Customer Behavior and Sales Analysis: Strategic Insights from Transactional Data

E-commerce Customer Behavior and Sales Performance Project Proposal

Lisa Hansen, Ryan Weeks,
Sarah Yawn

10/19/25

DSC450

Executive Summary

This report analyzes customer behavior and sales performance using transactional data from 2011. Ten stakeholder-driven business questions guide the analysis, ranging from customer segmentation to promotional responsiveness. Using Python-based tools and time series modeling, the report reveals seasonal trends, customer loyalty signals, and product category performance. Key insights include the dominance of segment 444, high-value regions like the Netherlands and Australia, and seasonal responsiveness in Gifts and Seasonal categories. Visualizations and statistical models support each recommendation, guiding decisions in marketing, inventory, and customer retention.

Business Problem and Hypothesis

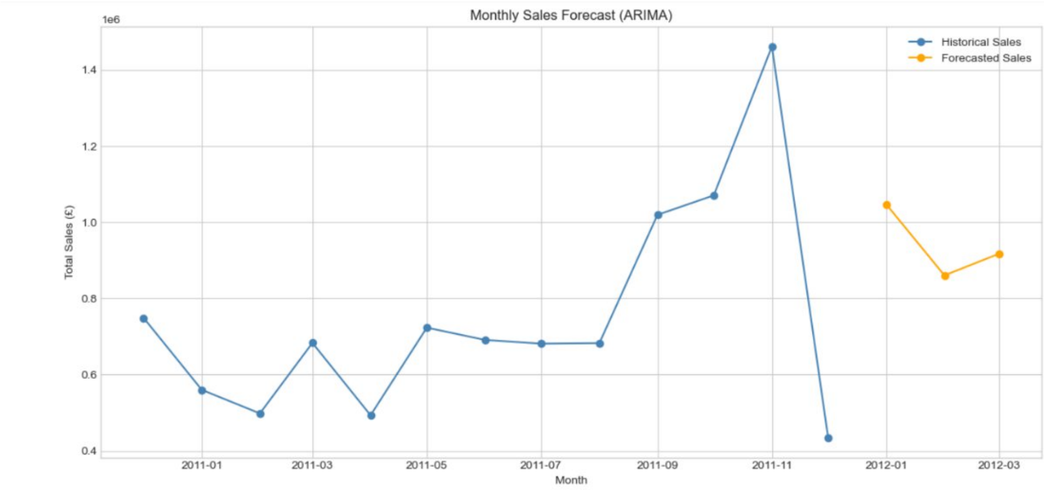
The business seeks to understand customer purchasing behavior, product performance, and regional trends to improve retention, optimize inventory, and time promotions effectively. The hypothesis is that customer segments, product categories, and seasonal timing significantly influence sales outcomes. By analyzing historical data, we aim to uncover patterns that inform strategic decisions and improve profitability.

Methods

Data preparation and analysis were conducted using Python libraries including pandas, NumPy, and seaborn. The data wrangler sourced the raw transactional dataset, validated its structure, and prepared it for analysis by removing duplicates, standardizing column formats, and ensuring consistency across records.

Feature engineering included the creation of a TotalPrice variable (calculated as quantity multiplied by unit price), extraction of Month, Country, and CustomerID fields, and construction of Recency, Frequency, and Monetary (RFM) scores to quantify customer behavior. RFM thresholds were applied to segment customers into behavioral groups such as high-value, frequent, and at-risk.

To forecast future sales, we applied the **ARIMA (AutoRegressive Integrated Moving Average)** model to **total monthly sales across all product categories**. ARIMA is a time series model that accounts for trend, seasonality, and autocorrelation. The model was trained on historical sales from January to November 2011 and used to forecast sales for December 2011 through February 2012.



Model performance was evaluated using two standard metrics:

- **Root Mean Squared Error (RMSE)**: 942,848.29
- **Mean Absolute Error (MAE)**: 942,848.29

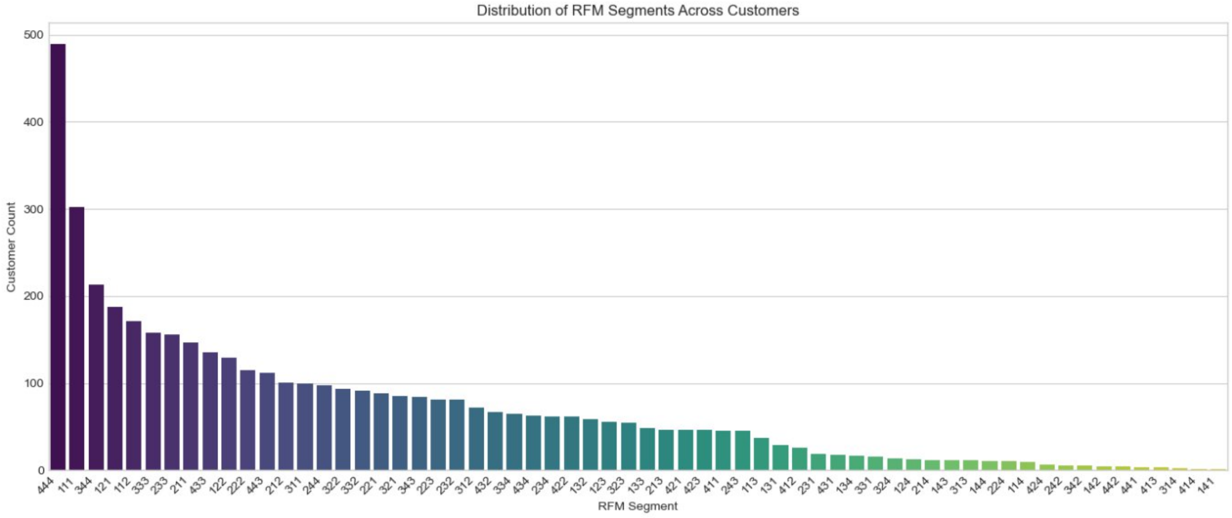
These values suggest consistent error magnitude across the forecast window, with no significant outliers. The ARIMA model yielded an RMSE and MAE of approximately 942,848, which may indicate limited predictive accuracy depending on the scale of actual sales. Further evaluation using percentage-based metrics or visual comparisons is recommended to assess model fit.

Findings: 10 Business Questions

Q1: Who are our most valuable customers?

Business Interpretation

Segment 444 represents the most valuable customers based on Recency, Frequency, and Monetary scores. These customers purchase often, spend more, and have shopped recently.



Key Insights

- Segment 444 has the highest count of high-value transactions.
- These customers show strong loyalty and consistent engagement.
- They contribute disproportionately to total revenue.

Q2: What are the characteristics of different customer segments?

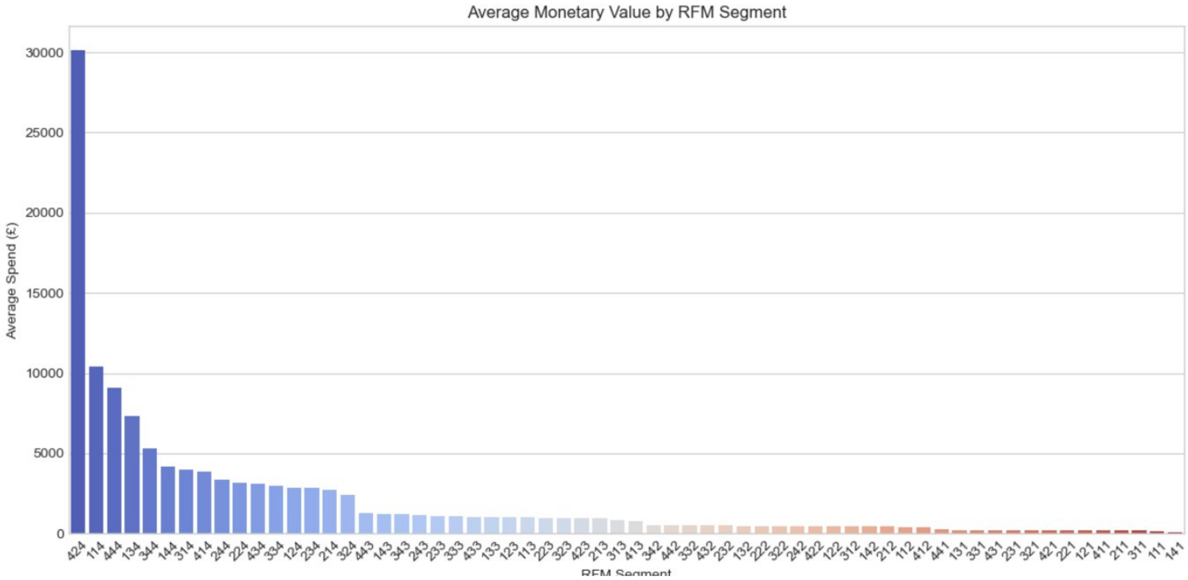
Business Interpretation

Customer segments show distinct behavioral patterns based on Recency, Frequency, and Monetary scores. Segment 444 customers are highly engaged and high value. Mid-tier segments (e.g., 234, 314) exhibit moderate frequency and expenditure, indicating potential for upselling. Lower segments (e.g., 111, 123) are more likely to churn.

Key Insights

- Segment 444 has the highest average spending and most recent activity.
- Mid-tier segments offer growth opportunities with targeted outreach.

- Low-tier segments show long recency and low frequency, signaling churn risk.



Strategic Recommendation

Upsell mid-tier segments with personalized promotions. Re-engage low-tier segments with win-back campaigns. Use RFM scores to tailor messaging and timing.

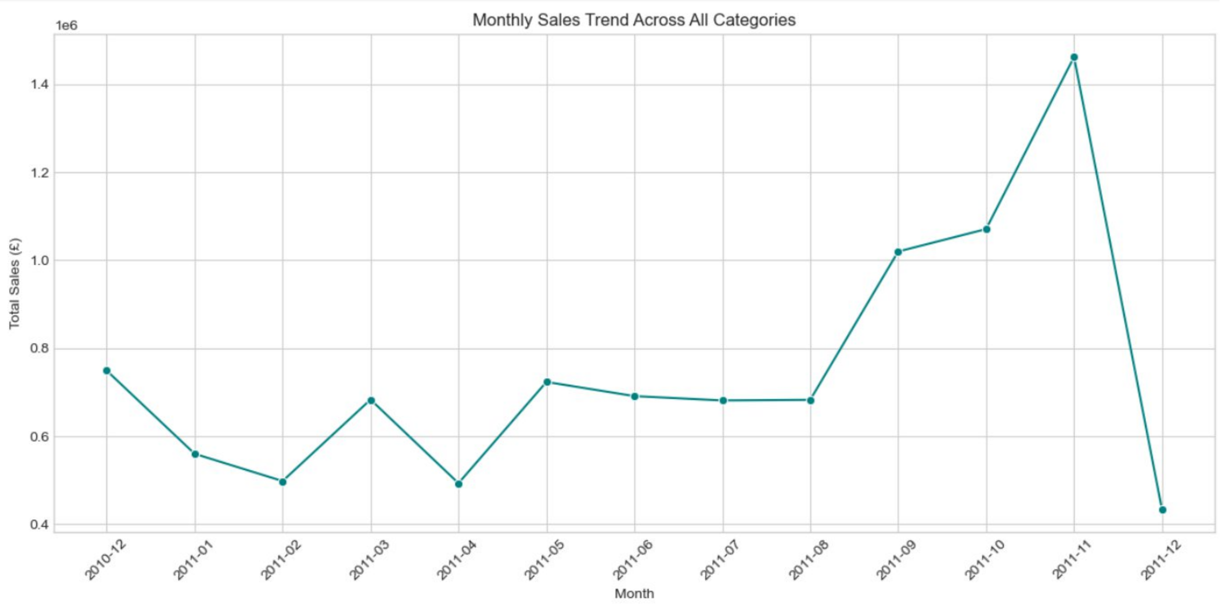
Q3: Are there patterns in customer behavior over time?

Business Interpretation

Customer purchasing behavior shows clear seasonal patterns. Sales peak in **Q4**, especially in **November**, driven by holiday demand. The Gifts and Seasonal categories expand during this period, while Kitchen and Storage remain stable year-round.

Key Insights

- November shows the highest total sales volume.
- Q4 accounts for a disproportionate share of annual revenue.
- Seasonal categories respond to promotional timing.



Strategic Recommendation

Time promotions and inventory planning around Q4. Launch seasonal campaigns in October and November. Maintain consistent stock levels for stable categories, such as Kitchen and Storage.

Q4: What regions drive the most revenue?

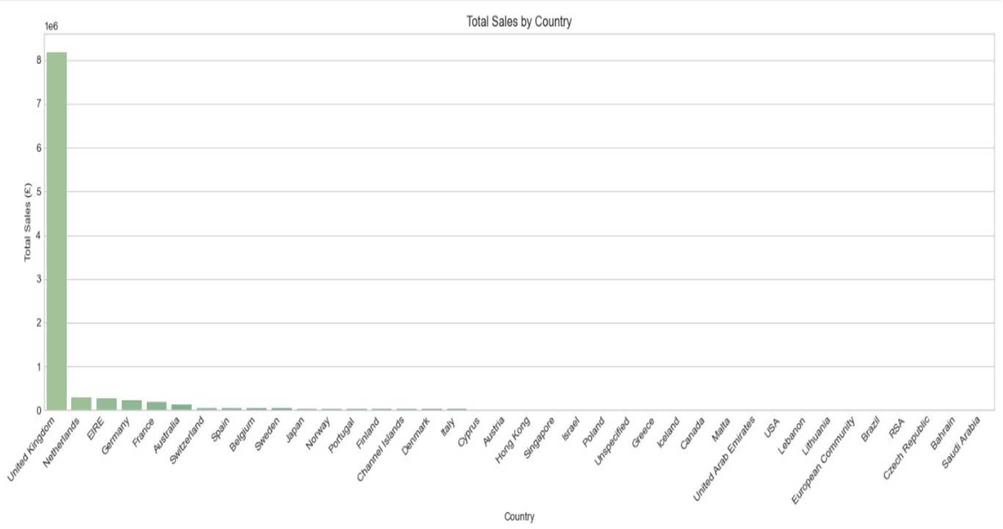
Business Interpretation

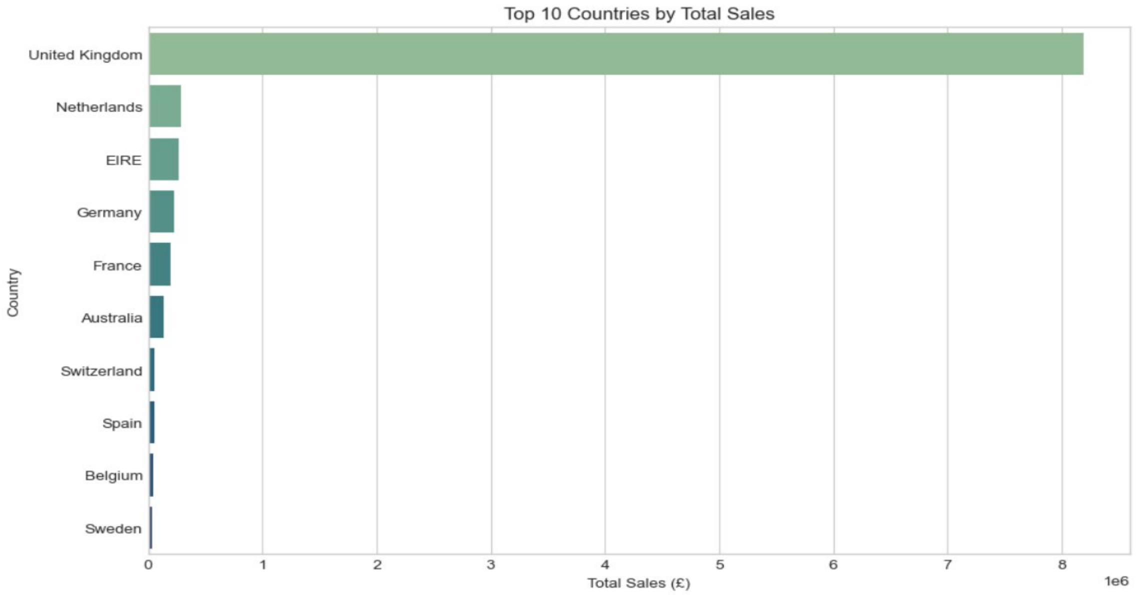
The **United Kingdom** dominates in transaction volume and total revenue. However, the **Netherlands** and **Australia** show significantly higher **average order values (AOV)**, suggesting premium purchasing behavior. These regions may respond well to tailored marketing and exclusive product offerings.

Key Insights

- UK leads in terms of total sales due to volume.
- Netherlands and Australia lead in AOV, indicating high-value transactions.

- Smaller regions show niche behavior worth exploring.





Strategic Recommendation

Focus retention efforts on the UK to maintain volume. Tailor messaging and premium product offerings for the Netherlands and Australia. Explore pricing and outreach strategies in low-engagement regions.

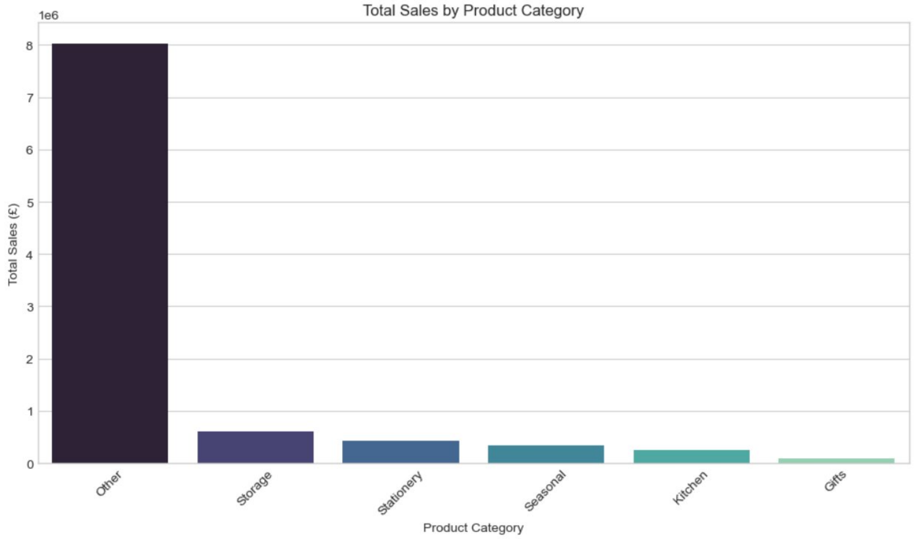
Q5: Which product categories perform best?

Business Interpretation

The **"Other"** category dominates overall sales, suggesting it contains high-performing products that may need sub-segmentation. **Kitchen**, **Stationery**, and **Storage** show consistent demand across months, while **Gifts** and **Seasonal** categories spike during promotional periods.

Key Insights

- "Other" category leads in total revenue.
- Kitchen and Storage maintain steady performance year-round.
- Gifts and Seasonal categories respond to Q4 promotions.



Strategic Recommendation

Maintain inventory for stable categories like Kitchen and Storage. Sub-segment "Other" to identify hidden product trends. Time promotions for Gifts and Seasonal categories around Q4.

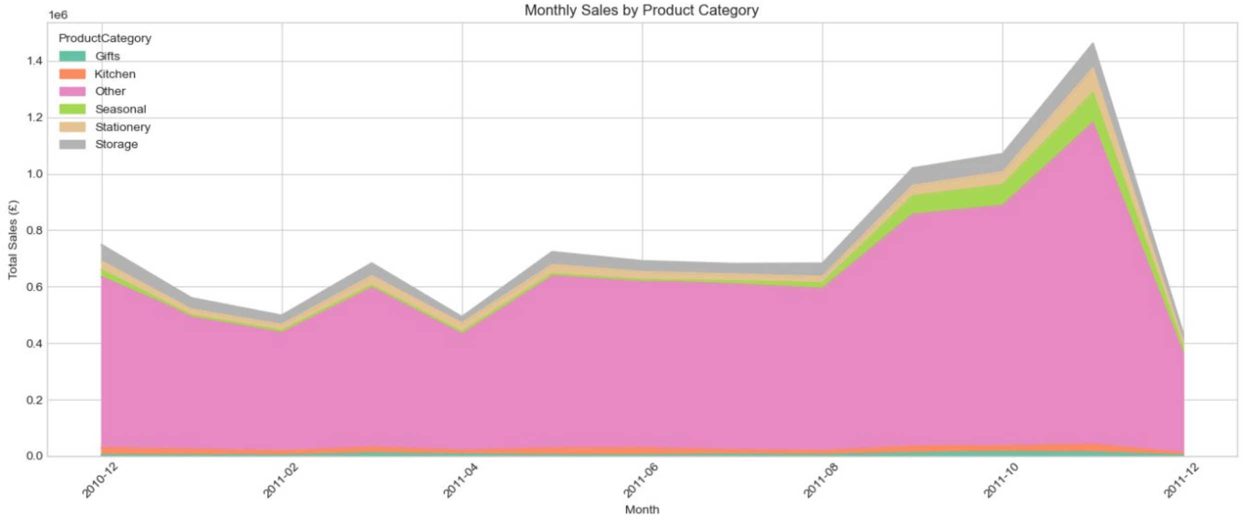
Q6: How do sales vary by time period?

Business Interpretation

Sales show clear seasonal variation, with a sharp peak in **November** driven by holiday demand. The **Gifts** and **Seasonal** categories expand dramatically during Q4, while **Kitchen**, **Storage**, and **Stationery** remain stable year-round. This confirms the need for seasonal inventory planning and promotional timing.

Key Insights

- November is the highest-grossing month.
- Q4 accounts for a disproportionate share of annual revenue.
- Seasonal categories respond strongly to promotional timing.



Strategic Recommendation

Launch seasonal campaigns in October and November. Stock Gifts and Seasonal items heavily in Q4. Maintain consistent inventory for stable categories like Kitchen and Storage.

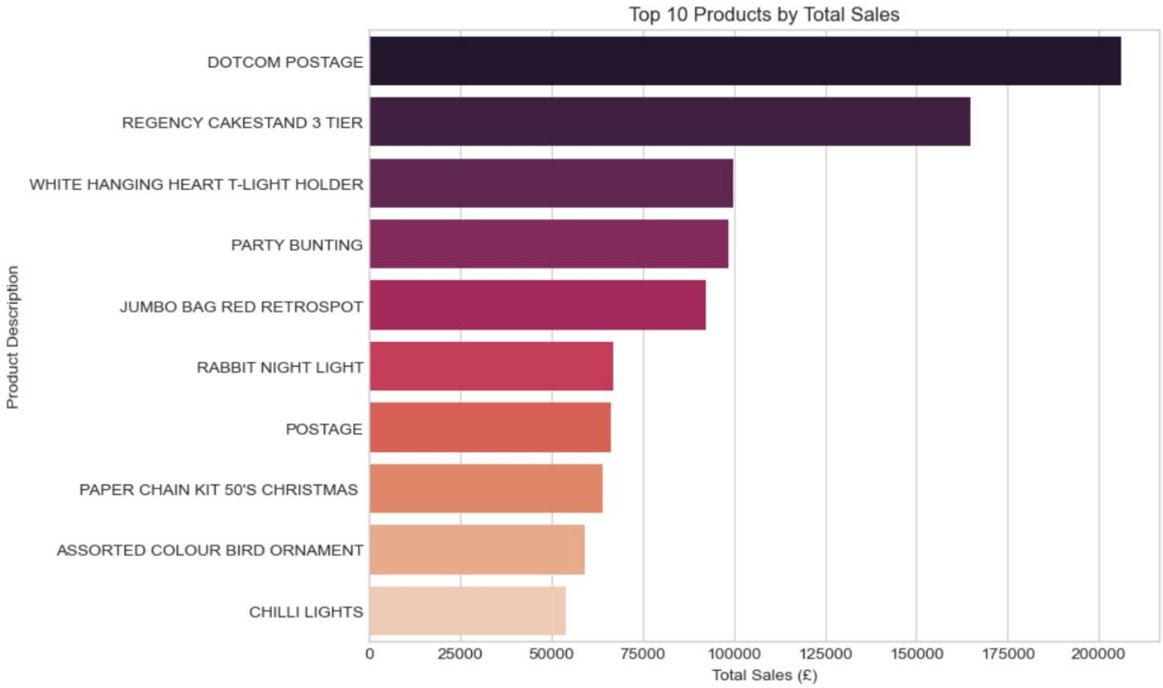
Q7: What are the top-selling products and categories?

Business Interpretation

The top-selling products include **DOTCOM POSTAGE**, **REGENCY CAKESTAND**, and **JUMBO BAG RED RETROSPOT**. These items consistently generate high revenue and appear across multiple invoices. The **“Other”** category leads in total sales, suggesting it contains high-performing products that may need further sub-segmentation.

Key Insights

- DOTCOM POSTAGE is the highest-grossing item, likely due to frequent shipping charges.
- Several top products are giftable or seasonal.
- The “Other” category dominates, but its contents are not clearly defined.



Strategic Recommendation

Sub-segment the “Other” category to identify hidden product trends. Promote top-selling items during seasonal campaigns. Consider bundling high-performing products for upsell opportunities.

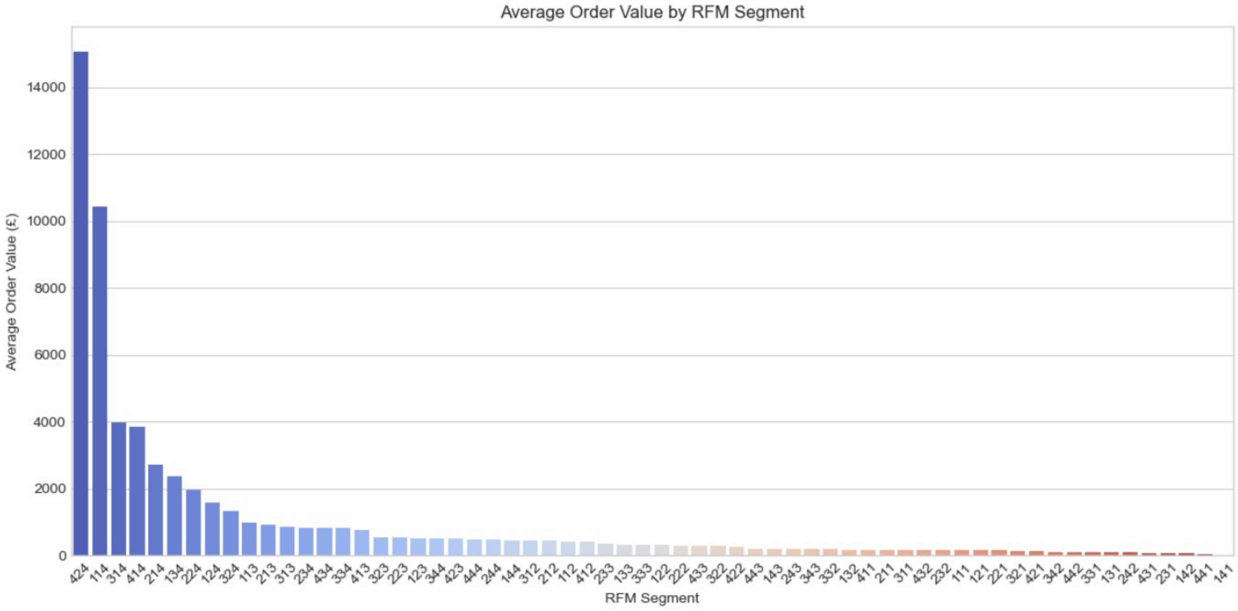
Q8: What is the average order value, and how does it vary?

Business Interpretation

Average Order Value (AOV) varies significantly across customer segments and regions. Customers in **Segment 444** and countries like the **Netherlands** and **Australia** consistently place higher-value orders. Mid-tier segments show moderate AOV, indicating upselling potential. Lower segments and regions show smaller transactions, often tied to single-item purchases.

Key Insights

- Segment 444 has the highest AOV, confirming premium behavior.
- Netherlands and Australia lead in regional AOV.
- Mid-tier segments offer growth opportunities with targeted promotions.



Strategic Recommendation

Upsell mid-tier segments with bundled offers and loyalty incentives. Tailor premium product campaigns to high-AOV regions. Monitor low-AOV segments for churn risk and reactivation opportunities.

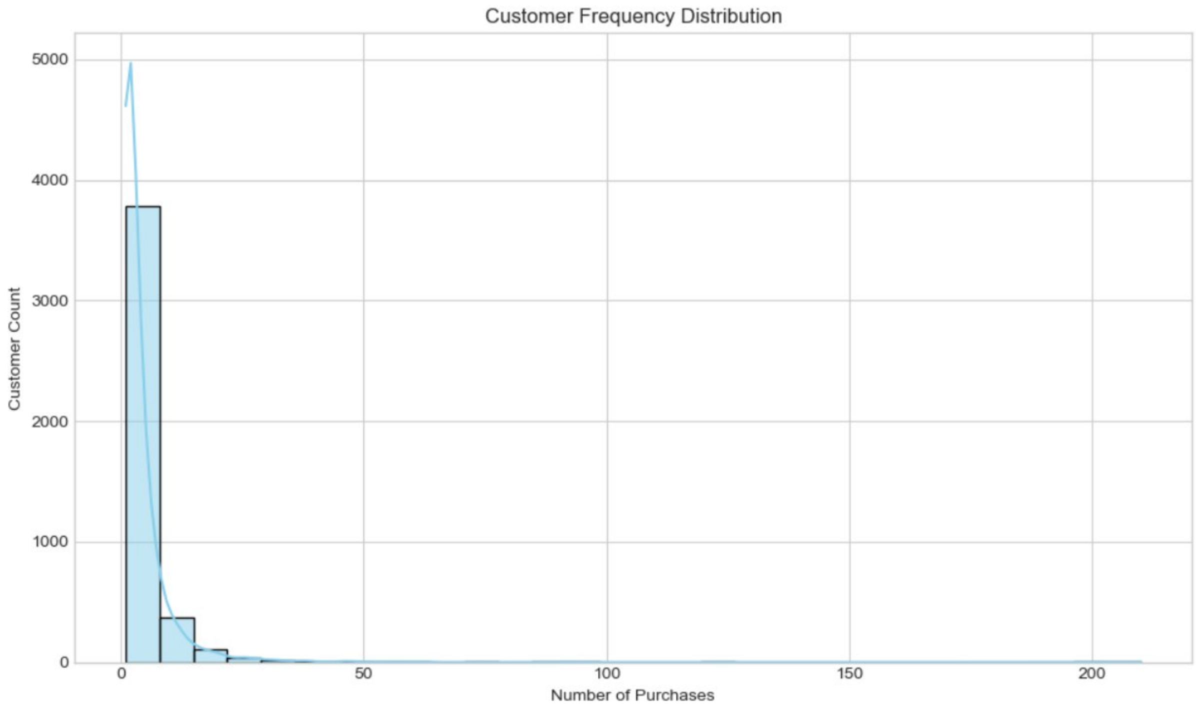
Q9: Are there loyalty or churn signals?

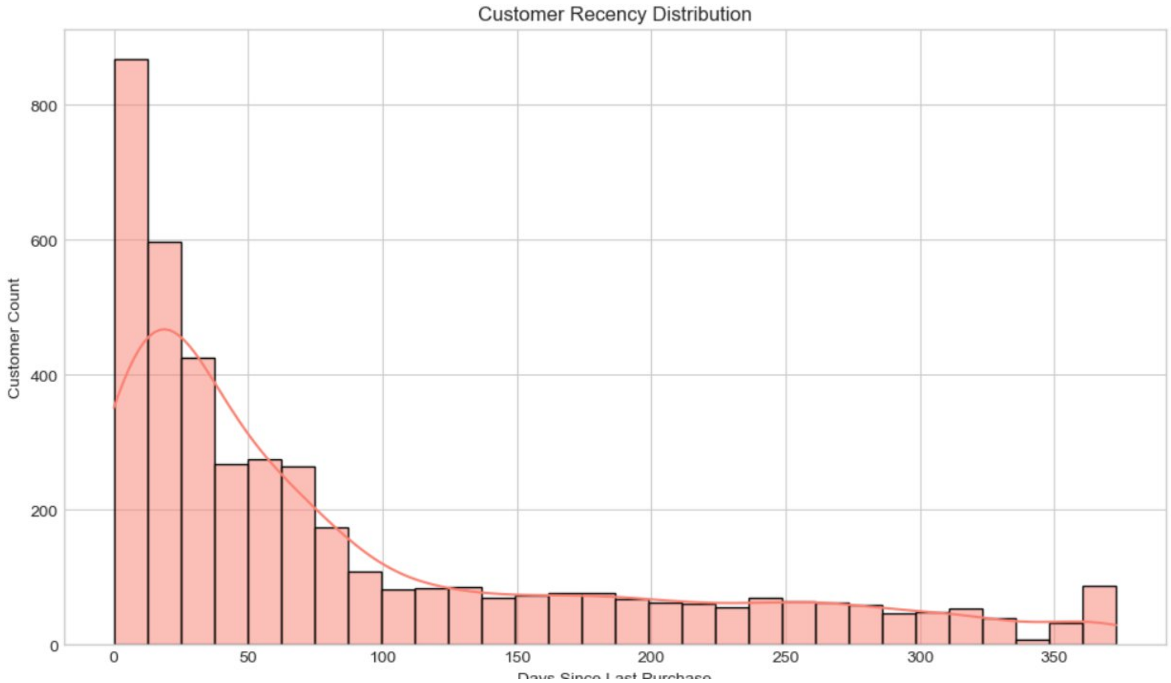
Business Interpretation

Customer behavior shows clear signals of loyalty and churn. Most customers purchase infrequently, with long gaps between transactions. However, loyal segments, especially Segment 444- exhibit high frequency and recent activity. Recency and frequency distributions reveal dormant customers who may be at risk of churn.

Key Insights

- The majority of customers have low frequency and long recency.
- Segment 444 shows strong loyalty with frequent, recent purchases.
- Dormant customers cluster in low-frequency, high-recency bins.





Strategic Recommendation

Re-engage dormant customers with win-back campaigns. Reward loyal customers with exclusive offers. Use recency and frequency thresholds to trigger automated retention workflows.

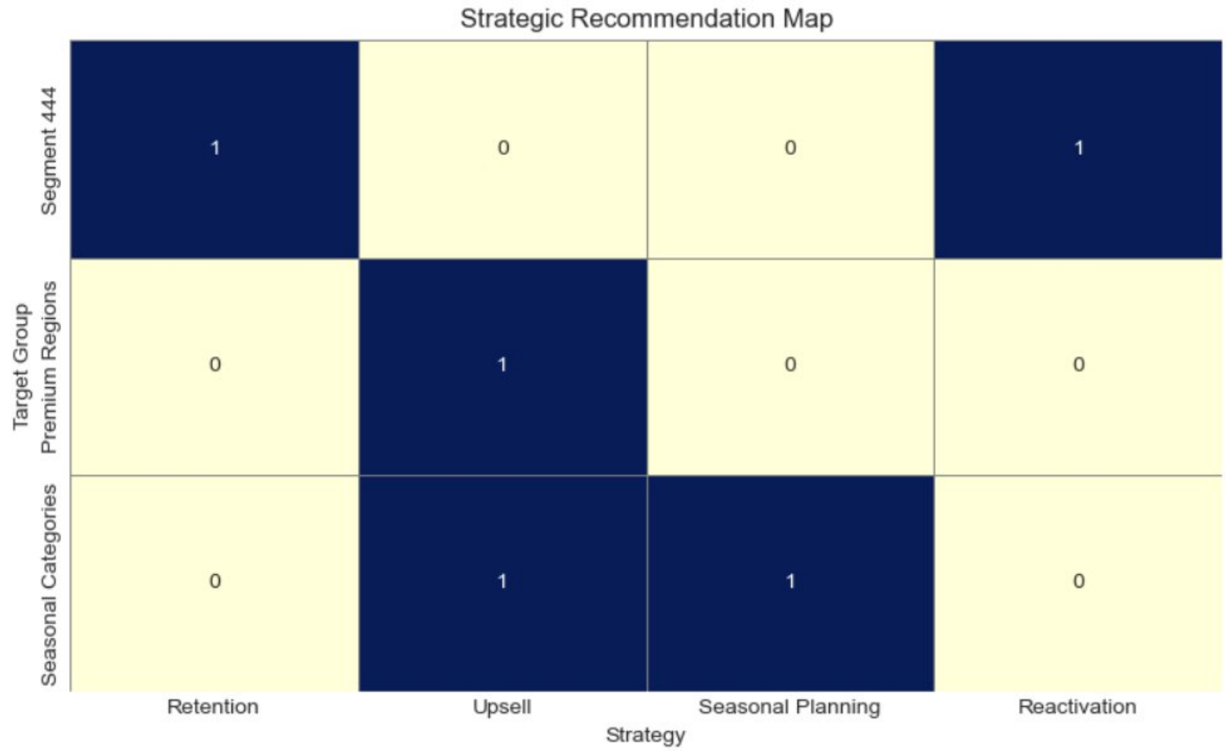
Q10: What recommendations can we make based on the data?

Business Interpretation

The analysis reveals clear opportunities for retention, upselling, and seasonal planning. Segment 444 customers are high-value and loyal. Q4 drives peak sales, especially in Gifts and Seasonal categories. Regions like Netherlands and Australia show premium behavior, while the UK leads in volume. Product trends highlight the dominance of “Other” and top-selling items like DOTCOM POSTAGE and REGENCY CAKESTAND.

Key Insights

- Segment 444 should be prioritized for retention and loyalty programs.
- Q4 promotions are critical for seasonal categories.
- Premium regions and products offer upsell potential.
- Dormant customers and low-AOV segments need reactivation strategies.



Strategic Recommendation

Bundle insights into a multi-pronged strategy:

- **Retention**: Loyalty campaigns for Segment 444 and UK customers.
- **Upsell**: Premium product bundles for the Netherlands and Australia.
- **Seasonal Planning**: Launch promotions in October–November for Gifts and Seasonal categories.
- **Reactivation**: Win-back campaigns for dormant segments with low frequency and high recency.

Strategic Recommendations

Customer Strategy

- Prioritize Segment 444 with personalized rewards and retention campaigns.
- Upsell mid-tier segments with targeted promotions and reactivation nudges.
- Improve checkout flow to reduce anonymous transactions and abandoned carts.

Product & Inventory Strategy

- Maintain steady inventory for Kitchen, Stationery, and Storage.
- Time campaigns around Q4 for Gifts and Seasonal categories.
- Sub-segment "Other" to uncover hidden product trends.

Regional Strategy

- Focus retention efforts on the UK.
- Tailor messaging for niche regions like the Netherlands and Australia.
- Adjust pricing or outreach in low-engagement regions.

Forecasting & Planning

- Use ARIMA forecasts to guide Q1 inventory and staffing.
- Validate models with RMSE and MAE.
- Apply forecasting across stable categories for precision planning.

References

American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). https://doi.org/10.1037/0000165-000

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering, 9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, 51–56. https://doi.org/10.25080/Majora-92bf1922-00a

Waskom, M. L. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software, 6*(60), 3021. https://doi.org/10.21105/joss.03021

Pandas Development Team. (2020). *pandas-dev/pandas: Pandas (Version 1.1.3)* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.3509134

Python Software Foundation. (2023). *Python (Version 3.10)* [Computer software]. https://www.python.org/

Appendix

A1. RFM Scoring Logic

- **Recency**: Number of days since a customer's last purchase, calculated from the most recent invoice date.
- **Frequency**: Count of unique invoices per customer.
- **Monetary**: Total purchase value per customer.
- **Scoring**: Each metric is scored into quartiles (1–4), with higher scores indicating more favorable behavior (e.g., 4 = most recent, most frequent, highest spend).
- **Segment Code**: Concatenated RFM scores (e.g., "444" represents top-tier customers across all dimensions).

A2. Python Libraries Used

- pandas: Data manipulation and aggregation
- matplotlib: Static plotting for visualizations
- seaborn: Statistical plotting with enhanced aesthetics

A3. Visualization Code Index

Each business question includes a corresponding Python visualization:

- **Q1**: RFM Segment Distribution
- **Q2**: Average Monetary Value by Segment
- **Q3**: Monthly Sales Trend
- **Q4**: Total Sales and AOV by Country
- **Q5**: Total Sales by Product Category
- **Q6**: Monthly Category Sales (Stacked Area Chart)
- **Q7**: Top 10 Products by Sales
- **Q8**: AOV by RFM Segment
- **Q9**: Frequency and Recency Distributions with KDE
- **Q10**: Strategic Recommendation Heatmap

A4. Data Cleaning Notes

- Removed rows with missing CustomerID
- Filtered out negative Quantity values
- Converted InvoiceDate to datetime format

- Created MonthYear and Season columns for time-based analysis
- Calculated TotalPrice as Quantity × UnitPrice
- Removed canceled transactions (invoices starting with “C”)

Submission Notes

All project files are available at https://github.com/lhansen77/DSC450_Sprint2, including the APA report, Jupyter notebook, and supporting visuals.