Nonlinear Factor Models in Financial Markets: A Comparative Study of Autoencoders and PCA

Jason Yi
730604615
jasonyi@unc.edu

Cole Beitscher
730652572
cbei@unc.edu

Rishabh Dey
730558424
rdey@unc.edu

Rohan Kashyap
730468430
rkashyap@unc.edu

Abstract

This project investigates financial return data using multiple data sources, including the Yahoo Finance API, Alpha Vantage API, and Polygon.io. We focus on standardized asset returns from the S&P 500 index, NASDAQ Composite index, and a portfolio of large-cap technology stocks. Exploratory analysis includes summary statistics, distributional analysis, and anomaly detection via traditional methods (IQR, Z-Score) and machine learning methods (Isolation Forest). These analyses highlight the presence of heavy tails, volatility clustering, and rare extreme events in market returns. In the next phase, we aim to apply dimensionality reduction techniques, including Principal Component Analysis (PCA) and neural autoencoders, to extract latent factors and perform clustering and anomaly detection with improved sensitivity to nonlinear structure.

GitHub Repository: `https://github.com/yi-json/comp562-final-project`

1 Introduction

Understanding the latent patterns in asset returns is a fundamental problem in financial modeling, with direct applications to risk management, portfolio optimization, and anomaly detection. Financial markets are influenced by a wide variety of economic, political, and behavioral factors, and identifying structure in asset returns can provide crucial insights for both theoretical and applied finance.

In this project, we study a multivariate dataset of U.S. equity returns, aiming to uncover hidden regularities and detect unusual patterns that may suggest structural shifts or market anomalies. The high dimensionality and inherent noise in financial data make unsupervised learning methods particularly appealing, as they can extract patterns without requiring labeled data.

To work towards this goal, we compare the effectiveness of various unsupervised learning techniques in revealing the underlying structure in U.S. equities. Specifically, we analyze the performance of Principal Component Analysis (PCA), non-linear PCA, K-means clustering, and autoencoders. These methods are used to reduce dimensionality and group assets based on latent factors that may influence return co-movement.

Additionally, we investigate anomaly detection by applying the Isolation Forest algorithm to identify outliers in the return distributions. We compare its performance against traditional univariate techniques, such as z-score filtering and the interquartile range (IQR) method, to assess whether the nuanced machine learning approach offers a more insightful understanding of financial anomalies.

Through this analysis, we aim to evaluate whether modern unsupervised learning methods can supplement or even outperform traditional statistical tools in understanding complex financial systems.

Preprint. Work in progress.

2 Related Work

Dimensionality reduction has long been used in finance to extract latent factors that explain asset returns, with Principal Component Analysis (PCA) being the most classical approach (Connor & Korajczyk, 1986). PCA linearly decomposes returns into orthogonal components that explain maximum variance, and has applications in risk modeling, portfolio construction, and factor investing.

However, PCA's linearity limits its ability to capture complex patterns in financial data. To address this, neural network-based autoencoders have been proposed as nonlinear generalizations of PCA (Hinton & Salakhutdinov, 2006). Autoencoders learn compressed representations through bottleneck architectures and have shown promise in quantitative trading, market regime detection, and latent factor extraction.

Beyond factor modeling, unsupervised learning methods like K-Means clustering have been used to group assets based on latent features. Studies have demonstrated that K-Means can enhance portfolio diversification by grouping stocks with similar return behavior (Nanda et al., 2010; Baser & Saini, 2019). However, K-Means struggles when the underlying structure is highly nonlinear.

Other work explores variations of PCA to better capture structural nuances, such as combining functional data analysis with PCA (Khoo et al., 2017) or applying autoencoders directly to trading data for latent feature extraction (Choi et al., 2020).

In this project, we build on these ideas by comparing PCA and autoencoder-derived latent factors, evaluating their effectiveness for return variance explanation, clustering, and anomaly detection. Our results highlight the strengths of autoencoders in modeling nonlinear relationships, particularly as model complexity increases.

3 Methodologies

3.1 Data Collection and Preprocessing

We collected historical financial data from Yahoo Finance, Alpha Vantage, and Polygon.io. From Yahoo Finance, we downloaded adjusted daily close prices for the S&P 500 index (`ĜSPC`) from 2000 to 2025 and computed daily log returns. Tech stock data (META, AMZN, AAPL, NFLX, GOOG, MSFT, NVDA, TSLA) was obtained from the Alpha Vantage `TIME_SERIES_DAILY_ADJUSTED` API, with returns calculated using adjusted close prices. Due to rate limits, only a subset of symbols was processed.

NASDAQ Composite index data was retrieved from Polygon.io and similarly transformed into daily log returns. All return series were standardized using z-score normalization (`StandardScaler` from `scikit-learn`) to prepare for outlier detection and dimensionality reduction analysis.

3.2 Anomaly Detection Techniques

3.2.1 Traditional Methods

We first implemented classical statistical methods for anomaly detection, including Z-score and Interquartile Range (IQR).

The **Z-score** measures how many standard deviations a data point x is from the mean μ:

$$Z = \frac{x - \mu}{\sigma}$$

Observations with $|Z| > 3$ are typically flagged as outliers under the assumption of normality.

The **Interquartile Range (IQR)** method flags values outside the range:

$$[\text{Q1} - 1.5 \cdot \text{IQR}, \quad \text{Q3} + 1.5 \cdot \text{IQR}]$$

where IQR $=$ Q3 $-$ Q1. This method is robust to skewed distributions but still fails to capture multivariate structure.

3.2.2 Machine Learning Methods

To address the limitations of univariate approaches, we implemented **Isolation Forests**, a tree-based ensemble method that identifies anomalies by recursively partitioning data and measuring the path length required to isolate a point. Isolation Forests do not rely on assumptions of normality and are effective in high-dimensional settings. They also implicitly account for correlations between features, enabling more reliable detection of multivariate anomalies.

3.3 Dimensionality Reduction for Unsupervised Learning

3.3.1 Limitations of PCA

Traditional **Principal Component Analysis (PCA)** identifies directions of maximum variance in the data via eigenvalue decomposition of the covariance matrix. However, it is inherently linear and sensitive to noise and outliers, making it less suitable for financial data with nonlinear dependencies.

We evaluate PCA and its alternatives using two metrics: the coefficient of determination (R^2) and the Mean Squared Error (MSE).

The **coefficient of determination** is defined as

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}.$$

where y_i is the actual return, $\hat{y}_i$ is the predicted return from the model, and $\bar{y}$ is the mean of the actual returns.

The **Mean Squared Error (MSE)** is calculated as

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

These metrics allow for a direct comparison between PCA and autoencoder-based reconstructions.

3.3.2 Nonlinear Alternatives: Autoencoders

To overcome PCA's limitations, we explored **autoencoders**, which are neural networks trained to reconstruct input data via a low-dimensional bottleneck. This approach is data-driven and can capture complex nonlinear relationships. Our architecture consisted of an encoder and decoder with a single hidden layer of 128 neurons. The latent factors learned from the bottleneck were used as nonlinear analogs to principal components and evaluated against PCA using the R^2 and MSE metrics described above.

4 Data Exploration

4.1 Daily Return Distributions

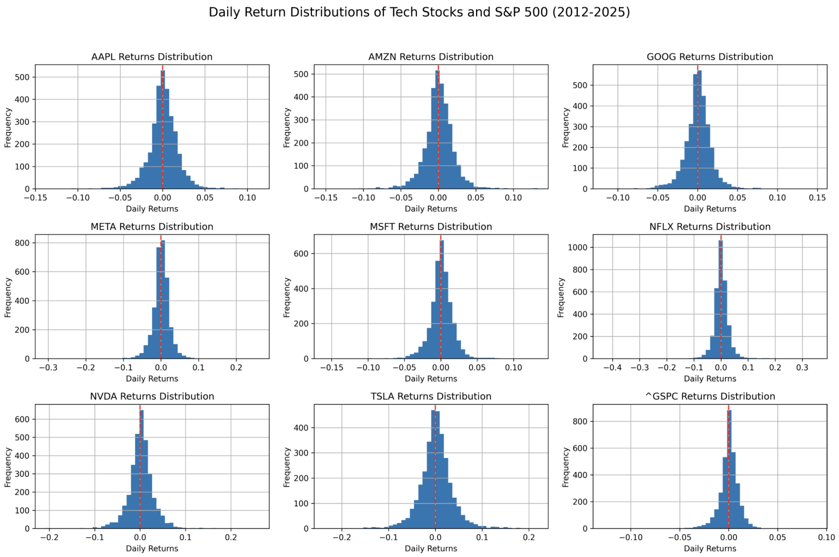


Figure 1: Daily return distributions of tech stocks and the S&P 500 index

Figure 1 displays histograms of daily log returns for eight major tech stocks and the S&P 500 index. All distributions are centered near zero, consistent with the efficient market hypothesis. However, they exhibit strong leptokurtosis (peakedness and fat tails), indicating a higher probability of extreme returns than predicted by a normal distribution. Tesla and Netflix show particularly heavy tails, reflecting their higher volatility. In contrast, the S&P 500 distribution is narrower and more symmetric, consistent with diversified market behavior.

4.2 Correlation Analysis

The correlation matrix (Figure 2) shows that returns across tech stocks are generally positively correlated, with most correlations between 0.4 and 0.7. Microsoft and Google exhibit the highest correlations with the S&P 500, while Tesla and Netflix are less correlated, suggesting a greater influence of idiosyncratic factors on their returns.

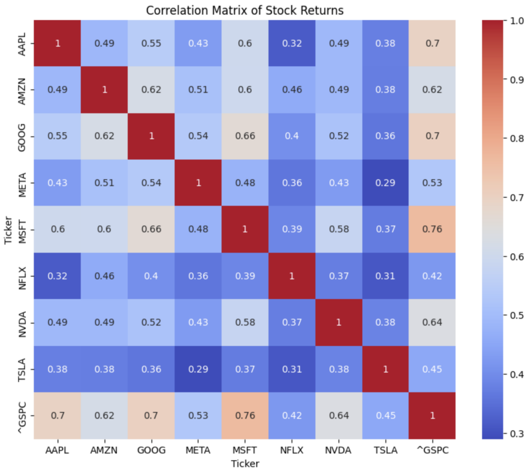


Figure 2: Correlation heatmap of daily returns across tech stocks and the S&P 500 index

4.3 Exploratory Histograms by Data Source

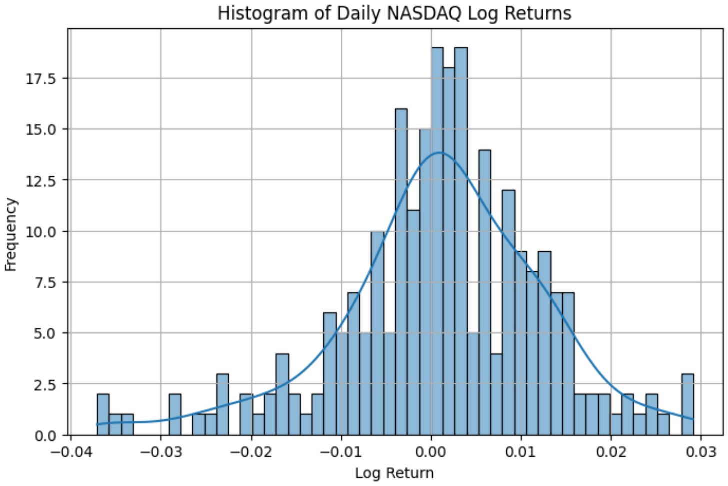


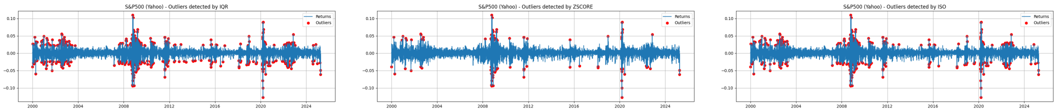
Figure 3: Histogram of NASDAQ Composite log returns from Polygon.io

Figures **??** and 3 show that Alpha Vantage and Polygon.io data exhibit similar distributional characteristics, namely fat tails and slight asymmetry. This confirms that heavy-tailed behavior is robust across different data sources.

5 Outlier Detection

We applied three methods to detect outliers in daily log returns for both the S&P 500 index (Yahoo Finance) and NASDAQ Composite index (Polygon.io): Interquartile Range (IQR), Z-Score, and Isolation Forest (ISO).

5.1 S&P 500 Outliers



(a) IQR

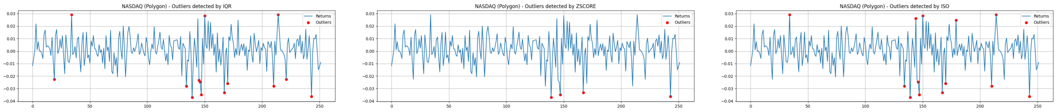
(b) Z-Score

(c) Isolation Forest

Figure 4: S&P 500 daily returns outlier detection results

Figures 4 illustrate the behavior of each outlier detection method on the S&P 500 daily returns. IQR captures a large number of points during high-volatility periods, notably around the 2008 financial crisis and the COVID-19 crash in 2020. However, because IQR is based solely on spread without considering distribution shape, it flags both extreme shocks and moderate volatility expansions. Z-Score, being stricter, identifies fewer points — mainly the most extreme returns far from the mean — and thus misses moderate but still abnormal volatility bursts. Isolation Forest captures a middle ground, detecting sharp shocks as well as volatility clusters, and shows better adaptability by identifying anomalies not necessarily tied to extreme magnitude but to irregular local patterns. This highlights the importance of using more sophisticated models when simple spread-based metrics like IQR fail to capture clustering effects.

5.2 NASDAQ Composite Outliers



(a) IQR (b) Z-Score (c) Isolation Forest

Figure 5: NASDAQ Composite daily returns outlier detection results

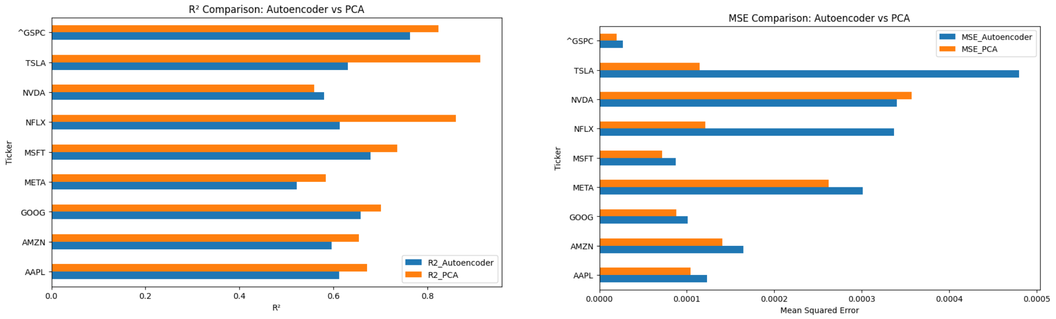
Figures 5 show broadly similar patterns for the NASDAQ Composite index, though with greater dispersion in detected outliers due to the higher inherent volatility in NASDAQ returns. IQR again captures a wide set of points but struggles to distinguish isolated shocks from sustained elevated volatility. Z-Score picks up only the most severe negative returns, especially around sharp market drawdowns. Isolation Forest demonstrates its flexibility by detecting both isolated spikes and clusters of unusual behavior, which may not be significant individually but form anomalous periods collectively. This behavior suggests that Isolation Forest could be particularly valuable in real-world trading environments where clusters of small anomalies often precede larger structural shifts.

5.3 Summary

Overall, the results highlight important trade-offs across methods. While IQR is simple and effective for symmetric, stable markets, it tends to overflag during general volatility increases. Z-Score is useful for strict extreme event detection but may miss important structural breaks. Isolation Forest, although more complex, provides richer insights by adapting to both magnitude and temporal anomaly patterns, making it a powerful tool for volatility regime detection in financial markets.

6 Autoencoder vs PCA Results

We compared the performance of Principal Component Analysis (PCA) and a neural network-based autoencoder in reconstructing standardized daily returns across major tech stocks and the S&P 500 index. Evaluation was based on two metrics: coefficient of determination (R^2) and mean squared error (MSE).



(a) R^2 scores comparison (b) Mean squared error comparison

Figure 6: Autoencoder vs PCA reconstruction performance across different assets.

6.1 Detailed Analysis

Figure 6 highlights key differences between PCA and autoencoder reconstructions. PCA consistently achieves higher R^2 values and lower MSE across all securities, particularly for highly volatile stocks such as Tesla (TSLA) and Netflix (NFLX). These results suggest that the linear structure extracted by PCA captures the dominant sources of variance in financial returns more effectively than the nonlinear representations learned by the autoencoder.

Interestingly, the autoencoder performs relatively closer to PCA on more stable assets like AAPL, AMZN, and MSFT, where return dynamics are smoother and less noisy. This indicates that, in low-noise environments, simple nonlinear compression via an autoencoder can approximate linear techniques reasonably well.

However, for assets characterized by sharp jumps, heavy tails, or sudden volatility shocks, the autoencoder struggles. Neural networks, including autoencoders, often require large amounts of diverse training data to generalize effectively. In this case, the relatively small dataset (daily returns across 13 years) may not have provided sufficient samples for the autoencoder to fully learn meaningful latent representations without overfitting.

Moreover, the nature of financial returns — typically dominated by noise and abrupt regime changes — may favor robust linear decompositions like PCA, especially when the goal is variance explanation rather than high-dimensional feature abstraction.

6.2 Summary

Overall, PCA demonstrated superior reconstruction performance on this dataset. While autoencoders theoretically offer an advantage by capturing nonlinear relationships, practical challenges such as limited data, model complexity, and sensitivity to noise limited their effectiveness in this context. Future work could explore more complex architectures (e.g., denoising autoencoders or variational autoencoders) or augment training data to better exploit the potential of nonlinear models in financial applications.

7 Clustering on Latent Features

7.1 Clustering on Latent Features

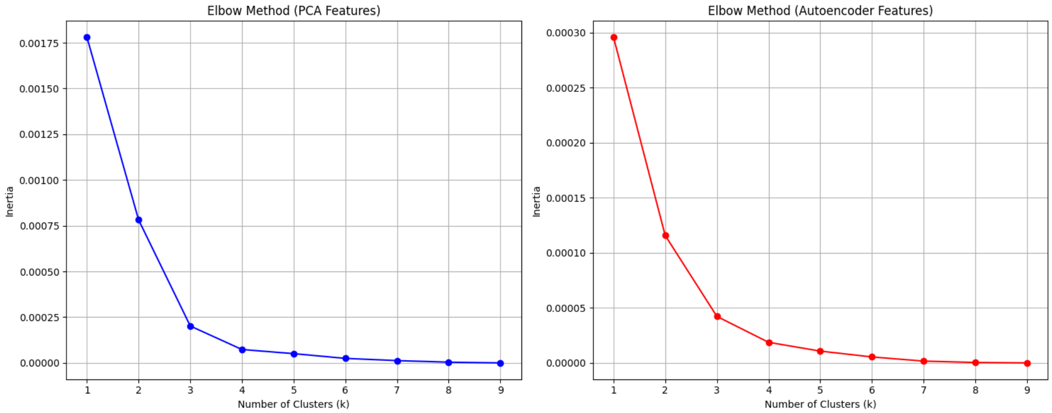


Figure 7: Results for latent feature clustering using the elbow method

Next, we performed clustering on the latent features extracted from both PCA and autoencoder methods to identify natural groupings in our stock market data. Using the elbow method, we plotted inertia against different numbers of clusters to find the optimal number of clusters where adding more clusters yields diminishing returns. This step was crucial for discovering which stocks behave similarly based on their latent factor exposures rather than relying on predefined sector classifications. The analysis revealed that three clusters optimally describe both our PCA and autoencoder embeddings, suggesting that despite their different mathematical approaches, both methods capture similar underlying market structures among these tech stocks.

Following our clustering analysis using the elbow method, we applied K-means clustering (with k=3) to both our PCA and autoencoder factor exposures to identify natural groupings among the stocks. This step goes beyond simple dimensionality reduction by organizing stocks into groups based on similar behavior in the latent space. While factor analysis tells us how stocks relate to underlying market drivers, clustering reveals which stocks move together based on these relationships, potentially uncovering market segments that traditional sector classifications might miss. The results

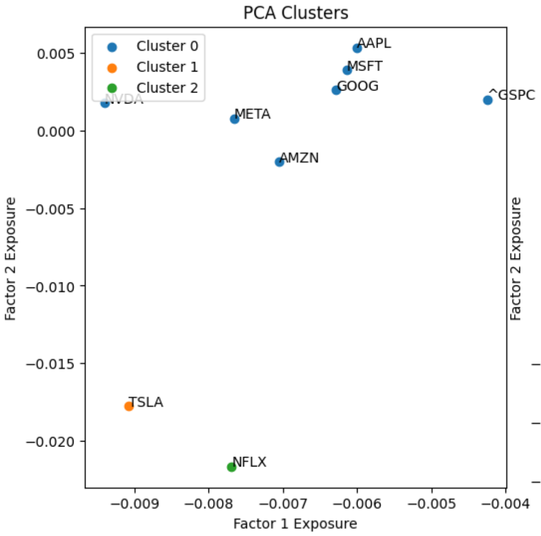


Figure 8: Comparison of K-means clustering results on PCA and autoencoder factor exposures

reveal interesting differences between the linear and nonlinear approaches: In the PCA results, most major tech stocks (AAPL, AMZN, GOOG, META, MSFT, NVDA) cluster with the market index (GSPC) in Cluster 0, while TSLA and NFLX each form their own singleton clusters (Clusters 1 and 2 respectively). This suggests that from a linear perspective, most tech stocks behave similarly, while TSLA and NFLX have distinctive return patterns. In contrast, the autoencoder clustering identifies different relationships: META and NVDA form their own cluster (Cluster 0), most large-cap tech stocks (AAPL, AMZN, GOOG, MSFT) plus the market index form Cluster 1, and NFLX and TSLA are grouped together in Cluster 2. This suggests the nonlinear autoencoder is capturing subtler relationships between stocks that the linear PCA approach misses, particularly similarities between NFLX and TSLA that weren't apparent in the linear analysis. These findings support our project hypothesis that nonlinear dimensionality reduction techniques can uncover market structures that classical linear methods might overlook.

8 Conclusion

In this project, we compared classical PCA and nonlinear autoencoders for modeling latent factors in financial return data, and evaluated outlier detection techniques including Z-score filtering, IQR, and Isolation Forests. We found that while PCA achieved higher reconstruction accuracy overall, autoencoders uncovered richer latent structures when clustering assets, identifying relationships that linear methods overlooked. Additionally, Isolation Forests detected multivariate anomalies more effectively than traditional univariate methods.

Our results suggest that nonlinear models can offer complementary insights to classical techniques in finance, particularly in identifying subtle dependencies and shifts in market structure. However, their effectiveness is sensitive to dataset size, noise levels, and model complexity, which can limit their advantages without careful tuning and validation. This study highlights that combining nonlinear dimensionality reduction with machine learning-based anomaly detection can enhance our understanding of financial return behavior.

Future work could explore deeper autoencoder architectures such as denoising or variational autoencoders, and apply these methods to larger, higher-frequency datasets to test their robustness across different market conditions. Incorporating regime-switching behavior, market microstructure effects, or cross-asset relationships could further extend the application of nonlinear latent factor models in financial analytics.

References

Below are all resources that were used.

[1] Connor, G., & Korajczyk, R. A. (1986). Performance measurement with the arbitrage pricing theory: A new framework for analysis. *Journal of Financial Economics*, 15(3), 373–394. https://www.sciencedirect.com/science/article/pii/0304405X86900279

[2] Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504–507. https://doi.org/10.1126/science.1127647

[3] Nanda, S. R., Mahanty, B., & Tiwari, M. K. (2010). Clustering Indian stock market data for portfolio management. *Expert Systems with Applications*, 37(12), 8793–8798. https://doi.org/10.1016/j.eswa.2010.06.026

[4] Baser, P., & Saini, J. R. (2015). Agent-based stock clustering for efficient portfolio management. *International Journal of Computer Applications*, 113(3), 6–12. https://doi.org/10.5120/20317-2381

[5] Khoo, T. H., Pathmanathan, K., & Dabo-Niang, S. (2023). Spatial autocorrelation of global stock exchanges using functional areal spatial principal component analysis. *Mathematics*, 11(3), 674. https://doi.org/10.3390/math11030674

[6] Choi, S., Gwak, M., Song, I., & Chang, W. (2020). Stock market network based on bi-dimensional histogram and autoencoder. *Intelligent Data Analysis*, 24(4), 793–807. https://doi.org/10.3233/IDA-215819