# Unsupervised ML: Capstone Project

## Online Retail Customer Segmentation


**Team Members**

Keerthana Reddy
Bhanu Pratap Shahi

# Contents

- Introduction
- Problem Statement
- Data Preview
- Data Summary
- EDA Insights
- RFM Modeling
- Clustering
- Conclusion

# Introduction

Customer segmentation is a vital process of grouping customers based on shared behaviors or attributes. The goal is to identify a high-value customer base with significant growth potential. Insights gained inform tailored marketing campaigns and overall strategy. The decision to segment and the chosen criteria depend on the company's philosophy and offerings, shaping business operations and strategy formulation.

# Problem Statement

**Overview:**

The project focuses on identifying major customer segments within a UK-based non-store online retail's transactional dataset. Specializing in unique all-occasion gifts, the company serves both individual customers and wholesalers.

**Objective:**

Utilizing the RFM (Recency, Frequency, Monetary) model, the primary objective is to delineate distinct customer segments based on transactional data. By uncovering meaningful clusters, the project aims to understand purchasing behaviors comprehensively. Through RFM-based segmentation, the goal is to enable targeted marketing, personalized experiences, and informed decision-making, ultimately enhancing the ability to address key business questions.

# Data Preview

- **InvoiceNo:** A 6-digit integral number uniquely assigned to each transaction. 'c' at the beginning denotes a cancellation.
- **StockCode:** A 5-digit integral number unique to each item.
- **Description:** Item name.
- **Quantity:** Quantity of each product per transaction.
- **InvoiceDate:** Date and time of each transaction.
- **UnitPrice:** Product price per unit.
- **CustomerID:** A 5-digit integral number unique to each customer.
- **Country:** Name of the country where each customer resides.

# Data Summary

- The dataset contains 8 columns and 541909 rows.
- There are some nan values as well. E.g. Null % in Description column is approx 31%.

Object Summary:

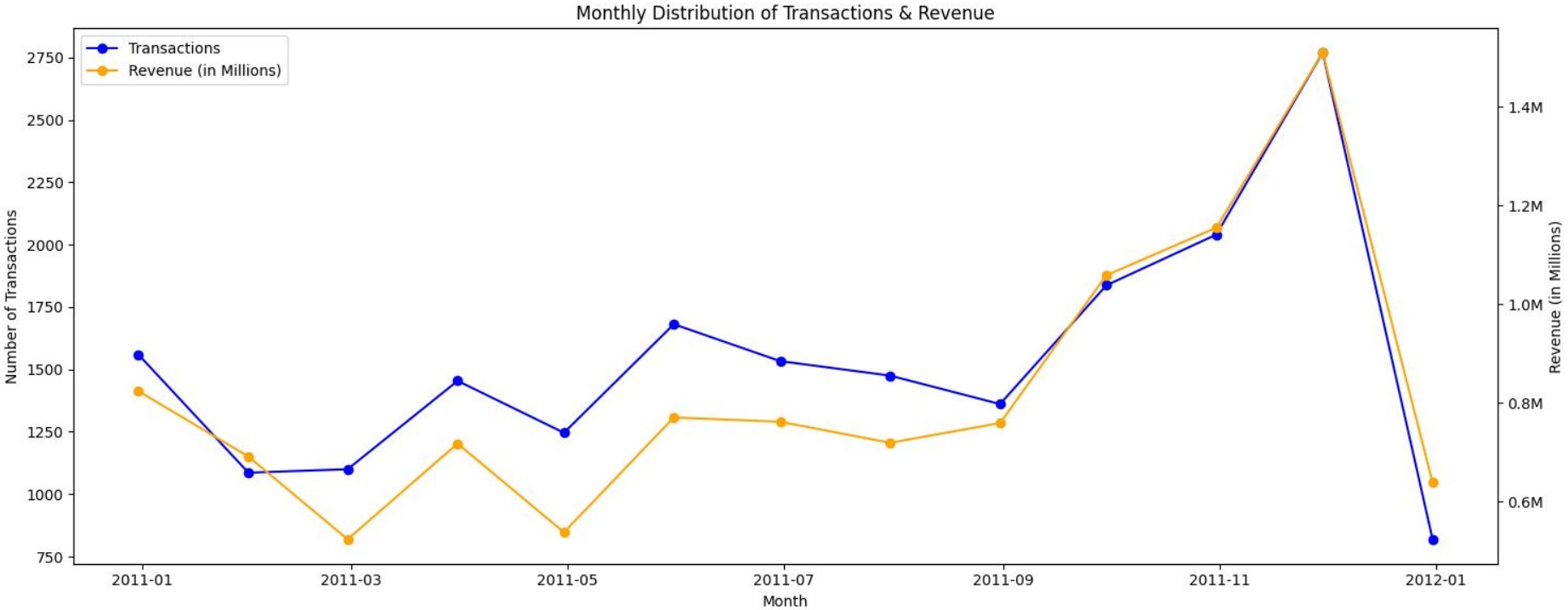
| | InvoiceNo | StockCode | Description | Country |
|---|---|---|---|---|
| count | 541909 | 541909 | 540455 | 541909 |
| unique | 25900 | 4070 | 4223 | 38 |
| top | 573585 | 85123A | WHITE HANGING HEART T-LIGHT HOLDER | United Kingdom |
| freq | 1114 | 2313 | 2369 | 495478 |

Numeric Summary:

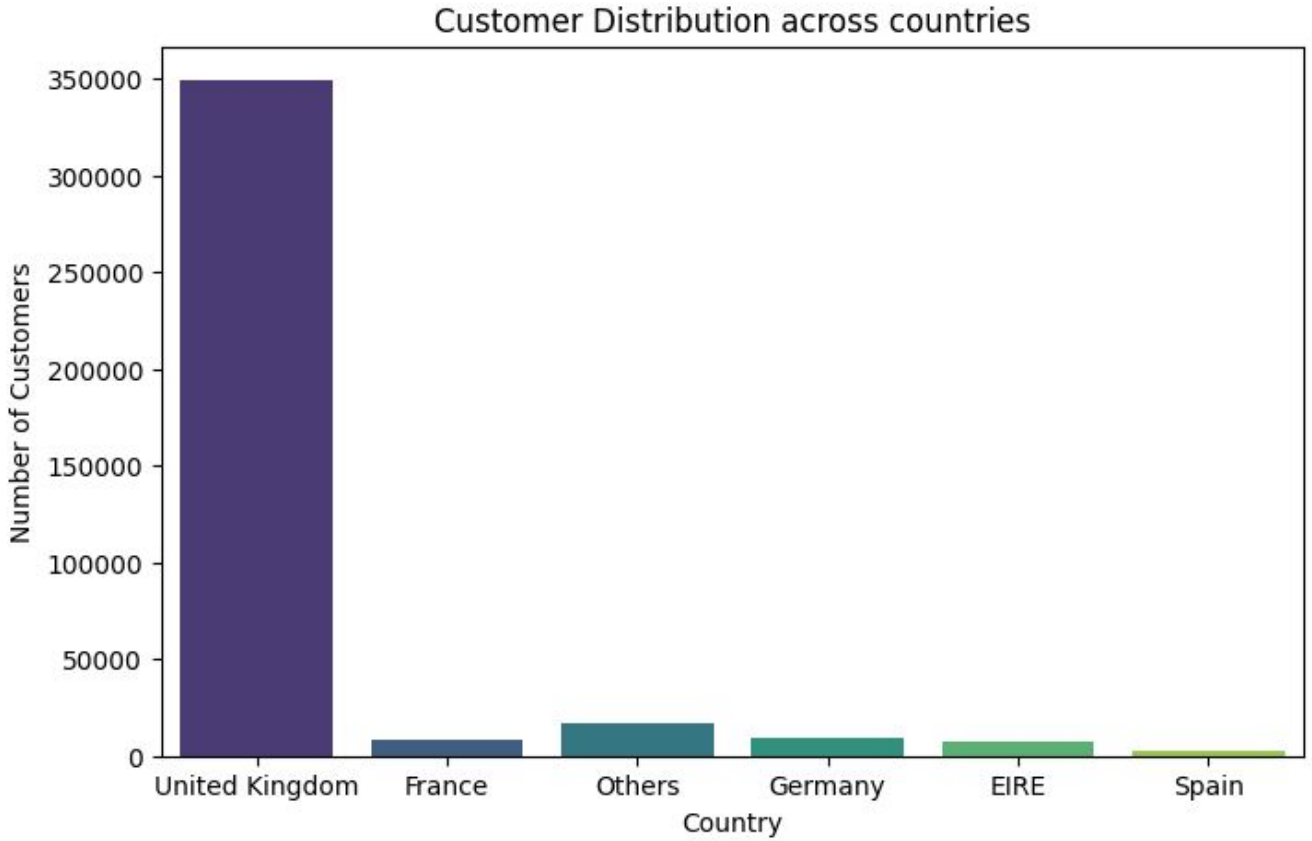
| | Quantity | UnitPrice | CustomerID |
|---|---|---|---|
| count | 541909.000000 | 541909.000000 | 406829.000000 |
| mean | 9.552250 | 4.611114 | 15287.690570 |
| std | 218.081158 | 96.759853 | 1713.600303 |
| min | -80995.000000 | -11062.060000 | 12346.000000 |
| 25% | 1.000000 | 1.250000 | 13953.000000 |
| 50% | 3.000000 | 2.080000 | 15152.000000 |
| 75% | 10.000000 | 4.130000 | 16791.000000 |
| max | 80995.000000 | 38970.000000 | 18287.000000 |

# EDA Insights

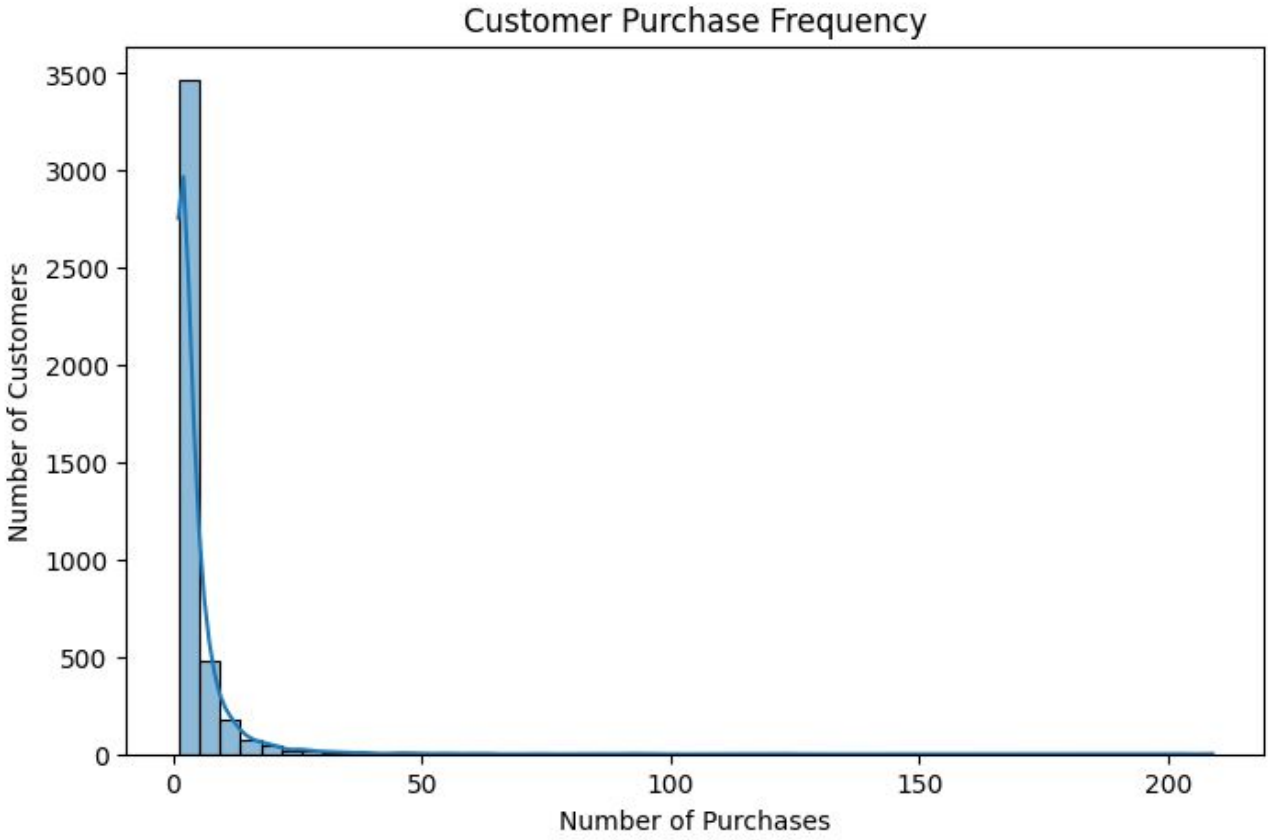
## Distribution of Transactions & Monthly Revenue Over Time



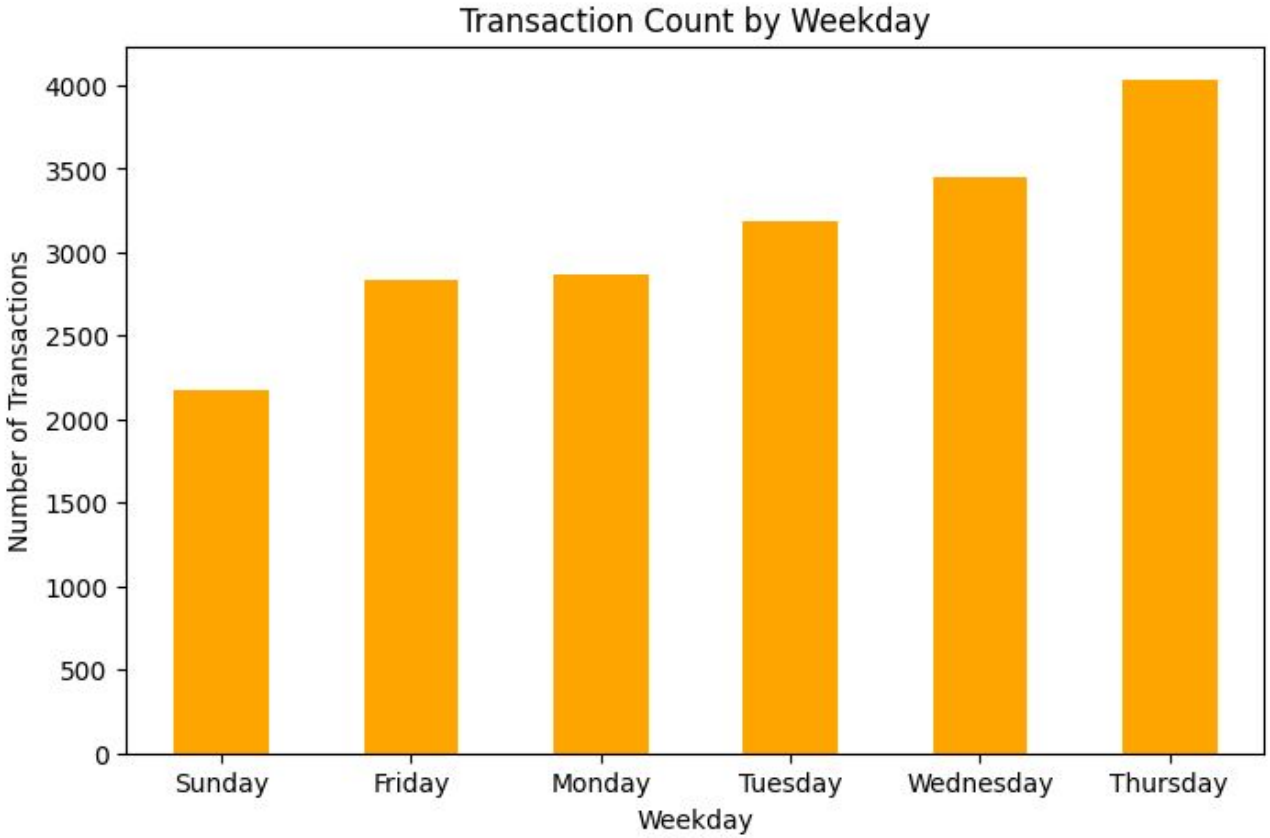
# Distribution of Customers Across Countries



# Customer Purchase Frequency



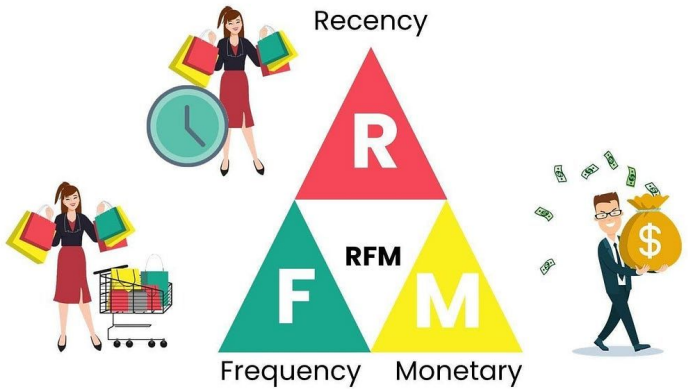
# Distribution of transaction by Weekday



# RFM Modelling

**Definition:** RFM, or Recency, Frequency, and Monetary, is a powerful technique in customer segmentation. It assigns scores to each customer based on the recency of their last transaction, the frequency of transactions in the last year, and the monetary value of those transactions.

**Significance:** RFM analysis addresses key questions: Who is our most recent customer? How frequently do they make purchases? What is the total value of their transactions? This information is critical for evaluating a customer's impact on the company, aiding in strategic decision-making.



# Feature Engineering for RFM Modelling

1. Create Revenue Feature:Calculate revenue by multiplying Quantity and Unit Price.
2. GroupBy Customer ID for RFM Metrics: Obtain Recency, Frequency, and Monetary values by grouping data based on Customer ID.
3. Calculate Recency (R):Extract the number of days from the last purchase date to the maximum date available.
4. Quantile-Based Scoring (1 to 4): Assign quantile scores (1 to 4) to Recency, Frequency, and Monetary values individually.

| **Quantiles --->** | **0** | **0.25** | **0.5** | **0.75** |
|---|---|---|---|---|
| Recency | 0 | 18 | 51 | 142 |
| Frequency | 0 | 1 | 2 | 5 |
| Monetary | 0 | 306 | 668 | 1660 |

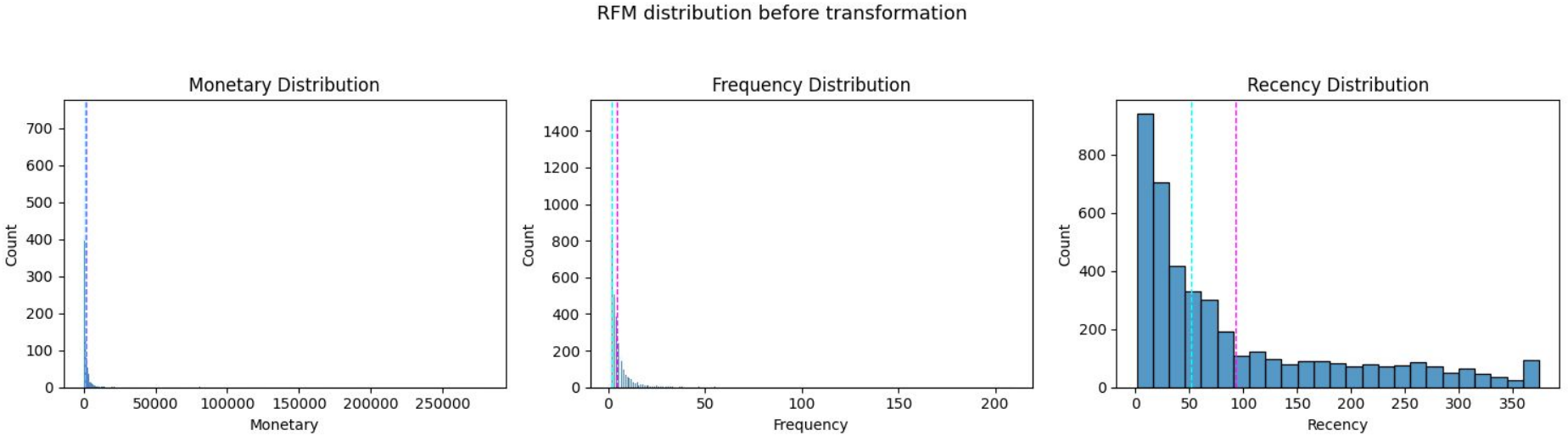
# Feature Engineering for RFM Modelling

5. Calculate Total RFM Score: Sum the quantile scores to get the total RFM score. Directly these scores can be used for segmentation. Segment customers based on total RFM scores into Low-Risk/High-Value; Medium-Risk/Medium-Value; and High-Risk/Low-Value segments.

6. Using R, F, M values directly with clustering techniques

1. **K-Means** - K-Means is a centroid-based clustering algorithm that partitions data into K clusters. It minimizes the sum of squared distances between data points and their assigned cluster centroids.
2. **Hierarchical Clustering** - Hierarchical clustering builds a tree of clusters, where each node represents a cluster. It can be agglomerative (bottom-up) or divisive (top-down).
3. **DBSCAN** - DBSCAN identifies clusters based on dense regions separated by sparser areas. It is effective at discovering clusters of arbitrary shapes.

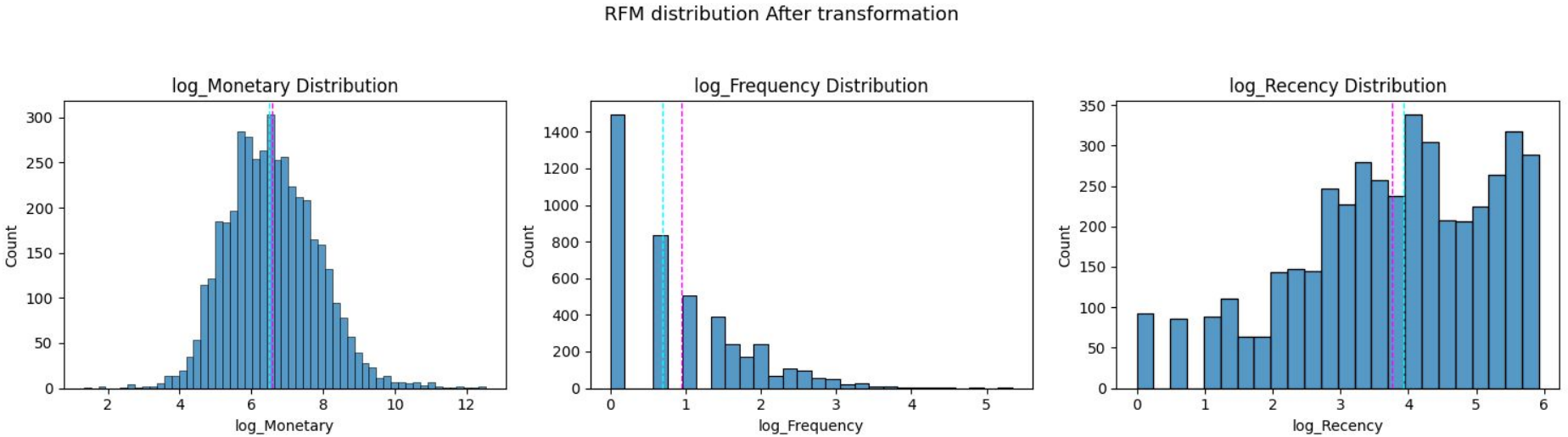
# RFM Modelling




**By seeing above plots we can say that Recency, Frequency and Monetary values are skewed & needs some transformation.**

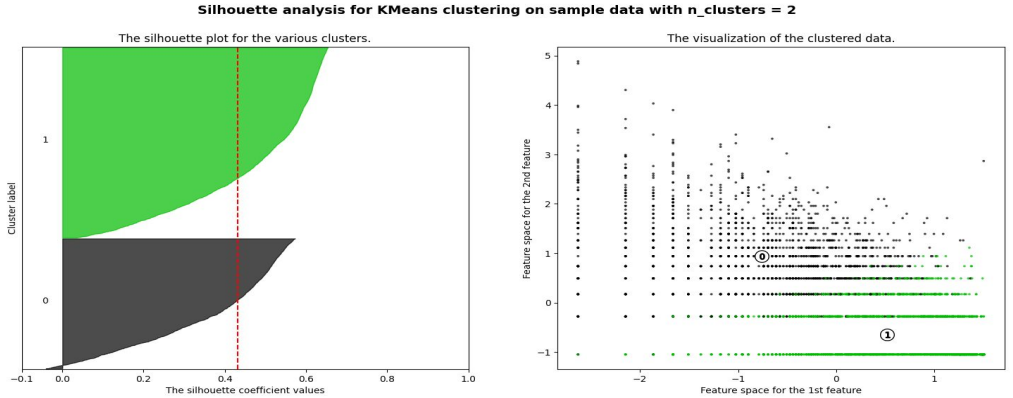
# RFM Modelling

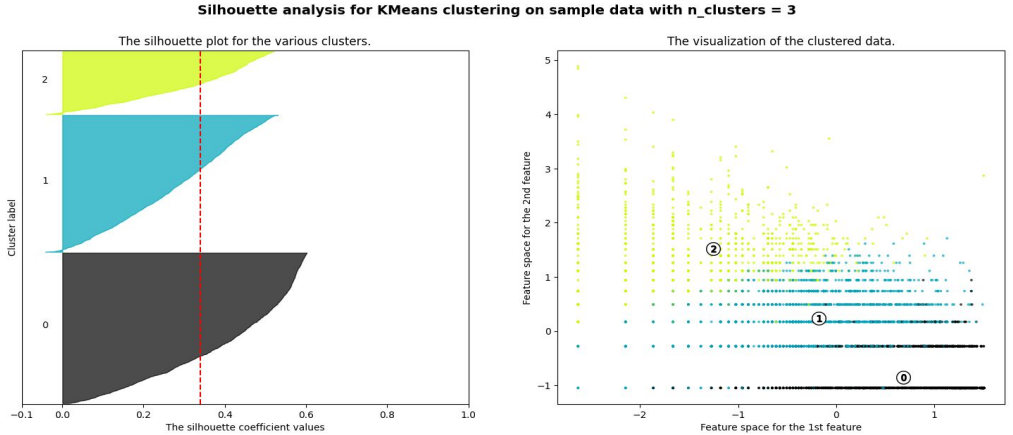
## Reduced skewness by applying log transformation to Recency, Frequency & Monetary values



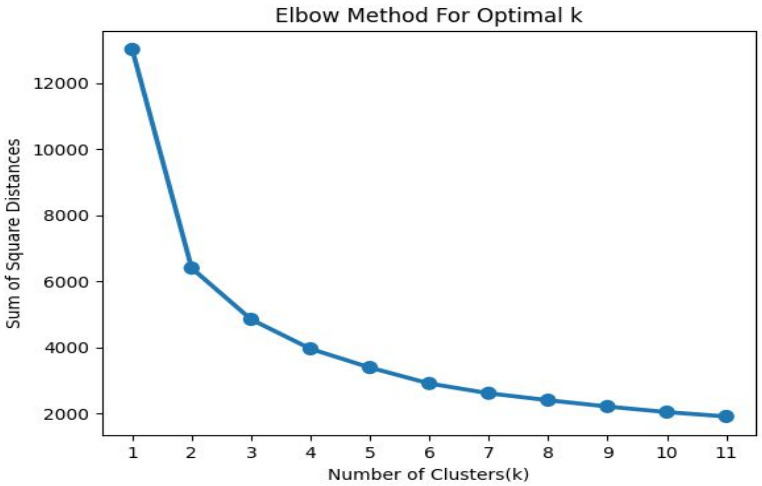
# Clustering

**K-Means with silhouette score at n_clusters = 2 (0.43) & n_clusters = 3 (0.33)**

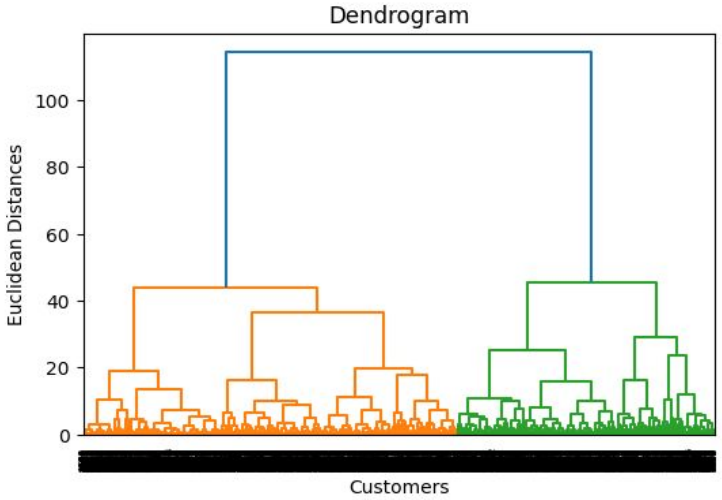


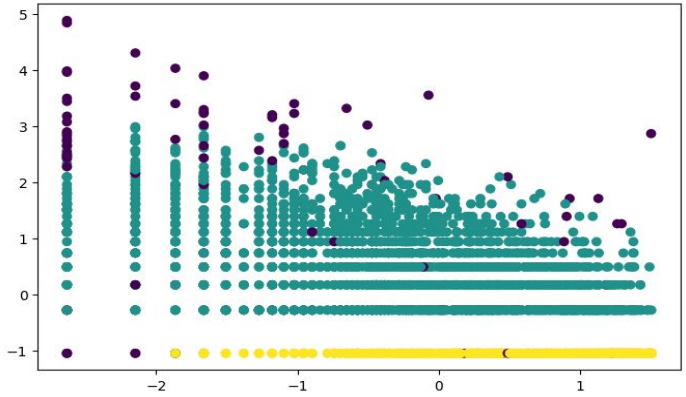
## K-Means with Elbow Method




## Hierarchical Clustering



**DBSCAN** > DBSCAN is a clustering algorithm identifying dense regions in data. It's robust to noise, excels in various shapes, and automatically determines cluster count, making it valuable for various applications.




## Overall Results

| SL No. | Model Name | No.of Clusters |
|---|---|---|
| 1 | K-Means with Silhouette score | 2 |
| 2 | K-Means with Elbow method | 2 |
| 3 | Hierarchical Clustering | 2 |
| 4 | DBSCAN | 3 |

# Conclusion

The **"Online Retail Customer Segmentation"** project has successfully categorized customers based on Recency, Frequency, and Monetary values, enabling targeted marketing strategies. By identifying high-value segments, the business can optimize resource allocation and enhance customer experiences.

After RFM modelling we have got the below customer segments.

**1. High-Value Segment:**

**Objective**: Retain and maximize value.

**Strategies**: Offer personalized promotions or loyalty programs. Provide exclusive access to premium products/services. Gather feedback to enhance their experience.

**2. Churn-Prone Segment:**

**Objective**: Prevent churn and re-engage.

**Strategies**: Implement targeted marketing campaigns to re-engage. Provide incentives or discounts for repeat purchases. Understand reasons for potential churn through surveys.

**3. Medium-Value or Average Segment**

**Objective**: Increase engagement and loyalty.

**Strategies**: Introduce loyalty programs or tiered rewards. Cross-sell or upsell relevant products/services. Monitor customer feedback and address concerns.


THANK YOU!