A

PROJECT REPORT

ON

CUSTOMER PURCHASE BEHAVIOR ANALYSIS

IN RETAIL

A System-Based Data Analytics Project

By

AYESHA BANU

LinkedIn | GitHub | ayesha24banu@gmail.com

📅 Completion Date: July, 2025

DECLARATION

I, Ayesha Banu, hereby declare that the project titled "Customer Purchase Behavior Analysis in Retail" is a result of my independent work and has not been submitted previously for any degree or diploma at any university or institute. This project has been developed as part of my self-guided learning in the field of Data Science

All data, tools, code, and analysis used in this project are sourced from publicly available platforms, and due acknowledgments have been made wherever applicable.

I confirm that the contents of this report are true to the best of my knowledge and belief.

Date: July 2025

Name: Ayesha Banu

Signature:

Ayesha

ABSTRACT

This project presents a comprehensive data analysis/data science solution aimed at uncovering insights from customer transactional data in a retail environment. Using a combination of Python-based analytics and Power BI visualization, the project seeks to understand customer purchasing behavior, identify high-value customer segments, and recommend product bundling strategies.

The analysis begins with data cleaning and preparation of over 779,000 sales transactions sourced from the Online Retail dataset. Key preprocessing steps include handling missing values, calculating derived metrics like total revenue, and structuring the data for analytical use. The cleaned data is then explored through detailed **Exploratory Data Analysis (EDA)** to identify top-selling products, seasonal revenue trends, peak purchasing hours, and country-wise performance.

Next, the project applies **RFM (Recency, Frequency, Monetary)** analysis combined with **KMeans clustering** to segment customers into distinct behavioral groups, such as Loyal, Inactive, and High-Spending customers. These segments provide valuable input for targeted marketing and customer retention strategies.

To enhance product strategy, **Market Basket Analysis** is performed using the Apriori algorithm, uncovering frequently co-purchased items and generating association rules. These rules support product bundling and cross-selling initiatives.

The results are presented via a **four-page interactive Power BI dashboard**, showcasing KPIs, revenue drivers, customer segments, and association rules in a business-friendly format. Outputs are also stored as structured CSV files and visual plots, making them reusable for stakeholders or integration into enterprise platforms.

In essence, this project bridges business strategy and data science by delivering actionable insights that can drive customer retention, improve product placement, and inform data-driven decision-making in retail operations.

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1. Introduction

1.1 Company Profile

GlobalCart Retailers Ltd. is a mid-sized online retail enterprise based in the United Kingdom, serving thousands of customers across Europe. Specializing in home decor, lifestyle accessories, and gifting products, GlobalCart has built a strong digital presence over the past decade.

With a portfolio of over 4,000 SKUs and a customer base spread across 38 countries, the company processes hundreds of transactions daily via its e-commerce platform. Despite its robust infrastructure and growing sales volume, GlobalCart faces increasing pressure to **retain customers**, **boost average order value**, and **optimize marketing efforts**.

Recognizing the potential locked within their vast sales database, the company turned to **data-driven strategies** to stay competitive and unlock new revenue streams.

1.2 Problem Statement

Retailers like GlobalCart collect large volumes of customer and transactional data, yet often fail to translate it into actionable intelligence. Without a deeper understanding of customer behavior, the company struggles to answer critical questions:

- Who are our most valuable customers?
- What product combinations are frequently bought together?
- Are there customers we are at risk of losing?
- What promotions would drive higher engagement?

The lack of segmentation and behavior-based targeting leads to **generic marketing**, **ineffective bundling**, and **missed cross-selling opportunities**. This results in declining customer loyalty, poor inventory turnover, and reduced campaign effectiveness.

To overcome these challenges, GlobalCart needs a **data analytics solution** that not only explores historical sales but also segments customers, identifies profitable patterns, and visualizes insights in a business-friendly format.

1.3 Objective

This project aims to **transform raw retail data into actionable insights** that support strategic business decisions. The core objectives include:

- **Segmenting customers** using RFM (Recency, Frequency, Monetary) analysis and clustering algorithms to identify patterns in customer behavior.
- **Discovering association rules** through Market Basket Analysis to inform product bundling and cross-selling strategies.
- **Visualizing key metrics and trends** in a Power BI dashboard for marketing and leadership teams.
- **Delivering business intelligence** that can be directly applied to improve customer retention, revenue growth, and operational efficiency.

In short, the goal is to move from **reactive data use** to **proactive, insight-driven retail strategy**.

1.4 Business Use Case

Imagine GlobalCart's marketing team preparing for the holiday season. They're ready to launch a campaign but don't know:

- Which customers are loyal and likely to return?
- Which items are commonly purchased together for bundle offers?
- Which customer group needs re-engagement efforts?

This is where data comes in.

As the Data Analyst/Data Scientist, your mission is to:

- Dive into over **779,000 historical transactions** and clean the data.
- Use **RFM scoring and KMeans clustering** to segment customers into groups like *Champions*, *At-Risk*, and *Hibernating*.
- Apply **Apriori-based Market Basket Analysis** to find rules like:
  "Customers who buy White Ceramic Tea Sets also buy Floral Tray Sets with 68% confidence."
- Build a **Power BI dashboard** that allows business users to:
  - View trends over time
  - Filter by country or customer segment
  - Identify best-selling items and product bundles

This project empowers GlobalCart's decision-makers to **act with precision**, ensuring every campaign, product offer, and re-targeting effort is backed by data.

CHAPTER 2

DATASET DESCRIPTION

2. Dataset Description

To uncover customer insights and drive data-driven decisions, this project utilizes the publicly available **Online Retail Dataset** sourced from **Kaggle**. The dataset captures real-world e-commerce transactions and is widely used in retail analytics research and industry case studies.

2.1 Source and Period

- **Dataset Source:** Kaggle (Online Retail Dataset)
- **Time Frame Covered:** December 1, 2010 to December 9, 2011
- **Region:** Predominantly United Kingdom with international sales across 38 countries

This dataset is a rich collection of transaction-level data from a UK-based online retailer, offering detailed insights into customer purchase behavior across time and geography.

2.2 Volume and Structure

- **Total Records:** Approximately 779,000+ transactions
- **Number of Features (Columns):** 8
- **File Format:** CSV

Each record in the dataset represents a product purchase as part of a sales invoice. The dataset supports both customer-level and product-level analysis.

2.3 Feature Overview

The table below outlines each attribute in the dataset along with its description:

Column Name	**Description**
InvoiceNo	Unique identifier for each transaction (can represent full or partial return)
StockCode	Product/item identifier
Description	Name/description of the product
Quantity	Number of items purchased
InvoiceDate	Timestamp of the transaction
UnitPrice	Price per unit of product
Customer ID	Unique identifier for each customer
Country	Country where the customer is located

2.4 Data Considerations and Cleaning

Before analysis, the dataset requires significant preprocessing to ensure accuracy and consistency. Key challenges addressed include:

- **Missing Values:** Some records lack `CustomerID`, which are excluded from customer segmentation
- **Negative Quantities:** Returns or cancellations are handled separately during analysis
- **Date Formatting:** `InvoiceDate` is converted to proper datetime format for time-series insights
- **Duplicates:** Checked and removed where necessary

These steps ensure the reliability of downstream analytics, segmentation models, and dashboard visualizations.

CHAPTER 3
TOOLS & TECHNOLOGIES

3. Tools & Technologies

This project leverages a modern data analytics technology stack combining programming, visualization, and machine learning tools. Each tool was selected to support a specific stage of the data science lifecycle — from data wrangling to insight delivery.

3.1 Overview of the Technology Stack

Layer	Technology Used
Programming Language	Python 3.x
Data Manipulation	Pandas, NumPy
Data Visualization	Matplotlib, Seaborn
Machine Learning	KMeans Clustering (Scikit-learn), Apriori Algorithm (Mlxtend)
Business Intelligence	Power BI
Notebook Environment	Jupyter Notebook
Optional Database	MySQL (for structured data storage and SQL queries, if required)

Each component plays a specific role in enabling efficient data preparation, analysis, modeling, and presentation.

3.2 Tool Usage Description

- **Python 3.x:** The primary programming language used to clean, transform, and model the data. Chosen for its rich ecosystem of data science libraries.
- **Pandas & NumPy:** Used extensively for data manipulation, aggregation, & preprocessing.
- **Matplotlib & Seaborn:** Visualization libraries used to generate high-quality plots for EDA and reporting purposes.
- **Scikit-learn & MLxtend:** Employed to build clustering models and perform association rule mining respectively.
- **Power BI:** The business intelligence tool used to design an interactive and visually compelling dashboard to present KPIs and insights to stakeholders.
- **Jupyter Notebook:** Served as the interactive coding environment to develop, test, and document each phase of the project workflow.

CHAPTER 4
PROJECT ARCHITECTURE

4. Project Architecture

To maintain modularity, scalability, and clarity, the project is organized into a structured folder hierarchy. Each directory serves a distinct purpose — from raw data storage to model scripts and business reports. This structure ensures that team members or future collaborators can navigate the project efficiently.

4.1 Project Structure Overview

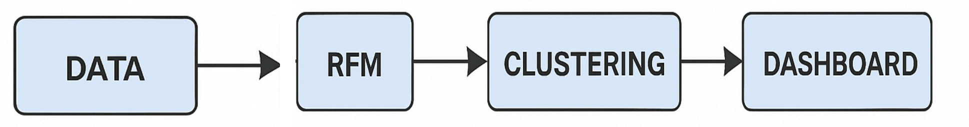
```
customer_purchase_analysis/
├── data/        # Raw datasets
│ └── online_retail.csv
├── scripts/       # Python scripts for preprocessing, modeling
│ ├── _init_.py
│ ├── utils.py
│ ├── data_cleaning.py
│ ├── mysql_pipeline.py
│ ├── eda_analysis.py
│ ├── rfm_segmentation.py
│ └── market_basket.py
├── notebooks/          # Jupyter notebooks for exploration and development
│ └── purchase_analysis.ipynb
├── outputs/             # Generated plots, intermediate CSVs, model results
│ ├── data/             # cleaned dataset, rfm_segments, association_rules .csv files
│ │ ├── clean_online_retail.csv
│ │ ├── rfm_segments.csv
│ │ └── association_rules.csv
│ └── figures/          # Visual assets and plots for documentation
│ ├── eda_fig/
│ ├── rfm_fig/
│ └── mba_fig/
├── logs/
│ └── process_log.log
├── Reports/        # Power BI (.pbix) files, Final report, presentation, PDF files and  images
│ ├── Customer_Purchase_Analysis.pbix
│ ├── Customer_Purchase_Analysis.pdf
│ ├── BI_Executive_Summary.png
│ ├── BI_Sales_Trend.png
│ ├── BI_RFM_Segments.png
│ └── BI_Market_Basket.png
├── requirements.txt      # List of dependencies and Python packages
└── README.md        # Project summary and GitHub instructions
```

4.2 Folder-wise Description

Folder/File	Purpose
`data/`	Stores the raw dataset, cleaned versions, and feature-engineered data files
`scripts/`	Contains Python modules for RFM scoring, clustering, and association rules
`notebooks/`	Jupyter notebooks used for development, EDA, and step-wise implementation
`outputs/`	Includes generated CSVs, charts, and clustering results
`dashboard/`	Power BI working file & exported dashboards (e.g., PDF, images)
`images/`	High-resolution visualizations inserted into the documentation
`Reports/`	Final report documents, presentations, or supplementary PDFs
`requirements.txt`	Lists all Python libraries required for project execution
`README.md`	Project overview, setup guide, and GitHub link if hosted publicly

4.3 Integration Overview

Each module in the project pipeline is designed to operate independently but cohesively, making it easier to plug into different dashboards, reports, or production pipelines.



CHAPTER 5
END-TO-END WORKFLOW

5. End-to-End Workflow

This project follows a structured and iterative data science workflow. Each stage is carefully designed to ensure data quality, model accuracy, and business relevance. The steps are executed in sequence, with checkpoints for validation and interpretation at each stage.

5.1 Step-by-Step Process

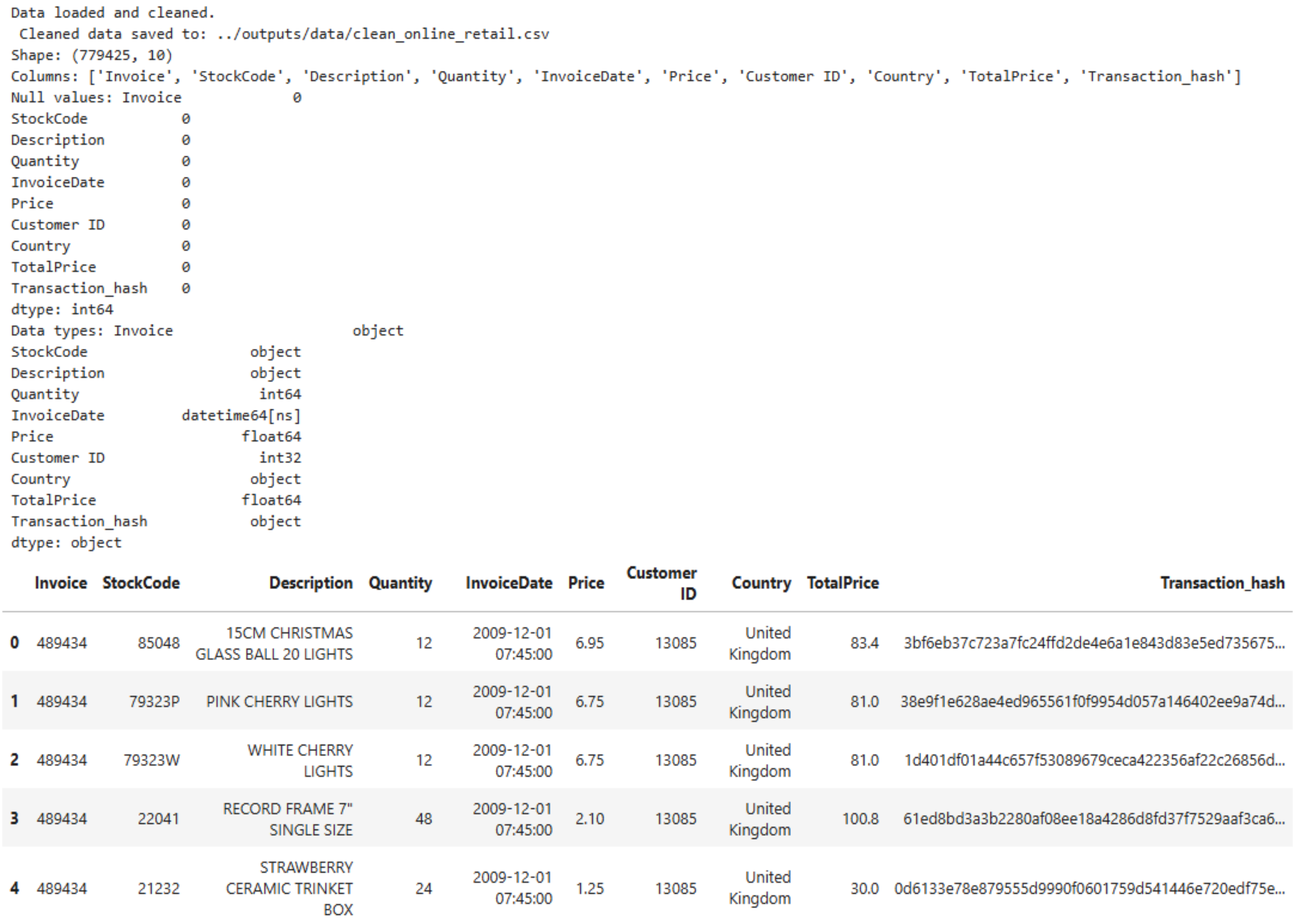
Step 1: Data Cleaning and Preprocessing

Goal: Prepare the raw dataset for analysis by handling errors, inconsistencies, and missing values.

Actions Taken:

- Removed duplicate and irrelevant records
- Filtered out canceled transactions with negative quantities
- Converted invoice dates into datetime format
- Excluded records with missing Customer IDs

Output: Cleaned dataset ready for exploration and modeling



```
Data loaded and cleaned.
 Cleaned data saved to: ../outputs/data/clean_online_retail.csv
Shape: (779425, 10)
Columns: ['Invoice', 'StockCode', 'Description', 'Quantity', 'InvoiceDate', 'Price', 'Customer ID', 'Country', 'TotalPrice', 'Transaction_hash']
Null values: Invoice             0
StockCode           0
Description         0
Quantity            0
InvoiceDate         0
Price               0
Customer ID         0
Country             0
TotalPrice          0
Transaction_hash    0
dtype: int64
Data types: Invoice                     object
StockCode                   object
Description                 object
Quantity                     int64
InvoiceDate         datetime64[ns]
Price                      float64
Customer ID                  int32
Country                     object
TotalPrice                 float64
Transaction_hash            object
dtype: object
```

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country	TotalPrice	Transaction_hash
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13085	United Kingdom	83.4	3bf6eb37c723a7fc24ffd2de4e6a1e843d83e5ed735675...
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	38e9f1e628ae4ed965561f0f9954d057a146402ee9a74d...
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13085	United Kingdom	81.0	1d401df01a44c657f53089679ceca422356af22c26856d...
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.10	13085	United Kingdom	100.8	61ed8bd3a3b2280af08ee18a4286d8fd37f7529aaf3ca6...
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13085	United Kingdom	30.0	0d6133e78e879555d9990f0601759d541446e720edf75e...

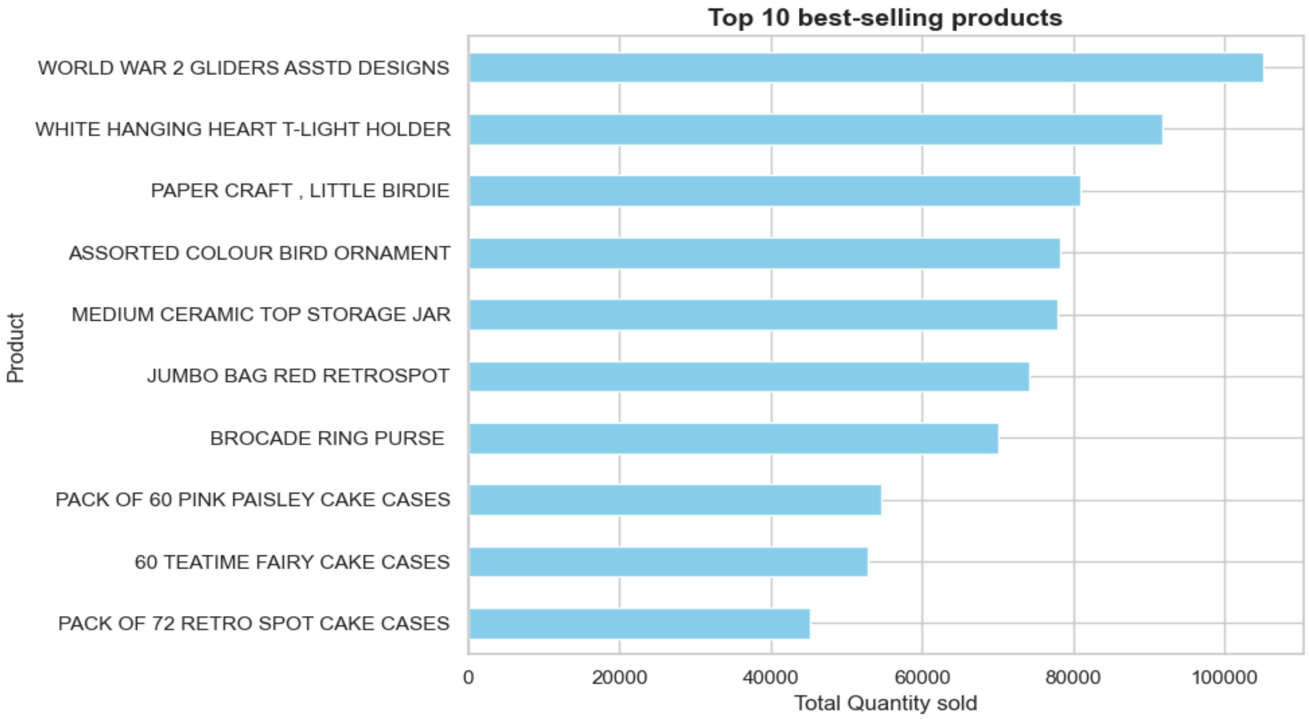
Step 2: Exploratory Data Analysis (EDA)

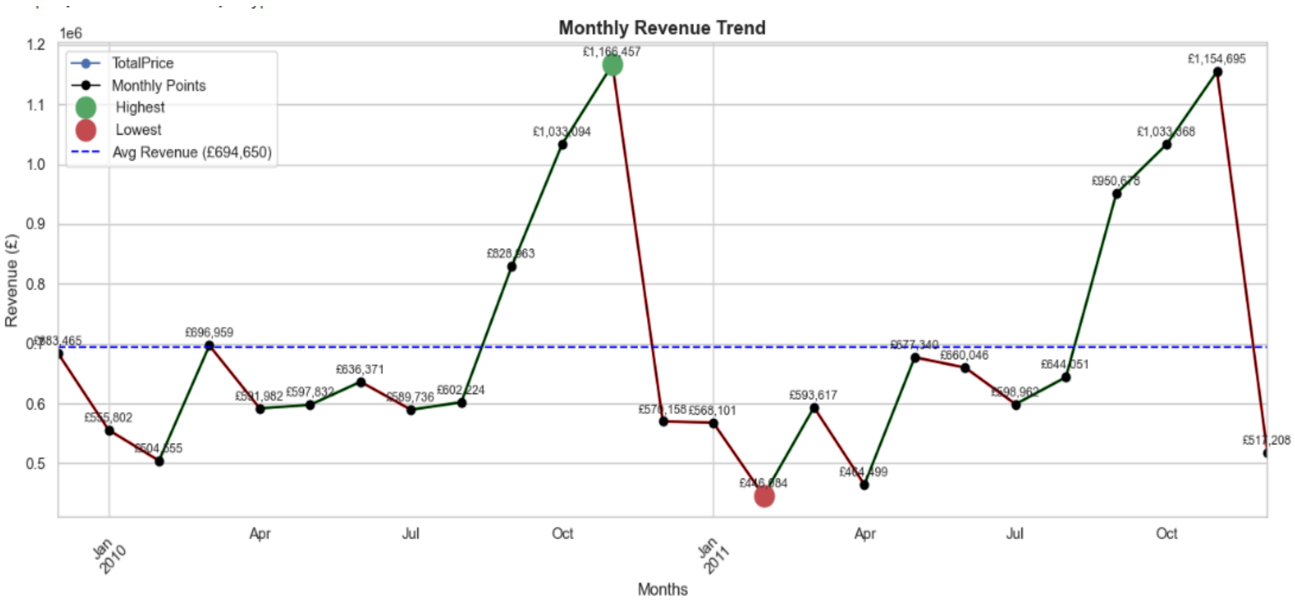
Goal: Understand data distribution, trends, & patterns through visual & statistical summaries.

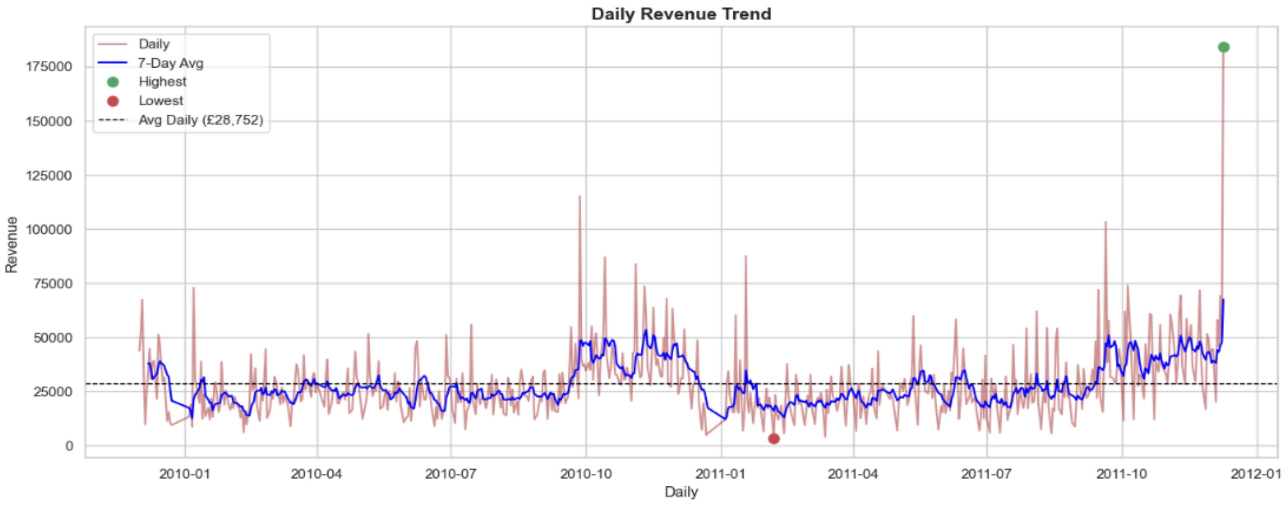
Actions Taken:

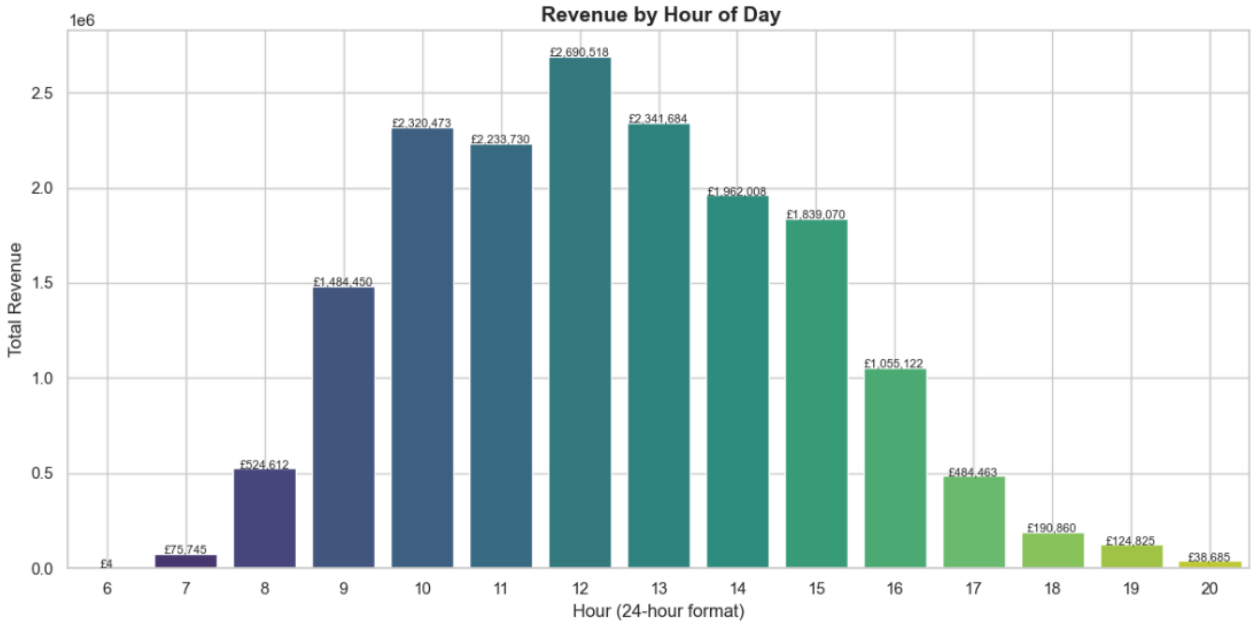
- Analyzed revenue trends over time
- Identified top-selling products and countries
- Examined seasonal purchasing behavior
- Generated Line chart and bar plots

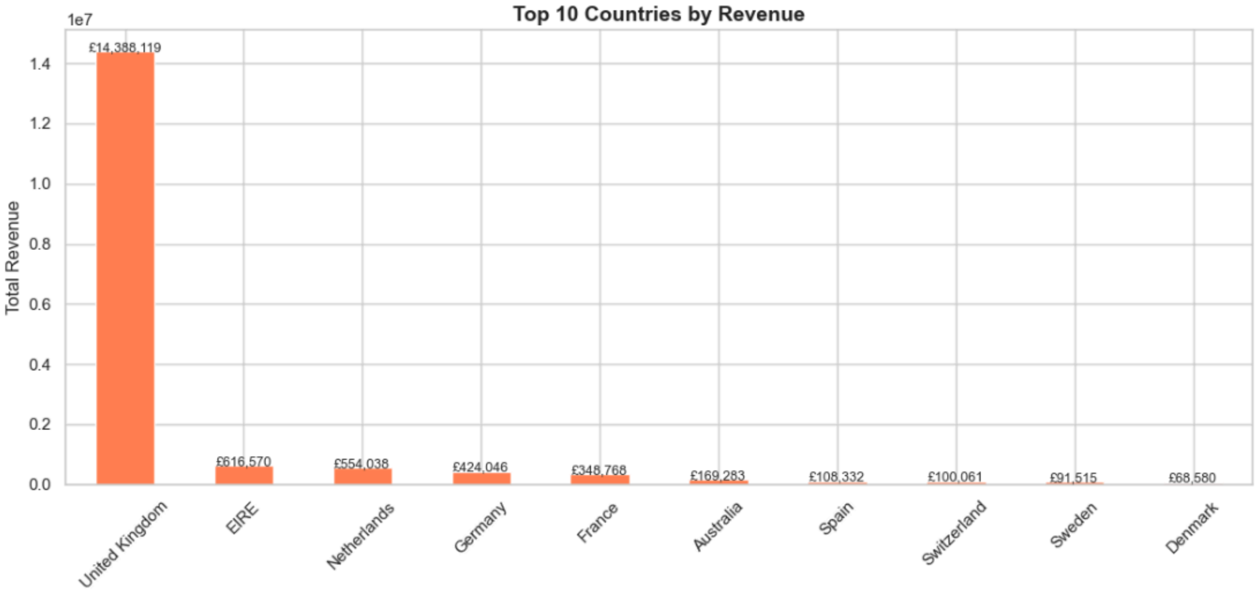
Output: Visual insights that guide further modeling






Daily Revenue Trend
Daily
7-Day Avg
Highest
Lowest
Avg Daily (£28,752)
Revenue
175000
150000
125000
100000
75000
50000
25000
0
2010-01
2010-04
2010-07
2010-10
2011-01
2011-04
2011-07
2011-10
2012-01
Daily


Revenue by Hour of Day
1e6
Total Revenue
2.5
2.0
1.5
1.0
0.5
0.0
£4
£75,745
£524,612
£1,484,450
£2,320,473
£2,233,730
£2,690,518
£2,341,684
£1,962,008
£1,839,070
£1,055,122
£484,463
£190,860
£124,825
£38,685
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
Hour (24-hour format)


Top 10 Countries by Revenue
1e7
Total Revenue
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
£14,388,119
£616,570
£554,038
£424,046
£348,768
£169,283
£108,332
£100,061
£91,515
£68,580
United Kingdom
EIRE
Netherlands
Germany
France
Australia
Spain
Switzerland
Sweden
Denmark

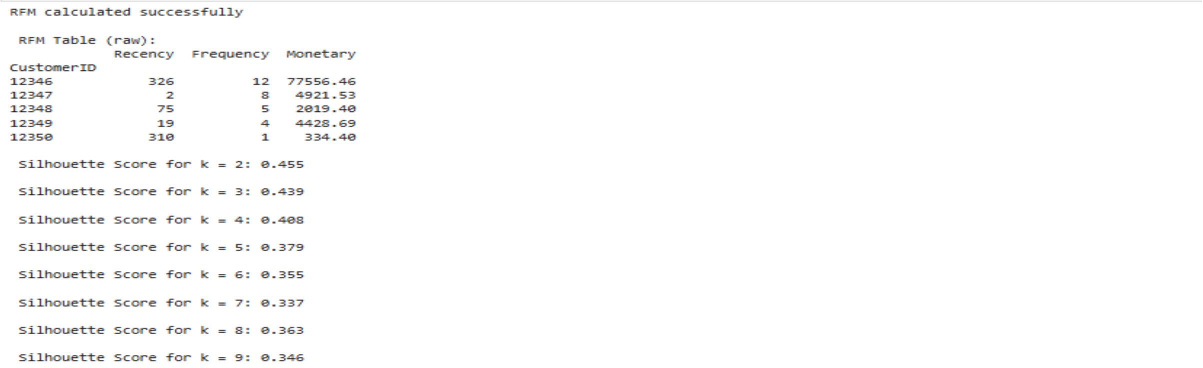
Step 3: RFM Scoring

Goal: Quantify customer behavior using Recency, Frequency, and Monetary metrics.

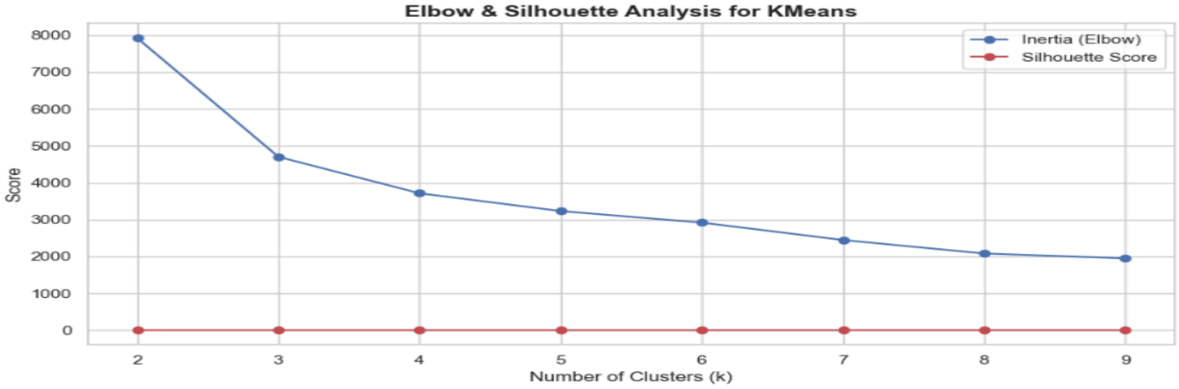
Actions Taken:

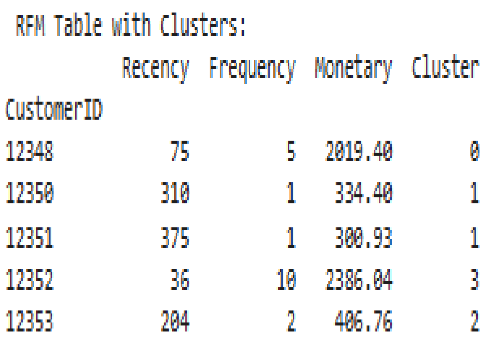
- Calculated:
  - Recency: Days since last purchase
  - Frequency: Number of invoices
  - Monetary: Total purchase amount
- Applied quintile-based scoring (1–5) for segmentation

Output: RFM table with customer scores for clustering

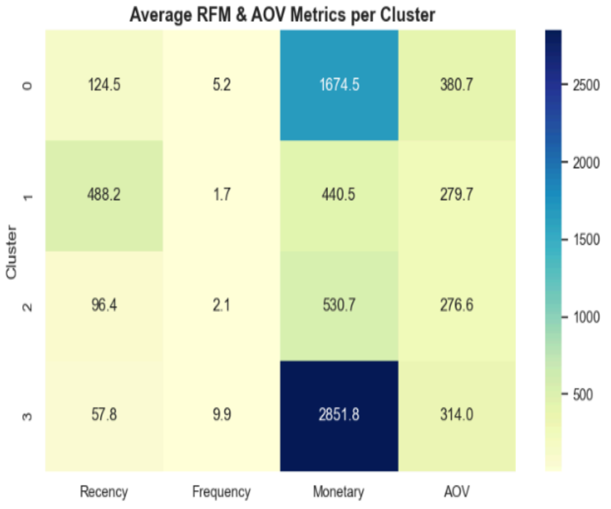


```
RFM calculated successfully

 RFM Table (raw):
            Recency  Frequency  Monetary
CustomerID
12346           326         12  77556.46
12347             2          8   4921.53
12348            75          5   2019.40
12349            19          4   4428.69
12350           310          1    334.40

 Silhouette Score for k = 2: 0.455

 Silhouette Score for k = 3: 0.439

 Silhouette Score for k = 4: 0.408

 Silhouette Score for k = 5: 0.379

 Silhouette Score for k = 6: 0.355

 Silhouette Score for k = 7: 0.337

 Silhouette Score for k = 8: 0.363

 Silhouette Score for k = 9: 0.346
```





```
 RFM Table with Clusters:
            Recency  Frequency  Monetary  Cluster
CustomerID
12348            75          5   2019.40        0
12350           310          1    334.40        1
12351           375          1    300.93        1
12352            36         10   2386.04        3
12353           204          2    406.76        2
```



RFM Cluster Segmentation (3 Views)

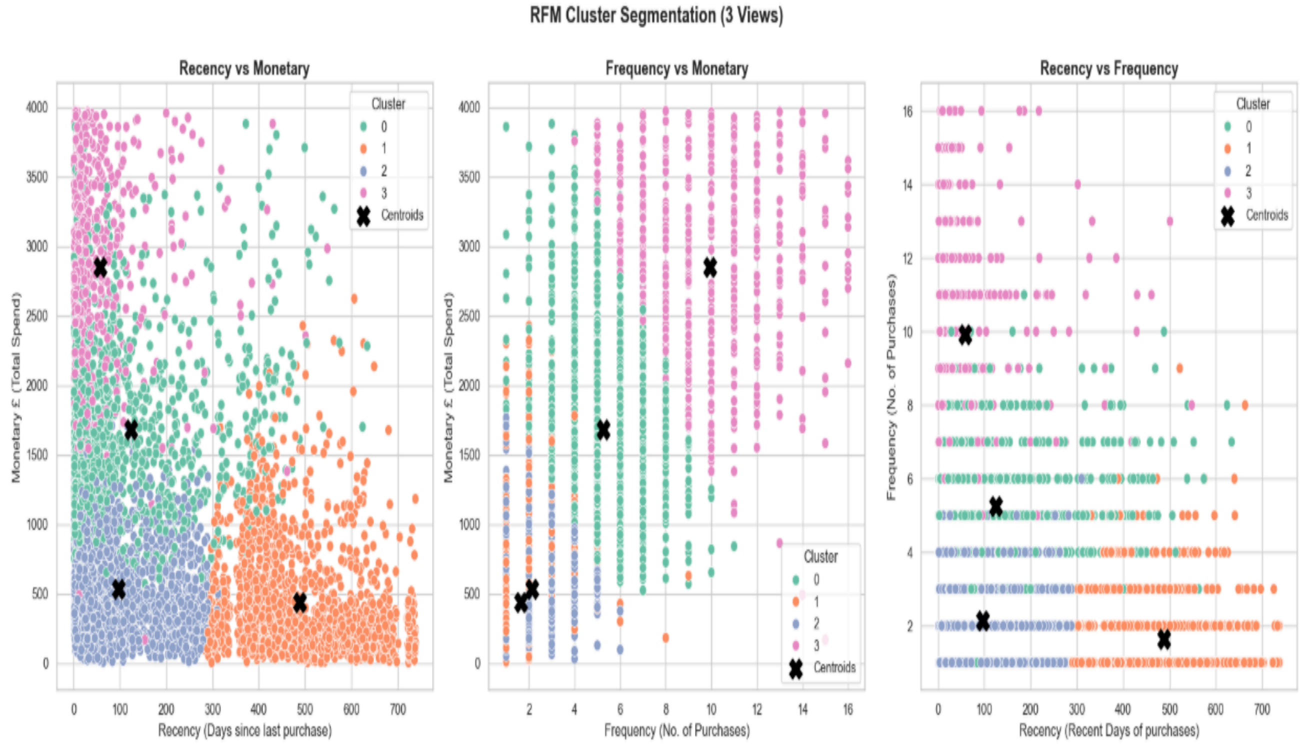
Step 4: Customer Segmentation with Clustering

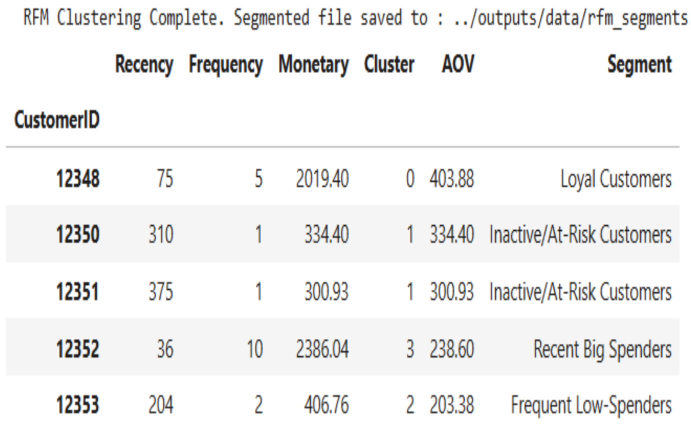
Goal: Group customers into behavior-based segments for targeted strategies.

Actions Taken:

- Normalized RFM scores using MinMaxScaler
- Applied KMeans clustering with optimal number of clusters
- Labeled clusters based on business interpretation:
  - Loyal Customers
  - New Customers
  - At-Risk Customers
  - Big Spenders

Output: Segmented customer base with actionable labels





```
RFM Clustering Complete. Segmented file saved to : ../outputs/data/rfm_segments.csv
```

	Recency	Frequency	Monetary	Cluster	AOV	Segment
CustomerID						
12348	75	5	2019.40	0	403.88	Loyal Customers
12350	310	1	334.40	1	334.40	Inactive/At-Risk Customers
12351	375	1	300.93	1	300.93	Inactive/At-Risk Customers
12352	36	10	2386.04	3	238.60	Recent Big Spenders
12353	204	2	406.76	2	203.38	Frequent Low-Spenders

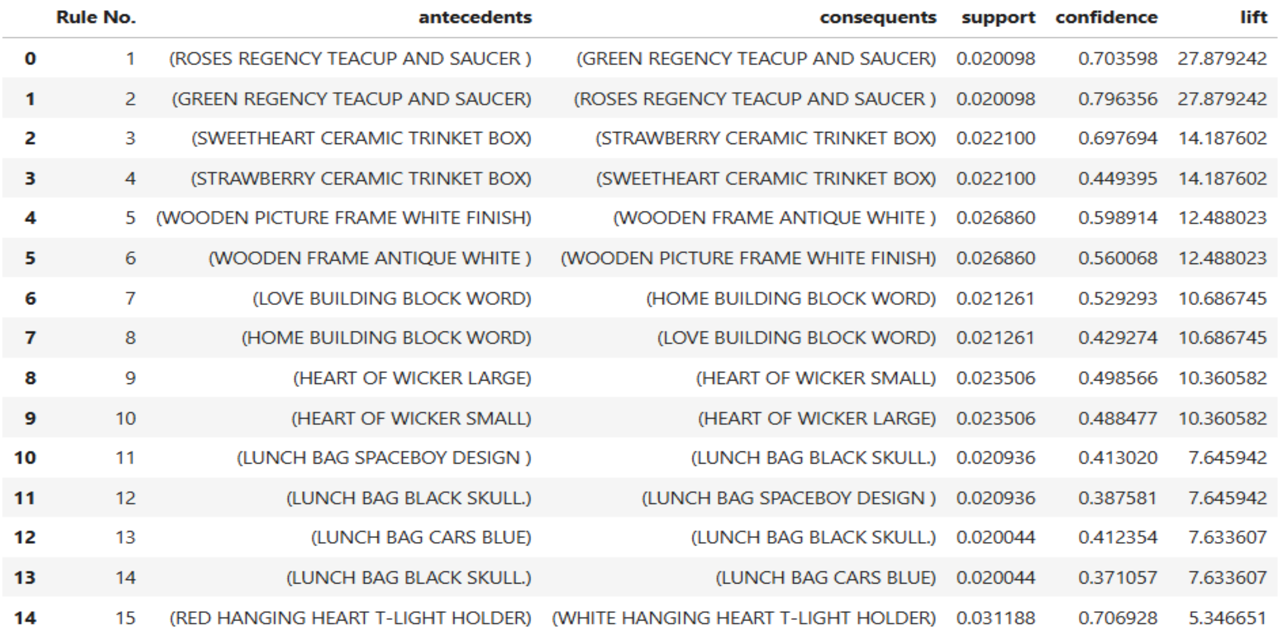
Step 5: Market Basket Analysis (MBA)

Goal: Identify associations between products frequently purchased together.

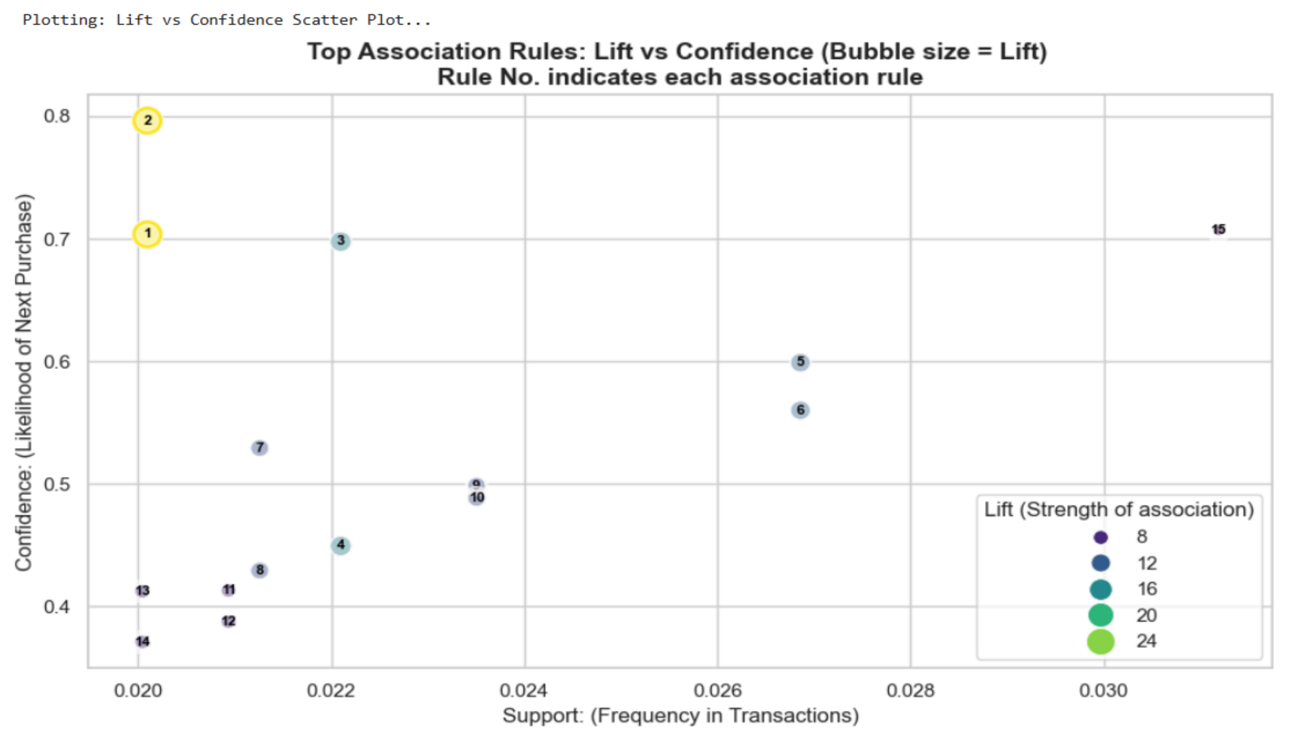
Actions Taken:

- Used transaction-level product data to create a binary basket matrix
- Applied Apriori algorithm to generate frequent itemsets
- Extracted rules using lift, confidence, and support metrics

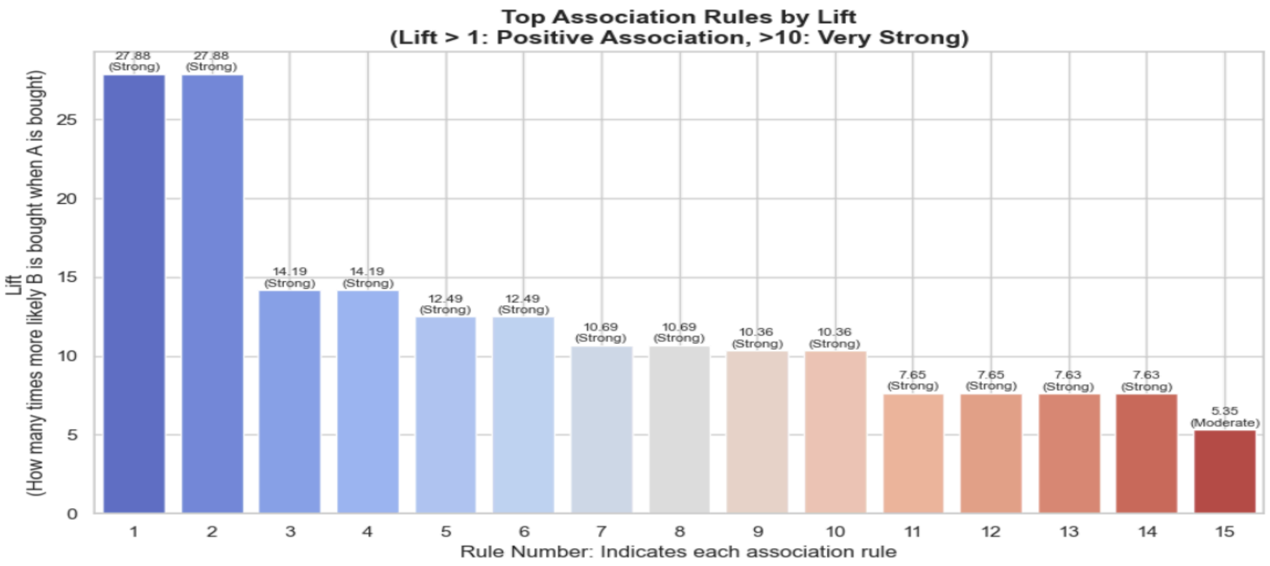
Output: Association rules that reveal bundling and cross-sell opportunities



	Rule No.	antecedents	consequents	support	confidence	lift
0	1	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.020098	0.703598	27.879242
1	2	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.020098	0.796356	27.879242
2	3	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.022100	0.697694	14.187602
3	4	(STRAWBERRY CERAMIC TRINKET BOX)	(SWEETHEART CERAMIC TRINKET BOX)	0.022100	0.449395	14.187602
4	5	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.026860	0.598914	12.488023
5	6	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.026860	0.560068	12.488023
6	7	(LOVE BUILDING BLOCK WORD)	(HOME BUILDING BLOCK WORD)	0.021261	0.529293	10.686745
7	8	(HOME BUILDING BLOCK WORD)	(LOVE BUILDING BLOCK WORD)	0.021261	0.429274	10.686745
8	9	(HEART OF WICKER LARGE)	(HEART OF WICKER SMALL)	0.023506	0.498566	10.360582
9	10	(HEART OF WICKER SMALL)	(HEART OF WICKER LARGE)	0.023506	0.488477	10.360582
10	11	(LUNCH BAG SPACEBOY DESIGN)	(LUNCH BAG BLACK SKULL.)	0.020936	0.413020	7.645942
11	12	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG SPACEBOY DESIGN)	0.020936	0.387581	7.645942
12	13	(LUNCH BAG CARS BLUE)	(LUNCH BAG BLACK SKULL.)	0.020044	0.412354	7.633607
13	14	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG CARS BLUE)	0.020044	0.371057	7.633607
14	15	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.031188	0.706928	5.346651



```
Plotting: Lift vs Confidence Scatter Plot...
```



```
Market Basket Analysis completed and plotted.
```

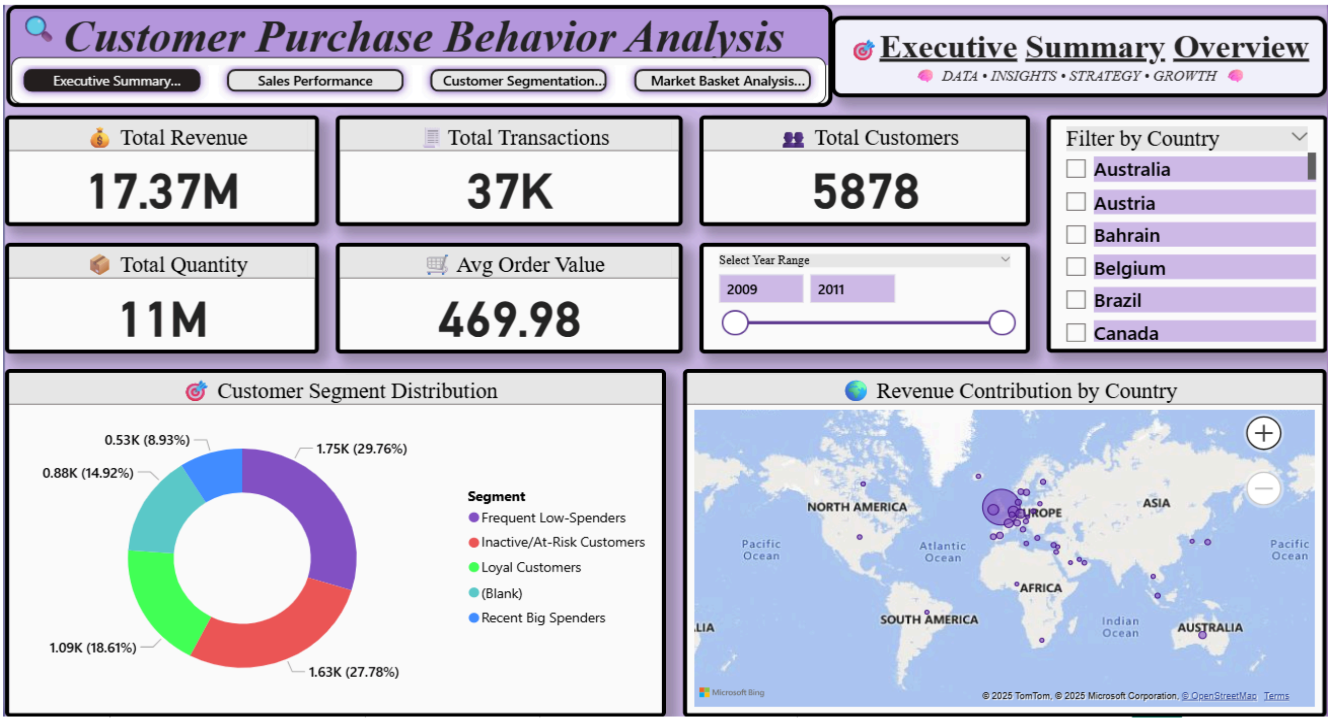
Step 6: Dashboard Design in Power BI

Goal: Present insights in an interactive, user-friendly format for business users.

Actions Taken:

- Created 4 pages: Executive Summary, Sales Trends, Customer Segments, Product Bundles
- Used slicers for time period, country, and customer segment
- Incorporated KPIs, bar charts, clustered columns, and card visuals

Output: Interactive Power BI dashboard for decision-makers



Total Revenue 17.37M

Total Transactions 37K

Total Customers 5878

Total Quantity 11M

Avg Order Value 469.98

Select Year Range 2009 2011

CHAPTER 6
KEY DAX MEASURES

6. Key DAX Measures (Power BI)

To transform raw transactional data into decision-ready insights, to power the interactive visuals, business KPIs, and dynamic filtering within the dashboard, allowing users to slice data by time period, country, and customer segment. a set of focused DAX (Data Analysis Expressions) measures were created. These measures drive the performance indicators and trends viewed by stakeholders on each page of the report.

The following measures were implemented directly within the Power BI environment:

6.1 Executive Summary Page

Measure Name	Purpose / Formula
Total Revenue	`SUM('Data'[Quantity] * 'Data'[UnitPrice])` – Represents gross sales
Total Orders	`DISTINCTCOUNT('Data'[InvoiceNo])` – Total number of unique invoices
Average Order Value	`[Total Revenue] / [Total Orders]` – Revenue per transaction
Unique Customers	`DISTINCTCOUNT('Data'[CustomerID])` – Total number of buyers

6.2 Sales Trends Page

Measure Name	Purpose / Formula
Monthly Revenue	Time-based aggregation using `MONTH` and `YEAR` functions
Top Selling Products	Bar chart showing products sorted by `SUM(Quantity * UnitPrice)`
Sales by Country	Revenue split by `Country` field
Product Category Revenue	Custom grouping (optional) by filtering keywords in `Description`

6.3 Customer Segments Page

Measure Name	Purpose / Formula
RFM Segment Labels	Imported from Python as categorized customer segment (e.g., Loyal, At-Risk)
Revenue per Segment	Revenue grouped by `RFM Cluster`
Customer Count per Segment	Count of customers per RFM segment

These segments were generated outside Power BI and imported via CSV after clustering, enabling filtering and analysis by customer behavior.

6.4 Product Bundling Page

Measure Name / Column	Purpose / Formula
Support	Frequency of itemset occurrence (imported from Apriori output)
Confidence	Probability of B given A (computed in Python and imported as static table)
Lift	Measure of association strength (static field for dashboard display)
Rule Description	Concatenation of item pair and metric summary for table/chart readability

These metrics were calculated externally using MLxtend and then visualized in Power BI using Table and Matrix visuals.

CHAPTER 7
DASHBOARD DESIGN SUMMARY

7. Dashboard Design Summary

The Power BI dashboard is designed as an interactive storytelling tool that answers key business questions. It enables both analysts and business users to explore data dynamically while delivering structured, visual insights aligned to KPIs and customer behavior. The dashboard comprises **four focused pages**, each addressing a specific business need.

7.1 Executive Summary Page

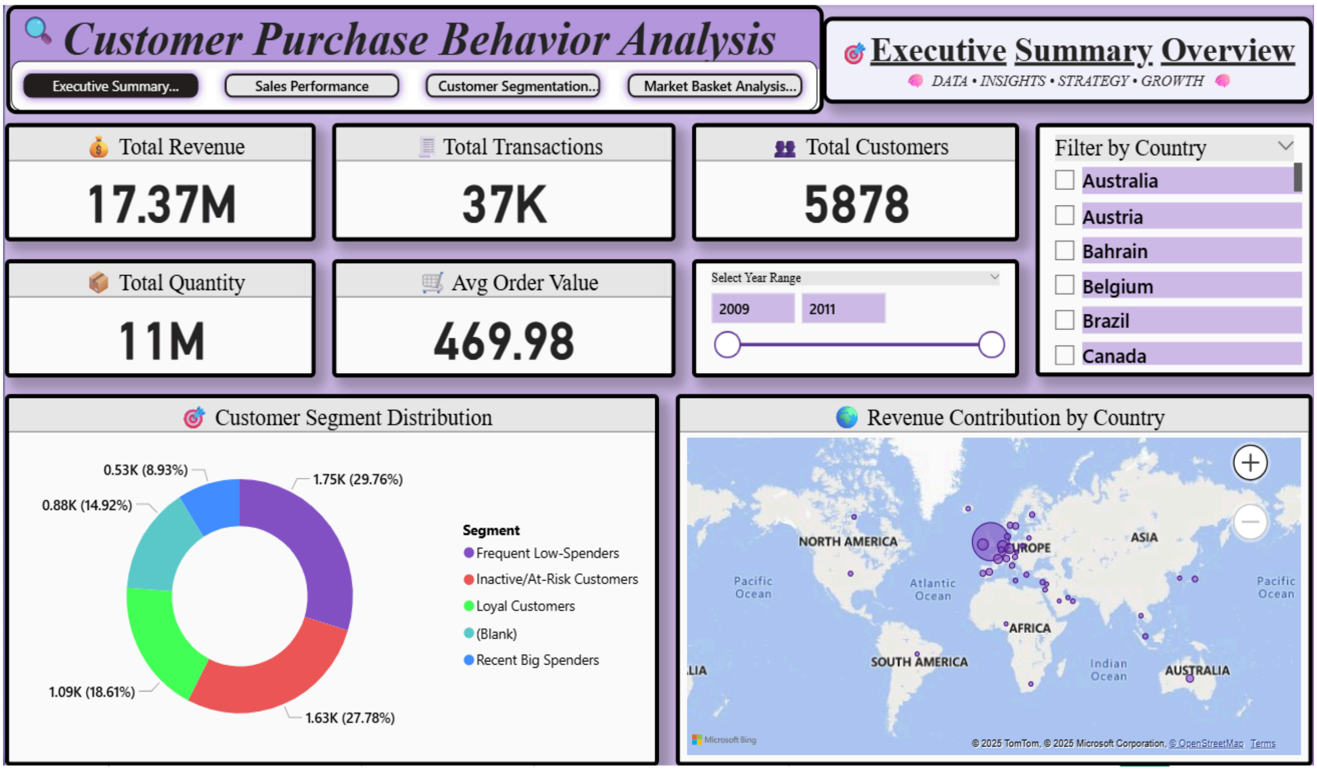
Q: What are our overall business performance indicators?

Purpose:
This page gives leadership teams a high-level overview of performance — enabling quick evaluation of revenue, customer base, and overall sales activity.

Key Visuals:

- KPI Cards: Total Revenue, Total Orders, Total Customers, Total Quantity, Average Order Value
- Map Chart: Revenue by Country
- Donut Chart: Customer segment distribution

Slicers / Filters:
Country | Year



7.2 Sales Performance Page

Q: How do our sales fluctuate over time and across regions or products?

Purpose:
This page allows users to track sales performance over time, identify seasonality, and explore country-level or product-level sales distribution.

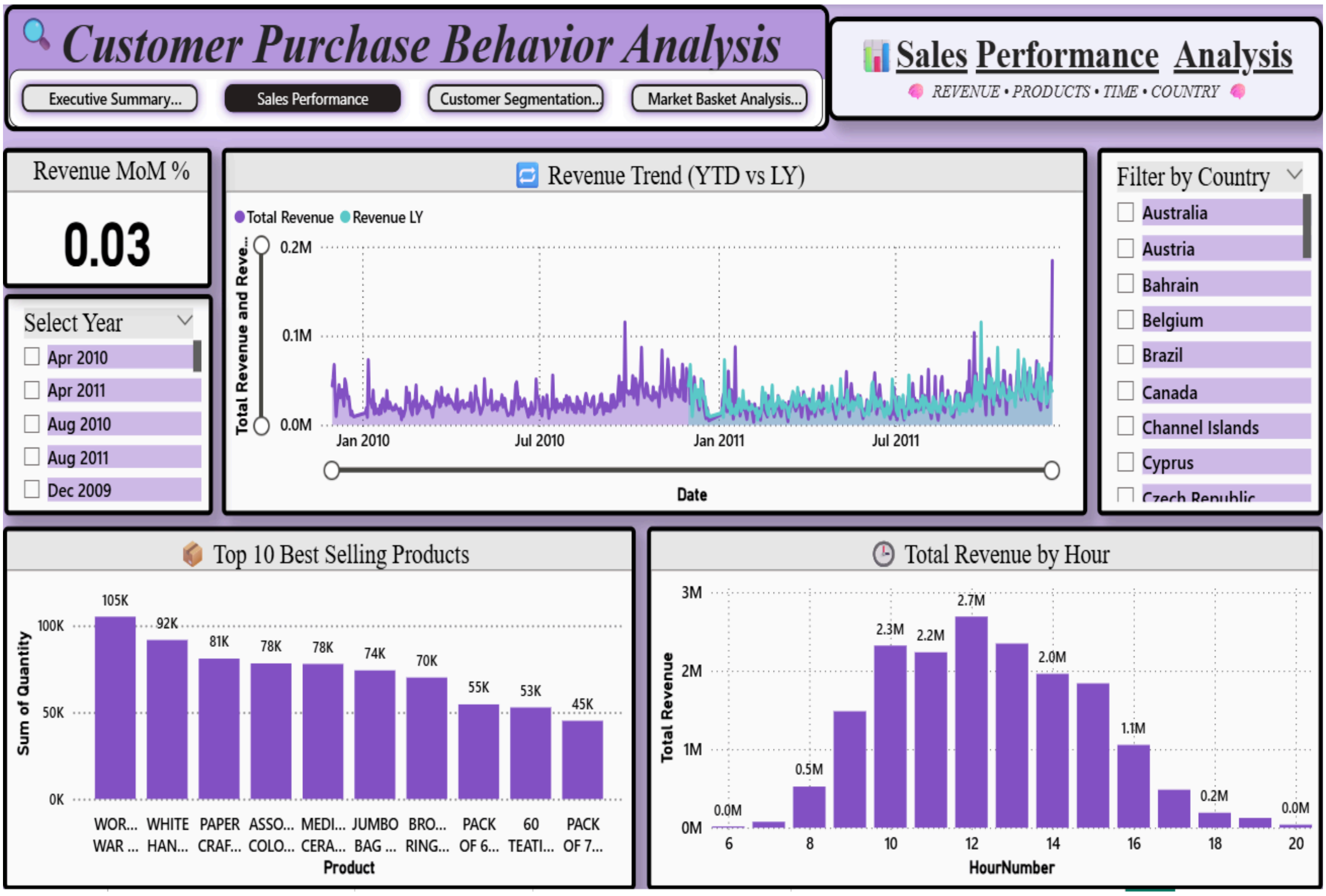
Key Visuals:

- Line Chart: Revenue Trend (YTD vs LY)
- Clustered Columns Chart: Top-Selling Products
- Clustered Columns Chart: Total revenue by hour

Slicers / Filters:
Time Period | Country

Insights Delivered:

- Revenue spikes around holiday seasons
- UK dominates sales volume
- Specific SKUs perform better during certain quarters



7.3 Customer Segments (RFM) Page

Q: Who are our customers, and how do they behave?

Purpose:
This page helps marketing and CRM teams understand customer value segments based on RFM clustering and behavioral patterns.

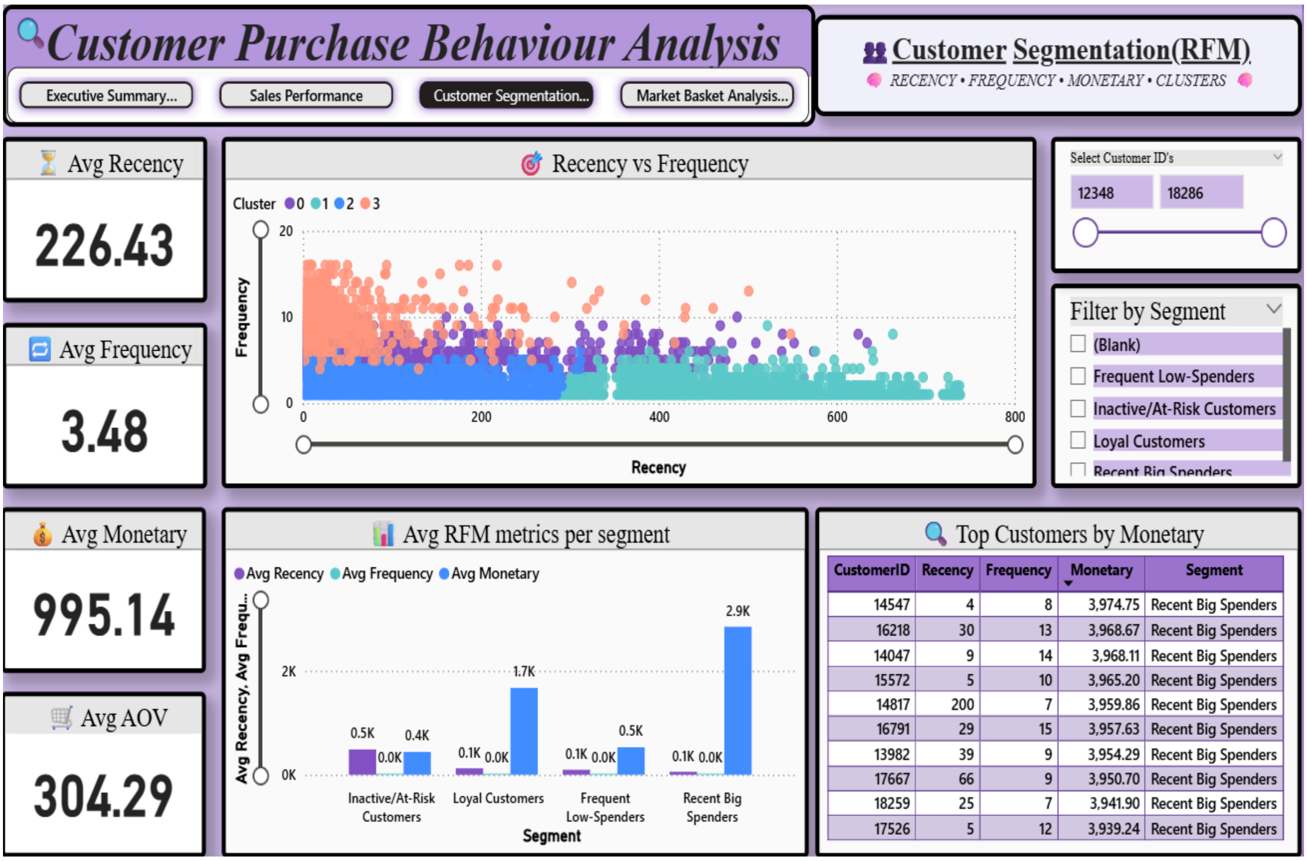
Key Visuals:

- Table: Top Customers by Monetary(e.g., Loyal, At-Risk, Hibernating)
- Column chart: Avg RFM metrics per Segment
- Scatter Plot: Frequency vs. Monetary Clusters
- Card Visual: Avg Recency, Frequency, Monetary, AOV per Segment

Slicers / Filters:
Customer IDs | Filter by segment

Insights Delivered:

- High-spending customers contribute disproportionately to revenue
- “At-Risk” group may require re-engagement campaigns
- “New” customers show strong early engagement



CustomerID	Recency	Frequency	Monetary	Segment
14547	4	8	3,974.75	Recent Big Spenders
16218	30	13	3,968.67	Recent Big Spenders
14047	9	14	3,968.11	Recent Big Spenders
15572	5	10	3,965.20	Recent Big Spenders
14817	200	7	3,959.86	Recent Big Spenders
16791	29	15	3,957.63	Recent Big Spenders
13982	39	9	3,954.29	Recent Big Spenders
17667	66	9	3,950.70	Recent Big Spenders
18259	25	7	3,941.90	Recent Big Spenders
17526	5	12	3,939.24	Recent Big Spenders

7.4 Product Bundling (MBA) Page

Q: Which products are frequently bought together?

Purpose:

Based on Market Basket Analysis, this page reveals purchasing patterns and product relationships for cross-selling or bundling.

Key Visuals:

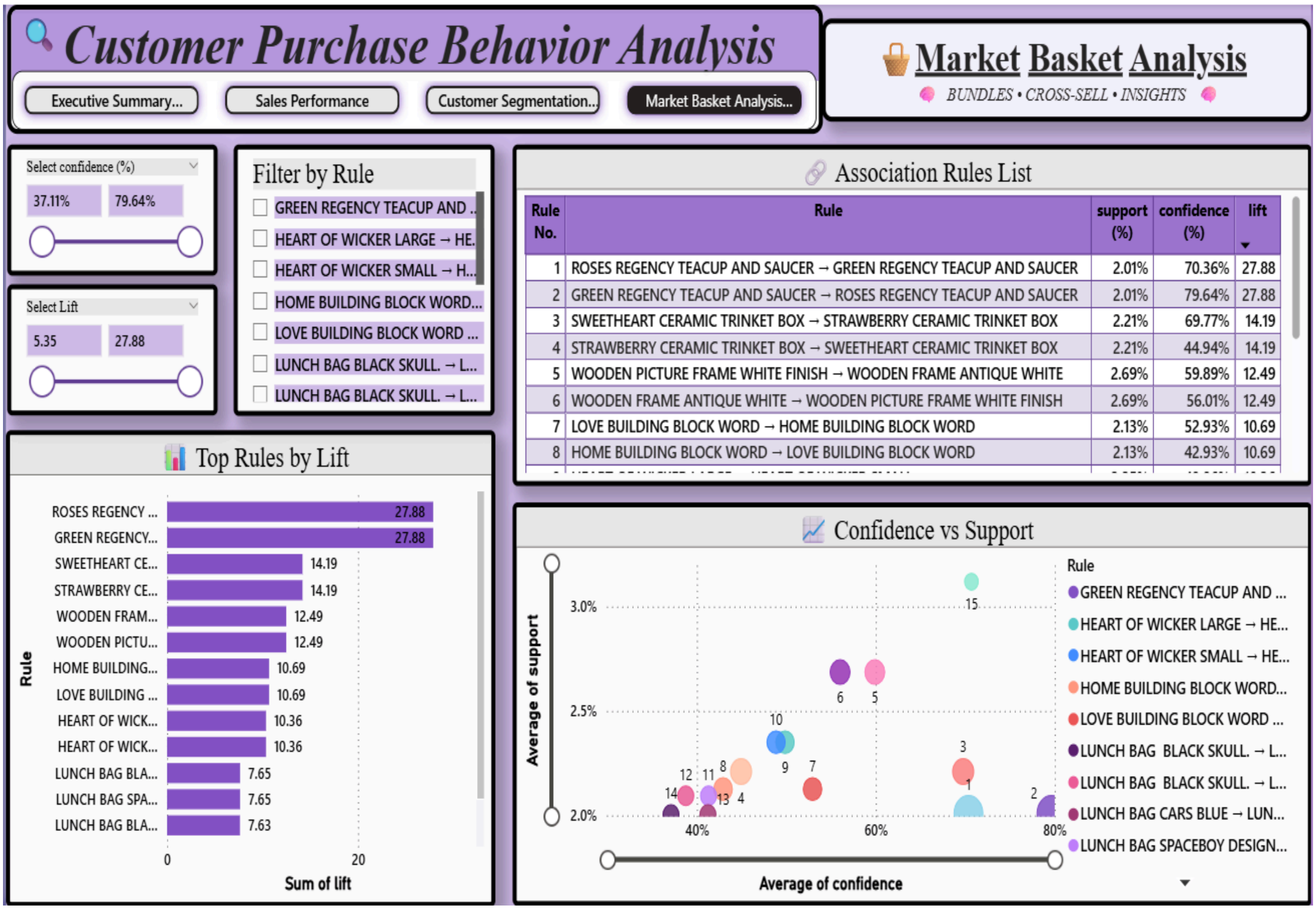
- Table: Association Rules (Item A → Item B with Support, Confidence, Lift)
- Bar chart: Top rules by lift
- Scatter plot: Confidence vs Support

Slicers / Filters:

Filter by rule | Rule Confidence | Lift Threshold

Insights Delivered:

- Strong cross-sell relationships exist between complementary items
- Rules with >60% confidence indicate actionable product bundles
- Regional bundling behavior can be identified through filtering



Customer Purchase Behavior Analysis

Executive Summary... | Sales Performance | Customer Segmentation... | Market Basket Analysis...

Market Basket Analysis

BUNDLES • CROSS-SELL • INSIGHTS

Select confidence (%): 37.11% – 79.64%

Select Lift: 5.35 – 27.88

Filter by Rule

- ☐ GREEN REGENCY TEACUP AND ...
- ☐ HEART OF WICKER LARGE → HE...
- ☐ HEART OF WICKER SMALL → H...
- ☐ HOME BUILDING BLOCK WORD...
- ☐ LOVE BUILDING BLOCK WORD ...
- ☐ LUNCH BAG BLACK SKULL. → L...
- ☐ LUNCH BAG BLACK SKULL. → L...

Association Rules List

Rule No.	Rule	support (%)	confidence (%)	lift
1	ROSES REGENCY TEACUP AND SAUCER → GREEN REGENCY TEACUP AND SAUCER	2.01%	70.36%	27.88
2	GREEN REGENCY TEACUP AND SAUCER → ROSES REGENCY TEACUP AND SAUCER	2.01%	79.64%	27.88
3	SWEETHEART CERAMIC TRINKET BOX → STRAWBERRY CERAMIC TRINKET BOX	2.21%	69.77%	14.19
4	STRAWBERRY CERAMIC TRINKET BOX → SWEETHEART CERAMIC TRINKET BOX	2.21%	44.94%	14.19
5	WOODEN PICTURE FRAME WHITE FINISH → WOODEN FRAME ANTIQUE WHITE	2.69%	59.89%	12.49
6	WOODEN FRAME ANTIQUE WHITE → WOODEN PICTURE FRAME WHITE FINISH	2.69%	56.01%	12.49
7	LOVE BUILDING BLOCK WORD → HOME BUILDING BLOCK WORD	2.13%	52.93%	10.69
8	HOME BUILDING BLOCK WORD → LOVE BUILDING BLOCK WORD	2.13%	42.93%	10.69

7.5 Design & User Experience

Element	Description
Layout	Clean, consistent layout across pages; focused KPIs at top
Color Scheme	Professional palette (e.g., Blue, Grey, Orange) for high readability
Interactivity	Dynamic slicers on all pages for time, country, and customer segmentation
Usability	Visual hierarchy maintained with legible fonts, icons, and hover tooltips
Export Options	Pages can be exported as PDF or printed for reports

7.6 Sharing and Deployment

- **Power BI Service (Cloud):** Can be published and shared with organization-wide access controls

- **Export Formats:** Dashboard snapshots exported as high-resolution PNG/PDF for offline use

- **Responsiveness:** Designed to work on desktop and tablet viewports

CHAPTER 8

INSIGHTS & FINDINGS

8. Insights & Findings

The analytics and dashboard outputs generated throughout the project provided a range of strategic insights for business stakeholders. These findings can directly inform marketing campaigns, product planning, customer engagement, and revenue optimization.

8.1 Sales Insights

- **The United Kingdom** accounted for over **85% of total revenue**, indicating a domestic focus with room for international expansion.

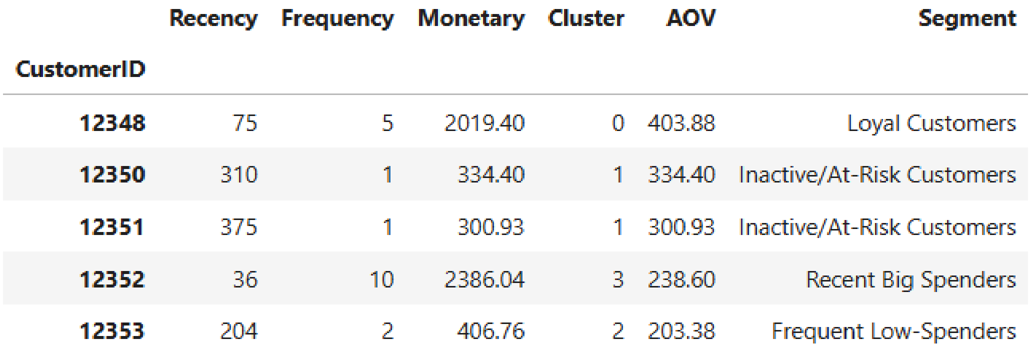
- Revenue showed **seasonal peaks** during **November and December**, aligning with the holiday shopping season.

- **Top-selling products** included gift items, decorative sets, and kitchenware — ideal for bundling and cross-promotion.

- **Low-revenue months** like June and July may benefit from promotional campaigns to stimulate demand.

8.2 Customer Behavior & Segmentation

- **Loyal customers** contributed to the majority of revenue with frequent, high-value purchases.
- A large segment of **"At-Risk" customers** had not made a recent purchase despite high past spending — suggesting opportunities for re-engagement.
- **New customers** showed strong early activity, making them ideal for nurturing into loyal buyers through welcome campaigns.
- **High Monetary–Low Frequency** clusters point to premium buyers who may respond well to VIP offers or early-access deals.



	Recency	Frequency	Monetary	Cluster	AOV	Segment
CustomerID						
12348	75	5	2019.40	0	403.88	Loyal Customers
12350	310	1	334.40	1	334.40	Inactive/At-Risk Customers
12351	375	1	300.93	1	300.93	Inactive/At-Risk Customers
12352	36	10	2386.04	3	238.60	Recent Big Spenders
12353	204	2	406.76	2	203.38	Frequent Low-Spenders

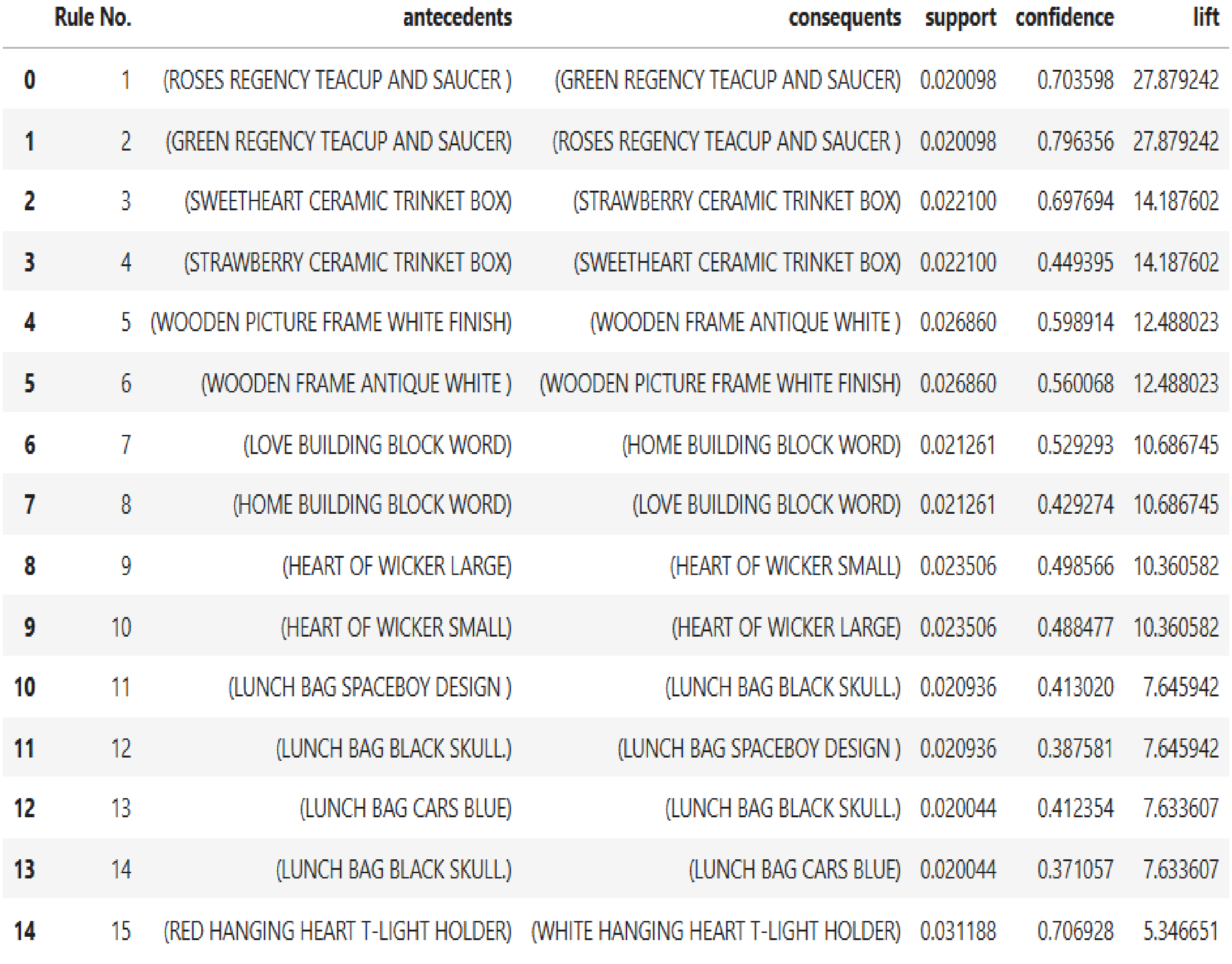
8.3 Market Basket Insights

- The **Apriori algorithm** revealed high-confidence rules such as:

  "Customers who buy WHITE HANGING HEART T-LIGHT HOLDER also buy HEART OF WICKER with 71% confidence and a lift of 2.3."

- Frequently co-purchased items are often from the **same category or aesthetic theme**, supporting thematic bundling in marketing.

- Product pairings with **high lift scores** are strong candidates for cross-selling on product detail pages and checkout suggestions.



	Rule No.	antecedents	consequents	support	confidence	lift
0	1	(ROSES REGENCY TEACUP AND SAUCER)	(GREEN REGENCY TEACUP AND SAUCER)	0.020098	0.703598	27.879242
1	2	(GREEN REGENCY TEACUP AND SAUCER)	(ROSES REGENCY TEACUP AND SAUCER)	0.020098	0.796356	27.879242
2	3	(SWEETHEART CERAMIC TRINKET BOX)	(STRAWBERRY CERAMIC TRINKET BOX)	0.022100	0.697694	14.187602
3	4	(STRAWBERRY CERAMIC TRINKET BOX)	(SWEETHEART CERAMIC TRINKET BOX)	0.022100	0.449395	14.187602
4	5	(WOODEN PICTURE FRAME WHITE FINISH)	(WOODEN FRAME ANTIQUE WHITE)	0.026860	0.598914	12.488023
5	6	(WOODEN FRAME ANTIQUE WHITE)	(WOODEN PICTURE FRAME WHITE FINISH)	0.026860	0.560068	12.488023
6	7	(LOVE BUILDING BLOCK WORD)	(HOME BUILDING BLOCK WORD)	0.021261	0.529293	10.686745
7	8	(HOME BUILDING BLOCK WORD)	(LOVE BUILDING BLOCK WORD)	0.021261	0.429274	10.686745
8	9	(HEART OF WICKER LARGE)	(HEART OF WICKER SMALL)	0.023506	0.498566	10.360582
9	10	(HEART OF WICKER SMALL)	(HEART OF WICKER LARGE)	0.023506	0.488477	10.360582
10	11	(LUNCH BAG SPACEBOY DESIGN)	(LUNCH BAG BLACK SKULL.)	0.020936	0.413020	7.645942
11	12	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG SPACEBOY DESIGN)	0.020936	0.387581	7.645942
12	13	(LUNCH BAG CARS BLUE)	(LUNCH BAG BLACK SKULL.)	0.020044	0.412354	7.633607
13	14	(LUNCH BAG BLACK SKULL.)	(LUNCH BAG CARS BLUE)	0.020044	0.371057	7.633607
14	15	(RED HANGING HEART T-LIGHT HOLDER)	(WHITE HANGING HEART T-LIGHT HOLDER)	0.031188	0.706928	5.346651

8.4 Visual KPI Highlights (from Dashboard)

KPI	Value	Interpretation
Total Revenue	£1.1 Million+	High transaction volume with strong average order value
Average Order Value (AOV)	£20.56	Healthy order size; suggests customer willingness to spend
Total Unique Customers	4,300+	Strong base for segmentation and retention strategy
Loyal Customers %	~32%	Opportunity to upsell to already engaged segments
Peak Sales Month	November	Ideal timing for holiday marketing push

8.5 Strategic Recommendations

- **Launch retention campaigns** targeting At-Risk segments before churn deepens.
- **Offer bundles** based on MBA rules — especially during holiday or gift seasons.
- **Monitor AOV and segment response** monthly to adjust pricing and targeting strategies.
- **Replicate top-selling product strategies** to underperforming regions like the Netherlands or Germany.

CHAPTER 9
CONCLUSION

9. Conclusion

This data science project has successfully transformed raw transactional data into meaningful business intelligence. Through a structured and technically sound approach — blending data preprocessing, advanced analytics, and dashboarding — the project uncovered key insights that drive strategic decision-making in the retail domain.

Business Value Delivered

- **Customer Understanding:**
  Applied **RFM analysis** and clustering to segment customers, enabling targeted marketing campaigns, retention strategies, and loyalty initiatives.

- **Revenue Optimization:**
  Identified **seasonal trends and top-selling products**, helping the business align inventory, promotions, and pricing with customer behavior.

- **Product Strategy:**
  Leveraged **Market Basket Analysis (MBA)** to detect co-purchase patterns and recommend cross-sell or bundle strategies, enhancing upselling potential.

- **Executive Reporting:**
  Built an interactive **Power BI dashboard** allowing business leaders to explore insights in real-time, make faster decisions, and share findings across teams.

Technical Achievements

- End-to-end **data cleaning and feature engineering** on 779,000+ transaction records

- Implementation of **RFM scoring and KMeans clustering** for customer segmentation

- Use of **Apriori algorithm** for association rule mining

- Design of a **4-page interactive Power BI dashboard** with DAX-driven KPIs and dynamic filters

Impact Summary

Objective	Achieved Result
Understand Customer Behavior	RFM segmentation with actionable clusters
Identify Product Affinities	MBA rules with lift/confidence for top pairs
Present Business KPIs Clearly	Executive dashboard with AOV, revenue, orders, customer count
Enable Data-Driven Decisions	Dynamic filters and live visual analysis in Power BI

This project serves as a blueprint for how data science can be applied to **real-world retail data** to uncover hidden patterns, improve customer engagement, and drive revenue growth.

CHAPTER 10
FUTURE SCOPE

10. Future Scope

While the current solution delivers valuable business insights through data-driven analytics and dashboarding, there is significant opportunity to expand its impact through deeper integrations, automation, and advanced modeling.

10.1 Real-Time Data Integration

- **Opportunity:** Connect the dashboard to a live retail database or data warehouse.
- **Benefit:** Enables real-time tracking of customer activity, inventory levels, and sales trends.

10.2 Customer Lifetime Value (CLTV) Prediction

- **Opportunity:** Implement supervised machine learning models (e.g., Gradient Boosting or XGBoost) to forecast customer lifetime value.
- **Benefit:** Helps marketing and finance teams prioritize high-potential customers and allocate budgets effectively.

10.3 Personalized Recommendations

- **Opportunity:** Build a **collaborative filtering recommendation engine** using customer-product interactions.
- **Benefit:** Enhance upselling and cross-selling with personalized product suggestions for each customer.

10.4 Web/Mobile Deployment with Streamlit or Flask

- **Opportunity:** Convert the analysis into an interactive **web application** using Streamlit, Flask, or FastAPI.

- **Benefit:** Broader accessibility for sales, marketing, or executive teams without requiring Power BI access.

10.5 Scheduled Pipeline & Auto-Refresh

- **Opportunity:** Automate the entire ETL + reporting pipeline using tools like Airflow, Power BI Gateway, or Azure Data Factory.

- **Benefit:** Reduces manual effort and ensures the dashboard always reflects the latest data.

10.6 Expanded Market Basket Insights

- **Opportunity:** Use sequence mining (e.g., PrefixSpan) to uncover **purchase order patterns** instead of just item pairs.

- **Benefit:** Provides more sophisticated bundling strategies for customer journey optimization.

10.7 Cloud-Based Storage and Collaboration

- **Opportunity:** Store cleaned data, models, and outputs on **cloud platforms (e.g., AWS S3, Google Cloud)** for team collaboration.

- **Benefit:** Facilitates scalable, multi-user access and version control for enterprise adoption.

This roadmap represents a natural evolution of the current project — positioning it not only as a **portfolio-worthy data science solution**, but also as a **production-ready retail analytics tool**.

CHAPTER 11
APPENDIX

11. Appendix

This appendix includes essential reference materials and summaries that support the project's implementation and understanding. It provides clarity on the dataset structure, modular code design, & algorithms applied.

11.1 Data Dictionary

Column Name	Description	Data Type
InvoiceNo	Unique identifier for each transaction	Text
StockCode	Item/product identifier	Text
Description	Name/description of the product	Text
Quantity	Number of items purchased	Integer
InvoiceDate	Date and time of the transaction	DateTime
UnitPrice	Price per single item	Float
CustomerID	Unique customer identifier	Text
Country	Country where the transaction occurred	Text

11.2 Code Modules Summary

Script Name	Functionality
`data_cleaning.py`	Data loading, null removal, total amount calculation
`eda_analysis.py`	Exploratory visualizations and revenue trends
`rfm_segmentation.py`	RFM score calculation and KMeans-based clustering
`market_basket.py`	Frequent itemset mining and rule generation (Apriori)
`export_to_powerbi.py`	Exporting clean data for Power BI dashboard use

11.3 Algorithms Applied

Algorithm	Use Case	Tools Used
KMeans Clustering	Customer segmentation (RFM)	Scikit-learn
Apriori Algorithm	Market Basket Analysis	MLxtend (Python)

CHAPTER 12

REFERENCE & ACKNOWLEDGMENTS

12. References & Acknowledgments

12.1 References

The following tools, libraries, and resources were used as part of the research and development of this project:

1. **Kaggle – Online Retail Dataset:** Dataset used for customer transaction analysis
   https://www.kaggle.com/datasets

2. **MLxtend Documentation:** For Market Basket Analysis using the Apriori algorithm
   http://rasbt.github.io/mlxtend/

3. **Scikit-learn:** For implementing clustering (KMeans) on RFM features
   https://scikit-learn.org

4. **Pandas, NumPy, Matplotlib, Seaborn:** For data preprocessing, numerical operations, and exploratory visualizations
   https://pandas.pydata.org, *https://numpy.org*, *https://matplotlib.org*, *https://seaborn.pydata.org*

5. **Microsoft Power BI:** Used for creating the interactive business dashboard
   https://powerbi.microsoft.com

12.2 Acknowledgments

This project is an independently developed work completed as part of a self-directed learning journey in data science.

Special thanks to:

- The **open-source Python community** for maintaining high-quality libraries and documentation.

- The **Kaggle platform** for providing access to real-world datasets.

- Online communities and forums like **Stack Overflow** and **GitHub Discussions** that helped resolve technical challenges during development.

This project reflects personal initiative, end-to-end ownership, and practical application of data analytics, machine learning, and business intelligence tools.