# CAPSTONE PROJECT 2
# REGRESSION MODEL

**TITLE : Yes Bank Stock Prediction**

Prepared By⟶ Praveen.S

(COHORT EVEREST)

# Let's go through the Defaulters:

1. Defining the Problem Statement
2. EDA and feature selection
3. Data Visualization
4. Normalization of data
5. Preparing dataset for modeling
6. Applying the model
7. Cross Validation and Hyperparamter Tuning

# Problem Statement

Yes Bank is a well known private sectored bank in the indian financial domain which was founded in 2004 headquartered in mumbai.Since 2018,it has been in the news because of the fraud case involving the founders.Owing to the fact,it was interesting to see the how that impacted the stock price of the company.The main objective is to predict the stock's closing price of the upcoming months.

# Dataset Understanding:

## COLUMNS:

- **Date:** It denotes the Month and Year of each Observation
- **Open:** It denotes the starting stock value of that month
- **High:** It denotes the highest stock value of that month
- **Low:** It denotes the lowest stock value of that month
- **Close:** It denotes the Closing stock value of that month

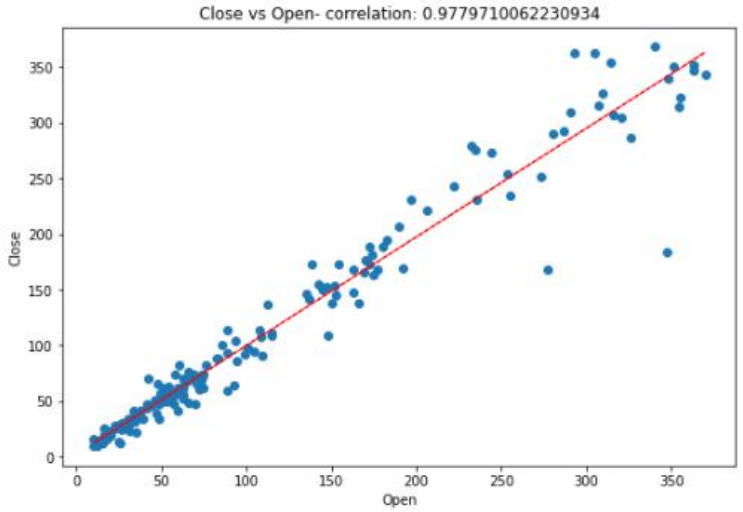
# Data Preprocessing:

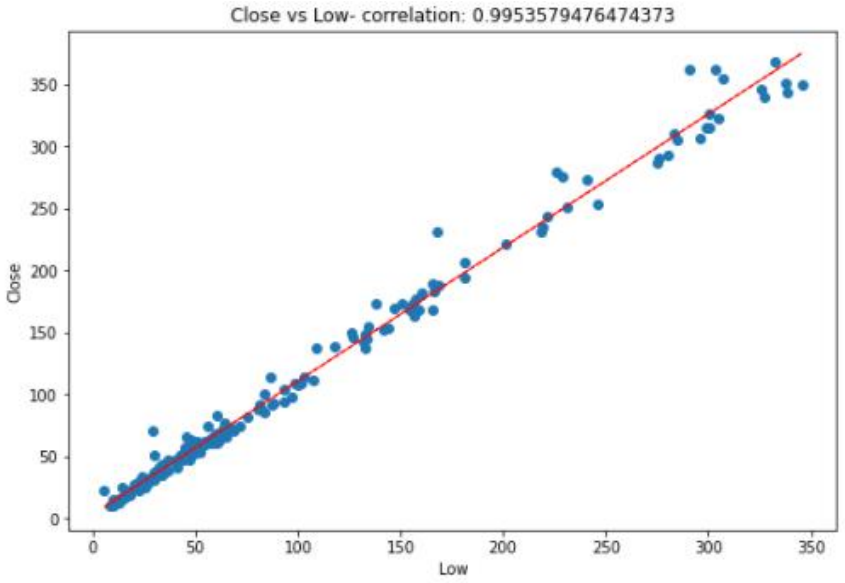
- No Duplicate values were found in this data
- Dropped the “Date” column as the target value is not dependent of it
- There is neither missing values nor null values in the dataset

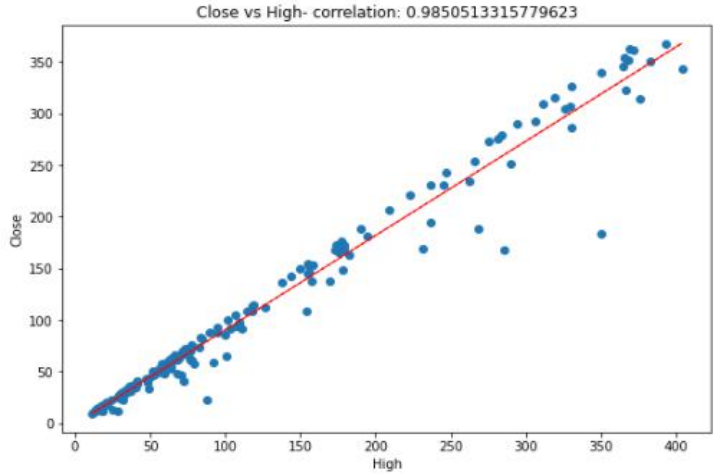
# Data Visualization:

- Correlation of independent features with the target variable.
- Distribution of all features in the dataset
- Distribution of each  features with their mean and median value
- Correlation of all features using a heat map
- Comparison of Starting  and closing stock value of last 3 years
- Comparison of High and low stock value of the last 3 years

# Correlation of each features with the target feature:

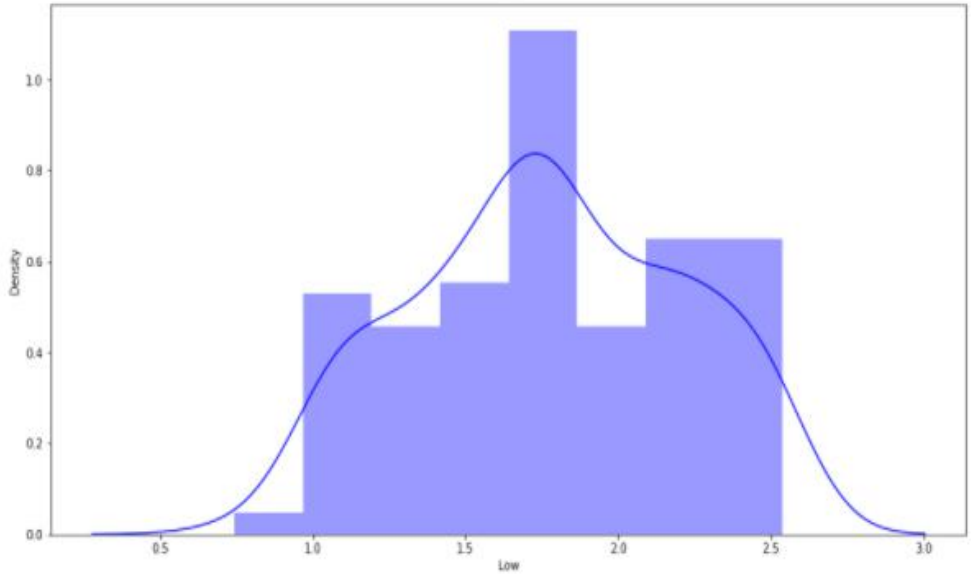


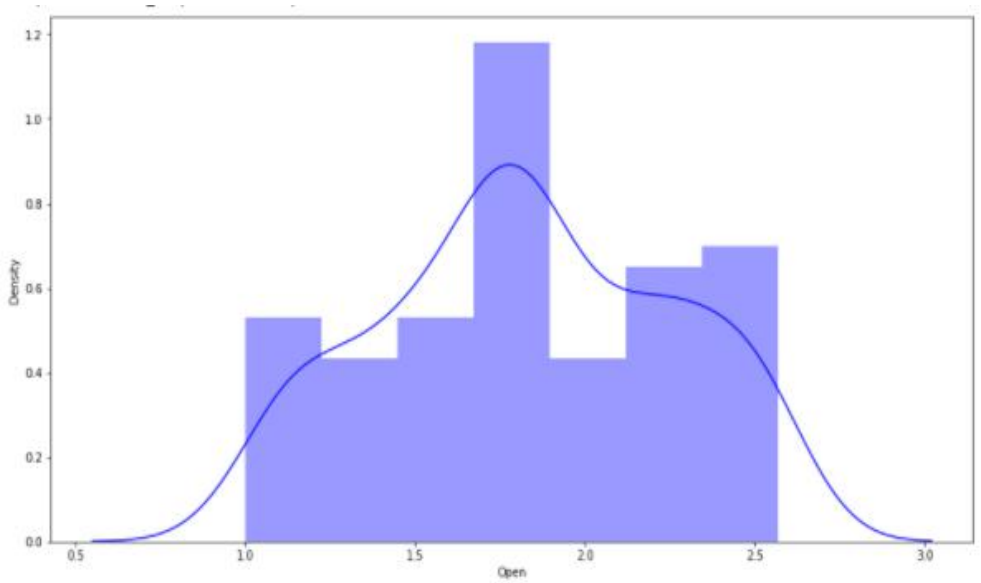



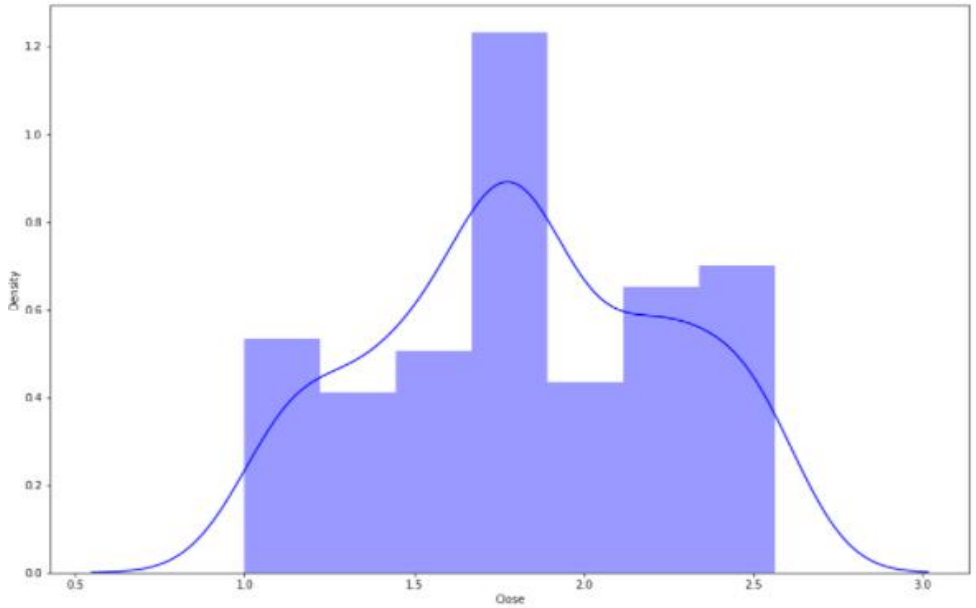



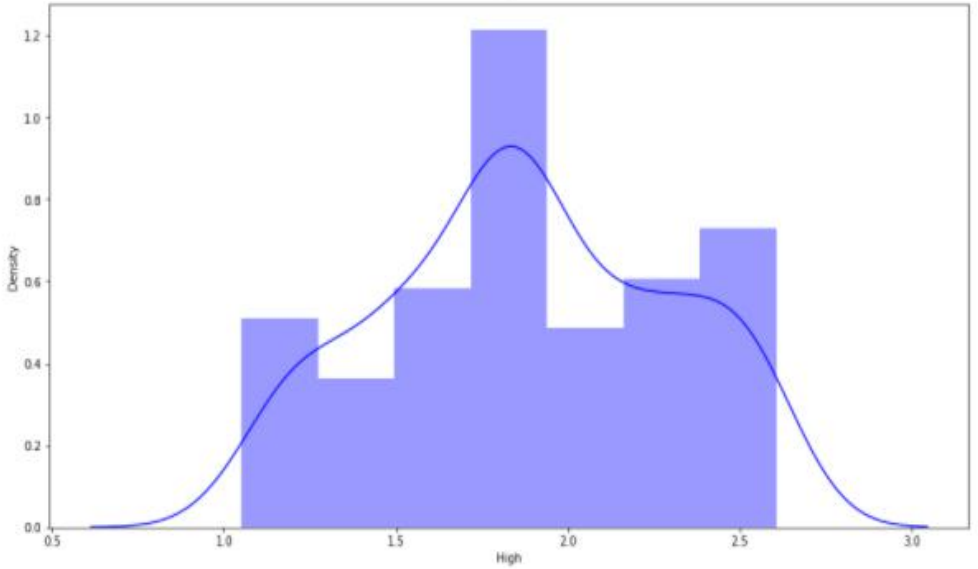
# Distribution of all the features:

All of them are normally distributed

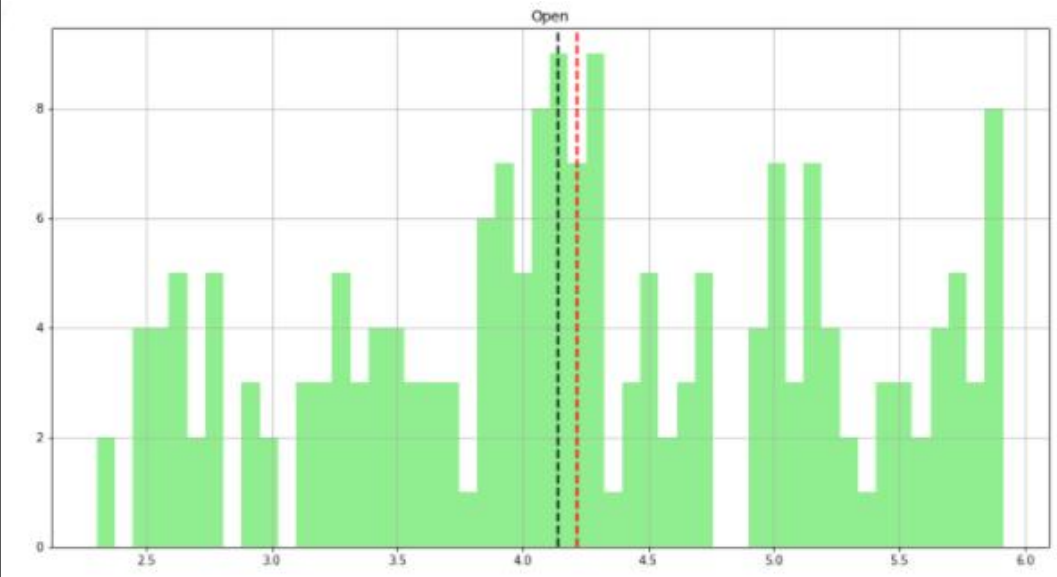


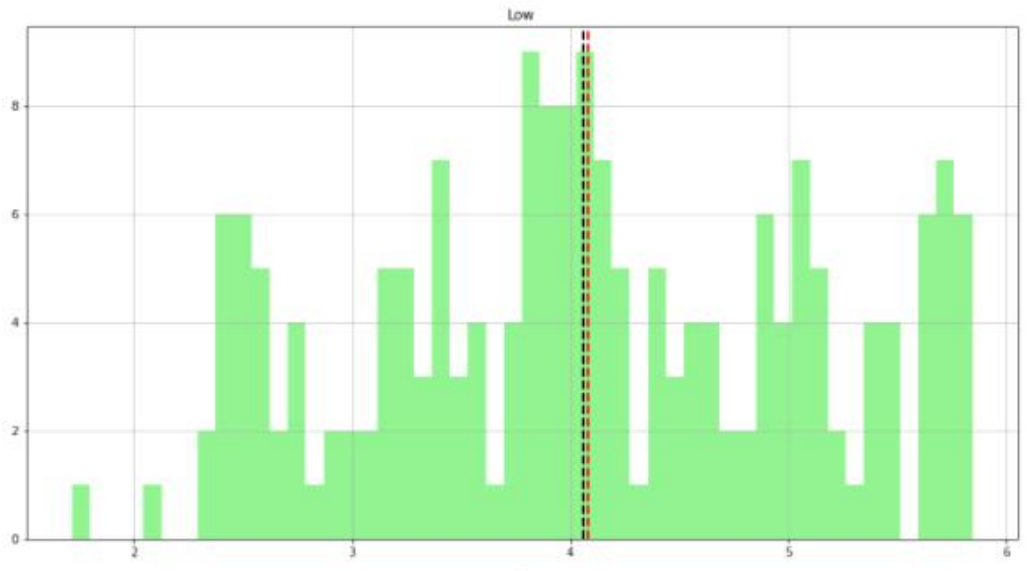



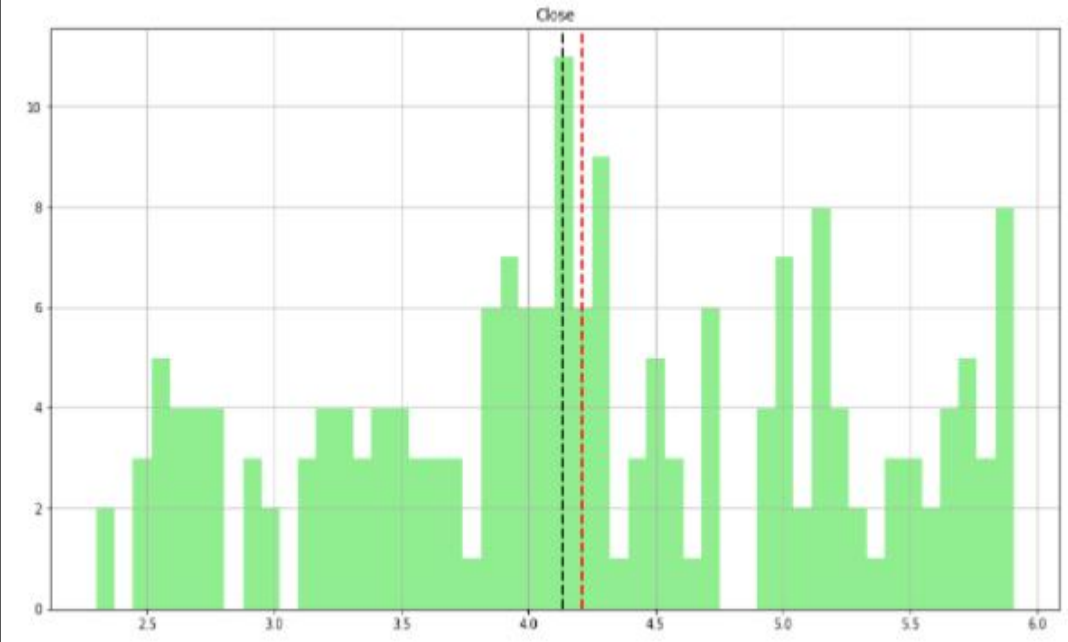



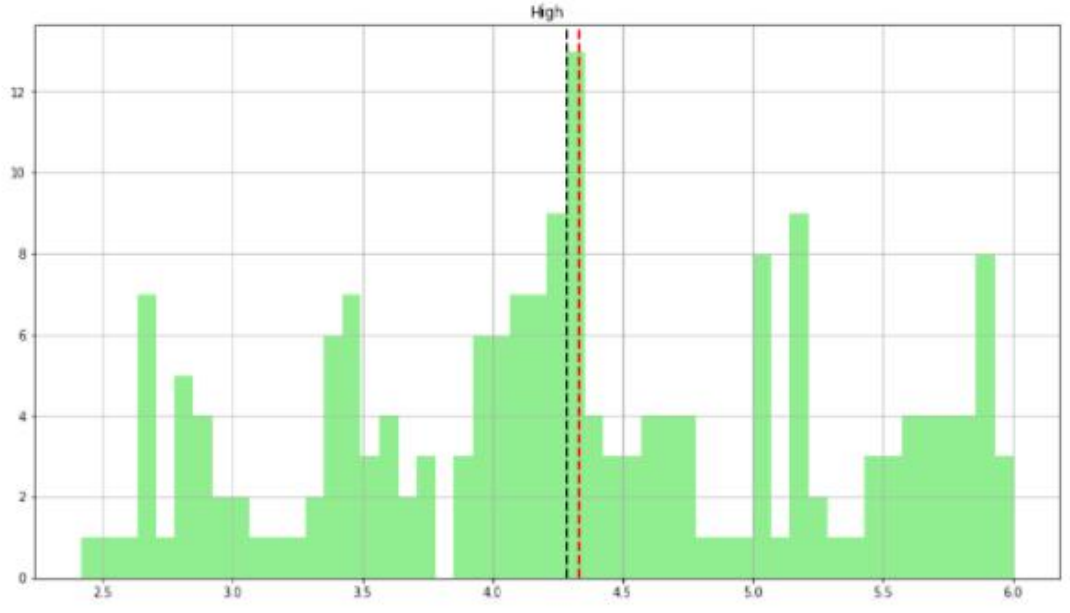




**Since all of them are normally distributed, the mean and median happens to be placed closely**



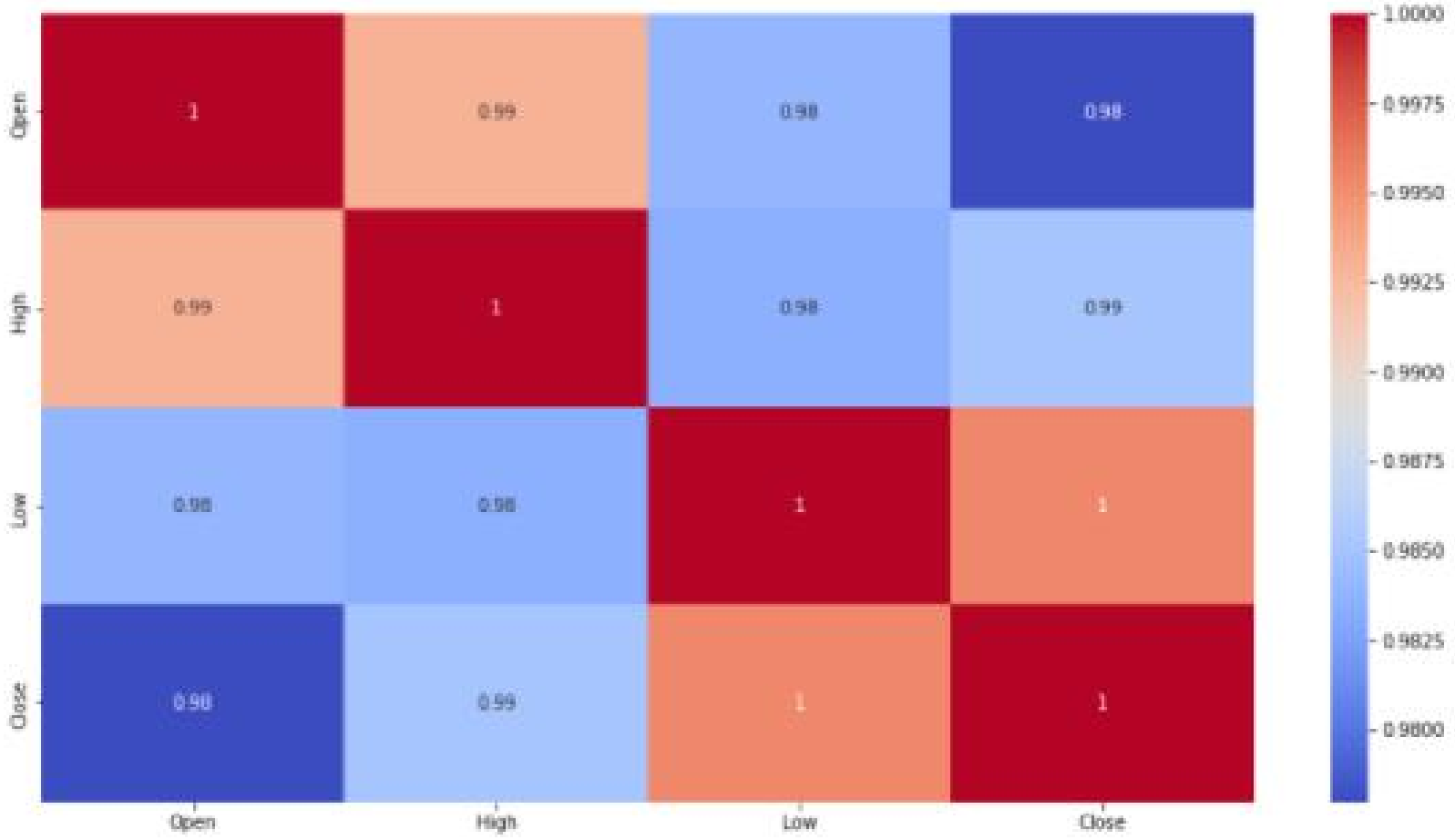



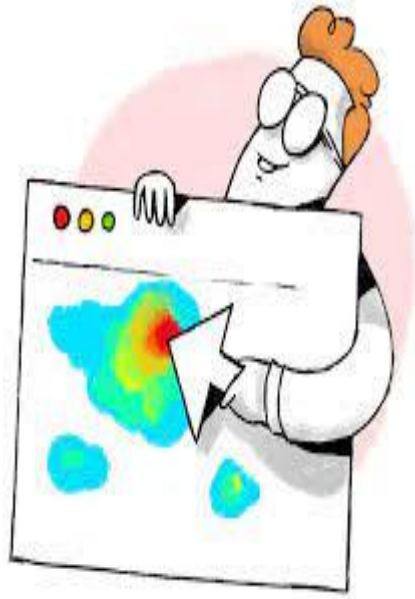



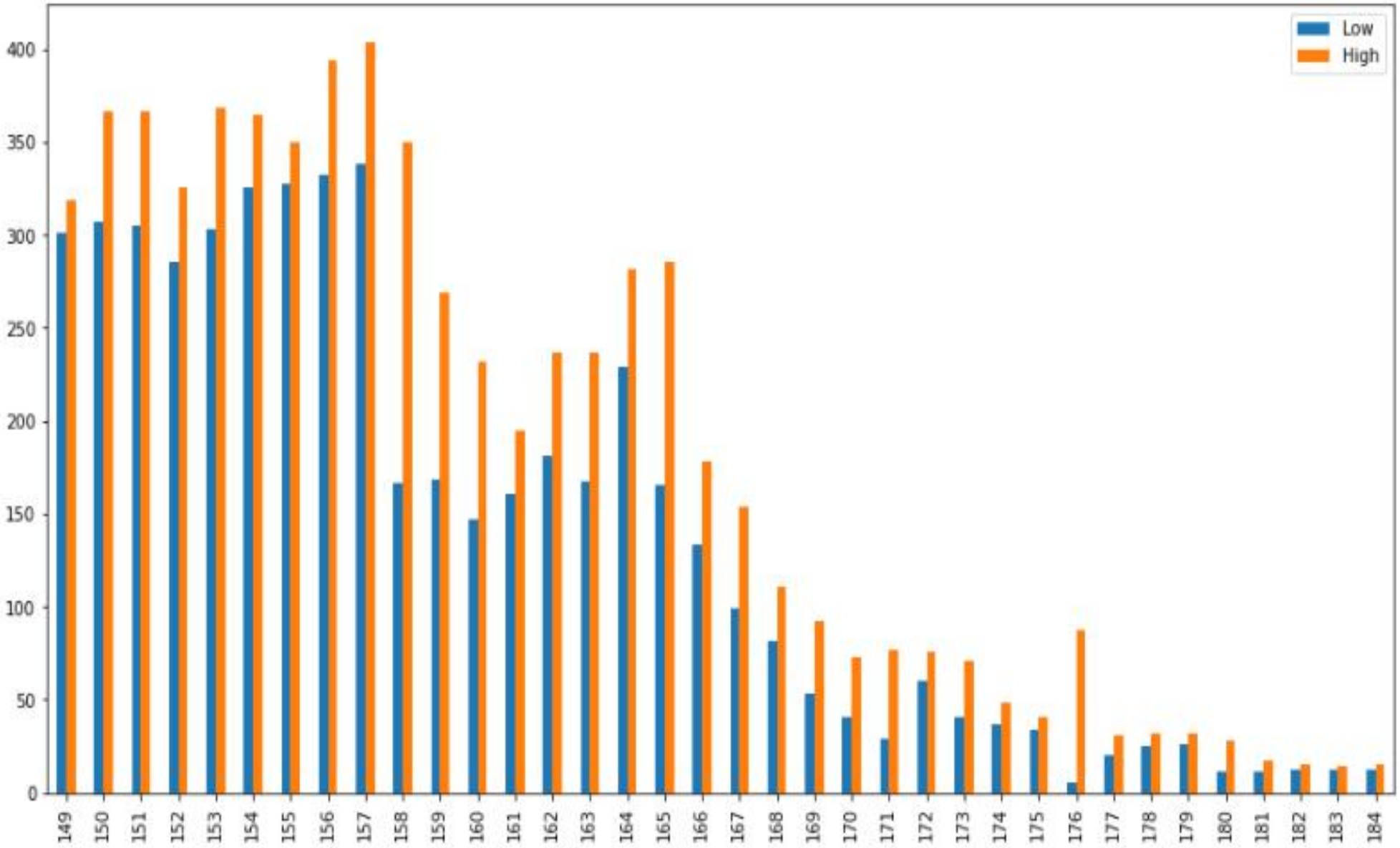
# Correlation Map:

As per the heat map below,we can easily say that the dataset is so correlated as its features collinearity are approximately equal to 1

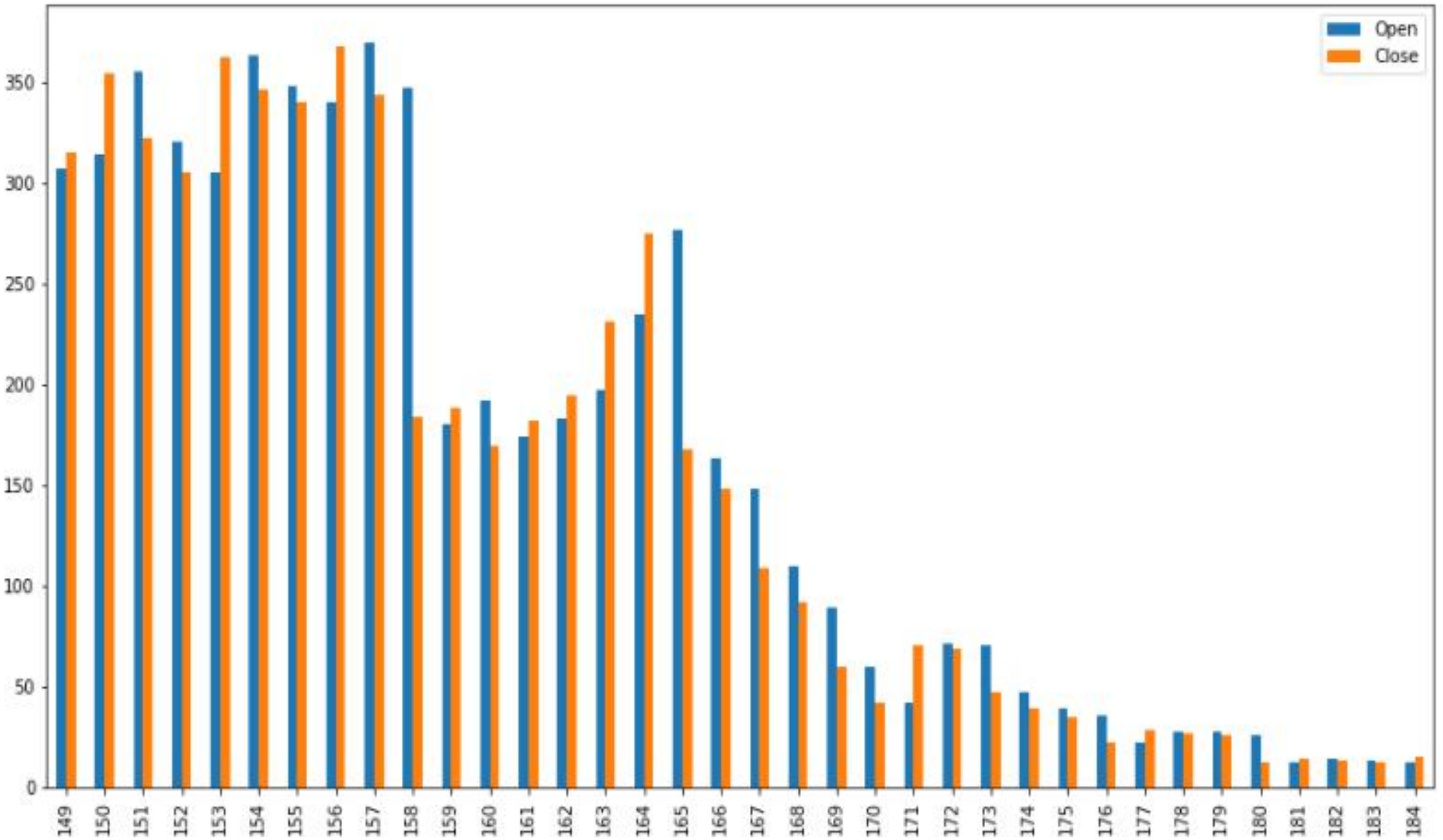


# Last 3 years highest and lowest stock values:



# Last 3 years starting and Closing stock value:



# Preparing dataset for modeling:

| | 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 |

train_test_split:

X_train:-(148, 3)

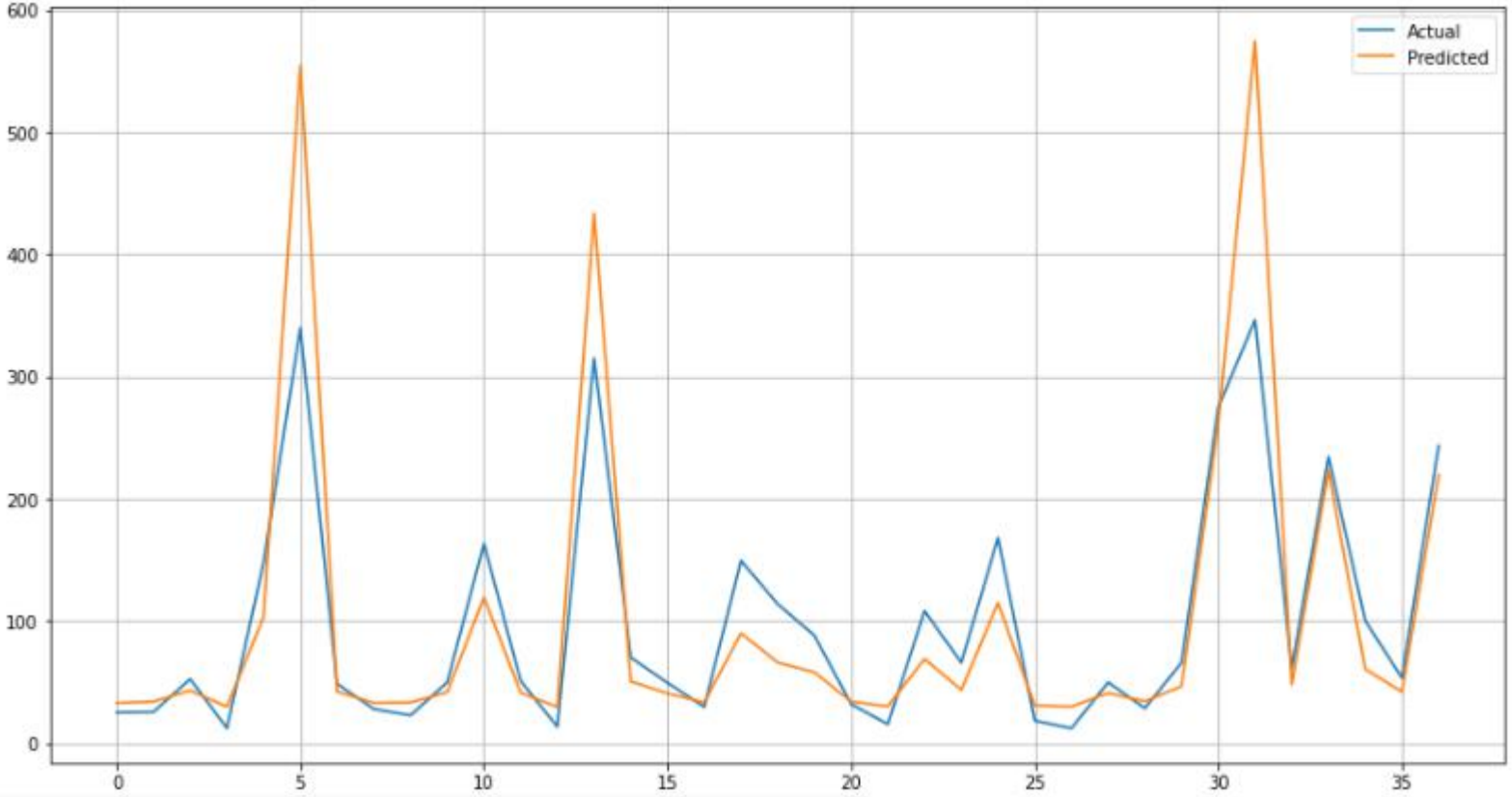
X_test:-(37, 3)

y_train:-(148,1)

y_test:-(37,1)

# Baseline model:

## Linear Regression




_Train Performance_
Mean Squared Error : 0.033659395765946 67
Root Mean Squared Error : 0.1834649714957781
Mean Absolute Error : 0.15591697668200555
R2 : 0.8146653424416905

_Test Performance_
Mean Squared Error : 0.031582518930487385
Root Mean Squared Error : 0.1777147121948191
Mean Absolute Error : 0.15128511034606282
R2 : 0.8225699915389754

Train Accuracy:
0.81466.....

Test Accuracy:
0.82256.....

# Model Validation & Selection:

| | Model | MAE | MSE | RMSE | R2 | RMSLE | MAPE | TT (Sec) |
|---|---|---|---|---|---|---|---|---|
| lr | Linear Regression | 4.0323 | 48.0047 | 6.4864 | 0.9920 | 0.0792 | 0.0584 | 0.029 |
| ridge | Ridge Regression | 5.3222 | 94.1389 | 8.8923 | 0.9854 | 0.0910 | 0.0696 | 0.024 |
| lasso | Lasso Regression | 6.0199 | 111.1297 | 9.7552 | 0.9817 | 0.1031 | 0.0815 | 0.028 |
| en | Elastic Net | 15.2899 | 459.0094 | 19.9194 | 0.9447 | 0.3082 | 0.2979 | 0.027 |

# Observations:

**Observation 1:**

As seen in the above tables,Linear Regression is giving great results on the basis of R2 value and least interpretability.

**Observation 2:**

We can able to perform lasso and ridge regression as they support continous values significantly and also both have similarly performed well when compared to their base line model.

**Observation 3:**

From the above observation we come to conclusion that we can also implement Elastic Net Regression as it is supported by both Ridge and Lasso regressions.

# Model Validation & Selection:

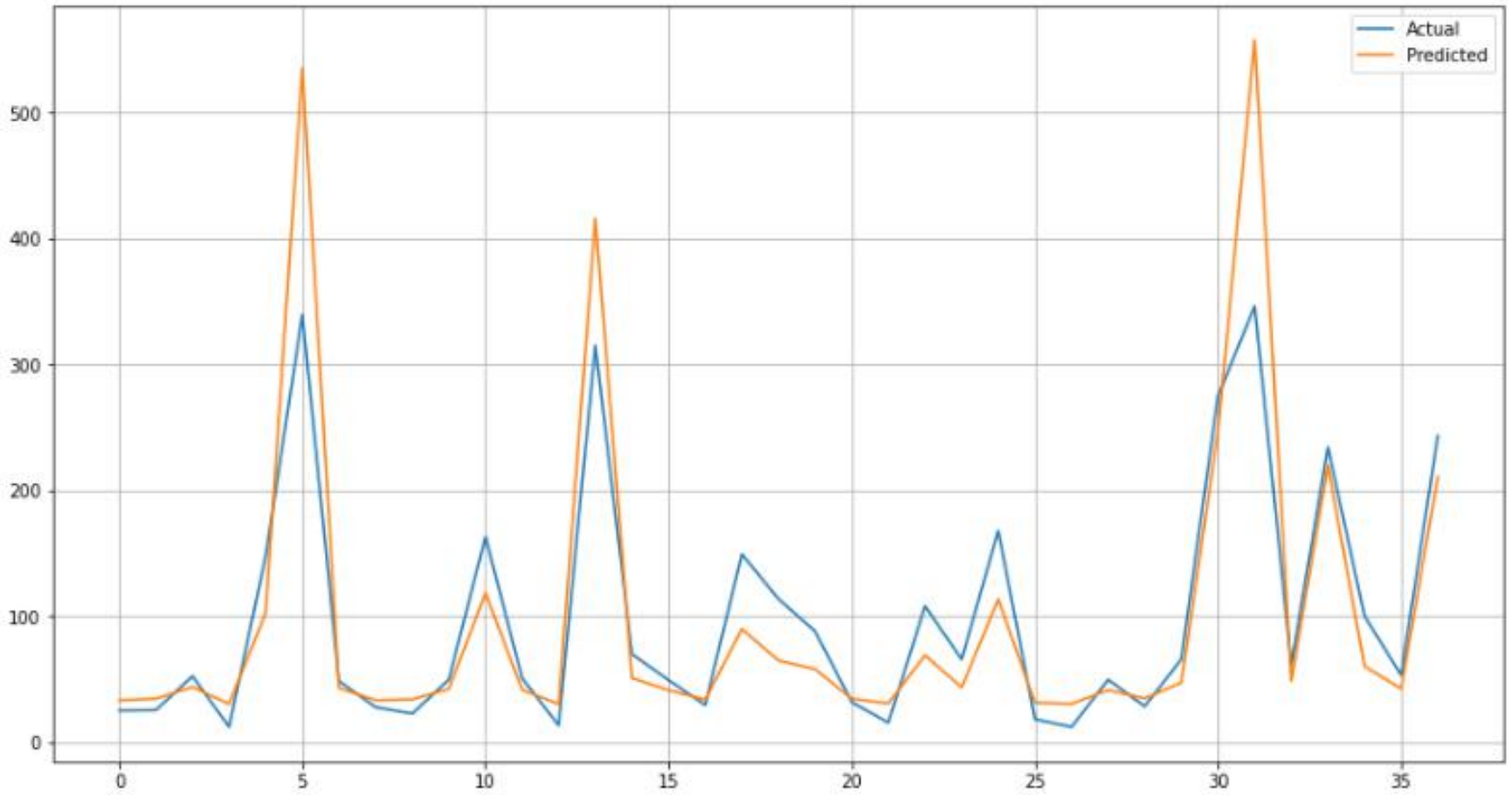
## Lasso Regression

```
GridSearchCV(cv=3, estimator=Lasso(),
             param_grid={'alpha': [1e-15, 1e-13, 1e-10, 1e-08, 1e-05, 0.0001,
                                   0.001, 0.01, 0.1, 1, 5, 10, 20, 30, 40, 45,
                                   50, 55, 60, 100, 0.0014]},
             scoring='neg_mean_squared_error')
```

```
The best fit alpha value is found out to be : {'alpha': 0.01}

Using  {'alpha': 0.01}  the negative mean squared error is:  -0.03515384844892758
```

# Lasso Regression (Actual vs Predicted)




MSE : 0.032222733360720 56
RMSE : 0.17950691730604856
MAE : 0.15277731058358668
R2 : 0.8189732786857935

# Model Validation & Selection:

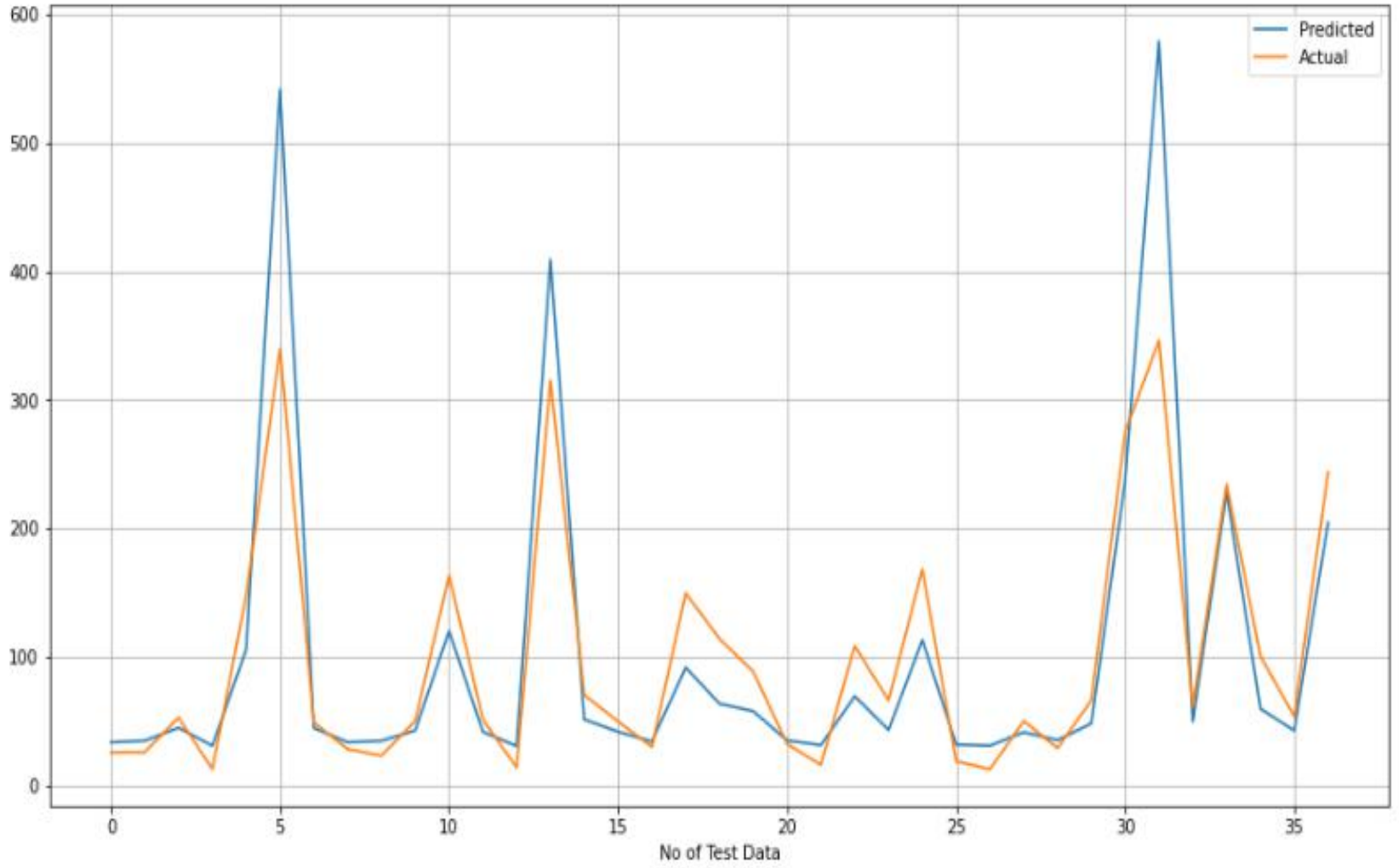
## Ridge Regression

```
GridSearchCV(cv=3, estimator=Ridge(),
             param_grid={'alpha': [1e-15, 1e-13, 1e-10, 1e-08, 1e-05, 0.0001,
                                   0.001, 0.01, 0.1, 1, 5, 10, 20, 30, 40, 45,
                                   50, 55, 60, 100]},
             scoring='neg_mean_squared_error')
```

```
The best fit alpha value is found out to be : {'alpha': 10}

Using  {'alpha': 10}  the negative mean squared error is:  -0.035198971562466846
```

# Ridge Regression (Actual vs Predicted)




MSE : 0.03253593988266965
RMSE : 0.18037721553086924
MAE : 0.15307727568266652
R2 : 0.8172136902260576

# Model Validation & Selection:

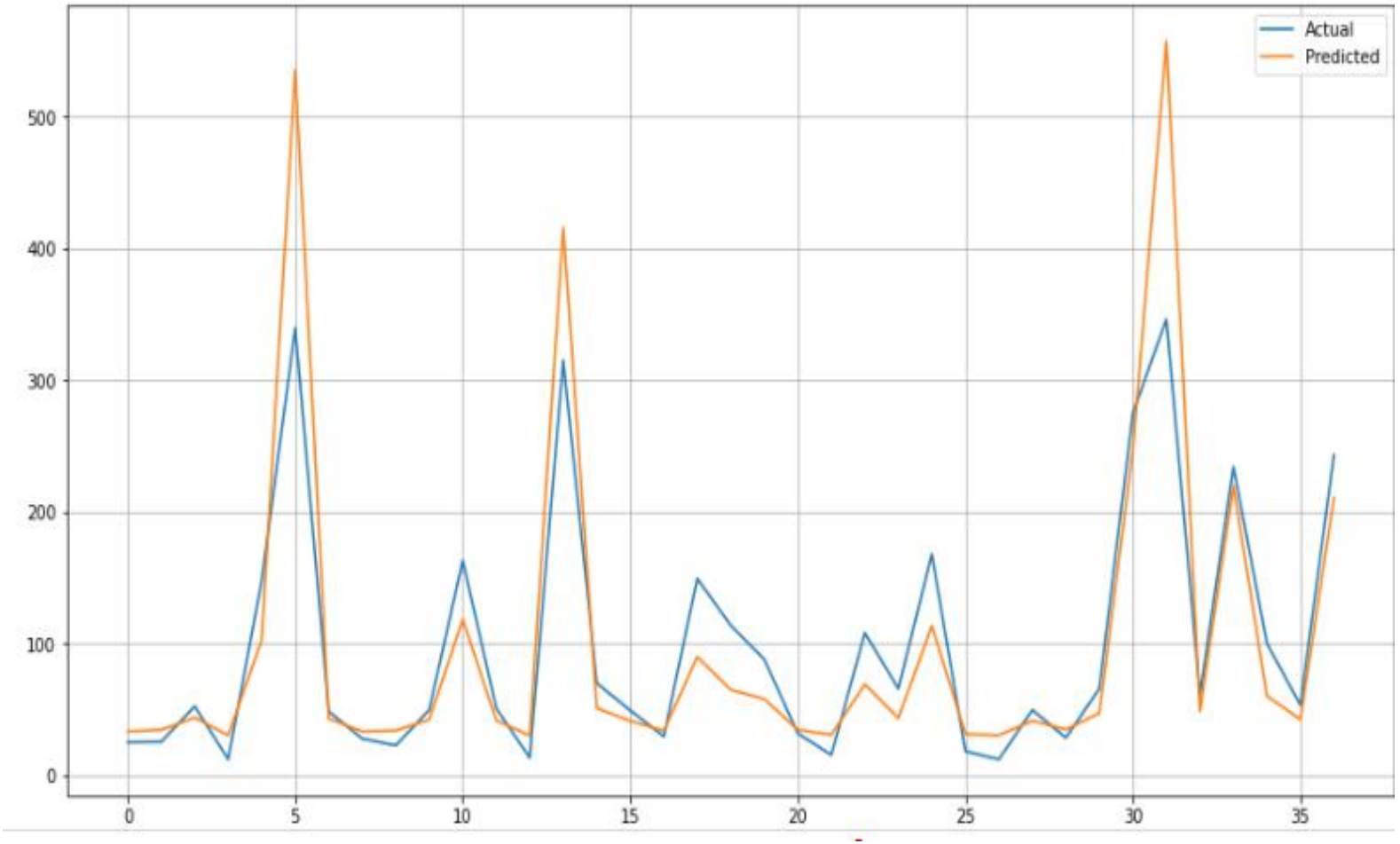
## Elastic Net Regression

```
GridSearchCV(cv=3, estimator=ElasticNet(),
             param_grid={'alpha': [1e-15, 1e-13, 1e-10, 1e-08, 1e-05, 0.0001,
                                   0.001, 0.01, 0.1, 1, 5, 10, 20, 30, 40, 45,
                                   50, 55, 60, 100],
                         'l1_ratio': [0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 1, 2]},
             scoring='neg_mean_squared_error')
```

```
The best fit alpha value is found out to be : {'alpha': 0.01, 'l1_ratio': 1}

Using  {'alpha': 0.01, 'l1_ratio': 1}  the negative mean squared error is:  -0.03515384844892758
```

# Elastic Net Regression (Actual vs Predicted)




MSE : 0.03222273336072056
RMSE : 0.17950691730604856
MAE : 0.15277731058358668
R2 : 0.8189732786857935

AI

# Conclusion:

- ✓ All the features except ‘Date’ are revelant in predicting Stock Closing Prediction and all of them are found to be so correlated to the target feature which gives high accuracy when it is implemented on the baseline model with the score of 82%.
- ✓ Only the Lasso, Ridge and Elastic Regression models are used but there were many good ones out there.The models can also be improved further by tuning finer on hyperparameters.
- ✓ Last two years of the stock values are dropped significantly due to the fraud accusation of the Owner and the model has been predicting low closing stock values for the upcoming months.

# Challenges:

- There was a higher possibility of Overfitting as the dataset has low number of data and the features are very multicollinear
- Hyper parameter tuning did gave a best result but did took more computation time

# Thank You