Report

# Introduction

This project is to analyze the data regarding the selected dataset of Microsoft stock values for time series trends and forecasting. The final outcome for the dataset could be like how well they can be illustrated for forecasting of future values.

To achieve the desired result, I was supposed to analyse the scenario on how the given points can be linearized with the help of methods of visualizing the data and various statistics. The visualization of dataset on various attributes was done.

# Description of Dataset

The stock values from the year 2015 till 2021 were downloaded from the stocks website for Microsoft Inc. This datasets was used in the analysis. The dataset later splitted into the datasets are train and test for forecasting the ARIMA model.

The primary dataset file contains 1511 rows with 6 attributes. Following information is furnished in the dataset:

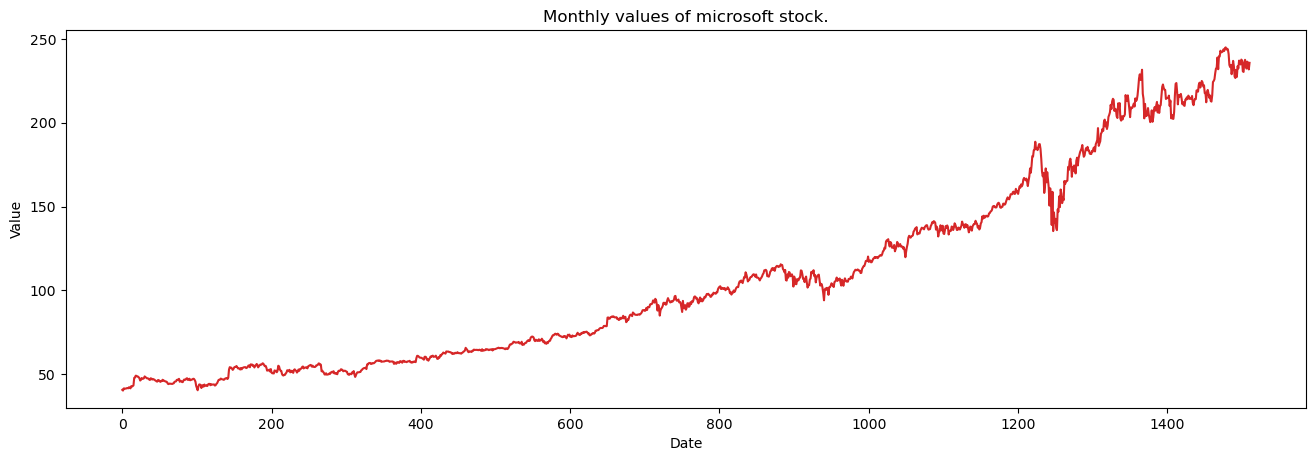
* Date – on which date the stock value recorded.
* Open – what was the opening value
* High – highest intraday value during the date
* Low – lowest intraday value during the date
* Close – closing value of the stock on the date
* Volume – total volume of stocks on the date

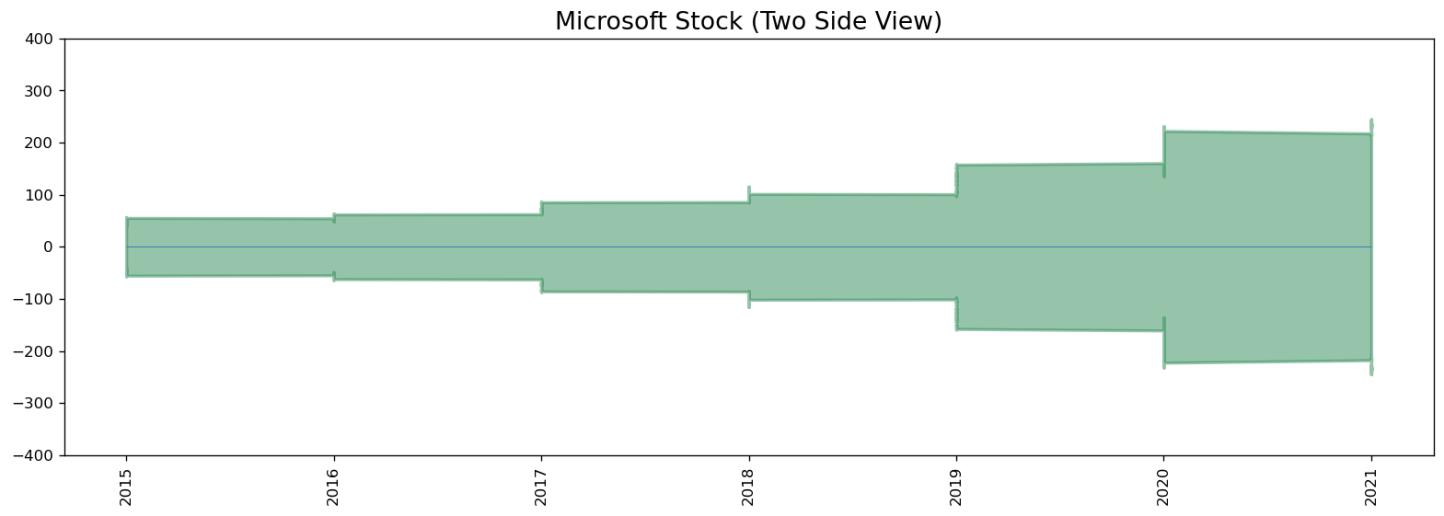
# Handling Missing Values

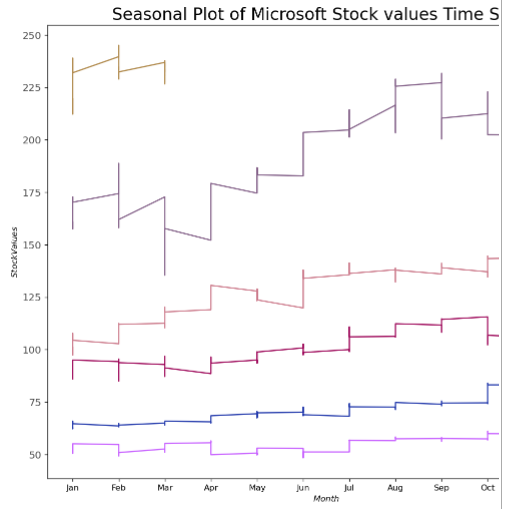
The dataset has no missing values in any of the fields in any of the dataset.

# Visualizations

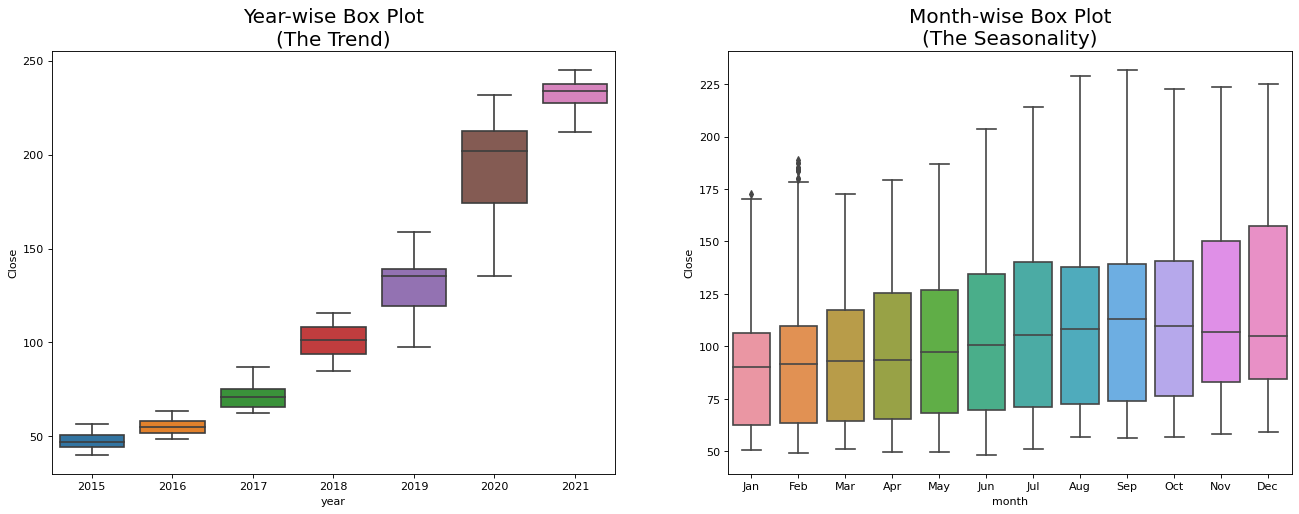
On the selected dataset some visualization was performed.



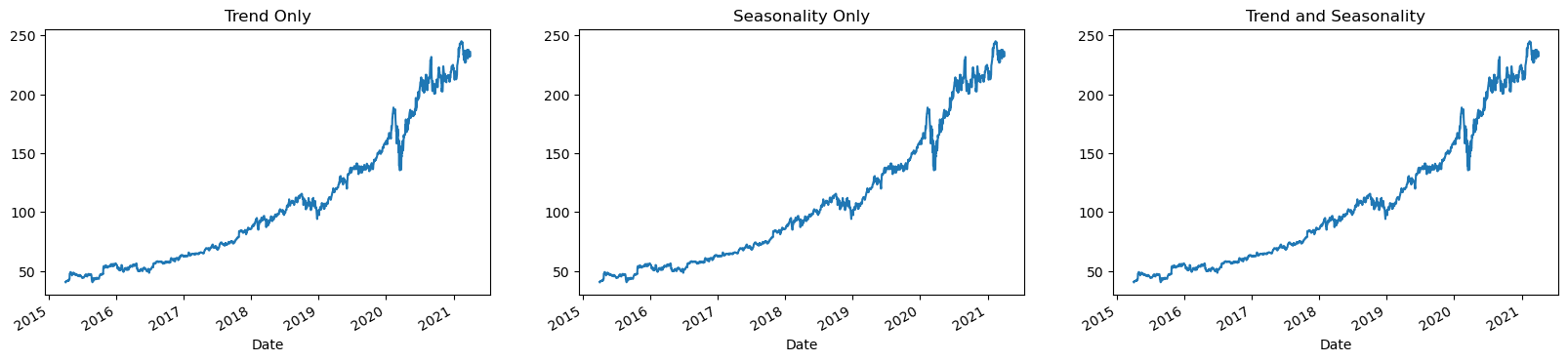




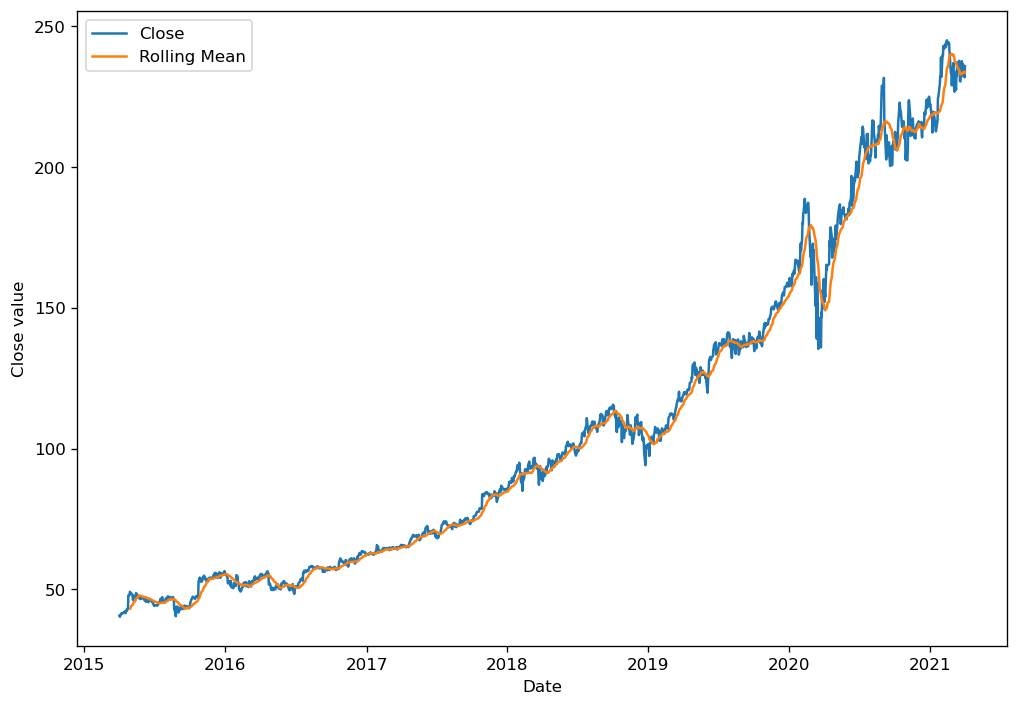
Trend and seasonality plots were built



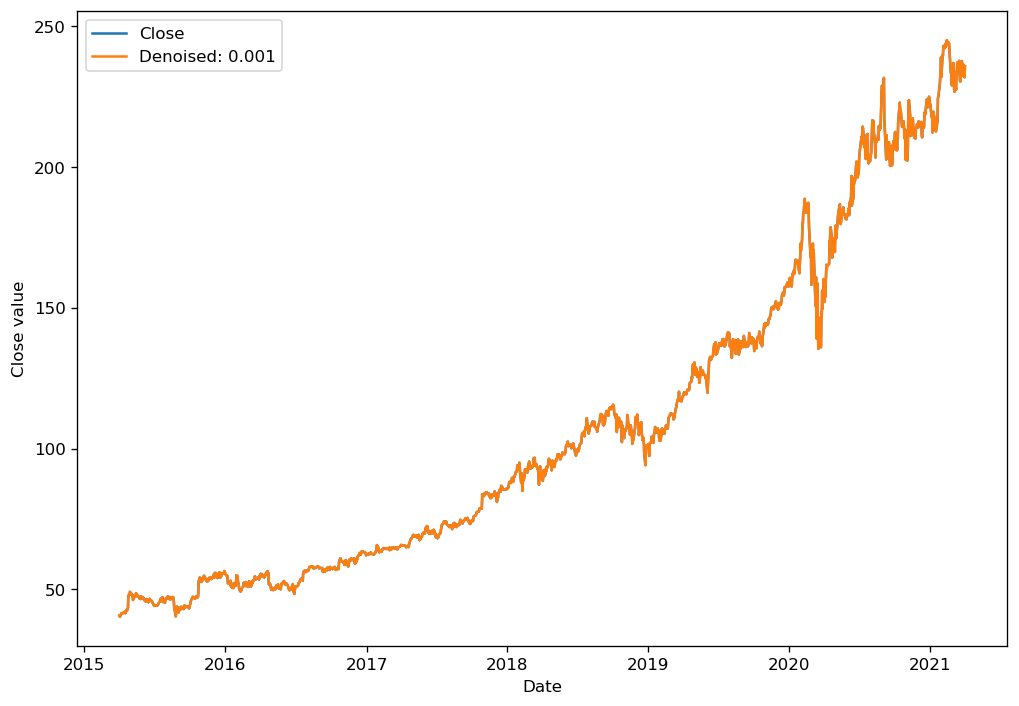
Trend, Seasonality and Trend & seasonality plot



Plot with rolling means



Plot with noise removal



# MODEL BUILDING

Checking whether time series data is stationary or not using Augmented Dickey-Fuller test:

# Ho (Null Hypothesis): The time series data is non-stationary

# H1 (alternate Hypothesis): The time series data is stationary

Augmneted Dickey\_fuller Statistic: 1.737136

p-value: 0.998216

critical values at different levels:

1%: -3.435

5%: -2.863

10%: -2.568

We can see from the p-value, it is clearly greater than 0.05 so we fail to reject the null hypothesis which says the time series data is non-stationary or in other words data is strongly non-stationary.

ARIMA model was created for time series data.

**ARIMA:**

SARIMAX Results

==============================================================================

Dep. Variable: Close No. Observations: 1284

Model: ARIMA(3, 2, 2) Log Likelihood -2586.912

Date: Thu, 28 Sep 2023 AIC 5185.825

Time: 22:56:24 BIC 5216.762

Sample: 0 HQIC 5197.441

- 1284

Covariance Type: opg

==============================================================================

coef std err z P>|z| [0.025 0.975]

------------------------------------------------------------------------------

ar.L1 -1.2170 0.016 -74.840 0.000 -1.249 -1.185

ar.L2 -0.2940 0.018 -16.271 0.000 -0.329 -0.259

ar.L3 0.0890 0.012 7.537 0.000 0.066 0.112

ma.L1 -0.0844 0.014 -6.159 0.000 -0.111 -0.058

ma.L2 -0.9125 0.014 -67.530 0.000 -0.939 -0.886

sigma2 3.2956 0.050 66.173 0.000 3.198 3.393

===================================================================================

Ljung-Box (L1) (Q): 0.00 Jarque-Bera (JB): 13850.98

Prob(Q): 0.96 Prob(JB): 0.00

Heteroskedasticity (H): 10.82 Skew: -0.91

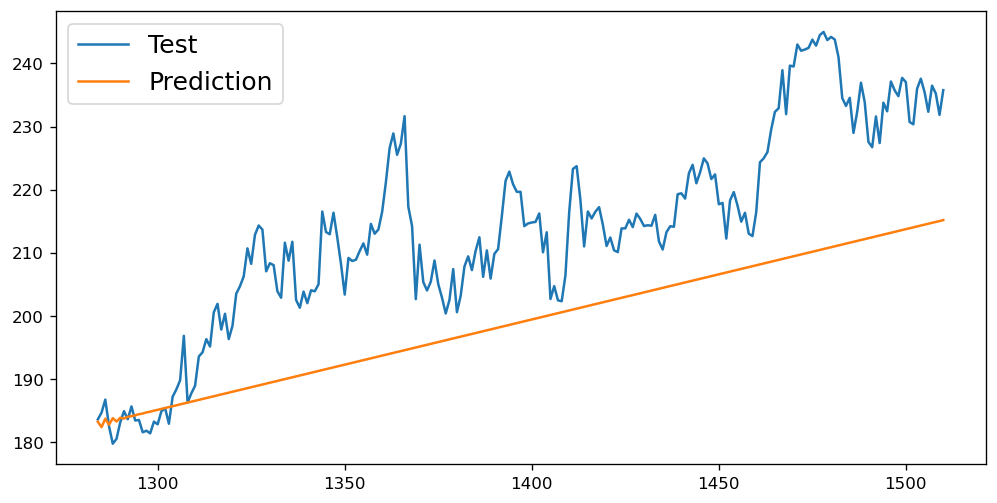
Prob(H) (two-sided): 0.00 Kurtosis: 19.00

===================================================================================

Warnings:

[1] Covariance matrix calculated using the outer product of gradients (complex-step).

**The plot of the model is as follows:**

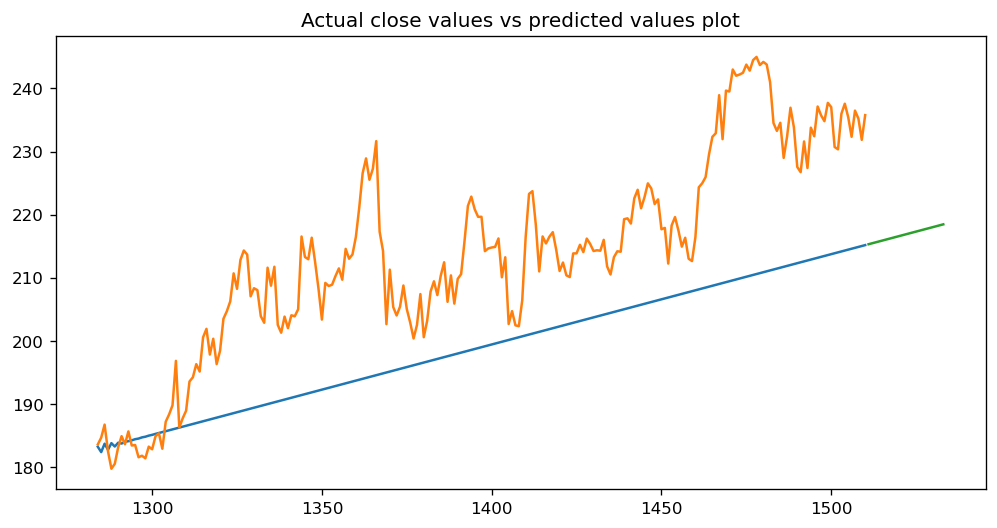
****

Mean Squared error of the model is

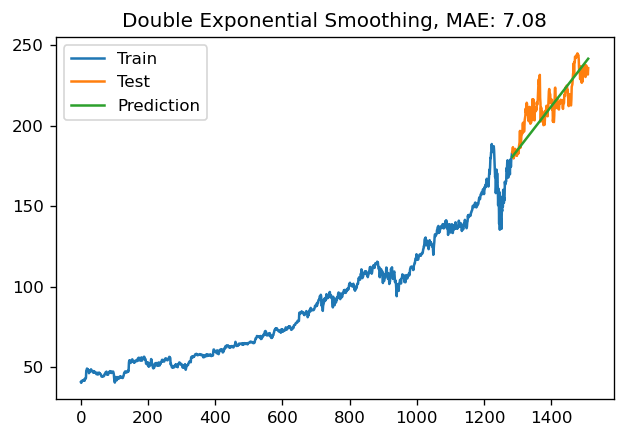
Mean Squared Error:

17.293829441533234

**The plot of the actual vs predicted values is as follows:**



**Smoothing of the time series data:**

****

# Discussion

Analysis was done by first performing the data visualization using various plots on attributes on the dataset. The train and test data sets were used for creating the model. ARIMA model was created.