Ghulam Ishaq Khan Institute of Engineering Sciences and Technology, Topi, Pakistan.

Muhammad Adeel	2022331
Nauman Ali Murad	2022479

Course: DS341 – Data Mining
Instructor: Dr. Ayaz Umer, Assistant Professor
Emails: ayaz.umer, u2022331, u2022479 {**@giki.edu.pk**}

Data of Submission: 10-May-2025
Semester Project Report

Retail Customer Analytics Using Advanced Data Mining Techniques

Project Overview

This project investigates transactional data from a UK-based online retail company spanning December 2009 to December 2011. The company specializes in unique giftware, mainly serving wholesale clients. The dataset, with over 1 million records, provided a strong basis for deep analysis.

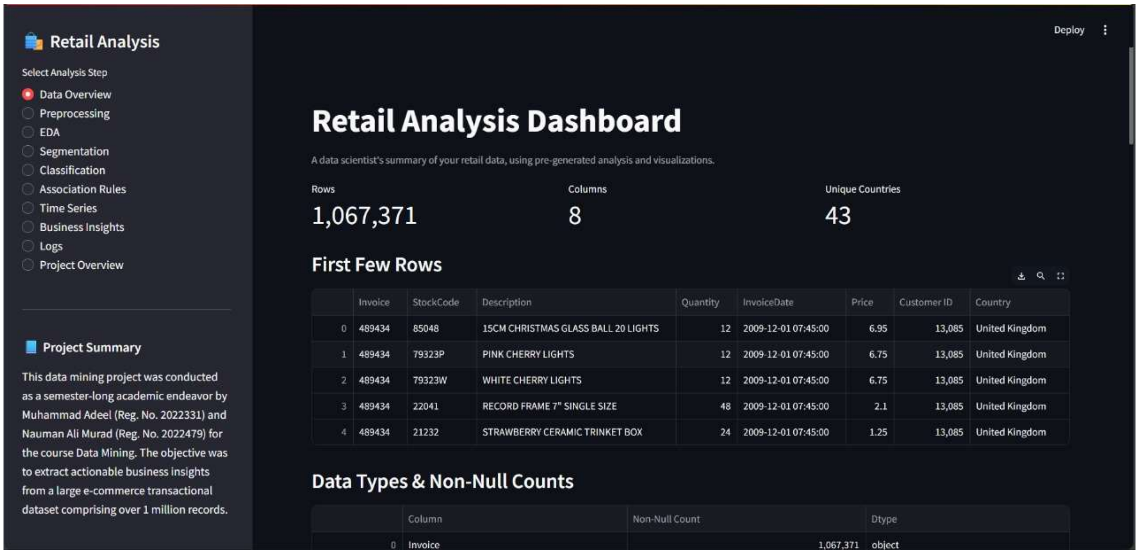
We began with data cleaning and preprocessing to ensure accuracy, followed by exploratory data analysis (EDA) to identify sales trends, seasonal patterns, and customer behavior. Using RFM segmentation, we categorized customers by purchasing habits, enabling targeted marketing insights.

Predictive models were developed to forecast sales trends and customer churn, while market basket analysis revealed product bundling opportunities. The project concluded with actionable business recommendations focused on improving customer retention, optimizing operations, and driving profitability through data-driven strategies.

Dataset Overview

- **Source:** Online Retail II Dataset
- **Duration:** December 1, 2009 – December 9, 2011
- **Number of Records:** 1,067,371 transactions
- **Attributes:**

Attribute	Description
InvoiceNo	Unique invoice number; 'C' indicates cancellation
StockCode	Unique product/item code
Description	Product name
Quantity	Number of items purchased per transaction
InvoiceDate	Date and time of transaction
UnitPrice	Price per unit in sterling (£)
CustomerID	Unique customer identifier
Country	Customer’s country of residence



Project Summary

This data mining project was conducted as a semester-long academic endeavor by Muhammad Adeel (Reg. No. 2022331) and Nauman Ali Murad (Reg. No. 2022479) for the course Data Mining. The objective was to extract actionable business insights from a large e-commerce transactional dataset comprising over 1 million records.

Retail Analysis Dashboard

A data scientist's summary of your retail data, using pre-generated analysis and visualizations.

Rows 1,067,371 | Columns 8 | Unique Countries 43

First Few Rows

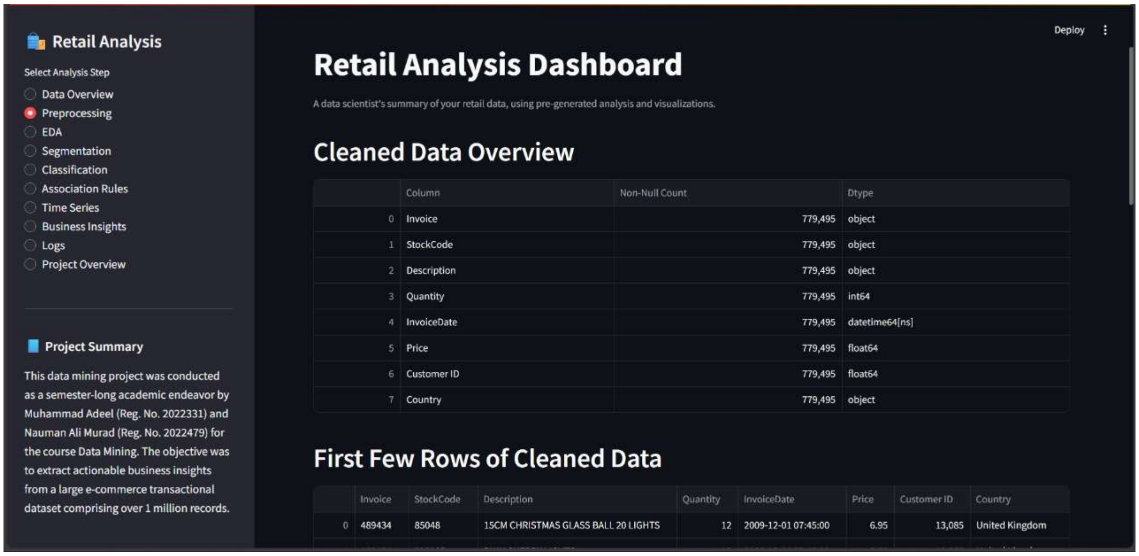
	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13,085	United Kingdom
1	489434	79323P	PINK CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13,085	United Kingdom
2	489434	79323W	WHITE CHERRY LIGHTS	12	2009-12-01 07:45:00	6.75	13,085	United Kingdom
3	489434	22041	RECORD FRAME 7" SINGLE SIZE	48	2009-12-01 07:45:00	2.1	13,085	United Kingdom
4	489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	2009-12-01 07:45:00	1.25	13,085	United Kingdom

Data Types & Non-Null Counts

	Column	Non-Null Count	Dtype
0	Invoice	1,067,371	object

1. Data Preprocessing

1. **Cleaning:**
    - Removed cancelled transactions (InvoiceNo starting with 'C')
    - Dropped rows with missing CustomerID and Description
    - Removed duplicate entries
    - Converted InvoiceDate to proper datetime format
2. **Feature Engineering:**
    - Created a new feature: TotalPrice = Quantity × UnitPrice
    - Extracted Year, Month, Day, Weekday, and Hour from InvoiceDate



Project Summary

This data mining project was conducted as a semester-long academic endeavor by Muhammad Adeel (Reg. No. 2022331) and Nauman Ali Murad (Reg. No. 2022479) for the course Data Mining. The objective was to extract actionable business insights from a large e-commerce transactional dataset comprising over 1 million records.

Retail Analysis Dashboard

A data scientist's summary of your retail data, using pre-generated analysis and visualizations.

Cleaned Data Overview

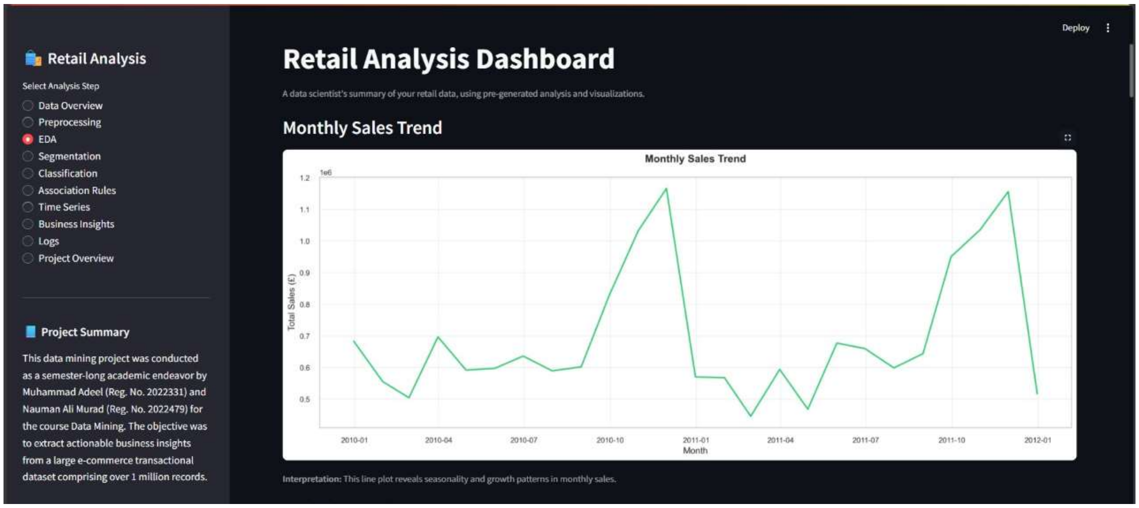
	Column	Non-Null Count	Dtype
0	Invoice	779,495	object
1	StockCode	779,495	object
2	Description	779,495	object
3	Quantity	779,495	int64
4	InvoiceDate	779,495	datetime64[ns]
5	Price	779,495	float64
6	Customer ID	779,495	float64
7	Country	779,495	object

First Few Rows of Cleaned Data

	Invoice	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
0	489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	2009-12-01 07:45:00	6.95	13,085	United Kingdom

2. Exploratory Data Analysis (EDA)

1. **Top Countries by Transaction Volume:**
   - United Kingdom
   - Ireland
   - Germany
   - France
   - Netherlands
2. **Popular Product Categories:**
   - Mini cases and polkadot designs
   - Bakelike alarm clocks
   - Children’s breakfast and cutlery sets
   - Cake stands and baking sets
   - Charlotte bags and lunch boxes
3. **Sales Trends:**
   - Significant seasonal peaks in December, indicating strong holiday-related shopping activity
   - Consistent sales patterns across weekdays and times of the day



Retail Analysis Dashboard

A data scientist's summary of your retail data, using pre-generated analysis and visualizations.

Monthly Sales Trend

Project Summary

This data mining project was conducted as a semester-long academic endeavor by Muhammad Adeel (Reg. No. 2022331) and Nauman Ali Murad (Reg. No. 2022479) for the course Data Mining. The objective was to extract actionable business insights from a large e-commerce transactional dataset comprising over 1 million records.

Interpretation: This line plot reveals seasonality and growth patterns in monthly sales.

3. Customer Segmentation Using RFM and K-Means Clustering

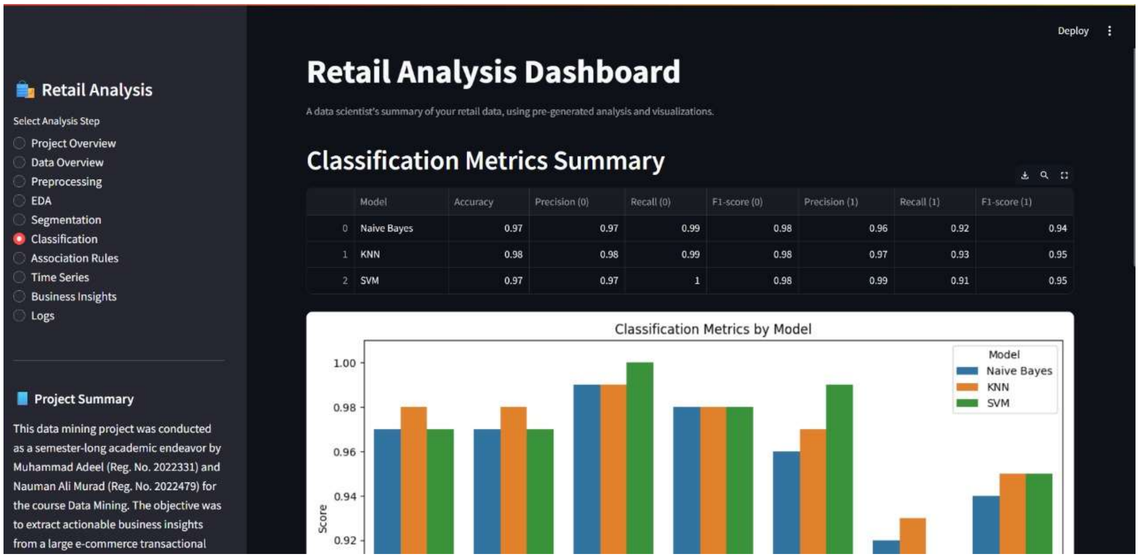
1. **Cluster 0: High-Value, Frequent Buyers**
   - Recency: ~66 days
   - Frequency: ~7.6 purchases
   - Monetary: ~£3,134 average spend

    - **Recommended Strategy:** Implement loyalty programs and exclusive offers

2. **Cluster 1: Dormant, Low-Value Customers**
    - Recency: ~462 days
    - Frequency: ~2.2 purchases
    - Monetary: ~£746 average spend
    - **Recommended Strategy:** Initiate re-engagement campaigns
3. **Cluster 2: VIP Customers**
    - Recency: ~23 days
    - Frequency: ~143 purchases
    - Monetary: ~£173,123 average spend
    - **Recommended Strategy:** Provide premium services and dedicated account managers

4. Classification Modeling

1. **Techniques Applied:**
    - Naive Bayes
    - K-Nearest Neighbors (KNN)
    - Support Vector Machine (SVM)
2. **Performance Evaluation:**
    - All models demonstrated strong accuracy
    - SVM provided the highest precision for minority classes
    - F1-scores confirmed balanced performance across clusters



Retail Analysis Dashboard

A data scientist's summary of your retail data, using pre-generated analysis and visualizations.

Classification Metrics Summary

	Model	Accuracy	Precision (0)	Recall (0)	F1-score (0)	Precision (1)	Recall (1)	F1-score (1)
0	Naive Bayes	0.97	0.97	0.99	0.98	0.96	0.92	0.94
1	KNN	0.98	0.98	0.99	0.98	0.97	0.93	0.95
2	SVM	0.97	0.97	1	0.98	0.99	0.91	0.95

Project Summary

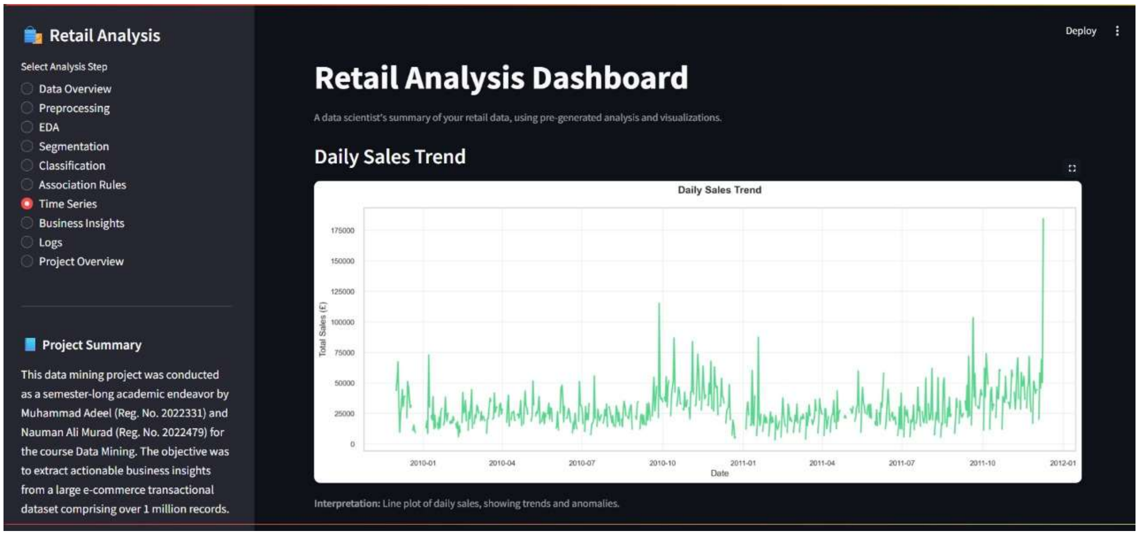
This data mining project was conducted as a semester-long academic endeavor by Muhammad Adeel (Reg. No. 2022331) and Nauman Ali Murad (Reg. No. 2022479) for the course Data Mining. The objective was to extract actionable business insights from a large e-commerce transactional

5. Association Rule Mining

1. **Techniques:** Apriori and FP-Growth algorithms
2. **Key Findings:**
   - Identified strong associations between complementary products
   - Opportunities for cross-selling and bundling strategies were uncovered
3. **Visualization:** Product relationship graphs helped illustrate actionable insights

6. Time Series Forecasting

- **ARIMA Model:**
  - Captured trend and seasonality components effectively
  - Suitable for short-term sales forecasting
- **Prophet Model:**
  - Handled outliers and missing data robustly
  - Provided accurate long-term forecasts with weekly seasonality considerations



Retail Analysis Dashboard

A data scientist's summary of your retail data, using pre-generated analysis and visualizations.

Daily Sales Trend

Interpretation: Line plot of daily sales, showing trends and anomalies.

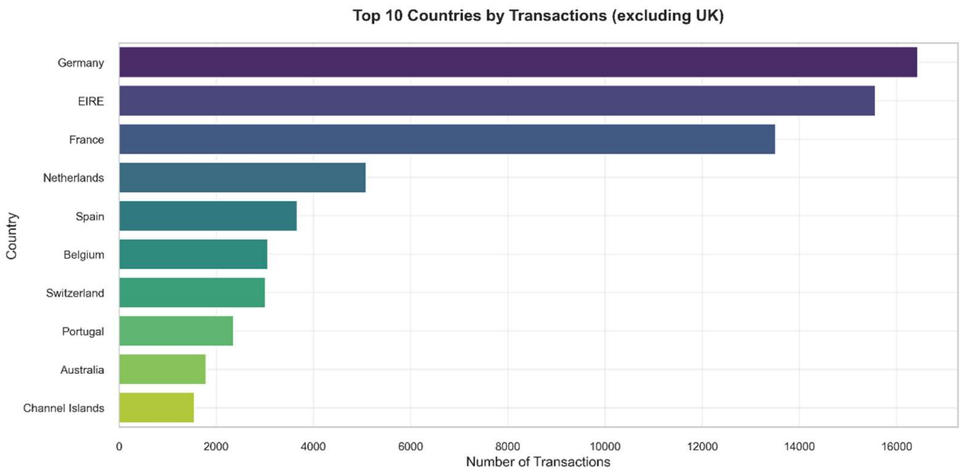
Project Summary

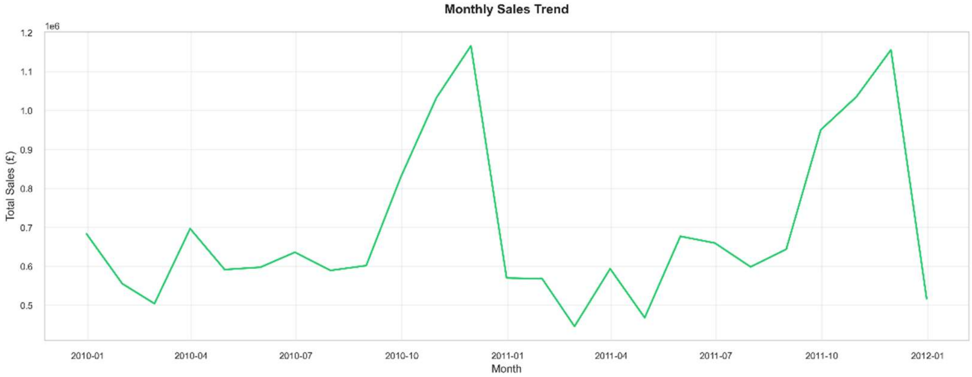
This data mining project was conducted as a semester-long academic endeavor by Muhammad Adeel (Reg. No. 2022331) and Nauman Ali Murad (Reg. No. 2022479) for the course Data Mining. The objective was to extract actionable business insights from a large e-commerce transactional dataset comprising over 1 million records.

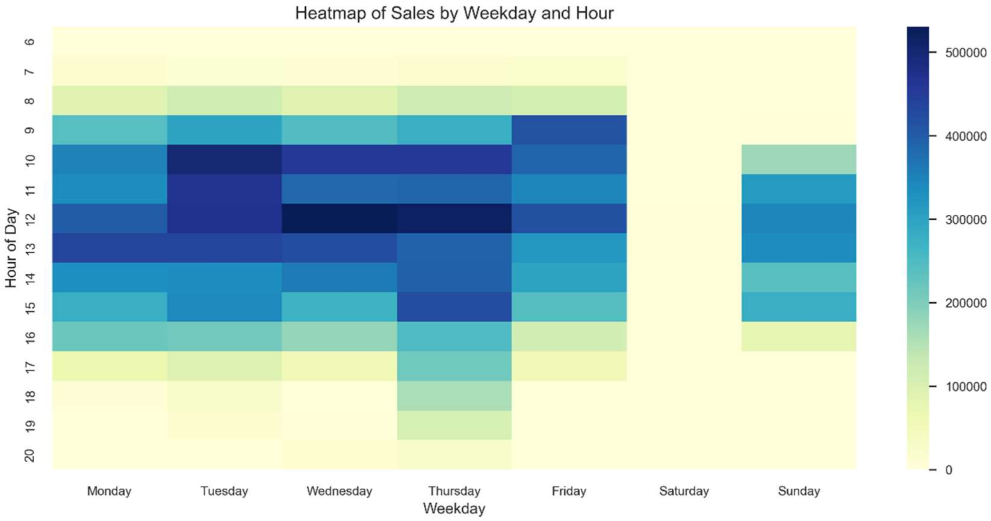
7. Business Insights & Strategic Recommendations

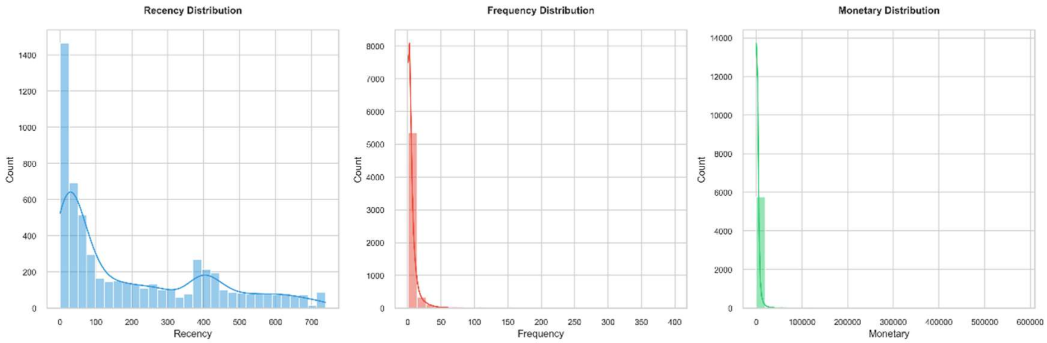
1. **Customer Retention:** Focus on high-value customer segments with personalized marketing and loyalty programs.
2. **Market Expansion:** Prioritize growth in the top-performing countries while exploring emerging markets.
3. **Product Strategy:** Maintain healthy stock levels of top-performing products and develop targeted product bundles.
4. **Operational Efficiency:** Implement dynamic pricing models and optimize supply chains based on sales forecasts.

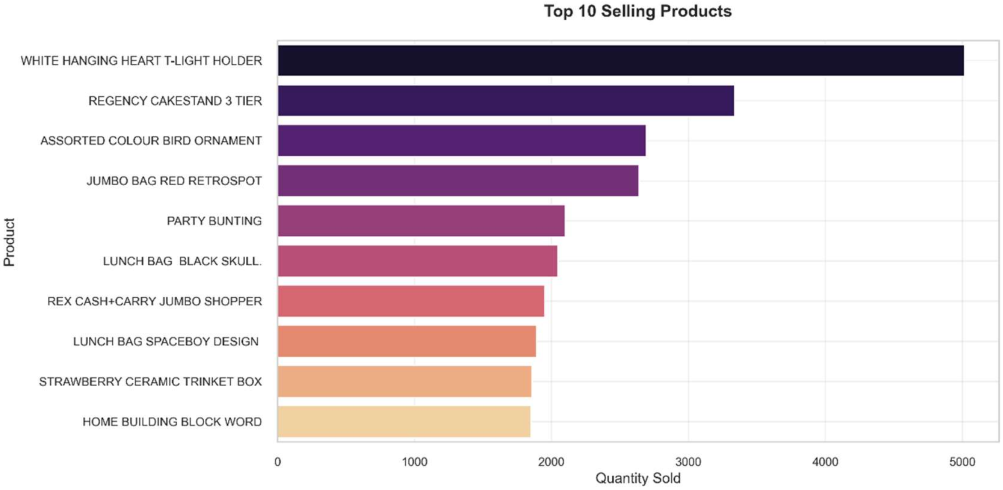
Project Insights

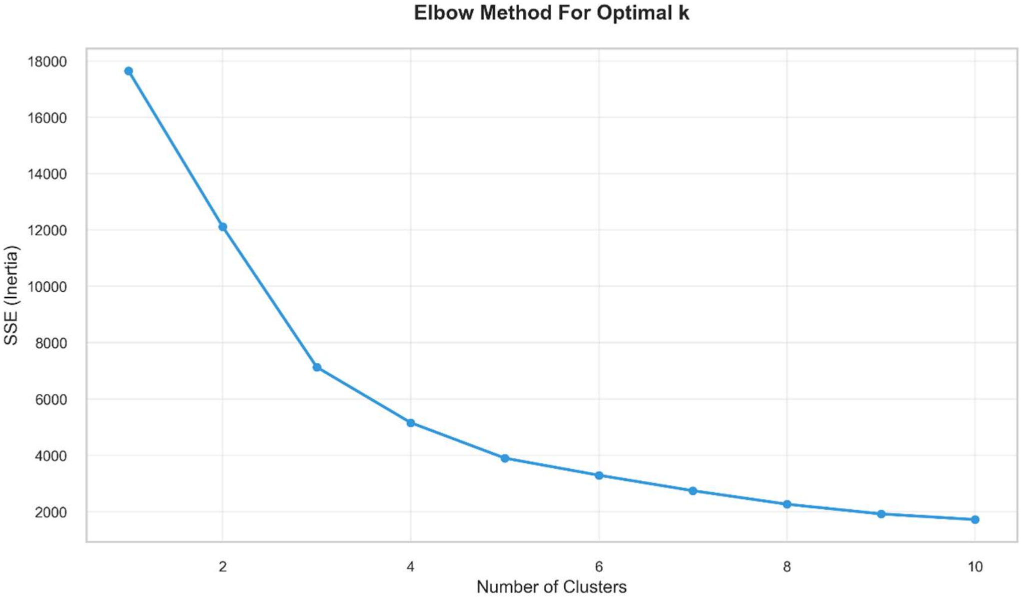


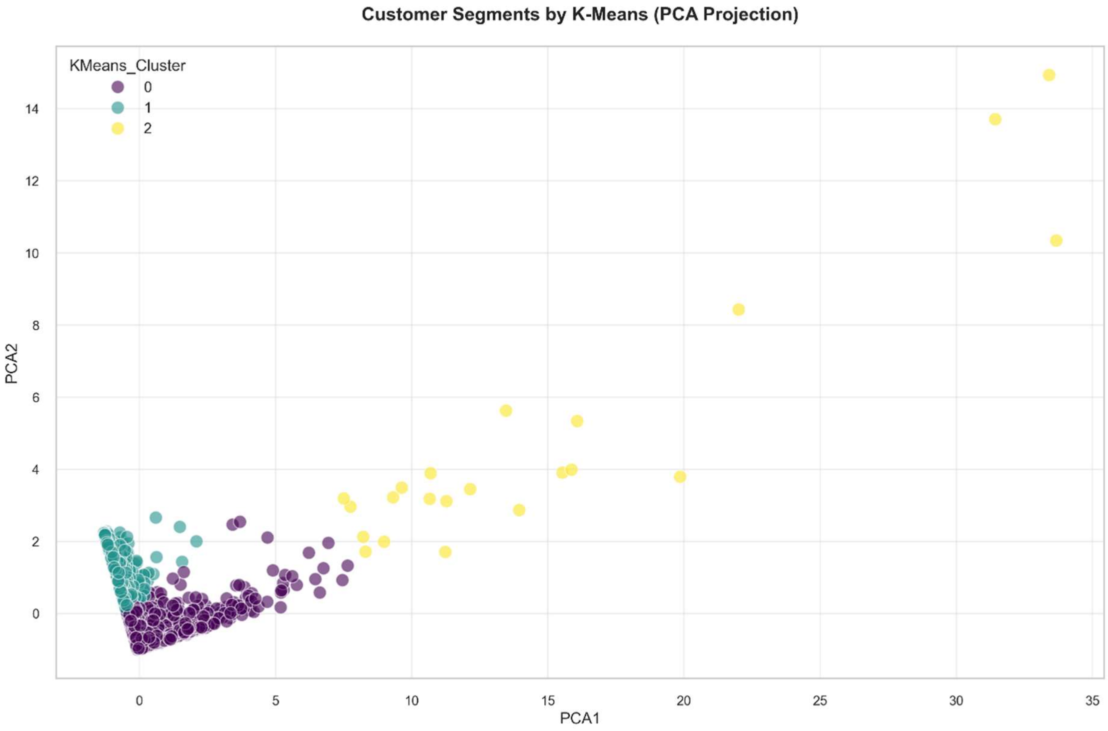


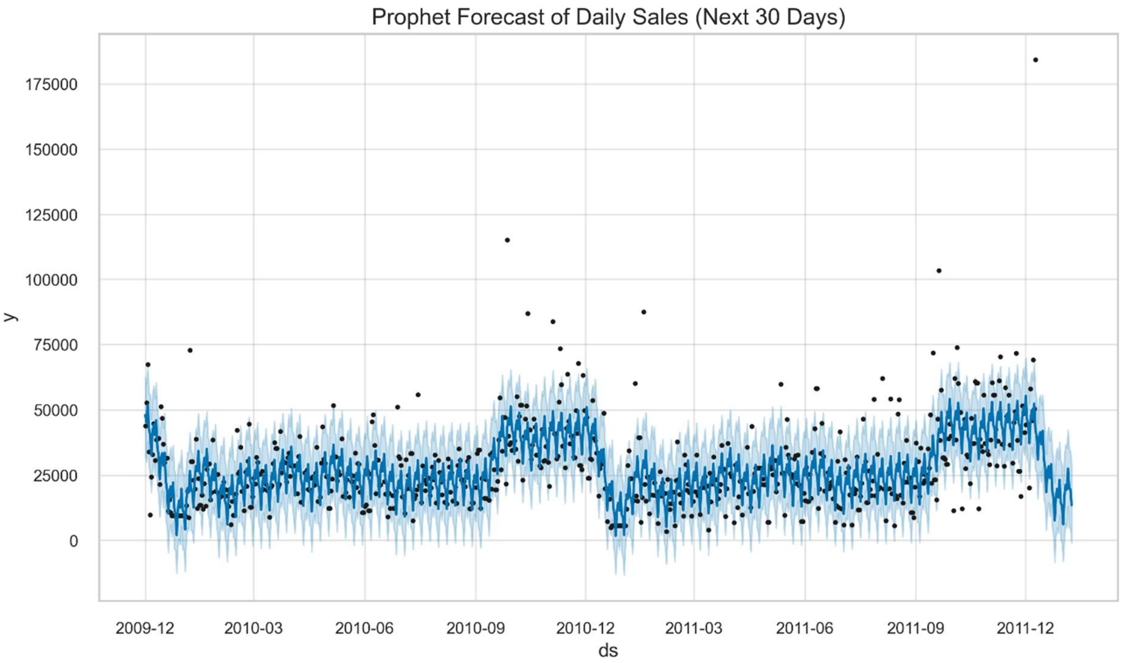


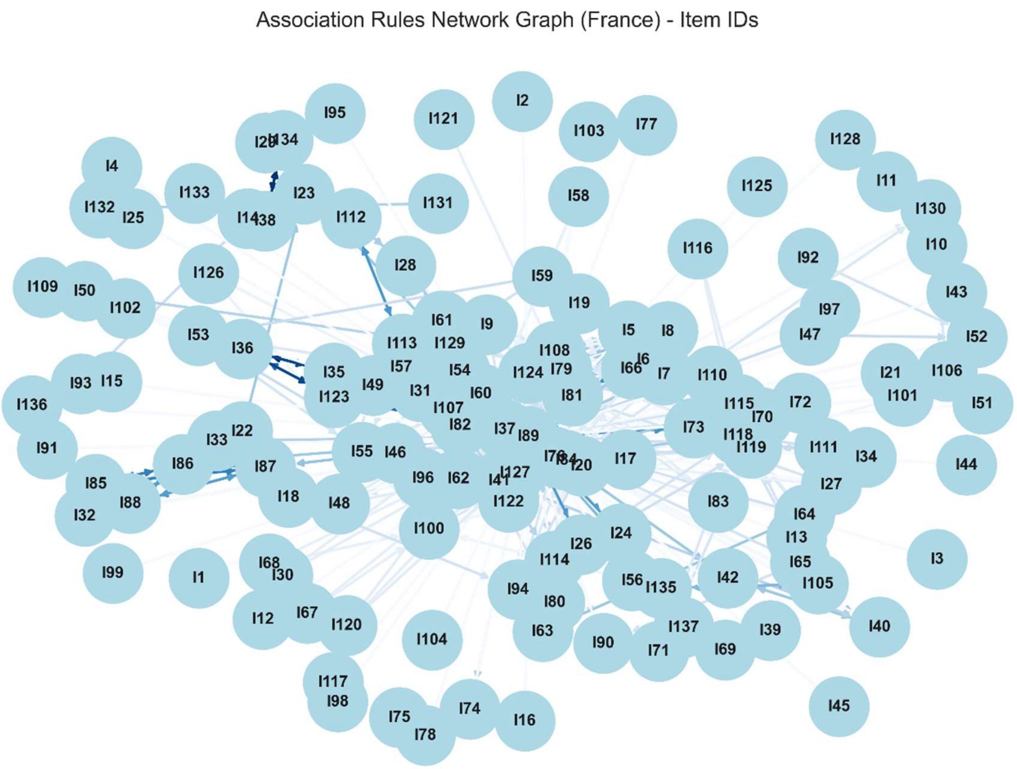

Recency Distribution
Count
0
200
400
600
800
1000
1200
1400
0
100
200
300
400
500
600
700
Recency
Frequency Distribution
Count
0
1000
2000
3000
4000
5000
6000
7000
8000
0
50
100
150
200
250
300
350
400
Frequency
Monetary Distribution
Count
0
2000
4000
6000
8000
10000
12000
14000
0
100000
200000
300000
400000
500000
600000
Monetary


Top 10 Selling Products
Product
WHITE HANGING HEART T-LIGHT HOLDER
REGENCY CAKESTAND 3 TIER
ASSORTED COLOUR BIRD ORNAMENT
JUMBO BAG RED RETROSPOT
PARTY BUNTING
LUNCH BAG BLACK SKULL.
REX CASH+CARRY JUMBO SHOPPER
LUNCH BAG SPACEBOY DESIGN
STRAWBERRY CERAMIC TRINKET BOX
HOME BUILDING BLOCK WORD
0
1000
2000
3000
4000
5000
Quantity Sold


Elbow Method For Optimal k
SSE (Inertia)
18000
16000
14000
12000
10000
8000
6000
4000
2000
2
4
6
8
10
Number of Clusters


Customer Segments by K-Means (PCA Projection)
KMeans_Cluster
0
1
2
PCA2
PCA1
0
2
4
6
8
10
12
14
0
5
10
15
20
25
30
35


Prophet Forecast of Daily Sales (Next 30 Days)
y
ds
175000
150000
125000
100000
75000
50000
25000
0
2009-12
2010-03
2010-06
2010-09
2010-12
2011-03
2011-06
2011-09
2011-12



Future Work

- Develop a dynamic data exploration platform that automatically detects and processes datasets, applying data mining techniques and generating insights without manual configuration.
- Upgrade the UI/UX by migrating from Streamlit to a MERN stack (MongoDB, Express, React, Node.js), enhancing flexibility, scalability, and interactivity.
- Integrate Flask as the backend to handle business logic, data processing, and API requests for seamless integration and performance.
- Automate model training and evaluation for customer segmentation, predictive modeling, and forecasting, allowing users to interact with results dynamically.
- Enable real-time data stream integration for up-to-date analytics, ensuring stakeholders have access to the latest insights.
- Implement interactive dashboards using JavaScript libraries like D3.js or Plotly for better visualization and user interaction with real-time analytics.
- Add A/B testing capabilities to measure marketing strategy effectiveness, automating data collection, hypothesis testing, and results visualization.
- Enhance security with user authentication and role-based access control for tailored access to features and data.

Acknowledgment

We extend our sincere gratitude to Dr. Ayaz Umer for his invaluable guidance, mentorship, and expertise throughout the course and the completion of this project. His clear explanation of complex data mining concepts and his encouragement in applying these techniques practically have greatly enhanced our understanding of the subject. Dr. Umer's dedication to teaching and his expertise in artificial intelligence, natural language processing, and data mining were instrumental in our success, and we deeply appreciate the opportunity to learn under his supervision.

References

[1] Chen, D., Sain, S.L., & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. *Journal of Database Marketing & Customer Strategy Management*, 19(3), 197-208.

[2] Chen, D., Guo, K., & Ubakanma, G. (2015). Predicting customer profitability over time based on RFM time series. *International Journal of Business Forecasting and Marketing Intelligence*, 2(1), 1-18.

[3] Chen, D., Guo, K., & Li, B. (2019). Predicting Customer Profitability Dynamically over Time: An Experimental Comparative Study. *24th Iberoamerican Congress on Pattern Recognition (CIARP 2019)*, Havana, Cuba.

[4] Singh, R., Graves, J.A., Talbert, D.A., & Eberle, W. (2018). Prefix and Suffix Sequential Pattern Mining. *Industrial Conference on Data Mining 2018: Advances in Data Mining. Applications and Theoretical Aspects*, 309-324.