**RFM Analysis Documentation**

**1. Project Overview**

**Objective:**

The objective of this project was to perform an RFM (Recency, Frequency, Monetary) analysis on customer transaction data to segment customers based on their purchasing behavior. This analysis helps identify customer segments such as "Champions," "Loyal Customers," "At Risk," etc., for targeted marketing and customer retention strategies.

**2. Data Loading**

**Datasets Used:**

* Customer\_Dataset\_1
* Customer\_Dataset\_2

**SQL Server Database:**

The datasets were loaded into the `*CustomerAnalysisDB*` database in MSSQL Server.

Tables Created:

```sql

CREATE TABLE [dbo].[Customer\_Dataset\_1](

[transaction\_id] [int] NOT NULL,

[customer\_id] [int] NOT NULL,

[transact\_date] [date] NOT NULL,

[revenue] [smallint] NOT NULL,

[discount\_amount] [smallint] NOT NULL

) ON [PRIMARY];

CREATE TABLE [dbo].[Customer\_Dataset\_2](

[transaction\_id] [int] NOT NULL,

NOT NULL,

[transact\_date] [date] NOT NULL,

[revenue] [smallint] NOT NULL

) ON [PRIMARY];

```

**Data Import:**

Data was imported into the above tables using MSSQL Server's data import tools.

**3. RFM Calculation**

**Recency:**

Calculated as the difference between the most recent transaction date and the analysis date.

```sql

SELECT

customer\_id,

DATEDIFF(DAY, MAX(transact\_date), @AnalysisDate) AS Recency

FROM

dbo.Customer\_Dataset\_1

GROUP BY

customer\_id;

```

**Frequency:**

Calculated as the total number of transactions made by each customer.

```sql

SELECT

customer\_id,

COUNT(transaction\_id) AS Frequency

FROM

dbo.Customer\_Dataset\_1

GROUP BY

customer\_id;

```

**Monetary:**

Calculated as the total revenue generated by each customer.

```sql

SELECT

customer\_id,

SUM(revenue) AS Monetary

FROM

dbo.Customer\_Dataset\_1

GROUP BY

customer\_id;

```

**4. RFM Scoring**

**Scoring Method:**

Customers were scored on a scale of 1 to 5 for each of Recency, Frequency, and Monetary values using the `NTILE` function.

NB: NTILE() is a window function that distributes rows of an ordered partition into a specified number of approximately equal groups, or buckets. It assigns each group a bucket number starting from one. For each row in a group, the NTILE() function assigns a bucket number representing the group to which the row belongs.

```sql

SELECT

customer\_id,

NTILE(5) OVER (ORDER BY Recency) AS RecencyScore,

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

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

FROM

RFM\_Table\_Customer\_Dataset\_1;

```

**5. Customer Segmentation**

**Segmentation Logic:**

Customers were segmented into different groups based on their RFM scores.

```sql

SELECT

customer\_id,

RecencyScore,

FrequencyScore,

MonetaryScore,

CASE

WHEN RecencyScore = 5 AND FrequencyScore = 5 AND MonetaryScore = 5 THEN 'Champions'

WHEN RecencyScore >= 4 AND FrequencyScore >= 3 AND MonetaryScore >= 3 THEN 'Potential Loyalists'

WHEN FrequencyScore <= 2 AND MonetaryScore >= 4 THEN 'Big Spenders'

ELSE 'Others'

END AS Segment

FROM

RFM\_Scores\_Customer\_Dataset\_1;

```

**Segment Definitions:**

* **Champions:** Highest RFM scores across all three metrics.
* **Potential Loyalists:** High Recency and Frequency, medium to high Monetary.
* **Big Spenders:** Low Frequency but high Monetary.
* **Others**: Customers who don’t fit into the above categories.

**6. Advanced Analysis**

**Custom Segmentation**:

Additional segments were defined using more specific criteria (e.g., "New Customers").

**Weighted Scoring:**

Weights were applied to the RFM scores to emphasize certain aspects like Monetary value.

```sql

SELECT

customer\_id,

(0.5 \* RecencyScore) + (0.3 \* FrequencyScore) + (0.2 \* MonetaryScore) AS WeightedRFMScore

FROM

RFM\_Scores\_Customer\_Dataset\_1;

```

**7. Automation**

**Stored Procedure:**

The entire RFM calculation and segmentation process was encapsulated in a stored procedure for automation.

```sql

CREATE PROCEDURE dbo.CalculateRFM

AS

BEGIN

-- Insert the detailed RFM calculation and segmentation logic here

INSERT INTO dbo.Customer\_Segmentation

SELECT

customer\_id,

RecencyScore,

FrequencyScore,

MonetaryScore,

Segment

FROM

RFM\_Scores\_Customer\_Dataset\_1;

END

GO

```

**Execution:**

The stored procedure was executed to refresh the analysis periodically.

```sql

EXEC dbo.CalculateRFM;

```

**8. Visualization**

**Tools Used:**

- **Power BI:** To create visual dashboards for:

* RFM score distribution.
* Customer segment analysis.
* Revenue per segment.

**Example Visualizations:**

* Heatmap: Showing customer distribution across RFM scores.
* Pie Chart: Displaying segment distribution.
* Bar Chart: Comparing average revenue per segment.

**9. Conclusion**

This document outlines the full process of performing an RFM analysis, from data loading to customer segmentation and visualization. The results can be used to drive targeted marketing efforts, improve customer retention, and enhance overall business strategy.

**Additional Questions:**

1. **Explain why you set the thresholds to the values that you did.**

**My Approach:**

* I used the `NTILE(5)` function to divide the customers into five buckets (percentiles) for each of the RFM metrics (Recency, Frequency, and Monetary).
* This results in scores ranging from 1 to 5, where 5 typically represents the most favorable score for each metric (for example, most recent for Recency, most frequent for Frequency, and highest spending for Monetary).

**Explanation:**

* Reason for Using NTILE(5): The decision to use NTILE(5) is rooted in the desire to segment the customer base into quintiles, which provides a balanced distribution of customers across the RFM scores.
* This method is effective for analyzing large datasets because it ensures each segment contains a roughly equal number of customers.
* Thresholds Rationale: By using NTILE(5), you inherently establish thresholds based on the distribution of the data. This means that the thresholds for each segment are dynamically determined by the data, allowing for a more data-driven segmentation process.

1. **Say which set(s) of customer group descriptions you decided to use and explain your reasoning for why you picked that set of descriptions.**

**My Approach:**

* I have opted for a set of customer group descriptions that closely align with Set X from the instructions. Examples include 'Champions', 'Loyal Customers', 'At Risk', and 'Lost Customers'.

**Explanation:**

* Reason for Choosing Set X: Set X is a more conventional set of customer segments that are commonly used in RFM analysis. These segments are intuitive and widely recognized in business contexts, making them easier to interpret and act upon.
* Alignment with Business Needs: These labels offer clear insights into customer behavior, which is essential for devising targeted marketing strategies. For instance, 'Champions' represent your best customers who need to be rewarded,
* while 'At Risk' customers may require re-engagement strategies.

1. **Do you observe any behavioral differences between Customer Dataset 1 and Customer Dataset 2? Please explain.**

**a.Customer Distribution (Customer Count):**

* Customer Dataset 1 -> has significantly higher customer counts across all segments compared to 'Customer Dataset 2'. For instance, the "Others" segment has 664 customers in Dataset 1, whereas Dataset 2 has only 76.
* The "Loyal Customers" and "Champions" segments also show a noticeable difference, with Dataset 1 having 159 and 57 customers respectively, while Dataset 2 has only 27 and 15.
* The "Potential Loyalists" segment is particularly striking, with 88 customers in Dataset 1 but only 2 in Dataset 2.

**b. Segment Composition and Scores:**

* Both datasets have similar AvgMonetaryScore, AvgFrequencyScore, and AvgRecencyScore across corresponding segments, indicating similar spending, purchasing frequency, and recency behavior within each segment.
* However, there are minor variations, such as the "Potential Loyalists" in Dataset 2 having an AvgMonetaryScore of 2 compared to 1 in Dataset 1.

**c. Segment Presence:**

* Both datasets include the same segments, but the customer distribution within these segments is vastly different. This could imply different stages of customer engagement or the effectiveness of retention strategies in different periods or customer cohorts.

**d. Behavioral Insights:**

* Dataset 1 has a larger and more diverse customer base with a significant number of "Loyal Customers" and "Champions." This might indicate better customer engagement and retention.
* Dataset 2 shows a more condensed customer base, with fewer customers across all segments. The sharp reduction in the "Potential Loyalists" segment could suggest challenges in converting new or borderline customers into loyal ones.

**Summary:**

* Dataset 1 likely represents a broader and more established customer base with strong loyalty and engagement, whereas "Dataset 2"could represent a smaller, perhaps newer, or more segmented customer base with less overall engagement and loyalty.