UCI Online Retail: *Geo-temporal Patterns, RFM Segments and Practical Recommendations*

Table of Contents

1. Introduction

Retail decision-makers need evidence-based insight to plan capacity, focus marketing and improve customer outcomes. This report analyses the UCI Online Retail dataset to link customer behaviour with geography and time, segment customers with RFM, and identify highvalue propensity for targeted actions.

The dataset provides invoice timestamps, customer IDs, country, quantity and unit price, enabling robust customer- and order-level analysis. Using these fields, we examine four questions: revenue concentration by country and diversification opportunities; monthly and weekday cycles that shape inventory, fulfilment and promotion timing; the distribution of customer value across recency, frequency and monetary contribution, including which segments drive disproportionate revenue; and which customers are most likely to become high value next and warrant targeted activation.

Our objectives align directly with those questions: ensure data quality for reliable customer metrics; quantify geo-temporal demand to guide calendar planning and operational capacity; derive actionable behavioural segments using RFM and clustering for differentiated retention and growth tactics; and build a predictive model of high-value propensity to prioritise outreach. Our approach proceeds in three steps: data cleaning and validation; geo-temporal exploratory analysis and RFM-based segmentation with K-means; and a supervised propensity model to rank customers by high-value likelihood. The results translate into targeted marketing, better peak-period planning and a clearer path to geographic revenue diversification.

2. Data Understanding and Cleaning

This section details the characteristics of the selected Online Retail dataset and outlines the comprehensive preprocessing pipeline developed to ensure data quality and suitability for subsequent customer segmentation analysis.

2.1 Dataset Characteristics

The analysis utilises the Online Retail dataset, which contains **541,909 transactions** from a UK-based online retailer. While the dataset's official metadata claimed no missing values, a direct programmatic inspection was performed to verify data quality and inform the cleaning strategy.

2.2 Data Preprocessing and Cleaning Steps

A multi-step cleaning process was executed to address data quality issues identified during the initial inspection.

2.2.1 Handling Missing Values

Contrary to its metadata, programmatic inspection revealed that **135,080 records** lacked a CustomerID as shown in Figure 2-1. As this identifier is essential for any customer-level analysis, all rows with a null CustomerID were removed. This foundational step reduced the working dataset to **406,829 records**.

```
--- Count of Missing Values ---
InvoiceNo          0
StockCode          0
Description     1454
Quantity           0
InvoiceDate        0
UnitPrice          0
CustomerID    135080
Country            0
```

Figure 2-1: Programmatic Check of Missing Values in the Raw Dataset

2.2.2 Addressing Outliers and Invalid Data

```
--- Descriptive Statistics for Numerical Data ---
            Quantity      UnitPrice     CustomerID
count  541909.000000  541909.000000  406829.000000
mean        9.552250       4.611114   15287.690570
std       218.081158      96.759853    1713.600303
min    -80995.000000  -11062.060000   12346.000000
25%         1.000000       1.250000   13953.000000
50%         3.000000       2.080000   15152.000000
75%        10.000000       4.130000   16791.000000
max     80995.000000   38970.000000   18287.000000
```

Figure 2-2: Descriptive Statistics of Raw Numerical Data

Descriptive statistics of the raw data revealed invalid entries, including negative Quantity and UnitPrice values which signify returns or cancelled orders as shown in Figure 2-2. To ensure the analysis was based solely on valid sales, records with non-positive Quantity or UnitPrice, as well as transactions identified as cancellations (e.g., InvoiceNo starting with 'C'), were filtered out. This refined the dataset to **397,884 transaction records**.

2.2.3 Correcting Data Types and Removing Duplicates

Data types were corrected to facilitate analysis, notably converting InvoiceDate to a datetime object for temporal calculations and CustomerID to an integer. A final de-duplication step then removed **5,192 duplicate entries**, resulting in a fina l dataset of **392,692 unique transactions**.

```
After removing duplicates, the dataset has 392692 rows remaining.
<class 'pandas.core.frame.DataFrame'>
Int64Index: 392692 entries, 0 to 541908
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    392692 non-null  object
 1   StockCode    392692 non-null  object
 2   Description  392692 non-null  object
 3   Quantity     392692 non-null  int64
 4   InvoiceDate  392692 non-null  datetime64[ns]
 5   UnitPrice    392692 non-null  float64
 6   CustomerID   392692 non-null  int32
 7   Country      392692 non-null  object
dtypes: datetime64[ns](1), float64(1), int32(1), int64(1), object(4)
memory usage: 25.5+ MB
```

Figure 2- 3: Structure of the Final Cleaned Dataset

2.3 Sampling Strategy

The entire cleaned dataset was used for this analysis rather than a sample. This decision was driven by the requirements of the intended RFM (Recency, Frequency, Monetary) model, which depends on a complete transaction history per customer to accurately calculate scores. Sampling would compromise the integrity of these metrics, so the full dataset of 392,692 records was retained to ensure analytical validity.

3. Exploratory Data Analysis

This section addresses Objective 2 by examining temporal and geographic demand patterns and lays the foundation for Objective 3 on customer segmentation.

3.1 Analysis Scope

This EDA examines 18,532 orders from 4,338 customers across 38 countries spanning December 2010 to December 2011, as shown in Figure 3-1. The analysis employs eleven visualizations across four dimensions: geographic distribution, temporal patterns, customer segmentation, and product performance, providing systematic insights into business patterns and strategic opportunities.

Total Revenue	8,887,209
Total Orders	18,532
Avg Order Value	480
Total Customers	4,338
Total Products	3,665

Figure 3-1: Key Metrics

3.2 Geographic Distribution

As illustrated in Figure 3-2, the UK dominates with £7,285,025, representing 82% of total sales, while Netherlands ranks second at £285,446, accounting for 4% of revenue. The

remaining 36 countries contributed only 14%, with 17 countries generating under £10,000 each. This longtail distribution presents both risk through single market over-reliance and opportunity via untapped international potential.

Implication: The 82% single-market concentration represents Critical Risk to business continuity. Priority action: diversify into Netherlands, Germany, France, Ireland, Switzerland to reduce UK dependency below 60% within 12 months.

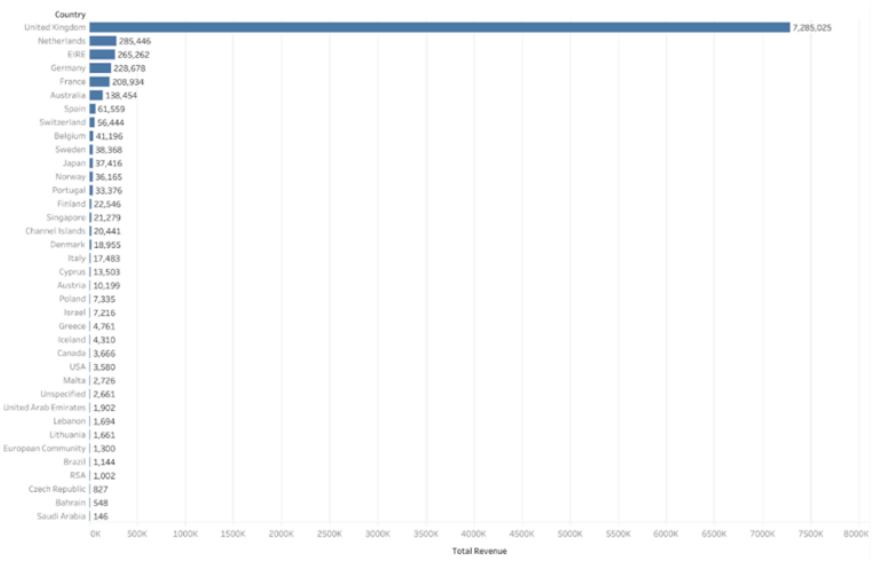


Figure 3-2: Country Revenue Ranking

3.3 Temporal Patterns

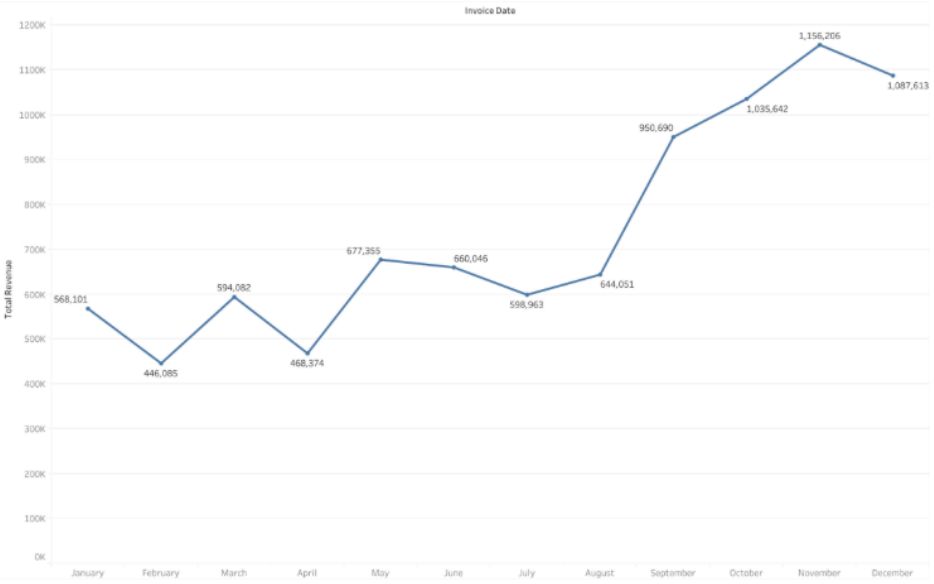


Figure 3-3: Monthly Revenue Trend

Figure 3-3 demonstrates that revenue fluctuated between £446,085-£677,355 during most of the year until September's 47% surge to £950,690, peaking at £1,156,206 in November. The

fourth quarter accounts for 37% of annual revenue, indicating critical dependence on Christmas shopping.

Implication: The 37% fourth-quarter concentration creates cash flow volatility. Actions: increase Sept-Nov inventory by 40%, launch pre-season promotions, and develop counterseasonal products for Q1-Q3.

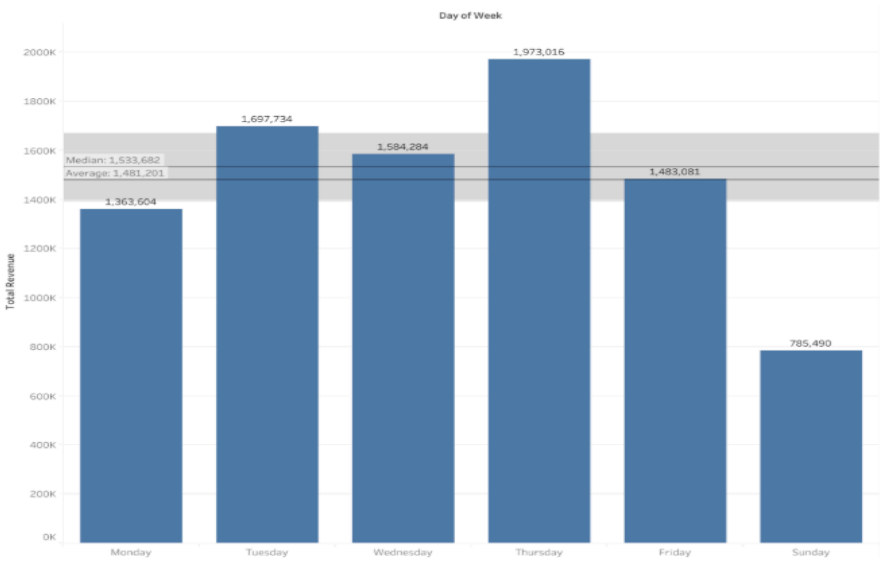


Figure 3-4: Weekly Sales Pattern

As shown in Figure 3-4, Thursday leads at £1,973,016, representing 33% above average, followed by Tuesday at £1,697,734 and Wednesday at £1,584,284. Sunday drops to £785,490, reaching only 53% of the average. This pattern enables predictable resource allocation.

Implication: Sunday revenue of £785,490 represents only 40% of Thursday's level of £1,973,016, suggesting operational inefficiency. Consider implementing Sunday-specific promotions or adjusted operating hours.

3.4 Customer Segmentation Analysis

Figure 3-5 presents four behavioral segments identified by Kumar and Reinartz in 2018: Occasional customers comprise 92.2% with 4,001 individuals, Regular customers account for 6.3% with 274 individuals, Frequent customers represent 1.1% with 48 individuals, and Super Customers constitute 0.3% with 14 individuals. The small Super and Frequent cohort of 1.4% warrants disproportionate attention given their higher revenue contribution.

Implication: The large Occasional customer base of 92.2%, representing 4,001 customers, indicates opportunity for frequency-building programs, while the small high-frequency cohort of 1.4%, comprising 62 customers, warrants premium service and retention focus given their likely higher per-customer value.

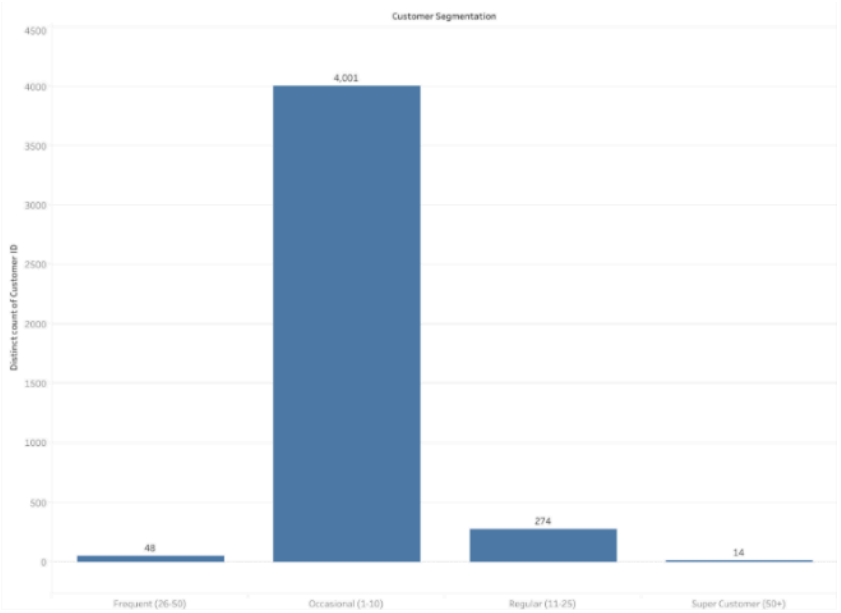


Figure 3-5: Customer Segmentation Distribution

As depicted in Figure 3-6, Super Customers cluster in high spending ranges with outliers exceeding £250,000, while Occasional customers concentrate in low-spending quadrants. Clear separation with minimal overlap validates the classification and highlights the need for differentiated engagement strategies.

Implication: Clear visual separation between segments confirms distinct customer behaviors, providing foundation for segment-specific engagement approaches.

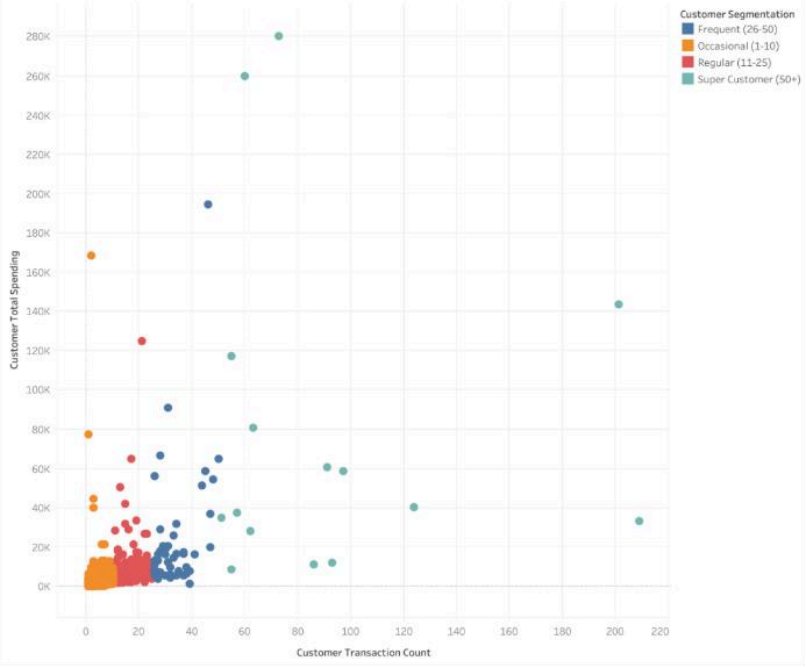


Figure 3-6: Customer Value Analysis

Figure 3-7 reveals that the Super Customer median of £40,520 is 67 times higher than the

Occasional customer median of £600. The steepest value increase occurs at the Occasional-toRegular transition, showing an 8.3-fold increase. The wide Super Customer interquartile range spanning £25,000 to £100,000 indicates sub-segmentation opportunity.

Implication: Prioritize Occasional activation programs for highest ROI; implement tiered retention strategies aligned with exponential value growth.

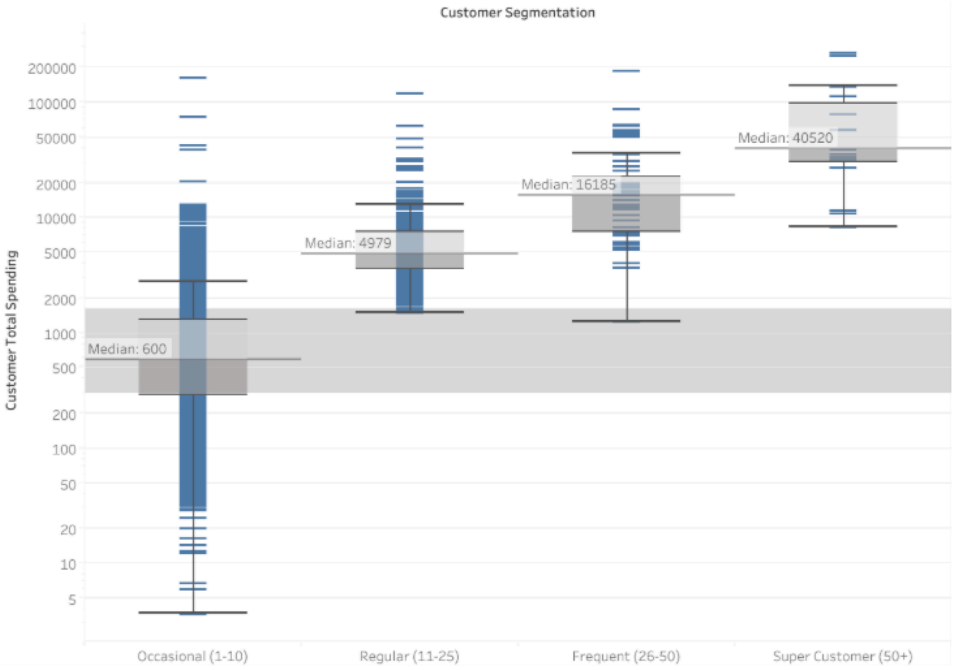


Figure 3-7: Customer Lifetime Value by Segment

3.5 Product Analysis

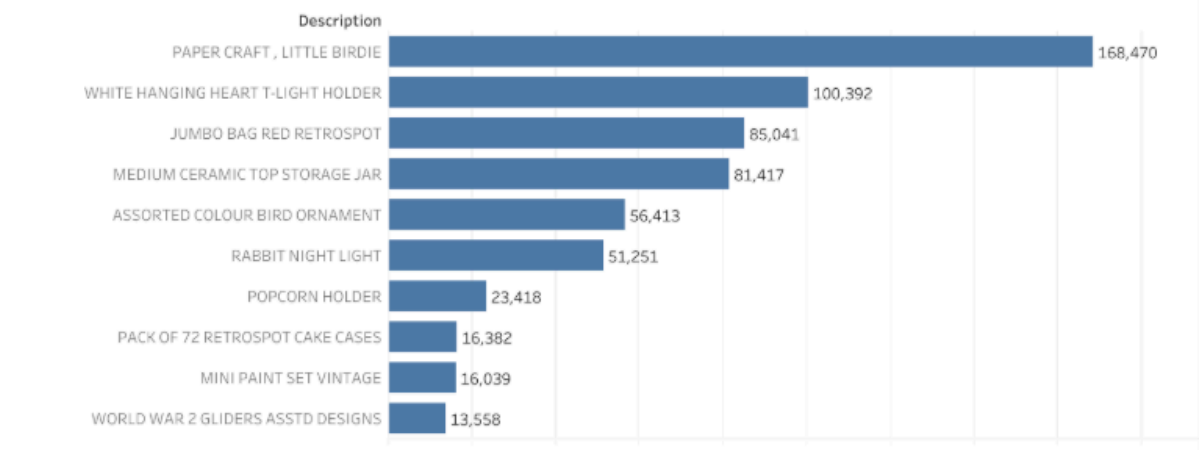


Figure 3-8: Top 10 Products by Revenue (Global)

According to Figure 3-8, Paper Craft Little Birdie dominates at £168,470, followed by White Hanging Heart T-Light Holder at £100,392. A sharp drop to £13,558 by the tenth-ranked product indicates hero product concentration and vulnerability.

Implication: Single product concentration of £168,470 creates portfolio risk. Priority: develop five to seven complementary craft-themed products to diversify revenue sources.

3.6 Key Findings Summary

Geographic Risk (Critical): 82% UK concentration creates single-market vulnerability. Immediate diversification into Netherlands, Germany, and France required.

Seasonal Risk (High): The fourth quarter accounts for 37% of annual revenue, peaking at £1.16 million in November. September's 47% surge demands proactive inventory planning.

Operational Efficiency (Quick Win): Thursday revenue of £1.97 million is 2.5 times higher than Sunday revenue of £0.79 million, enabling day-specific promotional strategies.

Customer Value Concentration (High): Super customers exhibit a median lifetime value of £40,520, representing a 67-fold differential compared to Occasional customers at £600. With 92.2% of the customer base in the Occasional segment, substantial activation potential exists.

Segmentation ROI (High): The Occasional-to-Regular transition demonstrates an 8.3-fold value increase, representing the highest return on investment for conversion initiatives. Super customer spending ranges from £25,000 to £100,000 based on interquartile range analysis, supporting premium sub-segmentation strategies.

Product Risk (Medium): Leading product generates £168,000 in revenue, with sharp decline to £13,600 at tenth position, indicating portfolio vulnerability requiring diversification.

4. Advanced Analysis

This section aims to further explore customer segmentation, completing two modules: customer consumption behaviour modelling and high-value potential customer prediction. To achieve this goal, we applied clustering and predictive modelling techniques.

4.1 Feature Engineering

Features were derived to transform raw transactional factors into numerical variables suitable for modelling (*Table 4-1*).

Table 4-1: Summary of Engineered Features and Their Business Implications

Feature Name	Statement	Value Type	Practical Business Implications
firstPurchaseDate	First purchase time	POSIXct	Time when the customer places the first order
lastPurchaseDate	Last purchase time	POSIXct	Time when the customer places the last order
customer_lifetime	Customer Lifecycle	number	The time span between the first and last purchase

frequency	Purchase frequency	integer	The number of orders after deduplication reflects the level of activity
totalSpent	Total consumption amount	number	The sum of purchase amounts in all orders, measuring customer value
recency	The number of days since the last purchase	number	Reflects the customer's "dormantness" or "activity" (the larger the value, the less active it is)
productVariety	Number of different product types	integer	Indicates the diversity of products that customers are interested in
avgPurchaseInterval	Average purchase interval	number	If the customer purchases multiple times, it reflects the average consumption frequency; if the customer only purchases once, it is 0

4.2 Clustering to Explore Customer Consumption Patterns

To explore customer consumption patterns, the *K-means* clustering algorithm was employed, leveraging 6 numerical features, recency, frequency, totalSpent, avgPurchaseInterval, productVariety, and customer_lifetime, all suitable for *Euclidean distance calculation*. These standardized features ensured equal contribution to clustering results.

The Elbow method identified **k = 4** as the optimal number of clusters (*Figure 4-1*). Clustering was performed using the kmeans() function, producing **4** distinct groups: Cluster 1 (949 observations), Cluster 2 (1,462 observations), Cluster 3 (17 observations), and Cluster 4 (1,910 observations). To visualize feature differences across clusters, a radar chart was created using fmsb() (*Figure 4-2*), helping interpret customer profiles such as high-spending, dormant, or diverse buyers. By integrating clustering results with customer country data, we identified the **top 3** countries associated with each cluster (*Figure 4-3*), enabling deeper geographic insight into customer behaviour.

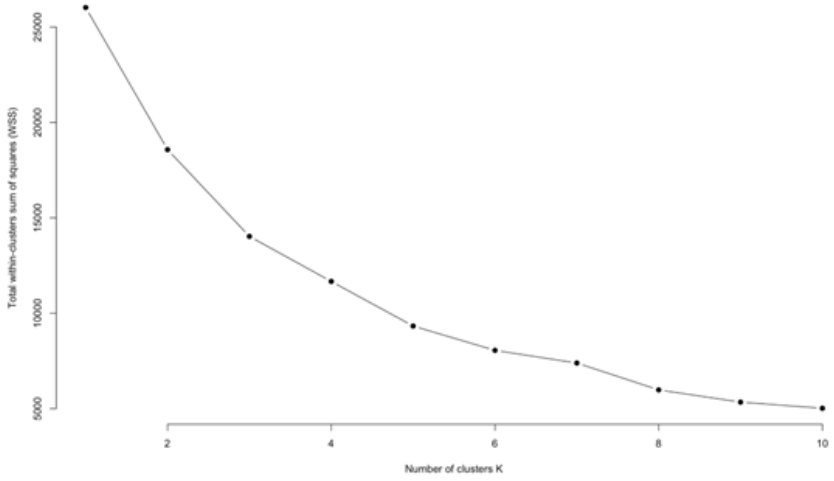


Figure 4-1: Elbow Method to Determine the Number of Clusters K

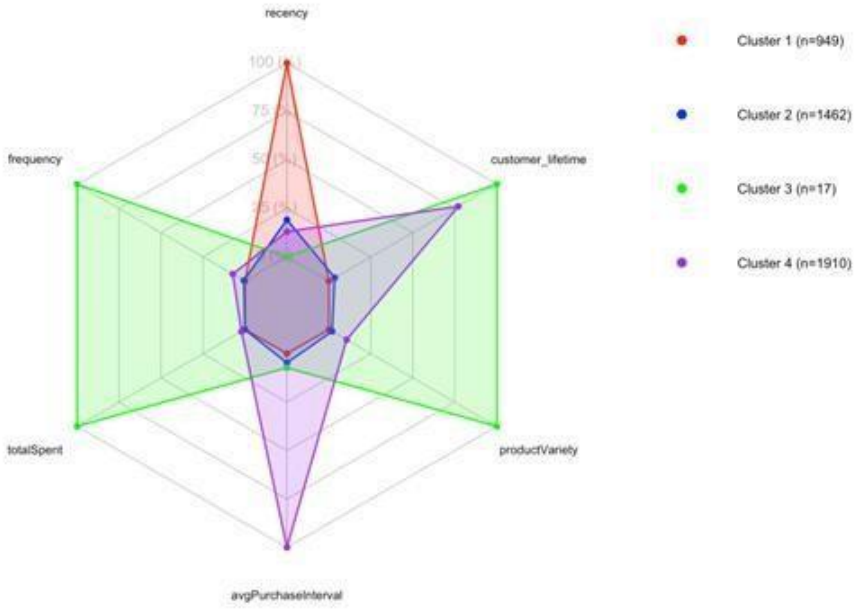


Figure 4-2: Customer Consumption Patterns Across Clusters

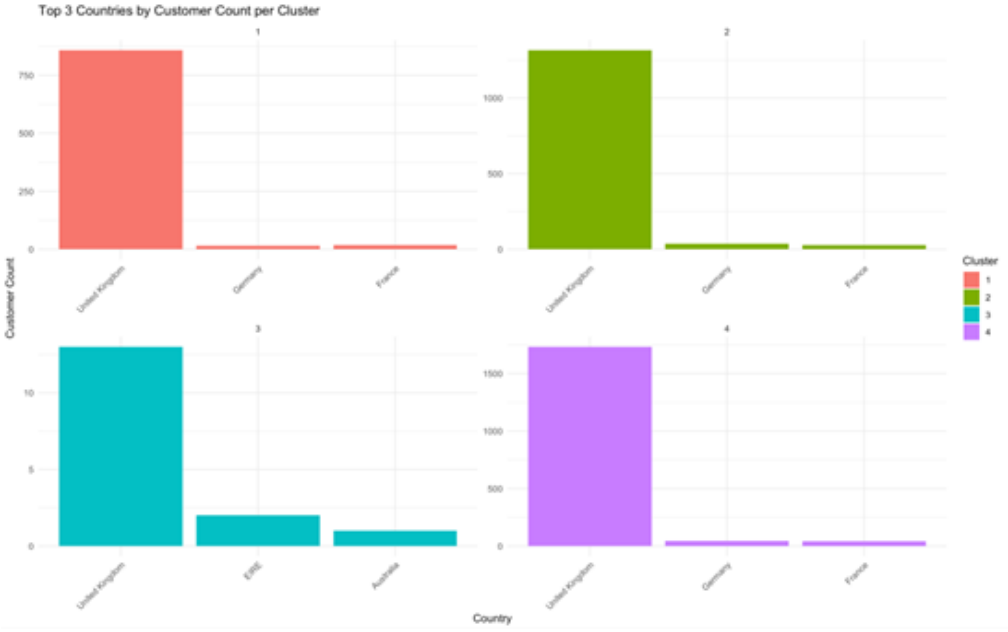


Figure 4-3: Distribution of Top 3 Countries by Customer Count per Cluster

4.2.1 Summary of Main Findings

The four clusters (VIPs, Active, Occasional Buyers, and Churned Customers) highlight distinct behavioural patterns for targeted marketing (*Table 4-2*).

Table 4-2: Consumer Segmentation Model

Cluster	Number of Customers	Main Feature	Top 3 Countries (by Customer Count)	Label

Cluster 1	949	High recency, low frequency, low spending, short lifetime	United Kingdom, Germany, France	Churned
Cluster 2	1462	Balanced across all behavioral indicators	United Kingdom, Germany, France	Active
Cluster 3	17	Very high frequency, spending, product variety; low recency	United Kingdom, EIRE, Australia	VIP
Cluster 4	1910	Long purchase intervals; moderate performance on other metrics	United Kingdom, Germany, France	Occasional Buyers

4.2.2 Evaluation of the K-means Method

The K-means algorithm provides an efficient and interpretable approach for identifying behavioural segments based on numerical features such as recency, frequency, and totalSpent. Its simplicity and scalability make it suitable for large customer datasets. However, reliance on Euclidean distance limits the inclusion of categorical factors like product category, meaning that differences driven by product types may not be fully reflected in the segmentation results.

4.3 Prediction for High-potential Customer

To predict potential high-value customers, an Artificial Neural Network (ANN) classifier was developed. The choice of ANN was based on its strong capability to model non-linear relationships and interactions between multiple customer behavioural attributes, which are likely to influence purchasing power.

4.3.1 The prediction process consisted of the following key steps

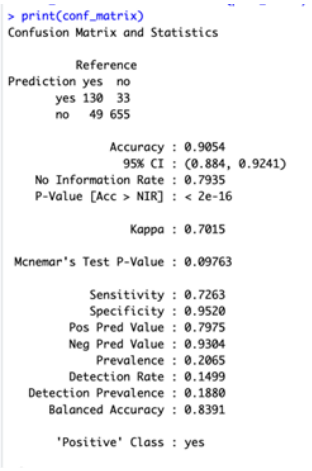
According to the quantile analysis of customer spending (*see Figure 4-4*), customers whose total spending exceeds approximately £2000 (the 80th percentile, ≈ £2055) were categorized as high-value customers, representing roughly the top 20% of the customer base. This threshold ensures that only the top-spending segment is classified as high-value, providing a data-driven definition for subsequent prediction.

```
> quantile(customer_data$totalSpent, probs = c(0.1, 0.2, 0.5, 0.8, 0.9))
    10%     20%     50%     80%     90%
155.296 249.344 668.570 2055.050 3640.841
```

Figure 4-4: Quantile Distribution of Total Spending Used for High-Value Customer Definition

The selected features included frequency, recency, productVariety, customer_lifetime, avgPurchaseInterval, which were standardized. *Model training was conducted* with **5-fold cross-validation** to evaluate performance. A grid search over hidden layer sizes (size = 3, 5, 7) and weight decay values (decay = 0.01, 0.1) was performed to optimize model performance, among them, the best parameters were size = 3 and decay = 0.1. Finally, the

trained model was applied to the test set, and predictions were compared with true labels, achieving an accuracy of **90.54%** and precision of **79.8%** (see *Figure 4-5*).



```
> print(conf_matrix)
Confusion Matrix and Statistics

          Reference
Prediction yes  no
       yes 130  33
       no   49 655

               Accuracy : 0.9054
                 95% CI : (0.884, 0.9241)
    No Information Rate : 0.7935
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.7015

 Mcnemar's Test P-Value : 0.09763

            Sensitivity : 0.7263
            Specificity : 0.9520
         Pos Pred Value : 0.7975
         Neg Pred Value : 0.9304
             Prevalence : 0.2065
         Detection Rate : 0.1499
   Detection Prevalence : 0.1880
      Balanced Accuracy : 0.8391

       'Positive' Class : yes
```

Figure 4-5: Confusion Matrix and Classification Performance Metrics for the ANN Model

4.3.2 Cross-Segment Analysis

Given that we have clustered the data into four segments (Cluster 1–4) and developed a highperforming ANN model to identify high-value customers. **Cluster 3** is composed entirely of high-value customers **(100%)**, indicating it represents a core profit group. In contrast, **Cluster 4** contains a significant proportion of high-value customers (40.7%), suggesting that this group includes users with untapped consumption potential (*see Figure 4-6*). Although their purchase intervals are relatively long, their extended customer lifetimes demonstrate loyalty and a strong likelihood of repeat purchases.

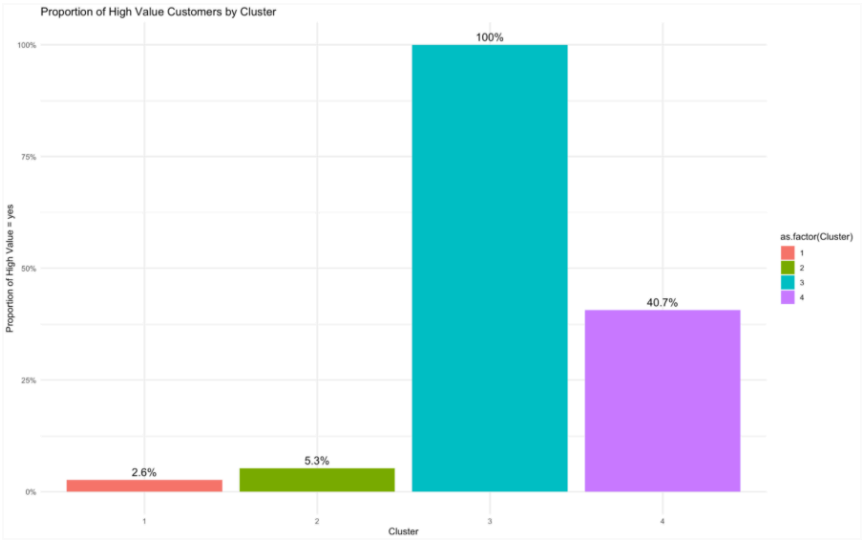


Figure 4-6: Proportion of High-Value Customers Across Clusters

4.3.3 Evaluation of the ANN Model

The Artificial Neural Network (ANN) achieved an overall accuracy of 90.54%, indicating strong predictive capability in distinguishing high-value customers. The model reached a

precision of 79.8% and a recall of 72.6%, suggesting a reasonable trade-off between identifying potential high-value customers and maintaining prediction reliability. However, as a black-box model, the ANN offers limited interpretability, making it difficult to directly explain individual predictions to business stakeholders. While this study focused on evaluating predictive performance, future work could apply techniques such as Self-Organizing Maps (SOM) or feature importance visualisations to explore the relationship between input variables and predicted outcomes.

5. Insights & Recommendations

This section presents evidence-based, data-driven commercial recommendations grounded entirely in the analytical results from Sections 3 and 4. The insights translate key findings from geographic, temporal, segmentation, and predictive analyses into actionable strategies to improve revenue stability, customer retention, and portfolio resilience. Limitations and directions for future work are also discussed to ensure transparency and credibility.

5.1 Geographic Market Optimisation

Evidence: The geographic analysis showed that 82% of sales revenue originates from the UK (Figure 3-2), while the Netherlands, Germany, and France together contribute less than 10%. Such a heavy dependence on a single market exposes the business to macroeconomic and regulatory risks.

Actionable Insight:

- Diversify operations by expanding targeted online marketing campaigns in the Netherlands, Germany, and France, which already show positive transaction trends.
- Use existing logistics and supplier networks to pilot regional promotions for topperforming product categories.
- Set a measurable goal to reduce UK market share to below 65% within 12 months, improving market resilience and revenue balance.

5.2 Temporal Sales Strategy

Evidence: Monthly revenue fluctuates significantly, with a 47% surge in September and a peak of £1.16M in November, indicating extreme Q4 seasonality. Weekly data also show strong Thursday concentration (33% above average) and weak Sunday performance (53% below average).

Actionable Insight:

- Increase production and inventory capacity by 40% from September to November to meet seasonal peaks.

- Launch pre-season campaigns (e.g., “Early Autumn Deals”) to smooth the Q4 revenue curve and reduce last-minute logistic pressure.
- Address low weekend sales by offering Sunday-only online discounts or flash promotions, leveraging underused operational capacity.

These evidence-based temporal adjustments can stabilise cash flow, enhance operational efficiency, and maintain customer engagement across the year.

5.3 Customer Segmentation and Retention

Evidence: The RFM and K-means analyses identified four customer segments: Cluster 1: Churned (inactive, short lifecycle) , Cluster 2: Active (balanced frequency and spending) , Cluster 3: VIP (high frequency, high spending, diverse products) , Cluster 4: Occasional (long purchase interval, moderate spending)

The ANN model (accuracy 90.54%) further revealed that Cluster 4 contains 40.7% high-value potential customers, representing the most promising target for marketing activation.

Actionable Insight:

- Develop personalised re-engagement campaigns for Cluster 4 (e.g., loyalty coupons, birthday offers, reminder emails).
- Strengthen VIP retention by offering exclusive bundles or early access to limited-edition craft items.
- Monitor the transition rate from Occasional → Regular → VIP as a key marketing performance indicator.
- Deploy predictive scores from the ANN model to prioritise marketing budgets toward customers with the highest probability of high-value conversion.

These targeted actions transform analytical segmentation into measurable business growth.

5.4 Product Portfolio Development

Evidence: Product analysis showed that revenue is heavily skewed toward one product — Paper Craft Little Birdie (£168,470) — while the tenth-ranked product contributes only £13,558. This steep drop indicates portfolio concentration risk.

Actionable Insight:

- Expand the “craft” product line by introducing 5–7 complementary items in the same thematic category (e.g., DIY kits, seasonal decorations).
- Use market basket analysis to identify frequent co-purchases and design bundle promotions to lift average order value.

- Phase out low-performing products gradually while introducing data-informed new designs that appeal to high-frequency clusters.

This diversification strategy enhances resilience against product-specific volatility and broadens customer appeal.

5.5 Limitations and Future Work

Despite generating strong analytical evidence, this study has several limitations:

- The dataset covers only one year (2010–2011), missing long-term or post-pandemic market dynamics.
- The absence of product category metadata restricts deeper analysis of cross-category purchasing behaviour.
- The ANN model, while accurate, functions as a "black box" and limits interpretability for decision-makers.

Future work should extend analysis to multi-year datasets, include categorical and regional macroeconomic features, and explore explainable models (e.g., decision trees or SHAP visualisation) to improve transparency and business trust. Additionally, A/B testing on targeted promotions for Cluster 4 customers could empirically measure uplift and validate model-driven recommendations.

6. Conclusion

Using the UCI Online Retail data, we established a geo-temporal baseline with UKconcentrated sales, pronounced Q4 peaks, clear weekday effects, and a steep value skew where a small cohort drives disproportionate revenue. RFM with K-means yielded four behavioural segments for targeted engagement, and a simple ANN flagged customers with higher propensity to become high value. Together, these results point to practical actions in market diversification, calendar and capacity planning, and prioritised outreach.

However, these insights have limits. The data covers one year of online transactions only; rows without customer IDs were excluded; classes are imbalanced; and features are largely numeric, which limits interpretability.

Future work should extend the analysis across multiple years and include offline and campaign data. Adding categorical and sequence features would give richer signals. We should report clustering validity metrics, explore more interpretable and cost-sensitive models, and run A/B tests to measure the real impact of the recommended actions.

References

Kumar, V., & Reinartz, W. (2018). Customer Relationship Management: Concept, Strategy, and Tools (3rd ed.). Springer. https://doi.org/10.1007/978-3-662-55381-7