Final Report

Members: Mu Ye Liu, Siyi Man, Borry Yu, Nikita Prabhu
Group P5
Student handing in R code and data set on Canvas: Mu Ye Liu

Introduction

We obtained the dataset from kaggle (https://www.kaggle.com/datasets/mikhail1681/walmart-sales?resource=download) and this dataset contains information on 45 Walmart stores across multiple regions.

Variable	Description	Variable Type	Unit
Weekly_Sales (response variable)	Weekly revenue generated from sales through the week	Continuous	USD
Holiday_Flag	Indicator for holiday week (1) or non-holiday week (0)	Categorical	/
Temperature	Temperature in the region of the store	Continuous	Degree Fahrenheit
Fuel_Price	Average cost of fuel in the store's region	Continuous	USD/gallon
CPI	Average Consumer Price Index throughout the week	Continuous	/
Unemployment	Average Unemployment rate throughout the week	Continuous	On a percentage scale

Description of Variables:

Weekly_Sales: The Weekly Revenue generated from sales through the week in USD.

- Holiday_Flag: Categorical dummy variable encoded as: 0 if no holiday present anytime within the week: 1 if there is at least one holiday present.
- Temperature: Average air temperature in the store's region throughout the week in degrees fahrenheit. For our analysis, it is converted into degrees celsius.
- Fuel_Price: Average cost of fuel in the store's region USD per gallon throughout the week.
- CPI: The average Consumer Price Index throughout the week.

Motivation:

Analyzing this dataset can reveal insights into the factors influencing retail sales at Walmart, such as holidays, weather, and broader economic conditions. Understanding these relationships is valuable for

developing data-driven strategies to optimize inventory, staffing, and marketing efforts. Insights from this dataset may also provide predictive value to other retail businesses seeking to enhance their sales forecasting and adapt to economic fluctuations. This dataset was scraped from the official website of Amazon through BeautifulSoup and WebDriver using Python.

Analysis:

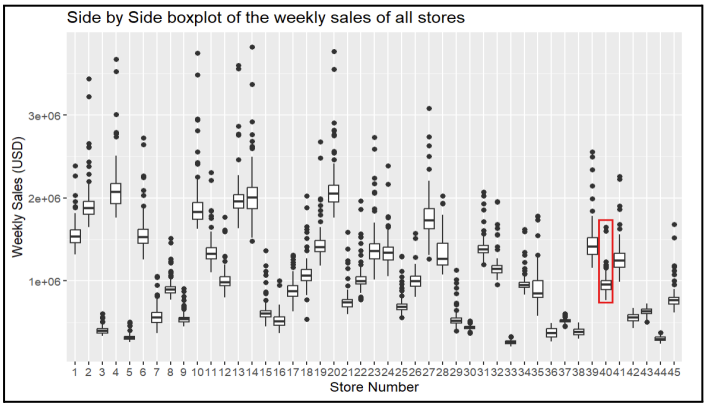
Data pre-processing:

Since the original dataset comprises 6,435 observations from 45 stores, conducting regression analysis on the entire dataset would violate the assumption of independence (i.e., observations are independent of one another). This is due to the fact that weekly sales within the same store are highly dependent, as they are influenced by shared factors such as location, recurring customer behavior patterns, local economic conditions, and store-specific operational strategies.

Therefore, we decide to choose one store as a representative of all the stores, which can reflect the characteristics of all stores.

To select a store that represents the overall characteristics of all stores effectively, we identify the store with the median rank based on a combination of its average weekly sales and variance in weekly sales. The calculations were performed using a series of steps. First, summary statistics, including the mean and variance of weekly sales, were computed for each store. The sum of these two ranks was calculated to determine each store's combined rank. Subsequently, the data was sorted based on the combined rank, and the store corresponding to the median rank was identified as the most representative.

In this analysis, store 40 was selected as the representative store. A boxplot of weekly sales across all stores confirmed that store 40's sales patterns are reasonable and reflective of the general trends observed in the dataset, since the mean and variance of store 40's weekly sales seem to be fairly average across all stores. Our store selection process ranks the stores by weekly sales, as well as its variance of weekly sales, and takes the sum of the ranks. The store with the median sum of ranks (store 40) is selected to increase the likelihood of the selected store being able to well represent the key statistical characteristics of the entire dataset.

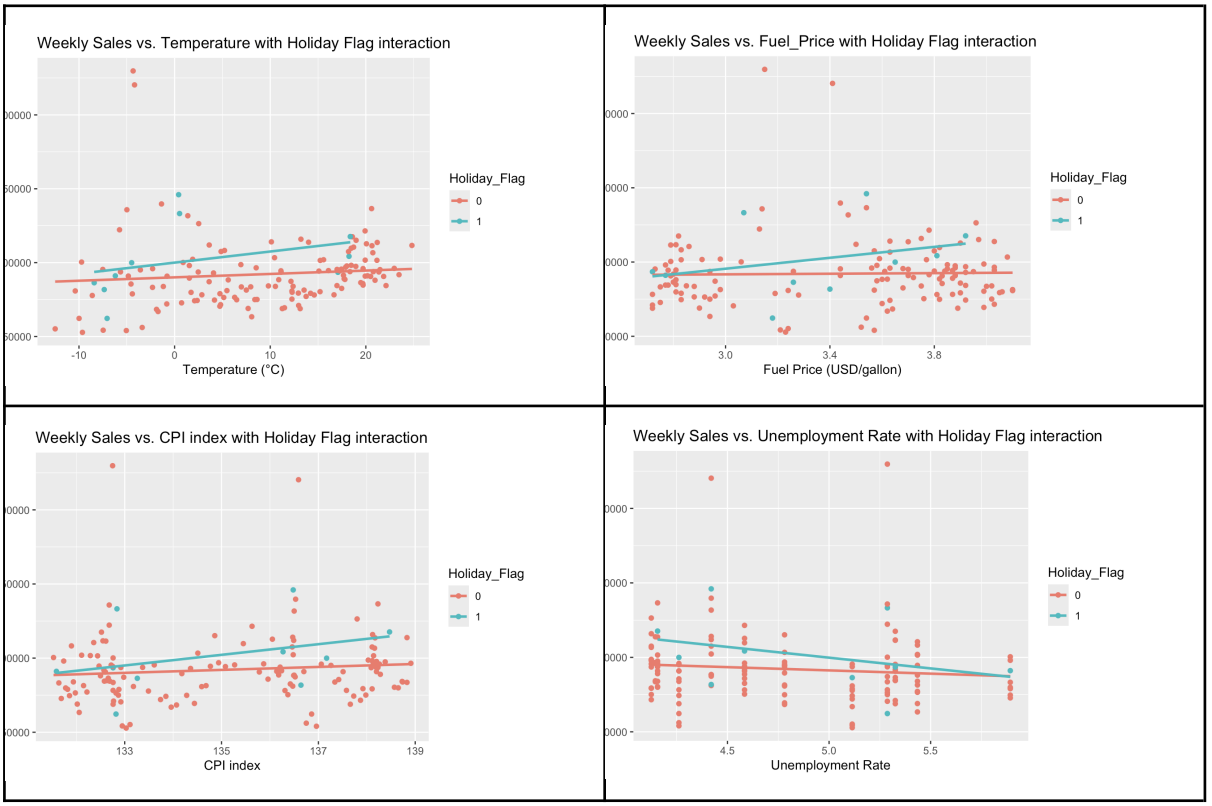


Also, 2 outlier points (data points that have a weekly sales above $1,500,000) were removed to avoid a biased model and have a more accurate model selection (performed below). This is because the 2 outliers will increase the standard errors of our estimated coefficients, thus increasing the p-value, and the likelihood that the covariate is removed during the model selection process (Type 2 error where the null hypothesis of the covariate having slope = 0 is failed to be rejected).

Exploratory Data Analysis:

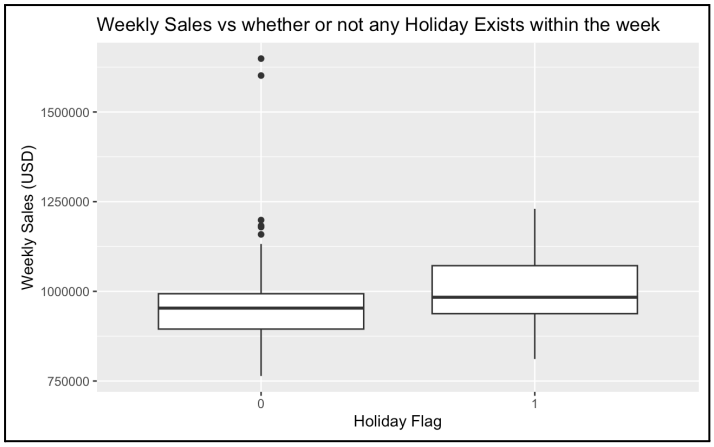
Next, we conduct data visualization of the response variable and covariates to do preliminary data analysis before using the regression model.

First, 4 scatter plots **Weekly Sales vs. the four continuous variables** (CPI, Unemployment, Fuel Price, and Temperature)**, categorized by Holiday_Flag** were created to preliminarily observe how each covariate (along with holiday flag interactions) affects the weekly sales.

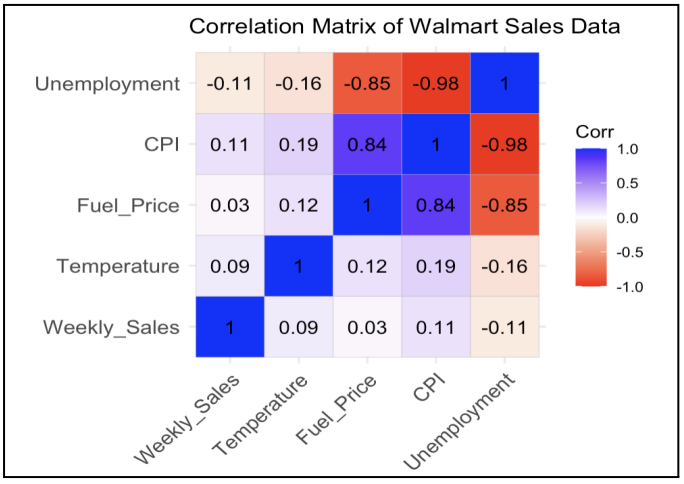


Based on early observations from the four plots above, we can initially draw a preliminary conclusion that interaction terms between the four continuous variables (CPI, Unemployment, Fuel Price, and Temperature) and the *Holiday_Flag* should be incorporated into our model. This is suggested by the differing slopes of the response variable (*y*) against these predictors when comparing holiday (yes) and non-holiday (no) conditions. However, as the differences in slopes are not particularly obvious and the number of data points for holidays is significantly smaller than for non-holidays, this conclusion is tentative and requires further validation. Additional analysis (backward selection) will be conducted later in the modeling process to confirm whether to include these interaction terms or not.

Moreover, we drew the boxplot of weekly sales, categorized by holiday flag (shown below), from which we can observe that there is significant difference of sales between weeks that have a holiday and weeks that do not have a holiday within. Thus, the variable Holiday Flag should be included in our model.



Furthermore, to assess the relationships between the different features in the dataset, a correlation matrix was computed for the continuous variables (**Temperature**, **Fuel Price**, **CPI**, **Unemployment rate** and **Weekly Sales)**. The correlation matrix was then visualized using a heatmap, where positive correlations are represented by cool colors (blues), while negative correlations are shown in warm colors (reds). Variables with strong correlations (close to 1 or -1) are indicated by more intense colors, while weak or no correlations (around 0) are shown by lighter shades.



From the heatmap, we observe that the response variable has positive relationship with Temperature, Fuel Price, CPI, and negative relationship with Unemployment rate. Further, Unemployment rate has strong correlation with Fuel Price and CPI, which may impose the risk of multicollinearity in our model if all these three variables are included. Therefore, a multicollinearity check is done below using the VIF.

Multicollinearity

First, the VIF is computed in the table below for all covariates together.

Covariate	CPI	Fuel Price	Temperature	Unemployment	Holiday Flag
VIF	23.584245	3.660348	1.094752	25.347647	1.046306

From the result, we can see that CPI and Unemployment are highly correlated (Both with VIF > 10), which would severely increase the sensitivity and error proneness of our fitted mosel. Based on the above heatmap, the correlation between weekly sales and CPI vs Unemployment is approximately the same, but based on the

scatterplots, Unemployment exhibits a slightly steeper slope, making it more plausible to keep Unemployment and remove CPI.

To follow up, the CPI was removed and the VIF was recalculated for the remaining variables, indicated in the table below. Now, the VIF for all variables is well below 10, indicating a low possibility for multicollinearity, and is appropriate to proceed with model selection.

Covariate	Fuel Price	Temperature	Unemployment	Holiday Flag
VIF	3.653674	1.063161	3.688905	1.040579

Model Selection

A backwards selection algorithm was used to select our model. A simpler and interpretable model is ideal. We began with the full model containing all the covariates and all the interaction terms of the variables with Holiday Flag. So we will have 7 covariates (or 8 beta parameters) in total: Temperature, Fuel_Price, Unemployment, Holiday_Flag, and the following interaction terms: Temperature:Holiday_Flag, Fuel_Price:Holiday_Flag, and Unemployment:Holiday_Flag. It was performed at a 15% significance level because it even kept variables that only slightly contribute to the model, reducing the probability of Type II error, where covariates that actually contribute to the model get removed. Also, a 15% significance level lowers the chance that all covariates get removed, and the final model being the null model. The parameter with a p-value higher than the threshold will be dropped. We then refitted the model using the remaining variables till we obtained a model with all statistically significant covariates.

As mentioned above in the data pre-processing, two outlier data points were firstly removed to ensure a more accurate, less biased model selection process.

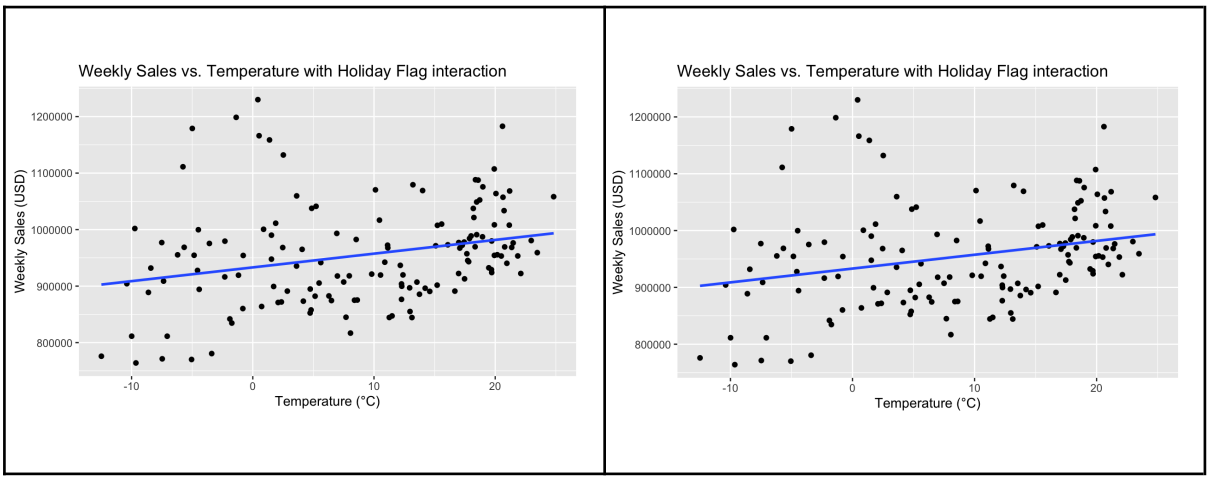
Variables removed in the Backward selection process

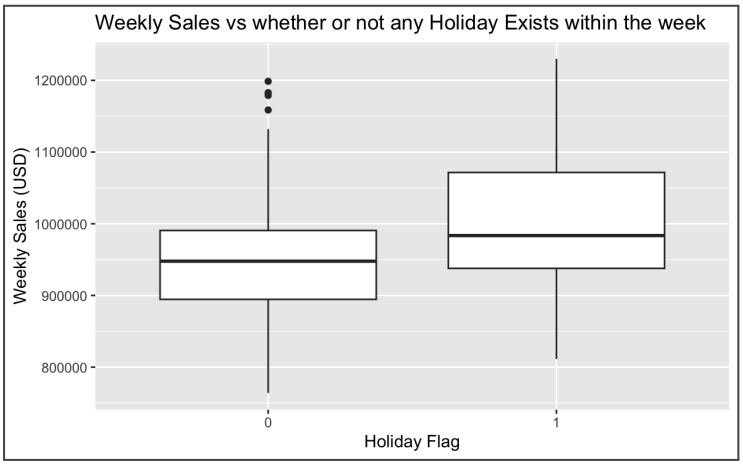
Step	**Variable Removed**	**P-value**
1	Fuel_Price:Holiday_Flag	0.931
2	Temperature:Holiday_Flag	0.850
3	Unemployment:Holiday_Flag	0.458
4	Fuel_Price	0.244

Our final model has only three variables- Temperature, Unemployment, and Holiday_Flag (without any interactions). These three variables are shown to be statistically significant, with p-values less than 0.15.

To verify our intuition from the preliminary observations in the scatterplots above, nearly identical scatterplots for the continuous covariates (Temperature and Unemployment), and a boxplot for the categorical variable (holiday flag) were recreated to visually verify whether or not the covariates we expected to be kept from the preliminary analysis were kept in the final model, this is true.

Replotted the scatterplots below to verify intuition:





From the recreated plots, we see that all covariates kept in the final model showed correlations in the scatter plots and boxplots respectively, successfully verifying intuition from the preliminary observations above.

Summary Statistics for the final model with 3 covariates without interactions:

```
Call:
lm(formula = Weekly_Sales ~ Temperature + Unemployment + Holiday_Flag,
    data = store_40)

Residuals:
    Min      1Q  Median      3Q     Max
-163807  -53483  -13022   46399  276879

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   1026548.8    65088.2  15.772  < 2e-16 ***
Temperature      2642.8      752.8   3.511 0.000606 ***
Unemployment   -21033.5    13194.0  -1.594 0.113201
Holiday_Flag1   78438.3    28163.9   2.785 0.006110 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 84230 on 137 degrees of freedom
Multiple R-squared:  0.1347,    Adjusted R-squared:  0.1158
F-statistic:  7.11 on 3 and 137 DF,  p-value: 0.0001785
```

The formula for our fitted model is: **y_hat = 1026548.8 + 2642.8*x1 - 21033.5*x2 + 78438.3*z**, where **y_hat** = fitted weekly sales, **x1** = Temperature, **x2** = Unemployment, **z** = Holiday flag dummy var.

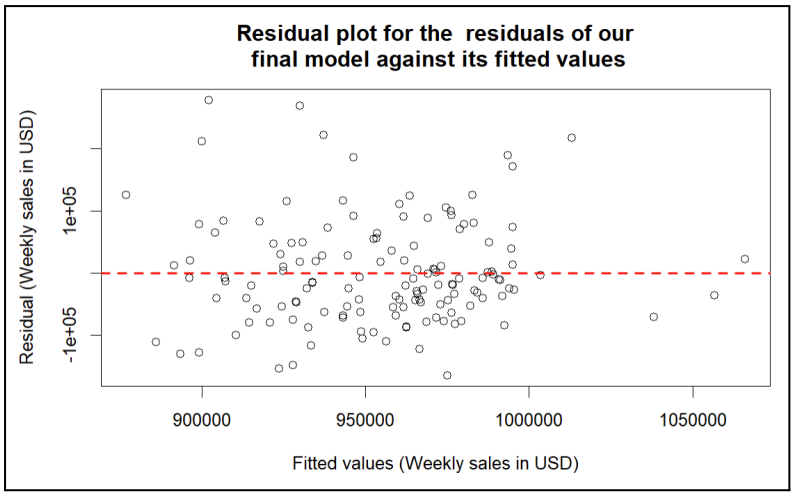
Based on the fitted model, in degrees Celsius increase in temperature, the average weekly sales is expected to increase by $2642.80 while holding other covariates constant. For each % increase Unemployment rate, the average weekly sales is expected to decrease by $21033.50 while holding other covariates constant. For a week that contains a holiday, the weekly sales are expected to be on average $78438.30 higher than a week with no holidays.

To examine the adequacy of the final model, we compare the R-squared and Adjusted R-squared values between the full model and the final simpler model.

Model	**R-squared**	**Adjusted R-squared**
full model	0.147	0.089
final simpler model	0.135	0.116

We observe that the R-squared and adjusted R-squared are low. R-squared is 0.135 for the simpler model. This means only 13.5% of the variance in the response variable, weekly sales, is explained by our model. For the full model, the model explains 14.7% of the variation in the response variable, weekly sales, which is normal because the R^2 never decreases as more covariates are added. However, given that the final model with less covariates explains most of the variation in the response, and has a higher adjusted R^2 compared to the full model, it is obviously desired over the full model because it is more interpretable, less complex, and fits the data more efficiently.

To assess the fit of our final model, we created a residual plot by plotting the residuals against the fitted values. This allows us to check the satisfaction of the assumption of linearity and homoscedasticity.



After examining the residual plot, we observe that the residuals are generally patternless and centered around 0, with the exception of a few outliers. Ignoring the few outliers, the residuals appear to be randomly scattered, suggesting that the assumption of linearity is likely satisfied. In regards to the assumption of homoscedasticity, although the residuals exhibit slightly higher variance for lower fitted values, this could be attributed to random variability within our dataset. However, ignoring some outliers on the right and the top of the plot, the remaining residuals appear to have relatively constant variance. This suggests that the assumption of constant variance, or homoscedasticity, is likely satisfied. Therefore, transformations and using a quadratic model is not needed.

Conclusion

In our project, to avoid the issue of multicollinearity, we had to extract data from a relatively representative Walmart store in the dataset. We then used a linear regression model to analyze the impact of existing variables on weekly sales. The final model achieved higher R^2 and adjusted R^2 values compared to the full model, indicating that the final model provided a better fit. **Therefore, compared to the full model, the final model retained only temperature, unemployment, and holiday flag. This means that among all the covariates in the full model, only these three covariates were considered to influence the weekly sales of a Walmart store in some way.**

Interpreting the correlations in our model contextually: The positive correlation between Temperature and weekly sales could indicate that consumers are more inclined to buy groceries when it is warmer, as certain weather conditions associated with cold temperatures, such as snowy or icy weather, could deter consumers from driving to Walmart those weeks with low temperatures, consequently lowering sales. The negative correlation between unemployment and sales could potentially mean that the result of a higher unemployment rate could be a weaker economy, causing people to spend less altogether. Finally, on weeks with holidays, higher weekly sales could be caused by the fact that there tend to be more parties and celebrations on holidays, encouraging people to buy more food and necessities for their parties, and consequently, spending more, resulting in higher weekly sales.

Important Limitations to note:

The R^2 and adjusted R^2 statistic is low, meaning our model (both full model and our final model) does not do a very good job at describing the variation in the response:
In real-world business practices, weekly sales are influenced by a wider range of factors, including local market competition, marketing spend, lifestyle habits, wealth disparities between cities or communities, transportation, mall size, and other quantifiable or non-quantifiable factors. However, the available dataset provides only a limited number of variables. As a result, our model had to omit many variables that may have a relationship with weekly sales, leading to poor model fit and limiting its ability to effectively explain the data.

Our model is not very generalizable:
Due to the limitations of our current knowledge, we were unable to effectively address the multicollinearity issue across different stores. As a result, we had to select data from a relatively representative store (the store with median sales) among all stores in an attempt to make the model relatively more generalizable. However, this approach still did not resolve the issue. Since our model fitting was based solely on the data from one store, its performance on data from other stores is poor, making it difficult to achieve generalizability.

Outliers were removed during model selection:
Although outliers were removed during model selection to allow for a more accurate, unbiased final model, it did have its drawbacks. Although outliers are annoying, they sometimes carry significant information about the dataset that hasn't been discovered yet, especially since they are usually not the results of entry errors. By removing outliers, you could be discarding valuable information that could be important statistically.

Improvements we could potentially make:

To better optimize our model, we could consider incorporating additional variables into the model. For factors that are inherently non-quantifiable, we can attempt to quantify them using a scoring approach. For instance, we could assign scores from 1 to 10 to represent the varying levels of high consumer willingness across different regions based on comparative analysis and incorporate this score as a variable into the model. At the same time, we may also explore potential nonlinear relationships between sales and the variables to enhance the model's ability to explain the dependent variable. Finally, we will explore new methods to address the issue of multicollinearity, aiming to incorporate data from more stores into our model fitting process.