	3	1	0	113931.57	1
4	15701354	Boni	699	France	Female	39	1	0	2	0	0	93826.63	0
5	15737888	Mitchell	850	Spain	Female	43	2	125510.82	1	1	1	79084.1	0
6	15574012	Chu	645	Spain	Male	44	8	113755.78	2	1	0	149756.71	1
7	15592531	Bartlett	822	France	Male	50	7	0	2	1	1	10062.8	0
8	15656148	Obinna	376	Germany	Female	29	4	115046.74	4	1	0	119346.88	1
9	15792365	He	501	France	Male	44	4	142051.07	2	0	1	74940.5	0
10	15592389	H?	684	France	Male	27	2	134603.88	1	1	1	71725.73	0

Table 1 – Data Sample

2.3. Dataset

Attached file is the dataset used for analysis and project workbook

3. Variables

3.1. Variable definition

The chosen data set has 14 variables defined as:

Variables	Description	Type of Variable	Level of Measurement
Row Number	Row index	Qualitative	Ordinal
Customer Id	Unique ID of the customer	Qualitative	Nominal
Surname	Last name of the customer	Qualitative	Nominal
Credit Score	Credit score of the customer. The range of credit score is from 350 to 850	Quantitative	Ratio
Geography	Country of the customer (France, Germany, Spain)	Qualitative	Nominal
Gender	Gender of the customer (Male, female)	Qualitative	Nominal
Age	Age of the customer. The range of customer's age is from 18 to 92	Qualitative	Ordinal
Tenure	Tenure of the customer in years. It is the customer has stayed with the bank.	Quantitative	Ratio
Balance	The amount of money available for withdrawal for a customer	Quantitative	Ratio
Number of Products	Number of products customer uses from bank such as credit card , savings account, salary account etc.	Quantitative	Ratio
Has Credit Card	Whether customer has credit card (1 for Yes, 0 for No)	Qualitative	Ordinal
Active Member	Whether customer is an active member (1 for Active, 0 for Not Active)	Qualitative	Ordinal
Estimated Salary	Estimated Salary of the customer	Quantitative	Ratio
Exited	Whether customer has churned or not (1 for exited & 0 for retained)	Qualitative	Ordinal

Table 2 – Variables Definition

4. Descriptive Analysis

This section will cover the analysis of the data with respect of churning of customers to identify important variables.

4.1. Data Visualization - Pie Charts

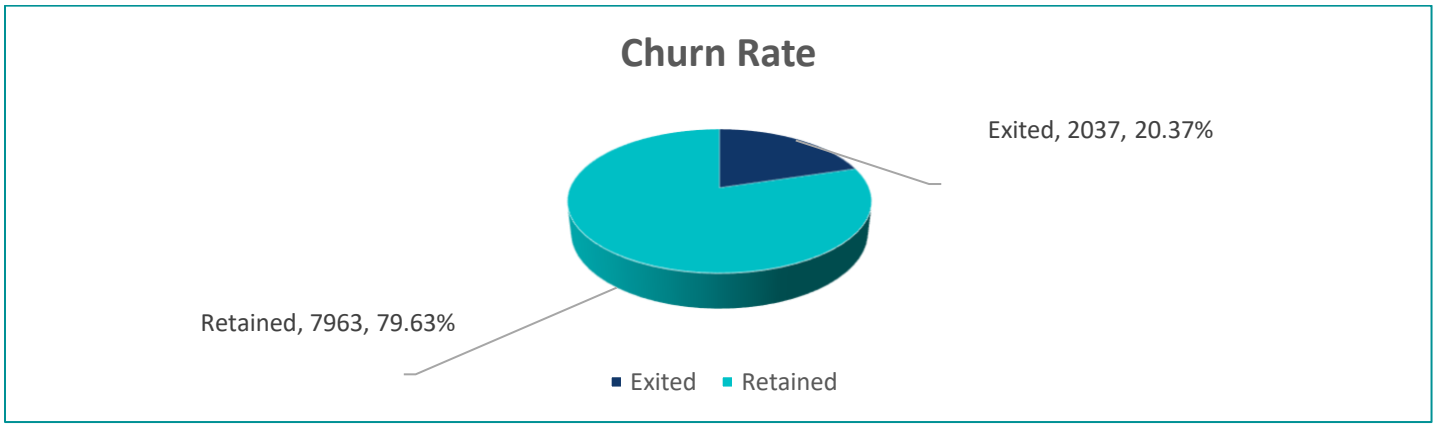

Chart 1- Churn Rate

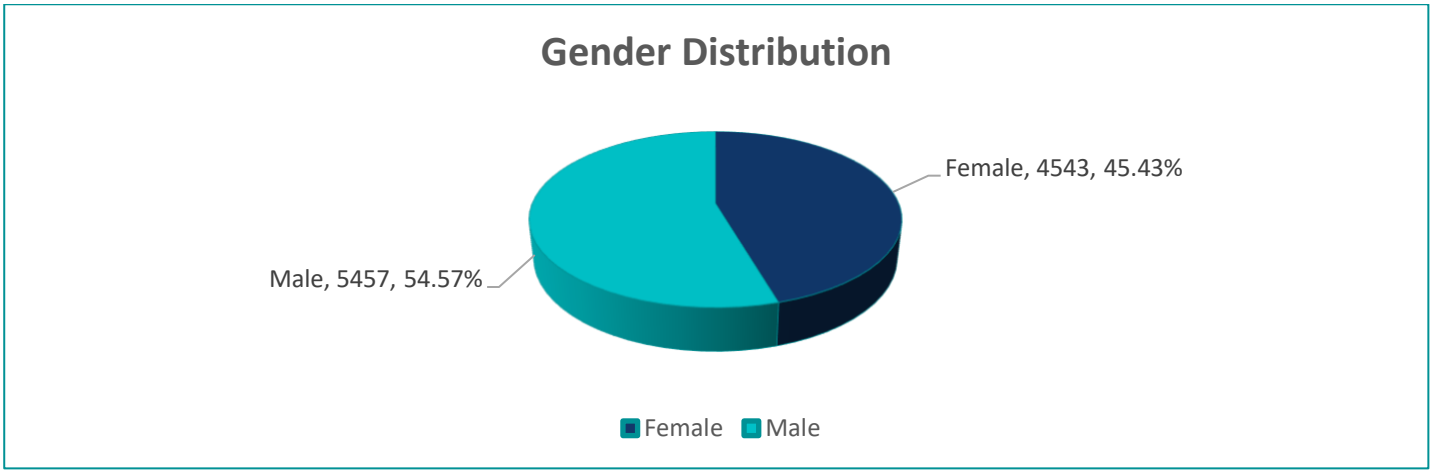

Chart 2 – Gender Distribution

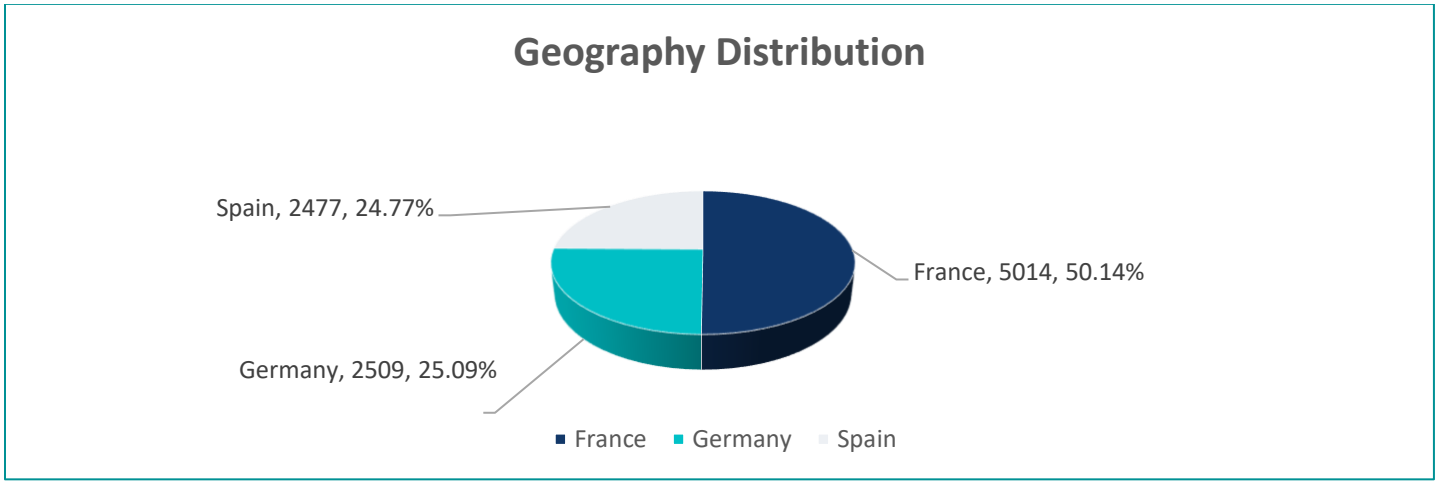

Chart 3 – Geography Distribution

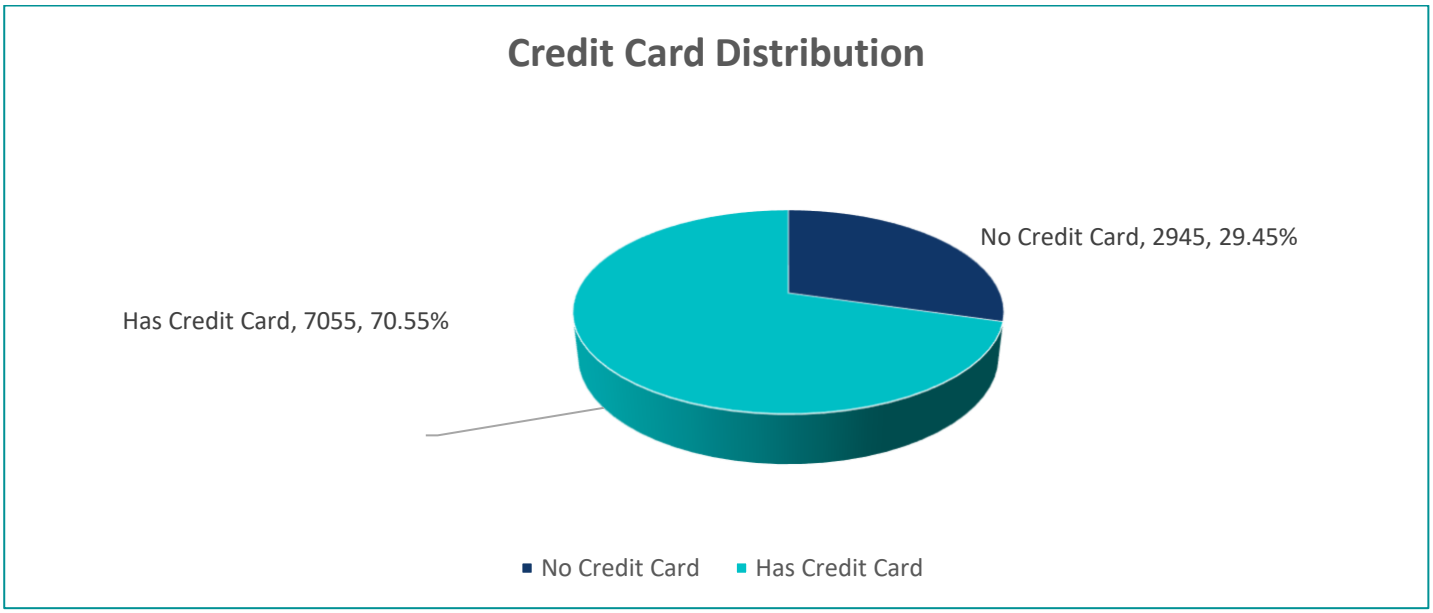

Chart 4 – Credit Card Distribution

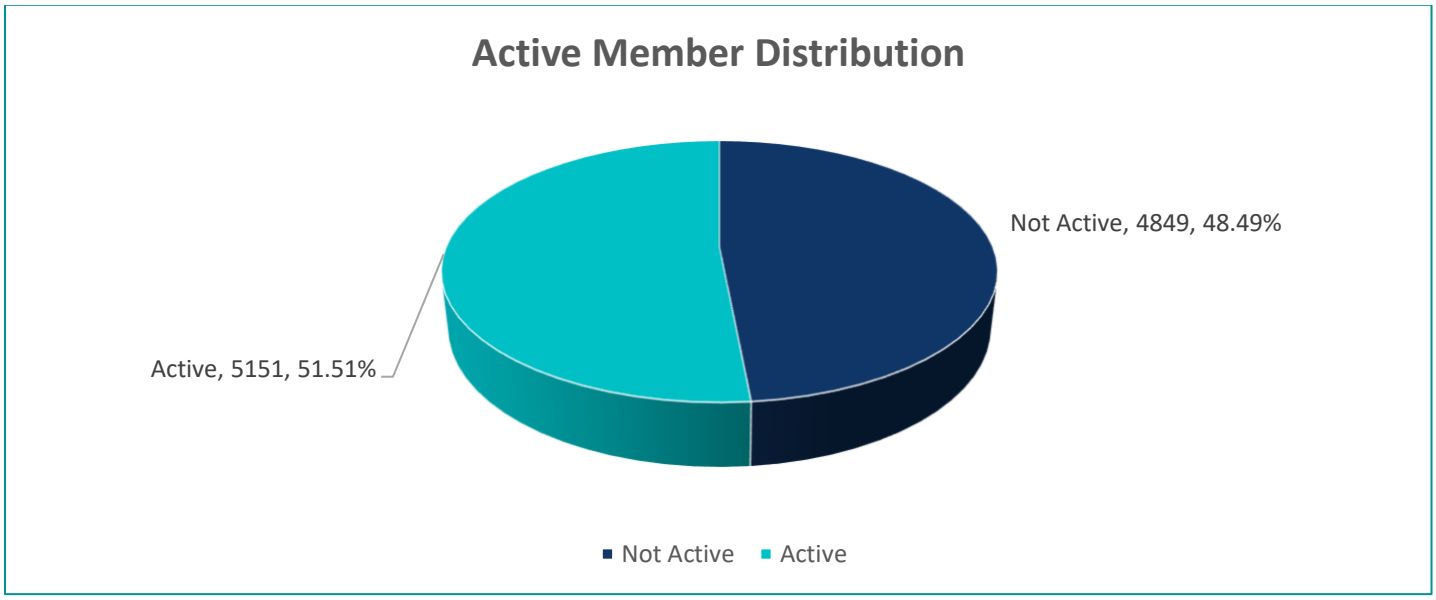

Chart 5 – Active Member Distribution

Observations

- **20.37%** of 10,000 customers have left the bank. Churn rate of 10,000 customers is 20.37%.
- Almost equal number of male and female customers. **50.57%** of 10,000 customers are **Male** and remaining **45.43%** are **females**.
- ABC bank have strong presence in France as **50.14%** of 10,000 customers are from **France** and rest **25.09% and 24.77%** are from **Germany and Spain** respectively.
- **Only 29.45%** of 10,000 customers **does not have credit card.**
- **48.49%** of 10,000 customers are **not active**. Might be a cause of worry for the bank.

4.2. Data Visualization – Bar Graphs

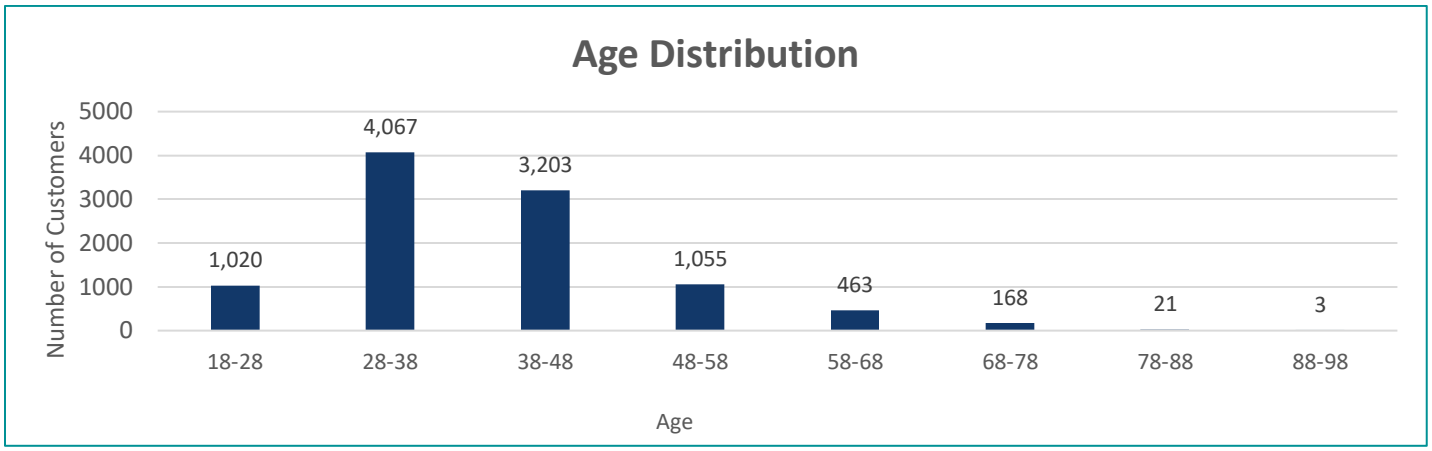

Chart 6 – Age Distribution

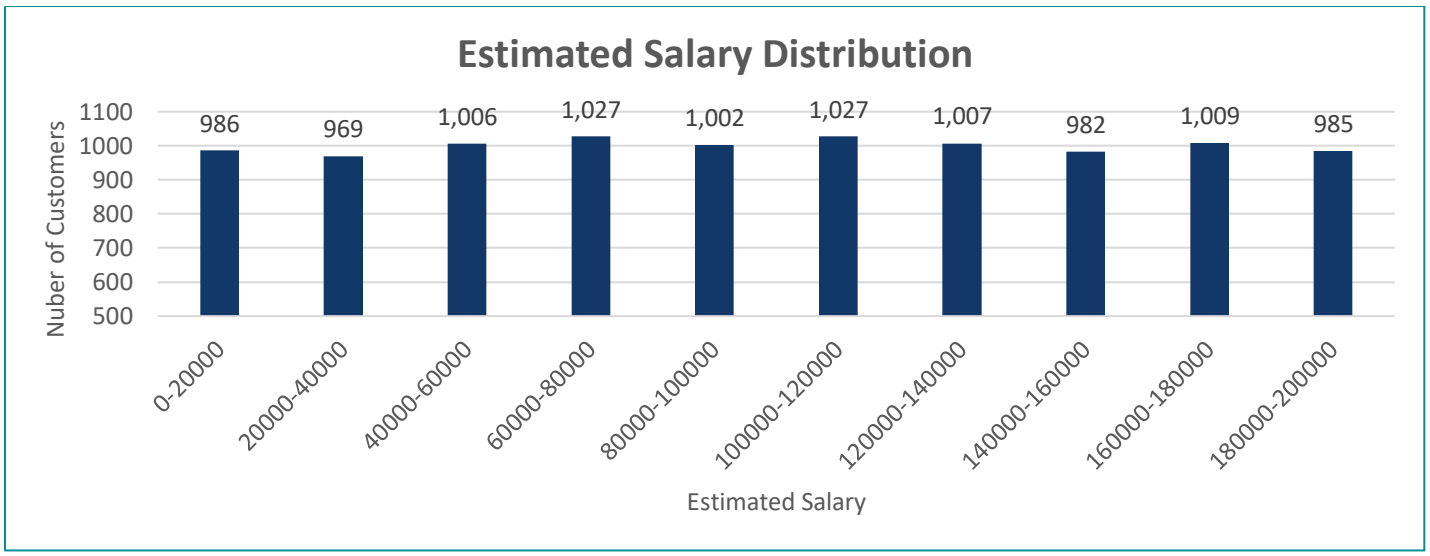

Chart 7 – Estimated Salary Distribution

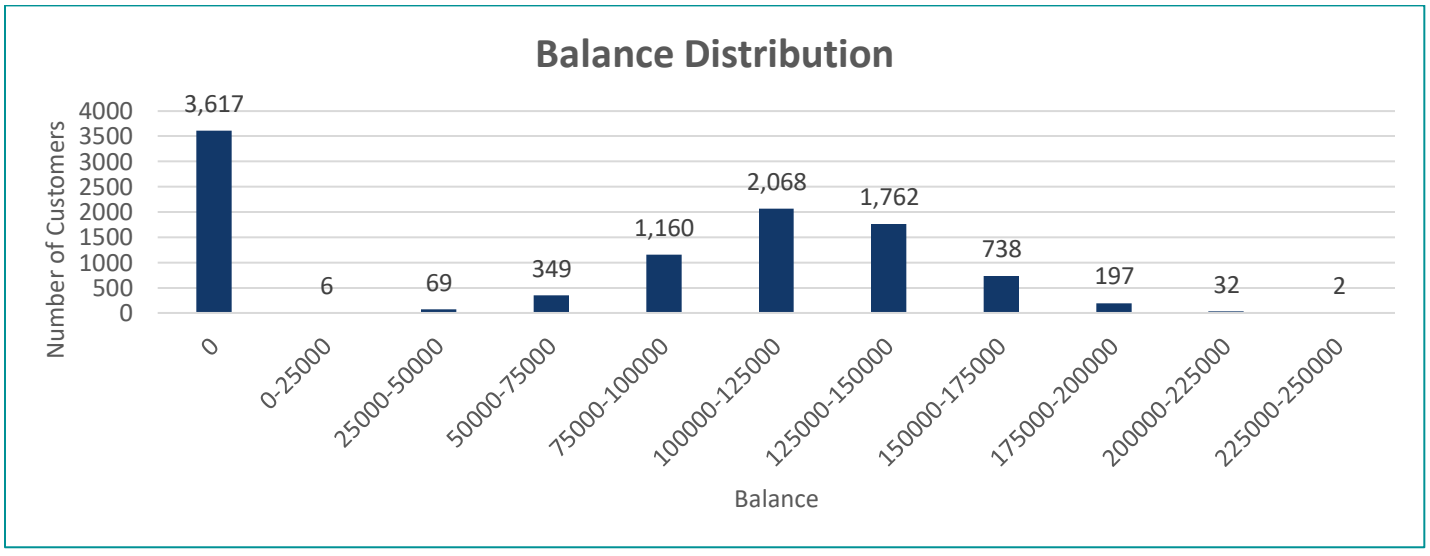

Chart 8 – Balance Distribution

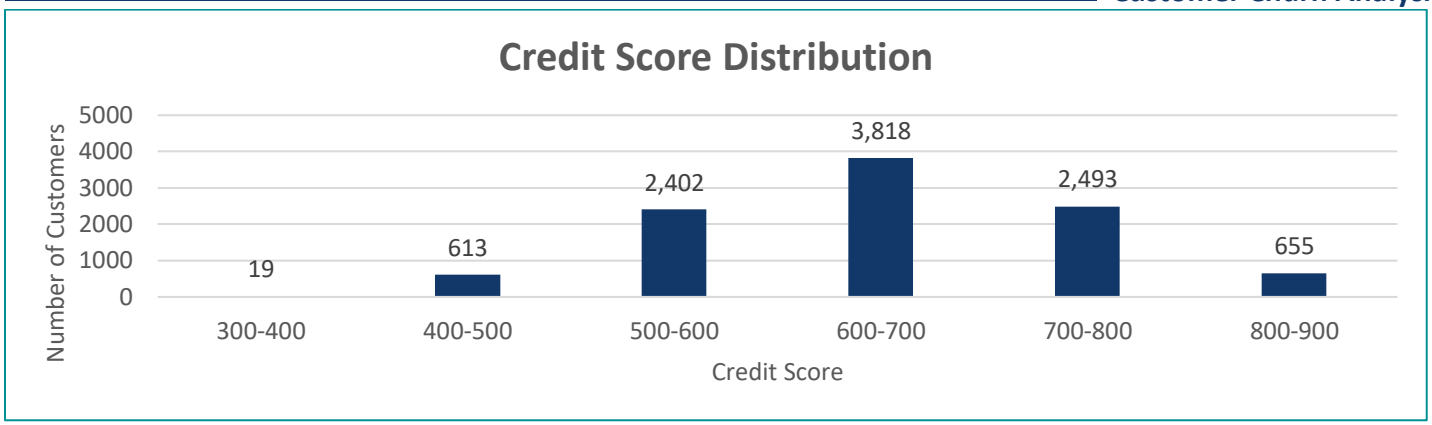

Chart 9 – Credit Score Distributions

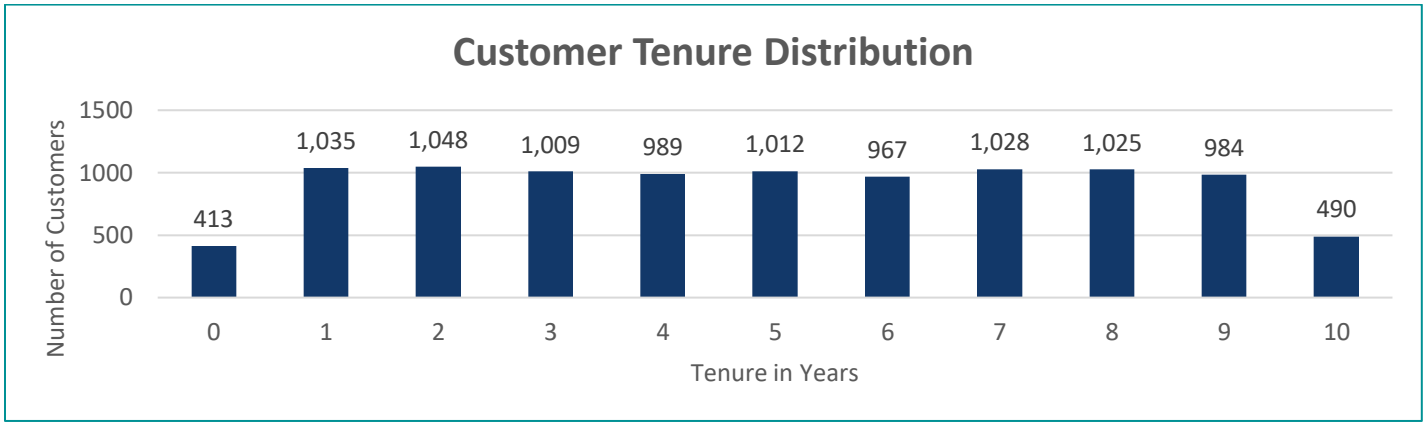

Chart 10 – Customer Tenure Distribution

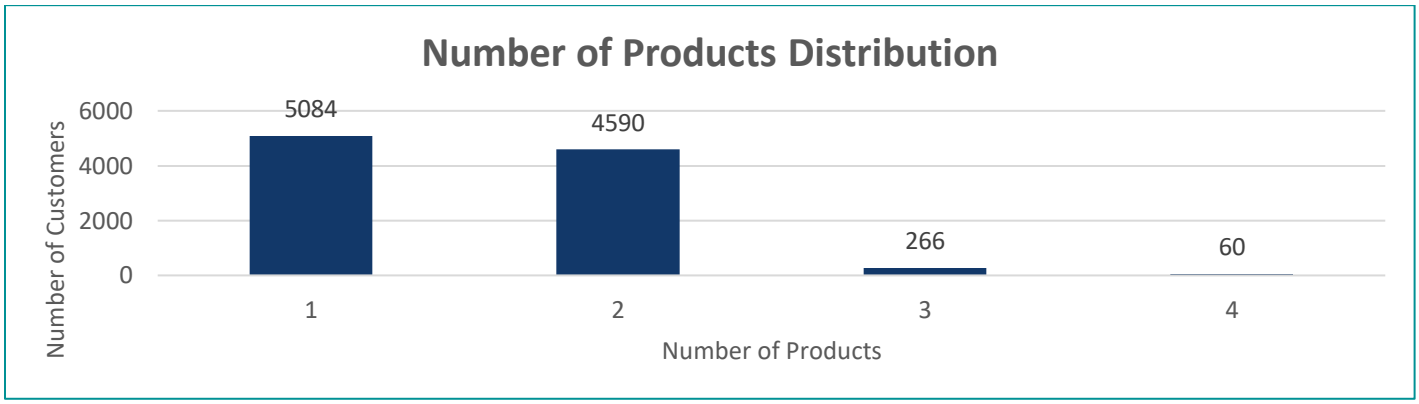

Chart 11 – Number of Products Distribution

Observations

- **3617** of 10,000 customers have Zero balance in bank.
- Bank customers are almost **evenly distributed** amongst various **salary ranges**.
- Age distribution is right skewed and showing **85%** of 10,000 customers are **under age 48**. This shows that **bank is not a preferred bank for customer older than 50 years**.
- Majority of 10,000 customers are within 500-800 credit score range.
- **90%** of 10,000 customers have **tenures less than 9 years** in bank.
- **96%** of 10,000 customers are using **either one or two products**, this shows that **bank is facing difficulty to sell more than two products to the customer**.

4.3. Data Visualization – Line Charts

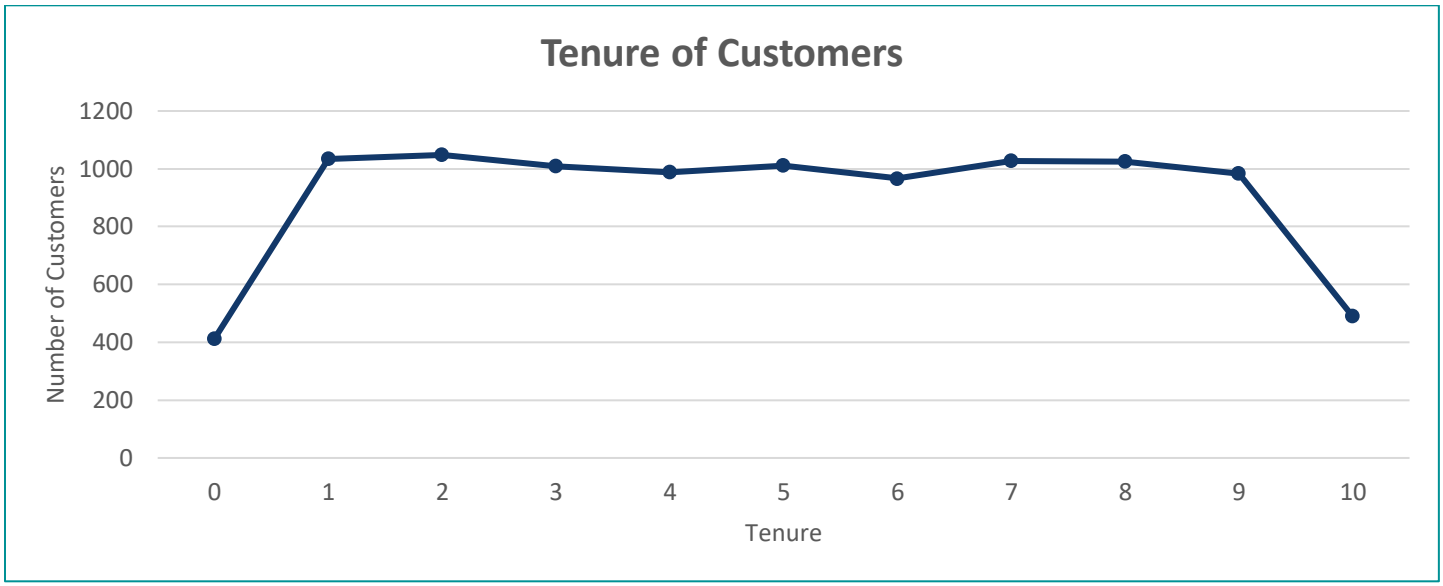

Chart 12 – Tenure of Customers

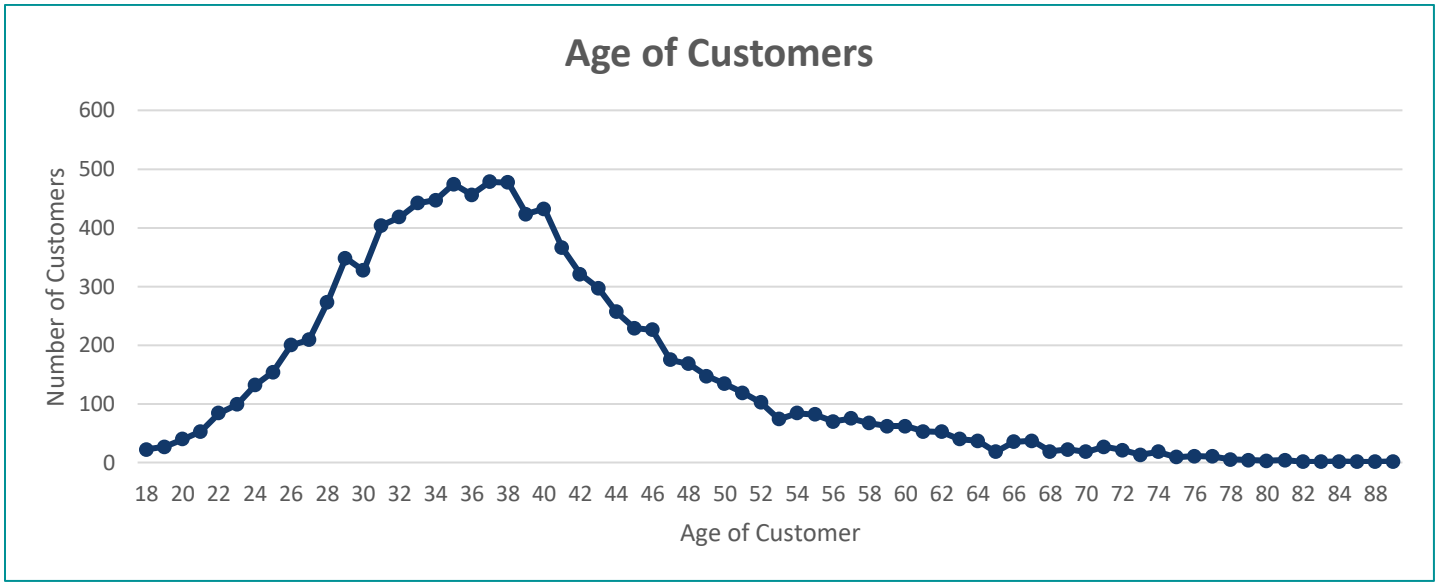

Chart 13 – Age of Customers

Observations

- After a tenure of 9 years there is sudden drop in number of customers.
- As soon as customer age go beyond 50, number of customers starts leaving the bank

4.4. Data Visualization – Side-by-Side Bar Graphs

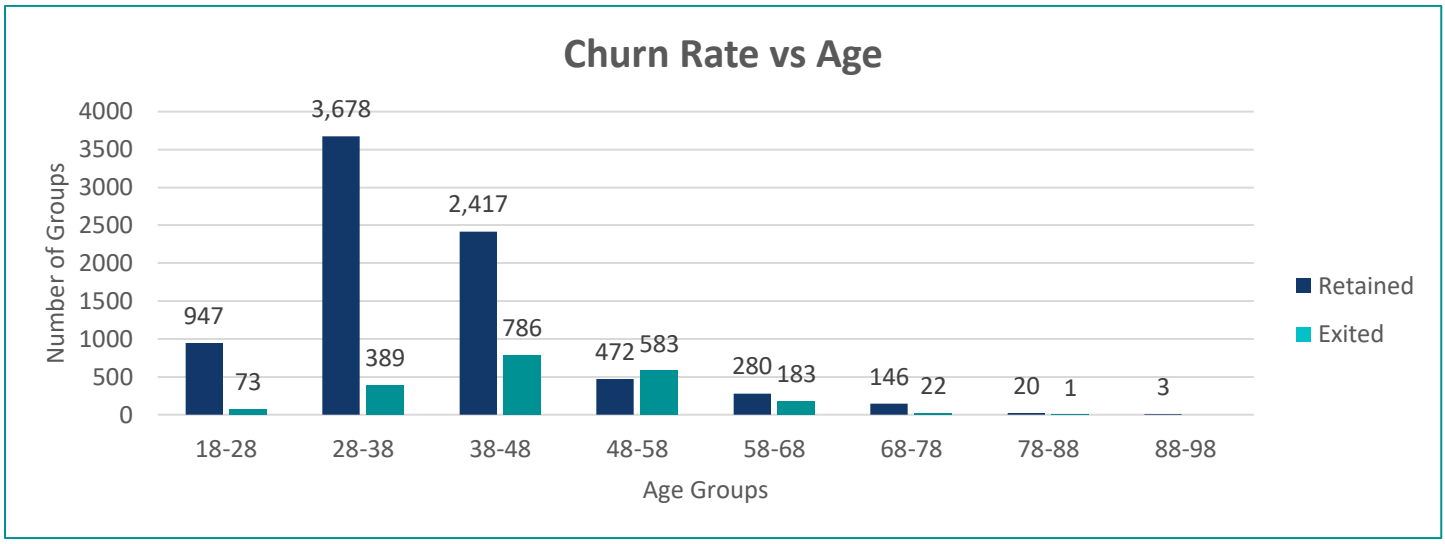

Chart 14 – Churn Rate vs Age

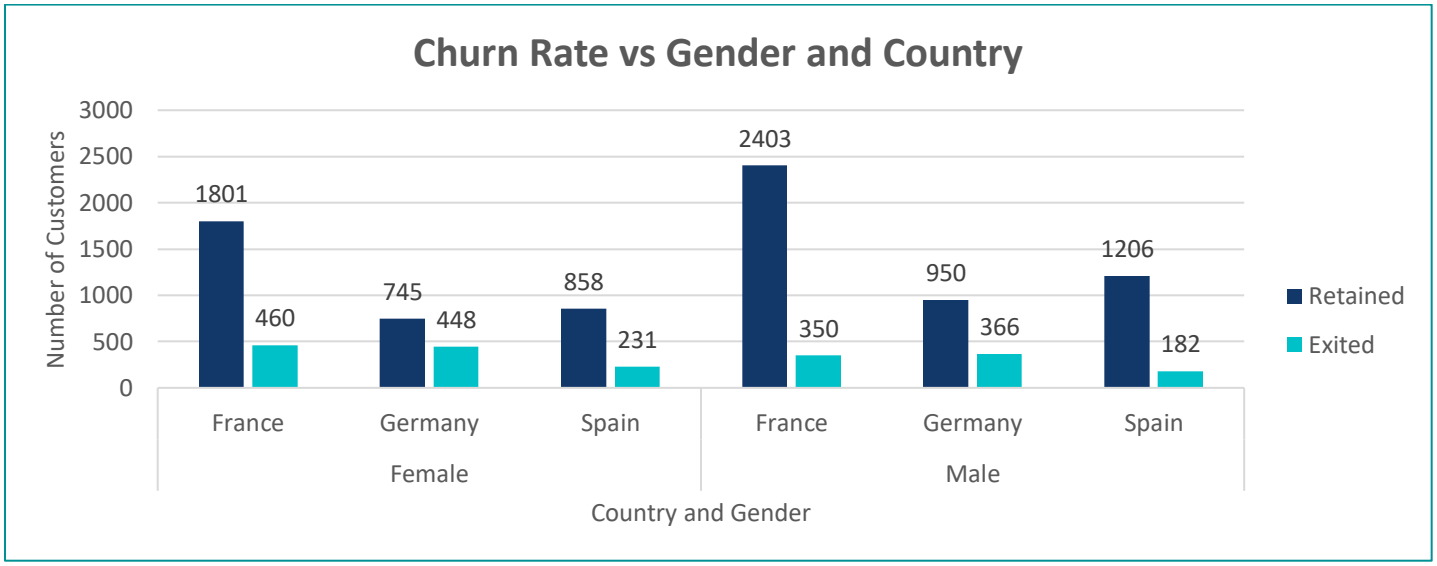

Chart 15 – Churn Rate vs Gender and Country

Observations

- **Maximum churning rate** is between **age 38-68** for 10,000 customers.
- Bank is able to retain younger generation but **not able to retain customers** where **age is greater than equal to 38.**
- Churning rate percentage is **highest** in **age group 48-58.**
- While the male churn rate is 16.46% and that of the **females is 25.07%** in 10,000 customers. In absolute terms also, rate of the **number of females exiting the bank is more than the males** even though the total numbers are much lesser.
- The percentage of **customers quitting the bank in Germany is significantly high at 32.44%** and out of these, the **females are quitting even more than males at 37.55%.**

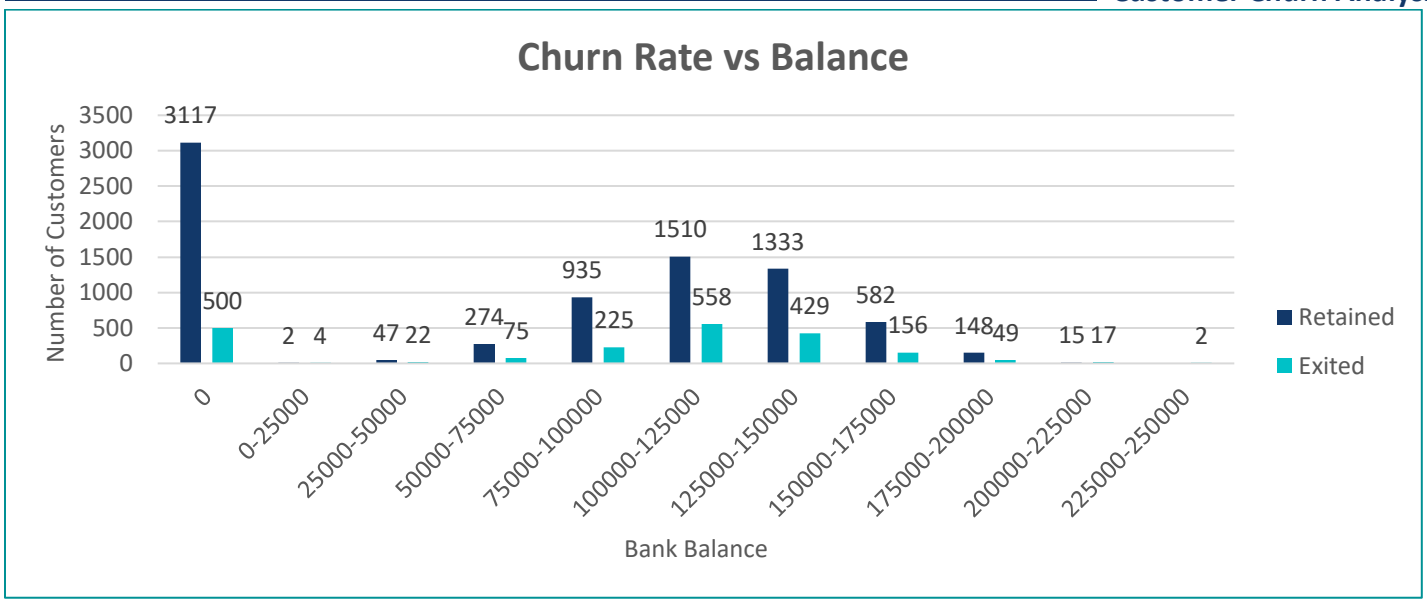

Chart 16 – Churn Rate vs Balance

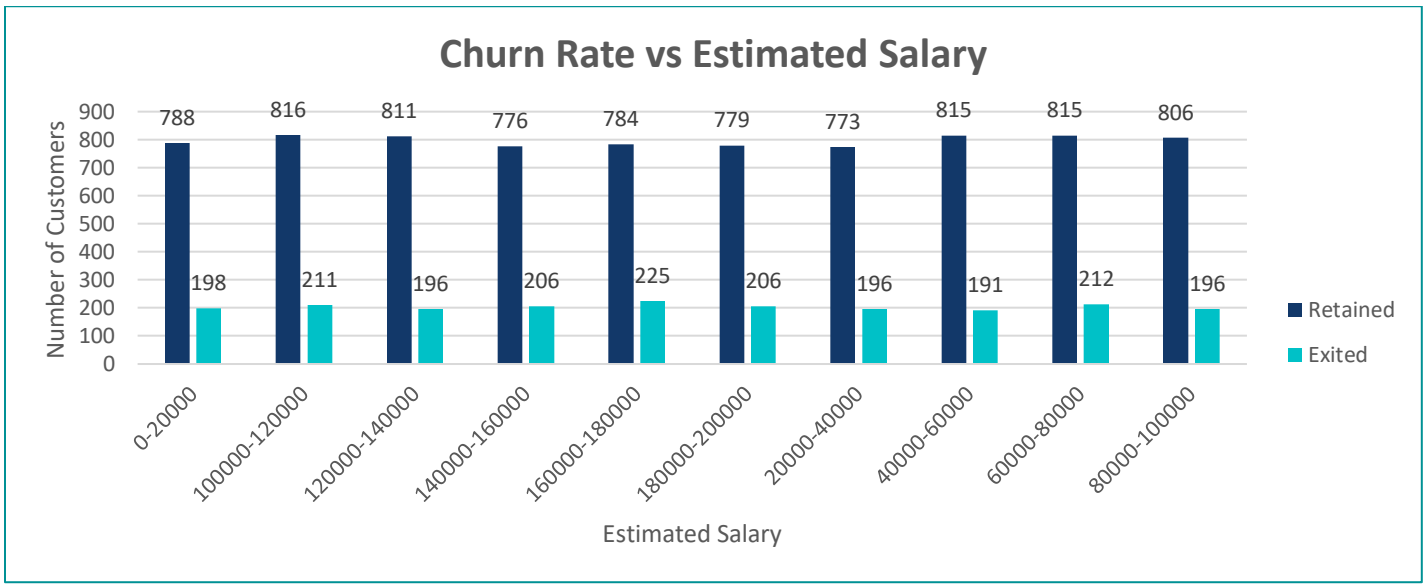

Chart 17 – Churn Rate vs Estimated Salary

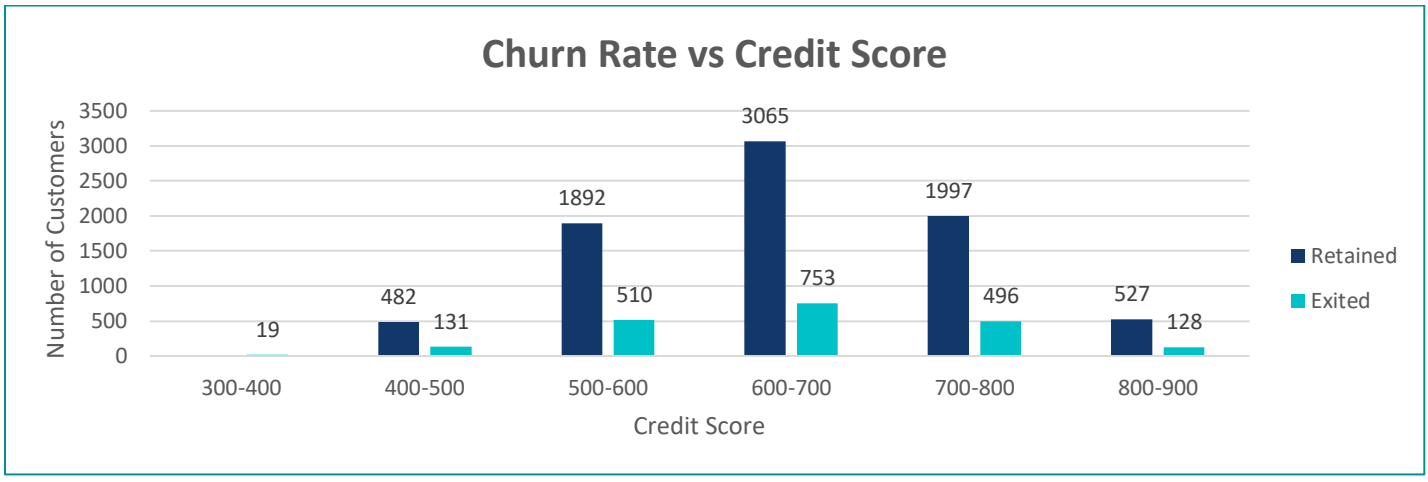

Chart 18 – Churn Rate vs Credit Score

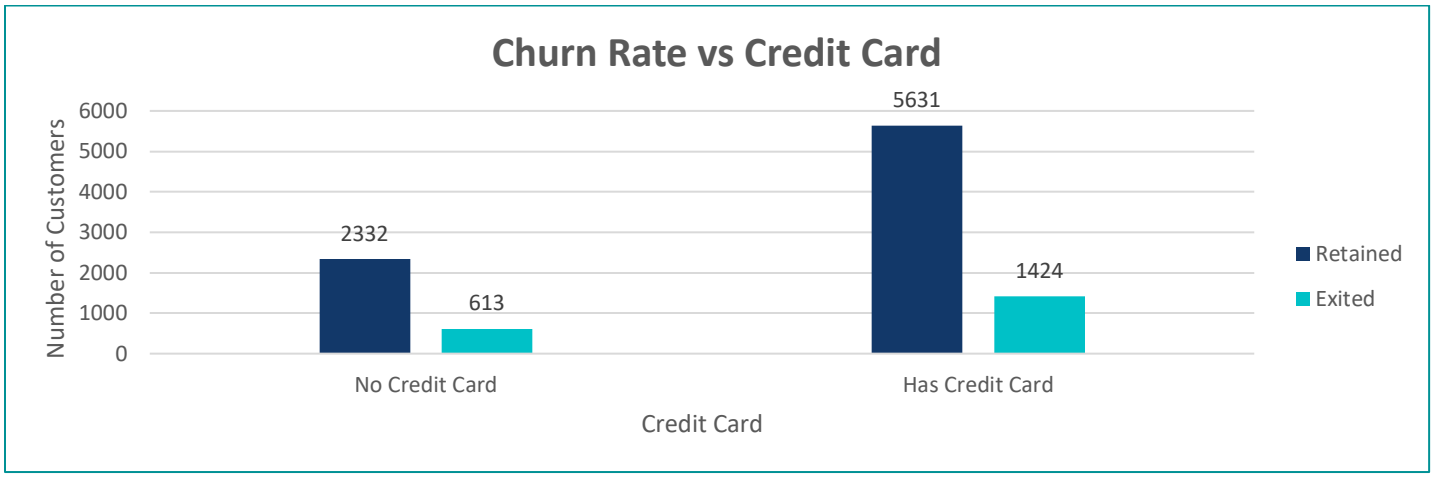

Chart 19 – Churn Rate vs Credit Card

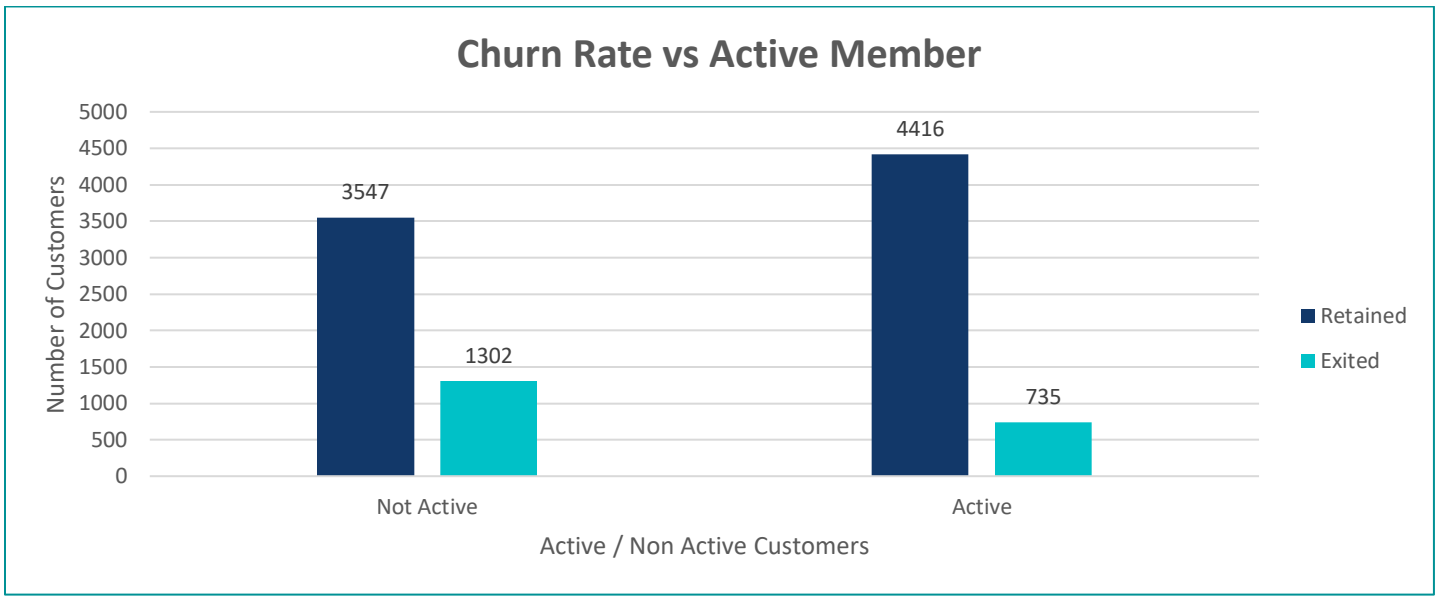

Chart 20 – Churn Rate vs Active Member

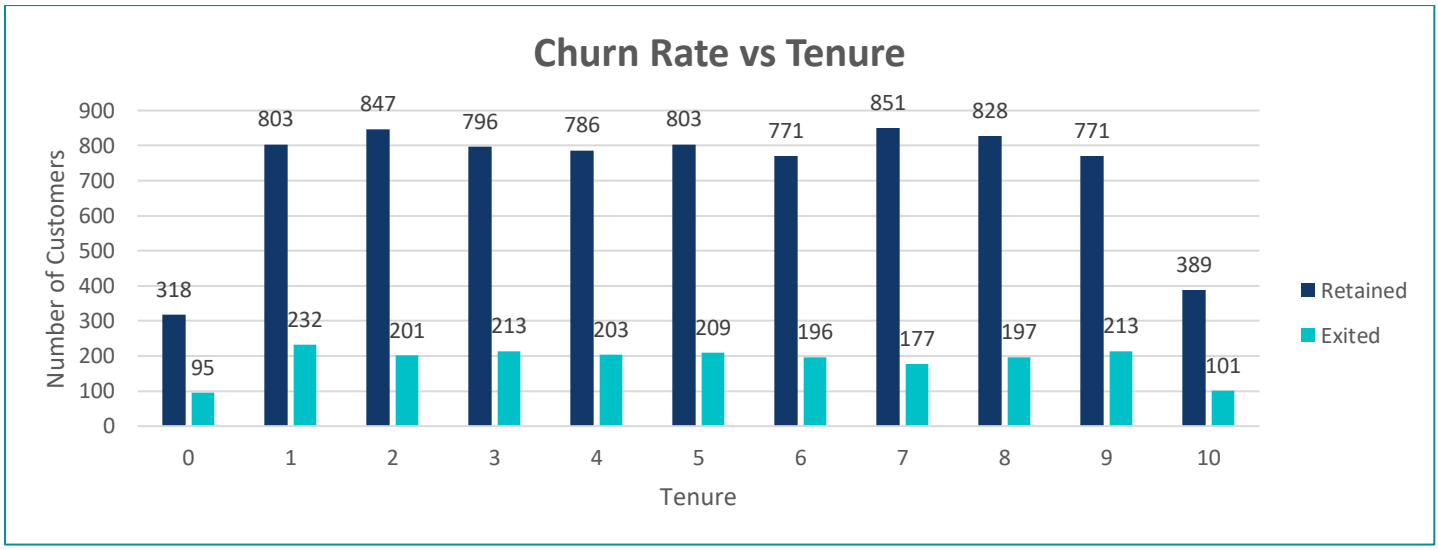

Chart 21 – Churn Rate vs Tenure

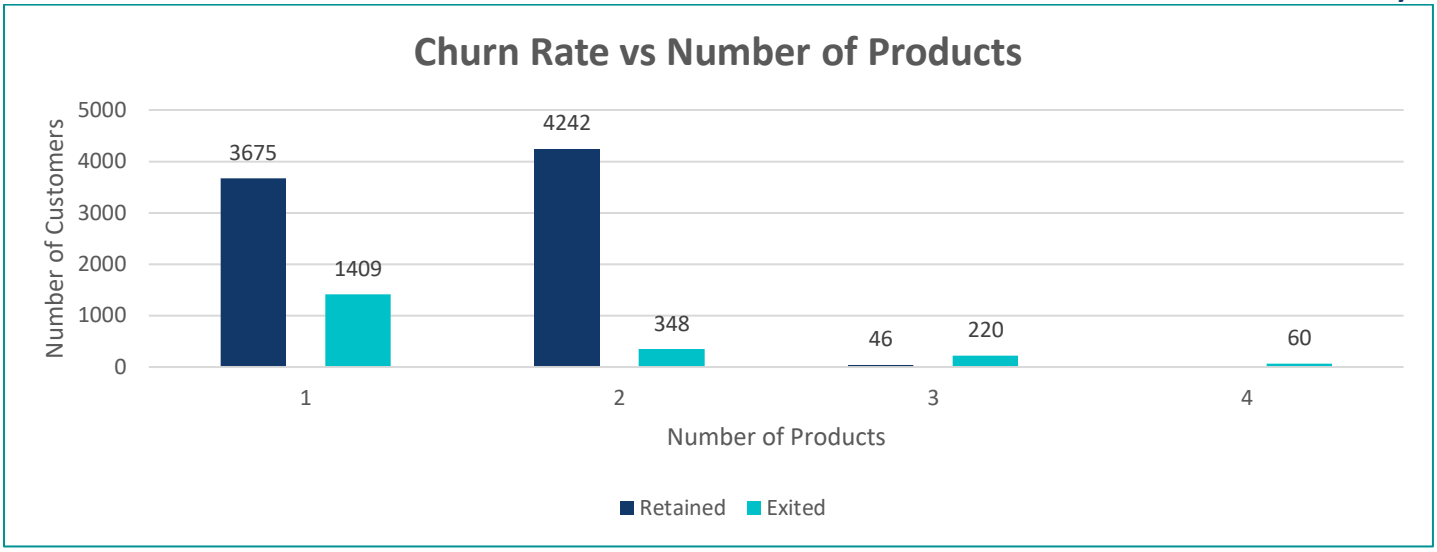

Chart 22 – Churn Rate vs Number of Products

Observations

- Bank balance, estimated salary and credit score does not seem to have any impact on churning as we can see a constant rate of churning.
- Customers having credit card or tenure with bank also seem to have no impact on churning.
- **26.85%** of 10,000 customers who are **not active** seems to leave bank more than **14.27%** of customers who are **active**.
- The churn rate with customers having **3 or 4 products** in the bank **is extremely high** for 10,000 customers. Churn rate is **7.58% is lowest for customers having 2 products**.

4.5. Data Visualization – Histograms

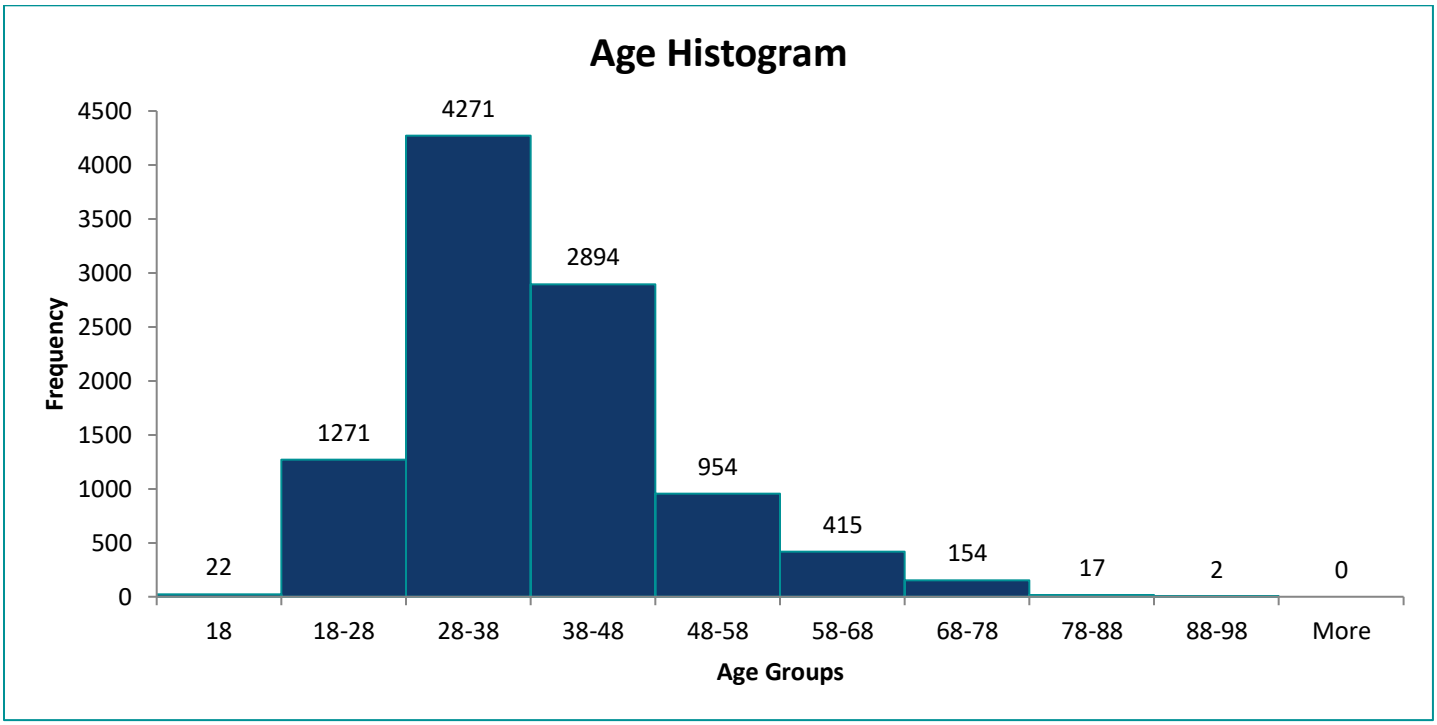

Chart 23 – Age Histogram

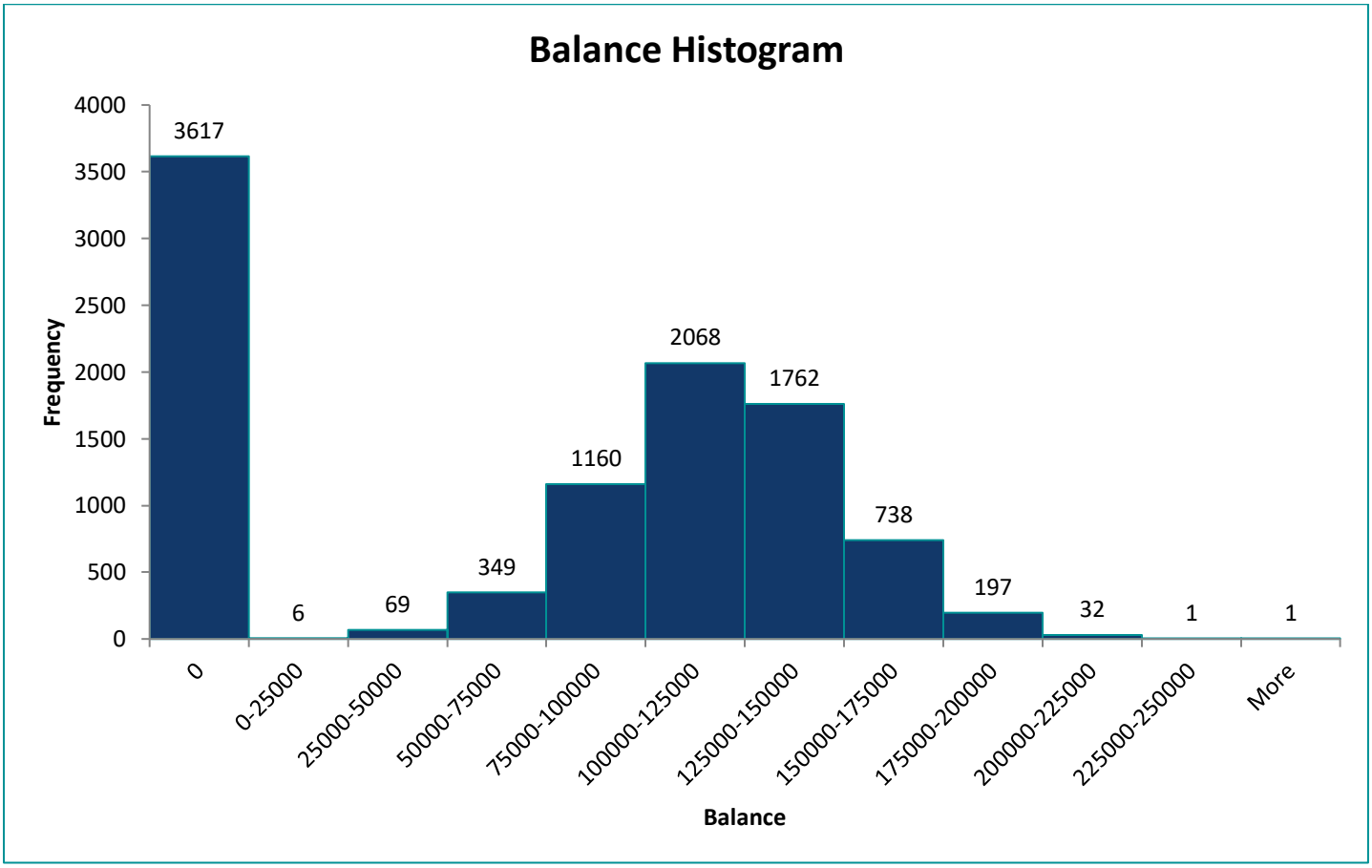

Chart 24 – Balance Histogram

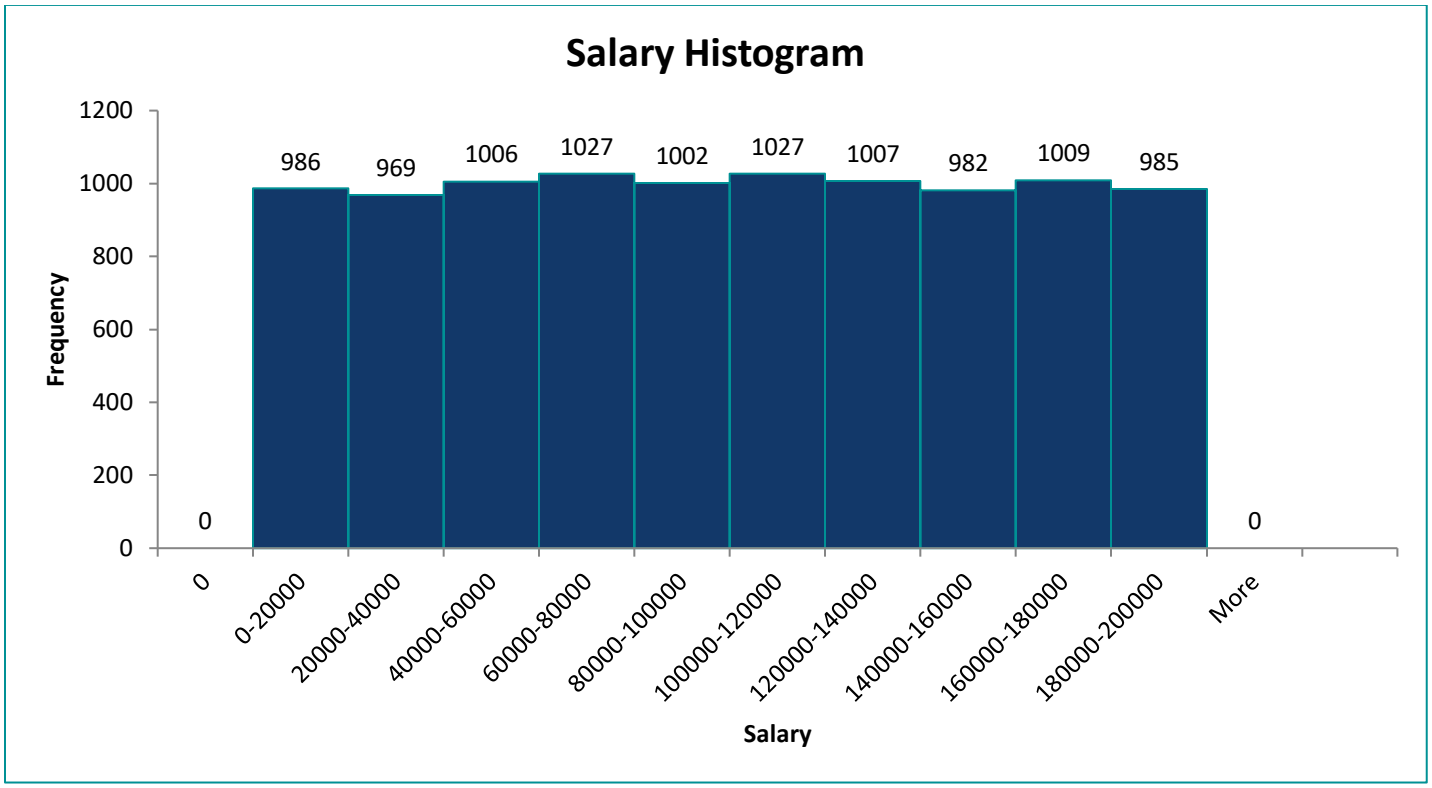

Chart 25 – Salary Histogram

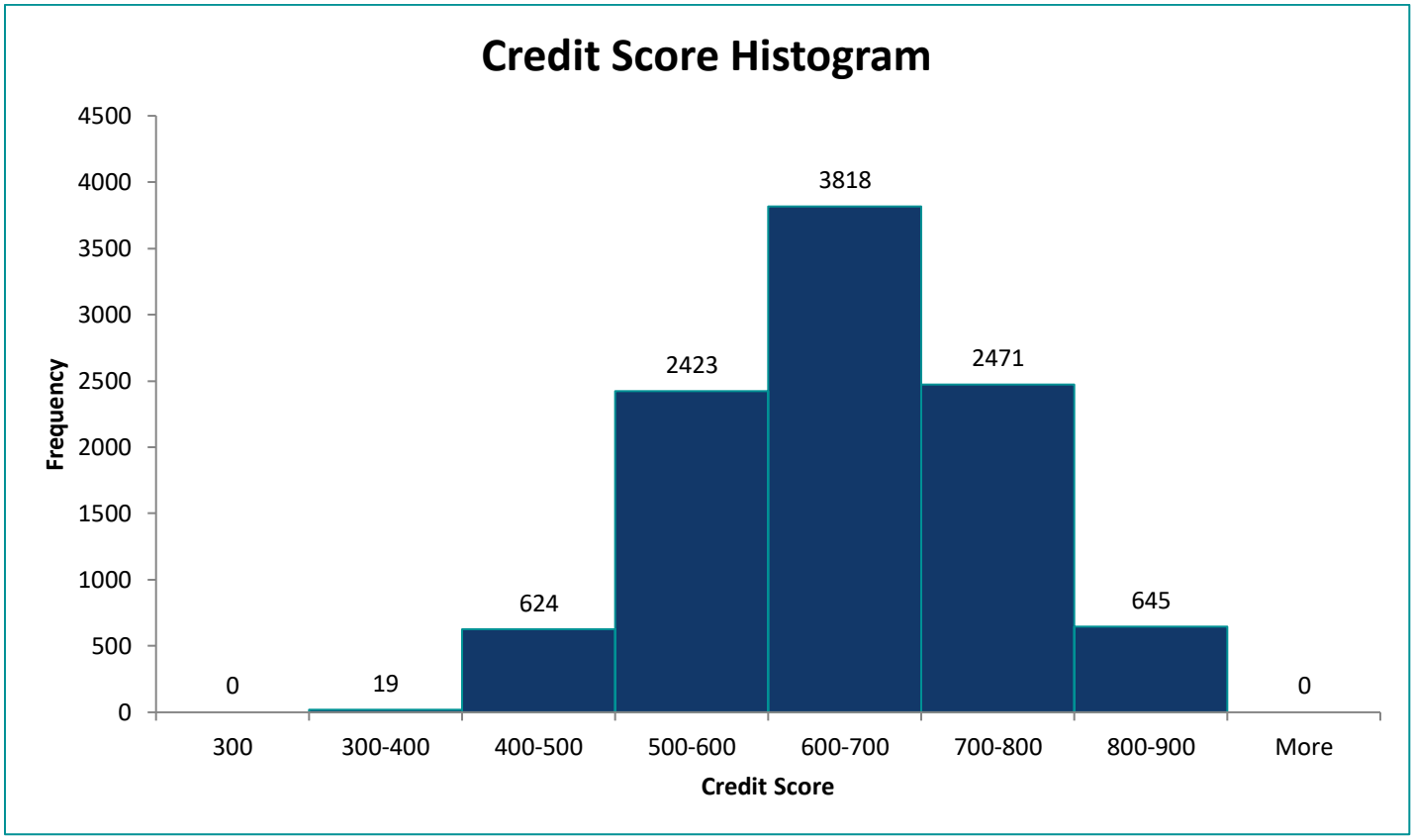

Chart 26 – Credit Score Histogram

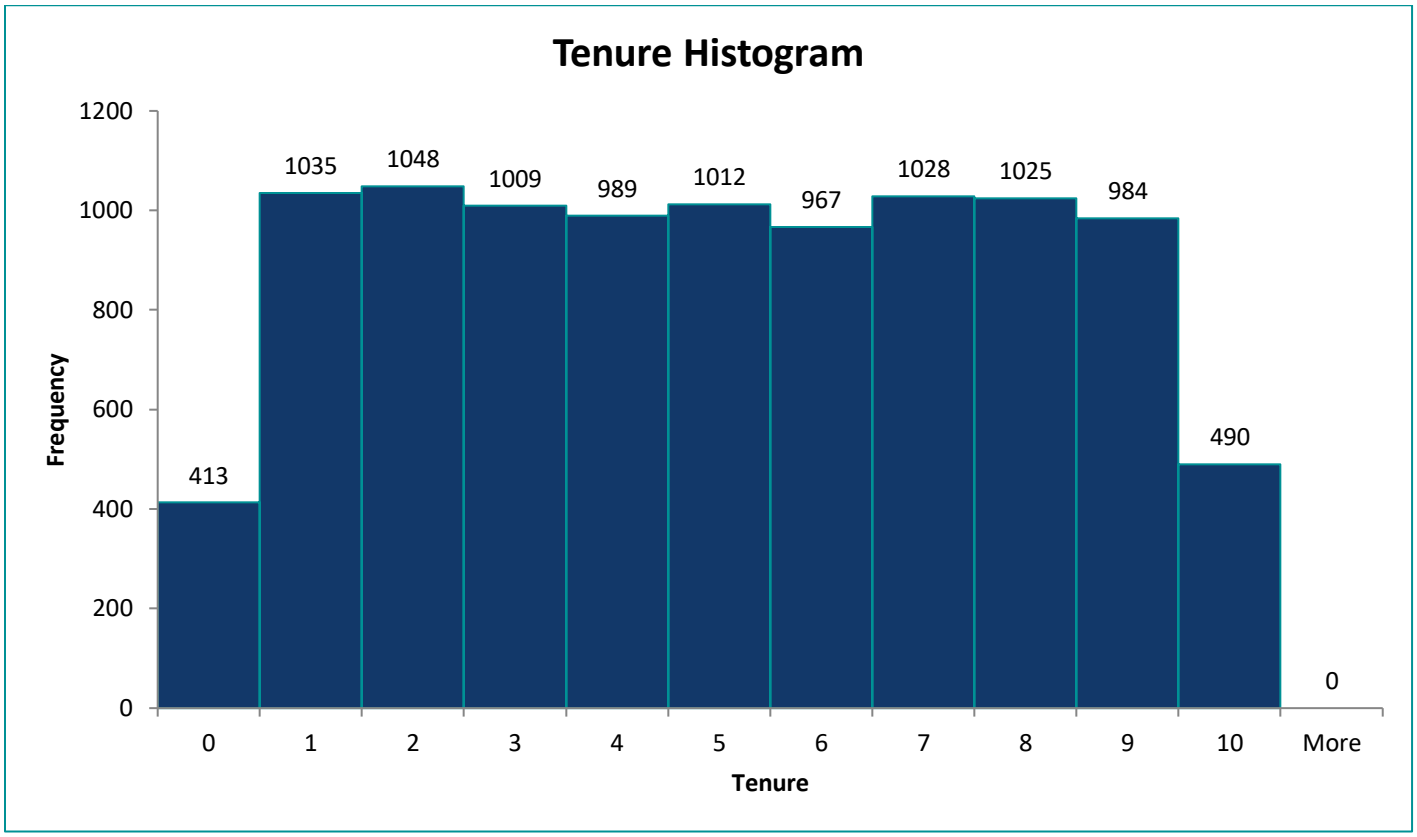

Chart 27 – Tenure Histogram

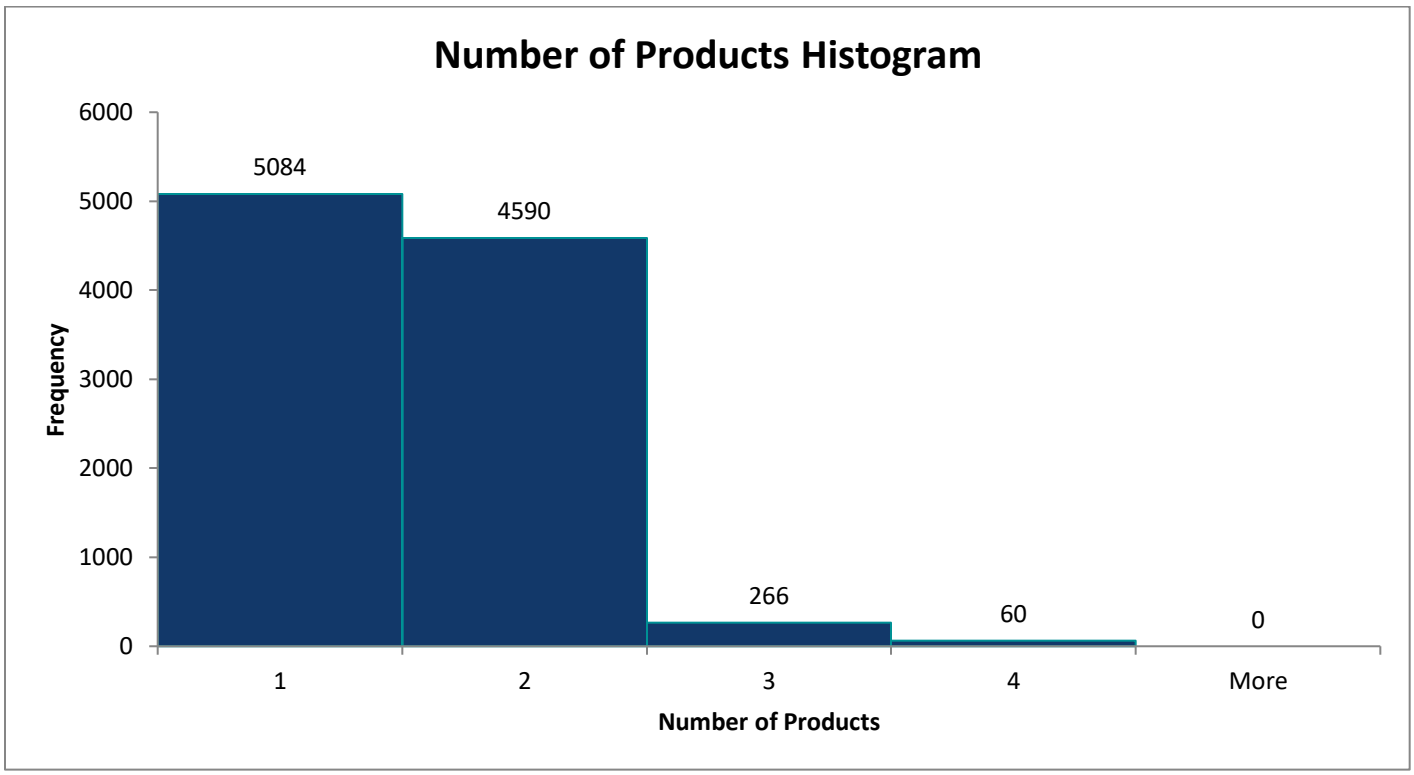

Chart 28 – Number of Products Histogram

Observations

- Above histograms represents the distribution of variables balance, estimated salary, credit score, tenure, number of products and age.
- **Age histogram is positively-skewed** and more frequencies are concentrated towards lower age group.
- **Majority of customers are below 48 years of age** with **mean age of 38 years**.
- **Balance histogram is symmetric** distribution, if we **ignore 0 balance**.
- Majority of customers are having balance between **75,000 to 150,000 with mean balance of 76,845.88**
- **Salary histogram is uniformly distributed** with each salary range have almost similar values.
- **Credit score histogram is negatively-skewed** and more frequencies are concentrated towards higher credit card scores.
- Majority of customers are having credit score between **600 to 700 with mean balance of 650**.
- **Tenure histogram is uniformly distributed** with each tenure have almost similar values.
- **Number of products histogram is positively-skewed** and more frequencies are concentrated towards lower number of products.
- Majority of customers are having 1 or 2 products with bank.

4.6. Scatter Plot

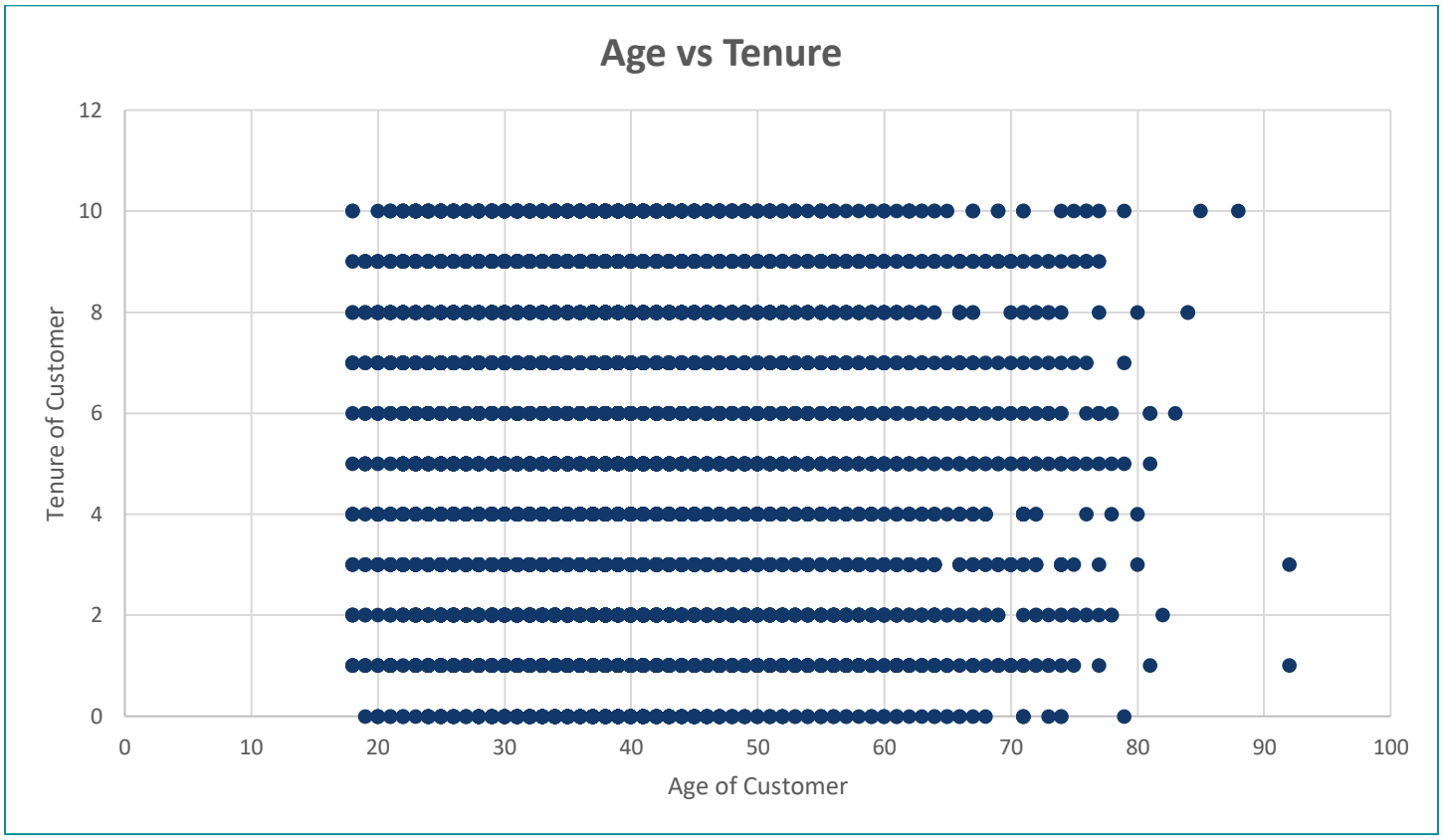

Chart 29 – Age vs Tenure

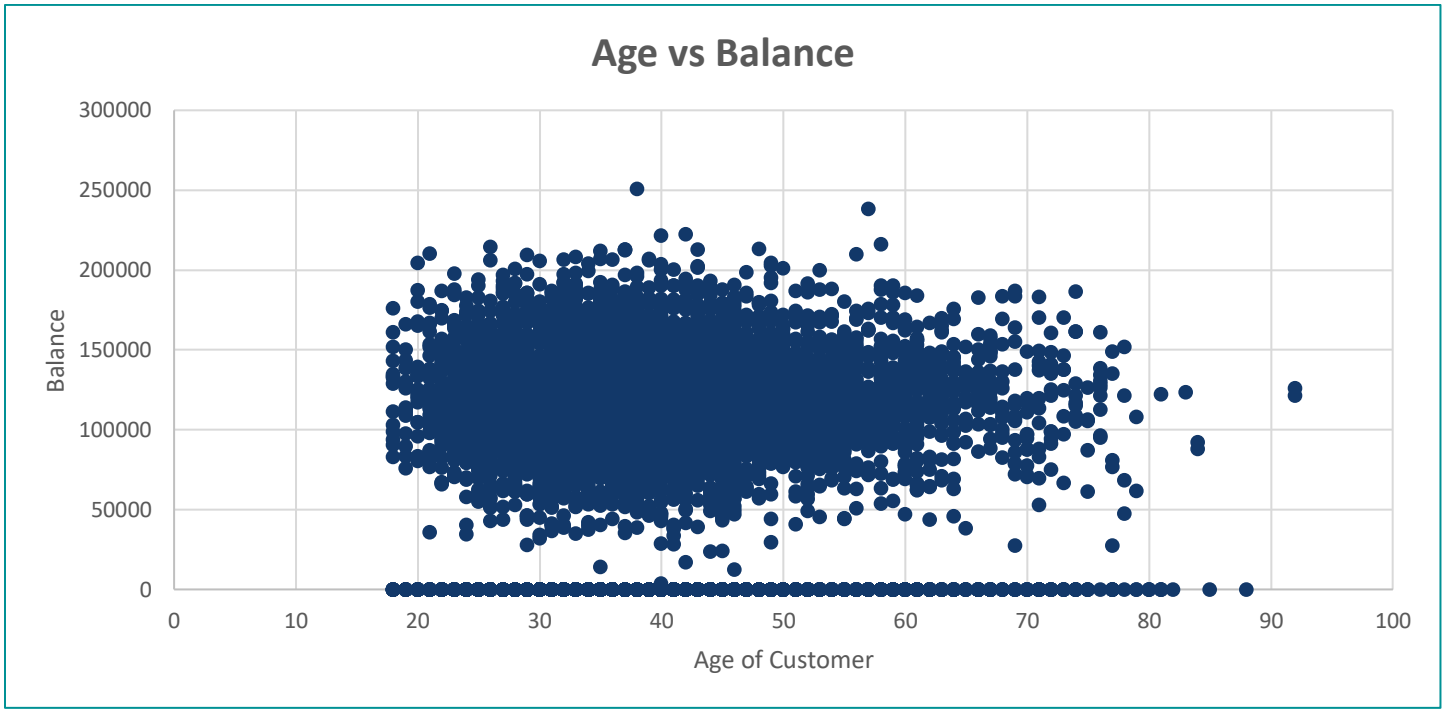

Chart 30 – Age vs Balance

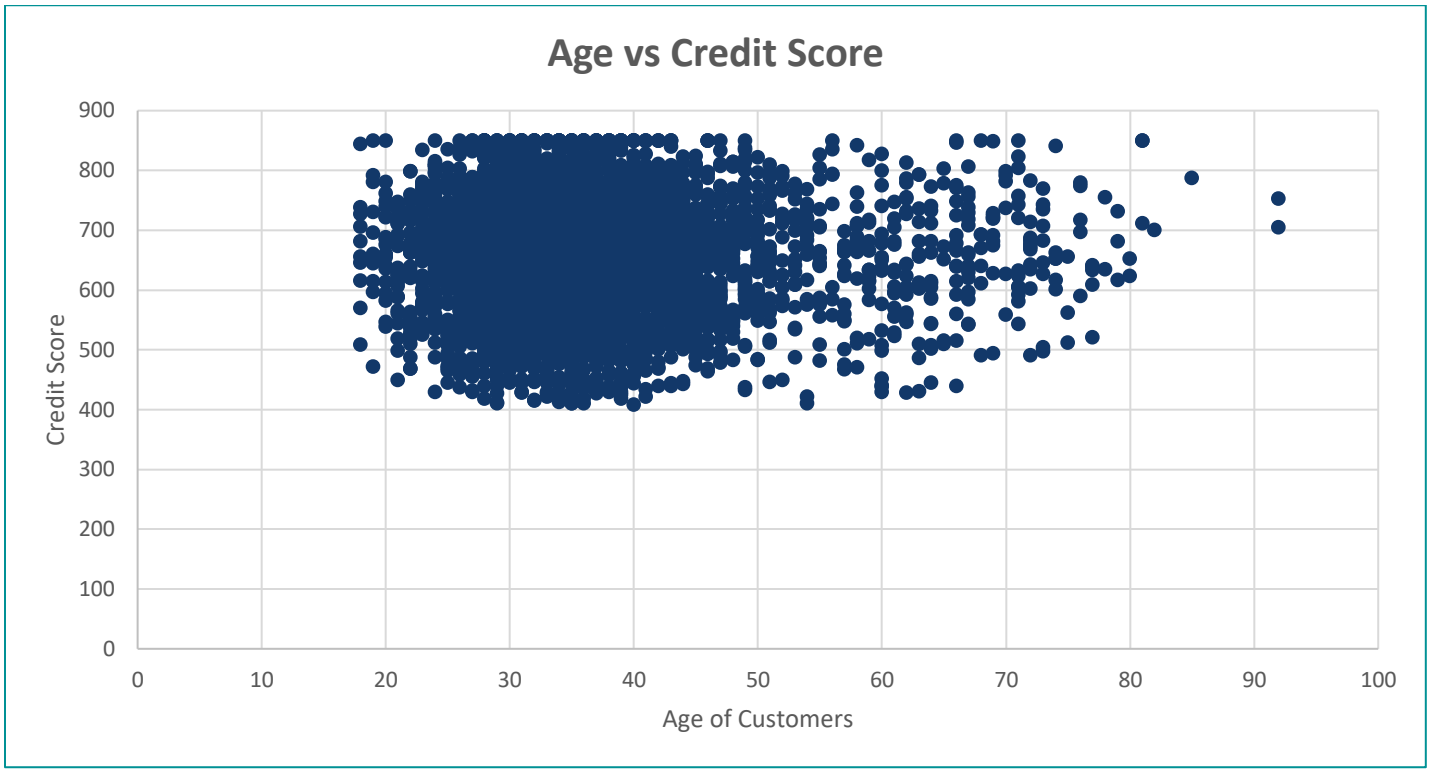

Chart 31 – Age vs Credit Score

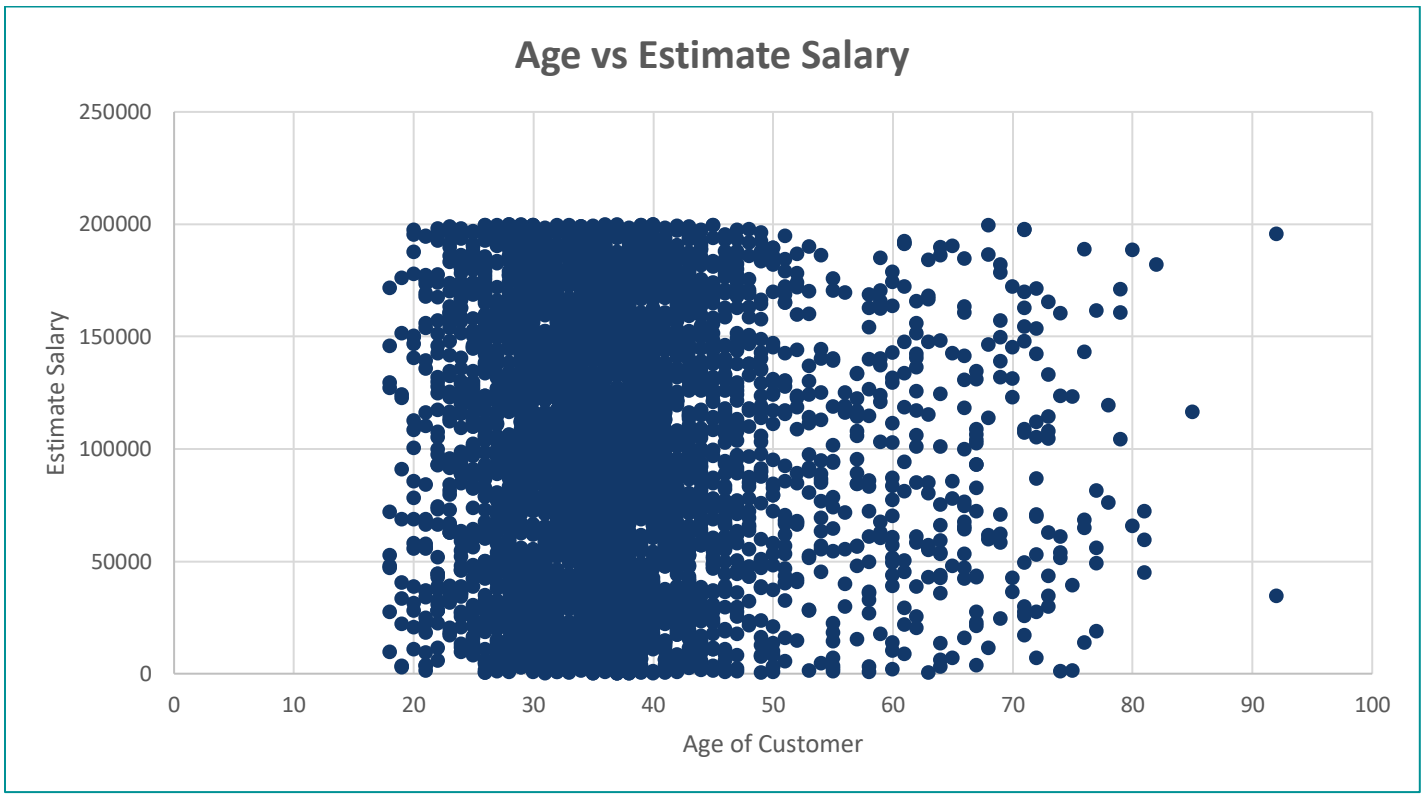

Chart 32 – Age vs Estimate Salary

Correlation Matrix

	Age	Credit Score	Tenure	Balance	Estimated Salary
Age	1				
Credit Score	-0.0039649	1			
Tenure	-0.0099968	0.00084194	1		
Balance	0.02830837	0.00626838	-0.0122539	1	
Estimated Salary	-0.007201	-0.0013843	0.00778383	0.0127975	1

Table 3 – Correlation Matrix

Observations

- The correlation coefficient value 'r' is close to zero and from the scatter plot, we can see its randomly distributed. **This indicates uncorrelated relationships between the variables**.

4.7. Numerical Summary of Variables

Parameters / Variables	Credit Score	Age	Tenure	Balance	Number Of Products	Estimated Salary
Mean	650.5288	38.9218	5.0128	76485.88929	1.5302	100090.2399
Standard Error	0.966532987	0.104878065	0.028921744	623.974052	0.005816544	575.1049282
Median	652	37	5	97198.54	1	100193.915
Mode	850	37	2	0	1	24924.92
Standard Deviation	96.65329874	10.48780645	2.892174377	62397.4052	0.581654358	57510.49282
Sample Variance	9341.860157	109.9940842	8.364672627	3893436176	0.338321792	3307456784
Kurtosis	-0.425725685	1.395347062	-1.165225227	-1.489411768	0.582980763	-1.181518447
Skewness	-0.071606608	1.011320263	0.010991458	-0.141108711	0.745567888	0.002085358
Range	500	74	10	250898.09	3	199980.9
Minimum	350	18	0	0	1	11.58
Maximum	850	92	10	250898.09	4	199992.48
Sum	6505288	389218	50128	764858892.9	15302	1000902399
Count	10000	10000	10000	10000	10000	10000

Table 4 – Summary Descriptive of Variables

Observations

Measure of Central Tendency

- **The average age of the customer is 39**. Half of the customers are below or equal to age 37 and remaining half are greater than 37. **Most customers are between age 28-38.**
- **The average credit score of the customers is 650.** Half of the customers are having credit score below or equal to 652 and remaining half are greater than 652.
- **The average tenure for the customers is 5 years.** Half of the customers are having tenure less than or equal to 5 years and remaining half are greater than 5 years. **Most customers are having tenure of 2 years**.
- **The average balance of the customer is 76485.88929.** Half of the customers are having balance below or equal to 97198.54 and remaining half have balance greater than 97198.54.
- **Most of the customers are having 1 product with the bank.**
- **The average estimate salary of the customer is 100090.2399.** Half of the customers are having estimate salary below or equal to 100193.915 and remaining half have balance greater than 100193.915.

Measure of Dispersion

- Approximate average distance/deviation of “Age” are “Credit Score” from the average are not too high and thus shows less variation in data.
- Approximate average distance/deviation of “Balance” are “Estimate Salary” from the average are 62397.4052 and 57510.49282 respectively. **This deviation is too high, means there is a high variation in data with respect to “Balance” and “Estimate Salary”** and thus are not good variables to analyze on.

Measure of Position

- P (25) of age is 32 years, meaning **25% of customers are less than or equal to age 32 years**.
- P (75) is 44, 25% of all the customers are greater than age 44 which means that **most of the customers are young and older customers are not part of bank anymore**.
- P (50) is 37, at least **half of the customers are between age 32 and 44 years of age**.

4.8. Box and Whisker Plot

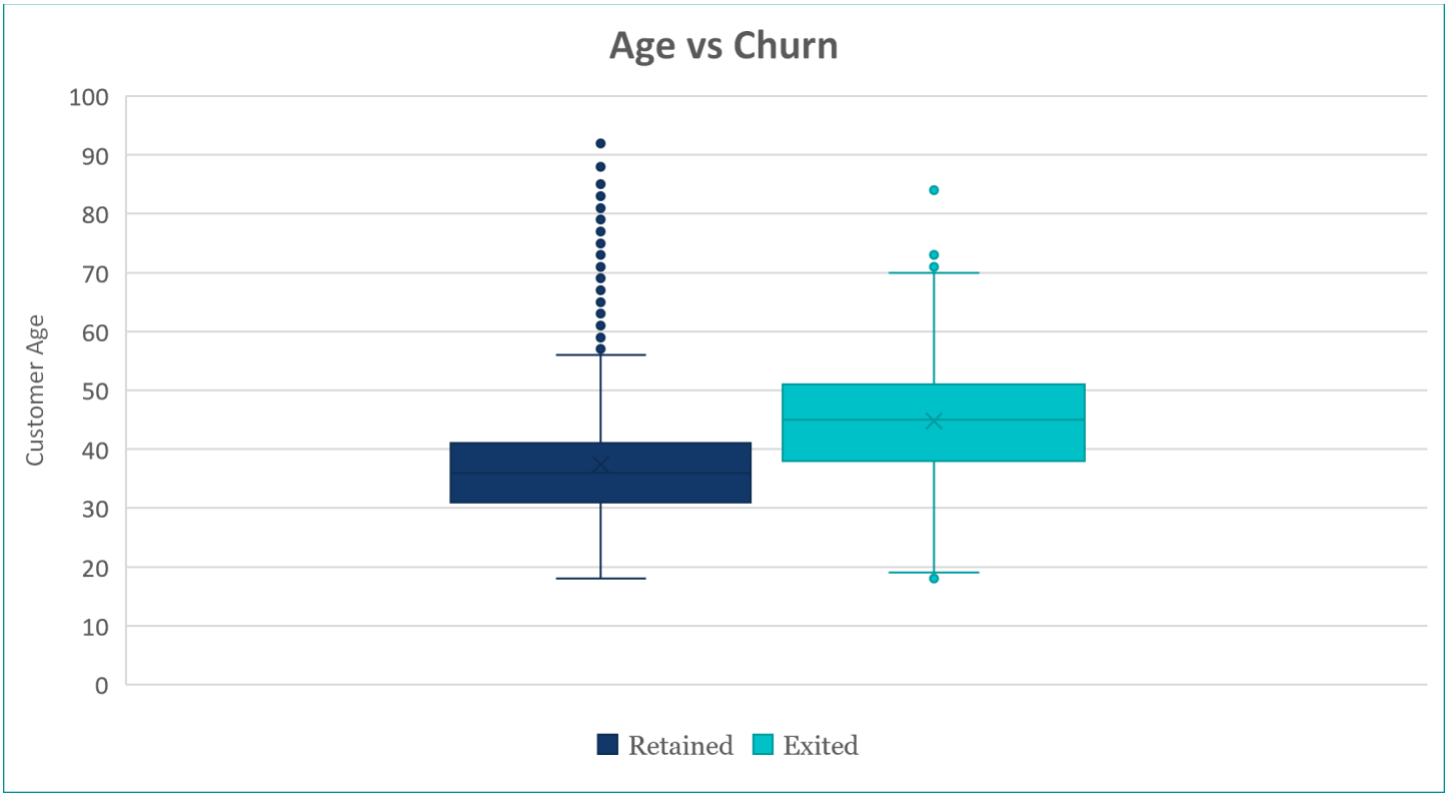

Chart 33 – Age vs Churn

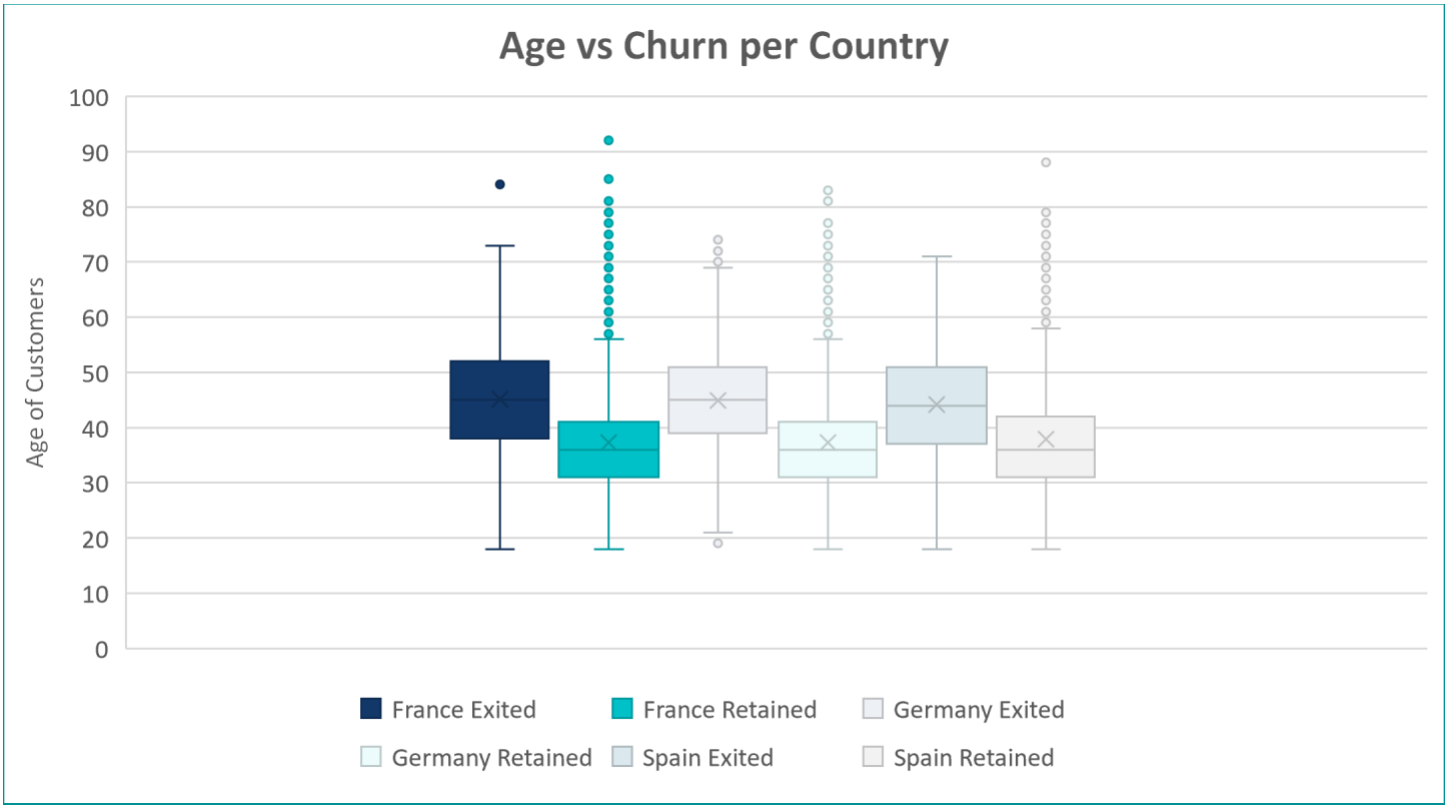

Chart 34 – Age vs Churn per Country

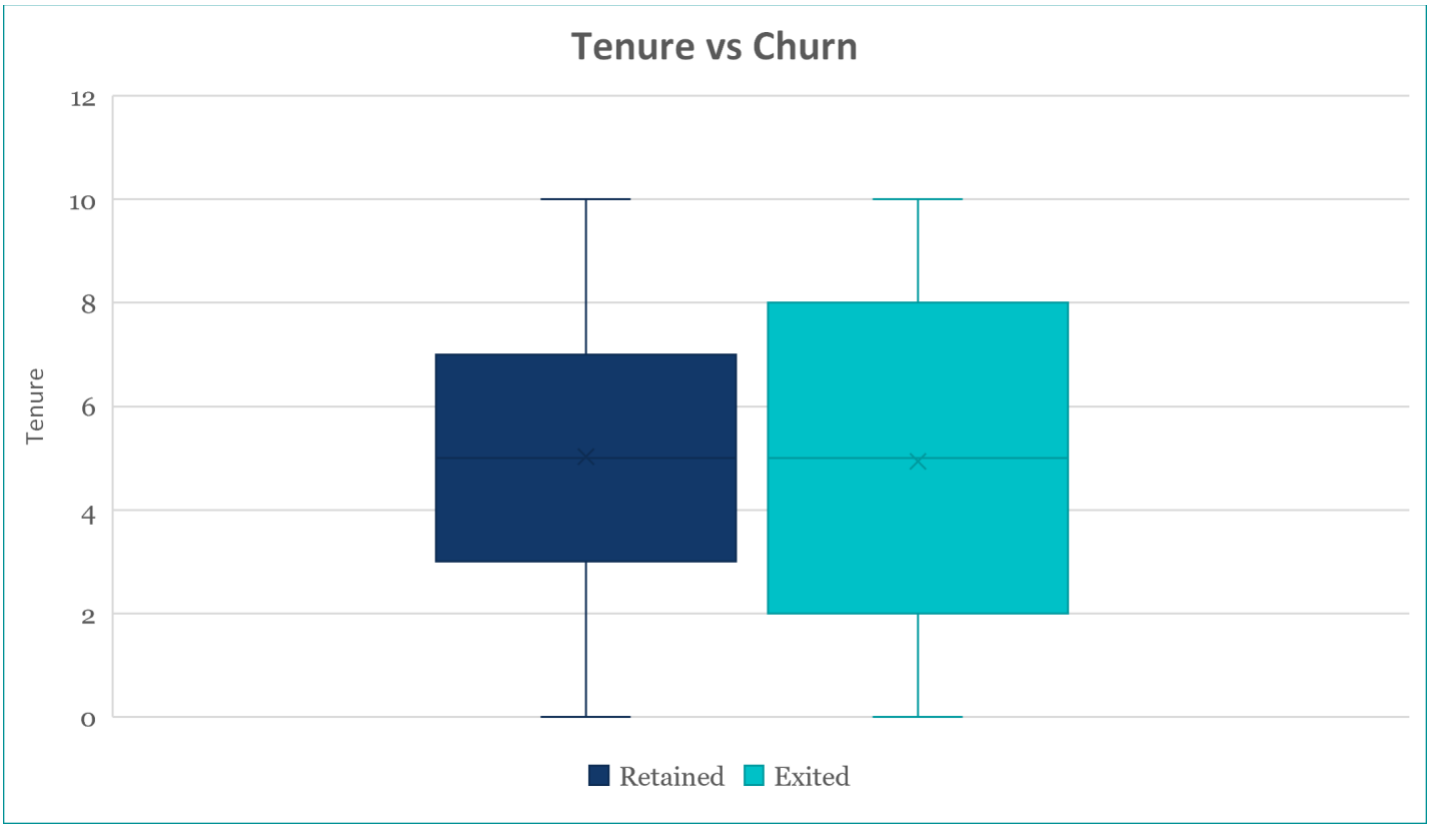

Chart 35 – Tenure vs Churn

Observations

- Customers with higher age have exited more as compared to customers with lower age.
- Inter quartile range of exited customers is 38 to 51 Years & for retained customers is 31 to 41 years.
- **Majority of the outliers lies within retained customers**.
- **Age behavior pattern is same across countries with respect to outliers**.
- Exited customers have broader inter quartile range for tenure as compared to retained customers.
- Exited customers inter quartile range can be derived by adding 0.5 year at the start and end of the inter quartile range of retained customers which could imply that **customers between tenure 2.5 to 3 years and between 7 to 7.5 years might be at verge of exit from bank.**

5. Probability

In this section we will try to again highlight some of the key findings from the descriptive analysis using conditional probability and through modelling variable with probability distribution

5.1. Operations with Probability

Since Age, Country, Gender and Credit score seems to be important variable to analyze the churn rate, we will use probability distributions on these variables to deduce more

Age Groups	Retained	Exited	Grand Total
18-28	947	73	1020
28-38	3678	389	4067
38-48	2417	786	3203
48-58	472	583	1055
58-68	280	183	463
68-78	146	22	168
78-88	20	1	21
88-98	3	0	3
Grand Total	**7963**	**2037**	**10000**

Table 5 – Age Group Distribution

- Probability of Customer exited
  - **P (Exited)** = 2037/10000 = **20.37%**
- Probability of customers retained
  - **P (Retained)** = 1 – P (Exited) = 1-.2037 = .7963 = **79.63%**
- Probability of customer being retained AND the customer is in age >= 58
  - **P (Retained/>=58)** = 449/10000 = **4.49%**
- Probability of customer being retained OR the customer is in age >= 58
  - **P (Retained\>=58)** = 655/10000 + 7963/10000 - 449/10000 = **81.69%**
- Probability of customer is in age group 28-38 OR customer is in age group 38-48
  - **P (28-38 \ 38-48)** = 4067/10000 + 3203/10000 = **70.9%**
- Probability of customer exited given that customer age >= 48
  - **P (Exited | >=48)** = (789/10000) / (1710/10000) = **46.14%**

Number of Products	Retained	Exited	Grand Total
One Product	3675	1409	5084
More than One Product	4288	628	4916
Grand Total	7963	2037	10000

Table 6 – Number of Products Distribution

- Probability (Exited | One Product) = 1409/5084 = 69%

Observations

- There is **69% chance that a customer will exit the bank given the customer has only one product** which implies **customers which have more than one product has less chance of exiting the bank**.
- Chances of customers **exiting the bank given that age is greater than equal to 48 is 46.14%.**
- Chances of customers **retained** with bank and age is greater than equal to 58 is only 4.49%.

5.2. Probability Distribution

The nature of data we have and the variables which are affecting the churn rate, we will not be able to apply Binomial, Poisson or Exponential distributions.

Age, Country, Gender and Credit score seems to be important variable to analyze the churn rate. Out of these variables, credit score follows the normal distribution and thus probability distribution can be applied

Customers with low credit scores churn out from the bank

- Based on the distribution analysis we know credit score follows normal distribution with mean 650.52 and standard deviation of 96.65. We will find the probability that any randomly selected customer will have credit score less than or equal to 500
- Probability (Credit Score <= 500) = NORM.DIST (500,650.52, 96.65, TRUE) = 6%

Maximum proportion of the customers have credit score between 600 & 700

- Based on the distribution analysis we know credit score follows normal distribution with mean 650.52 and standard deviation of 96.65. We will find the probability that any randomly selected customer will have credit score between 600 & 700
- Probability (600 <= Credit Score <= 700) = NORM.DIST (600,650.52, 96.65, TRUE) - NORM.DIST (700,650.52, 96.65, TRUE) = 40%

Customer with score greater than 700 are valuable customers and there churning will impact the bank

- Based on the distribution analysis we know credit score follows normal distribution with mean 650.52 and standard deviation of 96.65. We will find the probability that any randomly selected customer will have credit score greater than 700.
- Probability (Credit Score > 700) = 1 - NORM.DIST (700,650.52, 96.65, TRUE) = 30%

Observations

- Though chances are less customer having low credit score but bank must be watchful of the low credit score customers and their transactions to avoid any loss to bank.
- There are **40% chances that a randomly selected customer will have credit score between 600 & 700** which makes these customers to be monitored but not very closely.
- There is **30% probability that customer will have credit score higher than 700** which implies that bank need to focus on these 30% customers more as compared to other customers.

6. Inferential Statistics

6.1. Population and Parameter

Population

- P1: All customers of the bank
- P2: Collection of credit score of all the customers of the bank.
- P3: Collection of age of all the customers of the bank.

Parameter

- P1: u (mean) is the parameter – Average credit score of all the customers of the bank.
- P2: u (mean) is the parameter – Average age of all the customers of the bank.

6.2. Sample and Statistics

Sample

- P1: Credit score of 10,000 customers of the bank.
- P2: Age of 10,000 customers of the bank.

Statistics

- For P1:
    - First estimator (X) is the average credit score of 10,000 customers. Value of the estimator is 650.52.
    - Second estimator (S) is the standard deviation of credit score for the 10,000 customers. Value of the estimator is 96.65
- For P2:
    - First estimator (X) is the average age of 10,000 customers. Value of the estimator is 38.92.
    - Second estimator (S) is the standard deviation of age for the 10,000 customers. Value of the estimator is 10.48

6.3. Sampling Distribution of Statistic

- For P1:
    - The credit score has a normal distribution
    - The mean of the distribution is 650.52. The SD of the distribution is 96.65
    - The sample size is 10,000
    - Standard error is 0.9665
    - Note as sample size increases SD of sample decreases.
- For P1:
    - The age is not normal distribution
    - The mean of the distribution is 38.92. The SD of the distribution is 10.48
    - The sample size is large.
    - Standard error is 0.1048

6.4. 95% Confidence Interval of Parameter

- Interval for average credit score is (650.52 – _1.96 * 96.65/sqrt (10000), 650.52 + 1.96 * 96.65/sqrt (10000)) i.e. (**648.65,652.41)**.
- Interval for age is (38.92 – _1.96 * 10.48/sqrt (10000), 38.92 + 1.96 * 10.48/sqrt (10000)) i.e. **(38.71,39.12)**.

6.5. Hypothesis Testing

One Tail Test

Average credit score of the customers is ~650. We will test whether average credit score is significantly more than 650. Assume that credit score is normal

- H0 - Hypothesis will be that mean equal to 650.
- H1 - Alternate Hypothesis will be that mean greater than 650

t-Test: Two-Sample Assuming Unequal Variances		
	Credit Score	***Credit Score Default***
Mean	650.5288	0
Variance	9341.860157	0
Observations	10000	10000
Hypothesized Mean Difference	650	
df	9999	
t Stat	0.547110142	
P(T<=t) one-tail	0.292157637	
t Critical one-tail	1.645006033	
P(T<=t) two-tail	0.584315275	
t Critical two-tail	1.960201264	

Table 7 – Hypothesis Test for Credit Score

Observations

- Since **t-value =0.547 < 1.645**, we **fail to reject H0 at 5% level of significance** which means the **average credit score per customer at the bank is 650**.

Two Tail Test

Average age of the churned customers is greater than average age of the not churned customers. We will test if the average age of churned and not churned customers are same

- Hypothesis will be that mean age of churned customers is equal to mean age of non-churned customers.
- Alternate hypothesis will be that mean age of churned customers is not equal to mean age of non-churned customers.

t-Test: Two-Sample Assuming Unequal Variances		
	Age Churned	***Age Not Churned***
Mean	37.4083888	44.83799705
Variance	102.5229741	95.288084
Observations	7963	2037
Hypothesized Mean Difference	0	
df	3248	
t Stat	-30.4191972	
P(T<=t) one-tail	2.3582E-179	
t Critical one-tail	1.645322902	
P(T<=t) two-tail	4.7163E-179	
t Critical two-tail	1.960694631	

Table 8 - Hypothesis Test for Age

Observations

- Since **t-value =|-30.41| > 1.96**, we **reject H0 at 5% level of significance** which means the **mean age of churned customers is not equal to mean age of non-churned customers.**

7. Multiple Linear Regression

The bank wants to know if they can model credit score of the customers based on input parameters.

Following variables are used for the model

Variable	Predictor/Target
Credit Score	Target
Geography	Predictor
Gender	Predictor
Age	Predictor
Tenure	Predictor
Has Credit Card	Predictor
Estimate Salary	Predictor
Balance	Predictor
Number of Products	Predictor
Is Active Member	Predictor

Table 9 – Multiple Linear Regression Variables

7.1. Regression Coefficients

Regression Statistics	
Multiple R	0.031997539
R Square	0.001023843
Adjusted R Square	0.000123864
Standard Error	96.64731262
Observations	10000

Table 10 – Regression Coefficients

Observations

- R & Adjusted R square are both very small close top 0 which implies that **predictors are not able to explain the variability in the credit score.**

7.2. Model Assumption Validation

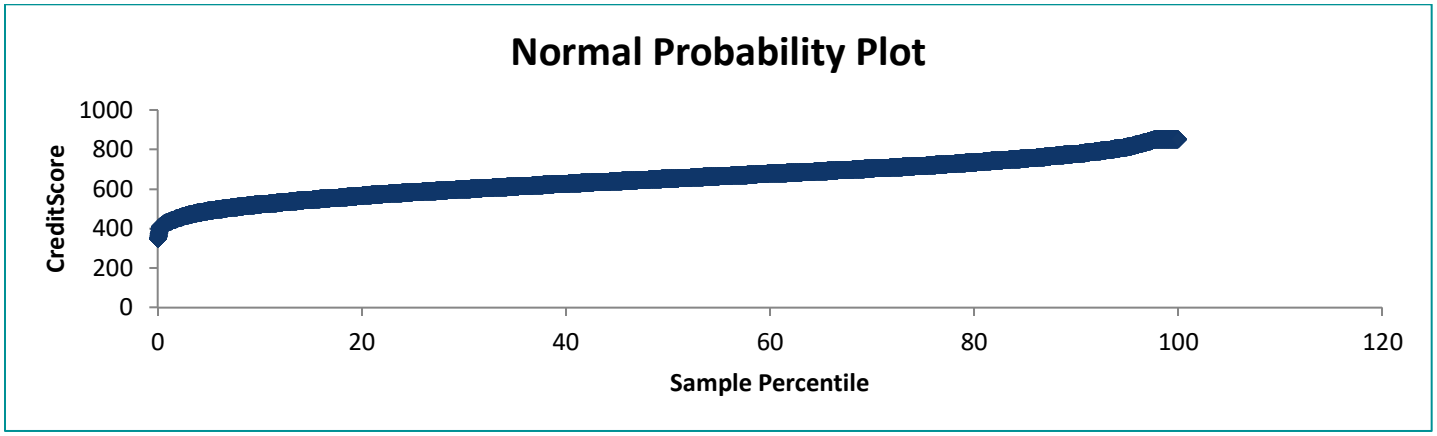

Chart 36 – Normal Probability Plot

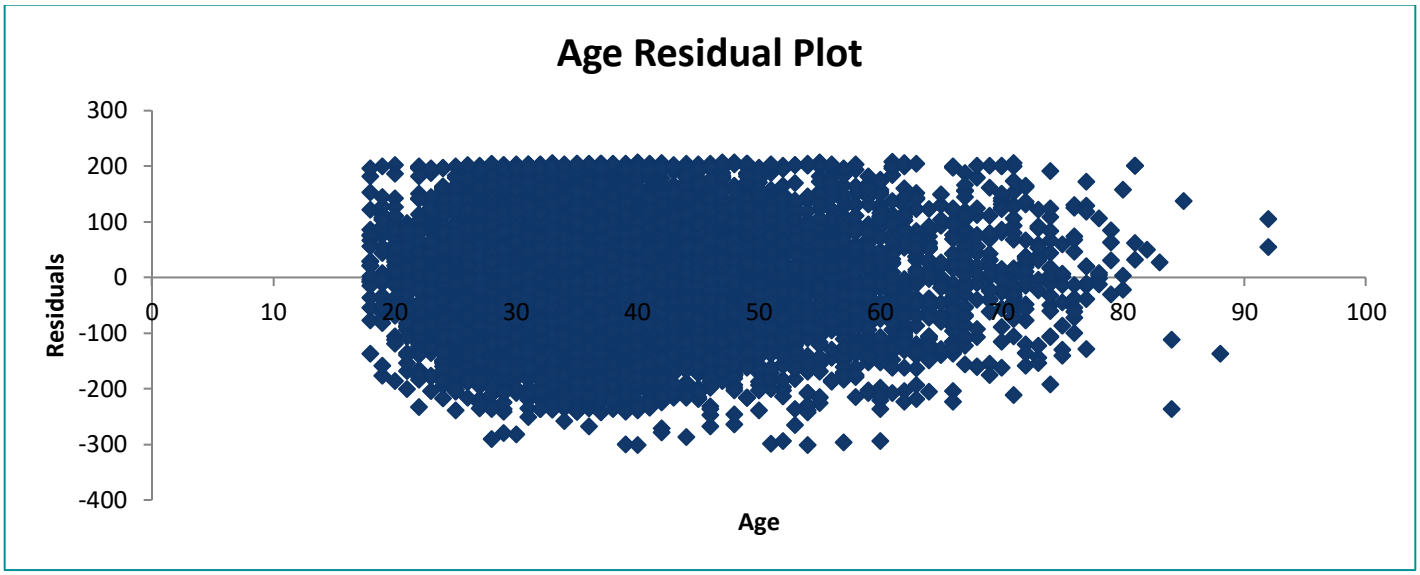

Chart 37 – Age Residual Plot

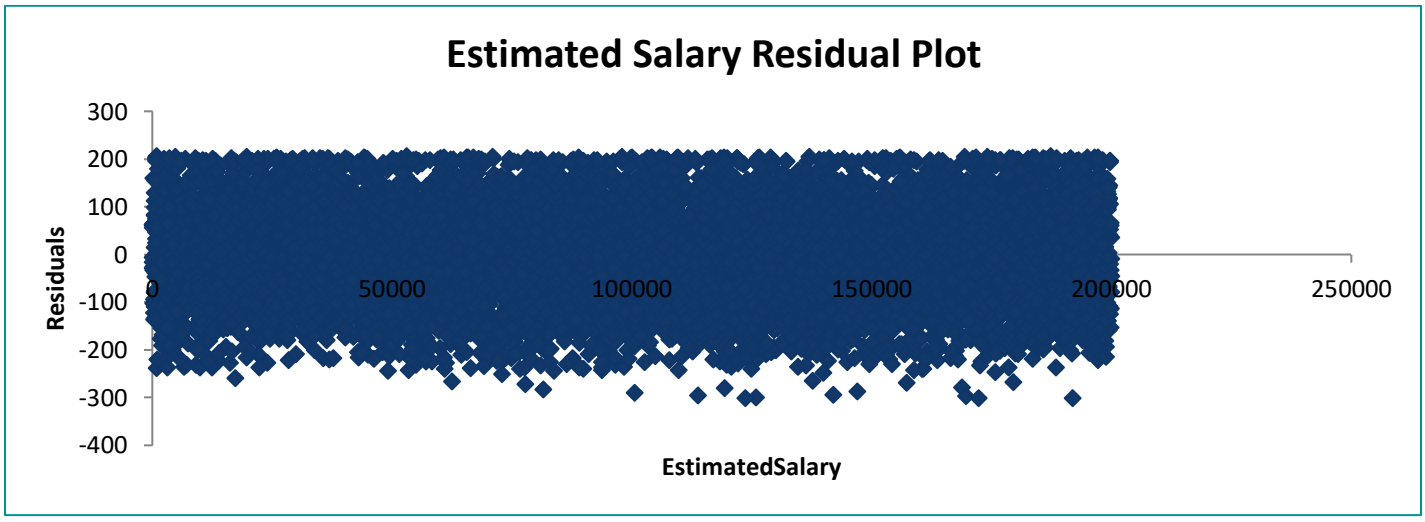

Chart 38 – Estimated Salary Residual Plot

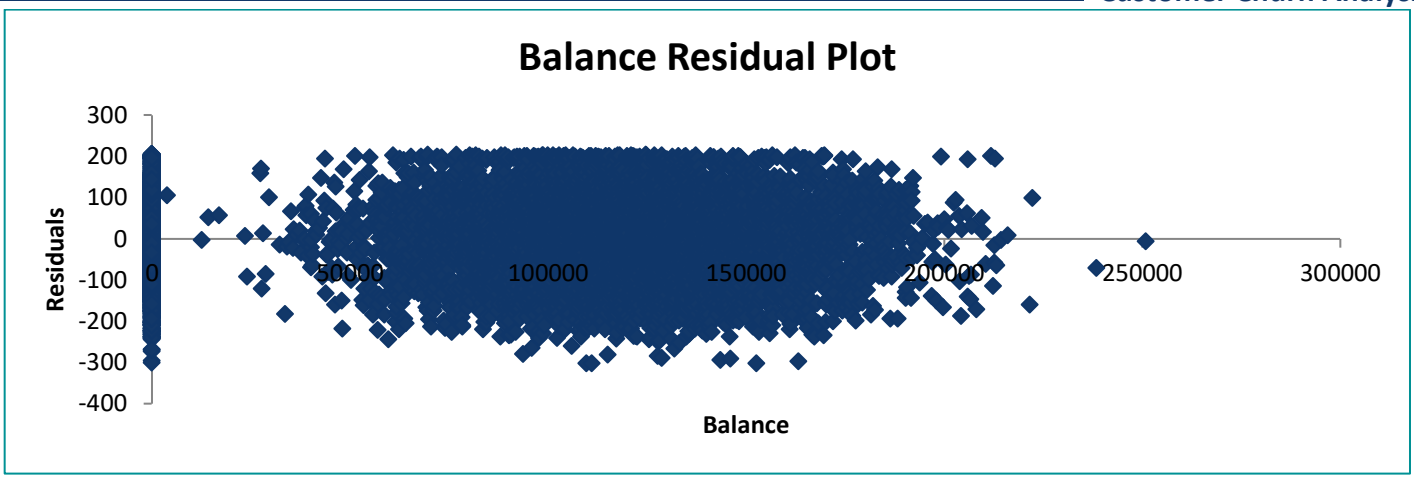

Chart 39 – Balance Residual Plot

Observations

- **Residual are homoscedastic** in nature for the variables.

7.3. Variables Summary Outcomes

Variable	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	644.1013566	6.479646091	99.40378649	0	631.3999447	656.8027684
Geography	0.674274107	1.251301851	0.538858076	0.589996795	-1.778529629	3.127077844
Gender	0.645877684	1.943652429	0.33230102	0.739668942	-3.164072678	4.455828046
Tenure	0.053776232	0.334507161	0.160762572	0.872283655	-0.6019252	0.709477663
Balance	1.41308E-05	1.74659E-05	0.80904912	0.418506161	-2.01059E-05	4.83674E-05
Number Of Products	2.416594833	1.756801006	1.375565488	0.168987039	-1.027089095	5.860278761
Has Credit Card	-1.100329675	2.121577981	-0.518637394	0.604025115	-5.259049968	3.058390617
Is Active Member	5.080070284	1.942958761	2.614605305	0.008946474	1.27147965	8.888660919
Estimated Salary	-2.66857E-06	1.68137E-05	-0.158714143	0.873897303	-3.56268E-05	3.02896E-05
Age	-0.059429672	0.092703688	-0.641071276	0.521491081	-0.241147578	0.122288235

Table 11 – Multiple Linear Regression Outcome

Observations

- **None of the factors are statistically significant individual predictor of Y as all p-values are more than 0.05.**
- We can observe that maximum prediction power is coming from intercept with coefficient 649 which means that if we even keep all variables 0 then also customer will have credit score equal to 649 which is very close to mean credit score.

8. ANOVA

ANOVA					
	df	***SS***	***MS***	***F***	***Significance F***
Regression	9	95636.37134	10626.26348	1.137629945	0.331640897
Residual	9990	93313623.33	9340.703036		
Total	9999	93409259.71			

Table 12 – ANOVA

Observations

- Since probability is 0.3316 which is greater than significance level of 0.05 it can be confirmed that model is not statistically significant in predicting credit score.

9. Conclusion from Analysis

ABC Bank CEO concludes that:

- Age, Credit Score, Geography, Gender and Number of Products are important factors for churn rates.
- Banks should develop specific programs and strategies to target customers who can still be saved. Those who are at old age, or those with poor credit scores.
- The age groups that are most likely to leave the bank are 38 to 78 years old. It looks that trying to retain the 38–48 and 48–58 age group has a higher return on value.
- Bank need to try to find more predictors to create and identify the model which will end up having more accurate model for analysis.
- Customers in Germany have the highest churn rate among three geographies. Spain and France female are more likely to churn than male customers. French and Spain banks need to allocate more resources for female customers and resolve their problems. Also, Germany banks needs to put more attention as why customer churn rate is so high.
- The most important factor for the bank is engagement and to improve customer service because we can see that as soon as age is increasing where probably customer earning is increasing and customer looking for more products other than account and credit cards (like retirements etc.), customers are leaving for other banks.

10. Contribution of Team Members

Section	Sub Section	Name and Roll Numbers
Introduction	Motivation	Group 5 - All
	Objective	Group 5 - All
	Outcomes	Group 5 - All
Dataset	Data Source	Group 5 - All
	Data Sample	Group 5 - All
	Dataset	Group 5 - All
Variables	Variable Definitions	Group 5 - All
Descriptive Analysis	Pie Charts	AKSHAT DAS (EPGP-13C-006), ANSHUMAN SINGH (EPGP-13C-014)
	Bar Graphs	AYUSH TYAGI (EPGP-13C-022), DIVYANSHU KATIYAR (EPGP-13C-030)
	Line Charts	JAYSHREE SOLANKI (EPGP-13C-038), MANDEEP SINGH (EPGP-13C-046)
	Side-by-Side Bar Graphs	NAGENDRA GANTI (EPGP-13C-054), PIYUSH PRIYADARSHI (EPGP-13C-062)
	Histograms	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079)
	Scatter Plot	ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
	Numerical Summary of Variables	SAURAV KUMAR (EPGP-13C-103), SMITA RANI TALUKDAR (EPGP-13C-111)
	Box and Whisker Plot	TUSHAR SINGHANIYA (EPGP-13C-119), AKASH SURESH (EPGP-13C-127)
Probability	Operations with Probability	AKSHAT DAS (EPGP-13C-006), ANSHUMAN SINGH (EPGP-13C-014), AYUSH TYAGI (EPGP-13C-022), DIVYANSHU KATIYAR (EPGP-13C-030)
	Probability with distributions	JAYSHREE SOLANKI (EPGP-13C-038), MANDEEP SINGH (EPGP-13C-046), NAGENDRA GANTI (EPGP-13C-054), PIYUSH PRIYADARSHI (EPGP-13C-062)
Inferential Statistics	populations and parameters	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079), ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
	Sample and Statistics	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079), ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
	Sampling distribution	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079), ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
	95% Confidence Interval of Parameter	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079), ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
	Hypothesis Testing - One Tail	SAURAV KUMAR (EPGP-13C-103), SMITA RANI TALUKDAR (EPGP-13C-111), TUSHAR SINGHANIYA (EPGP-13C-119), AKASH SURESH (EPGP-13C-127)
	Hypothesis Testing - Two Tail	SAURAV KUMAR (EPGP-13C-103), SMITA RANI TALUKDAR (EPGP-13C-111), TUSHAR SINGHANIYA (EPGP-13C-119), AKASH SURESH (EPGP-13C-127)
Multiple Linear Regressions	Regression Coefficients	AKSHAT DAS (EPGP-13C-006), ANSHUMAN SINGH (EPGP-13C-014), AYUSH TYAGI (EPGP-13C-022), DIVYANSHU KATIYAR (EPGP-13C-030)
	Model Assumption Validations	JAYSHREE SOLANKI (EPGP-13C-038), MANDEEP SINGH (EPGP-13C-046), NAGENDRA GANTI (EPGP-13C-054), PIYUSH PRIYADARSHI (EPGP-13C-062)
	Variable Summary Outcomes	RAJIV AMBASTHA (EPGP-13C-071), RAVI ROUSHAN KUMAR (EPGP-13C-079), ROHIT BAJPAI (EPGP-13C-087), SANDIP KUMAR (EPGP-13C-095)
ANOVA	ANOVA	SAURAV KUMAR (EPGP-13C-103), SMITA RANI TALUKDAR (EPGP-13C-111), TUSHAR SINGHANIYA (EPGP-13C-119), AKASH SURESH (EPGP-13C-127)
Conclusion	Conclusion	Group 5 - All
Review and Format	Review and Format	Group 5 - All