Stock Market Analysis

Using Data Science techniques in Excel and R

Mentor: Prof. Soumyakanti Chakraborty (XLRI)

Abstract

Project by:

Bhavdeep Singh Sachdeva

Acknowledgment

I would to like to thank Professor Soumyakanti Chakraborty Sir for his constant guidance and encouragement. He is an exceptional teacher with great passion for the subject and is a very good mentor. His invaluable support has strengthened my confidence and knowledge in the field of data science.

I would also like to thank the makers of analyticsvidhya.com, machinelearningmastery, and stackoverflow.com without whose extraordinary repertoire of blogs and answers I would not have been able to learn so much and complete the project.

Many thanks to the talent edge team for excellent transmission quality on the online classes and for the prompt uploading of the recording sessions.

Table of Contents

1 Introduction

1.1 Stock Market

Stocks – sometimes referred to as equity or equities – are issued by companies to raise capital in order to grow the business or undertake new projects. There are important distinctions between whether somebody buys shares directly from the company when it issues them (in the primary market) or from another shareholder (on the secondary market). When the corporation issues shares, it does so in return for money.

Stock Market, equity market or share market aggregation of buyers and sellers (a loose network of economic transactions, not a physical facility or discrete entity) of stocks (also called shares), which represent ownership claims on businesses; these may include securities listed on a public stock exchange as well as those only traded privately. The trends in the in the prices can be analysed to in order to make more informed decisions. There are two most important types of market that we need to keep in mind: "Primary" capital market and "Secondary" capital market.

When a company publicly sells new stocks and bonds for the first time, it does so in the primary capital market. In many cases, this takes the form of an initial public offering (IPO). When investors purchase securities on the primary capital market, the company offering the securities has already hired an underwriting firm to review the offering and create a prospectus outlining the price and other details of the securities to be issued.

The secondary market is where securities are traded after the company has sold all the stocks and bonds offered on the primary market. Markets such as the New York Stock Exchange (NYSE), London Stock Exchange or NASDAQ are secondary markets. On the secondary market, small investors have a better chance of buying or selling securities, because they are no longer excluded from IPOs due to the small amount of money they represent. Anyone can purchase securities on the secondary market as long as they are willing to pay the price for which the security is being traded.

1.2 The two analytical models for Stock analysis

There are many tools out there to analyze the markets. Majorly there are two basic methodologies investors rely upon when the objective of the analysis is to determine what stock to buy and at what price:

- Fundamental analysis is a method of evaluating securities by attempting to measure the intrinsic value of a stock. Fundamental analysts study everything from the overall economy and industry conditions to the financial condition and management of companies.
- Technical analysis is the evaluation of securities by means of studying statistics generated by market activity, such as past prices and volume. Technical analysts do not attempt to measure a security's intrinsic value but instead use stock charts to identify patterns and trends that may suggest what a stock will do in the future.

In the world of stock analysis, fundamental and technical analysis are on completely opposite sides of the spectrum. Earnings, expenses, assets and liabilities are all important characteristics to fundamental analysts, whereas technical analysts could not care less about these numbers.

Fundamental analysts examine earnings, dividends, assets, quality, ratio, new products, research and the like. Technicians employ many methods, tools and techniques as well, one of which is the use of charts. Using charts, technical analysts seek to identify price patterns and market trends in financial markets and attempt to exploit those patterns.

Technicians using charts search for archetypal price chart patterns, such as the well-known head and shoulders or double top/bottom reversal patterns, study technical indicators, moving averages, and look for forms such as lines of support, resistance, channels, and more obscure formations such as flags, pennants, balance days and cup and handle patterns.

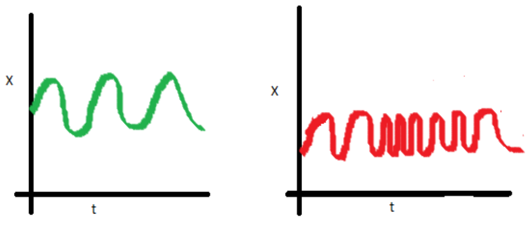
Technical analysts also widely use market indicators of many sorts, some of which are mathematical transformations of price, often including up and down volume, advance/decline data and other inputs. These indicators are used to help assess whether an asset is trending, and if it is, the probability of its direction and of continuation. Technicians also look for relationships between price/volume indices and market indicators. Examples include the moving average, relative strength index, and MACD.

2 Time Series Modelling

2.1.1 Stationary Series

There are three basic criterion for a series to be classified as stationary series:

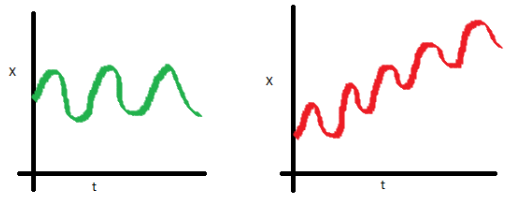
1. The mean of the series should not be a function of time rather should be a constant. The image below has the left hand graph satisfying the condition whereas the graph in red has a time dependent mean.



Stationary series

Non-Stationary series

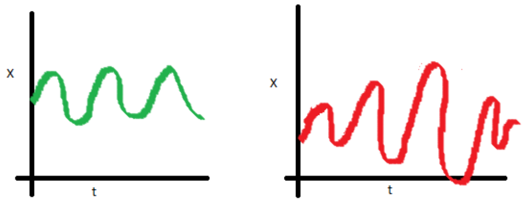
2. The variance of the series should not be a function of time. This property is known as homoscedasticity. Following graph depicts what is and what is not a stationary series. (Notice the varying spread of distribution in the right hand graph)



Stationary series

Non-Stationary series

3. The covariance of the i th term and the (i + m) th term should not be a function of time. In the following graph, you will notice the spread becomes closer as the time increases. Hence, the covariance is not constant with time for the 'red series'.



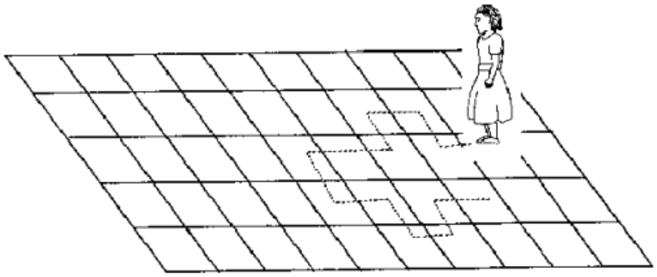
Stationary series

Non-Stationary series

Unless a time series is stationary, a time series model cannot be made on top of it. In cases where the stationary criterion are violated, the first requisite becomes to stationarize the time series and then try stochastic models to predict this time series. There are multiple ways of bringing this stationarity. Some of them are Detrending, Differencing etc.

2.2 Random Walk

Imagine, you are sitting in another room and are not able to see the girl. You want to predict the position of the girl with time. How accurate will you be? Of course you will become more and more inaccurate as the position of the girl changes. At t=0 you exactly know where the girl is. Next time, she can only move to 8 squares and hence your probability dips to 1/8 instead of 1 and it keeps on going down. Now let’s try to formulate this series:



X(t) = X(t-1) + Er(t)

Where Er(t) is the error at time point t. This is the randomness the girl brings at every point in time.

Now, if we recursively fit in all the Xs, we will finally end up to the following equation:

X(t) = X(0) + Sum(Er(1),Er(2),Er(3).....Er(t))

Because of the fact that Expectation of any Error will be zero as it is random, the mean is not dependent on time still the random walk is not a stationary process as it has a time variant variance and covariance.

Stationary testing and converting a series into a stationary series are the most critical processes in a time series modelling.

2.3 Basic Concept of Trends

The concept of trend is absolutely essential to the technical approach to market analysis. All of the tools used by the chartist— support and resistance levels, price patterns, moving averages, trend lines, etc.— have the sole purpose of helping to measure the trend of the market for the purpose of participating in that trend. We often hear such familiar expressions as "always trade in the direction of the trend," "never buck the trend," or "the trend is your friend."

In a general sense, the trend is simply the direction of the market, which way it's moving. But we need a more precise definition with which to work. First of all, markets don't generally move in a straight line in any direction. Market moves are characterized by a series of zigzags. These zigzags resemble a series of successive waves with fairly obvious peaks and troughs. It is the direction of those peaks and troughs that constitutes market trend. Whether those peaks and troughs are moving up, down, or sideways tells us the trend of the market. An uptrend would be defined as a series of successively higher peaks and troughs; a downtrend is just the opposite, a series of declining peaks and troughs; horizontal peaks and troughs would identify a sideways price trend.

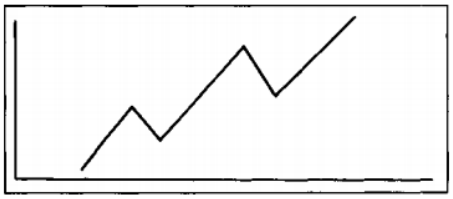


Figure 1 Example of an uptrend with ascending peaks and troughs.

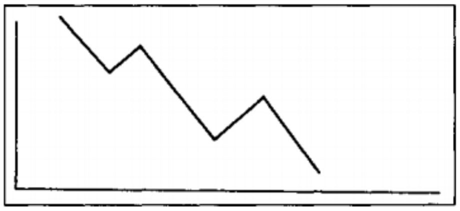


Figure 2 Example of a downtrend with ascending peaks and troughs.

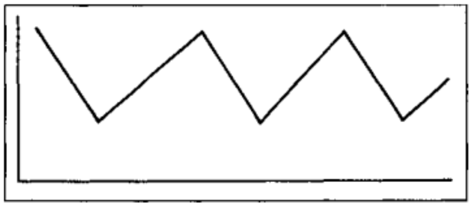


Figure 3 Example of a sideways trend with horizontal peaks and troughs. This type of a market is often referred to as "trendless"

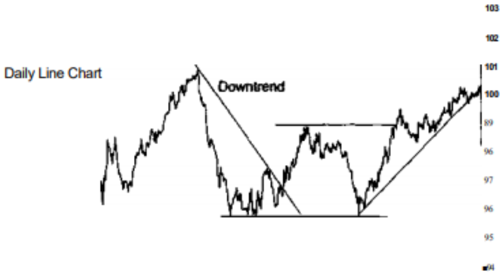


Figure 4 Example of a downtrend turning into an uptrend.

2.4 Moving Averages

2.4.1 INTRODUCTION

The moving average is one of the most versatile and widely used of all technical indicators. Because of the way it is constructed and the fact that it can be so easily quantified and tested, it is the basis for many mechanical trend-following systems in use today. Chart analysis is largely subjective and difficult to test. As a result, chart analysis does not lend itself that well to computerization. Moving average rules, by contrast, can easily be programmed into a computer, which then generates specific buy and sell signals.

As the second word implies, it is an average of a certain body of data. For example, if a 10 day average of closing prices is desired, the prices for the last 10 days are added up and the total is divided by 10. The term moving is used because only the latest 10 days' prices are used in the calculation. Therefore, the body of data to be averaged (the last 10 closing prices) moves forward with each new trading day. The most common way to calculate the moving average is to work from the total of the last 10 days' closing prices. Each day the new close is added to the total and the close 11 days back is subtracted. The new total is then divided by the number of days (10).

2.4.2 The Simple Moving Average

The simple moving average, or the arithmetic mean, is the type used by most technical analysts. But there are some who question its usefulness on two points. The first criticism is that only the period covered by the average (the last 10 days, for example) is taken into account. The second criticism is that the simple moving average gives equal weight to each day's price. In a 10 day average, the last day receives the same weight as the first day in the calculation. Each day's price is assigned a 10% weighting. In a 5 day average, each day would have an equal 20% weighting. Some analysts believe that a heavier weighting should be given to the more recent price action.

2.4.3 The Linearly Weighted Moving Average

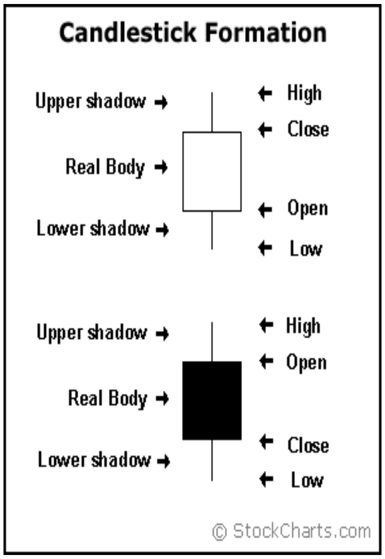
In an attempt to correct the weighting problem, some analysts employ a linearly weighted moving average. In this calculation, the closing price of the 10th day (in the case of a 10 day average) would be multiplied by 10, the ninth day by nine, the eighth day by eight, and so on. The greater weight is therefore given to the more recent closings. The total is then divided by the sum of the multipliers (55 in the case of the 10 day average: 10 + 9 + 8 + . . . + 1). However, the linearly weighted average still does not address the problem of including only the price action covered by the length of the average itself.

2.4.4 The Exponentially Smoothed Moving Average

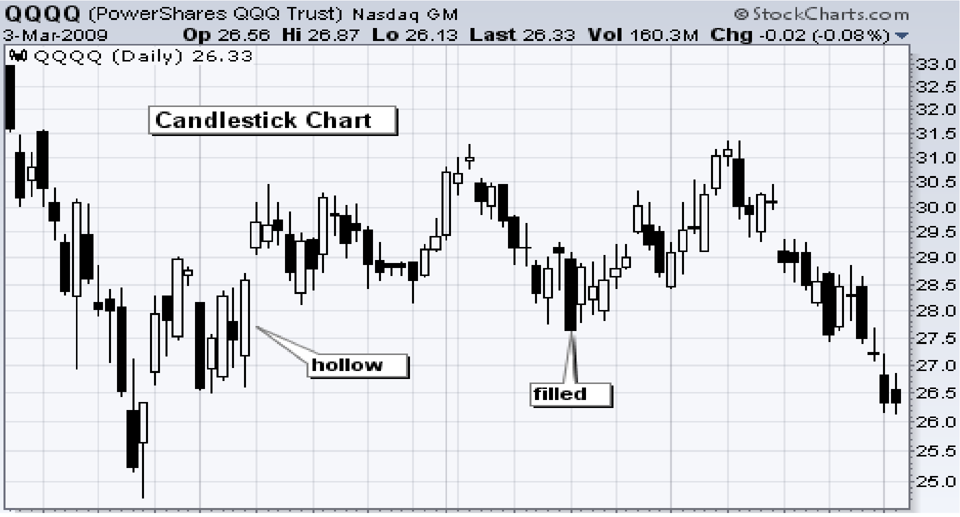
This type of average addresses both of the problems associated with the simple moving average. First, the exponentially smoothed average assigns a greater weight to the more recent data. Therefore, it is a weighted moving average. But while it assigns lesser importance to past price data, it does include in its calculation all of the data in the life of the instrument. In addition, the user is able to adjust the weighting to give greater or lesser weight to the most recent day's price. This is done by assigning a percentage value to the last day's price, which is added to a percentage of the previous day's value. The sum of both percentage values adds up to 100. For example, the last day's price could be assigned a value of 10% (.10), which is added to the previous day's value of 90% (.90). That gives the last day 10% of the total weighting. That would be the equivalent of a 20 day average. By giving the last day's price a smaller value of 5% (.05), lesser weight is given to the last day's data and the average is less sensitive. That would be the equivalent of a 40 day moving average.

2.5 Candlesticks

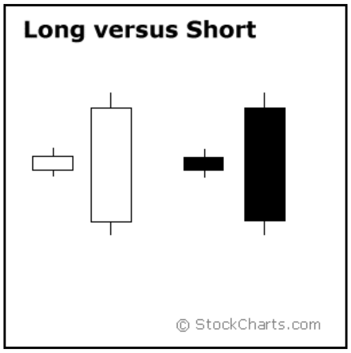
In order to create a candlestick chart, you must have a data set that contains open, high, low and close values for each time period you want to display. The hollow or filled portion of the candlestick is called “the body” (also referred to as “the real body”). The long thin lines above and below the body represent the high/low range and are called “shadows” (also referred to as “wicks” and “tails”). The high is marked by the top of the upper shadow and the low by the bottom of the lower shadow. If the stock closes higher than its opening price, a hollow candlestick is drawn with the bottom of the body representing the opening price and the top of the body representing the closing price. If the stock closes lower than its opening price, a filled candlestick is drawn with the top of the body representing the opening price and the bottom of the body representing the closing price.



Compared to traditional bar charts, many traders consider candlestick charts more visually appealing and easier to interpret. Each candlestick provides an easy-to-decipher picture of price action. Immediately a trader can compare the relationship between the open and close as well as the high and low. The relationship between the open and close is considered vital information and forms the essence of candlesticks. **Hollow** candlesticks, where the close is greater than the open, indicate buying pressure. **Filled** candlesticks, where the close is less than the open, indicate selling pressure.



2.5.1 Long versus Short Bodies

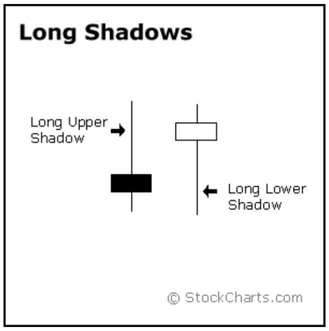


Long white candlesticks show strong buying pressure. The longer the white candlestick is, the further the close is above the open. This indicates that prices advanced significantly from open to close and buyers were aggressive. While long white candlesticks are generally bullish, much depends on their position within the broader technical picture. After extended declines, long white candlesticks can mark a potential turning point or support level. If buying gets too aggressive after a long advance, it can lead to excessive bullishness.

Long black candlesticks show strong selling pressure. The longer the black candlestick is, the further the close is below the open. This indicates that prices declined significantly from the open and sellers were aggressive. After a long advance, a long black candlestick can foreshadow a turning point or mark a future resistance level. After a long decline, a long black candlestick can indicate panic or capitulation.

2.5.2 Long versus Short Shadows

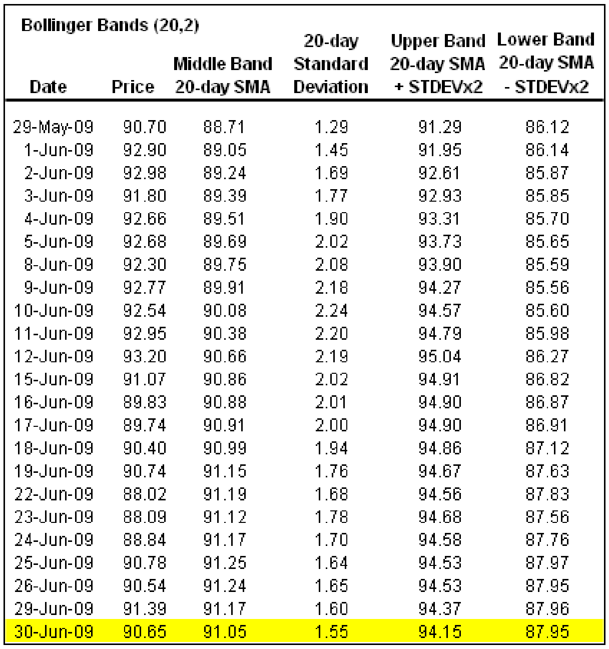
The upper and lower shadows on candlesticks can provide valuable information about the trading session. Upper shadows represent the session high and lower shadows the session low. Candlesticks with short shadows indicate that most of the trading action was confined near the open and close. Candlesticks with long shadows show that prices extended well past the open and close.



Candlesticks with a long upper shadow and short lower shadow indicate that buyers dominated during the session, and bid prices higher. However, sellers later forced prices down from their highs, and the weak close created a long upper shadow. Conversely, candlesticks with long lower shadows and short upper shadows indicate that sellers dominated during the session and drove prices lower. However, buyers later resurfaced to bid prices higher by the end of the session and the strong close created a long lower shadow.

2.6 Bollinger Bands

Developed by John Bollinger, Bollinger Bands are volatility bands placed above and below a moving average. Volatility is based on the standard deviation, which changes as volatility increases and decreases. The bands automatically widen when volatility increases and narrow when volatility decreases. This dynamic nature of Bollinger Bands also means they can be used on different securities with the standard settings. For signals, Bollinger Bands can be used to identify M-Tops and W-Bottoms or to determine the strength of the trend.

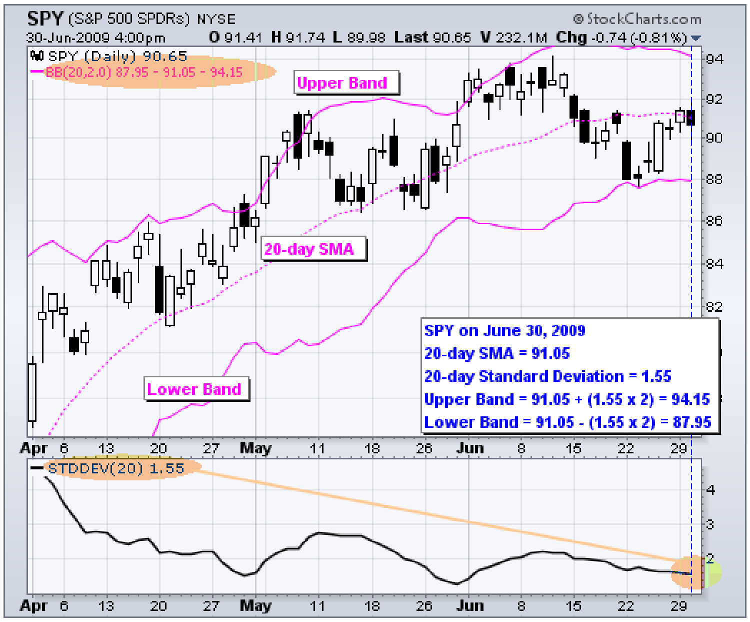


Bollinger Bands (20,2)

Date	Price	Middle Band 20-day SMA	20-day Standard Deviation	Upper Band 20-day SMA + STDEVx2	Lower Band 20-day SMA - STDEVx2
29-May-09	90.70	88.71	1.29	91.29	86.12
1-Jun-09	92.90	89.05	1.45	91.95	86.14
2-Jun-09	92.98	89.24	1.69	92.61	85.87
3-Jun-09	91.80	89.39	1.77	92.93	85.85
4-Jun-09	92.66	89.51	1.90	93.31	85.70
5-Jun-09	92.68	89.69	2.02	93.73	85.65
8-Jun-09	92.30	89.75	2.08	93.90	85.59
9-Jun-09	92.77	89.91	2.18	94.27	85.56
10-Jun-09	92.54	90.08	2.24	94.57	85.60
11-Jun-09	92.95	90.38	2.20	94.79	85.98
12-Jun-09	93.20	90.66	2.19	95.04	86.27
15-Jun-09	91.07	90.86	2.02	94.91	86.82
16-Jun-09	89.83	90.88	2.01	94.90	86.87
17-Jun-09	89.74	90.91	2.00	94.90	86.91
18-Jun-09	90.40	90.99	1.94	94.86	87.12
19-Jun-09	90.74	91.15	1.76	94.67	87.63
22-Jun-09	88.02	91.19	1.68	94.56	87.83
23-Jun-09	88.09	91.12	1.78	94.68	87.56
24-Jun-09	88.84	91.17	1.70	94.58	87.76
25-Jun-09	90.78	91.25	1.64	94.53	87.97
26-Jun-09	90.54	91.24	1.65	94.53	87.95
29-Jun-09	91.39	91.17	1.60	94.37	87.96
30-Jun-09	90.65	91.05	1.55	94.15	87.95

 * Middle Band = 20-day simple moving average (SMA)
 * Upper Band = 20-day SMA + (20-day standard deviation of price x 2)
 * Lower Band = 20-day SMA - (20-day standard deviation of price x 2)

Bollinger Bands consist of a middle band with two outer bands. The middle band is a simple moving average that is usually set at 20 periods. A simple moving average is used because the standard deviation formula also uses a simple moving average. The look-back period for the standard deviation is the same as for the simple moving average. The outer bands are usually set 2 standard deviations above and below the middle band.

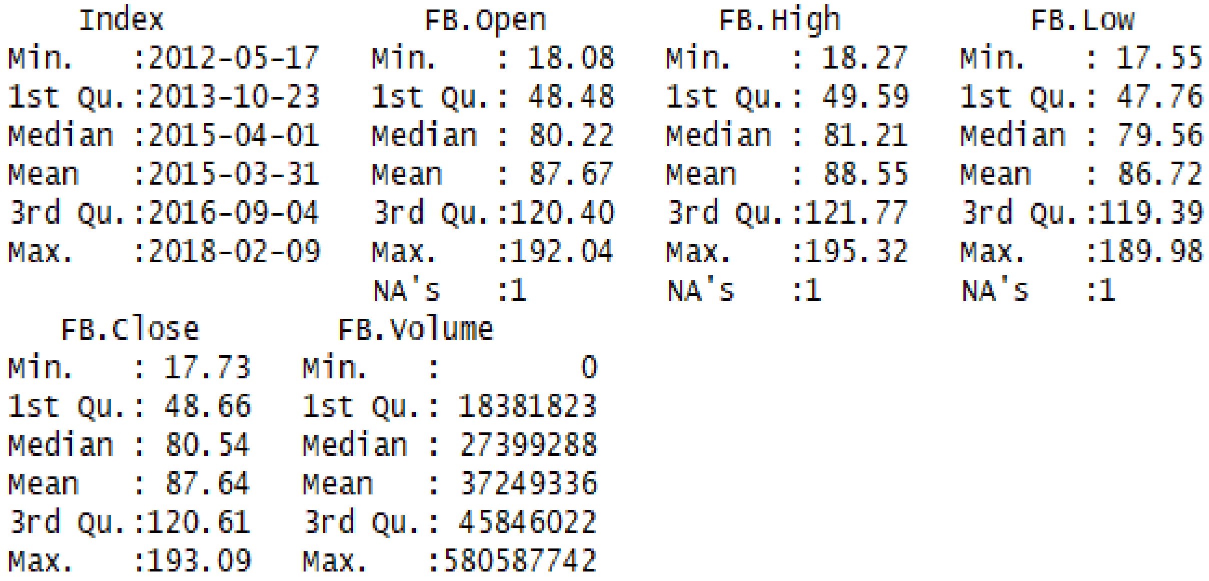


SPY on June 30, 2009
20-day SMA = 91.05
20-day Standard Deviation = 1.55
Upper Band = 91.05 + (1.55 x 2) = 94.15
Lower Band = 91.05 - (1.55 x 2) = 87.95

Bollinger recommends making small incremental adjustments to the standard deviation multiplier. Changing the number of periods for the moving average also affects the number of periods used to calculate the standard deviation. Therefore, only small adjustments are required for the standard deviation **multiplier**. An increase in the moving average period would automatically increase the number of periods used to calculate the standard deviation and would also warrant an increase in the standard deviation **multiplier**.

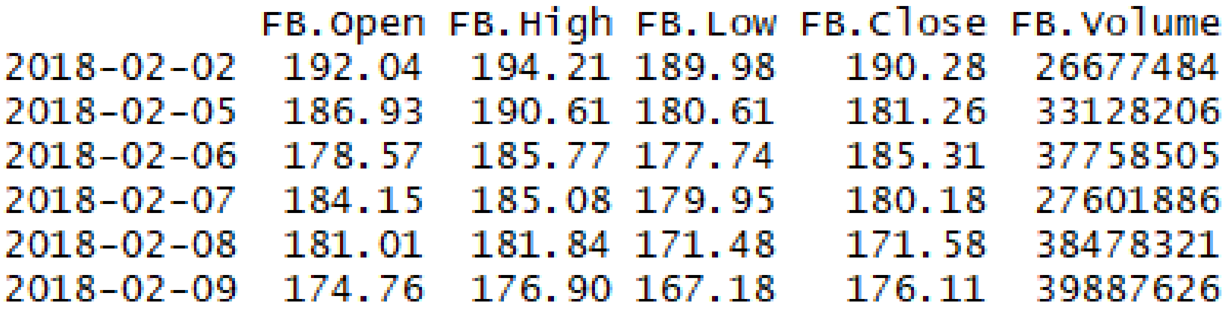
2.7 Data Exploration

2.7.1 Summary



```
     Index                 FB.Open           FB.High           FB.Low      
 Min.   :2012-05-17   Min.   : 18.08   Min.   : 18.27   Min.   : 17.55  
 1st Qu.:2013-10-23   1st Qu.: 48.48   1st Qu.: 49.59   1st Qu.: 47.76  
 Median :2015-04-01   Median : 80.22   Median : 81.21   Median : 79.56  
 Mean   :2015-03-31   Mean   : 87.67   Mean   : 88.55   Mean   : 86.72  
 3rd Qu.:2016-09-04   3rd Qu.:120.40   3rd Qu.:121.77   3rd Qu.:119.39  
 Max.   :2018-02-09   Max.   :192.04   Max.   :195.32   Max.   :189.98  
                      NA's   :1        NA's   :1        NA's   :1       
    FB.Close         FB.Volume        
 Min.   : 17.73   Min.   :        0  
 1st Qu.: 48.66   1st Qu.: 18381823  
 Median : 80.54   Median : 27399288  
 Mean   : 87.64   Mean   : 37249336  
 3rd Qu.:120.61   3rd Qu.: 45846022  
 Max.   :193.09   Max.   :580587742  
```

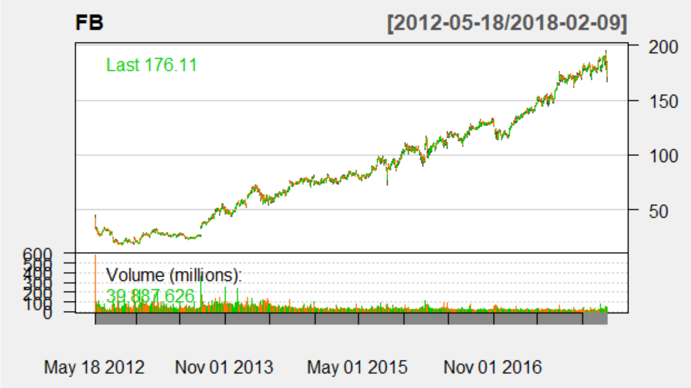
2.7.2 Values

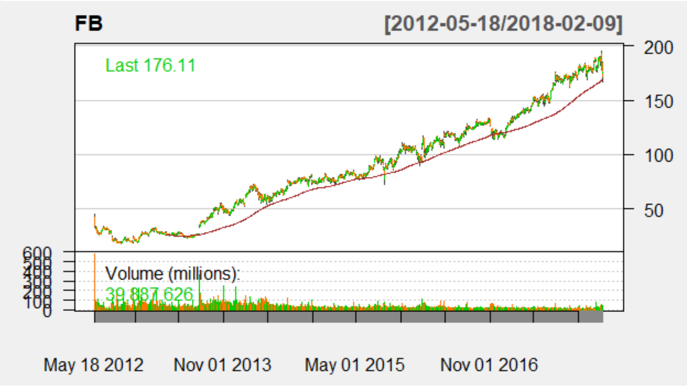


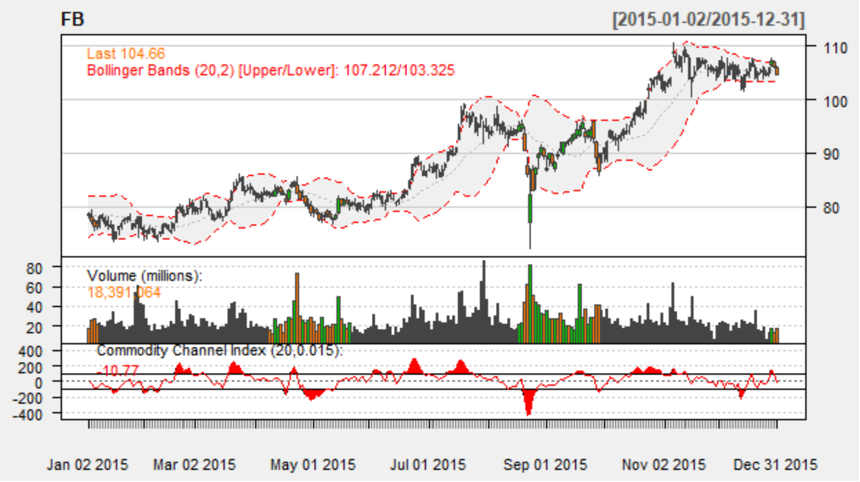
```
           FB.Open FB.High FB.Low FB.Close FB.Volume
2018-02-02  192.04  194.21 189.98   190.28  26677484
2018-02-05  186.93  190.61 180.61   181.26  33128206
2018-02-06  178.57  185.77 177.74   185.31  37758505
2018-02-07  184.15  185.08 179.95   180.18  27601886
2018-02-08  181.01  181.84 171.48   171.58  38478321
2018-02-09  174.76  176.90 167.18   176.11  39887626
```

2.7.3 Technical Indicators

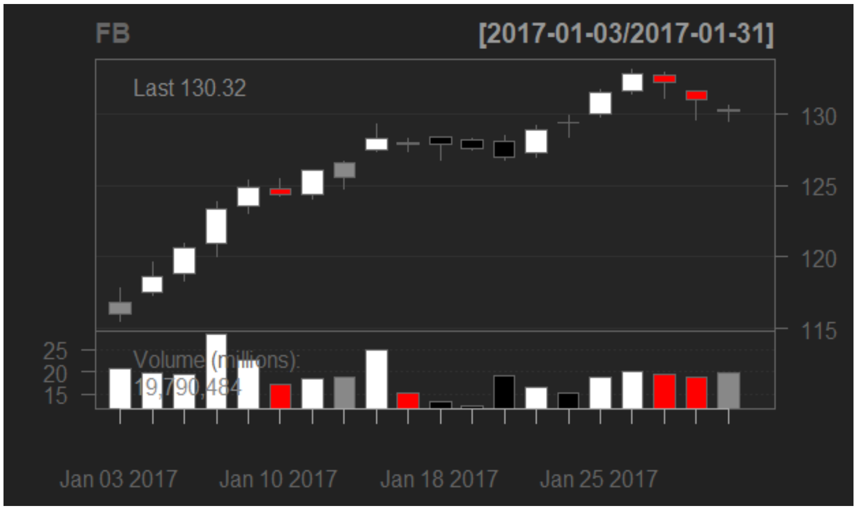
I experimented with a 200-period simple moving average (SMA) and a 10-period rate of change (ROC). The SMA will be overlaid directly on the chart as it shares the same scale while the ROC will live in a new pane underneath the main chart.



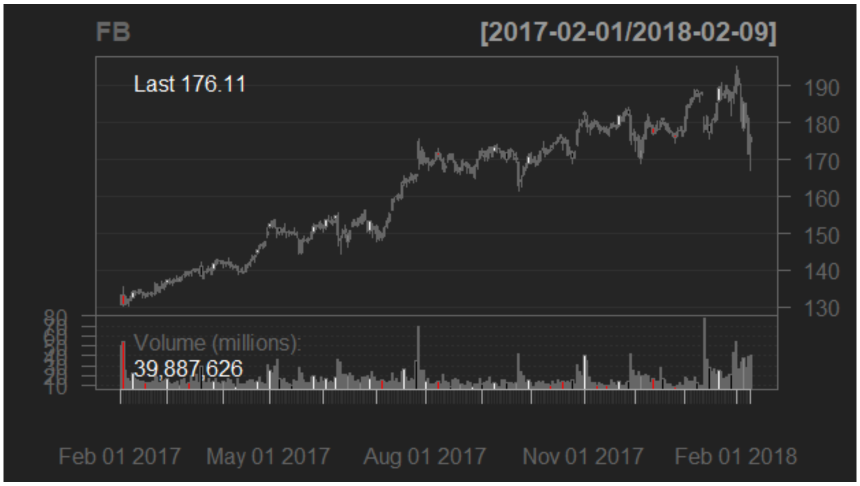




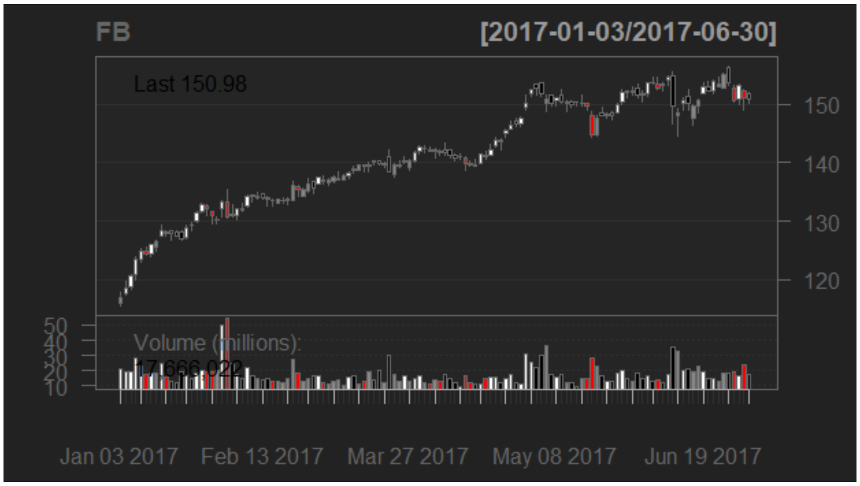
If we take a somewhat closer look at the data, we see a candle chart below focusing on the January 2017 alone.



Everything after February 2017

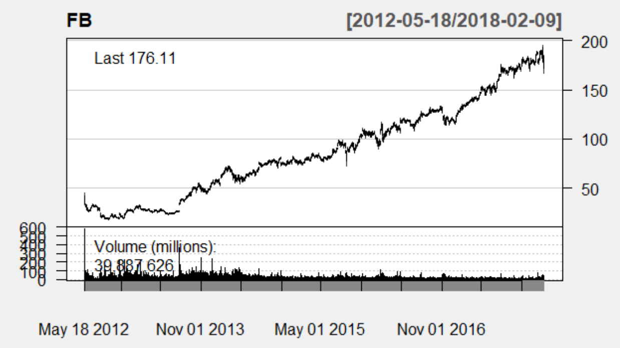


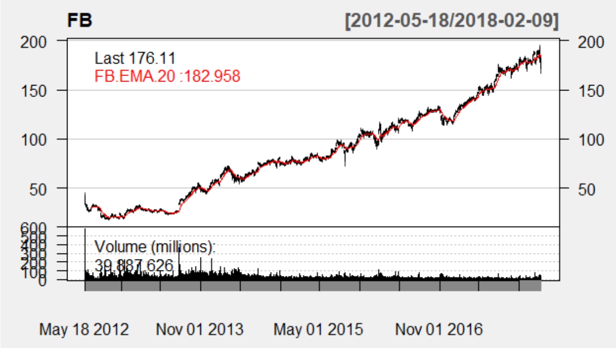
From January 2017 to June 2017

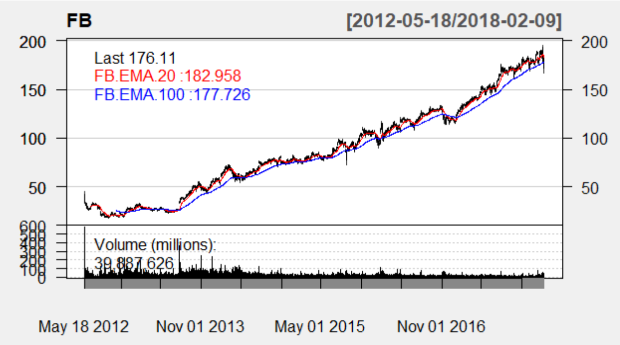


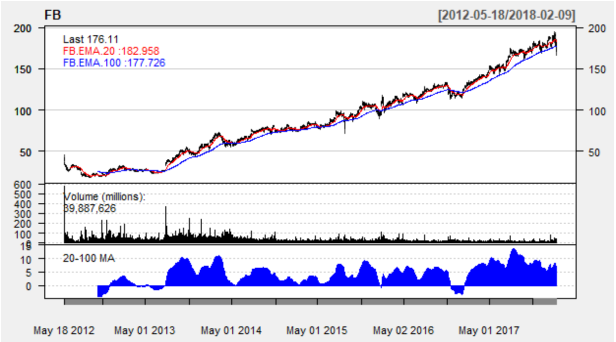
2.7.4 Custom Indicators

The indicator I'll build now will be the difference between two moving averages. We'll first create a chart, then a vector for each moving average that we'll overlay atop the main chart, and finally the difference between both moving averages in its own pane:









2.8 Autoregressive Integrated Moving Average (ARIMA)

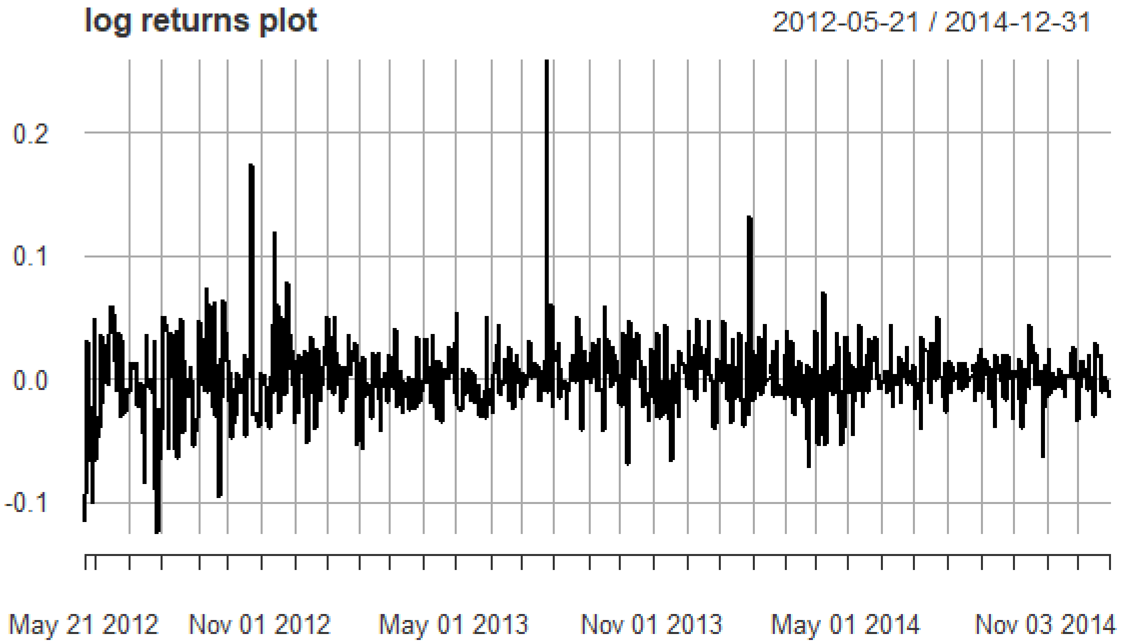
ARIMA stands for Autoregressive Integrated Moving Average. ARIMA is also known as Box-Jenkins approach.

The ARIMA model combines three basic methods:

- AutoRegression (AR) – In auto-regression the values of a given time series data are regressed on their own lagged values, which is indicated by the "p" value in the model.
- Differencing (I-for Integrated) – This involves differencing the time series data to remove the trend and convert a non-stationary time series to a stationary one. This is indicated by the "d" value in the model. If d = 1, it looks at the difference between two time series entries, if d = 2 it looks at the differences of the differences obtained at d =1, and so forth.
- Moving Average (MA) – The moving average nature of the model is represented by the "q" value which is the number of lagged values of the error term.

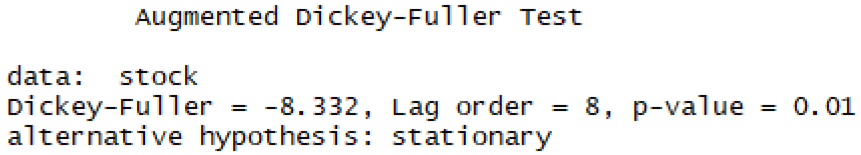
This model is called Autoregressive Integrated Moving Average or ARIMA(p,d,q)

To model a time series with the Box-Jenkins approach, the series has to be stationary. A stationary time series means a time series without trend, one having a constant mean and variance over time, which makes it easy for predicting values.



2.8.1 Testing for stationarity

We test for stationarity using the Augmented Dickey-Fuller unit root test. The p-value resulting from the ADF test has to be less than 0.05 or 5% for a time series to be stationary. If the p-value is greater than 0.05 or 5%, you conclude that the time series has a unit root which means that it is a non-stationary process.



```
        Augmented Dickey-Fuller Test

data:  stock
Dickey-Fuller = -8.332, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary
```

2.8.2 Differencing

To convert a non-stationary process to a stationary process, we apply the differencing method. Differencing a time series means finding the differences between consecutive values of a time series data. The differenced values form a new time series dataset which can be tested to uncover new correlations or other interesting statistical properties.

We can apply the differencing method consecutively more than once, giving rise to the “first differences”, “second order differences”, etc.

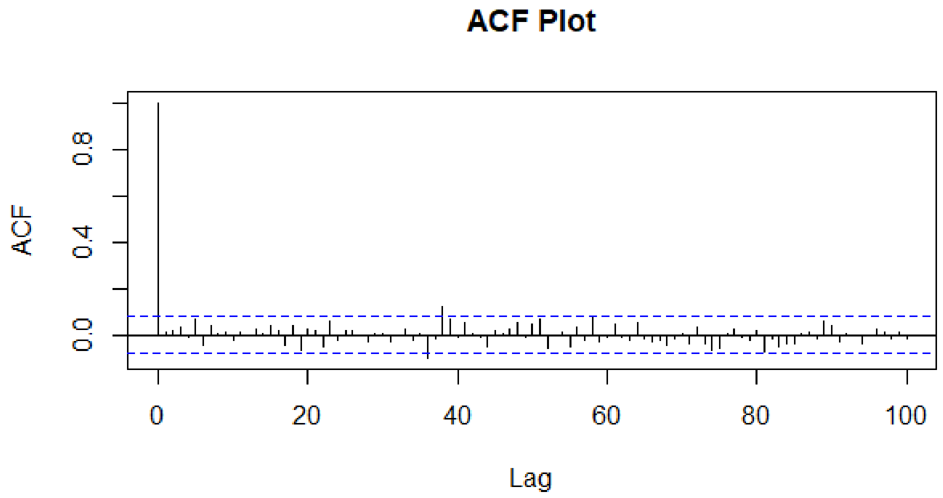
We apply the appropriate differencing order (d) to make a time series stationary before we can proceed to the next step.

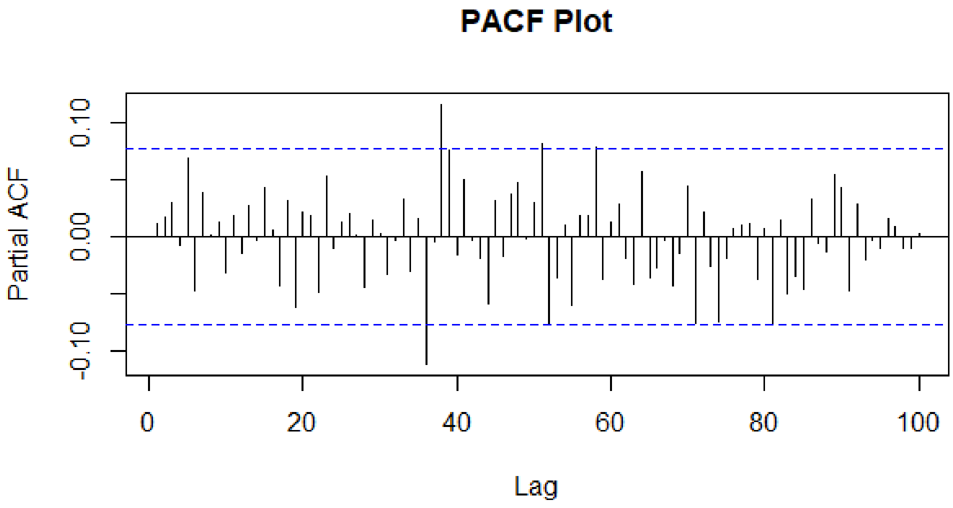
2.8.3 Identifying the p order of AR model

For AR models, the ACF will dampen exponentially and the PACF will be used to identify the order (p) of the AR model. If we have one significant spike at lag 1 on the PACF, then we have an AR model of the order 1, i.e. AR(1). If we have significant spikes at lag 1, 2, and 3 on the PACF, then we have an AR model of the order 3, i.e. AR(3).

2.8.4 Identifying the q order of MA model

For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order of the MA process. If we have one significant spike at lag 1 on the ACF, then we have an MA model of the order 1, i.e. MA(1). If we have significant spikes at lag 1, 2, and 3 on the ACF, then we have an MA model of the order 3, i.e. MA(3).





We can observe these plots and arrive at the Autoregressive (AR) order and Moving Average (MA) order.

We know that for AR models, the ACF will dampen exponentially and the PACF plot will be used to identify the order (p) of the AR model. For MA models, the PACF will dampen exponentially and the ACF plot will be used to identify the order (q) of the MA model. From these plots let us select AR order = 2 and MA order = 2. Thus, our ARIMA parameters will be (2,0,2).

2.8.5 Estimation and Forecasting

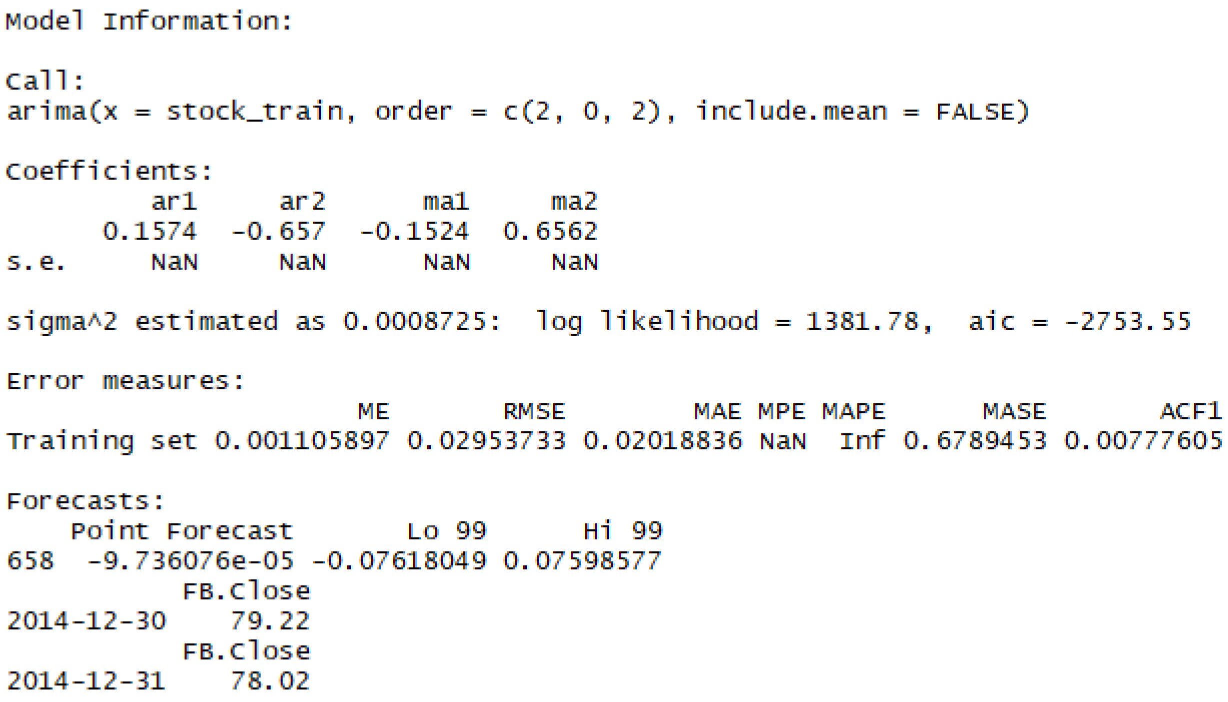
Once we have determined the parameters (p,d,q) we estimate the accuracy of the ARIMA model on a training data set and then use the fitted model to forecast the values of the test data set using a forecasting function. In the end, we cross check whether our forecasted values are in line with the actual values.

Our objective is to forecast the entire returns series from breakpoint onwards. We will make use of the For Loop statement in R and within this loop we will forecast returns for each data point from the test dataset.

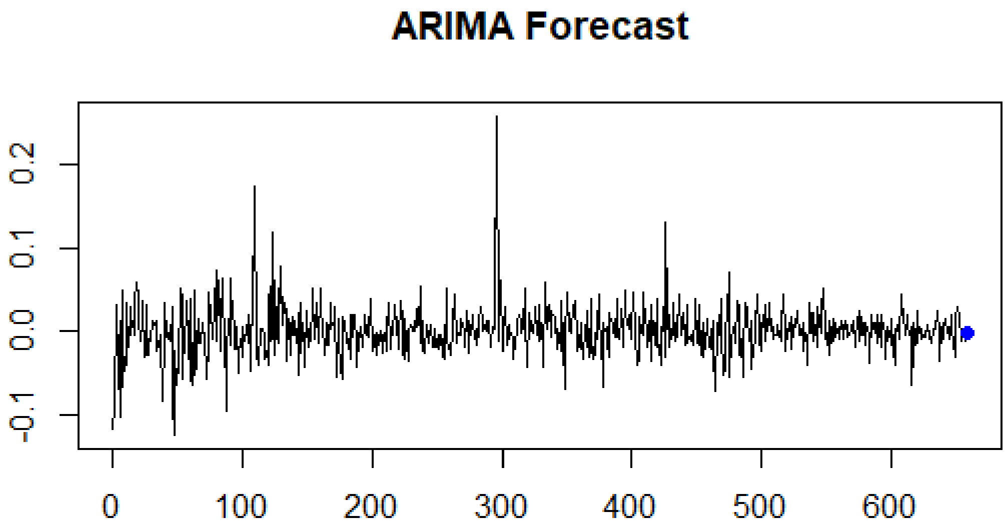
We first initialize a series which will store the actual returns and another series to store the forecasted returns. In the For Loop, we first form the training dataset and the test dataset based on the dynamic breakpoint.

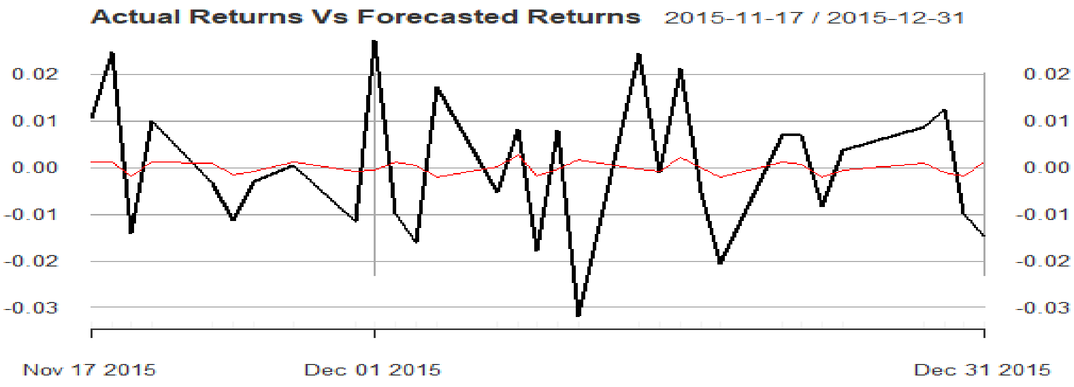
We call the arima function on the training dataset for which the order specified is (2, 0, 2). We use this fitted model to forecast the next data point by using the forecast. Arima function. The function is set at 99% confidence level. One can use the confidence level argument to enhance the model. We will be using the forecasted point estimate from the model. The “h” argument in the forecast function indicates the number of values that we want to forecast, in this case, the next day returns.

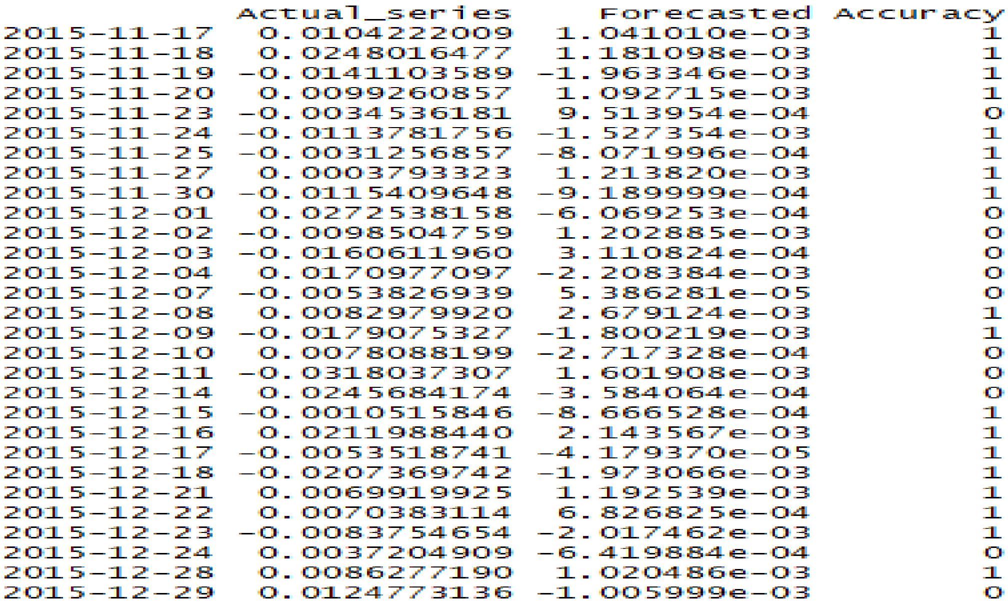
We can use the summary function to confirm the results of the ARIMA model are within acceptable limits.



```
Model Information:

Call:
arima(x = stock_train, order = c(2, 0, 2), include.mean = FALSE)

Coefficients:
         ar1     ar2      ma1     ma2
      0.1574  -0.657  -0.1524  0.6562
s.e.     NaN     NaN      NaN     NaN

sigma^2 estimated as 0.0008725:  log likelihood = 1381.78,  aic = -2753.55

Error measures:
                      ME       RMSE        MAE MPE MAPE      MASE       ACF1
Training set 0.001105897 0.02953733 0.02018836 NaN  Inf 0.6789453 0.00777605

Forecasts:
    Point Forecast       Lo 99      Hi 99
658  -9.736076e-05 -0.07618049 0.07598577
           FB.Close
2014-12-30    79.22
           FB.Close
2014-12-31    78.02
```







```
           Actual_series    Forecasted  Accuracy
2015-11-17  0.0104222009  1.041010e-03        1
2015-11-18  0.0248016477  1.181098e-03        1
2015-11-19 -0.0141103589 -1.963346e-03        1
2015-11-20  0.0099260857  1.092715e-03        1
2015-11-23 -0.0034536181  9.513954e-04        0
2015-11-24 -0.0113781756 -1.527354e-03        1
2015-11-25 -0.0031256857 -8.071996e-04        1
2015-11-27  0.0003793323  1.213820e-03        1
2015-11-30 -0.0115409648 -9.189999e-04        1
2015-12-01  0.0272538158 -6.069253e-04        0
2015-12-02 -0.0098504759  1.202885e-03        0
2015-12-03 -0.0160611960  3.110824e-04        0
2015-12-04  0.0170977097 -2.208384e-03        0
2015-12-07 -0.0053826939  5.386281e-05        0
2015-12-08  0.0082979920  2.679124e-03        1
2015-12-09 -0.0179075327 -1.800219e-03        1
2015-12-10  0.0078088199 -2.717328e-04        0
2015-12-11 -0.0318037307  1.601908e-03        0
2015-12-14  0.0245684174 -3.584064e-04        0
2015-12-15 -0.0010515846 -8.666528e-04        1
2015-12-16  0.0211988440  2.143567e-03        1
2015-12-17 -0.0053518741 -4.179370e-05        1
2015-12-18 -0.0207369742 -1.973066e-03        1
2015-12-21  0.0069919925  1.192539e-03        1
2015-12-22  0.0070383114  6.826825e-04        1
2015-12-23 -0.0083754654 -2.017462e-03        1
2015-12-24  0.0037204909 -6.419884e-04        0
2015-12-28  0.0086277190  1.020486e-03        1
2015-12-29  0.0124773136 -1.005999e-03        0
```

If the sign of the forecasted return equals the sign of the actual returns we have assigned it a positive accuracy score. The accuracy percentage of the model comes to around 61.29% which looks like a decent number.

3 Association Rule Mining

Association is the discovery of association relationships or correlations among a set of items. Market Basket analysis can also help retailers to plan which items to put on sale at reduced rates. For a set of items available at the supermarket, then each item has a Boolean variable representing the presence or absence of that item. Each basket can then be represented by a Boolean vector of the values assigned to these variables. The Boolean vector can be analyzed for buying patterns. These patterns can be represented in the form of association rules. Support and confidence are two measures of the rule interestingness. Association rules are considered interesting if both minimum support threshold and a minimum confidence threshold is satisfied. Association rule mining finds all the rules existing in the database that satisfy some minimum support and minimum confidence.

Association rules mining is an important subject in the study of data mining. Data mining is an interdisciplinary field, having applications in diverse areas like bioinformatics, medical informatics, scientific data analysis, financial analysis, consumer profiling, etc. In each of these application domains, the amount of data available for analysis has exploded in recent years. A time series data set consists of sequences of values or events that change with time. Time series data is popular in many applications, such as the daily closing prices of a hare in a stock market. Taking stock data as an example, it can be found the rules like “If stock A goes up and stock B goes up then stock C will goes up on the same day (5%, 75%)” with intra-transactional association rules, but it cannot be found the rules like “If stock A goes up on the first day and stock B goes up on the second day then stock C will goes up on the third day (5%, 75%)”. There is no time difference between the items in the intra-transactional association rules, so it cannot be used to predict the trend of time series.

Data mining is the process of finding valid, useful and understandable pattern in data. Due to the large size of databases, importance of information stored, and valuable information obtained, finding hidden patterns in data has become increasingly significant. The stock market provides an area in which large volumes of data is created and stored on a daily basis, and hence an ideal dataset for applying data mining techniques. The main objective of this study is to explore the suitability and a comparative study of the performance of item sets and association rules from a stock dataset.

3.1 Clean up and transform data

Not every stock symbol may have prices available for every day. Trading can be suspended for some reason, companies get acquired or go private, new companies form, etc.

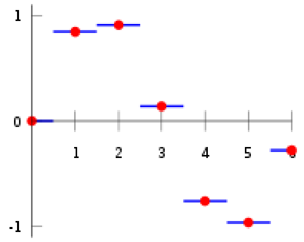
Going forward I have filled the missing values using the last reported price (piecewise constant interpolation) – a reasonable and widely used approach for stock price time series. After that, if there are still missing values, I have just removed the symbols that contain them, for the sake of simplicity.

3.2 Interpolation

Interpolation is a method of constructing new data points within the range of a discrete set of known data points. One often has a number of data points, obtained by sampling or experimentation, which represent the values of a function for a limited number of values of the independent variable. It is often required to interpolate (i.e., estimate) the value of that function for an intermediate value of the independent variable.

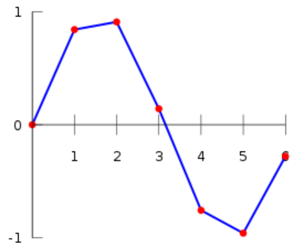
3.2.1 Piecewise constant interpolation

The simplest interpolation method is to locate the nearest data value, and assign the same value. In simple problems, this method is unlikely to be used, as linear interpolation is almost as easy, but in higher-dimensional multivariate interpolation, this could be a favourable choice for its speed and simplicity.



3.2.2 Linear interpolation

Linear interpolation on a set of data points (x0, y0), (x1, y1), ..., (xn, yn) is defined as the concatenation of linear interpolants between each pair of data points. This results in a continuous curve, with a discontinuous derivative.



Now that we have a universe of stocks with valid price data, convert those prices to log(returns) for the remaining analysis. The log(returns) are closer to normally distributed than prices especially in the long run. But why is getting the data closer to normal distribution important, it turns out it is somewhat necessary for the technique called partial correlation. That technique generally works better for normally distributed data than otherwise.

3.3 Finding Correlations

It's easy to convert the downloaded log(returns) data into a Pearson's sample correlation matrix X:

X = cor(log_returns)

The (i, j)th entry of the sample correlation matrix X above is a measurement of the degree of linear dependence between the log(return) series for the stocks in columns i and j.

There exist at least two issues that can lead to serious problems with the interpretation of the sample correlation values:

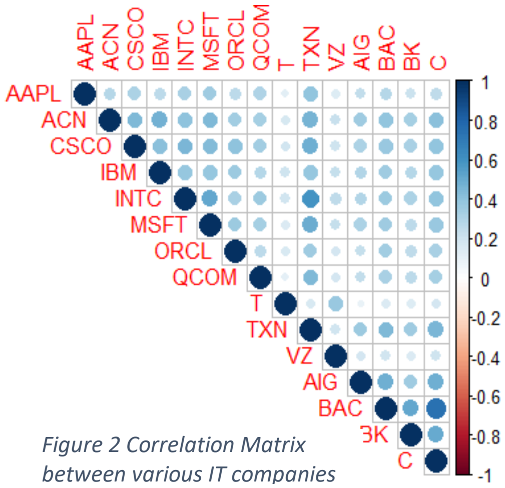


Figure 2 Correlation Matrix between various IT companies

- As Ledoit and Wolf point out, it's well-known that empirical correlation estimates may contain lots of error.
- Correlation estimates between two stock log(return) series can be misleading for many reasons, including spurious correlation or existence of confounding variables related to both series.

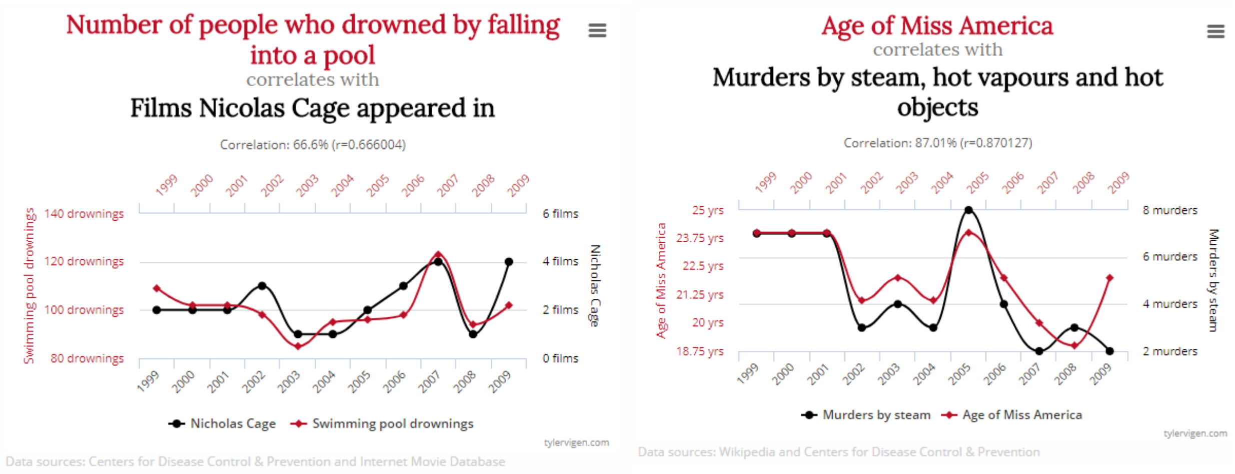


Figure 3. Spurious Correlations

3.3.1 Cointegration

If two or more series are individually integrated (in the time series sense) but some linear combination of them has a lower order of integration, then the series are said to be cointegrated. A common example is where the individual series are first-order integrated (I(1)) but some (cointegrating) vector of coefficients exists to form a stationary linear combination of them. For instance, a stock market index and the price of its associated futures contract move through time, each roughly following a random walk.

Cointegration is a wonderful but fairly technical topic. Instead, let's try a simpler approach.

We can try to address issue 2 above by controlling for confounding variables, at least partially. One approach considers partial correlation instead of correlation. That approach works best in practice with approximately normal data–one reason for the switch to log(returns) instead of prices. We will treat the entries of the precision matrix as measures of association in a network of stocks below.

3.3.2 Partial correlation

Partial correlation is the measure of association between two variables, while controlling or adjusting the effect of one or more additional variables. Partial correlations can be used in many cases that assess for relationship, like whether or not the sale value of a particular commodity is related to the expenditure on advertising when the effect of price is controlled.

The partial correlation coefficients between all stock log(returns) series are the entries of the inverse of the sample correlation matrix

As such inverting matrices is generally a bad idea and to make things even worse, issue 1 above says that our estimated correlation coefficients contains error (noise). Even tiny amount of noise can be hugely amplified if we invert the matrix. That's because the sample correlation matrix contains tiny eigenvalues and matrix inversion effectively divides the noise by those tiny values. Simply stated, dividing by a tiny number returns a big number–that is, matrix inversion tends to blow the noise up. This is a fundamental issue common to many inverse problems.

A sensible answer to reducing the influence of noise is regularization. Regularization replaces models with *different, but related*, models designed to reduce the influence of noise on their output. LW use a form of regularization related to ridge regression (a. k. a. Tikhonov regularization) with a peculiar regularization operator based on a highly structured estimate of the covariance.

3.4 Regularization

Regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting.

There are multiple ways to find the coefficients for a linear regression model. One of the widely used method is gradient descent. Gradient descent is an iterative method which takes some initial guess on coefficients and then tries to converge such that the objective function is minimized. Hence we work with partial derivatives on the coefficients.

$$\theta_j := \theta_j - \alpha \quad \frac{1}{m}\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})x_j^{(i)}$$

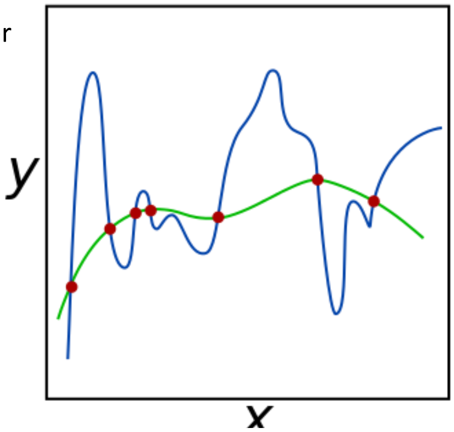
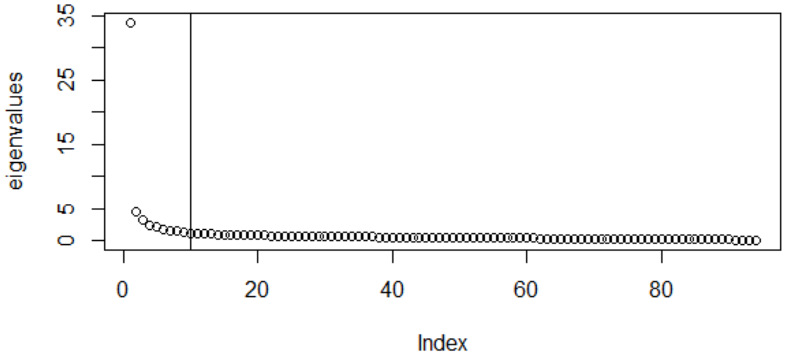


Figure 4 The green and blue functions both incur zero loss on the given data points. A learned model can be induced to prefer the green function, which may generalize better to more points drawn from the underlying unknown distribution, by adjusting, the weight of the regularization problem

R′s eigen() function takes care to return the (real-valued) eigenvalues of a symmetric matrix in decreasing order for us. (Technically, the correlation matrix is symmetric positive semi-definite, and will have only nonnegative real eigenvalues.)

Each eigenvector represents an orthogonal projection of the sample correlation matrix into a line (a 1-d shadow of the data); the first two eigenvectors define a projection of the sample correlation matrix into a plane (2-d), and so on. The eigenvalues estimate the proportion of information (or variability if you prefer) from the original sample correlation matrix contained in each eigenvector. Because the eigenvectors are orthogonal, these measurements of projected information are additive.



The eigenvalues fall off rather quickly in our example! That means that a lot of the information in the sample correlation matrix is contained in the first few eigenvectors.

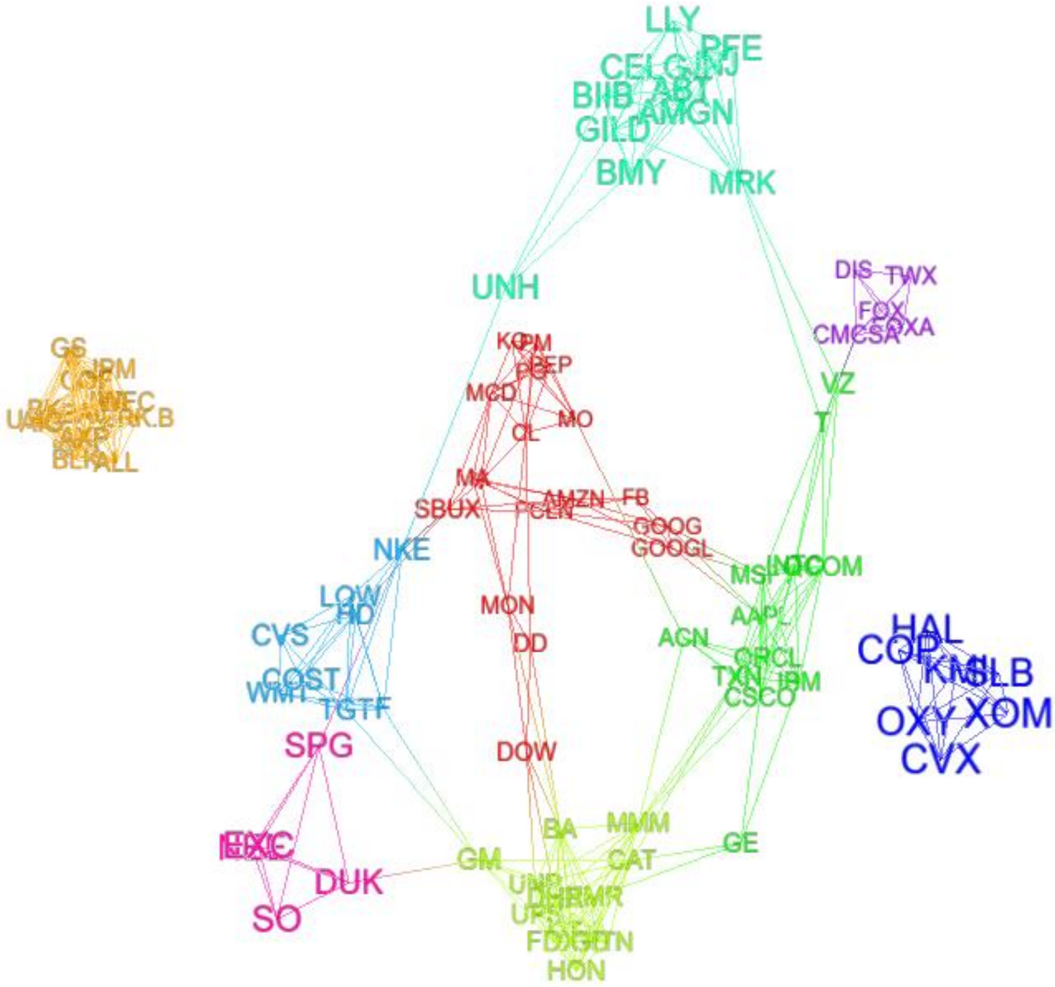
Regularization technique replaces the sample correlation matrix with an approximation defined by only its first few eigenvectors. Because they represent a large amount of the information content, the approximation can be pretty good. More importantly, because we assumed noise to be more or less equally represented across the eigenvector directions and weʹre cutting most of those off, this approximation tends to damp the noise more than the underlying information. Most importantly, weʹre cutting off the subspace associated with tiny eigenvalues, avoiding the problem of division by tiny values and significantly reducing amplified noise in the inverse of the sample correlation matrix.

The upshot is, we regularize the sample correlation matrix by approximating it by a low-rank matrix that substantially reduces the influence of noise on the precision matrix.

3.5 Networks and clustering

The (i, j)th entry of the precision matrix P is a measure of association between the log(return) time series for the stocks in columns i and j, with larger values corresponding to more association.

An interesting way to group related stocks together is to think of the precision matrix as an adjacency matrix defining a weighted, undirected network of stock associations. Thresholding entries of the precision matrix to include, say, only the top ten per cent results in a network of only the most strongly associated stocks.



The stock groups identified by this method are uncanny, but hardly all that surprising. Looking closely we will see that clusters are made up of bank-like companies (AIG, BAC, BK, C, COF, GS, JPM, MET, MS, USB, WFC), pharmaceutical companies (ABT, AMGN, BIIB, BMY, CELG, GILD, JNJ, LLY, MRK, PFE), computer/technology-driven companies (AAPL, ACN, CSCO, IBM, INTC, MSFT, ORCL, QCOM, T, TXN, VZ, and so on.

The groups more or less correspond to what we already know!

The FB, GOOG, AMZN, PCLN (Facebook, Alphabet/Google, Amazon, Priceline) group is interesting–it includes credit card companies V (Visa), MA (MasterCard). Perhaps the returns of FB, GOOG and AMZN are more closely connected to consumer spending than technology!

4 Bibliography

https://www.investopedia.com/university/stocks/stocks1.asp

https://en.wikipedia.org/wiki/Stock_market

https://economictimes.indiatimes.com/definition/stock-market

https://www.investopedia.com/ask/answers/012615/whats-difference-between-primary-and-secondary-capital-markets.asp

http://stockcharts.com/school/doku.php?id=chart_school:chart_analysis:introduction_to_candlesticks

http://stockcharts.com/school/doku.php?id=chart_school:technical_indicators:bollinger_bands

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc41.htm

https://fxborssa.com/wp-content/uploads/2017/09/Technical-Analysis-of-the-Futures-Markets-John-J.-Murphy-fxborssa.pdf

https://en.wikipedia.org/wiki/Commodity_channel_index

https://finance.google.com/finance?hl=en&tab=ee

http://www.itl.nist.gov/div898/handbook/pmc/section4/pmc4.htm

https://www.analyticsvidhya.com/blog/2015/12/complete-tutorial-time-series-modeling/

http://www.tylervigen.com/spurious-correlations

https://www.analyticsvidhya.com/blog/2017/08/mining-frequent-items-using-apriori-algorithm/

https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

https://pdfs.semanticscholar.org/058f/0f59a0bbf94a229a51f7e7f2588a89e997c4.pdf

http://www.lingayasuniversity.edu.in/ijdatpm/Pdf_paper/Paper_7.pdf