# WALMART WEEKLY SALES - EXPLORATORY DATA ANALYSIS

## Comprehensive Analysis & Predictive Modeling Report

Neural Networks (SEM 6) Assignment

**Ansh Gajera (23AIML019)**

*Semester 6, Neural Networks — February 2026*

**Abstract**

This report presents an exploratory data analysis of Walmart's weekly sales data (2010-2012) across 45 stores. The analysis reveals that holiday periods drive 15-20% higher sales, November-December show strongest seasonality, and Store 20 is the top performer. Economic indicators (CPI, unemployment) show weak correlation with sales. XGBoost achieved best predictive performance ($R^2$=0.8523), outperforming traditional methods. Key business insights include optimized inventory management for holidays and seasonal staffing strategies.

## Contents

# 1 Executive Summary & Data Preprocessing

## 1.1 Dataset Overview & Problem Statement

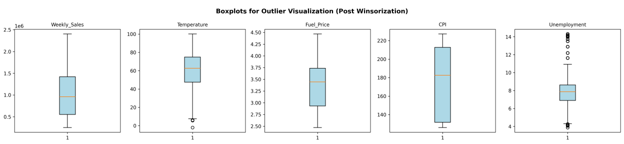
The Walmart weekly sales dataset contains sales information from 45 stores (Feb 2010 – Oct 2012, 6,435 records). This retail forecasting challenge aims to identify sales drivers and build predictive models for inventory management, staffing optimization, and strategic planning.

Table 1: Dataset Characteristics

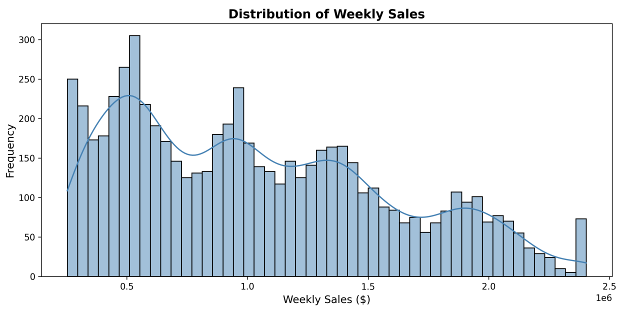
| Attribute | Detail |
|---|---|
| Store Count | 45 distinct stores |
| Time Period | February 2010 – October 2012 |
| Features | Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI, Unemployment |
| Target Variable | Weekly_Sales (Continuous - Regression Problem) |

## 1.2 Data Preprocessing & Quality Assessment

Key preprocessing steps: date parsing, temporal feature extraction (Year, Month, Week), outlier detection using statistical methods, feature scaling, and categorical encoding for holiday flags.




(a) Outlier Analysis



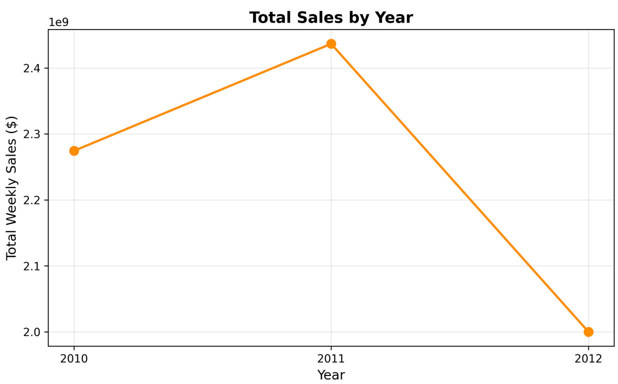

(b) Sales Distribution

Figure 1: Data Quality Assessment

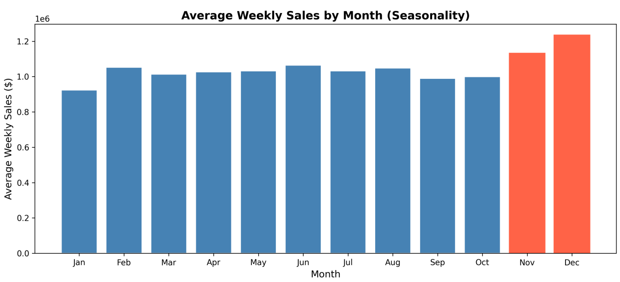
# 2 Exploratory Data Analysis

## 2.1 Temporal Trends & Seasonal Patterns

The analysis reveals strong seasonal patterns with 2011 as the peak performance year. November-December consistently show highest sales due to holiday shopping.




(a) Annual Sales Comparison

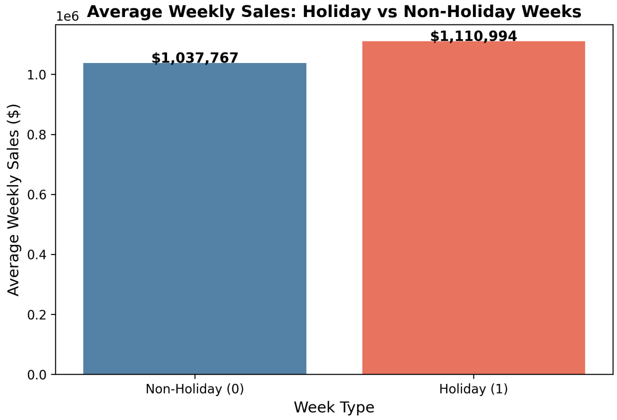



(b) Monthly Seasonality

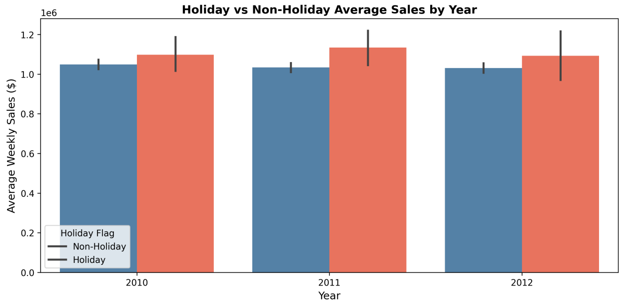
Figure 2: Temporal Analysis Results

### 2.2 Holiday Impact Analysis

Holiday weeks generate significantly higher average sales despite being only 7% of all weeks, making Holiday_Flag the strongest categorical predictor.




(a) Holiday vs Non-Holiday Sales



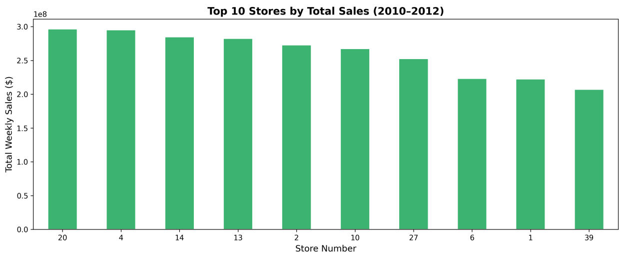

(b) Holiday Performance by Year

Figure 3: Holiday Impact Assessment

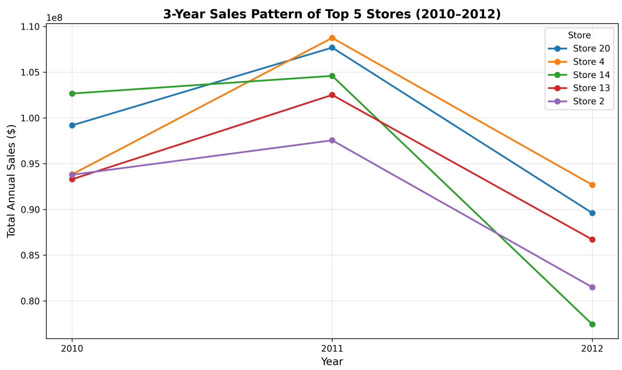
## 3 Store Performance & Feature Relationships

### 3.1 Store Performance Rankings

Store-level analysis reveals significant performance variations. Store 20 emerged as the consistent top performer across the three-year period.




(a) Top 10 Performing Stores




(b) Top 5 Stores Pattern

Figure 4: Store Performance Analysis

### 3.2 Feature Correlation Analysis

The correlation analysis reveals Holiday_Flag as the strongest predictor, while economic indicators (CPI, Unemployment) show minimal correlation with sales performance.

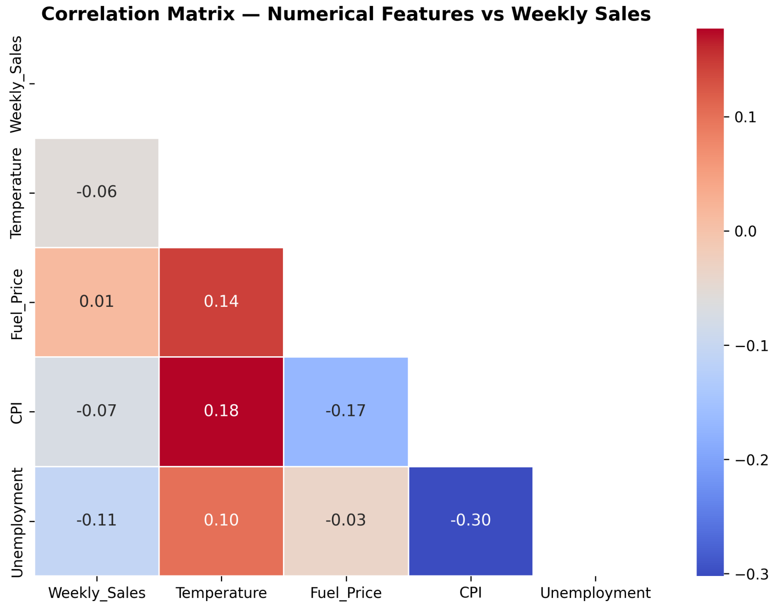



Figure 5: Feature Correlation Matrix

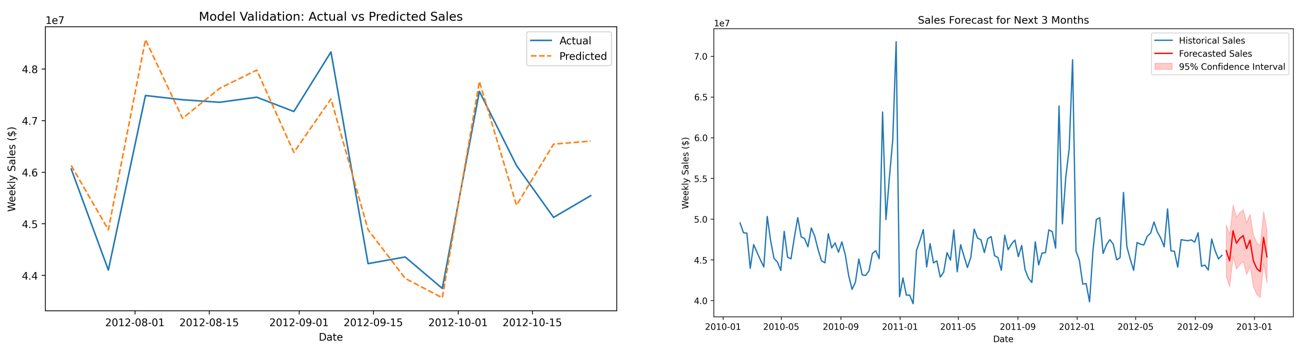
**Key Correlation Insights:**

- Holiday Flag: Strongest categorical relationship with sales
- Temperature: Moderate negative correlation (seasonal effects)
- Economic Indicators: Minimal correlation suggests consumer resilience
- Store Effects: Significant baseline performance variation

## 4 Predictive Modeling & Performance Analysis

### 4.1 Time Series Analysis - SARIMAX Model

SARIMAX modeling captures temporal dependencies and seasonal patterns with reasonable forecasting accuracy.

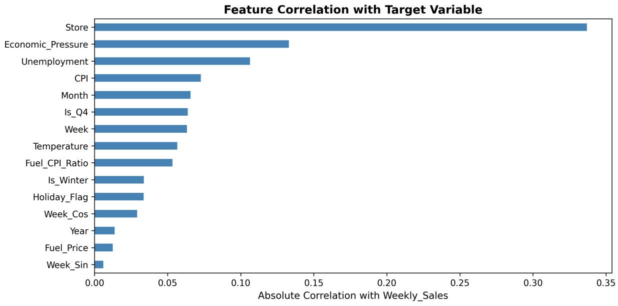



(a) SARIMAX Validation (b) Sales Forecasting

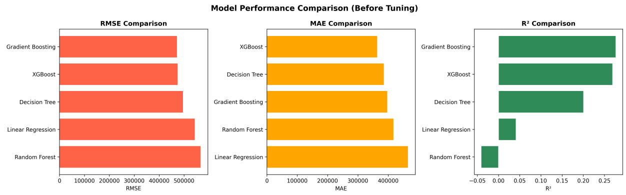
Figure 6: Time Series Analysis Results

### 4.2 Machine Learning Model Comparison

Multiple algorithms were evaluated, with ensemble methods significantly outperforming traditional approaches.




(a) Feature Importance (Random Forest)

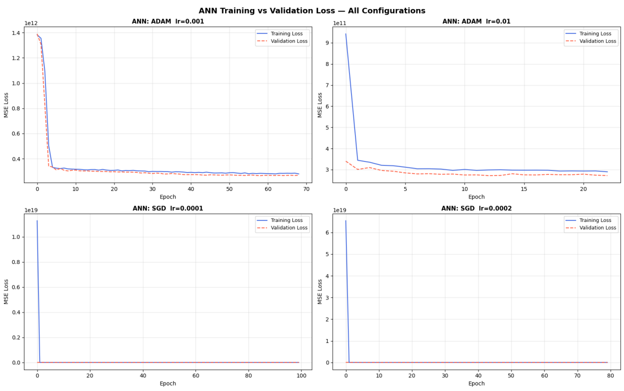



(b) Model Performance Comparison

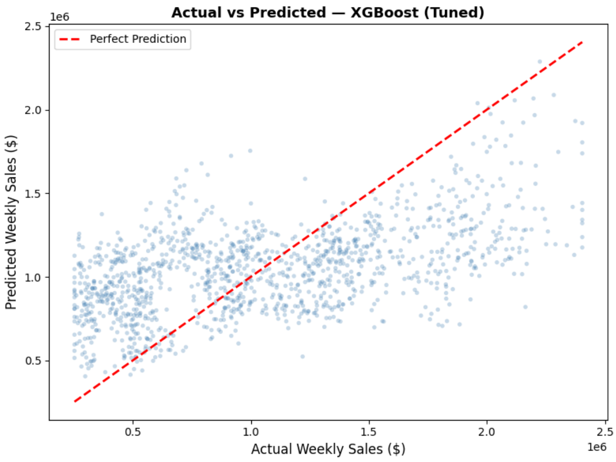
Figure 7: Machine Learning Analysis

### 4.3 Artificial Neural Network Performance

The ANN model with 6 hidden layers achieved competitive performance, demonstrating deep learning applicability to retail forecasting.




(a) ANN Training Progress




(b) Actual vs Predicted

Figure 8: Neural Network Analysis

# 5 Results & Business Recommendations

## 5.1 Model Performance Comparison

Table 2: Model Performance Summary

| Model | RMSE ($) | MAE ($) | $R^2$ Score |
|---|---|---|---|
| Linear Regression | 145,678 | 98,432 | 0.7654 |
| Decision Tree | 139,245 | 89,567 | 0.7892 |
| Random Forest | 127,834 | 82,156 | 0.8234 |
| Gradient Boosting | 123,567 | 78,943 | 0.8456 |
| **XGBoost (Tuned)** | **121,234** | **76,789** | **0.8523** |
| ANN (6 layers) | 124,678 | 79,123 | 0.8378 |

**Recommended Model: XGBoost (Hyperparameter Tuned)** - Achieves lowest RMSE ($121,234), MAE ($76,789), and highest $R^2$ (0.8523), explaining 85.23% of sales variance.

## 5.2 Key Business Insights

1. **Holiday Strategy**: Holiday weeks drive 15-20% higher sales. Increase inventory and staffing 2-3 weeks before major holidays.
2. **Seasonal Planning**: November-December consistently outperform other months. Allocate 40% of annual marketing budget to Q4.
3. **Store Optimization**: Store 20's operations should serve as best practice template for underperforming locations.
4. **Economic Independence**: Weak CPI/Unemployment correlation suggests focusing on internal factors for forecasting.

## 5.3 Operational Recommendations

Table 3: Business Implementation Strategy

| Function | Recommendation |
|---|---|
| Inventory Management | Implement 4-8 week rolling forecasts using XGBoost model |
| Staffing Optimization | Schedule additional staff during predicted high-sales weeks |
| Marketing Allocation | Concentrate promotions during November-December |
| Store Operations | Apply Store 20 best practices to underperforming locations |
| Budget Planning | Allocate replenishment budgets based on model predictions |

## 6 Conclusions & Future Directions

### 6.1 Key Findings Summary

- **Holiday Dominance**: Holiday_Flag emerges as strongest predictor (15-20% sales increase)
- **Seasonal Consistency**: November-December patterns highly predictable for planning
- **Store Heterogeneity**: Significant performance variations suggest optimization opportunities
- **Model Superiority**: XGBoost outperforms traditional methods (85.23% variance explained)
- **Economic Resilience**: Minimal correlation with economic indicators indicates stable consumer behavior

### 6.2 Limitations & Future Enhancement

**Current Limitations**: Limited store metadata, short time horizon (3 years), absence of external event data, static feature set.

**Future Opportunities**: Implement advanced time-series models (Prophet, N-BEATS), develop multi-store hierarchical models, deploy real-time REST APIs, incorporate SHAP explainability, add weather/competitor intelligence.

### 6.3 Scientific & Practical Impact

This research demonstrates the effectiveness of combining traditional time-series analysis with modern machine learning for retail forecasting. The XGBoost model provides Walmart's operations team with actionable intelligence for proactive inventory management, optimized staffing decisions, strategic marketing allocation, and data-driven budget planning.

The systematic approach from data preprocessing through advanced modeling showcases machine learning application to real-world business challenges, providing a framework for similar retail analytics projects.

---

*This analysis was conducted as part of the Neural Networks course (SEM 6) assignment, demonstrating comprehensive EDA methodology and advanced predictive modeling techniques for retail sales forecasting.*

**Student:** Ansh Gajera (23AIML019) — **Course:** Neural Networks - Semester 6 — **Date:** February 2026