Capstone Project

Yes Bank Stock Closing Price Prediction

Ankita Gupta

Outline

- 1. Overview & Objective.
- 2. Data outline.
- 3. Exploratory data analysis
- 4. Model implementation
- 5. Model Comparison via evaluation metrics.
- 6. Conclusion

Overview

- Yes bank is a well known bank in India which provide wide range of services and solutions right from bank accounts, deposits, cards, cash management, privilege banking, trade finance, Non-Resident India(NRI) banking, institutional banking, merchant acquiring, digital banking and agricultural banking solutions.

Objective

- As the data is all about stock price so, In this project I will analysing the patterns of dataset by performing Exploratory Data Analysis and try to build a model with the help of Machine Learning for predicting the closing stock price.

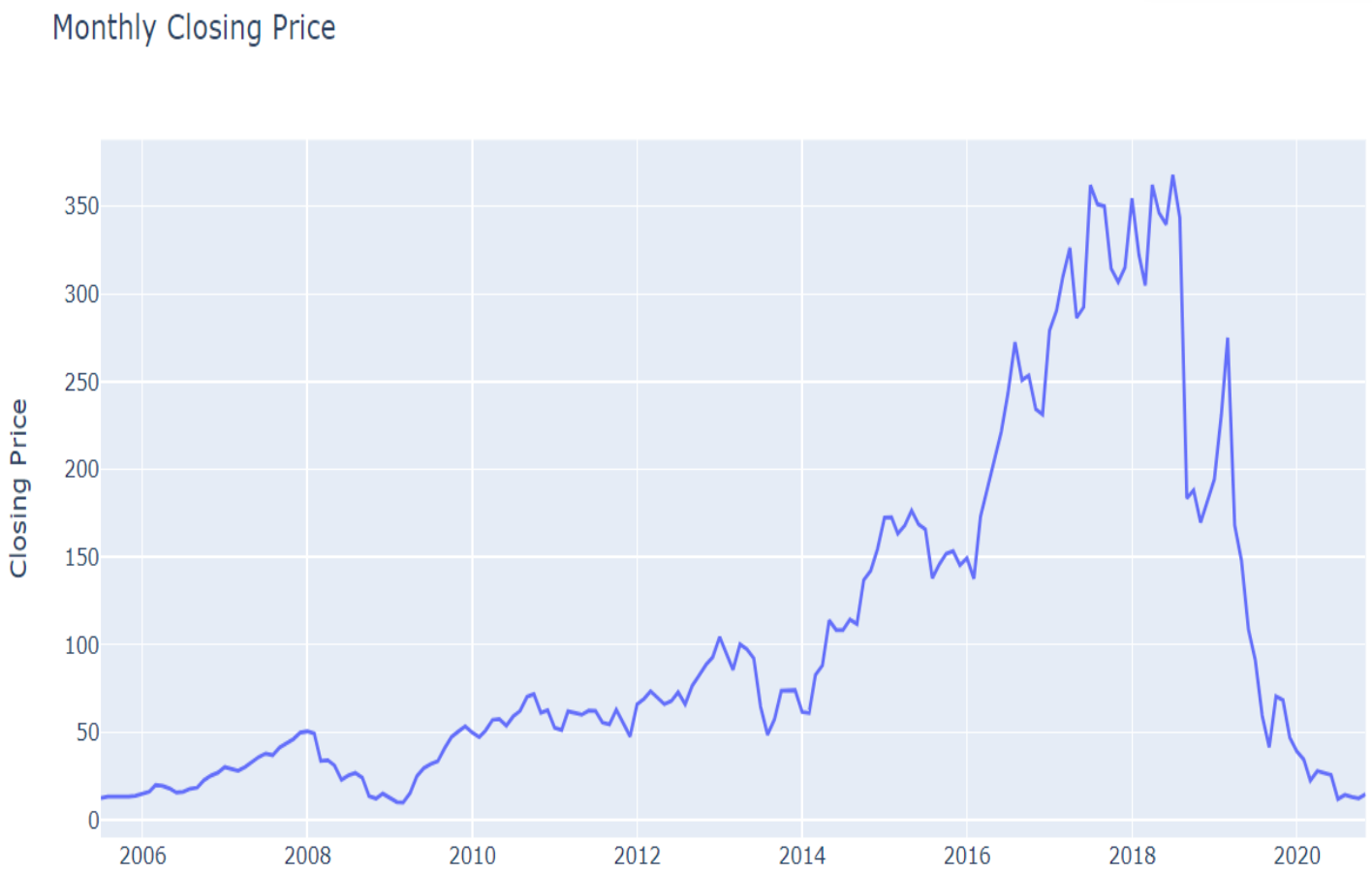
Data Outline

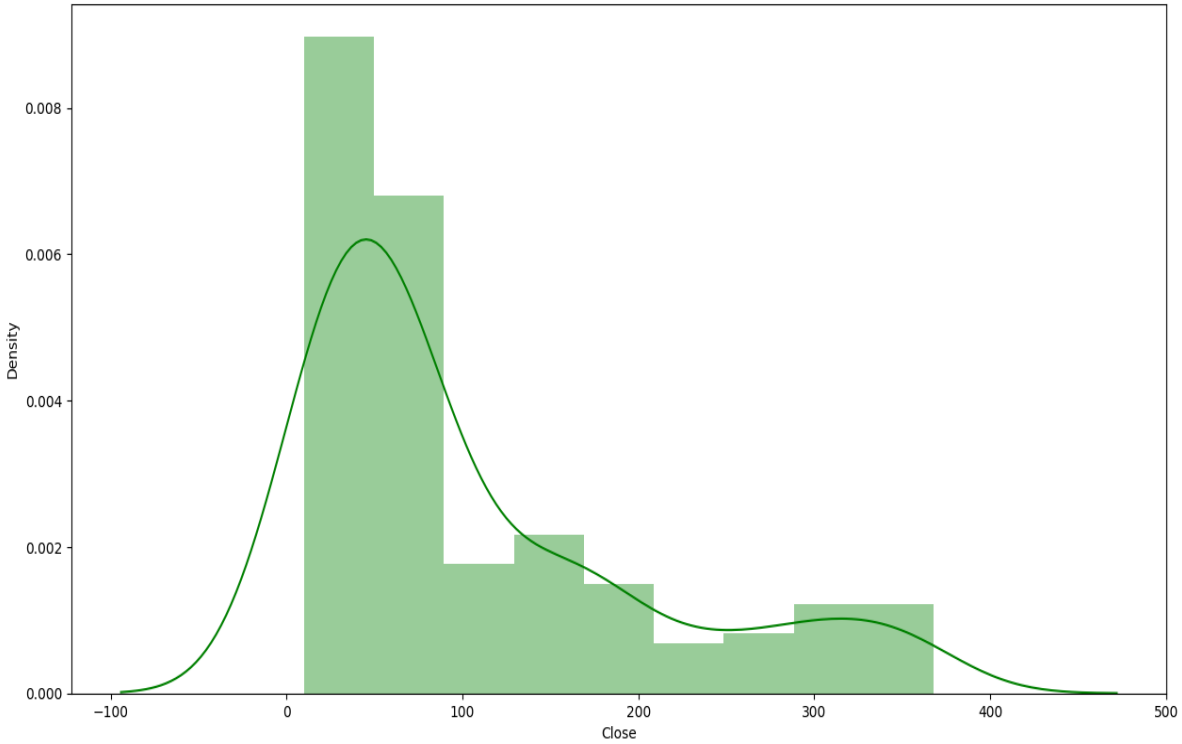
- We have a dataset which contains monthly stock prices of Yes bank shares since the opening of the bank. It contains multiple features like:-
- **Date** :- denotes the date (so we can see the price at a given date.)
- **Open** :- denotes the price at which a stock started trading.
- **High** :- highest price at which a stock traded during a period.
- **Low** :- the minimum price at which a stock traded during a period.
- **Close** :- the closing price refers to a stock's trading price closed at the end. (It's a dependent variable which we need to predict using ML models. The closing price is the price of the stock at the end of the month or the time period in consideration.)

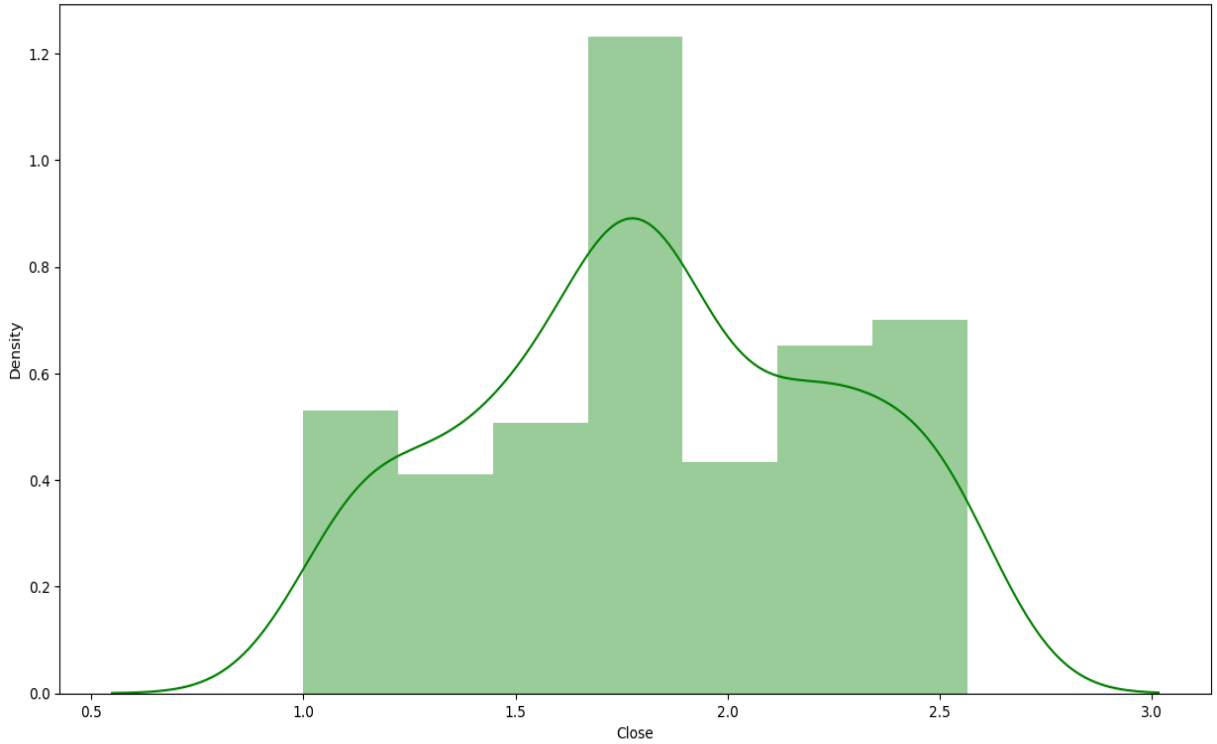
EDA : Visualizing our dependent variable.

- The graph demonstrates how closing price varies with each passing year.

- We can clearly see from the graph that around 2018, when the fraud case involving Rana kapoor came to light, a clear significant dip can be seen in the stock price of Yes Bank data

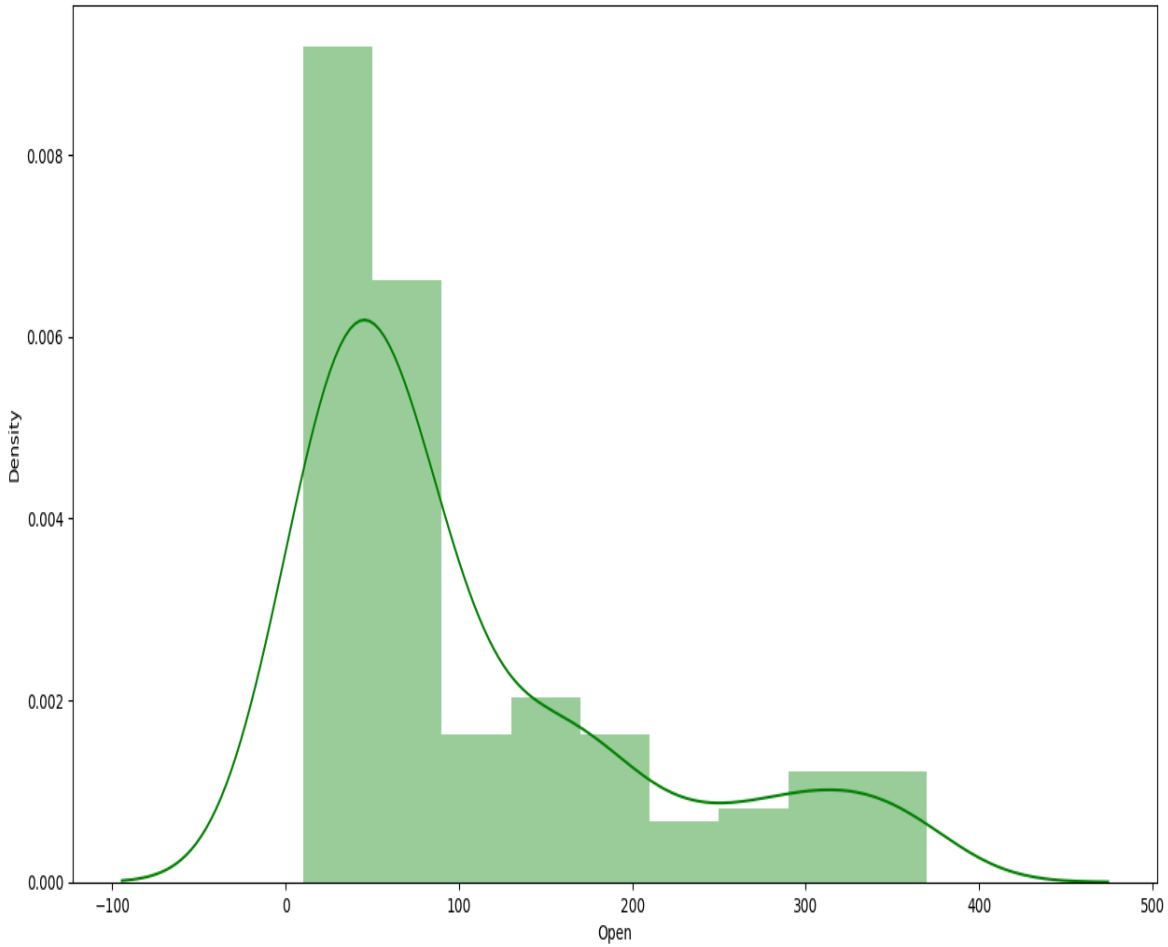


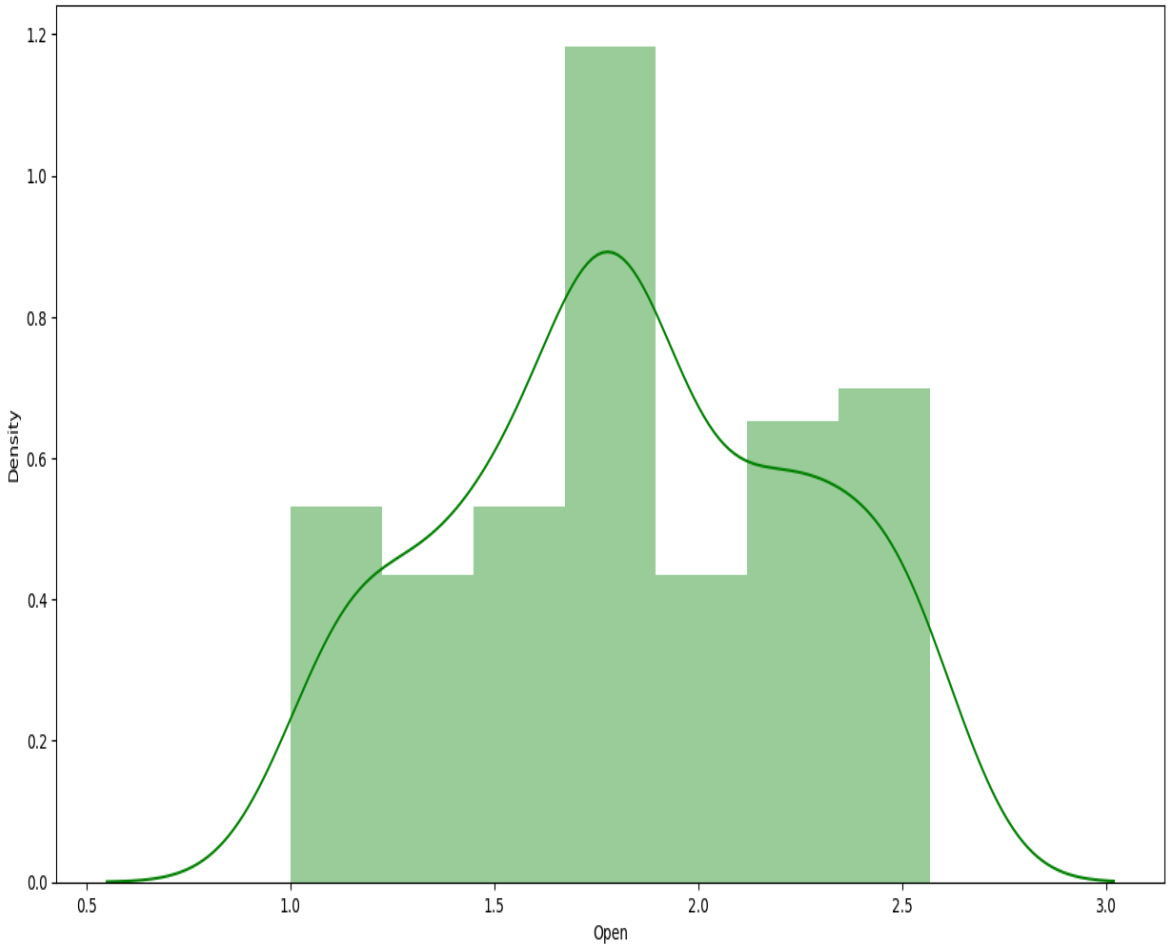




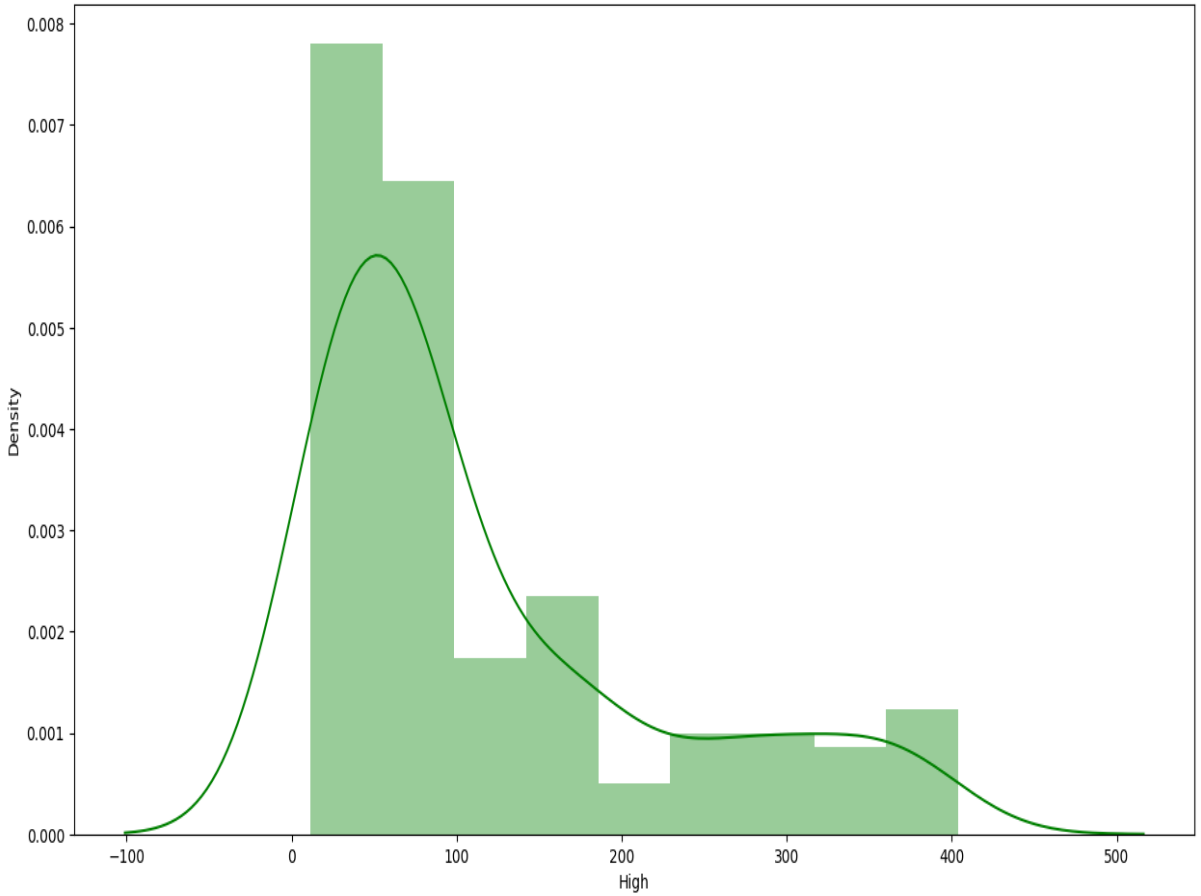
Plotting the dependent variable. We can see that our dependent variable close is positively skewed (as seen on the left). So we do a log transform on it and plot it as seen in the right chart.

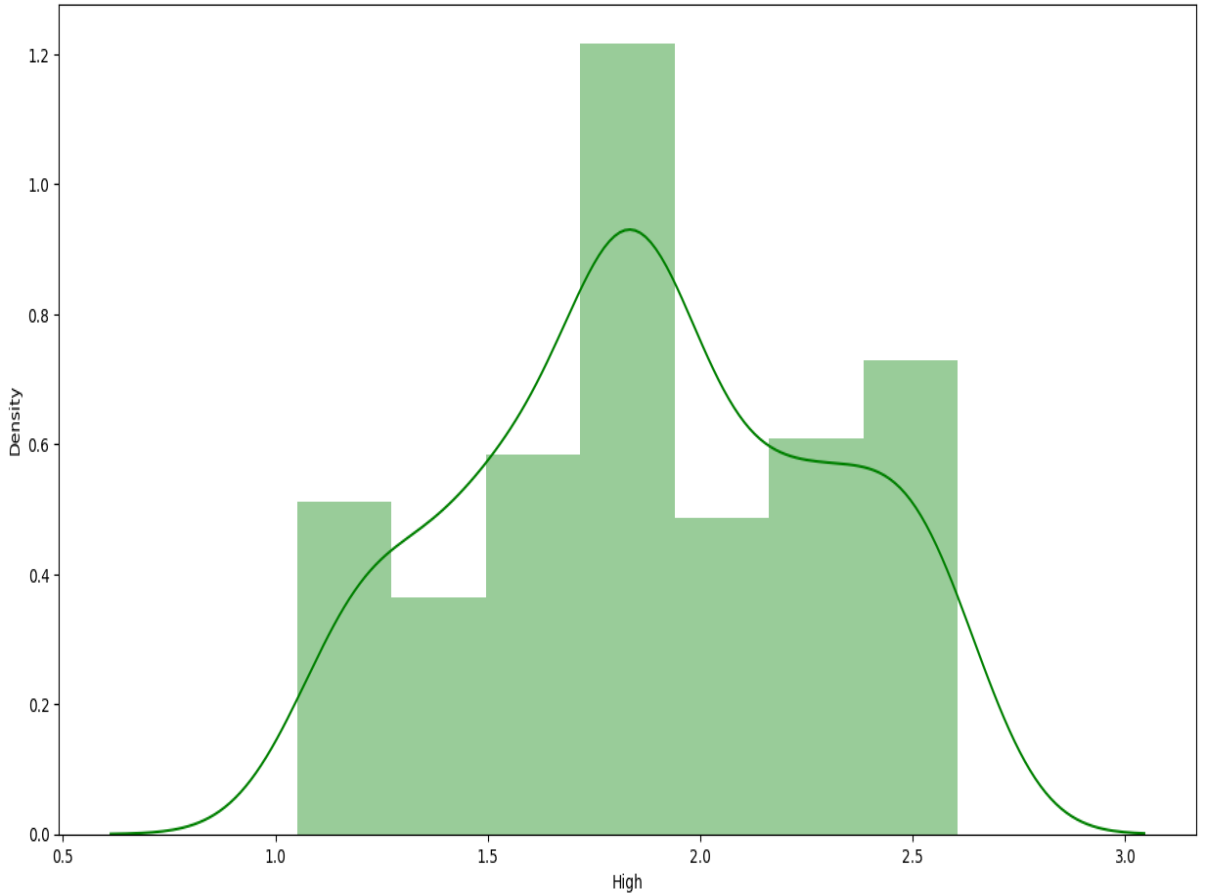
This makes it approximate normal distribution and is optimal for our model's performance. Now our mean and median are nearly equal.



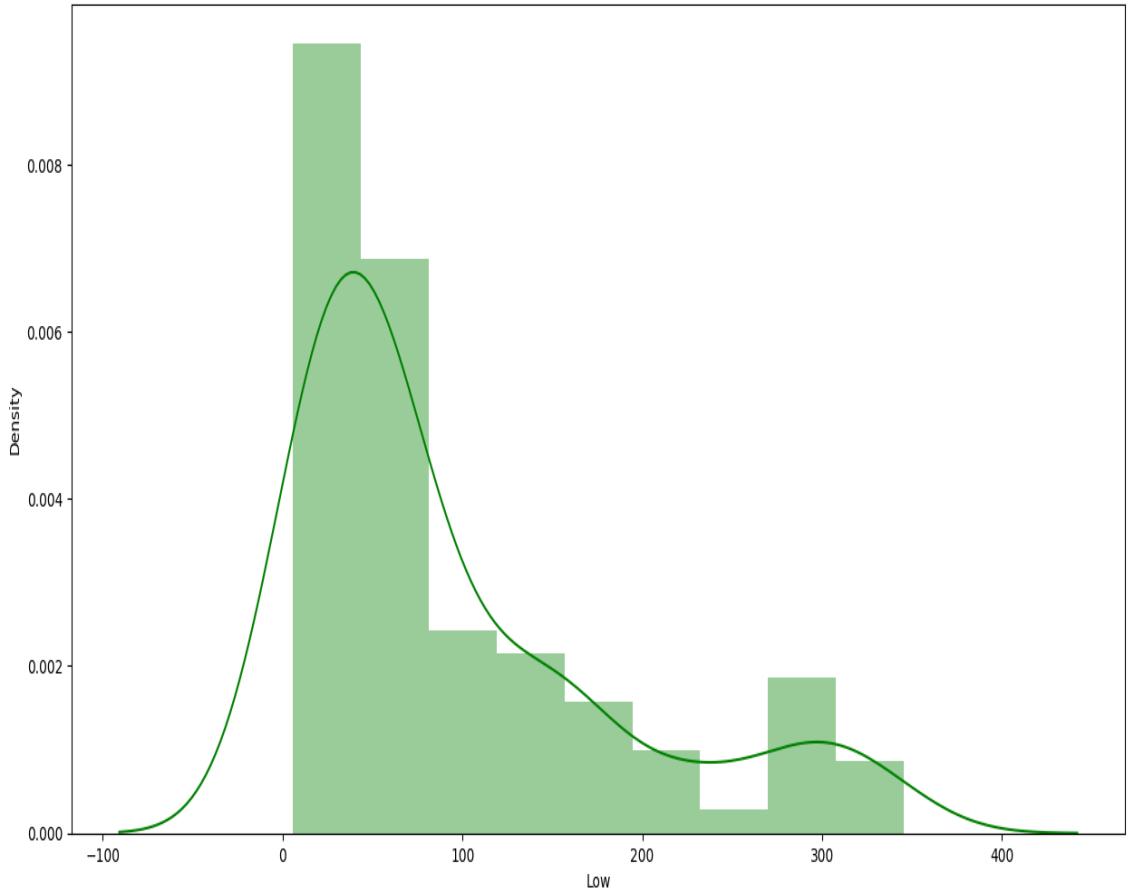


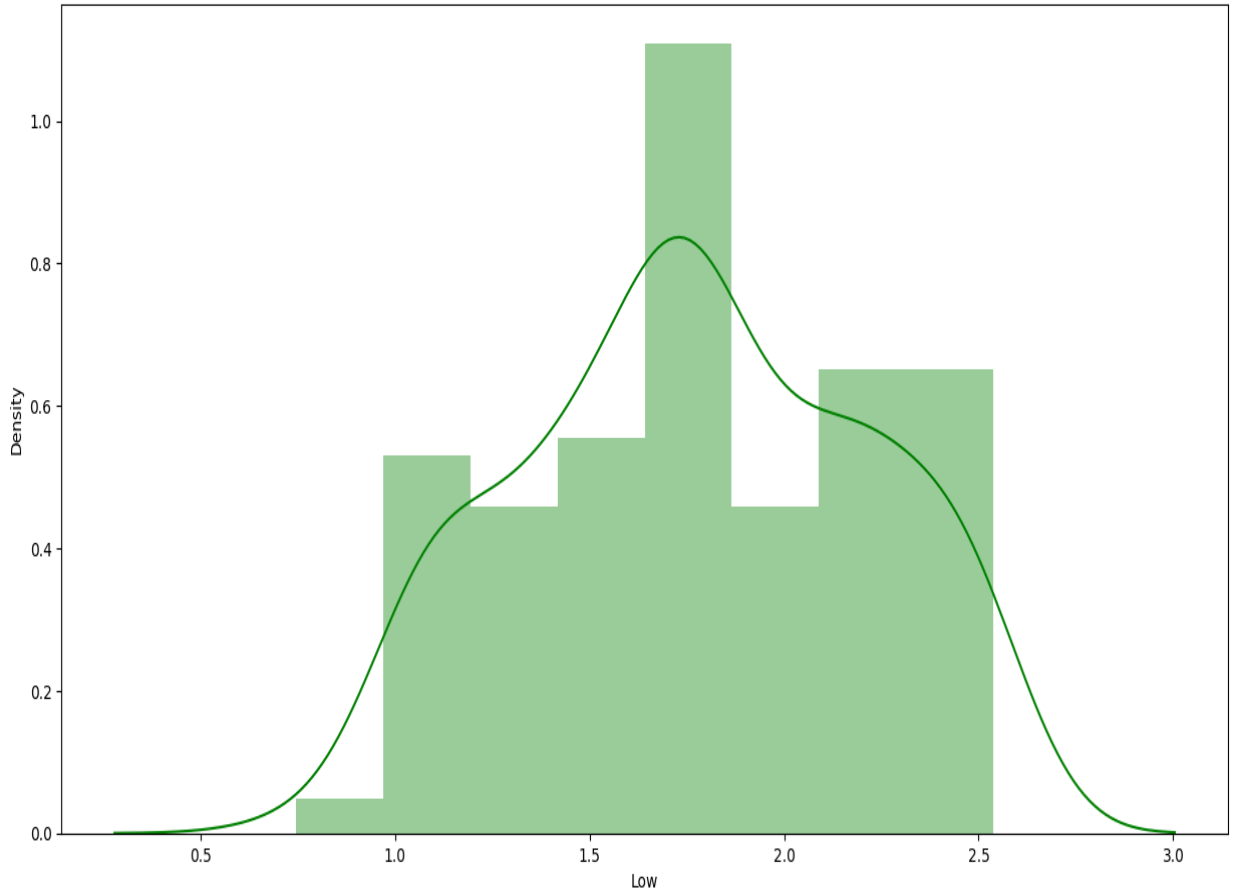
Plotting the independent variables. As we see in the left chart, data is positively skewed, so we perform a log transform on it. In the right chart, we can see the transformed distribution which is similar to a normal distribution.



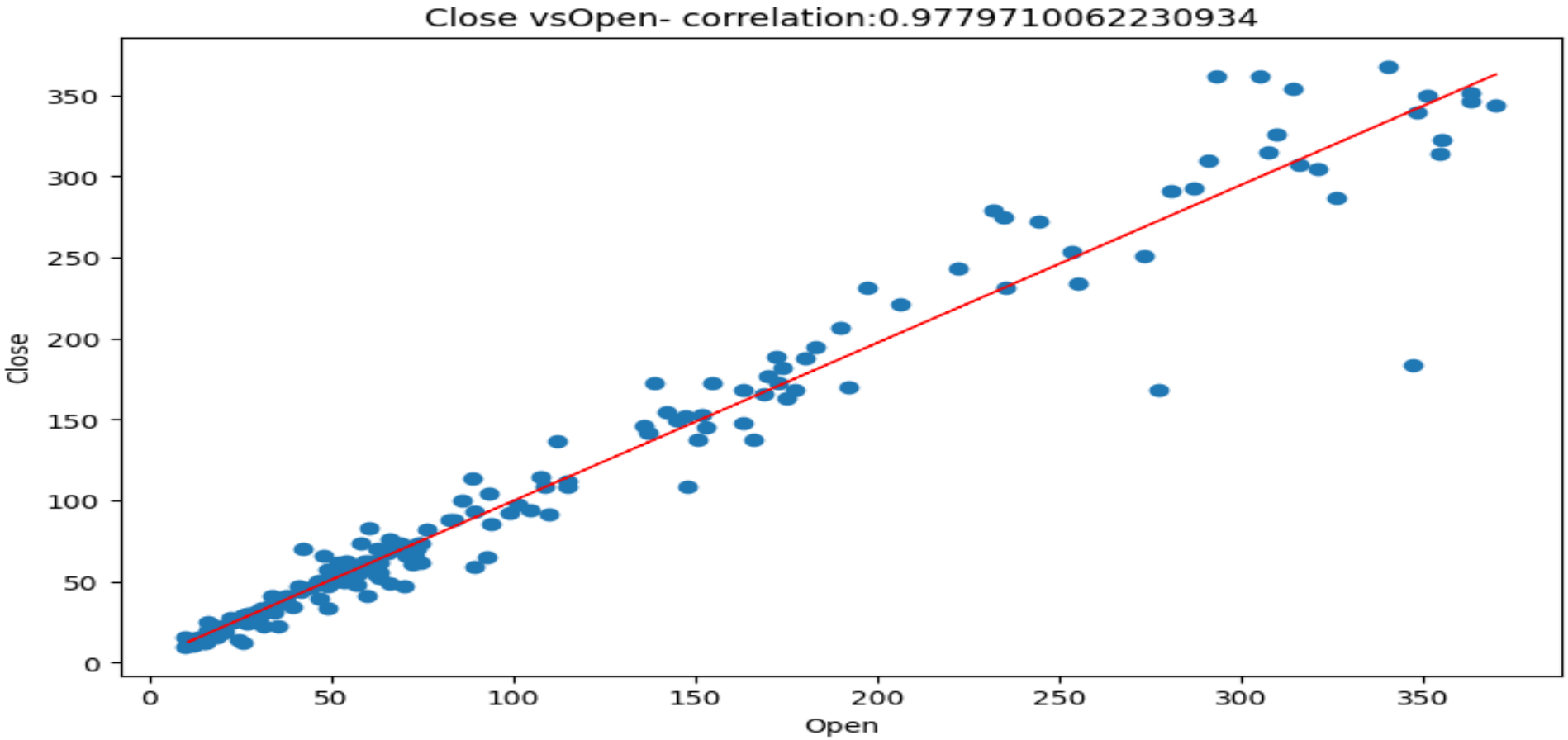


Distribution of independent variable High before and after applying log transform.



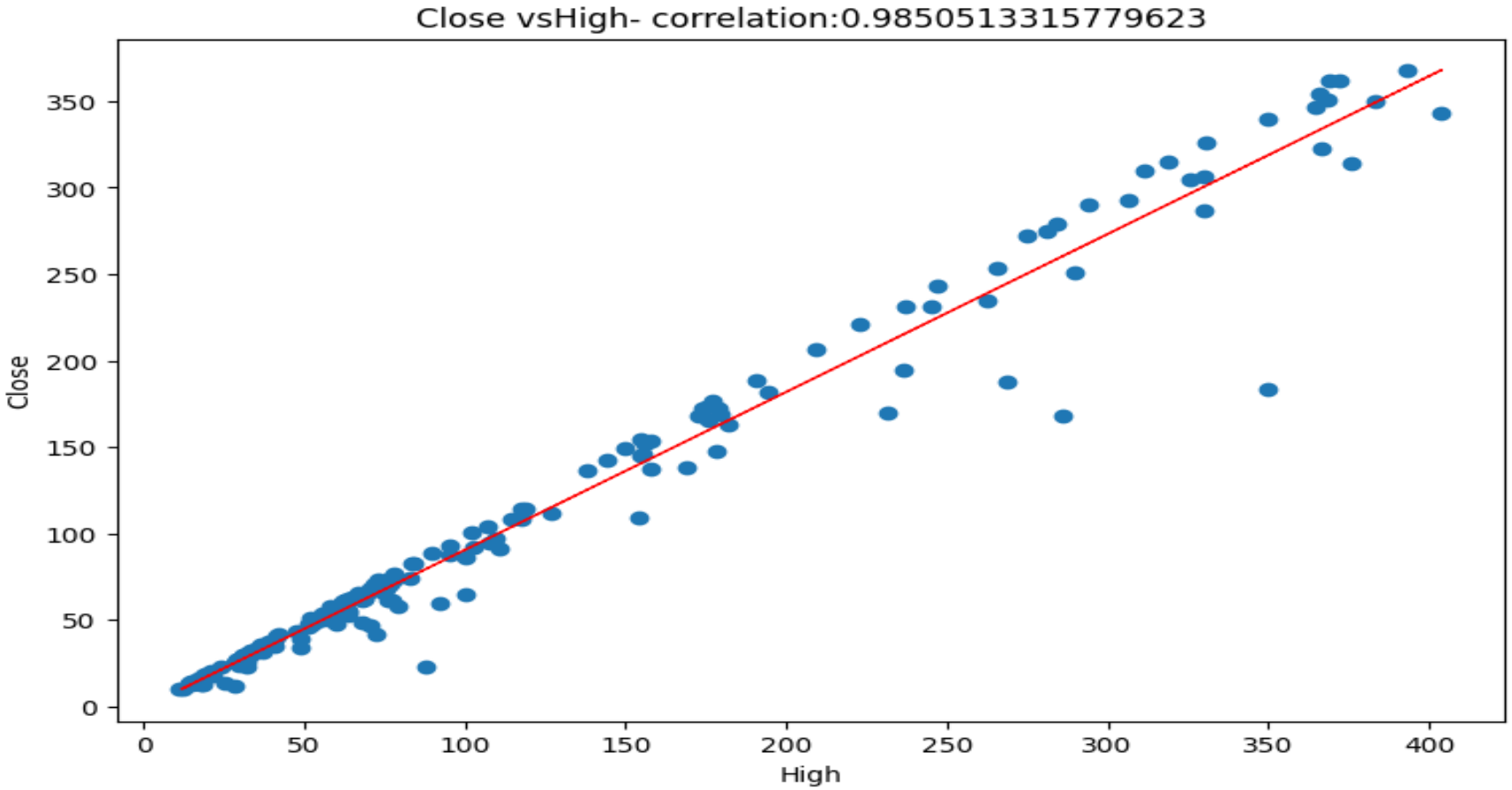


Distribution of independent variable before and after log transformation.



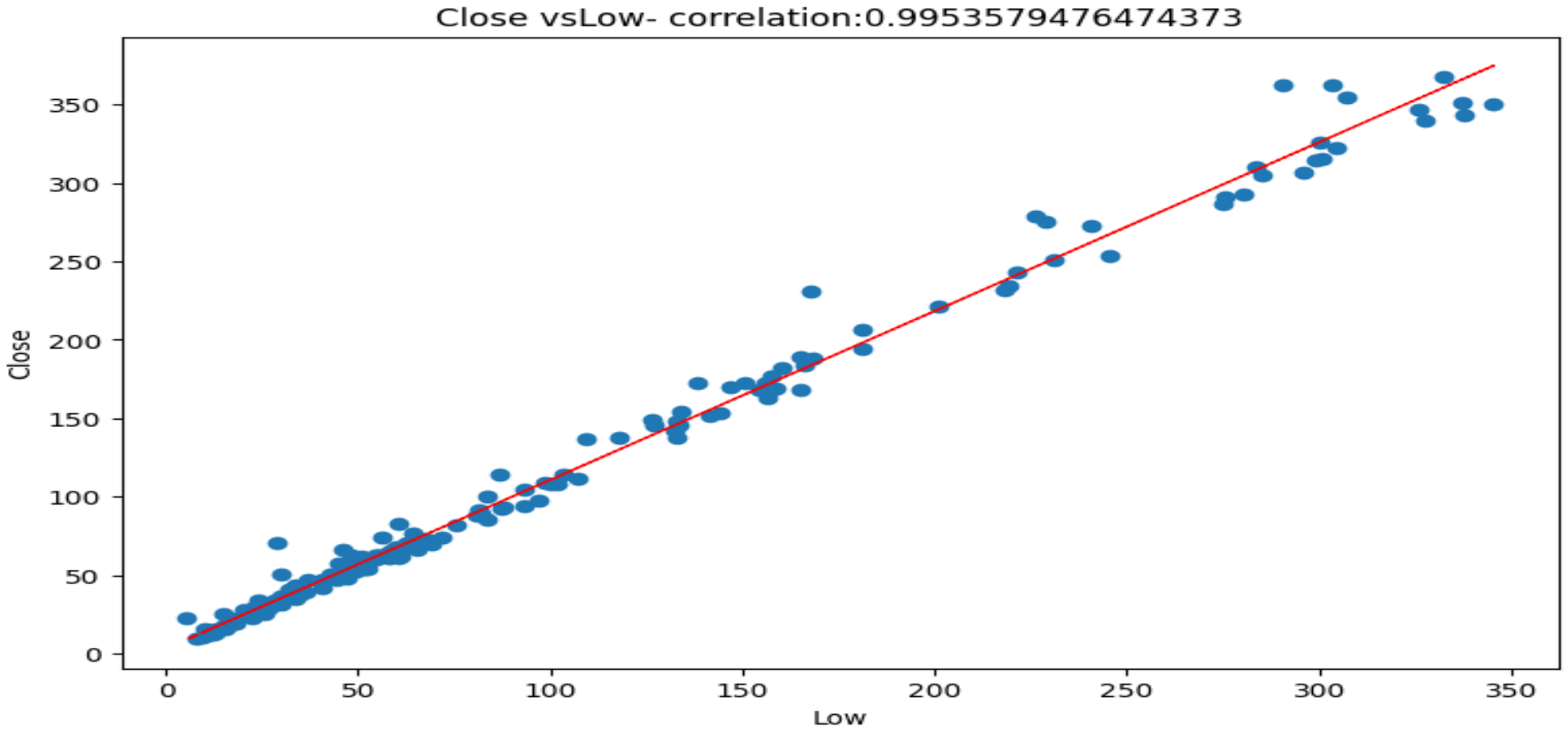
As we can see that the Open and Close data are Highly correlated therefore we can say that the closing price is very much dependent upon the Opening price of the stock.

Also we can see that the value of correlation between dependent variable Close and feature High is 0.977.



- **As we can see that the High and Close data are Highly correlated therefore we can say that the closing price is also very much dependent upon the High price of the stock.**

- **The value of correlation between Close and High is 0.985 .**

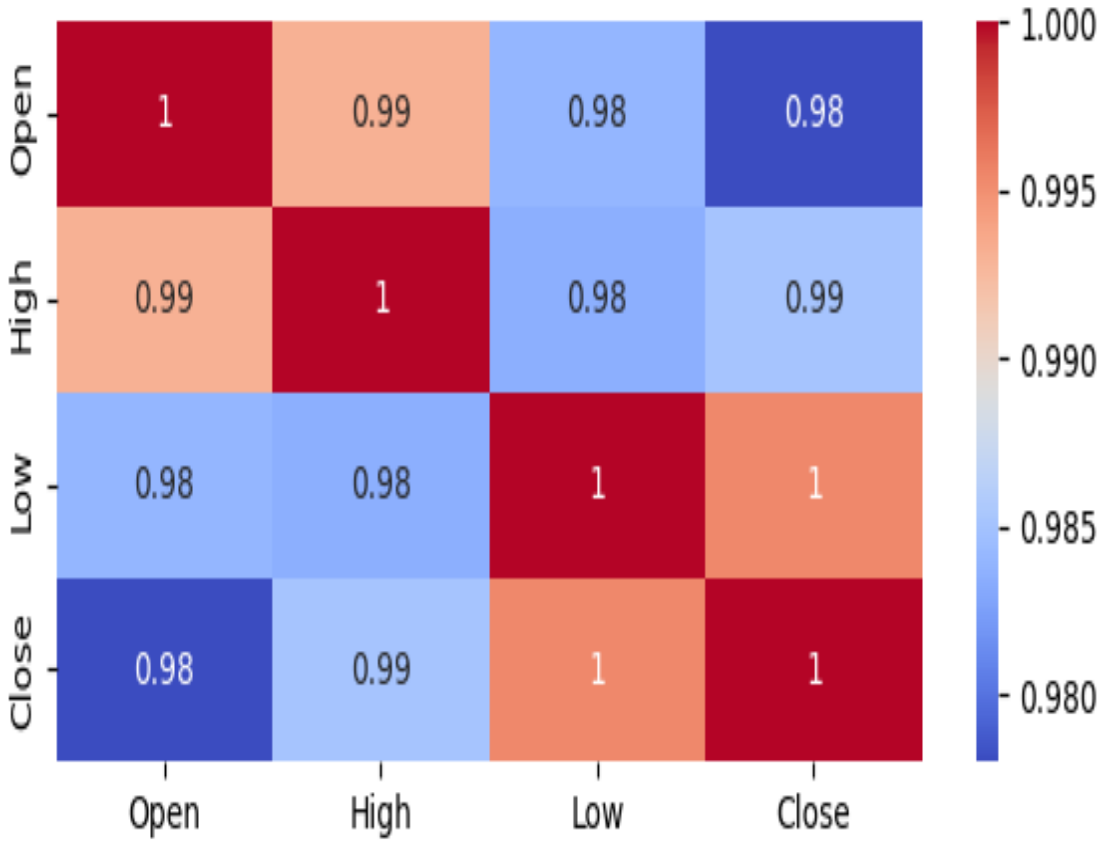


- **The high correlation between the low and price indicates that low price will also play an important role to have an idea about closing price.**

- **The value of correlation between Close and Low is 0.995.**

Correlation Heatmap

- The correlation matrix helps us visualize the correlation of each parameter with respect to every other parameter.
- The colors changes from blue to red for highest to the lowest correlation values and vice versa.
- From the above chart we can see that each and every feature is highly correlated to each other.



Model Implementation

- Based on the linear relationship between the dependent and independent variables present in our data, we implemented following models on our data.

❑Linear Regression

❑Lasso Regression with Cross-validation

❑Ridge Regression with Cross-validation

❑Elastic Net Regression with Cross-validation

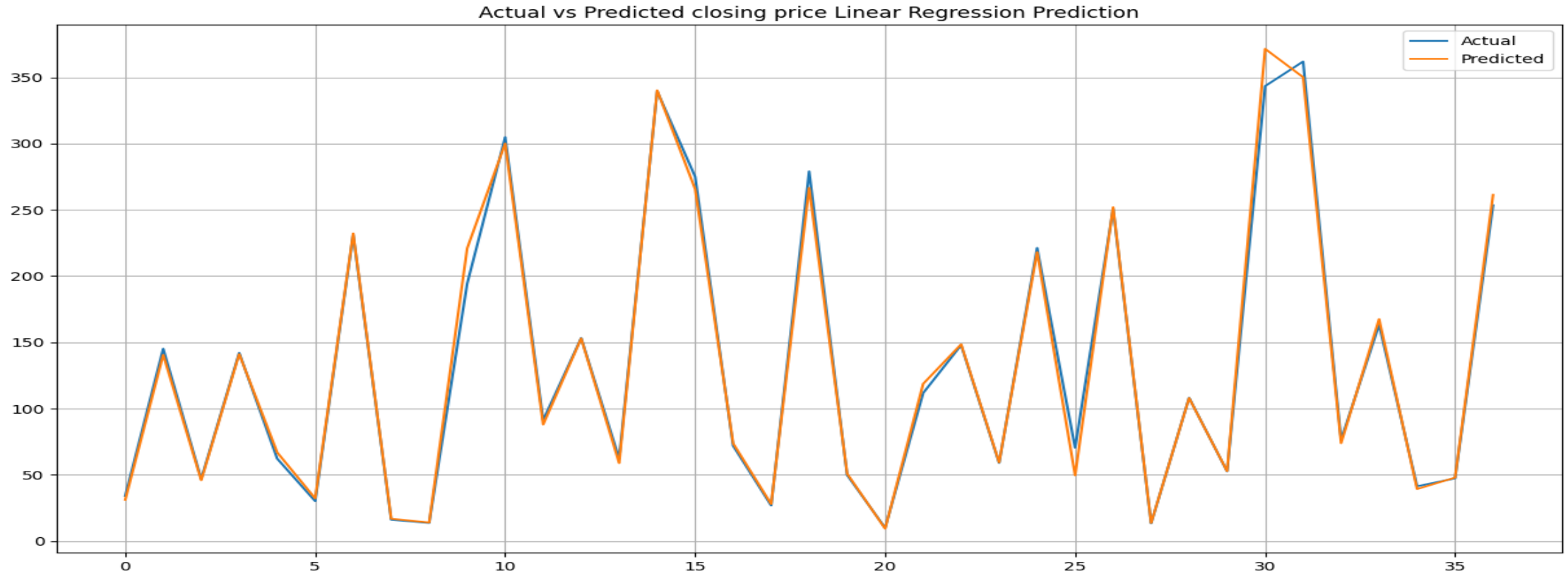
We fit these models on training data, learn the model parameters and then make predictions on test dataset. Then we check the performance of these models using various evaluation metrics such as :-

➢Mean Absolute error.

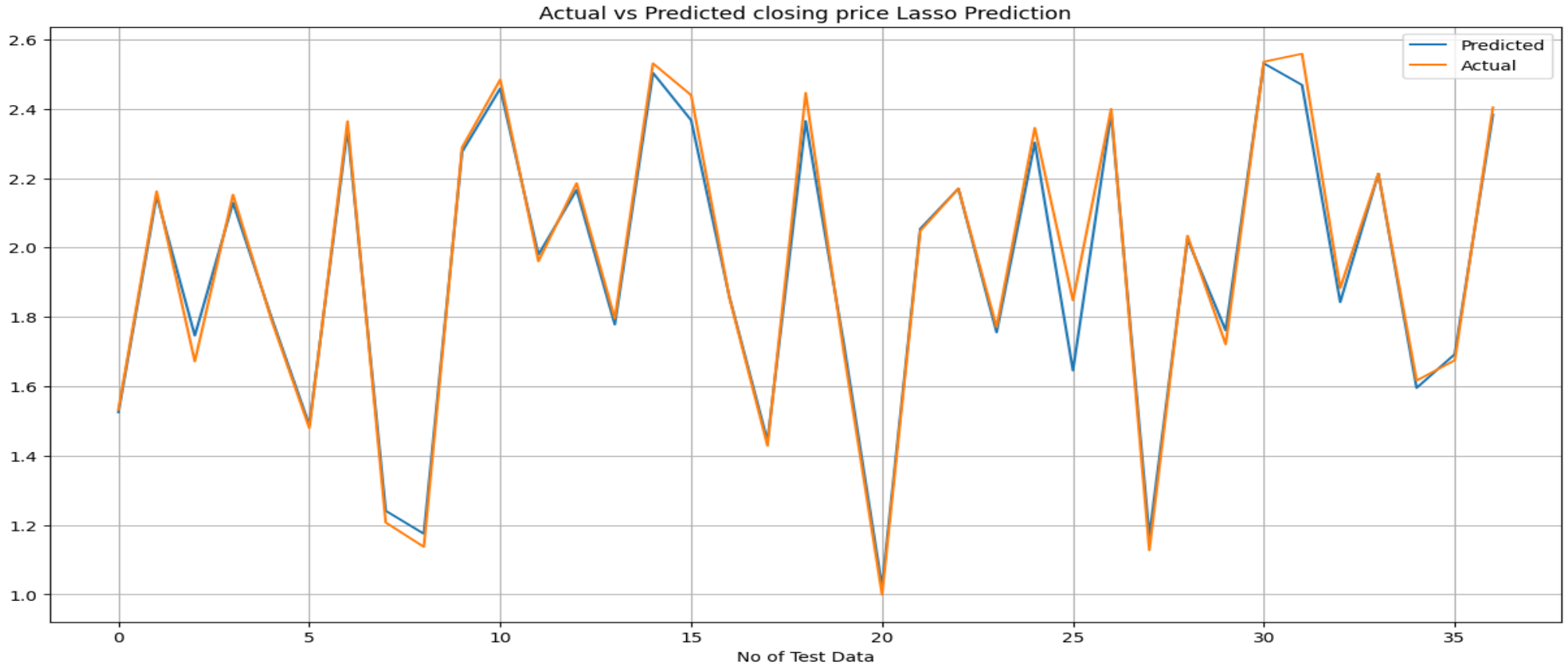
➢Mean squared error and RMSE

➢R-squared and Adjusted R-squared

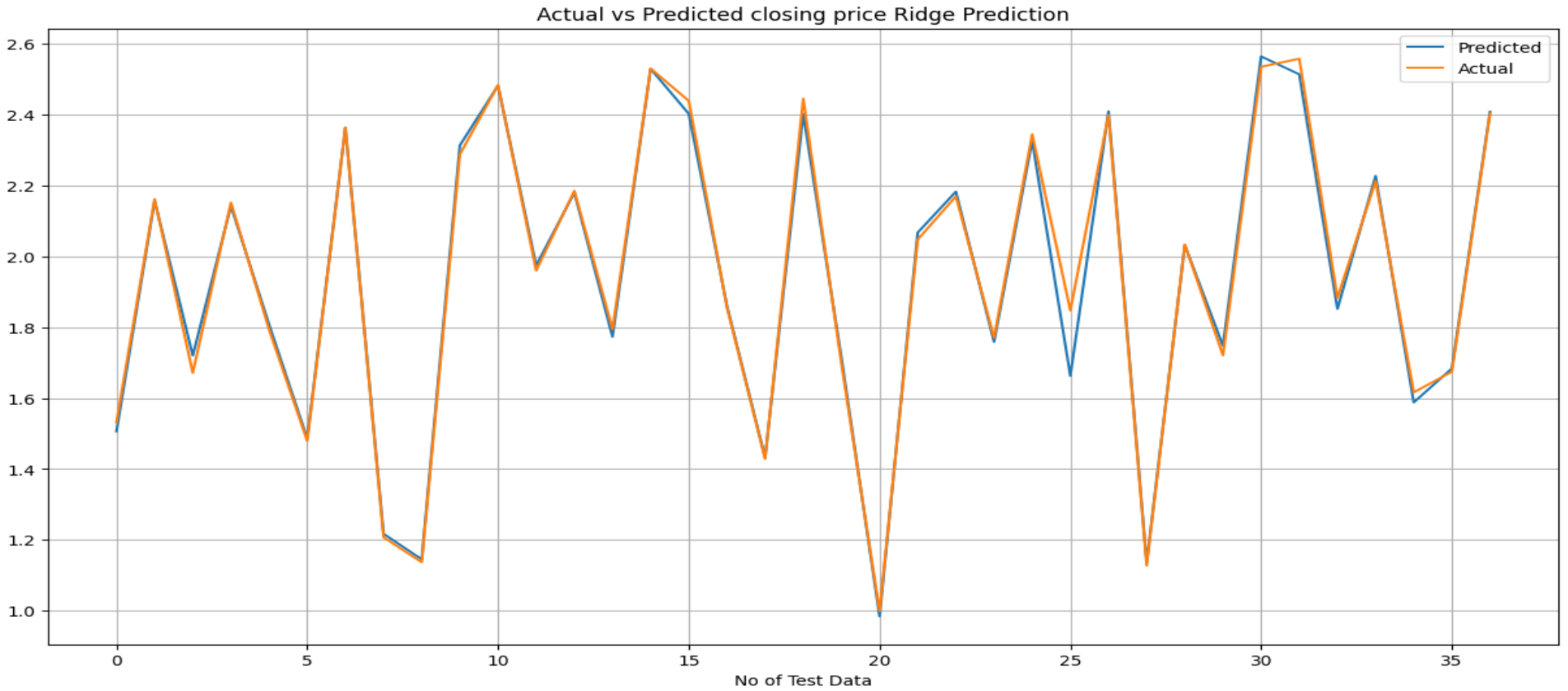
Finally, we select the best performing model based on these metrics.



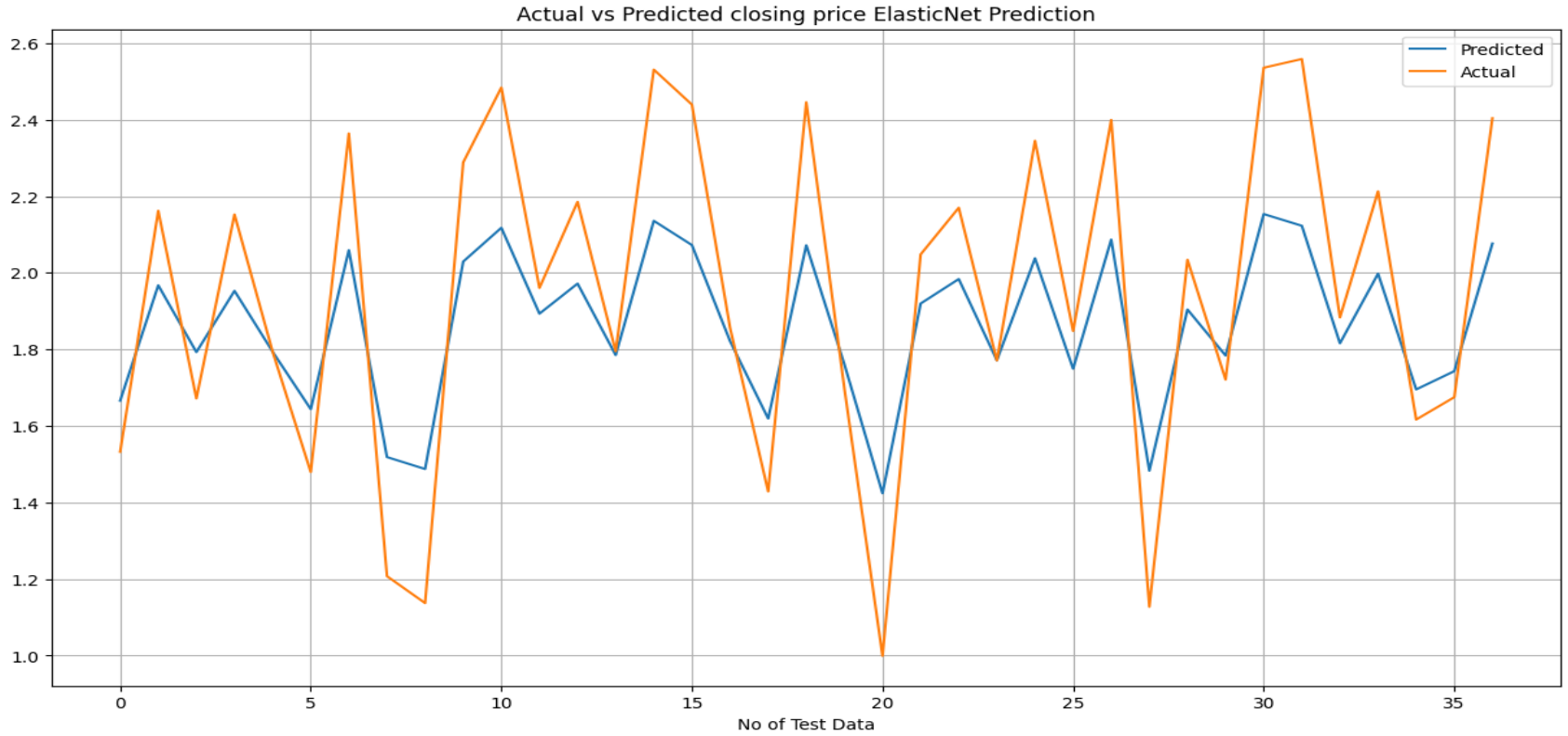
- Our simple Linear Regression Model predicted the closing price with Root Mean squared error(RMSE) of `0.03151`
- R2 score of this model is `0.994`
- Adjusted R2 score has the value 0.994 for this model. Which tells us that around 99.4 percent of the variance in our dependent variable is attributable to the independent variables.



- Our Lasso Regression Model predicted the closing price with Root Mean squared error of 0.04722.
- R2 score of this model is 0.988
- Adjusted R2 score has the value 0.998 for this model. Which tells us that around 99.8 percent of the variance in our dependent variable is attributable to the independent variables.

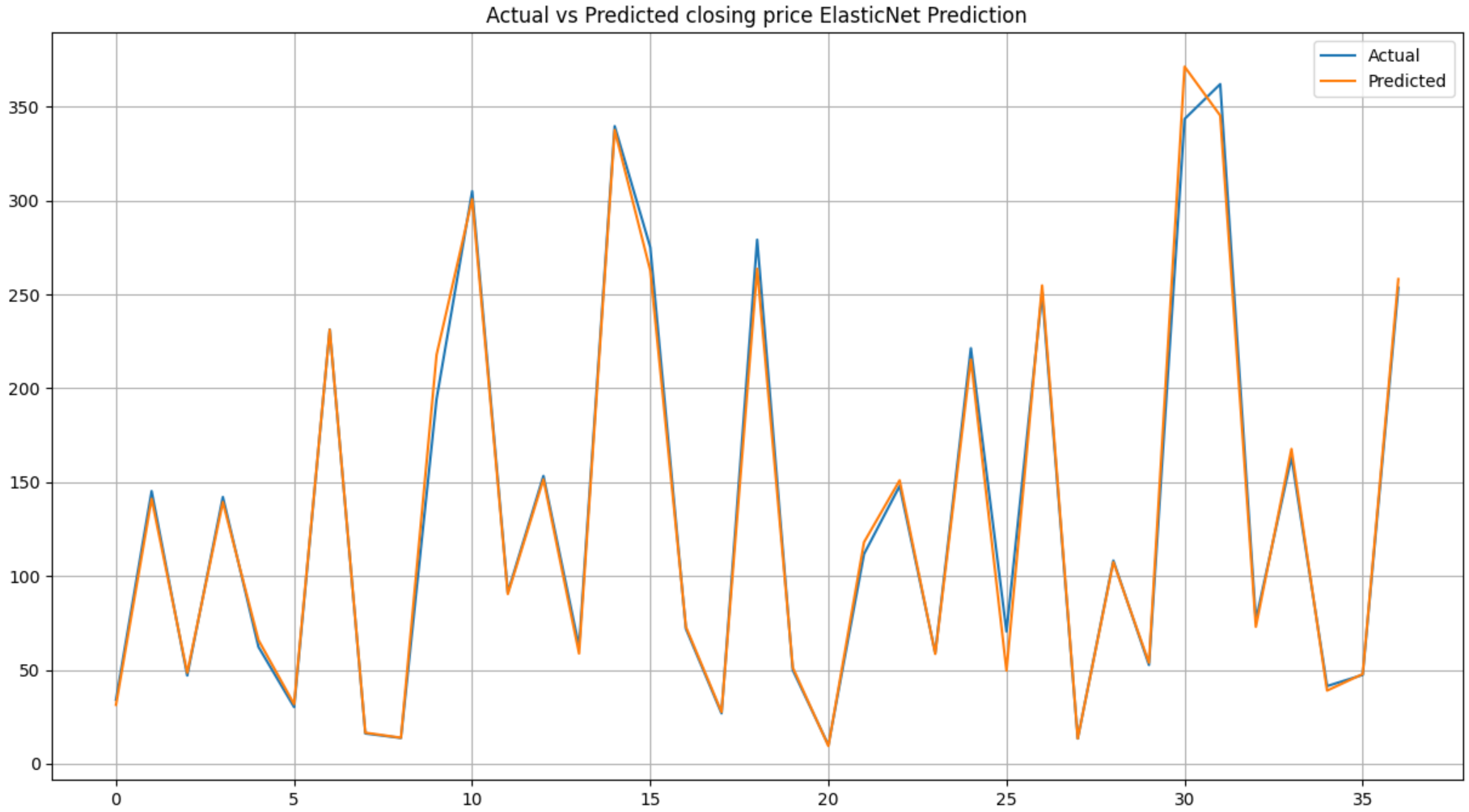


- Our Ridge Regression Model predicted the closing price with Root Mean squared error of 0.03662.
- R2 score of this model is 0.992.
- Adjusted R2 score has the value 0.992 for this model. Which tells us that around 99.2 percent of the variance in our dependent variable is attributable to the independent variables.



- Our Elastic Net Regression Model predicted the closing price with Root Mean squared error of **0.0307**
- R2 score of this model is **0.994**
- Adjusted R2 score has the value 0.994 for this model. Which tells us that around 99.4 percent of the variance in our dependent variable is attributable to the independent variables.

ElasticNet Prediction vs Actual (After Validification)



Evaluation Metrics:

	Linear Regression	Lasso	Ridge	Elastic-Net
MAE	0.01856	0.030441	0.01794	0.01814
MSE	0.0009933	0.002230	0.000933	0.00094
RMSE	0.0315190	0.04722	0.030555	0.03079
R-SQUARE	0.99466	0.98802	0.99498	0.99490
ADJUSTED R-SQUARE	0.995172	0.9954	0.9927	0.6723

From all the above models for Lasso and ElasticNet regression the evaluation metrices for test dataset are almost close to each other. So as per my understanding we can use elasticnet regressor for now.

Conclusion:

- The dataset does not have any null values/missing values as well as duplicate values which made the analysis easy and smooth.
- I started with univariate analysis in which it can be seen that all the variables were positively skewed.
- In the section of bivariate analysis it can be seen that all the independent variables are having linear relationship with the target variable.
- While analysing the close price with date it can be seen that there was huge fall in the stock prices after year 2018.
- In the correlation heatmap chart it can be clearly seen that all the variables are highly correlated to each other which is a problem for linear regression.
- In the box plot section it can be seen that the independent variables are having some outliers.
- Also the date column was formated to year-month-date-format.
- To tackle the outliers, skewness and multicollinearity problem the data was transformed to log10 value and a new feature as average which is the mean of the prices for each row was generated.
- At last I have tried to implement 5 models in order to predict the closing stock prices and finally found that ElasticNet regression model is the best performing model since it has better r2 score value as well as other evaluation metrics values.