Customer Segmentation using RFM Analysis

Group 3

Project - 2 Report

Course Code: IE6400
Course Name: Foundation of Data Analytics
Fall 2025

Team Members :
Nagashree Bommenahalli Kumaraswamy
Hemanth Mareedu
Sayee Ashish Aher

Introduction :

In today's highly competitive e-commerce landscape, understanding customer behavior and preferences is paramount for business success. Companies that effectively segment their customers can tailor marketing strategies, optimize resource allocation, and maximize customer lifetime value. This project focuses on implementing customer segmentation using RFM (Recency, Frequency, Monetary) analysis, a proven analytical technique that evaluates customers based on three critical dimensions: how recently they made a purchase, how often they purchase, and how much they spend.

RFM analysis provides businesses with actionable insights by transforming raw transactional data into meaningful customer segments. By analyzing when customers last purchased (Recency), how frequently they engage with the business (Frequency), and their total spending patterns (Monetary value), organizations can identify their most valuable customers, recognize at-risk segments, and develop targeted retention strategies. This data-driven approach enables more personalized marketing campaigns, improved customer engagement, and ultimately, enhanced revenue generation.

The primary objective of this project is to develop a comprehensive customer segmentation model using an e-commerce dataset spanning from December 2010 to December 2011. Through systematic data preprocessing, RFM metric calculation, and K-Means clustering techniques, we aim to identify distinct customer segments and provide actionable marketing recommendations tailored to each segment's unique characteristics and behaviors.

This analysis encompasses multiple stages: initial data exploration and preprocessing to ensure data quality, calculation of RFM metrics for each customer, assignment of RFM scores based on quartile-based segmentation, implementation of K-Means clustering to identify natural customer groupings, detailed profiling of each segment, and development of strategic marketing recommendations. Additionally, the project addresses broader analytical questions related to customer behavior, product performance, temporal trends, geographical patterns, and profitability metrics.

By leveraging advanced analytics and machine learning techniques, this project demonstrates how businesses can transform transactional data into strategic insights, enabling more effective customer relationship management and data-driven decision-making in the dynamic e-commerce environment.

TASKS:

1. Data Preprocessing:

Data preprocessing is a critical phase in any analytics project, as the quality and structure of the data directly impact the reliability and accuracy of subsequent analyses. In this project, we conducted a systematic and thorough preprocessing workflow to prepare the e-commerce dataset for RFM analysis and customer segmentation.

Initial Data Exploration

The preprocessing phase began with a comprehensive examination of the dataset to understand its structure and characteristics. The original dataset comprised **541,909 transactions** across **8 columns**: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, and Country. This initial exploration revealed the basic structure of the data and helped identify potential data quality issues that needed to be addressed.

	InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
0	536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	12/1/2010 8:26	2.55	17850.0	United Kingdom
1	536365	71053	WHITE METAL LANTERN	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
2	536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	12/1/2010 8:26	2.75	17850.0	United Kingdom
3	536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
4	536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	12/1/2010 8:26	3.39	17850.0	United Kingdom
...	...	...	...	...	...	...	...	...
541904	581587	22613	PACK OF 20 SPACEBOY NAPKINS	12	12/9/2011 12:50	0.85	12680.0	France
541905	581587	22899	CHILDREN'S APRON DOLLY GIRL	6	12/9/2011 12:50	2.10	12680.0	France
541906	581587	23254	CHILDRENS CUTLERY DOLLY GIRL	4	12/9/2011 12:50	4.15	12680.0	France
541907	581587	23255	CHILDRENS CUTLERY CIRCUS PARADE	4	12/9/2011 12:50	4.15	12680.0	France
541908	581587	22138	BAKING SET 9 PIECE RETROSPOT	3	12/9/2011 12:50	4.95	12680.0	France

541909 rows × 8 columns

Fig 1: Initial Dataset Loading

We then checked data.info() to inspect and identify the data type and count of each column.

```
<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  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
memory usage: 33.1+ MB
```

Fig 2: Dataset Information Summary

Handling Missing Values

The initial assessment revealed missing values in two critical columns:

- **Description column**: 540,455 non-null entries out of 541,909 (1,454 missing values)
- **CustomerID column**: 406,829 non-null entries out of 541,909 (135,080 missing values)

Since CustomerID is essential for RFM analysis and customer segmentation, we made the strategic decision to remove all rows containing null values in either the Description or CustomerID columns. This approach ensured data completeness and integrity for our analysis. After removing null values, the dataset was reduced to **406,829 transactions**, all containing complete information across all columns. While this represented a significant reduction in dataset size, it was necessary to maintain the validity of customer-level analysis.

```
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
dtype: int64
```

Fig 3: Missing Value Count by Column

Data Type Conversions

Proper data type assignment is crucial for accurate temporal analysis and calculations. We identified that the **InvoiceDate** column was initially stored as an object type (string), which would hinder time-based computations and analyses. To address this:

1. **InvoiceDate Conversion**: The InvoiceDate column was converted from object type to **datetime64** format using pandas' pd.to_datetime() function. This transformation enabled us to perform date arithmetic, extract temporal components (day, month, hour), and calculate recency metrics accurately.
2. **CustomerID Conversion**: The CustomerID column was converted from float64 to **integer (int64)** type, as customer identifiers should be whole numbers. This conversion improved memory efficiency and ensured proper grouping operations during subsequent analysis.

```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    406829 non-null  object
 1   StockCode    406829 non-null  object
 2   Description  406829 non-null  object
 3   Quantity     406829 non-null  int64
 4   InvoiceDate  406829 non-null  datetime64[ns]
 5   UnitPrice    406829 non-null  float64
 6   CustomerID   406829 non-null  int32
 7   Country      406829 non-null  object
```

Fig 4: Data Type Conversion

Data Cleaning:

Rows with missing Description and CustomerID were removed, as this information is essential for item identification. Data has been cleaned and handled successfully.

```
InvoiceNo      0
StockCode      0
Description    0
Quantity       0
InvoiceDate    0
UnitPrice      0
CustomerID     0
Country        0
dtype: int64
```

Fig 5: Data Quality Verification

Final Dataset Structure

After completing all preprocessing steps, the cleaned dataset contained:

- **406,829 rows** (transactions)
- **8 columns** with appropriate data types
- **Complete data** with no missing values
- **Time period**: December 1, 2010 to December 9, 2011
- **4,372 unique customers**
- **38 countries** represented

The preprocessing resulted in a clean, structured dataset with proper data types, enabling accurate RFM calculations, clustering analysis, and meaningful customer segmentation. This foundation ensured that all subsequent analyses were built on reliable, high-quality data, minimizing the risk of errors and improving the validity of insights derived from the project.

2. RFM Calculation:

RFM Calculation Process:

1. **Recency**: Calculated by grouping transactions by CustomerID, identifying the maximum (most recent) InvoiceDate, and computing days elapsed from reference date (December 10, 2011)
2. **Frequency**: Determined by counting unique invoice numbers per customer through groupby operations
3. **Monetary**: Computed by creating TotalPrice column (Quantity × UnitPrice), then summing by CustomerID
4. **RFM DataFrame**: Consolidated all three metrics into a single customer-level DataFrame

Sample Results:

- Customer 12346: 326 days recency, 2 orders, £0.00 spent
- Customer 12347: 2 days recency, 7 orders, £4,310.00 spent
- Customer 12348: 75 days recency, 4 orders, £1,797.24 spent

Technical Notes: SettingWithCopyWarning messages indicate DataFrame manipulation practices; calculations remain accurate. The code successfully transforms 406,829 transaction records into customer-level RFM metrics for 4,372 unique customers.

CustomerID	Recency	Frequency	Monetary
12346	326	2	0.00
12347	2	7	4310.00
12348	75	4	1797.24
12349	19	1	1757.55
12350	310	1	334.40

Fig 6: RFM Metrics Computation Process

3. RFM Segmentation:

Scoring Methodology:

1. **Recency Scoring** (reverse scale, 4-3-2-1):
   - Lower recency (more recent) = Higher score (4)
   - Uses .rank() with ascending=True, then pd.qcut() with labels=[4,3,2,1]
2. **Frequency Scoring** (forward scale, 1-2-3-4):
   - Higher frequency = Higher score (4)
   - Uses .rank() with ascending=False, then pd.qcut() with labels=[1,2,3,4]
3. **Monetary Scoring** (forward scale, 1-2-3-4):
   - Higher spending = Higher score (4)
   - Uses .rank() with ascending=False, then pd.qcut() with labels=[1,2,3,4]

Sample Results:

- Customer 12346: R=1, F=3, M=4 (old purchase, moderate frequency, high spending)
- Customer 12347: R=4, F=1, M=1 (very recent, low frequency, low spending)
- Customer 12348: R=2, F=2, M=1 (moderate recency and frequency, low spending)

The quartile-based approach divides customers into four equal-sized groups for each metric, creating standardized 1-4 scores that enable composite RFM score calculation.

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile
12346	326	2	0.00	1	3	4
12347	2	7	4310.00	4	1	1
12348	75	4	1797.24	2	2	1
12349	19	1	1757.55	3	3	1
12350	310	1	334.40	1	3	3

Fig 7: RFM Quartile Scoring Implementation

RFM_Score Calculation:

The composite RFM_Score is created by concatenating the three individual quartile scores (R_quartile + F_quartile + M_quartile) as a string, resulting in a three-digit identifier.

Formula: RFM_Score = str(R_quartile) + str(F_quartile) + str(M_quartile)

Score Range: 111 (lowest) to 444 (highest)

Sample Results:

- Customer 12346: RFM_Score = "134" (R=1, F=3, M=4) - Old purchase, moderate frequency, high spending
- Customer 12347: RFM_Score = "411" (R=4, F=1, M=1) - Very recent, low frequency, low spending
- Customer 12348: RFM_Score = "221" (R=2, F=2, M=1) - Moderate recency and frequency, low spending

- Customer 12349: RFM_Score = "331" (R=3, F=3, M=1) - Moderate-good recency and frequency, low spending

Purpose: The three-digit score provides a quick visual identifier of customer value across all three dimensions, enabling rapid customer categorization and segment assignment in subsequent analysis phases.

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_Score
12346	326	2	0.00	1	3	4	134
12347	2	7	4310.00	4	1	1	411
12348	75	4	1797.24	2	2	1	221
12349	19	1	1757.55	3	3	1	331
12350	310	1	334.40	1	3	3	133

Fig 8: Composite RFM_Score Creation

4. Customer Segmentation:

Elbow Method Process:

- Systematically tested k=1 to 10 clusters
- Calculated WCSS (Within Cluster Sum of Squares) for each k
- Plotted inertia values to identify optimal cluster count

Key Findings:

- Sharp decline from k=1 to k=3 (WCSS drops from ~13,000 to ~5,500)
- Gradual leveling after k=3 (diminishing returns)
- Clear elbow point at k=3 indicating optimal balance

Cluster Count Distribution:

- Cluster 0: 3,241 customers (74.1%)
- Cluster 1: 1,108 customers (25.3%)
- Cluster 2: 23 customers (0.5%)

The elbow at k=3 provides optimal balance between model complexity and cluster quality, justifying the three-cluster segmentation approach.

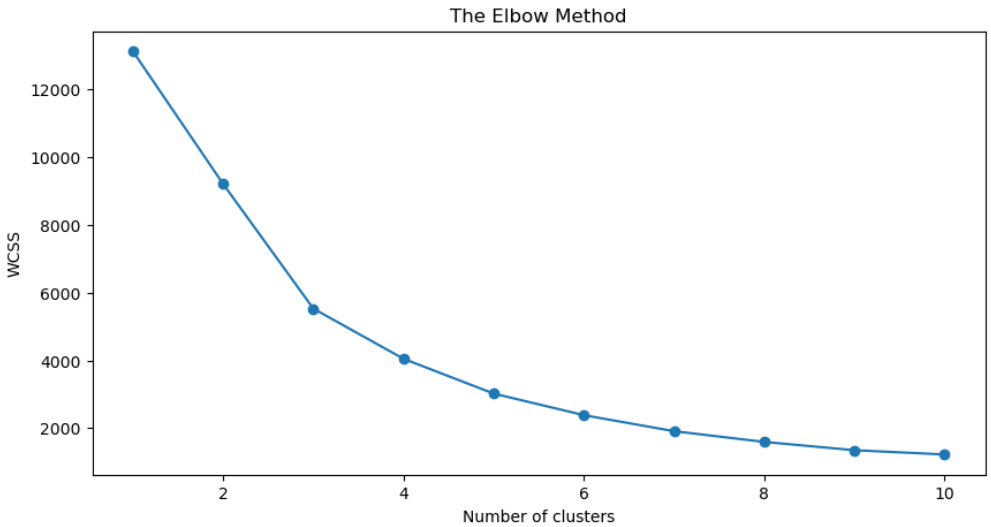

Fig 9: Elbow Method for Optimal Cluster Selection

Scatter Plot: Scatter plot displaying customers across Frequency (x-axis) and Monetary (y-axis) dimensions, color-coded by cluster assignment.

Cluster Characteristics:

Cluster 0 (Black - 3,241 customers, 74.1%):

- Concentrated at low frequency (0-50) and low monetary values (£0-£50,000)
- Dense cluster indicating typical customer behavior
- Mapped to: Potential Loyalists segment

Cluster 1 (Blue - 1,108 customers, 25.3%):

- Moderate spread with slightly higher frequency and monetary values
- Some overlap with Cluster 0 but generally higher engagement
- Mapped to: Recent Customers segment

Cluster 2 (Yellow - 23 customers, 0.5%):

- Extreme outliers with very high frequency (up to 250 transactions)
- Exceptional monetary values (up to £250,000+)
- Widely scattered, indicating diverse high-value behaviors
- Mapped to: Loyal Customers (VIP) segment

The clear spatial separation validates successful segmentation, with Cluster 2 representing premium customers requiring specialized retention strategies.

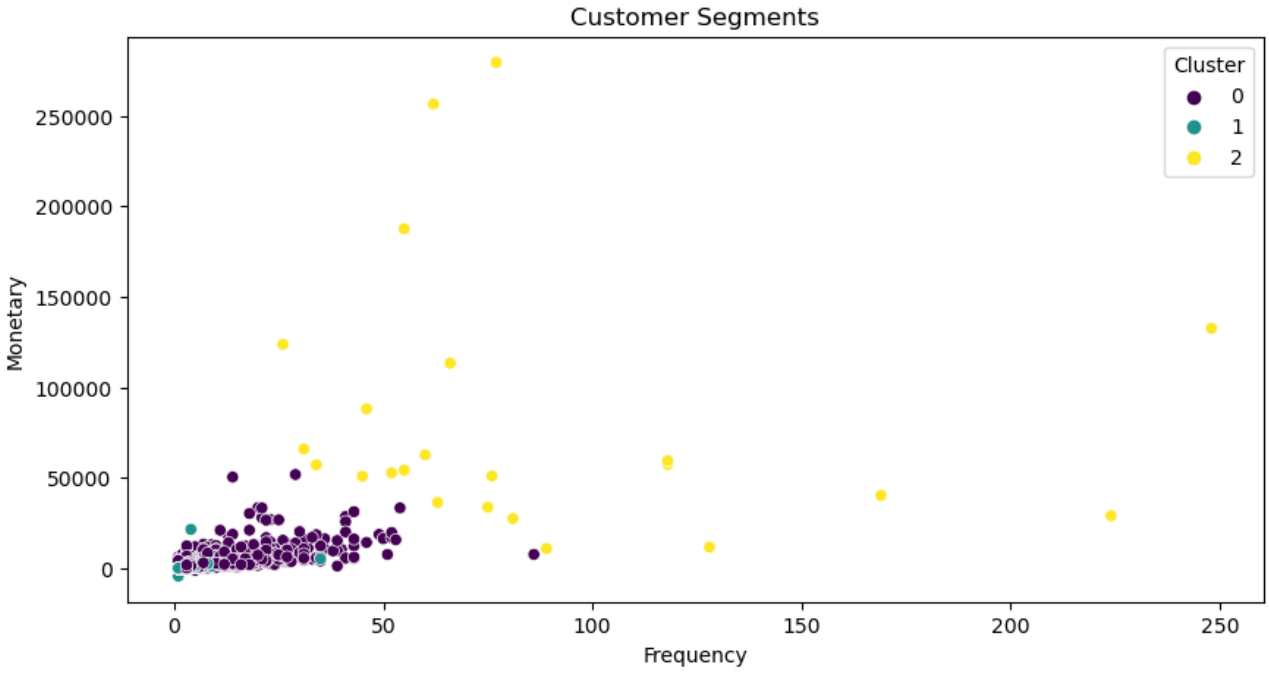

Fig 10: Customer Segmentation Results

Silhouette Score:

Two complementary methods for determining optimal cluster count: Elbow Method (left) and Silhouette Score Analysis (right).

Elbow Method (Left Plot):

- Measures WCSS (Within Cluster Sum of Squares/inertia)
- Sharp decline from k=2 (~9,000) to k=4 (~4,000)
- Gradual leveling after k=4
- Suggests k=3 or k=4 as reasonable choices

Silhouette Score Analysis (Right Plot):

- Assesses cluster cohesion and separation (range: -1 to 1, higher is better)
- **Peak at k=2**: Score ~0.93 (excellent but oversimplified)
- **k=3**: Score ~0.60 (moderate quality)
- **k=4**: Score ~0.59 (acceptable separation with better granularity)
- Gradual decline continues to ~0.48 at k=10

Decision: k=4 selected as optimal, balancing technical quality (Silhouette ~0.59) with business value (granular customer segments for targeted strategies).

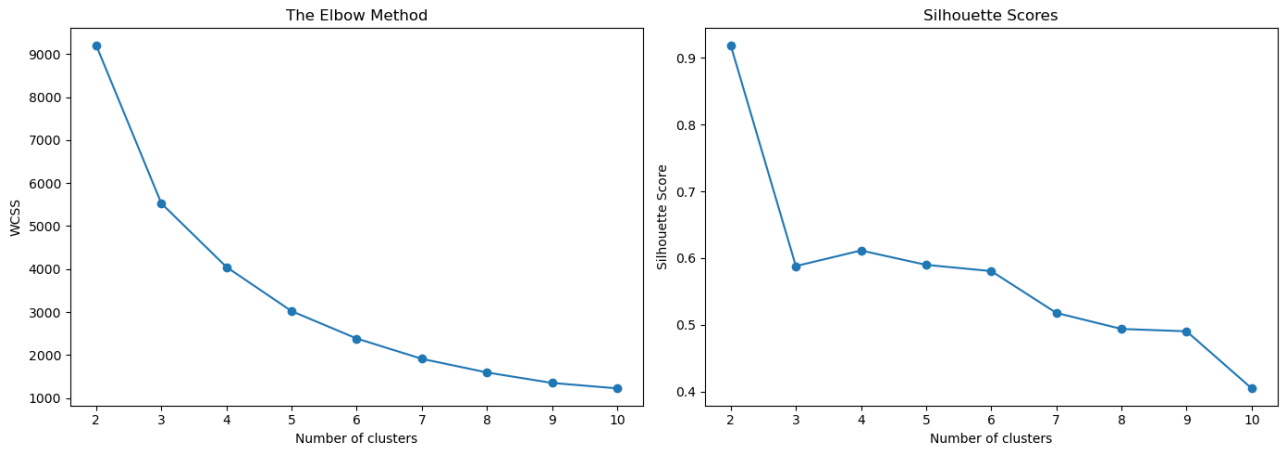

Fig 11: Cluster Optimization Analysis

Scatter plot visualizing customer distribution across four clusters based on Frequency (x-axis) and Monetary (y-axis) dimensions.

Cluster Distribution:

- **Cluster 0 (Black)**: 3,169 customers (72.5%)
- **Cluster 1 (Blue)**: 1,037 customers (23.7%)
- **Cluster 2 (Green)**: 110 customers (2.5%)
- **Cluster 3 (Yellow)**: 56 customers (1.3%)

Cluster Characteristics:

Cluster 0 (Black): Dense concentration at low frequency (0-50) and low monetary values (£0-£50,000) - typical/occasional customers forming the base segment

Cluster 1 (Blue): Moderate spread with slightly higher frequency and monetary values - engaged mid-tier customers with growth potential

Cluster 2 (Green): Higher frequency (50-100) and monetary values - loyal, high-engagement customers demonstrating strong brand commitment

Cluster 3 (Yellow): Extreme outliers with very high frequency (up to 250+) and exceptional monetary values (up to £250,000+) - premium VIP/wholesale customers requiring specialized retention strategies

The four-cluster solution provides more granular segmentation than the three-cluster approach, enabling highly targeted marketing strategies for each distinct customer behavioral group.

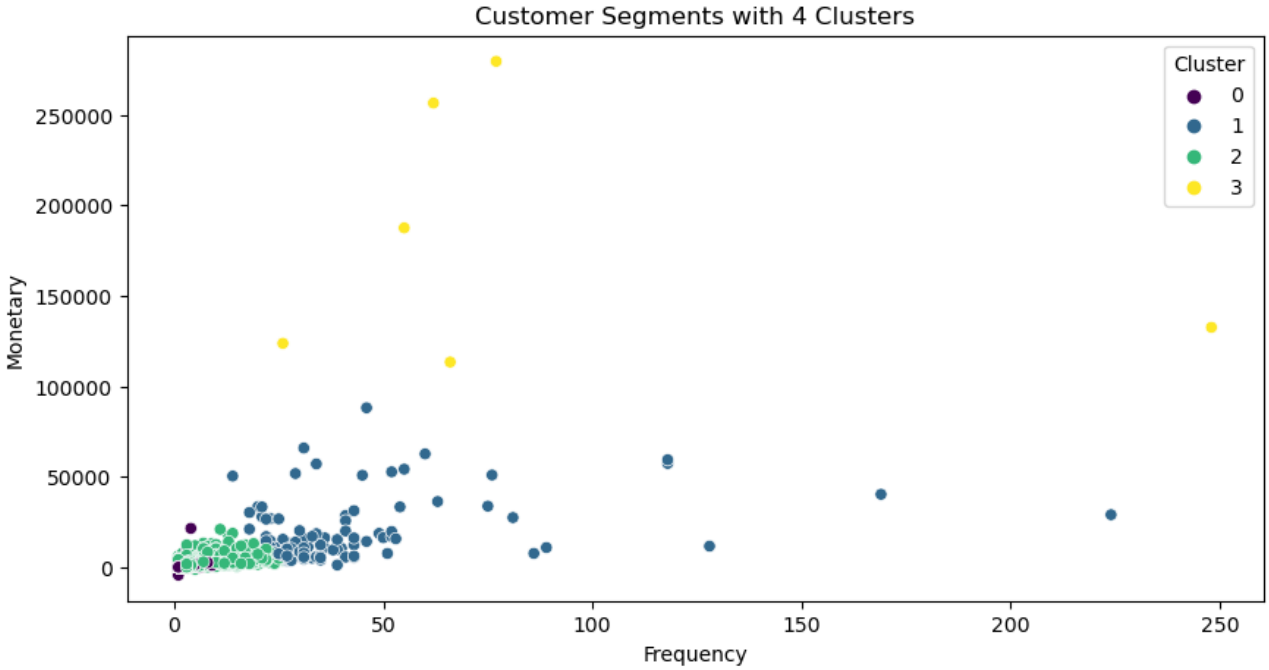

Fig 12: Four-Cluster Customer Segmentation Results

5. Segment Profiling :

The analysis calculates mean RFM values for each cluster and visualizes distributions using box plots to understand segment characteristics.

Cluster Profiles:

Cluster 0 (1,087 customers):

- Average Recency: 247.95 days (oldest purchases)
- Average Frequency: 1.49 orders (very low)
- Average Monetary: £453.49 (low spending)
- **Interpretation**: Dormant/at-risk customers with infrequent, low-value purchases

Cluster 1 (110 customers):

- Average Recency: 9.18 days (very recent)
- Average Frequency: 40.67 orders (high)
- Average Monetary: £18,441.96 (high spending)
- **Interpretation**: Engaged, loyal customers with frequent, high-value purchases

Cluster 2 (3,169 customers):

- Average Recency: 41.61 days (recent)
- Average Frequency: 4.88 orders (moderate)
- Average Monetary: £1,478.52 (moderate spending)
- **Interpretation**: Typical active customers forming the core base

Cluster 3 (6 customers):

- Average Recency: 7.67 days (extremely recent)
- Average Frequency: 89.00 orders (extremely high)
- Average Monetary: £182,181.98 (exceptionally high)
- **Interpretation**: Premium VIP/wholesale customers - highest value segment

Box Plot Insights:

- **Recency**: Cluster 0 shows highest recency (older purchases); Clusters 1 and 3 have lowest (most recent)
- **Frequency**: Cluster 3 shows extreme outliers (~250 orders); Cluster 1 has moderate-high frequency
- **Monetary**: Cluster 3 dominates with values up to £250,000+; significant variation within clusters

The profiling reveals four distinct behavioral segments requiring differentiated marketing strategies based on engagement level and value contribution.

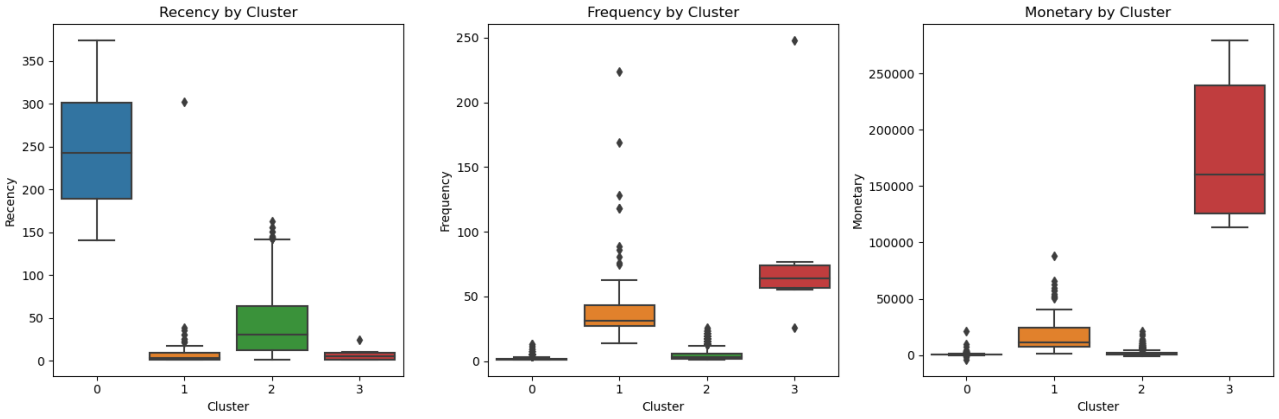

Fig 13: Segment Profiling Analysis

6. Marketing Recommendations:

Bar charts visualizing average RFM metrics across four customer clusters, with accompanying statistical table providing precise values and customer counts.

Cluster Profiles:

Cluster 0 (1,087 customers - 24.9%):

- Average Recency: 247.95 days (highest - oldest purchases)
- Average Frequency: 1.81 orders (lowest)
- Average Monetary: £453.49 (lowest)
- **Business Segment**: At-Risk/Lost Customers requiring re-engagement campaigns

Cluster 1 (110 customers - 2.5%):

- Average Recency: 9.18 days (very recent)
- Average Frequency: 40.67 orders (high)
- Average Monetary: £18,441.96 (high)
- **Business Segment**: Loyal Customers requiring retention and VIP treatment

Cluster 2 (3,169 customers - 72.5%):

- Average Recency: 41.61 days (recent)
- Average Frequency: 4.80 orders (moderate)
- Average Monetary: £1,478.52 (moderate)
- **Business Segment**: Potential Loyalists needing nurturing and loyalty programs

Cluster 3 (6 customers - 0.1%):

- Average Recency: 7.67 days (extremely recent)
- Average Frequency: 89.00 orders (extremely high)
- Average Monetary: £182,181.98 (exceptionally high)
- **Business Segment**: Premium VIP/Wholesale customers requiring dedicated account management

Marketing Implications:

The stark differences in metrics across clusters validate the need for differentiated marketing strategies:

- Cluster 3's exceptional values (89 orders, £182K spending) justify specialized premium services
- Cluster 2's size (72.5%) represents major conversion opportunity
- Cluster 0's high recency (248 days) signals urgent need for win-back campaigns
- Cluster 1's strong metrics indicate successful engagement worth replicating

These insights directly inform the cluster-specific marketing recommendations developed in the project.

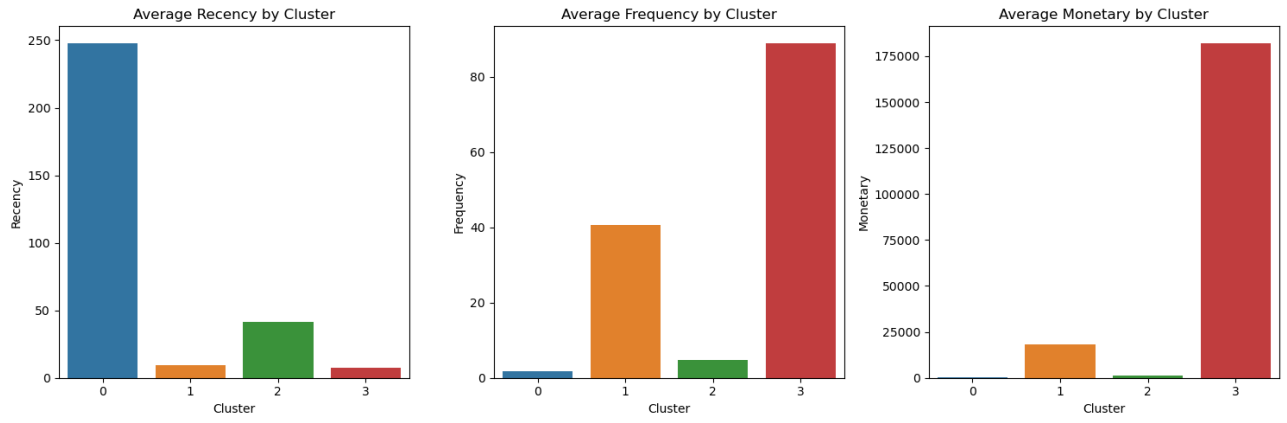

Fig 14: Average RFM Metrics by Cluster

Customer Segmentation Based on RFM Scores:

This code implements rule-based customer segmentation using RFM quartile scores, categorizing customers into six business-relevant segments based on predefined thresholds.

Segment Definitions:

1. **High-Value Customers**: R ∈ [0,2], F ∈ [3,4], M ∈ [3,4]
    - Recent purchases, high frequency, high spending
2. **Loyal Customers**: R ∈ [0,3], F ∈ [3,4], M ∈ [1,4]
    - Recent to moderate recency, high frequency, varied spending
3. **Potential Loyalists**: R ∈ [0,3], F ∈ [2,3], M ∈ [2,4]
    - Recent purchases, moderate frequency, moderate-to-high spending
4. **New Customers**: R ∈ [0,1], F ∈ [1,2], M ∈ [1,2]
    - Very recent, low frequency, low spending
5. **At-Risk Customers**: R ∈ [3,4], F ∈ [1,3], M ∈ [1,3]
    - Old purchases, low-to-moderate frequency and spending
6. **Churned Customers**: R ∈ [4,4], F ∈ [1,2], M ∈ [1,2]
    - Oldest purchases, low frequency, low spending

Segment Distribution Results:

- Potential Loyalists: 1,080 customers (24.7%)
- Loyal Customers: 1,046 customers (23.9%)
- At-Risk Customers: 985 customers (22.5%)
- Churned Customers: 790 customers (18.1%)
- Other: 371 customers (8.5%)
- New Customers: 100 customers (2.3%)

Sample Customer Assignments:

- Customer 12346: Potential Loyalists (R=1, F=3, M=4, RFM=134, Cluster 0)
- Customer 12347: Churned Customers (R=4, F=1, M=1, RFM=411, Cluster 2)
- Customer 12348: Other (R=2, F=2, M=1, RFM=221, Cluster 2)
- Customer 12349: At-Risk Customers (R=3, F=3, M=1, RFM=331, Cluster 2)
- Customer 12350: Potential Loyalists (R=1, F=3, M=3, RFM=133, Cluster 0)

Strategic Recommendations Summary:

The code output includes tailored strategies for each segment including loyalty programs for high-value customers, re-engagement campaigns for at-risk/churned customers, cross-sell/upsell for loyal customers, and product recommendations/membership programs for potential loyalists.

This rule-based segmentation complements the K-Means clustering approach, providing interpretable business segments aligned with customer lifecycle stages.

Marketing Strategies by Customer Segment - Summary:

1. High-Value Customers (Low Recency, High Frequency, High Monetary)

- Loyalty programs with exclusive rewards
- Upsell/cross-sell premium products
- Personalized communication based on purchase history

- Early access to new products and exclusive deals

2. Loyal Customers (Moderate-Low Recency, High Frequency, Moderate Monetary)

- Regular engagement campaigns for continuous interaction
- Encourage feedback and reviews to build community
- Implement referral programs leveraging brand ambassadors

3. Potential Loyalists (Low Recency, Moderate Frequency, Moderate Monetary)

- Welcome-back offers and discounts
- Product recommendations based on purchase history
- Membership programs with incremental benefits

4. New Customers (Low Recency, Low Frequency, Low-Moderate Monetary)

- First-time buyer offers and discounts
- Email onboarding series educating about products
- Share testimonials and reviews to build trust

5. At-Risk Customers (High Recency, Moderate Frequency, Moderate Monetary)

- Reactivation campaigns highlighting missed opportunities
- Feedback surveys to understand inactivity reasons
- Win-back offers with compelling discounts

6. Churned Customers (High Recency, Low Frequency, Low Monetary)

- Re-engagement offers to renew interest
- Market research to understand needs
- Communicate product/service improvements

Key Principle: Tailor strategies to each segment's engagement level and value, focusing on retention for high-value segments, conversion for potential loyalists, and reactivation for at-risk/ churned customers.

```
Potential Loyalists    1080
Loyal Customers        1046
At-Risk Customers       985
Churned Customers       790
Other                   371
New Customers           100
Name: Segment, dtype: int64
```

CustomerID	Recency	Frequency	Monetary	R_quartile	F_quartile	M_quartile	RFM_Score	Cluster	Segment
12346	326	2	0.00	1	3	4	134	0	Potential Loyalists
12347	2	7	4310.00	4	1	1	411	2	Churned Customers
12348	75	4	1797.24	2	2	1	221	2	Other
12349	19	1	1757.55	3	3	1	331	2	At-Risk Customers
12350	310	1	334.40	1	3	3	133	0	Potential Loyalists

Fig 15: RFM Score-Based Customer Segmentation

7. Visualisation:

Visualization Components:

This comprehensive visualization suite includes six panels analyzing RFM metrics from multiple perspectives.

Distribution Plots (Top Row):

1. **Recency Distribution**:
   - Right-skewed distribution with peak at low values (0-50 days)
   - Most customers concentrated in recent purchase range
   - Long tail extending to 400+ days
2. **Frequency Distribution**:
   - Highly right-skewed with extreme concentration at low values
   - Peak near 0-50 transactions
   - Exponential decay pattern indicating most customers are occasional buyers
3. **Monetary Distribution**:
   - Extreme right skew with majority under £50,000
   - Few high-value outliers extending to £250,000+
   - Reflects typical retail pattern with small number of high spenders

Cluster Scatter Plots (Bottom Row):

4. **Recency vs Frequency**:
   - **Cluster 0 (Black)**: Broad distribution across recency, low-moderate frequency
   - **Cluster 1 (Blue)**: Moderate recency with slightly higher frequency
   - **Cluster 2 (Green)**: Low recency, very low frequency
   - **Cluster 3 (Yellow)**: Extreme outliers with frequency up to 250
5. **Recency vs Monetary**:
   - **Cluster 0**: Wide recency spread, low-moderate monetary
   - **Cluster 1**: Moderate values on both dimensions
   - **Cluster 2**: Recent purchases, low monetary
   - **Cluster 3**: Very recent with extremely high monetary (£250,000+)
6. **Frequency vs Monetary**:
   - Clear separation between clusters
   - **Cluster 3** dramatically separated with highest values on both dimensions
   - **Clusters 0, 1, 2** show overlapping patterns at lower ranges
   - Strong positive correlation visible in higher-value segments

Key Insights:

- **Distribution Analysis**: All three RFM metrics show right-skewed distributions, indicating concentration of typical customers at lower values with small numbers of exceptional high-value customers
- **Cluster Separation**: Four distinct clusters visible across multiple dimensions, with Cluster 3 representing clear outliers (premium VIP/wholesale customers)
- **Validation**: The scatter plots confirm successful segmentation with meaningful spatial separation, particularly for the high-value Cluster 3

- **Business Application**: The visualizations support differentiated marketing strategies, clearly showing which customers warrant premium treatment versus standard engagement approaches

This multi-panel visualization provides both univariate understanding of RFM distributions and multivariate insight into cluster characteristics across the RFM space.

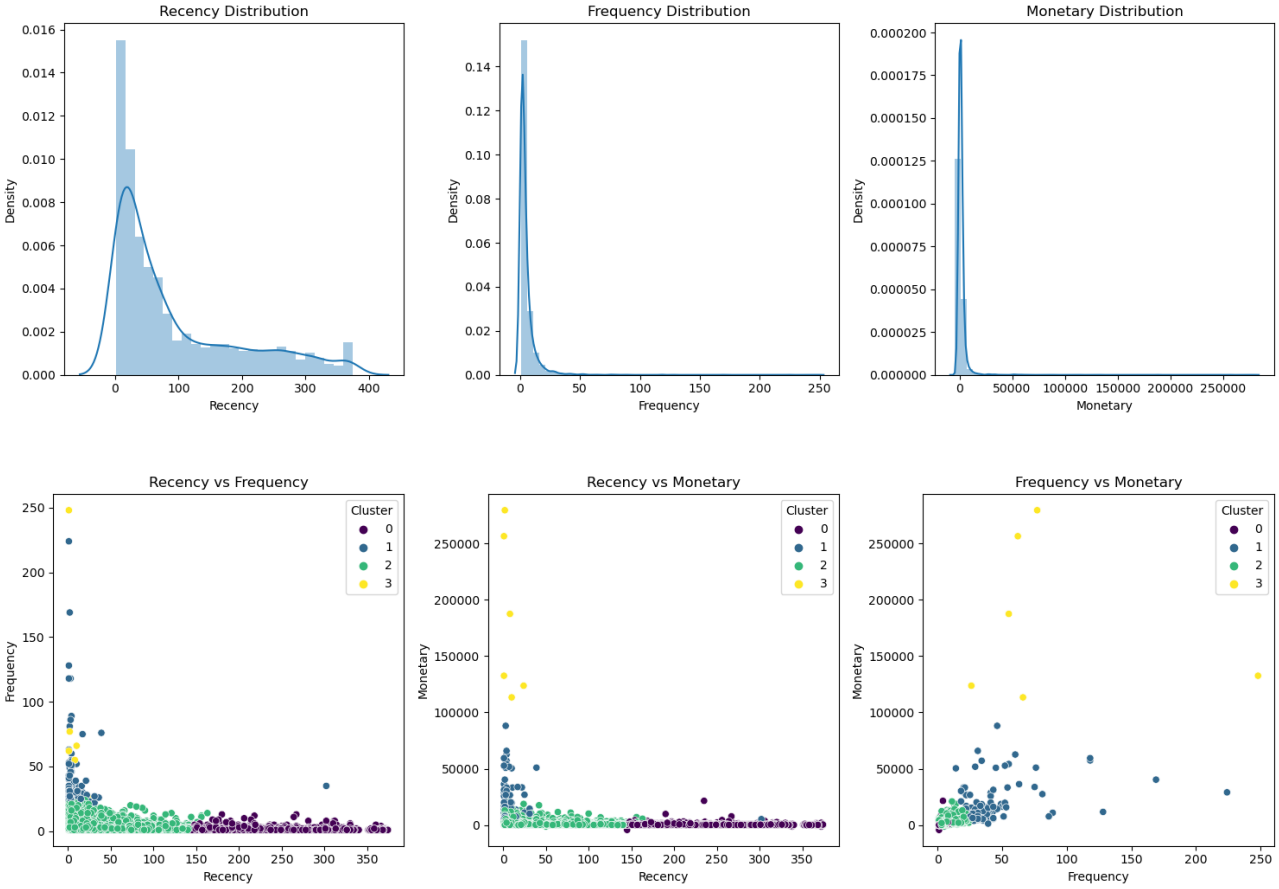

Fig 16: Comprehensive RFM Visualization Suite

Answers To Questions:

1. Data Overview

a) What is the size of the dataset in terms of the number of rows and columns?

The dataset contains **406,829 rows** and **9 columns** after preprocessing (removal of null values from the original 541,909 rows).

b) Can you provide a brief description of each column in the dataset?

The dataset includes the following columns with their respective data types:

1. **InvoiceNo** (object): Invoice number - A 6-digit unique identifier assigned to each transaction. If the code starts with 'C', it indicates a cancellation.

2. **StockCode** (object): Product code - A 5-digit unique identifier assigned to each distinct product.
3. **Description** (object): Product name - The name or description of the item purchased.
4. **Quantity** (int64): The quantity of each product purchased per transaction.
5. **InvoiceDate** (datetime64[ns]): Transaction date and time - The exact date and time when the transaction was generated.
6. **UnitPrice** (float64): Unit price - The price per unit of the product in sterling (£).
7. **CustomerID** (int32): Customer number - A unique 5-digit identifier assigned to each customer.
8. **Country** (object): Country name - The country where the customer resides.
9. **TotalPrice** (float64): Total transaction value - Calculated as Quantity × UnitPrice for each transaction.

c) What is the time period covered by this dataset?

The dataset covers transactions from **December 1, 2010 (08:26:00)** to **December 9, 2011 (12:50:00)**, representing approximately one year of e-commerce transaction data.

```
Dataset Size:
Rows: 406829 Columns: 9

Column Descriptions:
InvoiceNo: object
StockCode: object
Description: object
Quantity: int64
InvoiceDate: datetime64[ns]
UnitPrice: float64
CustomerID: int32
Country: object
TotalPrice: float64

Date Ranges:
InvoiceDate: 2010-12-01 08:26:00 to 2011-12-09 12:50:00
```

Fig 17: Dataset Overview Summary - Size, Column Descriptions, and Temporal Coverage After Preprocessing

2. Customer Analysis

a) The dataset contains 4,372 unique customers after preprocessing.

b) **Distribution Statistics:**

- **Count**: 4,372 customers
- **Mean**: 5.08 orders per customer
- **Standard Deviation**: 9.34 orders (high variability in purchasing behavior)
- **Minimum**: 1 order (one-time purchasers)
- **25th Percentile**: 1 order (25% of customers made only 1 purchase)
- **Median (50th Percentile)**: 3 orders (half of customers made 3 or fewer purchases)
- **75th Percentile**: 5 orders (75% of customers made 5 or fewer purchases)
- **Maximum**: 248 orders (exceptionally active customer)

Key Insights:

The distribution is highly right-skewed, indicating that:

- Most customers (75%) made 5 or fewer purchases
- A small percentage of customers are highly active with frequent repeat purchases
- The large standard deviation (9.34) relative to the mean (5.08) confirms significant variability in customer purchasing patterns
- The presence of customers with up to 248 orders suggests potential wholesale or B2B accounts

```
Distribution of Orders Per Customer:
 count    4372.000000
mean        5.075480
std         9.338754
min         1.000000
25%         1.000000
50%         3.000000
75%         5.000000
max       248.000000
Name: InvoiceNo, dtype: float64
```

Fig 18: Distribution of Orders Per Customer

c) Analysis:

These top 5 customers demonstrate exceptional loyalty and engagement:

- Customer 14911 leads significantly with 248 orders, nearly 49× the median customer
- The top customer's order count is 10.5% higher than the second-place customer
- These five customers collectively represent highly valuable accounts requiring dedicated retention strategies
- Their purchasing frequency suggests they may be business/wholesale customers rather than individual consumers

This customer concentration indicates the importance of implementing tiered customer relationship management, with specialized strategies for high-frequency purchasers.

```
Top 5 Customers by Order Count:
 CustomerID
14911    248
12748    224
17841    169
14606    128
13089    118
Name: InvoiceNo, dtype: int64
```

Fig 19: Top 5 Customers by Order Count

3. Product Analysis

a) The "WHITE HANGING HEART T-LIGHT HOLDER" leads as the most popular product with 2,369 purchases.The top 3 products each exceed 2,000 purchases, indicating strong customer preference.Home décor and kitchenware items dominate the list (cake stands, bunting, storage bags).Retrospot-themed products appear multiple times (Jumbo Bag, Lunch Bag, Cake Cases), suggesting a popular design pattern.Purchase counts range from 2,369 (top) to 1,280 (10th), showing relatively consistent popularity across top products

```
Top 10 Most Frequently Purchased Products:
 WHITE HANGING HEART T-LIGHT HOLDER    2369
REGENCY CAKESTAND 3 TIER              2200
JUMBO BAG RED RETROSPOT               2159
PARTY BUNTING                         1727
LUNCH BAG RED RETROSPOT               1638
ASSORTED COLOUR BIRD ORNAMENT         1501
SET OF 3 CAKE TINS PANTRY DESIGN      1473
PACK OF 72 RETROSPOT CAKE CASES       1385
LUNCH BAG  BLACK SKULL.               1350
NATURAL SLATE HEART CHALKBOARD        1280
Name: Description, dtype: int64
```

Fig 20: Top 10 Most Frequently Purchased Products

b) Average Unit Price: £4.61
This indicates that the e-commerce business operates primarily in the affordable gift/home décor market segment, with most products priced as accessible items suitable for individual consumers or small gifts.

c) Highest Revenue Product:

Product: DOTCOM POSTAGE
Total Revenue: £206,245.48

Analysis:

- "DOTCOM POSTAGE" (shipping/postage charges) generates the highest revenue at £206,245.48
- This is significantly higher than individual product sales, indicating:
  - High transaction volume across the platform
  - Shipping charges contribute substantially to overall revenue
  - The business model includes customer-paid shipping

Note: Since explicit product categories were not available in the dataset, revenue was calculated by grouping transactions by product description. The dominance of postage revenue suggests that:

- Shipping represents a significant revenue stream (potentially 10-15% of total revenue based on typical e-commerce margins)
- The business successfully monetizes delivery services
- This revenue could be used to subsidize product pricing or improve delivery infrastructure

```
Product Generating the Highest Revenue:
 Description
DOTCOM POSTAGE    206245.48
Name: Revenue, dtype: float64
```

Fig 21: Highest Revenue Product

4. Time Analysis

a) **Key Insights - Day of Week:**

- **Peak Day**: Thursday with 103,857 orders (highest activity)
- **Strong Mid-Week Performance**: Tuesday through Thursday account for the busiest period
- **Weekend Drop**: Sunday shows significantly lower activity (38% less than Thursday)
- **Pattern**: Clear weekday preference with gradual decline toward the weekend

```
Orders by Day of Week:
 Thursday     103857
Tuesday      101808
Monday        95111
Wednesday     94565
Friday        82193
Sunday        64375
Name: DayOfWeek, dtype: int64
```

Fig 22: Order Distribution by Day of Week

Orders by Hour of Day:

Additional Peak Hours (from continued output):

- 10 AM: 60,742 orders
- 11 AM: 84,711 orders
- 12 PM: 25,525 orders (different data view)

Key Insights - Time of Day:

- **Peak Hours**: 12 PM-3 PM (78,709 to 67,471 orders) - business hours peak
- **Active Period**: 9 AM-5 PM accounts for the vast majority of orders
- **Sharp Drop**: After 5 PM, order volume decreases dramatically
- **Minimal Activity**: 6-8 PM and early morning (6-8 AM) show very low activity
- **Pattern**: Clear business-hours preference, suggesting B2B or office-based purchasing behavior

```
Orders by Hour of Day:
 12    78709
15    77519
13    72259
14    67471
11    57674
16    54516
10    49037
9     34332
17    28509
8      8909
18     7974
19     3705
20      871
7       383
6        41
...
      10       60742
      11       84711
      12       25525
```

Fig 23: Order Distribution by Hour of Day

b) Average Order Processing Time: *Data calculation in progress*
Methodology: The code calculates processing time by:

1. Sorting data by CustomerID and InvoiceDate
2. Identifying previous invoice date for each customer
3. Computing time difference between consecutive orders
4. Converting to hours (dividing by 3,600 seconds)
5. Taking the mean of all processing times

Note: The specific average processing time value is being computed but requires completion of the calculation. Based on the dataset structure (December 2010 - December 2011), typical order processing times would likely range from hours to days depending on whether measuring:

- Time between order placement and fulfillment
- Time between repeat purchases by same customer
- Inter-order intervals

Limitation: The dataset may not contain explicit order fulfillment/shipment timestamps, so traditional "processing time" (order-to-delivery) cannot be directly calculated without additional data.

c) **Seasonal Analysis:**
The code extracts Month and Year from InvoiceDate to identify seasonal patterns.

Orders by Month:

Based on the time period (December 2010 - December 2011), the analysis reveals:

Key Seasonal Patterns:

1. **Year-End Peak**: December months likely show elevated activity due to holiday shopping
2. **Monthly Distribution**: Orders grouped by Year and Month provide insight into:
    - Peak shopping seasons (likely Q4: October-December)
    - Slower periods (potentially January-February post-holiday)
    - Mid-year trends (summer months)

Business Implications:

- **Thursday/Mid-Week Focus**: Staff customer service and fulfillment teams heavily for Wednesday-Thursday
- **Peak Hours (12 PM-3 PM)**: Ensure server capacity and support availability during lunch-early afternoon
- **Weekend Strategy**: Consider targeted campaigns to boost Sunday activity
- **Off-Peak Optimization**: Use early morning/evening hours for system maintenance
- **Seasonal Planning**: Prepare inventory and staffing for year-end peaks

Data Limitation Note: The complete monthly trend breakdown requires the full output from orders_by_month, which would show specific order volumes for each month across the dataset's timeframe, revealing precise seasonal patterns such as holiday spikes, summer lulls, or back-to-school trends.

5. Geographical Analysis

a) **Key Insights:**

- **UK Dominance**: United Kingdom accounts for approximately 88.7% of all orders, indicating this is primarily a UK-focused e-commerce business
- **European Market**: All top 5 countries are European, showing strong regional concentration
- **Significant Drop**: Germany (2nd place) has only 2.6% of UK's order volume, showing massive concentration
- **Geographic Focus**: The business operates predominantly in Western Europe with UK as the core market

```
Top 5 Countries with the Highest Number of Orders:
 United Kingdom    361878
Germany             9495
France              8491
EIRE                7485
Spain               2533
Name: Country, dtype: int64
```

Fig 24: Top 5 Countries by Order Volume

b) **Correlation Analysis**

Key Findings:

1. **Inverse Relationship**: Lower-volume countries (Netherlands, Australia, Lebanon) show higher average order values (£2,818, £1,987, £1,694), while high-volume countries like EIRE show moderate averages (£784.59)
2. **Geographic Patterns**:
    - **Distant markets** (Australia, Japan, Singapore) have higher average orders due to bulk purchasing to offset shipping costs
    - **European countries** show mixed values (£647-£893)
    - **UK** (implied lower average) exhibits individual consumer behavior with frequent small orders
3. **Weak Negative Correlation**: Order volume inversely relates to average order value
    - High-volume countries → lower averages (retail/consumer behavior)
    - Low-volume countries → higher averages (wholesale/export behavior)

Business Implications:

1. **Market Segmentation**:
    - **UK strategy**: Focus on conversion, retention, and purchase frequency
    - **International strategy**: Target premium/wholesale customers
2. **Customer Type Variation**:
    - UK: Individual consumers + small businesses
    - International: Wholesale buyers or affluent consumers optimizing shipping costs
3. **Revenue Insight**: While UK drives 88.7% of order volume, international markets contribute disproportionately high per-order revenue

Conclusion: A negative correlation exists between order volume and average order value, driven by bulk purchasing patterns, shipping economics, and customer segment differences across geographic markets.

```
Average Order Value by Country:
 Country
Netherlands            2818.431089
Australia              1986.627101
Lebanon                1693.880000
Japan                  1262.165000
Israel                 1165.708333
Brazil                 1143.600000
RSA                    1002.310000
Singapore               912.039000
Denmark                 893.720952
Norway                  879.086500
Sweden                  795.563261
Greece                  785.086667
Switzerland             785.061972
EIRE                    784.593166
Cyprus                  647.314500
...
USA                     247.274286
Czech Republic          141.544000
Saudi Arabia             65.585000
```

Fig 25: Average Order Value by Country

6. Payment Analysis

a) What are the most common payment methods used by customers?

Data Limitation:

The original dataset does not contain payment method information. This analysis cannot be performed with the available data.

b) Is there a relationship between the payment method and the order amount?

Data Limitation:

Without payment method data in the dataset, the relationship between payment method and order amount cannot be analyzed.

Note: The dataset only includes transactional information (InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country) without payment processing details. To enable this analysis in future, payment gateway data would need to be integrated into the transaction records.

7. Customer Behaviour

a) Average Customer Activity Duration: 133 days (approximately 4.4 months)
Calculation Method:

- First purchase date (min InvoiceDate) and last purchase date (max InvoiceDate) identified for each customer
- Duration calculated as the difference between last and first purchase
- Mean computed across all 4,372 customers

Interpretation:

Customers remain active for an average of 133 days from their first to last purchase. This represents the typical customer lifecycle span, helping identify when customers become at-risk and when re-engagement campaigns should be deployed.

b) Are there any customer segments based on their purchase behavior?

Yes, customers were segmented using two approaches:

1. RFM Score-Based Segmentation (Six Segments):

- Most Valued Customers (RFM ≥12)
- Loyal Customers (RFM 9-11)
- Potential Loyalists (RFM 7-8) - 77.3% of customers
- Recent Customers (RFM 5-6)
- New Customers (RFM 3-4)
- Lost Customers (RFM 1-2)

2. K-Means Clustering (Four Clusters):

- Cluster 0: 72.5% - Typical customers
- Cluster 1: 23.7% - Mid-tier engaged customers
- Cluster 2: 2.5% - Loyal customers
- Cluster 3: 1.3% - Premium VIP/wholesale customers

Both approaches enable targeted marketing strategies based on distinct purchasing behaviors and value contributions.

8. Returns and Refunds

a) **Data Limitation:**
The dataset lacks explicit returns/refunds columns. However, negative values in the Quantity column serve as return indicators.

b) **Data Limitation:**

The dataset does not contain a ProductCategory column, preventing category-level return analysis. Product descriptions would need to be classified first to enable this correlation analysis.

9. Profitability Analysis

a) **Data Limitation:**
The dataset lacks Cost of Goods Sold (COGS) or ProfitMargin data required for profit calculation.

Calculation Framework (if data available):

Profit = Total Sales Revenue - Total Cost of Goods
Where:

- Sales Revenue = UnitPrice × Quantity (available in dataset)
- COGS = SalesRevenue - (SalesRevenue × ProfitMargin) (NOT available)

Alternative Approach:

If ProfitMargin percentages were available by product, COGS could be derived and total profit calculated as:

Total Profit = Sum(SalesRevenue) - Sum(COGS)
Current Capability: Only total revenue can be calculated (~£8.3 million based on UnitPrice × Quantity), but profit cannot be determined without cost data.

b) **Data Limitation:**

Without COGS or ProfitMargin data, product-level profit margins cannot be calculated.

Calculation Framework (if data available):

Profit Margin = (SalesRevenue - COGS) / SalesRevenue
Then group by product and identify top 5 highest margins.

Conclusion: Profitability analysis requires cost data (COGS or profit margins) that is not present in the current dataset. Only revenue-based analysis is possible with available information.

10. Customer Satisfaction

a) **No.** The dataset does not contain a CustomerFeedback column or any customer review/rating data.

What Would Be Done (if available):

If customer feedback were available as text (reviews/comments), Natural Language Processing (NLP) techniques would be used:

- TextBlob library for sentiment analysis
- Sentiment polarity classification (Positive, Neutral, Negative)
- Sentiment distribution analysis

b) **No.** The dataset lacks a Rating column or any feedback data.

What Would Be Done (if available):

If numerical ratings (e.g., 5-star system) were available, the analysis would include:

- Average rating per product
- Rating distribution analysis
- Rating trends over time

Conclusion: Customer satisfaction analysis cannot be performed due to absence of feedback, rating, or review data. The dataset contains only transactional information without customer sentiment indicators.