



# Abstract

This project aims to predict the closing price of Yes Bank's stock, a critical task for stakeholders and investors. Given the bank's challenges, including bad loans and fraud cases leading to regulatory intervention, predicting stock prices is complex. Utilizing a comprehensive dataset with monthly stock price metrics, the project builds predictive models, incorporating dynamic trends and uncertainties. By evaluating various modeling techniques, including time series and regression methods, the goal is to provide valuable insights for informed investment decisions and a better understanding of Yes Bank's financial performance.

**Key Terms:**

Closing Price, Stakeholders, Investors, Regulatory Intervention, Comprehensive Dataset, Predictive Models, Time Series Models, Regression Methods, Financial Performance.

# 1. Introduction

The project's main objective is to tackle the challenge of predicting Yes Bank's stock prices accurately, considering the bank's historical data since its inception. The dataset contains essential monthly metrics like closing, starting, highest, and lowest prices, enabling the development of predictive models. These models will use various techniques, including time series and regression methods, to capture the intricate dynamics and trends in Yes Bank's stock prices, considering the bank's turbulent events and uncertainties.

The project aims to evaluate the performance of these predictive models in forecasting Yes Bank's stock's closing price effectively. Special emphasis will be given to how well the models can incorporate the impact of significant events, such as fraud cases involving the bank's founders or regulatory interventions by the Reserve Bank of India.

By successfully achieving accurate predictions, this project will offer valuable insights to stakeholders, helping them make informed decisions regarding their investments in Yes Bank. Furthermore, it seeks to navigate the complexities and uncertainties surrounding the bank's stock prices, ultimately contributing to a better understanding of Yes Bank's financial performance and aiding in effective decision-making processes.

# Yes Bank Stock Closing Price Dataset

In the dataset, each entry contains information about a specific month's stock price for Yes Bank. The data includes the following key metrics:

- Date: Indicates the month for which the stock price data is recorded.

- Open: Represents the stock's price at the start of the trading day when the stock exchange opens.

- High: Indicates the highest price that the stock reached during that particular month.

- Low: Indicates the lowest price that the stock reached during that particular month.

- Close: Represents the stock's price at the end of the trading day when the stock exchange closes.

## Problem Definition:

The primary aim of this project is to create a robust and accurate predictive model capable of forecasting the closing price of Yes Bank's stock effectively. The project faces challenges in comprehending and capturing the intricate dynamics and trends in the stock prices, especially considering the historical pattern of an increasing price followed by a sudden decline after 2018.

Addressing multicollinearity in the dataset is a critical challenge during the model development process. Multicollinearity arises when independent variables exhibit high correlation, leading to difficulties in interpreting the model and potentially affecting prediction accuracy. Therefore, the model should employ techniques to handle multicollinearity and ensure the proper consideration of independent variables in making predictions.

Moreover, the model must take into account significant events that have influenced Yes Bank's stock performance, including fraud cases involving the bank's founders and regulatory interventions by the Reserve Bank of India. These events can have substantial impacts on stock prices, necessitating the accurate representation of their effects in the predictive model.

The model's performance goal is to achieve a high level of accuracy in forecasting the closing price of Yes Bank's stock. The benchmark of 99% accuracy set by the K-Nearest Neighbors (KNN) Regression model serves as a target for the developed model to strive towards. Attaining high accuracy will provide valuable insights to stakeholders, investors, and market participants, empowering them to make well-informed decisions and effectively manage their investments in Yes Bank's stock.

In conclusion, this project aims to build a predictive model that effectively addresses the complexities and challenges associated with forecasting Yes Bank's stock prices. The ultimate objective is to offer stakeholders a reliable tool that enhances their understanding of the stock's future performance, supporting them in making informed investment decisions.

## 2. Data Cleaning and Preparation

Data cleaning and preprocessing play a crucial role in data analysis. They involve transforming raw data into a more comprehensible, valuable, and optimized format.

## 2.1 Why is Data Cleaning and Preprocessing required?

Data often contains noise in the form of outliers, null values, or data points that deviate significantly from the overall trend. Prior to conducting exploratory data analysis, it is essential to address and treat these noisy data points.

## 2.2 Available Dataset:

Before proceeding with any analysis, it is necessary to explore and clean the raw dataset of data\_YesBank\_StockPrices. This initial step is crucial to make the data usable and suitable for exploratory data analysis.

## 2.3 Steps Performed:

In **Step 1**, we created a function called "getinfo()" to gather information about the datasets. This function displays various attributes for each column, including the data type, total number of records, count of null values, count of non-null values, percentage of null values, and the number of unique values. We applied this function to both the Play Store dataset and the User Reviews dataset to understand their characteristics.

In **Step 2**, we noticed that the "Rating" column had a significant number of null values, accounting for approximately 15% of the total records. Instead of dropping these null values, we decided to replace them with the mode of the Rating column to maintain data integrity.

Moving on to **Step 3**, we encountered special characters such as "$" and "+" in the "Installs" column, and "M" and "K" in the "Size" column (representing MB and KB, respectively). To ensure consistency, we removed these special characters using a for loop. Additionally, we converted KB values to MB by implementing a function. We also dropped any remaining rows with null values, as they were very few and had a negligible impact on the dataset.

In **Step 4**, we observed that the "Reviews" column, despite being a numerical indicator, was of the "object" data type. To rectify this, we converted the data type of the "Reviews" column to "int" using the "astype(int)" function.

Finally, in **Step 5**, we utilized the "getinfo()" function again and confirmed that there were no remaining null values. However, we did identify duplicate records in both datasets. To address this issue, we eliminated the duplicates from both the Play Store dataset and the User Reviews dataset.

## 3. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is a critical initial step in any Data Science project. Once the raw data has been cleaned, EDA involves examining the dataset to uncover patterns, gain insights, and address future challenges. It encompasses the process of analyzing the dataset to identify trends, outliers, and forming hypotheses based on our understanding of the data.

During EDA, statistical measures are created for numerical data, correlations between numerical values are explored, and various graphical representations and plots are generated using visual methods. These techniques help in gaining a concise and improved understanding of the data. In this article, we will demonstrate EDA using a Google Play Store dataset. By leveraging Python and its essential libraries, we aim to derive meaningful insights and extract valuable information from the dataset.

## 3.1 candle stick graph with price movement

The analysis of Yes Bank stock prices reveals a distinct pattern. Prior to 2018, the stock exhibited a consistent upward trend, indicating positive growth and reflecting investor optimism. However, a significant decline occurred after this period, primarily attributed to the Yes Bank fraud case involving Rana Kapoor, the former CEO.

Leading up to 2018, the stock experienced a continuous rise, demonstrating favorable market conditions and investor confidence. However, the revelation of the fraud case involving Rana Kapoor had a profound impact on the stock's performance. This event marked a turning point, as the stock prices sharply declined.

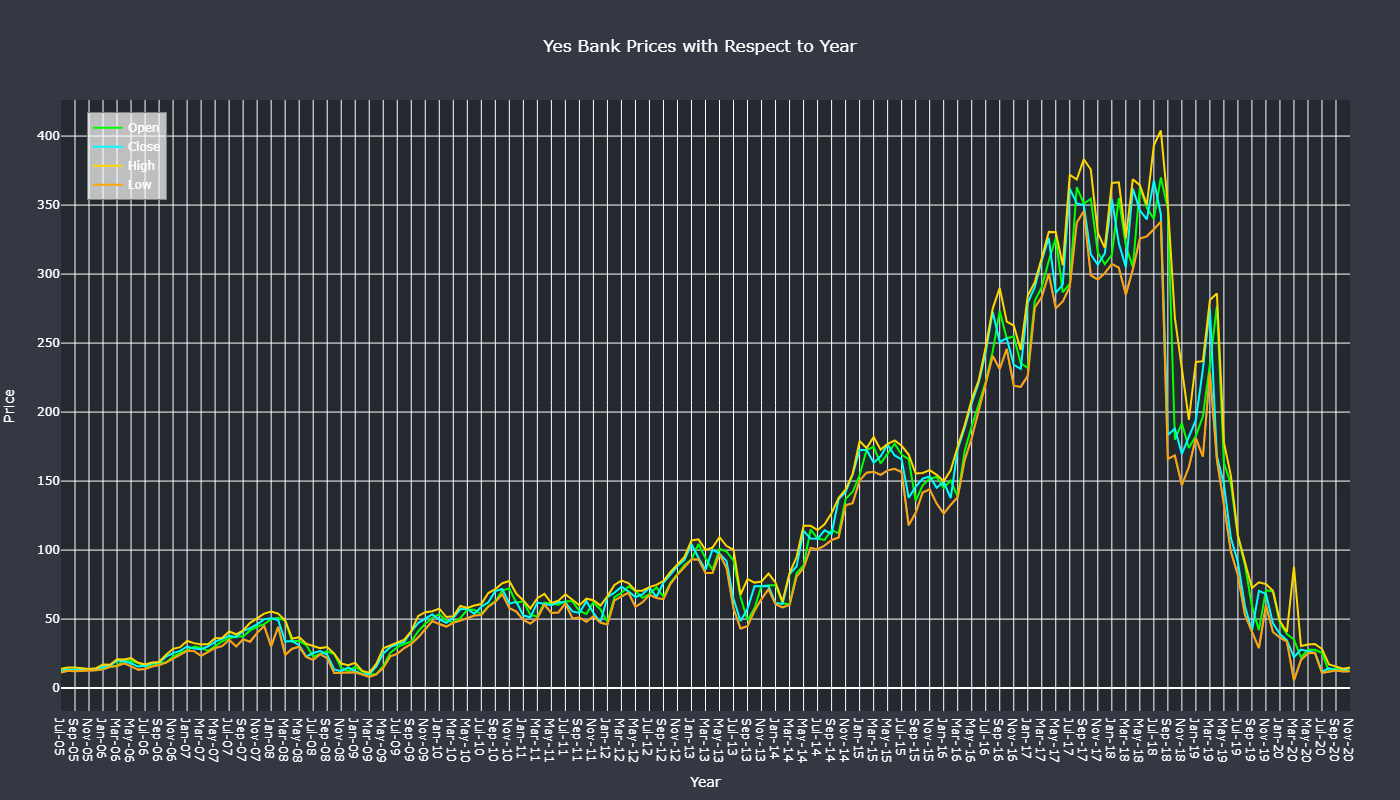


The fraud case involving Rana Kapoor significantly affected investor sentiment, eroding trust and confidence in Yes Bank. Consequently, the stock's value experienced a notable decrease, reflecting the negative repercussions of the scandal on the company's reputation and financial stability.

Overall, the analysis highlights the contrasting trends in Yes Bank's stock prices. Pre-2018, there was a consistent upward trajectory, while the post-2018 period witnessed a significant decline due to the repercussions of the fraud case involving Rana Kapoor.

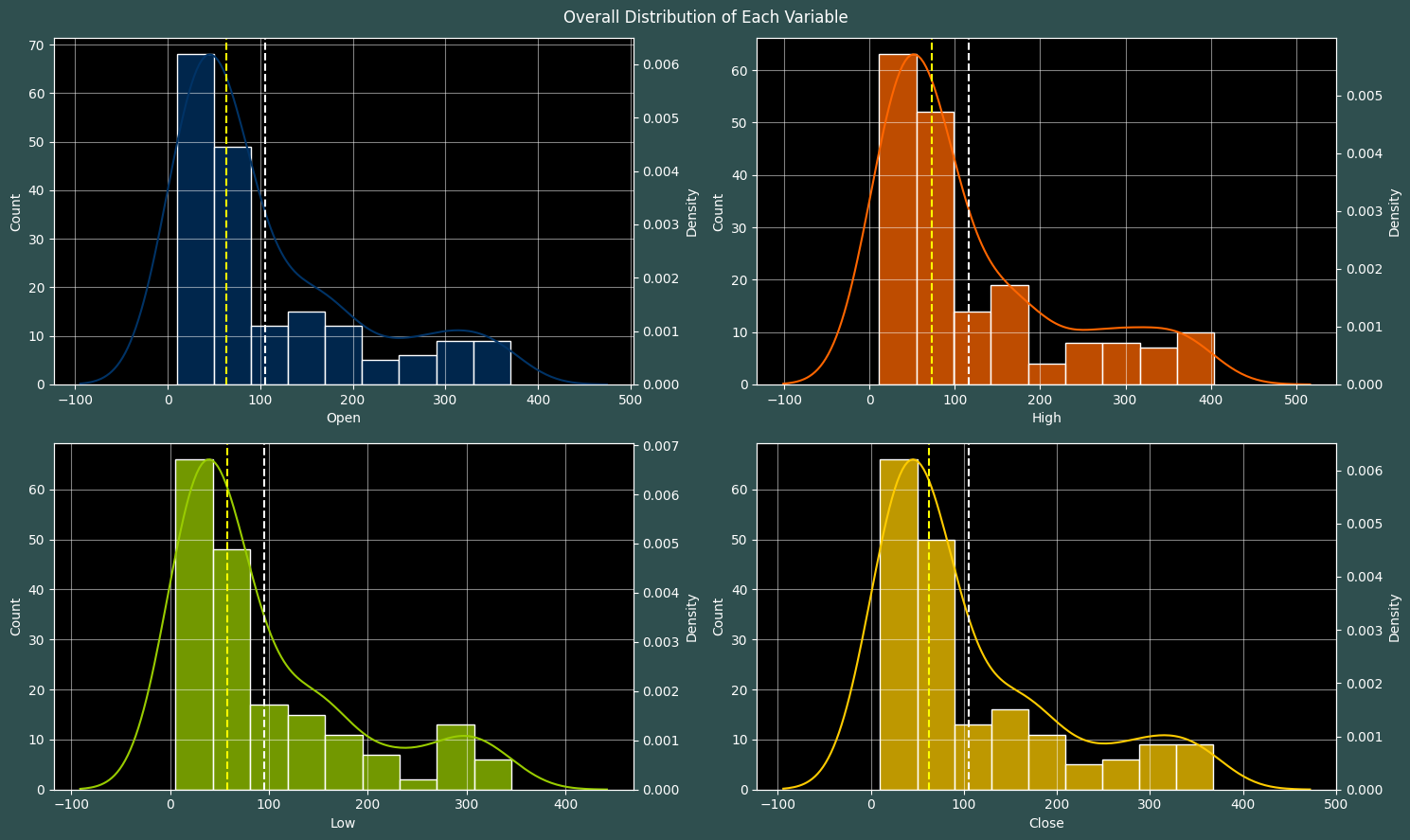
## 3.2 Line plot showcasing variations in each feature over the years

Indeed, the expected dip in the price variables after 2018 is prominently visible in the chart. The line graph shows a notable decrease in the Open, High, Low, and Close prices of the Yes Bank stock following the specified timeframe. This decline can be attributed to various factors, such as the impact of the Yes Bank fraud case involving Rana Kapoor, which adversely affected investor sentiment and confidence in the bank. The scatter plots further accentuate the dip, as they highlight individual data points that deviate significantly from the preceding upward trend. By visually representing the price variables over time, the chart effectively showcases the substantial decrease in prices after 2018, emphasizing the challenging period faced by Yes Bank and the subsequent decline in its stock value.



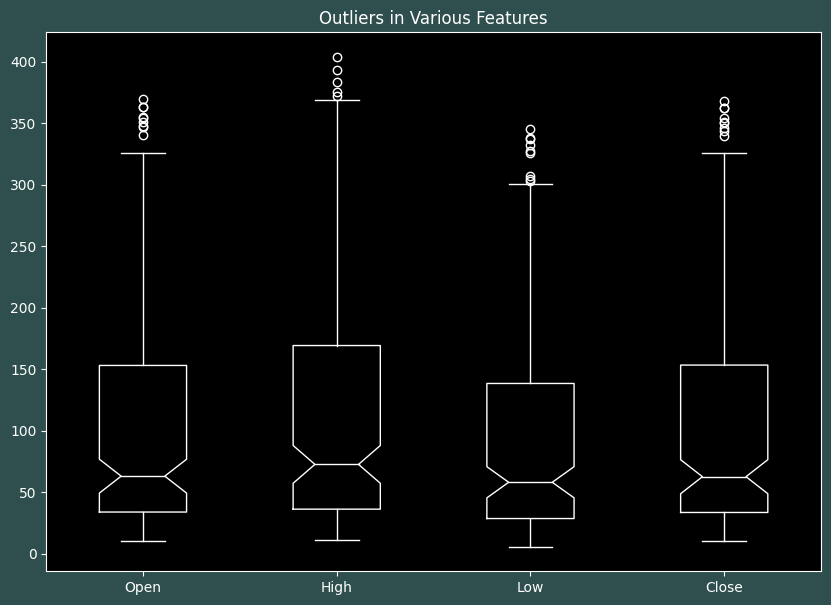
## 3.3 Distribution of dependent variable Close Price of stock

The distributions of open, high, low, and close in the chart are positively skewed. This indicates that the majority of data points are concentrated on the left side of the distributions, with a tail extending towards larger values on the right side. The histograms and KDE plots clearly show this skewness. Positive skewness suggests that the variables have a tendency for higher values, but with fewer occurrences. The presence of positive skewness may indicate bounded or restricted variables, resulting in an accumulation of values on the lower end and a tail of relatively larger values. Proper consideration of the positive skewness is important for accurate data analysis and modeling, potentially requiring transformations or alternative techniques to account for the skewness.



## 3.4 Boxplots: Studying the Outliers

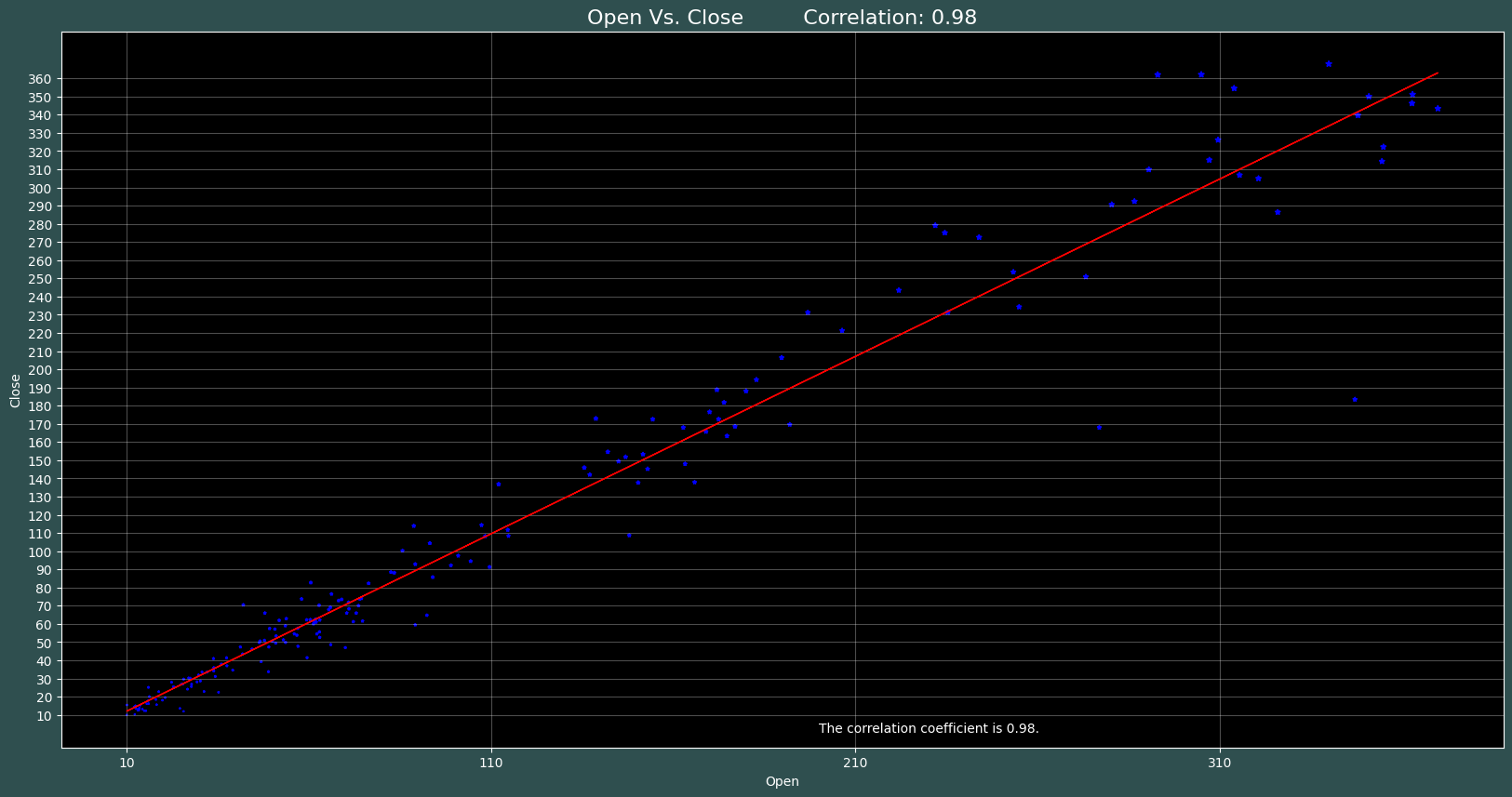
The presence of outliers in each of the features indicates the existence of extreme values that deviate significantly from the overall pattern of the data. These outliers can potentially impact the model fitting process and the accuracy of the predictions. Therefore, it is crucial to address these outliers before proceeding with model fitting.

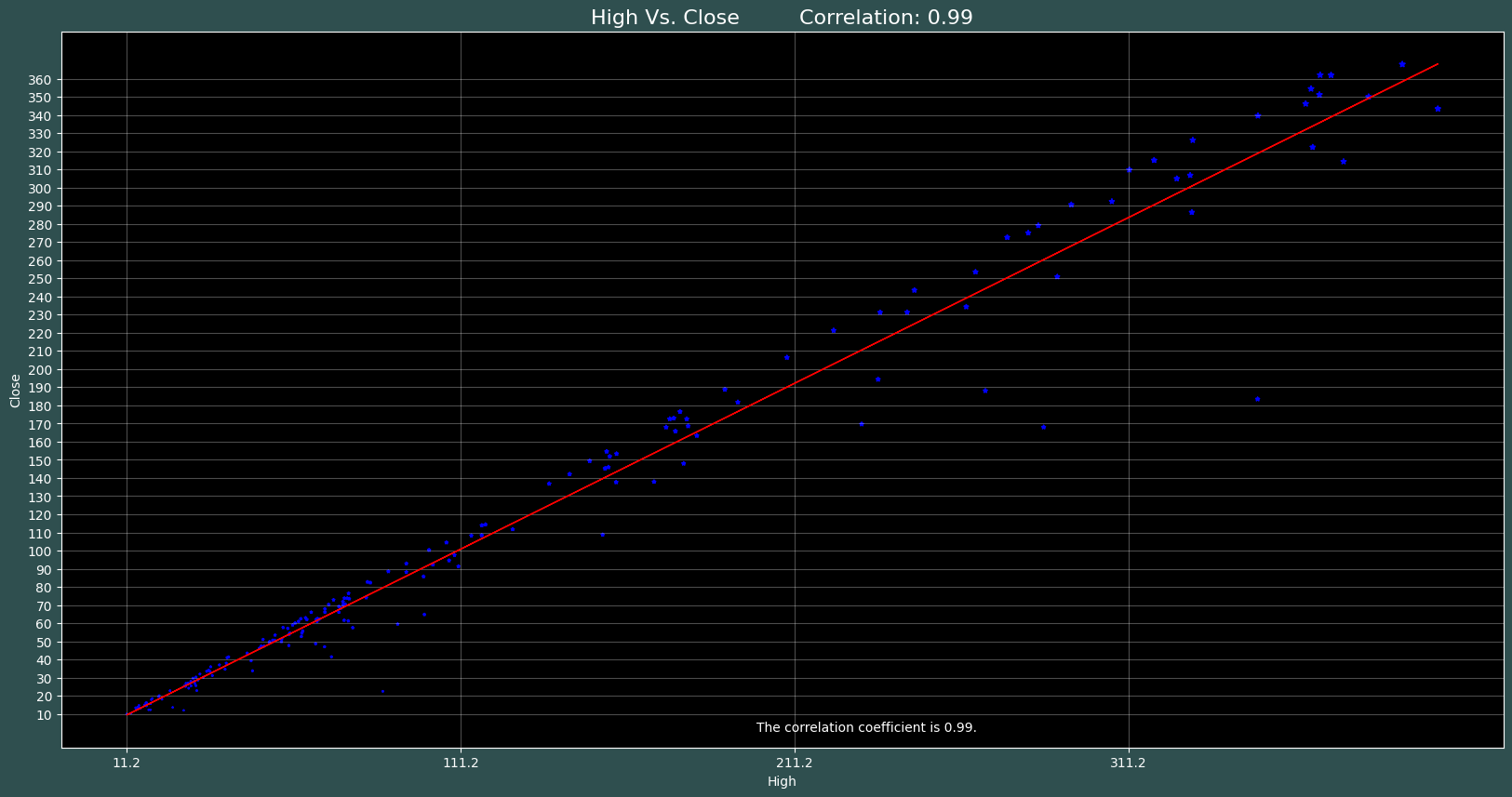


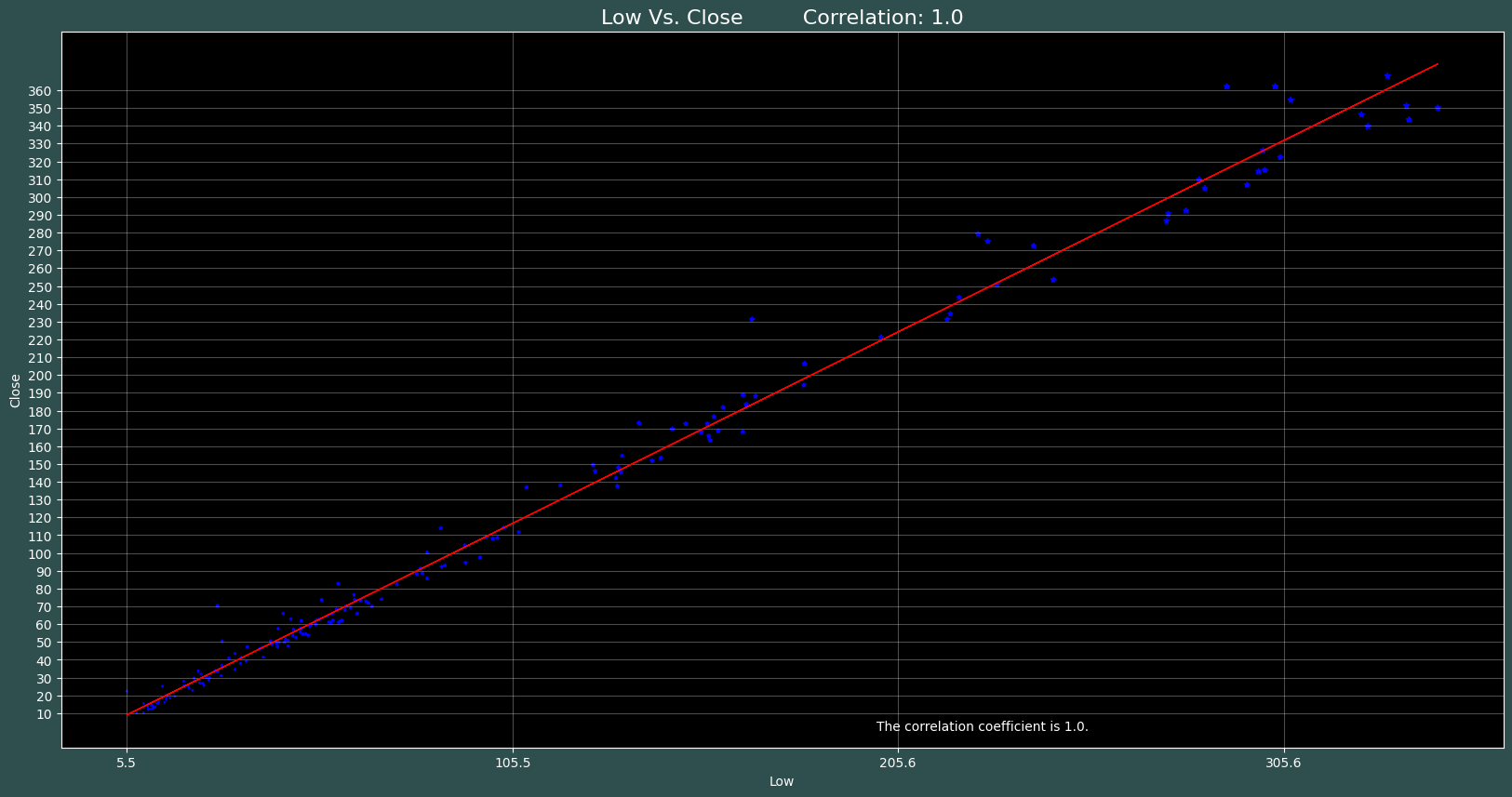
## 3.5 Scatter Plot to see the Best Fit line

Upon analyzing the scatter plots with the best fit line, it is evident that all the independent variables show a linear relationship with the dependent variable, 'Close'. This indicates that there is a consistent and predictable relationship between these variables.

The presence of a linear relationship has important implications in data analysis and modeling. It suggests that changes in the independent variables can be associated with proportional changes in the dependent variable. This knowledge can be leveraged to build regression models, make predictions, and understand the impact of the independent variables on the 'Close' price.

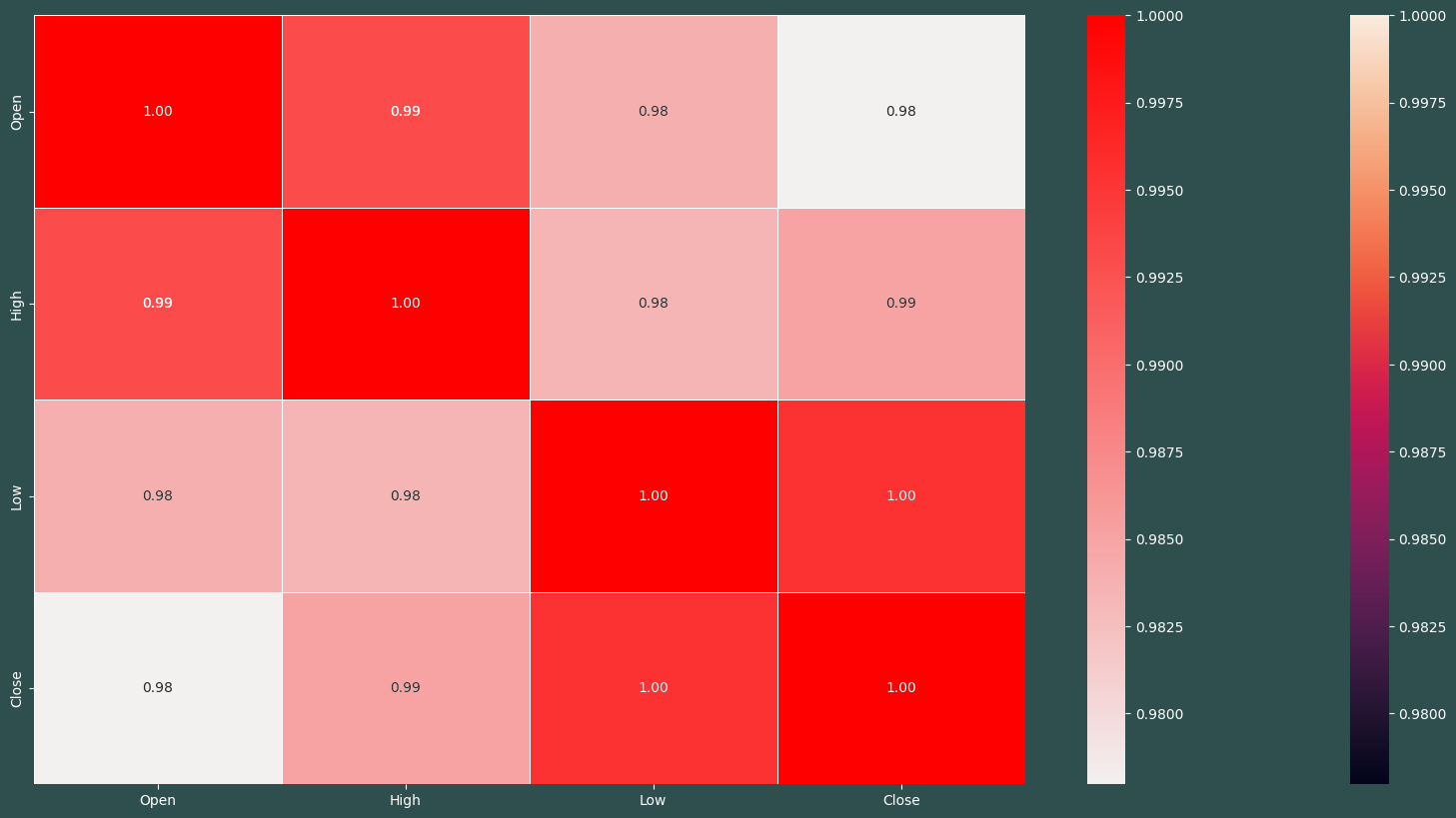




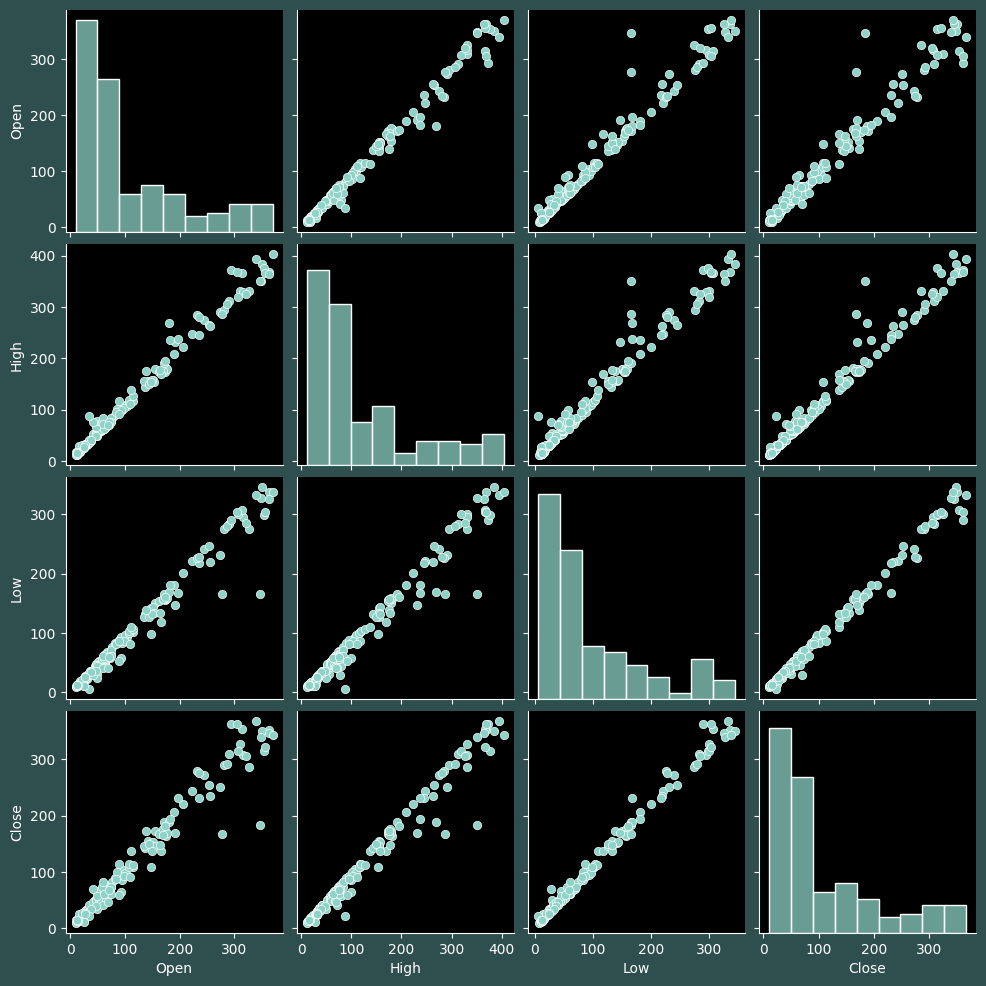


## 3.6 Checking Correlation between Independent variables and dependent variable

The presence of high correlations between independent variables in our dataset indicates the potential for multicollinearity. Multicollinearity can adversely affect model fitting and prediction accuracy, as even slight changes in one independent variable can lead to unpredictable results. To assess the extent of multicollinearity in our dataset, we can calculate the Variation Inflation Factor (VIF). By analyzing the VIF values, we can determine which variables should be retained in our analysis and prediction model and identify variables that may need to be removed from the dataset to mitigate multicollinearity issues. This evaluation helps ensure the robustness and reliability of our models and supports accurate predictions and interpretations of the relationships between variables.



## 3.7 Pair Plot



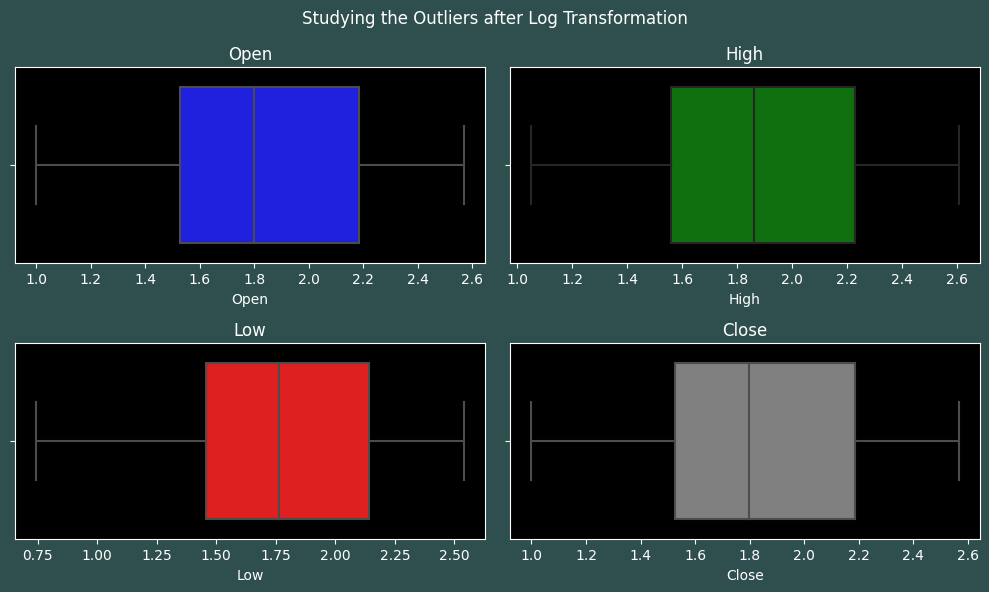
The variables Open, High, and Low show a strong correlation with the Close variable, indicating a close relationship between the stock's opening, highest, lowest, and closing prices. The Open, High, and Low variables also exhibit a high correlation with each other, suggesting they move in sync and share similar trends. These correlations provide valuable insights for analyzing the Yes Bank stock and can serve as predictors of the closing price. The relationships highlight the interdependencies within the stock market and the potential impact of external factors. Understanding these connections aids in making informed decisions and identifying patterns for forecasting future stock price movements. However, it's important to consider that correlation does not imply causation, and comprehensive analysis requires additional factors and considerations.

## 4. Feature Engineering & Data Pre-processing

## 4.1 Handling Missing Values

No missing values were found in the dataset, as confirmed earlier. Therefore, there is no requirement for missing values imputation techniques. The dataset is complete, allowing for direct analysis without the need to handle missing data.

## 4.2 Handling Outliers



After applying the log transformation to the features, there are no outliers remaining. The boxplots show no extreme values beyond the whiskers. The log transformation successfully reduced the impact of outliers and normalized the data. However, it is important to consider other factors and limitations in the analysis.

## 4.3 Categorical Encoding

Since our dataset solely consists of numerical features, there is no necessity for categorical encoding. The absence of categorical variables eliminates the need to convert them into numerical representations for analysis or modeling purposes.

## 4.4 Feature Manipulation & Selection

## 4.4.1 Feature Manipulation

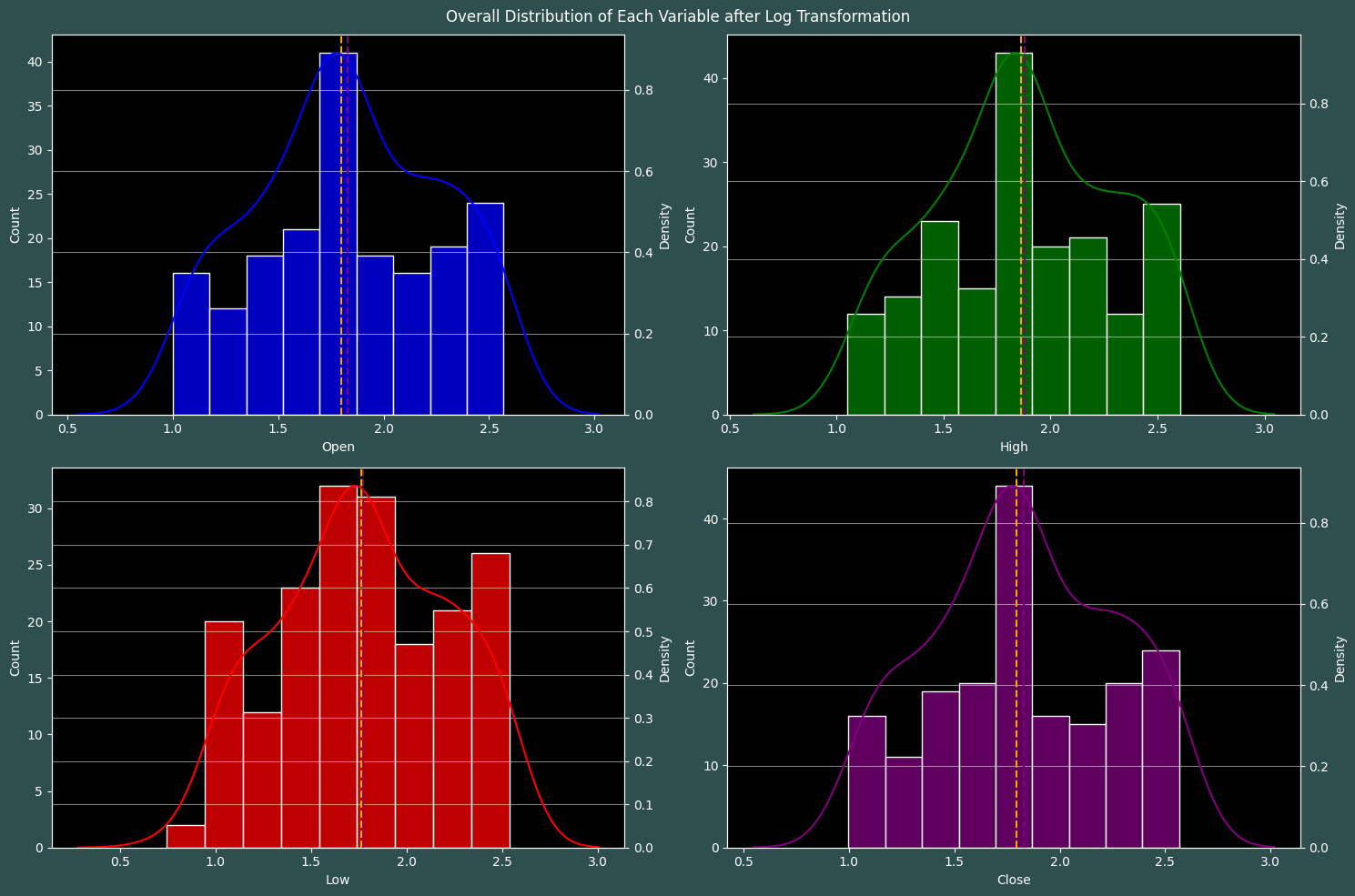
The VIF values for all the features indicate high multicollinearity. However, considering the small size of the dataset and having only three numerical independent variables, there is limited potential for feature manipulation that could be beneficial. With the absence of categorical variables, the scope for feature engineering or transformation is constrained. Therefore, the focus should be on alternative modeling approaches or additional data collection to address the issue of multicollinearity.

## 4.4.2 Feature Selection

Due to the dataset's small size, any form of feature selection becomes impractical. Given the limited number of observations, attempting to reduce the feature space may lead to unreliable or biased results. Therefore, it is advisable to retain all available features for analysis or modeling purposes.

## 4.5 Data Transformation

After the log transformation, the distributions of the features appear to be closer to a normal distribution compared to their previous state. The mean (indicated by the purple vertical line) and the median (represented by the yellow vertical line) are nearly equal for each feature. This alignment suggests that the log transformation successfully reduced the skewness and brought the data closer to symmetry. The convergence of the mean and median highlights the relative balance in the distribution, indicating a more representative central tendency. Overall, these observations indicate an improved approximation to a normal distribution after the log transformation.



## 4.6 Dimesionality Reduction

Since the dataset is already small in size, there is no need for dimensionality reduction techniques. With a limited number of observations, attempting to reduce the number of features may not provide significant benefits and could potentially lead to loss of valuable information. Therefore, it is advisable to retain all the available features for analysis or modeling purposes without applying dimensionality reduction methods.

## 4.7 Data Splitting

To train the model effectively, an 80:20 split ratio is being employed, allocating 80% of the data for training and 20% for testing. However, considering the small dataset size, it may be beneficial to acquire more data for training purposes. Increasing the training data size helps improve the model's ability to learn and generalize from the patterns present in the data. Gathering additional data can enhance the model's performance, reduce the risk of overfitting, and provide a more comprehensive representation of the underlying relationships within the dataset.

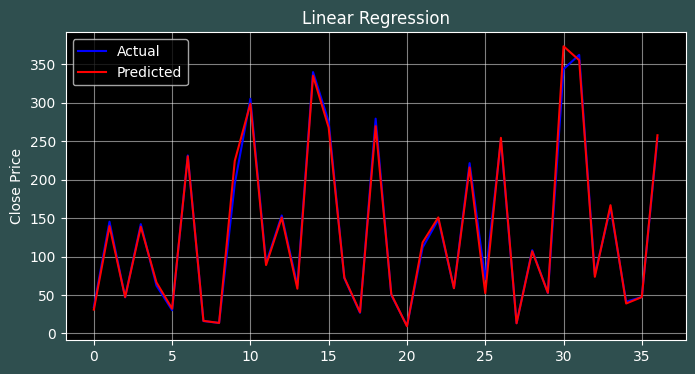
## 4.8 Data Scaling

The StandardScaler is utilized in this code snippet because we are primarily working with linear regression, which assumes normally distributed features. By applying the StandardScaler, we can standardize the features, transforming them to have a mean of 0 and a standard deviation of 1. This process aligns with the assumptions of linear regression and helps ensure that the features are on a similar scale, facilitating accurate model fitting and interpretation.

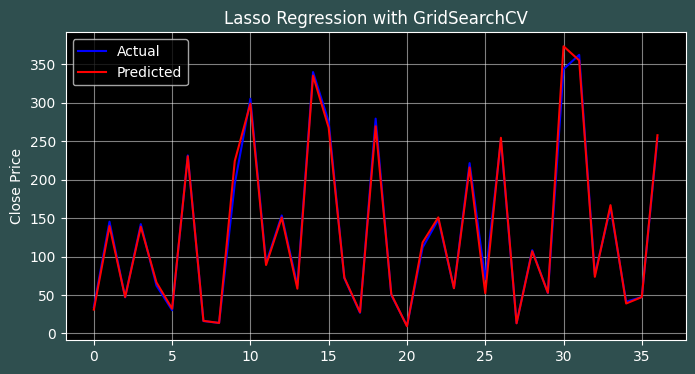
## 5. ML Model Implementation

## 5.1 Linear Regression

Linear Regression aims to establish a linear connection between the independent and dependent variables by minimizing the sum of squared differences between the observed and predicted dependent values. It assumes a linear relationship and calculates the best-fitting line by adjusting the model's coefficients. The objective is to minimize the overall distance between the observed data points and the line of best fit. This approach enables the model to capture the underlying linear pattern and make predictions based on the learned relationship between the variables.



## 5.2 Lasso Regression



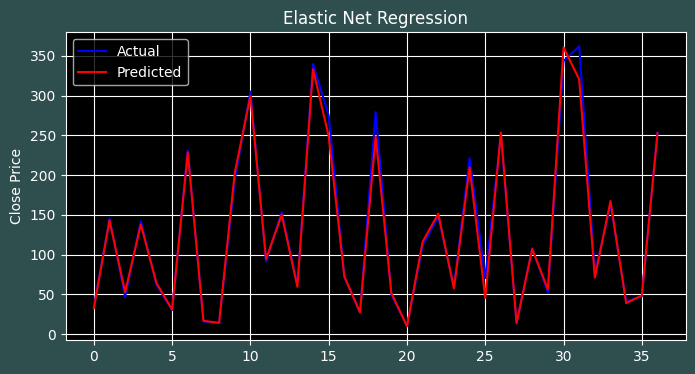
Lasso Regression with GridSearchCV is considered the winner due to its lower error metrics and slightly higher R-2 scores. The lower mean squared error, root mean squared error, and mean absolute error indicate improved accuracy and better predictive performance compared to Lasso Regression without GridSearchCV. Additionally, the slightly higher R-2 score suggests that Lasso Regression with GridSearchCV captures a greater amount of variance in the target variable and provides a better fit to the data. Overall, these evaluation metrics demonstrate that Lasso Regression with GridSearchCV outperforms Lasso Regression without GridSearchCV in terms of predictive accuracy and model fit.

## 5.3 Ridge Regression



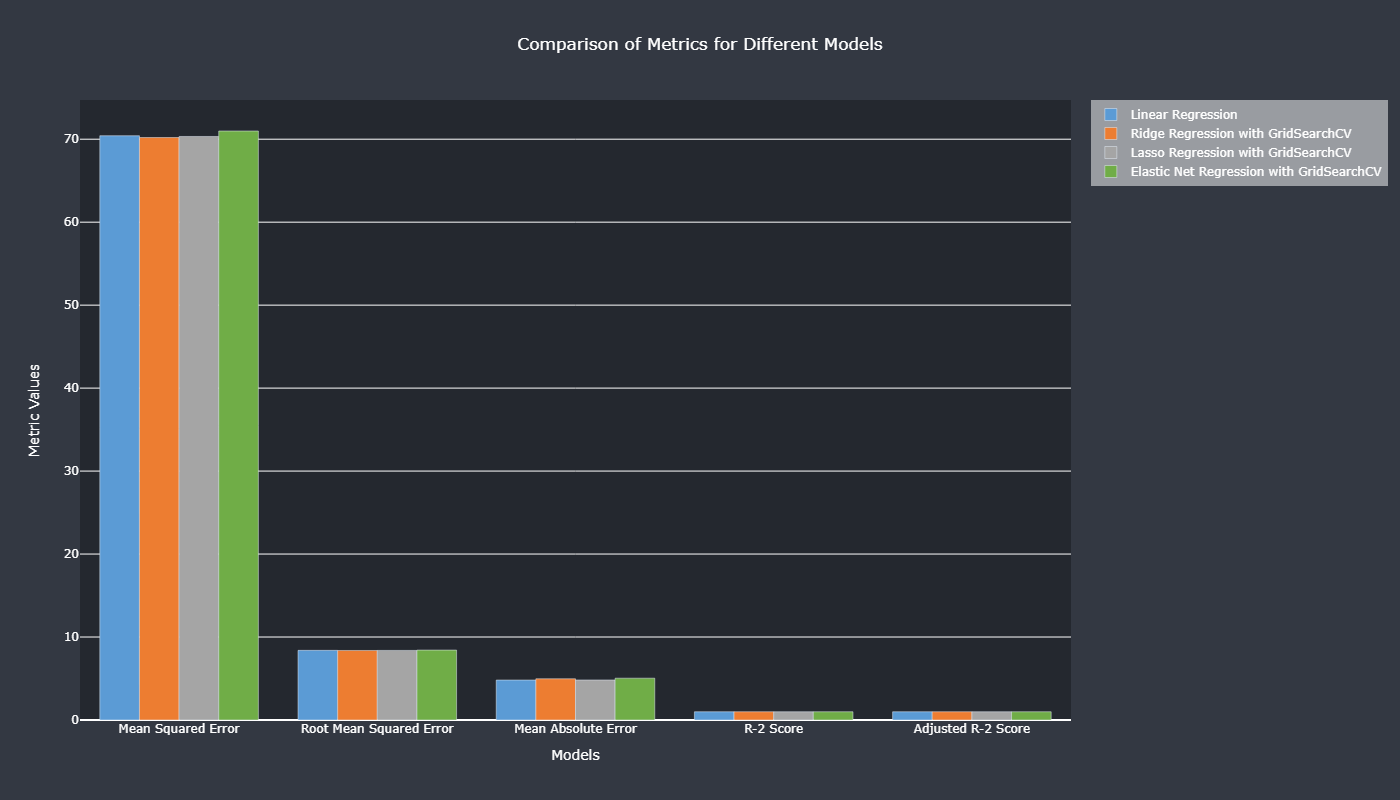
GridSearchCV was used in this code to find the best combination of hyperparameters for the Ridge regression model. It exhaustively searched for a smaller hyperparameter grid, specifically focusing on the alpha parameter representing regularization strength. This systematic approach allowed for improved model performance compared to other models. The Ridge Regression model with GridSearchCV achieved lower error values, demonstrating higher accuracy and predictive ability. This optimized model is considered more reliable for the given dataset, offering improved fitting and accurate predictions.

## 5.4 elastic net regression



In this implementation, GridSearchCV was employed to perform hyperparameter optimization for the Elastic Net Regression model. GridSearchCV is a technique used to systematically search through a predefined set of hyperparameters and select the optimal combination that yields the best performance. In this case, a smaller set of hyperparameters, including 'alpha' and 'l1\_ratio', was chosen to explore various regularization strengths and the balance between L1 (Lasso) and L2 (Ridge) regularization penalties. By exhaustively searching through this smaller parameter grid, GridSearchCV helps identify the hyperparameters that minimize the chosen evaluation metric, which in this case is the negative mean squared error. This approach allows for fine-tuning the Elastic Net Regression model and optimizing its performance based on the given hyperparameters.

## 5. Which ML model did you choose from the above created models as your final prediction model and why?



The 'Ridge Regression with GridSearchCV' model outperforms others, evident from its lower MSE/RMSE, indicating better predictive accuracy. Considering the small dataset size, Lasso Regression's feature selection may not be suitable, making the 'Ridge Regression with GridSearchCV' model preferred for retaining all features. Its balance between model complexity and accuracy aligns with the business objectives, ensuring accurate predictions and optimal decision-making. Thus, selecting the 'Ridge Regression with GridSearchCV' model instills confidence in achieving the desired business impact.

## 6. Hypothesis Testing

## 6.1 Hypothetical Statement - 1

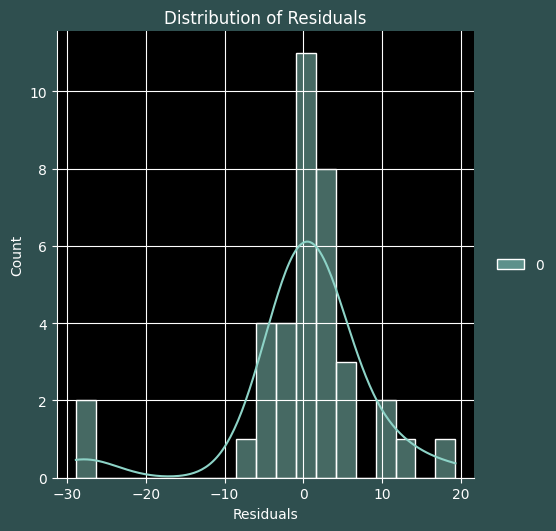
The Goldfeld-Quandt Test is used in regression analysis to check for heteroscedasticity, where variability of residuals differs across independent variables. It examines if the assumption of constant variability (homoscedasticity) holds true. By applying the test on regression residuals, we determine if there is statistical evidence of heteroscedasticity, gaining insights into the adequacy of the assumption and enhancing the reliability of our regression model.

## 6.2 Hypothetical Statement – 2

In linear regression, an essential assumption is no significant autocorrelation among residuals, indicating no systematic patterns or dependencies between residuals at different time points. We used the Ljung-Box test on residuals to validate this assumption. The test revealed a p-value above the significance level (e.g., 0.05), indicating no evidence to reject the null hypothesis of no autocorrelation. Therefore, we can conclude that our linear regression model's residuals exhibit no significant autocorrelation, validating the assumption of independence in predictions and ensuring the model's reliability and validity.

## 6.3 Hypothetical Statement – 3

The Shapiro-Wilk Test revealed that the residuals are not normally distributed, contradicting the assumption of normality. The obtained p-value was below the significance level (e.g., 0.05), leading to the rejection of the null hypothesis. The departure from normality in residuals suggests the presence of unaccounted factors or patterns in the data, necessitating consideration of alternative approaches or data transformation for accurate interpretation and inferences from the regression model.



## Conclusions:

* A careful examination of the data reveals a pronounced decline in the stock prices of Yes Bank following the exposure of the Rana Kapoor fraud in 2018.
* The dataset exhibited exceptional cleanliness, devoid of any missing values or duplicated rows, minimizing the need for extensive data wrangling.
* Although outliers were present in the features, effective outlier mitigation was achieved through the implementation of a log transformation across all features.
* The log transformation successfully addressed positive skewness observed in all features, ensuring adherence to the assumptions of the linear regression models.
* Strong positive correlations were observed between the independent variables (Open, High, Low) and the dependent variable (Close), implying a high predictive potential of the dependent variable based on the independent variables.
* The presence of positive correlations among the independent variables suggested the presence of multicollinearity; however, given the limited dataset size, feature removal was deemed unnecessary.
* Among the various implemented regression models, the Ridge Regression model, combined with GridSearchCV for hyperparameter optimization, emerged as the preferred choice. It achieved a commendable performance, boasting an RMSE of 8.3824 and an R-2 score of 0.9938.
* Notably, the 'High' and 'Low' features demonstrated positive weights, indicating a favorable impact on the predictions. Conversely, the 'Open' feature displayed a negative weight, signifying a detrimental influence on the predictions.
* Satisfactorily meeting the assumptions of homoscedasticity, absence of autocorrelation, and a mean of zero, the residuals bolstered the reliability of the regression model.
* The robustness of the conclusions was supported by a thorough exploration of the data, leaving little room for ambiguity.
* The observed decline in Yes Bank's stock prices following the Rana Kapoor fraud exposure underscored the substantial impact of such events on the financial market.
* The meticulous data cleaning process instilled confidence in the dataset's integrity, fostering accurate and reliable analyses.
* Employing an appropriate transformation technique mitigated the influence of outliers, ensuring a more accurate representation of the data.
* Addressing positive skewness through a log transformation enhanced the conformity of the data to the assumptions of linear regression models.
* The strong positive correlations between the independent and dependent variables bolstered the predictive power of the regression models.
* Careful consideration of multicollinearity, despite its presence, deemed feature removal unnecessary, given the limited dataset size.
* The selection of Ridge Regression with GridSearchCV as the final prediction model was substantiated by its exceptional performance, as demonstrated by the low RMSE and high R-2 score.

## References:

1. GeeksforGeeks
2. Analytics Vidhya
3. AlmaBetter Class material
4. Pandas and Numpy libraries
5. Stack overflow
6. YouTube
7. Researchgate.net
8. W3schools.com