Stock Prediction Models: Collecting Data using QuantMod, Forecasting with ARIMA

Brandon Huang

Table of Contents

File Structure of post02

Create the following file structure inside post02 folder:

- data

Prologue

The following package installations are required for the reproducibility of this post:

```
install.packages("forecast")
install.packages("ggplot2")
install.packages("quantmod")
```

Use `library()` to load pacakge into current RStudio.

```
library(forecast)
```

```
## Warning: package 'forecast' was built under R version 3.4.2
```

```
## Warning in as.POSIXlt.POSIXct(Sys.time()): unknown timezone 'default/
## America/Los_Angeles'
```

```
library(ggplot2)
library(quantmod)
```

```
## Warning: package 'quantmod' was built under R version 3.4.2
```

```
## Loading required package: xts
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: TTR
```

```
## Version 0.4-0 included new data defaults. See ?getSymbols.
```

Introduction

Motivation

I have been investing in stocks since 2015 ever since my mom showed me her portfolio. I am interested in stocks because of my appeal to the process of developing the unqiue analytical skills needed to profit in the market. The concept of betting is a zero sum game that rewards the winner for his/her confidence in his/her decision. The fundamentals of betting comes down to raising the probability of winning through a person's analytical skills.

Unlike STEM classes' analytical skills which are primarily developed through hard skills, the analytical skills required for a successful investment are much more broad because the information channels for the markets are so broad. For example, to increase the probability of successfully building a Shiny APP, the tools required are Shiny, knowledge of R, and other packages needed for the Server section of Shiny. However, for investing, knowledge to increase the probability of success can be gained through technical skills and soft skills.

I have been primarily investing in Asia stocks through soft skills without much development of technical market skills. In essence, I have traveled back and forth between USA and China so many times that I am able to identify the analogies between the two countries' companies. And since China's economy is still in the process of catching up to USA's, I am able to look at a China company's infrastructure and value to determine which USA company it is most similar too. Then, I identify if the company's security is a good buy or not based on its long term growth.

The motivation behind this blog post is to back up an investment with both technical skills and soft skills. Since I already have developed some soft skills, I will use this post to develop some technical skills. Keep in mind, both skills are analytical so the code chunks in this blog post will be to mainly generate supportive information that makes my arguments for my bets stronger.

Content

I am going to first show you how to download stock data from yahoo finance. Then I will show you how to acquire a company's financials with `quantmod`. Then, I will explain how the equation of ARIMA works so that we have a better grasp of the equation when we plot the data and manipualte the data sets. We will manipulate our stock data so that we can create relavent plots from our data. Finally, we will use ARIMA to predict the stock prices for the next two years.

Data Preparation

Downloading Historical Quote Data

The three stocks I will be analyzing are:

1. https://finance.yahoo.com/quote/TCEHY/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo
2. https://finance.yahoo.com/quote/GELYF/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo
3. https://finance.yahoo.com/quote/BZUN/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo

To download the csv for each URL:

1. Enter the link into the browser.
2. Select a time period of 5Y and Apply.
3. Click on Download Data.
4. Name the file as the stock symbol. Example: For tcehy, name the file as "tcehy.csv".
5. Download the file into the data folder.
6. Import each dataset into R Studio.

```
#setwd to your data folder
setwd("/Users/brandon/Desktop/stat133/Post/post02/data")

#import tcehy data
tcehy <- read.csv("tcehy.csv")

#import gelyf data
gelyf <- read.csv("gelyf.csv")

#import bzun data
bzun <- read.csv("bzun.csv")
```

Collecting Financials with QuantMod.

Quantmod downloads income statements, balance sheets, and cash flow statements from Google Finance. tcehy and gelyf's financials are not available on Google Finance so this portion will only be for bzun. The downloaded container for the financial statements is a custom S3 object of type "financials".

```
#download financials of baozun
getFinancials("bzun")
```

```
## [1] "bzun.f"
```

`viewFinancials()` is a function from the package `quantmod`. `viewFinancials()` is comprised of three elements.

1. custom S3 object of type “financials”.
2. type = c(‘BS’, ‘IS’, ‘CF’)
    - BS = Balance Sheet
    - IS = Income Statement
    - CF = Cash Flow
3. period = c(‘A’, ‘Q’)
    - A = Annual
    - Q = Quarterly

Lets extract the annual balance sheet for bzun into a dataframe so that we can manipulate the data or print it into an app.

```
#extract annual balance sheet into a data frame
bzunBS <- data.frame(viewFinancials(bzun.f, type = ('BS'), period = "A"))
```

```
## Annual Balance Sheet for bzun
```

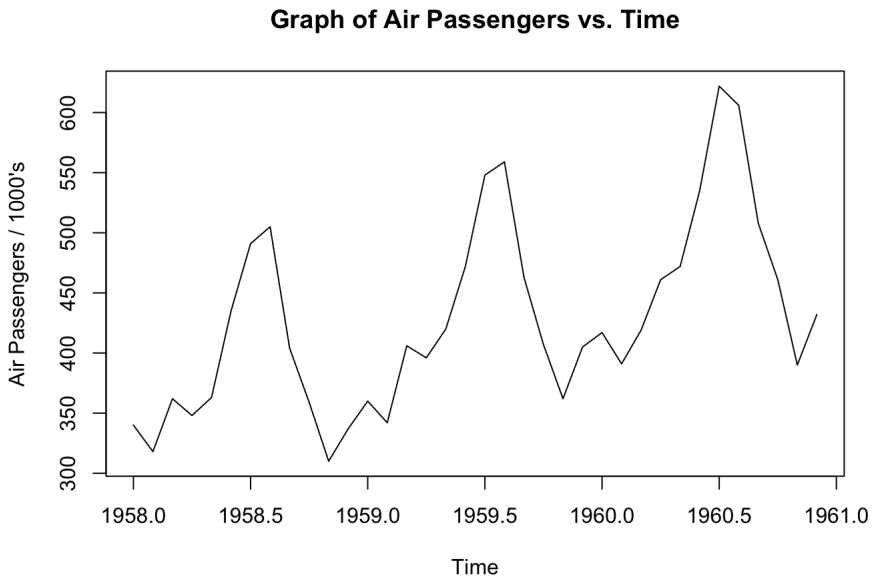
```
#omit rows with NA in bzunBS
bzunBS <- na.omit(bzunBS)

#write bzun into data folder
write.csv(bzunBS, file = 'bzunBS.csv')
```

ARIMA: Autoregressive Integrated Moving Average

ARIMA is a time series equation that predicts stock prices over time. In general, time series model relationships using data collected over time. Time series involves decomposition of the data into a trend, seasonal, cyclical, or irregular component. The decomposition depends on the patterns found in the data. For example, air passengers travel due to seasonality.

For example, the graph of air passengers vs. time is the decomposition of the data into seasons because a large proportion of air passengers decide to travel due to seasons. For example, during thanksgiving and winter, more people travel due to holidays. The following data is not reproducible because it is an example of what a seasonal decomposition would look like and it is not needed for our post.

```
setwd("/Users/brandon/Desktop/stat133/Post/post01/Data")
cbe <- read.table(file = 'cbe.dat', header = T) #create matrix cbe

data(AirPassengers) #acquire airpassengers
AP <- AirPassengers

elec_ts <- ts(cbe$elec, start = 1958, freq = 12) #create ts for electricity

#time series: intersect air passengers and electricity
ap_elec <- ts.intersect(AP, elec_ts)

plot(ap_elec[ , 1],
     ylab = "Air Passengers / 1000's",
     main = 'Graph of Air Passengers vs. Time') #plot x = time, y = air passengers
```

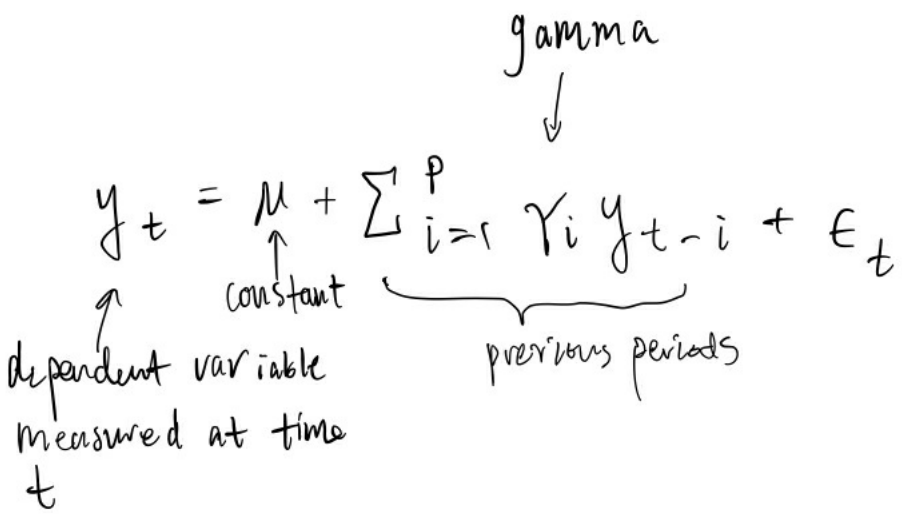


Lags

Lags are a very important aspect of regressive models because regressive models should not only take into account the independent and dependent variables but also “lag”. Lag is the relation between current and past data. For example, when we compute the stock prediction; we

must also take into account the relation of current price compared to the past price and not only the price (dependent variable) and the time (independent variable). In essence, a regression equation that does not take lag into account is overestimating the weights of the independent and dependent variable. Therefore, a regression without lag would not be as accurate. In the following segments, I will specifically note where the equation takes lag into the equation.

AutoRegressive

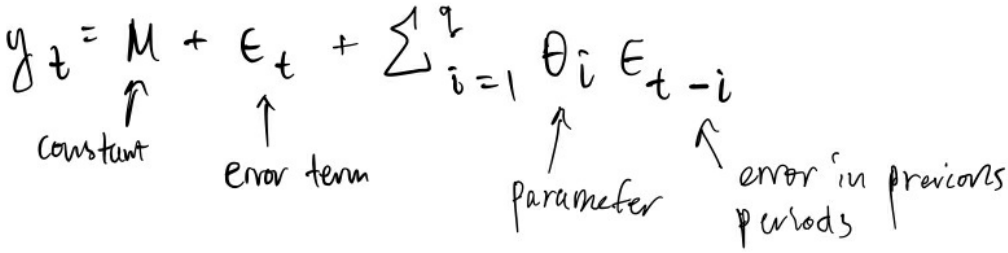


AR(p) is an AR model with p lags

“p” is the past periods that we need to take into account to predict future prices. “p” realizes that the value in one period is related to the value in a previous period. The \(\gamma\) in the equation denotes how much of the past data to retain. For example, if the \(\gamma\) was set at 0.8, then the next prediction would have to retain 0.8 of the past data. Therefore, if the \(\gamma\) was -0.8, would the data oscillate more or less compared to 0.8? If you answered oscillate more, then you are correct because -0.8 would always mean you predict in the opposite direction because you are retaining the opposite of the previous data.

Moving Average

The moving average is the average of the data throughout a certain period of time. This is very important in forecasting situations and is widely used in stock markets for technical analysis. When using moving averages in stock markets, you want to set a couple different moving averages that takes into account different time periods. For example, you want to set one at 20, 100, and 180 so that the moving averages are broader and you can see the trend for each different time period. If the stock is rapidly oscillating but overall performing at a positive rate, then the 20 moving average would follow the oscillation of the stock but the 180 moving average would have a positive slope because in the long run, the stock will remain positive in earnings.



MA(q) is a MA model with q lags

AutoRegressive and Moving Average

To obtain ARIMA, we combine the AR and MA equations.

$$y_t = \mu + \sum_{i=1}^{p} \gamma_i y_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

AutoRegressive and Moving Average Equation

Predicting Stock Prices with ARIMA

1. Convert stock price data to time series. The “start” in `ts()`

refers to the starting date for the stock prices. For example, the tcehy data starts in December so the start is 2012 and 0 for December. The bzun data starts in April so the start is 2012 and 4 for April. “freq” denotes that there are 12 months in one year.

```
# convert data frame of tcehy to time series
tcehy.ts <- ts(tcehy$Close, start = c(2012, 0), freq = 12)

# convert data frame of tcehy to time series
gelyf.ts <- ts(gelyf$Close, start = c(2012, 0), freq = 12)

# convert data frame of tcehy to time series
bzun.ts <- ts(bzun$Close, start = c(2015, 7), freq = 12)
```

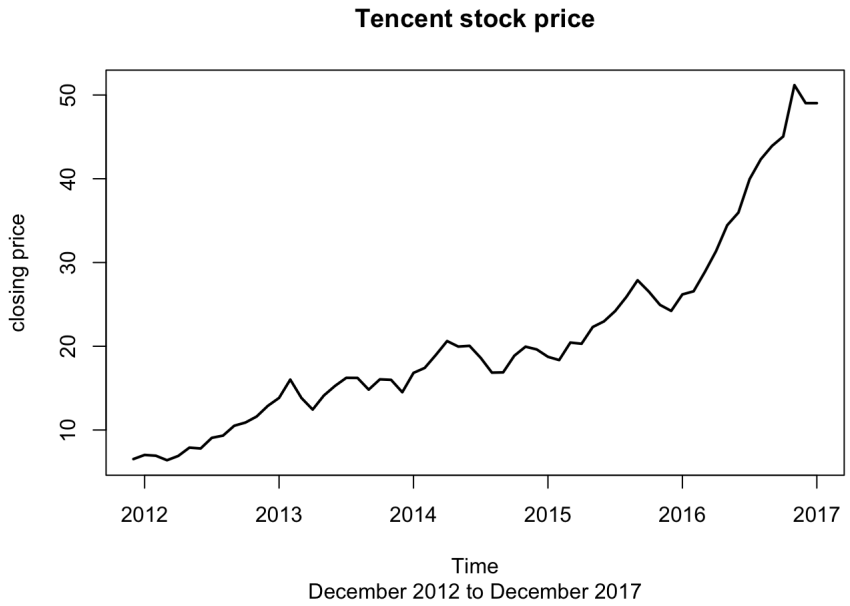
2. Convert each stock data into a data frame with closing prices and

log of closing prices. This is so that when we graph the values, the x axis is the actual date of the closing stock prices and not the increasing values of the number of months.

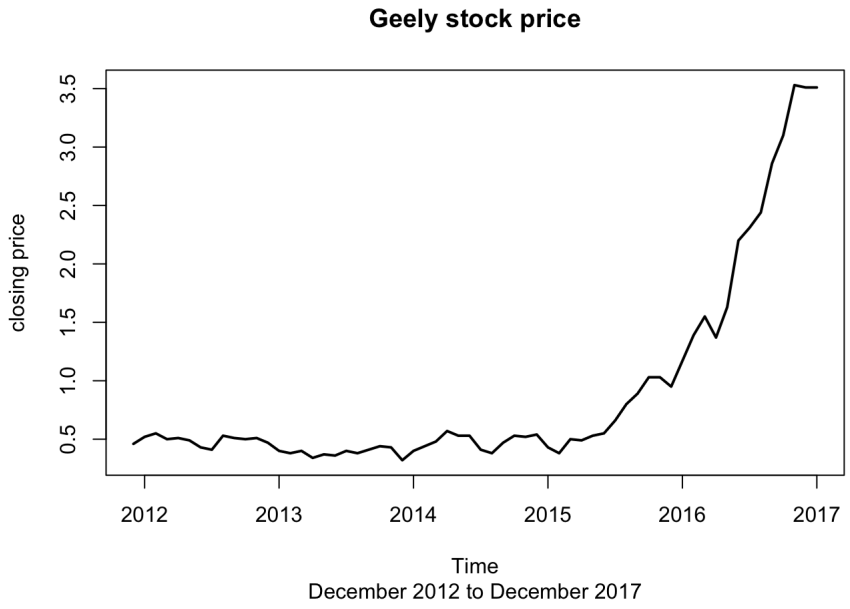
```
# convert tcehy into a data frame with log of closing prices and closing prices
tcehy <- data.frame(closing = tcehy.ts, lclosing = log(tcehy.ts))
save(tcehy, file = "tcehy.df.RData")
load("tcehy.df.RData")

# convert gelyf into a data frame with log of closing prices and closing prices
gelyf <- data.frame(closing = gelyf.ts, lclosing = log(gelyf.ts))
save(gelyf, file = "gelyf.df.RData")
load("gelyf.df.RData")

# convert bzun into a data frame with log of closing prices and closing prices
bzun <- data.frame(closing = bzun.ts, lclosing = log(bzun.ts))
save(bzun, file = "bzun.df.RData")
load("bzun.df.RData")
```

3. Plot the stock values

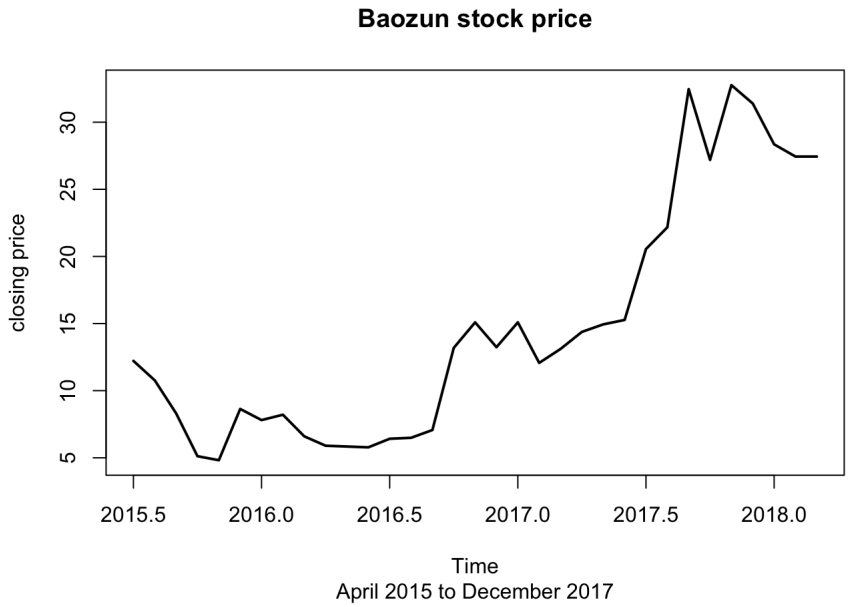
```
# line plot for tcehy
plot(tcehy$closing,
     main = "Tencent stock price",
     lwd = 2,
     type = "l",
     sub = "December 2012 to December 2017",
     ylab = "closing price")
```




```
# line plot for gelyf
plot(gelyf$closing,
     main = "Geely stock price",
     lwd = 2,
     type = "l",
     sub = "December 2012 to December 2017",
     ylab = "closing price")
```



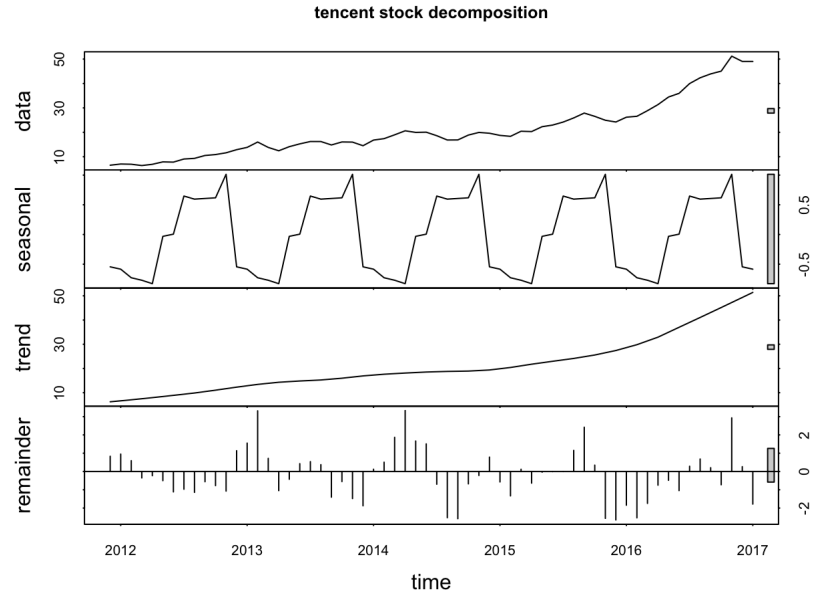

```
# line plot for bzun
plot(bzun$closing,
     main = "Baozun stock price",
     lwd = 2,
     type = "l",
     sub = "April 2015 to December 2017",
     ylab = "closing price")
```



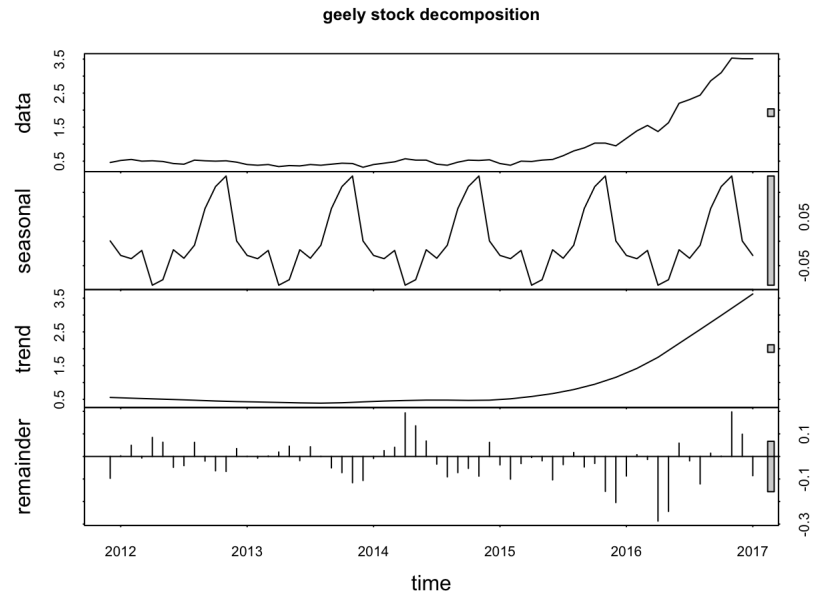
5. Decomposition: Earlier we talked about the decomposition of a dataset.

We can decompose each stock data set according to its seasonal, trend, and remainder patterns. `stl()` allows us to create a data frame with the patterns.

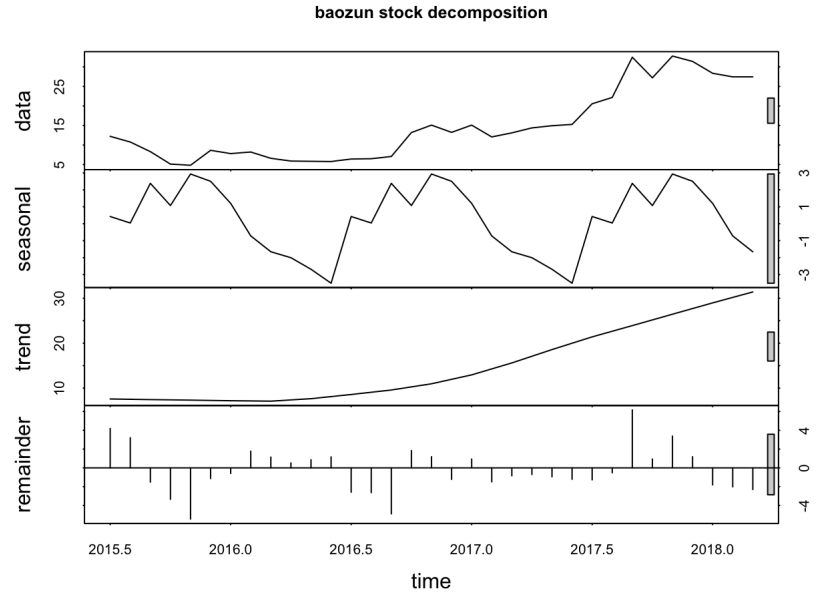
```
# decompose tcehy data set
tcehy.stl = stl(tcehy$closing, s.window = "periodic")
# plot tcehy stock decomposition data
plot(tcehy.stl, main = "tencent stock decomposition")
```




```
# decompose tcehy data set
gelyf.stl = stl(gelyf$closing, s.window = "periodic")
# plot tcehy stock decomposition data
plot(gelyf.stl, main = "geely stock decomposition")
```

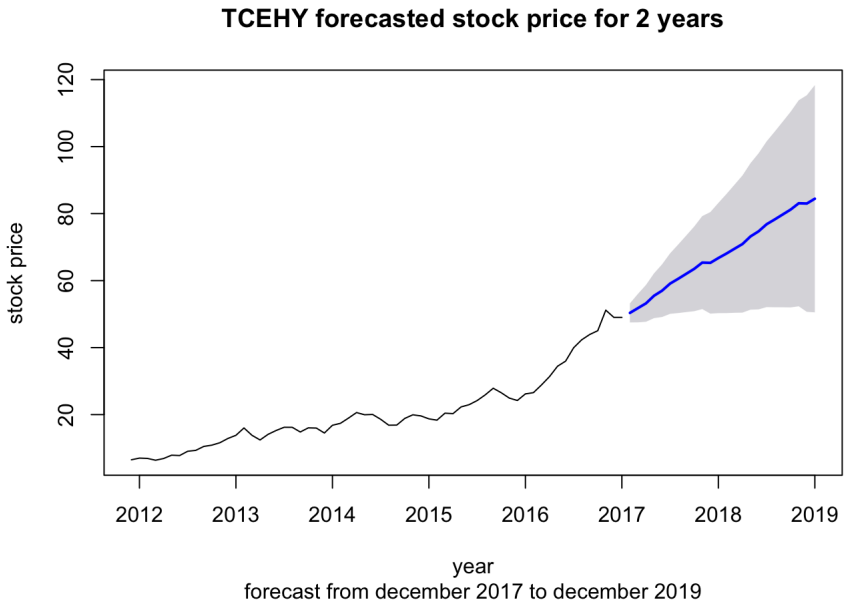



```
# decompose tcehy data set
bzun.stl = stl(bzun$closing, s.window = "periodic")
# plot tcehy stock decomposition data
plot(bzun.stl, main = "baozun stock decomposition")
```

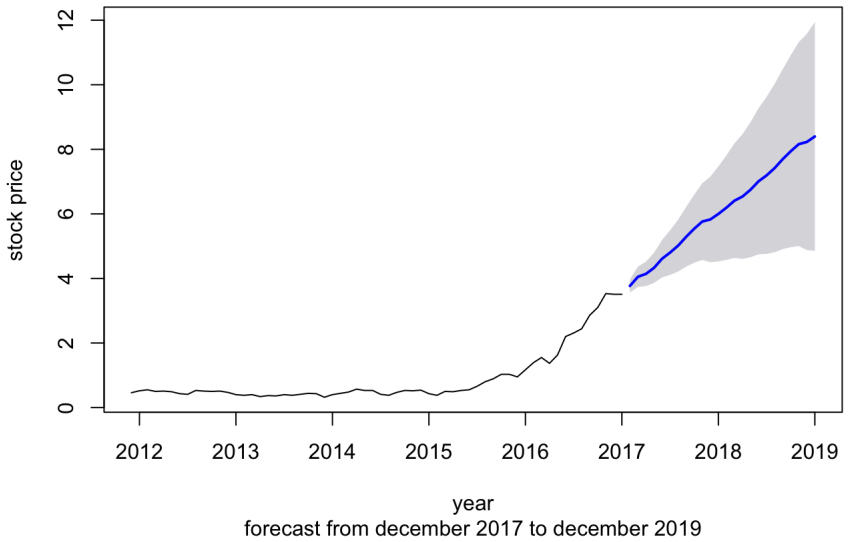


6. Forecasting

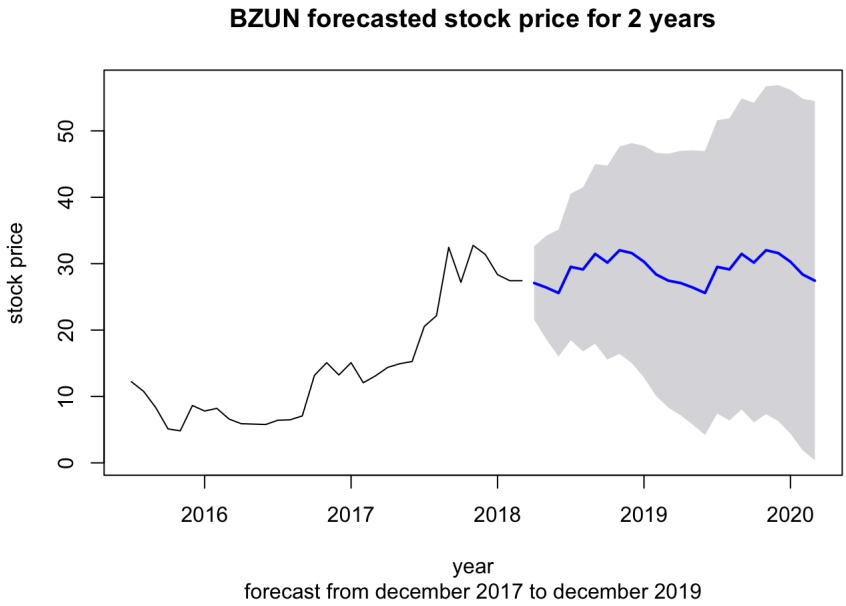
```
# TENCENT
# h = 24 is 24 months which is 2 years, 95% confidence interval
tcehy.f = forecast(tcehy.stl,
                   method = "arima",
                   h = 24,
                   level = 95)
# plot the forecasted values
plot(tcehy.f,
     ylab = "stock price",
     xlab = "year",
     main = "TCEHY forecasted stock price for 2 years",
     sub = "forecast from december 2017 to december 2019")
```




```
#GEELY
# h = 24 is 24 months which is 2 years, 95% confidence interval
gelyf.f = forecast(gelyf.stl,
                   method = "arima",
                   h = 24,
                   level = 95)
# plot the forecasted values
plot(gelyf.f,
     ylab = "stock price",
     xlab = "year",
     main = "GELYF forecasted stock price for 2 years",
     sub = "forecast from december 2017 to december 2019")
```




```
#BAOZUN
# h = 24 is 24 months which is 2 years, 95% confidence interval
bzun.f = forecast(bzun.stl,
                   method = "arima",
                   h = 24,
                   level = 95)
# plot the forecasted values
plot(bzun.f,
     ylab = "stock price",
     xlab = "year",
     main = "BZUN forecasted stock price for 2 years",
     sub = "forecast from december 2017 to december 2019")
```




```
#display baozun's financials too
print(bzunBS)
```

```
##                                          X2016.12.31 X2015.12.31
## Cash & Equivalents                            917.32      787.26
## Cash and Short Term Investments               957.32      837.26
## Accounts Receivable - Trade, Net              663.59      402.35
## Total Receivables, Net                        683.49      419.73
## Total Inventory                               312.07      334.35
## Prepaid Expenses                              164.32      124.72
## Other Current Assets, Total                    50.83       66.65
## Total Current Assets                         2168.03     1782.70
## Property/Plant/Equipment, Total - Gross       155.66       87.07
## Accumulated Depreciation, Total               -54.77      -27.86
## Intangibles, Net                               26.98       20.13
## Long Term Investments                          33.44       13.31
## Other Long Term Assets, Total                  38.91       13.83
## Total Assets                                 2368.26     1889.17
## Accounts Payable                              526.46      464.96
## Accrued Expenses                              110.35      120.69
## Notes Payable/Short Term Debt                 115.14       31.09
## Other Current liabilities, Total               44.30       37.96
## Total Current Liabilities                     796.25      654.70
## Total Long Term Debt                            0.00        0.00
## Total Debt                                    115.14       31.09
## Total Liabilities                             796.25      654.70
## Common Stock, Total                             0.10        0.09
## Additional Paid-In Capital                   1761.43     1535.66
## Retained Earnings (Accumulated Deficit)      -233.87     -320.50
## Other Equity, Total                            44.35       19.21
## Total Equity                                 1572.01     1234.47
## Total Liabilities & Shareholders&#39; Equity 2368.26     1889.17
## Total Common Shares Outstanding               159.41      151.47
##                                          X2014.12.31 X2013.12.31
## Cash & Equivalents                            206.39      154.16
## Cash and Short Term Investments               206.39      154.16
## Accounts Receivable - Trade, Net              244.65      113.59
## Total Receivables, Net                        269.04      121.20
## Total Inventory                               242.98      133.35
## Prepaid Expenses                               59.47       49.19
## Other Current Assets, Total                    41.68       37.24
## Total Current Assets                          819.56      495.14
## Property/Plant/Equipment, Total - Gross        45.79       27.47
## Accumulated Depreciation, Total               -15.57       -8.13
## Intangibles, Net                               14.67        9.90
## Long Term Investments                           5.62        5.62
## Other Long Term Assets, Total                   2.44        1.45
## Total Assets                                  872.51      531.45
## Accounts Payable                              307.48      173.81
## Accrued Expenses                               46.74       37.42
## Notes Payable/Short Term Debt                  17.00        0.00
## Other Current liabilities, Total               22.25       13.85
## Total Current Liabilities                     393.46      225.08
## Total Long Term Debt                            0.00        0.00
## Total Debt                                     17.00        0.00
## Total Liabilities                             393.46      225.08
## Common Stock, Total                             0.02        0.02
## Additional Paid-In Capital                      3.75       -0.02
## Retained Earnings (Accumulated Deficit)      -327.20     -232.33
## Other Equity, Total                             1.20       -0.04
## Total Equity                                  479.06      306.37
## Total Liabilities & Shareholders&#39; Equity  872.51      531.45
## Total Common Shares Outstanding               112.70      112.70
```

Conclusion

Take Home Message

The ARIMA forecast show that among the three stocks in my portfolio: TCEHY, GELYF, and BZUN, I should retain TCEHY and GELYF and sell BZUN because BZUN shows no rate of growth.

To conclude, without the technical analysis of my portfolio, I would have kept BZUN for an X amount of time and it would have showed no signs of growth but I would not know what to do with the stock because I do not have a technical analysis of the stock. The take home message is that technical skills is just as important as soft skills when it comes to investing and having both is the most important aspect of investing successfully. Statistics plays a huge role in predicting the future with current and past data. Therefore, it is very important to integrate statistics into anything that can benefit from time-series analysis.

References

1. https://finance.yahoo.com/quote/TCEHY/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo

2. https://finance.yahoo.com/quote/GELYF/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo
3. https://finance.yahoo.com/quote/BZUN/history?period1=1354521600&period2=1512288000&interval=1mo&filter=history&frequency=1mo
4. https://www.linkedin.com/pulse/using-r-easily-bulk-scrape-financial-statements-matt-lunkes/
5. https://www.youtube.com/watch?v=10cuDKGytMw
6. https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials
7. https://www.r-bloggers.com/forecasting-stock-returns-using-arima-model/
8. https://pocfarm.wordpress.com/2016/06/15/share-price-prediction-using-r/
9. http://rmarkdown.rstudio.com/authoring_basics.html
10. https://www.youtube.com/watch?v=FYwTl9s4IVc