Statistical Analysis Report: Walmart Sales Data

1. Business Problem

1.1. Overview

The prime business idea for this analysis is to identify the factors affecting Walmart's weekly sales. Walmart, being one of the largest retail corporations globally, requires precise and data-driven insights to optimize its sales strategies. Specifically, the company is interested in understanding the impact of economic indicators (such as unemployment rates and CPI) and special events (holidays) on weekly sales across different stores.

1.2. Key Business Questions

1. Do holidays significantly impact Walmart's weekly sales?
2. How do economic indicators such as CPI and unemployment influence weekly sales?
3. Can we build a predictive model to estimate weekly sales based on economic factors and holiday flags?

2. Dataset Description

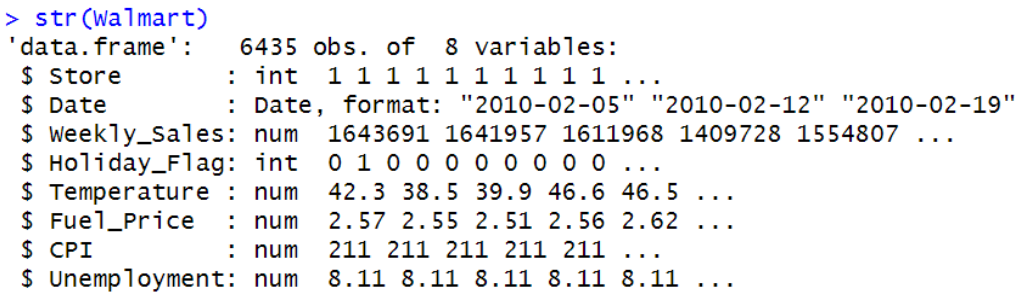
2.1. Data Source

The dataset used for this analysis is the Walmart Sales data, which includes 6,435 observations and 8 variables. The variables included Store, Date, Weekly_Sales, Holiday_Flag, Temperature, Fuel_Price, CPI and Unemployment.

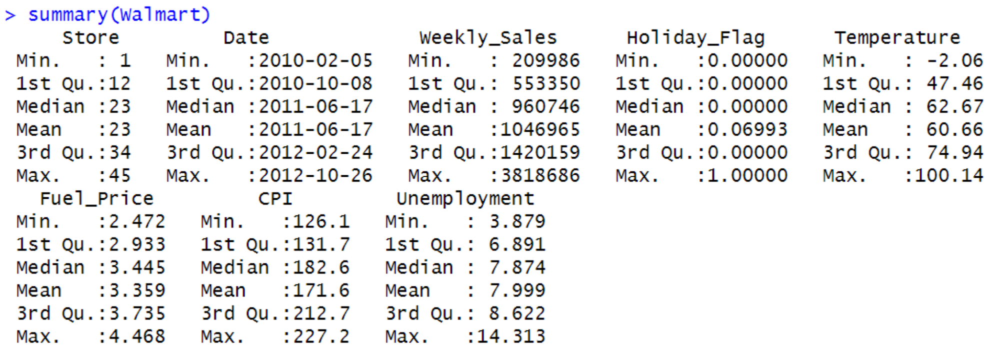
2.2. Initial Data Inspection

A summary of the data structure and initial descriptive statistics were performed. The data revealed the following:

- The dataset contains no missing values or duplicate rows.
- The sales data (Weekly_Sales) is highly variable, ranging from $209,986 to $3,818,686.
- There is a relatively low frequency of holidays (Holiday_Flag mean = 0.07).



```
> str(Walmart)
'data.frame':	6435 obs. of  8 variables:
 $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Date        : Date, format: "2010-02-05" "2010-02-12" "2010-02-19"
 $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
 $ Holiday_Flag: int  0 1 0 0 0 0 0 0 0 0 ...
 $ Temperature : num  42.3 38.5 39.9 46.6 46.5 ...
 $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
 $ CPI         : num  211 211 211 211 211 ...
 $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
```



```
> summary(Walmart)
     Store         Date             Weekly_Sales      Holiday_Flag      Temperature    
 Min.   : 1   Min.   :2010-02-05   Min.   : 209986   Min.   :0.00000   Min.   : -2.06  
 1st Qu.:12   1st Qu.:2010-10-08   1st Qu.: 553350   1st Qu.:0.00000   1st Qu.: 47.46  
 Median :23   Median :2011-06-17   Median : 960746   Median :0.00000   Median : 62.67  
 Mean   :23   Mean   :2011-06-17   Mean   :1046965   Mean   :0.06993   Mean   : 60.66  
 3rd Qu.:34   3rd Qu.:2012-02-24   3rd Qu.:1420159   3rd Qu.:0.00000   3rd Qu.: 74.94  
 Max.   :45   Max.   :2012-10-26   Max.   :3818686   Max.   :1.00000   Max.   :100.14  
   Fuel_Price         CPI          Unemployment   
 Min.   :2.472   Min.   :126.1   Min.   : 3.879  
 1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891  
 Median :3.445   Median :182.6   Median : 7.874  
 Mean   :3.359   Mean   :171.6   Mean   : 7.999  
 3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622  
 Max.   :4.468   Max.   :227.2   Max.   :14.313  
```

3. Data Preparation and Cleaning

3.1. Data Conversion

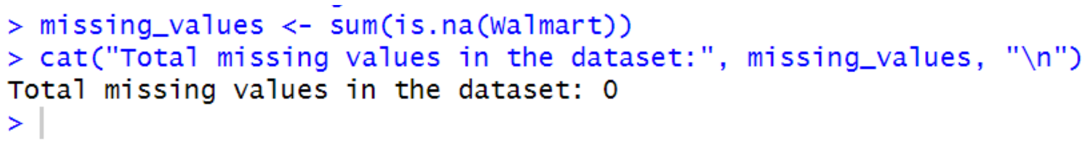
The Date variable was converted from string format to a date format for better handling in time series analysis. The data types of the other variables were confirmed to be appropriate for analysis.

3.2. Outlier Detection and Removal

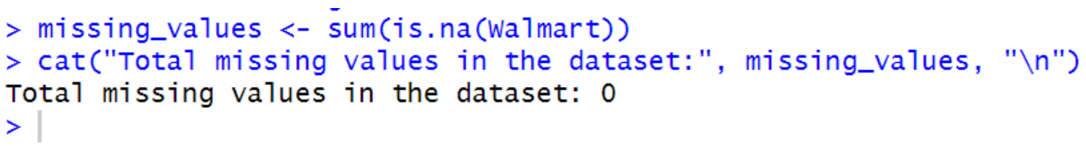
Outliers in Weekly_Sales were identified using the Interquartile Range (IQR) method. Observations with weekly sales outside the range of 1.5 times the IQR from the 1st quartile (Q1) and the 3rd quartile (Q3) were removed. This step reduced the dataset slightly but improved the robustness of subsequent analyses.

3.3. Final Dataset Summary

After cleaning, the dataset was re-summarized. The range of Weekly_Sales was narrowed (minimum of $209,986 and maximum of $2,685,352), ensuring that extreme outliers were effectively removed without discarding useful information.



```
> missing_values <- sum(is.na(Walmart))
> cat("Total missing values in the dataset:", missing_values, "\n")
Total missing values in the dataset: 0
> 
```



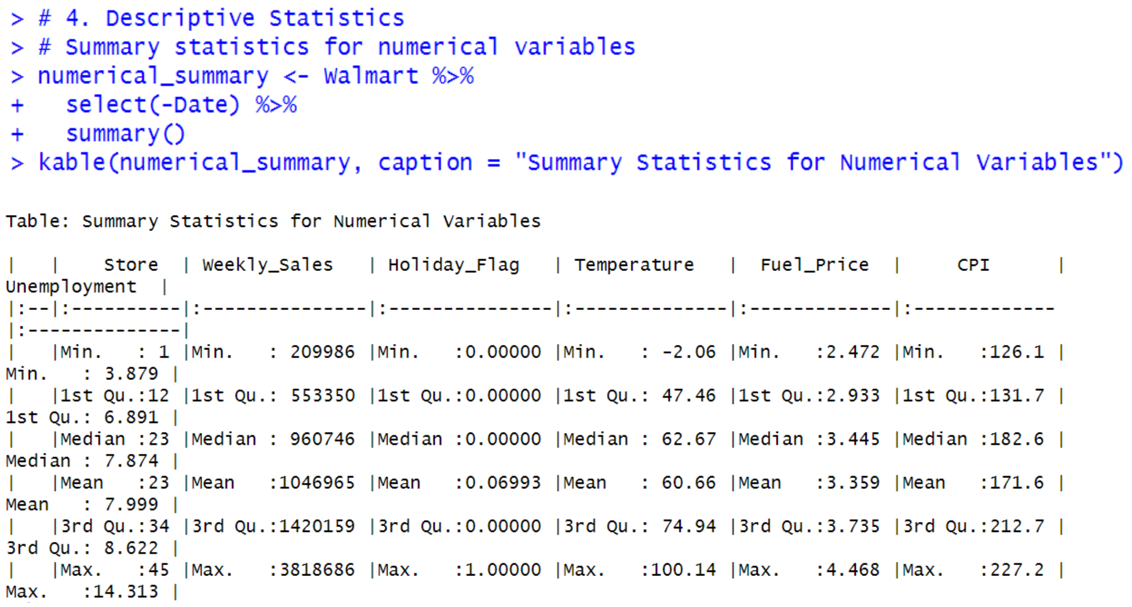
```
> missing_values <- sum(is.na(Walmart))
> cat("Total missing values in the dataset:", missing_values, "\n")
Total missing values in the dataset: 0
> 
```

4. Descriptive Statistics and Exploratory Data Analysis

4.1. Summary Statistics

Descriptive statistics were computed for all numerical variables. The key insights include:

- **Store**: The dataset covers 45 stores.
- **Temperature**: The range of temperature is between -2.06°F to 100.14°F, beside a mean of 60.66°F.
- **Weekly_Sales**: The average weekly sales across all stores are approximately $1,046,965.
- **Economic Indicators**: CPI and unemployment show considerable variation, indicating potential economic instability during the observed period.



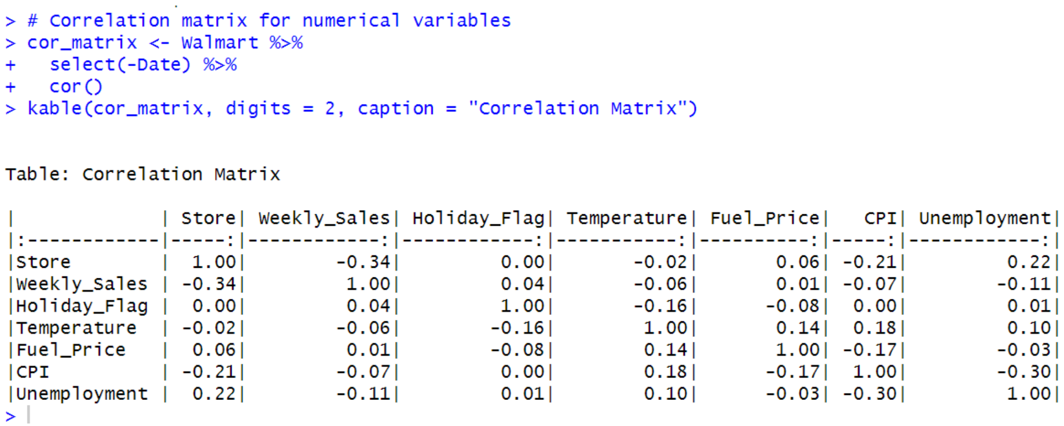
```
> # 4. Descriptive Statistics
> # Summary statistics for numerical variables
> numerical_summary <- Walmart %>%
+   select(-Date) %>%
+   summary()
> kable(numerical_summary, caption = "Summary Statistics for Numerical Variables")

Table: Summary Statistics for Numerical Variables

|   |    Store   | Weekly_Sales   | Holiday_Flag   | Temperature    |  Fuel_Price   |      CPI      |
Unemployment  |
|:--|:----------|:---------------|:---------------|:--------------|:-------------|:-------------
|:--------------|
|   |Min.   : 1 |Min.   : 209986 |Min.   :0.00000 |Min.   : -2.06 |Min.   :2.472 |Min.   :126.1 |
Min.   : 3.879 |
|   |1st Qu.:12 |1st Qu.: 553350 |1st Qu.:0.00000 |1st Qu.: 47.46 |1st Qu.:2.933 |1st Qu.:131.7 |
1st Qu.: 6.891 |
|   |Median :23 |Median : 960746 |Median :0.00000 |Median : 62.67 |Median :3.445 |Median :182.6 |
Median : 7.874 |
|   |Mean   :23 |Mean   :1046965 |Mean   :0.06993 |Mean   : 60.66 |Mean   :3.359 |Mean   :171.6 |
Mean   : 7.999 |
|   |3rd Qu.:34 |3rd Qu.:1420159 |3rd Qu.:0.00000 |3rd Qu.: 74.94 |3rd Qu.:3.735 |3rd Qu.:212.7 |
3rd Qu.: 8.622 |
|   |Max.   :45 |Max.   :3818686 |Max.   :1.00000 |Max.   :100.14 |Max.   :4.468 |Max.   :227.2 |
Max.   :14.313 |
```

4.2. Correlation Analysis

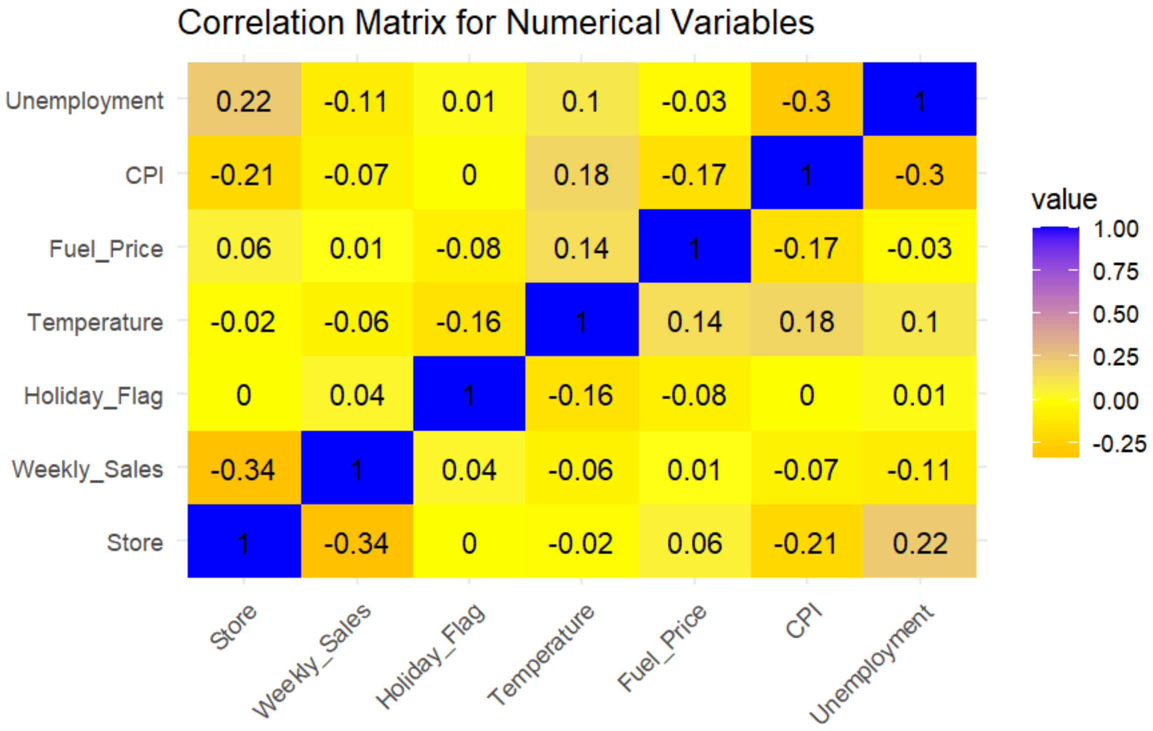
A correlation matrix is given to see associations between numerical variables. Notable correlations include:

- A reasonable negative correlation involving Store and Weekly_Sales (-0.34), suggesting that some stores consistently perform better or worse.

- A significant negative correlation between CPI and Unemployment (-0.30), which is consistent with economic theory.



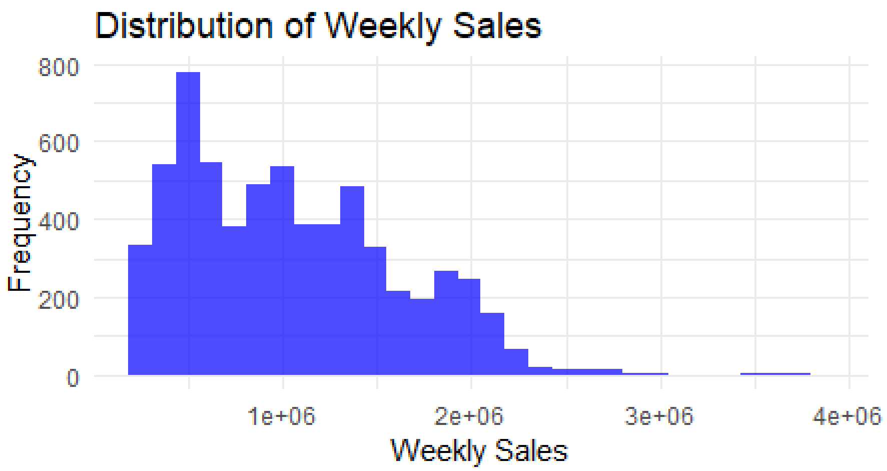
```
> # Correlation matrix for numerical variables
> cor_matrix <- Walmart %>%
+   select(-Date) %>%
+   cor()
> kable(cor_matrix, digits = 2, caption = "Correlation Matrix")


Table: Correlation Matrix

|             | Store| Weekly_Sales| Holiday_Flag| Temperature| Fuel_Price|   CPI| Unemployment|
|:------------|-----:|------------:|------------:|-----------:|----------:|-----:|------------:|
|Store        |  1.00|        -0.34|         0.00|       -0.02|       0.06| -0.21|         0.22|
|Weekly_Sales | -0.34|         1.00|         0.04|       -0.06|       0.01| -0.07|        -0.11|
|Holiday_Flag |  0.00|         0.04|         1.00|       -0.16|      -0.08|  0.00|         0.01|
|Temperature  | -0.02|        -0.06|        -0.16|        1.00|       0.14|  0.18|         0.10|
|Fuel_Price   |  0.06|         0.01|        -0.08|        0.14|       1.00| -0.17|        -0.03|
|CPI          | -0.21|        -0.07|         0.00|        0.18|      -0.17|  1.00|        -0.30|
|Unemployment |  0.22|        -0.11|         0.01|        0.10|      -0.03| -0.30|         1.00|
> |
```



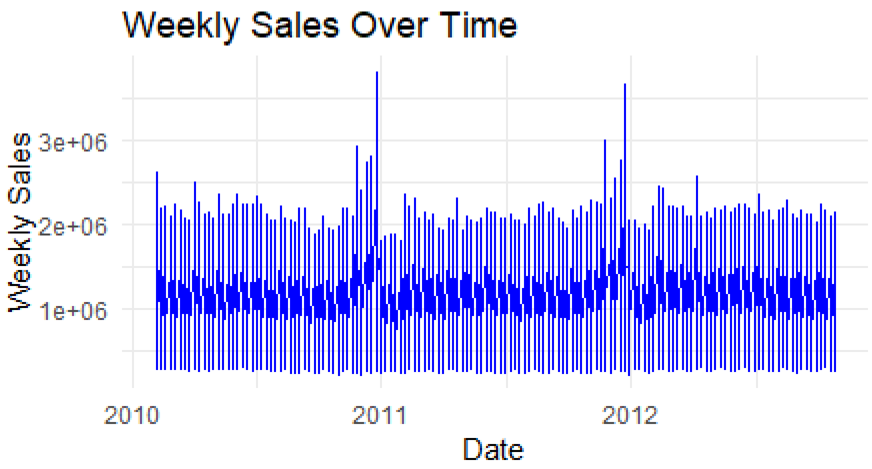
4.3. Visualizations

Several visualizations were created:

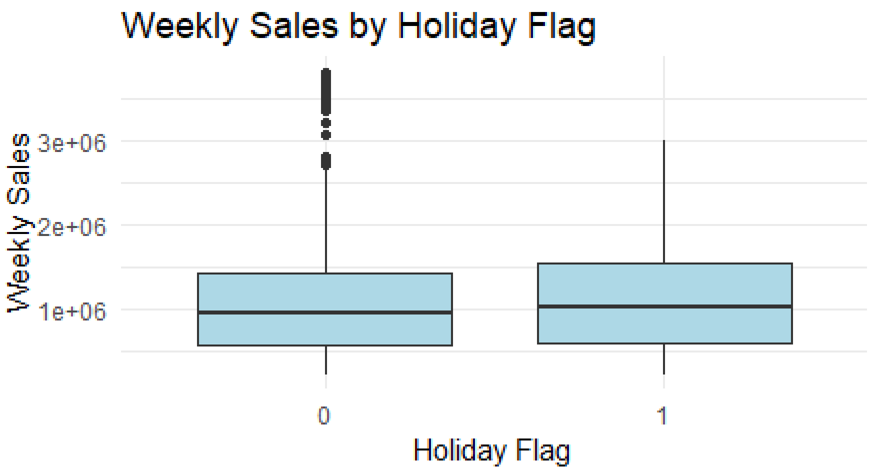
- **Histogram**: Distribution of Weekly_Sales, revealing a right-skewed distribution.



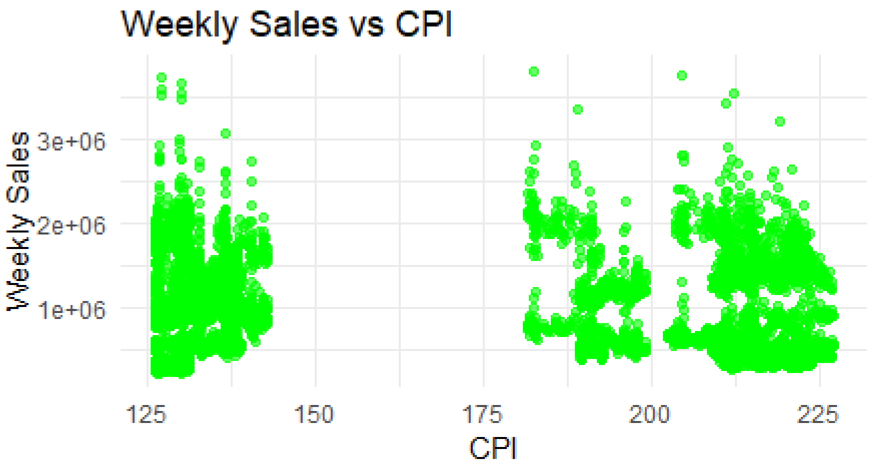
- **Time Series Plot**: Weekly sales over time, showing seasonal patterns.



- **Boxplot**: Comparison of weekly sales between holiday and non-holiday weeks.



- **Scatter Plots**: Relationship between weekly sales and economic indicators (CPI and Unemployment).

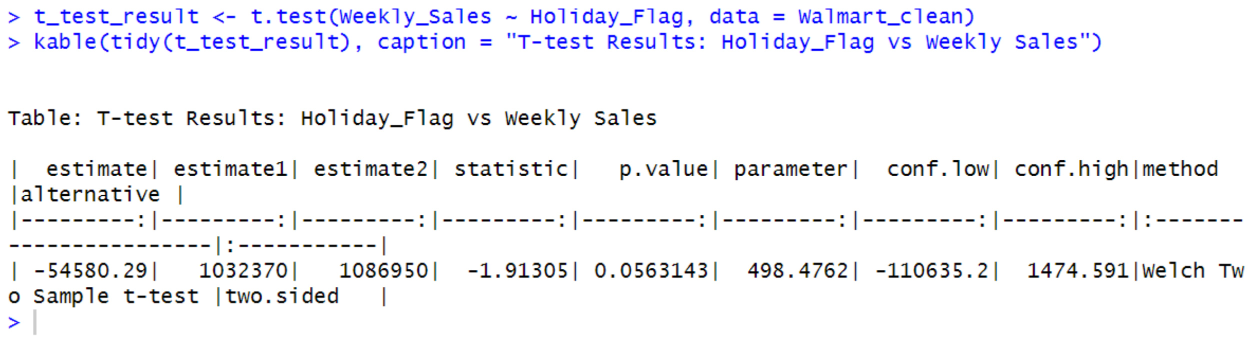


5. Hypothesis Testing and Inferential Statistics

5.1. Hypothesis 1: Holiday Impact on Sales

Null Hypothesis (H0): Holidays do not significantly impact weekly sales.
Alternative Hypothesis (H1): Holidays significantly impact weekly sales.

A t-test was performed to compare weekly sales during holiday and non-holiday weeks. The p-value was 0.056, slightly above the p-value of 0.05. Thus, we fail to eliminate the null hypothesis, indicating that holidays do not have a statistically substantial impact on weekly sales.

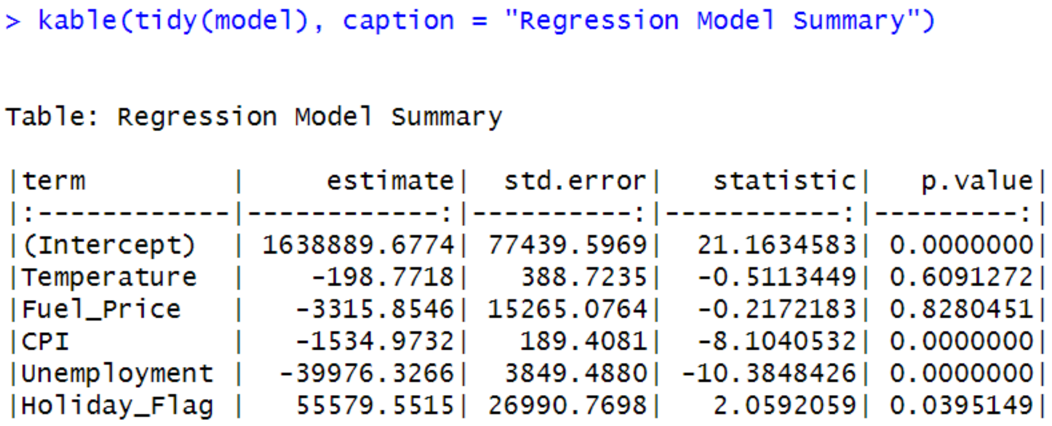


```
> t_test_result <- t.test(Weekly_Sales ~ Holiday_Flag, data = Walmart_clean)
> kable(tidy(t_test_result), caption = "T-test Results: Holiday_Flag vs Weekly Sales")


Table: T-test Results: Holiday_Flag vs Weekly Sales

|  estimate| estimate1| estimate2| statistic|   p.value| parameter|  conf.low| conf.high|method
|alternative |
|---------:|---------:|---------:|---------:|---------:|---------:|---------:|---------:|:-------
----------------|:-----------|
| -54580.29|   1032370|   1086950|  -1.91305| 0.0563143|  498.4762| -110635.2|  1474.591|Welch Tw
o Sample t-test |two.sided   |
> |
```

5.2. Hypothesis 2: Economic Indicators' Impact on Sales

Null Hypothesis (H0): Economic indicators (CPI, Unemployment) do not significantly impact weekly sales.
Alternative Hypothesis (H1): Economic indicators significantly impact weekly sales.

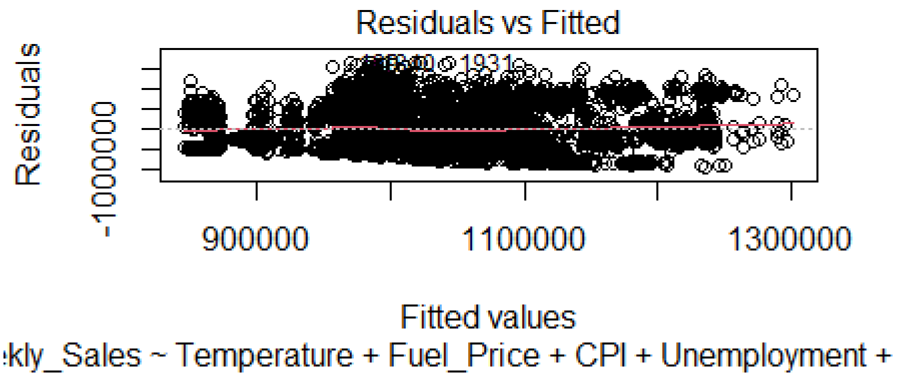
A linear regression model was built with Weekly_Sales as the dependent variable and Temperature, Fuel_Price, CPI, Unemployment, and Holiday_Flag as independent variables. The model showed that CPI and Unemployment are statistically significant predictors of weekly sales, with intercept of 0.000.



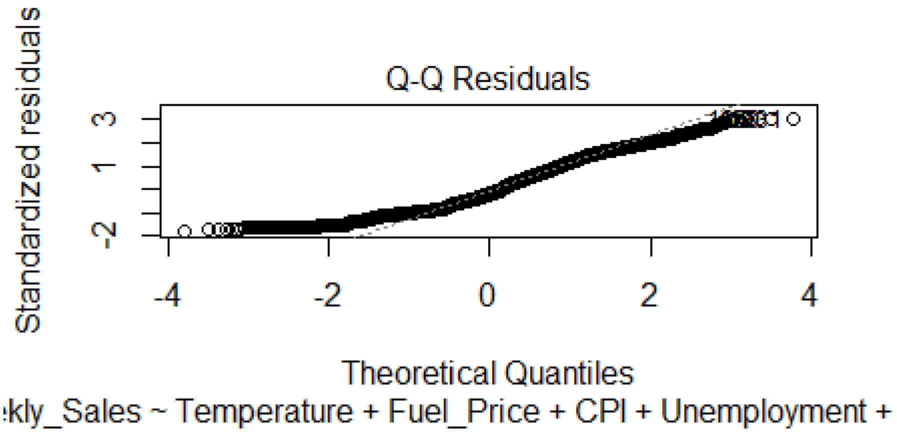
```
> kable(tidy(model), caption = "Regression Model Summary")


Table: Regression Model Summary

|term         |     estimate|  std.error|   statistic|   p.value|
|:------------|------------:|----------:|-----------:|---------:|
|(Intercept)  | 1638889.6774| 77439.5969|  21.1634583| 0.0000000|
|Temperature  |    -198.7718|   388.7235|  -0.5113449| 0.6091272|
|Fuel_Price   |   -3315.8546| 15265.0764|  -0.2172183| 0.8280451|
|CPI          |   -1534.9732|   189.4081|  -8.1040532| 0.0000000|
|Unemployment |  -39976.3266|  3849.4880| -10.3848426| 0.0000000|
|Holiday_Flag |   55579.5515| 26990.7698|   2.0592059| 0.0395149|
```

5.3. Model Assumptions and Diagnostics

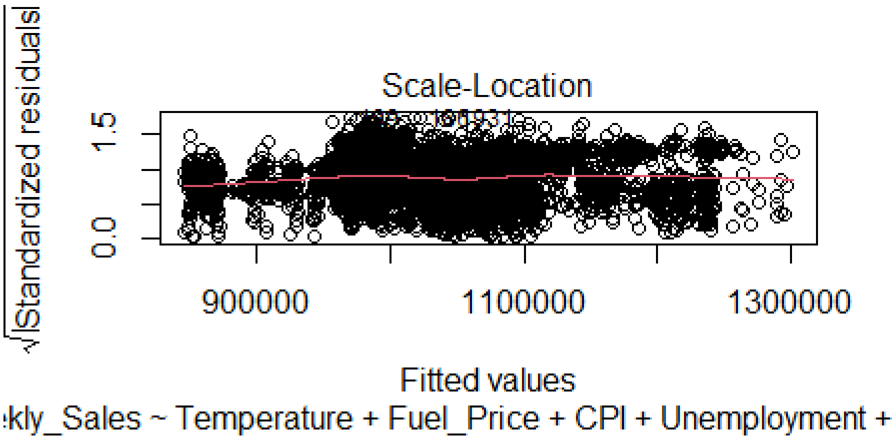
- **Linearity**: Residuals vs. Fitted plot showed no clear pattern, supporting linearity.



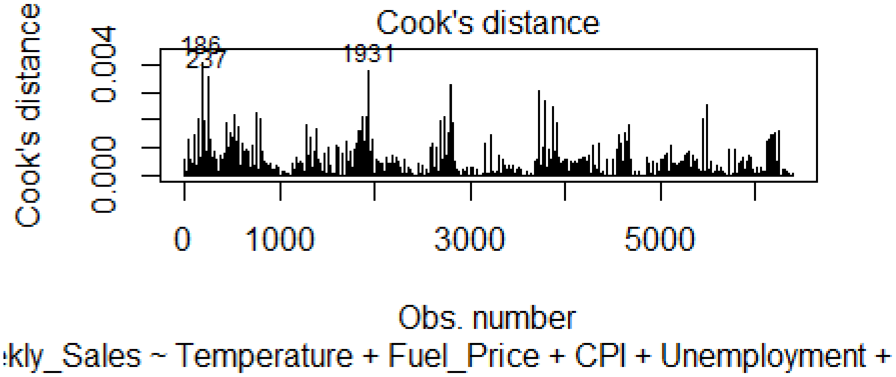
- **Normality**: The Normal Q-Q plot indicated that residuals are approximately normally distributed.



- **Homoscedasticity**: The Scale-Location plot did not show a funnel shape, indicating constant variance.



- **Multicollinearity**: Variance Inflation Factor (VIF) values were all below 1.3, suggesting no significant multicollinearity.



5.4. Model Summary

The regression model explains only about 2.3% of the variability in weekly sales (R-squared = 0.023). The significant variables were CPI, Unemployment, and Holiday_Flag. However, the low R-squared indicates that other factors not included in the model may also be influencing sales.

6. Conclusion and Recommendations

6.1. Key Findings

- Holidays do not have a statistically significant impact on weekly sales.
- Economic indicators such as CPI and Unemployment do significantly affect weekly sales, though the effect size is small.
- Despite identifying significant economic indicators, the model's low R-squared suggests that other factors should be considered for a more comprehensive understanding of sales variability.

6.2. Business Implications

The findings suggest that Walmart's weekly sales are more influenced by broader economic conditions than by holidays. This insight can guide Walmart’s strategic decisions, such as pricing and inventory management during different economic periods.

6.3. Limitations and Future Work

- **Limitations**: The model has a low R-squared, indicating limited explanatory power. The analysis could be improved by including more variables, such as promotional activities or competitor data.

- **Future Work**: Future analysis could explore non-linear models or machine learning algorithms to better capture the complexity of the data.

6.4. Final Recommendation

1. **Focus on Economic Indicators**
   Monitor economic indicators (like CPI and unemployment) to inform adaptive pricing, targeted promotions, and inventory management strategies.

2. **Enhance non-Holiday Sales**
   Develop seasonal campaigns and localized marketing strategies to boost sales during non-holiday periods.

3. **Explore Additional Factors**
   Incorporate more variables (e.g., promotions, weather, competitor pricing) in future analyses and utilize advanced analytics for better modeling.

4. **Invest in Data-Driven Tools**
   Implement real-time monitoring and predictive analytics tools to enable proactive decision-making and strategy adjustments.

5. **Optimize Store Performance**
   Tailor strategies for underperforming stores, establish performance benchmarks, and replicate best practices across the network.

6. **Continuous Improvement**
   Regularly review and refine strategies based on feedback, data, and market conditions, and provide ongoing training for staff.

These steps will help Walmart better align its strategies with economic conditions, optimize store performance, and adapt to changing market dynamics.

7. Supporting Files

The cleaned dataset and regression model summary are available in the GitHub repository [https://github.com/MuhammadAhmadJamil18/B105_Applied-Statistical-Modelling].