Customer Segmentation & Retention Analysis

Analytical Techniques for Data Analysts

A comprehensive reference guide — from first principles to business action

Covering: RFM · CLV · Cohort Analysis · Segmentation · Churn

Based on UCI Online Retail II Dataset | Dec 2009 – Dec 2011

Table of Contents

RFM Analysis

Score every customer on Recency, Frequency & Monetary value

What is RFM?

RFM is a behavioural scoring framework that evaluates every customer across three dimensions using nothing more than their transaction history. It answers the question: *among all our customers, who deserves our attention right now?* It is fast to compute, easy to explain to business stakeholders, and actionable without any machine-learning infrastructure.

The Three Dimensions

R — Recency

How many days have passed since the customer's last purchase, measured from a fixed reference date (usually the last date in the dataset or 'today' in a live system). **Lower = better.** A customer who bought 5 days ago is far more likely to buy again than one who bought 400 days ago. In the Online Retail II dataset, the reference date was 1 Dec 2011 (one day after the last transaction).

F — Frequency

The total number of distinct invoices placed by the customer during their entire tenure in the dataset. **Higher = better.** Note: this is a raw count, not a rate. A customer with 40 orders over 24 months and one with 40 orders in 3 months get the same frequency score — a known limitation that CLV and Cohort analysis address.

M — Monetary

The total revenue generated by the customer: sum of (Quantity x Price) across all their invoices. **Higher = better.** In a B2B wholesale context like our dataset, a single gift-shop buyer placing bulk orders can easily reach £10,000+, dwarfing casual retail buyers.

How RFM Scores Are Calculated

Each dimension is divided into quintiles (5 equal buckets) and assigned a score from 1 (worst) to 5 (best). For Recency, a smaller number of days = score 5. For Frequency and Monetary, a larger value = score 5. The three scores are then combined — either summed (RFM Total Score) or concatenated (e.g. '554') — to produce a customer profile.

RFM Score = R_score (1–5) + F_score (1–5) + M_score (1–5) → Total: 3 to 15

RFM Segment Labels

Once scored, customers are mapped to named segments using business rules. Below is the standard segment taxonomy used in the notebook:

Segment	RFM Profile	Recommended Action
Champions	5-5-5 / very high all round	Reward, upsell, ask for reviews
Loyal Customers	High F & M, good R	Loyalty programme, early access
Potential Loyalists	Recent, moderate F	Nurture with targeted offers
At Risk	Was high, R now poor	Win-back campaign urgently
Hibernating	Low on all three	Low-cost re-engagement or ignore
New Customers	High R, low F	Onboarding, second-purchase nudge
Lost	Very low R, low F & M	Write off or deep discount

Limitations of RFM

- Frequency is a raw count, not a rate — it does not capture purchase rhythm.
- Recency can unfairly penalise seasonal customers (e.g. skipping December).
- RFM is backward-looking; it scores past behaviour, not future potential.
- These limitations are the reason CLV and Cohort Analysis exist as complements.

CLV — Customer Lifetime Value

How much is this customer worth to the business over their entire relationship?

What is CLV?

Customer Lifetime Value answers a forward-looking question: *if this customer continues their current behaviour, how much revenue will they generate before they stop buying?* Where RFM scores the past, CLV quantifies the future. It is the single most important number for deciding how much to spend on acquiring, retaining, or winning back a customer.

The CLV Formula

CLV = Average Order Value × Purchase Frequency (per year) × Customer Lifespan (years)

Breaking Down Each Component

Average Order Value (AOV)

Total revenue from the customer divided by their total number of invoices. In the Online Retail II dataset, a wholesale gift-shop buyer might have an AOV of £300, while a casual one-time buyer has an AOV of £25.

Purchase Frequency

Unlike the raw count used in RFM, CLV uses frequency as a **rate** — how many times per year does this customer buy? A customer who placed 24 orders over 2 years has a frequency of 12/year. This is the corrected view of frequency that you rightly identified as more meaningful.

Customer Lifespan

The expected duration of the customer relationship in years. This can be estimated simply (average tenure across all customers), or probabilistically using the **BG/NBD model** (Buy-Till-You-Die), which uses each customer's own purchase gap to predict whether they are still 'alive'. In the notebook, a 90-day inactivity threshold was used as a practical churn proxy to estimate lifespan.

Worked Example from the Dataset

Lifespan= last purchase- first purchase
churn proxy was set to 90 days

Customer Type	AOV	Frequency	Lifespan	CLV
Customer A (Wholesale)	£300	12/year	3 years	**£10,800**
Customer B (Casual)	£25	1/year	0.5 years	**£12.50**
Customer C (At-Risk)	£150	6/year	0.2 years	**£180**

Customer A has a modest AOV but is a consistent repeat buyer — their CLV dwarfs Customer B who spent more per visit but never returned. This is the core insight CLV delivers that neither raw spend nor RFM alone can reveal.

CLV Tiers and Business Action

Tier	Definition	Action
High CLV	Top 20% by CLV score	Assign account managers, premium support, loyalty perks
Mid CLV	Middle 60%	Targeted upsell campaigns, frequency nudges
Low CLV	Bottom 20%	Low-cost automated marketing, assess acquisition cost

RFM vs CLV — The Key Difference

RFM tells you who was valuable. CLV tells you who will be valuable.

A new customer with only 3 invoices but very high AOV and short purchase gaps may have a higher CLV than a 2-year veteran whose order frequency is declining. Always use both together for a complete picture.

Cohort Analysis

Track how groups of customers behave over time from their acquisition month

What is Cohort Analysis?

A cohort is a group of customers who share a common starting point — usually the month they made their very first purchase. Cohort analysis tracks what percentage of each cohort returned to buy again in subsequent months. It is the most direct way to measure **retention** — how well you are keeping customers over time.

Why It Matters

RFM and CLV look at the customer in isolation. Cohort analysis looks at *groups across time*. This lets you answer questions like: Are customers acquired in Q4 (holiday season) better retained than those acquired in Q1? Did a product change or marketing campaign in mid-2011 improve retention? Is the business getting better or worse at keeping new customers?

How to Build a Cohort Analysis

Step 1 — Assign cohort month

Find each customer's first-ever purchase date. Group them by year-month. E.g. all customers who first bought in Jan 2010 form the 'Jan-2010 cohort'.

Step 2 — Calculate cohort index

For each subsequent purchase, calculate how many months after their first purchase it occurred. Month 0 = their first purchase month, Month 1 = one month later, and so on.

Step 3 — Build the retention matrix

Create a pivot table: rows = cohort month, columns = months since first purchase (0, 1, 2, ...), values = number of customers still active. Divide each value by the Month 0 count to get a retention percentage.

Step 4 — Visualise as a heatmap

A heatmap with colour intensity representing retention % makes patterns immediately visible. Dark cells = high retention, light = drop-off.

Reading the Cohort Matrix — Example

Cohort	Month 0	Month 1	Month 2	Month 3	Month 6
Jan-2010	100%	42%	35%	30%	28%
Apr-2010	100%	38%	29%	22%	20%
Jul-2010	100%	45%	40%	36%	33%
Dec-2010	100%	28%	18%	12%	—

The Jul-2010 cohort shows the strongest retention — 33% still active at Month 6. The Dec-2010 cohort drops sharply after Month 1, likely holiday one-time buyers. These insights directly inform when and how to invest in retention campaigns.

Key Metrics from Cohort Analysis

• **Month-1 Retention Rate** — the single most important early signal. If less than 20% of new customers return after their first purchase, something is broken in the onboarding or product experience.

• **Retention Curve Shape** — does it flatten out (healthy loyal base) or keep dropping to zero (transactional, no loyalty)?

• **Cohort Comparison** — are newer cohorts retaining better than older ones? This tells you if your retention efforts are working over time.

Cohort Analysis is the only technique that shows you retention trends over time — something RFM and CLV alone cannot do.

Customer Segmentation

Group customers into meaningful clusters to enable targeted business action

What is Customer Segmentation?

Segmentation is the process of dividing your entire customer base into distinct groups where customers within a group are similar to each other and different from those in other groups. It is the final step that makes all the preceding analysis actionable — instead of treating 5,000 customers identically, you create 5–8 meaningful groups and design a specific strategy for each.

Two Approaches Used in the Notebook

Approach 1 — Rule-Based RFM Segmentation

Business rules are manually defined to map RFM score combinations to segment labels. For example: if R_score >= 4 AND F_score >= 4 AND M_score >= 4 → 'Champion'. This approach is fully transparent, easy to explain to non-technical stakeholders, and directly tied to business intuition.

Approach 2 — KMeans Clustering (Machine Learning)

KMeans is an unsupervised ML algorithm that finds natural groupings in the data without any predefined rules. The process in the notebook was:

- **Scale the features** — RFM values have very different ranges (days vs invoice counts vs £). StandardScaler normalises them so no single dimension dominates.
- **Choose K** — the optimal number of clusters was selected using the Elbow Method (plot inertia vs K, pick the 'elbow' where gains diminish).
- **Fit and label** — KMeans assigns each customer to a cluster. Each cluster is then interpreted by examining the average RFM values within it.
- **Name the clusters** — e.g. Cluster 2 has low recency, high frequency, high monetary → 'Lapsed High-Value'. This naming step requires human judgement.

Rule-Based vs KMeans — When to Use Which

Dimension	Rule-Based	KMeans
Transparency	Full — rules are explicit	Low — algorithm decides
Business buy-in	Easy to present	Needs interpretation layer

Finds surprises?	No — only what you defined	Yes — reveals hidden patterns
Flexibility	Must manually update rules	Re-runs adapt to new data
Best for	Operational marketing actions	Exploratory discovery

Example Segments from the Dataset

Segment	RFM Profile	Action
Bulk Wholesale Buyers	Low R, low F, very high M	Large infrequent orders — quarterly outreach
Consistent Mid-Buyers	Good R, high F, mid M	Core base — loyalty rewards
Lapsed High-Value	Poor R, high F, high M	Top win-back priority
One-Time Holiday Buyers	High R (Dec), F=1, low M	Convert to second purchase
International High AOV	Moderate R & F, high M	Personalised account management

Segmentation is not the analysis — it is the output that makes all other analysis actionable. Every segment must have an owner and a strategy.

Churn Analysis

Identify customers who have stopped buying — and predict who is about to

What is Churn?

Churn is the loss of a customer — they have stopped purchasing and are unlikely to return without intervention. In subscription businesses, churn is explicit (a cancellation event). In transactional retail like our dataset, churn is **implicit** — we must define it using an inactivity threshold.

Defining Churn in the Online Retail Dataset

The notebook used a **90-day inactivity threshold**: any customer who has not placed an order in the last 90 days (relative to the reference date) is classified as churned. This threshold was also tested at 60 and 120 days to understand sensitivity — a technique called **threshold sensitivity analysis**.

Churn Definition: Last Purchase Date < Reference Date – 90 days → Churned

Why the Threshold Matters

Threshold	Effect	Risk
60 days	More customers flagged as churned	Aggressive — may waste retention budget on still-active customers
90 days	Balanced threshold (used in notebook)	Good default for most retail businesses
120 days	Fewer customers flagged	Conservative — may miss early churn signals

The right threshold depends on your industry's natural purchase cycle. A grocery retailer might use 14 days. A furniture retailer might use 365 days. In our wholesale gift context, 90 days was a reasonable middle ground.

Types of Churn Analysis

Descriptive Churn

Simply count: how many customers churned this month/quarter? What % of the total base? Track this over time. In the notebook, churn rate was reported alongside cohort retention to give a complete retention picture.

Churn by Segment

Break churn down by RFM segment, country, or acquisition channel. In the dataset, international customers had lower churn rates than UK customers — a signal that wholesale buyers are stickier than domestic retail buyers.

Churn Prediction (Advanced)

Use ML models (Logistic Regression, Random Forest, XGBoost) trained on RFM features, purchase gap history, and behavioural signals to predict which active customers are *likely to churn in the next 30/60/90 days*. This is the most actionable form — it lets you intervene before it happens. The notebook used the 90-day rule as a simpler proxy for this.

Churn vs RSS — The Connection

The 'Stop' category in RSS is essentially your churned segment. Churn Analysis goes deeper — it quantifies the rate, identifies the drivers, and (in predictive mode) scores customers by churn probability so you can prioritise retention spend on high-CLV customers who are showing early churn signals.

Recommended Actions by Churn Risk

Status	Definition	Action
Active	Bought within threshold	Reward, upsell, maintain engagement
At-Risk	Approaching threshold (60–89 days)	Targeted win-back offer NOW
Churned (Recent)	Just crossed 90-day mark	Strong incentive — discount, personalised outreach
Churned (Long)	180+ days inactive	Low-cost campaign or write off

The goal of churn analysis is not to count lost customers — it is to identify at-risk customers before they leave.

The Complete Framework — At a Glance

RFM	Score each customer: Recency, Frequency, Monetary	*Who is valuable right now?*
CLV	AOV × Frequency Rate × Lifespan	*Who will be valuable in the future?*
Cohort	Track retention % by acquisition month	*Are we keeping customers over time?*
Segmentation	Rule-based labels + KMeans clusters	*How do we act on each group?*
Churn	90-day inactivity threshold + predictive models	*Who is about to leave?*

The Analysis Pipeline

User Behavior Analysis understands HOW customers shop → RFM & CLV score and VALUE them → Cohort tracks retention OVER TIME → Segmentation GROUPS them → Churn identifies who is LEAVING → Business acts on each group with a targeted strategy.