Walmart Weekly Sales Analysis Part 2: Statistical Analysis with Python

by

Temidayo Olowoyeye

Data and Business Analyst

1. Introduction

In retail, accurately predicting future sales is essential to managing inventory, reallocating resources, and effectively improving business planning. As one of the largest retailers globally, Walmart will likely utilize sales forecast models to optimize its operations. In part one of this analysis, SQL and Power BI were adopted to provide insight into KPIs. This part looks into the numerical variables in the dataset to address questions such as the adequacy of featured variables for predicting future sales, the significance of independent variables on weekly sales, and identifying the best predictive models. Thus providing valuable insights that can enhance the decision-making process.

1.1. Objective

The aim of this analysis is to;

i. Analyze the trends of variables over time.

ii. Determine the model that best predicts Walmart's future weekly sales.

III. Develop a strategic business insight (SBI) that will promote business growth

1.2. Hypothesis Testing

To answer the question in the introductory part, the following hypothesis was formulated for testing.

- Hypothesis 1:
  - *H*0:- The featured variables will sufficiently predict future weekly sales at Walmart.
- Hypothesis 2:
  - *H*0:-Not all independent variables significantly affect Walmart's weekly sales.

1.3 About the Dataset

The data was extracted from link. It contains the Export sales records from 2010 to 2012, with 6435 rows and eight (8) columns.

1.4. Data Processing

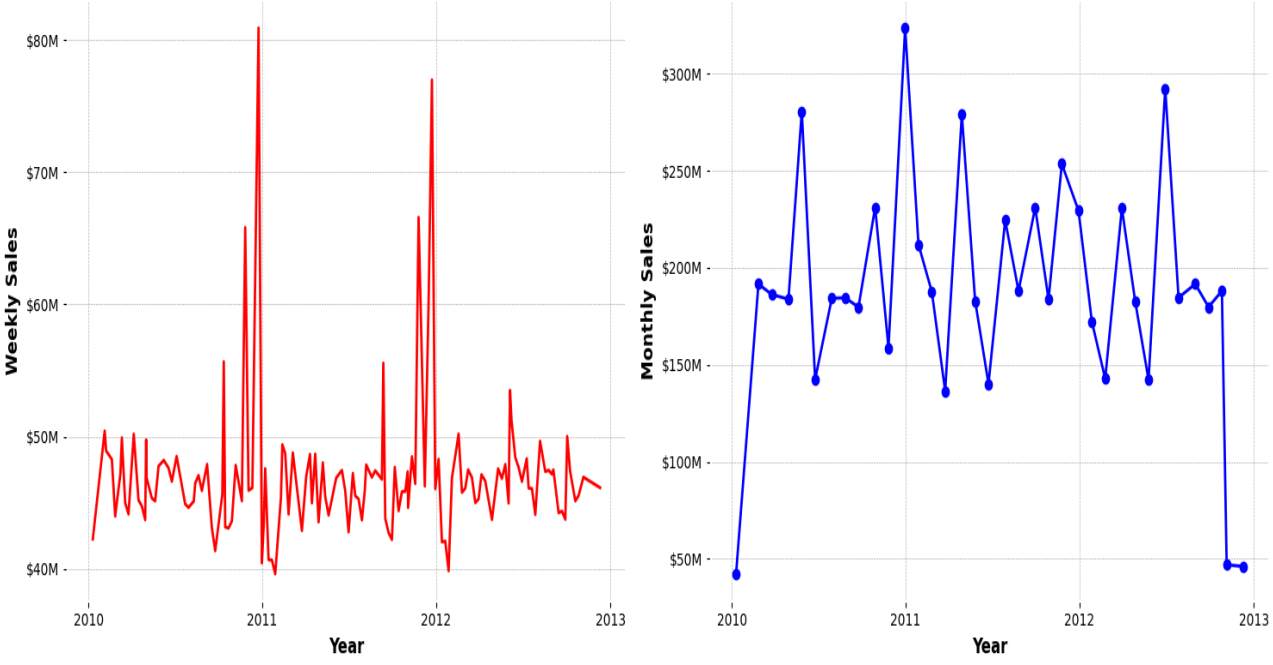
The dataset was cleaned and reshaped to fit the analysis's objective. After loading the data into the data frame, the following process was carried out: identify and handle missing values, remove duplicates, add and transform columns, and restructure the data frame for optimization. This practice ensures the dataset's accuracy, dependability, and readiness for subsequent analysis. This and all other analysis processes were done using Python, and the codes can be accessed via this link.

2. Data Analysis and Modeling

2.1. Trend Analysis

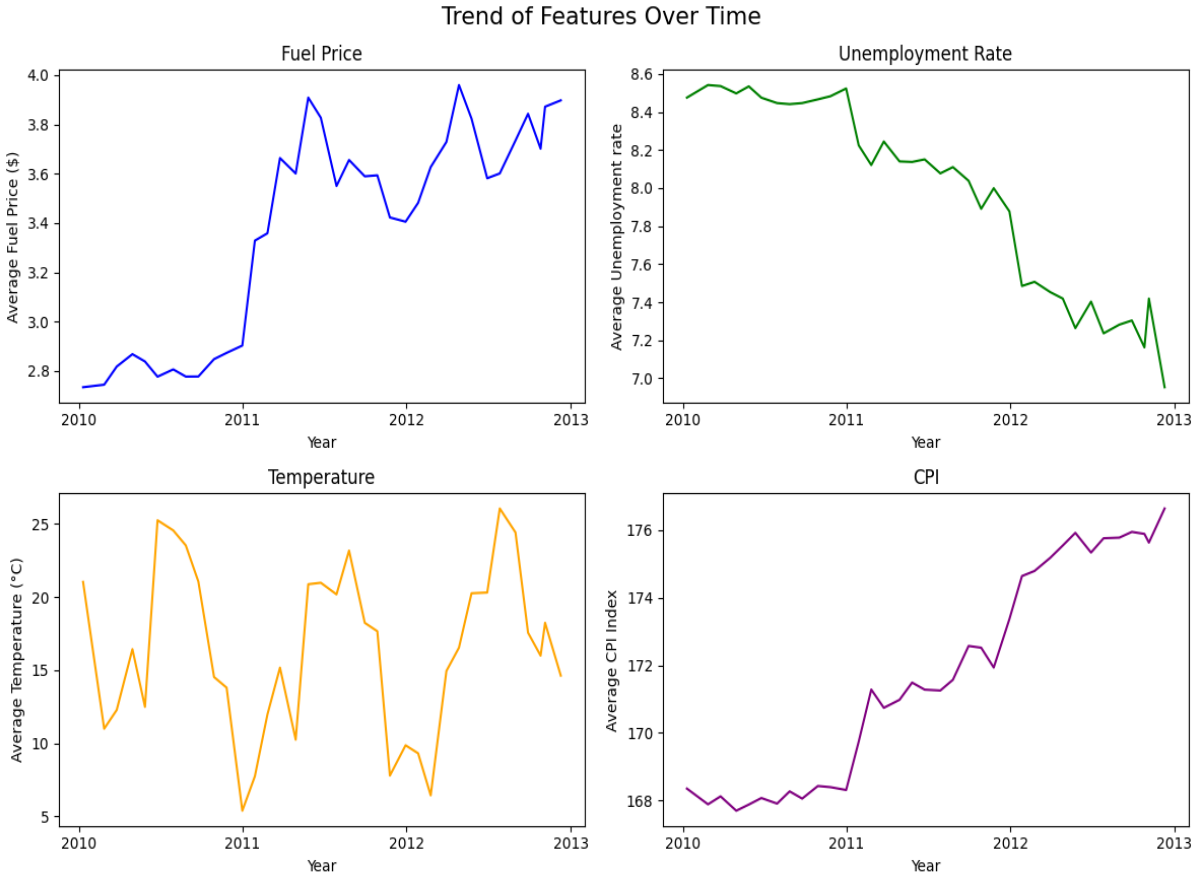
2.1.1. Sales Trend Over Time

Weekly sales exhibited relative stability from 2010 to 2012, consistently between the $40 million to $50 million threshold. However, notable spikes were observed during the fourth quarter of 2010 and 2011. These spikes can be attributed to various factors, including festive seasons and promotional activities. The increased consumer spending typically associated with holidays and special promotions likely contributed to the observed surge in sales during these periods. However, while examining monthly sales trends, a different pattern characterized by fluctuations emerges. December 2010 marked the peak in sales, reaching over $300 million, while December 2012 recorded the lowest sales at $48 million. This variability could be attributed to numerous factors, including competition, pricing strategies, seasonal trends, economic conditions, consumer behaviour and many more.



2.1.2. Trend of Features Overtime

The continuous increase in fuel prices from 2010 to 2012, representing an approximate 35.71% rise, mirrors a similar trend observed in the Customer Price Index (CPI), which showed a continuous increase of 4.76% during this period. Conversely, a consistent decline in the Unemployment rate was observed, marking a reduction of 21.43% over the same timeframe. However, it's worth noting that the temperature trend during this period exhibited fluctuations, indicating variability rather than a consistent pattern of change. This suggests that while economic indicators such as fuel prices, CPI, and unemployment rates followed relatively predictable trajectories, environmental factors like temperature displayed more erratic behaviour over the same period.



2.2. Exploratory Data Analysis (EDA)

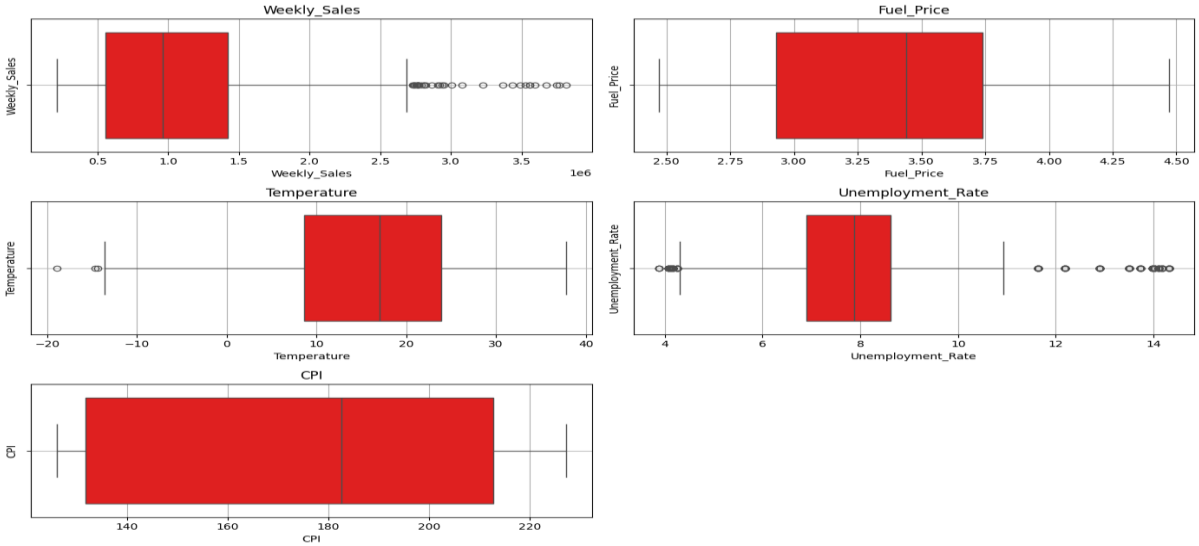
Before running models, the following analysis was conducted to understand our dataset.

2.2.1. Univariate Analysis

Histograms with Kernel Density Estimation (KDE) curve were utilized to analyze the distribution of the numerical dataset. From the output, it was observed that only the Unemployment Rate has a normal distribution pattern. This implies that the choice of models will be influenced. Since the assumption of normality is often associated with specific models like linear regression, having non-normally distributed predictors might limit the suitability of those models. However, other models like decision trees, random forests, support vector machines, and neural networks are more flexible and can handle non-normally distributed variables effectively.

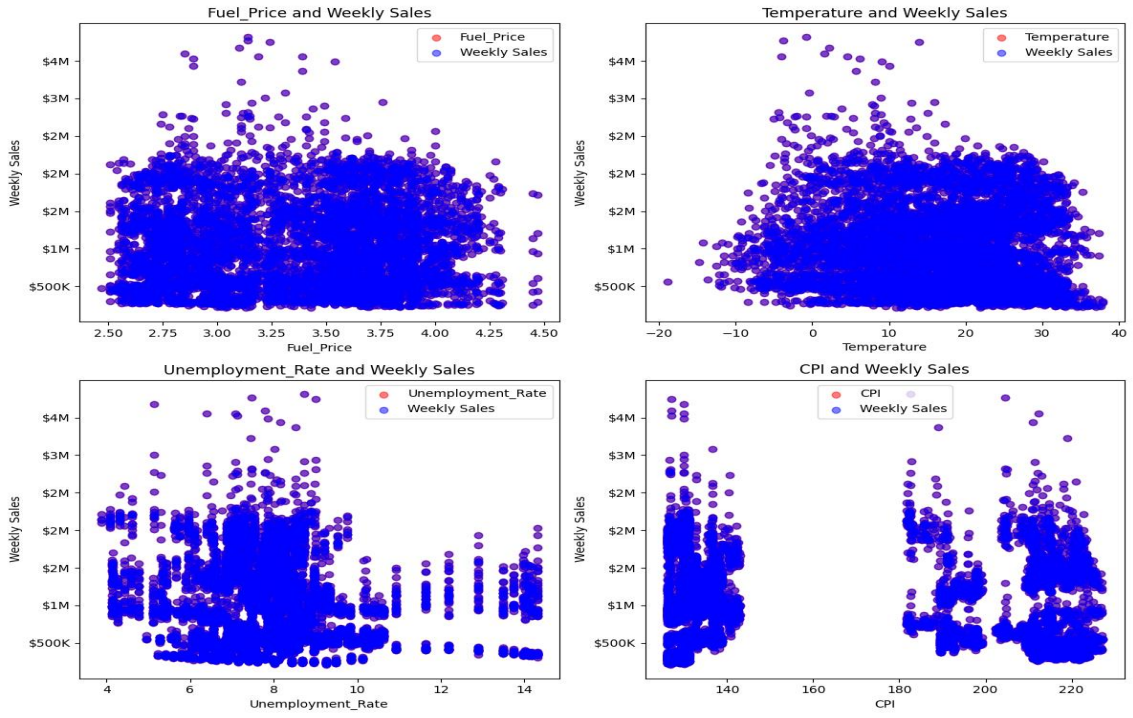
Additionally, Box plots were employed to assess the presence of outliers in the numerical dataset. Outliers were detected for variables such as weekly sales, temperature, and unemployment rate, with a higher prevalence for weekly sales. While it's common practice to consider removing outliers from a dataset, particularly to maintain data integrity, this approach was opted against for the current analysis. This decision is based on the understanding that weekly sales data may exhibit occasional spikes due to various factors,

and removing outliers could potentially lead to the loss of valuable information regarding sales fluctuations.



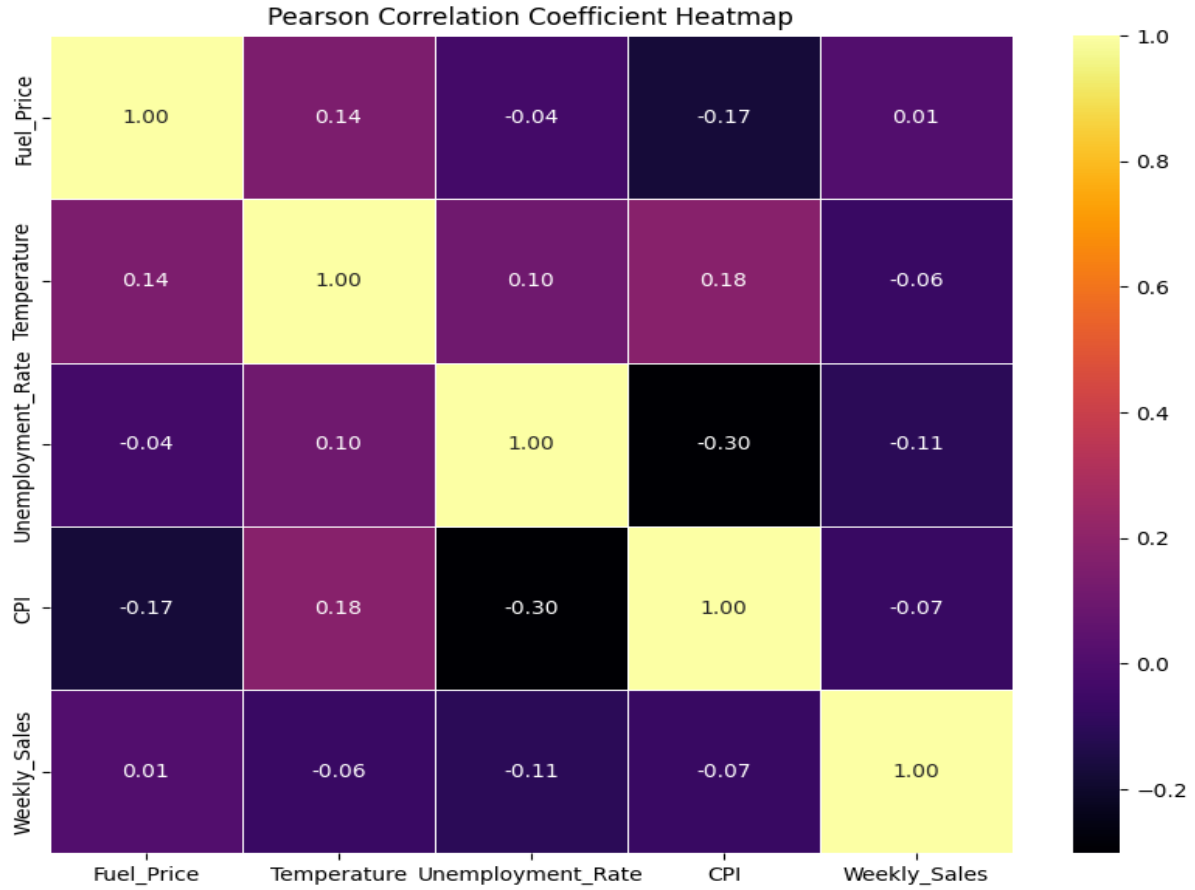
2.2.2. Bivariate Analysis

The relationship between the dependent and independent variables was examined using a Scatter plot. It was noted that the relationship between the dependent and feature variables is not linear. This suggests a weak association between these variables, indicating that changes in one variable do not lead to proportional changes in the other. Additionally, this observation will influence the choice of models to adopt, as linear regression may not be suitable given the non-linear relationship between the variables.



2.2.3. Correlation Analysis

Further analysis of the relationship between variables affirms our findings using scattered plots. A negative correlation coefficient suggests an inverse relationship between CPI, temperature, unemployment rate, and weekly sales. Conversely, fuel price exhibits a positive correlation with weekly sales. While these correlations imply negative and positive associations between the variables and weekly sales, their strengths are relatively weak, I.e., they only have a minor influence on weekly sales.



2.2.4. Check for Multicollinearity

it is essential to ascertain the relationship between variables to avoid overfitting; we adopt two methods for this analysis. Firstly, by checking with the correlation matrix and setting the threshold for multicollinearity to 0.7, any variable with a strong relationship above 0.7 will be considered for removal from the model. The second approach is to validate the correlation matrix using the Variance Inflation Factor (VIF). VIF quantifies how much the variance of an estimated coefficient for a particular independent variable is increased because of multicollinearity. A VIF value of 1 indicates no multicollinearity, while values greater than 1 suggest increasing levels of multicollinearity. The VIF threshold was set to 5 to identify levels of multicollinearity. From the two results, multicollinearity was not detected for the independent variables, suggesting that the variables included in the model are not highly correlated.

```
Multicollinearity Result for each independent variable:
No multicollinearity detected for 'Temperature'.
No multicollinearity detected for 'Fuel_Price'.
No multicollinearity detected for 'CPI'.
No multicollinearity detected for 'Unemployment_Rate'.
```

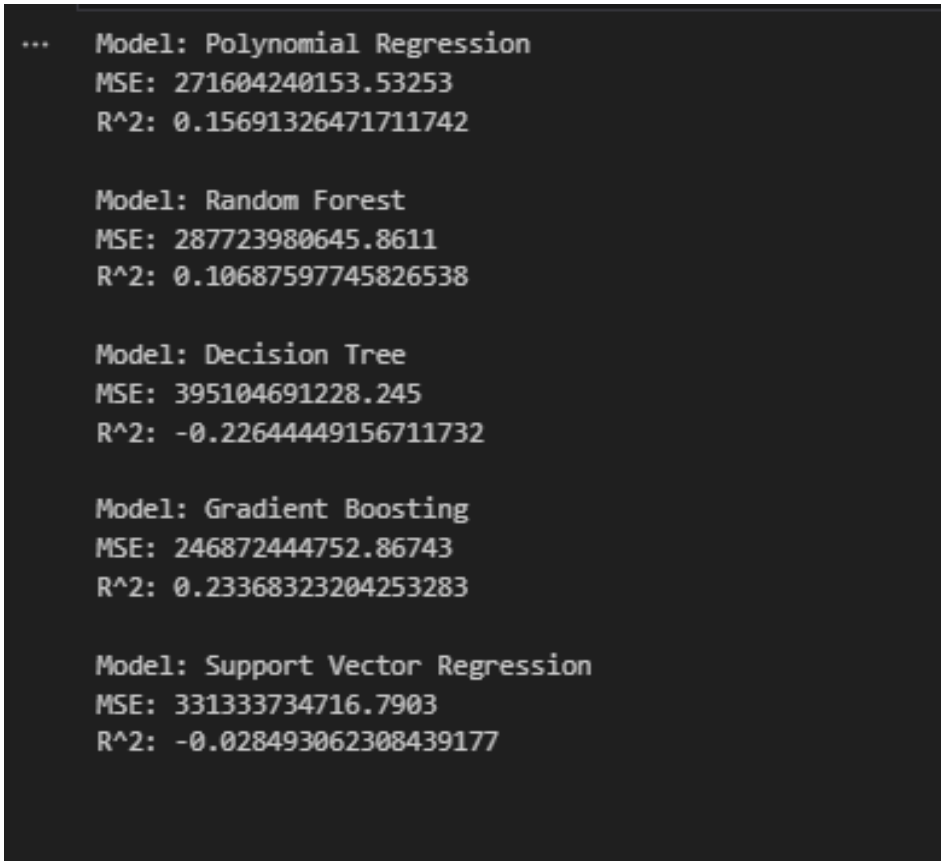
```
Variance Inflation Factor (VIF) for each independent variable:
            Variable       VIF Multicollinearity
0        Temperature  1.104939               No
1         Fuel_Price  1.081745               No
2                CPI  1.220793               No
3  Unemployment_Rate  1.149246               No
```

2.3. Modelling

Given the characteristics and relationships observed in the dataset, selecting models that align with these attributes is crucial. The following models were chosen for their adaptability and resilience in handling the dataset's characteristics, which include a non-normal distribution, lack of linear relationships, and weak/negative correlations:

- Polynomial Regression
- Random Forest
- Decision Tree
- Gradient Boosting
- Support Vector Regression

These models were run to determine their effectiveness in predicting future weekly Walmart sales. However, the results indicate that none of the models achieved satisfactory performance. The R-squared values for each model fell below 50%, suggesting that the variables in the dataset do not adequately explain the variation in weekly sales. Consequently, the hypothesis that the featured variables would sufficiently predict future weekly sales at Walmart is rejected. Instead, the alternative hypothesis, asserting that the featured variables cannot sufficiently predict future weekly sales at Walmart, is accepted.



To test hypothesis 2, which posits that "not all independent variables significantly affect Walmart's weekly sales," a Generalized Linear Model (GLM) was used. The results are as follows:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:          Weekly_Sales   No. Observations:                 5148
Model:                           GLM   Df Residuals:                     5143
Model Family:               Gaussian   Df Model:                            4
Link Function:              Identity   Scale:                     3.0961e+11
Method:                         IRLS   Log-Likelihood:                -75407.
Date:               Mon, 29 Apr 2024   Deviance:                  1.5923e+15
Time:                       23:14:21   Pearson chi2:                1.59e+15
No. Iterations:                    3   Pseudo R-squ. (CS):           0.02615
Covariance Type:           nonrobust
==============================================================================
                     coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.045e+06   7756.441    134.699      0.000    1.03e+06    1.06e+06
Temperature   -1.266e+04   8160.614     -1.552      0.121   -2.87e+04    3332.137
Fuel_Price    -1.019e+04   8102.502     -1.257      0.209   -2.61e+04    5692.899
CPI           -6.424e+04   8577.297     -7.489      0.000   -8.11e+04   -4.74e+04
Unemployment_Rate -8.18e+04   8316.071     -9.836      0.000   -9.81e+04   -6.55e+04
==============================================================================
```

Interpretation of p-values:

- Temperature and Fuel_Price: The p-values associated with these variables are greater than the typical significance level of 0.05, specifically 0.121 and 0.209, respectively. This suggests that these variables are not statistically significant predictors of weekly sales.
- CPI and Unemployment: In contrast, the p-values for CPI and Unemployment are less than 0.05, indicating statistical significance. Specifically, both p-values are approximately zero.

Conclusion: Based on the results, the alternate hypothesis (*H1*) is rejected because not all independent variables significantly affect weekly sales. Therefore, we accept the null hypothesis (*H0*), which states that not all independent variables significantly affect weekly sales.

3. Strategic Business Insight (SBI)

The following insight can be adopted to boost business growth;

- Maintain focus on marketing efforts and promotional campaigns based on factors impacting weekly sales. For example, during periods of economic downturn (indicated by high unemployment rates), target promotions and discounts can be launched to attract price-sensitive customers. Similarly, during periods of high CPI inflation, pricing strategies can be adjusted to remain competitive.
- Enhancing forecasting accuracy by integrating significant variables into forecasting models can help anticipate demand fluctuations and adjust inventory levels accordingly, optimizing resource allocation.
- Recognizing that certain factors influencing sales may not be captured in the analysis highlights the importance of identifying and incorporating additional data sources and variables to enhance forecasting accuracy.
- Streamlining datasets by removing non-significant variables can improve data quality and focus resources on collecting new variables with greater predictive power, thereby refining the forecasting process.
- Prioritizing statistically significant variables ensures high-quality data for key predictors, leading to more reliable and accurate analyses and forecasts.

4. Conclusion

Analyzing Walmart's weekly sales dataset reveals that not all independent variables significantly affect sales. However, further analysis provides Strategic Business Insights (SBI) to optimize operations and drive sales growth. Recommendations from the analysis include aligning marketing efforts with economic indicators, enhancing forecasting accuracy, and prioritizing statistically significant variables for more reliable analyses and forecasts. These strategies aim to improve operational efficiency and sales performance.