School of Computing and Mathematics
University of South Wales

DS70N25A - Predictive Analysis of Online Retail Sales Trends and Customer behavior

Student name: Nithyashree Thimmegowda
Student ID: 30119304

Contents:

Background

- E-commerce growth has shifted retail from physical stores to online platforms.
- Massive transactional data is generated but often underutilized.
- Traditional decision-making struggles with pattern detection and forecasting.
- Predictive analytics and machine learning help extract actionable insights.
- This study analyzes 541,909 UK online retail transactions (2010–2011).
- Techniques used: Random Forest, Logistic Regression, ARIMA, SARIMA, and Apriori for segmentation, churn prediction, forecasting, and association analysis.

Problem Statement:

- E-commerce generates complex, noisy, and incomplete transactional data.
- Businesses struggle to extract actionable insights for customer segmentation, forecasting, and churn prediction.
- Challenges: missing customer IDs, class imbalance, seasonality effects.
- Lack of integrated and interpretable predictive frameworks limits data-driven decisions.
- This study develops a robust ML-based framework to address these issues.

Aim & Objectives

- **Aim:**

  Develop a predictive analytics framework using ML & statistical models to analyze customer behavior, predict churn, forecast sales, and identify product associations.

- **Objectives:**
  - Clean & preprocess online retail data.
  - Perform EDA on temporal, geographic & product trends.
  - Segment customers via RFM & K-means clustering.
  - Build churn prediction models (Logistic Regression, Random Forest).
  - Forecast sales (ARIMA, SARIMA, Prophet).
  - Apply Apriori for association rule mining.
  - Evaluate models using accuracy, recall, RMSE, MAPE.

Research Questions

- What sales trends (temporal & product-level) exist in the dataset?
- How can behavioral customer segmentation improve marketing?
- Which behavioral features predict customer churn?
- Which products have strong co-purchase relationships for cross-selling?
- Which forecasting models best predict future sales?
- How do ML models (Logistic Regression vs. Random Forest) compare for churn prediction?

Literature review

Study	Techniques Used	Dataset Type	Validation/Accuracy	Unique Contribution	Key Limitation
Kaya & Saleem (2023)	Clustering, Assoc. Rules, Trend Analysis	Genuine business data	Not reported	Focus on temporal behavior	Lack of model validation
Alghanam et al. (2022)	K-means, J48, C4.5, Apriori	Northwind (synthetic)	95.2% (claimed)	Hybrid modelling for recommendations	Reproducibility not addressed
Thomas (2024)	Overview: Clustering, Assoc. Rules, NLP	Conceptual	N/A	Sentiment analysis and ethics	No implementation or dataset evaluation
Sri Darshan et al. (2024)	Prefix Span, Apriori, Prophet	Synthetic retail data	MAE, RMSE used	Time series + sequential behavior	Not customer-segment focused
Alawadh & Barnawi (2024)	Clustering, C4.5, Apriori	1M+ real transactions	Not reported	Practical deployment with customer segments	Theoretical grounding and validation lacking

Data preparation and Cleaning

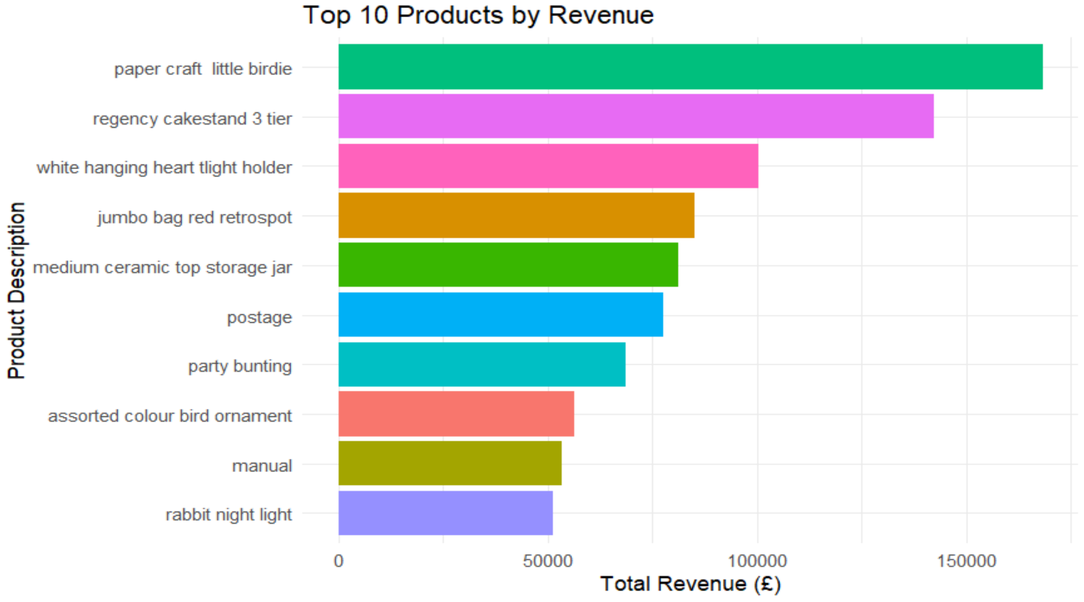
- **Methodology:** CRISP-DM framework (Business Understanding → Deployment)
- **Dataset:** 541,909 UK online retail transactions (Dec 2010–Dec 2011)
- **Data Cleaning:**
  - Removed missing CustomerID (~25%) and Description fields.
  - Removed duplicates, cancelled transactions (InvoiceNo starting with 'C').
  - Eliminated invalid Quantity & UnitPrice (≤0).
  - Standardized data types (InvoiceDate, CustomerID).

Feature Engineering

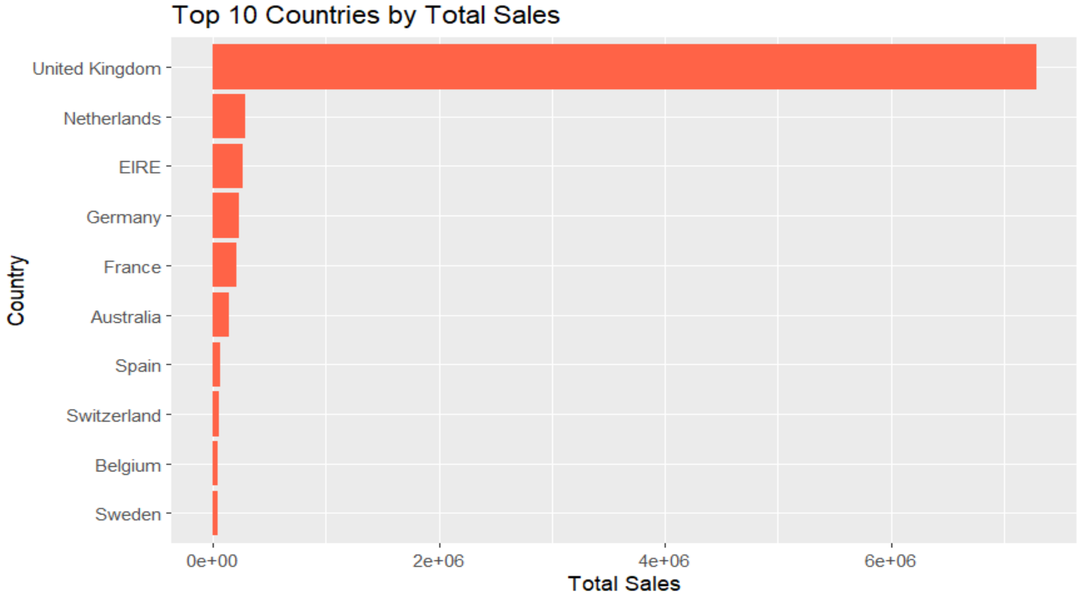
- Created **TotalPrice** (Quantity × UnitPrice).
- Extracted temporal features (Month, Day, Hour).
- Calculated **RFM metrics** (Recency, Frequency, Monetary).
- Computed Basket Size.
- Created Churn Labels (inactive >90 days).

Exploratory Visualizations

- **Top Products by Revenue:**
‘Paper Craft Little Birdie’, ‘Regency Cakestand 3 Tier’ are top sellers.

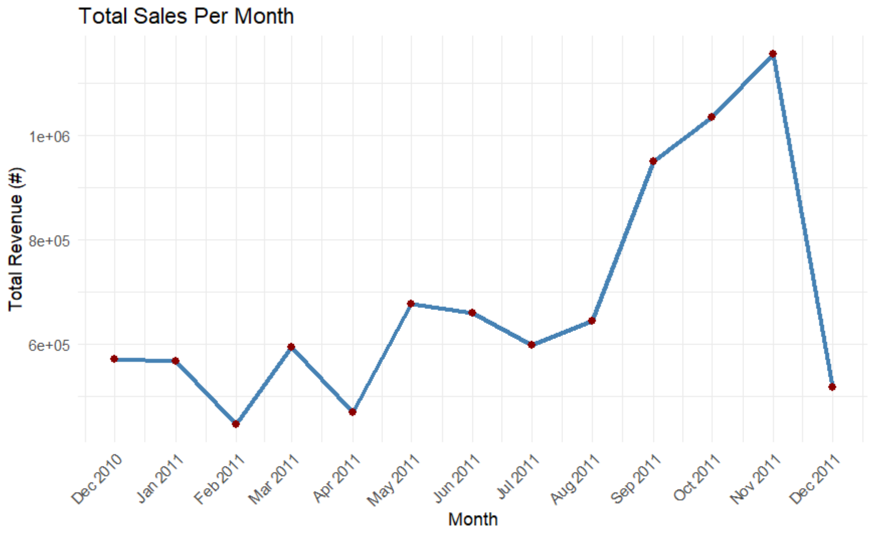


- **Country Insights:**
UK: 90% of sales; EIRE: highest average revenue per customer.

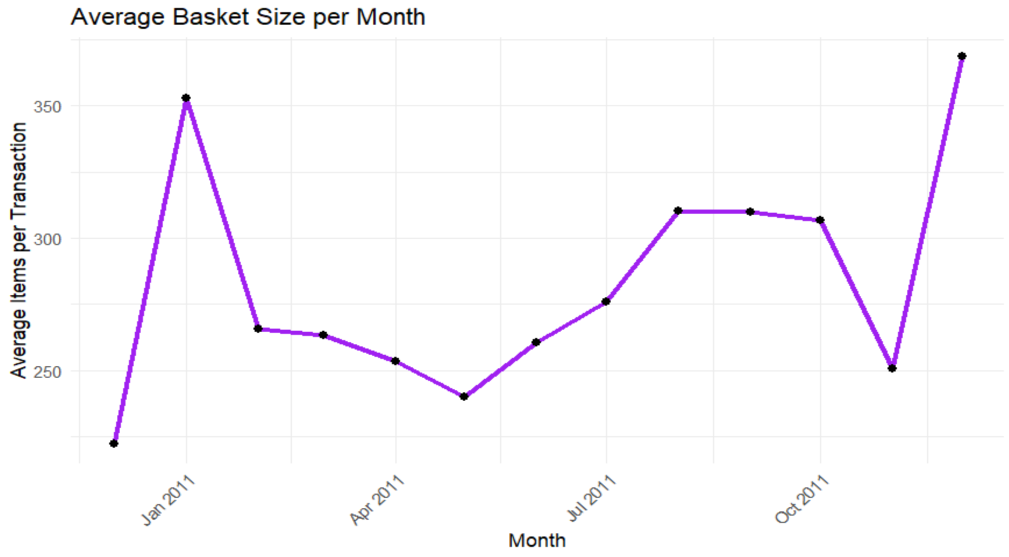


Exploratory Visualizations continued

- **Monthly Sales Trends:**
  - Peak sales: Nov 2011 (holiday season).
  - Lowest sales: Feb & Apr 2011.



- **Basket Size Trends:**
  Highest basket size in December 2011.

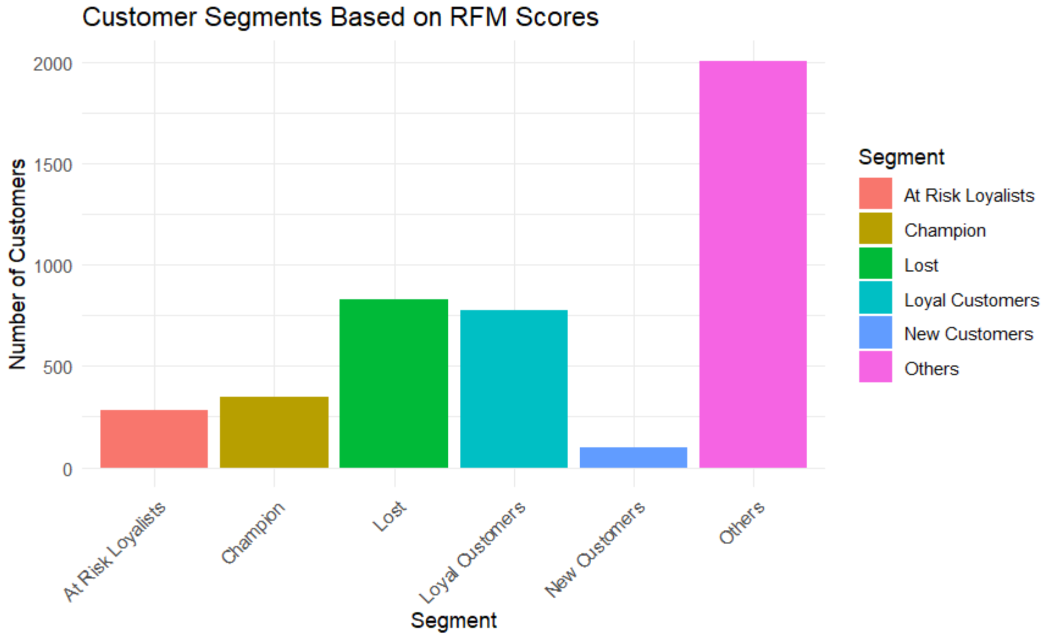


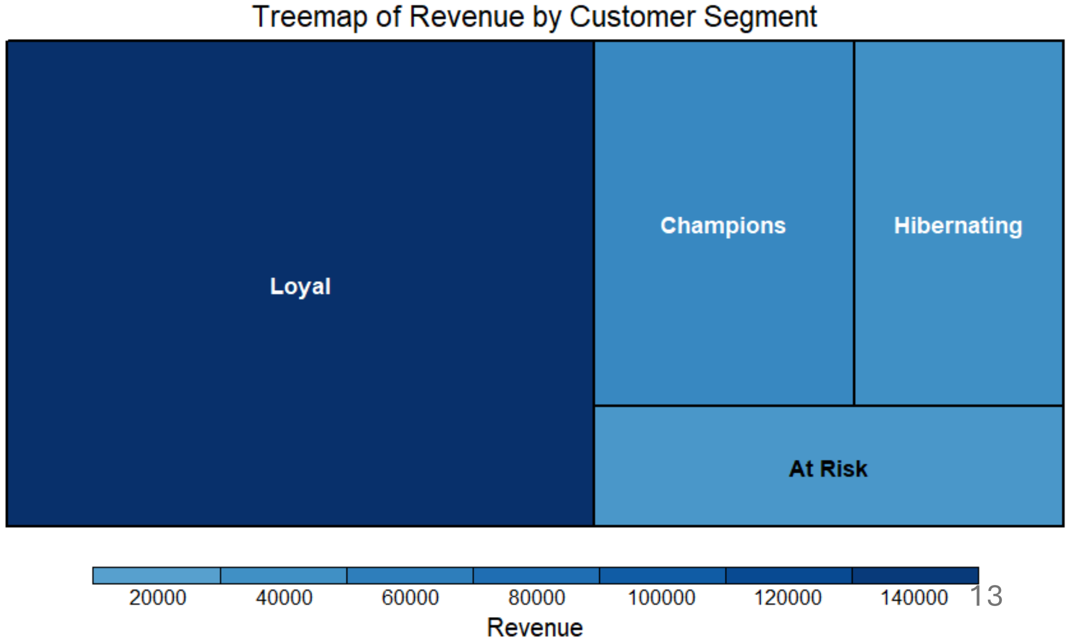
Methodology

- Adopted **CRISP-DM framework** for entire modeling pipeline.
- Tools: R (dplyr, ggplot2, caret, forecast, randomForest, arules).
- Methods applied:
    - Customer Segmentation
    - Churn Prediction
    - Time Series Forecasting
    - Association Rule Mining
    - Interactive Dashboards

Methodology

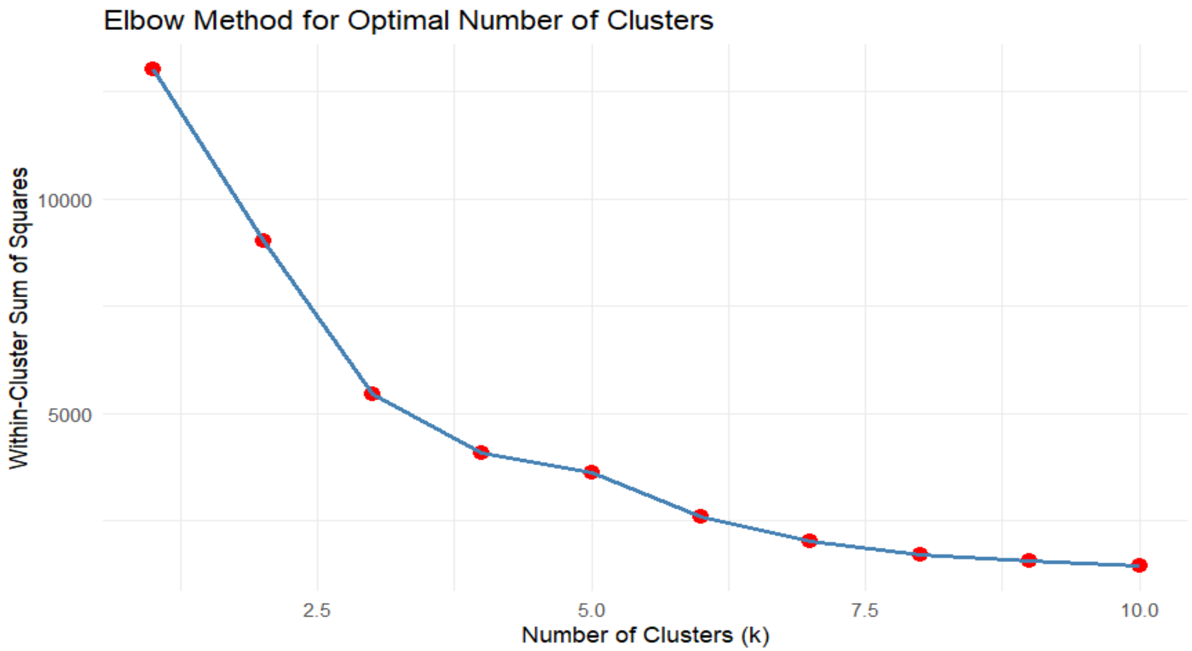
- **Customer Segmentation (RFM)**
- Used **Recency, Frequency, Monetary (RFM)** model.
- RFM scores: 1 (low) to 5 (high) for each dimension.
- Segments identified:
    - **Champions**
    - **Loyal Customers**
    - **At Risk**
    - **Hibernating**

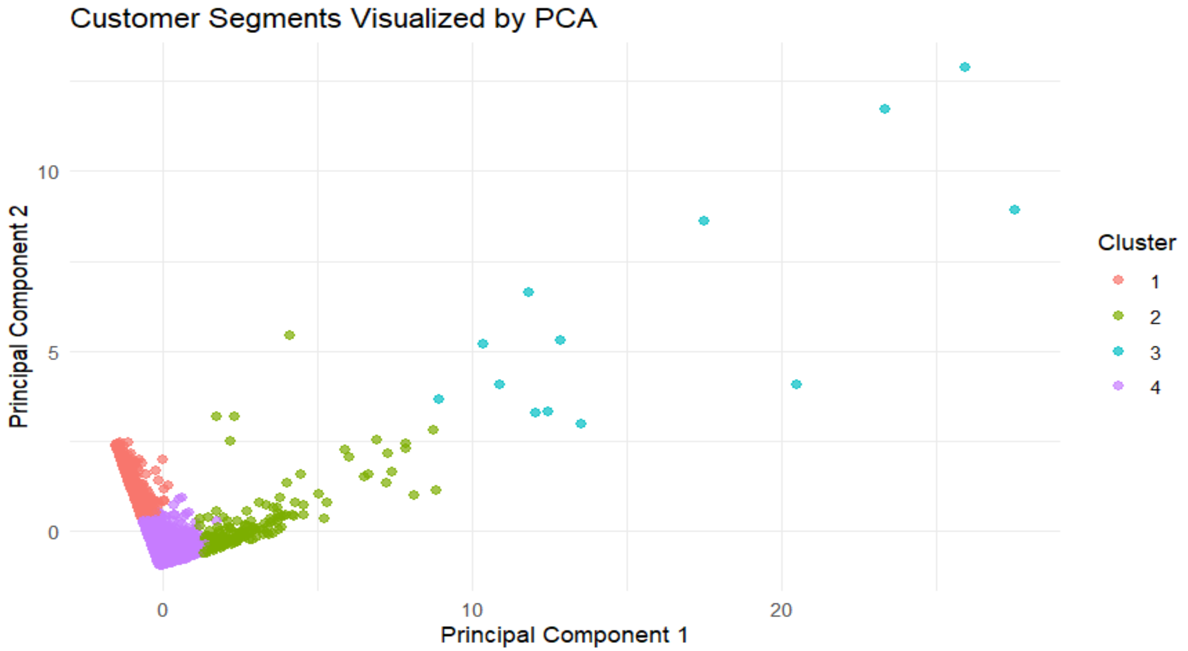




Methodology

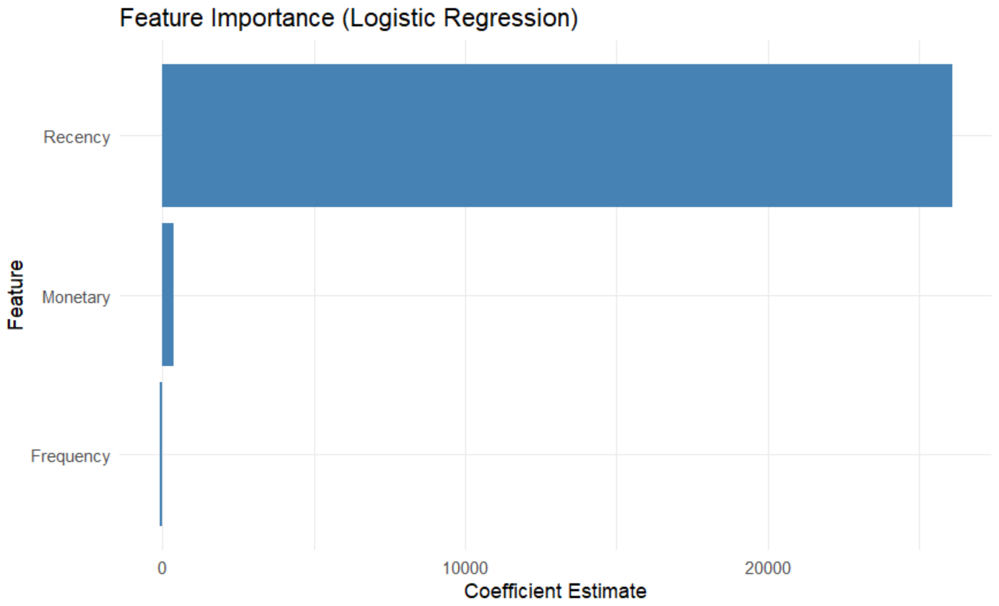
- **Clustering: Persona Development**
- Applied **K-Means Clustering** on normalized RFM data.
- Optimal clusters determined using Elbow Method (**k=4**).
- Resulting Personas:
    - **Bargain Hunters**
    - **Big Spenders**
    - **Occasional Buyers**
    - **Frequent Buyers**

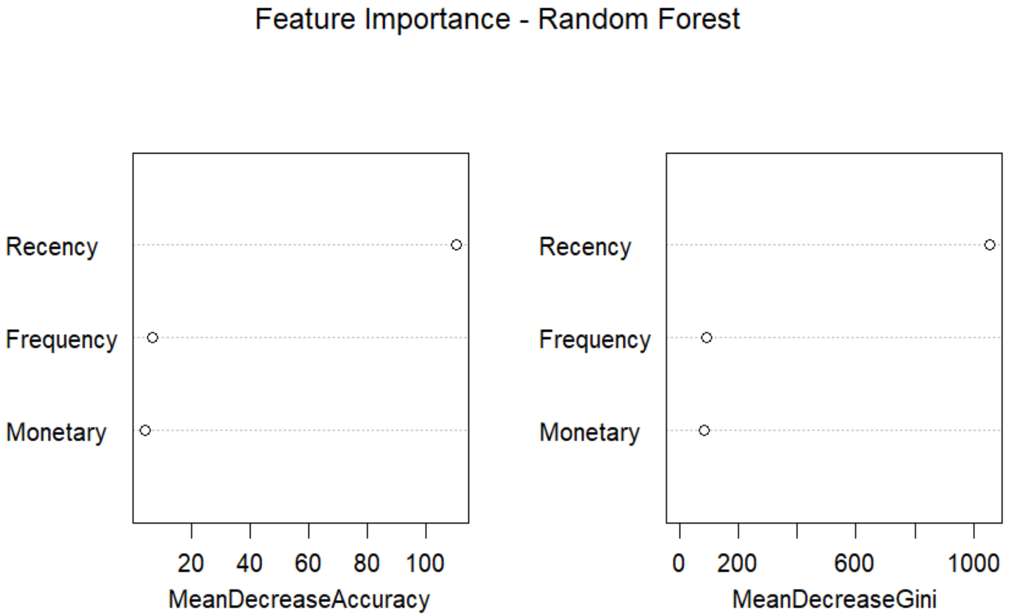




Methodology

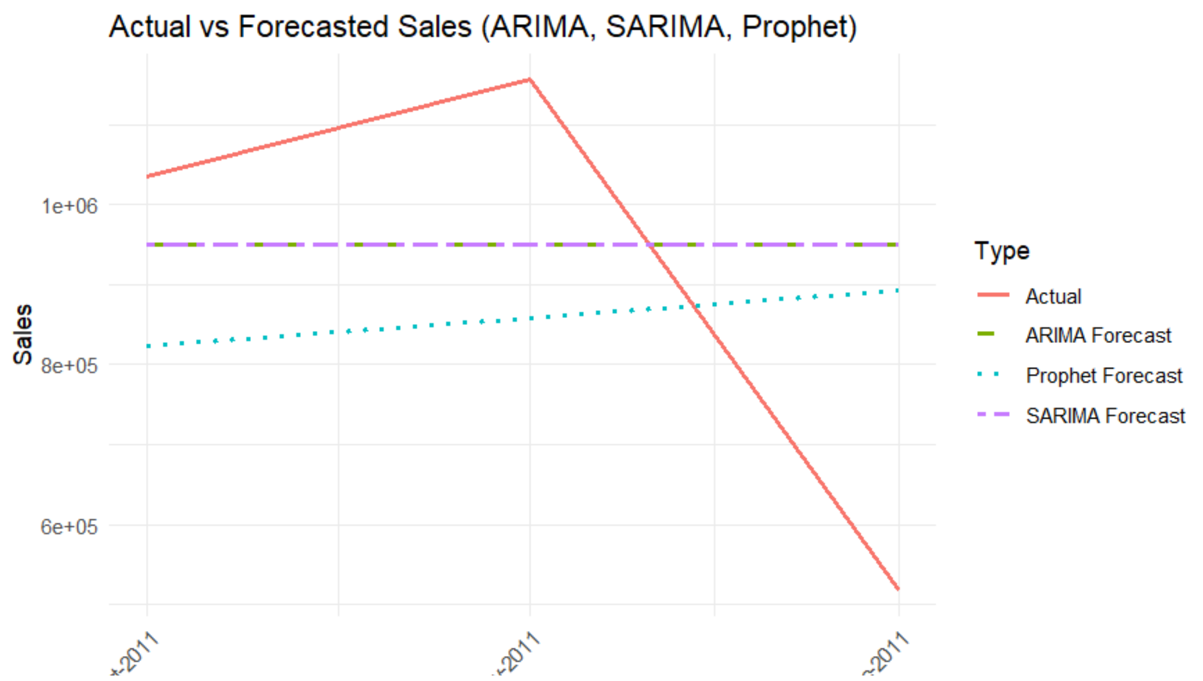
- **Churn Prediction**
- Defined churn: No purchase in last **90 days**.
- Features: Recency, InvoiceCount, TotalSpent.
- Models used:
  - **Logistic Regression**
  - **Random Forest** (better performance)
- Random Forest Results:
  - Accuracy: **82%**
  - Recall: **76%**
  - F1-Score: **0.78**
  - AUC: **>0.80**

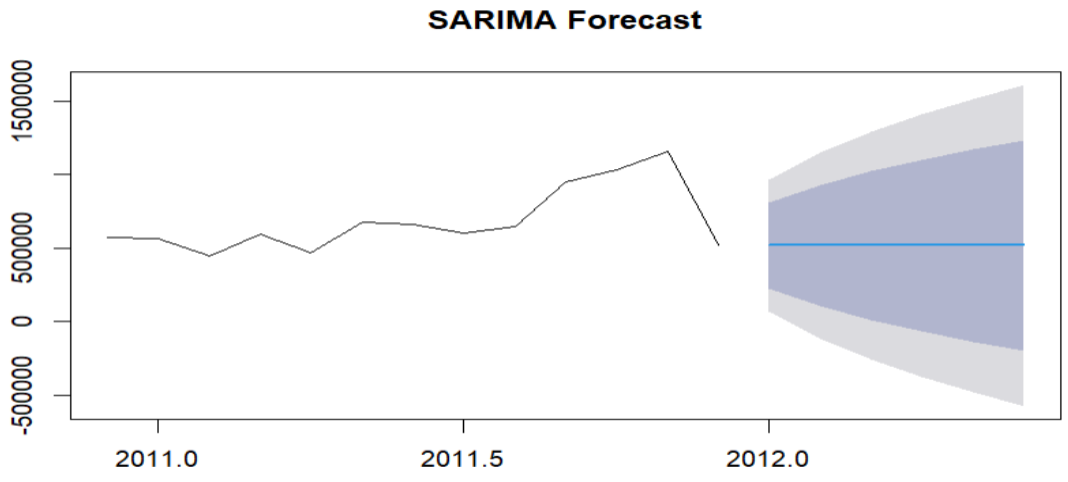




Methodology

- **Time Series Forecasting**
- Models applied:
  - **ARIMA**
  - **SARIMA**
  - **Prophet**
- Best performing model: **SARIMA**(with seasonality)
  - RMSE: **281,292**
  - MAPE: **36.6%**
- Seasonal patterns successfully captured.





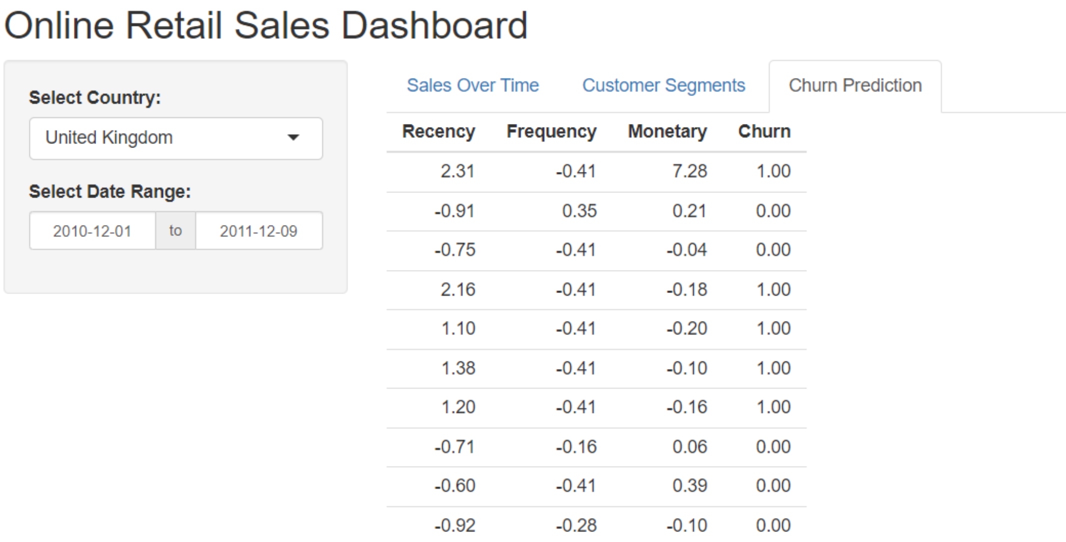
Methodology

- **Association Rule Mining**
- Applied **Apriori algorithm**.
- Extracted frequent product pairs for cross-selling.
- Identified high-lift rules to support bundling & promotion strategies.

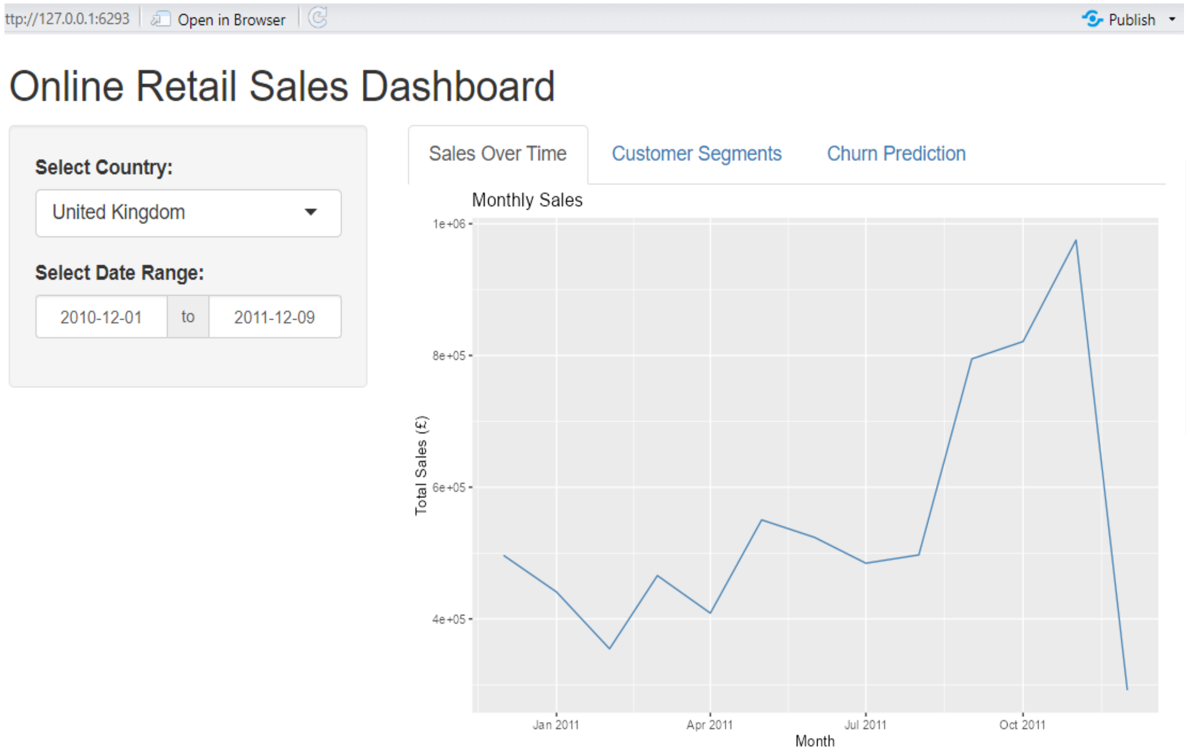
Rule (If customer buys…)	**…then also buys**	**Support**	**Confidence**	**Lift**	**Insight**
pink knitted egg cosy	blue knitted egg cosy	0.0012	88.0%	**652.3**	Strong complementary purchase – almost always bought together
blue knitted egg cosy	pink knitted egg cosy	0.0012	88.0%	**652.3**	Same as above – bi-directional relationship
pantry hook balloon whisk + spatula	tea strainer	0.0011	90.9%	**601.7**	High-value kitchen bundle
balloon whisk + tea strainer	spatula	0.0011	83.3%	**594.0**	Another strong pantry set rule
pantry hook spatula	tea strainer	0.0012	84.6%	**560.0**	Consistently paired tools
tea strainer	spatula	0.0012	78.6%	**560.0**	Very high frequency pairing
spatula + tea strainer	balloon whisk	0.0011	90.9%	**543.5**	Suggests promoting as a set
tea strainer	balloon whisk	0.0013	85.7%	**512.4**	Strong unidirectional affinity
balloon whisk	tea strainer	0.0013	77.4%	**512.4**	Mirrors above pattern
party pizza dishes (3 styles)	green polkadot dish	0.0011	87.5%	**506.7**	Party dish set frequently bought as full color range

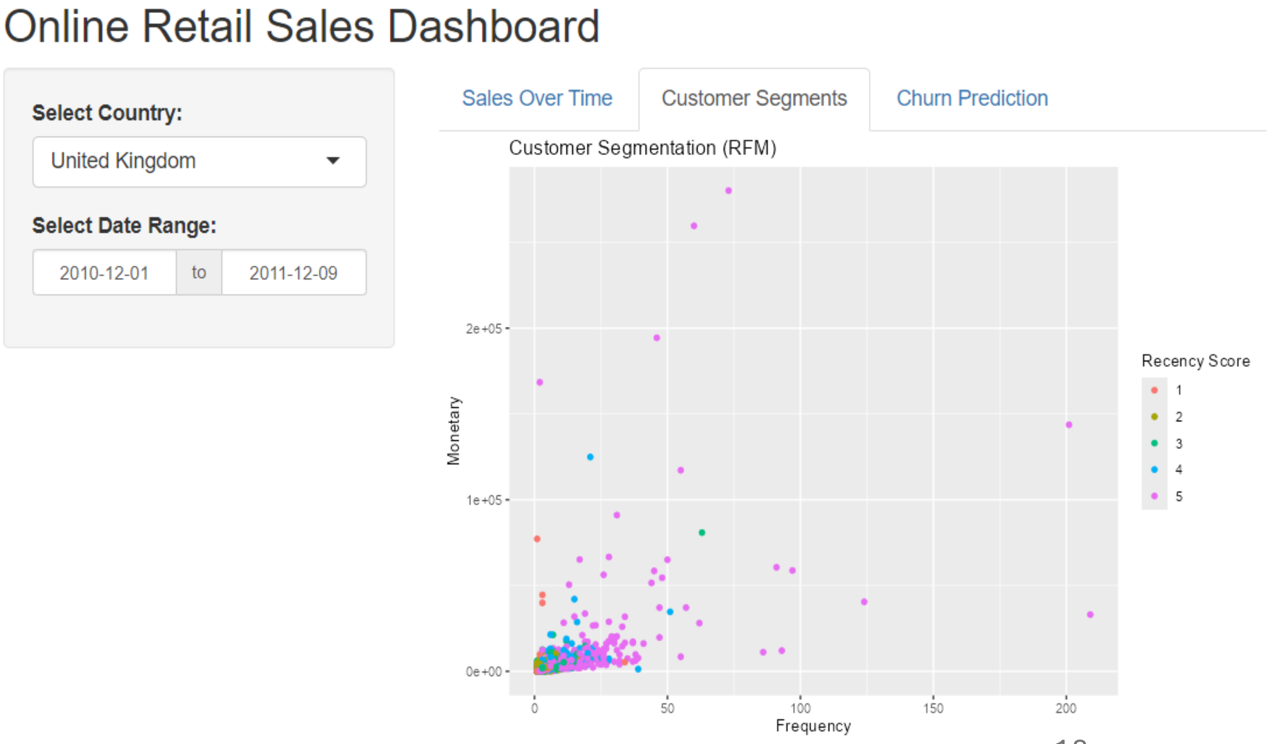
Methodology

- **Interactive Dashboards:**
  - Developed for business users to explore sales, segmentation & churn predictions.



Recency	Frequency	Monetary	Churn
2.31	-0.41	7.28	1.00
-0.91	0.35	0.21	0.00
-0.75	-0.41	-0.04	0.00
2.16	-0.41	-0.18	1.00
1.10	-0.41	-0.20	1.00
1.38	-0.41	-0.10	1.00
1.20	-0.41	-0.16	1.00
-0.71	-0.16	0.06	0.00
-0.60	-0.41	0.39	0.00
-0.92	-0.28	-0.10	0.00





Conclusion

- Developed an integrated **predictive analytics framework** for online retail.
- Enabled actionable insights for:
  - Customer engagement
  - Retention strategies
  - Sales forecasting
  - Cross-selling opportunities
- Framework applicable to real-world business decision-making.

Future Work

- Incorporate **real-time data** & **deep learning models**.
- Apply **fairness-aware analytics**.
- Integrate external data sources: IoT, social media, customer feedback.
- Extend framework to longer time periods for seasonal stability.


Thank
You


ANY
questions?