CAPSTONE PROJECT

WALMART – SALES PREDICTION

BY
MANASA S

TABLE OF CONTENTS

PROBLEM STATEMENT

One of the leading retail stores in the US, Walmart, would like to predict the sales and demand accurately. There are certain events and holidays which impact sales on each day. There are sales data available for different stores of Walmart. The business is facing a challenge due to unforeseen demands. An ideal ML algorithm can predict demand accurately and ingest factors like economic conditions including CPI, Unemployment Index, etc. Historical sales data for 45 Walmart stores located in different regions are available that covers sales from 2010-02-05 to 2012-10-26.

The aim is to come up with strategies that can improve sales in various stores.

PROJECT OBJECTIVE

1. To come up with useful insights that can be used by each of the stores to improve in various areas.
2. To make prediction models to forecast the sales using all features.
3. To forecast the sales for each store for the next 12 weeks using time series.

DATA DESCRIPTION

The ‘walmart.csv’ contains 6435 rows and 8 columns.

S.No	FEATURE NAME	DESCRIPTION
1.	Store	The store number
2.	Date	The week in which sales happened
3.	Weekly_Sales	Sales for the given store in that week.
4.	Holiday_Flag	Whether the week is a special holiday week. 1 – Holiday week and 0 – Non-holiday week
5.	Temperature	Temperature on the day of sales.
6.	Fuel_Price	Cost of the fuel in the region.
7.	CPI	Prevailing consumer price index
8.	Unemployment	Prevailing unemployment rate

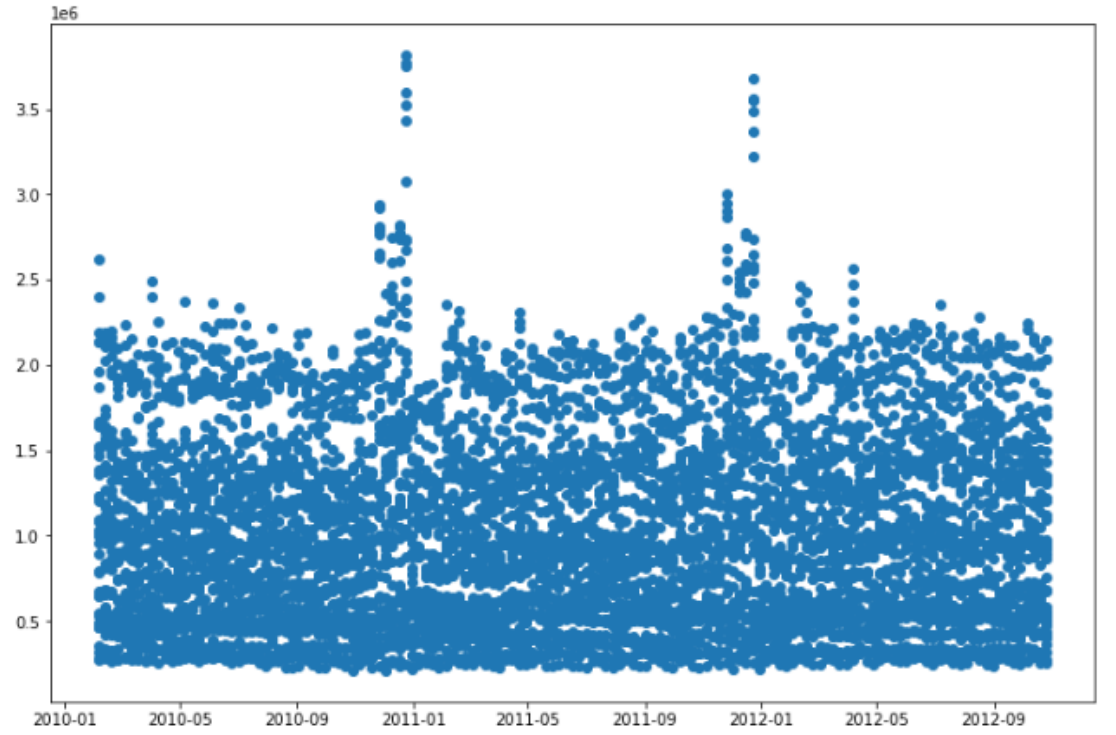
DATA PREPROCESSING STEPS AND INSPIRATION

1. EDA:

(a) It is observed that few weeks sales are really high may be due to festival and other reasons. So the information about the date is extracted (i.e.) day, month and year.

```
2 plt.scatter(data['Date'],data['Weekly_Sales'])
```

```
<matplotlib.collections.PathCollection at 0xb122f70>
```



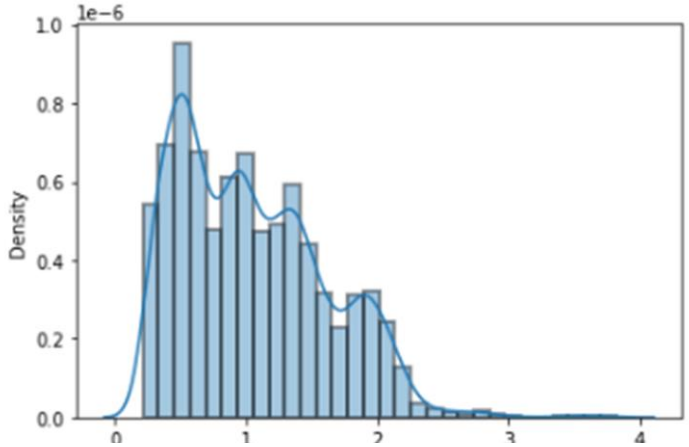
(b) (i) Non holiday weeks have shown higher sales than holiday weeks for a few weeks

(ii) Higher the unemployment rate, lower the sales

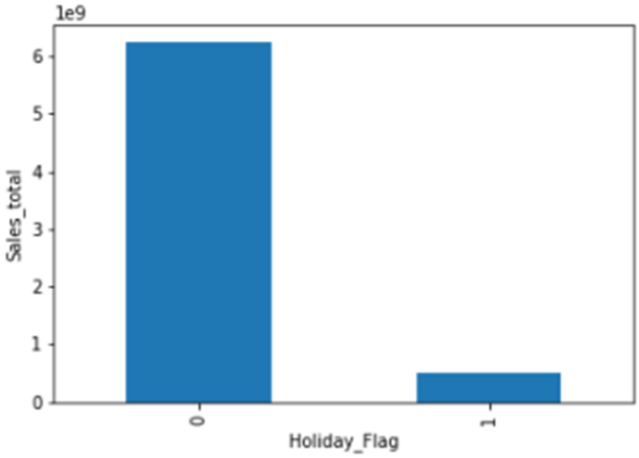
(iii) Among all the months December generally shows the highest sale

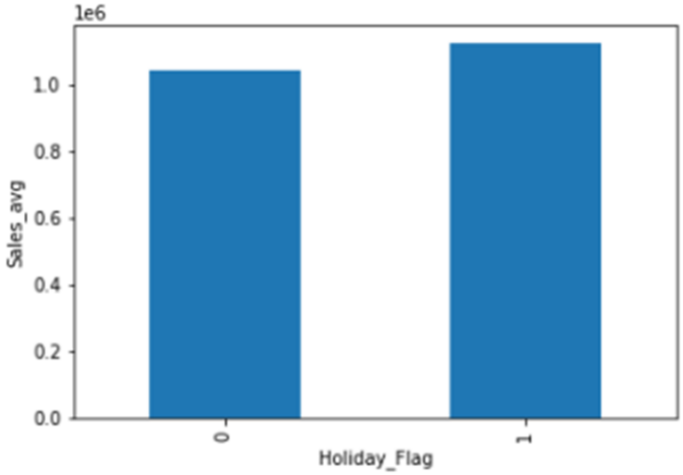
(iv) Year on year the maximum sale amount has decreased.

(c) Weekly sales data is not normally distributed.

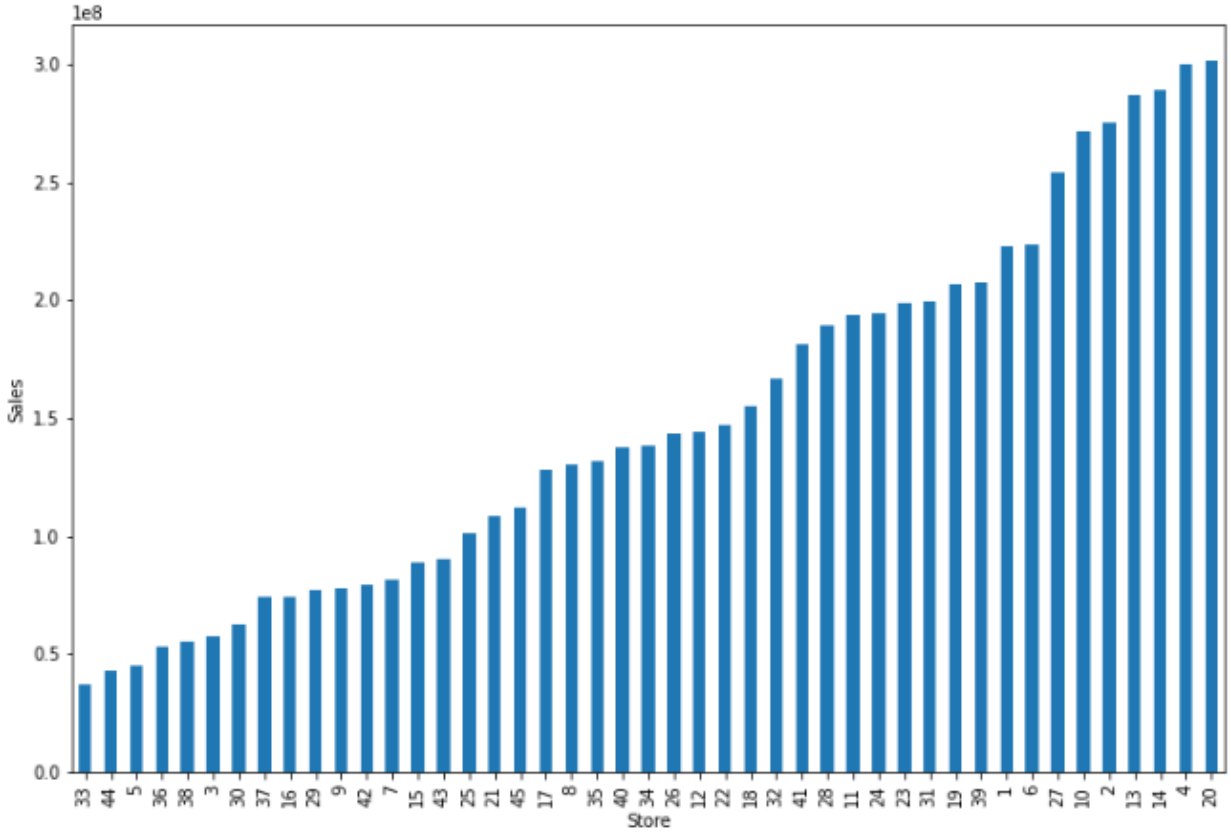


(d) Weeks with non-holidays are observed to have higher sales due to higher number of weeks without holidays. But the average sales on a holiday are observed to be higher.

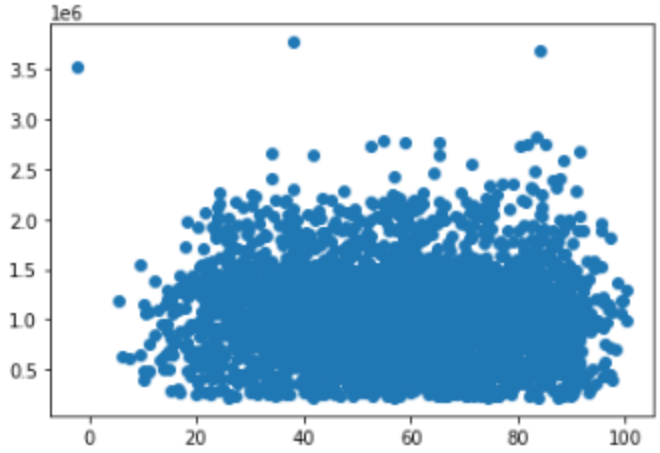




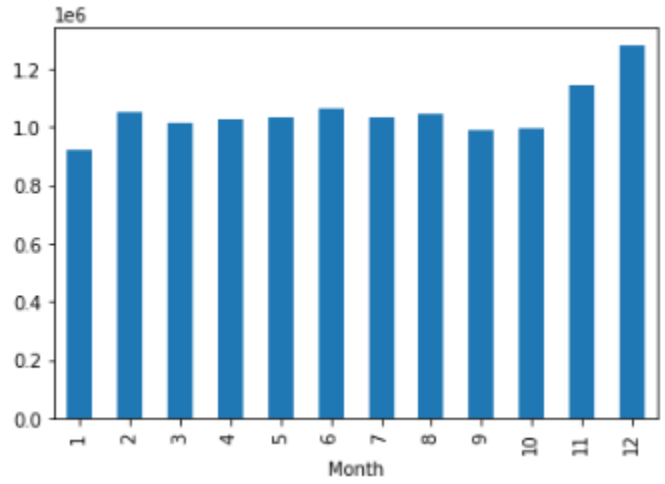
(e) Less than half the stores have total sales more than 1.5e+08. (Maximum sum of sales being 3e+08 by stores 4 and 20).



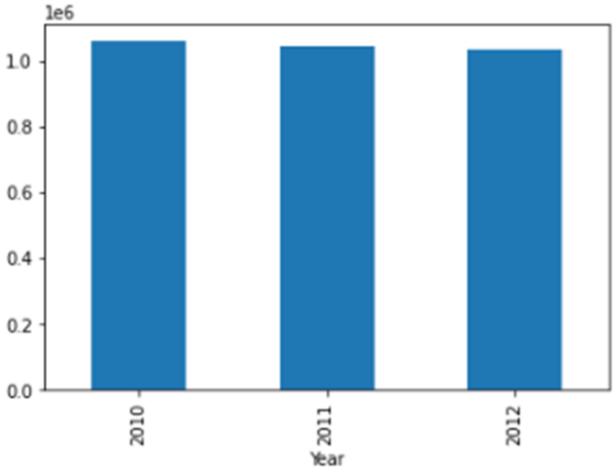
(f) Weekly sales don't seem to have been impacted by temperature much.



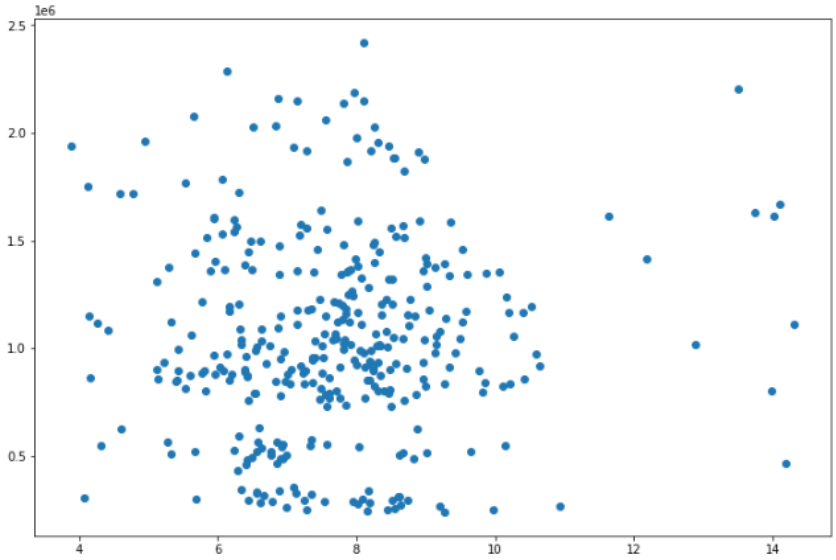
(g) The average sales over years have been the same between February and October. November and December show higher sales obviously due to Christmas.



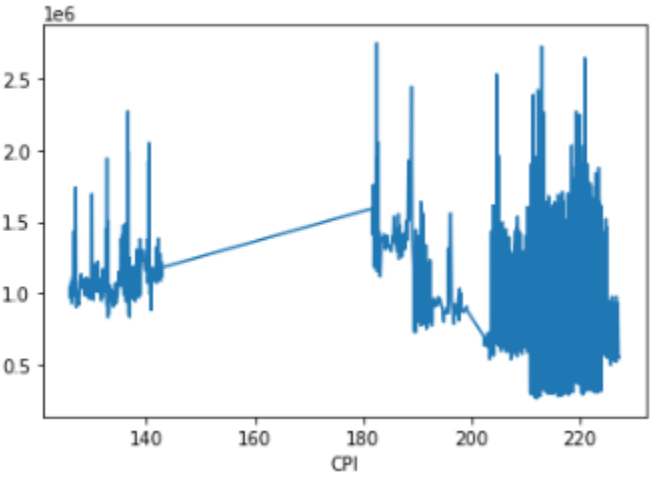
(h) Almost the same sales have happened each year.



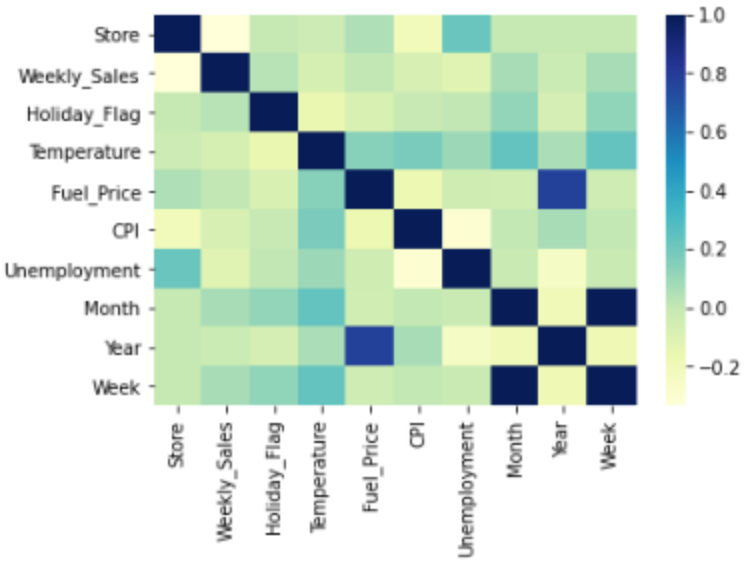
(i) The unemployment doesn't seem to have correlation with weekly sales. But, as the unemployment rate has gone beyond 10, sales have reduced.



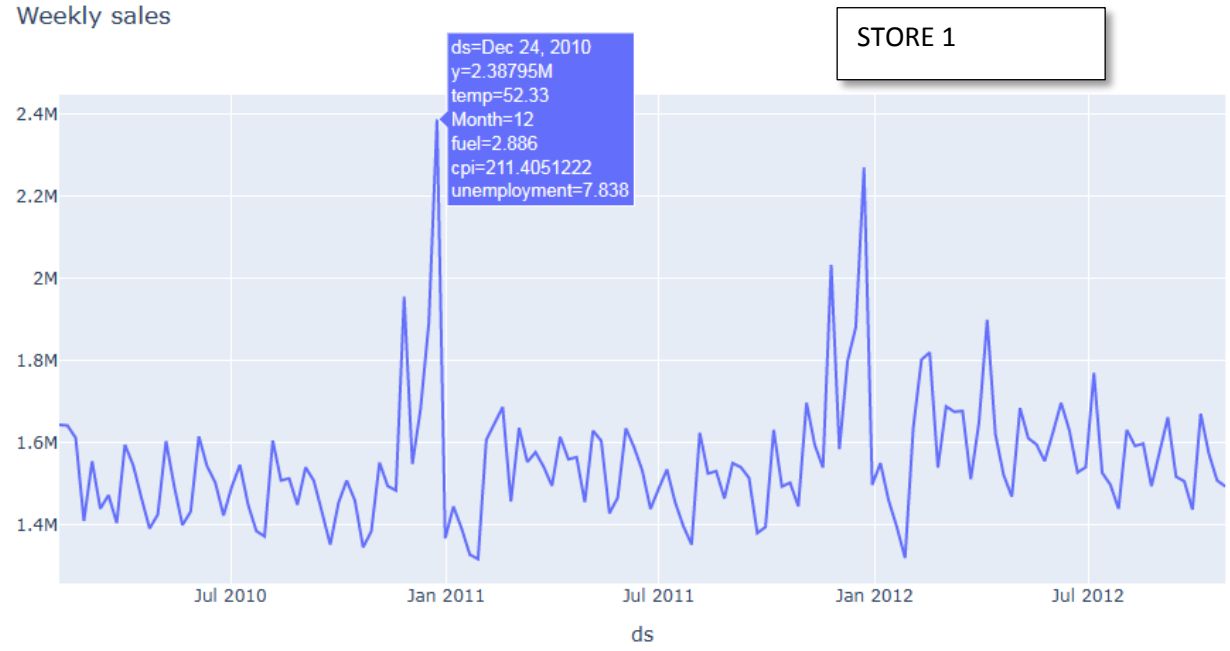
(j) High CPI is indicating High sales.



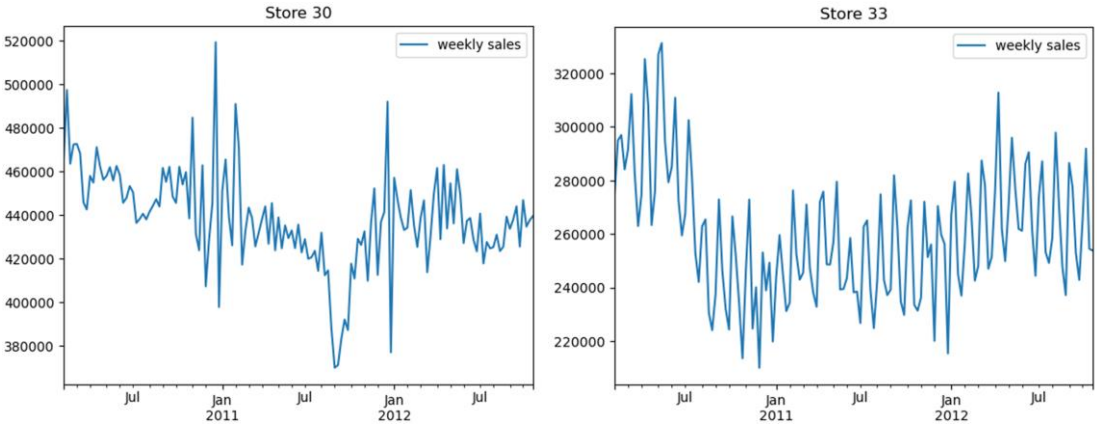
(k) We see that fuel price and year are correlated. So are month and week. (Problem of multicollinearity)



(I) Weekly sales for stores were analyzed for trend and seasonality.



There is seasonality observed for most of the stores. Trend is observed for a few. A few stores show no specific pattern like stores 33, 30 etc.



2. It is noted that date is an object column. Hence it is first changed to date time.

```
1 data['Date']=pd.to_datetime(data['Date'],dayfirst=True)
```

```
1 data.head()
```

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment
0	1	2010-02-05	1643690.90	0	42.31	2.572	211.096358	8.106
1	1	2010-02-12	1641957.44	1	38.51	2.548	211.242170	8.106
2	1	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106
3	1	2010-02-26	1409727.59	0	46.63	2.561	211.319643	8.106
4	1	2010-03-05	1554806.68	0	46.50	2.625	211.350143	8.106

3. Day, month, year has been extracted from date

	Store	Date	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Day	Month	Year
0	1	2010-02-05	1643690.90	0	42.31	2.572	211.096358	8.106	4	2	2010
1	1	2010-02-12	1641957.44	1	38.51	2.548	211.242170	8.106	4	2	2010
2	1	2010-02-19	1611968.17	0	39.93	2.514	211.289143	8.106	4	2	2010
3	1	2010-02-26	1409727.59	0	46.63	2.561	211.319643	8.106	4	2	2010
4	1	2010-03-05	1554806.68	0	46.50	2.625	211.350143	8.106	4	3	2010

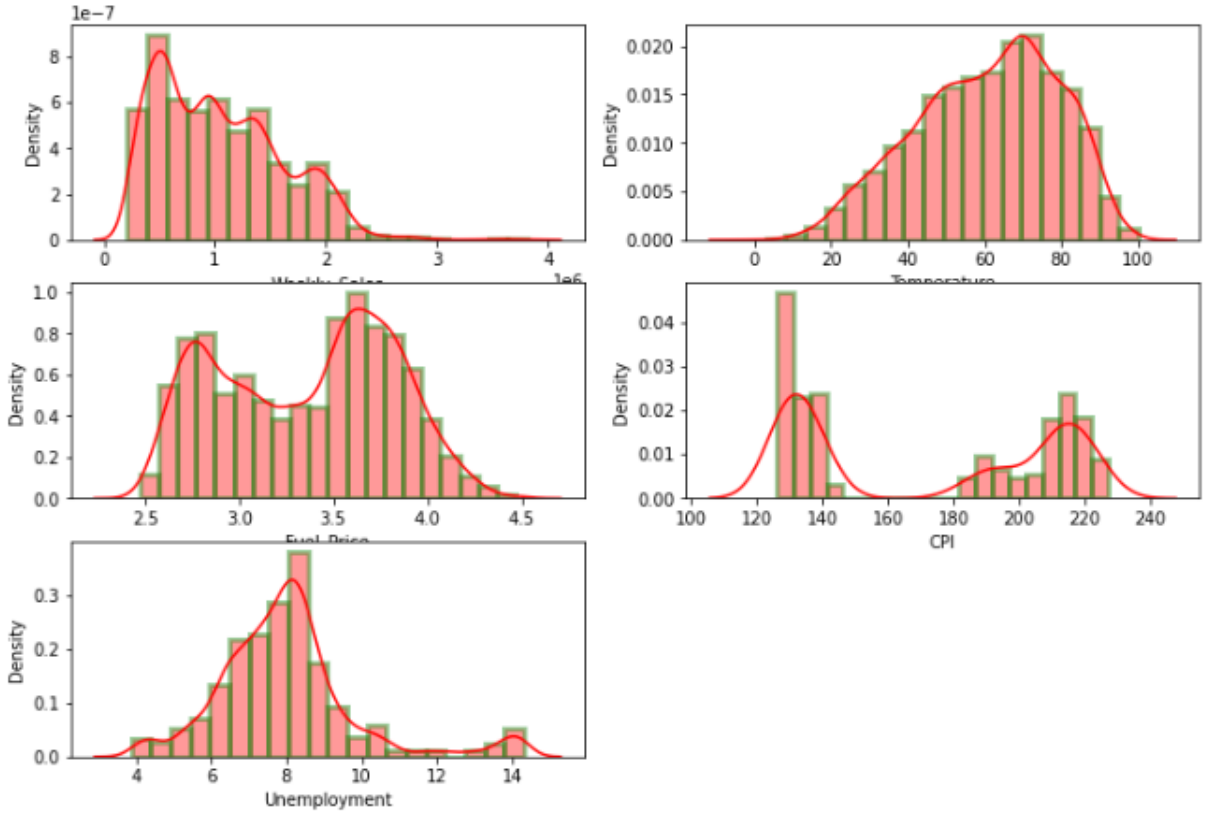
4. Since all data are given for Friday, the only unique value in weekday column is 4. Hence we will drop it instead talk about which week of the month it is.

	Store	Weekly_Sales	Holiday_Flag	Temperature	Fuel_Price	CPI	Unemployment	Month	Year	Week
0	1	1643690.90	0	42.31	2.572	211.096358	8.106	2	2010	5
1	1	1641957.44	1	38.51	2.548	211.242170	8.106	2	2010	6
2	1	1611968.17	0	39.93	2.514	211.289143	8.106	2	2010	7
3	1	1409727.59	0	46.63	2.561	211.319643	8.106	2	2010	8
4	1	1554806.68	0	46.50	2.625	211.350143	8.106	3	2010	9

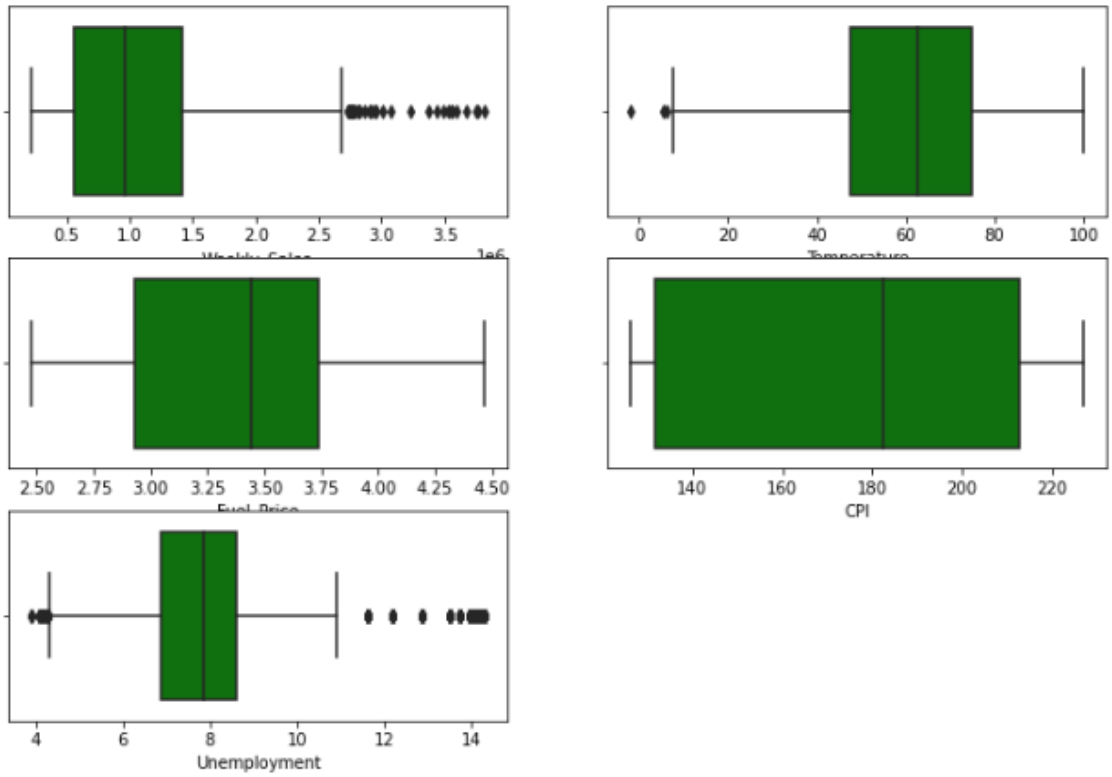
5. We filter out discrete and continuous variables.

```
Discrete variable : ['Store', 'Holiday_Flag', 'Month', 'Year', 'Week']
Continuous Variables : ['Weekly_Sales', 'Temperature', 'Fuel_Price', 'CPI', 'Unemployment']
```

6. Continuous variables are not all normally distributed. Highly skewed variables like cpi, fuelprice needs to be normalized and hence has to be normalized.



7. Independent Variables like unemployment and temperature have outliers which need to be dealt with either by removing or normalizing.



8. Since month feature is very well captured by week attribute, the month column is removed.

9. Categorical features are one hot encoded.

Store_7	Store_8	Store_9	Store_10	...	Week_48	Week_49	Week_50	Week_51	Week_52	Weekly_Sales	Temperature	Fuel_Price	CPI	Unemployment
0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1643690.90	42.31	2.572	211.096358	8.106
0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1641957.44	38.51	2.548	211.242170	8.106
0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1611968.17	39.93	2.514	211.289143	8.106
0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1409727.59	46.63	2.561	211.319643	8.106
0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	1554806.68	46.50	2.625	211.350143	8.106

10. Data has been transformed in 2 ways: (i) using box-cox (ii) removing outliers. Both of them were scaled using standard scaler and min max scaler.

11. Multicollinearity was handled using statsmodel.api.OLS and VIF. VIF did not seem to perform well on both data sets as so much of data was lost. In statsmodel case if $p<0.05$, then the multicollinearity scope is very less. This way number of features reduced to 83.

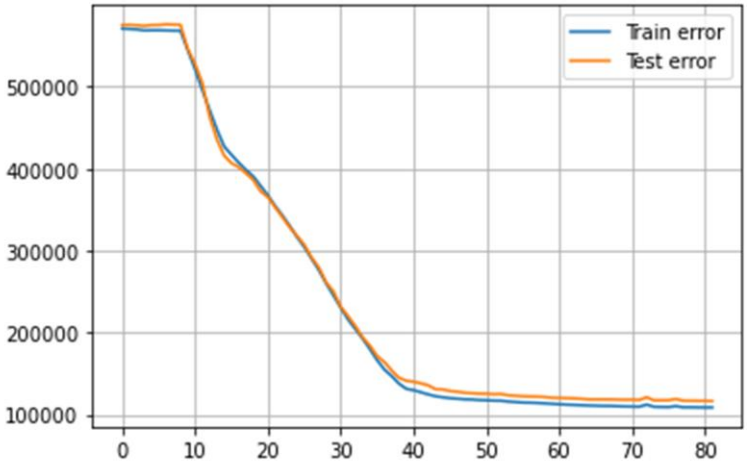
	Feature	p values
0	Week_35	0.041195
1	Week_6	0.028492
2	Store_40	0.026515
3	Week_20	0.011654
4	Fuel_Price	0.010973
...	...	...
78	Store_21	0.000000
79	Store_25	0.000000
80	Store_27	0.000000
81	Store_30	0.000000
82	const	0.000000

	Features	Variance Inflation
0	Fuel_Price	11.242389
1	Unemployment	12.039647
2	Temperature	16.288615
3	CPI	1142.111391
4	Week_14	1507.587730
...	...	...
101	Store_22	inf
102	Store_40	inf
103	Store_21	inf
104	Store_31	inf
105	Store_1	inf

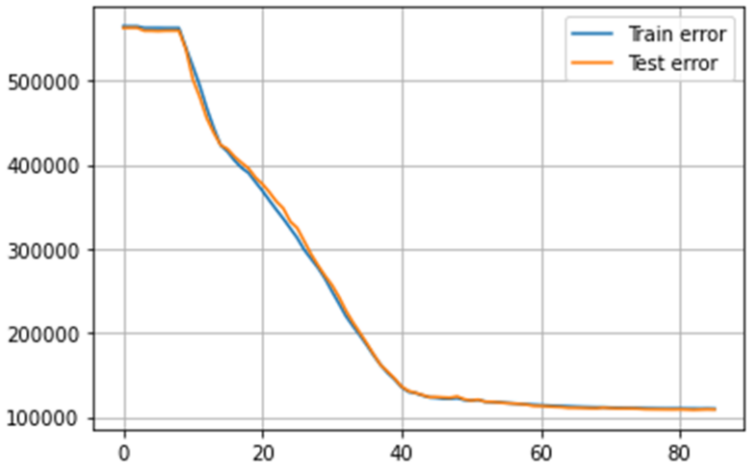
	Features	Variance Inflation
0	Unemployment	1.175196
1	CPI	1.586661
2	Temperature	3.552606
3	Fuel_Price	7.404788
4	Week_15	inf
...	...	...
56	Week_16	inf
57	Week_37	inf
58	Week_45	inf
59	Week_25	inf
60	Week_33	inf

	Features	Variance Inflation
0	Fuel_Price	1.078226
1	Unemployment	1.084436
2	Temperature	1.096695
3	CPI	1.159851

12. Using RFE, important features were alone selected. With this 70 features in outliers removed data and 76 features in box cox data were selected.



Outlier removed data



Box cox data

13. Dataframes with dates as index and weekly sales as a column, for each store seperately has been created and stored in a list.

```
[            weekly sales
 Date
 2010-02-05    1643690.90
 2010-02-12    1641957.44
 2010-02-19    1611968.17
 2010-02-26    1409727.59
 2010-03-05    1554806.68
 ...                  ...
 2012-09-28    1437059.26
 2012-10-05    1670785.97
 2012-10-12    1573072.81
 2012-10-19    1508068.77
 2012-10-26    1493659.74
              weekly sales
 Date
 2010-02-05    2136989.46
 2010-02-12    2137809.50
 2010-02-19    2124451.54
 2010-02-26    1865097.27
 2010-03-05    1991013.13
 ...                  ...
 2012-09-28    1746470.56
 2012-10-05    1998321.04
 2012-10-12    1900745.13
 2012-10-19    1847990.41
 2012-10-26    1834458.35
              weekly sales
 Date
 2010-02-05     461622.22
 2010-02-12     420728.96
 2010-02-19     421642.19
 2010-02-26     407204.86
 2010-03-05     415202.04
 ...                  ...
 2012-09-28     389813.02
 2012-10-05     443557.65
 2012-10-12     410804.39
 2012-10-19     424513.08
 2012-10-26     405432.70
```

14. Dataframes with sales, and also other factors like temperature, CPI, Unemployment, Month, Fuel price etc. are extracted for each store seperately and stored in a list.

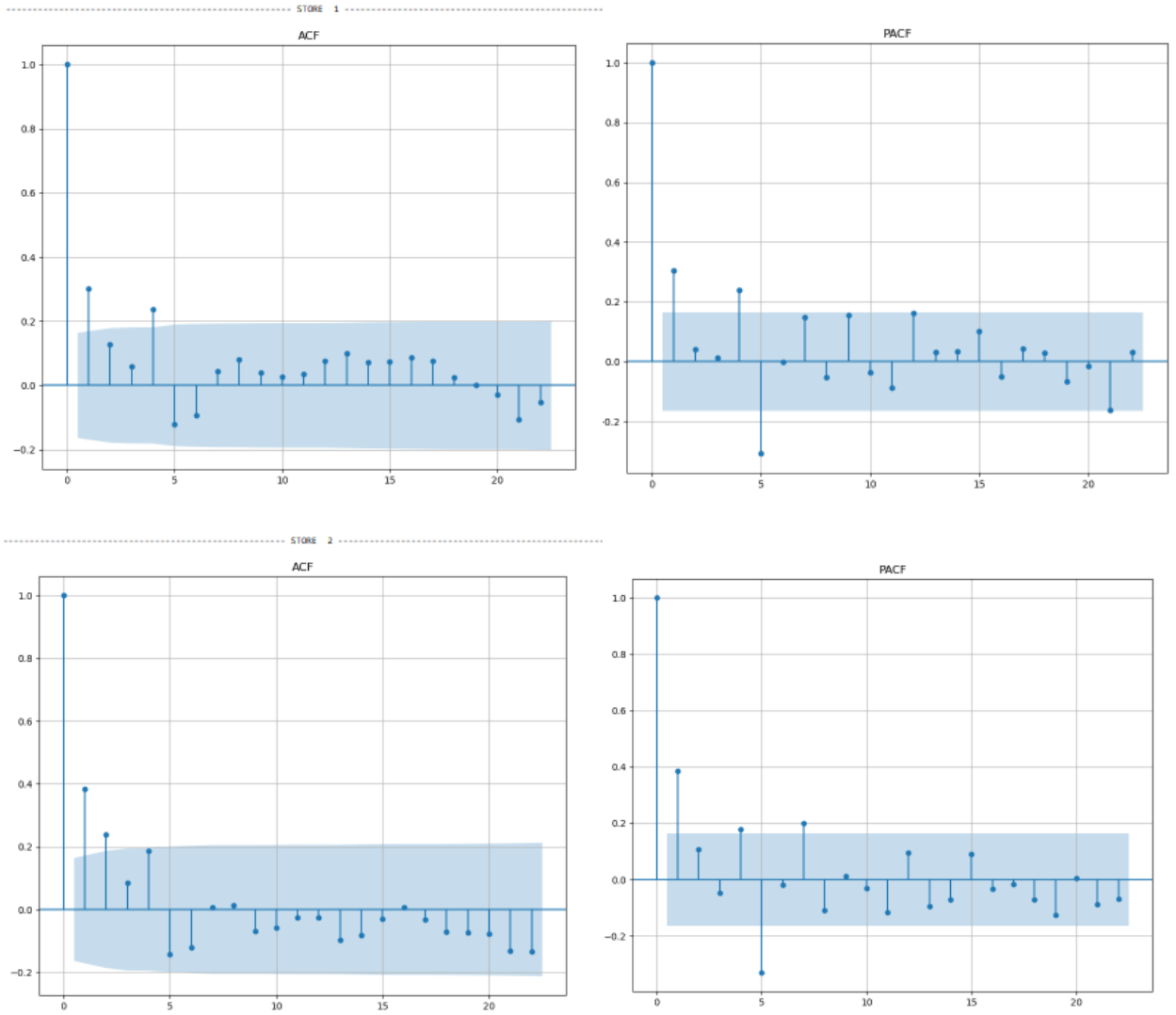
```
[           ds           y   temp   fuel         cpi  unemployment  Month
 0   2010-02-05  1643690.90  42.31  2.572  211.096358         8.106      2
 1   2010-02-12  1641957.44  38.51  2.548  211.242170         8.106      2
 2   2010-02-19  1611968.17  39.93  2.514  211.289143         8.106      2
 3   2010-02-26  1409727.59  46.63  2.561  211.319643         8.106      2
 4   2010-03-05  1554806.68  46.50  2.625  211.350143         8.106      3
 ..         ...         ...    ...    ...         ...           ...    ...
 138 2012-09-28  1437059.26  76.08  3.666  222.981658         6.908      9
 139 2012-10-05  1670785.97  68.55  3.617  223.181477         6.573     10
 140 2012-10-12  1573072.81  62.99  3.601  223.381296         6.573     10
 141 2012-10-19  1508068.77  67.97  3.594  223.425723         6.573     10
 142 2012-10-26  1493659.74  69.16  3.506  223.444251         6.573     10

            ds           y   temp   fuel         cpi  unemployment  Month
 0   2010-02-05  2136989.46  40.19  2.572  210.752605         8.324      2
 1   2010-02-12  2137809.50  38.49  2.548  210.897994         8.324      2
 2   2010-02-19  2124451.54  39.69  2.514  210.945160         8.324      2
 3   2010-02-26  1865097.27  46.10  2.561  210.975957         8.324      2
 4   2010-03-05  1991013.13  47.17  2.625  211.006754         8.324      3
 ..         ...         ...    ...    ...         ...           ...    ...
 138 2012-09-28  1746470.56  79.45  3.666  222.616433         6.565      9
 139 2012-10-05  1998321.04  70.27  3.617  222.815930         6.170     10
 140 2012-10-12  1900745.13  60.97  3.601  223.015426         6.170     10
 141 2012-10-19  1847990.41  68.08  3.594  223.059808         6.170     10
 142 2012-10-26  1834458.35  69.79  3.506  223.078337         6.170     10
```

15. It is observed that except for 5 to 6 stores, the weekly sales is observed to be stationary for the rest of them.

```
---------------------------------------------------------- STORE  1 ----
ADF Statistics: -5.102186145192288
p- value: 1.3877788330759434e-05
Data is stationary.
---------------------------------------------------------- STORE  2 ----
ADF Statistics: -3.708862572618916
p- value: 0.003990207089066256
Data is stationary.
---------------------------------------------------------- STORE  3 ----
ADF Statistics: -2.9638677455113216
p- value: 0.038409261798312666
Data is stationary.
---------------------------------------------------------- STORE  4 ----
ADF Statistics: -2.8793819840147084
p- value: 0.047798662236698805
Data is stationary.
---------------------------------------------------------- STORE  5 ----
ADF Statistics: -4.310974424060914
p- value: 0.00042517056141923293
```

16. ACF and PACF plots of different stores' weekly sales were analyzed.



17. The weekly sales data was decomposed into its seasonal components, to understand trend, residual and seasonality.

CHOOSING THE ALGORITHM FOR THE PROJECT

The given data which has been transformed into data with outliers removed and one with box cox applied to reduce the outliers, were analyzed mainly in 2 ways: Regression and Time series.

The summary of the observations are as follows:

1. Regression models:

   (a) Linear Regression:

      (i) Results with outliers removed data

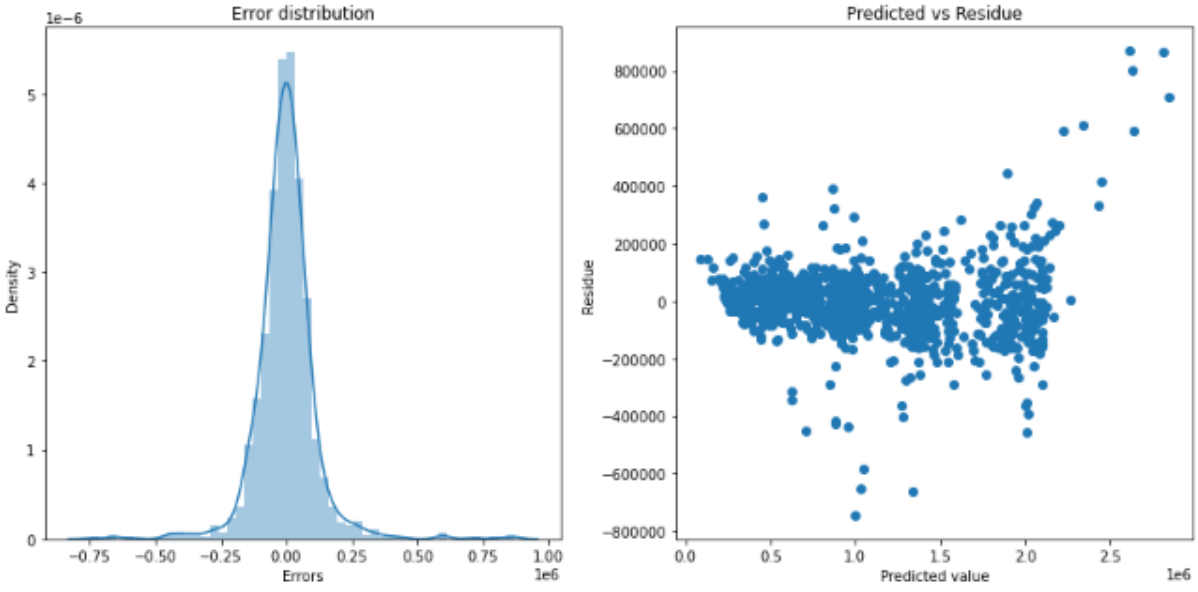
```
------------------------------ TRAINING SET ------------------------------

RMSE: 109841.41210025577
MAE: 70842.20263167468
R2 Score: 0.9630817155627436
Adjusted R square: 0.9625590469782676

------------------------------ TESTING SET ------------------------------

RMSE: 117989.08103715842
MAE: 73863.37886720811
R2 Score: 0.957803950156088
Adjusted R square: 0.9582253127275254

------------------------------ RESIDUAL PLOTS ------------------------------
```



      (ii) Results with box cox applied data

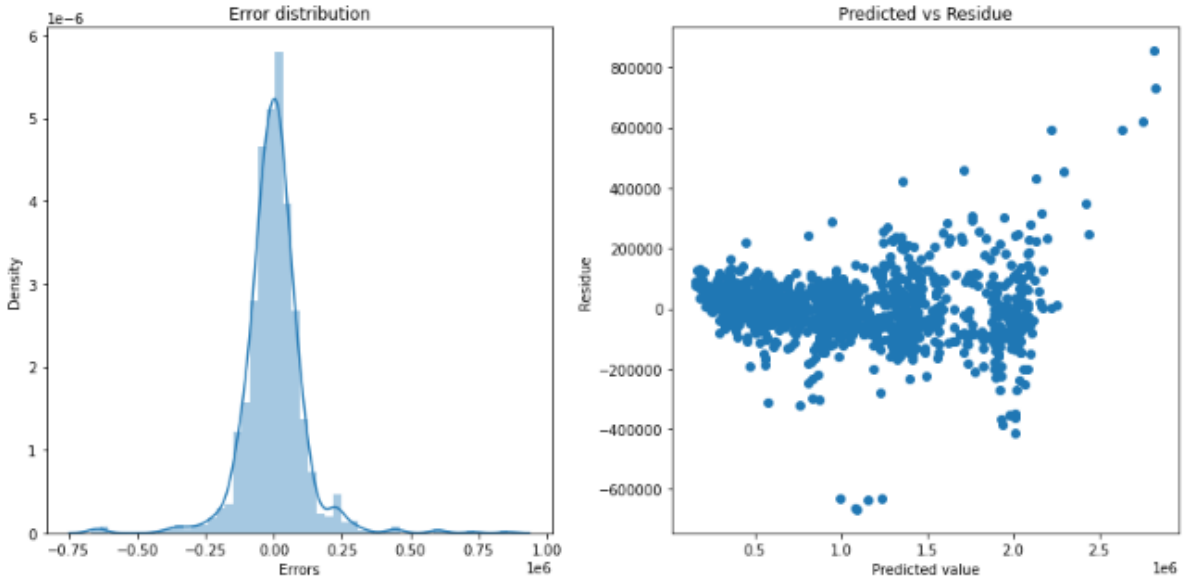
```
------------------------------ TRAINING SET ------------------------------

RMSE: 110623.21240470343
MAE: 71312.63424528798
R2 Score: 0.9616442812183323
Adjusted R square: 0.9611253954178199

------------------------------ TESTING SET ------------------------------

RMSE: 109661.09665615612
MAE: 70782.44587773865
R2 Score: 0.9619519603935076
Adjusted R square: 0.9631214946531416

------------------------------ RESIDUAL PLOTS ------------------------------
```



(b) Ridge regression : (i) Results with outliers removed data

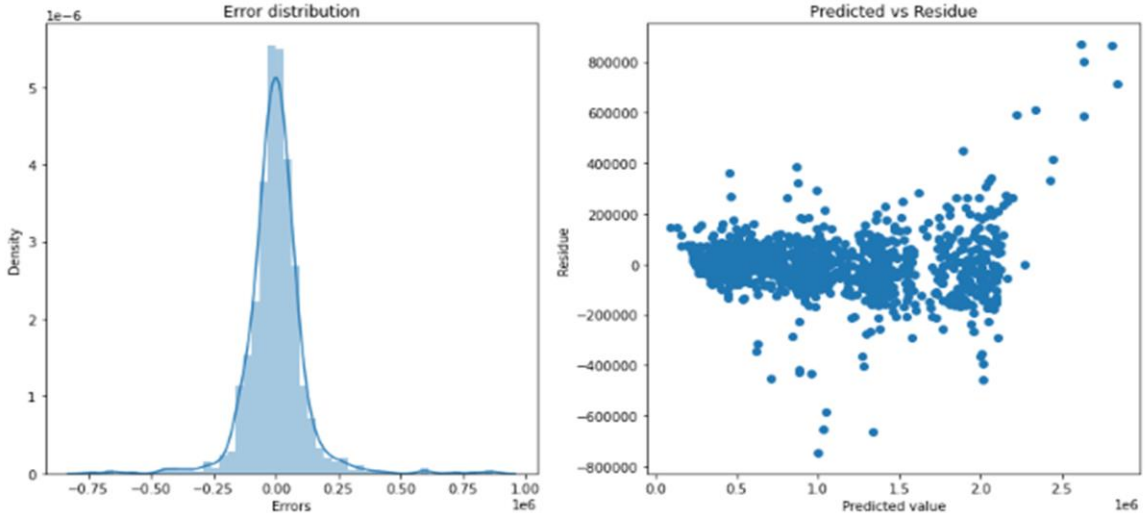
```
------------------------------ TRAINING SET ------------------------------

RMSE: 109835.5630581741
MAE: 70833.86521453483
R2 Score: 0.9630856472465381
Adjusted R square: 0.9625590469782676

------------------------------ TESTING SET ------------------------------

RMSE: 117967.17533569141
MAE: 73837.8775579688
R2 Score: 0.9578196168306227
Adjusted R square: 0.9582253127275254

------------------------------ RESIDUAL PLOTS ------------------------------
```



(ii) Results with box cox applied data

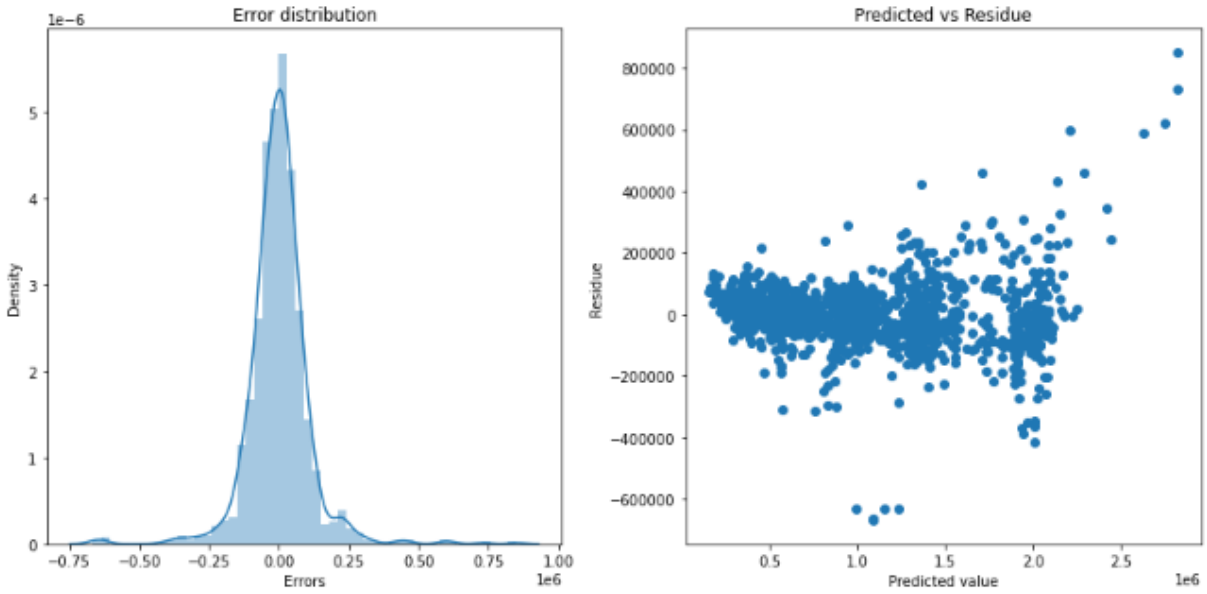
```
 ------------------------------ TRAINING SET ------------------------------

RMSE: 110577.14069203357
MAE: 71160.24292467578
R2 Score: 0.9616762228992497
Adjusted R square: 0.9611253954178199

 ------------------------------ TESTING SET ------------------------------

RMSE: 109446.11891257318
MAE: 70594.5665507084
R2 Score: 0.962100991626886
Adjusted R square: 0.9631214946531416

 ------------------------------ RESIDUAL PLOTS ------------------------------
```



(c) Lasso Regression : (i) Results with outliers removed data

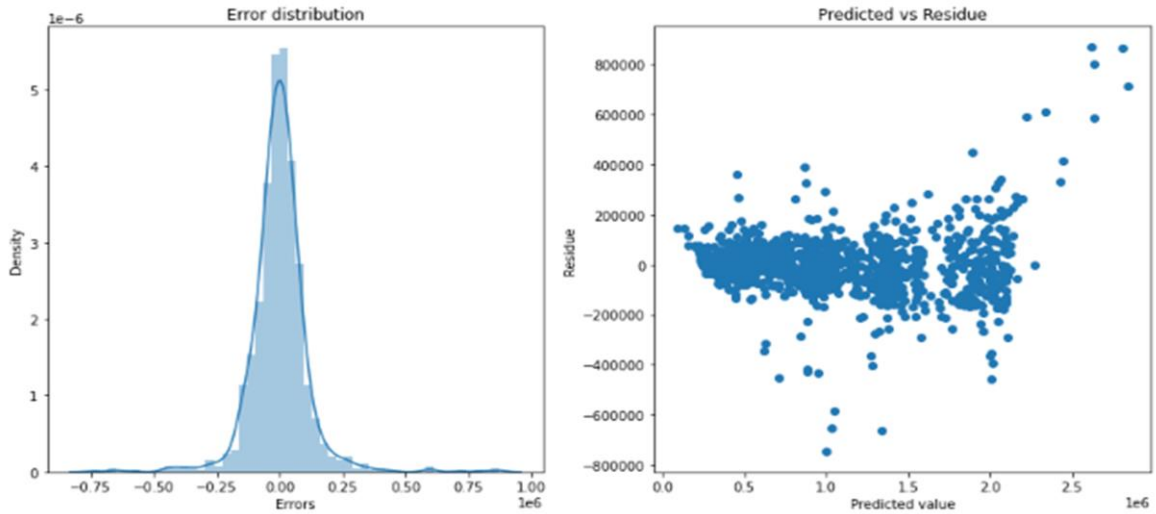
```
 ------------------------------ TRAINING SET ------------------------------

RMSE: 109834.79746872085
MAE: 70861.62524168406
R2 Score: 0.963086161854737
Adjusted R square: 0.9625590469782676

 ------------------------------ TESTING SET ------------------------------

RMSE: 117964.84774785339
MAE: 73874.51705386846
R2 Score: 0.9578212813203931
Adjusted R square: 0.9582253127275254

 ------------------------------ RESIDUAL PLOTS ------------------------------
```



(ii) Results with box cox applied data

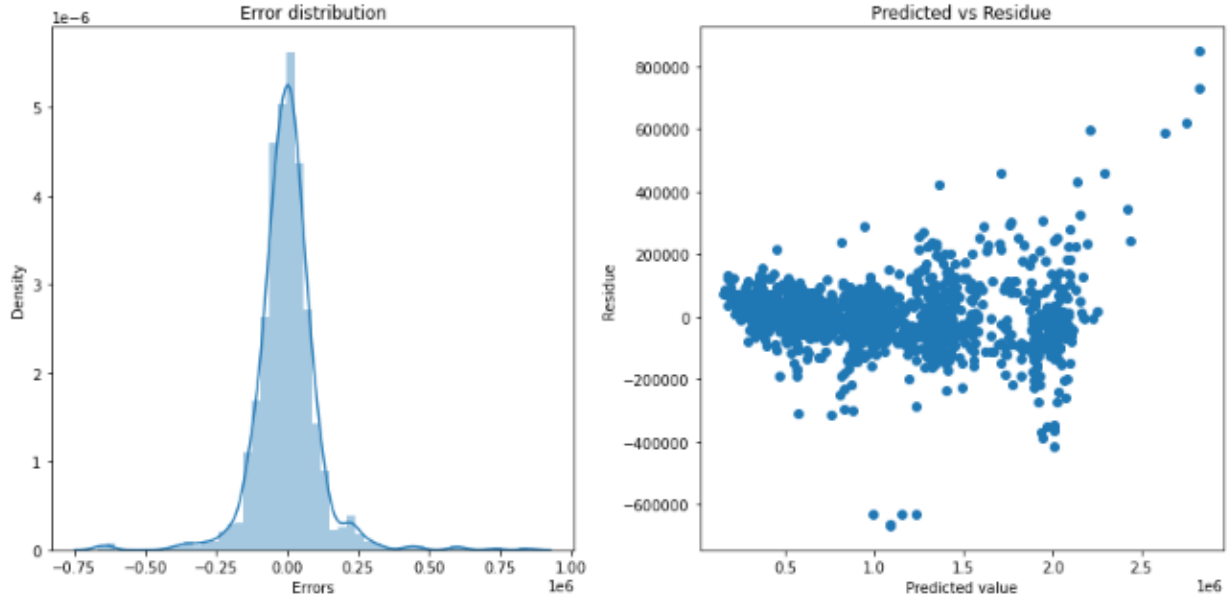
```
------------------------------ TRAINING SET ------------------------------

RMSE: 110576.37489403572
MAE: 71188.82388920887
R2 Score: 0.9616767537172826
Adjusted R square: 0.9611253954178199

------------------------------ TESTING SET ------------------------------

RMSE: 109435.37590884525
MAE: 70618.3662481085
R2 Score: 0.9621084314376316
Adjusted R square: 0.9631214946531416

------------------------------ RESIDUAL PLOTS ------------------------------
```



(d) Elastic net regression : (i) Results with outliers removed data

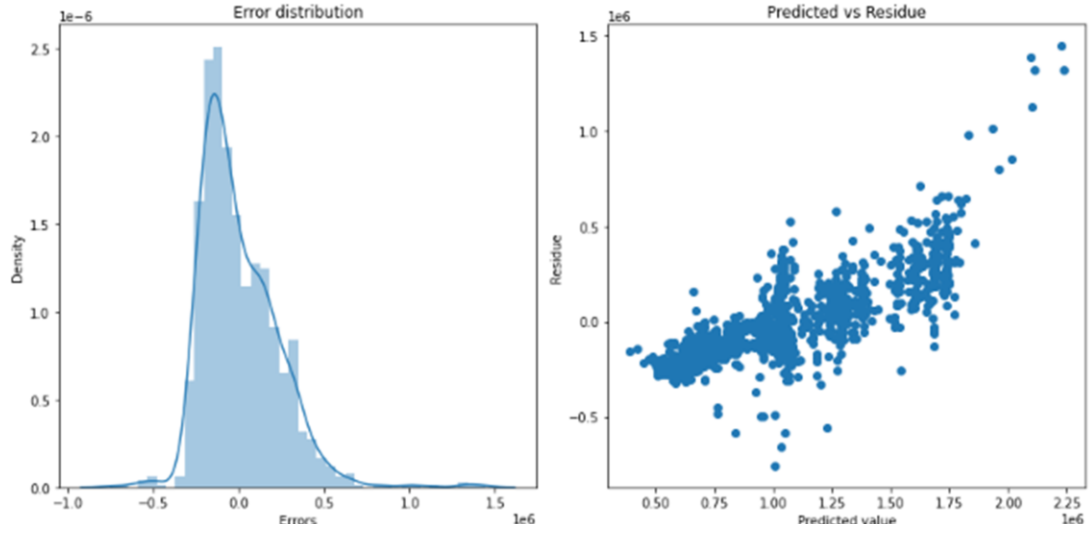
```
------------------------------ TRAINING SET ------------------------------

RMSE: 222295.61567505382
MAE: 177118.6052547865
R2 Score: 0.8487932805969619
Adjusted R square: 0.9625590469782676

------------------------------ TESTING SET ------------------------------

RMSE: 228450.56309676138
MAE: 175835.26884443642
R2 Score: 0.8418123314847639
Adjusted R square: 0.9582253127275254

------------------------------ RESIDUAL PLOTS ------------------------------
```



(ii) Results with box cox applied data

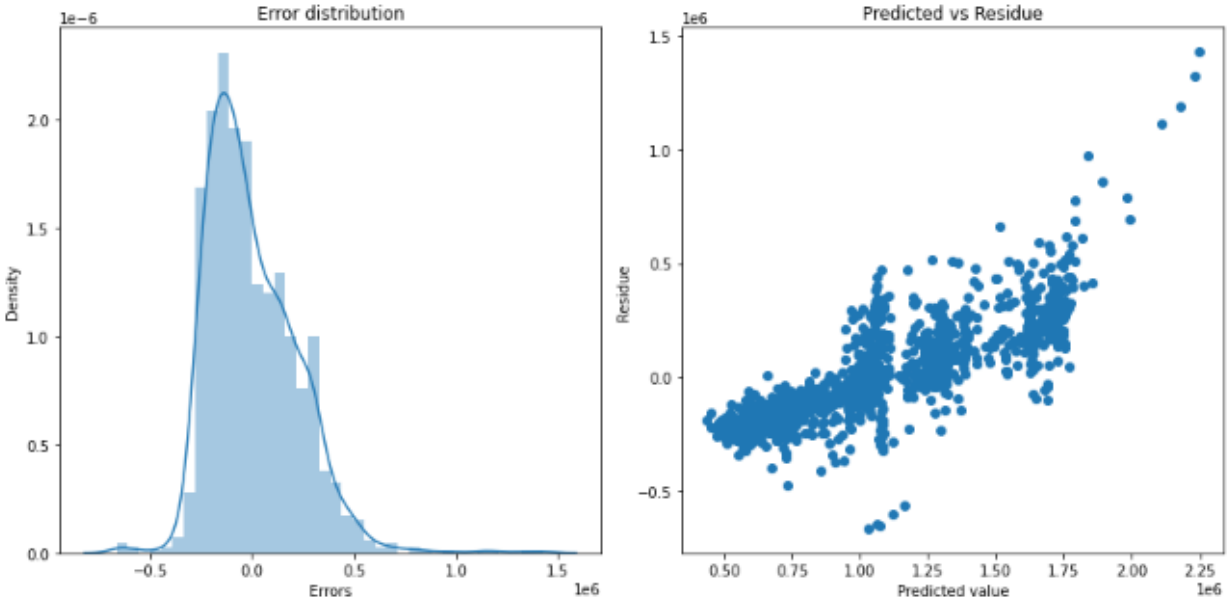
```
 ------------------------------ TRAINING SET ------------------------------

RMSE: 220732.67502886403
MAE: 174882.27619802198
R2 Score: 0.8472888150583748
Adjusted R square: 0.9611253954178199

 ------------------------------ TESTING SET ------------------------------

RMSE: 220414.9011032499
MAE: 173482.85070618085
R2 Score: 0.8462875592094325
Adjusted R square: 0.9631214946531416

 ------------------------------ RESIDUAL PLOTS ------------------------------
```



(e) Polynomial regression : (i) Results with outliers removed data

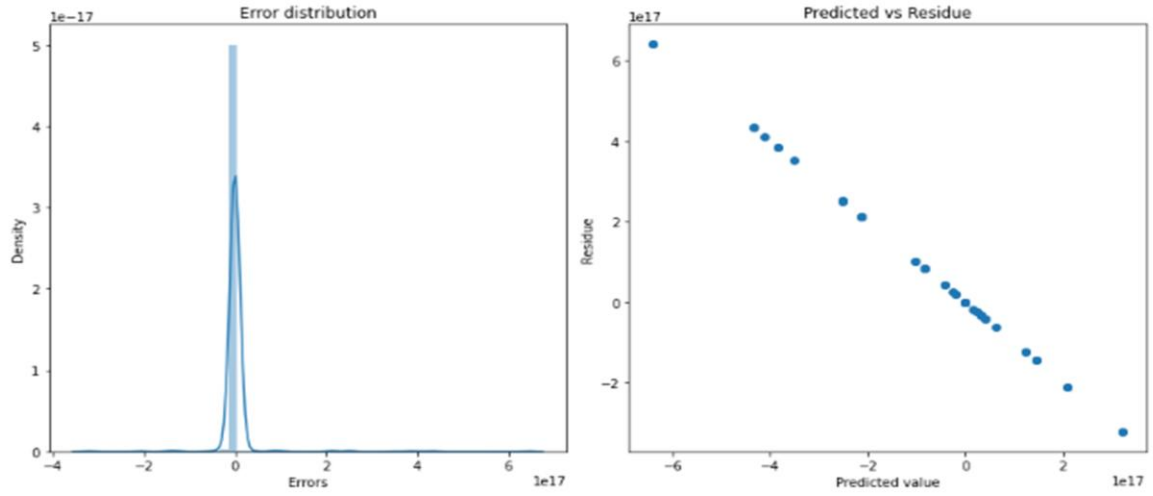
```
 ------------------------------ TRAINING SET ------------------------------

RMSE: 49713.54671835281
MAE: 33482.83732563025
R2 Score: 0.9924376148359884
Adjusted R square: 0.9900329482113012

 ------------------------------ TESTING SET ------------------------------

RMSE: 4.6384815585947736e+16
MAE: 6498189505676537.0
R2 Score: -6.521384816843231e+21
Adjusted R square: 0.9874356188841401

 ------------------------------ RESIDUAL PLOTS ------------------------------
```



(ii) Results with box cox applied data

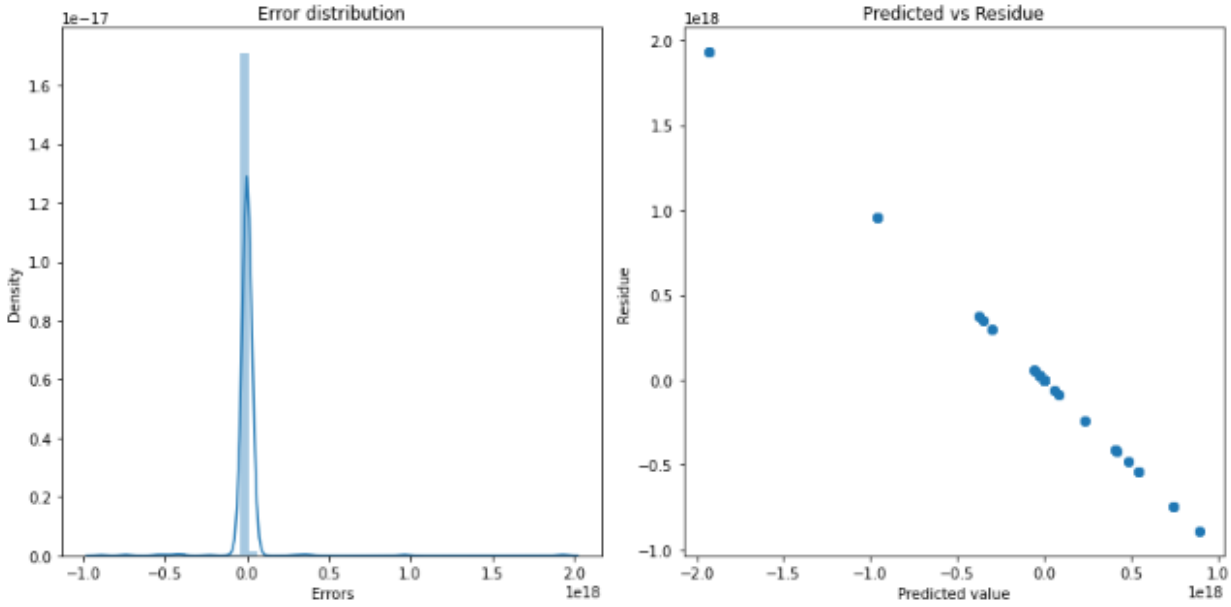
```
------------------------------ TRAINING SET ------------------------------

RMSE: 49889.25651948188
MAE: 34870.07722416473
R2 Score: 0.9921989752137823
Adjusted R square: 0.9902894398973756

------------------------------ TESTING SET -------------------------------

RMSE: 1.2562638687551758e+17
MAE: 1.5955726401733496e+16
R2 Score: -4.993313736001955e+22
Adjusted R square: 0.988051452106637

------------------------------ RESIDUAL PLOTS -----------------------------
```



With respect to regression models, the summary of the results are as follows:

(i) With outliers removed data:

	Model	RMSE_train	RMSE_test	MAE_train	MAE_test	R2_score_train	R2_score_test	Adjusted_R2_score_train	Adjusted_R2_score_test
0	LinearRegression()	109841.412100	1.179891e+05	70842.202632	7.386338e+04	0.963082	9.578040e-01	0.962559	0.958225
1	Lasso()	109834.797469	1.179648e+05	70861.625242	7.387452e+04	0.963086	9.578213e-01	0.962559	0.958225
2	Ridge()	109835.563058	1.179672e+05	70833.865215	7.383788e+04	0.963086	9.578196e-01	0.962559	0.958225
3	ElasticNet()	222295.615675	2.284506e+05	177118.605255	1.758353e+05	0.848793	8.418123e-01	0.962559	0.958225
4	Polynomial Regression()	49713.546718	4.638482e+16	33482.837326	6.498190e+15	0.992438	-6.521385e+21	0.990033	0.987436

(ii) With box-cox applied data:

	Model	RMSE_train	RMSE_test	MAE_train	MAE_test	R2_score_train	R2_score_test	Adjusted_R2_score_train	Adjusted_R2_score_test
0	LinearRegression()	110623.212405	1.096611e+05	71312.634245	7.078245e+04	0.961644	9.619520e-01	0.961125	0.963121
1	Lasso()	110576.374894	1.094354e+05	71188.823889	7.061837e+04	0.961677	9.621084e-01	0.961125	0.963121
2	Ridge()	110577.140692	1.094461e+05	71160.242925	7.059457e+04	0.961676	9.621010e-01	0.961125	0.963121
3	ElasticNet()	220732.675029	2.204149e+05	174882.276198	1.734829e+05	0.847289	8.462876e-01	0.961125	0.963121
4	Polynomial Regression()	49889.256519	1.256264e+17	34870.077224	1.595573e+16	0.992199	-4.993314e+22	0.990289	0.988051

Comparing the RMSE and R2 scores of train and test models of both entries, simple linear regression, Lasso and Ridge regression models' metrics are comparable in both cases. Errors in these cases were normally distributed and heteroskedacticity is less. Polynomial regression has over fit the data. There is a problem of under fitting with Elastic Net.

The data with outliers removed is comparatively a little better than box cox applied one.

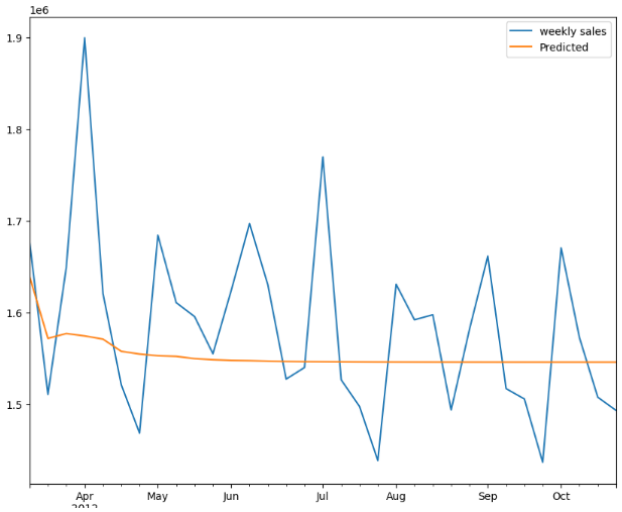
2. Time series analysis using the weekly sales.

Here ARIMA, SARIMAX, Univariate FB Prophet, Multivariate FB Prophet were used.

(a) ARIMA model:

The ARIMA model performed very bad on the data points and hence can't be used for prediction.

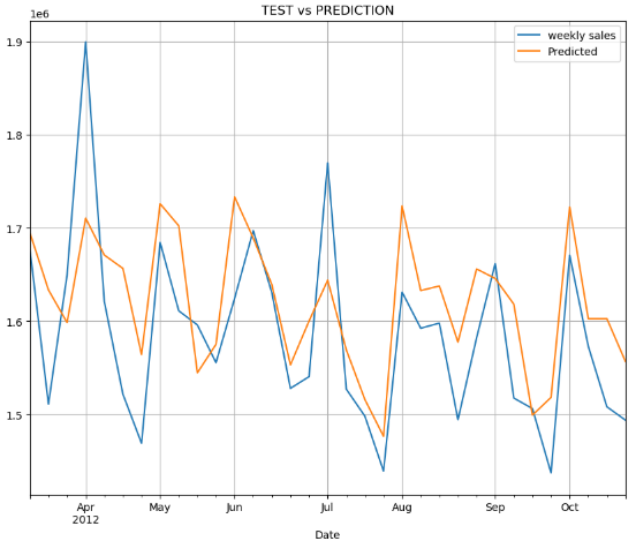
	order	rmse
0	(0, 0, 0)	1.039919e+05
1	(0, 0, 1)	1.032029e+05
2	(0, 0, 2)	1.026412e+05
3	(0, 0, 3)	1.025579e+05
4	(0, 0, 4)	9.952155e+04
..	...	...
92	(7, 0, 4)	1.115793e+05
93	(7, 1, 0)	1.589584e+06
94	(7, 1, 1)	1.589550e+06
95	(7, 1, 2)	1.589491e+06
96	(7, 2, 0)	1.588576e+06

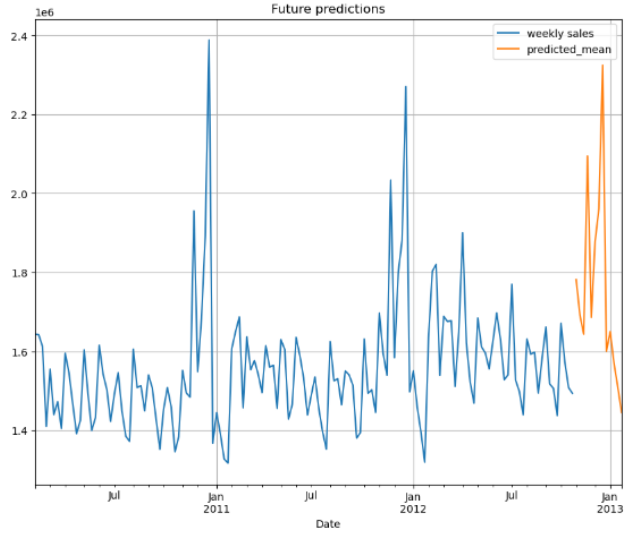


(b) SARIMAX model:

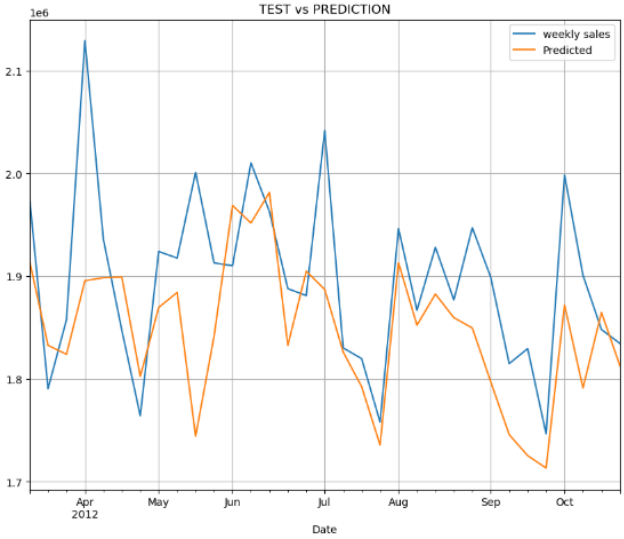
The SARIMAX model performed better than the ARIMA model, but for few stores the extremes peaks were not modeled properly. The predictions for few of them are shown below.

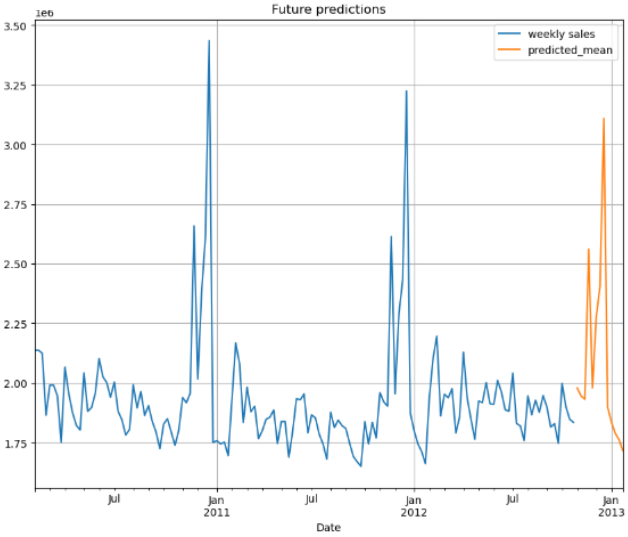
STORE 1



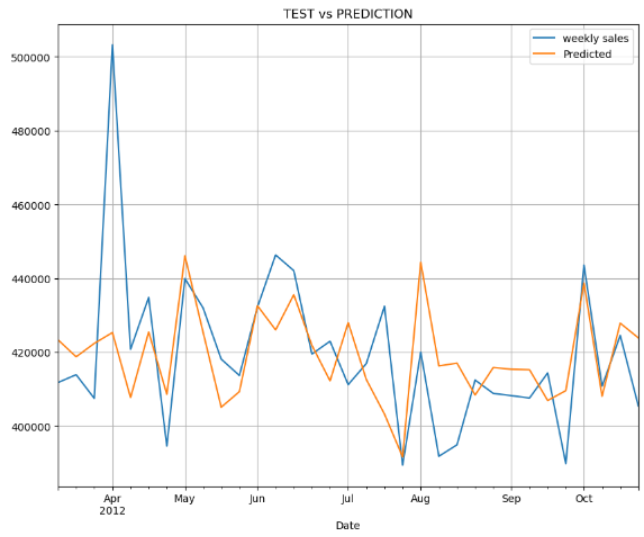


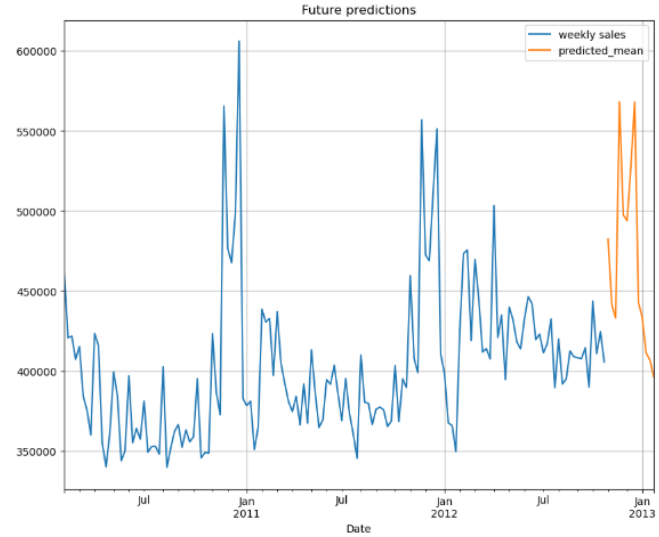
STORE 2





STORE 3

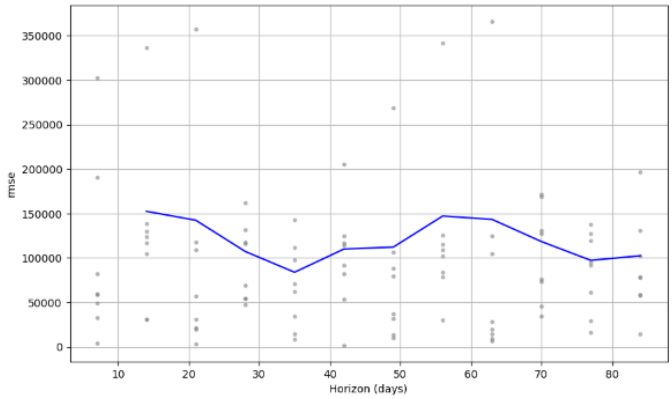


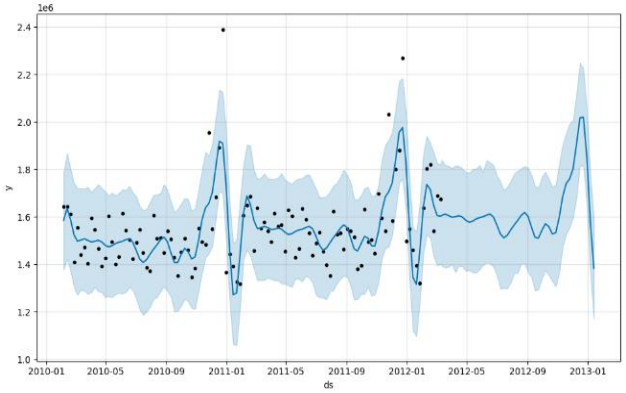


(c) Univariate FB Prophet:

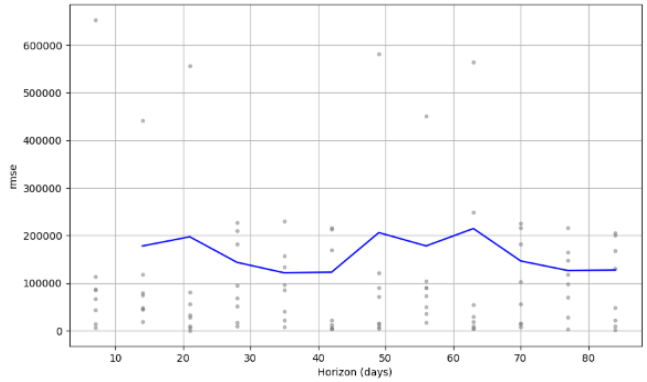
The FB Prophet using weekly sales data alone was able to predict the values well within a range though not the actual value.

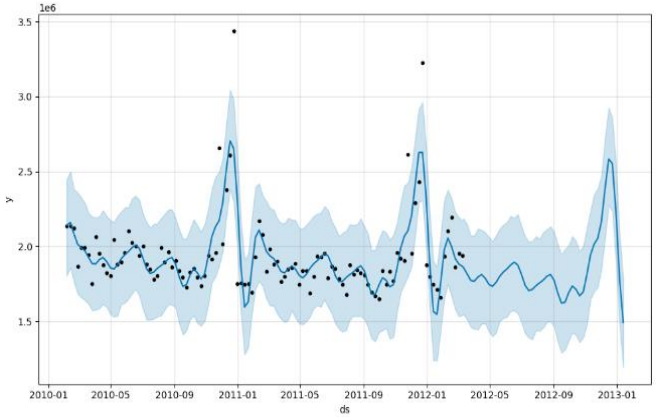
STORE 1



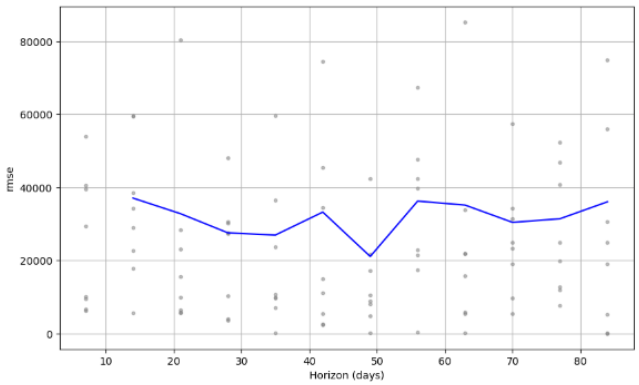


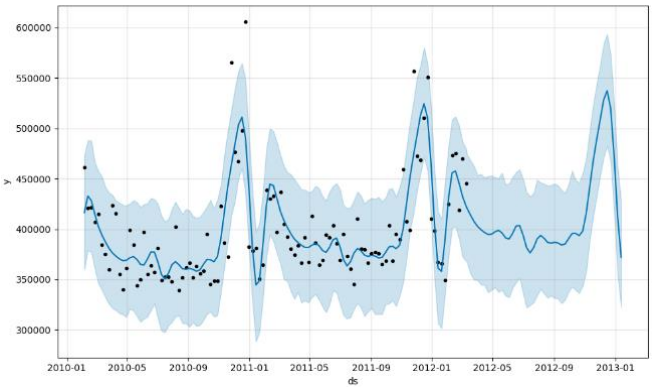
STORE 2





STORE 3

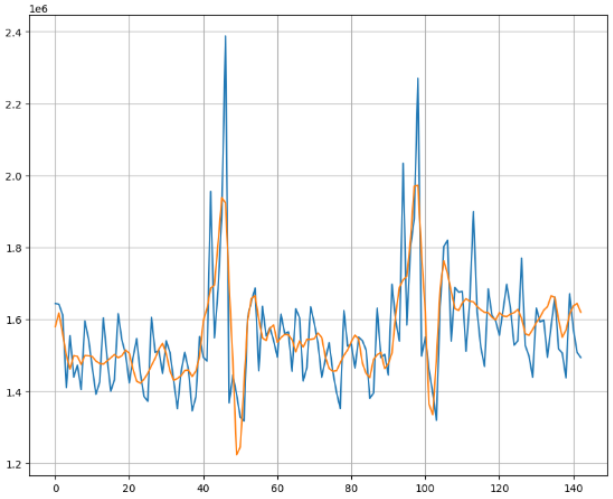


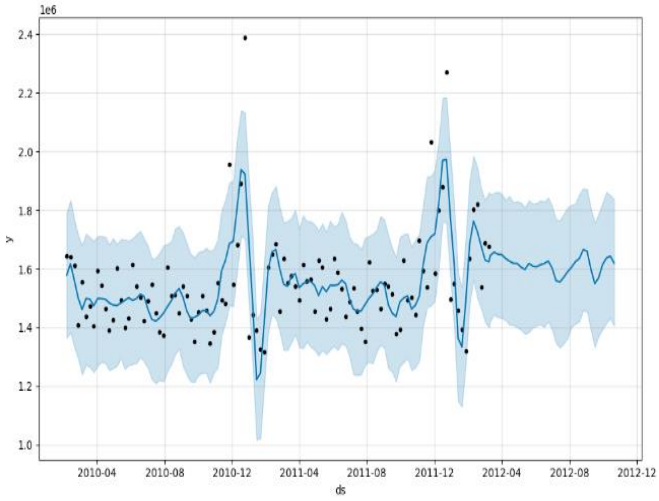


(d) Multivariate FB Prophet:

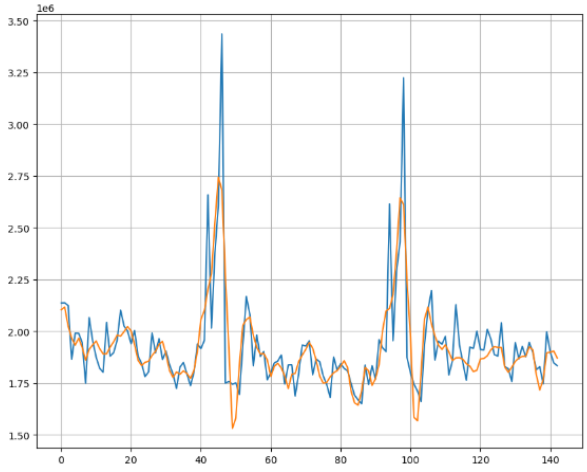
The model is able to grab a few more points which were originally in outlier region by Univariate FB Prophet, but can still improve may be after tuning the other input parameters better.

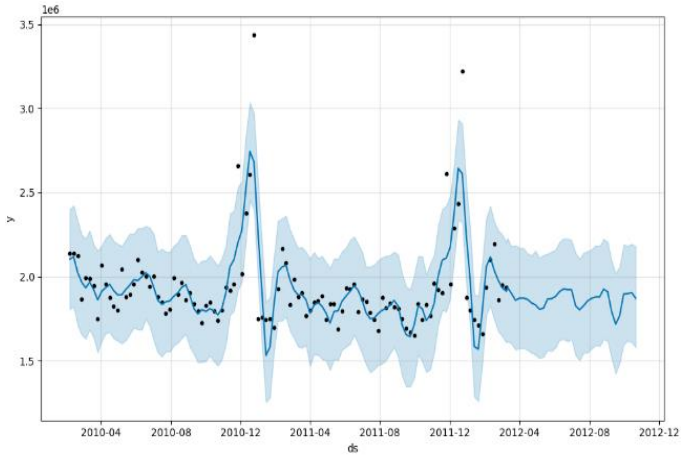
STORE 1



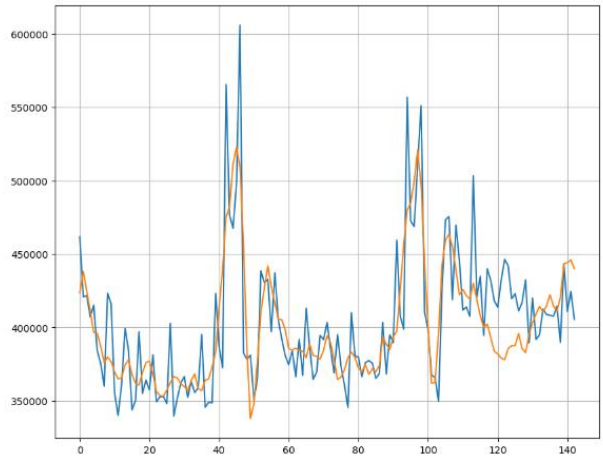


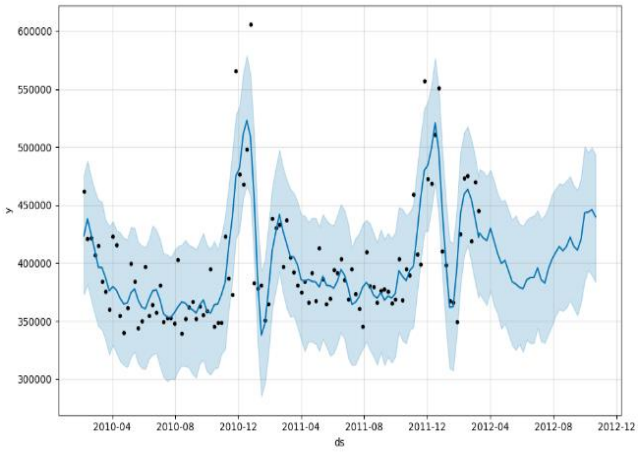
STORE 2





STORE 3





It is observed that out of the 4 models, multivariate FB Prophet seems to perform better when compared to the other models keeping in mind the RMSE and MAPE scores.

ASSUMPTIONS

1. Linear regression (Ridge and Lasso):
   (a) Independence of observations (No multicollinearity)
   (b) Linear relationship
   (c) Normality of residuals
   (d) Homoscedasticity
2. Time Series:
   (a) Future trends hold similar pattern as that of historical data

MODEL EVALUATION AND TECHNIQUE

The model was evaluated using:

(a) RMSE score
(b) R2 score, adjusted R2 scores
(c) Plots to visualize the prediction/error

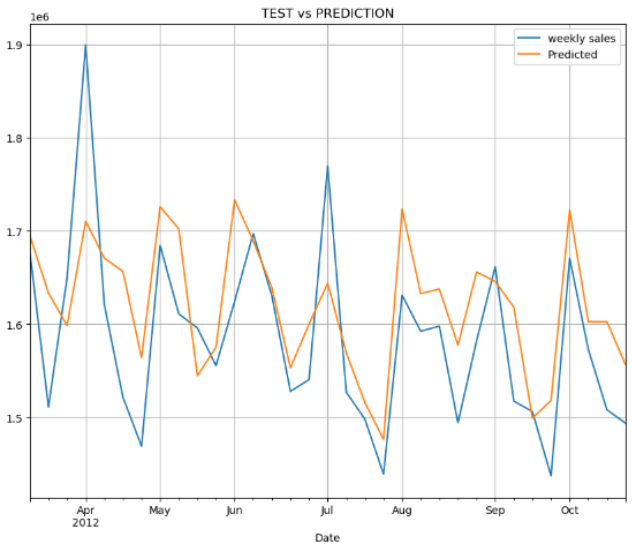
Outlier removed data:

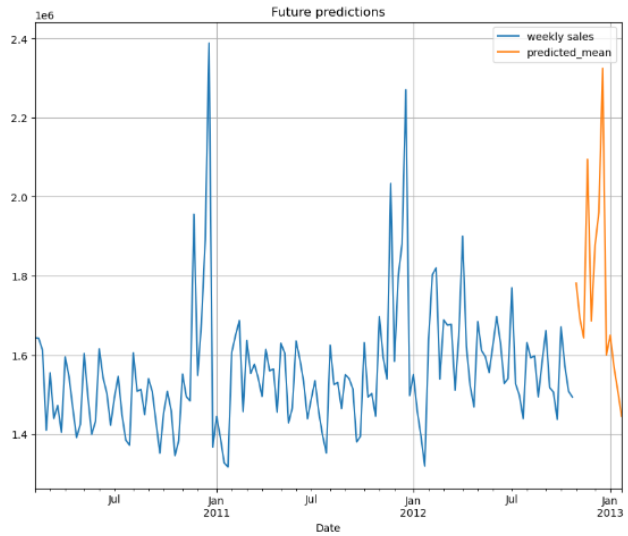
	Model	RMSE_train	RMSE_test	MAE_train	MAE_test	R2_score_train	R2_score_test	Adjusted_R2_score_train	Adjusted_R2_score_test
0	LinearRegression()	109841.412100	1.179891e+05	70842.202632	7.386338e+04	0.963082	9.578040e-01	0.962559	0.958225
1	Lasso()	109834.797469	1.179648e+05	70861.625242	7.387452e+04	0.963086	9.578213e-01	0.962559	0.958225
2	Ridge()	109835.563058	1.179672e+05	70833.865215	7.383788e+04	0.963086	9.578196e-01	0.962559	0.958225
3	ElasticNet()	222295.615675	2.284506e+05	177118.605255	1.758353e+05	0.848793	8.418123e-01	0.962559	0.958225
4	Polynomial Regression()	49713.546718	4.638482e+16	33482.837326	6.498190e+15	0.992438	-6.521385e+21	0.990033	0.987436

Box cox data

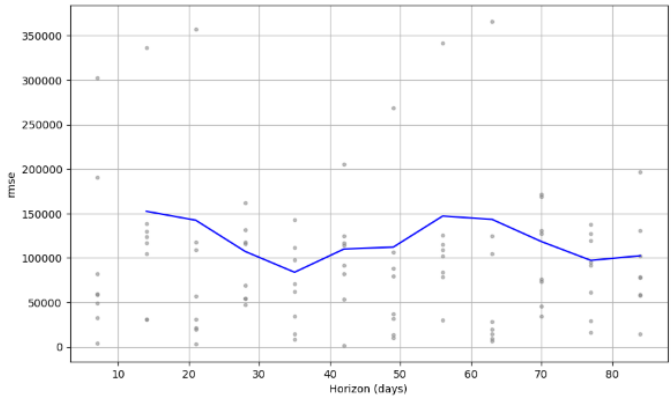
	Model	RMSE_train	RMSE_test	MAE_train	MAE_test	R2_score_train	R2_score_test	Adjusted_R2_score_train	Adjusted_R2_score_test
0	LinearRegression()	110623.212405	1.096611e+05	71312.634245	7.078245e+04	0.961644	9.619520e-01	0.961125	0.963121
1	Lasso()	110576.374894	1.094354e+05	71188.823889	7.061837e+04	0.961677	9.621084e-01	0.961125	0.963121
2	Ridge()	110577.140692	1.094461e+05	71160.242925	7.059457e+04	0.961676	9.621010e-01	0.961125	0.963121
3	ElasticNet()	220732.675029	2.204149e+05	174882.276198	1.734829e+05	0.847289	8.462876e-01	0.961125	0.963121
4	Polynomial Regression()	49889.256519	1.256264e+17	34870.077224	1.595573e+16	0.992199	-4.993314e+22	0.990289	0.988051

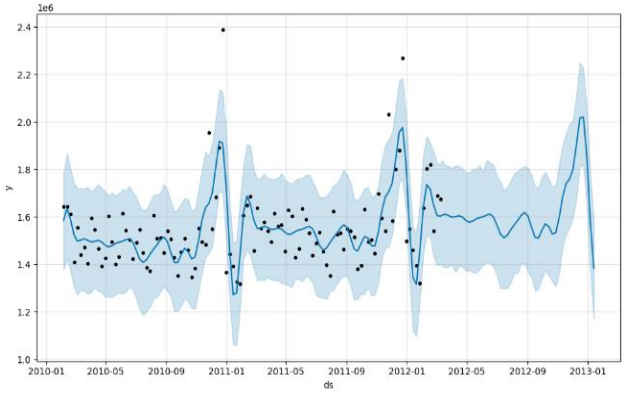
Time series – SARIMAX (STORE 1)





Time series – FB Prophet (Store 1)



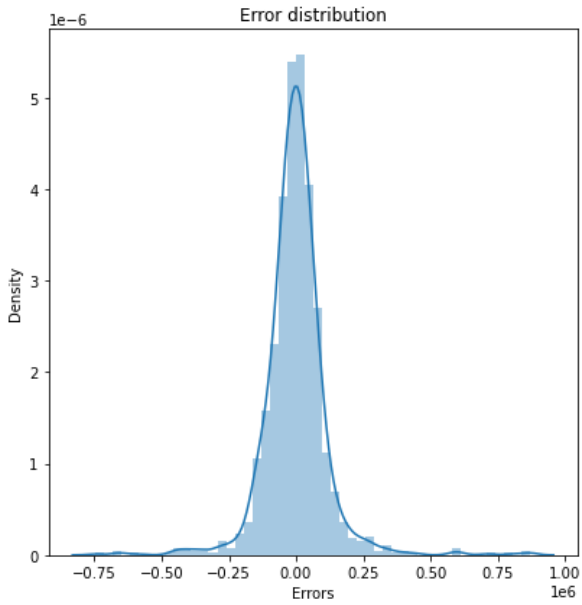


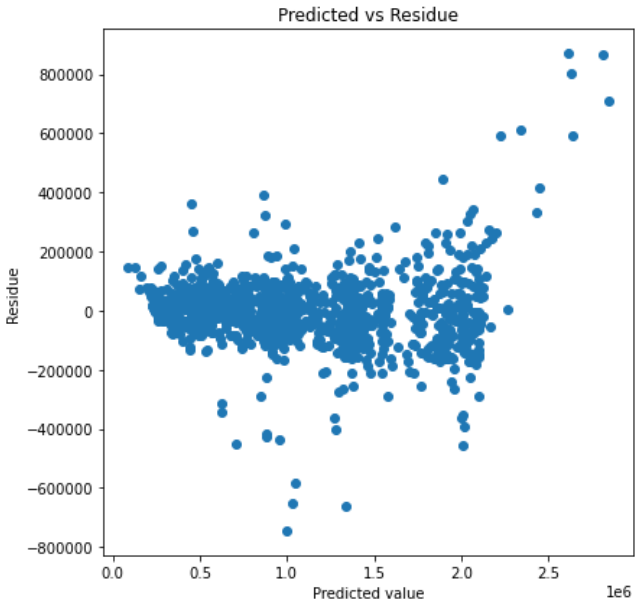
INFERENCES FROM THE PROJECT

1. Looking at the accuracy scores of the different models for both datasets, we note that RMSE score of train and test data are relatively low in case of Linear regression/Lasso/Ridge. Polynomial regression has over fit the data as the train RMSE is very less when compared to test RMSE. There is a problem of under fitting with Elastic Net because the train and test RMSE seems to be very high.

2. i. Simple Linear regression:

   (a) With outliers removed:

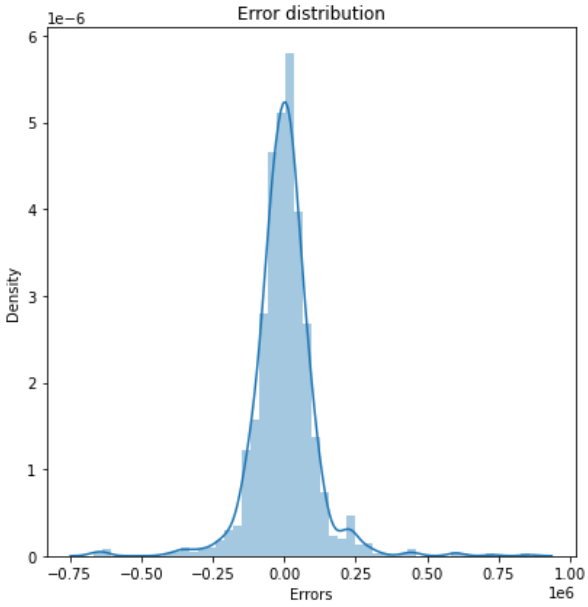


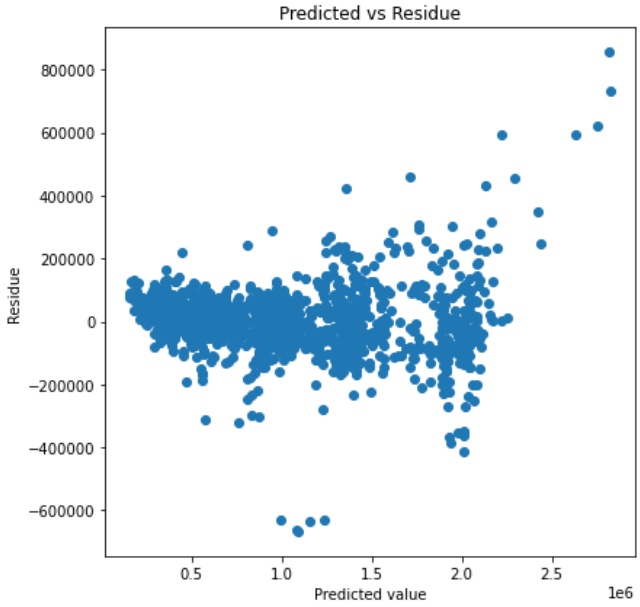


We notice that the model has error values normally distributed and also the right hand side graph suggests homoscedasticity since except for a few point the scatter plot is a constant line.

This suggests that this model can be relied upon even for further addition of data points.

(b) With box cox data:



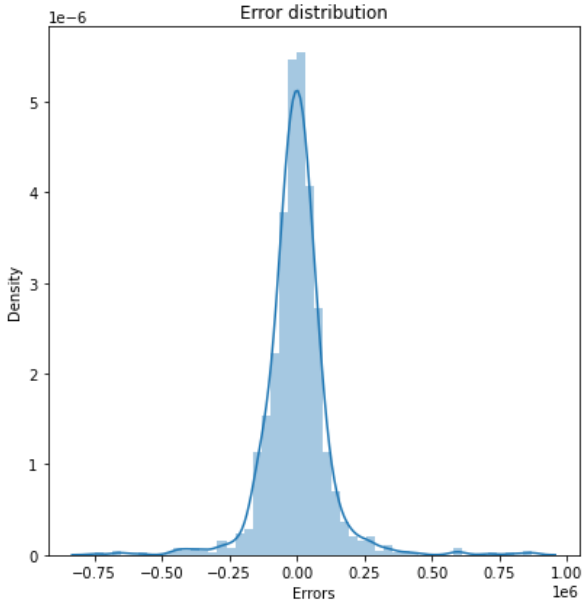


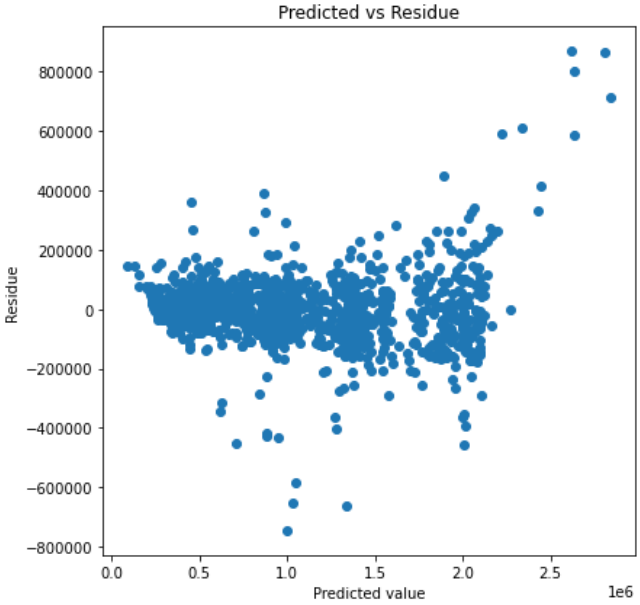
We notice that the model has error values normally distributed and also the right hand side graph suggests homoscedasticity since except for a few point the scatter plot is a constant line.

This suggests that this model can be relied upon even for further addition of data points.

ii) Lasso regression

(a) With outliers removed data:

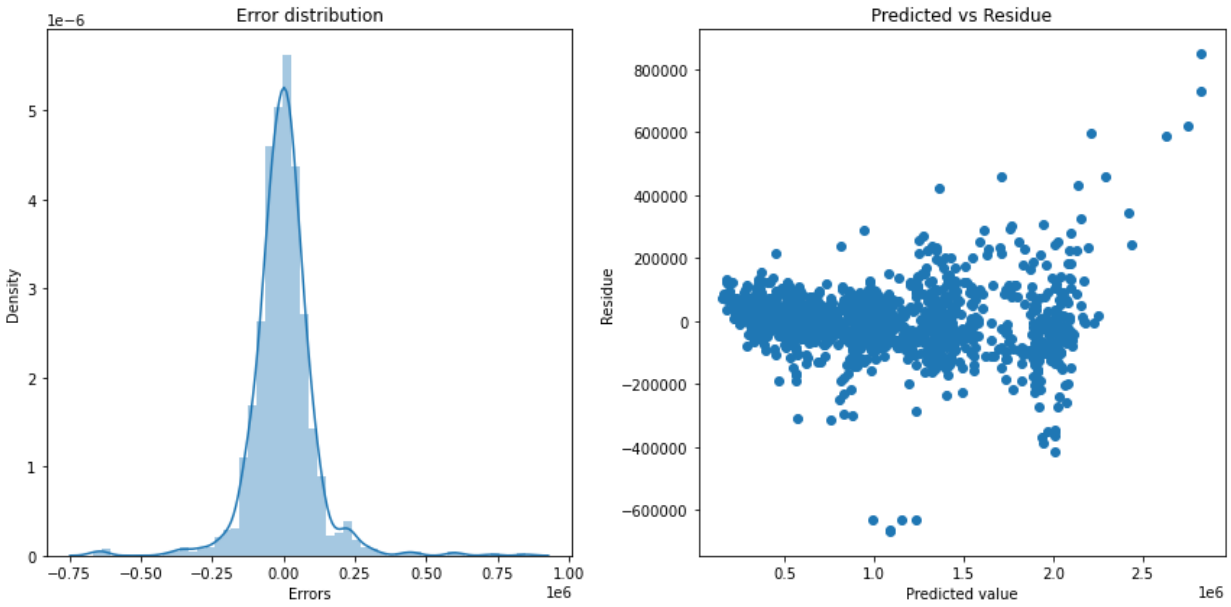




We notice that the model has error values normally distributed and also the right hand side graph suggests homoscedasticity since except for a few point the scatter plot is a constant line.

This suggests that this model can be relied upon even for further addition of data points.

(b) With Box cox data:

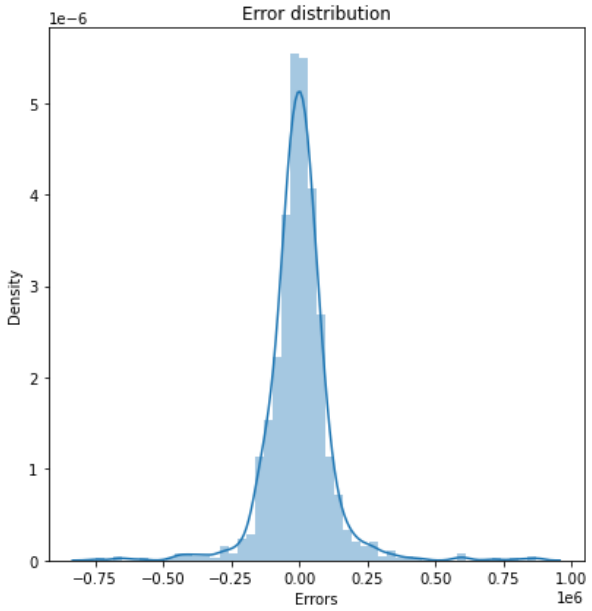


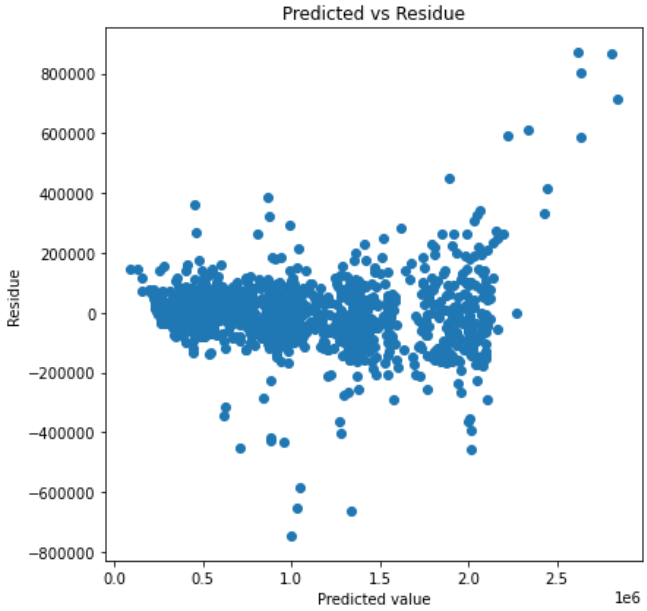
We notice that the model has error values normally distributed and also the right hand side graph suggests homoscedasticity since except for a few point the scatter plot is a constant line.

This suggests that this model can be relied upon even for further addition of data points.

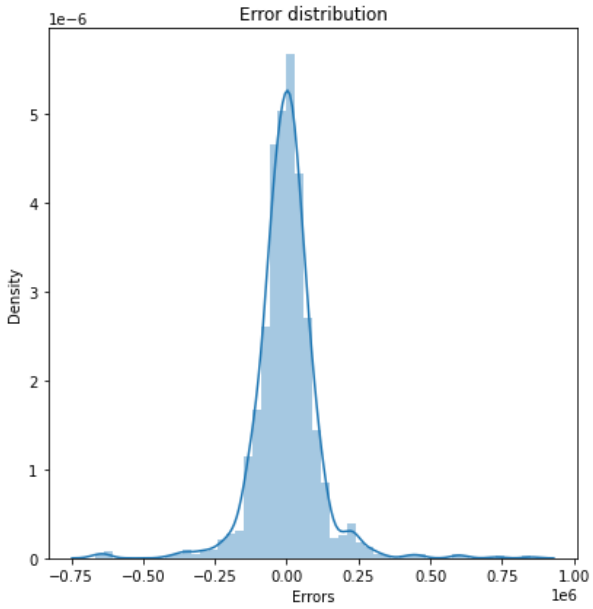
iii. Ridge regression:

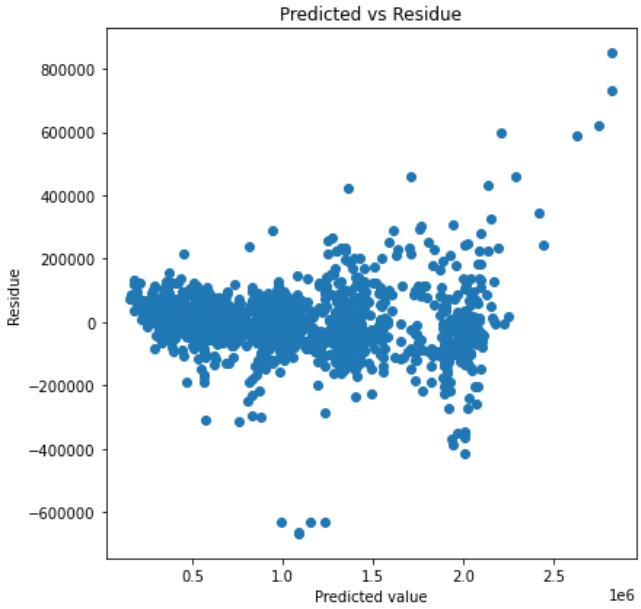
(a) With outliers removed data:





(b) With box cox data:

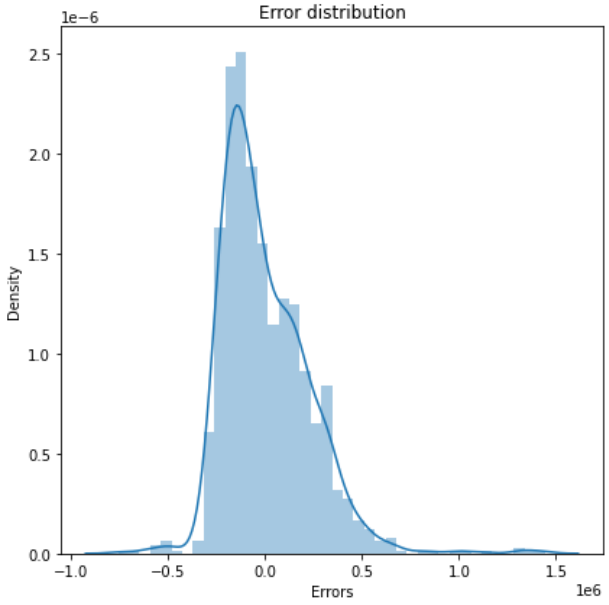


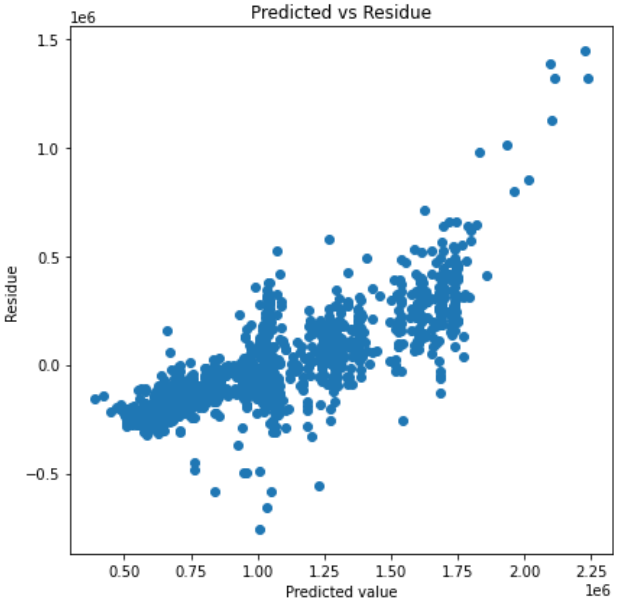


In both cases we notice same kind of conclusion as the previous ones; making it a reliable model.

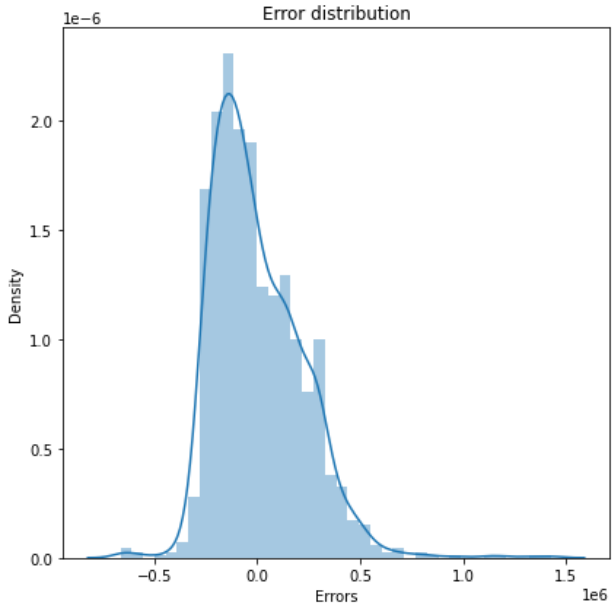
iv. Elastic Net:

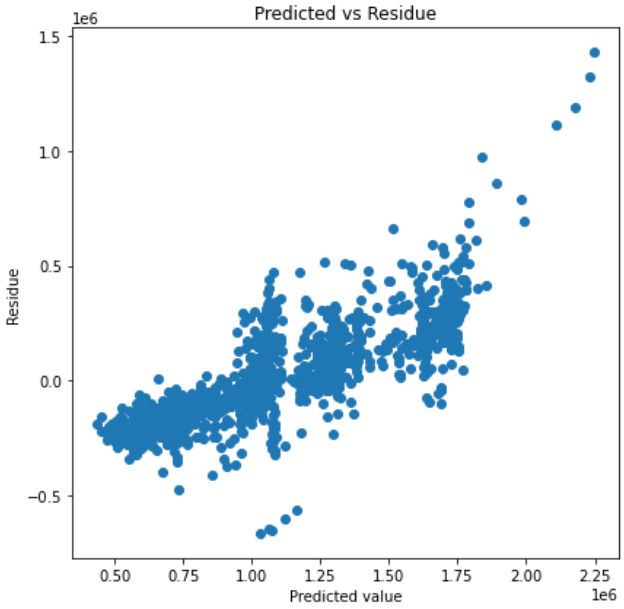
(a) With outliers removed data:





(b) With box cox data:

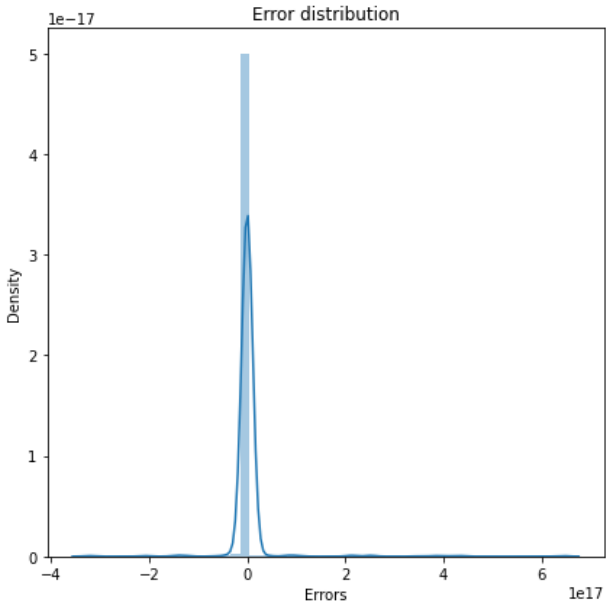


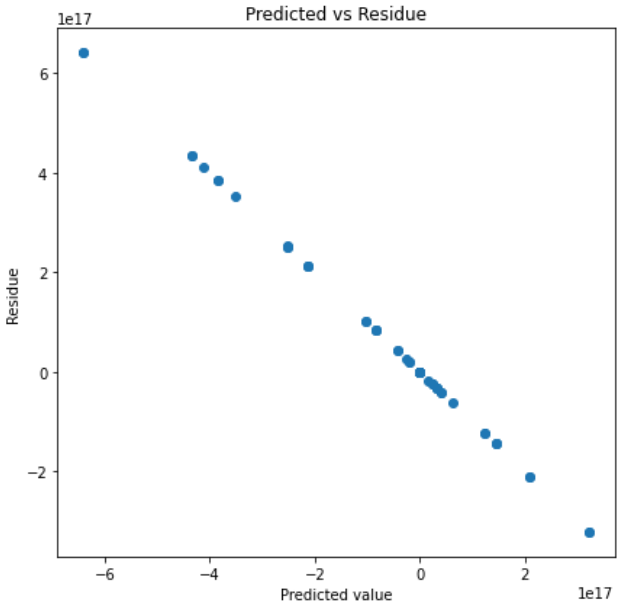


In both types of data we notice that the error distribution is right skewed. Also the variance is heteroscedastic as the predicted value vs error graph is not a constant. This implies that the model is not reliable if more data points are added to the existing one.

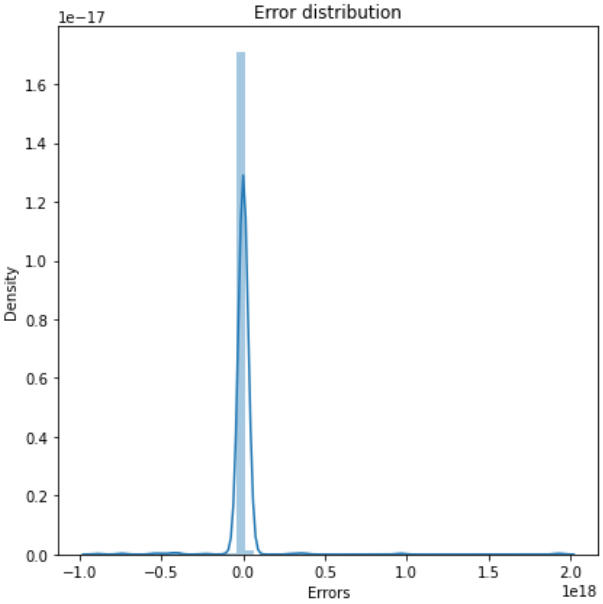
v. Polynomial regression:

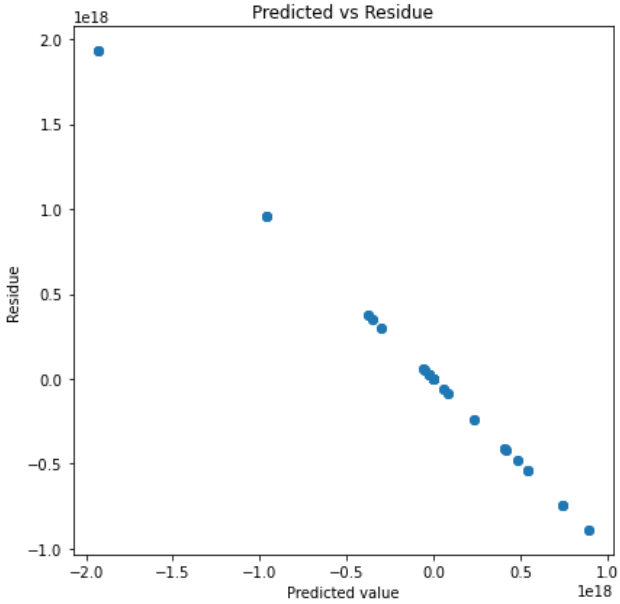
(a) With outliers removed data:





(b) With box cox data:

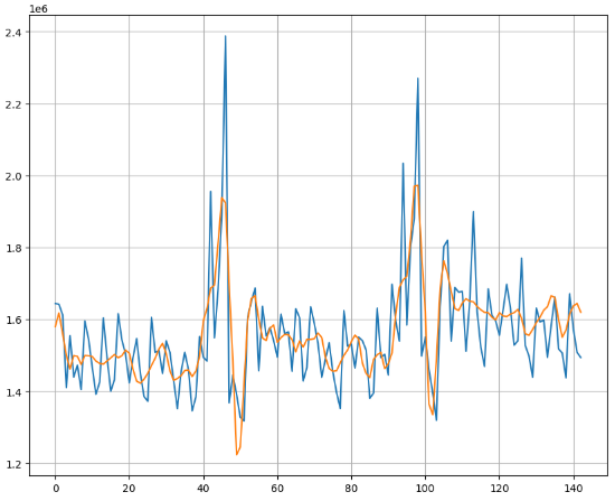


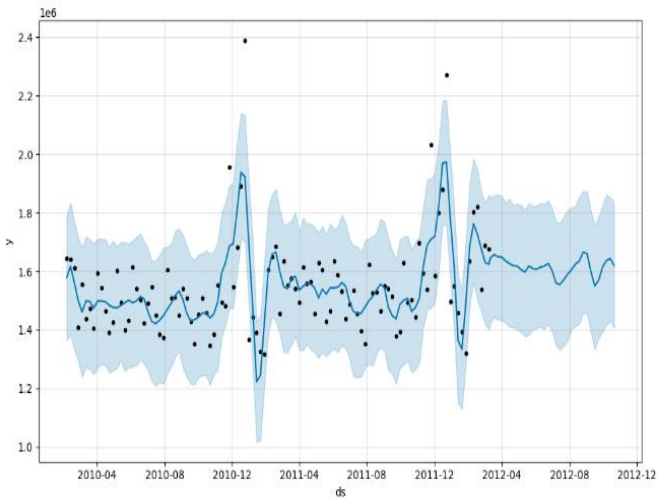


Here in both cases the error is skewed on both sides and error is heteroscedastic, suggesting model is not reliable.

So to predict sales using regression models we choose to go with Simple Linear regression (Ridge/Lasso).

3. In case of Time series models, FB Prophet works better as it covers all the predictions well within the range of the confidence interval.


1e6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
0
20
40
60
80
100
120
140


1e6
2.4
2.2
2.0
1.8
1.6
1.4
1.2
1.0
y
2010-04
2010-08
2010-12
2011-04
2011-08
2011-12
2012-04
2012-08
2012-12
ds

FUTURE POSSIBILITIES

The models can be used with fresh incoming data points in future and the best algorithm can be chosen based on the R2 scores, RMSE scores and adjusted R2 scores. The way outliers are treated can be handled automatically to give the best scores.

If the sales are going to follow the past trend, then the time series forecasting can be used to help forecast the future sales.

In either of the cases given the data, the project should select the best model for the new data and predict.

The independent variables in multivariate FB Prophet can be modeled further to improve the predictions.

CONCLUSION

In this project we analyzed the Walmart sales data by using different regression techniques and Time series forecasting. In regression, different techniques like ridge, lasso etc. were analyzed and predictions based on the same were performed.

On the other hand, forecasting the data with Time series alone has also been performed based on the data pertaining to each store separately.

REFERENCES

1. Linear, Ridge and Lasso Regression comprehensive guide for beginners (analyticsvidhya.com)

2. Ridge and Lasso Regression: L1 and L2 Regularization | by Saptashwa Bhattacharyya | Towards Data Science

3. Prophet vs SARIMA — Time Series Forecasting | by Rishabh Sharma | MLearning.ai | Medium