**SQL (Data) Engineer - Silvertreebrands**

**Introduction and Overview**

1. **Overview of the Task:**

The task involved performing RFM (Recency, Frequency, Monetary) analysis on two customer datasets. The goal was to create customer classification columns that categorize customers based on their purchasing behavior.

1. **Initial Setup:**

To start, I uploaded the customer datasets into a SQL database using MSSQL Server. This allowed me to efficiently query and manipulate the data.

**Step-by-Step Explanation of Approach**

# SECTION A

1. **Calculating RFM Metrics:**

* **Recency Calculation:**
  + For Recency, I calculated the number of days since each customer’s last order. This was done by taking the difference between the most recent transaction date for each customer and a reference date (typically the current date).

```sql

SELECT CustomerID, DATEDIFF(day, MAX(TransactionDate), GETDATE()) AS Recency

FROM Transactions

GROUP BY CustomerID;

```

* + This metric helps identify how recently a customer has engaged with the business, which is crucial in understanding their current engagement level.
* **Frequency Calculation:**
  + For Frequency, I counted the total number of transactions each customer made. This gives insight into how often a customer returns to make a purchase.

```sql

SELECT CustomerID, COUNT(TransactionID) AS Frequency

FROM Transactions

GROUP BY CustomerID;

```

* + Higher frequency indicates a more loyal customer who frequently interacts with the business.
* **Monetary Calculation:**
  + For Monetary, I summed the total value of transactions for each customer. This metric reflects the overall revenue each customer has contributed.

```sql

SELECT CustomerID, SUM(TransactionAmount) AS Monetary

FROM Transactions

GROUP BY CustomerID;

```

* + The monetary value is key to identifying high-value customers who contribute significantly to the business's revenue.

**Conclusion on Customer Behavior:**

var SI = Skyrate Investment (Customer Dataset 2)

var MT = Munemo Trading (Customer Dataset 1)

**1. Transaction Volume vs. Revenue:**

* + SI had a slightly higher number of transactions (27,090) compared to MT (24,960), but the revenue generated by SI was significantly higher at 20,055,288 compared to MT’s 5,589,665.
  + Despite having a lower transaction volume and offering discounts, MT's total revenue is much lower than SI's, which indicates that the average revenue per transaction is significantly lower for MT.

**2. Customer Base Size:**

* + MT had a larger customer base with 1,454 customers compared to SI's 200 customers. This suggests that MT is catering to a broader but possibly less lucrative customer base.
  + The fact that SI generates significantly more revenue with fewer customers suggests that SI’s customers are either more engaged, have higher spending power, or are more loyal.

**3. Impact of Discounts:**

* + MT offered discounts, which likely attracted more customers but may have also reduced the overall profitability. The large number of transactions does not seem to translate into proportional revenue, likely due to the discounts reducing the net revenue.
  + SI, on the other hand, did not offer any discounts yet achieved much higher revenue. This could indicate that SI's customers are less price-sensitive, more loyal, or perceive higher value in SI's offerings.

**4. Revenue Per Customer:**

* + The revenue per customer for SI is approximately 100,276 (20,055,288 / 200), whereas for MT, it is approximately 3,844 (5,589,665 / 1,454).
  + This stark difference indicates that SI’s customers are significantly more valuable on a per-customer basis.

**Advice for the Two Businesses:**

**1. MT (Focus on Increasing Profitability):**

* + Reevaluate Discount Strategy: MT should assess whether the discounts are truly driving profitability. If the discounts are not leading to enough additional revenue to justify the lower margins, it might be beneficial to reduce the discounts or offer them more selectively.
  + Segment Customers: MT should consider segmenting its customers to identify high-value customers who might not need discounts to make a purchase. Target these customers with tailored offers that emphasize value rather than price.
  + Enhance Customer Value: Focus on strategies to increase the average transaction value, such as upselling, cross-selling, or offering premium products or services.
  + Customer Loyalty Programs: Consider implementing loyalty programs that reward repeat customers without heavily relying on discounts. This could increase the lifetime value of each customer.

**2. SI (Maintain and Enhance Customer Loyalty):**

* + Maintain Current Strategy: SI seems to be doing well with its current pricing strategy, as evidenced by the high revenue with no discounts. Maintaining this approach while ensuring that the perceived value of the products or services remains high will likely continue to be beneficial.
  + Focus on Customer Experience: SI could further enhance customer loyalty by focusing on the overall customer experience, including excellent service, quality products, and personalized offers.
  + Expand Customer Base: While SI’s current customers are highly valuable, growing the customer base with similar high-value customers could further increase revenue. SI could consider targeted marketing campaigns to attract more customers who are less price-sensitive and value the offerings.
  + Data-Driven Insights: Leverage data analytics to understand the preferences and behaviors of these high-value customers, allowing for more personalized marketing efforts and product recommendations.

**Final Thoughts:**

- MT should consider optimizing its discount strategy and finding ways to increase the profitability of its large customer base.

- SI should continue its successful approach while exploring ways to grow its high-value customer base and further strengthen customer loyalty.

# SECTION B

**2. Setting Thresholds and Bucketing:**

* **Determining Buckets:**
  + After calculating the RFM scores, I needed to categorize customers into different behavioral segments. To do this, I divided the RFM scores into buckets. I used SQL's `NTILE` function to evenly distribute customers across these buckets.

```sql

SELECT CustomerID,

NTILE(5) OVER (ORDER BY Recency ASC) AS RecencyBucket,

NTILE(5) OVER (ORDER BY Frequency DESC) AS FrequencyBucket,

NTILE(5) OVER (ORDER BY Monetary DESC) AS MonetaryBucket

FROM RFMTable;

```

* + Each bucket represents a different segment of customers based on their RFM scores. For example, customers in the top Recency bucket are those who have made recent purchases, while those in the top Monetary bucket are the highest spenders.
* **Choosing Group Descriptions (Set X or Set Y):**
  + I then used these buckets to assign each customer to a specific group. I chose to use [mention Set X or Set Y] because [provide your reasoning]. For instance, I found that the group names in [Set X/Set Y] more accurately reflected the customer behaviors I observed in the data.

```sql

CASE

WHEN RecencyBucket = 5 AND FrequencyBucket = 5 AND MonetaryBucket = 5 THEN 'Champions'

-- other conditions for other groups

ELSE 'Others'

END AS RFMGroup

```

* + The 'Champions' group, for example, represents customers who are recent, frequent, and high-value purchasers. These customers are crucial to the business, as they are the most engaged and provide the most revenue.

**3. Behavioral Insights and Differences Between Datasets:**

* **Differences Observed:**
  + After categorizing the customers, I analyzed the distribution of these groups across both datasets. I observed [mention any specific trends, e.g., Dataset 1 had a higher proportion of 'Champions' compared to Dataset 2, or Dataset 2 had more 'At Risk' customers].
  + These differences indicate [provide an interpretation, such as a possible difference in customer engagement or loyalty between the two datasets]. Understanding these variations can help tailor marketing strategies to address the needs of different customer segments.

1. **Technical Considerations and Optimization:**

* **Query Optimization:**
  + I ensured that my SQL queries were optimized for performance. For example, I made use of indexes on key columns such as `CustomerID` and `TransactionDate` to speed up the aggregation queries. Additionally, I avoided unnecessary subqueries and used window functions to efficiently calculate the RFM buckets.
  + These steps were crucial in handling potentially large datasets and ensuring that the queries executed quickly.

**5. Summary and Final Output:**

* **Final Deliverables:**
  + The final output includes the categorized customer data, showing the RFM group for each customer, along with aggregated results that display the number of customers in each group.
  + This allows for a clear understanding of the customer base and provides actionable insights for targeted marketing and customer retention strategies.

**Conclusion:**

* + Overall, the RFM analysis provided a robust framework for understanding customer behavior and segmenting the customer base. By focusing on Recency, Frequency, and Monetary metrics, the business can prioritize efforts on high-value and at-risk customers to maximize engagement and revenue.

### \*\*Behavioral Differences Between Customer Dataset 1 (MT) and Customer Dataset 2 (SI):\*\*

After segmenting both customer datasets based on RFM scores, the results reveal distinct differences in customer behavior between MT (Customer Dataset 1) and SI (Customer Dataset 2). Below is a detailed analysis of these behavioral differences:

# OVERAL OBSERVATION

**1. Distribution of Customer Segments:**

* **Key Observations:**
  + **Customer Base Size:** MT has a significantly larger customer base in each segment compared to SI, which aligns with the earlier analysis showing MT’s larger overall customer count.
  + **Champions and Loyal Customers:** While MT has more Champions and Loyal Customers in absolute numbers, SI’s percentages are higher, indicating a more concentrated base of high-value customers. This suggests that SI's customers are more engaged and generate higher value on average.
  + **Lost Customers:** Both businesses have a substantial proportion of Lost Customers, but MT's larger size means it has more customers in this category, which may indicate a challenge in customer retention.

**2. Monetary, Frequency, and Recency Scores:**

* **Key Observations:**
  + Champions: Both MT and SI have high scores across all RFM metrics for their Champions segment, reflecting the value and frequency of these top-tier customers. However, the fact that SI has a smaller but highly concentrated group of Champions further supports the earlier conclusion about SI having more high-value customers.
  + At Risk Segment: Both MT and SI have customers in the At Risk segment who have relatively high monetary and frequency scores but low recency scores. This indicates customers who were once active but have not made recent purchases, signaling a potential for churn.
  + Low Value and Lost Customers: Both companies show similar patterns in the Low Value and Lost Customers segments, with low scores across all RFM metrics. These segments are likely contributing less to overall revenue and may require re-engagement strategies or might be less worth investing in.

**3. Conclusion and Advice:**

* **For MT:**
  + Focus on Retention: With a large base of Lost Customers and At Risk customers, MT should prioritize customer retention strategies. Targeted marketing campaigns, reactivation offers, and customer loyalty programs could help retain these customers.
  + Evaluate Segment-Specific Strategies: Given the larger customer base and varying customer values, MT could benefit from more refined customer segmentation strategies, focusing on converting Potential Loyalists to Loyal Customers or Champions.
* **For SI:**
  + Maintain and Grow High-Value Segments: SI's smaller but more valuable customer base suggests that the company should continue nurturing its high-value customers. SI should also explore opportunities to expand its customer base with similar high-value profiles.
  + Monitor At Risk Customers: Although the At Risk segment is small, it's important for SI to monitor these customers closely and implement measures to prevent churn.

In summary, MT should focus on retention and enhancing the value of its large but less lucrative customer base, while SI should continue to nurture its high-value customers and explore opportunities for growth within that high-value segment.

# Mock Questions & Answers

*1. Q: How did you decide on the thresholds for the RFM buckets?*

- A: I used a combination of absolute values and percentiles to determine the thresholds. This approach allowed me to categorize customers in a way that was both statistically sound and meaningful for business insights.

*2. Q: Why did you choose Set X/Set Y for the customer group descriptions?*

- A: I chose Set [X/Y] because the group names in this set more intuitively described the behaviors I observed in the data. For example, 'Champions' in Set X clearly represents the most engaged and valuable customers, which aligned well with the RFM metrics I calculated.

*3. Q: What were the key differences you noticed between the two customer datasets?*

- A: I noticed that Dataset [1/2] had a higher proportion of [specific group], which suggests that [interpretation of this trend]. This difference could indicate varying levels of customer engagement or loyalty between the two datasets.

*4. Q: How did you ensure your SQL queries were efficient?*

- A: I optimized my queries by indexing key columns, using window functions for bucketing, and avoiding redundant subqueries. These techniques ensured that the analysis could be scaled to handle large datasets without performance issues.

# Technical Questions

**1. SQL-Based Transformations:**

*Q: Describe your approach to building SQL-based transformations for data pipelines.*

* + A: I focus on creating efficient and scalable SQL queries that transform raw data into meaningful insights. My approach involves thorough data profiling to understand the structure and quality of the data. I then use CTEs (Common Table Expressions) and window functions to create intermediate steps that are easy to debug and maintain. I ensure the final output is clean, well-structured, and optimized for downstream consumption, whether that be for dashboards, reporting, or feeding into machine learning models.

**2. Cloud-Based Data Warehouses:**

*Q: What is your experience with cloud-based data warehouses like BigQuery?*

* + A: I've worked with cloud-based data warehouses where I've used BigQuery to handle large datasets efficiently. My experience includes writing optimized SQL queries, managing data ingestion processes, and leveraging BigQuery’s advanced features like partitioning and clustering to improve query performance. Additionally, I’ve integrated BigQuery with other tools in the data pipeline, ensuring seamless data flow from ingestion to analysis.

**3. ELT and Data Integration:**

*Q: Can you explain how you’ve used ELT tools like Data stream or Airbyte in your previous roles?*

* + A: In my previous roles, I've used ELT tools to extract data from various sources, such as transactional databases and APIs, and load it into data warehouses. With Data stream or Airbyte, I've set up connectors to pull data efficiently and ensure it’s transformed and loaded in a way that maintains data integrity. I also monitored these pipelines for any failures or anomalies, implementing error-handling mechanisms to minimize downtime.

**4. Power BI and Dashboarding:**

*Q: How do you collaborate on Power BI-driven dashboards and reports?*

* + A: I collaborate closely with BI teams to understand the business requirements and KPIs that need to be visualized. I then ensure that the data models supporting these dashboards are well-structured and performant. I use SQL to create calculated fields and aggregate data, making it easy for BI developers to connect Power BI to the data sources. I also participate in the iterative design process to refine dashboards based on stakeholder feedback, ensuring the final product is both insightful and user-friendly.

**5. AI and Data Science Integration:**

*Q: Have you integrated SQL data pipelines with AI or data science tools?*

* + A: Yes, I’ve worked on projects where SQL-based pipelines were integrated with AI and data science tools. I’ve prepared datasets for machine learning models by performing feature engineering and data cleaning in SQL. The transformed data was then used as input for models built in Python, which were deployed to production environments. This integration allowed for real-time data-driven decisions and personalized customer experiences.

**Behavioral and Scenario-Based Questions**

**1. Cross-Functional Collaboration:**

*Q: How do you ensure effective collaboration with non-technical teams like marketing?*

* + A: I prioritize clear communication and understanding of the business goals. When collaborating with teams like marketing, I first ensure that I fully grasp their objectives and the type of insights they need. I then translate those requirements into data solutions, keeping them informed throughout the process. This approach ensures that the data outputs are aligned with their expectations and can be seamlessly integrated into their workflows.

**2. Handling Large Datasets:**

*Q: Describe a time when you optimized a query or data pipeline to handle large datasets efficiently.*

* + A: In a previous project, I had to optimize a pipeline that was processing millions of rows daily. The original query was running slowly due to lack of indexing and inefficient joins. I reviewed the query execution plan and added necessary indexes, restructured the joins to reduce the dataset size earlier in the process, and used BigQuery's partitioning and clustering features. This optimization reduced the query time by over 50%, significantly improving the overall pipeline performance.

**3. Adapting to New Tools:**

*Q: How do you approach learning and integrating new tools or technologies into your workflow?*

* + A: I have a proactive approach to learning new tools. I start by exploring the official documentation and tutorials to understand the core functionalities. Then, I experiment with small, non-critical tasks to get hands-on experience. I also engage with online communities and forums to learn best practices and common pitfalls. This approach ensures that I can quickly adapt and effectively integrate new tools into my workflow.

**4. Commercial Focus:**

*Q: Can you provide an example of how you've applied a commercial focus in your data engineering work?*

* + A: In a project where we were analyzing customer purchasing patterns, I focused not just on the technical implementation but also on how the insights could drive business outcomes. By identifying key customer segments and their purchasing behaviors, we were able to tailor marketing strategies that led to a 15% increase in conversion rates. My commercial focus ensured that the data engineering efforts translated into tangible business benefits.

**Questions You Can Ask**

**1. On Data Challenges:**

* + What are the biggest data challenges Silvertreebrands is currently facing, especially in terms of integrating data from multiple brands like UCOOK and Faithful to Nature?

**2. On AI Integration:**

* + How does Silvertreebrands plan to leverage AI and machine learning in the near future? Are there specific areas where you see SQL-driven data pipelines playing a critical role in supporting these initiatives?

**3. On Collaboration:**

* + How closely does the SQL/Data Engineering team collaborate with the AI and data science teams? Can you provide an example of a cross-functional project that successfully utilized both teams?