**Final Project Report**

**Linear Regression to Predict Weekly Sales**

By: Jacob Haussler, Bradley Haren, Tien Ho, Muhammed Abubaker Syed

Logo

Description automatically generated

**Table of Contents**

1 Introduction........................................................................................................................ 3

1.1 Introduction …………………………………………………………………………. 3

1.2 Objectives …………………………………………………………………………… 3

1.3 Research Questions ………………………………………………………………….. 4

2 Methodology …………………………………………………………………………….. 4

3 Results …………………………………………………………………………………… 5

3.1 Setting up the MLR Model ………………………………………………………….. 5

3.2 Transformation of the Data ………………………………………………………….. 6

3.3 Stepwise Methods …………………………………………………………………… 7

3.4 Influential/ Leverage Point Analysis ………………………………………………... 7

3.5 Residual Analysis ……………………………………………………………...…… 10

4 Conclusion ………………………………………………………………………………13

4.1 Conclusion …………………………………………………….….….……………...13

Appendices ………………………………………………………….…………………...14

**Chapter 1**

**Introduction**

**1.1 Introduction:**

We have decided to use a data set[[1]](#footnote-1) consisting of Walmart stores with variables Store, Date, Weekly\_Sales, Holiday\_Flag, Temperature, Fuel\_Price, CPI, and Unemployment. The dataset was available on Kaggle and the description of the dataset is “Historical sales data for 45 Walmart stores located in different regions are available. There are certain events and holidays which impact sales on each day. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to inappropriate machine learning algorithms. Walmart would like to predict the sales and demand accurately”. However, we will use a particular subset of the dataset to predict weekly sales. The rest of the data that is not the response variable or the regressors are discarded since they do not present any real meaning to the model.

**1.2 Objectives:**

1. We want to predict the Weekly\_Sales based on the other variables within the data set.
2. We want to use a Multiple Linear Regression model that regresses Weekly\_Sales onto Holiday\_Flag, Temperature, Fuel\_Price, CPI, and Unemployment.
3. Use log transformation to increase the R-squared value and fix the linearity.
4. Use stepwise methods for variable selection in order to find the statistically significant variables.
5. Find the high leverage points as well as outliers and influential points on the transformed data set.
6. Compute Residual Analysis on the Data.

**1.3 Research Questions:**

1. What variables are most significant?
2. What variables are in the model after all of the transformation?
3. What are the most influential points in the data?

**Chapter 2**

**Methodology**

**2.1 Methodology:**

First, using the pairs plot in R in order to see the data variables related to one another. Then to limit the amount of data we must use filter in R to change the data set to one store only which would be store 36. The main methodology that is going to be used throughout this project is Multiple Linear Regression model. We will use this model to model Weekly\_sales against the other variables in the data set. The fitted Regression for this MLR model would be:

Model: y = β0 + β1 x4+β2 x5+β3 x6+β4 x7+β5 x8

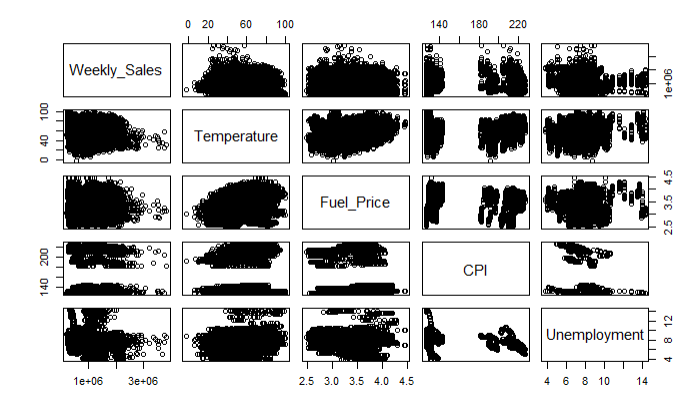
With the MLR model we will also be using different methods of transformation including log transformation to transform the data to be more linear, stepwise methods to determine which variables are statistically significant, Influential/leverage point analysis to determine high leverage points as well as finding outliers and influential points, and Residual Analysis.

**Chapter 3**

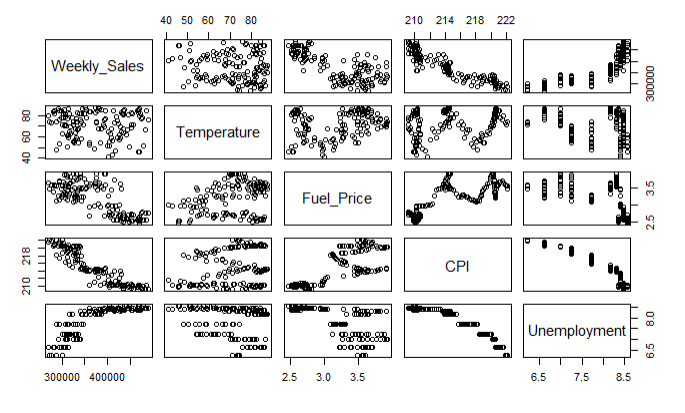
**Results**

**3.1 Setting up the MLR Model:**

First, we must use the pairs function in R to determine the linearity of the variables compared to Weekly\_Sales.



We notice that this is too much data contained here so we want to limit the amount of data. To do this we will look at one store at a time. This store we decided to be store 36. Now the pairs plot after filtering the data to only store 36 looks like this:



This is a much easier data set to work with now and will be better to analyze. We then need to create a Multiple Linear Regression model with Weekly\_Sales being Regressed on the variables Holiday\_Flag, Temperature, Fuel\_PRice, CPT, and Unemployment. In doing this we get the adjusted R-squared value of 0.8473 with the fitted regression line being:

WeeklySales\_hat = 4295736.1 - 9053.8(Holiday\_Flag) + 242.8(Temperature) + 1733.3(Fuel\_Price) - 17367.1(CPI) - 27347.8(Unemployment)

**3.2 Transformation of the Data:**

We noticed the pairs plot in the section above shows that some of the variables related to Weekly\_Sales are not quite linear. Since this is the case we will use log transformation on the data set. In doing so, this slightly improved our adjusted R-squared to 0.8587 which means the transformation was appropriate. The new fitted regression line is:

log(WeeklySales\_hat) = 22.6896722 - 0.0235910(Holiday\_Flag) + 0.0005696(Temperature) + 0.0156304(Fuel\_Price) - 0.0444870(CPI) - 0.0519033(Unemployment)

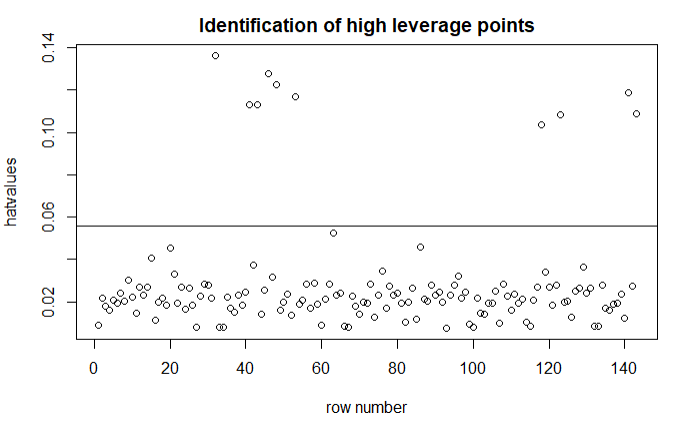
**3.3 Stepwise Methods:**

We notice that not all of the variables are statistically significant. For this reason we will use stepwise methods for variable selection in this model and see how it affects the results. We will first try dropping Fuel\_Price since it seems to be the least significant in the MLR model. After dropping Fuel\_Prices the model increases the adjusted R-squared to 0.8593, which suggests that Fuel\_Price does not belong in the model. However, the VIF for CPI and Unemployment are around 9, which suggest that one of these may need to be dropped due to multicollinearity. We decide to Remove Unemployment since it is less significant in the MLR model than CPI. After removing the Unemployment variable it leads to a small decline in the adjusted R-squared to 0.8574, but the VIF results show that we no longer have multicollinearity in our model. The new regression line after the stepwise methods is:

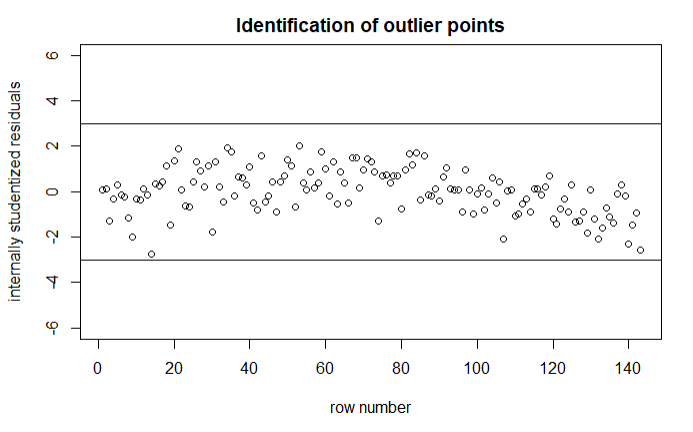
log(Weekly\_Sales)\_hat = 20.3530144 - 0.0353244(CPI) + 0.0007238(Temperature) - 0.0256877(HolidayFlag)

**3.4 Influential/ Leverage Point Analysis:**

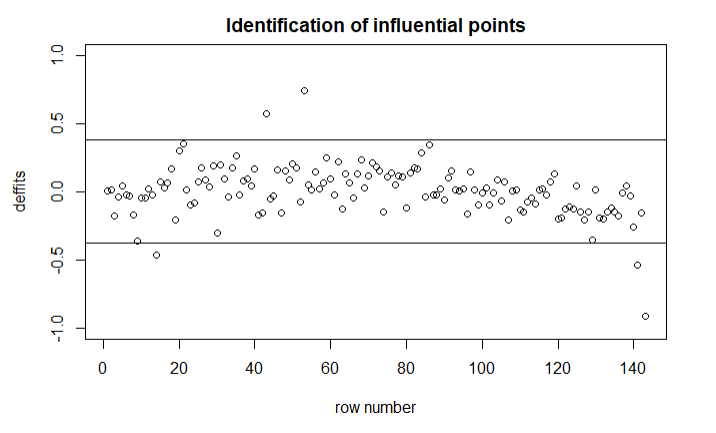
First in this section we want to find the high leverage points in the data. In doing this we find that the high leverage points are rows 32, 41, 43, 46, 48, 53, 118, 123, 141, 143. As shown in the graph below:



Next, we want to find the outliers in the data using internally standardized residuals. In our calculations we find that there are no outliers in our data after all the transformation we have done to it. As shown in the graph below:



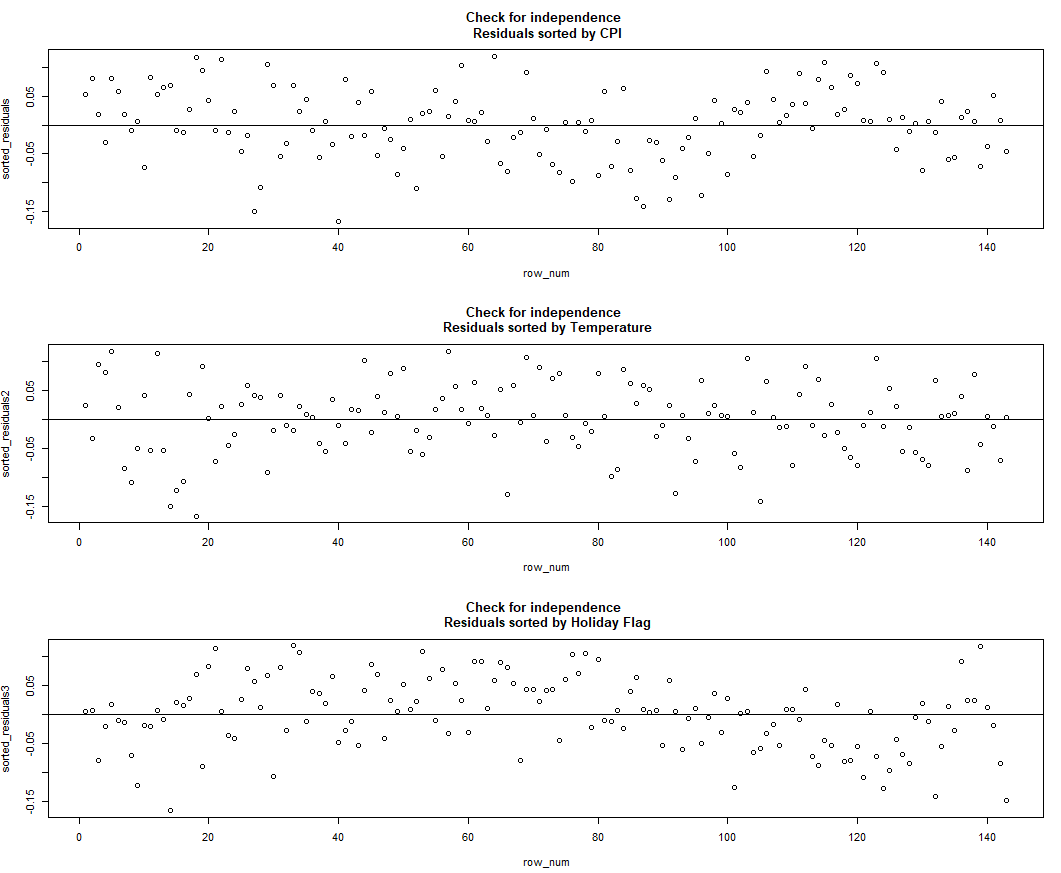
Last thing we want to do in this section is find the most influential points in the data set using (dffits). In our calculations we notice that the most influential points are rows 14, 43, 53, 141, 143. As shown in the graph below:

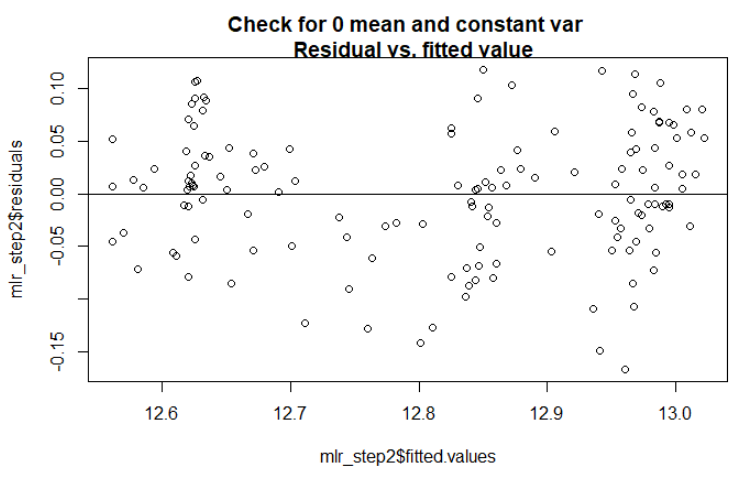


All leverage points and influential points are as follows: 14, 32, 41, 43, 46, 48, 53, 118, 123, 141, 143. To perform a complete analysis, it would be helpful to see what these outliers look like. Looking at these rows shows that most of them have a Holiday\_Flag = 1, meaning that the Weekly\_Sales data comes from a week containing a holiday. This may suggest that our data could be clustered into two groups: one with holidays, and one without. A possibility is that our model is labeling these as outliers since most of the data in the model has a Holiday\_Flag = 0, since there are many more weeks in a year without a holiday compared to weeks with one, which is reflected in our sample.

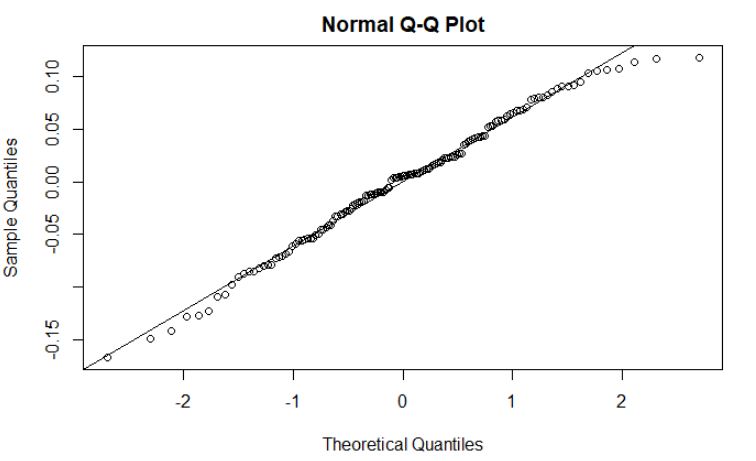
**3.5 Residual Analysis:**

Now that we have a finished model, let us check the residuals and ensure that everything looks good. Our first check is going to confirm that this model has independence of random error. This is referenced from the following graphs (one for each variable present in the model):



Overall, it appears that each of these is centered on 0 and is homoscedastic in terms of variance. There is a slight jump in the check for Holiday Flag residuals, but this is most likely attributed to what was mentioned in the previous section. A check is also necessary for 0 mean and constant variance for the model as a whole. Here are the results for this:

Overall, it looks again as if there is a general centering around a 0 mean, and relatively constant variance throughout (minus the one small subsection located between 12.7 and 12.8 in the fitted values). Lastly, a look at the Normal Q-Q Plot:



This plot is centered almost perfectly on the normal line. This gives some evidence that our sample data is pulled from a population that is normally distributed.

**Chapter 4**

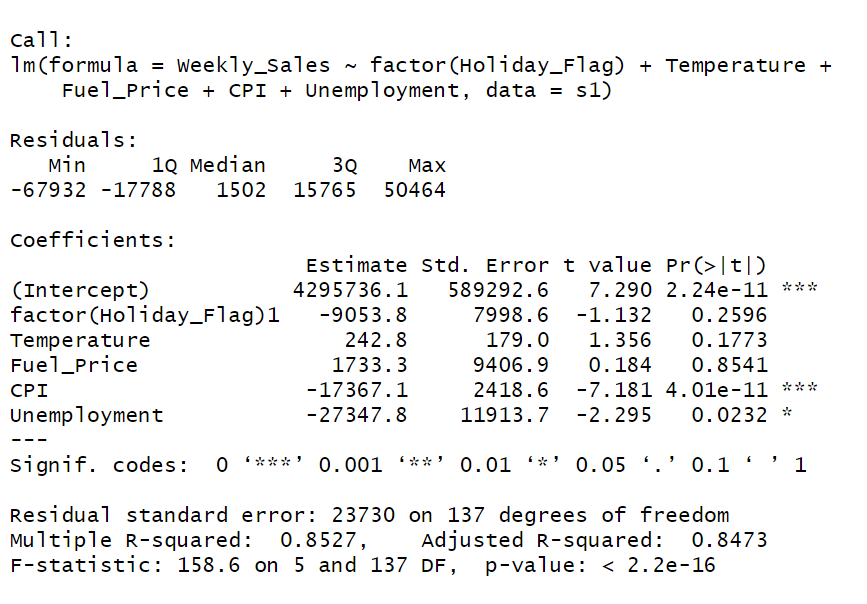
**Conclusion**

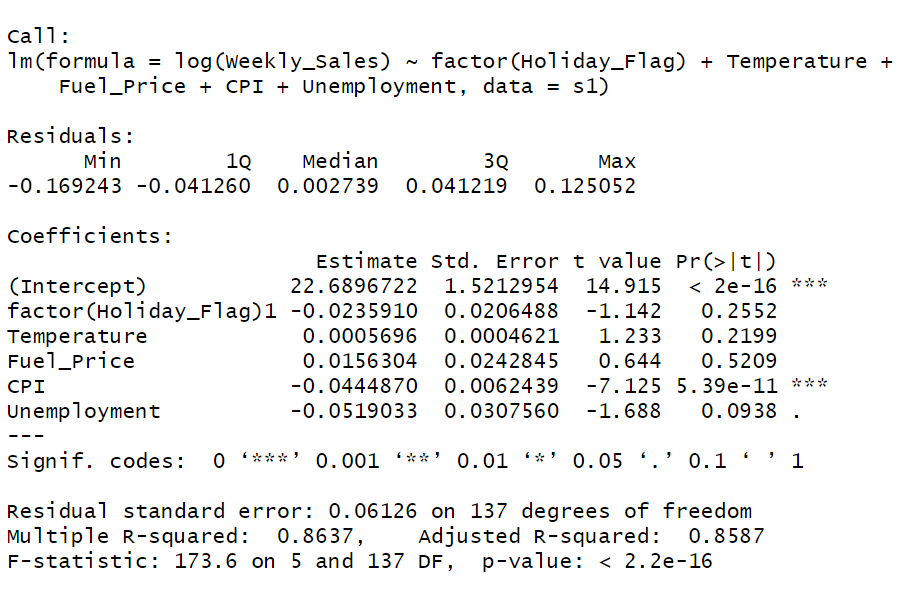
**4.1 Conclusion:**

In conclusion, we have found that the best variables for Weekly\_Sales to be regressed on are CPT, Temperature, and Holiday\_Flag. These are the most significant variables in our data set when comparing them to the Weekly\_Sales. In our calculations we find that the most influential points in our data set are the rows 14, 43, 53, 141, 143. In doing this project we have learned how to use supply chain management to complete real world problems and come up with complex solutions. So, all in all, we figured out which model was the best for predicting the Weekly\_Sales and which variables had the most significant factor in doing so.

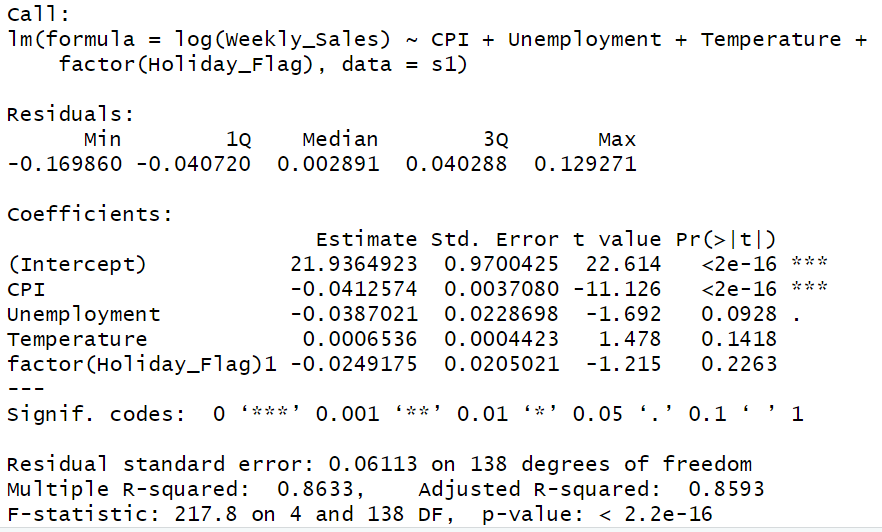
**Appendices**

**Appendix A: Summary of the MLR model (section 3.1):**

****

**Appendix B: Summary of the log MLR model (section 3.2):**

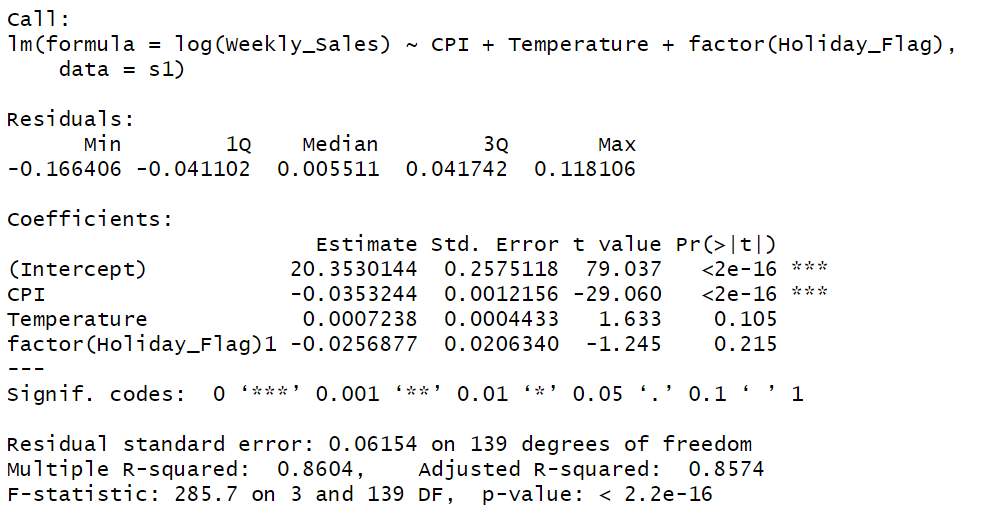
**Appendix C: Summary of the stepwise MLR model (section 3.3):**

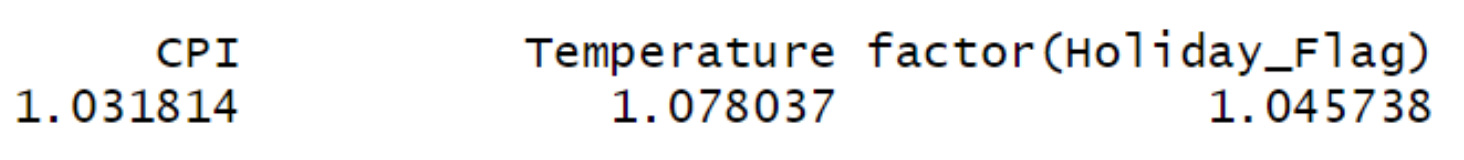
****

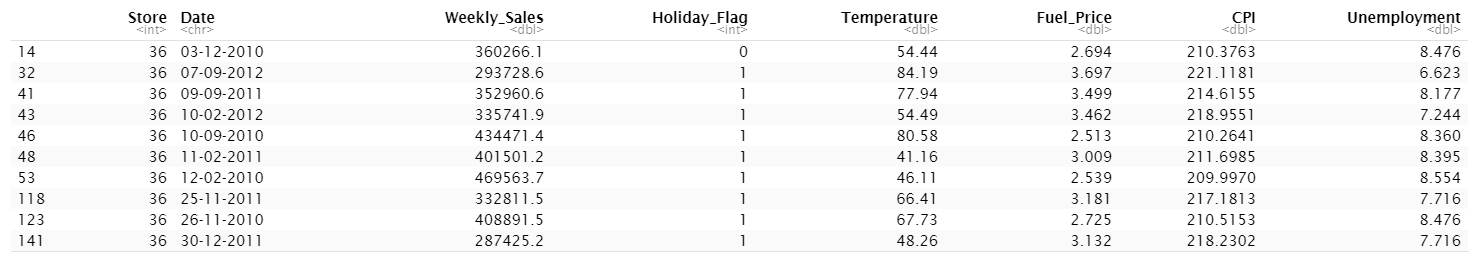
**Appendix D: Summary of the VIF (section 3.3):**



**Appendix E: Summary of the second stepwise MLR model (section 3.3):**

****

**Appendix F: Summary of the second VIF (section 3.3):**

**Appendix G: Dataframe containing outliers (section 3.4):**

1. Rutu Patel. 2021, Historical sales data of the Walmart store, <https://www.kaggle.com/rutuspatel/walmart-dataset-retail> [↑](#footnote-ref-1)