Imperial College London

Signal Processing and Machine Learning in Finance

Yee Hong Low
YHL116
01202613

Contents

1 Regression Methods

1.1 Processing stock price in Python

1. The natural-log of the time-series is show in Figure 1.

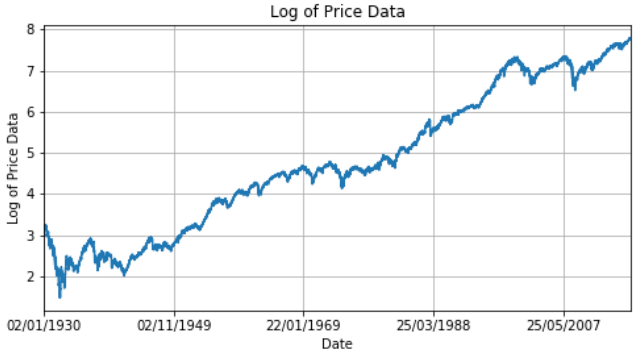

Figure 1: Log of Price Data.

2. The rolling mean and rolling standard deviation of both the price data and log-price data can be seen in Figure 2.

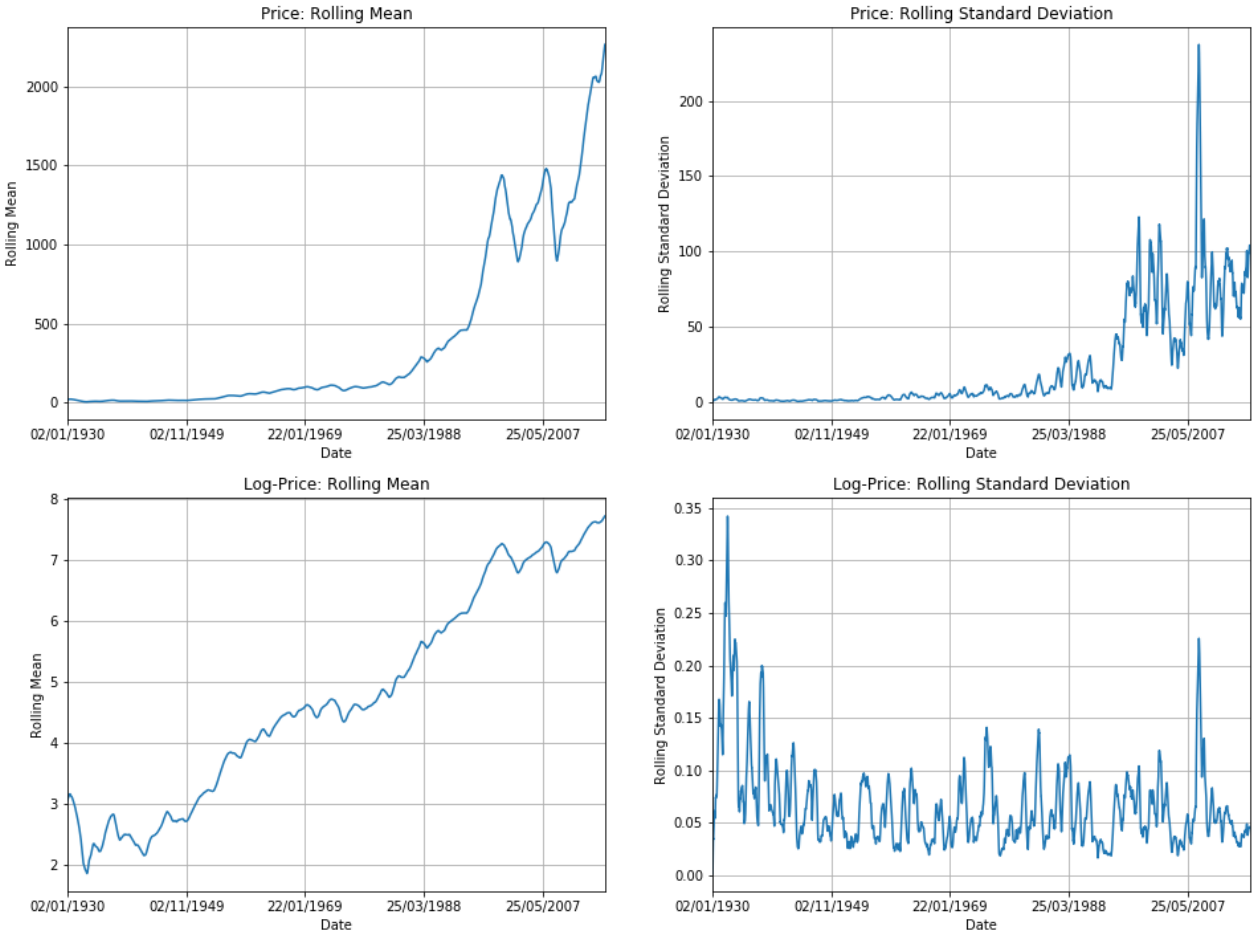

Figure 2: Rolling mean and rolling standard deviation of log of price data.

A stationary process is one where the mean and variance stays constant. For both the price and log-price, it is seen that the mean has an upward trend while the standard deviation stays relitively constant. It is hence inferred that both the price and log-price data are non-stationary. To further prove this, the Augmented Dickey-Fuller test is performed on the log price and price. Both the price and log-price give p-values close to 1, which proves that both the data are not stationary.

3. The rolling mean and rolling standard deviation, with 252 days window, of both the log return and simple return can be seen in Figure 3.

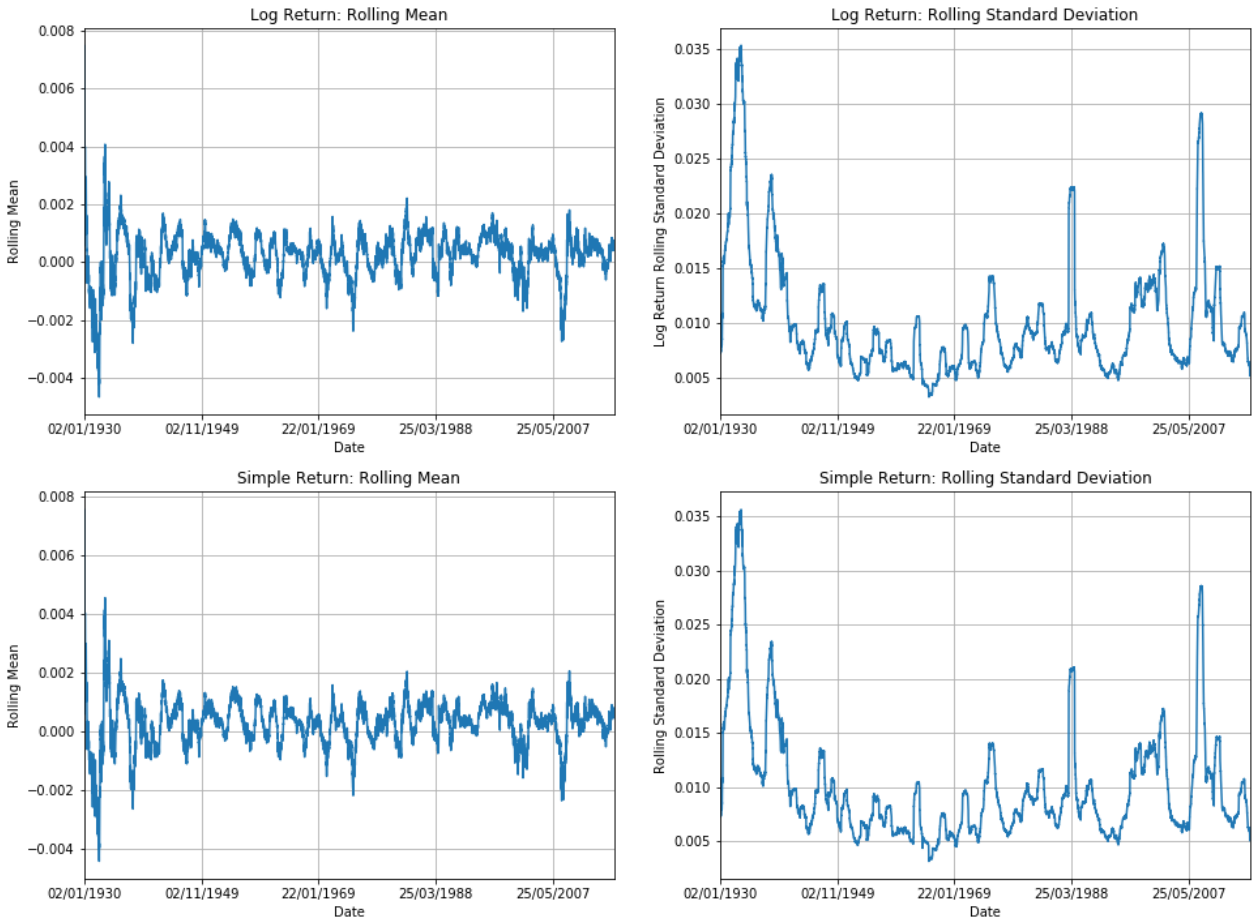

Figure 3: Rolling mean and rolling standard deviation of log return and simple return.

Both the log return and simple return are stationary as their mean and standard deviation stay relatively constant.

4. Log-return has a a few benefits over simple return which includes the following:

- Log-returns are often distributed log normally, then $log(1+r_i)$ is normally distributed because: $1+r_i = \frac{p_i}{p_j} = e^{log(\frac{p_i}{p_j})}$. This is beneficial as many classic statistics presumes normality.
- When returns are very small, log-return can be approximated as the actual return i.e. $log(1+r) \approx r$ when $r \ll 1$.
- Log-return is time-additive. The compounding return is given by:

$$(1+r_1)(1+r_2)\dots(1+r_3) = \prod_i (1+r_i)$$

which is undesirable as the product of normally distributed variables is not normal. The sum of normally distributed variables however is normally distributed (when all variables are uncorrelated) and is desirable. This can be obtained by applying the log function on the above equation, giving:

$$\sum_i log(1+r_i) = log(1+r_1) + log(1+r_2) + \dots + log(1+r_n)$$

With this, the compound returns can be added across periods, which is easier to compute and maintain the normality of the return. This allows for easier mathematical operations for signal processing and analysis.

- From the time-additivity of log-return, we note that it is numerically safe to do addition or subtraction of small numbers but multiplying small numbers is not.

The Jarque–Bera test is a goodness-of-fit test of whether sample data have the skewness and kurtosis matching a normal distribution. The null hypothesis of the test is that the input random variable is normally distributed. For price data, the Jarque-Bara test gives a p-value of zero. So we reject the null hypothesis and conclude that price data is not normally distributed.

Log returns assumes data to be independent and identically distributed (i.i.d) normal, which in this case is not true as the data is skewed. The data is skewed, which is expected as log-normality over long time-scale for financial price is unrealistic. However, log returns still have the benefits of time-additivity over simple returns.

5. The log returns are 0.69314 and -0.69314, whereas the simple returns are 1.0 and -0.5. This illustrates the desirable symmetric property of log-returns. When aggregating returns, the log-return can simply be calculated by summing up all the returns which in this case gives 0.69314 - 0.69314 = 0, which corresponds to 0 net return. There is however no simple way of aggregating simple returns across time.

6. We should use simple return when aggregating across assets in a portfolio. The simple return of a portfolio is the weighted sum of the simple returns of the constituents of the portfolio. Besides that, the assumption of log-normality over long time-scales is unrealistic. A positive skew is assumed in log-normal distributions, but most financial data is negatively skewed in long time-scales, owing to financial crashes.

1.2 ARMA vs. ARIMA Models for Financial Applications

1. The rolling mean and rolling standard deviation of SNP and log SNP are plotted in Figure 4.

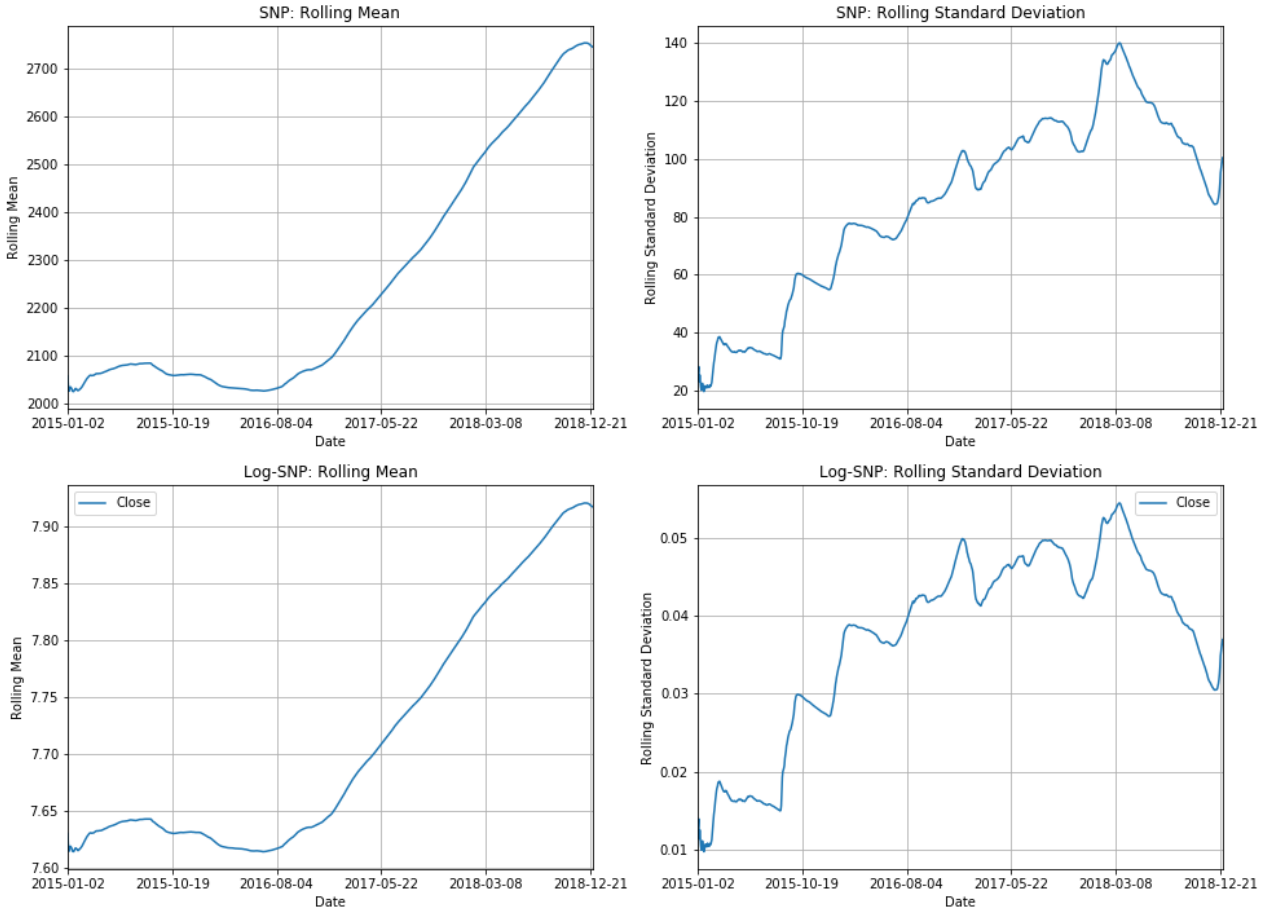

Figure 4: Rolling mean and rolling standard deviation of SNP and log SNP.

It is inferred that both SNP and log-SNP are non-stationary as it is obvious from the plot that both the time-series exhibit an upward trend in their mean and standard deviation. This is further confirmed with the Augmented Dickey-Fuller test in which both the time-series give a p-value of 0.667598. It is hence concluded that both SNP and log-SNP are non-stationary. An ARIMA model is hence more suitable to model the data.

2. The ARMA(1,0) model and the `True` data is plotted in Figure 5.

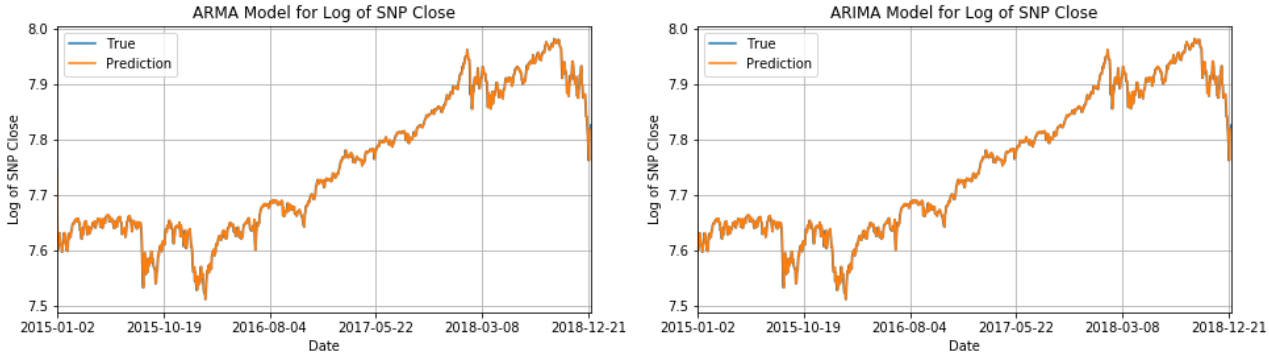

Figure 5: ARMA and ARIMA model of log-SNP 500.

The ARMA(1,0) model is essentially an AR model with degree 1 in which it can be written as

$$y_t = a_1 y_{t-1} + \eta$$

which for the ARMA(1,0) model of SP 500, $a_1 = 0.997359$ and $\eta = 7.739997$. These values can be heuristically deduced from the plot of y_t against y_{t-1} in Figure 6. Since $a_1 = 0.997359 \approx 1$, the model is basically taking the last value to predict the next value which might not be the most effective prediction.

However, these findings are not useful at gauging the performance of the model. It is given that price data is often non-stationary in which the ARMA(1,0) model will not give good prediction. A more useful approach is to plot an autocorrelation plot of the data. If there is a positive correlation between y_t and y_{t-1} then an AR model would be appropriate. Conversely, if there is negative correlation between y_t and y_{t-1} then an MA model would be more appropriate. From Figure 6, it is seen that y_t and y_{t-1} of SP data exhibits strong positive correlation which explains why the ARMA(1,0) model (which is technically an AR model) fits the data well with a low mean squared error of 8.627e-05 despite being non-stationary. This is further supported by the plot of y_t vs y_{t-1} in Figure 6 which shows that y_t and y_{t-1} exhibits a linear relationship i.e. $y_t \approx m y_{t-1} + c$.

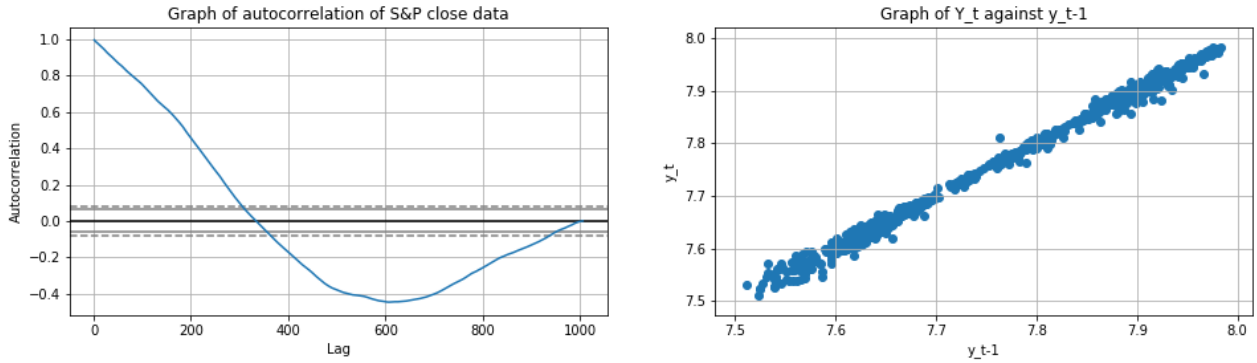

Figure 6: Correlation between y_t vs y_{t-1} and y_t vs y_{t-1}.

3. The ARMA(1,1,0) model and the `True` data is plotted in Figure 5. For the ARIMA(1,1,0) can be represented with the mathematical equation:

$$y_t - y_{t-1} = a_1(y_{t-1} - y_{t-2}) + \eta$$
$$y_t = y_{t-1} + a_1(y_{t-1} - y_{t-2}) + \eta$$

which in the case of SP data, $a_1 = -0.008752$ and $\eta = 0.000196$.

The ARIMA(1,1,0) model is more reliable and practical than the ARMA(1,0) model as it uses differencing to remove the stationary in the data. In the SP dataset, the ARIMA(1,1,0) model gives better prediction with a mean squared error of 7.4283e-5 as compared to the ARMA(1,0) model which gave a mean squared error of 8.6272e-5.

4. By taking the log of the prices, we essentially change the return to log return:

$$y_t - y_{t-1} = a_1(y_{t-1} - y_{t-2}) + \eta$$

Taking log of prices instead gives us:

$$log(y_t) - log(y_{t-1}) = a_1(log(y_{t-1}) - log(y_{t-2})) + \eta$$
$$log\_return_t = a_1 log\_return_{t-1} + \eta$$

Where y is price data. This introduces benefits which was previously discussed in 1.1.4. Taking the log of prices also makes the data more stationary which is desired as the model requires the input data to be stationary.

1.3 Vector Autoregressive (VAR) Models

1. Given

$$\mathbf{B} = \begin{bmatrix} \mathbf{c} & \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \end{bmatrix}$$

and

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

In matrix form, $\mathbf{y}_t$ is given by:

$$\mathbf{y}_t = \begin{bmatrix} \mathbf{c} & \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{y}_{t-1} \\ \mathbf{y}_{t-2} \\ \cdots \\ \mathbf{y}_{t-p} \end{bmatrix} + \mathbf{e}_t$$

and $\mathbf{y}_t$ can be deduced as:

$$\mathbf{y}_{t-1} = \begin{bmatrix} \mathbf{c} & \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \end{bmatrix} \begin{bmatrix} 1 \\ \mathbf{y}_{t-2} \\ \mathbf{y}_{t-3} \\ \cdots \\ \mathbf{y}_{(t-1)-p} \end{bmatrix} + \mathbf{e}_{t-1}$$

where $\mathbf{y}_t, \mathbf{y}_{t-1} \in \mathbb{R}^{K\times 1}$. If we define $\mathbf{Y}$ as:

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_t & \mathbf{y}_{t-1} & \cdots & \mathbf{y}_{t-(T-1)} \end{bmatrix}$$

then

$$\mathbf{BZ} + \mathbf{U} = \begin{bmatrix} \mathbf{c} & \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \end{bmatrix} \begin{bmatrix} 1 & 1 & \dots & 1 \\ \mathbf{y}_{t-1} & \mathbf{y}_{t-2} & \cdots & \mathbf{y}_{(t-1)-(T-1)} \\ \mathbf{y}_{t-2} & \mathbf{y}_{t-3} & \cdots & \mathbf{y}_{(t-2)-(T-1)} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{t-p} & \mathbf{y}_{(t-1)-p} & \cdots & \mathbf{y}_{(t-p)-T} \end{bmatrix} + \begin{bmatrix} \mathbf{e}_t & \mathbf{e}_{t-1} & \dots & \mathbf{e}_{(t-p)-(T-1)} \end{bmatrix}$$

Hence, Equations (2)-(3) can be represented in a concise matrix form as

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

where $\mathbf{Y} \in$

$$\mathbf{Y} = \begin{bmatrix} \mathbf{y}_t & \mathbf{y}_{t-1} & \cdots & \mathbf{y}_{t-(T-1)} \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} \mathbf{c} & \mathbf{A}_1 & \mathbf{A}_2 & \dots & \mathbf{A}_p \end{bmatrix}$$

$$\mathbf{Z} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ \mathbf{y}_{t-1} & \mathbf{y}_{t-2} & \cdots & \mathbf{y}_{t-T} \\ \mathbf{y}_{t-2} & \mathbf{y}_{t-3} & \cdots & \mathbf{y}_{(t-1)-T} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{y}_{t-p} & \mathbf{y}_{(t-1)-p} & \cdots & \mathbf{y}_{(t-p)-T} \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} \mathbf{e}_t & \mathbf{e}_{t-1} & \dots & \mathbf{e}_{t-(T-1)} \end{bmatrix}$$

and $\mathbf{Y} \in \mathbb{R}^{K\times 1}, \mathbf{B} \in \mathbb{R}^{K\times(KP+1)}, \mathbf{Z} \in \mathbb{R}^{(KP+1)\times T}$ and $\mathbf{U} \in \mathbb{R}^{K\times T}$.

2. To find $\mathbf{B}_{opt}$, we would like to minimise the error term $\mathbf{U}$ i.e. we would like to approximate $\mathbf{Y} \approx \mathbf{B}_{opt}\mathbf{Z}$. We do this by first assuming $\mathbf{U} = \mathbf{0}$. Then,

$$\mathbf{Y} = \mathbf{B}\mathbf{Z}$$

To find $\mathbf{B}_{opt}$, we use the least squares projection matrix:

$$\begin{aligned} \mathbf{Y}\mathbf{Z}^T &= \mathbf{B}_{opt}\mathbf{Z}\mathbf{Z}^T \\ \mathbf{B}_{opt} &= \mathbf{Y}\mathbf{Z}^T(\mathbf{Z}\mathbf{Z}^T)^(- 1) \end{aligned}$$

3. The VAR(1) process is given as:

$$\begin{aligned} \mathbf{y}_t &= \mathbf{A}\mathbf{y}_{t-1} + \mathbf{e}_t \\ &= \mathbf{A}(\mathbf{A}\mathbf{y}_{\mathbf{t-2}} + \mathbf{e}_{t-1}) + \mathbf{e}_t \\ &= \mathbf{A}^2\mathbf{A}\mathbf{y}_{\mathbf{t-2}} + \mathbf{A}\mathbf{e}_{t-1} + \mathbf{e}_t \\ &= \mathbf{A}^N\mathbf{A}\mathbf{y}_{\mathbf{t-N}} + \sum_{i=0}^{N-1} \mathbf{A}^i\mathbf{e}_{t-i} \end{aligned}$$

For stability, we need to ensure that $\mathbf{y}_t$ is bounded, in which we need to ensure $||\mathbf{A}^N|| < 1$. The singular value decomposition of $\mathbf{A}^N$ is given by the following:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\mathbf{\Sigma}\mathbf{V}^* \\ \mathbf{A}^N &= (\mathbf{U}\mathbf{\Sigma}\mathbf{V}^*)^N \\ &= \mathbf{U}\mathbf{\Sigma}^N\mathbf{V}^* \end{aligned}$$

Hence to ensure $||\mathbf{A}^N|| < 1$, we need to ensure $||\mathbf{\Sigma}^N|| < 1$ i.e. all the eigenvalues of $\mathbf{A}$ needs to be less than 1.

4.

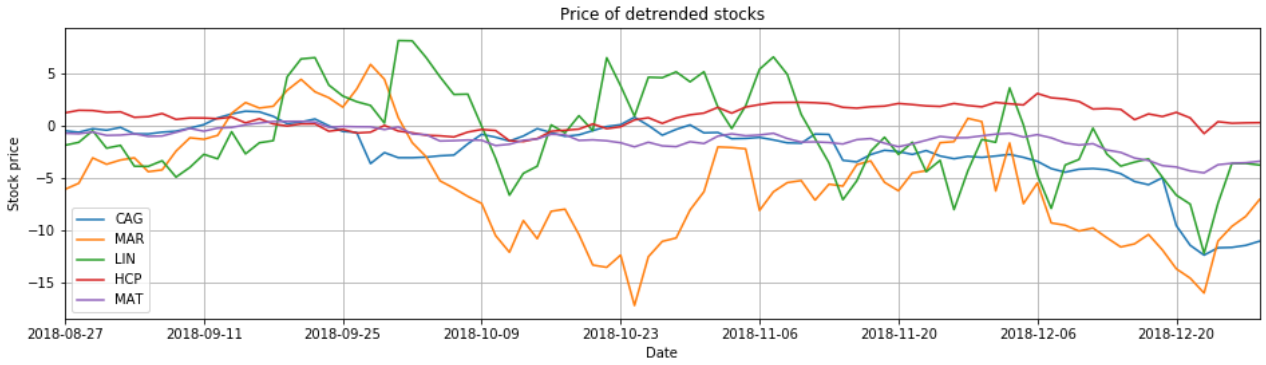

Figure 7: Price of Detrended Stocks.

It is not sensible to construct a portfolio using these stocks. They are chosen randomly and do not exhibit correlation between one another. This is evident from the plot in Figure 7. Stocks from different sector have different factors affecting its dynamics, making it difficult to predict them using VAR.

Besides that, it is also seen that one of the eigenvalues of the $\mathbf{A}$ matrix have a larger absolute value of 1.006, which is greater than 1. This will cause the prediction to become unstable as time progresses.

5. It is generally recommended to build a portfolio by grouping stocks by sectors. This is because they usually are more related and have similar dynamics hence leading to a more accurate VAR model prediction. This is further proven from the eigenvalues of the $\mathbf{A}$ matrix where all sector but 1 have a largest eigenvalue

smaller than 1. This would ensure that the prediction made using the VAR model will not go unstable over time.

Sector	**Largest Eigenvalue**
Industrials	0.991721
Health Care	0.994153
Information Technology	0.992738
Communication Services	0.982263
Consumer Discretionary	0.990650
Utilities	0.985648
Financials	1.004340
Materials	0.991744
Real Estate	0.982785
Consumer Staples	0.991508
Energy	0.985577

Table 1: Largest eigenvalue of A matrix of each sector.

2 Bond Pricing

2.1 Examples of bond pricing

1a. Annual compounding:

$$1,100 = 100(1+r)$$
$$r = 0.1/annum$$

Percentage return per annum, r = 10%

1b. Semiannual compounding:

$$1,100 = 1,000(1+\frac{r}{2})^2$$
$$r = 0.976$$

Percentage return per annum, r = 9.76%

1c. Monthly compounding:

$$1,100 = 1,000(1+\frac{r}{12})^{12}$$
$$r = 0.09569$$

Percentage return per annum, r = 9.57%

1d. Continuous compounding:

$$1,100 = 1,000e^r$$
$$r = 0.09531$$

Percentage return per annum, r = 9.53%

2. Let r = interest per annum

$$(1+\frac{0.15}{12})^{12} = e^r$$
$$r = 0.14907$$

Percentage interest per annum = 14.91% when continuous compounding is 15% per annum with monthly compounding.

3.

$$10,000e^{\frac{0.12}{4}} = 10,304.54$$
$$10,000e^{\frac{0.12}{2}} = 10,618.37$$
$$10,000e^{\frac{0.12\times 3}{4}} = 10941.74$$
$$10,000e^{0.12} = 11,274.97$$

The interest paid per quarter is hence given by:

$$\text{1st Quarter} = 10,304.54 - 10,000.00 = 304.54$$
$$\text{2nd Quarter} = 10,618.37 - 10,304.54 = 313.83$$
$$\text{3rd Quarter} = 10,941.74 - 10,618.37 = 323.37$$
$$\text{4th Quarter} = 11,274.97 - 10,941.74 = 333.23$$

2.2 Forward rates

a. I can be happy with the extra 9% but I might lose out on the possibility that there will be an interest rate higher than 9% in the 2nd year.

b. The 5% and 7% strategies are straight-forward. It can be calculated and the returns are guaranteed. The 9% forward rate strategy usually involves a contract where the 9% interest is guaranteed for the 2nd year regardless of the market condition. This is so that we can minimize the risk taken and maximize our return.

c. As discussed in (b), the forward rate of 9% usually involves a contract so that we can hedge the risk i.e. we are guaranteed an interest rate of 9%. However, the disadvantages is that we will potentially lose out on a higher interest rate in the 2nd year.

d. For me to go from a one year investment to a 2 year investment, it will depends on 2 conditions:

1. I am confident that the interest rate for the second year will be lower than 9% in which if I were for the 2 year investment, I will lose out on some potential income.
2. I am confident that I will not urgently need the amount in which I have invested in the next two years.

2.3 Duration of coupon-bearing bond

a. Duration = 0.0124 + 0.0236 + 0.0337 + 0.0428 + 0.0510 + 0.0583 + 6.5377 = 6.7595

b.

$$Modified\ duration = \frac{duration}{1 + yield} = \frac{6.7595}{1 + 0.05} = 6.4376$$

The modified duration is an adjusted version of the Macaulay duration, which accounts for changing yield to maturities. It is observed that the modified duration gives a slightly smaller value than the duration (by 6.7595 - 6.4376 = 0.3219). This means that the duration of the bond will decrease by 0.3219 years if the yield to maturity increases by 1% (from 1% to 2%).

c. Both bonds and pension liabilities are inversely proportional to interest rates. An increase in interest rates will decrease the liability while a decrease in interest rates will increase the liability. Hence similarly to bonds, the increase and decrease of the pension liabilities can also be estimated using the duration of the liabilities. Hence the duration is a helpful tool which can protect investors from unexpected changes in interest rates.

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

1. The mean market returns per day is plotted in Figure 8:

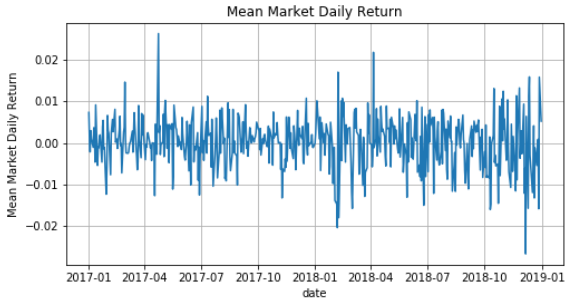

Figure 8: Mean market return per day, R_{m_t}.

2. The rolling beta, $\beta_{i,t}$, for five companies i, with rolling window of 22 days are plotted in Figure 9: The volatility of $\beta_{i,t}$ can be interpreted as follows:

- $\beta = 1$: as volatile as the market
- $\beta > 1$: more volatile than the market

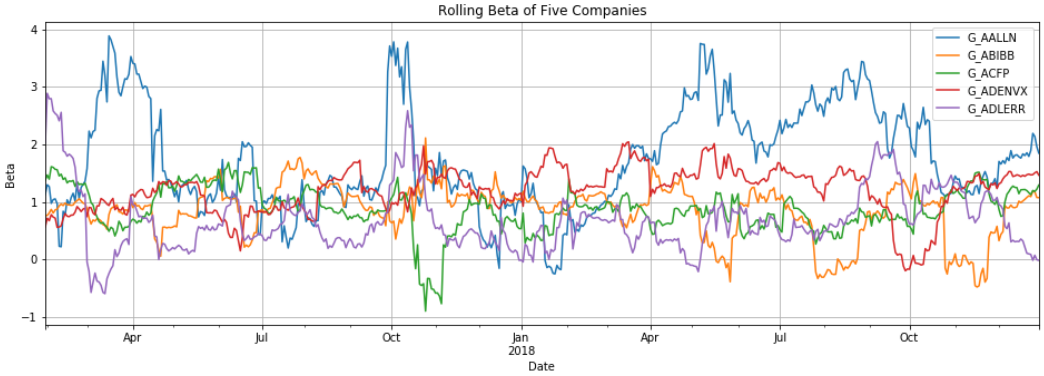

Figure 9: The rolling beta, $\beta_{i,t}$, for five companies i, with rolling window of 22 days.

- $\beta < 1$: less volatile than the market
- $\beta = 0$: uncorrelated to the market
- $\beta < 0$: negatively correlated to the market

Each of the beta values for each companies are grouped according to the above bins. The results are shown in Table 2:

Companies	**More Volatile**	**Less Volatile**	**Uncorrelated**	**Negatively Correlated**
G_AALLN	410	90	0	10
G_ABIBB	211	289	0	44
G_ACFP	173	327	0	13
G_ADENVX	358	142	0	10
G_ADLERR	101	399	0	2
⋮	⋮	⋮	⋮	⋮

Table 2: Binning of five out of 157 companies to show the properties of the beta values of each company.

Table 2 can be interpreted as follows: take the company G_AALLN for example. Across the observation period (of 500 samples), it is on average more volatile than the market (410 out of 500 times). It is never uncorrelated to the market and is rarely (10 out of 500 times) negatively correlated to the market. The same analysis can be performed any of the 157 companies. The mean and standard deviation of each of the bins are then calculated:

	More Volatile	**Less Volatile**	**Uncorrelated**	**Negatively Correlated**
Mean	242.496815	257.503185	24.808917	26.146497
Standard Deviation	129.163861	129.163861	93.889060	30.844485

Table 3: Statistics of $\beta_{i,t}$.

From Table 3, it can be deduced that across the observation period, on average the companies are 242 out of 500 times more volatile than the market. About 4.96% ($\frac{24.8}{500} \times 100\%$) of the time, the companies are uncorrelated to the market whereas about 5.22% ($\frac{26.15}{500} \times 100\%$) of the time they are negatively correlated to the market.

3. The cap-weighted market return, R_m, is plotted against the mean market return, R_{m_t}, in Figure 10:
The cap-weighted market return, R_m is a type of market index with individual components that are weighted according to their total market capitalization. The weighting coefficient $\sum_i mcap_i$ represents the total market capitalisation. The components with a higher market cap carry a higher weighting percentage in the index. Hence, it is a good measure of a companies influence on the market.

4. The beta values are again calculated by using the cap-weighted market return. All of the beta values,

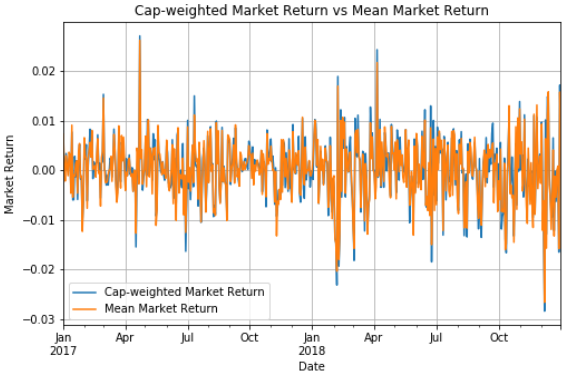

Figure 10: Cap-weighted market return vs mean market return.

$\beta_{m_{i,t}}$ are again grouped into bins per in 2.4.2 and its statistics analysed. The following gives the mean and variance of $\beta_{m_{i,t}}$:

	More Volatile	**Less Volatile**	**Uncorrelated**	**Negatively Correlated**
Mean	224.993631	275.006369	24.808917	29.070064
Standard Deviation	120.699537	120.699537	93.889060	35.914449

Table 4: Statistics for $\beta_{m_{i,t}}$.

Again, we can analyse these results per 2.4.2. Across the observation period, on average the companies are 224 out of 500 times more volatile than the cap-weighted market. About 4.96% ($\frac{24.8}{500} \times 100\%$) of the time, the companies are uncorrelated to the cap-weighted market whereas about 5.81% ($\frac{29.07}{500} \times 100\%$) of the time they are negatively correlated to the cap-weighted market.

With cap-weighted returns, the companies with higher market cap carry a higher weighting percentage in the index. Hence, investors can use cap-weighted beta to find potential returns above the market average. However, the overweighting towards larger companies give a distorted view of the market.

Over time, companies can grow to the extent that they make up an inordinate amount of the weighting in an index. As a company grows, index designers are obligated to appoint a greater percentage of the company to the index, which can endanger a diversified index by placing too much weight on one individual stock's performance.

Also, index funds or exchange-traded funds buy additional shares of a stock as its market capitalization increases or as the share price increases. In other words, as the stock price is rising, the funds are purchasing more shares at the higher prices, which can be counterintuitive to the investing mantra of buying low and selling high.

If a company's stock is overvalued from a fundamental standpoint, the purchasing of the stock as its market-cap and price increases can create a bubble in the stock's price. As a result, purchasing stocks based on market-cap weightings can lead to a stock market bubble and increase the risk of the bubble bursting sending stock prices into free fall.

5a. The plot for a, R_m and R_s are in Figure 11:

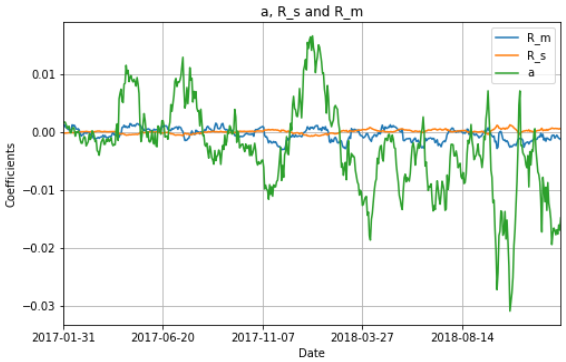

Figure 11: a, R_m and R_s.

5b. The magnitude and variance of a, R_m and R_s are given below:

	R_m	R_s	a
Mean	7.946136e-03	0.024550	0.183809
Variance	1.078719e-07	0.000001	0.000060

Table 5: Statistics for $\beta_{m_{i,t}}$.

Here, a represents the risk-free rate, R_s and R_m are the risk premium factor and b_m and b_i are the level of volatility of the corresponding asset price with respect to R_s and R_m respectively.

We observe that R_s has the lowest magnitude among the three parameters, it has the least influence on return, r_i. In contrast, the magnitude of a is the biggest and hence has the biggest influence on the return.

5c. The correlation through time for five companies are given below:

Company	**Covariance with $\varepsilon_\mathbf{i}$**
G_AALLN	0.922100
G_ABIBB	0.927952
G_ACFP	0.930458
G_ADENVX	0.893420
G_ADLERR	0.963189
⋮	⋮

Table 6: Covariance of each company with $\varepsilon_\mathbf{i}$ which gives the specific return of each company.

5d. The covariance matrix is given as:

	R_m	R_s
R_m	6.350191e-07	-1.271291e-07
R_s	-1.271291e-07	7.465433e-08

Table 7: Covariance matrix between R_m and R_s.

The magnitude (Forbenius norm) of the covariance matrix is 6.641883618076104e-07. It can also be deduced that since each of the element in the covariance matrix is smaller than 1, it has a magnitude smaller than 1. Since its magnitude is less than 1, it is concluded that the covariance matrix is stable.

5e. PCA is a statistical procedure that uses an orthogonal transformation to decompose a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The eigenvector captures the information of the specific return, whereas the eigenvalue represents the weight of the information.

The first principal component explains the highest amount of variance (12.07%) to best represent the specific returns. In this case, we can say the first principal component represents 12.07% of the specific returns.

3 Portfolio Optimization

3.1 Adaptive minimum-variance portfolio optimization

1. Given that:

$$J'(\mathbf{w}, \lambda, \mathbf{C}) = \frac{1}{2}\mathbf{w}^T\mathbf{C}\mathbf{w} - \lambda(\mathbf{w}^T\mathbf{1} - 1)$$

Differentiating both sides with respective to $\mathbf{w}$ and setting $\frac{\partial J'}{\partial \mathbf{w}} = 0$:

$$\begin{aligned} \frac{\partial J'}{\partial \mathbf{w}} &= \mathbf{C}\mathbf{w} - \lambda\mathbf{1} = 0 \\ \mathbf{C}\mathbf{w} &= \lambda\mathbf{1} \end{aligned} \tag{1}$$

and also differentiating both sides with respect to λ and again setting $\frac{\partial J'}{\partial \lambda} = 0$:

$$\begin{aligned} \frac{\partial J'}{\partial \lambda} &= -\mathbf{w}^T\mathbf{1} - 1 = 0 \\ \mathbf{w}^T\mathbf{1} &= 1 \end{aligned} \tag{2}$$

Given these two constraints constraints, recall the definition of variance, $\bar{\sigma}^2$:

$$\begin{aligned} \bar{\sigma}^2 &= \mathbf{w}^T\mathbf{C}\mathbf{w} \\ &= \mathbf{w}^T\lambda\mathbf{1} \\ &= \lambda\mathbf{w}^T\mathbf{1} \\ &= \lambda 1 \\ &= \lambda \end{aligned}$$

Hence, the theoretical variance of returns if we were to apply the minimum variance estimator is given by $\bar{\sigma}^2 = \lambda$.

2. To find the optimal weights, $\mathbf{w}_{opt}$, we have to solve the two equations from 3.1.1:
From (1):

$$\begin{aligned} \mathbf{C}\mathbf{w}_{opt} &= \lambda\mathbf{1} \\ \mathbf{C}\mathbf{w}_{opt} - \lambda\mathbf{1} &= \mathbf{0} \\ \begin{bmatrix}\mathbf{C} & \mathbf{-1}\end{bmatrix}\begin{bmatrix}\mathbf{w}_{opt} & \lambda\end{bmatrix}^T &= \mathbf{0} \end{aligned} \tag{3}$$

Combining (3) with (2), we get

$$\begin{bmatrix}\mathbf{C} & \mathbf{-1} \\ \mathbf{1} & 0\end{bmatrix}\begin{bmatrix}\mathbf{w}_{opt} \\ \lambda\end{bmatrix} = \begin{bmatrix}\mathbf{0} \\ 1\end{bmatrix}$$

$$\begin{bmatrix}\mathbf{w}_{opt} \\ \lambda\end{bmatrix} = \begin{bmatrix}\mathbf{0} \\ 1\end{bmatrix}\begin{bmatrix}\mathbf{C} & \mathbf{-1} \\ \mathbf{1} & 0\end{bmatrix}^{-1}$$

$\mathbf{w_{opt}}$ was calculated using the train set (first half of the data) and is used in the test set (second half of the data). Its performance is compared to $\mathbf{w_{equal}}$ where all the companies are weighted equally. The cumulative return using the two different strategies are plotted below and the variance for each strategy is given in Table:

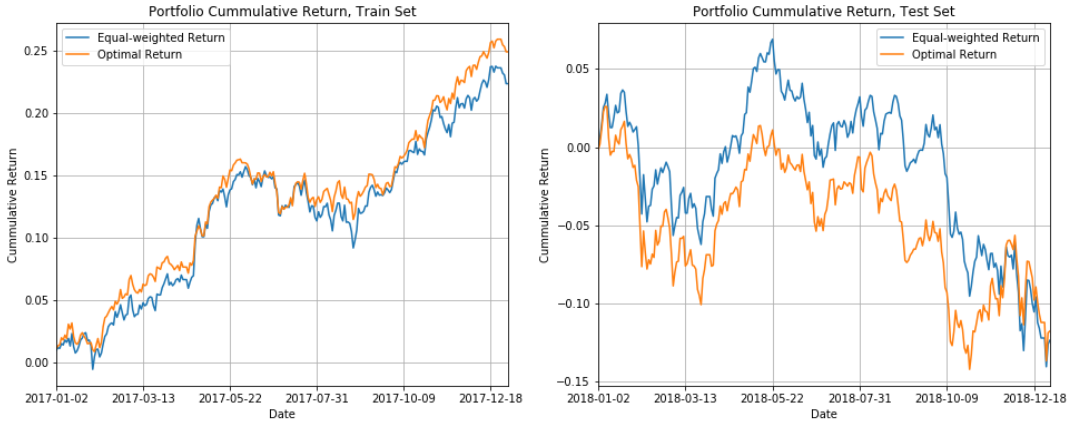

Figure 12: Cumulative return using the two different strategies.

	$\mathbf{w_{opt}}$, Train Set	$\mathbf{w_{opt}}$, Test Set	$\mathbf{w_{equal}}$, Test Set
Variance	0.000029	0.000082	0.00008

Table 8: Variance of Different Strategies.

It is seen that in the training set, the return using $\mathbf{w_{opt}}$ outperforms the return using $\mathbf{w_{equal}}$ but the opposite is true for the test set. The same is also true when looking at the variance of each strategies. The variance is a benchmark of risk of a portfolio and should be ideally kept as low as possible. It is seen that the variance is lowest in the training set when $\mathbf{w_{opt}}$ is used but in the test set, the $\mathbf{w_{equal}}$ strategy outperforms the $\mathbf{w_{opt}}$.

This is due to the fact that $\mathbf{w_{opt}}$ was trained in a period of time where the return for each companies are generally positive but was used in a period where the return was generally negative. It also showed that return is difficult predict and that using a static $\mathbf{w}$ to manage a portfolio is generally a bad strategy.

3. From 13, it is seen that the adaptive minimum variance strategy outperforms both the strategies using $\mathbf{w_{opt}}$ and $\mathbf{w_{equal}}$. From Table 9, it is seen that it has the highest mean return, highest minimum return, smaller variance (less risk), and a higher median. It is hence concluded that the adaptive minimum variance strategy is the best overall strategy.

The better performance is due to the fact that the prediction is made based on a closer past history. Conventionally, time-series data are more highly correlated to closer past history. Updating the prediction recursively makes predictions more accurate and more sensitive to change.

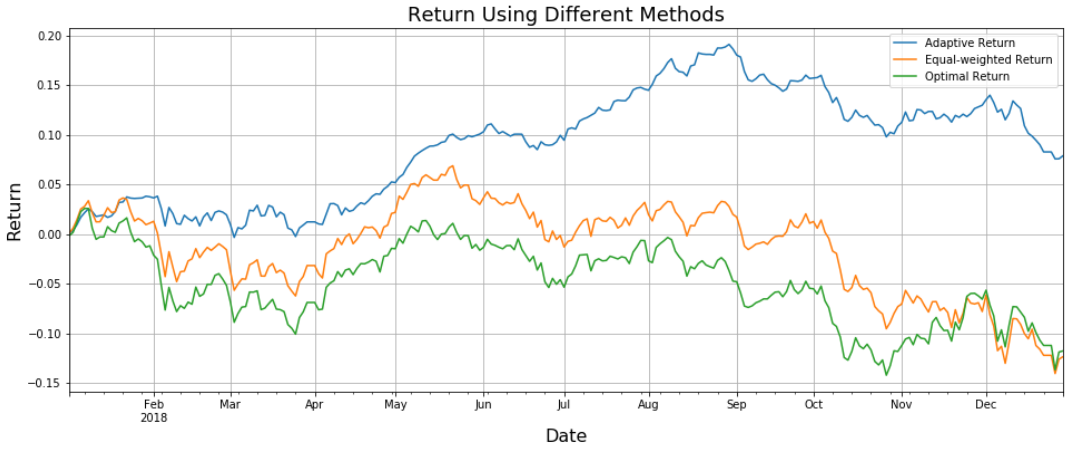

Figure 13: Cumulative return using different strategies.

	Adaptive Return	**Equal-Weighted Return**	**Optimal Return**
Mean	0.000303	-0.000475	-0.000452
Variance	0.000031	0.000079	0.000082
Median	0.000564	0.000138	-0.000088
Minimum	-0.017477	-0.024691	-0.026380

Table 9: Performance of Different Strategies.

4 Robust Statistics and Non Linear Methods

4.1 Data Import and Exploratory Data Analysis

1. The statistics of the required stocks are given in tables below:

	Open	**High**	**Low**	**Close**	***Adj. Close***	**Volume**
Mean	187.766640	189.646154	185.891336	187.786357	186.223776	3.272384e+07
Standard Deviation	22.314686	22.450839	22.179512	22.330877	22.077041	1.427659e+07
Min	143.979996	145.720001	142.000000	142.190002	141.582779	1.251390e+07
25%	171.220001	172.839996	169.625000	170.944999	170.209107	2.283110e+07
50%	186.350006	187.809998	185.100006	186.440002	184.411102	2.901770e+07
75%	207.529999	209.640000	206.379997	208.240005	206.232834	3.938885e+07
Max	230.779999	233.470001	229.779999	232.070007	230.275482	9.624670e+07

Table 10: Statistics of AAPL.

	Open	**High**	**Low**	**Close**	***Adj. Close***	**Volume**
Mean	138.451822	139.490770	137.314332	138.356599	134.840205	5.214225e+06
Standard Deviation	12.212233	12.009360	12.302672	12.125288	10.746462	3.346461e+06
Min	108.000000	111.000000	105.940002	107.570000	106.331108	1.963200e+06
25%	130.909996	131.894996	128.404999	130.285004	127.164894	3.454950e+06
50%	142.929993	144.080002	142.179993	143.000000	138.566391	4.239900e+06
75%	146.674995	147.300003	145.615005	146.419998	141.873985	5.389000e+06
Max	160.059998	162.000000	159.639999	160.910004	153.671936	2.206370e+07

Table 11: Statistics of IBM.

	Open	**High**	**Low**	**Close**	***Adj. Close***	**Volume**
Mean	108.773036	109.717530	107.737004	108.663684	107.297936	1.470041e+07
Standard Deviation	5.377210	5.218206	5.459230	5.322690	4.862787	5.341134e+06
Min	92.690002	94.220001	91.110001	92.139999	91.397758	6.488400e+06
25%	104.694999	105.465000	103.680001	104.689999	104.1399	1.081290e+07
50%	109.449997	110.730003	107.809998	109.089996	107.506935	1.374490e+07
75%	113.370003	114.310001	112.599998	113.360000	111.431881	1.704650e+07
Max	119.129997	119.239998	118.080002	118.629997	116.856049	4.131390e+07

Table 12: Statistics of JPM.

The simple return and log return of each of the stocks are calculated and added as a new column to each dataframe:

Date	**Log Return**	**Simple Return**
2018-03-16	0.000000	0.000000
2018-03-19	-0.018325	-0.018158
2018-03-20	-0.007335	-0.007309
2018-03-21	0.003132	0.003137
2018-03-22	-0.029797	-0.029357
⋮	⋮	⋮

Table 13: Log return and simple return of AAPL.

Date	Log Return	Simple Return
2018-03-16	0.000000	0.000000
2018-03-19	-0.007914	-0.007883
2018-03-20	0.000960	0.000960
2018-03-21	0.000872	0.000872
2018-03-22	-0.042643	-0.041747
⋮	⋮	⋮

Table 14: Log return and simple return of IBM.

Date	Log Return	Simple Return
2018-03-16	0.000000	0.000000
2018-03-19	-0.013544	-0.013453
2018-03-20	-0.004717	-0.004728
2018-03-21	-0.001820	-0.001818
2018-03-22	-0.029789	-0.029350
⋮	⋮	⋮

Table 15: Log return and simple return of JPM.

2. The histogram and probability density function of the *adj. close* and log-return are plotted in Figure 14:

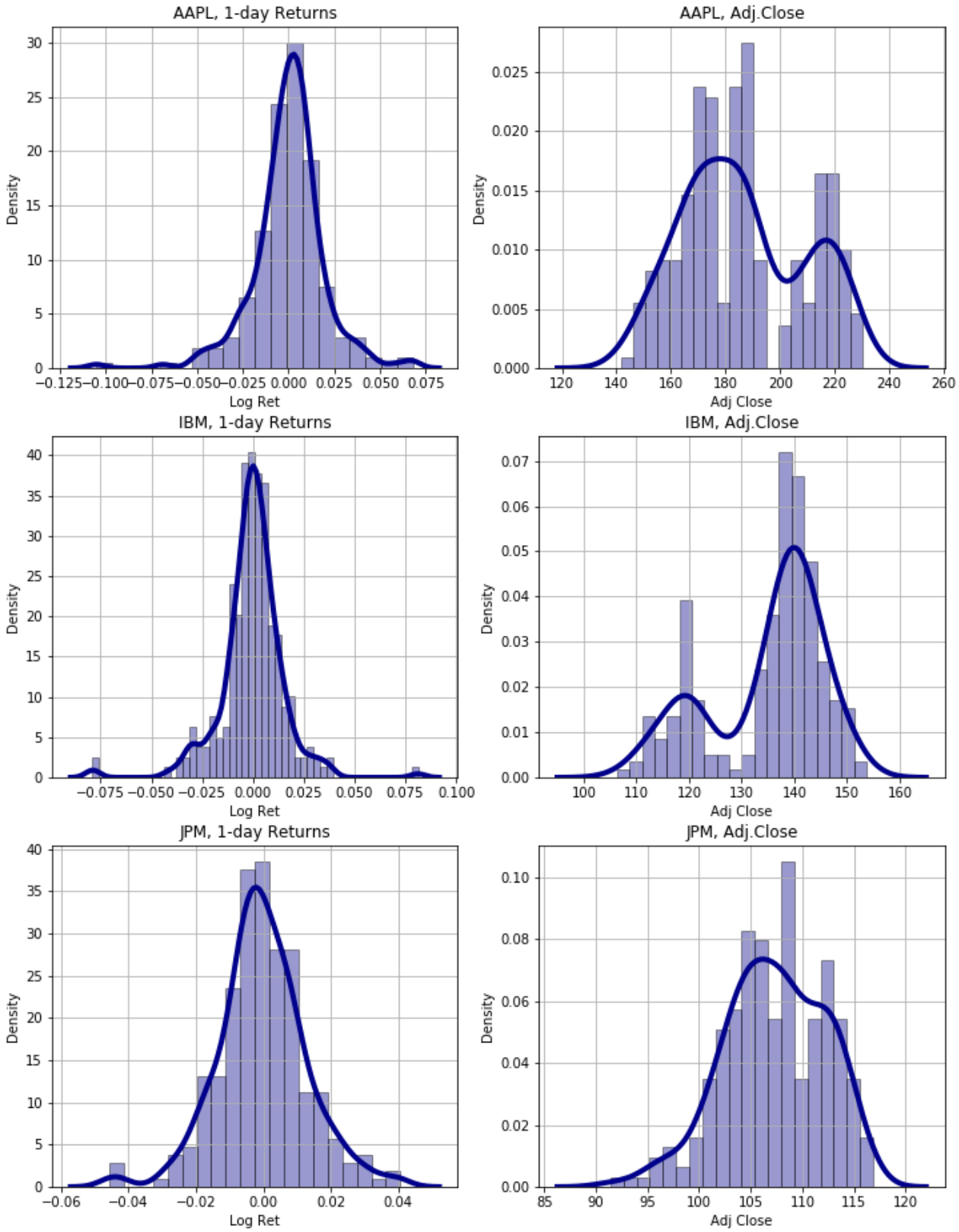

Figure 14: *Adj. close* and log-return of all required stocks.

It is observed that the log-returns is normally distributed i.e. has a bell-shaped distribution with a mean of around 0 where as the *adj. close* does not.

3. The rolling mean and rolling median of *adj. close* with a 5-day window of each of the stocks are plotted in Figure 15

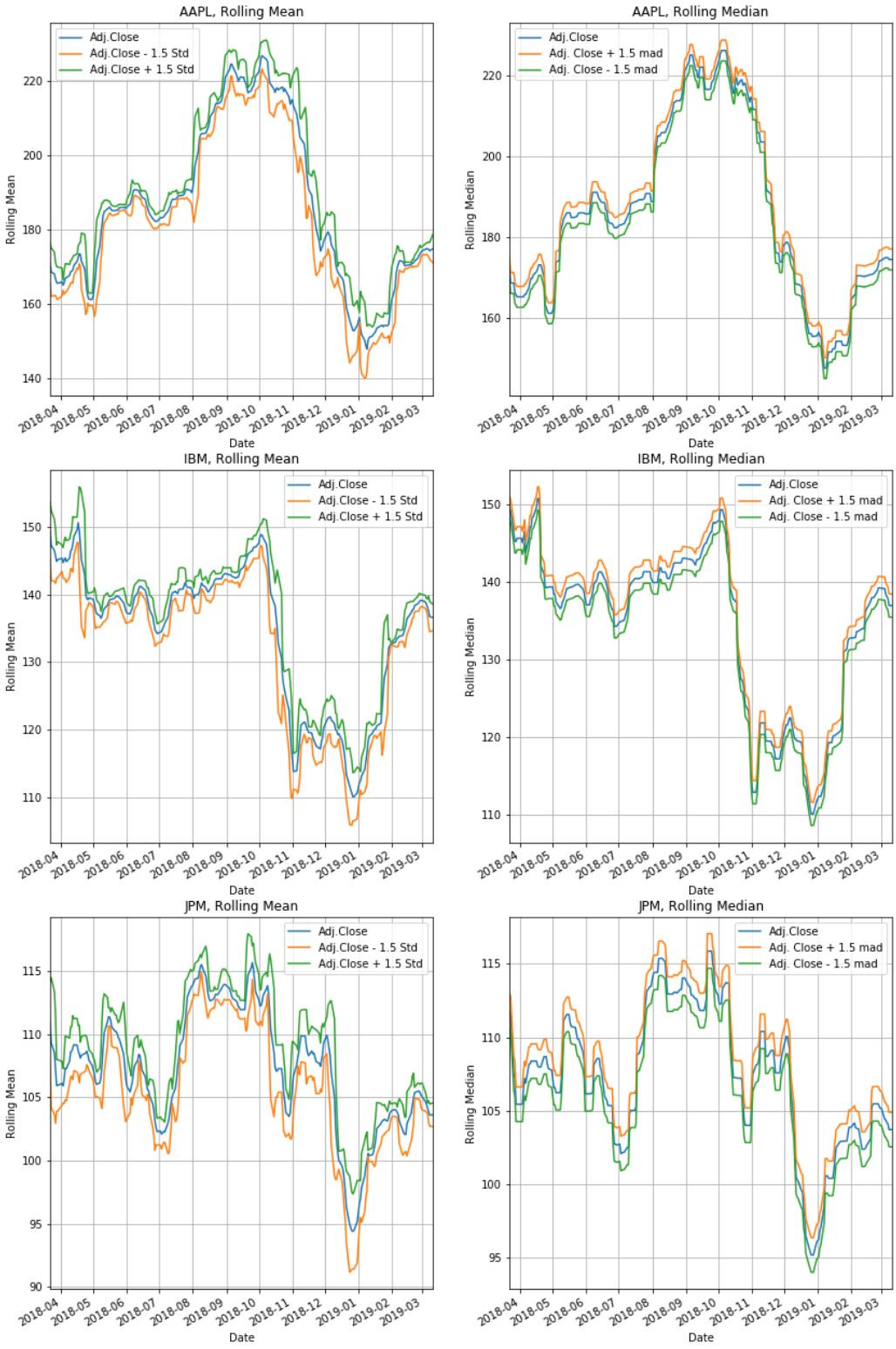

Figure 15: Rolling mean and rolling median of *adj. close* with a 5-day window.

It is observed that the rolling mean and rolling median for each of the stocks are very similar. Analytically, they have very similar trend and it is difficult to judge if which one statistics (mean or median) gives a better representation of the stocks.

4. After introducing some outliers, the rolling mean and rolling median of *adj. close* with a 5-day window of each of the stocks are again plotted in Figure 16:

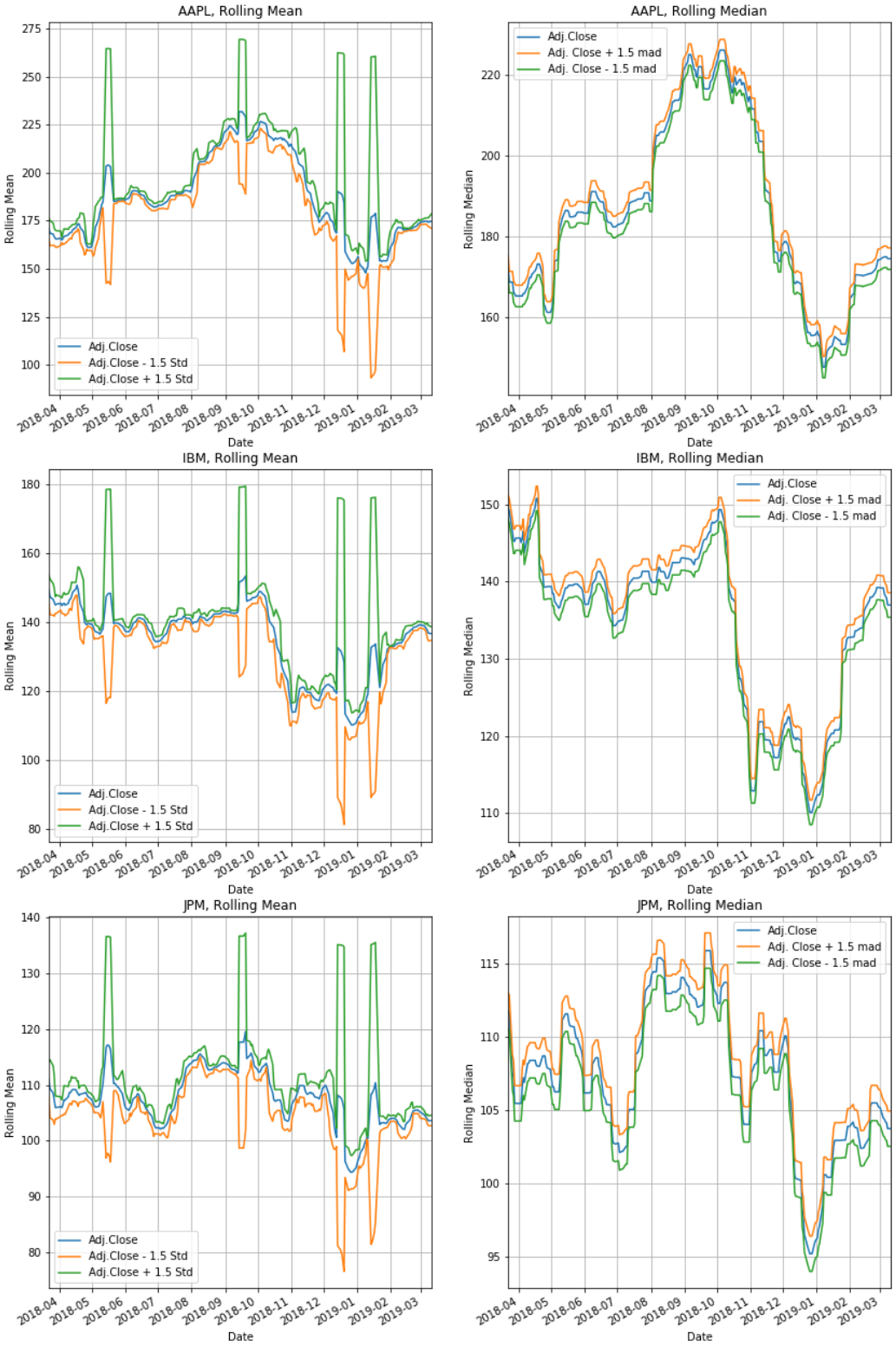

Figure 16: Rolling mean and rolling median of *adj. close* with a 5-day window and added outliers.

It is observed that the rolling median with outliers has a more similar trend to the original data as compared to the rolling mean. Besides that, there is obvious spikes in the rolling mean where there is an outlier. This suggests that rolling median is more robust to outliers.

5. Figure 17 illustrates how a box plot is interpreted in comparison to a distribution plot. It also shows the box plot of normal distribution in which we will use to compare the distribution of each stocks.

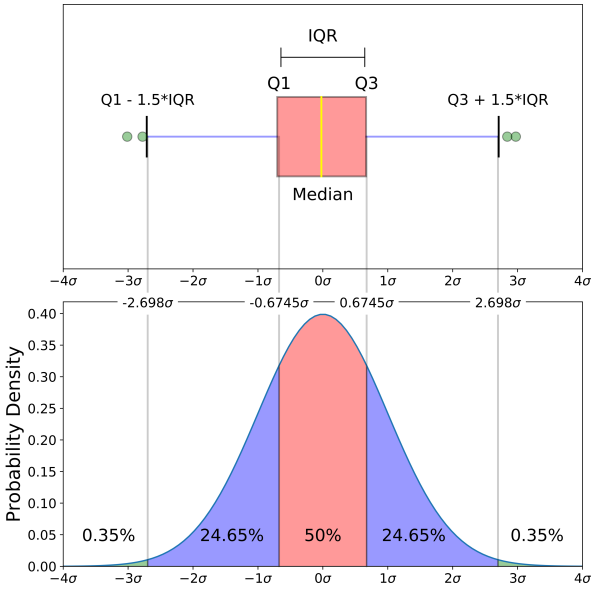

Figure 17: Box plot explained. (Galarnyk 2018)

and Figure 18 shows the boxplot of each of the stocks.

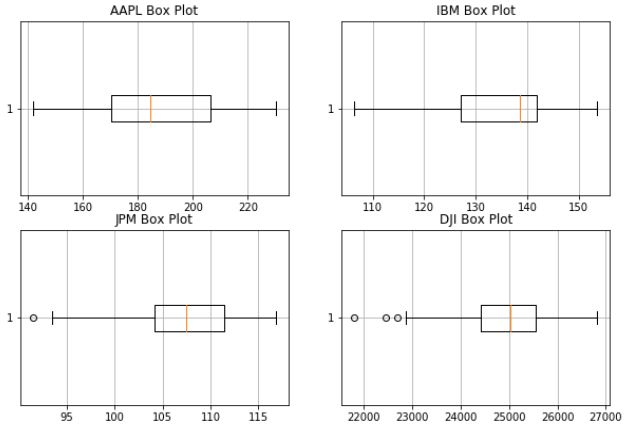

Figure 18: Boxplot of each of the stocks.

From a boxplot, we get a compact view of a distribution as well as a clearer view of the outliers. It is observed that all the stocks are not normally distributed as the median is not in the centre of the distribution. It is observed that IBM and JPM are skewed to the right while AAPL is slightly skewed to the left.

Out of the three stocks, AAPL has the biggest interquartile range meaning it distribution is the flattest.

4.2 Robust Estimators

1. The robust location estimator (median) and robust scale estimator (IQR (Interquartile range) and MAD (Median Absolute Deviation)) of the required stocks are calculated and presented below:

Stocks	Median	IQR	MAD
AAPL	184.411102	36.023727	15.709945
IBM	138.566391	14.709091	4.543045
JPM	107.506935	7.291882	3.626938

Table 16: All estimators for each stocks.

2. To look for the median, MAD and IQR, we need to sort the data. Since, sorting algorithms has a complexity of $O(nlog(n))$ and that to look for the median and IQR, the data needs to be sorted only once, both of the calculations has a computational complexity of $O(nlog(n))$.

To calculate the MAD on the other hand, we need to sort the data twice, hence it has a computational complexity of $2O(nlog(n))$.

3. To assess the breakdown points of each estimator, we plot the absolute error of each estimator against the percentage of outliers:

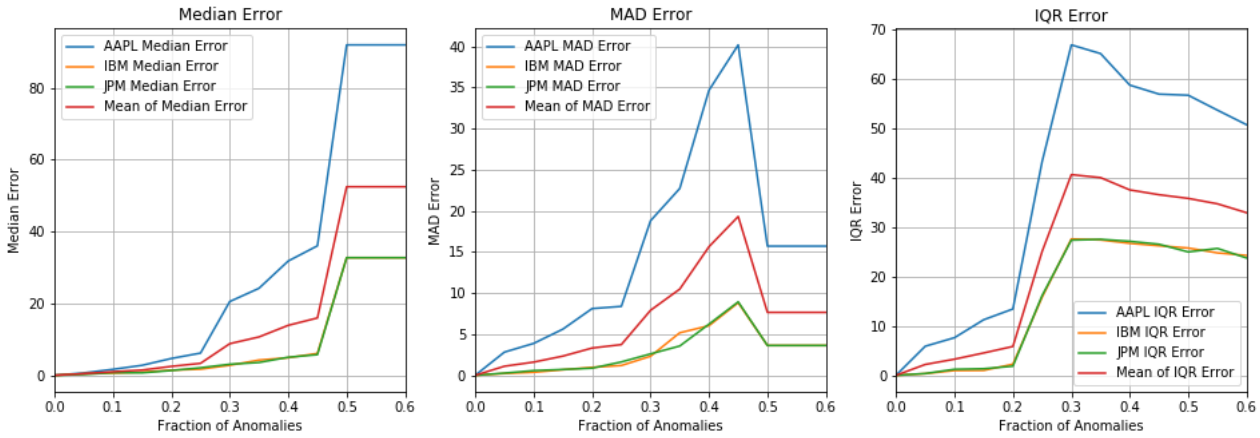

Figure 19: Absolute error of eatimator vs percentage of outliers.

It is seen that the IQR has a sharp increase when the anomalies reaches 20% and is the least robust out of the three statistics. Comparing the median error and MAD error, it is seen that both the statistics has an absolute error of 20 around 35% anomalies. However, the median error goes over 80 when there is about 50% anomalies. Hence, we can deduce the order of robustness of the three statistics, from most robust to least robust as: MAD, median and IQR.

4.3 Robust and OLS regression

DJI (The Dow Jones Industrial Average) is a stock market index which measures the daily price movements of 30 large American companies. Here, we are studying the performance of the three stocks (AAPL, IBM and JPM) in relation to the DJI index using regression. Below shows the plot of each stock vs the DJI market index with an ordinary least squared (OLS) regression:

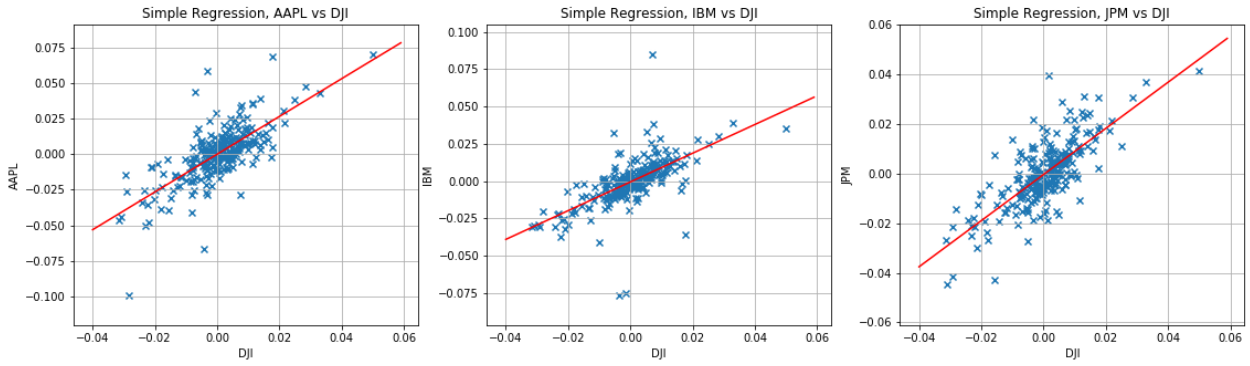

Figure 20: Individual stock vs DJI, ordinary least squared regressor.

The properties of the ordinary least squared regressors are given below. "Score" is a measure of how well the regressor fits the actual data. Note that "Score" has a range of ≤ 1 and that the higher "Score" the better the regressor fits the data.

Stock	Coefficient	Intercept	Score
AAPL	1.327464	0.000047	0.518874
IBM	0.961929	-0.000468	0.418597
JPM	0.930549	-0.000377	0.556493

Table 17: Properties of ordinary least squared regressors.

2. We then do the same analysis but instead of using OLS regression, we use Huber regression which is the similar to OLS regression but with an added parameter, epsilon. Epsilon is used to regularise the model and prevent overfitting. It is essentially a parameter which determines how much of the data should be considered as outliers. The smaller epsilon, the more robust the model will be to outliers. Hence, it is inferred that the Huber regressor is less prone to effects from outliers. The plot of both the Huber regressor and the OLS regressor are given below:

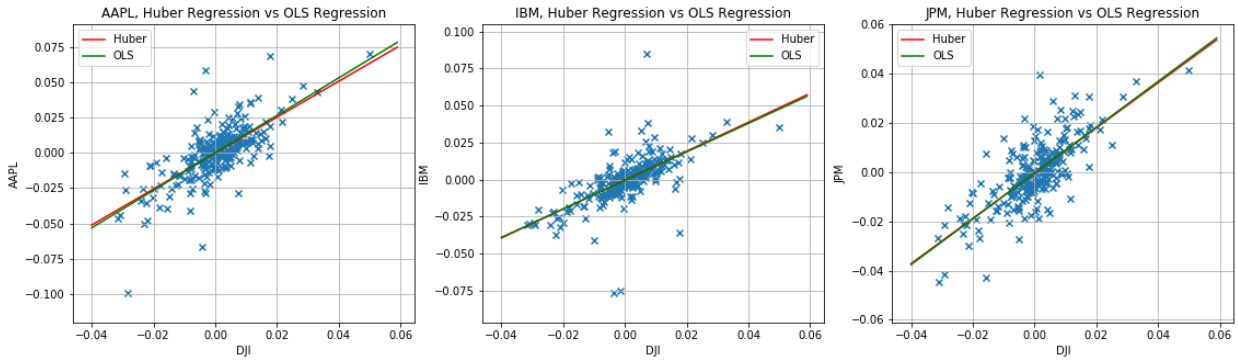

Figure 21: Huber regression vs OLS regression for each individual stocks.

The properties of the Huber regressors are given below:

Stock	Coefficient	Intercept	Score
AAPL	1.272508	-0.000136	0.517697
IBM	0.976089	-0.000266	0.418488
JPM	0.918433	-0.000439	0.554913

Table 18: Properties of huber regressors.

3. Analysing the plots in Figure 21, it is seen that the two regression methods resulted in very similar plots. To properly compare the performance of the two regression methods, the data was split into 80-20 for training and testing. The error of each methods are given below:

Error	AAPL	IBM	JPM	Mean
OLS Training Error	0.000160	0.000129	0.000078	0.000123
OLS Test Error	0.000275	0.000194	0.000070	0.000180
Huber Training Error	0.000162	0.000129	0.000078	0.000123
Huber Test Error	0.000265	0.000194	0.000071	0.000177

Table 19: Training error and test error of OLS regression and Huber regression.

From the training error and test, it is seen that both the OLS regression and Huber regression gave very similar results. This is because there are no obvious outliers in the data. However, looking at the mean of the error, it is seen that though the two methods have the same training error, the Huber method has a slightly lower test error, suggesting that it is the better method. However, the results are not conclusive as the differences are negligible.

To further test the two methods, we simulate the presence of outliers by artificially adding outliers. We do this multiplying 20% of the samples with a random number between 1.2 and 2. The same analysis is performed per the previous step. The plots are given in Figure 22. The training error and test error are given in Table 20.

Figure 22: Individual stock vs DJI, ordinary least squared regressor and huber regressor, with artificially added outliers.

Error	AAPL	IBM	JPM	Mean
OLS Training Error	0.191910	0.160743	0.129132	0.160595
OLS Test Error	0.239409	0.218829	0.076199	0.178146
Huber Training Error	0.196935	0.168618	0.134928	0.166827
Huber Test Error	0.248774	0.214393	0.065880	0.176349

Table 20: Training error and test error of OLS regression and Huber regression, with artificially added outliers.

After the outliers are added, the difference between the two regression methods are now visibly distinguishable from the plots. Looking at the training and test error, it is seen that the OLS training error is lower than the Huber training error. However, the OLS test error is higher than the Huber test error. This suggests that the OLS regression overfits the data. Hence, we can conclude that the Huber regression method is more robust against outliers. This is further justified by the change in the properties of the regressor after outliers are introduced:

Stock	Percentage Change in Coefficient (%)	Percentage Change in Intercept (%)
AAPL	76.488785	91.946750
IBM	401.644468	92.925314
JPM	259.101116	95.216036
Mean	245.744790	93.362700

Table 21: Percentage change in the OLS regressor after outliers are introduced.

Stock	Percentage Change in Coefficient (%)	Percentage Change in Intercept (%)
AAPL	4.641351	99.964096
IBM	9.618628	100.013638
JPM	0.966580	100.035304
Mean	5.075519	100.004346

Table 22: Percentage change in the Huber regressor after outliers are introduced.

If a regression method is robust to outliers, its regressor parameters should remain relatively constant even after outliers are introduced. From the table above, we see that for both the methods, there is significant change in the intercept of the regressor. However, when it comes to the coefficient of the regressor, the Huber regressor only has an average of 5% change as compared to about 246% of the OLS regressor. This further reinforce the deduction that the Huber regression method is more robust to outliers as compared to the OLS regression method.

4.4 Robust Trading Strategies

1. The *adj. close* for each individual stock, as well as its 20-day moving average and 50-day average are plotted below on the left. This is compared to the same analysis but performed on data corrupted with outliers. The outliers are introduced by artificially setting 10% of the stock as $1.5 \times max(adj.\ close)$. These are plotted on the right:

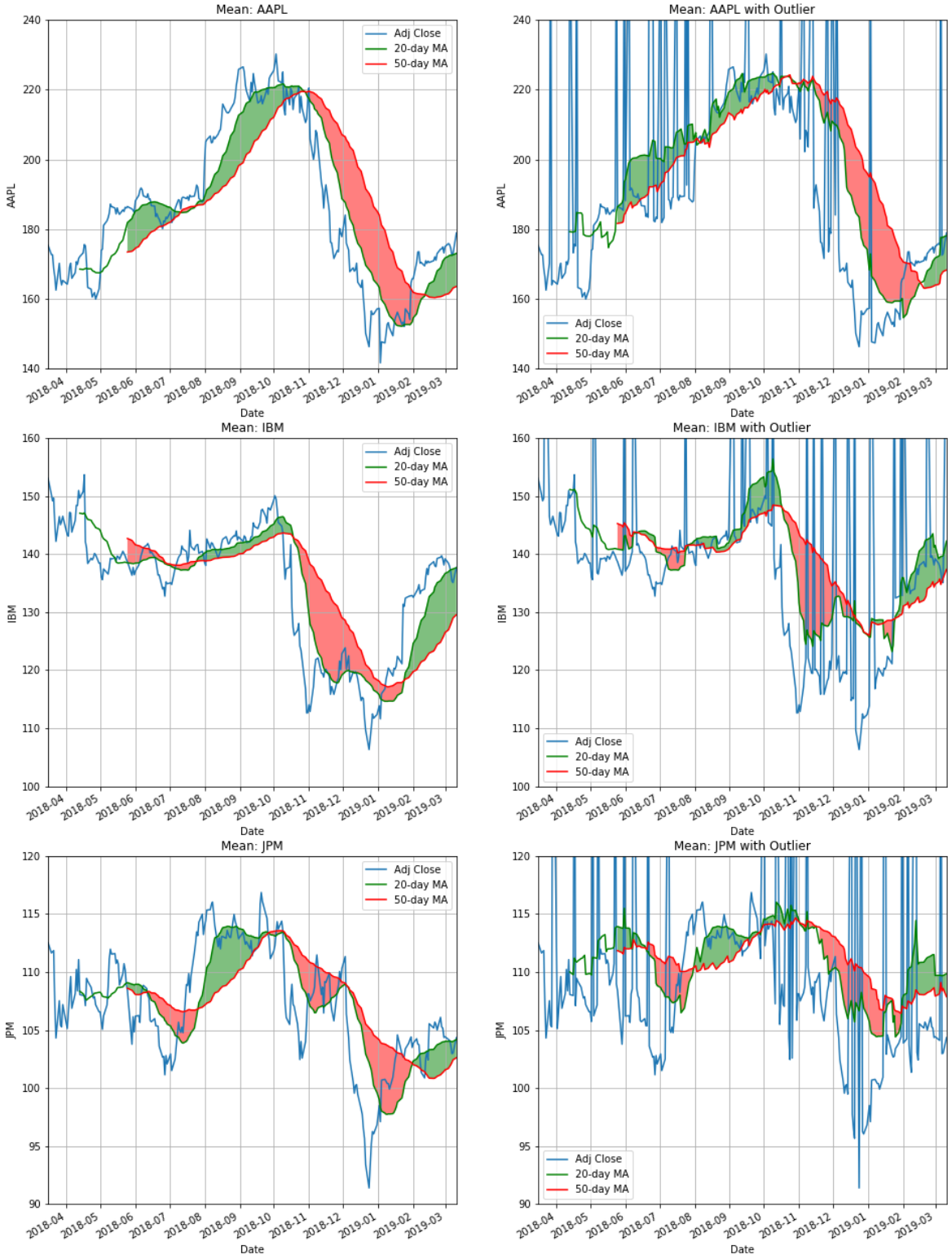

Figure 23: 20-day moving-mean and 50-day moving average of AAPL, IBM and JPM. Left: original data; Right: data corrupted with outliers.

2. The analysis in 4.4.1 is repeated but instead of using rolling-mean, rolling median is used. The plot is given below:

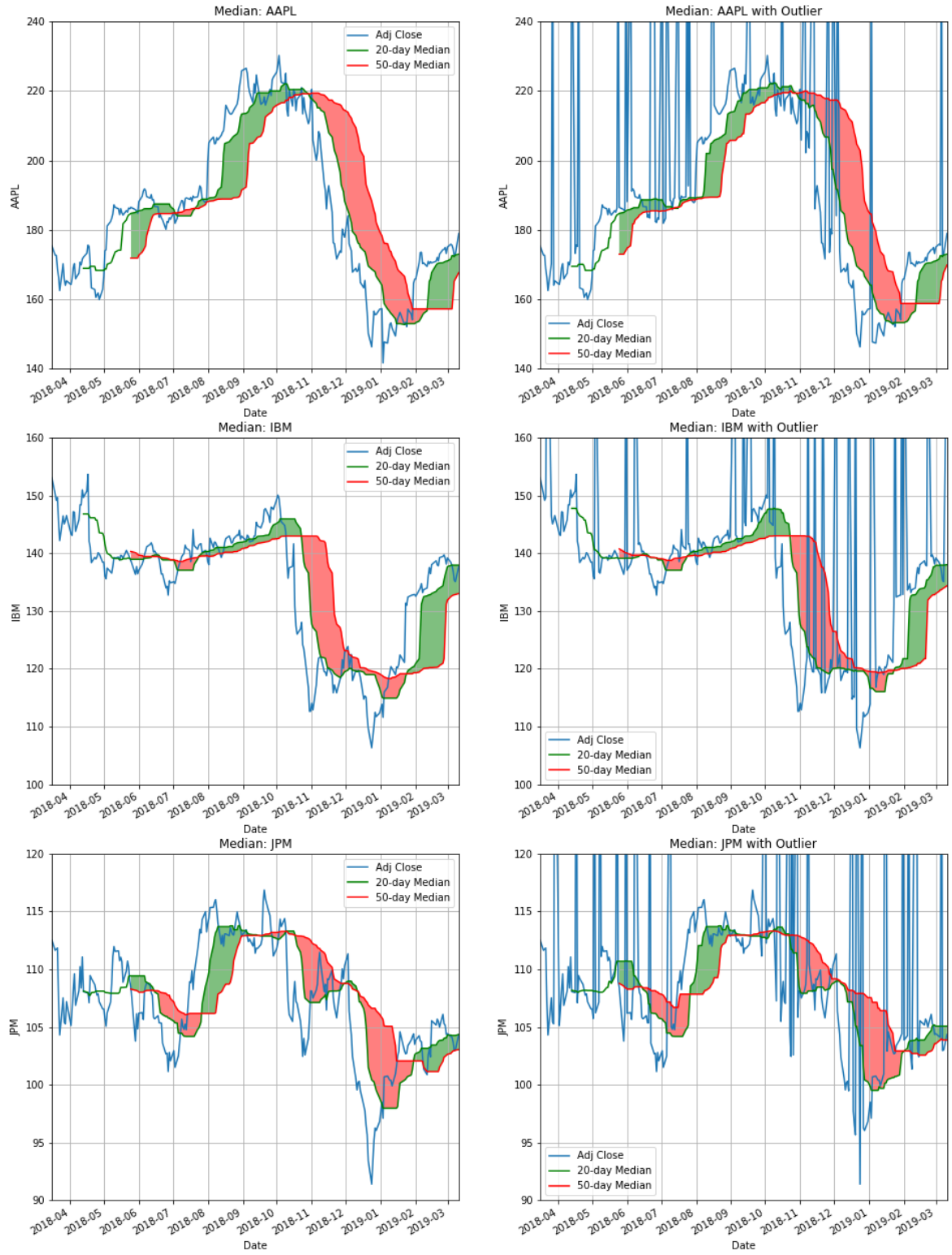

Figure 24: 20-day moving-median and 50-day moving average of AAPL, IBM and JPM. Left: original data; Right: data corrupted with outliers.

It is observed that when outliers are introduced, the rolling-mean strategy change more significantly as compared to the rolling-median strategy. This means that rolling-median is more robust to outliers. However, it is also observed that rolling-median is less sensitive to a sudden change in price and hence has a slower response time which is also a major drawback when it comes to financial strategies.

5 Graphs in Finance

1. To build a portfolio, there are two criteria we need to optimise:

    1. We need to choose a selection of diverse stocks to diversity our risks.

    2. We also need to choose stocks which have high return and have low risks (small variance).

Since stocks from different sectors are usually non-correlated to each other, we aim to choose a selection of stocks from a few high-performing sectors. Our strategy prioritises high return over diversity and hence we first narrow down on high performing stocks before considering their diversity. We do this by first looking at the performance of each sector. The mean cumulative return of all the stocks in each sector are plotted below:

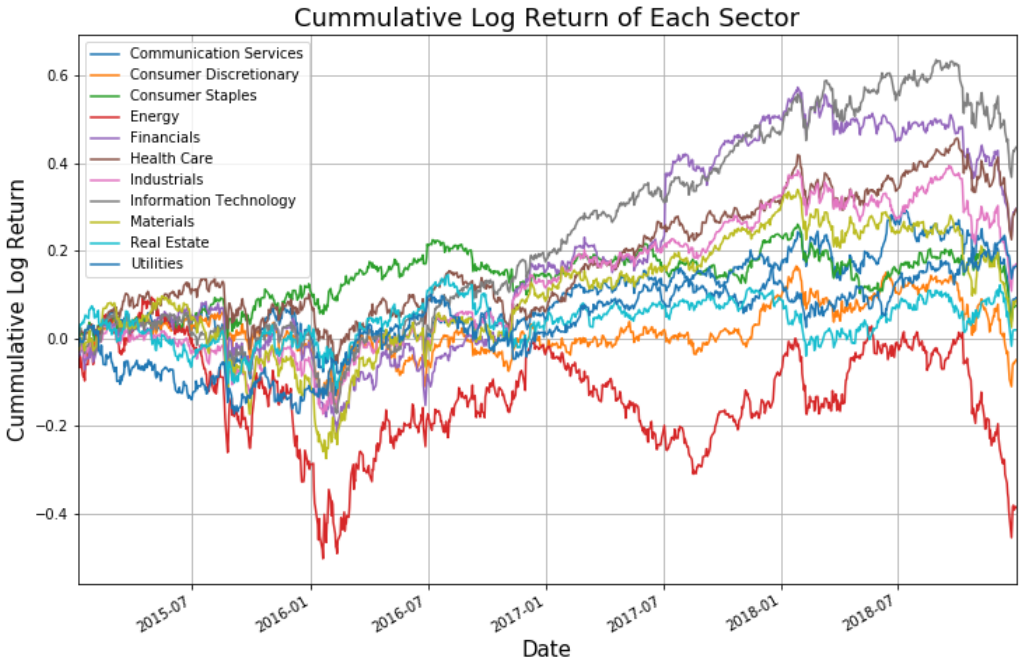

Figure 25: Cumulative return of stocks in each sector.

We consider two factors in determining if a sector is high-performing:

1. The sector's relative cumulative return ranking compared to the other sectors.

2. The percentage of times where the sectors's cumulative return is positive. This is a measure of stability and but not necessarily mean high return.

The mean ranking of each sector is given below:

Sector	Relative Mean Ranking
Health Care	2.790050
Information Technology	3.083582
Consumer Staples	4.040796
Financials	4.467662
Industrials	5.737313
Communication Services	5.949254
Materials	6.061692
Real Estate	7.368159
Utilities	7.934328
Consumer Discretionary	8.152239
Energy	10.414925

Table 23: Relative mean ranking of each sector as compared to all the other sectors in the market.

The percentage of times where the stock has a positive cumulative return is given below:

Sector	Percentage of Positive cumulative Return (%)
Consumer Staples	99.502982
Health Care	97.514911
Communication Services	86.481113
Information Technology	84.890656
Materials	77.435388
Real Estate	76.938370
Financials	72.962227
Industrials	70.874751
Utilities	66.500994
Consumer Discretionary	61.332008
Energy	11.232604

Table 24: Percentage of times the sector has a positive cumulative return.

From these two analysis, it is observed that **Health Care**, **Consumer Staples** and **Information Technology** have high ranking and have positive cumulative return almost all the time so these three sectors is the clear choice. Now we choose about three stocks per sector to make up our portfolio. We do this by considering the correlation between the top 10 stocks for each of the chosen sector and choose stocks which are non-correlated to each other. The top 10 stocks are chosen based on their average relative ranking to all the other stocks in the same sector. The correlation matrix for the top 10 stocks from each of the three chosen sectors are given below:

Stock	ABMD	HCA	EW	RMD	IDXX	BSX	ZTS	ISRG	ILMN	HUM
ABMD	1.000000	0.321079	0.625685	0.456121	0.595400	0.518321	0.513833	0.618545	0.565473	0.321971
HCA	0.321079	1.000000	0.425115	0.324982	0.456980	0.529485	0.488610	0.470236	0.454065	0.432983
EW	0.625685	0.425115	1.000000	0.473391	0.627573	0.572661	0.571713	0.670362	0.648146	0.378275
RMD	0.456121	0.324982	0.473391	1.000000	0.530097	0.525396	0.457241	0.594032	0.531564	0.353146
IDXX	0.595400	0.456980	0.627573	0.530097	1.000000	0.615835	0.629013	0.717784	0.698816	0.416635
BSX	0.518321	0.529485	0.572661	0.525396	0.615835	1.000000	0.635066	0.681290	0.640594	0.453666
ZTS	0.513833	0.488610	0.571713	0.457241	0.629013	0.635066	1.000000	0.584844	0.611466	0.463843
ISRG	0.618545	0.470236	0.670362	0.594032	0.717784	0.681290	0.584844	1.000000	0.724275	0.487239
ILMN	0.565473	0.454065	0.648146	0.531564	0.698816	0.640594	0.611466	0.724275	1.000000	0.420619
HUM	0.321971	0.432983	0.378275	0.353146	0.416635	0.453666	0.463843	0.487239	0.420619	1.000000

Table 25: Correlation matrix of the top 10 performing stocks in **Health Care**.

Stock	AMD	NVDA	AMAT	LRCX	AVGO	IDXX	ADSK	ANET	IPGP	GLW
AMD	1.000000	0.436901	0.382220	0.359832	0.295197	0.302496	0.332777	0.288108	0.285068	0.294934
NVDA	0.436901	1.000000	0.542010	0.485837	0.417633	0.433719	0.424984	0.415368	0.359448	0.383506
AMAT	0.382220	0.542010	1.000000	0.838504	0.527319	0.689005	0.497073	0.434167	0.440061	0.480213
LRCX	0.359832	0.485837	0.838504	1.000000	0.508011	0.768501	0.525700	0.414618	0.429826	0.484178
AVGO	0.295197	0.417633	0.527319	0.508011	1.000000	0.429853	0.382352	0.337219	0.350023	0.411429
KLAC	0.302496	0.433719	0.689005	0.768501	0.429853	1.000000	0.437607	0.361586	0.294564	0.431493
ADSK	0.332777	0.424984	0.497073	0.525700	0.382352	0.437607	1.000000	0.469163	0.352653	0.461261
ANET	0.288108	0.415368	0.434167	0.414618	0.337219	0.361586	0.469163	1.000000	0.364941	0.357774
IPGP	0.285068	0.359448	0.440061	0.429826	0.350023	0.294564	0.352653	0.364941	1.000000	0.382847
GLW	0.294934	0.383506	0.480213	0.484178	0.411429	0.431493	0.461261	0.357774	0.382847	1.000000

Table 26: Correlation matrix of the top 10 performing stocks in **Information Technology**.

Stock	LW	EL	COST	CLX	MNST	STZ	CHD	BF-B	MKC	WMT
LW	1.000000	0.180040	0.216745	0.216726	0.157032	0.237846	0.210414	0.229148	0.271430	0.156731
EL	0.180040	1.000000	0.296877	0.279527	0.258291	0.206047	0.359539	0.329124	0.247611	0.215134
COST	0.216745	0.296877	1.000000	0.282564	0.235999	0.185208	0.270576	0.279337	0.357578	0.460177
CLX	0.216726	0.279527	0.282564	1.000000	0.274454	0.163345	0.720428	0.301871	0.530875	0.302839
MNST	0.157032	0.258291	0.235999	0.274454	1.000000	0.243161	0.288227	0.288487	0.310416	0.272818
STZ	0.237846	0.206047	0.185208	0.163345	0.243161	1.000000	0.189598	0.327811	0.175875	0.140046
CHD	0.210414	0.359539	0.270576	0.720428	0.288227	0.189598	1.000000	0.312941	0.475689	0.255202
BF-B	0.229148	0.329124	0.279337	0.301871	0.288487	0.327811	0.312941	1.000000	0.357214	0.227889
MKC	0.271430	0.247611	0.357578	0.530875	0.310416	0.175875	0.475689	0.357214	1.000000	0.293771
WMT	0.156731	0.215134	0.460177	0.302839	0.272818	0.140046	0.255202	0.227889	0.293771	1.000000

Table 27: Correlation matrix of the top 10 performing stocks in **Consumer Staples**.

Finally we choose the stock which are weakly correlated to each other from each of the three sectors. They are:

- Health Care: ABMD, HCA, HUM
- Information Technology: AMD, AVGO
- Consumer Staples: EL, LW, STZ

The cumulative return of the eight chosen stocks are plotted below:

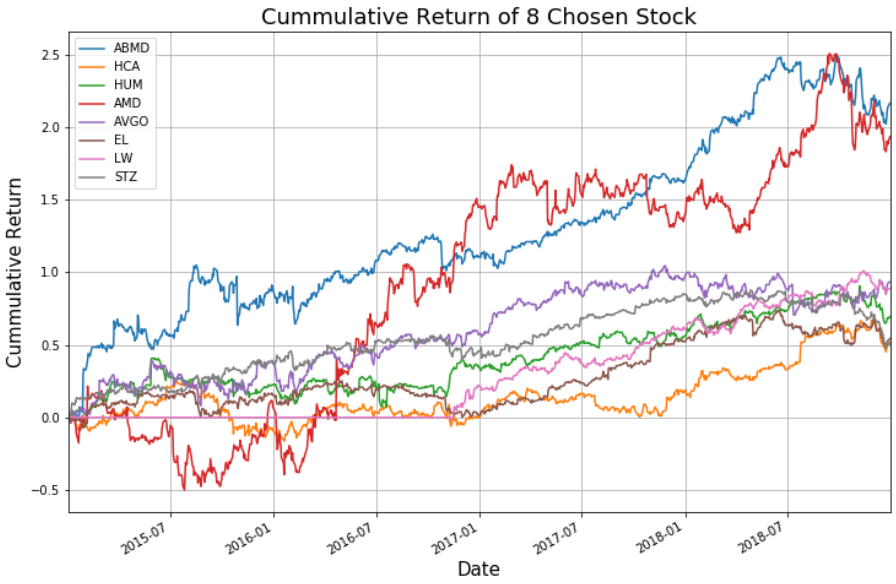

Figure 26: Cumulative return of the eight chosen stocks.

It is observed that all the chosen stocks are indeed high performing with high overall cumulative return and have high percentage of positive cumulative return.

2. Now we have to consider the correlation of between all pairs of the chosen stocks to make sure that they are indeed not strongly correlated to each other. To do this, we construct the correlation matrix:

Stock	**ABMD**	**HCA**	**HUM**	**AMD**	**AVGO**	**EL**	**LW**	**STZ**
ABMD	1.000000	0.243607	0.172734	0.189414	0.278935	0.237274	0.104637	0.251820
HCA	0.243607	1.000000	0.328667	0.191557	0.299671	0.232230	0.089526	0.285997
HUM	0.172734	0.328667	1.000000	0.095141	0.204707	0.157882	0.066254	0.177742
AMD	0.189414	0.191557	0.095141	1.000000	0.282439	0.210468	0.093811	0.109505
AVGO	0.278935	0.299671	0.204707	0.282439	1.000000	0.279596	0.131966	0.267001
EL	0.237274	0.232230	0.157882	0.210468	0.279596	1.000000	0.135583	0.267256
LW	0.104637	0.089526	0.066254	0.093811	0.131966	0.135583	1.000000	0.166336
STZ	0.251820	0.285997	0.177742	0.109505	0.267001	0.267256	0.166336	1.000000

Table 28: Correlation matrix of the eight chosen stocks.

The correlation matrix gives the correlation between all pairs of stocks which were chosen. As expected, the stocks from different sectors exhibit low correlation from each other. Since we chose stocks in the same sector to have low correlation, the correlation between all the pairs of chosen stocks (even the ones from the same sector) are low. The strongest correlation is observed to be between *HUM* and *HCA* at 0.328677. This is desirable as a low correlation means that the risk in our portfolio is diversified.

To construct a correlation graph, we would like to present as much information as possible but at the same time make sure that the visualisation is not congested which makes the graph difficult to interpret. In order to achieve this, a good strategy would be to plot the distribution of all the correlation and determine the threshold for the correlation to be included in the plot. The distribution of the correlation is plotted below:

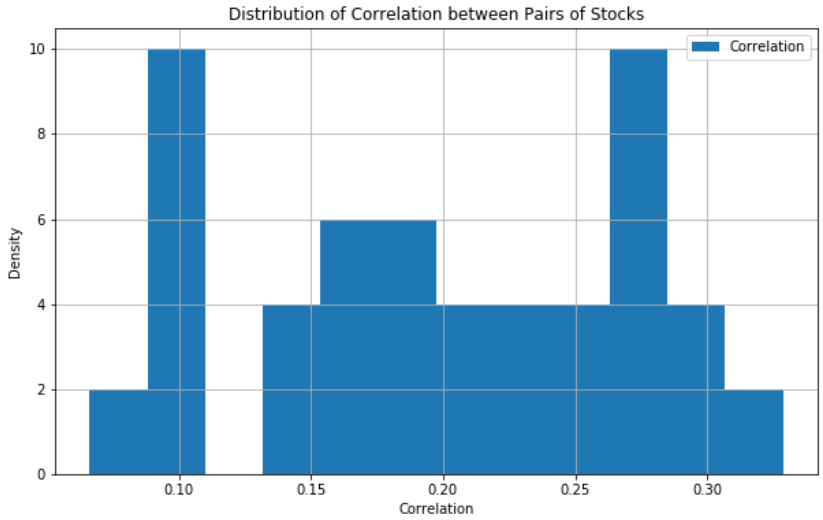

Figure 27: Distribution of correlation between pairs of selected stock.

It is seen that there is a spike of correlation at about 0.125. Hence, we choose the threshold to be 0.125, and only plot the correlation which are larger than this threshold. The graph plot is plotted below:

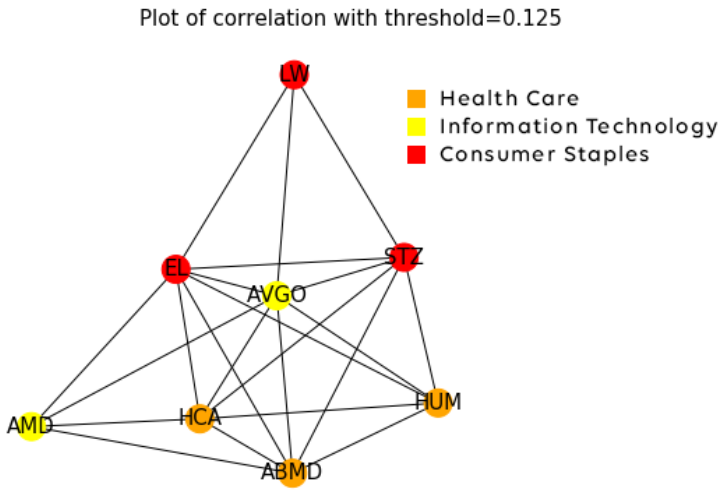

Figure 28: Graph of correlation between the chosen stocks with a threshold = 0.125.

Here, all the edges mean that there is a correlation higher than 0.125 between the two connected stocks. Note that the closer the nodes are together, the more strongly correlated they are. This way the strongly correlated stocks will form a cluster. The length of the edges is proportional to (1 - correlation). From this graph, it is observed that each of the different sectors form a cluster. The stock, *LW* is not very strongly correlated to any other stock as it is quite far from the other stocks. The stock *AVGO* on the other hand is strongly correlated strongly correlated to stocks which are not within the same sector, particularly *EL* and *STZ*. With this graph, we can have an understanding of how the dynamics of one of the stock affect the other stocks.

3. The topology of the graph is mainly dictated by the how the edges are defined which in this case is the correlation between each nodes. The position of each node and the orientation of the graph might differ. However, the length of each edges is dictated by the magnitude of the correlation between the nodes. The edges are formed when the correlation is larger than 0.125, in which the threshold can be adjusted. Hence, the reordering of the vertices will not affect the results as the relationship between each nodes still hold.

If we reorder the time-series of all the stocks in the same order, the resulting graph will still be the same as the correlation between each of the stocks will remain the same. If we however, only shuffle the time-series of only one or some of the stocks, the graph will be different. This can be seen by looking at the definition of the Pearson correlation function. The Pearson correlation coefficient of two random variables, x and y, is define as:

$$\rho(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{4}$$

From the equation, it is seen that as long as all pairs of x_t and y_t remain the same, the correlation between the two stocks will remain unchanged. Reordering the time-series of just some of the stocks will change the pairs of x_t and y_t, hence changing the correlation between the stocks which in turn change the topology of the graph.

4. We choose the mean absolute distance as the metric of choice. The formula for the mean absolute distance (MAD) is given below:

$$d(x,y) = \frac{1}{N}\sum_{i=1}^{N}|x_i - y_i| \tag{5}$$

Note that the Pearson's correlation is a shape based measure between two time-series. This means that it measures the overall difference across the entire time-series. Intuitively, the Pearson's correlation can be thought of as if the two time-series move up or down (have the same trend) at the same time. It completely neglects the difference between each individual time point. Another way to look at Pearson's correlation is how well the two time-series can be modelled as a linear model i.e.

$$A = mB + c$$

where A and B are the two time-series while m and c are arbitrary constants.

The MAD on the other hand is a point-based analysis where intuitively, it looks at the difference between every time point of the two time-series. It does not care about the overall shape of the two time-series and does not consider them as a whole. As it is so different from the Pearson's correlation measure, it is able to provide insights and intuitions regarding the stocks which we would otherwise lack. To plot the MAD graph, we again calculate the MAD matrix of the stocks and look at the distribution of the MAD between different stocks:

Stock	**ABMD**	**HCA**	**HUM**	**AMD**	**AVGO**	**EL**	**LW**	**STZ**
ABMD	0.000000	0.000857	0.000920	0.002039	0.000916	0.000821	0.000848	0.000804
HCA	0.000857	0.000000	0.000355	0.001677	0.000493	0.000351	0.000325	0.000320
HUM	0.000920	0.000355	0.000000	0.001802	0.000554	0.000381	0.000328	0.000364
AMD	0.002039	0.001677	0.001802	0.000000	0.001617	0.001619	0.001680	0.001725
AVGO	0.000916	0.000493	0.000554	0.001617	0.000000	0.000459	0.000464	0.000460
EL	0.000821	0.000351	0.000381	0.001619	0.000459	0.000000	0.000240	0.000269
LW	0.000848	0.000325	0.000328	0.001680	0.000464	0.000240	0.000000	0.000223
STZ	0.000804	0.000320	0.000364	0.001725	0.000460	0.000269	0.000223	0.000000

Table 29: MAD matrix of the eight chosen stocks.

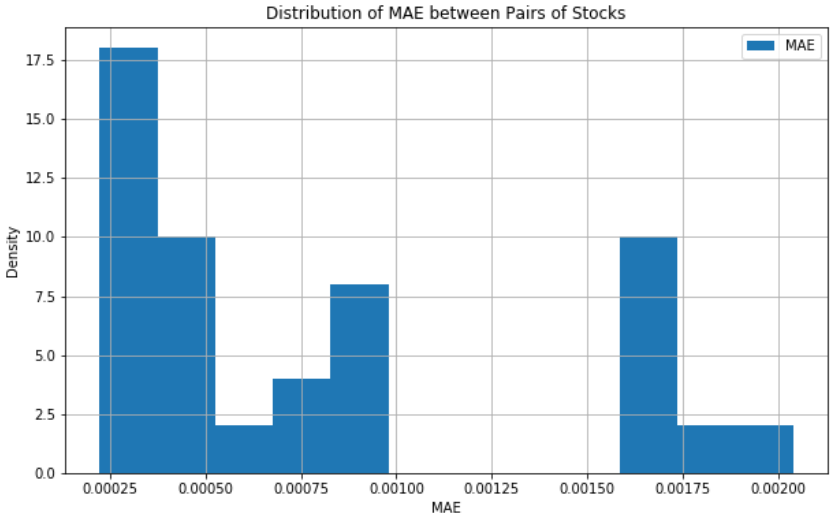

Figure 29: Distribution of MAD between pairs of selected stock.

Note that the smaller the MAD, the more similar the two stocks are. Hence we only would like to analyse pairs of stocks with low MAD. From Figure 29, we see that there is a spike in MAD at threshold < 0.0005 and we choose this value to plot the graph.

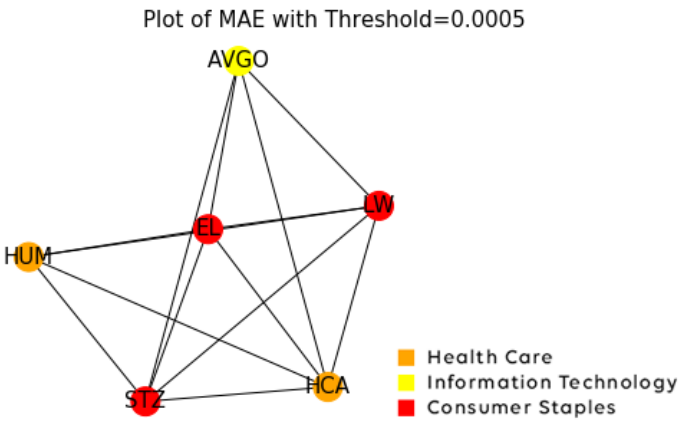

Figure 30: MAD graph of chosen stocks, threshold = 0.0005.

Since the graph is slightly congested, we choose a a spike which is slightly lower than 0.0005. The next spike is observed at about threshold < 0.000375. This graph is plotted below:

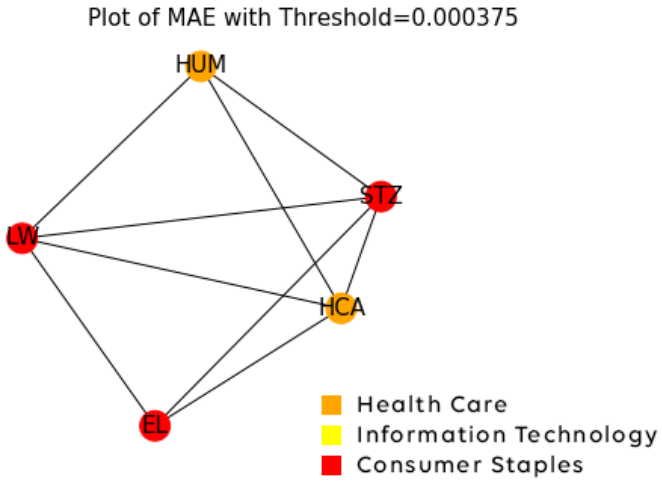

Figure 31: MAD graph of chosen stocks, threshold = 0.000375.

Note that again, the closer the stocks are in the graph, the more "similar" they are. Here we observe that the stocks STZ and HCA are similar to quite a number of other stocks and that the "similarities" are quite strong. Hence, if we were to further diversity the risk of our portfolio, we should consider replacing the STZ and HCA stocks with some other stocks which are not as similar to the other stocks in which we have chosen.

5. The correlation matrix of the chosen stocks are given below. It is observed that the correlation between the raw prices of the stocks are a higher than of their log returns. This is due to the fact that raw-prices are non-stationary and that for any statistics (including correlation) to make sense, the underlying data has to be stationary. Besides that, prices at different time-step are not independent of each other i.e. prices between adjacent days are highly correlated to each other. Returns on the other hand have equal importance at every time-step. Hence we conclude that using the raw price of the stocks for any financial strategy is not recommended.

Stock	ABMD	HCA	HUM	AMD	AVGO	EL	LW	STZ
ABMD	1.000000	0.852377	0.935828	0.822597	0.700428	0.922605	0.907558	0.819681
HCA	0.852377	1.000000	0.859229	0.801239	0.515945	0.757530	0.877479	0.563938
HUM	0.935828	0.859229	1.000000	0.881000	0.802697	0.895960	0.940116	0.838363
AMD	0.822597	0.801239	0.881000	1.000000	0.787343	0.690949	0.628246	0.753307
AVGO	0.700428	0.515945	0.802697	0.787343	1.000000	0.739201	0.377021	0.911940
EL	0.922605	0.757530	0.895960	0.690949	0.739201	1.000000	0.875080	0.862433
LW	0.907558	0.877479	0.940116	0.628246	0.377021	0.875080	1.000000	0.662055
STZ	0.819681	0.563938	0.838363	0.753307	0.911940	0.862433	0.662055	1.000000

Table 30: Correlation matrix of the raw prices of the eight chosen stocks.

References

Galarnyk, M. (2018), 'Understanding boxplots'.URL:https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51