Time Series Analysis of Dow-Jones

Team 14
Maxime Alos, Yi-Shin Gan, Ramesh Aadhithya, Sanchit Dighe

Summary

As more and more people choose to become amateur investors, it becomes imperative to analyze and identify key strategies that can help us maximize our chances of making a profit.

There is approximately 2800 companies listed on the `New York Stock Exchange (NYSE)`, and with around 1.46 billion shares traded each day, it becomes virtually impossible to evaluate and track the performance of every single stock.[1]

As we intend to try to understand how to optimize stock price predictions and develop trading strategies, we have chosen to analyse the *Dow Jones Industrial Average Index.*

> The Dow Jones Industrial Average (DJIA), also known as the Dow 30, is a stock market index that tracks 30 large, publicly-owned blue-chip companies trading on the New York Stock Exchange (NYSE) and Nasdaq.[2]

We will primarily work with data recorded over 15 years and implement multiple algorithms trying to predict future stock prices, the direction of movement, and the volatility. The models we fit and analyzed are

- ARIMA
- GARCH
- VAR

The results obtained were consistent with our initial intuition as one would expect of any financial dataset. The the data contained both trend and heteroskedasticity, and hence the ARMA-GARCH model performed better than the other models. The VAR model performed well at predicting the volatility range, however, it also couldn't predict the exact value accurately as expected of any reasonable model.

[1] `https://www.advfn.com/nyse/newyorkstockexchange.asp`

[2] `https://www.investopedia.com/terms/d/djia.asp`

TABLE OF CONTENTS

1 INTRODUCTION

1.1 The Stock Market Problem

A stock exchange is a marketplace where investors can buy and sell stocks, or shares, of publicly traded companies. The price of each share is predominantly driven by supply and demand: the more the buyers, the higher the price rises; conversely, the more the number of sellers, the lower the price drops.

Stock markets exist in most countries today, but the first stock market appeared in 17^{th} century Amsterdam. Nowadays, stock exchange generates tremendous revenues and cash flows, and many parties are interested in the market: *companies* raise money by selling their shares which they can reinvest to further their business ventures; *regulating* entities make a business of facilitating the exchange of stocks and on tracking financial information; *and speculators* use the fluctuations on the prices of stocks to buy and resell with a margin of profit.

Because so many actors have a vested interest in the exchange of stocks, it is important to be able to accurately keep track, and even predict stock prices. However, this "simple" task is far more daunting than it seems. As prices are influenced by demand and supply, they can be subject to huge fluctuations in short amounts of time, often in unpredictable ways.

1.2 The Analyst's Initial Intuition

In this paper, we will implement different algorithms to predict future stock prices and we will evaluate the accuracy of those different methods. Given the nature of the financial data series, we expect to see a lot of variation over time, particularly an exponential trend and heteroskedasticity resulting in a non-constant mean and variance. And because the stock prices are subject to strong fluctuations, we expect to find that most predictions will be inaccurate, and that the most interesting metric that can be studied will be the volatility of the data, as being able to predict stock volatility would be a great way of building efficient investment strategies based on optimizing a risk/reward ratio.

1.3 The Dow Jones

To perform our analysis, we will use the Dow Jones Industrial Average Dataset. This dataset contains daily information of the stock prices of 30 large American companies on the Nasdaq and the New York Stock Exchange.

Our version of the Dow Jones dataset is extracted from Yahoo and contains 3855 entries of daily stock price information ranging from January 3rd 2007 to April 25th 2022 (current day). Each day has information on the highest and lowest price reached by the stock during that day, the price of the stock at the times of opening and closing of the market, and the volume, which is the number of shares that were traded during that day. We choose to use the Daily Closing Prices for our primary analysis.

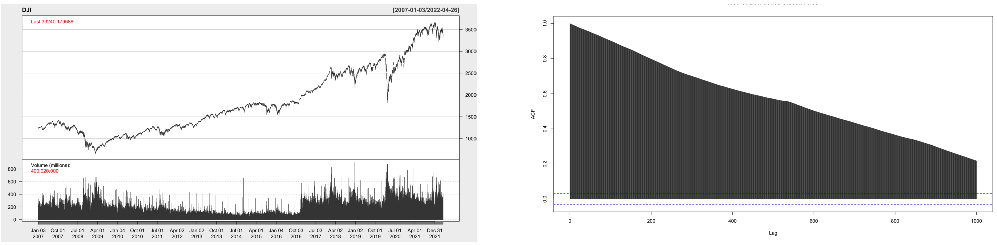
We used the following code to retrieve the data

```
getSymbols("^DJI", src="yahoo")
```

2 EXPERIMENTATION

2.1 Exploratory Data Analysis

We will begin by first exploring our data to identify its main characteristics so that we can better understand it and accurately predict its future values. The first step was to plot the daily stock closing prices and its ACF.



(a) Time Series Plot

(b) ACF

Figure 1: Dow Jones Index - Closing Price

Figure 1a contains the plot of the stock prices at the closing time of the market, coupled with the volume of shares that were exchanged each day below it. Figure 1b is the corresponding ACF plot.

As surmised in our initial intuition 1.2, the original time series plot exhibits an exponential upward trend with no seasonality. The ACF plot with autocorrelation that slowly but steadily decreases with lags confirms the presence of a trend. There appears to be heteroskedasticity in the data as well as the variability of the data seems to be dependent on time. In particular, we can observe high periods of volatility during the 2008 market crash period, and the COVID pandemic period. Since the data series contains trend and heteroskedasticity, it would be ideal to fit an ARMA-GARCH model, however, for the purpose of this study, we'll fit all above mentioned models and compare them.

The trend in the data is further highlighted by the following plot corresponding to the fit of splines regression model.

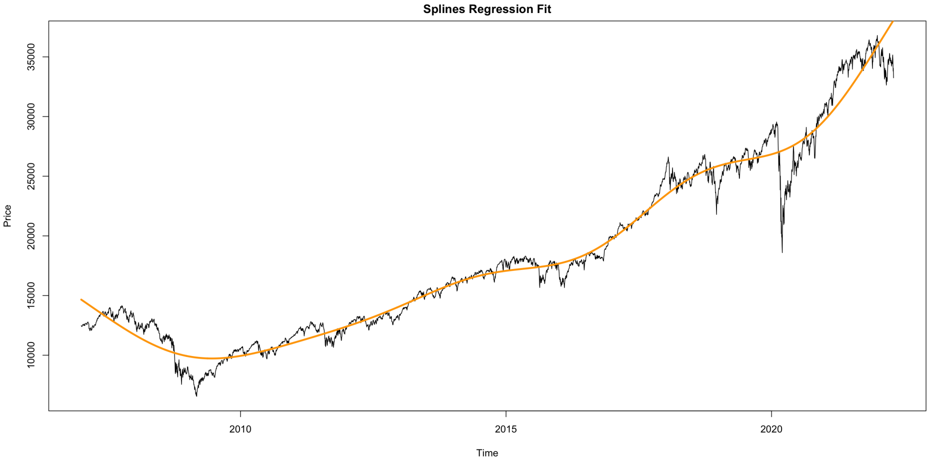
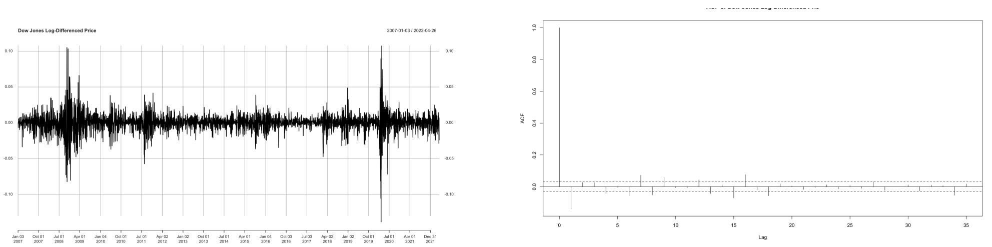


Figure 2: Splines Trend Fitting on Dow Jones

This graphical analysis indicates that the Dow Jones dataset is heteroskedastic and has a clear upward trend but shows no sign of seasonality. Another interesting point in figure 1a is that there is a sharp increase in the volume of exchanges around 2017 that seems to coincide with a faster increase in stock prices as well as increase in stock price variability.

Looking at the differenced time series data, we observe:



(a) Time Series Plot

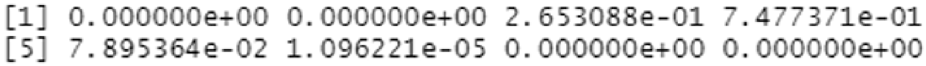
(b) ACF

Figure 3: Dow Jones Index - Log-Differenced Closing Price

The differenced time series is centered around zero, having a constant mean and the ACF mainly has small autocorrelations that are mostly within the confidence bands. Under those conditions we can consider the differenced data to be stationary but with heteroskedasticity.

2.2 The ARIMA Model

The first model we tried to fit to our data to predict future values is ARIMA. We chose the order of our ARIMA model by performing a grid search and minimizing the AIC values for model fits. We have found the best fit to be an ARIMA(6,1,2) model. We looked at the p-values for this ARIMA model and found them to be very close to zero, showing that all coefficients are statistically significant.



```
[1] 0.000000e+00 0.000000e+00 2.653088e-01 7.477371e-01
[5] 7.895364e-02 1.096221e-05 0.000000e+00 0.000000e+00
```

Figure 4: ARIMA Residual p-Values

In our first experiment, we trained the ARIMA model on rolling windows of 30 days to predict the value of the upcoming day. The resulting plot is in figure 5.

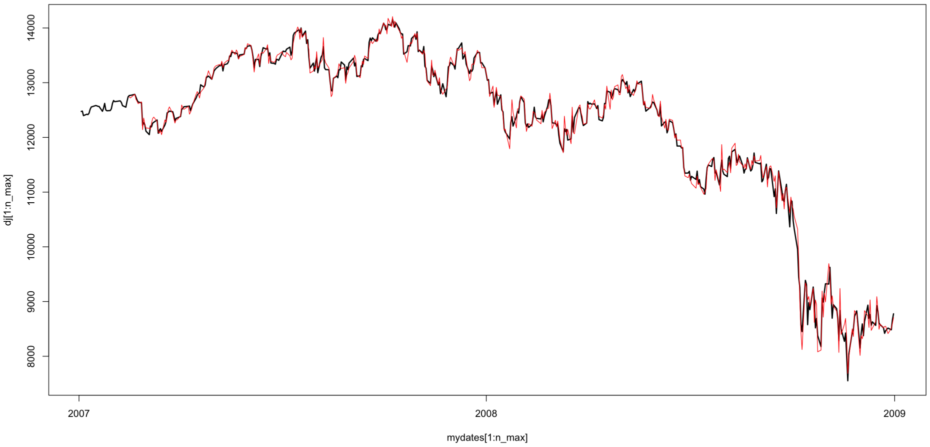
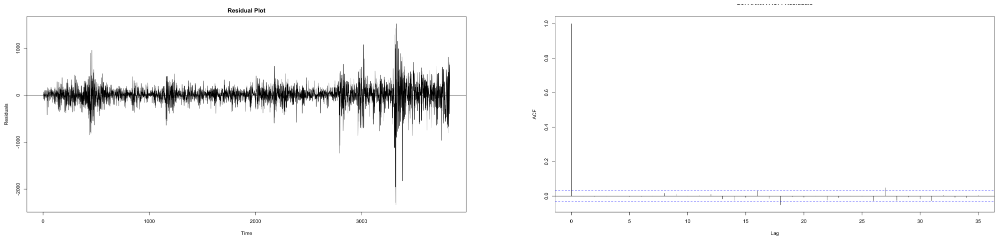


Figure 5: ARIMA 30-Day Rolling Prediction

We clearly see that the predictions are largely aligned with the corresponding expected value for the data. This precision is confirmed by an extremely low MAPE of 0.00732233.

Residual Analysis
The residual plots and thier ACF are shown in figure 6.



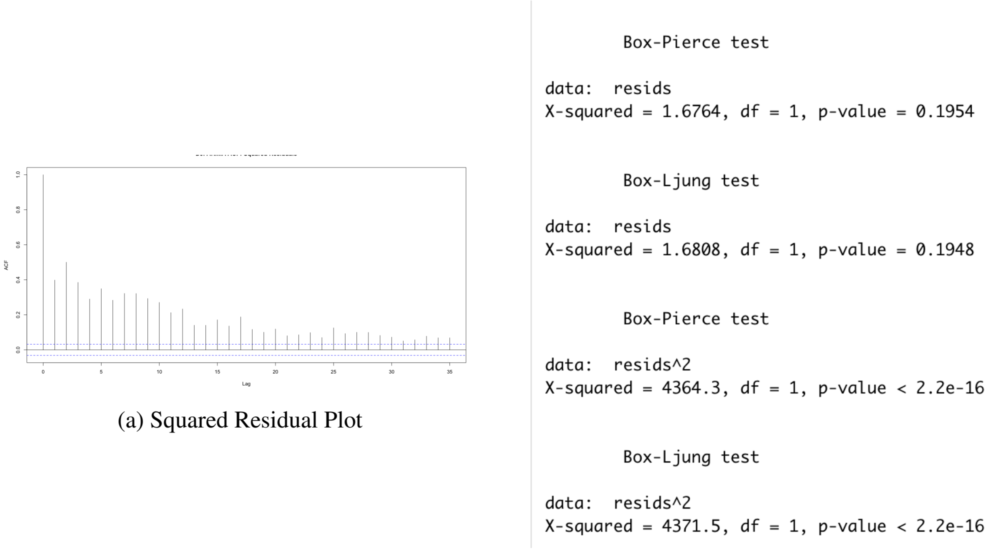
(a) Residual Plot

(b) ACF of Residuals

Figure 6: Dow Jones Index - ARIMA Model

At first glance, the residuals look stationary as the plot shows a constant variance, and the ACF looks like that of a White Noise process.

The ACF of the squared residuals and the Box-Tests of the residuals and the squared residuals are shown in figure 7.



(a) Squared Residual Plot

```
        Box-Pierce test

data:  resids
X-squared = 1.6764, df = 1, p-value = 0.1954


        Box-Ljung test

data:  resids
X-squared = 1.6808, df = 1, p-value = 0.1948


        Box-Pierce test

data:  resids^2
X-squared = 4364.3, df = 1, p-value < 2.2e-16


        Box-Ljung test

data:  resids^2
X-squared = 4371.5, df = 1, p-value < 2.2e-16
```

(b) ARIMA Residual Box-Tests

Figure 7: Dow Jones Index - ARIMA Model Residual Analysis

By looking at the ACF of the squared residuals, we can observe that they are serially correlated. The Box-Tests confirm our hypothesis, as we obtain an extremely small p-value, hence rejecting the null hypothesis of no serial correlation.

To conclude, the ARIMA model produces residuals that are uncorrelated, but not independent violating our assumption of i.i.d. residuals.

In our second experiment we predicted stock prices 25 days ahead based on the ARIMA model. However, this experiment yielded very poor results as can be seen in figure 8. While the true or expected values are within the 95% confidence interval of the predictions, we notice that ARIMA performs poorly in anticipating the variations in the data. The predicted values are mostly equal to the last observed value.

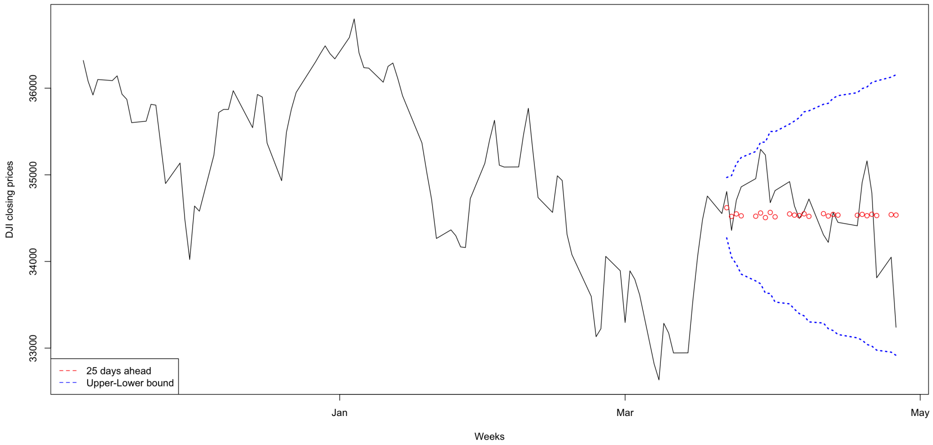


Figure 8: ARIMA Forecast

This second experiment shows the limitations of ARIMA based predictions. While ARIMA performs well for near-future predictions, on datasets like stock prices that show high levels of volatility, the results of ARIMA are unreliable.

Stock prices often show random variations, induced by the demands of the market that cannot be determined easily from past patterns, which ARIMA relies upon. For this reason, instead of trying to forecast the baseline price of a stock, it might be interesting to instead evaluate the volatility of the dataset, which would allow traders to accurately determine the risks of investing in a specific stock.

2.3 The GARCH Model

The next model we tried to implement was ARMA-GARCH. The best fit for ARMA was ARMA(6,2) and after refining our orders for GARCH, we found the best model to be GARCH(2,0). For that fit, the Ljung- Box test returns p-values almost equal to 0. We therefore know that there is serial correlation. Here is the plot of the squared residuals' ACF:

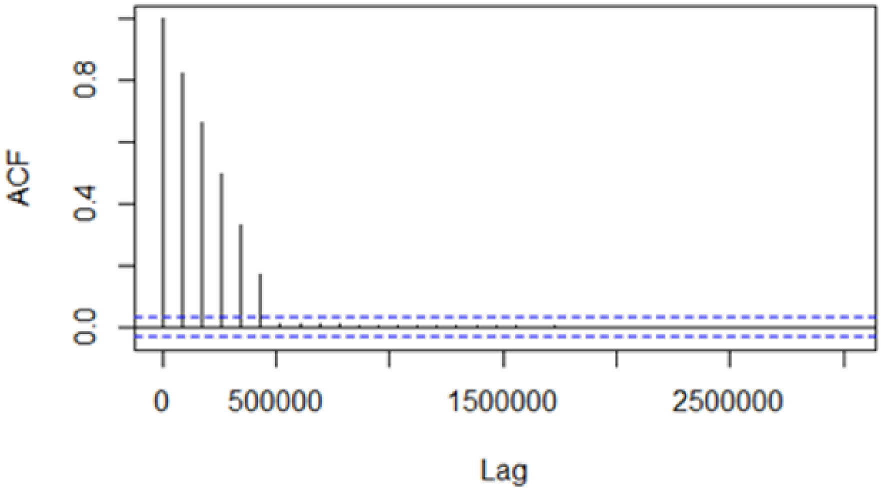


Figure 9: ACF of Squared Residuals

The fact that the first 6 spikes are significant corresponds to the selection of the ARIMA orders. Like we did with ARIMA, we will first perform predictions using a training a rolling window in order to predict the upcoming day, and then try a prediction on a test dataset of 25 days. Here is a plot of GARCH predictions trained on a rolling window to predict future values one by one:

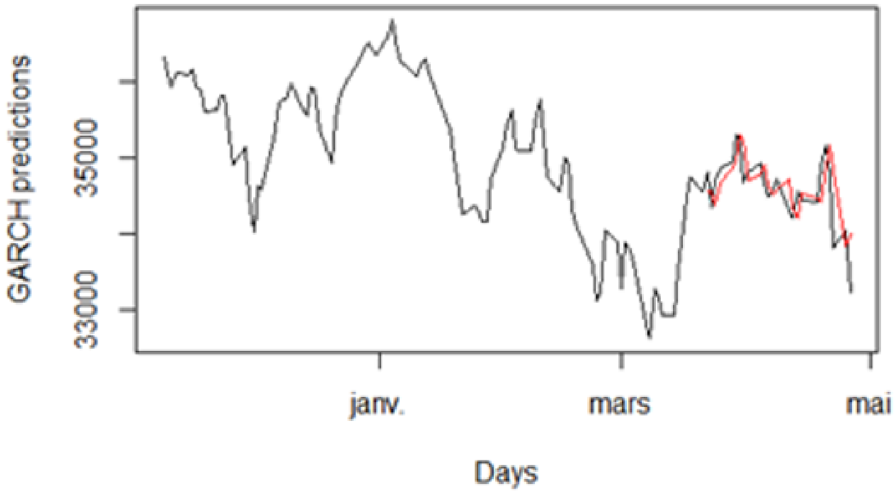


Figure 10: GARCH Rolling Window Prediction

It can be seen that predictions are accurate but somewhat lagged. The obtained MAPE is lower than for ARIMA. However, the real gain of GARCH implementation is the prediction of the volatility of the time series. This becomes apparent when looking at our second implementation of GARCH to predict stock price data 25 days ahead:

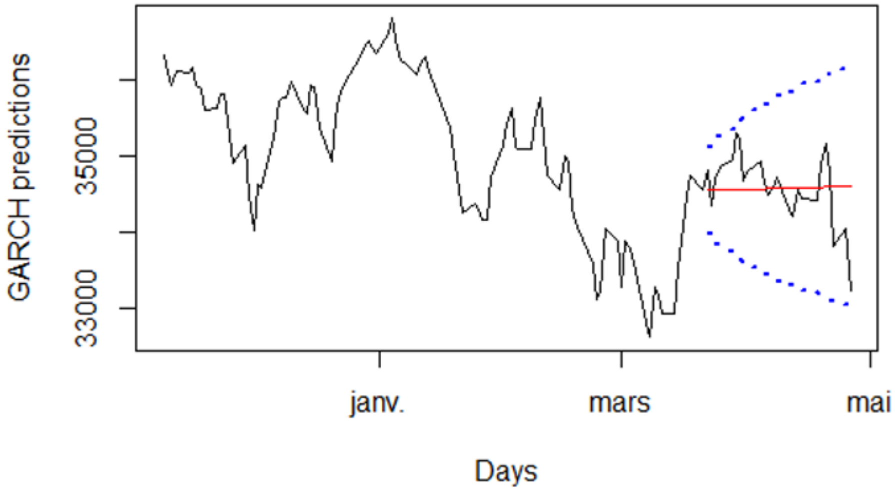


Figure 11: GARCH Prediction 25 Days Ahead

While the value prediction is constant and very poor, **it is interesting to note the results of the upper and lower bounds appearing in blue.** They represent the 95% confidence of GARCH on the values it predicts. It can be noted that all real values of the dataset are indeed within that confidence interval for all 25 days predicted. Because it is too difficult to directly predict future stock prices, a strong alternative that GARCH allows is to give confidence intervals on its prediction, which allows investors to estimate the risks of investing in certain stocks.

2.4 The VAR Model

In this section, we are going to perform multivariate time series analysis. We will use quarterly time series data started from 2007/01/01 to 2021/04/01. The time series data chosen were the Consumer Price Index(CPI), the Crude Oil Price(WTI), and the Unemployment Rate (UER). We will work on return price data instead of original data.

Here is the plot of 5 time series (return price data):

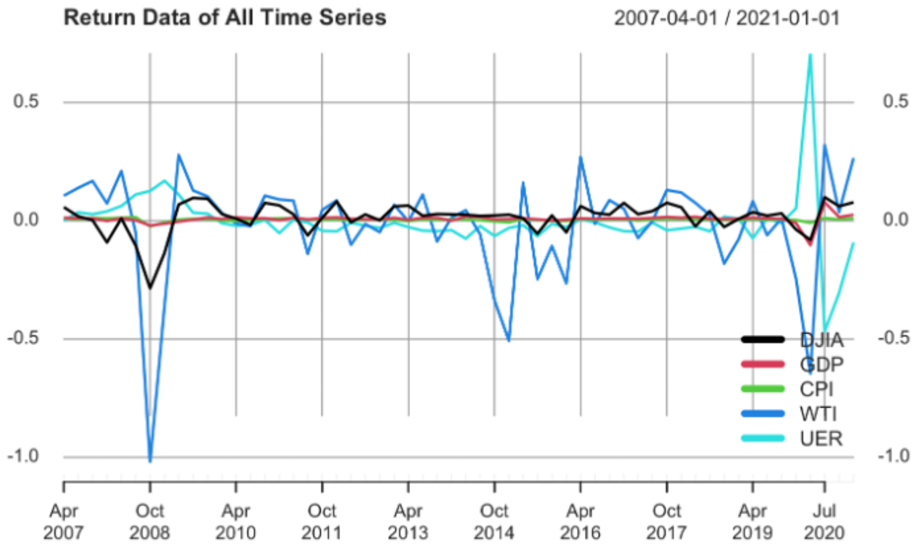


Figure 12: Retrun Price for All Time Series

Since GDP and CPI are in smaller scale, we plot these two time series along with the DJIA in a separate figure:

Table 1: Co-Integration

	Time Series			
	GDP	CPI	WTI	UER
None	**-1.9499*(5%)**	**-2.7319*(5%)**	-1.327	0.5809
Trend	-1.5006	-2.9055	-3.1073	-0.7904
Drift	-1.9383	-2.7392	-1.2566	0.8128

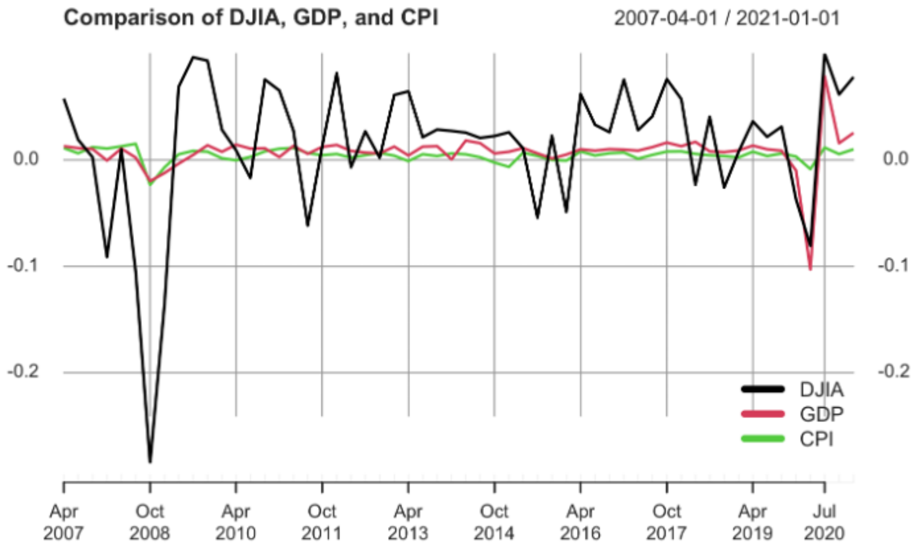


Figure 13: Retrun Price for GDP & CPI

We can see that DJIA and the crude oil price (WTI) follow certain similar patterns, but crude oil price seems to have larger volatility. The unemployment rate acts inversely compared to DJIA. GDP and CPI show very similar drops and growth during 2008, economic crisis, and 2020 2021, the pandemic year.

Co-integration

We evaluate whether there is co-integration between the DJIA and the other four time series. We use the original aggregated quarterly data instead of return price data. Linear regression will be applied first, following we will use ur.df() command to test the co-integration after the trend, drift, or none is removed.

The table is the result of co-integration test under trend, drift, or none is removed. We marked the significant result in bold text and also show the confidence level.

We use two linear models which gdp and cpi are regressors to fit the data and see the prediction power of each. We split the data into a training set with 54 data points and the remaining 4 data points for the testing set.

Regressor: GDP

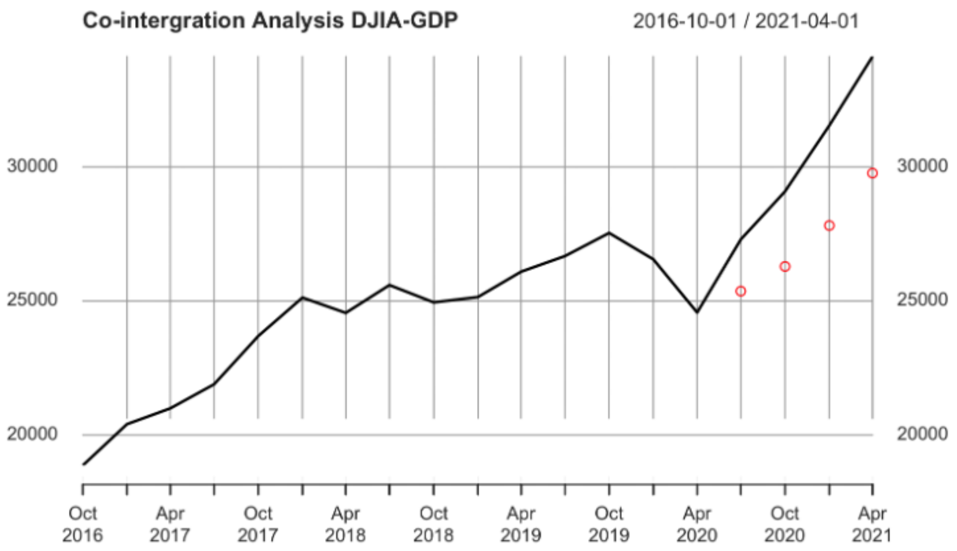


Figure 14: Co-Integration Analysis

MAPE = 0.1032367
PM = 1.681177

Regressor: CPI

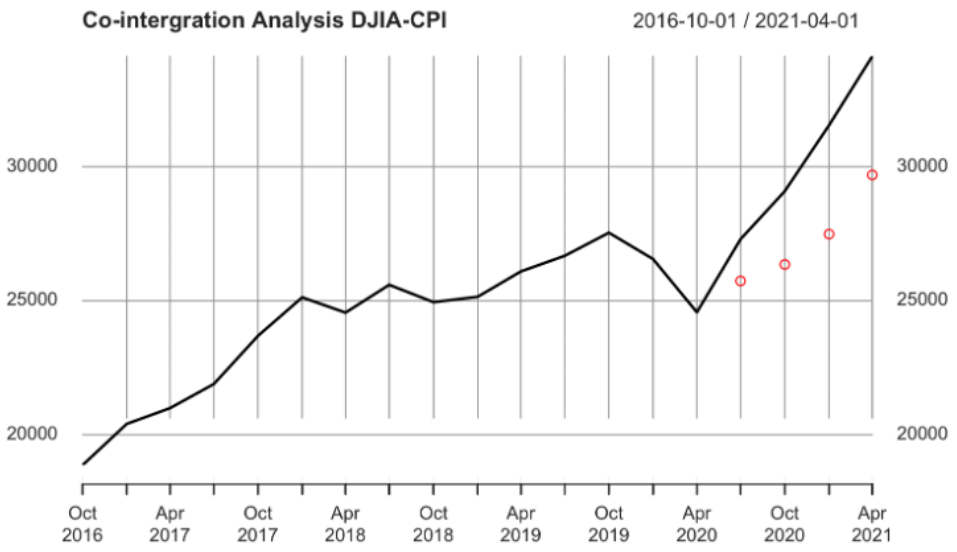


Figure 15: Co-Integration Analysis - CPI

MAPE = 0.1024536
PM = 1.741222

We try to evaluate the correlation between DJIA and 4 other time series data. In this section, we will work on return price data. First, we use VARselect() command to select the lag with lowest AIC value. Following, we perform the rolling prediction with restricted model that is selected above to predict 4 data points. The results are shown in table 3

Table 2: Co-Integration Analysis

	Original	lm.gdp	lm.cpi
2020/07/01	27299.04	25368.26	25739.61
2020/10/01	29091.59	26289.16	26353.69
2021/01/01	31550.58	27816.89	27487.24
2021/04/01	34121.48	29769.33	29692.87

Table 3: Predictions

	Original	restrict(var.aic)
2020/07/01	-0.08073211	-0.34912713
2020/10/01	0.09993783	0.07795508
2021/01/01	0.06161762	0.01334224
2021/04/01	0.07793815	0.01334224

MAPE = 1.289189
PM = 3.936347

**We would like to note that while performing the VARselect(), the results return –Inf for AIC value where lag is greater than 7. The –Inf AIC indicate the model is over fitting, which is not appropriate to use for prediction.*

3 CONCLUSION

After our analysis, we understand that while investors would ideally like to have an accurate prediction of a stock's value for the foreseeable future, that is unrealistic. Because of the complex interactions of the stock exchange markets with traders, that are influenced by the economical context of the moment, stock prices are highly volatile. In the short-term especially they can fluctuate and show periods of decreases and increases in quick successions. For those reasons, predictive models that base themselves on the past like ARIMA, and even models such as VAR which use extraneous variables to complement its analysis cannot accurately predict future evolutions of a stock's price. Because of that, the best alternative remains to use models such as GARCH which predict volatility to estimate risks associated with a specific stock. Being able to efficiently manage the risks of investing in a stock and balance those risks with potential rewards are at the core of modern investment strategies and volatility prediction is an excellent way of giving indications of those risks.

References

References follow the acknowledgments. Use unnumbered first-level heading for the references. Any choice of citation style is acceptable as long as you are consistent. It is permissible to reduce the font size to `small` (9 point) when listing the references. Note that the Reference section does not count towards the page limit.

[1] Consumer Price Index https://fred.stlouisfed.org/series/CPIAUCSL

[2] Crude Oil Prices – West Texas Intermediate https://fred.stlouisfed.org/series/MCOILWTICO

[3] Unemployment Rate https://fred.stlouisfed.org/series/UNRATE

[4] GDP Class Material

[5] Dow Jones Industrial Average Yahoo resource

[6] https://www.forecast-chart.com/dow-links2.html

[7] https://www.longtermtrends.net

[8] https://www.investopedia.com/terms/d/djia.asp

[9] https://www.advfn.com/nyse/newyorkstockexchange.asp