Summary of objectives and data

I examined weekly sales across Walmart stores using the Kaggle Walmart dataset years 2010 to 2012. The analysis focused on three linked objectives :

i. to characterize their distribution and seasonality of weekly sales.
ii. to measure the relationship between weekly sales and economic/operational predictors (temperature, fuel price, CPI, unemployment, holiday flag).
iii. to estimate marginal effects and run diagnostics to check assumptions.

The dataset covers weekly observations and includes Weekly Sales, Temperature, Fuel_Price, CPI, Unemployment, and Holiday_Flag. Data cleaning steps (de-duplication, date conversion, and simple missing-value handling) were applied prior to analysis.

Limitations

- The analysis used aggregated weekly sales. A finer temporal (daily) or SKU-level data may reveal stronger relationships indepth.
- Some variables that likely drive sales (promotions, competitor pricing, inventory stockouts, advertising) were not present. Without them, omitted variable bias is a risk.
- OLS conclusions are conditional on the current variable set; adding high-value predictors will change coefficients and diagnostic behavior.

Results and Discussions

1. **Distribution and time-series patterns (figures: boxplots and time series)**

What I plotted: boxplots of weekly sales by year and a line plot of average weekly sales over time.

Key findings:

- The median weekly sales is stable across years and sits near the same baseline (~$0.9–1.0M). This suggests the retail core demand is consistent year-to-year.
- There are large, recurring spikes around late November–December consistent with holiday shopping (Black Friday/Christmas) and promotional activity. These appear as extreme outliers on the boxplots and sharp peaks on the time series.
- Outside holiday spikes, the series shows weak long-term trend but strong seasonal shocks (regular annual peaks). This pattern implies that forecasting models must explicitly model seasonality or include holiday indicators to capture these predictable spikes.

the business should treat holiday weeks as structurally different regimes — separate forecast models or explicit holiday dummies will reduce forecast errors and prevent misleading coefficient estimates that try to “explain” seasonal spikes with ordinary predictors.

2. Correlation structure (figure: correlation heatmap)

What I measured: Pearson correlations between Weekly Sales and predictors.

Key findings:

- Correlation coefficients between Weekly Sales and individual predictors are **very weak** (roughly ±0.10 or smaller). No predictor showed a strong linear relationship with weekly sales in raw form.
- The strongest correlations are **not** between sales and a single economic predictor, but among predictors themselves (e.g., CPI and unemployment show meaningful negative association in our sample).

Interpretation: Weak pairwise linear correlations imply that sales are influenced by **multiple interacting factors** and non-linearities, or that the predictors in the dataset (temperature, fuel price, CPI, unemployment) are insufficiently granular (no SKU-level promotion flags, no day-of-week effects, no store demographics). Contemporary retail forecasting literature emphasizes the need for high-dimensional features (promotions, loyalty data, intra-category effects) to materially improve forecast accuracy.

3. OLS regression results and diagnostics

Model form used: Weekly_Sales ~ Temperature + Fuel_Price + CPI + Unemployment + Holiday_Flag (constant included). Depending on whether CPI was used to create an inflation-adjusted sales measure, results are similar qualitatively.

Summary of numeric results (high level):

- Coefficient magnitudes are generally small relative to baseline sales and several are not statistically significant at conventional levels.
- Holiday_Flag is strongly positive and statistically significant — as expected, the holidays explain a large portion of peak sales. Other predictors show limited explanatory power in the linear specification.

Diagnostics (figures: residuals vs fitted, Q-Q plot, scale-location, residual histogram):

- **Heteroscedasticity:** Residuals vs fitted shows non-constant variance (variance increases with fitted sales). This violates the OLS assumption of homoscedastic errors and suggests standard errors may be biased — robust (White) standard errors or heteroscedasticity-consistent inference should be used for hypothesis testing. (See standard diagnostics literature and tests for heteroscedasticity.) stat.cmu.edu
- **Non-normal residuals:** Q–Q and histogram reveal skewness and heavy tails. Large positive residuals correspond to holiday spikes; such heavy tails reduce the reliability of p-values.
- **Nonlinearity / omitted variables:** Structure in residuals indicates OLS did not capture key dynamic patterns — seasonality, promotions, or interaction terms are likely important.

Practical consequence: OLS is useful for transparent, interpretable marginal effects (especially for policy-style interpretation), but it is not an ideal forecasting engine in this retail

context. Forecast accuracy and correctness of inference will improve after addressing heteroscedasticity and nonlinearity (e.g., transform the target, include seasonal dummies, or use non-linear models).

Integrated interpretation (putting plots, correlation and regression together)

The **boxplots & time series** show pronounced seasonality and outliers driven by holidays; the **heatmap** shows that common macro predictors in the file are weak linear drivers; the **OLS diagnostics** show OLS is interpretable but mis-specified for forecasting.

Together, the evidence suggests: retail sales are driven largely by **idiosyncratic and contextual factors** (promotions, product mix, store events, day-of-week, SKU effects), which are not captured by the small set of macro predictors in this dataset. Similar conclusions are reached in applied retail forecasting literature: adding SKU-level promotions, intra-category features, and richer event indicators materially improves performance.