CAPSTAN PROJECT

YES BANK STOCK CLOSING PRICE PREDICTION

BY SAMBIT PANI

PROJECT DETAILS

Problem

- Yes Bank is a well known bank. Since 2018, it has been in the news of the fraud case involving Rana Kapoor.
- Owing to this fact, it was interesting to see how that impacted stock price.

Given

- Monthly stock prices of Yes Bank since inception.
- Open,High,Low and Closing prices of every month.

Objective

- Use of machine learning models.
- Find the monthly closing price of the stock.

DATASET DESCRIPTION

- The dataset comprises of 185 rows and 5 columns. Columns are as follows:-
- **Date :** It is the date taken as the beginning of the month from which readings has been taken.
- **Open :** It is the month opening price. i.e. the price at the beginning of the month.
- **Close :** It is the closing price of the month. i.e. the price at the end of the month.
- **High :** It is the highest price throughout the month.
- **Low :** It is the lowest price throughout the month.
- Here “Close” is considered as the dependent variable.

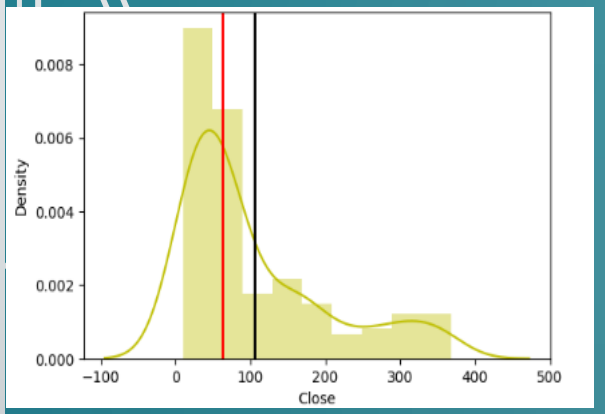
DATA WRANGLING

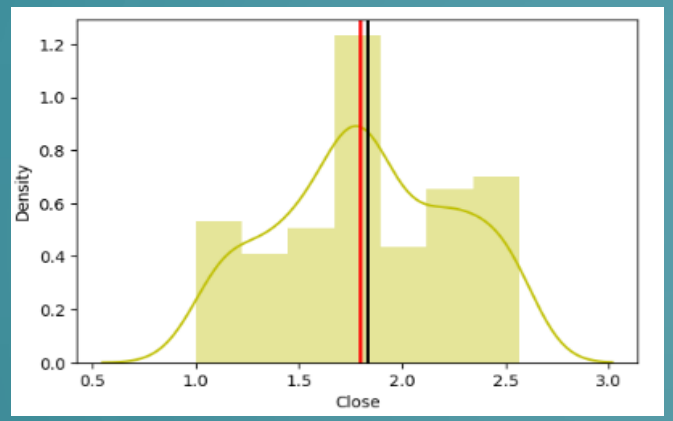
- The data is clean and doesn't have any null values.
- The Date column is of object data-type. So changed to datetime.
- Also I have added three new columns.
- One is year as 1st year,2nd year etc.
- Another is month number. As month affects the cyclic price behavior and financial announcements.
- Last one is pivot. It is just the mean of High, Low & Open.

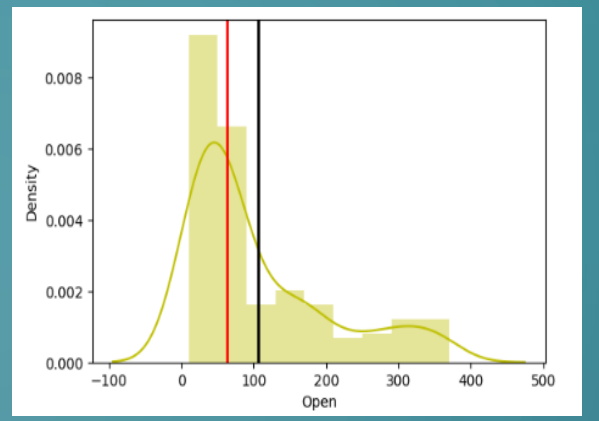
```
[ ]  # Write your code to make your dataset analysis ready.
     df['Date'] = df['Date'].apply(lambda x: datetime.strptime(x,"%b-%y"))

[ ]  # pivot point is the mean of high,low and close. It acts crucial role in stock prices behavior.
     df['pivot'] = round((df['High']+df['Low']+df['Open'])/3,2)

     # adding month number from date as there is a quarterly result associated with the company. SO month is important.
     # adding year from starting of the stock.
     df['month'] = df['Date'].dt.month
     df['year'] = df['Date'].dt.year-2004
```

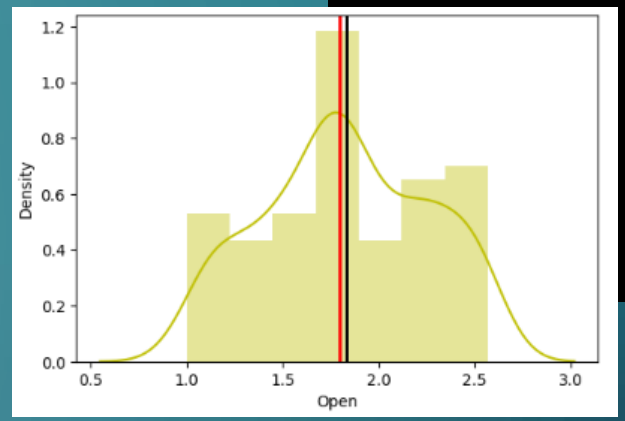
DEPENDENT VARIABLE "CLOSE"

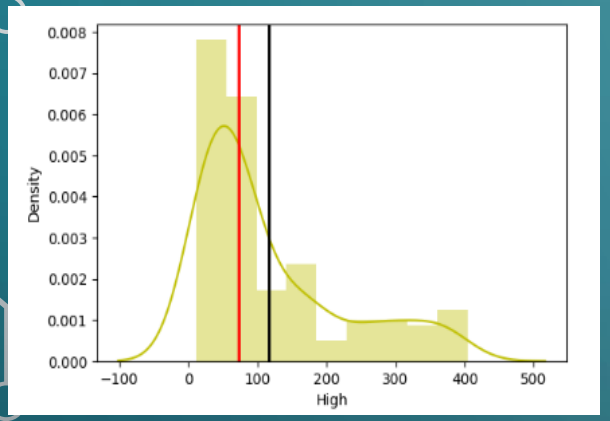
- The stock keeps rising after the inception of stock.
- It can be seen clearly that the closing price starts falling after 2018 fraud case.
- It keeps falling drastically till the listing price.

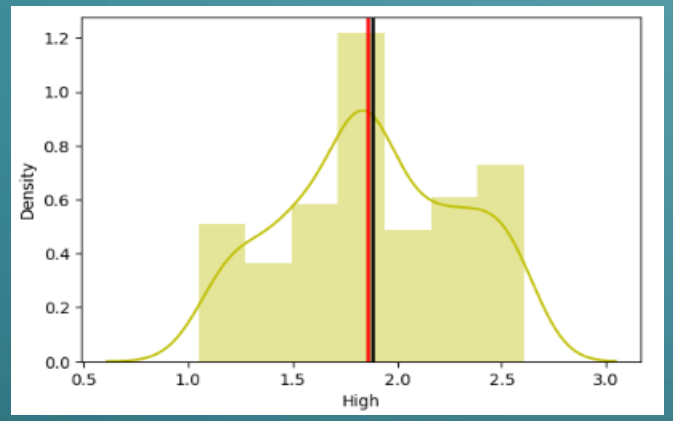


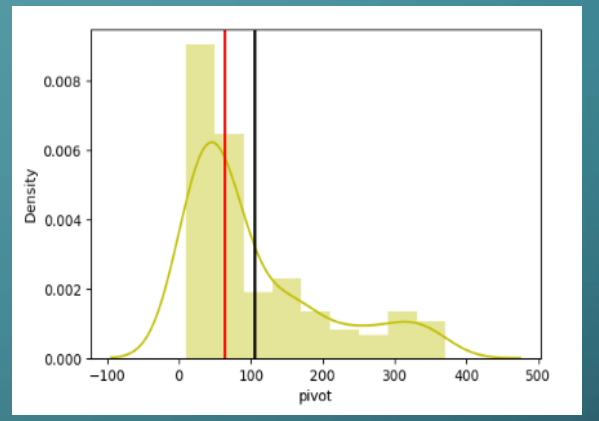


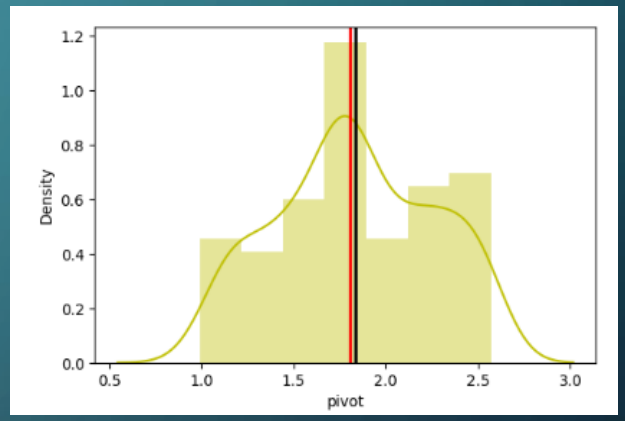






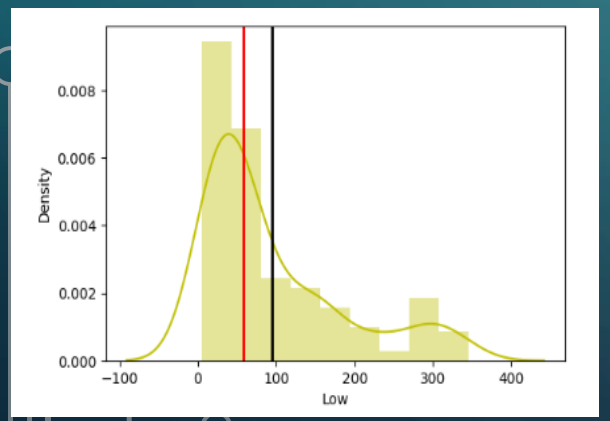


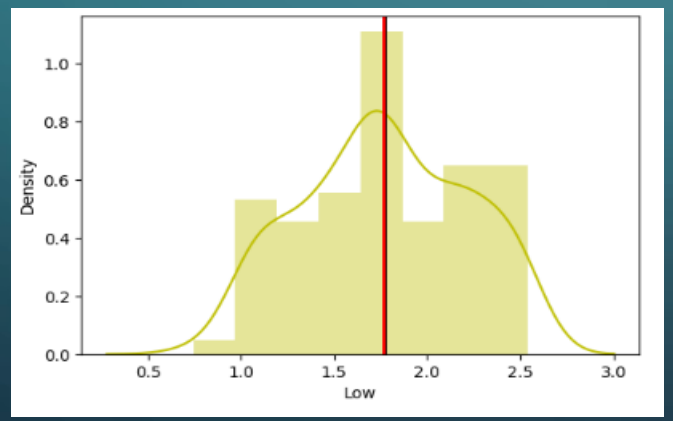




BEFORE

AFTER



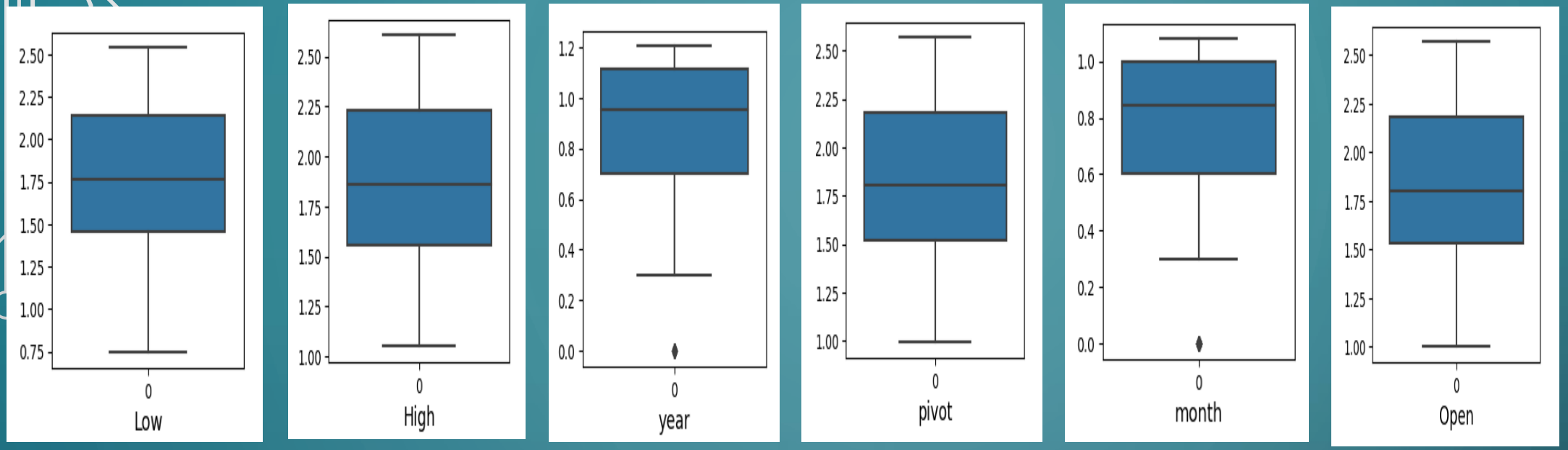


BEFORE

AFTER

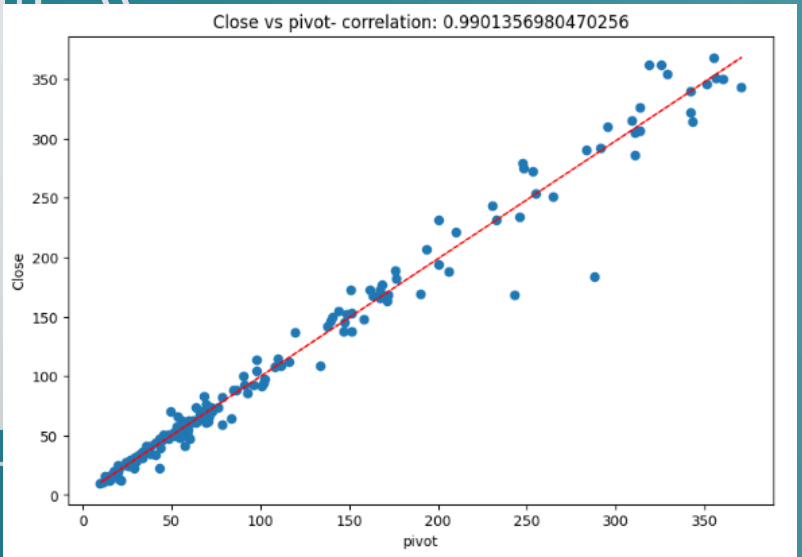
LOG TRANSFORMATION

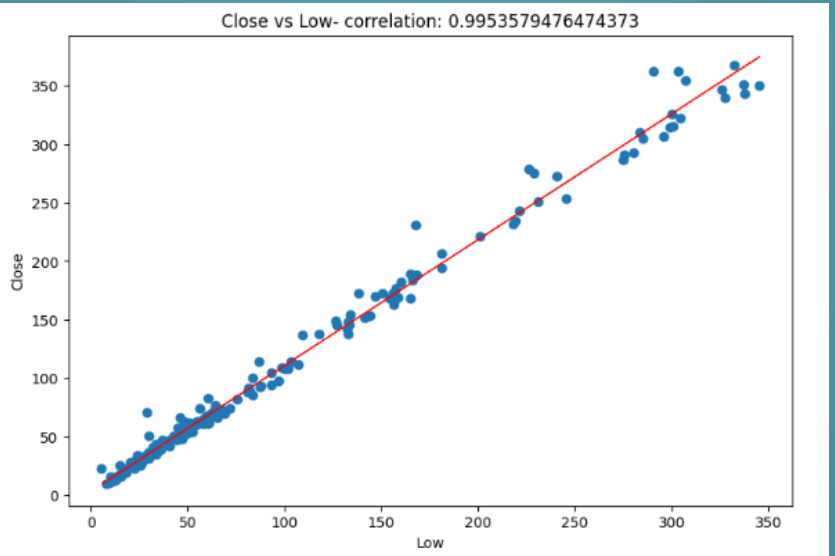
- Here we find some positively skewed data.
- So we apply log transformation.
- Here are the before and after plots of Close, High, Low, Open, Pivot log transformation.

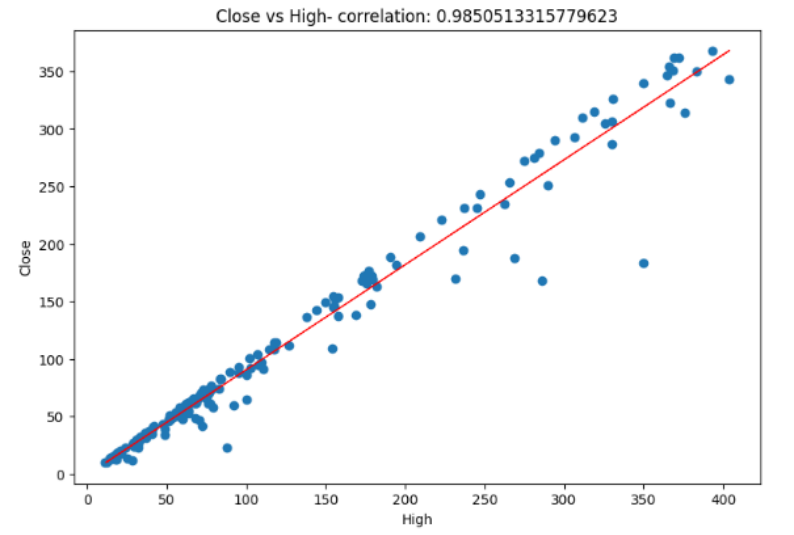


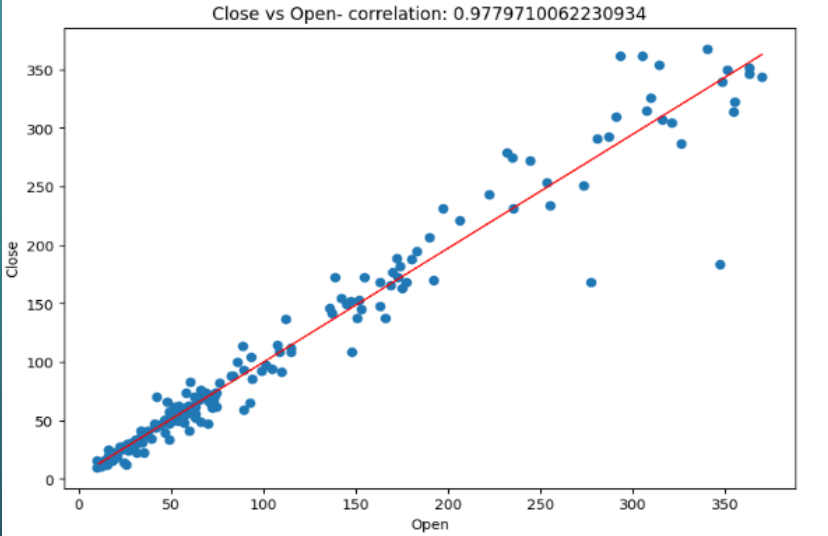
BOX PLOT

- Box plots shows that the data set has no outliers. So we can proceed with our next operations.



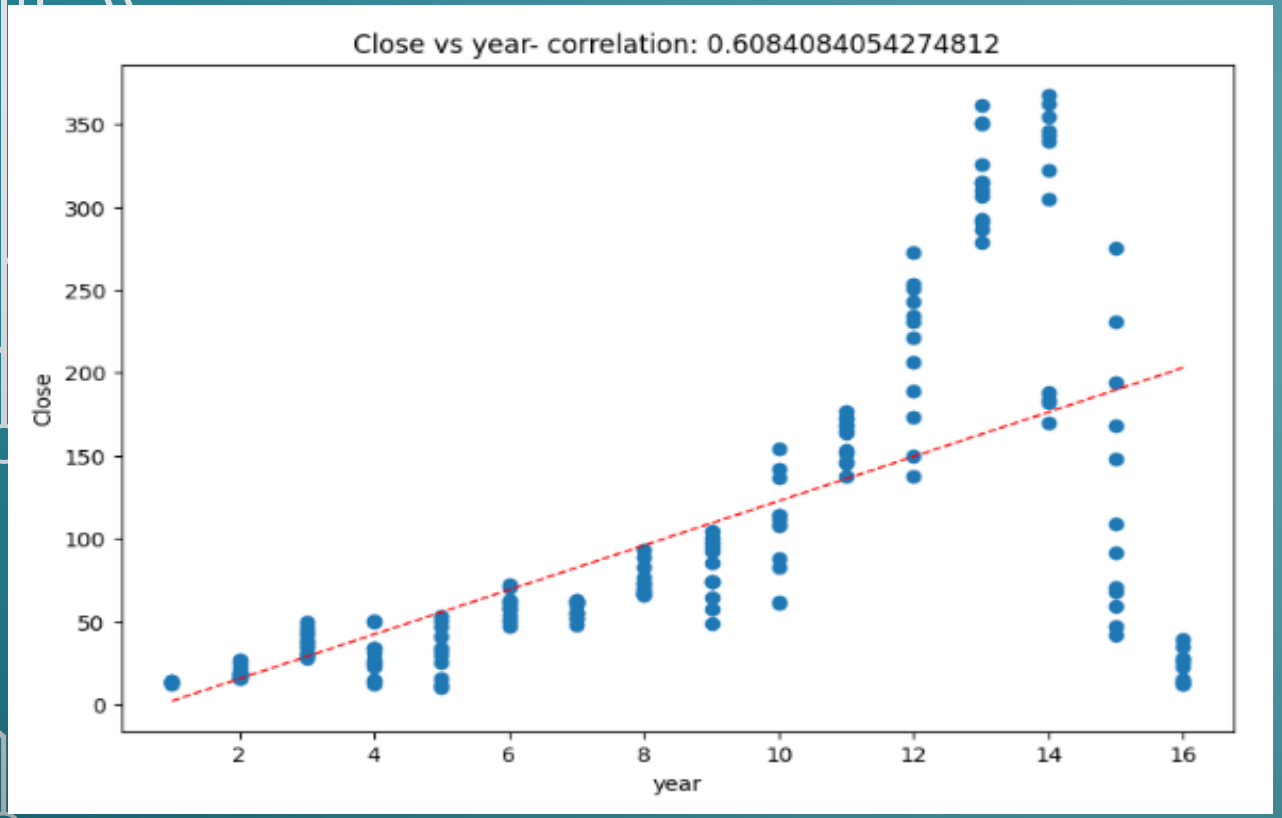


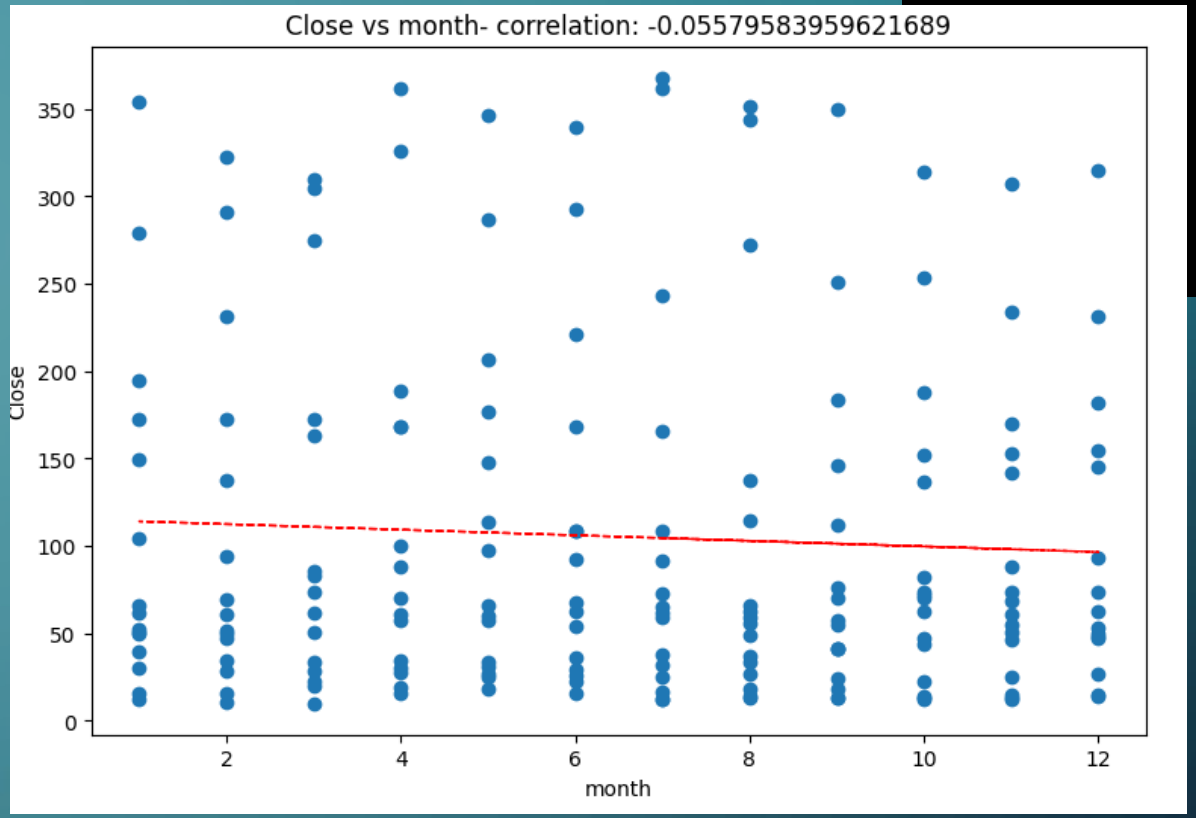




Correlation(with Close)	Coefficient
Open	0.9780
High	0.9850
Low	0.9953
Pivot	0.9901

- These plots shows that Close is highly correlated to Open, High, Low and Pivot.
- That means predictability will be more efficient.





- These plot shows that Close is more correlated to Year as Corr = 0.6084 and less correlated to Month as Corr = -0.05579 .
- I will keep these features assuming both may contribute even a little bit to the output.

CORRELATION HEATMAP

- Here we see every features' correlation with respect to each other.
- It has been found that High, Low, Open & Pivot are highly correlated to each other and with Close. So we will keep Pivot as it is the mean of Open, High & Low.
- We will also keep month and year although the correlations less.

VIF

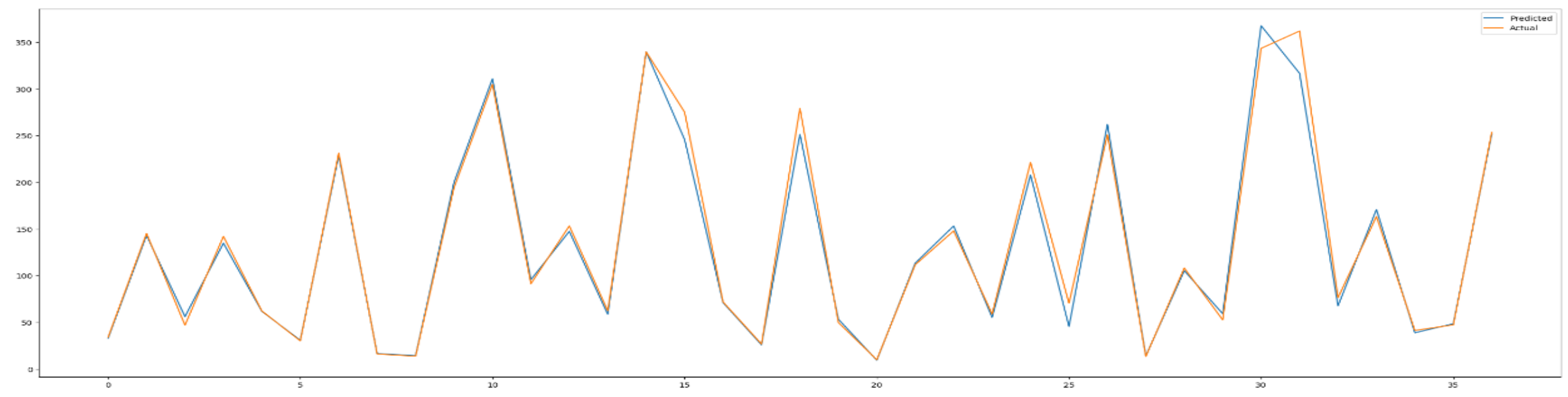
- After selecting independent variables Vif has been found out to be less than 6. So now we will proceed for model implementation.
- The numeric features are Pivot, month & year.

	variables	VIF
0	pivot	3.521748
1	month	2.513890
2	year	5.251269

MODEL IMPLEMENTATION

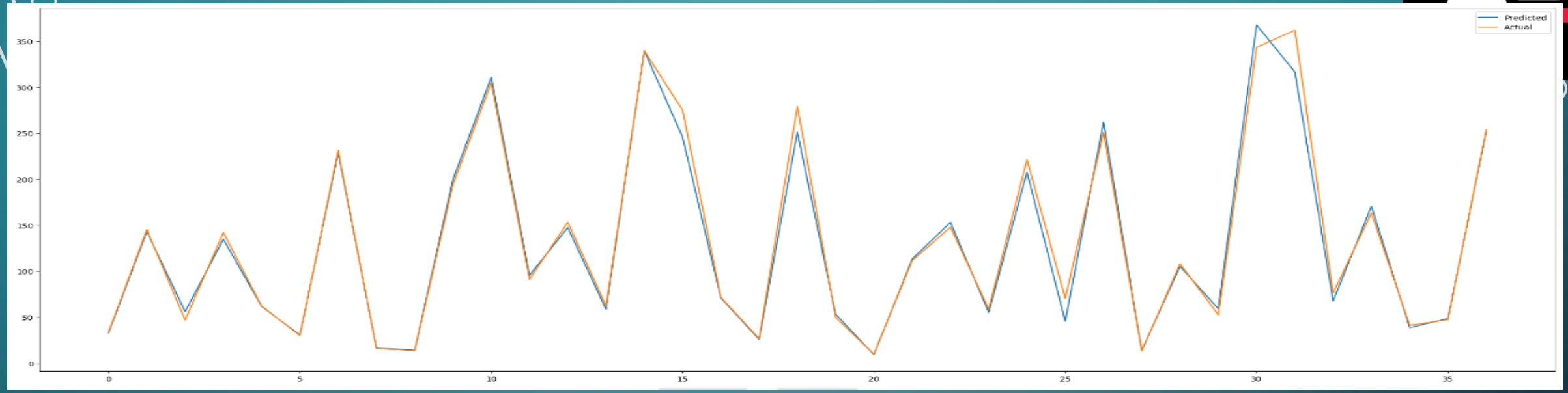
- Now we will split the data into 80% and 20%. Using 80% data we will learn the trend and pattern inside the data through different regression models are as follows : -
    - Linear Regression
    - Lasso Regression
    - Ridge Regression
    - Elastic-Net Regression
- We will test the output using the remaining 20% data using different metrics such as:-
    - Mean Squared Error
    - Root Mean Squared Error
    - R2 score
    - Adjusted R2 score

LINEAR REGRESSION PREDICTION



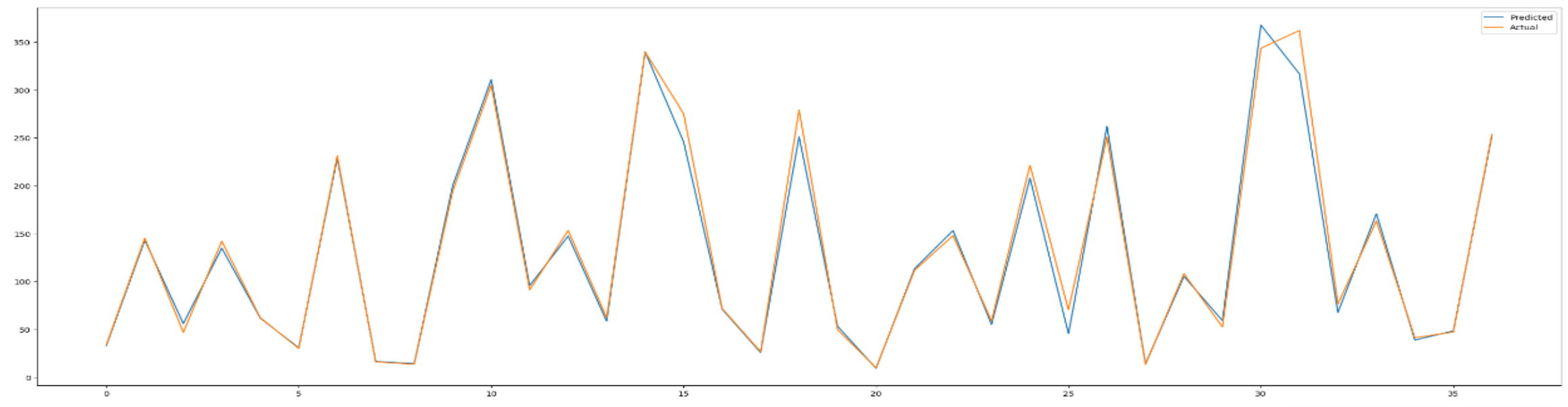
- In this model RMSE value is 12.4638
- Here the predicted values are highly congruent with the actual values.
- R2 score is 0.9850 which means 98.5% variance of dependent variable can be predictable from independent variables.

LASSO REGRESSION PREDICTION



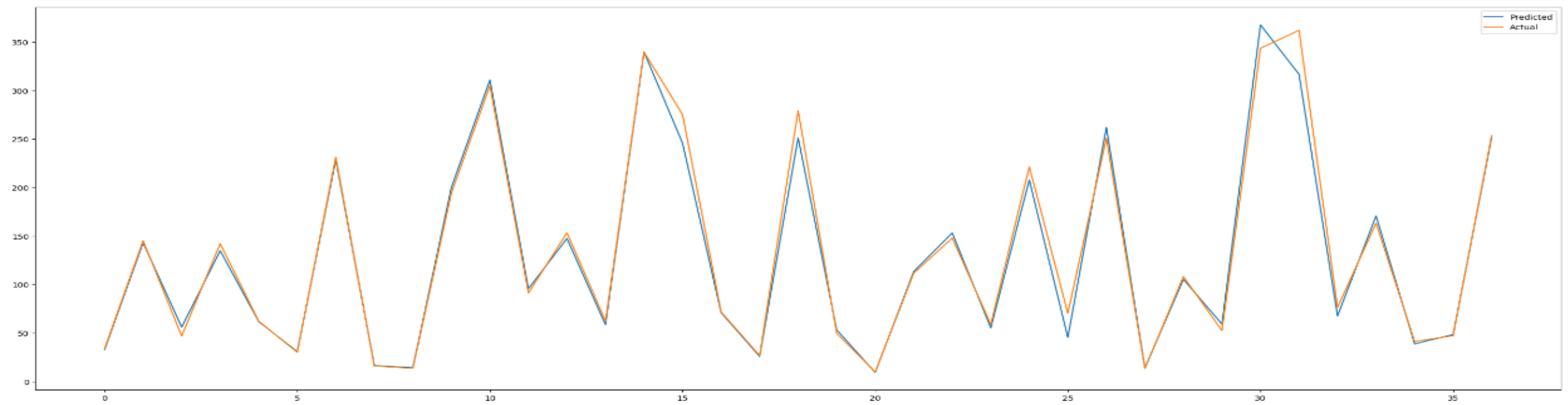
- In this model RMSE value is 12.4637
- Here also the predicted values are highly congruent with the actual values.
- R2 score is 0.9861 which means 98.61% variance of dependent variable can be predictable from independent variables.

RIDGE REGRESSION PREDICTION



- In this model RMSE value is 12.4637
- Here the predicted values are highly congruent with the actual values.
- R2 score is 0.9862 which means 98.62% variance of dependent variable can be predictable from independent variables.

ELASTIC-NET REGRESSION PREDICTION



- In this model RMSE value is 12.4637
- Here the predicted values are highly congruent with the actual values.
- R2 score is 0.9849 which means 98.49% variance of dependent variable can be predictable from independent variables.

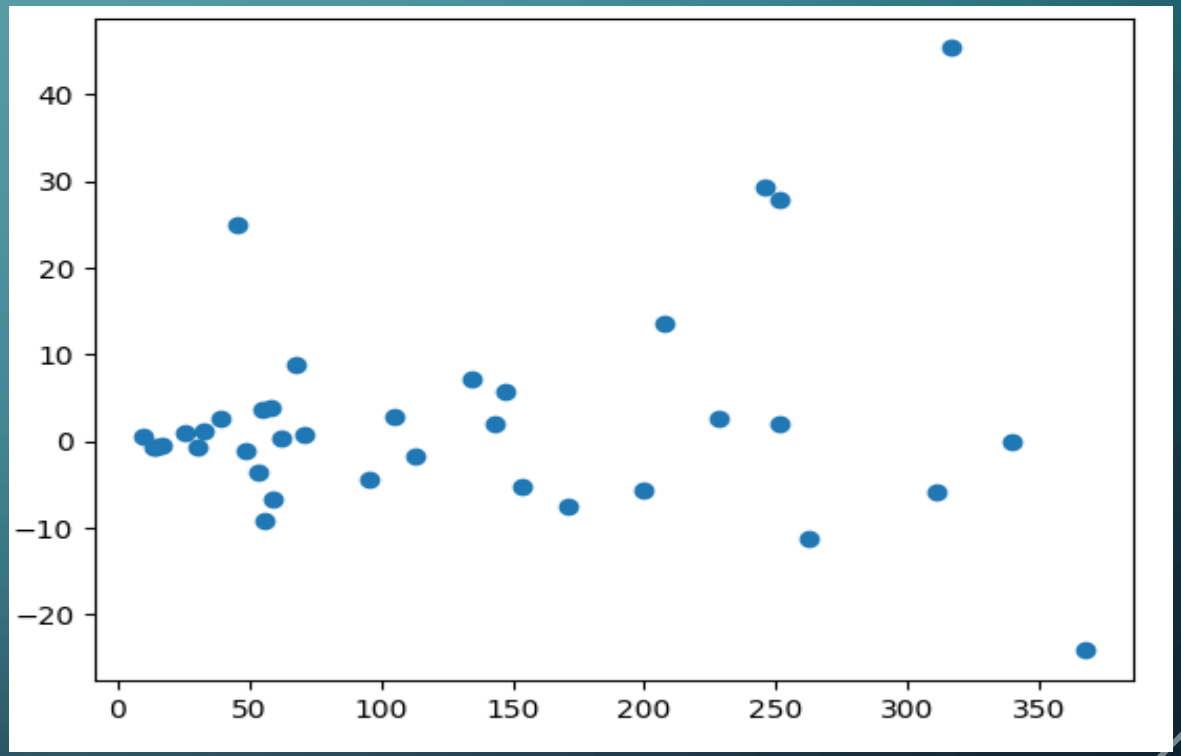
COMPARISON OF METRICS

- By comparing the metrics it has been found out that Elastic-Net Regression model is best among all models.
- Also we can see that R2 score is high.
- So we can confidently apply this model to predict the future value also.

	Metrics	Linear	Lasso	Ridge	ElasticNet
0	MSE	155.3457	155.345664	155.345664	155.345664
1	RMSE	12.4638	12.463774	12.463774	12.463774
2	R2	0.9862	0.986199	0.986199	0.986199
3	Adj R2	0.9850	0.984944	0.984944	0.984936

RESIDUAL PLOT ELASTIC NET MODEL

- In this plot we are going to find the heteroscedasticity of our data exists or not.
- Since the graph is quiet symmetrical about zero line. We can say that our data is free from heteroscedasticity.
- Hence the assumption of homoscedasticity is valid.



FINAL RESULT

1. It has been seen that the stock price fall after 2008 fraud case.
2. Again i found that the data comprises of 5 variables only.
3. We created two more variables. One is month number as it plays a crucial role as there is financial result announcements. Another is Year and pivot,
4. We found no null values in the data.
5. Data is positively skewed so log transformation is applied.
6. There is no outliers observed so the data is clean.
7. There is no categorical column so no need for dummies.
8. Then we model the data with 4 types of regressions.
9. As per the above data Elastic net regression is quiet better in all the metrics. So we will select this model for upcoming predictions.
10. Also it has been found out that R2 value is more correlated with pivot.
11. By seeing the model accuracy it can be used confidently for upcoming predictions.

Thank you !