CEROP: Cross-Border Enterprise Risk & Operations Predictor

Week 3: Getting and Exploring the Data

Group 4

Ankit Kumar, Arthunya Kanoklertwongse, Aswin Shriram Thiagarajan, and Divya Gopal

Westcliff University

AIT 506: Machine Learning

Professor Oueichek Ibba

August 10, 2026

Part I: Financial Distress Layer

Getting and Exploring the Data

Financial Distress Layer — Taiwan Bankruptcy Dataset (Arthunya)

Data Loading

The Taiwanese Bankruptcy Prediction Dataset was loaded via the UCI Machine Learning Repository, confirming the expected structure of 6,819 firms across 96 columns (95 financial ratio features plus the binary target variable, Bankrupt?). An initial completeness check returned zero missing values across all columns.

```
# ---- 1. Basic inspection ----
print("Shape:", df.shape)
print(df.columns.tolist()[:10], "...")
df.head()

Show hidden output

# ---- 2. Confirm target column & class balance ----
target_col = "Bankrupt?"  # check colum name
print(df[target_col].value_counts())
print(df[target_col].value_counts(normalize=True) * 100)
```

Code Snippet 1. Loading the dataset and confirming target variable / class balance.

Target Variable and Class Balance

The target variable confirms a heavily imbalanced class distribution: 96.77% non-bankrupt (6,599 firms) versus 3.23% bankrupt (220 firms), consistent with the figure cited from Liang et al. (2016) in the Week 2 submission. This confirms the need for the Precision, Recall, F1-score, and AUC-ROC evaluation approach already committed to in the Frame the Problem document, rather than relying on raw accuracy.

Correlation Analysis

Correlation analysis identified Debt ratio % (r = 0.250), Current Liability to Assets (r = 0.194), and Borrowing dependency (r = 0.177) as the features most strongly associated with bankruptcy, alongside Net Income to Total Assets (r = -0.315) and ROA(A) before interest and % after tax (r = -0.283) as the strongest negative predictors. Two features, Current Liabilities/Equity and Current Liability to Equity, returned identical correlation values (r = 0.154), indicating likely multicollinearity that will need to be addressed during Week 4 feature selection.

Visualizations

Histograms of the six features most strongly correlated with bankruptcy (by absolute value) were generated to visually assess separability between bankrupt and non-bankrupt firms.

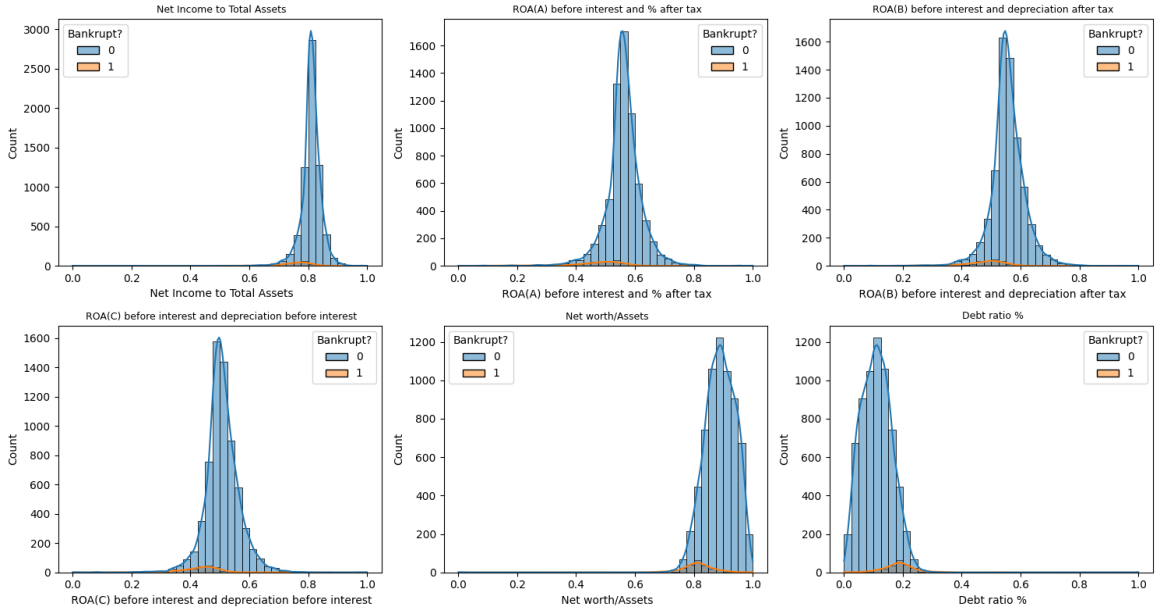


Figure 1. Distribution of top correlated financial ratios by bankruptcy status (0 = non-bankrupt, 1 = bankrupt).

A correlation heatmap of the eleven most relevant features was also produced to visualize relationships among the strongest predictors and the target variable.

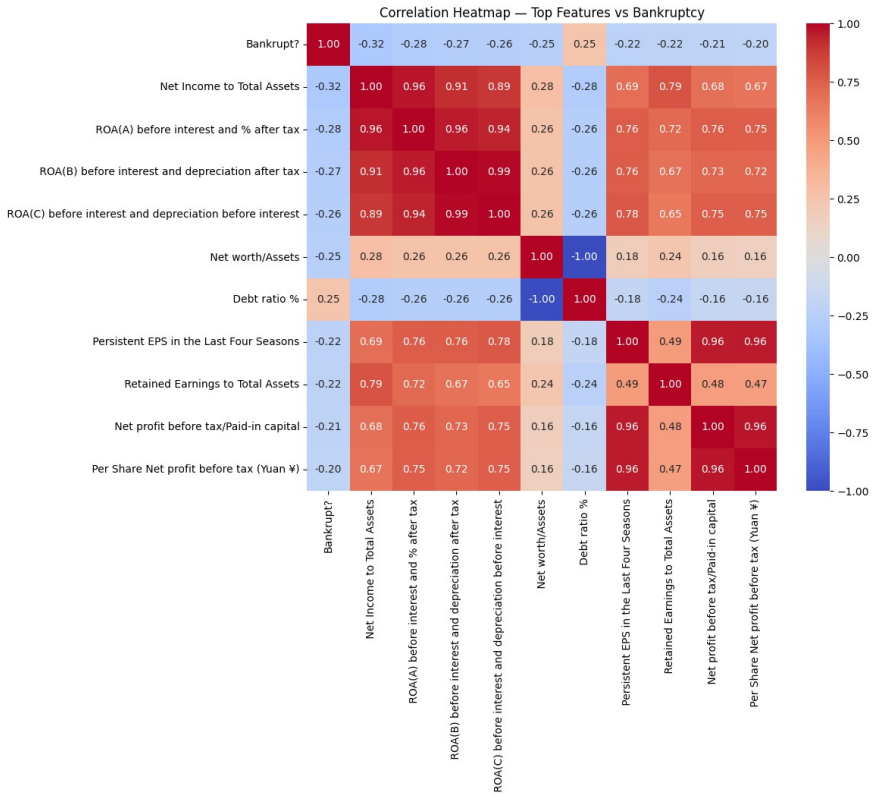


Figure 2. Correlation heatmap of the top 11 features most associated with bankruptcy.

Data Quality Deep-Dive: Outlier and Sentinel Value Analysis

A more detailed outlier check was conducted after the six top-correlated features were found to be cleanly bounded between 0 and 1 — a result that prompted verification across the full feature set rather than assuming this pattern held throughout.

```python
# ---- 7B. Check normalization range across ALL numeric columns ----
numeric_cols = df.select_dtypes(include=[np.number]).columns
numeric_cols = numeric_cols.drop(target_col)  # not include target column

range_summary = pd.DataFrame({
    "min": df[numeric_cols].min(),
    "max": df[numeric_cols].max()
})

# check column [0, 1] match (if so, normalized)
range_summary["normalized_0_1"] = (
    (range_summary["min"] >= 0) & (range_summary["max"] <= 1)
)

print(f"Total numeric columns (excl. target): {len(numeric_cols)}")
print(f"Columns within [0, 1] range: {range_summary['normalized_0_1'].sum()}")
print(f"Columns NOT within [0, 1] range: {(~range_summary['normalized_0_1']).sum()}")

print("\n--- Columns that are NOT normalized (min/max outside [0,1]) ---")
not_normalized = range_summary[~range_summary["normalized_0_1"]]
print(not_normalized)
```

Code Snippet 2. Full-column check for min-max normalization across all 95 features.

Extending the range check to all 95 numeric features found that 71 features fell within a normalized [0, 1] range, while 24 did not. Critically, these 24 non-normalized features showed maximum values in the billions despite a minimum of 0, a pattern inconsistent with realistic financial ratios and indicative of sentinel or placeholder values used in place of true missing data — which explains the apparent absence of missing values noted earlier.

Row-level analysis separated these 24 features into three tiers based on the percentage of records affected:

Tier	# Features	Description
Minor (< 2% affected)	15	Isolated data-entry errors; correctable by capping at the 99th percentile (e.g., Current Ratio 0.01%, Quick Ratio 0.13%, Cash/Current Liability 0.67%)
Borderline (~3% affected)	1	Interest-bearing debt interest rate (3.24% of records affected); flagged for closer review before deciding to cap or exclude
Severe (18–88% affected)	8	Majority of records affected; too little reliable data remains to justify

imputation (see list below)

The eight severely affected features are: Total Asset Growth Rate (88.2% of records), Cash Turnover Rate (62.2%), Research and Development Expense Rate (59.1%), Inventory Turnover Rate (43.2%), Quick Asset Turnover Rate (34.9%), Operating Expense Rate (33.5%), Current Asset Turnover Rate (18.1%), and Fixed Assets Turnover Frequency (18.0%). Importantly, none of these eight features appear among the top correlated predictors identified above, so this data quality issue does not compromise the features prioritized for modeling.

Implications for Week 4 (Preparing Data)

Based on this analysis, the financial-distress layer requires a tiered data cleaning strategy rather than a single uniform approach: (1) the 71 already-normalized features and the 15 minor-outlier features will be capped at the 99th percentile where needed and scaled normally; (2) the borderline feature (Interest-bearing debt interest rate) will be reviewed individually before a capping or exclusion decision is made; and (3) the 8 severely affected features will be flagged as candidates for exclusion from the final modeling feature set, since imputing values for columns where the majority of data is unreliable would introduce more noise than signal.

Operations Layer — DataCo Smart Supply Chain Dataset (Aswin)

Data Loading and Shape

The DataCo Smart Supply Chain dataset (Constante et al., 2019) was loaded via Kaggle, confirming 180,519 order-line records across 53 raw columns. Seven columns containing personally identifiable or non-analytical information (customer names, email, password, street address, product image, and description fields) were removed as an initial structural pass, bringing the working dataset to 37 columns ahead of the deeper cleaning planned for Week 4.

Missing Value Check

A completeness check across the retained columns confirmed the dataset was largely complete, with missing values concentrated in a small number of fields (notably Order Zipcode and Product Description) flagged for handling during Week 4 preprocessing rather than at the exploration stage.

Target Variable Distribution

Late_delivery_risk, the operations-layer target, showed a moderate class split: roughly 55% of orders were flagged as late, versus approximately 45% delivered on time. This near-even balance stands in sharp contrast to the severe 3.23%/96.77% imbalance found in the Financial layer, meaning the two layers ultimately need different evaluation strategies — the Operations layer can rely more heavily on standard accuracy-adjacent metrics, while the Financial layer cannot.

Descriptive Statistics and Outlier Check

An IQR-based outlier scan across the numeric columns found the largest concentration of extreme values in Benefit per order (18,942 outliers, 10.49% of records) and Sales per customer (1,943 outliers, 1.08% of records). Unlike the Financial layer's sentinel-value problem — placeholder values in the billions masquerading as clean data — these outliers reflect genuine business variance, such as orders with unusually large losses or unusually large per-customer sales, and were flagged for IQR-based clipping rather than removal in Week 4, so real transactions are not discarded.

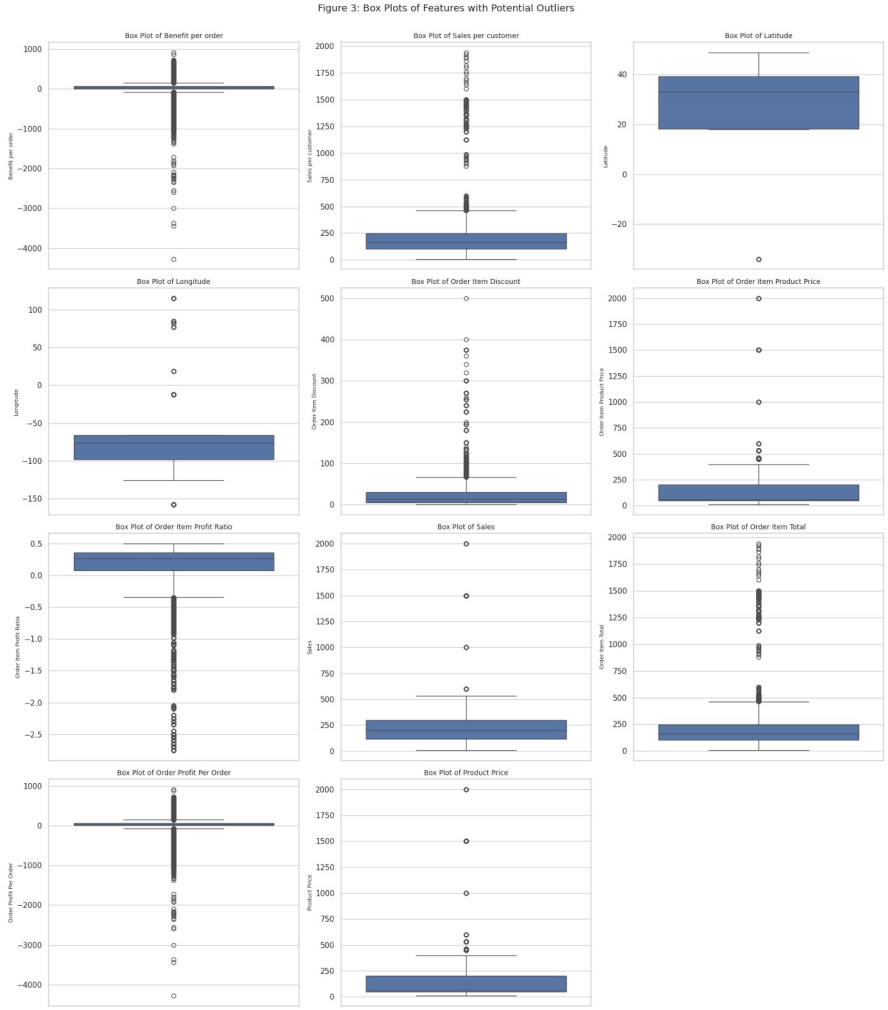


Figure 3. Box plots of key numeric operations features showing the spread of values before outlier treatment.

Univariate Analysis: Histograms

Histograms of key numeric features (Benefit per order, Sales per customer, Days for shipping (real), Order Item Discount, Order Item Product Price, and Order Item Quantity) showed right-skewed distributions typical of transactional order data, with most orders clustered at low values and a long tail of larger orders.

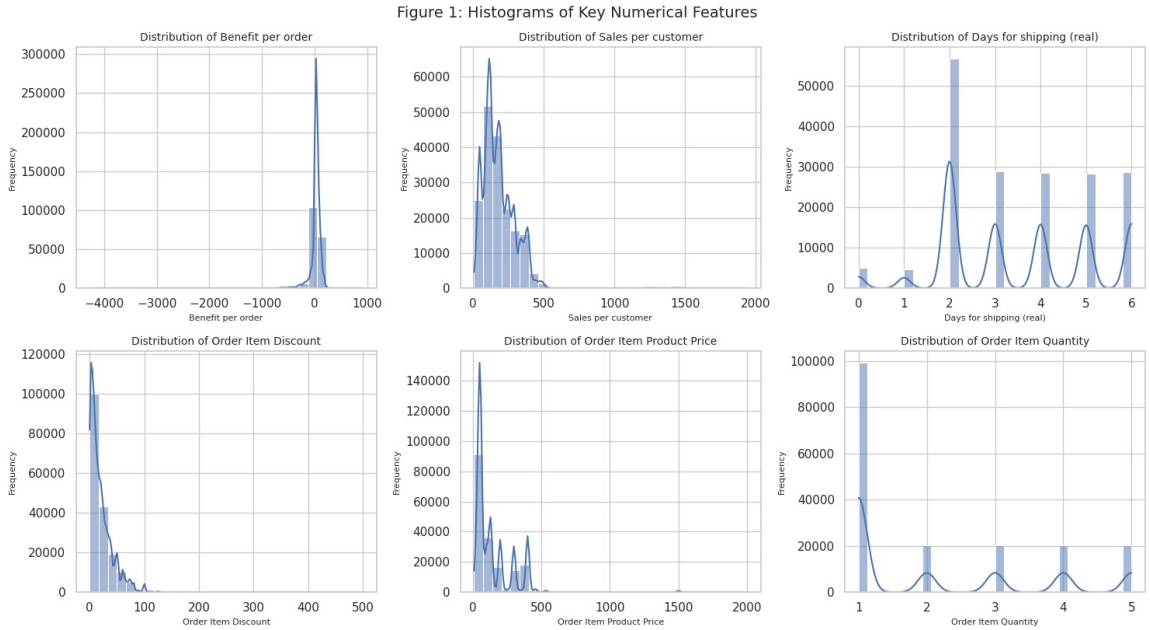


Figure 4. Distribution of six key numeric operations features.

Bivariate Analysis: Correlation Matrix

A correlation heatmap against Late_delivery_risk identified Days for shipping (real) (r = 0.40) and Days for shipment (scheduled) (r = -0.37) as by far the strongest linear associations with late delivery; every other numeric feature (Sales, Order Item Total, Product Price, discount and profit fields) correlated at or near r = 0.00. This finding carries a direct warning forward to Week 4: Days for shipping (real) records the actual number of days a shipment took, which is only known once the delivery has already happened. Using it as a predictor would leak the outcome directly into the model, which is why this column — along with the related Delivery

Status field — was excluded during Week 4 preprocessing. Days for shipment (scheduled), by contrast, is known at the time the order is placed and was retained as a legitimate predictor.

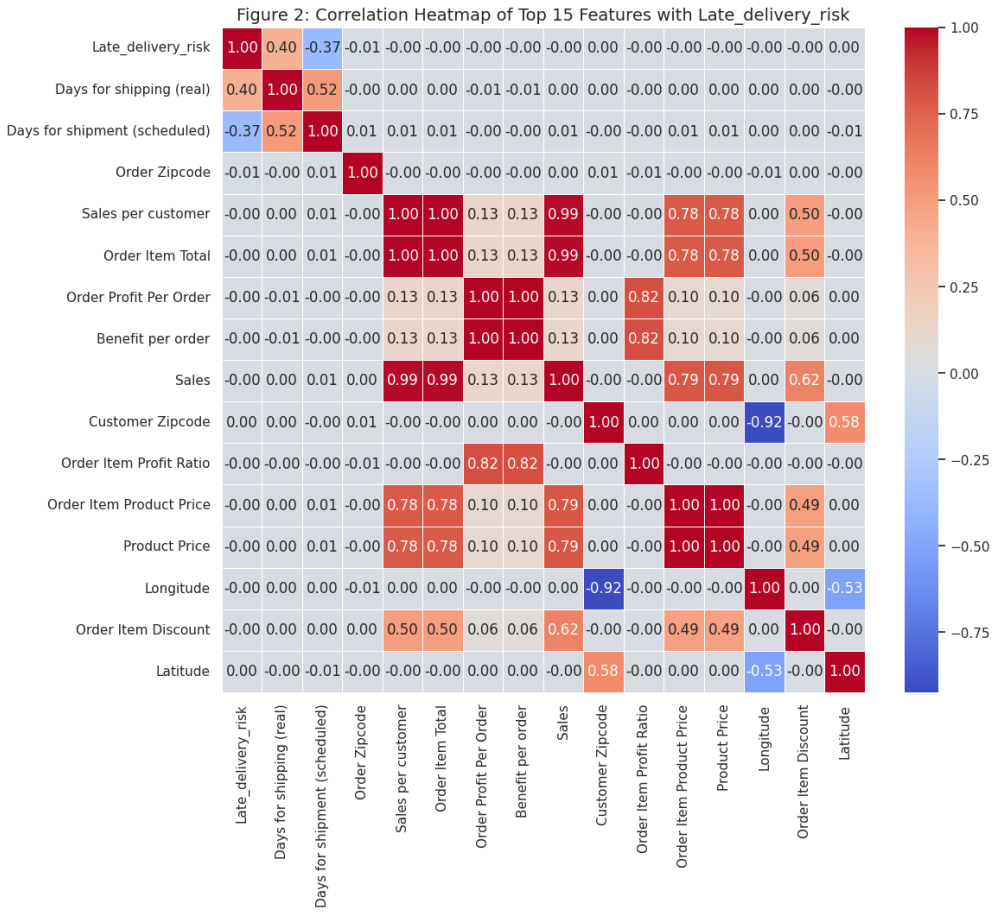


Figure 5. Correlation heatmap of the top 15 features most associated with Late_delivery_risk.

Geographic and Temporal Patterns

Late-delivery risk was further broken down by country and year. The highest average late-delivery rates were observed for orders shipped to the United States and Puerto Rico (both approximately 54–57%), broadly in line with the dataset's overall late-delivery rate rather than showing an extreme geographic outlier.

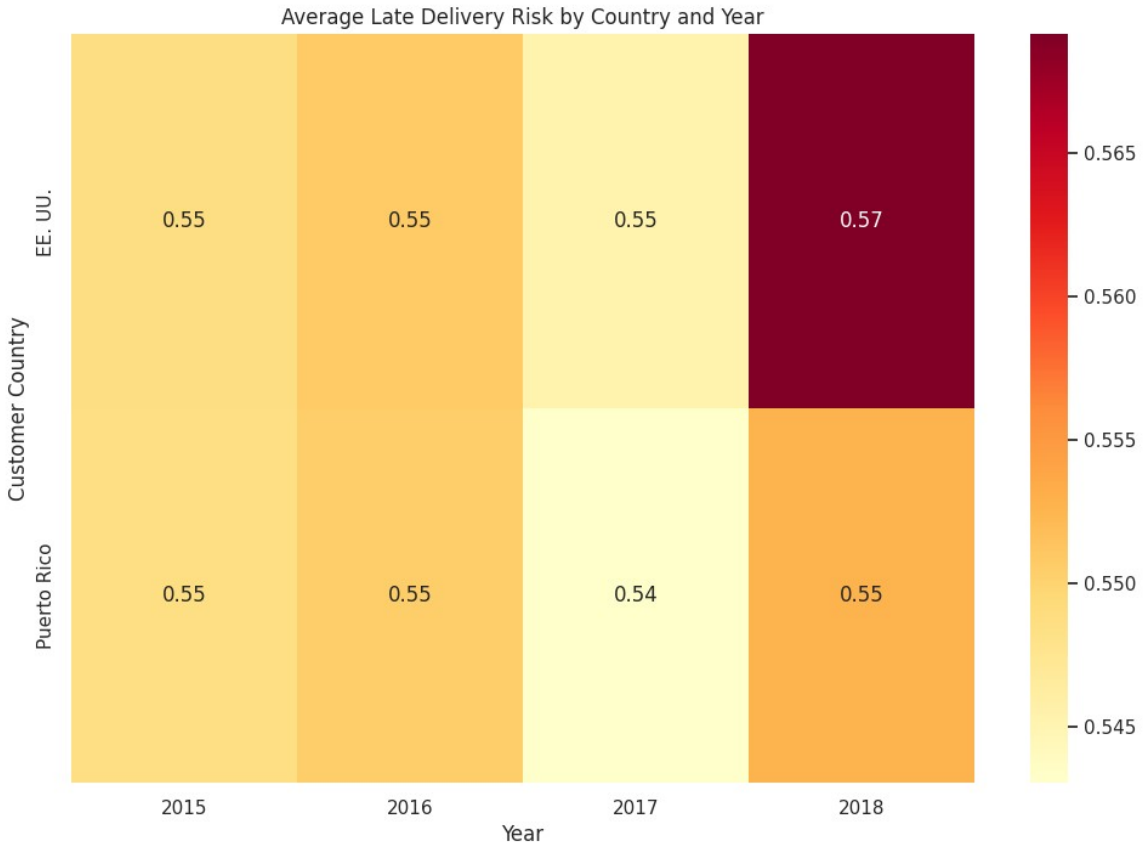


Figure 6. Average late-delivery risk by customer country and year.

A daily time-series view of total Sales, total Profit, and the daily Late_delivery_risk rate was also produced to check whether the data behaved consistently across the multi-year window it covers. Sales and profit held a stable daily pattern from January 2015 through roughly September 2017, then showed sharp, erratic spikes and drops after that point — a signal that the data collection became unreliable toward the end of the observed period. The daily late-delivery rate itself stayed noisy but centered around 50–60% throughout, with no strong long-run trend.

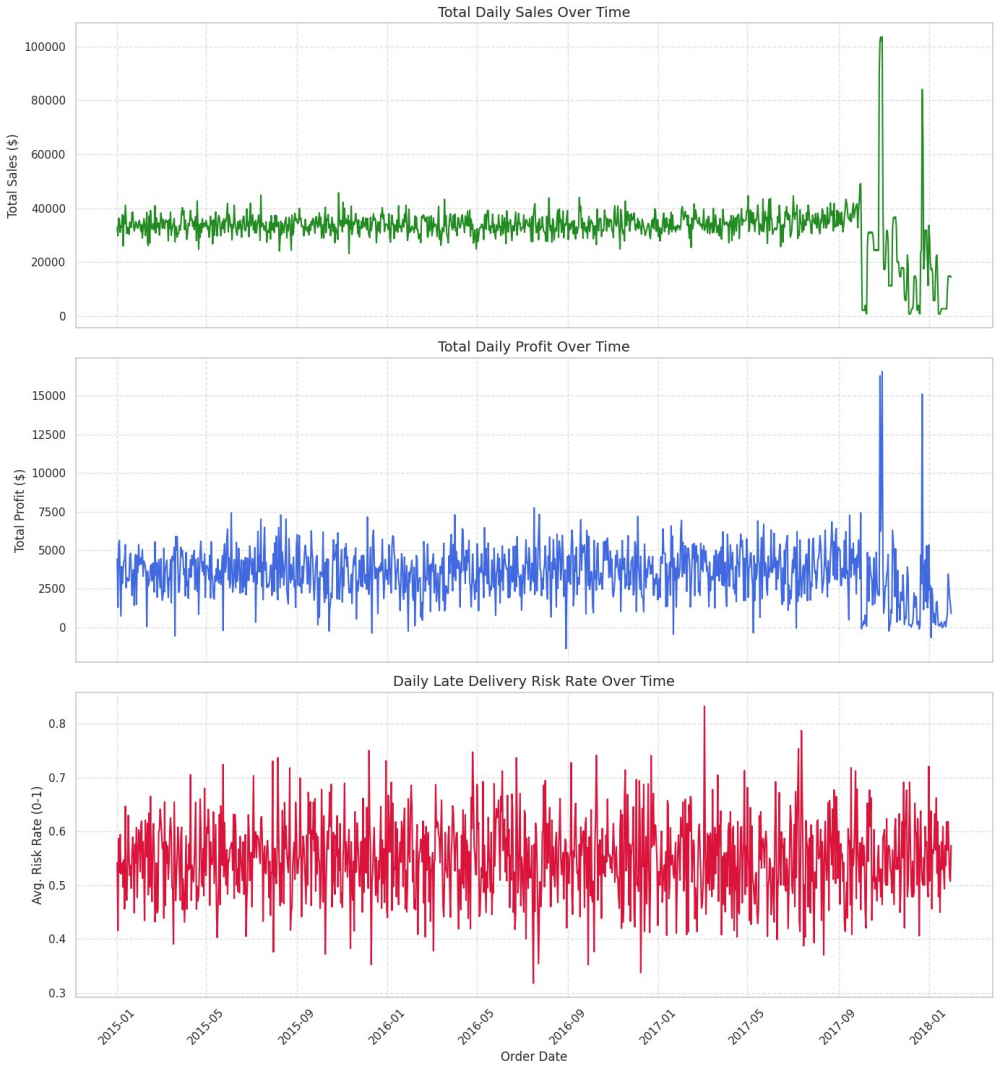


Figure 7. Daily total sales, total profit, and late-delivery risk rate over the full date range of the dataset.

Implications for Week 4

The Operations layer enters data preparation with three concrete tasks flagged from EDA: (1) exclude Days for shipping (real) and Delivery Status as target-leakage columns, since both are only known after a delivery outcome is determined; (2) resolve the identified outliers in Benefit per order and related financial fields via IQR clipping fitted on training data only; and (3) exclude the unreliable tail of the dataset identified in the time-series view — later implemented

in Week 4 as the removal of records after 2017-10-01, which showed the sudden spikes visible in Figure 7.

References

Constante, F., Silva, F., & Pereira, A. (2019). DataCo smart supply chain for big data analysis [Data set]. Mendeley Data. https://doi.org/10.17632/8gx2fvg2k6.5

Taiwanese Bankruptcy Prediction [Data set]. (2020). UCI Machine Learning Repository. https://doi.org/10.24432/C5004D