# EconoMetrics: Banking ER Analytics

Sanjukta Baruah
*School of Engg. and Applied Sciences*
*Data Science*
Buffalo, United States
sanjuktabaruah5@gmail.com

***Abstract*—The banking sector continuously seeks innovative strategies to enhance customer relationship management, transaction security, performance evaluation, and strategic planning. This project identifies seven critical business propositions addressable through Structured Query Language (SQL) including customer management, transaction monitoring, performance analysis, and compliance reporting. By developing an Entity-Relationship (ER) diagram and ensuring data integrity through Boyce-Codd Normal Form (BCNF), the study utilizes a hybrid data curation method. This combines synthetic data generation with the Faker library and the integration of authentic statewise data to simulate realistic scenarios. The creation of PostgreSQL tables and subsequent data importation via SQL Alchemy sets the stage for comprehensive data analysis aimed at solving these business proposals.**

## I. Problem Statement

The banking industry faces numerous challenges ranging from customer relationship management to compliance reporting and risk assessment. Addressing these challenges requires efficient data management and analysis tools. However, existing solutions often lack the flexibility and comprehensiveness needed to tackle diverse business requirements effectively. Hence, there is a need for an integrated analytics system capable of providing insights into various banking functions while ensuring data accuracy, compliance, and usability. This project aims to develop such a system, leveraging SQL-based analytics techniques to address critical banking challenges and enhance operational efficiency and decision-making processes within the banking sector.

## II. Data Generation and Expansion for Enhanced Query Interpretability and Optimization

Data curation was approached using a hybrid method involving the generation of synthetic data through Faker library and incorporation of real-world data obtained from credible sources, maintaining the authenticity of crucial attributes.

### *A. Synthetic Data Generation with Faker*

Utilized the Faker library to generate synthetic data for the ”df_customer” dataframe. For instance, attributes such as ”first_name,” ”last_name,” ”dob” (date of birth), and ”gender” were populated with realistic data to simulate customer profiles.

### *B. Skewed Distribution Implementation*

Implemented skewed distributions in certain dataframes to reflect real-world scenarios. In the ”df_transaction” dataframe, transactions typically falling within the range of 1 to 700 dollars were simulated using random.beta, ensuring a skewed distribution that aligns with common transaction patterns. Similarly, in the ”df_card” dataframe, the attribute ”is_blocked” was skewed to predominantly represent unblocked cards, mimicking the usual scenario where the majority of cards are functional with only a few being blocked.

### *C. Integration of Authentic Statewise Data*

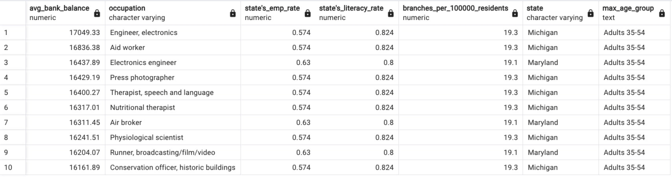
Integrated actual datasets sourced from internet(link) about diverse states into the ”df_statewise” dataframe. This dataframe aggregates information from consolidated CSV files containing crucial real-world data from various states. Attributes such as ”state_name” and relevant data pertaining to demographics, financial indicators, or other pertinent information were included to facilitate comprehensive analysis and decision-making processes.

The next step involved the creation of tables in PostgreSQL using SQL queries. Subsequently, the data from the prepared dataframes was successfully imported into these tables utilizing SQL Alchemy.

## III. Queries to solve the Business Problems

### *A. Customer Relation Management*

```
SELECT
  ROUND(AVG(a.bank_balance),2) AS avg_bank_balance,
  c.occupation,
  s."employment_rate(%)" AS "state’s_emp_rate",
  s."literacy_rate(%)" AS "state’s_literacy_rate",
  s.branches_per_100000_residents,
  s.state,
  CASE
        WHEN s."children_0_to_18(%)" = GREATEST
        (s."children_0_to_18(%)",
        s."adults_19_to_25(%)",
        s."adults_26_to_34(%)",
        s."adults_35_to_54(%)",
        s."adults_55_to_64(%)",
        s."65+(%)")
            THEN ’Children 0-18’
        WHEN s."adults_19_to_25(%)" = GREATEST
        (s."children_0_to_18(%)",
        s."adults_19_to_25(%)",
        s."adults_26_to_34(%)",
        s."adults_35_to_54(%)",
        s."adults_55_to_64(%)",
        s."65+(%)")
            THEN ’Adults 19-25’
        WHEN s."adults_26_to_34(%)" = GREATEST
        (s."children_0_to_18(%)",
        s."adults_19_to_25(%)",
        s."adults_26_to_34(%)",
        s."adults_35_to_54(%)",
        s."adults_55_to_64(%)",
        s."65+(%)")
            THEN ’Adults 26-34’
        WHEN s."adults_35_to_54(%)" = GREATEST
        (s."children_0_to_18(%)",
        s."adults_19_to_25(%)",
        s."adults_26_to_34(%)",
        s."adults_35_to_54(%)",
        s."adults_55_to_64(%)",
        s."65+(%)")
            THEN ’Adults 35-54’
        WHEN s."adults_55_to_64(%)" = GREATEST
        (s."children_0_to_18(%)",
        s."adults_19_to_25(%)",
        s."adults_26_to_34(%)",
        s."adults_35_to_54(%)",
        s."65+(%)")
            THEN ’Adults 55-64’
        ELSE ’65+’
    END AS max_age_group
FROM account a
INNER JOIN customer c ON a.customer_id = c.customer_id
INNER JOIN branch b ON c.branch_id = b.branch_id
INNER JOIN statewise s ON b.state = s.state
GROUP BY c.occupation,
s."employment_rate(%)",
s."literacy_rate(%)",
s.branches_per_100000_residents,
s."children_0_to_18(%)",
    s."adults_19_to_25(%)",
    s."adults_26_to_34(%)",
    s."adults_35_to_54(%)",
    s."adults_55_to_64(%)",
    s."65+(%)",s.state
ORDER BY avg_bank_balance DESC
LIMIT 10;
```



| | avg_bank_balance<br>numeric | occupation<br>character varying | state's_emp_rate<br>numeric | state's_literacy_rate<br>numeric | branches_per_100000_residents<br>numeric | state<br>character varying | max_age_group<br>text |
|---|---|---|---|---|---|---|---|
| 1 | 17049.33 | Engineer, electronics | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 2 | 16836.38 | Aid worker | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 3 | 16437.89 | Electronics engineer | 0.63 | 0.8 | 19.1 | Maryland | Adults 35-54 |
| 4 | 16429.19 | Press photographer | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 5 | 16400.27 | Therapist, speech and language | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 6 | 16317.01 | Nutritional therapist | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 7 | 16311.45 | Air broker | 0.63 | 0.8 | 19.1 | Maryland | Adults 35-54 |
| 8 | 16241.51 | Physiological scientist | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |
| 9 | 16204.07 | Runner, broadcasting/film/video | 0.63 | 0.8 | 19.1 | Maryland | Adults 35-54 |
| 10 | 16161.89 | Conservation officer, historic buildings | 0.574 | 0.824 | 19.3 | Michigan | Adults 35-54 |

Fig. 1. Customer Relationship Mgmt

This table provides valuable insights into how different occupations might correlate with average bank balances, and how these figures interact with broader socioeconomic indicators at the state level.

**Analytical insights:**

1) **Occupation and Bank Balance Correlation**
    - **Financial Behavior Analysis:** By examining the link between occupations and average bank balances, banks can analyze financial behaviors specific to different professions, identifying trends and opportunities for targeted product offerings.
2) **State Demographic Insights**
    - **Economic Health Assessment:** Through state data on employment and literacy rates, banks can assess the economic health of different regions, analyzing how it affects banking needs and access to inform strategic decisions on service provision.
3) **Age Demographic Financial Tailoring**
    - **Life Stage Financial Planning:** By knowing the main age demographic, banks can analyze the financial needs at different life stages, allowing them to develop age-specific products and advisory services that align with their customers' life milestones.
4) **Strategy Goals**
    - **Service Customization Analysis:** Banks can utilize customer data to analyze and identify opportunities for service customization, ensuring that each customer receives a personalized banking experience that meets their unique financial needs.
    - **Financial Literacy Initiatives Analysis:** The impact of financial education on customer behavior and satisfaction can be analyzed to tailor financial literacy initiatives effectively, ensuring they are relevant and beneficial to the targeted demographic.
    - **Branch Network Optimization Analysis:** An analysis of the distribution of bank branches relative to population demographics can inform an optimization strategy for branch locations, enhancing physical accessibility and customer convenience.
    - **Customer Engagement and Wellbeing Analysis:** By integrating all these data points, banks can analyze the drivers of customer engagement and financial wellbeing to refine CRM strategies and ensure they are creating value for customers and the bank.

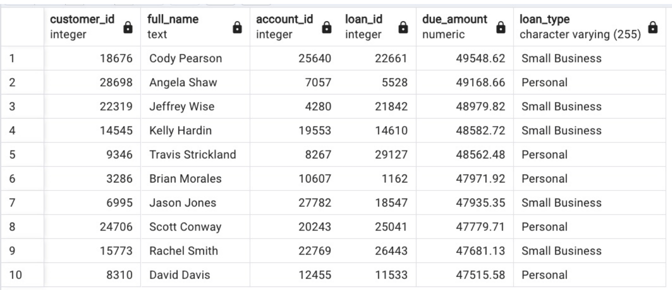
### *B. Risk Assessment Management*

```
SELECT
    c.customer_id,
    c.first_name || ’ ’ || c.last_name as full_name,
    a.account_id,
    l.loan_id,
    (lt.base_amount - l.amount_paid) as due_amount,
    lt.loan_type
FROM
    customer c
JOIN
    account a ON c.customer_id = a.customer_id
JOIN
    loan l ON a.account_id = l.account_id
JOIN
    loan_type lt ON l.loan_type_id = lt.loan_type_id
WHERE
    EXTRACT(YEAR FROM AGE(c.date_of_birth)) > 65
    AND l.due_date BETWEEN CURRENT_DATE
    AND CURRENT_DATE + INTERVAL ’10 years’
    AND (lt.base_amount - l.amount_paid) > (a.bank_balance / 2)
ORDER BY
    due_amount DESC
LIMIT 10;
```

This table(Fig 2) provides valuable insights into how different occupations might correlate with average bank balances, and how these figures interact with broader socioeconomic indicators at the state level.

**Analytical insights**

1) **Occupation and Bank Balance Correlation**
    - **Financial Behavior Analysis:** By examining the link between occupations and average bank balances, banks can analyze financial behaviors specific to different professions, identifying trends and opportunities for targeted product offerings.



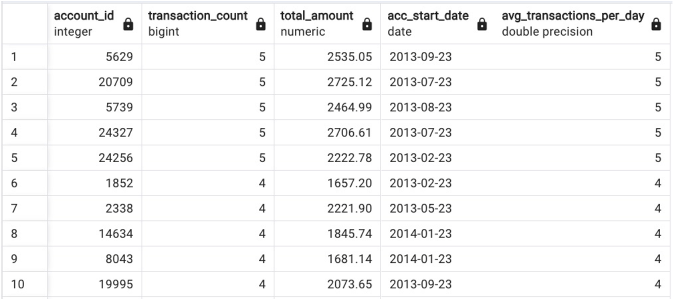
| | customer_id<br>integer | full_name<br>text | account_id<br>integer | loan_id<br>integer | due_amount<br>numeric | loan_type<br>character varying (255) |
|---|---|---|---|---|---|---|
| 1 | 18676 | Cody Pearson | 25640 | 22661 | 49548.62 | Small Business |
| 2 | 28698 | Angela Shaw | 7057 | 5528 | 49168.66 | Personal |
| 3 | 22319 | Jeffrey Wise | 4280 | 21842 | 48979.82 | Small Business |
| 4 | 14545 | Kelly Hardin | 19553 | 14610 | 48582.72 | Small Business |
| 5 | 9346 | Travis Strickland | 8267 | 29127 | 48562.48 | Personal |
| 6 | 3286 | Brian Morales | 10607 | 1162 | 47971.92 | Personal |
| 7 | 6995 | Jason Jones | 27782 | 18547 | 47935.35 | Small Business |
| 8 | 24706 | Scott Conway | 20243 | 25041 | 47779.71 | Personal |
| 9 | 15773 | Rachel Smith | 22769 | 26443 | 47681.13 | Small Business |
| 10 | 8310 | David Davis | 12455 | 11533 | 47515.58 | Personal |

Fig. 2. Risk Assessment

2) **State Demographic Insights**
    - **Economic Health Assessment:** Through state data on employment and literacy rates, banks can assess the economic health of different regions, analyzing how it affects banking needs and access to inform strategic decisions on service provision.
3) **Age Demographic Financial Tailoring**
    - **Life Stage Financial Planning:** By knowing the main age demographic, banks can analyze the financial needs at different life stages, allowing them to develop age-specific products and advisory services that align with their customers' life milestones.
4) **Strategy Goals**
    - **Service Customization Analysis:** Banks can utilize customer data to analyze and identify opportunities for service customization, ensuring that each customer receives a personalized banking experience that meets their unique financial needs.
    - **Financial Literacy Initiatives Analysis:** The impact of financial education on customer behavior and satisfaction can be analyzed to tailor financial literacy initiatives effectively, ensuring they are relevant and beneficial to the targeted demographic.
    - **Branch Network Optimization Analysis:** An analysis of the distribution of bank branches relative to population demographics can inform an optimization strategy for branch locations, enhancing physical accessibility and customer convenience.
    - **Customer Engagement and Wellbeing Analysis:** By integrating all these data points, banks can analyze the drivers of customer engagement and financial wellbeing to refine CRM strategies and ensure they are creating value for customers and the bank.

### *C. Transaction Tracking and Fraud Detection*

```
SELECT
    t.account_id,
    COUNT(*) as transaction_count,
    SUM(t.transaction_amount) as total_amount,
    a.acc_start_date,
    COUNT(*) / (date_part('day',
    age(CURRENT_DATE, a.acc_start_date)) + 1)
    as avg_transactions_per_day
FROM
    transaction t
JOIN
    account a ON t.account_id = a.account_id
GROUP BY
    t.account_id, a.acc_start_date
HAVING
    COUNT(*) / (date_part('day',
    age(CURRENT_DATE, a.acc_start_date)) + 1) > 0.26
ORDER BY avg_transactions_per_day DESC;
```



| | account_id<br>integer | transaction_count<br>bigint | total_amount<br>numeric | acc_start_date<br>date | avg_transactions_per_day<br>double precision |
|---|---|---|---|---|---|
| 1 | 5629 | 5 | 2535.05 | 2013-09-23 | 5 |
| 2 | 20709 | 5 | 2725.12 | 2013-07-23 | 5 |
| 3 | 5739 | 5 | 2464.99 | 2013-08-23 | 5 |
| 4 | 24327 | 5 | 2706.61 | 2013-07-23 | 5 |
| 5 | 24256 | 5 | 2222.78 | 2013-02-23 | 5 |
| 6 | 1852 | 4 | 1657.20 | 2013-02-23 | 4 |
| 7 | 2338 | 4 | 2221.90 | 2013-05-23 | 4 |
| 8 | 14634 | 4 | 1845.74 | 2014-01-23 | 4 |
| 9 | 8043 | 4 | 1681.14 | 2014-01-23 | 4 |
| 10 | 19995 | 4 | 2073.65 | 2013-09-23 | 4 |

Fig. 3. Transaction tracking and Fraud Detection

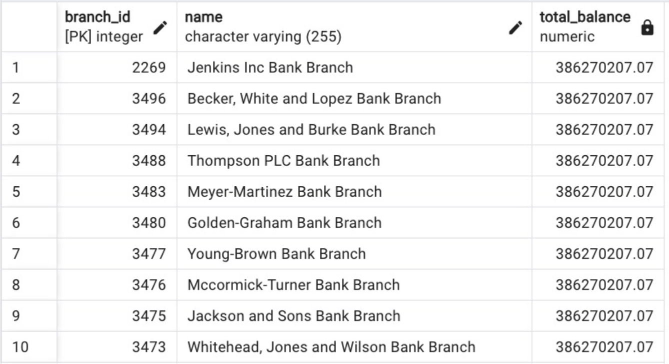
The table displays a bank's loan account data, including customer IDs, names, loan details, and due amounts, categorized by personal and small business loans, useful for analyzing transaction patterns and detecting fraud.

**Analytical insights:**

1) **Transaction Activity Analysis:**
    - By comparing the daily transactions of each account against the national average transaction rate (0.26 transactions per person per day), we can pinpoint accounts with higher-than-average activity. Accounts that consistently exceed this benchmark could be flagged for further investigation.
2) **Temporal Analysis:**
    - We can enhance our scrutiny by taking into account the length of time each account has been open, using the acc_start_date field. By calculating the average number of transactions per day since each account's inception, we can identify any accounts that show an unusual increase in activity over time or right from the start, which might suggest fraudulent behavior.
3) **Pattern Recognition:**
    - Beyond the frequency of transactions, analyzing the type, timing, and amounts can also reveal patterns indicative of fraud. Unusual patterns, such as high-value transactions occurring at odd hours, or transactions that do not fit the typical profile of the account holder based on their loan type (e.g., Personal vs. Small Business), could warrant a closer look.

### *D. Branch Performance*

```
SELECT b.branch_id, b.name, SUM(a.bank_balance) as total_balance
FROM customer c
JOIN branch b ON b.branch_id = c.branch_id
JOIN account a ON a.customer_id = a.customer_id
GROUP BY b.branch_id, b.name
ORDER BY total_balance
```



| | branch_id<br>[PK] integer | name<br>character varying (255) | total_balance<br>numeric |
|---|---|---|---|
| 1 | 2269 | Jenkins Inc Bank Branch | 386270207.07 |
| 2 | 3496 | Becker, White and Lopez Bank Branch | 386270207.07 |
| 3 | 3494 | Lewis, Jones and Burke Bank Branch | 386270207.07 |
| 4 | 3488 | Thompson PLC Bank Branch | 386270207.07 |
| 5 | 3483 | Meyer-Martinez Bank Branch | 386270207.07 |
| 6 | 3480 | Golden-Graham Bank Branch | 386270207.07 |
| 7 | 3477 | Young-Brown Bank Branch | 386270207.07 |
| 8 | 3476 | Mccormick-Turner Bank Branch | 386270207.07 |
| 9 | 3475 | Jackson and Sons Bank Branch | 386270207.07 |
| 10 | 3473 | Whitehead, Jones and Wilson Bank Branch | 386270207.07 |

Fig. 4. Account Management

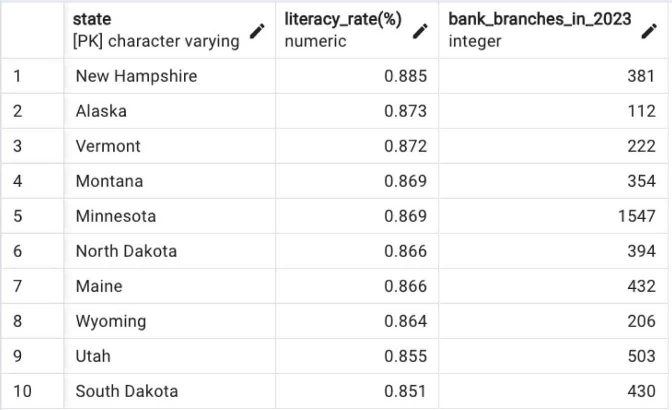
"The bank contains various bank branches with their corresponding identifiers and total balances."

**Analytical insights:**

1) **Total Balance Analysis:**
   - The total balance column in the table gives an indication of the amount of money deposited at each branch. A higher total balance could suggest that a branch is doing well in terms of attracting deposits, which is often a sign of good performance.

### *E. Strategic Planning and Analysis*

```
SELECT state, "literacy_rate(%)", bank_branches_in_2023
FROM statewise
ORDER BY "literacy_rate(%)" DESC
LIMIT 10;
```



| | state<br>[PK] character varying | literacy_rate(%)<br>numeric | bank_branches_in_2023<br>integer |
|---|---|---|---|
| 1 | New Hampshire | 0.885 | 381 |
| 2 | Alaska | 0.873 | 112 |
| 3 | Vermont | 0.872 | 222 |
| 4 | Montana | 0.869 | 354 |
| 5 | Minnesota | 0.869 | 1547 |
| 6 | North Dakota | 0.866 | 394 |
| 7 | Maine | 0.866 | 432 |
| 8 | Wyoming | 0.864 | 206 |
| 9 | Utah | 0.855 | 503 |
| 10 | South Dakota | 0.851 | 430 |

Fig. 5. Strategic Planning and Assessment

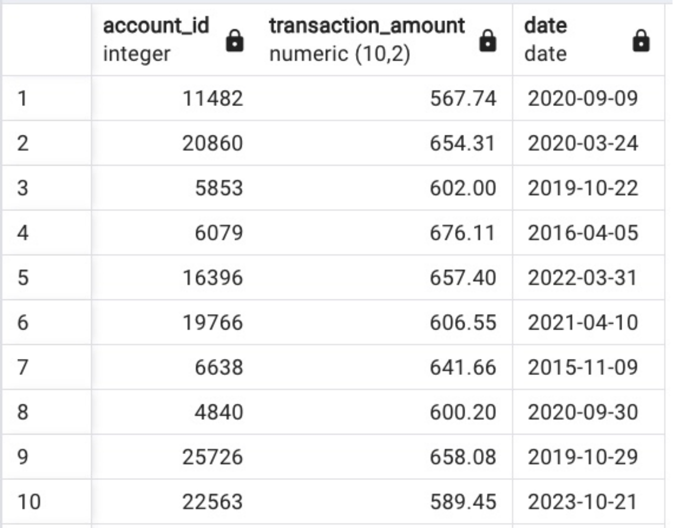
The table contains demographic and economic data, which could be used for strategic planning, such as where to open new branches or which services to focus on in certain regions.

**Analytical insights:**

1) **Branch Efficiency:**
   - Evaluating the effectiveness of each branch in attracting and managing deposits, as a uniform total balance may indicate operational anomalies or data inconsistencies.
2) **Strategic Planning:**
   - Identifying opportunities for growth and resource allocation to optimize network performance and customer reach.
3) **Market Understanding:**
   - Gaining insights into market conditions and customer preferences, which can guide decisions on service improvements and expansion.

### *F. Compliance and Reporting*

```
SELECT a.account_id, t.transaction_amount, t.date
FROM account a
JOIN transaction t ON a.account_id = t.account_id
WHERE t.transaction_amount > (SELECT MAX(transaction_amount) * 0.80 FROM
```



| | account_id<br>integer | transaction_amount<br>numeric (10,2) | date<br>date |
|---|---|---|---|
| 1 | 11482 | 567.74 | 2020-09-09 |
| 2 | 20860 | 654.31 | 2020-03-24 |
| 3 | 5853 | 602.00 | 2019-10-22 |
| 4 | 6079 | 676.11 | 2016-04-05 |
| 5 | 16396 | 657.40 | 2022-03-31 |
| 6 | 19766 | 606.55 | 2021-04-10 |
| 7 | 6638 | 641.66 | 2015-11-09 |
| 8 | 4840 | 600.20 | 2020-09-30 |
| 9 | 25726 | 658.08 | 2019-10-29 |
| 10 | 22563 | 589.45 | 2023-10-21 |

Fig. 6. Compliance and Reporting

The table displayed appears to list individual financial transactions, capturing the account ID, the amount of the transaction, and the date it occurred.

**Analytical insights**

1) **Transaction Trends Analysis:**
   - By examining the transaction amounts and dates, we can identify patterns over time, such as peak transaction periods or changes in transaction values. Based on this observation, the bank can decide the release of different products or offers to attract customers.
2) **Audit Compliance:**
   - We can generate reports that isolate transactions exceeding certain thresholds, such as 80% of the maximum transaction value, to comply with internal audits and regulatory requirements. Such reports can also be crucial in detecting outliers or unusual transactions that may warrant further investigation for error or fraud detection.

## IV. Key Learnings from SQL Implementation in Business Solutions

1) **Adaptation to Database-Specific Functions:**

- *Learning:* Adapting to different database environments requires familiarity with compatible functions. While certain functions like DATEFROMPARTS and DATEDIFF are commonly used in SQL Server, PostgreSQL may require alternative functions such as AGE and NOW. This experience highlights the importance of understanding database-specific functionalities and adapting queries accordingly to ensure compatibility and achieve desired results.

2) **Syntax Compatibility in Query Construction:**
   - *Learning:* Constructing SQL queries involves attention to syntax nuances specific to the database platform being used. During query construction, issues such as mismatched column projections in subqueries may arise, necessitating iterative refinement of SQL statements. This underscores the significance of thorough syntax understanding and meticulous query construction to avoid errors and ensure query effectiveness across different database environments.

3) **Comprehensive Project Execution:**
   - *Learning:* The project's success relied on a comprehensive approach encompassing various stages from proposal identification to data curation, SQL query formulation, and troubleshooting. This experience emphasizes the importance of a structured project workflow, encompassing diverse tasks and meticulous execution to effectively utilize SQL in addressing multifaceted business challenges.

## References

[1] https://www.geeksforgeeks.org/python-pandas-working-with-dates-and-times/.

[2] https://www.geeksforgeeks.org/python-faker-library/

[3] https://www.geeksforgeeks.org/types-of-keys-in-relational-model-candidate-super-primary-alternate-and-foreign/

[4] https://www.holistics.io/blog/top-5-free-database-diagram-design-tools/

[5] https://app.diagrams.net/