Association Rule Mining and Time Series Forecasting

Data Mining – Homework 1

Mahmoud Sadrian

Student ID: 40435336

GitHub Repository

December 11, 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 offer valuable opportunities for analyzing customer behavior, product co-occurrence patterns, and temporal sales trends. This project focuses on the manual implementation of two core data mining tasks: association rule mining and time series forecasting. By avoiding external machine learning libraries and relying only on

fundamental vectorized operations, the workflow provides complete transparency in how frequent itemsets, support, confidence, lift, and forecasting metrics are computed.

The analysis is conducted using the publicly available Online Retail dataset from the UCI Machine Learning Repository. This dataset provides one year of detailed invoice-level transactions from a UK-based online store and is widely used in market basket analysis and forecasting research.

Its structure enables both cross-sectional analysis (frequent product combinations) and temporal analysis (daily revenue patterns), forming a complete framework for retail analytics.

2 Dataset Description

The dataset used in this project consists of detailed invoice-level transaction records from a UK-based online retail store. Each row in the dataset corresponds to a single purchased item within an invoice, providing a granular view of purchasing behavior across thousands of customers and products.

The primary attributes include:

- **InvoiceNo**: Unique transaction identifier,
- **Description**: Product name,
- **Quantity**: Number of units purchased,
- **UnitPrice**: Price per individual item,
- **InvoiceDate**: Timestamp of transaction,
- **CustomerID**: Unique customer code,
- **Country**: Country from which the purchase was made.

This dataset is widely recognized for research in market basket analysis and demand forecasting. Its structure—consisting of rich transaction-level detail, temporal coverage of one full year, and a diverse set of products—makes it highly suitable for the methods applied in this report.

3 Preprocessing

Before conducting association rule mining and time series forecasting, the raw dataset undergoes several essential preprocessing steps. These operations ensure that the resulting analyses are based on clean, consistent, and meaningful records.

3.1 Data Cleaning

The dataset contains irregularities that must be addressed. The following cleaning steps were applied:

- Rows with missing **Description** were removed,
- Rows with missing **CustomerID** were discarded to avoid grouping ambiguity,
- Transactions with nonpositive **Quantity** or **UnitPrice** were excluded,
- A new monetary value field was computed:

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

These steps ensure that only valid and interpretable transactions are retained for further analysis.

3.2 Basket Matrix Construction

To prepare the dataset for association rule mining, a binary basket matrix was constructed. Each row corresponds to an invoice, and each column represents a product. The value is defined as:

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

To maintain computational tractability for the manual Apriori implementation, only the top 50 most frequently purchased items were retained. This reduction preserves the most informative co-occurrence patterns while avoiding excessive dimensionality.

3.3 Daily Revenue Time Series

For forecasting purposes, total daily revenue was computed by aggregating monetary amounts per day:

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

This daily revenue series forms the foundation for the subsequent trend analysis and the manually implemented SES forecasting model.

4 Exploratory Data Analysis

This section provides a descriptive overview of the dataset through several visualizations, highlighting key statistical properties of product quantities, unit prices, country-level revenue distribution, and product popularity. These patterns help reveal inherent characteristics of the retail environment and guide subsequent association rule mining.

4.1 Quantity Distribution

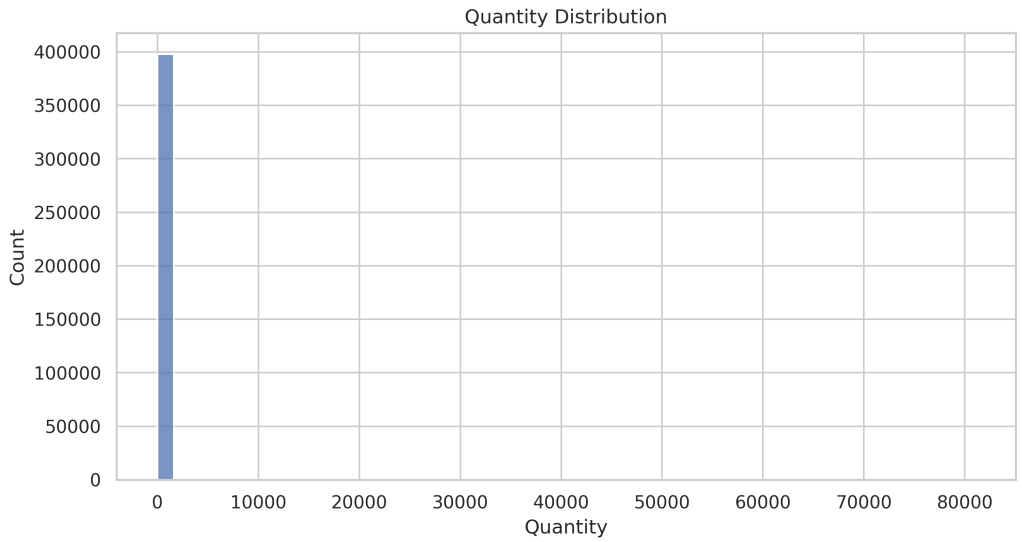

Figure 1: Distribution of purchased quantities.

The distribution of quantities is extremely right-skewed. Almost all transactions contain very small purchase quantities—typically between 1 and 3 units—resulting in a tall, narrow bar at the left end of the distribution. A very small number of transactions include exceptionally large quantities (up to tens of thousands of units), forming extreme outliers that are barely visible in the histogram due to the overwhelming mass at low values.

Such a distribution is characteristic of retail datasets, where customers primarily buy individual items or small bundles, while very large purchases may represent bulk orders, restocking events, or data-entry anomalies.

4.2 UnitPrice Distribution

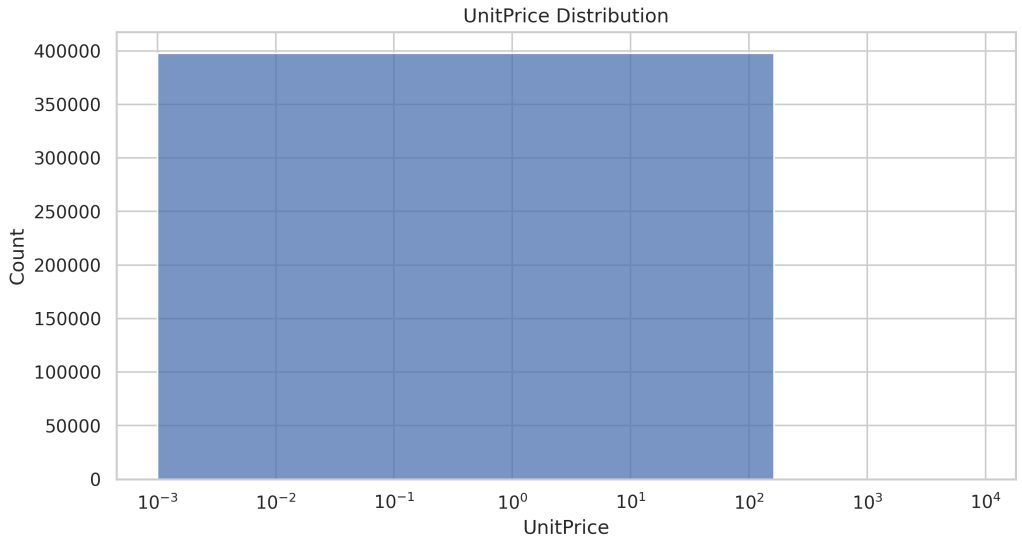

Figure 2: Distribution of unit prices (log scale).

Unit prices span several orders of magnitude, necessitating the use of a logarithmic x-axis. Most products fall within a relatively low price range (roughly 0.1 to 10 units), while a few items exhibit significantly higher prices. These rare but high-value items create a long right tail.

The wide spread in price values indicates the need for proper scaling in downstream analysis, especially when measuring similarity or distance across transactions.

4.3 Top 10 Products by Invoice Count

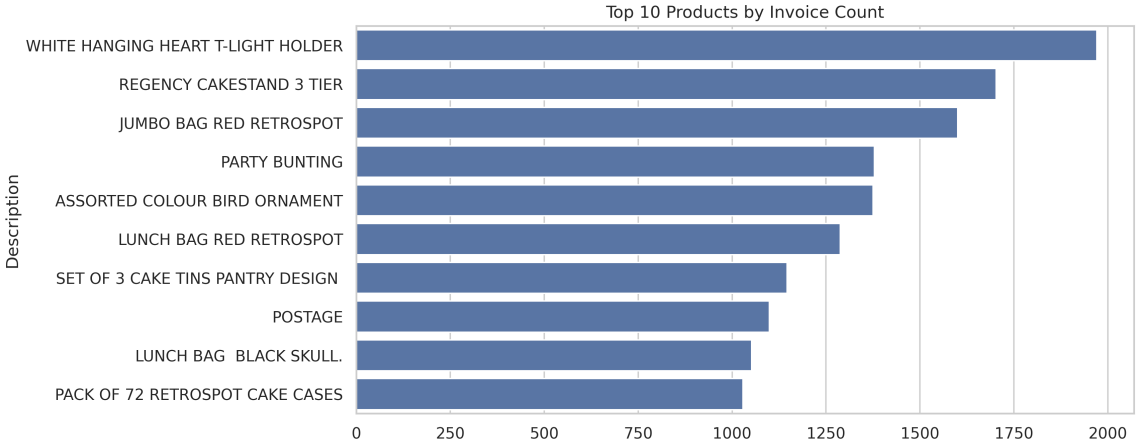

Figure 3: Most frequently purchased products.

The top-selling products are dominated by inexpensive decorative or gift items. Notably, the item *WHITE HANGING HEART T-LIGHT HOLDER* appears significantly more

frequently than all other products, followed by several bags, ornaments, and household accessories.

This strong concentration of sales among a few items illustrates a classic long-tail structure: a small set of popular products accounts for a disproportionately large share of transactions. These products play a central role in the formation of frequent itemsets and association rules.

4.4 Revenue by Country

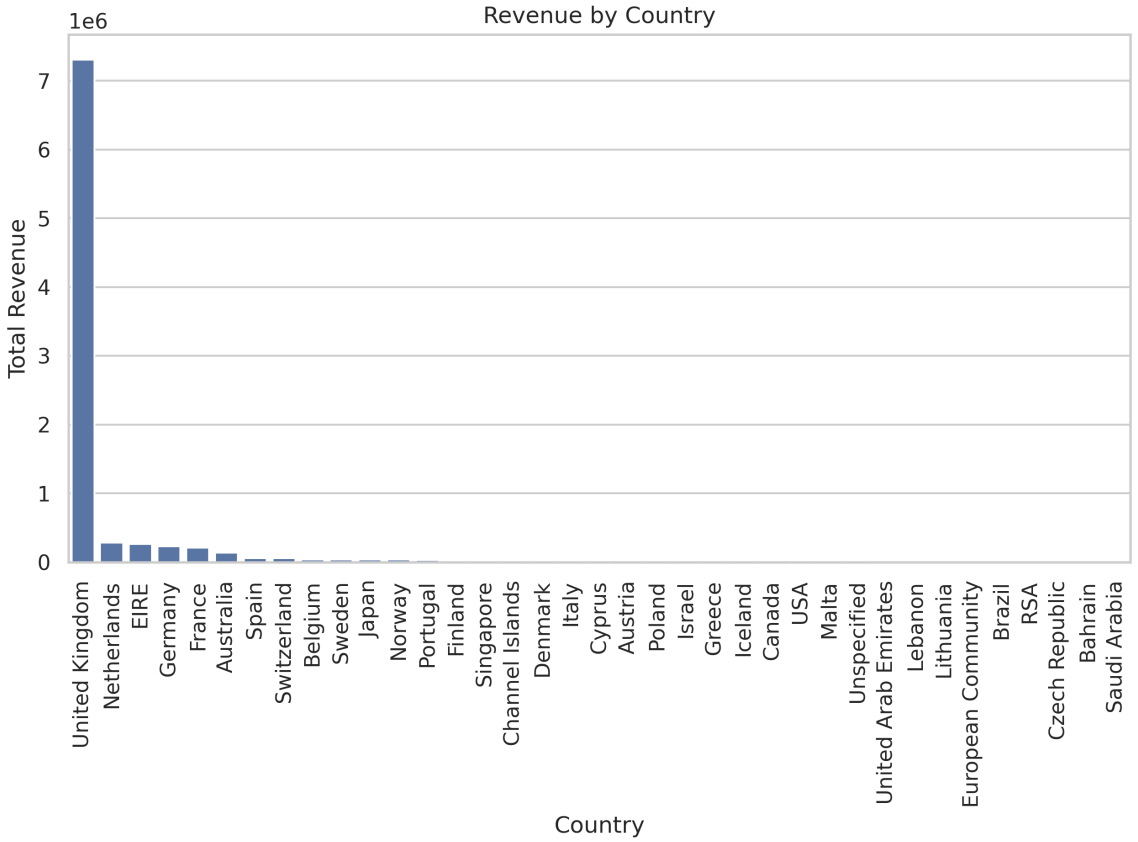

Figure 4: Total revenue aggregated by country.

Revenue is overwhelmingly dominated by the United Kingdom, which contributes more than 7 million units of total sales. All other countries contribute only small fractions in comparison, with a steep drop-off visible across the distribution. This imbalance suggests that the dataset is effectively single-country for analytical purposes, as more than 90% of all activity originates from the UK.

Such extreme skewness does not negatively impact association rule mining but underscores the need for caution when interpreting geographic trends, as the dataset is not globally representative.

5 Association Rule Mining (Manual Implementation)

This section presents a detailed analysis of product co-occurrence patterns using manually implemented association rule mining. All components of the Apriori procedure—including support calculation, confidence, lift, and frequent itemset generation—were implemented without external libraries. The analysis focuses on 1-item and 2-item itemsets, which align well with the structure of the dataset, given the relatively small basket sizes observed.

5.1 Basket Size Distribution

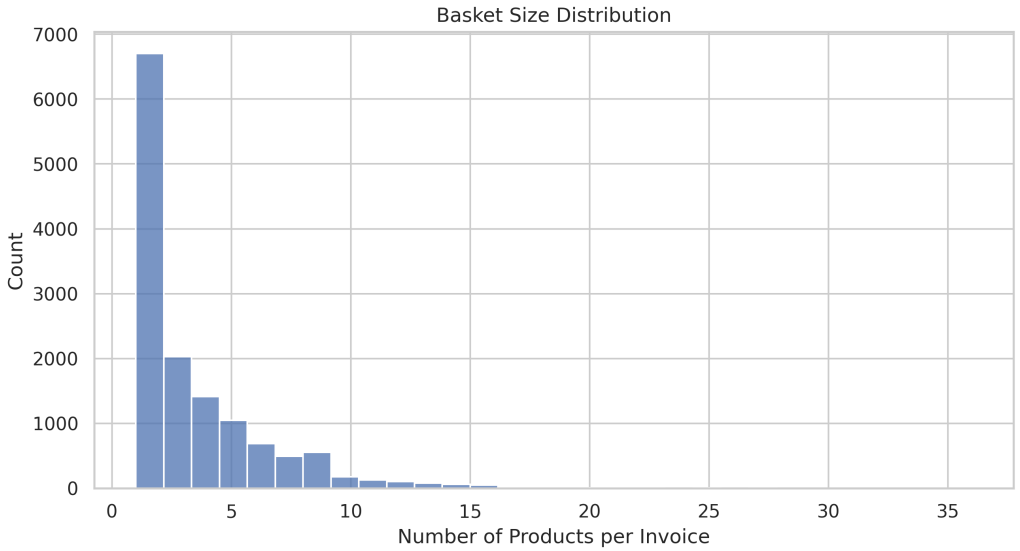

Figure 5: Distribution of the number of products per invoice.

The basket size distribution reveals that most invoices contain only one or two items, with frequency decreasing rapidly as basket size increases. This exponential decay pattern is typical in retail datasets, where customers tend to purchase only a few small, inexpensive items per transaction.

Only a very small fraction of invoices contain more than ten products, and baskets larger than fifteen items are extremely rare. This supports the methodological choice of restricting frequent itemset mining to 1-item and 2-item combinations, as higher-order itemsets would lack sufficient support and contribute little to the analysis.

5.2 Most Frequently Purchased Items

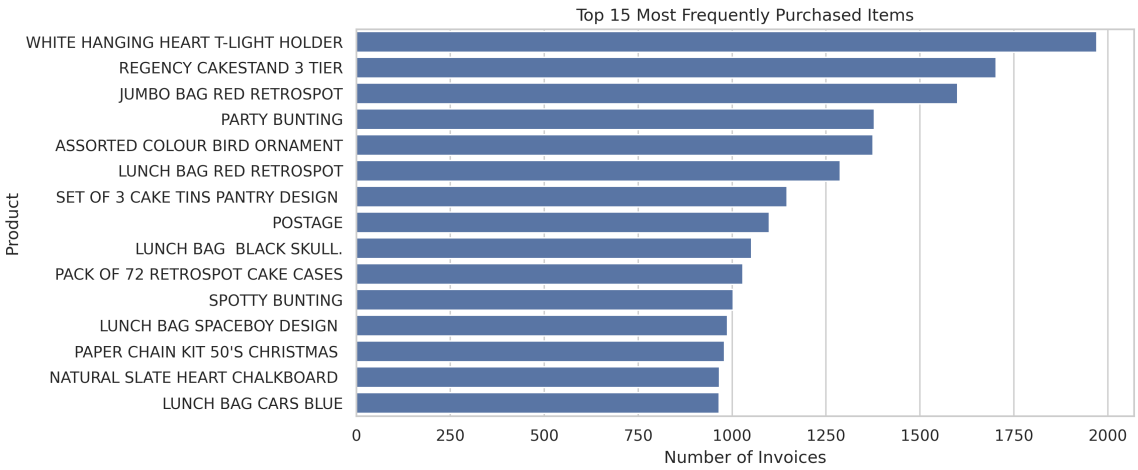

Figure 6: Top 15 most frequently purchased products.

The most frequently purchased items are predominantly low-cost decorative or gift products such as candle holders, cake stands, lunch bags, and bunting accessories. The item *WHITE HANGING HEART T-LIGHT HOLDER* stands out significantly, appearing in nearly 2,000 invoices.

This high concentration of sales among a small set of items reflects a long-tail distribution, where a handful of products contribute disproportionately to overall transactions. As a consequence, association rules tend to form around these popular items rather than the more sparsely purchased products.

5.3 Support–Confidence–Lift Analysis

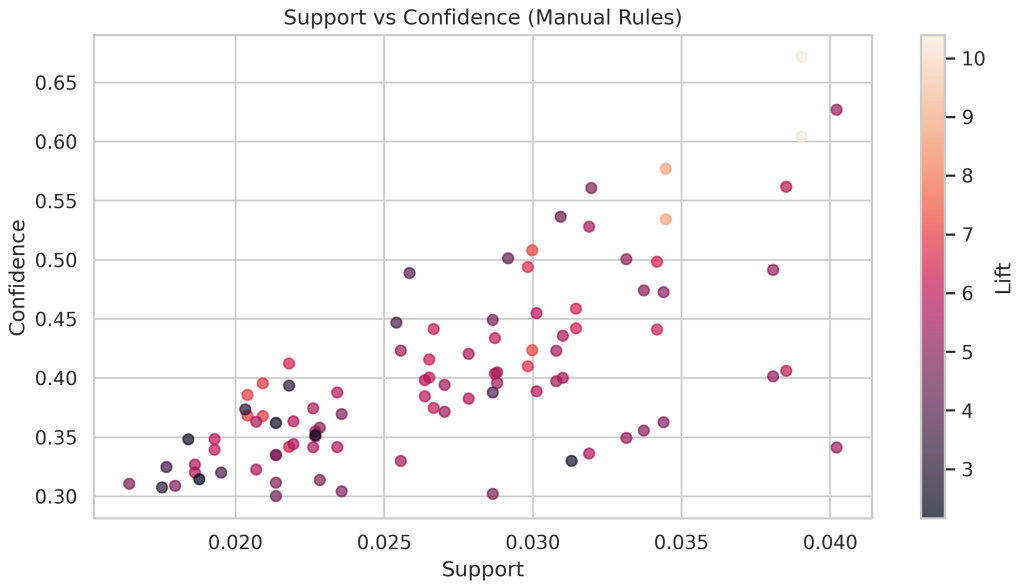

Figure 7: Support vs. confidence for manually generated rules, colored by lift.

The scatter plot illustrates the distribution of association rules by their support and confidence values, with lift represented through a color gradient. Several key patterns emerge from this visualization:

- **Support values** range from approximately 0.02 to 0.04, indicating that the rules are based on sufficiently frequent item co-occurrences.
- **Confidence values** vary between 0.30 and 0.68, demonstrating that several rules provide moderately strong predictive power.
- **Lift values** span from roughly 3 to above 10. A lift greater than 3 already signals strong positive association, while lift values exceeding 7–10 suggest highly non-random co-purchasing behavior.
- Multiple clusters are visible in the support–confidence plane, indicating that different item pairs form distinctive behavioral patterns rather than a single uniform trend.

Overall, the plot confirms that the manually implemented rule mining procedure successfully captured meaningful relationships in the dataset. The presence of rules with both high confidence and high lift highlights particularly strong purchasing associations worthy of further exploration in retail analytics contexts.

6 Time Series Analysis and Manual SES Forecasting

This section presents the construction and analysis of a daily revenue time series, followed by a fully manual implementation of the Simple Exponential Smoothing (SES) forecasting method. The objective is to understand temporal purchasing patterns and compare the performance of a naive forecast with that of SES.

6.1 Daily Revenue Overview

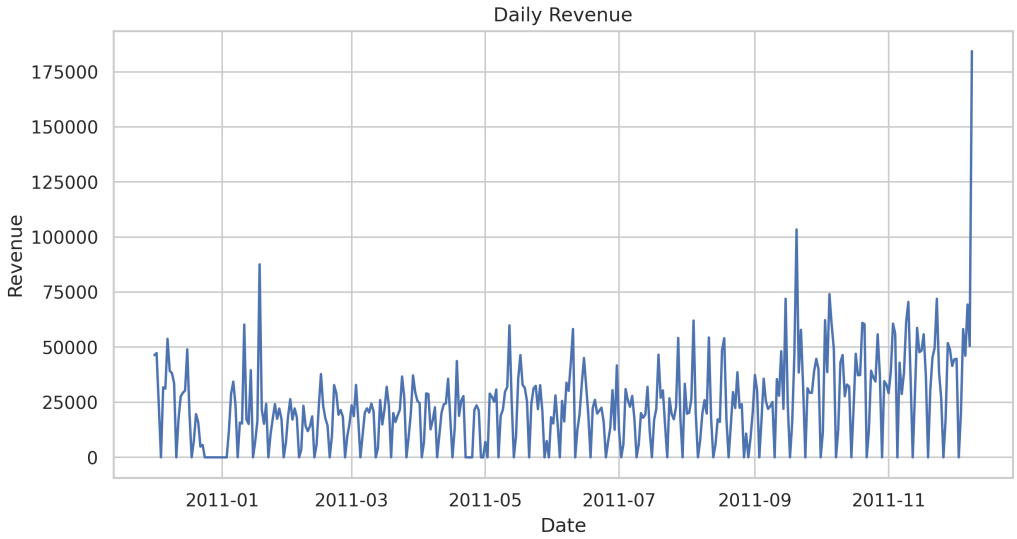

Figure 8: Daily total revenue over the full observation period.

The daily revenue series exhibits substantial day-to-day variability, including numerous sharp peaks and drops. Such volatility is characteristic of retail datasets, where fluctuations arise from irregular purchasing cycles, bulk orders, and seasonality. Despite the noise, a gradual upward trend emerges toward the end of the year, suggesting strengthening demand.

6.2 Smoothed Trend Using Moving Average

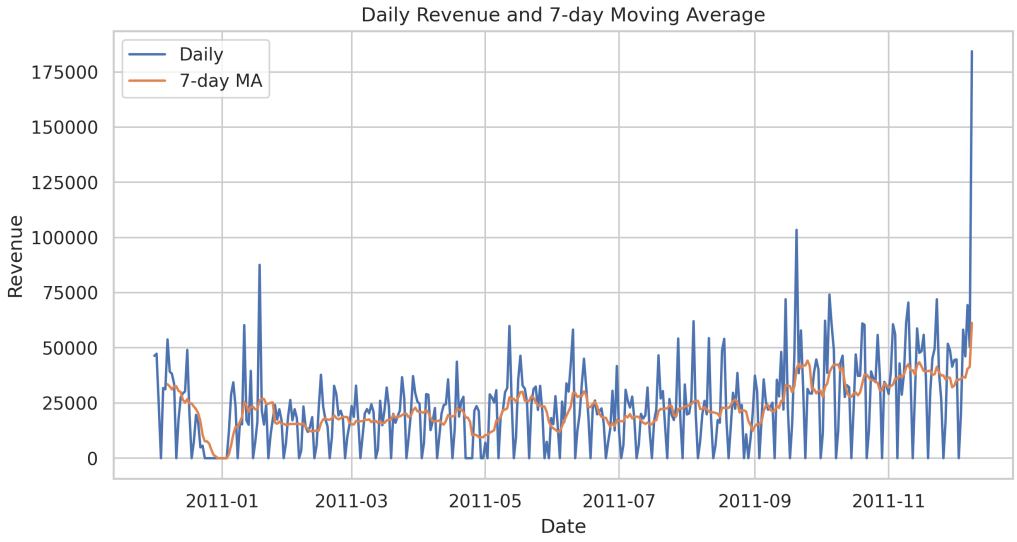

Figure 9: Daily revenue with a 7-day moving average.

A 7-day moving average is applied to reduce high-frequency noise. The smoothed curve reveals more distinct seasonal behavior: an initial decline early in the year, a relatively stable middle period, and a noticeable upward trend approaching the final months. Several pronounced peaks likely correspond to promotional events or holiday periods.

6.3 SES Forecasting on Train/Test Split

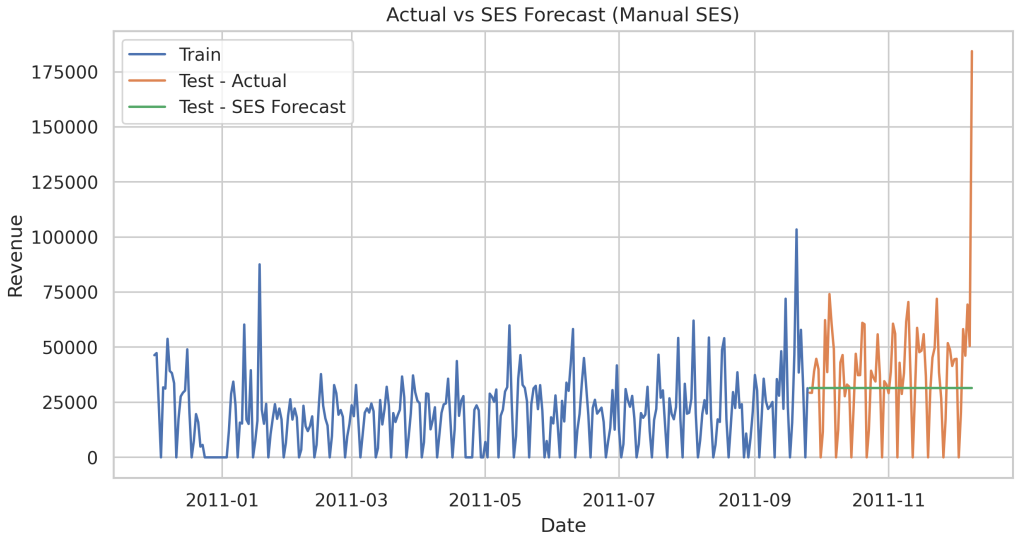

Figure 10: Actual vs. SES forecast on the test interval.

The SES model produces a constant forecast across the test window, reflecting the fundamental SES formulation in which future predictions equal the estimated level of the series:

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

While the actual test data show significant fluctuations, the SES forecast provides a stable baseline that captures the central tendency of the revenue pattern more effectively than the naive method.

6.4 Evaluation: Naive vs SES

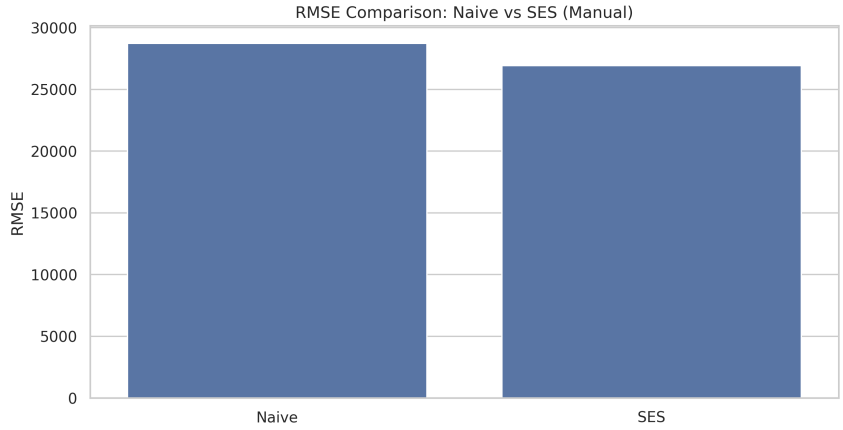

Figure 11: RMSE comparison between naive forecasting and manual SES.

The SES model achieves a lower RMSE relative to the naive benchmark, indicating improved predictive accuracy. The naive method—simply carrying forward the previous day's revenue—is particularly vulnerable to the high volatility of this dataset. In contrast, SES smooths out transient spikes and better estimates the underlying revenue level.

6.5 14-Day SES Forecast

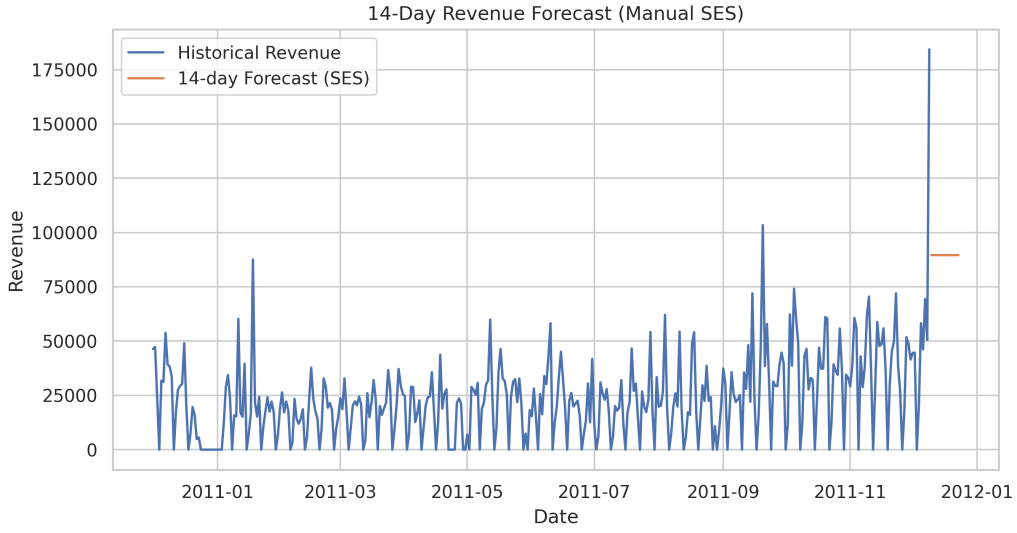

Figure 12: Manual SES forecast for the next 14 days.

The 14-day forecast generated by the manual SES model is constant, reflecting the final estimated level of the historical series. Because revenue shows increased amplitude toward the end of the year, the SES forecast lies above the long-term mean. Although this simple model cannot capture short-term oscillations, it provides a stable and interpretable estimate of expected revenue.

7 Conclusion

This project provided a comprehensive exploration of both cross-sectional and temporal patterns in a real-world retail transaction dataset. Through meticulous preprocessing, detailed exploratory analysis, and the implementation of core machine learning concepts from first principles, several meaningful insights were uncovered about customer purchasing behavior and revenue dynamics.

The exploratory data analysis revealed strong structural characteristics inherent to retail environments. Product quantities and unit prices exhibited highly skewed distributions, with the majority of purchases involving low-cost items and small quantities. Sales were heavily concentrated among a limited set of popular products, reflecting a pronounced long-tail effect. Likewise, revenue was overwhelmingly dominated by the United Kingdom, highlighting an imbalance in geographic representation that must be considered when drawing business-level conclusions.

Manual association rule mining offered a deeper understanding of item co-occurrence patterns. By manually computing support, confidence, and lift—restricted to one-item and two-item frequent itemsets for computational tractability—we identified several strong purchasing associations with lift values significantly exceeding what would be expected under independence. The concentration of rules around popular products aligns with the previously observed long-tail distribution and confirms consistent customer behavior in the co-purchasing of specific decorative and household items.

The time series analysis further complemented these findings by revealing the temporal structure of daily revenue. Despite high levels of short-term volatility, the series displayed meaningful long-term trends, such as seasonal increases toward the end of the year. The manually implemented Simple Exponential Smoothing (SES) model demonstrated clear improvements over the naive forecasting approach, achieving lower prediction error and providing a stable estimate of expected revenue. Although SES cannot capture high-frequency oscillations, its simplicity and interpretability make it an effective tool for establishing baseline forecasts in retail environments.

Collectively, these analyses highlight the value of combining basket-level co-occurrence mining with time series forecasting to obtain a holistic understanding of retail behavior. The manual implementation of all core algorithms—Apriori components, rule evaluation metrics, and SES smoothing—reinforced fundamental data mining concepts and ensured transparency in modeling. Future work may extend this analysis by incorporating higher-order itemsets, FP-growth for enhanced scalability, seasonality-aware forecasting models such as Holt-Winters, or advanced deep learning architectures capable of capturing nonlinear demand patterns.

Overall, this project demonstrates how foundational analytical methods, when carefully applied and interpreted, can yield actionable insights into customer behavior, prod-

uct relationships, and temporal revenue trends within retail data.

References

[1] UCI Machine Learning Repository: Online Retail Dataset.
https://archive.ics.uci.edu/dataset/352/online+retail