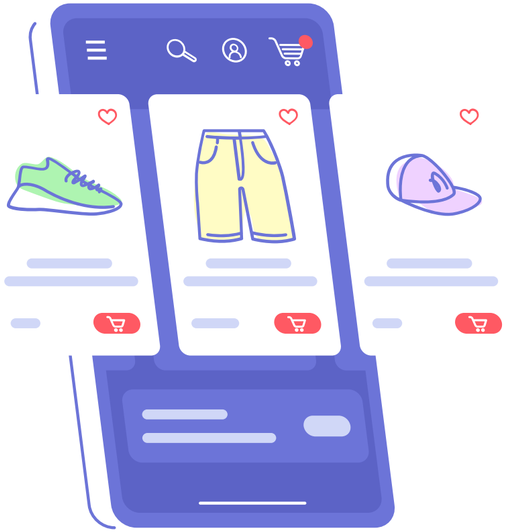
CUSTOMER SEGMENTATION OF ONLINE RETAIL USING RFM ANALYSIS AND K-MEANS CLUSTERING

BY HARISH MUHAMMAD

BUSINESS PROBLEM & BACKGROUND

Customer centric to win in high competitive environment



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

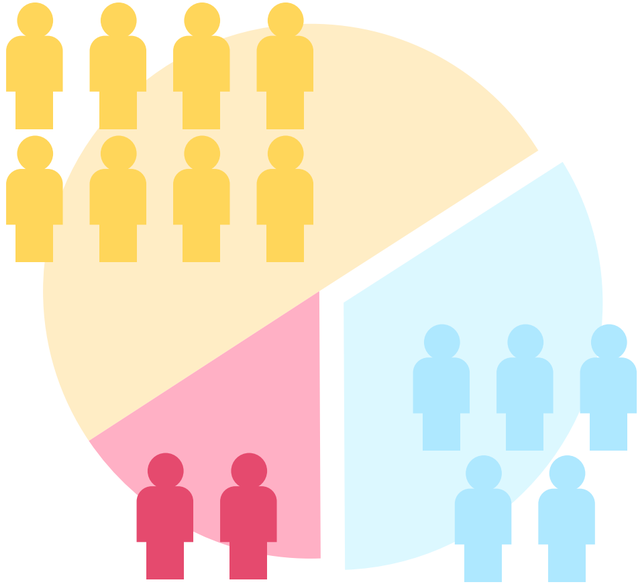
Customer-centric can be the way to win the completion

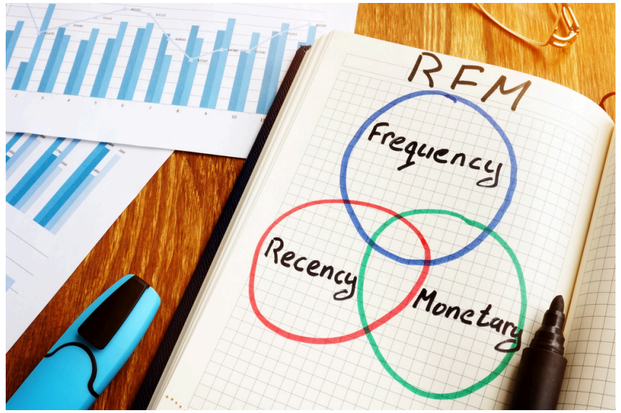
- Understanding customers' needs, perceptions, and expectations is crucial for stronger customer relationships & better business outcomes

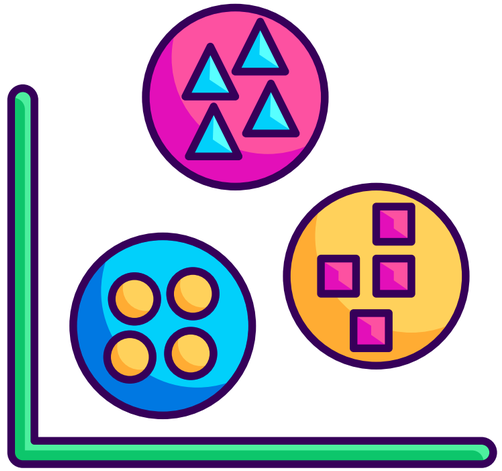
Challenge: Vast amounts of data:

- 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

BUSINESS AND TECHNICAL OBJECTIVES

Business objective

🎯 Business Objective

Goal:

To support targeted marketing strategies by grouping customers into actionable group segments

Why?

- Improve customer retention by identifying loyal and high-value customers.
- Design personalized promotions based on customer purchase behavior.
- Reduce marketing costs by focusing on the right segments.
- Identify at-risk or inactive customers for reactivation campaigns.

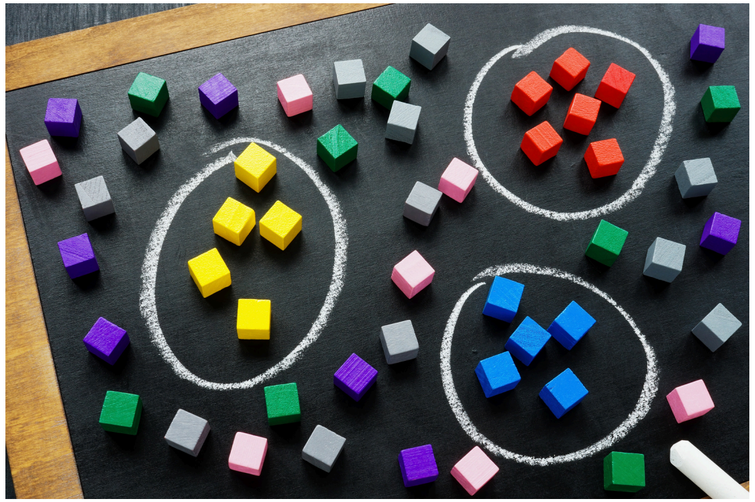
Business Benefit

- Better understanding of the customer base to create a targeted marketing campaign or product differentiation based on each customer → Stronger customer engagement and loyalty.

Technical data science objective

🎯 Technical Objective

Technical objective:
To build an unsupervised learning model using RFM analysis and K-Means clustering to segment customers into behavior-based groups.



Technical Keys

- Implement RFM (Recency, Frequency, Monetary) feature engineering
- Apply data preprocessing, transformation, and clustering
- Visualize and interpret clustering results

EXAMINING DATA (MENELAAH DATA)

Dataset source: UK-based online retail 2010-2011

Scientific paper

Also available on

UC Irvine Machine Learning Repository

Online Retail
Donated on 11/5/2015

This is a transnational data set which contains all the transactions occurring between 01/12/2010 and 09/12/2011 for a UK-based and registered non-store online retail.

Dataset Characteristics	Subject Area	Associated Tasks
Multivariate, Sequential, Time-Series	Business	Classification, Clustering
Feature Type	**# Instances**	**# Features**
Integer, Real	541909	6

Data Summary

```
<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
```

- Dimension: 8 columns and 541909 rows
- Date range: Dec 2010 - Dec 2011

Statistical descriptive

	count	mean	min	25%	50%	75%	max	std
Quantity	541909.0	9.55225	-80995.0	1.0	3.0	10.0	80995.0	218.081158
InvoiceDate	541909	2011-07-04 13:34:57.156386048	2010-12-01 08:26:00	2011-03-28 11:34:00	2011-07-19 17:17:00	2011-10-19 11:27:00	2011-12-09 12:50:00	NaN
UnitPrice	541909.0	4.611114	-11062.06	1.25	2.08	4.13	38970.0	96.759853
CustomerID	406829.0	15287.69057	12346.0	13953.0	15152.0	16791.0	18287.0	1713.600303

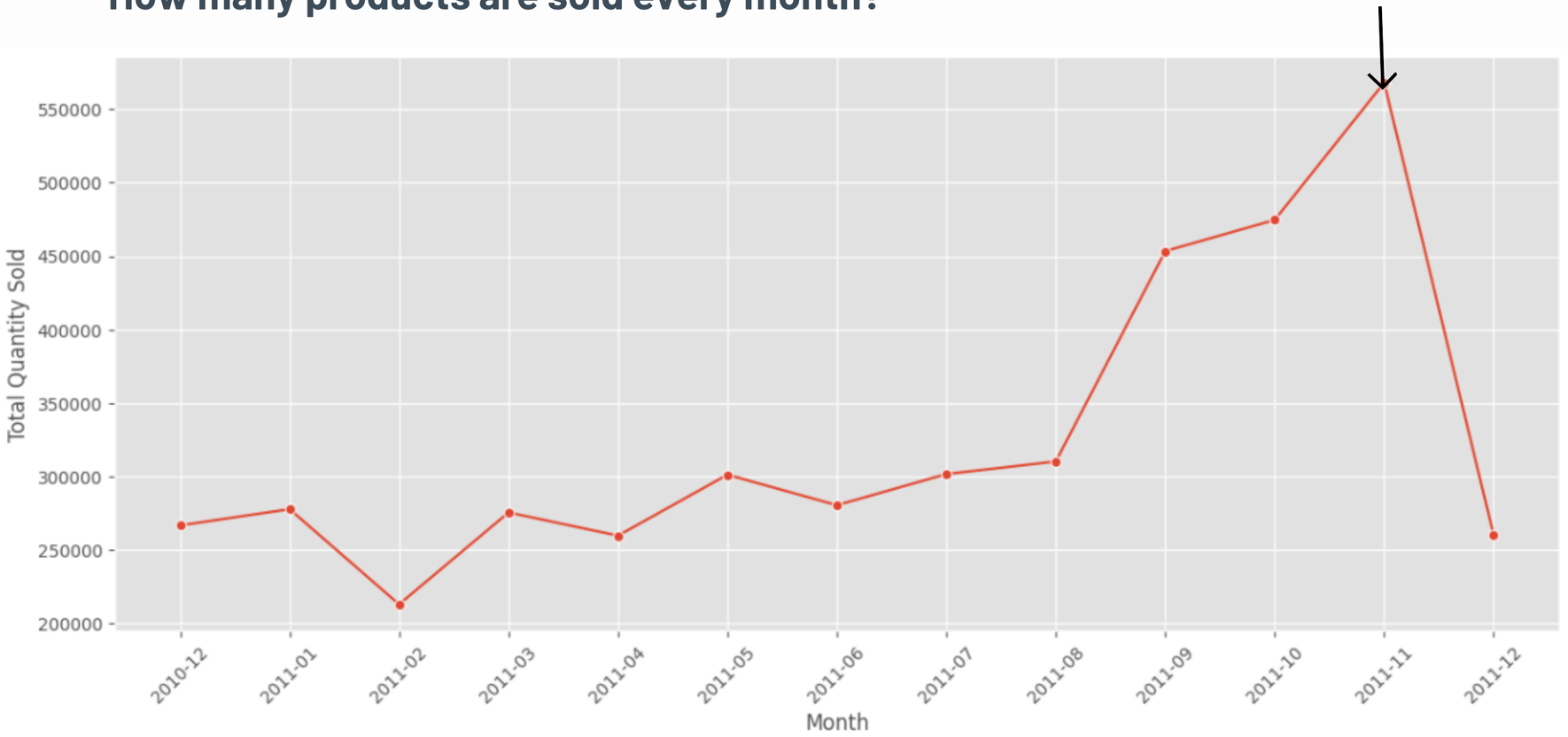
	count	unique	top	freq
InvoiceNo	541909	25900	573585	1114
StockCode	541909	4070	85123A	2313
Description	540455	4223	WHITE HANGING HEART T-LIGHT HOLDER	2369
Country	541909	38	United Kingdom	495478

- There are 38 countries mentioned. But the majority of consumers are from the United Kingdom (more than 90%).
- The min and max value for 'Quantity' is -80995 and 80995, this could attributed to cancelled or returned orders.
- The 'UnitPrice' has negative values that may represent cancelled orders or bad debt by retailers.
- These anomalies will be handled in next steps.

Total product sold per month overtime

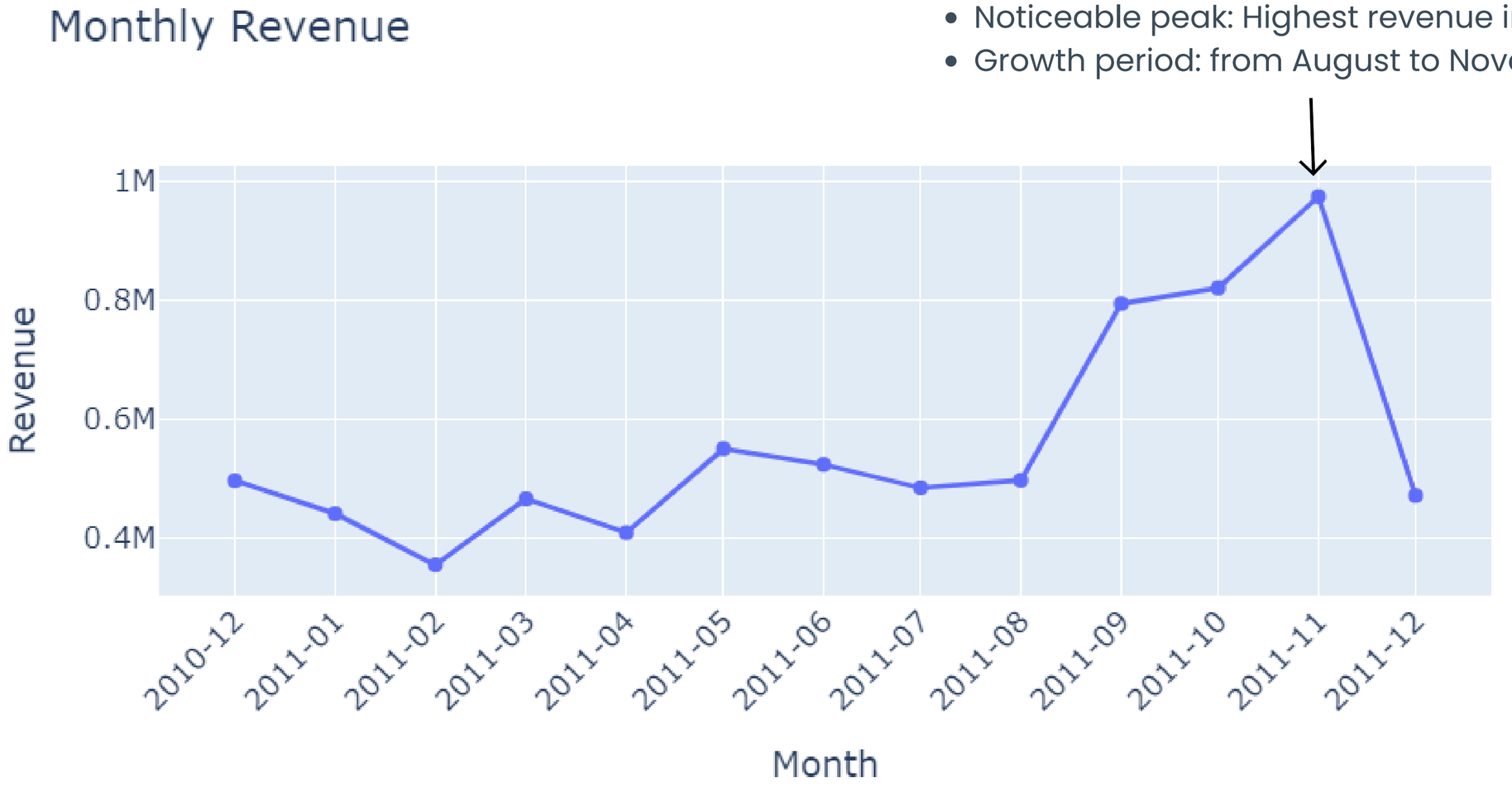
How many products are sold every month?

The highest number of product sold on November



- Product sold in November → 13,41% of transactions per year.
- Therefore, the business team can increase sales this month → by increasing promotions to customers.

Total revenue overtime



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

- Marketing strategy can be implemented by observing seasonal or annual pattern
- Increase marketing efforts a few months before November.

VALIDATE DATA (MEMVALIDASI DATA)

Validate data

🎯 Validation Actions:

- Validate that the InvoiceDate format is consistent
- Remove transactions with negative Quantity or UnitPrice
- Remove entries with missing CustomerID

Correcting data type format of 'InvoiceDate'

```
# converting data type from the columns of 'InvoiceDate' into datetime
df['InvoiceDate'] = pd.to_datetime(df['InvoiceDate'], format='%d.%m.%Y %H:%M:%S')
```

```
# Checking the data type of the converted columns
print(f"Data type of 'InvoiceDate' column: {df['InvoiceDate'].dtypes}")
```

```
Data type of 'InvoiceDate' column: datetime64[ns]
```

Validate data

Removing UnitPrice that less than zero

```
# Removing rows with anomaly in 'UnitPrice'
df = df[df['UnitPrice'] > 0]
```

Removing Quantity that less than zero

```
[ ] # Displaying the rows with anomaly in 'Quantity'
    df_quantity_anomaly = df[df.Quantity < 0]
    print(f'The number of rows with Quantity less than zero are {len(df_quantity_anomaly)} or {round(len(df_quantity_anomaly)
    df_quantity_anomaly.head(3)
```

```
The number of rows with Quantity less than zero are 9192 or 1.86% of total dataset
```

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
141	C536379	D	Discount	-1	2010-12-01 09:41:00	27.50	14527.0	United Kingdom
154	C536383	35004C	SET OF 3 COLOURED FLYING DUCKS	-1	2010-12-01 09:49:00	4.65	15311.0	United Kingdom
235	C536391	22556	PLASTERS IN TIN CIRCUS PARADE	-12	2010-12-01 10:24:00	1.65	17548.0	United Kingdom

```
# Removing rows with anomaly in 'Quantity'
df = df[df['Quantity'] > 0]
```

DETERMINE OBJECT DATA (MENENTUKAN OBJEK DATA)

Determine object data

Primary Entity (Object):

- CustomerID - the unit of analysis for segmentation.
- Filter: only UK transactions
- Each row in the dataset is a transaction, but segmentation is done per customer.

Derived Metrics:

- Aggregate transactions by customer to calculate:
  - Recency: Days since last purchase
  - Frequency: Number of transactions
  - Monetary: Total amount spent

RFM Variables Constructed From:

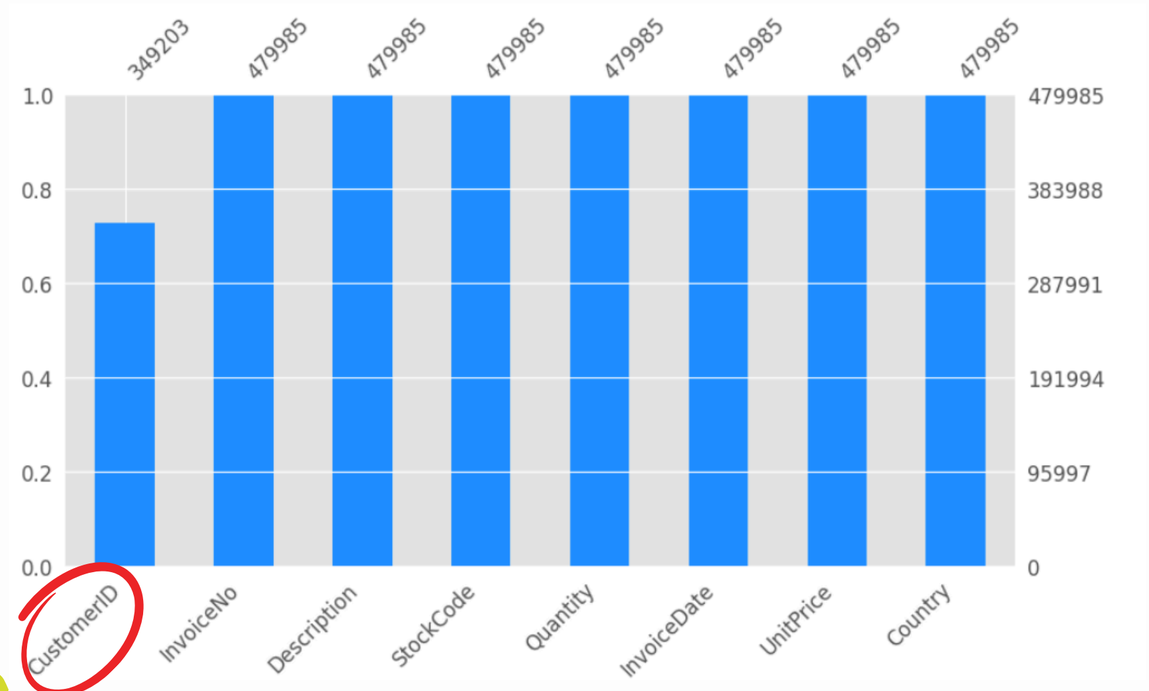
- InvoiceDate → Recency
- Count of invoices → Frequency
- Quantity × UnitPrice → Monetary

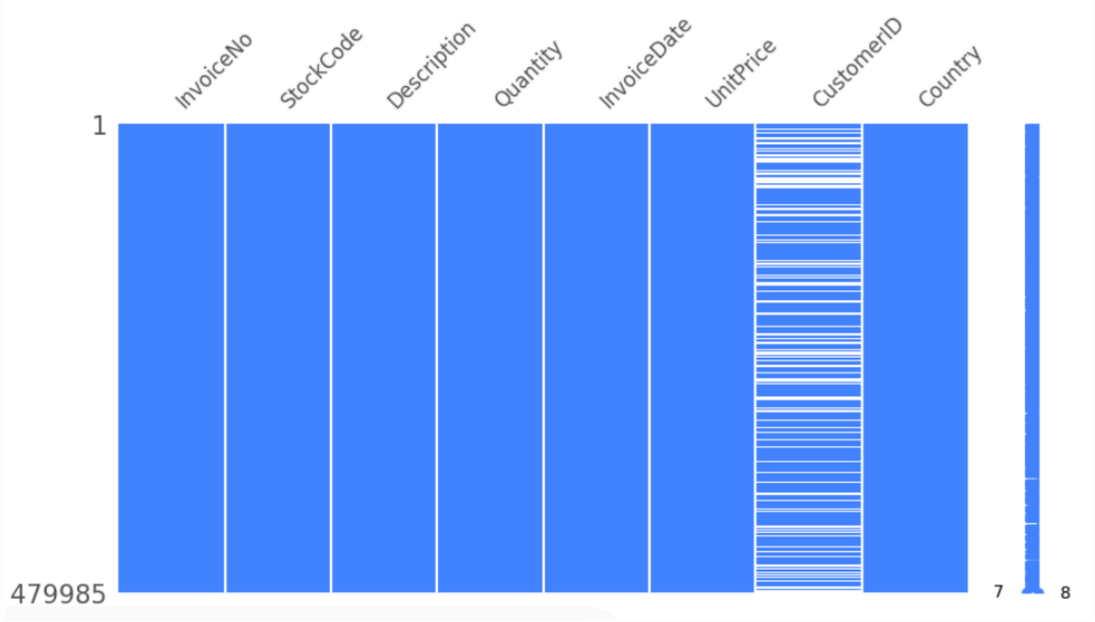
DATA CLEANING (MEMBERSIHKAN DATA)

Handling Missing Values

- Remove transactions with missing CustomerID

	Percentage of Missing values
InvoiceNo	0.000000
StockCode	0.000000
Description	0.000000
Quantity	0.000000
InvoiceDate	0.000000
UnitPrice	0.000000
CustomerID	27.247101
Country	0.000000





Handling Duplicates

```
# Checking the number of duplicated data
print(f"Total number of duplicated data: {df.duplicated().sum()}")
print(f"Percentage of duplicated data: {df.duplicated().sum()/len(df)*100:.2f}%")
```

```
Total number of duplicated data: 5138
Percentage of duplicated data: 1.06%
```

```
# Handling duplicated data

# Dropping duplicates and resetting the index
df = df.drop_duplicates(ignore_index=True)
```

DATA CONSTRUCTION (MENGKONSTRUKSI DATA)

RFM Analysis

RFM analysis allows us to segment customers by the frequency and value of purchases and identify those customers who spend the most money

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

Construct the data

Define Reference Date:

- Chosen as 1 December 2011 (the day after the last transaction).
- Used to calculate Recency.

Calculate RFM Metrics Per Customer

- Recency = Days since last purchase (Reference Date – Last InvoiceDate)
- Frequency = Number of unique invoices per customer
- Monetary = Total spend per customer
- (Quantity × UnitPrice, aggregated)

Construct Final RFM Table:

- Each row = one unique customer
- Columns = Recency, Frequency, Monetary

Transformation

- Applied log or Box-Cox transformation to reduce skewness

Scaling (Standardization)

- Applied StandardScaler to normalize features for clustering

RFM Score

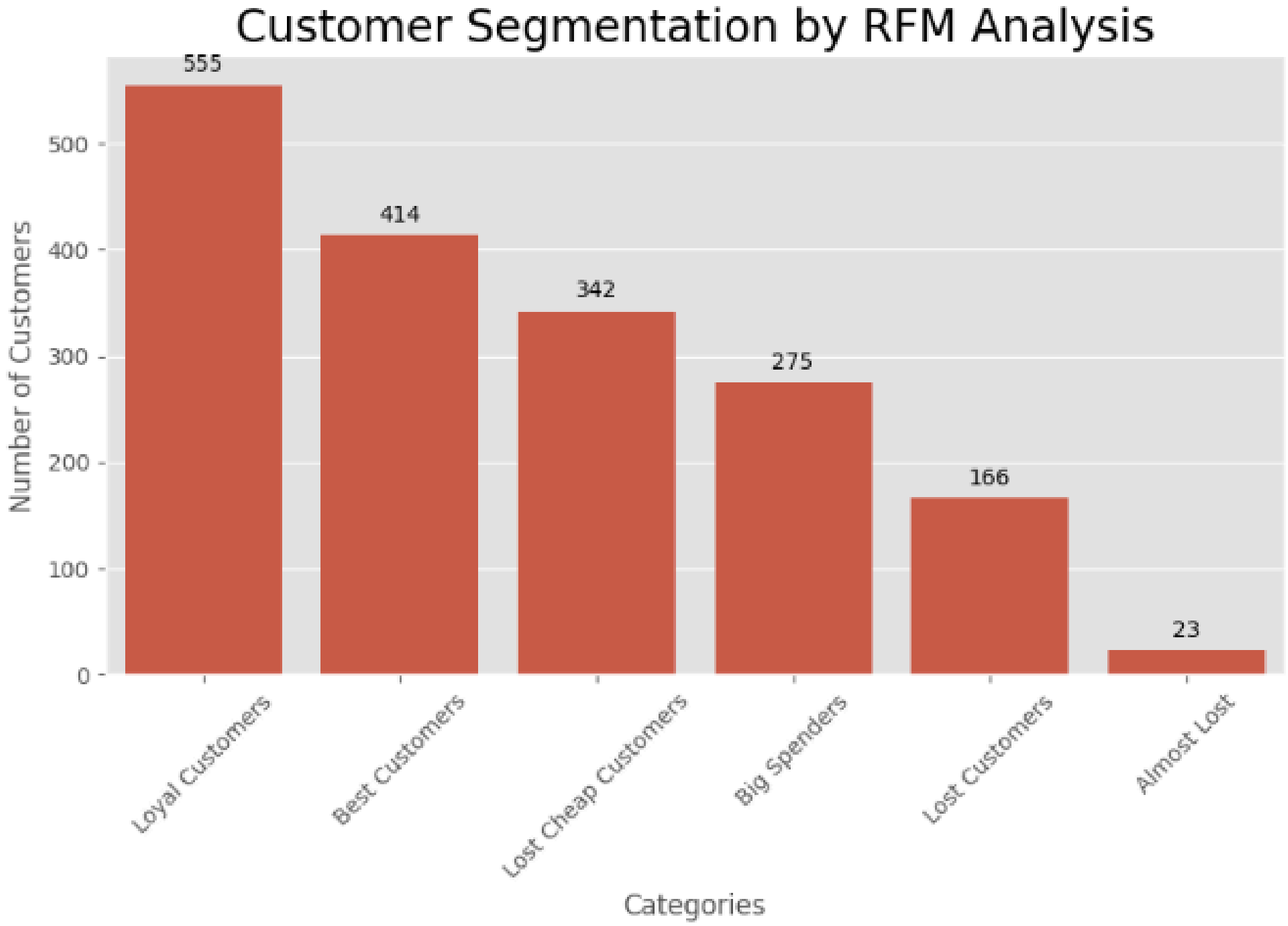
- RFM metrics are segmented based on quantiles
- Each metric (recency, frequency, monetary) is assigned a score from 1 to 4, (1 for the highest value, 4 for the lowest value)
- RFM score (overall) is calculated by combining individual RFM score.

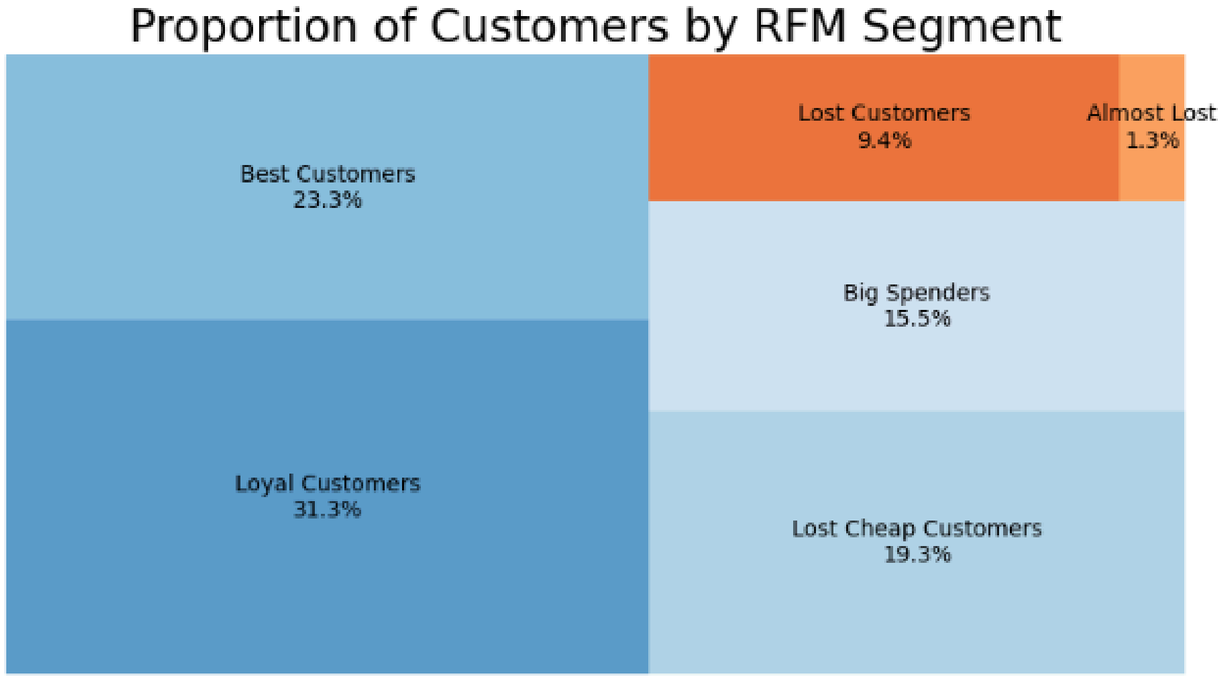
Segment	RFM Score
Best Customers	111
Loyal Customers	F = 1
Big Spenders	M = 1
Almost lost	134
Lost Customers	344
Lost Cheap Customers	444

RFM Table

	CustomerID	customer_age	recency	frequency	monetary	avg_spend_per_product	R_quartile	F_quartile	M_quartile	RFM_segment	RFM_score	RFM_label
0	12346.0	326	326	1	77183.60	1.040000	4	4	1	441	9	Big Spenders
1	12747.0	369	2	103	4196.01	4.367864	1	1	1	111	3	Best Customers
2	12748.0	374	1	4412	33053.19	2.671874	1	1	1	111	3	Best Customers
3	12749.0	213	4	199	4090.88	4.999950	1	1	1	111	3	Best Customers
4	12820.0	327	3	59	942.34	1.904746	1	2	2	122	5	Others

Customer segmentation by RFM





Feature engineering for K-means Clustering

Optimum conditions for K-means clustering:

Normal/symmetric distribution

Minimized outliers

Low Skewness

Standardised data

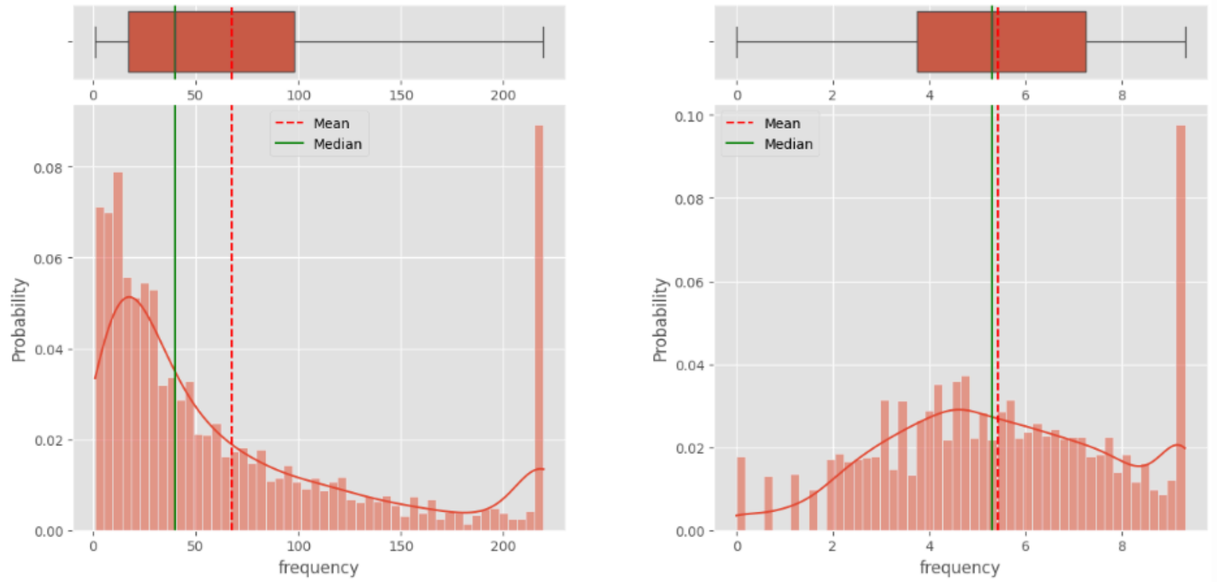
Data distribution & skewness check:

Data are not normally distributed and highly skewed

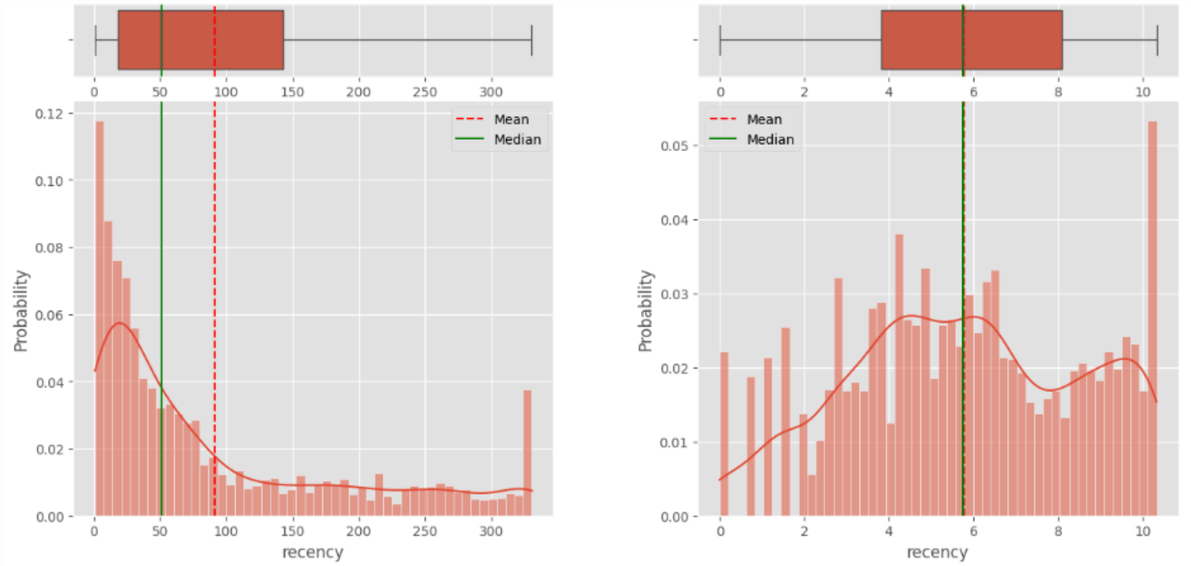
Feature	D'Agostino-Pearson Statistic	P-value	Distribution	Skewness	Skewness Type
recency	583.200928	2.288377e-127	Not Normally Distributed	1.163469	Right Skew
frequency	589.373112	1.045334e-128	Not Normally Distributed	1.170132	Right Skew
monetary	589.807769	8.411440e-129	Not Normally Distributed	1.173867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

Before & After:
Data Transformation

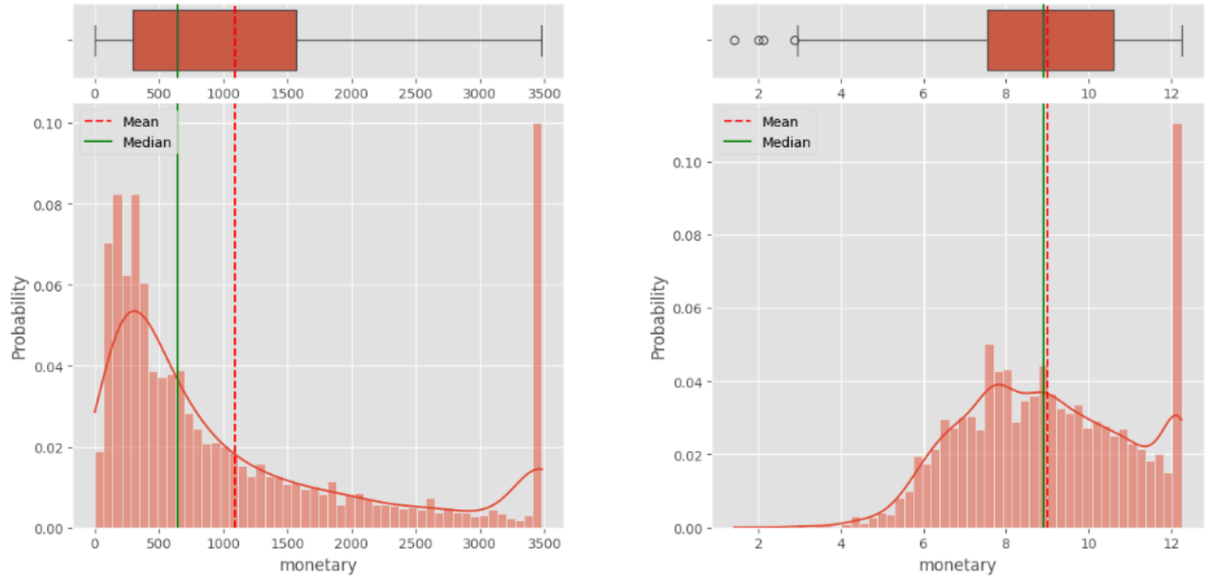
Frequency → Transformed Frequency



Recency → Transformed Recency



Monetary → Transformed Monetary



Data pre-processing: Outlier handling, Transformation, Standardisation

Handling outliers

Winsorization

Scaling

Robust Scaler

Transformation

Box Cox - method

BUILDING A MODEL SCENARIO (MEMBANGUN SKENARIO MODEL)

Model scenario

Modeling strategy

- Model type: unsupervised clustering
- Algorithm: K-means clustering
- Features: recency, frequency, and monetary

Determine optimal K

- Elbow Method (minimize within-cluster sum of squares)
- Silhouette Score (measure separation quality)
- Davies-Bouldin Index (Measures intra-cluster similarity and inter-cluster differences)

Run simulations with various K

- Test K values from 2 to 15
- Use a combination of Elbow, Silhouette, and DBI to select the optimal K
- Choose K with balanced cohesion, separation, and business interpretability

MODEL BUILDING (MEMBANGUN MODEL)

Model Building

Fitting the model

- Applied to scaled RFM data
- Cluster centroids are automatically optimized

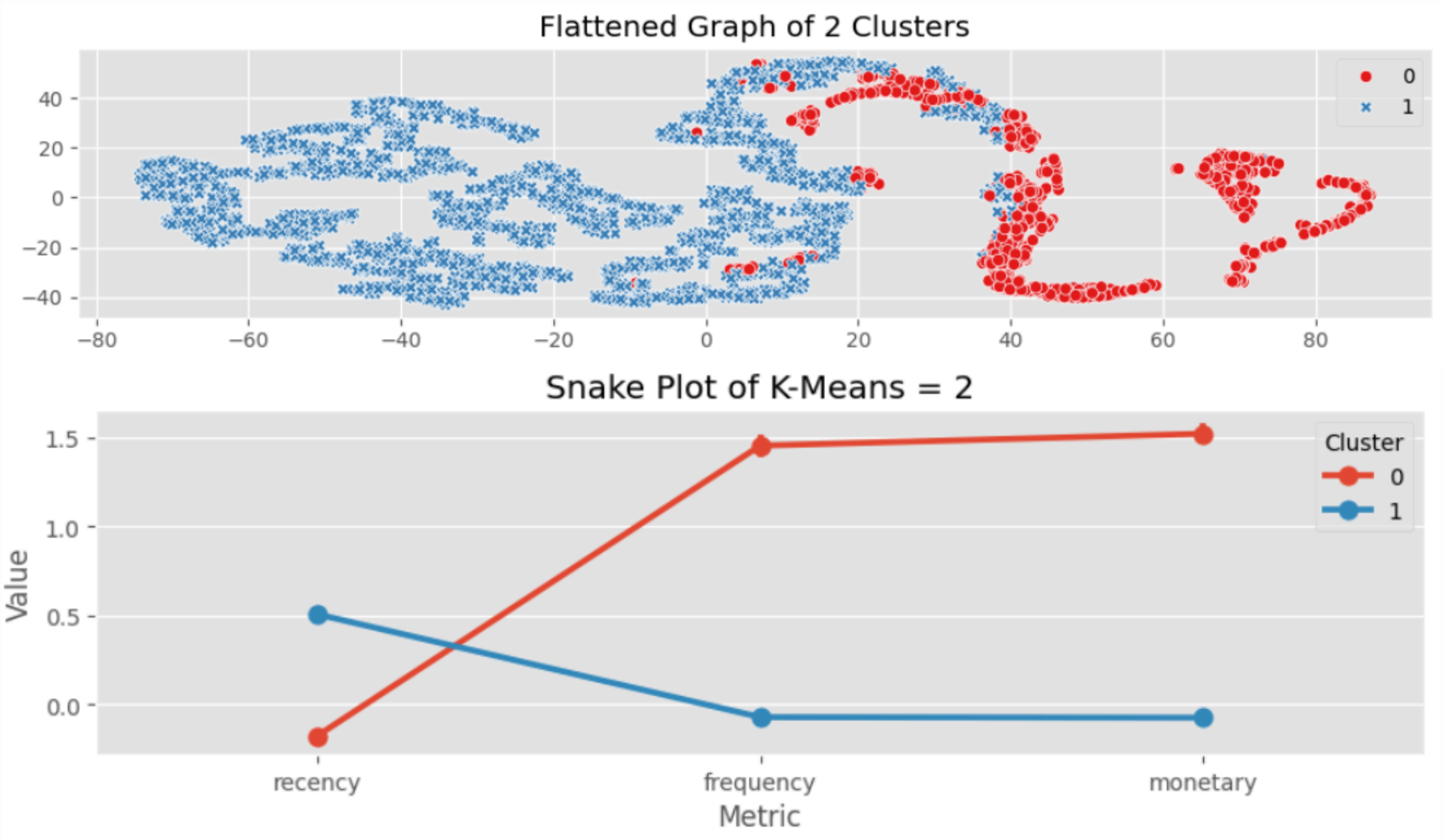
Assign Cluster Labels to customers

- Each customer is assigned a Cluster ID
- Added as a new column in the dataset

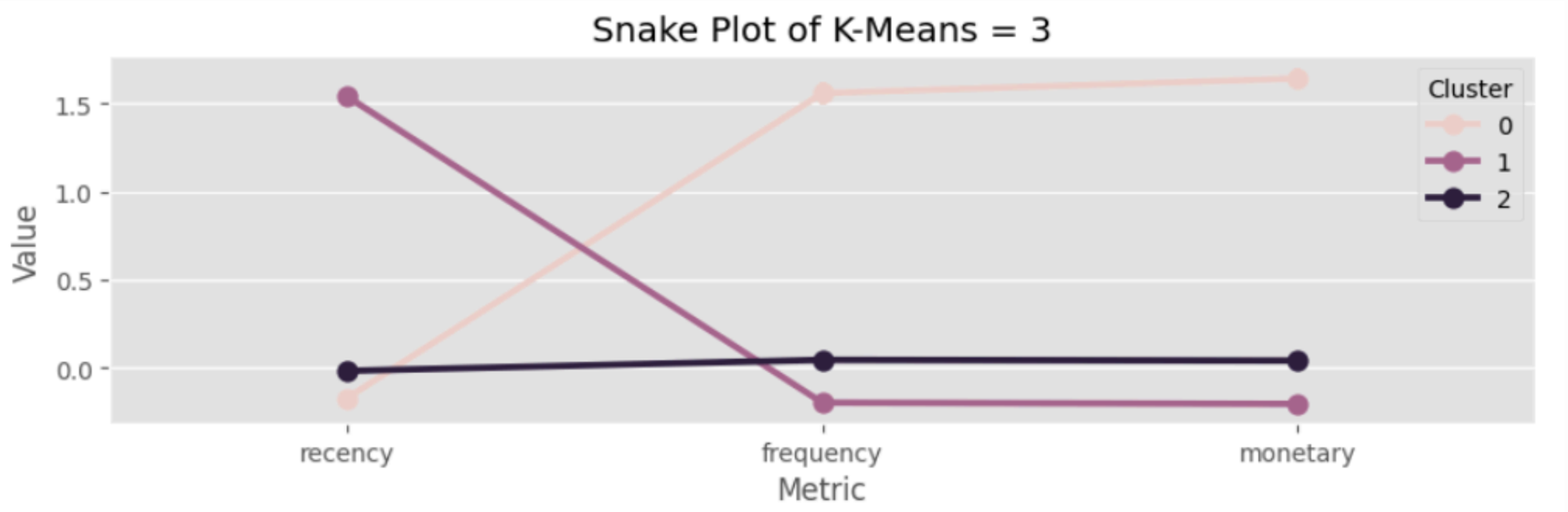
Model output and business integration

- Cluster output will be a guide for the marketing campaign
  - High-value: loyalty offers
  - Low-frequency: re-engagement promotions
  - At-risk: win-back initiatives

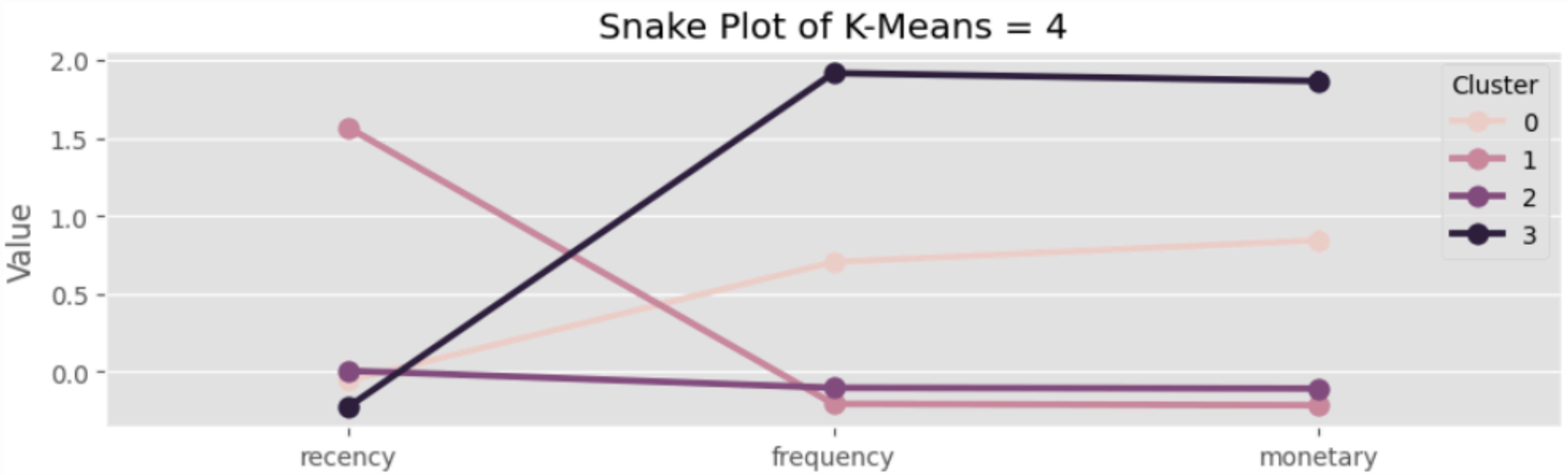
Data Modelling: K-means (Cluster = 2)



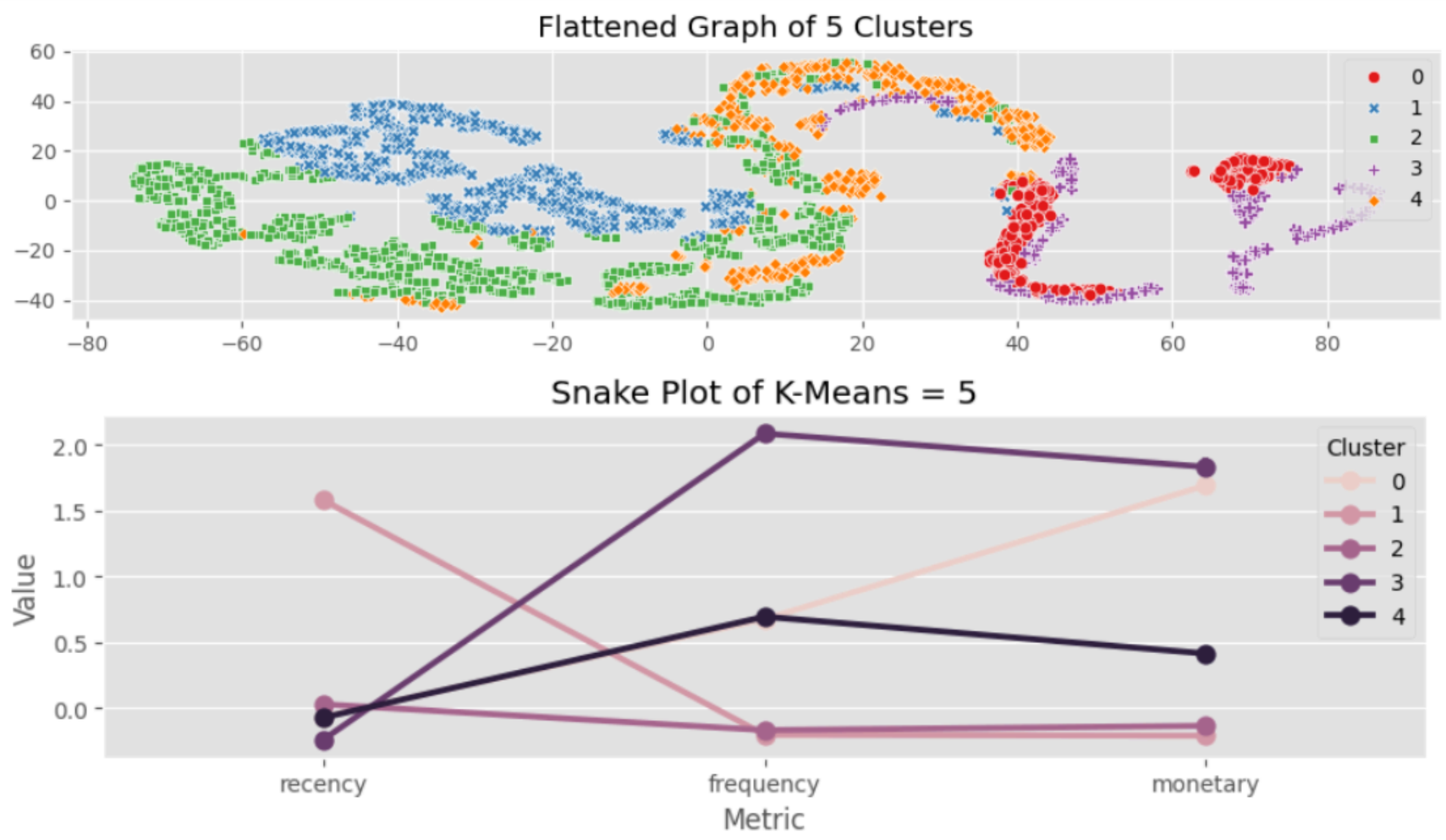
Data Modelling: K-means (Cluster = 3)



Data Modelling: K-means (Cluster = 4)



Data Modelling: K-means (Cluster = 5)



MODEL EVALUATION (MENGEVALUASI MODEL)

Model Evaluation

Evaluation metrics used

- **Silhouette Score**
  - Measures how well samples are clustered with similar ones
  - Range: -1 to 1 (higher = better)
- **Davies-Bouldin Index (DBI)**
  - Evaluates average similarity between each cluster and its most similar one
  - Lower DBI = better separation

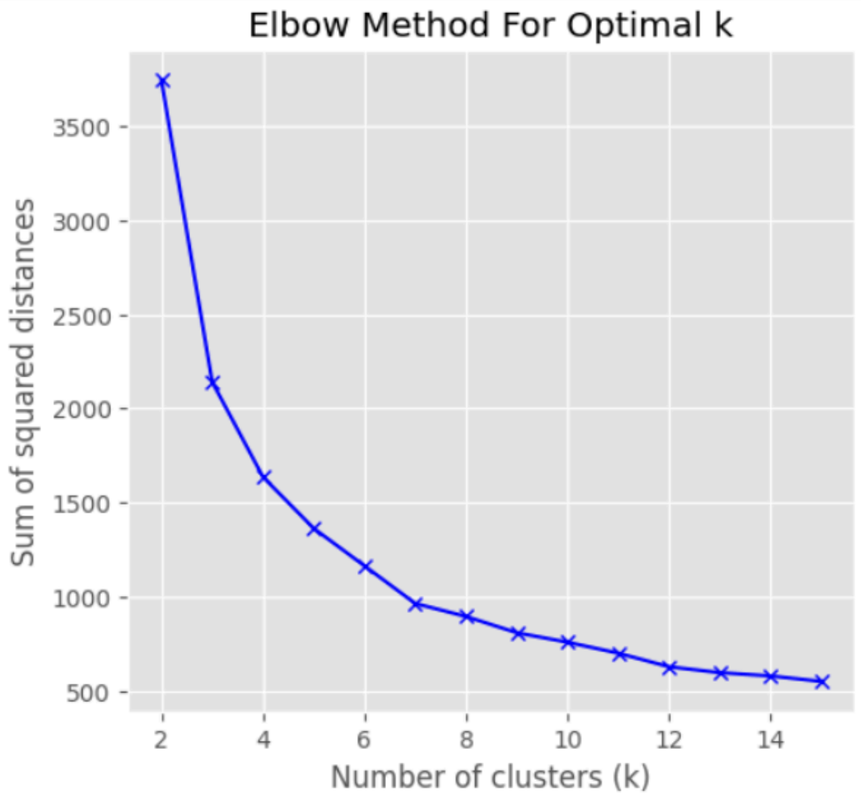
- Applied to scaled RFM data
- Cluster centroids are automatically optimized

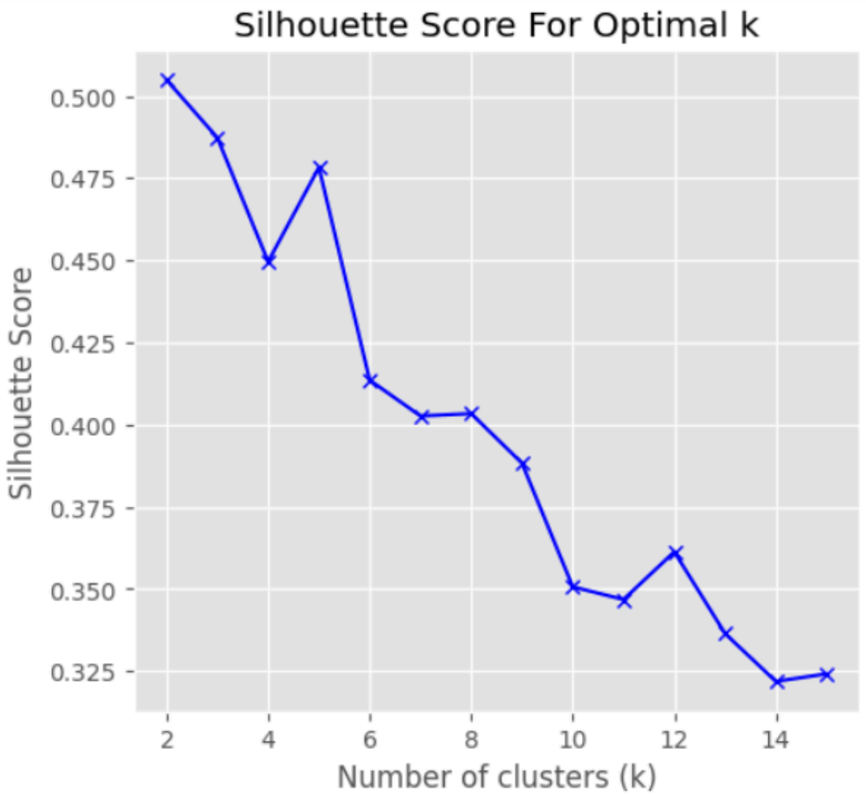
Assign Cluster Labels to customers

- Each customer is assigned a Cluster ID
- Added as a new column in the dataset

Model output and business integration

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 separation → better clustering

The model with 3 clusters is optimal for balancing:

- Statistical clustering performance
- Interpretability
- Actionability for marketing strategy

Cluster Evaluation

Cluster = 0

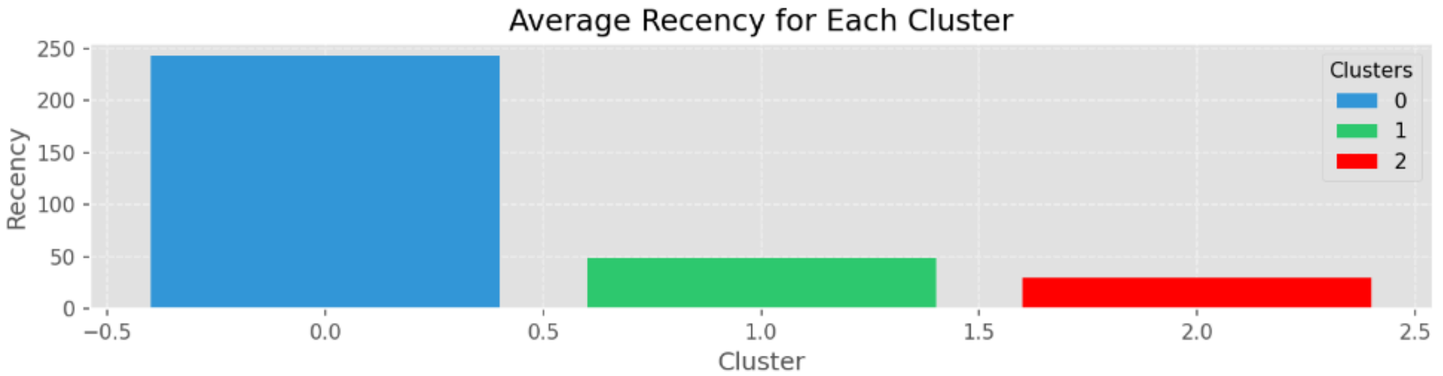
- High recency, low frequency, & low monetary

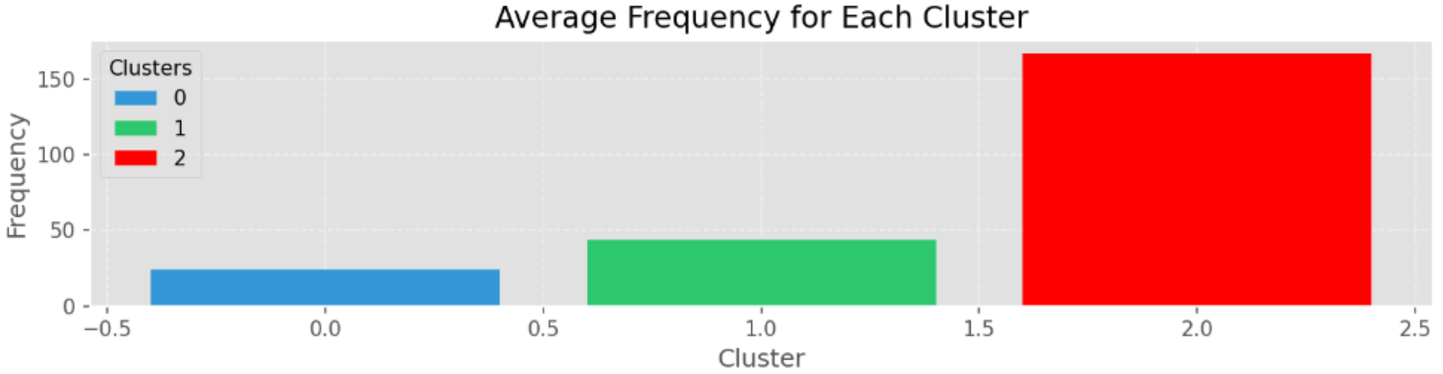
Cluster = 1

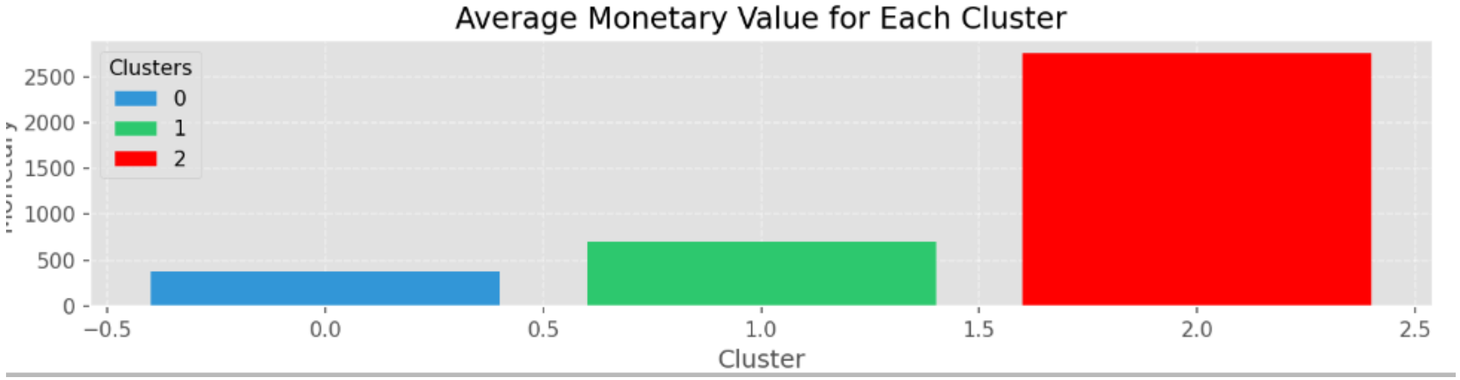
- Low recency, low frequency, & low monetary

Cluster = 2

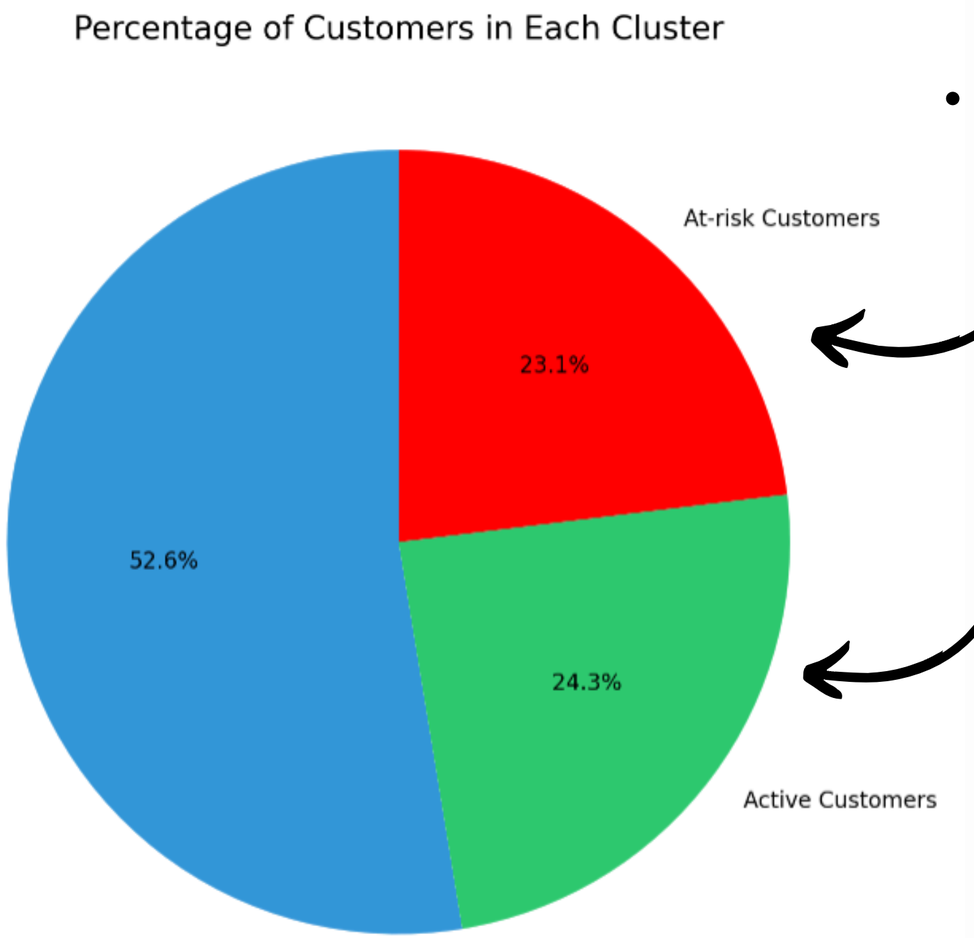
- Low recency, highest frequency, & highest monetary







Cluster profiling & business-relevant check



Cluster = 0 or 'At-risk customers'

- High recency, low frequency, & low monetary
- Represent inactive customers → need re-engegement or 'at risk' of being lost

Cluster = 1 or 'Active Customers'

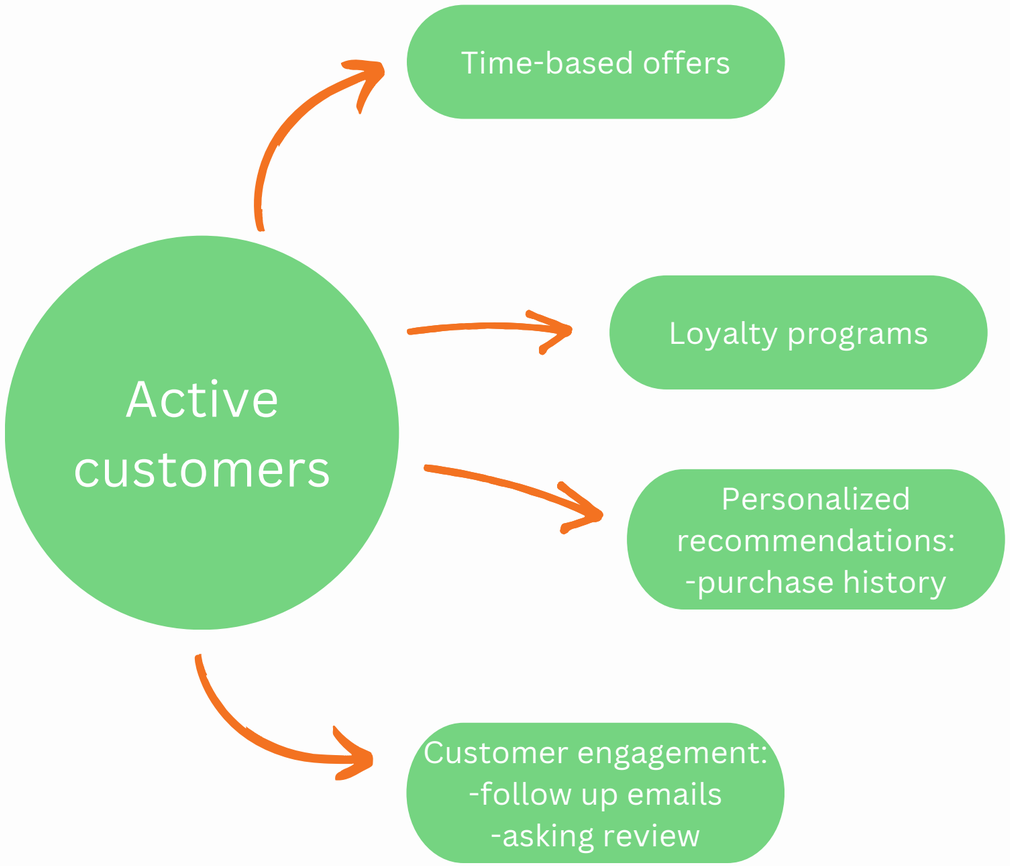
- Low recency, low frequency, & low monetary
- Somehow engaged with offerings, making regular purchases, but not at the highest frequency or monetary value

Cluster = 2 or 'Best Customer'

- Low recency, highest frequency, & highest monetary
- Highly engaged and contributes significantly to the revenue → focus for retention & loyalty programs

Actionable recommendations

Keep & improve the Loyalty of Active Customers



Re-Engagement of At-Risk Customers

Actionable recommendations

Retention & Loyalty programs for best customers

REVIEWING THE MODELING PROCESS (MELAKUKAN REVIEW PEMODELAN)

Review the modeling process

Alignment with Business Objectives

- **Successfully segmented customers** into 3 actionable groups
  - High-value, best customers
  - Active customers
  - At-risk customers
- Enables targeted strategies for retention, reactivation, and upselling

Alignment with Technical Objectives

- Used RFM (Recency, Frequency, Monetary) framework as planned
- Applied K-Means clustering as a scalable, unsupervised model
- Validated model using:
  - Silhouette Score (0.487)
  - Davies-Bouldin Index (0.708)
- Model results are documented, reproducible, and interpretable

Opportunities for Improvement

- Alternative algorithms (e.g., DBSCAN, Hierarchical) may uncover non-spherical clusters
- Perform domain expert review earlier to refine cluster labeling
- Future work: evaluate cluster stability over time


THANK YOU