

FINAL PROJECT

T

Multiple Linear Regression & Machine Learning-Model
Building

Komalla

Cemerla

Kondreddy

Kothwal

Varsha

Neha

Sujith

Praneeth Reddy



Walmart

Introduction



Aim of the project is to increase the sales of Walmart

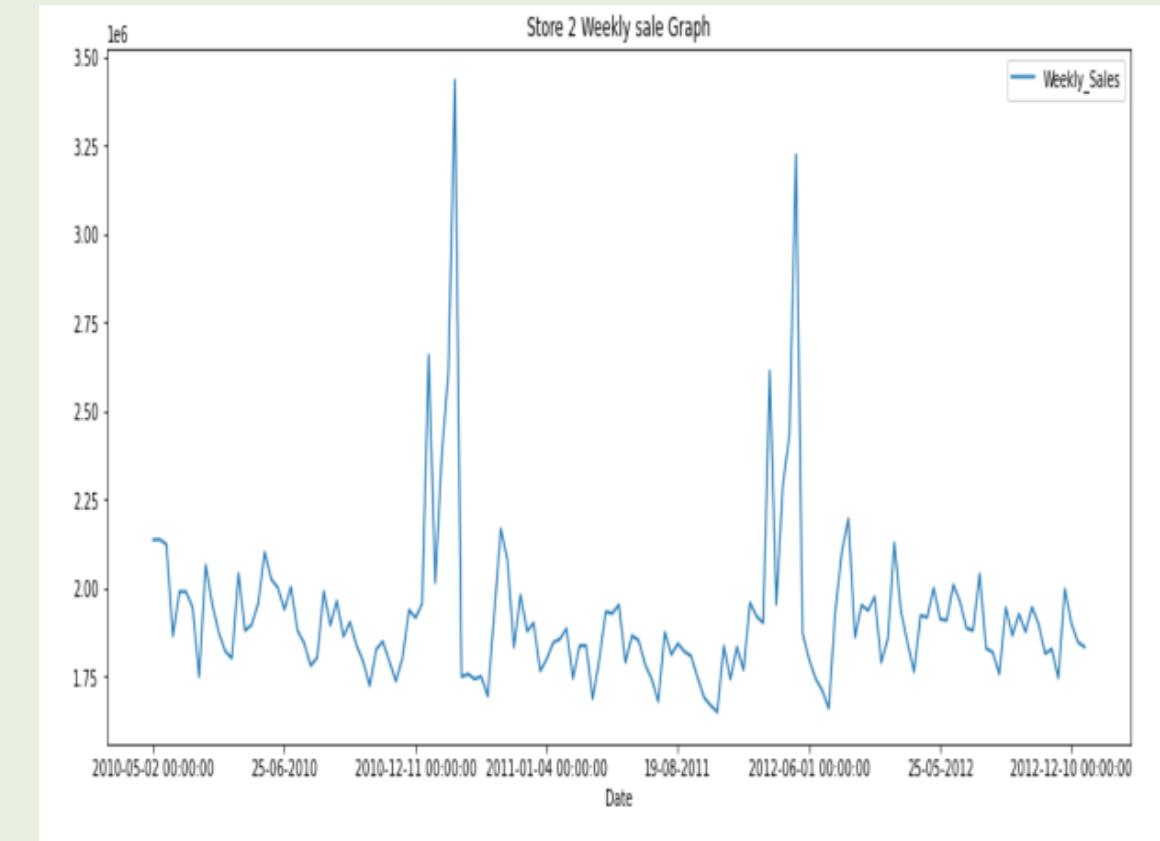
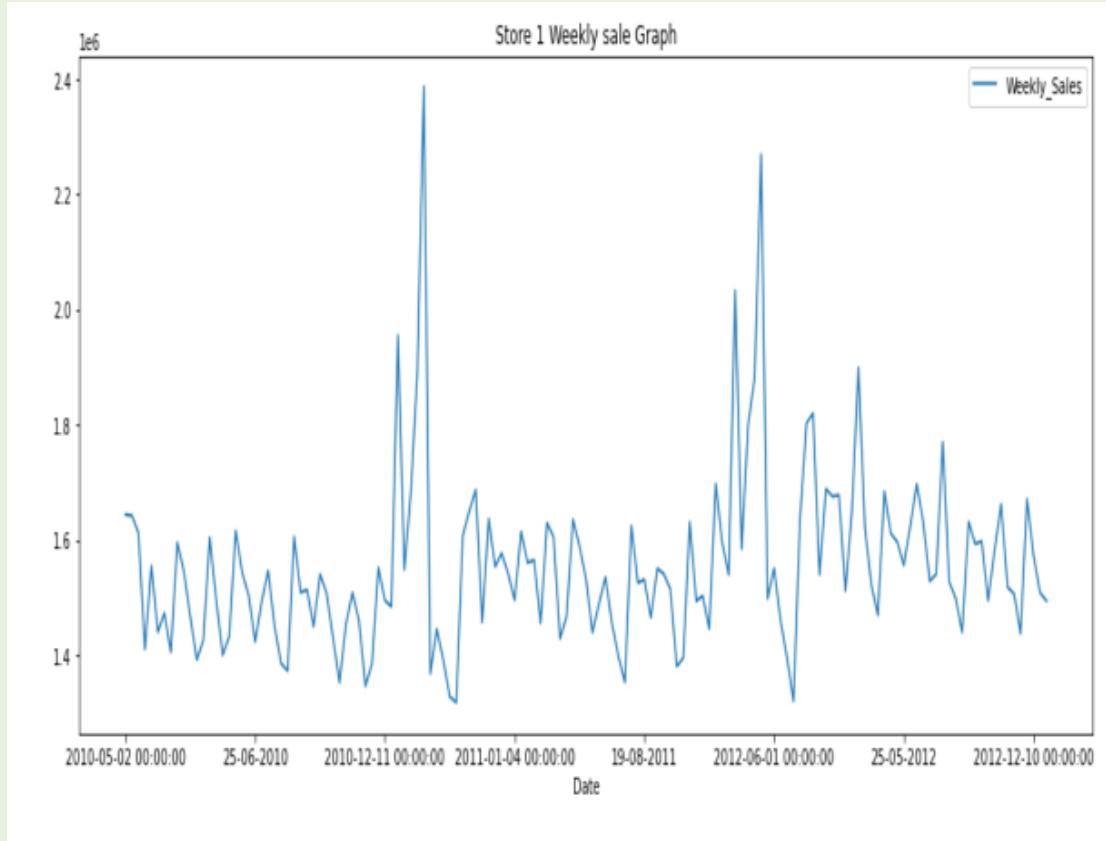


To analyze what factors effect the sales

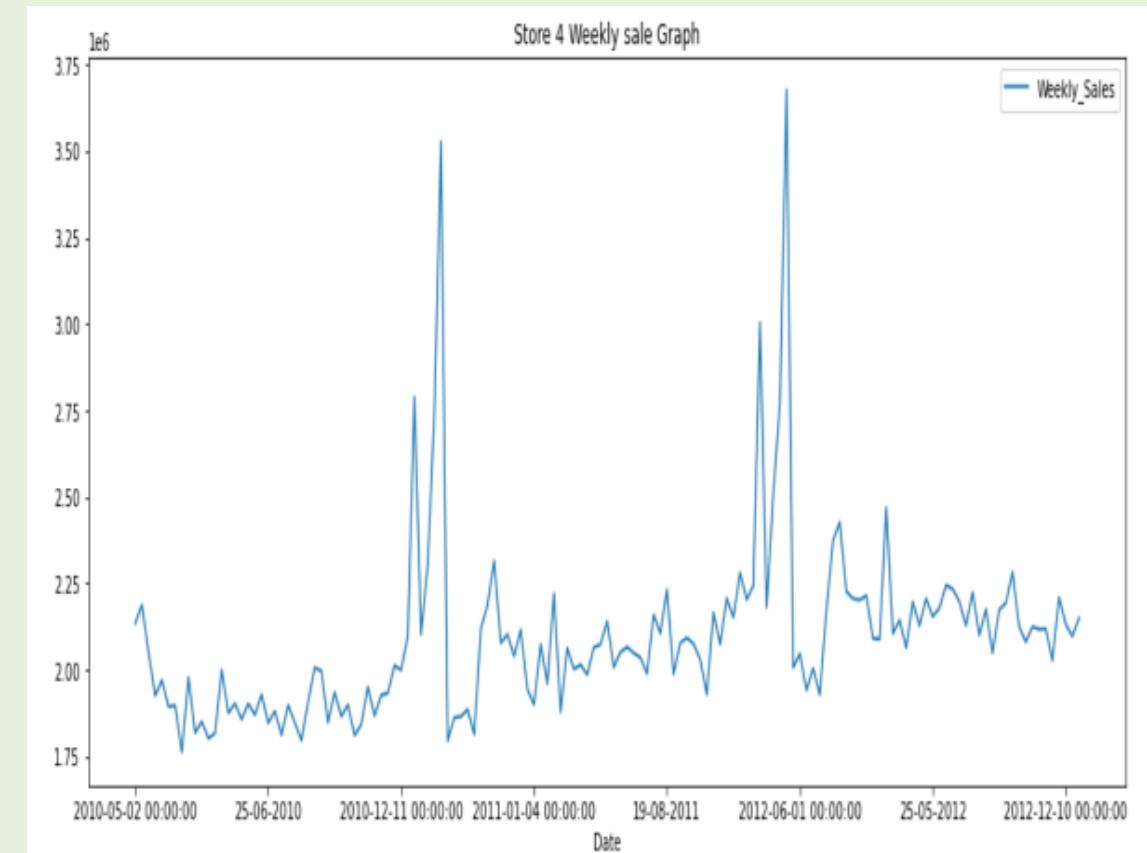
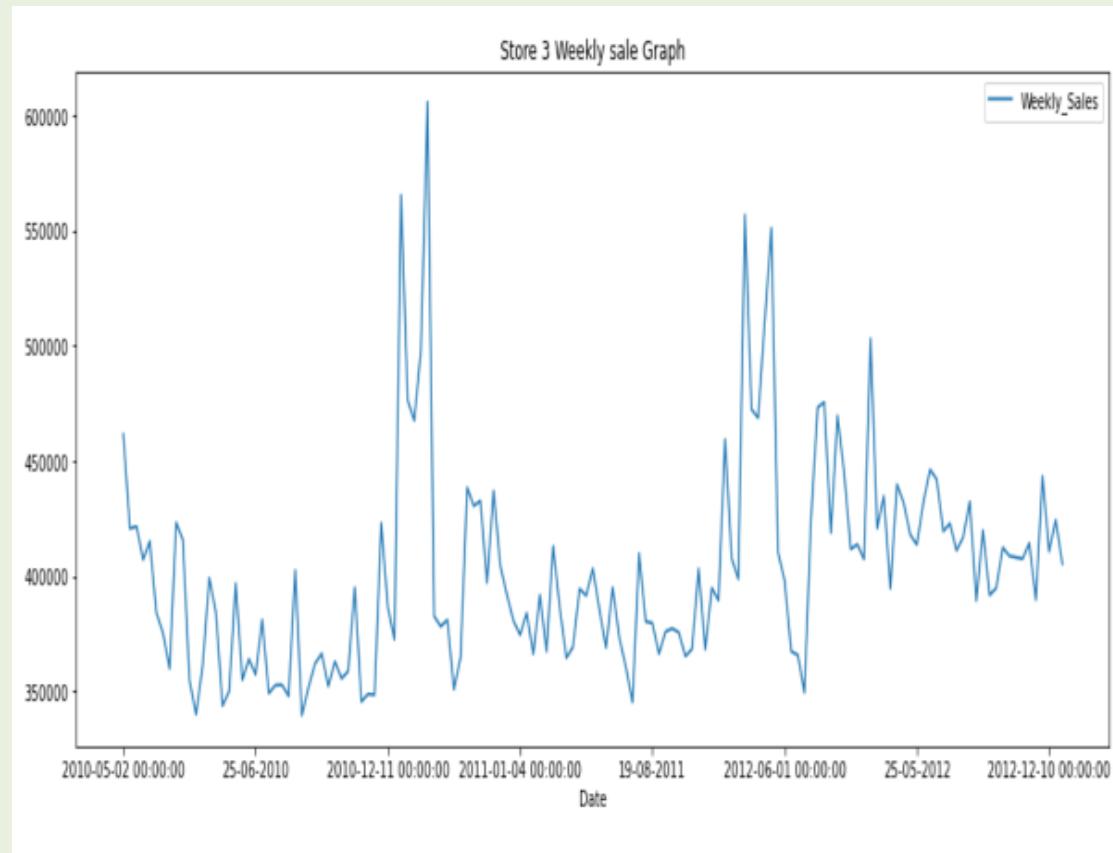
Process of Model Building

- Data Cleaning
- Created plots in excel to represent EDA in the form of graphs
- Checked for the linear regression assumptions in the dataset. Transformed the data if there were any violations.
- Represented the data in the form of plots to describe the relation between dependent and independent variables.

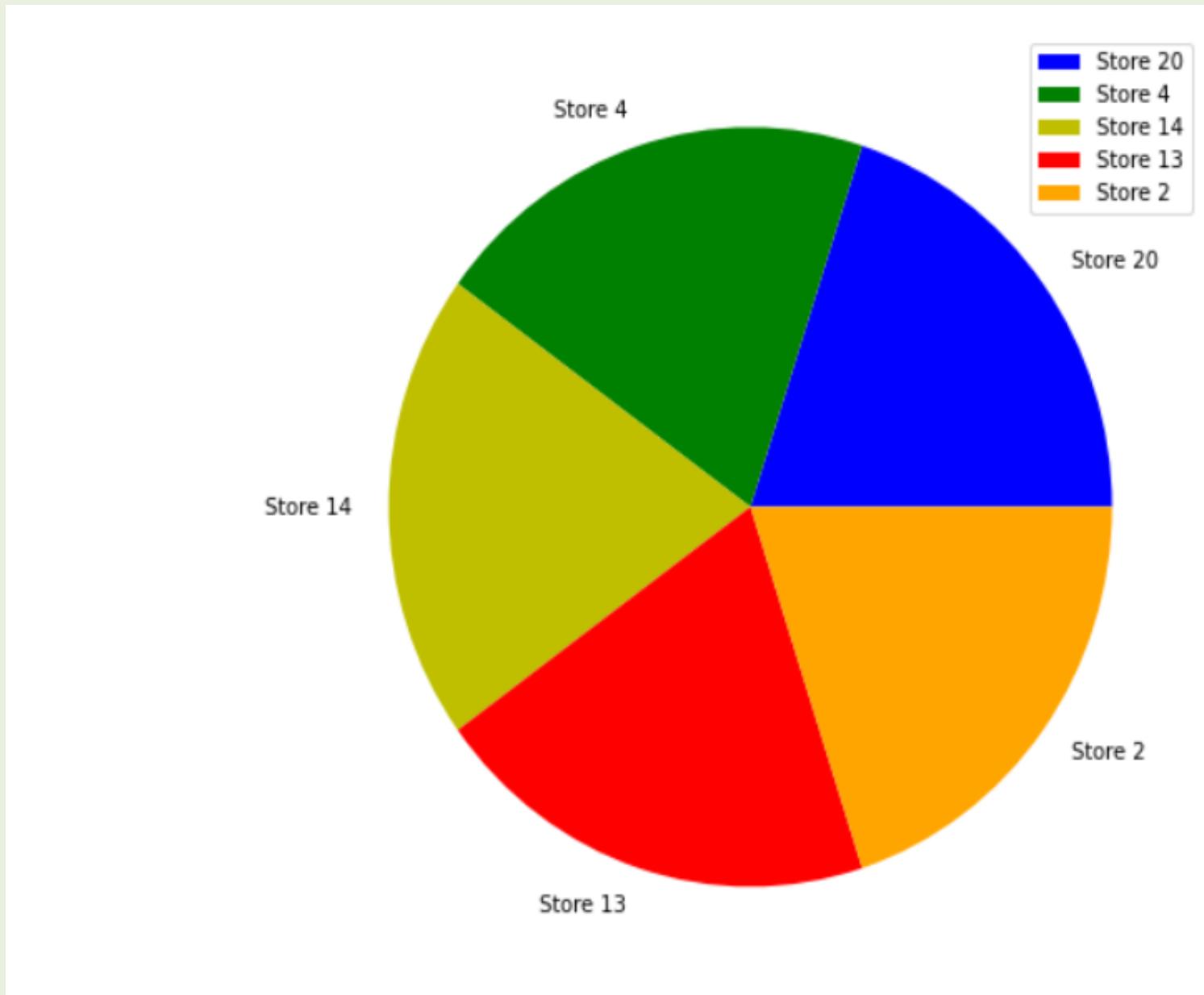
WEEKLY GRAPHS OF STORE 1 AND STORE 2



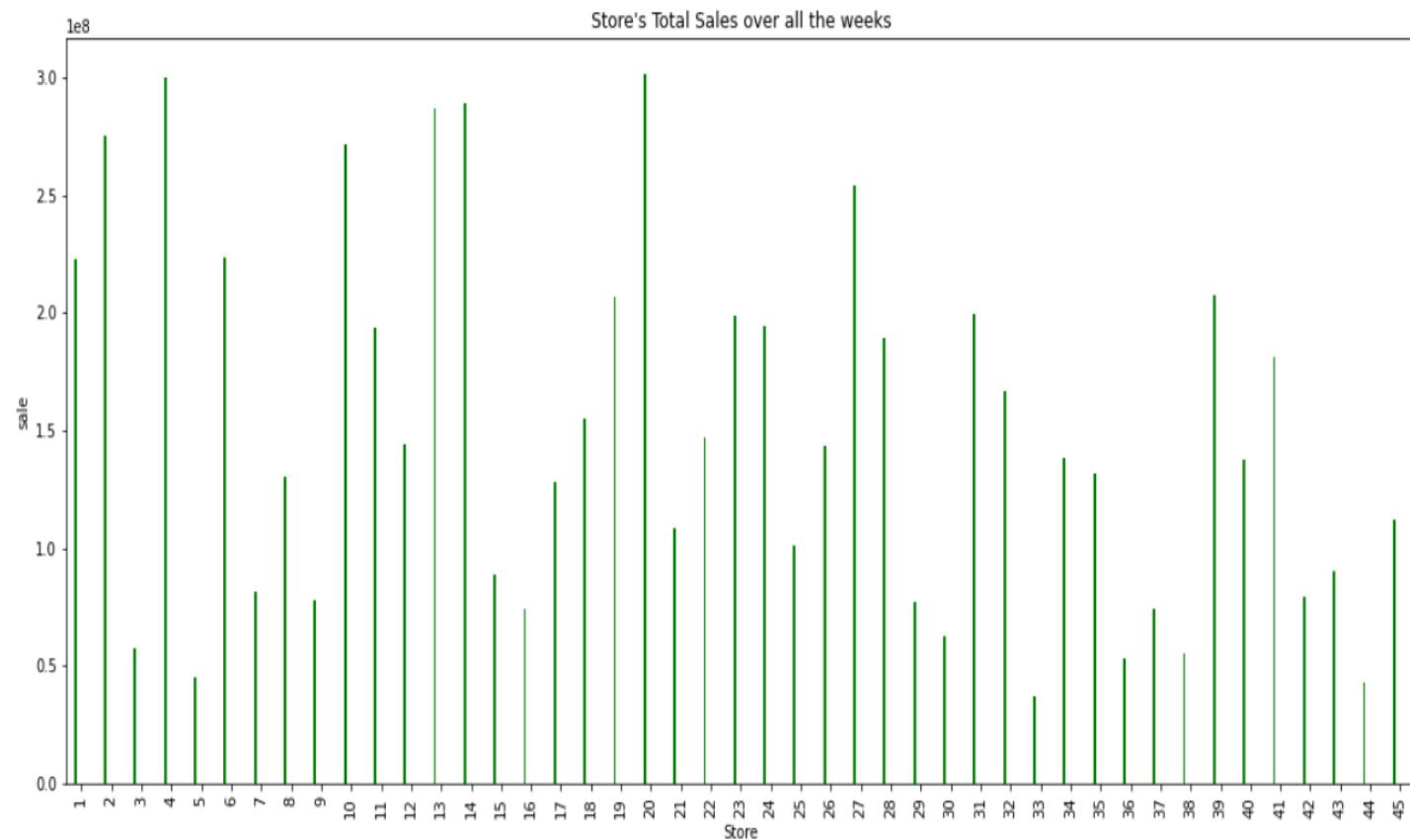
WEEKLY GRAPHS OF STORE 3 AND STORE 4



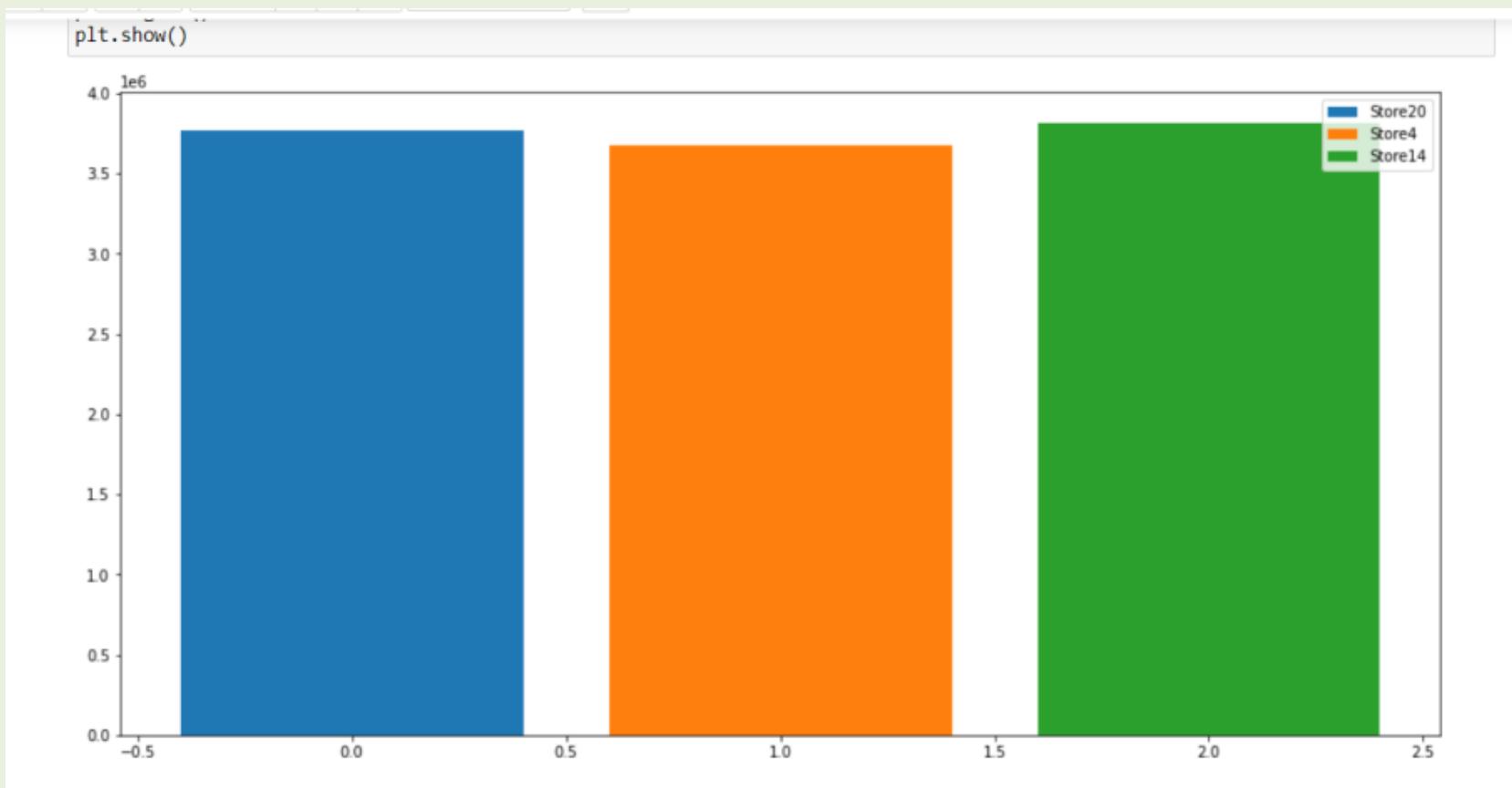
Revenue of various Walmart stores

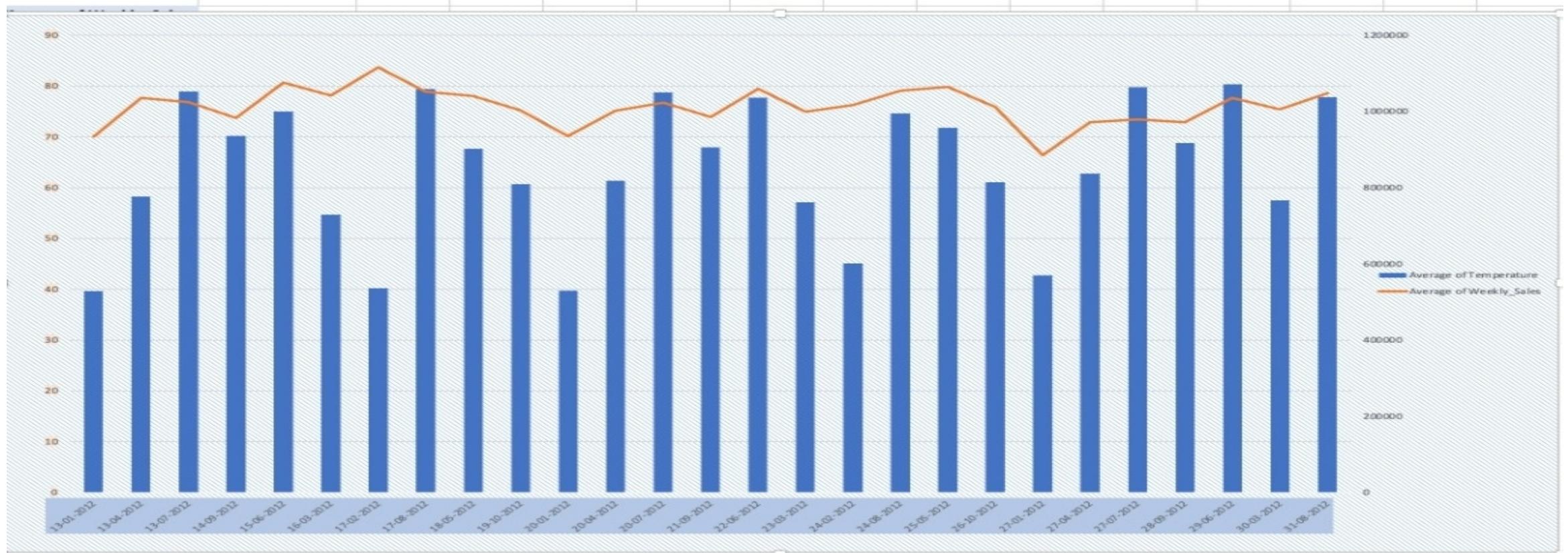


Total Store data in all weeks



Purchase stock in various stores

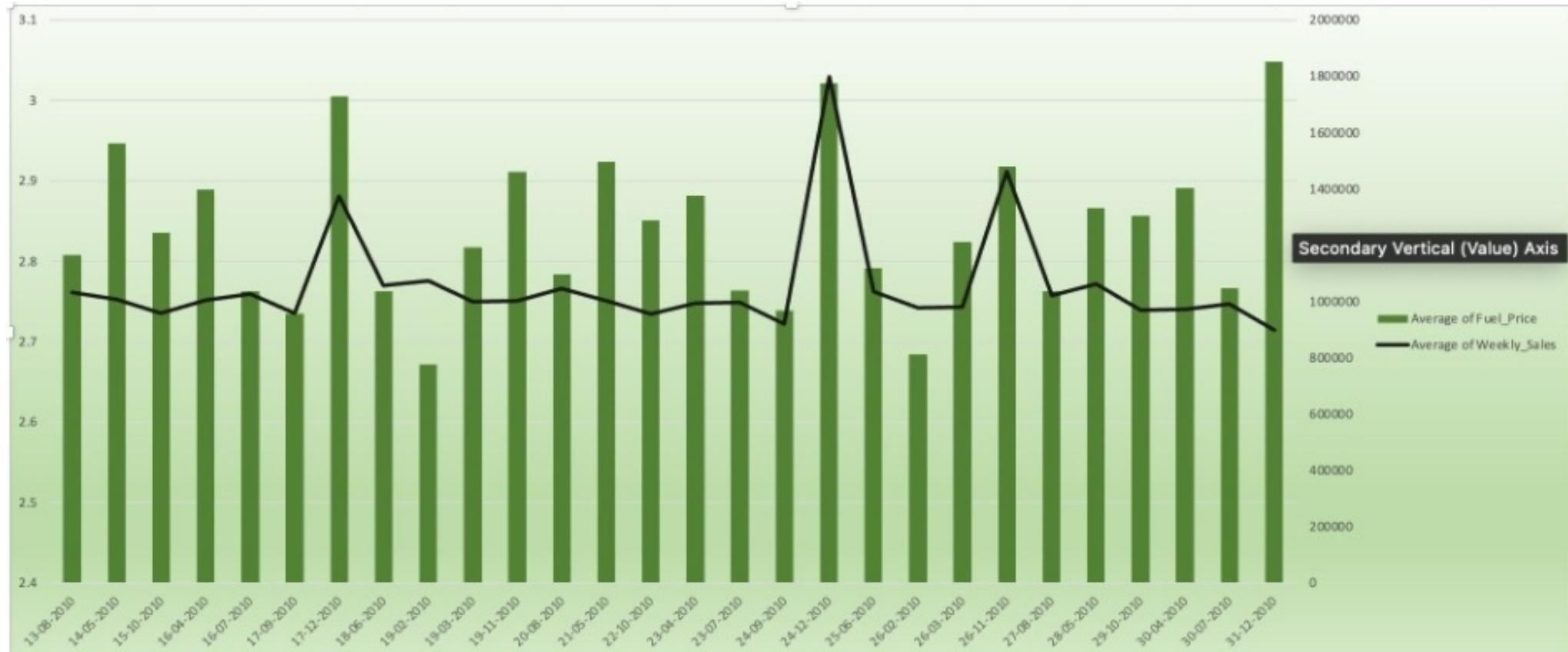




SALES VS TEMPERATURE

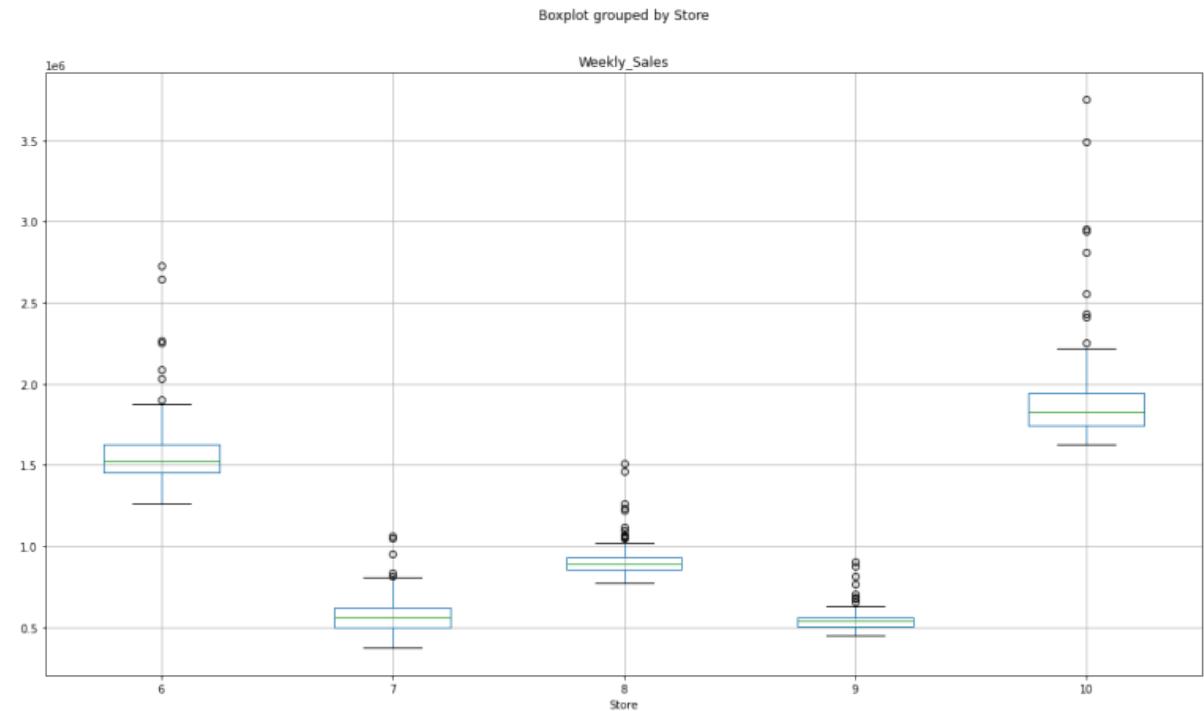
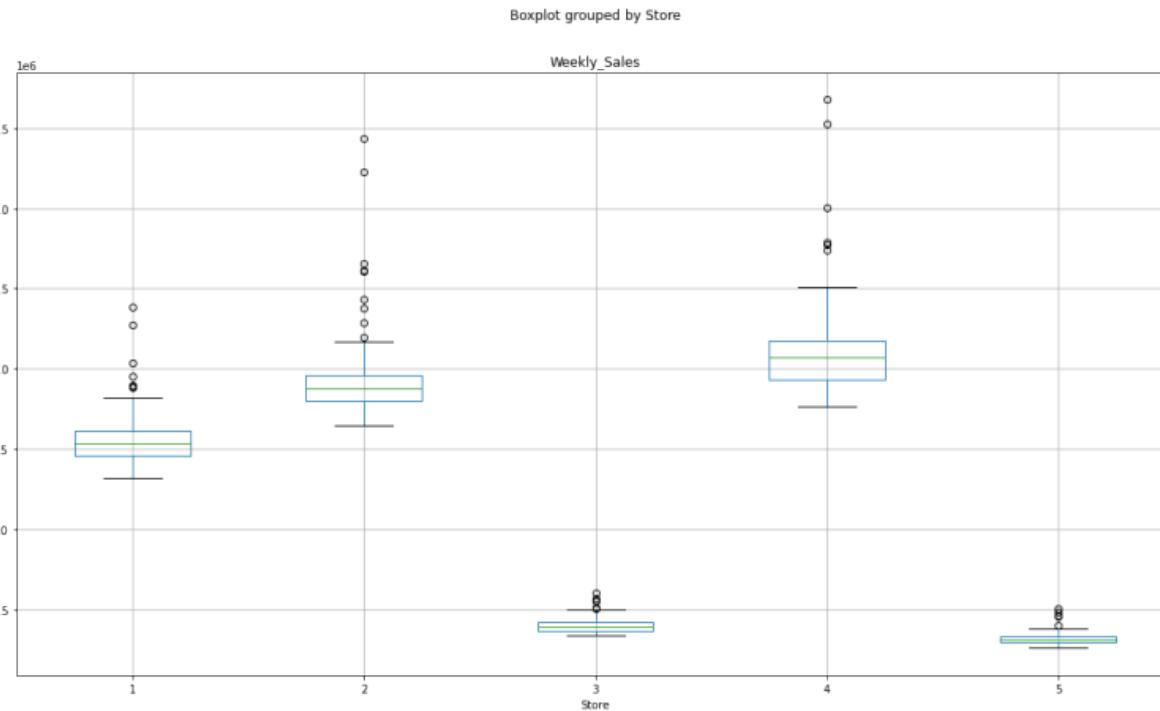
UNEMPLOYMENT RATE





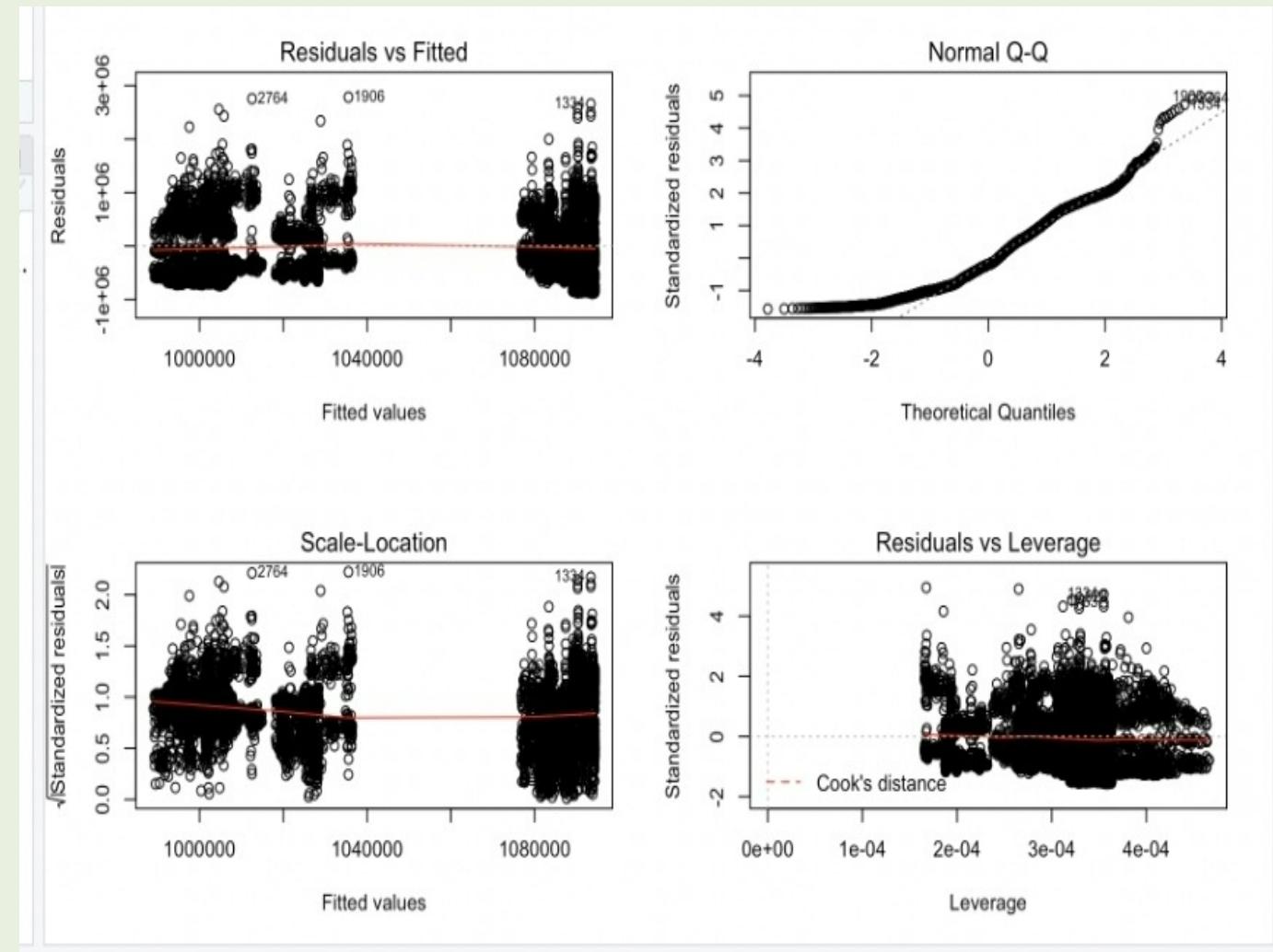
SALES VS FUEL PRICE

WEEKLY SALES GROUPED BY STORE

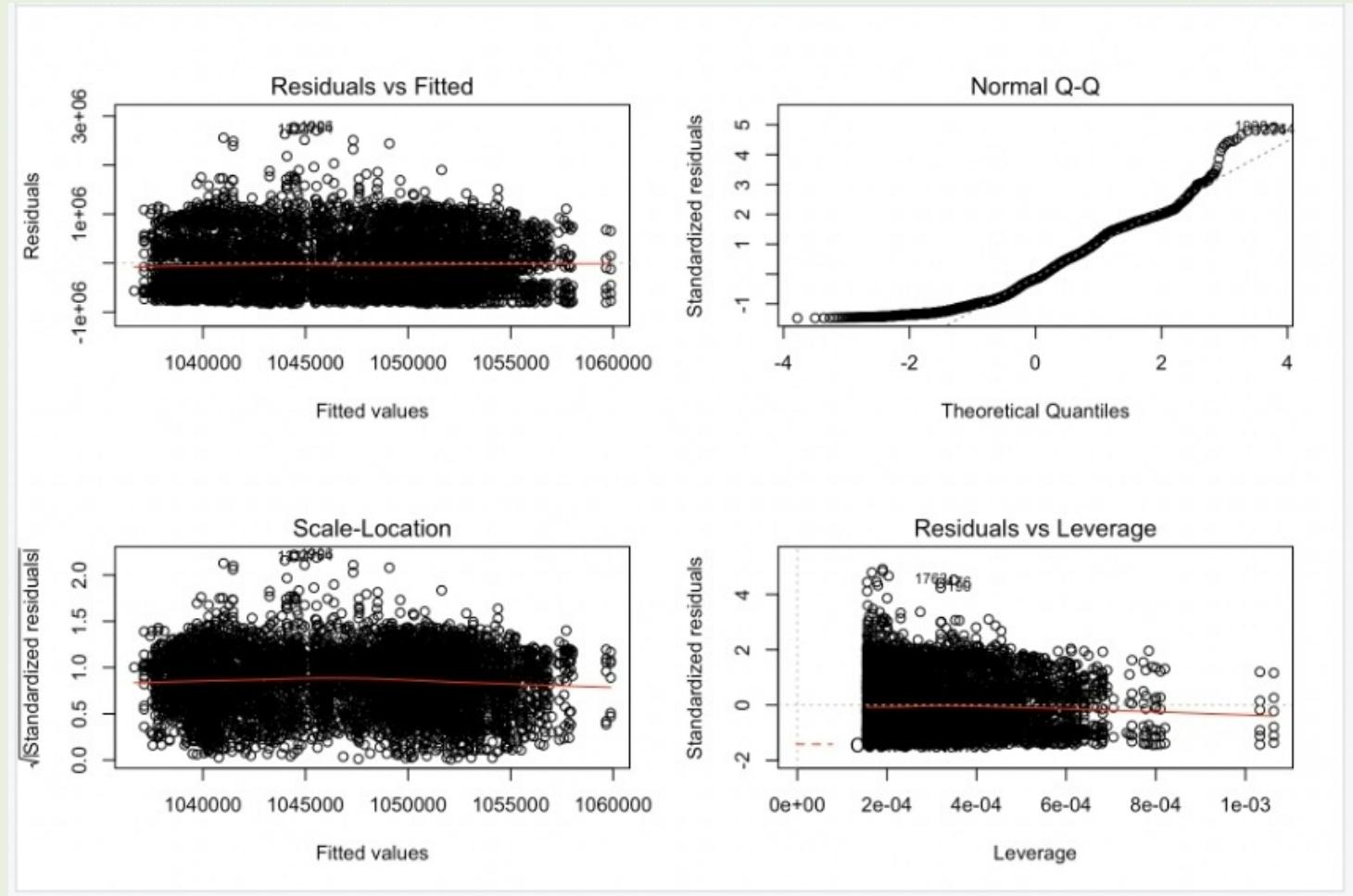


REGRESSION MODELS

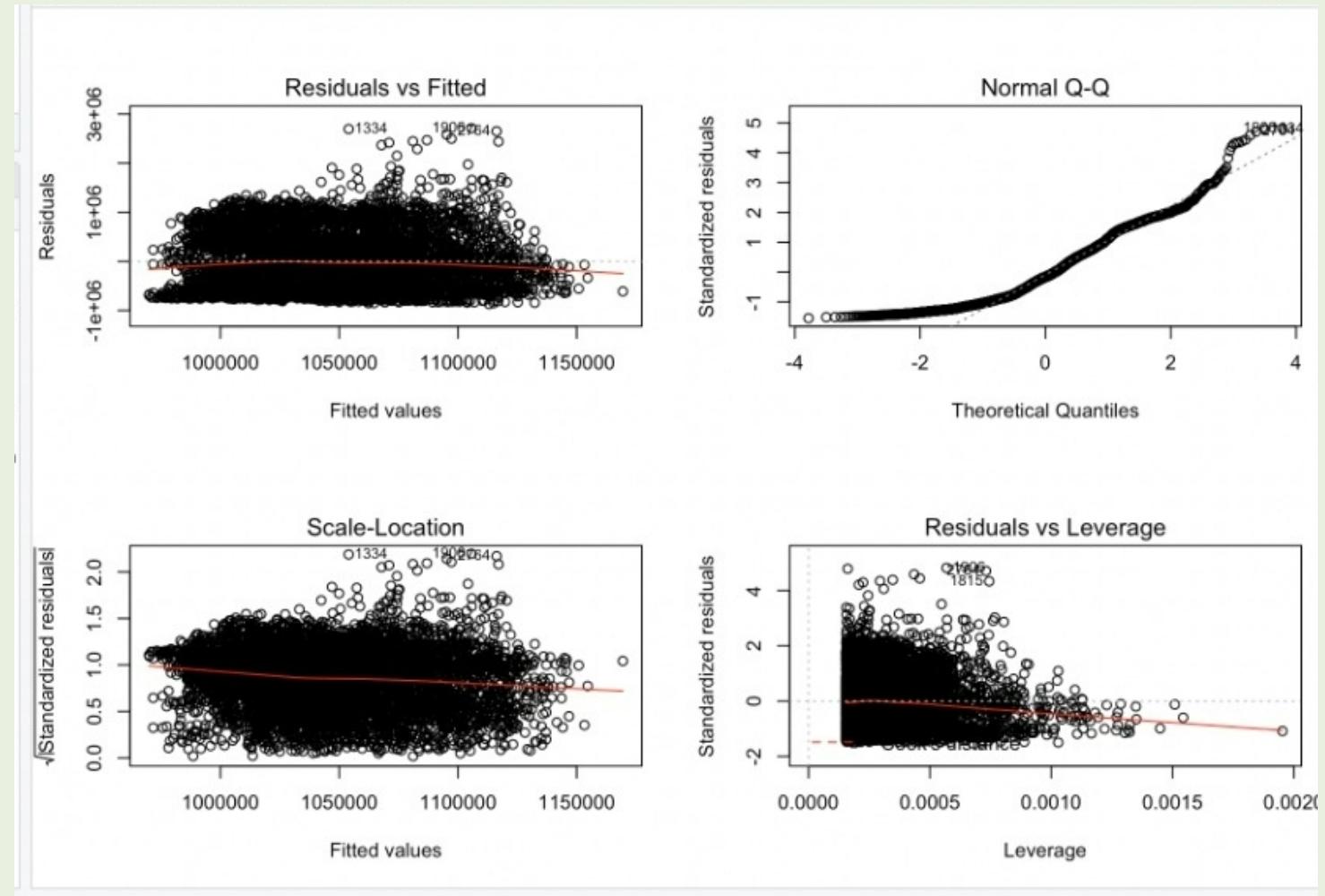
WEEKLY
SALES VS
TEMPERATUR
E



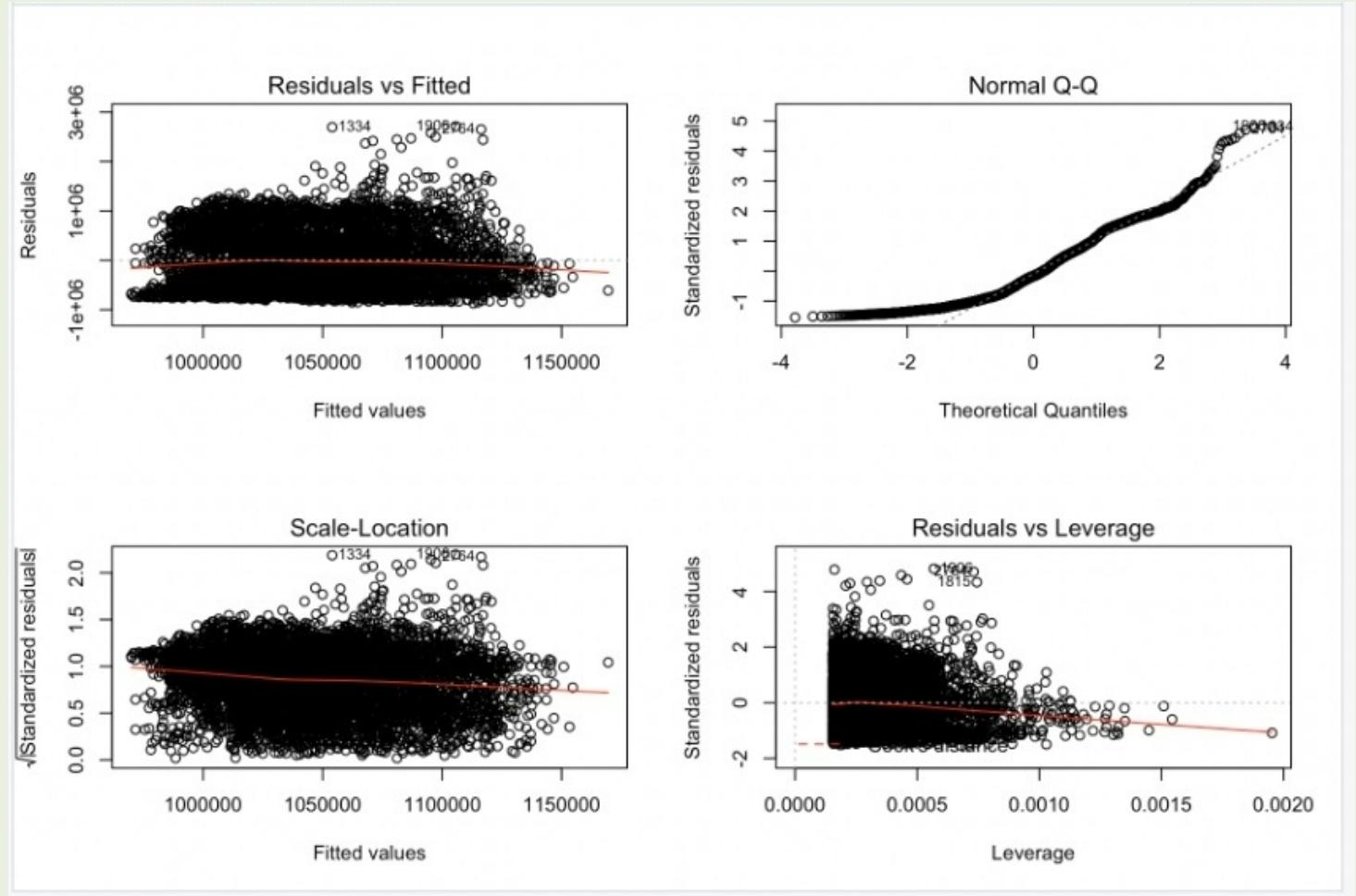
WEEKLY SALES VS FUEL PRICES



WEEKLY SALES VS CPI

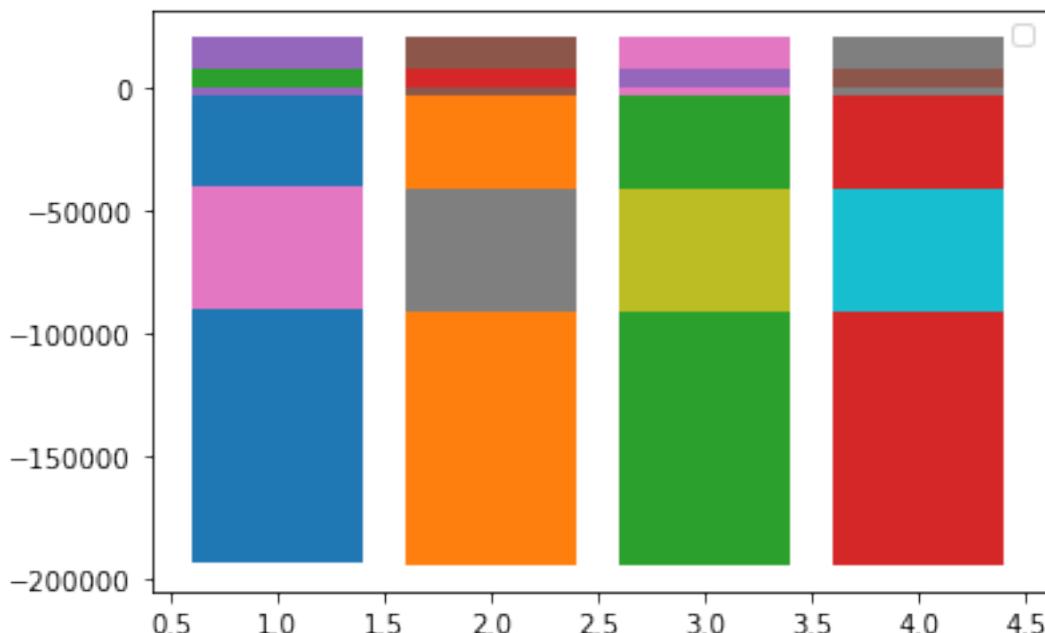


WEEKLY SALES VS TEMPERATURE, FUEL PRICES AND CPI



MACHINE LEARNING

Mean Squared Error Coefficient of all models:



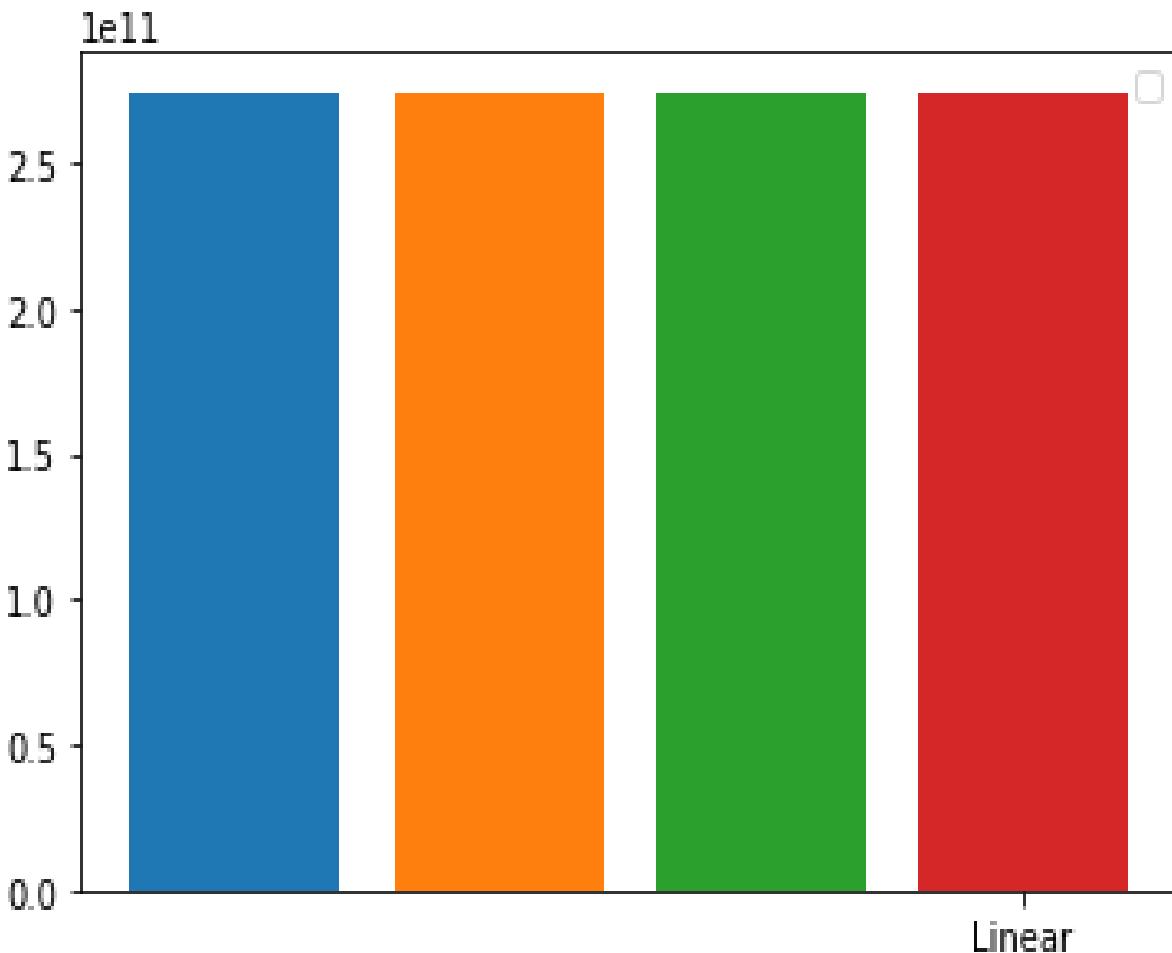
The coefficients of multiple statistical methods were calculated and then plotted in the above graph which just shows the range of values of coefficients where they fall after training or learning.

```
ax=plt.subplot()
for x in range(7):
    ax.bar(1,all_cofs[0][x])
    ax.bar(2,all_cofs[1][x])
    ax.bar(3,all_cofs[2][x])
    ax.bar(4,all_cofs[3][x])

plt.legend()
plt.show()
```

Comparison of MSE of all above 4 models

using K fold cross validation:



The plot above shows the mean error of the regression models in 5-fold which were EN, Ridge, Lasso and Rigid regression and all of them had the same affect as in they have same value of mean error.

```
ax=plt.subplot()
plt.title="Mean Error of all models in 5 folds"
plt.bar(1,np.mean(EN_Err),tick_label="EN")
plt.bar(2,np.mean(Ridge_Err),tick_label="Ridge")
plt.bar(3,np.mean(Lasso_Err),tick_label="Lasso")
plt.bar(4,np.mean(Linear_Err),tick_label="Linear")
plt.legend()
plt.show()
```

CONCLUSION

- Sales are high for the month December in 2010 for all the stores
- Sales are low for the month June in 2012 for all the stores
- Stores 4, 14, 20 have the highest sales among all.
- Sales are less when the temperatures are low.
- Unemployment rate is very low during the holiday season.





THANK YOU