Brian Park

Dylan Maray

Hangting Lu

He Li

Walmart Weekly Sales

I. Introduction:

In today's retail economy, understanding the factors that drive weekly sales is essential for retailers seeking to enhance profitability and ensure sustained growth. Walmart, as one of the largest retail chains in the world, experiences regular fluctuations in sales that can be attributed to various factors.

Several elements contribute to the variability in Walmart's weekly sales. Time-specific events, such as holidays or seasonal changes, can lead to significant changes in consumer behavior. For example, sales often increase during holiday seasons, while certain months may see a dip in traffic and purchases. The location of each store also plays a role, as different regions may experience varying demands based on local preferences, climate, purchasing, and more.

External economic conditions are another crucial factor. Weather can influence foot traffic to stores, while rising fuel prices may discourage customers from making frequent trips. Economic indicators like the Consumer Price Index (CPI) and unemployment rates may also impact consumer spending patterns. A higher CPI, signaling inflation, might make consumers more cautious with their purchases, while higher unemployment could reduce income and spending.

By analyzing these factors, retailers like Walmart can gain valuable information about consumer behavior and sales trends. Understanding how different variables interact allows retailers to better demand, optimize inventory, and adapt marketing strategies. This information has an important impact on weekly sales, which is crucial for making informed business decisions and maintaining a competitive edge in the retail business.

The data set we found on Kaggle is Walmart.csv. This data set covers sales from February 5, 2010, to November 1, 2012. Within this file, the following fields are Store (the store number), Date (the week of sales), Weekly_Sales (sales for the given store), Holiday_Flag (whether the week is a special holiday week: 1 – Holiday week, 0 – Non-holiday week), Temperature (Temperature on the day of sale), Fuel_Price (Cost of fuel in the region), CPI (Prevailing consumer price index), and Unemployment (Prevailing unemployment rate).

In this report, we will focus on analyzing the weekly sales data of a randomly selected store (Store 1 in our case). Our primary objective is to understand the relationship between the Consumer Price Index (CPI) and weekly sales performance. The first hypothesis explores whether changes in CPI have an impact on sales. The null hypothesis H_0 suggests that CPI does not influence sales, meaning that fluctuations in CPI have no significant effect on consumer purchasing behavior. The alternative hypothesis H_1, however, suggests that CPI does affect sales, indicating a possible correlation.

The second hypothesis examines whether the relationship between CPI and sales varies depending on the week. Specifically, we want to determine whether the effect of CPI on sales is consistent throughout the year or if it changes during certain periods, such as holidays or specific seasons. The null hypothesis H_0 assumes that there is no interaction between the weeks and CPI, meaning the effect of CPI on sales remains constant over time. The alternative hypothesis H_1 suggests that there is an interaction, indicating that the impact of CPI may differ depending on the time of year, influenced by factors such as seasonal shopping patterns.

II. Literature Review:

In reviewing the code of other analyses conducted on the Walmart dataset, most followed a structured approach that included data exploration, exploratory data analysis (EDA), data preprocessing, data manipulation, feature selection, predictive modeling, and concluding the results. These analyses were typically performed in Python using machine learning models. Our focus, however, is on examining the conclusions drawn from the predictive modeling phase, specifically how the models' outputs were interpreted and applied to understand the best models utilized for this dataset.

In an analysis by Shahjhan Alam, he applied models such as Linear regression, Lasso Regression, Ridge Regression, Polynomial Regression, ElasticNet Regression, Decision tree Regressor, Random Forest Regressor, and XGB Regressor. He calculated the Mean Squared Error, Mean Absolute Error, and R^2 score for each model. After comparing each model, although the XGB Regressor came out on top, Linear Regression was very close considering its high R^2 score even though its MSE and MAE were bigger than the XGB's.

In another analysis by Gurpreet Singh, he built multiple regressions and compared their evaluation metrics to choose the best-fit model for both training and testing sets. The models included Multiple Linear Regression, Ridge Regression, Lasso Regression, Elastic-Net Regression, and Polynomial Regression. He calculated the Mean Squared Error, Root Mean Squared Error, and R^2 score for each model. After comparing each model, he was able to conclude that simple Multiple Linear Regression gave the best results.

In conclusion, while the analyses by Shahjhan Alam and Gurpreet Singh focused on multiple stores and a wider range of variables, our analysis specifically concentrated on a single Walmart store. Despite this narrower focus, their analysis provides crucial information about the effectiveness of different models. Both analyses highlighted that Linear Regression performed well in predicting sales, with Alam noting its competitive R^2 score and Singh finding it to be the best-performing model in his study. Based on these observations, we can confidently conclude that Multiple Linear Regression is a strong model for this dataset, even when applied to a single store.

III. Methods:

After uploading the dataset, we know that each store has a very different group mean. Each different labeled week also has a different group mean. Then, we focus on store 1 weekly sales to avoid multiple categorical variables.

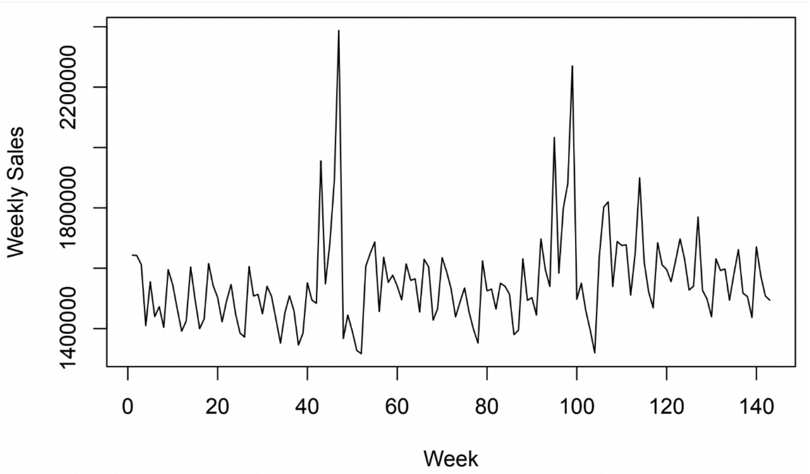


Fig. 1: Weekly Sales vs. Weeks in consecutive three years

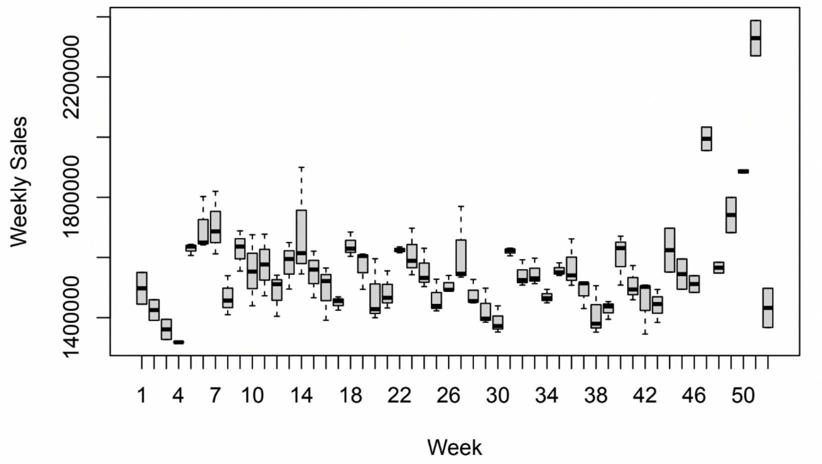
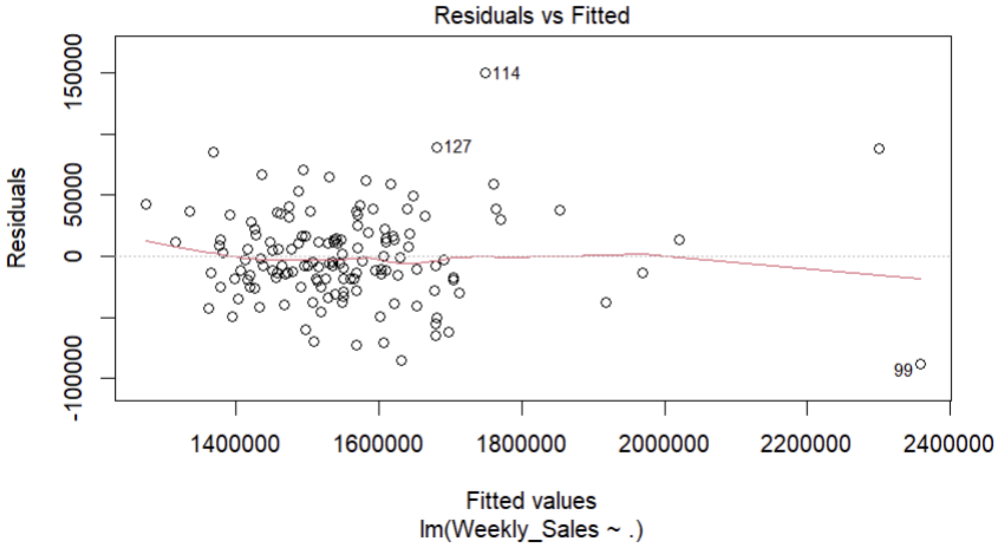
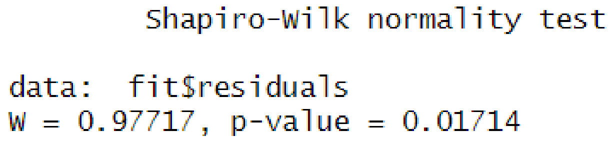


Fig. 2: Weekly Sales Grouped by Weeks

From Figure 1, we can see that the weekly sales of Store 1 show a similar trend each year. From Figure 2, we can observe that the weekly sales are very close across all years, until the weeks from November to December, when sales reach their peak. Therefore, we can organize our dataset by weeks, which will serve as our categorical variable.

Next, we created a data frame for only store 1 with weeks as our categorical variable. Then, we fitted a linear model of Weekly Sales on this data frame to determine whether the 4 assumptions of Linear Regression hold: Homoscedasticity, Linearity, Normality, and Independence.





```
               Shapiro-Wilk normality test

       data:  fit$residuals
       W = 0.97717, p-value = 0.01714
```

Fig. 3: Residuals vs. Weekly Sales Model and Shapiro-Wilk Normality test of Weekly Sales Model

Here, we could see that homoscedasticity and linearity hold but not normality. The W statistic is close to 1. However, the p-value is less than 0.05 level of significance. Then, we can conclude that the residuals are not normally distributed. This conveys that a transformation is needed.

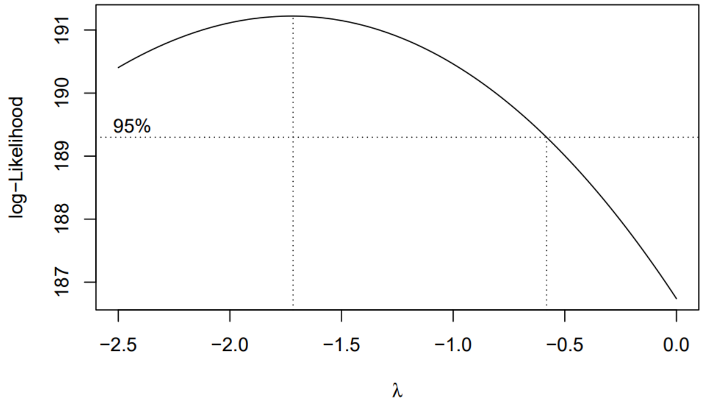


Fig. 4: Box-Cox Transformation

To check the best fit linear model we applied the Box-Cox transformation of our data set. By observing that -2 is within the confidence interval and with a good fit of our transformation, we use -2 as the power of our transformation. Then, we apply this transformation to our model, fit a new linear model, and check for the four assumptions again.

A). Testing the 4 Assumptions of Linear Regression

1. Linearity & Homoscedasticity (from plot, pattern):

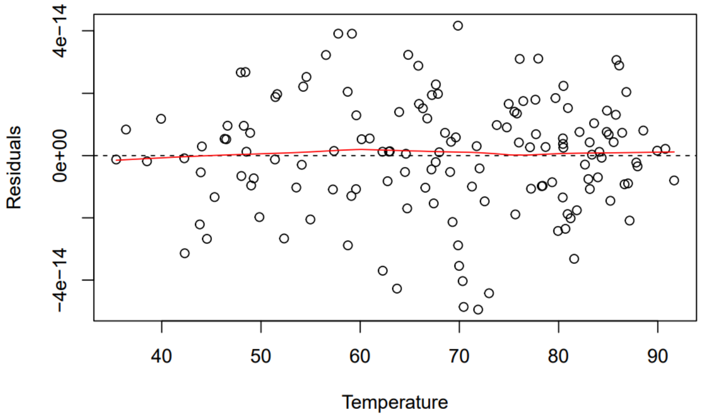


Fig. 5: Residual plot vs. Temperature

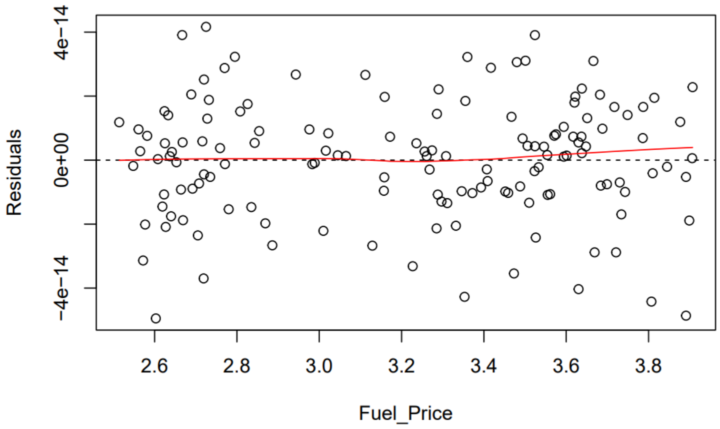


Fig. 6: Residuals vs Fuel Price

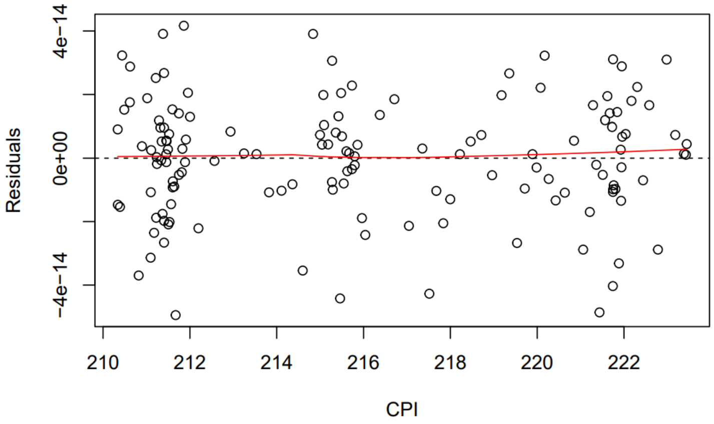


Fig. 7: Residuals vs. CPI

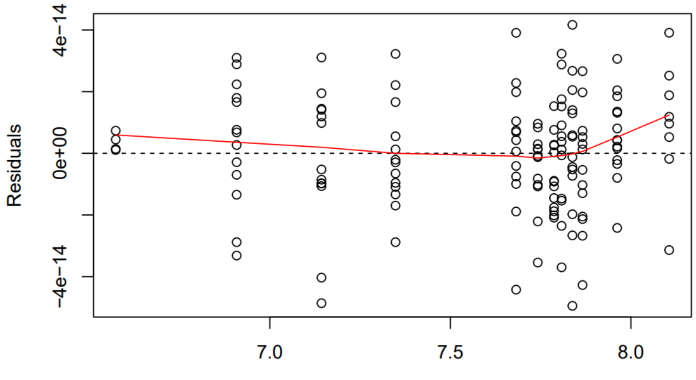


Fig. 8: Residuals vs. Unemployment

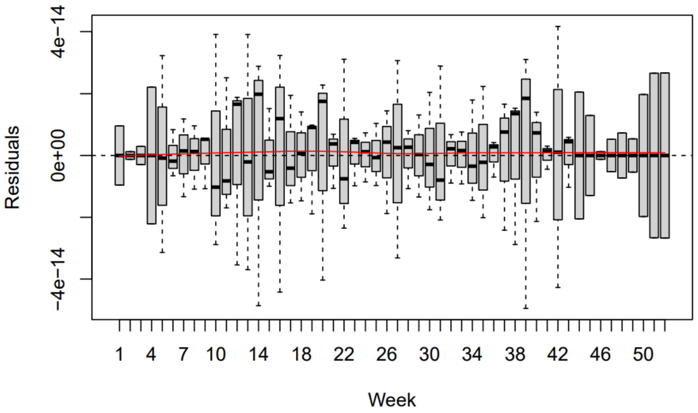
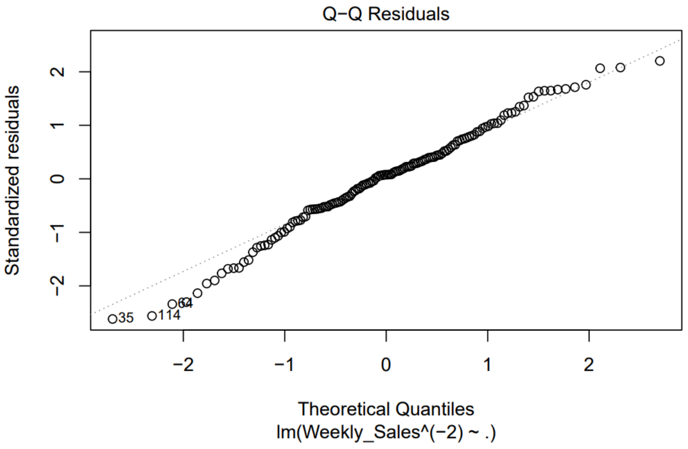
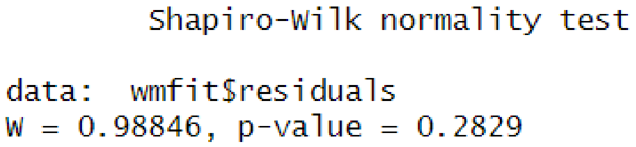


Fig. 9: Residuals vs. Week

Based on these plots, we observe that the range of variances remains nearly constant across all variables, indicating that our model satisfies the homoscedasticity assumption. Additionally, linearity holds for all variables.

2. <u>Normality (Shapiro-Wilk test null hypothesis: Normally Distributed; Q-Q plot):</u>





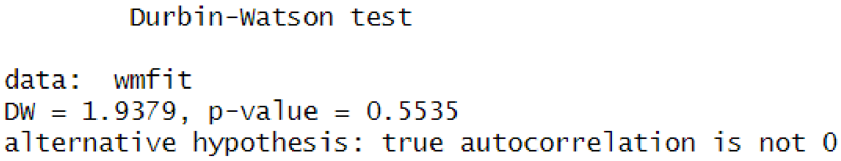
```
                  Shapiro-Wilk normality test

data:  wmfit$residuals
W = 0.98846, p-value = 0.2829
```

Fig. 10: QQ-plot of Weekly Sales after Box-Cox transformation and Sharp-Wilk Normality test of Transformed Weekly Sales Model

From the plot, the pattern of residuals fits the property of the distribution well. We applied the Shapiro-Wilk normality test as well, and the W statistic is close to 1. The p-value is greater than 0.05 level of significance. Then, we can conclude that the residuals are normally distributed.

3. Independence:



```
        Durbin-Watson test

data:  wmfit
DW = 1.9379, p-value = 0.5535
alternative hypothesis: true autocorrelation is not 0
```

Fig. 11: Durbin-Watson Test of Transformed Weekly Sales Model

The DW statistic is around 2, suggesting no autocorrelation. The p-value is much higher than 0.05. Then, we conclude that the residuals are independently distributed.

After reviewing the tests for the four assumptions, we can apply linear regression to the model.

B). Identifying unusual observations

To identify outlier points, we applied a Student's t-test and identified the point with the highest T-value. However, this maximum T-value is less than the extreme T-value allowed for a Type I error, as calculated using the Bonferroni correction. Therefore, there are no outlier points in our model.

To identify leverage points, we compare the diagonal value of hat matrix $H(H_{ii})$ with $\frac{2p}{n}$ (where p is the number of degrees of freedom and n is the size of the data.) Based on this criterion, we found no leverage points.

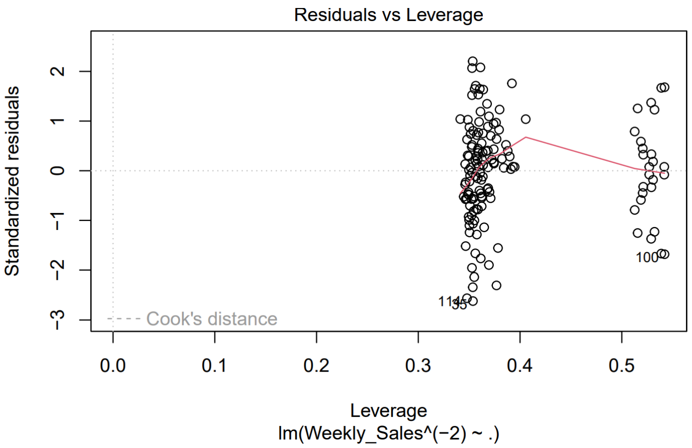


Fig. 12: Residuals vs. Transformed Weekly Sales Model with Cook's distance measurement

We used Cook's distance to identify influential points. After plotting the values, we observed that no points fall beyond the red dashed lines. (The red dashed lines are not visible in the plot as they are far from our data points.) Therefore, our model does not include any influential points.

C). Model selection

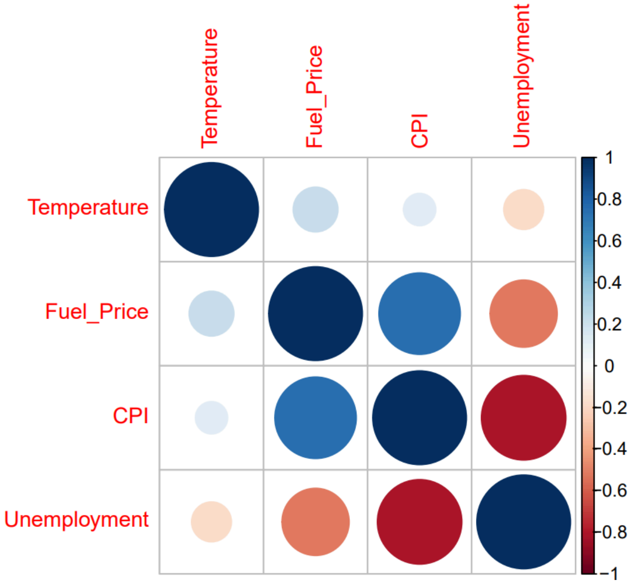


Fig. 13: Correlation Plot

We have to drop some variables to control the model complexity. Then we introduce the Variance-Covariance Matrix to check the correlations. There are correlations among the numerical variables, especially between pairs such as (Fuel_Price, CPI), (Fuel_Price, Unemployment), and (CPI, Unemployment). Observing that Temperature has low correlations with all other variables, we selected it for our model. Among the other three variables, CPI has the lowest correlation with Temperature but a higher correlation with the other two variables, so we chose CPI to represent this group.

We have three choices for the model:

Model 1: Weekly_Sales ~ Week + Temperature

Model 2: Weekly_Sales ~ Week + CPI

Model 3: Weekly_Sales ~ Week + CPI + Temperature

BIC penalizes the model complexity. We apply this function to find the best-fitted model. The output values for models are -8224.578, -8368.978, and -8364.743, respectively. Model 2 has the lowest BIC value, hence we chose it as our linear model.

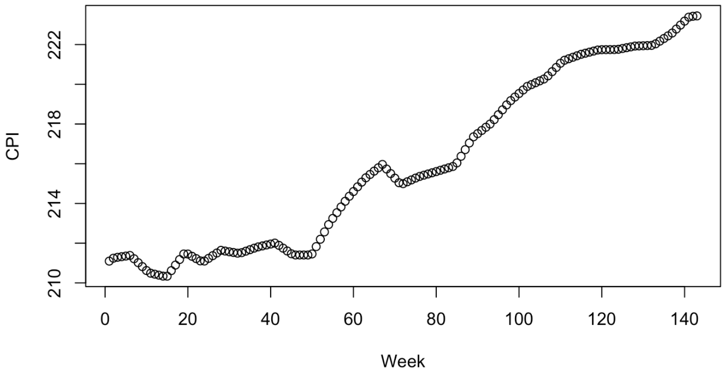


Fig. 14: CPI vs. Week

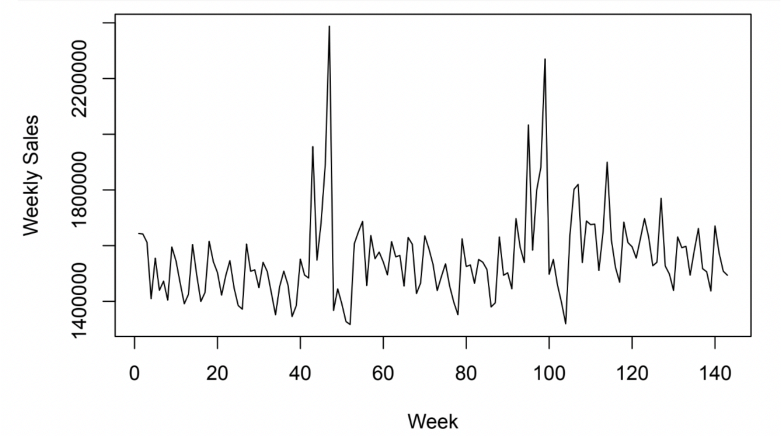


Fig. 15: Weekly Sales vs. Week

The CPI shows an increasing trend over time. But, the Weekly Sales does not present a certain upward trend as time. Then we want to know if it is reasonable to include CPI as a variable in our model.

IV: Hypothesis

A). First hypothesis: Should we consider CPI for analyzing weekly sales?

We use the nested model comparison method at this point to check whether there is a significant difference if we use CPI as the independent variable.

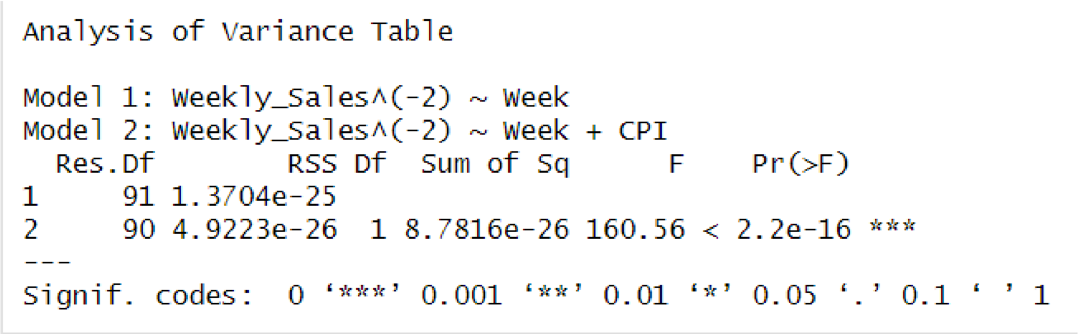
Null hypothesis: Weekly sales are not affected by CPI.

$\mathbf{M}_1$: $Weekly\ Sales^{(-2)} \sim Week$

Alternative hypothesis: Weekly sales are affected by CPI.

$\mathbf{M}_{1+2}$: $Weekly\ Sales^{(-2)} \sim Week + CPI$

Result:



```
Analysis of Variance Table

Model 1: Weekly_Sales^(-2) ~ Week
Model 2: Weekly_Sales^(-2) ~ Week + CPI
  Res.Df        RSS Df  Sum of Sq      F    Pr(>F)    
1     91 1.3704e-25                                   
2     90 4.9223e-26  1 8.7816e-26 160.56 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Fig. 16: ANOVA table for $\boldsymbol{M}_1$ and $\boldsymbol{M}_{1+2}$ for First Hypothesis

According to the ANOVA table, the p-value is less than 0.05 level of significance. Also, the F statistic is greater than $F_{0.05,\,1,\,90}$. Then, we reject the null hypothesis, concluding that the effect of CPI is significant. Therefore, CPI should be included as a variable.

Then, for further improvement of the model, we care about the interaction between CPI and time (Week), i.e. if the regression equations for each different week are parallel.

B). Second Hypothesis: Is CPI affected by each different week (week i)? (i = 1, 2, ..., 52)

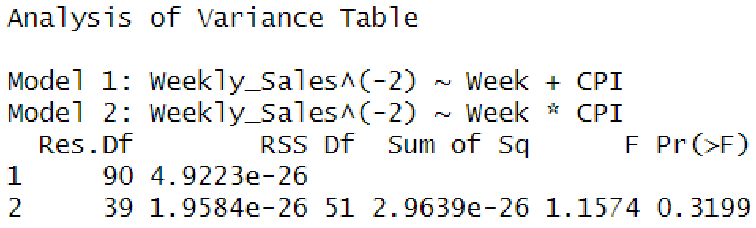
Null hypothesis: There is no interaction between week i and CPI. (coefficient of week i: CPI is not significant)

$\mathbf{M_1}$: $Weekly\ Sales^{(-2)} \sim Week + CPI$

Alternative hypothesis: There is an interaction between week_i and CPI.

$\mathbf{M_{1+2}}$: $Weekly\ Sales^{(-2)} \sim Week + CPI + Week{:}CPI$

Result:



```
Analysis of Variance Table

Model 1: Weekly_Sales^(-2) ~ Week + CPI
Model 2: Weekly_Sales^(-2) ~ Week * CPI
  Res.Df        RSS Df  Sum of Sq      F Pr(>F)
1     90 4.9223e-26
2     39 1.9584e-26 51 2.9639e-26 1.1574 0.3199
```

Fig. 17: ANOVA Table for $\boldsymbol{M_1}$ *and* $\boldsymbol{M_{1+2}}$ *for Second Hypothesis*

According to the ANOVA table, since the p-value is greater than 0.05, we are not able to reject the null hypothesis. Hence, the interaction between week i and CPI is not significant.

V: Conclusion

In conclusion, the analysis of Walmart's weekly sales data reveals that the Consumer Price Index (CPI) plays a significant role in influencing sales. The results of the first hypothesis test using nested model comparison showed that the inclusion of CPI significantly improves the model's performance, with a p-value below 0.05 and an F statistic exceeding the critical value, allowing us to reject the null hypothesis. This indicates that CPI has a measurable impact on weekly sales and should be considered in models aimed at forecasting sales.

Furthermore, the results from the second hypothesis test suggest that the effect of CPI on sales remains consistent over time. The interaction term between CPI and specific weeks was not significant, as the p-value exceeded the 0.05 threshold. This implies that while CPI affects weekly sales overall, its impact does not vary meaningfully across different weeks or seasons. Therefore, seasonal changes or specific weeks do not alter the relationship between CPI and sales in this dataset.

Overall, these findings indicate that CPI is a valuable predictor of Walmart's weekly sales, though its effect is stable across the year. Incorporating CPI in sales prediction models can help Walmart and similar retailers better understand how economic conditions impact consumer spending, leading to more accurate demand predictions and inventory planning.

Works Cited

GurpreetSingh2512. 🌟🚀 *Walmart Prediction*🌟🚀, www.kaggle.com/code/bunny11/walmart-prediction/notebook.

H, M Yasser. “Walmart Dataset.” *Kaggle*, 26 Dec. 2021, www.kaggle.com/datasets/yasserh/walmart-dataset/code

Shahjhanalam. “🌟Walmart Prediction🌟: Various ML Model: (94%).” *Kaggle*, Kaggle, 10 Aug. 2024, www.kaggle.com/code/shahjhanalam/walmart-prediction-various-ml-model-94