Part 1 Summary

1. Find All Variables and Understand Them

All 2,121 transactions and 21 attributes from X store sales file were imported with pandas. Data types span integers, floats and objects; key numeric measures (Sales, Quantity, Discount, Profit) show means of 350, 3.8, 0.17 and 8.7, with standard deviations confirming strong variation in Sales (≈ 503) and Profit (≈ 136). There is no missing data in dataframe.

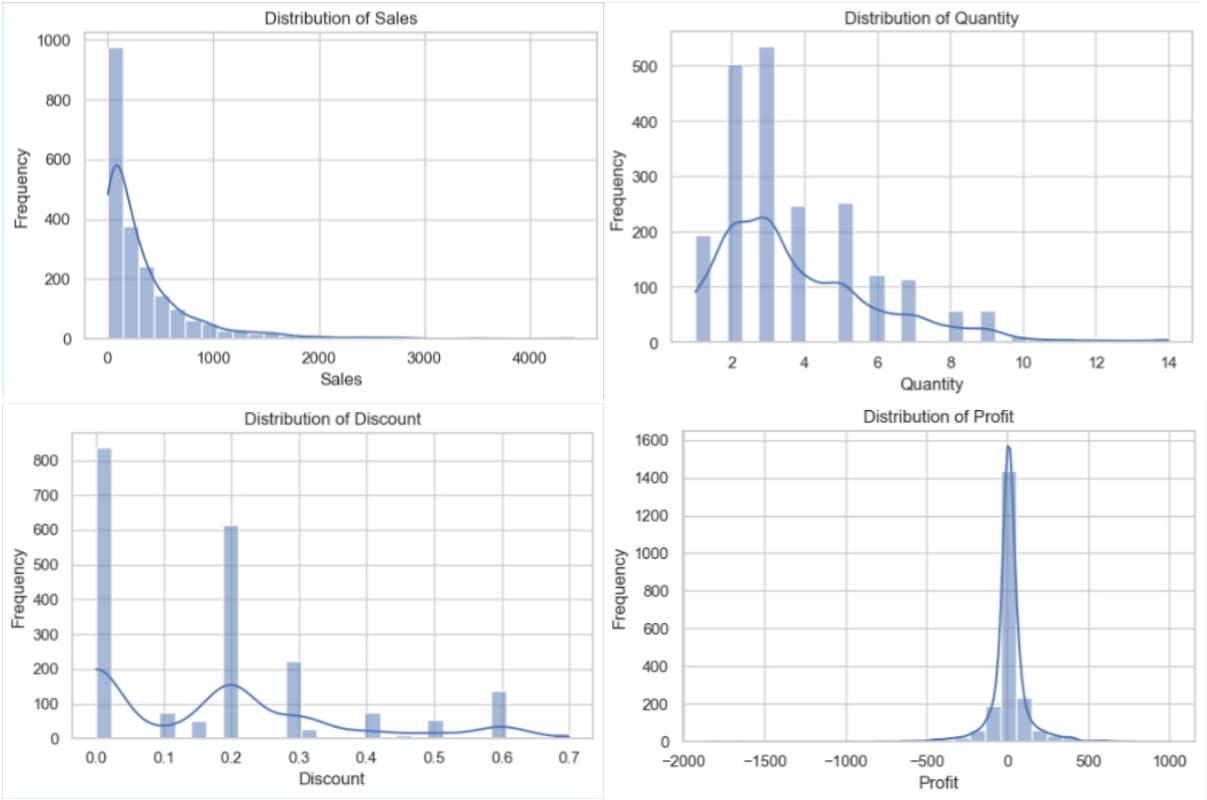

Figure 1: Distributions of Key Numeric Variables (Sales, Quantity, Discount, and Profit)

As show in Figure 1, histograms reveal Sales, Quantity, and Discount are right-skewed, but a few very high ones pull the tail. Profit shows an approximately normal distribution, centred around zero, with most values clustered near the mean and fewer extreme values on either side.

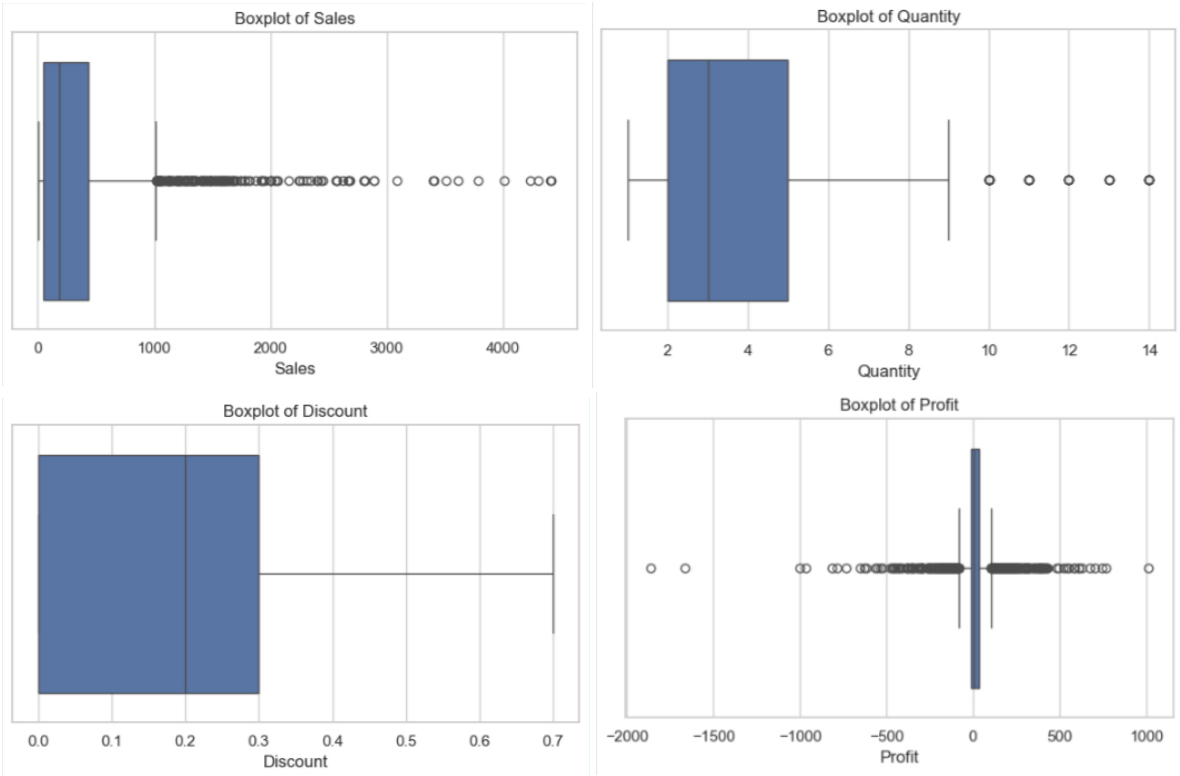

Figure 2: Boxplots of Sales, Quantity, Discount, and Profit

In addition, box plots (Figure 2) echo this tale, highlighting several distant outliers in both tails.

2. Visualise Data

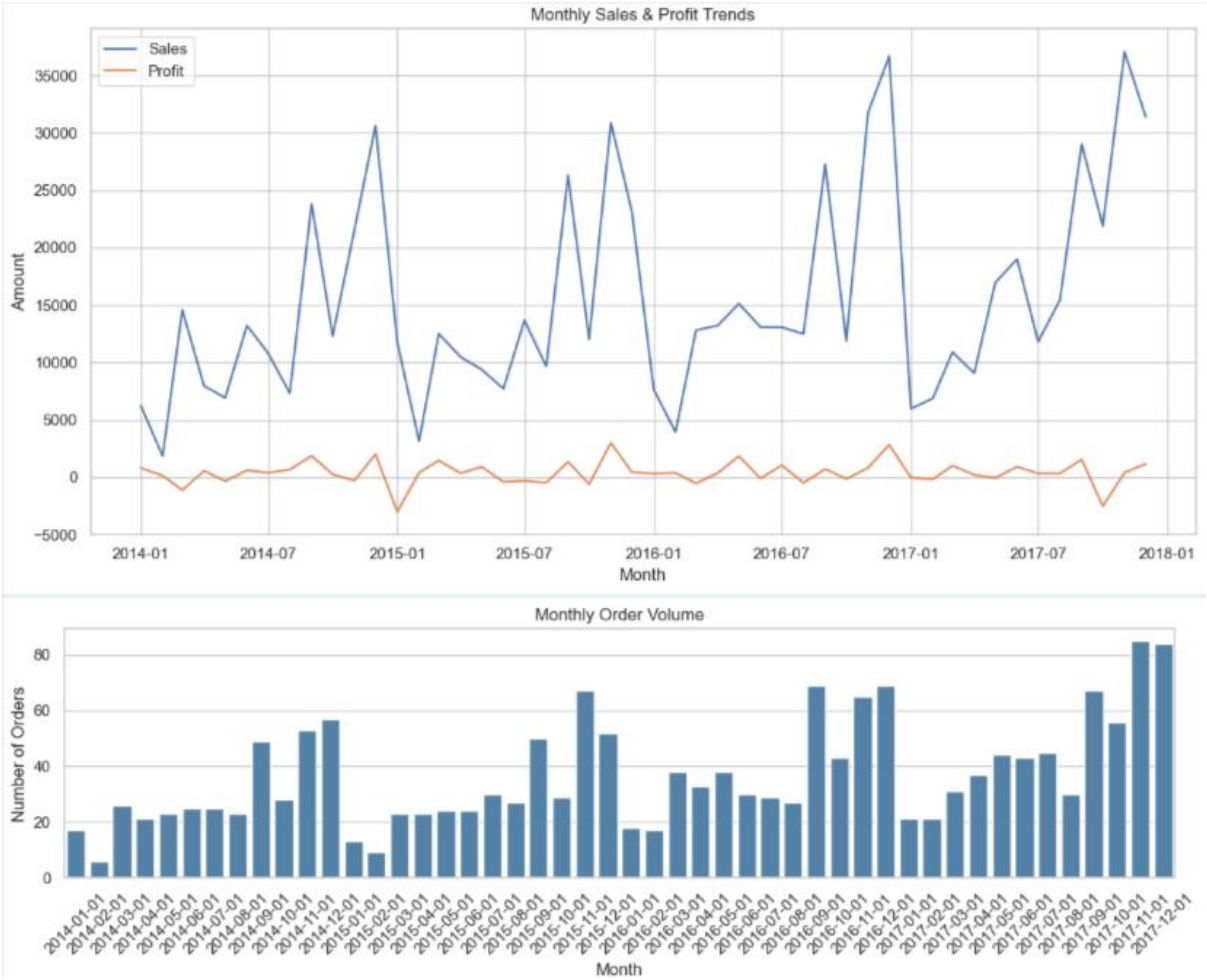

Figure 3: Monthly Trends of Sales, Profit, and Order Volume (2014–2017)

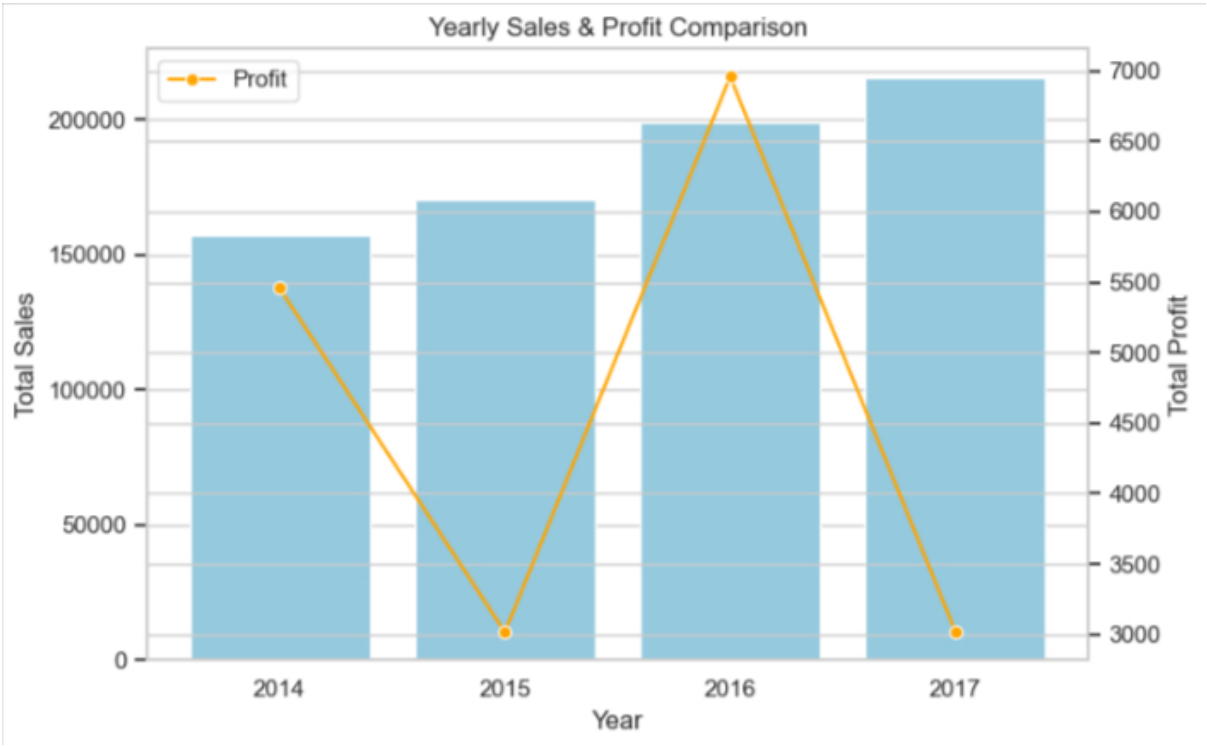

Figure 4: Yearly Sales and Profit Comparison (2014–2017)

A monthly line chart (Figure3) exposes a clear twelve-month rhythm: revenue and order volume climb from September to December, then retreat each January. Overlaying the annual curves in Figure 4 reveals significant fluctuations in profits—steady growth in 2014, a clear decline in 2015, recovery in 2016, and sharp growth in 2017—suggesting that cost structure or pricing policy might be changing or unknown reasons.

3. Clean Data

I confirmed that the table contained no missing entries, so no imputation was needed. Extreme points in Sales, Quantity, Discount and Profit were capped at the 1.5 × IQR limits to stop rare deals impacting the scale. All numeric fields were then standardised with a z-score, ensuring each feature contributes fairly in later modelling. Segment, Ship Mode and Region—each with few categories—were label-encoded, while high-cardinality columns such as City and Region were one-hot encoded to retain detail without ordinals. Finally, mutual-information filtering removed weak predictors.

4. Identify Correlated Variables

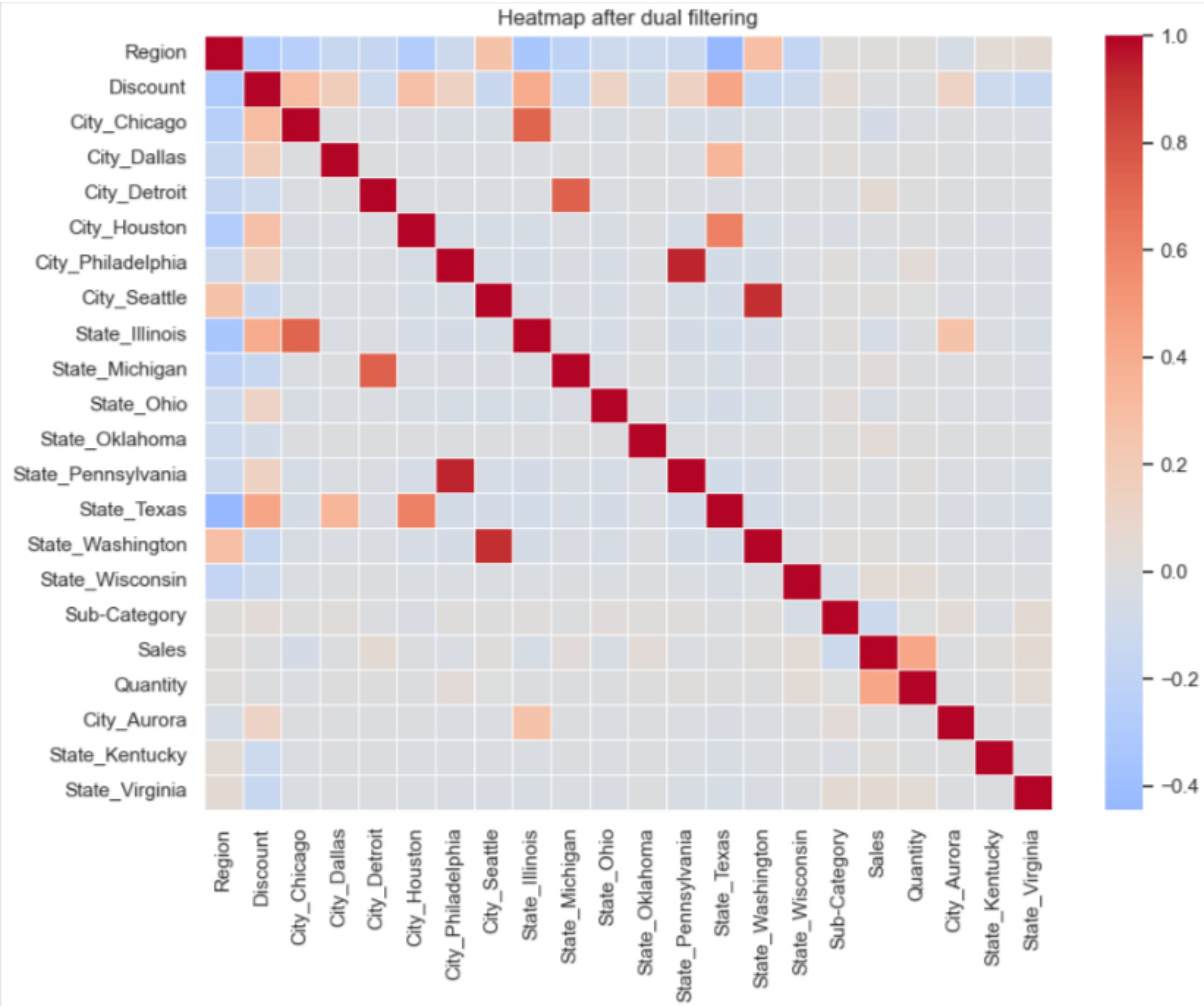

Figure 5: Heatmap of Correlation Between Final 22 Features.

For each correlated group, the variable with the highest correlation to Profit was retained to preserve predictive value. I built a full Pearson matrix on the 435 scaled predictors, masked the upper triangle and kept pairs with $0.10 \le |r| \le 0.85$. That step removed isolated and over-collinear fields, leaving 324 variables. A second screen compared each survivor with Profit; only those whose target correlation also lay in the same band were retained, shrinking the set to 22 predictors. The heat-map in Figure 5 shows modest positive chains between matched City_ and State_ variables, and one strong negative link between Discount and Profit ($r \approx -0.63$). No extreme multicollinearity remains.

Conclusion

The workflow loaded 2,121 sales lines, explored distributions, capped IQR outliers and z-scaled numerics. Low-cardinality columns were label-encoded; high-cardinality categories were expanded into separate binary columns, creating 435 features. Dual correlation filtering first removed weak or redundant relationships between variables, then discarded those with little link to Profit. The final 22-feature table is free from missing data, balanced in scale, and mostly independent, offering a compact and meaningful base for future regression or forecasting work.