CONESTOGA
Connect Life and Learning

CONESTOGA COLLEGE, DOON CAMPUS
PREDICTIVE ANALYTICS (1498)

STATISTICAL FORECASTING (STAT8041)

PROJECT REPORT-1: INDIVIDUAL PROJECT 1

DATE: 06/23/2024

SURAJ KUMAR MISHRA - 8902749

GUIDED BY: PROF. JONATHAN PLUMBTREE

1. Data

Time Series Data Set of Tesla Stock Price Analysis

The dataset chosen for the analysis is the daily closing prices of Tesla stocks. This data provides a comprehensive overview of the stock price trends and patterns over time. The source of this dataset is Kaggle, a well-known platform for data science and machine learning datasets. You can access the dataset through the following:

URL: https://www.kaggle.com/datasets/manavpatel571/tslacsv

This dataset includes historical stock prices from **2019-2022,** which are crucial for performing time series analysis and forecasting. The detailed data helps in applying various forecasting models to predict future stock prices. So that we can find out the insights.

Data Discussion:

The dataset was imported from a CSV file. Initial exploration of the dataset revealed that it contains the following columns:

- **Date:** The date of the stock price.
- **Open:** The opening price of the stock on a given day.
- **High:** The highest price of the stock on a given day.
- **Low:** The lowest price of the stock on a given day.
- **Close:** The closing price of the stock on a given day.
- **Volume:** The number of shares traded on a given day.
- **Dividends**: The dividends paid by the company on a given day
- **Stock.Splits:** Information on any stock splits that occurred on a given day.

Structure of the dataset before preprocessing–

data.frame': 758 obs. of 8 variables:
$ Date : chr "2019-05-21" "2019-05-22" "2019-05-23" "2019-05-24" .$ Open : num 39.6 39.8 38.9 40 38.2 ...
$ High : num 41.5 40.8 39.9 40 39 ...
$ Low : num 39.2 38.4 37.2 37.8 37.6 ...
$ Close : num 41 38.5 39.1 38.1 37.7 ...
$ Volume : int 90019500 93426000 132735500 70683000 51564500 59843000 39632500 520 33500 65322000 69037500 ...
$ Dividends : int 0 0 0 0 0 0 0 0 0 0 ...
$ Stock.Splits: num 0 0 0 0 0 0 0 0 0 0 ...

Data Quality Checks:

Before proceeding with the analysis, the dataset was checked for null values by using **sum(is.na(tesla))** and duplicates by using **sum(duplicated(tesla)).** It was confirmed that there are no missing values or duplicate entries in the dataset. This ensures the reliability and accuracy of the subsequent analysis.

Data Cleaning and Preprocessing:

Imported the dataset by using **read.csv ().** The '**Date**' column was converted to the Date data type to facilitate time series analysis. The dataset was then transformed into a **tsibble** object with '**Date**' as the index. To handle any irregularities and ensure the tsibble is regular, row numbers were added as the index. This step involves adding a new column **'rowid'** and updating the tsibble to use this column as the index. This is particularly useful for handling irregular time series data and ensures consistency in subsequent analysis.

'data.frame': 758 obs. of 8 variables:
$ Date : Date, format: "2019-05-21" "2019-05-22" "2019-05-23" "2019-05-24" ...
$ Open : num 39.6 39.8 38.9 40 38.2 ...
$ High : num 41.5 40.8 39.9 40 39 ...
$ Low : num 39.2 38.4 37.2 37.8 37.6 ...
$ Close : num 41 38.5 39.1 38.1 37.7 ...
$ Volume : int 90019500 93426000 132735500 70683000 51564500 59843000 39632500 52033500 65322000 69037500 ...
$ Dividends : int 0 0 0 0 0 0 0 0 0 0 ...
$ Stock.Splits: num 0 0 0 0 0 0 0 0 0 0 ...

For this analysis, we focused on the **'Date'** and **'Close'** columns as it is a **"Univariate Analysis"** to investigate the trend and forecast future stock prices.

Problem Statement:

The main objective of this analysis is to develop a reliable forecasting model for Tesla's stock prices. Accurate forecasts can aid investors and analysts in making informed decisions regarding stock trading. Specifically, the goal is to identify the underlying trends and seasonal patterns in Tesla's stock prices and use this information to predict future stock prices. By doing so, we aim to address the practical problem of maximizing returns and minimizing risks in stock investments through informed decision-making based on robust forecasting models.

2. Visualization

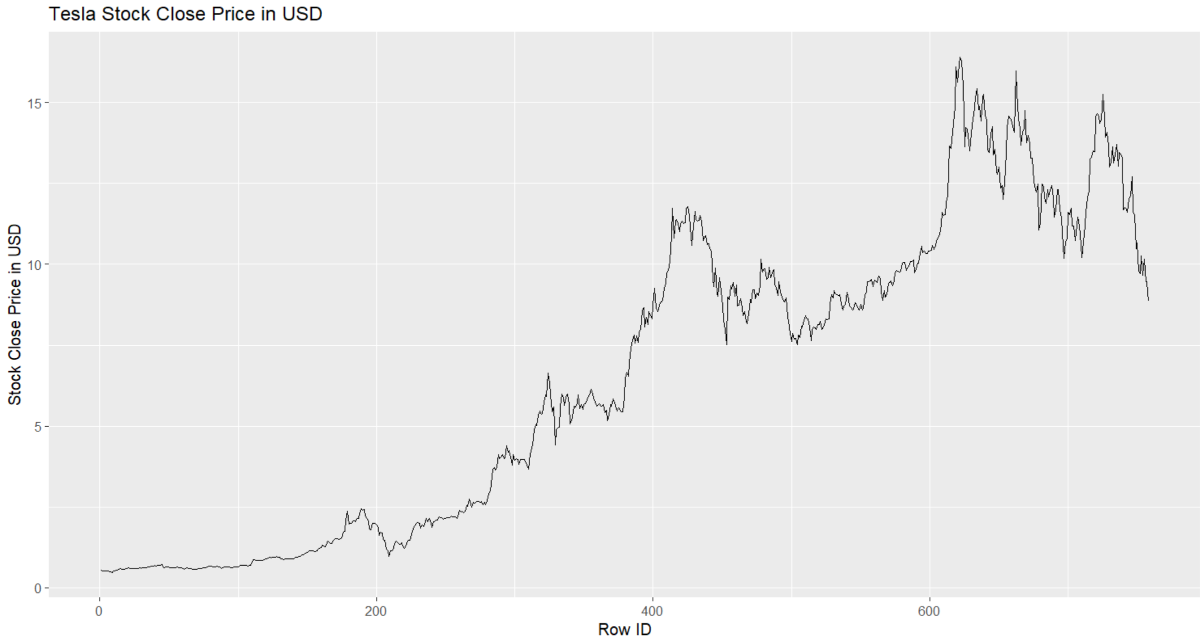
Time Plot:
A time plot of Tesla's stock closing prices over the given time period generated. The plot illustrates the trend and variability in Tesla's stock prices.

Initial Plot (Before Transformation)

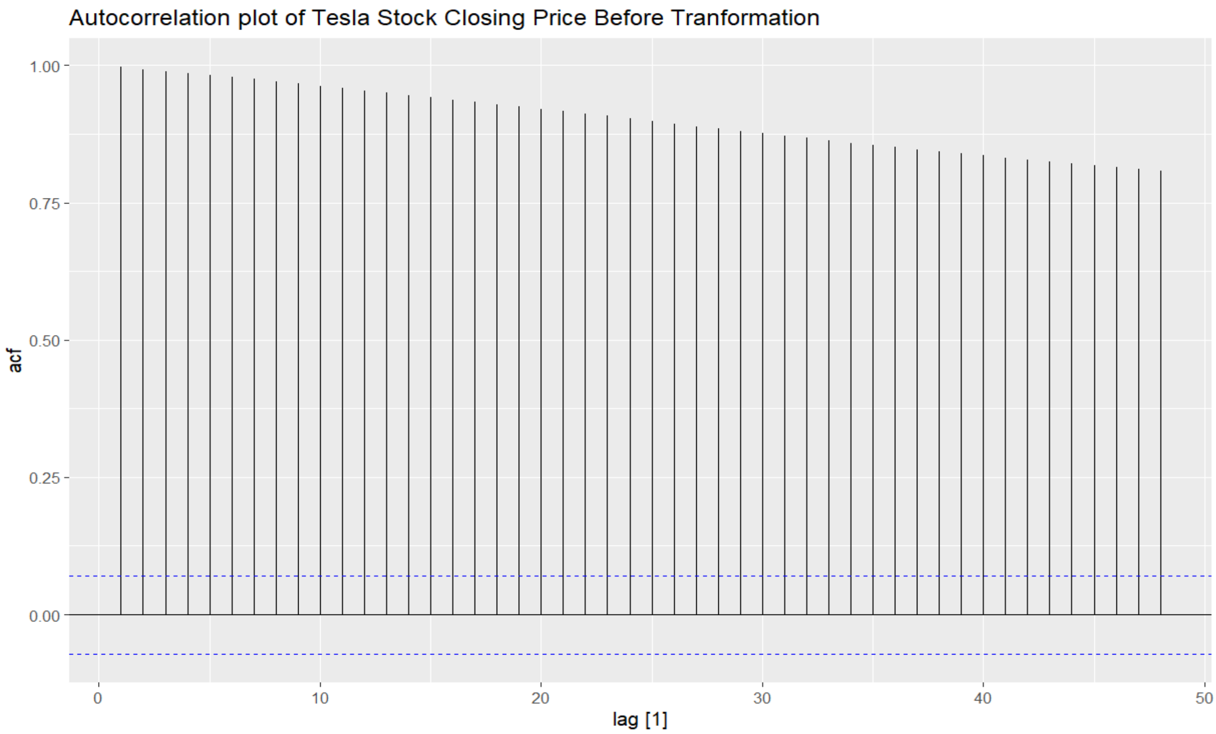
The initial plot represents the raw closing prices of Tesla stock in USD over time. Here are the key observations:

This plot depicts a four-year period of data **from 2019 to 2022,** with a general upward trend from left to right. While the overall movement suggests an increase over time, it's important to note the significant fluctuations throughout as it is steady at the beginning. There are periods with higher values (peaks) and periods with lower values (dips).

1. **Variability and Non-Normality:** The original data exhibits significant variability and non-normality. The prices range from approximately **$0 to over $15**, with noticeable peaks and troughs.
2. **Trend and Volatility:** The plot shows a general upward trend with periods of rapid growth, followed by declines. There are significant fluctuations in the data, especially in the later periods, indicating high volatility



Autocorrelation Plot Interpretation:



The Autocorrelation Function (ACF) plot for Tesla's stock closing prices reveals essential information about the time dependencies within the data. On the Y-axis, the autocorrelation coefficient ranges from -1 to 1, **with values closer to 1 indicating a strong positive correlation**. The X-axis represents the number of time steps (lags), where a lag of 1 indicates the correlation between today's and yesterday's prices. The bars in the plot are initially high at lower lags and gradually decrease, indicating strong correlations between recent prices. The blue dashed lines mark statistical significance, and bars above these lines signify significant autocorrelations.

Interpreting the ACF plot, we observe a high initial autocorrelation, showing that today's price is highly predictive of tomorrow's price. The gradual decline in autocorrelation suggests that while the influence of past prices diminishes over time, it remains significant even at higher lags. This pattern reveals strong temporal dependencies in Tesla's stock prices, emphasizing the value of historical data for forecasting. The plot indicates that past stock prices have a lasting influence, making historical data a reliable basis for predictive modelling.

3. Transformations

Box- Cox Transformations –

The Box-Cox transformation is a statistical technique used to stabilize variance and make the data more normally distributed. This transformation is particularly useful when dealing with time series data that exhibit heteroscedasticity (i.e., when the variance of the residuals is not constant) or non-normality.

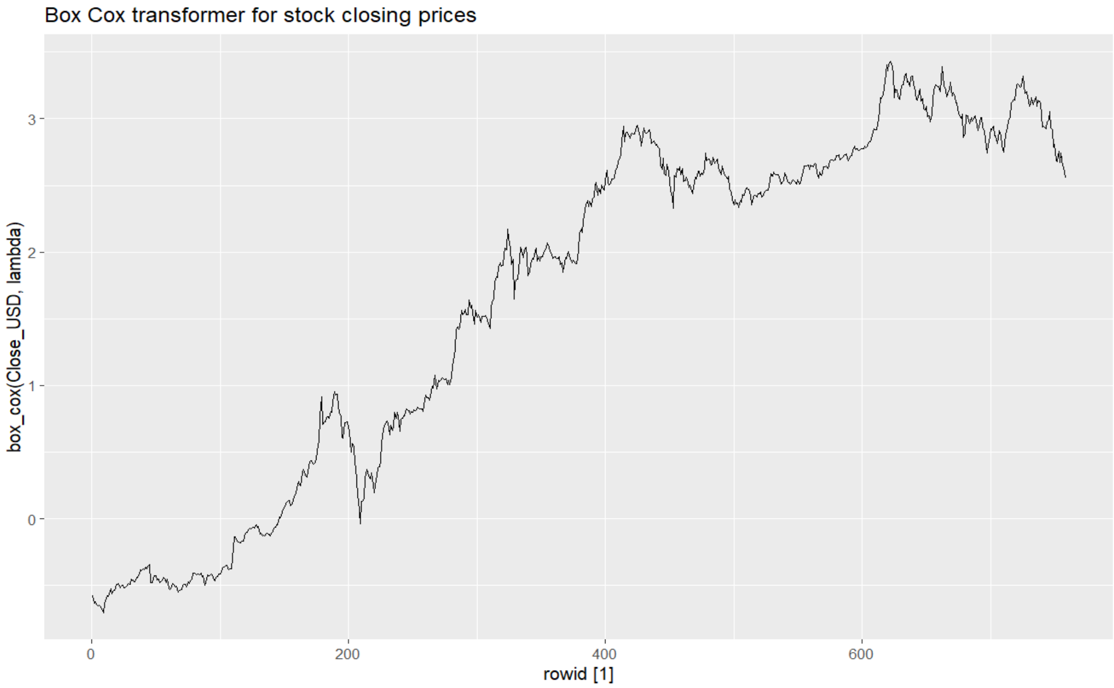
Why Do We Use Box-Cox Transformation?

1. **Stabilize Variance:** To reduce the impact of heteroscedasticity.
2. **Normalize Data:** To make the data more closely follow a normal distribution.
3. **Improve Model Accuracy:** To enhance the performance of statistical models that assume normality and constant variance.

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \ln(y) & \lambda = 0 \end{cases}$$

- y is the original data.
- λ lambda is the transformation parameter.

> lambda
[1] 0.1406559 (Check the source code)



The transformed plot represents the closing prices after applying the Box-Cox transformation. Here's a detailed analysis:

1. **Stabilized Variance:**

- The transformed values start near 0 and increase to over 3. The transformation effectively stabilizes the variance, making the data smoother and less volatile.

2. **Normalized Data:**

- The transformed data approximates a normal distribution better than the original data. This normalization is beneficial for statistical modelling.

3. **Trend Clarity:**

- The upward trend is clearly visible with reduced fluctuations. The transformation reveals a smoother and more linear growth pattern.

<u>Detailed Phase Comparison</u>

1. **Initial Phase (Row ID 0 to 100):**

   Before Transformation: The prices are relatively low, with gradual increases and some small fluctuations.

After Transformation: The values start near 0 and show a gradual increase. The small fluctuations are smoothed out, making the initial trend clearer.

2. **Middle Phase (Row ID ~100 to ~400):**

   Before Transformation: This period shows rapid growth with high volatility and significant peaks and troughs.

   After Transformation: The values exhibit a steep, smooth increase, indicating a period of rapid growth with stabilized variance. The transformation reduces volatility, highlighting the significant upward trend.

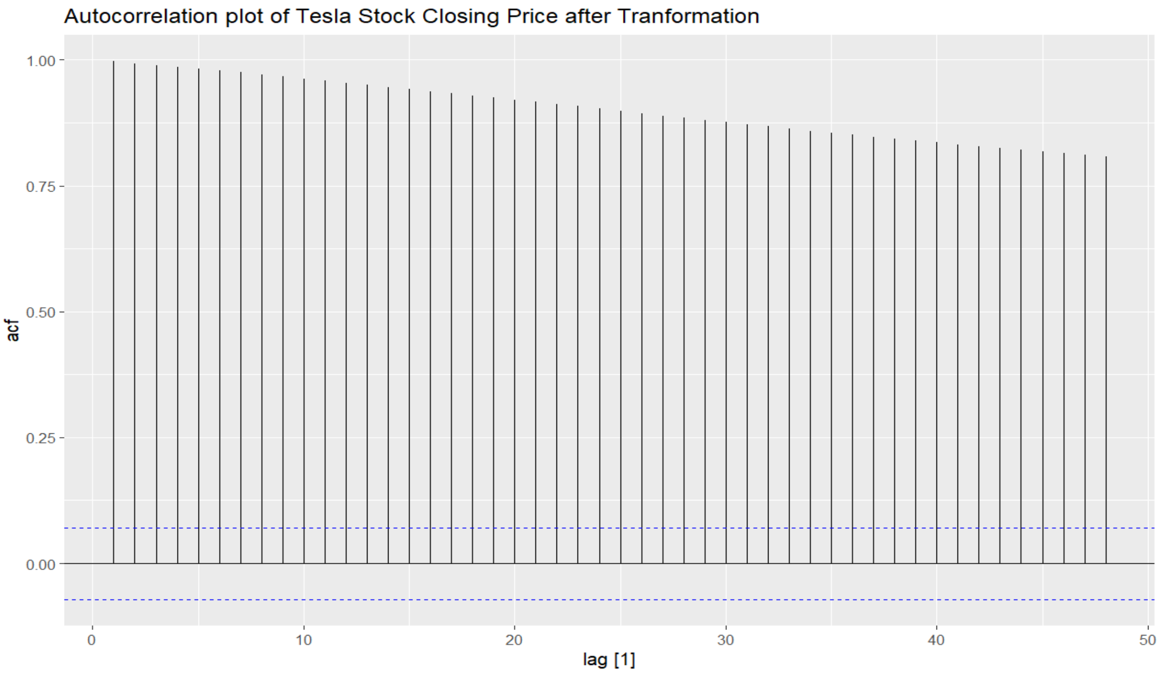
3. **Later Phase (Row ID ~400 to end):**

   Before Transformation: The prices peak above $15, followed by a decline with high variability.

   After Transformation: The values continue to rise, peaking above 3, and then show a slight decline. The transformation normalizes these fluctuations, reducing the impact of erratic changes and making the overall trend more discernible.

ACF PLOT AFTER BOX-COX TRANSFORMATION –

In the ACF plot there is slightly change as the it is more gradually decreasing. However, its not that significant though.

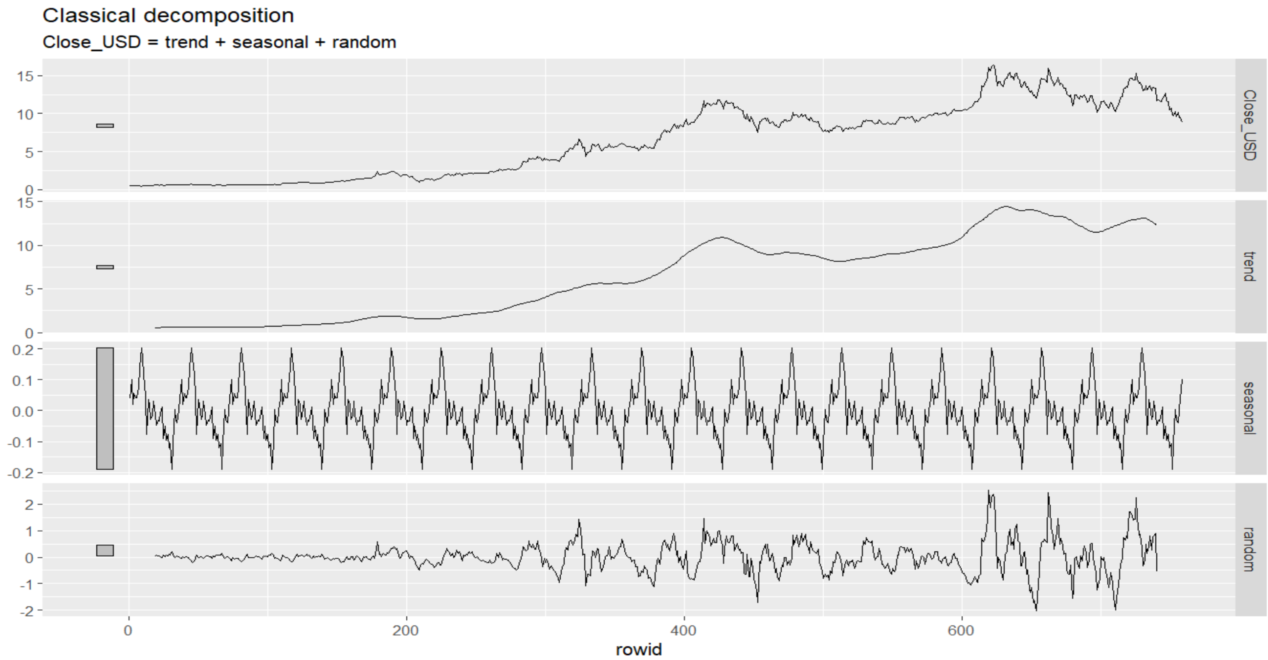


Decomposition:

1. Classical Decomposition – Classical decomposition is a method used in time series analysis to break down a time series into three distinct components: trend, seasonal, and random (or residual). This technique helps to understand and analyze the underlying patterns and behaviors in the data by separating it into these components.

- **Trend Component:** This represents the long-term progression of the series. It shows the overall direction, whether increasing or decreasing, without the influence of seasonal or irregular variations.
- **Seasonal Component:** This captures the regular, repeating patterns or cycles in the data, which occur at consistent intervals (e.g., monthly, quarterly, yearly).

- **Random (Residual) Component:** This is the remainder after removing the trend and seasonal components. It includes irregular or random fluctuations that cannot be attributed to the other two components.



Interpretation of the Classical Decomposition Plot -

1. Original Data (Close_USD)

The top plot represents the actual closing price of the asset in USD over time. The observed range of the closing price is from approximately $0 to $15. The data shows significant fluctuations, indicating periods of growth and decline throughout the period analysed.

2. Trend Component

The second plot illustrates the long-term trend of the closing price. **The trend starts near $0 and shows a gradual increase, peaking at around $8 before a slight decline to about $7.5**

towards the end of the period. This indicates a steady long-term growth of approximately $8, followed by a minor decrease of $0.5.

3. <u>Seasonal Component</u>

The third plot captures the seasonal variations, which are regular and cyclical fluctuations in the closing price. These seasonal changes oscillate between -$0.2 and +$0.2, indicating a consistent cyclical pattern. **The total range of these predictable cyclic variations is about $0.4**, **reflecting periodic increases and decreases around the trend**.

4. <u>Random Component</u>

The bottom plot shows the random or residual component, which represents the irregular fluctuations that remain after removing the trend and seasonal effects. **This component varies between -$2 and +$1.5, highlighting periods of higher volatility and unexpected deviations**. These random fluctuations indicate that unpredictable events can cause the closing price to vary significantly within this range.

<u>Classical Decomposition Components of Tesla's Stock Closing Prices -</u>

This table shows the observed closing prices and their corresponding seasonal adjustments,

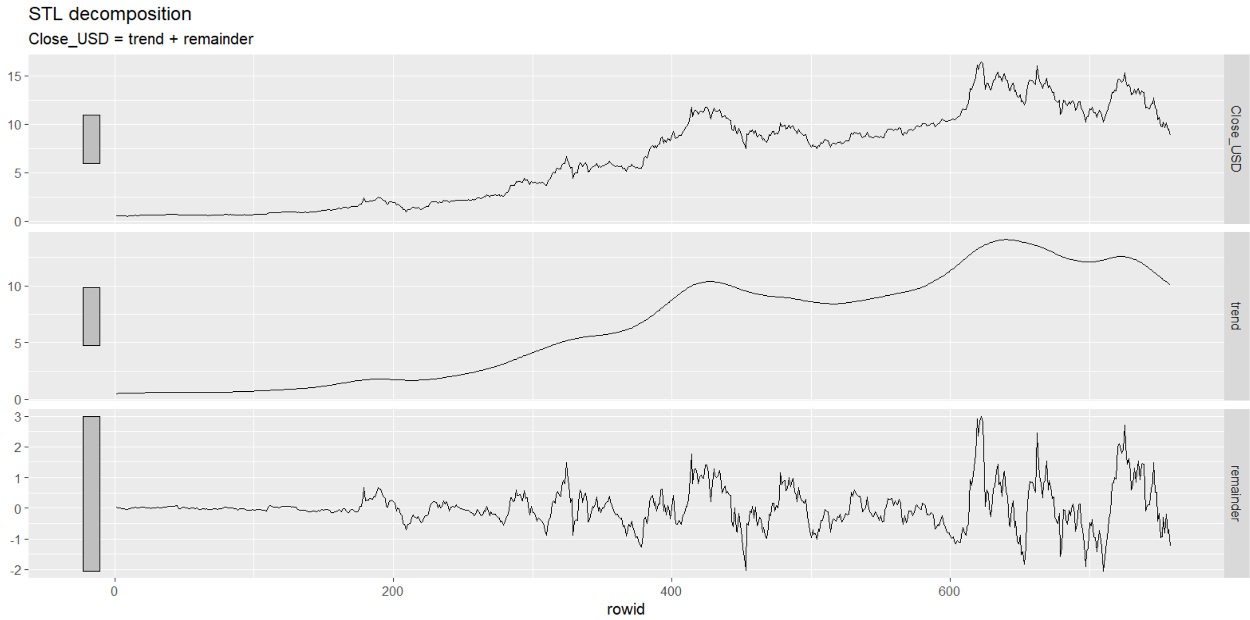
.model	rowid	Close_USD	trend	seasonal	random	season_adjust
classical	753	10.261200	NA	0.0041048	NA	10.257096
classical	754	9.658267	NA	-0.0279817	NA	9.686248
classical	755	10.154800	NA	-0.0390840	NA	10.193884
classical	756	9.464133	NA	0.0022608	NA	9.461873
classical	757	9.458933	NA	0.0389593	NA	9.419974
classical	758	8.852000	NA	0.1007109	NA	8.751289

2. Seasonal Decomposition – STL (Seasonal and Trend decomposition using Loess) is a method used to decompose a time series into three components:

1. **Trend**: Long-term movement or direction in the data.
2. **Seasonal**: Repeating short-term cycle within the data.
3. **Remainder**: Irregular fluctuations not captured by the trend or seasonal components.

- **Trend Analysis**: Understands the long-term direction of the data.
- **Seasonal Patterns**: Identifies recurring patterns.
- **Anomaly Detection**: Isolates irregular variations.
- **Forecasting**: Enhances accuracy by clarifying data components.

Interpretation of the STL Plot –



The plot shows the STL decomposition of the closing price of USD (Close_USD).

Top Panel: Original Time Series (Close_USD)

The initial value is near 0 and gradually increases. It peaks around rowid 400 at approximately 15 USD, then declines but remains above 10 USD.

Middle Panel: Trend

The trend starts near 0 USD and rises gradually. It peaks at around 10 USD around rowid 550, showing a slight decline but staying above 5 USD towards the end.

Bottom Panel: Remainder

The remainder fluctuates around 0 USD, indicating irregular variations. There are noticeable spikes around rowid 400 and 550, with values between 3 USD and -2 USD.

Summary

The trend analysis shows an upward movement, peaking at around 10 USD around rowid 550, indicating long-term growth before a slight decline. The irregular variations exhibit high volatility around rowid 400 and 550, likely due to external factors affecting the price.

STL Decomposition Components of Tesla's Stock Closing Prices –

.model	rowid	Close_USD	trend	remainder	season_adjust
stl	753	10.261200	10.55173	-0.2905301	10.261200
stl	754	9.658267	10.45764	-0.7993729	9.658267
stl	755	10.154800	10.36355	-0.2087488	10.154800
stl	756	9.464133	10.26946	-0.8053244	9.464133
stl	757	9.458933	10.17537	-0.7164337	9.458933
stl	758	8.852000	10.08128	-1.2292755	8.852000

4. Forecasting and Analysis

Train and Test

Filter data for training and testing sets - The data is divided by using **filter()** into training and testing sets based on the year. The training set consists of data from the **year 2021**, while the testing set includes data from the **year 2022**. This division helps in training the models on historical data and validating their performance on future data.

Fit the models -Various forecasting models are fitted to the training data to predict the closing price of Tesla stock. The models used include:

1. Mean Model: Predicts future values based on the average of past values, assuming that the best estimate of future prices is the mean of the historical data.

2. Naive Model: Assumes that the most recent observed value is the best predictor of the future value. For example, if the last closing price is 700 USD, the model will predict 700 USD for all future points.

3. Drift Model: A variant of the Naive model that incorporates a drift component, accounting for a constant change over time. It projects future values based on the average change in the data. For instance, if the price increases by an average of 5 USD per day, this model will add 5 USD to the last observed value for each subsequent day.

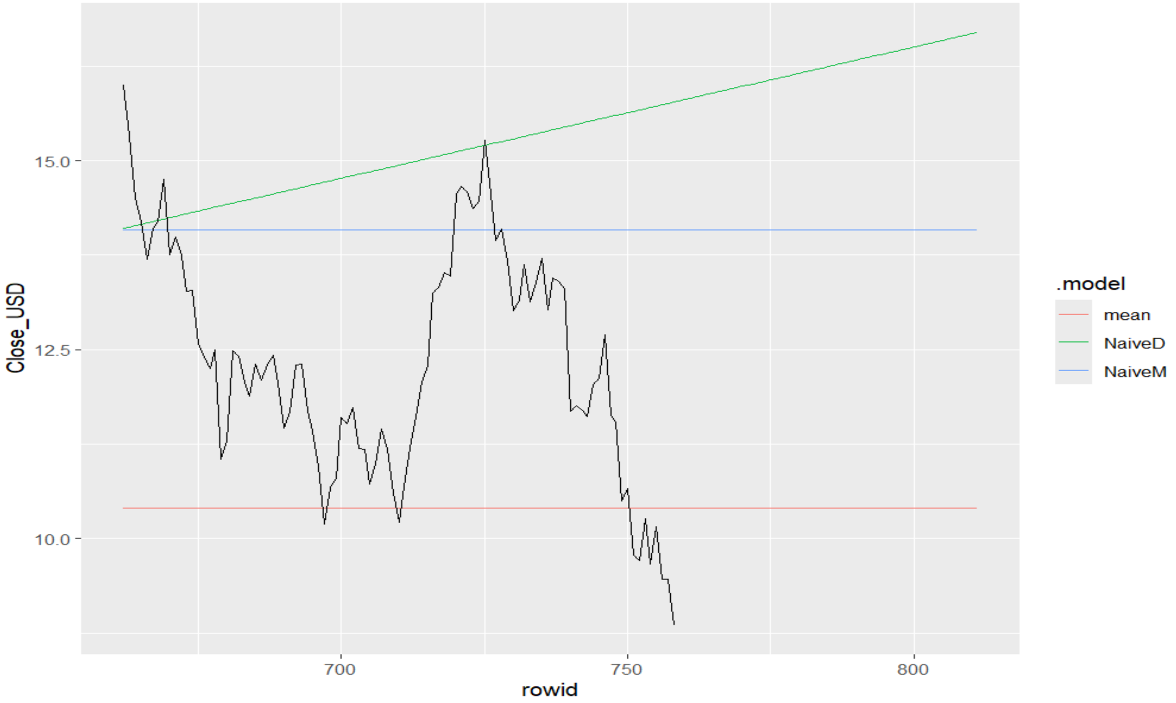
Forecast Using the Fitted Models

After fitting the forecasting models to the training data, the next step is to generate forecasts for a specified future period. In this case, we generate forecasts for the next 150 days. This involves using the fitted models to predict the future closing prices of Tesla stock.

- **Generate Forecasts**: Using the fitted models (Mean, Naive, and Drift models), we make forecasts for the next **150 days**. This involves projecting the closing prices of Tesla stock into the future based on the patterns and trends identified in the training data.
- **Forecast Horizon**: The forecast horizon (h) is set to 150, indicating that the models will predict the closing prices for the next 150 days.

By generating these forecasts, we obtain a set of predicted values for the closing price of Tesla stock. These forecasts can then be compared against the actual values in the testing set (2022 data) to assess the performance and accuracy of the model

Forecast Plot –



The plot provides a comparison of actual and forecasted closing prices of Tesla stock. It includes the following elements:

1. **Actual Data (Black Line)**: The black line represents the actual closing prices of Tesla stock in the test dataset, measured in USD. This line shows the observed values over the given time period.
2. **Forecasted Data**:

- **Mean Model (Red Line)**: The red line shows the forecasted values using the mean model. This model predicts the future values as the average of the historical closing prices. In this case, the mean closing price is approximately USD 11.50. The red line is a horizontal line at USD 11.50, **indicating that the model predicts a constant future price of USD 11.50.**
- **Naive Model (Blue Line)**: The blue line shows the forecasted values using the naive model. This model predicts the future values to be the same as the last observed value in the training dataset. The last observed value in the training set is approximately USD 12.50. The blue line is a horizontal line at USD 12.50, indicating that the model predicts a constant future price of USD 12.50.
- **Naive Model with Drift (Green Line)**: The green line shows the forecasted values using the naive model with drift. This model assumes a linear trend based on the historical rate of change. The forecast starts from the last observed value (USD 12.50) and increases linearly, reflecting an upward drift. The line starts at USD 12.50 and gradually rises, indicating a predicted increase over time.

Interpretation of the Plot

- **Mean Model**: The mean model forecasts a constant price of USD 11.50. This model does not account for any trends or patterns in the data, leading to a flat prediction line.
- **Naive Model**: The naive model predicts a constant price of USD 12.50, based on the last observed value in the training dataset. Similar to the mean model, it does not consider trends or patterns.
- **Naive Model with Drift**: The naive model with drift predicts an upward trend starting from USD 12.50. This model captures the historical rate of change and assumes it will continue linearly.

Accuracy of Forecasting Models

The accuracy of the forecasting models is evaluated using two key metrics: RMSE and MAE.

- **Root Mean Squared Error (RMSE)**: RMSE is a measure of the differences between predicted and actual values. It is the square root of the average of the squared differences between forecasted and actual values. RMSE provides a measure of how spread out these differences are, with higher values indicating larger errors.
- **Mean Absolute Error (MAE)**: MAE is the average of the absolute differences between predicted and actual values. It measures the average magnitude of the errors without considering their direction (i.e., whether they are positive or negative). Lower MAE values indicate better model accuracy.

The table below shows the RMSE and MAE for each forecasting model:

	.model	RMSE	MAE
	<chr>	<dbl>	<dbl>
1	NaiveD	3.13	2.69
2	NaiveM	2.31	1.92
3	mean	2.46	2.07

Interpretation of Accuracy Metrics

- **Naive Model with Drift (NaiveD)**:
  RMSE: 3.13
  MAE: 2.69
  This model has the highest RMSE and MAE values, indicating that its predictions are the least accurate among the three models. The upward drift predicted by this model does not match the actual downward trend in the data.

- **Naive Model (NaiveM)**:
  RMSE: 2.31
  MAE: 1.92
  This model has the lowest RMSE and MAE values, indicating that its predictions are the most accurate among the three models. Despite predicting a constant price, this model's forecast is closer to the actual values compared to the other models.

- **Mean Model**:
  RMSE: 2.46
  MAE: 2.07

This model's RMSE and MAE values are slightly higher than those of the naive model, indicating that its predictions are less accurate but still better than the naive model with drift.

The evaluation metrics (RMSE and MAE) indicate that the naive model (NaiveM) provides the most accurate forecasts for Tesla's closing prices in this case, despite its simplicity. The mean model performs moderately well, while the naive model with drift (NaiveD) performs the worst due to its incorrect assumption of an upward trend.

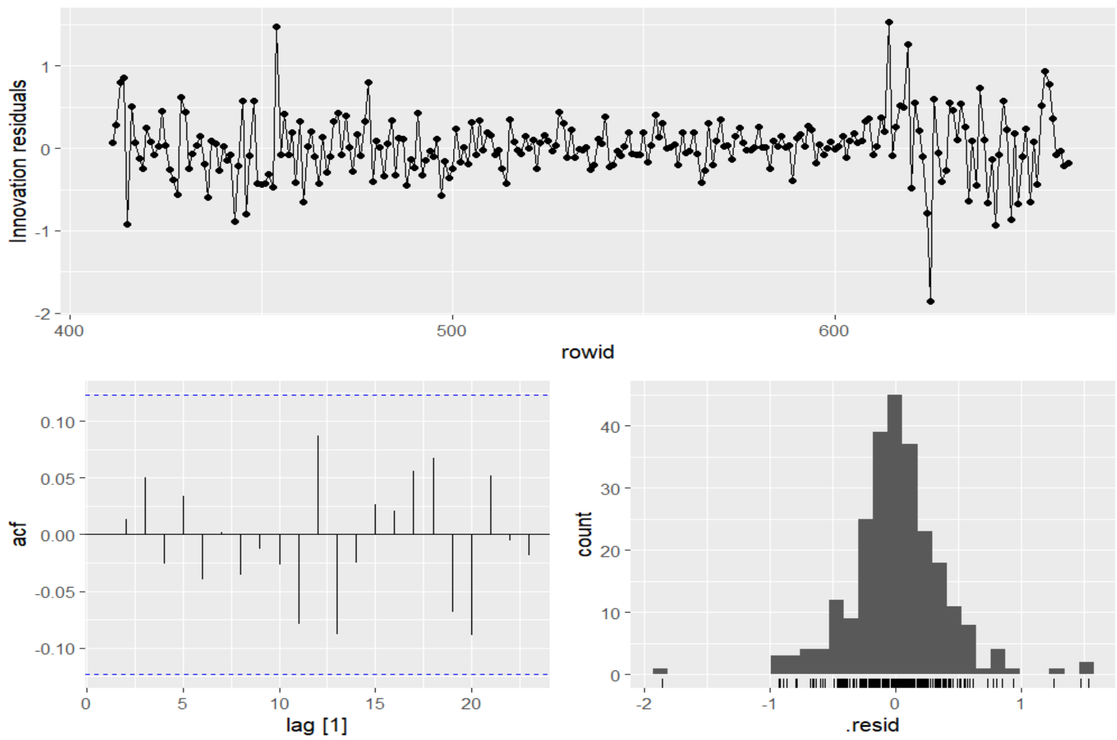
Residual Analysis of the Naive Model

Time Series Plot of Residuals

In the top plot, we present the time series of residuals, which are the differences between the actual values and the predicted values from the naive model. The residuals are cantered around zero, indicating that our model's predictions are unbiased on average. We observe spikes, such as those around rowid 460 and 620, where the residuals reach approximately 1.5 and -1.5, respectively, indicating significant deviations from actual values at these points. Despite these spikes, the residuals do not exhibit a clear pattern over time, suggesting that our model's errors are randomly distributed. However, some clustering of residuals, particularly between rowid 500 and 600, hints at potential autocorrelation that we should consider addressing.

Autocorrelation Function (ACF) Plot of Residuals

The bottom left plot displays the Autocorrelation Function (ACF) of the residuals, which helps us understand if the residuals are correlated with their past values (lags). Ideally, residuals should not show significant autocorrelation. Most of the autocorrelation values in our plot fall within the blue dashed lines, representing the 95% confidence intervals, indicating no significant autocorrelation at most lags. However, there are notable spikes at lag 5 and lag 10, where the autocorrelation values reach approximately 0.10 and 0.08, respectively. These findings suggest potential autocorrelation at these specific lags, indicating that while our model performs well, it could be improved by addressing these dependencies.



Histogram of Residuals

The bottom right plot is a histogram of the residuals, which shows their distribution. For an ideal model, the residuals should follow a normal distribution cantered around zero. Our histogram reveals that the residuals are roughly normally distributed with a mean close to zero. Most of residuals fall between -1 and 1, with the highest frequency around zero, indicating most errors are small. There are a few outliers on both ends of the distribution, such as those at -2 and 2, which indicate extreme prediction errors. Despite these outliers, the concentration of residuals around zero is a positive sign of model performance.

Overall Evaluation of the Naive Model Residuals

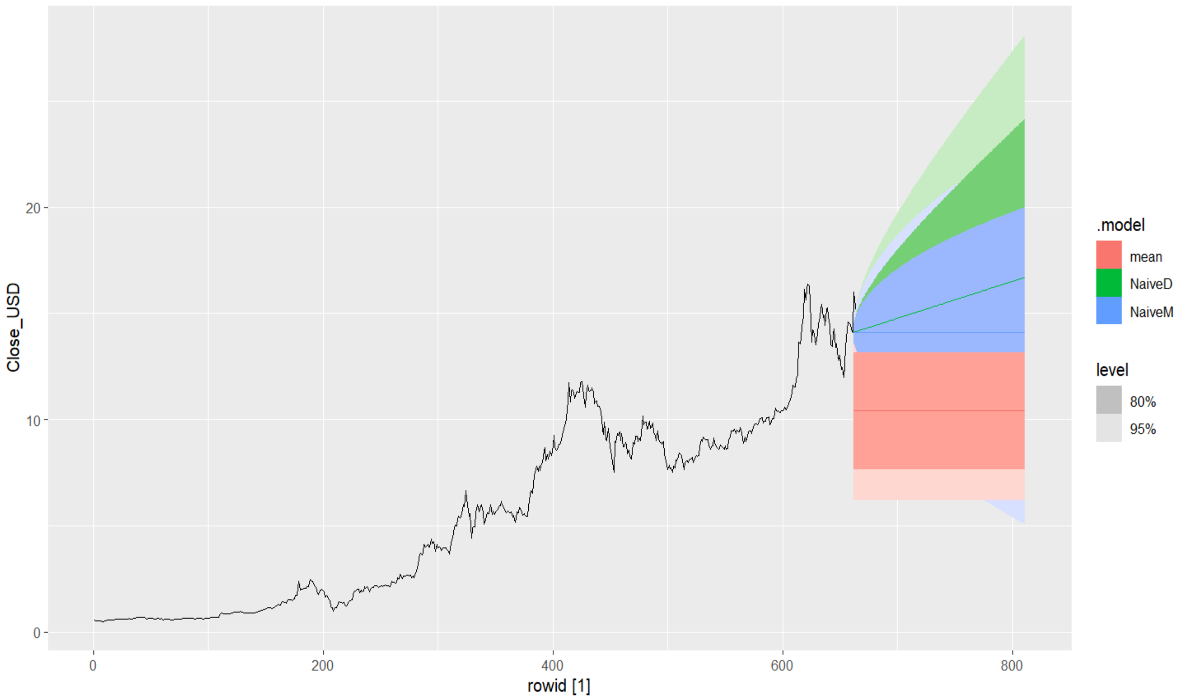
From our residual analysis, we conclude that the naive model (NaiveM) performs reasonably well. The residuals are unbiased and randomly distributed, with a roughly normal distribution. The presence of minor autocorrelation at specific lags (5 and 10) suggests that our model could be further refined to improve accuracy. Overall, this analysis reinforces our earlier findings that the naive model is the best performer among the three models we evaluated for forecasting Tesla's stock prices.

Comparison with Other Models

When comparing the naive model with drift (NaiveD) and the mean model, we find that the naive model (NaiveM) outperforms them both. The residuals from the naive model with drift likely show a pattern due to the incorrect assumption of an upward trend, reflected in its higher RMSE (3.13) and MAE (2.69) values. Similarly, the mean model's residuals would likely be cantered around zero but exhibit more significant deviation patterns due to the simplistic approach of predicting a constant mean value, resulting in intermediate RMSE (2.46) and MAE (2.07) values.

Forecasting Tesla Stock Prices

In this section, we present the forecast of Tesla's stock prices using three different models: the mean model, the naive model, and the naive model with drift. The accompanying plot shows the historical stock prices and the forecasts for each model, along with their 80% and 95% prediction intervals.



Historical Stock Prices

The black line represents the historical stock prices of Tesla (Close_USD) over time, ranging from rowid 1 to approximately rowid 600. We can observe significant fluctuations in the stock prices, with notable upward trends and volatility. This historical data is crucial as it forms the basis for our forecasting models.

Forecast Models and Prediction Intervals

We have implemented three forecasting models to predict Tesla's stock prices:

1. **Mean Model (red shaded area):**

- The mean model forecasts future prices as the average of past prices. The red line represents the mean forecast, while the shaded areas indicate the 80% (darker red) and 95% (lighter red) prediction intervals.
- The prediction intervals for the mean model are wide, reflecting high uncertainty in the forecast. This is because the model does not account for trends or patterns in the data, resulting in a static prediction that extends horizontally from the end of the historical data.

2. **Naive Model (blue shaded area):**

- The naive model assumes that future prices will be the same as the last observed price. The blue line represents the naive forecast, and the shaded areas show the 80% (darker blue) and 95% (lighter blue) prediction intervals.
- The prediction intervals for the naive model are narrower than those of the mean model, indicating relatively lower uncertainty. This model's simplicity makes it useful for short-term forecasting but less effective for longer-term predictions where trends may change.

3. **Naive Model with Drift (green shaded area):**

- The naive model with drift incorporates a trend component, assuming that future prices will follow a linear trend based on the historical data. The green line represents the naive with drift forecast, with the shaded areas indicating the 80% (darker green) and 95% (lighter green) prediction intervals.
- The prediction intervals for the naive model with drift are the widest, reflecting the increasing uncertainty as the forecast horizon extends. This model accounts for the upward trend observed in the historical data, leading to an increasing forecast over time.

Interpretation and Evaluation

- **Prediction Intervals**: The prediction intervals provide a range within which we expect the future stock prices to lie, with 80% and 95% confidence levels. Wider intervals indicate higher uncertainty, while narrower intervals suggest more confident predictions.
- **Model Comparison**: The naive model (NaiveM) provides a balance between simplicity and reasonable prediction accuracy for short-term forecasts. It has narrower prediction intervals compared to the mean model, indicating less uncertainty. The naive model with drift (NaiveD) is useful for capturing trends and providing long-term forecasts, although it comes with wider prediction intervals, reflecting higher uncertainty as the forecast horizon increases. **The mean model** has the widest prediction intervals and does not account for any trends or patterns, making it the least effective for capturing the dynamic nature of Tesla's stock prices.

Conclusion

Our analysis using three different forecasting models reveals the strengths and limitations of each approach. **The naive model (NaiveM) is suitable for short-term forecasts due to its simplicity and narrower prediction intervals. The naive model with drift (NaiveD) is beneficial for capturing trends over the long term but comes with increased uncertainty**. The mean model, while the simplest, is the least effective due to its inability to capture trends or patterns. These insights help us understand the potential future movements of Tesla's stock prices and guide our decision-making in selecting the appropriate forecasting model.

Areas for Improvement

In our analysis of Tesla's stock prices using the naive, mean, and naive with drift models, we identified several areas for improvement. Firstly, the presence of minor autocorrelation in the residuals suggests that our models could benefit from more sophisticated techniques, such as **ARIMA (AutoRegressive Integrated Moving Average) models**, which account for autocorrelation and provide a more accurate forecast. Additionally, incorporating seasonality into our models could capture periodic fluctuations in stock prices, further improving forecast accuracy. We also noted some outliers in the residuals, indicating extreme prediction errors. Addressing these outliers by applying robust statistical methods or transformations can stabilize the variance and enhance model performance. Another significant improvement would be the inclusion of exogenous variables such as market indicators, economic factors, and news sentiment, which can provide additional context and information for better predictions. Finally, combining forecasts from multiple models using ensemble methods can leverage the strengths of different approaches, resulting in a more robust and reliable forecast. By implementing these enhancements, we aim to refine our predictive capabilities and provide more accurate and insightful forecasts for Tesla's stock prices