RFM Cohort Analysis

Project Overview

The purpose of this project is to **develop and implement an RFM and Cohort Analysis system** that segments customers and evaluates their retention trends.
This system helps businesses understand **customer value, purchase frequency, and loyalty patterns**, enabling **data-driven marketing and retention strategies**.

The project includes:

- Python-based implementation of **RFM segmentation** and **Cohort retention analysis**,
- Detailed data preprocessing and feature engineering steps,
- Visualization of customer behavior and retention trends, and
- Key insights to improve customer engagement and business performance.

Dataset Description

The dataset contains **transactional sales records** from an online retail business. Each record corresponds to an individual item purchased by a customer in a specific invoice. The data is used to measure **recency (last purchase date)**, **frequency (number of purchases)**, and **monetary value (total spend)** for each customer, as well as to build **monthly cohorts** for retention analysis.

Key Columns and Their Meanings

Column Name	Description	Example	Role in Analysis
InvoiceNo	A unique identifier for each transaction (invoice). Canceled transactions may have a prefix like “C”.	536365	Used to identify unique purchase events.
StockCode	Product (item) code.	85123A	Helps in tracking which items were purchased.
Description	Name/description of the product.	WHITE HANGING HEART T-LIGHT HOLDER	Used for exploratory analysis or product-level aggregation.

Quantity	Number of units of the product purchased in that transaction.	6	Used in calculating total purchase value.
InvoiceDate	Date and time of the transaction.	2010-12-01 08:26:00	Used for calculating Recency and forming Cohorts.
UnitPrice	Price per unit of the product.	2.55	Used to compute monetary value.
CustomerID	Unique ID assigned to each customer.	17850	Used for grouping transactions by customer.
Country	Customer's country of residence.	United Kingdom	Useful for geographical segmentation or filtering.

Derived Features

New Column	Description	Formula/Logic
TotalAmount	Total value of a single transaction.	Quantity × UnitPrice
InvoiceMonth	Month and year of the transaction.	Extracted from InvoiceDate
CohortMonth	Month of the customer's first purchase (cohort assignment).	Minimum InvoiceMonth per CustomerID
Recency	Days since last purchase (based on snapshot date).	SnapshotDate - LastPurchaseDate
Frequency	Total number of unique invoices per customer.	Count of distinct InvoiceNo
Monetary	Total amount spent by the customer.	Sum of TotalAmount

Data Preprocessing Steps

1. **Handling Missing Values**
    - Dropped rows with missing CustomerID (since RFM requires customer-level tracking).
    - Checked for missing values in InvoiceDate, Quantity, and UnitPrice.
2. **Removing Invalid Transactions**
    - Excluded transactions where:
        - Quantity <= 0 (returns or cancellations).
        - UnitPrice <= 0 (invalid or promotional records).
    - Dropped invoices with prefix "C" (canceled orders).
3. **Feature Engineering**
    - Created **TotalAmount** = Quantity × UnitPrice.
    - Converted InvoiceDate into datetime format for time-based calculations.
    - Extracted **InvoiceMonth** and **CohortMonth** for cohort analysis.
4. **Reference Date Definition**
    - Choose a **snapshot date** (usually one day after the last transaction in the dataset).
    - Used this date to compute **Recency** (days since last purchase).
5. **Data Validation**
    - Verified total transaction count before and after cleaning.
    - Ensured customer IDs were unique and consistent.
    - Checked for outliers in Quantity and TotalAmount columns.

Methodology

A. RFM Analysis

Step 1: Metric Calculation

- **Recency (R):** Number of days since the customer's last purchase.
- **Frequency (F):** Total number of purchases made by the customer.

- **Monetary (M):** Total amount spent by the customer.

CustomerID	Recency	Frequency	MonetaryValue
12346.0	326	1	77183.60
12347.0	2	182	4310.00
12348.0	75	31	1797.24
12349.0	19	73	1757.55
12350.0	310	17	334.40

Step 2: Scoring
Each RFM variable was scored on a **1–5 scale**, where:

- Higher **R** score = more recent purchase.
- Higher **F** score = more frequent buyer.
- Higher **M** score = higher spender.

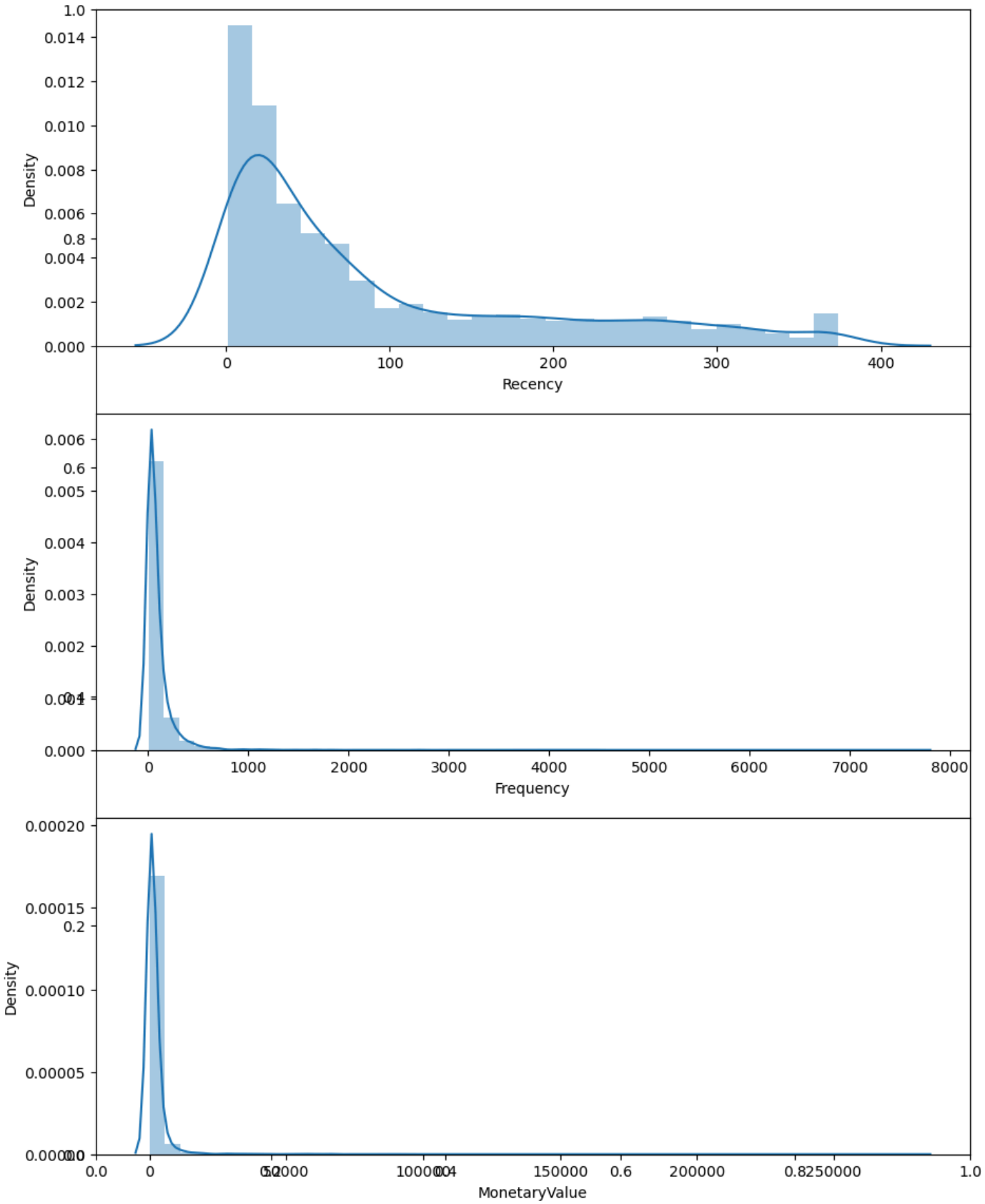
Step 3: Segmentation
Customers were classified into segments such as:

Segment	Description
Champions	Recent, frequent, high spenders.
Loyal Customers	Frequent buyers with moderate spending.
At Risk	Previously active but haven't purchased recently.
New Customers	Recently joined, low frequency.
Lost Customers	No purchases in a long time.

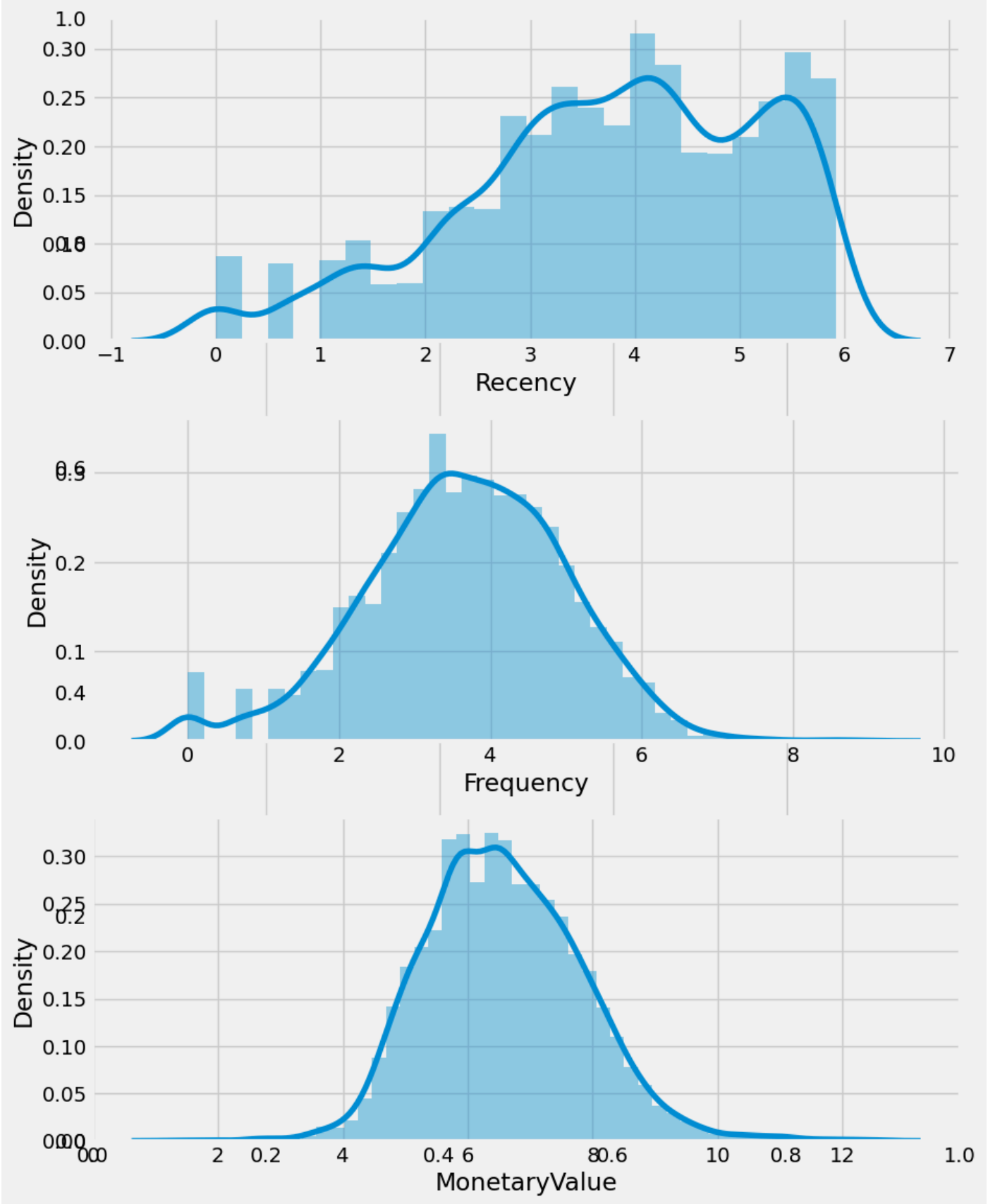
Step 4: Visualization

- Distribution plots for R, F, and M values.
- Segment-level heatmaps showing revenue contributions.

Assymmetric distribution of variables (data skewed)



Logarithmic transformation (positive values only) will manage skewness



B. Cohort Analysis

Step 1: Cohort Creation

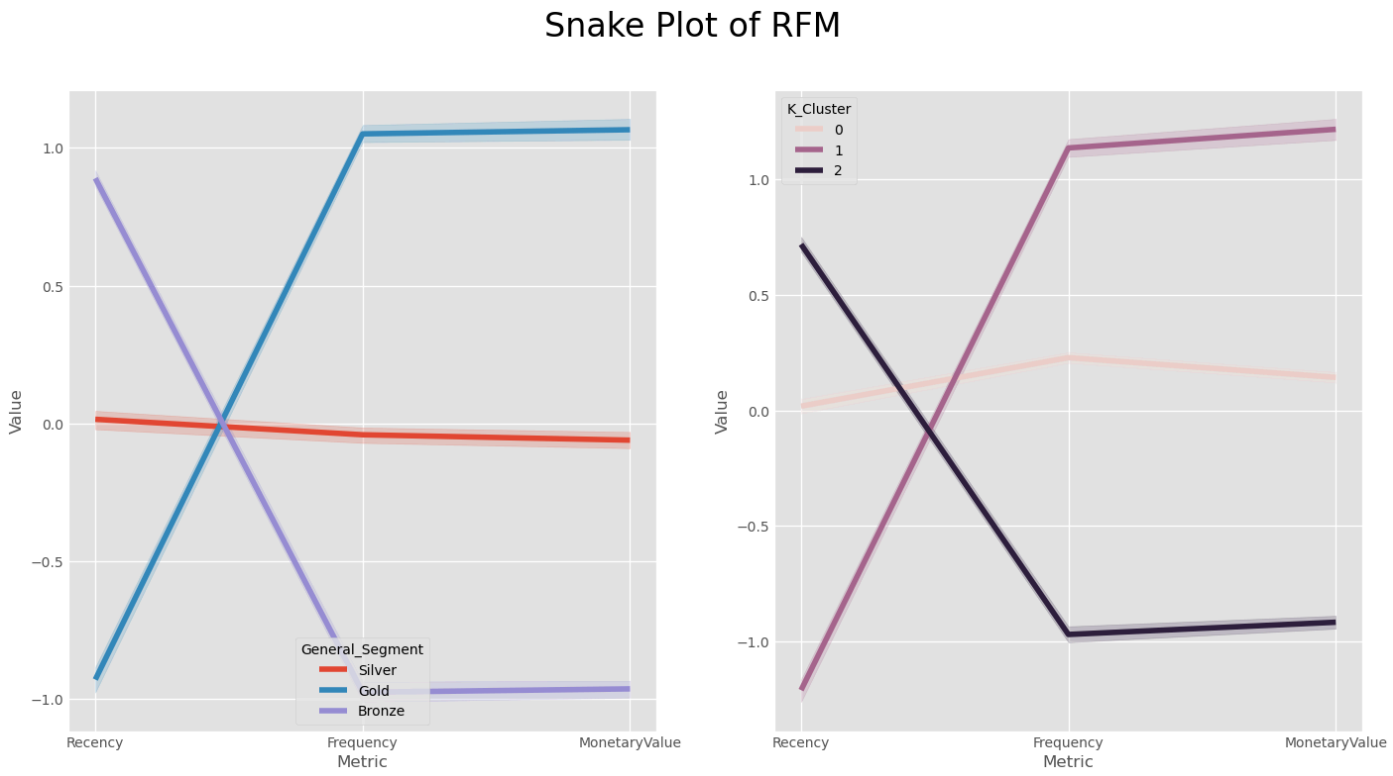
- Cohorts were formed based on each customer's **first purchase month**.

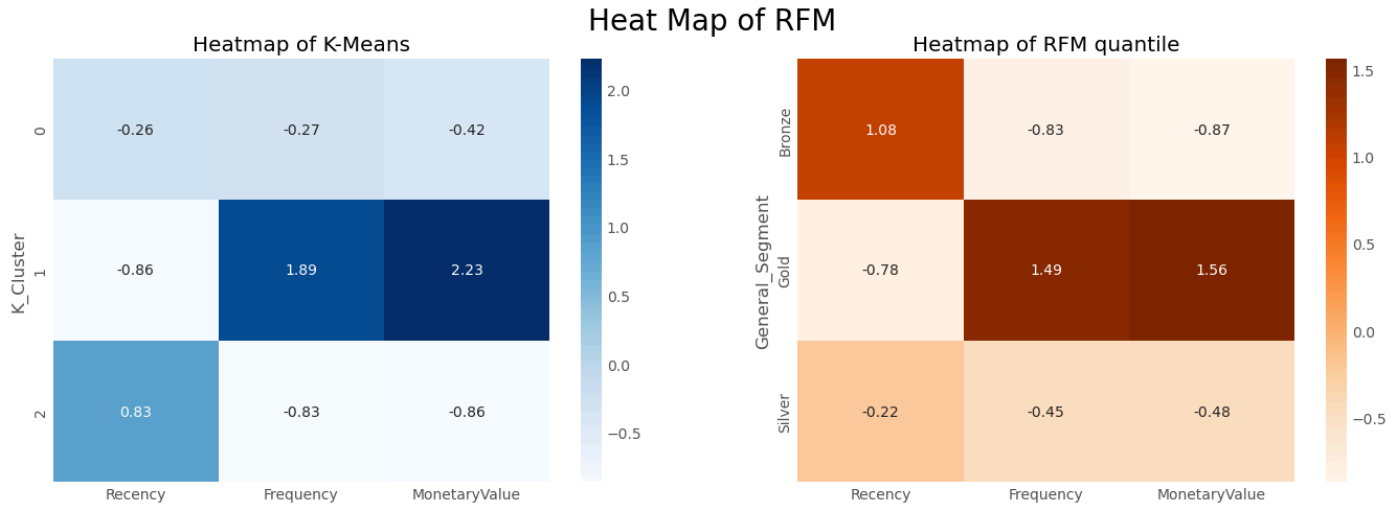
Step 2: Retention Matrix

- Calculated **retention rate** = (Customers retained in a given period) / (Customers in cohort's first month).
- Created a retention table showing customer return behavior over months.

Step 3: Visualization

- **Cohort heatmap**: Highlights how retention decreases over time.
- **Cohort curve**: Shows the trend of customer retention by cohort.





Data Analysis Implementation

The analysis was implemented in Python using the following libraries:

- **Pandas** – data manipulation
- **NumPy** – numerical calculations
- **Matplotlib / Seaborn** – visualization
- **Datetime** – date transformations
- **Jupyter Notebook** – interactive coding environment

Key Implementation Steps:

1. Imported and cleaned the dataset.
2. Conducted exploratory data analysis (EDA) to understand purchase patterns.
3. Computed RFM metrics and segmented customers.
4. Built Cohort retention matrix and generated heatmaps.
5. Derived insights from visualizations.

Results and Insights

RFM Segmentation Insights

- **Champions** (top 20% customers) contribute the majority of revenue.
- **At Risk** customers represent a significant recovery opportunity — previously high-spending but inactive recently.
- **Loyal Customers** exhibit consistent purchase behavior with medium-to-high monetary values.
- **Hibernating and Lost Customers** show long inactivity and minimal engagement, indicating potential churn.

Cohort Analysis Insights

- Strong **Month 0 (initial) retention**, followed by typical drop-off after 2–3 months.
- Gradual **improvement in newer cohorts**, suggesting improved onboarding or marketing.
- The **majority of repeat purchases** occur within the first 60 days after acquisition.

Visual Findings

- RFM histograms show skewness toward low recency (many inactive customers).
- Cohort heatmaps reveal retention patterns by acquisition month.
- Revenue concentration visualizations highlight the 80/20 pattern (Pareto principle).

Business Implications

- **Retention Strategy:**
  Target *At Risk* and *Potential Loyalist* customers with personalized reactivation campaigns.
- **Customer Value Maximization:**
  Offer loyalty rewards to *Champions* and *Loyal Customers* to strengthen brand attachment.
- **Churn Prevention:**
  Use early indicators (high recency + low frequency) to identify likely churners.
- **Acquisition vs. Retention Optimization:**
  Cohort analysis helps evaluate whether new acquisition strategies are improving long-term retention.

Conclusion

The developed RFM and Cohort Analysis system successfully:

- Quantifies customer behavior using RFM scoring,
- Segments customers for targeted marketing, and
- Tracks retention trends over time using cohort analysis.

This framework helps businesses:

- Prioritize high-value customers,
- Design data-driven retention programs, and
- Optimize marketing expenditure through segmentation-based strategies.

The combination of **RFM metrics and cohort-based retention tracking** gives a comprehensive view of customer lifecycle value, enabling **proactive decision-making and long-term growth**.