From the weekly, monthly, quarterly, and yearly sales trends, we observe consistent seasonal spikes around November and December, likely due to holiday promotions such as Thanksgiving and Christmas. This is especially evident in the monthly and weekly plots, where sharp peaks appear annually at year-end.

The monthly seasonal plot further supports this, showing that December consistently records the highest average weekly sales across all three years (2010–2012), followed by noticeable increases in November. In contrast, January and September often have the lowest sales, indicating post-holiday slowdowns or seasonal demand dips.

The boxplot comparing holiday vs. non-holiday weeks shows a slightly higher median and greater spread in weekly sales during holidays, though there is overlap. This suggests that while holidays tend to boost sales, some outlier spikes also occur during non-holiday periods—potentially tied to markdown events or store openings.

Overall, the data exhibits strong seasonality, with year-end months being critical for high sales. Incorporating time-based features (like month or holiday indicators) is likely to improve forecasting performance.

To prepare the datasets for modeling, I began by merging three sources (features.csv, sales.csv, and stores.csv) using Store and Date as join keys. I performed key statistical analysis using describe() and visualized numerical data distributions with histograms and boxplots, helping to understand skewness and outliers. Date variables were decomposed into Year, Month, and Day for easier feature extraction.

For categorical handling, the only boolean feature IsHoliday was converted to numerical format (0/1). I visualized time-series patterns of sales across weekly, monthly, and yearly levels to uncover trends and seasonality.

Missing values were primarily concentrated in markdown columns, which were expected due to non-promotional periods. I replaced them with 0 to preserve business logic. Outliers were detected using IQR and STL decomposition, then cross-validated by checking markdown activity. Only genuine anomalies (with all markdowns = 0) were removed to retain legitimate spikes.

For feature selection, I applied backward elimination, which iteratively removes features with the highest p-values until only statistically significant variables remain. This method improves model simplicity and avoids including irrelevant predictors. The final selected features were CPI, MarkDown3, Year, and Day, all showing strong relationships with weekly sales.

Finally, correlation analysis revealed strong multicollinearity between CPI & Fuel_Price, and MarkDown1 & MarkDown4. Fortunately, backward selection excluded one from each correlated pair, suggesting that it effectively reduced redundancy and preserved only independently predictive variables.

This thorough process ensured the dataset was clean, structured, and well-prepared for robust modeling.