Segmentation Project - RFM Analysis

Author: Klaudia Rapa
Date: 01.12.2025

Summary

The Vintage Haven has grown from a small London boutique into an international online retailer, resulting in a much larger and more diverse customer base. To use its marketing budget efficiently and maximise revenue, the business needs a clearer understanding of how different customers behave and which groups contribute the most value. This project analyses one year of transactional data from the UCI Online Retail dataset using RFM segmentation to assess customer recency, frequency, and monetary value. The analysis identifies the highest-value customers, those at risk of churn, and the segments most worth targeting in future campaigns. These insights provide a data-driven foundation for focusing marketing spend where it will deliver the greatest impact.

Problem

The Vintage Haven is a London boutique specialising in quirky home décor, handmade accessories, and seasonal gifts. What started as a small neighbourhood shop has expanded into a global online retailer, attracting thousands of customers across different countries and purchasing behaviours. With this growth, the owners need a clearer understanding of how their customer base has changed and which types of buyers drive the most revenue.

The business now serves a wide mix of customers - from one-off individual shoppers to **small B2B buyers and large wholesalers**. These groups behave very differently, making it difficult to see who brings long-term value, who is beginning to disengage, and where marketing investment should be focused to achieve the best return. Treating all customers the same risks inefficient spending and missed opportunities to retain or grow profitable segments.

This project uses transactional data to help The Vintage Haven:

- identify its **most and least valuable** customer groups,

- detect **behavioural patterns and early signs of churn**,
- understand differences between **retail, small B2B, and wholesale buyers**,
- and determine **which segments should be prioritised in the next marketing campaign**.

By generating these insights, the shop can allocate its marketing budget more strategically and focus future campaigns on the segments that will maximise revenue.

Inputs

This project is based on a single transactional dataset from the UCI Machine Learning Repository. The data contains all online retail transactions for a UK-based giftware company over a one-year period, providing a rich view of customer behaviour, purchasing patterns, and revenue distribution.

Data Source

Chen, D. (2015). *Online Retail* [Dataset]. UCI Machine Learning Repository.

https://doi.org/10.24432/C5BW33

The dataset covers all transactions between **1 December 2010** and **9 December 2011**, capturing both domestic and international customers. It includes product-level and customer-level information for an online retailer selling unique, all-occasion gifts — a natural fit for analysing The Vintage Haven's business model.

Dataset Overview

- **Rows:** 541,909 transactional records
- **Columns:** 8
- **Time period:** 1 year
- **Unique customers:** varies after cleaning (many missing IDs)
- **Unique products:** ~3,600 after cleaning and deduplication

Key Fields

Column	Description
InvoiceNo	Unique identifier for each transaction (including cancellations).

Column	Description
StockCode	Product SKU.
Description	Product name/description.
Quantity	Number of units purchased (negative for returns).
InvoiceDate	Timestamp of purchase.
UnitPrice	Price per item in GBP.
CustomerID	Unique customer identifier (missing for ~25% of records).
Country	Customer’s country at time of purchase.

Why this dataset works for the project

This dataset enables:

- customer-level analysis (RFM scoring, segmentation),
- product-level exploration (what different groups buy),
- behavioural investigation (recency, frequency, spend patterns),
- international comparison (UK vs abroad), and
- identification of high-value or at-risk customers.

It provides everything needed to build a realistic segmentation and marketing prioritisation strategy for The Vintage Haven.

Discovery

Data Cleaning & Exploration

The dataset contains several data quality issues:

- **Missing Customer IDs** (~25% of rows)
- **Negative quantities** (representing returns or refunds)
- **Duplicate rows**
- **Inconsistent product descriptions**
- **Outliers** (e.g., extremely large quantities and unit prices)

To ensure meaningful analysis, we performed the following steps:

- Dropped rows with **missing Customer IDs**, **cancelled invoices**, and **duplicates**, as these would not contribute value or could skew results. *(Potentially, missing Customer IDs could be imputed by joining with other data sources using Invoice No.)*
- Cleaned **product descriptions** by identifying similar names and consolidating them.
- **Outliers** were retained because there was insufficient information to determine if they were errors.

Additionally, we created a few columns:

- **Customer Type** (based on quantity purchased)
- **Total Price** (Quantity × Unit Price)
- **Description_Clean**
- **Date** and **Time** components
- **Recency, Frequency, and Monetary scores** to support customer segmentation.

Customer Segmentation & RFM Preparation

Exploration of **Quantity** and **Unit Price** have highly skewed distributions. Based on data characteristics and discussions with stakeholders, we divided customers into three groups to improve the reliability of the RFM analysis. Without segmentation, the RFM scoring would heavily favour wholesalers due to their larger transaction volumes. Tailoring marketing strategies per group ensures more meaningful insights.

The customer groups are:

1. **Individual customers:** average purchase fewer than 20 units
2. **Small B2B businesses:** average purchase between 20 and 100 units
3. **Wholesalers:** average purchase more than 100 units

Geographic Insights

The majority of customers are based in the **UK**, followed by **Germany, France, and Ireland**. Marketing and engagement strategies can be tailored to these key markets.

November is the busiest month across all customer groups, likely due to the Christmas season, while January and February are the slowest, suggesting the company should focus on marketing and stock in September–October to prepare for the year-end peak.

Product Insights

At the start of the analysis, the store had 3,877 products listed. But some of these were really the same items with slightly different names — things like the words being in a different order. After cleaning that up, the number of actual unique products dropped to 3,608.

When we looked closer at sales, we found that a lot of these products didn't sell very much. Some were sold only once all year, and about 1,000 of them sold fewer than 100 units. That's a big chunk of inventory that isn't bringing in much revenue.

Unless these slow sellers are special one-off vintage finds that naturally sell quickly and only once, it might make sense to offer fewer products overall. Keeping unpopular items around just takes up storage space and costs money. A good move would be to discount these low-demand products and highlight them in a marketing campaign to help clear them out and make room for items people actually want.

A good next step would be a deeper product analysis. We could look at how many units of each item are currently in storage, when each product was added to the catalog, and how often it actually sells. This would give a clearer picture of which products are worth keeping and which ones might be tying up space for no reason. However, it is out of scope of the current analysis.

Execution

Jupyter Notebook

RFM Analysis Introduction

To better understand customer behaviour and guide marketing strategies, we applied **RFM (Recency, Frequency, Monetary) analysis**, which evaluates:

- **Recency (R):** How recently a customer made a purchase. More recent purchases are more valuable.
- **Frequency (F):** How often a customer purchases. Frequent buyers are more engaged.
- **Monetary (M):** How much a customer spends. High spenders contribute more to revenue.

Customers who purchase recently, often, and spend more tend to be the most valuable — and the most likely to return.

Those who haven't bought in a long time or only purchased once have a higher chance of drifting away.

By scoring each customer on these three behaviours, RFM creates clear segments such as **Champions**, **Loyal**, **Potential Loyalists**, and **At Risk** groups.

This makes it far easier to decide **where the marketing budget should go**, which customers deserve more attention, and which groups have the strongest potential for future revenue.

Since the dataset contains very different customer types- ranging from individual consumers to wholesalers-we calculate RFM scores **within each customer type**. This ensures fair comparisons and prevents large-volume buyers (wholesalers) from dominating the scoring.

Step 1: Dividing Customers into Groups

We grouped customers based on their average order quantity:

```
def classify_customer(avg_qty):
    if avg_qty < 20:
        return "Individual"
    elif avg_qty < 100:
        return "Small B2B"
    else:
        return "Wholesaler"

customer_quantity['Customer Type'] = customer_quantity['mean'].apply(class
ify_customer)
```

Why this matters:

Dividing customers into **Individuals, Small B2B, and Wholesalers** ensures that each group is scored relative to its peers. Without this segmentation, wholesalers with large order volumes would skew RFM metrics, making it impossible to identify high-value individuals or small businesses. Tailoring RFM analysis per group makes marketing insights actionable for each customer segment.

Step 2: Calculating RFM Values

We aggregate the transactional data to calculate **Recency, Frequency, and Monetary values** for each customer:

```
rfm = clean_data.groupby('CustomerID').agg({
    'InvoiceDate': lambda x: (ref_date - x.max()).days,  # Recency
```

```
    'InvoiceNo': 'nunique',                    # Frequency
    'TotalPrice': 'sum'                       # Monetary
}).reset_index()

rfm.rename(columns={
    'InvoiceDate': 'Recency',
    'InvoiceNo': 'Frequency',
    'TotalPrice': 'Monetary'
}, inplace=True)
```

- **Recency:** days since the last purchase (smaller = better)
- **Frequency:** number of unique invoices
- **Monetary:** total spending per customer

Step 3: Assigning RFM Scores

RFM values are converted into **quintiles** within each customer type:

```
def score_quintile_rank(x, ascending=True):
    pct = x.rank(method="average", pct=True)         # percentile rank (0–1)
    if ascending:                                    # invert for Recency
        pct = 1 - pct
    return (pct * 5).apply(lambda v: min(max(int(np.ceil(v)), 1), 5))
```

```
rfm['R_score'] = rfm.groupby('Customer Type')['Recency'].transform(lambda
x: score_quintile_rank(x, ascending=False))
rfm['F_score'] = rfm.groupby('Customer Type')['Frequency'].transform(lambd
a x: score_quintile_rank(x, ascending=True))
rfm['M_score'] = rfm.groupby('Customer Type')['Monetary'].transform(lambd
a x: score_quintile_rank(x, ascending=True))
```

- Recency is inverted so that more recent buyers get higher scores.

- Frequency and Monetary are scored normally (higher = better).
- Calculating scores **per customer type** ensures fairness and comparability across diverse purchase behaviours.

Step 4: Customer Segmentation

We assign each customer to a **segment** based on their RFM scores:

```
def segment_customer(df):
    R, F, M = df['R_score'], df['F_score'], df['M_score']

    if R >= 4 and F >= 4 and M >= 4:
        return 'Champions'
    elif R >= 4 and F >= 3:
        return 'Loyal'
    elif R >= 3 and F >= 3:
        return 'Potential Loyalist'
    elif R >= 4 and F <= 2:
        return 'Recent Customers'
    elif R <= 2 and (F >= 4 or M >= 4):
        return 'At Risk but Valuable'
    elif R <= 2 and F <= 2:
        return 'At Risk / Lost'
    else:
        return 'Others'

rfm['Segment'] = rfm.apply(segment_customer, axis=1)
```

- **Champions:** High in all three dimensions; most valuable.
- **Loyal:** Frequently buying and recently active.
- **Potential Loyalist:** Fairly engaged, but not top-tier.
- **Recent Customers:** Bought recently but not frequently.
- **At Risk but Valuable:** Low recency but previously high spenders.

- **At Risk / Lost:** Low recency and low engagement.
- **Others:** Remaining customers not fitting other categories.

Segmenting this way allows targeted marketing campaigns: for instance, “Champions” can be rewarded, “At Risk” can be re-engaged, and “Recent Customers” can be nurtured into loyalty.

Step 5: Integration with Transactional Data

Finally, RFM scores and segments are merged back into the full dataset for further analysis and visualisation:

```
clean_data_rfm = clean_data.merge(
    rfm[['CustomerID', 'Recency', 'Frequency', 'Monetary', 'R_score', 'F_score',
'M_score', 'Segment']],
    on='CustomerID',
    how='left'
)
```

This allows exploration by **customer type, segment, country, and product**, and supports building interactive dashboards (e.g., in Power BI).

RFM Integration and Export

After calculating the **Recency, Frequency, and Monetary (RFM) scores** and assigning each customer to a segment, the results were merged back into the full transactional dataset. This allows each transaction to carry **customer behaviour insights**, enabling analysis by **customer type, segment, country, product, and time period**.

The merged dataset contains **392,732 rows** and includes both the original transaction columns and the newly created features:

- **Customer Type** (Individual, Small B2B, Wholesaler)
- **RFM values** (Recency, Frequency, Monetary)
- **RFM scores** (R_score, F_score, M_score, 1–5 scale)
- **Segment** (Champions, Loyal, Potential Loyalist, etc.)

By combining transactional and RFM data, we can **visualise trends, monitor segment performance, and tailor marketing strategies** for each customer group.

Finally, the cleaned datasets were exported for further exploration in **Power BI**:

```
rfm.to_csv("RFM_Segments.csv", index=False)
clean_data_rfm.to_csv("Clean_Transactions.csv", index=False)
```

This ensures that both the **customer-level RFM insights** and the **full transactional context** are available for reporting and downstream analysis.

Dashboard

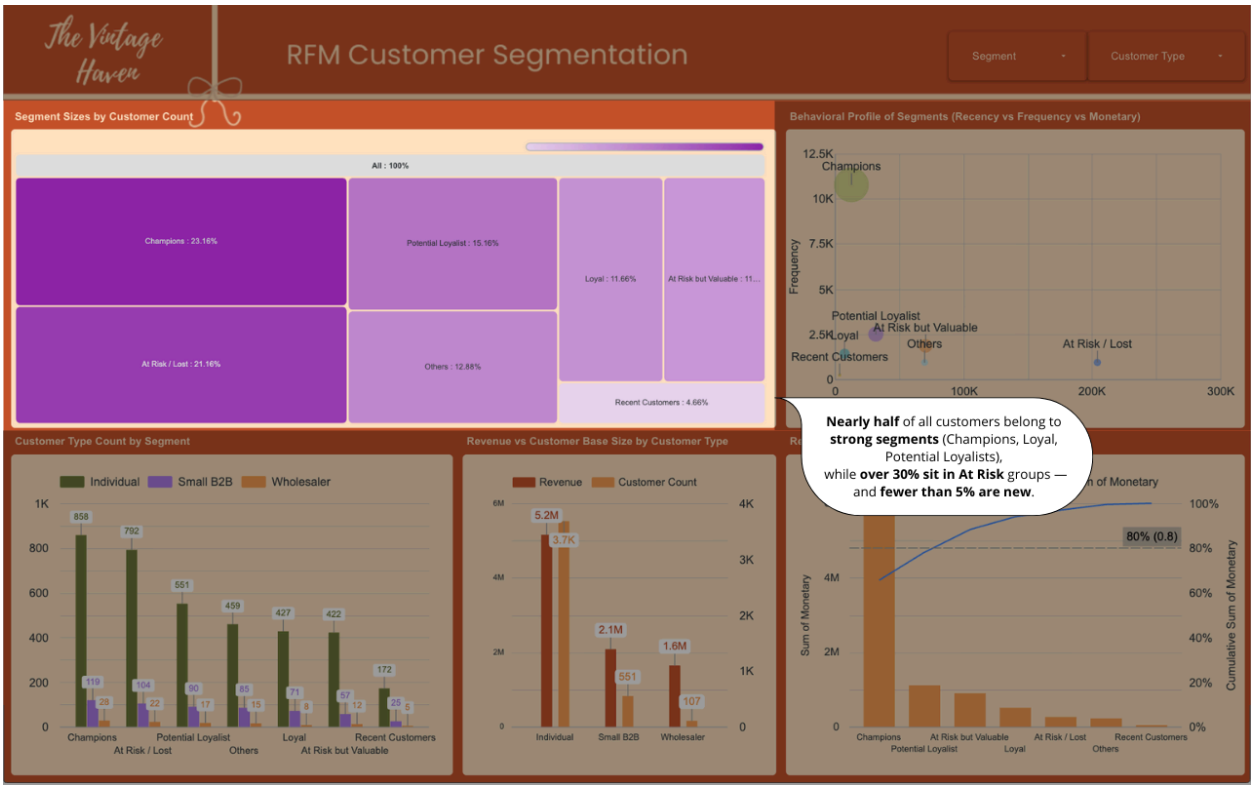
Link to the dashboard: https://lookerstudio.google.com/reporting/92f2c237-6ef2-44f3-849e-93c0e4171fea

This dashboard was designed to answer a central business question:

Which customers should we focus on next?

Using RFM insights, it identifies the segments that drive the most value and those with the strongest potential for future growth. The goal is to guide the marketing team toward the groups most worth investing in, ensuring the campaign budget is used where it will have the greatest impact.

Chart I: Segment Sizes by Customer Count



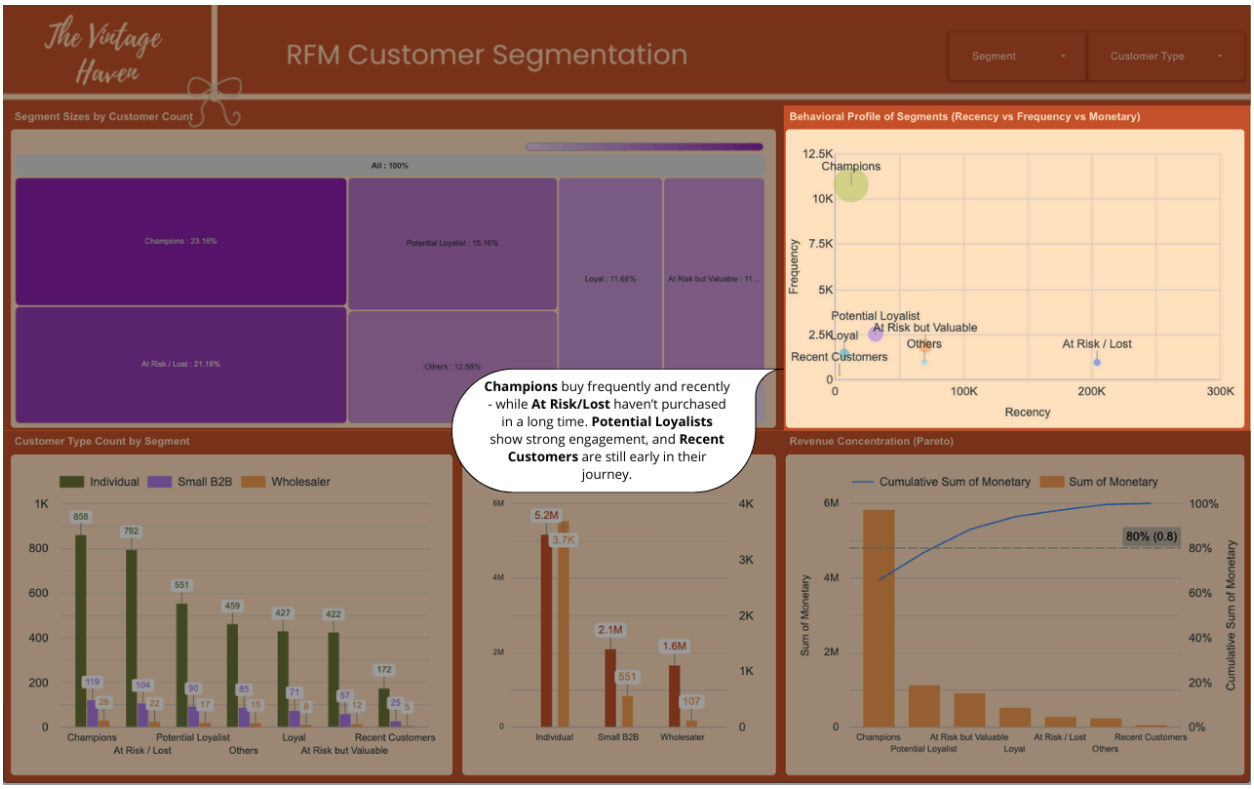
This chart shows how the customer base is distributed across RFM segments. Three groups dominate: **Champions**, **Potential Loyalists**, and **At Risk / Lost**, each representing a significant share of the total customers.

What makes this visual important is the imbalance it highlights:

- **Champions + Potential Loyalists + Loyal** together make up **nearly 40% of the customer base**. These are the customers who buy often, spend more, and return consistently.
- **At Risk / Lost + At Risk but Valuable** together exceed **30%**, revealing a large pool of customers who were previously active but have stopped purchasing.
- **Recent Customers account for less than 5%**, suggesting limited new customer growth compared to the size of the disengaged segments.

This distribution suggests a mixed health profile: **the business has a solid core of valuable customers but is also losing a substantial number who once had strong potential**. The imbalance between “strong” and “declining” groups will influence which segment deserves priority in the next campaign.

Chart II: Behavioural Profile of Segments (Recency vs Frequency vs Monetary)



The **Behavioural Profile** chart maps each RFM segment by two crucial behaviours:

- **Recency** → how long since last purchase
- **Frequency** → how often they buy
- **Bubble size** → monetary value

This visual makes the behavioural differences between segments immediately clear.

1. Champions: high-frequency, high-value and active

- Highest purchase frequency (~11–12K)
- Low recency → they buy often and recently
- Largest bubble size = highest monetary value

They are large in number, profitable and engaged. **This segment should be protected and rewarded to maintain their purchasing momentum.**

2. Potential Loyalists: engaged but not consistent yet

Potential Loyalists show:

- Mid-high frequency (~3K)
- Moderate recency (~40–50K)
- Strong monetary potential

They are close to becoming Champions but inconsistencies in purchase timing hold them back. **A targeted win-back or loyalty programme could convert them into top-tier customers.**

3. Loyal: steady but less valuable

Loyal customers appear with:

- **Steady but low frequency**
- **Low recency (active recently)**

Even though they shop recently and reliably, their monetary bubble is smaller. **Upsell/cross-sell campaigns could increase their value without major effort.**

4. At Risk but Valuable: valuable but slipping away

This segment shows **worrying behaviour**:

- Decent past frequency
- Moderate recency (~80–100K)
- Mid-size bubble = historically strong monetary value

They used to buy well — but now they're drifting. **This group offers the highest revenue upside for a targeted reactivation campaign.**

5. At Risk / Lost: very high recency, almost no repeat behaviour

This group is positioned far right (> **200K recency**), meaning they haven't purchased in **a long time**.

Frequency is extremely low, and monetary value weak. **Low priority for personalised campaigns - focus only on scalable, mass email flows.**

6. Others & Recent Customers: low frequency and very low value

These segments cluster at the bottom:

- Low frequency
- Low monetary value
- Recent Customers have low recency but no repeat pattern yet

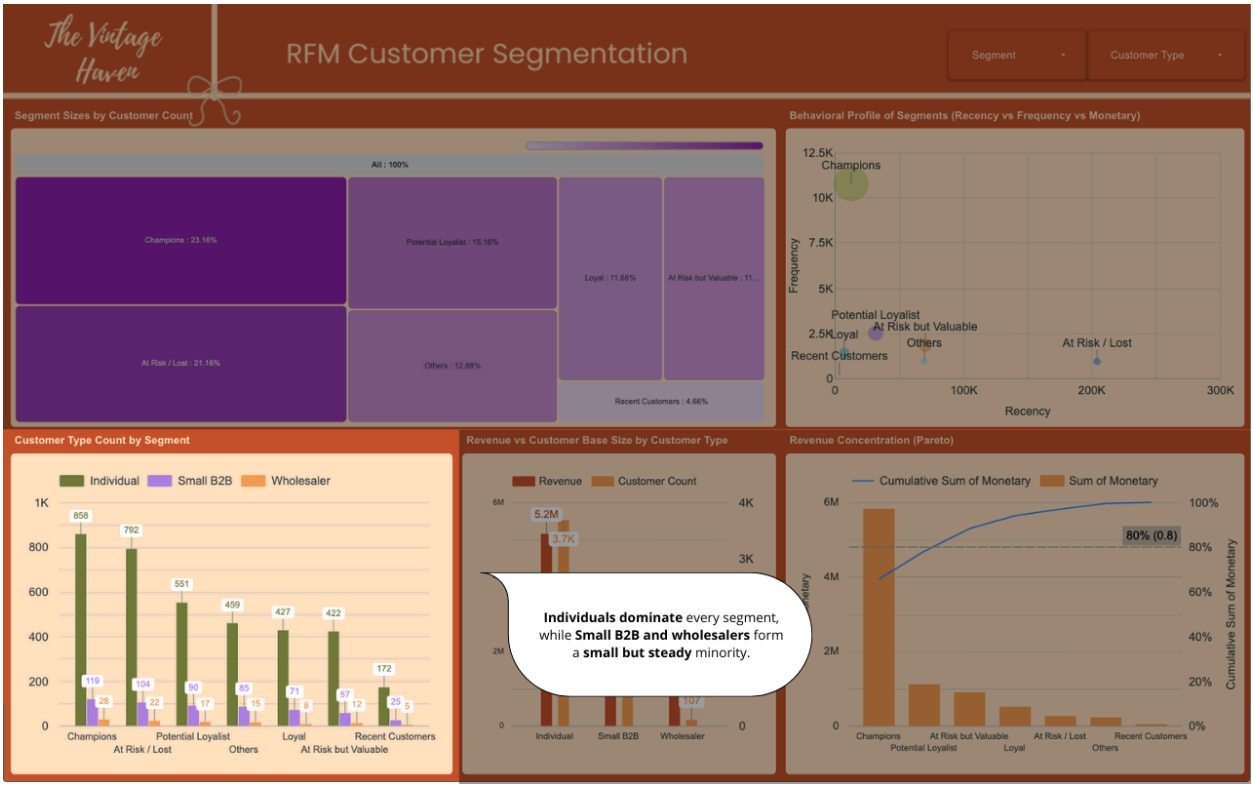
These groups require nurturing, not investment.

Based on the visual, the strategy should be following:

- Champions → maintain
- Potential Loyalists → grow
- At Risk but Valuable → rescue

The rest should be nurtured but not prioritised in revenue-driving campaigns.

Chart III: Customer Type Count by Segment



This visual shows **how many customers of each type** (Individual, Small B2B, Wholesaler) belong to each RFM segment.

From the bars, we can see:

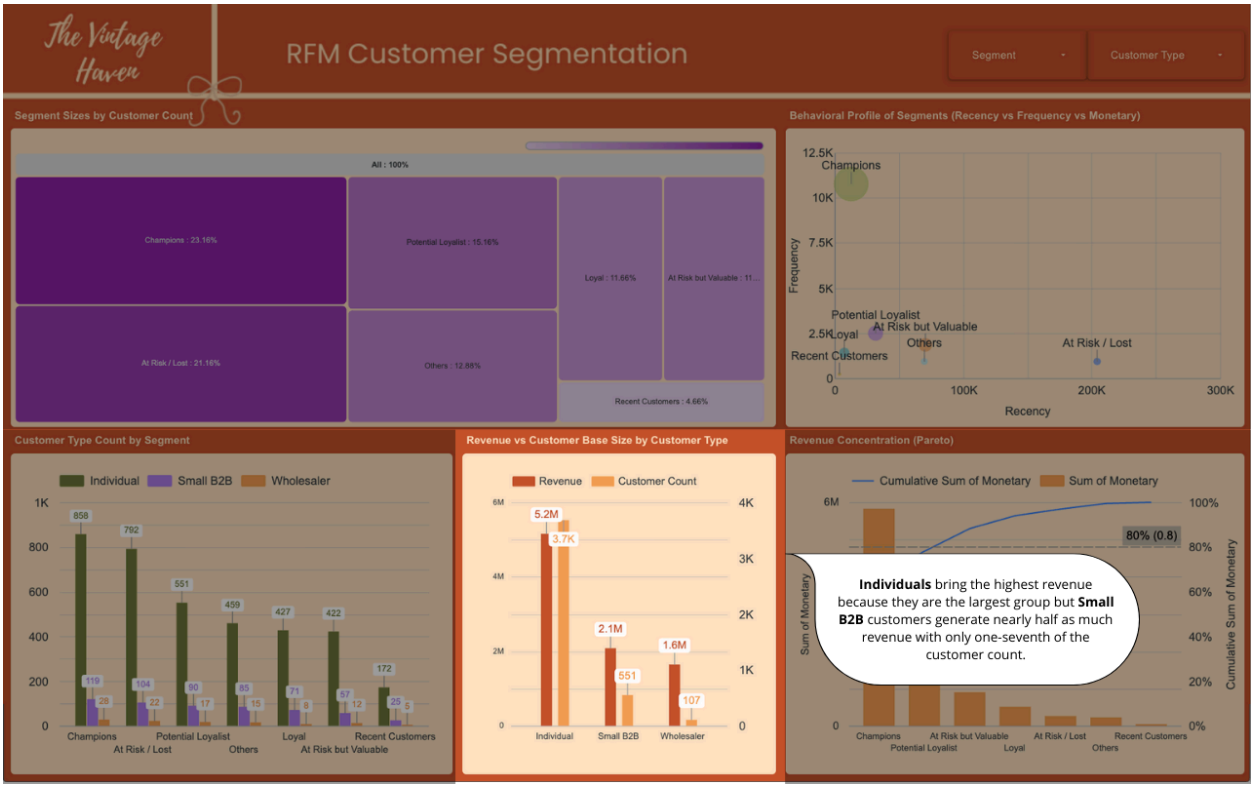
- In **every segment**, Individuals are by far the largest group.
- **Small B2B customers appear consistently in all segments**, usually at about 10–15% of the Individual count.
- **Wholesalers are the smallest group in all segments**, but still present everywhere

Key Takeaways:

- The segmentation is **not driven by customer type composition** each segment has a similar shape: mostly Individuals, then a smaller B2B portion, then a small number of Wholesalers.
- This confirms that the **RFM segments describe behaviour rather than customer category**. All three customer types can be Champions, Loyal, or At Risk - they just do so at very different scales.

For the marketing decision later, this means we should choose target segments based primarily on **behaviour (RFM)**, then refine tactics within those segments by type (Individual vs B2B vs Wholesaler), rather than assuming certain segments are “wholesale-only” or “retail-only”.

Chart IV: Revenue vs Customer Base Size by Customer Type



This visual shows how each **customer type** contributes to overall revenue and how that compares to their size.

1. Individuals — high revenue due to high volume

With **3,700+ customers**, Individuals generate **£5.2M**, but their contribution is driven by sheer scale rather than high spend per customer.

2. Small B2B — the most revenue-efficient group

Only **~550 customers** contribute **£2.1M**, giving them almost **3× higher average revenue per customer** than Individuals.

3. Wholesalers — small group, high value per order

Just **107 customers** bring in **£1.6M**, but their purchasing behaviour is more volatile and less predictable.

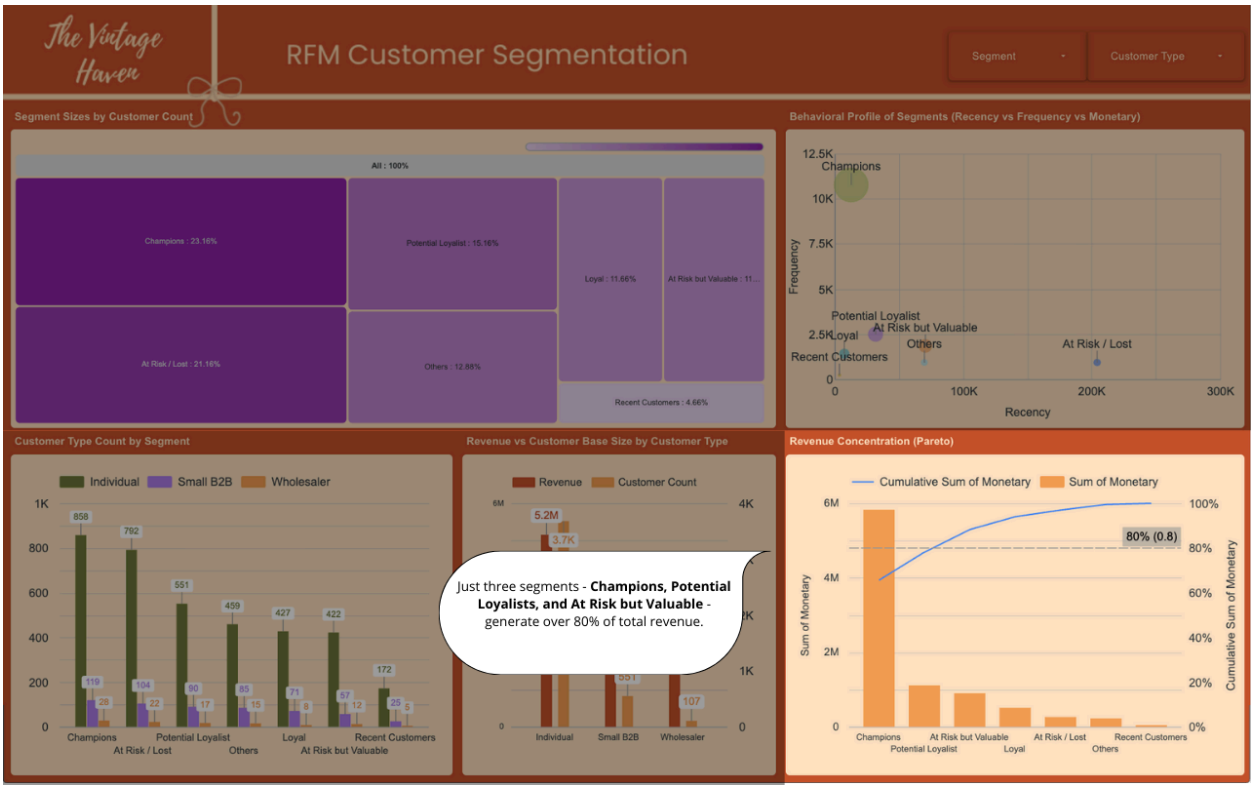
Key Takeaways

The most cost-effective marketing targets are:

- **Small B2B** — high ROI per customer
- **High-value Individuals** — large group worth nurturing
- **Wholesalers** — personalised, low-frequency outreach

This ensures the next campaign focuses on customers who deliver the **highest value per marketing pound**, not just the largest groups.

Chart V: Revenue Concentration (Pareto)



This Pareto chart shows how revenue is concentrated across RFM segments.

1. Champions drive the largest share of revenue

They contribute nearly **£6M** alone, making them the most valuable and predictable group to retain.

2. Potential Loyalists + At Risk but Valuable complete the 80/20 rule

Together with Champions, these three segments generate **over 80% of total revenue**, showing that value is heavily concentrated in a small set of behaviourally strong customers.

3. All remaining segments contribute only 20% of revenue

Loyal, Others, At Risk/Lost, and Recent Customers represent nearly half the customer base but deliver minimal and inconsistent spend.

4. Strategic takeaway

- Prioritise **Champions** (retain),
- **Potential Loyalists** (grow),
- **At Risk but Valuable** (reactivate).

Lower-value groups should receive lighter-touch, scalable messaging.

Marketing Recommendations

The RFM analysis shows that revenue is heavily concentrated: **Champions, Potential Loyalists, and At Risk but Valuable customers generate over 80% of total revenue**, while most other segments contribute little and require minimal investment. Marketing efforts should therefore prioritise the groups that combine high value with high potential for growth or recovery.

Priority Segments

1. Champions — Retain

Your most valuable and reliable buyers.

Actions: early access to collections, loyalty perks, personalised thank-you messages.

Goal: keep them engaged and prevent churn.

2. Potential Loyalists — Grow

Engaged and promising but not yet consistent.

Actions: follow-up emails, personalised recommendations, small incentives.

Goal: convert them into Champions.

3. At Risk but Valuable — Reactivate

High past spend but declining activity.

Actions: win-back emails, time-limited offers, reminders of favourite items.

Goal: recover lost revenue from high-value customers.

Secondary Segments

4. Loyal — Increase Spend

Active but low monetary value.

Actions: upselling, bundles, spend-based rewards.

5. At Risk / Lost — Automate

Low engagement and low value.

Actions: passive newsletter or seasonal emails only.

6. Recent Customers - Nurture

Too new to evaluate.

Actions: post-purchase follow-ups, review requests.

Tailoring by Customer Type

Use RFM to choose who to target, then adapt messaging to customer type:

- **Individuals:** seasonal promotions, product suggestions
- **Small B2B:** reorder reminders, bundle discounts, curated assortments
- **Wholesalers:** personalised outreach, early stock notifications

Results

Through RFM analysis and dashboard exploration, the project successfully identified which customer groups contribute the most value and should be prioritised in the next marketing campaign. The segmentation provided clear behavioural insights, highlighting Champions, Potential Loyalists, and At Risk but Valuable customers as the most impactful segments for future marketing efforts.

One key learning was that the **manual segmentation based on purchase quantity (Individuals, Small B2B, Wholesalers)** did not reveal strong behavioural differences within RFM segments. While useful for understanding revenue efficiency by customer type, it did not meaningfully improve the RFM scoring process and was ultimately unnecessary for segment definition.

Overall, the project achieved its core goals:

- Identifying high-value and at-risk customer groups
- Understanding how value is distributed across the customer base
- Providing clear, data-driven marketing recommendations

What I would do differently in a real-world environment

In a production setting, this work would be extended through:

- **Collaboration with marketing teams** to ensure segment definitions align with business priorities

- **Secure data handling** through governed storage (e.g., BigQuery, Snowflake) instead of local files
- **Automated dashboards** connected to scheduled ETL pipelines, ensuring recency and reliability
- **Role-based access controls** for sensitive customer data

Next steps to extend the project

Several valuable analyses remain that would further strengthen marketing decisions:

- **Behavioural deep dive**: understand which products each segment prefers, seasonal buying patterns, and cross-selling opportunities.
- **Seasonality analysis**: identify peak months and predict demand by segment to support inventory and campaign planning.
- **Campaign design**: develop targeted promotions, email flows, or loyalty programs tailored to high-priority segments.
- **Churn prediction model**: build a machine-learning model to proactively flag customers at risk before they fall into the At Risk/Lost group.
- **Customer lifetime value (CLV)** prediction: estimate long-term revenue per segment to prioritise budgets more precisely.

Together, these steps would transform the insights from this project into a fully actionable, data-driven customer strategy.