ANLY500_Project_AAnand

Abhishek Anand

2022-11-17

Load Packages

Problem Statement

ADD DETAILS for each below: Exploring the variability in retail sales around holiday periods compared to regular periods
Predicting increased sales phenomena for staffing and inventory planning
Provide recommendations for preparations to compete well against other giants
Background on Walmart
Size, volume of sales
Background on Holiday Sales
Overall volume
Share by competitors

Shortcomings of Current analyses:
Describe what summaries on Walmart sales we find
What is missing is the ratios / confidence of what rates of growth occur per holiday period over years?
Describe the business value my analyses will provide

Data Description

DESCRIBE DATA SOURCE

LIST AND DESCRIBE VARIABLES

DESCRIBE HOW THE DATA PRESENTS [Range, Levels, etc.]

Data Screening

Accuracy

```
noerr <- walmart_data
noerr$Holiday_Flag <- as.factor(noerr$Holiday_Flag)
noerr$Date <- as.Date(noerr$Date, tz="UTC", format = "%d-%m-%Y")
summary(noerr)
```

```
##      Store         Date             Weekly_Sales     Holiday_Flag
##  Min.   : 1   Min.   :2010-02-05   Min.   : 209986   0:5985
##  1st Qu.:12   1st Qu.:2010-10-08   1st Qu.: 553350   1: 450
##  Median :23   Median :2011-06-17   Median : 960746
##  Mean   :23   Mean   :2011-06-17   Mean   :1046965
##  3rd Qu.:34   3rd Qu.:2012-02-24   3rd Qu.:1420159
##  Max.   :45   Max.   :2012-10-26   Max.   :3818686
##   Temperature       Fuel_Price         CPI         Unemployment
```

```
##  Min.   : -2.06   Min.   :2.472   Min.   :126.1   Min.   : 3.879
##  1st Qu.: 47.46   1st Qu.:2.933   1st Qu.:131.7   1st Qu.: 6.891
##  Median : 62.67   Median :3.445   Median :182.6   Median : 7.874
##  Mean   : 60.66   Mean   :3.359   Mean   :171.6   Mean   : 7.999
##  3rd Qu.: 74.94   3rd Qu.:3.735   3rd Qu.:212.7   3rd Qu.: 8.622
##  Max.   :100.14   Max.   :4.468   Max.   :227.2   Max.   :14.313
```

We see that the data looks relatively accurate:
- The number of store ranges from 1 to 45, the number of stores the dataset is meant to have data for
- Date ranges from Feb 05, 2010 to Oct 26, 2012
- The Weekly Sales ranges from $209,986 to $3,818,686, which seems normal for a chain of Walmart's size
- There are two levels for the Holiday Flag, i.e., days with holidays and days without
- The temperature ranges from -2F to 100F, which indicates the temperatures are within observed values for the region
- Fuel price ranges from $2.47 to $4.47, also within reason
- CPI ranges from 126 to 227, which seems a little weird. The US CPI for 2010 was approx. 218 (229 for 2012). Because the data seems inaccurate, I will remove this variable fully to keep analysis grounded in reality.
- Unemployment rate ranges from 3.9% to 14.3%, which matches our knowledge of the 2010-2012 period of the rise out of 2008 recession and into economic growth

```
noerr <- noerr[,c(1:6,8)]
```

```
percentmiss <- function(x){sum(is.na(x))/length(x)*100}
missing <- apply(noerr, 1, percentmiss)
table(missing)
```

```
## missing
##    0
## 6435
```

Because there is no missing values, our data is complete and we do not have to apply estimation to extrapolate and fill missing values if MCAR.

```
noerr$year <- as.numeric(format(as.Date(noerr$Date, format="%Y-%m-%d"),"%Y"))
noerr$month <- as.numeric(format(as.Date(noerr$Date, format="%Y-%m-%d"),"%m"))
#noerr$day <- as.numeric(format(as.Date(noerr$Date, format="%Y-%m-%d"),"%d"))
noerr$Holiday_Flag <- as.numeric(noerr$Holiday_Flag)
```

I believe that year and month may be important factor for sales volume. However, because the data is collected each Saturday, the specific day itself would make little sense as a predictor.

Outliers:

```
mahal <- mahalanobis(noerr[ , -c(1:2,4,8:9)],
                     colMeans(noerr[ , -c(1:2,4,8:9)], na.rm=TRUE),
                     cov(noerr[ , -c(1:2,4,8:9)], use ="pairwise.complete.obs"))
cutmahal <- qchisq(1-.001, ncol(noerr[,-c(1:2,4,8:9)]))

badmahal <- as.numeric(mahal >cutmahal)
table(badmahal)
```

```
## badmahal
##    0    1
## 6425   10
```

We see that there are 10 outliers using a general mahalnobis approach.

Let us also try the outlier based on leverage and cooks.

```
noerr$year <- as.factor(noerr$year)
noerr$month <- as.factor(noerr$month)
noerr$Holiday_Flag <- as.factor(noerr$Holiday_Flag)
noerr$Store <- as.factor(noerr$Store)
```

```
model_outlier <- lm(Weekly_Sales ~ Temperature + Fuel_Price + Unemployment,
                    noerr)
```

```
k <- 3
leverage <- hatvalues(model_outlier)
cutleverage <- (2*k+2)/nrow(noerr)
badleverage <- as.numeric(leverage > cutleverage)
table(badleverage)
```

Leverage

```
## badleverage
##    0    1
## 6021  414
```

According to leverage, there are 414 outliers.

```
cooks <- cooks.distance(model_outlier)
cutcooks <- 4 / (nrow(noerr) - k - 1)
badcooks <- as.numeric(cooks > cutcooks)
table(badcooks)
```

```
## badcooks
##    0    1
## 6282  153
```

According to Cooks', there are no 153.

```
totalout <- badmahal + badleverage + badcooks
table(totalout)
```

```
## totalout
##    0    1    2
## 5891  511   33
```

We see that there are 33 outliers across the 3 outlier tests (if we have a cutoff that an outlier is only if two tests result them in outlier). However, upon examination of the outliers, it may make sense, for the purpose of our analysis, to leave these outliers in to understand whether there is actual difference in sales volume between holiday and no-holiday flags. If we removed outliers, we would remove high sales volume data from no-holiday period and low sales volume data from holiday period.

Hence, we will only remove outliers that meet all three outlier tests, which for our dataset, is none.

```
noout_v1 <- subset(noerr, totalout<2) # intermediate outlier removal
noout_v2 <- subset(noerr, totalout<3) # minimal outlier removal
justout <- subset(noerr, totalout>=2)
```

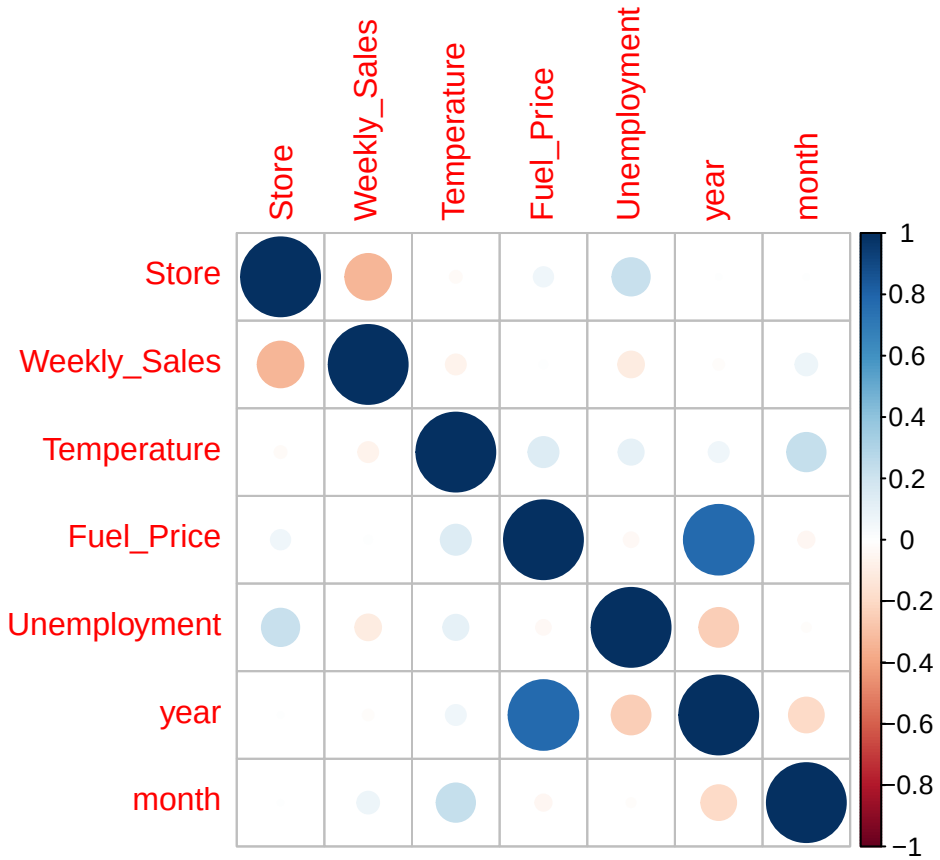
Additivity:

```
noout_v2$year <- as.numeric(noout_v2$year)
noout_v2$month <- as.numeric(noout_v2$month)
noout_v2$Holiday_Flag <- as.numeric(noout_v2$Holiday_Flag)
noout_v2$Store <- as.numeric(noout_v2$Store)
```

```
cor(noout_v2[,-c(2,4)])
```

```
##                      Store Weekly_Sales Temperature   Fuel_Price Unemployment
## Store          1.00000000 -0.335332015 -0.02265908  0.060022955   0.22353127
## Weekly_Sales -0.33533201  1.000000000 -0.06381001  0.009463786  -0.10617609
## Temperature  -0.02265908 -0.063810013  1.00000000  0.144981806   0.10115786
## Fuel_Price    0.06002295  0.009463786  0.14498181  1.000000000  -0.03468374
## Unemployment  0.22353127 -0.106176090  0.10115786 -0.034683745   1.00000000
## year          0.00000000 -0.018377543  0.06426923  0.779470302  -0.24181349
## month         0.00000000  0.076143320  0.23586176 -0.042155900  -0.01274559
##                      year       month
## Store          0.00000000  0.00000000
## Weekly_Sales -0.01837754  0.07614332
## Temperature   0.06426923  0.23586176
## Fuel_Price    0.77947030 -0.04215590
## Unemployment -0.24181349 -0.01274559
## year          1.00000000 -0.19446452
## month        -0.19446452  1.00000000
```

```
corrplot(cor(noout_v2[,-c(2,4)]))
```

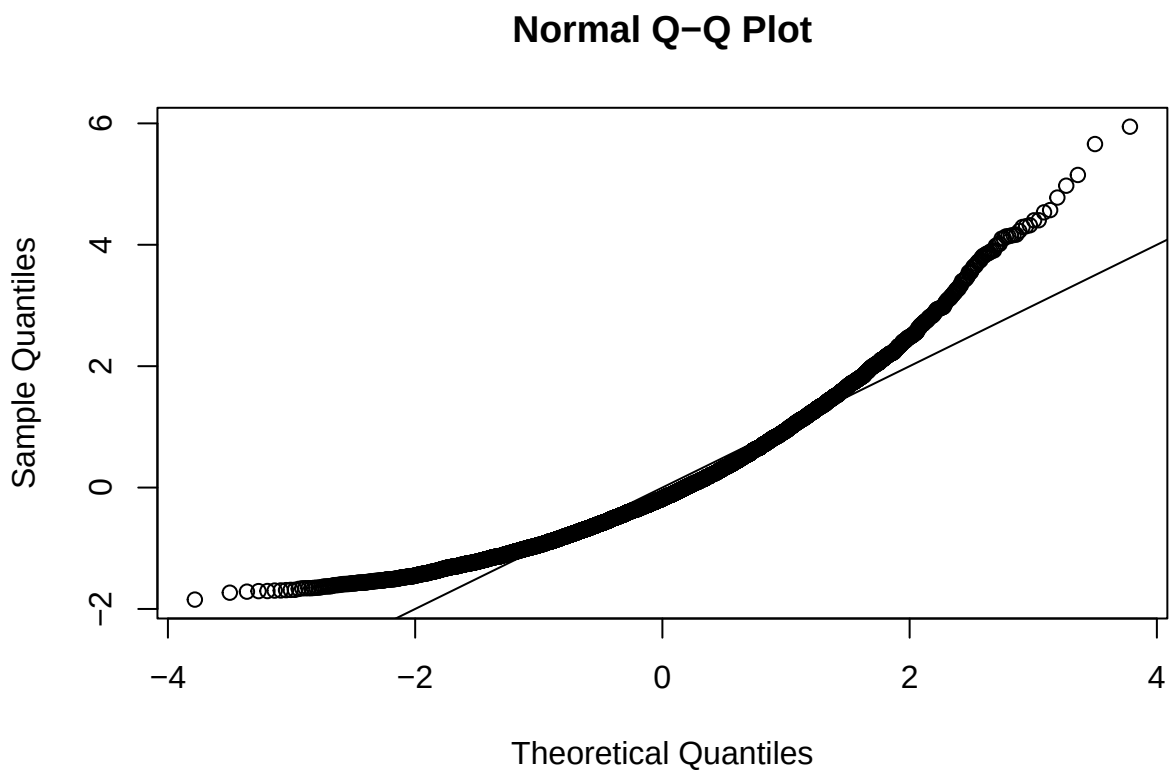


The assumption of Additivity is met because of lack of colinearity.

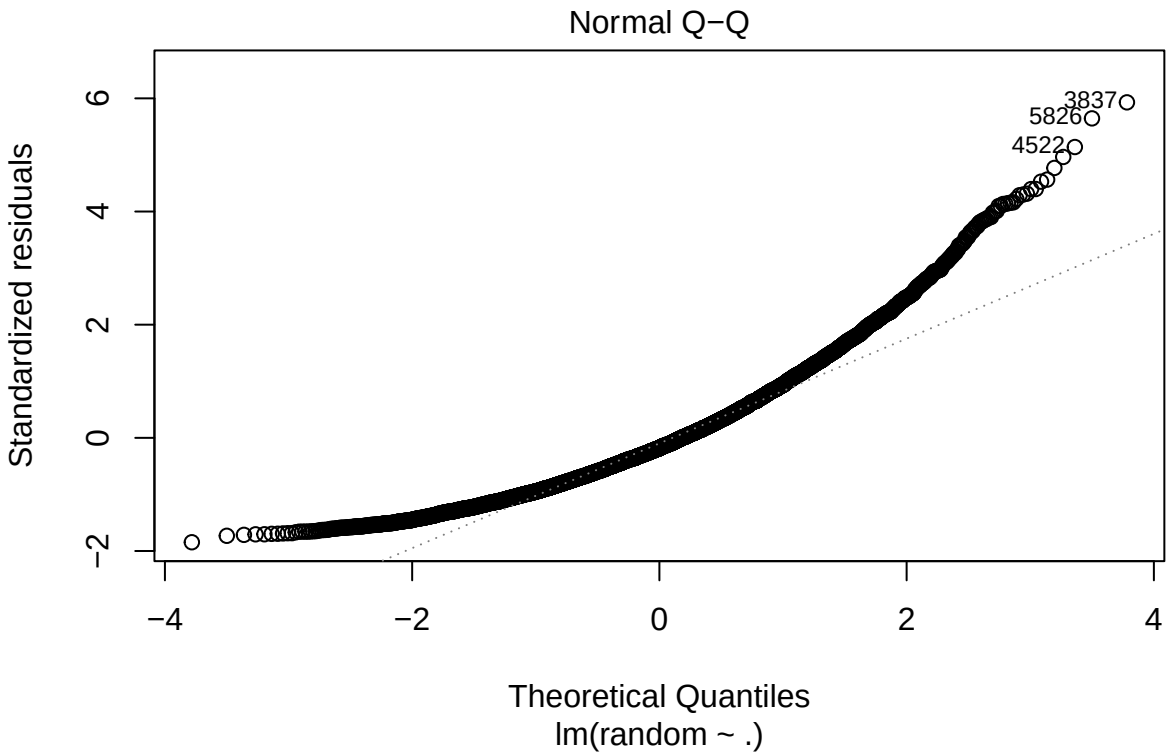
Linearity:

```
random <- rchisq(nrow(noout_v2), 7)
fake <- lm(random ~ .,
           data = noout_v2)
standardized <- rstudent(fake)
fitvalues <- scale(fake$fitted.values)
```

```
{qqnorm(standardized)
abline(0,1)}
```



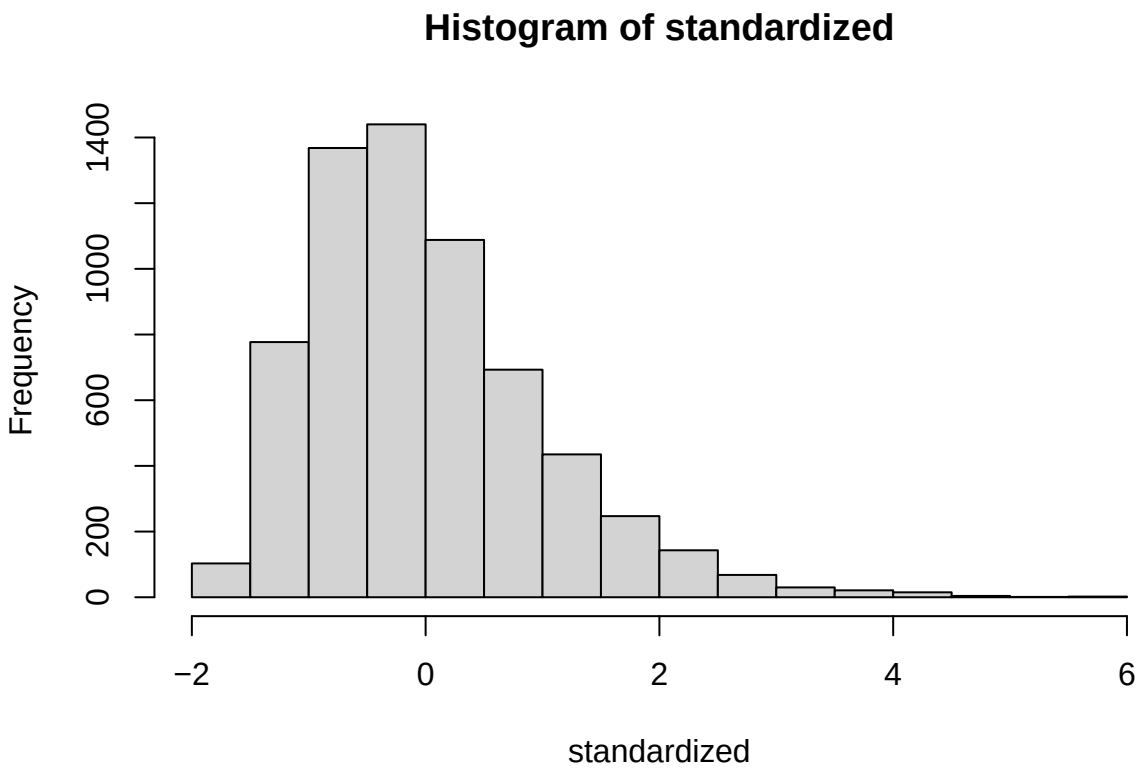

```
plot(fake, 2)
```



The linearity assumption is met. The data lays primarily on the line between -2 to 2.

Normality

```
hist(standardized, breaks = 15)
```



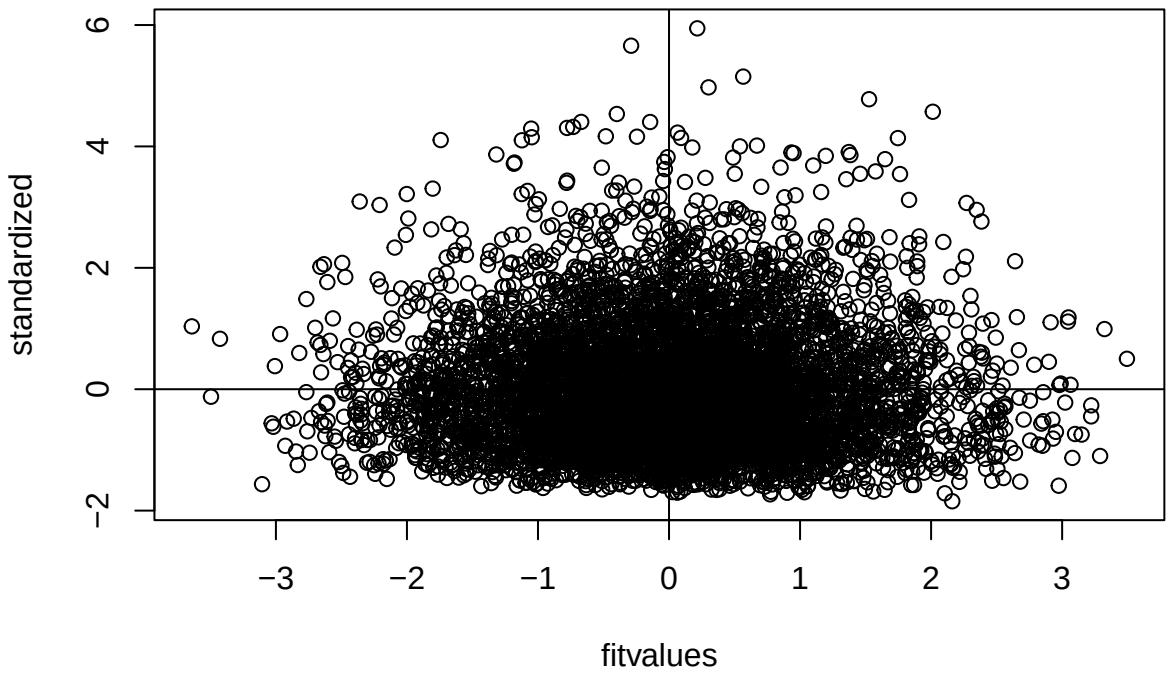

```
round(mean(standardized),5)
```

```
## [1] 8e-05
```

We see that the histogram of standardized values is centered around the mean of 0, with much of the spread contained between -2 and 2. Only a few values cause a tail towards the positive x-axis, but the distribution is still quite normal-looking. Hence, we can say that the assumption for normality is met.

Homogeneity/Heteroscedasticity

```
{plot(fitvalues, standardized)
abline(0,0)
abline(v = 0)}
```



The spread of data looks even between -2 and 2 across the x- and y-axes. Hence, the assumption of homogeneity and homoscedascity appears to be met.

```
walmart_clean <- noout_v2
```

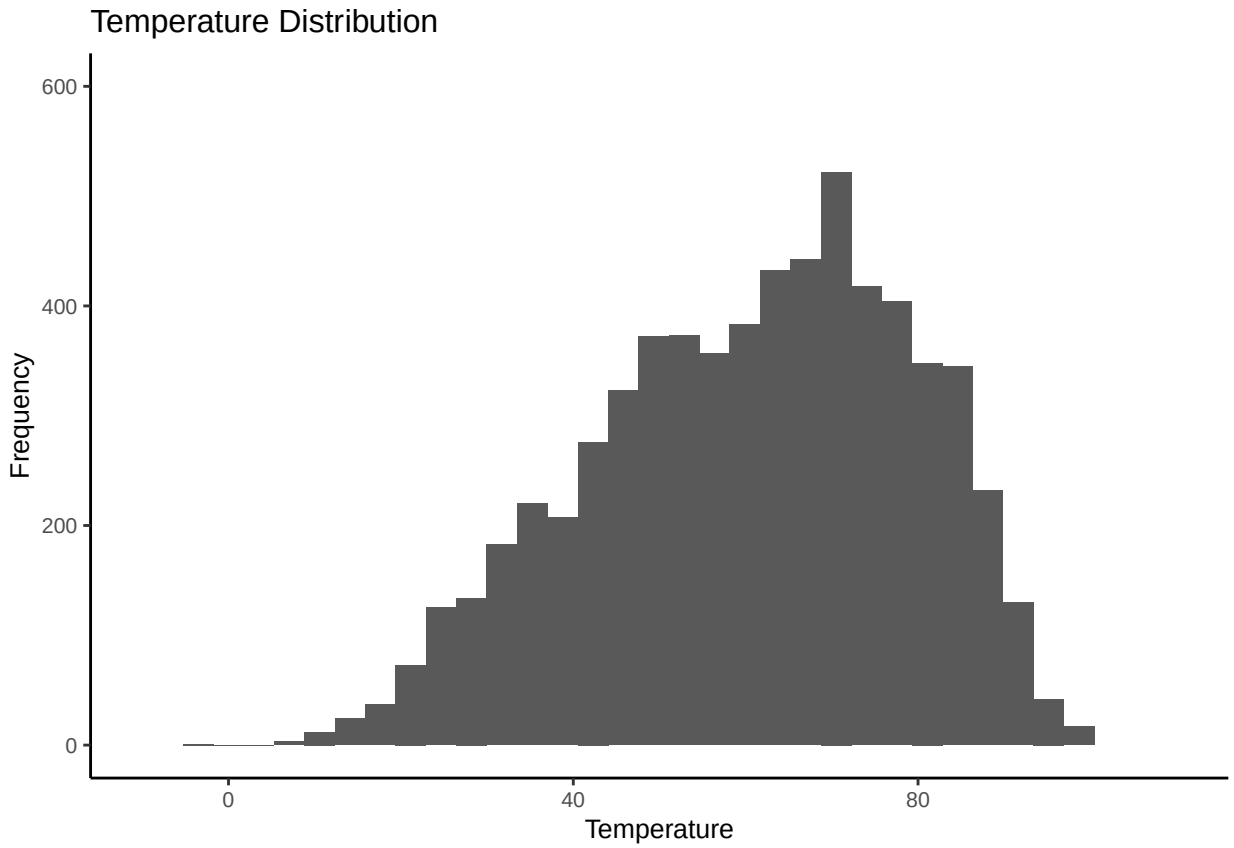
Exploratory Data Analyses

```
walmart_clean$Holiday_Flag <- factor(walmart_clean$Holiday_Flag,
                                     levels = c(1,2),
                                     labels = c("No_Holiday", "Holiday"))
walmart_clean$Store <- as.factor(walmart_clean$Store)
walmart_clean$year <- factor(walmart_clean$year,
                             levels = c(1,2,3),
                             labels = c(2010,
                                        2011,
                                        2012)) #break out date into year
walmart_clean$month <- factor(walmart_clean$month,
                             levels = c(1,2,3,4,5,6,7,8,9,10,11,12),
                             labels = c("Jan",
                                        "Feb",
                                        "Mar",
                                        "Apr",
                                        "May",
                                        "Jun",
                                        "Jul",
                                        "Aug",
```

```
                                        "Sep",
                                        "Oct",
                                        "Nov",
                                        "Dec")) #break out date into month
```

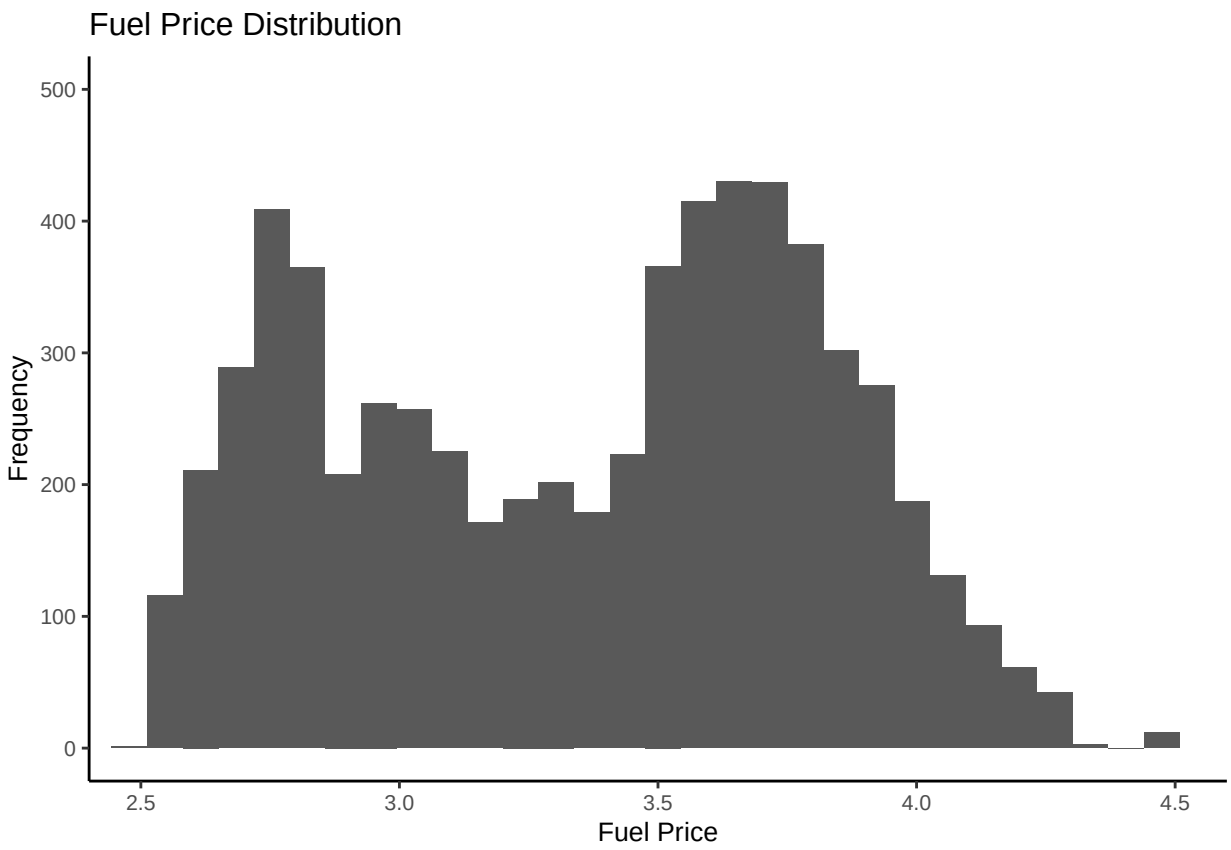
Data Distribution Plots

```
walmart_clean%>%
  ggplot(aes(Temperature))+
  geom_histogram(bins=30)+
  labs(title = 'Temperature Distribution',
       y='Frequency',
       x='Temperature')+
  cleanup +
  coord_cartesian(xlim = c(-10,110), ylim = c(0,600))
```



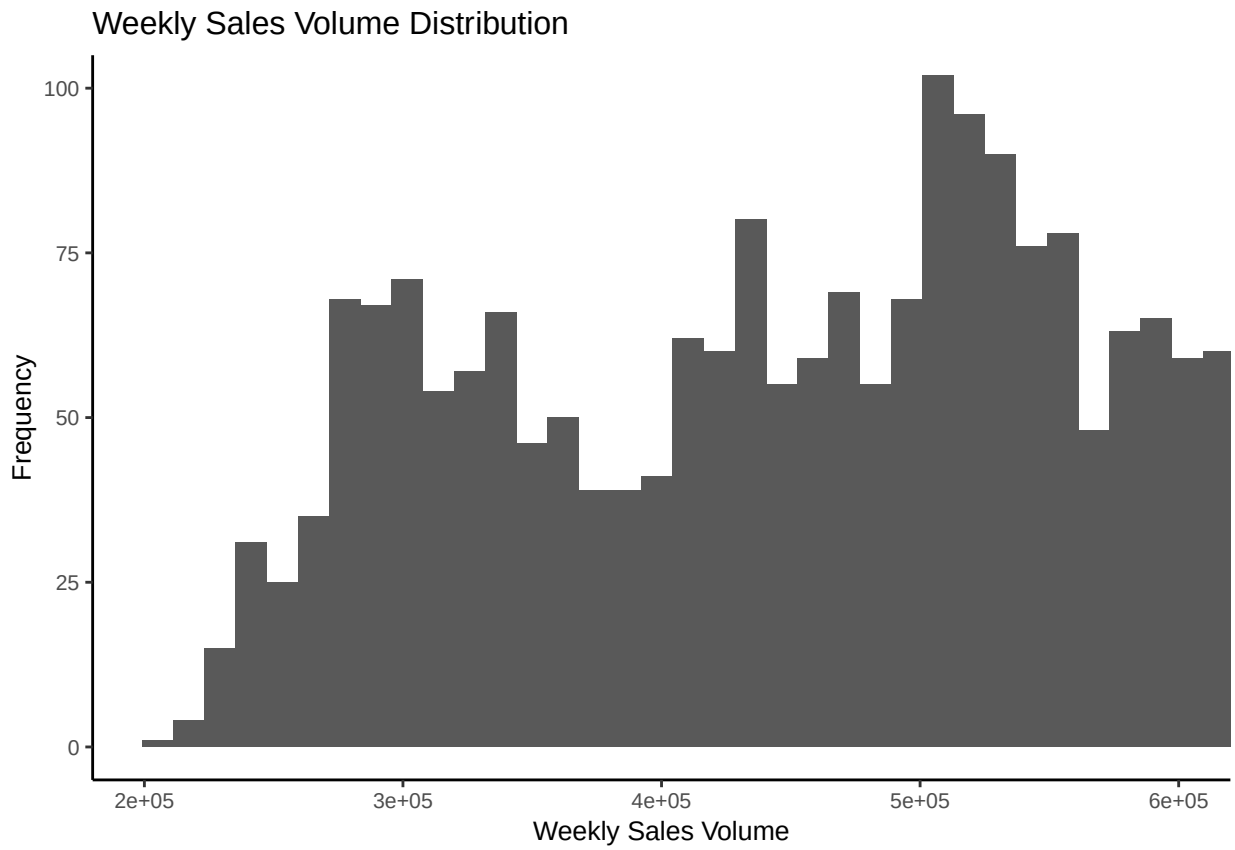
Temperature

```
walmart_clean%>%
  ggplot(aes(Fuel_Price))+
  geom_histogram(bins=30)+
  labs(title = 'Fuel Price Distribution',
       y='Frequency',
       x='Fuel Price')+
  cleanup +
  coord_cartesian(xlim = c(2.5,4.5), ylim = c(0,500))
```



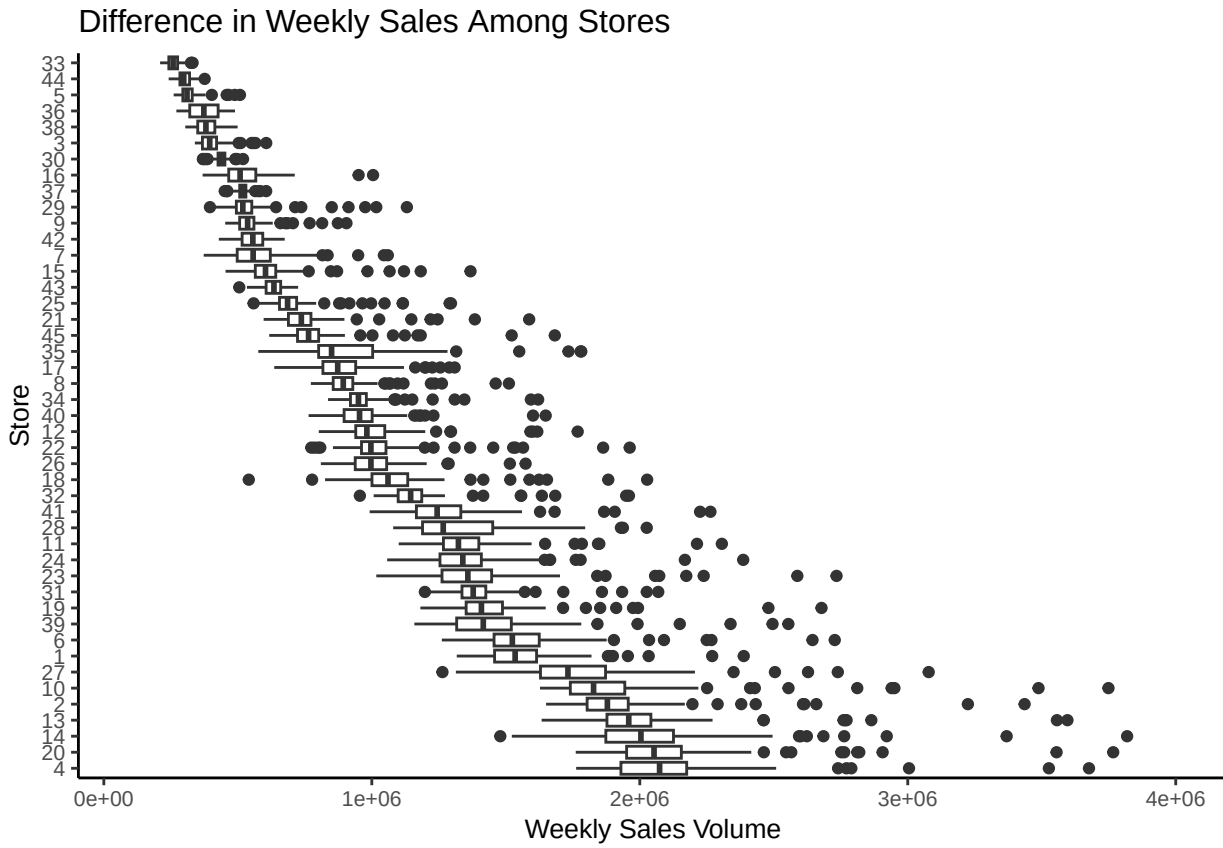
Fuel Price

```
walmart_clean%>%
  ggplot(aes(Weekly_Sales))+
  geom_histogram(bins=300)+
  labs(title = 'Weekly Sales Volume Distribution',
       y='Frequency',
       x='Weekly Sales Volume')+
  cleanup +
  coord_cartesian(xlim = c(200000,600000), ylim = c(0,100))
```



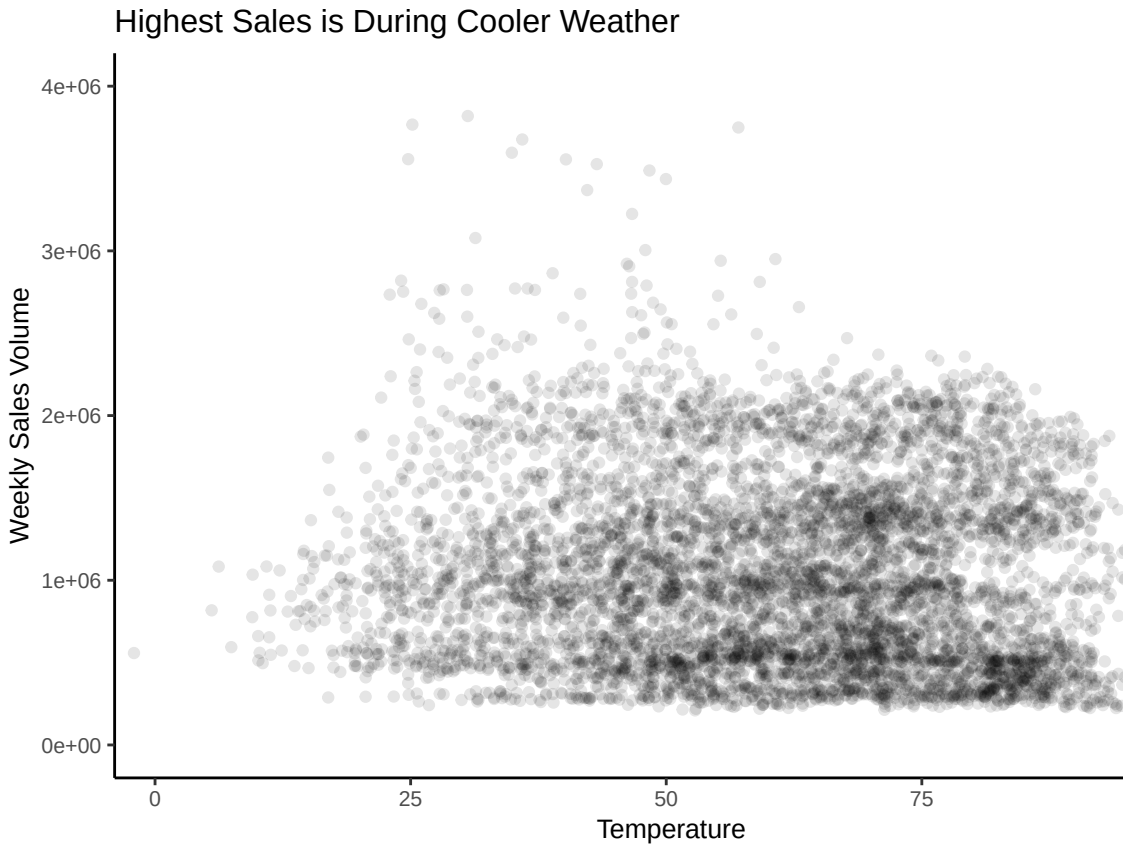
Relationship Plots

```
walmart_clean%>%
  ggplot(aes(Weekly_Sales, reorder(Store,
                                   FUN = median, Weekly_Sales,
                                   decreasing = TRUE)))+
  geom_boxplot()+
  labs(title = 'Difference in Weekly Sales Among Stores',
       x='Weekly Sales Volume',
       y='Store')+
  cleanup +
  coord_cartesian(xlim = c(100000,4000000))
```



Sales By Store

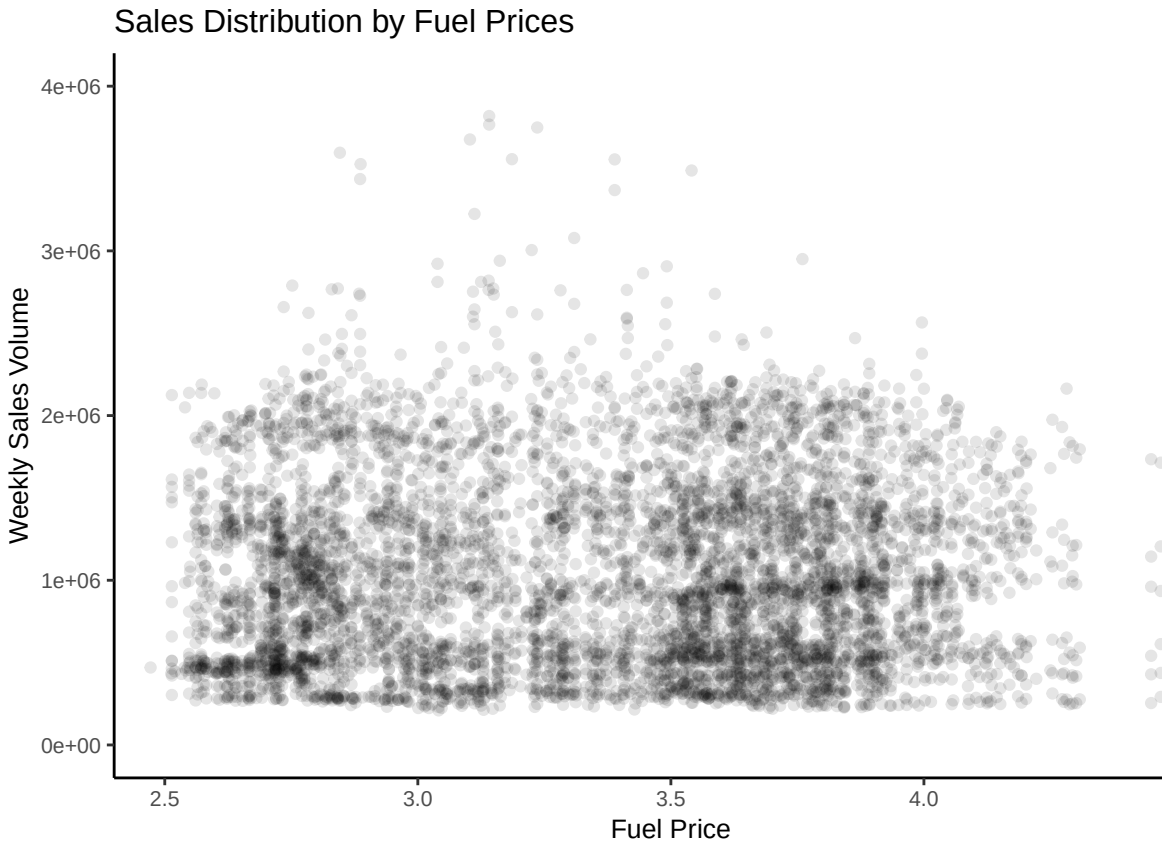
```
walmart_clean%>%
  ggplot(aes(Temperature, Weekly_Sales))+
  geom_point(alpha = 0.1) +
  #geom_histogram(bins=75)+
  labs(title = 'Highest Sales is During Cooler Weather',
       y='Weekly Sales Volume',
       x='Temperature')+
  cleanup +
  coord_cartesian(xlim = c(1,100), ylim = c(0,4000000))
```



Sales By Temperature

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

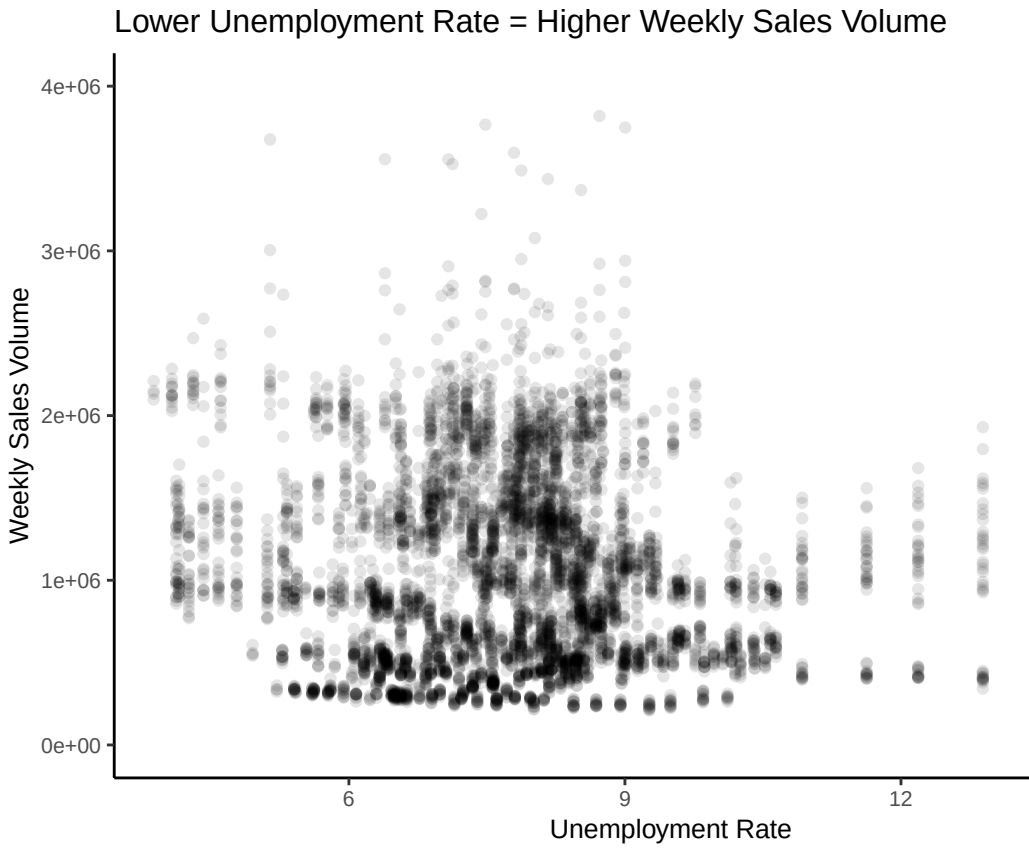
```
walmart_clean%>%
  ggplot(aes(Fuel_Price, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Sales Distribution by Fuel Prices',
       y='Weekly Sales Volume',
       x='Fuel Price')+
  cleanup +
  coord_cartesian(xlim = c(2.5,4.5), ylim = c(0,4000000))
```



Sales By Fuel Price

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

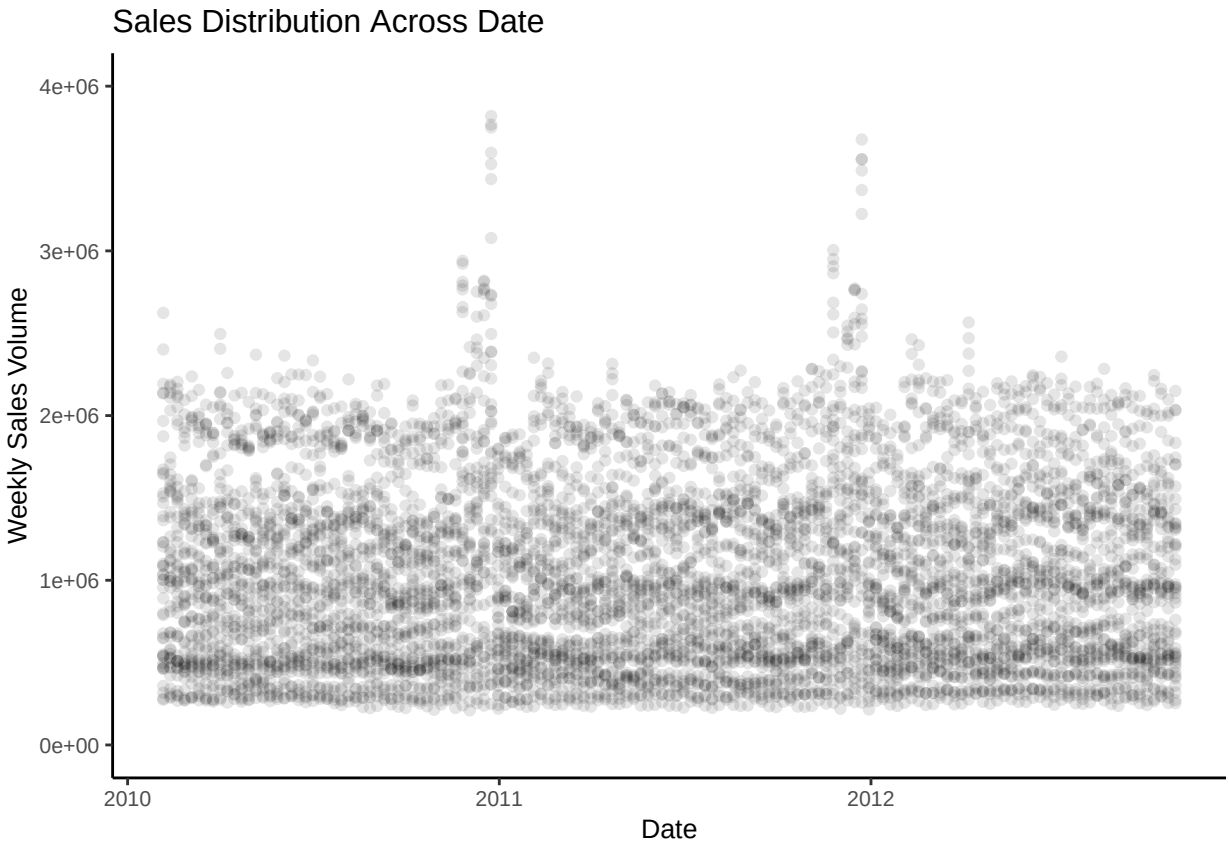
```
walmart_clean%>%
  ggplot(aes(Unemployment, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Lower Unemployment Rate = Higher Weekly Sales Volume',
       y='Weekly Sales Volume',
       x='Unemployment Rate')+
  cleanup +
  coord_cartesian(xlim = c(4,15), ylim = c(0,4000000))
```



Sales by Unemployment Rate

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

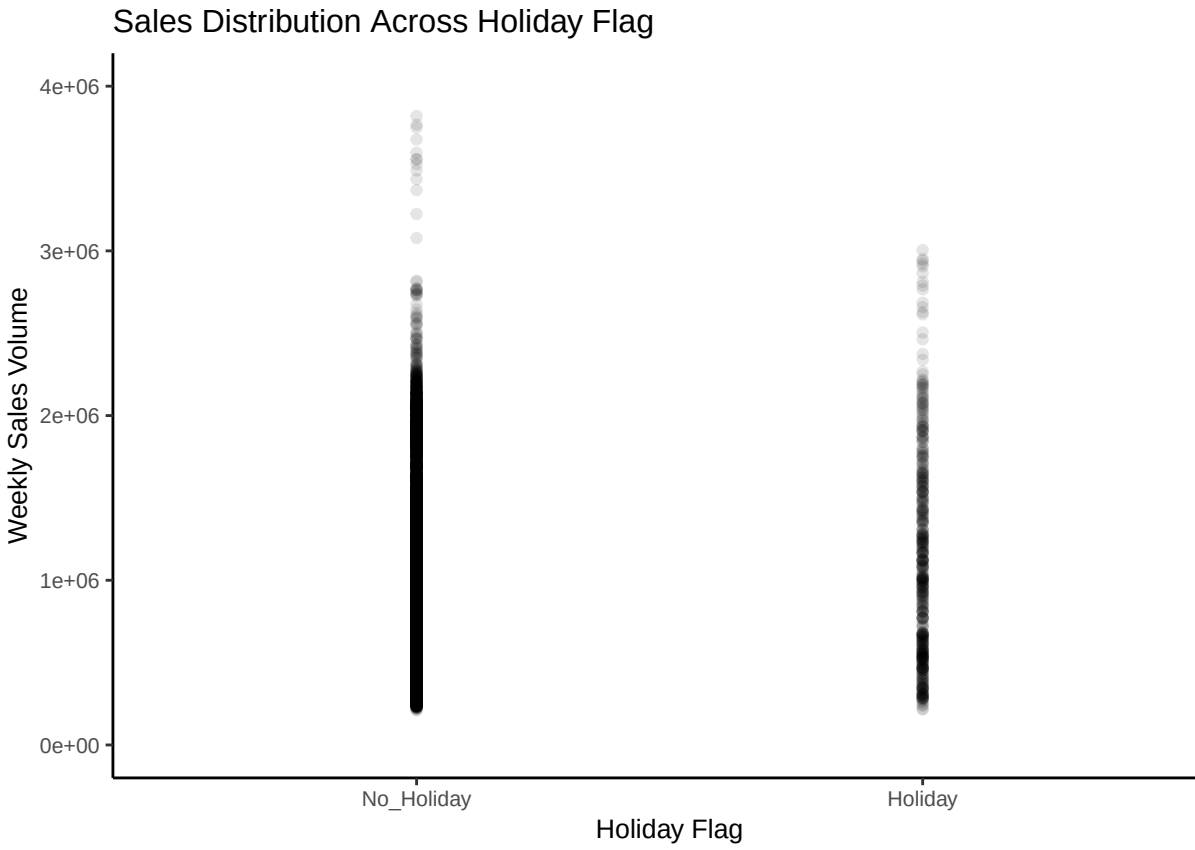
```
walmart_clean%>%
  ggplot(aes(Date, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Sales Distribution Across Date',
       y='Weekly Sales Volume',
       x='Date')+
  cleanup +
  coord_cartesian(ylim = c(0,4000000))
```



Sales by Date

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

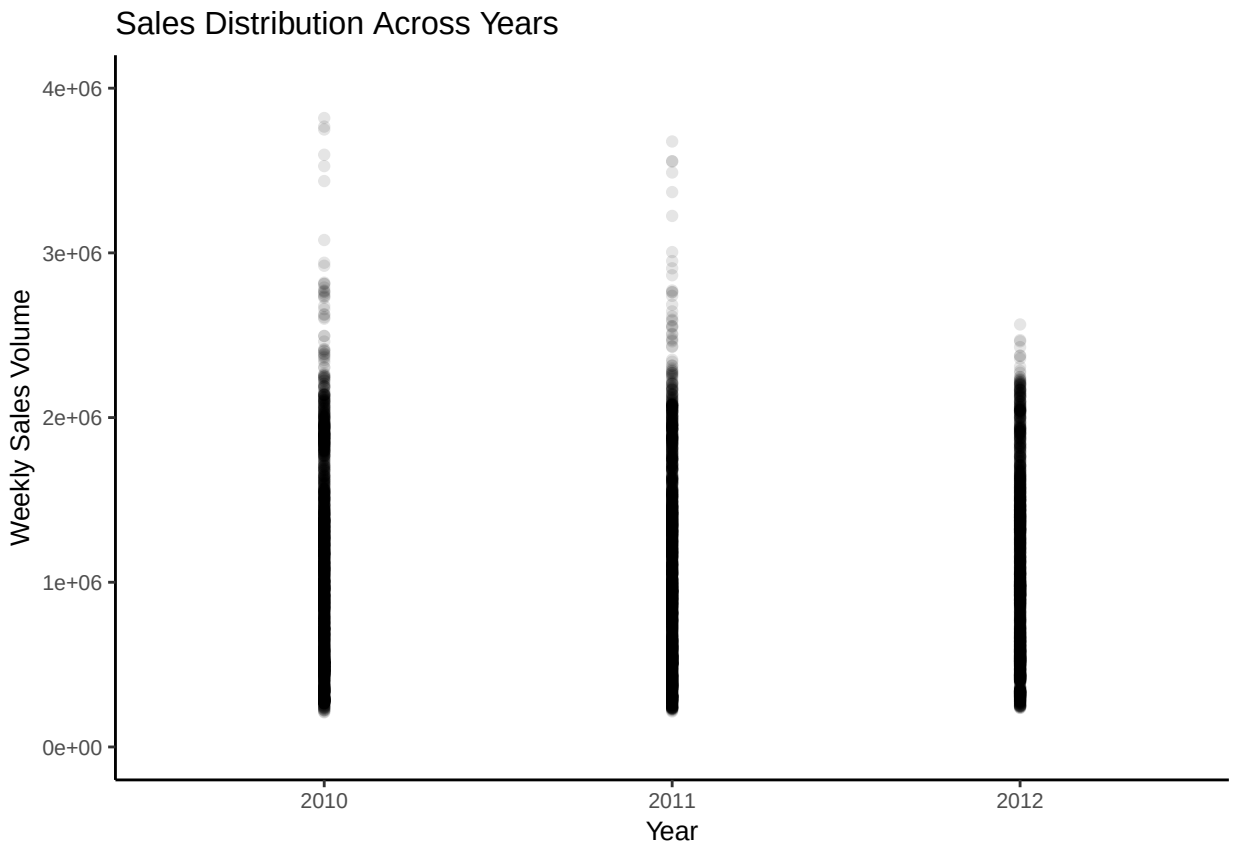
```
walmart_clean%>%
  ggplot(aes(Holiday_Flag, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Sales Distribution Across Holiday Flag',
       y='Weekly Sales Volume',
       x='Holiday Flag')+
  cleanup +
  coord_cartesian(ylim = c(0,4000000))
```



Sales by Holiday

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

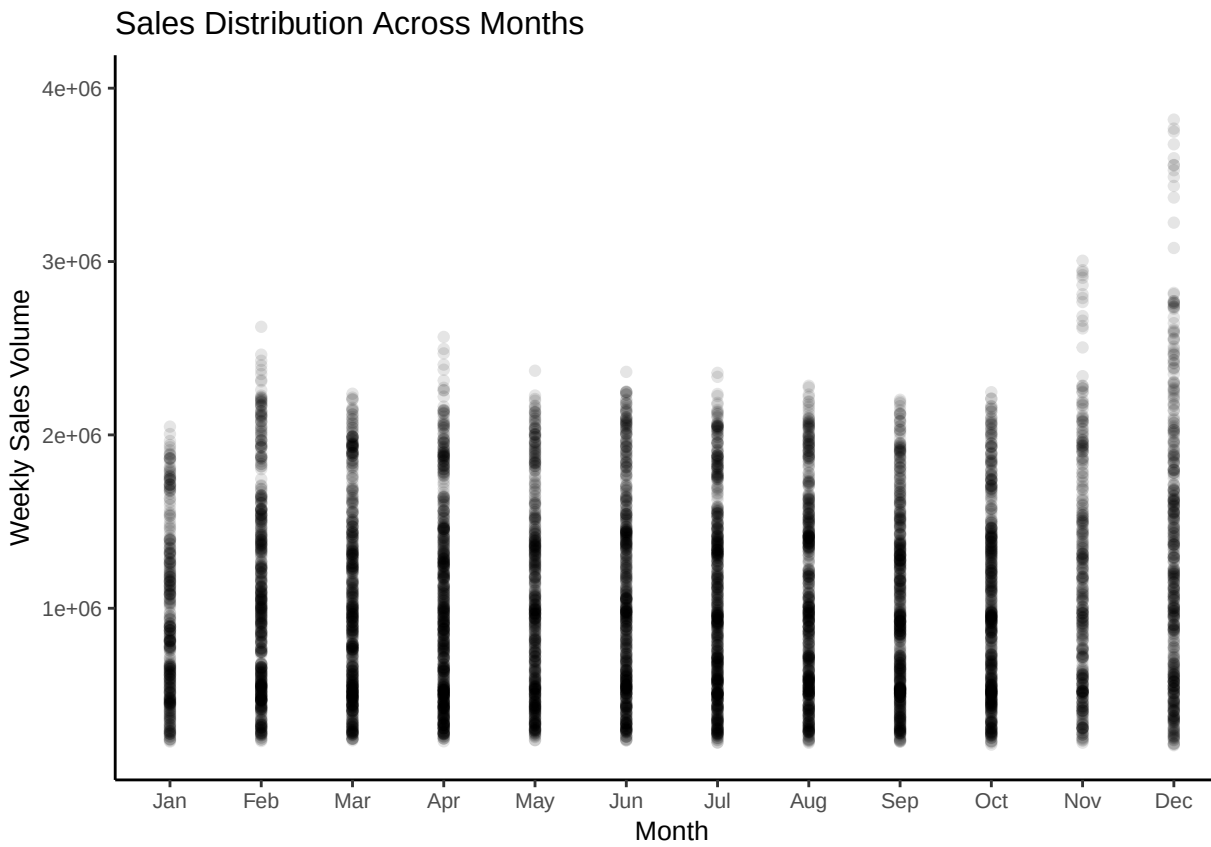
```
walmart_clean%>%
  ggplot(aes(year, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Sales Distribution Across Years',
       y='Weekly Sales Volume',
       x='Year')+
  cleanup +
  coord_cartesian(ylim = c(0,4000000))
```



Sales by Year

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

```
walmart_clean%>%
  ggplot(aes(month, Weekly_Sales))+
  #geom_histogram(bins=70)+
  geom_point(alpha = 0.1) +
  labs(title = 'Sales Distribution Across Months',
       y='Weekly Sales Volume',
       x='Month')+
  cleanup +
  coord_cartesian(ylim = c(200000,4000000))
```



Sales by Month

```
#+ geom_smooth(method = 'lm', se=FALSE, color='navyblue')
```

Means

```
summary(walmart_clean$Weekly_Sales)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  209986  553350  960746 1046965 1420159 3818686
```

```
summary(walmart_clean$Weekly_Sales[walmart_clean$Holiday_Flag=="No_Holiday"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  209986  551378  956211 1041256 1414344 3818686
```

```
summary(walmart_clean$Weekly_Sales[walmart_clean$Holiday_Flag=="Holiday"])
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  215359  575866 1018538 1122888 1555213 3004702
```

Plots

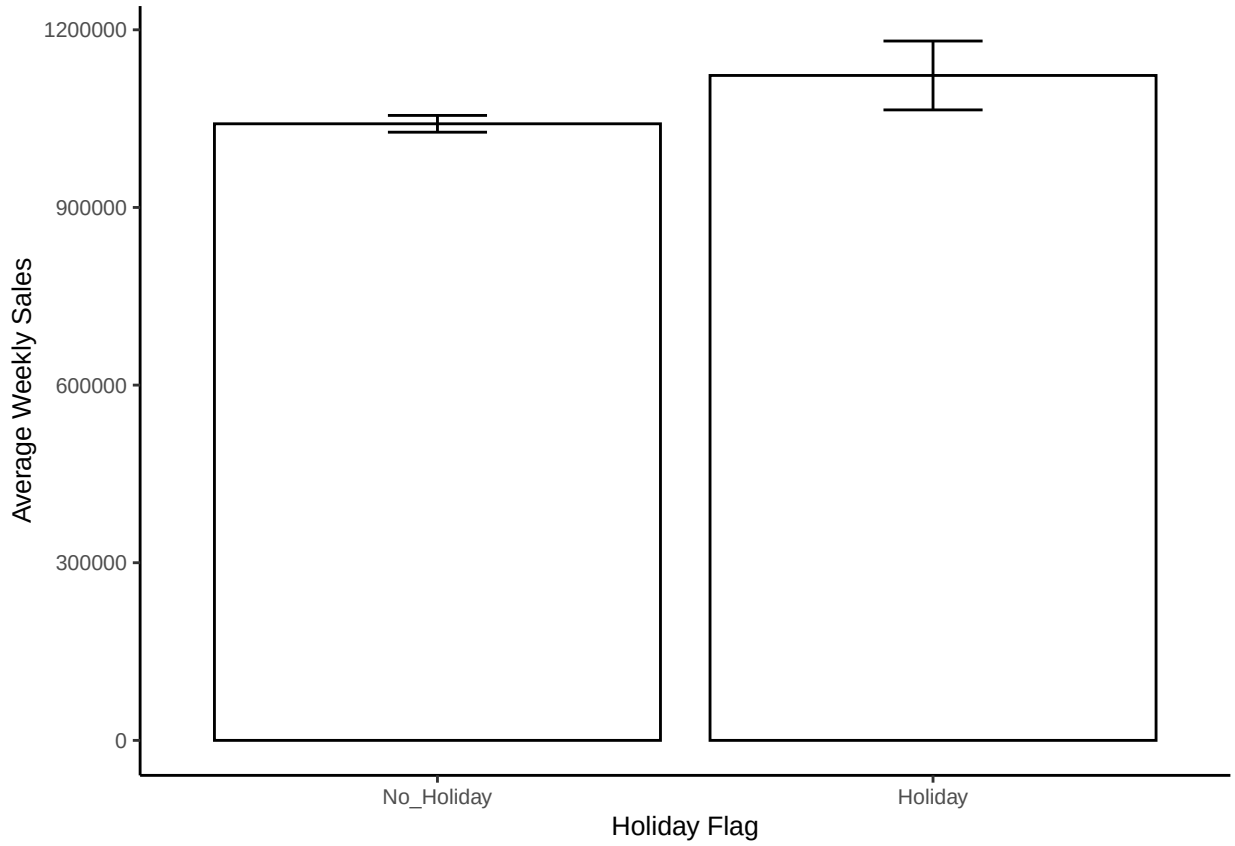
```
bargraph <- ggplot(walmart_clean, aes(Holiday_Flag, Weekly_Sales))

bargraph +
  cleanup +
  stat_summary(fun.y = mean,
```

```
               geom = "bar",
               fill = "White",
               color = "Black") +
  stat_summary(fun.data = mean_cl_normal,
               geom = "errorbar",
               width = .2,
               position = "dodge") +
  xlab("Holiday Flag") +
  ylab("Average Weekly Sales")
```

Holiday vs. No Holiday

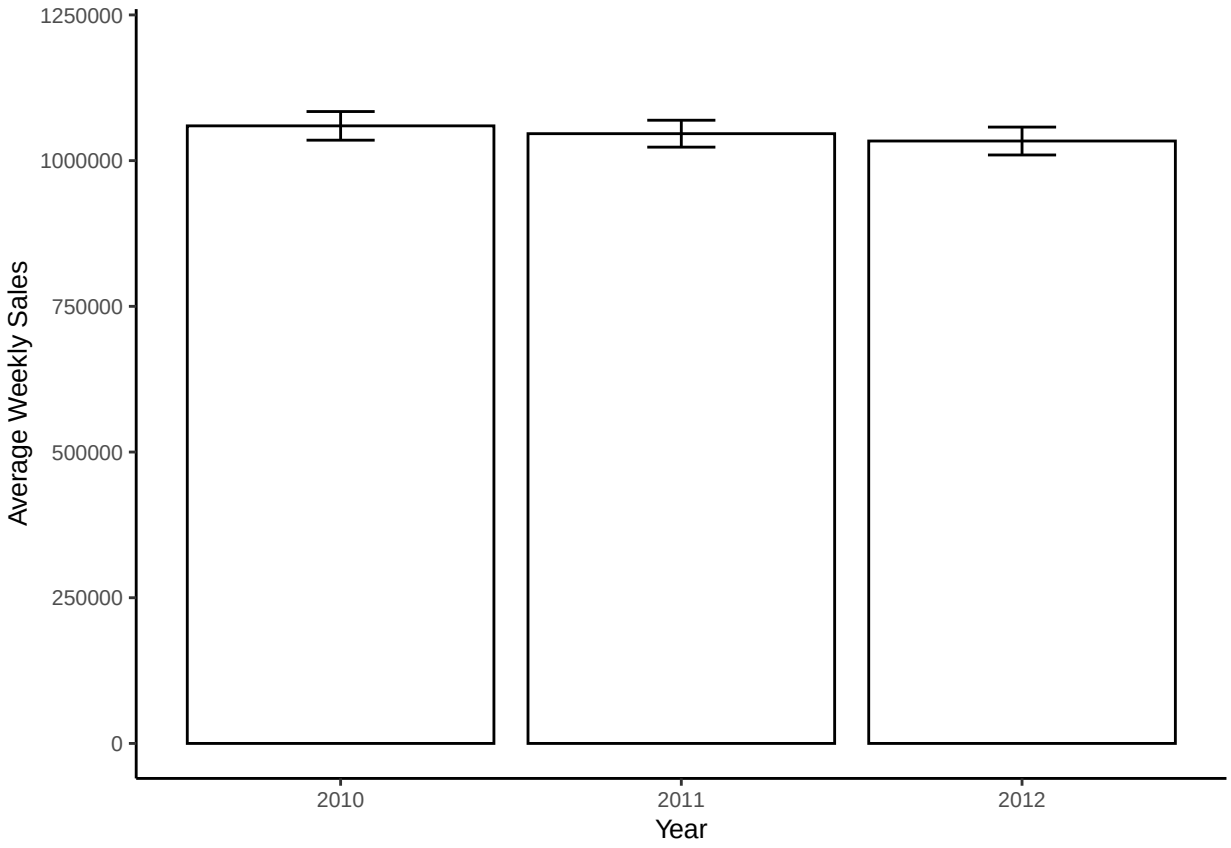
```
## Warning: The `fun.y` argument of `stat_summary()` is deprecated as of ggplot2 3.3.0.
## i Please use the `fun` argument instead.
```




```
bargraph2 <- ggplot(walmart_clean, aes(year, Weekly_Sales))

bargraph2 +
  cleanup +
  stat_summary(fun.y = mean,
               geom = "bar",
               fill = "White",
               color = "Black") +
  stat_summary(fun.data = mean_cl_normal,
               geom = "errorbar",
```

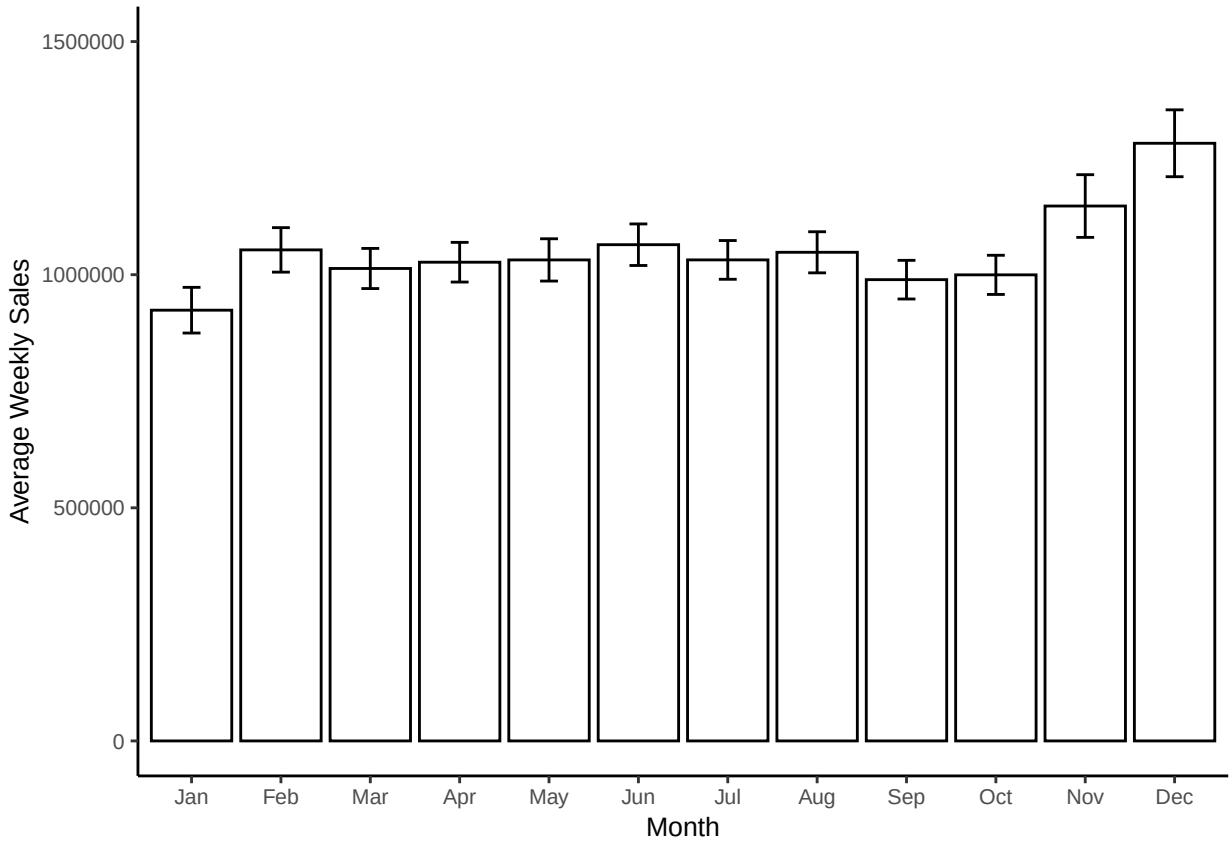
```
              width = .2,
              position = "dodge") +
  xlab("Year") +
  ylab("Average Weekly Sales") +
  coord_cartesian(ylim = c(0,1200000))
```



Years

```
bargraph3 <- ggplot(walmart_clean, aes(month, Weekly_Sales))

bargraph3 +
  cleanup +
  stat_summary(fun.y = mean,
               geom = "bar",
               fill = "White",
               color = "Black") +
  stat_summary(fun.data = mean_cl_normal,
               geom = "errorbar",
               width = .2,
               position = "dodge") +
  xlab("Month") +
  ylab("Average Weekly Sales") +
  coord_cartesian(ylim = c(0,1500000))
```



Months

Technical Approach

Outline the steps I will follow:

Collinearity [NOT USEFUL]

```
#walmart_clean$Holiday_Flag <- as.numeric(walmart_clean$Holiday_Flag)
#walmart_clean$Store <- as.numeric(walmart_clean$Store)
#walmart_clean$year <- as.numeric(walmart_clean$year)
#walmart_clean$month <- as.numeric(walmart_clean$month)
```

```
#cocor(~Weekly_Sales + year | Weekly_Sales + Fuel_Price,
#      data = walmart_clean)
```

Weekly Sales and Year vs. Weekly Sales and Fuel Price [NOT USEFUL?]

```
#new <- subset(walmart_clean, Holiday_Flag == 1)
#old <- subset(walmart_clean, Holiday_Flag == 2)
#ind_data <- list(new,old)
#cocor(~ Weekly_Sales + Fuel_Price | Weekly_Sales + Fuel_Price,
#      data = ind_data)
```

Holiday Flag vs. No Holiday Flag [NOT USEFUL]

```
#pcor(walmart_clean[,-c(2)], method = "pearson")
```

Partial Correlations [NOT USEFUL]

Linear Models - Hierarchical Regression VERSION 1

I believe that certain known variables have a greater effect on weekly sales. I will use stepwise regression, and carry out ANOVA to test the significance after each step as outlined below: - First variable: Unemployment - Second variable: Fuel_Price - Third Variable: Temperature - Fourth Variable: Month - Fifth Variable: Holiday_Flag

STEP 1:

```
model_hr_v1_1 <- lm(Weekly_Sales ~ Unemployment,
                data = walmart_clean)
summary(model_hr_v1_1)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -844415 -481049  -69658  369648 2794876
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1302485      30645  42.503   <2e-16 ***
## Unemployment   -31944       3730  -8.564   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 561200 on 6433 degrees of freedom
## Multiple R-squared:  0.01127,    Adjusted R-squared:  0.01112
## F-statistic: 73.35 on 1 and 6433 DF,  p-value: < 2.2e-16
```

We see that effect of unemployment on weekly sales is large and significant.

```
model_hr_v1_2 <- lm(Weekly_Sales ~ Unemployment + Fuel_Price,
                data = walmart_clean)
summary(model_hr_v1_2)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment + Fuel_Price, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -842594 -482167  -68252  370749 2796381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1278101      60586  21.096   <2e-16 ***
## Unemployment   -31883       3732  -8.543   <2e-16 ***
## Fuel_Price       7117      15253   0.467    0.641
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 561300 on 6432 degrees of freedom
## Multiple R-squared:  0.01131,    Adjusted R-squared:  0.011
## F-statistic: 36.78 on 2 and 6432 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v1_1, model_hr_v1_2)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Unemployment
## Model 2: Weekly_Sales ~ Unemployment + Fuel_Price
##   Res.Df        RSS Df  Sum of Sq      F Pr(>F)
## 1   6433 2.0262e+15
## 2   6432 2.0261e+15  1 6.8575e+10 0.2177 0.6408
```

```
model_hr_v1_3 <- lm(Weekly_Sales ~ Unemployment + Fuel_Price + Temperature,
                data = walmart_clean)
summary(model_hr_v1_3)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment + Fuel_Price + Temperature,
##     data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -886407 -484304  -81258  380499 2746024
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1333071.1    61759.2  21.585  < 2e-16 ***
## Unemployment  -30100.5     3748.6  -8.030 1.15e-15 ***
## Fuel_Price     17303.3    15403.5   1.123    0.261
## Temperature    -1705.2      385.1  -4.428 9.66e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 560400 on 6431 degrees of freedom
## Multiple R-squared:  0.01431,    Adjusted R-squared:  0.01385
## F-statistic: 31.13 on 3 and 6431 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v1_1, model_hr_v1_2, model_hr_v1_3)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Unemployment
## Model 2: Weekly_Sales ~ Unemployment + Fuel_Price
## Model 3: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6433 2.0262e+15
## 2   6432 2.0261e+15  1 6.8575e+10  0.2183    0.6403
## 3   6431 2.0200e+15  1 6.1590e+12 19.6086 9.661e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_hr_v1_4 <- lm(Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month,
                data = walmart_clean)
summary(model_hr_v1_4)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment + Fuel_Price + Temperature +
##     month, data = walmart_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1097180  -472949   -83257   377809  2535871
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1134918.0    68344.8  16.606  < 2e-16 ***
## Unemployment  -30578.1     3787.3  -8.074 8.07e-16 ***
## Fuel_Price     32912.6    15762.5   2.088 0.036834 *
## Temperature    -2025.0      641.1  -3.159 0.001592 **
## monthFeb      142834.8    37853.2   3.773 0.000162 ***
## monthMar      114500.1    38080.0   3.007 0.002650 **
## monthApr      139731.6    39025.3   3.581 0.000345 ***
## monthMay      156930.0    42032.1   3.734 0.000190 ***
## monthJun      212964.3    44350.4   4.802 1.61e-06 ***
## monthJul      193082.2    45533.5   4.240 2.26e-05 ***
## monthAug      201912.9    45853.5   4.403 1.08e-05 ***
## monthSep      130706.0    43434.4   3.009 0.002629 **
## monthOct      117658.5    40305.5   2.919 0.003522 **
## monthNov      257486.4    42138.6   6.110 1.05e-09 ***
## monthDec      373225.8    39341.0   9.487  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 555500 on 6420 degrees of freedom
## Multiple R-squared:  0.03316,    Adjusted R-squared:  0.03106
## F-statistic: 15.73 on 14 and 6420 DF,  p-value: < 2.2e-16
```

```r
anova(model_hr_v1_1, model_hr_v1_2, model_hr_v1_3, model_hr_v1_4)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Unemployment
## Model 2: Weekly_Sales ~ Unemployment + Fuel_Price
## Model 3: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature
## Model 4: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6433 2.0262e+15
## 2   6432 2.0261e+15  1 6.8575e+10  0.2222    0.6374
## 3   6431 2.0200e+15  1 6.1590e+12 19.9568 8.057e-06 ***
## 4   6420 1.9813e+15 11 3.8634e+13 11.3805 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
model_hr_v1_5 <- lm(Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month + Holiday_Flag,
                data = walmart_clean)
```

```
summary(model_hr_v1_5)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment + Fuel_Price + Temperature +
##     month + Holiday_Flag, data = walmart_clean)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1120213  -472457   -83259   378095  2541871
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1134008.2    68351.6  16.591  < 2e-16 ***
## Unemployment         -30600.4     3787.4  -8.080 7.70e-16 ***
## Fuel_Price            33012.9    15762.9   2.094 0.036268 *
## Temperature           -2004.7      641.4  -3.125 0.001784 **
## monthFeb             135585.4    38588.5   3.514 0.000445 ***
## monthMar             114240.3    38081.1   3.000 0.002711 **
## monthApr             139303.6    39028.0   3.569 0.000361 ***
## monthMay             156335.1    42036.8   3.719 0.000202 ***
## monthJun             212186.6    44357.9   4.784 1.76e-06 ***
## monthJul             192223.8    45542.4   4.221 2.47e-05 ***
## monthAug             201051.4    45862.4   4.384 1.18e-05 ***
## monthSep             123300.3    44104.6   2.796 0.005195 **
## monthOct             117160.4    40309.0   2.907 0.003667 **
## monthNov             250009.0    42842.3   5.836 5.62e-09 ***
## monthDec             367395.7    39800.4   9.231  < 2e-16 ***
## Holiday_FlagHoliday   28957.6    29943.4   0.967 0.333542
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 555500 on 6419 degrees of freedom
## Multiple R-squared:  0.03331,    Adjusted R-squared:  0.03105
## F-statistic: 14.74 on 15 and 6419 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v1_1, model_hr_v1_2, model_hr_v1_3, model_hr_v1_4, model_hr_v1_5)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Unemployment
## Model 2: Weekly_Sales ~ Unemployment + Fuel_Price
## Model 3: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature
## Model 4: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month
## Model 5: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month +
##     Holiday_Flag
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6433 2.0262e+15
## 2   6432 2.0261e+15  1 6.8575e+10  0.2222    0.6374
## 3   6431 2.0200e+15  1 6.1590e+12 19.9565 8.058e-06 ***
## 4   6420 1.9813e+15 11 3.8634e+13 11.3804 < 2.2e-16 ***
## 5   6419 1.9810e+15  1 2.8863e+11  0.9352    0.3335
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_hr_v1_6 <- lm(Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month + Holiday_Flag + Sto
                data = walmart_clean)
summary(model_hr_v1_6)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Unemployment + Fuel_Price + Temperature +
##     month + Holiday_Flag + Store, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -655599  -62142   -3954   44741 1631032
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1714595.8    43394.6  39.512  < 2e-16 ***
## Unemployment          -33178.9     3535.4  -9.385  < 2e-16 ***
## Fuel_Price            -21094.1     5283.1  -3.993 6.60e-05 ***
## Temperature              712.4      334.7   2.128   0.0334 *
## monthFeb              125979.3     9886.9  12.742  < 2e-16 ***
## monthMar               90601.0    10420.4   8.695  < 2e-16 ***
## monthApr               99718.4    11608.2   8.590  < 2e-16 ***
## monthMay               99143.1    13491.8   7.348 2.26e-13 ***
## monthJun              119999.1    15759.9   7.614 3.04e-14 ***
## monthJul               81618.9    16927.3   4.822 1.46e-06 ***
## monthAug               97027.6    16840.3   5.762 8.72e-09 ***
## monthSep               38384.3    15213.8   2.523   0.0117 *
## monthOct               57678.6    12411.3   4.647 3.43e-06 ***
## monthNov              213301.6    11594.8  18.396  < 2e-16 ***
## monthDec              355910.1    10224.6  34.809  < 2e-16 ***
## Holiday_FlagHoliday    32352.6     7661.9   4.223 2.45e-05 ***
## Store2                370996.8    16782.2  22.107  < 2e-16 ***
## Store3              -1169168.8    16882.2 -69.255  < 2e-16 ***
## Store4                489100.1    17895.1  27.331  < 2e-16 ***
## Store5              -1281669.4    17415.8 -73.593  < 2e-16 ***
## Store6                -24824.4    17157.1  -1.447   0.1480
## Store7               -931460.7    19601.3 -47.520  < 2e-16 ***
## Store8               -692772.6    17739.9 -39.052  < 2e-16 ***
## Store9              -1061023.1    17613.4 -60.239  < 2e-16 ***
## Store10               373840.1    17308.8  21.598  < 2e-16 ***
## Store11              -216235.7    16906.8 -12.790  < 2e-16 ***
## Store12              -356813.6    26679.8 -13.374  < 2e-16 ***
## Store13               439973.2    17612.0  24.981  < 2e-16 ***
## Store14               511830.9    17692.8  28.929  < 2e-16 ***
## Store15              -899850.7    17938.0 -50.165  < 2e-16 ***
## Store16             -1056545.6    18958.8 -55.728  < 2e-16 ***
## Store17              -679868.9    18703.3 -36.350  < 2e-16 ***
## Store18              -414482.2    18239.5 -22.724  < 2e-16 ***
## Store19               -78492.7    17890.1  -4.387 1.17e-05 ***
## Store20               557815.5    17372.4  32.109  < 2e-16 ***
## Store21              -799135.1    16783.1 -47.615  < 2e-16 ***
## Store22              -497298.9    17563.8 -28.314  < 2e-16 ***
## Store23              -240201.8    20377.8 -11.787  < 2e-16 ***
## Store24              -153016.5    18002.2  -8.500  < 2e-16 ***
```

```
## Store25              -840779.6    17686.1 -47.539  < 2e-16 ***
## Store26              -521427.6    18824.4 -27.700  < 2e-16 ***
## Store27               248333.3    17452.3  14.229  < 2e-16 ***
## Store28               -42293.0    26679.8  -1.585   0.1130
## Store29              -928716.3    19319.2 -48.072  < 2e-16 ***
## Store30             -1116624.6    16783.1 -66.533  < 2e-16 ***
## Store31              -159302.8    16783.1  -9.492  < 2e-16 ***
## Store32              -344790.7    17899.5 -19.263  < 2e-16 ***
## Store33             -1263272.7    17621.3 -71.690  < 2e-16 ***
## Store34              -504430.4    18935.6 -26.639  < 2e-16 ***
## Store35              -584441.4    17848.8 -32.744  < 2e-16 ***
## Store36             -1175531.2    16833.2 -69.834  < 2e-16 ***
## Store37             -1029816.0    16835.3 -61.170  < 2e-16 ***
## Store38              -980083.6    26679.8 -36.735  < 2e-16 ***
## Store39               -97647.1    16825.4  -5.804 6.80e-09 ***
## Store40              -665132.8    20508.1 -32.433  < 2e-16 ***
## Store41              -293639.8    18209.7 -16.125  < 2e-16 ***
## Store42              -969180.6    17308.8 -55.994  < 2e-16 ***
## Store43              -845225.9    18688.9 -45.226  < 2e-16 ***
## Store44             -1269730.6    17746.7 -71.547  < 2e-16 ***
## Store45              -723166.1    17692.8 -40.874  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141900 on 6375 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9368
## F-statistic:  1617 on 59 and 6375 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v1_1, model_hr_v1_2, model_hr_v1_3, model_hr_v1_4, model_hr_v1_5, model_hr_v1_6)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Unemployment
## Model 2: Weekly_Sales ~ Unemployment + Fuel_Price
## Model 3: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature
## Model 4: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month
## Model 5: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month +
##     Holiday_Flag
## Model 6: Weekly_Sales ~ Unemployment + Fuel_Price + Temperature + month +
##     Holiday_Flag + Store
##   Res.Df        RSS Df  Sum of Sq         F    Pr(>F)
## 1   6433 2.0262e+15
## 2   6432 2.0261e+15  1 6.8575e+10    3.4054 0.0650303 .
## 3   6431 2.0200e+15  1 6.1590e+12  305.8542 < 2.2e-16 ***
## 4   6420 1.9813e+15 11 3.8634e+13  174.4159 < 2.2e-16 ***
## 5   6419 1.9810e+15  1 2.8863e+11   14.3335 0.0001545 ***
## 6   6375 1.2837e+14 44 1.8527e+15 2090.9690 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear Models - Hierarchical Regression VERSION 2

I believe that certain known variables have a greater effect on weekly sales. I will use stepwise regression, and carry out ANOVA to test the significance after each step as outlined below: - First variable: Store - Second variable: Month - Third Variable: Holiday_Flag - Fourth Variable: Unemployment - Fifth Variable:

Fuel_Price - Sixth Variable: Temperature

STEP 1:

```
model_hr_v2_1 <- lm(Weekly_Sales ~ Store,
                data = walmart_clean)
summary(model_hr_v2_1)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -543795  -67567  -15838   32373 1849633
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1555264      13609 114.280  < 2e-16 ***
## Store2        370487      19246  19.250  < 2e-16 ***
## Store3      -1152560      19246 -59.885  < 2e-16 ***
## Store4        539449      19246  28.029  < 2e-16 ***
## Store5      -1237253      19246 -64.285  < 2e-16 ***
## Store6          9464      19246   0.492    0.623
## Store7       -984647      19246 -51.160  < 2e-16 ***
## Store8       -646515      19246 -33.592  < 2e-16 ***
## Store9      -1011284      19246 -52.544  < 2e-16 ***
## Store10       344160      19246  17.882  < 2e-16 ***
## Store11      -198881      19246 -10.333  < 2e-16 ***
## Store12      -546263      19246 -28.383  < 2e-16 ***
## Store13       448356      19246  23.296  < 2e-16 ***
## Store14       465714      19246  24.198  < 2e-16 ***
## Store15      -931952      19246 -48.422  < 2e-16 ***
## Store16     -1036017      19246 -53.829  < 2e-16 ***
## Store17      -661683      19246 -34.380  < 2e-16 ***
## Store18      -470546      19246 -24.449  < 2e-16 ***
## Store19      -110265      19246  -5.729 1.06e-08 ***
## Store20       552413      19246  28.702  < 2e-16 ***
## Store21      -799195      19246 -41.525  < 2e-16 ***
## Store22      -526763      19246 -27.370  < 2e-16 ***
## Store23      -165400      19246  -8.594  < 2e-16 ***
## Store24      -198509      19246 -10.314  < 2e-16 ***
## Store25      -848543      19246 -44.089  < 2e-16 ***
## Store26      -552353      19246 -28.699  < 2e-16 ***
## Store27       219952      19246  11.428  < 2e-16 ***
## Store28      -231742      19246 -12.041  < 2e-16 ***
## Store29     -1015813      19246 -52.780  < 2e-16 ***
## Store30     -1116685      19246 -58.021  < 2e-16 ***
## Store31      -159363      19246  -8.280  < 2e-16 ***
## Store32      -388696      19246 -20.196  < 2e-16 ***
## Store33     -1295403      19246 -67.306  < 2e-16 ***
## Store34      -588483      19246 -30.576  < 2e-16 ***
## Store35      -635539      19246 -33.021  < 2e-16 ***
## Store36     -1181752      19246 -61.401  < 2e-16 ***
## Store37     -1036364      19246 -53.847  < 2e-16 ***
```

```
## Store38     -1169533      19246 -60.767  < 2e-16 ***
## Store39      -104596      19246  -5.435 5.69e-08 ***
## Store40      -591136      19246 -30.714  < 2e-16 ***
## Store41      -287139      19246 -14.919  < 2e-16 ***
## Store42      -998861      19246 -51.899  < 2e-16 ***
## Store43      -921940      19246 -47.902  < 2e-16 ***
## Store44     -1252516      19246 -65.078  < 2e-16 ***
## Store45      -769283      19246 -39.970  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 162700 on 6390 degrees of freedom
## Multiple R-squared:  0.9174, Adjusted R-squared:  0.9168
## F-statistic:  1613 on 44 and 6390 DF,  p-value: < 2.2e-16
```

We see that store is a huge predictor of weekly sales volume.

```
model_hr_v2_2 <- lm(Weekly_Sales ~ Store + month,
                data = walmart_clean)
summary(model_hr_v2_2)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store + month, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -632161  -62923   -5369   45602 1614734
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1432184      14036 102.039  < 2e-16 ***
## Store2        370487      16928  21.887  < 2e-16 ***
## Store3      -1152560      16928 -68.088  < 2e-16 ***
## Store4        539449      16928  31.868  < 2e-16 ***
## Store5      -1237253      16928 -73.091  < 2e-16 ***
## Store6          9464      16928   0.559    0.576
## Store7       -984647      16928 -58.168  < 2e-16 ***
## Store8       -646515      16928 -38.193  < 2e-16 ***
## Store9      -1011284      16928 -59.742  < 2e-16 ***
## Store10       344160      16928  20.331  < 2e-16 ***
## Store11      -198881      16928 -11.749  < 2e-16 ***
## Store12      -546263      16928 -32.271  < 2e-16 ***
## Store13       448356      16928  26.487  < 2e-16 ***
## Store14       465714      16928  27.512  < 2e-16 ***
## Store15      -931952      16928 -55.055  < 2e-16 ***
## Store16     -1036017      16928 -61.203  < 2e-16 ***
## Store17      -661683      16928 -39.089  < 2e-16 ***
## Store18      -470546      16928 -27.798  < 2e-16 ***
## Store19      -110265      16928  -6.514 7.88e-11 ***
## Store20       552413      16928  32.634  < 2e-16 ***
## Store21      -799195      16928 -47.213  < 2e-16 ***
## Store22      -526763      16928 -31.119  < 2e-16 ***
## Store23      -165400      16928  -9.771  < 2e-16 ***
## Store24      -198509      16928 -11.727  < 2e-16 ***
```

```
## Store25      -848543      16928 -50.128  < 2e-16 ***
## Store26      -552353      16928 -32.630  < 2e-16 ***
## Store27       219952      16928  12.994  < 2e-16 ***
## Store28      -231742      16928 -13.690  < 2e-16 ***
## Store29     -1015813      16928 -60.009  < 2e-16 ***
## Store30     -1116685      16928 -65.968  < 2e-16 ***
## Store31      -159363      16928  -9.414  < 2e-16 ***
## Store32      -388696      16928 -22.962  < 2e-16 ***
## Store33     -1295403      16928 -76.526  < 2e-16 ***
## Store34      -588483      16928 -34.765  < 2e-16 ***
## Store35      -635539      16928 -37.545  < 2e-16 ***
## Store36     -1181752      16928 -69.812  < 2e-16 ***
## Store37     -1036364      16928 -61.223  < 2e-16 ***
## Store38     -1169533      16928 -69.090  < 2e-16 ***
## Store39      -104596      16928  -6.179 6.85e-10 ***
## Store40      -591136      16928 -34.921  < 2e-16 ***
## Store41      -287139      16928 -16.963  < 2e-16 ***
## Store42      -998861      16928 -59.008  < 2e-16 ***
## Store43      -921940      16928 -54.464  < 2e-16 ***
## Store44     -1252516      16928 -73.992  < 2e-16 ***
## Store45      -769283      16928 -45.445  < 2e-16 ***
## monthFeb      129315       9739  13.278  < 2e-16 ***
## monthMar       89425       9588   9.327  < 2e-16 ***
## monthApr      102877       9457  10.879  < 2e-16 ***
## monthMay      107830       9739  11.072  < 2e-16 ***
## monthJun      140440       9588  14.647  < 2e-16 ***
## monthJul      107863       9457  11.406  < 2e-16 ***
## monthAug      124133       9588  12.946  < 2e-16 ***
## monthSep       65451       9588   6.826 9.52e-12 ***
## monthOct       75748       9588   7.900 3.26e-15 ***
## monthNov      223381      10669  20.938  < 2e-16 ***
## monthDec      357979      10121  35.369  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143100 on 6379 degrees of freedom
## Multiple R-squared:  0.9362, Adjusted R-squared:  0.9357
## F-statistic:  1703 on 55 and 6379 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v2_1, model_hr_v2_2)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Store
## Model 2: Weekly_Sales ~ Store + month
##   Res.Df        RSS Df  Sum of Sq      F    Pr(>F)
## 1   6390 1.6924e+14
## 2   6379 1.3069e+14 11 3.8548e+13 171.04 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_hr_v2_3 <- lm(Weekly_Sales ~ Store + month + Holiday_Flag,
                data = walmart_clean)
summary(model_hr_v2_3)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store + month + Holiday_Flag, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -657548  -62111   -5197   45378 1621081
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1432184      14018 102.167  < 2e-16 ***
## Store2                370487      16906  21.914  < 2e-16 ***
## Store3              -1152560      16906 -68.173  < 2e-16 ***
## Store4                539449      16906  31.908  < 2e-16 ***
## Store5              -1237253      16906 -73.182  < 2e-16 ***
## Store6                 9464      16906   0.560    0.576
## Store7              -984647      16906 -58.241  < 2e-16 ***
## Store8              -646515      16906 -38.241  < 2e-16 ***
## Store9             -1011284      16906 -59.816  < 2e-16 ***
## Store10              344160      16906  20.357  < 2e-16 ***
## Store11             -198881      16906 -11.764  < 2e-16 ***
## Store12             -546263      16906 -32.311  < 2e-16 ***
## Store13              448356      16906  26.520  < 2e-16 ***
## Store14              465714      16906  27.547  < 2e-16 ***
## Store15             -931952      16906 -55.124  < 2e-16 ***
## Store16            -1036017      16906 -61.279  < 2e-16 ***
## Store17             -661683      16906 -39.138  < 2e-16 ***
## Store18             -470546      16906 -27.832  < 2e-16 ***
## Store19             -110265      16906  -6.522 7.47e-11 ***
## Store20              552413      16906  32.675  < 2e-16 ***
## Store21             -799195      16906 -47.272  < 2e-16 ***
## Store22             -526763      16906 -31.158  < 2e-16 ***
## Store23             -165400      16906  -9.783  < 2e-16 ***
## Store24             -198509      16906 -11.742  < 2e-16 ***
## Store25             -848543      16906 -50.191  < 2e-16 ***
## Store26             -552353      16906 -32.671  < 2e-16 ***
## Store27              219952      16906  13.010  < 2e-16 ***
## Store28             -231742      16906 -13.707  < 2e-16 ***
## Store29            -1015813      16906 -60.084  < 2e-16 ***
## Store30            -1116685      16906 -66.051  < 2e-16 ***
## Store31             -159363      16906  -9.426  < 2e-16 ***
## Store32             -388696      16906 -22.991  < 2e-16 ***
## Store33            -1295403      16906 -76.622  < 2e-16 ***
## Store34             -588483      16906 -34.808  < 2e-16 ***
## Store35             -635539      16906 -37.592  < 2e-16 ***
## Store36            -1181752      16906 -69.900  < 2e-16 ***
## Store37            -1036364      16906 -61.300  < 2e-16 ***
## Store38            -1169533      16906 -69.177  < 2e-16 ***
## Store39             -104596      16906  -6.187 6.52e-10 ***
## Store40             -591136      16906 -34.965  < 2e-16 ***
## Store41             -287139      16906 -16.984  < 2e-16 ***
## Store42             -998861      16906 -59.082  < 2e-16 ***
## Store43             -921940      16906 -54.532  < 2e-16 ***
## Store44            -1252516      16906 -74.085  < 2e-16 ***
```

```
## Store45              -769283      16906 -45.502  < 2e-16 ***
## monthFeb              121382       9916  12.241  < 2e-16 ***
## monthMar               89425       9576   9.338  < 2e-16 ***
## monthApr              102877       9445  10.892  < 2e-16 ***
## monthMay              107830       9727  11.086  < 2e-16 ***
## monthJun              140440       9576  14.666  < 2e-16 ***
## monthJul              107863       9445  11.420  < 2e-16 ***
## monthAug              124133       9576  12.963  < 2e-16 ***
## monthSep               58127       9740   5.968 2.53e-09 ***
## monthOct               75748       9576   7.910 3.01e-15 ***
## monthNov              215448      10828  19.897  < 2e-16 ***
## monthDec              351632      10225  34.389  < 2e-16 ***
## Holiday_FlagHoliday    31735       7701   4.121 3.82e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143000 on 6378 degrees of freedom
## Multiple R-squared:  0.9364, Adjusted R-squared:  0.9358
## F-statistic:  1677 on 56 and 6378 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v2_1, model_hr_v2_2, model_hr_v2_3)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Store
## Model 2: Weekly_Sales ~ Store + month
## Model 3: Weekly_Sales ~ Store + month + Holiday_Flag
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6390 1.6924e+14
## 2   6379 1.3069e+14 11 3.8548e+13 171.473 < 2.2e-16 ***
## 3   6378 1.3035e+14  1 3.4704e+11  16.981 3.823e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_hr_v2_4 <- lm(Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment,
                data = walmart_clean)
summary(model_hr_v2_4)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment,
##     data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -659101  -61641   -5316   44260 1633746
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          1623053      25667  63.236  < 2e-16 ***
## Store2                370828      16805  22.067  < 2e-16 ***
## Store3              -1163574      16851 -69.052  < 2e-16 ***
## Store4                497628      17456  28.508  < 2e-16 ***
## Store5              -1270669      17223 -73.776  < 2e-16 ***
## Store6                -16037      17050  -0.941    0.347
```

```
## Store7               -959863      17036 -56.342  < 2e-16 ***
## Store8               -685105      17361 -39.463  < 2e-16 ***
## Store9              -1049669      17355 -60.483  < 2e-16 ***
## Store10               363284      16943  21.442  < 2e-16 ***
## Store11              -209896      16851 -12.456  < 2e-16 ***
## Store12              -406344      23067 -17.616  < 2e-16 ***
## Store13               432890      16895  25.622  < 2e-16 ***
## Store14               492100      17067  28.834  < 2e-16 ***
## Store15              -922147      16841 -54.755  < 2e-16 ***
## Store16             -1064794      17116 -62.210  < 2e-16 ***
## Store17              -688645      17078 -40.323  < 2e-16 ***
## Store18              -439343      17170 -25.587  < 2e-16 ***
## Store19              -100461      16841  -5.965 2.57e-09 ***
## Store20               546339      16819  32.484  < 2e-16 ***
## Store21              -798854      16805 -47.537  < 2e-16 ***
## Store22              -515101      16856 -30.558  < 2e-16 ***
## Store23              -236919      18645 -12.707  < 2e-16 ***
## Store24              -177250      16975 -10.442  < 2e-16 ***
## Store25              -854617      16819 -50.813  < 2e-16 ***
## Store26              -545704      16822 -32.441  < 2e-16 ***
## Store27               229896      16842  13.650  < 2e-16 ***
## Store28               -91823      23067  -3.981 6.95e-05 ***
## Store29              -960010      17948 -53.490  < 2e-16 ***
## Store30             -1116344      16805 -66.430  < 2e-16 ***
## Store31              -159022      16805  -9.463  < 2e-16 ***
## Store32              -363912      17036 -21.361  < 2e-16 ***
## Store33             -1271954      17012 -74.768  < 2e-16 ***
## Store34              -529416      18080 -29.281  < 2e-16 ***
## Store35              -605600      17142 -35.329  < 2e-16 ***
## Store36             -1175180      16821 -69.863  < 2e-16 ***
## Store37             -1029792      16821 -61.220  < 2e-16 ***
## Store38             -1029614      23067 -44.636  < 2e-16 ***
## Store39               -98024      16821  -5.827 5.90e-09 ***
## Store40              -662655      18645 -35.541  < 2e-16 ***
## Store41              -303327      16904 -17.944  < 2e-16 ***
## Store42              -979737      16943 -57.826  < 2e-16 ***
## Store43              -862873      18080 -47.725  < 2e-16 ***
## Store44             -1274746      16991 -75.024  < 2e-16 ***
## Store45              -742897      17067 -43.529  < 2e-16 ***
## monthFeb              127510       9880  12.905  < 2e-16 ***
## monthMar               94286       9534   9.889  < 2e-16 ***
## monthApr              107089       9400  11.392  < 2e-16 ***
## monthMay              110947       9675  11.468  < 2e-16 ***
## monthJun              142379       9521  14.954  < 2e-16 ***
## monthJul              109647       9390  11.677  < 2e-16 ***
## monthAug              123339       9519  12.957  < 2e-16 ***
## monthSep               59046       9682   6.099 1.13e-09 ***
## monthOct               73228       9523   7.690 1.70e-14 ***
## monthNov              221580      10785  20.545  < 2e-16 ***
## monthDec              357758      10187  35.118  < 2e-16 ***
## Holiday_FlagHoliday    31600       7655   4.128 3.70e-05 ***
## Unemployment          -25412       2870  -8.855  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 142100 on 6377 degrees of freedom
## Multiple R-squared:  0.9372, Adjusted R-squared:  0.9366
## F-statistic:  1669 on 57 and 6377 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v2_1, model_hr_v2_2, model_hr_v2_3, model_hr_v2_4)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Store
## Model 2: Weekly_Sales ~ Store + month
## Model 3: Weekly_Sales ~ Store + month + Holiday_Flag
## Model 4: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6390 1.6924e+14
## 2   6379 1.3069e+14 11 3.8548e+13 173.555 < 2.2e-16 ***
## 3   6378 1.3035e+14  1 3.4704e+11  17.187 3.431e-05 ***
## 4   6377 1.2876e+14  1 1.5832e+12  78.410 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_hr_v2_5 <- lm(Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment + Fuel_Price,
                data = walmart_clean)
summary(model_hr_v2_5)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment +
##     Fuel_Price, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -660148  -62308   -4176   45167 1634802
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1745433      40915  42.660  < 2e-16 ***
## Store2               370935      16787  22.097  < 2e-16 ***
## Store3             -1167016      16857 -69.232  < 2e-16 ***
## Store4               484503      17769  27.266  < 2e-16 ***
## Store5             -1281113      17419 -73.548  < 2e-16 ***
## Store6               -24007      17158  -1.399 0.161802
## Store7              -951675      17151 -55.488  < 2e-16 ***
## Store8              -697164      17624 -39.557  < 2e-16 ***
## Store9             -1061665      17616 -60.268  < 2e-16 ***
## Store10              376463      17270  21.799  < 2e-16 ***
## Store11             -213338      16857 -12.656  < 2e-16 ***
## Store12             -354798      26670 -13.303  < 2e-16 ***
## Store13              429401      16902  25.406  < 2e-16 ***
## Store14              504348      17345  29.078  < 2e-16 ***
## Store15             -911832      17036 -53.522  < 2e-16 ***
## Store16            -1073345      17242 -62.251  < 2e-16 ***
## Store17             -695728      17160 -40.545  < 2e-16 ***
## Store18             -425102      17549 -24.224  < 2e-16 ***
## Store19              -90145      17036  -5.291 1.25e-07 ***
```

```
## Store20               548443      16810  32.627  < 2e-16 ***
## Store21              -798748      16787 -47.582  < 2e-16 ***
## Store22              -506966      16971 -29.872  < 2e-16 ***
## Store23              -254780      19198 -13.272  < 2e-16 ***
## Store24              -163355      17339  -9.421  < 2e-16 ***
## Store25              -852513      16810 -50.716  < 2e-16 ***
## Store26              -539136      16890 -31.920  < 2e-16 ***
## Store27               240255      17039  14.100  < 2e-16 ***
## Store28               -40277      26670  -1.510 0.131050
## Store29              -938080      18817 -49.854  < 2e-16 ***
## Store30             -1116237      16787 -66.495  < 2e-16 ***
## Store31              -158915      16787  -9.467  < 2e-16 ***
## Store32              -355724      17151 -20.741  < 2e-16 ***
## Store33             -1257423      17411 -72.222  < 2e-16 ***
## Store34              -511012      18687 -27.347  < 2e-16 ***
## Store35              -592242      17473 -33.894  < 2e-16 ***
## Store36             -1173440      16809 -69.809  < 2e-16 ***
## Store37             -1027738      16812 -61.133  < 2e-16 ***
## Store38              -978067      26670 -36.672  < 2e-16 ***
## Store39               -95970      16812  -5.709 1.19e-08 ***
## Store40              -680516      19198 -35.448  < 2e-16 ***
## Store41              -307944      16929 -18.191  < 2e-16 ***
## Store42              -966557      17270 -55.969  < 2e-16 ***
## Store43              -844414      18690 -45.179  < 2e-16 ***
## Store44             -1280349      17036 -75.157  < 2e-16 ***
## Store45              -730649      17345 -42.125  < 2e-16 ***
## monthFeb              127310       9870  12.899  < 2e-16 ***
## monthMar               99192       9610  10.322  < 2e-16 ***
## monthApr              113771       9550  11.913  < 2e-16 ***
## monthMay              118717       9874  12.023  < 2e-16 ***
## monthJun              146643       9576  15.314  < 2e-16 ***
## monthJul              111592       9394  11.879  < 2e-16 ***
## monthAug              126558       9546  13.258  < 2e-16 ***
## monthSep               63274       9734   6.500 8.62e-11 ***
## monthOct               74634       9520   7.840 5.25e-15 ***
## monthNov              222428      10776  20.642  < 2e-16 ***
## monthDec              358075      10177  35.186  < 2e-16 ***
## Holiday_FlagHoliday    31266       7647   4.089 4.39e-05 ***
## Unemployment          -33353       3535  -9.434  < 2e-16 ***
## Fuel_Price            -20221       5269  -3.838 0.000125 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141900 on 6376 degrees of freedom
## Multiple R-squared:  0.9373, Adjusted R-squared:  0.9367
## F-statistic:  1644 on 58 and 6376 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v2_1, model_hr_v2_2, model_hr_v2_3, model_hr_v2_4, model_hr_v2_5)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Store
## Model 2: Weekly_Sales ~ Store + month
## Model 3: Weekly_Sales ~ Store + month + Holiday_Flag
## Model 4: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment
```

```
## Model 5: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment +
##     Fuel_Price
##   Res.Df        RSS Df  Sum of Sq       F    Pr(>F)
## 1   6390 1.6924e+14
## 2   6379 1.3069e+14 11 3.8548e+13 173.928 < 2.2e-16 ***
## 3   6378 1.3035e+14  1 3.4704e+11  17.224 3.365e-05 ***
## 4   6377 1.2876e+14  1 1.5832e+12  78.579 < 2.2e-16 ***
## 5   6376 1.2847e+14  1 2.9680e+11  14.731 0.0001252 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_hr_v2_6 <- lm(Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment + Fuel_Price + Temperatu
                data = walmart_clean)
summary(model_hr_v2_6)
```

```
##
## Call:
## lm(formula = Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment +
##     Fuel_Price + Temperature, data = walmart_clean)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -655599  -62142   -3954   44741 1631032
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1714595.8    43394.6  39.512  < 2e-16 ***
## Store2               370996.8    16782.2  22.107  < 2e-16 ***
## Store3             -1169168.8    16882.2 -69.255  < 2e-16 ***
## Store4               489100.1    17895.1  27.331  < 2e-16 ***
## Store5             -1281669.4    17415.8 -73.593  < 2e-16 ***
## Store6               -24824.4    17157.1  -1.447   0.1480
## Store7              -931460.7    19601.3 -47.520  < 2e-16 ***
## Store8              -692772.6    17739.9 -39.052  < 2e-16 ***
## Store9             -1061023.1    17613.4 -60.239  < 2e-16 ***
## Store10              373840.1    17308.8  21.598  < 2e-16 ***
## Store11             -216235.7    16906.8 -12.790  < 2e-16 ***
## Store12             -356813.6    26679.8 -13.374  < 2e-16 ***
## Store13              439973.2    17612.0  24.981  < 2e-16 ***
## Store14              511830.9    17692.8  28.929  < 2e-16 ***
## Store15             -899850.7    17938.0 -50.165  < 2e-16 ***
## Store16            -1056545.6    18958.8 -55.728  < 2e-16 ***
## Store17             -679868.9    18703.3 -36.350  < 2e-16 ***
## Store18             -414482.2    18239.5 -22.724  < 2e-16 ***
## Store19              -78492.7    17890.1  -4.387 1.17e-05 ***
## Store20              557815.5    17372.4  32.109  < 2e-16 ***
## Store21             -799135.1    16783.1 -47.615  < 2e-16 ***
## Store22             -497298.9    17563.8 -28.314  < 2e-16 ***
## Store23             -240201.8    20377.8 -11.787  < 2e-16 ***
## Store24             -153016.5    18002.2  -8.500  < 2e-16 ***
## Store25             -840779.6    17686.1 -47.539  < 2e-16 ***
## Store26             -521427.6    18824.4 -27.700  < 2e-16 ***
## Store27              248333.3    17452.3  14.229  < 2e-16 ***
## Store28              -42293.0    26679.8  -1.585   0.1130
## Store29             -928716.3    19319.2 -48.072  < 2e-16 ***
```

```
## Store30             -1116624.6    16783.1 -66.533  < 2e-16 ***
## Store31              -159302.8    16783.1  -9.492  < 2e-16 ***
## Store32              -344790.7    17899.5 -19.263  < 2e-16 ***
## Store33             -1263272.7    17621.3 -71.690  < 2e-16 ***
## Store34              -504430.4    18935.6 -26.639  < 2e-16 ***
## Store35              -584441.4    17848.8 -32.744  < 2e-16 ***
## Store36             -1175531.2    16833.2 -69.834  < 2e-16 ***
## Store37             -1029816.0    16835.3 -61.170  < 2e-16 ***
## Store38              -980083.6    26679.8 -36.735  < 2e-16 ***
## Store39               -97647.1    16825.4  -5.804 6.80e-09 ***
## Store40              -665132.8    20508.1 -32.433  < 2e-16 ***
## Store41              -293639.8    18209.7 -16.125  < 2e-16 ***
## Store42              -969180.6    17308.8 -55.994  < 2e-16 ***
## Store43              -845225.9    18688.9 -45.226  < 2e-16 ***
## Store44             -1269730.6    17746.7 -71.547  < 2e-16 ***
## Store45              -723166.1    17692.8 -40.874  < 2e-16 ***
## monthFeb              125979.3     9886.9  12.742  < 2e-16 ***
## monthMar               90601.0    10420.4   8.695  < 2e-16 ***
## monthApr               99718.4    11608.2   8.590  < 2e-16 ***
## monthMay               99143.1    13491.8   7.348 2.26e-13 ***
## monthJun              119999.1    15759.9   7.614 3.04e-14 ***
## monthJul               81618.9    16927.3   4.822 1.46e-06 ***
## monthAug               97027.6    16840.3   5.762 8.72e-09 ***
## monthSep               38384.3    15213.8   2.523   0.0117 *
## monthOct               57678.6    12411.3   4.647 3.43e-06 ***
## monthNov              213301.6    11594.8  18.396  < 2e-16 ***
## monthDec              355910.1    10224.6  34.809  < 2e-16 ***
## Holiday_FlagHoliday    32352.6     7661.9   4.223 2.45e-05 ***
## Unemployment          -33178.9     3535.4  -9.385  < 2e-16 ***
## Fuel_Price            -21094.1     5283.1  -3.993 6.60e-05 ***
## Temperature              712.4      334.7   2.128   0.0334 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 141900 on 6375 degrees of freedom
## Multiple R-squared:  0.9374, Adjusted R-squared:  0.9368
## F-statistic:  1617 on 59 and 6375 DF,  p-value: < 2.2e-16
```

```
anova(model_hr_v2_1, model_hr_v2_2, model_hr_v2_3, model_hr_v2_4, model_hr_v2_5, model_hr_v2_6)
```

```
## Analysis of Variance Table
##
## Model 1: Weekly_Sales ~ Store
## Model 2: Weekly_Sales ~ Store + month
## Model 3: Weekly_Sales ~ Store + month + Holiday_Flag
## Model 4: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment
## Model 5: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment +
##     Fuel_Price
## Model 6: Weekly_Sales ~ Store + month + Holiday_Flag + Unemployment +
##     Fuel_Price + Temperature
##   Res.Df        RSS Df  Sum of Sq        F    Pr(>F)
## 1   6390 1.6924e+14
## 2   6379 1.3069e+14 11 3.8548e+13 174.0246 < 2.2e-16 ***
## 3   6378 1.3035e+14  1 3.4704e+11  17.2338 3.348e-05 ***
## 4   6377 1.2876e+14  1 1.5832e+12  78.6220 < 2.2e-16 ***
```

```
## 5   6376 1.2847e+14  1 2.9680e+11  14.7389 0.0001247 ***
## 6   6375 1.2837e+14  1 9.1211e+10   4.5295 0.0333539 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

T-TEST - Holiday Flag

```
t.test(Weekly_Sales ~ Holiday_Flag,
       data = walmart_clean,
       var.equal = TRUE,
       paired = FALSE)
```

```
##
##  Two Sample t-test
##
## data:  Weekly_Sales by Holiday_Flag
## t = -2.9609, df = 6433, p-value = 0.003079
## alternative hypothesis: true difference in means between group No_Holiday and group Holiday is not e
## 95 percent confidence interval:
##  -135677.70  -27585.32
## sample estimates:
## mean in group No_Holiday    mean in group Holiday
##                  1041256                  1122888
```

```
t.test(Weekly_Sales ~ Holiday_Flag,
       data = walmart_clean,
       var.equal = FALSE,
       paired = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  Weekly_Sales by Holiday_Flag
## t = -2.6801, df = 504, p-value = 0.007602
## alternative hypothesis: true difference in means between group No_Holiday and group Holiday is not e
## 95 percent confidence interval:
##  -141473.17  -21789.85
## sample estimates:
## mean in group No_Holiday    mean in group Holiday
##                  1041256                  1122888
```

ANOVA - Store

```
walmart_clean$partno <- 1:nrow(walmart_clean)
ezANOVA(data = walmart_clean,
        dv = Weekly_Sales,
        between = Store,
        wid = partno,
        type = 3,
        detailed = T)$`Levene's Test for Homogeneity of Variance`
```

```
## Coefficient covariances computed by hccm()
```

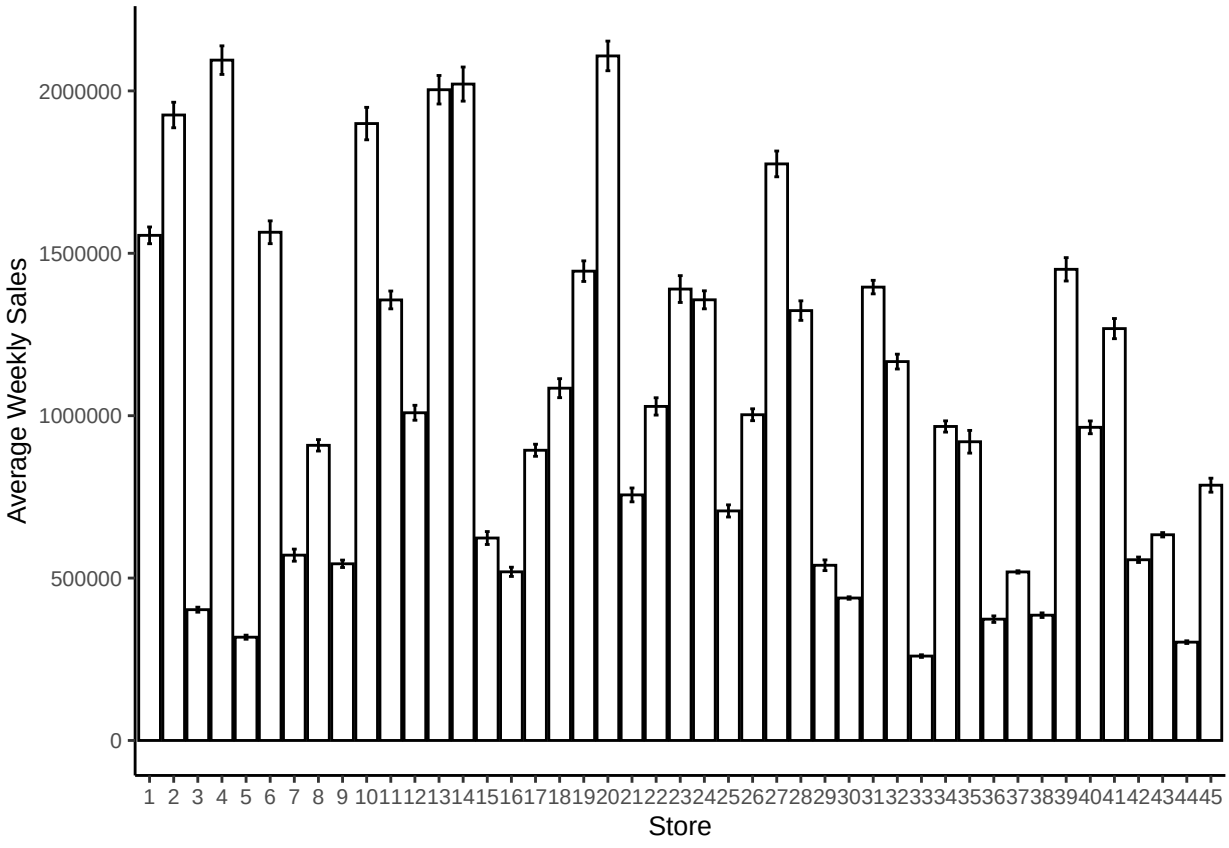
```
##   DFn  DFd          SSn          SSd        F             p p<.05
## 1  44 6390 1.455522e+13 1.111227e+14 19.02234 3.059212e-137     *
```

We see that Levene's test is highly significant. Hence, we will run a one-way test.

```
oneway.test(Weekly_Sales~Store, data = walmart_clean)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Weekly_Sales and Store
## F = 2749.3, num df = 44.0, denom df = 2222.6, p-value < 2.2e-16
```

```
bargraph4 <- ggplot(walmart_clean, aes(Store, Weekly_Sales))
bargraph4 +
  cleanup +
  stat_summary(fun.y = mean,
               geom = "bar",
               fill = "White",
               color = "Black") +
  stat_summary(fun.data = mean_cl_normal,
               geom = "errorbar",
               width = .2,
               position = "dodge") +
  xlab("Store") +
  ylab("Average Weekly Sales")
```



ANOVA - Month

```
ezANOVA(data = walmart_clean,
        dv = Weekly_Sales,
        between = month,
        wid = partno,
        type = 3,
        detailed = T)$`Levene's Test for Homogeneity of Variance`
```

```
## Warning: Converting "partno" to factor for ANOVA.
```

```
## Warning: Data is unbalanced (unequal N per group). Make sure you specified a
## well-considered value for the type argument to ezANOVA().
```

```
## Coefficient covariances computed by hccm()
```

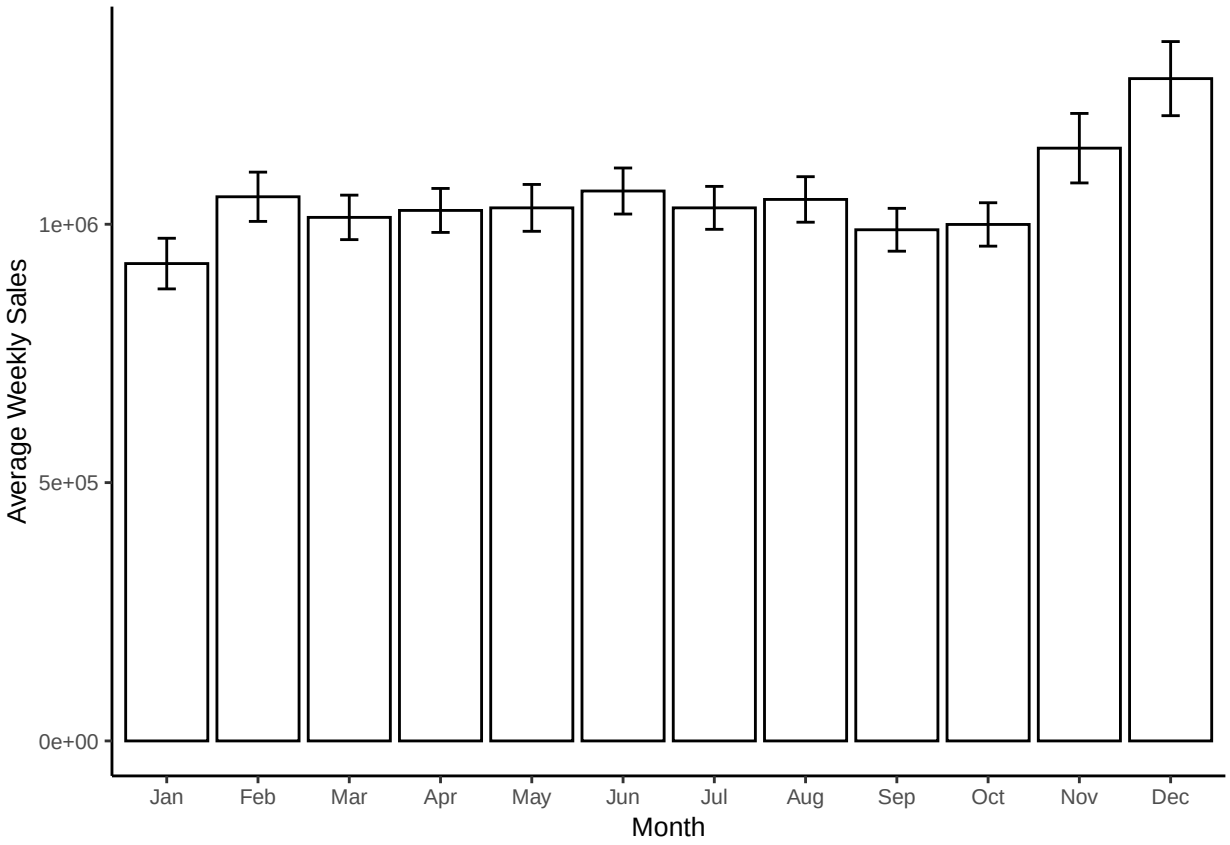
```
##   DFn  DFd          SSn        SSd        F            p p<.05
## 1  11 6423 1.576587e+13 6.6845e+14 13.77191 1.609411e-26     *
```

We see that Levene's test is highly significant. Hence, we will run a one-way test.

```
oneway.test(Weekly_Sales~month, data = walmart_clean)
```

```
##
##  One-way analysis of means (not assuming equal variances)
##
## data:  Weekly_Sales and month
## F = 7.9354, num df = 11.0, denom df = 2418.7, p-value = 1.05e-13
```

```
bargraph5 <- ggplot(walmart_clean, aes(month, Weekly_Sales))
bargraph5 +
  cleanup +
  stat_summary(fun.y = mean,
               geom = "bar",
               fill = "White",
               color = "Black") +
  stat_summary(fun.data = mean_cl_normal,
               geom = "errorbar",
               width = .2,
               position = "dodge") +
  xlab("Month") +
  ylab("Average Weekly Sales")
```


Average Weekly Sales
1e+06
5e+05
0e+00
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month