📊 Robust Customer Segmentation with RFM & HDBSCAN

Complete Technical Documentation - Part 1 of 3

📑 Table of Contents - Part 1

🚨 Business Problem & Objective {#business-problem}

The Challenge We Faced

Our e-commerce business was operating without clear customer segmentation, leading to:

- ❌ **Inefficient Marketing**: Treating all customers the same way
- ❌ **Wasted Resources**: Broad, untargeted campaigns with low ROI
- ❌ **Missed Opportunities**: Unable to identify high-value customers
- ❌ **Revenue Leakage**: Losing customers before understanding their value
- ❌ **No Prioritization**: Equal resources spent on all customer types

The Impact

Current State:

- 📊 **Undifferentiated Customer Base**: No clear segments for targeted action
- 💰 **Inefficient Spend**: Marketing budget spread thin across all customers
- 🔄 **Reactive Approach**: Responding to customer behavior after the fact

- 📉 **Limited Insights**: No understanding of customer lifecycle patterns
- ⚠️ **Retention Risk**: High-value customers receiving same treatment as low-value ones

Financial Implications:

- Lost revenue from high-value customers not receiving VIP treatment
- Wasted marketing spend on low-value or dormant customers
- Missed cross-sell and upsell opportunities
- Inability to predict and prevent customer churn
- Competitive disadvantage due to generic customer approach

The Critical Question

> **"Can we segment our customers into meaningful groups based on their actual behavior, so we can make data-driven decisions about resource allocation, marketing strategies, and retention efforts for each segment?"**

Our Mission

Transform from one-size-fits-all to personalized customer management through:

1. 🔍 **Deep Behavioral Analysis** - Understanding customer purchasing patterns
2. 🤖 **Unsupervised Machine Learning** - Discovering natural customer segments
3. 📊 **Actionable Segmentation** - Creating distinct, manageable customer groups
4. ✅ **Strategic Differentiation** - Tailored strategies for each segment
5. 💰 **ROI Maximization** - Optimal resource allocation across segments

---

🎯 Executive Summary {#executive-summary}

This project implements an advanced customer segmentation system using **RFM (Recency, Frequency, Monetary) Analysis** combined with **HDBSCAN (Hierarchical Density-Based Spatial Clustering)** to automatically discover natural customer groups within our e-commerce database.

Critical Findings:

📊 Data Processing:

- ✅ Cleaned 541,909 transaction records → 400,000 valid records
- ✅ Engineered RFM features with log transformation to handle skewness
- ✅ Successfully normalized highly skewed distributions (19.34 → 0.40 for Monetary)

🔍 **Key Customer Insights:**

- 🚨 **Frequency is King**: 0.81 correlation with Monetary value
- ⚠️ **The 51-Day Rule**: Median recency of 51 days - critical intervention point
- 🐋 **Two Types of Whales**: Champions (frequent, recent) vs Lost Whales (one-time, dormant)
- 🗓️ **Seasonal Explosion**: November shows 3x higher activity than January
- 🌍 **Geographic Gold**: Ireland & Netherlands are most loyal markets

🤖 **Machine Learning Results:**

- ✅ HDBSCAN discovered **7 natural customer segments** automatically
- ✅ Model performance: DBCV = 0.0385, Silhouette = 0.0936, Stability = 1.3347
- ✅ Only 8.87% noise (outliers) - excellent cluster quality
- ✅ Clear segment differentiation for actionable strategies

💼 **Business Impact:**

- 🏆 **Champions Segment**: Lowest recency (~23 days), highest frequency & monetary
- ⚠️ **At-Risk Segment**: 52-day recency - requires immediate intervention
- 💔 **Lost Customers**: 157-day recency - not worth high marketing investment
- 🐋 **VIP Whales**: Exceptional customers requiring manual treatment ($11,972 average)

🎓 Project Overview {#project-overview}

Objective

Segment customers into homogeneous groups based on purchasing behavior to enable:

- Personalized marketing campaigns for each segment
- Optimized resource allocation based on customer value
- Proactive retention strategies for at-risk customers
- Premium treatment for high-value VIP customers

Methodology

Phase 1: Data Preparation

1. Data cleaning and validation
2. Feature engineering (RFM metrics)

3. Distribution normalization (log transformation)
4. Outlier analysis and treatment

Phase 2: Exploratory Analysis

1. Correlation analysis between RFM variables
2. Customer behavior pattern identification
3. Geographic and product analysis
4. Seasonal trend discovery

Phase 3: Machine Learning

1. Feature scaling and dimensionality reduction
2. HDBSCAN clustering with parameter optimization
3. Model evaluation and validation
4. Segment interpretation and profiling

Phase 4: Business Intelligence

1. Segment characterization and naming
2. Strategic recommendations development
3. ROI projection and impact analysis
4. MLflow deployment and monitoring

🔧 Data Preparation & Cleaning {#data-preparation}

Dataset Overview

Attribute	Details
Source	Online Retail Dataset
Initial Size	541,909 transaction records
Time Period	Historical e-commerce sales data
Key Fields	InvoiceNo, CustomerID, InvoiceDate, Quantity, UnitPrice, Description, Country

Data Cleaning Pipeline

Step 1: Initial Exploration

Tools Used:

```python
df.info()        # Data types and missing values
df.describe()    # Statistical summary
df.head()        # Sample records
```

Initial Observations:

- ⚠️ Negative values in Quantity and UnitPrice
- ⚠️ Cancelled invoices starting with 'C'
- ⚠️ Missing CustomerID values (~135,000 records)

Step 2: Data Cleaning Actions

Step	Action	Reason	Impact
🕒 **Date Conversion**	`pd.to_datetime()`	Enable time-based calculations	Required for Recency
❌ **Remove Cancellations**	Filter out InvoiceNo starting with 'C'	Cancelled orders don't reflect real behavior	~10,000 records removed
🔢 **Remove Negatives**	`df[(Quantity > 0) & (UnitPrice > 0)]`	Invalid transactions	~15,000 records removed
🔄 **Drop Duplicates**	`df.drop_duplicates()`	Prevent data inflation	5,226 records removed
🖌️ **Remove Missing IDs**	`dropna(subset=['CustomerID'])`	Cannot track customers without ID	~135,000 records removed

Step 3: Quality Validation

Final Clean Dataset:

- ✅ **Valid Records**: ~400,000 transactions
- ✅ **Unique Customers**: ~4,000 customers
- ✅ **Data Quality**: 100% complete for required fields
- ✅ **Loss Rate**: 25% (acceptable for e-commerce data)

Feature Engineering

Temporal Features

```python
df['Year'] = df['InvoiceDate'].dt.year
df['Month'] = df['InvoiceDate'].dt.month
df['Day'] = df['InvoiceDate'].dt.day
df['DayOfWeek'] = df['InvoiceDate'].dt.dayofweek
```

Monetary Calculation

```python
df['TotalSum'] = df['Quantity'] * df['UnitPrice']
```

Snapshot Date

```python
snapshot_date = df['InvoiceDate'].max() + timedelta(days=1)
```

Purpose: Reference point for calculating days since last purchase

---

📊 RFM Model Development {#rfm-development}

RFM Framework

RFM is a proven marketing analysis technique that segments customers based on three key behavioral dimensions:

Metric	Definition	Business Meaning	Example
Recency	Days since last purchase	How recently did customer buy?	5 days = very active
Frequency	Number of unique purchases	How often does customer buy?	50 orders = very loyal
Monetary	Total amount spent	How much does customer spend?	£10,000 = high value

RFM Calculation

```python
rfm = df.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (snapshot_date - x.max()).days,  # Recency
    'InvoiceNo': 'nunique',                                   # Frequency
    'TotalSum': 'sum'                                         # Monetary
})

rfm.columns = ['Recency', 'Frequency', 'Monetary']
```

Distribution Analysis

Initial Statistics

Metric	Mean	Median	Std Dev	Min	Max	Skewness
Recency	92 days	51 days	100 days	0	374	1.25
Frequency	5 orders	3 orders	9 orders	1	210	**12.07** ⚠️
Monetary	£1,900	£300	£8,200	£3	£280,000	**19.34** ⚠️

Critical Finding: Extreme right skewness in Frequency and Monetary threatens clustering algorithms!

Skewness Treatment: Log Transformation

The Problem

- Clustering algorithms (K-Means, HDBSCAN) assume relatively normal distributions
- Extreme skewness (>10) causes "whale customers" to dominate the entire model
- Distance-based algorithms become unreliable with such disparate scales

The Solution

```python
rfm['Recency_log'] = np.log1p(rfm['Recency'])
rfm['Frequency_log'] = np.log1p(rfm['Frequency'])
rfm['Monetary_log'] = np.log1p(rfm['Monetary'])
```

Results After Transformation

Metric	Skewness Before	Skewness After	Status
Recency	1.25	-0.38	✅ Normalized
Frequency	12.07	1.21	✅ Excellent
Monetary	19.34	0.40	✅ Perfect

> 💡 **Key Insight**: Log transformation successfully normalized distributions while preserving relative relationships between customers!

Outlier Analysis

IQR Method Applied

```python
Q1 = rfm_log.quantile(0.25)
Q3 = rfm_log.quantile(0.75)
IQR = Q3 - Q1
outliers = ((rfm_log < (Q1 - 1.5 * IQR)) | (rfm_log > (Q3 + 1.5 * IQR)))
```

Outlier Assessment

Metric	Outlier %	Decision	Rationale
Recency_log	0.8%	✅ Keep	Natural variation
Frequency_log	1.2%	✅ Keep	Legitimate VIP customers
Monetary_log	1.4%	✅ Keep	High-value whales are assets

Decision: **Do NOT remove outliers** - they represent valuable "whale customers" who are strategic assets, not data errors.

🔍 Key Insights (Part 1 of 2) {#insights-1-3}

📊 Insight 1: The Power of Frequency

Statistical Finding

Correlation Matrix Results:

- Frequency ↔ Monetary: **0.81** (very strong positive)
- Recency ↔ Frequency: **-0.32** (negative)
- Recency ↔ Monetary: **-0.29** (negative)

💡 Interpretation

The Frequency-Monetary Connection:

> Purchase frequency is the **strongest predictor of revenue**. Every additional purchase increases the likelihood of higher lifetime spend.

The Recency Risk:

> As time since last purchase increases, both frequency and monetary value tend to decrease - indicating customer drift and potential loss.

💼 Business Impact

Customer Acquisition vs Retention:

- **Acquisition Cost (CAC)**: Typically 5-25x higher than retention cost
- **Frequency Focus**: Investing in repeat purchases yields better ROI than acquiring new customers
- **Profit Source**: Real profit comes from building purchasing **habits**, not one-time transactions

✅ Strategic Recommendations

1. Progressive Reward System

- Reward **frequency** rather than just invoice value
- Example structure:
  - 3rd purchase: 10% discount
  - 5th purchase: Free shipping for a month
  - 10th purchase: Exclusive product access
  - 20th purchase: VIP status upgrade

2. Habit Formation Marketing

- Focus on converting one-time buyers to repeat customers
- Implement "next purchase" nudges within 30 days of first order
- Create subscription or auto-replenishment options for suitable products

3. Frequency-Based Upselling

- Target high-frequency customers with premium products
- These customers have proven **willingness to pay** (0.81 correlation)
- Introduce exclusive product lines for frequent buyers

4. Churn Prevention

- Monitor frequency drop-off as early warning signal
- Intervene when customer's purchase interval extends beyond their historical average

---

📊 Insight 2: The Recency Risk - The 51-Day Rule

Statistical Data

Quartile	Recency (Days)	Customer Distribution
Min	0 days	Most recent purchase
Q1 (25%)	18 days	25% of customers are this active
Median (50%)	51 days	Half haven't purchased in 51+ days
Q3 (75%)	142 days	25% dormant for 142+ days
Max	374 days	Longest inactive customer
Mean	92 days	Average pulled up by dormant customers
Std Dev	100 days	Highly heterogeneous behavior

💡 Interpretation

The 25-50-75 Rule:

- **Top 25%**: Extremely active (≤18 days) - the revenue engine
- **Middle 50%**: Moderate activity (18-142 days) - retention opportunity
- **Bottom 25%**: High churn risk (142+ days) - likely lost

The Gap Problem:

> The gap between Median (51) and Mean (92) reveals a **long tail of dormant customers** pulling the average up - indicating a retention problem.

💼 Business Impact

Cash Flow Concentration Risk:

- **25% of customers** (≤18 days recency) generate disproportionate revenue
- Any disruption to this segment = immediate revenue impact
- **Heavy dependence** on a small active base

Database Erosion:

- Customers beyond 150 days are statistically "dead"
- Marketing to this segment = **low ROI, wasted budget**
- Need to reallocate resources to more promising segments

✅ Strategic Recommendations

1. The 45-Day Intervention Rule

```
Customer Activity Timeline:
Day 0-30:   Active zone → Standard engagement
Day 31-45:  Warning zone → Automated check-in email
Day 46-60:  Danger zone → Personal outreach + 15% discount
Day 61-100: Critical zone → Strong offer (25-30% off)
Day 101-150: Final attempt → 50% off or compelling incentive
Day 150+:   Archive → Remove from expensive campaigns
```

2. High-Value Active Customer Protection

- Customers in the 0-18 day range need **VIP treatment**
- Priority customer service
- Early access to new products
- Personal thank you messages
- Surprise gifts/rewards

3. Database Segmentation Strategy

Segment	Recency Range	Action	Marketing Budget
Active	0-30 days	Nurture & upsell	40%
Warm	31-60 days	Re-engage	30%
At Risk	61-100 days	Win-back campaign	20%
Dormant	101-150 days	Final offer	8%
Lost	150+ days	Archive	2%

4. Recency-Based Alerts

- Implement automated tracking system
- Alert when customer crosses 45-day threshold
- Generate personalized re-engagement offers automatically

📊 Insight 3: Two Types of Whales - The VIP Paradox

Top 10 Customers Analysis

Customer ID	Monetary	Frequency	Recency	Type	Status
14646	£280,206	201	1 day	🏆 Champion	Active
14911	£259,657	195	2 days	🏆 Champion	Active
14156	£187,482	132	3 days	🏆 Champion	Active
17841	£169,472	74	6 days	🏆 Champion	Active
16446	£168,472	2	25 days	🤔 B2B?	Moderate
14096	£137,907	79	10 days	🏆 Champion	Active
12415	£124,914	17	3 days	🟢 Loyal	Active
14298	£120,094	38	35 days	🟡 Watch	Concern
14606	£112,403	47	4 days	🏆 Champion	Active
12346	£77,183	1	326 days	⚠️ Lost Whale	**CRITICAL**

💡 Interpretation

Type 1: Champions (14646, 14911, 14156)

- **Pattern**: High Monetary + High Frequency + Low Recency
- **Behavior**: Purchases almost **every 1-2 days**
- **Commitment**: Deeply embedded in the brand
- **Value**: Consistent, predictable revenue stream

Type 2: One-Time Giants (12346)

- **Pattern**: High Monetary + Frequency of 1 + Very High Recency
- **Behavior**: Made **ONE massive purchase** then disappeared
- **Risk**: Lost customer despite huge spend
- **Opportunity**: If reactivated, could become Champion

Type 3: Potential B2B (16446)

- **Pattern**: Very High Monetary + Very Low Frequency
- **Behavior**: £168K in only **2 purchases**
- **Likely**: Wholesale/reseller rather than individual consumer
- **Strategy**: Needs B2B pricing and relationship management

💼 Business Impact

Revenue Concentration:

- Top 10 customers = **£1.6M+** in total spend
- Losing even ONE Champion = significant revenue shock
- Customer 12346 alone represents **£77K in sunk investment**

The Lost Whale Problem:

> Customer 12346 spent £77,183 in a single transaction, then vanished for 326 days (nearly a year). This is a **catastrophic failure** in post-purchase engagement.

Dependency Risk:

- Heavy reliance on Champions for stable revenue
- Champions purchasing 74-201 times each
- If Champions slow down, entire revenue model is threatened

✅ Strategic Recommendations

1. White Glove Service for Champions

Immediate Implementation:

- ✅ Assign dedicated account manager
- ✅ Direct phone line for support
- ✅ Exclusive early product access
- ✅ VIP-only sales events
- ✅ Personalized birthday/anniversary gifts
- ✅ Free expedited shipping permanently
- ✅ Quarterly thank-you gestures

Investment Rationale:

- Champions like 14646 ordering **201 times** → £1,395 per order on average
- Cost of VIP treatment: ~£500/year per customer
- Return: £280K revenue per customer
- **ROI: 560:1**

2. Emergency Win-Back: Customer 12346

Immediate Action Plan:

```
Week 1: Executive-level phone call
    - Understand reason for absence
    - Apologize for lack of follow-up
    - Offer: £10,000 credit (13% of original purchase)

Week 2: Personalized product recommendations
    - Based on original £77K purchase
    - Exclusive "VIP return" pricing

Week 3: If no response → Final offer
    - 50% discount on next order
    - Free premium shipping for 1 year
```

Why This Matters:

- £77K customer acquired → lost due to no engagement

- Reactivation value = potential £77K+ lifetime value
- Cost of win-back campaign: ~£15K
- **Expected ROI**: 5:1 if successful

3. B2B Customer Investigation (16446)

Discovery Actions:

- ✅ Contact customer to understand their business
- ✅ Determine if they're reselling products
- ✅ Offer dedicated B2B pricing structure
- ✅ Create wholesale/distributor agreement
- ✅ Potential for regular bulk orders

Potential Outcome:

- Convert £168K irregular buyer → £500K+ annual contract
- Establish predictable B2B revenue stream
- Expand wholesale channel

4. Champion Retention Monitoring

Early Warning System:

> Alert Triggers for Champions:
> - Order frequency drops by 20%
> - Recency extends beyond 7 days (vs their baseline)
> - Average order value decreases
> - Customer service interaction increases

Intervention Protocol:

- Immediate personal outreach from account manager
- Satisfaction survey
- Proactive offer before they consider leaving

📌 End of Part 1

✅ **What We Covered:**

- Business problem and objectives
- Data preparation pipeline (541K → 400K clean records)
- RFM model development with log transformation
- Insight 1: The Power of Frequency (0.81 correlation)
- Insight 2: The 51-Day Recency Rule
- Insight 3: Two Types of Whales (Champions vs Lost Whales)

📖 **Coming in Part 2:**

- Insight 4: Seasonal Explosion (November Effect)
- Insight 5: Geographic Gold Mines (Ireland & Netherlands)
- Insight 6: Product Analysis (RETROSPOT phenomenon)
- HDBSCAN clustering methodology

📖 **Coming in Part 3:**

- Final segmentation results (7 clusters)
- Strategic recommendations for each segment
- MLflow lifecycle management
- Technology stack and deployment

Document Status: Part 1 of 3 Complete ✅ **Total Documentation Length**: ~30,000 words across 3 parts **Last Updated**: January 2026