Walmart sales forcasting
Jayaprakash (prakashjz121@gmail.com)

Walmart Retail Business Use case

Problem Statement

- *A retail store that has multiple outlets across the country are facing issues in managing the inventory - to match the demand with respect to supply. You are a data scientist, who has to come up with useful insights using the data and make prediction models to forecast the sales for X number of months/years.*

Project Objective

- Find out useful insights from given dataset for each stores to improve various areas,
- Understand the dataset and features , and
- Use suitable Data Preprocessing and Feature Selection/Engineering Methods

Data Description

- Given dataset contains 6435 rows and 8 columns,
- Primary key is Store ,

Data Pre-processing Steps and Inspiration

Data Exploration

- During data analysis , it is observed that there are NO null values in the given dataset.
- There are few unique records / rows observed for each features.

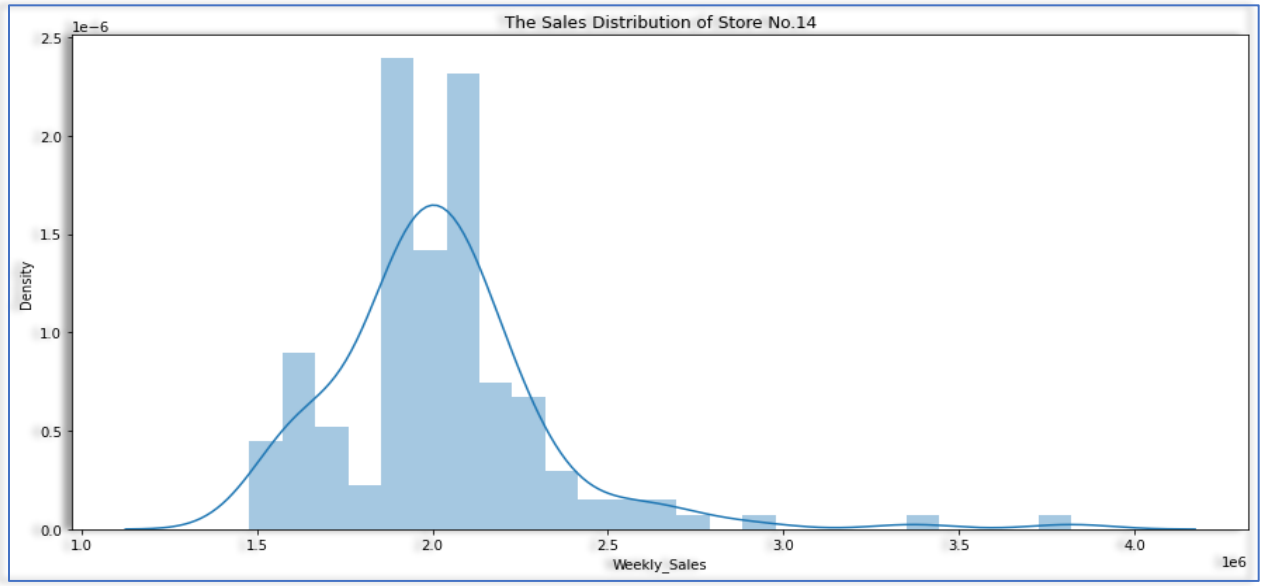
```
df.nunique().sort_values()
```

```
Holiday_Flag        2
Store              45
Date              143
Unemployment      349
Fuel_Price        892
CPI              2145
Temperature      3528
Weekly_Sales     6435
dtype: int64
```

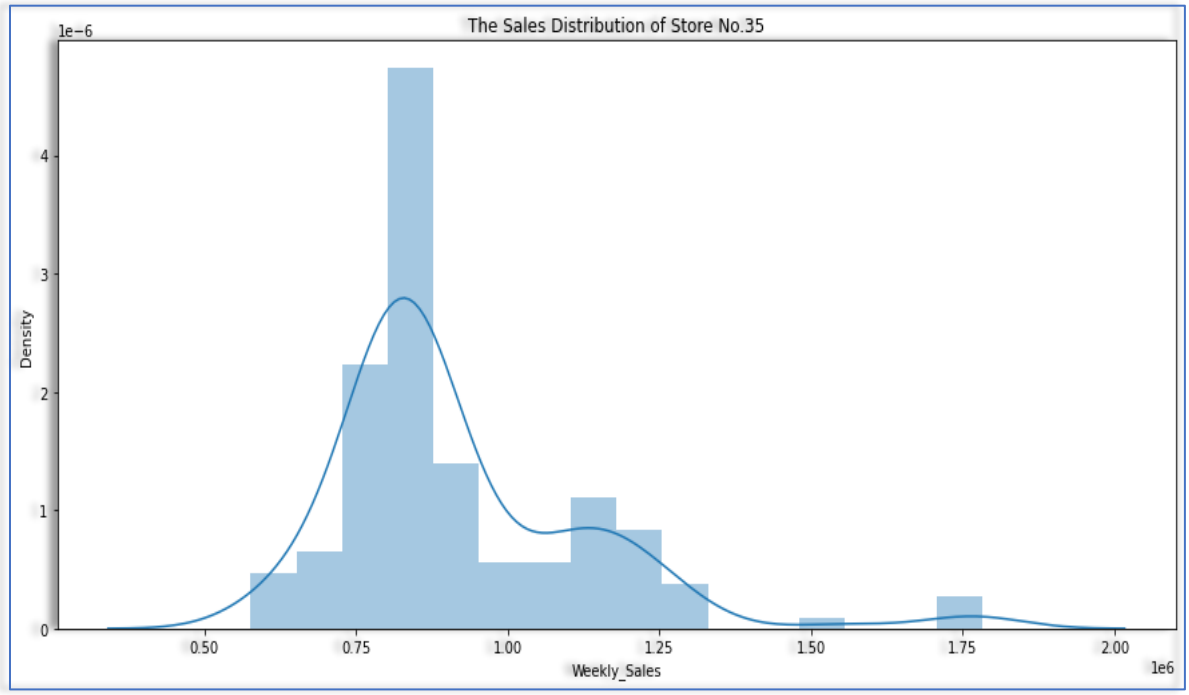
- Split the data into features and target variable & date column into day,month,year for better data analysis , and
- Give dataset has 6 numerical & 5 categorical features.

Exploratory Data Analysis (EDA)

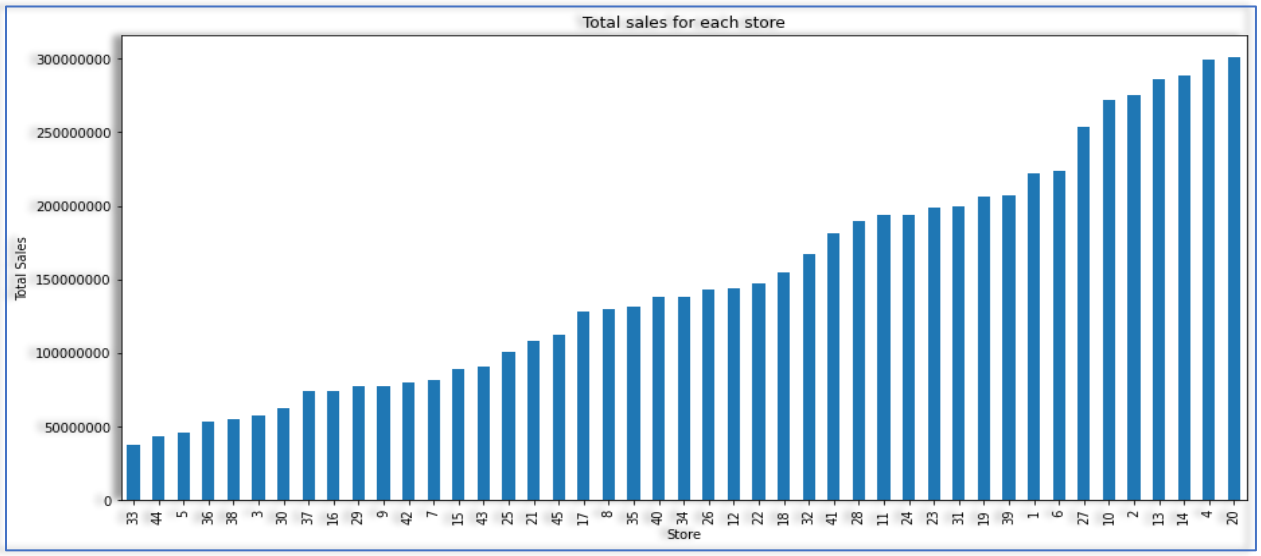
- In given dataset , target Variable (weekly Sales) seems to be normally distributed, averaging around 20 units,



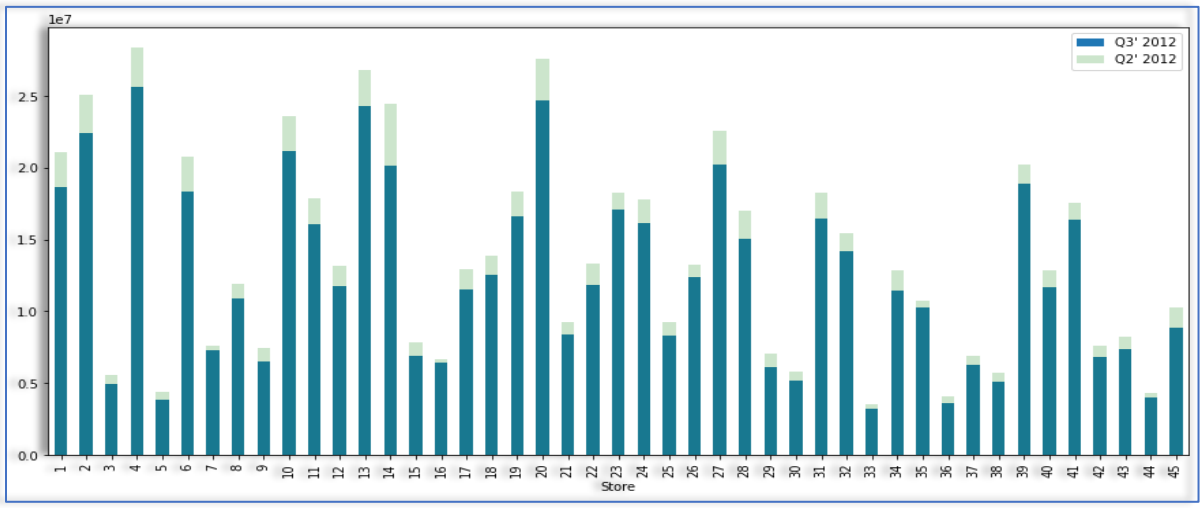
- Store-35 has maximum coefficient of mean to standard deviation , hence sales variation in that store is more compared to other stores ,



- **Sales analysis - Max/ min sales stores** :
    - Store-20 has maximum sales and Store-33 has minimum sales ,
    - Store-14 has maximum standard deviation , thus it has sales more sales variation,



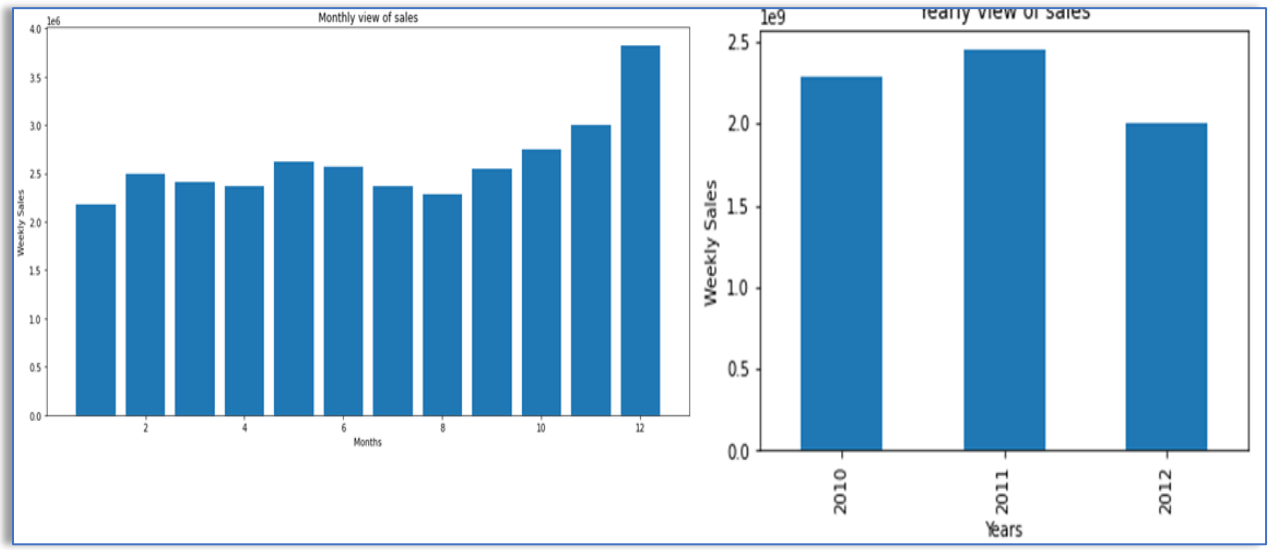
- **Data analysis - 2012 Q3&Q4** : Store 4 has maximum sales in Q3'2012 ,



- **Holidays impact on Sales** : from the given dataset ,looks like sales was Higher in thanks giving day compared to other Holidays ,

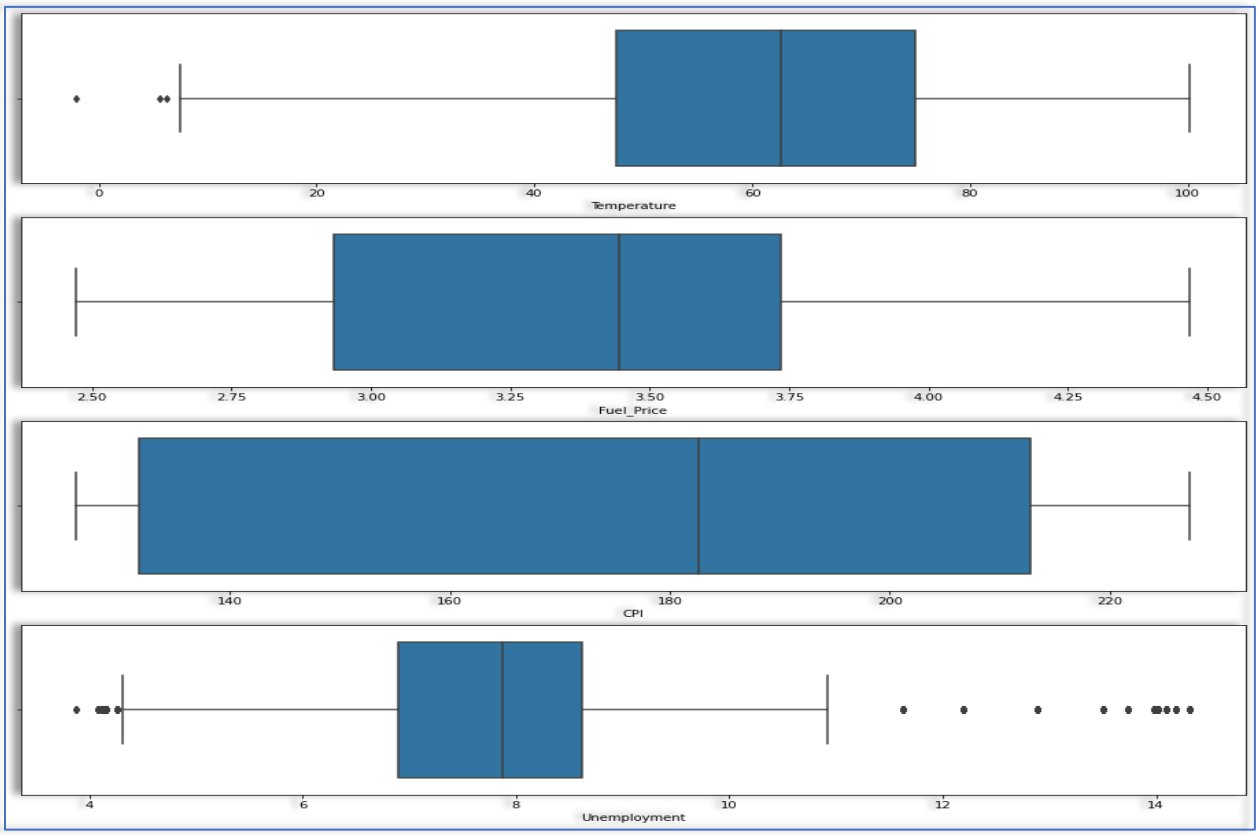
```
{'Super_Bowl_Sales': 1079127.9877037038,
 'Labour_Day_Sales': 1042427.293925926,
 'Thanksgiving_Sales': 1471273.427777778,
 'Christmas_Sales': 960833.1115555555,
 'Non_Holiday_Sales': 1041256.3802088555}
```

- **Monthly & Yearly Sales :** Overall monthly sales are higher in the month of December while the yearly sales in the year 2011 are the highest.

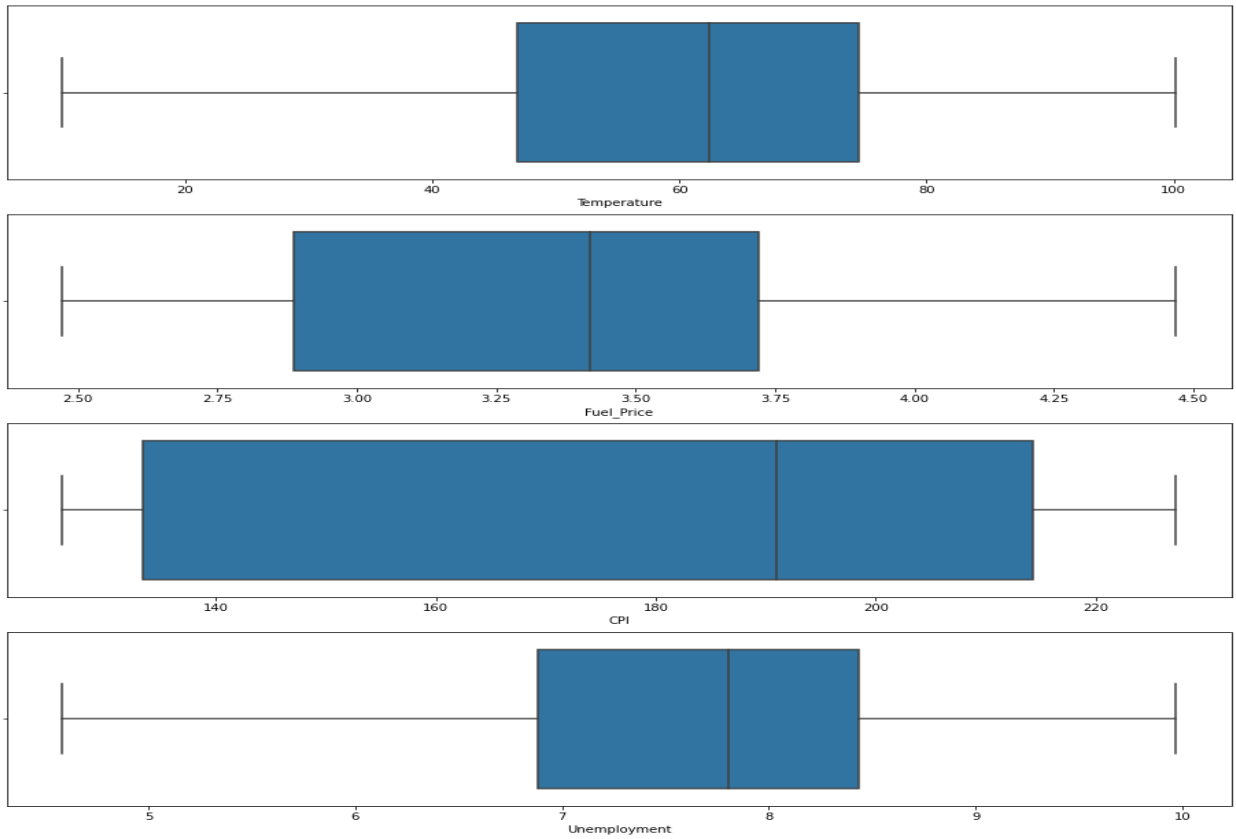


Choosing the Algorithm for the Project

- Look at the outliers and remove them for better data analysis,

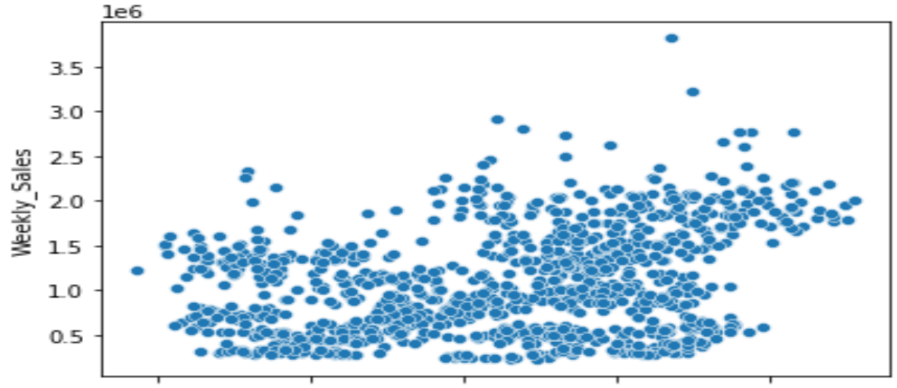


- There were outliers w/ few variable which can be removed for better analysis,

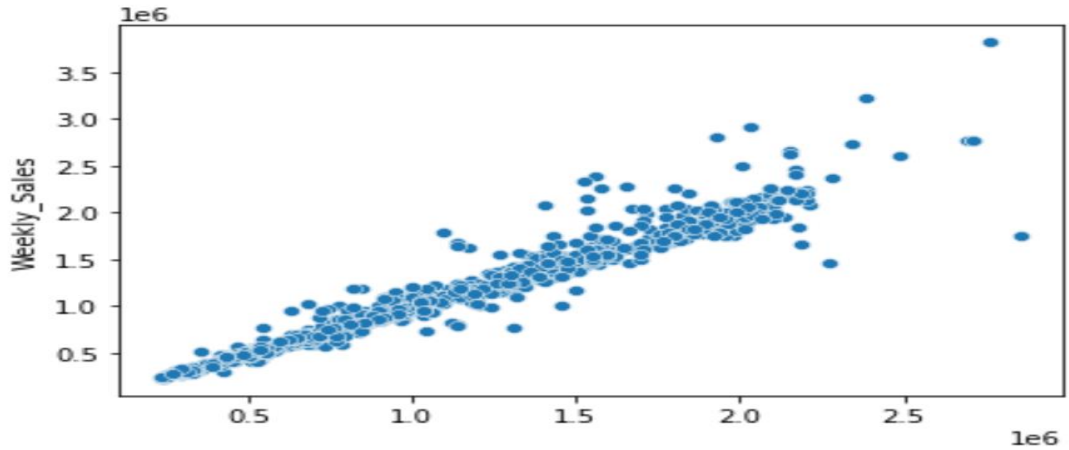


- Have tried Linear Regression & Random Forest Model for the given problem and dataset , and
- Linear Regression is not an appropriate model to use which is clear from it's low accuracy (14%). However, Random Forest Regression gives accuracy of over 98% , so, it is the best model to forecast.

```
Linear Regression:
-------------------
-------------------
Accuracy: 13.986765595483108
Mean Absolute Error: 454302.458649398
Mean Squared Error: 299938590243.44037
Root Mean Squared Error: 547666.4954545242
```



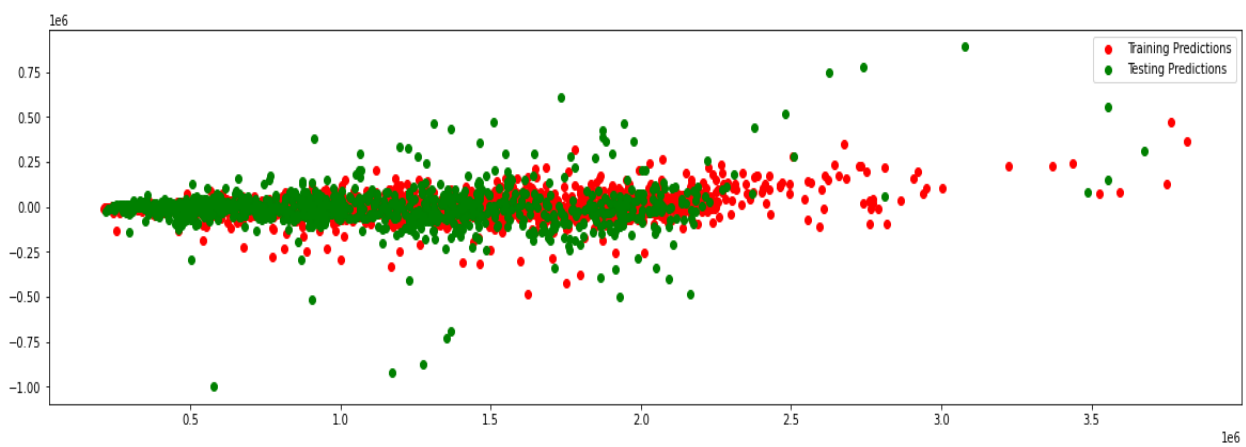
```
Random Forest Regressor:
--------------------
--------------------
Accuracy: 95.0255702346395 5
Mean Absolute Error: 63781.16694310955
Mean Squared Error: 16282283486.600037
Root Mean Squared Error: 127602.05126329293
```



- From the above chart , accuracy for random forest mode is 95 % when compared to Linear regression (13%) , so let's use Random forest for further analysis,
- Fine tune the selected mode w/ hyper parameters ,

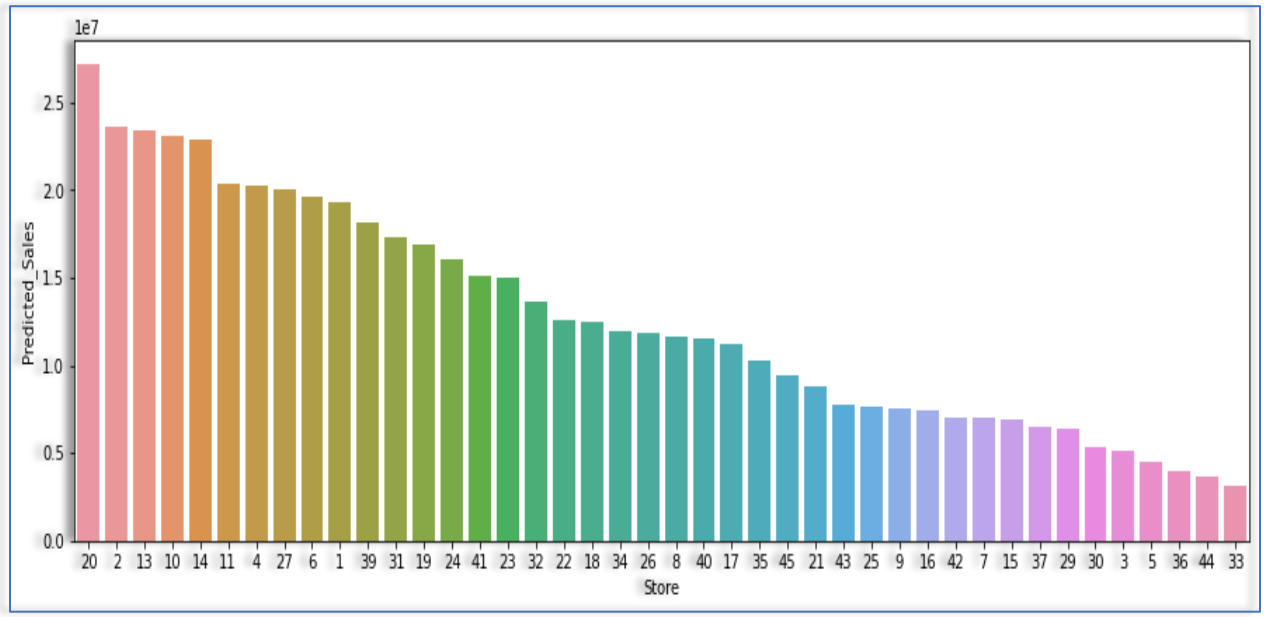
```
For Training Data
 ---------------------------------
MAE:   23200.58232336495
MSE:   1858452353.8690917
r2:   0.9944019291484907
RMSE:   43109.77097908422

For Test Data
 ---------------------------------
MAE:   62368.5441530919
MSE:   14000771193.26318
r2:   0.9581931281303442
RMSE:   118324.8545034524
```

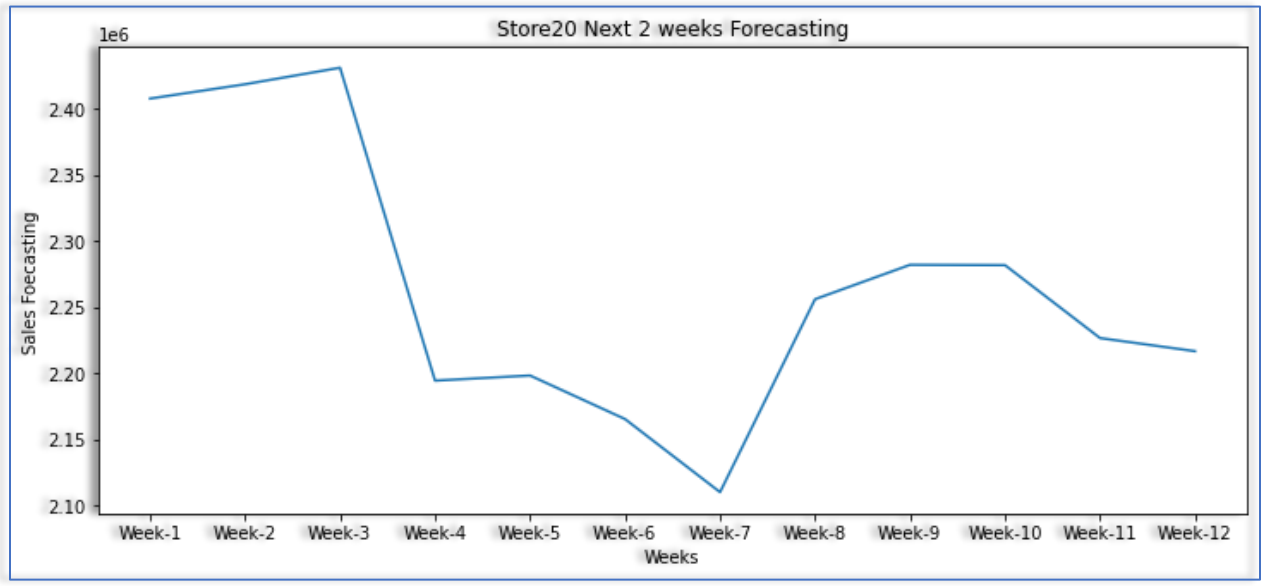


Forecasting the Sales for next 12 weeks

- Forecasting shows the store-20 and 33 have max , min sales respectively.



- Let's look at the forecasting for Store-20 , it is evident that week-3 will have max sales and week-7 will have low sales.



Motivation and Reasons For Choosing the Algorithm

- Given dataset was labelled and our problem statement was of prediction, hence we have used different supervised learning algorithms used for prediction,
- All the algorithms used in this project are :
    1. Linear Regression

2. Decision Tree
3. Random Forest

Assumptions

NA

Model Evaluation and Techniques

Inferences from the Same

- Best insight to focus is that both Thanks Giving and Christmas has a lot of impact on sales,
- It is possible to notice that some variables like temperature and IsHolliday have the biggest sales accumulated at some ranges, but not much of linear relationships.
- The low selling stores should look forward to increasing their size and capacity to store more items and consumer products,
- January sales are significantly less than other months. This is the result of November and December high sales. After two high sales month, people prefer to pay less on January.
- The low selling stores should look forward to increasing their size and capacity to store more items and consumer products.
- Special discount coupons can be distributed during low selling periods to attract more customers
- CPI, temperature, unemployment rate and fuel price have no pattern on weekly sales, and
- Data analysis shows that Christmas, Thanksgiving and Super Bowel are very important than other weeks for sales and 5th important time is 22nd week of the year and it is end of the May, when schools are closed. Most probably, people are preparing for holiday at the end of the May.

Future Possibilities of the Project

- Special discount coupons can be distributed during low selling periods to attract more customers,
- Sales are likely to fluctuate during holidays. Special offers can be given during festive season accompanied with suitable marketing to keep the sales high during holidays as well ,
- To check into the store that have poor sales and check deep what makes those bad ,
- To further improve the predictive model using the ensembling method to combine models and come with better model , and
- Take the data to Store level and to predict the store level sales which would help to solve the inventory management issues and supply chain management.