Week 2 Report

Sales Forecasting Based on Past Sales Statistics

Nguyen Minh Duc - 20225437

Week 2 Objectives:

- Exploratory Data Analysis: trends, seasonality, decomposition, category comparisons.
- Analyze relationships between behavioral data and sales.

---

1 Exploratory Data Analysis

1.1 Univariate analysis

We examined distributions for numerical and categorical variables to understand ranges, skewness, and outliers.

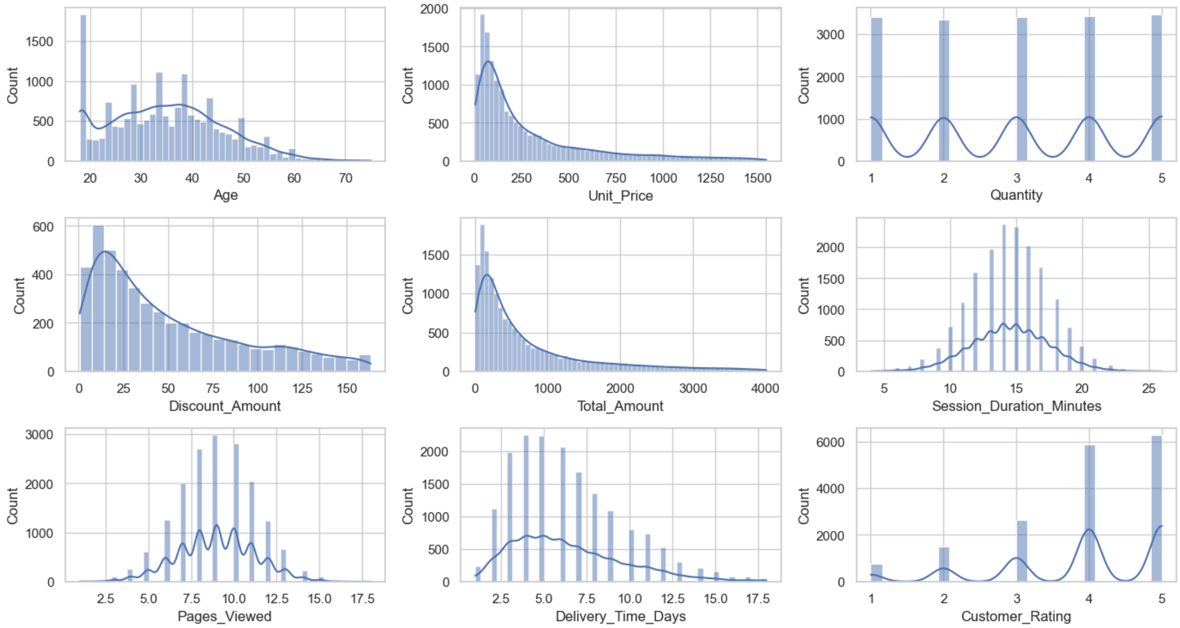


Figure 1: Univariate distribution of numerical columns

Key observations:

- Monetary variables (Unit_Price, Discount_Amount, Total_Amount) amd Age are strongly right-skewed; many zeros appear in `Discount_Amount`.
- Behavioral metrics (Session_Duration_Minutes, Pages_Viewed) show more symmetric / near-normal shapes.

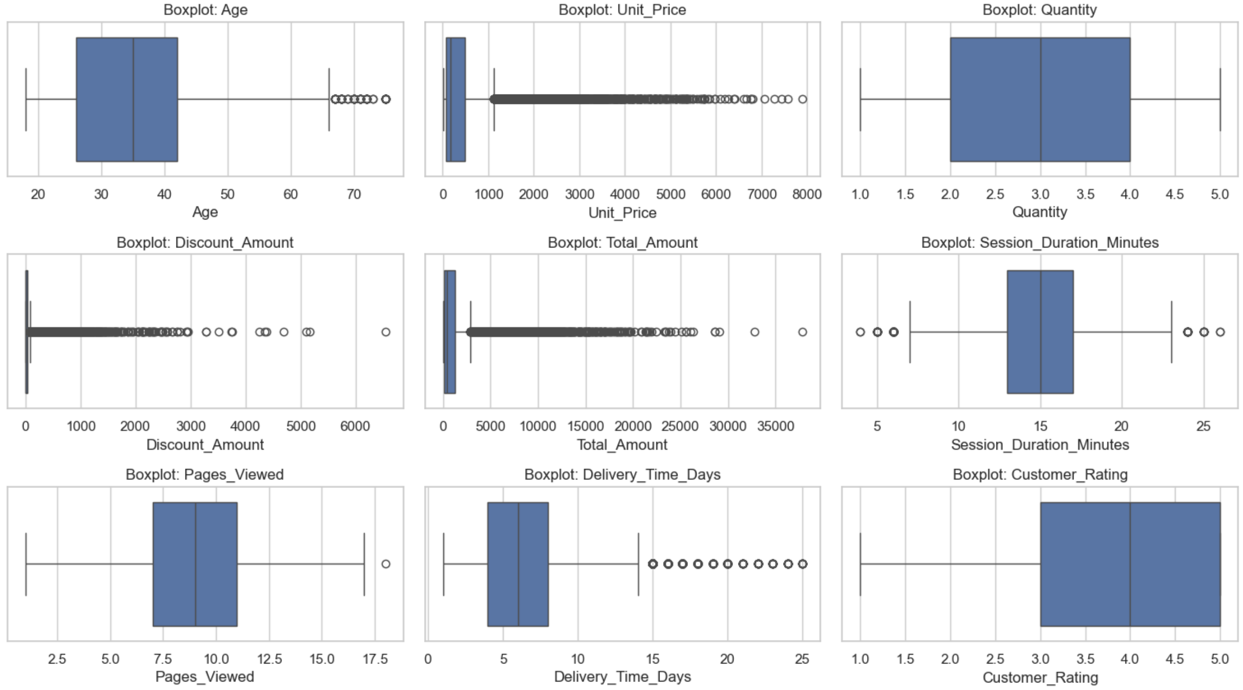


Figure 2: Boxplots of numerical columns

1.2 Categorical distributions

Categorical variables were summarized by proportion to identify dominant categories and potential class imbalance.

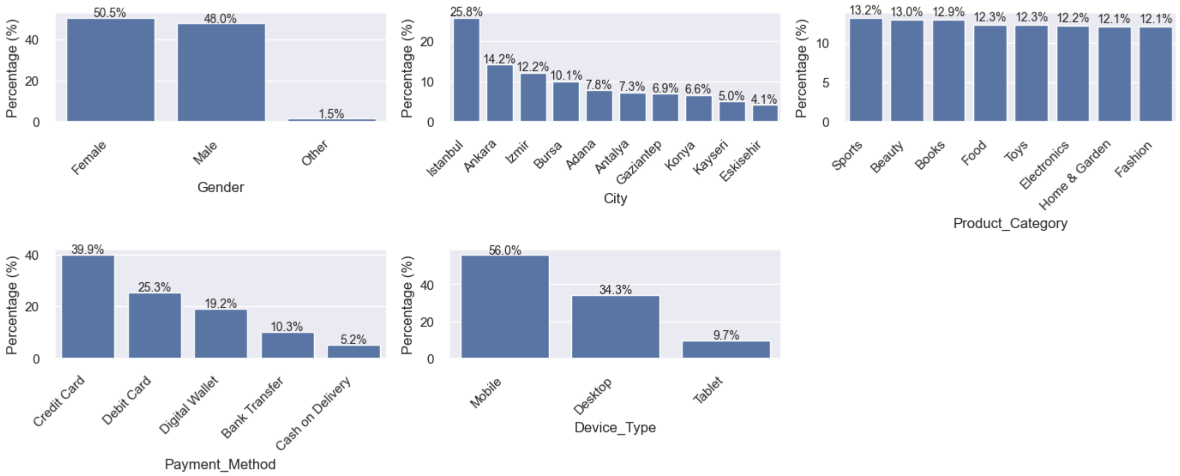


Figure 3: Distribution of categorical columns

Key observations:

- Gender and Product_Category display a roughly even distribution between categories.
- For City, Payment_Method and Device_Type, certain items are more prominent than others.

1.3 Correlation analysis

A correlation heatmap between numeric variables was computed.

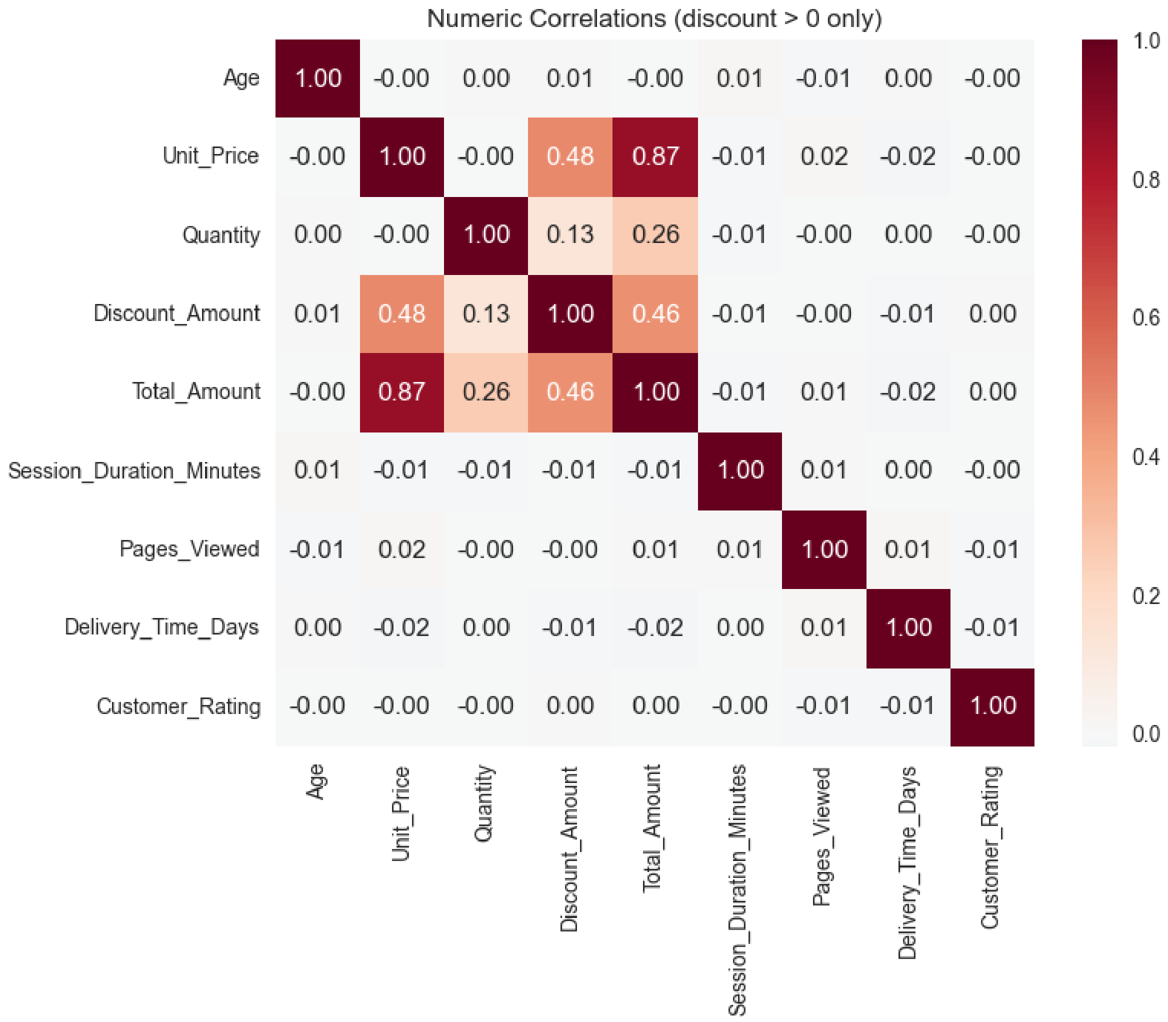


Figure 4: Correlation heatmap between numerical columns

Note that for the correlation heatmap above, correlations involving Discount_Amount only consider entries where Discount_Amount is non-zero to prevent spurious correlations. Key observations:

- Most numerical columns do not correlate with others.
- The only correlations recorded were with monetary columns (Unit_Price, Quantity, Discount_Amount and Total_Amount).

1.4 Daily revenue and rolling statistics

Daily revenue was computed by aggregating order totals per day. Rolling mean and rolling standard deviation were plotted to visualize short-term trends and volatility.

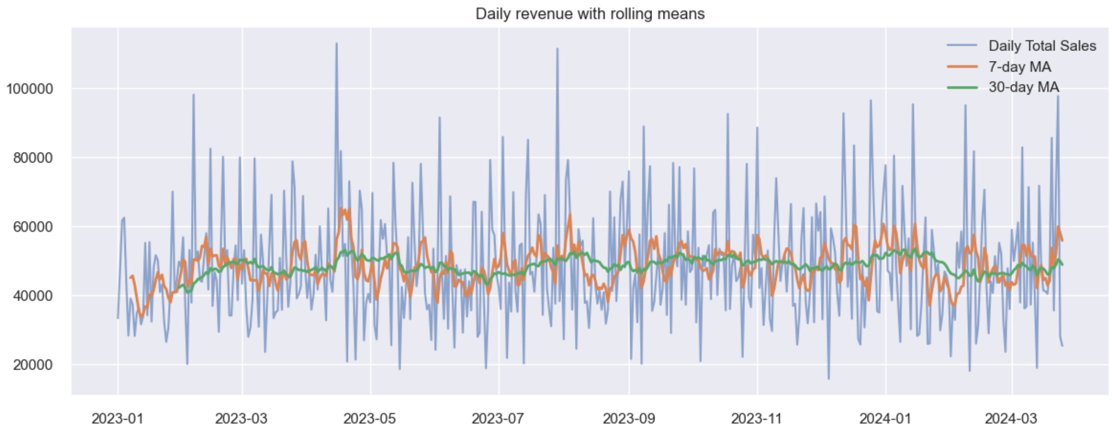


Figure 5: Rolling means of daily revenue

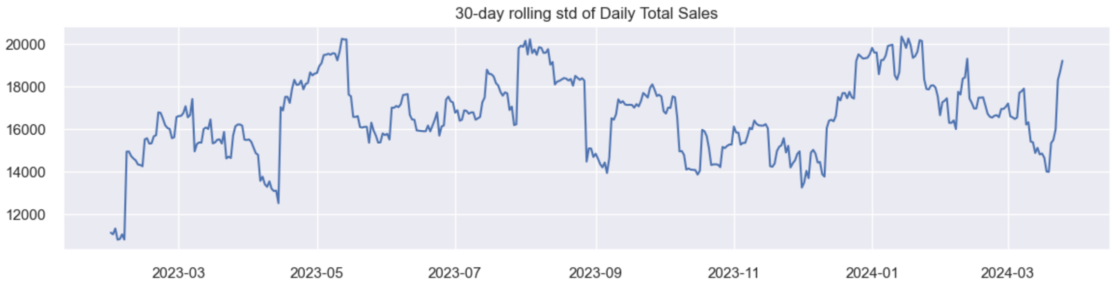
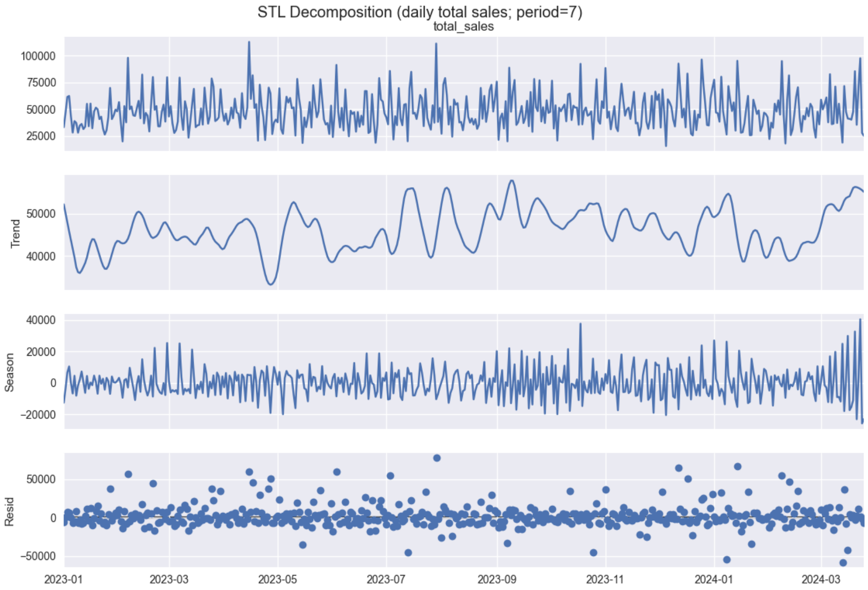
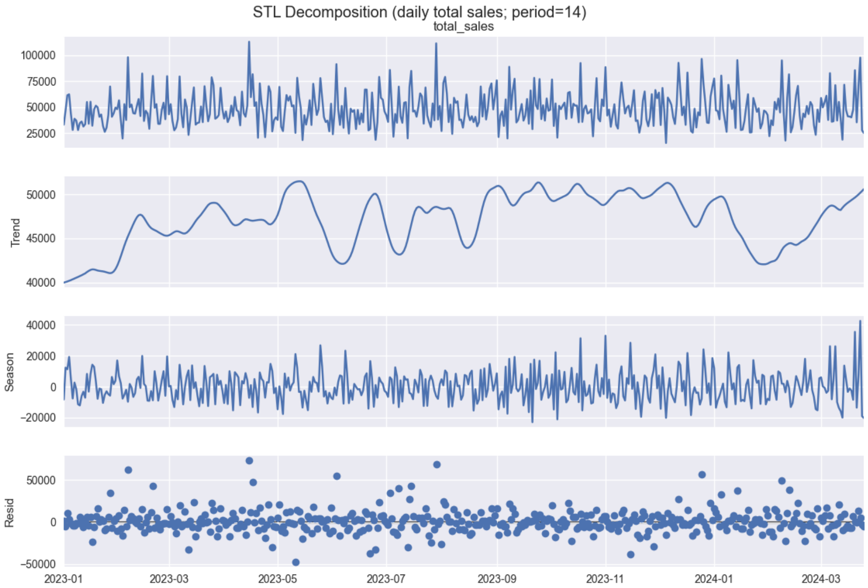


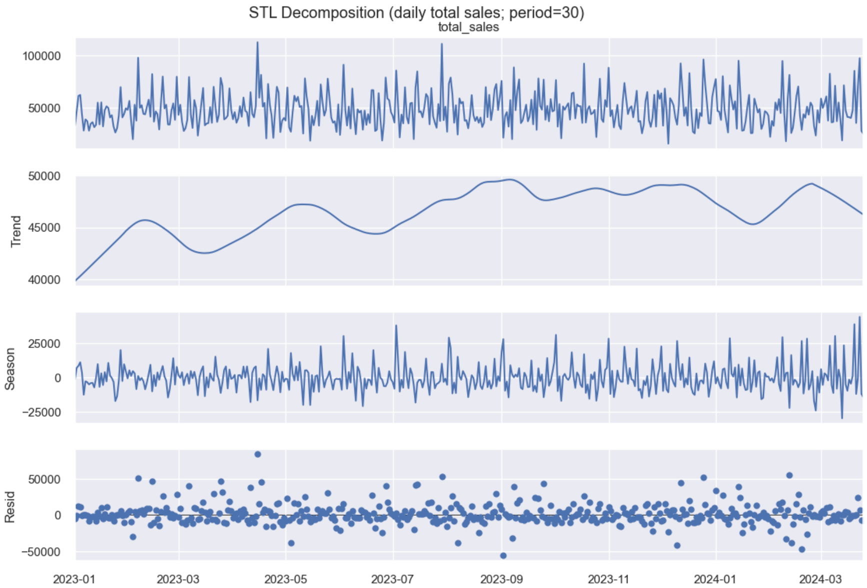
Figure 6: Rolling STD of daily revenue

1.5 STL Decomposition at multiple seasonalities

STL decomposition was run with periods of 7, 14 and 30 days to inspect seasonal components at different granularities.







Interpretation: STL results indicate only weak weekly seasonality (small repeating effects) and almost no meaningful monthly or multi-week seasonality. Residuals dominate the variance.

1.6 Autocorrelation analysis (ACF & PACF)

ACF and PACF plots were computed on the daily total sales series. The results reinforce the findings of STL decomposition, where only minor ACF & PACF spikes were recorded at 7 and 14-day marks. No major spikes indicate a lack of temporal relationship within the daily sales amount.

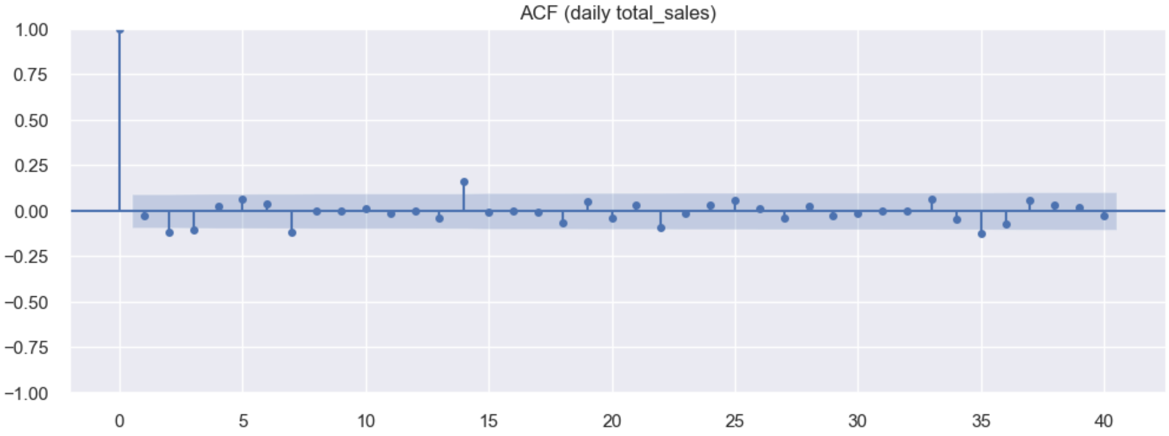


Figure 7: ACF of the daily total sales series

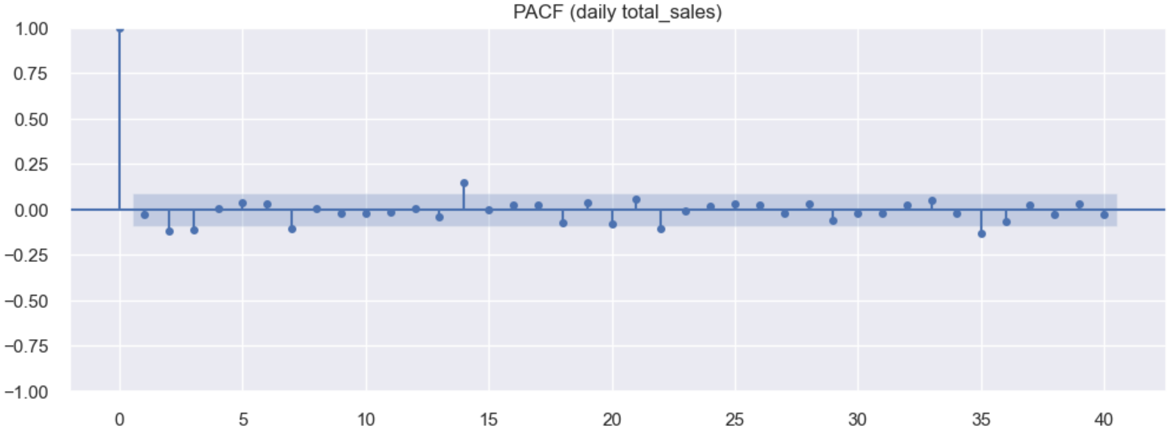


Figure 8: PACF of the daily total sales series

1.7 Spectral analysis (Periodogram)

A periodogram was computed to find dominant frequencies. The results show only weak spectral peaks.
Interpretation: The strongest spectral peak corresponds to a period of approximately 4.7 days. However, these peaks are weak and not sharply defined; the spectral evidence does not imply clear, stable periodic behavior.

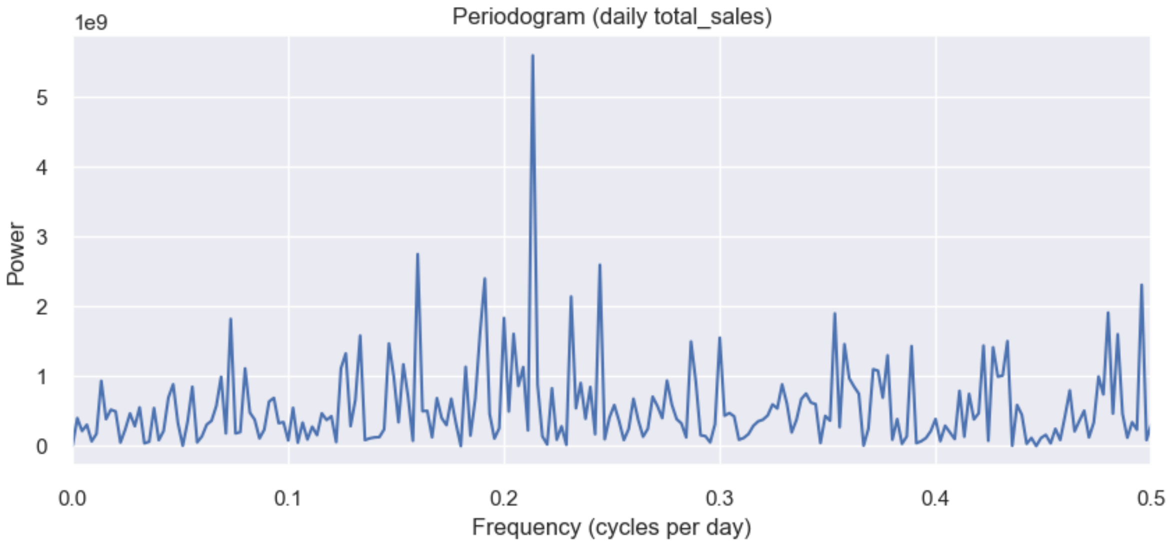


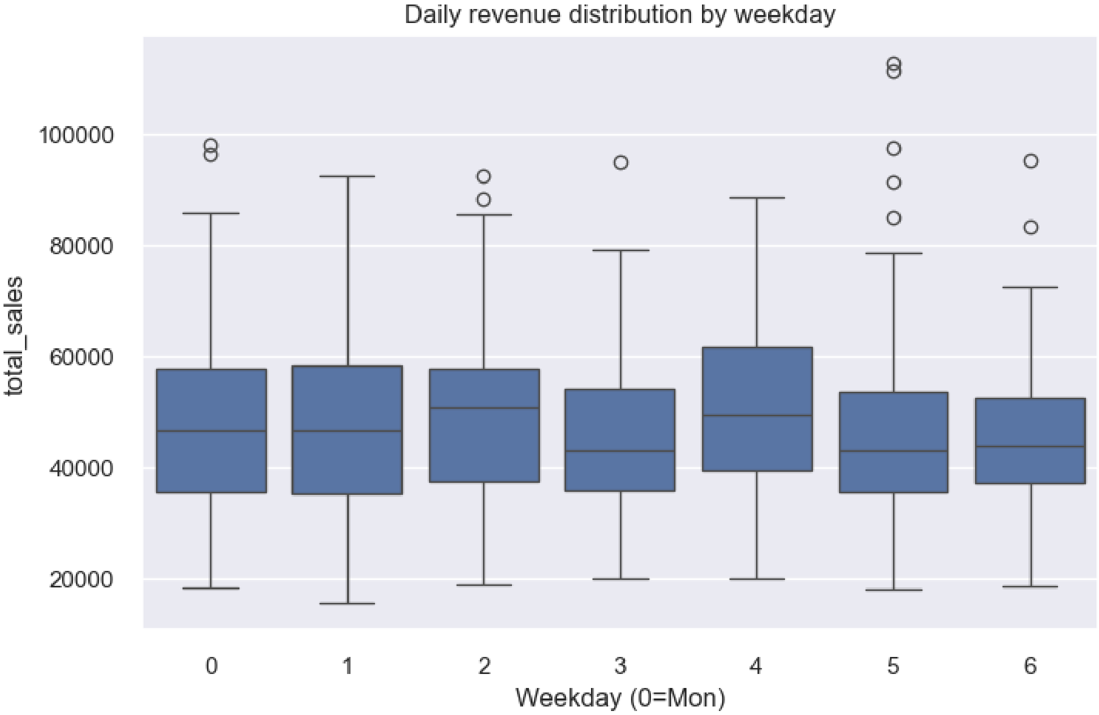
Figure 9: Periodogram Visualization

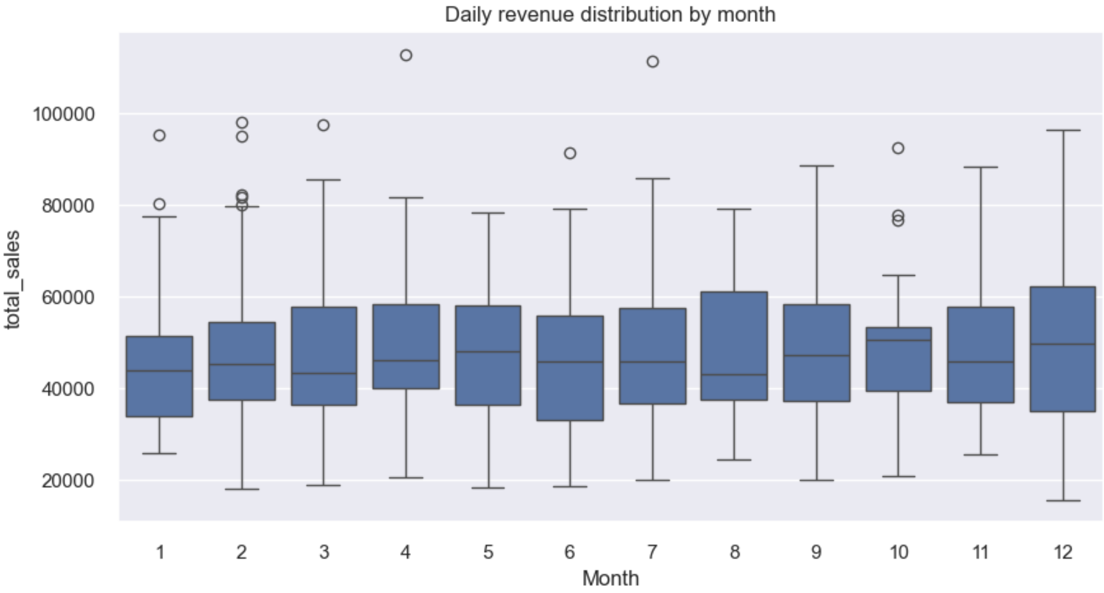
Table 1: Top periods (days) by spectral power from the periodogram.

Frequency (cycles/day)	Period (days)	Power
0.2133	4.7	5.598464695×10^9
0.1600	6.2	2.751054482×10^9
0.2444	4.1	2.597852328×10^9
0.1911	5.2	2.400743403×10^9
0.4956	2.0	2.311156409×10^9
0.2311	4.3	2.142924413×10^9
0.4800	2.1	1.912746901×10^9
0.3533	2.8	1.900237103×10^9

1.8 Revenue by weekday and by month

We plotted daily revenue grouped by weekday and by calendar month to detect intra-week and intra-year patterns.





Interpretation: The weekday boxplot shows small differences between weekdays; however variance is large and overlapping, suggesting any weekday effect is weak compared to daily noise. Monthly boxplots do not show strong, repeating monthly peaks.

1.9 Holiday analysis

Holiday impact analysis (Turkish public holidays in 2023–2024) produced mixed results: some holidays show very large positive lifts on specific dates while many holidays show negligible or negative lifts. This inconsistency suggests that holidays are not reliable predictors of daily sales spikes.

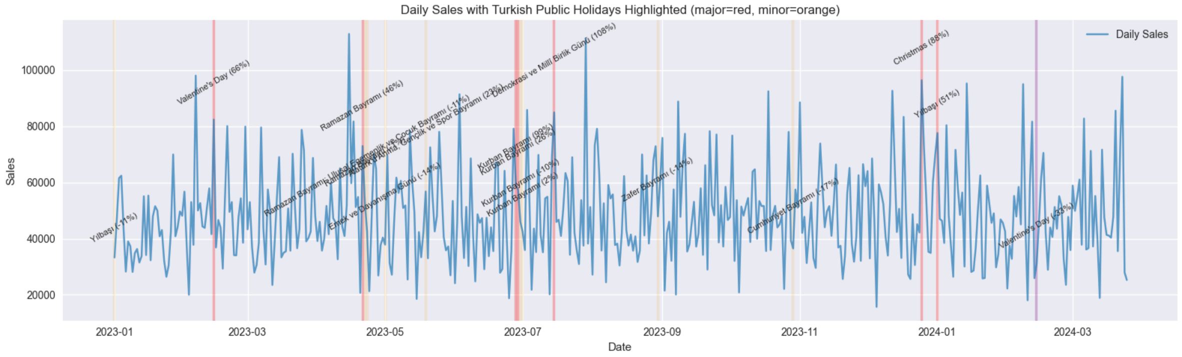


Figure 10: Daily Sales with national and international holidays highlighted

2 Conclusions and implications for modeling

2.1 Main conclusions from EDA

- The daily sales series is **highly noisy** and dominated by irregular fluctuations; neither strong trend nor stable seasonality is evident.
- STL, ACF/PACF and periodogram show only weak weekly signals (if any) and little evidence of robust periodicities. Spectral peaks at periods around 2–6 days exist but are weak and not stable.
- Monetary fields are strongly right-skewed; log-transformations (or other variance-stabilizing transforms) will likely improve model behavior.
- Behavioral features at aggregate (daily) level show weak correlation; consider using them at the order or customer level instead.

2.2 Implications and recommended modeling strategy

Based on the EDA, the following modeling decisions are recommended:

1. **Aggregation choice:** Use daily aggregation for business reporting, but evaluate weekly aggregation to reduce noise (weekly series may reveal more structure). Model selection should be aligned with the forecast horizon (daily vs weekly).

2. **Baseline models:** Start with simple baselines (naïve, seasonal naïve with week lag) to set a benchmark.

3. **Transformations:** Apply `log1p` to monetary series (Total_Amount, Unit_Price) to stabilize variance and reduce skewness.

4. **Feature engineering:**

   - Lag features (1, 7, 14, 30) and rolling statistics (mean, std) at multiple windows.
   - Categorical encodings for `Product_Category`, `City`, `Payment_Method`, `Device_Type`.
   - Customer-level features: returning-customer flag, average basket size, recency/frequency.

5. **Model families to try:**

   - **Tree-based models** (LightGBM/XGBoost) on engineered features - expected to perform well on noisy, tabular datasets.
   - **Gradient-boosted regression with time features** (lags + rolling stats + categorical indicators).
   - **Prophet / SARIMAX** with exogenous regressors (holidays, promotions) - useful as explainable baselines, but ARIMA/SARIMA may be weak since autocorrelation is low.
   - **Neural models (LSTM/GRU)** if non-linear, long-range dependencies are suspected; these need careful tuning and sufficient data.
   - **Hybrid / ensembles** combining statistical and ML models to capture different signal components.

6. **Validation:** Use time-series cross-validation (rolling-origin / expanding window). Evaluate using multiple metrics (RMSE, MAE, MAPE) and inspect residuals.

7. **Feature selection:** Conduct permutation importance / SHAP analysis on tree models to identify useful predictors (discounts, returning-customer, session metrics, category effects).