Timothy Compton
Project: Linear Regression with R (R Studio)

Analysis and Prediction of Sales: Walmart USA

Business Scenario:

One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to the inappropriate machine learning algorithm.

Objectives:

Provide a statistical analysis of historical sales of Walmart stores. Use data collected to develop a statistical model that can be used to accurately predict future sales for Store 1.

Summary of Data used for Analysis:

Historical data for 45 Walmart stores located in different regions is available from 2010-02-05 to 2012-11-01. Historical data includes:

- Store - the store number
- Date - the week of sales
- Weekly_Sales - sales for the given store
- Holiday_Flag - whether the week is a special holiday week 1 – Holiday week 0 – Non-holiday week
- Temperature - Temperature on the day of sale
- Fuel_Price - Cost of fuel in the region
- CPI – Prevailing consumer price index
- Unemployment - Prevailing unemployment rate

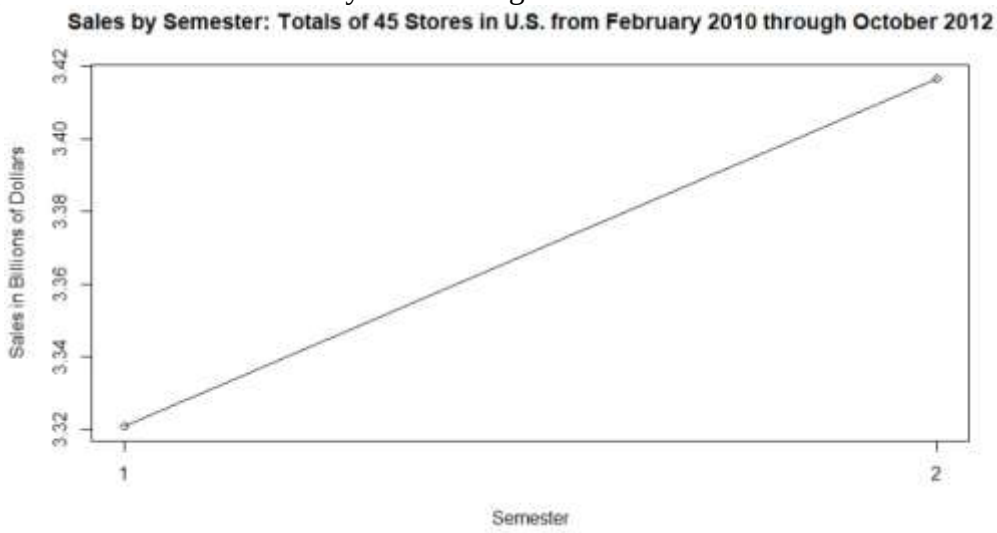
Solutions Summary:

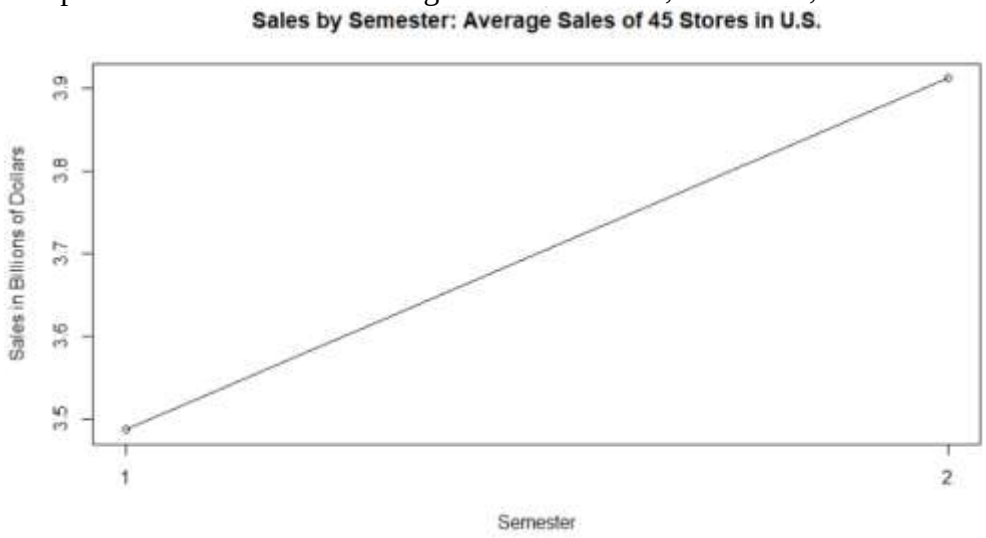
Statistical Analysis of all 45 Stores:

- Maximum Weekly Sales: Store #14
- Maximum Standard Deviation of Weekly Sales: Store #14
- Greatest CV of Weekly Sales: Store #35
- Quarterly Growth Rate from Q2 to Q3 2012:
    - Best: Store #7
    - Second-Best: Store #16
- Holiday Sales: Top 4 Holidays
    - Weekly Sales greater than mean of non-holiday weeks: Super Bowl, Labor Day, Thanksgiving
    - Weekly Sales less than mean of non-holiday weeks: Christmas

- Monthly View of Sales:
    - Using provided data from February 2010 through October 2012:
        - 

    - After data updated to account for missing months Jan/2010, Nov/2012, Dec/2012:
        - 

- Semester View of Sales:
    - Using provided data from February 2010 through October 2012:
        - 

    - After data updated to account for missing months Jan/2010, Nov/2012, Dec/2012:
        - 

Statistical Modeling for Store 1:

- All dates were replaced by a value that represented the number of days that had passed since February 4, 2010. This resulted in February 5, 2010 being replaced by a 1, February 6, 2010 being replaced by a 2, et cetera.
- My Hypotheses made before modeling the data were as follows:
    - CPI: will not impact sales
        - Incorrect: CPI was the most significant variable in predicting sales
    - Unemployment: will impact sales
        - Incorrect: Unemployment was one of the least significant variables
    - Fuel Price: will impact sales, but not in a linear manner
        - Incorrect, Fuel Price was one of the least significant variables
- Model developed with the best accuracy: Fit3
    - Fit3: lm(formula = Weekly_Sales ~ CPI + month +semester, data = Store1cLR)
    - Fit3 R^2 Value: 20.1
    - Fit3 MAPE Value: 4.47%
- Summary of All Models Considered:

Data	Approach	Model Name	Detail	R2	adj R2	std err	R2 - adj R2	MAPE	Priority
Store1aLR	VIF, Step	Fit1	-year -day - quarter	24.31	21.55	138200	2.76	6.04%	3
Store1bLR	Manual, VIF, Step	Fit2	Replace Holiday Week Sales w/ Average Sales, -year -day -quarter	23.68	21.47	127200	2.21	5.59%	2
Store1cLR	Manual, VIF, Step	Fit3	Replace Holiday Week Sales w/ Average Sales, Replace December Sales w/ Average Sales, -year -day -quarter	20.1	18.37	87640	1.73	4.47%	1

Summarize Additional Data Fields Introduced and Modified:

The data that was initially provided required some wrangling to be useful in the Fit3 Model. For instance, month and semester were very important independent variables that had to be pulled from the dates provided. Also, replacing outliers that were the result of a holiday or holiday season was important. Replacing weekly sales from the holidays of the Super Bowl, Labor Day, Thanksgiving, and Christmas, as well as the entire month of December (Christmas season), resulted in a measured 25% reduction in error of the final model.

Statistical Algorithm Execution:

```
#TimothyCompton
#April 8, 2021
#03DSwR
#Project1: Retail Analysis with WalMart Data

#=========================================================
#Load and Set Up Data:
#=========================================================
#Set Up Environment:
rm(list=ls())
library(lubridate)
library(dplyr)
library(sp)
library(raster)
library(usdm)

#Read Data File:
DF1=read.csv("C:/Users/Tim/Documents/Certs & Tests/DSci 2021/Simplilearn 03062021/Courses/03 Data Science with R/Projects-Assessment/Project1_RetailAnalysisWithWalmartData/Walmart_Store_sales.csv")

#View Data:
View(DF1)
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
1	1	05-02-2010	1643691	0	42.31	2.572	211.0964	8.106
2	1	12-02-2010	1641957	1	38.51	2.548	211.2422	8.106

```
#Check for NAs:
summary(DF1) #No NAs found
```

```
> #Check for NAs:
> summary(DF1) #No NAs found
     Store        Date           Weekly_Sales      Holiday_Flag      Temperature       Fuel_Price
 Min.   : 1   Length:6435        Min.   : 209986   Min.   :0.00000   Min.   : -2.06   Min.   :2.472
 1st Qu.:12   Class :character   1st Qu.: 553350   1st Qu.:0.00000   1st Qu.: 47.46   1st Qu.:2.933
 Median :23   Mode  :character   Median : 960746   Median :0.00000   Median : 62.67   Median :3.445
 Mean   :23                      Mean   :1046965   Mean   :0.06993   Mean   : 60.66   Mean   :3.359
 3rd Qu.:34                      3rd Qu.:1420159   3rd Qu.:0.00000   3rd Qu.: 74.94   3rd Qu.:3.735
 Max.   :45                      Max.   :3818686   Max.   :1.00000   Max.   :100.14   Max.   :4.468
      CPI         Unemployment
 Min.   :126.1   Min.   : 3.879
 1st Qu.:131.7   1st Qu.: 6.891
 Median :182.6   Median : 7.874
 Mean   :171.6   Mean   : 7.999
 3rd Qu.:212.7   3rd Qu.: 8.622
 Max.   :227.2   Max.   :14.313
```

```
#Reformat Initial Data:
DF1$Date=as.Date(DF1$Date,format="%d-%m-%Y")

#Add all new columns to DF1:
DF1$year=year(DF1$Date)
DF1$month=month(DF1$Date)
DF1$quarter=quarter(DF1$Date)
```

```
DF1$semester=semester(DF1$Date)

View(DF1)
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester
1	1	2010-02-05	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1
2	1	2010-02-12	1641957	1	38.51	2.548	211.2422	8.106	2010	2	1	1

```
#=====================================================
#Basic Statistics Tasks:
#=====================================================
#New Data File for Questions 1-4: Stats1
Stats1=summarize(group_by(DF1,Store), Max_WeeklySales=max(Weekly_Sales),
Sum_WeeklySales=sum(Weekly_Sales), Mean_WeeklySales=mean(Weekly_Sales),
StdDev_WeeklySales=sd(Weekly_Sales), Sales_2012.Q2=sum(Weekly_Sales[quarter==2&year==2012]),
Sales_2012.Q3=sum(Weekly_Sales[quarter==3&year==2012]))
Stats1$CV_WeeklySales=Stats1$StdDev_WeeklySales/Stats1$Mean_WeeklySales*100
Stats1=Stats1[, colnames(Stats1)[c(1:5,8,6,7)]]
Stats1$GrowthRate_Q3.Q2=Stats1$Sales_2012.Q3/Stats1$Sales_2012.Q2*100
View(Stats1)
```

	Store	Max_WeeklySales	Sum_WeeklySales	Mean_WeeklySales	StdDev_WeeklySales	CV_WeeklySales	Sales_2012.Q2	Sales_2012.Q3	GrowthRate_Q3.Q2
1	1	2387950.2	222402809	1555264.4	155980.77	10.029212	20978760	20253948	96.54502
2	2	3436007.7	275382441	1925751.3	237683.69	12.342388	25083605	24303355	96.88940

```
#New Data File for Question 5: Stats2
Metric=c("TotalSales")
Stats2=data.frame(Metric)
Stats2$Non_Holiday=sum(DF1$Weekly_Sales[DF1$Holiday_Flag==0])/sum(DF1$Holiday_Flag==0)
Stats2$SuperBowl=sum(DF1$Weekly_Sales[DF1$Holiday_Flag==1&DF1$month==2]/sum(DF1$Holiday_Flag==1&DF1$month==2))
Stats2$LabourDay=sum(DF1$Weekly_Sales[DF1$Holiday_Flag==1&DF1$month==9]/sum(DF1$Holiday_Flag==1&DF1$month==9))
Stats2$Thanksgiving=sum(DF1$Weekly_Sales[DF1$Holiday_Flag==1&DF1$month==11]/sum(DF1$Holiday_Flag==1&DF1$month==11))
Stats2$Christmas=sum(DF1$Weekly_Sales[DF1$Holiday_Flag==1&DF1$month==12]/sum(DF1$Holiday_Flag==1&DF1$month==12))
View(Stats2)
```

	Metric	Non_Holiday	SuperBowl	LabourDay	Thanksgiving	Christmas
1	TotalSales	1041256	1079128	1042427	1471273	960833.1

```
#New Data File for Question 6: Stats3
Stats3=summarize(group_by(DF1,month),MonthlySales=sum(Weekly_Sales))
Stats3$MonthlySales.Millions=Stats3$MonthlySales/1000000
Stats3$SalesJanNovDec=Stats3$MonthlySales #Add new column to calculate sales for months missing
Stats3[1,"SalesJanNovDec"]=Stats3[1,"MonthlySales"]+Stats3[1,"MonthlySales"]/2 #Account for January 2010 sales missing. Add average of 2011 and 2012 sales
Stats3[11,"SalesJanNovDec"]=Stats3[11,"MonthlySales"]+Stats3[11,"MonthlySales"]/2 #Account for November 2012 sales missing. Add average of 2010 and 2011 sales
Stats3[12,"SalesJanNovDec"]=Stats3[12,"MonthlySales"]+Stats3[12,"MonthlySales"]/2 #Account for December 2012 sales missing. Add average of 2010 and 2011 sales
Stats3$MonthlySalesJanNovDec.Millions=Stats3$SalesJanNovDec/1000000
```

View(Stats3)

	month	MonthlySales	MonthlySales.Millions	SalesJanNovDec	MonthlySalesJanNovDec.Millions
1	1	332598438	332.5984	498897658	498.8977
2	2	568727890	568.7279	568727890	568.7279
3	3	592785901	592.7859	592785901	592.7859
4	4	646859785	646.8598	646859785	646.8598
5	5	557125572	557.1256	557125572	557.1256
6	6	622629887	622.6299	622629887	622.6299
7	7	650000977	650.0010	650000977	650.0010
8	8	613090209	613.0902	613090209	613.0902
9	9	578761179	578.7612	578761179	578.7612
10	10	584784788	584.7848	584784788	584.7848
11	11	413015725	413.0157	619523588	619.5236
12	12	576838635	576.8386	865257953	865.2580

#New Data File for Question 6: Stats4
Stats4=summarize(group_by(DF1,semester),SemesterSales=sum(Weekly_Sales))
Stats4$SemesterSales.Billions=Stats4$SemesterSales/1000000000
Stats4$SemesterSalesJanNovDec=Stats4$SemesterSales
Stats4[1,"SemesterSalesJanNovDec"]=Stats4[1,"SemesterSales"]+Stats3[1,"MonthlySales"]/2
Stats4[2,"SemesterSalesJanNovDec"]=Stats4[2,"SemesterSales"]+Stats3[11,"MonthlySales"]/2+Stats3[12,"MonthlySales"]/2
Stats4$SemesterSalesJanNovDec.Billions=Stats4$SemesterSalesJanNovDec/1000000000
View(Stats4)

	semester	SemesterSales	SemesterSales.Billions	SemesterSalesJanNovDec	SemesterSalesJanNovDec.Billions
1	1	3320727474	3.320727	3487026693	3.487027
2	2	3416491513	3.416492	3911418693	3.911419

#1. Which Store Has Maximum Sales?
#14 had the highest week of sales.
print(Stats1[which.max(Stats1$Max_WeeklySales),1])

```
A tibble: 1 x 1
Store
<int>
   14
```

#2. Which Store has Maximum Standard Deviation?
#14 has the highest standard deviation among its Weekly Sales.
print(Stats1[which.max(Stats1$StdDev_WeeklySales),1])

```
A tibble: 1 x 1
Store
<int>
   14
```

#3. Which Store has the Greatest CV (Coefficient of Mean to Standard Deviation)?
#35 has the greatest CV for its Weekly Sales.
print(Stats1[which.max(Stats1$CV_WeeklySales),1])

```
A tibble: 1 x 1
Store
<int>
   35
```

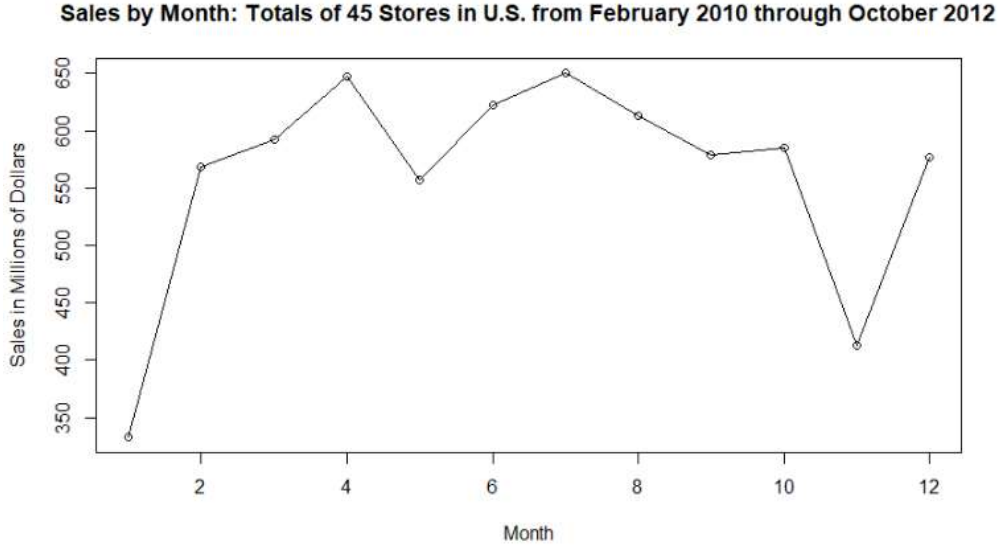
#4 Which store/s has good quarterly growth rate in Q3'2021?
#7 has the best quarterly growth rate from Q2 to Q3 2021.
#Store #16 had the second best quarterly growth rate.
print(Stats1[which.max(Stats1$GrowthRate_Q3.Q2),1])

```
A tibble: 1 x 1
Store
<int>
    7
```

#5 Which holidays have higher sales than the mean sales in non-holiday season (all stores)
#Super Bowl, Labor Day, and Thanksgiving all have higher sales than average non-holiday periods
#Christmas has lower sales than non-holiday periods
View(Stats2)

	Metric	Non_Holiday	SuperBowl	LabourDay	Thanksgiving	Christmas
1	TotalSales	1041256	1079128	1042427	1471273	960833.1

#6 Monthly and Semester View of Sales
#Monthly
plot(Stats3$month,Stats3$MonthlySales.Millions,type="o",main="Sales by Month: Totals of 45 Stores in U.S. from February 2010 through October 2012",xlab="Month",ylab="Sales in Millions of Dollars")



plot(Stats3$month,Stats3$MonthlySalesJanNovDec.Millions,type="o",main="Sales by Month: Average Sales of 45 Stores in U.S.",xlab="Month",ylab="Sales in Millions of Dollars")

#Insights

#This is very interesting. I have a few thoughts:

- If little is going on and the weather is warm, people seem to spend more money (May is a crunch to complete the school year)

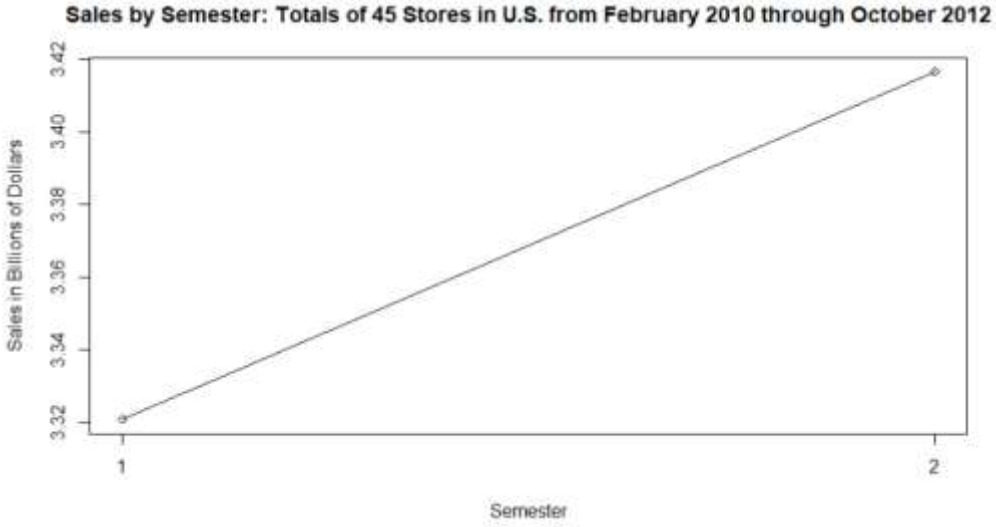
- Christmas seems to create dismal January sales (people try to spend less in general because of high spendings on Christmas gifts/activities).

- Although the week of Christmas yields low sales, sales in the month of Christmas (December) benefit from Christmas.

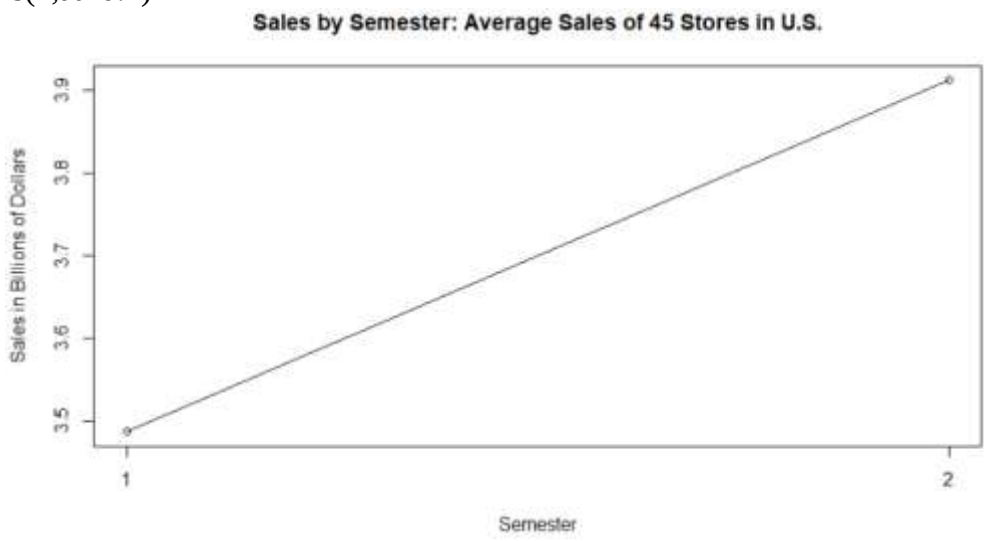
#Semester

plot(Stats4$semester,Stats4$SemesterSales.Billions,type="o",xaxt="none",main="Sales by Semester: Totals of 45 Stores in U.S. from February 2010 through October 2012",xlab="Semester",ylab="Sales in Billions of Dollars")

axis(1,at=0:2)



plot(Stats4$semester,Stats4$SemesterSalesJanNovDec.Billions,type="o",xaxt="none",main="Sales by Semester: Average Sales of 45 Stores in U.S.",xlab="Semester",ylab="Sales in Billions of Dollars")
axis(1,at=0:2)



#Insights:
#Not too much insight from me here.
- There is a nearly 11% difference between Semester 1 and Semester 2 across all stores and years.
- The presence of Christmas in Semester 2 is a likely culprit for this difference.

#==
#Statistical Model Section:
#==
#For Store 1: Build Prediction Models to Forecast Demand, and
Select the Linear Regression model which gives best accuracy

#Create new DF for Store 1
Store1=filter(DF1,Store==1)
View(Store1)

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester
1	1	2010-02-05	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1
2	1	2010-02-12	1641957	1	38.51	2.548	211.2422	8.106	2010	2	1	1

#Restructure Dates as 1 for 5 Feb 2010:
Store1$Day=as.numeric(Store1$Date)-14645+1 #Convert all Dates into Numeric values, where Day1 (5 Feb 2010) has value of 1465
View(Store1)

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester	Day
1	1	2010-02-05	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1	1
2	1	2010-02-12	1641957	1	38.51	2.548	211.2422	8.106	2010	2	1	1	8

Store1LR=Store1[,-c(1,2)] #Removes Store and Date columns
View(Store1LR)

	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester	Day
1	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1	1
2	1641957	1	38.51	2.548	211.2422	8.106	2010	2	1	1	8

#My Hypotheses (before looking at graphs, data, etc.):
CPI: Will not impact sales
Unemployment: Will impact sales
Fuel Price: will impact sales, but not in a linear manner

#--------------------------
#Linear Regression Model 1:
#--------------------------
Store1aLR=Store1LR

#Check for NAs: none
summary(Store1aLR)

```
  Weekly_Sales      Holiday_Flag      Temperature      Fuel_Price         CPI
 Min.   :1316899   Min.   :0.00000   Min.   :35.40   Min.   :2.514   Min.   :210.3
 1st Qu.:1458105   1st Qu.:0.00000   1st Qu.:58.27   1st Qu.:2.764   1st Qu.:211.5
 Median :1534850   Median :0.00000   Median :69.64   Median :3.290   Median :215.5
 Mean   :1555264   Mean   :0.06993   Mean   :68.31   Mean   :3.220   Mean   :216.0
 3rd Qu.:1614892   3rd Qu.:0.00000   3rd Qu.:80.48   3rd Qu.:3.594   3rd Qu.:220.5
 Max.   :2387950   Max.   :1.00000   Max.   :91.65   Max.   :3.907   Max.   :223.4
  Unemployment        year          month           quarter         semester
 Min.   :6.573   Min.   :2010   Min.   : 1.000   Min.   :1.000   Min.   :1.000
 1st Qu.:7.348   1st Qu.:2010   1st Qu.: 4.000   1st Qu.:2.000   1st Qu.:1.000
 Median :7.787   Median :2011   Median : 6.000   Median :2.000   Median :1.000
 Mean   :7.610   Mean   :2011   Mean   : 6.448   Mean   :2.483   Mean   :1.497
 3rd Qu.:7.838   3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:3.000   3rd Qu.:2.000
 Max.   :8.106   Max.   :2012   Max.   :12.000   Max.   :4.000   Max.   :2.000
      Day
 Min.   :  1.0
 1st Qu.:249.5
 Median :498.0
 Mean   :498.0
 3rd Qu.:746.5
 Max.   :995.0
```

#Check for Correlations with DV (Weekly_Sales): Poor correlation among all variables
cor(Store1aLR)

```
              Weekly_Sales Holiday_Flag Temperature  Fuel_Price         CPI Unemployment         year
Weekly_Sales   1.00000000   0.19490521 -0.22270056  0.12459158  0.22540766  -0.09795539  0.15239570
Holiday_Flag   0.19490521   1.00000000 -0.20054304 -0.08590253 -0.02891916   0.08294894 -0.05678257
Temperature   -0.22270056  -0.20054304  1.00000000  0.22849268  0.11850334  -0.18069498  0.06884342
Fuel_Price     0.12459158  -0.08590253  0.22849268  1.00000000  0.75525865  -0.51394406  0.80976859
CPI            0.22540766  -0.02891916  0.11850334  0.75525865  1.00000000  -0.81347056  0.94814064
Unemployment  -0.09795539   0.08294894 -0.18069498 -0.51394406 -0.81347056   1.00000000 -0.79814895
year           0.15239570  -0.05678257  0.06884342  0.80976859  0.94814064  -0.79814895  1.00000000
month          0.20218780   0.12299577  0.24641700 -0.10125622  0.05095169   0.04082096 -0.19446452
quarter        0.13500354   0.08136344  0.25141226 -0.10121457  0.04692421   0.01148622 -0.18523825
semester       0.04150098   0.11160194  0.29377885 -0.12743109  0.05440567   0.01628344 -0.13192950
Day            0.21453922  -0.01328524  0.15406940  0.78178912  0.97394350  -0.79122155  0.94166795
                   month     quarter    semester         Day
Weekly_Sales  0.20218780  0.13500354  0.04150098  0.21453922
Holiday_Flag  0.12299577  0.08136344  0.11160194 -0.01328524
Temperature   0.24641700  0.25141226  0.29377885  0.15406940
Fuel_Price   -0.10125622 -0.10121457 -0.12743109  0.78178912
CPI           0.05095169  0.04692421  0.05440567  0.97394350
Unemployment  0.04082096  0.01148622  0.01628344 -0.79122155
year         -0.19446452 -0.18523825 -0.13192950  0.94166795
month         1.00000000  0.96707047  0.86052064  0.14565116
quarter       0.96707047  1.00000000  0.88550939  0.14392240
semester      0.86052064  0.88550939  1.00000000  0.16161742
Day           0.14565116  0.14392240  0.16161742  1.00000000
```

#Check for Multicollinearity: Removes year, Day, quarter
vifstep(Store1aLR[,-1],th=10)

```
3 variables from the 10 input variables have collinearity problem:

year Day quarter

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( semester ~ Unemployment ):  0.01628344
max correlation ( semester ~ month ):  0.8605206

---------- VIFs of the remained variables --------
     Variables      VIF
1 Holiday_Flag 1.082907
2  Temperature 1.417153
3   Fuel_Price 3.244455
4          CPI 6.787102
5 Unemployment 3.679463
6        month 3.904681
7     semester 4.173579
```

#Generate Best Model, minimize AIC
Fit1=step(lm(Weekly_Sales~Temperature+CPI+month+Holiday_Flag+Fuel_Price+Unemployment+semester,data=Store1aLR))

```
Start:  AIC=3394.34
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Fuel_Price +
    Unemployment + semester

               Df  Sum of Sq        RSS    AIC
- Fuel_Price    1 3.0137e+08 2.6031e+12 3392.4
- Unemployment  1 1.1447e+10 2.6142e+12 3393.0
<none>                      2.6028e+12 3394.3
- Holiday_Flag  1 4.8355e+10 2.6512e+12 3395.0
- CPI           1 6.3317e+10 2.6661e+12 3395.8
- Temperature   1 1.3311e+11 2.7359e+12 3399.5
- semester      1 1.6660e+11 2.7694e+12 3401.2
- month         1 3.2546e+11 2.9283e+12 3409.2

Step:  AIC=3392.36
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Unemployment +
    semester

               Df  Sum of Sq        RSS    AIC
- Unemployment  1 1.1814e+10 2.6149e+12 3391.0
<none>                      2.6031e+12 3392.4
- Holiday_Flag  1 4.8474e+10 2.6516e+12 3393.0
- CPI           1 1.2693e+11 2.7300e+12 3397.2
- Temperature   1 1.6137e+11 2.7645e+12 3399.0
- semester      1 1.7174e+11 2.7748e+12 3399.5
- month         1 3.2593e+11 2.9290e+12 3407.2

Step:  AIC=3391.01
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + semester

               Df  Sum of Sq        RSS    AIC
<none>                      2.6149e+12 3391.0
- Holiday_Flag  1 5.0687e+10 2.6656e+12 3391.8
- semester      1 1.7233e+11 2.7872e+12 3398.1
- Temperature   1 1.8170e+11 2.7966e+12 3398.6
- CPI           1 2.1490e+11 2.8298e+12 3400.3
- month         1 3.4175e+11 2.9567e+12 3406.6
```

summary(Fit1)

```
Call:
lm(formula = Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag +
    semester, data = Store1aLR)

Residuals:
    Min      1Q  Median      3Q     Max
-367200  -87799   -6388   64929  740790

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  -193898.6   576419.4  -0.336  0.73710
Temperature    -2723.2      882.6  -3.085  0.00246 **
CPI             9007.4     2684.4   3.355  0.00103 **
month          29680.2     7014.3   4.231 4.23e-05 ***
Holiday_Flag   76740.7    47091.9   1.630  0.10548
semester     -138400.7    46060.2  -3.005  0.00316 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 138200 on 137 degrees of freedom
Multiple R-squared:  0.2431,    Adjusted R-squared:  0.2155
F-statistic: 8.801 on 5 and 137 DF,  p-value: 2.901e-07
```

#month(***),Temperature+CPI+semester(**),Holiday_Flag( )
#Multiple R-squared 0.2431

#MAPE
Fit1$residuals

Fit1$residuals/Store1LR$Weekly_Sales #PE

abs(Fit1$residuals/Store1LR$Weekly_Sales) #APE

mean(abs(Fit1$residuals/Store1LR$Weekly_Sales)) #MAPE
#Returns 0.06042918

```
[1] 0.06042918
```

#Predict
a1=as.data.frame(predict(Fit1,newdata = Store1aLR))
names(a1)[1]="Predicted"
a1$Actual=Store1LR$Weekly_Sales
a1$Difference=a1$Predicted-a1$Actual
a1$absDifference=abs(a1$Difference)
a1$Percent_Error=a1$Difference/a1$Actual*100
a1$absPercent_Error=abs(a1$Percent_Error)

```
#Predict
a1=as.data.frame(predict(Fit1,newdata = Store1aLR))
#Predict
a1=as.data.frame(predict(Fit1,newdata = Store1aLR))
names(a1)[1]="Predicted"
a1$Actual=Store1LR$Weekly_Sales
a1$Difference=a1$Predicted-a1$Actual
a1$absDifference=abs(a1$Difference)
a1$Percent_Error=a1$Difference/a1$Actual*100
a1$absPercent_Error=abs(a1$Percent_Error)
```

```
#--------------------------
#Linear Regression Model 2:
#--------------------------
#Create Second DF for Store1, this time considering Holidays to be outliers
#Replace all values for Weekly_Sales on Holidays with the mean of Weekly_Sales
Store1bLR=Store1aLR
Store1bLR$Weekly_Sales=ifelse(Store1bLR$Holiday_Flag>0,mean(Store1bLR$Weekly_Sales),Store1bLR$Weekly_Sales)
View(Store1bLR)
```

	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester	Day
1	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1	1
2	1555264	1	38.51	2.548	211.2422	8.106	2010	2	1	1	8

```
summary(Store1bLR) #Check for NAs: none
```

```
  Weekly_Sales      Holiday_Flag      Temperature      Fuel_Price         CPI         Unemployment
Min.   :1316899   Min.   :0.00000   Min.   :35.40   Min.   :2.514   Min.   :210.3   Min.   :6.573
1st Qu.:1459505   1st Qu.:0.00000   1st Qu.:58.27   1st Qu.:2.764   1st Qu.:211.5   1st Qu.:7.348
Median :1540164   Median :0.00000   Median :69.64   Median :3.290   Median :215.5   Median :7.787
Mean   :1547538   Mean   :0.06993   Mean   :68.31   Mean   :3.220   Mean   :216.0   Mean   :7.610
3rd Qu.:1604365   3rd Qu.:0.00000   3rd Qu.:80.48   3rd Qu.:3.594   3rd Qu.:220.5   3rd Qu.:7.838
Max.   :2387950   Max.   :1.00000   Max.   :91.65   Max.   :3.907   Max.   :223.4   Max.   :8.106
      year          month           quarter         semester          Day
Min.   :2010   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :  1.0
1st Qu.:2010   1st Qu.: 4.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:249.5
Median :2011   Median : 6.000   Median :2.000   Median :1.000   Median :498.0
Mean   :2011   Mean   : 6.448   Mean   :2.483   Mean   :1.497   Mean   :498.0
3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:746.5
Max.   :2012   Max.   :12.000   Max.   :4.000   Max.   :2.000   Max.   :995.0
```

```
cor(Store1bLR) #Check for Correlations with DV (Weekly_Sales): Poor correlation among all variables
```

```
              Weekly_Sales Holiday_Flag Temperature  Fuel_Price         CPI Unemployment        year
Weekly_Sales   1.00000000   0.01480667 -0.20798106  0.14433550  0.22913200  -0.11380170  0.15690552
Holiday_Flag   0.01480667   1.00000000 -0.20054304 -0.08590253 -0.02891916   0.08294894 -0.05678257
Temperature   -0.20798106  -0.20054304  1.00000000  0.22849268  0.11850334  -0.18069498  0.06884342
Fuel_Price     0.14433550  -0.08590253  0.22849268  1.00000000  0.75525865  -0.51394406  0.80976859
CPI            0.22913200  -0.02891916  0.11850334  0.75525865  1.00000000  -0.81347056  0.94814064
Unemployment  -0.11380170   0.08294894 -0.18069498 -0.51394406 -0.81347056   1.00000000 -0.79814895
year           0.15690552  -0.05678257  0.06884342  0.80976859  0.94814064  -0.79814895  1.00000000
month          0.20694043   0.12299577  0.24641700 -0.10125622  0.05095169   0.04082096 -0.19446452
quarter        0.13076982   0.08136344  0.25141226 -0.10121457  0.04692421   0.01148622 -0.18523825
semester       0.03257009   0.11160194  0.29377885 -0.12743109  0.05440567   0.01628344 -0.13192950
Day            0.21948872  -0.01328524  0.15406940  0.78178912  0.97394350  -0.79122155  0.94166795
                   month     quarter    semester         Day
Weekly_Sales  0.20694043  0.13076982  0.03257009  0.21948872
Holiday_Flag  0.12299577  0.08136344  0.11160194 -0.01328524
Temperature   0.24641700  0.25141226  0.29377885  0.15406940
Fuel_Price   -0.10125622 -0.10121457 -0.12743109  0.78178912
CPI           0.05095169  0.04692421  0.05440567  0.97394350
Unemployment  0.04082096  0.01148622  0.01628344 -0.79122155
year         -0.19446452 -0.18523825 -0.13192950  0.94166795
month         1.00000000  0.96707047  0.86052064  0.14565116
quarter       0.96707047  1.00000000  0.88550939  0.14392240
semester      0.86052064  0.88550939  1.00000000  0.16161742
Day           0.14565116  0.14392240  0.16161742  1.00000000
```

vifstep(Store1bLR[,-1],th=10) #Check for Multicollinearity: Removes year, Day, quarter

```
3 variables from the 10 input variables have collinearity problem:

year Day quarter

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( semester ~ Unemployment ):  0.01628344
max correlation ( semester ~ month ):  0.8605206

---------- VIFs of the remained variables --------
     Variables      VIF
1 Holiday_Flag 1.082907
2  Temperature 1.417153
3   Fuel_Price 3.244455
4          CPI 6.787102
5 Unemployment 3.679463
6        month 3.904681
7     semester 4.173579
```

#Generate Best Model, minimize AIC. year, Day, quarter excluded
Fit2=step(lm(Weekly_Sales~Temperature+CPI+month+Holiday_Flag+Fuel_Price+Unemployment+semester,data=Store1bLR))

```
Start:  AIC=3371.34
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Fuel_Price +
    Unemployment + semester

               Df  Sum of Sq        RSS    AIC
- Fuel_Price    1 9.0307e+08 2.2170e+12 3369.4
- Unemployment  1 3.6559e+09 2.2198e+12 3369.6
- Holiday_Flag  1 1.1762e+10 2.2279e+12 3370.1
<none>                       2.2161e+12 3371.3
- CPI           1 3.4451e+10 2.2506e+12 3371.5
- semester      1 1.5730e+11 2.3734e+12 3379.1
- Temperature   1 1.5893e+11 2.3751e+12 3379.2
- month         1 3.4356e+11 2.5597e+12 3390.0

Step:  AIC=3369.4
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Unemployment +
    semester

               Df  Sum of Sq        RSS    AIC
- Unemployment  1 6.0957e+09 2.2231e+12 3367.8
- Holiday_Flag  1 1.1858e+10 2.2289e+12 3368.2
<none>                       2.2170e+12 3369.4
- CPI           1 9.6268e+10 2.3133e+12 3373.5
- semester      1 1.7074e+11 2.3878e+12 3378.0
- Temperature   1 1.7536e+11 2.3924e+12 3378.3
- month         1 3.4303e+11 2.5601e+12 3388.0

Step:  AIC=3367.79
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + semester

               Df  Sum of Sq        RSS    AIC
- Holiday_Flag  1 1.1147e+10 2.2343e+12 3366.5
<none>                       2.2231e+12 3367.8
- semester      1 1.7116e+11 2.3943e+12 3376.4
- CPI           1 1.8248e+11 2.4056e+12 3377.1
- Temperature   1 1.9211e+11 2.4152e+12 3377.6
- month         1 3.5569e+11 2.5788e+12 3387.0

Step:  AIC=3366.51
Weekly_Sales ~ Temperature + CPI + month + semester

              Df  Sum of Sq        RSS    AIC
<none>                      2.2343e+12 3366.5
- semester     1 1.7636e+11 2.4106e+12 3375.4
- Temperature  1 1.8096e+11 2.4152e+12 3375.6
- CPI          1 1.8338e+11 2.4177e+12 3375.8
- month        1 3.5014e+11 2.5844e+12 3385.3
```

summary(Fit2)

```
Call:
lm(formula = Weekly_Sales ~ Temperature + CPI + month + semester,
    data = Store1bLR)

Residuals:
    Min      1Q  Median      3Q     Max
-205183  -87253  -10794   65120  740306

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -53746.6   530637.5  -0.101  0.91947
Temperature   -2636.2      788.5  -3.343  0.00107 **
CPI            8320.2     2472.2   3.365  0.00099 ***
month         30002.3     6451.5   4.650 7.67e-06 ***
semester    -139810.3    42360.9  -3.300  0.00123 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 127200 on 138 degrees of freedom
Multiple R-squared:  0.2368,    Adjusted R-squared:  0.2147
F-statistic:  10.7 on 4 and 138 DF,  p-value: 1.386e-07
```

#CPI+month(***),Temperature+semester(**)
#Multiple R-squared 0.2368

#MAPE
Fit2$residuals

Fit2$residuals/Store1LR$Weekly_Sales #PE

abs(Fit2$residuals/Store1LR$Weekly_Sales) #APE

mean(abs(Fit2$residuals/Store1LR$Weekly_Sales)) #MAPE
#Returns 0.05594411

```
[1] 0.05594411
```

#Predict
a2=as.data.frame(predict(Fit2,newdata = Store1bLR))
names(a2)[1]="Predicted"
a2$Actual=Store1LR$Weekly_Sales
a2$Difference=a2$Predicted-a2$Actual
a2$absDifference=abs(a2$Difference)
a2$Percent_Error=a2$Difference/a2$Actual*100
a2$absPercent_Error=abs(a2$Percent_Error)

```
> #Predict
> a2=as.data.frame(predict(Fit2,newdata = Store1bLR))
> names(a2)[1]="Predicted"
> a2$Actual=Store1LR$Weekly_Sales
> a2$Difference=a2$Predicted-a2$Actual
> a2$absDifference=abs(a2$Difference)
> a2$Percent_Error=a2$Difference/a2$Actual*100
> a2$absPercent_Error=abs(a2$Percent_Error)
```

```
#--------------------------
#Linear Regression Model 3:
#--------------------------
#Create Third DF for Store1, this time considering all sales in December to be outliers
#Replace all values for Weekly_Sales in December with the mean of Weekly_Sales
Store1cLR=Store1bLR
Store1cLR$Weekly_Sales=ifelse(Store1cLR$month==12,mean(Store1cLR$Weekly_Sales),Store1cLR$Weekly
_Sales)
View(Store1cLR)
```

	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	year	month	quarter	semester	Day
1	1643691	0	42.31	2.572	211.0964	8.106	2010	2	1	1	1
2	1555264	1	38.51	2.548	211.2422	8.106	2010	2	1	1	8

```
summary(Store1cLR) #Check for NAs: none
  Weekly_Sales      Holiday_Flag      Temperature      Fuel_Price         CPI          Unemployment
 Min.   :1316899   Min.   :0.00000   Min.   :35.40   Min.   :2.514   Min.   :210.3   Min.   :6.573
 1st Qu.:1459505   1st Qu.:0.00000   1st Qu.:58.27   1st Qu.:2.764   1st Qu.:211.5   1st Qu.:7.348
 Median :1540164   Median :0.00000   Median :69.64   Median :3.290   Median :215.5   Median :7.787
 Mean   :1528798   Mean   :0.06993   Mean   :68.31   Mean   :3.220   Mean   :216.0   Mean   :7.610
 3rd Qu.:1590679   3rd Qu.:0.00000   3rd Qu.:80.48   3rd Qu.:3.594   3rd Qu.:220.5   3rd Qu.:7.838
 Max.   :1899677   Max.   :1.00000   Max.   :91.65   Max.   :3.907   Max.   :223.4   Max.   :8.106
      year          month           quarter         semester          Day
 Min.   :2010   Min.   : 1.000   Min.   :1.000   Min.   :1.000   Min.   :  1.0
 1st Qu.:2010   1st Qu.: 4.000   1st Qu.:2.000   1st Qu.:1.000   1st Qu.:249.5
 Median :2011   Median : 6.000   Median :2.000   Median :1.000   Median :498.0
 Mean   :2011   Mean   : 6.448   Mean   :2.483   Mean   :1.497   Mean   :498.0
 3rd Qu.:2012   3rd Qu.: 9.000   3rd Qu.:3.000   3rd Qu.:2.000   3rd Qu.:746.5
 Max.   :2012   Max.   :12.000   Max.   :4.000   Max.   :2.000   Max.   :995.0
```

```
cor(Store1cLR) #Check for Correlations with DV (Weekly_Sales): Poor correlation among all variables
              Weekly_Sales Holiday_Flag Temperature  Fuel_Price         CPI Unemployment        year
Weekly_Sales   1.00000000   0.07069549 -0.05206692  0.31053147  0.36787844  -0.29099691  0.34422409
Holiday_Flag   0.07069549   1.00000000 -0.20054304 -0.08590253 -0.02891916   0.08294894 -0.05678257
Temperature   -0.05206692  -0.20054304  1.00000000  0.22849268  0.11850334  -0.18069498  0.06884342
Fuel_Price     0.31053147  -0.08590253  0.22849268  1.00000000  0.75525865  -0.51394406  0.80976859
CPI            0.36787844  -0.02891916  0.11850334  0.75525865  1.00000000  -0.81347056  0.94814064
Unemployment  -0.29099691   0.08294894 -0.18069498 -0.51394406 -0.81347056   1.00000000 -0.79814895
year           0.34422409  -0.05678257  0.06884342  0.80976859  0.94814064  -0.79814895  1.00000000
month         -0.02613775   0.12299577  0.24641700 -0.10125622  0.05095169   0.04082096 -0.19446452
quarter       -0.08140531   0.08136344  0.25141226 -0.10121457  0.04692421   0.01148622 -0.18523825
semester      -0.14702781   0.11160194  0.29377885 -0.12743109  0.05440567   0.01628344 -0.13192950
Day            0.32174005  -0.01328524  0.15406940  0.78178912  0.97394350  -0.79122155  0.94166795
                  month     quarter     semester         Day
Weekly_Sales -0.02613775 -0.08140531 -0.14702781  0.32174005
Holiday_Flag  0.12299577  0.08136344  0.11160194 -0.01328524
Temperature   0.24641700  0.25141226  0.29377885  0.15406940
Fuel_Price   -0.10125622 -0.10121457 -0.12743109  0.78178912
CPI           0.05095169  0.04692421  0.05440567  0.97394350
Unemployment  0.04082096  0.01148622  0.01628344 -0.79122155
year         -0.19446452 -0.18523825 -0.13192950  0.94166795
month         1.00000000  0.96707047  0.86052064  0.14565116
quarter       0.96707047  1.00000000  0.88550939  0.14392240
semester      0.86052064  0.88550939  1.00000000  0.16161742
Day           0.14565116  0.14392240  0.16161742  1.00000000
>
```

vifstep(Store1cLR[,-1],th=10) #Check for Multicollinearity: Removes year, Day, quarter

```
3 variables from the 10 input variables have collinearity problem:

year Day quarter

After excluding the collinear variables, the linear correlation coefficients ranges between:
min correlation ( semester ~ Unemployment ):  0.01628344
max correlation ( semester ~ month ):  0.8605206

---------- VIFs of the remained variables --------
     Variables      VIF
1 Holiday_Flag 1.082907
2  Temperature 1.417153
3   Fuel_Price 3.244455
4          CPI 6.787102
5 Unemployment 3.679463
6        month 3.904681
7     semester 4.173579
> |
```

#Generate Best Model, minimize AIC. year, Day, quarter excluded

Fit3=step(lm(Weekly_Sales~Temperature+CPI+month+Holiday_Flag+Fuel_Price+Unemployment+semester,data=Store1cLR))

```
Start:  AIC=3265.22
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Fuel_Price +
    Unemployment + semester

               Df  Sum of Sq        RSS    AIC
- Unemployment  1 9.2562e+07 1.0552e+12 3263.2
- Fuel_Price    1 5.9065e+08 1.0557e+12 3263.3
- Temperature   1 1.3468e+09 1.0565e+12 3263.4
- Holiday_Flag  1 8.9639e+09 1.0641e+12 3264.4
<none>                       1.0551e+12 3265.2
- CPI           1 2.2803e+10 1.0779e+12 3266.3
- month         1 4.7364e+10 1.1025e+12 3269.5
- semester      1 7.3458e+10 1.1286e+12 3272.8

Step:  AIC=3263.24
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + Fuel_Price +
    semester

               Df  Sum of Sq        RSS    AIC
- Fuel_Price    1 4.9816e+08 1.0557e+12 3261.3
- Temperature   1 1.2555e+09 1.0565e+12 3261.4
- Holiday_Flag  1 8.8993e+09 1.0641e+12 3262.4
<none>                       1.0552e+12 3263.2
- month         1 4.7413e+10 1.1026e+12 3267.5
- CPI           1 6.7228e+10 1.1225e+12 3270.1
- semester      1 7.4327e+10 1.1296e+12 3271.0

Step:  AIC=3261.3
Weekly_Sales ~ Temperature + CPI + month + Holiday_Flag + semester

               Df  Sum of Sq        RSS    AIC
- Temperature   1 9.0409e+08 1.0566e+12 3259.4
- Holiday_Flag  1 8.9149e+09 1.0646e+12 3260.5
<none>                       1.0557e+12 3261.3
- month         1 4.7644e+10 1.1034e+12 3265.6
- semester      1 8.0215e+10 1.1359e+12 3269.8
- CPI           1 1.9145e+11 1.2472e+12 3283.1
```

```
Step:  AIC=3259.43
Weekly_Sales ~ CPI + month + Holiday_Flag + semester

               Df  Sum of Sq        RSS    AIC
- Holiday_Flag  1 1.0998e+10 1.0676e+12 3258.9
<none>                       1.0566e+12 3259.4
- month         1 4.7658e+10 1.1043e+12 3263.7
- semester      1 8.5714e+10 1.1423e+12 3268.6
- CPI           1 1.9076e+11 1.2474e+12 3281.2

Step:  AIC=3258.91
Weekly_Sales ~ CPI + month + semester

           Df  Sum of Sq        RSS    AIC
<none>                   1.0676e+12 3258.9
- month     1 5.0287e+10 1.1179e+12 3263.5
- semester  1 8.4982e+10 1.1526e+12 3267.9
- CPI       1 1.8774e+11 1.2554e+12 3280.1
```

summary(Fit3)

```
Call:
lm(formula = Weekly_Sales ~ CPI + month + semester, data = Store1cLR)

Residuals:
    Min      1Q  Median      3Q     Max
-229236  -61118   -4454   53243  305653

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -209085     365370  -0.572  0.56807
CPI             8370       1693   4.944 2.17e-06 ***
month          11369       4443   2.559  0.01158 *
semester      -95732      28780  -3.326  0.00113 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 87640 on 139 degrees of freedom
Multiple R-squared:  0.201,     Adjusted R-squared:  0.1837
F-statistic: 11.65 on 3 and 139 DF,  p-value: 7.374e-07
```

#CPI(***),semester(**),month(*)

#Multiple R-squared 0.201

#MAPE
Fit3$residuals

Fit3$residuals/Store1LR$Weekly_Sales #PE

abs(Fit3$residuals/Store1LR$Weekly_Sales) #APE

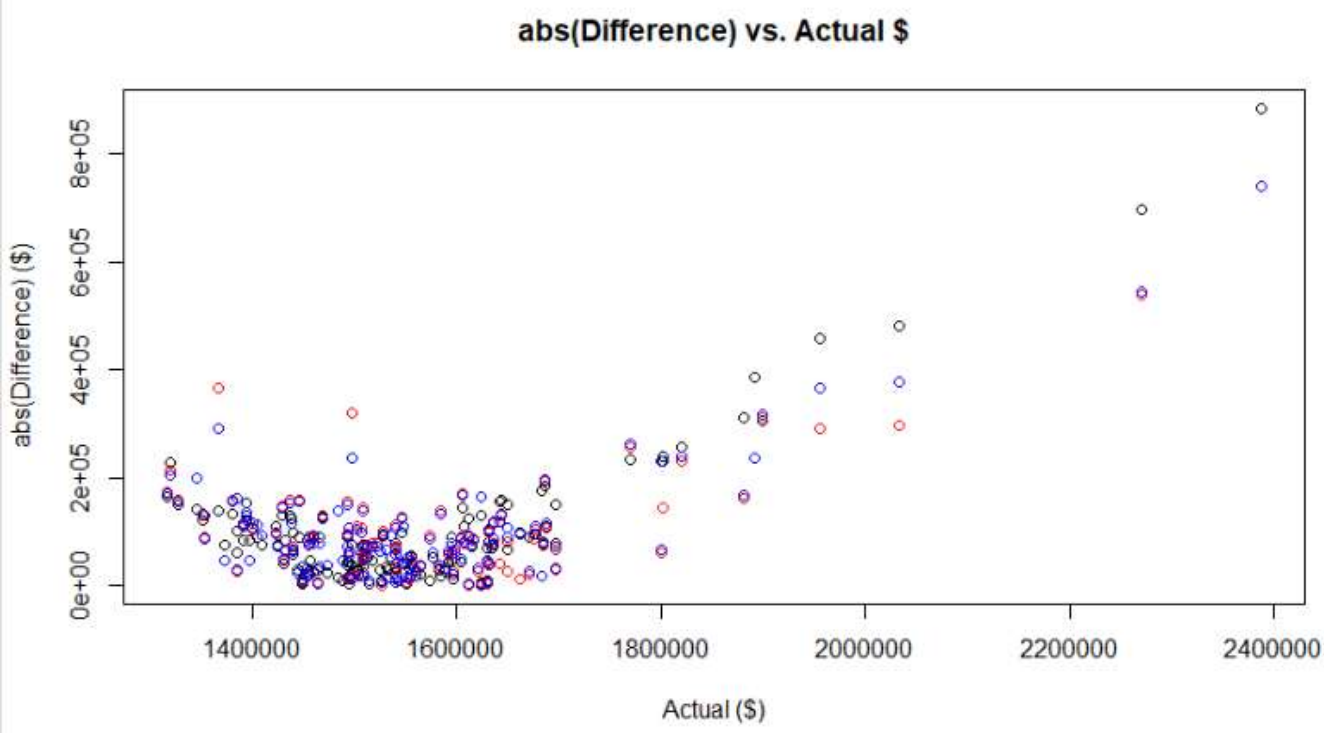
mean(abs(Fit3$residuals/Store1LR$Weekly_Sales)) #MAPE
#Returns 0.04471027

```
[1] 0.04471027
```

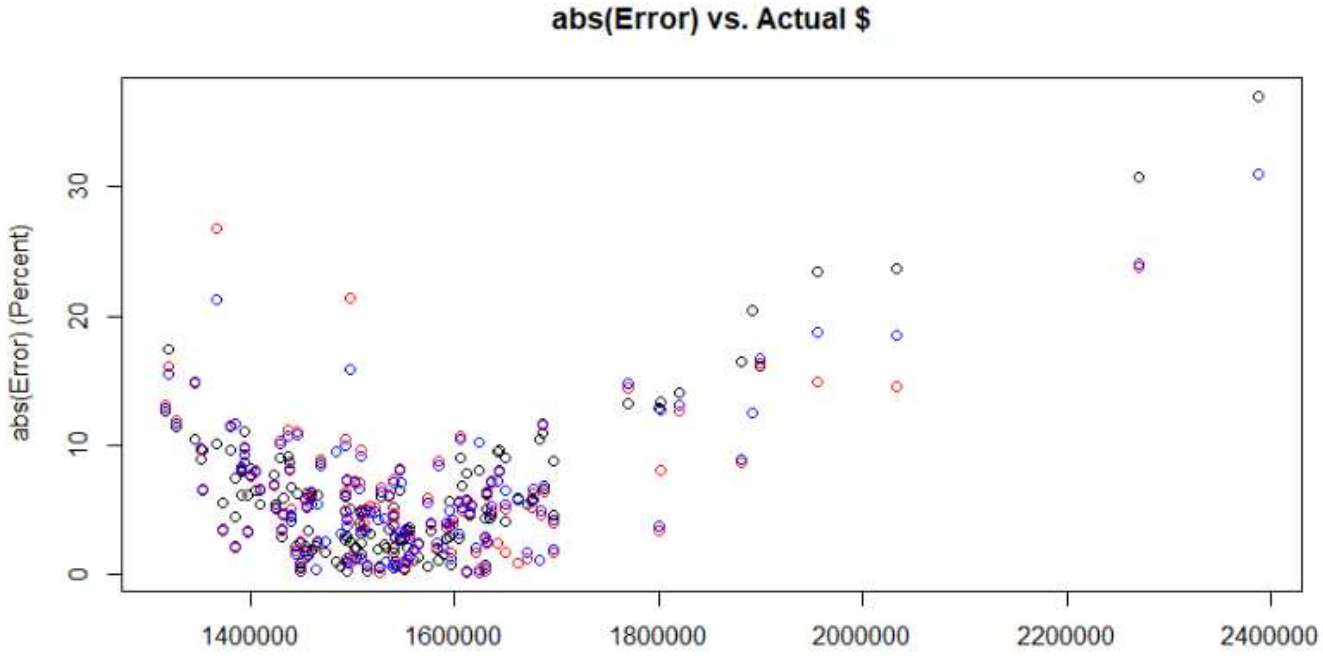
#Predict
a3=as.data.frame(predict(Fit3,newdata = Store1cLR))
names(a3)[1]="Predicted"
a3$Actual=Store1LR$Weekly_Sales
a3$Difference=a3$Predicted-a3$Actual
a3$absDifference=abs(a3$Difference)
a3$Percent_Error=a3$Difference/a3$Actual*100
a3$absPercent_Error=abs(a3$Percent_Error)


```
#Predict
a3=as.data.frame(predict(Fit3,newdata = Store1cLR))
names(a3)[1]="Predicted"
a3$Actual=Store1LR$Weekly_Sales
a3$Difference=a3$Predicted-a3$Actual
a3$absDifference=abs(a3$Difference)
a3$Percent_Error=a3$Difference/a3$Actual*100
a3$absPercent_Error=abs(a3$Percent_Error)
```

```
#--------------------------
#Visualization of Results: Models 1, 2, 3
#--------------------------
plot(a3$Actual,a3$absDifference,type="p",main="abs(Difference) vs. Actual $",xlab="Actual
($)",ylab="abs(Difference) ($)")
points(a1$Actual,a1$absDifference,type="p",col="red")
points(a2$Actual,a2$absDifference,type="p",col="blue")
```




```
plot(a3$Actual,a3$absPercent_Error,type="p",main="abs(Error) vs. Actual $",xlab="Actual
($)",ylab="abs(Error) (Percent)")
points(a1$Actual,a1$absPercent_Error,type="p",col="red")
points(a2$Actual,a2$absPercent_Error,type="p",col="blue")
```



Summary of Logistic Models

I created three models to predict weekly sales of Store 1. The models were Fit1, Fit2, and Fit3. Fit1 took all supplied data as-is. Fit2 took data where some outliers were smoothed out. Lastly Fit3 took data where there were additional outliers smoothed out. With each smoothing of outliers came an increase in accuracy. Fit3 was selected as the most accurate model because it resulted in lower Standard Error and MAPE values. A model summary is displayed below.

Data	Approach	Model Name	Detail	R2	adj R2	std err	R2 - adj R2	MAPE	Priority
Store1aLR	VIF, Step	Fit1	-year -day - quarter	24.31	21.55	138200	2.76	6.04%	3
Store1bLR	Manual, VIF, Step	Fit2	Replace Holiday Week Sales w/ Average Sales, -year -day -quarter	23.68	21.47	127200	2.21	5.59%	2
Store1cLR	Manual, VIF, Step	Fit3	Replace Holiday Week Sales w/ Average Sales, Replace December Sales w/ Average Sales, -year -day -quarter	20.1	18.37	87640	1.73	4.47%	1

Conclusions:

There is no denying it, predicting retail sales is no simple task. Even if predicted weekly sales for a store were 5% off of the actual sales of $1.5million, then the prediction would be $75k off. That $75k could represent the annual salary two or three employees. Of course, the metrics used for the prediction model developed in this exercise are very high-level. If the employer were hoping to develop a model with greater accuracy then they may look for more specific variables to track, for instance trends in sales by department. Moving forward, it would be interesting to apply the model developed for Store 1 toward predicting sales in the other 44 stores. I would anticipate regional differences among stores that would yield less accurate predictions for the other 44 stores. Perhaps the regional differences could be captured in a new variable?

If a store did not previously have any means to predict sales, then the model that I developed could provide the management with some value. It would at least represent a starting point that could certainly be improved upon to better predict sales.