Capstone Project 2

Yes Bank Stock Closing Price Prediction

Contribution - Individual
Project By - Saidul Mondal

Contents

• Introduction

To determine the YES bank's stock’s future value on the national stock exchange by making machine learning model of linear regression. The advantage of a successful prediction of a stock's future price could results insignificant profit. The efficient market hypothesis recommends that stock costs mirror all right now accessible data and any value changes that are not founded on recently uncovered data subsequently are an unpredictable. We have to build model which help us to predict the future stock closing prices.

Data Summary

We have Yes bank stock price dataset in this project. In this dataset, we have 185 rows and 4 (attributes) columns.
Four attributes are Open, High, Low, and Close price.

Attributes

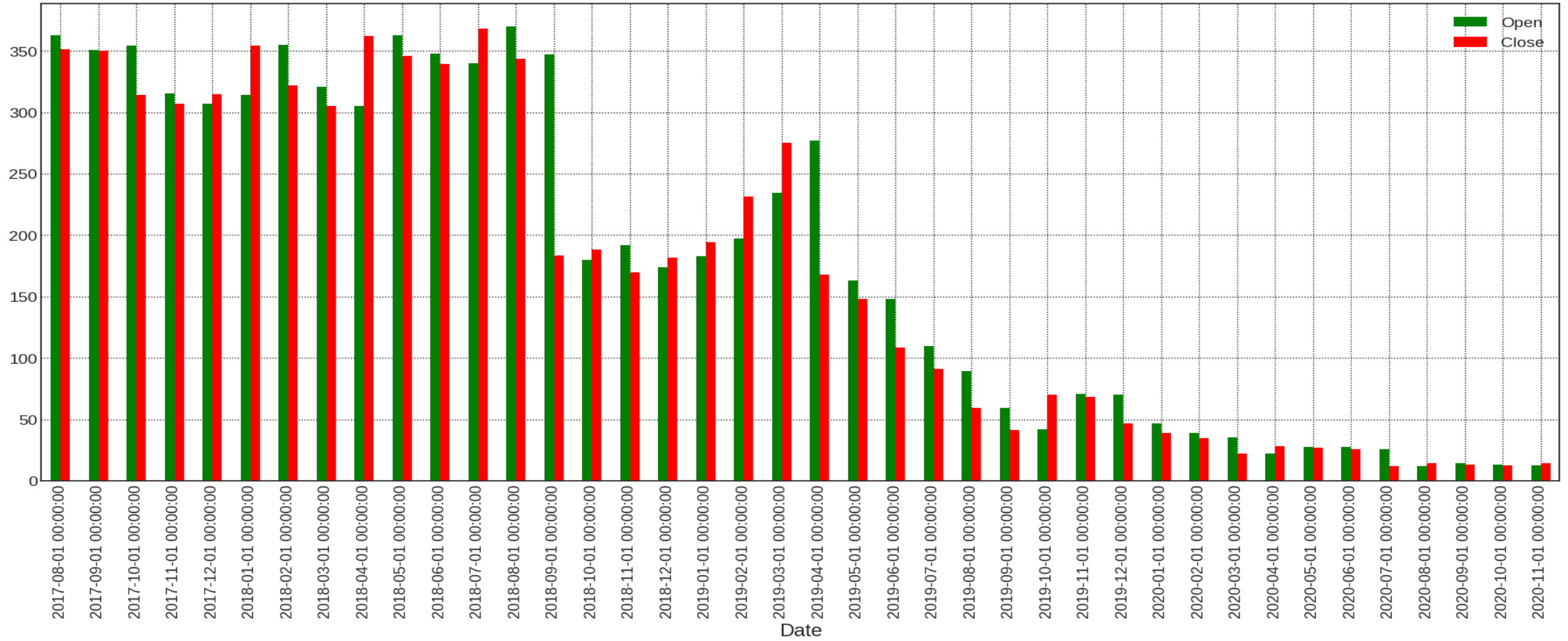
ROWS

	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

• Variable Analysis

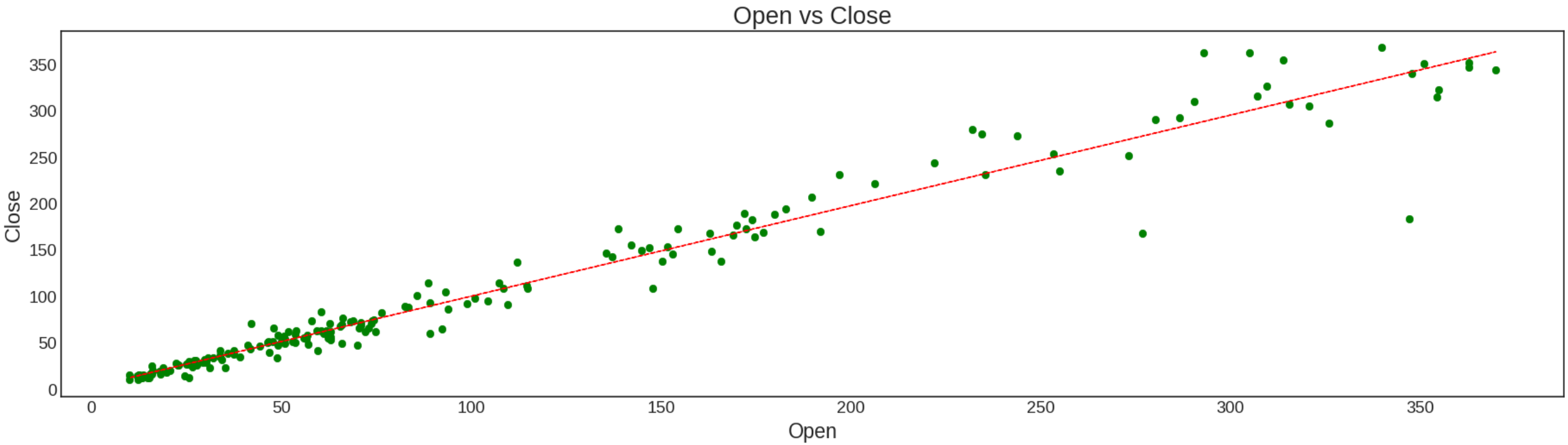
- Here, Yes bank opening price and Yes bank closing price has same result. Opening price started increasing in year 2014 and it was at peak in year 2018. But after 2018 it started falling down continuously and came at 0 in year 2020 same as Yes bank closing price.

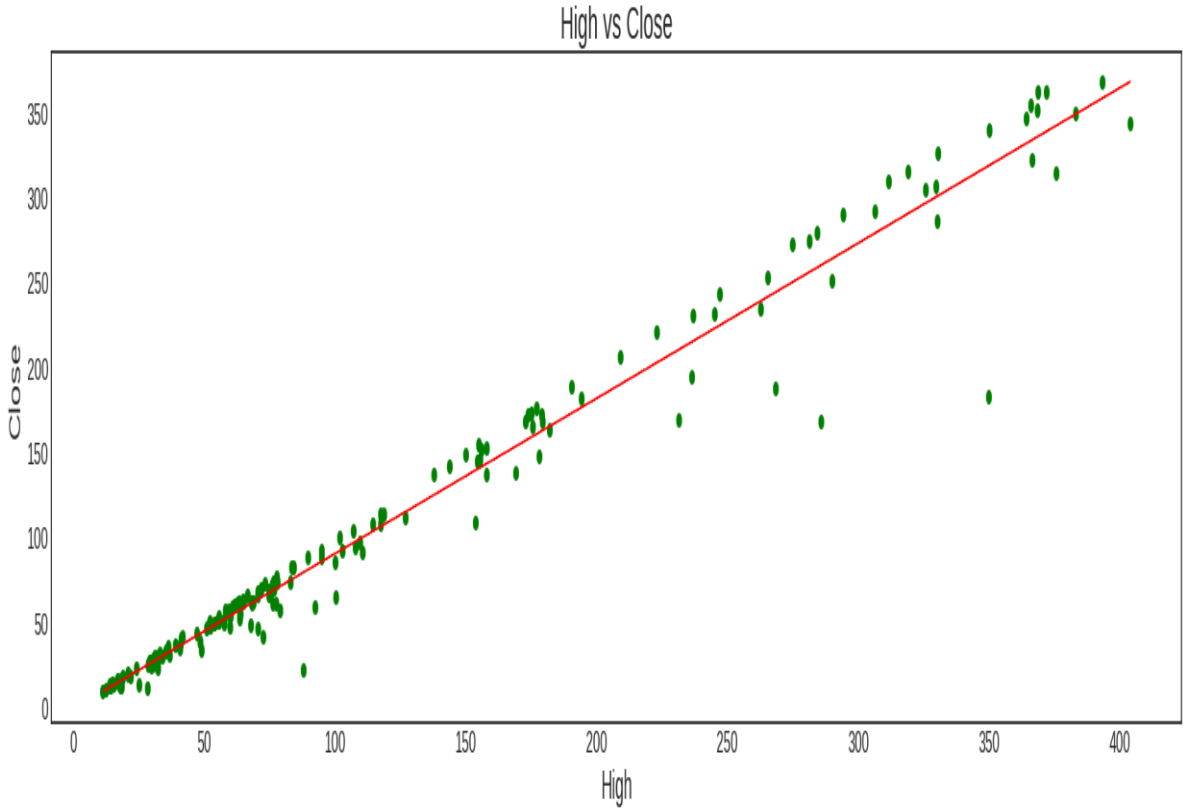
- **Graph analysis between Open and Close prices**

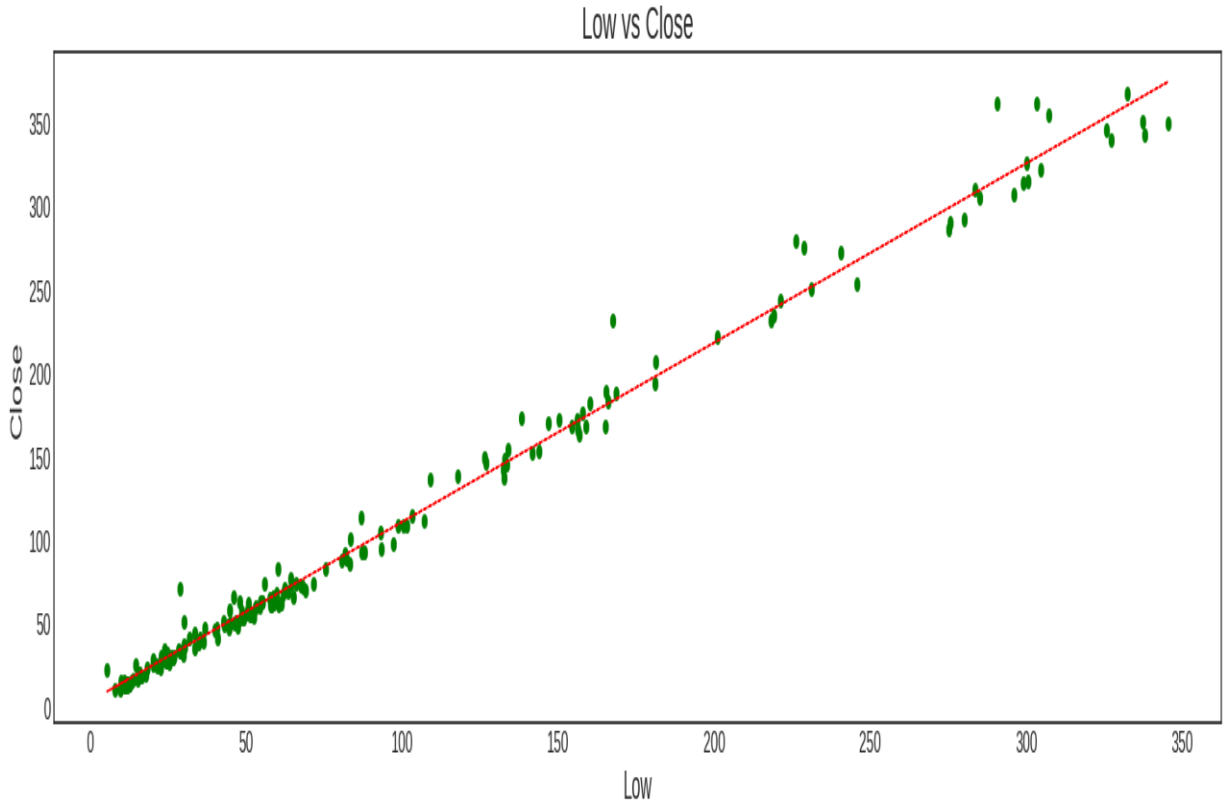


- From the above graph we can conclude the point that the stock price of the YES BANK falls down after the year 2018 and it is not beneficial for investors to invest their money.

• Bivariate Analysis

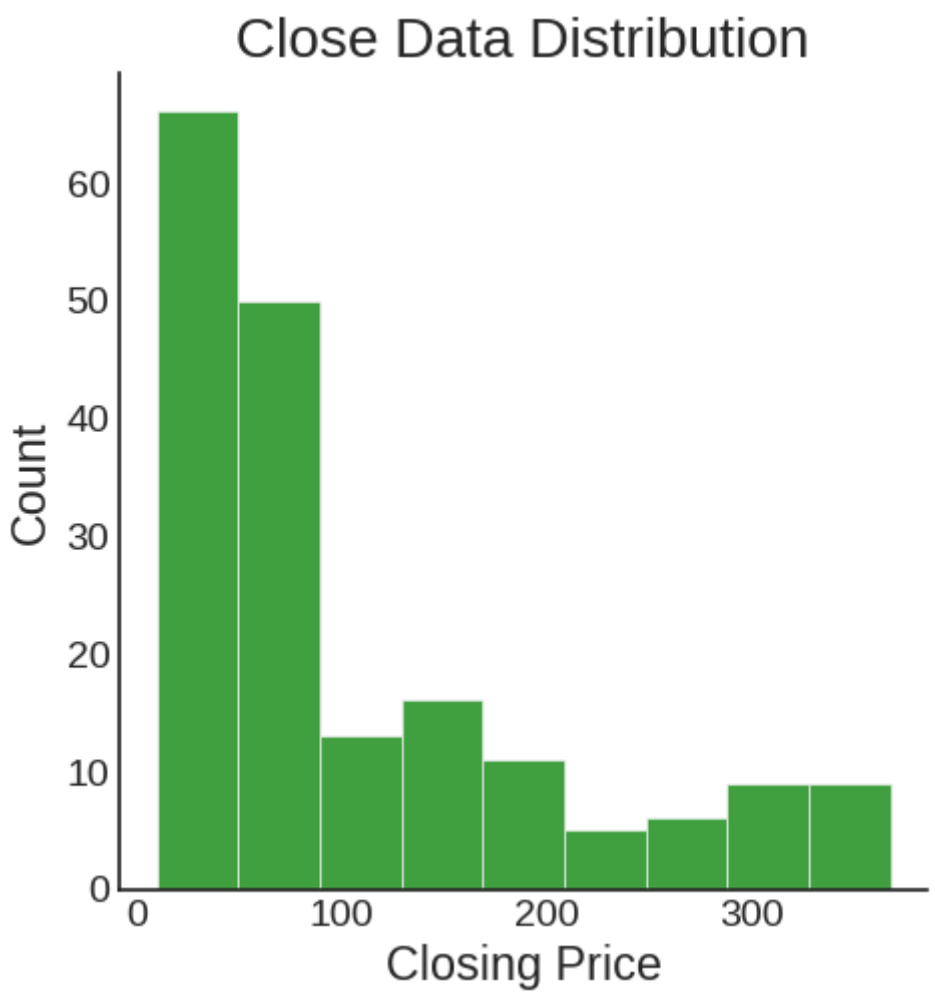


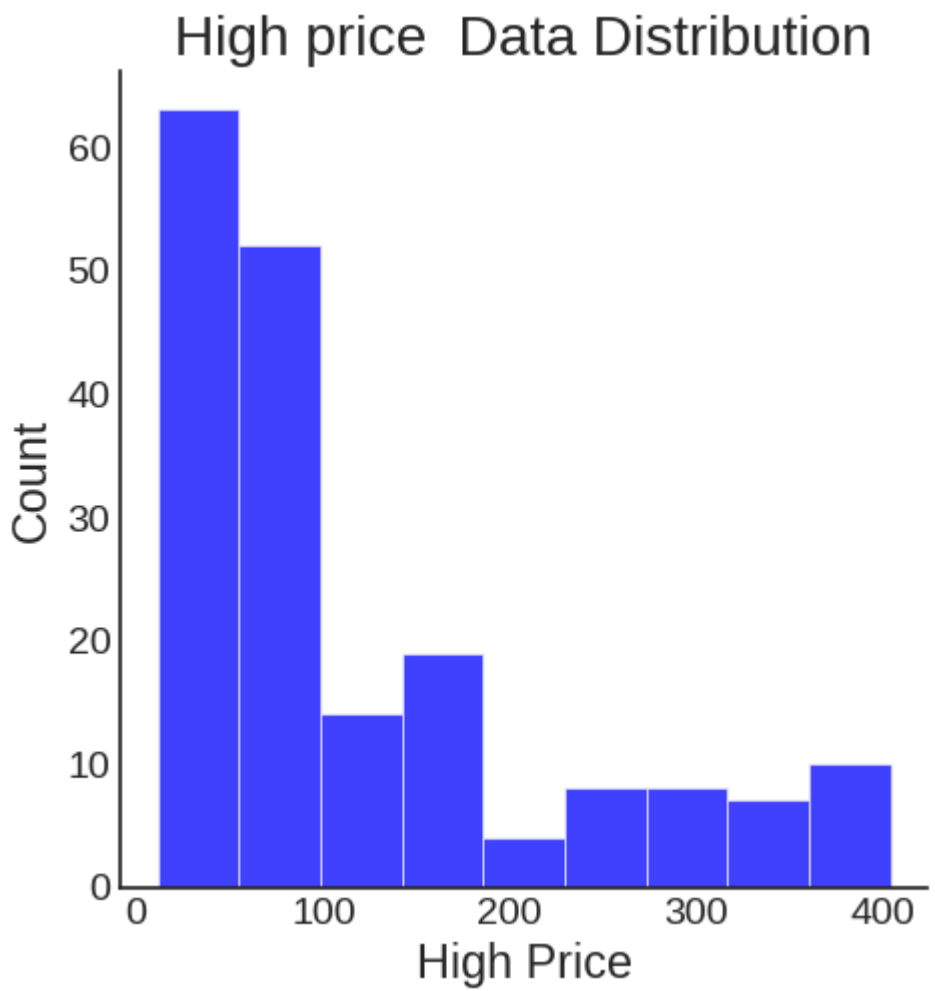


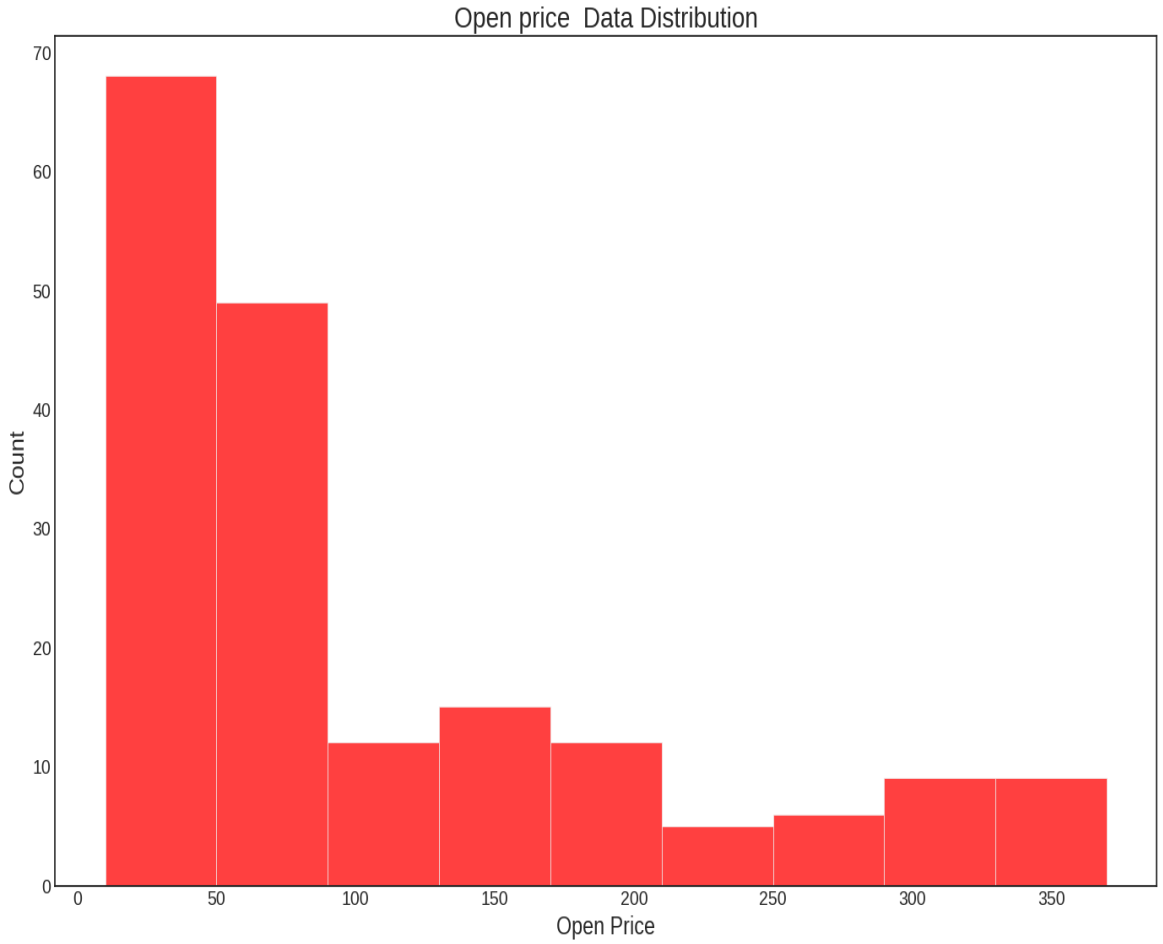


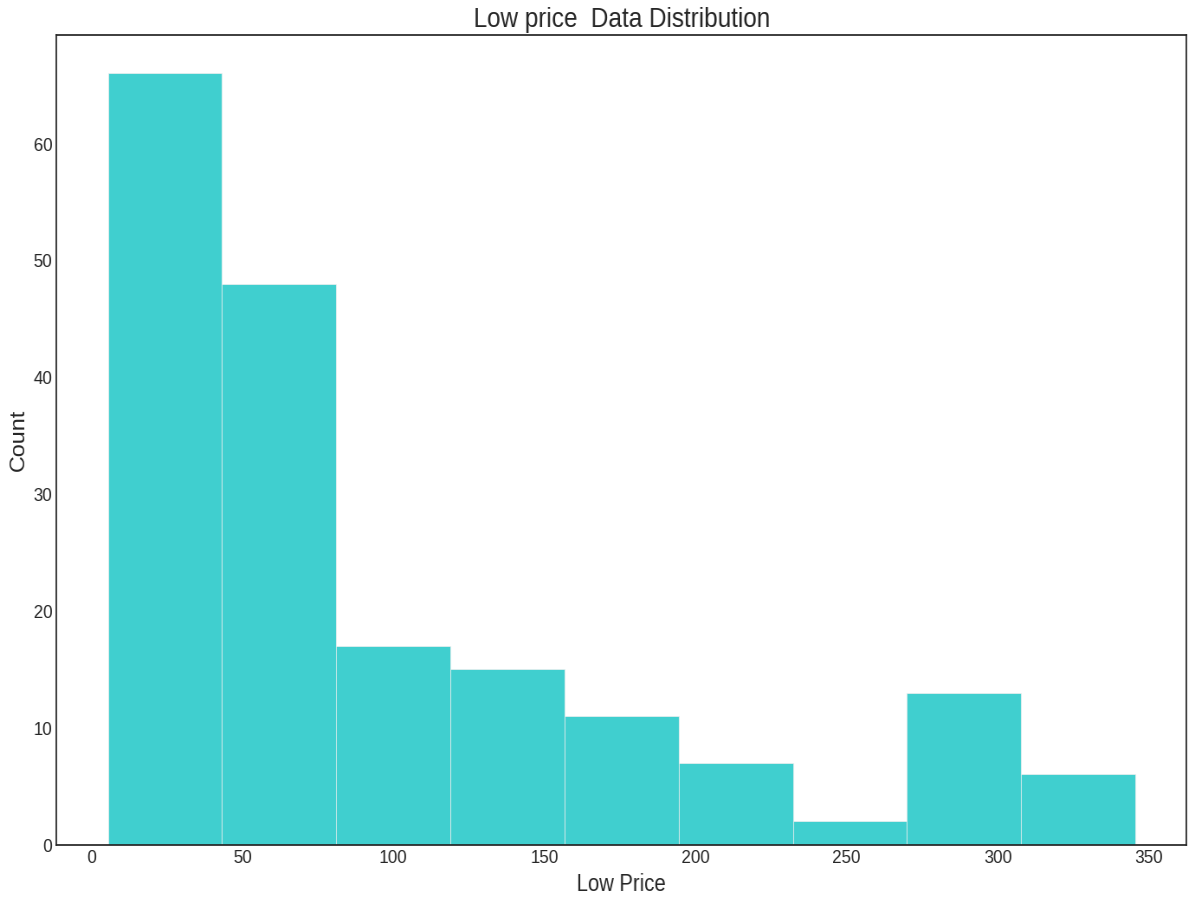
- In all above scatter plot we can conclude that bivariate analysis shows high correlation of close price with other features, and other features also shows correlation between each other.

Data Analysis



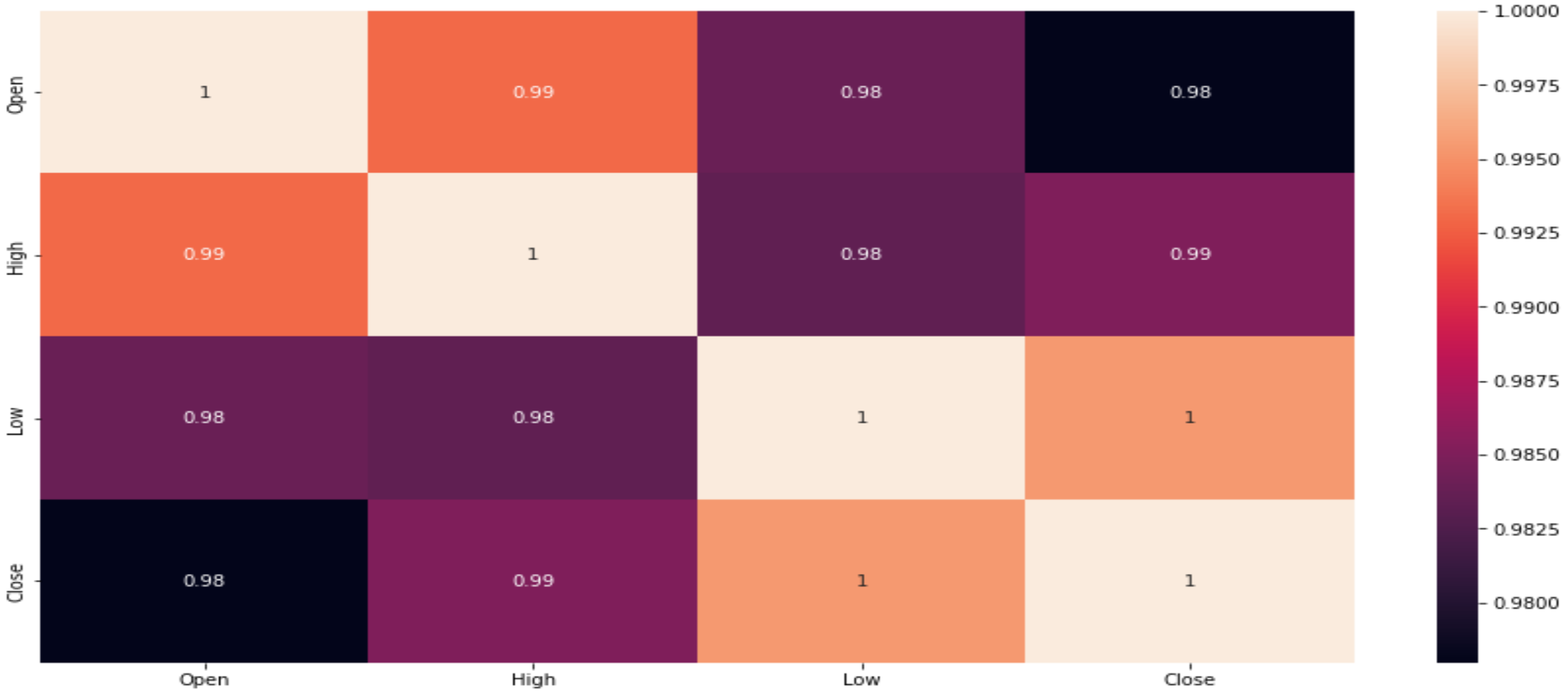






- We can see in all histogram plot that they all are right skewed.

• Collinearity graph between all variables



- From the above heatmap we can conclude that all the features showing high correlation between each other.

4.Data Cleaning

In Data cleaning, we are importing datetime so that we can convert the date in to proper format of date.We have given date in mmm-yy format then it converted in proper format of yyyy-mm-dd and given date column has dtype as object converting it into date time format.

Before Data Cleaning

	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

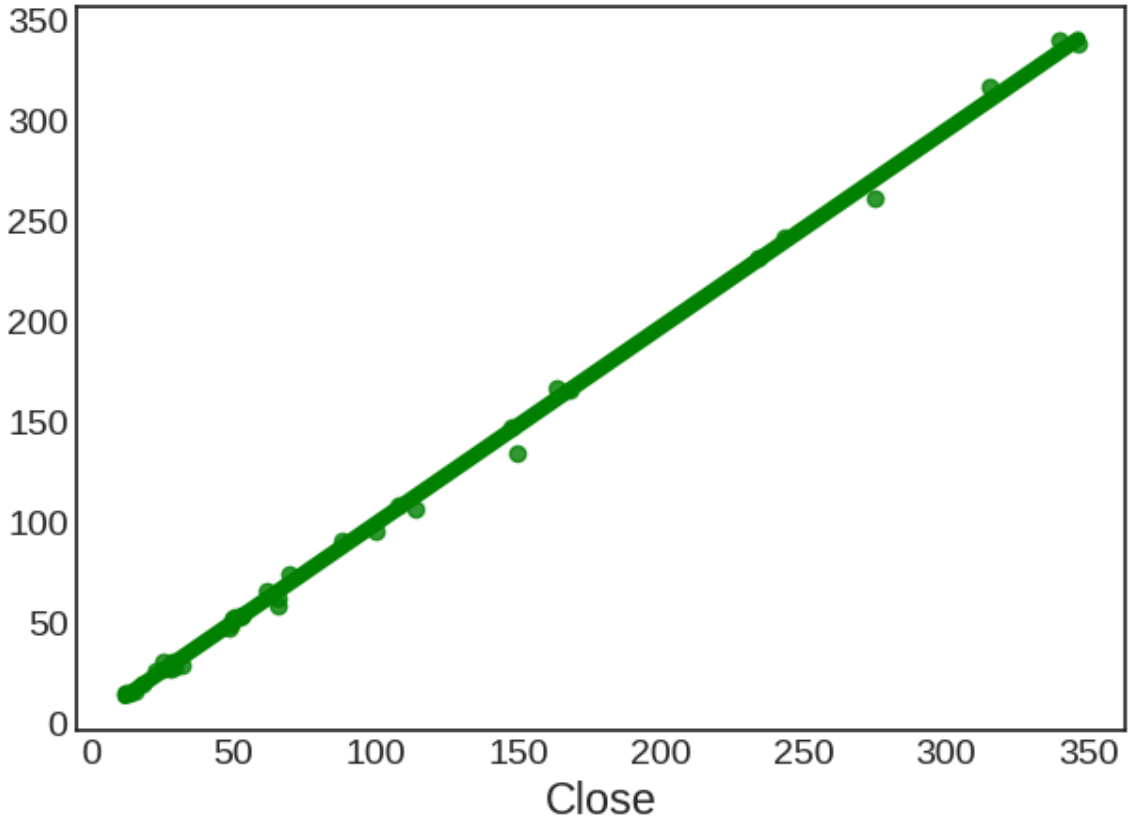
After Data Cleaning

Date				
2005-07-01	13.00	14.00	11.25	12.46
2005-08-01	12.58	14.88	12.55	13.42
2005-09-01	13.48	14.87	12.27	13.30
2005-10-01	13.20	14.47	12.40	12.99
2005-11-01	13.35	13.88	12.88	13.41

6.Model Training by Regression Problem

• Linear Regression

The term regression is used when you try to find the relationship between variables.Linear regression uses the relationship between the data-points to draw a straight line through all them.This line can be used to predict future values.

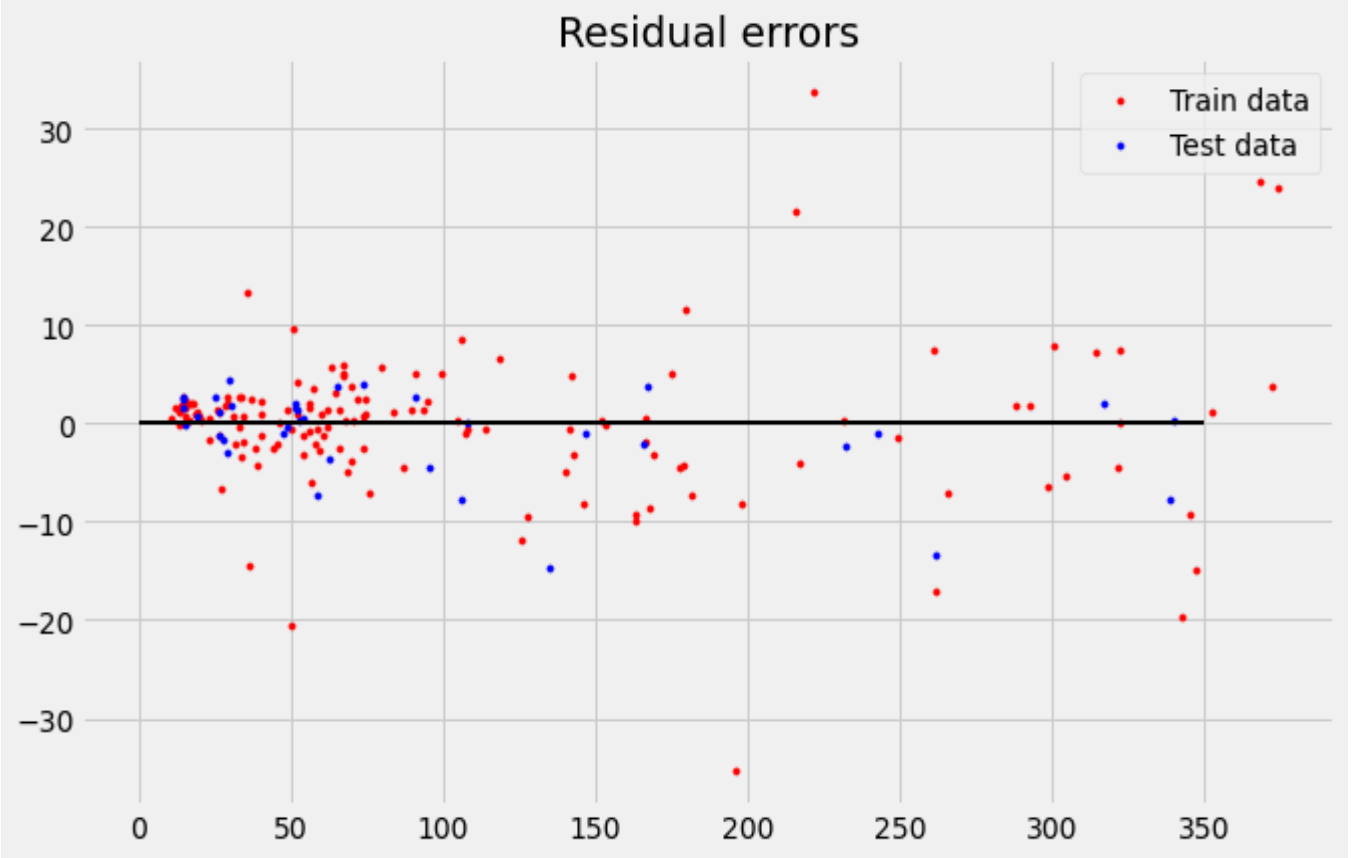


How it works - Python has methods for finding a relationship between data-points and to draw a line of linear regression. We will show you how to use these methods instead of going through the mathematical formula.

A linear relationship between a dependent (y)(in our case is Close Price) and one or more independent (in our case Open,Low,high) variables, hence called as linear regression.

Residual Error

A residual is a measure of how far away a point is vertically from the regression line. It is the error between a predicted value and the observed actual value.
Here, in this residual plot it has a high density of points close to the origin and a low density of points away from the origin and also it is symmetric about the origin.
This linear model is a good fit for relatively small x-value,but is not a good predictor of larger x-values.



Lasso Regression

Lasso regression analysis is a shrinkage and variable selection method for linear regression models. The goal of lasso regression is to obtain the subset of predictors that minimizes prediction error for a quantitative response variable. The lasso does this by imposing a constraint on the model parameters that causes regression coefficients for some variables to shrink toward zero.

MSE : 20.214848381869473 RMSE : 4.496092568205138

MAE : 3.073530700265413 R2 : 0.9978168172187769

Cross Validation

In cross validation we can perform our model on the new dataset or we can say test dataset.So that we can check our model performance.

So the conclusion, the R squared value for the test data was 99.7%. This is almost same as in the score from the training dataset which proves that in a dataset we achieve the best fit model.

MSE : 20.06685135300319 RMSE : 4.4796039281395394

MAE : 3.058968183477561 R2 : 0.9978328007452911

Ridge Regression

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values being far away from the actual values.

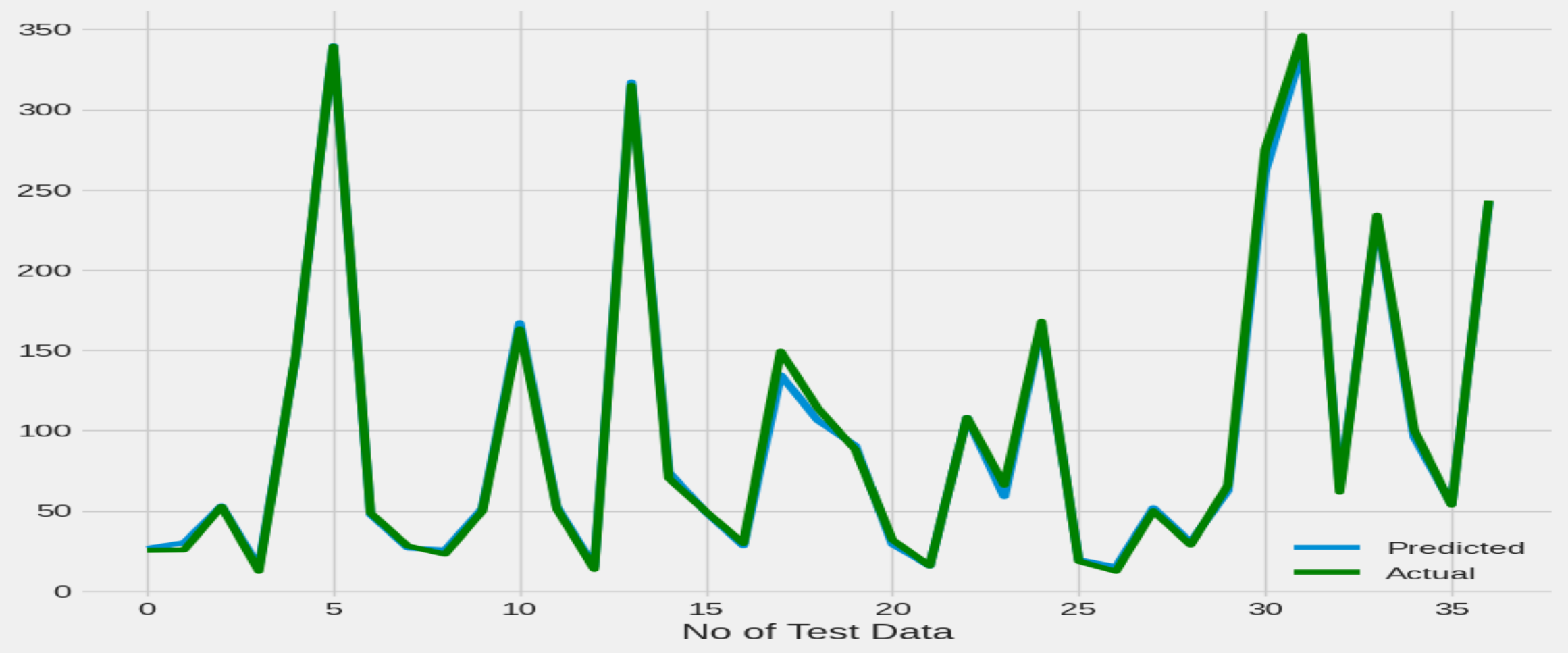
MSE : 29.595509126508347 RMSE : 5.440175468356545
MAE : 3.6721812916399585 R2 : 0.9968037155309819

Cross Validation

MSE : 20.06685135300319 RMSE : 4.4796039281395394
MAE : 3.058968183477561 R2 : 0.9978328007452911

After implementing the best parameters best R^2 score we have 99.78% for Ridge regression model.

• Linear regression model performance visualization



From the above linear regression model visualization we can say that our model is perfect fit.

- **Challenges**

1. Small dataset and that dataset is in improper manner.
2. In data cleaning, we had to change into proper dd/mm/year format.
3. All the features showing high correlation between each other.

Conclusions

- At first I did the data wrangling and then data cleaning and after that I did the EDA part.
- In EDA part I conclude from our dataset that
  - Stock close price decreased after year 2018 it is mainly because of Rana Kapoor case and hitted the stock price badly.
  - The graph for Yes bank opening price and Yes bank closing price has same result.
  - The point that the stock price of the YES BANK falls down after the year 2018 and it is not beneficial for investors to invest their money.
  - From scatter plot I can conclude that bivariate analysis shows high correlation of close price with other features.
  - All histogram plot shows that all are right skewed.
  - From heatmap I can conclude that all the features showing high correlation between each other.
- I implemented linear regression and the accuracy of our linear regression model is 99.78%.
- After that I visualise the performance of our linear regression model and the graph shows that I achieve the almost best fit model for our dataset.

Thank You!