Retail Analytics: Uncovering Sales Drivers in Walmart's Weekly Data

Brian Park

Dylan Maray

Hangting Lu

He Li

Overview

Why is this important?

Understanding factors that drive weekly sales is an important key point for retailers to maintain and enhance profitability and ensure sustained growth. Without consistent weekly sales to sustain cash flow, business expenses (i.e employee wages, inventory, rent, etc.) build up which can lead to poor financial health of the company.

What can such information do for the company?

By understanding factors that drive weekly sales, companies would be able to take various actions and strategies (i.e create new marketing strategies, improve inventory management, supply-chain optimization, choosing suppliers, etc.) to prevent poor financial health of the company.

Factors that Impact Weekly Sales

1). Store:	1). Store can have different average weekly sales compared with others
2). Time:	2). Weekly sales fluctuate as time. Around Holidays or special weeks, people are more likely to purchase items for their family/friends
3). Temperature:	3). Temperature may impact the consumer's ability to travel. Extreme weather conditions such as blizzards, cold snaps, and heat waves can keep consumers away from the store.
4). Consumer Price Index (CPI):	4). CPI determine a consumer's purchasing power. A high CPI indicates that consumers have less money to spend.
5). Unemployment Rate:	5). Unemployment rate can show the consumer's' ability to spend money. High unemployment can indicate that more people have less of an ability to spend money.
6). Fuel Price	6). Fuel Price can impact the consumer's ability to travel to the store as well as the delivery vehicle's ability to bring goods from suppliers to stores or from stores to consumers.

Introduction

Objective:

The purpose of this analysis is to understand the relationship between chosen independent variables (Consumer Price Index & Week) and Weekly Sales.

First Hypothesis:

H_0: CPI does not influence sales.

H_1: CPI does affect sales.

Second Hypothesis:

H_0: There is no interaction between the weeks and CPI.

H_1: There is an interaction between the weeks and CPI.

- If the H_1 of Hypothesis 1 is true, then CPI does affect sales, indicating a possible correlation between the two variables.
- If the H_1 of Hypothesis 2 is true, 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.

Additional Notes:

- To avoid multiple categorical variables, all stores assigned number 1 in the dataset were chosen candidates for the study.

Methods

- Figure 1, we can see that the weekly sales of Store 1 show a similar trend each year.
- 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.
- After observing Figure 1 and Figure 2, we see that we can organize our dataset by weeks which will serve as our categorical variable.

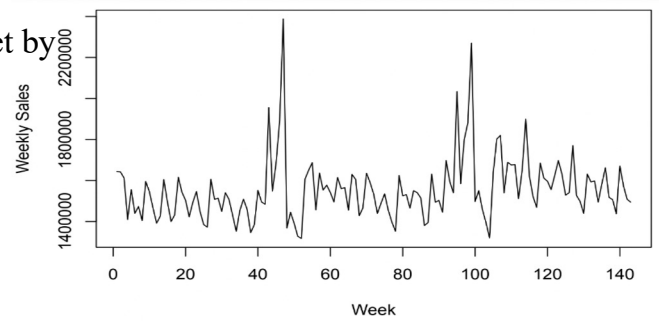


Fig. 1: Weekly Sales vs. Weeks in consecutive three years

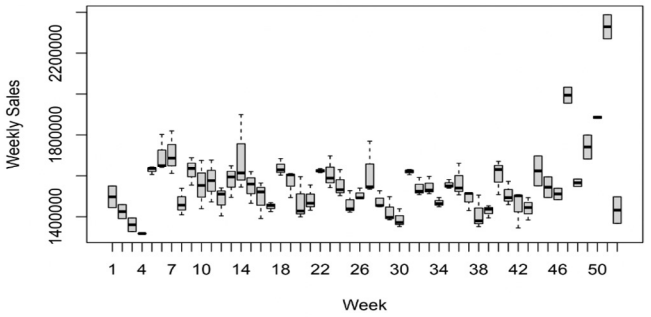
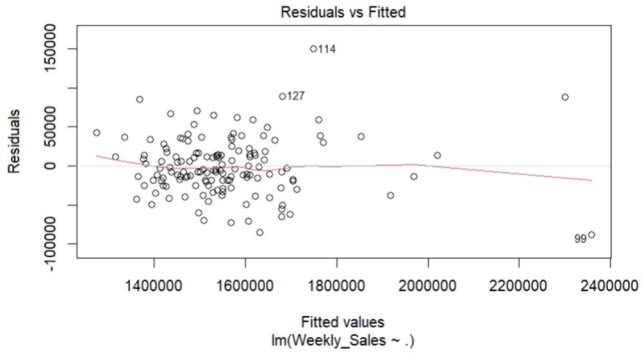


Fig. 2: Weekly Sales Grouped by Weeks




```
        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

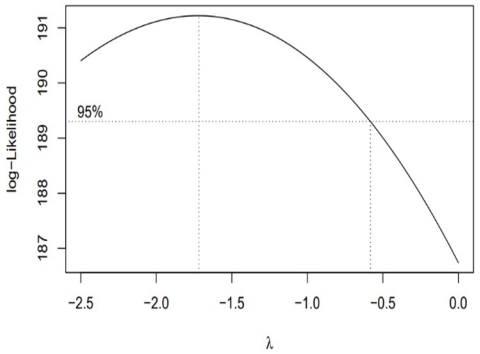


Fig. 4: Box-Cox Transformation

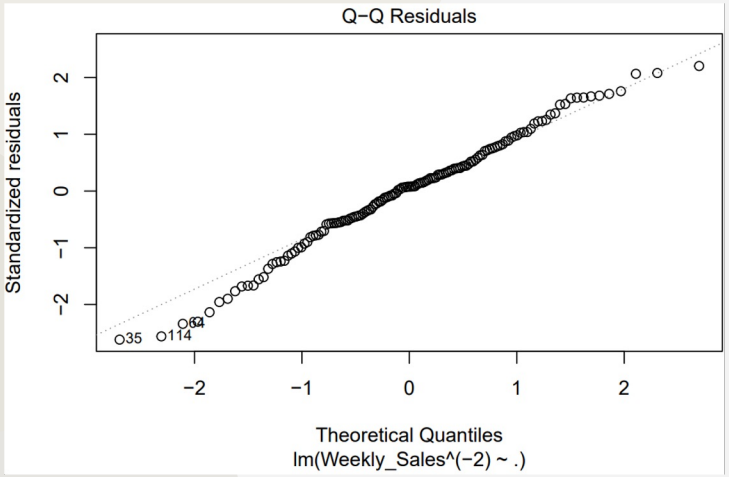
Figure 3:

- From the Residuals vs. Weekly Sales plot, we see that Homoscedasticity and Linearity hold for the Weekly Sales Model.
- The W-value of the Shapiro-Wilk Test is close to 1, but the p-value of 0.01714 is than $\alpha = 0.05$, meaning that the data is not-normal. To normalize the data, apply the Box-Cox Transformation.

Figure 4:

- Observe that -2 is within the 95% confidence interval indicating a good fit of our transformation.
- Apply -2 as the transformation for our weekly sales model.

Testing the 4 Assumptions of Linear Regression for our Transformed Model:




```
        Shapiro-Wilk normality test

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

Normality:

- Observe that the Q-Q plot of the transformed model is linear.
- The W-value of the Shapiro-Wilk Test is close to 1 and the p-value of 0.2829 is greater than $\alpha = 0.05$.
- Based on the Q-Q plot and the Shapiro-Wilk Test, we can conclude that the transformation applied to the model normalized the data.

Independence:

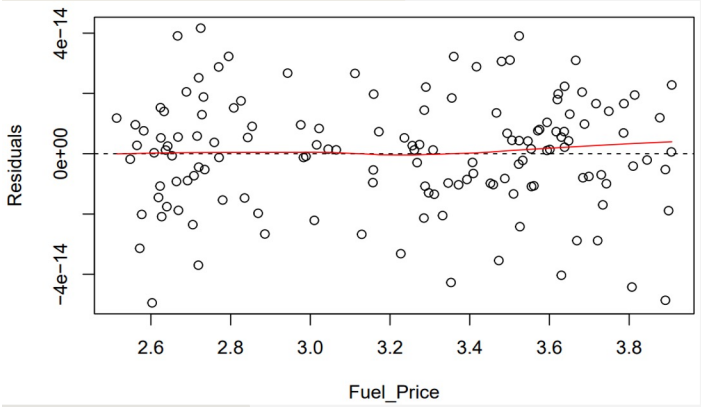
The Durbin-Watson statistic is around 2, suggesting no autocorrelation. The p-value is much higher than 0.05. Then, we conclude that the residuals are independently distributed.

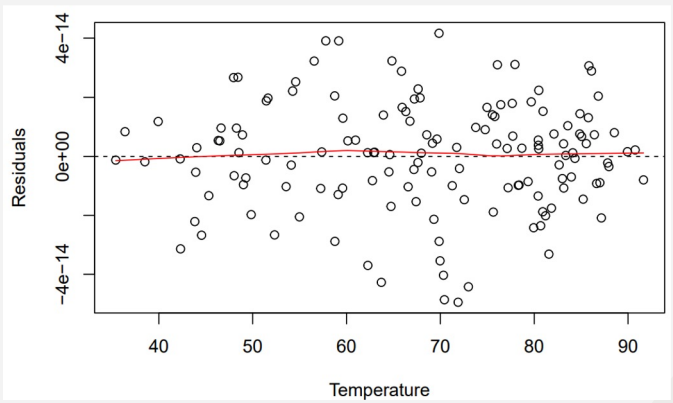
```
        Durbin-Watson test

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

Linearity & Homoscedasticity:

- Observe that the range of variances remains nearly constant across all independent variables (Fuel Price, Temperature, CPI, Unemployment Rate, Week), indicating that our model satisfies the homoscedasticity assumption. Additionally, linearity holds for all variables.

Now that the transformed model satisfies the 4 Assumptions, we can apply linear regression.





From top to bottom, left to right:

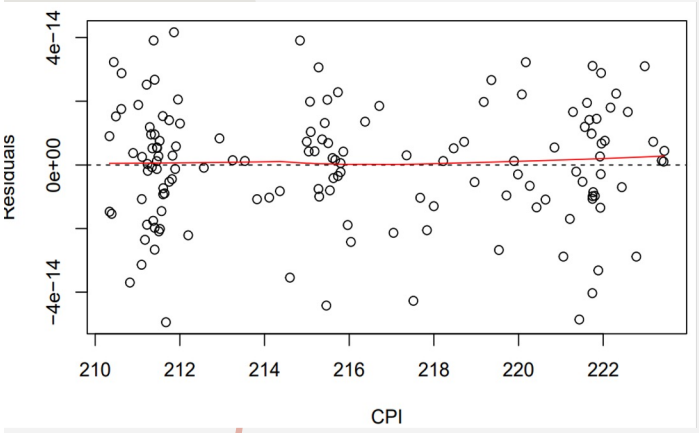
Residual vs. Fuel Price,

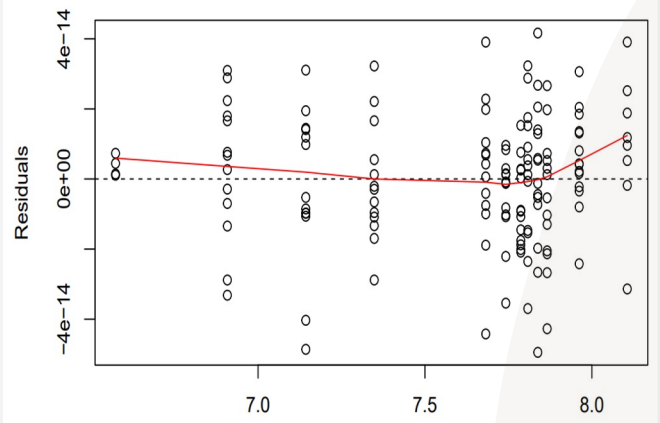
Residuals vs Temperature ,

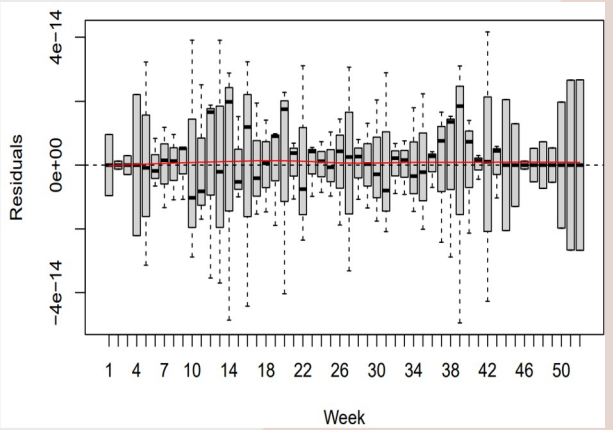
Residuals vs CPI,

Residuals vs Unemployment Rate,

Residuals vs Week

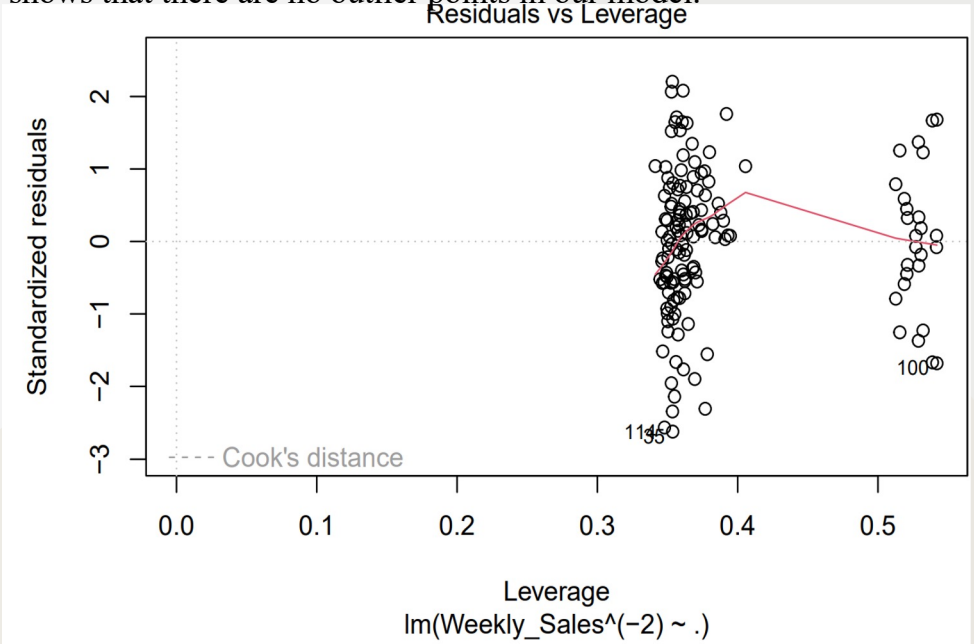






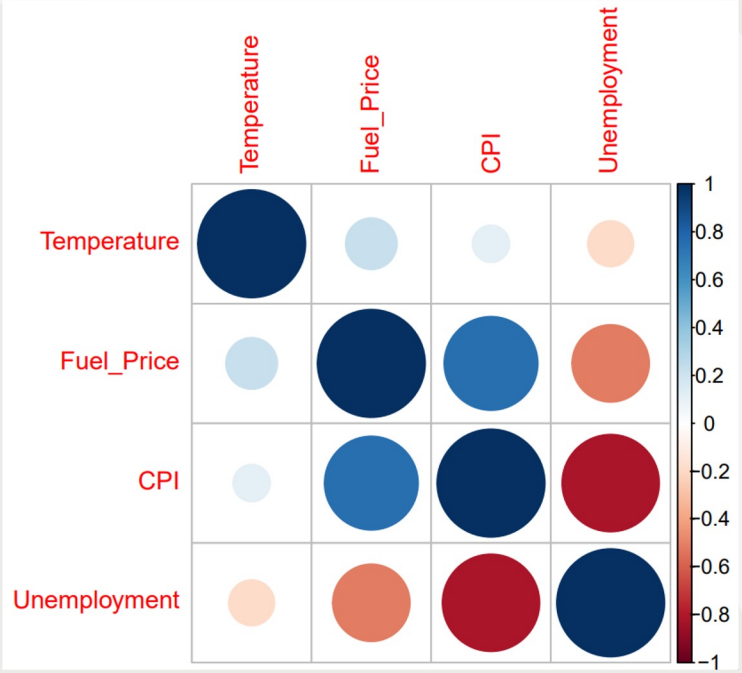
Identifying unusual observations:

- We applied a Student's t-test in R to identify outlier points .
- The point with the highest T-value was identified. However, this maximum T-value is less than the extreme T-value allowed for a Type I error, as calculated using the Bonferroni correction. This shows that there are no outlier points in our model.
- To identify leverage points in our data, we compared the diagonal values of hat matrix H(H_{ii}) with 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.
- 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.



Model Selection

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.



Model Choices:

Model 1: Weekly_Sales ~ Week + Temperature

Model 2: Weekly_Sales ~ Week + CPI

Model 3: Weekly_Sales ~ Week + CPI + Temperature

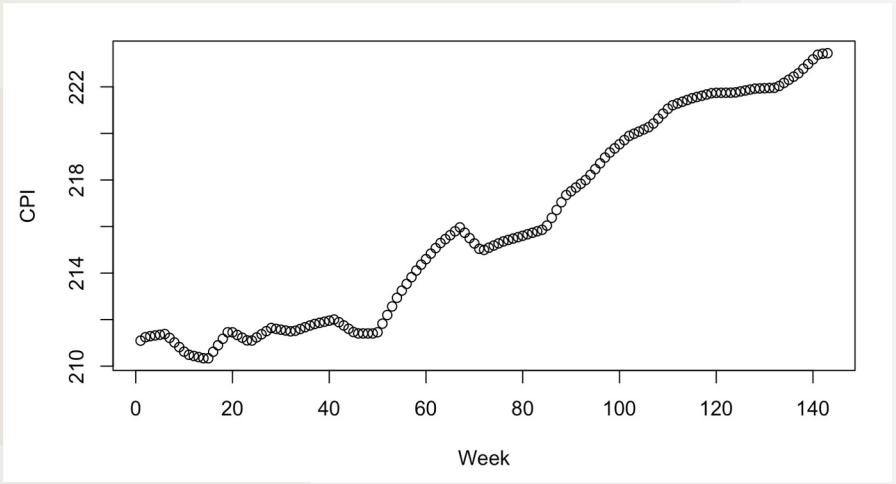
```
wmfit_1 = lm(Weekly_Sales^(-2) ~ Week + Temperature, data = walmart_s1t)
wmfit_2 = lm(Weekly_Sales^(-2) ~ Week + CPI, data = walmart_s1t)
wmfit_3 = lm(Weekly_Sales^(-2) ~ Week + Temperature + CPI, data = walmart_s1t)
BIC(wmfit_1); BIC(wmfit_2); BIC(wmfit_3)
```

```
## [1] -8224.578
## [1] -8368.978
## [1] -8364.743
```

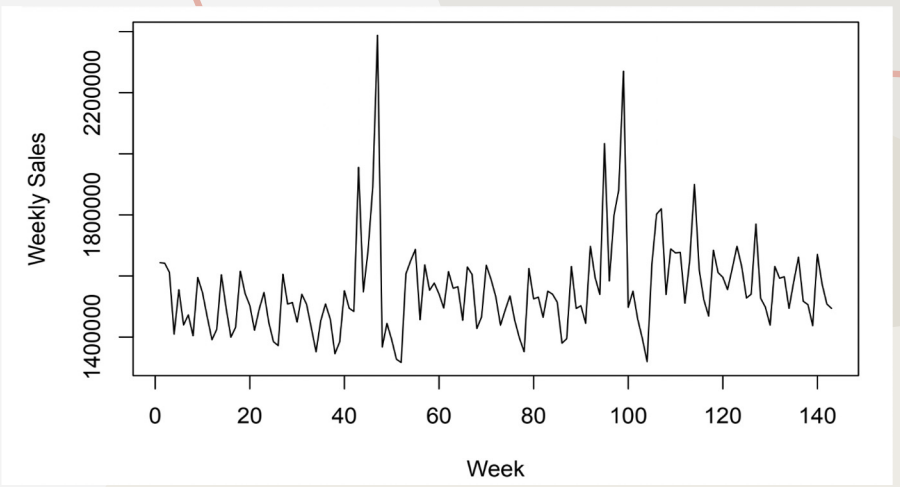
We choose the second model (with the lowest BIC value).

Applying 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.**



CPI vs. Week



Weekly Sales vs. Week

From the given plots generated in R, 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.

Testing First Hypothesis

First Hypothesis: Should we consider CPI for analyzing weekly sales?

H_0: Weekly sales are not affected by CPI.

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

H_1: Weekly sales are affected by CPI.

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

```
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
```

Result: 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.

Testing Second Hypothesis

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

***H_0*:** There is no interaction between week *i* and CPI. (coefficient of week i: CPI is not significant)

M_1: *Weekly Sales$^{(-2)}$ ~ Week + CPI*

***H_1*:** There is an interaction between week *i* and CPI.

M_{1+2}: *Weekly Sales$^{(-2)}$ ~ Week + CPI + Week::CPI*

```
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
```

Result: 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.

Conclusion

First Hypothesis:

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 shows that CPI has a measurable impact on weekly sales and should be considered in models aimed at forecasting sales. Overall, CPI plays a significant role in influencing sales.

Second Hypothesis:

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 suggesting the effect of CPI on sales remains consistent over time.