Dogecoin to the ground*

A forecasting of the Dogecoin price

Siming Shan

27 April 2022

Abstract

One of the world's most influential men, Elon Musk, has a huge impact not only on technology but also on the financial markets. In the last year, Elon Musk purchased a large amount of Dogecoin (a cryptocurrency), and highly recommended it on his twitter, saying that it is the people's cryptocurrency. As a result, Dogecoin's price soared dramatically in a short period of time. One year later, is Dogecoin still worth investing in? This paper focuses on developing an ARIMA model to fit Dogecoin price data, and then predicting price for the next 10 days. In our analysis, we found that Dogecoin's price fits into an ARIMA(1,1,1) model, and we predicted that Dogecoin's price would continue decreasing for the next ten days using this model.

keywords: Elon Musk, Bogecoin, ARIMA, Finance, Prediction, Cryptocurrency

Contents

*Code and data supporting this analysis is available at: https://github.com/SimingShan/Dogecoin-price-prediction.git

1 Introduction:

Tesla Motors CEO Elon Musk is known as one of the richest men in the world. In March 14th, Elon Musk proposed to buy Twitter for 43 billion dollars, which brought him to the attention of the public. Nonetheless, Elon Musk bought 9% of Twitter and this news caused the twitter shares to jump more than 20% (Morris 2022). For Elon Musk, this isn’t the first time he has had an impact on the finance market.

Elon Musk advertised a crypto currency he bought last year - Dogecoin - a lot, making it a big year for crypto currency. Developed by Billy Markus and Jackson Palmer (Frankenfield 2022b), Dogecoin is an open-source cryptocurrency. The cryptocurrency is a digital or virtual currency which is secured by crypotography(Frankenfield 2022a). Since Satoshi Nakamoto created the first cryptocurrency bitcoin in the world in 2018, decentralized networks have been gaining more and more traction around the globe. Cryptocurrency investment has now become the first choice of many investors, since it is secure, decentralized, and stable.

Musk tweeted that “dogecoin is the people’s crypto” on May 10th 2021, and claimed SpaceX will launch a satellite called “Doge-1” to the moon, thus “Doge to the moon” has become a popular meme in the crypto currency community. In fact, the doge coin did soar after Elon Musk’s tweets. After dogecoin’s wave passed, and Elon Musk stopped tweeting about dogecoin, the price of dogecoin fell rapidly. Is dogecoin still worth investing in? The study uses the statistical programming language R(R Core Team 2022) and focuses on using the auto-regressive integrated moving average(ARIMA) model to model the dogecoin price for the period 1st May 2021 to 26th April 2022. The data was collected from Yahoo finance by using the function `getSymbols` from the `quantmod` package(Ryan and Ulrich 2020). The original dataset contains 361 observations with six variables each, including the open price, the high price, the low price, the close price, the volume, and the adjusted price. We focus on the close price since it represents the last price during a regular trading day(Hayes 2021).

ARIMA is a statistical analysis model for forecasting time series data such as unemployment rates, stock prices, and global temperature. In the section on methods, we will discuss how the ARIMA model can predict a future value based on past values and past errors. Our analysis of our data shows that dogecoin prices fit the ARIMA(1,1,1) model, and by predicting based on this model the price of dogecoin will continuously decrease for the next ten days.

2 Data

2.1 Time series data introduction

When analyzing time series data, a sample size of at least 100 observations is preferred for a model to be accurate. From 1st May 2021 to 26th April 2022, the dataset collected from Yahoo finance records the open price, the high price, the low price, the close price, the volume, and the adjusted price of dogecoin, with a total of 361 observations, so if there was an underlying relationship between observations, we could build a relatively precise model. Dogecoin was first released in 2013, however, we are not using all the observations for modeling because the price of Dogecoin was so low for a long period of time, as shown in Figure 1 shown, which would largely affect the accuracy of the model, thus we focus on the price of dogecoin between 1st May 2021 and 26th April 2022. Figure 2, observed that although all five different prices have varying values, their trends and shapes are quite similar, the prices were declining rapidly from around 0.7 USD to around 0.2 USD after May 2021, after that the prices fluctuated a little but the overall trend was decreasing. For modeling, we only have to analyze one of these prices, in this case, the closing price. In Figure3, the volume of Dogecoin for the past 361 days is shown. Although volume can provide some insight into whether an investor should enter the market or not, this paper will only examine the price.

2.2 Data visualization

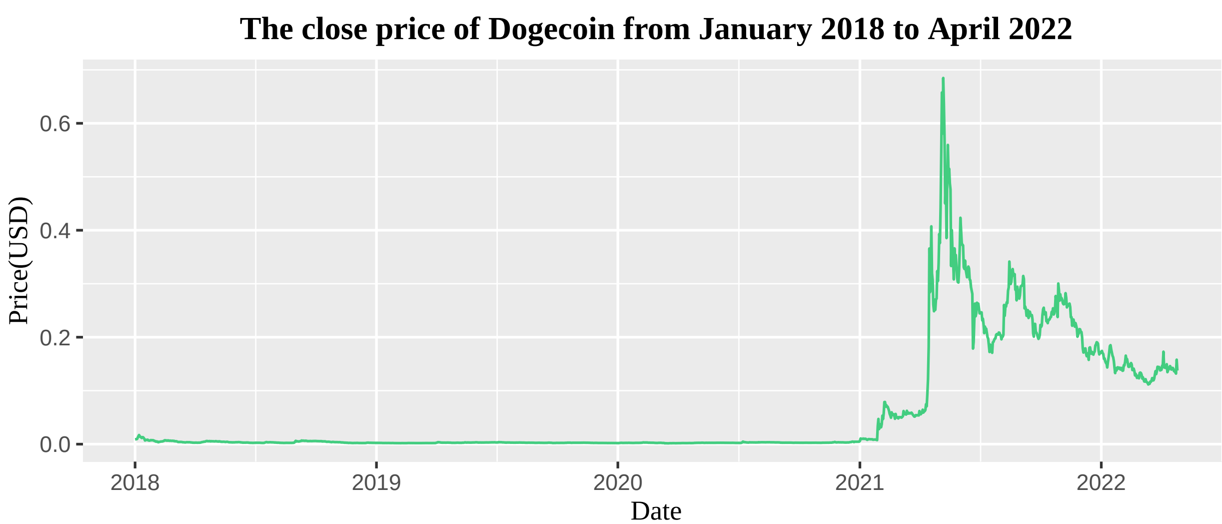

Figure 1: The close price of Dogecoin from 2018 to 2022

Figure 2: Five different prices of the Dogecoin from May 2021 to April 2022

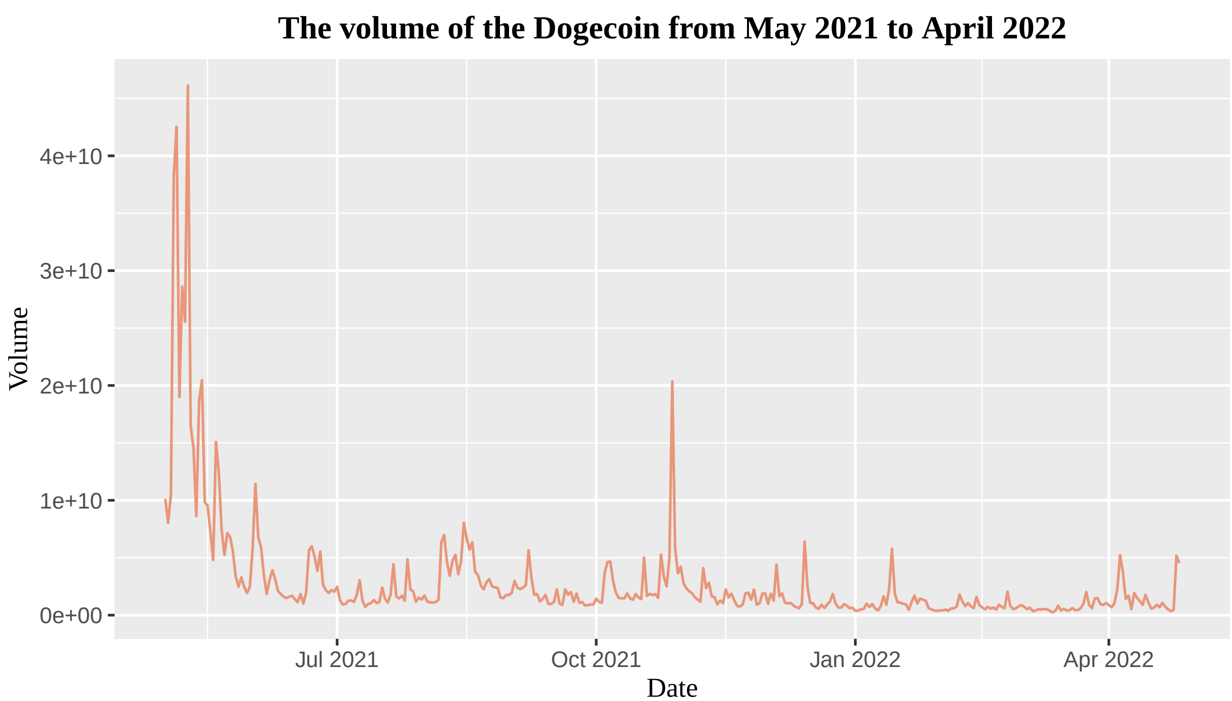

Figure 3: The volume of the Dogecoin from May 2021 to April 2022

3 Method

3.1 ARIMA model

Autoregressive integrated moving average model ARIMA(p,d,q) is one of the most common and useful models to predict future values based on past values and past errors, where p represents the autoregressive order, d is the dependent order(the number of difference transformations needed), and q is the moving average order. The mathematical equation of an ARIMA model can be expressed as follows:

$$X_t = \phi_0 + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \phi_3 X_{t-3} + ... + \phi_p X_{t-p} + \theta_0 + \theta_1 W_{t-1} + \theta_2 W_{t-2} + \theta_3 W_{t-3} + ... + \theta_q W_{t-q}$$

In the above equation, X_t is the value to predict, $X_{t-1}, X_{t-2}, X_{t-3} ... X_{t-p}$ are past values, ϕ_p are the coefficients of the past values, $W_{t-1}, W_{t-2}, W_{t-3} ... W_{t-p}$ are past white noise errors, θ_p are the coefficients of the past errors.One important assumption to build a ARIMA model is that the time series data is stationary with no seasonality. A stationary time series is defined as time series that have:

- Constant mean over time t
- Constant variance over time t
- The Auto-covariance function between two observations X_{t1} and X_{t2} only depends on the interval t_1 and t_2.

Changing the data into stationary without seasonality is therefore necessary before fitting a model to the time series data. There are generally two ways to do this. The first method is Applying difference transformation: transform each observations at time t into the difference between observation at time t and observation at time (t-1),the mathematical equation is defined as $\Delta X_t = X_t - X_{t-1}$; the second method is applying Box-Cox transformation.

As shown in Figure 4, there is an obvious trend, therefore the time series does not seem to be stationary. Hence, a first difference transformation was applied to the data. The Figure 5 shows the price of the Dogecoin after the first difference transformation. We observed that the mean became constant over time, however, the variance is still obviously not constant over time since the plot exhibits a fanning pattern. Thus we applied a Box-Cox transformation to stabilize the variance. The Figure 6 shows the price of the Dogecoin after both transformations. We observed that both mean and variance have been stabilized, thus we would use this Box-Cox transformed data to fit a model. At this point, we can determine the dependence order of the ARIMA(p,d,q) is d=1.

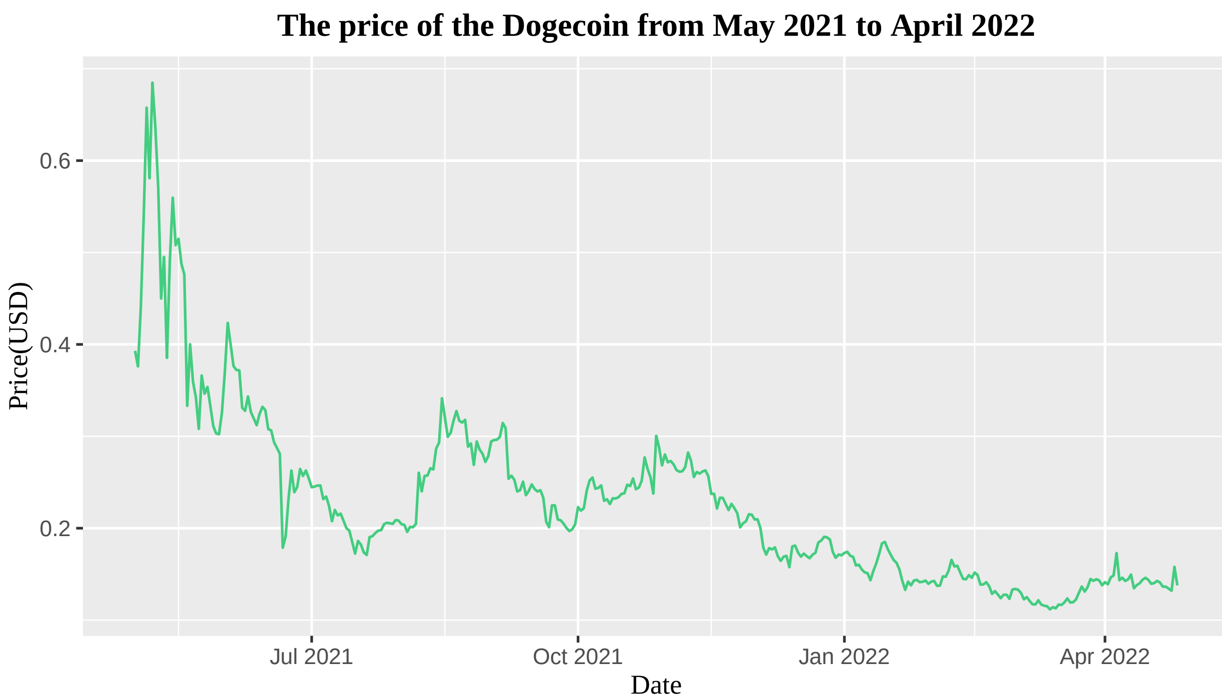

Figure 4: The price of the Dogecoin from May 2021 to April 2022

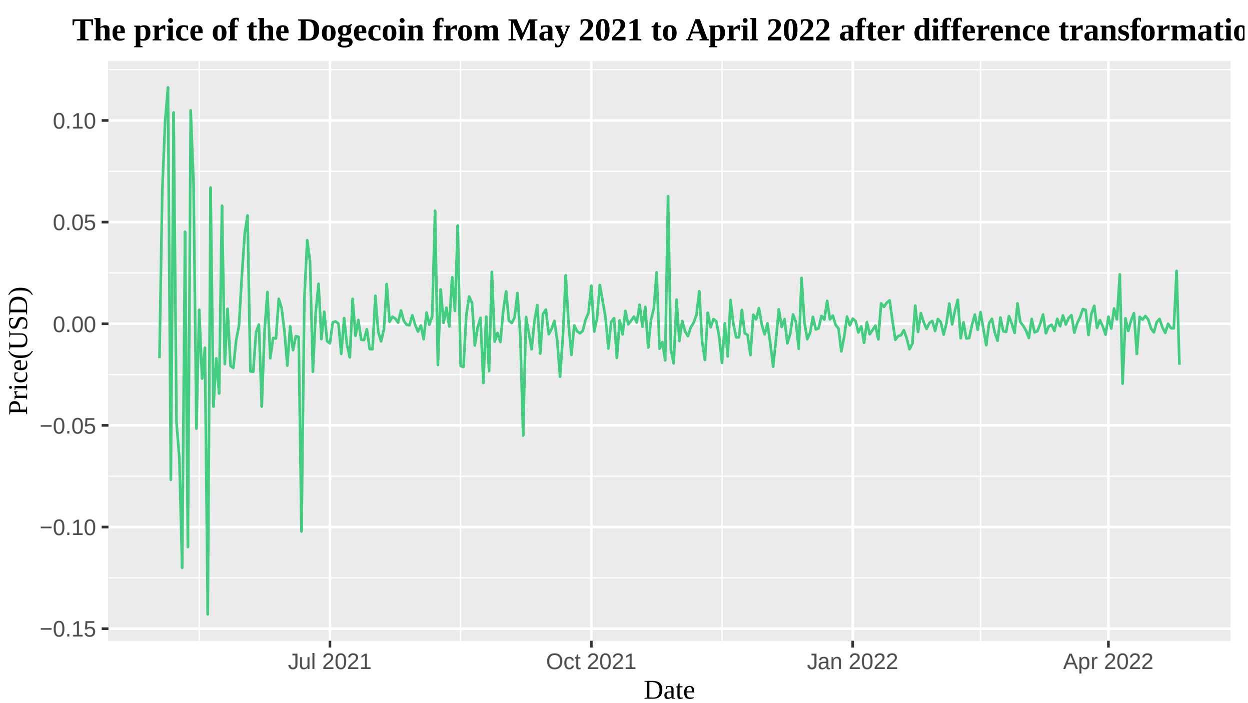

Figure 5: The price of the Dogecoin from May 2021 to April 2022 after difference transformation

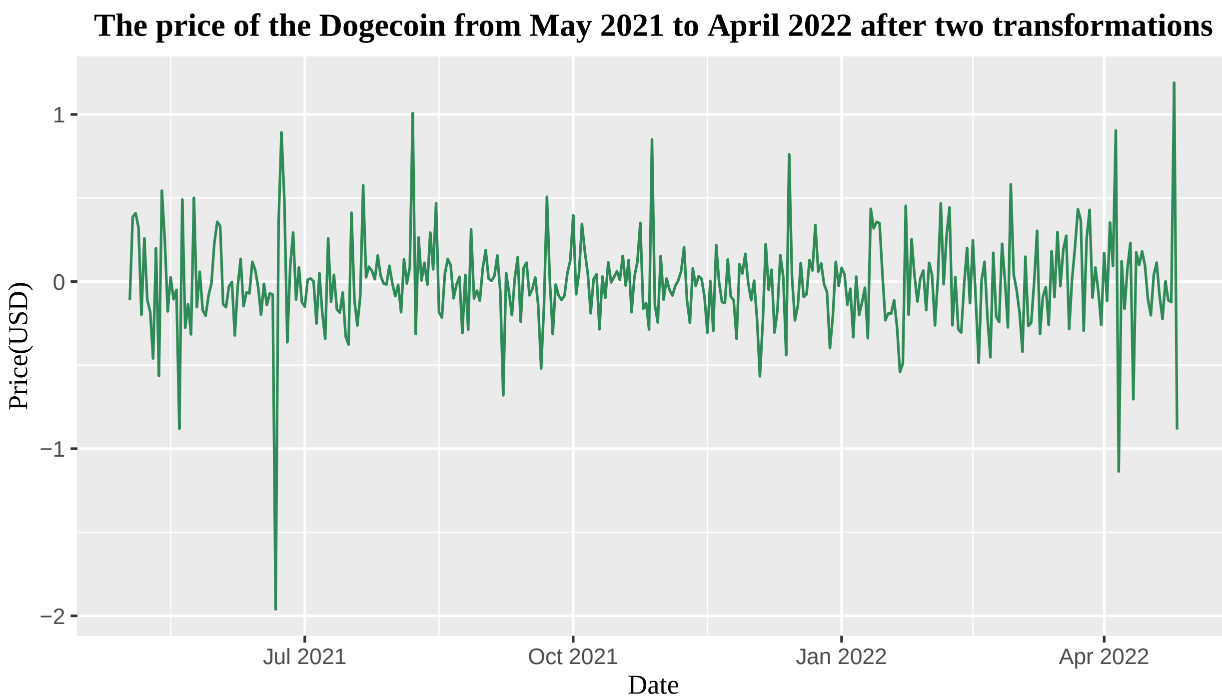

Figure 6: The price of the Dogecoin from May 2021 to April 2022 after difference and Box-Cox transformation

3.2 ACF and PACF

An Autocorrelation function (ACF) plot and a Paritial autocorrelation function (PACF) plot were used to determine the AR and MA order. Autocorrelation function calculates the correlation between one observation(X_t) with it's lag(X_{t-h}), while PACF also measures the correlation between X_t and its lag at time t X_{t-h}, in addition each correlation controls for any correlation between observations of a shorter lag length(Abigail-Kate 2021). Figure 7 shows the ACF and PACF plots of the transformed time series data. The ACF plots show a gradually declining to zero after the first lag. The PACF also shows a gradual decline to zero after the first lag. So the AR order of the model could be 1 or 2 since PACF is also quite large at second lag. The MA order of the model could be 1 since the ACF declines rapidly after lag 1. Thus two possible models can be introduced, ARIMA(1,1,1) and ARIMA(2,1,1)

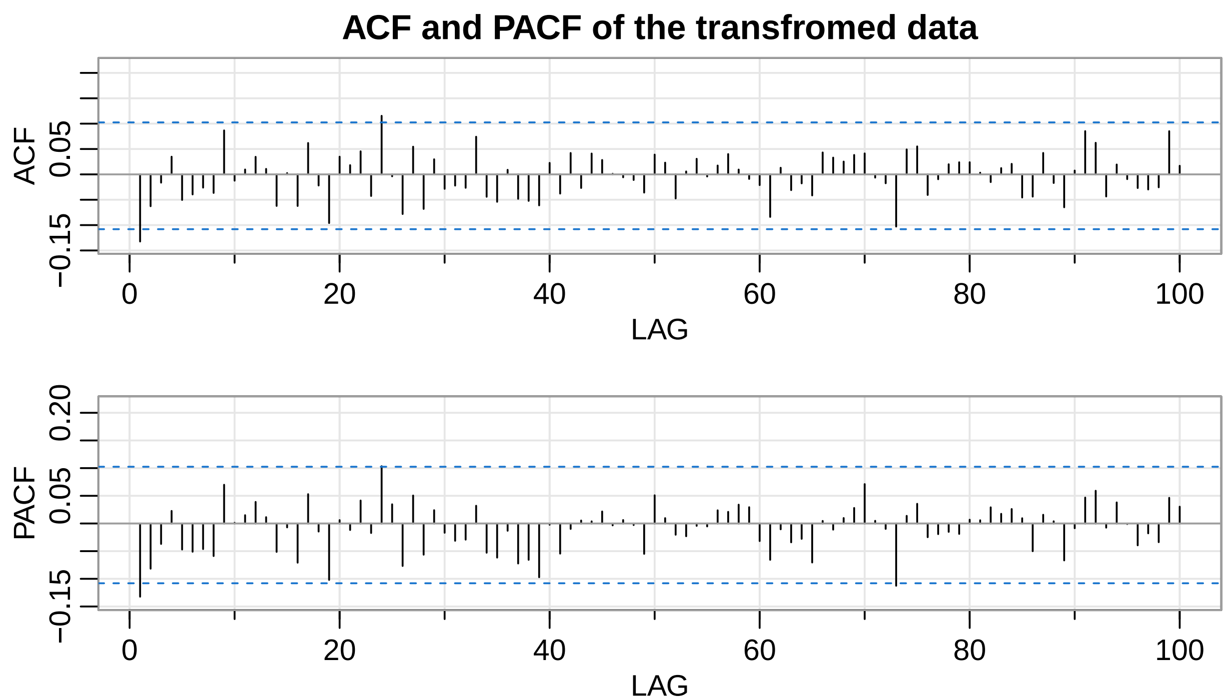

Figure 7: TACF and PACF of the transfromed data

3.3 Model selection

For model diagnostics, as Figure 8 and Figure 9 shown, we plot a residual plot, a ACF of residual plot, a normal-QQ plot, and a Ljung-Box statistic plot to examine whether the model violated any model assumptions or not. Then we compare the Akaike information criterion (AIC) of each plot, finally, we compare the significance of their coefficients to find the most suitable model.

3.3.1 Model assumptions

A valid model should not violate any of the assumptions such as normality, independence, and randomness. We can use a standard residual plot and ACF residual plot to test whether the model violates the randomness assumption. Use Normal Q-Q plot to determine whether the model violates the normal assumption, and use Ljung-Box statistics plot to determine whether the model violates the independent assumption. Figure 8 and Figure 9 shows the necessary model assumptions of two models.

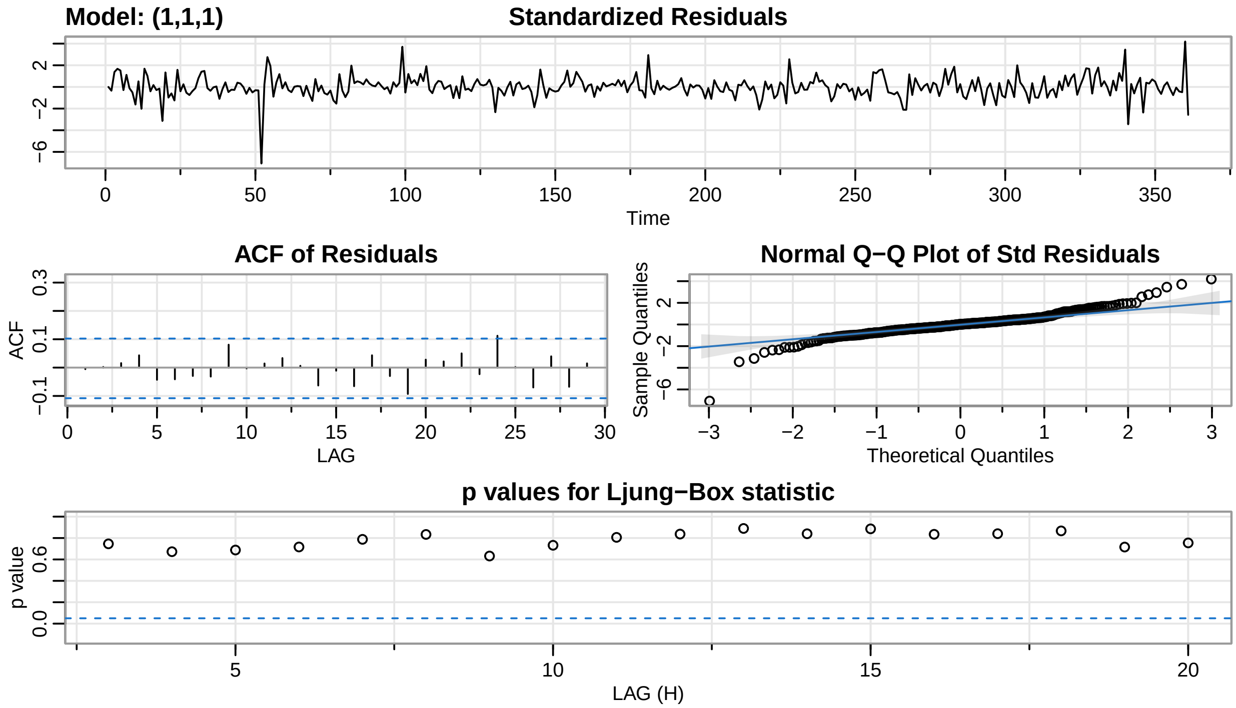

Figure 8: Model diagnostics for ARIMA(1,1,1)

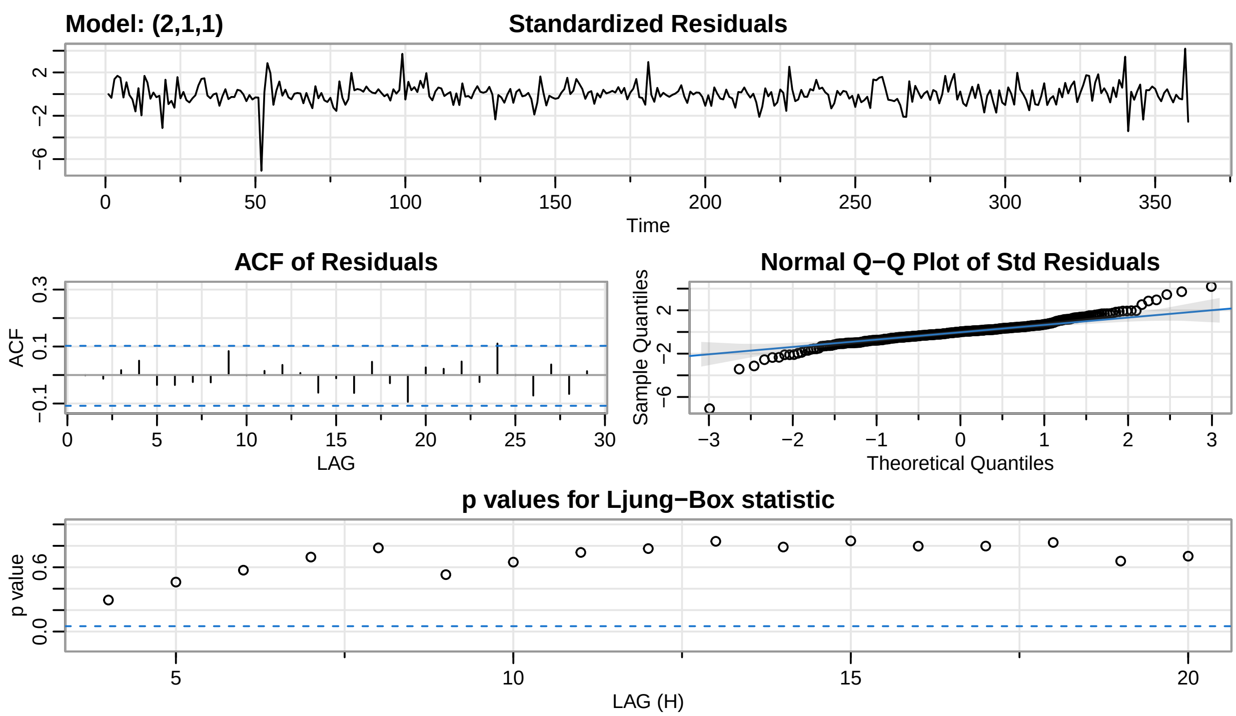

Figure 9: Model diagnostics for ARIMA(2,1,1)

3.3.2 AIC

Akaike information criterion (AIC) is a statistical tool to examine how a model fit the data, and the model's simpilicity, and it has been widely used to compare different model for selecting the best model(Bevans 2020). The lower AIC, the better the model fits the data. The AIC of ARIMA(1,1,1) is 110.95 and ARIMA(2,1,1) model is 112.81.

3.3.3 Parameters significance

To test the significance of the parameter we set a null hypothesis and an alternative hypothesis as follow:

- null hypothesis H_0 :the coefficients of the parameters are not significant different from 0
- alternative hypothesis H_A:the coefficients of the parameters are significant different from 0

Below tables shows the coefficients of each parameters and their p-values.

	Estimate	SE	t.value	p.value
ar1	0.4945	0.2760	1.7915	0.0741
ma1	-0.6375	0.2472	-2.5789	0.0103
constant	-0.0128	0.0106	-1.2073	0.2281

	Estimate	SE	t.value	p.value
ar1	0.5966	0.2971	2.0079	0.0454
ar2	0.0346	0.0889	0.3892	0.6974
ma1	-0.7453	0.2923	-2.5497	0.0112
constant	-0.0128	0.0102	-1.2565	0.2098

4 Result

4.1 ARIMA(1,1,1)

Figure 8 is the model diagnostic plots for the ARIMA(1,1,1) model. We observed that the standard residuals showed no obvious patterns, and there were no outliers. The ACF Residuals plot shows only one spike. Hence, we conclude that there are no violations of the model randomness assumption. In addition, there are no heavy tails on both ends of the Normal Q-Q Plot of Residuals, which supports that there are no apparent violations of the model normality assumption. Besides, all the p-values for Ljung-Box statistics are above the significant level, so there is no violation of the model independent assumption. So no model assumptions are violated. The ARIMA(1,1,1) model is a valid model with a Akaike information criterion (AIC) of 110.95.

Figure 9 is the model diagnostic plots for the ARIMA(1,1,1) model, we observed that the standard residuals shows no obvious patterns, and there are no outliers. The ACF Residuals plot shows only one spike. Hence, we conclude that there are no violations of the model randomness assumption. In addition, there are no heavy tails on both ends of the Normal Q-Q Plot of Residuals, it supports that there are no apparent violations of the model normality assumption. Besides, all the p-values for Ljung-Box statistics are above the significant level, so there is no violation of the model independent assumption. So no model assumptions are violated. The ARIMA(1,1,1) model is a valid model with a Akaike information criterion (AIC) of 112.81.

So neither model violated any model assumptions. However, ARIMA(1,1,1) has a smaller AIC(110.95) than ARIMA(2,1,1) which has a AIC of 112.81. Thus from the AIC perspective, ARIMA(1,1,1) is the better model which may fit better to the data, and has more simplicity. In addition, the parameter tables suggest that the parameter of ARIMA(1,1,1) has p-values of 0.0741 and 0.0103, thus we have no evidence to reject our null hypothesis, thus all the parameters in ARIMA(1,1,1) are significant. On the other hand, the parameters of

ARIMA(2,1,1) have p-values of 0.0454,0.6974, and 0.0112 respectively, while the first AR and MA parameters have small p-values, the second AR parameter has a p-value that is larger than 0.5, which means we cannot reject the null hypothesis, so the second AR parameter in the ARIMA(2,1,1) model is not significant.

So in general, we found the ARIMA(1,1,1) is the better model to fit the time series data.

4.2 Predicting the future price

The final step is to use the ARIMA(1,1,1) with $\phi = 0.4945$ and $\theta = -0.6375$ to predict the price of Dogecoin for the next ten days. Figure 10 visualize our prediction, however since we Box-Cox transformed our data previously, the value of prediction has also been Box-Cox transformed.

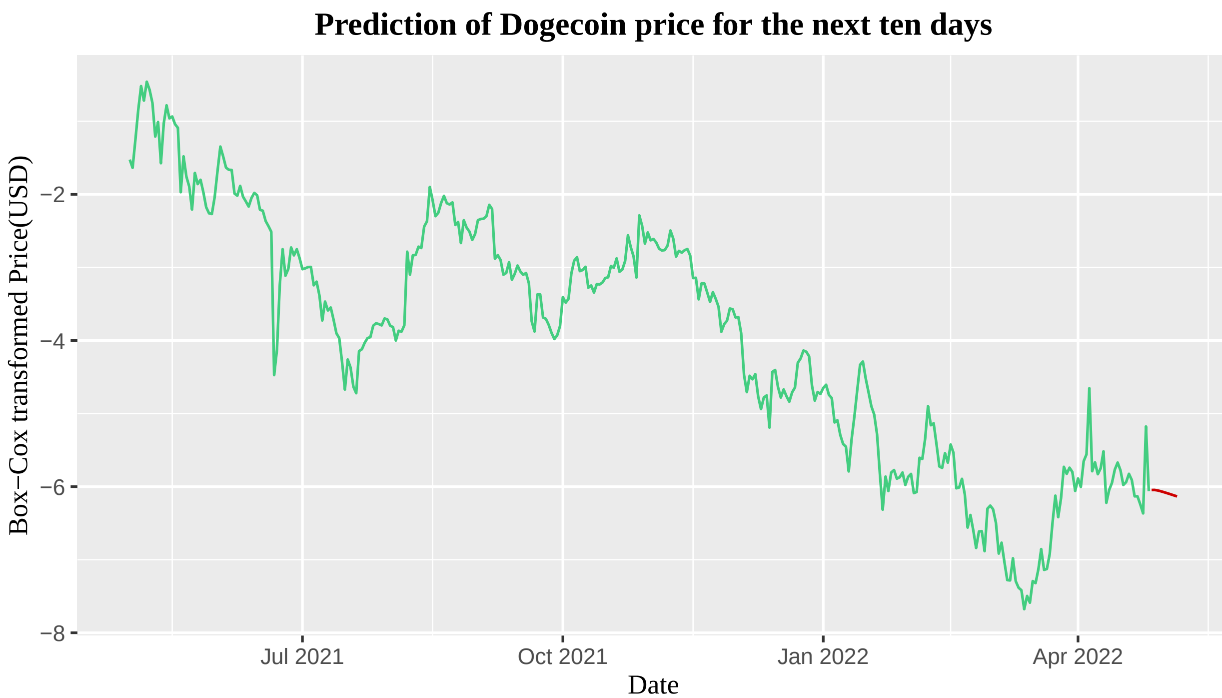

Figure 10: Dogecoin price prediction for the next ten days

We calculate the actual predicted price which has not been Box-Cox transformed by doing an inverse Box-Cox transformation. Figure 11 shows the actual prediction of the Dogecoin price in the next ten days as from 26th April 2022. We observed that the price will continue to decline for the next ten days.

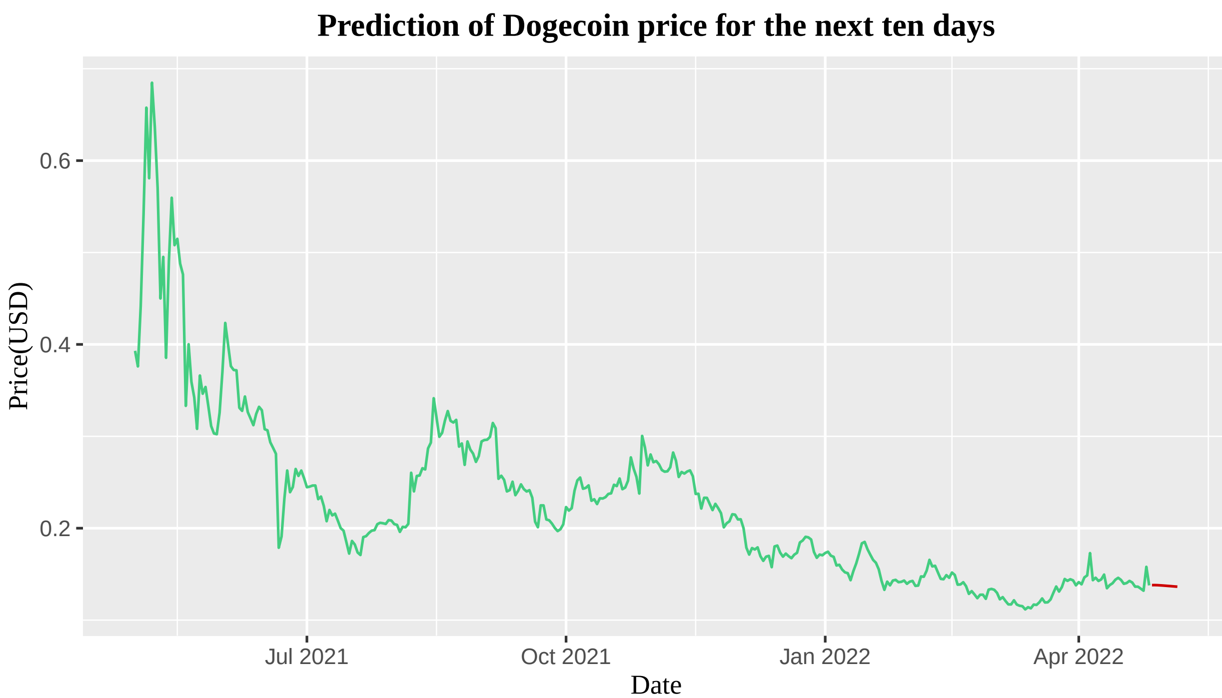

Figure 11: Dogecoin price prediction for the next ten days

5 Disscussion

5.1 Conclusion

We found that a ARIMA(1,1,1) model with coefficient $\phi = 0.4945$ and $\theta = -0.6375$ can fit the Dogecoin data from May 2021 to April 2022, then if we apply this model to predict the future price of the Dogecoin we can find that the price will be continuously decreasing for the next ten days. So for investors who are looking for investment opportunities, investing in Dogecoin may be a risky choice. In addition, the decreasing trend may also imply that Elon Musk is losing his interest in Dogecoin and Cryptocurrency.

5.2 "All models are wrong"

A famous British statistician George E. P. Box once said "All models are wrong, but some are useful"(Barroso 2019). Due to Dogecoin's volatility, its price may be influenced by many factors such as government policy, financial influencer involvement, and so on. In some cases, a cryptocurrency may rise today, and disappear tomorrow. So our model certainly can not predict the actual future value of the Dogecoin, however the prediction can show investors some insights about the overall trend of the price to help them decide whether to invest it or not.

5.3 Next step

Except for the fact that cryptocurrency is naturally challenging to predict, in our data, we only have useful 361 observations. In some cases, 361 observations is enough for data set to build a decent model, however, for financial prediction the more informative observations the better. So for the next step, I will continue monitoring the Dogecoin price to see if the actual price matches our prediction, if not, I will try to build another model to fit the data.

5.4 Relevant research

I think another research which focuses on the building the relationship between the number of Elon Musk's tweets about Dogecoin and the price of the Dogecoin could be made, to see if there is actually a relationship between them, if so, how can we use this relationship to predict the Dogecoin price and make investment advice.

6 References

knitr::write_bib(c("knitr,""base,""float,""quantmod,""dplyr,""janitor,""ggplot2,""ggthemes,""astsa,""aTSA,""forecast,""fp
knitr::write_bib(c("knitr,""base,""float,""quantmod,""dplyr,""janitor,""ggplot2,""ggthemes,""astsa,""aTSA,""forecast,""fp
knitr::write_bib(c("knitr,""base,""float,""quantmod,""dplyr,""janitor,""ggplot2,""ggthemes,""astsa,""aTSA,""forecast,""fp

Abigail-Kate, Abigail-Kate. 2021. "Learn about Time Series ACF and PACF in Stata with Data from the USDA Feed Grains Database (1876–2015)." https://doi.org/10.4135/9781529710298.

Arnold, Jeffrey B. 2021. *Ggthemes: Extra Themes, Scales and Geoms for Ggplot2*. https://github.com/jrnold/ggthemes.

Barroso, Guillem. 2019. *"All Models Are Wrong, but Some Are Useful." George e. P. Box*. https://www.lacan.upc.edu/admoreWeb/2018/05/all-models-are-wrong-but-some-are-useful-george-e-p-box/.

Bevans, Rebecca. 2020. *Akaike Information Criterion | When & How to Use It*. https://www.scribbr.com/statistics/akaike-information-criterion/#:~:text=The%20Akaike%20information%20criterion%20(AIC,best%20fit%20for%20the%20data.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Frankenfield, Jake. 2022a. *Cryptocurrency*. https://www.investopedia.com/terms/c/cryptocurrency.asp.

———. 2022b. *Dogecoin (DOGE)*. https://www.investopedia.com/terms/d/dogecoin.asp#:~:text=Dogecoin%20is%20an%20open%2Dsource,proof%2Dof%2Dwork%20technology.

Hayes, Adam. 2021. *Closing Price*. https://www.investopedia.com/terms/c/closingprice.asp#:~:text=Key%20Takeaways,stock%20dividends%2C%20or%20stock%20splits.

Hyndman, Rob. 2020. *Fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition)*. https://CRAN.R-project.org/package=fpp2.

Hyndman, Rob J, and Yeasmin Khandakar. 2008. "Automatic Time Series Forecasting: The Forecast Package for R." *Journal of Statistical Software* 26 (3): 1–22. https://doi.org/10.18637/jss.v027.i03.

Hyndman, Rob, George Athanasopoulos, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Mitchell O'Hara-Wild, Fotios Petropoulos, Slava Razbash, Earo Wang, and Farah Yasmeen. 2022. *Forecast: Forecasting Functions for Time Series and Linear Models*. https://CRAN.R-project.org/package=forecast.

Morris, Chris. 2022. *Elon Musk Added 20*. https://fortune.com/2022/04/04/elon-musk-twitter-investment-stock-jump-hisotry-of-company-tweets/.

Qiu, Debin. 2015. *aTSA: Alternative Time Series Analysis*. https://CRAN.R-project.org/package=aTSA.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Ryan, Jeffrey A., and Joshua M. Ulrich. 2020. *Quantmod: Quantitative Financial Modelling Framework*. https://CRAN.R-project.org/package=quantmod.

Schmidt, Drew. 2017. *Introducing the float Package: 32-Bit Floats for R*. https://cran.r-project.org/package=float.

———. 2022. "float: 32-Bit Floats." https://cran.r-project.org/package=float.

Schmidt, Drew, and Wei-Chen Chen. 2022. *Float: 32-Bit Floats*. https://github.com/wrathematics/float.

Stoffer, David. 2021. *Astsa: Applied Statistical Time Series Analysis*. https://CRAN.R-project.org/package=astsa.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2021. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Re-*

producible Computational Research, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

———. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. https://yihui.org/knitr/.

———. 2022. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.