Data Mining Project

Project Title: Online Retail Customer Segmentation

Business Context and Problem

Business Context

Our project focuses on a UK-based online retail store aiming to optimize customer segmentation. By understanding customer behavior, the company can deliver personalized marketing strategies, improve customer retention, and maximize revenue.

Business Problem Statement

How can we group customers effectively based on their purchasing behavior to enhance targeted promotions, increase customer satisfaction, and boost overall sales performance?

Tackling the Business Problem

Challenges with Conventional Managerial Insights:

Traditional managerial approaches categorize customers broadly, using generic promotions. This often fails to uncover nuanced patterns such as loyalty, spending, and inactivity. (e.g., Broad campaigns lead to overspending on low-value customers)

Our Approach:

1. Focus on **RFM-based segmentation** to uncover customer behavior.
2. Apply **unsupervised learning techniques** to identify unique customer groups.
3. Combine **data-driven insights** with tailored business strategies to maximize engagement and revenue.

Available Resources:

- **Dataset** Source from Kaggle – contains 500,000+ transactions from a UK-based online retailer (Dec 2010-Dec 2011).

lers.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Count
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/10 8:26	2.55	17850.0	Unit Kingdc
1	536365	71053	WHITE METAL LANTERN	6	12/1/10 8:26	3.39	17850.0	Unit Kingdc
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/10 8:26	2.75	17850.0	Unit Kingdc
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/10 8:26	3.39	17850.0	Unit Kingdc
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/10 8:26	3.39	17850.0	Unit Kingdc

Data Overview and Preprocessing

Data Description

- Includes transactional data containing 541,909 transactions from an online retailer with 8 features as in table.
- Contains Unique, all-occasion gifts with a significant wholesale customer base.

NAME	DESCRIPTION	TYPE
InvoiceNo	Unique identifier for each invoice	Numeric
StockCode	Unique identifier for each product	Numeric
Description	Description of the product	Textual
Quantity	Quantity of the product purchased	Numeric
InvoiceDate	Date and time when the invoice was generated	Datetime
UnitPrice	Price of a single unit of the product	Numeric
CustomerID	Unique identifier for the customer	Numeric
Country	Country where the customer is located	Categorical

Preprocessing

- **Data Cleaning**: Removed missing values for CustomerID
- **Feature Engineering**:
- Added **TotalPrice** (Quantity * UnitPrice) for spending analysis.
- Extracted **Month**, **Quarter**, **Day of Week** and **Hour** from InvoiceDate
- **Aggregated data to create RFM Metrics**:

Exploratory Data Analysis (EDA)

Frequency: Number of purchase invoices.

Monetary: Total spending.

Trends: Peak sales observed in **October-December (Q4)**.

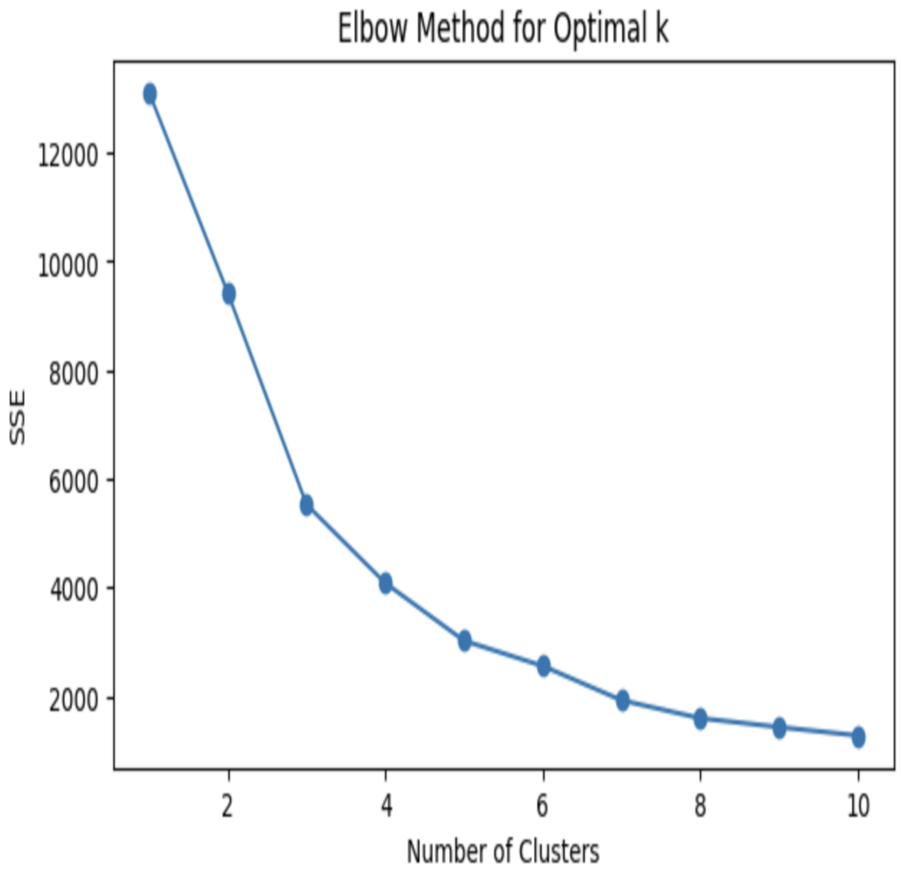
Patterns: High customer activity on **Thursdays**. **Afternoon hours** show the most engagement.

Data Mining Models

Data Aggregation → Exploratory Data Analysis (EDA) → Clustering for Segmentation → Outlier removal & Model refinement → Insights and Recommendations

K-Means Clustering

- Used to segment customers into distinct groups based on their Recency, Frequency, and Monetary (RFM) metrics.
- Customers were grouped by minimizing differences within clusters and maximizing differences between clusters. Centroids represent each cluster's average traits for easy interpretation. The **Elbow Method** identified **3 clusters** [domain expertise decide to take 3] as optimal for business relevance.



Hierarchical Clustering

- Used to validate K-Means results and visualize customer relationships through a **dendrogram**.
- Customers were grouped by splitting clusters based on similarities using **Ward linkage** to minimize variance. A dendrogram guided the selection of 3 clusters for consistency with K-Means.

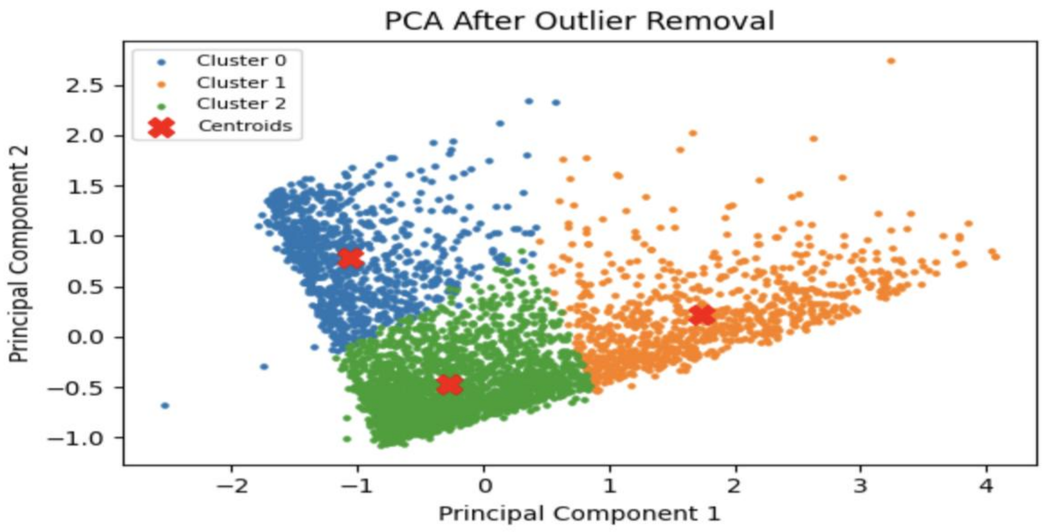
Detailed Model Description

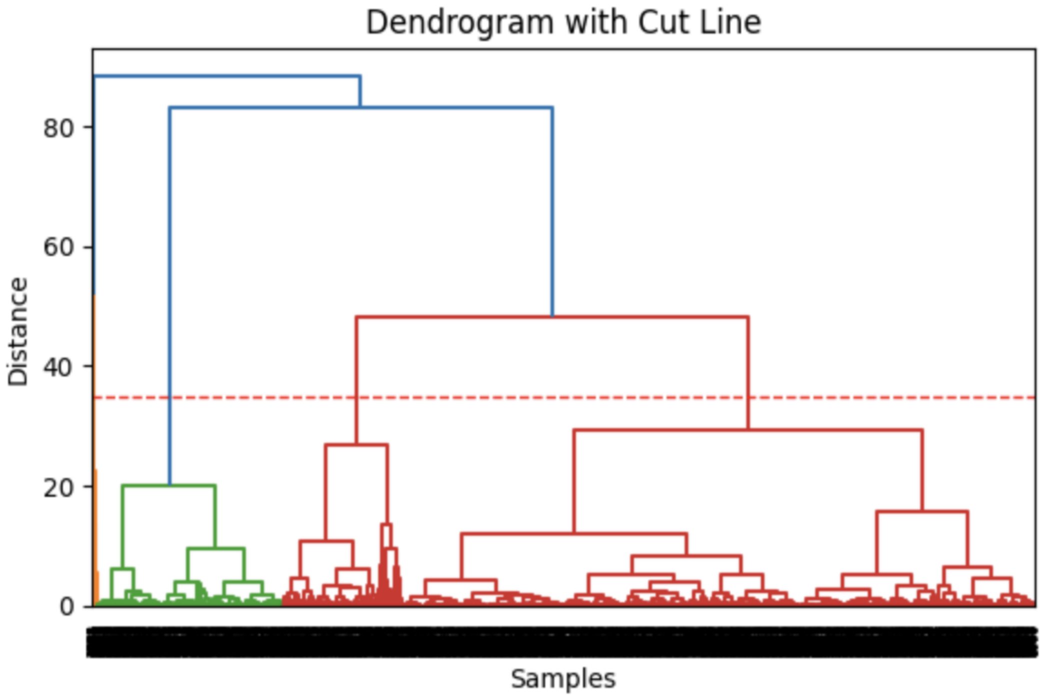
K-Means Clustering:

- With optimal k=3, used **PCA** to reduce dimensions and visualize clusters in 2D space.
- Identified initial rectangular edges in PCA plots, smoothed out clusters by **removing outliers (~692 customers).**
- Post-outlier removal, clusters showed **improved** cohesion (3.35) and separation (9.61).
- Why: Scalable, efficient for large datasets, and interprets customer

Hierarchical Clustering:

- Generated a dendrogram using **Ward's linkage &** **Euclidean distance** to assess cluster relationships.
- Horizontal **cut at 35** distance indicated a 3-cluster solution consistent with K-Means results.
- Applied **Agglomerative Clustering** for grouping customers into 3 hierarchical clusters.
- Why: Validates K-Means and visualizes customer relationships with dendrograms.



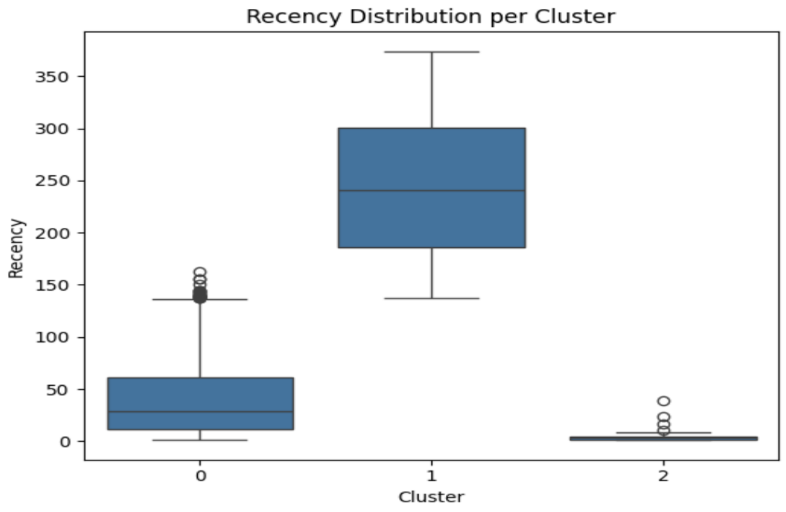


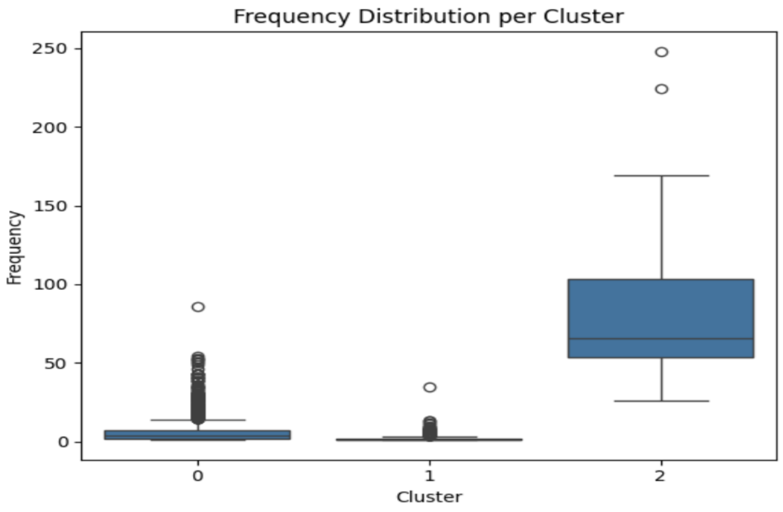
Results and Model Evaluation

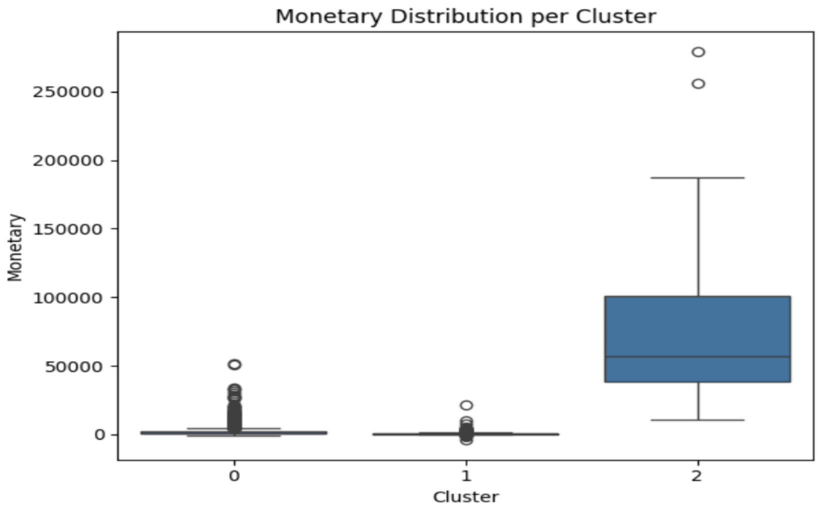
Aspect	K-Means Clustering	Hierarchical Clustering
Optimal Clusters	3 (determined via Elbow Method)	3 (validated via dendrogram with horizontal cut at 35 distance)
Silhouette Score	0.332(before), 0.435(after)	0.588
Cohesion (Intra-Cluster)	3.35	N/A
Separation (Inter-Cluster)	9.61	N/A
Cluster 0 (Moderate Buyers)	Avg. Monetary ~£1,822, steady but less frequent purchases	Avg. Monetary ~£1,709, steady spenders
Cluster 1 (Inactive Buyers)	Avg. Monetary ~£459, low engagement and spending	Avg. Monetary ~£450, disengaged customers
Cluster 2 (High-Value Buyers)	Avg. Monetary ~£81,836, frequent and loyal customers	Avg. Monetary ~£78,233, frequent and loyal customers

Model Selection:

- **K-Means** was chosen for its scalability, computational efficiency, and clear centroid-based clusters align with business strategies, offering actionable insights.
- Hierarchical Clustering, while offering slightly better silhouette scores, is less suitable for large datasets and lacks adaptability for real-time updates.

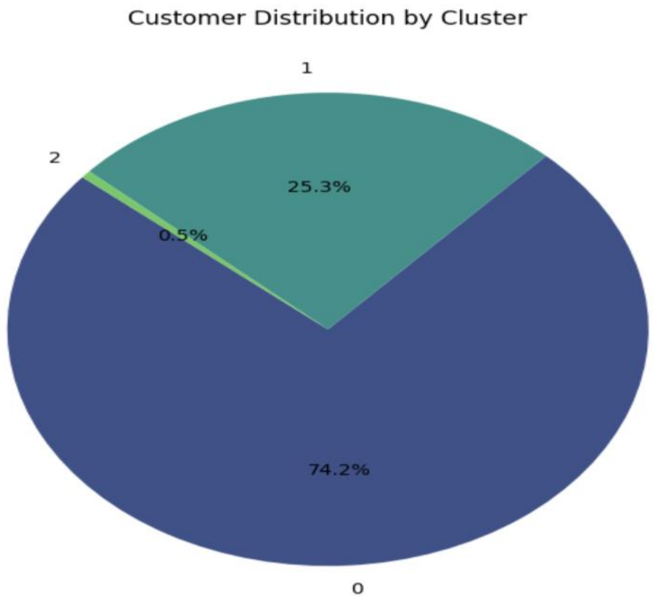


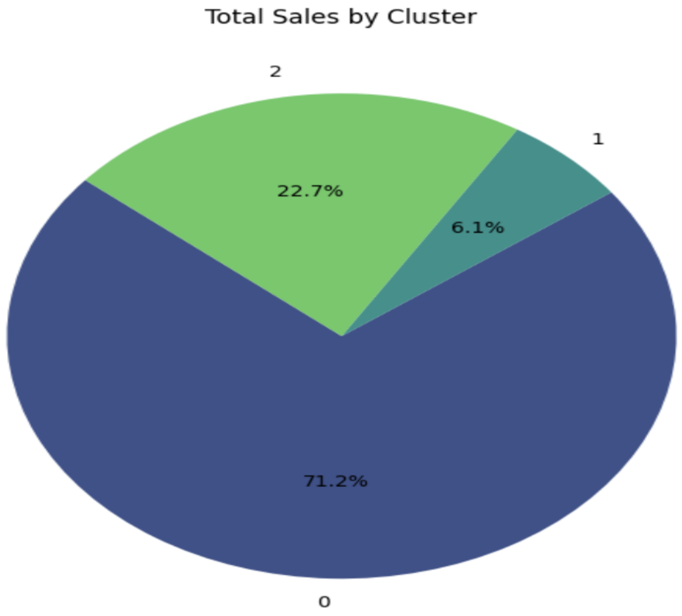




Interpreting Results in Business Context

Cluster	Key Insights
High-Value Buyers (Cluster 2)	Smallest group (0.5% of customers) generating 22.7% of total revenue. High-frequency purchases with substantial monetary value (£81,836)
Moderate Buyers (Cluster 0)	Largest group (74.2% of customers) contributing ~71.2% of revenue.
	Stable purchasing patterns.
Inactive Buyers (Cluster 1)	Contribute the least (~6.1% of revenue).
	Minimal engagement, high churn risk.
Beyond Managerial Intuition	Revenue contributions reveal disproportionate impact of certain groups.
	Patterns of disengaged buyers (Cluster 1) and concentrated revenue in Cluster 2 challenge conventional intuition.
	- Seasonal purchasing spikes and outlier patterns revealed hidden trends.





Managerial Insights and Recommendations

Actionable Insights

- High-Value Buyers(Cluster 2) : **Exclusive loyalty programs and personalized offers.**
- Moderate Buyers(Cluster 0): **Cross-sell, upsell, and seasonal promotions.**
- Inactive Buyers (Cluster 1): **Reactivation campaigns and engagement surveys.**
- Overall: Develop data-driven **pricing strategies** and monitor customer feedback for continuous improvement.

Implementation Ideas:

- Use **real-time segmentation dashboards** to track customer behavior.
- Establish a **quarterly review** process to reassess cluster assignments and adjust strategies accordingly.
- **Align marketing budgets with high-ROI** segments.

Strategic Value:

- **Boost revenue** through tailored customer strategies.
- **Enhance retention and reduce churn** through proactive engagement.
- **Strengthen customer relationships** by addressing specific needs and behaviors.