CUSTOMER SEGMENTATION OF ONLINE RETAIL USING RFM ANALYSIS AND K-MEANS CLUSTERING

BY HARISH MUHAMMAD

Overview

Content of this presentation

Framework

Business understanding

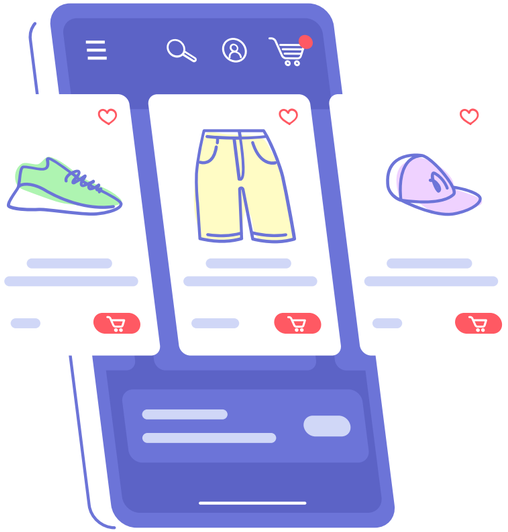
Data understanding & Data Cleansing

Exploratory Data Analysis

Developing model & Evaluation

Insights & Recommendations

Customer centric to win in high competitive environment



Online retail or e-commerce operates in a **highly competitive environment.**

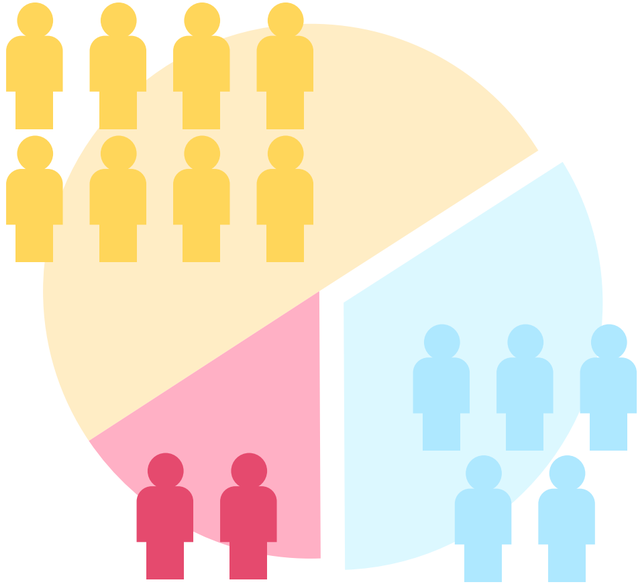
Customer-centric can be the way to win the completion

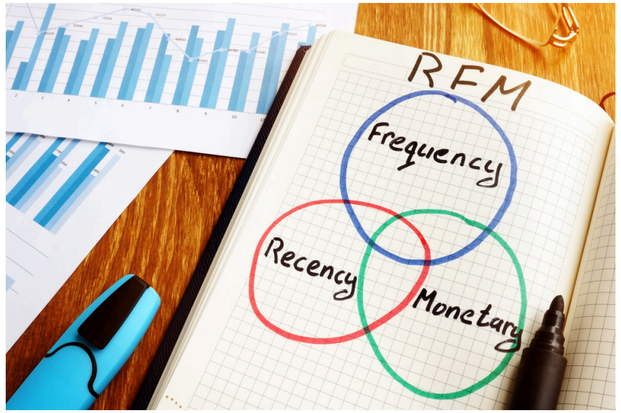
- Understanding customers' needs, perceptions, and expectations is crucial for stronger customer relationships & better business outcomes

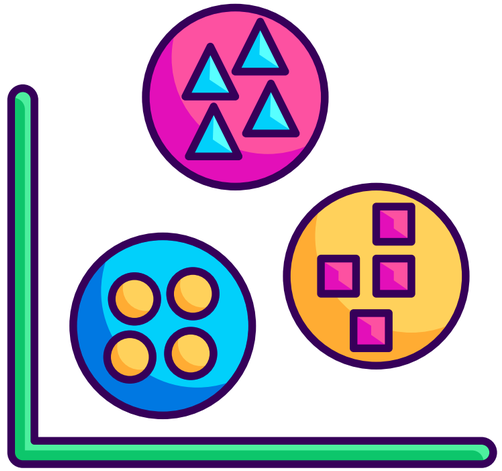
Challange: Vast amounts of data:

- purchase history,
- transactional data,
- demographic information
- etc.

Customer segmentation to identify and prioritize high-value customers







Customer Segmentation
crucial for online retail because it allows businesses to understand and cater to the diverse needs of their customers.

Leveraging data science techniques:

- **RFM analysis**
- Machine learning: **K-means clustering**

can help to separate and identify which segment to focus in generating revenue

Customer analytics and segmentation goals

Goal 01

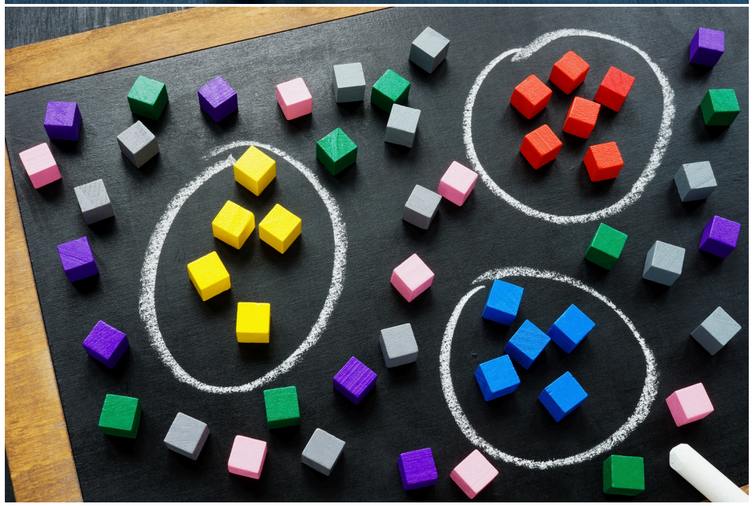
Gaining an understanding of how many products are sold and revenue

Goal 02

Developing segmentation models that categorizes customers into distinct segments

Goal 03

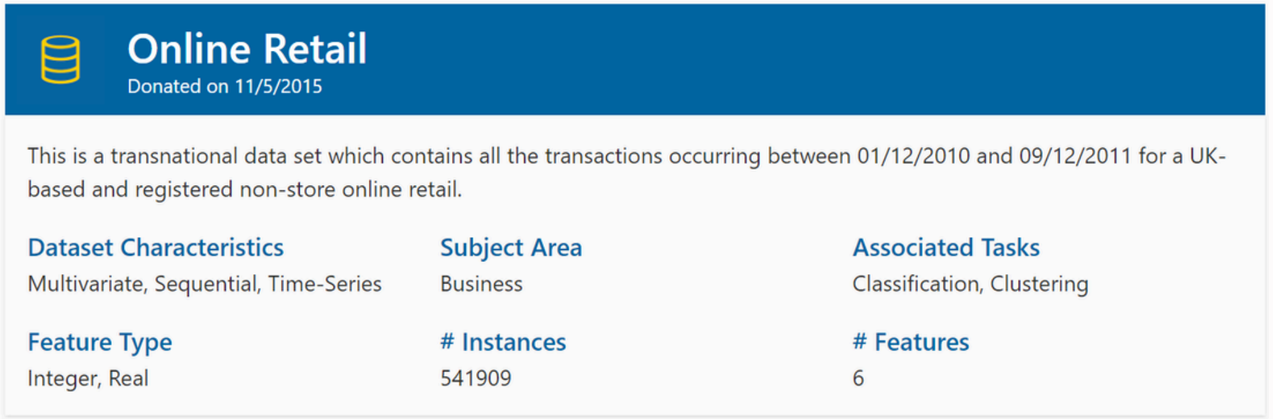
Providing insights and recommendations for each segment



Dataset source: UK-based online retail 2010-2011

Scientific paper

Also available on



Data Understanding

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 541909 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  datetime64[ns]
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: datetime64[ns](1), float64(2), int64(1), object(4)
memory usage: 33.1+ MB
```

Total data: 541909
Numerical columns: 3
Categorical columns: 4

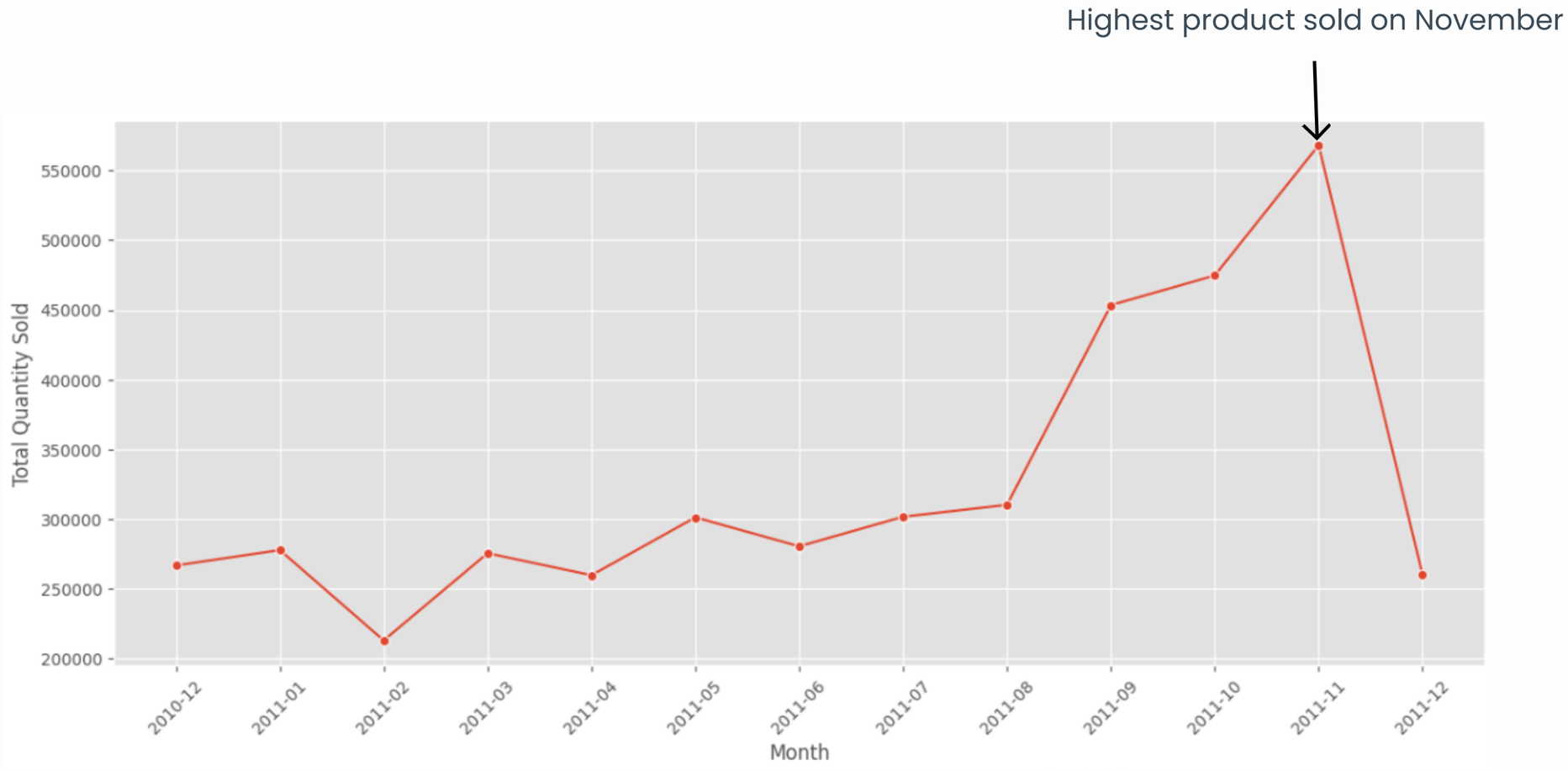
Data distribution:
Not normally distributed

Data Cleaning

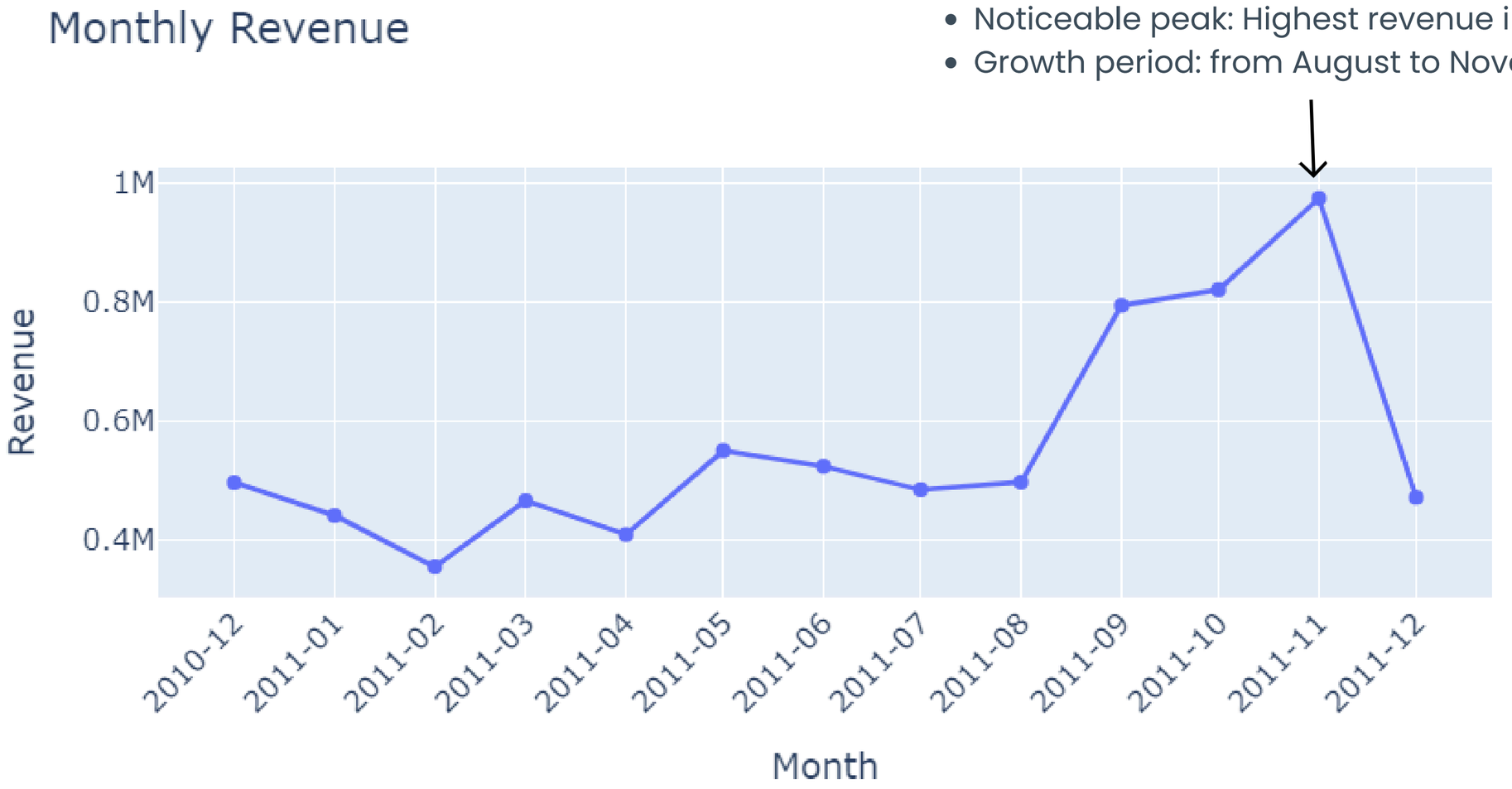
- **Handling duplicates**
- **Handling missing values**
- **Correcting data type format**
- **Removing irrelevant features**
- **Handling anomalies**
- **Handling outliers**

Missing values in a column: 27%
Duplicated data: 1%
Outliers: 5.8-8.3%

Total product sold per month overtime



Total revenue overtime



- Noticeable peak: Highest revenue in November
- Growth period: from August to November

RFM Analysis

Recency

Measures the time elapsed since a customer last made a purchase

Frequency

Assesses how often a customer makes a purchase.

Monetary

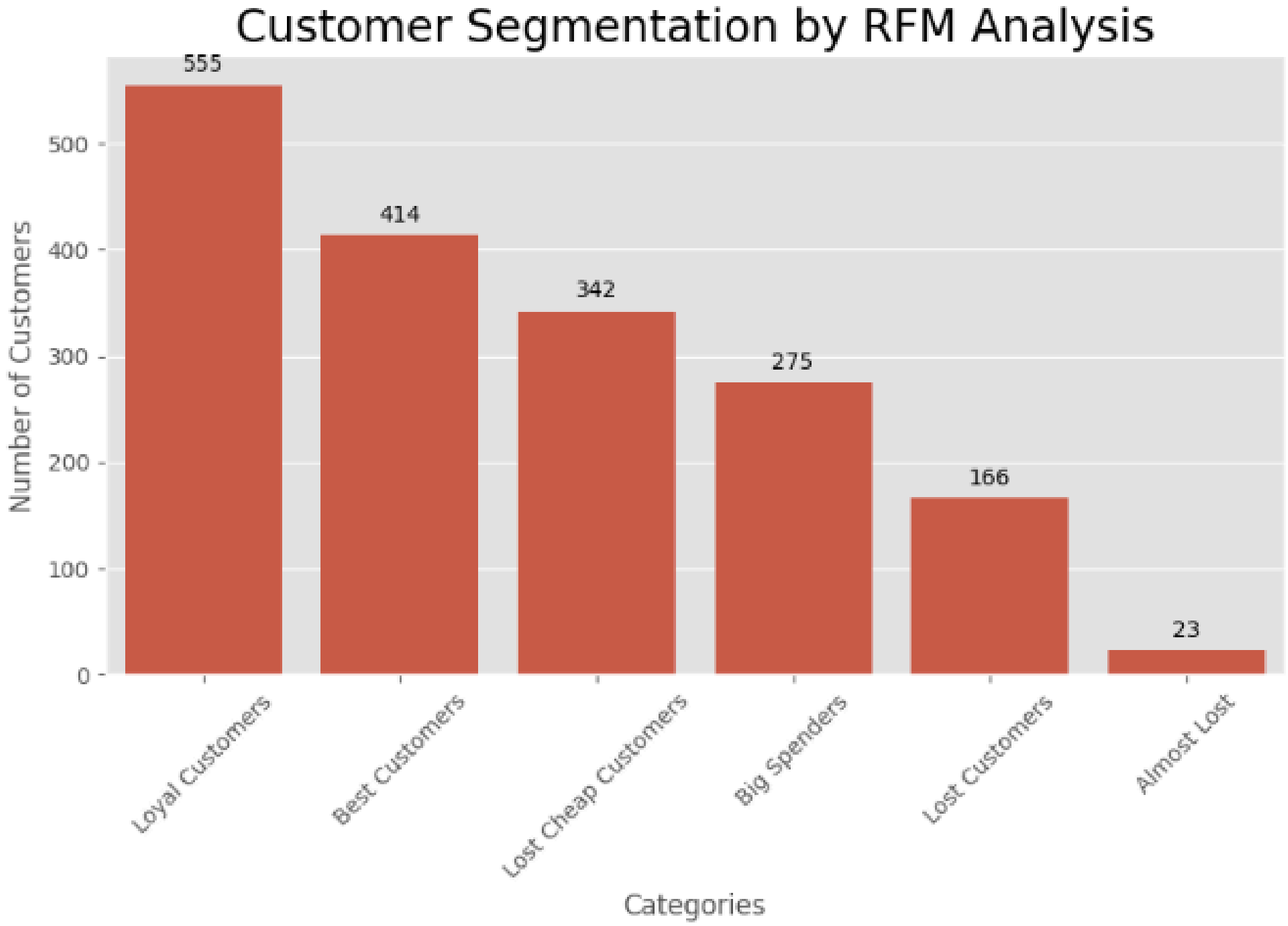
Calculates the total amount of money a customer has spent

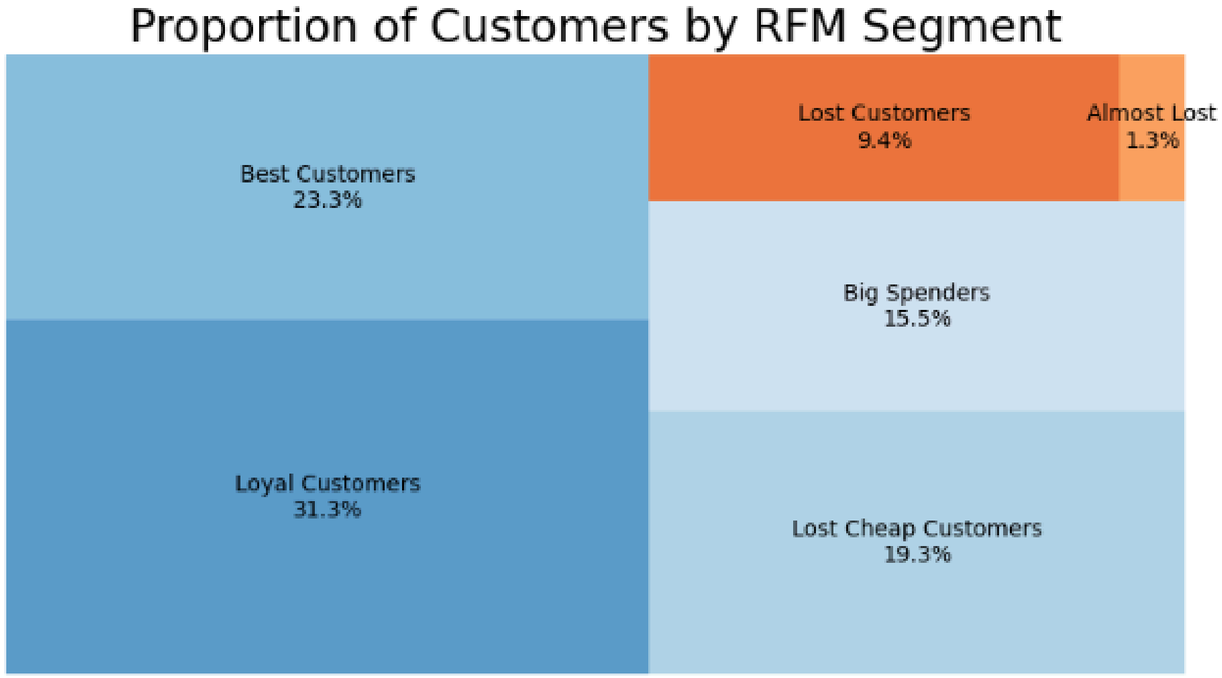
Data Modelling: RFM Quantiles

- RFM metrics are segmented based on quantiles
- Each metric (recency, frequency, monetary) is assigned a score from 1 to 4, (1 for the highest value, 4 for the lowest value)
- RFM score (overall) is calculated by combining individual RFM score.

Segment	RFM Score
Best Customers	111
Loyal Customers	F = 1
Big Spenders	M = 1
Almost lost	134
Lost Customers	344
Lost Cheap Customers	444

Customer segmentation by RFM





Data Modelling: K-means Clustering

Optimum conditions for K-means clustering:

Normal/symmetric distribution

Minimized outliers

Low Skewness

Standardised data

Data distribution & skewness check:

Data are not normally distributed and highly skewed

Feature	D'Agostino-Pearson Statistic	P-value	Distribution	Skewness	Skewness Type
recency	583.200928	2.288377e-127	Not Normally Distributed	1.163469	Right Skew
frequency	589.373112	1.045334e-128	Not Normally Distributed	1.170132	Right Skew
monetary	589.807769	8.411440e-129	Not Normally Distributed	1.173867	Right Skew

Data pre-processing: Outlier handling, Transformation, Standardisation

Handling outliers

Winsorization

Scaling

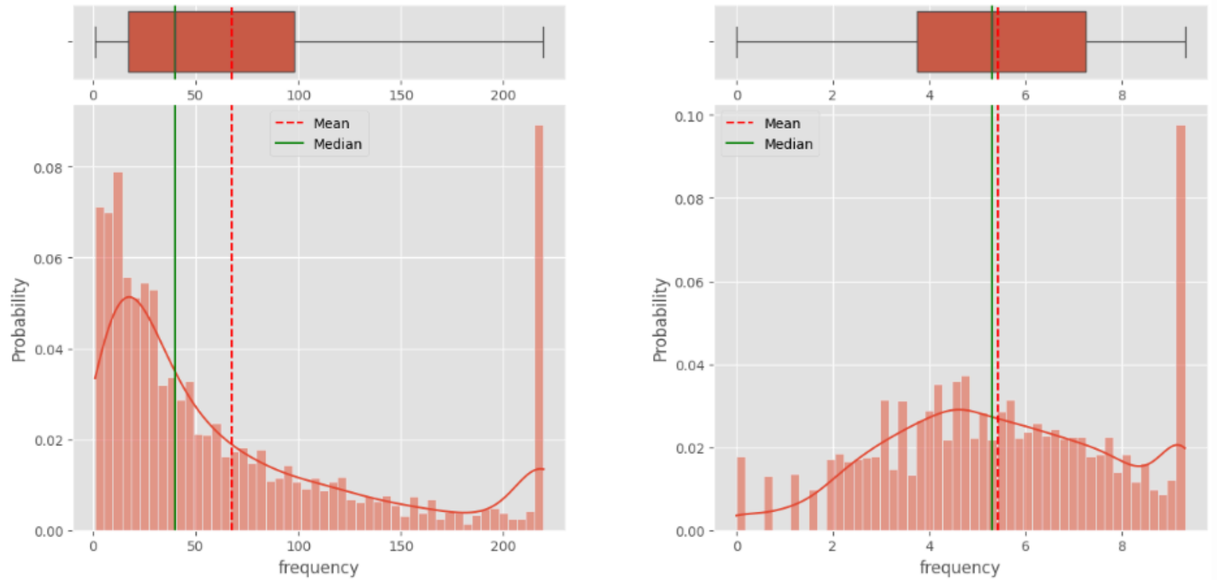
Robust Scaler

Transformation

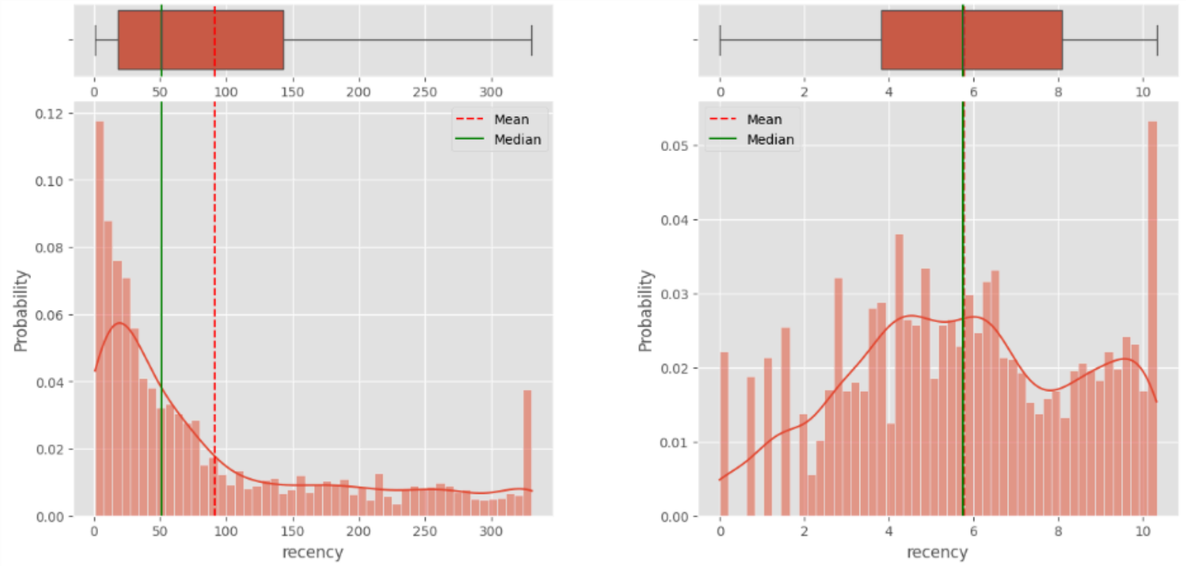
Box Cox - method

Before & After:
Data Transformation & scaling

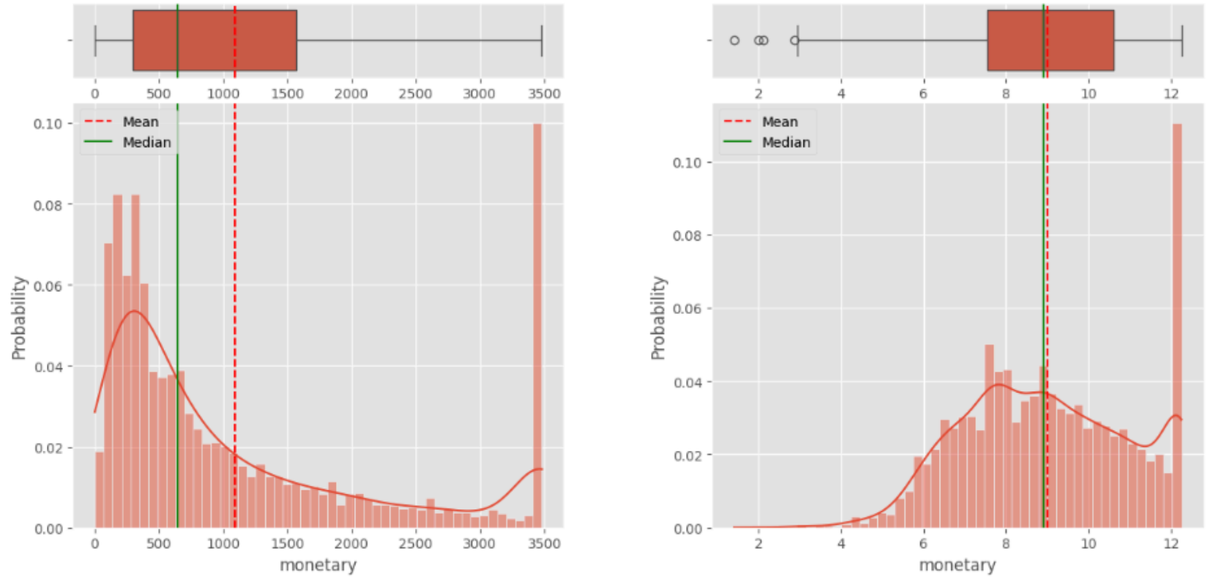
Frequency → Transformed Frequency



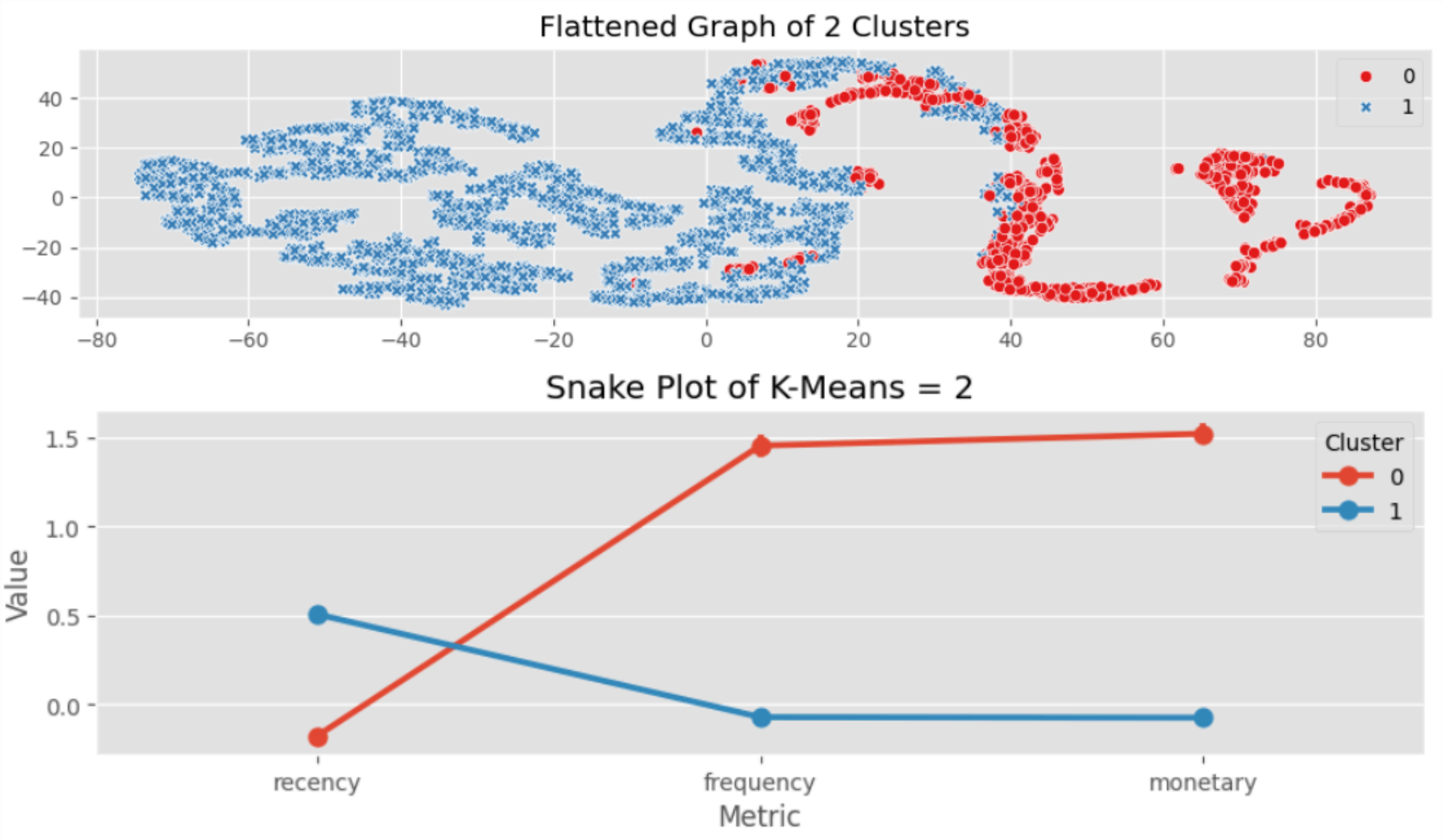
Recency → Transformed Recency



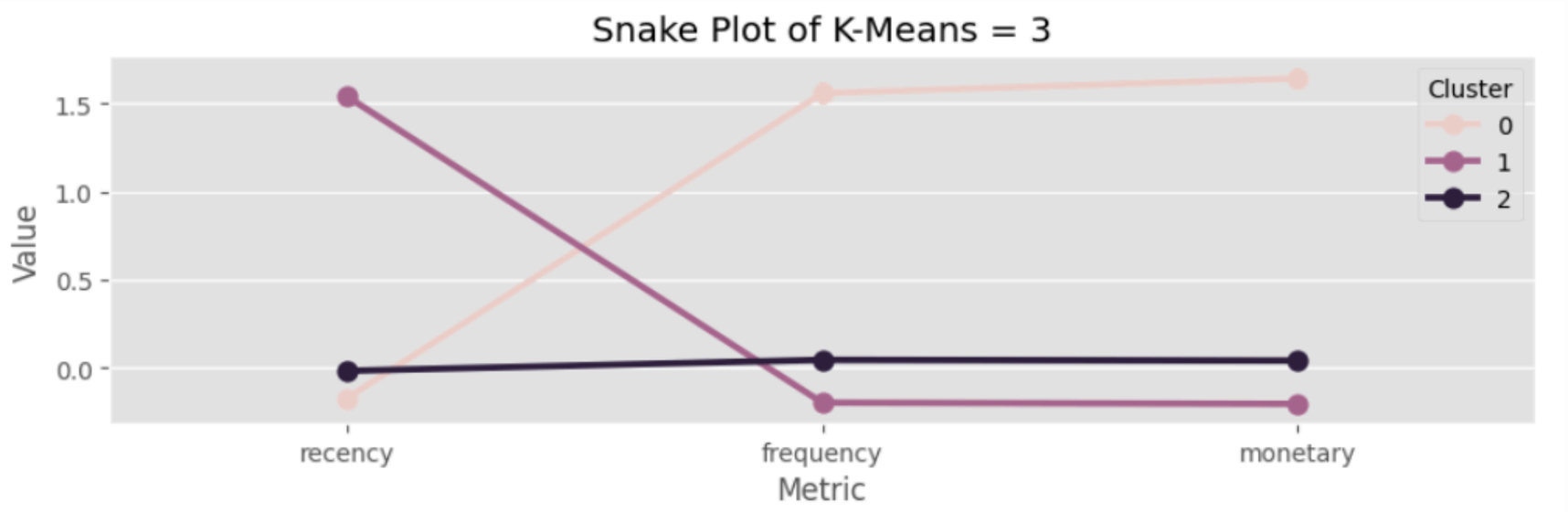
Monetary → Transformed Monetary



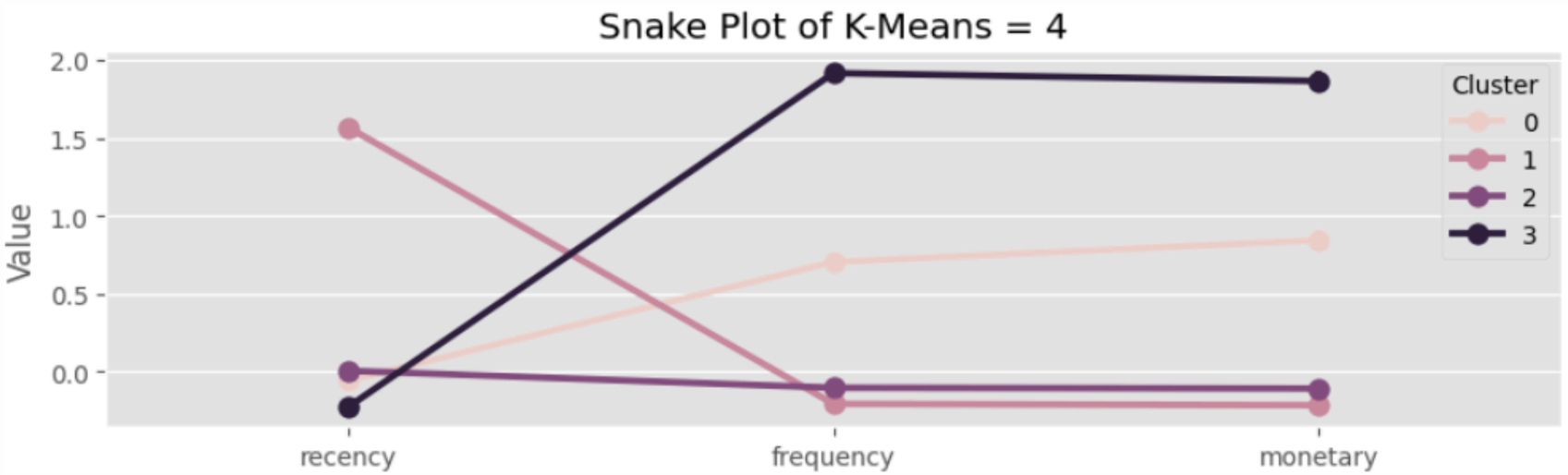
Data Modelling: K-means (Cluster = 2)



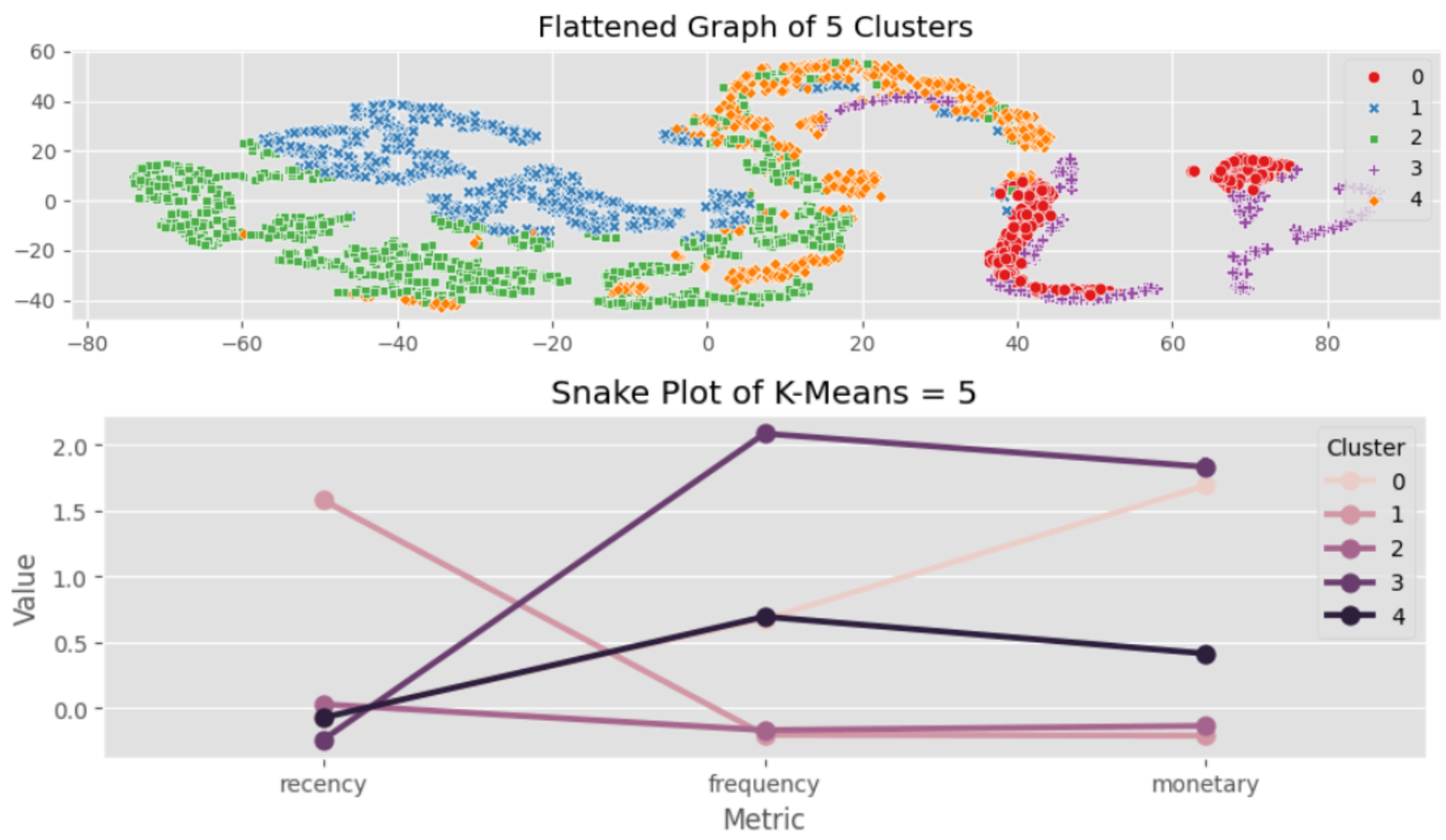
Data Modelling: K-means (Cluster = 3)



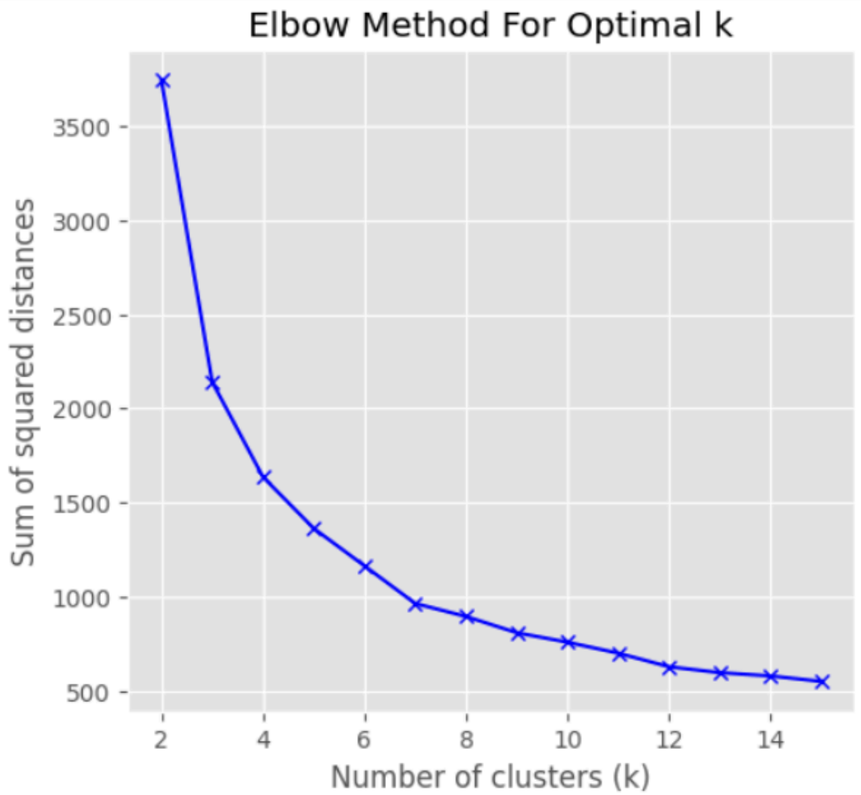
Data Modelling: K-means (Cluster = 4)

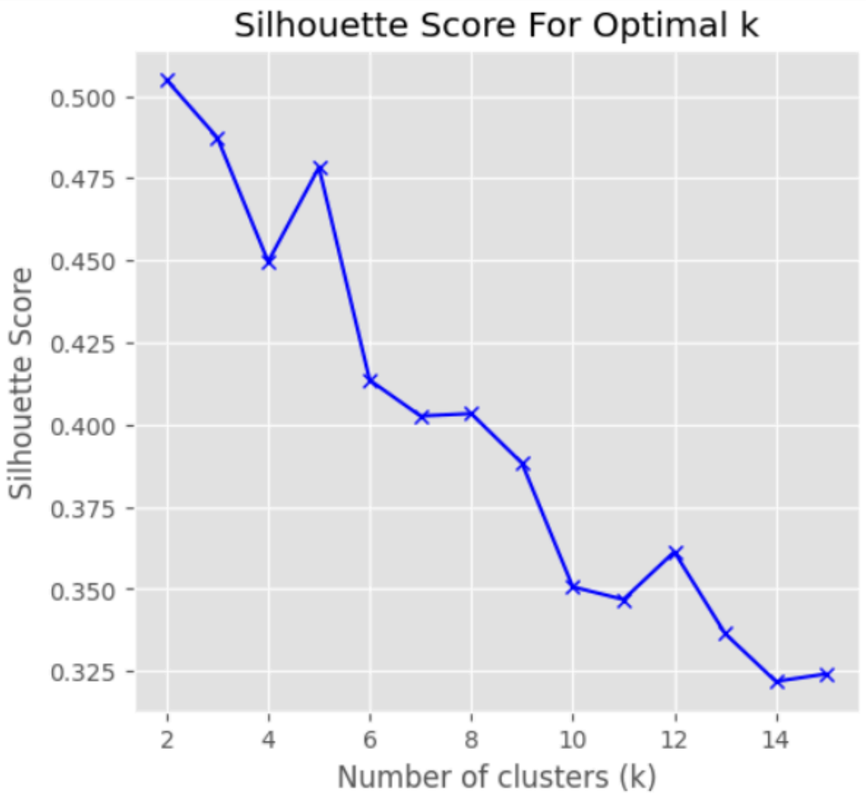


Data Modelling: K-means (Cluster = 5)



Finding optimal number of cluster





The "elbow point," where the rate of decrease sharply slows down, suggests an optimal number of clusters.

Model evaluation: K-Means Clustering

Optimal cluster k = 3

Clusters	Silhouette Score	Davies-Bouldin Index
2	0.504892	0.802572
3	0.487308	0.708335
4	0.449600	0.869746
5	0.478286	0.845857

- Davies-Bouldin index:
  - Similarity ratio of clusters.
  - The lowest values indicate better clustering.

Cluster Evaluation

Cluster = 0

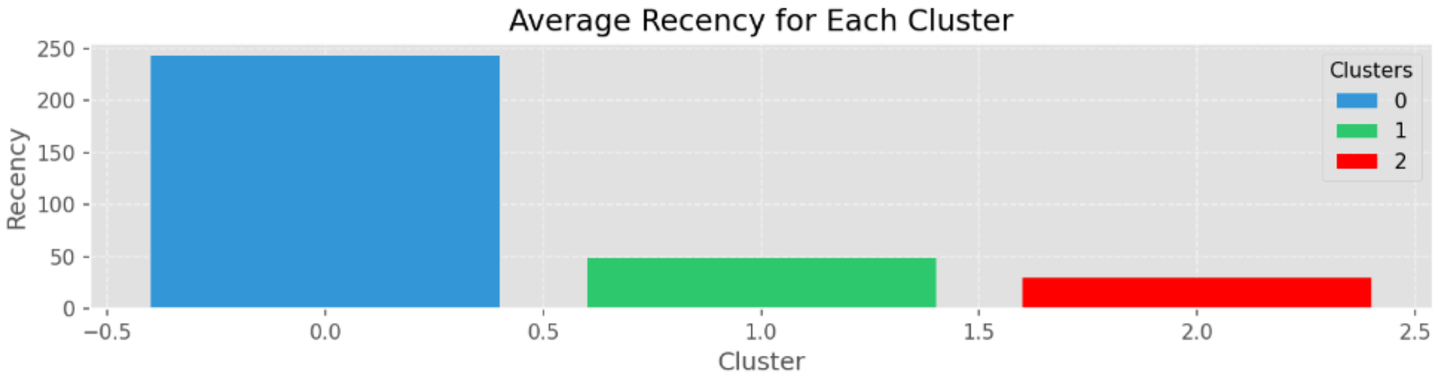
- High recency, low frequency, & low monetary

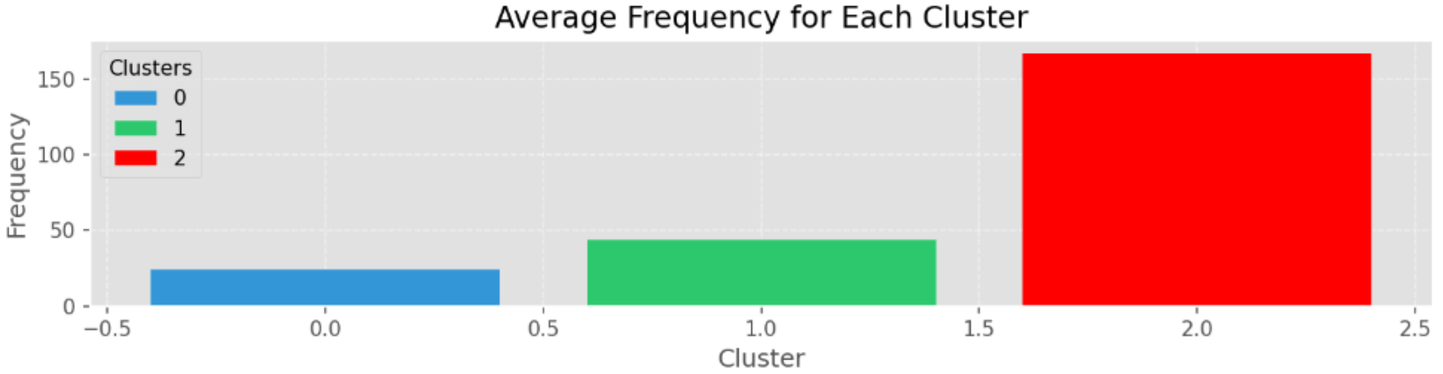
Cluster = 1

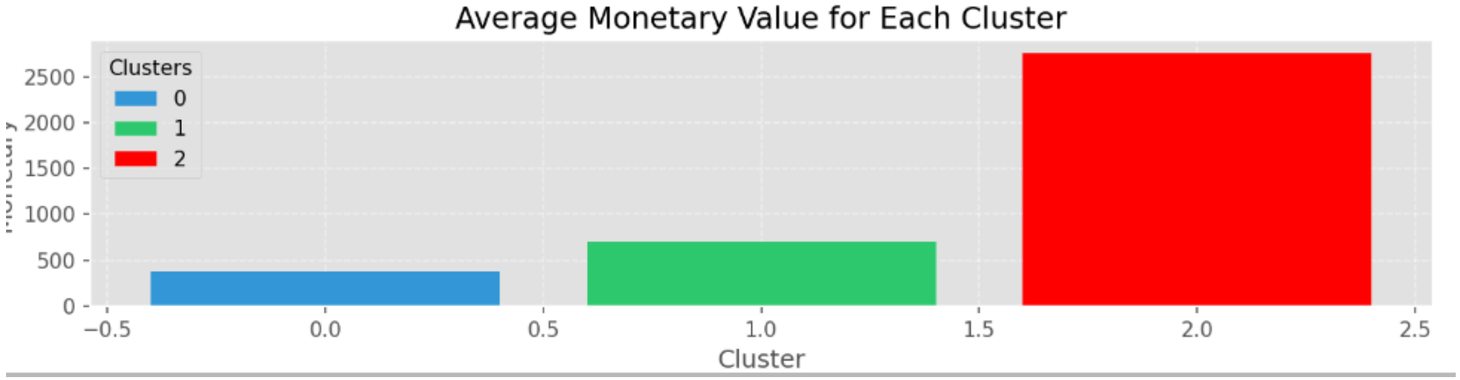
- Low recency, low frequency, & low monetary

Cluster = 2

- Low recency, highest frequency, & highest monetary



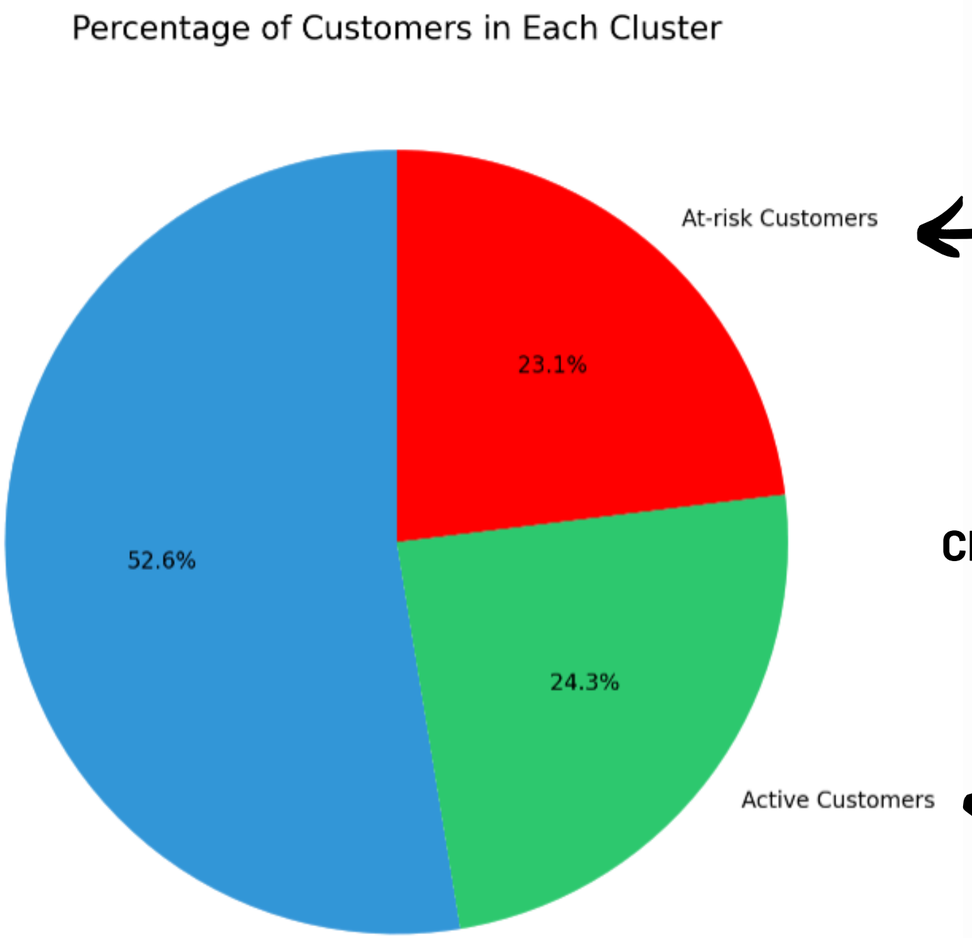




Insights

Cluster = 0 or ‘At risk customers’

- High recency, low frequency, & low monetary



Cluster = 2 or ‘Best Customer’

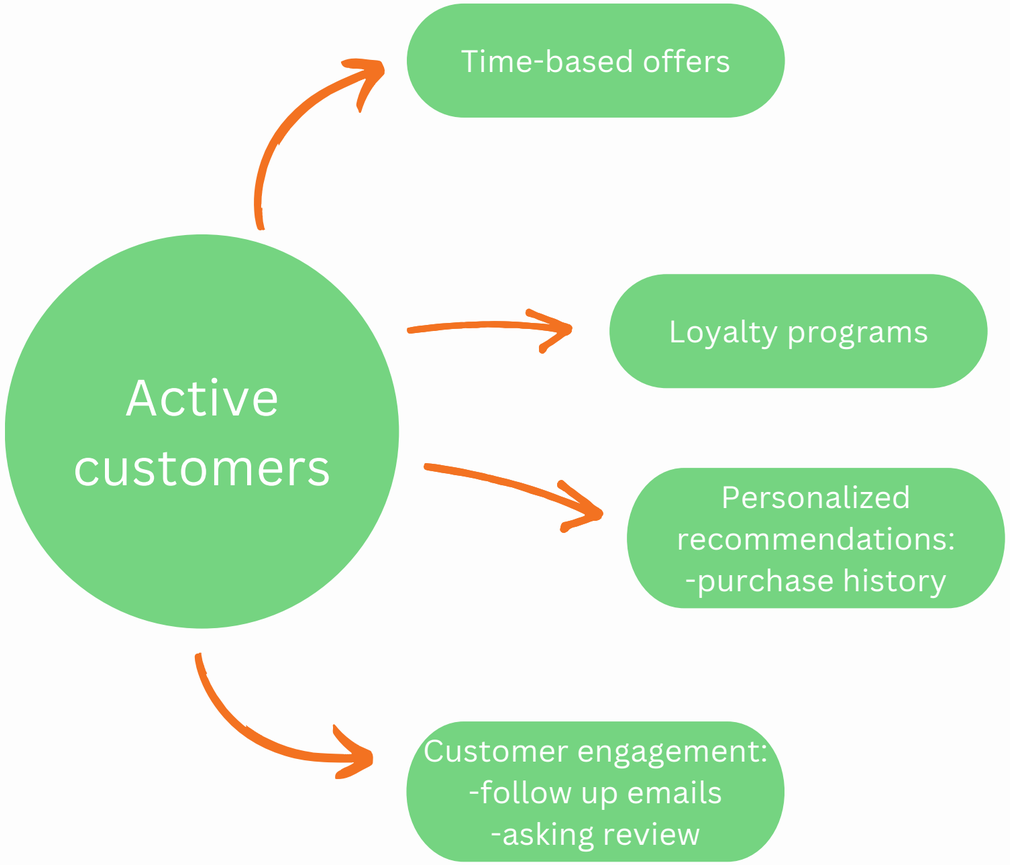
- Low recency, highest frequency, & highest monetary

Cluster = 1 or ‘Active Customers’

- Low recency, low frequency, & low monetary

Recommendation: Focus on the best customers

Keep & improve the Loyalty of Active Customers



Re-Engagement of At-Risk Customers


THANK YOU