Customer Segmentation and Basket Analysis

Insights from Online Retail Dataset

BY: Ruslan Maystrenko

Agenda

01 Overview of Online Retail Dataset

02 Exploratory Data Analysis findings

03 Customer Segmentation

04 Market Basket Analysis

05 Recommendations and Impact

Dataset Overview

Understanding Our Data: Online Retail Transactions

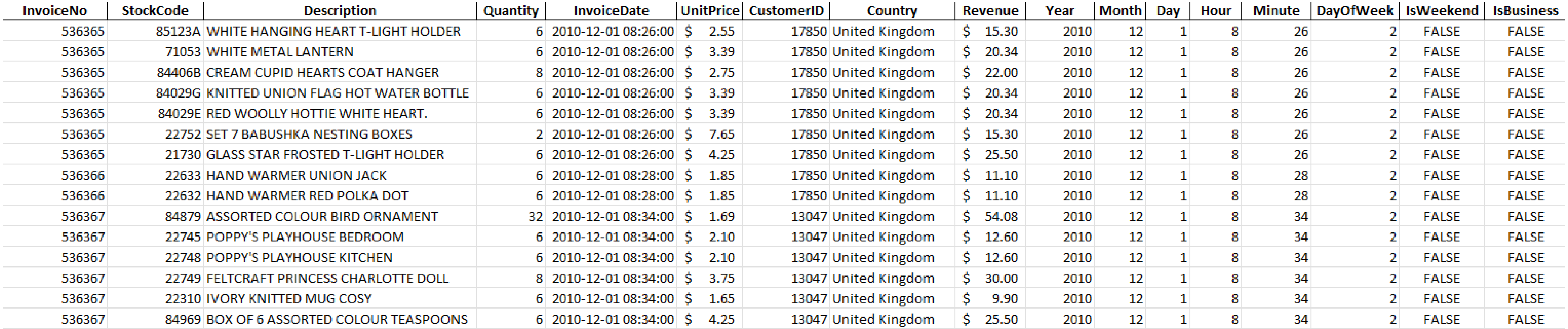
Source: UCI Machine Learning Repository (541,909 transactions, Dec 2010–Dec 2011)

njhg+ nhjn

Original Attributes: InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice, CustomerID, Country

Feature Engineered Columns:

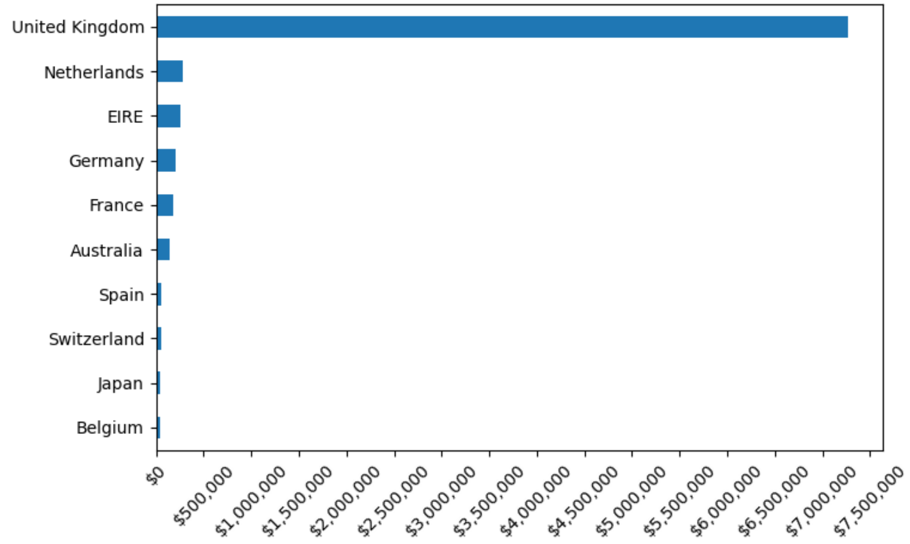
- *Revenue:* Calculated as UnitPrice * Quantity for total transaction value
- *Year, Month, Day, Hour, Minute:* Extracted from InvoiceDate for temporal analysis
- *DayOfWeek:* Day of the week for purchase pattern analysis
- *IsWeekend:* Boolean indicating weekend purchases (Saturday/Sunday)
- *IsBusiness:* Boolean identifying high spend customers (revenue > $30K) as businesses



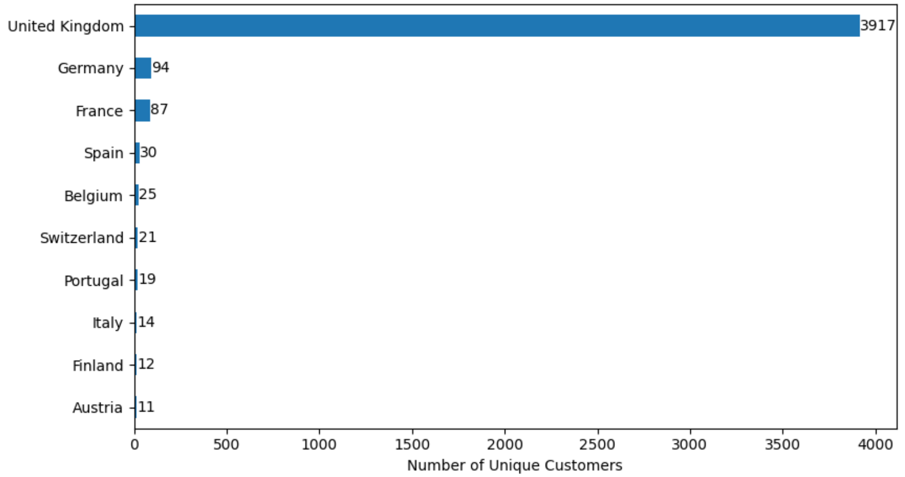
InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country	Revenue	Year	Month	Day	Hour	Minute	DayOfWeek	IsWeekend	IsBusiness
536365	85123A	WHITE HANGING HEART T-LIGHT HOLDER	6	2010-12-01 08:26:00	$ 2.55	17850	United Kingdom	$ 15.30	2010	12	1	8	26	2	FALSE	FALSE
536365	71053	WHITE METAL LANTERN	6	2010-12-01 08:26:00	$ 3.39	17850	United Kingdom	$ 20.34	2010	12	1	8	26	2	FALSE	FALSE
536365	84406B	CREAM CUPID HEARTS COAT HANGER	8	2010-12-01 08:26:00	$ 2.75	17850	United Kingdom	$ 22.00	2010	12	1	8	26	2	FALSE	FALSE
536365	84029G	KNITTED UNION FLAG HOT WATER BOTTLE	6	2010-12-01 08:26:00	$ 3.39	17850	United Kingdom	$ 20.34	2010	12	1	8	26	2	FALSE	FALSE
536365	84029E	RED WOOLLY HOTTIE WHITE HEART.	6	2010-12-01 08:26:00	$ 3.39	17850	United Kingdom	$ 20.34	2010	12	1	8	26	2	FALSE	FALSE
536365	22752	SET 7 BABUSHKA NESTING BOXES	2	2010-12-01 08:26:00	$ 7.65	17850	United Kingdom	$ 15.30	2010	12	1	8	26	2	FALSE	FALSE
536365	21730	GLASS STAR FROSTED T-LIGHT HOLDER	6	2010-12-01 08:26:00	$ 4.25	17850	United Kingdom	$ 25.50	2010	12	1	8	26	2	FALSE	FALSE
536366	22633	HAND WARMER UNION JACK	6	2010-12-01 08:28:00	$ 1.85	17850	United Kingdom	$ 11.10	2010	12	1	8	28	2	FALSE	FALSE
536366	22632	HAND WARMER RED POLKA DOT	6	2010-12-01 08:28:00	$ 1.85	17850	United Kingdom	$ 11.10	2010	12	1	8	28	2	FALSE	FALSE
536367	84879	ASSORTED COLOUR BIRD ORNAMENT	32	2010-12-01 08:34:00	$ 1.69	13047	United Kingdom	$ 54.08	2010	12	1	8	34	2	FALSE	FALSE
536367	22745	POPPY'S PLAYHOUSE BEDROOM	6	2010-12-01 08:34:00	$ 2.10	13047	United Kingdom	$ 12.60	2010	12	1	8	34	2	FALSE	FALSE
536367	22748	POPPY'S PLAYHOUSE KITCHEN	6	2010-12-01 08:34:00	$ 2.10	13047	United Kingdom	$ 12.60	2010	12	1	8	34	2	FALSE	FALSE
536367	22749	FELTCRAFT PRINCESS CHARLOTTE DOLL	8	2010-12-01 08:34:00	$ 3.75	13047	United Kingdom	$ 30.00	2010	12	1	8	34	2	FALSE	FALSE
536367	22310	IVORY KNITTED MUG COSY	6	2010-12-01 08:34:00	$ 1.65	13047	United Kingdom	$ 9.90	2010	12	1	8	34	2	FALSE	FALSE
536367	84969	BOX OF 6 ASSORTED COLOUR TEASPOONS	6	2010-12-01 08:34:00	$ 4.25	13047	United Kingdom	$ 25.50	2010	12	1	8	34	2	FALSE	FALSE

Exploratory Data Analysis

Revenue per Country



Customer Count per Country



Insights into what drives the revenue for Netherlands and EIRE as customer count is exceedingly low

Netherland *(9 Customers)*

CustomerID	Revenue
14646	**$279,138**
12759	$1,411
12775	$1,281

EIRE *(3 Customers)*

CustomerID	Revenue
14911	**$136,275**
14156	**$116,729**
14016	$4,291

Germany *(94 Customers)*

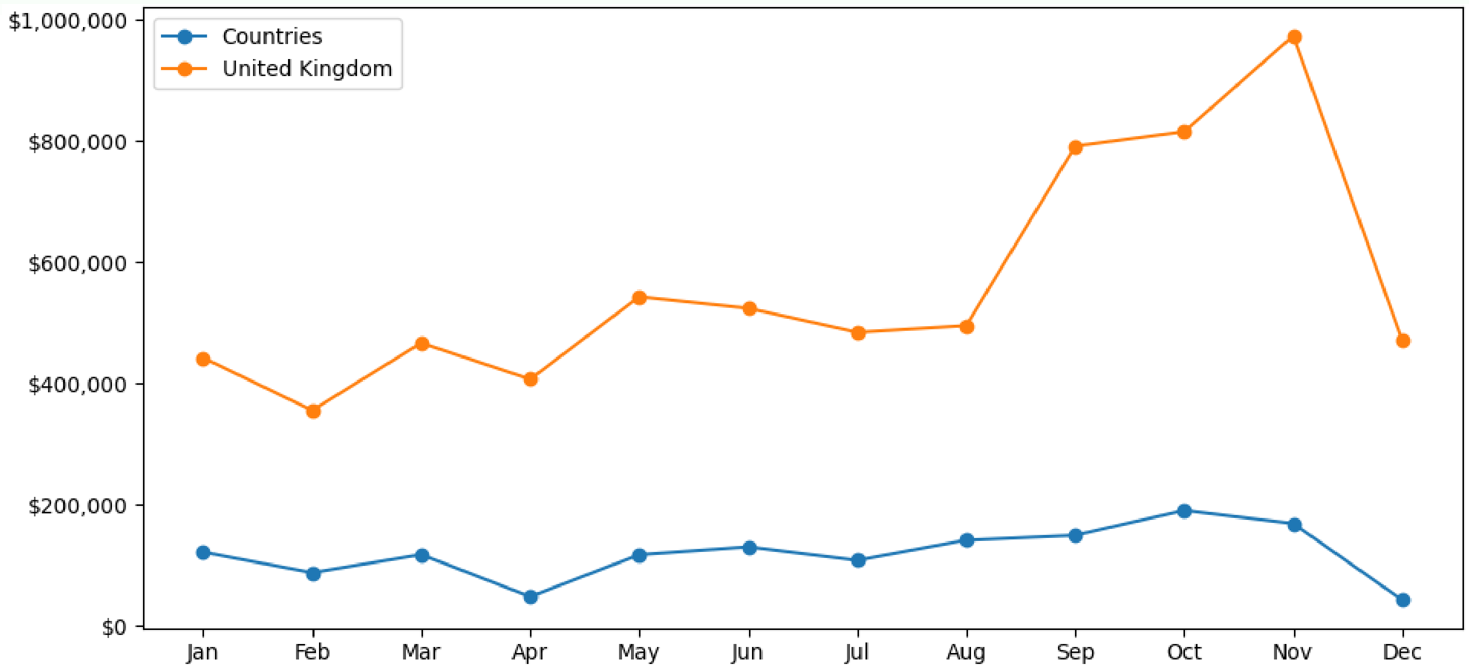
CustomerID	Revenue
12471	$17,424
12477	$13,219
12621	$12,411

EDA Insights

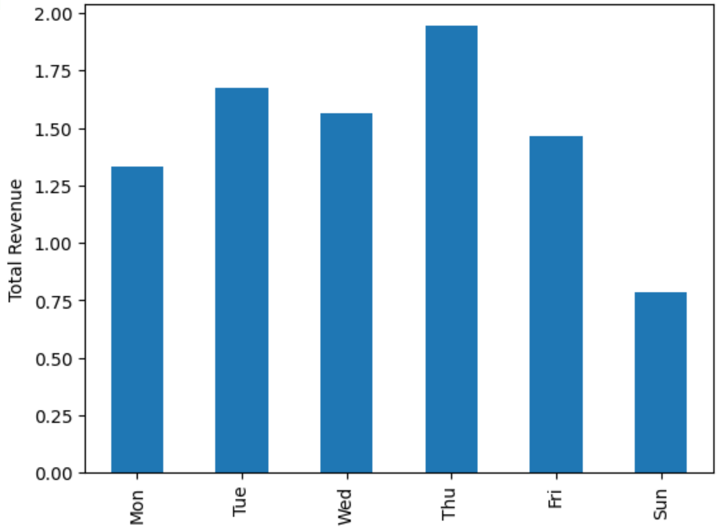
- We see that the UK is heavily **dominated** in this dataset exceeding their **revenue** and **customer count** compared to the rest of the countries
- Further analysis into why Netherlands and EIRE appear as 2nd and 3rd ranked in revenue although don't appear in top customer counts
  - Netherlands is dominated by one customer: 14646, that makes up **98.33% of their revenue**
  - EIRE contains two customers out of 3 also making up **98% of its revenue**

Exploratory Data Analysis

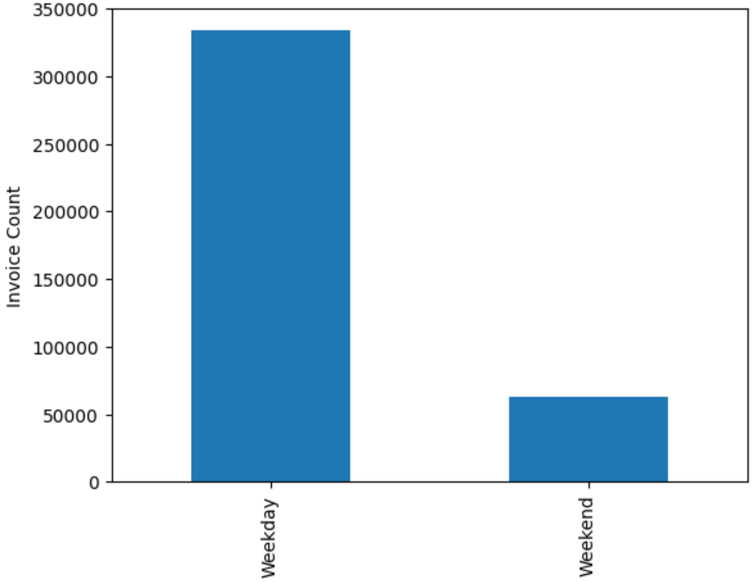
Monthly Revenue



Day of Week Revenue



Weekday vs Weekend

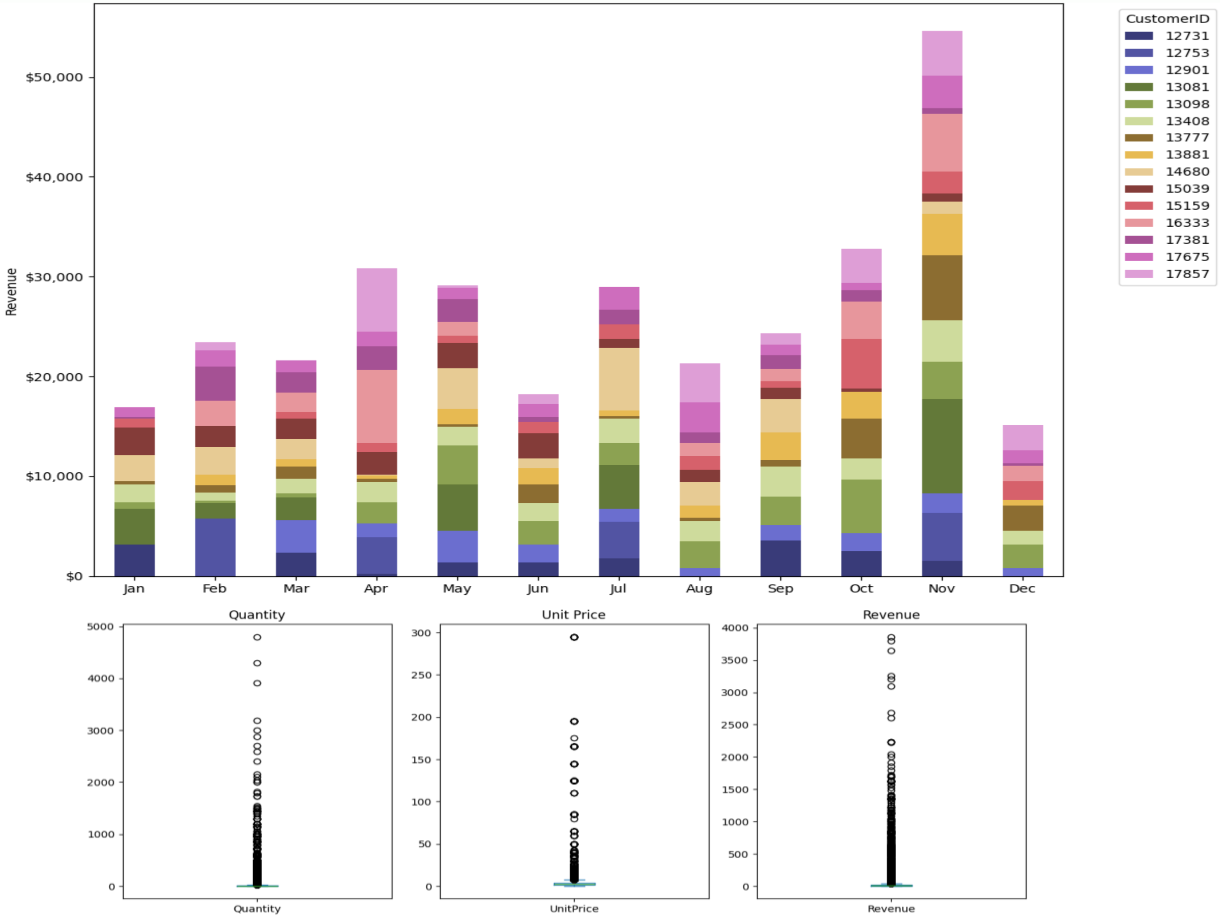


Revenue Insights

- The UK **exceeds** all other countries in **monthly revenue**; we even see seasonal trends during the winter holiday season
- Looking at the daily revenue, we see **no transactions** on **Saturdays** which lead to a large disparity between weekday and weekend revenue

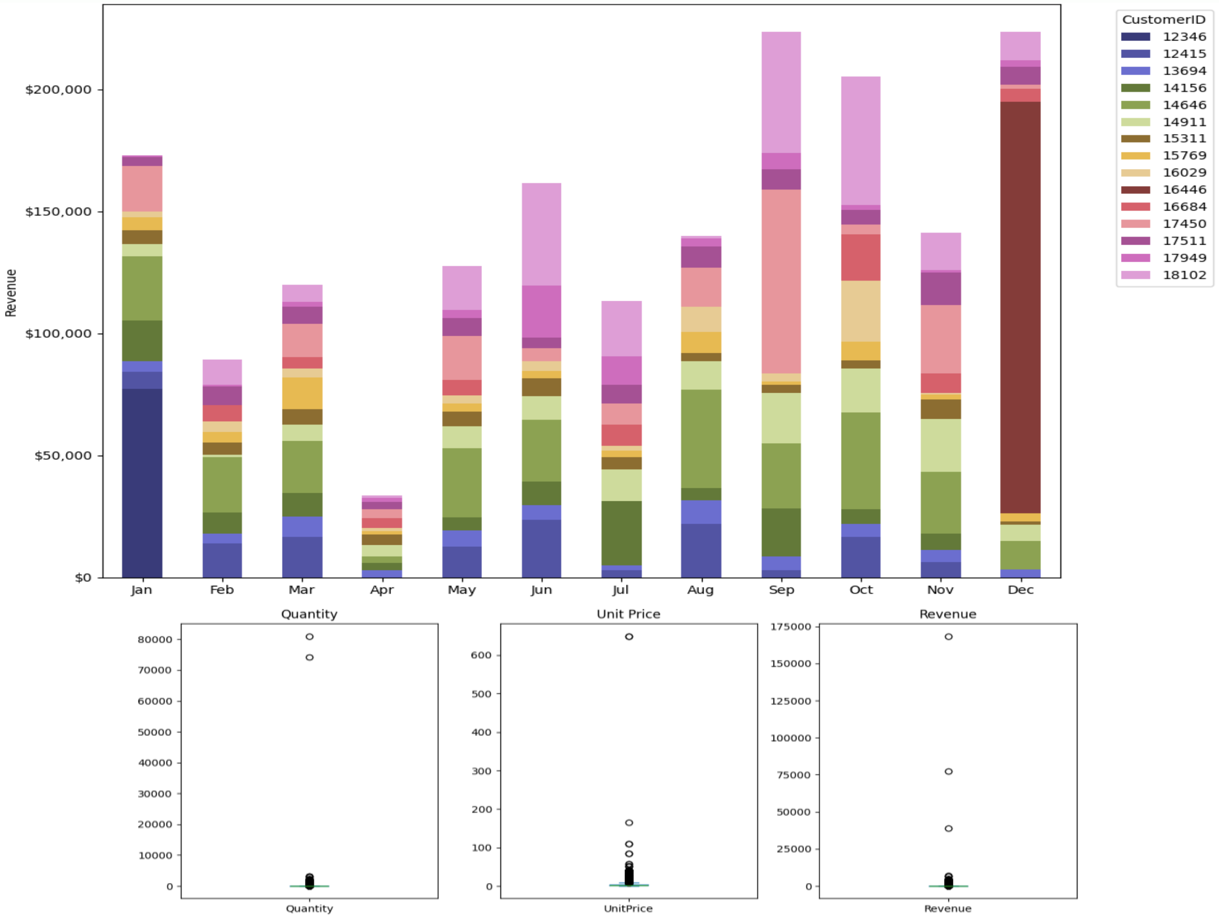
Exploratory Data Analysis

Non-Business Customers



For non-business customers, we observe a **uniformly** distributed revenue pattern with no discernible outliers

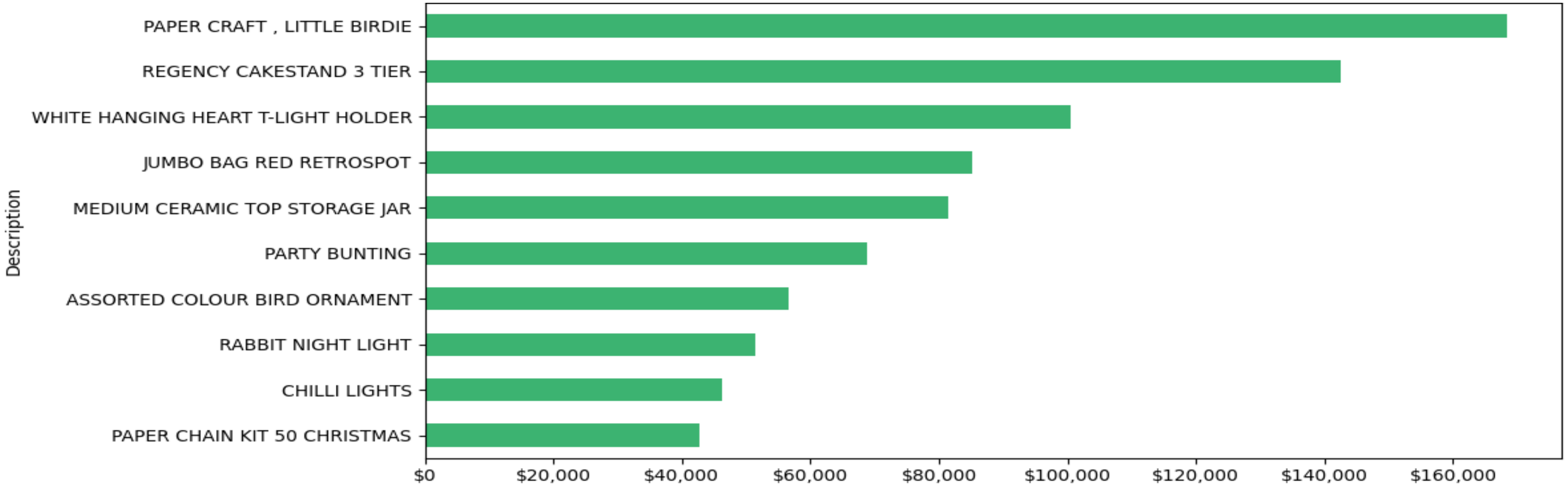
Non-Business Customers



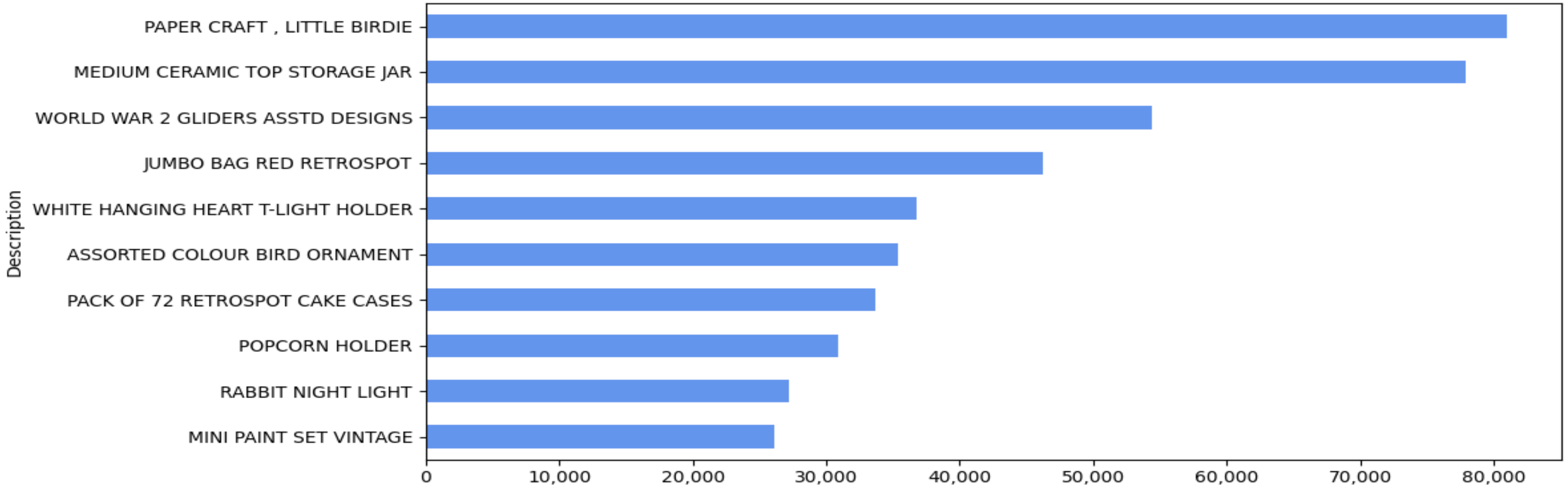
For business customers, the box plot **reveals outliers** (CustomerID: 16446 and 12346), which correspond to businesses placing bulk orders at the end/beginning of the year, likely driven by seasonal demand, holiday shopping, or tax related reasons

Exploratory Data Analysis

Top 10 Products by Revenue



Top 10 Products by Quantity



Product Insights

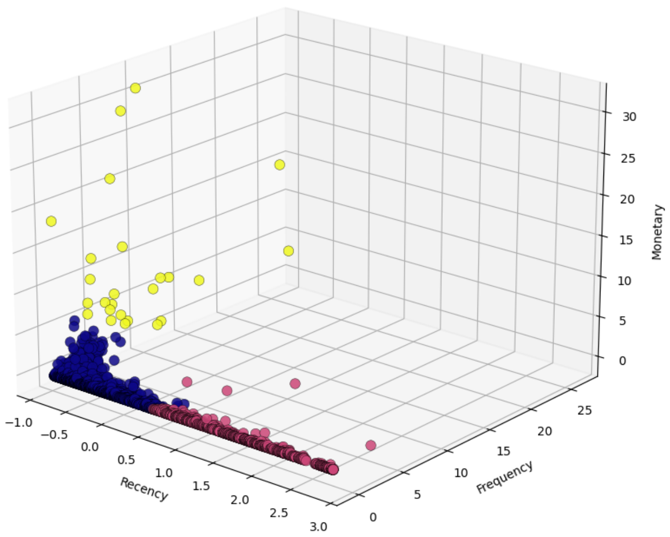
- The product **"PAPER CRAFT , LITTLE BIRDIE"** dominates both revenue and quantity sold
- Other products like **"REGENCY CAKESTAND 3 TIER"** appear in the revenue chart but not in quantity, suggesting a high price point despite fewer units sold
- **"WORLD WAR 2 GLIDERS ASSTD DESIGNS"** shows up in the quantity chart but not revenue, implying it's a lower priced, high-volume product

Customer Segmentation

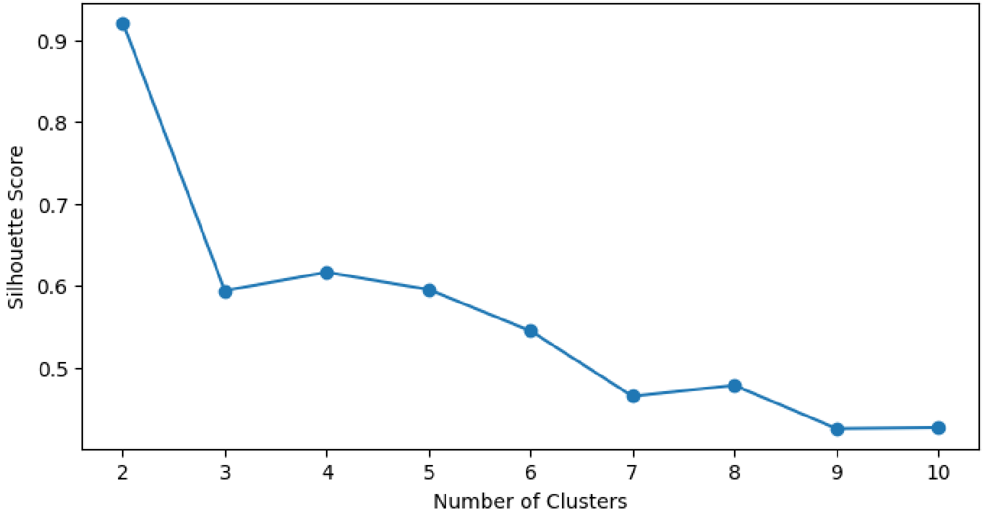
Segmenting Customers for Targeted Marketing

- **Methodology:** Use K-Means clustering on standardized RFM scores, with number of clusters determined by Elbow Method and Silhouette Scores
- **RFM Analysis:** Measures **Recency** (days since last purchase), **Frequency** (transaction count), **Monetary** (total Revenue)
  - *Why RFM?*: Proven framework **to identify high-value customers** and prioritize **marketing efforts**
- **Determining K value:**
  - **Elbow Method:** Indicates that **3 clusters** offers the best trade off between compactness and interpretability
  - **Silhouette Score:** Indicates a peak at 2 clusters, likely due to overfitting, with an improvement observed at 4 clusters; however, the score stabilizes and remains relatively consistent from 3 to 5 clusters, suggesting that **3 clusters** is the optimal choice between balance and simplicity

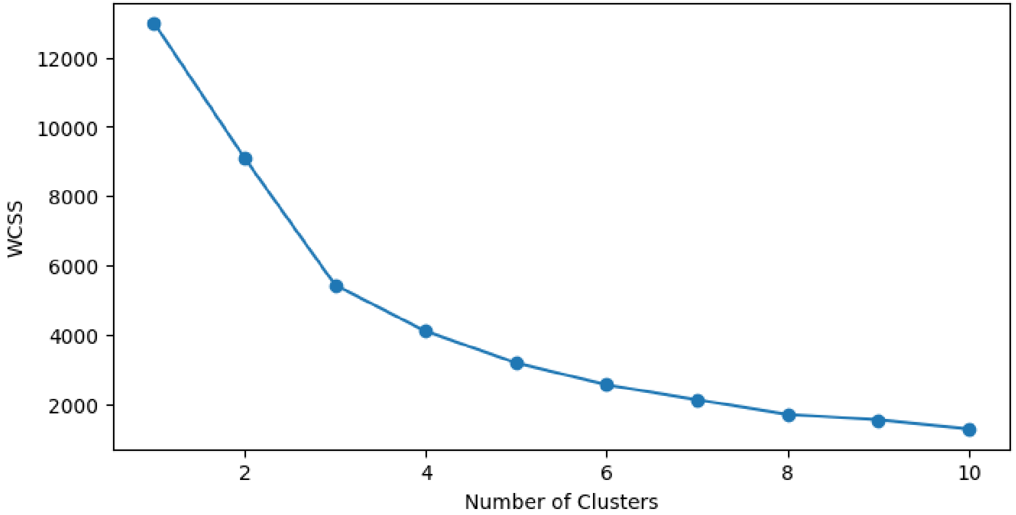
3D K-Means Clustering



Silhouette Score

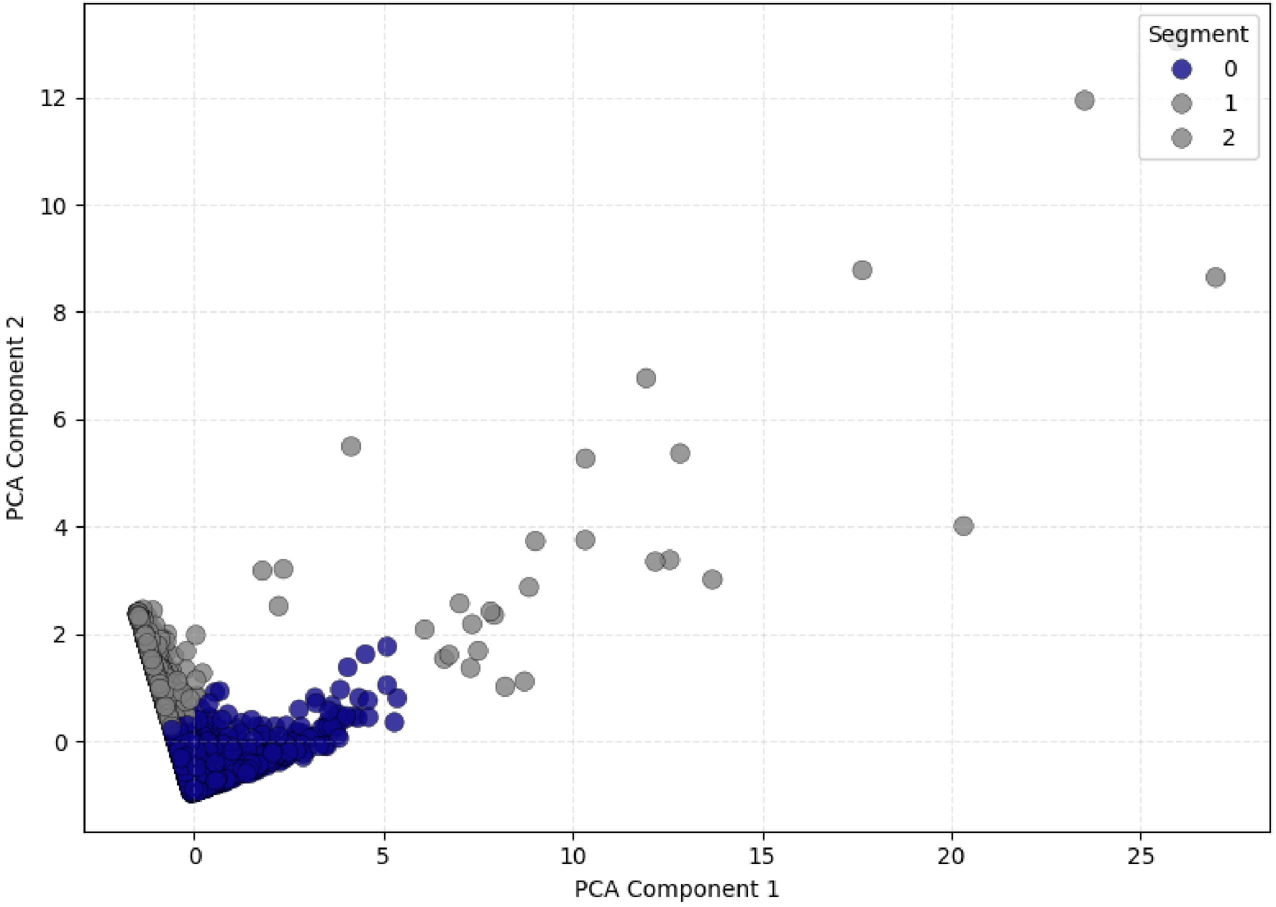


Elbow Method



Customer Segmentation

Segment 0 (Mid-Value, Active Customers)



Segment 0 (3,226 customers)

Recency	Frequency	Monetary
41 days	4.65 times	$1,839

Characteristics

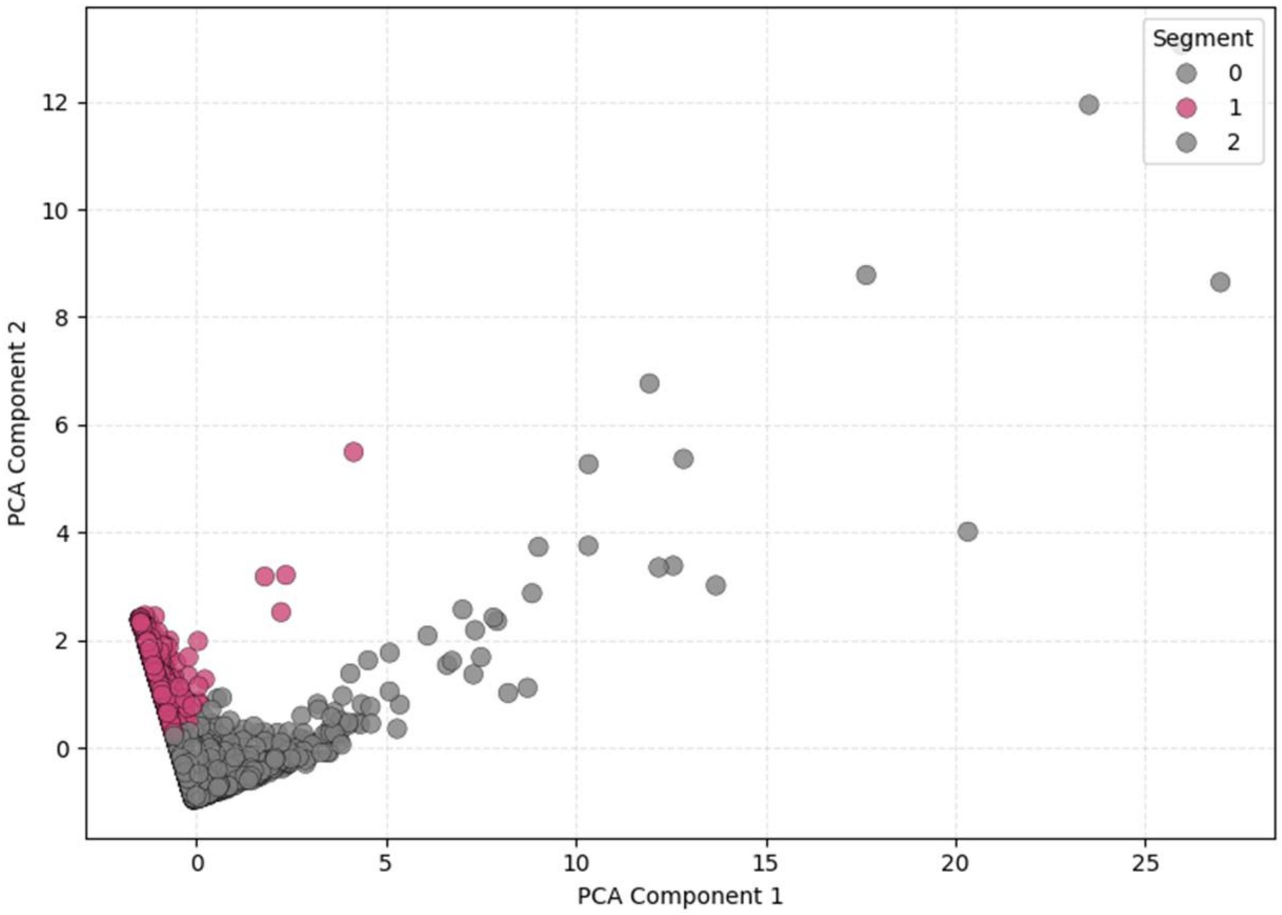
- These customers are **active** and **engaged**, purchasing regularly but not at the highest frequency or spend levels
- They represent a **broad**, **stable** customer base with potential for increased spending through upselling or cross-selling

Marketing Strategies

- Promote higher value products with **targeted discounts** via email to increase average order value (AOV)
- Implement a **loyalty program** with rewards for repeat purchases to boost frequency
- Launch **holiday campaigns** with special offers on popular items via digital ads to maximize sales during peak seasons

Customer Segmentation

Segment 1 (Low-Value, Inactive Customers)



Segment 1 (1,084 customers)

Recency	Frequency	Monetary
247 days	1.57 times	$626

Characteristics

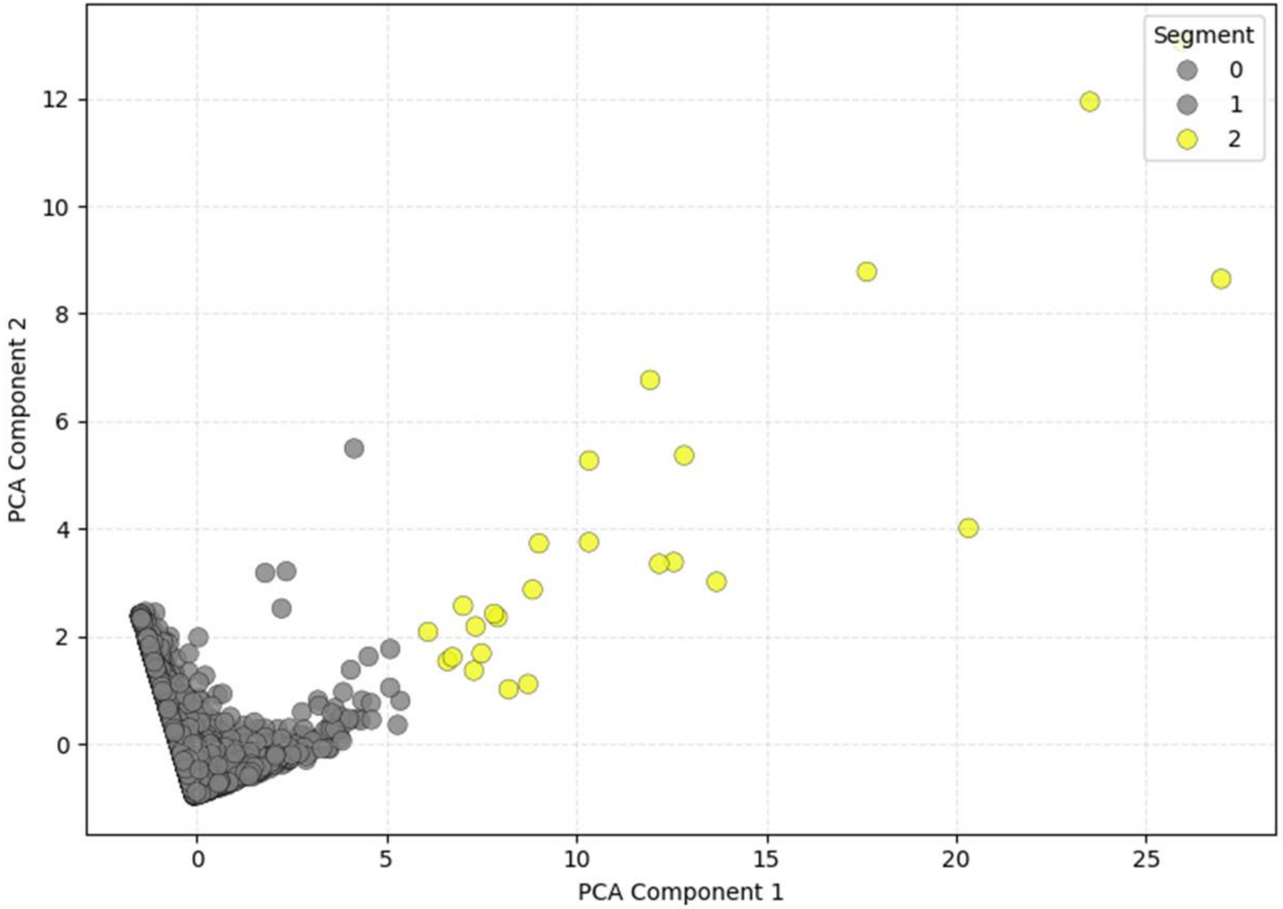
- These customers are largely **disengaged**, with purchases occurring rarely and long ago
- **Re-engaging these customers** could recover lost revenue and reduce churn, potentially converting them into mid-value customers

Marketing Strategies

- Send personalized **reactivation offers** via email to encourage return purchases
- Provide low-cost incentives like **discounts** via retargeting ads to stimulate repeat purchases
- **Collect feedback** through surveys with small incentives to understand disengagement

Customer Segmentation

Segment 2 (High-Value, Loyal Customers)



Segment 2 (25 customers)

Recency	Frequency	Monetary
6 days	67.88 times	$85,942

Characteristics

- These are the **most loyal** and **profitable** customers, purchasing frequently and spending significantly, often in bulk
- Despite their small size, this segment drives a **disproportionate share of revenue** due to high spend and frequency, therefore retaining these customers is critical

Marketing Strategies

- Offer a **premium loyalty program** with exclusive benefits to retain loyalty
- Provide **personalized product recommendations** via email to encourage larger orders
- Engage with dedicated **account support** for seasonal **bulk orders** to secure high value transactions

Market Basket Analysis

Segment 0

Antecedents	Consequents	Confidence	Lift
PINK REGENCY TEACUP AND SAUCER	ROSES & GREEN REGENCY TEACUP AND SAUCER	0.75	18.81

Strong link to coordinated teacup sets, ideal for cross-selling to boost mid-value customer revenue

Segment 1

Antecedents	Consequents	Confidence	Lift
GREEN REGENCY TEACUP AND SAUCER	ROSES & PINK REGENCY TEACUP AND SAUCER	0.88	25

Similar teacup pairing trend as Segment 0 offering reactivation opportunities for inactive customers

Segment 2

Antecedents	Consequents	Confidence	Lift
HERB MARKER ROSEMARY	HERB MARKER PARSLEY	0.81	68

Unique herb marker association, reflecting niche business purchases for Segment 2

Basket Mix Insights

Rules for Segments 0 and 1 are similar focusing on teacup sets, suggesting a unified consumer trend that can be leveraged for broad **cross-selling campaigns**. While Segment 2 differs with herb markers, highlighting a niche business market offering a **targeted opportunity** to secure high-value bulk orders

Business Summary

Key Findings:

Customer Segmentation: Identified Mid-Value Active, Low-Value Inactive, and High-Value Loyal customers using RFM analysis, revealing diverse engagement and spending patterns for targeted strategies

Market Basket Insights: Teacup set preferences in Mid-Value and Inactive segments support cross-selling, while High-Value's herb marker rule indicates niche business demand.

Business Impact:

- **Upselling** Mid-Value and **reactivating** Inactive customers can enhance revenue through tailored campaigns
- **Improved marketing efficiency** by aligning efforts with segment specific behaviors, reducing wasted ad spend
- Strengthened retention with **personalized approaches**, addressing churn across all segments

Next Steps:

- Launching **new marketing campaigns** for each segment
- **Monitor KPI's** (revenue, retention rate, order value) post-campaign to evaluate success
- Explore **real-time analytics** for continuous improvement