# Visual Analytics - Similarities Between Equities and Financial Time Series Forecasting Optimization

Axil Sudra | Visual Analytics | The school of Mathematics, Computer Science and Engineering
City, University of London

## 1. Motivation, Data and Research Questions

### Motivation

Following the recent significant economic events such as Brexit, investors' interests in examining the operations of global corporations have intensified. The search for similarities between publicly traded equities has become a key objective when investing; for example, an investor that leverages this information can hedge risk by selling (short position) a similar equity to the equity or equities they hold if they are declining in value. Another key objective when investing, perhaps focused more towards institutional investors than private investors, is forecasting the future price of an equity or equities. Forecasting is utilised for numerous purposes; for example, an investor might forecast to analyse the effects of expected earnings declarations on the price of an equity.

Given the investing objectives above, this study will aim to explore the application of state-of-the-art visual analytics to identify similarities in equities listed on the Dow Jones Industrial Average 30 (DJIA30), and effectively build a forecasting model for equity pricing.

### Data

To analyse equities listed on the DJIA30, multiple time series datasets were retrieved from 'Yahoo Finance' using an API call from [1] and saved to comma-separated value (csv) files. Note that the data for each DJIA30 equity contains the following attributes:

1. Date – each dataset contained daily data from the beginning of 2015 to the end of 2017; this amounts to 756 rows.
2. High – the maximum price of an equity on a given date denoted in dollars ($).
3. Low – the minimum price of an equity on a given date denoted in dollars ($).
4. Open – the opening price of an equity on a given date denoted in dollars ($).
5. Close – the closing price of an equity on a given date denoted in dollars ($).
6. Adjusted Close – the closing price of an equity that accounts for any corporate activity during non-trading periods; the adjusted closing price is calculated at the beginning of the next trading day.
7. Volume – the trading volume of an equity on a given date.

Each of the attributes listed above will be used to investigate the investing objectives mentioned in the previous section (Motivation) as these attributes are fundamental when analysing the characteristics of equities. The log transformation of pricing data of each equity was applied for various experiments to scale the difference in the magnitude of each DJIA30 equity.

### Research Questions

The aim of this study is to investigate techniques to identify similarities between equities listed on the DJIA30 using visual analytics. Furthermore, experiment with building an equity pricing forecast model and using visual analytics to evaluate performance. Thus, suitable research questions for this study are:

- Which DJIA30 equities were considerably similar in quantitative behaviour between 2015 and 2018?
- How can we forecast the price of an equity? What methods can be used to tune the forecasting model and evaluate performance?

## 2. Tasks and Approach

Several libraries in **Python** were utilised to complete data processing, data analysis and visualisation. These have been outlined in the text where necessary.

### Data Retrieval and Data

To commence the study of our research questions, the datasets for equities listed on the DJIA30 (and the DJIA30 index dataset) were retrieved using an API call loop from [1] and stored as individual data frames; this was accomplished using libraries 'pandas_datareader' and 'pandas' respectively. Each dataset was inspected for missing values, formatted where necessary (i.e. dates were converted to date-time format using the package 'datetime') and visualised. Fortunately, most of the datasets retrieved were in ideal shape. Note that the retrieval and wrangling of data was fundamentally a preliminary task rather than an analytical task.

Data Exploration

For an initial overview and a ‘non-technical’ exploration of our first research question detailed in section 1, the trend (over a 3-year period – 2015 to 2018) of the DJIA30 index and various equities listed on the index were visualised to distinguish any obvious similarities. Visualisations in this study were produced using the interactive library ‘plotly’ [2]. Due to the challenging nature of displaying multiple financial time series as described by Ziegler et al. [3], the average annualised returns (and annualised volatility) for each equity is computed and visualised on a scatter plot to reveal any clear performance similarities.

Multidimensional Scaling Similarity Identification

To further explore our first research question, the feature neutral (i.e. results are not dependent on a specific dataset) technique of multidimensional scaling (MDS) [4] will be adopted to model the similarity between equities listed on the DJIA30. New features such as annualised volatility, average adjusted closing price and average volume will be computed for each equity and grouped with annualised returns. These features will be used as input into distancing algorithms. These algorithms will be parameters to the MDS models. Note that the ‘euclidean_distances’ and ‘manhattan_distances’ functions from the ‘sklearn’ library will be used and compared. The distances of each equity will be displayed using a ‘plotly’ heatmap. Once these distances have been computed, they will be put through the MDS function in ‘sklearn’ which will give 3-dimensional space coordinates for each equity; these will be visualised using a scatter plot. Both Euclidean and Manhattan MDS models will be evaluated using the raw stress metric [5].

Equity Price Forecasting Model

Here we will implement a widely used statistical forecasting model called Autoregressive Integrated Moving Average (ARIMA) [6] to forecast the adjusted closing price of Visa (V) equity. Note that the ‘statsmodels’ library will be imported to use the ARIMA function. Before forecasting the price of Visa equity, we will conduct a random search to optimise the ARIMA model’s hyperparameters (p – the model order (i.e. number of time lags in the model), d – the degree of differencing and q – period of moving average) and visualise the best (optimal) combination. Each hyperparameter combination will be used as input to a ARIMA model and the mean squared error(MSE) for each model will be noted to identify the best hyperparameters.

## 3. Analytical Steps

Data Exploration – Trend Visual Analysis

To explore our first analytical task detailed in section 2, we can visually examine the trend movements of the DJIA30 index and some of its listed equities between 2015 and 2018 (as shown in figure 1).

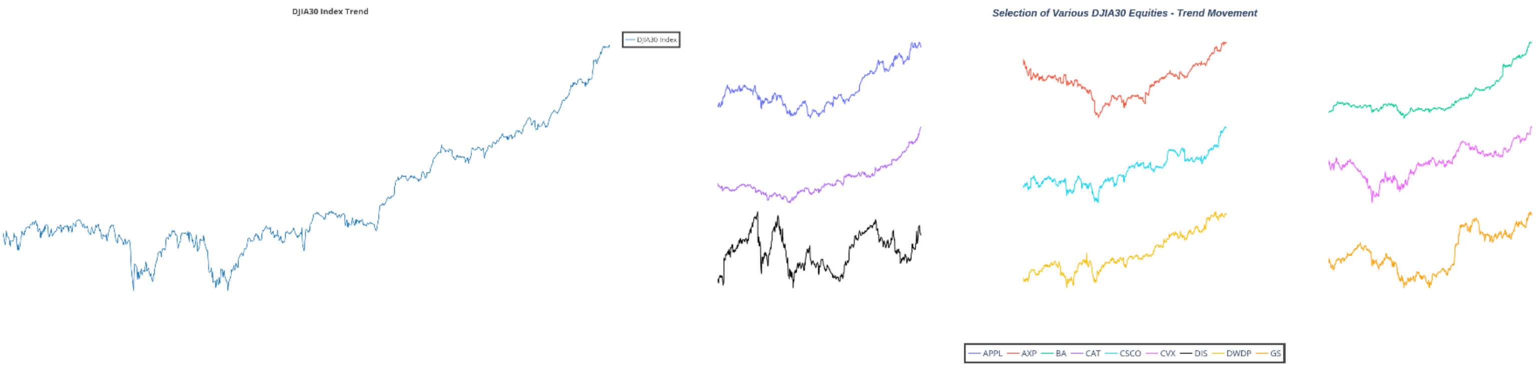
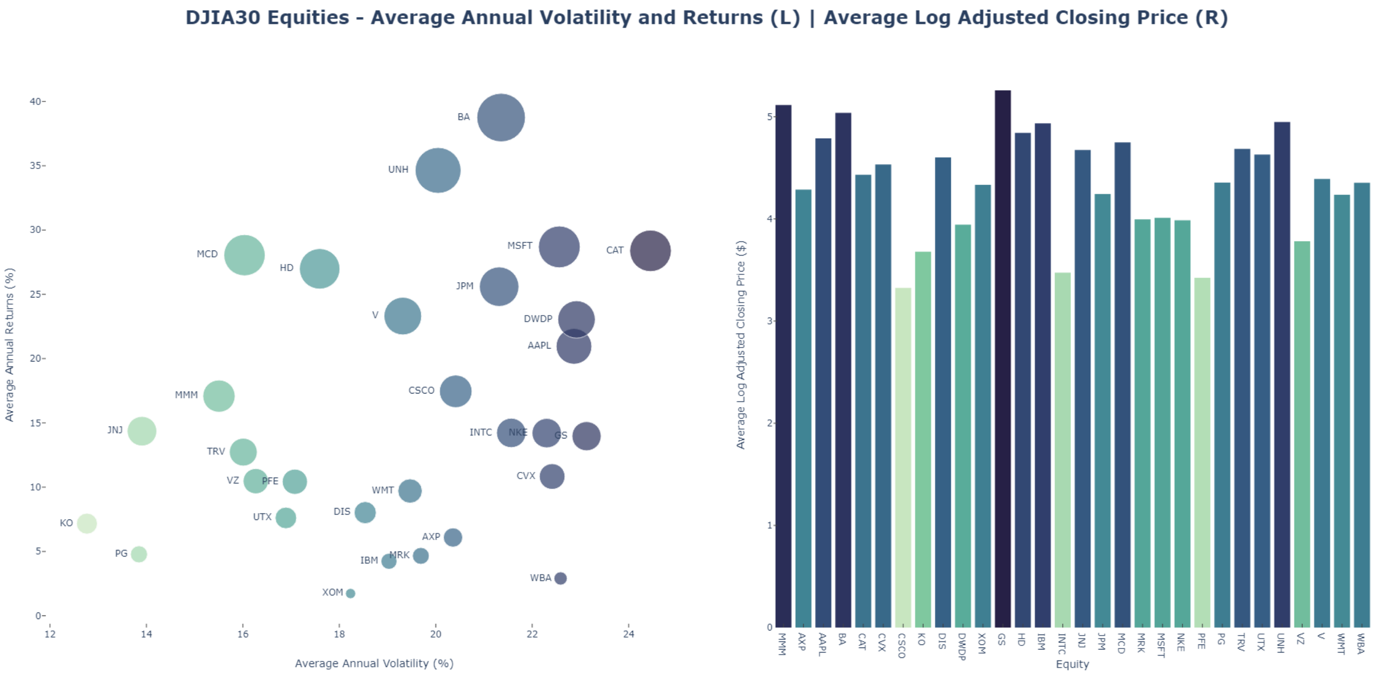


*Figure 1: DJIA30 index trend (2015 to 2018) (L) and various DJIA30 listed equities’ trends (2015 to 2018) (R).*

The general trend amongst the displayed equities in figure 1 seem to follow the movement of the DJIA30 index, although equities such as Disney (DIS) and Goldman Sachs (GS) appear to be more unpredictable in behaviour. Therefore, we could draw to the conclusion that most equities in the DJIA30 are similar by interpreting their trend movements, although this doesn’t seem practicable. Due to the challenging task of visually interpreting multiple financial time-series at once, we can only analyse a handful of equities listed on the DJIA30 index at one time which is not ideal.

Data Exploration – Annualised Returns

For a more suitable means of analysing the behaviour and similarities of equities listed on the DJIA30, we can calculate their respective average annual returns and volatility and visualise them on a scatter plot (as shown in figure 2).



DJIA30 Equities - Average Annual Volatility and Returns (L) | Average Log Adjusted Closing Price (R)

*Figure 2: DJIA30 listed equities' average annual returns and volatility (L) and average log adjusted close prices ($) (R).*

Equities listed on the DJIA30 with high average annual returns are denoted by large points on the scatter plot in figure 2; equities with high annual volatility are denoted by dark shaded points than equities with low annual volatility (which are shaded lightly). From figure 2, the equity Boeing (BA) has the greatest average annual return and the equity Caterpillar (CAT) is the most volatile. Apple (AAPL) and Dow Du Pont (DWDP) appear to be the 'most' similar in respect to return and volatility characteristics.

The bar chart on the right in figure 2 shows the average 3-year average log adjusted closing price ($) for each equity listed on the DJIA30; the shading of each bar corresponds to the value of the 3-year average log adjusted closing price ($) (the darker the shade, the higher the value). From the bar chart, we deduce that Goldman Sachs (GS) has the highest 3-year average log adjusted closing price ($). In contrast, Cisco Systems (CSCO) has the lowest 3-year average log adjusted closing price ($).

Multidimensional Scaling Similarity Identification – Distance Measurements

Identifying behaviour similarities between equities using annualised computations of returns and volatility is a simple, although effective method for private investors with little quantitative ability. However, the method only utilises one feature in our datasets, namely the equity's adjusted closing price.

Through the application of multidimensional scaling (MDS), we can take advantage of most (if not all) of the features in our datasets. In a nutshell, "MDS techniques develop spatial representations of psychological stimuli or other complex objects about which people make judgements (e.g. preference, (dis)similarity)" [7]. MDS allocates each object (in our case, equities listed on the DJIA30) a point in a specified dimension; however, before allocating points to each object, the MDS model requires distances between these objects. In this study we compute the Euclidean distance and Manhattan distance between each equity listed on the DJIA30 using various features in our dataset (average annual returns, annual volatility, 3-year average adjusted closing price and 3-year average volume). The formulas for Euclidean and Manhattan distances are given below [4]:

$$\text{Euclidean Distance} = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \mid \text{Manhattan Distance} = |x_1 - x_2| + |y_1 - y_2|$$

The Euclidean distances and Manhattan distances of equities listed on the DJIA30 have been visualised using 'plotly' heatmaps (as shown in figure 3). Note that small distances between any two equities correspond to high correlative behaviour [7].

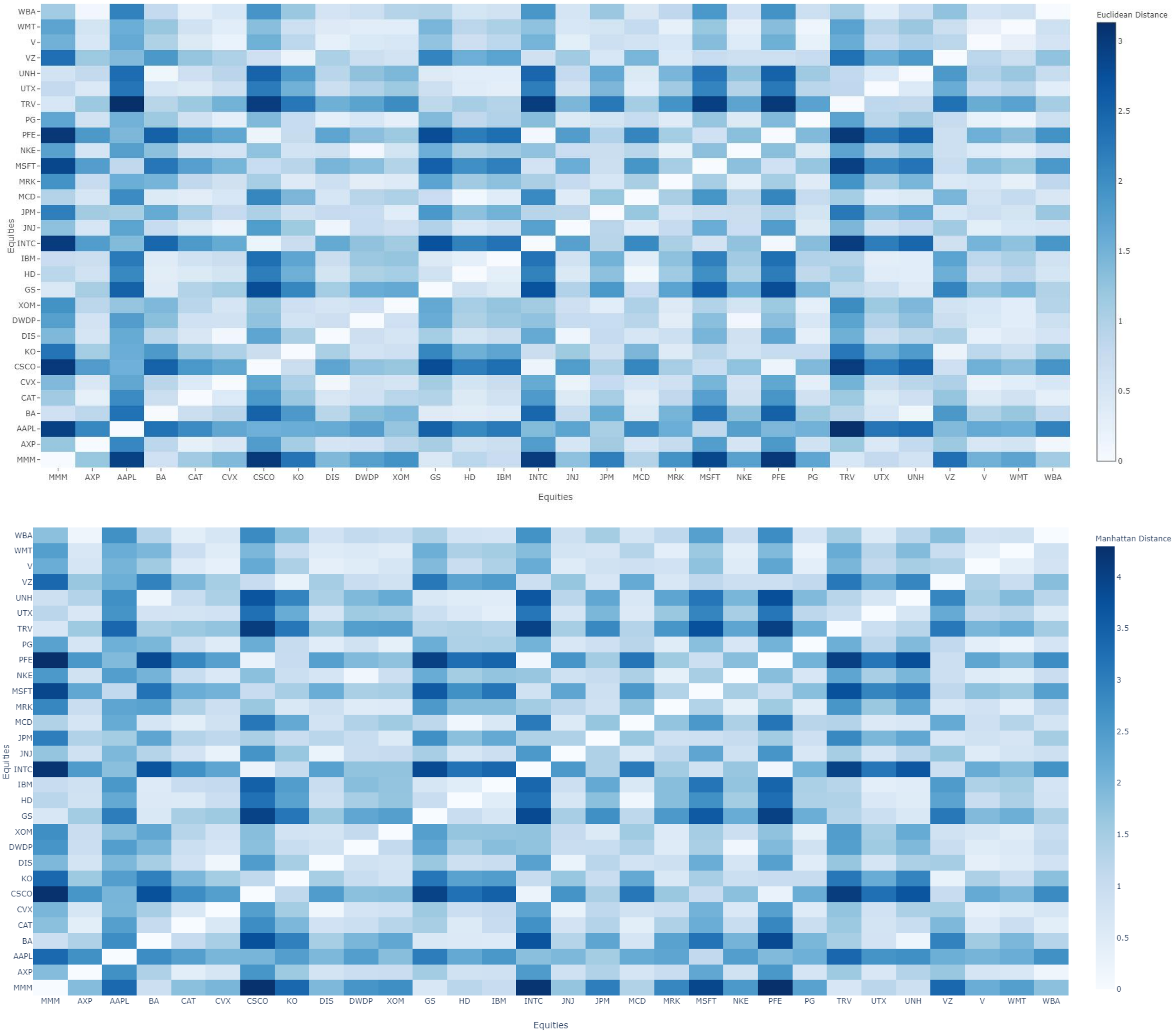


*Figure 3: Euclidean distance (T) and Manhattan distance (B) heatmaps of DJIA30 equities.*

Through the computation of the minimum value in each heatmap (that is greater than 0), we find that the Euclidean distance suggests that Intel (INTC) and Pfizer (PFE) are the 'closest' in terms of distance (distance of 0.08358 5.d.p.). In contrast, the Manhattan distance suggests that Dow Du Pont (DWDP) and Nike (NKE) are the 'closest' (distance of 0.15271 5.d.p.).

<u>Multidimensional Scaling Similarity Identification – Application and Visualisation</u>

By obtaining the corresponding distance matrices (Euclidean and Manhattan), we can now use MDS to achieve dimension reduction of the distance matrices for distances between equities listed on the DJIA30 (as shown in figure 4). Note that the axes metrics are unknown in MDS and we have labelled them 'distance'. To evaluate the 'goodness-of-fit' of each MDS model, we have computed the raw stress scores for each distance algorithm. These are given below:

Euclidean distance raw stress score = 0.31127 (5.d.p.)
Manhattan distance raw stress score = 3.02671 (5.d.p.)

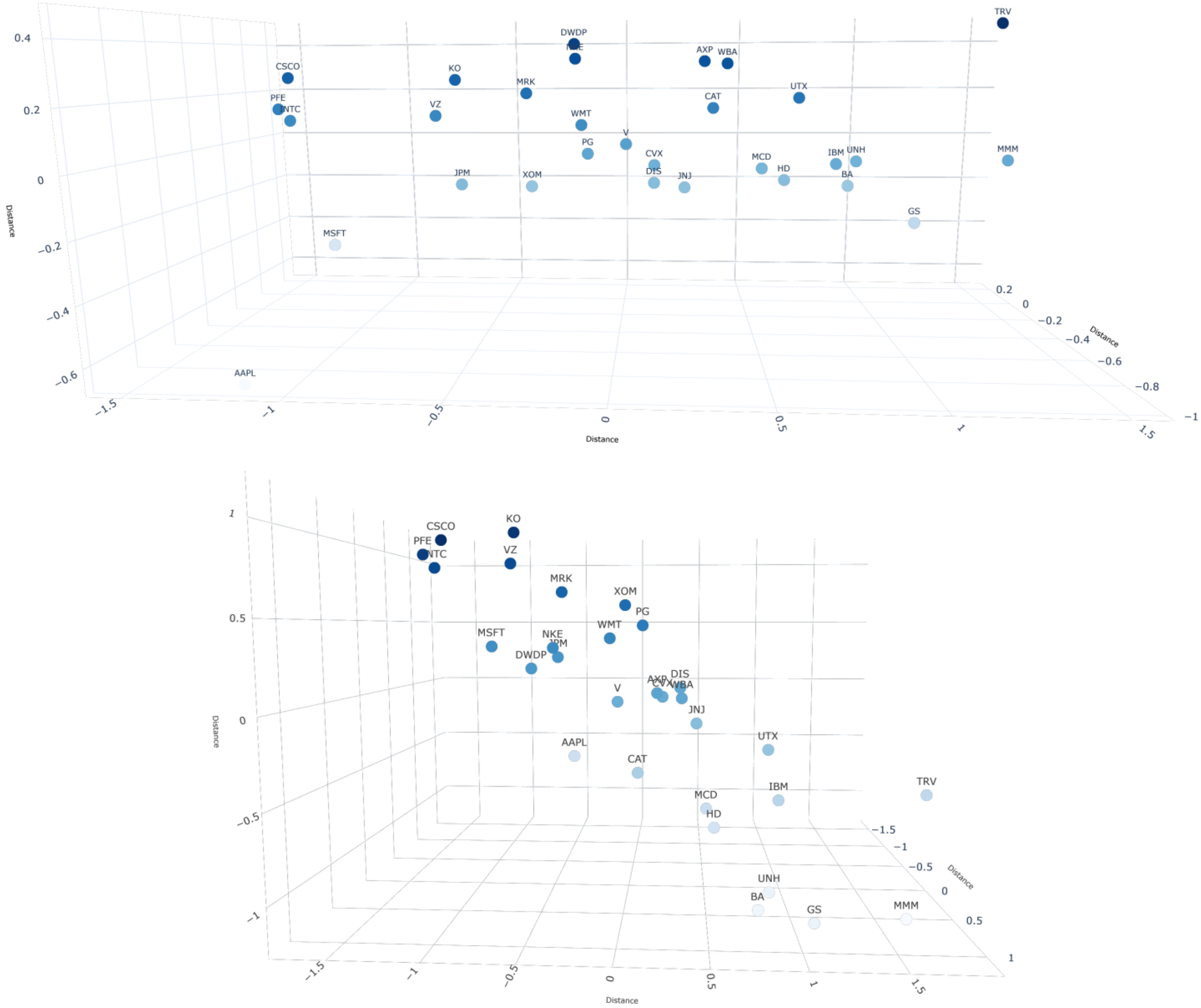


*Figure 4: MDS scatter plots of Euclidean (T) and Manhattan (B) distances.*

### Equity Price Forecasting Model

To explore our final analytical task outlined in section 2, we construct an Autoregressive Integrated Moving Average (ARIMA) model using the dataset for the equity Visa (V). Alternative forecasting models do exist [6]; however, due to economic, political etc. factors, financial time series may exhibit non-stationary behaviour (i.e. they follow trends) and thus we use the ARIMA model to make the time series stationary with a ‘differencing’ (d) hyperparameter. Using the ‘statsmodels’ and ‘random’ libraries, we construct a random search to obtain an optimal set of hyperparameters (p, d, q) for the model. The random grid search is shown below in figure 5.

The random grid search has been displayed on a 3-dimensional scatter plot, where each axis corresponds to either p, d or q. The mean squared error (MSE) for each model has been denoted by a colourmap; the lower the MSE, the darker the point’s colour. Each point has been labelled with it’s corresponding MSE to make it easier to identify the best (optimal) hyperparameters. Through inspection of figure 5 and confirmation from the Python coding script, we find that the best (optimal) model has the hyperparameter values p = 2, d = 1 and q = 1 and MSE of 0.60887 (5.d.p.).

The final step to forecasting the price of the equity Visa (V) is to construct an ARIMA model with the best (optimal) hyperparameters found during the random search. Note that the dataset has been split into a training and test set; 70% of observations are contained in the training set and the remaining observations are contained in the test set. The results of the prediction are shown in figure 6. The blue line in figure 6 denoted data contained in the training set, the red line denotes data in the test set and the black dotted line denotes the ARIMA model's prediction. The deviations between test and predicted values are shown in the bar chart below the line chart.

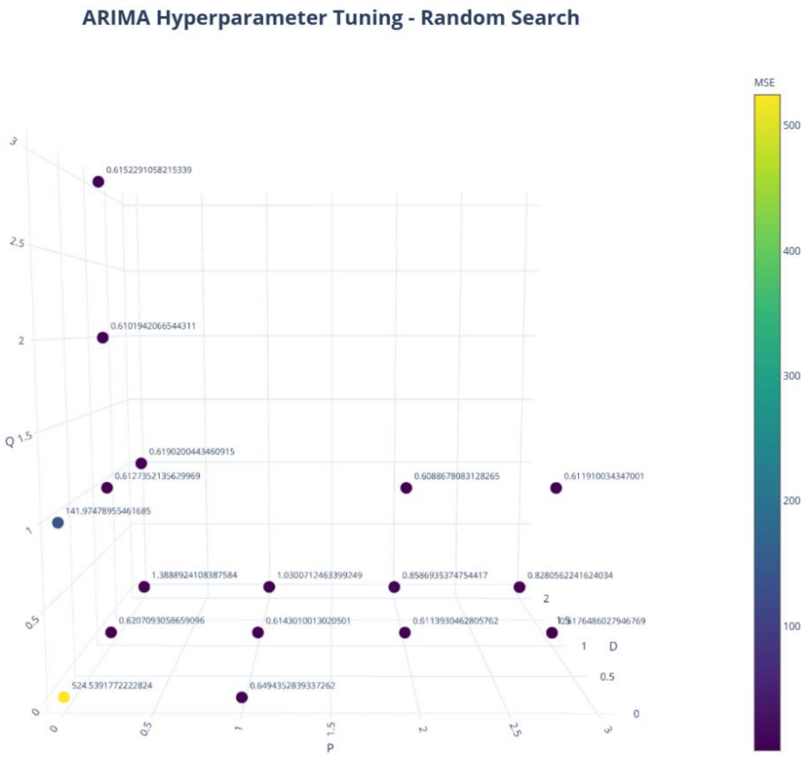


*Figure 5: Results of hyperparameter random grid search for ARIMA model.*

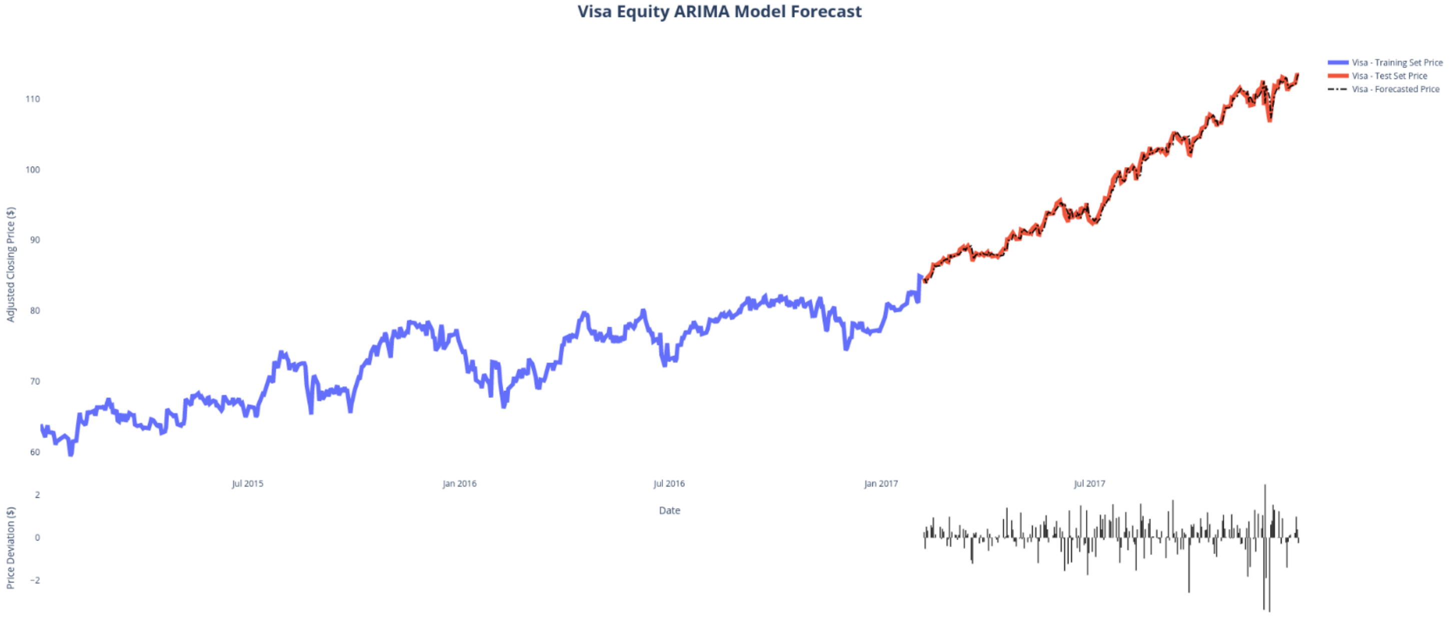


*Figure 6: Application of optimal hyperparameters to Visa (V) equity.*

## 4. Findings

Through the creation of an annual average returns and volatility scatter plot (shown in figure 2), we have identified that the equities Apple (AAPL) and Dow Du Pont (DWDP) are highly similar. However, using annual average returns and volatility as a measurement of similarity is not ideal due to only factoring two features (returns and volatility) of an equity. A more analytically rigorous method of measuring similarity between equities listed on the DJIA30 was using various distance algorithms (Euclidean and Manhattan) to find an 'unknown' dimensional relationship and then applying the data to an MDS model. The application of an MDS model allowed us to perform dimension reduction on the distance matrices and visualisation of the results.

From the MDS scatter plots (figure 4), we revealed that Intel (INTC) and Pfizer (PFE) had the greatest similarities for the Euclidean distance algorithm; Dow Du Pont (DWDP) and Nike (NKE) had the greatest similarities for the Manhattan distance algorithm. To evaluate the performance of the MDS models (goodness-of-fit), the raw stress score was computed. The Euclidean and Manhattan MDS model scored 0.31127 (5.d.p.) and 3.02671 (5.d.p.) respectively. Krushal [5] states that the following boundaries are a good indication of how to judge raw stress scoring:

| Raw Stress Score | Goodness-of-Fit |
|---|---|
| 0.200 | Poor |
| 0.100 | Fair |
| 0.050 | Good |
| 0.025 | Excellent |
| 0.000 | Perfect |

This table suggests that both of our MDS models (Euclidean and Manhattan) performed poorly; the Manhattan distance MDS model performed significantly poorly with a raw stress score greater than 1!

To explore our second analytical question, we choose the ARIMA model as our forecasting tool of choice. To begin, a random search was conducted on multiple ARIMA models with a combination of different hyperparameters (p, d, q). Note that the random search fitted 16 models. The following hyperparameter combination was the best (optimal):

p – model order = 2
d – differencing = 1
q – moving average = 1

These were identified using a 3-dimensional scatter plot of the hyperparameter random search (and confirmation from the Python coding script). This model had an evaluation MSE of 0.60887 (5.d.p.), which is a significantly close (good) fit. Applying the ARIMA model to the Visa (V) equity showed perfect results (as shown in figure 6) as highlighted by the small MSE.

## 5. Critical Reflection

Visual Analytics Approaches – Answering Research Questions

This study aimed to explore the research questions outlined in section 1 with feasible techniques and experiments. To investigate which DJIA30 equities were similar in quantitative behaviour between 2015 and 2018, the of average annual return and volatility computations were visualised using a scatter plot with 'feature defining' properties; the colour shade and size of each point was used to give investors an idea of the level of annual volatility and average annual returns a specific equity has endured respectively (i.e. the darker the shade of the point, the more volatile and the bigger the point, the higher the returns). This method was feasible, although it doesn't use any additional information about an equity, and thus measures similarity in terms of returns and volatility only.

Using different distance measures (Euclidean and Manhattan) and the application of an MDS model, we made use of additional features to make similarity results more valid. Both distance MDS models produced different results (i.e. one model suggested two equities are similar while the other model suggested another two different equities are similar). Evaluating these models with the use of a raw stress score, showed that each model performed badly, although the Manhattan was significantly poor. For future similarity experiments, using more features (such as P/E ratios of a company or market capitalization) could potentially improve model performance.

To investigate how we can forecast the price of an equity, we deployed the ARIMA model onto the Visa (V) equity. Through the visualisation of a random search to identify the best (optimal) hyperparameters, we were able to assess model performance with various hyperparameter combinations.

Implications of Findings for Chosen Application/Domain

The process of investing in investment securities, whether equities or any other asset class, and constructing a financial portfolio is challenging for investors [8]. The results produced by the application of Euclidean and Manhattan distancing on equities listed on the DJIA30 to essentially quantify similarity between equities was surprising. These distancing algorithms suggested that two equities (Intel and Pfizer for Euclidean and Nike and Dow Du Pont for Manhattan) from completely different industries were the most similar; these results are clearly distorted and wouldn't have any fundamental implications for the investment management industry or private investors. However, with more robust and granular equity data (i.e. earnings, leverage ratios, etc.) this method of identifying similarities in equity markets could be successful.

Regarding the implications of our ARIMA forecasting experiment, we used methods (i.e. random search to optimise hyperparameters) that are already active in the investment management industry today, and in financial series in general. However, the visualisation of the prediction against the test values seem to good to be true with an MSE of roughly 0.6.

### Generalisability to Other Domains

The visual and computational techniques used in this study could potentially be applied to several domains (e.g. health, engineering, climate assessment etc.). For example, the use of distancing algorithms and MDS could be applied to detecting if cancerous cells are spreading in a patient's body. Furthermore, the ARIMA forecasting technique and hyperparameter optimisation visualisation could and is used in various domains today (e.g. climate forecasting [9]).

## References

[1] Yahoo Finance: official site providing free equity quotes, the latest news, portfolio management resources, international market data and other financial resources. API call to DJIA30 equities and index data. Available: https://uk.finance.yahoo.com/ (Accessed: December 2018).

[2] VanderPlas, J. (2016) 'Visualisation with Matplotlib: Further Resources', in VanderPlas, J. (ed.) *Python Data Science Handbook.* Sebastopol CA: O'Reilly Media, Inc., pp. 329-330.

[3] Ziegler, H. et al. (2010) 'Visual Market Sector Analysis for Financial Time Series Data', *2010 IEEE Symposium on Visual Analytics Science and Technology*, pp. 83-90.

[4] Machado, J.T. et al. (2011) 'Analysis of Stock Market Indices Through Multidimensional Scaling', *Communications in Nonlinear Science and Numerical Simulation*, 16(12), pp. 4610-4618.

[5] Krushal, J. (1964) 'Multidimensional Scaling by Optimizing Goodness of Fit to a Nonmetric Hypothesis', *Psychometrika*, 29(1), pp. 1-27.

[6] Bögl, M. et al. (2013) 'Visual Analytics for Model Selection in Time Series Analysis', *IEEE transactions on visualization and computer graphics*, 19(12), pp. 2237-2246.

[7] Esmalifalak, H., et al. (2015) '(Dis)integration Levels Across Global Stock Markets: A Multidimensional Scaling and Cluster Analysis', *Expert Systems with Applications*, 42(22), pp. 8393-8402.

[8] Savikhin, A., et al. (2011) 'An Experimental Study of Financial Portfolio Selection with Visual Analytics for Decision Support', *2011 44th Hawaii International Conference on System Sciences*, pp. 1-10.

[9] Robenson, S.M. et al. (1990) 'Evaluation and Comparison of Statistical Forecast Models for Daily Maximum Ozone Concentration', *Atmospheric Environment. Part B. Urban Atmosphere*, 24(2), pp. 303-312.