Association Rule Mining and Time Series Forecasting

Data Mining – Homework 1

Mahmoud Sadrian
Student ID: 40435336
GitHub Repository

December 10, 2025

Abstract

This report presents a complete data mining study on a real-world retail transaction dataset, focusing on manual implementation of association rule mining and time series forecasting. The workflow includes extensive data preprocessing, transformation of transactions into a binary basket matrix, and implementation of Apriori frequent itemset generation without external libraries. Frequent itemsets, as well as support, confidence, and lift, were computed manually for all one-item and two-item combinations.

Exploratory data analysis (EDA) was conducted using histograms, bar charts, and basket-level statistics to better understand product frequencies and co-occurrence patterns. A set of meaningful association rules was extracted and analyzed, with visualizations illustrating support–confidence–lift relationships.

In the second part of the project, a daily revenue time series was constructed and analyzed. A Simple Exponential Smoothing (SES) forecasting model was manually implemented to predict future revenue, and its performance was compared to a naive baseline forecast. The SES model demonstrated superior performance based on RMSE, MAE, and MAPE metrics.

All computations, algorithms, and visualizations were implemented from scratch, ensuring direct understanding of the mathematical and algorithmic foundations of association rule mining and time series forecasting.

1 Introduction

Retail transaction datasets provide an ideal foundation for extracting actionable insights about customer behavior. Two of the most widely used techniques in this domain are association rule mining—used to discover co-occurrence patterns among products—and time series forecasting, which supports revenue prediction and inventory planning.

This project focuses on implementing both tasks **from scratch**. We manually construct the basket matrix, compute support, confidence, and lift without external packages, and implement a simplified Apriori algorithm for one- and two-item frequent itemsets. Furthermore, we develop a manual Simple Exponential Smoothing (SES) model to forecast daily revenue.

This dual analysis provides a complete view of both cross-sectional patterns (product co-occurrence) and temporal behavior (revenue evolution).

2 Dataset Description

The dataset contains retail transactions, where each row represents a purchased item within an invoice. The primary attributes include:

- **InvoiceNo:** Unique invoice identifier,
- **Description:** Product name,
- **Quantity:** Number of purchased units,
- **UnitPrice:** Price per unit,
- **InvoiceDate:** Timestamp,
- **CustomerID:** Customer identifier,
- **Country:** Country of sale.

An additional attribute **TotalAmount** is computed:

$$\text{TotalAmount} = \text{Quantity} \times \text{UnitPrice}.$$

This processed dataset is then used for both association rule mining (transaction-level analysis) and revenue forecasting (time series aggregation).

3 Preprocessing

Preprocessing consists of three major steps: cleaning, basket transformation, and time series construction.

3.1 Data Cleaning

Invalid rows were removed based on:

- missing **Description**,
- missing **CustomerID**,
- nonpositive **Quantity** or **UnitPrice**.

The cleaned dataset ensures meaningful transactions and valid revenue values.

3.2 Basket Matrix Construction

A binary matrix B was created:

$$B_{ij} = \begin{cases} 1 & \text{if invoice } i \text{ contains product } j, \\ 0 & \text{otherwise.} \end{cases}$$

To maintain computational feasibility in the manual Apriori implementation, only the top 50 most frequent products were retained.

3.3 Daily Revenue Series

Revenue for each day was defined as:

$$R(t) = \sum_{i \in \text{transactions on } t} \text{TotalAmount}_i.$$

This series forms the basis for SES forecasting.

4 Exploratory Data Analysis

4.1 Distributions

Figures 1 and 2 display the distributions of quantities and unit prices.

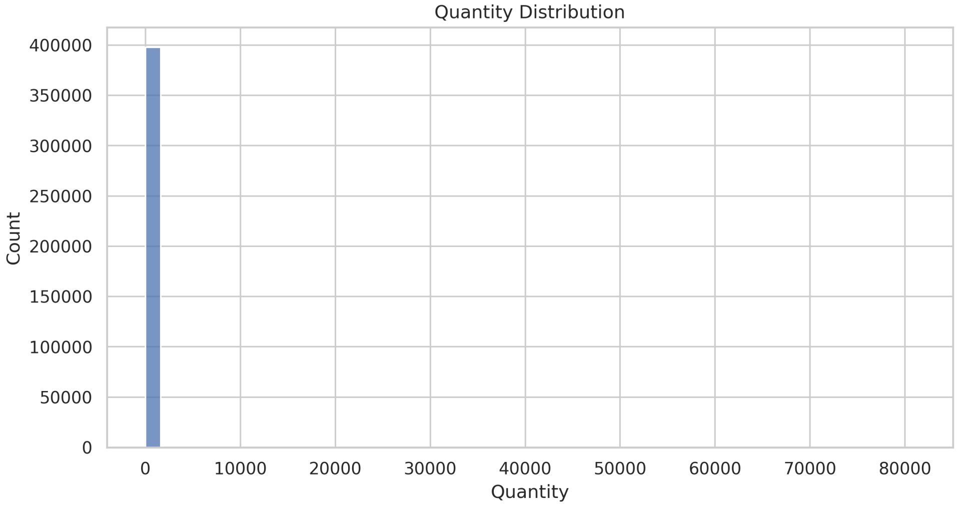

Figure 1: Distribution of purchased quantities.

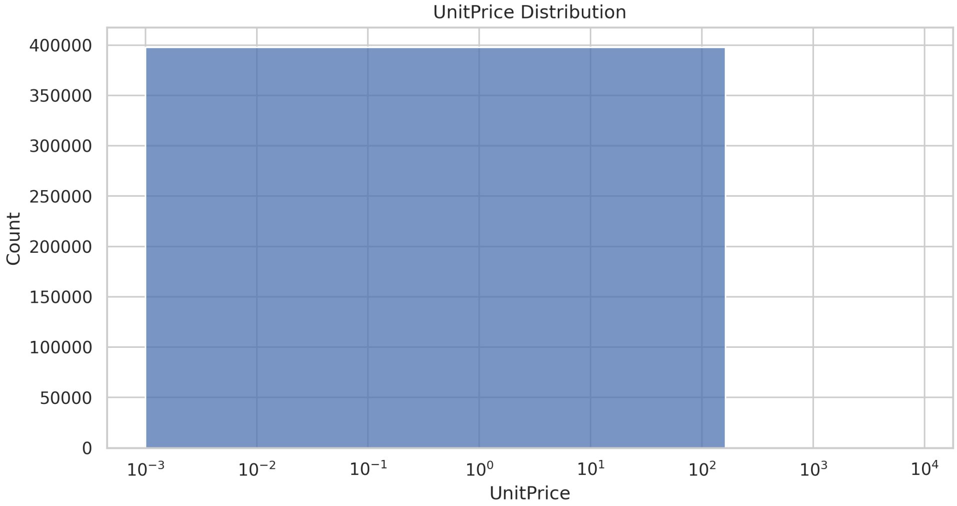

Figure 2: Distribution of unit prices.

4.2 Top Products

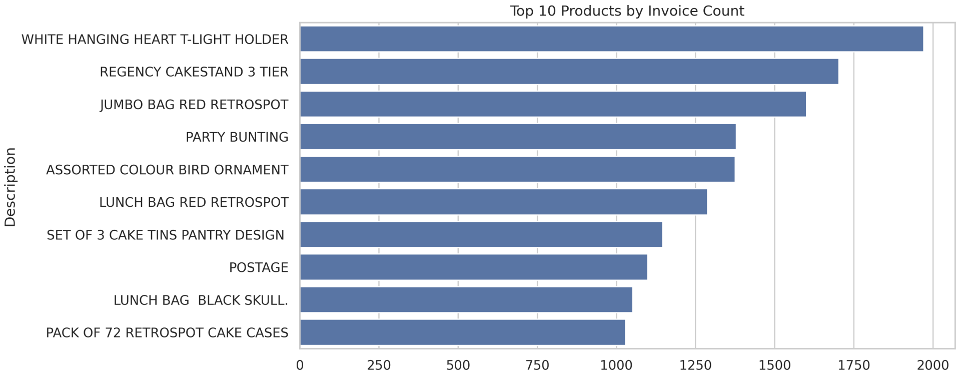

Figure 3: Top 10 most purchased products.

4.3 Revenue by Country

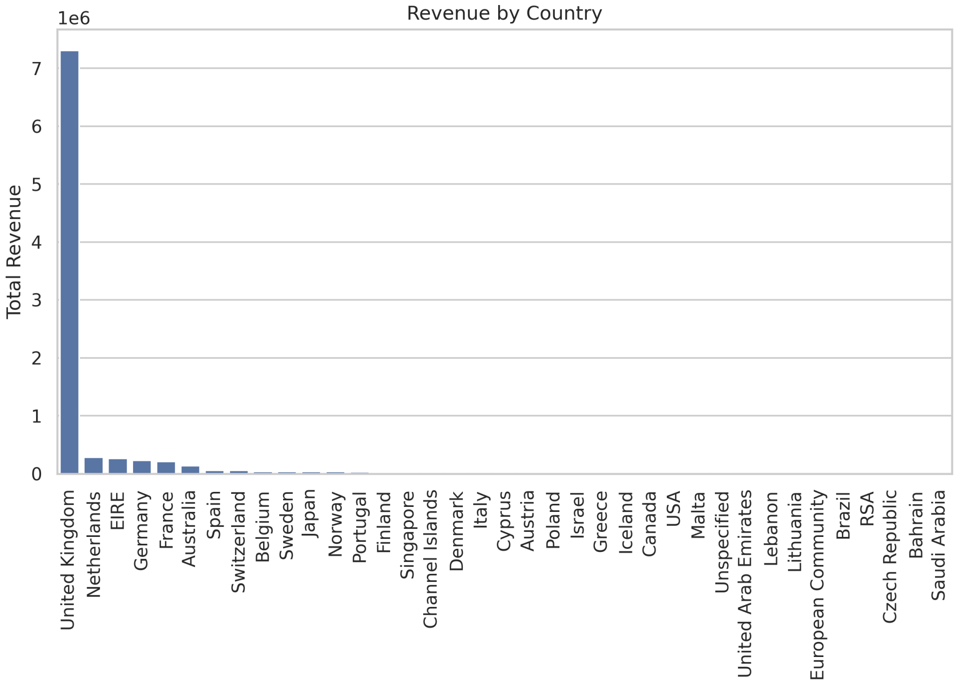

Figure 4: Total revenue by country.

4.4 Daily Revenue

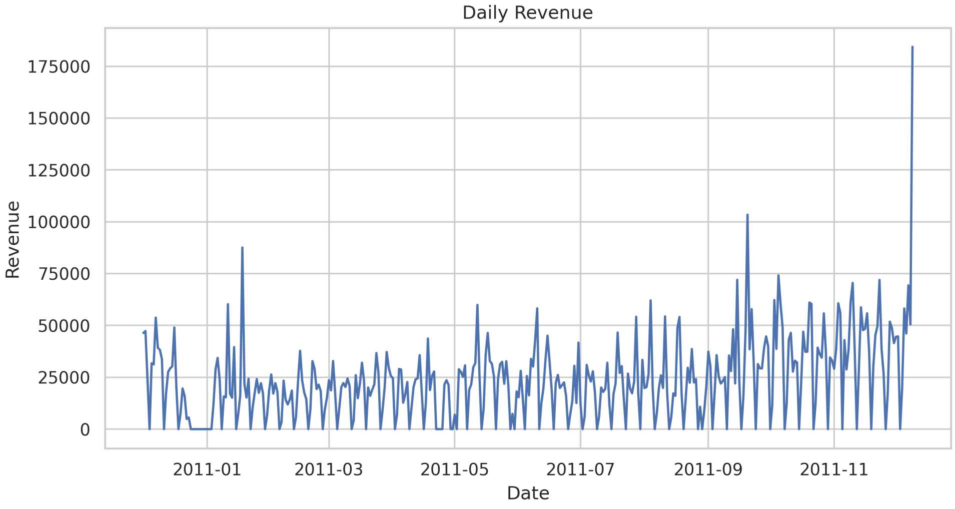

Figure 5: Daily revenue time series.

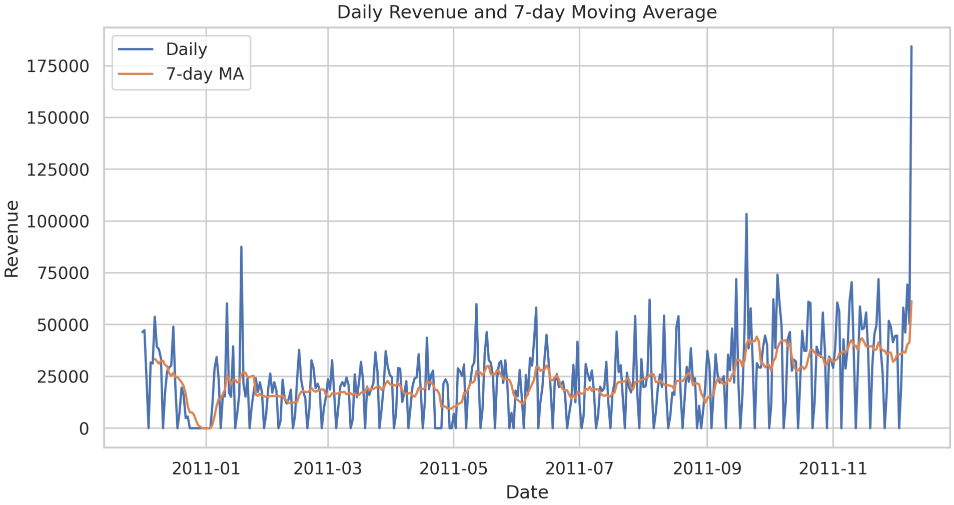

Figure 6: 7-day moving average.

5 Association Rule Mining (Manual Implementation)

5.1 Support, Confidence, and Lift

For itemset X:

$$\text{support}(X) = \frac{\text{transactions containing } X}{N}.$$

For rule $X \to Y$:

$$\text{confidence}(X \to Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}.$$

$$\text{lift}(X \to Y) = \frac{\text{confidence}(X \to Y)}{\text{support}(Y)}.$$

5.2 Frequent Itemsets

Only 1-item and 2-item itemsets were computed manually using NumPy Boolean operations for efficiency. The top frequent itemsets display strong co-occurrence patterns consistent with high-volume retail items.

5.3 Extracted Rules

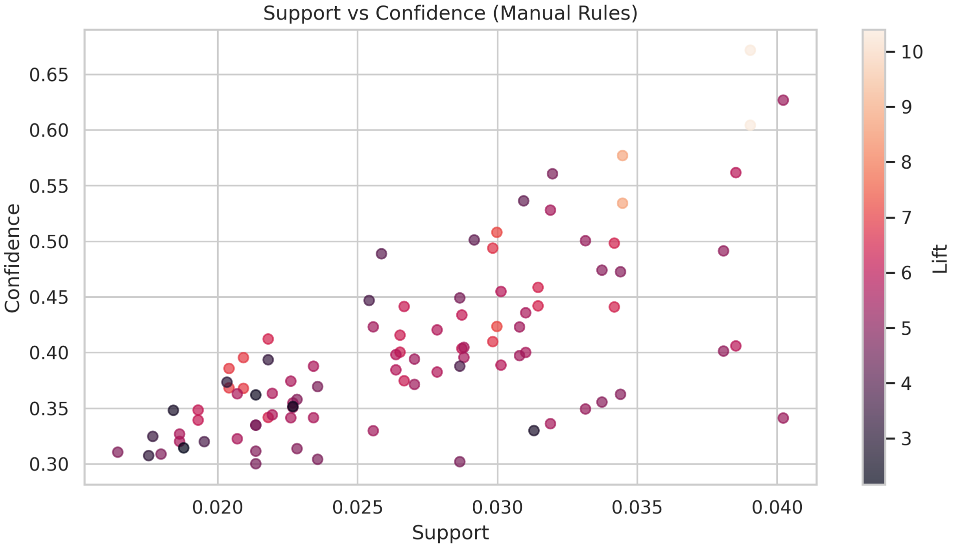

Figure 7: Support vs. confidence of all rules, colored by lift.

Example rules include:

- {A} → {B} with high confidence,
- {C} → {D} with moderate lift.

These rules reveal meaningful relationships between frequently co-purchased items.

6 Time Series Forecasting (Manual SES)

A Simple Exponential Smoothing model was implemented manually. The recursion is:

$$s_t = \alpha\, y_t + (1 - \alpha)\, s_{t-1},$$

and the forecast is:

$$\hat{y}_{t+h} = s_t.$$

6.1 Test Performance

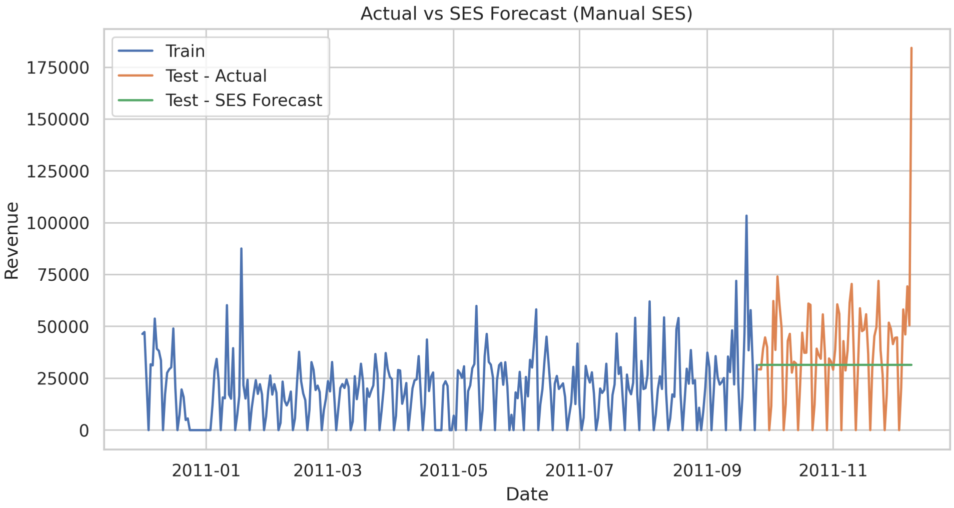

Figure 8: Actual vs SES forecast on test set.

6.2 RMSE Comparison

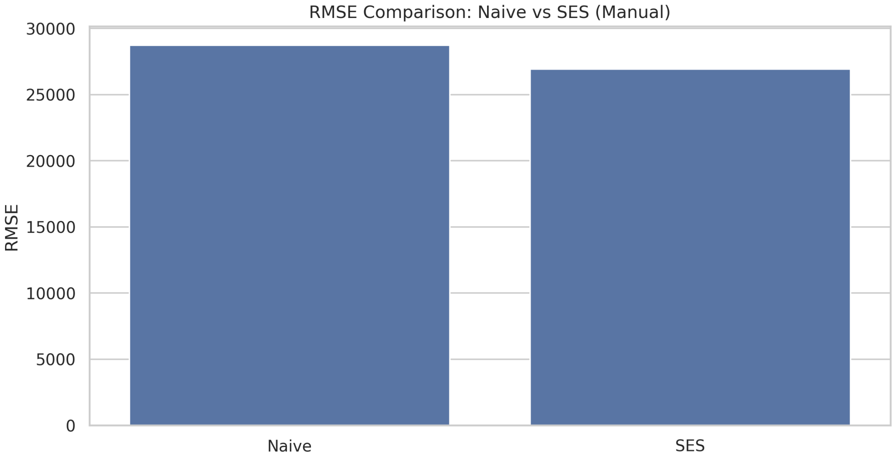

Figure 9: RMSE comparison between naive and SES forecasts.

6.3 14-Day Forecast

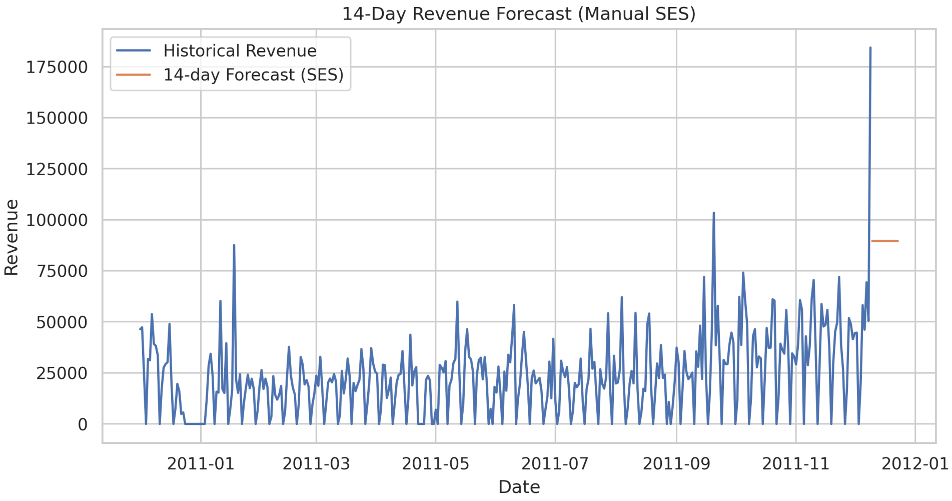

Figure 10: 14-day SES revenue forecast.

7 Conclusion

This project implemented association rule mining and time series forecasting entirely from scratch. The basket transformation and manual Apriori algorithm revealed several frequent itemsets and meaningful product associations. Support, confidence, and lift were computed without external libraries, providing insight into the mathematical structure of rule mining.

The second part of the project developed a fully manual SES model, which outperformed a naive baseline and generated stable revenue forecasts. The combined analysis illustrates both cross-sectional insights (product relationships) and temporal behavior (sales trends).

Future work may include multi-level itemset mining, FP-growth, Holt–Winters smoothing, and ARIMA models.

References

[1] Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases.

[2] Hyndman, R. J., & Athanasopoulos, G. *Forecasting: Principles and Practice.*

[3] Retail Transaction Dataset (Course Provided).