Final Report

Prepared for: Dr. Chuan-Fa Tang

Prepared by:
Ved Nigam (vxn190021)
Saketh Dontikurti (sxd190083)
Mitchell Craven(mac190011)

10 May, 2023

Data Introduction

Describing the Data

Walmart is the largest and the most profitable retail chain in the world containing 10,500 stores in 20 countries. Everywhere we go, we see walmart. Whether it's in one of the largest cities in the world or a rural town in the middle of nowhere, a Walmart will always be there serving the needs of customers from groceries to the newest technology available for consumption. We here at team great value want to determine the factors that make Walmart one of the most profitable retail stores we have seen. Specifically when it comes to Weekly Sales.

We have discovered our dataset on *Kaggle*, but it is not originally sourced from anywhere. This data set has historics sales data on 45 Walmart Stores located in different regions of the United States from the years 2010 to 2012. This set also includes 6435 observations with 8 variables. We will not be using all of the 8 variables as some of the variables will be used as response and others will be used as predictor variables. We will have some variables not be used as either, for, we don't have any use for a predictor and response variable. We will have some use for it later when we want to explore our data.

Reflective Process

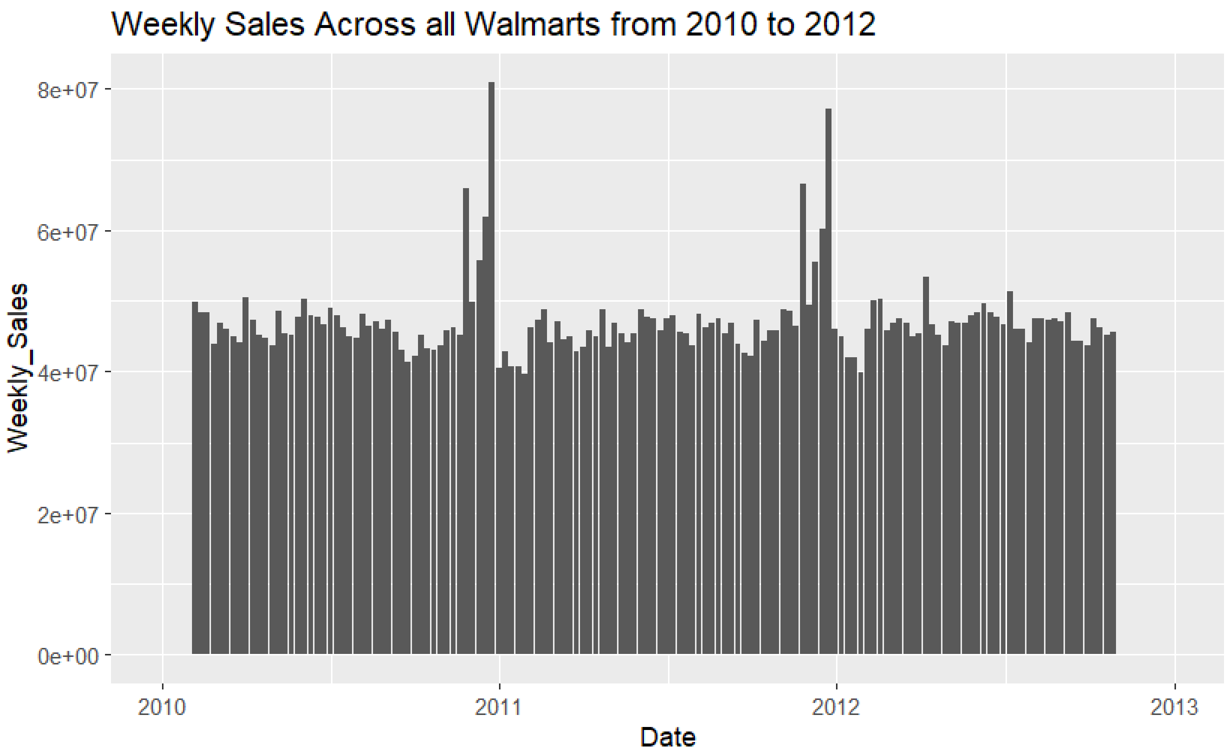
The Walmart data set contains a total of 8 variables. The Weekly Sales and Consumer Price Index are our possible response variables. Doing Weekly Sales is quite obvious since a store's success is determinant of how well it is able to sell and make money. CPI measures the quality of the store; we are also determining what factor makes that store good. CPI measures the average price paid for groceries in the United States during a specific time period; for our case, the CPI was remeasured every week since we are considering weekly sales. Overall, we want to identify the best response variable to determine how well a store is performing. CPI will directly have an effect on our weekly sales because the trends seen in weekly sales will be the same seen in the CPI. After an initial analysis of the data, we picked Weekly Sales and the CPI as our two possible response variables.

We will be considering 4 predictor variables. The first predictor variable is **Temperature**. We want to determine whether the temperature overall determines the consumers' shopping habits for that week. It could be that warmer temperatures encourage more shopping than cooler temperatures; but, this may not remain true for the holiday seasons in November and December in the United States during winter. Temperature will be measured in Fahrenheit. We will have temperatures ranging from 0 degrees Fahrenheit to temperatures close to 100 degrees. Our second predictor will be the **Unemployment**. We see strong variation in our unemployment rate. The unemployment will be around 3 percent to 14 percent. We want to determine if a high unemployment rate means a lower weekly sales or if a low unemployment rate means a high weekly sales. The next predictor is **Holiday_Flag** covering the weeks close to or on Super Bowl Weekend, Labour Day, Thanksgiving and Christmas. Our last predictor we will be using is the **Fuel_Price** going from around 2.50 dollars to around 4.30 dollars. We want to see if a higher fuel price suggests lower weekly sales and a lower fuel price suggests a higher weekly sales or CPI.

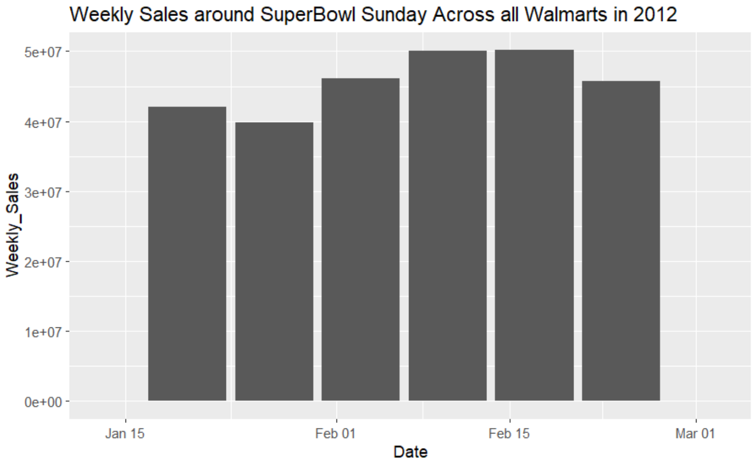
Our response variable **Weekly_Sales** is the revenue collected from the store from that specific week. In our data set, it generally ranges from 240k to 4 million dollars per week and is measured in USD. Our **CPI** ranges from 126 to 227 and is measured by comparing the current cost of goods to the previous year's cost in goods all multiplied by 100. Generally, a higher CPI is bad for a store and a lower CPI is good for a store. We will have two full models, one for Weekly_Sales and one for CPI. We will use 4 predictor variables in two separate models to create our linear regression models to see which have a positive, negative, or no correlation with respect to our CPI and Weekly_Sales.

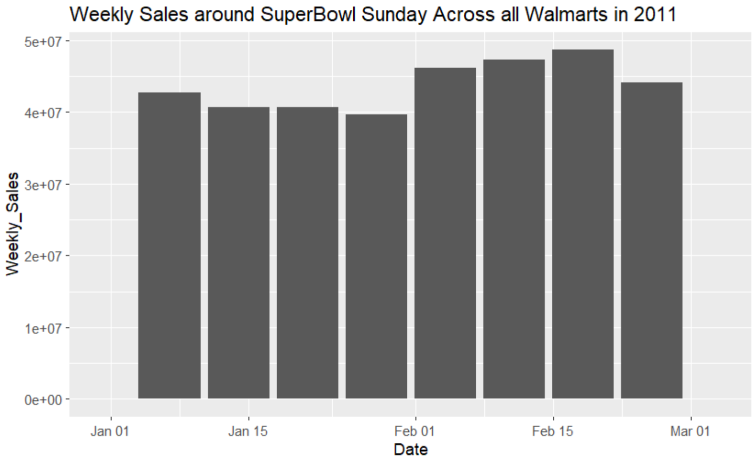
Data Analysis

Exploratory Analysis

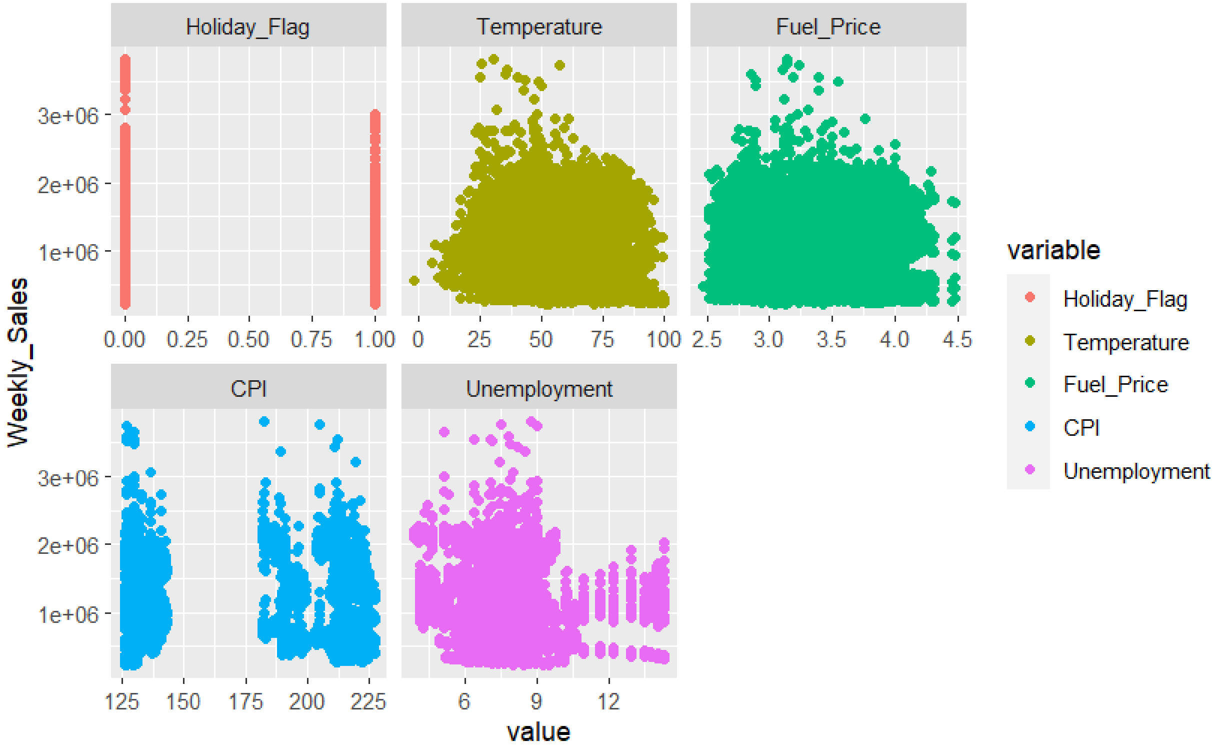


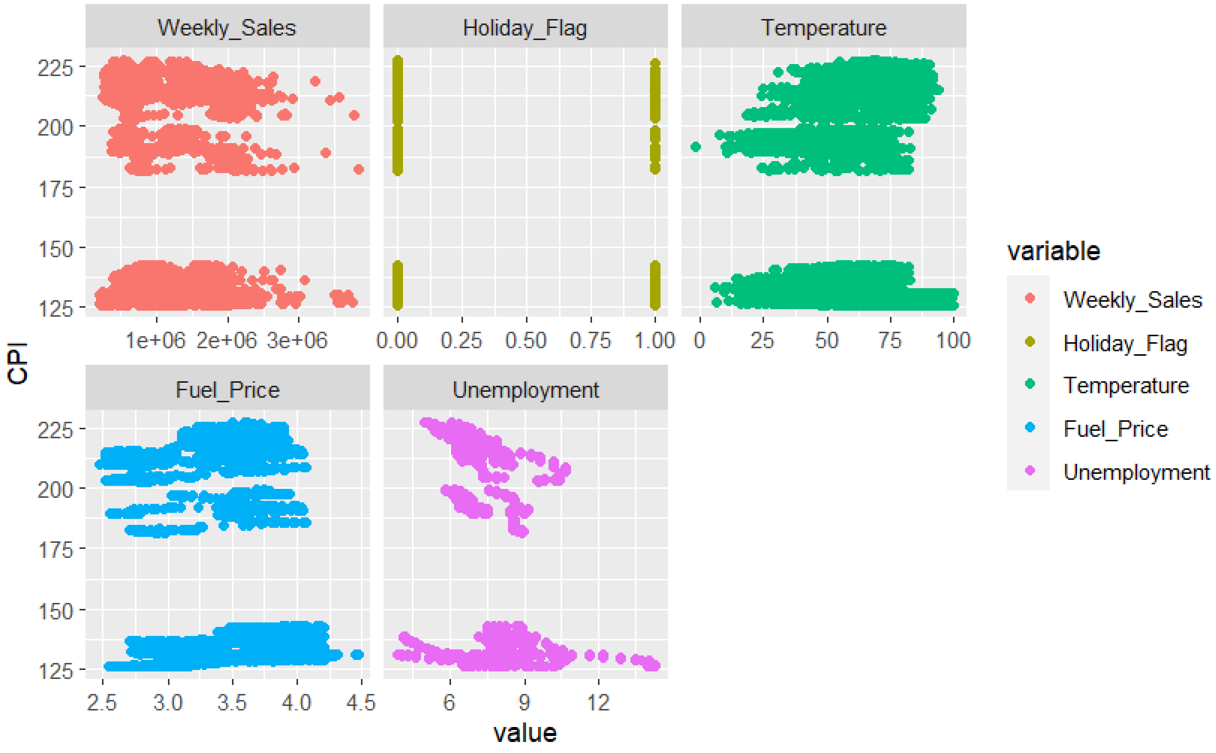
In the above graph, we see that the Weekly_Sales are consistent until the end of the year. Towards the end of the year, we holidays such as Thanksgiving, Black Friday, Christmas, and New Years heavily affect sales. That is when we started to see a spike in Weekly Sales in 2010 and 2011. In 2012 however, the sales stayed roughly the same with the largest spikes happening around July 4th and good friday. The reason for this is because the data stops until October of 2012 rather than all the way to the end of the year making it seem odd that there is no spike towards the end of the year. Based on how the previous years performed, we do expect to see a spike towards the end of the year.



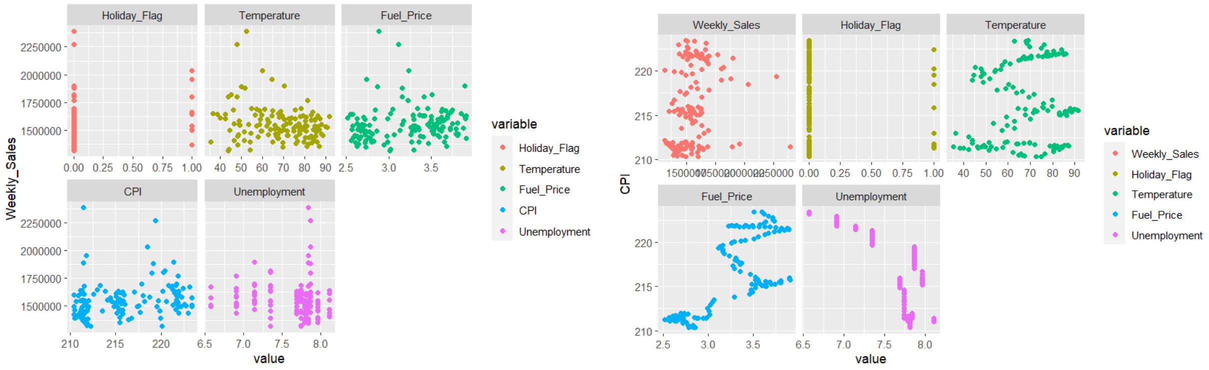


In the above graphs, there is a small bump between January to March in both 2011 and 2012. That bump represents when the Super Bowl happened, but the reason you don’t see a huge spike compared to towards the end of the year is that stores are closed during Thanksgiving, Christmas and New Years, whereas, a lot of people go out during the Super Bowl and July 4th.

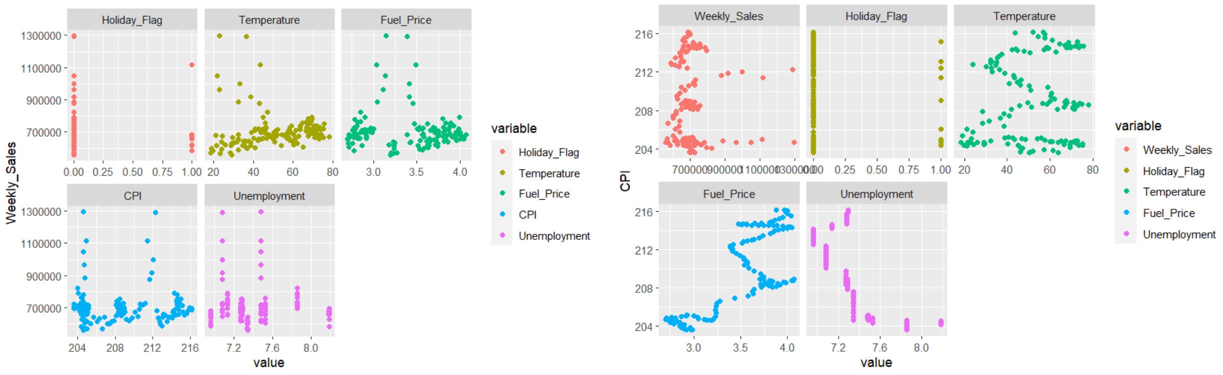




We used scatterplots to display the distribution of our Weekly_Sales, as seen above, based on all the predictor variables before we did any transformation on the data set. We performed the same process for CPI. Below, we see that it is difficult to determine the correlation of each variable compared to both the Weekly_Sales and CPI.



Store 1 Scatter Plots



Store 25 Scatter Plots

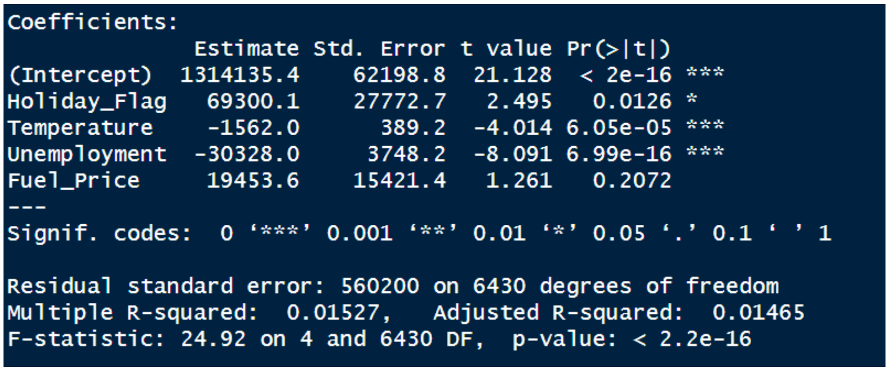
Because of the difficulty in finding the trends, we decided to find the scatter plots for specific stores. We decided to do Stores 1 and 25 to see if there's any similarities or differences to determine how our overall scatter plots are being affected. In Store 25, we see that there is a stronger positive correlation with temperature and Weekly_Sales if you ignore the handful of outliers in the beginning. You can also see a somewhat of a negative correlation with the Unemployment rate and Weekly_Sales. However with Store 1, we don't see much correlation at least when it comes to Weekly_Sales. In CPI, the story is a bit more different. We see that the fuel price and unemployment rate is quite consistent with the CPI, thus making it look as if CPI is a more viable option as a response variable. Due to such variation in each store, it would be biased to consider a specific store. Our goal is to find trends regarding Walmart sales as a country, not in a specific region. For this reason, we did not subset our data in any way. We will be continuing with the original dataset we have.

Picking a response variable

Our plan is to consider Weekly_Sales and CPI as response variables, and use the rest of the variables in the dataset as predictors to fit a multi-linear regression model. Another possible approach is to build separate linear regression models for each predictor to see which attribute most closely has a relationship with the Weekly_Sales. We also used Holidays as a response variable to fit it with an multilogistic regression model and predict whether a week is classified as a holiday week or not. In that case, we used Weekly_Sales and CPI as predictor variables. Since this part of the project focusses on finding linear trends in a dataset, that research will be part of our future work on this datatset. After discussion with the team, we decided to compare Weekly_Sales and CPI to pick the best response variable for our research.

Variable Selection

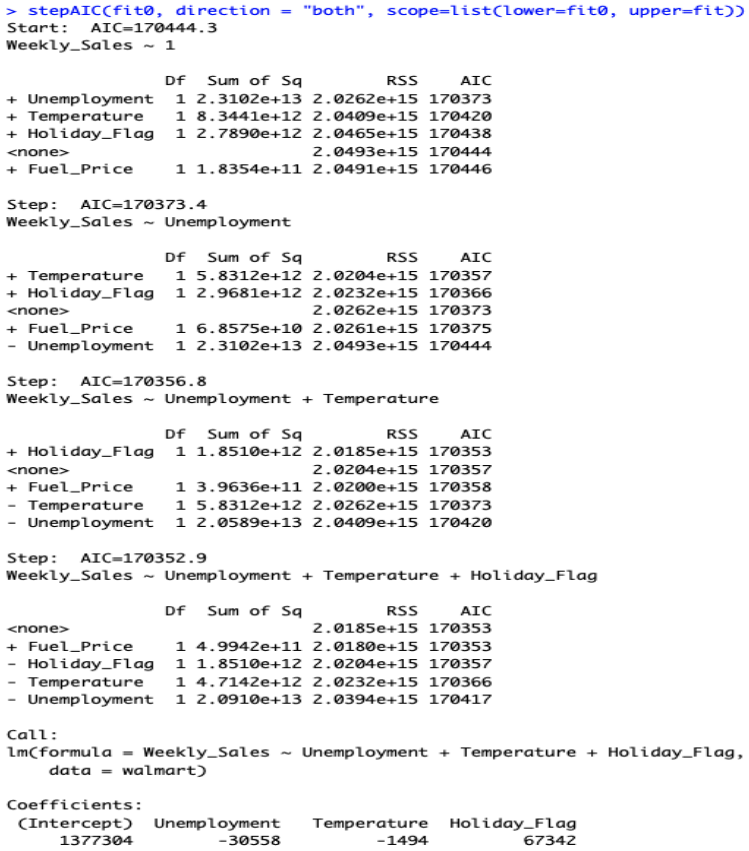
One of the first tasks we do before fitting an initial linear model os look at which variables we do not want. We removed Date due to its irrelevance in what we want to determine with the CPI and Weekly Sales. Next thing we do is fit a linear model with the predictors:



```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1314135.4    62198.8  21.128  < 2e-16 ***
Holiday_Flag   69300.1    27772.7   2.495   0.0126 *  
Temperature    -1562.0      389.2  -4.014 6.05e-05 ***
Unemployment  -30328.0     3748.2  -8.091 6.99e-16 ***
Fuel_Price     19453.6    15421.4   1.261   0.2072    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560200 on 6430 degrees of freedom
Multiple R-squared:  0.01527,	Adjusted R-squared:  0.01465 
F-statistic: 24.92 on 4 and 6430 DF,  p-value: < 2.2e-16
```

Looking at the results from the first result, we see that all of the variables beside Fuel_Price are statistically significant. Another thing to notice is the R-squared value. We notice that it is incredibly close to 0 signifying that the model does not rteflect a linear relationship.



```
> stepAIC(fit0, direction = "both", scope=list(lower=fit0, upper=fit))
Start:  AIC=170444.3
Weekly_Sales ~ 1

               Df  Sum of Sq        RSS    AIC
+ Unemployment  1 2.3102e+13 2.0262e+15 170373
+ Temperature   1 8.3441e+12 2.0409e+15 170420
+ Holiday_Flag  1 2.7890e+12 2.0465e+15 170438
<none>                       2.0493e+15 170444
+ Fuel_Price    1 1.8354e+11 2.0491e+15 170446

Step:  AIC=170373.4
Weekly_Sales ~ Unemployment

               Df  Sum of Sq        RSS    AIC
+ Temperature   1 5.8312e+12 2.0204e+15 170357
+ Holiday_Flag  1 2.9681e+12 2.0232e+15 170366
<none>                       2.0262e+15 170373
+ Fuel_Price    1 6.8575e+10 2.0261e+15 170375
- Unemployment  1 2.3102e+13 2.0493e+15 170444

Step:  AIC=170356.8
Weekly_Sales ~ Unemployment + Temperature

               Df  Sum of Sq        RSS    AIC
+ Holiday_Flag  1 1.8510e+12 2.0185e+15 170353
<none>                       2.0204e+15 170357
+ Fuel_Price    1 3.9636e+11 2.0200e+15 170358
- Temperature   1 5.8312e+12 2.0262e+15 170373
- Unemployment  1 2.0589e+13 2.0409e+15 170420

Step:  AIC=170352.9
Weekly_Sales ~ Unemployment + Temperature + Holiday_Flag

               Df  Sum of Sq        RSS    AIC
<none>                       2.0185e+15 170353
+ Fuel_Price    1 4.9942e+11 2.0180e+15 170353
- Holiday_Flag  1 1.8510e+12 2.0204e+15 170357
- Temperature   1 4.7142e+12 2.0232e+15 170366
- Unemployment  1 2.0910e+13 2.0394e+15 170417

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

Coefficients:
 (Intercept)  Unemployment   Temperature  Holiday_Flag  
     1377304        -30558         -1494         67342  
```

Another way we did variable selection was using the stepAIC model selection, as show above. One of the values we looked at while doing stepwise selection was the AIC score. We want a decreasing AIC score when we add more variables. Looking at the model, we see that all of the variables were included except the Fuel_Price. By confirming with the stepwise model selection, we decided to remove Fuel_Price as a predictor variable.

Model Fitting

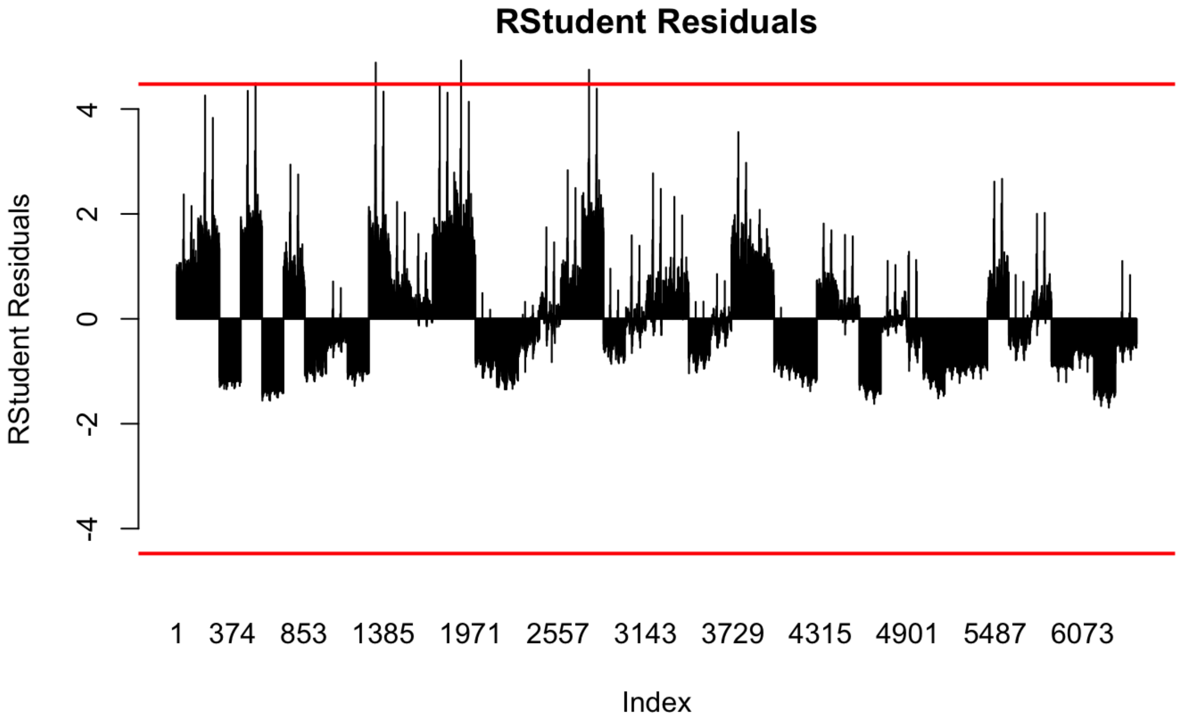
We tried transforming our response variable and choosing different combinations of predictor variables, but we were not able to find a linear relationship in the Walmart dataset. We tried to take the square root of weekly sales, but that barely increased the R-squared value to around .15. We then tried to take the log of Weekly Sales, but that had a very similar effect to the square root. We then thought about removing one of our predictor variables. After doing the AIC variable selection, we were left with Holiday Flags, Temperature, and Unemployment. Even after, our R-squared value told us that there is no significant linear relationship between Weekly Sales of a Walmart with Holidays, Temperature, or Unemployment throughout the year.

Since our dataset included data for 3 years, we subsetted the data into just one year. This kept almost all our metrics the same, meaning the data throughout the years resembles similar patterns. Thus, there was no linear pattern found again. To continue to confirm that there is no linear relationship in the data, we performed other analyses on the dataset to confirm that if there was a linear relationship in the data, we would be able to see it.

Residual Analysis

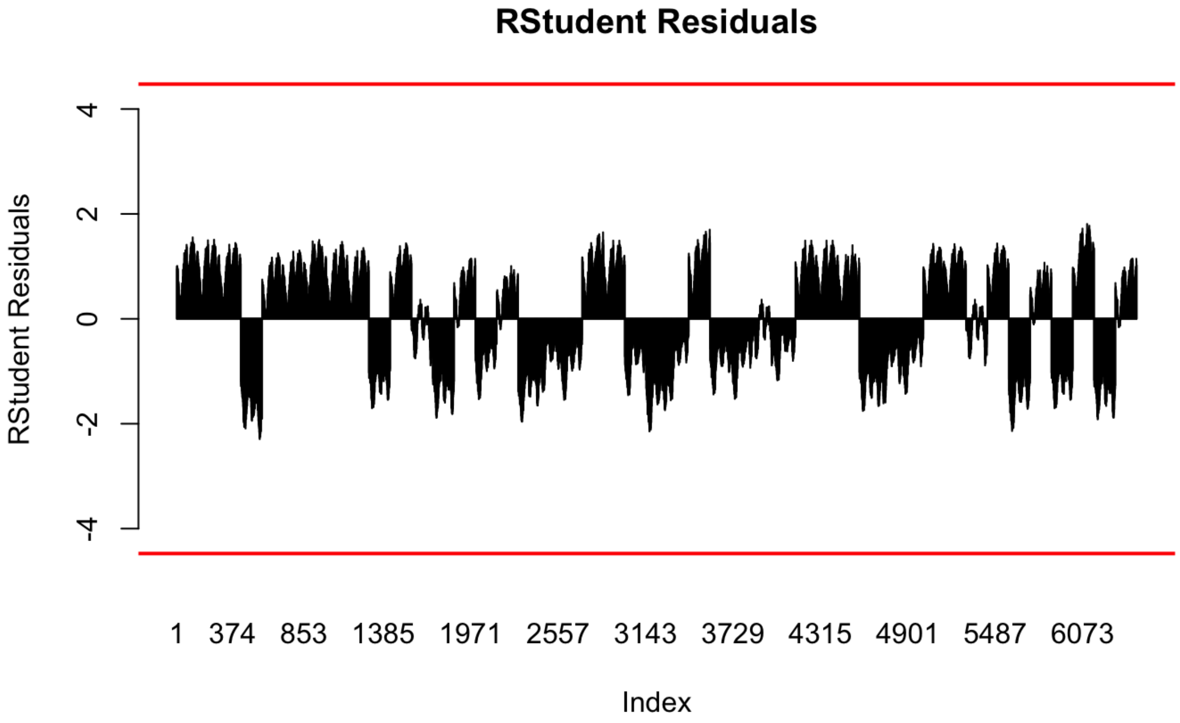
One of the most important reasons that we performed residual analysis on the dataset was to identify whether there is a relationship between the residuals of a dataset. These would include spikes in Weekly Sales which will happen around certain times of the year. Some of them are shown in our initial analysis of the data: there will be spikes in Weekly Sales around special times of the year. These weeks include Super Bowl week, Easter, Thanksgiving, Black Friday, Christmas, and many other days that may cause grocery or merchandise sales.

After performing a RStudent residual analysis and making a barplot of the dataset, this is what we see:



The above barplot has plotted the residuals for our dataset on a 95% confidence interval (the horizontal red line) of how influential they will be on increasing the standard deviation of our beta value predictions. Based on the output, we don’t have too many points that will cause a significant deviation.

We also performed residual analysis with CPI as our predictor. At a 95% confidence interval, we got a very nice plot:

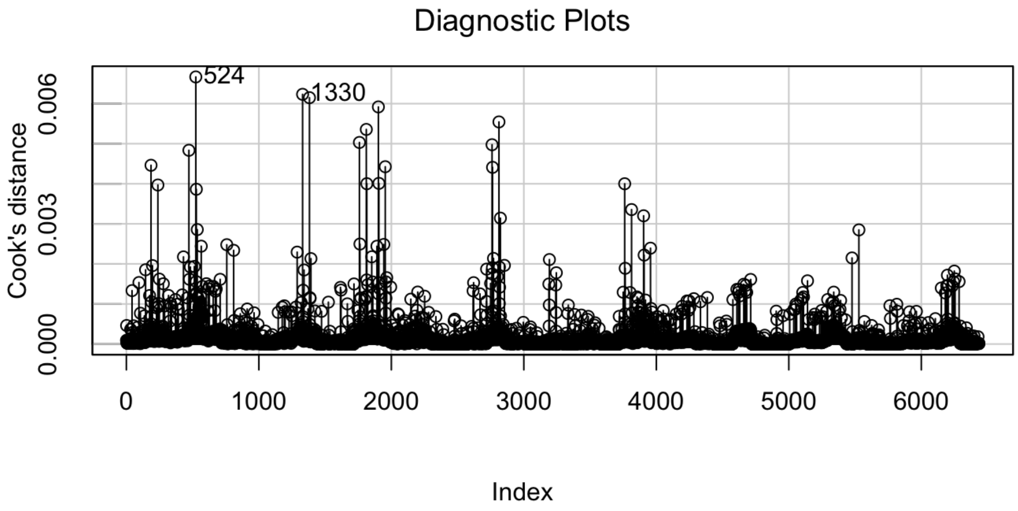


The residuals are beautiful, as there are none statistically significant. Initially, we wanted to switch our response variable from Weekly Sales to CPI based on the above residuals plot, but we continued with Weekly Sales as a response variable after doing an influential analysis on the data.

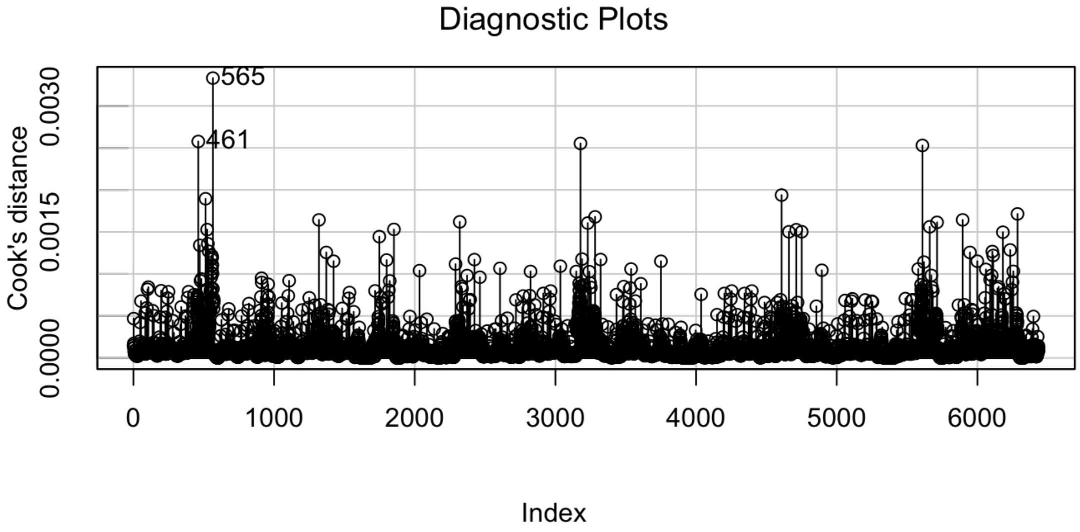
We will perform the influential analysis to confirm if the residuals that cross the red line in the residual analysis for Weekly Sales will not cause significant deviation to our final model.

Influential Analysis

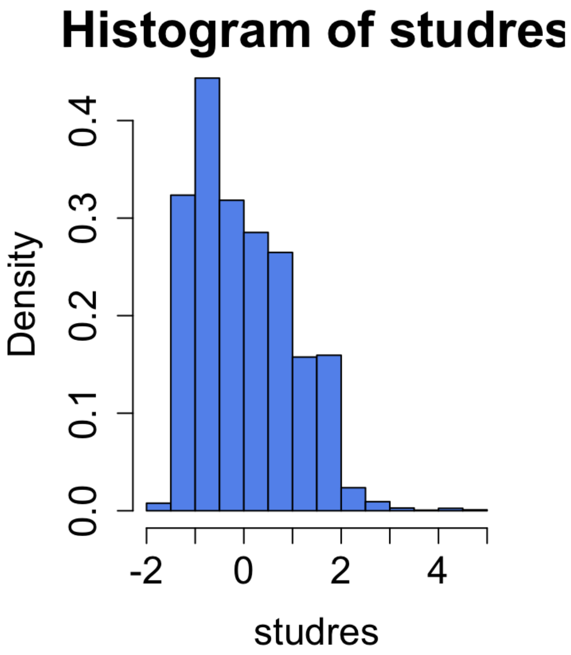
Influential analysis is able to identify significant observations in the data to conclude whether a dataset will be able to provide a significant linear regression model or not. The first significant image of our influential analysis is the following Cook's plot:

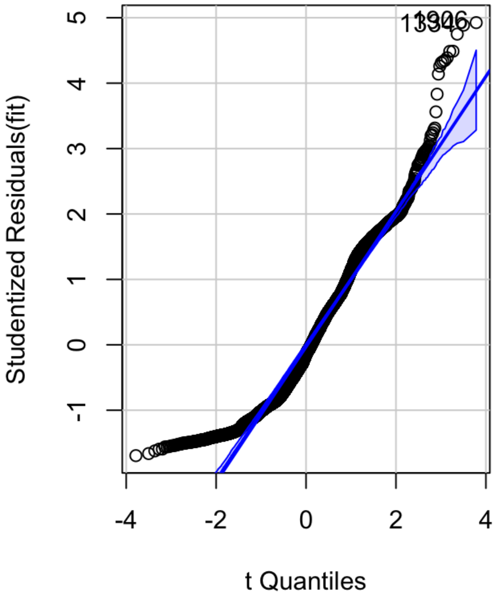


Outlying points on a Cook's diagnostic plot determine if there are significant observations in the dataset to more accurately predict beta values for the regression model. For our case, we performed an influential analysis on Weekly Sales and CPI to pick a more accurate response variable. The following is the Cook's diagnostic plot with CPI as the response variable:

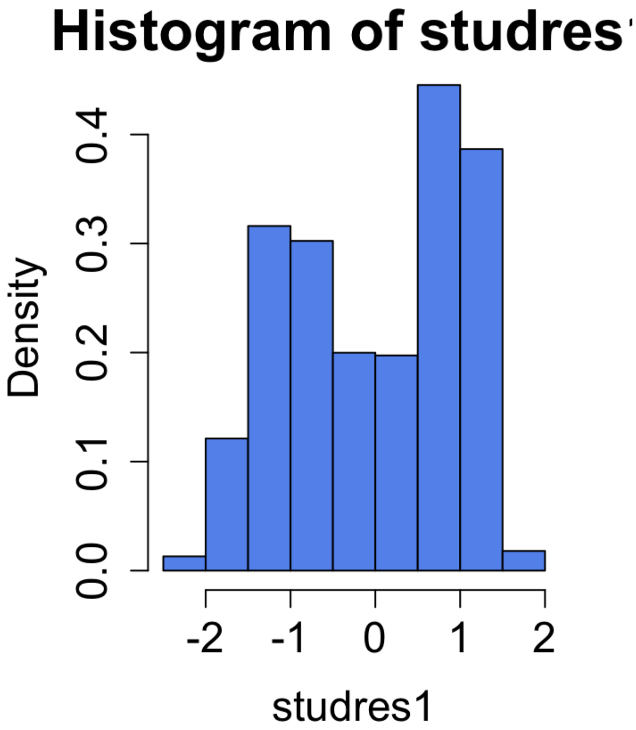


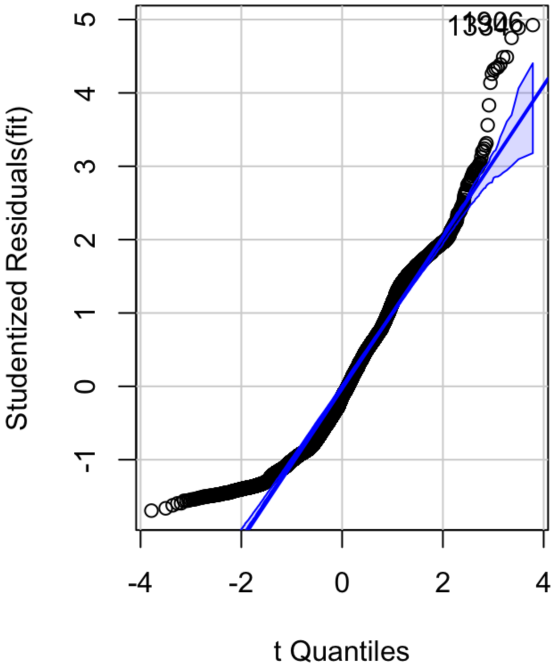
What stood out as we compared the two plots was that Weekly Sales seem to have more influential points than CPI as a response variable. The next plot we looked at was the histogram of studentized residuals. We want these to be normally distributed. The following is the histogram and a QQplot for Weekly Sales as a response variable:





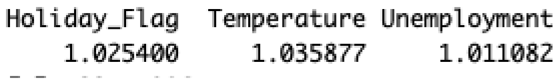
Next is the histogram and QQplot for CPI as a response variable:





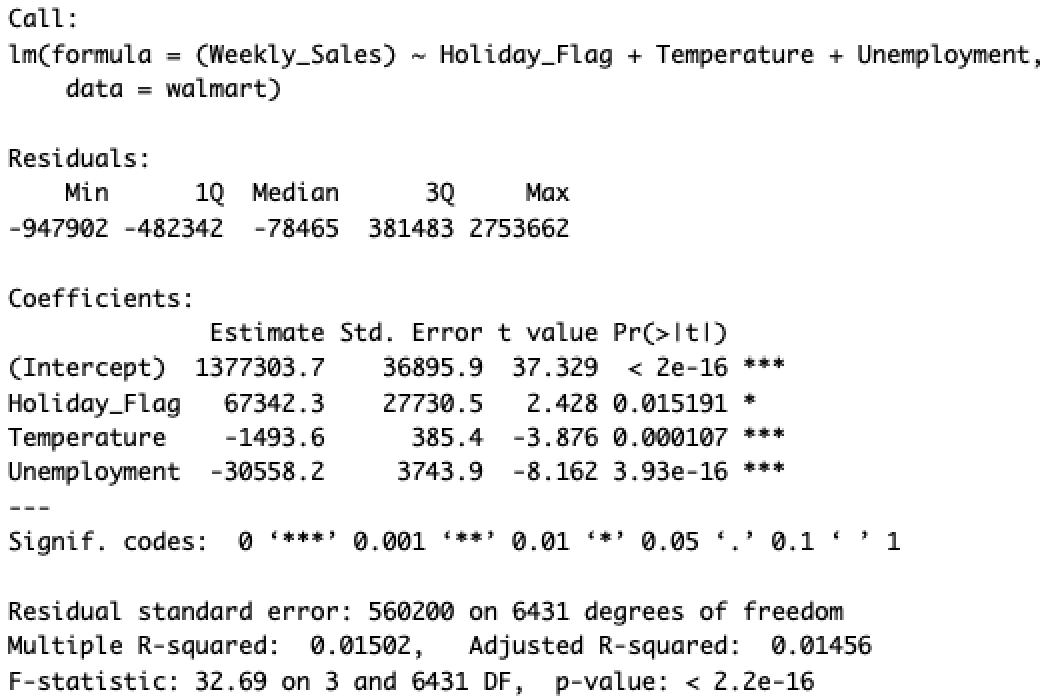
After comparing the two, the Weekly Sales response is more normally distributed than the CPI. Although the histogram for CPI is more centered, the data does not seem to follow a normal distribution. The histogram for Weekly Sales is positively skewed, but is more normally distributed. Due to this, we concluded to use Weekly Sales as our Response variable for our conclusion.

The following are the variance inflation factors for the predictor variables that we used:



```
Holiday_Flag  Temperature Unemployment 
    1.025400     1.035877     1.011082 
```

None of these are above 10, or even close to 10. This concludes that the data is proper for a linear regression and that the predictors we are using are significant. Despite this, we did not find a linear pattern in the data. Our final model is our initial model:



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

Residuals:
    Min      1Q  Median      3Q     Max
-947902 -482342  -78465  381483 2753662

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  1377303.7    36895.9  37.329  < 2e-16 ***
Holiday_Flag   67342.3    27730.5   2.428 0.015191 *
Temperature    -1493.6      385.4  -3.876 0.000107 ***
Unemployment  -30558.2     3743.9  -8.162 3.93e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 560200 on 6431 degrees of freedom
Multiple R-squared:  0.01502,   Adjusted R-squared:  0.01456
F-statistic: 32.69 on 3 and 6431 DF,  p-value: < 2.2e-16
```

Conclusion

Future Direction

- Do a similar analysis like Target in the same way we looked at Walmart and compare results
- It would be interesting to look at Sams club ,whose parent company is Walmart, and how a store that sells items in bulk compares
- Look at the geographical location of stores and how Sales and CPI vary among locations
- Use unemployment as an possible response variable
- Compare Sales and CPI on weekdays and weekends

Conclusion

When comparing CPI and Weekly sales as response variables, we notice that CPI is not normally distributed and that Weekly Sales is a more appropriate response variable. Also, we notice that the ideal model for Weekly sales included Holiday flag, unemployment, and temperature disregarding fuel price by using step AIC. After many transformations, subsets, and adjustments to the variables, we were not able to find a statistically significant linear relationship between the performance of Walmart as a company during the year. Shopping patterns vary throughout the year, which does not make this very suitable for linear regression.

Appendix

References

Rutu Patel(Collaborator). “Historic Sales Data of the Walmart Store.” *Kaggle*, https://www.kaggle.com/datasets/rutuspatel/walmart-dataset-retail

Xu L, Poon WY, Lee SY. Influence analysis for the factor analysis model with ranking data. Br J Math Stat Psychol. 2008 May;61(Pt 1):133-61. doi: 10.1348/000711006X169991. PMID: 18482479.

Team Responsibilities

Saketh Dontikurti- Data Introduction and Exploration

Ved Nigam- Model Fitting, Influential Analysis, and Residual Analysis

Mitchel Craven- Response Variable, Variable Selection, and Conclusion

Code

Data Exploration

```
#Scatter Plots and Histograms
library(MASS)
library(glmnet)
library(olsrr)
library(ggplot2)
library(reshape)
library(car)
library(reshape2)

#Scatter Plot for Weekly sales for all stores
Walmart2 <- melt(Walmart[,3:8], id.vars = 'Weekly_Sales')
ggplot(Walmart2) +
  geom_jitter(aes(value, Weekly_Sales, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")
#Scatter Plot for CPI for all stores
Walmart2 <- melt(Walmart[,3:8], id.vars = 'CPI')
ggplot(Walmart2) +
  geom_jitter(aes(value, CPI, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")
#Scatter Plot for Weekly sales for Store 1
store_1 <- subset(Walmart, Store == 1)
store_1_melted <- melt(store_1[,3:8], id.vars = 'Weekly_Sales')
ggplot(store_1_melted) +
  geom_jitter(aes(value, Weekly_Sales, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")
#Scatter Plot for CPI for Store 1
store_1_melted <- melt(store_1[,3:8], id.vars = 'CPI')
ggplot(store_1_melted) +
  geom_jitter(aes(value, CPI, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")

#Scatter Plot for Weekly sales for Store 25
store_25 <- subset(Walmart, Store == 25)
store_25_melted <- melt(store_25[,3:8], id.vars = 'Weekly_Sales')
ggplot(store_25_melted) +
  geom_jitter(aes(value, Weekly_Sales, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")
#Scatter Plot for CPI for Store 25
```

```
store_25_melted <- melt(store_25[,3:8], id.vars = 'CPI')
ggplot(store_25_melted) +
  geom_jitter(aes(value, CPI, colour = variable),) +
  facet_wrap(~variable, scales = "free_x")
#Reformatting
Walmart$Date <- as.Date(Walmart$Date, format ="%d-%m-%Y")
#Entirety of Weekly Sales from all 45 Stores throughout the 2010-2012
ggplot(Walmart) +
  aes(x = Date, y = Weekly_Sales) +
  geom_bar(stat = "identity") +
  xlim(as.Date(c("2010-01-01", "2012-12-31"))) +
  ggtitle("Weekly Sales Across all Walmarts from 2010 to 2012")

# Super Bowl Sunday
ggplot(Walmart) +
  aes(x = Date, y = Weekly_Sales) +
  geom_bar(stat = "identity") +
  xlim(as.Date(c("2012-1-13", "2012-3-3"))) +
  ggtitle("Weekly Sales around SuperBowl Sunday Across all Walmarts in 2012")

# Super Bowl Sunday
ggplot(Walmart) +
  aes(x = Date, y = Weekly_Sales) +
  geom_bar(stat = "identity") +
  xlim(as.Date(c("2010-12-31", "2011-3-04"))) +
  ggtitle("Weekly Sales around SuperBowl Sunday Across all Walmarts in 2011")
```

Variable Selection

```
fit <- lm((Weekly_Sales) ~ Holiday_Flag + Temperature + Unemployment + Fuel_Price, data = Walmart)
summary(fit)
```

Residual Analysis on Weekly Sales

```
nrow(is.na(Walmart)) # Checking to see if there are NA in data



stdres <- stdres(fit)
studres <- studres(fit)
Rstudent <- rstudent(fit)

range(stdres)

barplot(height = stdres, names.arg = 1:6435,
     main = "Standardized Residuals", xlab = "Index",
     ylab = "Standardized Residuals", ylim=c(-5,5))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

barplot(height = studres, names.arg = 1:6435,
     main = "Studentized Residuals", xlab = "Index",
     ylab = "Studentized Residuals", ylim=c(-5, 5))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

quantile <- qt(.05/(2*6435), 6429, lower.tail = F)
quantile

barplot(height = Rstudent, names.arg = 1:6435,
        main = "RStudent Residuals", xlab = "Index",
        ylab = "RStudent Residuals", ylim=c(-5,5))
abline(h=quantile, col = "Red", lwd=2)
abline(h=-quantile, col = "Red", lwd=2)
```

Residual Analysis on CPI

```
stdres1 <- stdres(fit1)
studres1 <- studres(fit1)
Rstudent1 <- rstudent(fit1)

range(stdres1)

barplot(height = stdres1, names.arg = 1:6435,
        main = "Standardized Residuals", xlab = "Index",
        ylab = "Standardized Residuals", ylim=c(-5,5))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

barplot(height = studres1, names.arg = 1:6435,
        main = "Studentized Residuals", xlab = "Index",
        ylab = "Studentized Residuals", ylim=c(-5, 5))
abline(h=3, col = "Red", lwd=2)
abline(h=-3, col = "Red", lwd=2)

quantile <- qt(.05/(2*6435), 6429, lower.tail = F)
quantile

barplot(height = Rstudent1, names.arg = 1:6435,
        main = "RStudent Residuals", xlab = "Index",
        ylab = "RStudent Residuals", ylim=c(-5,5))
```

```
abline(h=quantile, col = "Red", lwd=2)
abline(h=-quantile, col = "Red", lwd=2)
```

Measures of Influence on Weekly Sales

```
myInf <- influence.measures(fit)
myInf

influenceIndexPlot(fit, vars = c("hat"))

influenceIndexPlot(fit, vars=c("Cook"))

dfbetasPlots(fit,intercept=T)

vif(fit)

par(mfrow=c(1,2))
hist(studres, breaks=10, freq=F, col="cornflowerblue",
    cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(fit)

residualPlot(fit, type="rstudent", quadratic=F, col = "dodgerblue",
            pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)
```

Measures of Influence on CPI

```
myInf <- influence.measures(fit1)
myInf

influenceIndexPlot(fit1, vars = c("hat"))
influenceIndexPlot(fit1, vars=c("Cook"))

dfbetasPlots(fit1, intercept=T)

vif(fit1)

par(mfrow=c(1,2))
hist(studres1, breaks=10, freq=F, col="cornflowerblue",
   cex.axis=1.5, cex.lab=1.5, cex.main=2)
qqPlot(fit)

residualPlot(fit1, type="rstudent", quadratic=F, col = "dodgerblue",
        pch=16, cex=1.5, cex.axis=1.5, cex.lab=1.5)
```