Capstone Project - 2

Yes Bank Stock Price Prediction

Team Members

Sanjay Malviya
Biswajeet Sethi
Aleem Khan
PVN Malleswara Rao

Who doesn’t want to predict stock prices?

Let’s go ahead and try!

1. Premise
2. Objective
3. Data pipeline/pre-processing
4. EDA
5. Models
8. Prediction metrics
9. Conclusion

Premise

Yes Bank is a household name in India, at least after the fiasco it was involved in recently. It was alleged that the owner had authorised reckless sanctioning of loans which were later siphoned off to his own personal accounts. Whatever it may be, the health of the company was being reflected in its stock price movements. This makes the stock price a good indicator of a company's health. And at the same time, it also begs the question whether we can predict its prices in some way to reap the benefits?

© picture-alliance/AP Photo/A. Solanki

Objective

We have been provided with a dataset with the monthly stock price details for Yes Bank. The objective of this project has been to apply different models to check whether the prices/movement of the stock can be predicted using certain features and/or past performance.

Looking at the various features of the dataset, we can understand the relationships between them and accordingly pass the required parameters in the model to train it and to ultimately predict the closing price according to the trained data.

Data Pipeline

- **Data Preprocessing:** At this stage, we check for duplicate values and missing values and treat them if any. Furthermore, we check the datatype of the features present in our dataset and transform them if necessary
- **Exploratory Data Analysis (EDA):** At this stage, we conduct an EDA on the selected features in order to better understand their spread, pattern and relationship with the other features. It gives us an intuition as to what is going on in the dataset.
- **Model Building:** At this stage, we apply various models to understand which one will give us the best result.

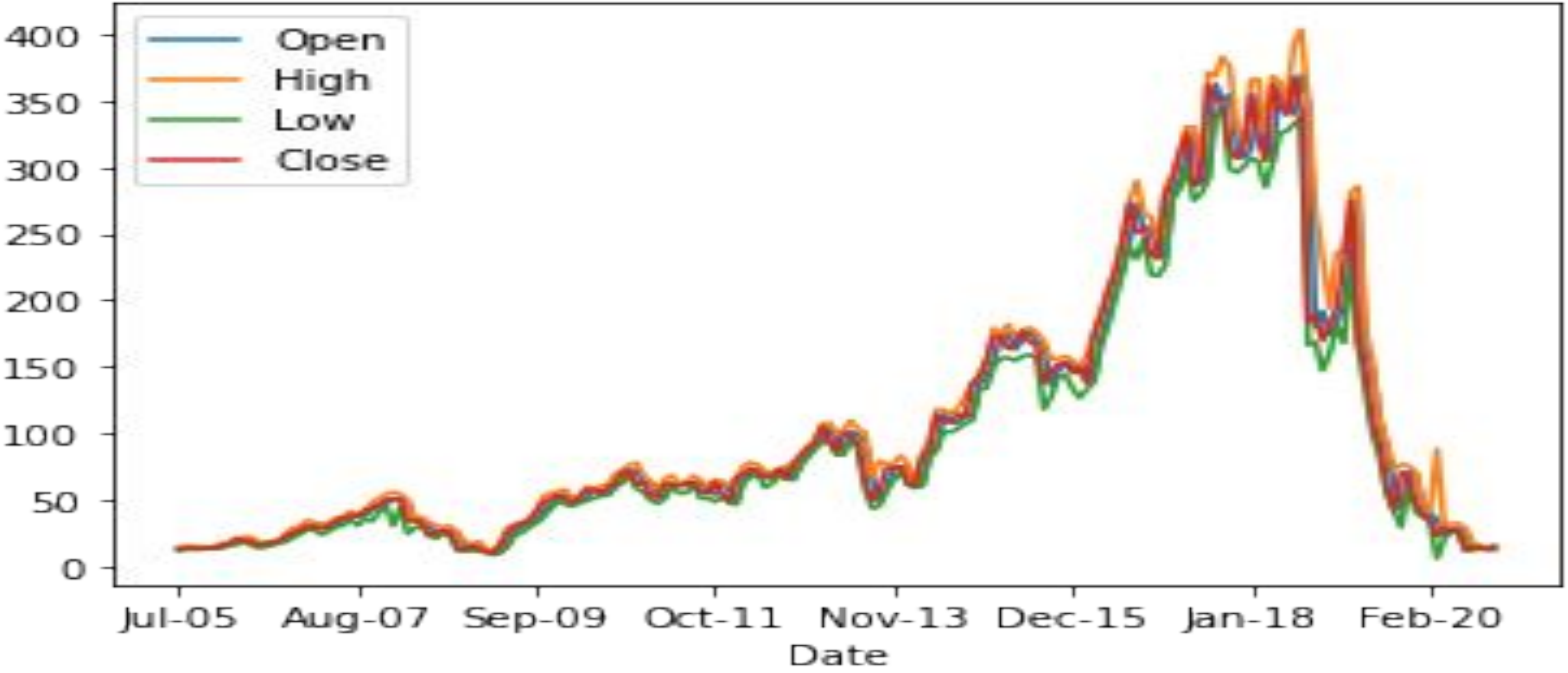
Data Summary

- **Date:** This feature contains the month and year of the respective stock price details. They are in an ascending order in the dataset
- **Open:** This is the price at which the stock opens for the day/week/month. As we have monthly dataset, the opening price will be corresponding to the opening price of the month
- **High:** This is the stock's highest price reached for the day/week/month. As we have monthly dataset, this price will be corresponding to the highest price of the month
- **Low:** This is the stock's lowest price reached for the day/week/month. As we have monthly dataset, this price will be corresponding to the lowest price of the month

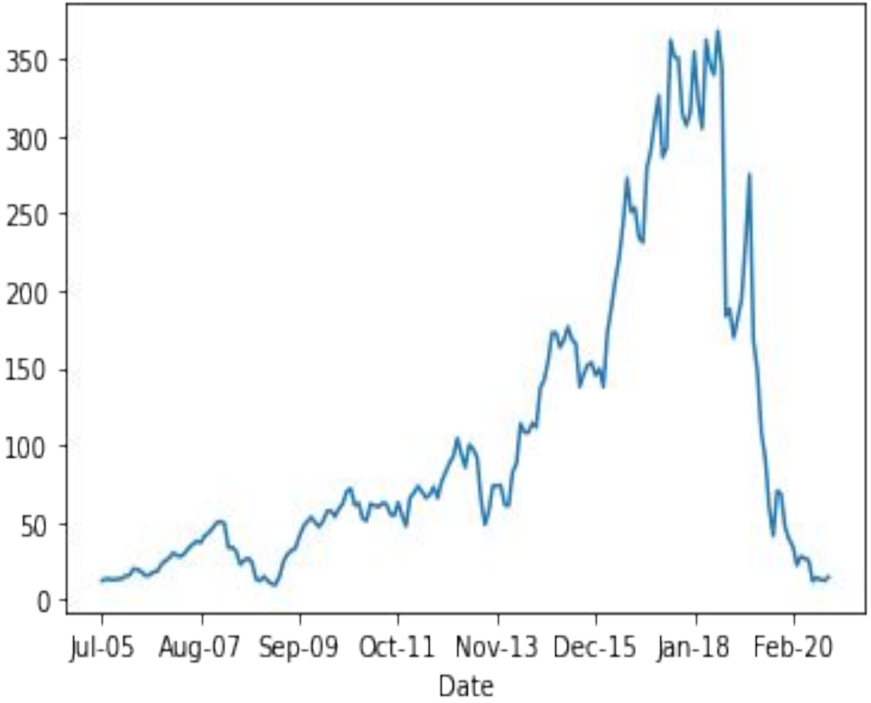
Data Summary

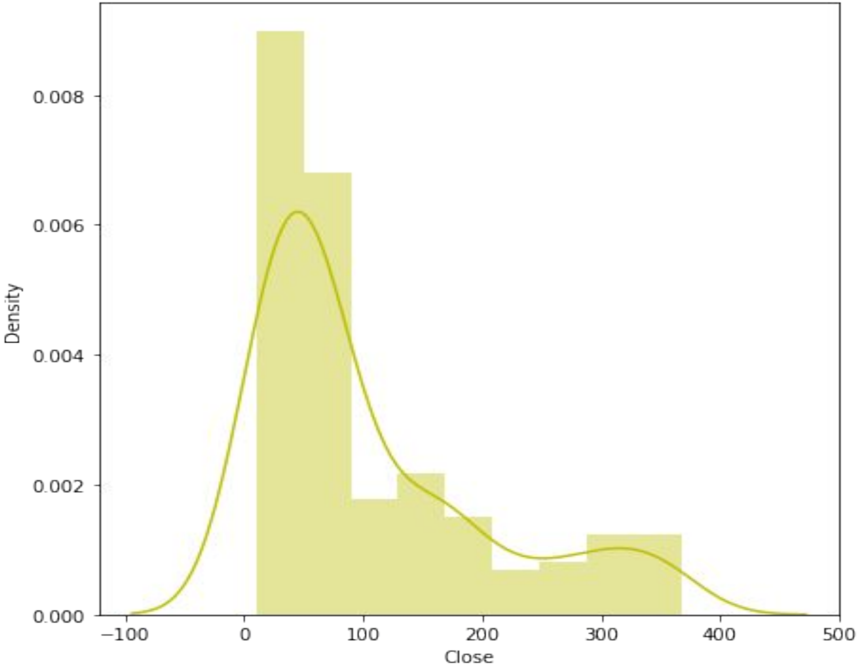
- **Close:** This is the price at which the stock closes for the day/week/month. As we have monthly dataset, the closing price will be corresponding to the final price of the month. We are predicting the final price of the month, hence we will be treating this feature as the dependent variable and the rest of the features and independent variable.

Exploratory Data Analysis (EDA)

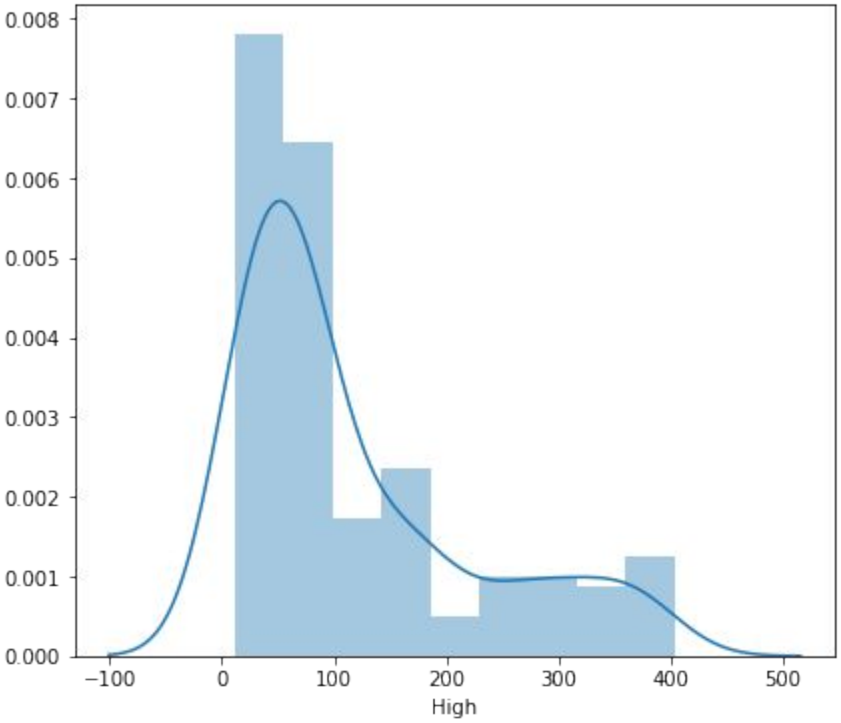


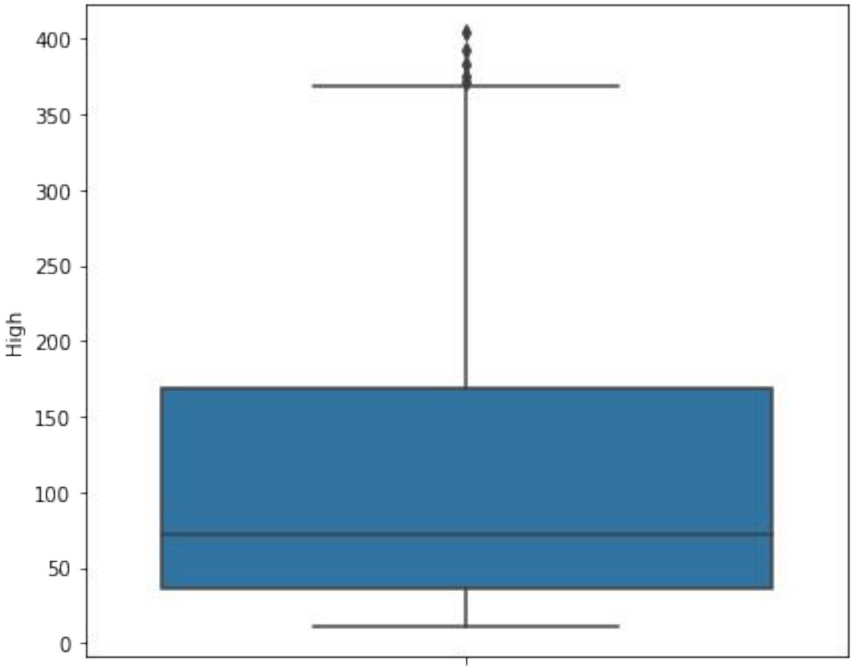
EDA (Continued..)



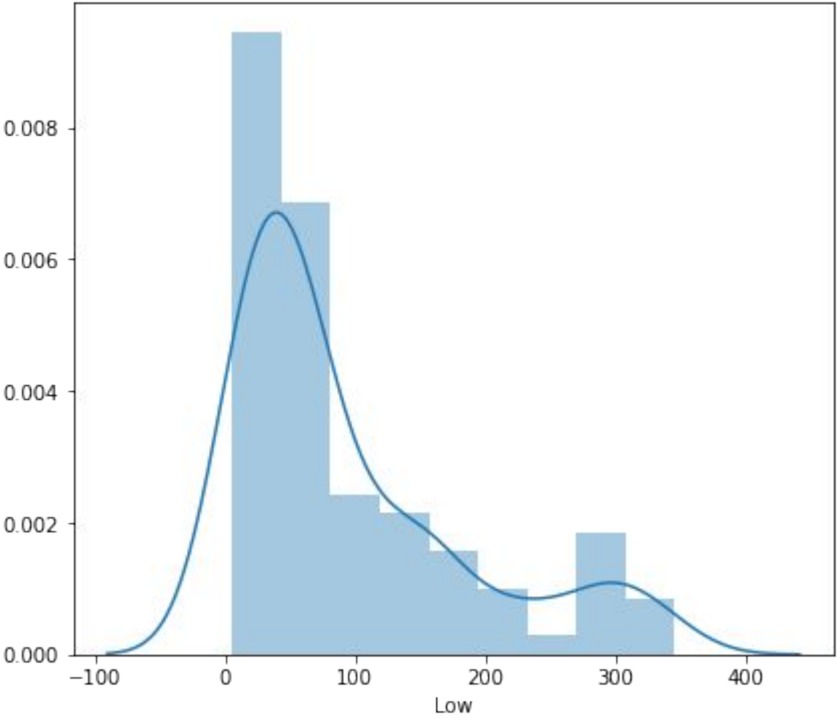


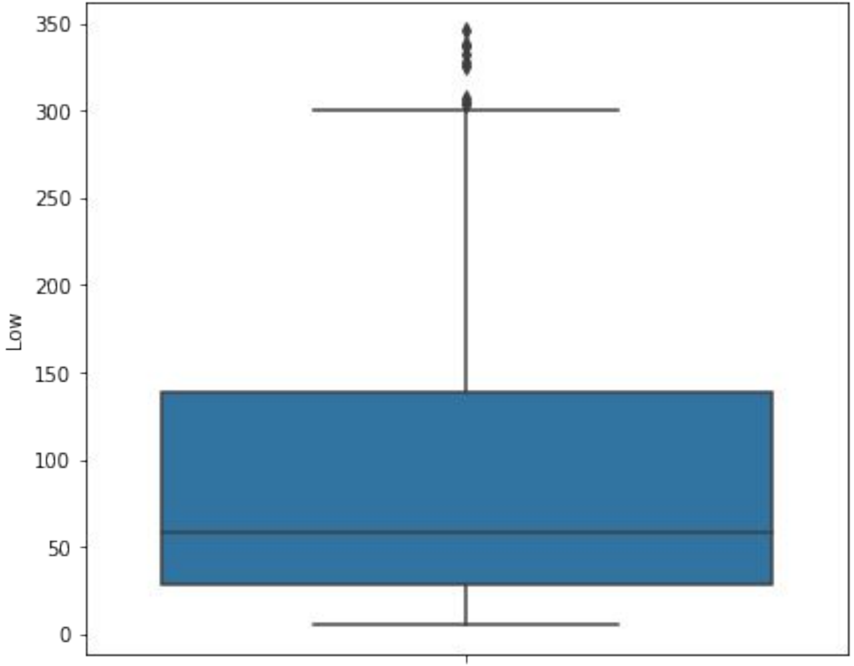
EDA (Continued..)



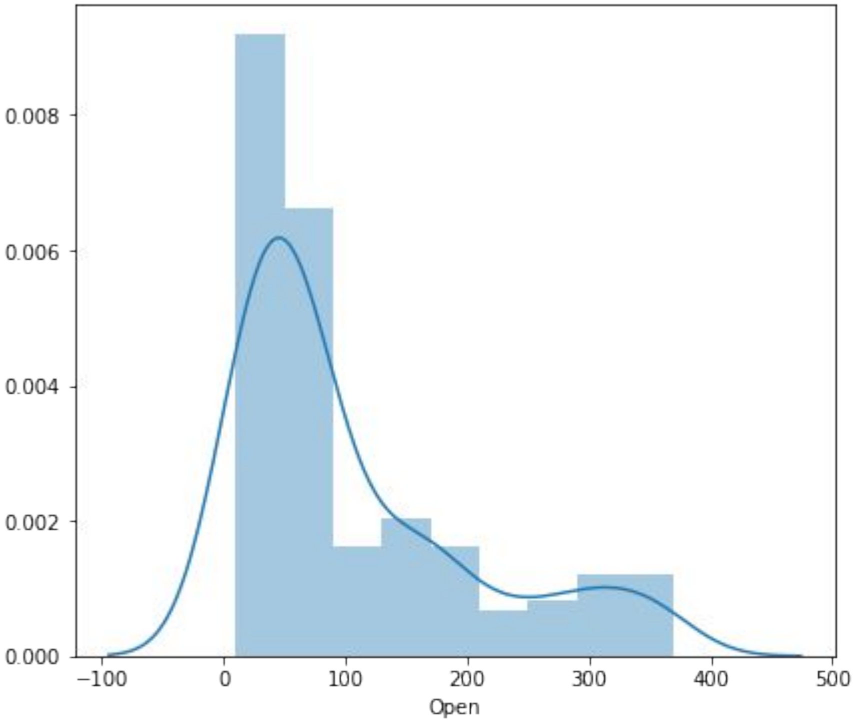


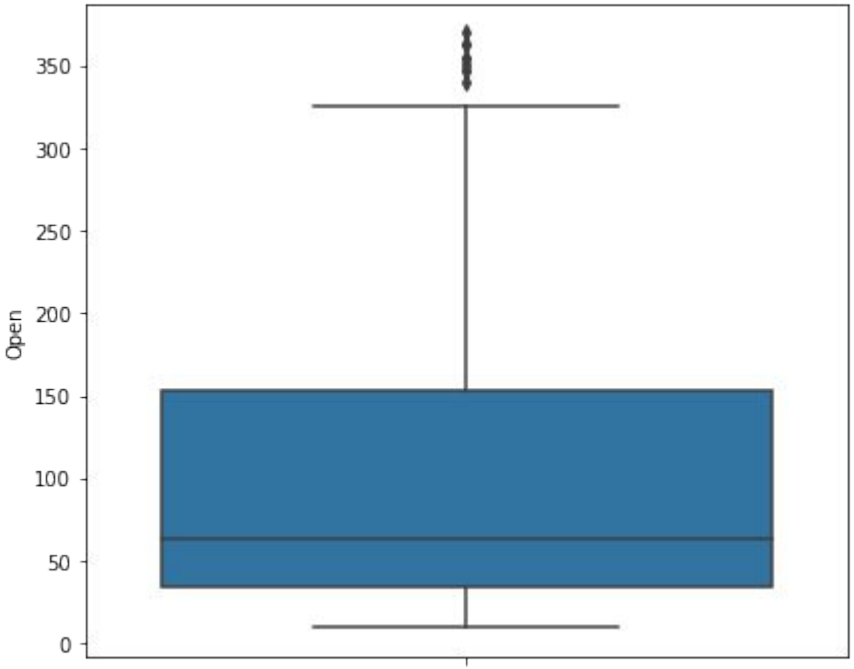
EDA (Continued..)





EDA (Continued..)



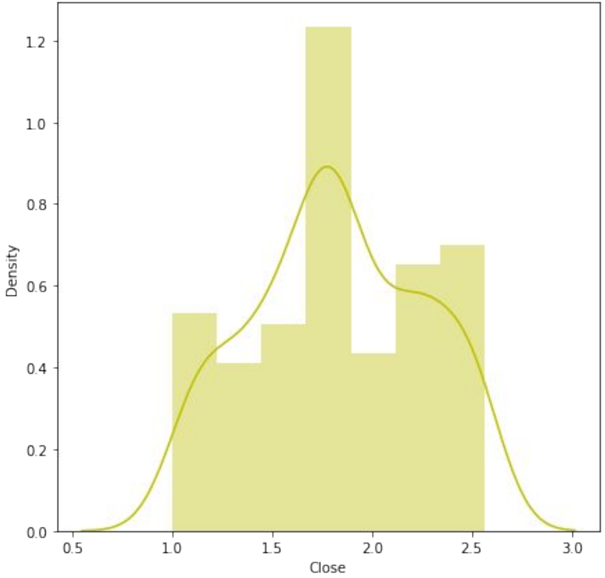


EDA (Continued..)

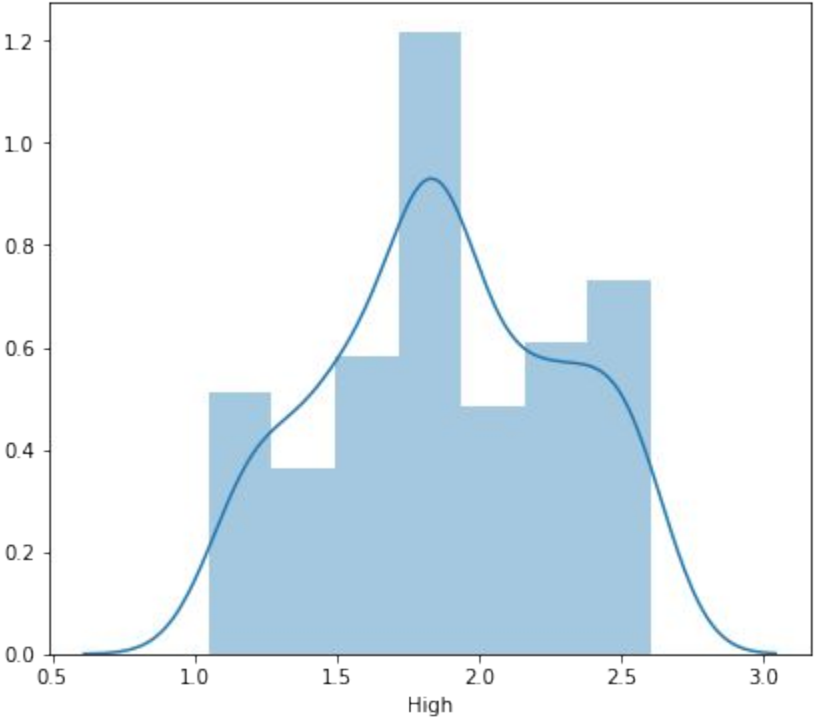
Data Transformation

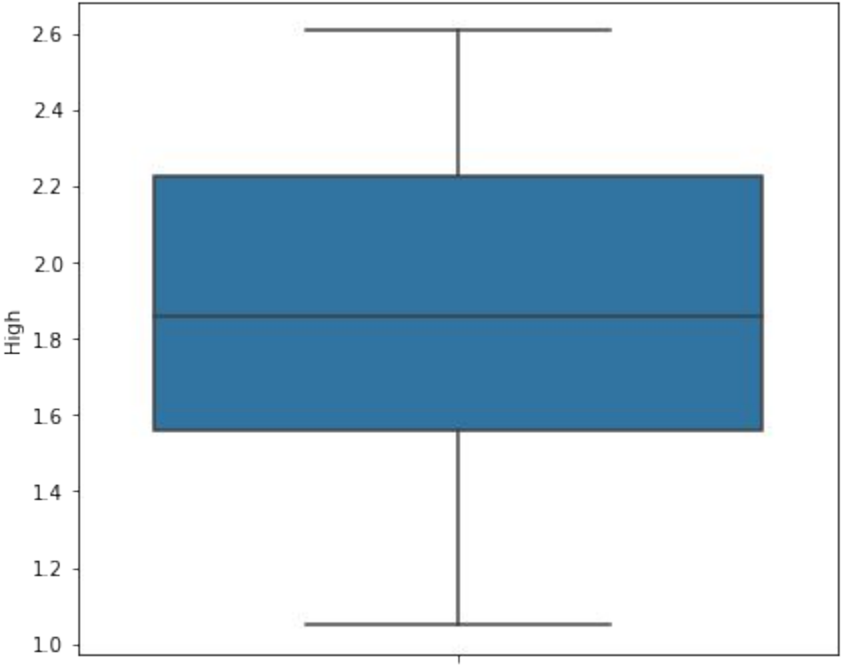
As observed in the preceding slides, the observed data was found to be skewed. We will transform the data to make it uniform before passing it into our machine learning models. Let's have a look at how they will look once the transformation is applied to them. The image on the right shows how the distribution of our close price would look after a log transformation is applied to it.

Following slides gives the shows the distribution and boxplots of the normalized data of the independent variables.

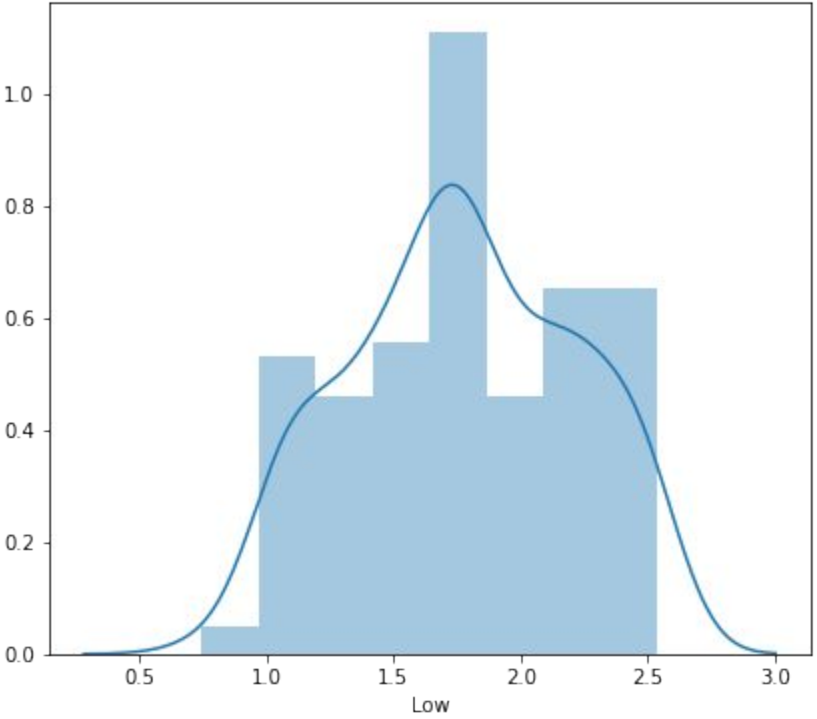


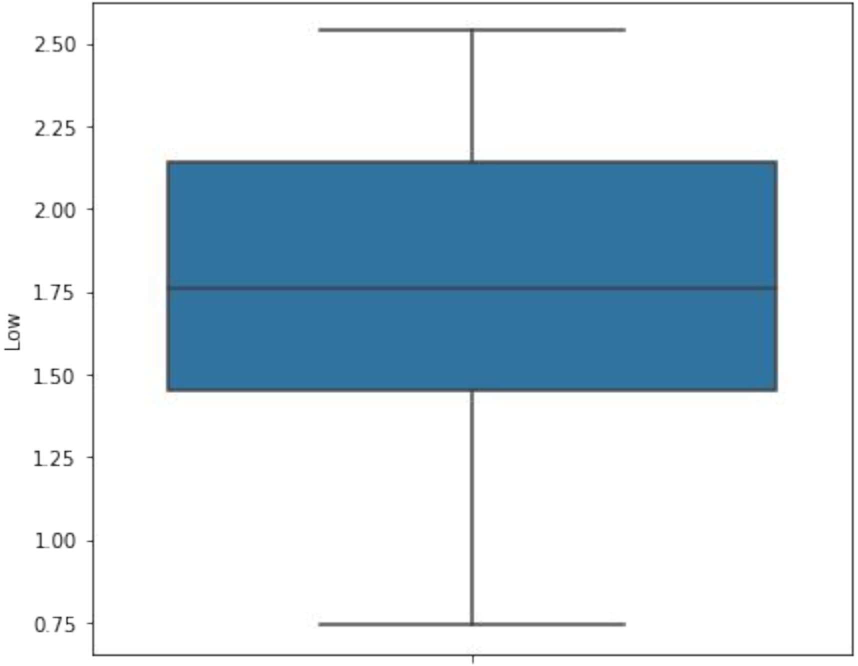
EDA (Continued..)



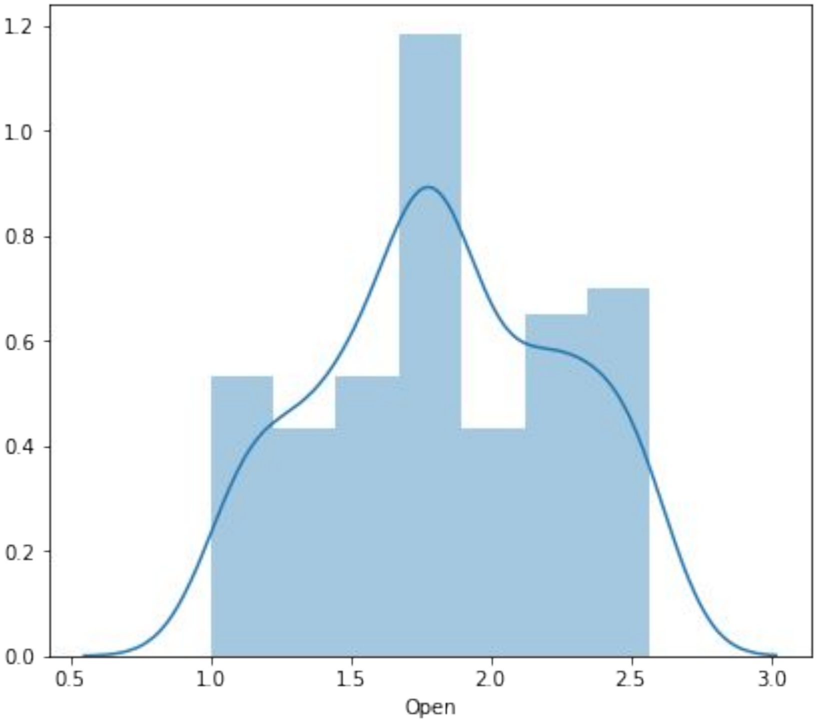


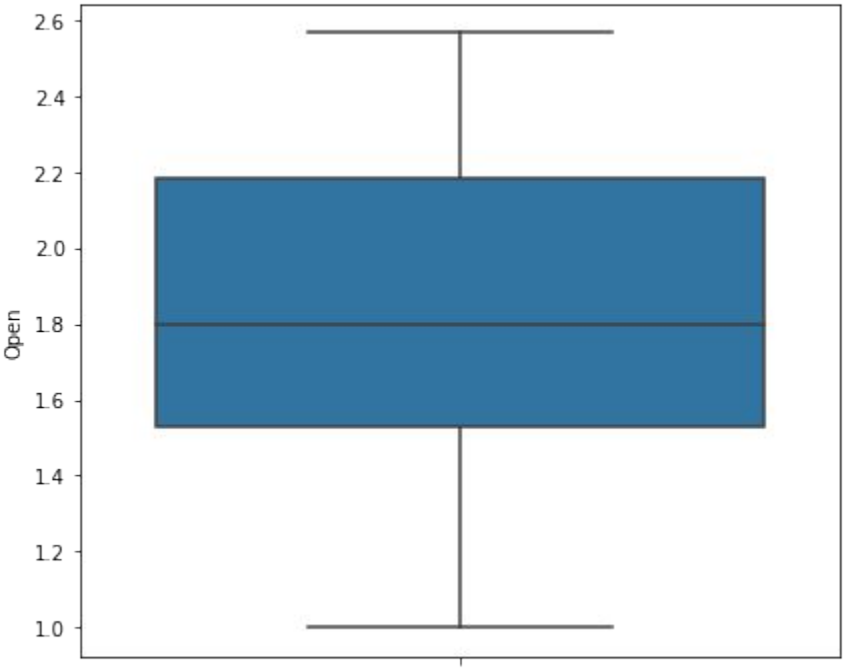
EDA (Continued..)





EDA (Continued..)



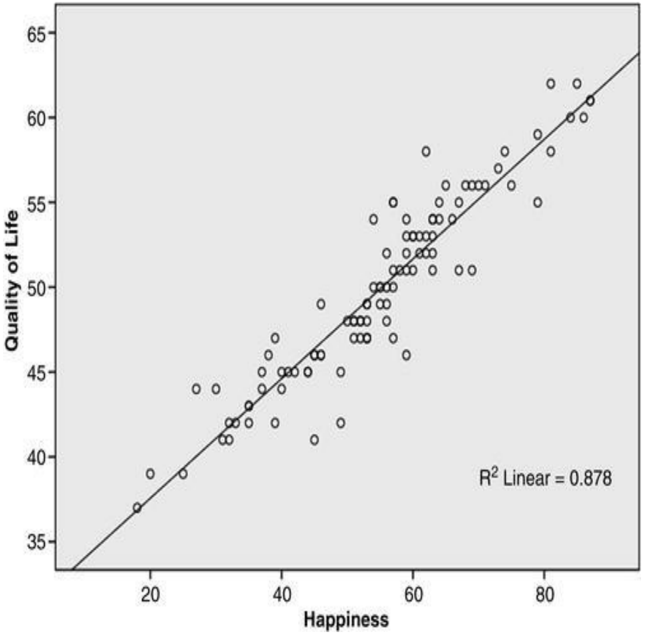


EDA (Continued..)

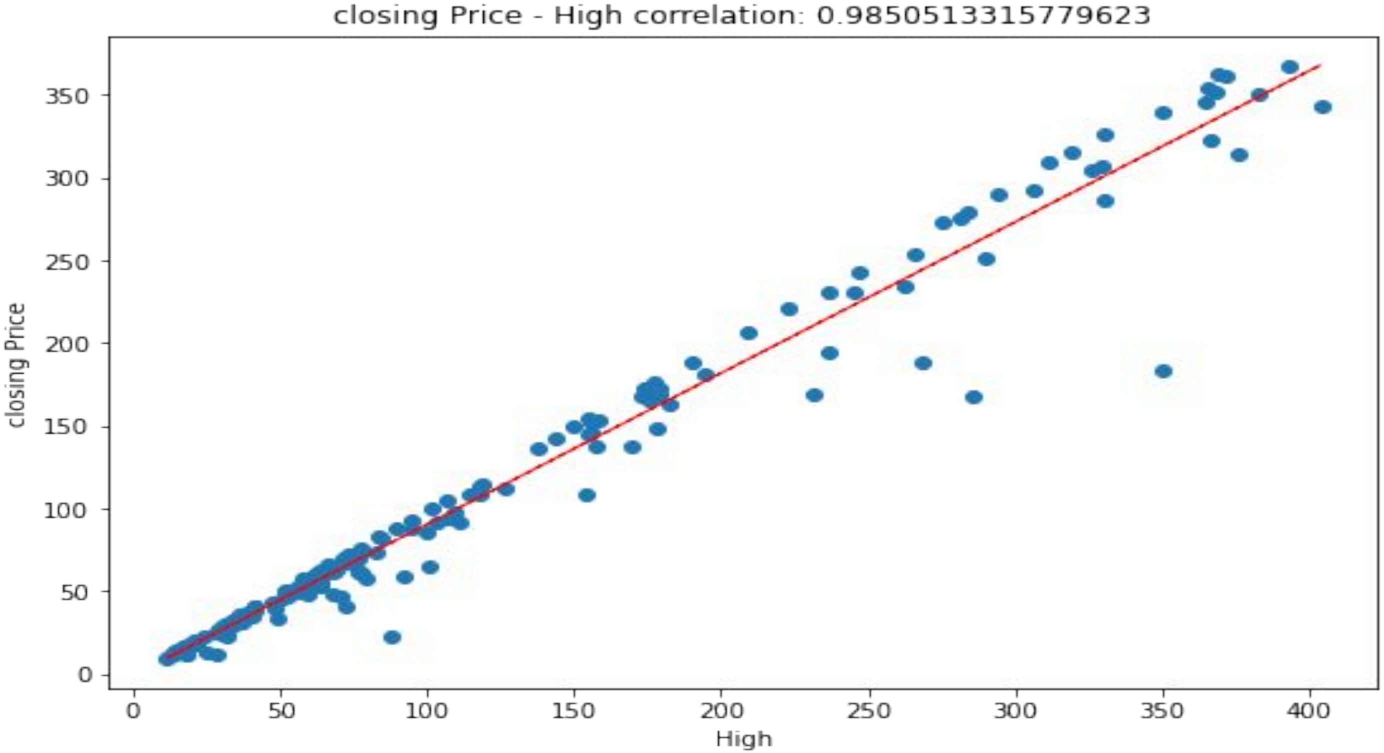
Bivariate Analysis: Bivariate analysis is one of the simplest forms of statistical analysis. It involves the analysis of two variables for the purpose of determining the empirical relationship between them.

Bivariate analysis can be helpful in testing hypothesis of association. Bivariate analysis can help determine to what extent it becomes easier to know and predict a value for one variable if we know the value of the other variable.

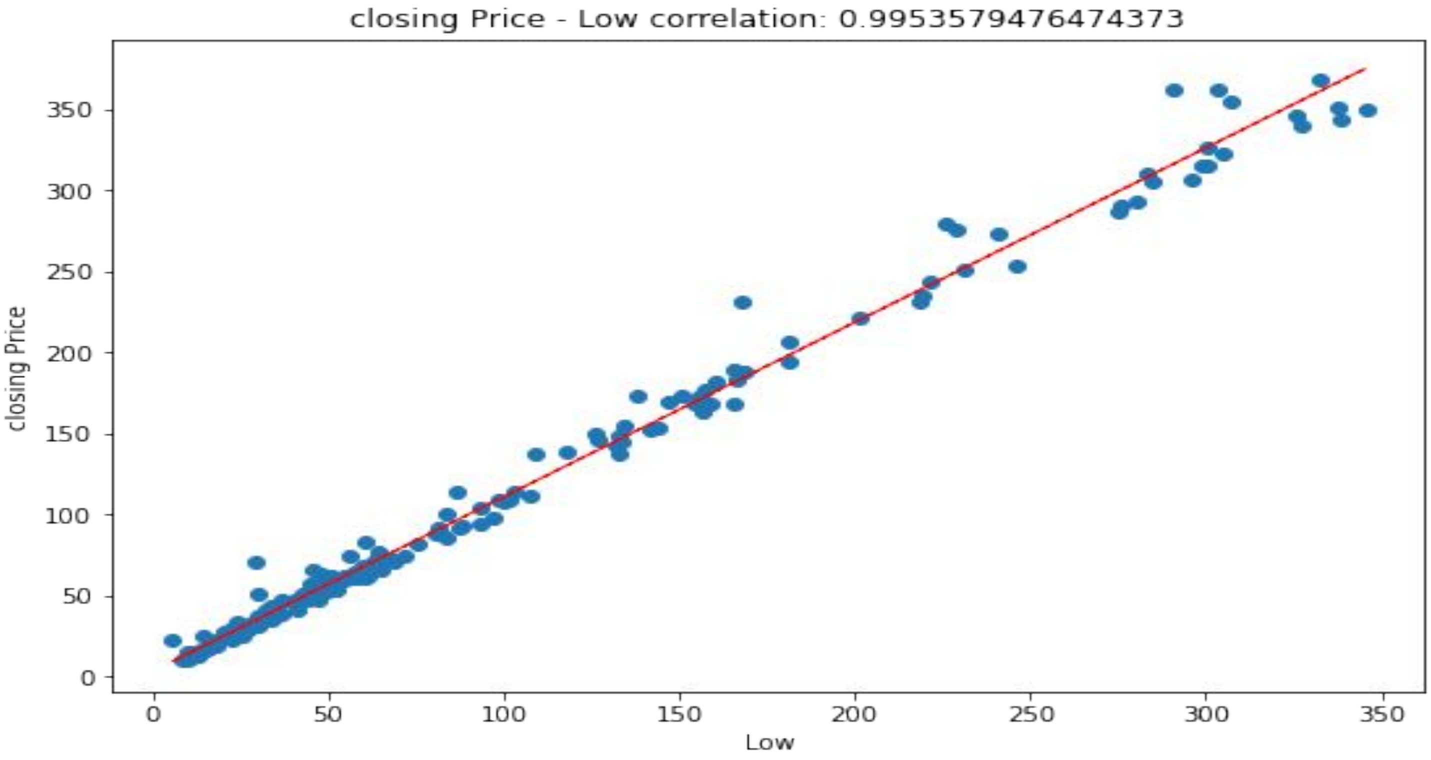
In the following slides, we shall see the dependent variable (close price) being correlated with each of the other independent variables



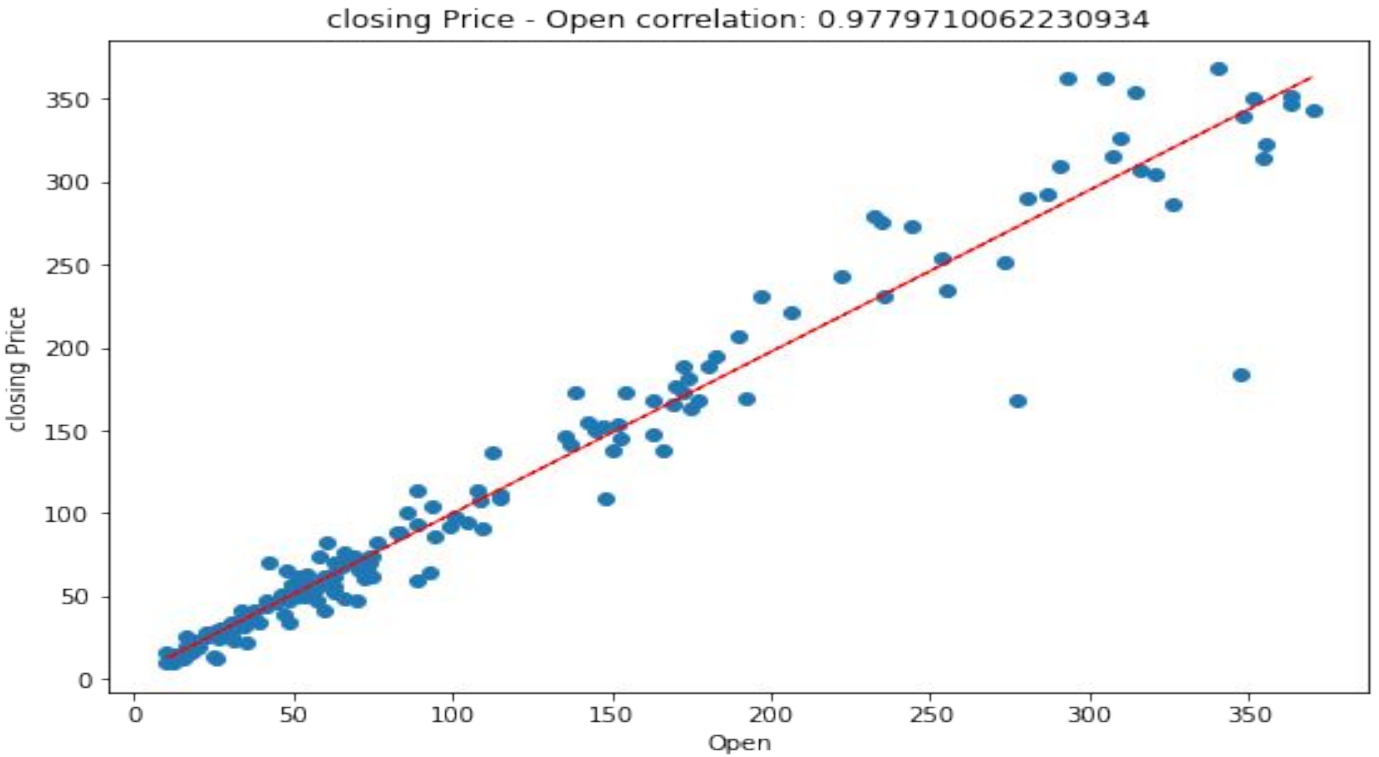
Correlation between the high and the closing price:



Correlation between the low and the closing price:

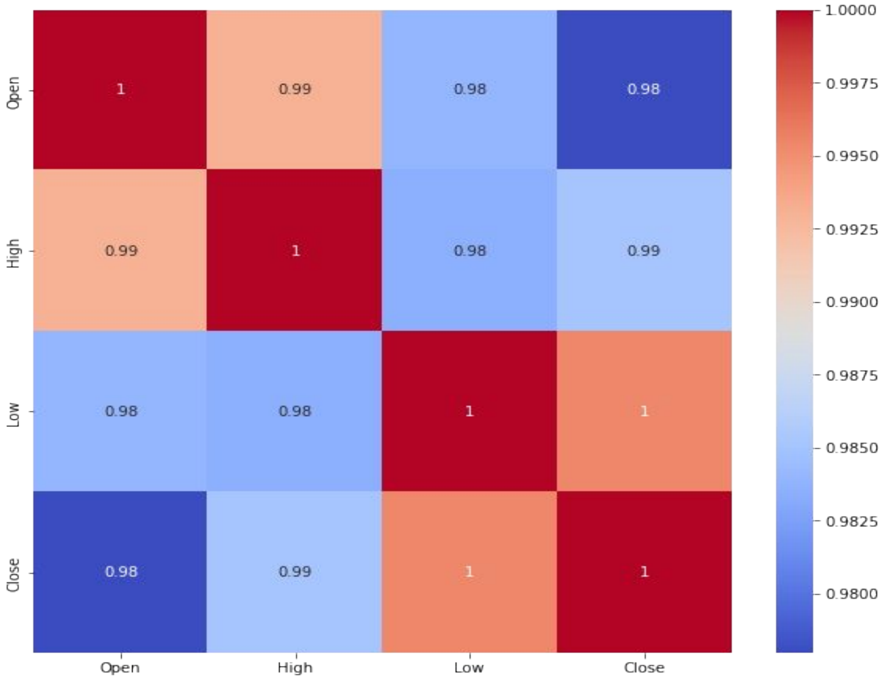


Correlation between the open and the closing price:



Correlation Matrix:

- The correlation matrix helps us visualise the correlation of each parameter with respect to every other parameter.
- The shades changes from the highest to lowest (or vice versa) correlations.
- We can see in the matrix on this slide that our dependent variable (close price) is highly correlated with all the other independent variables



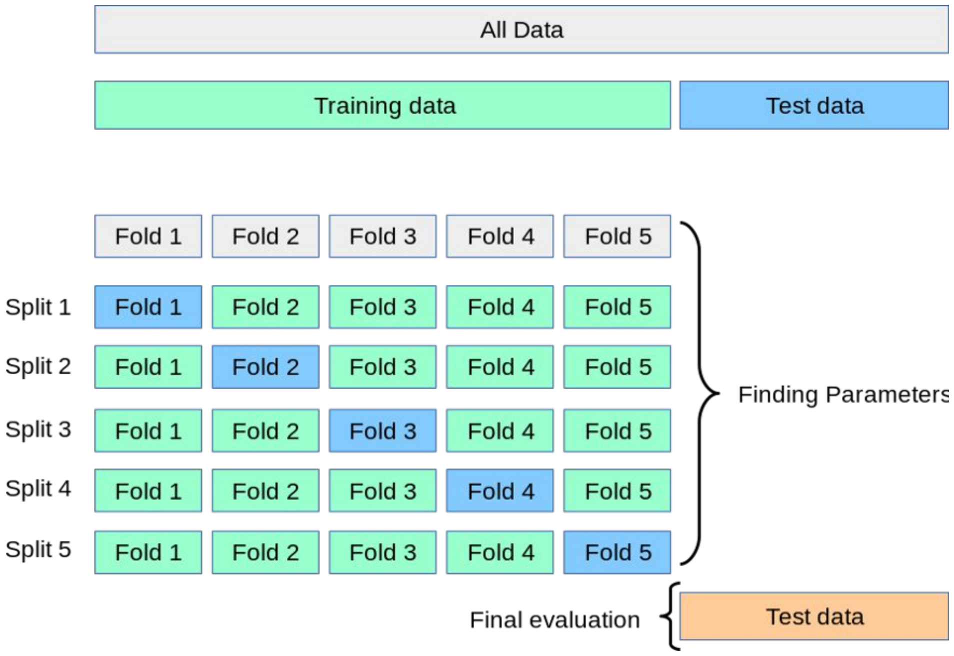
Model Selection

- After transforming our data using MinMaxScaler, we split the data into the training and testing set.
- Furthermore, we passed the data into the Linear Regression model, the Lasso regression model and the Ridge Regression model.
- We checked the the performance of the model across various parameters.

Grid Search Cross Validation

Grid Search CV tries all combinations of parameters grid for a model and returns with the best set of parameters having the best performance score.

We conducted a Cross-Validation in the ridge and lasso regressions to check if any of their set can perform better than the Linear Regression.



All Data

Training data | Test data

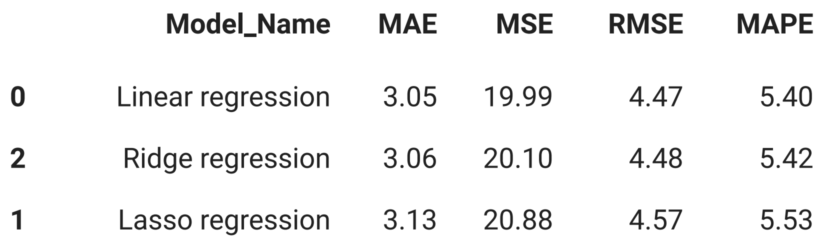
	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 1	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 2	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 3	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 4	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Split 5	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5

Finding Parameters

Final evaluation | Test data

Cross Validation

- For better results we cross checked our models with grid search cross validation
- In grid search we take random data in five different ways from the dataset and we get the best values for both lasso and ridge regression.
- In each iteration, the train and test splits dataset are varied and the model gives the respective outputs
- The best output is then taken as final output of the model for both lasso and ridge regression
- Linear regression still performed the best of all the models.

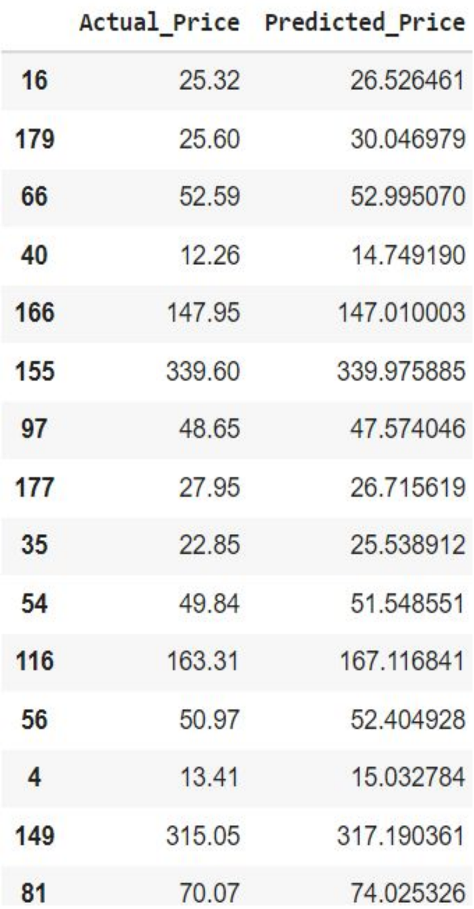


	Model_Name	MAE	MSE	RMSE	MAPE
0	Linear regression	3.05	19.99	4.47	5.40
2	Ridge regression	3.06	20.10	4.48	5.42
1	Lasso regression	3.13	20.88	4.57	5.53

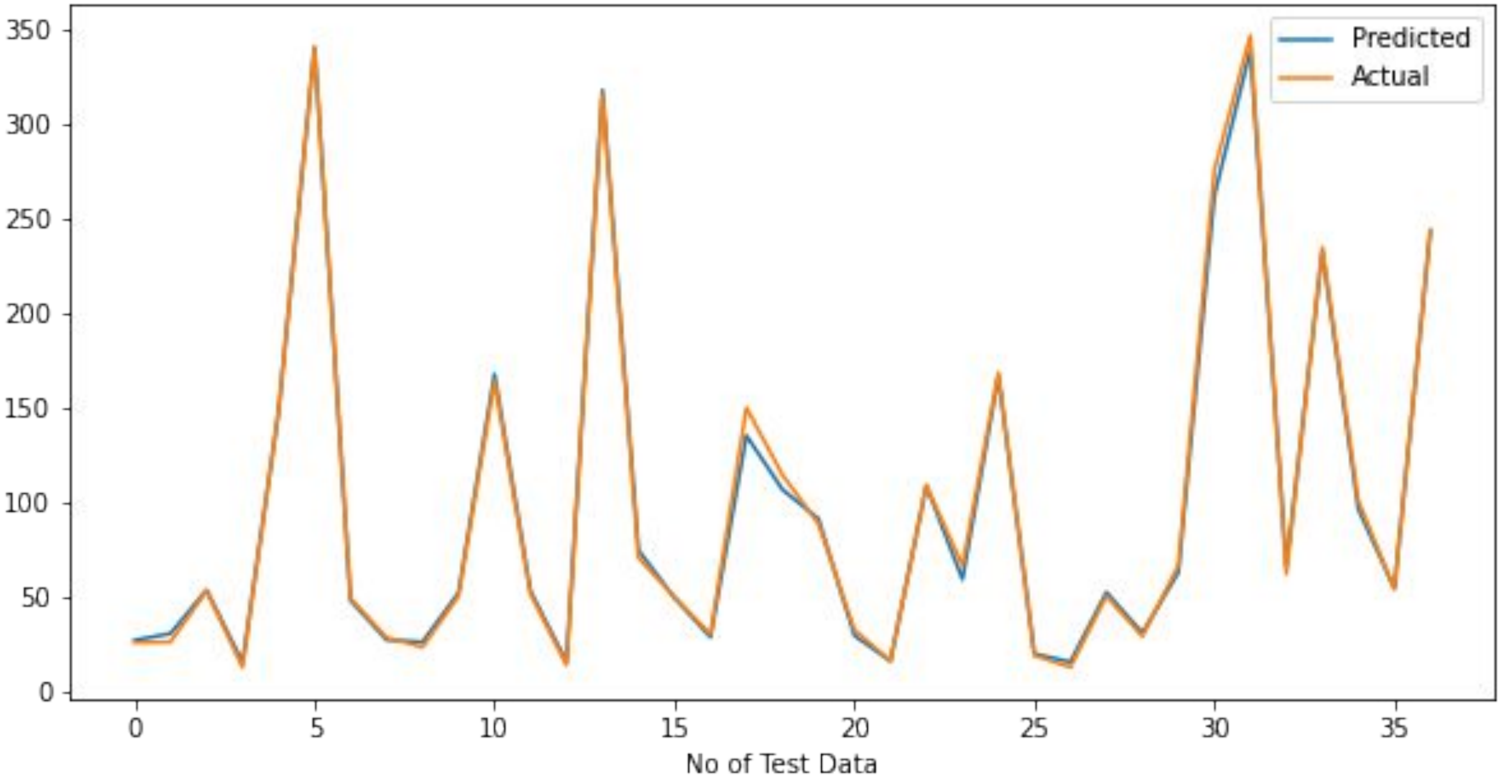
Model Prediction

- We have seen the prediction metrics and it can be concluded from it that the Linear Regression model performs the best so far. The KNN and the XGBoost model has given similar results.

- On this slide can be seen a table with both the actual values and the predicted values of the test set



	Actual_Price	Predicted_Price
16	25.32	26.526461
179	25.60	30.046979
66	52.59	52.995070
40	12.26	14.749190
166	147.95	147.010003
155	339.60	339.975885
97	48.65	47.574046
177	27.95	26.715619
35	22.85	25.538912
54	49.84	51.548551
116	163.31	167.116841
56	50.97	52.404928
4	13.41	15.032784
149	315.05	317.190361
81	70.07	74.025326


350
300
250
200
150
100
50
0
0
5
10
15
20
25
30
35
No of Test Data
Predicted
Actual

Conclusion

- The dependent variable is highly correlated with all the independent variables. Hence, we used all the parameters given in the data set as independent variables to train our models
- Of all the models, Linear Regression performed the best, with lowest MAE, MSE, RMSE and MAPE scores.
- Ridge regression shrunk the parameters to reduce complexity and multicollinearity, but ended up affecting the evaluation metrics.
- Lasso regression did feature selection and ended up giving up worse results than ridge which again reflects the fact that each feature is important
- KNN AND XGBoost have given similar results. We are already getting good results with simple models like Linear Regression. If it was otherwise, models like XG Boost could have further improved the model with relevant hyperparameter tuning.

Challenges

- We have used the open price, high and low prices of the month as the independent variable to predict the dependent variable.
- This approach, though might work in theoretical sense, is not a very practical approach for real world problem solving.
- If we want to predict the closing price of a particular month (current or future), there is no way of knowing their high and low prices in order to enter those in our model.
- It is possible that such problems can be better tackled by using time series forecasting