# DS502
# Final Report

- Group 6

Xiuhan li, Enbo Tian, Ziyang Xu

# 1. Introduction

Warmart is one of the largest supermarkets in the United States. There are stores located in different cities. Large dataset of sales is generated each year. There are problems with how to manage the data. We want to focus on the dataset beyond 2010 to 2012, defining the related data and predict a model for the following time.

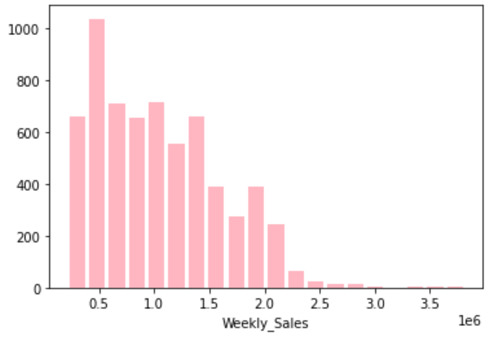
# 2. Data Processing

## Introduction of the dataset

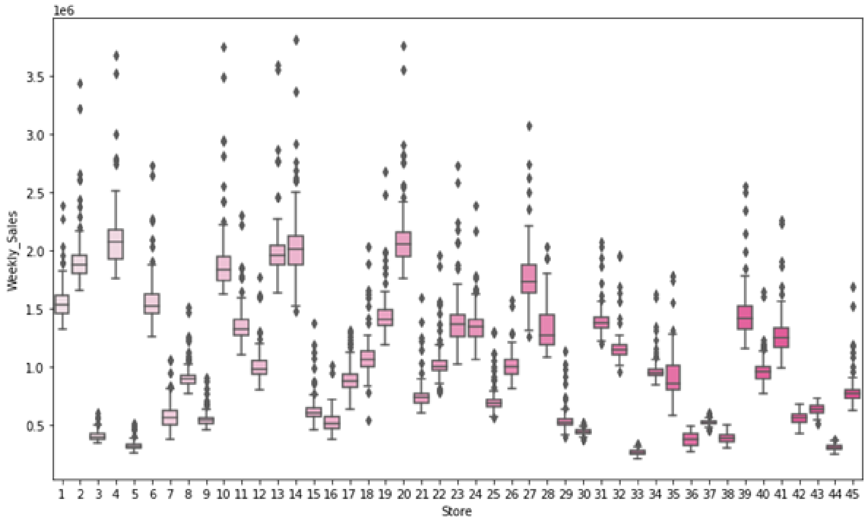
Our original dataset contains 6435 rows and 8 columns without any duplicated rows. There are 3 categorical predictors and 4 numeric predictors. The target label of our dataset is column 'weekly_sales', which makes our dataset have only one column of label. The predictors include date, holiday flag, store information, unemployment and temperature.

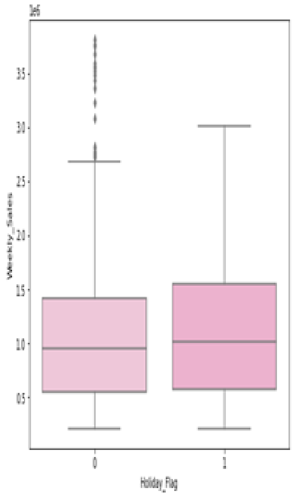
## Data Exploration

We first explored the distribution of our target label. Which can be seen in Figure 1.

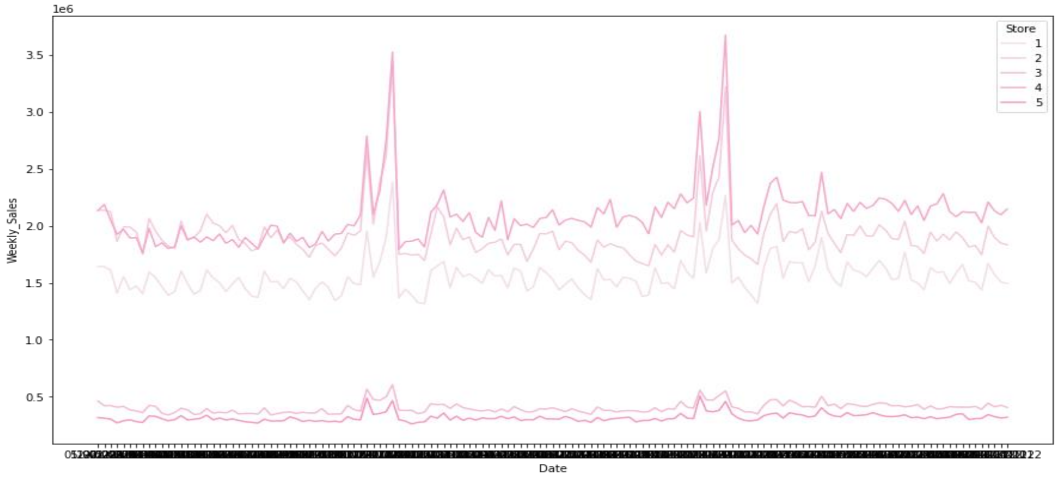


Next we researched the relationship between label and predictors. And the boxplot of sales versus stores as well as holiday can be seen in Figure 2. As can be seen in the figure, the variance of sales between stores is quite large, and there' s no relatively obvious linear relationship between store and sales, so we would consider one-hot encoding in later modeling. Also we may consider one hot encoding for holiday flag.

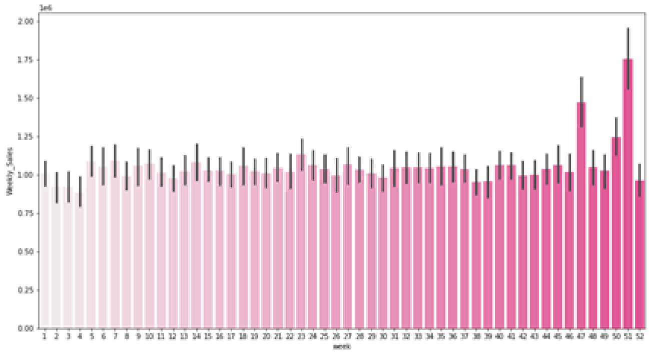


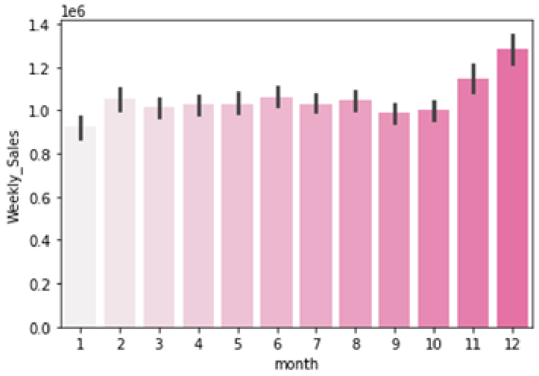


To see the change of sales along with time. We first checked the average sales of all stores, from 2010-02-05 to 2012-11-01, which showed some time varying patterns. To see sales of each store over time, we plotted the first 5 stores as an example, and the same pattern appeared again. (Figure 3)

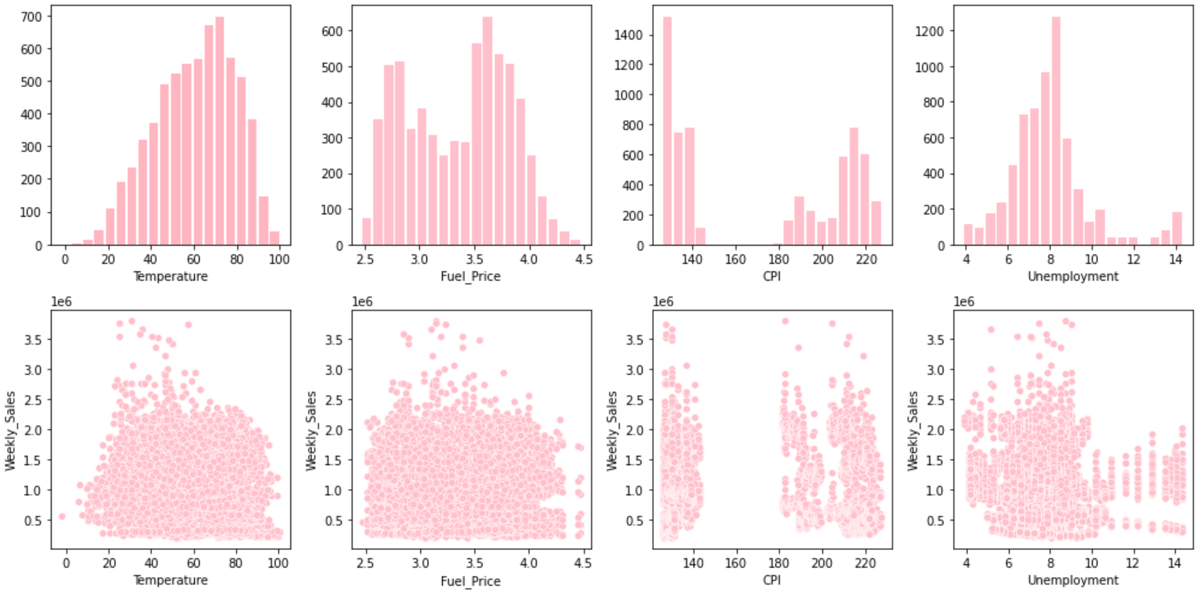


To further explore the time varying pattern, we further divided the predictor “date” into “month” and “week” to see their relationship. As shown in Figure 4, high sales tend to appear in the last 2 months of a year.

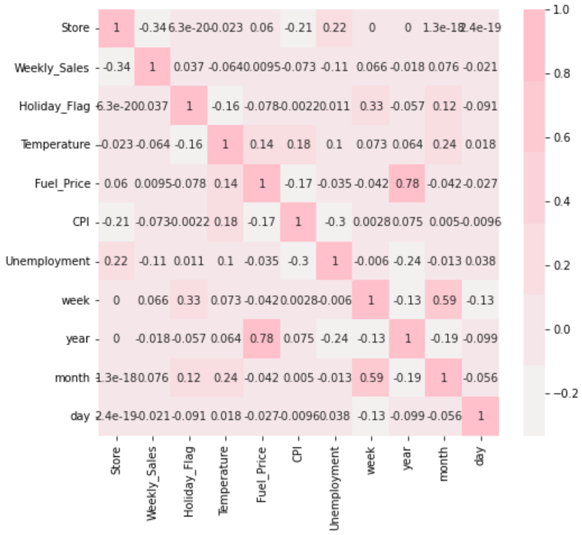




Then we displayed the numerical predictors in Figure 5, where the upper half are the distribution of the predictors and the bottom half are the scatterplot of predictors together with labels. From the distribution plot we could see that, except f or the predictor 'CPI', all other predictors are relatively normally distributed, so there's no skewness correction needed. And from the scatter plot, obvious positive relationship can be discovered in none of the predictors, which would be further verified by Figure 6.



Finally the correlation heatmap of all features is shown in Figure 6.



# 3. Feature Engineering

## 1. Missing Values

Since the missing values only make up less than 5% percent of total dataset. We drop rows with missing values.

## 2. Categorical variables

**a. one-hot encoding:** We perform one-hot encoding on predictor 'holiday flag' and predictor 'store'.

b. **change into numerical:** For the predictor “date”, we split the date into year, week, month to further capture more features from date.

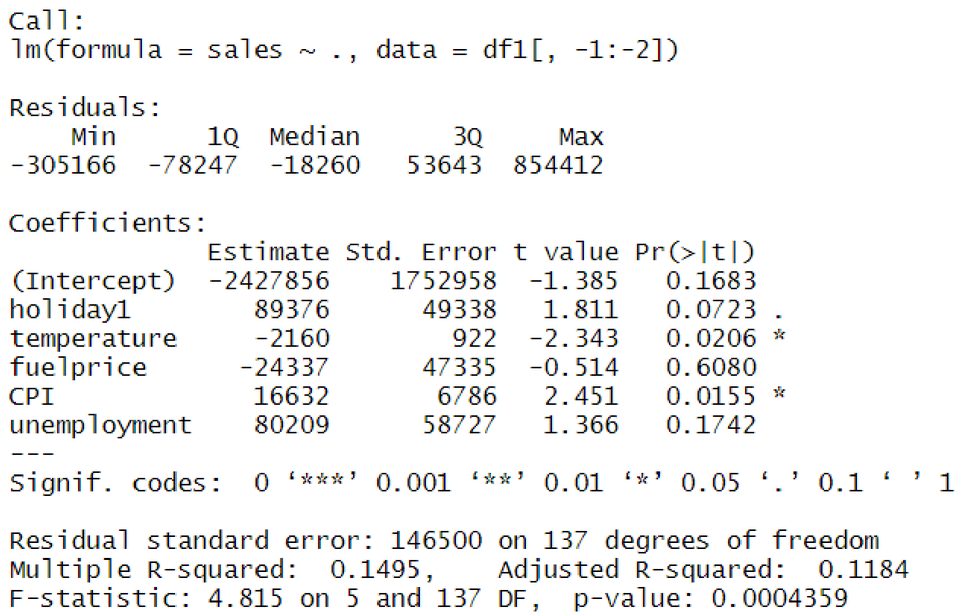
3. Outliers

   We deducted outliers for predictors “unemployment” and “temperature”.

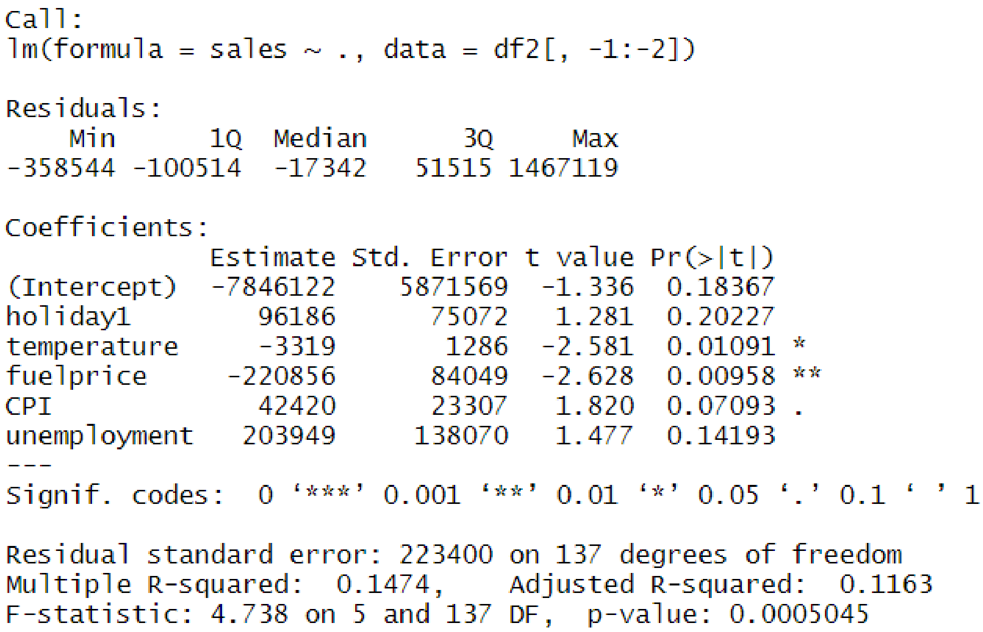
# 4. Model

## 1. Regression

We are trying to fit a linear regression Model, and make the regression Model more accurate. It is necessary to show the linear model for each store, Since for each store there may be a different relationship between sales and the factors. Here is an example for store 1, store 2:



```
Call:
lm(formula = sales ~ ., data = df1[, -1:-2])

Residuals:
    Min      1Q  Median      3Q     Max 
-305166  -78247  -18260   53643  854412 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)  
(Intercept)  -2427856    1752958  -1.385   0.1683  
holiday1        89376      49338   1.811   0.0723 .
temperature     -2160        922  -2.343   0.0206 *
fuelprice      -24337      47335  -0.514   0.6080  
CPI             16632       6786   2.451   0.0155 *
unemployment    80209      58727   1.366   0.1742  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 146500 on 137 degrees of freedom
Multiple R-squared:  0.1495,    Adjusted R-squared:  0.1184 
F-statistic: 4.815 on 5 and 137 DF,  p-value: 0.0004359
```

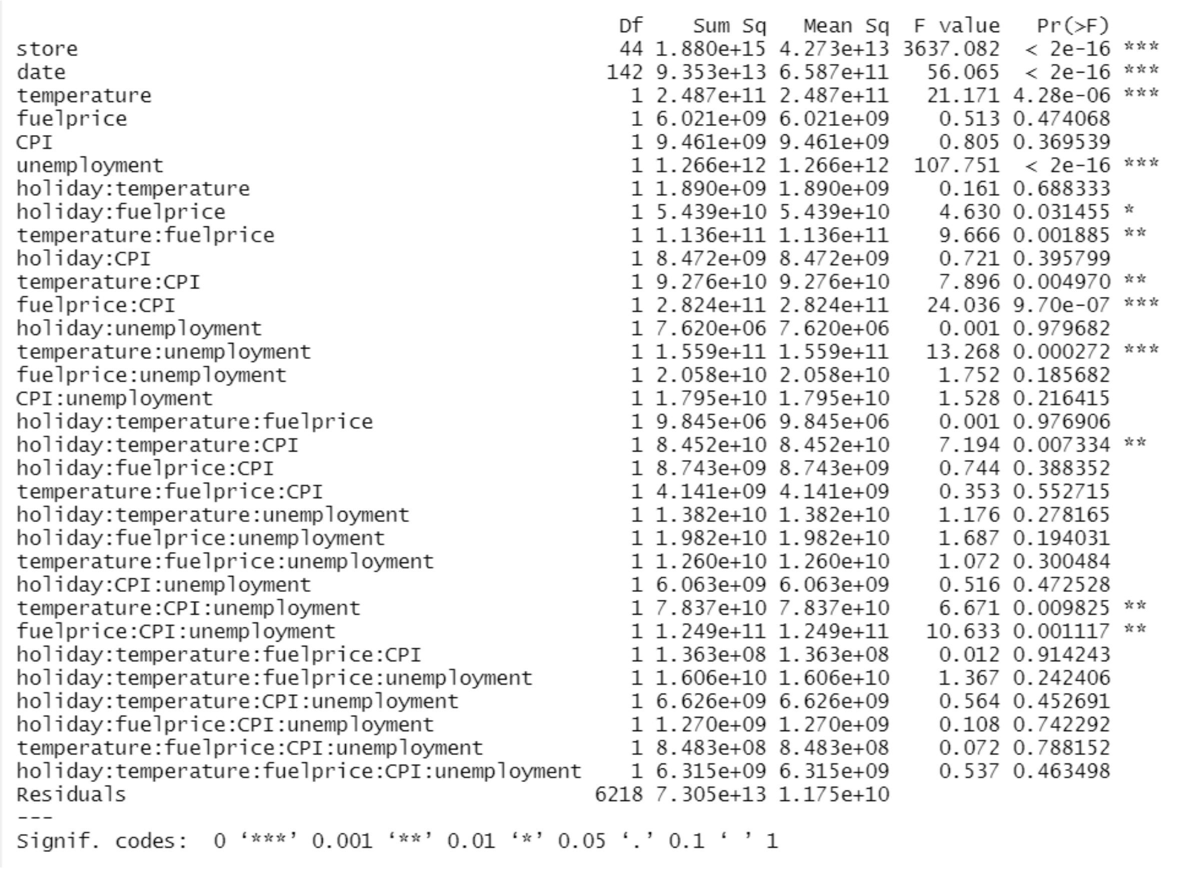


```
Call:
lm(formula = sales ~ ., data = df2[, -1:-2])

Residuals:
    Min      1Q  Median      3Q     Max 
-358544 -100514  -17342   51515 1467119 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)   
(Intercept)  -7846122    5871569  -1.336  0.18367   
holiday1        96186      75072   1.281  0.20227   
temperature     -3319       1286  -2.581  0.01091 * 
fuelprice     -220856      84049  -2.628  0.00958 **
CPI             42420      23307   1.820  0.07093 . 
unemployment   203949     138070   1.477  0.14193   
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 223400 on 137 degrees of freedom
Multiple R-squared:  0.1474,	Adjusted R-squared:  0.1163 
F-statistic: 4.738 on 5 and 137 DF,  p-value: 0.0005045
```

From the T test, It shows that temperature and CPI are the main factors for the sales of Store 1. Fuel price temperature are the main factors for the sales of Store 2. We can see that the main factor is changing for different stores. So we would like to count the times of importance of each factor for each store, and drop the factors which have significantly less count than the others. We find that the factor holiday has a count of 12, temperature has a count of 19, fuel price has a count of 4, CPI has a count of 20, and the factor of unemployment has a count of 12. This shows that Fuel price has much less count than the other factors. If we can not fit it with a transform of log or quadratic, we may drop it from the linear model. We find that the log transform of fuel price has a count of 3, and the quadratic has a count of 4.

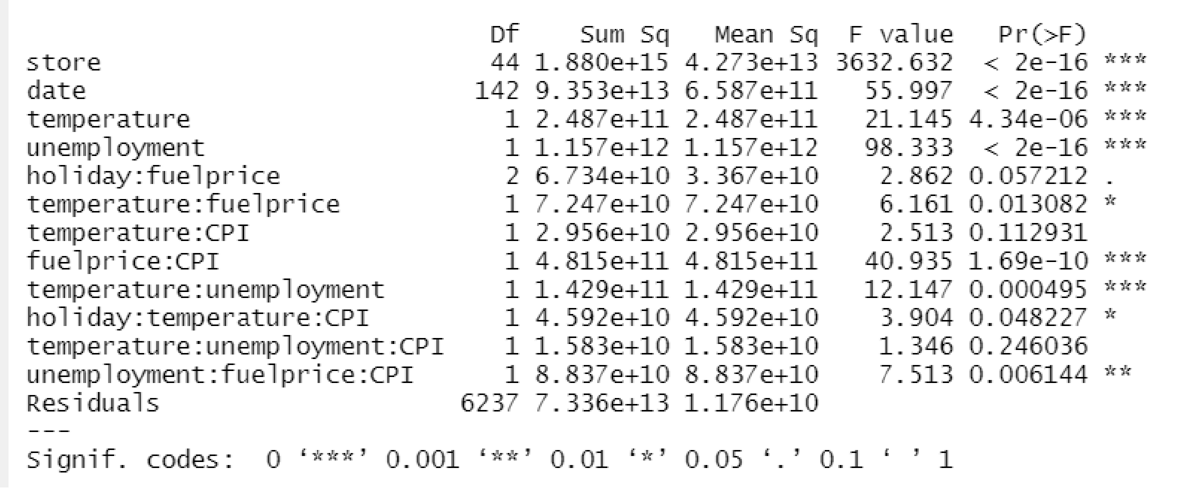
As a conclusion, Fuel price can be dropped from the linear model of each store, but when we fit a model for all of the stores, the factor of Fuel price remains. The following step is to fit a model of all stores.

Firstly we want to fit an anova table to show which of the factors and interactions are important. In this model, stores and the week date are also factors, because we combine all the stores together, and there are the same week days for each of the stores. Here is the anova table:

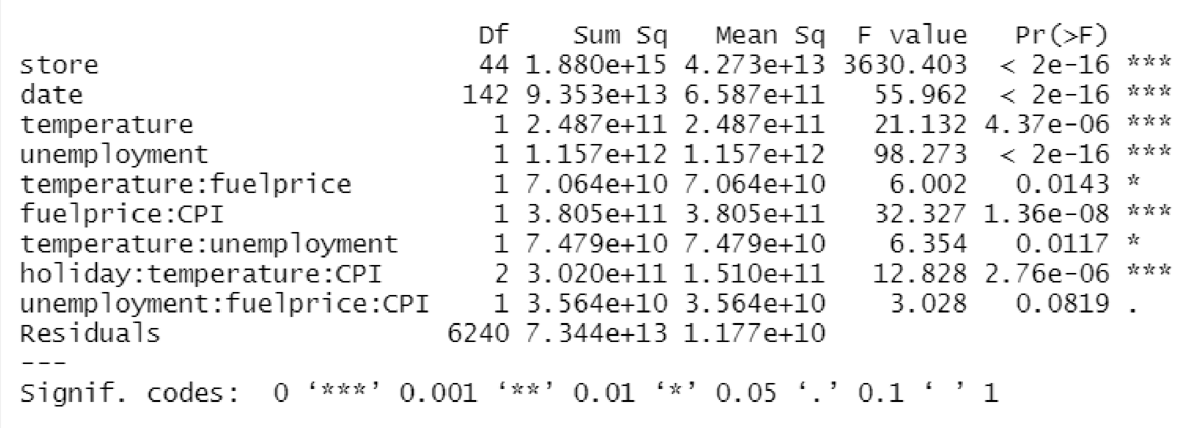


```
                                              Df    Sum Sq   Mean Sq  F value   Pr(>F)    
store                                         44 1.880e+15 4.273e+13 3637.082  < 2e-16 ***
date                                         142 9.353e+13 6.587e+11   56.065  < 2e-16 ***
temperature                                    1 2.487e+11 2.487e+11   21.171 4.28e-06 ***
fuelprice                                      1 6.021e+09 6.021e+09    0.513 0.474068    
CPI                                            1 9.461e+09 9.461e+09    0.805 0.369539    
unemployment                                   1 1.266e+12 1.266e+12  107.751  < 2e-16 ***
holiday:temperature                            1 1.890e+09 1.890e+09    0.161 0.688333    
holiday:fuelprice                              1 5.439e+10 5.439e+10    4.630 0.031455 *  
temperature:fuelprice                          1 1.136e+11 1.136e+11    9.666 0.001885 ** 
holiday:CPI                                    1 8.472e+09 8.472e+09    0.721 0.395799    
temperature:CPI                                1 9.276e+10 9.276e+10    7.896 0.004970 ** 
fuelprice:CPI                                  1 2.824e+11 2.824e+11   24.036 9.70e-07 ***
holiday:unemployment                           1 7.620e+06 7.620e+06    0.001 0.979682    
temperature:unemployment                       1 1.559e+11 1.559e+11   13.268 0.000272 ***
fuelprice:unemployment                         1 2.058e+10 2.058e+10    1.752 0.185682    
CPI:unemployment                               1 1.795e+10 1.795e+10    1.528 0.216415    
holiday:temperature:fuelprice                  1 9.845e+06 9.845e+06    0.001 0.976906    
holiday:temperature:CPI                        1 8.452e+10 8.452e+10    7.194 0.007334 ** 
holiday:fuelprice:CPI                          1 8.743e+09 8.743e+09    0.744 0.388352    
temperature:fuelprice:CPI                      1 4.141e+09 4.141e+09    0.353 0.552715    
holiday:temperature:unemployment               1 1.382e+10 1.382e+10    1.176 0.278165    
holiday:fuelprice:unemployment                 1 1.982e+10 1.982e+10    1.687 0.194031    
temperature:fuelprice:unemployment             1 1.260e+10 1.260e+10    1.072 0.300484    
holiday:CPI:unemployment                       1 6.063e+09 6.063e+09    0.516 0.472528    
temperature:CPI:unemployment                   1 7.837e+10 7.837e+10    6.671 0.009825 ** 
fuelprice:CPI:unemployment                     1 1.249e+11 1.249e+11   10.633 0.001117 ** 
holiday:temperature:fuelprice:CPI              1 1.363e+08 1.363e+08    0.012 0.914243    
holiday:temperature:fuelprice:unemployment     1 1.606e+10 1.606e+10    1.367 0.242406    
holiday:temperature:CPI:unemployment           1 6.626e+09 6.626e+09    0.564 0.452691    
holiday:fuelprice:CPI:unemployment             1 1.270e+09 1.270e+09    0.108 0.742292    
temperature:fuelprice:CPI:unemployment         1 8.483e+08 8.483e+08    0.072 0.788152    
holiday:temperature:fuelprice:CPI:unemployment  1 6.315e+09 6.315e+09    0.537 0.463498    
Residuals                                   6218 7.305e+13 1.175e+10                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From the F test of the ANOVA table, we can find the factors with a significant effect, Which have at least a (*) at the end of the table. We reduce the other factors and use the significant factors to fit another ANOVA table to get our final result:

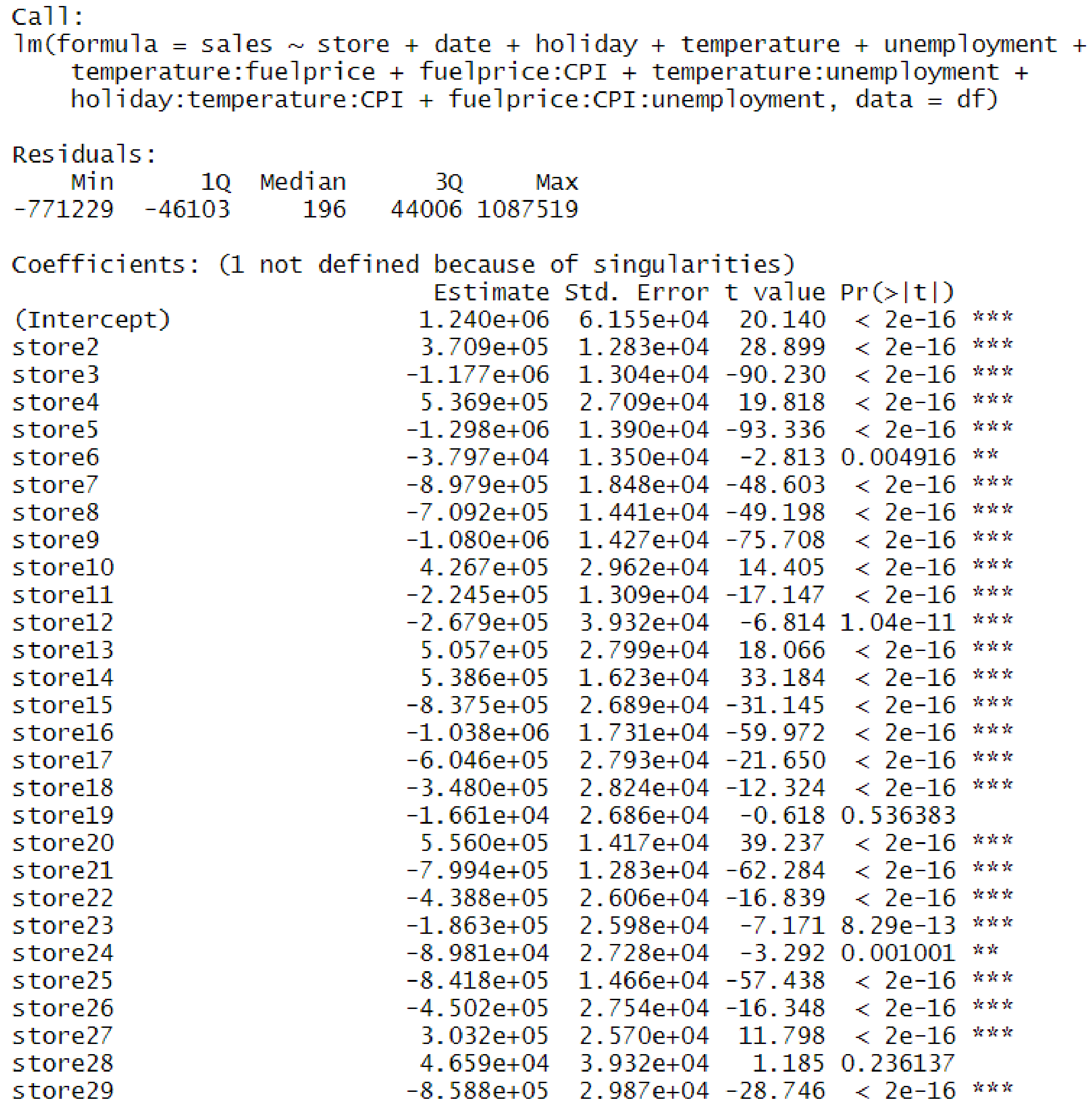


```
                               Df    Sum Sq   Mean Sq  F value   Pr(>F)    
store                          44 1.880e+15 4.273e+13 3632.632  < 2e-16 ***
date                          142 9.353e+13 6.587e+11   55.997  < 2e-16 ***
temperature                     1 2.487e+11 2.487e+11   21.145 4.34e-06 ***
unemployment                    1 1.157e+12 1.157e+12   98.333  < 2e-16 ***
holiday:fuelprice               2 6.734e+10 3.367e+10    2.862 0.057212 .  
temperature:fuelprice           1 7.247e+10 7.247e+10    6.161 0.013082 *  
temperature:CPI                 1 2.956e+10 2.956e+10    2.513 0.112931    
fuelprice:CPI                   1 4.815e+11 4.815e+11   40.935 1.69e-10 ***
temperature:unemployment        1 1.429e+11 1.429e+11   12.147 0.000495 ***
holiday:temperature:CPI         1 4.592e+10 4.592e+10    3.904 0.048227 *  
temperature:unemployment:CPI    1 1.583e+10 1.583e+10    1.346 0.246036    
unemployment:fuelprice:CPI      1 8.837e+10 8.837e+10    7.513 0.006144 ** 
Residuals                    6237 7.336e+13 1.176e+10                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



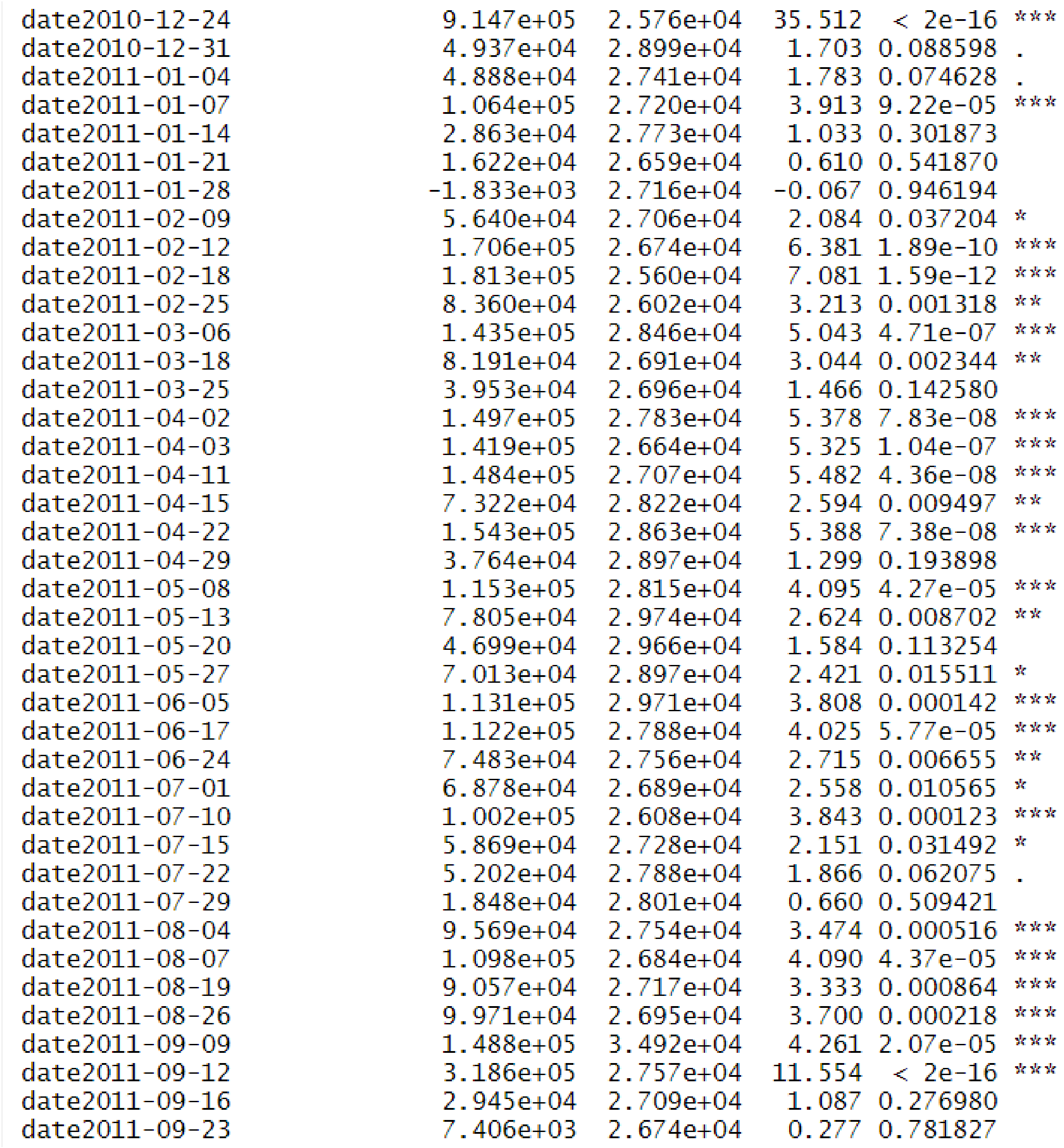
```
                              Df    Sum Sq   Mean Sq  F value    Pr(>F)    
store                         44 1.880e+15 4.273e+13 3630.403  < 2e-16 ***
date                         142 9.353e+13 6.587e+11   55.962  < 2e-16 ***
temperature                    1 2.487e+11 2.487e+11   21.132 4.37e-06 ***
unemployment                   1 1.157e+12 1.157e+12   98.273  < 2e-16 ***
temperature:fuelprice          1 7.064e+10 7.064e+10    6.002   0.0143 *  
fuelprice:CPI                  1 3.805e+11 3.805e+11   32.327 1.36e-08 ***
temperature:unemployment       1 7.479e+10 7.479e+10    6.354   0.0117 *  
holiday:temperature:CPI        2 3.020e+11 1.510e+11   12.828 2.76e-06 ***
unemployment:fuelprice:CPI     1 3.564e+10 3.564e+10    3.028   0.0819 .  
Residuals                   6240 7.344e+13 1.177e+10                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We can now fit a model with the factors and interactions, Since Stores and date have a degree of freedom 44, and 142, we have 44+142 estimators for the Stores and date.

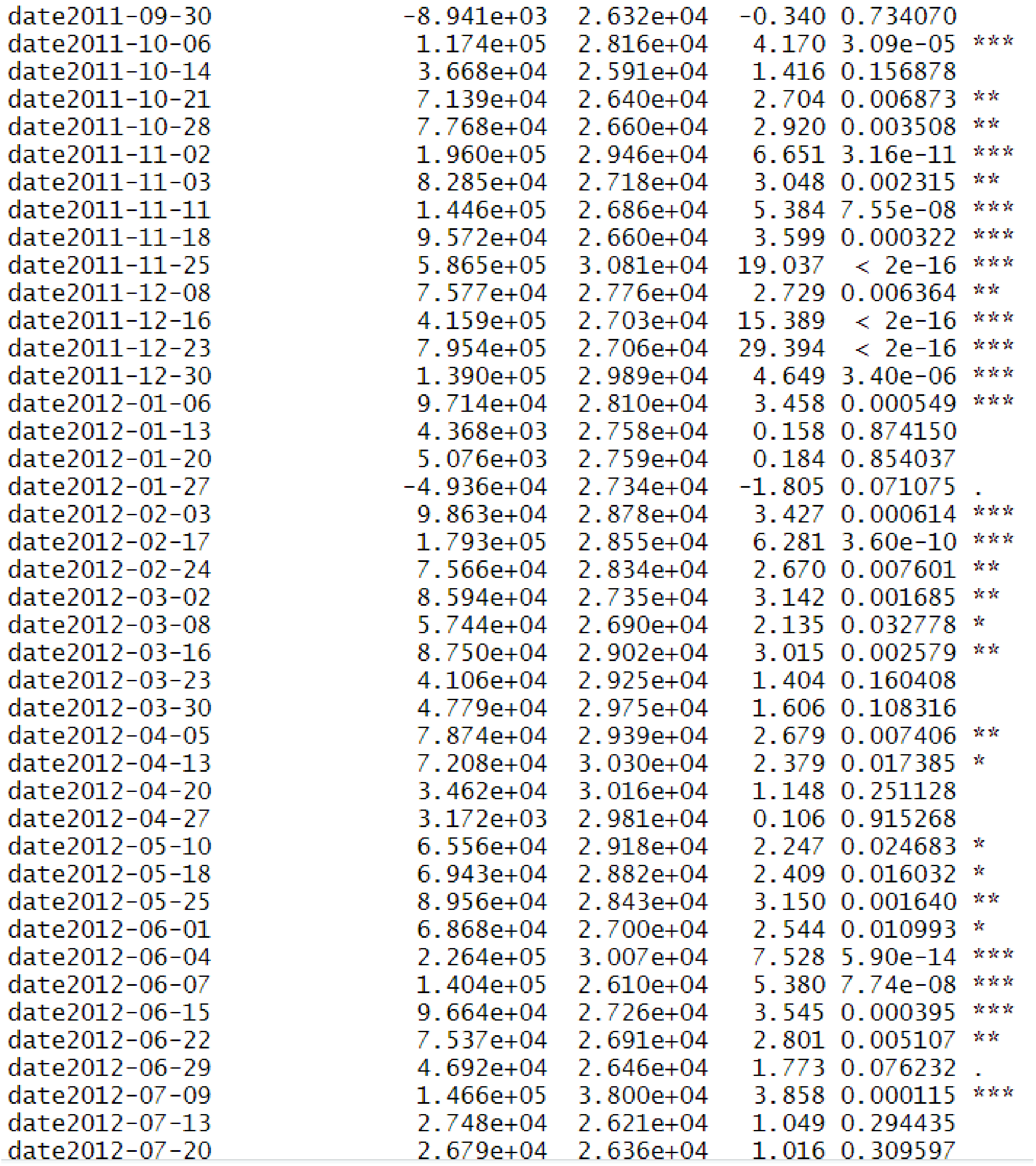


```
Call:
lm(formula = sales ~ store + date + holiday + temperature + unemployment + 
    temperature:fuelprice + fuelprice:CPI + temperature:unemployment + 
    holiday:temperature:CPI + fuelprice:CPI:unemployment, data = df)

Residuals:
    Min      1Q  Median      3Q     Max 
-771229  -46103     196   44006 1087519 

Coefficients: (1 not defined because of singularities)
                             Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 1.240e+06  6.155e+04  20.140  < 2e-16 ***
store2                      3.709e+05  1.283e+04  28.899  < 2e-16 ***
store3                     -1.177e+06  1.304e+04 -90.230  < 2e-16 ***
store4                      5.369e+05  2.709e+04  19.818  < 2e-16 ***
store5                     -1.298e+06  1.390e+04 -93.336  < 2e-16 ***
store6                     -3.797e+04  1.350e+04  -2.813 0.004916 ** 
store7                     -8.979e+05  1.848e+04 -48.603  < 2e-16 ***
store8                     -7.092e+05  1.441e+04 -49.198  < 2e-16 ***
store9                     -1.080e+06  1.427e+04 -75.708  < 2e-16 ***
store10                     4.267e+05  2.962e+04  14.405  < 2e-16 ***
store11                    -2.245e+05  1.309e+04 -17.147  < 2e-16 ***
store12                    -2.679e+05  3.932e+04  -6.814 1.04e-11 ***
store13                     5.057e+05  2.799e+04  18.066  < 2e-16 ***
store14                     5.386e+05  1.623e+04  33.184  < 2e-16 ***
store15                    -8.375e+05  2.689e+04 -31.145  < 2e-16 ***
store16                    -1.038e+06  1.731e+04 -59.972  < 2e-16 ***
store17                    -6.046e+05  2.793e+04 -21.650  < 2e-16 ***
store18                    -3.480e+05  2.824e+04 -12.324  < 2e-16 ***
store19                    -1.661e+04  2.686e+04  -0.618 0.536383    
store20                     5.560e+05  1.417e+04  39.237  < 2e-16 ***
store21                    -7.994e+05  1.283e+04 -62.284  < 2e-16 ***
store22                    -4.388e+05  2.606e+04 -16.839  < 2e-16 ***
store23                    -1.863e+05  2.598e+04  -7.171 8.29e-13 ***
store24                    -8.981e+04  2.728e+04  -3.292 0.001001 ** 
store25                    -8.418e+05  1.466e+04 -57.438  < 2e-16 ***
store26                    -4.502e+05  2.754e+04 -16.348  < 2e-16 ***
store27                     3.032e+05  2.570e+04  11.798  < 2e-16 ***
store28                     4.659e+04  3.932e+04   1.185 0.236137    
store29                    -8.588e+05  2.987e+04 -28.746  < 2e-16 ***
```

```
store30                      -1.117e+06  1.283e+04 -87.021  < 2e-16 ***
store31                      -1.596e+05  1.283e+04 -12.432  < 2e-16 ***
store32                      -3.161e+05  1.617e+04 -19.546  < 2e-16 ***
store33                      -1.213e+06  3.015e+04 -40.212  < 2e-16 ***
store34                      -4.317e+05  3.219e+04 -13.409  < 2e-16 ***
store35                      -5.238e+05  2.686e+04 -19.500  < 2e-16 ***
store36                      -1.173e+06  1.292e+04 -90.795  < 2e-16 ***
store37                      -1.028e+06  1.291e+04 -79.590  < 2e-16 ***
store38                      -8.912e+05  3.932e+04 -22.665  < 2e-16 ***
store39                      -9.510e+04  1.290e+04  -7.374 1.87e-13 ***
store40                      -6.099e+05  2.603e+04 -23.433  < 2e-16 ***
store41                      -2.744e+05  1.639e+04 -16.745  < 2e-16 ***
store42                      -9.163e+05  2.962e+04 -30.932  < 2e-16 ***
store43                      -8.146e+05  1.591e+04 -51.196  < 2e-16 ***
store44                      -1.205e+06  2.777e+04 -43.403  < 2e-16 ***
store45                      -6.964e+05  1.623e+04 -42.909  < 2e-16 ***
date2010-02-04                2.180e+05  2.387e+04   9.135  < 2e-16 ***
date2010-02-07                1.332e+05  2.301e+04   5.789 7.43e-09 ***
date2010-02-19                2.077e+05  2.601e+04   7.985 1.66e-15 ***
date2010-02-26                1.079e+05  2.572e+04   4.196 2.75e-05 ***
date2010-03-09                9.488e+04  2.301e+04   4.123 3.78e-05 ***
date2010-03-12                2.261e+05  2.546e+04   8.884  < 2e-16 ***
date2010-03-19                1.072e+05  2.424e+04   4.422 9.96e-06 ***
date2010-03-26                9.097e+04  2.444e+04   3.722 0.000200 ***
date2010-04-06                1.717e+05  2.291e+04   7.495 7.56e-14 ***
date2010-04-16                9.173e+04  2.353e+04   3.898 9.78e-05 ***
date2010-04-23                8.273e+04  2.359e+04   3.507 0.000456 ***
date2010-04-30                5.946e+04  2.354e+04   2.526 0.011574 *
date2010-05-02                2.476e+05  2.671e+04   9.269  < 2e-16 ***
date2010-05-03                1.638e+05  2.512e+04   6.521 7.52e-11 ***
date2010-05-11                1.143e+05  2.391e+04   4.780 1.80e-06 ***
date2010-05-14                9.097e+04  2.346e+04   3.878 0.000106 ***
date2010-05-21                7.495e+04  2.313e+04   3.240 0.001200 **
date2010-05-28                1.244e+05  2.295e+04   5.420 6.17e-08 ***
date2010-06-08                1.118e+05  2.316e+04   4.827 1.42e-06 ***
date2010-06-18                1.143e+05  2.291e+04   4.988 6.27e-07 ***
date2010-06-25                8.158e+04  2.308e+04   3.534 0.000412 ***
date2010-07-05                1.543e+05  2.324e+04   6.640 3.39e-11 ***
date2010-07-16                6.651e+04  2.319e+04   2.868 0.004146 **
date2010-07-23                3.553e+04  2.324e+04   1.529 0.126267
date2010-07-30                3.301e+04  2.312e+04   1.428 0.153400
date2010-08-10                8.001e+04  2.309e+04   3.464 0.000535 ***
date2010-08-13                7.461e+04  2.312e+04   3.226 0.001260 **
date2010-08-20                8.724e+04  2.313e+04   3.772 0.000164 ***
date2010-08-27                6.745e+04  2.301e+04   2.931 0.003388 **
date2010-09-04                1.387e+05  2.344e+04   5.917 3.45e-09 ***
date2010-09-07                1.065e+05  2.309e+04   4.613 4.05e-06 ***
date2010-09-17                1.511e+04  2.288e+04   0.660 0.508968
date2010-09-24               -2.377e+04  2.288e+04  -1.039 0.298793
date2010-10-09                1.224e+05  3.063e+04   3.998 6.47e-05 ***
date2010-10-12                3.569e+05  2.597e+04  13.741  < 2e-16 ***
date2010-10-15                3.973e+04  2.323e+04   1.710 0.087244 .
date2010-10-22                4.114e+04  2.335e+04   1.762 0.078172 .
date2010-10-29                5.318e+04  2.337e+04   2.275 0.022933 *
date2010-11-06                1.176e+05  2.293e+04   5.127 3.03e-07 ***
date2010-11-19                1.071e+05  2.459e+04   4.355 1.35e-05 ***
date2010-11-26                6.058e+05  2.863e+04  21.158  < 2e-16 ***
date2010-12-02                2.416e+05  2.870e+04   8.419  < 2e-16 ***
date2010-12-03                1.335e+05  2.450e+04   5.448 5.30e-08 ***
date2010-12-11                1.261e+05  2.422e+04   5.205 2.00e-07 ***
date2010-12-17                4.913e+05  2.585e+04  19.010  < 2e-16 ***
```



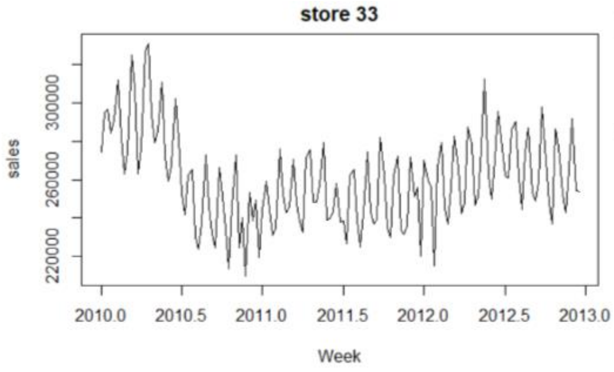
```
date2010-12-24                 9.147e+05  2.576e+04  35.512  < 2e-16 ***
date2010-12-31                 4.937e+04  2.899e+04   1.703 0.088598 .
date2011-01-04                 4.888e+04  2.741e+04   1.783 0.074628 .
date2011-01-07                 1.064e+05  2.720e+04   3.913 9.22e-05 ***
date2011-01-14                 2.863e+04  2.773e+04   1.033 0.301873
date2011-01-21                 1.622e+04  2.659e+04   0.610 0.541870
date2011-01-28                -1.833e+03  2.716e+04  -0.067 0.946194
date2011-02-09                 5.640e+04  2.706e+04   2.084 0.037204 *
date2011-02-12                 1.706e+05  2.674e+04   6.381 1.89e-10 ***
date2011-02-18                 1.813e+05  2.560e+04   7.081 1.59e-12 ***
date2011-02-25                 8.360e+04  2.602e+04   3.213 0.001318 **
date2011-03-06                 1.435e+05  2.846e+04   5.043 4.71e-07 ***
date2011-03-18                 8.191e+04  2.691e+04   3.044 0.002344 **
date2011-03-25                 3.953e+04  2.696e+04   1.466 0.142580
date2011-04-02                 1.497e+05  2.783e+04   5.378 7.83e-08 ***
date2011-04-03                 1.419e+05  2.664e+04   5.325 1.04e-07 ***
date2011-04-11                 1.484e+05  2.707e+04   5.482 4.36e-08 ***
date2011-04-15                 7.322e+04  2.822e+04   2.594 0.009497 **
date2011-04-22                 1.543e+05  2.863e+04   5.388 7.38e-08 ***
date2011-04-29                 3.764e+04  2.897e+04   1.299 0.193898
date2011-05-08                 1.153e+05  2.815e+04   4.095 4.27e-05 ***
date2011-05-13                 7.805e+04  2.974e+04   2.624 0.008702 **
date2011-05-20                 4.699e+04  2.966e+04   1.584 0.113254
date2011-05-27                 7.013e+04  2.897e+04   2.421 0.015511 *
date2011-06-05                 1.131e+05  2.971e+04   3.808 0.000142 ***
date2011-06-17                 1.122e+05  2.788e+04   4.025 5.77e-05 ***
date2011-06-24                 7.483e+04  2.756e+04   2.715 0.006655 **
date2011-07-01                 6.878e+04  2.689e+04   2.558 0.010565 *
date2011-07-10                 1.002e+05  2.608e+04   3.843 0.000123 ***
date2011-07-15                 5.869e+04  2.728e+04   2.151 0.031492 *
date2011-07-22                 5.202e+04  2.788e+04   1.866 0.062075 .
date2011-07-29                 1.848e+04  2.801e+04   0.660 0.509421
date2011-08-04                 9.569e+04  2.754e+04   3.474 0.000516 ***
date2011-08-07                 1.098e+05  2.684e+04   4.090 4.37e-05 ***
date2011-08-19                 9.057e+04  2.717e+04   3.333 0.000864 ***
date2011-08-26                 9.971e+04  2.695e+04   3.700 0.000218 ***
date2011-09-09                 1.488e+05  3.492e+04   4.261 2.07e-05 ***
date2011-09-12                 3.186e+05  2.757e+04  11.554  < 2e-16 ***
date2011-09-16                 2.945e+04  2.709e+04   1.087 0.276980
date2011-09-23                 7.406e+03  2.674e+04   0.277 0.781827
```



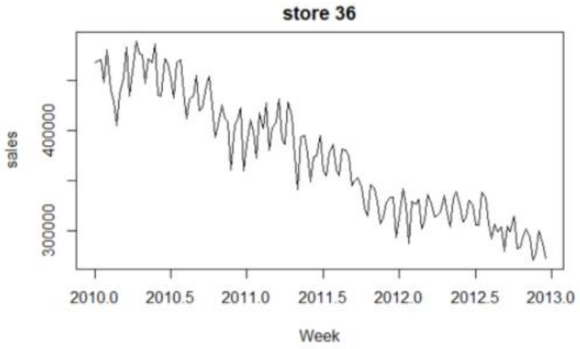
```
date2011-09-30              -8.941e+03  2.632e+04  -0.340 0.734070    
date2011-10-06               1.174e+05  2.816e+04   4.170 3.09e-05 ***
date2011-10-14               3.668e+04  2.591e+04   1.416 0.156878    
date2011-10-21               7.139e+04  2.640e+04   2.704 0.006873 ** 
date2011-10-28               7.768e+04  2.660e+04   2.920 0.003508 ** 
date2011-11-02               1.960e+05  2.946e+04   6.651 3.16e-11 ***
date2011-11-03               8.285e+04  2.718e+04   3.048 0.002315 ** 
date2011-11-11               1.446e+05  2.686e+04   5.384 7.55e-08 ***
date2011-11-18               9.572e+04  2.660e+04   3.599 0.000322 ***
date2011-11-25               5.865e+05  3.081e+04  19.037  < 2e-16 ***
date2011-12-08               7.577e+04  2.776e+04   2.729 0.006364 ** 
date2011-12-16               4.159e+05  2.703e+04  15.389  < 2e-16 ***
date2011-12-23               7.954e+05  2.706e+04  29.394  < 2e-16 ***
date2011-12-30               1.390e+05  2.989e+04   4.649 3.40e-06 ***
date2012-01-06               9.714e+04  2.810e+04   3.458 0.000549 ***
date2012-01-13               4.368e+03  2.758e+04   0.158 0.874150    
date2012-01-20               5.076e+03  2.759e+04   0.184 0.854037    
date2012-01-27              -4.936e+04  2.734e+04  -1.805 0.071075 .  
date2012-02-03               9.863e+04  2.878e+04   3.427 0.000614 ***
date2012-02-17               1.793e+05  2.855e+04   6.281 3.60e-10 ***
date2012-02-24               7.566e+04  2.834e+04   2.670 0.007601 ** 
date2012-03-02               8.594e+04  2.735e+04   3.142 0.001685 ** 
date2012-03-08               5.744e+04  2.690e+04   2.135 0.032778 *  
date2012-03-16               8.750e+04  2.902e+04   3.015 0.002579 ** 
date2012-03-23               4.106e+04  2.925e+04   1.404 0.160408    
date2012-03-30               4.779e+04  2.975e+04   1.606 0.108316    
date2012-04-05               7.874e+04  2.939e+04   2.679 0.007406 ** 
date2012-04-13               7.208e+04  3.030e+04   2.379 0.017385 *  
date2012-04-20               3.462e+04  3.016e+04   1.148 0.251128    
date2012-04-27               3.172e+03  2.981e+04   0.106 0.915268    
date2012-05-10               6.556e+04  2.918e+04   2.247 0.024683 *  
date2012-05-18               6.943e+04  2.882e+04   2.409 0.016032 *  
date2012-05-25               8.956e+04  2.843e+04   3.150 0.001640 ** 
date2012-06-01               6.868e+04  2.700e+04   2.544 0.010993 *  
date2012-06-04               2.264e+05  3.007e+04   7.528 5.90e-14 ***
date2012-06-07               1.404e+05  2.610e+04   5.380 7.74e-08 ***
date2012-06-15               9.664e+04  2.726e+04   3.545 0.000395 ***
date2012-06-22               7.537e+04  2.691e+04   2.801 0.005107 ** 
date2012-06-29               4.692e+04  2.646e+04   1.773 0.076232 .  
date2012-07-09               1.466e+05  3.800e+04   3.858 0.000115 ***
date2012-07-13               2.748e+04  2.621e+04   1.049 0.294435    
date2012-07-20               2.679e+04  2.636e+04   1.016 0.309597    
```

```
date2012-07-27              -1.646e+04  2.698e+04  -0.610 0.541974
date2012-08-06               1.276e+05  2.757e+04   4.629 3.75e-06 ***
date2012-08-17               5.964e+04  2.821e+04   2.115 0.034511 *
date2012-08-24               6.697e+04  2.838e+04   2.360 0.018307 *
date2012-08-31               5.755e+04  2.859e+04   2.013 0.044154 *
date2012-09-03               1.082e+05  2.896e+04   3.738 0.000187 ***
date2012-09-14               3.004e+02  2.913e+04   0.010 0.991772
date2012-09-21               5.339e+03  2.958e+04   0.181 0.856743
date2012-09-28              -9.996e+03  2.907e+04  -0.344 0.730983
date2012-10-02               2.090e+05  3.108e+04   6.726 1.90e-11 ***
date2012-10-08               5.599e+04  2.752e+04   2.034 0.041981 *
date2012-10-19               1.797e+04  2.948e+04   0.610 0.542172
date2012-10-26               2.726e+04  2.875e+04   0.948 0.342988
date2012-11-05               7.440e+04  2.910e+04   2.557 0.010578 *
date2012-12-10               4.228e+04  2.971e+04   1.423 0.154817
holiday1                            NA         NA      NA       NA
temperature                  8.486e+03  1.273e+03   6.665 2.88e-11 ***
unemployment                -2.028e+04  7.384e+03  -2.746 0.006051 **
temperature:fuelprice       -1.181e+03  2.815e+02  -4.195 2.77e-05 ***
fuelprice:CPI                5.593e+02  8.387e+01   6.669 2.79e-11 ***
temperature:unemployment    -1.595e+02  4.710e+01  -3.387 0.000712 ***
holiday0:temperature:CPI    -1.094e+01  2.543e+00  -4.303 1.71e-05 ***
holiday1:temperature:CPI    -1.536e+01  3.158e+00  -4.862 1.19e-06 ***
unemployment:fuelprice:CPI -1.938e+01  1.114e+01  -1.740 0.081894 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 108500 on 6240 degrees of freedom
Multiple R-squared:  0.9642,    Adjusted R-squared:  0.963
F-statistic: 865.3 on 194 and 6240 DF,  p-value: < 2.2e-16
```

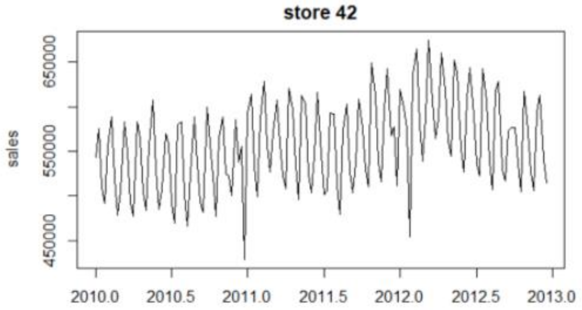
We can see that now all of the dates and stores are necessary for the model. This also can be shown from the special time series plot, which has much difference than the other Stores.



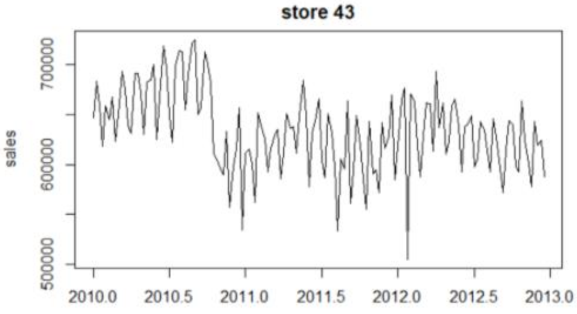
store 33



store 36

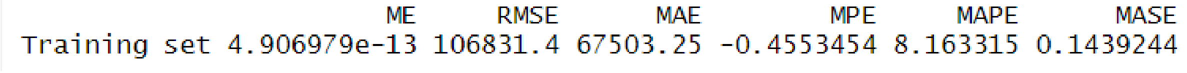


store 42



store 43

To make a comparison with other models, We would like to find the accuracy of Linear Regression Model:



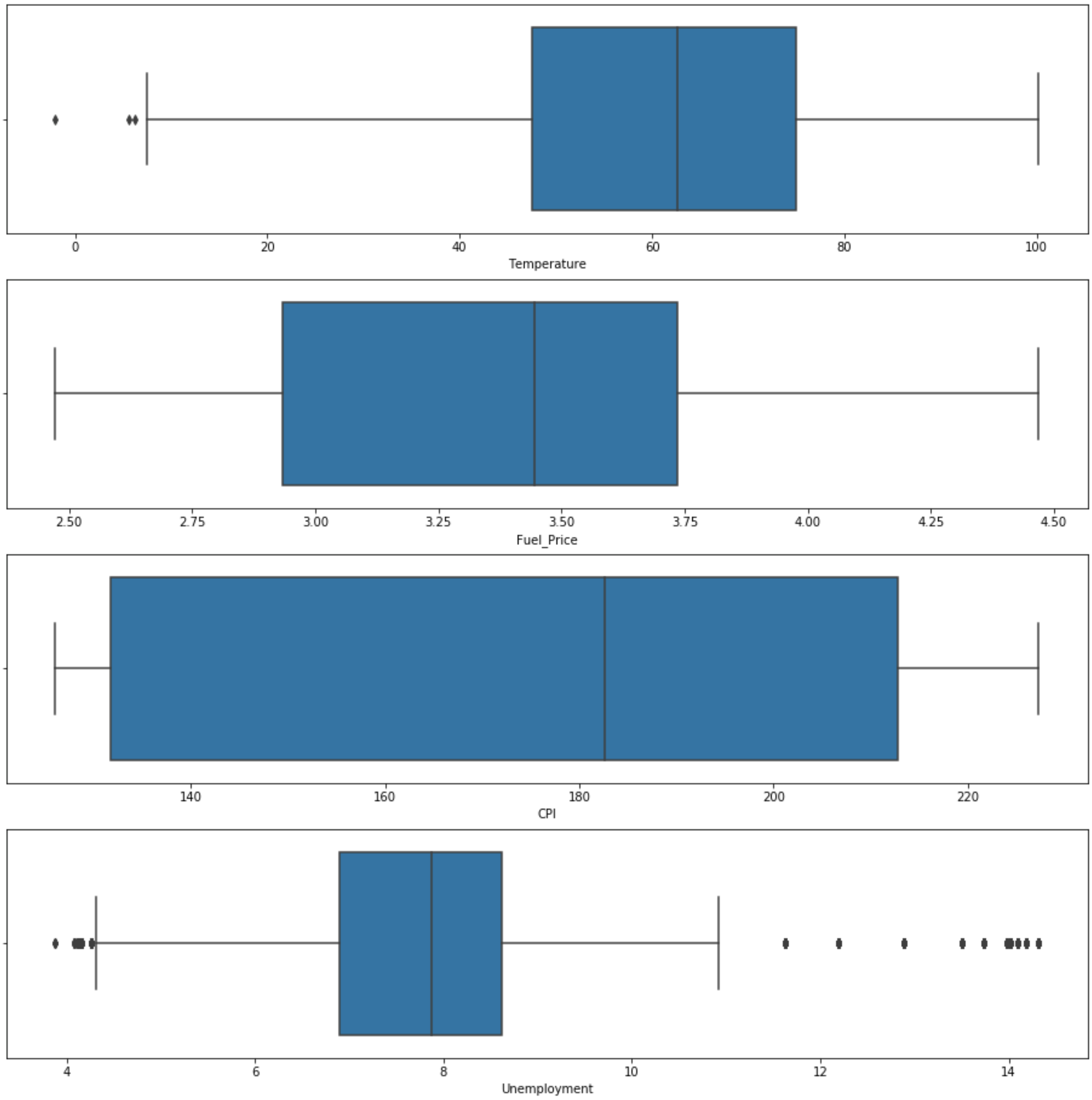
```
                       ME     RMSE      MAE        MPE     MAPE      MASE
Training set 4.906979e-13 106831.4 67503.25 -0.4553454 8.163315 0.1439244
```

Also we have the MSE of 7.344*10^13 from the ANOVA table.

## 2. Tree

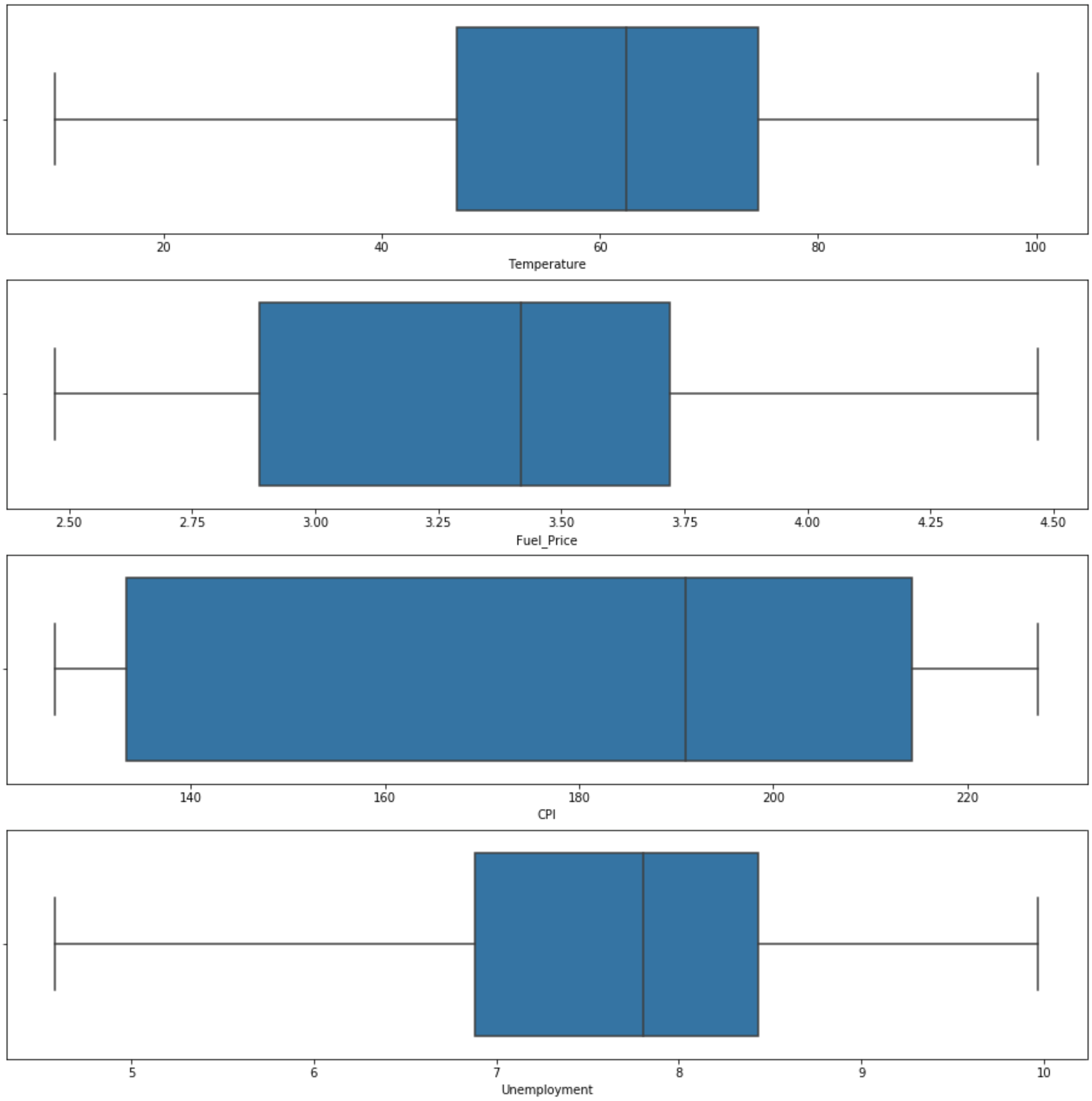
### 2.1. Outliers

Here, we are trying to fit a Decision tree regressor Model. We first use the box plot to detect whether there are any outliers for the numerical features.



Based on the plot, we can clearly see that “Temperature” and “Unemployment” have some outliers. Therefore, we drop the points where “Temperature” is smaller than 10 and “Unemployment” smaller than

4.5 and higher than 10. After dropping the outliers, we plot the boxplot again.



We can see there are no more outliers, we now obtained a cleaner data set for feature work.

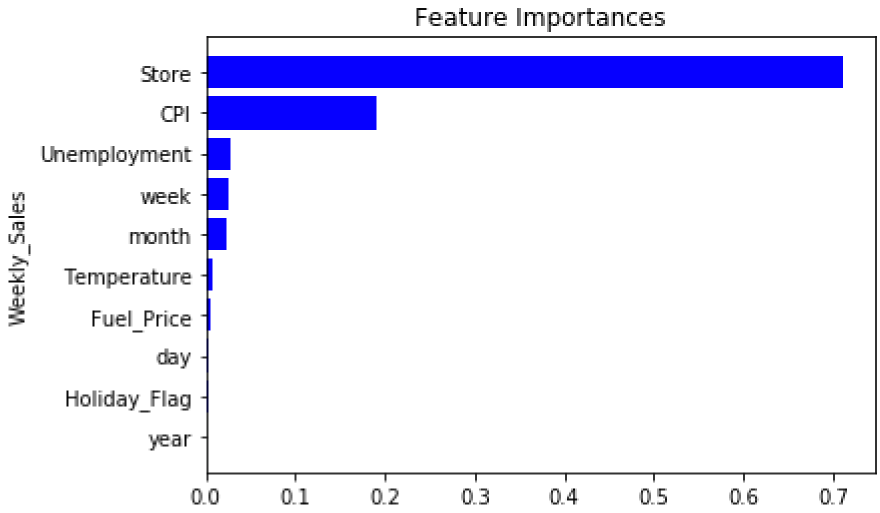
## 2.2. Feature Selection

First, we build a decision tree regressor with all the features, which are ’Store’, ’Holiday_Flag’, ’Temperature’, ’Fuel_Price’, ’CPI’, ’Unemployment’, ’week’, ’day’, ’year’, ’month’. And put ’Weekly_Sales’ as our prediction target. The result is shown below.

```
Decision Tree Regressor with X:
Mean Absolute Error: 81579.22962014136
Mean Squared Error: 23707178610.1199
```

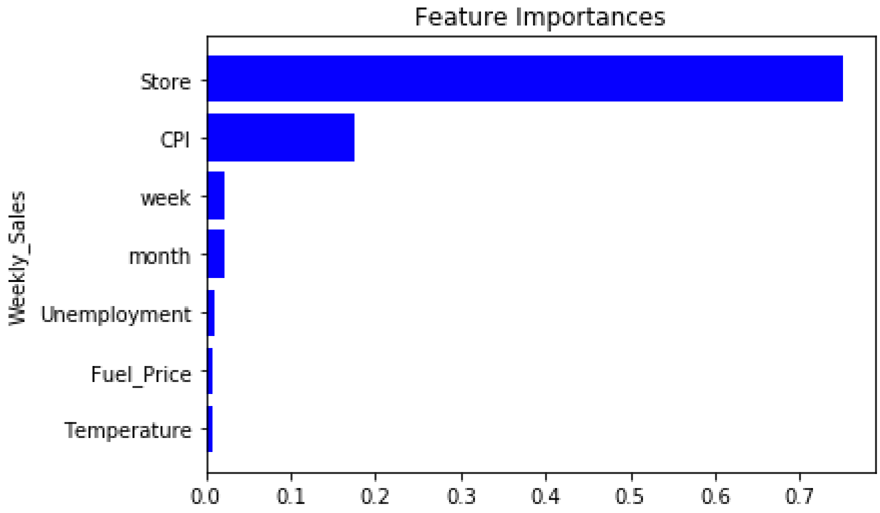
```
Root Mean Squared Error: 153971.35645995947
```

Then we graph the plot of feature importance of this model.



we drop the least three important features “year”, “Holiday_Flag”, and “day”, and use the rest of the features to build a tree again. Then the result is shown below.

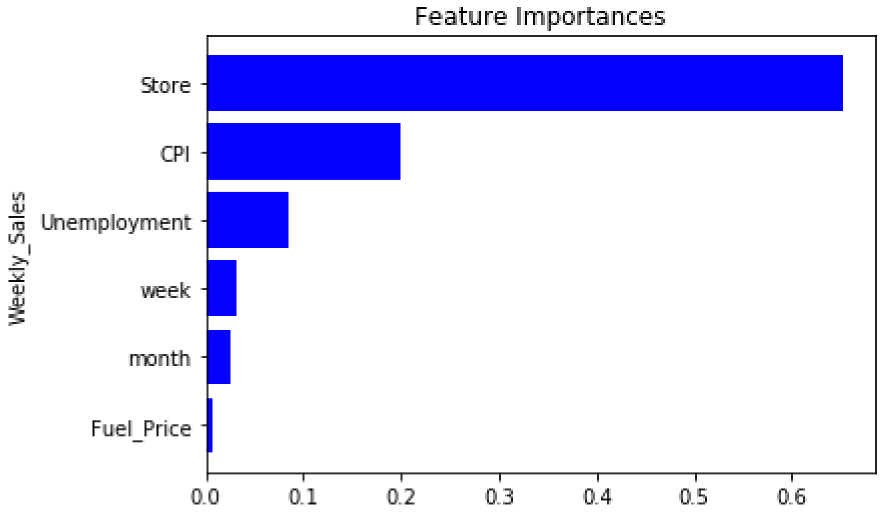
```
Decision Tree Regressor with X:
Mean Absolute Error: 76584.22007067138
Mean Squared Error: 21548856565.854446
Root Mean Squared Error: 146795.28795521485
```



We can see MAE, MSE, and RMSE all decreased. Hence, the accuracy of our model has improved. For the next step, we continue to drop the least important feature “Temperature”. And the result is shown below.

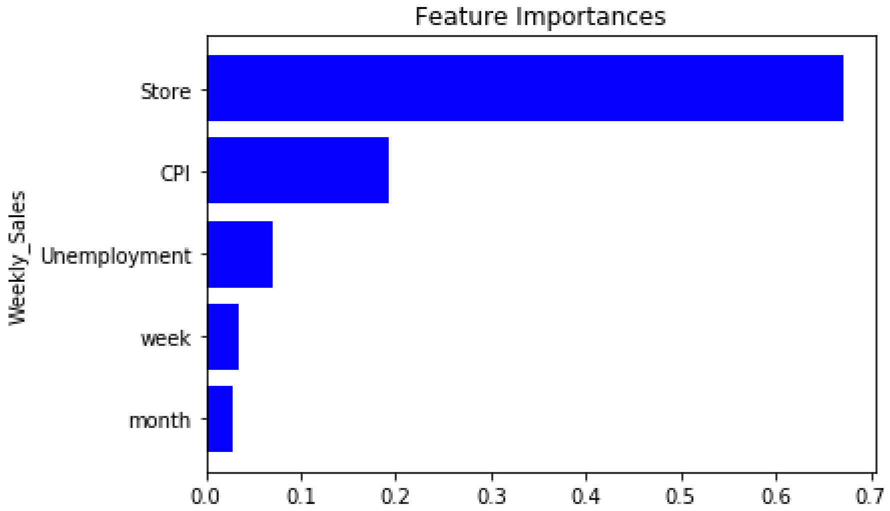
```
Decision Tree Regressor with X:
```

```
Mean Absolute Error: 76350.61183745583
Mean Squared Error: 19375118624.50201
Root Mean Squared Error: 139194.53518188855
```



We can see MAE, MSE, and RMSE are keep decreasing on the test set. Hence, the accuracy of our model keeps improving. For the next step, we continue to drop the least important feature “Fuel_price”. And the result is shown below.

```
Decision Tree Regressor with X:
Mean Absolute Error: 72027.80410777385
Mean Squared Error: 15711243645.870707
Root Mean Squared Error: 125344.49986286079
```



We can see our model get a better score than last time. And all the features are fairly important to our model based on the feature importances plot. Therefore, we will use “Store”, “CPI”, “Unemploymen t”, “week”, and “month” as our features to build the decision tree regressor.

#### 2.3. Result

By using the features we previously decided, we now build a decision tree regressor. We ran the model 50 times, and printed the mean score for MAE, MSE, and RMSE. The final result is shown below.

```
Decision Tree Regressor:
Mean Absolute Error: 72870.70425971733
Mean Squared Error: 16390844791.99989
Root Mean Squared Error: 127947.106865705
```

### 3. Randomforest

By using all the features, we build a random forest regressor to predict weelk_sales. The result is shown below.

```
Random Forest Regressor with X:
Mean Absolute Error: 65997.55455653711
Mean Squared Error: 15189204958.608917
Root Mean Squared Error: 123244.49260964531
```

## 5. Result and Evaluation

| | Linear Regression | Decision Tree Regressor | Random Forest Regressor |
|---|---|---|---|
| Mean Absolute Error | 6.7503e+4 | 7.2870e+4 | 6.5997e+4 |
| Mean Squared Error | 1.176e+10 | 1.6390e+10 | 1.51892e+10 |
| Root Mean Squared Error | 1.06831e+5 | 1.2794e+5 | 1.23244e+5 |

Based on the table, random forest regressors have the best MAE. However, linear regression models have the best MSE and RMSE among all three different models.