**Capstone Project:**

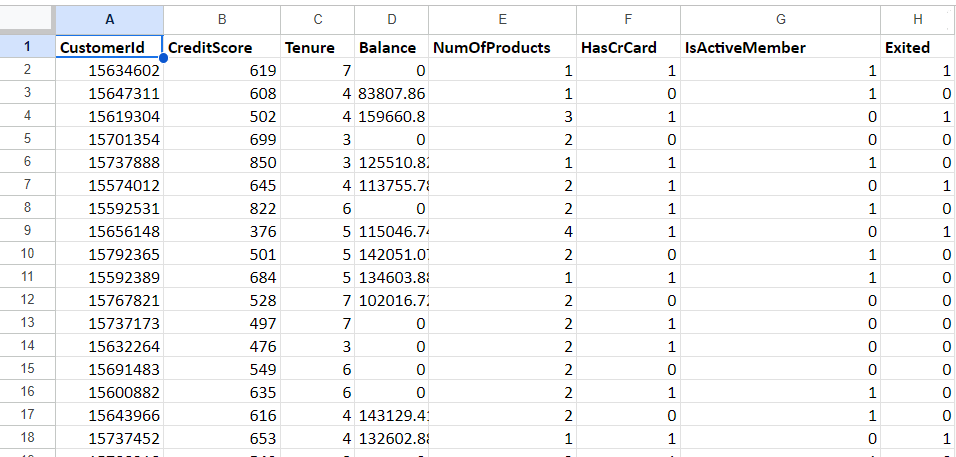
**Analytical CRM Development for a Bank**

**Newton School – By Arti Awasthi**

**Problem Statement**

Problem Statement

You are an analytical CRM (Customer Relationship Management) specialist hired by a bank to extract meaningful insights from various customer-related datasets. The bank aims to reduce customer churn, improve service delivery, and enhance customer satisfaction. They have provided you with datasets including customer demographics, transaction details, customer exit information, and active customer profiles.



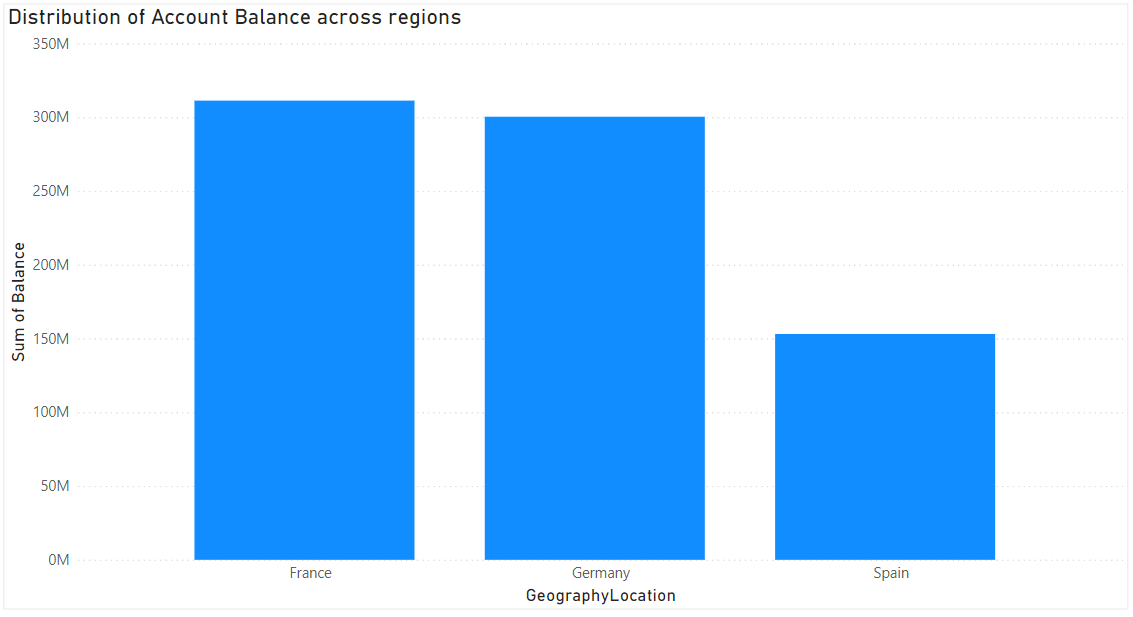
The image above displays details about Bank Customer data, including:

* **RowNumber:** The row number in the dataset, likely used for reference or indexing.
* **CustomerId:** A unique identifier for each customer.
* **CreditScore:** A numerical representation of the customer's creditworthiness.
  + Credit score:
    - Excellent: 800–850
    - Very Good: 740–799
    - Good: 670–739
    - Fair: 580–669
    - Poor: 300–579
* **GeographyID:** A numerical identifier that likely corresponds to a geographical location, such as a country or region**.**
* **GenderID:** A numerical identifier for the customer's gender, where for example, '1' could represent male and '2' could represent female.
* **Age:** The age of the customer.
* **Tenure:** The number of years the customer has been with the bank.
* **Balance:** Current balance in the customer's account.
* **NumOfProducts:** refers to the number of products that a customer has purchased through the bank.
* **HasCrCard:** denotes whether or not a customer has a credit card. This column is also relevant, since people with a credit card are less likely to leave the bank.
  + 1 represents credit card holder
  + 0 represents non credit card holder
* **IsActiveMember**: active customers are less likely to leave the bank (as per the criteria defined by the bank for identifying the activeness).
  + 1 represents Active Member
  + 0 represents Inactive Member
* **Estimated Salary**: as with balance, people with lower salaries are more likely to leave the bank compared to those with higher salaries.
* **Exited**: whether or not the customer left the bank.
  + 0 represents Retain
  + 1 represents Exit
* **Bank DOJ**: date when the Customer associated/joined with the bank.

**Objective Questions**

1. **What is the distribution of account balance across different regions?**

For this first I connected the data in the mysql database and then connected the database with PowerBI. After this I made all the necessary relationships between all the tables which were imported in the PowerBI. Once all of this was done, then I created a Bar chart to visualise this data. For this, I choose GeographyLocation from the Geography table and the Balance Column from the bank\_churn table in Sum as an Aggregator. This graph shows that the highest balance is in France, followed by Germany and then Spain. This is mainly due to the fact that France has the highest number of accounts. If we switch the aggregation from Sum to Avg then the highest number of Avg balance per Customer will be Germany, followed by France and then Spain.



1. **Identify the top 5 customers with the highest Estimated Salary in the last quarter of the year. (SQL)**

For this particular use case, I wrote a SQL query to find out which user has the highest Estimate salary in the last quarter of the year (Oct, Nov, Dec).  
  
select a.﻿CustomerId, a.EstimatedSalary

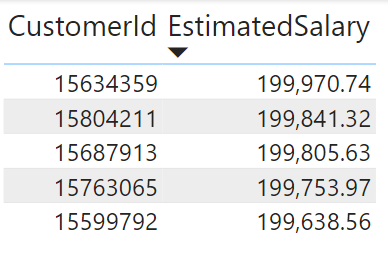
from capstone.customerinfo a

where extract(month from a.`Bank DOJ`) between 10 and 12

order by 2 desc

Limit 5

Then I connected the Mysql connection in PowerBI and entered the above query to create a connection. Now the data has been imported, I renamed the table which was created with the query and then created a table in PowerBI. Then I ordered the table in descending order of EstimatedSalary column.



1. **Calculate the average number of products used by customers who have a credit card. (SQL)**

For this use case, I’ve created a SQL query to find out the average number of products used by the customers who have credit card.

SELECT round(avg(NumOfProducts),2) avg\_number\_of\_products

FROM capstone.bank\_churn

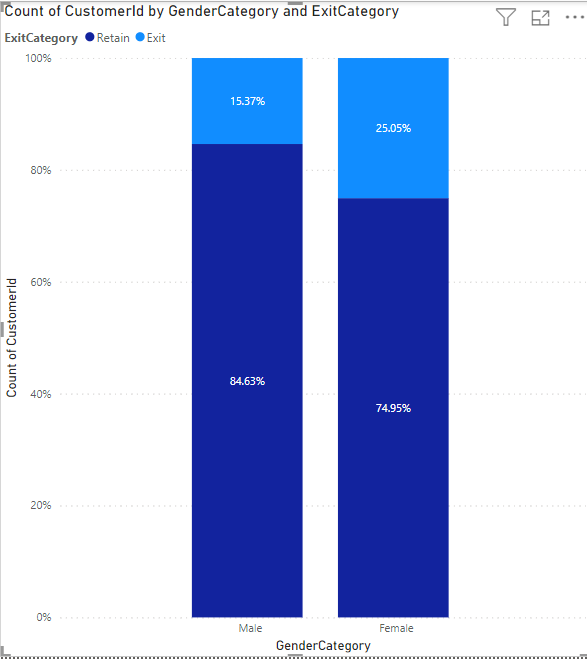
where HasCrCard = 1

Then we connected this query with the power bi to added the result as a table in powerbi. Then we created a card in Power Bi to display the result, which is coming out to be 1.53



1. Determine the churn rate by gender for the most recent year in the dataset.

For this, we created a stacked column chart in Power BI, where we placed GenderCategory from gender table on the x-axis, ExitCategory on the Legend and count of CustomerId on the y-axis. With this we can see that the male have 15.37% and females have 25.05% of churn rate.



1. Compare the average credit score of customers who have exited and those who remain. (SQL)

To calculate this, we have written the below SQL query to extract the data.

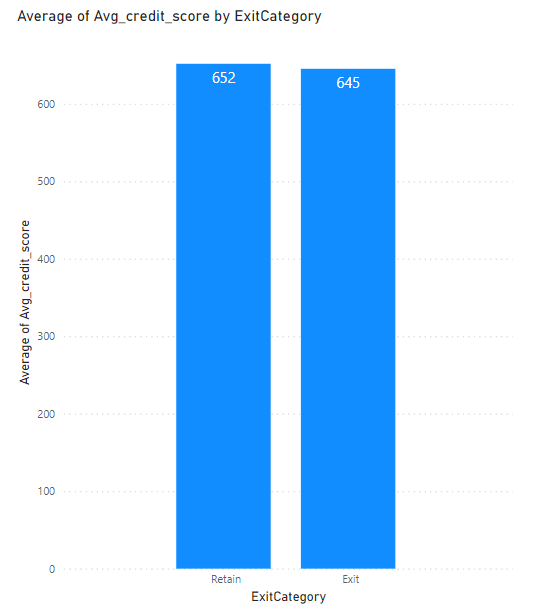
select b.ExitCategory, avg(a.CreditScore) Avg\_credit\_score

from capstone.bank\_churn a

join capstone.exitcustomer b on a.Exited = b.﻿ExitID

group by 1

Then we connected the query with the PowerBI and extracted and stored the data in a table format. Then with the extracted data we have created the bar chart for where we placed ExitCategory in the x-axis and Avg\_credit\_score in the y-axis. Here we can see that the Avg credit score of the Retained users is 652 whereas for the exited users it is 645.



1. Which gender has a higher average estimated salary, and how does it relate to the number of active accounts? (SQL)

For this we’ve created a SQL query to extract the data which is needed for the calculation.

select b.GenderCategory, round(avg(a.EstimatedSalary),2) Avg\_estimated\_salary from capstone.customerinfo a

join capstone.gender b on a.GenderID = b.﻿GenderID

group by 1

This query will help us in finding out the average estimated salary of customers as per their gender category. Here we can see that the average salary of the female customers is 101K and for male customers is 100K. For the correlation with this and the active accounts, we need another query to find out the active accounts share for different gender category.

select

b.GenderCategory,

d.ActiveCategory,

count(distinct a.﻿CustomerId) customers

from capstone.customerinfo a

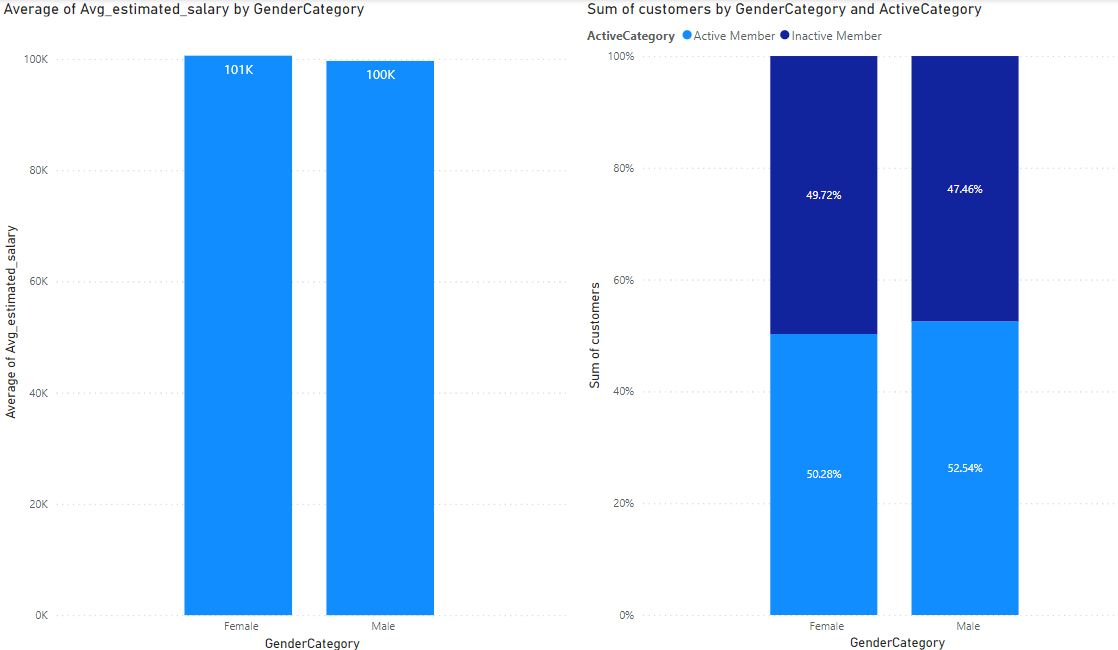
join capstone.gender b on a.GenderID = b.﻿GenderID

join capstone.bank\_churn c on a.﻿CustomerId = c.﻿CustomerId

join capstone.activecustomer d on c.IsActiveMember = d.﻿ActiveID

group by 1,2

After storing the data from this query in PowerBI as a table,and visualising the data in PowerBI as a stacked column chart, we can see that the Females have only 50.28% as Active customers and Males have 52.54% of active customers or accounts.



1. Segment the customers based on their credit score and identify the segment with the highest exit rate. (SQL)

To solve this problem, I’ve written this query to fetch the desired output. For the credit bucket, I’ve took the bucket logic which was given in the data introduction.

select

case when a.CreditScore between 300 and 579 then 'Poor'

when a.CreditScore between 580 and 669 then 'Fair'

when a.CreditScore between 670 and 739 then 'Good'

when a.CreditScore between 740 and 799 then 'Very\_Good'

when a.CreditScore between 800 and 850 then 'Excellent' else 'Others' end credit\_score\_bucket,

b.ExitCategory,

count(distinct a.﻿CustomerId) customers

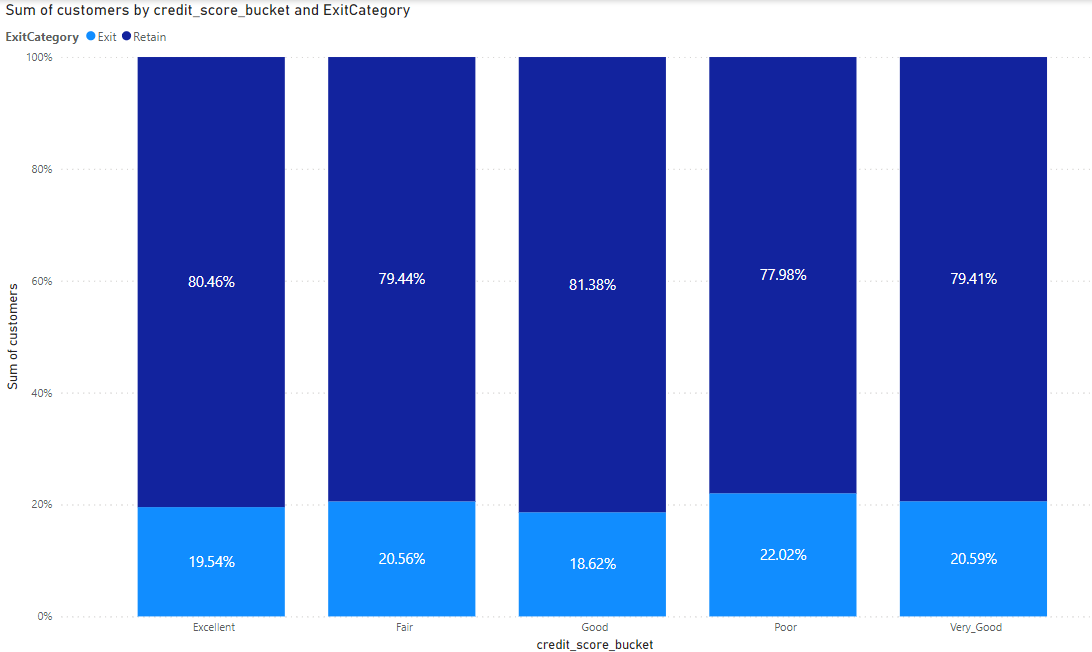
from capstone.bank\_churn a

join capstone.exitcustomer b on a.Exited = b.﻿ExitID

group by 1,2

order by 1,2

Then storing the data which we’re getting form the query as a table in PowerBI and then visualising the data as in stacked column chart, we can see that the Poor Credit score bucket has the highest churn rate of 22.02%.



1. Find out which geographic region has the highest number of active customers with a tenure greater than 5 years. (SQL)  
     
   To solve this question, we need to write the following query to fetch the desired dataset.

select c.GeographyLocation, count(distinct a.﻿CustomerId) customers

from capstone.customerinfo a

join capstone.bank\_churn b on a.﻿CustomerId = b.﻿CustomerId

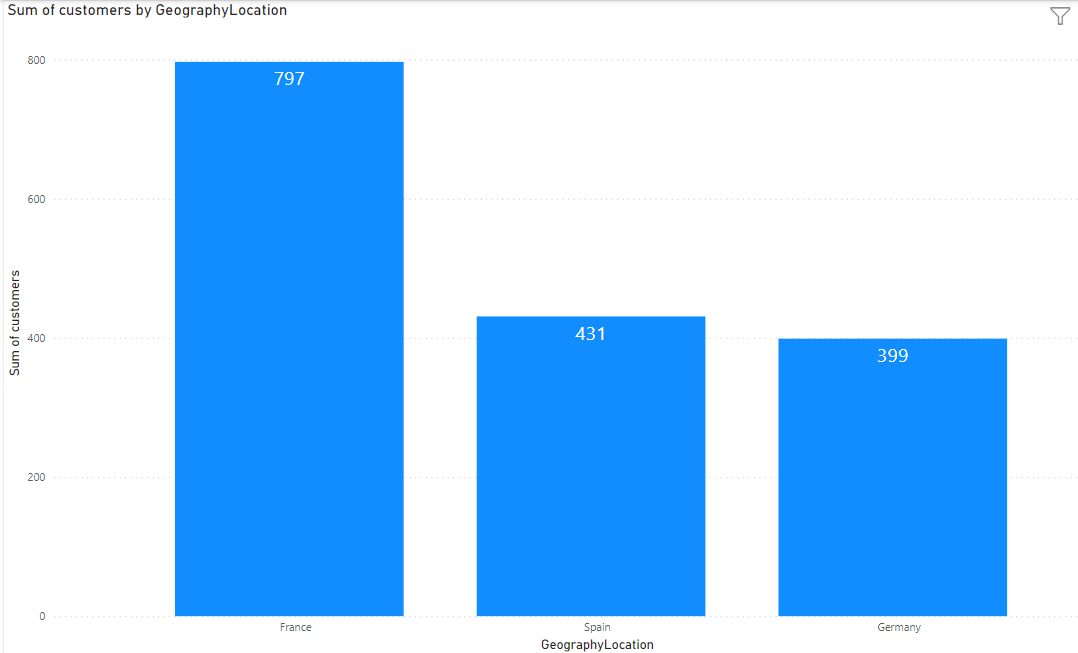
join capstone.geography c on a.GeographyID = c.﻿GeographyID

join capstone.activecustomer d on d.﻿ActiveID = b.IsActiveMember

where d.ActiveCategory = 'Active Member' and b.Tenure > 5

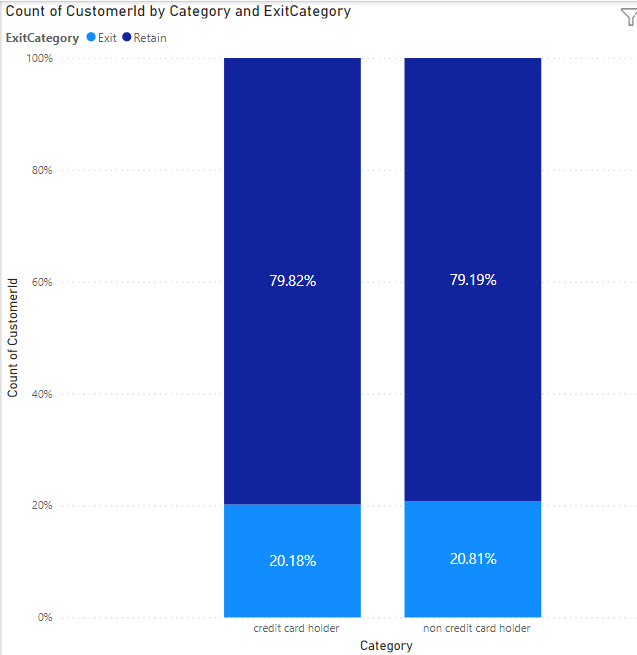
group by 1

After saving the data from this query result in a table we can visualize the data as a bar chart format. In this we can see that the France has the highest number of active users of 797



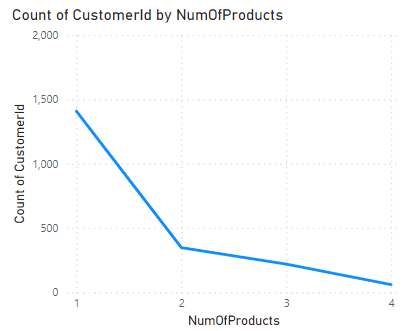
1. What is the impact of having a credit card on customer churn, based on the available data?  
     
   For this, we have created a stacked column chart, where Category from creditcard table was placed in x-axis, count of customerid from bank\_churn table was placed in y-axis and ExitCategory from exitcustomer table was placed as a Legend to visualize how the churn rate of customer is getting impacted by whether the customer is having the credit card or not.

Here we can see that the customers who are having the credit card has a churn rate of 20.18% and those who don’t have a credit card have a churn rate of 20.81%.



1. For customers who have exited, what is the most common number of products they had used?

To visualize this data, i’ve used the line chart where we placed NumOfProducts from the bank\_churn table in the x-axis and Count of customerId from the same table in the y-axis. Here we can see that the most of the customers are only using 1 product.



1. Examine the trend of customer joining over time and identify any seasonal patterns (yearly or monthly). Prepare the data through SQL and then visualize it.  
     
   For this, we’ve create a SQL query to fetch the data in the format which can show both yearly and monthly trend of all the customers with just changing the data hierarchy in PowerBI visualisation.  
     
   select

date\_format(`Bank DOJ`,'%Y-%m-01') as bank\_DOJ,

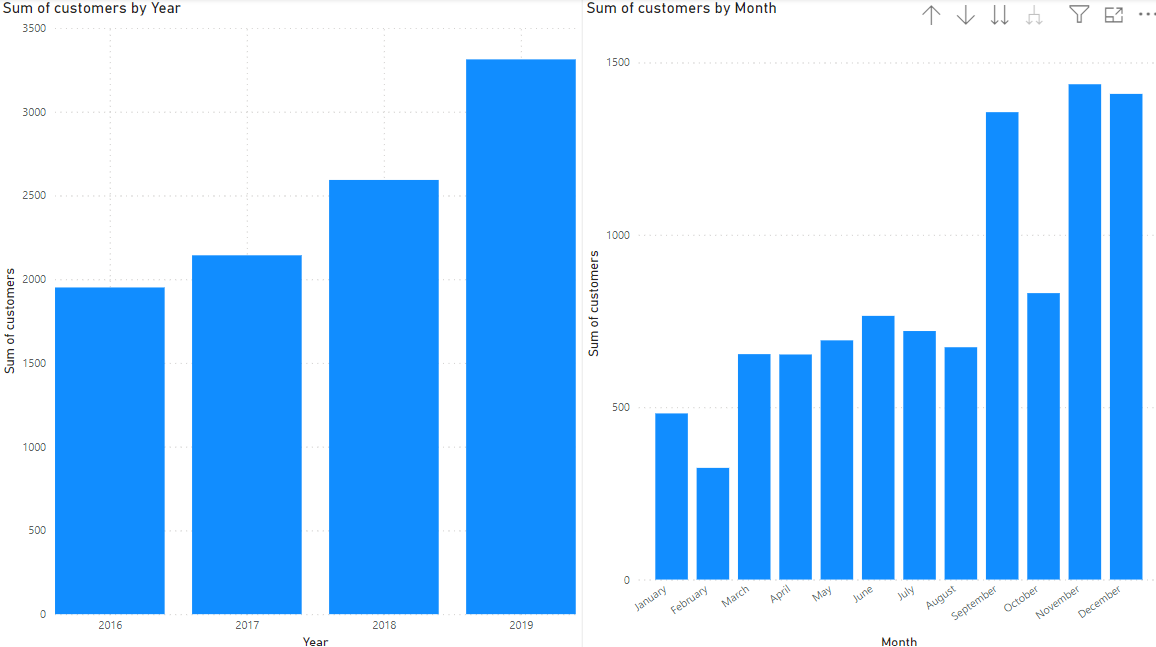
count(distinct a.﻿CustomerId) customers

from capstone.customerinfo as a

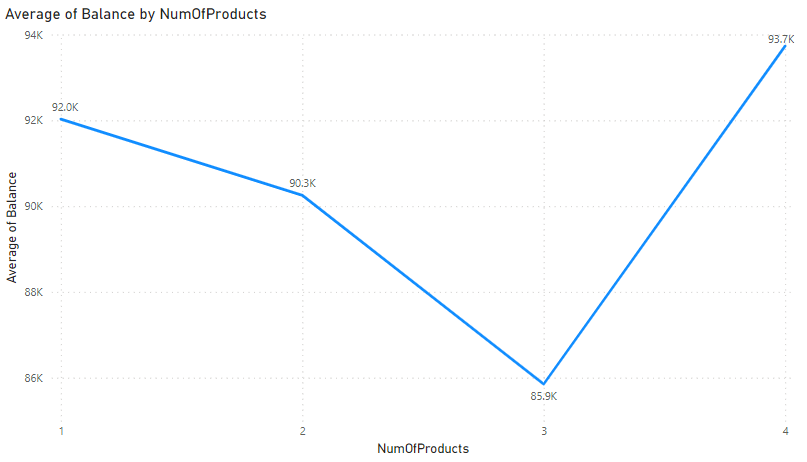
group by 1

order by 1

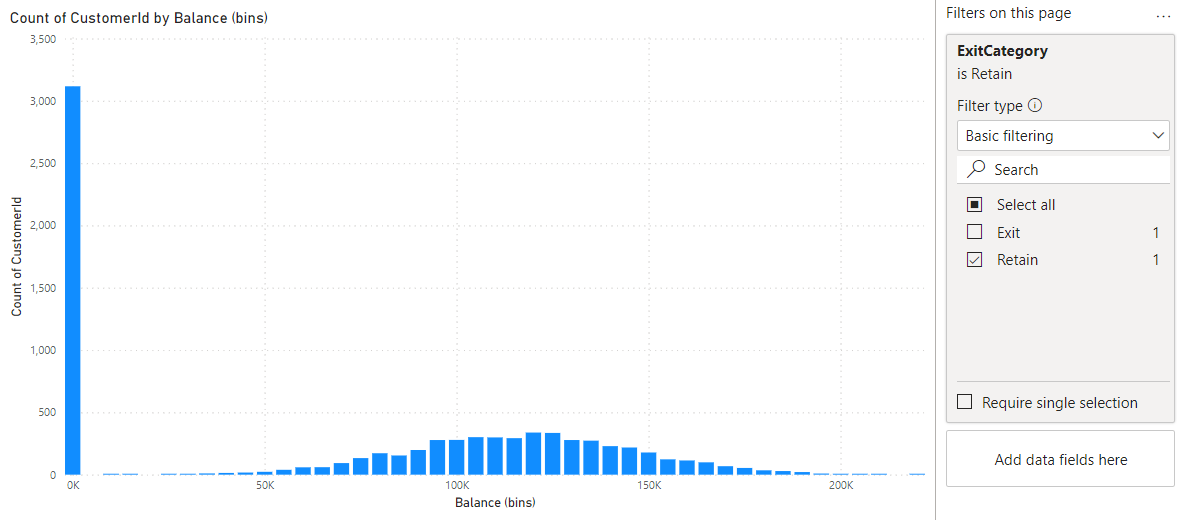
After storing the data from the query as table in the PowerBI and then visualise the data in a Bar chart. Here we can see that the number of customers who are registering in the bank has been increasing y-o-y where the maximum was in the year 2019. If we visualise the data in the monthly pattern, we can see that in the September, November and December we can see the higher number of customer date of joining as compared to the average of the all month.



1. Analyze the relationship between the number of products and the account balance for customers who have exited.  
     
   For this question, we plotted the line chart where NumOfProduct from the bank\_churn table and average of Balance on the y-axis. Here we can see that the users who has 3 products has the lowest of the average balance. The trend is in the downward trajectory except where the balance is also high for the users who has 4 product. But if we see from 1 to 3 number of product the avg balance of customer is in downward trend.



1. Identify any potential outliers in terms of balance among customers who have remained with the bank.

For this, we have transform the data of the bank\_churn table where we grouped the data of Balance into the bin of 5000. With this we can easily create the different buckets and see which balance group has how many customers. To analyse this, we have created a bar chart to see the distribution of customers. Here on x-axis, I’ve place the Balance bin and on y-axis, I’ve placed the count of customer id. After this, we’ve also place a filter for this page on ExitCategory to get the retained users only. Here we can see that there are around 3117 customers in the 0 balance bin, after that, there is a proper bell curve distribution maxing at around 120K of the balance.   
  


1. How many different tables are given in the dataset, out of these tables which table only consist of categorical variables?  
     
   There are a total of 7 tables which were provided to us for this project, out of which there are 5 tables which contains only categorical data which are as follows

* Activecustomer
* Gender
* Geography
* Exitcustomer
* creditcard

1. Using SQL, write a query to find out the gender wise average income of male and female in each geography id. Also rank the gender according to the average value. (SQL)  
     
   For this, I’ve written the query to find ot the gender wise average income in each geography along with the ranking of all those rows.  
     
   with a as

(

select

c.GeographyLocation,

b.GenderCategory,

round(avg(a.EstimatedSalary),2) as avg\_salary

from capstone.customerinfo a

join capstone.gender b on a.GenderID = b.﻿GenderID

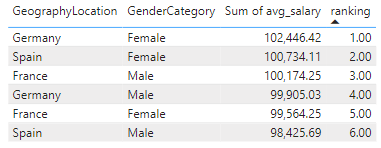
join capstone.geography c on a.GeographyID = c.﻿GeographyID

group by 1,2

)

select \*, rank()over(order by avg\_salary desc) as ranking

from a   
  
After importing the results in the PowerBI, the results are as follows:



1. Using SQL, write a query to find out the average tenure of the people who have exited in each age bracket (18-30, 30-50, 50+).  
     
   For this, we’ve created a query to fetch the desired result:  
     
   select

case when a.Age between 18 and 30 then '18-30'

when a.Age between 31 and 50 then '30-50' else '50+' end age\_bracket, round(avg(b.Tenure),2) avg\_tenure

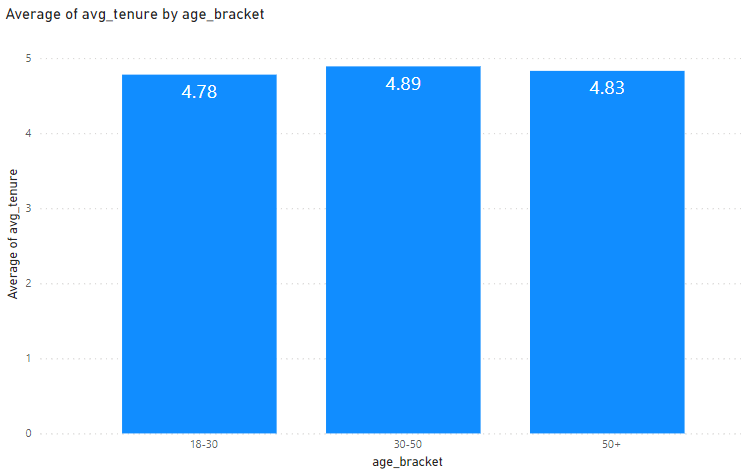
from capstone.customerinfo a

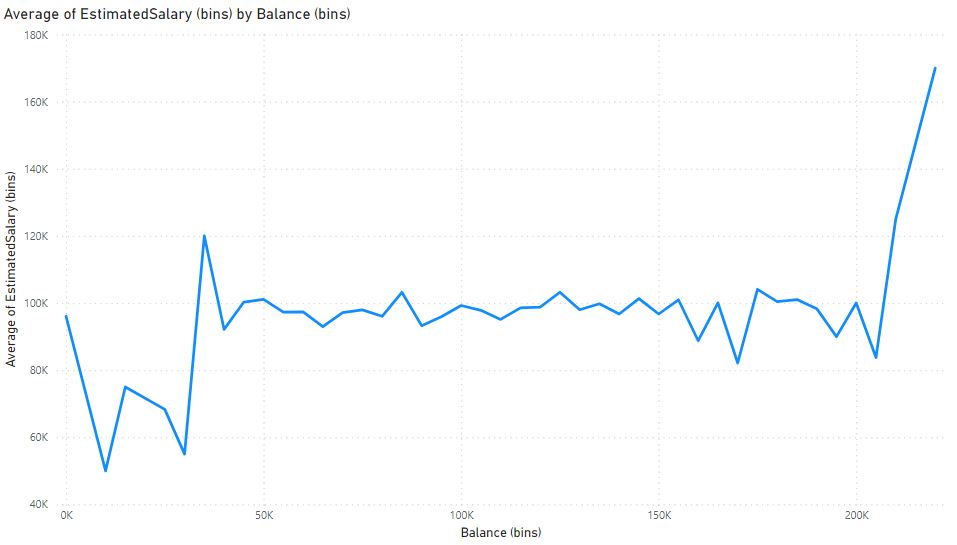
join capstone.bank\_churn b on a.﻿CustomerId = b.﻿CustomerId

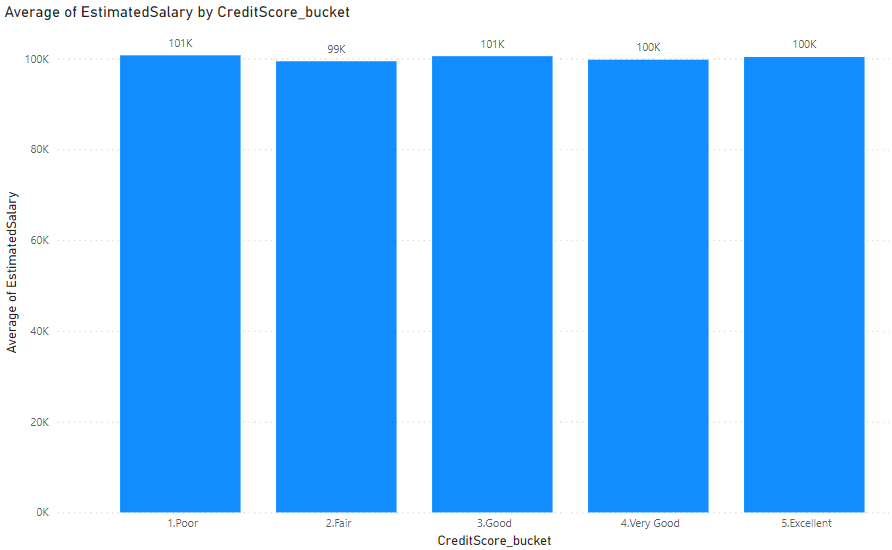
where b.Exited = 1

group by 1

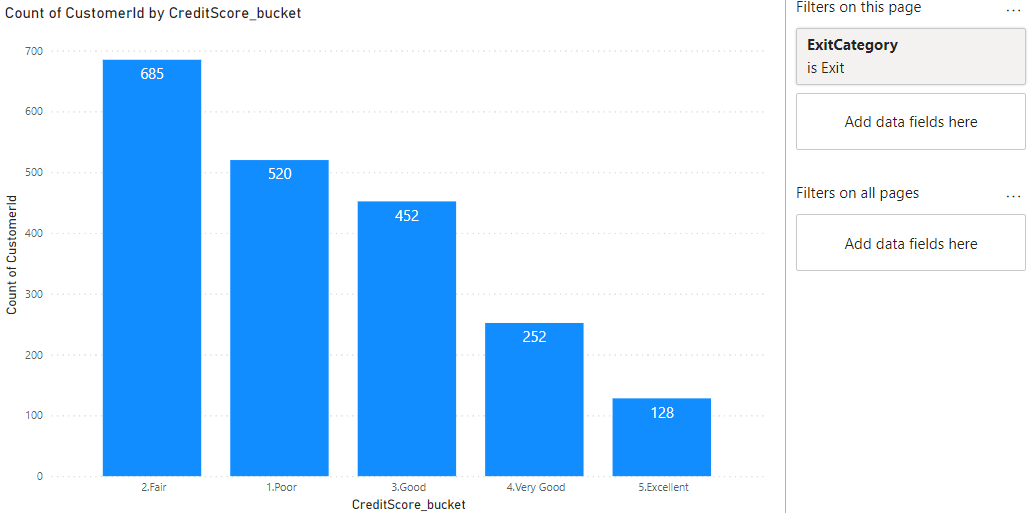
order by 1

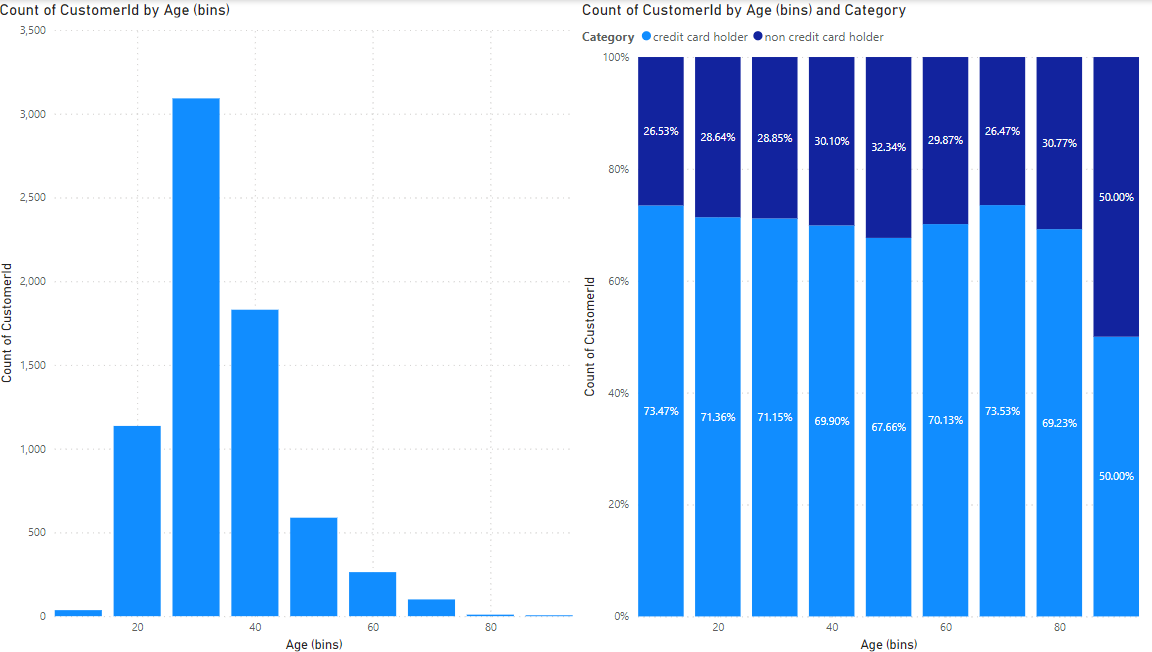
After importing the data in PowerBI as a table, and after visualising the data in bar chart, we can see that the Age\_bracket of 30-50 has the highest tenure of 4.89, whereas the other age brackets are not not much falling behind with 4.83 and 4.78 for age\_bracket 50+ and 18-30 respectively.  
  


1. Is there any direct correlation between salary and balance of the customers? And is it different for people who have exited or not?  
     
   For this analysis, we can re-utilize the previous balance bin and then plotting it in the line chart where we used the balance bin in the x-axis and average of Estimatedsalary in the y-axis. With this we can see that till the balance bin of 30K the average salary is around the 70K mark. After that, the average estimated salary increases to around 100K. After the 205K balance mark, the average salary increases rapidly.  
     
   
2. Is there any correlation between salary and Credit score of customers?  
     
   For this, we have created a conditional table with the condition to break the credit scores into different category as per mentioned in the starting of the document. With that done, we created a bar chart, where we used the creditscore\_bucket on the x-axis and average of estimatedsalary on the y-axis. Through this we can clearly see that there is not a much of a difference in the average salary of customers who are falling in these buckets, only 1K difference in the bucket. This represents that there is no correlation between the salary and the credit score.



1. Rank each bucket of credit score as per the number of customers who have churned the bank.

For this question, we will plot a bar chart where we will place the customerid from bank\_churn count in the y-axis and Credit score buckets which we created earlier on the x-axis. After this, we will place ExitCategory from the exitcustomer table and only selected Exit. Here we can see that Fair credit score bucket has the highest churn users whereas the Excellent has the lowest number of churn users. This might also be due to the fact that the the fair credit score users consist of the majority of the customer base.  
  


1. According to the age buckets find the number of customers who have a credit card. Also retrieve those buckets who have lesser than average number of credit cards per bucket.  
     
   For this question, I’ve created an age buckets by grouping the data in each a bucket of 10 steps. After doing that, I’ve created a bar chart where x-axis is the Age bucket which we just created and unique count of customerid from the bank churn table in y-axis. After that I placed a Filter of credit card category from the creditcard table and only selected credit card holder. After that we can analyse the data and we can see that the 30-40 age bucket is the highest.  
     
   After that to analyse the second part of the question, we need to plot the stacked bar chart, where we placed the age bucket on the x-axis and customerid count on y-axis. After that we placed Category from the creditcard table on the legend. Here we can see that the age group 90, 50, 80, 40,60 have less than average credit card holders, as compared to the rest of the group.  
   
2. Rank the Locations as per the number of people who have churned the bank and average balance of the learners.  
     
   with a as

(

SELECT

c.GeographyLocation,

count(distinct case when a.Exited = 1 then a.﻿CustomerId end) churn\_users,

round(avg(a.Balance),2) avg\_balance

FROM capstone.bank\_churn as a

join capstone.customerinfo b on a.﻿CustomerId = b.﻿CustomerId

join capstone.geography c on c.﻿GeographyID = b.GeographyID

group by 1

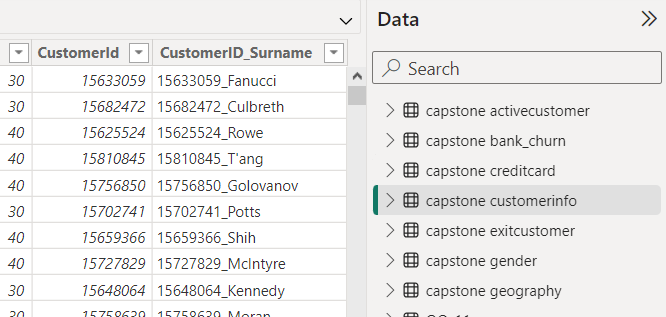
)

select \*, row\_number()over(order by churn\_users desc, avg\_balance desc) ranking

from a

order by 4 asc

With this query we can find the ranking of the locations as per the number of people who have churned the bank and the average balance of the customers.

1. As we can see that the “CustomerInfo” table has the CustomerID and Surname, now if we have to join it with a table where the primary key is also a combination of CustomerID and Surname, come up with a column where the format is “CustomerID\_Surname”.  
     
   For this, we selected the transform table from the menu bar on the home page. After that customerinfo table and selected the customerid column and Surname column from the table and then clicked on the merge column option from the add column section from the menu bar. After that we selected separator as “\_” and game the new column name as “CustomerID\_Surname”. Then we closed and apply the changes which was done on the transform data page. Now the new merged column in the table is visible.  
   
2. Without using “Join”, can we get the “ExitCategory” from ExitCustomers table to Bank\_Churn table? If yes do this using SQL.  
     
   For this question, please find the query mentioned below:  
     
   SELECT bc.\*, (SELECT ec.ExitCategory FROM capstone.exitcustomer ec WHERE ec.﻿ExitID = bc.Exited) AS ExitCategory

FROM capstone.bank\_churn bc  
  
With this we can get all the columns from the bank\_churn table along with ExitCategory from the exitcustomers table.

1. Were there any missing values in the data, using which tool did you replace them and what are the ways to handle them?  
     
   There were no missing values which was present in the tables shared to us.
2. Write the query to get the customer ids, their last name and whether they are active or not for the customers whose surname ends with “on”.  
     
   For this question, please find the query mentioned below  
     
   select a.﻿CustomerId, a.Surname, c.ActiveCategory

from capstone.customerinfo a

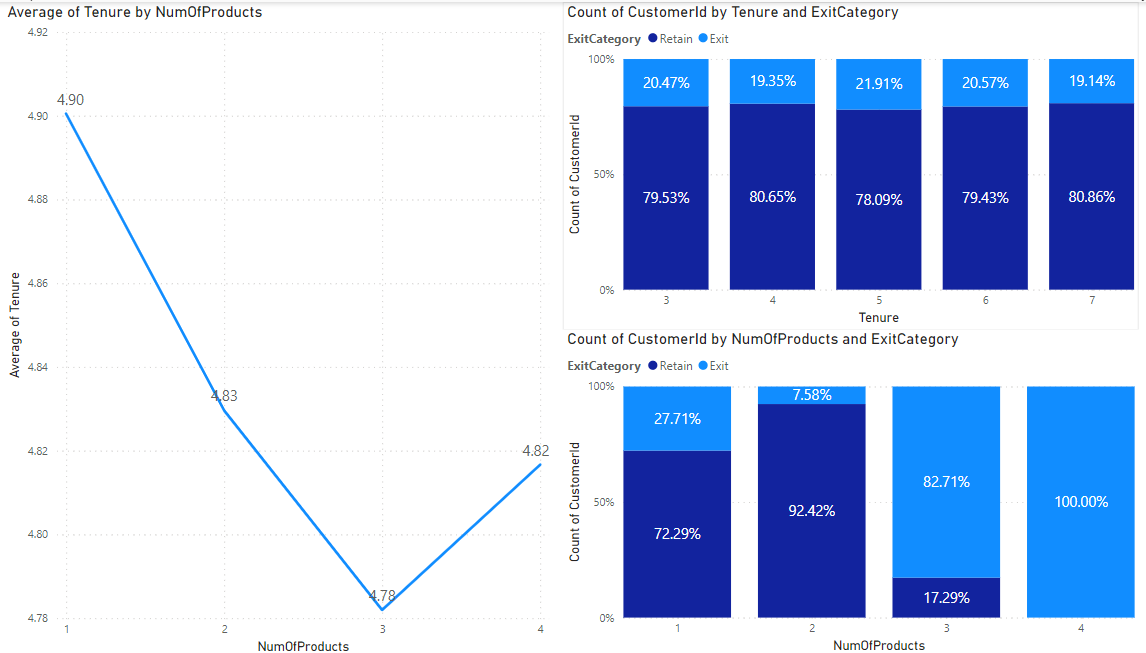
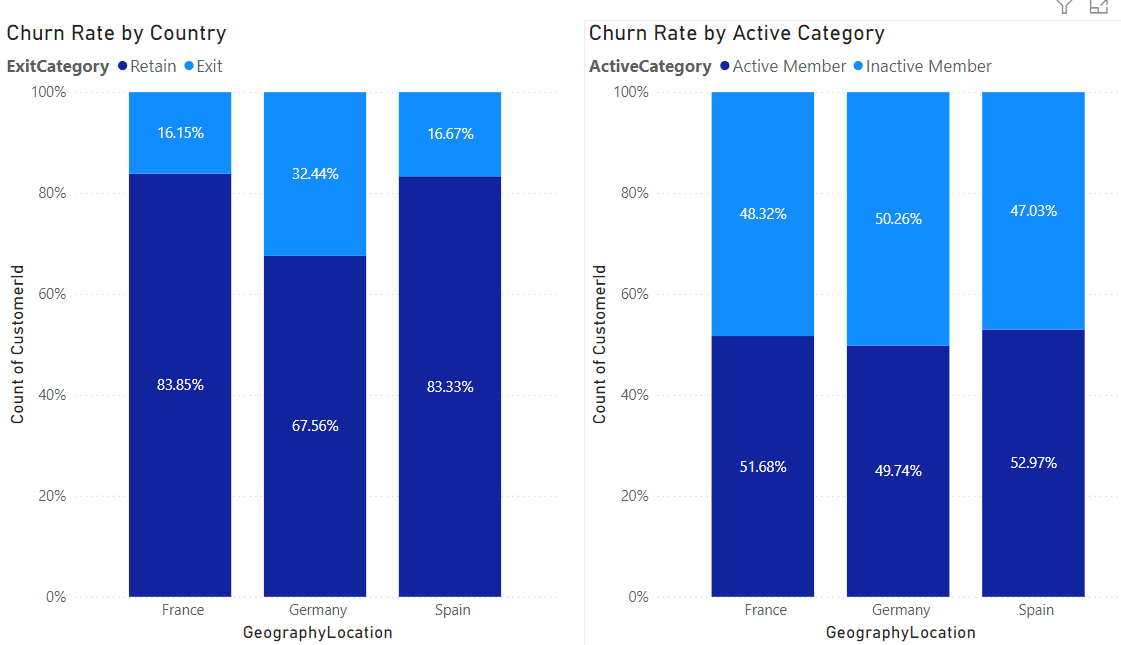
join capstone.bank\_churn b on a.﻿CustomerId = b.﻿CustomerId

join capstone.activecustomer c on b.IsActiveMember = c.﻿ActiveID

where a.Surname like '%on%

1. Can you observe any data disrupency in the Customer’s data?  
     
   There are no data discrepancy which is directly visible in the customer’s dataset which was provided to us. But there was one piece of information which in there are certain set of customers who have been exited from the bank but are showing as active in the database. These shouldn’t be happening as if the users has already exited then it can’t be active within the system.

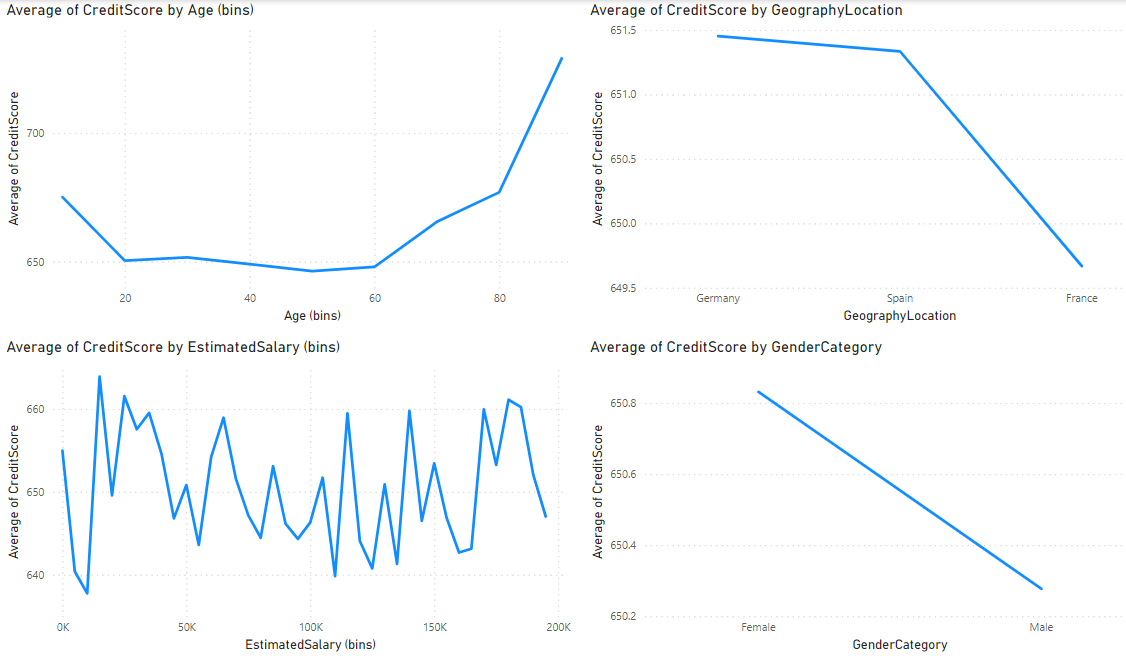
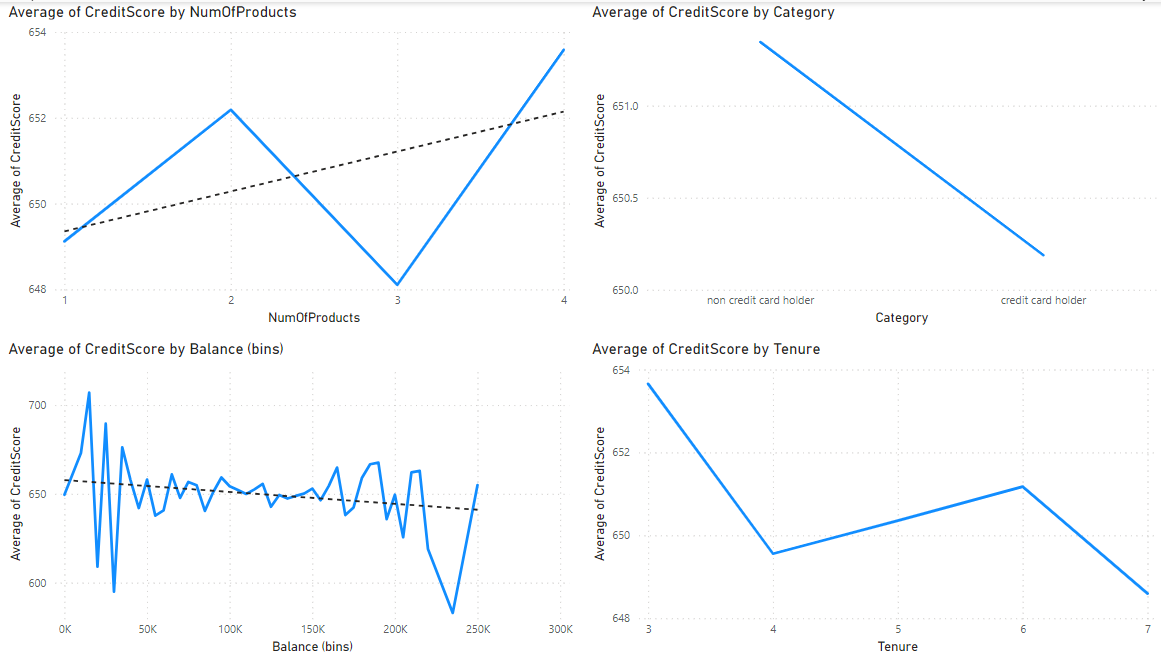
**Subjective Questions**

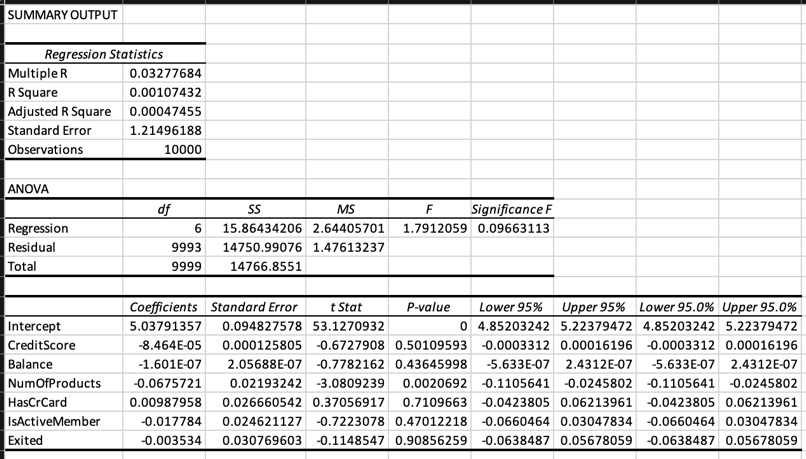
1. Customer Behavior Analysis: What patterns can be observed in the spending habits of long-term customers compared to new customers, and what might these patterns suggest about customer loyalty?  
     
   For this particular section, we want to analyse how the tenure of customer gets affected by the number of products, how the tenure is affecting the customer churn rate and how no of products are affecting the exit of the employees.   
   Firstly, we create a line chart where we placed the NumofProduct on the x-axis and average of Tenure on the y-axis. Here we can see that as the number of product increases the tenure of customer decreases. After this, we created a stacked bar chart where we plotted the Tenure on the x-axis and count of customer id on the y-axis along with this exit category on the legend. With this we cannot see any significant trend in the behaviour. After this we created a stacked bar chart and added numofproduct on the x-axis and count of customerid on the y-axis and Exit category on the legend. Here we can see that the best num of product to be sold is 2 as after that the churn rate of the users increases drastically.   
   
2. Product Affinity Study: Which bank products or services are most commonly used together, and how might this influence cross-selling strategies?  
     
   There is no not sufficient data data which is available in the dataset on the different products or services sold to the customers. So there is no analysis which can be done.
3. Geographic Market Trends: How do economic indicators in different geographic regions correlate with the number of active accounts and customer churn rates?  
     
   To analyse this, we created a stacked bar chart where we placed GeographyLocation on the x-axis and count of customerid on the y-axis and ExitCategory on the Legend. Here we can see that the germany has the highest churn rate of 32.44% as compared to other country which is around 16%. This shows that Germany has around 2 times the churn rate as compared to the other countries. Apart from this, we created another stacked bar chart where we plotted GeographyLocation is on the x-axis and count of customerid on the y-axis and Active Category on the legend. Here also we can see that the number of active users in Germany is higher than the other country which is around 2-3% higher than that of other countries. This shows that there is some genuine issue which is happening in Germany banks location due to which the Churn rate is higher and inactive users are also higher.  
   
4. Risk Management Assessment: Based on customer profiles, which demographic segments appear to pose the highest financial risk to the bank, and why?  
     
   Risk management assessment of the customer is generally measured by the credit score of the customer. If the credit score of the customer is high, it is potentially low risk customer as this shows that the customer will pay his credits on time. If the customers have low credit score then that means that the there have been some late payments due to which users credit scores have been dropped. In order to check which particular demographic have low credit score, we need to check on various different criterias like age, geography, gender, salary etc. For that we need to create multiple charts to understand this.  
     
   For this, we created a line chart of Age bin and the avg of credit score, here we can see that the age bin from 20 to 60 have the lower credit scores as compared to the rest of the age bins.  
   Then we created the line chart for the Geography location and the average of credit score. Here we can see that France have lower credit score as compared to the rest but the difference is not so much to raise any alarm.  
   Then we created the line chart for the Estimated salary bin and the avg of credit score. Here we can see that for each salary bin the avg credit score goes up or down by only 10 points. There is no clear indication or pattern as such to find any conclusion.  
   Then we created another line chart for the Gender category and avg of credit score. From this we can see that the Male has lower than Female customers, but the difference is negligible.

Then we created another line chart of the Num of Product vs the avg credit score. Here after plotting the trendline, we can see that as the num of product increases the avg credit score increases, but by very small margin.

Then we created another line chart of Credit card category vs the avg credit score. Here we can see that the users who have credit card has lower avg credit score by a very small margin.

Then we created line chart of Balance bin vs the average credit score. Here we can see that the graph is very fluctuating so there is no clear insight, but once we plot the trend line, we can see that as the balance of the customer increases the avg credit score decreases.

And lastly we also created a line chart for the Customer Tenure vs the avg of credit score. Here we can see that as the tenure of the customer increases the avg credit score of the customer decreases.  
  
  


1. Customer Lifetime Value Forecast: How would you use the available data to model and predict the lifetime (tenure) value of different customer segments?   
     
   In order to this analysis, we’ve ran the Regression analysis in the Excel which is present in the Data analysis section in Data Tab. Here are the insight which we got from there  
   

The output is divided into several sections: Regression Statistics, ANOVA, and the Coefficients table.

1. Regression Statistics  
   This section provides an overview of the fit of the regression model.  
   1. Multiple R: 0.03277684  
     
   This is the correlation coefficient, which measures the strength and direction of a linear relationship between the observed and predicted values. A value close to 0 indicates a very weak linear relationship.

2. R Square: 0.00107432  
  
This is the coefficient of determination, which indicates how well the independent variables explain the variability of the dependent variable. An R Square of 0.00107432 means that only 0.1% of the variability in the dependent variable is explained by the model.  
  
3. Adjusted R Square: 0.00047455  
  
This adjusts the R Square value for the number of predictors in the model. It is slightly lower than the R Square, which is expected when adding more predictors does not improve the model much.  
  
4. Standard Error: 1.21496188  
  
This is the standard deviation of the residuals, which gives an idea of the average distance that the observed values fall from the regression line.  
  
5. Observations: 10000  
  
This indicates the number of observations or data points used in the regression analysis.

1. ANOVA (Analysis of Variance)  
     
   This section provides an analysis of the variance in the data.  
     
   - df (Degrees of Freedom):  
    - Regression: 6 (number of independent variables)  
    - Residual: 9993 (total observations - number of independent variables - 1)  
    - Total: 9999 (total observations - 1)  
     
   - SS (Sum of Squares):  
    - Regression: 15.86434206  
    - Residual: 14750.99076  
    - Total: 14766.8551 (this is the sum of Regression SS and Residual SS)

- MS (Mean Square):  
 - Regression: 2.64405701 (Regression SS / df for regression)  
 - Residual: 1.47613237 (Residual SS / df for residual)

* F: 1.7912059  
  This is the F-statistic for the overall regression model. It tests whether at least one of the predictor variables has a non-zero coefficient. A higher F-value indicates that the model is more significant.

- Significance F: 0.09663113  
This is the p-value for the F-statistic. A p-value greater than 0.05 suggests that the overall model is not statistically significant at the 5% significance level.

1. Coefficients Table  
   This section provides detailed information about each predictor in the model.  
     
   - Coefficients:  
   These are the estimated coefficients for the regression equation. For example, the coefficient for Intercept is 5.03791357, and for CreditScore it is -8.464E-05.  
     
   - Standard Error:  
   The standard error of each coefficient, which measures the variability of the coefficient estimate.  
     
   - t Stat:  
   The t-statistic for each coefficient, which tests whether the coefficient is significantly different from zero.  
     
   - P-value:  
   The p-value for each t-statistic. A p-value less than 0.05 typically indicates that the coefficient is statistically significant.  
   For example: NumOfProducts has a p-value of 0.0020692, which is less than 0.05, indicating it is a significant predictor.  
   Other variables like CreditScore, Balance, HasCrCard, IsActiveMember, and Exited have p-values greater than 0.05, indicating they are not significant predictors in this model.  
     
   - Lower 95% and Upper 95%:  
   These are the lower and upper bounds of the 95% confidence interval for each coefficient. If the interval includes zero, the coefficient is not statistically significant at the 95% confidence level.

Summary

* The model explains a very small proportion of the variance in the dependent variable (R Square is very low).
* Most predictors are not statistically significant except NumOfProducts.
* The overall model is not statistically significant (Significance F > 0.05).

This suggests that the predictors used in this regression model do not adequately explain the variability in the dependent variable. Further investigation and possibly the inclusion of other variables may be needed to improve the model.

1. Marketing Campaign Effectiveness: How could you assess the impact of marketing campaigns on customer retention and acquisition within the dataset? What extra information would you need to solve this?  
     
   Marketing campaign can be of various types, it can be TV ads, mobile ads, newspaper ads, Billboards, poster, etc., or it could even be as simple as telephonic call or whatsapp message. At first, we would like to understand, what kind of campaign was ran, via what medium, to which users, for how many days, what exact aspect was targetted from those ads. Apart from this, would also want to know how much spending happened on those campaigns to get the ROI if there is a possibility. These data points can help us understand the marketing campaign effectiveness.
2. Customer Exit Reasons Exploration: Can you identify common characteristics or trends among customers who have exited that could explain their reasons for leaving?  
     
   For this analysis, we tried to analyse the exit % of customers based on the following criterias in stacked column chart:   
    - Num of products

- Estimated Salary bin

- Tenure

- Age bin

- Balance bin

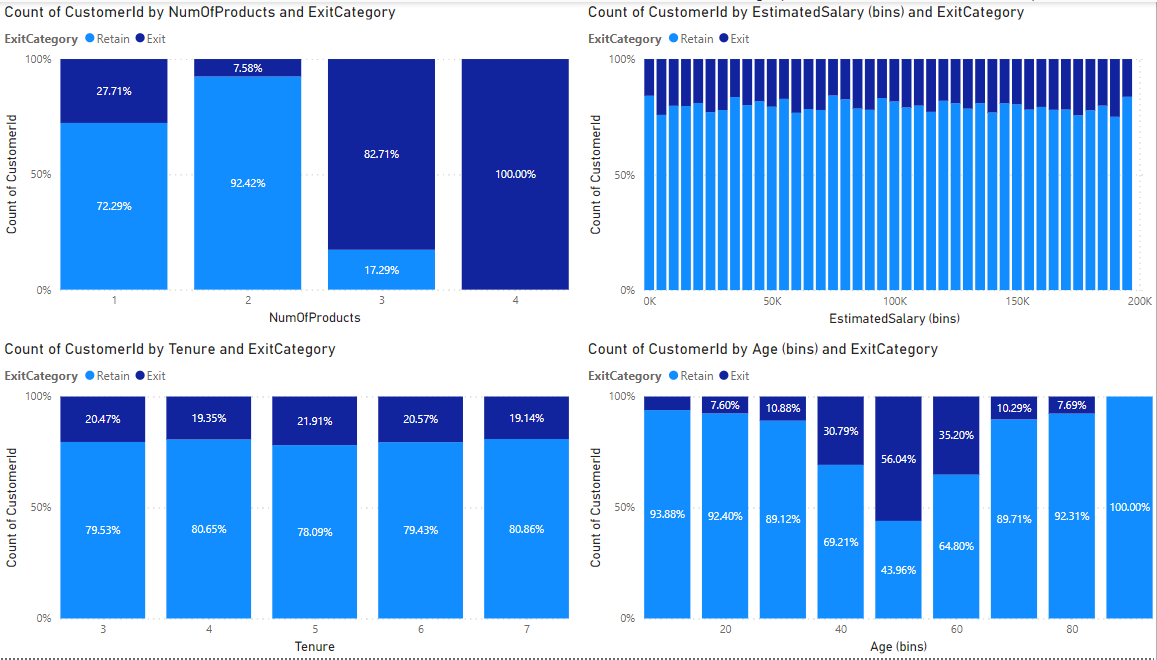
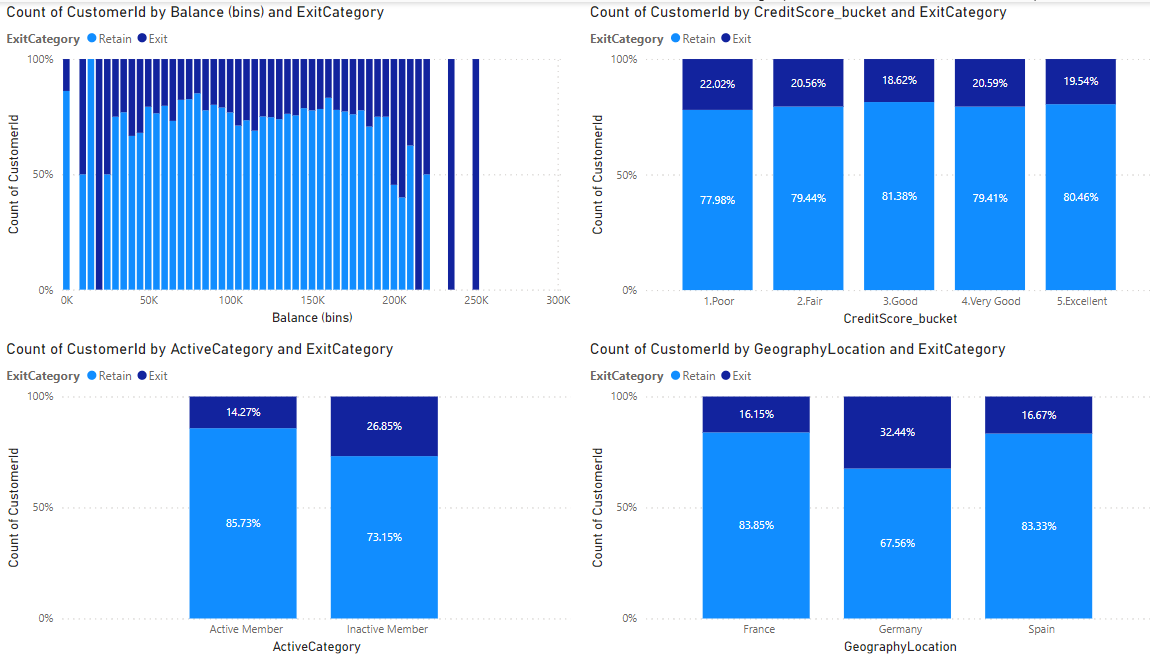
- Credit score bucket

- Active Category

- Geography Location

In this, there are majorly four factors for which the exit % is higher in some cohort or varies majorly for good volume, those are

* Num of Product - as the number of products sold to the customer increases the exit % of the customer increases
* Age bin - as the customer is of 40 to 60 bin, the exit % of the customer is much higher as compared to the other cohort
* Active Category - the inactive users have a higher chance of being churned as compared
* Geography Location - The customers from Germany has the higher exit % as compared to the rest other countries

1. Are 'Tenure', 'NumOfProducts', 'IsActiveMember', and 'EstimatedSalary' important for predicting if a customer will leave the bank?  
     
   As mentioned in the above answer, the num of Product and Is Active Member data is important for predicting whether the customer will leave the bank of not. But the Estimated salary has not shown any specific pattern which tells us about the exit or churn rate.
2. Utilize SQL queries to segment customers based on demographics and account details.  
   To write the above query, I’m currently targeting the age bucket for the demographics at the moment

select case when a.Age between 10 and 19 then '10-20'

when a.Age between 20 and 29 then '20-30'

when a.Age between 30 and 39 then '30-40'

when a.Age between 40 and 49 then '40-50'

when a.Age between 50 and 59 then '50-60'

when a.Age between 60 and 69 then '60-70'

when a.Age between 70 and 79 then '70-80'

when a.Age between 80 and 89 then '80-90' else 'Other' end Age\_bucket,

round(avg(Balance),2) avg\_balance

from capstone.customerinfo a

join capstone.bank\_churn b on a.﻿CustomerId = b.﻿CustomerId

group by 1

order by 1

1. How can we create a conditional formatting setup to visually highlight customers at risk of churn and to evaluate the impact of credit card rewards on customer retention?

Here are the steps for the coditional formatting in PowerBI:

1. Select the Table Visual.

2. Go to the Format Pane.

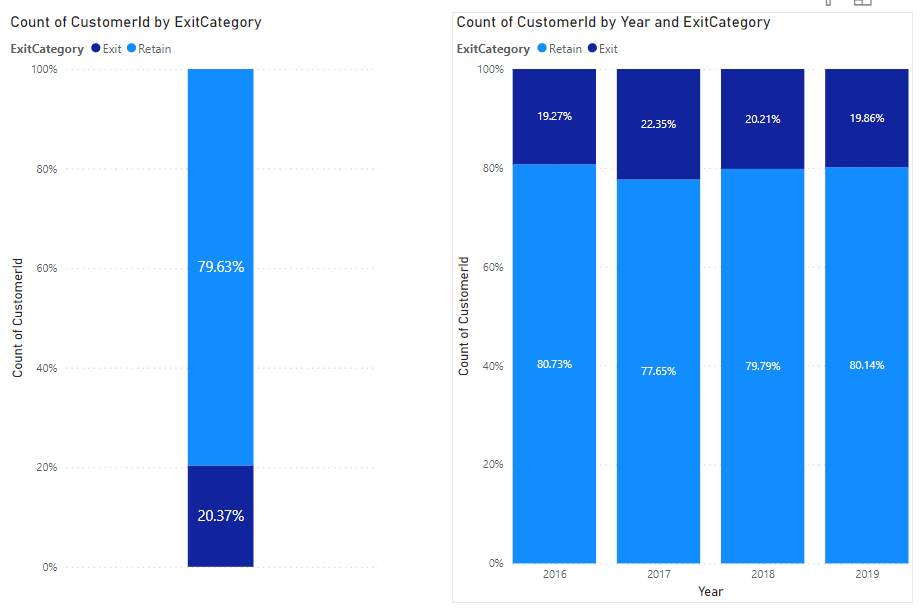
3. Expand the Conditional Formatting Section.

4. Select the Field to Format (e.g., Churn Risk).

5. Choose a Color Scale or Rule-Based Formatting

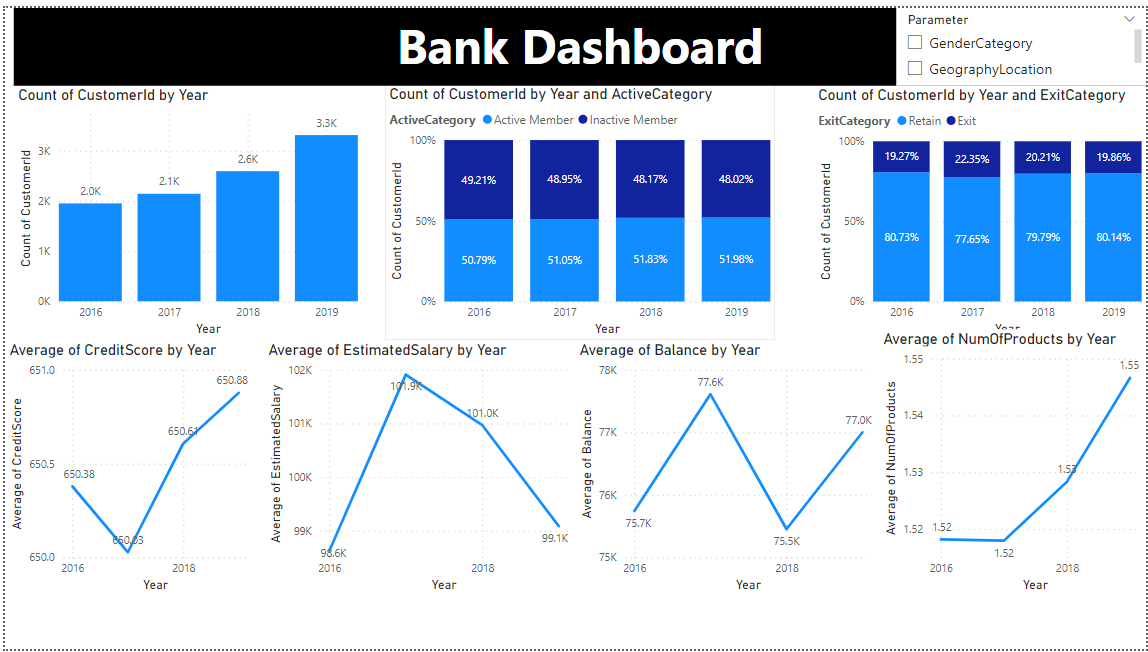
1. What is the current churn rate per year and overall as well in the bank. Can you suggest some insights to the bank about which kind of customers are more likely to churn and what are the different strategies that can be used to decrease the churn rate.

To analyse this, Firstly we created a stacked bar chart where we placed the exit category on the legend and the count of customer id on the y-axis. Through this we can see that the overall churn rate is around 20.4%. Then duplicating the gharph and adding the year in the x-axis, which will give us the year level year level churn rate. For the Churn rate reasons, I’ve already mentioned in the above subjective question number 7.



1. Create a dashboard incorporating all the KPIs and visualization related metrics. Use a slicer in order to assist in selection in the dashboard.

Here we have created the dashboard with the following graphs, customer id count, active and inactive member split, retain and exit members split, avg of credit score, avg of estimated salary, avg of balance and number of products sold. Apart from this I’ve created a parameter which contains 3 things, year, Gender category and Geography location, which is placed in the x-axis in all the graphs.



1. How would you approach this problem, if the objective and subjective questions weren't given?

If there were no instruction given then, I would’ve checked the churn rate reasons first, and then the risk assessment would have been performed. These activity helps us in understanding the underlying problems if there are any with the existing customers which could be helpful in the prediction of upcoming customers and offerings.

1. In the “Bank\_Churn” table how can you modify the name of “HasCrCard” column to “Has\_creditcard”?  
     
   In Power BI, you can rename columns directly within the data view. Here's how you can modify the name of the "HasCrCard" column to "Has\_creditcard":
2. Open Power BI Desktop: Launch Power BI Desktop.
3. Load Data: Load the "Bank\_Churn" table into Power BI. You can do this by connecting to your data source and selecting the "Bank\_Churn" table.
4. Navigate to Data View: In the left-hand panel, click on the "Data" icon to switch to the data view.
5. Locate the Column: Find the column named "HasCrCard".
6. Rename the Column:
   * Right-click on the column header (HasCrCard).
   * Select "Rename" from the context menu.
   * Type "Has\_creditcard" as the new column name and press Enter.
7. Close and Apply Changes: After renaming the column, click on the "Close & Apply" button in the Home tab to save the changes and close the Power Query Editor.

Once you apply these steps, the column name will be updated to "Has\_creditcard" in your Power BI dataset.