Walmart sales prediction

Dhiraj Bankar

Bellevue University

Abstract

For project 3, I choose to go with the retail dataset. I worked with Sears Holding corporation, who is pioneers of shopping malls, credit cards, layaway, home services and many more. I am here because of sears and I would like to make prediction model for retail though I don't have data from Sears but I found the data for Walmart, I want to explore a predictive model-based dataset and the current dataset provided me the necessary scope required for the project. If sears would have used the Data science and leveraged the power of analytics to save retails industry. Till last year Sears was a sinking boat like a Titanic.

Context

Historical sales data for 45 Walmart stores located in different regions are available. There are certain events and holidays which impact sales on each day. The business is facing a challenge due to unforeseen demands and runs out of stock sometimes, due to inappropriate machine learning algorithm. Walmart would like to predict the sales and demand accurately. An ideal ML algorithm will predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc. The objective is to determine the factors affecting the sales and to analyze the impact of markdowns around holidays on the sales.

Table of Contents

ABSTRACT....2

CONTEXT....2

DATASET:....3
PROPOSED METHODS....5
ANTICIPATED ISSUES:....5

REFERENCES....15

APPENDIX A....16

Dataset:

Below is the URL address for the heart disease dataset.

https://www.kaggle.com/rutuspatel/retail-analysis-with-walmart-sales-data

The dataset has about 8 attributes.

- 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

The advantage with this dataset is that it can be used to build the predictive model as the goal is to predict the sale of the day for particular Walmart store. However, the challenge with this dataset is that the amount of data is very limited and it has only 6435 rows and 8 attributes including the sales which is what I am going to predict.

Holiday Events

Super Bowl: 12-Feb-10, 11-Feb-11, 10-Feb-12, 8-Feb-13

Labour Day: 10-Sep-10, 9-Sep-11, 7-Sep-12, 6-Sep-13

Thanksgiving: 26-Nov-10, 25-Nov-11, 23-Nov-12, 29-Nov-13

Christmas: 31-Dec-10, 30-Dec-11, 28-Dec-12, 27-Dec-13

Analysis Tasks

Basic Statistics tasks

1) Which store has maximum sales

2) Which store has maximum standard deviation i.e., the sales vary a lot. Also, find out the coefficient of mean to standard deviation

3) Which store/s has good quarterly growth rate in Q3’2012

4) Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together

5) Provide a monthly and semester view of sales in units and give insights

Statistical Model

For Store 1 – Build prediction models to forecast demand (Linear Regression – Utilize variables like date and restructure dates as 1 for 5 Feb 2010 (starting from the earliest date in order). Hypothesize if CPI, unemployment, and fuel price have any impact on sales.) Change dates into days by creating new variable.

Select the model which gives best accuracy.

Proposed Methods

There are 3 different models that are being used on the current dataset. Given it's a prediction problem, all the chosen models were regression models. As with any prediction problem, the first model is the multiple linear regression. Followed by the multiple linear regression, the next model used here is the polynomial regression just to make sure and catch any patterns that does not perform really well on the polynomial regression. The third model that was chosen for this dataset is the random forest regression model. The rationale for this model is to make sure to try a regression model which different from traditional regression models such as multiple linear regression or the polynomial regression model.

Anticipated Issues:

As mentioned above, one of the anticipated issue that I might run into with this dataset is the dataset size. Since the data set is only of the size with 6435 rows and 8 columns, I need to be really careful so that the data is **not over fit.** However, I still decided to go with the dataset as I want to learn different ways by which I can work with a smaller dataset and still get a better accuracy. The

other issue I anticipate is to deal with different accuracies I get from models and how do I choose one over the other when these individual models give me vastly different accuracies.

Project Code and actual Analysis:

Pandas Library was used and within that read csv method was used to load the data

Data Information

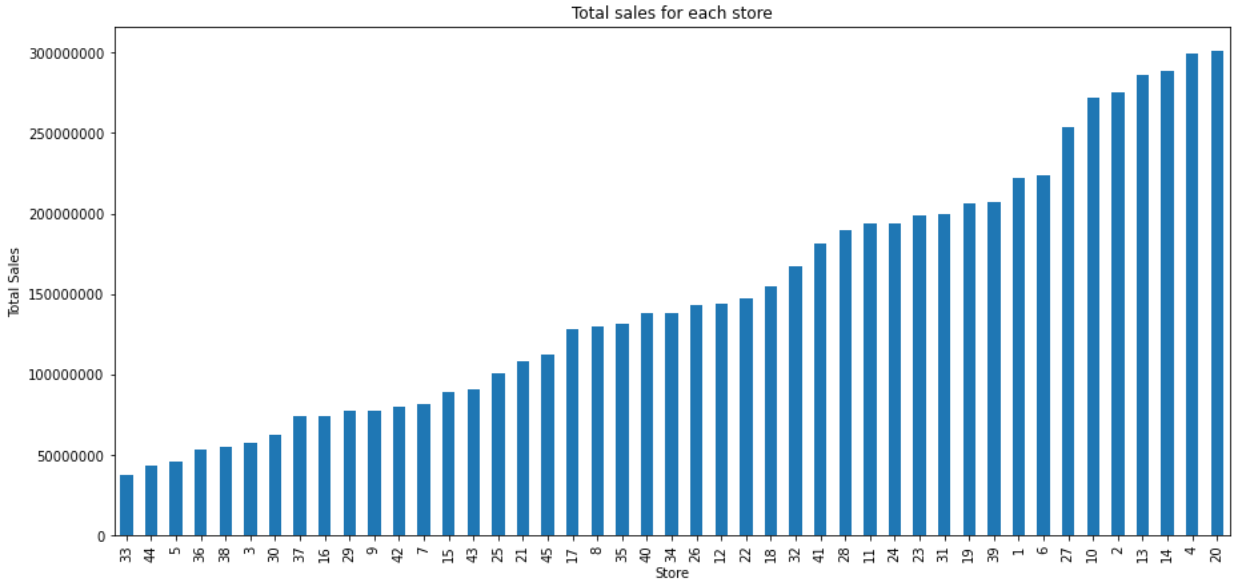
Given dataset has 6435 entries and has total 8 columns.

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 6435 entries, 0 to 6434
Data columns (total 8 columns):
 #   Column        Non-Null Count  Dtype
---  ------        --------------  -----
 0   Store         6435 non-null   int64
 1   Date          6435 non-null   datetime64[ns]
 2   Weekly_Sales  6435 non-null   float64
 3   Holiday_Flag  6435 non-null   int64
 4   Temperature   6435 non-null   float64
 5   Fuel_Price    6435 non-null   float64
 6   CPI           6435 non-null   float64
 7   Unemployment  6435 non-null   float64
dtypes: datetime64[ns](1), float64(5), int64(2)
memory usage: 402.3 KB
```

There are zero missing values, so dataset is neat and very clean and ready to use without data cleaning.

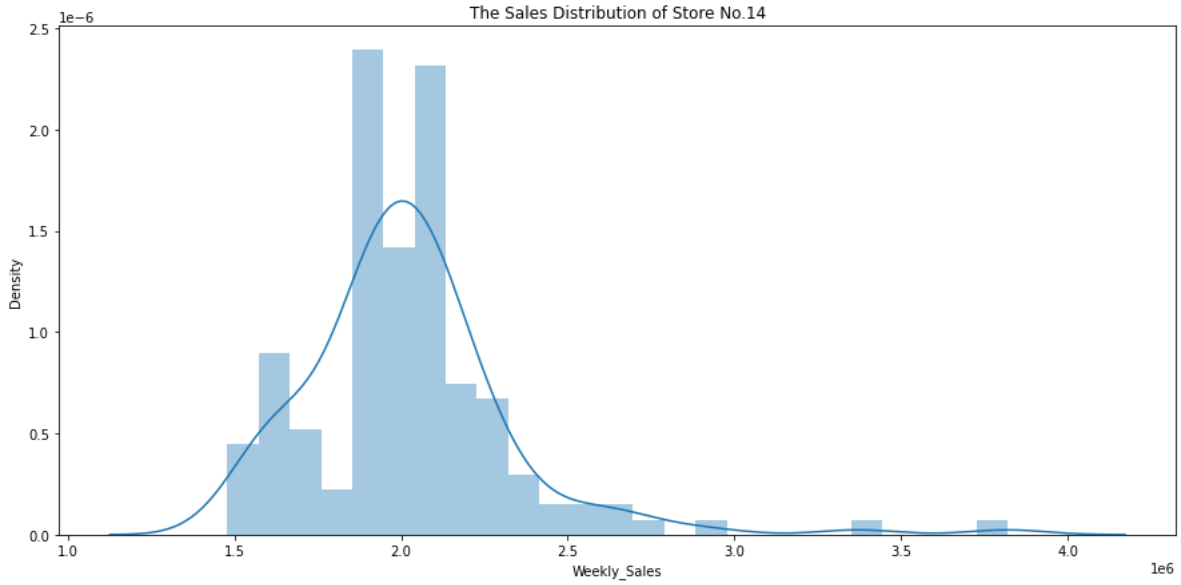
Visualizations

Find the store has maximum sales



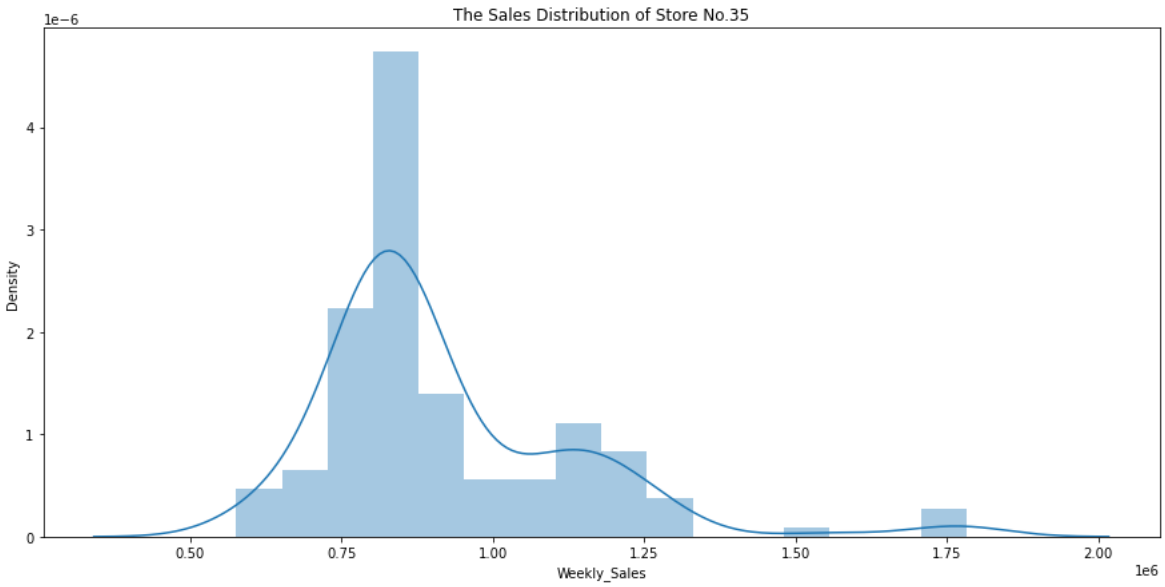
Clearly, from the above graph, it is visible that the store which has maximum sales is store number 20 and the store which has minimum sales is the store number 33.

Which store has maximum standard deviation? i.e. the sales vary a lot. Also, find out the coefficient of mean to standard deviation.



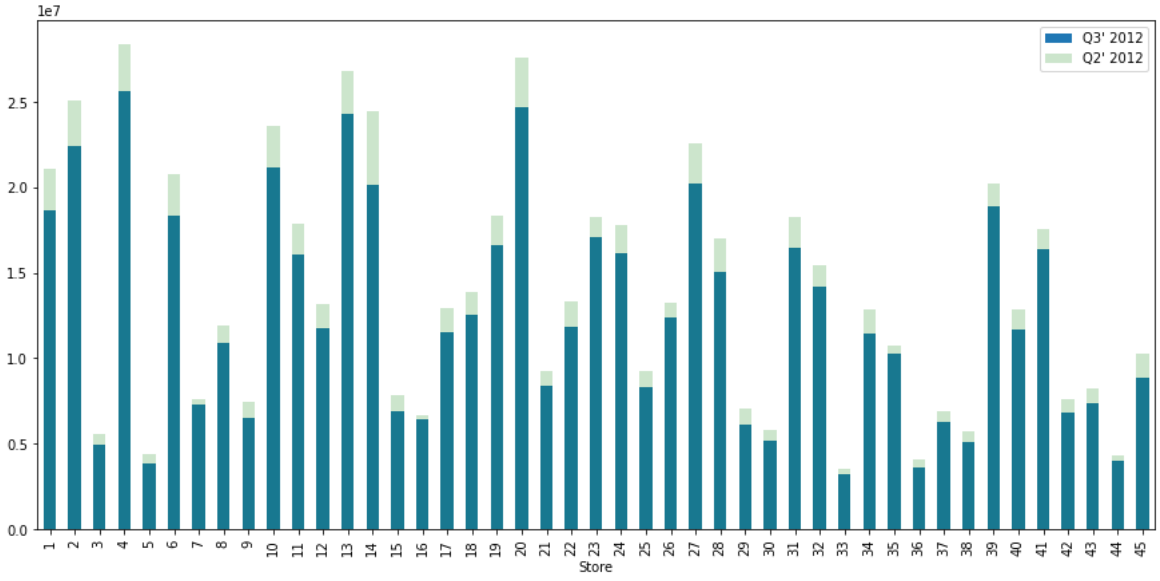
Thus, the store which has maximum standard deviation is store number 14.

Calculating the coefficient of mean to standard deviation



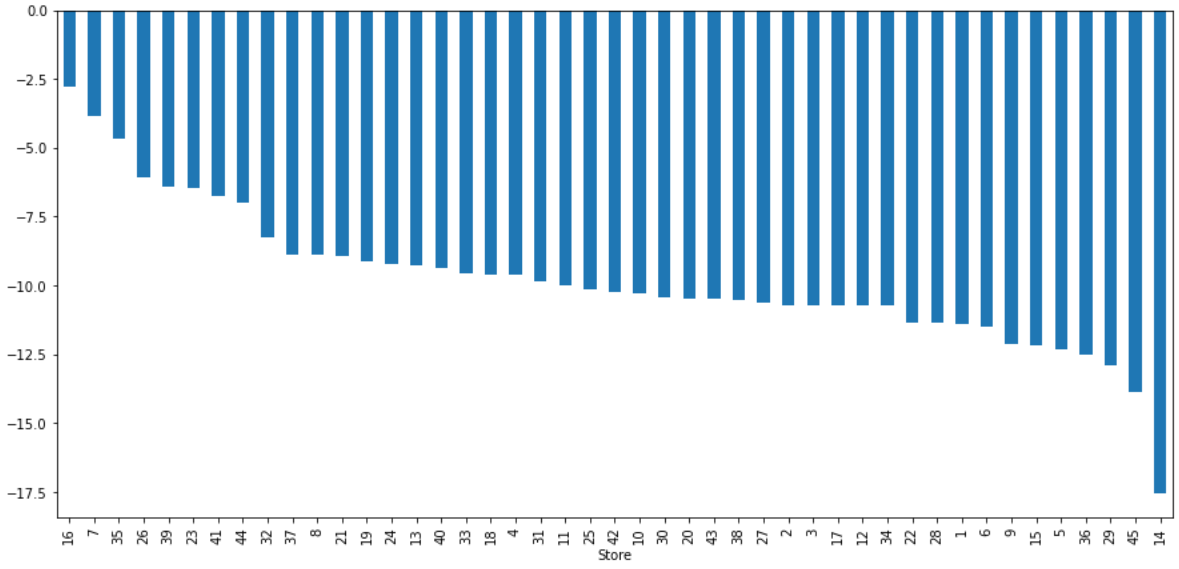
The store which has maximum coefficient of mean to standard deviation is store number 35.

Which store/s has good quarterly growth rate in Q3'2012 ?



Clearly, from the above graph, it is evident that the store which has good quarterly sales in Q3'2012 is store no. 4.

Performance of store by quarter?



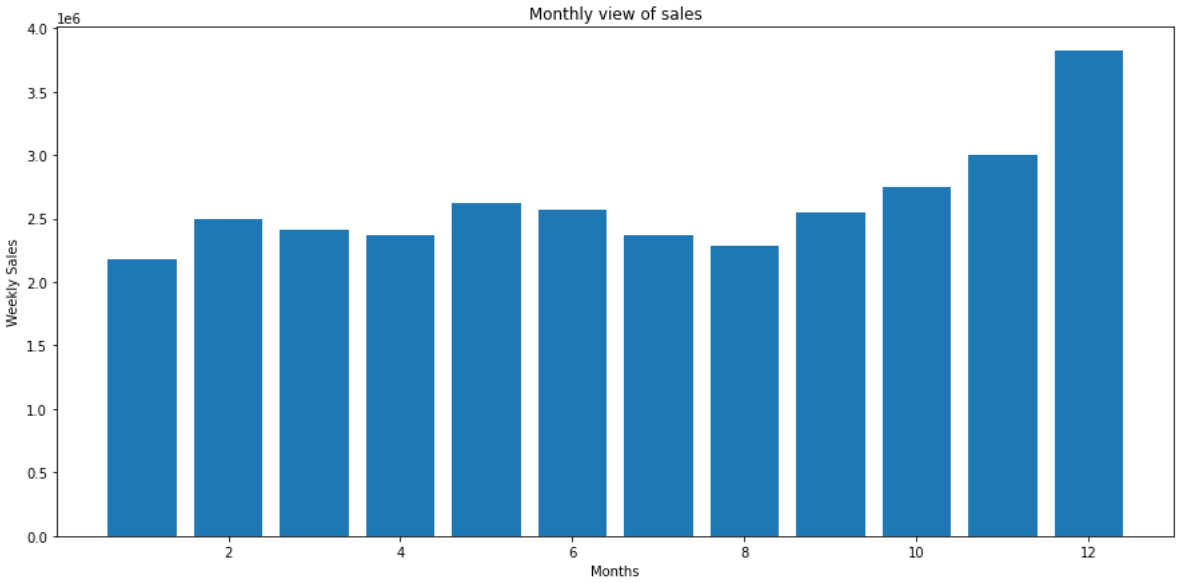
Here, there is no store which has performed better in the 3rd quarter as compared to the 2nd quarter.

Some holidays have a negative impact on sales. Find out holidays which have higher sales than the mean sales in non-holiday season for all stores together.

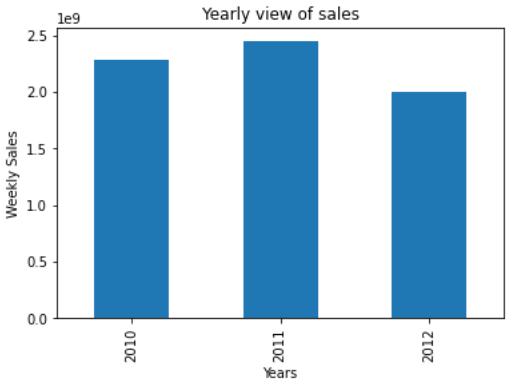
```
{'Super_Bowl_Sales': 1079127.9877037033,
 'Labour_Day_Sales': 1042427.2939259257,
 'Thanksgiving_Sales': 1471273.427777778,
 'Christmas_Sales': 960833.1115555551,
 'Non_Holiday_Sales': 1041256.3802088564}
```

After analysis it clearly says that, Thanksgiving has higher sales than the mean sales on non-holidays.

Monthly view of sales



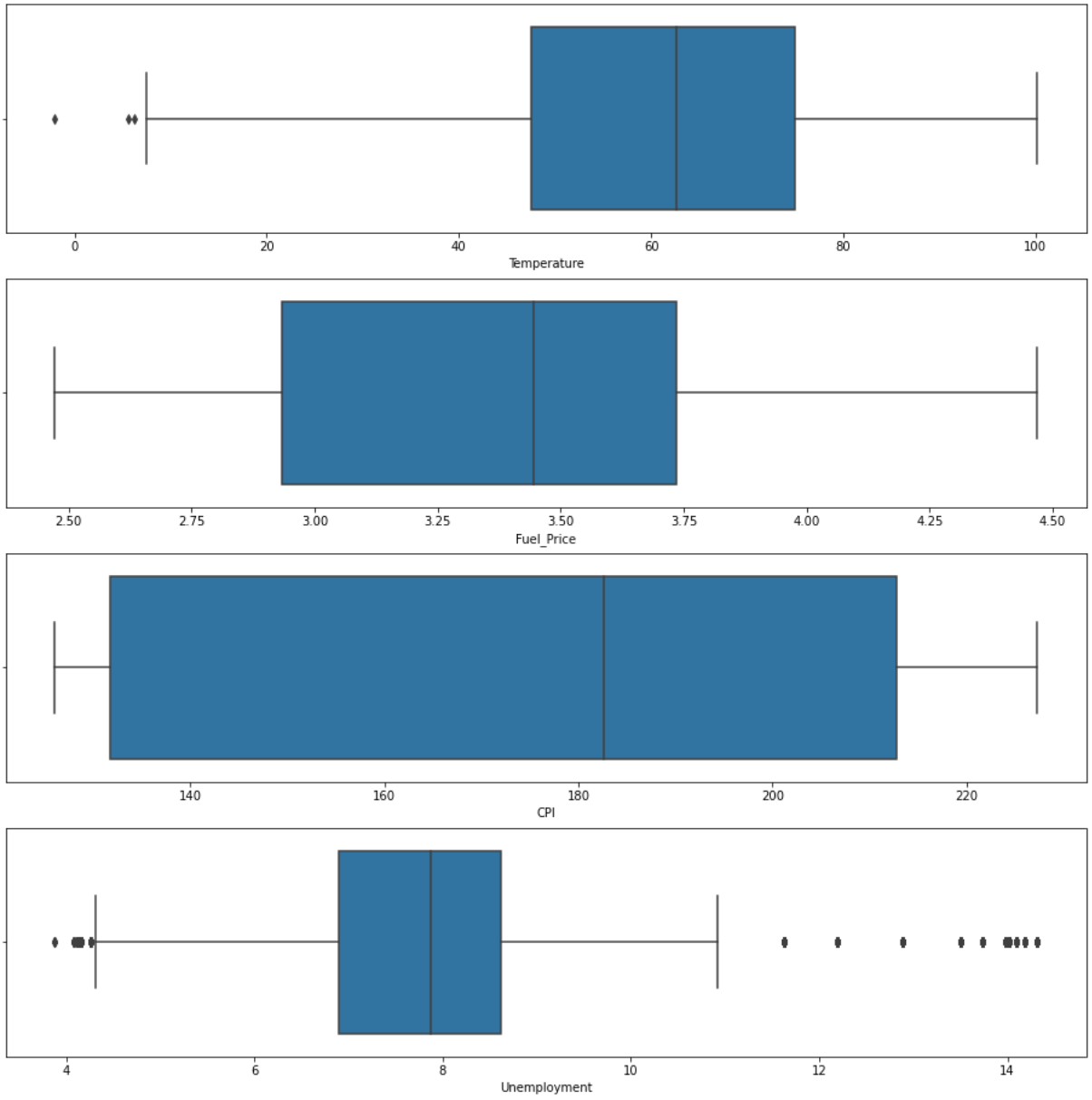
Yearly sales



Here, overall monthly sales are higher in the month of December while the yearly sales in the year 2011 are the highest.

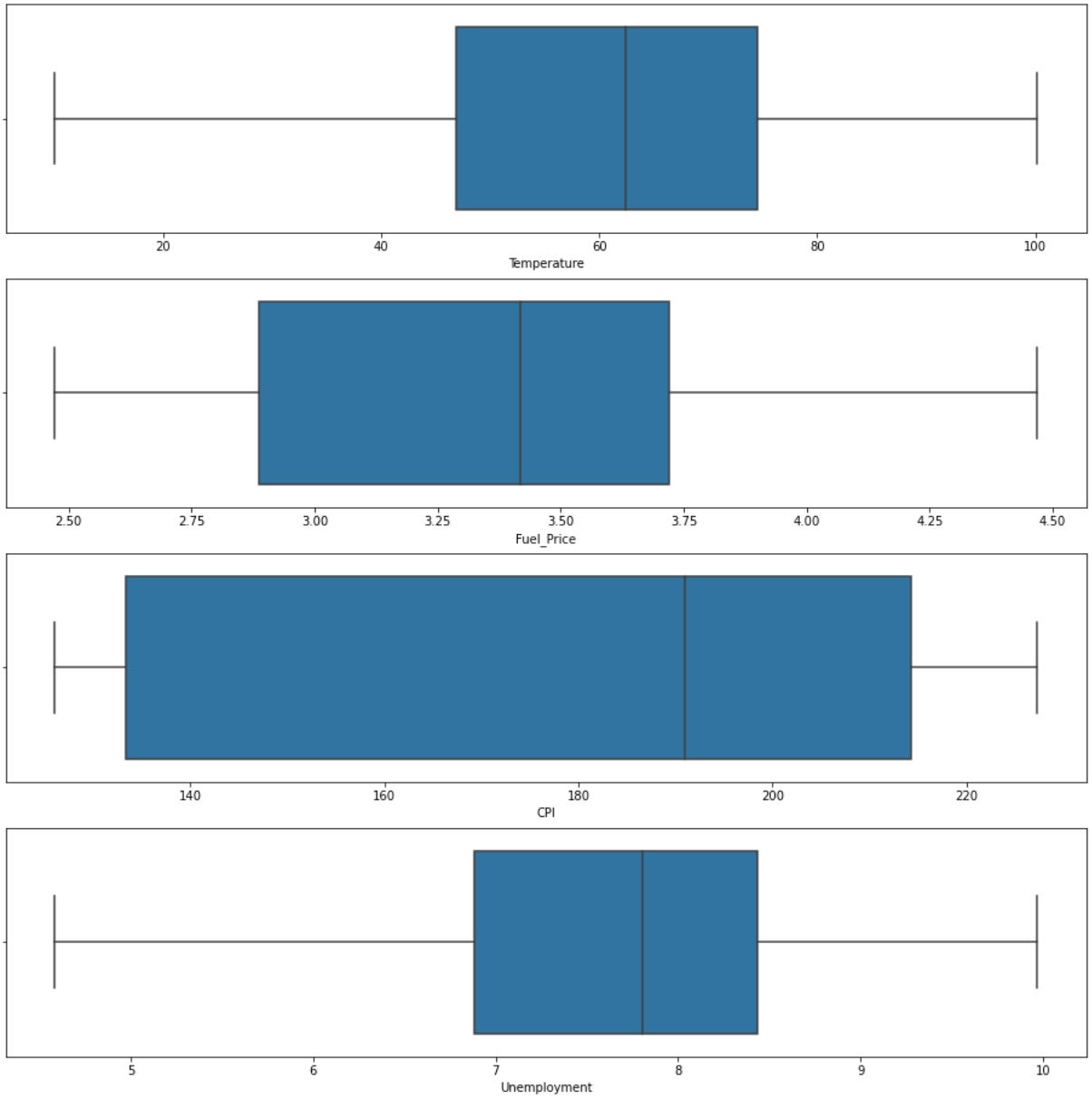
Outliner analysis

Just to data analysis and find the outlines I considered only store 1 with temperature, Fuel price , CPI and unemployment.



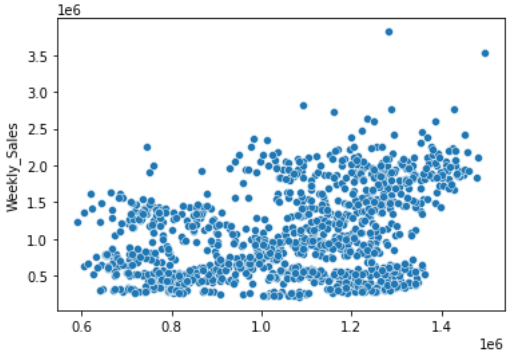
The most of the outliner are there with unemployment data. So I removed the unemployment data greater that 10 and less than 4.5.

After outliner data clean-up

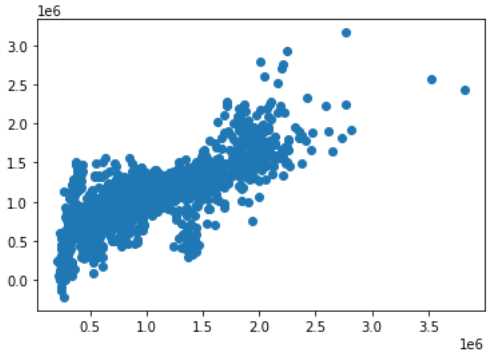


Model fitting

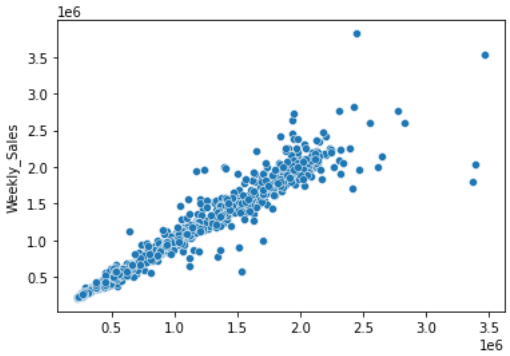
Linear Regression



Polynomial Regression Model



Random Forest regressor



Conclusion

Linear Regression is not an appropriate model to use which is clear from it's low accuracy as well as Polynomial Regression model has accuracy of 59%. However, Random Forest Regression gives accuracy of over 95% , so, it is the best model to forecast demand.

Questions

1) What do the CPI column represent?

Answer: Prevailing consumer price index The Consumer Price Index (CPI) is a measure that examines the weighted average of prices of a basket of consumer goods and services, such as transportation, food, and medical care. It is calculated by taking price changes for each item in the predetermined basket of goods and averaging them. Changes in the CPI are used to assess price changes associated with the cost of living.

2) Does the holiday affects the sells for particular store?

   Answer: Yes correct.

3) What will be the final outcome of this dataset, What will be the driving factor for this model?

   Answer: The Random forest model has fantastic accuracy 95%

4) What is CPI role for this model?

   Answer: The CPI is one of the most frequently used statistics for identifying periods of inflation or deflation. It may be compared with the producer price index (PPI), which instead of considering prices paid by consumers looks at what businesses pay for inputs.

5) Does number unemployment the model?

   Answer- It could be, It will get clarified at the end of model. Still I am working on multiple factors and model.

6) Does temperature affects the model?

   Answer – Yes

7) What is most sales day in all holidays?

   Answer: **Clearly, Thanksgiving has higher sales than the mean sales on non-holidays.**

8) How would you treat ‘Unknown’ label?

Answer: replace them with nulls first, since they don't provide any information. Thereafter imputed them with values based on K nearest neighbor to remove all nulls. However, all three have minimal effect on the dependent variable, so they can be left alone as well.

9) Can this model is applicable to all geographic conditions?

   Answers: This dataset is only analyzed or considered for United States of America with 46 stores only.

10) Does geographic conditions change the sales of the store?

    Answer: Defiantly it affects the sales depending of location of store.

11) Will this model is ready to use by any store of Walmart?

    Answer: I would say model is ready to use in any Walmart store, but it would be much better if there is one more eye and some more testing of the model.

References

Applied Text Analysis with Python, Benjamin Bengfort, Rebecca Bilbro & Tony Ojeda

Machine Learning with Python Cookbook, Chris Albon

Dataset:

https://www.kaggle.com/rutuspatel/retail-analysis-with-walmart-sales-data

https://www.investopedia.com/terms/c/consumerpriceindex.asp

Appendix A

Categorical data understanding

https://medium.com/hugo-ferreiras-blog/dealing-with-categorical-features-in-machine-learning-1bb70f07262d

https://en.wikipedia.org/wiki/Feature_(machine_learning)

LinearRegression:

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html?highlight=linear regression#sklearn.linear_model.LinearRegression

RandomForestClassifier:

https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

PolynomialFeatures:

https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.PolynomialFeatures.html