Department of Electrical and Electronic Engineering
Imperial College London

EE4-69 Signal Processing and Machine Learning for Finance
Coursework

Steven Wong
syw215@ic.ac.uk
01067352

April 18, 2019

Table of Contents

1 Regression Methods **3**
1.1 Processing Stock Price Data in Python . . . 3
1.1.1 S&P 500 index daily price . . . 3
1.1.2 The mean and variance of the price and log-price time series . . . 3
1.1.3 The simple and log return time series from the price data . . . 4
1.1.4 Suitability of log returns over simple returns for signal processing purposes . . . 5
1.1.5 Simple and logarithmic returns example . . . 6
1.1.6 Circumstances when simple return is better than logarithmic return . . . 6
1.2 ARMA vs ARIMA Models for Financial Applications . . . 6
1.2.1 S&P 500 time-series and the suitability of ARMA and ARIMA . . . 6
1.2.2 Autoregressive moving average model . . . 7
1.2.3 Autoregressive integrated moving average model . . . 7
1.2.4 The necessity of taking the log of the prices for the ARIMA analysis . . . 8
1.3 Vector Autoregressive (VAR) Models . . . 9
1.3.1 Concise matrix representation of VAR models . . . 9
1.3.2 Optimal set of coefficients . . . 10
1.3.3 Stability of VAR model . . . 10
1.3.4 VAR model of the stocks with tickers CAG, MAR, LIN, HCP, MAT . . . 10
1.3.5 VAR model on stocks within the same sector . . . 12

2 Bond Pricing **14**
2.1 Examples of bond pricing . . . 14
2.2 Forward rates . . . 14
2.3 Duration of a coupon-bearing bond . . . 15
2.4 Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) . . . 15
2.4.1 Market returns . . . 15
2.4.2 Sliding beta . . . 16
2.4.3 Cap-weighted market return . . . 16
2.4.4 Cap-weighted beta . . . 17
2.4.5 Arbitrage pricing theory . . . 17

3 Portfolio Optimisation **21**
3.1 Adaptive minimum-variance portfolio optimisation . . . 21
3.1.1 Derive optimal weights . . . 21
3.1.2 Compare optimal weights and equally-weighted portfolio . . . 21
3.1.3 Adaptive weights portfolio . . . 22

4 Robust Statistics and Non Linear Methods **25**
4.1 Data Import and Exploratory Data Analysis . . . 25
4.1.1 Stock key statistics . . . 25
4.1.2 Histogram and probability density function of the stocks . . . 25
4.1.3 Anomaly detection . . . 27
4.1.4 Testing anomaly detection performance . . . 28
4.1.5 Box plot . . . 30
4.2 Robust Estimators . . . 31
4.3 Robust and OLS regression . . . 31
4.4 Robust Trading Strategies . . . 34

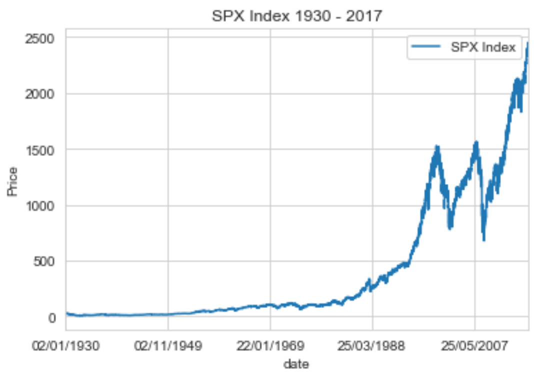
5 Graphs in Finance **42**
5.1 Graph Theory . . . 42
5.2 Impact of Crude Oil Reports on Crude Oil Price . . . 43
5.3 Impact of Crude Oil Price on ADX and S&P500 . . . 43
5.4 Impact of Crude Oil Price and ADX & S& P500 Indices on Forex Rates . . . 44
5.5 Potential of Dutch Disease . . . 45
5.6 Advance Commodity/Forex Graph . . . 46
5.7 Summary . . . 46

1 Regression Methods

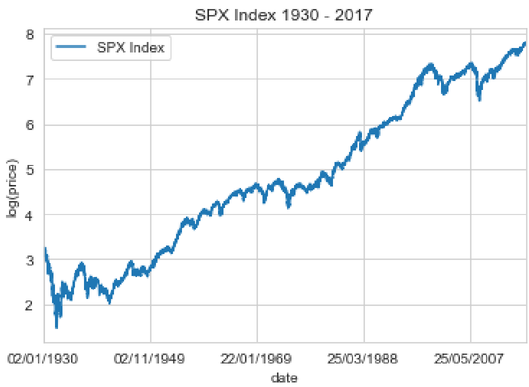
1.1 Processing Stock Price Data in Python

Python will be used throughout this coursework for signal processing and data visualisation. The dataset used in this section is the S&P 500 (SPX) Index daily price from 1930 to 2017. The daily price, mean and variance of the price, log and simple return of the dataset will be shown and discussed.

1.1.1 S&P 500 index daily price



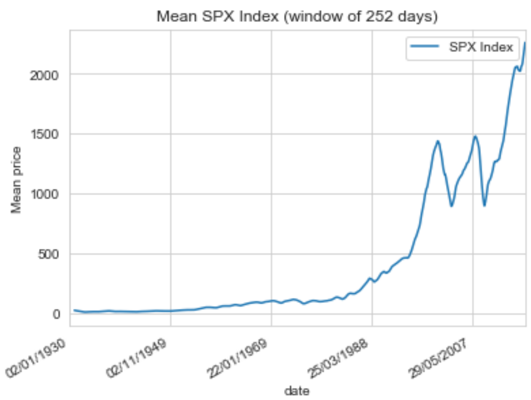
(a) Price time-series



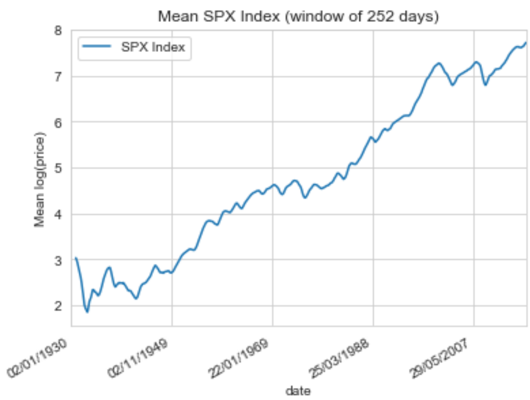
(b) Log-price time-series

Figure 1: S&P 500 Index daily price from 1930 to 2017

1.1.2 The mean and variance of the price and log-price time series

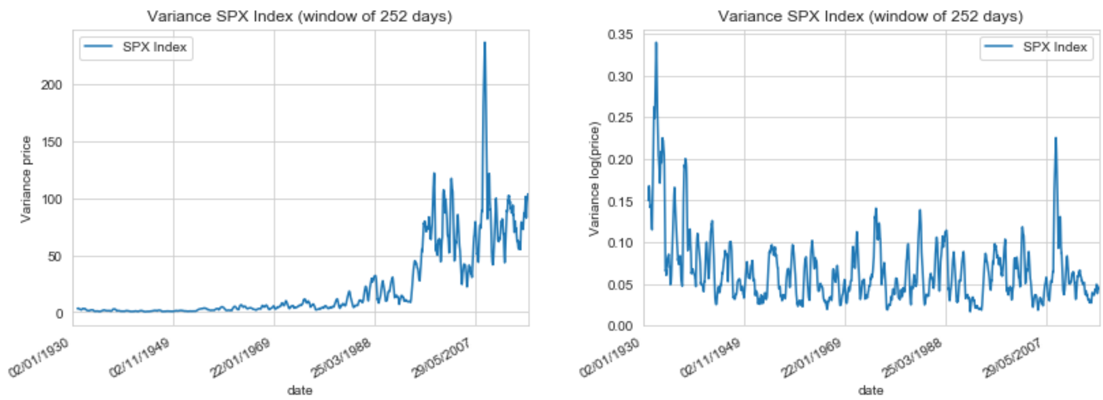


(a) Mean price



(b) Mean log-price

Figure 2: S&P 500 Index yearly rolling mean price from 1930 to 2017



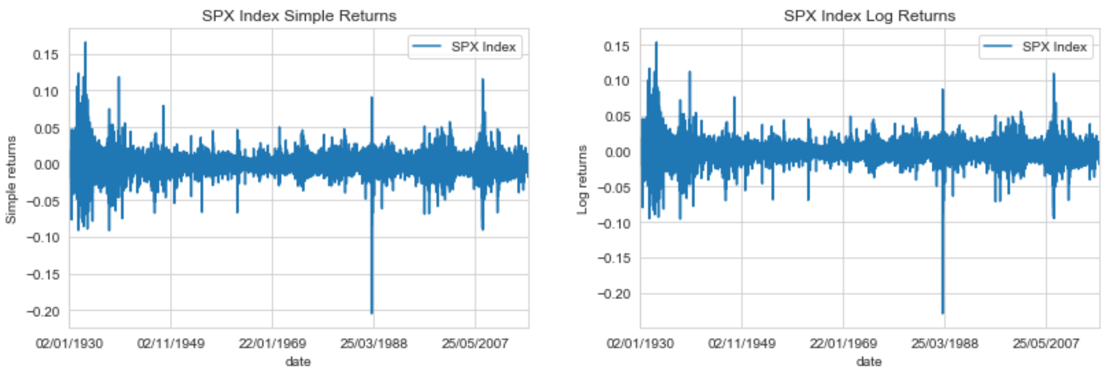
(a) Price variance

(b) Log-price variance

Figure 3: S&P 500 Index yearly rolling price variance from 1930 to 2017

From figure 2(a) and 3(a), the rolling mean price and price variance suggest that the S&P 500 Index could be stationary in the first 30 years. But this is wrong because the graph has stretched up to $250, the price changes in that period could be relatively large regarding that time, but too small compared to the maximum price it has reached. From figure 2(b) and 3(b), we can observe that the rolling mean log-price and log-price variance is not constant. Hence, the S&P 500 time-series is non-stationary.

1.1.3 The simple and log return time series from the price data



(a) Simple returns

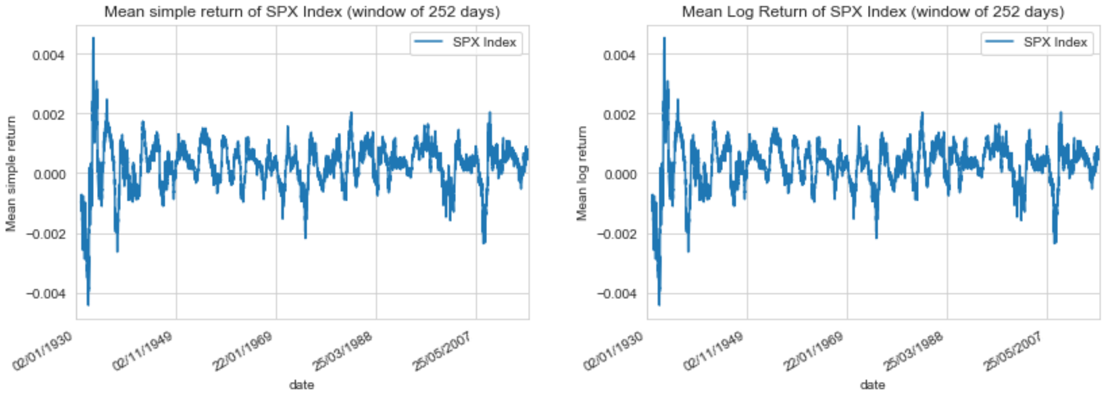
(b) Log returns

Figure 4: S&P 500 Index simple and log return from 1930 to 2017

By converting the prices to returns, the time series is said to be stationarised. The simple return and log return can be calculated with the following equations:

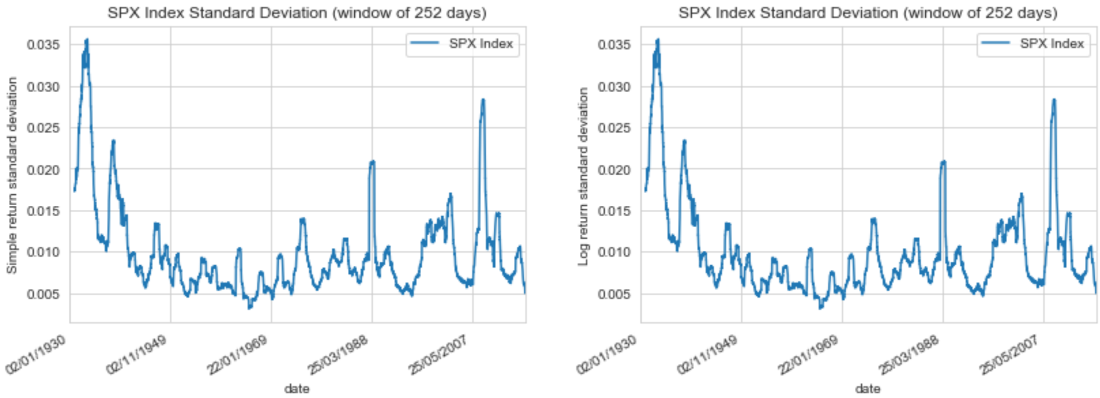
$$\text{Simple return} = R[t] = \frac{p[t]}{p[t-1]} - 1 \tag{1}$$

$$\text{Log return} = r[t] = \log(p[t]) - \log(p[t-1]) \tag{2}$$



(a) Simple returns (b) Log returns

Figure 5: S&P 500 Index rolling simple and log return mean from 1930 to 2017



(a) Simple returns (b) Log returns

Figure 6: S&P 500 Index simple and log return rolling variance from 1930 to 2017

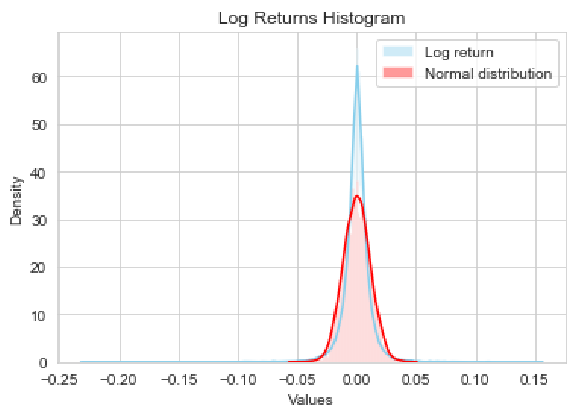
The simple and log return is more stationary compare to the price and log-price because their mean and variance are more stable.

1.1.4 Suitability of log returns over simple returns for signal processing purposes

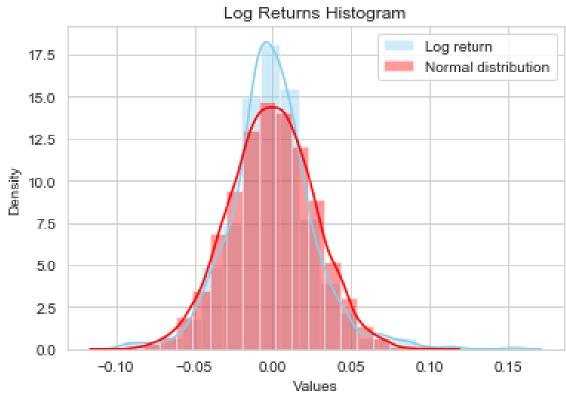
The logarithmic function is a monotonic function which preserves the relative ordering ($a > b$ then $\log a > \log b$), and log compresses the range of data. For small returns, the simple return is approximately equal to the log return which can be observed in figure 4.

The Jarque-Bera test has been performed on both simple and log returns, the output of their p-value both equal to zero which suggests the simple and log returns time-series does not have a normal distribution confirming the histogram shown in figure 7a. The normal distribution plotted below has the same mean, variance, and the number of samples as the log return as a normal distribution reference.

Another reason that causes both log and simple returns to fail the normality test is about the number of outliers in the data with the long-term trends are causing a shift in the skewness of the return. A shorter duration of returns of the histogram shows a shape closer to the normal distribution in figure 7b, as the data has fewer outliers and the samples are all on roughly the same trend.



(a) 1930 to 2017



(b) 1000 samples

Figure 7: Log return and a normal distribution with the same mean and variance

1.1.5 Simple and logarithmic returns example

The log return has the time additivity property over simple return which can be shown with the following example; if I were to purchase a stock for £1, then the next day its value goes up to £2 and the following day back to £1. The above table shows the differences between simple and log returns. From the above example, it can be concluded that the logarithmic return is a monotonic function which preserves the relative ordering which simple return does not.

Table 1: Electrode connections based on analogue switch control logic

Day	One	Two	Three	**Total returns**
Stock (£)	1	2	1	
Simple returns		1	-0.5	+0.5
Log returns		0.693	-0.693	0

1.1.6 Circumstances when simple return is better than logarithmic return

With the above advantages with log returns, there are circumstances when simple returns should be used over log returns. It is preferred not to use log returns over long periods because the assumption of log-normality over long time-scales is unrealistic and a positive skew is assumed in log-normal distributions, but most financial data is negatively skewed in long time scales. In addition, log returns are not linearly additive across assets whereas simple returns are linearly additive.

1.2 ARMA vs ARIMA Models for Financial Applications

ARMA and ARIMA stand for autoregressive-moving-average process and autoregressive-integrated-moving-average respectively. The autoregressive part regresses the variable $x[t]$ on its own delayed values uses to describe the momentum and mean reversion effects observed in the trading markets.The moving-average part represents the error term as a linear combination of error terms from the past, $y[t-1], ..., y[t-q]$, can be used to capture the shock effects (unexpected events) observed as white noise.

The ARMA(p,q) can be represented by the following equation:

$$x[t] = \sum_{i=1}^{p} a_i x[t-i] + \sum_{i=1}^{q} b_i \eta[t-i] + \eta[t] \tag{3}$$

where η is the white noise error terms, a_i and b_i are the parameters of the model.

The following graph shows the S&P 500 index daily log closing prices from 2015 to 2019.

1.2.1 S&P 500 time-series and the suitability of ARMA and ARIMA

The following graph shows the S&P 500 index daily log closing prices from 2015 to 2019.

Figure 8: S&P 500 Log-closing price from 2015 to 2018

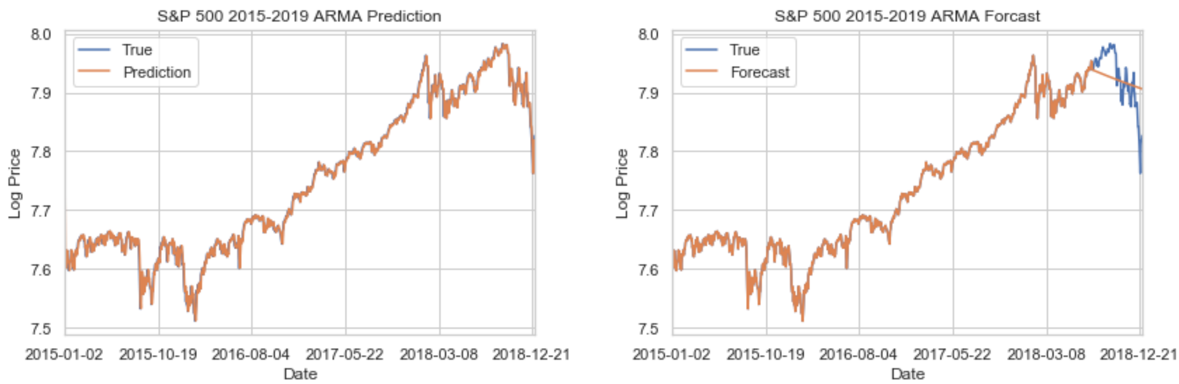
From figure 8, it is clear that the log price is non-stationary. Since the ARMA model assumes the data is stationary, an ARIMA model would be more appropriate. ARIMA model uses an initial differencing to remove elements of non-stationary in the data.

1.2.2 Autoregressive moving average model

The ARMA model has the model parameters of 7.74 and 0.9974 as the constant and autoregressive variable respectively. The first value is the constant represents the constant value in the model; the second value is the slope coefficient suggest the S&P price is dropping.

The method the data is fitted using regression of the data, then the prediction is formed by using the initial data hence the prediction, and the true signal is overlapped in figure 9a (predicted data are meaningless), a mean-squared error of -40.6 dB. Figure 9b shows the forecasted data using roughly 90% of the data. As expected, the ARMA model expects the price to continue the previous day's return in the ARMA model, which steadily decreases due to the slope is smaller than 1.

The AR(1) model only considers the previous day's return; it cannot predict a value that is outside its current linear trend which is not very useful in the real world.



(a) ARMA model

(b) ARMA forecasted signal

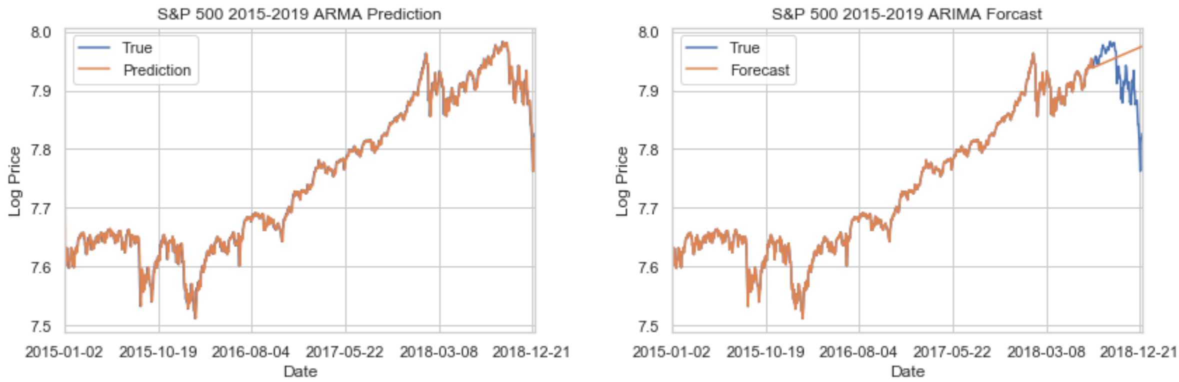
Figure 9: S&P 500 True and predicted with forecasted signal using ARMA from 2015 to 2018

1.2.3 Autoregressive integrated moving average model

The ARIMA model has the model parameters of 0.0002 and -0.0088 as the constant and difference variable respectively. The constant parameter is small, suggest the ARIMA model can handle the non-stationarity of the data. The differences parameter suggest the latest trend is drifting downwards.

The mean of ARMA and ARIMA model prediction error is 0.005974 and 0.005848 respectively. Therefore, the ARIMA model performs slightly better than the ARMA model. However, the predicted stock prices are across the past data; out-of-sample data is required to test the models predictions accurately. In addition, both models are only somewhat accurate on predicting one step ahead, with every new data, a correction was made on the model parameters to ensure the next prediction is as accurate as possible.

The following figure shows when the model is fitted with the first 90% of data and forecast the futures prices with its latest ARIMA model parameters. It can be seen the forecast is blindly following the latest information in the model, supporting the ARIMA model is for short term predictions only.



(a) ARIMA model

(b) ARIMA forecasted signal

Figure 10: S&P 500 True and predicted with forecasted signal using ARIMA from 2015 to 2018

Although both ARMA and ARIMA model forecast is incorrect, ARIMA model analysis would provide more information about the latest trend due to it uses the differences between past data where ARMA model uses a constant parameter and attempt to predict with a parameter based on all past data.

1.2.4 The necessity of taking the log of the prices for the ARIMA analysis

As mentioned in section 1.1.4, taking the log of the prices can preserve the relative ordering and compresses the range of data, and ARIMA analysis requires using log to compress the range of data hence the differencing process can remove the non-stationarity effectively and provide accurate model parameters.

In addition, the differencing term in the ARIMA model essentially becomes log return; the predictions become more linear and stationary as the overall variance of the difference model parameter become smaller which more stable.

1.3 Vector Autoregressive (VAR) Models

Vector autoregressive is a multivariate extension of the AR processes given by:

$$\mathbf{y}_t = \mathbf{c} + \mathbf{A}_1\mathbf{y}_{t-1} + \mathbf{A}_2\mathbf{y}_{t-2} + \cdots + \mathbf{A}_p\mathbf{y}_{t-p} + \mathbf{e}_t \tag{4}$$

The expanded matrix notation:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ \vdots \\ y_{k,t} \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_k \end{bmatrix} + \begin{bmatrix} a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} \\ a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ \vdots \\ y_{k,t-1} \end{bmatrix} + \cdots + \begin{bmatrix} a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ a^p_{2,1} & a^p_{2,2} & \cdots & a^p_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ \vdots \\ y_{k,t-p} \end{bmatrix} + \begin{bmatrix} e_{1,t} \\ e_{2,t} \\ \vdots \\ e_{k,t} \end{bmatrix} \tag{5}$$

1.3.1 Concise matrix representation of VAR models

The above matrix expression can be represented in the vector form of the following:

$$\begin{aligned} y_{1,t} &= c_1 + a^1_{1,1}y_{1,t-1} + a^1_{1,2}y_{2,t-1} + \cdots + a^1_{1,k}y_{k,t-1} + \cdots + a^p_{1,1}y_{1,t-p} + a^p_{1,2}y_{2,t-p} + \cdots + a^p_{1,k}y_{k,t-p} + e_{1,t} \\ y_{2,t} &= c_2 + a^1_{2,1}y_{1,t-1} + a^1_{2,2}y_{2,t-1} + \cdots + a^1_{2,k}y_{k,t-1} + \cdots + a^p_{2,1}y_{1,t-p} + a^p_{2,2}y_{2,t-p} + \cdots + a^p_{2,k}y_{k,t-p} + e_{2,t} \\ &\vdots \\ y_{k,t} &= c_k + a^1_{k,1}y_{1,t-1} + a^1_{k,2}y_{2,t-1} + \cdots + a^1_{k,k}y_{k,t-1} + \cdots + a^p_{k,1}y_{1,t-p} + a^p_{k,2}y_{2,t-p} + \cdots + a^p_{k,k}y_{k,t-p} + e_{k,t} \end{aligned}$$

The VAR(p) with k variables can be represented in a concise matrix form as:

$$\mathbf{Y} = \mathbf{BZ} + \mathbf{U} \tag{6}$$

Where $\mathbf{Y} \in \mathbb{R}^{K\times T}$, $\mathbf{B} \in \mathbb{R}^{K\times(KP+1)}$, $\mathbf{Z} \in \mathbb{R}^{(KP+1)\times T}$, and $\mathbf{U} \in \mathbb{R}^{K\times T}$ with the following representations:

$$\mathbf{Y} = \begin{bmatrix} y_p & y_{p+1} & \cdots & y_T \end{bmatrix} = \begin{bmatrix} y_{1,p} & y_{1,p+1} & \cdots & y_{1,T} \\ y_{2,p} & y_{2,p+1} & \cdots & y_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p} & y_{k,p+1} & \cdots & y_{k,T} \end{bmatrix} \tag{7}$$

$$\mathbf{B} = \begin{bmatrix} c & A_1 & A_2 & \cdots & A_p \end{bmatrix} = \begin{bmatrix} c_1 & a^1_{1,1} & a^1_{1,2} & \cdots & a^1_{1,k} & \cdots & a^p_{1,1} & a^p_{1,2} & \cdots & a^p_{1,k} \\ c_2 & a^1_{2,1} & a^1_{2,2} & \cdots & a^1_{2,k} & \cdots & a^p_{2,1} & a^p_{2,2} & \cdots & a^p_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \cdots & \vdots & \vdots & \ddots & \vdots \\ c_k & a^1_{k,1} & a^1_{k,2} & \cdots & a^1_{k,k} & \cdots & a^p_{k,1} & a^p_{k,2} & \cdots & a^p_{k,k} \end{bmatrix} \tag{8}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{p-1} & y_p & \cdots & y_{T-1} \\ y_{p-2} & y_{p-1} & \cdots & y_{T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_0 & y_1 & \cdots & y_{T-p} \end{bmatrix} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ y_{1,p-1} & y_{1,p} & \cdots & y_{1,T-1} \\ y_{2,p-1} & y_{2,p} & \cdots & y_{2,T-1} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-1} & y_{k,p} & \cdots & y_{k,T-1} \\ y_{1,p-2} & y_{1,p-1} & \cdots & y_{1,T-2} \\ y_{2,p-2} & y_{2,p-1} & \cdots & y_{2,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,p-2} & y_{k,p-1} & \cdots & y_{k,T-2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{1,0} & y_{1,1} & \cdots & y_{1,T-p} \\ y_{2,0} & y_{2,1} & \cdots & y_{2,T-p} \\ \vdots & \vdots & \ddots & \vdots \\ y_{k,0} & y_{k,1} & \cdots & y_{k,T-p} \end{bmatrix} \tag{9}$$

$$\mathbf{U} = \begin{bmatrix} e_p & e_{p+1} & \cdots & e_T \end{bmatrix} = \begin{bmatrix} e_{1,p} & e_{1,p+1} & \cdots & e_{1,T} \\ e_{2,p} & e_{2,p+1} & \cdots & e_{2,T} \\ \vdots & \vdots & \ddots & \vdots \\ e_{k,p} & e_{k,p+1} & \cdots & e_{k,T} \end{bmatrix} \tag{10}$$

1.3.2 Optimal set of coefficients

The optimal set of coefficients $\mathbf{B}$, denoted by $\mathbf{B}_{opt}$ can be obtained by assuming $\mathbf{U} = \mathbf{0}$, hence:

$$\mathbf{Y} = \mathbf{B}_{opt}\mathbf{Z} \tag{11}$$

In order to inverse $\mathbf{Z}$, it first has to be transformed into a square matrix by multiplying both sides by $\mathbf{Z}^T$.

$$\mathbf{Y}\mathbf{Z}^T = \mathbf{B}_{opt}\mathbf{Z}\mathbf{Z}^T \tag{12}$$

Then apply the inverse of $\mathbf{Z}\mathbf{Z}^T$ to obtain the optimal set of coefficients $\mathbf{B}$, $\mathbf{B}_{opt}$.

$$\mathbf{B}_{opt} = \mathbf{Y}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^{-1} \tag{13}$$

1.3.3 Stability of VAR model

For a VAR(1) process:

$$\mathbf{y}_t = \mathbf{A}\mathbf{y}_{t-1} + \mathbf{e}_t \tag{14}$$

With the previous time instant:

$$\mathbf{y}_{t-1} = \mathbf{A}\mathbf{y}_{t-2} + \mathbf{e}_{t-1} \tag{15}$$

Then

$$\mathbf{y}_t = \mathbf{A}^2\mathbf{y}_{t-2} + \mathbf{A}\mathbf{e}_{t-1} + \mathbf{e}_t \tag{16}$$

$$\mathbf{y}_t = \mathbf{Q}\Lambda^k\mathbf{Q}^{-1}\mathbf{y}_{t-k} + \sum_{i=0}^{k-1}\mathbf{A}^i e_{t-1} \tag{17}$$

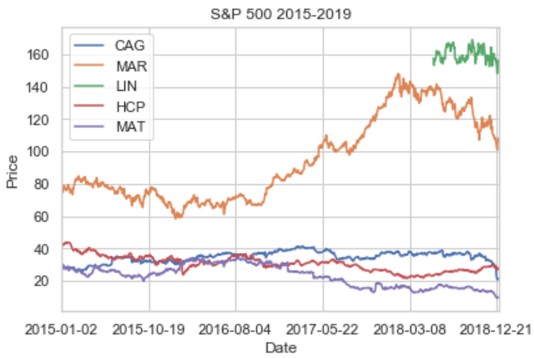
$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{A}_2 & \cdots & \mathbf{A}_{p-1} & \mathbf{A}_p \\ \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} \end{bmatrix} \tag{18}$$

Where Λ is the diagonal matrix with the eigenvalues λ, $\mathbf{Q}$ is the eigenvector matrix.

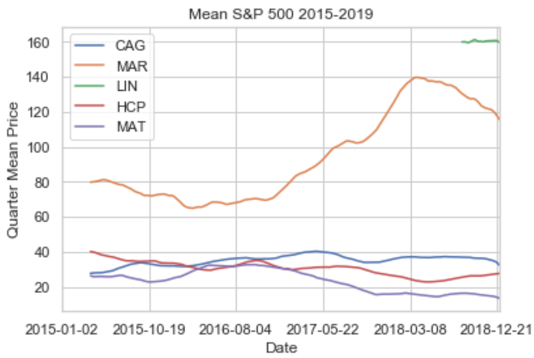
In order for the VAR(1) process to be stable, the roots of $|\mathbf{A} - \lambda\mathbf{I}| = 0$ has to be within the unit circle which all the eigenvalues of the matrix $\mathbf{A}$ must be less than 1 in absolute value for the process to be stable.

1.3.4 VAR model of the stocks with tickers CAG, MAR, LIN, HCP, MAT

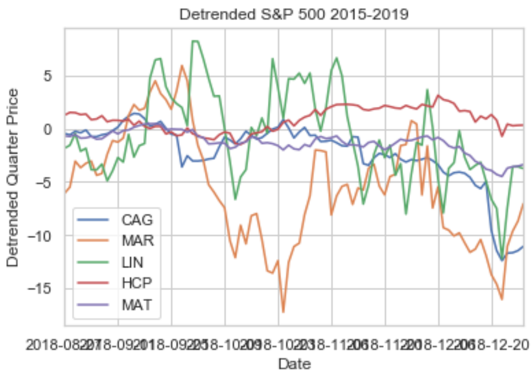
The following graphs are the original price, rolling seasonal mean price and rolling seasonal detrended price of stock CAG, MAR, LIN, HCP, and MAT. From figure 13, it is near impossible to interpret the stationarity and similarity between stocks. The rolling seasonal mean of each stock is plotted; they are the smoother version of its stock price. The detrended price is the original stock price with the linear trend and mean removed. All five lines cross each other at random occasions and with no similar trends indicates if these stocks were to use to construct a portfolio, the risk of the portfolio would be spread and diversified as half of the stocks can rise and the other half could drop over the same period. Therefore, it would make sense to construct a low-risk portfolio with these stocks.



(a) Simple returns



(b) Log returns



(c) Log returns

Figure 11: S&P 500 Index simple and log return from 1930 to 2017

The following table shows the stock symbols with their absolute eigenvalues modelled by the VAR(1) process. All of the stocks except for stock LIN have an absolute eigenvalue of lower or equal to 1; this suggests the stock LIN is unstable and should not be used to construct a portfolio.

Table 2: Caption

Symbol	Absolute Eigenvalue
CAG	0.726094
MAR	0.726094
LIN	1.006360
HCP	0.860519
MAT	0.911445

The following colour map shows the correlation between each stock, the green or red colour suggests there is a positive or negative correlation between the stocks across the x and y-axis respectively. The off-white colour suggests there is zero correlation between the stocks. The correlation colour map suggests the stock HCP has zero to negative correlation to the rest of the stocks shown in the colour map; the green colour suggests there is a strong correlation between the stocks MAT, CAG, and MAR, all three of the stocks belong to the consumer sector. Since there is a correlation between these stocks, in order to spread the risk, less than 50% of the portfolio should be placed on those three stocks to spread out the overall risk.

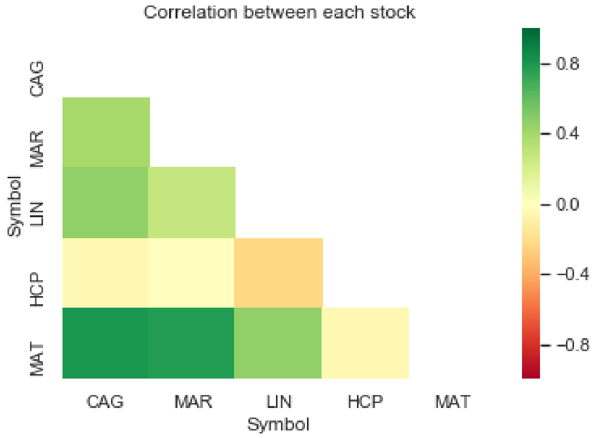
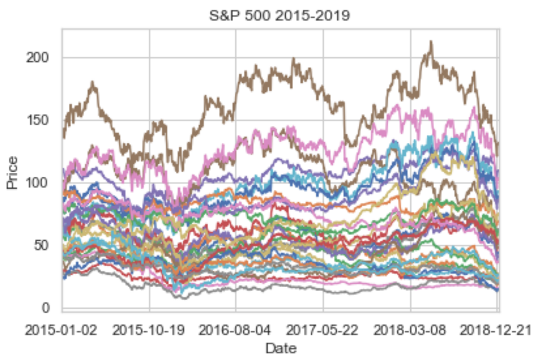


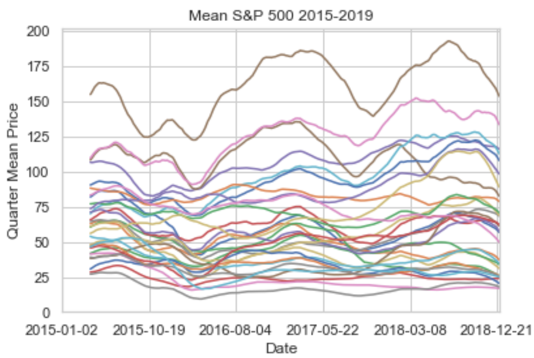
Figure 12: Correlation between selected stocks

1.3.5 VAR model on stocks within the same sector

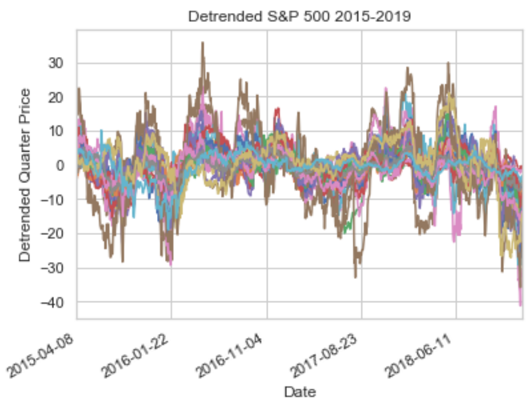
Figure 13a, 13b, 13c show the price, rolling seasonal mean price, and rolling seasonal detrended price of the stocks in the energy sector. In figure 13c, all of the stocks have roughly the same rises and drops (excluding the magnitude) suggest they are all influenced by the same noise source (news).



(a) Simple returns



(b) Log returns



(c) Log returns

Figure 13: S&P 500 Index simple and log return from 1930 to 2017

The correlation colour map below suggests all the stocks in the energy is somewhat correlated to each other supporting all of the stocks in the energy sector get influenced by the same type of news. Depending on the type of portfolio, grouping stocks of the same sector could be good because it only requires tracking on one type of news, the energy-related news in this case. Therefore, if there is a report such as a release of a weekly inventory report of some energy-related commodities, all of the stocks in the portfolio are very likely to move in the same direction, a quick execution of short or long can minimise the losses and increase the profit of the portfolio.

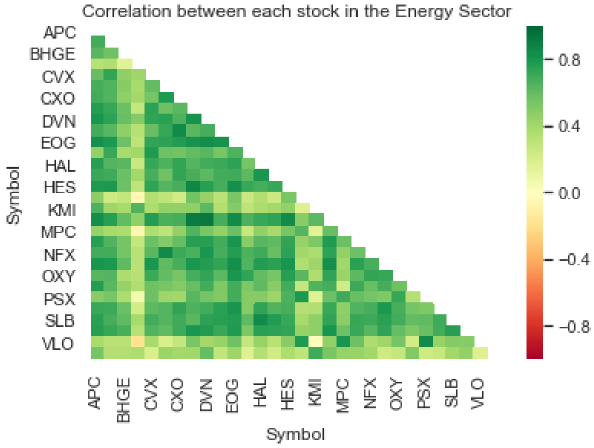


Figure 14: Correlation between each energy sector stock

For a portfolio with different sectors of stocks, it is possible to use stocks that in the same sector in the market to model the trend of the stock in the portfolio since it's been shown that same sector of stocks is correlated.

2 Bond Pricing

2.1 Examples of bond pricing

1. An investor receives USD 1,100 in one year in return for an investment of USD 1000 now. The investor receives USD r with an initial investment of USD a with a risk-free rate constant at $i\%$ paid k times a year. The return can be calculated with the following equation.

$$r = a \times \left(1 + \frac{i}{k}\right)^k \tag{19}$$

The compounding interest can be calculated with the following equation.

$$i = k \times \left[\exp\left(\frac{1}{k}\ln\left(\frac{r}{a}\right)\right) - 1\right] \tag{20}$$

(a)

$$\text{Annual compounding interest} = \frac{1100}{1000} - 1 = 10\% \tag{21}$$

(b)

$$\text{Semiannual compounding interest} = 2 \times \left[\exp\left(\frac{1}{2}\ln\left(\frac{1100}{1000}\right)\right) - 1\right] = 9.762\% \tag{22}$$

(c)

$$\text{Monthly compounding interest} = 12 \times \left[\exp\left(\frac{1}{12}\ln\left(\frac{1100}{1000}\right)\right) - 1\right] = 9.569\% \tag{23}$$

(d)

$$\text{Continuous compounding interest} = \ln(1.1) = 9.531\% \tag{24}$$

2. What rate of interest with continuous compounding is equivalent to 15% per annum with monthly compounding?

$$\text{Continuous compounding interest} = \ln\left(\left(1 + \frac{15\%}{12}\right)^{12}\right) = 14.907\% \tag{25}$$

3. A deposit account pays 12% per annum with continuous compounding, but interest is actually paid quarterly. How much interest will be paid each quarter on a USD 10,000 deposit?

$$\text{Annual compounding interest} = \exp(12\%) - 1 = 12.750\% \tag{26}$$

$$\text{Quarterly interest payment} = \$10000 \times \frac{12.750\%}{4} = \$318.74 \tag{27}$$

2.2 Forward rates

1. Given that the one-year interest rate, r_1 is 5%, and the two-year interest rate, r_2 is 7%. The extra return that I would earn for the second year is $\frac{1.07^2}{1.05} - 1 = 9\%$.

   (a) I would be happy to earn that extra 9% for investing for two years rather than one given that that money is my saving account and the interest rate at the time is higher than the inflation or I would be losing money using the present value of money.

   (b) The 5% and 7% investment strategies are the amount of money that would grow if the investor invests in one year and two years respectively. The 9% is the forward rate that one would earn between both years; it exists to avoid arbitrage opportunities.

   (c) Forward rate depends on the current interest rate curve which is not fixed; it changes as time goes, which is there might have interest rate risk. The advantage is that it void arbitrage opportunities, but the downside is there exist a chance where the one-year interest rate might increase after one year resulting the overall payout is higher than the 7% two-year interest rate.

   (d)

2.3 Duration of a coupon-bearing bond

1. The duration is the weighted average of the times to each of the cash payment where the weight for each year is the present value of the cash flow received at that time divided by the total present value of the bond.

$$\text{Duration} = \sum_{t=1}^{T} \frac{t \times PV(C_t)}{PV} \tag{28}$$

   (a) The duration for the 1% bond

$$= 0.0124 + 0.0236 + 0.0337 + 0.0428 + 0.0510 + 0.0583 + 6.5377 = 6.7595$$

   Which is sensible as it is less than the maturity date which is 7 and close enough.

   (b)

$$\text{Modified duration} = \frac{6.7595}{1 + 5\%} = 6.438\% \tag{29}$$

   The modified duration the yield would fluctuate by a rate of 6.438%. Note that as the maturity is proportional to the duration and volatility of the bond, as the bond coupon increases, its duration decrease causing the bond to become less volatile. At last, as the interest rates increase, duration decreases and the bond's sensitivity to further interest rate increases goes down.

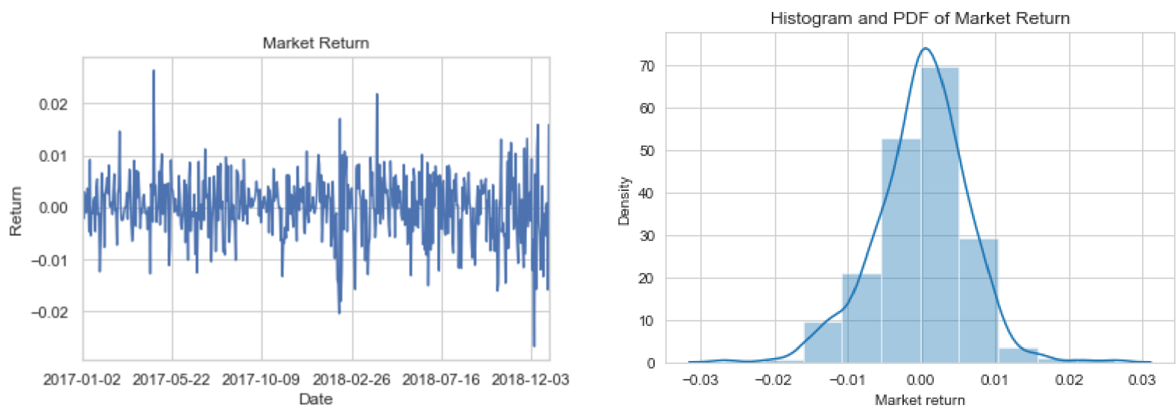
   (c) Modified duration takes into account yield which affects the present value which affects the pricing of the bond. Therefore, for an unexpected change of interest rate, by using the modified duration to calculate the amount of time to reach an obligation is the preferable option.

2.4 Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT)

The capital asset pricing model (CAPM) describes the relationship between systematic risk and expected return for assets, in this case, stocks. It is used for pricing risky securities and generating expected returns for assets given the risk of those stocks. The data used in this section consist of the daily returns of 157 European companies from the beginning of 2017 to the end of 2018.

2.4.1 Market returns

The following figure shows the mean return (market return) over 157 European companies and its histogram.



(a) (b)

Figure 15: Market return over time and its histogram

2.4.2 Sliding beta

The beta of each stock in CAPM is a measure of how much risk the investment will add to a portfolio relative to the market. A high beta suggests that stock is riskier than the market, and a beta value of less than one would reduce the risk of the portfolio. It represents the gradient of the regression data from the corresponding stock returns against the market.

$$\text{Beta coefficient } \beta = \frac{\text{Covariance}(R_i, R_m)}{\text{Variance}(R_m)} \tag{30}$$

where the index i refers to a particular company, R_i is the daily return with a rolling window of 22 days, R_m is the market return. The covariance of R_i and R_m describe how the stock of company i return is related to the market return.

The following figures show the histogram of mean beta of each stock; it shows the beta of stocks has a sliding beta centred around 1 (market volatility), and the majority of stocks are around 0.5 to 1.5 times less and more volatile as the market.

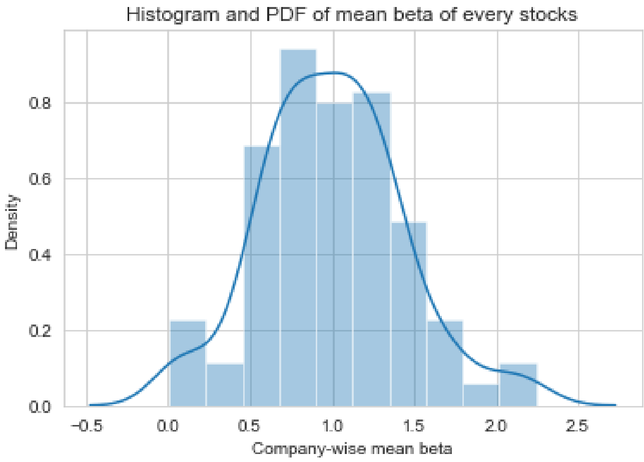
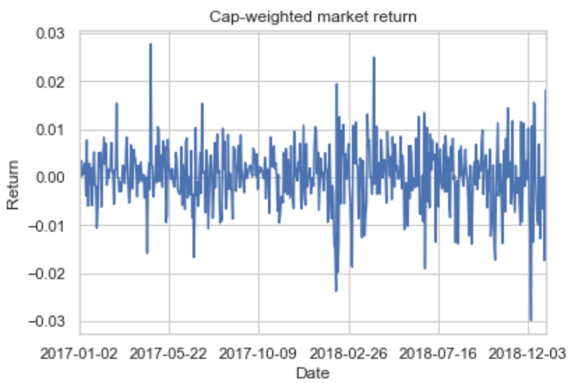


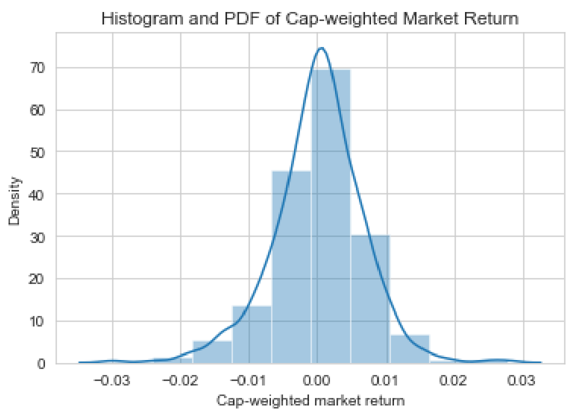
Figure 16: Mean of each stock's sliding monthly beta

2.4.3 Cap-weighted market return

The cap-weighted market return takes into account the company size before calculating the corresponding stock return by using the weighting coefficient $\sum_i \text{mcap}_i$ giving a more robust market return. The following figure shows the cap-weighted market return for every day. It can be seen that the cap-weighted market return is a weighted version of the market return shown in figure 15a.



(a)



(b)

Figure 17: Cap-weighted market return over time and its histogram

2.4.4 Cap-weighted beta

As the cap-weighted market return is used to calculate the cap-weighted beta, it is expected that the new beta values have a lower variance and leaning a beta value of below 1; as the larger companies have more effect on the market returns, the volatility experienced by the smaller companies does not affect the overall market, and larger companies have a more stable stock price. The following figure shows the histogram of the mean cap-weighted beta of each stock. Both of the betas are centred around 1; the cap-weighted beta shows that only 4 bins in the histogram have a density equal or higher than 0.6.

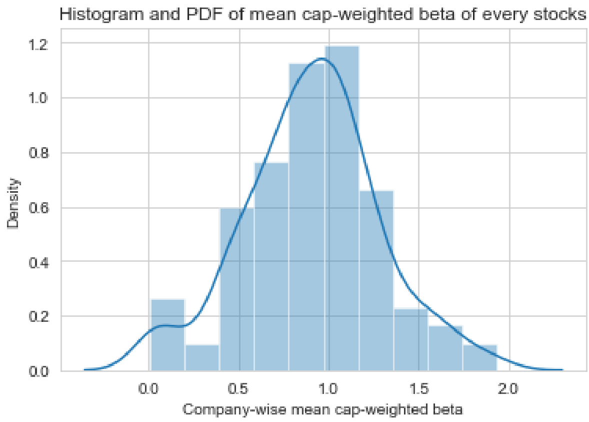


Figure 18: Mean of each stock's sliding monthly beta

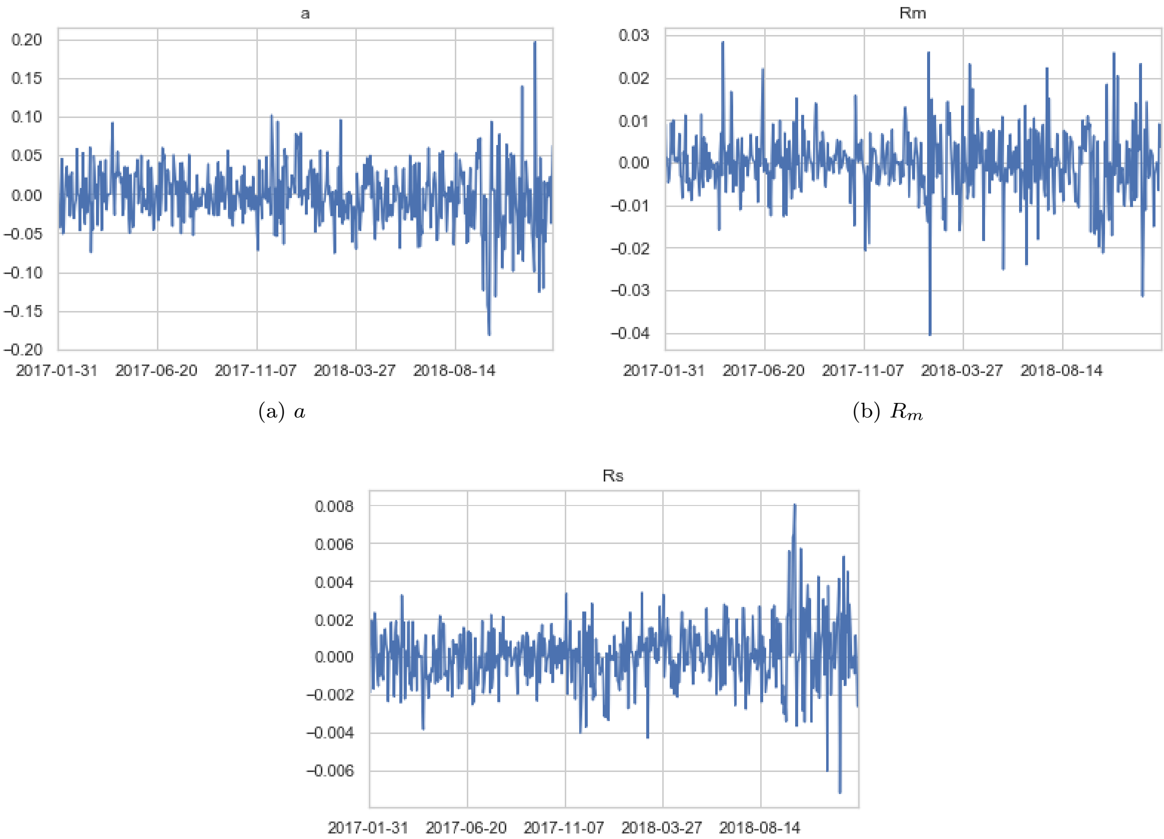
2.4.5 Arbitrage pricing theory

Arbitrage pricing theory (APT) is a multi-factor asset (stock) pricing model based on a stock's return can be predicted using the linear relationship between the stock's expected return and a number of macroeconomic variables that capture systematic risk. APT assumes markets sometimes misprice securities before the market eventually corrects add securities move back to fair value.

The arbitrage pricing theory describe here uses a two-factor model using the exposure b_{s_i} ln(size) and cap-weighted beta β_{m_i} as the cross-sectional regression coefficients for company i at each time instant t, r_i is the return per company and ϵ_i is the residual of this regression (specific return).

$$\begin{bmatrix} r_1(t) \\ r_2(t) \\ \vdots \\ r_i(t) \end{bmatrix} = \begin{bmatrix} 1 & b_{m_1}(t) & b_{s_1}(t) & \epsilon_1(t) \\ 1 & b_{m_2}(t) & b_{s_2}(t) & \epsilon_2(t) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & b_{m_i}(t) & b_{s_i}(t) & \epsilon_i(t) \end{bmatrix} \begin{bmatrix} a \\ R_m \\ R_s \\ 1 \end{bmatrix} \tag{31}$$

The following figures show the estimated a, R_m, and R_s throughout 2017 to 2018.



(a) a

(b) R_m

(c) R_s

Figure 19: Arbitrage pricing theory two-factor model coefficients

Table 3: Regression coefficients mean and variance

Coefficients	Mean	Variance
a	-0.0022989	0.0015541
R_m	-0.00030324	6.1911e-05
R_s	0.00011159	2.7483e-06

Coefficient a represents the estimated risk-free rate of return; its magnitude ranges from -0.2 to 0.2 with a mean of -0.0022989. This suggests around -0.223% of the returns are risk-free but negative. The risk-free rate of return variance suggest its volatility, and it is relatively higher than the variance of R_m and R_s.

Coefficient R_m is the estimated market return; its range of magnitudes closely matches the cap-weighted market return shown in figure 17a. The estimated market return mean and variance is small suggest the volatility of the market do not have a significant impact on each stock's return.

Coefficient R_s has the smallest variance compared to the other two coefficients suggest the exposure (size) of the company does not have a large impact on predicting the return of its company's stock.

Since the mean of b_s and b_m is around 23.4 and 1 respectively, the majority the return described in equation 31 comes from the specific return $\epsilon_i(t)$. Figure 20 shows the correlation between each company-specific return and its daily return, and figure 21 shows the corresponding histogram, the high correlation shown in both figures suggest each stock's daily return is highly specific to its own company's performance.

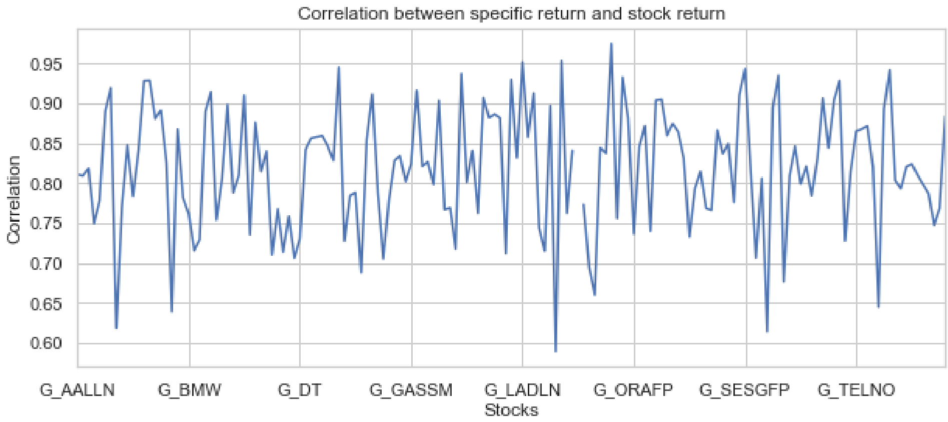


Figure 20: Correlation between specific return and each stock's return

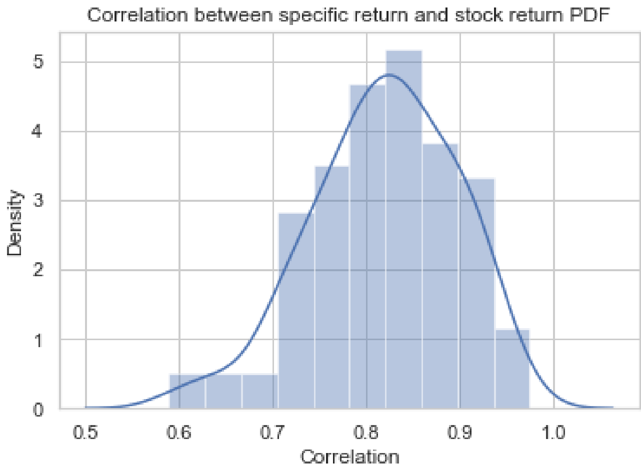


Figure 21: Histogram of the correlation between specific return and each stock's return

The coefficients R_m and R_s are used to create the matrix $\mathbf{R}$.

$$\mathbf{R} = \begin{bmatrix} R_{m_1} & R_{s_1} \\ \vdots & \vdots \\ R_{m_{500}} & R_{s_{500}} \end{bmatrix} \tag{32}$$

A rolling window of 22 days is applied before calculating its covariance matrix. Each covariance matrix forms a Hurwitz matrix where it is used to describe the stability of the covariance matrix. In figure 22, both of the eigenvalues magnitude are positive suggesting the system (stock market) is unstable throughout the measured duration.

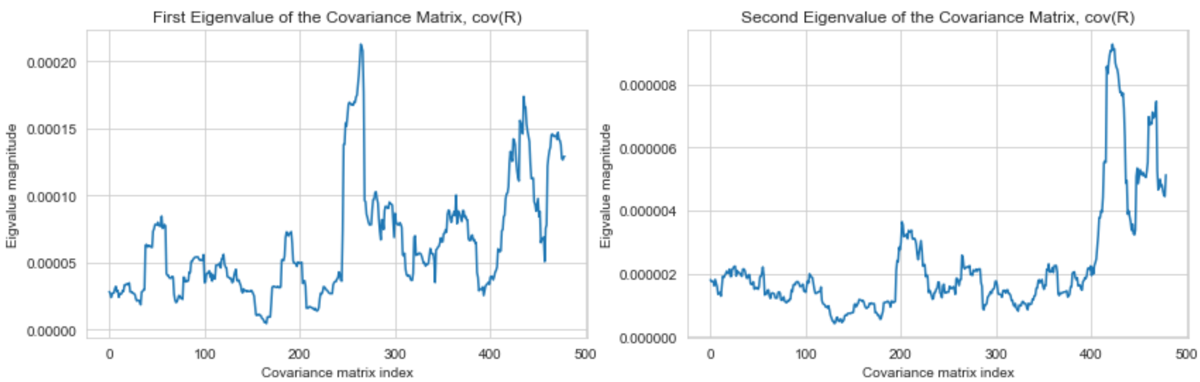


Figure 22: Covariance matrix eigenvalues

The obtained specific return is used to form the following matrix **E**.

$$\mathbf{E} = \begin{bmatrix} \epsilon_{1,t=0} & \cdots & \epsilon_{157,t=0} \\ \epsilon_{1,t=1} & \cdots & \epsilon_{157,t=1} \\ \vdots & \ddots & \vdots \\ \epsilon_{1,t=500} & \cdots & \epsilon_{157,t=500} \end{bmatrix} \tag{33}$$

The following figure shows the percentage of variance each principal component (stock) carries.

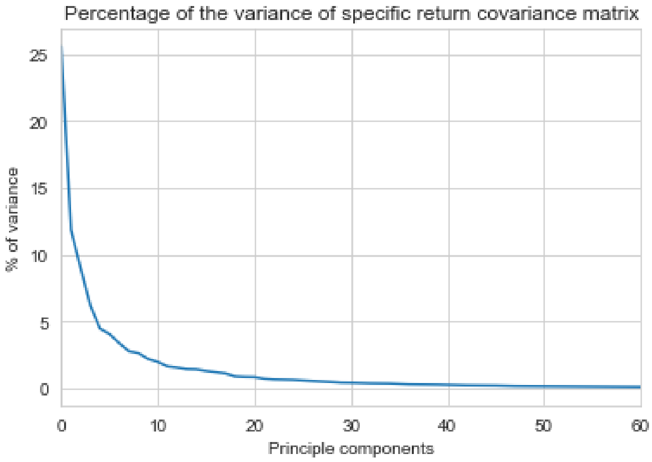


Figure 23: Percentage of variance on each principle components of matrix **E**

The top four principal components contribute 25.58%, 11.86%, 9.03%, and 6.20% of the variance; this suggests that only four of the stocks are responsible for more than 50% of the variance (risk). In portfolio management, we would like to choose weights to make the expected return large and the risk to be as small as possible. In other words, we would want the % of the variance of each principal component carries are roughly the same or equal, spreading the risk over the whole portfolio.

3 Portfolio Optimisation

3.1 Adaptive minimum-variance portfolio optimisation

The holdings of a portfolio can be represented by a set of weights $\mathbf{w} = [w_1, \cdots, w_M]^T$ that we apply to our assets. The weight w_m represents the percentage of our capital which is allocated to asset m. The return of the portfolio at time t can be calculated with $\bar{r}[t] = \mathbf{w}^T\mathbf{r}[t]$. Minimum variance portfolio is used in portfolio management to minimise the risk (variance) of the portfolio. The set of optimum weights can be found by minimising the following cost function J.

$$\min_{\mathbf{w}} \quad J(\mathbf{w}, \mathbf{C}) = \frac{1}{2}\mathbf{w}^T\mathbf{C}\mathbf{w} \tag{34}$$

$$\text{subject to} \quad \mathbf{w}^T\mathbf{e} = 1 \tag{35}$$

where $\mathbf{e}$ is a column vector of ones, $\mathbf{e} \in R^{M\times 1}$.

3.1.1 Derive optimal weights

Using the Lagrange optimisation, equation 34 and 35 becomes the following.

$$\min_{\mathbf{w},\lambda} \quad J(\mathbf{w}, \lambda, \mathbf{C}) = \frac{1}{2}\mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{e} - 1) \tag{36}$$

Differentiate equation 36 with respect to w and λ give the following.

$$\frac{\partial J}{\partial \mathbf{w}} = \mathbf{w}^T\mathbf{C} - \lambda\mathbf{e}^T = 0 \tag{37}$$

$$\frac{\partial J}{\partial \lambda} = \mathbf{w}^T\mathbf{e} - 1 = 0 \tag{38}$$

Solving for $\mathbf{w}$ and λ, note that because $\mathbf{C}$ is a covariance matrix, $\mathbf{C}^T = \mathbf{C}$.

$$\mathbf{w}^T\mathbf{C} = \lambda\mathbf{e}^T$$

$$\mathbf{C}\mathbf{w} = \lambda\mathbf{e}$$

$$\mathbf{w} = \lambda\mathbf{C}^{-1}\mathbf{e} \tag{39}$$

$$\mathbf{w}^T = \lambda\mathbf{e}^T\mathbf{C}^{-1} \tag{40}$$

Substituting equation 40 into equation 35.

$$\lambda\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e} = 1$$

$$\lambda = \frac{1}{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}} \tag{41}$$

The optimal weights can be obtained by substituting λ into equation 39.

$$\mathbf{w}_{opt} = \frac{\mathbf{C}^{-1}\mathbf{e}}{\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e}} \tag{42}$$

The theoretical variance when optimal weights are used:

$$\sigma_{opt}^2 = \mathbf{w}_{opt}^T\mathbf{C}\mathbf{w}_{opt} = \lambda\mathbf{e}^T\mathbf{C}^{-1}\mathbf{C}\lambda\mathbf{C}^{-1}\mathbf{e} = \lambda^2\mathbf{e}^T\mathbf{C}^{-1}\mathbf{e} = \lambda = \frac{1}{\text{sum}\big(\mathbf{C}^{-1}\big)} \tag{43}$$

3.1.2 Compare optimal weights and equally-weighted portfolio

The last 10 stocks from section 2.4 are used in this section; the first half of the data is used as training for the optimal weights, then the other half of the data is used for testing. The optimal weights are compared with equally-weighted portfolio with $w_m = \frac{1}{M}$ for all m. The performance of each strategy over a time-horizon of T days is measured by the portfolio variance $\bar{\sigma}^2$ and the cumulative return.

$$\text{Cumulative return:} \quad R[T] = \sum_{t=1}^{T}\bar{r}[t] \tag{44}$$

The following figure shows the cumulative return over all the data using optimal weight and equally-weighted portfolio.

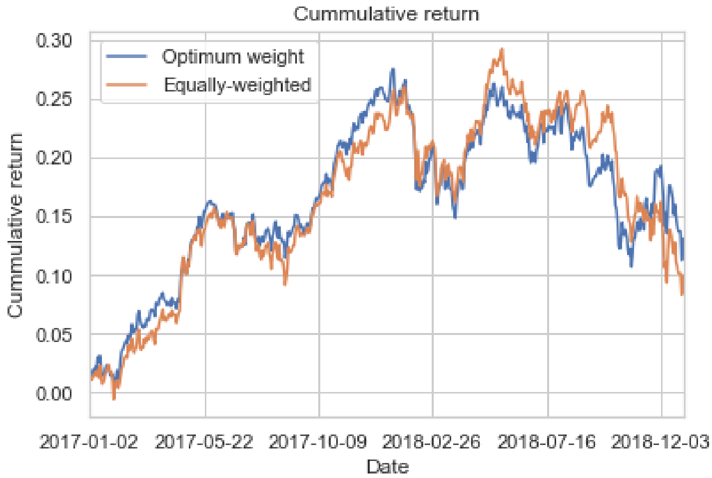


Figure 24: Cumulative return on 10 stocks using optimum weights and equal weights

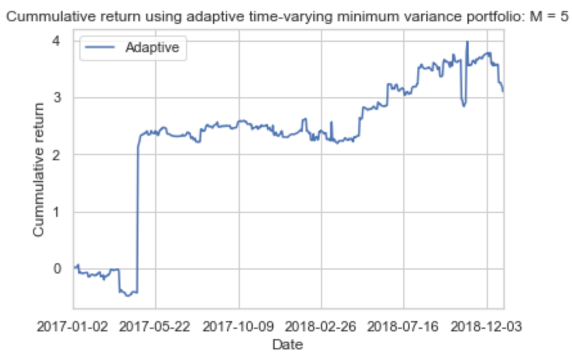
Table 4: Different strategy's portfolio variance comparison

Strategy	Data	
	Training	Testing
Optimal weights	0.00741	0.0213
Equally-weighted	0.00971	0.0206

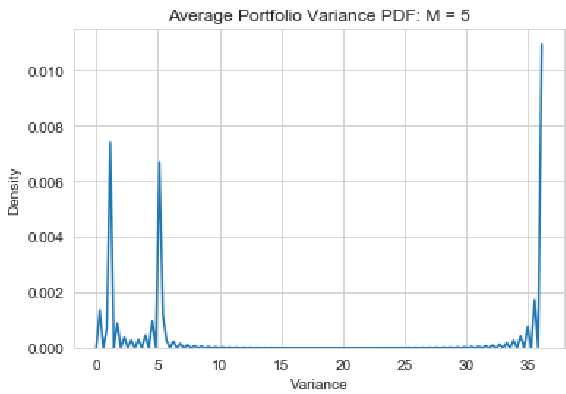
The optimal weights portfolio has a higher cumulative return over the first half of the plot in figure 24 than using the equally-weighted portfolio. Then in the second half of the data, it can be seen that both strategies perform approximately well as shown in table 4 as both strategies' variance is approximately the same. The optimal weights are overfitted to the training dataset. Hence it is not the best weights for the testing dataset as the theoretical minimum variance for the testing data is 0.0152 where the equally-weighted portfolio variance is closer to the theoretical minimum variance. This is the case as the equally-weighted portfolio assumes all of the stocks have equal risk and chance of obtaining returns, the optimal weights assume the each stock's return follows the same trend in the training data.

3.1.3 Adaptive weights portfolio

A recursive update of portfolio weights strategy (adaptive time-varying minimum variance portfolio) is used, figure 25a shows different rolling window length M applied. The past M^{th} days daily return is used to calculate the optimal weights for the current day.



(a) Cumulative return

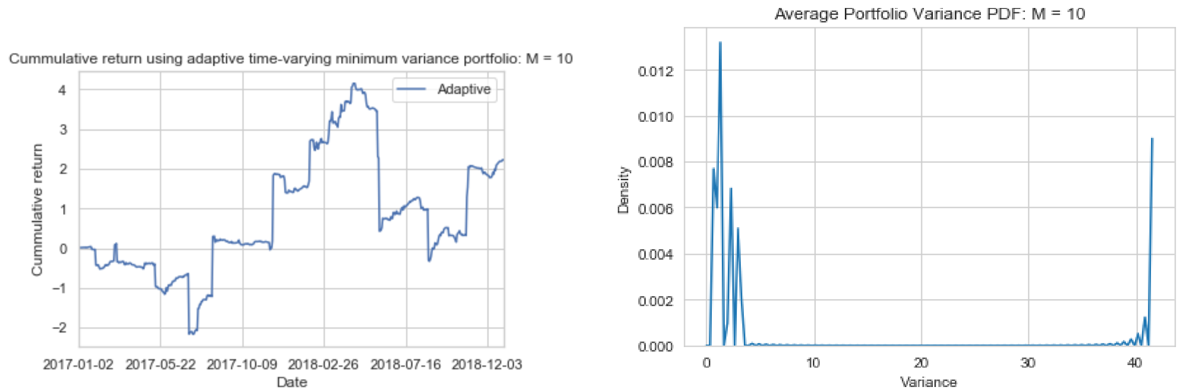


(b) Variance histogram

Figure 25: Adaptive time-varying minimum variance portfolio using a weekly rolling window

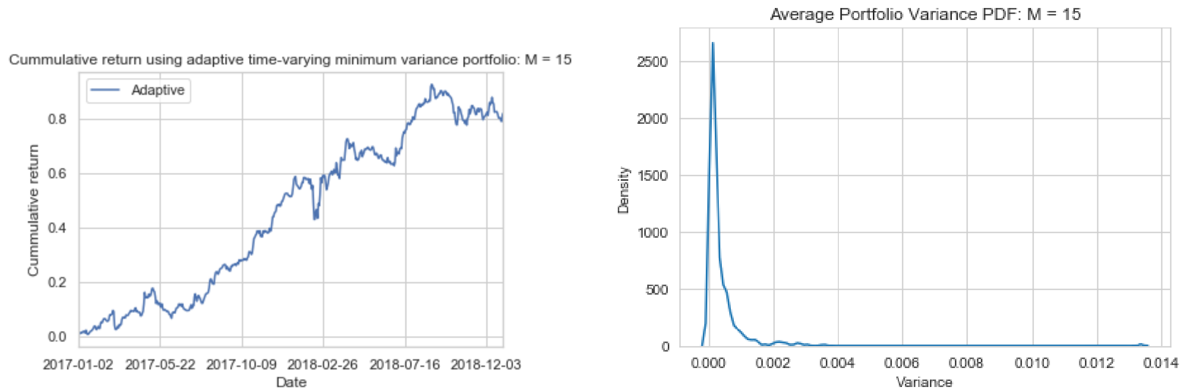
Figure 25a shows that $M = 5$ has the highest final cumulative return; however, this window length has performed some risky trades as seen in figure 25b. The one weight combination with portfolio variance near 35 is responsible for the big jump in cumulative return as a high-risk with high return.

The following figure shows an example where the high-risk results in a high lost situation. The portfolio variance of roughly 40 corresponds to the sharp drop of cumulative return resulting in an approximate 300% lost. Therefore, a balance of risk and return need to establish to minimise lost and maximise the rate of return.



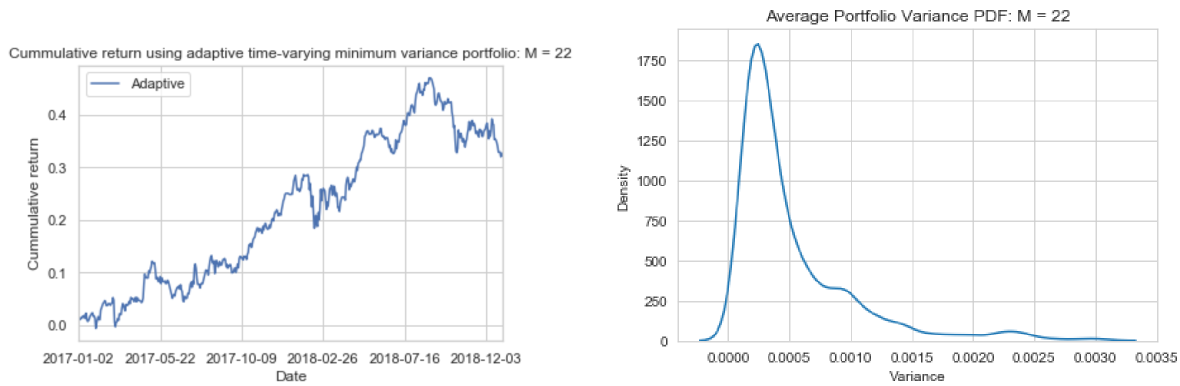
(a) Cumulative return (b) Variance histogram

Figure 26: Adaptive time-varying minimum variance portfolio using a two weeks rolling window



(a) Cumulative return (b) Variance histogram

Figure 27: Adaptive time-varying minimum variance portfolio using a three weeks rolling window



(a) Cumulative return (b) Variance histogram

Figure 28: Adaptive time-varying minimum variance portfolio using a monthly rolling window

It can be seen that as the rolling window length increase to three weeks, the majority of the portfolio variance of all weights used is below 0.001, suggest that this window length has a low-risk. At the same time, the final cumulative return is dropped.

The following figure shows the relationship between the final cumulative return with corresponding variance and window length. It can be seen that the final cumulative return is higher for a shorter window length, this indicates that the daily return is influenced by a shorter period of time, but at the same time it is riskier. It is found that $M = 15$ is the optimal window length for adaptive time-varying minimum variance portfolio. As the window length increases, irrelevant information is used increasing the risk. In general, the adaptive time-varying minimum variance portfolio performs much better than equally-weighted and single optimal weight calculation in terms of the rate of return and risk.

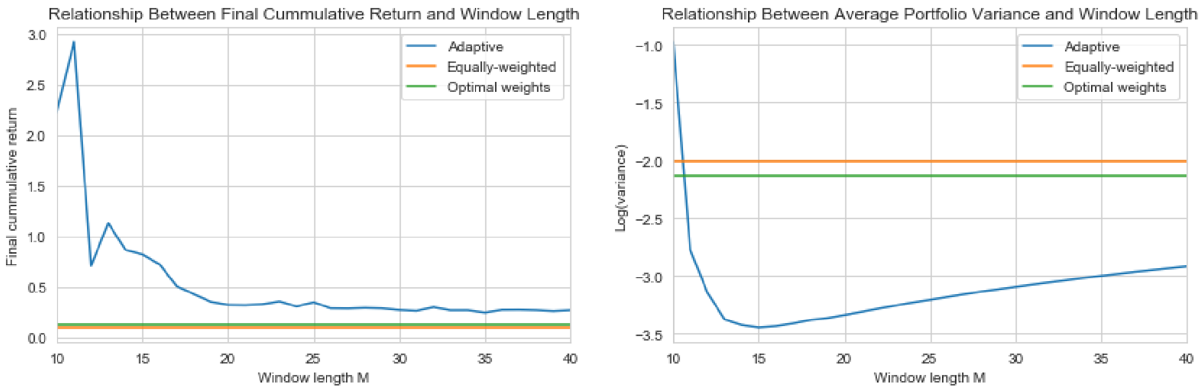


Figure 29: Relationship between window length and cumulative return

Calculating the covariance is the most computational expensive calculation in each weight update iteration, the covariance matrix calculation has a computational complexity of $\mathcal{O}(NM^2)$ with a window length M and N number of stocks. The sample covariance matrix can be estimated by using maximum likelihood estimator to reduce the computational complexity, but at the same time introduce bias and error into the covariance matrix.

4 Robust Statistics and Non Linear Methods

4.1 Data Import and Exploratory Data Analysis

4.1.1 Stock key statistics

Table 5: Key descriptive statistics of stock data

(a) Open

Stock	Mean	Median	Std
AAPL	187.687	186.29	22.146
IBM	138.454	142.81	12.114
JPM	108.708	109.18	5.359
DJI	25001.3	25025.60	858.835

(b) High

Stock	Mean	Median	Std
AAPL	189.562	187.40	22.282
IBM	139.492	143.99	11.913
JPM	109.652	110.53	5.203
DJI	25142	25124.1	815.204

(c) Low

Stock	Mean	Median	Std
AAPL	185.824	184.94	22.009
IBM	137.329	142.06	12.205
JPM	107.683	107.79	5.433
DJI	24846	24883	903.302

(d) Close

Stock	Mean	Median	Std
AAPL	187.712	186.12	22.161
IBM	138.363	142.71	12.028
JPM	108.607	109.02	5.300
DJI	24999.2	25044.3	859.132

(e) Adj Close

Stock	Mean	Median	Std
AAPL	186.174	184.352	21.905
IBM	134.903	138.566	10.672
JPM	107.263	107.219	4.833
DJI	24999.2	25044.3	859.132

(f) Volume

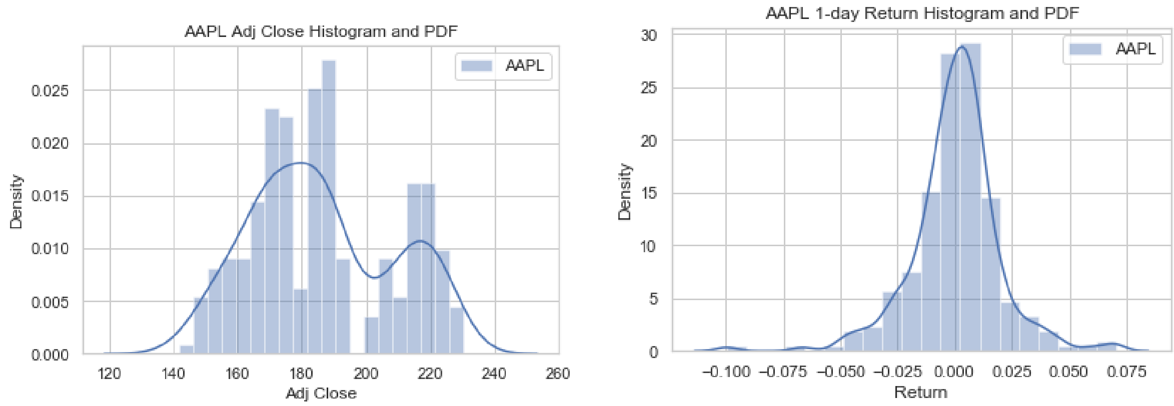
Stock	Mean	Median	Std
AAPL	3.27048e+07	29184000	1.4180e+07
IBM	5.19894e+06	4237900	3.3290e+06
JPM	1.47007e+07	13633000	5.3498e+06
DJI	3.32889e+08	313790000	9.4078e+07

(g) One-day return

Stock	Mean	Median	Std
AAPL	0.00042555	0.0016114	0.019323
IBM	-0.00025162	0.0004095	0.015562
JPM	-0.00013302	-0.0006026	0.013088
DJI	0.00019681	0.0003745	0.010476

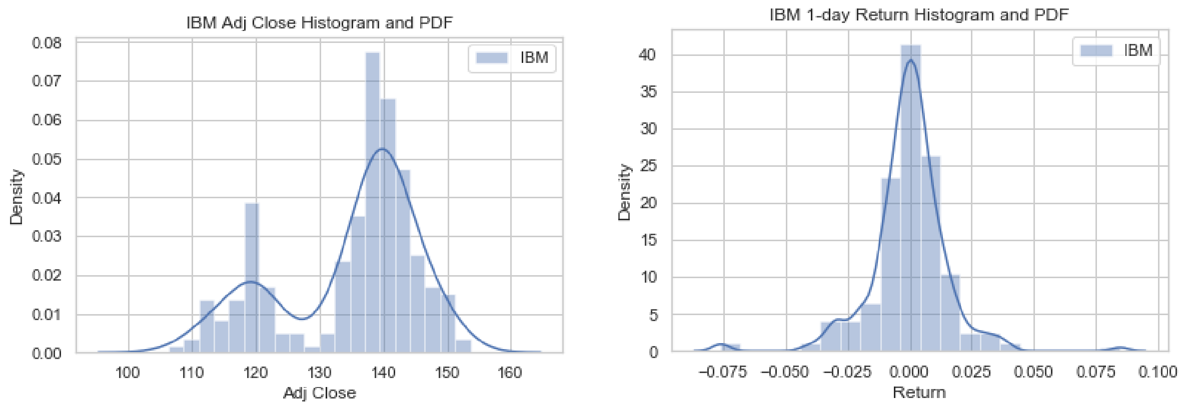
4.1.2 Histogram and probability density function of the stocks

The following figures show the histogram and probability density function of the adj. close and one-day return of AAPL, IBM, and JPM stocks and DJI index. The one-day return histogram's distribution is close to a Gaussian distribution than the adj. close histogram. Most of the adj. close distributions have more than one peak and the adj. close values have different mean and a large variance. The one-day distribution is centred around zero with a comparable variance in different stocks and indices. However, the one-day return histogram has more outliers than the adj. close histogram.



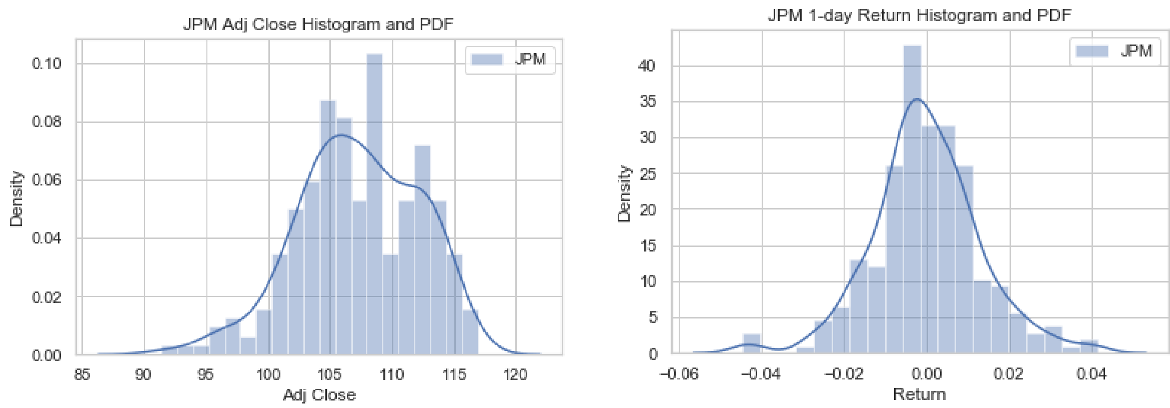
(a) Adj. close (b) One-day return

Figure 30: AAPL stock histogram and probability density function



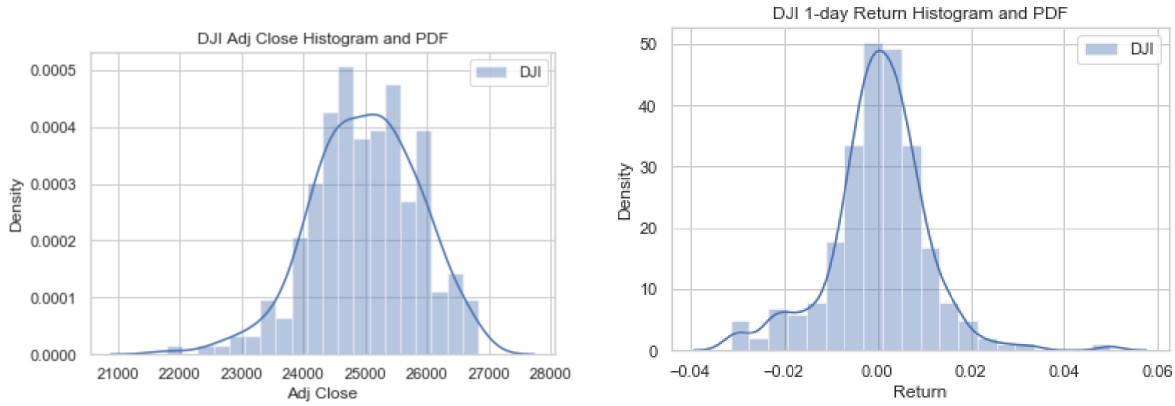
(a) Adj. close (b) One-day return

Figure 31: IBM stock histogram and probability density function



(a) Adj. close (b) One-day return

Figure 32: JPM stock histogram and probability density function



(a) Adj. close

(b) One-day return

Figure 33: DJI stock histogram and probability density function

4.1.3 Anomaly detection

In the anomaly detection with moving average, the rolling mean with a weekly window and the relative $\pm 1.5\times$ standard deviations are plotted. Another method used is to apply a rolling median with a weekly window and $\pm 1.5\times$ median absolute deviation (MAD) relative to the rolling median. They are applied to 4 different stocks and observations were made.

The lower and upper bound tightly wrapping the adj close of each stock in the rolling mean method, where the MAD's bound is more comprehensive, allowing more room for stock market fluctuations. As the median is known as a robust estimator, using the median to detect outlier is more reliable as a small number of outliers would not affect the bound for outlier detections. The rolling mean shows a smoother curve than the rolling median curve as the moving average is a low pass filter.

Both of the methods can be used as a trading strategy, as the price hits the lower or higher bound, a long or short order can be placed respectively with the centre (rolling mean or median) as the take profit. This bounds can be used as an indicator of if the current price is too high or low if the stock is the market is moving sideways instead of trending.

The following figures show both anomaly detection methods applied to 4 different stocks.

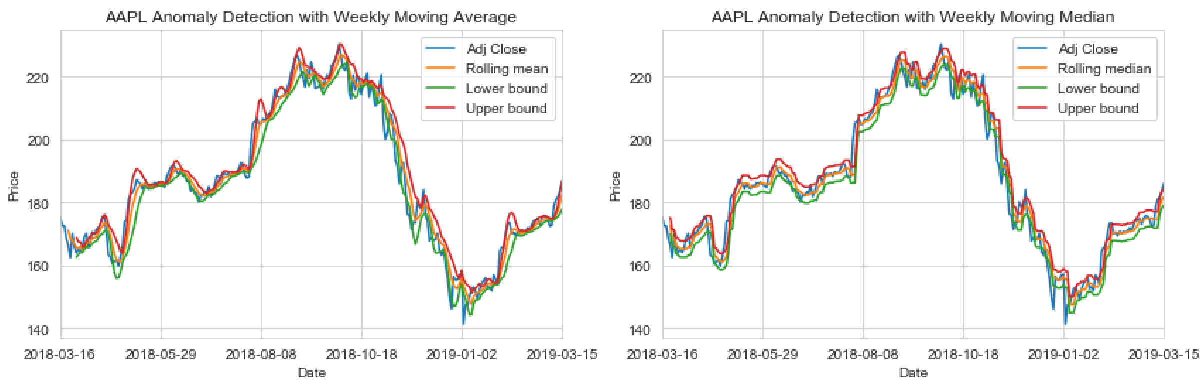


Figure 34: AAPL Anomaly detection using moving average and median

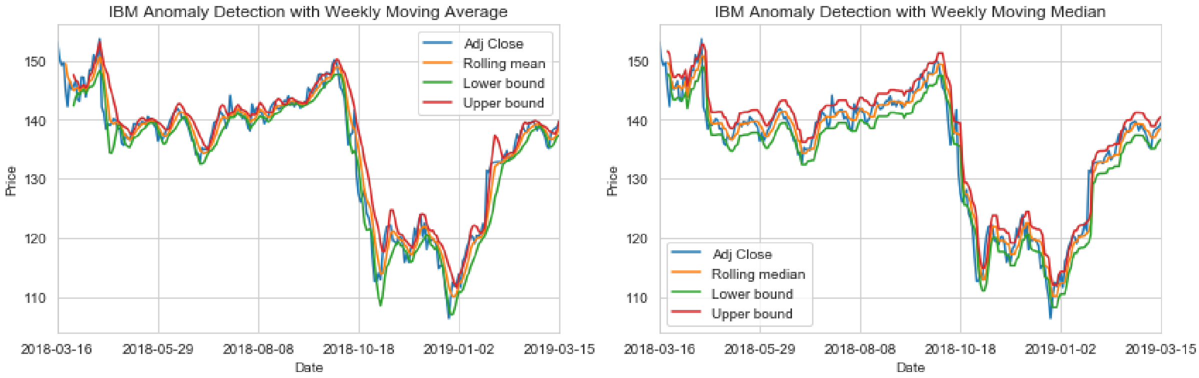


Figure 35: IBM Anomaly detection using moving average and median

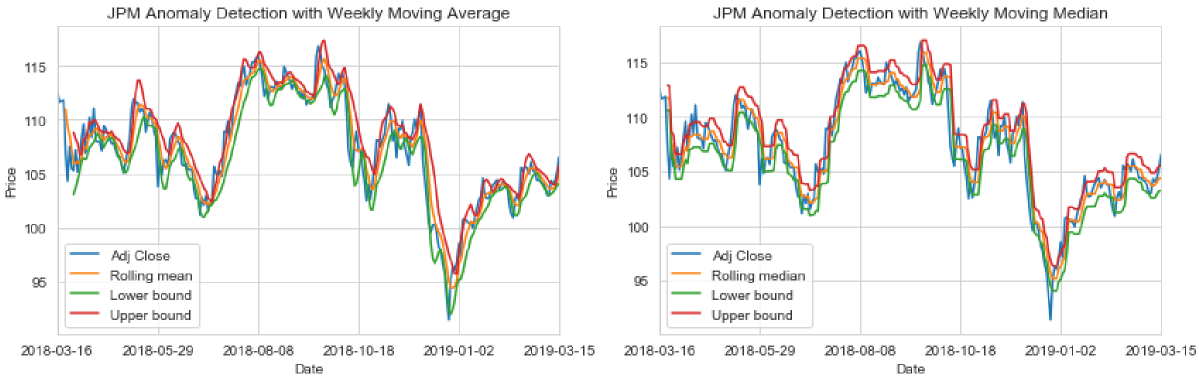


Figure 36: JPM Anomaly detection using moving average and median

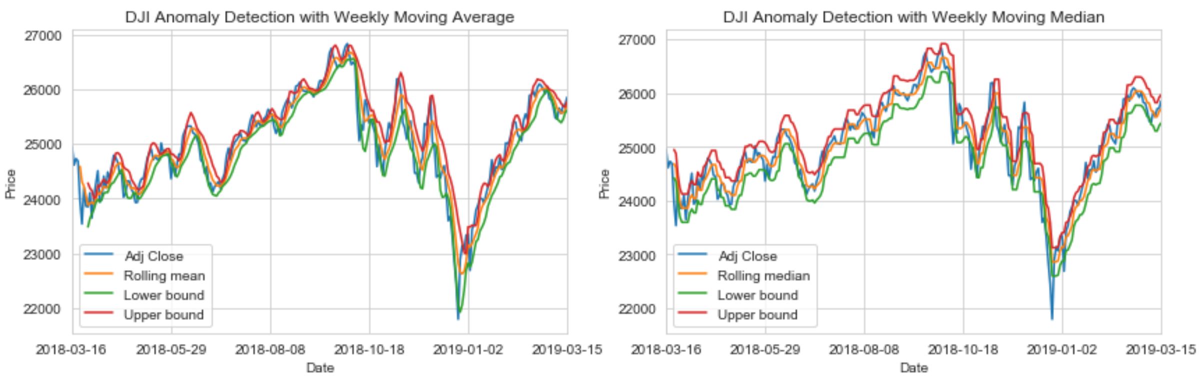


Figure 37: DJI Anomaly detection using moving average and median

4.1.4 Testing anomaly detection performance

Outliers are introduced to the adj close data in four dates: 2018-05-14, 2018-09-14, 2018-12-14, 2019-01-14 with a value equal to 1.2× the maximum price throughout time for each stock. It can be seen from the figures 38, 39, 40, 41 that the magnitude of the bound changes as the outlier is within the weekly window range in the moving average method, and the MAD's bound is approximately the same as shown in the last section. The rolling mean and its bound is highly offset by the outlier in the next four days period that's defined by the window used, making anomaly detection faulty in that period. With the rolling median, the median won't be offset by a significant amount. Although both methods capable of detecting all of the introduced outliers, it is clear that the rolling median with MAD bound anomaly detection method is more robust.

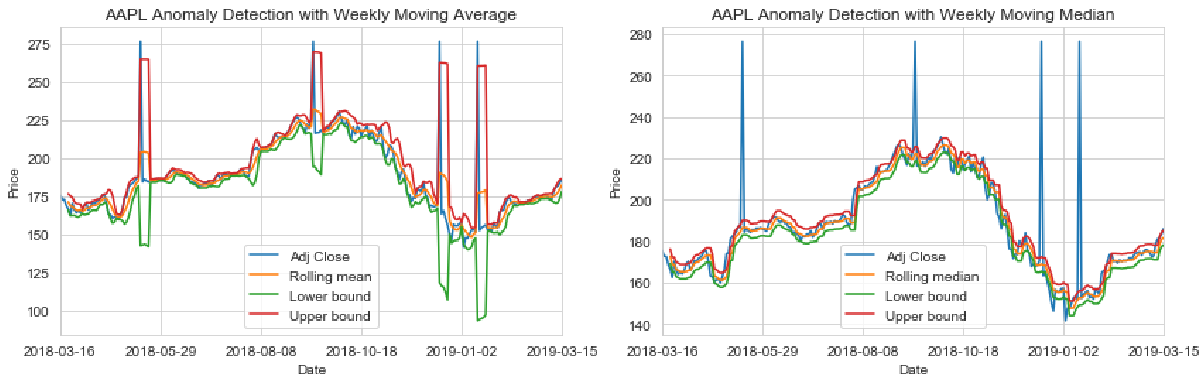


Figure 38: AAPL Anomaly detection using moving average and median

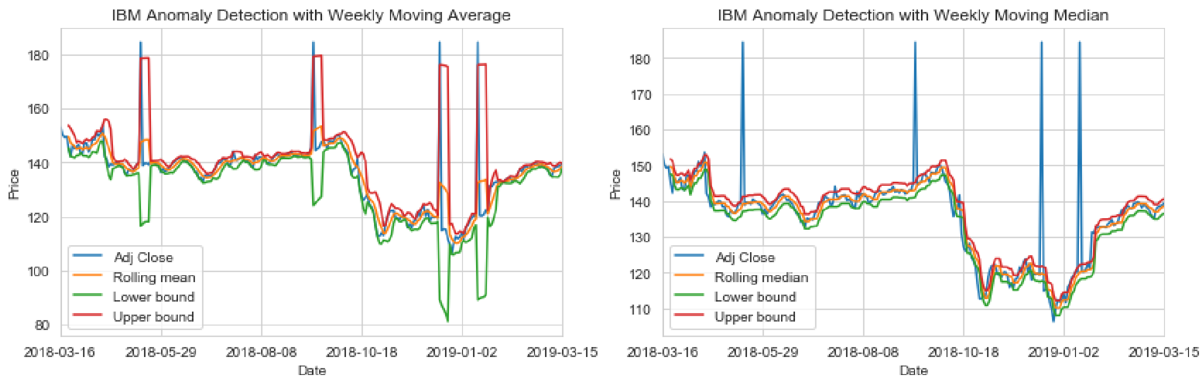


Figure 39: IBM Anomaly detection using moving average and median

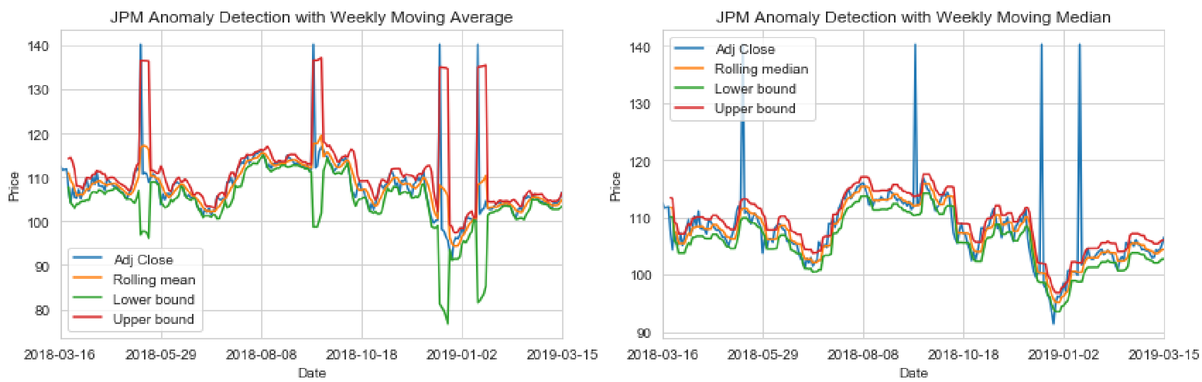


Figure 40: JPM Anomaly detection using moving average and median

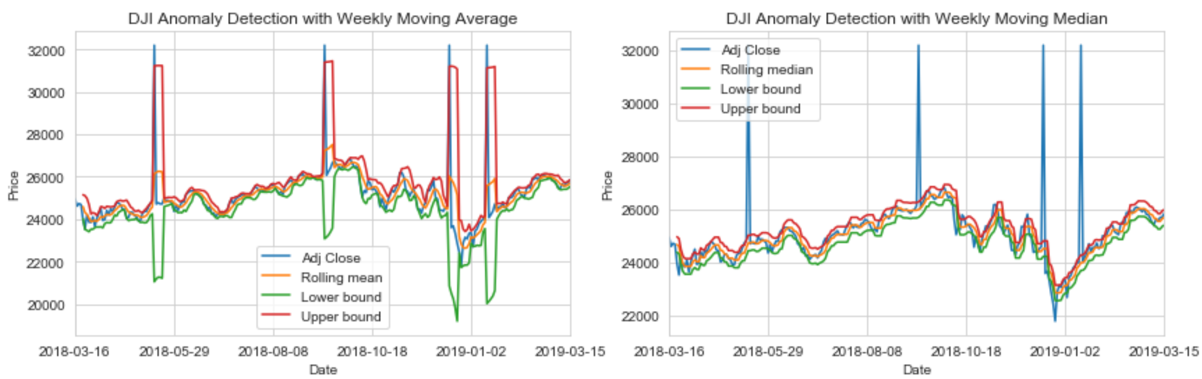
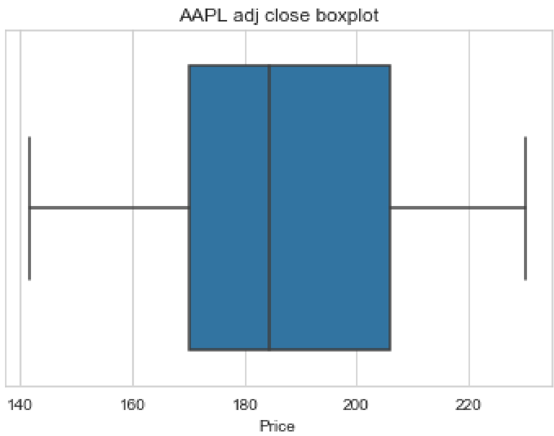


Figure 41: DJI Anomaly detection using moving average and median

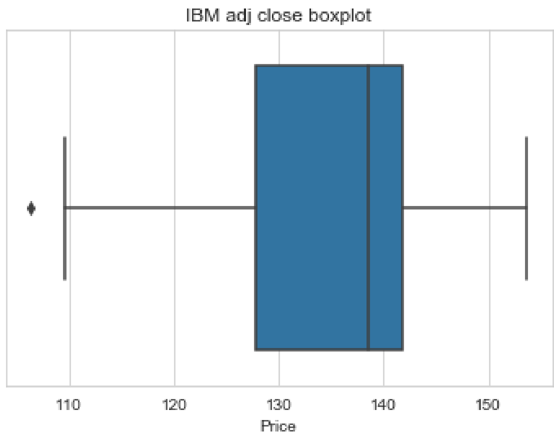
4.1.5 Box plot

A box plot can be used to describe the statistical information in a set of data.

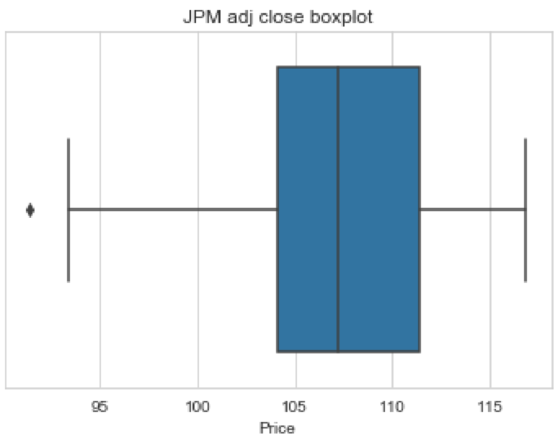
- The line splitting the box is the location of the median ($Q2$)
- The skewness of the data can be seen from the location of the median relative to the edge of the box 25^{th} percentile ($Q1$) and 75^{th} percentile ($Q3$)
- The length of the box ($Q3 - Q1$) represents the interquartile range (IQR) covering 50% of data
- The location of the minimum ($Q1 - 1.5 \times IQR$) and maximum ($Q3 + 1.5 \times IQR$) values with outliers (shown as navy diamonds).



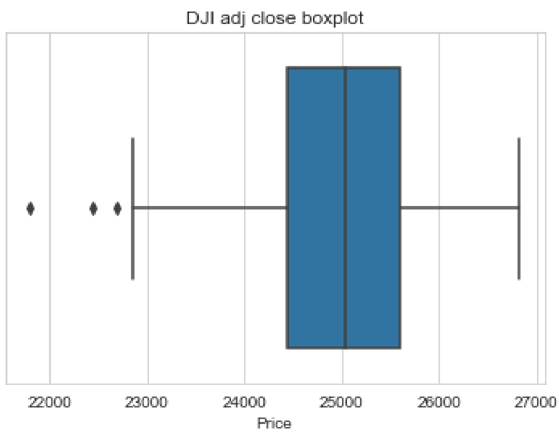
(a) AAPL



(b) IBM



(c) JPM



(d) DJI

Figure 42: Box plot for adj. close for each stock

The above box plots show that AAPL price has a slightly positively skewed adj close price with no outliers from 2017 to 2018 with a large IQR suggest the stock price is volatile with a high variance. IBM stock price is highly negatively skewed with one of the daily adj close price being too low (outlier), JPM stock price is slightly positively skewed with one of the daily adj close price being too low like IBM's stock price, both of the stocks have roughly the same volatility as their IQR are similar. At last, DJI's stock adj close price has a normal distribution shape as the median is roughly at the centre of Q1 and Q3 with several outliers, since its IQR is smaller, this suggests its stock price is less volatile with a lower variance.

However, the box plot does not provide the information on the data size as most of the descriptions are based on percentile and the shape of the data distribution as a histogram or PDF function can.

4.2 Robust Estimators

The following Python code shows the functions used to calculate the median, interquartile range (IQR), and the median absolute deviations (MAD).

```python
import numpy as np

# Robust location estimator: median
def location_estimator(x):
    return np.median(x)

# Robust scale estimator: IQR(interquartile range)
def iqr(x):
    q3, q1 = np.percentile(x, [75 ,25])
    return q3 - q1

# Robust scale estimator: MAD (median absolute deviation)
def mad(x):
    return np.median(np.abs(x-np.median(x)))
```

Robust location estimator: the median of an unsorted list can be found using different selection algorithms such as by sorting the list and obtain the center value which has a best case of $\mathcal{O}(n)$ with an average and worst case of $\mathcal{O}(n \log n)$, or an algorithm called median-of-medians which divide the list into sublists and finding each sublist's median to obtain the unsorted list's median which has a computational complexity of $\mathcal{O}(n)$ for all cases.

Interquartile range: to obtain the IQR of an unsorted list, the list first has to be sorted. The 75-th percentile and 25-th percentile values are taken out and subtracted. The computational complexity is dominated by the sorting algorithm which $\mathcal{O}(n \log n)$.

Median absolute deviation: the MAD function is considered the most computational expensive function to the rest described here, it requires sorting the unsorted list, n subtractions, n absolute values which is equal to setting all values sign bit to positive, and at last finding the median again. Depending on the algorithm used, the computational complexity could be $\mathcal{O}(4n)$ or $\mathcal{O}(n \log n)$ as n is large.

The breakdown point is a measurement of the robustness; the finite sample breakdown point of an estimator is the percentage of data that can be given arbitrary values (outliers) without making the estimator arbitrarily inaccurate. Hence, the maximum breakdown point is defined as 50%. The breakdown point for the median is $\frac{n-1}{2n} \approx 50\%$ because the median can handle up to 50% of data being one of the outliers before becoming inaccurate.

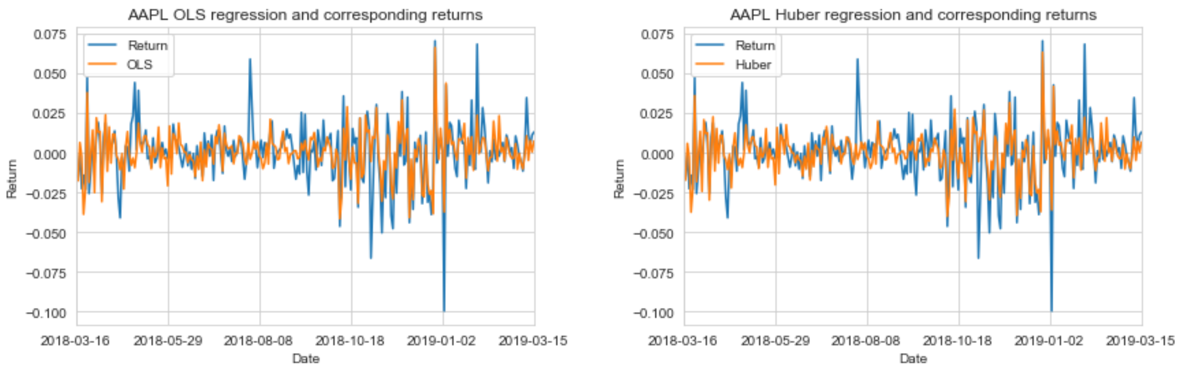
The breakdown point of IQR is 25% because the IQR value is calculated by splitting the data into the largest and smallest set of 25% of data where the maximum allowable outliers in either top or bottom 25% of the data is precisely 25%. The breakdown point of the MAD is limited by the breakdown point of the median which is 50%, as the first median is successfully applied and being an accurate estimator, the second median has no effect on the robustness. Hence, the breakdown point of MAD is 50%.

4.3 Robust and OLS regression

The ordinary least squares (OLS) regression as known as linear regression and Huber regression, a robust linear regression is used to estimate the linear relationship between the stocks: AAPL, IBM, JPM return and the DJI index return. The gradient and intercept are estimated using both linear regression methods, the following figures show the original return with the corresponding OLS and Huber regression model where the OLS regression is trying to minimise the squared-error, and Huber regression is trying to minimise the following equation.

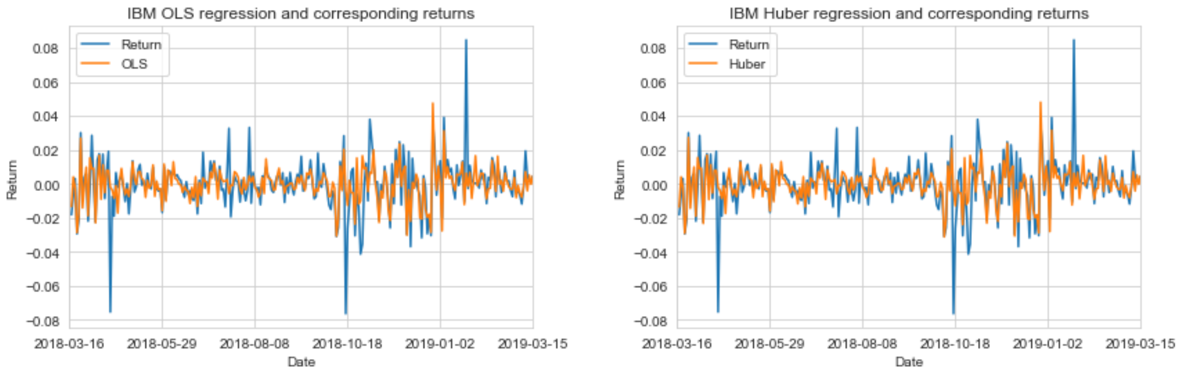
$$\frac{y - xw}{\sigma} < \epsilon \tag{45}$$

where σ and w are the parameters to be optimised, σ is the scale factor to the output y to prevent additional changes to ϵ, w is the input weights coefficient, and ϵ is an input parameter that controls the number of samples that should be classified as outliers.



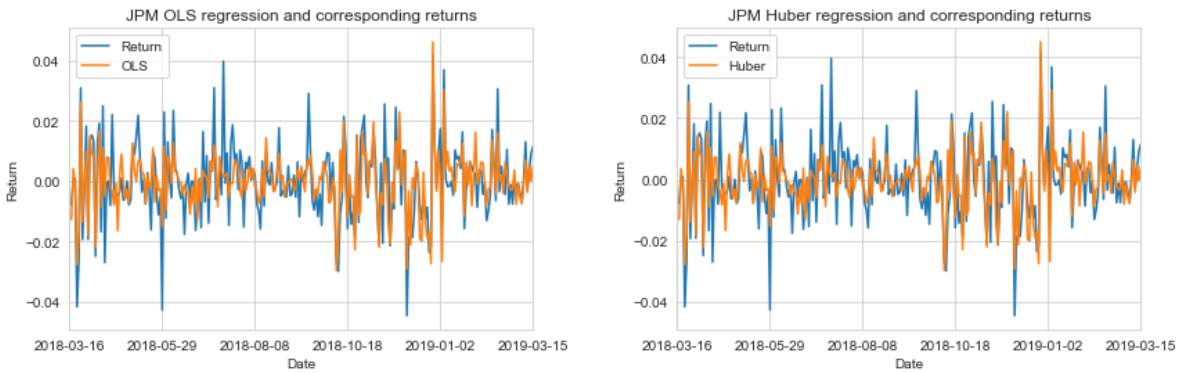
(a) OLS regression (b) Huber regression

Figure 43: AAPL regression of one-day returns against one-day returns of DJI



(a) OLS regression (b) Huber regression

Figure 44: IBM regression of one-day returns against one-day returns of DJI

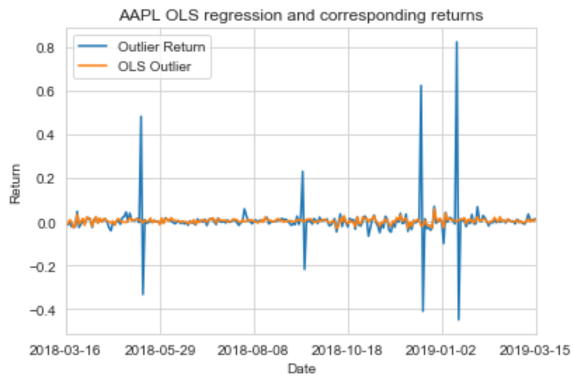


(a) OLS regression (b) Huber regression

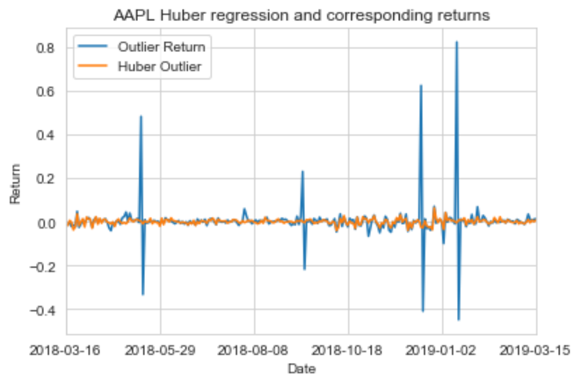
Figure 45: JPM regression of one-day returns against one-day returns of DJI

Table 6: Mean-squared error of fitted regression models

Stocks	Regression mean-squared error (dB)	
	OLS	Huber
AAPL	-37.45209781110427	-37.44175508110574
IBM	-38.525382803015795	-38.524633928819334
JPM	-41.20489288840111	-41.190462136552554

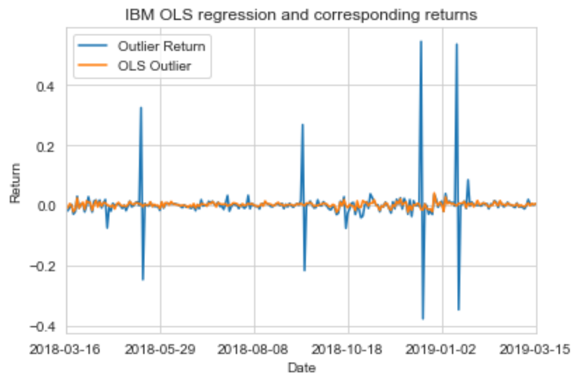


(a) OLS regression

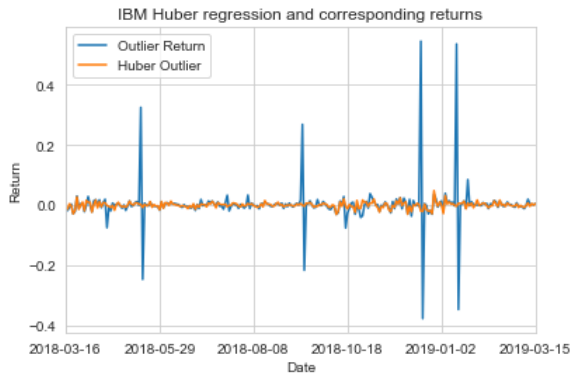


(b) Huber regression

Figure 46: AAPL regression of one-day returns against one-day returns of DJI

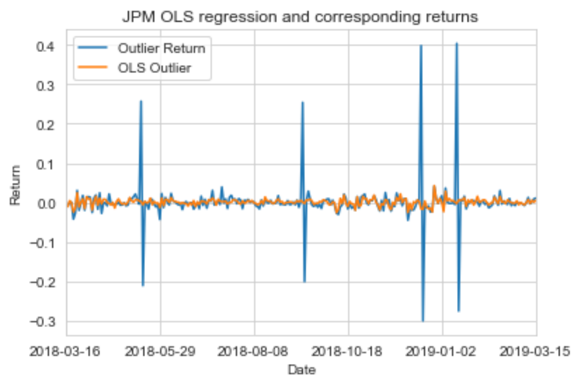


(a) OLS regression

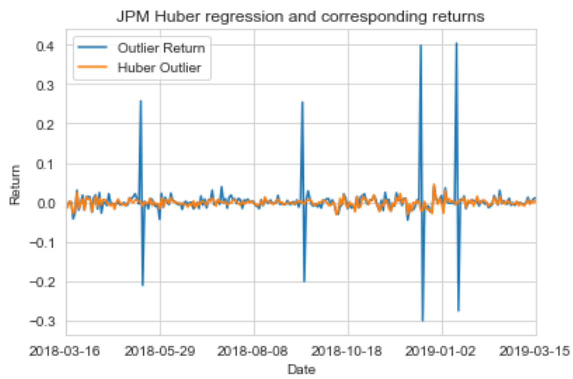


(b) Huber regression

Figure 47: IBM regression of one-day returns against one-day returns of DJI



(a) OLS regression



(b) Huber regression

Figure 48: JPM regression of one-day returns against one-day returns of DJI

Table 7: Mean-squared error of fitted regression models with outliers

Stocks	Regression mean-squared error (dB)	
	OLS	Huber
AAPL	-37.05485718036936	-37.44311114308877
IBM	-38.29486571824614	-38.52461640846538
JPM	-41.05605281180091	-41.17898607206752

Both of OLS and Huber regression models look very similar; the mean-squared error is calculated to represent the error in both models to the true one-day return signal. Table 6 shows that without synthetic outliers introduced, OLS and Huber regression have approximately the same error. As synthetic outliers are introduced, table 7 shows that Huber's model has approximately the same error before the outliers are introduced, not affected by the outliers. However, it can be seen that OLS regression has an increase in error caused by the outliers. Hence, Huber regression is more robust than OLS regression.

4.4 Robust Trading Strategies

The moving average (MA) crossover trading strategy is used on AAPL, IBM, JPM, and DJI with the following rules:

- Buy X shares of a stock when its 20-day MA > 50-day MA
- Sell X shares of a stock when its 20-day MA < 50-day MA

The following figures show the cumulative return, and the corresponding price with 20 and 50 MA plotted, each red dot represent a crossover of the moving averages. X is set as 1 unit throughout this section.

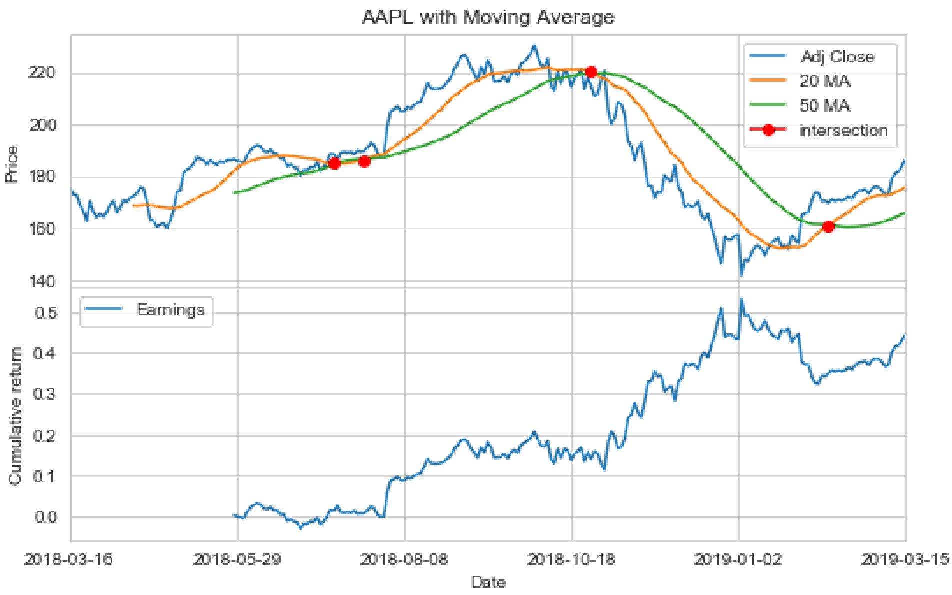


Figure 49: Robust trading strategy on AAPL using moving average

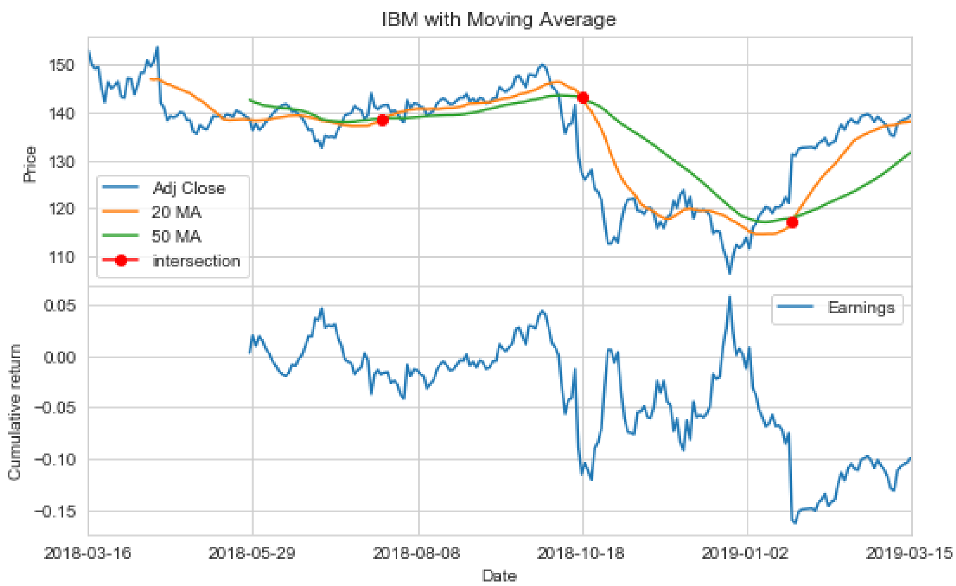


Figure 50: Robust trading strategy on IBM using moving average

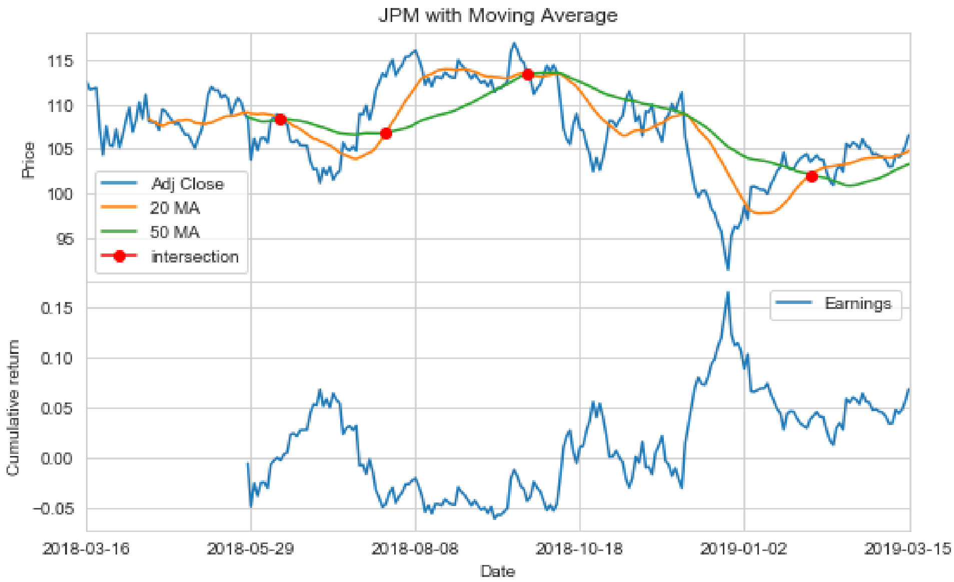


Figure 51: Robust trading strategy on JPM using moving average

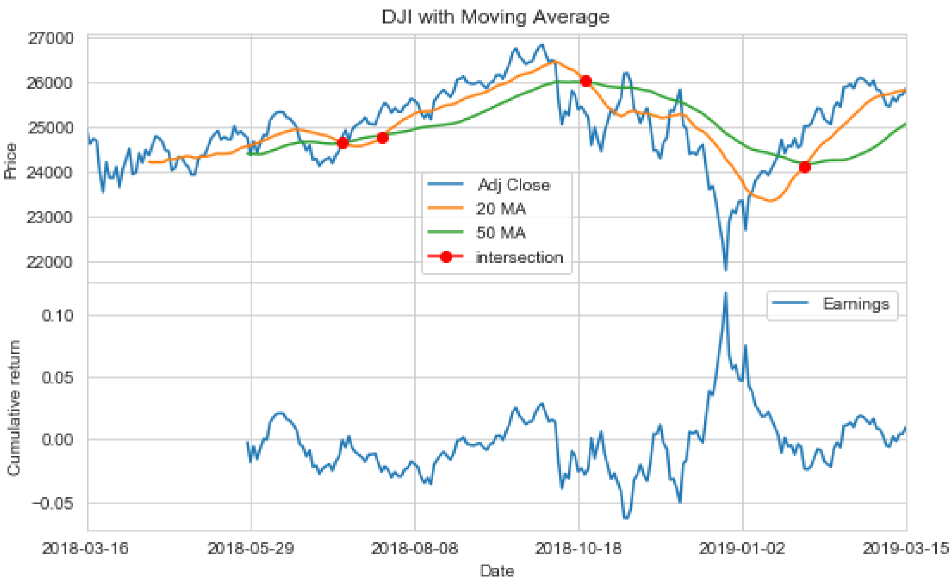


Figure 52: Robust trading strategy on DJI using moving average

It can be seen that the AAPL stock is the only trade with significant positive returns, the IBM stock has sharp price changes causing negative return from the changing of long and short action is too late to be profited by the trends. JPM and DJI stocks are moving sideways, less volatile, resulting in a close to zero return.

Synthetic outliers are introduced in the historical data; the same moving average crossover strategy is applied. The following figures show how the outliers affect the moving averages and cumulative returns.

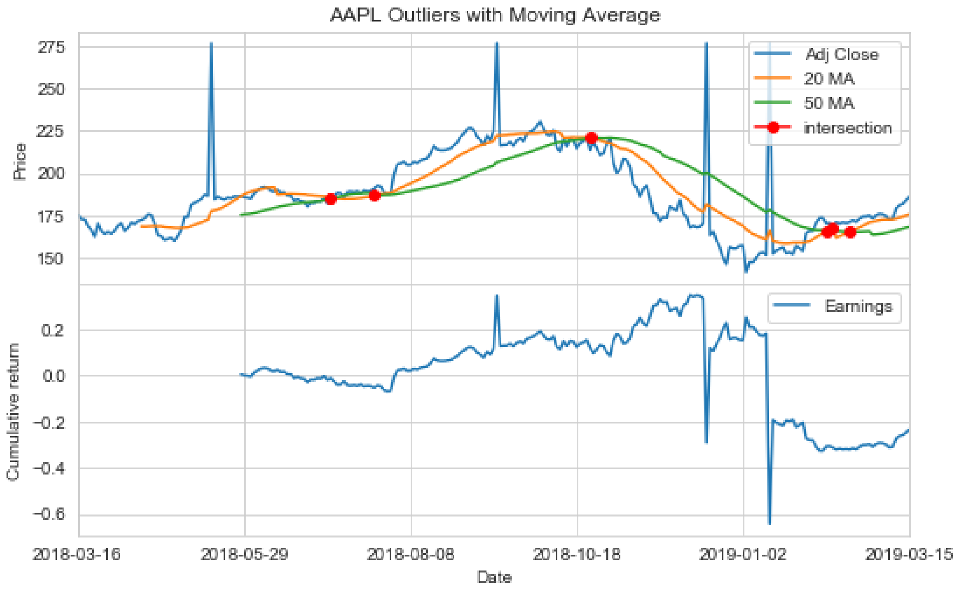


Figure 53: Robust trading strategy on AAPL with outlier using moving average

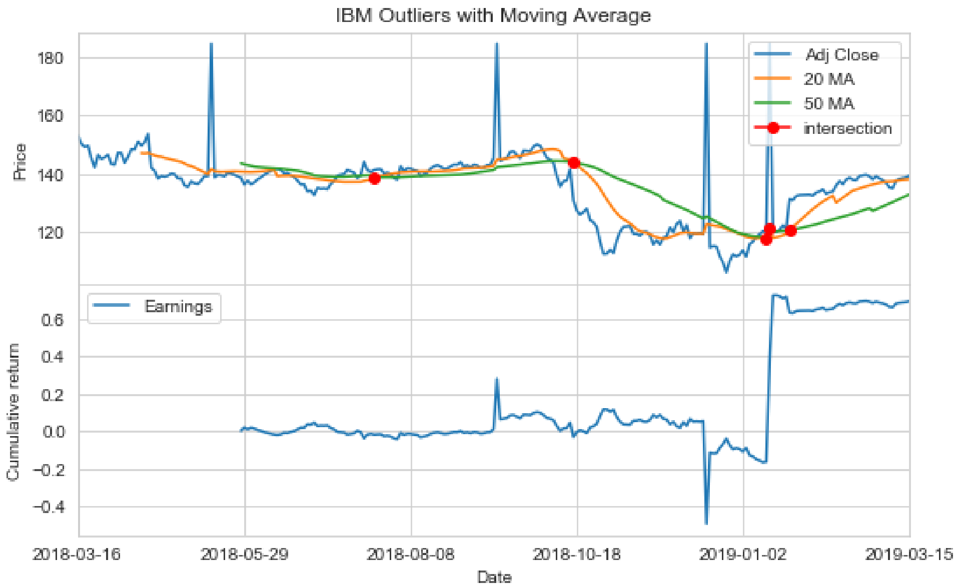


Figure 54: Robust trading strategy on IBM with outlier using moving average

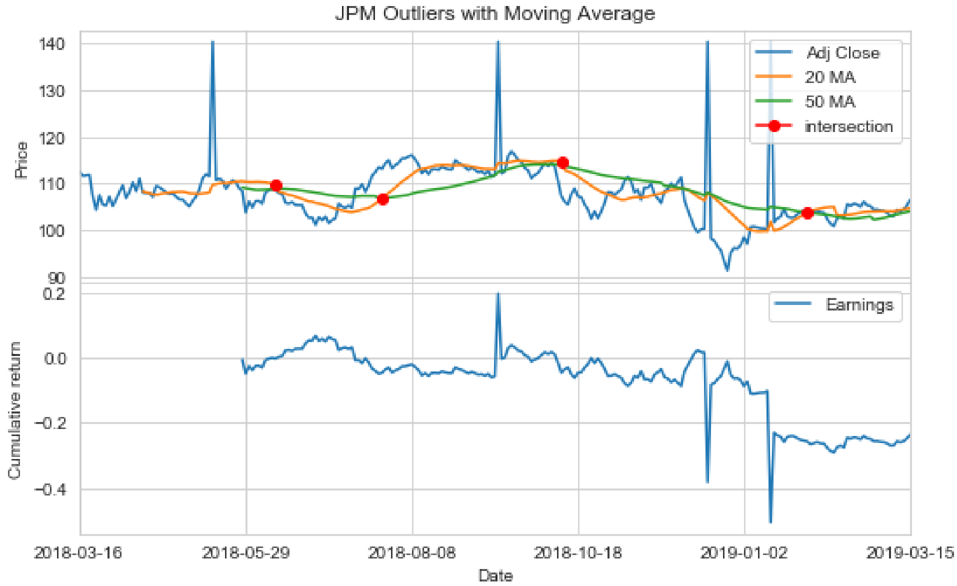


Figure 55: Robust trading strategy on JPM with outlier using moving average

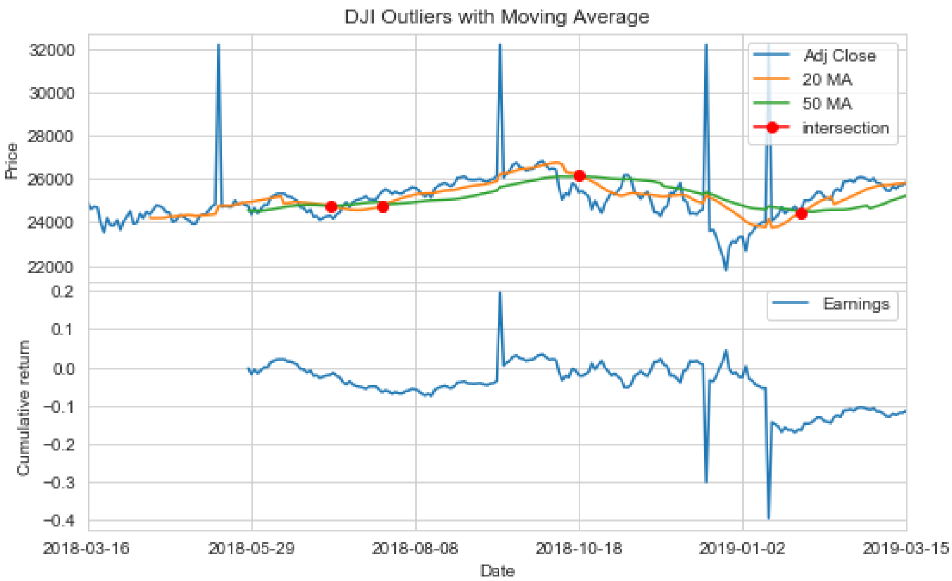


Figure 56: Robust trading strategy on DJI with outlier using moving average

The above figures show the outliers significantly reducing the cumulative return in all stocks and index except for IBM, as the outliers causing extra intersection points, having the trader changes it's original long/short order. The reason IBM has a higher cumulative return is that the additional intersection point is aligned with one of the outliers resulting in a high return in one daily trade.

The following figures show when moving median is used instead of moving average.

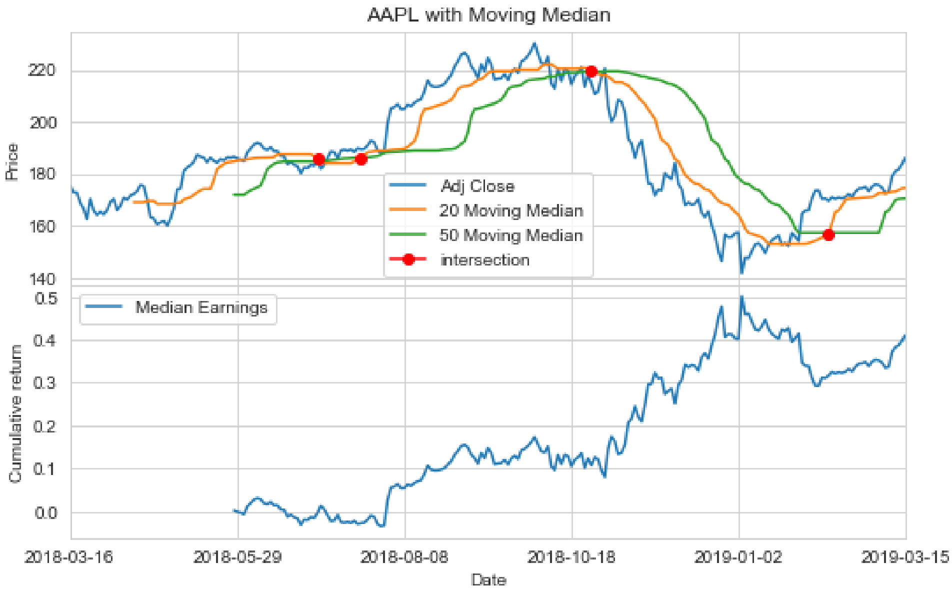


Figure 57: Robust trading strategy on AAPL using moving median

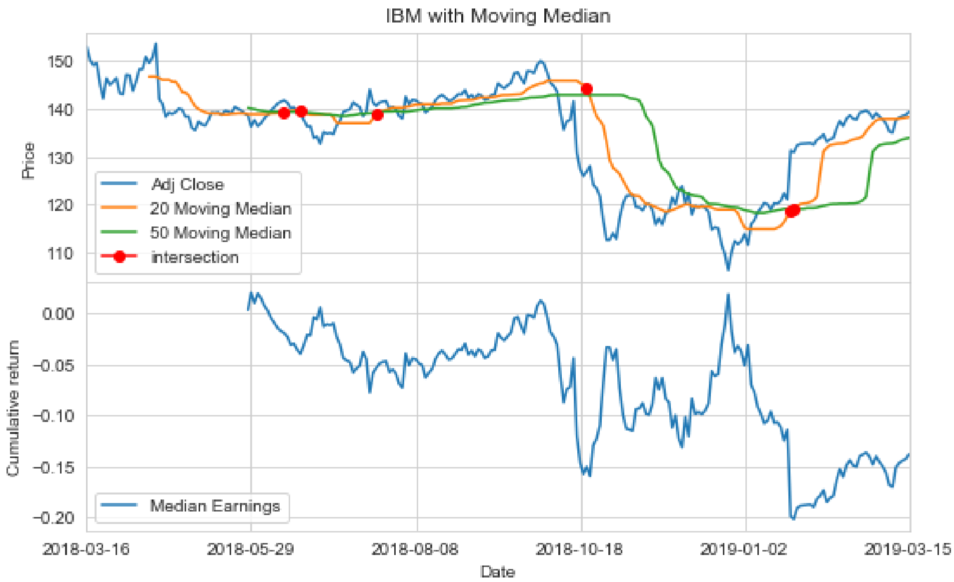


Figure 58: Robust trading strategy on IBM using moving median

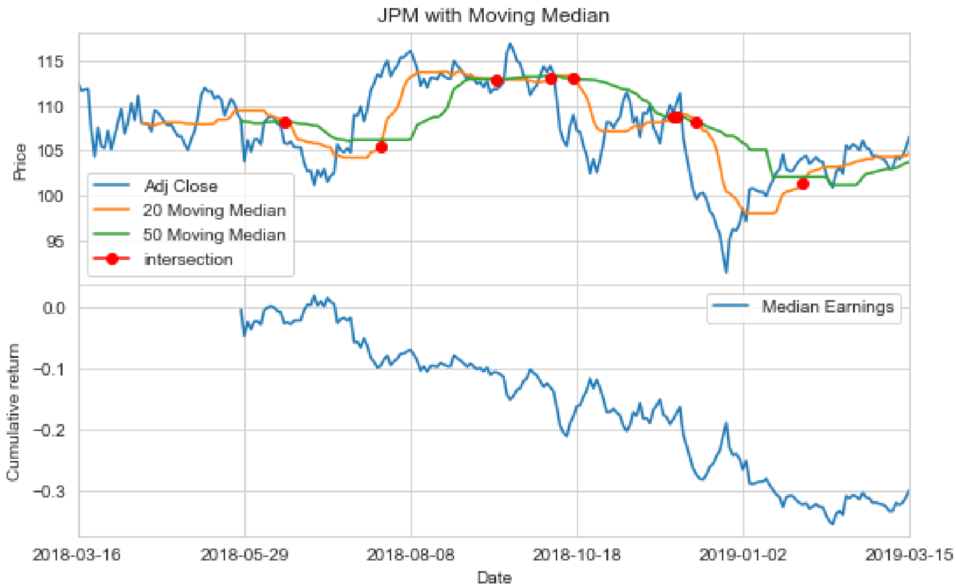


Figure 59: Robust trading strategy on JPM using moving median

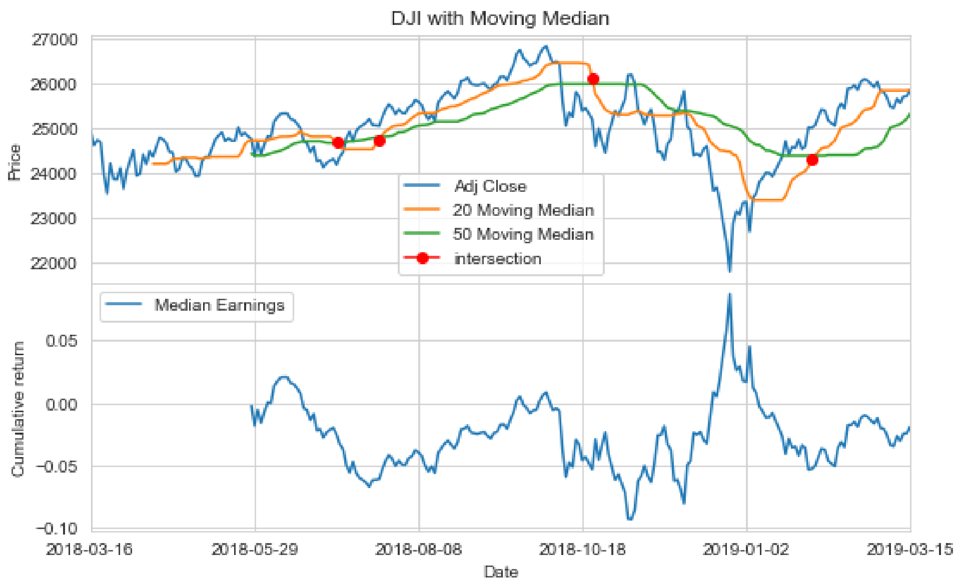


Figure 60: Robust trading strategy on DJI using moving median

It can be seen that the moving median is able to have an earlier crossover as the median estimator is more robust than the mean, hence reducing the response speed, but at the same time resulting in more losses as the response time is not fast enough to earn from the volatility of the stocks.

Synthetic outliers are introduced in the historical data; the moving median crossover strategy is applied. The following figures show how the outliers affect the moving medians and cumulative returns.

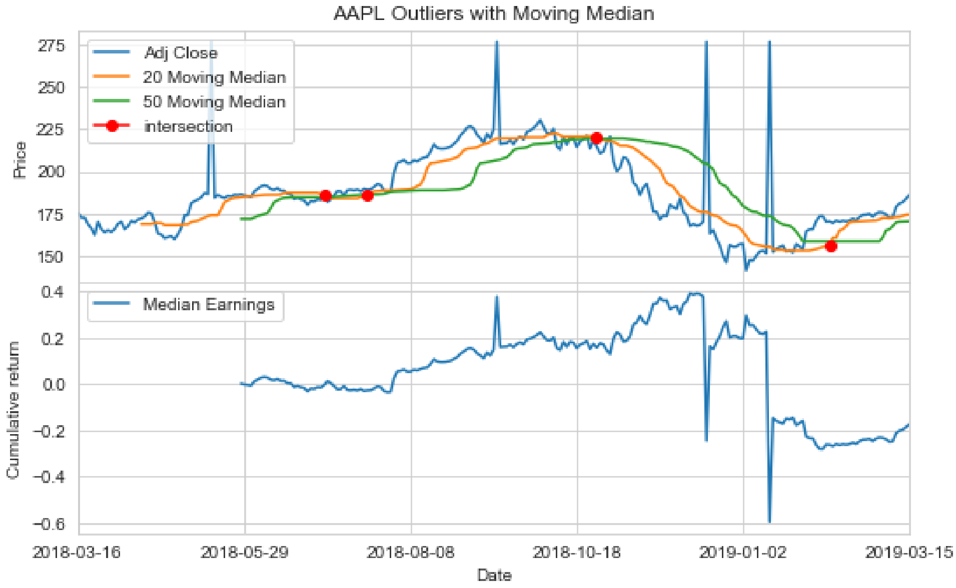


Figure 61: Robust trading strategy on AAPL with outlier using moving median

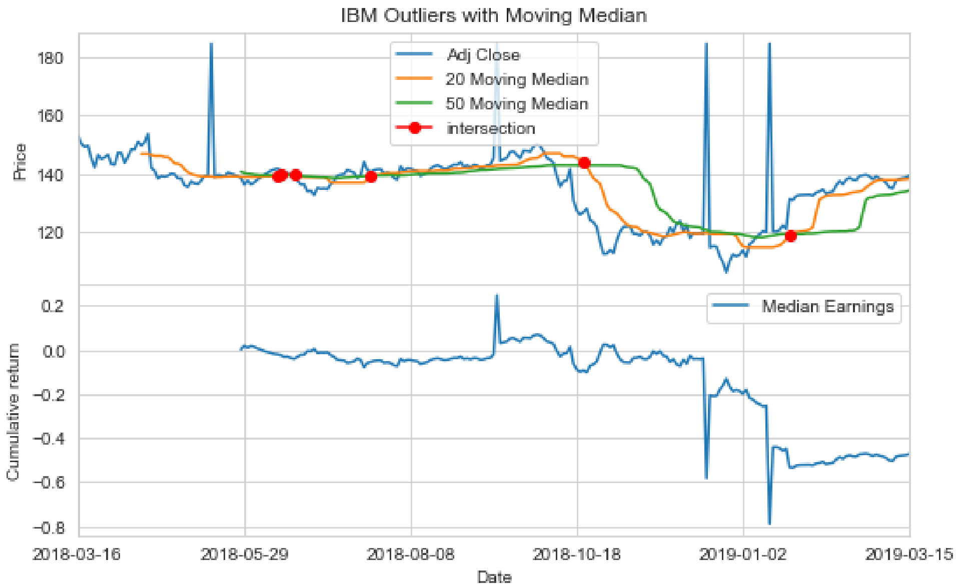


Figure 62: Robust trading strategy on IBM with outlier using moving median

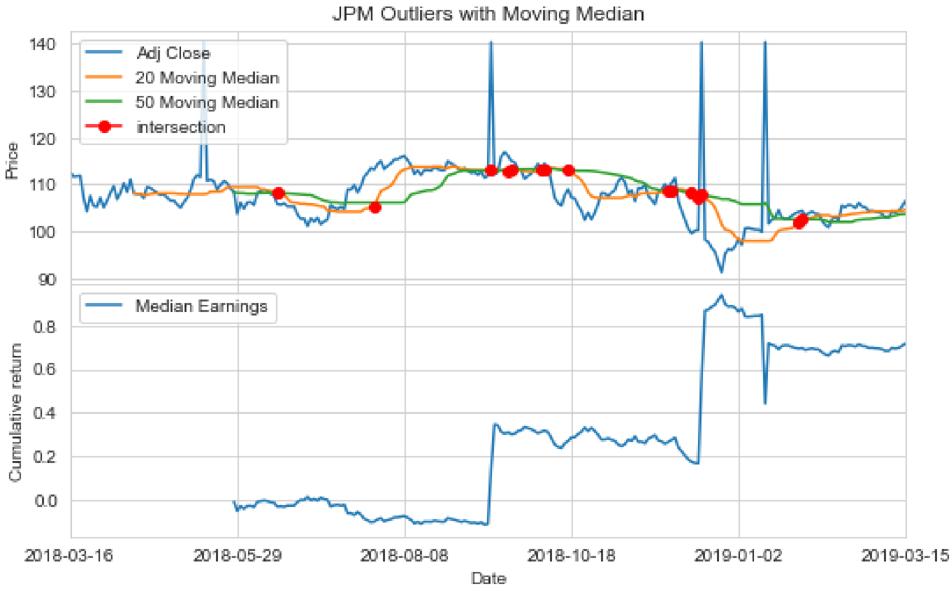


Figure 63: Robust trading strategy on JPM with outlier using moving median

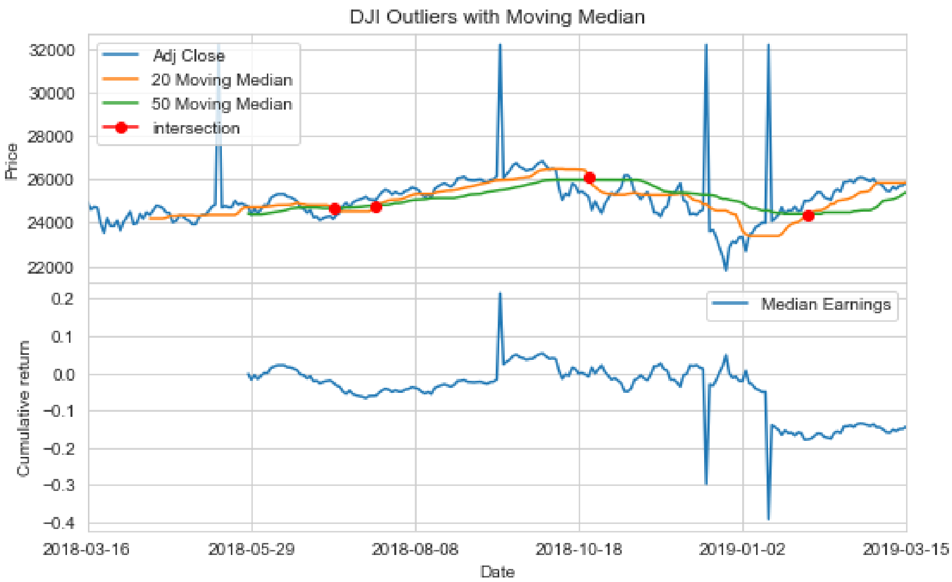


Figure 64: Robust trading strategy on DJI with outlier using moving median

In general, the moving median is more robust than moving average with outliers, resulting in a reduction in losses with outliers. The IBM stock's cumulative return is high only because the change from sell to buy and buy to sell happens to align with two of the outliers which result in two jumps in cumulative return.

5 Graphs in Finance

Partner: Suraj Tirupati

5.1 Graph Theory

A graph is a mathematical structure consisting of numerous nodes, or *vertices*, that contain information regarding different objects. Each vertice may or may not have an effect, *or weightage*, upon another vertice within the graph. The extent to which a node impacts another is described as the *weight* of the *edge*. With the edge being a line drawn between the two vertices that illustrates a relationship the two nodes have with one another. Graphs are used to model some spatial systems, e.g. sensor arrays, crossing of bridges and airport flight times between airports - to name a few illustrative examples. However, they are also used in finance to model complex relationships between different entities and drawing deductions from data such as the formation of financial cliques, risk management, portfolio selection and foreign exchange trading systems to name a few.

This section gives a practical example of **utilising a graph to represent and analyse the effect that Crude Oil Inventory Reports have over the Crude Oil Price**. Further developments are made to the model to incorporate the knock-on, or reflexive, relationship that the changing oil price has on the US and UAE economies. The UAE economy was chosen as its primary export is Crude Oil. A further graph node is introduced to monitor the relationship that the foreign exchange rate between **USD/AED** has on the oil price and vice versa. This analysis can lead insight into how when the price of a foreign export, that is the primary export of a particular nation changes, how its foreign currency exchange rate also subsequently changes and whether its economy responds in a similar manner. It should be noted that crude oil is an obvious example and it should be expected that a reduction in crude oil price will negatively impact the economies of the US and UAE because crude oil is a necessity for all world economies.

This analysis may also be useful in displaying the extent to which countries depend on their natural resources and when the price of those resources drops, how their economies respond. It can serve as an analytical tool for identifying which countries are prone to the *Dutch Disease*: defined as a causal relationship between the development of the economy in one sector (e.g. crude oil/natural resources) and the subsequent degradation in other sections leading to a heavy reliance on the prospering sector with the country's currency appreciates relative to other world currencies and then their primary export becoming too expensive for other nations to purchase. This leads to a decrease in economic wealth for the country in question since other sectors were not developed in parallel. The term was invented by The Economist 20 years after the Netherlands discovered large gas reserves in their country but quickly fell victim to the above definition.

To summarise: the following graph theory example illustrates the effect weekly crude oil production estimates have on the corresponding crude oil price. The effect of the oil price on two nation's economies is represented; with one nation having crude oil as its primary export. The foreign exchange rate of the nation is also represented as a node. A final node is added to represent the total percentage of the GDP that natural resources account for. The analysis can be utilised to track the potential of the nation in question falling victim to the "Dutch Disease". The assumption being made is that the commodity in question is being relied upon as a primary export of the nation in question.

5.2 Impact of Crude Oil Reports on Crude Oil Price

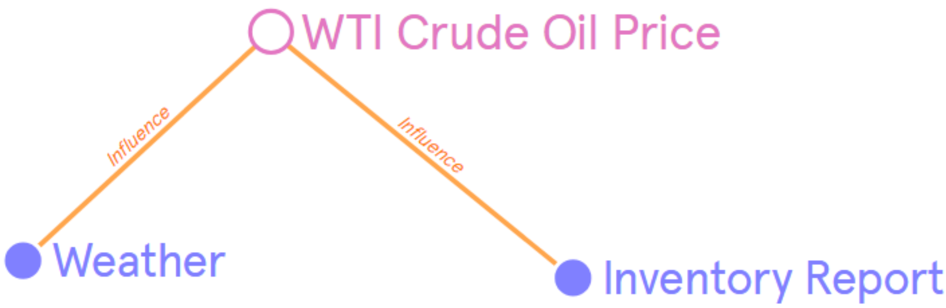


Figure 65: WTI crude oil price influencers graph

As we can see, the weekly crude oil reports from the *Crude Oil inventories* has an effect on the crude oil price. The edge in the graph is represented as a functional influence between the inventory report and the price of the oil. It cannot be stated that a high correlation corresponds to *causation* but it can be ascertained that the effect of the report has some effect on determining the market price for crude oil so creating a function between the two may lead to discovering the relationship one has on the other.

This effect can be studied over time with statistics such as β (volatility) and variance being measured. It may be found that the volatility of the report estimates is greater than the price of the crude oil itself is indicating a lower degree of correlation between the two. Or perhaps the crude oil price will indicate a high volatility just like the errors on the inventory estimation. This may suggest oil is a volatile resource and may be dangerous to invest in for a safety first type of investor.

Other factors such as the weather affect natural resource commodity production too. These factors are accounted for within the graph as well.

5.3 Impact of Crude Oil Price on ADX and S&P500

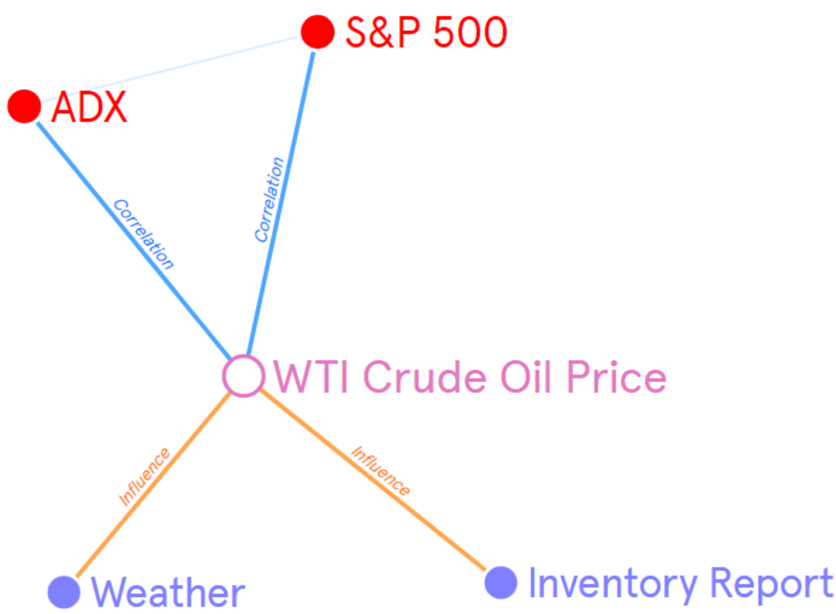


Figure 66: Indices influence on crude oil price graph

The following graph now adds the nodes of the **ADX** (*Abu Dhabi Security Exchange*) index price and the **S&P 500** index price. Correlations between the two stock markets and the oil price are measured. Their relationships are of interest to the purpose of this graph.

It is known that oil is a primary commodity within the world economy. It is required for energy which powers every aspect of the economy. However, higher oil prices imply higher production costs for each production layer of the world's economies. Higher oil prices may indicate a decrease in S&P 500 prices but an increase in the prices of the ADX index in the UAE. A potential reason for this is due may be because increasing oil prices mean more profit for oil manufacturers. However, this may not be the case. Increasing oil prices may decrease global demand causing the UAE's economy to drop. Irrespective, the relationship between oil price and stock market performance is a parameter to be measured and studied from this.

If an economic advisor within the UAE were to have access to this graph model, then they could counteract the increasing oil prices by reducing national interest rates which in turn decrease foreign investment, devaluing their currency. The devaluation of the currency means foreign nations will still be able to afford the conversion into AED to purchase the oil. Conversely, if oil prices are decreasing and this is negatively affecting the state of the *USD/AED* rate, then the UAE government could withhold the production of their oil so that demand increases once more causing an increase in oil price. This highlights the potential for this kind of analysis when in the hands of macroeconomic advisors working closely with world governments.

5.4 Impact of Crude Oil Price and ADX & S& P500 Indices on Forex Rates

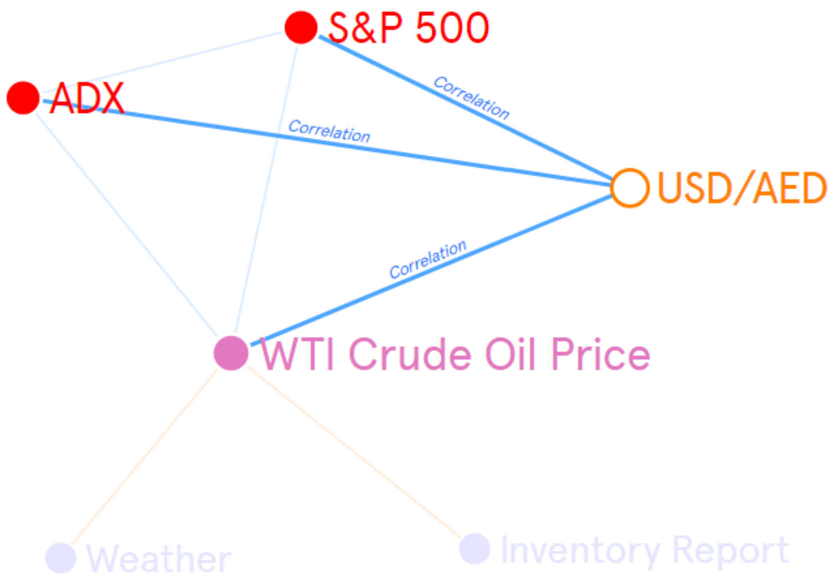


Figure 67: Crude oil back influence on Forex rates graph

The introduction of the foreign exchange rates within the above graph allows for the analysis of the effect of crude oil price changes on the *USD/AED* exchange rate. The correlation between the exchange rate and the oil price and stock markets are taken. Its relationship is of interest here and is attempted to be studied by viewing the correlation each factor has on one another.

Greater oil demand can lead to increased prices and an increase in the GDP of the UAE. This increase in GDP may make the countries currency stronger with respect to the dollar. The effect of a stronger currency on the overall performance of the stock market of the nation can be analysed as well. Questions such as *"Does the UAE's economy historically perform better when oil prices are up?"* can be asked and answered through the graph's technical analysis. This information can not only be utilised by governmental economic advisors in whether they should balance the strength of the currency through inflation with changing oil prices but also by private hedge fund managers who can hedge their bets on rising or falling oil prices by maintaining a short position the United Arab Emirate's Dirham.

Understanding the relationship between oil prices and the global economy is also something hedge fund managers can use effectively to manage their portfolios for times of economic hardship.

5.5 Potential of Dutch Disease

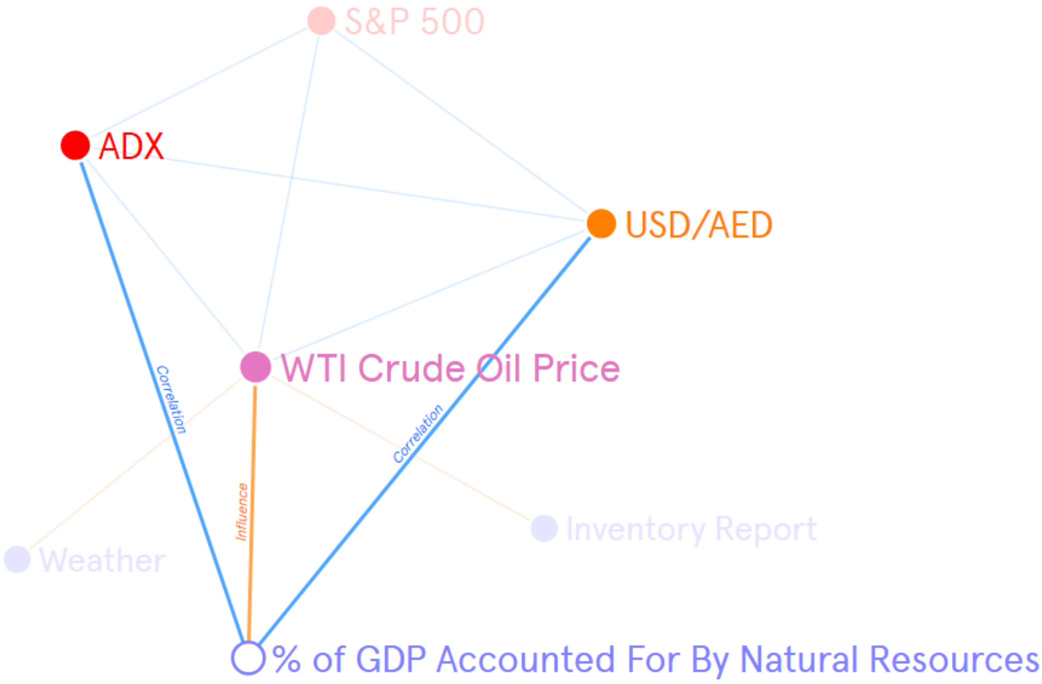


Figure 68: Graph of impact from % of GDP accounted for by natural resources

The above graph indicates a further node additionally placed within the graph. This node is the percentages that the natural resource sector of the UAE's economy accounts for within its GDP. The % is influenced by the price of oil; high oil prices mean the % will be increased while low oil prices mean the % will be decreased.

This is a careful parameter to be watched over by economic advisors advising a nation. The purpose of this is because rising oil prices due to demand may cause an increase in the value of the *Dirham*. If this increase in value is not balanced out through inflation, then the threat to the UAE's economy is that foreign nation's are no longer able to afford the expense of buying their oil. If the country has not been developing other areas of its economy, then the ultimate threat is that the nation is no longer able to consistently export the same value of goods causing its currency to become devalued with the economy going into recession and the participants suffering due to economic hardship. This is known as the Dutch Disease. Countries such the UAE are prone to it if they depend heavily on one sector (the most common sector being natural resources).

The graph is particularly useful in analysing this as it gives correlations and volatility fluctuations for a UAE economic advisor based on the price of oil, it's the effect on the rest of the economy and foreign exchange rates and the percentage of GDP that the natural resource (oil) sector is accounting for. It should be noted that some of these effects are not simply one way - for example the performance of the ADX may have a highly correlated effect on the strength of the USD/AED conversion; this, in turn, reduces the correlation of increasing demand in oil (measured by crude oil price) is having on the performance of the ADX. The graph is clearly highly linked with its individual factors all playing separate, but related, roles within their macroeconomic environment.

5.6 Advance Commodity/Forex Graph

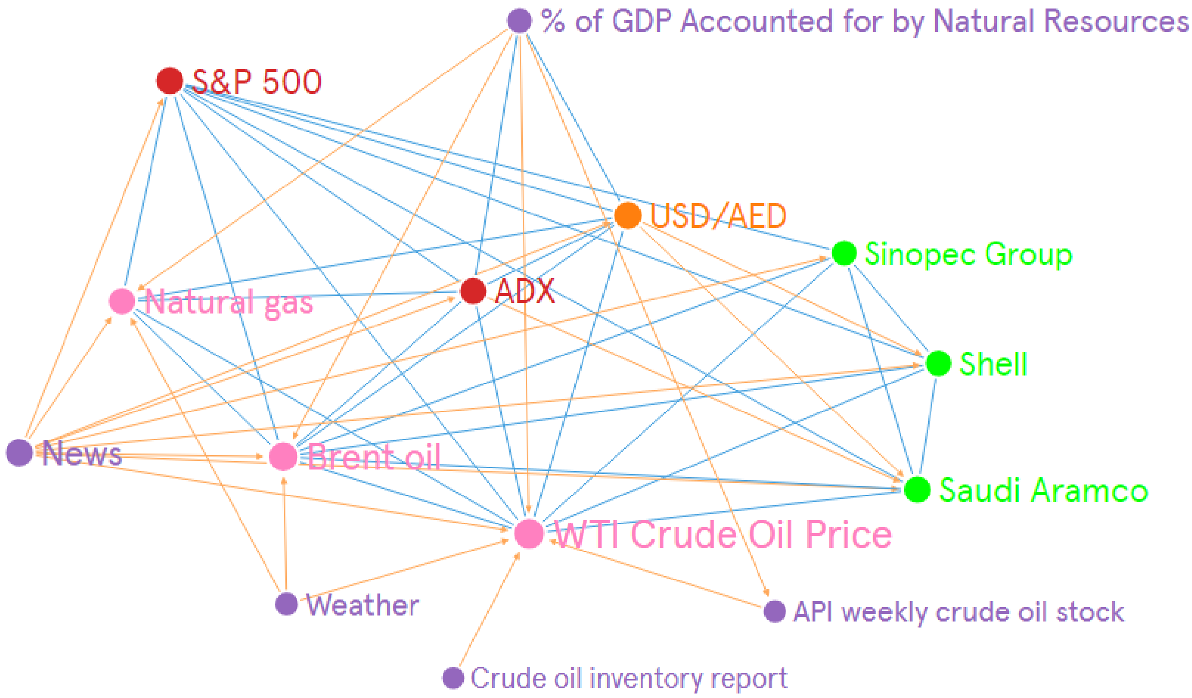


Figure 69: Graph centred on WTI crude oil

This section contains a scale-up of the commodity graph that takes into account other natural resources, some of the largest oil company stock prices and other external factors such as the news and crude oil stock. It is a mockup of what an actual graph that performs the functions described in this section would look like.

A further step can be taken into account by analysing macroeconomic factors such as commodity prices and foreign exchange markets. These can be looked at by scaling up the graph in this section to include other commodities such as other natural resources (e.g. Brent oil, natural gas etc.), *the price of average food items (FAO - Food Price Index), base metals, natural resources and raw materials*. Then the graph can be updated to include forex rates of at least 8 major world economies: *EUR, USD, JPN, YEN, RUP, GBP, CAD and RUB.*

Exchange rates between the major players can be represented on the graph, and their relationships with changing commodity prices can be analysed. The greater the effect of the commodity price on a foreign exchange rate, the larger the vertice between the two. The vertice will once again represent the correlation between the two. This will not give causal relationships between commodity prices and forex rates, but it will give a forex investor a detailed insight into how fluctuating commodity prices can impact major world economies. If one then takes their study into understanding what factors affect the price of certain commodities, then the historical change of their prices can be factored into future investments and hedged bets across major world economies.

5.7 Summary

A few final comments. The Dutch Disease is a potentially catastrophic outcome for a nation's economy if allowed to progress. Government advisors and world governments who have a susceptibility to relying too heavily on their natural resource, or any other sector, need to put in place policies and schemes to prevent its outcome from damaging their nation's prosperity and their civilian's wealth. Utilising a graph that provides a strong visual representation of the relationships oil price, currency exchange rates, % of GDP due to natural resources and the national and global stock market could be very useful and insightful for leaders to make intelligent decisions.

As Warren Buffet wrote to the board of Berkshire Hathaway in 2004 "Be greedy when others are fearful and be fearful when others are greedy" - the context is very different here, but government officials should be wary when one of their sectors begins to outshine and out-prosper the others. In times like this, it is important to maintain a level head and take up balancing acts to ensure long term sustenance and sustainability of a nation's economy. The temptation to overcommit and overly rely on a particular sector is easy to fall into. Having a visual representation makes it all the easier for officials to spot trends and act on them accordingly.

The UAE was given as an example in this section due to the large exports of crude oil from Abu Dhabi. However, Sheikh Mohammed bin Rashid Al Maktoum of Dubai had a strong response to the potentiality of Dubai is falling doom to the Dutch Disease during the 2000s when he heavily invested in the tourism industry of Dubai turning it into a mystical land for nationals across to world to visit, or even live. Even to this day, Dubai is reducing the bank waivers international tourism companies require to provide so that it is easier to set up companies there. These are great examples of shrewdness from a leader who foresaw the potential of falling victim to the Dutch Disease. Other leaders and governments, particularly those who come into a fortune, may not be as diligent in their approach. For them, a graph of this nature and a team to help with development and analysis could provide critical in long term sustenance.