Jessy Huang

Time Series Data

Professor Hurvich

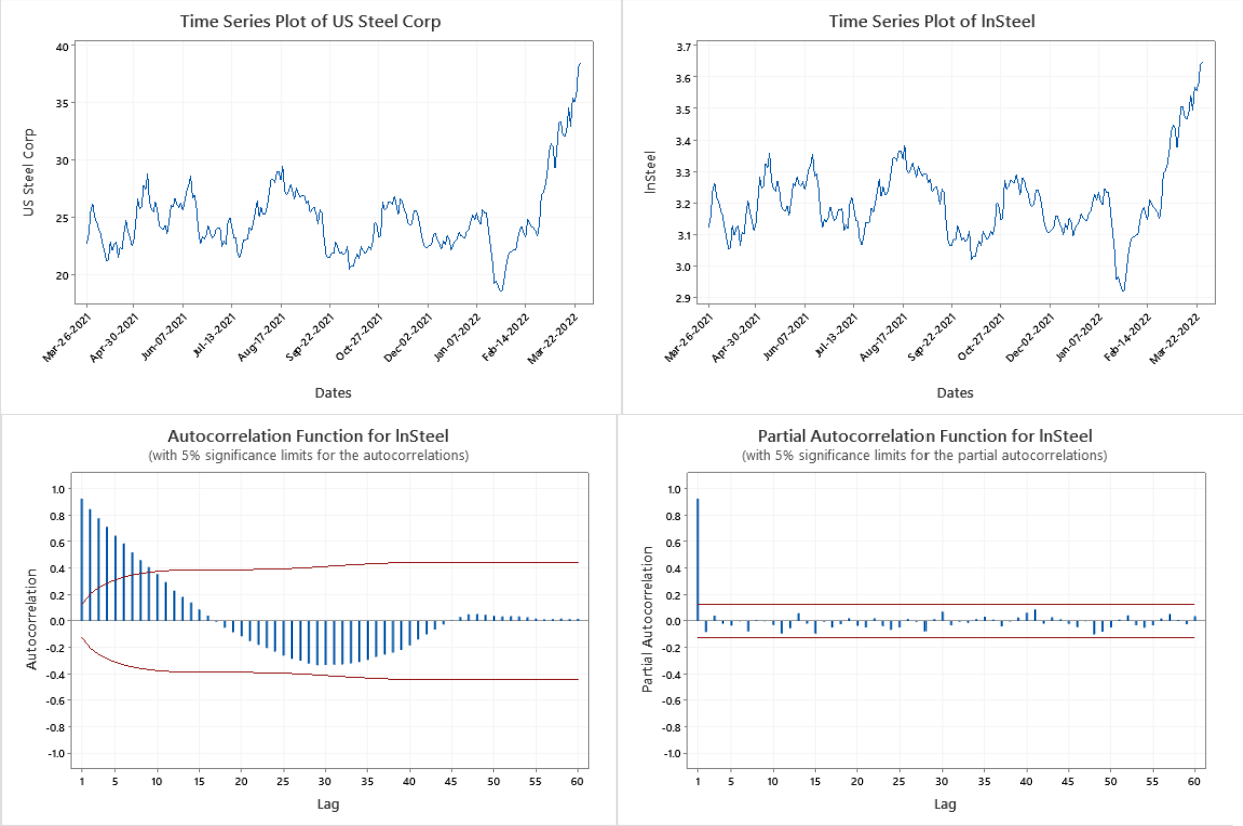
4/4/22

ARIMA Model on Time Series Forecasting

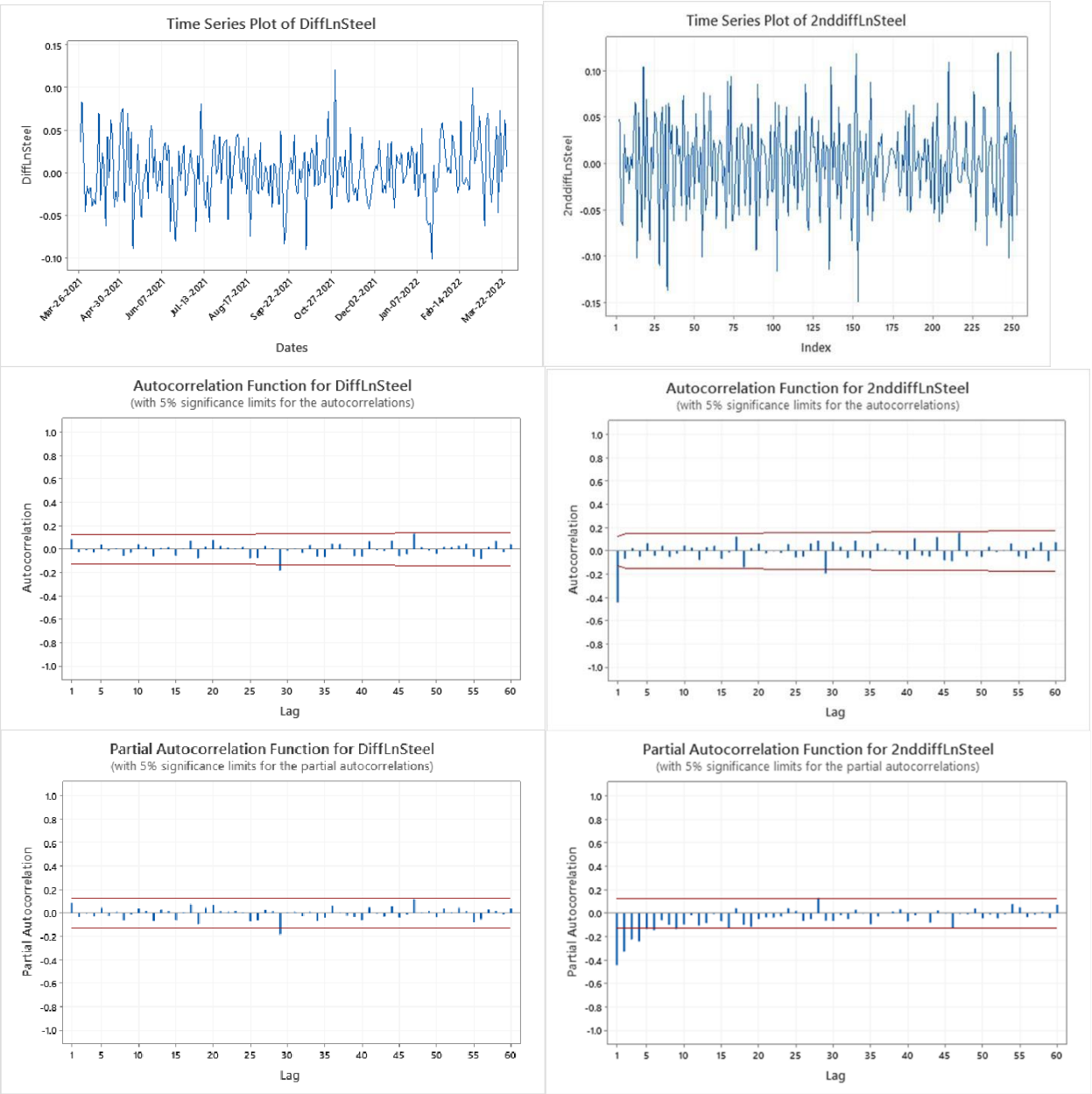
This project analyzes and interprets the stock price of United States Steel Corporation (NYSE: X) over the course of the past year, from May 26, 2021 to March 22, 2022, which is 253 datapoints. The data is sources from CapitalIQ (United States Steel Corporation (NYSE:X) > Chart Builder (capitaliq.com)).

I think it is a good idea to work with the natural log of the pricing data because often, while comparing the time-series plots of non-logged price and differenced prices, the volatility in price tracks with the level of price. So, following the time series plot we construct the natural log plot.

ARIMA Model Identification:



From the time series plot of the natural log price, we can see that the natural log price is not stationary; it appears to have some volatility. The ACF of the natural log price shows a “sine” wave pattern and exponential decay. The PACF of natural log price cut off at first lag, suggesting that only the first lag is significant. Therefore, we should differentiate the natural log price at least once in order to remove most of the volatility from the model.



I plotted the first and second differentiation of natural log of pricing data as well as their respective ACF and PACF to determine whether the first or second differentiation is required for ARIMA modelling. Based on the results, we can see that the first differentiation significantly improves the level of volatility, although some remains. Taking the second differentiation seems to be over differencing, as the ACF and PACF for the second differentiation show a strong pattern of negative autocorrelation in some lags. Therefore, we will use d=1 for the ARIMA model.

The ACF and PACF of the first differences of log price are insignificant at all lags except lag 29 which is perhaps slightly significant. Overall, this suggests a random walk, i.e., an ARIMA(0,1,0) model for the log prices. However, we will test several options.

To determine the range of p and q values to test, we must consider the parsimony rule in model identification: “The total number of parameters in the model should be as small as possible”.