Capstone Project: 2

YES BANK STOCK CLOSING PRICE PREDICTION

Team Members

Suraj Kumar

Shreya Ranjan

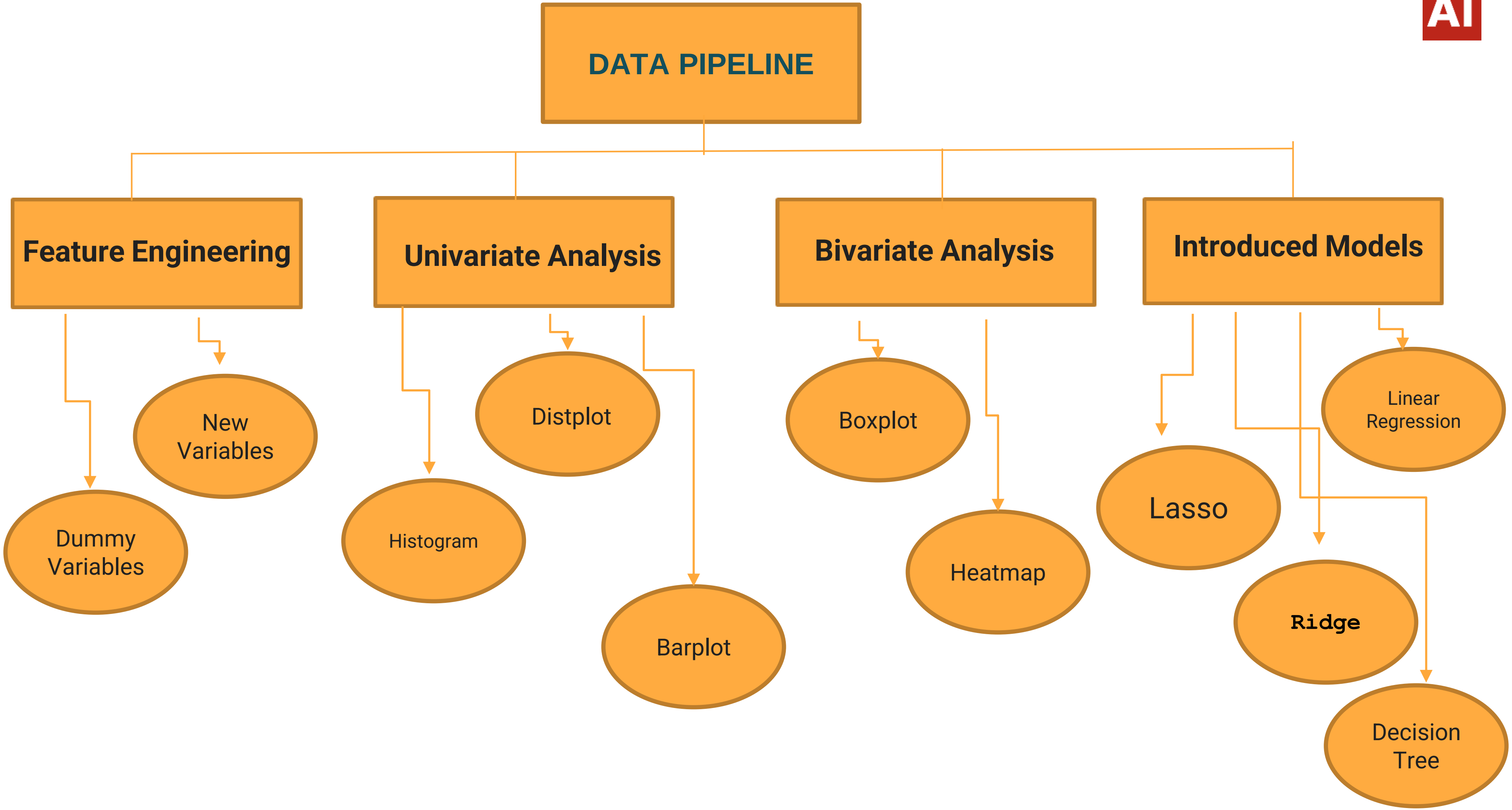
YES BANK

Yes Bank is a well-known bank in the Indian financial domain. It has been in the headlines since 2018 as a result of the Rana kapoor fraud case. Due to this, it was interesting to observe how it affected the company's stock prices and whether Time series models or other prediction models could properly reflect for such circumstances. Since the bank's founding, this dataset has included closing, starting, highest, and lowest stock prices for each month.

YES BANK STOCK CLOSING PRICE PREDICTION DATASET

We have 185 rows and 5 columns in our dataset. Here our dependent variable is Close and Independent variable is Open, High and Low.

Date :- It denotes the month and year for a specific pricing.
Open :- The price at which a stock started trading that month is referred to as the "Open."
High :- The highest price for that particular month.
Low :- It describes the monthly minimum price.
Close :- It refers to the final trading price for that month, which we have to predict using regression.


DATA PIPELINE
Feature Engineering
Univariate Analysis
Bivariate Analysis
Introduced Models
New Variables
Dummy Variables
Distplot
Histogram
Barplot
Boxplot
Heatmap
Linear Regression
Lasso
Ridge
Decision Tree

Libraries:

1} NumPy
2} Panda
3} Matplotlib
4} Seaborn
5} Datetime
6} Sklearn

Data Wrangling:

- Shape of the Data

```
Df.shape
```

```
(185, 5)
```

- Datatype in Data Frame

```
Df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 185 entries, 0 to 184
Data columns (total 5 columns):
 #   Column  Non-Null Count  Dtype
---  ------  --------------  -----
 0   Date    185 non-null    object
 1   Open    185 non-null    float64
 2   High    185 non-null    float64
 3   Low     185 non-null    float64
 4   Close   185 non-null    float64
dtypes: float64(4), object(1)
memory usage: 7.4+ KB
```

Data Wrangling(cont.)

Finding the Null values:

```
Df.isnull().sum()
```

```
Date     0
Open     0
High     0
Low      0
Close    0
```

Starting 10 Values

	Date	Open	High	Low	Close
0	Jul-05	13.00	14.00	11.25	12.46
1	Aug-05	12.58	14.88	12.55	13.42
2	Sep-05	13.48	14.87	12.27	13.30
3	Oct-05	13.20	14.47	12.40	12.99
4	Nov-05	13.35	13.88	12.88	13.41
5	Dec-05	13.49	14.44	13.00	13.71
6	Jan-06	13.68	17.16	13.58	15.33
7	Feb-06	15.50	16.97	15.40	16.12
8	Mar-06	16.20	20.95	16.02	20.08
9	Apr-06	20.56	20.80	18.02	19.49

Data Wrangling(cont.)

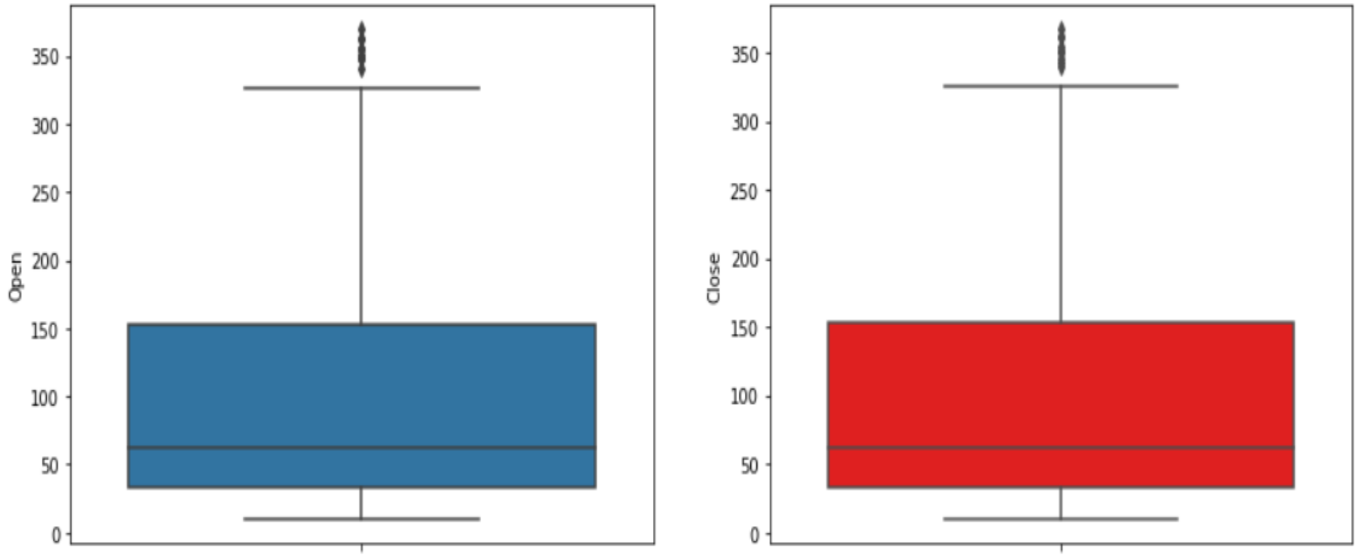
Last 5 value in the datasets

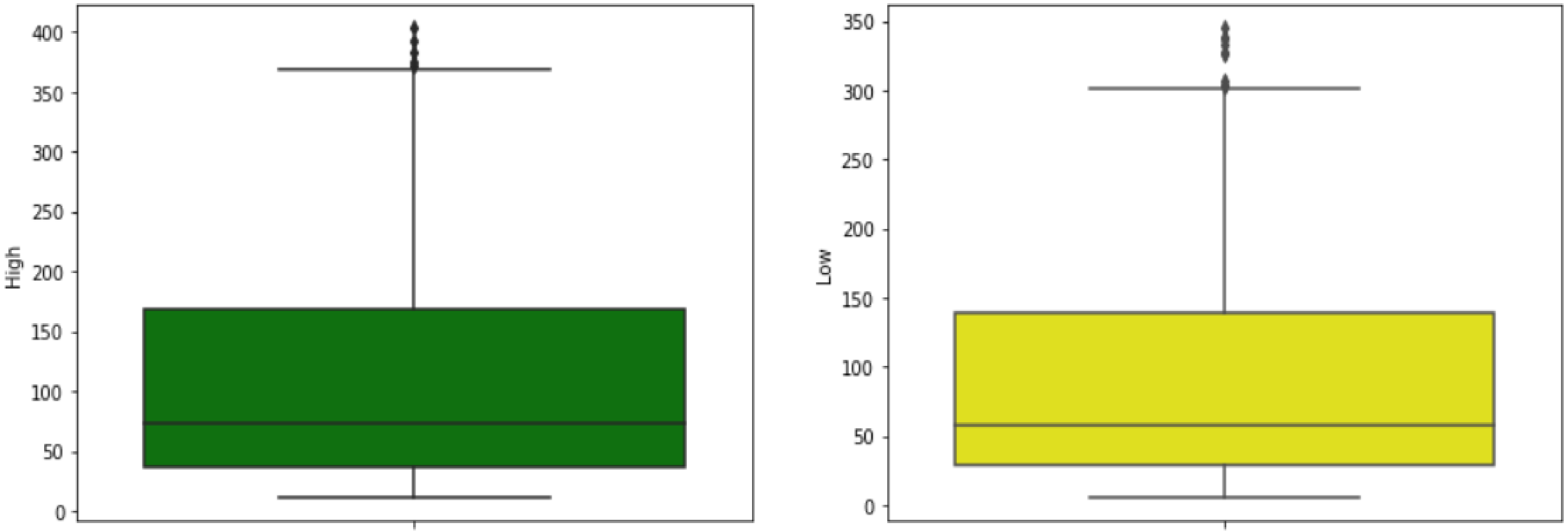
	Date	Open	High	Low	Close
180	Jul-20	25.60	28.30	11.10	11.95
181	Aug-20	12.00	17.16	11.85	14.37
182	Sep-20	14.30	15.34	12.75	13.15
183	Oct-20	13.30	14.01	12.11	12.42
184	Nov-20	12.41	14.90	12.21	14.67

Description of datasets

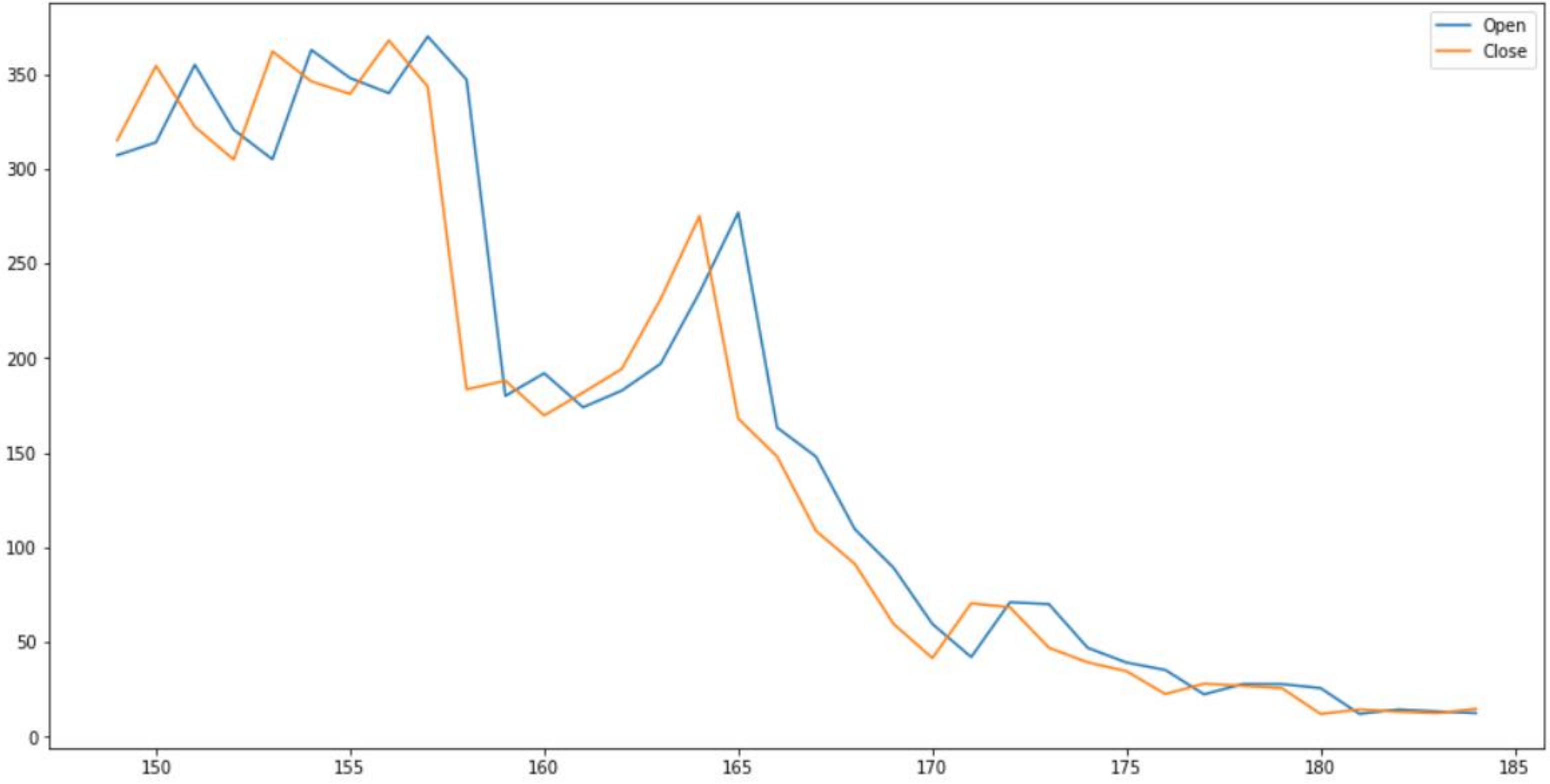
	Open	High	Low	Close
count	185.000000	185.000000	185.000000	185.000000
mean	105.541405	116.104324	94.947838	105.204703
std	98.879850	106.333497	91.219415	98.583153
min	10.000000	11.240000	5.550000	9.980000
25%	33.800000	36.140000	28.510000	33.450000
50%	62.980000	72.550000	58.000000	62.540000
75%	153.000000	169.190000	138.350000	153.300000
max	369.950000	404.000000	345.500000	367.900000

Exploratory Data Analysis:

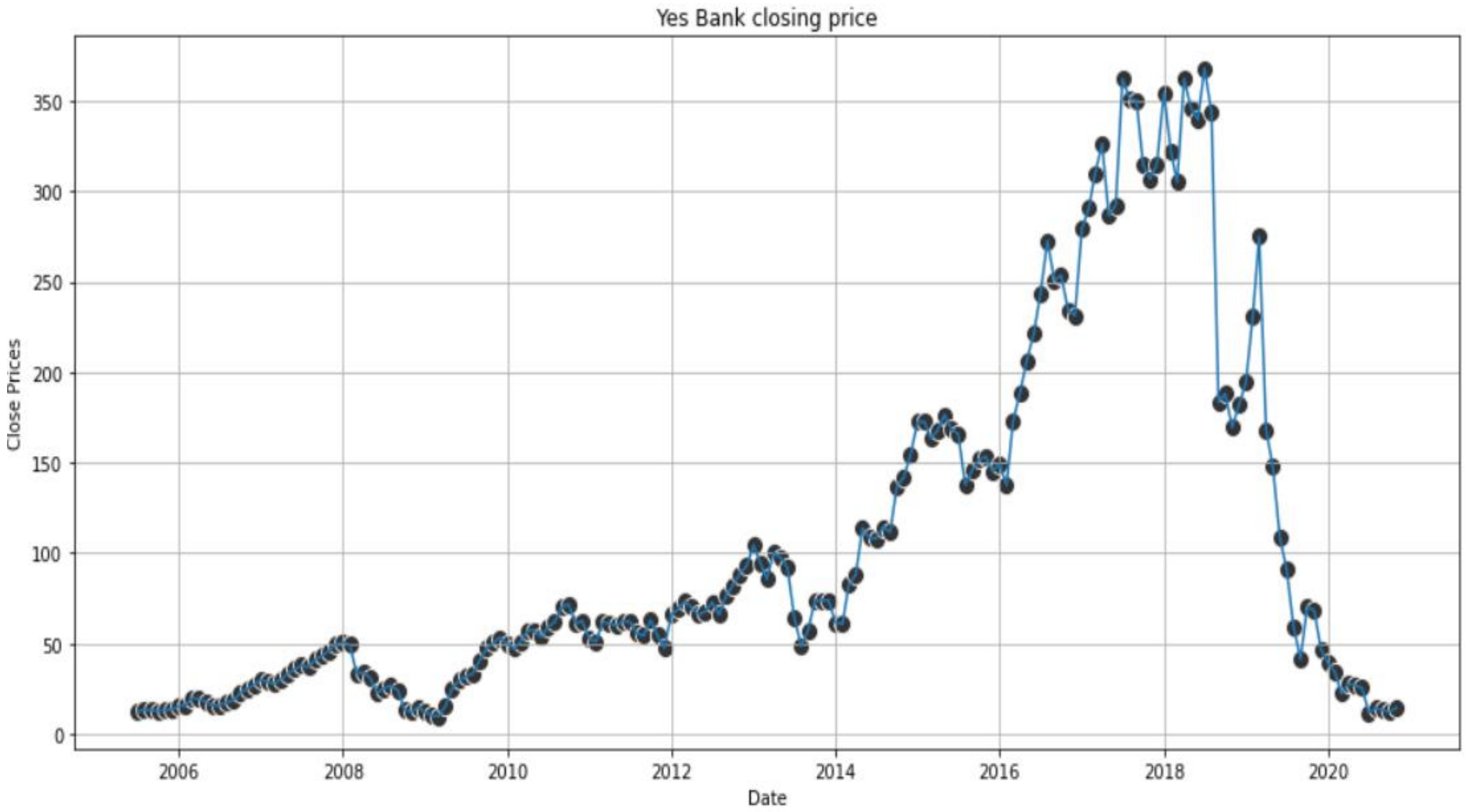




Outliners in the dataset

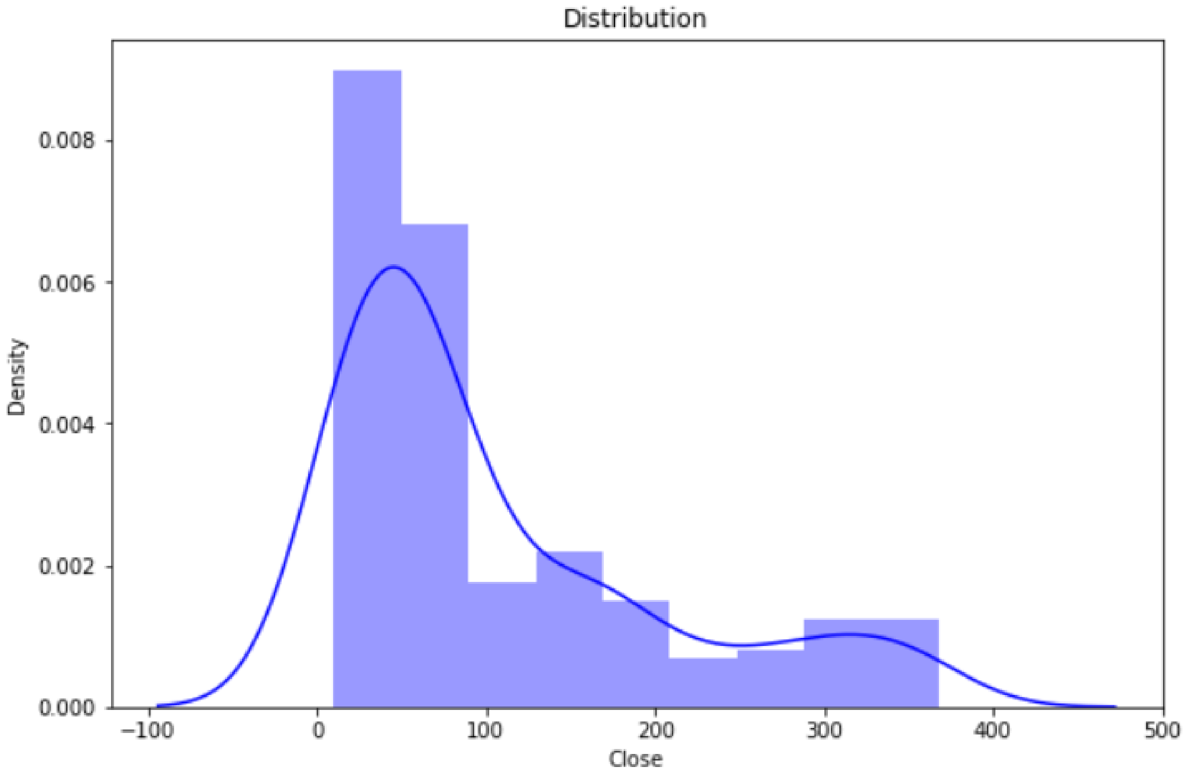


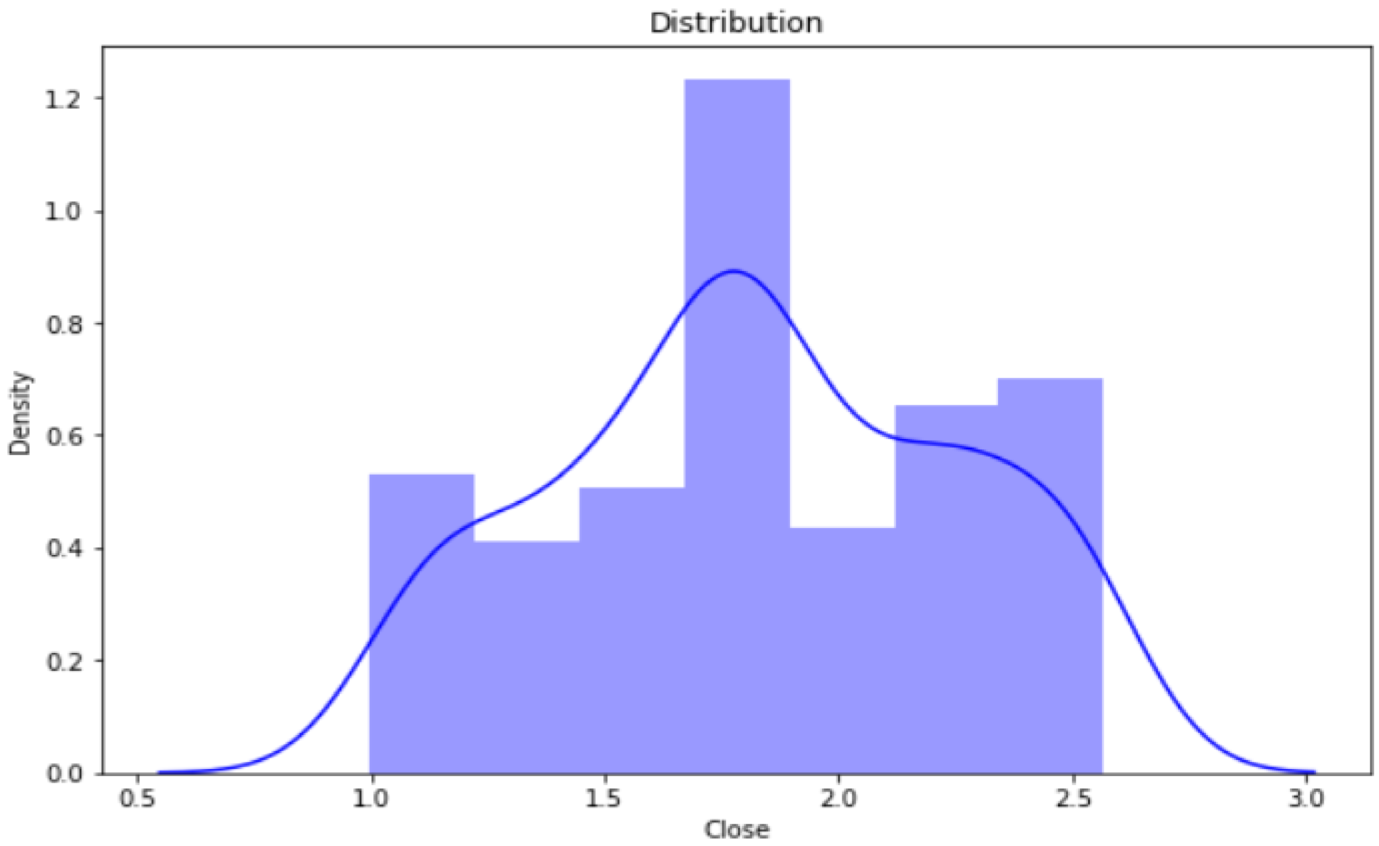
Last three year record of opening and closing stock price



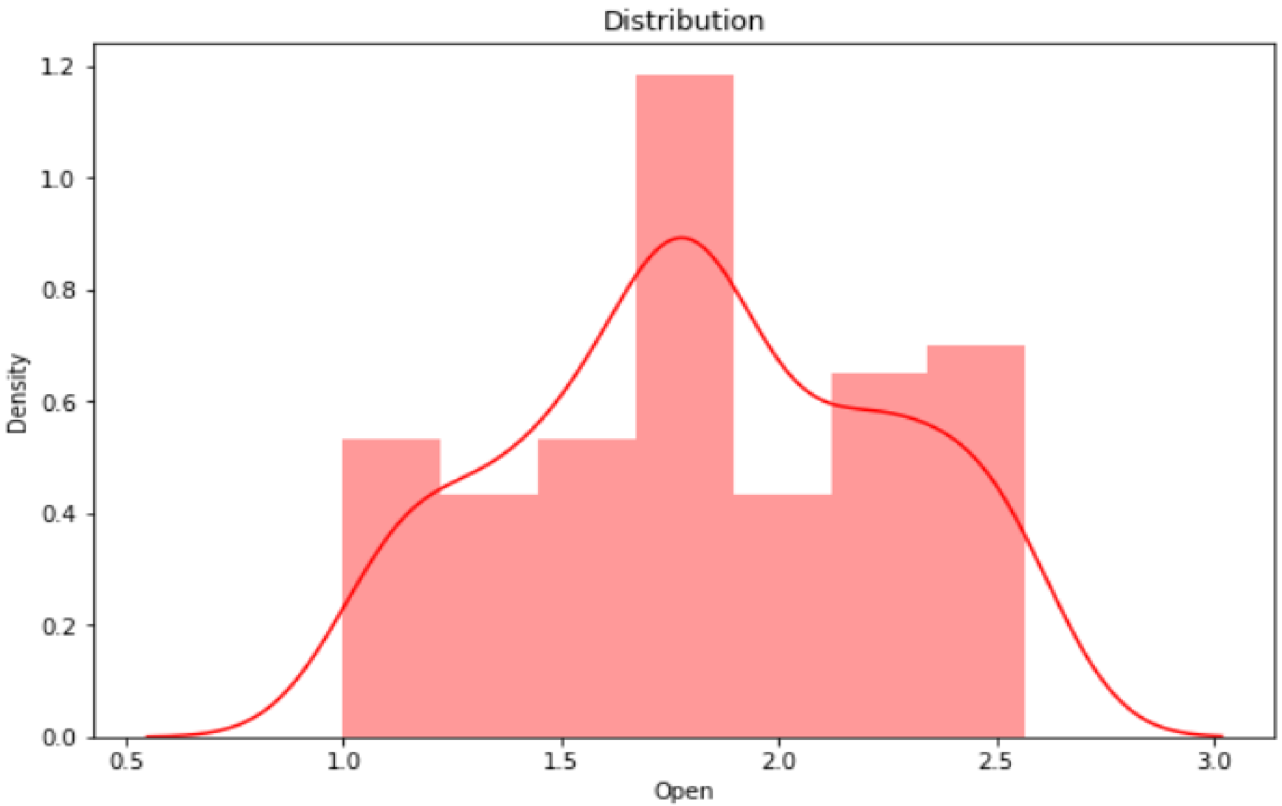
Yes Bank Closing Price Yearly

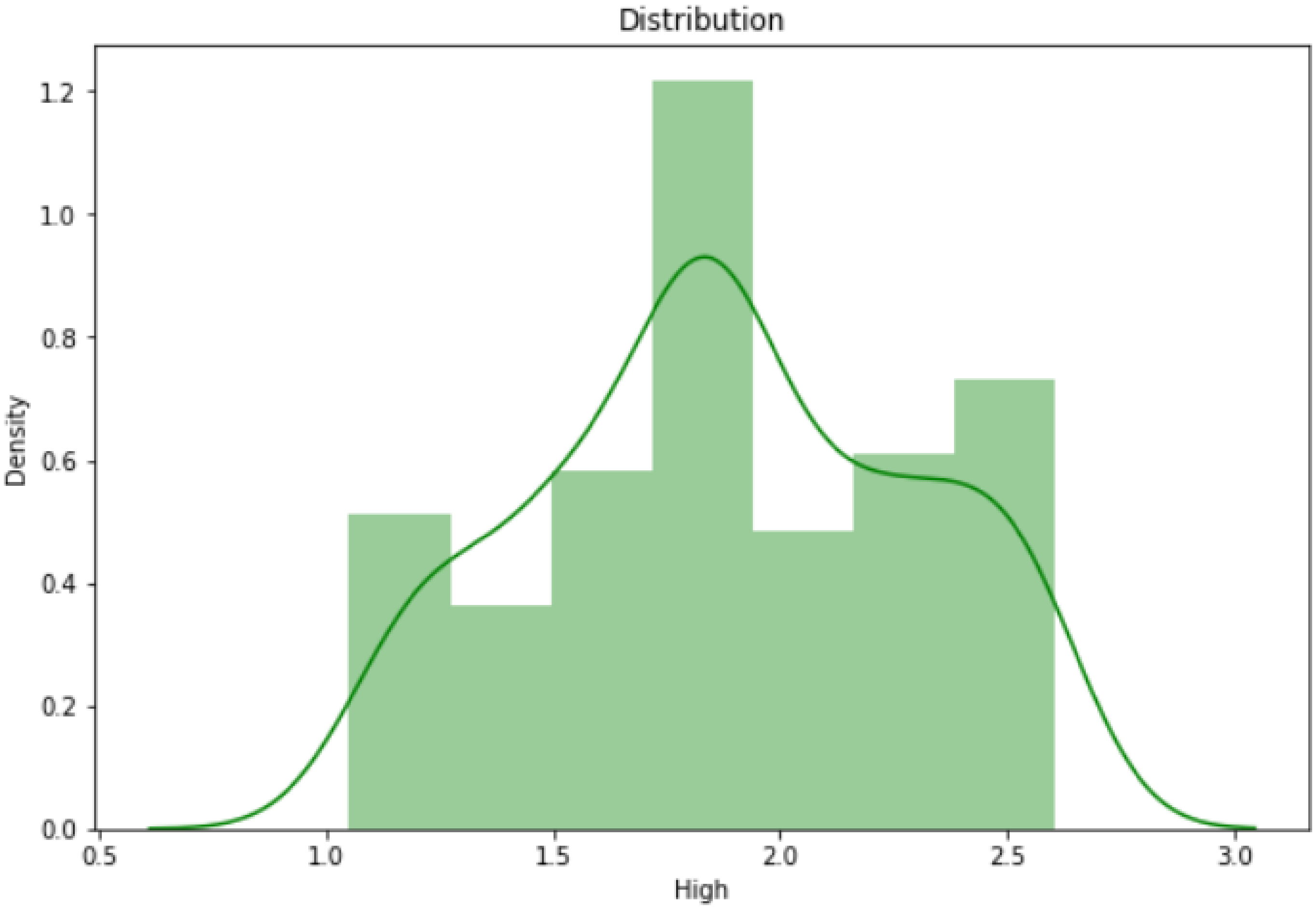
Visualization of Dependent Variable of Closing Price:

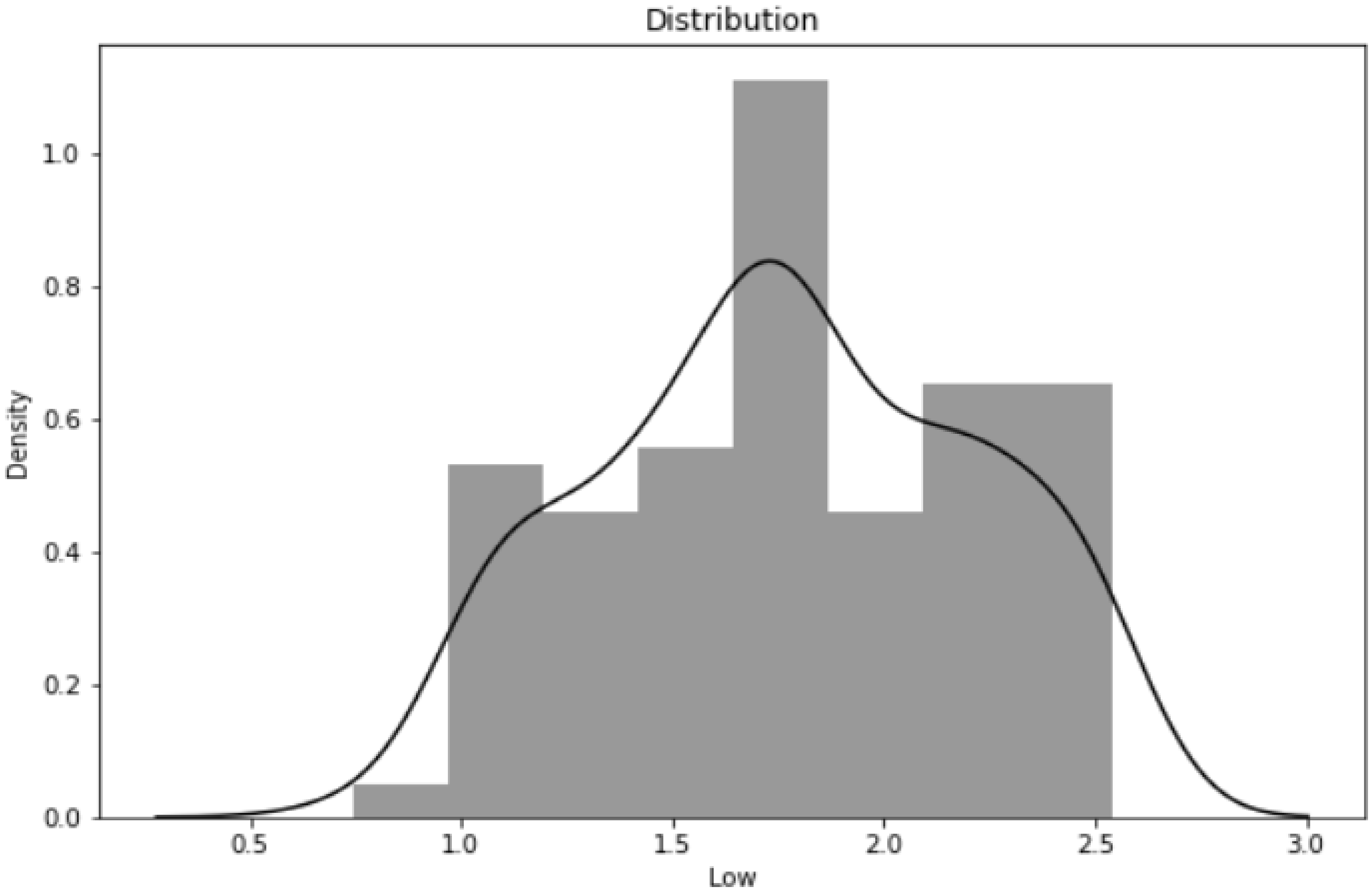



Distribution
Density
Close
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.5
1.0
1.5
2.0
2.5
3.0

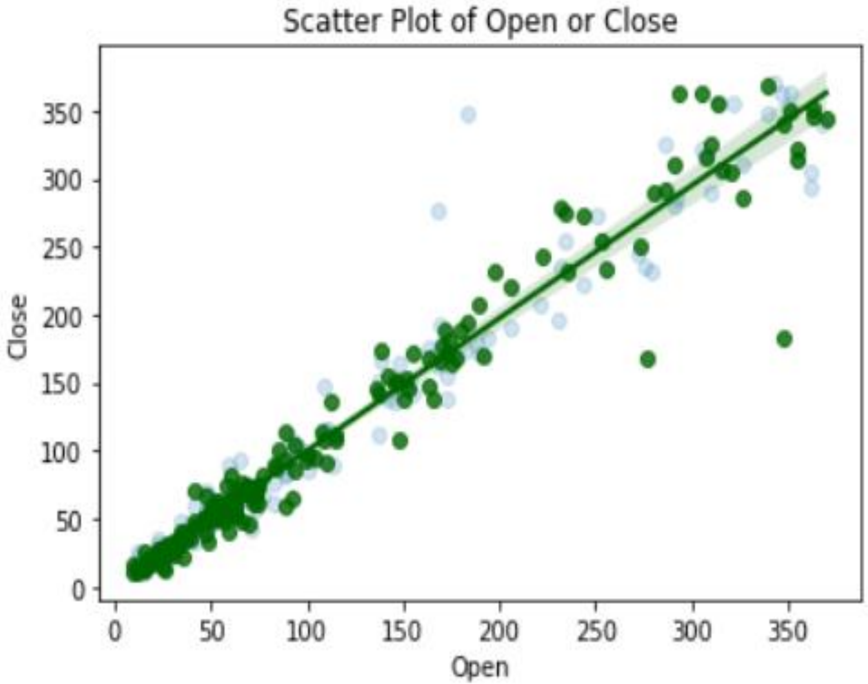
Visualization of Independent Variable Open , High and Low price of stock:

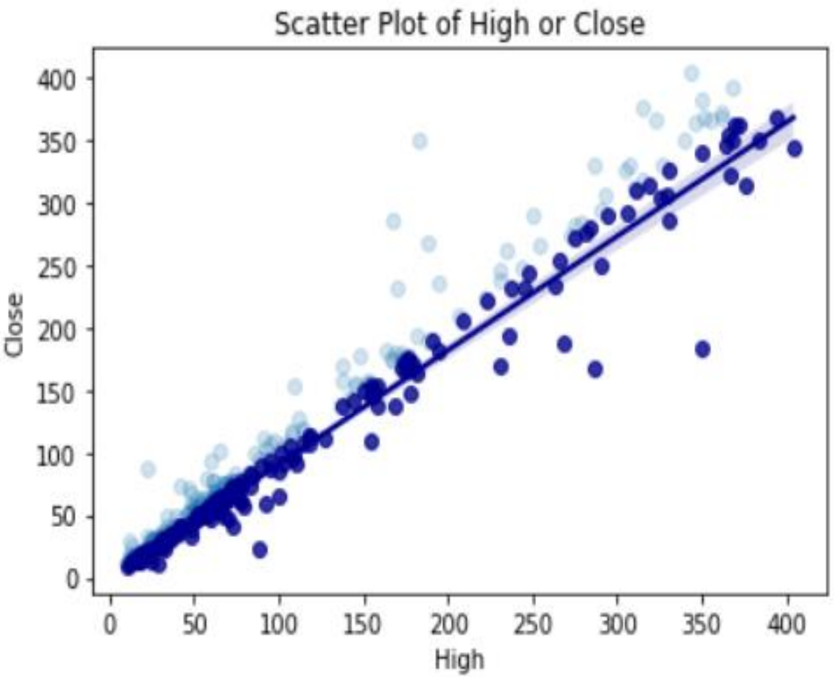


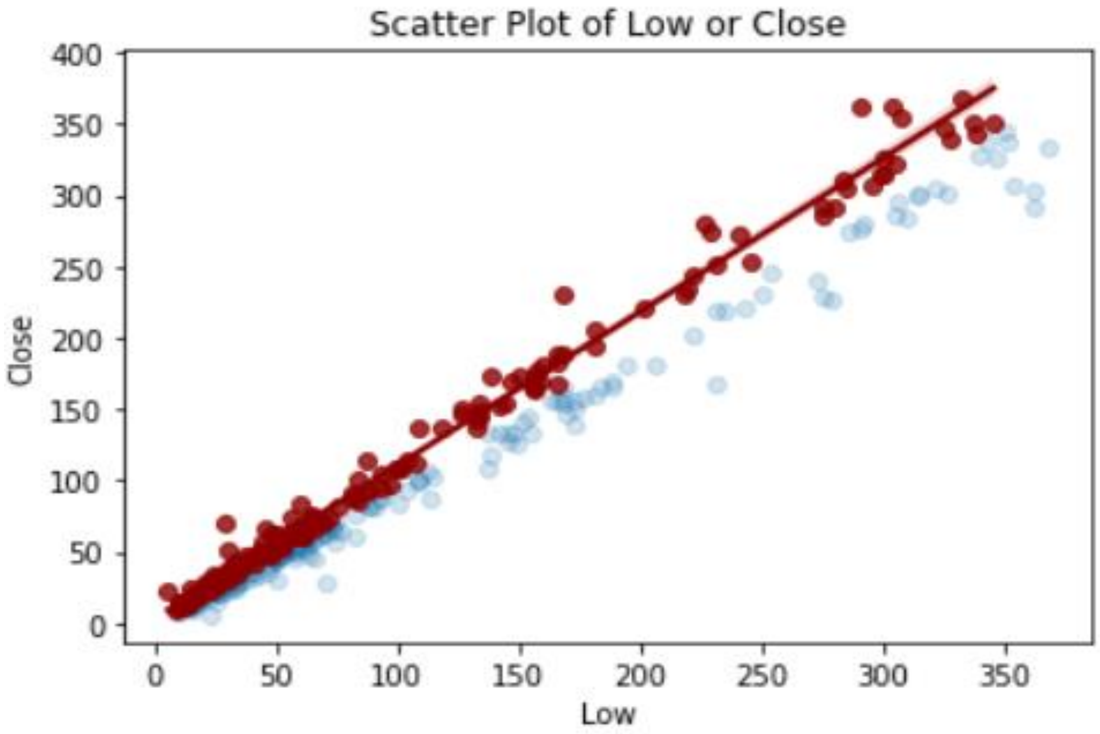

Distribution
Density
High
0.0
0.2
0.4
0.6
0.8
1.0
1.2
0.5
1.0
1.5
2.0
2.5
3.0


Distribution
Density
Low
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.0
1.5
2.0
2.5
3.0

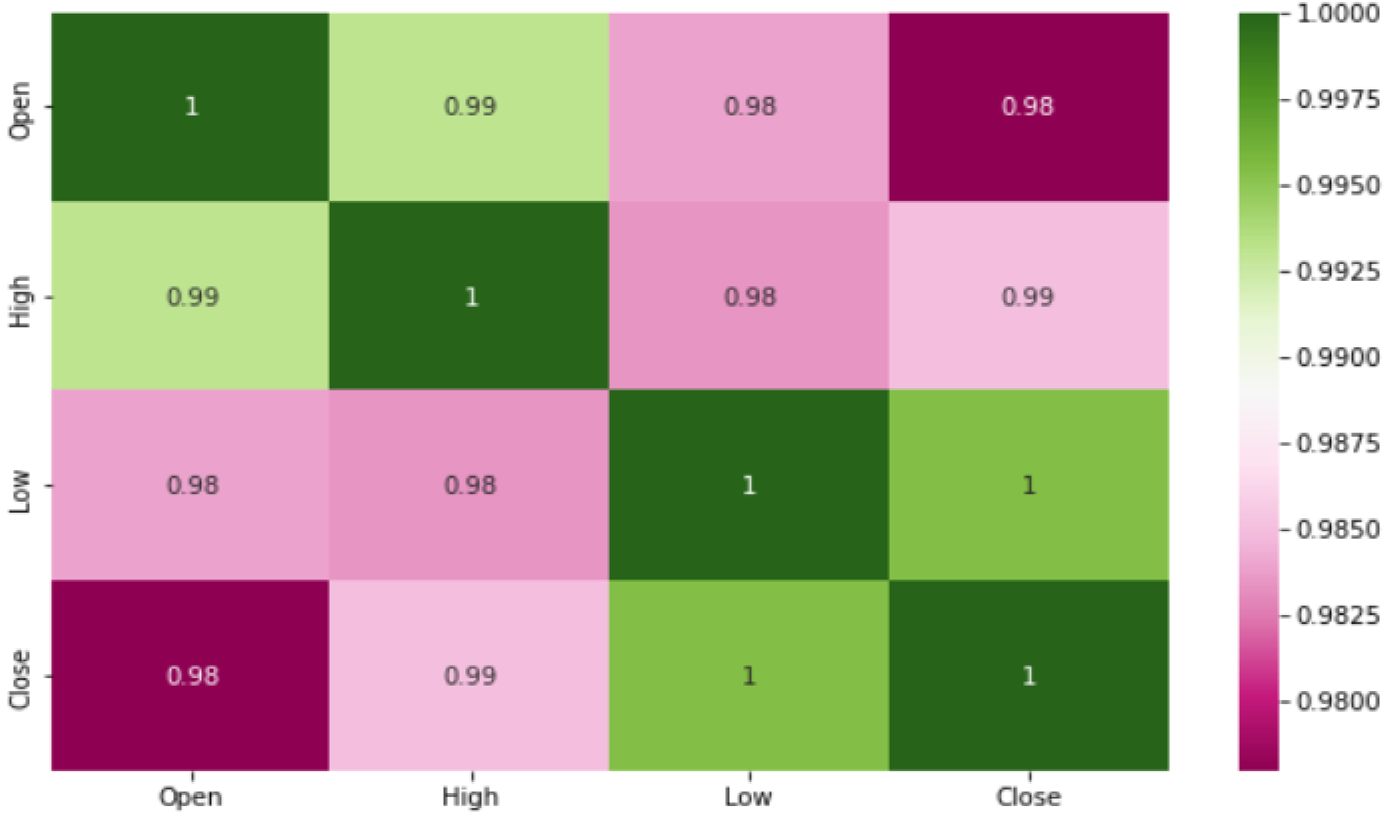
Relation between the Dependent Variable and Independent Variable:

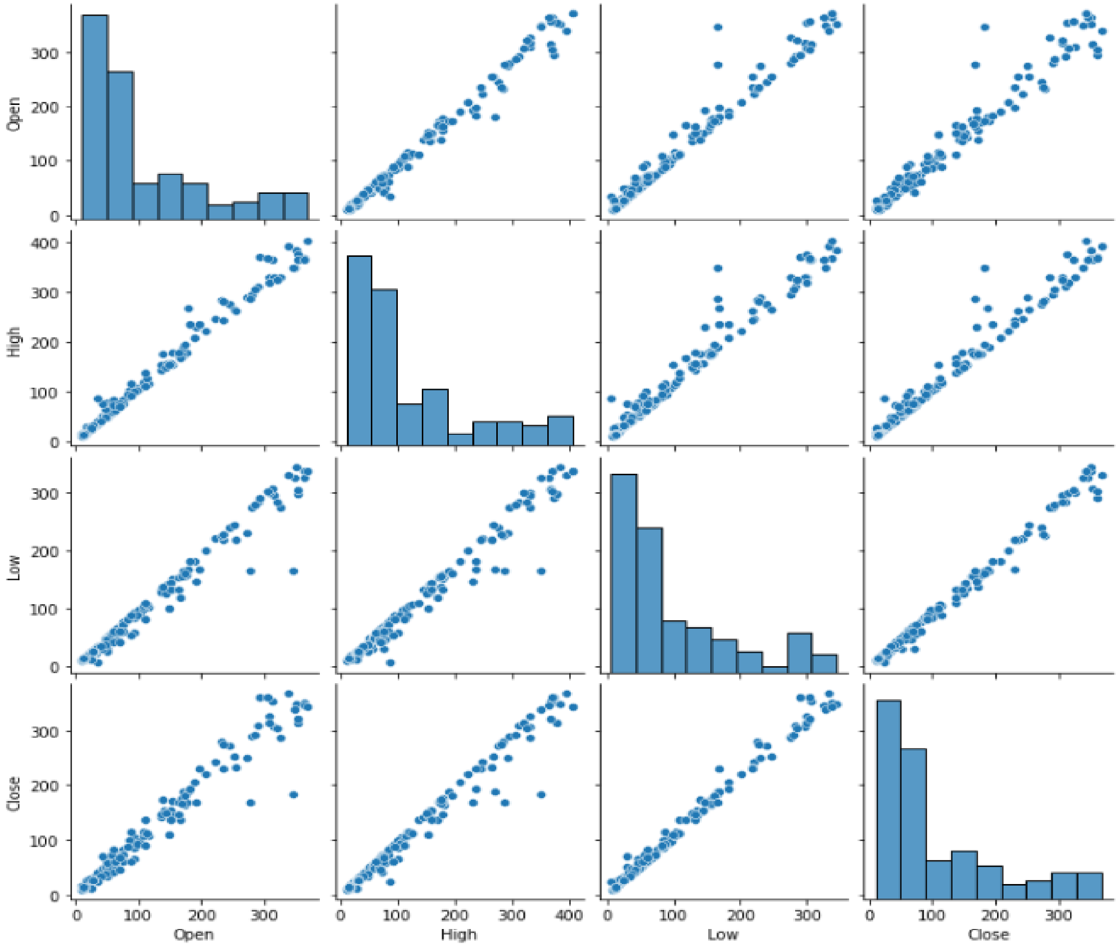





Scatter Plot of Low or Close
Close
Low
400
350
300
250
200
150
100
50
0
0
50
100
150
200
250
300
350

Correlation:





Visualization of every single column of our Df against every other column.

Multicollinearity:

Even though we have strong VIF ratings, we won't do feature engineering because each feature is critical for this specific use case. Most indicators in the real world consider each of these characteristics to predict future values.

Due to the fact that each column is equally crucial for prediction, we are not deleting any columns.

Column removal resulted in the loss of important data (features) that are necessary for the model to make correct predictions. It produces a poor model.

Therefore, we are not removing any features from the dataset while we attempt to predict the outcome, assess the model's performance with respect to multicollinearity, and make adjustments as necessary.

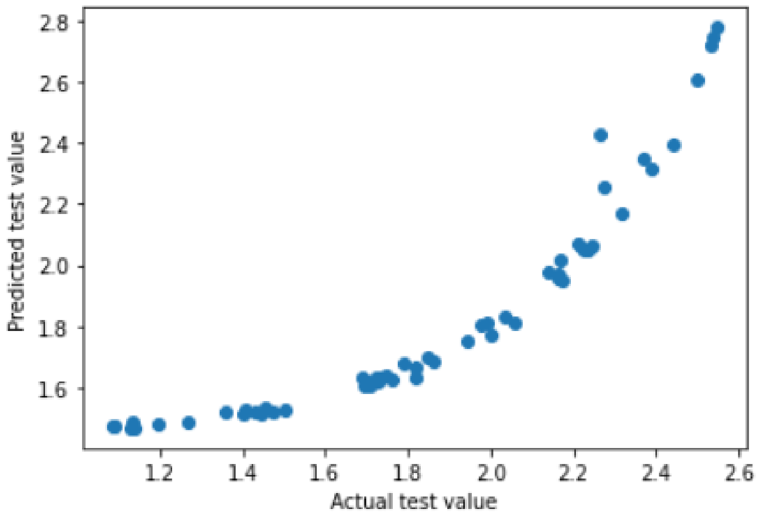
	Variables	VIF
0	Open	175.185704
1	High	167.057523
2	Low	71.574137

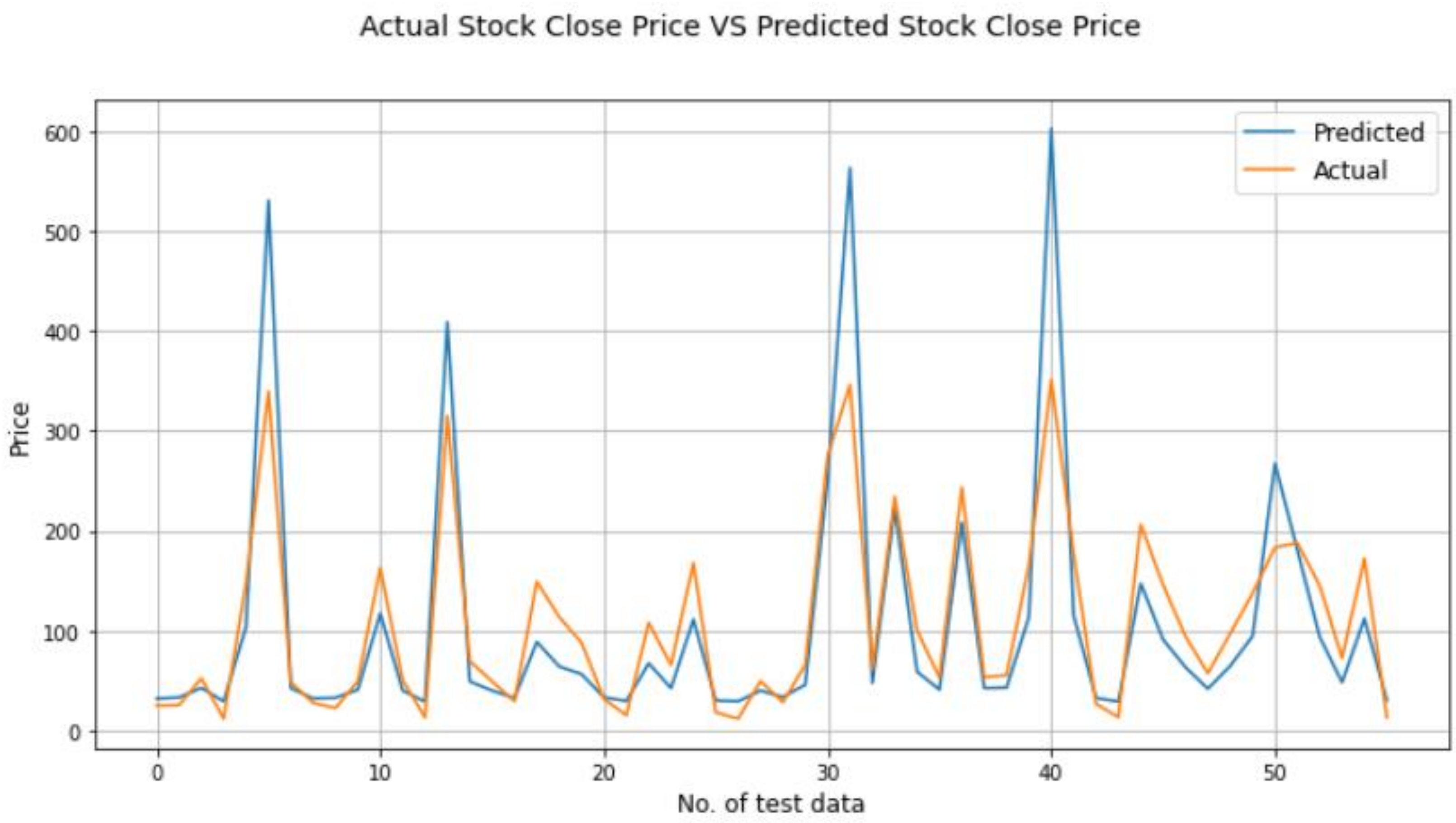
Regression Model:

Linear Regression:

Linear regression is the most basic machine learning approach that can be applied to this data. The result of the linear regression model is an equation showing how the independent variables a nd dependent variable related to each other.

```
Performance of Linear Regression Model
MSE : 0.0329
RMSE : 0.1814
MAE : 0.1594
MAPE : 0.0964
R2 : 0.8103
```



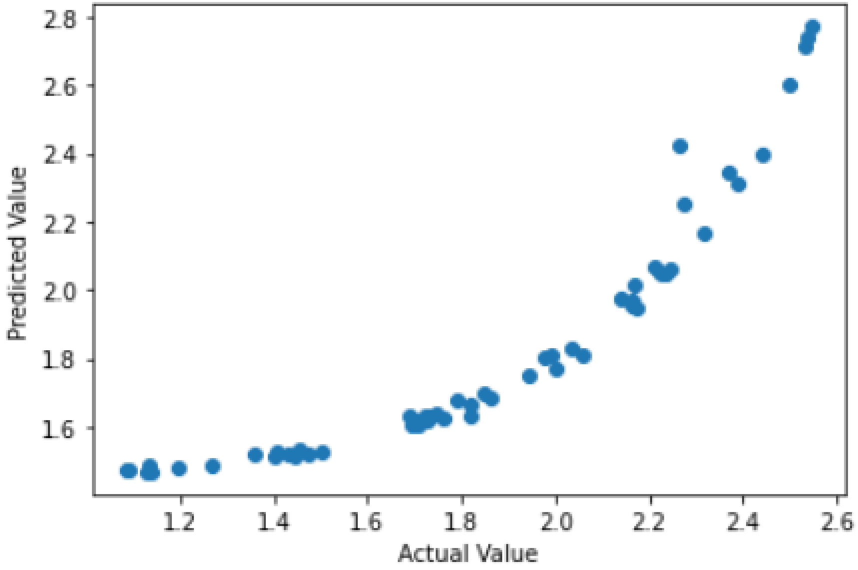


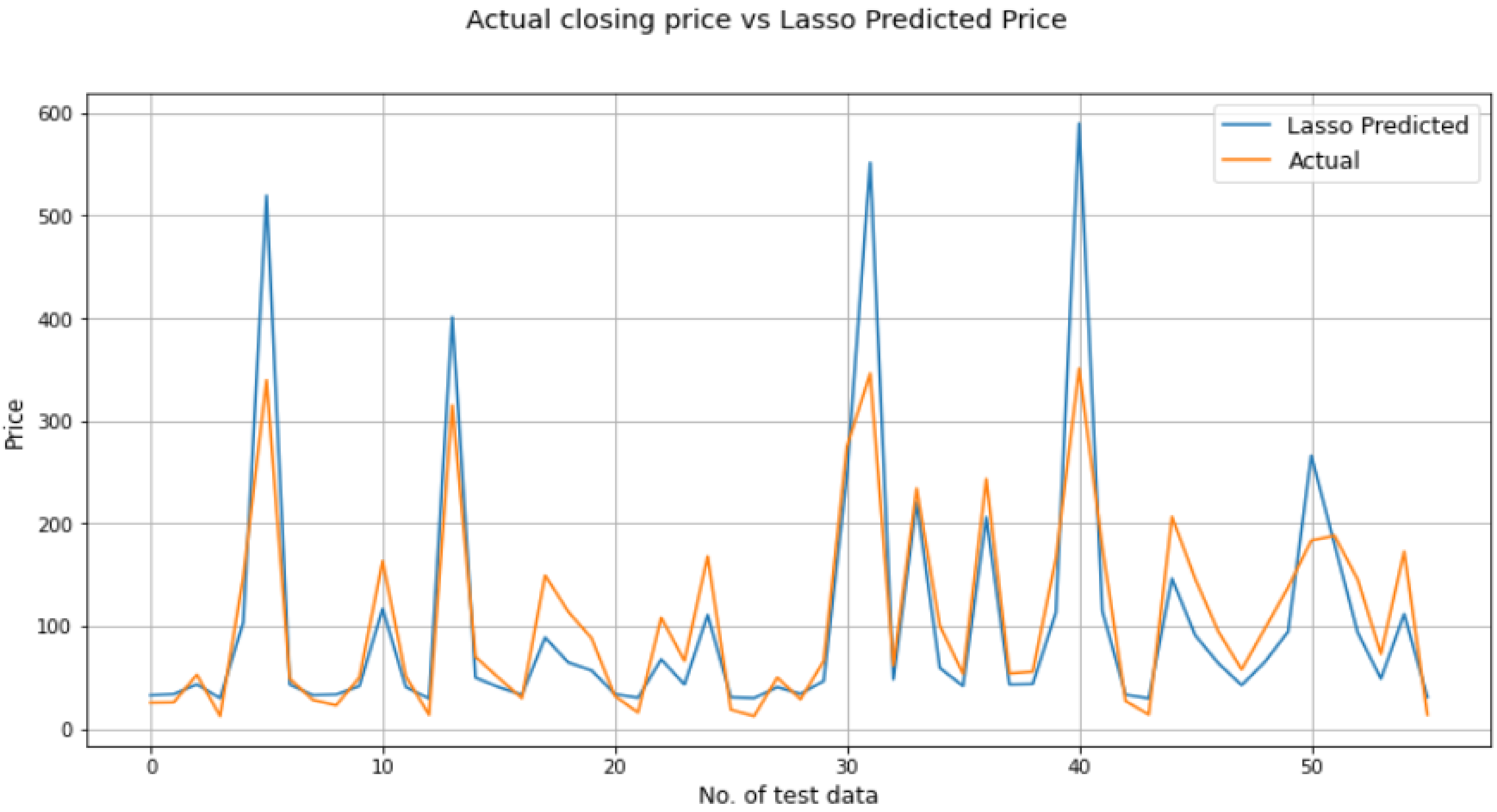
Actual Stock Close Price VS Predicted Stock Close Price
Predicted
Actual
Price
No. of test data
0
100
200
300
400
500
600
0
10
20
30
40
50

Lasso Regression:

Lasso(least absolute shrinkage and selection operator) regression is another technique of Parameter estimation regression method. This method is usually used in machine learning for the selection of the subset of variables. It provides greater prediction accuracy as compared to other regression models. Lasso Regularization enhances the accessibility of models.

```
Performance of Lasso Regression Model
MSE : 0.0331
RMSE : 0.1818
MAE : 0.1598
MAPE : 0.0968
R2 : 0.8094
```

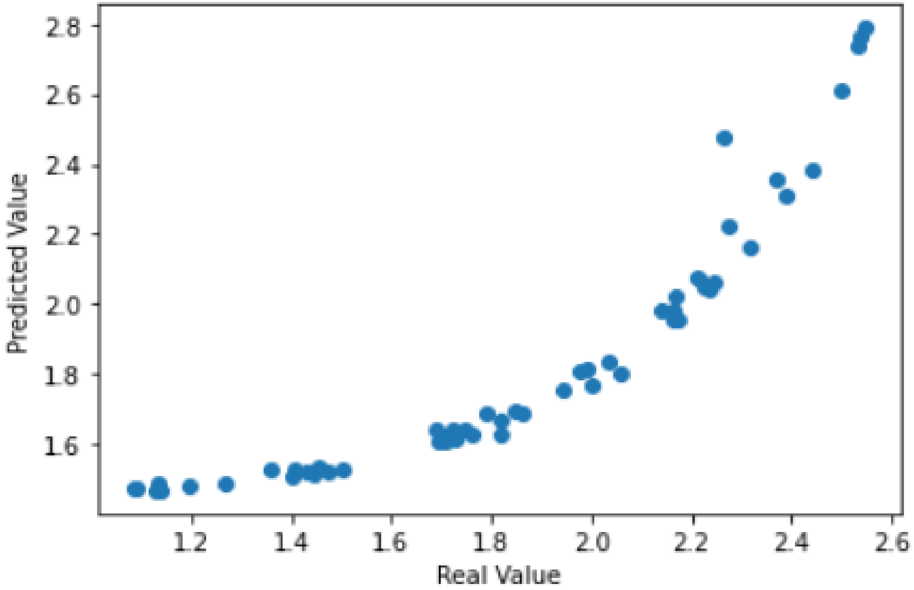


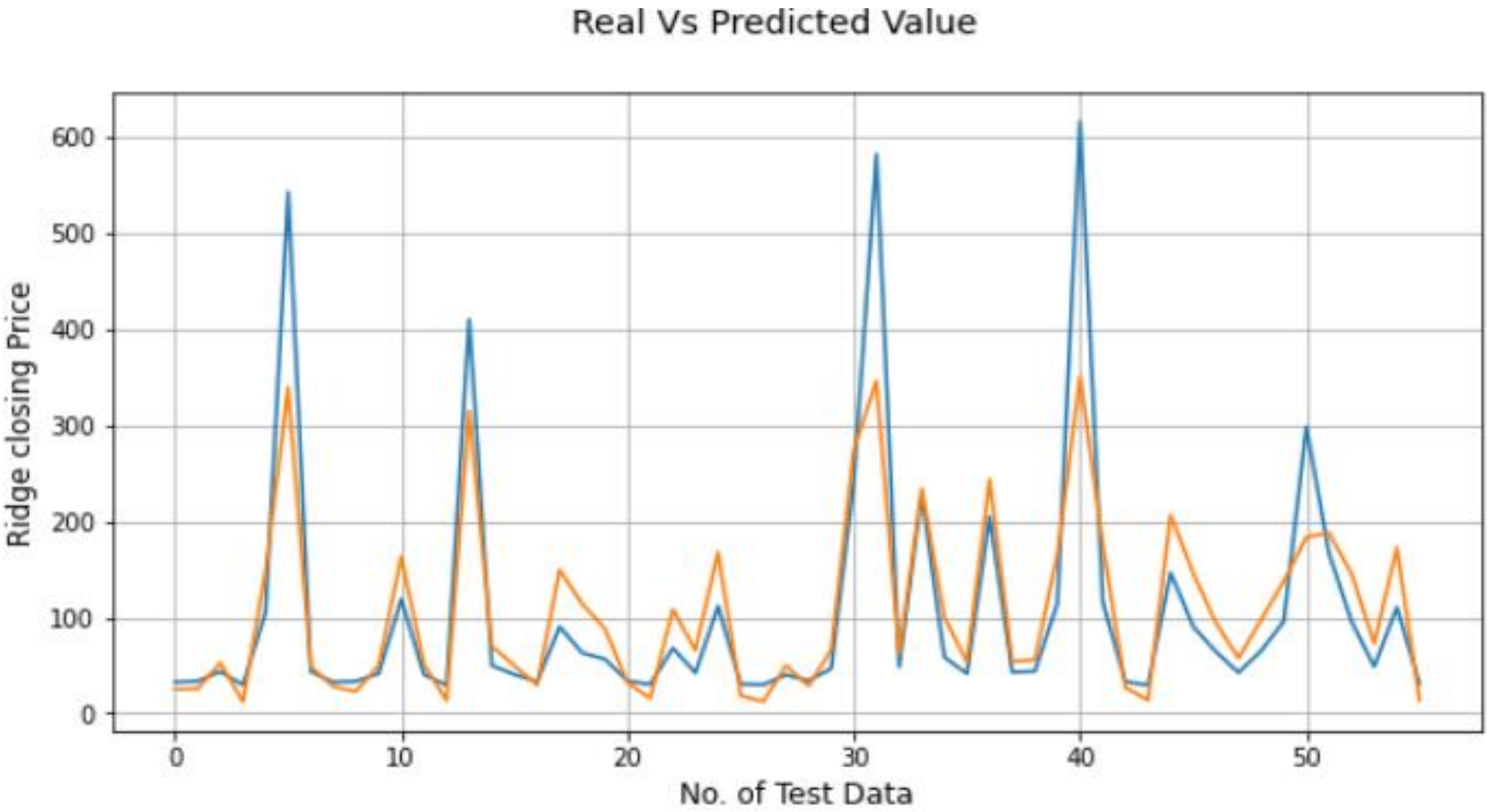

Actual closing price vs Lasso Predicted Price
Lasso Predicted
Actual
Price
No. of test data
0
100
200
300
400
500
600
0
10
20
30
40
50

Ridge Regression:

Ridge regression is a model-tuning technique that is used to analyse any multicollinear data. L2 regularization is done using this technique. The projected values vary significantly from the actual values when the problem of multicollinearity is present, least-squares are unbiased, and variances are large.

```
Performance of Ridge Regression Model
MSE : 0.0337
RMSE : 0.1835
MAE : 0.1614
MAPE : 0.0973
R2 : 0.8058
```

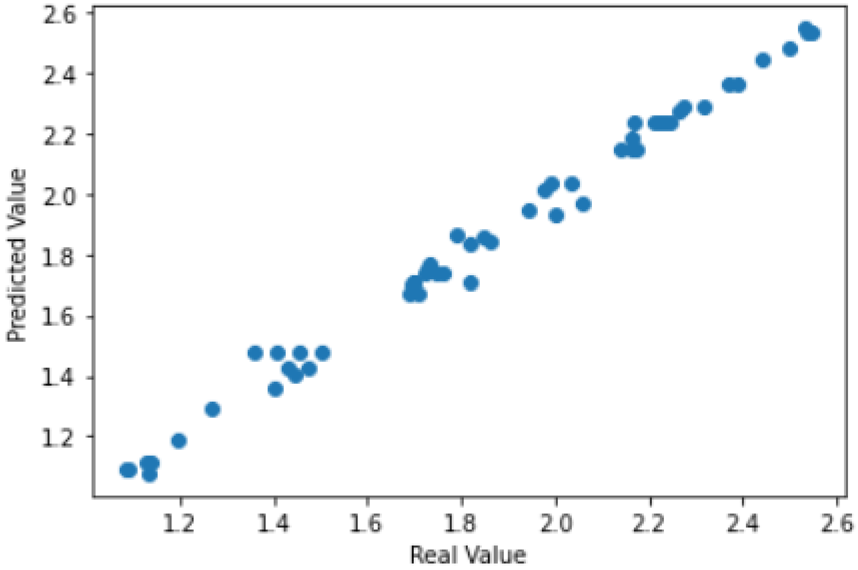



Real Vs Predicted Value
600
500
400
300
200
100
0
Ridge closing Price
0
10
20
30
40
50
No. of Test Data

Decision Tree Regression:

Decision tree regression trains a model in the form of a tree to predict data in the future and generate useful continuous output by observing the properties of an item.

```
Performance of Decision tree Regression Model
MSE : 0.002
RMSE : 0.0447
MAE : 0.0308
MAPE : 0.0175
R2 : 0.9885
```




Real Vs Predicted Close Price
Predicted
Actual
Decision tree Closing Price
Test Data number
0
50
100
150
200
250
300
350
0
10
20
30
40
50

Conclusion:

1. The trend of the price of Yes Bank's stock increased until 2018 and then Close, Open, High, Low price decreased.
2. Based on the open vs. close price graph, we concluded that Yes Bank's stock fell significantly after 2018.
3. Both duplicate and null values are absent, as we have seen. But object data type values are available for the Date feature. Therefore, we transformed it to the correct date format, YYYY-MM-DD.
4. The dependent and independent values were found to be linearly related.
5. The data contained a significant amount of multicollinearity.

6.Decision Tree regression Is best model for yes bank stock closing price data this model use for further prediction

7. Visualization has allowed us to notice that the closing price of the stock has suddenly fallen starting in 2018. It seems reasonable that the Yes Bank stock price was significantly impacted by the Rana Kapoor case fraud.

In this work, we create 5 regression models for our data:-

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Decision Tree Regression

These four models gives us the following results: High, Low, Open are directly correlate with the closing price of stocks.

Thank You