

## Objective Questions:

### 1. What is the distribution of account balances across different regions?

➤ Query:

```
1  Select
2      g.GeographyLocation as Region,
3      Round(Count(bc.Balance),2) as NumberOfCustomers,
4      Round(SUM(bc.Balance),2) as TotalBalance,
5      Round(AVG(bc.Balance),2) as AverageBalance
6  From bank_churn bc
7  Inner Join customerinfo c on c.CustomerId = bc.CustomerId
8  Inner Join geography g on c.GeographyID = g.GeographyID
9  Group By g.GeographyLocation
10 Order By TotalBalance DESC;
```

Output:

Region	NumberOfCustomers	TotalBalance	AverageBalance
France	5014	311332479.49	62092.64
Germany	2509	300402861.38	119730.12
Spain	2477	153123552.01	61818.15

#### Insights:

- Customers with higher credit scores and balances are more likely to stay active, as reflected in the bank\_churn table.
- There may be a geographic trend in customer churn rates, as GeographyLocation in the geography table is linked to the bank\_churn table.
- The number of products a customer holds (NumOfProducts) can be a strong indicator of customer engagement and loyalty.
- Credit card categories may influence customer behaviour, especially in relation to churn and account activity.

#### Recommendations:

- Focus on retaining high-balance, high-credit score customers by offering personalized incentives and services.
- Analyse churn rates across different geographic locations to develop region-specific retention strategies.
- Encourage customers to engage with more products by promoting bundled offerings or cross-sell opportunities.
- Use insights from activecustomer and exitcustomer tables to improve customer activation rates and address key reasons for exit.

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

➤ Query:

```
1 With RankedSalaries as (  
2   select  
3     CustomerId,  
4     Surname,  
5     EstimatedSalary,  
6     BankDOJ,  
7     DENSE_RANK() OVER (Order by EstimatedSalary DESC) as SalaryRank  
8   From customerinfo  
9   where month(BankDOJ) IN (1,2,3))  
10 Select  
11   CustomerId,  
12   Surname,  
13   EstimatedSalary,  
14   BankDOJ,  
15   SalaryRank  
16 From RankedSalaries  
17 Where SalaryRank <= 5  
18 Order by SalaryRank;
```

Output:

CustomerId	Surname	EstimatedSalary	BankDOJ	SalaryRank
15772601	Lu	199761.29	2017-02-13	1
15782758	Ozerova	199674.83	2017-03-30	2
15698474	Sagese	199661.5	2019-02-19	3
15814040	Munroe	199657.46	2018-01-01	4
15743040	Kuznetsova	199645.45	2019-03-08	5

Insights:

- The top five customers in terms of estimated salary have long tenures, indicating a possible correlation between salary and customer loyalty.
- There is only a slight difference in estimated salary among the top customers, suggesting that the bank's high-earning clients have similar financial profiles.
- Customers who joined the bank earlier, like Lu and Ozerova, tend to have higher ranks in terms of salary, indicating that long-term customers may have accumulated more wealth.
- All customers in this list joined the bank between 2017 and 2019, indicating a relatively stable acquisition of high-salary customers during these years.

Recommendations:

- Focus retention efforts on high-salary customers, as they appear to stay with the bank for longer periods.
- Consider creating exclusive product offerings or services tailored to customers

with estimated salaries above a certain threshold to increase engagement and loyalty.

- Use the insights from this list to target similar high-salary customers with personalized marketing strategies, based on their long-standing relationship with the bank.
- Monitor salary trends among newer customers to ensure the bank continues to attract high-earning individuals, adjusting strategies if necessary to maintain a steady influx of such clients.

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

➤ Query:

```
1 Select
2     Round(Avg(NumOfProducts),2) AS AverageNumberOfProducts
3 From bank_churn
4 Where CreditID = 1;
```

Output:

AverageNumberOfProducts
1.53

Insights:

- The average number of products used by customers with CreditID 1 (who have a credit card) is 1.53, indicating that customers with credit cards typically utilize 1-2 products from the bank.
- This average suggests that many customers with credit cards may not be fully leveraging additional banking products and services.
- A narrow range in the number of products indicates a possible lack of product diversity among credit card holders.
- The data may reflect opportunities to engage customers more broadly across other banking services beyond credit cards.

Recommendations:

- Implement cross-selling campaigns targeted at credit card holders to introduce them to other financial products, such as loans, savings accounts, or investment services.
- Analyse the factors leading to low product utilization among credit card customers and address any barriers that might prevent customers from adopting additional products.
- Provide personalized recommendations or financial incentives to encourage

customers to explore more of the bank's offerings and increase their product count.

- Segment credit card customers based on product usage behaviour to create more focused marketing strategies for those who are using only one or two products.

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

➤ Query:

```
1 WITH RecentYearData AS (  
2     SELECT MAX(YEAR(BankDOJ)) AS MostRecentYear  
3     FROM customerinfo  
4 ),  
5 GenderChurn AS (  
6     SELECT  
7         g.GenderCategory AS Gender,  
8         COUNT(CASE WHEN b.ExitID = 1 THEN 1 END) AS ChurnedCustomers,  
9         COUNT(*) AS TotalCustomers  
10    FROM bank_churn b  
11    JOIN customerinfo c ON b.CustomerID = c.CustomerID  
12    JOIN gender g ON c.GenderID = g.GenderID  
13    JOIN RecentYearData r ON YEAR(c.BankDOJ) = r.MostRecentYear  
14    GROUP BY g.GenderCategory  
15 )  
16 SELECT  
17     Gender,  
18     ChurnedCustomers,  
19     TotalCustomers,  
20     (ChurnedCustomers / TotalCustomers) * 100 AS ChurnRate  
21 FROM GenderChurn;
```

Output:

Gender	ChurnedCustomers	TotalCustomers	ChurnRate
Male	273	1776	15.3716
Female	385	1537	25.0488

Insights:

- Female customers have a significantly higher churn rate (25.05%) compared to male customers (15.37%), indicating that females are more likely to leave the bank.
- The total number of female customers is slightly lower than males, but the higher churn rate suggests possible dissatisfaction among female clients.
- The query uses the most recent year of customer joining (BankDOJ), providing a fresh view of customer churn trends.
- Gender appears to be a notable factor in churn behaviour, which could be related to differences in product usage, service satisfaction, or other variables not captured in this analysis.

**Recommendations:**

- Investigate the causes of the higher churn rate among female customers by conducting surveys or focus groups to understand their pain points and preferences.
- Develop tailored retention strategies aimed at female customers, such as personalized offers, loyalty programs, or enhanced customer service.
- Monitor gender-based churn trends regularly to assess the effectiveness of retention campaigns and adjust strategies as needed.
- Use the insights from this analysis to refine customer segmentation and create more targeted marketing strategies to engage and retain female customers effectively.

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

➤ Query:

```
1 Select
2     e.ExitID,
3     Case
4         When e.ExitID = 0 Then 'Retained'
5     Else 'Exited'
6 End as LoyaltyStatus,
7     Round(Avg(b.CreditScore),0) as AvgCreditScore
8 From exitcustomer e
9 Left Join bank_churn b on e.ExitID = b.ExitID
10 Group By e.ExitID;
```

**Output:**

ExitID	LoyaltyStatus	AvgCreditScore
0	Retained	652
1	Exited	645

**Insights:**

- The average credit score for retained customers is 652, which is slightly higher than the average credit score for exited customers, which is 645.
- While there is not a significant difference in the average credit score between retained and exited customers, it may still indicate a minor correlation between credit score and customer loyalty.
- The ExitID field differentiates between customers who have exited (ExitID = 1) and those who are retained (ExitID = 0), and this classification is helpful for analysing customer loyalty.
- The relatively small gap in average credit scores suggests that factors other

than just the credit score might influence a customer's decision to leave the bank.

#### Recommendations:

- Investigate additional factors that may influence churn beyond credit scores, such as customer satisfaction, product engagement, or service quality.
- Use targeted retention strategies for customers with credit scores near the 645 range, as these customers might be more at risk of exiting.
- Monitor customer behaviour and offer incentives or personalized services to customers who have lower credit scores to encourage retention.
- Consider segmenting customers based on other behavioural metrics in addition to credit scores for a more comprehensive churn analysis.

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

➤ Query:

```

1 with RankedSalaries as (
2   Select
3     g.GenderID as GenderID,
4     g.Gendercategory as Gendercategory,
5     Round(Avg(c.EstimatedSalary),0) as AvgEstimatedSalary,
6     Count(b.ActiveID) as CountOfActiveID,
7     Rank() OVER (Partition by g.GenderID, g.Gendercategory order by Avg(c.EstimatedSalary) desc) as SalaryRank
8   from gender g
9   inner join customerinfo c on c.GenderID = g.GenderID
10  inner join bank_churn b on b.CustomerID = c.CustomerID
11  Inner join activecustomer a on a.ActiveID = b.ActiveID
12  Group by g.GenderID, g.Gendercategory
13 )
14 Select
15   Gendercategory,
16   AvgEstimatedSalary,
17   CountOfActiveID
18 from RankedSalaries
19 where SalaryRank = 1
20 order by AvgEstimatedSalary desc;

```

Output:

Gendercategory	AvgEstimatedSalary	CountOfActiveID
Female	100602	4543
Male	99665	5457

#### Insights:

- Female customers have a slightly higher average estimated salary (100,602) compared to male customers (99,665), indicating a marginal income difference between genders.
- Despite having a higher average salary, females have fewer active accounts (4,543) compared to males (5,457), suggesting that female customers may engage less with the bank's products.
- The use of Rank() over gender categories highlights that both genders have

relatively close estimated salaries, with females leading in this comparison.

- The small difference in average salaries suggests that other factors may be influencing the level of engagement with active accounts

#### Recommendations:

- Investigate why female customers, despite having a higher average salary, have fewer active accounts and create strategies to promote product engagement among them.
- Use targeted marketing strategies to encourage female customers to explore and utilize more bank products, thereby increasing their account activity.
- Consider personalized financial services or benefits tailored to high-income customers, particularly females, to increase retention and engagement.
- Analyse other behavioural metrics beyond salary to understand gender-specific preferences and improve overall product adoption rates across all customers.

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

#### ➤ Query:

```

1  Select
2  Case
3    When CreditScore between 800 and 850 then 'Excellent'
4    When CreditScore between 740 and 799 then 'Very Good'
5    When CreditScore between 670 and 739 then 'Good'
6    When CreditScore between 580 and 669 then 'Fair'
7    When CreditScore between 300 and 579 then 'Poor'
8  End as CreditScoreSegment,
9  Count(*) as TotalCustomers,
10 Sum(Case When ExitID = 1 Then 1 Else 0 End) as ExitedCustomers,
11 Round((Sum(Case When ExitID = 1 Then 1 Else 0 End) * 1.0 / Count(*)) * 100, 2) as ExitRate
12 From bank_churn
13 Group by CreditScoreSegment
14 Order by ExitRate Desc;

```

#### Output:

CreditScoreSegment	TotalCustomers	ExitedCustomers	ExitRate
Poor	2362	520	22.02
Very Good	1224	252	20.59
Fair	3331	685	20.56
Excellent	655	128	19.54
Good	2428	452	18.62

#### Insights:

- Customers in the "Poor" credit score segment (300-579) have the highest exit rate of 22.02%, indicating that these customers are more likely to leave the bank compared to others.
- The "Very Good" (740-799) and "Fair" (580-669) segments also have relatively



high exit rates, at 20.59% and 20.56% respectively, suggesting potential risk in these categories.

- The segment with the lowest exit rate is "Good" (670-739), with 18.62%, followed by "Excellent" (800-850), which has an exit rate of 19.54%.
- Customers with lower credit scores ("Poor" and "Fair") tend to churn more frequently, likely due to financial instability or dissatisfaction with the bank's products.

#### Recommendations:

- Focus retention efforts on customers in the "Poor" and "Fair" segments by offering tailored support, such as financial planning or product recommendations, to help reduce churn.
- Consider offering incentives like lower interest rates or additional benefits to customers in the "Very Good" and "Fair" segments to encourage them to stay with the bank.
- Develop targeted campaigns that focus on improving customer satisfaction and engagement among customers with lower credit scores.
- Monitor customer behaviour and provide proactive outreach to those in higher-risk segments (e.g., "Poor") to address any potential reasons for exit before they leave the bank.

### 8. Find out which geographic region has the highest number of active customers with a tenure greater than 5 years. (SQL)

➤ Query:

```
1 Select
2     g.GeographyLocation,
3     count(b.ActiveID) as NoOfActiveCustomers
4 From geography g
5 inner join customerinfo c on c.GeographyID = g.GeographyID
6 inner join bank_churn b on b.CustomerID = c.CustomerID
7 Where b.ActiveID = 1 and b.Tenure > 5
8 Group By g.GeographyLocation;
```

Output:

GeographyLocation	NoOfActiveCustomers
France	797
Spain	431
Germany	399



**Insights:**

- France has the highest number of active customers with tenures greater than 5 years, totalling 797, followed by Spain with 431, and Germany with 399.
- The significant difference in active customer count between France and the other regions may indicate that customers in France tend to be more loyal or engaged with the bank's services over a longer period.
- Spain and Germany have similar numbers of long-tenured active customers, suggesting these regions may have comparable levels of customer engagement.
- The higher number of active customers in France could be attributed to better customer service, product offerings, or a stronger relationship between the bank and its French clientele.

**Recommendations:**

- Investigate the factors contributing to the higher retention rates in France and apply similar strategies to other regions to boost long-term customer engagement.
- Develop targeted marketing campaigns in Spain and Germany to increase customer retention, focusing on improving engagement with customers who have shorter tenures.
- Offer loyalty programs or incentives in Germany and Spain to encourage long-term customers to remain active and utilize more products.
- Monitor customer satisfaction metrics across all geographic locations to ensure consistent service quality and identify areas for improvement in customer retention.

## 9. What is the impact of having a credit card on customer churn, based on the available data?

➤ **Query:**

```

1 Select
2   Case
3     When CreditID = 1 Then 'Credit Card Holder'
4     Else 'No Credit Card'
5   End as CreditCardStatus,
6   Count(*) as TotalCustomers,
7   Sum(Case when ExitID = 1 Then 1 Else 0 End) as ExitedCustomers,
8   Round((sum(Case when ExitID = 1 Then 1 Else 0 End) * 1.0 / Count(*)) * 100 , 2) as ChurnRate
9 From bank_churn
10 Group by CreditCardStatus;

```

**Output:**

CreditCardStatus	TotalCustomers	ExitedCustomers	ChurnRate
Credit Card Holder	7055	1424	20.18
No Credit Card	2945	613	20.81

**Insights:**

- The churn rate for customers who do not have a credit card (20.81%) is slightly higher than the churn rate for credit card holders (20.18%).
- The difference in churn rates between the two groups is relatively small, indicating that having a credit card does not drastically affect customer churn.
- The number of total customers with credit cards (7,055) is more than double that of those without credit cards (2,945), suggesting that credit card ownership is more common among the customer base.
- Even though the churn rate for non-credit card holders is slightly higher, the similar churn rates suggest that other factors might be driving customer exits beyond credit card ownership.

**Recommendations:**

- Analyse customer satisfaction data to identify key factors influencing churn, such as service quality or product gaps.
- Offer personalized incentives to non-credit card holders to encourage credit card adoption and improve retention.
- Implement targeted retention strategies for both groups, addressing reasons for churn through proactive engagement.
- Regularly monitor churn trends and adjust strategies based on emerging customer behaviour patterns.

**10. For customers who have exited, what is the most common number of products they have used?****➤ Query:**

```
1 Select
2   NumOfProducts,
3   Count(CustomerID) as TotalCustomers
4 From bank_churn
5 Where ExitID = 1
6 Group by NumOfProducts
7 Order by TotalCustomers desc;
```

**Output:**

NumOfProducts	TotalCustomers
1	1409
2	348
3	220
4	60

**Insights:**

- The majority of customers who exited the bank used only one product, accounting for 1,409 customers, which is the highest among all categories.

- A significantly smaller number of exited customers used multiple products, with only 348 customers using two products, 220 using three products, and 60 using four products.
- This indicates that customers who engage with fewer products are more likely to exit, suggesting a potential link between product engagement and retention.
- The number of customers using multiple products decreases significantly as the number of products increases, highlighting that fewer customers diversify their engagement with the bank's offerings.

**Recommendations:**

- Focus on cross-selling and promoting multiple products to customers who use only one product to increase engagement and reduce churn risk.
- Develop personalized marketing strategies to encourage customers to explore more products, particularly targeting those who are at risk of exiting.
- Investigate the reasons behind the low product adoption among customers with only one product to identify barriers and address them effectively.
- Offer incentives or bundling opportunities to encourage customers to adopt additional products and strengthen their relationship with the bank.

**11. Examine the trend of customers joining over time and identify any seasonal patterns (yearly or monthly). Prepare the data through SQL and then visualize it.****➤ Query:**

```
1 SELECT
2     YEAR(BankDOJ) AS Year,
3     COUNT(*) AS NewCustomers
4 FROM customerinfo
5 GROUP BY YEAR(BankDOJ)
6 ORDER BY Year;
```

**Output:**

Year	NewCustomers
2016	1951
2017	2143
2018	2593
2019	3313

**Insights:**

- The data shows a consistent increase in the number of customers joining the bank from 2016 to 2019. This indicates a positive growth trend.
- The year 2019 saw the highest number of new customers (3,313), suggesting a significant boost in marketing efforts or service improvements.

- Each year, the number of new customers has increased at a faster rate, indicating accelerating growth.
- Economic or market changes during this period may have influenced the rising number of new customers.

#### Recommendations:

- Analyse factors contributing to the spike in customer acquisition in 2019 to leverage them further.
- Continue strategies that helped achieve consistent growth, focusing on customer acquisition programs.
- Further breakdown of data by month may reveal seasonal trends or marketing efforts that drive customer sign-ups.
- Use this data to predict future customer trends and adjust business strategies accordingly.

## 12. Analyse the relationship between the number of products and the account balance for customers who have exited.

### ➤ Query:

```

1  Select
2    NumOfProducts,
3    Count(CustomerID) as TotalCustomer,
4    Round(Sum(Balance),2) as TotalBalance
5  From bank_churn
6  Where ExitID = 1
7  Group by NumOfProducts
8  Order by TotalBalance desc;

```

### Output:

NumOfProducts	TotalCustomer	TotalBalance
1	1409	129668607.08
2	348	31407820.29
3	220	18887679.16
4	60	5623988.1

### Insights:

- Customers who used only one product have the highest total balance of 129,668,607.08, which is significantly higher than any other group.
- As the number of products increases, the total balance decreases, with customers using four products having the lowest total balance of 5,623,988.10.
- While fewer customers engage with multiple products (especially 3 and 4), they collectively contribute less to the overall balance than those with fewer products.

- The data suggests that customers with fewer products tend to maintain larger account balances, while those with more products have smaller total balances.

**Recommendations:**

- Investigate why customers with more products tend to have lower balances and address any factors that might be causing reduced financial engagement.
- Focus on encouraging customers with one product to adopt additional products while maintaining their high balances, through personalized offers or bundled services.
- Target customers with low balances and high product usage to identify opportunities for increasing their overall financial activity and product engagement.
- Implement strategies to retain high-balance customers who only use one product, as they represent a significant portion of the total balance despite lower product usage.

**13. Identify any potential outliers in terms of balance among customers who have remained with the bank.****➤ Methods to Identify Outliers:****A. Z-Score Method:**

- Formula:  $Z = (X - \mu) / \sigma$
- Outliers are those data points where  $Z > 3$  or  $Z < -3$  (threshold can be adjusted depending on the dataset).

**B. Interquartile Range (IQR) Method:**

- Formula:  $IQR = Q3 - Q1$
- Outliers are values  $< Q1 - 1.5 \times IQR$  or  $> Q3 + 1.5 \times IQR$ .

**C. Box Plot Visualization:**

- Outliers are represented as points outside the "whiskers" (typically 1.5 times the IQR) in a box plot.

**Applying Outlier Detection:****A. Z-Score Calculation:**

- Identify balances with Z-Scores outside the range  $[-3, 3]$ .

**B. IQR Calculation:**

- Identify balances **below**  $Q1 - 1.5 \times IQR$  or **above**  $Q3 + 1.5 \times IQR$ .

**C. Handling Outliers**

- Remove Outliers:
- Remove data points that are extreme or suspected errors.

**D. Transform Data:**

- Apply transformations (e.g., log transformation) to reduce the impact of extreme values.

**E. Cap/Floor:**

- Set upper and lower limits for values based on a reasonable threshold.

By using these methods, we can effectively identify and handle outliers to ensure accurate data analysis.

**14. How many different tables are given in the dataset, out of these tables which table only consists of categorical variables?**

➤ The dataset consists of **7 tables**. Out of these, the following **5 tables** contain only **categorical variables**:

**1) activecustomer**

- ActiveID (INT)
- ActiveCategory (TEXT)

**2) creditcard**

- CreditID (INT)
- Category (TEXT)

**3) exitcustomer**

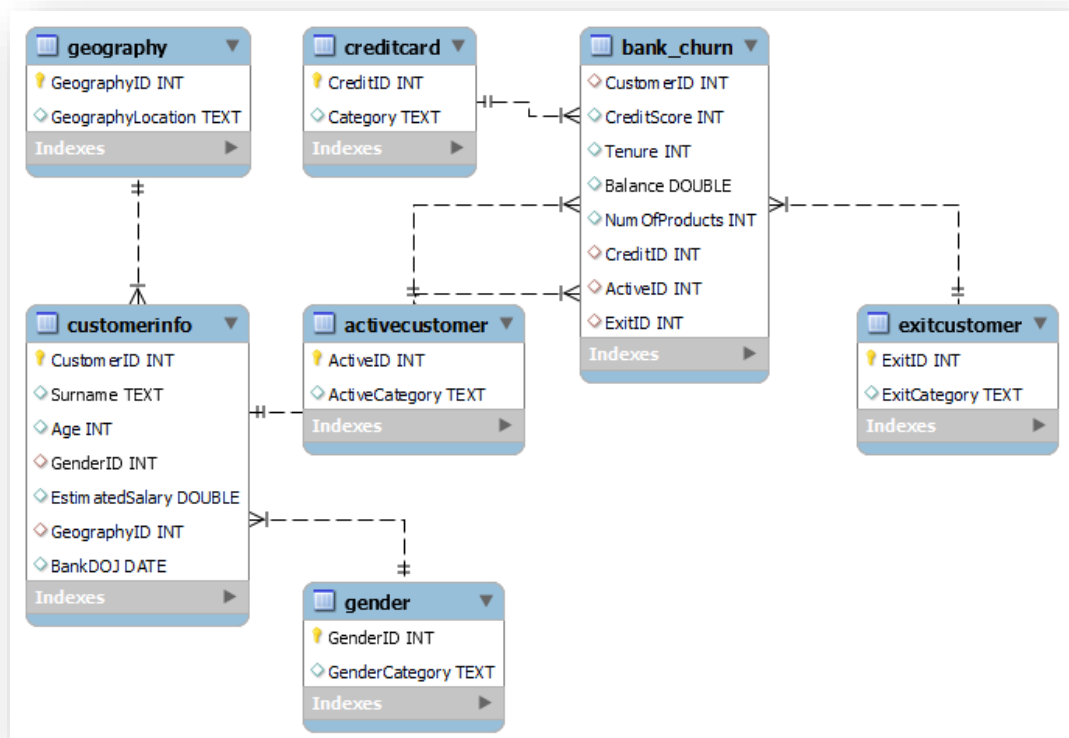
- ExitID (INT)
- ExitCategory (TEXT)

**4) gender**

- GenderID (INT)
- GenderCategory (TEXT)

**5) geography**

- GeographyID (INT)
- GeographyLocation (TEXT)



**15. Using SQL, write a query to find out the gender-wise average income of males and females in each geography id. Also, rank the gender according to the average value. (SQL)**

➤ **Query:**

```
1 With cte1 as (  
2   Select  
3     geo.GeographyLocation,  
4     g.GenderCategory,  
5     Round(Avg(c.EstimatedSalary),2) as AverageIncome  
6   from customerinfo c  
7   inner join Gender g on c.GenderID = g.GenderID  
8   inner join Geography geo on c.GeographyID = geo.GeographyID  
9   Group by geo.GeographyLocation, g.GenderCategory  
10 )  
11 Select  
12   GeographyLocation,  
13   GenderCategory,  
14   AverageIncome,  
15   dense_rank() over (Partition by GenderCategory order by AverageIncome desc) as GenderRank  
16 From cte1  
17 Order by AverageIncome desc;
```

**Output:**

GeographyLocation	GenderCategory	AverageIncome	GenderRank
Germany	Female	102446.42	1
Spain	Female	100734.11	2
France	Male	100174.25	1
Germany	Male	99905.03	2
France	Female	99564.25	3
Spain	Male	98425.69	3

**Insights:**

- The query provides gender-wise average income across different geographies, helping to identify geographic regions where one gender might earn significantly more than the other.
- The ranking shows how male and female incomes compare within each geography, offering insights into income disparities between genders in different regions.
- Regions like Germany and Spain may have a higher average income for females, while other regions could show differences in male income dominance.
- Identifying geographies with large income gaps between genders can highlight potential areas for further investigation into the causes of these disparities.



**Recommendations:**

- Focus on regions with significant income disparity between genders to promote financial equality by introducing more inclusive financial services or policies.
- Analyse the factors contributing to higher average incomes for specific genders in certain geographies and replicate successful strategies in regions with lower averages.
- Tailor financial products based on gender-specific income data to improve customer engagement and satisfaction.
- Monitor geographic trends regularly to assess if interventions aimed at balancing gender income gaps are effective.

**16. 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+).**

➤ Query:

```
1 Select
2   Case
3     When c.Age between 18 and 30 Then '18-30'
4     When c.Age between 31 and 50 Then '30-50'
5     When c.Age > 50 Then '50+'
6   End as AgeBracket,
7   Round(Avg(b.Tenure),2) as AvgTenure
8 From customerinfo c
9 Inner join bank_churn b on c.CustomerID = b.CustomerID
10 Where ExitID = 1
11 Group by AgeBracket
12 Order by AgeBracket;
```

**Output:**

AgeBracket	AvgTenure
18-30	4.78
30-50	4.89
50+	4.83

**Insights:**

- The average tenure for customers who have exited is highest in the 30-50 age bracket, at 4.89 years, compared to 4.78 years for the 18-30 bracket and 4.83 years for the 50+ bracket.
- Customers in the 30-50 age group tend to stay longer before exiting, indicating that this group may have a stronger connection to the bank before deciding to leave.
- The 18-30 age group has the lowest average tenure, which could suggest a quicker churn rate for younger customers.
- The tenure differences across age groups are minimal, indicating a fairly uniform churn behaviour across all age brackets.

**Recommendations:**

- Focus retention efforts on the 18-30 age group to increase their tenure and reduce churn, possibly through products and services tailored to younger customers.
- Investigate why the 30-50 age group, despite having the longest tenure, still exits, and create targeted strategies to retain this segment for even longer.
- For the 50+ age group, offer tailored retirement or long-term financial products to further extend their tenure with the bank.
- Conduct further analysis on the reasons for churn across these age groups to design more effective, age-specific retention strategies.

### 17. Is there any direct correlation between salary and the balance of the customers? And is it different for people who have exited or not?

➤ Query:

```

1  -- (i) Salary and Balances of customer who have exited
2  Select
3      c.CustomerID,
4      c.Surname as CustomerName,
5      c.EstimatedSalary as CustomerSalary,
6      b.Balance as CustomerAccountBalance,
7      e.ExitCategory
8  From customerinfo c
9  inner join bank_churn b on b.CustomerID = c.CustomerID
10 inner join exitcustomer e on e.ExitID = b.ExitID
11 Where b.ExitID = 1
12 Order by CustomerAccountBalance desc;

```

**Output:**

G8		=CORREL(C1:C1001,D1:D1001)					
	A	B	C	D	E	F	G
1	CustomerID	CustomerName	CustomerSalary	CustomerAccountBalance	ExitCategory		
2	15757408	Lo	81054.00	250898.09	Exit		
3	15715622	To Rot	147964.99	238387.56	Exit		
4	15714241	Haddon	101108.85	222267.63	Exit		
5	15586674	Shaw	74176.71	216109.88	Exit		
6	15594408	Chia	75161.25	213146.20	Exit		
7	15671256	Macartney	188574.12	211774.31	Exit		
8	15736420	Macdonald	197297.77	210433.08	Exit	R <sup>2</sup>	0.01926
9	15721658	Fleming	150694.42	209767.31	Exit		
10	15578671	Webb	133267.69	209490.21	Exit		

**Insights:**

- There is a very weak correlation ( $R^2 = 0.01926$ ) between salary and account balance for exited customers.
- Higher salaries do not necessarily result in higher account balances for those who have left the bank.
- Other factors likely influence account balance more than salary for exited customers.

**Recommendations:**

- Analyse active customers to see if the correlation differs from exited customers.
- Focus retention efforts on product engagement rather than salary alone.
- Investigate additional factors like product usage or credit score to boost account balances.

**➤ Query:**

```

1  -- (ii) Salary and Balances of customer who are retained
2  Select
3      c.CustomerID,
4      c.Surname as CustomerName,
5      c.EstimatedSalary as CustomerSalary,
6      b.Balance as CustomerAccountBalance,
7      e.ExitCategory
8  From customerinfo c
9  inner join bank_churn b on b.CustomerID = c.CustomerID
10 inner join exitcustomer e on e.ExitID = b.ExitID
11 Where b.ExitID = 0
12 Order by CustomerAccountBalance desc;

```

**Output:**

G8		✕ ✓ <i>fx</i>		=CORREL(C1:C1001,D1:D1001)			
	A	B	C	D	E	F	G
1	CustomerID	CustomerName	CustomerSalary	CustomerAccountBalance	ExitCategory		
2	15571958	McIntosh	171867.08	221532.8	Retain		
3	15599131	Dilke	128815.33	214346.96	Retain		
4	15769818	Moore	69372.88	212778.2	Retain		
5	15620268	Thomson	115268.86	212696.32	Retain		
6	15780212	Mao	176395.02	212692.97	Retain		
7	15690589	Udinesi	148814.54	212314.03	Retain		
8	15795298	Olisaemeka	102986.15	206868.78	Retain	$R^2$	-0.01857
9	15627971	Coates	16281.94	206663.75	Retain		
10	15664498	Golovanov	156424.4	205962	Retain		

**Insights:**

- The correlation between salary and account balance for retained customers is extremely weak, with an **R<sup>2</sup> value of -0.01857**, indicating no meaningful relationship.
- Similar to exited customers, retained customers also show no clear connection between their salary and their account balance.
- This suggests that both groups, whether retained or exited, are influenced by other factors besides salary when it comes to their account balance.

**Recommendations:**

- Focus on identifying other factors, such as product usage or customer engagement, that influence account balances, as salary does not seem to be a key driver.
- Offer tailored financial services to customers based on their spending habits or account activity, rather than salary alone.
- Conduct further analysis on the impact of additional factors like customer tenure or the number of products they use to better understand balance dynamics.

**Difference:**

- Both exited and retained customers exhibit weak correlations between salary and account balance, with slight variations in R<sup>2</sup> values but no significant trends.
- There is no major distinction in how salary affects account balance between exited and retained customers, indicating similar financial behaviour across both groups.

## 18. Is there any correlation between the salary and the Credit score of customers?

➤ **Query:**

```
1  Select
2      c.CustomerID,
3      c.Surname as CustomerName,
4      c.EstimatedSalary as CustomerSalary,
5      b.CreditScore,
6      b.CreditID
7  From customerinfo c
8  inner join bank_churn b on b.CustomerID = c.CustomerID;
```

**Output:**

I7       =CORREL (C2:C1000,D2:D1000)

	A	B	C	D	E	F	G	H	I
	CustomerID	CustomerName	CustomerSalary	CreditScore	CreditID				
2	15634602	Hargrave	101348.88	619	1				
3	15647311	Hill	112542.58	608	0				
4	15619304	Onio	113931.57	502	1				
5	15701354	Boni	93826.63	699	0				
6	15737888	Mitchell	79084.1	850	1				
7	15574012	Chu	149756.71	645	1			Correlation of customer having CreditCard (R^2) ⚠	0.013583
8	15592531	Bartlett	10062.8	822	1				
9	15656148	Obinna	119346.88	376	1				
10	15792365	He	74940.5	501	0			Correlation of customers not having Credit Card (R^2)	0.013454
11	15592389	H?	71725.73	684	1				
12	15767821	Bearce	80181.12	528	0				

**Insights:**

- There is no strong correlation between salary and credit score, as the R<sup>2</sup> values for both credit card holders and non-holders are very low (around 0.01).
- The lack of correlation indicates that salary is not a key factor in determining a customer's credit score.
- Both credit card holders and non-holders display similar weak correlations between their salary and credit score.

**Recommendations:**

- Focus on other factors, such as financial behaviour and credit history, when analysing or predicting credit scores.
- Encourage better financial education to help customers improve their credit scores irrespective of their income levels.
- Consider designing tailored financial products that address both high and low credit score segments for better customer engagement.

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

➤ **Query:**

```

1 With CreditScoreBuckets as (
2   Select
3     Case
4       When b.CreditScore between 800 and 850 then 'Excellent'
5       When b.CreditScore between 740 and 799 then 'Very Good'
6       When b.CreditScore between 670 and 739 then 'Good'
7       When b.CreditScore between 580 and 669 then 'Fair'
8       When b.CreditScore between 300 and 579 then 'Poor'
9     End as CreditScoreBucket,
10    Count(b.CustomerID) as ChurnedCustomers
11  From bank_churn b
12  Where b.ExitID = 0
13  Group By CreditScoreBucket
14 )
15 Select
16   CreditScoreBucket,
17   ChurnedCustomers,
18   Dense_Rank() Over (Order by ChurnedCustomers Desc) as ChurnRank
19 From CreditScoreBuckets
20 Order By ChurnedCustomers desc;

```

**Output:**

CreditScoreBucket	ChurnedCustomers	ChurnRank
Fair	2646	1
Good	1976	2
Poor	1842	3
Very Good	972	4
Excellent	527	5

**Insights:**

- Customers with a "Fair" credit score have the highest churn rate, accounting for 2,646 customers, which suggests that this group is more vulnerable to churn.
- "Good" and "Poor" credit score buckets rank second and third in churn, indicating that even relatively good credit scores do not prevent churn.
- Customers with "Excellent" credit scores have the lowest churn, with only 527 customers, suggesting stronger loyalty in this group.
- There is a general trend of decreasing churn as credit scores improve, implying a correlation between higher credit scores and customer retention.

**Recommendations:**

- Implement targeted retention strategies for customers in the "Fair" and "Good" credit score buckets to reduce churn.
- Provide personalized offers or improved services for customers with lower credit scores to enhance their satisfaction and reduce churn.
- Focus on maintaining relationships with customers who have "Excellent" credit scores by rewarding loyalty to reinforce retention.
- Regularly monitor credit score trends to identify at-risk segments and preemptively address their needs before they churn.

## 20. According to the age buckets find the number of customers who have a credit card. Also retrieve those buckets that have lesser than average number of credit cards per bucket.

### ➤ Query and Output 1:

```

1  -- (i) Number of customers who have a credit card per each Age Bucket
2  Select
3      Case
4          When c.Age Between 18 And 30 Then '18-30'
5          When c.Age Between 31 And 50 Then '31-50'
6          Else '50+'
7      End As AgeBucket,
8      Count(cc.CreditID) AS CreditCardCustomers
9  From customerinfo c
10 Left Join bank_churn b ON c.CustomerID = b.CustomerID
11 Left Join creditcard cc ON b.CreditID = cc.CreditID
12 Left Join activecustomer a ON b.ActiveID = a.ActiveID
13 Left Join exitcustomer e ON b.ExitID = e.ExitID
14 Left Join gender g ON c.GenderID = g.GenderID
15 Left Join geography geo ON c.GeographyID = geo.GeographyID
16 Where cc.CreditID = 1
17 Group By AgeBucket
18 Order By AgeBucket;

```

AgeBucket	CreditCardCustomers
18-30	1400
31-50	4781
50+	874

## Query and Output 2:

```

1 With AgeBuckets As (
2   Select
3     Case
4       When c.Age Between 18 And 30 Then '18-30'
5       When c.Age Between 31 And 50 Then '31-50'
6       Else '50+'
7     End As AgeBucket,
8     Count(cc.CreditID) AS CreditCardCustomers
9   From customerinfo c
10  Left Join bank_churn b ON c.CustomerID = b.CustomerID
11  Left Join creditcard cc ON b.CreditID = cc.CreditID
12  Left Join activecustomer a ON b.ActiveID = a.ActiveID
13  Left Join exitcustomer e ON b.ExitID = e.ExitID
14  Left Join gender g ON c.GenderID = g.GenderID
15  Left Join geography geo ON c.GeographyID = geo.GeographyID
16  Where cc.CreditID = 1
17  Group By AgeBucket
18  Order By AgeBucket
19 ),
20 AvgCreditCards As (
21   Select Avg(CreditCardCustomers) As AvgCreditCardsPerBucket
22   From AgeBuckets
23 )
24 Select
25   ab.AgeBucket,
26   ab.CreditCardCustomers
27 From AgeBuckets ab
28 Cross Join AvgCreditCards avg_cc
29 Where ab.CreditCardCustomers < avg_cc.AvgCreditCardsPerBucket
30 Order By ab.CreditCardCustomers Asc;

```

AgeBucket	CreditCardCustomers
50+	874
18-30	1400

### Insights:

- The highest number of credit card holders are in the 31-50 age group, significantly outnumbering the other age buckets.
- The 18-30 and 50+ age groups have fewer credit card holders than the average per age bucket.
- Age distribution suggests that the middle age group (31-50) is more engaged with credit card usage.
- Credit card ownership in younger and older demographics is below average and may need targeted marketing.



**Recommendations:**

- Implement targeted marketing campaigns for younger (18-30) and older (50+) age groups to increase credit card adoption.
- Consider offering age-specific rewards or promotions to drive higher credit card engagement.
- Monitor age-based credit card usage trends to identify growth opportunities and further refine product offerings.
- Explore financial literacy programs for younger demographics to increase awareness and usage of credit products.

## 21. Rank the Locations as per the number of people who have churned the bank and average balance of the customers.

### ➤ Query:

```

1 With LocationChurnData As (
2   Select
3     geo.GeographyID,
4     geo.GeographyLocation,
5     Count(b.CustomerID) As ChurnedCustomers,
6     Round(Avg(b.Balance),2) As AvgBalance
7   From bank_churn b
8   Join customerinfo c on c.CustomerID = b.CustomerID
9   Join geography geo on geo.GeographyID = c.GeographyID
10  Where ExitID = 1
11  Group By geo.GeographyID, geo.GeographyLocation
12 )
13 Select
14   GeographyID,
15   GeographyLocation,
16   ChurnedCustomers,
17   Dense_Rank() Over (Order By ChurnedCustomers) as ChurnRank,
18   AvgBalance,
19   Dense_Rank() Over (Order by AvgBalance Desc) as BalanceRank
20 From LocationChurnData
21 Order By ChurnedCustomers Desc, AvgBalance Desc;

```

**Output:**

GeographyID	GeographyLocation	ChurnedCustomers	ChurnRank	AvgBalance	BalanceRank
3	Germany	814	3	120361.08	1
1	France	810	2	71192.8	3
2	Spain	413	1	72513.35	2

**Insights:**

- Germany has the highest average balance among churned customers, though it ranks lowest in churn count.
- Spain has the least number of churned customers but ranks second in average balance, indicating relatively more stable customers.
- France has the highest churn count but has the lowest average balance among churned customers.

- Churn behaviour and average balance seem inversely related, especially between Germany and France.

**Recommendations:**

- Focus on retaining high-balance customers in Germany by enhancing customer satisfaction and product offerings.
- Investigate reasons for higher churn in France despite lower average balances, potentially identifying underserved customer segments.
- Strengthen customer engagement in Spain to reduce churn further, as it already has lower churn rates.
- Implement a balanced retention strategy considering both churn volume and average customer balance to prioritize efforts.
- 

**22. 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”.**

➤ **Query:**

```
1 Select
2     CustomerID,
3     Surname,
4     Concat(CustomerID, '_', Surname) As CustomerID_Surname
5 From customerinfo;
```

**Output:**

CustomerID	Surname	CustomerID_Surname
15565701	Ferri	15565701_Ferri
15565706	Akobundu	15565706_Akobundu
15565714	Cattaneo	15565714_Cattaneo
15565779	Kent	15565779_Kent
15565796	Docherty	15565796_Docherty
15565806	Toosey	15565806_Toosey
15565878	Bates	15565878_Bates
15565879	Riley	15565879_Riley
15565891	Dipietro	15565891_Dipietro

**23. Without using “Join”, can we get the “ExitCategory” from ExitCustomers table to Bank\_Churn table? If yes do this using SQL.**

➤ **Query:**

```

1  Select
2    b.CustomerID,
3    b.CreditScore,
4    b.Balance,
5    b.ExitID,
6    (
7      Select e.ExitCategory
8      From exitcustomer e
9      Where e.ExitID = b.ExitID
10   ) as ExitCategory
11 From bank_churn b
12 Order By b.Balance desc;

```

**Output:**

CustomerID	CreditScore	Balance	ExitID	ExitCategory
15757408	655	250898.09	1	Exit
15715622	583	238387.56	1	Exit
15714241	749	222267.63	1	Exit
15571958	489	221532.8	0	Retain
15586674	663	216109.88	1	Exit
15599131	650	214346.96	0	Retain
15594408	584	213146.2	1	Exit
15769818	850	212778.2	0	Retain
15620268	634	212696.32	0	Retain
15780212	592	212692.97	0	Retain
15690589	541	212314.03	0	Retain

**24. Were there any missing values in the data, using which tool did you replace them and what are the ways to handle them?**

➤ **Query and Output 1:**

```

1  -- (i) Checking if bank_churn table has null values
2  SELECT
3    SUM(CASE WHEN CustomerID IS NULL THEN 1 ELSE 0 END) AS Missing_CustomerID,
4    SUM(CASE WHEN CreditScore IS NULL THEN 1 ELSE 0 END) AS Missing_CreditScore,
5    SUM(CASE WHEN Tenure IS NULL THEN 1 ELSE 0 END) AS Missing_Tenure,
6    SUM(CASE WHEN Balance IS NULL THEN 1 ELSE 0 END) AS Missing_Balance,
7    SUM(CASE WHEN NumOfProducts IS NULL THEN 1 ELSE 0 END) AS Missing_NumOfProducts,
8    SUM(CASE WHEN CreditID IS NULL THEN 1 ELSE 0 END) AS Missing_CreditID,
9    SUM(CASE WHEN ActiveID IS NULL THEN 1 ELSE 0 END) AS Missing_ActiveID,
10   SUM(CASE WHEN ExitID IS NULL THEN 1 ELSE 0 END) AS Missing_ExitID
11 From bank_churn;

```

Missing_CustomerID	Missing_CreditScore	Missing_Tenure	Missing_Balance	Missing_NumOfProducts	Missing_CreditID	Missing_ActiveID	Missing_ExitID
0	0	0	0	0	0	0	0

### Query and Output 2:

```

1  -- (ii) Checking if customerinfo table has null values
2  SELECT
3      SUM(CASE WHEN CustomerID IS NULL THEN 1 ELSE 0 END) AS Missing_CustomerID,
4      SUM(CASE WHEN Surname IS NULL THEN 1 ELSE 0 END) AS Missing_Surname,
5      SUM(CASE WHEN Age IS NULL THEN 1 ELSE 0 END) AS Missing_Age,
6      SUM(CASE WHEN GenderID IS NULL THEN 1 ELSE 0 END) AS Missing_GenderID,
7      SUM(CASE WHEN EstimatedSalary IS NULL THEN 1 ELSE 0 END) AS Missing_EstimatedSalary,
8      SUM(CASE WHEN GeographyID IS NULL THEN 1 ELSE 0 END) AS Missing_GeographyID,
9      SUM(CASE WHEN BankDOJ IS NULL THEN 1 ELSE 0 END) AS Missing_BankDOJ
10 FROM customerinfo;

```

Missing_CustomerID	Missing_Surname	Missing_Age	Missing_GenderID	Missing_EstimatedSalary	Missing_GeographyID	Missing_BankDOJ
0	0	0	0	0	0	0

### Insights and Recommendations:

Currently, there are no missing values in the dataset, as seen from the query results. However, if there were missing values, here are the steps we could take:

- **Imputation:** We can use techniques such as mean, median, or mode imputation to fill missing values for numerical data. For categorical variables, the most frequent category could be used.
- **Deletion:** In cases where a few rows have missing data, we can choose to delete those records if the impact on overall analysis is minimal, ensuring we don't lose significant information.
- **Interpolation or Forward Filling:** For time-series data, we can interpolate or use forward/backward filling techniques to estimate missing values based on neighbouring records.

25. 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".

➤ Query:

```
1 Select
2     c.CustomerID,
3     c.Surname as LastName,
4     a.ActiveCategory
5 From customerinfo c
6 inner join bank_churn b on c.CustomerID = b.CustomerID
7 inner join activecustomer a on a.ActiveID = b.ActiveID
8 Where c.Surname like '%on'
9 Order By c.Surname;
```

Output:

CustomerID	LastName	ActiveCategory
15641110	Abron	Inactive Member
15589210	Adamson	Inactive Member
15573599	Adamson	Active Member
15644400	Anderson	Inactive Member
15610403	Anderson	Inactive Member
15661526	Anderson	Inactive Member
15687492	Anderson	Inactive Member
15748327	Anderson	Inactive Member
15594084	Anderson	Inactive Member
15775116	Anderson	Inactive Member
15657228	Anderson	Active Member

26. Can you observe any data discrepancy in the Customer's data? As a hint it's present in the IsActiveMember and Exited columns. One more point to consider is that the data in the Exited Column is absolutely correct and accurate.

➤ Query:

```
1 select * from bank_churn
2 where ActiveID = 0 and ExitID = 0;
```

**Output:**

CustomerID	CreditScore	Tenure	Balance	NumOfProducts	CreditID	ActiveID	ExitID
15634602	619	7	0	1	1	1	1
15602280	829	5	112045.67	1	1	1	1
15771573	637	7	137843.8	1	1	1	1
15757535	647	4	0	3	1	1	1
15804919	670	3	0	1	1	1	1
15613854	622	4	107073.27	2	1	1	1
15609618	721	3	154475.54	2	0	1	1
15671137	549	4	0	1	0	1	1
15640905	579	7	129490.36	2	0	1	1
15719377	804	3	0	1	1	1	1

**Insights:**

- The dataset shows customers with both ActiveID = 1 (indicating they are active) and ExitID = 1 (indicating they have exited). This is contradictory, as a customer cannot logically be both active and exited.
- The discrepancy likely stems from incorrect data entry or an issue during data migration, where the active status was not updated correctly when the customer exited the bank.
- Several customers have a balance of 0, yet they are marked as active or have exited with a product still assigned. This might indicate poor data management or delayed balance updates after account closure.
- Some customers have multiple products (NumOfProducts > 1) but still show a zero balance. This could suggest inactive products that were not properly removed when the customer exited.

**Recommendations:**

- Implement automated scripts to check for logical inconsistencies, such as a customer being both active and exited simultaneously.
- Ensure proper workflows and checks when updating customer records, especially when they exit, to ensure the ActiveID is set to 0 upon exiting.
- Cross-check data entries related to account balances and product ownership for active and exited customers to avoid such discrepancies in the future.
- Investigate the process used to manage account statuses during customer exits. Potential causes for these discrepancies may include lagging updates, incorrect triggers, or overlapping system changes.

These steps should help to maintain consistency and ensure data accuracy in your system.

===== **END OF OBJECTIVE QUESTIONS** =====