# Data Transformation Plan for RFM Analysis

## Business need

Business need identified by the marketing team:

- Identify four segments of customers. These segments should include:
    - High-value customers: customers who spend a lot of money and are frequent customers
    - Loyal customers: customers who make frequent purchases, even if they don't spend the most money
    - At-risk customers: customers who have not made a purchase recently
    - Persuadable customers: customers who have made a recent purchase in any amount
- Use the customer segments to better target ads aimed at increasing customer loyalty, spending, and frequency of purchases.

## Data

Summary of the data that will be used and the timeframe for the data:

- The data we'll use is the TheLook eCommerce online store dataset.
- The fields we will need include the customer IDs, dates of purchase, products bought, and total amounts spent.
- The timeframe for this analysis is customer behavior in 2022.
- Before beginning the transformation, we'll clean the data to make sure the data is consistent, errors are removed, and any missing information is filled in.

## Data transformation

Summary of the definitions and logic used to transform the data:

- The RFM score ranks customers based on recency, frequency, and monetary value.
    - Recency is defined as how recently a customer made a purchase of any size.
    - Frequency is defined as how often the customers made a purchase within the established timeframe.
    - Monetary value is defined as how much the customers spent on all purchases over the established timeframe.
- To calculate the RFM score, we'll rank the customers based on recency, frequency, and monetary value.
- Rankings will then be used to sort the customers into four segments based on the following conditions:
    - High-value customers are customers with highest monetary value and frequency scores.
    - Loyal customers are customers with high-frequency scores but lower monetary value scores.
    - At-risk customers are the customers with the lowest recency scores.
    - Persuadable customers are the customers with the highest recency scores who do not fit into other groups.

Summary of the transformation techniques that will be used:

- Data aggregation:
    - Summarizes multiple orders and items into single rows per customer
    - Calculates recency, frequency, and monetary value
- Derivation:
    - Creates new recency, frequency, monetary, and quantiles columns
- Joining:

    - Combines data from orders and order items tables
- Common Table Expressions (CTEs):
    - Organizes query logic into temporary result sets for clarity and reuse
- Window functions:
    - Calculates quantiles and divides data into groups for each customer's metrics
- Conditional logic:
    - Assigns customer segments based on specific rules involving derived values

## Tools

Based on availability and transformation needs, these are the proposed tools for the analysis:

- BigQuery
    - Stores the dataset for each access
    - Provides a SQL workspace
- SQL
    - Used to write the step-by-step data transformation steps
    - Useful for data querying and exploration

## Testing

Proposed testing procedure to ensure data quality:

- Double-check calculations