Statistical Data Science Project 1

1. Time series

A time series is a collection of information or observations that were gathered throughout time at regular, repeated periods. The sequence of the data points in a time series is important since they are normally gathered in chronological order. Depending on the situation and the underlying phenomena being studied, time series data can be gathered at different rates, such as daily, monthly, quarterly, or annually.

For this task I used yfinance, a well-known Python library which provides a user-friendly interface for gaining access to financial data from Yahoo Finance. For my project I chose the financial stock prices from Apple.



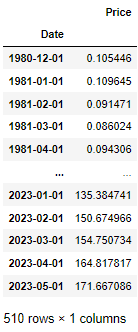
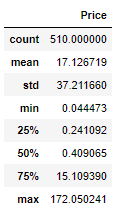
The dataset contains everyday stock prices from 1980 to present. As we have different columns with information regarding Open prices, Highest and lowest price of the day and Close and Adj Close prices and even volume, we will use only one column.

Two often used measures for the closing price when examining stock data are "Close Price" and "Adjusted Close Price." Let's see how these two measurements differ from one another:

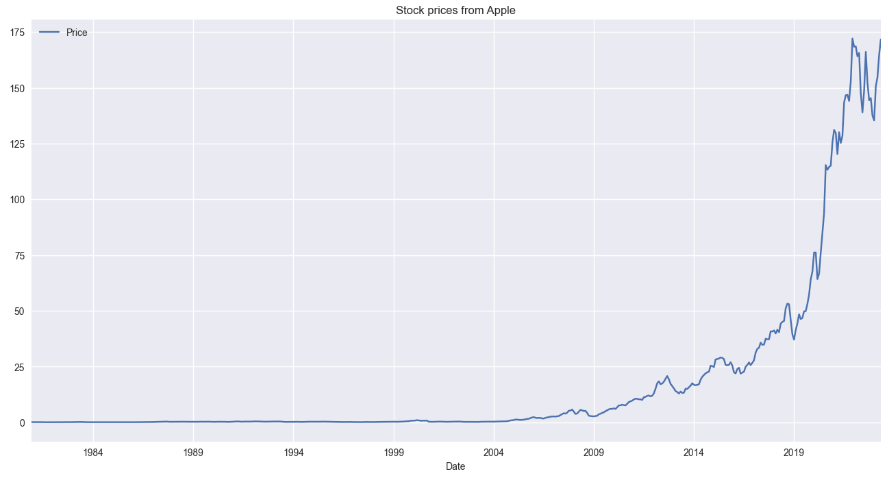
* The "close price," often referred to as the "closing price" or simply "Close," is the stock's last traded price at the conclusion of a certain trading day. It is the cost at which the final transaction for the day took place. When examining daily performance, the close price for a stock is often the one that is most frequently used as a reference.
* The adjusted close price takes into consideration business decisions that might affect the stock price, such as stock splits, dividends, and other alterations. It represents the price that takes these aspects into account to give a more realistic picture of the stock's actual performance over time. The impact of dividends is taken into account by changing the historical prices to reflect the adjusted close price, which also takes into account changes in the number of shares outstanding brought on by stock splits.

So, for a more complete picture of a stock's past performance, taking into account any factors that could have affected the price we will chose Adj Close column.

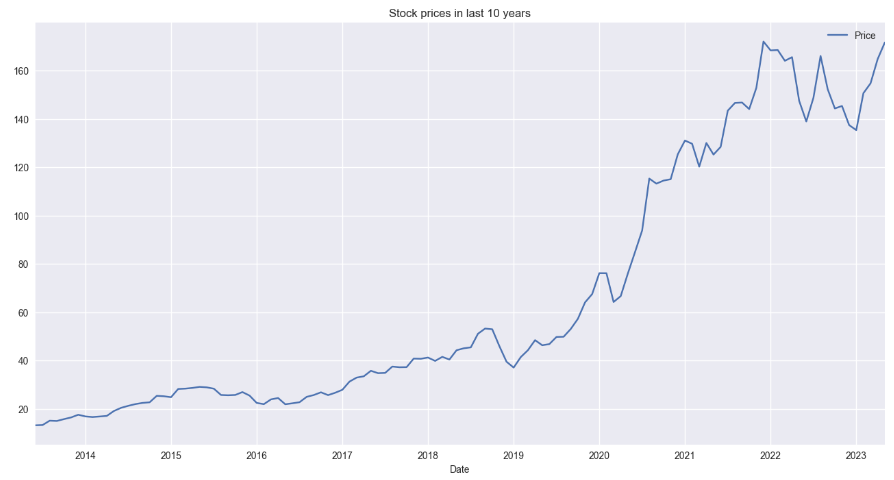
Also, we will resample the data from its original frequency to a monthly frequency and will use the average Adj closing price for that month. The final dataset looks like this:

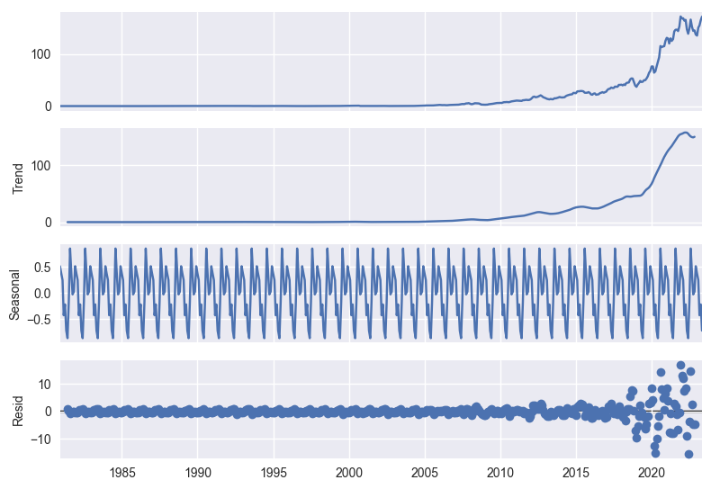
At first look we can observe that the stock prices went from around 10 cents to around 171 dollars in almost 40 years.



As we can see, we have an ascending trend in time, let’s have a better look what happened in last 10 years.



Around middle of 2020 we have a big increase in stock prices and this could be the cause of covid 19 when different technology companies had a rise of stock prices. “The pandemic turned 2020 into a year of unprecedented events — not the least of which was the swift crash and then record-fast recovery of the stock market.” (<https://www.cnbc.com/2020/12/30/how-the-pandemic-drove-massive-stock-market-gains-and-what-happens-next.html>)



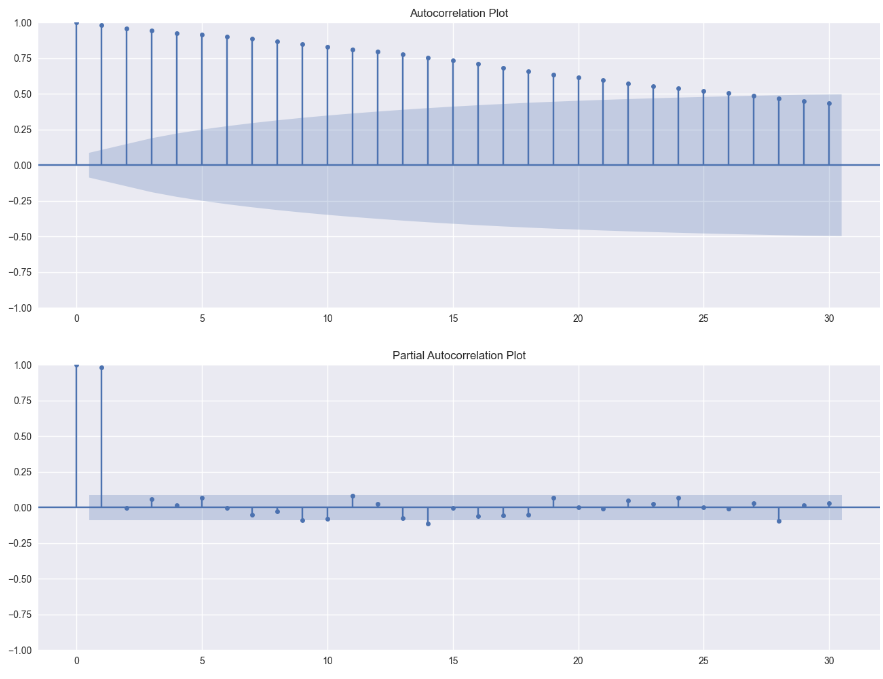
From previous analyses we observed the ascendent trend but now we can see better also the seasonality so our data is not stationary. “Non-stationary behaviors can be trends, cycles, random walks, or combinations of the three. Non-stationary data, as a rule, are unpredictable and cannot be modeled or forecasted.”

The following can indicate a series' stationary behavior:

* the mean and constant mean It doesn't rely on time
* constant variance and a non-time-dependent variance
* constant covariance and non-time-dependent covariance

(<https://www.investopedia.com/articles/trading/07/stationary.asp>)

Autocorrelation Function (ACF) and Partial Autocorrelation Function (PACF)



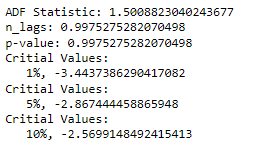
This is an essential step since we may utilize those two functions to select the ARIMA model's parameters p and q.

The ACF calculates the correlation between a time series' lagged values and an observation. It assists in locating any patterns, trends, or dependencies that may be present in the data at various delays.

The PACF takes into account correlations at shorter lags when measuring the correlation between an observation and its delayed values. By eliminating the indirect impact of intervening lags, it makes it easier to see the direct link between observations at various delays.

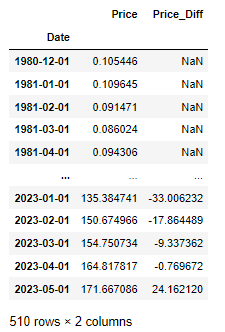
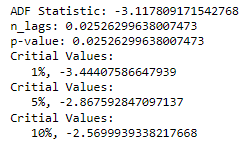
(<https://statisticsbyjim.com/time-series/autocorrelation-partial-autocorrelation/>)

Testing for Stationarity we can do it with a visual approach or with a hypothesis testing named Augmented Dickey Fuller Test. This test consists in having 2 hypotheses where the null hypothesis can be rejected if the p-value is below a set significance level (0.05).



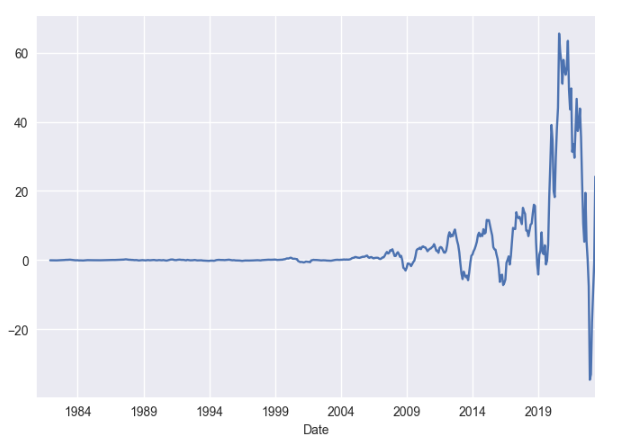
The p-value is obtained is greater than significance level of 0.05 and the ADF statistic is higher than any of the critical values.

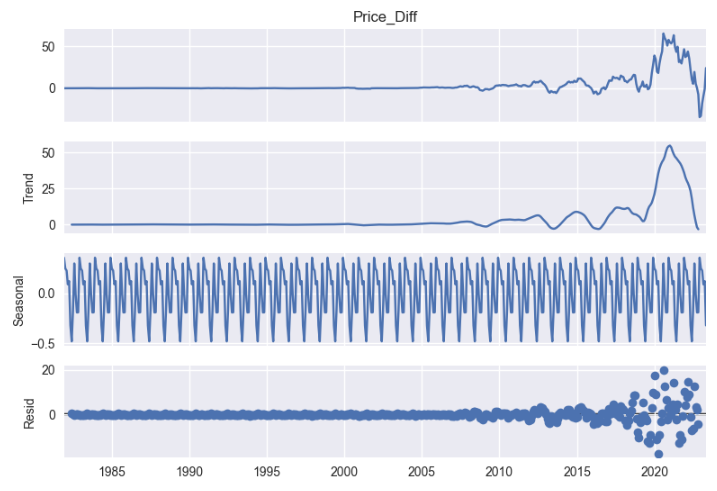
We apply a 12 shift differencing to remove annual seasonality and the data looks like this and we apply ADF test again:

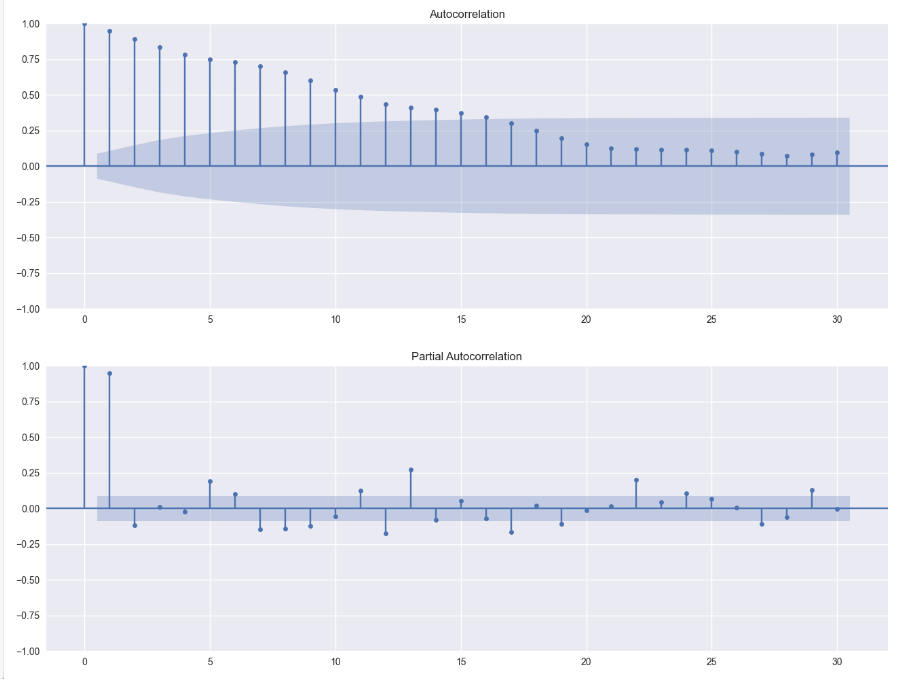
The p-value is very less than the significance level of 0.05 and hence we can reject the null hypothesis and take that the series is stationary.

(<https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/>)





Because of the substantial seasonality of our data, a seasonal ARIMA model will be used in this instance rather than a regular ARIMA.



SARIMAX (Seasonal Autoregressive Integrated Moving Average with Exogenous factors) combines the ARIMA model with seasonality and external variables. It is frequently employed in many different fields, including anticipating demand, economics, and finance.

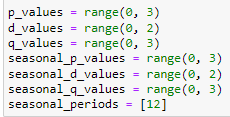
Autoregressive (AR) Order (p): The autoregressive order, indicates how many lag-time observations are incorporated in the model. It illustrates how the present observation relates to earlier observations. Analyzing the partial autocorrelation function (PACF) plot provides the parameter p. (1,2,6)

Integrated (I) Order (d): The integrated order, shows the amount of differencing necessary to ensure stationarity in the time series. In our case we discovered that a single difference by one shift must be applied in order to establish stationarity. So, our d is 1.

Moving Average (MA) Order (q): The number of lags in the model's predicted mistakes is set by the moving average order. It illustrates the connection between the most recent and past prediction inaccuracies. A longer recall of prior mistakes is indicated by a greater value for q. The ACF plot is examined to determine the parameter q. In our case from the plot, I am not sure what is the best lag to choose.

Seasonal Cycle Length (s): The seasonal cycle length (s) is the quantity of time steps that make up a single seasonal period. Understanding the underlying seasonal trend in the data helps to determine it. In our case we use 12.

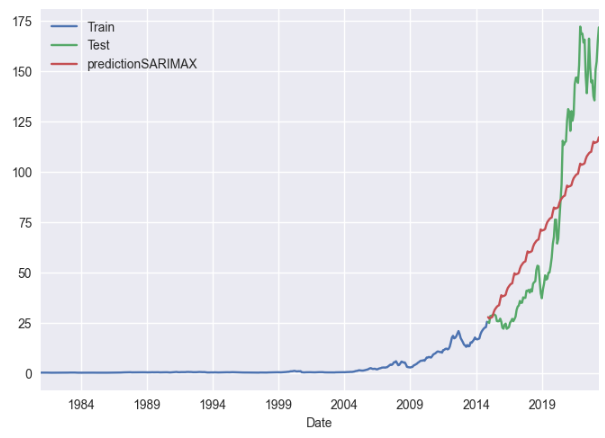
For the best results I built a function called sarimax\_param\_tuning with splits our data in 80% train data and 20% test data and we perform a parameter tunning and evaluation is done on predicted forecasts through RMSE (root mean squared error).

We use 324 combinations to see which paramters berform the best.

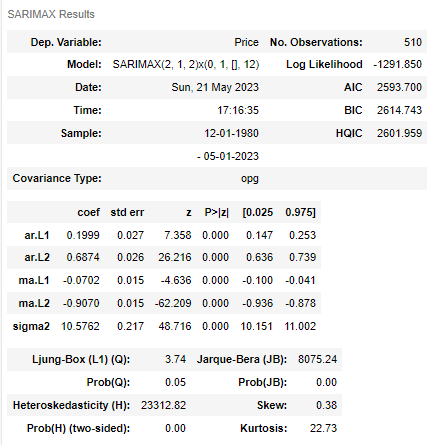


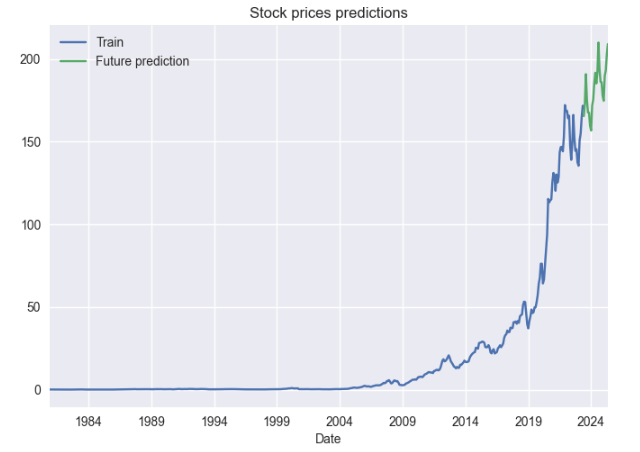
After almost 7 minutes we obtained the best parameters with the lowest rmse.

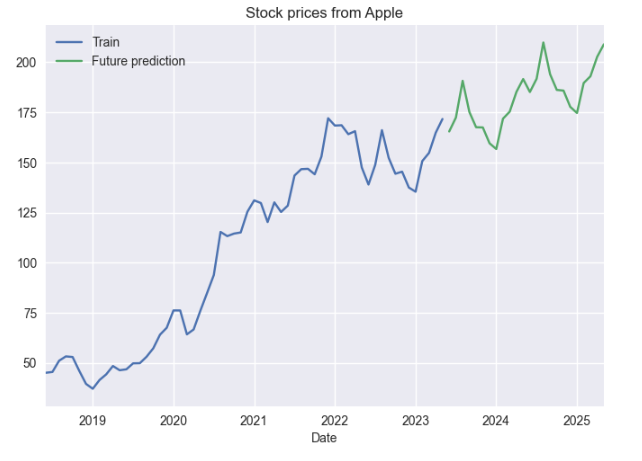


This is how our prediction looks like on training data.

Let’s apply our parameters on a new SARIMAX model to predict the future 24 months of stock prices.







Task 2  
The Durbin-Levinson (D-L) algorithm is a recursive method for solving linear equations in an autoregressive (AR) model that is utilized in signal processing and statistics. It is made particularly to solve the Yule-Walker equations, which link an AR model's parameters to a time series' autocorrelation function. (<https://www.sciencedirect.com/science/article/pii/S2405844018331888>)

The autocorrelation function of a stationary time series is related to the coefficients of an autoregressive (AR) model by the Yule-Walker equations. The Yule-Walker equations can be expressed as follows for an AR(p) model:

R(k) = φ(1)R(k-1) + φ(2)R(k-2) + ... + φ(p)R(k-p)

R(k) is the autocorrelation at lag k

φ(p) is coefficient of the AR model

p is the order of AR model

(<http://www-stat.wharton.upenn.edu/~steele/Courses/956/Resource/YWSourceFiles/YW-Eshel.pdf>)

AR(2) with the parameters Φ =(1,-.9), 200 observations.

