# Data Cleaning and Analysis Report, Store Sales (2013–2017)

## 1) Introduction

Milestone 2 delivers a clean, analysis-ready dataset and a clear audit trail of the cleaning process. Starting from the raw daily sales and auxiliary tables (oil prices, holidays, stores, transactions), we unify sources, resolve missingness, remove duplicates, and address heavy skew in key variables so that modeling in the next milestone rests on reliable inputs.

## 2) Data Sources

We use the Store Sales dataset (2013–2017) comprising:

- Daily sales per store and product family, with promotions (onpromotion).
- Oil price series (dcoilwtico).
- Holiday/event calendar.
- Store metadata and daily transactions (used only for exploration; dropped for consistency).

Time coverage in the consolidated training data spans from 01-01-2013 to 15-08-2017.

## 3) Data Cleaning Process

The cleaning began with the consolidation of sales, oil prices, holidays, and store metadata into a unified dataset. The initial merged table contained just over 3.05 million rows and 13 columns, but further checks revealed 30,294 duplicate rows caused by overlaps during joining. These were removed, resulting in a clean base of 3.02 million records. After restructuring, the final training dataset reached a size of 3,000,888 rows and 12 columns, and a second round of checks confirmed that no additional duplicates remained.

Missing values presented the largest challenge. The oil price series contained extensive gaps, with more than 850,000 null entries in the training data and a smaller number in the test set. To address this, missing values were imputed using a forward fill method, with backfill applied to the earliest missing stretches to ensure continuity from the beginning of the series.

The day_type field also contained over 2.5 million null entries, which were filled with the most representative category, “Work Day”, to maintain consistency. After these steps, both the oil price and day type variables were complete with no missing values in either training or test data.

To align schema between datasets, the transactions column was dropped from both train and test, as it was not consistently available and did not serve the modeling objective. This adjustment reduced the training dataset to 11 columns, with a standardized structure across all files. Data types were also verified, ensuring that dates, numeric measures, and categorical fields were all correctly formatted.

Finally, distribution checks revealed that both sales and onpromotion were highly skewed, with long right tails that distorted summary statistics and visualizations. Rather than removing these outliers, as many of which represented real holiday-driven spikes, we decided to apply a log(1+x). This step preserved the natural variation while reducing skewness, producing smoother histograms and more stable patterns for analysis. The dataset retained 33 product families and a complete set of store and day type categories after these refinements.

# 4) Challenges Encountered

Several issues were encountered during cleaning. Extended gaps in the oil price series required a combination of forward and backward filling to avoid biasing the early portion of the data. Day type classifications also posed difficulties, as overlaps between holidays and bridge days created inconsistencies. A standardized hierarchy was applied, and missing values were set to “Work Day” to avoid fragmented categories. The heavy skew in sales and promotions complicated early descriptive analysis, but the log transformation balanced this issue while retaining meaningful spikes.

# 5) Insights from the Cleaned Dataset

The cleaned dataset now offers reliable coverage across all key fields. Seasonal patterns remain visible and consistent, and the treatment of missing values allows oil price and day type to be used in feature engineering without interruption. Store types and product families are harmonized, confirming that there are 33 consistent categories across the dataset. The transformation of sales and promotions has created distributions that are easier to interpret and model, while still capturing the holiday surges that drive our dataset demand. With the removal of the transactions column, the training and test sets now share a unified structure, avoiding mismatches in later modeling stages.

# 6) Conclusion

Through systematic cleaning, the dataset has been brought to a state of consistency and completeness. Duplicates were eliminated, missing values in critical fields were resolved, categorical inconsistencies were corrected, and skewed variables were transformed for stability. The final training dataset contains 3,000,888 rows and 11 columns, fully aligned with the test set and ready for use in forecasting models. These steps ensure that the foundation for Milestone 3 is

both transparent and dependable, enabling robust feature engineering and accurate predictive modeling.