# **Data Exploration Report — Store Sales (2013–2017)**

## **Overview**

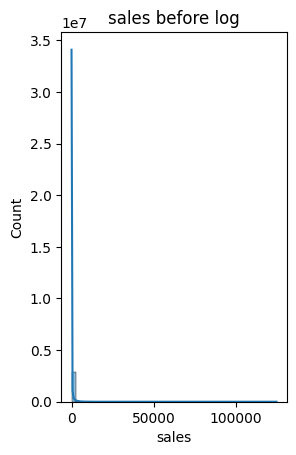
This exploration analyzes daily store-family sales from 1 January 2013 through 15 August 2017, joined with store descriptors (city, state, store\_type, cluster), holiday/event information as a categorical day\_type, and WTI oil prices (dcoilwtico). After merging and column clarification, the working training table contains 3,000,888 rows and 11 columns: date, store\_nbr, family, sales, onpromotion, city, state, store\_type, cluster, dcoilwtico, and day\_type, covering 54 stores across 22 cities and 16 states, organized into 17 clusters and 33 product families. Uniqueness checks highlighted 1,684 distinct dates; 33 families; 362 unique onpromotion values; 22 cities; 16 states; 5 store types; 17 clusters; 994 oil price points; and 6 day types. By using the figures attached to our notebook, we were able to highlight the following key points: the day-type charts (“Average Sales per Day Type,” “Max Sales per Day Type”) show Transfer days lead on average while Holidays create the biggest spikes, which is helpful for planning promotions; the store-type view (“Insights from Sales by Store Type”) shows the clear ranking D > C > A > B > E; the multi-year line (“Store Sales Analysis (2013–2017)”) shows steady growth to 2016 and a dip in 2017; the monthly panel (“Monthly Sales Analysis (2013–2017)”) makes December highs, February lows, and the 2017 break (record May, sharp August drop) easy to see; the oil chart (“dcoilwtico Timeline by Year”) gives outside market context; and the transform checks (“sales before/after log,” “onpromotion before/after log”) show the log step makes the numbers more even and reduces extremes.

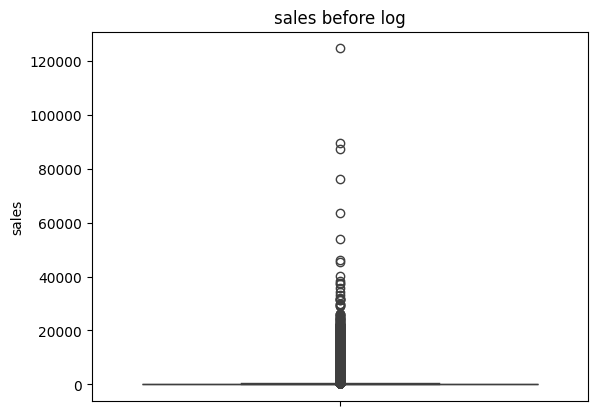
## **Data Quality and Preparation**

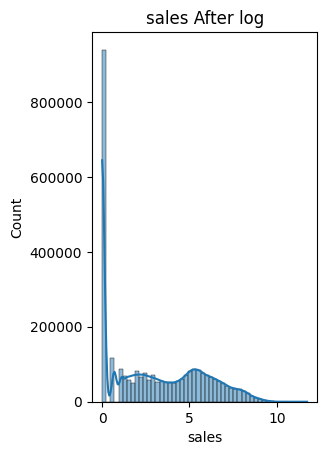
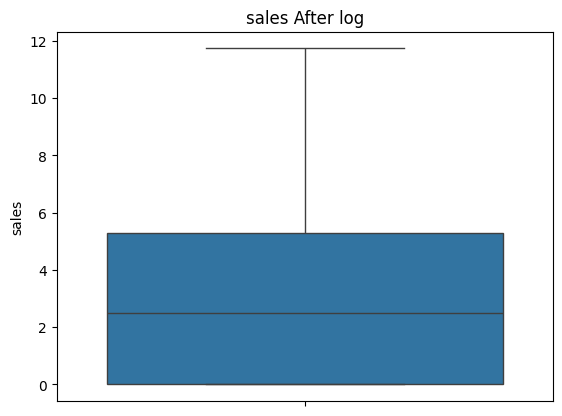
Before cleaning, two fields had a lot of missing data: dcoilwtico had 880,308 empty values (~28.8%), and day\_type had 2,551,824 empty values (~83.5%). We filled oil price gaps by carrying the last known value forward (and backfilled at the very start if needed). The missing day\_type was set to “Work Day.” We also cleaned text labels (trimmed spaces, used Title Case) to avoid duplicate categories. Duplicates were removed in two steps: first 30,294 rows were dropped, then a second pass (ignoring day\_type) produced the final table of 3,000,888 × 11. Data types are set as expected: dates for date; categories for family, city, state, store\_type, day\_type; and numbers for store\_nbr, sales, onpromotion, cluster, dcoilwtico. Sales data is right skewed: min 0.0, Q1 0.0, median 11.0, Q3 195.8473, mean ~357.78, max 124,717.0. Oil prices vary widely (about $26.19–$110.62; median $53.41; mean ~$67.86; SD ~$25.66). Using the IQR rule, we found 447,105 outliers in sales and 611,329 in onpromotion (none in store\_nbr, cluster, or dcoilwtico). To make the numbers easier to model, we applied a log transform to sales and onpromotion and confirmed by using the existing before/after plots, that this reduced skew and shrank extreme values.

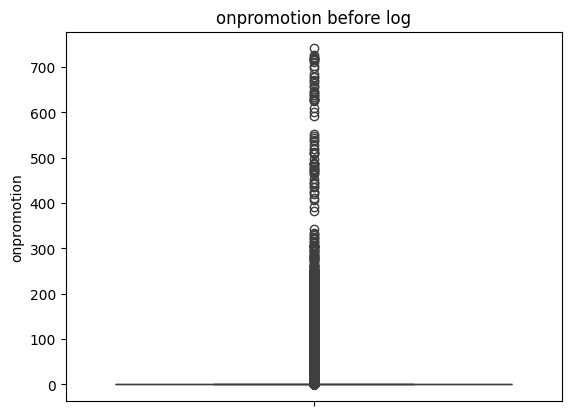
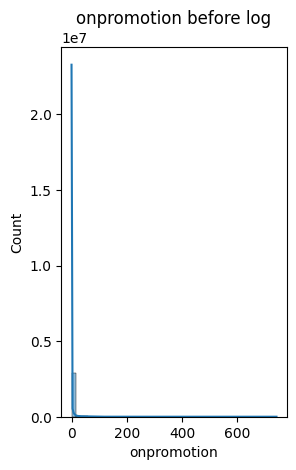
## **Engineered Features and Transformations**

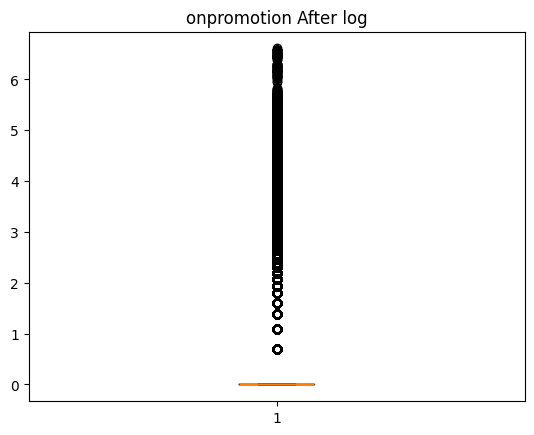
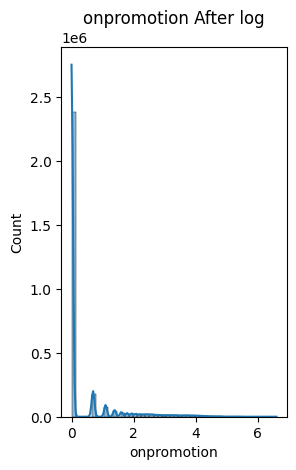
We added calendar fields (Year, Month, Day, Week\_day) and a Work\_day flag from day\_type to support time-based summaries. We applied a log transform to sales and onpromotion to reduce skew and shrink extremes; the change is visible in the existing diagnostics (“sales before/after log,” “onpromotion before/after log”). We also demonstrated one-hot encoding for categoricals and scaling for numeric fields for modeling readiness; in the encoding demo, “Work Day” was dropped as the baseline. No other features were created (no lags, rolling stats, interactions, or extra holiday flags). The log versions of sales and onpromotion are used only for diagnostics/modeling.









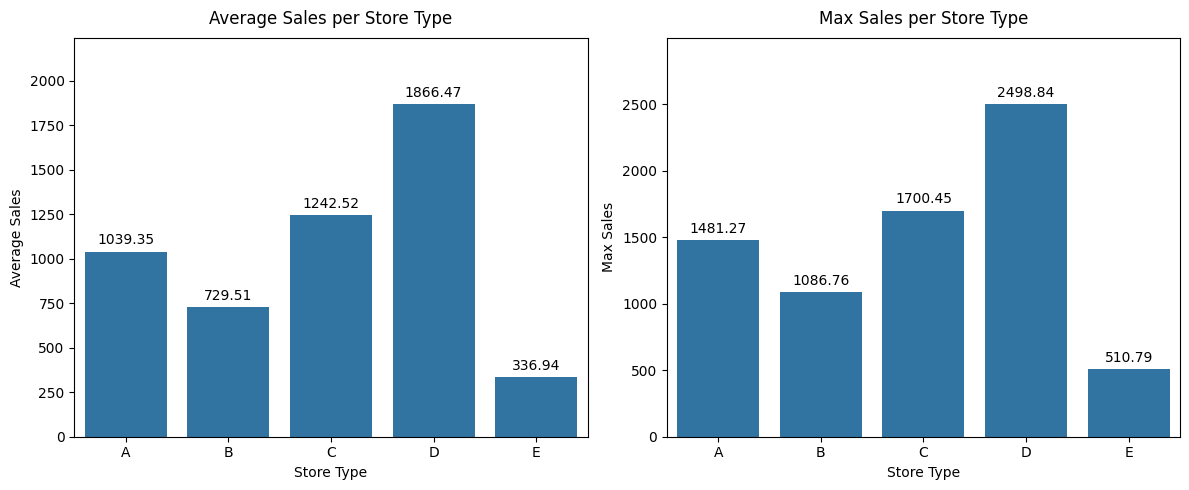


## **Trends by Day Type**

Day type clearly affects sales. In “Average Sales per Day Type,” Transfer days have the highest average (~6,107), then Additional (~5,800), Event (~5,698), Bridge (~5,681), Work Day (~5,192), and Holiday (~5,076). In “Max Sales per Day Type,” Holiday days hit the biggest single-day highs (~7,107), with Additional (~7,055) and Work Day (~7,040) close behind; then Transfer (~6,929), Event (~6,760), and Bridge (~6,271). In short: Transfer wins on averages, Holidays create the biggest spikes, Work Days are steady with upside, and Bridge is the weakest.

## **Store-Type Performance**

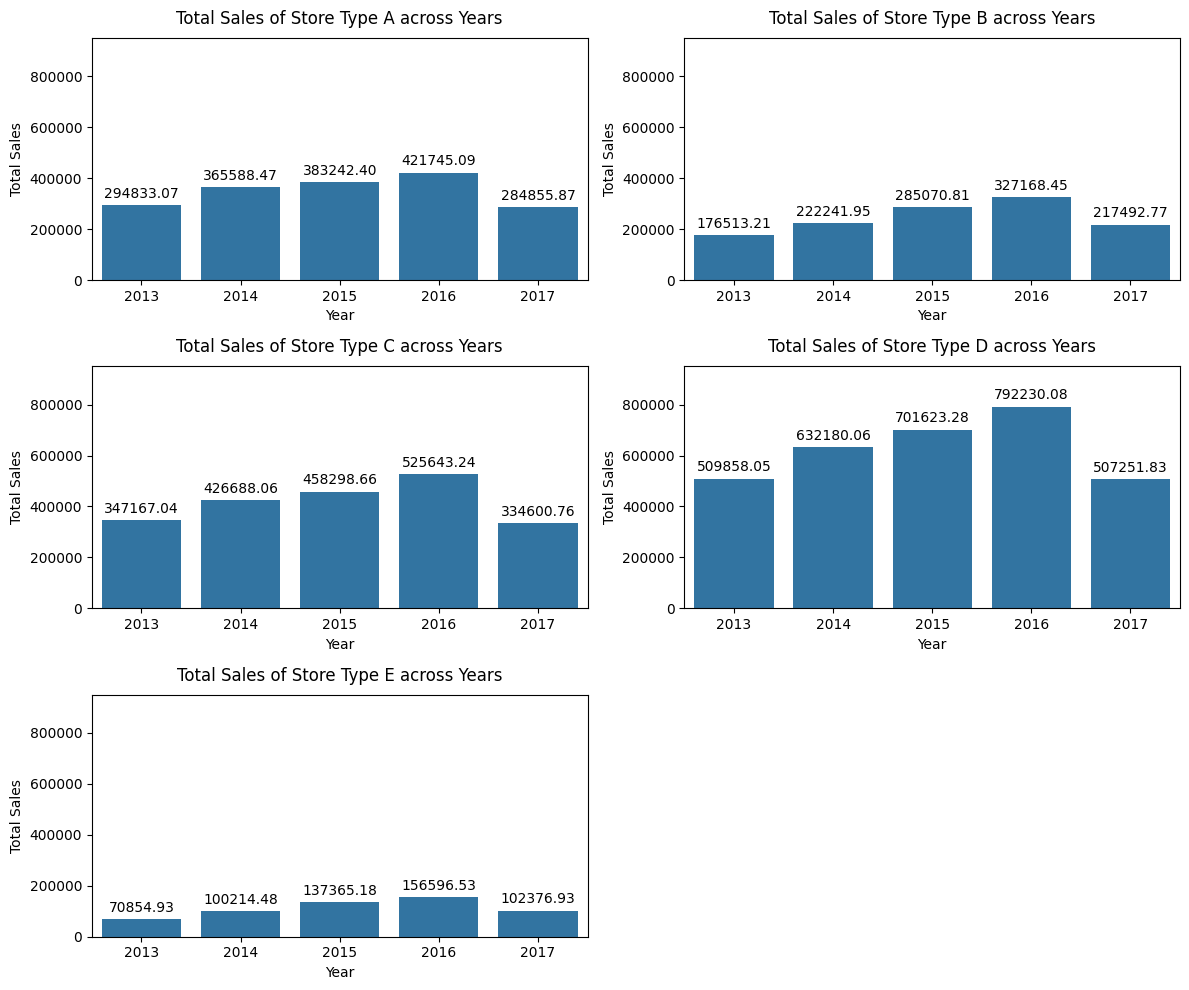
Store types show a clear, repeatable gap in performance. In “Insights from Sales by Store Type,” Type D is the top performer on both metrics with about 1,866 average sales and 2,499 max, followed by Type C (~1,242; ~1,700) and Type A (~1,039; ~1,481). Type B is lower (~729; ~1,087), and Type E is the weakest (~337; ~511). The ranking is identical for averages and peaks (D > C > A > B > E), so this isn’t a one-off spike; it’s a stable pattern in the data. Practically, D and C look like the main volume drivers, A provides a steady base, and B. Especially E, needs review to understand what’s holding results back before expecting big gains. This hierarchy is useful for setting inventory levels, staffing, and promotion intensity by store type.

**

## 

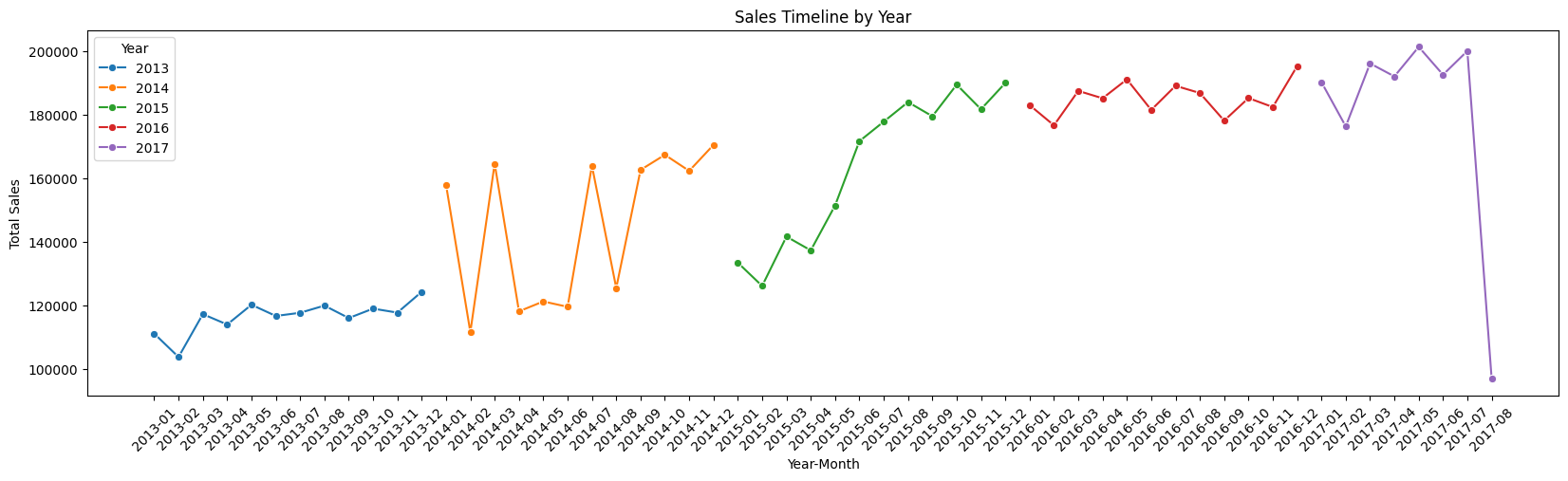
## **Multi-Year Trajectory (2013–2017)**

The year-by-store-type table shows the same pattern across all types: steady growth from 2013 to 2016, then a drop in 2017. For example, Type D totals are 509,858 (2013), 792,230 (2016), and 507,252 (2017). Type C is 347,167; 525,643; and 334,601. Type A is 294,833; 421,745; and 284,856. Type B is 176,513; 327,168; and 217,493. Type E is 70,855; 156,597; and 102,377. A summary table in the notebook also marks 2016 as the peak year for every type (e.g., D ≈ 95.6M, A ≈ 91.9M) and 2013 as the lowest (e.g., D ≈ 45.8M, A ≈ 47.0M). The “Store Sales Analysis (2013–2017)” chart makes this rise through 2016 and the broad pullback in 2017 clear.



## **Seasonality**

Monthly results show a clear, steady pattern. From 2013 to 2016, total sales grow from about 1.40M to 2.22M. December is the top month each year (124,412; 170,628; 190,106; 195,397), and February is the lowest (103,894; 111,748; 126,258; 176,784). This makes the seasonal shape predictable: a strong year-end lift and a softer early-year dip. In 2017, that shape changes. The yearly total is about 1.45M, May hits a record 201,465, and August falls to 97,005. This is different from the earlier years where December led. The notebook also notes that late-2017 months may be missing, so results should be checked against the usual September–December bump once data is confirmed.



## **Conclusion and Key Findings**

The data shows a clear, repeatable pattern up to 2016: sales peak in December and dip in February. In 2017 that pattern breaks as May hits a record high, then sales fall sharply over the summer. By day type, Transfer days have the best averages, Holidays create the biggest single-day spikes, Work Days are steady with room to rise, and Bridge days are the weakest. By store type, D leads across the board, C is strong, A is steady, and B and E lag. Data issues were mainly in the add-on fields. We filled missing oil prices by carrying the last value forward, set missing day\_type to “Work Day,” removed duplicates, and cleaned category labels. Because sales and promotions are very skewed, we used a simple log step to tame extremes. Overall, the EDA confirms solid seasonality before 2017, clear differences by day and store type, and a clean base for forecasting. The break in 2017 and the weak performance of Bridge, B, and especially E gurantees closer review in modeling.