Task One Summary

Le Yu

University of Canterbury

Note: Section numbers refer to the submitted notebook "Part 1-submission".

This summary outlines key steps and findings from the data cleaning, exploration, and feature engineering process to prepare the W Store Sales dataset for sales forecasting.

Dataset Overview

This analysis uses the "W Store Sales Dataset" covering January 6, 2014, to December 30, 2017. It includes 2,121 records and 21 fields, with details on orders, customers, products, and sales performance across U.S. cities. *(Section 1.1)*

Data Preprocessing

1. Column types were adjusted as needed—for example, identifiers like "Row ID" and "Postal Code" were treated as categorical, and "Order Date" was converted to datetime. *(Section 1.2)*
2. Basic statistics were calculated for numerical fields. *(Section 1.2)*

3. No missing values were found in any column. *(Section 2.1)*

4. Date completeness check revealed that 566 of 1,455 daily dates were missing (38.9%), **but weekly data is complete — supporting weekly aggregation for modeling**. (Section 2.1)

5. Outliers in numeric fields were identified using the IQR method. No logical issues found in date fields; rare categories posed no concern. *(Section 2.2)*

6. Identified and removed non-significant features including identifiers, single-valued columns (e.g., Country, Category), and high-cardinality fields (e.g., City, State) to avoid dimensionality issues in encoding. *(Section 2.3)*

7. Four categorical variables—Segment, Region, Sub-Category, and Ship Mode—were retained and one-hot encoded. *(Section 2.4).* There are clear differences in sales and profit distributions *(Section 3.2)*

EDA Highlights

Points 2–4 focus on initial feature exploration using daily-level data. From point 4 onward, analysis shifts to the weekly level to identify valuable modeling features.

1.Yearly comparison of sales shows clear seasonality, with total sales increasing steadily year over year. *(Section 1.3)*

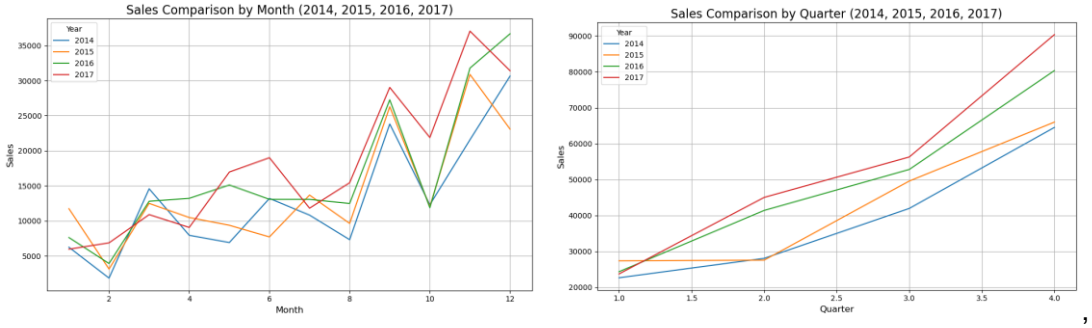

Figure 1

2.Daily Sales is right-skewed (Figure 2) with spikes during holidays (Figure 3) — supporting the use of a holiday flag. *(Section 1.3 & Section 2.2)*

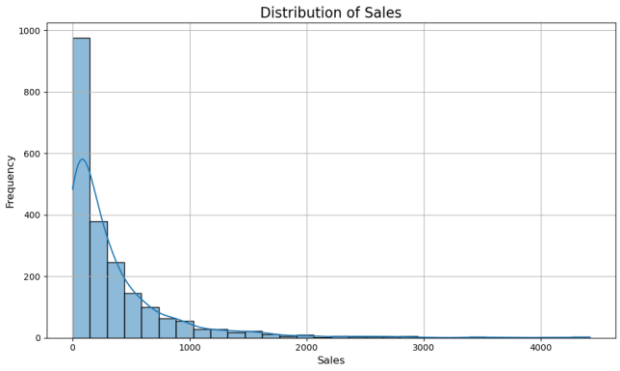

Figure 2

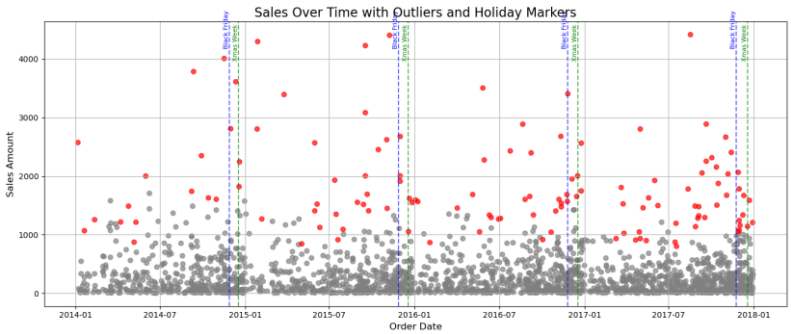

Figure 3

(The red dots indicate statistical outliers in correlation strength)

3. Daily Profit (Figure 4) is roughly normal with some negative and extreme values. *(Section 1.3)*

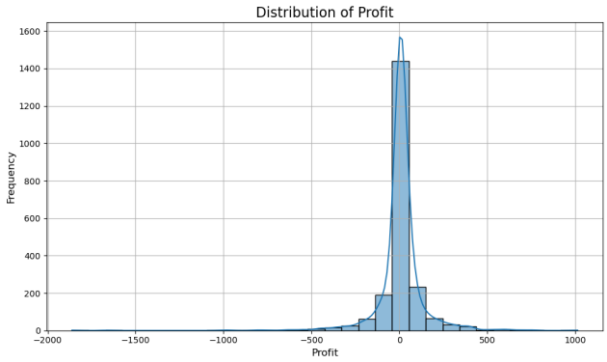

Figure 4

4. Figure 5 below shows clear class imbalance across all four categorical variables, which may impact model training and should be accounted for. *(Section 1.3)*

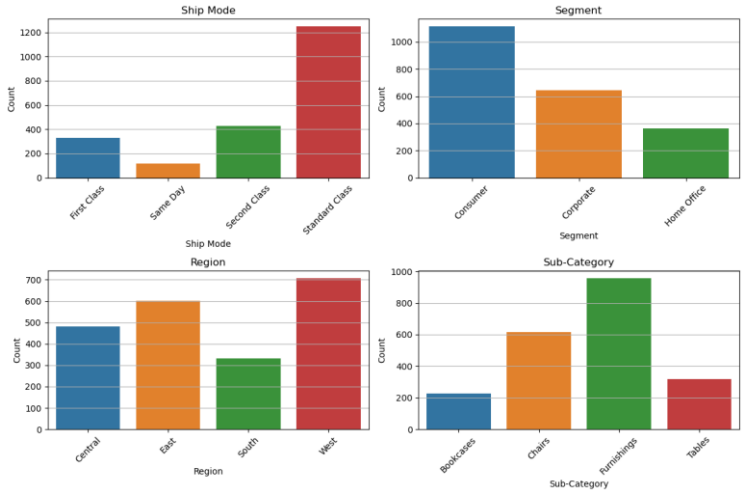

Figure 5

5. As Figure 6 shows, discounts consistently reduce profit, they don't always boost sales(Section 3.2). High negative correlations imply that Discount interactions with category variables may be useful features for profit prediction.

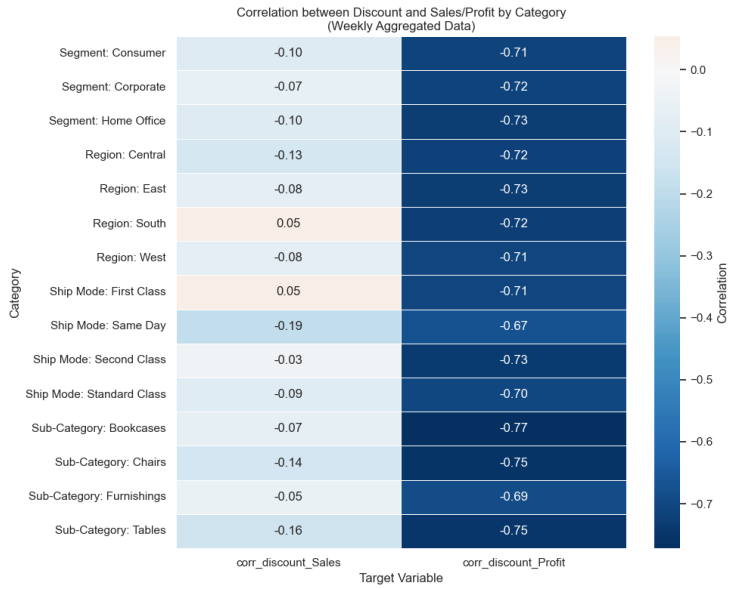

Figure 6

6. Many category indicators show strong positive correlations with weekly sales, indicating high predictive value. *(Section 4.1)*

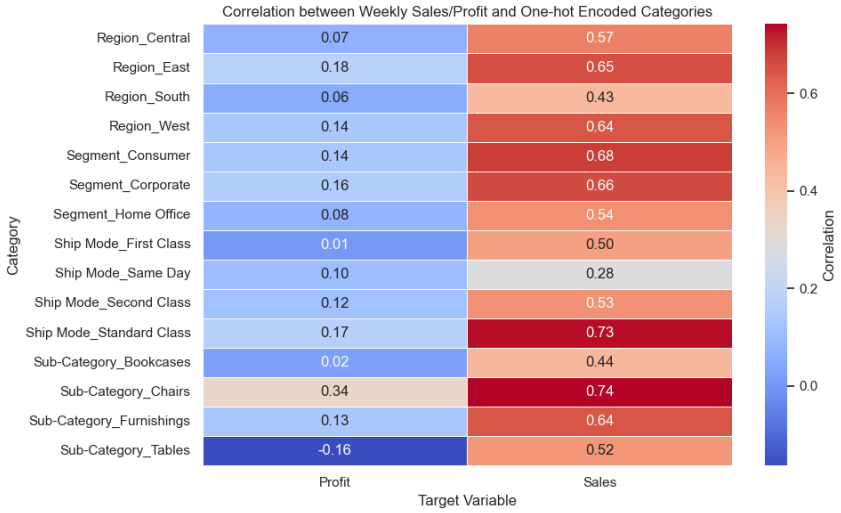

Figure 7

7. Figure 8 shows while discounts reduce weekly profit, they don't always boost weekly sales *(Section 4.1)*

Figure 8

8. STL (Figure 9) shows upward trend and holiday seasonality of weekly sales, explaining 76% of variance — supporting time related features. *(Section 1.3)*

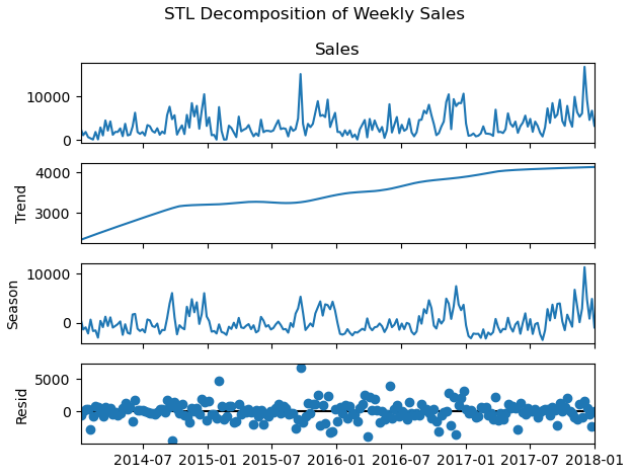

Figure 9.

ACF/PACF (Figure 10) reveal short-term autocorrelation, supporting lag and rolling features. *(Section 1.3)*

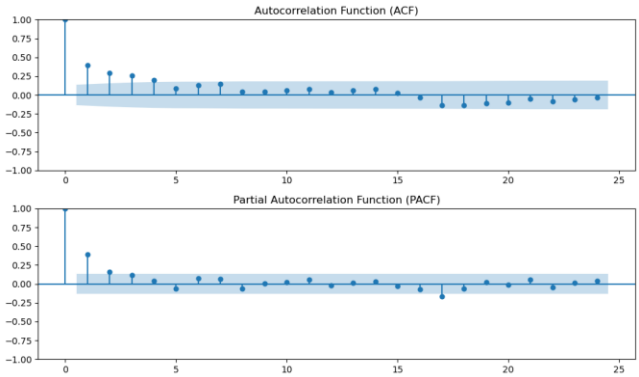

Figure 10. ACF/PACF

Feature Engineering

Created holiday flags for the weeks preceding Christmas and Black Friday. Time-based features such as week, month, and quarter were added. Lag and rolling features (Figure 11) showed strong correlations with weekly sales (>0.7), confirming their relevance. *(Section 4.3)*

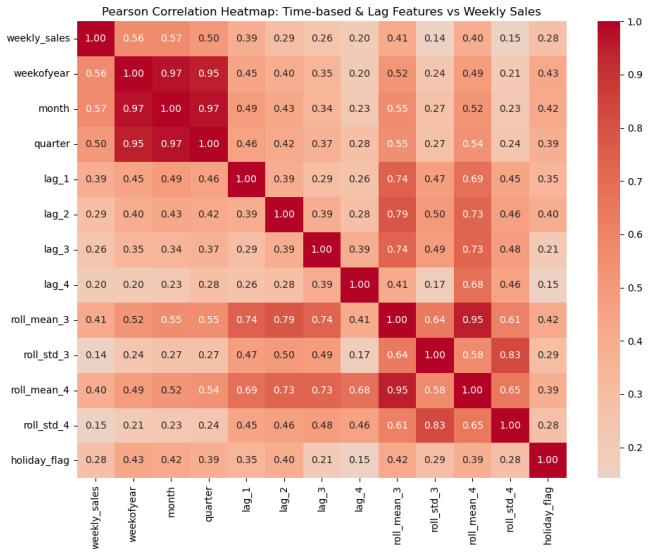

Figure 11

Conclusion

The dataset is clean and exhibits strong seasonality, making it suitable for weekly-level forecasting. Time features, holiday flags, and selected numerical and categorical variables should be retained.

(Section 5 lists all key variables retained or created during preprocessing and feature engineering.)