Retail Stock Market Behavior

TEAM - DATA SCOUTS

TEAM DATA SCOUTS

1. MEESALA SREE SAI NATH

2. AKULA JITHENDRANATH

3. S ANSAR TEJMUL MOVIN

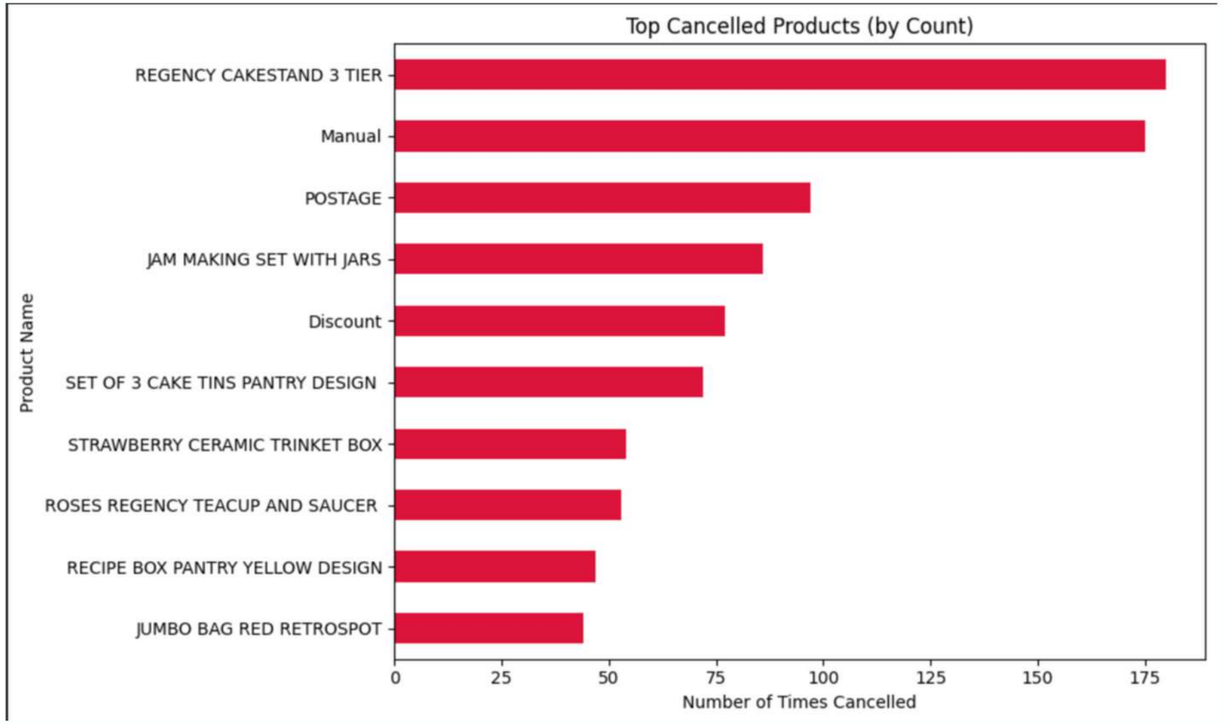
Exploratory Data Analysis

Dataset Overview

```
Data columns (total 8 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   InvoiceNo    541909 non-null  object
 1   StockCode    541909 non-null  object
 2   Description  540455 non-null  object
 3   Quantity     541909 non-null  int64
 4   InvoiceDate  541909 non-null  object
 5   UnitPrice    541909 non-null  float64
 6   CustomerID   406829 non-null  float64
 7   Country      541909 non-null  object
dtypes: float64(2), int64(1), object(5)
```

- 541,909 → 333,234 transactions (after cleaning)
- 4,191 unique customers
- 3,575 unique products
- 37 countries
- £4.3 Million total revenue
- Dec 2010 – Dec 2011 (UK-based online retail)

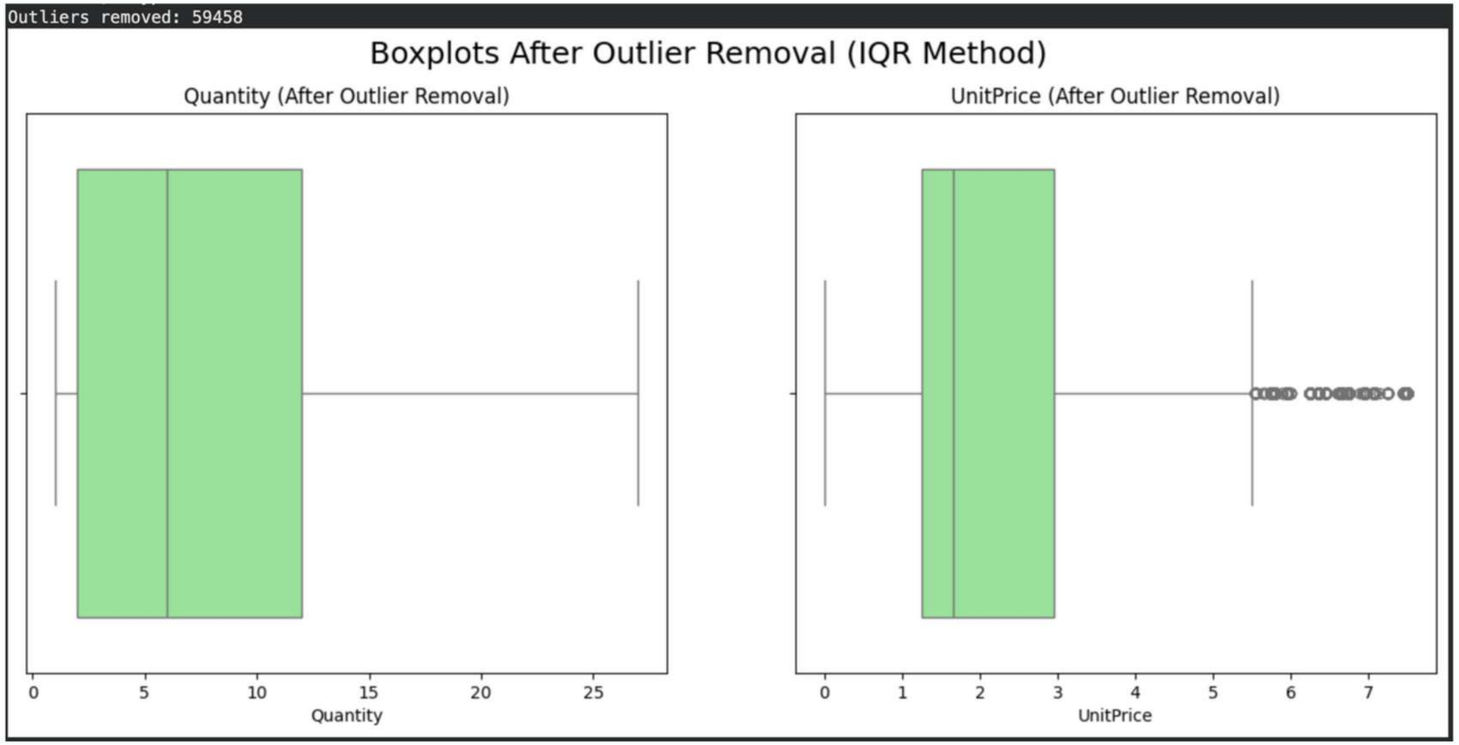
Exploratory Data Analysis



- No of Duplicates entries: 5268
- No of cancelled invoices: 8872
- No of invalid Quantity/Price rows: 40

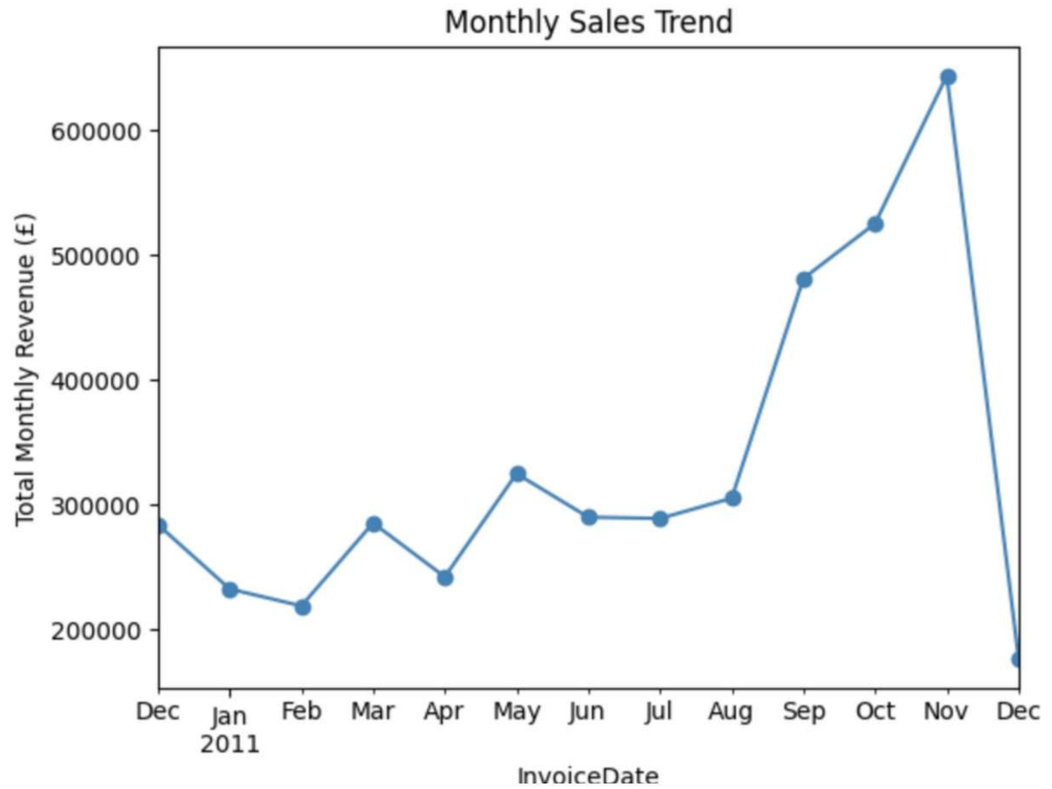
Invoices beginning with “C” represent cancelled transactions. We exclude them to prevent double-counting and revenue misrepresentation.

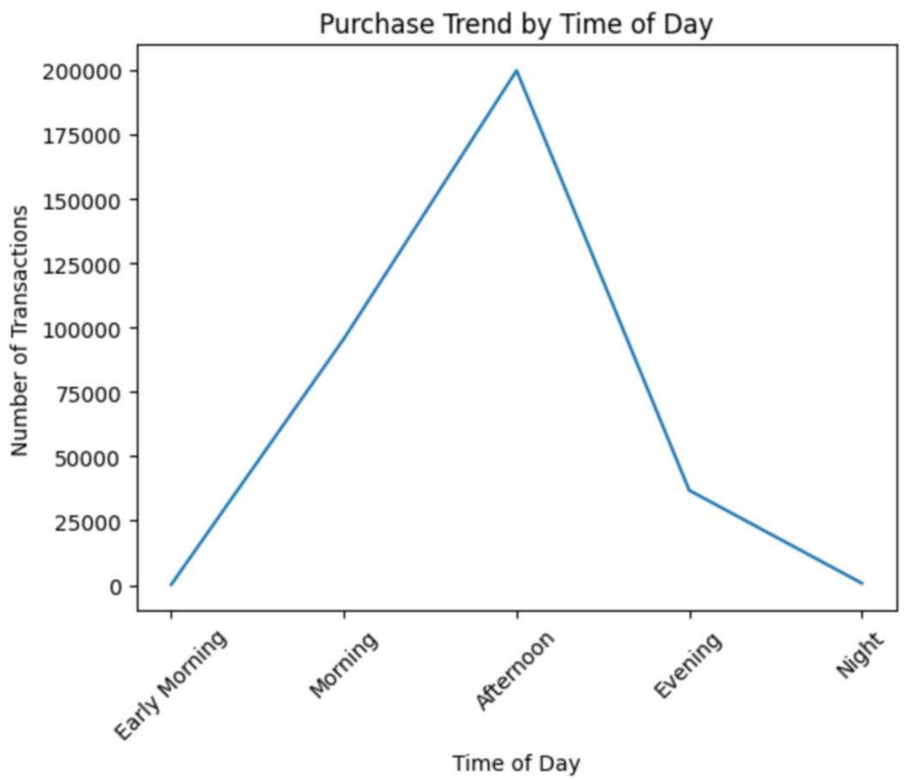
Exploratory Data Analysis



- Removed outliers using IQR Method

Exploratory Data Analysis

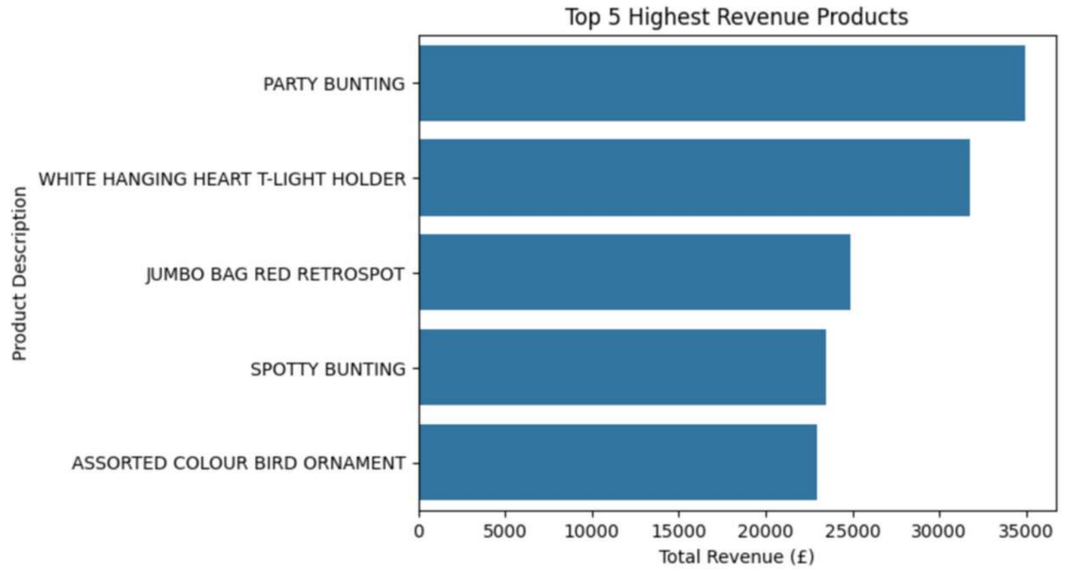


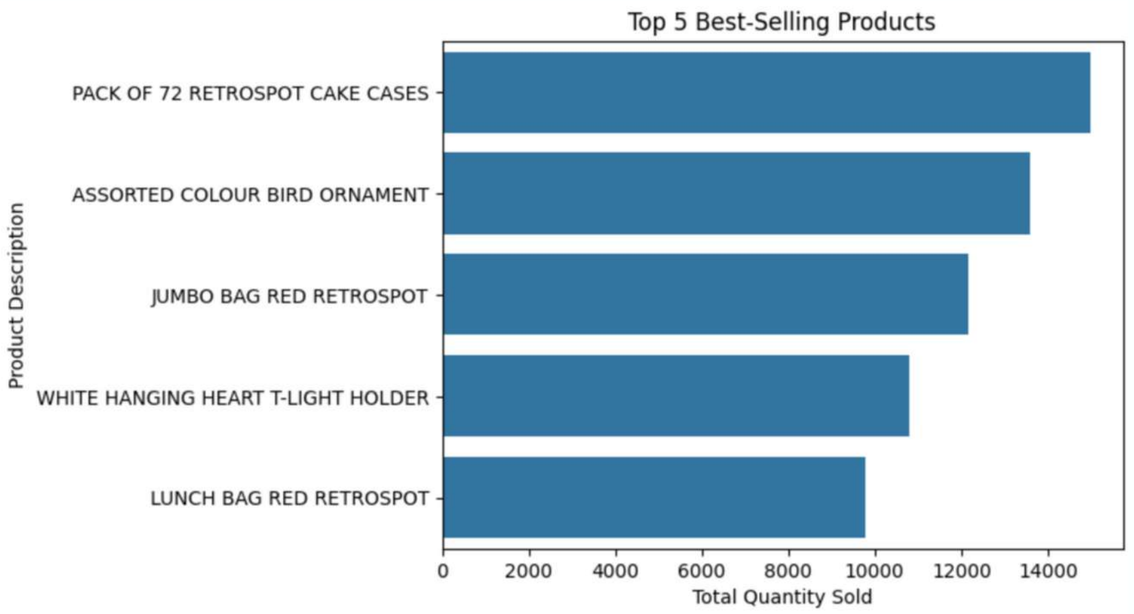


- Peak at Nov-Dec 2011 (Christmas)
- Q4 Christmas Peak = Massive Revenue Jump

- Peak hours at Afternoon
- No Orders between Night and Early Mornings

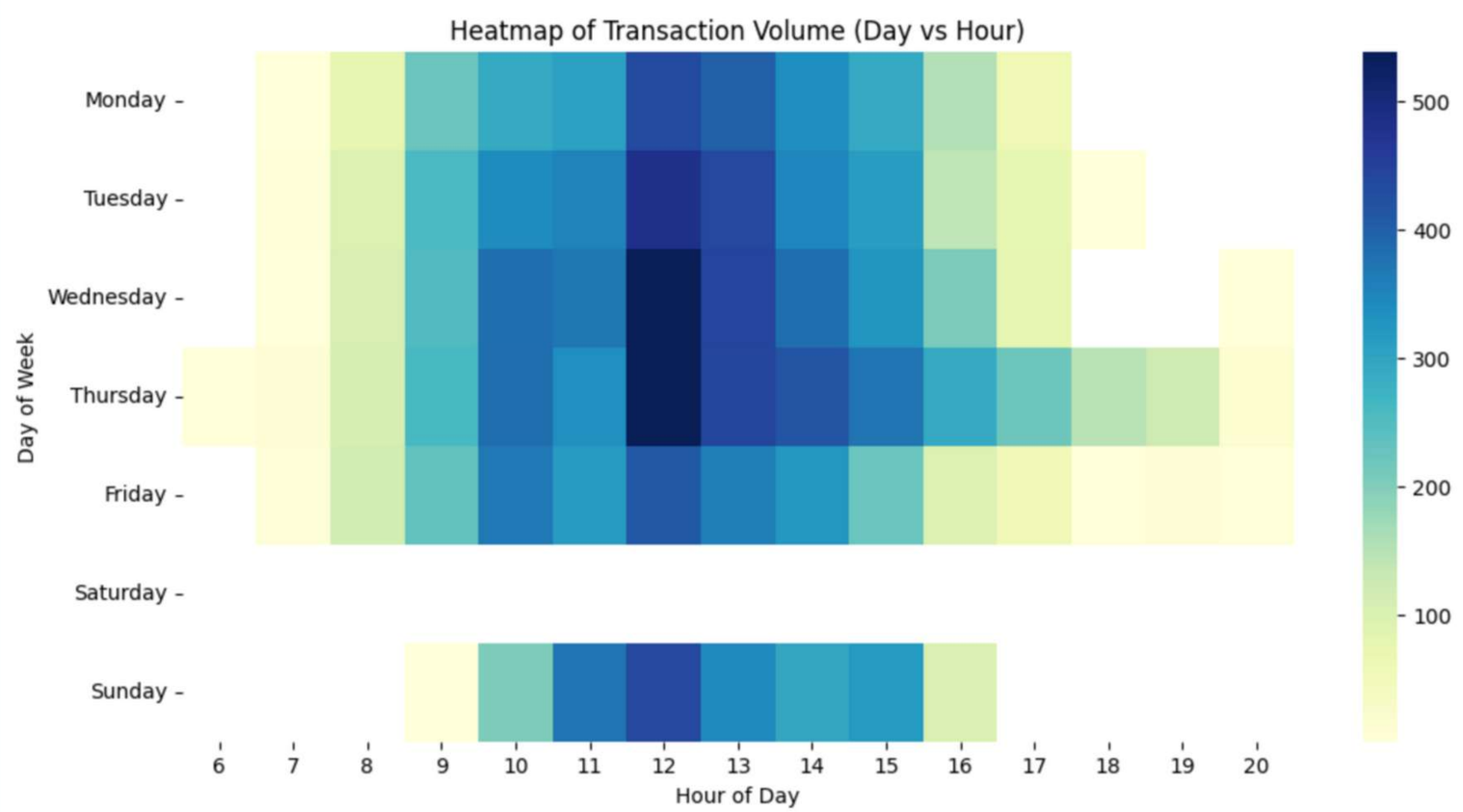
Exploratory Data Analysis





- Here we got to know that , the Best Selling Product is not the Top 5 of the Highest Revenue Products

Exploratory Data Analysis

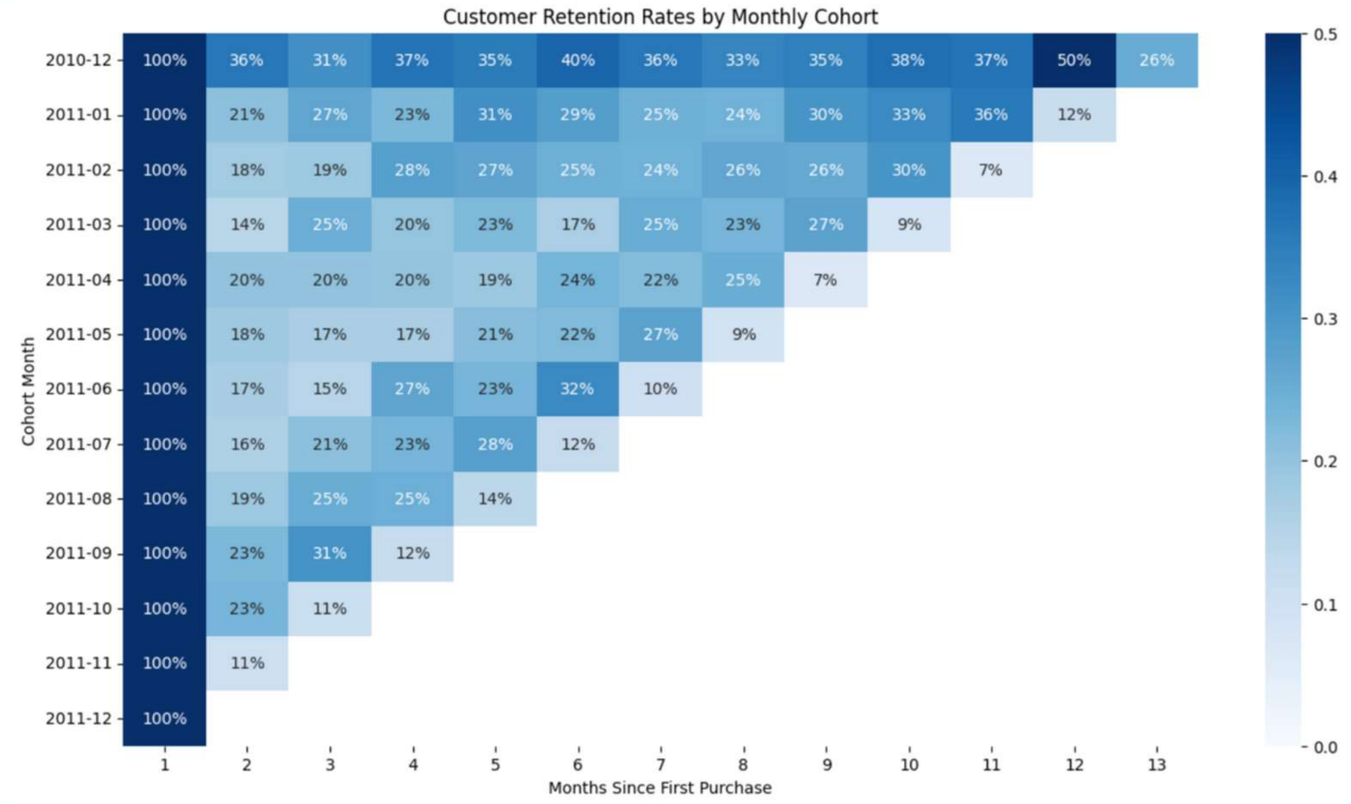


- Peak transaction volume occurs on weekdays from 8 AM–3 PM, especially around noon
- Activity nearly vanishes after 4 PM and on weekends

Exploratory Data Analysis

- Product word cloud dominated by "BAG", "HEART", "LIGHT HOLDER", "RED", "PINK", "VINTAGE", "CHRISTMAS".
- Store focuses on quirky, colorful, heart-themed gifts and seasonal decor and not on everyday essentials.

Exploratory Data Analysis

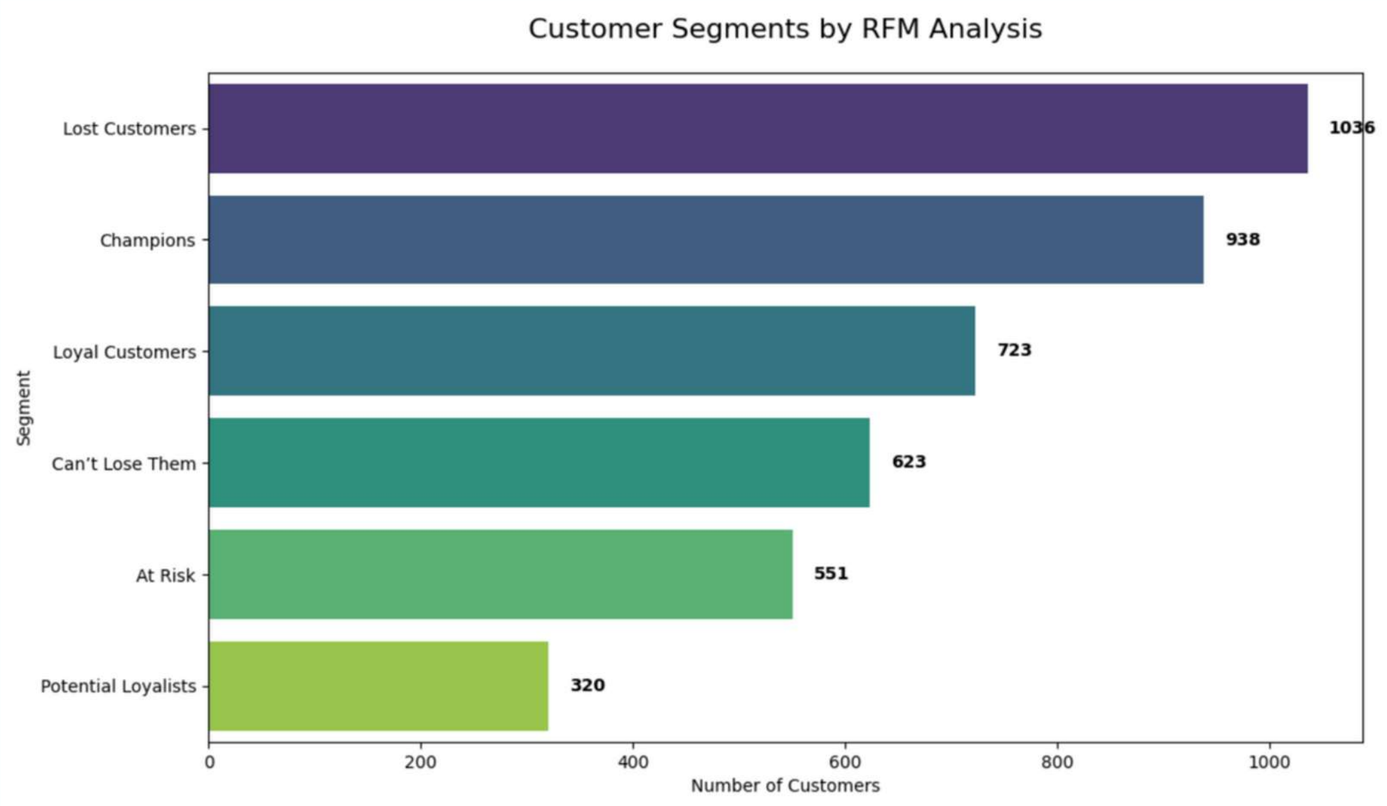


Customer Retention Crisis

- 95% customers lost within 6 months
- Christmas 2010 cohort retains 50% after 12 months
- Massive retention problem. Immediate action required by retailers

Github link

Exploratory Data Analysis



Lost Customers - 1036 , Champions - 938 , Loyal Customers - 723,
Can't Lose Them - 623, At Risk - 551 , Potential Loyalists - 320

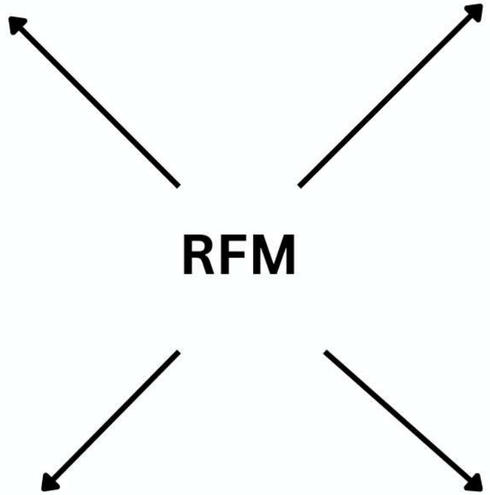
Exploratory Data Analysis

The "Champions" (High Value)

- Who: Recent, frequent, high spenders (Score 4+).
- Action: "Reward them. They don't need discounts, they need exclusivity and early access."

The "Can't Lose Them" (Immediate Alert)

- Who: Used to buy often, but haven't seen them lately (R ≤ 2, F ≥ 3).
- Action: "Reactivate immediately. Send aggressive offers or surveys to find out why they left."



The "Potential Loyalists" (Growth Opportunity)

- Who: Recent buyers with average frequency (R ≥ 4, F ≥ 2).
- Action: "Upsell them. Offer membership programs to turn them into Champions."

The "Lost Customers" (Cut Losses)

- Who: Haven't bought in a long time and rarely visited (R ≤ 2, F ≤ 2).
- Action: "Stop spending marketing budget here. Focus resources on acquiring new users instead."

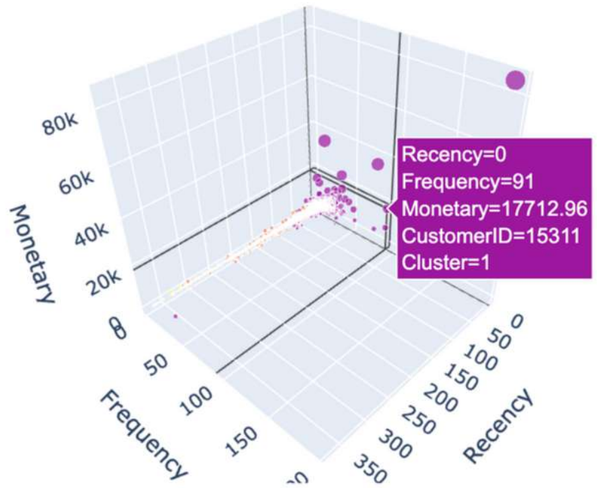
Descriptive & Predictive Analysis

K-Means Clustering: Optimal K & Final Customer Segments

By using **Plotly** on Recency , Frequency, Monetary

FINAL CUSTOMER SEGMENTS

Cluster	Recency	Frequency	Monetary	CustomerID	%	Segment
1	10.0	12.7	3461.4	668	15.9	Loyal Customers (Frequent, Good Spenders)
2	63.3	4.1	1137.7	1192	28.4	At Risk (Spent big, but long ago)
0	20.8	1.9	336.3	819	19.5	Champions (High Value, Recent, Frequent)
3	188.0	1.3	235.3	1512	36.1	Lost / Hibernating (Low everything)



Using RFM-based K-Means clustering, we identified 4 distinct customer groups that enable targeted marketing, better retention, and revenue optimization.

Descriptive & Predictive Analysis

Market Basket Analysis & Apriori:

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift
0	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.024234	0.039356	0.021527	0.888283	22.570744
1	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER , PINK REGENCY TEACUP AND SAUCER)	0.039356	0.024234	0.021527	0.546980	22.570744
2	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.030243	0.031630	0.021527	0.711790	22.503870
3	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER, ROSES REGENCY TEACUP AND SAUCER)	0.031630	0.030243	0.021527	0.680585	22.503870
4	(GREEN REGENCY TEACUP AND SAUCER)	(PINK REGENCY TEACUP AND SAUCER)	0.039356	0.031630	0.025819	0.656040	20.741281
5	(PINK REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.031630	0.039356	0.025819	0.816284	20.741281

- All Regency Teacup & Saucer variants show extremely strong cross-selling behavior.
- Buying one variant strongly predicts purchasing the others (lift > 20), making this the most powerful association in the dataset.

Descriptive & Predictive Analysis

Predict Big Orders (> £500) – XGBoost Classifier

The model shows strong performance

- Accuracy 96.9% and F1-score 0.844 indicate reliable big-order prediction.

Price and quantity drive predictions

- UnitPrice and Quantity are the most influential features.

WORK PLANNING & DIVISION

DATE RANGE	PHASE	FOCUS AREA	KEY DELIVERABLES	PHASE LEAD
Nov 6 - Nov 20	Phase 2	Data Exploration & Modeling	EDA Notebooks, Processed Dataset Trained Models	Akula Jithendranath

TEAM PLANNING:

AKULA JITHENDRANATH - DESCRIPTIVE & PREDICTIVE MODELING

MEESALA SREE SAI NATH - PREPROCESSING & EDA

S ANSAR TEJMUL MOVIN - RESEARCH & PREPROCESSING

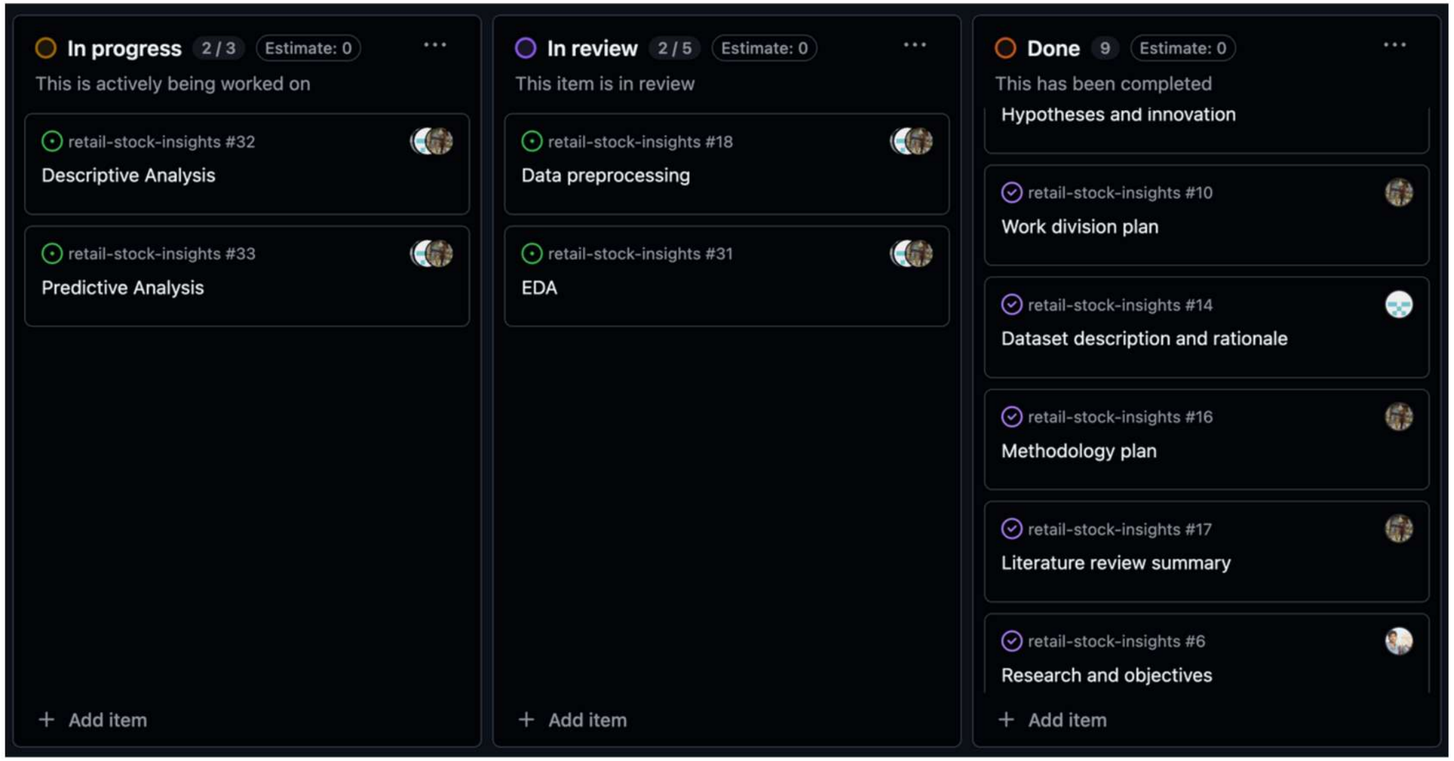
The primary objective of this plan is to ensure:

- Equal technical and documentation contributions across all members.
- Clear accountability through leadership rotation.
- Systematic workflow aligned with data-mining process stages.

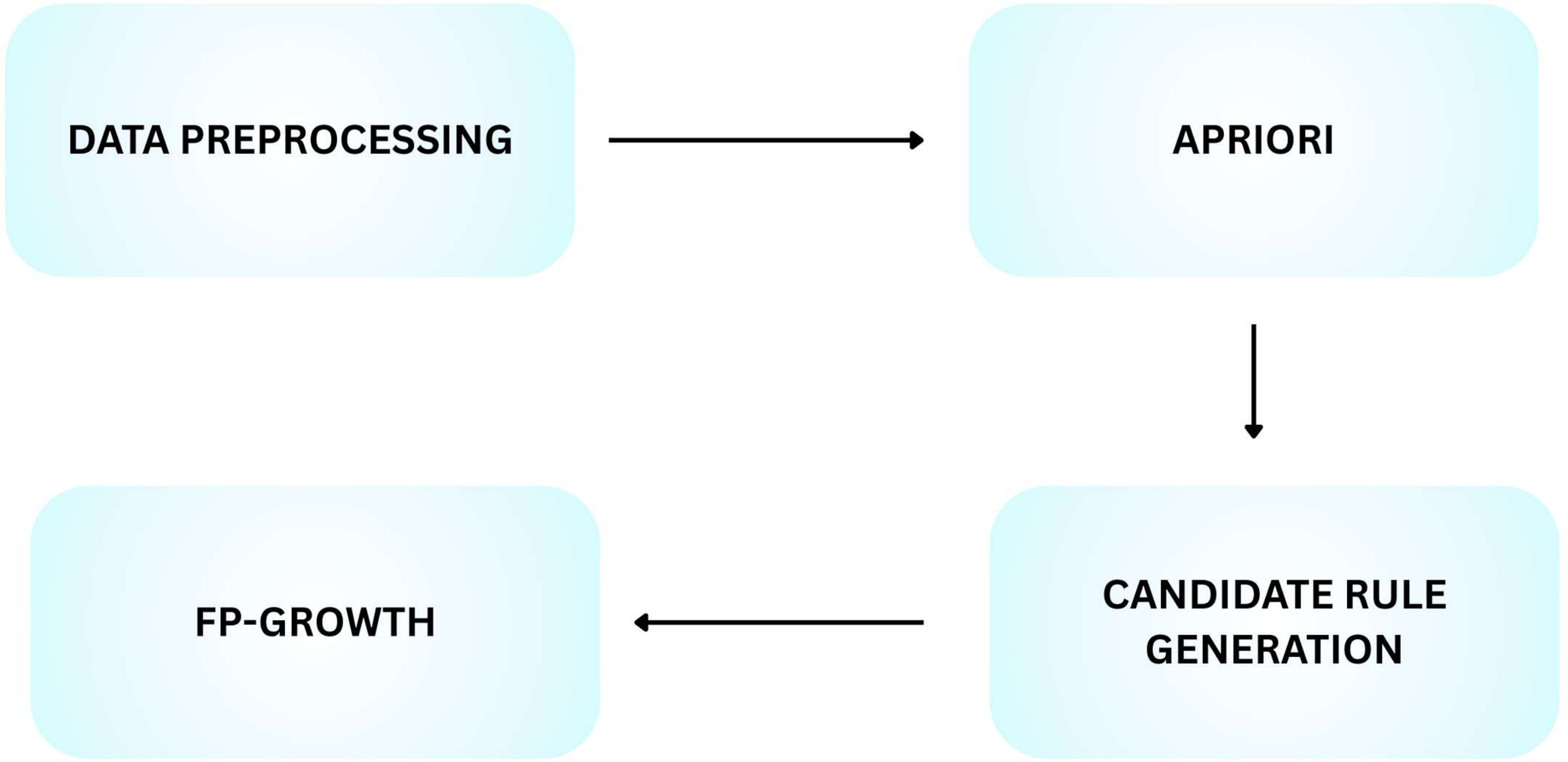
WORK PLANNING & DIVISION

Phase Leadership and Workflow Management:

In each phase, the designated Phase Lead will be responsible for reviewing the Pull Requests (PRs) raised by team members. Task assignments, progress tracking, and updates will be managed using the Kanban board, ensuring organized workflow and clear accountability throughout the project lifecycle.



RESEARCH METHODOLOGY



RESEARCH METHODOLOGY

Why did we also do FP - Growth and what we achieved ?

- Research shows FP-Growth is faster on large datasets because it avoids candidate generation.
- In our experiment, Apriori (2s) was faster than FP-Growth (6s).
- This happened because our dataset is small, so Apriori's bottlenecks never appeared.
- FP-Growth has higher overhead (FP-tree building + recursion), which slows it down on small data.

Conclusion: Apriori is faster for small datasets; FP-Growth scales better for large datasets.

Thankyou

~ Team DATA SCOUTS