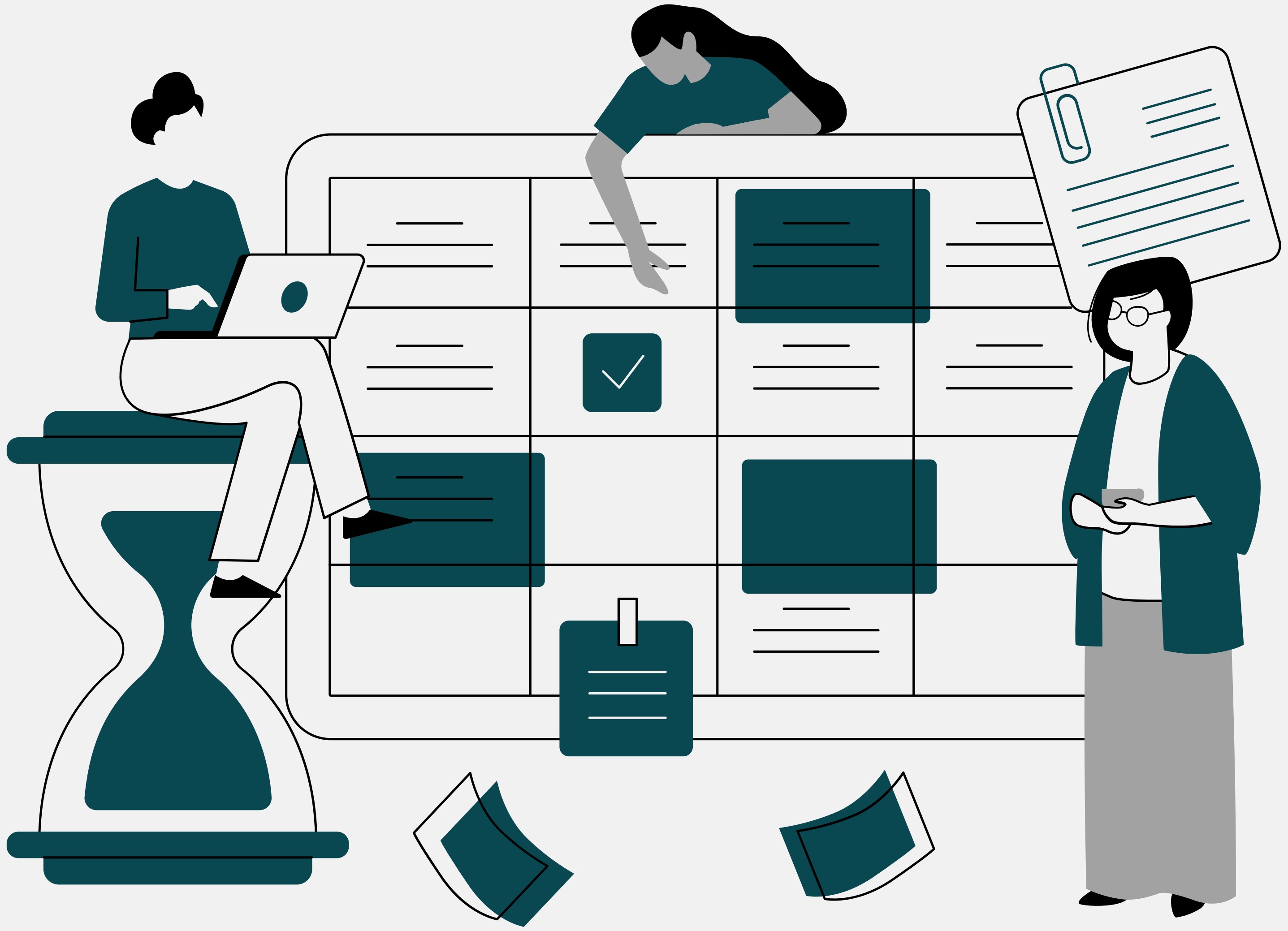
Predicting Equity price with Time-Series Analysis and Machine Learning

General Assembly
DSIR - 1101

MASON LEE

Introduction

A brief look at what we will discuss on this project



01 Problem Statement

02 Data Collection

03 Exploratory Data Analysis

04 Data pre-processing & Modeling

05 Results & Next Steps

06 Questions

Problem Statement

- **Data-Collection using API**
- **Time-Series Analysis and Machine Learning:**
  - **AutoRegressive Integrated Moving Average (ARIMA) model**
  - **VectorAutoregressive (VAR) model**
  - **Long Short-Term Memory (LSTM) model**
- **Outcome Discussion**

The goal in this project is to collect historical stock trading data using an API, and perform Time-Series analysis and Machine Learning with ARIMA model, VAR model, and LSTM model to predict and forecast future price of the equity.

The success of outcome will be measured based on how accurately the models predict equity price of interest and compared to the actual prices of the equities using MSE and RMSE.

Data Collection

Yahoo Finance API

Historical trading data of equities of interest was obtained using Yahoo Finance API. It provides easy-to-use tool to collect OHLC, Adjusted Close, and Volume of given equity.

Period: 01/01/2011 - 12/31/2021

Time-Series data was collected for all **TRADING DAYS** from 2011 to 2021.

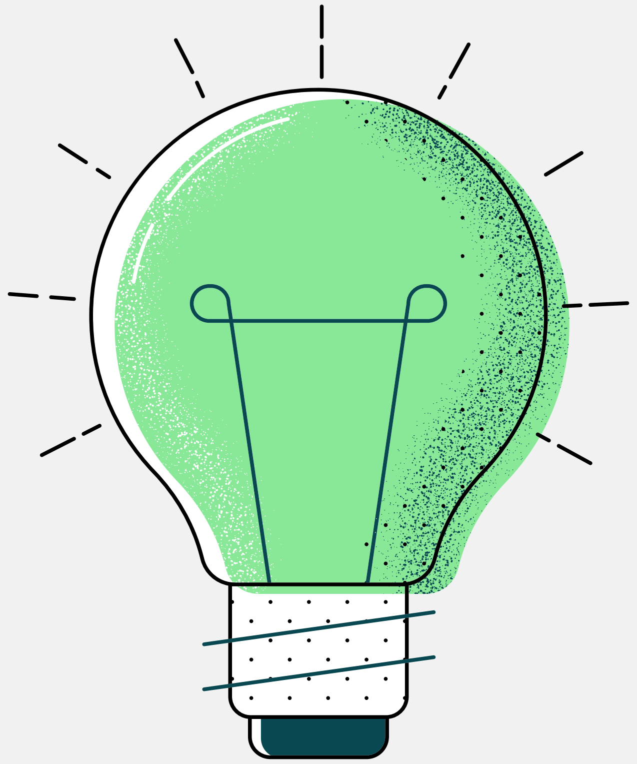
Collected Equity Names

SPY - SPDR S&P 500 ETF Trust
QQQ - Invesco QQQ Trust Series 1, NASDAQ ETF
AAPL - Apple, Inc.
AMZN - Amazon, Inc.
GOOG - Alphabet, Inc. (Google)
MSFT - Microsoft, Inc.
TSLA - Tesla, Inc.
VIX - Chicago Board Options Exchange's CBOE Volatility Index.

Glossary of Terms

TERM	DESCRIPTION
Open	Opening price of the equity for a given period
High	Highest price of the equity for a given period
Low	Lowest price of the equity for a given period
Close	Closing price of the equity for a given period
Adj Close	Adjusted closing price of the equity for a given period, accounting for any corporate actions, such as stock splits, dividends, and rights offerings.
VWAP	Volume-Weighted Average Price of the equity, calculated by taking the total dollar value of trading in the security and dividing it by the volume of trades for a given period.
Daily_pct_change	Daily percentage change of the equity, also referred as daily returns
log_Adj_close	Log-transformed adjusted closing price of the equity for a given period, utilized to better compare the performance of the stocks. Log transformation reduces/removes the skewness of the original data.
log_VWAP	Log-transformed VWAP of the equity for a given period, utilized to better compare the performance of the stocks. Log transformation reduces/removes the skewness of the original data.

Exploratory Data Analysis



Cleaning Dataset

- No null values in the initially obtained data
- VWAP calculation
  - VIX does not have traded volume: no VWAP

Plotting Charts

- Plotting daily charts with 50 MA and 200 EMA
  - Observe Golden Cross & Death Cross
- Plotting all charts together to compare **prices**
- Plotting all charts together to compare **performance**

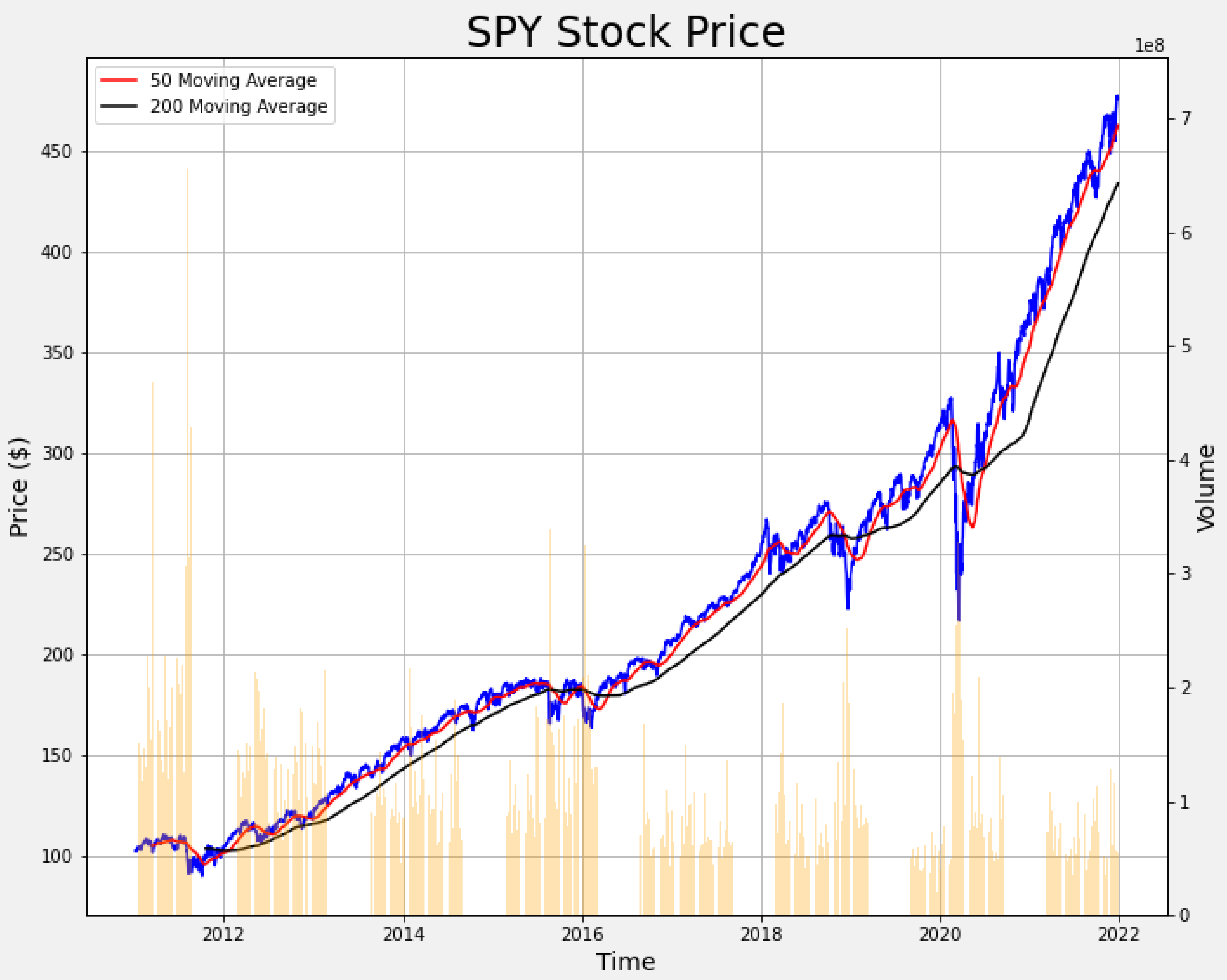
Feature Engineering & Plotting

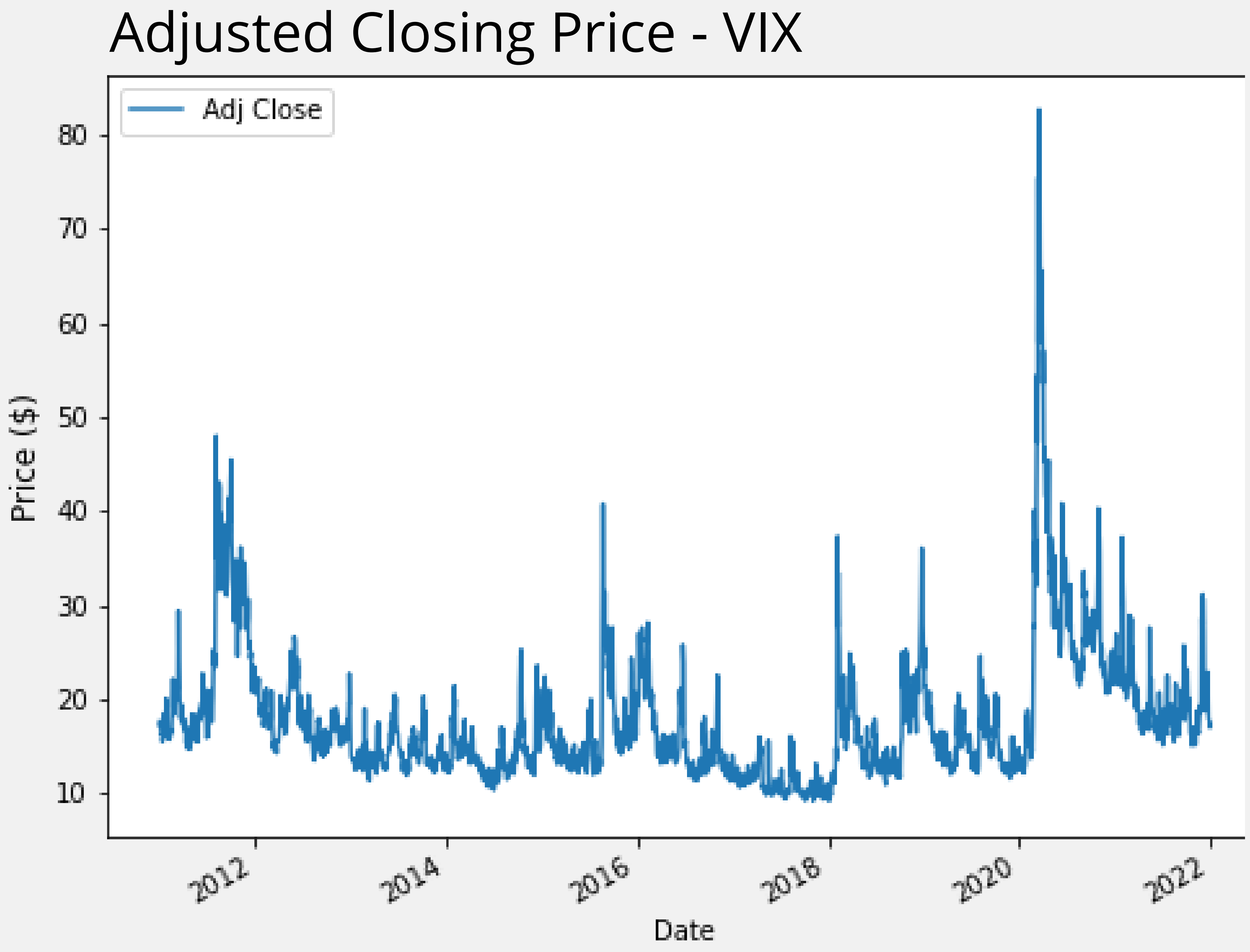
- Traded Volume Analysis
- Daily Percentage Change Analysis
- Histogram & KDE Plots for daily percentage changes for the distribution of data

Relationship between features

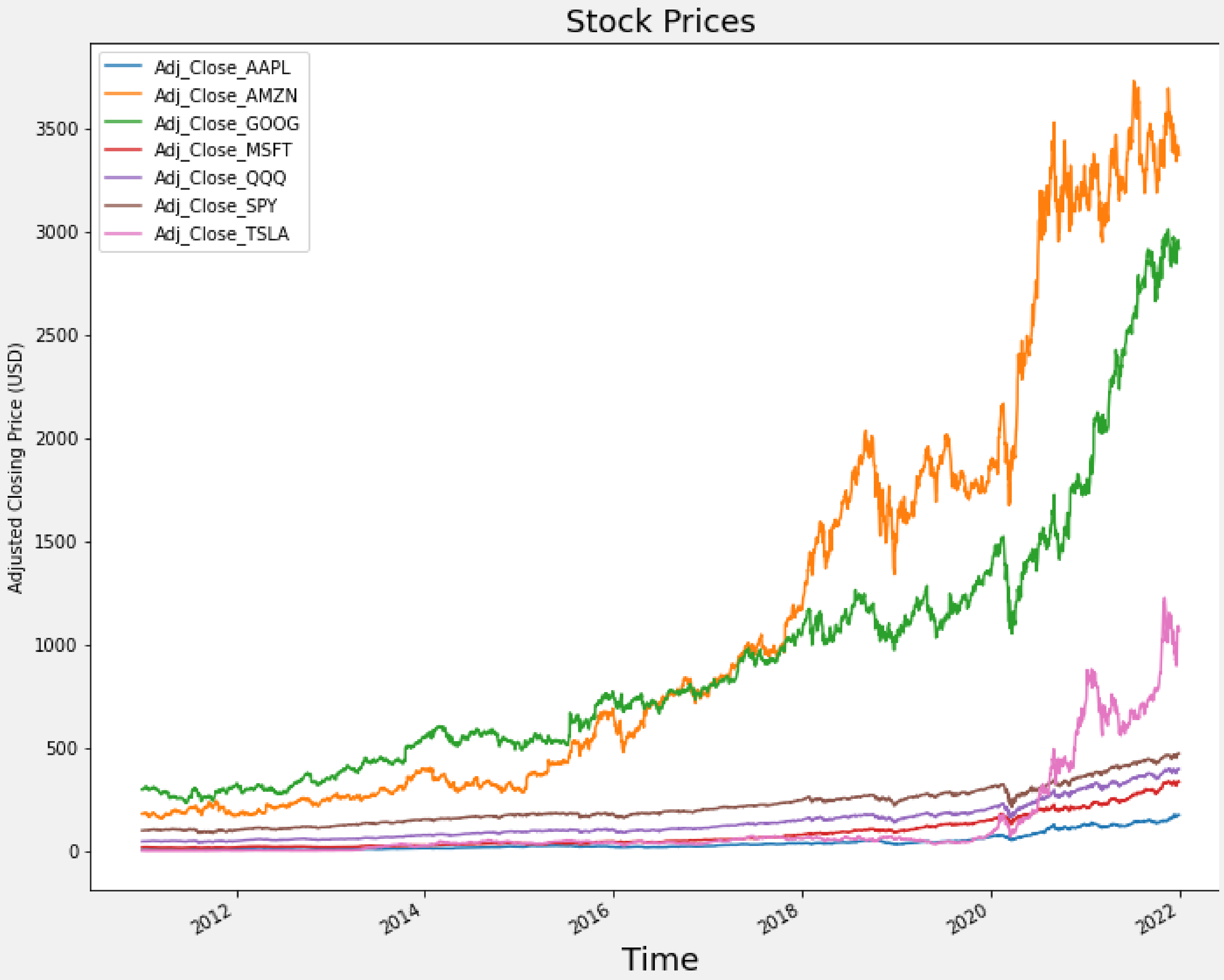
- Correlation Matrices

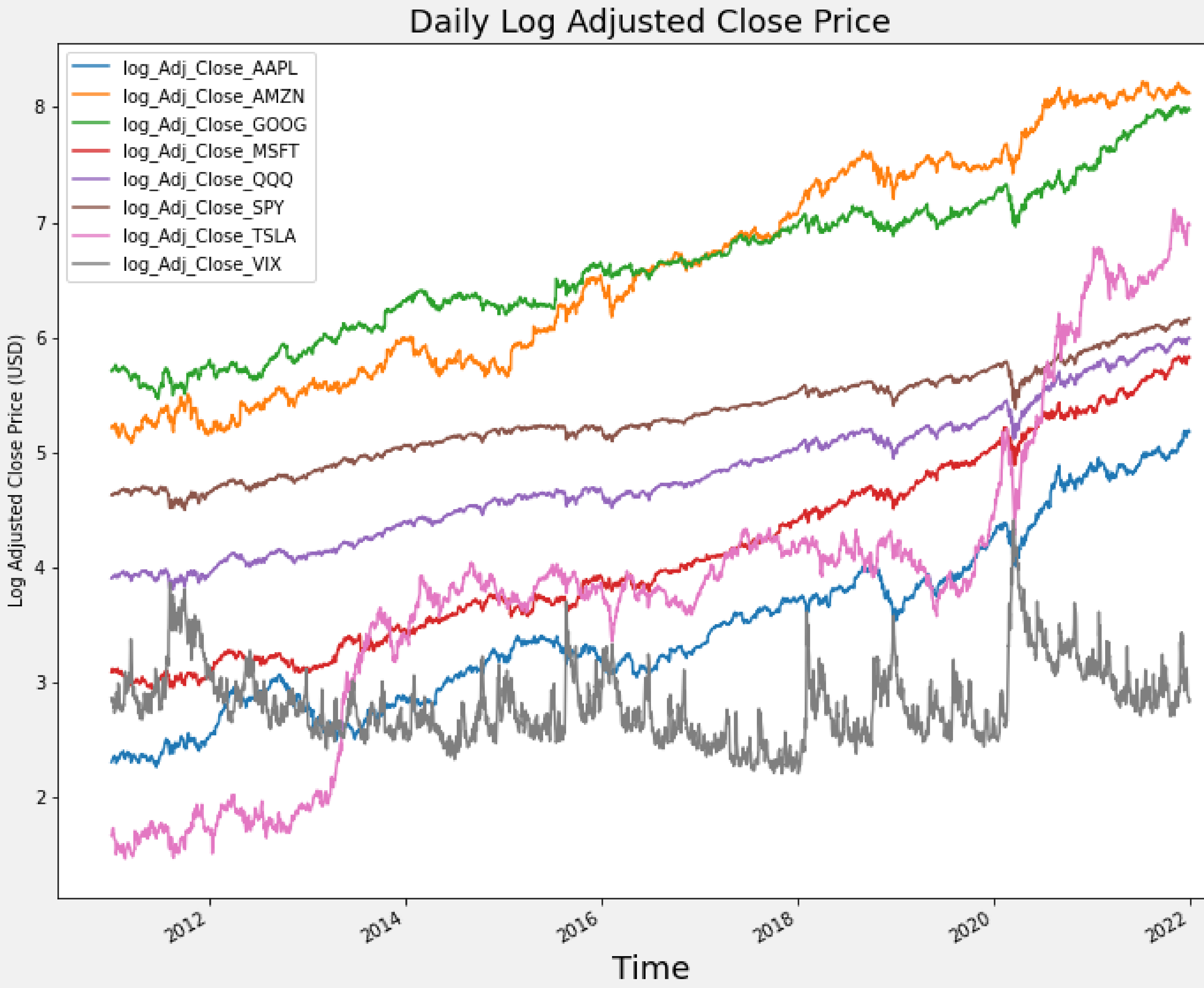
CHARTS - SPY & VIX



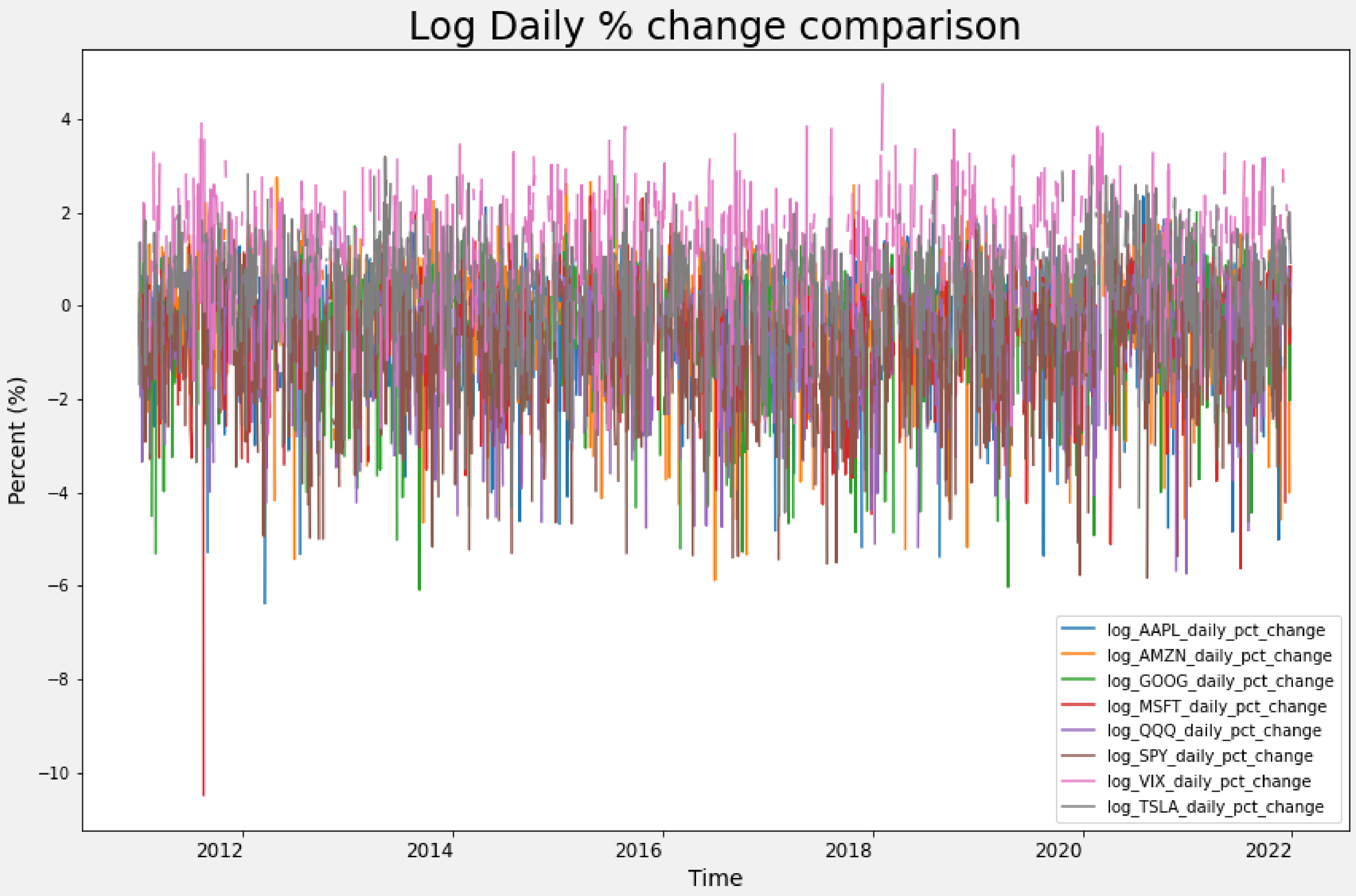


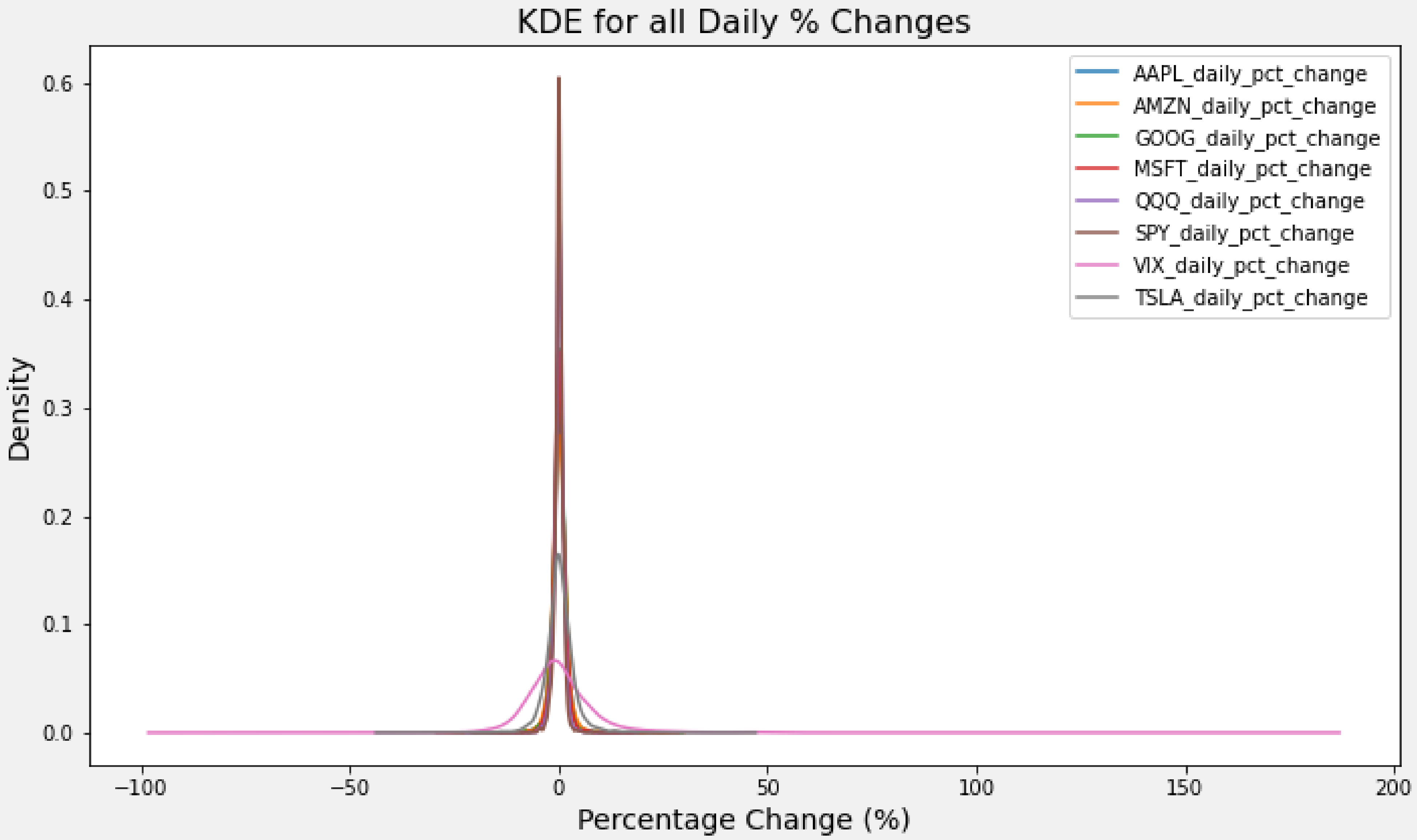
COMBINED CHARTS & PERFORMANCE





CHARTS - DAILY % CHANGE & DISTRIBUTION

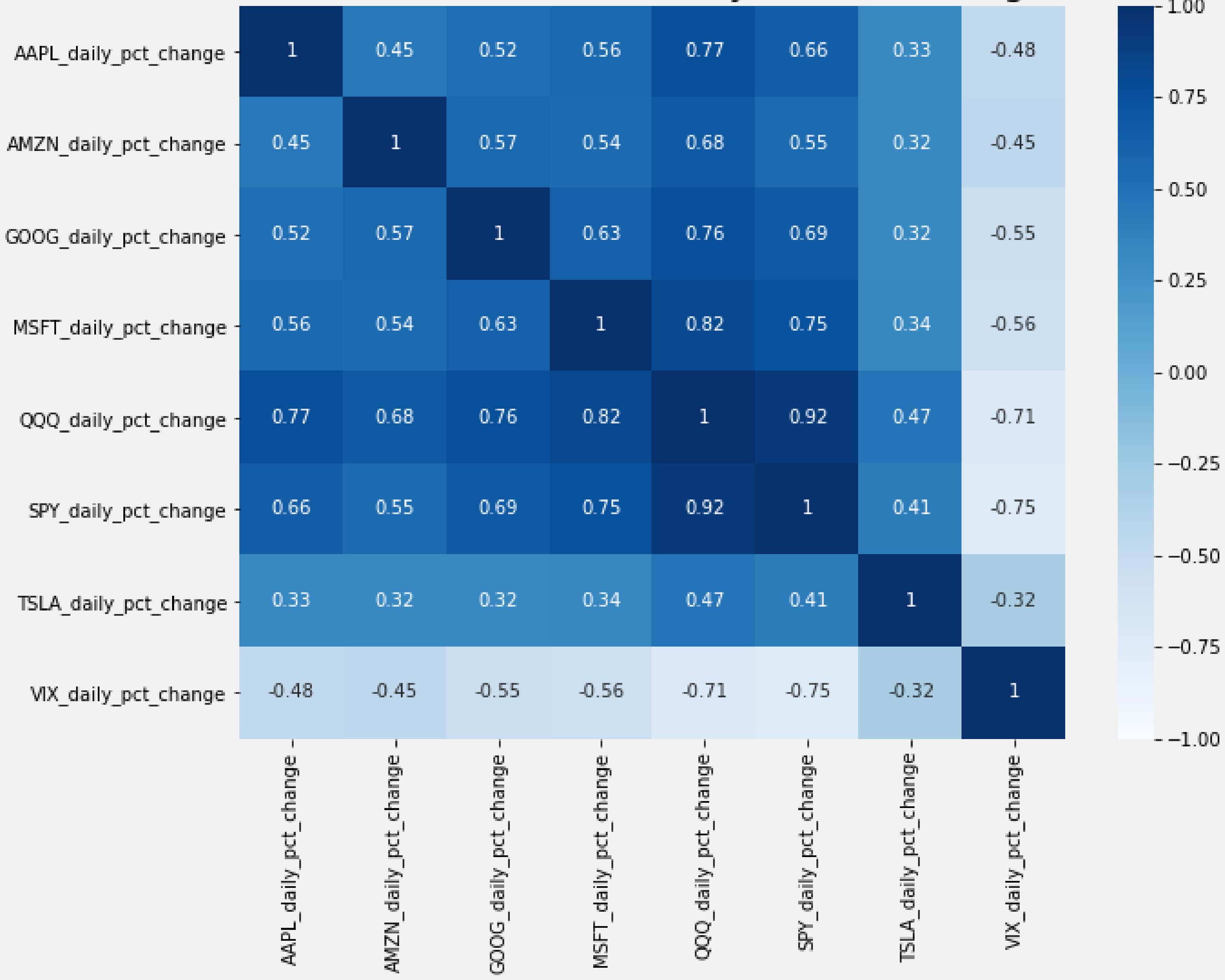




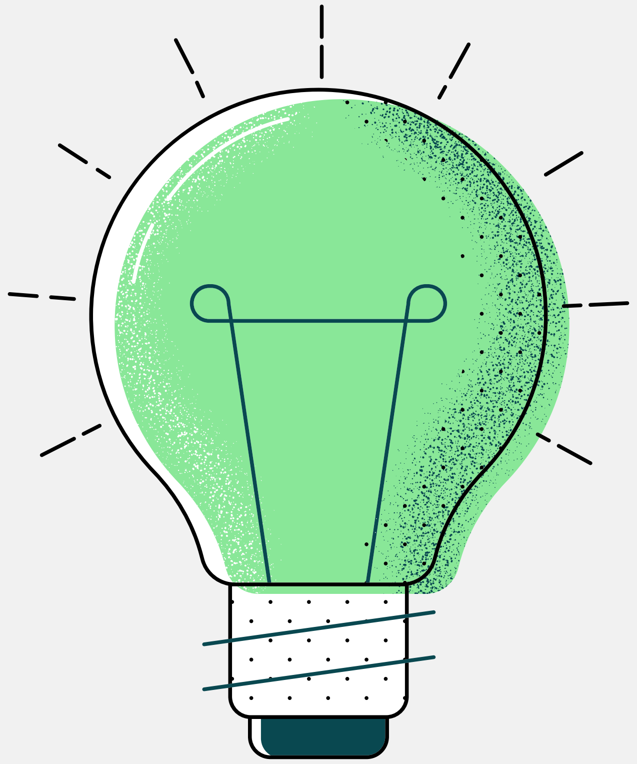
CORRELATION MATRICES

Correlation Matrix for Adj Close Price

Correlation Matrix for Daily % Price Change



Data pre-processing & Modeling



Data pre-processing

- Keeping adjusted closing price, volume, VWAP, and daily percentage change features only
- Data cleaning - no null values
- Splitting Dataset - 80:20 split for time-series: no shuffling

Time-Series Analysis

- Granger-Causality Test
- ACF & PACF Plots
- Seasonal Decomposition
- Augmented Dickey-Fuller (ADFuller) Test
- Differencing Data for Stationarity

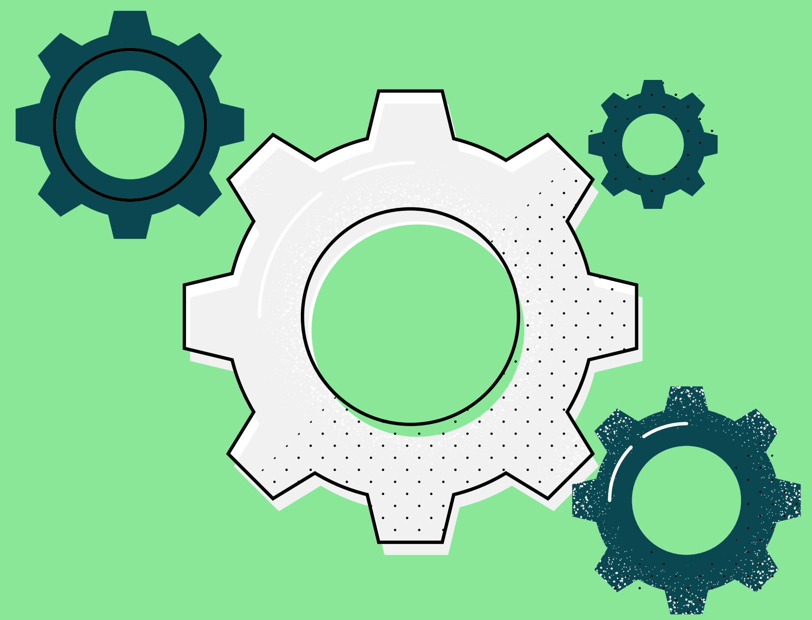
Modeling & Predictions

- Auto-ARIMA model
- Vector AutoRegressive model
- Long Short-Term Memory model

Repeating with smaller dataset

- Auto-ARIMA
- VAR
- LSTM

Granger-Causality Test



Do we have useful data for forecasting?

```
Granger Causality
number of lags (no zero) 253
ssr based F test:         F=1.9598  , p=0.0000  , df_denom=2008, df_num=253
ssr based chi2 test:   chi2=621.0166, p=0.0000  , df=253
likelihood ratio test: chi2=555.0117, p=0.0000  , df=253
parameter F test:         F=1.9598  , p=0.0000  , df_denom=2008, df_num=253
{253: ({'ssr_ftest': (1.9597847571491793, 3.227089671006535e-15, 2008.0, 253),
   'ssr_chi2test': (621.0165548058949, 6.33786496993289e-33, 253),
   'lrtest': (555.0116533643959, 1.1116884052556068e-24, 253),
   'params_ftest': (1.9597847571493563,
    3.2270896709775744e-15,
    2008.0,
    253.0)},
  [<statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7f7be7115d30>,
   <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x7f7be7115be0>,
   array([[0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 0., 0., 0.],
          ...,
          [0., 0., 0., ..., 0., 0., 0.],
          [0., 0., 0., ..., 1., 0., 0.],
          [0., 0., 0., ..., 0., 1., 0.]])])}
```

- What is Granger-Causality Test?
- Null Hypothesis: Time Series A does not Granger-Cause Time Series B

Summary:

We can conclude that knowing the price of SPY is useful for predicting the future prices of stocks: AAPL, AMZN, GOOG, MSFT, QQQ, TSLA, and VIX.

ACF & PACF plots

Find AutoRegressive and/or Moving Average components

What does ACF & PACF plots tell us?

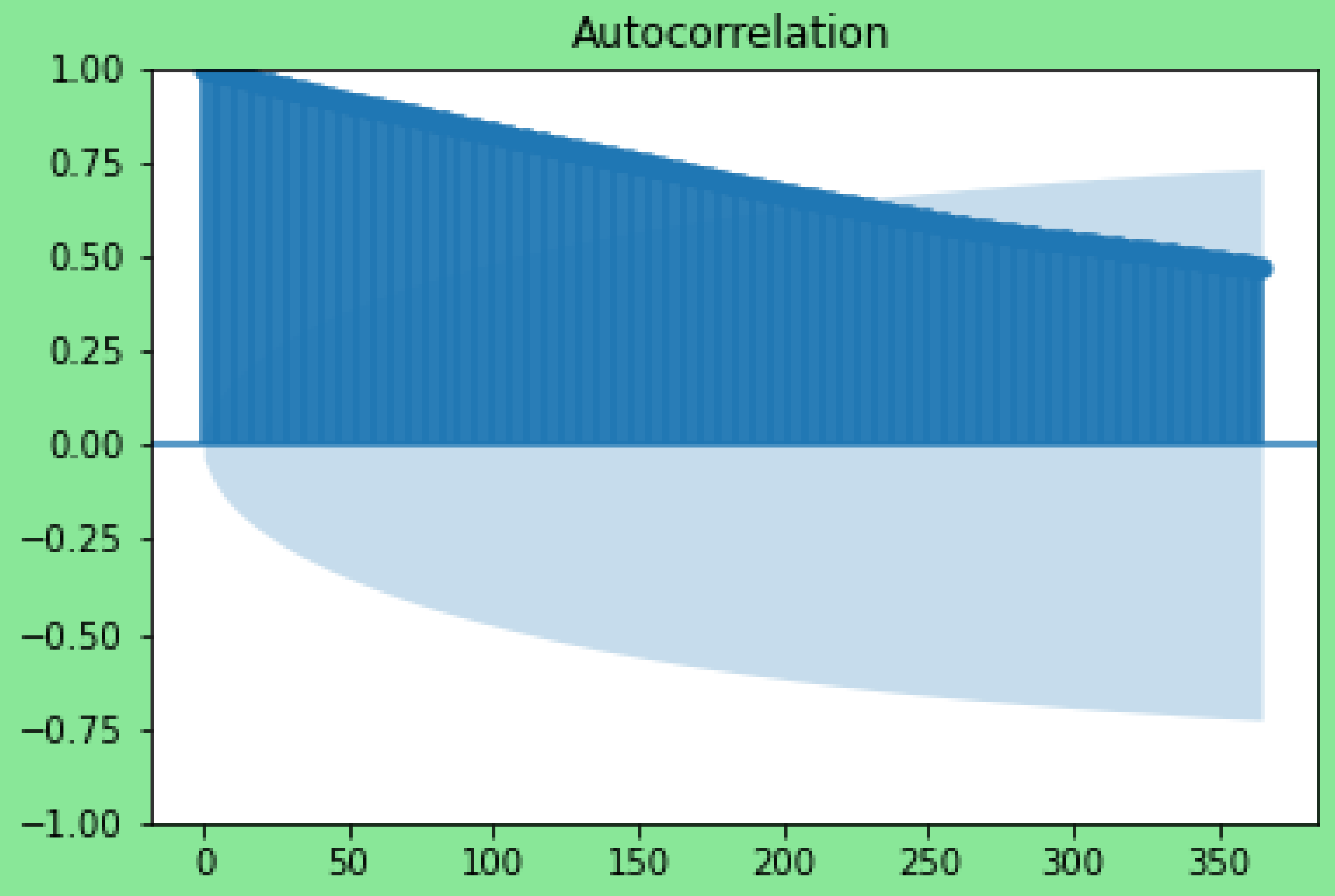
- AutoRegressive (AR) and/or Moving Average (MA) components
- Stationarity
- Lag orders

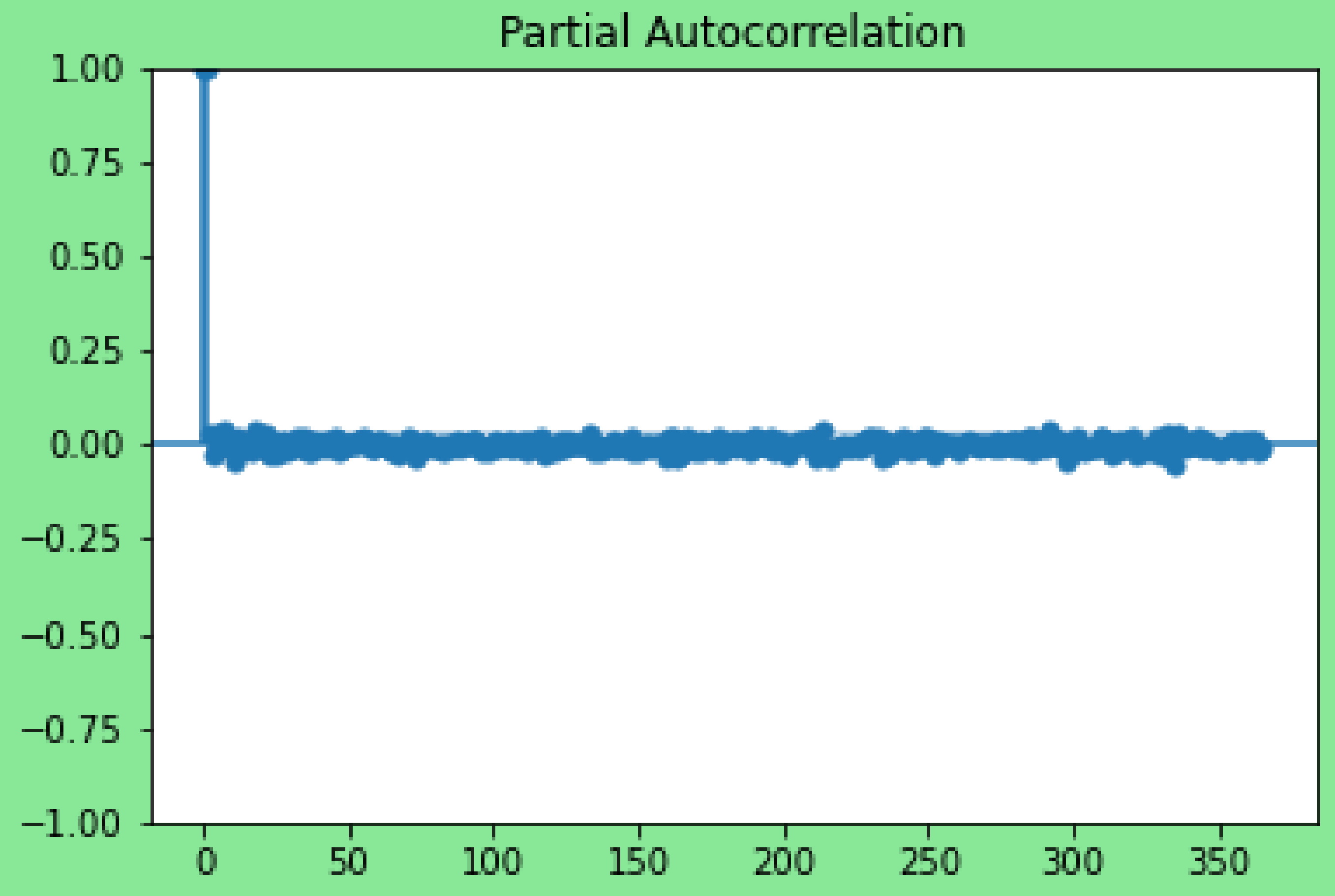
Summary:

Based on ACF & PACF plots of each stocks, adjusted closing prices of AAPL, AMZN, GOOG, MSFT, QQQ, SPY, and TSLA are shown to be non-stationary.

VIX requires more investigation.

Example: ACF & PACF plots of SPY

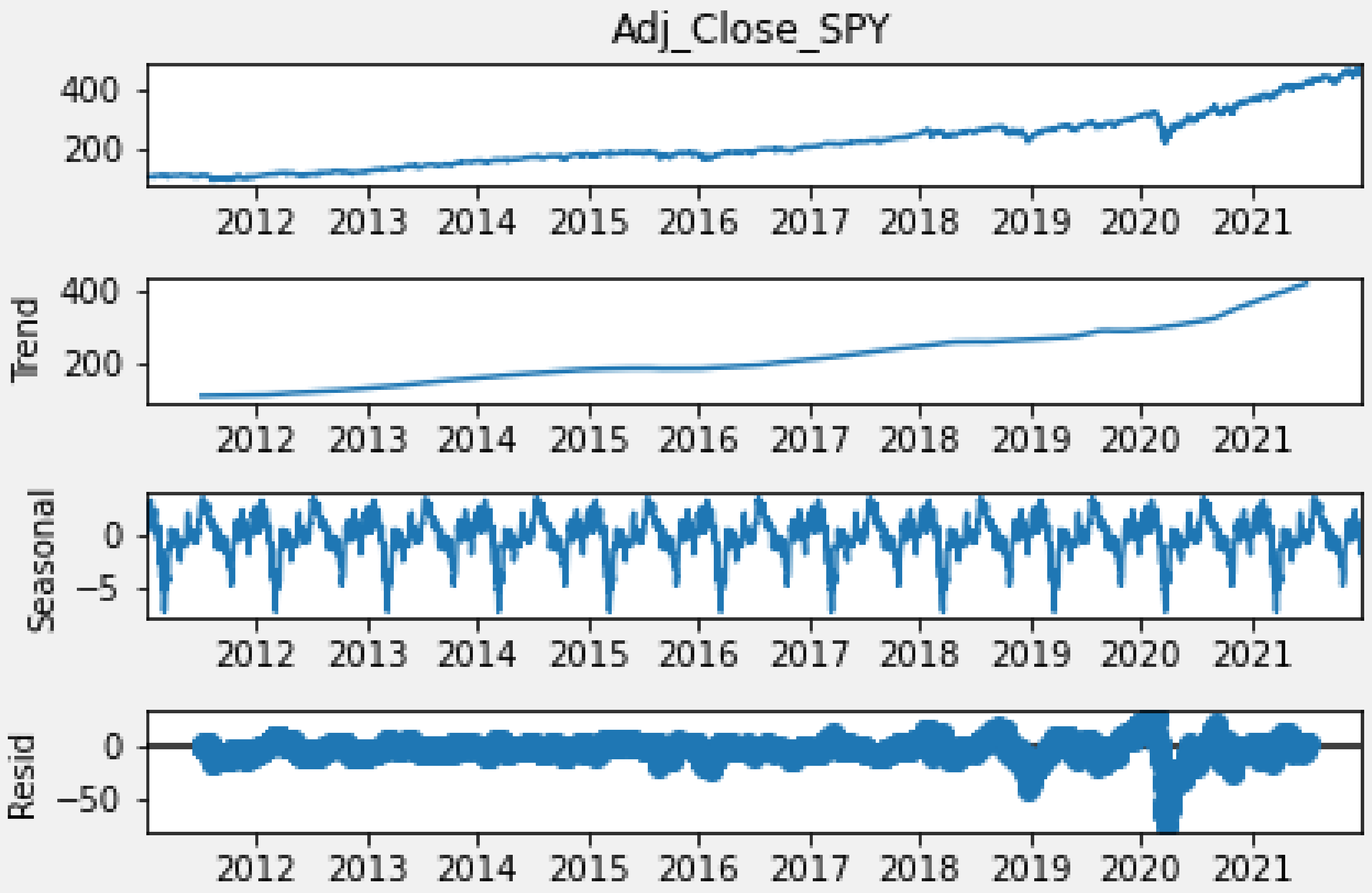




Seasonal Decomposition

Seasonality & Trend

- What is affecting the Time-Series?
- Is there a fixed and known frequency?
- Assumption of Linear trend.



- **Seasonality**: periodic repetition
- **Trend**: time-series behavior over time
- **Residual (Noise)**: Variability in the data unexplained by the model

ADFuller Test & Stationarity

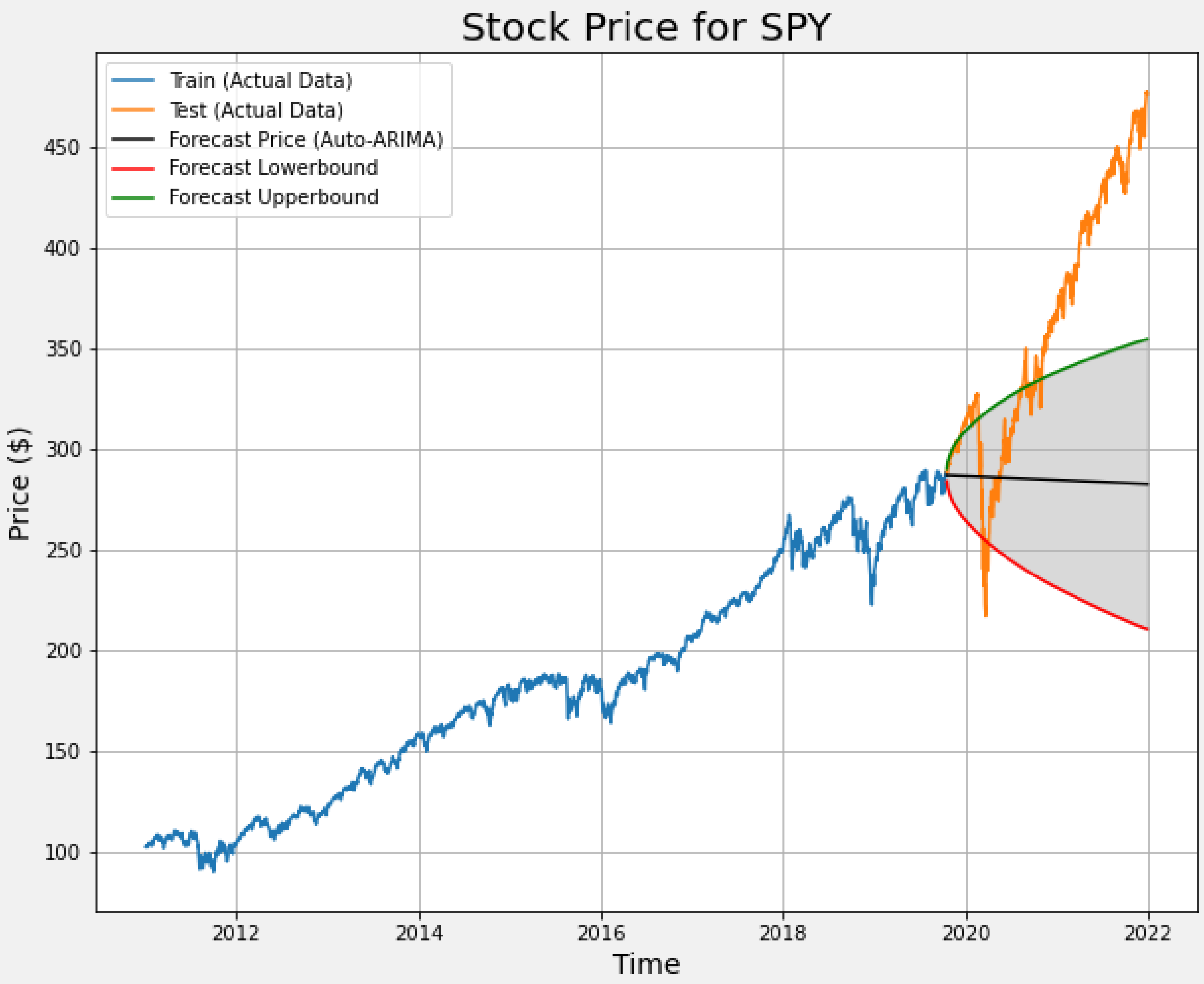
Tests for Stationarity

- Null hypothesis: not stationary
- Alternative hypothesis: stationary
- compare p-value with alpha to determine whether or not to reject the null hypothesis

	Test_Stat_(Adj_close)	p-value_(Adj_close)	Hypothesis_Adj_Close	Test_Stat_(VWAP)	p-value_(VWAP)	Hypothesis_VWAP
AAPL	0.192695	0.971849	Reject	-0.043369	0.954795	Reject
AMZN	0.395822	0.981338	Reject	0.488344	0.984514	Reject
GOOG	0.080608	0.964712	Reject	0.203343	0.972448	Reject
MSFT	2.817629	1.000000	Reject	2.542088	0.999062	Reject
QQQ	0.591425	0.987402	Reject	0.395331	0.981319	Reject
SPY	0.191992	0.971809	Reject	-0.135185	0.945819	Reject
TSLA	-1.689426	0.436569	Reject	-1.592528	0.487346	Reject
VIX	-4.531121	0.000173	Accept	NaN	NaN	NaN

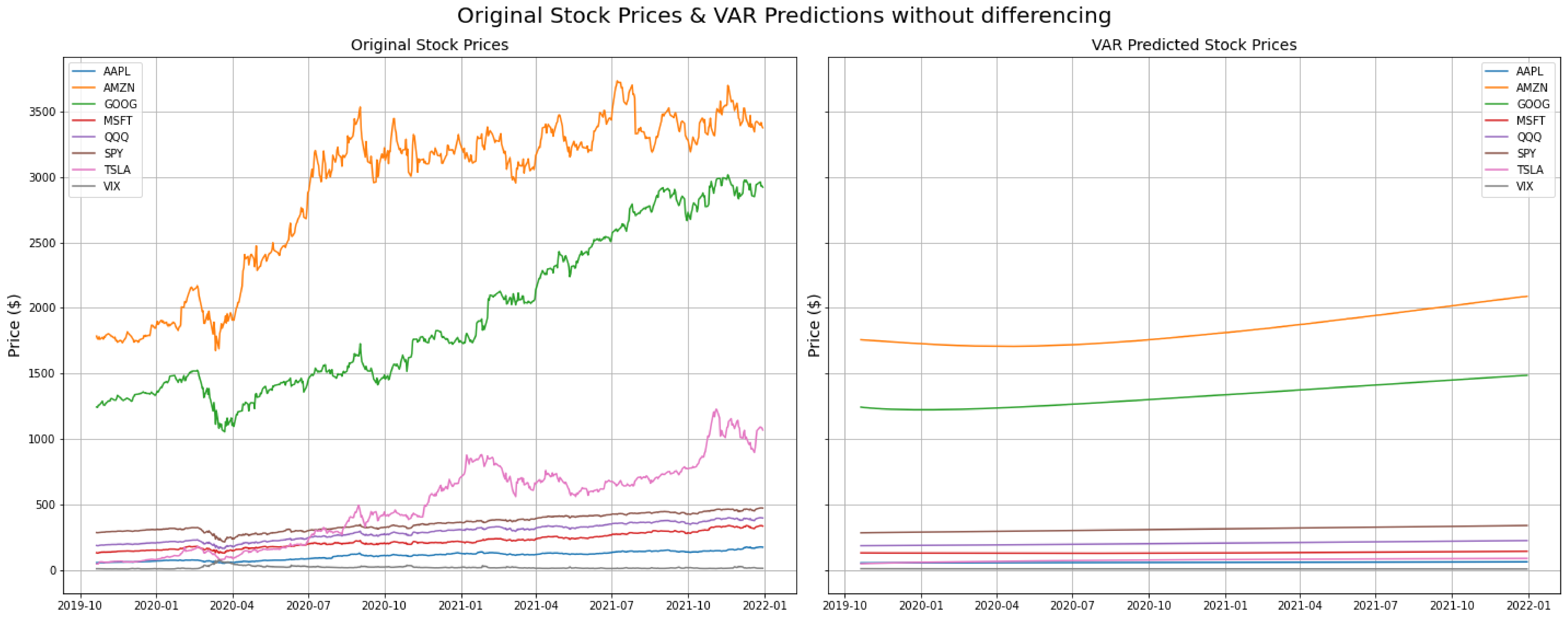
	Test_Stat_(Adj_close)_diff	p-value_(Adj_close)_diff	Hypothesis_Adj_Close_diff	Test_Stat_(VWAP)_diff	p-value_(VWAP)_diff	Hypothesis_VWAP_diff
AAPL	-9.977448	2.158722e-17	Accept	-9.135621	2.941233e-15	Accept
AMZN	-12.271639	8.613920e-23	Accept	-12.163043	1.478570e-22	Accept
GOOG	-11.501078	4.490178e-21	Accept	-10.666556	4.272548e-19	Accept
MSFT	-13.656420	1.550678e-25	Accept	-13.714356	1.220812e-25	Accept
QQQ	-12.945363	3.464890e-24	Accept	-12.943558	3.493685e-24	Accept
SPY	-11.219334	2.031932e-20	Accept	-11.261913	1.614215e-20	Accept
TSLA	-9.863107	4.182127e-17	Accept	-9.921775	2.977837e-17	Accept
VIX	-12.301656	7.426742e-23	Accept	NaN	NaN	NaN

ARIMA

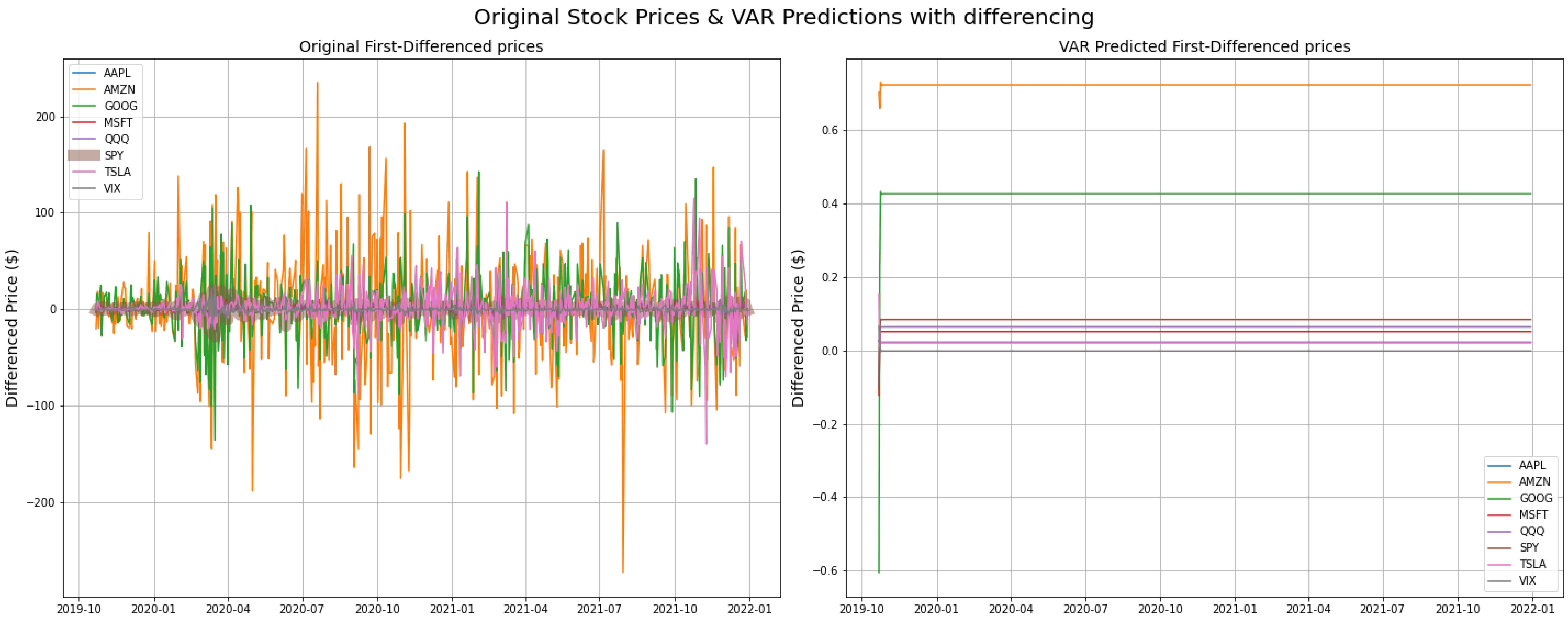


Stock	Model	Note on the result
AAPL	SARIMAX(0,1,0)	Exponential growth above the forecast upper boundary noted. No AR or MA component.
AMZN	SARIMAX(3,1,2)	Exponential growth above the foreaset upper boundary noted.
GOOG	SARIMAX(2,1,2)	Price hovering around the forecast price range. Exponential growth noted after mid 2020.
MSFT	SARIMAX(2,1,2)	Price movement with quite a volatility. Exponential growth noted as well afer mid 2020.
QQQ	SARIMAX(4,1,4)	Prive movement with quite a volatility both above and below forecast price range. Exponential growth noted after mid 2020.
SPY	SARIMAX(1,0,1)	Prive movement with quite a volatility both above and below forecast price range. Exponential growth noted after mid 2020.
TSLA	SARIMAX(0,1,0)	Exponential growth above the forecast upper boundary since late 2019. No AR or MA component.
VIX	SARIMAX(3,0,2)	The only model that has most price covered in the forecast price range. Hugh spikes in 2020-2021 can be explained by the general market tumult due to COVID-19 pandemic related market shifts and policy changes.

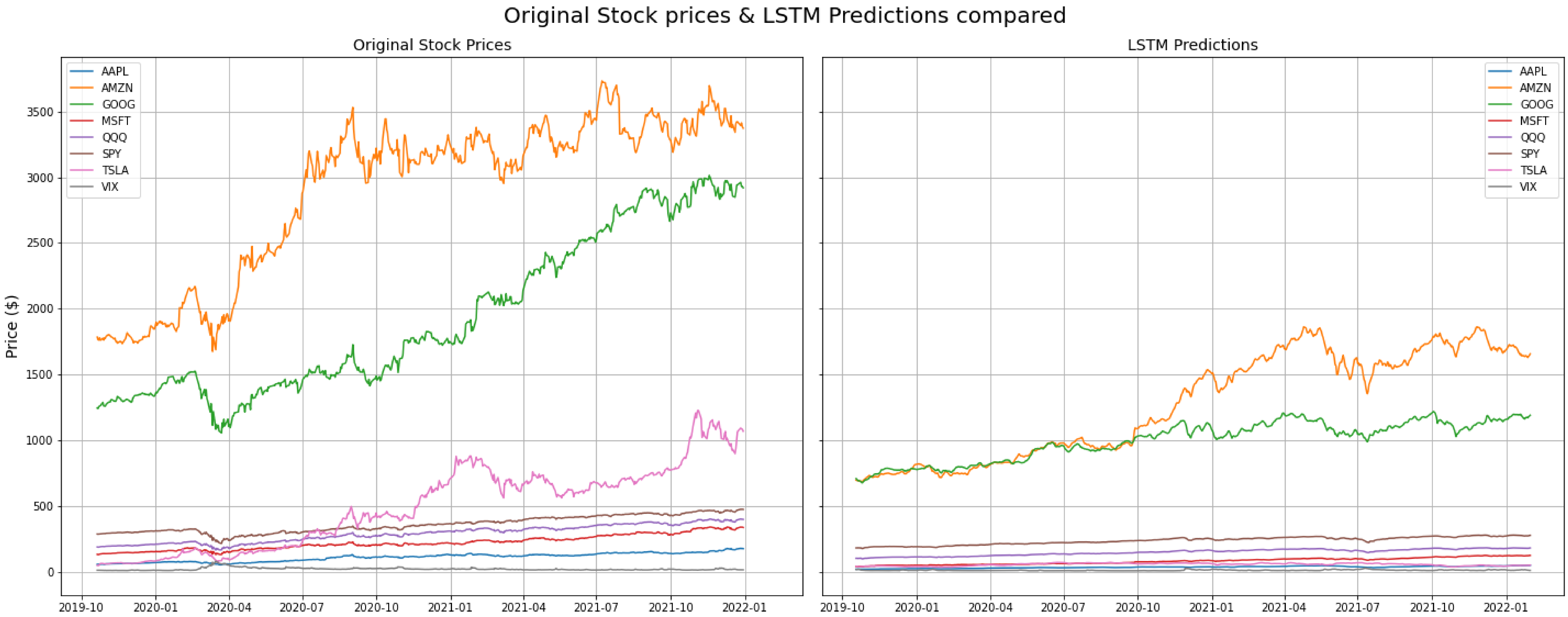
Vector Autoregressive Model



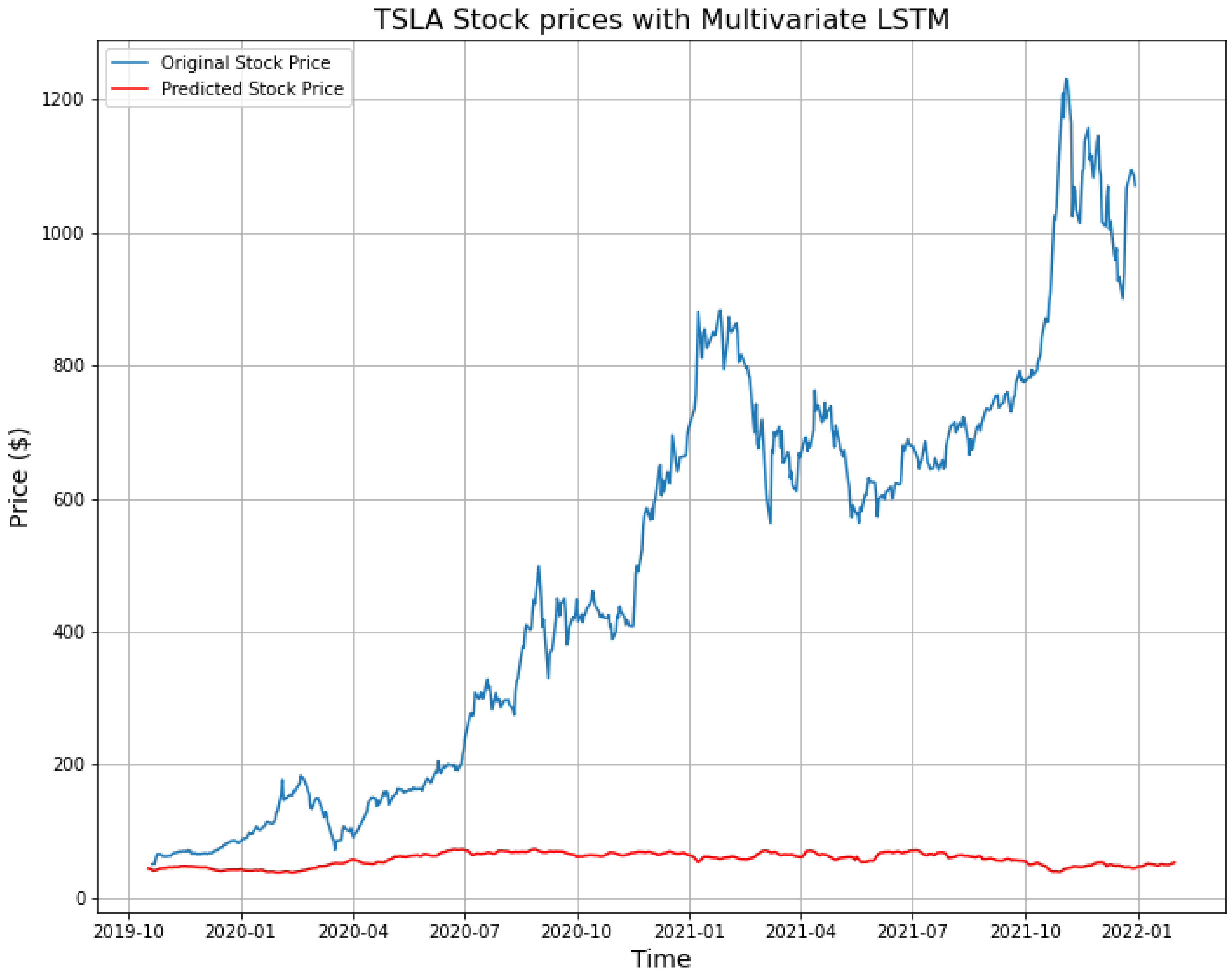
Vector Autoregressive Model (cont)

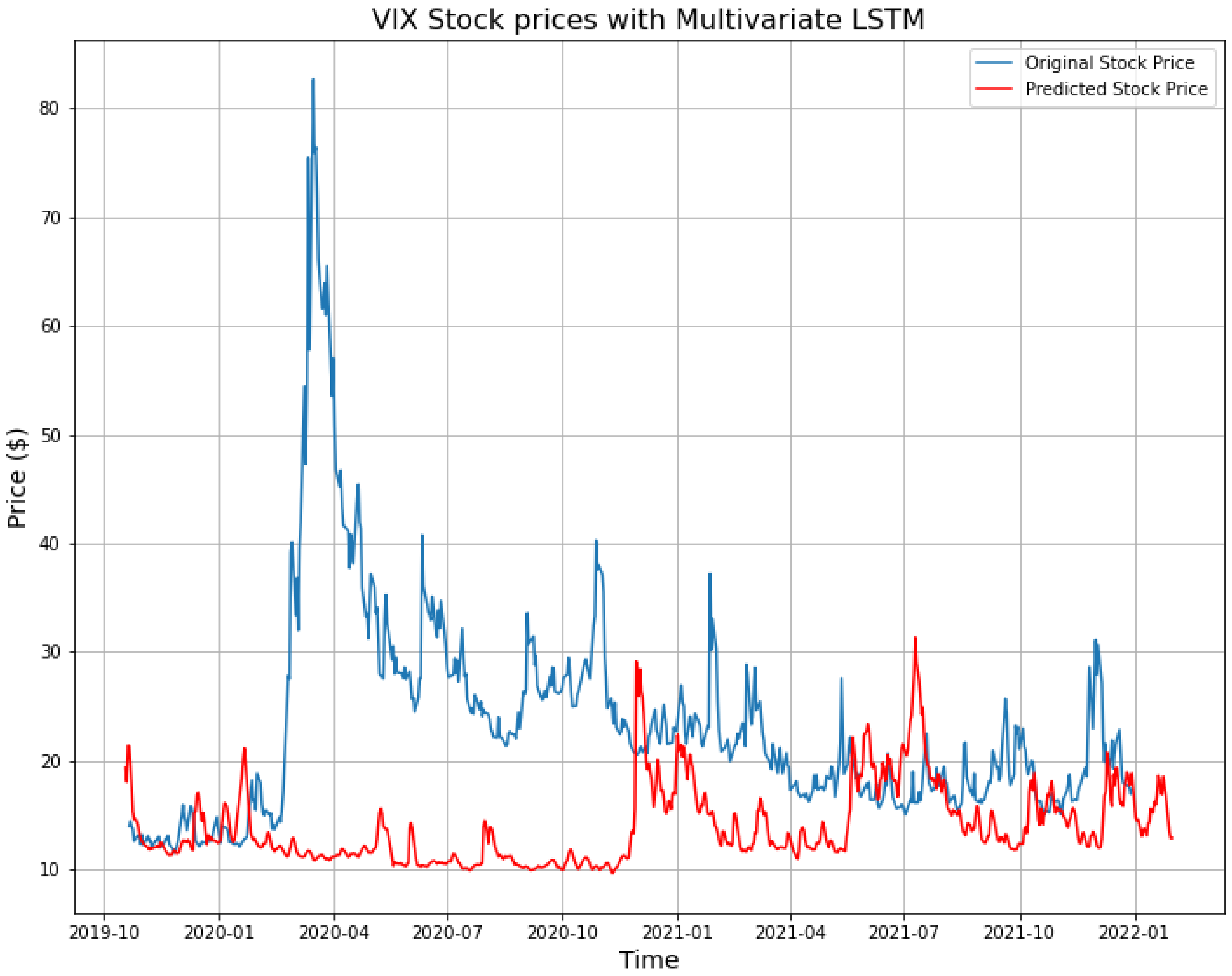


Long Short-Term Model (LSTM)



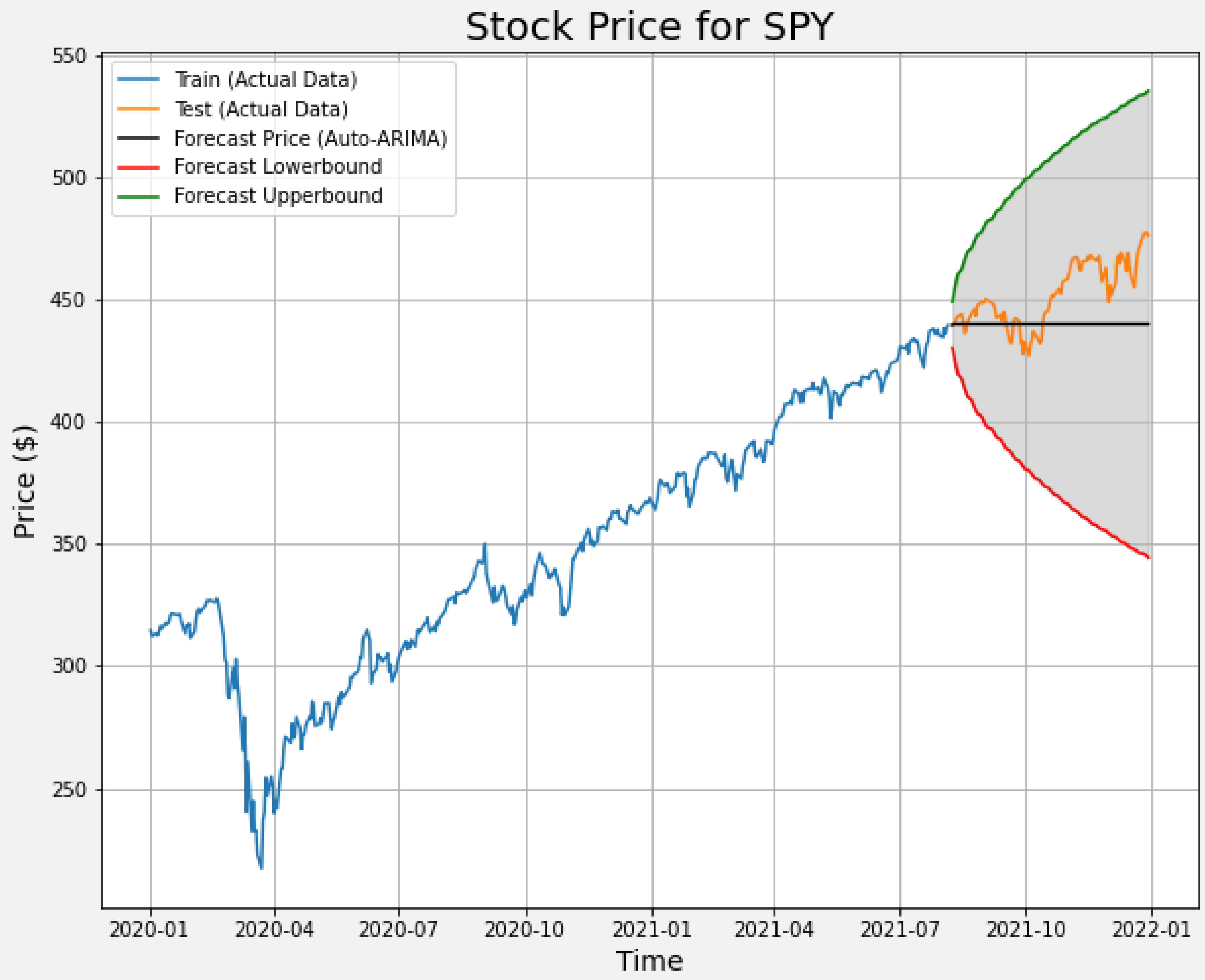
Long Short-Term Model (LSTM)

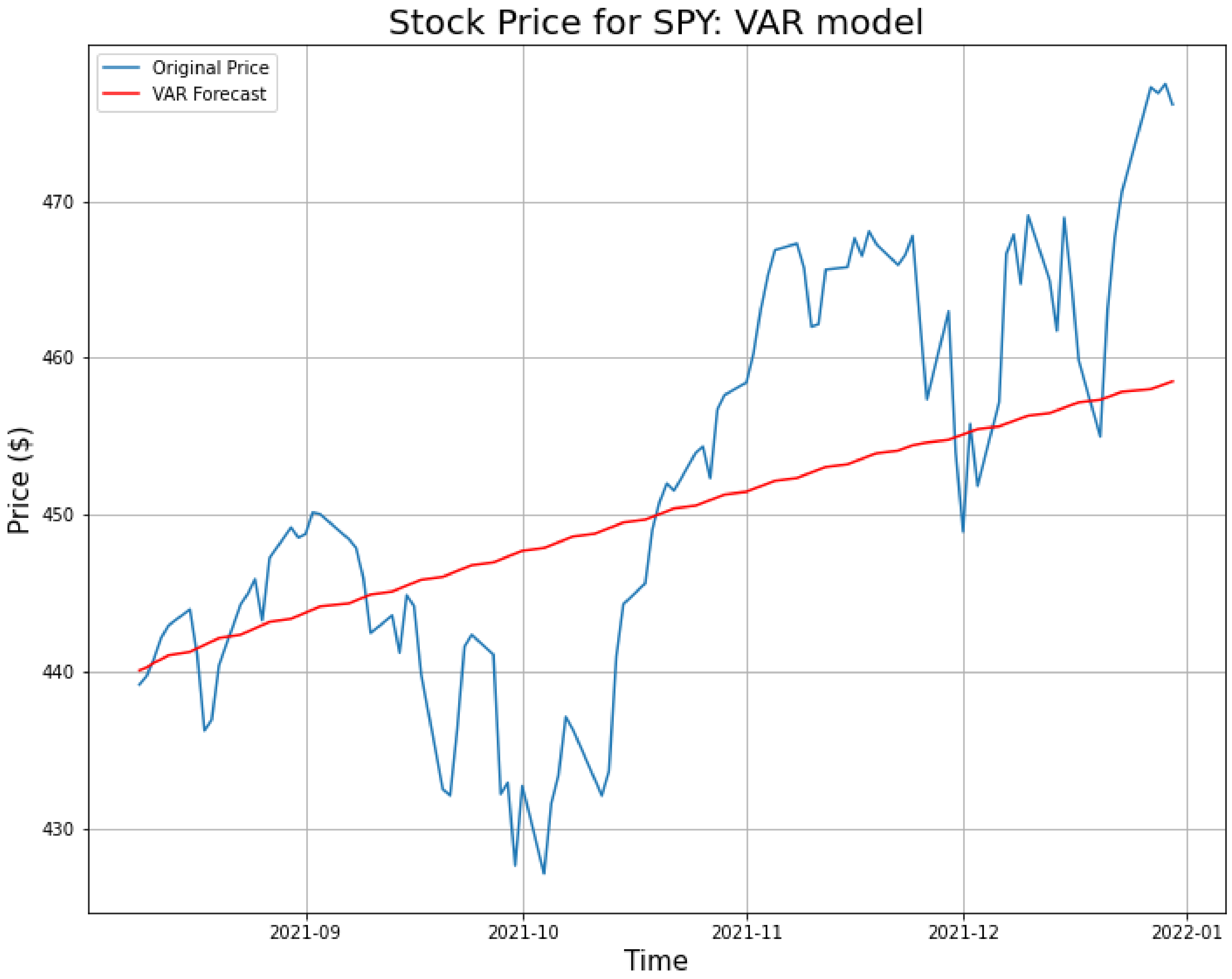


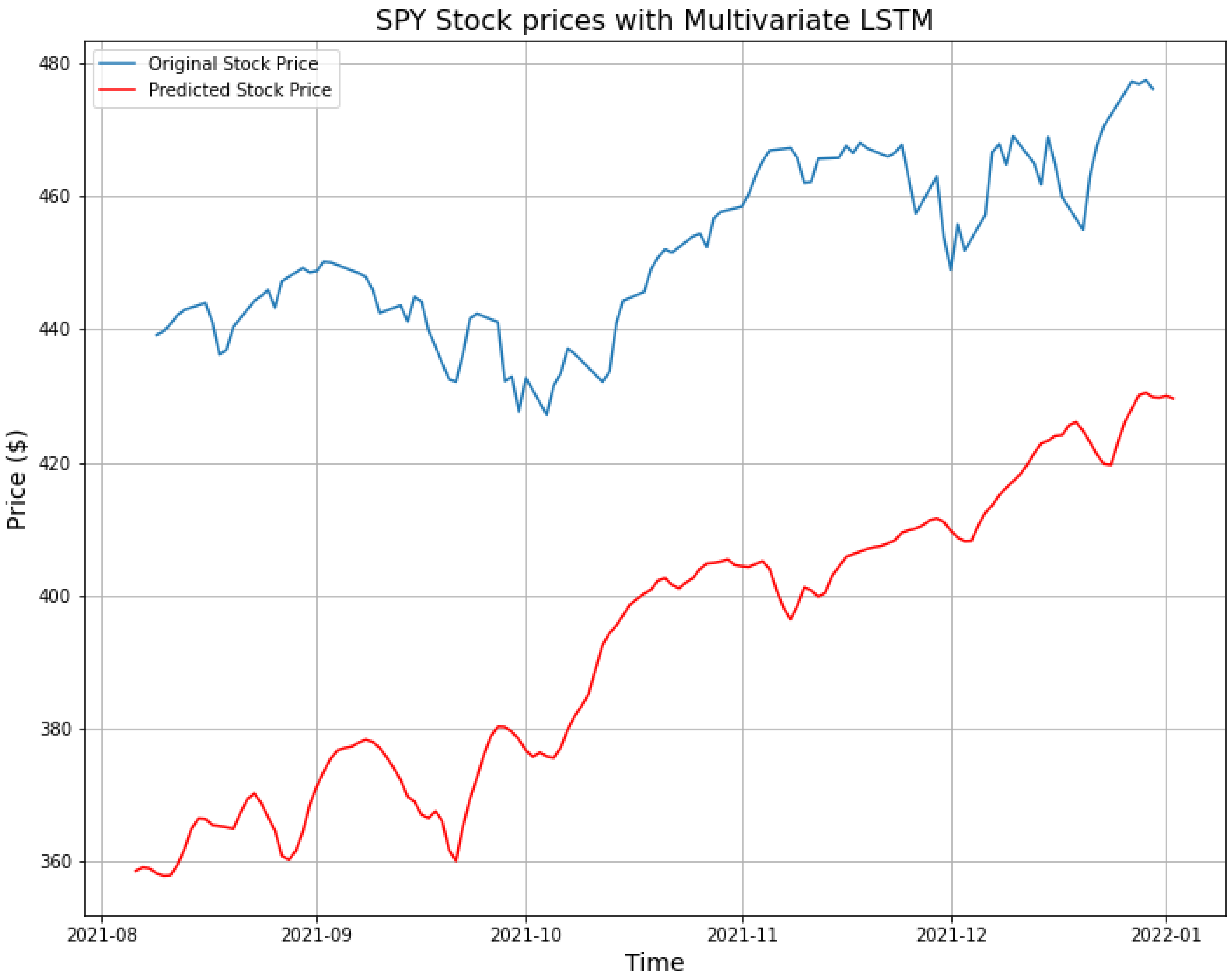


Models with shorter time period "Auto-ARIMA"

Despite not accurate price prediction, forecast boundaries with confidence intervals do cover the actual price movements.







VAR model

Incorrect trend observed
MSE: 94.55513030271806
RMSE: 9.72394623096602

LSTM model

Similar up-trend pattern observed
Price difference is big again.
MSE: 4985.517976039366
RMSE: 70.60820048719117

Discussion of results & Next Steps

Poor performing models

Slightly better performance with smaller data

Incorporate external events for volatility

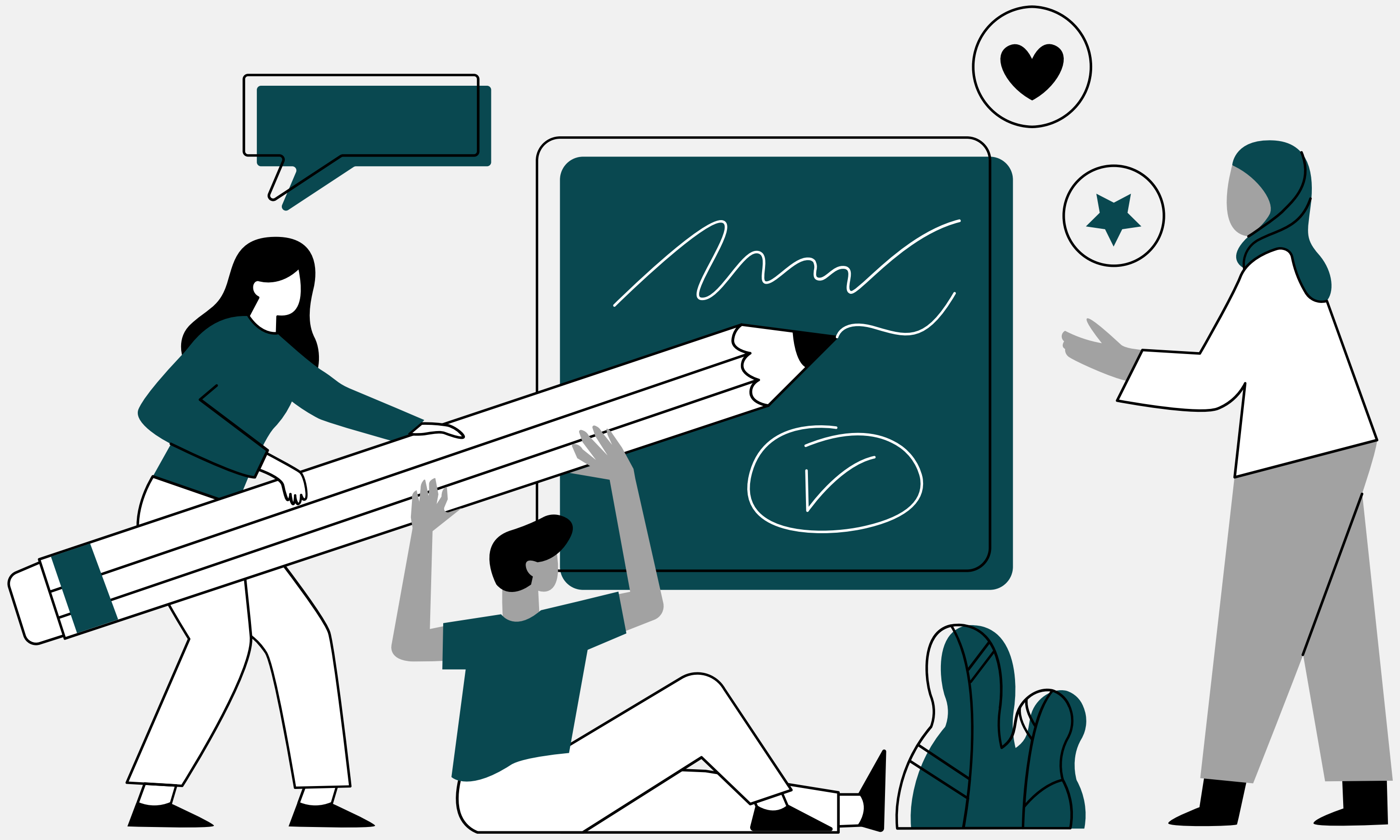
Prediction/Forecast with other available models: FbProphet (Prophet), Kats or Greykite

Prophet (feat. Streamlit)

Showcasing Streamlit App built with Prophet.
(If remaining time is enough)

Do you have any questions?

Contact me anytime!



GitHub Repository

https://github.com/mh0805/Stock-price-prediction-and-forecast-with-Time-series-analysis-and-Machine-Learning

Email Address

mhoon0805@gmail.com

LinkedIn

www.linkedin.com/in/mason-lee85