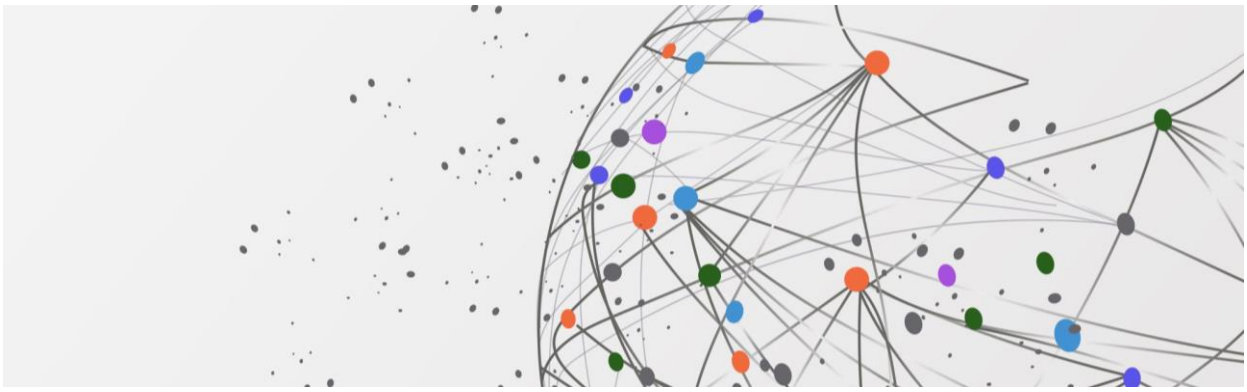


IE 6700: DATA MANAGEMENT FOR ANALYTICS

USE CASE: MILESTONE #3

Enhancing customer understanding and Service Improvement for a Banking System

---

GROUP NO. 12

GROUP MEMBERS:

MALLIKA GAIKWAD

SHARVARI PRAVIN DESHPANDE

PROBLEM STATEMENT:

Objective:

The primary objective of this project is to empower the banking system with precise insights into customer behavior, preferences, and risk profiles. By utilizing advanced analytical techniques, the bank aims to refine its understanding of clients and formulate specific strategies for service improvement. The bank seeks a robust and well-structured approach to ensure convincing results.

IDENTIFYING AND IMPLEMENTING SQL QUERIES:

1.Identifying the district with the highest average balance

SQL QUERY:

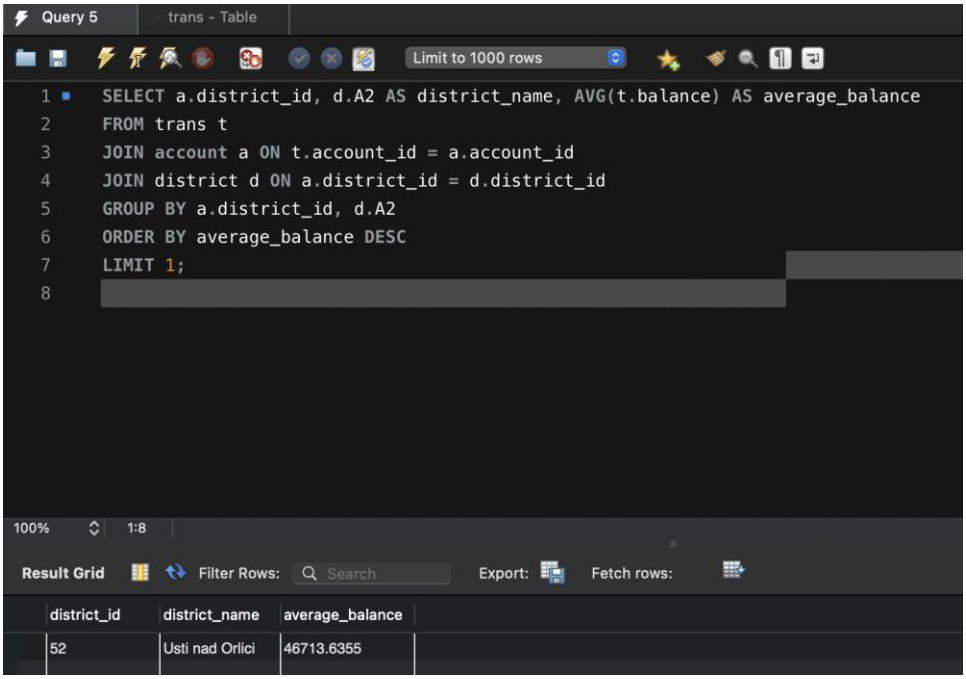
```
SELECT a.district_id, d.A2 AS district_name, AVG(t.balance) AS average_balance

FROM trans t

JOIN account a ON t.account_id = a.account_id
```

JOIN district d ON a.district_id = d.district_id

GROUP BY a.district_id, d.A2

ORDER BY average_balance DESC

LIMIT 1;



The query aims to find the district with the highest average balance based on transactions.

It achieves this by joining the **trans** table with the **account** table using the **account_id** column, and then further joining with the **district** table using the **district_id** column from the **account** table.

Grouping by **district_id** and **A2** ensures that the average balance is calculated for each district.

Ordering by **average_balance** in descending order ensures that the district with the highest average balance appears at the top.

The **LIMIT 1** ensures that only the top row (district with the highest average balance) is returned in the result set.

The calculation of the average balance per district based on transactions could provide valuable insights into the financial behavior and characteristics of different regions.

- **Regional Financial Analysis:**
    - Understanding the average balance per district allows to perform a financial analysis on a regional level. This analysis could help identify areas with higher or

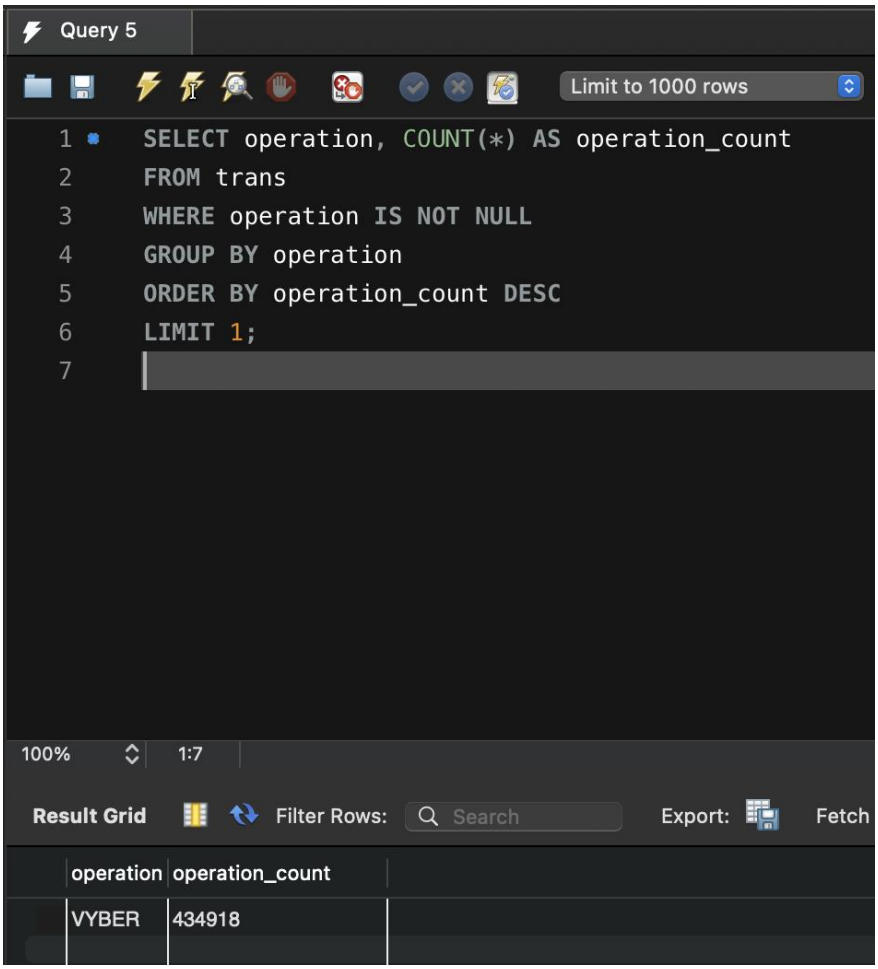
lower average balances, which might be indicative of economic disparities or varying financial behaviors among districts.

- **Risk Assessment:**
  - Higher average balances might indicate more stable financial situations in certain districts, while lower average balances might suggest potential financial challenges. This information could be useful for risk assessment and decision-making in financial services.
- **Business Strategy:**
  - Businesses, especially those in the financial sector, can use this information to tailor their services and marketing strategies to better suit the financial needs and behaviors of different regions.
- **Resource Allocation:**
  - Government agencies or organizations involved in resource allocation may find it beneficial to understand the financial landscape of different districts. This information can help in directing resources, planning, and policy making.

2.Determine the most common operation type in transactions

SQL QUERY:

```
SELECT operation, COUNT(*) AS operation_count

FROM trans

WHERE operation IS NOT NULL

GROUP BY operation

ORDER BY operation_count DESC

LIMIT 1;
```



The query aims to identify the most common operation type in financial transactions.

Filtering for **operation IS NOT NULL** ensures that only transactions with a specified operation type are considered.

Grouping by **operation** allows for the count of occurrences of each unique operation type.

Ordering the result set in descending order of **operation_count** ensures that the most common operation type appears first.

The **LIMIT 1** ensures that only the top row (most common operation type) is returned in the result set.

Analyzing the most common operation types in financial transactions can provide valuable insights for various purposes.

- **Understanding Transaction Patterns:**
    - Identifying the most common operation types helps in understanding the dominant patterns of financial transactions. This insight is crucial for financial

institutions and businesses to adapt their services to meet the prevalent needs of customers.

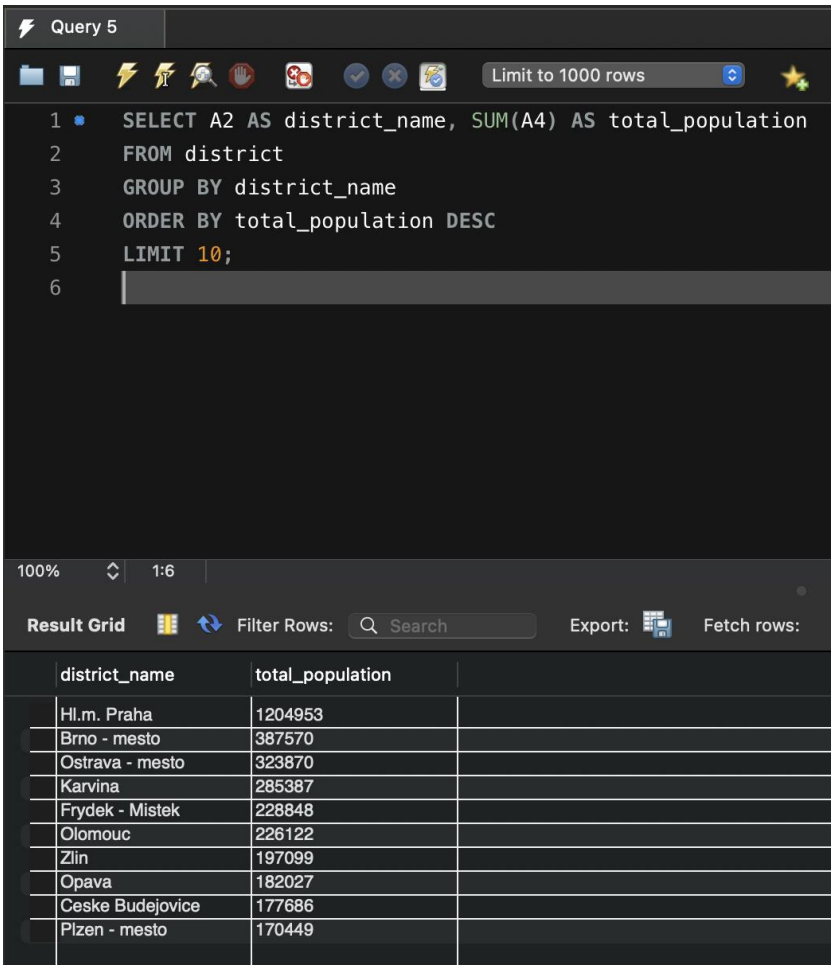
- **Risk Management:**
  - Certain types of financial operations may be associated with higher or lower levels of risk. Analyzing the frequency of different operation types can assist in assessing and managing risks more effectively.
- **Service Optimization:**
  - Financial institutions may use this analysis to optimize their services. For example, if a specific operation type is significantly more common, the institution may choose to streamline processes or introduce targeted services related to that operation.
- **Marketing and Product Development:**
  - Identifying the most common operation types can guide marketing efforts and product development. Financial institutions can tailor their marketing strategies and introduce new products or features that align with the prevailing transaction patterns.

---

3. List the top 10 districts with the highest total population

SQL QUERY:

```
SELECT A2 AS district_name, SUM(A4) AS total_population

FROM district

GROUP BY district_name

ORDER BY total_population DESC

LIMIT 10;
```



The query aims to identify and list the top 10 districts with the highest total populations based on the data in the **district** table.

The **SUM(A4)** calculates the total population for each district by summing up the values in the **A4** column, which presumably represents the population.

The **GROUP BY district_name** ensures that the total population is calculated for each distinct district.

The **ORDER BY total_population DESC** sorts the result set in descending order of total population, so the districts with the highest populations appear first.

The **LIMIT 10** restricts the output to only the top 10 districts, providing a concise list of the most populous districts.

Analyzing the total population by district can be done for various reasons, and the insights gained from such an analysis can inform decision-making processes in areas such as urban

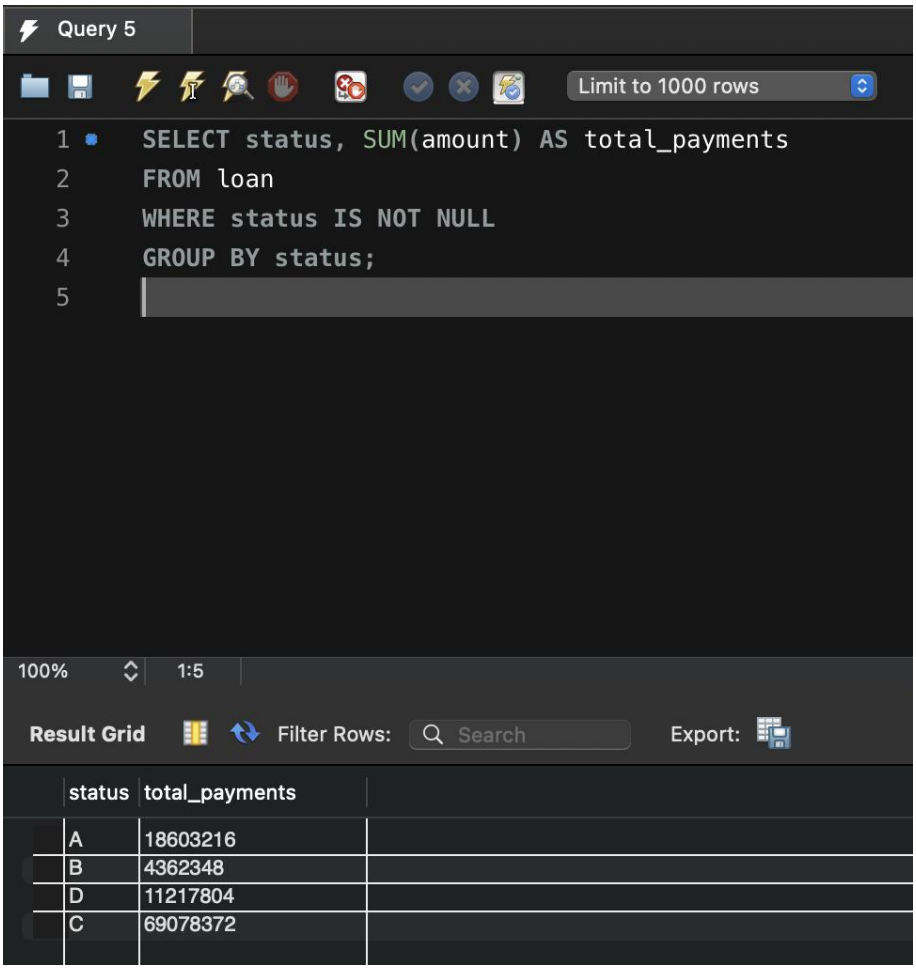
planning, resource allocation, and policy development. Here are some common reasons why this type of analysis might be performed:

- **Resource Allocation:**
  - Governments and organizations can use population data to allocate resources effectively. Districts with higher populations may require more infrastructure, services, and public resources.
- **Urban Planning:**
  - Urban planners use population data to design and optimize city layouts. Understanding the distribution of population across districts helps in planning for housing, transportation, and public spaces.
- **Infrastructure Development:**
  - The total population per district can guide decisions related to the development of essential infrastructure such as schools, hospitals, public transportation, and utilities.
- **Public Services Provision:**
  - Governments can use population data to ensure that public services such as healthcare, education, and emergency services are adequately provided to meet the needs of each district.

4.Calculate the total amount of payments for each loan status

SQL QUERY:

```
SELECT status, SUM(amount) AS total_payments

FROM loan

WHERE status IS NOT NULL

GROUP BY status;
```



The query aims to calculate the total payments for loans based on their status.

The **SUM(amount)** function calculates the total amount paid for each unique loan status.

The **GROUP BY status** ensures that the calculation is performed separately for each distinct loan status.

The **WHERE status IS NOT NULL** ensures that only loans with a specified status are considered in the calculation.

Analyzing the total payments by loan status can provide valuable insights for various purposes

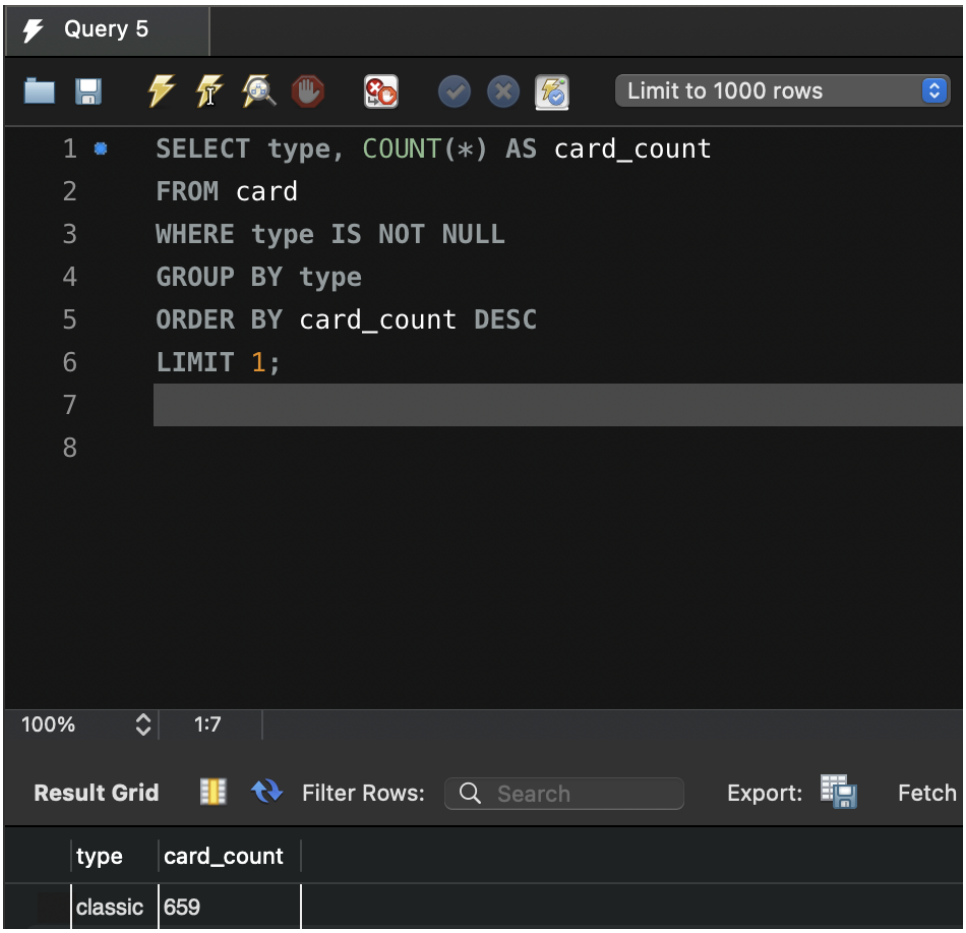
- **Loan Portfolio Performance:**
    - Assessing total payments by loan status helps financial institutions evaluate the performance of different loan categories. It provides a snapshot of how well loans in each status are being managed and repaid.
- **Risk Management:**
    - Understanding the distribution of payments across loan statuses assists in risk management. Financial institutions can identify trends, such as higher default rates in certain statuses, and take proactive measures to mitigate risks.

- **Decision-Making:**
    - Decision-makers can use this analysis to make informed decisions about loan approval processes, underwriting criteria, and overall lending strategies. It helps in optimizing the loan portfolio based on the historical performance of different loan statuses.
- **Financial Reporting:**
    - Financial institutions often report on the performance of their loan portfolios to regulators, investors, and internal stakeholders. Analyzing total payments by loan status contributes to comprehensive financial reporting.

5.Identify the most common card type among clients

SQL QUERY:

```
SELECT type, COUNT(*) AS card_count

FROM card

WHERE type IS NOT NULL

GROUP BY type

ORDER BY card_count DESC

LIMIT 1;
```



The query aims to identify and retrieve the most common card type based on the data in the **card** table.

The **COUNT(*)** function calculates the count of occurrences for each unique card type.

The **GROUP BY type** ensures that the count is calculated separately for each distinct card type.

The **WHERE type IS NOT NULL** ensures that only cards with a specified type are considered in the count.

The **ORDER BY card_count DESC** sorts the result set in descending order of the card count, so the card type with the highest count appears first.

The **LIMIT 1** ensures that only the top row (most common card type) is returned in the result set.

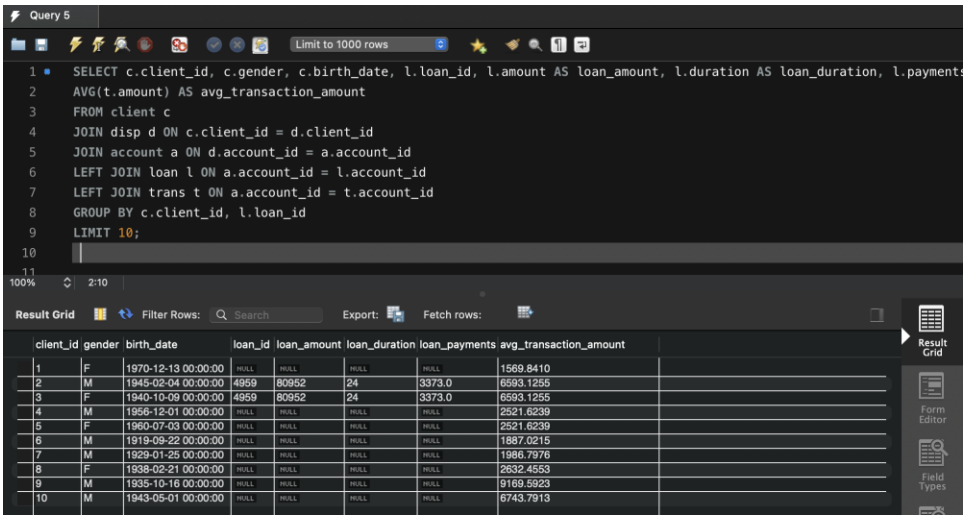
Analyzing the distribution of card types and identifying the most common card type is done for several reasons in the context of a financial institution or business that issues credit or debit cards. Here are some common reasons for performing such an analysis:

- **Product Performance Assessment:**
    - Understanding the popularity of different card types allows financial institutions to assess the performance of their card products. It helps in identifying which types of cards are more widely adopted by customers.
- **Marketing Strategy:**
    - Knowing the most common card type helps in developing targeted marketing strategies. Financial institutions can tailor their promotional efforts to highlight the features and benefits of the most popular card types.
- **Customer Segmentation:**
    - Analyzing card types assists in customer segmentation. Financial institutions can better understand the preferences and needs of different customer segments based on their choice of card type.
- **Product Development:**
    - Analysis can inform product development decisions. If a particular card type is significantly more popular, there may be opportunities to introduce variations or additional features to meet customer demands.

6.Retrieve Loan information, client demographics and transaction amounts

SQL QUERY:

```
SELECT c.client_id, c.gender, c.birth_date, l.loan_id, l.amount AS loan_amount, l.duration AS loan_duration, l.payments AS loan_payments,

AVG(t.amount) AS avg_transaction_amount

FROM client c

JOIN disp d ON c.client_id = d.client_id

JOIN account a ON d.account_id = a.account_id

LEFT JOIN loan l ON a.account_id = l.account_id

LEFT JOIN transaction t ON a.account_id = t.account_id

GROUP BY c.client_id, l.loan_id

LIMIT 10;
```



The query aims to provide information about clients, their loans, and the average transaction amount for each loan.

The use of LEFT JOINs ensures that clients without loans or transactions are still included in the result.

Grouping by **client_id** and **loan_id** allows the calculation of the average transaction amount for each loan separately.

The **LIMIT 10** clause restricts the result set to the first 10 rows, which can be useful for previewing or analyzing a small portion of the data.

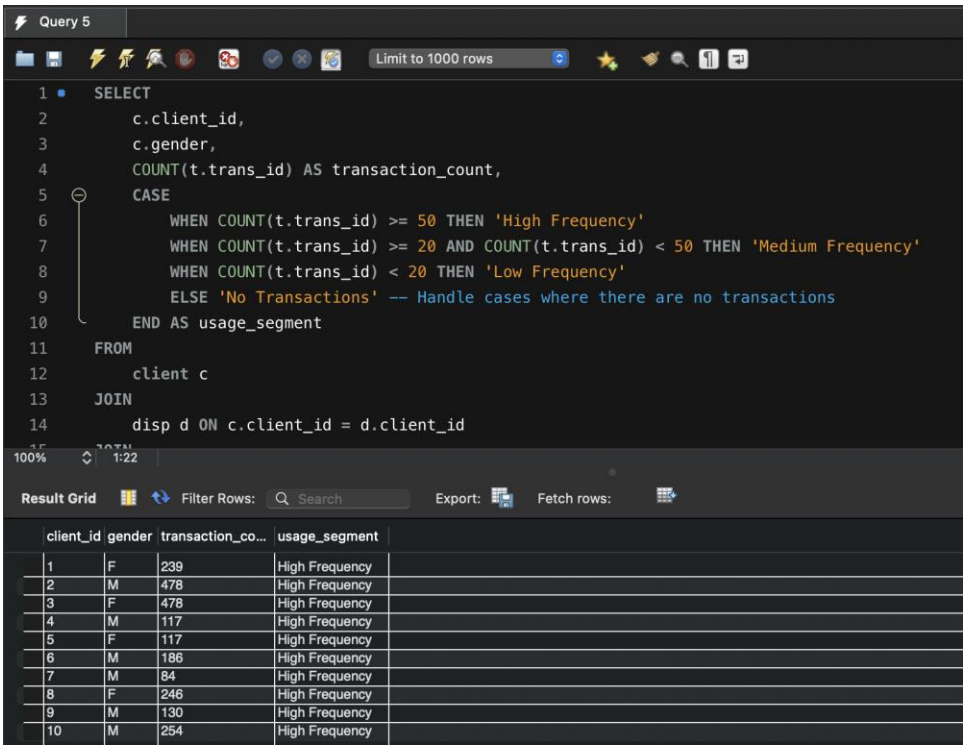
- **Client and Loan Analysis:**
  - The query combines information about clients and loans, allowing for a detailed analysis of the relationships between clients and the loans they have taken.
- **Loan Characteristics:**
  - By selecting loan-related information such as loan amount, duration, and payments, the query facilitates an analysis of the characteristics of each loan. This can be crucial for understanding the distribution of loan amounts and durations among clients.
- **Transaction Behavior:**
  - The inclusion of the average transaction amount per loan provides insights into the financial behavior of clients who have taken loans. Analyzing transaction data helps understand how clients with loans interact with their accounts.
- **Identification of Patterns:**
  - Grouping by **client_id** and **loan_id** enables the identification of patterns and trends in client behavior concerning loans. It allows for the recognition of commonalities or variations in transaction patterns among clients with different loans.

Analytical based Queries

7.Credit Card Usage Patterns: Can we identify segments based on card usage behaviors or transaction frequencies?

SQL QUERY:

```
SELECT c.client_id, c.gender,
COUNT(t.trans_id) AS transaction_count,
CASE
    WHEN COUNT(t.trans_id) >= 50 THEN 'High Frequency'
    WHEN COUNT(t.trans_id) >= 20 AND COUNT(t.trans_id) < 50 THEN 'Medium Frequency'
    WHEN COUNT(t.trans_id) < 20 THEN 'Low Frequency'
    ELSE 'No Transactions' -- Handle cases where there are no transactions
  END AS usage_segment
FROM client c
JOIN disp d ON c.client_id = d.client_id
JOIN account a ON d.account_id = a.account_id
LEFT JOIN trans t ON a.account_id = t.account_id
GROUP BY c.client_id, c.gender
LIMIT 10; -- Limit the output to 10 rows
```



The query joins the **client**, **disp**, **account**, and **trans** tables to get relevant information about clients, cardholders, accounts, and transactions.

It uses the **COUNT** function to calculate the number of transactions for each client.

The **CASE** statement categorizes clients into segments based on their transaction frequencies.

The result includes **client_id**, **gender**, **transaction_count**, and **usage_segment**.

This type of analysis is done for several reasons, often related to understanding and optimizing business strategies, marketing efforts, and customer relationships. Here are some reasons why this type of query might be performed:

- **Segmentation for Targeted Marketing:**
  - By categorizing clients into different frequency segments, businesses can tailor their marketing strategies to better meet the needs and preferences of each segment. For example, high-frequency users might be targeted with premium offers or loyalty programs.
- **Customer Behavior Understanding:**
  - Analyzing transaction frequencies provides insights into how clients interact with their credit cards. Understanding patterns in customer behavior helps businesses adapt their services and products to better align with customer preferences.
- **Risk Assessment:**

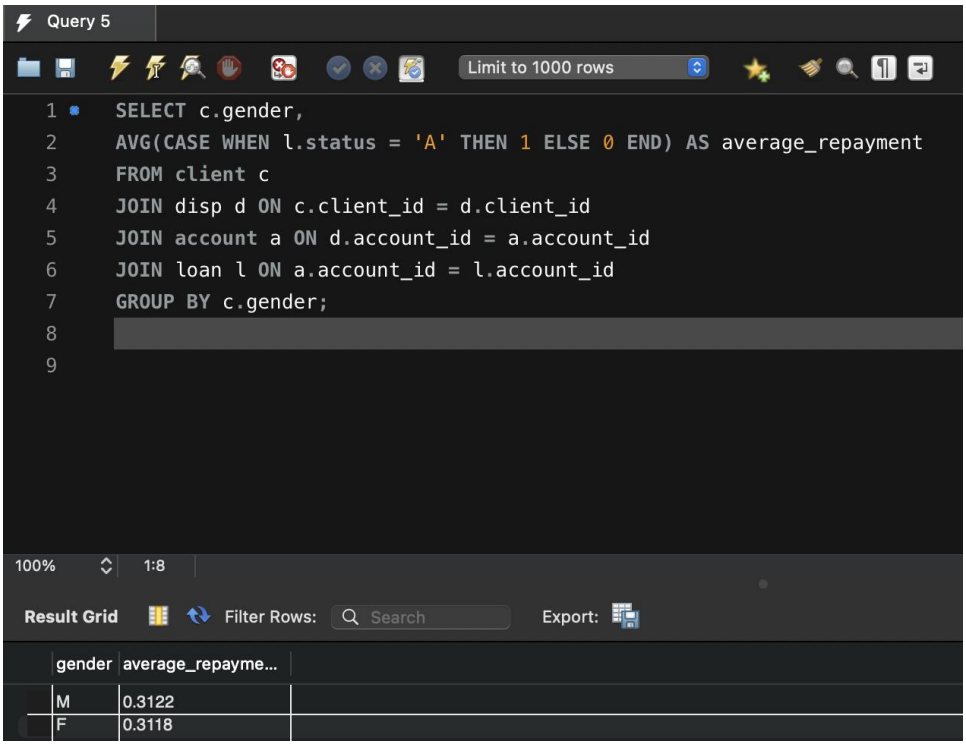
    - Transaction frequencies can be indicative of potential risks. Unusual or unexpected transaction patterns may signal fraud or irregular behavior, and such analysis can help in detecting and preventing fraudulent activities.
- **Product Development and Offerings:**
    - Businesses can use transaction data to refine existing products or develop new ones. For instance, if a segment of clients exhibits high transaction frequency, there may be an opportunity to create specialized offerings or benefits for this group.
- **Customer Retention:**
    - Knowing the transaction behavior of clients allows businesses to implement retention strategies. For instance, identifying clients with decreasing transaction frequency might prompt the development of targeted campaigns to retain their business.
- **Operational Efficiency:**
    - Understanding credit card usage patterns can help financial institutions optimize their operations. For example, it may inform decisions related to resource allocation, fraud detection systems, or customer support strategies.
- **Business Intelligence:**
    - Overall, the analysis provides valuable business intelligence. It helps stakeholders gain a comprehensive view of their customer base and make data-driven decisions to enhance customer satisfaction, loyalty, and overall business performance.

---

8. Loan Performance by Demographics: Are there patterns in loan repayment based on client demographics?

SQL QUERY:

```
SELECT c.gender,

AVG(CASE WHEN l.status = 'A' THEN 1 ELSE 0 END) AS average_repayment

FROM client c

JOIN disp d ON c.client_id = d.client_id

JOIN account a ON d.account_id = a.account_id

JOIN loan l ON a.account_id = l.account_id

GROUP BY c.gender;
```



We're selecting the **gender** column from the **client** table to represent the demographic information.

The **AVG** function is used with a **CASE** statement to calculate the average repayment status. It assigns a value of 1 for successful repayments (status 'A') and 0 for others.

The query joins the **client**, **disp**, **account**, and **loan** tables to get information about clients, their accounts, and loans.

The **GROUP BY** clause groups the results by gender.

Analyzing loan performance by demographics is done for several important reasons, each contributing to a better understanding of how different groups of clients manage their loans. Here are some key reasons why this type of analysis is valuable:

- **Risk Assessment and Management:**
  - By examining loan repayment patterns across different demographics, financial institutions can assess the risk associated with various client groups. Certain demographic factors may be correlated with higher or lower default rates, helping institutions adjust lending criteria and manage risk more effectively.
- **Tailored Product Offerings:**
  - Understanding how different demographic groups perform in terms of loan repayment allows financial institutions to tailor their product offerings. For example, if a particular demographic segment consistently demonstrates good

repayment behavior, there may be opportunities to offer specialized loan products or benefits to that group.

- **Regulatory Compliance:**
  - Financial institutions are often required to adhere to regulations related to fair lending practices. Analyzing loan performance by demographics helps ensure compliance with these regulations and helps identify and rectify any disparities in lending outcomes among different demographic groups.
- **Marketing Strategies:**
  - Insights into loan performance by demographics can inform marketing strategies. Financial institutions can target advertising and promotional efforts toward demographic groups that have historically shown good repayment behavior, attracting more reliable borrowers.
- **Customer Relationship Management:**
  - Knowing how different demographics perform in terms of loan repayment enables better customer relationship management. Financial institutions can proactively engage with clients who may be facing difficulties or offer incentives to those with good repayment histories.
- **Operational Efficiency:**
  - Analysis of loan performance by demographics can contribute to operational efficiency. It helps financial institutions allocate resources effectively, adjust underwriting criteria, and optimize collection strategies based on the characteristics of different client groups.
- **Product Innovation:**
  - The insights gained from this analysis can contribute to the development of new financial products. Understanding the preferences and behaviors of different demographic segments can guide the creation of innovative and targeted loan products that better meet the needs of specific customer groups.
- **Credit Scoring and Underwriting:**
  - Demographic data may be used as part of the credit scoring and underwriting process. Analyzing loan performance by demographics can inform the development or refinement of credit scoring models, leading to more accurate risk assessments.
- **Ethical and Social Responsibility:**
  - Financial institutions have an ethical responsibility to ensure fairness in lending practices. Analyzing loan performance by demographics helps identify and address any potential biases or disparities, contributing to social responsibility and ethical lending practices.

---

9. Loan Portfolio Analysis: What is the distribution of loan amounts, interest rates, and terms across different districts or customer segments?

SQL QUERY:

```
SELECT d.district_id,d.A2 AS district_name,
```

AVG(l.amount) AS average_loan_amount,

AVG(l.duration) AS average_loan_term

FROM district d

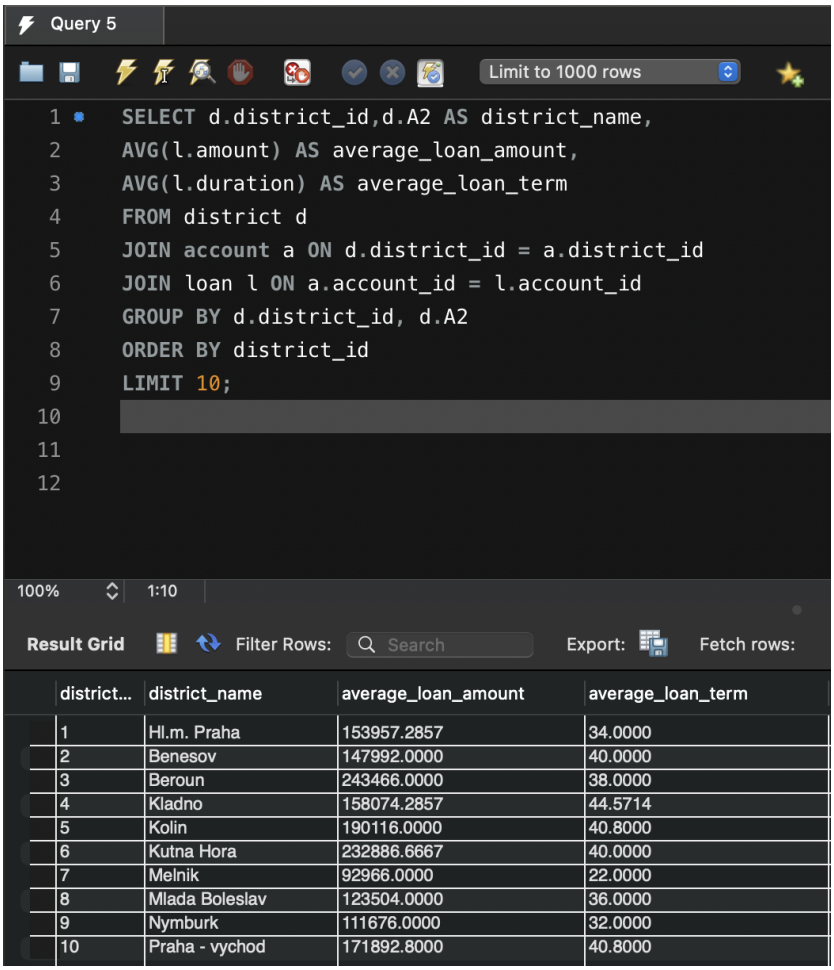
JOIN account a ON d.district_id = a.district_id

JOIN loan l ON a.account_id = l.account_id

GROUP BY d.district_id, d.A2

ORDER BY district_id

LIMIT 10;



district...	district_name	average_loan_amount	average_loan_term
1	Hl.m. Praha	153957.2857	34.0000
2	Benesov	147992.0000	40.0000
3	Beroun	243466.0000	38.0000
4	Kladno	158074.2857	44.5714
5	Kolin	190116.0000	40.8000
6	Kutna Hora	232886.6667	40.0000
7	Melnik	92966.0000	22.0000
8	Mlada Boleslav	123504.0000	36.0000
9	Nymburk	111676.0000	32.0000
10	Praha - vychod	171892.8000	40.8000

The analysis is focused on understanding the average loan characteristics in different districts. It helps identify districts with higher or lower average loan amounts and durations. The results could be used for strategic decision-making, such as targeting specific districts for marketing campaigns or adjusting loan terms based on regional trends. The grouping by district allows for a more granular analysis, revealing variations in loan patterns across different geographical areas.

By joining the **district**, **account**, and **loan** tables, the query establishes relationships between district information and loan details, enabling a comprehensive analysis.

Aggregating loan data using the **AVG** function allows for the calculation of district-level averages, providing a summary of the typical loan characteristics in each district.

The **GROUP BY** clause ensures that the averages are calculated for each unique district, preventing the results from being aggregated across all records.

The **ORDER BY** clause organizes the results by district ID, facilitating easier interpretation and comparison.

Reasons why calculating average of loan is important because:

1. **Regional Variations:** Understanding the average loan amounts and terms at the district level helps identify regional variations in borrowing patterns. This information is crucial for tailoring financial products to meet the specific needs and preferences of customers in different areas.

2. **Risk Assessment:** By analyzing loan data by district, financial institutions can assess the risk associated with lending in specific regions. Higher average loan amounts or longer terms may indicate different risk profiles, allowing for more informed decisions on risk management and mitigation strategies.

3. **Marketing and Targeting:** The analysis enables targeted marketing efforts. Financial institutions can use the insights to design marketing campaigns that address the unique characteristics and demands of each district. This can lead to more effective customer acquisition and retention strategies.

4. **Product Optimization:** Understanding the average loan amounts and terms in different districts helps in optimizing financial products. It allows institutions to adjust loan terms, interest rates, and other features to better align with the preferences and financial capabilities of customers in each region, ultimately improving customer satisfaction.

5. **Strategic Planning:** District-level loan analysis is valuable for strategic planning. It helps financial institutions identify growth opportunities in specific regions, allocate resources effectively, and make informed decisions about expanding or refining their services. This type of analysis supports long-term planning and ensures that business strategies are aligned with regional market dynamics.

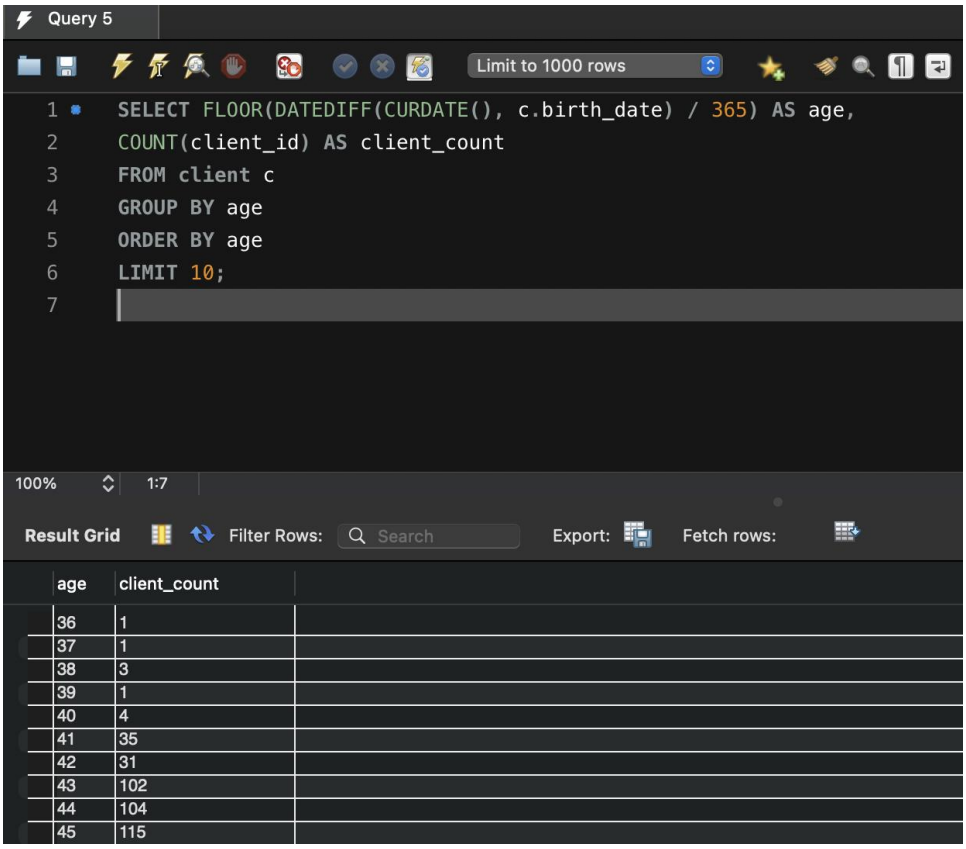
In summary, district-level analysis of loan data provides actionable insights that go beyond overall averages, allowing financial institutions to make data-driven decisions. These insights are crucial for risk management, product development, targeted marketing, and strategic planning, ultimately contributing to the overall success and sustainability of financial institutions.

---

10. Client Age Distribution: Analyzing the distribution of client ages

SQL QUERY:

```
SELECT FLOOR(DATEDIFF(CURDATE(), c.birth_date) / 365) AS age,
COUNT(client_id) AS client_count
FROM client c
GROUP BY age
ORDER BY age;
```



This query is designed to analyze the distribution of client ages in the **client** table.

The calculation of age is based on the current date and each client's birth date.

The **GROUP BY** clause groups the results by age, and the **COUNT** function counts the number of clients in each age group.

The **ORDER BY** clause sorts the results by age in ascending order, providing an age distribution.

Finally, the **LIMIT** clause ensures that only the top 10 age groups are displayed, making the output more manageable.

Reasons why the age and client analysis could be valuable:

1. **Demographic Analysis:** The query allows for a thorough analysis of the age distribution within the client base. Understanding the demographic composition is crucial for tailoring products, services, and marketing strategies to specific age groups.

2. **Marketing Targeting:** By knowing the age distribution of clients, businesses can create targeted marketing campaigns. Different age groups may have varying preferences, needs, and behaviors, and this information can be used to design more effective and personalized marketing messages.

3. **Product Customization:** The age data can inform the development or customization of financial products. Different age groups may have distinct financial goals, risk tolerances, and preferences. Tailoring products to suit the needs of specific age segments can enhance customer satisfaction and engagement.

4. **Strategic Planning:** Analyzing the age distribution aids in strategic planning for the future. Businesses can use this information to anticipate changes in their client base over time and make informed decisions about resource allocation, product development, and overall business strategy.

5. **Risk Management:** Age can be a factor in assessing risk, especially in the financial industry. Certain age groups may have different risk profiles, and understanding these variations is valuable for risk management strategies, including decisions related to lending, credit scoring, and financial planning.

In summary, this query provides valuable insights into the age demographics of the client base, enabling businesses to make data-driven decisions in areas such as marketing, product development, strategic planning, and risk management. Understanding the composition of the client base in terms of age is a fundamental aspect of customer-centric decision-making in various industries.
