Retail Data Analysis

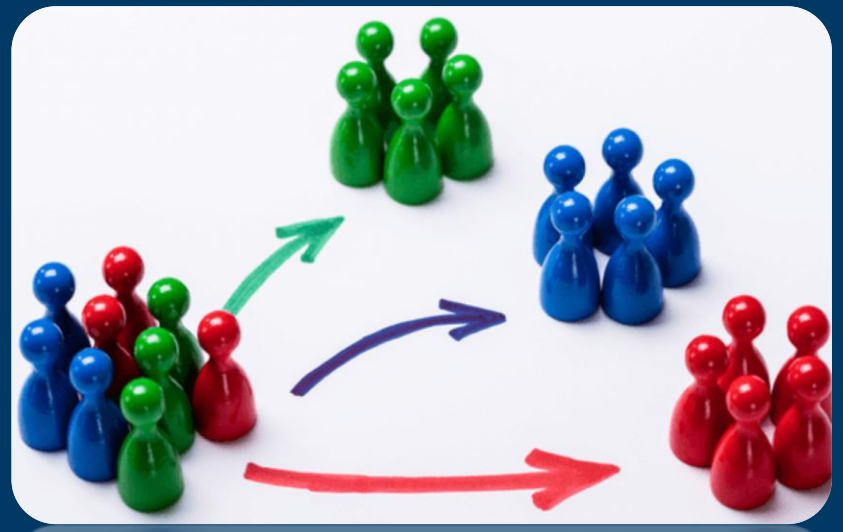
Customer Segmentation



Pamela Gatica
Marketing Analytics

October 2025

Introduction



- The retail industry generates massive amounts of transactional data every day.
- Through data analytics, companies can identify their most valuable customers and predict future sales trends.
- In this project, we apply **RFM (Recency, Frequency, Monetary)** analysis and machine learning techniques to uncover meaningful insights from retail data.

Data Overview

The dataset contains 541,909 rows and 8 columns (pre-cleaning) → Each row represent a unique product purchase within an invoice.

Total Unique Customers: 4372 / Total Unique Products: 3684

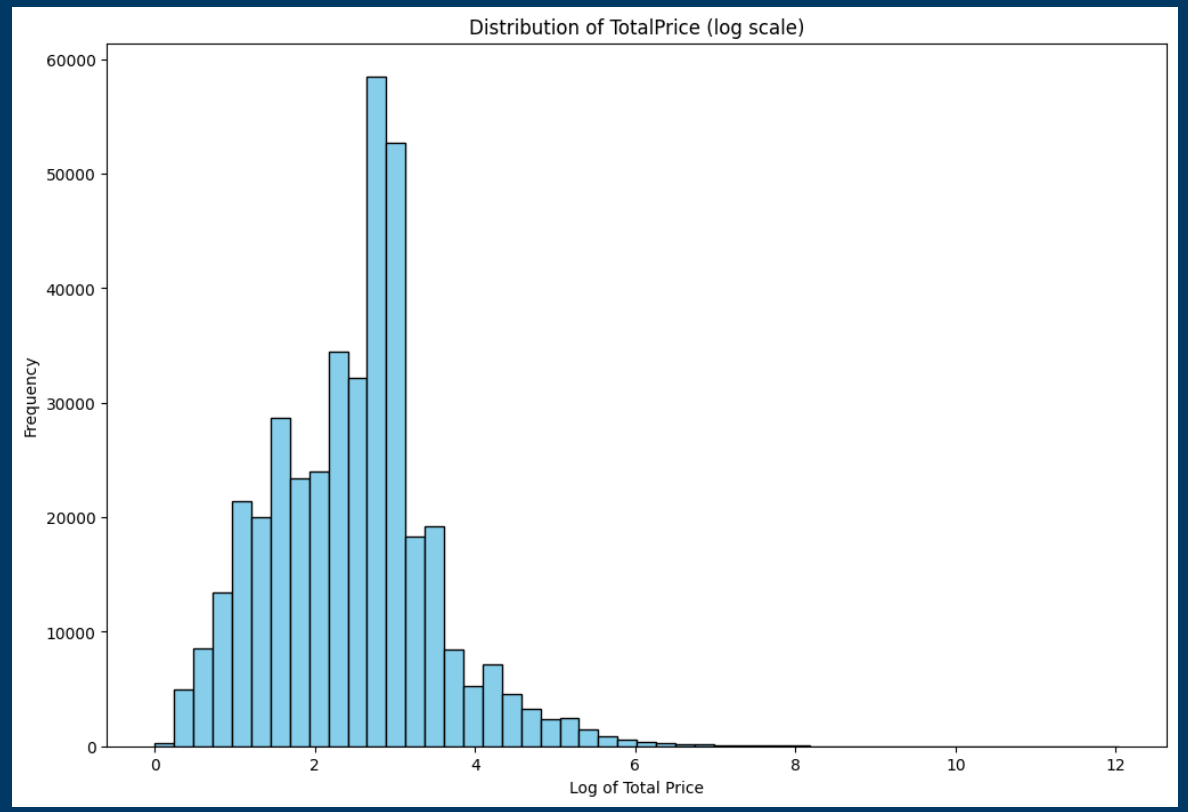
Covers 1 year of data → From December 2010 to December 2011.

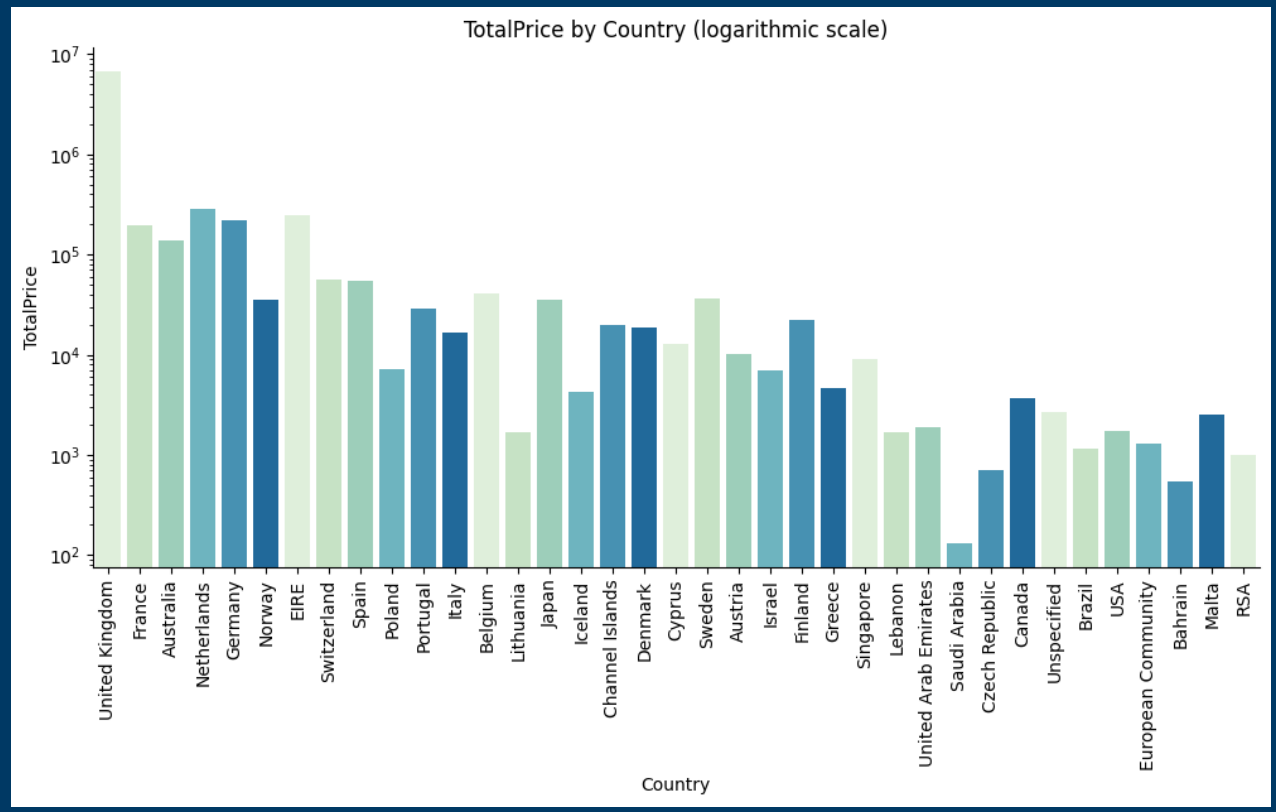
Includes transactions from 35 different countries.

Contains missing CustomerIDs from anonymous shoppers → Data cleaning was performed to remove returns (negative quantities) and missing values.



Exploratory Data Analysis (EDA)

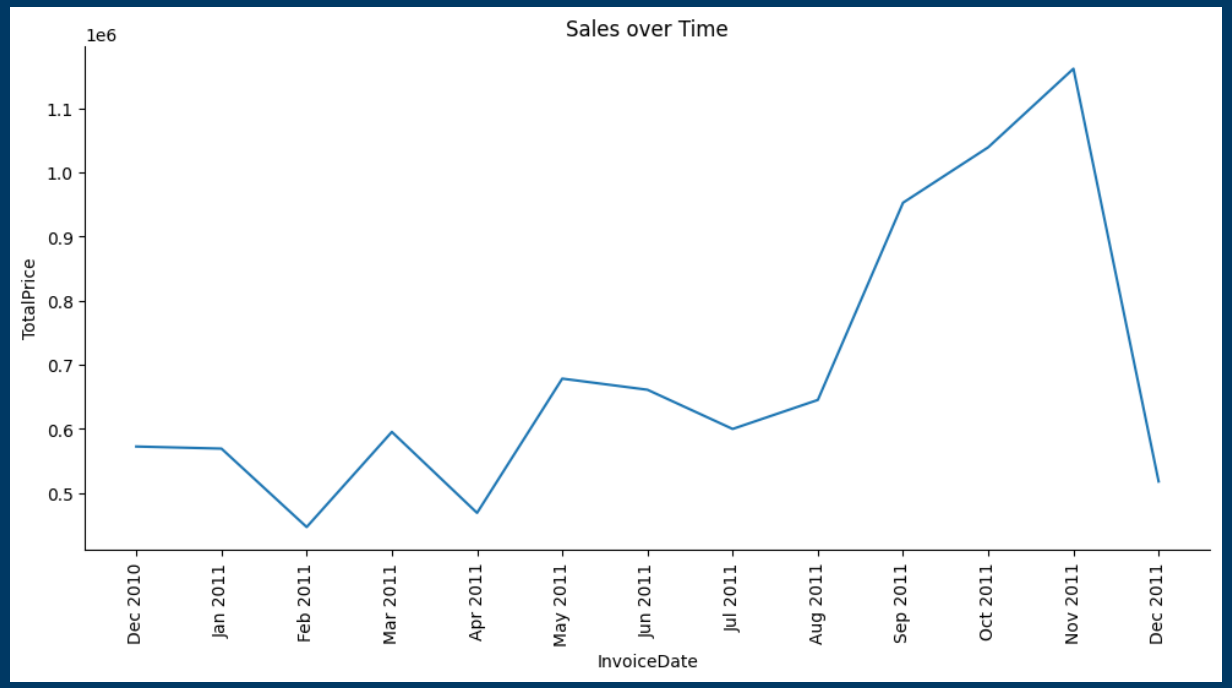




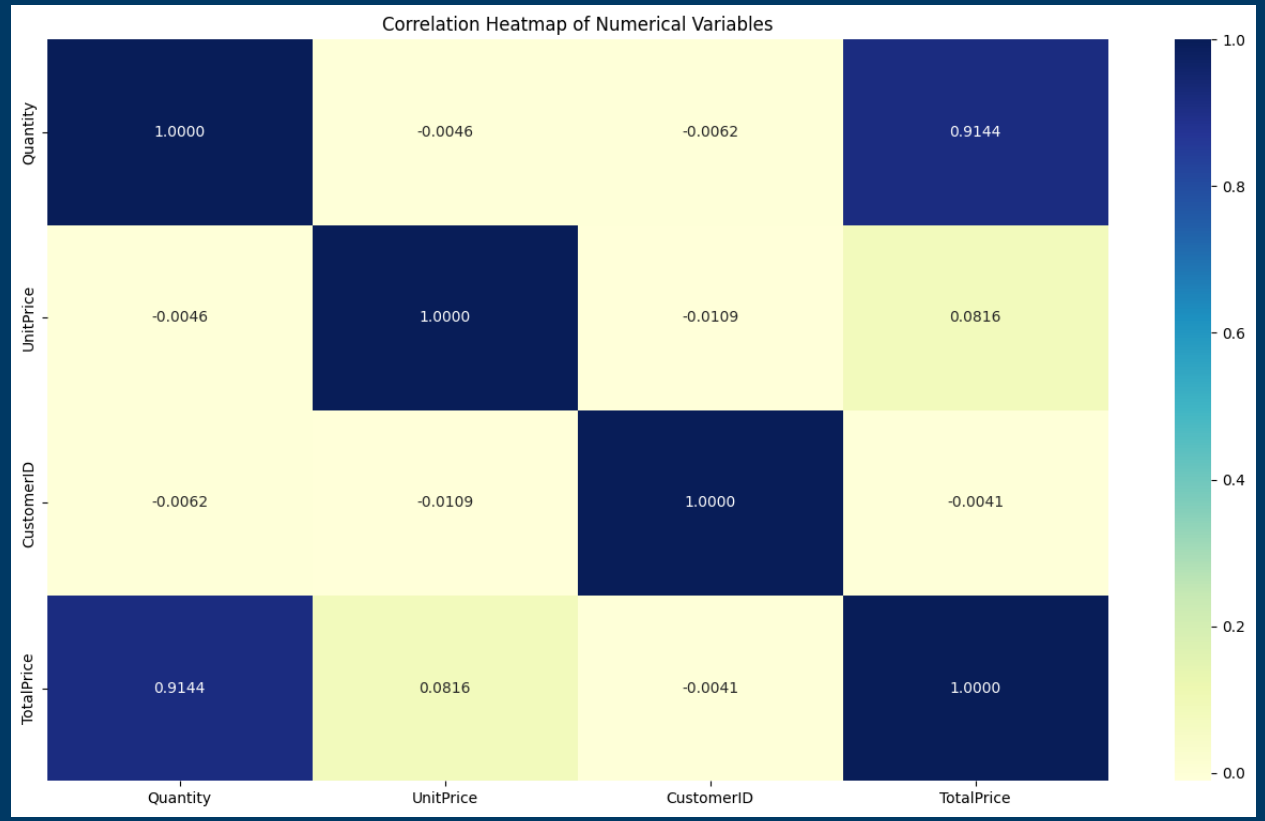
The majority of transactions are concentrated in the **lower range**, meaning most purchases are of **small total amounts**.

The United Kingdom generates the majority of sales, while other countries contribute relatively small amounts.

Exploratory Data Analysis (EDA)

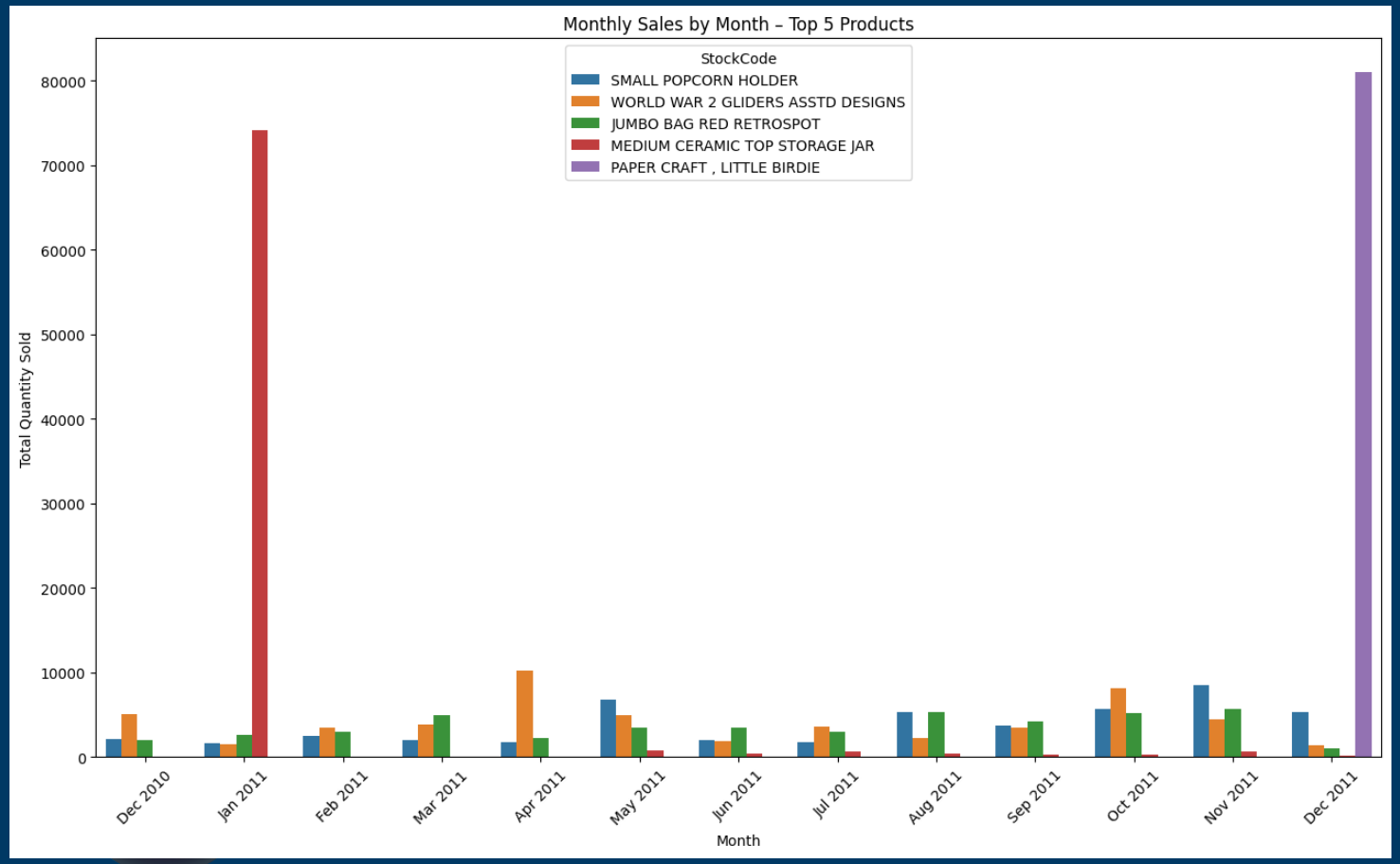


Sales show **fluctuations throughout the year**, with noticeable peaks during the **last quarter of 2011** (October–November).



Quantity is the main driver of TotalPrice, while UnitPrice has minimal impact.

Advanced Analysis



- Sales show clear seasonal fluctuations, with strong peaks in January and December.
- "MEDIUM CERAMIC TOP STORAGE JAR" and "PAPER CRAFT, LITTLE BIRDIE" dominate these months, suggesting seasonal or promotional demand.
- Other products maintain steady sales throughout the year.

RFM Matrix Analysis

•Customers with **low Recency** (12347) made purchases recently, indicating **active engagement**.

•Customers with **high Recency** (12346, 12350) have not purchased for a long time, showing **inactivity**.

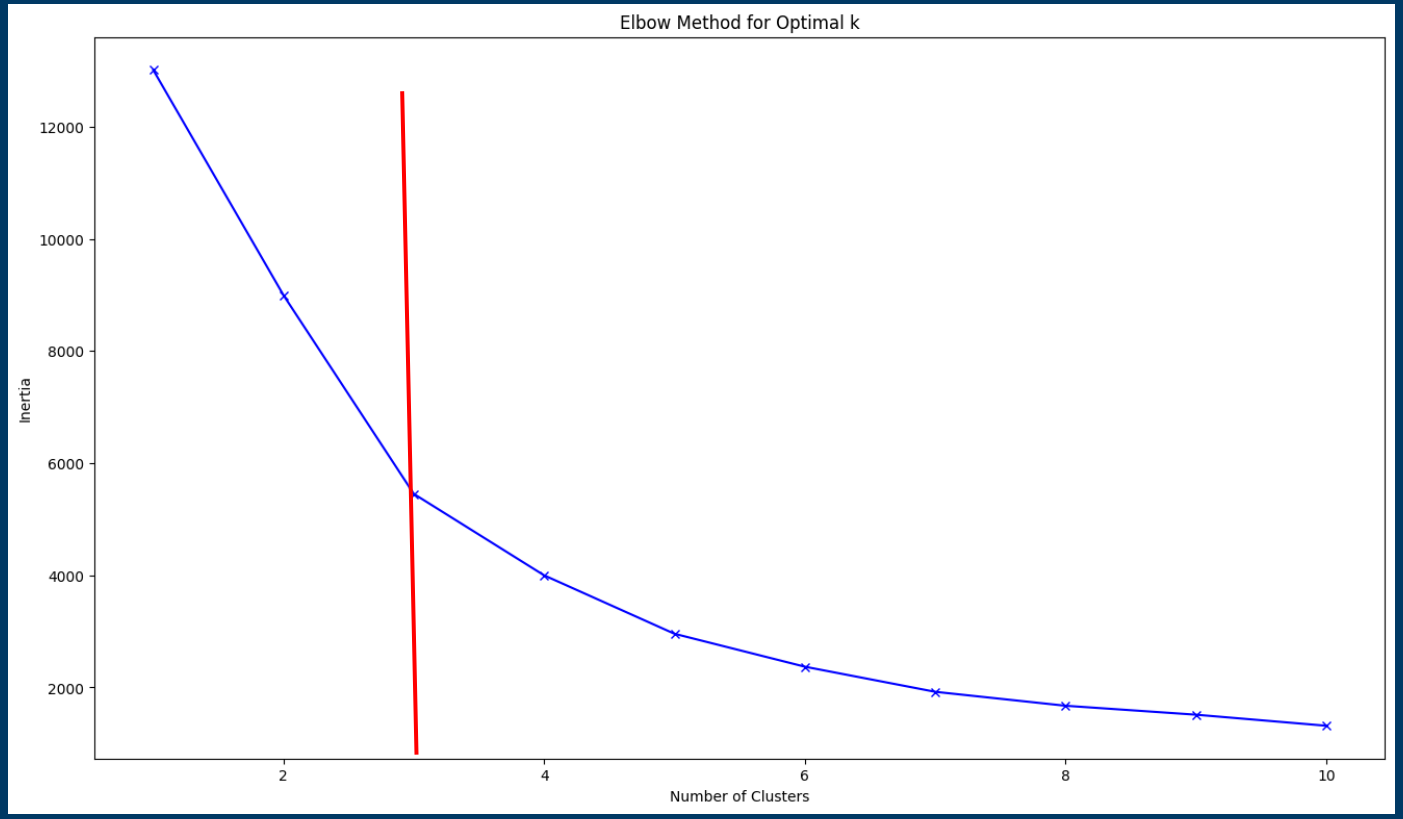
CustomerID	InvoiceDate (Recency)	InvoiceNo (Frequency)	TotalPrice (Monetary)
12346.0	325	1	77183.60
12347.0	1	182	4310.00
12348.0	74	31	1797.24
12349.0	18	73	1757.55
12350.0	309	17	334.40

•Higher **Frequency** and **Monetary values** (12347) suggest **loyal and valuable customers**.

•Lower **Frequency** and **Monetary values** represent **occasional or low-value buyers**.

KMeans clustering (Elbow Method)

The Elbow Method indicates that **3 clusters** provide a good balance between model simplicity and explained variance.



Predictive Analytics

Random Forest Classifier

	precision	recall	f1-score	support
0	1.00	1.00	1.00	79577
accuracy			1.00	79577
macro avg	1.00	1.00	1.00	79577
weighted avg	1.00	1.00	1.00	79577

- Predicted whether a customer would return an item.
- Achieved **perfect accuracy (100%)**, likely due to a **high class imbalance** — meaning there were very few or no returns in the dataset.
- Indicates the need for a **more balanced dataset** to obtain more realistic results.

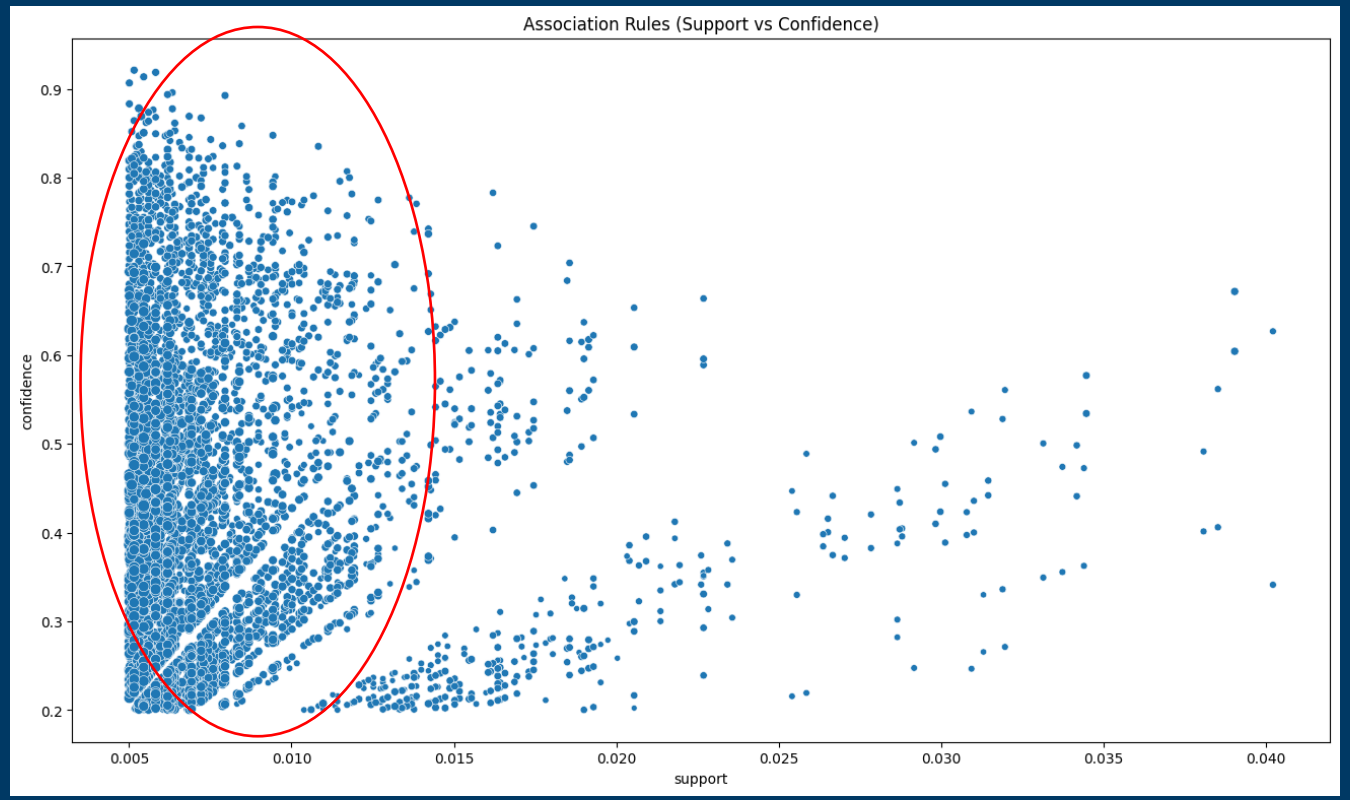
Linear Regression

```
RMSE: 55.08892423671942
MAE: 11.954276922845304
R^2: 0.29432829124450954
```

- Predicted the **invoice total price**.
- Achieved an **R^2 of 0.29**, explaining about **29% of price variance**.
- Suggests that **other factors**, such as product category or customer behavior, may influence the total price.

Association Rule Mining

	antecedents	consequents	support	confidence	lift
42	(ALARM CLOCK BAKELIKE GREEN)	(ALARM CLOCK BAKELIKE RED)	0.039039	0.671736	10.398510
43	(ALARM CLOCK BAKELIKE RED)	(ALARM CLOCK BAKELIKE GREEN)	0.039039	0.604333	10.398510
218	(BAKING SET 9 PIECE RETROSPOT)	(RETROSPOT TEA SET CERAMIC 11 PC)	0.017089	0.271663	4.481280
219	(RETROSPOT TEA SET CERAMIC 11 PC)	(BAKING SET 9 PIECE RETROSPOT)	0.017089	0.281896	4.481280
294	(HEART OF WICKER LARGE)	(HEART OF WICKER SMALL)	0.029979	0.508115	7.178114



- The Apriori algorithm identified products that are frequently bought together.
- For example, “Alarm Clock Bakelike Green” and “Red” show a strong association (lift = 10.39), meaning they are often purchased together.
- Most rules show **low support but high confidence**, indicating a few strong and specific product relationships.

Advanced EDA Techniques
Pareto Analysis (80/20 Rule)

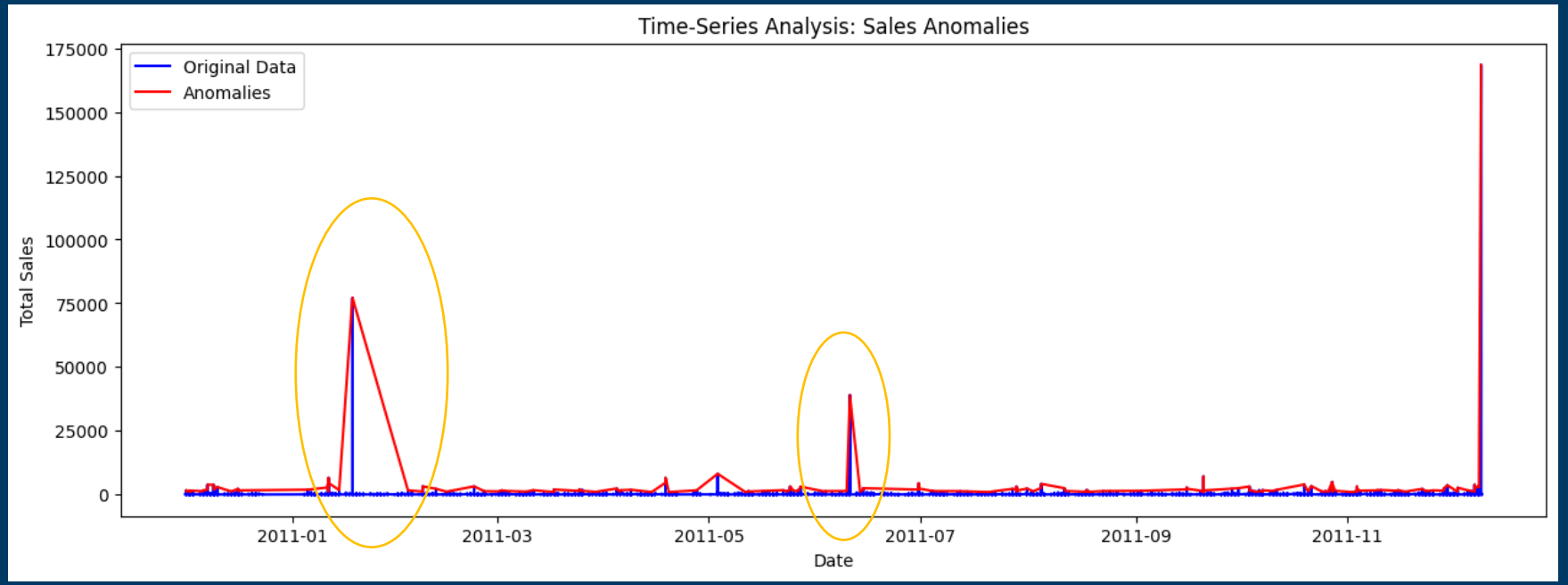
	StockCode	Description	TotalPrice	cum_pct
0	23843	PAPER CRAFT , LITTLE BIRDIE	168469.60	1.890494
1	22423	REGENCY CAKESTAND 3 TIER	142592.95	3.490611
2	85123A	WHITE HANGING HEART T-LIGHT HOLDER	100448.15	4.617797
3	85099B	JUMBO BAG RED RETROSPOT	85220.78	5.574108
4	23166	MEDIUM CERAMIC TOP STORAGE JAR	81416.73	6.487731
...	...	...	...	...
805	20973	12 PENCIL SMALL TUBE WOODLAND	2551.15	79.866139
806	21985	PACK OF 12 HEARTS DESIGN TISSUES	2550.29	79.894757
807	72802C	VANILLA SCENT CANDLE JEWELLED BOX	2545.37	79.923320
808	23266	SET OF 3 WOODEN STOCKING DECORATION	2543.39	79.951861
809	22072	RED RETROSPOT TEA CUP AND SAUCER	2539.19	79.980355

810 rows × 4 columns

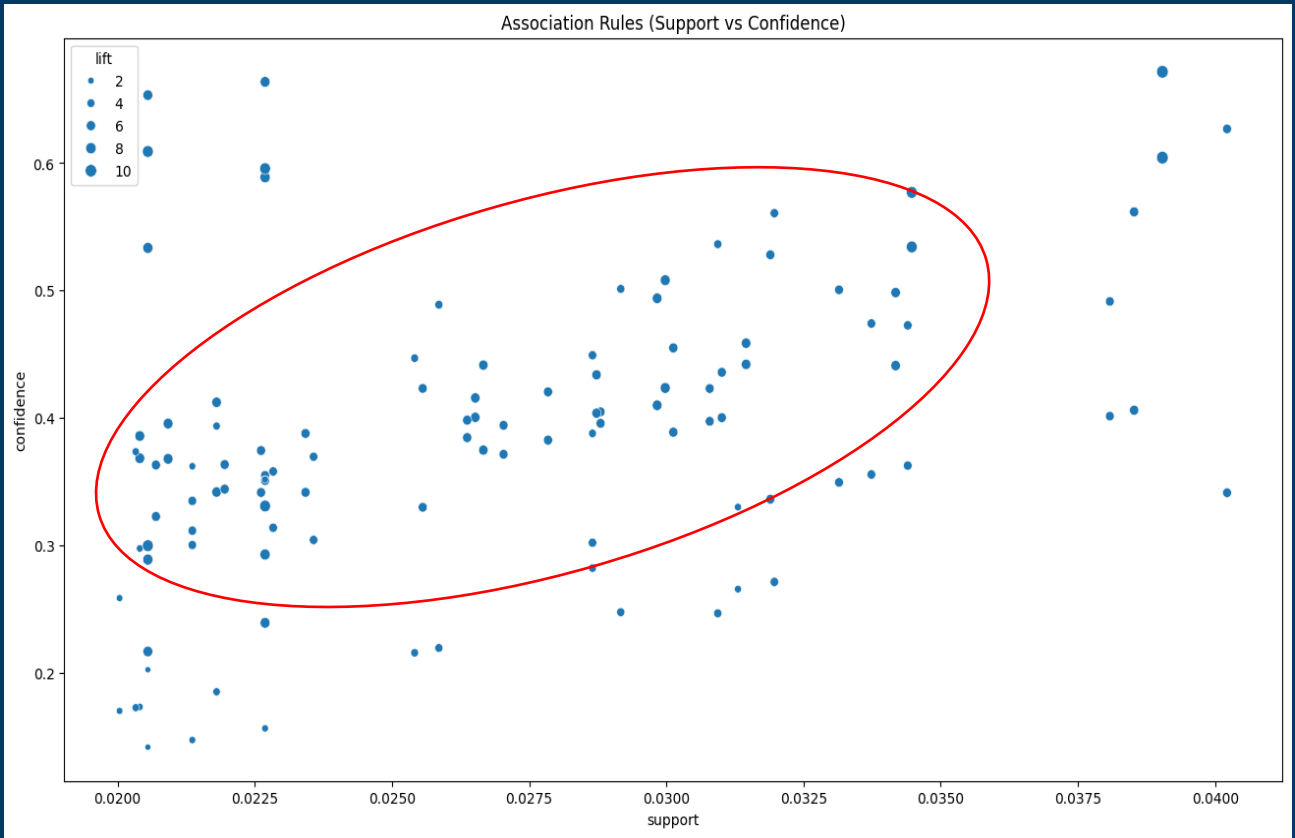
- The analysis shows that around 20% of the products account for 80% of total revenue.
- Top items like “PAPER CRAFT, LITTLE BIRDIE” and “REGENCY CAKESTAND 3 TIER” are the main revenue drivers.

Advanced EDA Techniques
Time-Series Anomalies

- The time-series analysis identified several sales spikes (in red), indicating days with unusually high revenue.
- These could be due to promotions, seasonal peaks, or data anomalies that need further investigation.



Advanced Modeling Techniques
Market Basket Analysis Enhancements



- The enhanced Apriori analysis explored product combinations with multiple items.
- Most rules show **moderate support but high confidence**, indicating products that are frequently bought together.

Conclusion

- The dataset was successfully cleaned, keeping only valid transactions.
- Some class imbalance affected model performance, especially in churn prediction.
- The data shows clear customer patterns, with a few products generating most of the revenue (Pareto rule).
- Clustering revealed three main customer groups with different purchasing habits.
- Association analysis found specific product combinations frequently bought together.
- The churn model had limited success due to class imbalance.



Recommendations



- Prioritize top-selling products (top 20%) in inventory planning, promotions, and marketing campaigns to maximize profits.
- Create product bundles based on frequently purchased combinations (e.g., “Alarm Clock Green + Alarm Clock Red”).
- Target high-value customers with loyalty programs or exclusive discounts to increase retention.
- Encourage low-frequency buyers with email reminders or small incentives to re-engage them.
- Improve data collection for missing or inactive customers to better predict churn in future analyses.