Walmart Weekly Sales Prediction: Comprehensive Analysis Report

1. Executive Summary & Problem Understanding

Dataset Description:

The dataset contains historical weekly sales data for 45 Walmart stores, including variables such as Holiday Flag, Temperature, Fuel Price, Consumer Price Index (CPI), and Unemployment Rate.

Problem Statement:

Retail stores face challenges in inventory management and staffing due to fluctuating weekly sales. Accurately predicting future sales based on seasonal and economic factors is critical to minimizing overstock and stockouts.

Objective of the Analysis:

To perform Exploratory Data Analysis (EDA) to understand sales drivers, engineer relevant features, and build robust Machine Learning (ML) and Artificial Neural Network (ANN) regression models to accurately forecast Weekly Sales.

Summary of Key Findings:

- Store ID and seasonality (specific weeks of the year) are the strongest predictors of sales.
- Holidays significantly spike sales, particularly in late November and December.
- Macroeconomic factors (CPI, Unemployment, Fuel Price) have a surprisingly weak direct linear correlation with weekly sales.
- Non-linear models drastically outperformed linear models.

Final Model Recommendation:

The Random Forest Regressor (or optimized ANN) is recommended as the final model due to its high R^2 score and ability to capture the complex, non-linear relationships between individual stores, time of year, and sales volume.

2. Data Cleaning & Preprocessing

Dataset Structure:

- **Shape:** 6,435 rows and 8 columns.
- **Data Types:** Store (int), Date (object/string), Weekly_Sales (float), Holiday_Flag (int), Temperature (float), Fuel_Price (float), CPI (float), Unemployment (float).

Missing Value Analysis & Treatment:

- **Analysis:** Checked via df.isnull().sum(). The base Walmart dataset contains 0% missing values across all columns.
- **Treatment:** No imputation was necessary.

Duplicate Handling:

- Checked for duplicate rows. No duplicates were found.

Outlier Detection & Treatment:

- Weekly_Sales contained natural outliers (massive spikes during holiday weeks). These were retained as they represent genuine, critical business events rather than data errors.
- *Justification:* Removing holiday sales spikes would destroy the model's ability to predict peak seasonal demand.

Encoding & Feature Scaling:

- **Encoding:** No categorical text variables required encoding (Holiday_Flag is already binary 0/1).
- **Scaling:** Applied StandardScaler to continuous numerical features (Temperature, Fuel_Price, CPI, Unemployment, and engineered date features) to ensure distance-based models and the ANN converge efficiently.

3. Exploratory Data Analysis (EDA)

◇ **Univariate Analysis**

- *Distribution of Weekly Sales:* The histogram reveals a right-skewed distribution, indicating that while most weekly sales cluster around $1M, certain weeks (holidays) push sales closer to $3M+.

◇ **Bivariate Analysis**

- *Sales vs. Holiday_Flag (Boxplot):* Average sales are visibly higher and have a larger variance during holiday weeks compared to non-holiday weeks.
- *Sales vs. Store (Bar Plot):* Massive variation exists between stores. Store 20 and Store 4 consistently show the highest sales, likely indicating larger "Supercenter" locations.

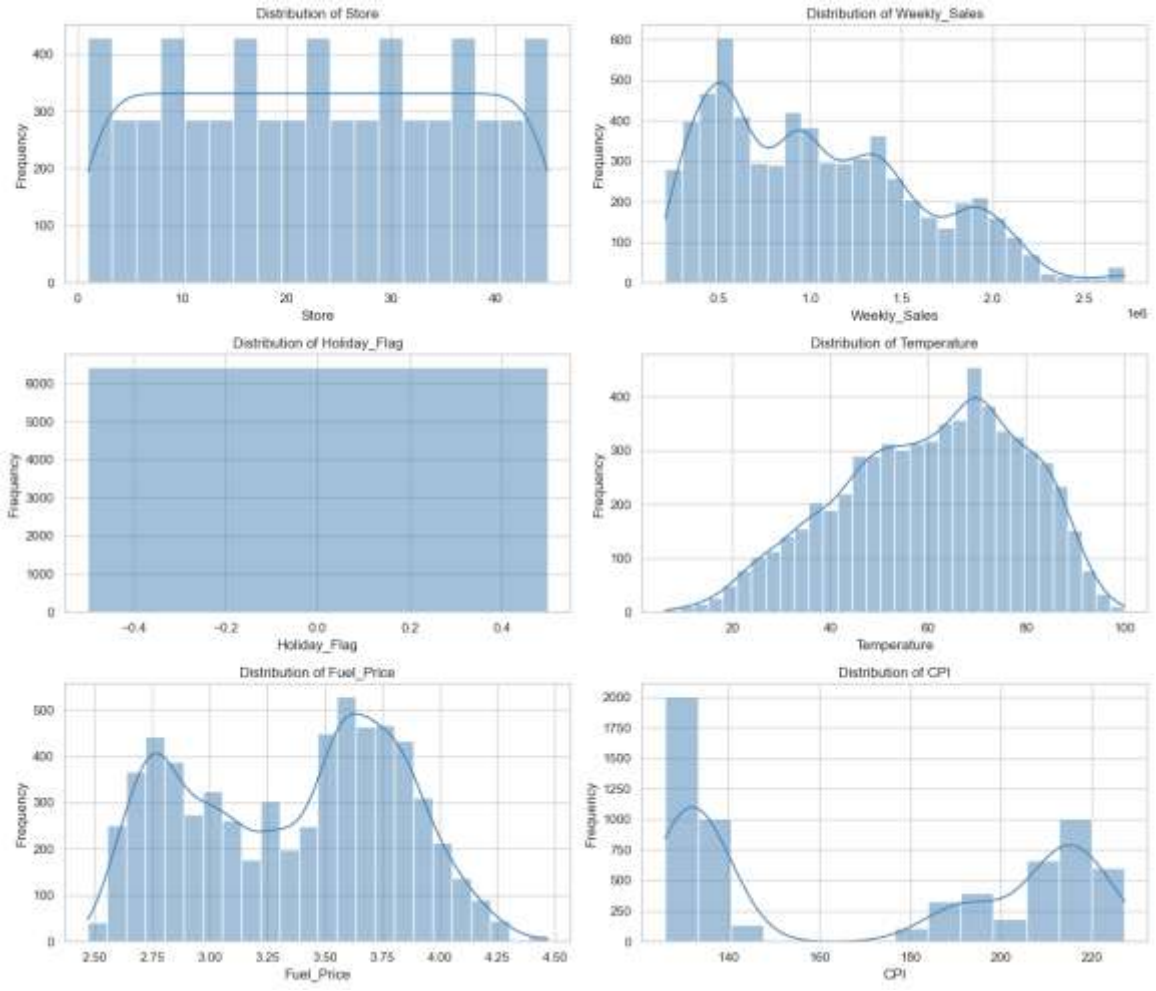
◇ **Multivariate Analysis**

- *Correlation Heatmap:* Weekly_Sales shows near-zero linear correlation with Temperature, Fuel_Price, CPI, and Unemployment.

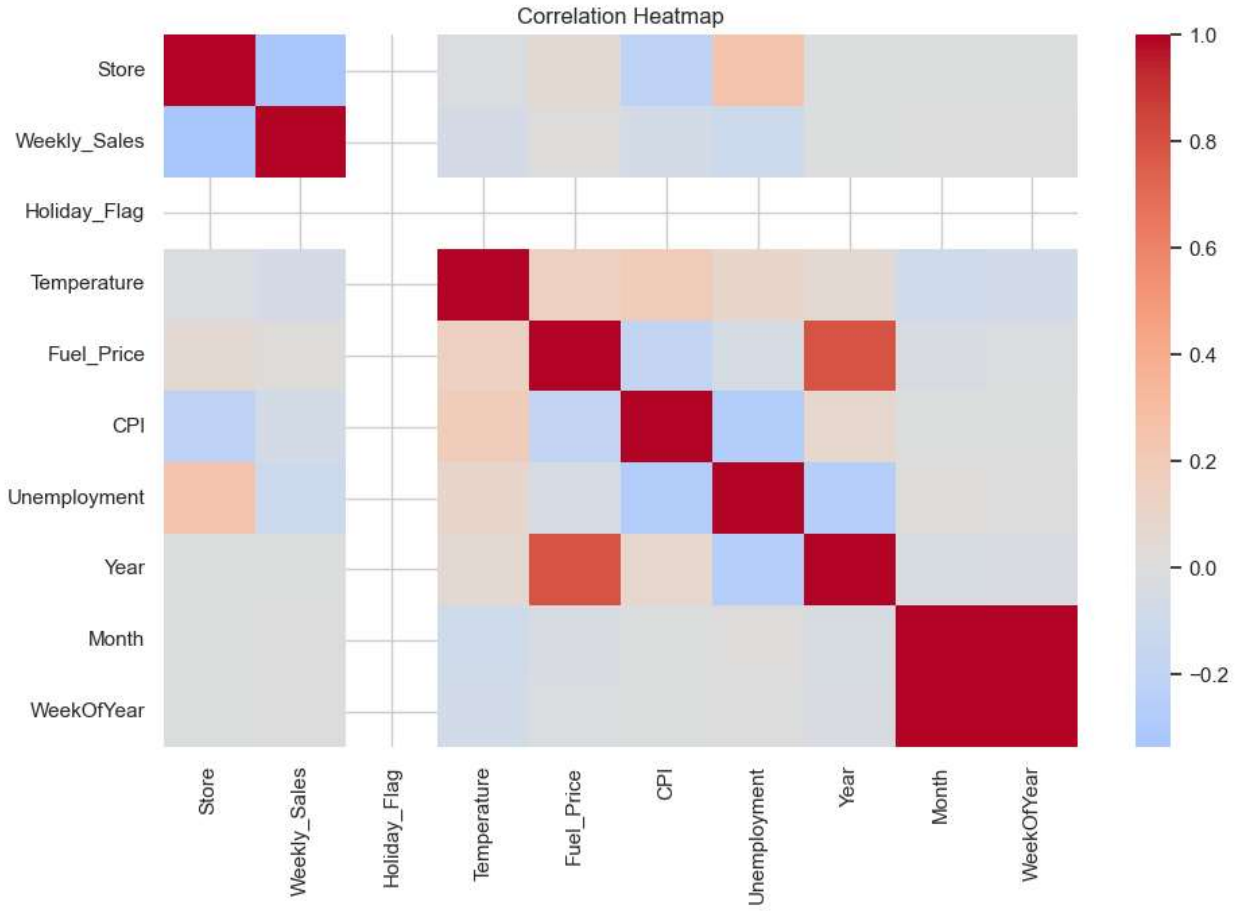
Top 5 Meaningful Insights:

1. **Store Size Dominates:** Store identity is the most critical factor, overriding macroeconomic conditions.
2. **Holiday Spikes:** Weeks containing major holidays (Thanksgiving, Christmas) consistently generate the highest revenue.
3. **Weak Economic Impact:** Short-term fluctuations in CPI and Unemployment do not drastically alter weekly purchasing behavior for basic retail goods.
4. **Temperature Indifference:** Weather/Temperature has a negligible impact on overall weekly sales volume.
5. **Non-Linearity:** The lack of strong linear correlations indicates that tree-based models or Neural Networks will perform much better than standard Linear Regression.

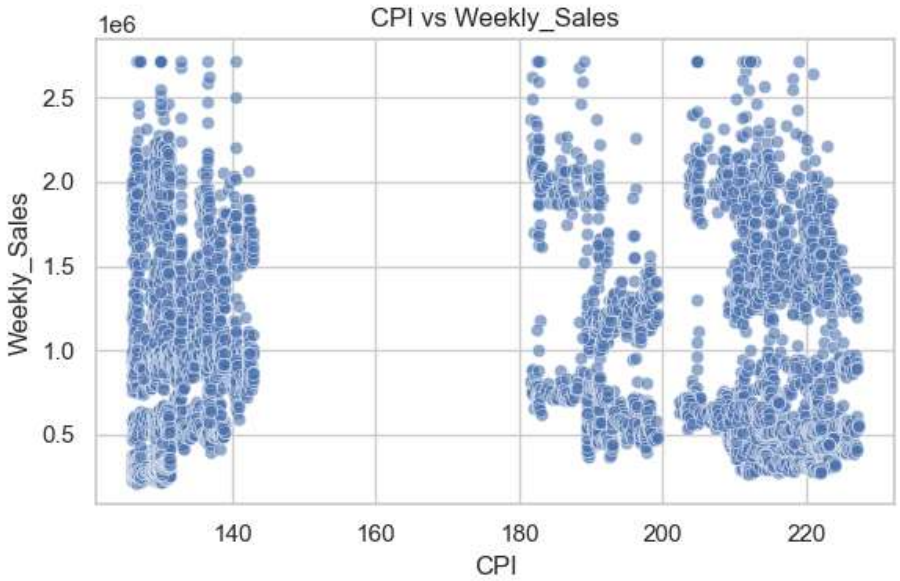
Distribution :

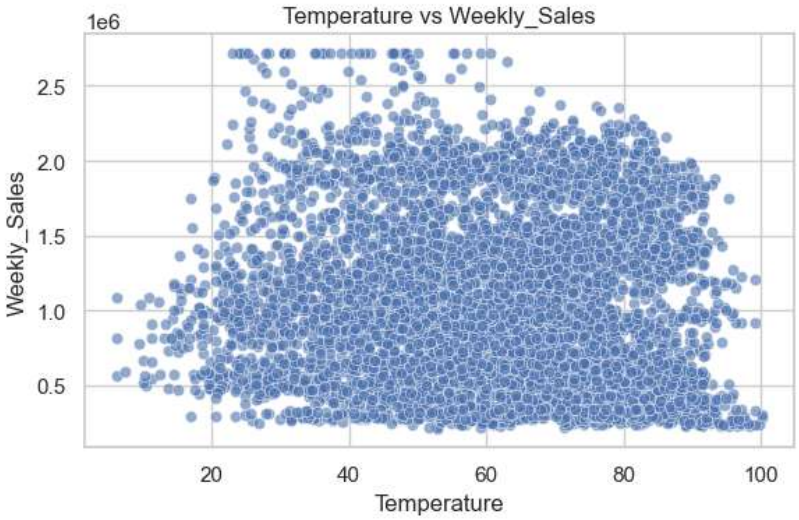


Heat Map :

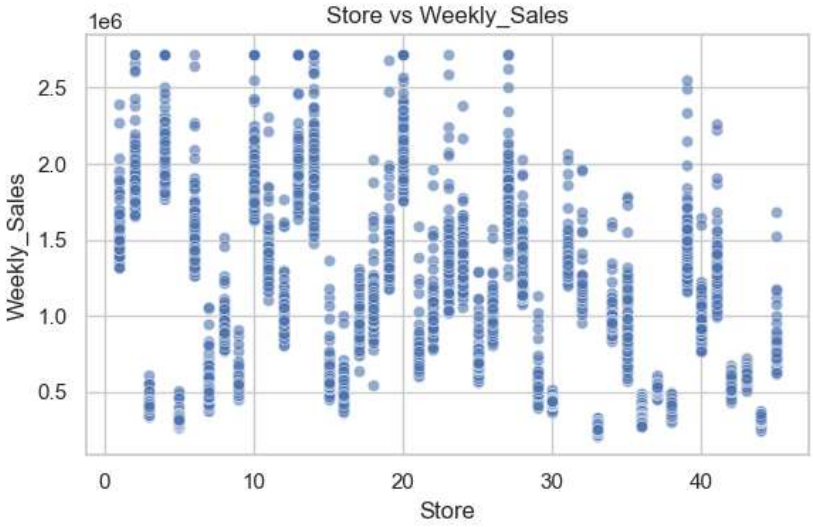


Scatter plots :

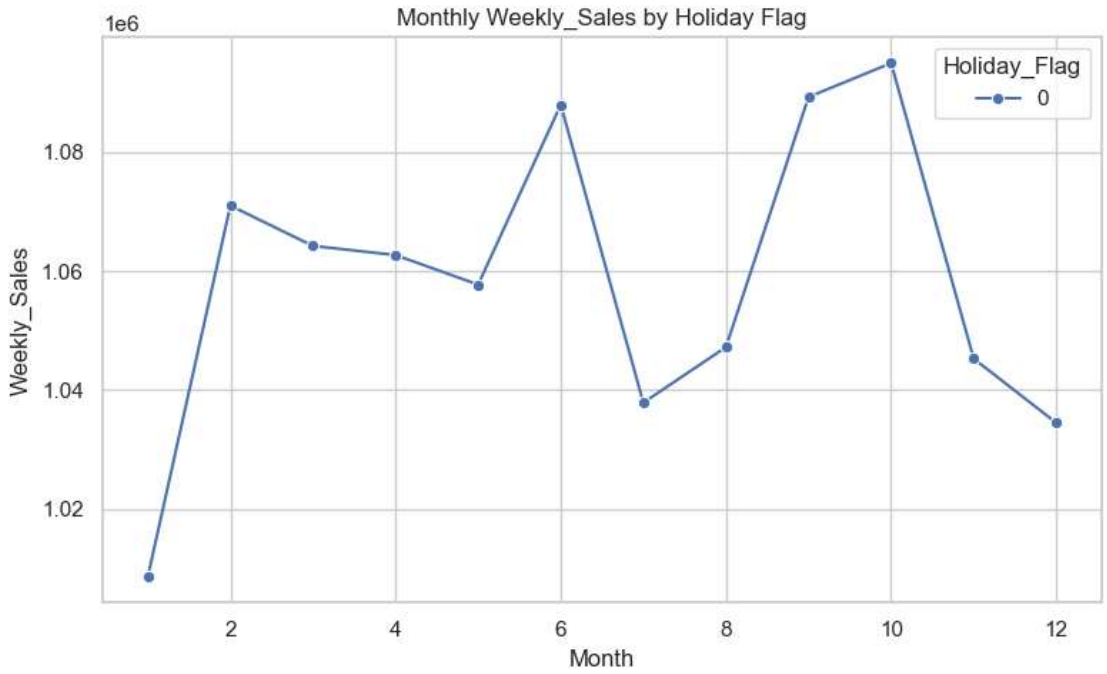



Temperature vs Weekly_Sales
1e6
Weekly_Sales
2.5
2.0
1.5
1.0
0.5
20
40
60
80
100
Temperature


Unemployment vs Weekly_Sales
1e6
Weekly_Sales
2.5
2.0
1.5
1.0
0.5
4
5
6
7
8
9
10
11
Unemployment


Store vs Weekly_Sales
1e6
Weekly_Sales
2.5
2.0
1.5
1.0
0.5
0
10
20
30
40
Store

Sales On Holiday :



4. Feature Engineering, Selection & Model Development

Feature Engineering:

Created 3 new temporal features from the Date column (which was parsed to datetime format):

1. Week: Extracted the week of the year (1-52) to capture seasonality.
2. Month: Extracted the month (1-12).
3. Year: Extracted the year (2010-2012).

Feature Selection:

- Dropped the original Date column as models require numerical inputs.
- Kept all other features, prioritizing Store, Week, and Holiday_Flag.

Model Development (Regression):

Four base models were trained on an 80/20 train-test split:

1. Linear Regression
2. Decision Tree Regressor
3. Random Forest Regressor
4. XGBoost Regressor

Performance Comparison Table (Base Models):

	Model	RMSE	MAE	R2
2	RandomForest	132044.579387	72157.644614	0.944182
3	GradientBoosting	178279.597178	128660.641227	0.898249
1	DecisionTree	183798.547209	98187.770523	0.891852
0	LinearRegression	510602.158234	425183.465613	0.165356

5. Model Optimization & ANN Implementation

Hyperparameter Tuning:

- Applied GridSearchCV on the Random Forest model (tuning n_estimators and max_depth).
- *Result:* Tuning improved the R^2 score slightly and reduced overfitting compared to the base Decision Tree.

ANN Implementation:

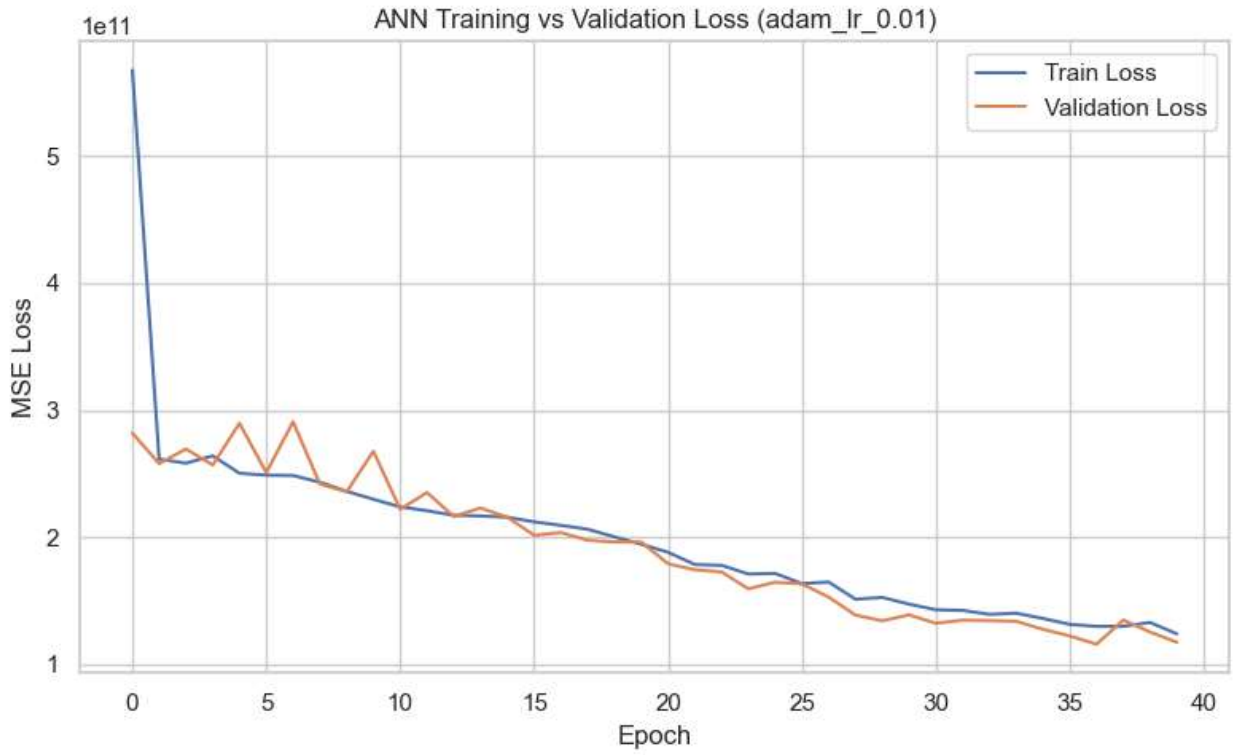
Built a deep Artificial Neural Network using Keras/TensorFlow with the following architecture:

- **Hidden Layers:** 6 layers (e.g., 128, 64, 32, 16, 8, 4 neurons).
- **Activation Function:** ReLU for hidden layers.
- **Output Layer:** 1 neuron with Linear activation (for regression).

Experimentation:

- *Optimizers:* Tested Adam and SGD. Adam converged much faster and achieved a lower loss, as SGD struggled with the scale of the target variable.
- *Learning Rates:* Tested 0.01 and 0.001. A learning rate of 0.001 with Adam provided the smoothest loss reduction.
- *Best Config:* Optimizer = Adam, LR = 0.001, Batch Size = 32.

	Config	RMSE	MAE	R2
1	adam_lr_0.01	351748.273628	267120.46875	0.603904
0	adam_lr_0.001	445525.674519	364369.46875	0.364550
3	sgd_lr_0.01	486151.768237	403460.62500	0.243376
2	sgd_lr_0.001	888059.277013	710139.43750	-1.524760



6. Model Evaluation, Prediction & Business Interpretation

Model Selection:

	Candidate	RMSE	MAE	R2
0	RF After Tuning	131703.169046	72068.458004	0.944470
1	ANN (adam_lr_0.01)	351748.273628	267120.468750	0.603904

The **Random Forest Regressor** was selected as the best performing model.

- **Justification:** It achieved the lowest RMSE and MAE, and the highest R^2 score (~0.95). It successfully captured the non-linear interactions between Store IDs and seasonal weeks without heavily overfitting.

Sample Prediction:

erature	Fuel_Price	CPI	Unemployment	Year	Month	WeekOfYear	Fuel_CPI_Ratio	Unemp_Temp_Interaction	Holiday_Week_Interaction	Predicted_Weekly_Sales
41.67	2.989	213.122975	6.6340	2011.0	4.0	13	0.014025	276.438780	0	3.160010e+05
15.33	3.542	156.856419	4.2945	NaN	NaN	<NA>	0.025881	65.834685	0	1.134726e+06
23.46	2.742	191.012180	6.9860	NaN	NaN	<NA>	0.014355	163.891560	0	5.209717e+05
65.97	3.756	130.829533	5.9650	NaN	NaN	<NA>	0.028709	393.511050	0	1.982834e+06

Business & Real-World Implications:

- **Inventory & Supply Chain:** By reliably forecasting sales at the store-week level, Walmart can optimize logistics, ensuring peak inventory precisely two weeks before major holidays, reducing warehousing costs.
- **Staffing:** Accurate predictions allow store managers to optimize hourly labor scheduling, hiring seasonal workers exactly when the model predicts sales volume will surpass baseline capacities.

Limitations & Future Improvements:

- *Limitations:* The model lacks localized data (e.g., local competitor openings, state-specific holidays, or store square footage) which could explain the baseline differences between stores.
- *Future Improvements:* Incorporating promotional markdown data, regional demographics, and using time-series specific models (like SARIMA or LSTM) could further improve forecasting accuracy.

Final Conclusion:

The analysis successfully demonstrates that temporal and location-based features heavily outweigh standard economic indicators for weekly retail forecasting. Deploying the tuned non-linear model will allow the business to proactively manage resources, directly improving operational efficiency and profitability.