***Customer Segmentation and Risk Profiling in Financial Data Using SQL***

In today's data-driven financial landscape, extracting meaningful insights from large volumes of financial data is critical for informed decision-making. This project, **" Customer Segmentation and Risk Profiling in Financial Data Using SQL”,** delves into the intricacies of financial and transaction data to uncover trends, optimize user segmentation, and detect potential risks or anomalies.

The project leverages Structured Query Language (SQL) to analyze datasets related to credit cards, transactions, users, and merchants. By performing exploratory data analysis (EDA), customer segmentation, risk profiling, and transaction behavior analysis, the project aims to provide actionable insights for stakeholders, such as financial institutions and businesses, to enhance decision-making and mitigate risks.

**Objectives:**

1. **Data Quality Assessment**: Identify missing, invalid, or inconsistent data across key datasets and address data quality issues.
2. **Exploratory Data Analysis (EDA)**: Explore patterns in credit card usage, transaction behavior, and user demographics to extract high-level trends.
3. **Customer Segmentation**: Perform segmentation based on RFM (Recency, Frequency, Monetary), demographics, and behavioral metrics for targeted marketing strategies.
4. **Risk Profiling**: Develop credit score-based risk categorization and detect high-risk transactions for fraud detection and prevention.
5. **Key Insights Extraction**: Analyze spending patterns by gender, income group, day of the week, and seasonality to identify influential factors in user behavior.
6. **Fraud Detection**: Identify anomalies such as duplicate transactions and abnormal spending patterns to detect potential fraudulent activity.

**Tools and Techniques:**

This project relies on SQL as the primary tool for querying, analyzing, and transforming data. Advanced SQL techniques, including subqueries, window functions, common table expressions (CTEs), and statistical computations, are used to derive insights.

**Key Outcomes:**

The project provides an end-to-end view of financial and transaction data, offering actionable insights such as:

* Identification of high-value and high-risk users.
* Detection of outliers in credit utilization.
* Detailed segmentation for personalized financial services.
* Fraud detection using data-driven approaches.

This SQL project highlights the importance of leveraging data analytics in the financial sector, showcasing how structured analysis can lead to better decision-making, customer engagement, and operational efficiency.

**Deliverables and Workflow**

This project focuses on analyzing customer behavior, segmenting users based on transaction patterns, and uncovering insights from transactional, card, and user data. The deliverables include SQL scripts, an Entity-Relationship Diagram (ERD), a comprehensive report, an interactive dashboard, and clean datasets ready for further analysis.

**Project Deliverables**

1. **SQL Query Scripts**:

* Queries to clean and transform data.
* Queries for exploratory data analysis (EDA).
* Queries to implement segmentation models (e.g., RFM analysis, risk profiling).
* Queries for insights such as spending patterns, fraud detection, and credit usage trends.

1. **Entity-Relationship Diagram (ERD)**:

* A clear representation of relationships between datasets (e.g., User Data, Card Data, Transaction Data).

1. **Comprehensive Report**:

* Detailed documentation of findings, SQL workflows, and insights derived from the analysis.
* Interpretation of patterns and trends observed in the data.

1. **Dashboard**:

* Visual representation of key insights such as:
* Spending patterns over time.
* Customer segmentation.
* Credit risk profiles.

1. **Cleaned Datasets**:

* Transformed datasets free of duplicates and inconsistencies, ready for additional analytics or visualization.

**Workflow Objectives and Tasks**

1. **Data Cleaning and Transformation**
   * Identify and handle missing data across all datasets.
   * Format columns for consistency (e.g., date formats, numeric fields).
   * Validate foreign key relationships between datasets.
   * Remove duplicate or erroneous entries.
2. **Exploratory Data Analysis (EDA)**
   * Analyze distributions of key metrics such as transaction amounts, credit scores, and yearly income.
   * Explore relationships between:
     + User demographics and transaction behavior.
     + Card type and transaction trends.
   * Detect outliers or unusual patterns.
3. **Customer Segmentation**
   * Implement RFM (Recency, Frequency, Monetary) segmentation.
   * Group users by credit risk (e.g., Low, Medium, High Risk) based on their credit scores and total debt.
   * Identify customer personas using demographic and transactional data.
4. **Spending and Fraud Analysis**
   * Identify transaction trends based on merchant categories (mcc) and geographic data.
   * Detect anomalies such as duplicate transactions or exceeding credit limits.
5. **Visualization and Reporting**
   * Use Power BI or Tableau to create an interactive dashboard with key performance indicators (KPIs).
   * Summarize insights in a well-structured report, highlighting patterns and actionable recommendations.

### ****Exploratory Data Analysis (EDA)****

The exploratory data analysis (EDA) process provides insights into the data's structure, distribution, and underlying trends. The following steps outline the EDA performed for this project, leveraging SQL queries to ensure comprehensive analysis.

**1. Understanding Data Distributions**

To assess the completeness and integrity of the datasets, SQL queries were used to identify missing values in critical columns across all datasets:

**Card Data:** A query was executed to check for missing values in key attributes such as card\_id, user\_id, card\_brand, and credit\_limit.

**Errors Data**: Missing descriptions in the errors dataset were flagged to ensure proper error categorization.

**Transaction Data:** Transactions with null values in essential fields like transaction\_id, user\_id, amount, or mcc\_code were identified.

**User Data:** Missing demographic data, such as current\_age, gender, or credit\_score, were flagged for further cleaning.

Additionally, summary statistics were calculated to understand the distributions of critical numerical fields:

**Credit Limit:** Maximum, minimum, average, median, and mode values were calculated for the credit\_limit column in the Card Data.

**Spending Trends**: Temporal spending patterns were analyzed by gender, with average monthly spending calculated for transactions after January 2019.

**2. Investigating Relationships**

EDA involved analyzing the relationships between various attributes to uncover meaningful patterns:

**Demographics and Income**: The average yearly income and credit scores were grouped by gender to assess variations.

**Chip Usage**: The presence of chips in cards (has\_chip) was evaluated to determine adoption rates.

**Card Type Usage**: The frequency of different card types (card\_type) was assessed to identify the most commonly used cards.

**Common Transaction Errors**: A join between the transaction\_data and errors datasets revealed the top three most frequent transaction errors, providing insights into system or user issues.

**3. Trend Analysis**

Temporal trends were evaluated to identify seasonal or long-term patterns:

**Card Usage Trends (Past 5 Years)**: The usage of cards was tracked by month to observe changes in transaction volumes over the past five years.

**Spending Patterns**: The average monthly spending by gender was calculated to detect differences in spending behavior.

**SQL Queries and Insights**

1. **Monthly Average Spending by gender**

There have been slight fluctuations in spending between males and females over the months. Certain months show higher averages for females (e.g., February 2019), while others show higher averages for males (e.g., November 2024). Spending patterns appear seasonal or event-driven, with noticeable variations in holiday periods (e.g., December).

1. **Summary Statistics**

The wide range of credit limits (0 to 151,223) reflects diverse customer profiles, from standard to high-net-worth clients. The low mode compared to the median and mean suggests most customers have lower limits, indicating a conservative lending strategy. Zero-limit accounts may warrant further investigation to determine their status or address potential data issues.

1. **Card with Chip**

The adoption of chip-enabled cards has consistently surpassed non-chip cards across all years, underscoring a strong preference for enhanced security in payment methods. Over time, non-chip cards have experienced a gradual decline, with a significant drop in 2024, possibly due to being phased out or replaced by newer technologies. While the count of chip-enabled cards remained relatively stable in previous years, there was a sharp decrease in 2024, indicating a potential market shift, regulatory changes, or a transition to alternative payment solutions.

1. **Most used card type**

The data indicates that debit cards dominate the distribution, accounting for the highest count among the card types, with 3,511 issued. Credit cards follow with 2,057 issued, showing significant usage but not surpassing debit cards. Prepaid debit cards make up the smallest portion, with 578 issued, reflecting a more niche use case compared to standard debit and credit cards. This distribution suggests a preference for traditional debit cards, potentially due to their ease of use and immediate access to funds.

1. **Most Common Error in Transaction**

The data reveals that the most common errors in card transactions are related to invalid card details, with "Bad Card Number, Bad Expiration" being the most frequent issue, occurring 7,375 times. Close behind are "Bad Expiration" errors, which occurred 7,370 times, and "Bad CVV" errors, with 7,342 instances. These errors highlight the importance of ensuring accurate card information is entered during transactions, as small mistakes in card numbers, expiration dates, or CVVs can lead to transaction failures and processing issues.

1. **Average Credit Score by Gender**

The data shows that the average income for females is slightly higher at $46,048 compared to males, who have an average income of $45,373. However, the difference in average income between genders is relatively small. Additionally, both groups have nearly identical round values, with females at 709 and males at 710, indicating minimal variance in the rounding or grouping of income data between genders. This suggests that income levels are somewhat similar across genders in this dataset, with a slight favor towards females.

1. **Card usage history in past 5 years**

Transaction usage is generally stable, with occasional fluctuations, such as higher usage in May and July and lower counts in February and November. November 2024 saw a sharp drop to 674 transactions, a significant anomaly. Overall, 2024 shows stable or slightly increased transaction counts compared to previous years, with a few seasonal variations.

### Customer Segmentation Analysis

**RFM Segmentation (Recency, Frequency, Monetary)**:  
This query assigns a recency, frequency, and monetary score to each user by analyzing the last transaction date, total number of transactions, and total amount spent. Each score is divided into four quartiles.

**Demographic Segmentation**:  
Users are categorized based on their age and income. The age groups are "Young" (under 30), "Middle-Aged" (30-50), and "Senior" (over 50). Income groups are "Low Income" (under 50K), "Middle Income" (50K-100K), and "High Income" (over 100K). This segmentation counts how many users fall into each demographic group.

**Behavioral Segmentation (Current Year)**:  
This query segments users based on their age, income, and their transaction count and total spending in 2024. It filters for users with over 100 transactions and a total amount spent over 10,000, then lists the top 5 segments with the highest transaction counts.

**Customer Lifetime Value (CLV) Segmentation**:  
Users are segmented based on their total spending across all transactions. Users with total spending over 500,000 are classified as "High CLV", those with spending between 100,000 and 500,000 as "Medium CLV", and others as "Low CLV".

**SQL Queries and Insights**

1. **RFM Segmentation(Recency, Frequency, Monetary)**

The RFM analysis segments users based on Recency (R), Frequency (F), and Monetary (M) scores, each ranging from 1 to 4. High-value users (R=4, F=4, M=4) are engaged and profitable, ideal for loyalty programs. At-risk users have low recency but high frequency and/or spending, requiring re-engagement campaigns. Frequent but low spenders are suited for upsell opportunities, while high spenders with low frequency can be targeted with exclusive offers to boost their transaction frequency. Recommendations include personalized campaigns for high-value users, re-engagement for at-risk users, and upsell or frequency incentives for other segments.

1. **Demographic Segmentation**

The majority of users are low-income, with the highest concentration in the Senior and Middle-Aged age groups. Middle-income users are fewer, and high-income users represent a small segment across all age groups. Targeted marketing could focus on addressing the needs of low-income and middle-income users.

1. **Behavioural Segmentation current year**

Low-income users, particularly in the Senior and Middle-Aged age groups, make up the highest transaction volume, with over 15,000 transactions. Young, low-income users are also active but with fewer transactions. Middle-income users, though fewer, still show significant transaction activity, highlighting potential for targeted engagement.

1. **Customer Lifetime Value (CLV) Segmentation**

The CLV segmentation results reveal that the majority of users fall into the **Low CLV** category (959 users), indicating that a large portion of users contribute less to revenue. A smaller group is in the **Medium CLV** (416 users), and a very small number (2 users) are in the **High CLV** category, suggesting a highly concentrated but valuable group of users who generate significant revenue. These insights suggest that there may be a significant opportunity to focus on increasing the CLV of the larger user base through personalized marketing, loyalty programs, or upselling strategies.

### ****Transactions Behaviour Segmentation****

**Transaction Count and Value by Day of Week**: The analysis of transactions by day of the week reveals which days contribute most to both transaction volume and total value. Days with the highest transaction value could be targeted for special promotions or campaigns.

**Seasonality Analysis**: This analysis helps identify months with higher spending trends, potentially tied to holidays, promotions, or seasonal events. It provides insight into when users are most likely to make large transactions.

**Transaction Frequency vs. Spending**: Users with more than five transactions tend to spend more. These users can be classified as highly engaged and may benefit from loyalty rewards, upselling, or cross-selling opportunities based on their spending patterns.

**Customer Retention Analysis**: Users who made purchases within the last 30 days are considered recent. This analysis can help segment customers into recent vs. lapsed categories, enabling targeted retention strategies for users who have not purchased recently.

**High-Value Transaction Identification**: Identifying transactions over $500 can highlight high-value users and their spending behavior, helping to create targeted marketing or loyalty programs aimed at retaining these high-value customers.

**Merchant Analysis: Revenue Contribution**: By identifying the top 10 merchants contributing to the highest revenue, this analysis can guide partnerships or marketing campaigns to further increase revenue from these merchants.

**User Credit Limit Utilization**: Tracking how much users are utilizing their credit limits can help identify high-risk users who might be close to maxing out their credit limits and may need additional monitoring or risk management strategies.

**SQL Queries and Insights**

1. **Transaction Count and Value by Day of Week**

Transactions are highest on Monday and Saturday, with significant values, indicating these days are key for user engagement and revenue. Sunday and Wednesday also show strong transaction volumes, suggesting potential for targeted promotions on weekends. Thursday and Friday show slightly lower activity, potentially indicating opportunities to increase engagement towards the end of the week.

1. **Seasonality Analysis**

The seasonality analysis shows stable spending throughout the year, with peaks in March, August, and October, indicating higher user engagement during these months. December and November show lower spending, suggesting a potential drop in activity during the holiday season. Targeted promotions and campaigns could help boost revenue during these slower months.

1. **Transaction Frequency Vs Spending**

This dataset shows a variety of user spending behaviors. Users with higher transaction counts (e.g., user 111 with 4177 transactions) tend to have higher total spending. On average, the spending per transaction remains consistent (around $140-$160), with a few users (e.g., user 317 with an average of $160) spending slightly more per transaction. Identifying high-frequency users could help focus on retention strategies for valuable customers.

1. **Customer Retention Analysis**

The retention analysis indicates that users with the highest transaction counts (e.g., user 1111 with 4218 transactions) are more likely to have made recent purchases. The majority of recent users have a significant number of transactions, suggesting their continued engagement. On the other hand, users categorized as "Not Recent" typically have lower transaction counts, which could be an area for targeted re-engagement strategies.

1. **Merchant Analysis, Revenue Contribution**

The merchant analysis reveals that the top merchants by total revenue are those with higher average transaction values, such as merchant 98349, with an average transaction value of 350.81. These merchants, despite fewer transactions, contribute significantly to the overall revenue. Merchants like 43482 and 34752, with a high number of transactions but slightly lower average transaction values, also play a substantial role in revenue generation. This suggests both transaction volume and value are critical for merchant success.

1. **Credit Utilization**

The analysis reveals that the majority of users (1377) have a credit utilization of less than 40%, indicating conservative usage of available credit. A smaller group (396) utilizes between 40-70% of their credit, while only 28 users have a utilization rate above 70%, signaling high credit reliance.

**Risk Detection**

**Credit Score-Based Risk Profiling**: Users are classified into high, moderate, or low risk based on their credit scores. Those with scores below 600 are categorized as high risk, while scores between 600 and 750 indicate moderate risk, and scores above 750 are considered low risk.

**High-Risk Transaction Detection**: Transactions over $5000 are flagged as potentially high-risk, helping identify large purchases that could indicate financial distress or fraud.

**Income Group vs. Spending Behavior**: Income groups are classified into low, middle, and high income based on yearly income. Spending behavior is tracked by the total amount spent, average transaction value, and transaction count, with the highest spenders being categorized accordingly.

**Fraud Detection Based on Abnormal Spending Patterns**: Users with more than 10 transactions and an average transaction amount over $1000 are flagged for potential fraud, as this may indicate unusual behavior.

**Duplicate Transactions**: Identifies duplicate transactions, highlighting potential fraudulent activity or system errors when a transaction appears more than once for the same card.

**SQL Queries and Insights**

1. **Credit Score-Based Risk Profiling**

Moderate Risk has the largest number of users, with a nearly equal gender distribution and an average income of $45,390. Moderate-High Risk has a similar income but fewer users and less gender disparity. Moderate-Low Risk shows a lower income and the oldest average age (48). Low Risk, with the smallest user count, has the highest income ($71,240) and the youngest average age (43). This suggests that higher income correlates with lower risk, while age plays a role in risk categorization.

1. **Income Group vs. Spending Behavior**

The analysis shows that **Low Income** users make the highest number of transactions (704,536) and spend the most overall ($105.68 million), with an average transaction of $150. **Middle Income** users have fewer transactions (269,686) and spend $40.38 million, with an average of $149.72 per transaction. **High Income** users make the fewest transactions (25,777) and spend the least ($3.86 million), but their average spend per transaction is similar to Middle Income at $149.73. Low-income individuals are more frequent, smaller spenders.

**Conclusion: Insights and Recommendations**

The "Customer Segmentation and Risk Profiling in Financial Data Using SQL" project provides a comprehensive understanding of financial data trends, customer behavior, and potential risk factors. By leveraging SQL, the project successfully uncovers actionable insights for businesses and financial institutions. The findings from data cleaning, exploratory data analysis (EDA), segmentation, and transaction behavior analysis pave the way for strategic decision-making and optimization. Here are the key conclusions drawn from the analysis:

**1. Spending Patterns and Demographic Trends**

**Key Findings:**

* Spending behavior is significantly influenced by demographic factors like gender, income, and age group.
* Certain user segments, such as senior low-income users, exhibit higher transaction volumes despite lower spending power, highlighting their engagement.
* Seasonal and day-of-week trends suggest spikes during weekends, holidays, and promotional seasons.

**Recommendations:**

* Tailor marketing campaigns and offers to specific user groups, such as seniors, who demonstrate high engagement despite financial constraints.
* Capitalize on seasonal and day-of-week patterns by aligning campaigns and promotions during high-spending periods like weekends and holidays.

**2. Customer Segmentation**

**Key Findings:**

* **RFM Analysis**: Identified high-value users who are consistent spenders, as well as at-risk users who require re-engagement strategies.
* **Demographic Segmentation**: The majority of users are in the low-income category, indicating potential for financial inclusivity initiatives.
* **CLV Segmentation**: A small group of users (High CLV) generates significant revenue, while most users contribute less, falling into the Low CLV category.

**Recommendations:**

* Develop loyalty programs for high-value (R=4, F=4, M=4) and high-CLV users to retain their engagement.
* Introduce targeted campaigns for low-frequency or low-value users to upsell and encourage higher spending.
* Use demographic segmentation to design personalized financial products catering to the unique needs of different user groups.

**3. Risk Profiling and Fraud Detection**

**Key Findings:**

* Transactions with invalid card details (e.g., CVV, expiration date) were the most frequent errors, highlighting operational inefficiencies.
* High-risk users were identified based on credit utilization patterns, with some users nearing their credit limits.
* Outlier transactions and duplicate entries were flagged, indicating potential fraud or system errors.

**Recommendations:**

* Enhance fraud detection mechanisms using machine learning models to identify anomalies and prevent high-risk transactions.
* Educate users on proper card usage to minimize errors and enhance transaction success rates.
* Monitor credit utilization and alert users when they approach risky thresholds to prevent defaults.

**4. Merchant and Transaction Trends**

**Key Findings:**

* Certain merchants contributed disproportionately to revenue, indicating key partnerships that drive value.
* Transactions show seasonality, with higher spending during peak shopping periods like December.

**Recommendations:**

* Strengthen relationships with top merchants through exclusive partnerships or joint promotional campaigns.
* Align inventory, staffing, and marketing efforts with identified seasonal trends to maximize revenue during peak periods.

**5. Fraud Detection and Anomaly Analysis**

**Key Findings:**

* Fraudulent activity, such as duplicate transactions and abnormal spending patterns, could not be identified.
* Users with zero credit limits or flagged transactions are potential areas of concern.

**Recommendations:**

* Use anomaly detection algorithms to automate fraud prevention and increase the system’s efficiency.
* Collaborate with merchants to enhance verification processes for card transactions, particularly for high-value purchases.

**6. Future Opportunities**

This project sets the foundation for further advanced analytics. The insights can be expanded upon through:

* **Machine Learning Models**: Build predictive models to classify users into high-risk categories or predict churn.
* **Dynamic Credit Allocation**: Use segmentation insights to optimize credit offerings for each user group.
* **Enhanced Dashboards**: Implement interactive, real-time visualizations for stakeholders to monitor KPIs.

**Strategic Impact**

This project underscores the value of data-driven insights for financial institutions and businesses. The use of SQL for comprehensive analysis highlights its power in identifying patterns and trends critical for decision-making. By leveraging the findings, organizations can enhance customer engagement, improve risk management, and drive operational efficiency.

**Final Thought:**

Understanding customer behavior through segmentation and profiling is not just about identifying patterns; it is about taking meaningful action. By adopting the insights and recommendations derived from this project, organizations can better align their strategies with user needs, ensure financial security, and foster a more inclusive, responsive financial ecosystem.