Predicting Daily Coffee Price Through Time Series Analysis

Amir Voloshin | amirvoloshin@ucsb.edu

6/14/2023

Abstract

This time series report presents the results and findings of predicting the daily closing stock price of coffee between August 18, 2022, and September 2, 2022, using SARIMA and GARCH models. Initially, a model was built using historical data dating back to 2002, but it proved to be inaccurate. To enhance accuracy, the models were trained using data from the past year and a half. The predicted points from both the SARIMA and GARCH models were similar but deviated from the actual closing prices during the observed days. However, all of the true closing prices fell within the confidence intervals of the predicted points, indicating successful predictions.

Considering the volatility and unpredictable nature of stock market prices, precise future stock price prediction remains challenging. The higher actual closing price compared to the predicted price may be attributed to various factors, including market volatility and the effects of inflation, particularly in the post-COVID-19 period.

Introduction

This project aims to analyze the daily stock price of coffee from January 3, 2000, to September 2, 2023, using time series analysis techniques. Specifically, two forecasting methods, SARIMA (Seasonal AutoRegressive Integrated Moving Average) and GARCH (Generalized AutoRegressive Conditional Heteroskedasticity), are employed to predict the future stock price of coffee.

The SARIMA model captures the temporal dependencies and seasonal patterns in the coffee stock price data, allowing for accurate short-term and long-term forecasting. It incorporates autoregressive (AR), moving average (MA), and differencing (I) components to account for trend, seasonality, and stationarity.

Additionally, the GARCH model is used to capture the volatility clustering observed in financial time series data, including stock prices. By considering the conditional variance as a function of past squared residuals, GARCH provides insights into the risk and uncertainty associated with the coffee stock price.

The project utilizes historical data to estimate the model parameters and validate the forecasting accuracy. The performance of SARIMA and GARCH models is evaluated based on various metrics, such as mean squared error (MSE) or root mean squared error (RMSE), to assess the predictive power of each method.

Ultimately, the project aims to provide reliable forecasts of the stock price of coffee. In the real world this is useful in assisting investors, traders, and other stakeholders in making informed decisions related to the coffee market. The results obtained from SARIMA and GARCH modeling provide us with a better understanding of the coffee market dynamics and its the possible risks associated with investing in this market.

Data

This dataset contains the Daily Coffee Prices in Cents per pound in USD from January 3, 2000 to September 2, 2023. The coffee prices available are the opening price, closing price, high price, and low price. The dimensions of this dataset are 5746 x 7, making it quite large.

I found this data set on kaggle, it has 4585 downloads, and 21 unique contributors, as this data set is open source. Since this data if open sourced, I decided to check the actually daily stock prices of coffee on the Nasdaq website and found that the data is the same, proving its reliability as real data. I have the two link below, the kaggle page which I found the dataset on, and the Nasdaq website of coffee. As it suggests, this data is collected from the stock market price of coffee.

https://www.kaggle.com/datasets/psycon/daily-coffee-price

https://www.nasdaq.com/market-activity/commodities/kt:nmx

This dataset is important because Coffee is one of the worlds most consumed drink, and has an interesting trend in price over time. Coffee beans are a crop which grows well in certain areas of the world, is imported, and then roasted (not always in that order). So climate and natural factors affect how much coffee can be grown, and the quality of the coffee, which in turn affects the price of coffee. All of these factors make it a fascinating subject to study using Time Series. Furthermore, I am personally very into coffee and studied about it in my free time in the past, so finding a dataset about coffee prices for this project was a perfect match.

Methodology

The SARIMA model captures the temporal dependencies and seasonal patterns in the coffee stock price data, allowing for accurate short-term and long-term forecasting. It incorporates autoregressive (AR), moving average (MA), and differencing (I) components to account for trend, seasonality, and stationarity.

The GARCH model is used to capture the volatility clustering observed in financial time series data, including stock prices. By considering the conditional variance as a function of past squared residuals, GARCH provides insights into the risk and uncertainty associated with the coffee stock price.

Libraries

```
library(dplyr)
library(tidyverse)
library(astsa)
library(forecast)
library(fGarch)
```

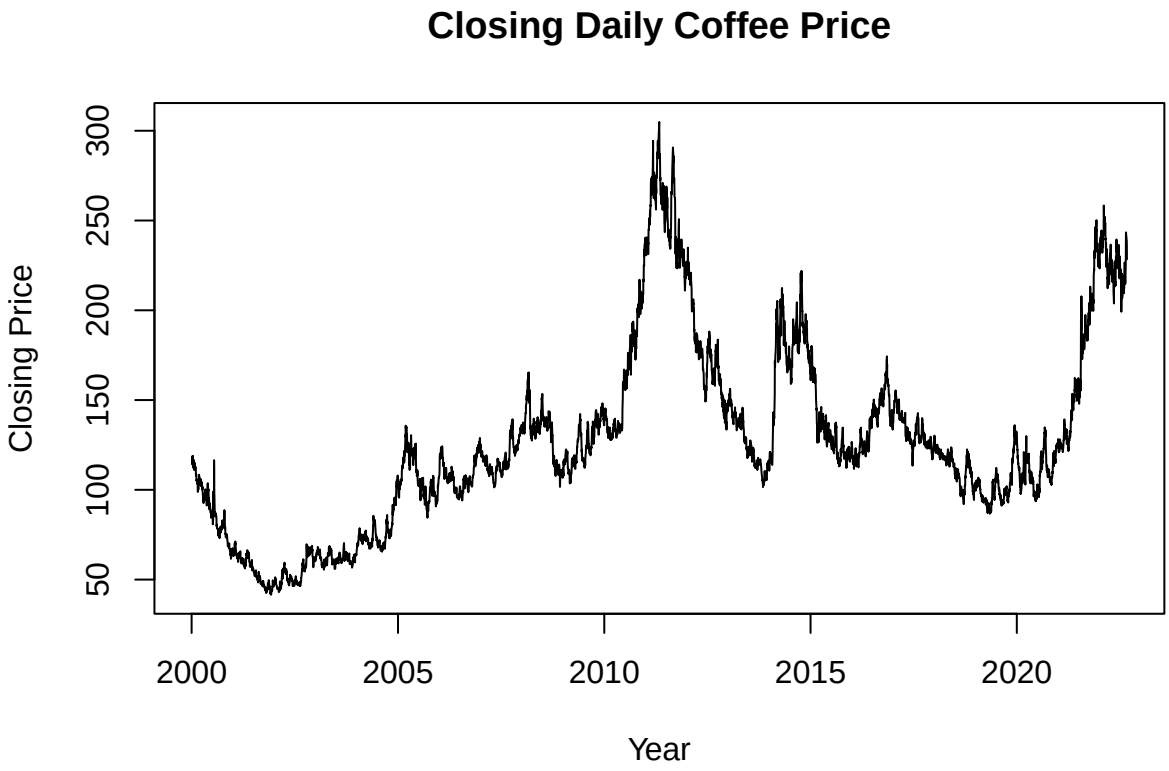
Results

Visualizing the Data and Diagnostics

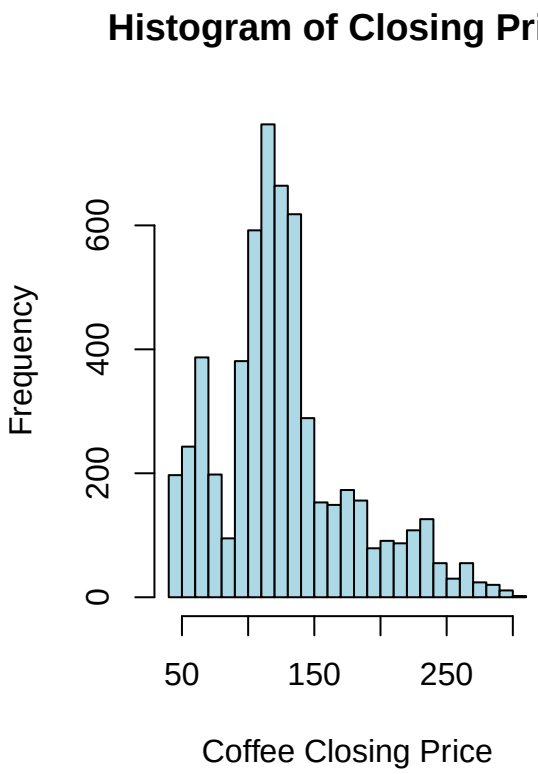
```
# Reading in the Daily Coffee price data
coffee <- read_csv('/Users/amirvoloshin/Desktop/PSTAT 174/Final Project/Coffee/coffee.csv')
head(coffee)
```

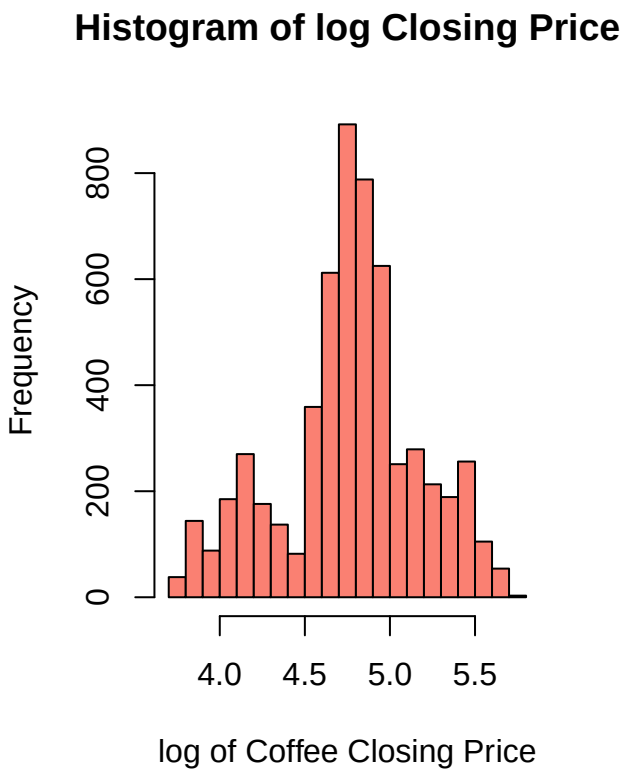
```
## # A tibble: 6 x 7
##   Date        Open  High   Low Close Volume Currency
##   <date>     <dbl> <dbl> <dbl> <dbl>  <dbl> <chr>
## 1 2000-01-03  122.  124   116.  116.   6640 USD
## 2 2000-01-04  116.  120.  116.  116.   5492 USD
## 3 2000-01-05  115   121   115   119.   6165 USD
## 4 2000-01-06  119   121.  116.  117.   5094 USD
## 5 2000-01-07  117.  118.  114.  114.   6855 USD
## 6 2000-01-10  124.  126   117.  118.   7499 USD
```

```
# Original Data
plot(x = coffee$Date, y = coffee$Close, type = 'l', xlab = 'Year', ylab = 'Closing Price', main = 'Clos
```



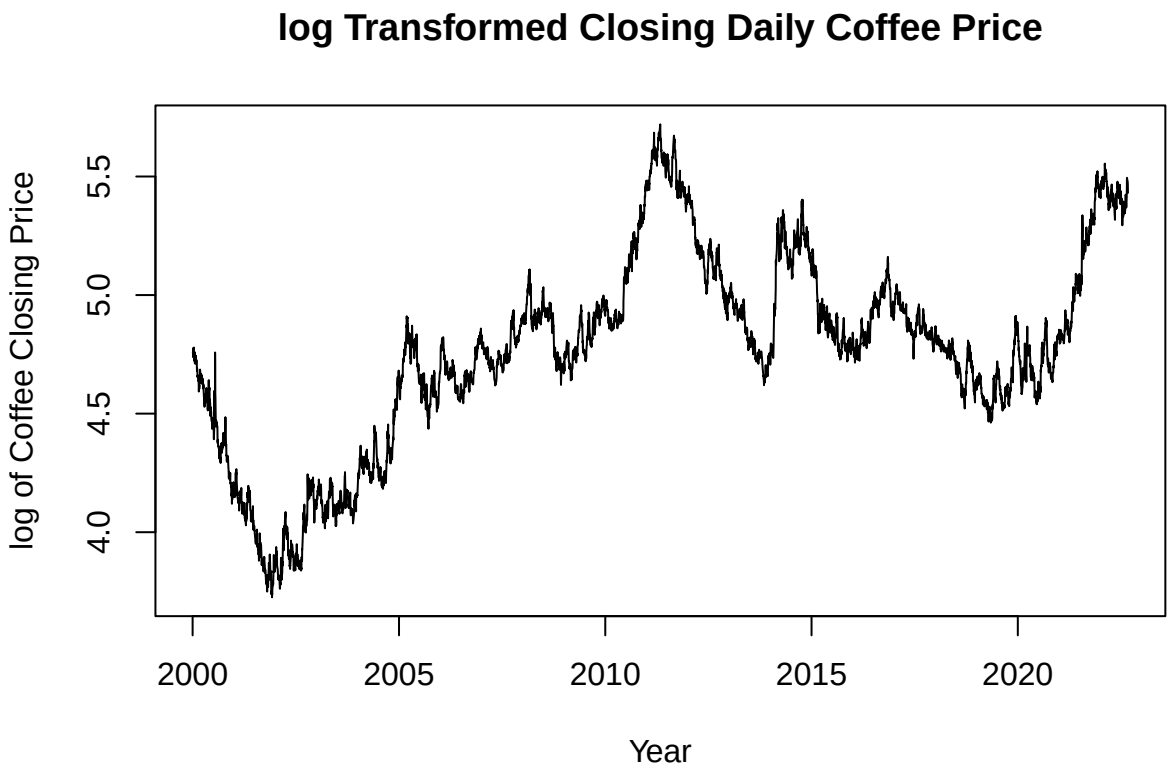

```
# Comparing Histogram of the data, and log of the data for normality
par(mfrow = c(1, 2))
hist(coffee$Close, main = 'Histogram of Closing Price', xlab = 'Coffee Closing Price', breaks = 20, col
hist(log(coffee$Close), main = 'Histogram of log Closing Price', xlab = 'log of Coffee Closing Price',b
```



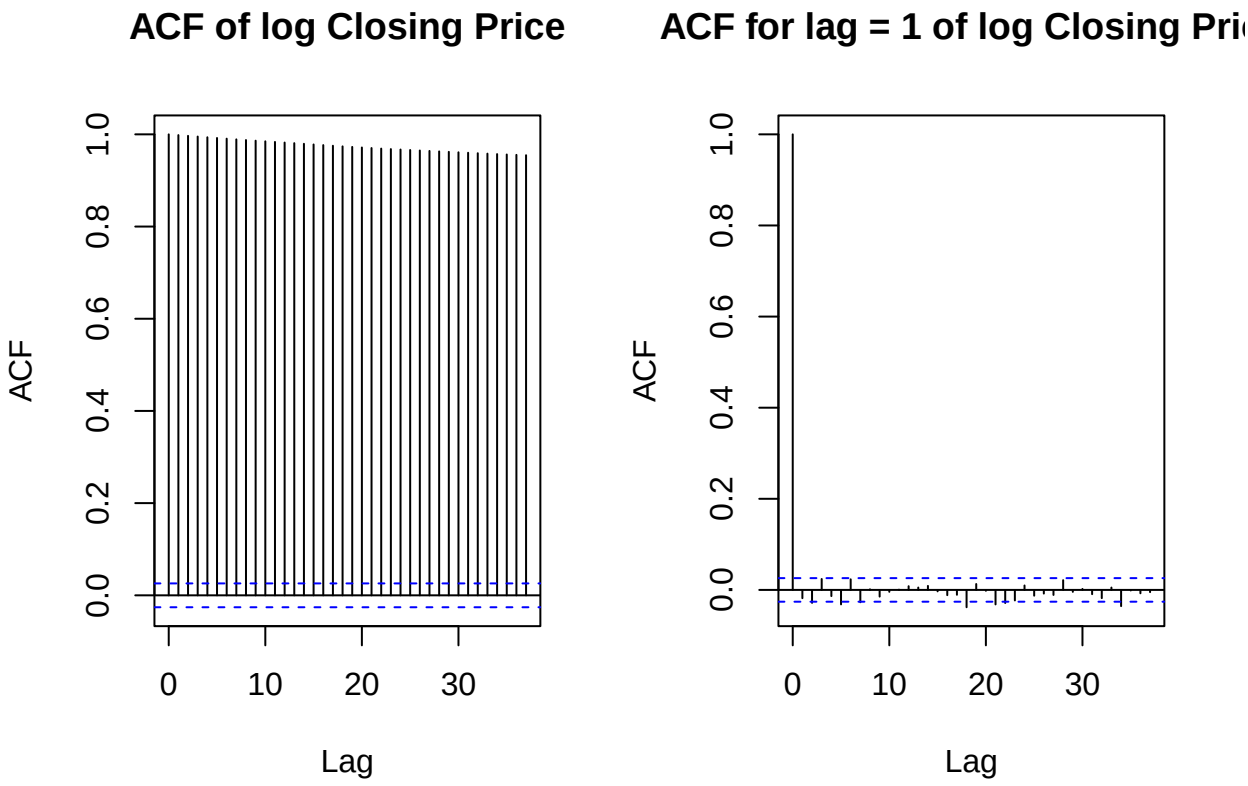


As we can see above when comparing the histogram, my original data set is not normally distributed but the log transformed data set is very close to normal. I will continue to use the transformed data because it is close to normal.

```
# Plotting transformed data
plot(x = coffee$Date, y = log(coffee$Close), type = 'l', main = 'log Transformed Closing Daily Coffee Pr
```

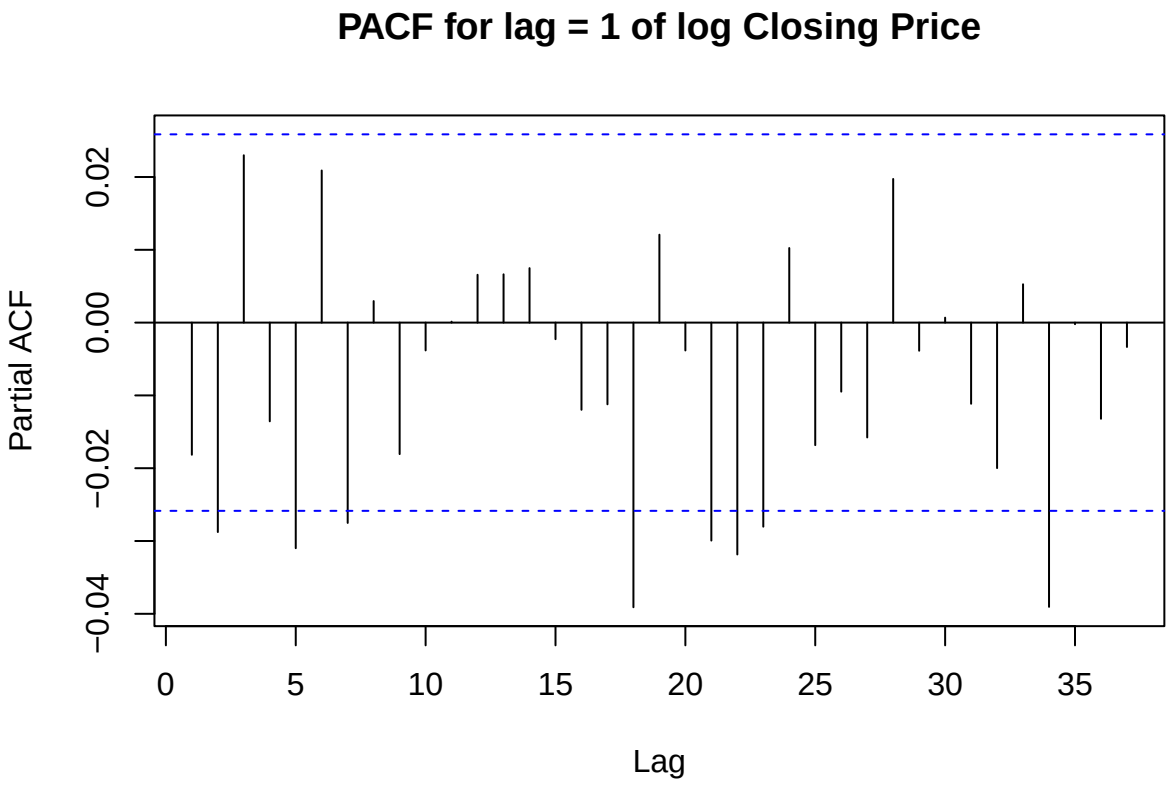



```
par(mfrow = c(1, 2))

# Auto-Correlation Plot
acf(log(coffee$Close), main = 'ACF of log Closing Price')

# Differenced Auto-Correlation Plot
acf(diff(log(coffee$Close), 1), main = 'ACF for lag = 1 of log Closing Price')
```



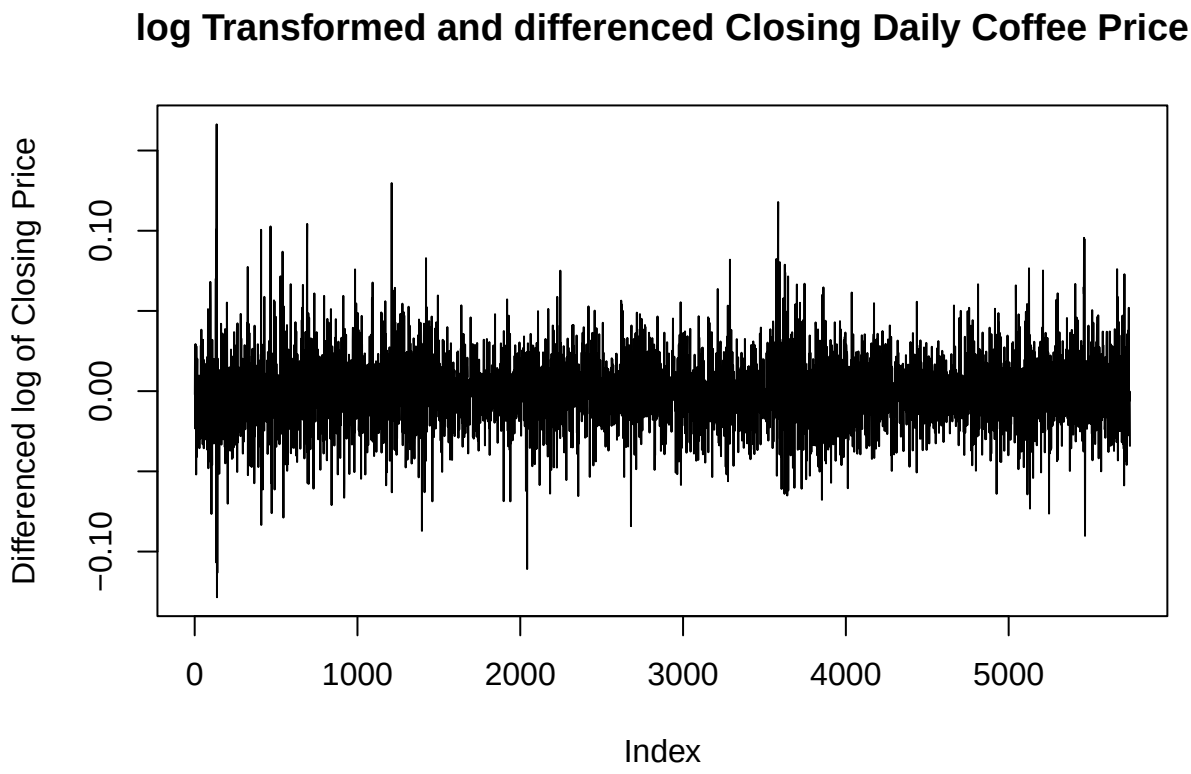
As we see above there is a very slow decay of the auto-correlation plot, this is an indication that there is high correlation between the points. On the other hand, the ACF plot with a lag of 1 has only white noise variation so I will continue to use the differenced data.

```
# Partial Auto-Correlation Plot
pacf(diff(log(coffee$Close), 1), main = 'PACF for lag = 1 of log Closing Price')
```



From the partial auto-correlation plot we can see a lot of the variation is simply due to white noise.

```
# Log tranformed and differenced plot of data
plot(diff(log(coffee$Close), 1) ,type = 'l', main = 'log Transformed and differenced Closing Daily Coff
```



Looking at the new plot of my transformed and differenced data, the data now looks to be stationary which will be very beneficial when forecasting. Stationarity is a property we want and it makes the models easier to build.

Fitting a SARIMA Model

```
# Adding log column to dataframe
coffee_log <- log(coffee$Close)
coffee <- cbind(coffee, coffee_log)

# Assigning train and test sets
n <- length(coffee$coffee_log)

# Using only the past 700 points to account for some trend
dates <- coffee$Date[5000:(n - 12)]
x.train <- coffee$coffee_log[5000:(n - 12)]
x.test <- coffee$coffee_log[(n - 11):n]
n <- length(x.train) + length(x.test)
```

```
# Model Selection to Select best SARIMA Model
df <- data.frame(expand.grid(P=0, Q=0, p=0:7, q=0:5), AIC=NA, BIC=NA)
for (i in 1:nrow(df)) {
  m <- df[i, ]
  fit <- arima(x.train, order=c(m$p, 1, m$q), # first differencing
```

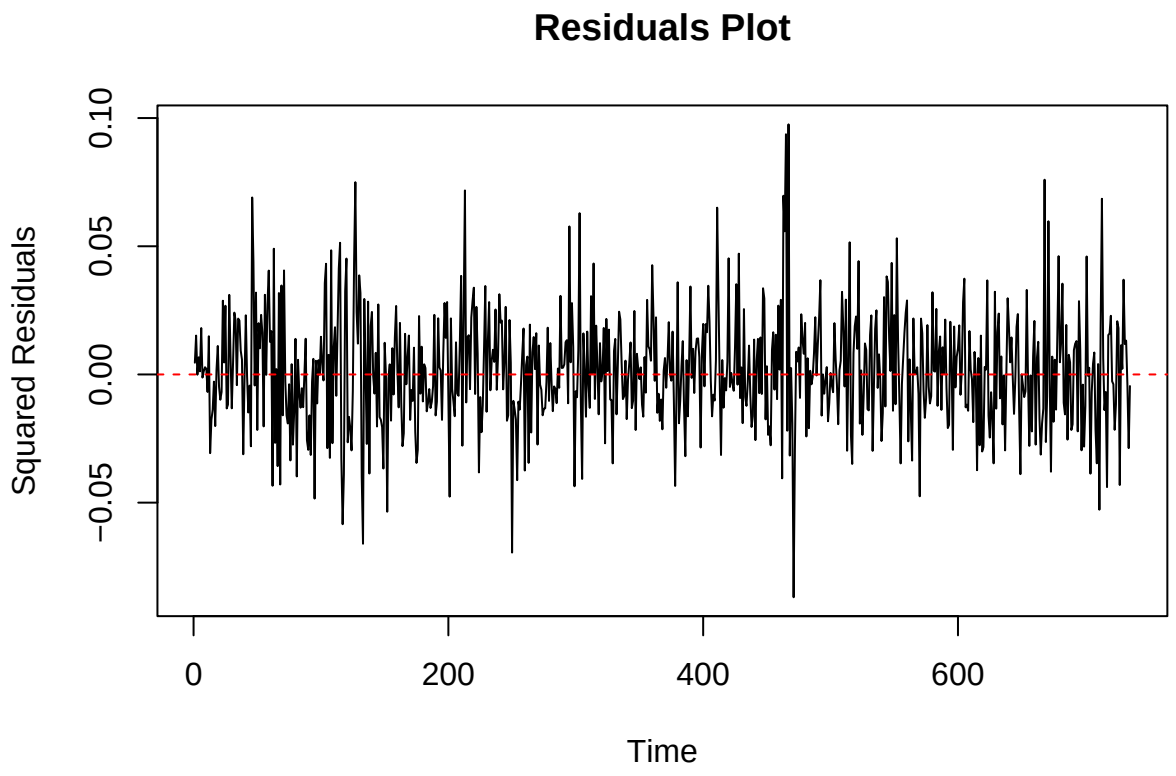
```
                method="ML")
  df[i, ]$AIC <- fit$aic; df[i, ]$BIC <- BIC(fit)
  residuals <- fit$residuals
}
df[order(df$AIC)[1:3], ]
```

```
##    P Q p q       AIC       BIC
## 37 0 0 4 4 -3463.751 -3422.364
## 46 0 0 5 5 -3463.465 -3412.881
## 1  0 0 0 0 -3463.422 -3458.824
```

A SARIMA Model with p = 4, d = 1, q = 4, P = 0, Q = 0, S = 0 with no differencing had the lowest AIC value of -3463.751 , so I will be using this SARIMA Model to forecast.

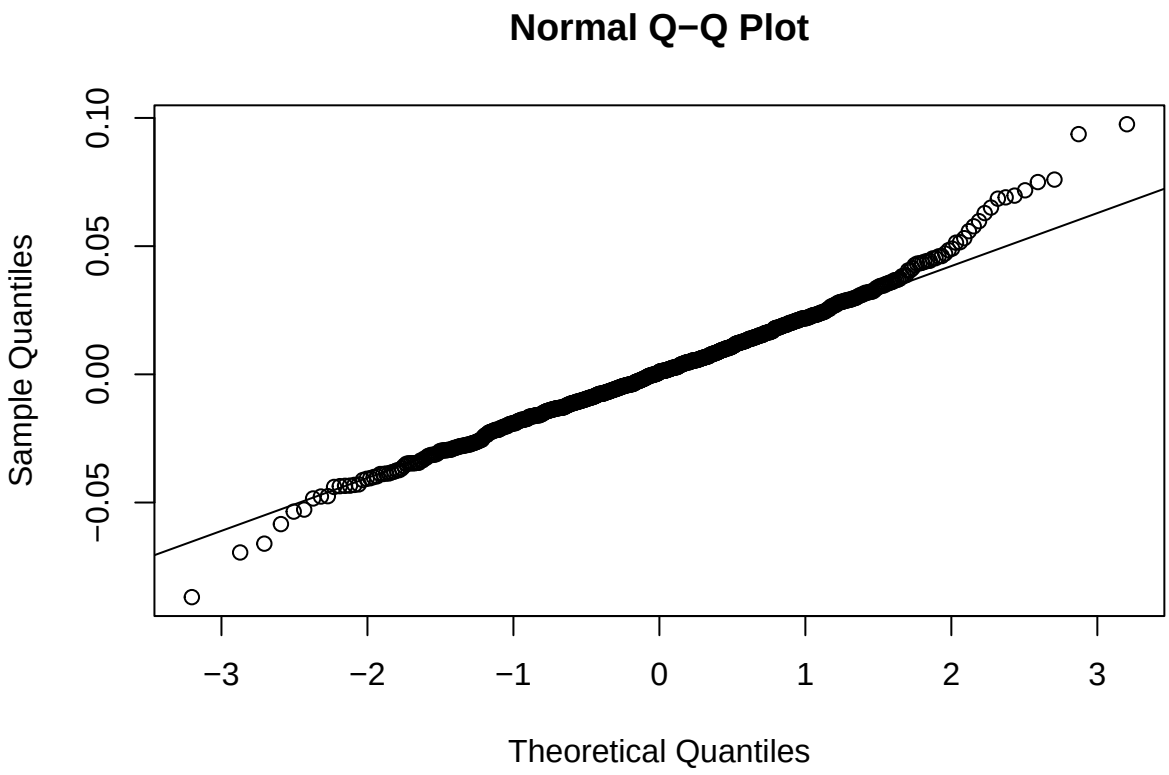
SARIMA Diagnostics

```
# Fitting Model and Calculating Residuals
sarima_model <- arima(x.train, order = c(4, 1, 4))
residuals <- sarima_model$residuals

# Residual Plot
plot(1:length(residuals), residuals, type = 'l', xlab = 'Time', ylab = 'Squared Residuals',
     main = 'Residuals Plot')
abline(h = 0, col = "red", lty = 2)
```



Looking at the residuals plot above there is no apparent pattern or trends which suggests that my model is capturing all the information present in the data. It is also a good sign that the residuals are randomly scattered around zero.

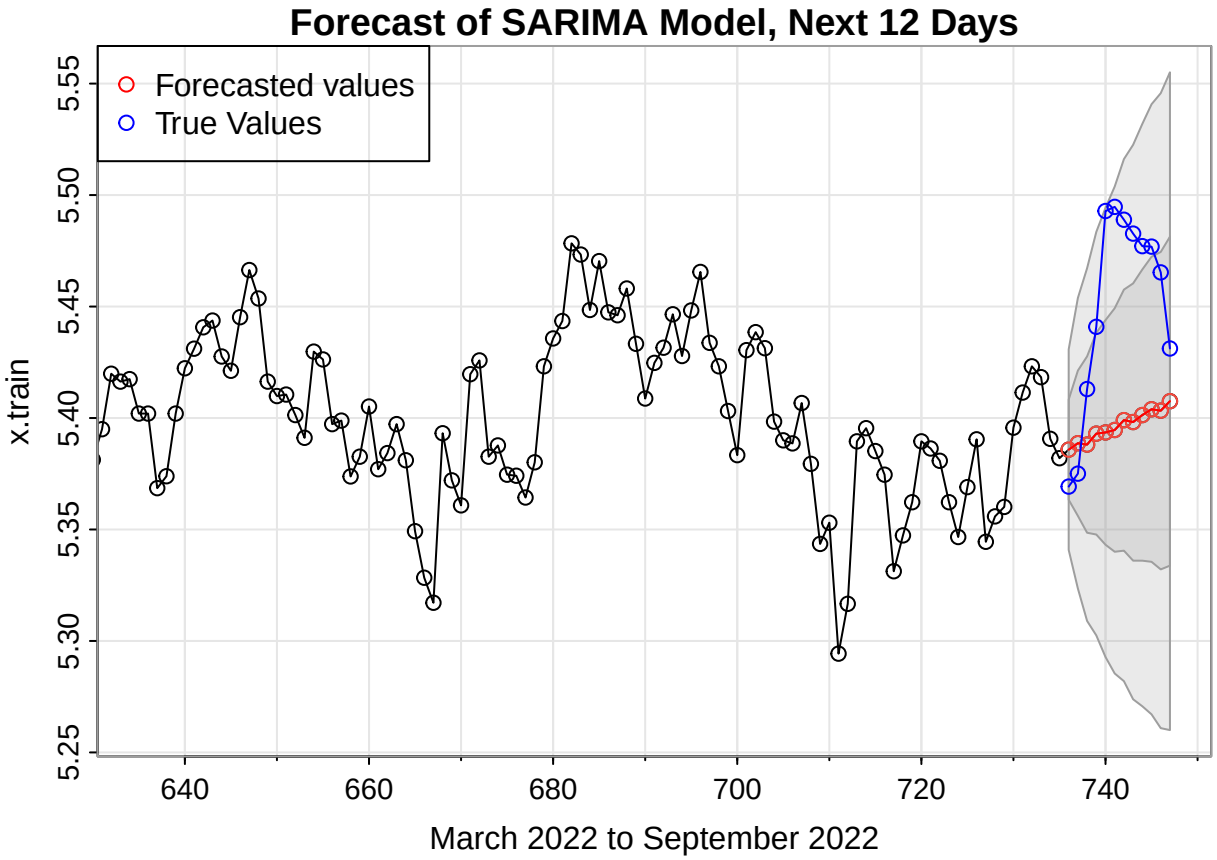
```
# QQ plot
qqnorm(residuals)
qqline(residuals)
```



Taking a look at the Normal Q-Q Plot we can see that my transformed data is close to normal given a vast majority of the observations lie on the linear line. Having close to normal data is very useful for analysis and prediction.

SARIMA Forecasting

```
# Predicting the Coffee Price for the next 12 days
pred.tr <- sarima.for(x.train, n.ahead = 12, plot.all = FALSE,
                      p = 4, d = 1, q = 4, xlab = 'March 2022 to September 2022')
lines((n - 11):n, pred.tr$pred, col = 'red')
lines((n - 11):n, x.test, col = 'blue')
points((n - 11):n, x.test, col = 'blue')
legend('topleft', pch = 1, col = c('red', 'blue'),
legend = c('Forecasted values', 'True Values'))
title('Forecast of SARIMA Model, Next 12 Days')
```



Looking at the graph of my forecasted prices, all of the true data points fell within my confidence interval of my predicted points which is a good sign. A likely reason for this spike in coffee prices during the end of 2022 is a result of inflation. Ever since the COVID-19 pandemic prices in various industries of the economy have significantly increased.

Fitting a GARCH (Generalized Autoregressive Conditional Heteroskedasticity) Model

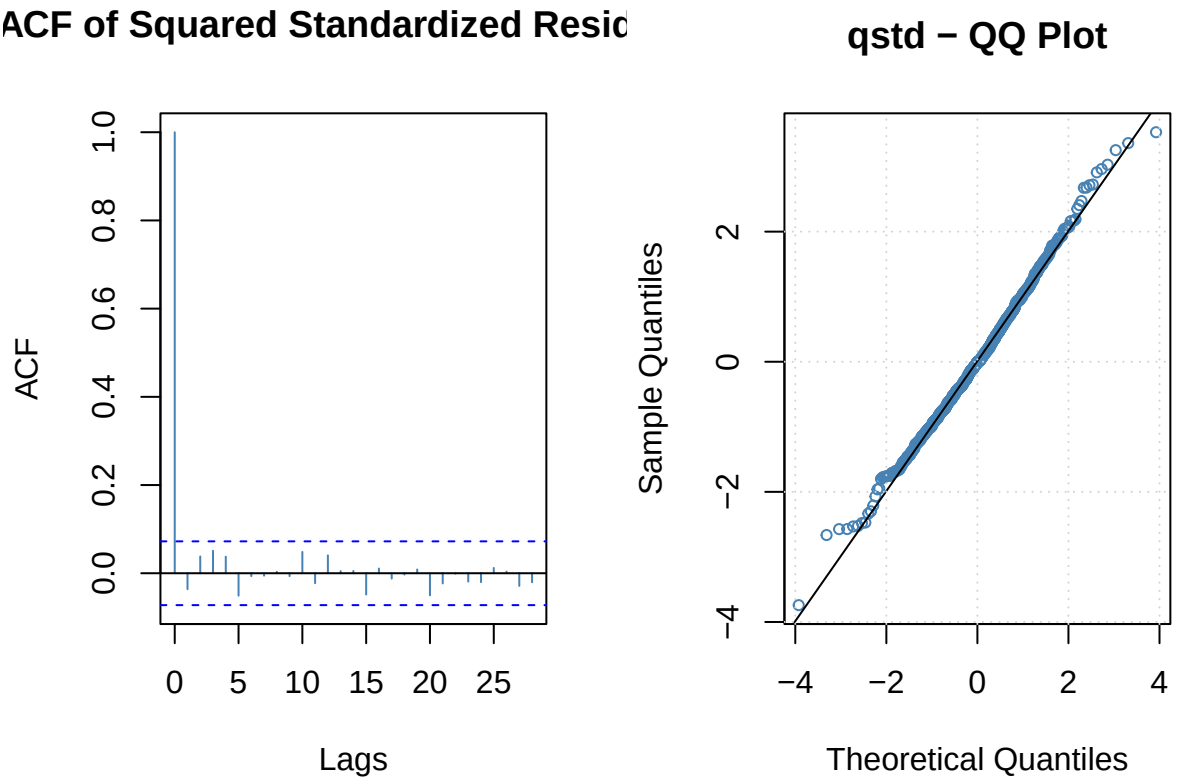
```
# Selecting best Garch Model
auto.arima(x.train, stationary = FALSE)
```

```
## Series: x.train
## ARIMA(0,1,0)
##
## sigma^2 = 0.0005213:  log likelihood = 1732.71
## AIC=-3463.42   AICc=-3463.42   BIC=-3458.82
```

```
# Fitting the model based on AIC
garch_fit <- garchFit(formula = ~ arma(1, 0) + garch(1, 2), x.train, cond.dist = 'std', trace = FALSE)
```

Further GARCH Diagnostics

```
par(mfrow = c(1, 2))
plot(garch_fit, which = c(11, 13))
```

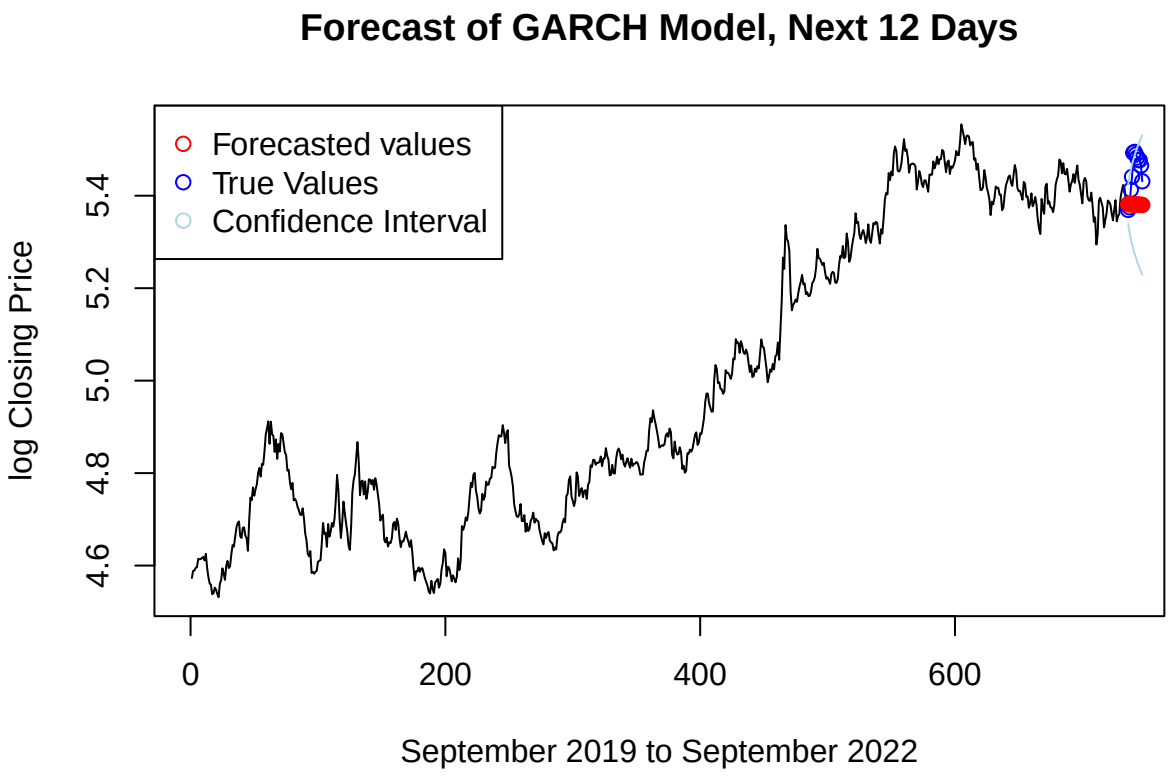


In the graphs above we can see the ACF of the squared standardized residuals only have white noise variation, and the QQ Plot is almost normal which is a good indication for the data.

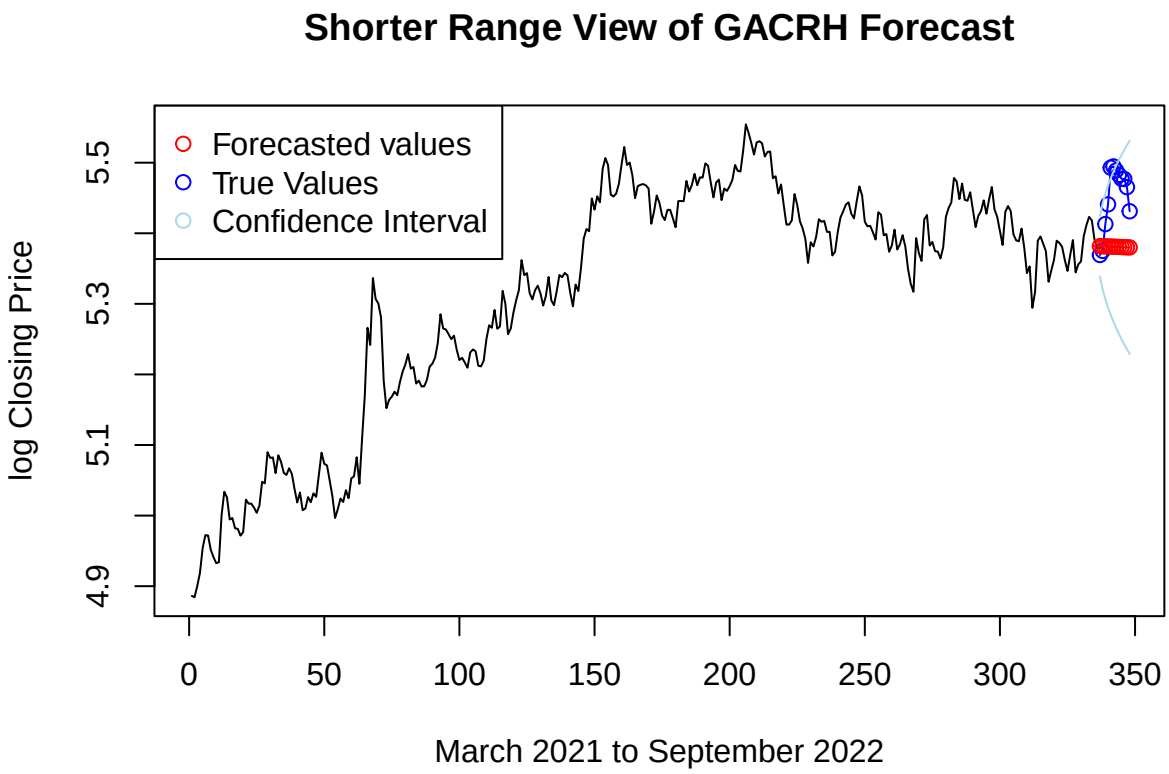
```
garch_for <- predict(garch_fit, n.ahead = 12, plot = TRUE)
```

GARCH Forecasting

```
plot(x.train, type = 'l', ylab = 'log Closing Price', xlab = 'September 2019 to September 2022')
lines((n - 11):n, x.test, col = 'blue') # Want to add dates, x = dates, y =
points((n - 11):n, x.test, col = 'blue')
lines((n - 11):n, garch_for$meanForecast, col = 'red')
points((n - 11):n, garch_for$meanForecast, col = 'red')
lines((n - 11):n, garch_for$upperInterval, col = 'lightblue')
lines((n - 11):n, garch_for$lowerInterval, col = 'lightblue')
legend('topleft', pch = 1, col = c('red', 'blue', 'lightblue'),
legend = c('Forecasted values', 'True Values', 'Confidence Interval'))
title('Forecast of GARCH Model, Next 12 Days')
```




```
# New n for zoomed in plot
n_1 <- length(x.train[400:735]) + length(x.test)

# Shorter range plot
plot(x.train[400:746], type = 'l', ylab = 'log Closing Price', xlab = 'March 2021 to September 2022')
lines((n_1 - 11):n_1, x.test, col = 'blue')
points((n_1 - 11):n_1, x.test, col = 'blue')
lines((n_1 - 11):n_1, garch_for$meanForecast, col = 'red')
points((n_1 - 11):n_1, garch_for$meanForecast, col = 'red')
lines((n_1 - 11):n_1, garch_for$upperInterval, col = 'lightblue')
lines((n_1 - 11):n_1, garch_for$lowerInterval, col = 'lightblue')
legend('topleft', pch = 1, col = c('red', 'blue', 'lightblue'),
legend = c('Forecasted values', 'True Values', 'Confidence Interval'))
title('Shorter Range View of GACRH Forecast')
```



Conclusion and Future Study

After running diagnostics and adequate model selection for my SARIMA and GARCH models I predicted the Daily Closing Stock price of Coffee between August 18, 2022 and September 2, 2022. I originally built a model using all of the past data points dating all the way back to 2002, but that model was very inaccurate. In order to improve my models accuracy I trained them using data from the past year and a half.The predicted points for both my SARIMA and GARCH models were very similar but at the same time they were a bit far off from the actual closing price during the observed days. That being said all of the true closing prices were within the confidence interval for the predicted points as pictures above in the plots. So considering all of the true prices were withing the confidence intervals I can conclude my predictions are successful.

Given the volatility and nature of stock market prices, it is very difficult to precisely predict the price of a future stock which can explain my significantly higher price of the true price as opposed to my predicted price. Additionally ever since COVID-19 inflation has been on the rise which could also be an explanation for the higher than predicted closing price of coffee.

Finally, for future study it would be very beneficial to have a better understanding of economics, the stock market, and specifically the factors which determine the coffee price. With a deeper understanding of Time Series Analysis and the economic factors of coffee, a more accurate model in the short term and long term is in store.

References

1. Time Series Analysis and its Applications with R Examples by H. Shumway and D. S. Stoffer
2. Class Lectures and Section Notes
3. https://www.kaggle.com/datasets/psycon/daily-coffee-price
4. https://www.nasdaq.com/market-activity/commodities/kt:nmx