An LSTM Autoencoder for Dogecoin Anomaly Detection

Matt Carswell
Georgetown University
12/10/24

Abstract

This project develops an LSTM (long short-term memory) autoencoder model to detect anomalous behavior in the Dogecoin cryptocurrency market using OHLC (open, high, low, close) and volume-related features. The model captures complex, temporal relationships between multiple variables, offering a more nuanced detection of market anomalies compared to traditional volume-based methods. Principal component analysis revealed that market anomalies in 2021 were influenced by a broad set of factors, while those in 2024 were primarily driven by shifts in trading volume. The LSTM autoencoder provides a robust framework for anomaly detection and can serve as the foundation for future market alert systems or high-frequency trading algorithms, with potential applications to other financial assets.

1. Introduction

Dogecoin, a cryptocurrency created as a parodic alternative to Bitcoin in 2013, has grown into a unique digital asset with a vibrant community. Initially introduced as a “meme coin”, Dogecoin features the Shiba Inu dog from the popular "Doge" meme as its logo. Despite its humorous origins, the cryptocurrency has gained real-world utility for tipping content creators, charitable donations, microtransactions, and even speculative investment.

Dogecoin's behavior in the market has been characterized by significant volatility, often driven by community sentiment, social media trends, and endorsements from high-profile individuals.[1] Notably, tweets and public comments from influential figures have caused sudden and dramatic price fluctuations. This speculative nature, combined with its relatively low transaction fees and high coin supply, sets Dogecoin apart from more traditional cryptocurrencies.

[1] https://coinmarketcap.com/academy/article/elon-musks-history-in-crypto-the-good-the-bad-and-the-doge

Understanding and predicting anomalous behavior in Dogecoin's market activity is critical for traders and analysts. Unusual price spikes or trading patterns may indicate shifts in investor sentiment or the influence of external factors, making anomaly detection an essential tool for navigating the unpredictable landscape of cryptocurrency trading. **In this paper, a method for the detection of anomalous market behavior in Dogecoin is developed using a long short-term memory autoencoder.**

2. Methods

2.1 Model Architecture

As mentioned, an LSTM autoencoder is the main method used in this project in order to predict anomalies in the Dogecoin market. An **autoencoder** is a type of neural network used for unsupervised learning (or self-supervised in some cases), primarily aimed at data compression or dimensionality reduction. It consists of two main components: an **encoder**, which compresses the input data into a lower-dimensional representation (latent space), and a **decoder**, which reconstructs the original input from this compressed form. The goal of an autoencoder is to minimize the difference between the input and the output, learning an efficient representation of the data in the process.

An **LSTM autoencoder** is a variant of the traditional autoencoder designed to handle sequential data by incorporating Long Short-Term Memory (LSTM) cells in both the encoder and decoder. LSTMs are specialized recurrent neural networks capable of capturing long-term dependencies in time series data. The encoder uses LSTM layers to process and compress sequential data into a fixed-length vector that captures temporal relationships, while the decoder reconstructs the original sequence from this compressed form, learning the underlying patterns in the data. The main benefits of an LSTM autoencoder include its ability to capture complex temporal dependencies and non-linear patterns in sequential data, making it ideal for time series anomaly detection. Additionally, the LSTM autoencoder is capable of generalizing across different time periods and identifying subtle deviations from normal behavior, making it highly effective for detecting anomalies in financial data, such as cryptocurrency markets.

Using an autoencoder for anomaly detection provides a comprehensive and data-driven approach compared to relying solely on EDA or flagging anomalies based on volume. Autoencoders analyze multiple variables (such as volume, log returns, and volatility, in our case) to capture complex, nonlinear relationships and dependencies that might indicate anomalies. Unlike volume-based methods, which may miss subtler market irregularities or over-flag normal fluctuations, an autoencoder learns a baseline of normal behavior from historical data and identifies deviations objectively through reconstruction errors. Additionally, LSTM-based autoencoders account for temporal patterns and evolving market dynamics, making them particularly suited for detecting nuanced anomalies in time series

data. This approach ensures a broader and more reliable detection of unusual market activity in Dogecoin's volatile environment.

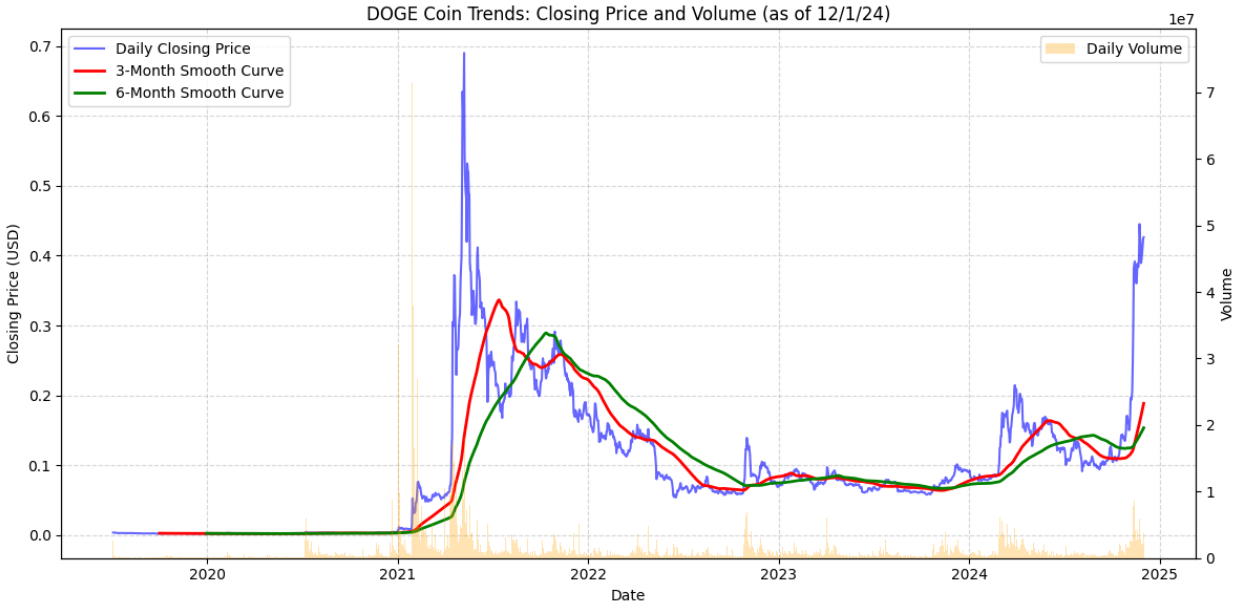


Figure 1: DOGE Closing Price and Volume over Time

2.2 Data Overview

Data for this project was taken from a Kaggle dataset[2] that is updated weekly via Binance's Cryptocurrency API[3] and contains OHLC (open, high, low, close) and volume data for the 50 most popular cryptocurrencies at the minute level. Just Dogecoin data (the 7th most popular cryptocurrency) was used. The data used in this paper's analysis was last updated December 1, 2024 and downloaded as a .csv file for analysis.

In order to preserve computational efficiency and avoid some of the pitfalls that come with anomaly detection dealing with noisy data at the minute-level, all data was aggregated to the hour-level. **Therefore, all anomalies that are predicted with the autoencoder represent anomalous hours of market activity.** We can observe our hourly aggregated data in the plot below.

Due to large fluctuations seen in 2021 and 2024 in the graph above, and based on what we know about the general consensus of Dogecoin being in its most volatile and speculative states in these years,[4] data from 2021 and 2024 were chosen as our validation and test datasets, respectively. Once our data was split, 9 features were selected through a selection process involving exploratory data analysis (EDA)[5] and Principal Component

[2] https://www.kaggle.com/datasets/kaanxtr/btc-price-1m
[3] https://www.binance.com/en/binance-api
[4] See Appendix A for more information.
[5] See Appendix B.

Analysis (PCA).[6] These 9 features include 'volume', 'quote_asset_volume', 'number_of_trades', 'relative_volume', 'log_return', 'volatility', 'range', 'change', and 'close_open_ratio'.

The variables used—such as relative volume, log return, and volatility—are designed to capture market momentum and changes relative to historical trends as opposed to absolute prices contained in OHLC data. Specifically, relative volume and log return are calculated based on the past 24 hours, allowing the model to detect unusual spikes or drops in trading activity, price changes, and overall market dynamics. These variables collectively highlight shifts in behavior that deviate from established patterns, enabling the identification of anomalies. Therefore, an anomaly in the Dogecoin market is defined as any significant deviation from baseline market activity, according to these variables, observed during the training period, which includes data from 2019, 2020, 2022, and 2023.

3. Analysis and Results

3.1 Optimal Model

Through a process of manual hyperparameter tuning, observations on how well our autoencoder learned the training data, and the extent to which our autoencoder predicted anomalies on our validation set, an optimal model architecture was found. Keras and TensorFlow served as the main packages for model compilation, training, and evaluation.

First off, our data was split into matrices of size (60,9) with each row . Therefore, our total train dataset was a rank-3 tensor with size (30553, 60, 9), with our validation set being size (8743, 60, 9), and the test set of size (8043, 60, 9). This structure serves as a quasi hour-level aggregation where each matrix represents one hour of data with each row representing one minute of data. Our data is thus being processed by our model in a way that allows it to detect anomalies hour-by-hour.

The final, optimal model can be described as such: the encoder consists of an LSTM layer with 64 units and a 'tanh' activation function, followed by a bottleneck layer that repeats the encoded output to match the original sequence length. The decoder mirrors the encoder with another LSTM layer, also with 64 units and 'tanh' activation, followed by a TimeDistributed Dense layer to reconstruct the input data. The model uses Mean Squared Error (MSE) as the loss function and is optimized with the Adam optimizer. Key hyperparameters include a sequence length of 60 and 9 features, with a batch size of 64 and a maximum of 200 epochs. Early stopping was implemented to monitor training loss (since we are attempting to learn the training data, baseline market activity, as much as possible), with a patience of 10 epochs and a minimum delta of 0.001, and the model

[6] See Appendix C.

stopped training at the 82nd epoch when no significant improvement in loss was observed.[7] The best weights were restored after early stopping. A summary of the model architecture can be found below.

Layer (type)	Output Shape	Param #
input_layer (InputLayer)	(None, 60, 9)	0
lstm (LSTM)	(None, 64)	18,944
repeat_vector (RepeatVector)	(None, 60, 64)	0
lstm_1 (LSTM)	(None, 60, 64)	33,024
time_distributed (TimeDistributed)	(None, 60, 9)	585

Total params: 52,553 (205.29 KB)
Trainable params: 52,553 (205.29 KB)
Non-trainable params: 0 (0.00 B)

Figure 2: LSTM Autoencoder Architecture

The encoder's transformation from 60,9 (60 timesteps and 9 features) to a 64-dimension dense vector helps to reduce the dimensionality and extract important features from the input sequence. The RepeatVector architecture, with a (60,64) structure, repeats this compressed representation, effectively allowing the model to learn both the temporal context and the relationships between features across time. The decoder then reconstructs the input data by learning additional weights through a second LSTM layer, creating a hidden output of (64,60) which is notably larger than our input data dimensions. Such a structure is often referred to as an "overcomplete" structure, meaning the decoder has more units than the input, and also in our case, the compressed representation. This overcompleteness can be beneficial because it allows the decoder to better model the complex temporal patterns and non-linear relationships within the data, ensuring that even subtle anomalies can be detected. By expanding the latent space, the model can reconstruct the data with higher fidelity, providing a better measure of reconstruction error for anomaly detection.

3.2 Anomaly Thresholding

With our optimal model, an anomaly was defined as the mean training reconstruction loss times four standard deviations of the training reconstruction loss. Typically, three standard deviations is the “norm” but in order to be a little more conservative in predicting anomalous market activity, four standard deviations was chosen. Therefore, any datapoint, an hour in the market, in the validation set (2021) and test set (2024) that records a reconstruction loss of more than 10.42, is classified as an anomalous

[7] See Appendix D

trading hour. The anomaly threshold derived from the training data and applied to the validation can be visualized below in the following plots.

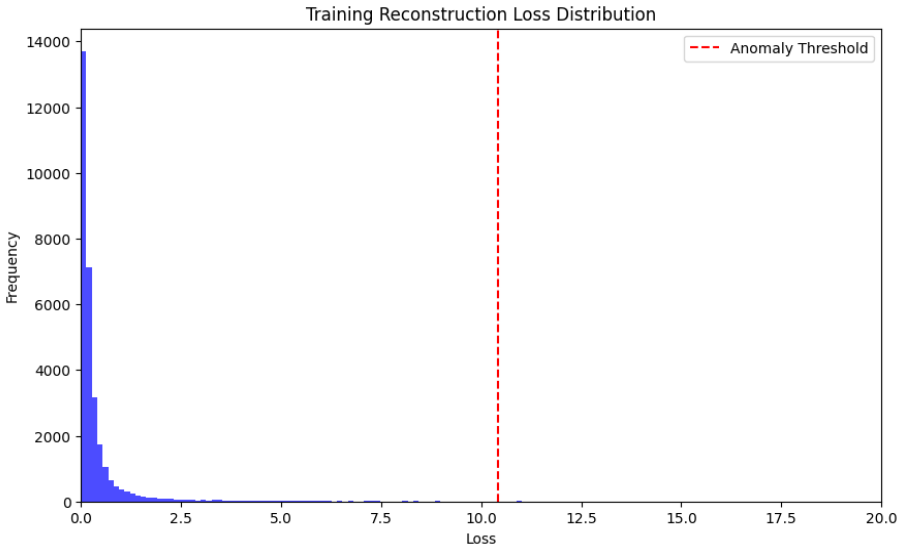


Figure 3: Training Reconstruction Loss

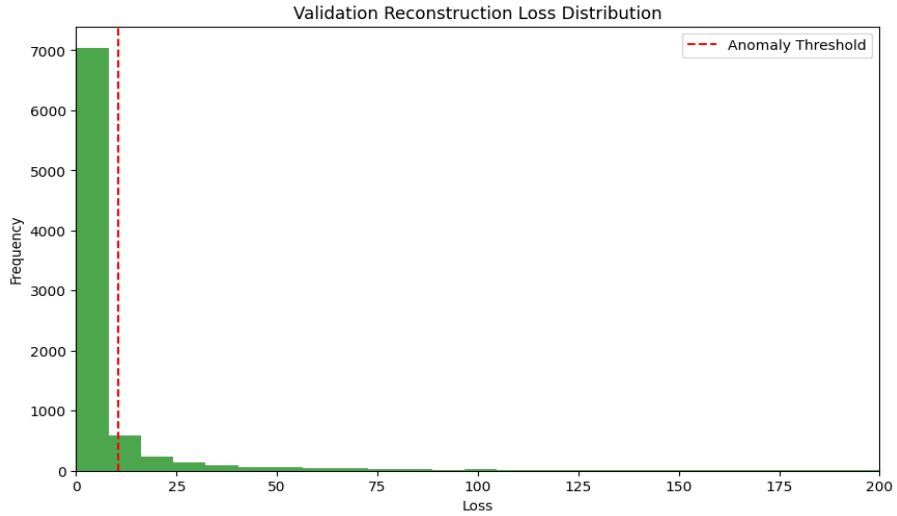


Figure 4: Validation Reconstruction Loss

3.3 Final Results of Validation and Test Data

With our optimal model and anomaly threshold determined, anomalies were calculated for both the validation and test data sets. Once again, the validation and test sets represent the years 2021 and 2024, respectively. **Out of the 8,743 hours in 2021, we identified 1,718 of those hours to present anomalous market activity**. These anomalies can be visualized in the below plot that contains the log returns and trade volume of Dogecoin in 2021.

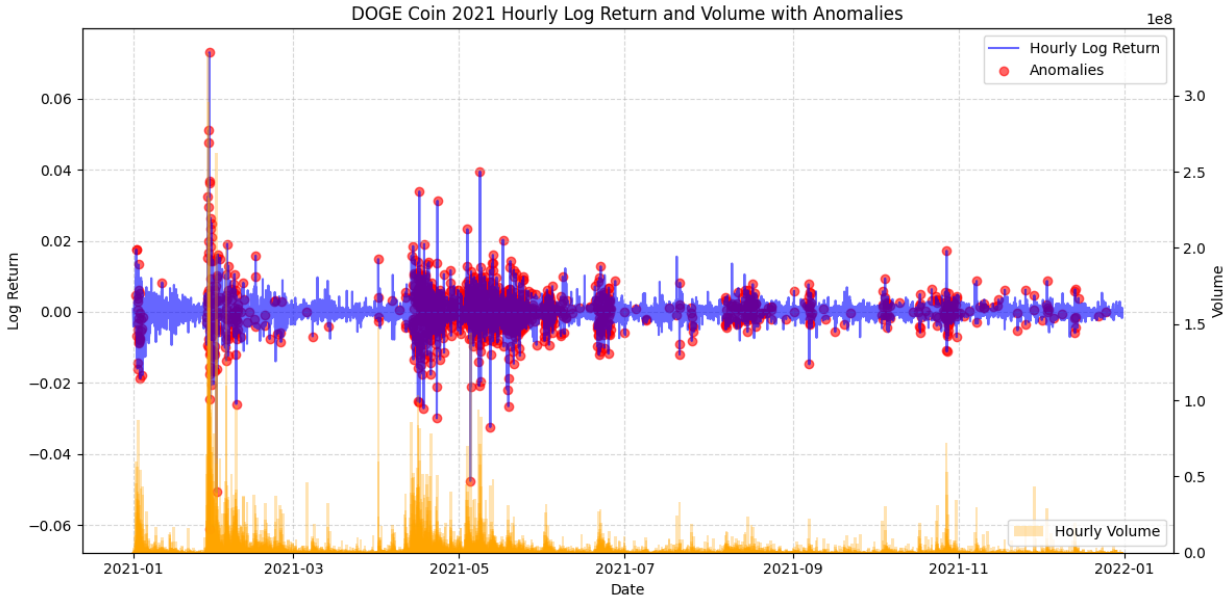


Figure 5: Dogecoin 2021 Hourly Log Return and Volume with Anomalies

First of all, it is important to note that log return and volume are not the only indicators of anomalies but visualizing these variables with anomalies shows us that anomalous hours are being flagged when the market shows more volatility, higher momentum in either price direction, and both large spikes and dips in market volume. Given the knowledge of Dogecoin's behavior in 2021, our model seems to be working correctly.

Similarly, the 2024 test set was visualized in the same way. **Out of the 8,043 hours in 2024, we identified 398 of those hours to present anomalous market activity.** These anomalies can be seen in the figure below.

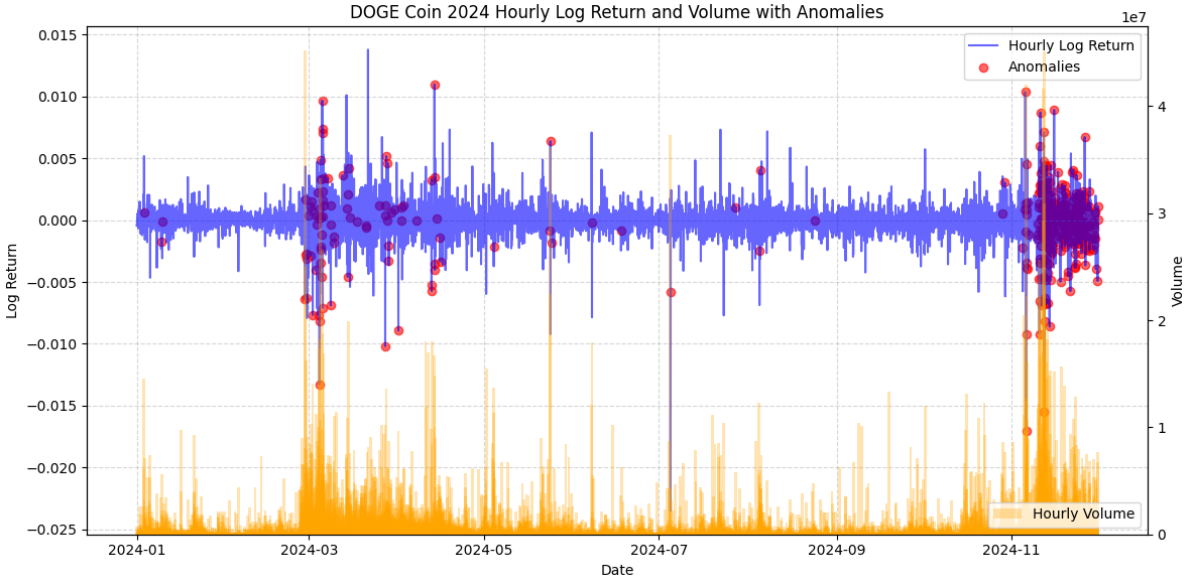


Figure 6: Dogecoin 2024 Hourly Log Return with Anomalies

The 2024 data presents more market stability than the 2021 data, but still presents anomalies around the beginning of Q2 (quarter 2) and also Q4. This observation makes intuitive sense as, at a high-level, 2024 has seemed to present anomalous market fluctuations similar to 2021, but not nearly as inordinate as 2021, especially in the middle parts of the year. The current "bull run" that has taken place in November 2024 is being clearly lit up by our model which is an extremely positive takeaway.

In order to pinpoint further the main drivers of market anomalies, PCA was performed on the anomalous data points for both 2021 and 2024.[8] 2021 anomalies showed to have three main principal components, with the following variables being the most important:

- PC1: number_of_trades, quote_asset_volume, range, volume
- PC2: log_return, close_open_ratio, change
- PC3: volatility, - relative volume

The 2024 anomalies only presented one really main principal component which is explained with the following variables:

- PC1: relative volume (explains most of the variance), number of trades

The analysis of market anomalies in both years, especially in 2024 where trading volume and activity dominate the primary principal component, might initially suggest that focusing solely on volume-based metrics would be sufficient for anomaly detection. However, while volume is a significant indicator, relying only on it risks overlooking the nuanced relationships and dynamics present in the market. The autoencoder provides a more comprehensive approach by accounting for multiple variables simultaneously—such as volatility, log return, and price movements—capturing complex, non-linear interactions between these factors. Unlike simple volume-based methods, which may miss subtle anomalies where volume alone doesn't fully explain market shifts, the autoencoder learns a holistic representation of normal market behavior across all relevant features. This allows for the detection of anomalies that stem from intricate patterns, not just isolated spikes or drops in volume, ensuring a more robust and accurate identification of unusual market conditions.[9]

Finally, in order to visualize a 2-dimensional representation space of our high-dimensional data, t-SNE (t-Distributed Stochastic Neighbor Embedding) was applied to the entire 2021 and 2024 datasets where we have aggregated the data by hour and have each hour labeled as anomalous or non-anomalous. t-SNE is a dimensionality reduction technique that visualizes high-dimensional data by preserving the pairwise similarities

[8] See Appendix E
[9] See Appendix F for further breakdown of features in anomalies versus non-anomalies.

between points, mapping them to a lower-dimensional space while maintaining local structure and revealing patterns or clusters. The visualizations can be seen below.

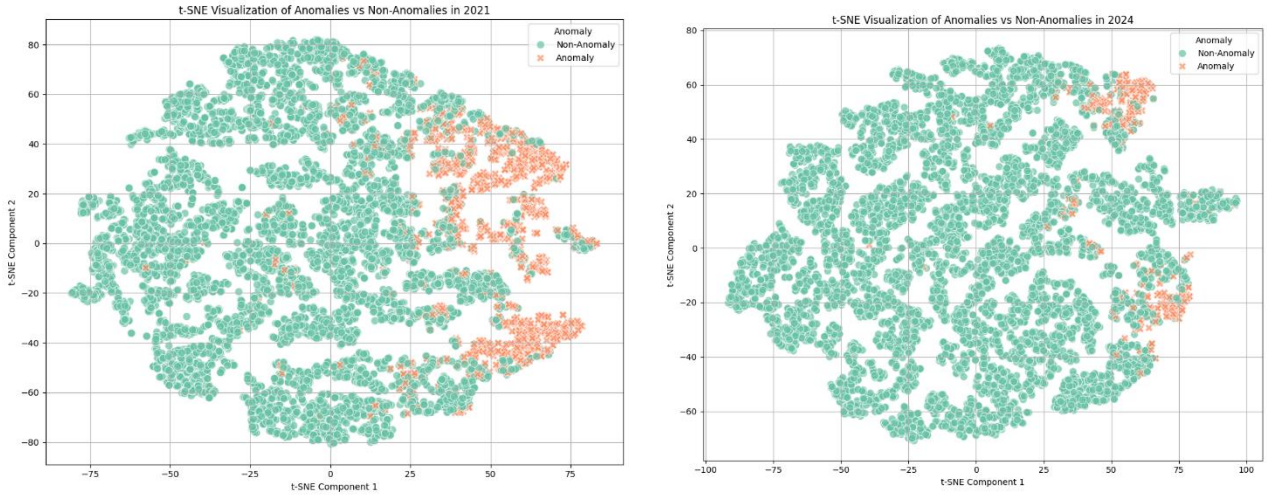


Figure 7: t-SNE Visualizations for 2021 and 2024

The clear separation between anomalies and non-anomalies in the t-SNE graphs indicates that the model has successfully identified distinct patterns in the data, with anomalies forming a separate cluster from normal market behavior. The two sub-clusters of anomalies likely represent different types of market irregularities or behaviors, possibly linked to specific factors. One sub-cluster may correspond to anomalies driven by trading volume and activity, as seen in the 2024 data, where relative volume and number of trades were the primary drivers. The other sub-cluster could represent anomalies influenced by a broader set of factors, as observed in 2021, where variables such as log return, volatility, and range played more significant roles. This differentiation suggests that market anomalies may manifest in different ways depending on the underlying market conditions at different times.

4. Conclusion

In this project, an LSTM autoencoder was developed to detect anomalous behavior in the Dogecoin market using a range of features, including trading volume, price changes, volatility, and return metrics. The autoencoder model, with its ability to capture complex temporal relationships and nonlinear dependencies across multiple variables, proved to be a powerful tool for identifying subtle deviations from normal market behavior. By analyzing the reconstructed error, the model was able to pinpoint anomalies more effectively than simple volume-based methods, offering a more comprehensive view of market irregularities.

Further, principal component analysis (PCA) revealed that market anomalies in 2021 were influenced by a broader range of factors, while anomalies in 2024 were primarily

driven by shifts in trading volume and activity. This distinction highlights the value of the autoencoder in capturing a wide spectrum of anomaly characteristics, ensuring that even the most nuanced changes in market behavior are detected. Ultimately, the LSTM autoencoder provided a robust framework for understanding and detecting anomalies in the volatile and ever-changing cryptocurrency market.

This LSTM autoencoder model has great potential for future applications, particularly in developing a market alert system that notifies users when anomalous behavior occurs, prompting them to check the market and potentially make buy or sell decisions. While this system is not designed to directly execute trades, it could serve as the foundation for a more advanced buy/sell system. Additionally, with more computing power, the model could be adapted for high-frequency trading, detecting minute-by-minute anomalies and executing rapid micro-trades. The same logic behind this approach could be applied to any financial asset with OHLC (Open, High, Low, Close) and volume data, whether it's stocks, commodities, or other cryptocurrencies. By analyzing historical price movements and volume trends, the autoencoder can identify anomalies that signal unusual market behavior, allowing for timely, informed decisions across a broad range of financial markets.

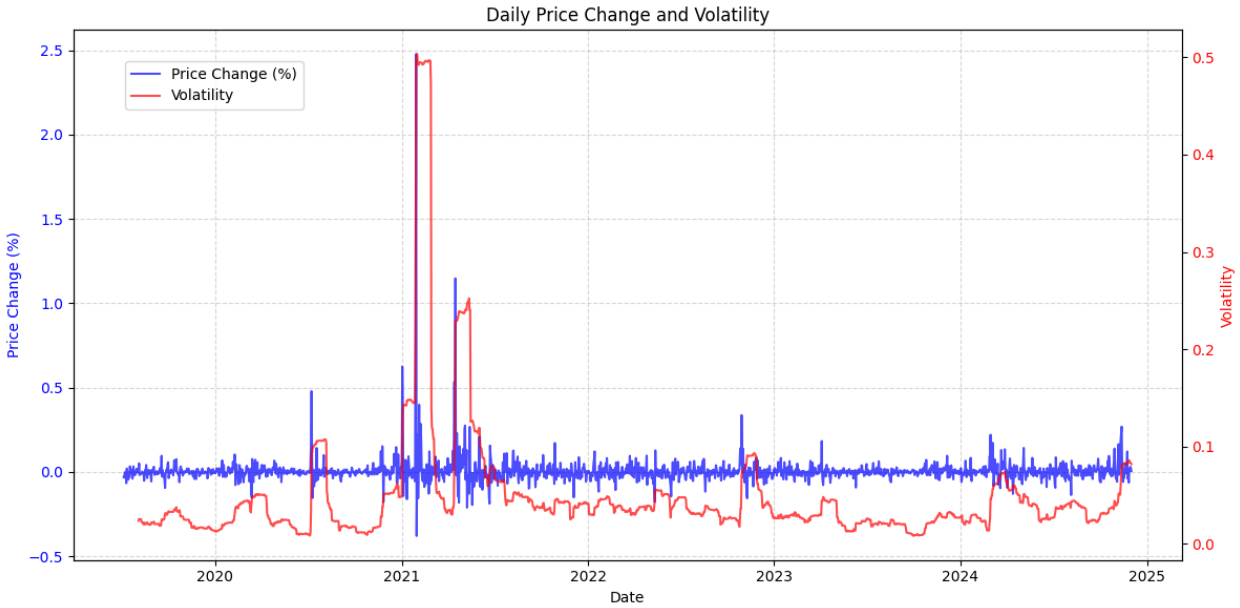
Additional References

Inzirillo, H., & De Villelongue, L. (2023, April 20). An attention free conditional Autoencoder for anomaly detection in cryptocurrencies. arXiv.org. https://arxiv.org/abs/2304.10614

Inzirillo, H., & De Villelongue, L. (2022) An attention free long short-term memory for time series forecasting. arXiv preprint arXiv:2209.09548,

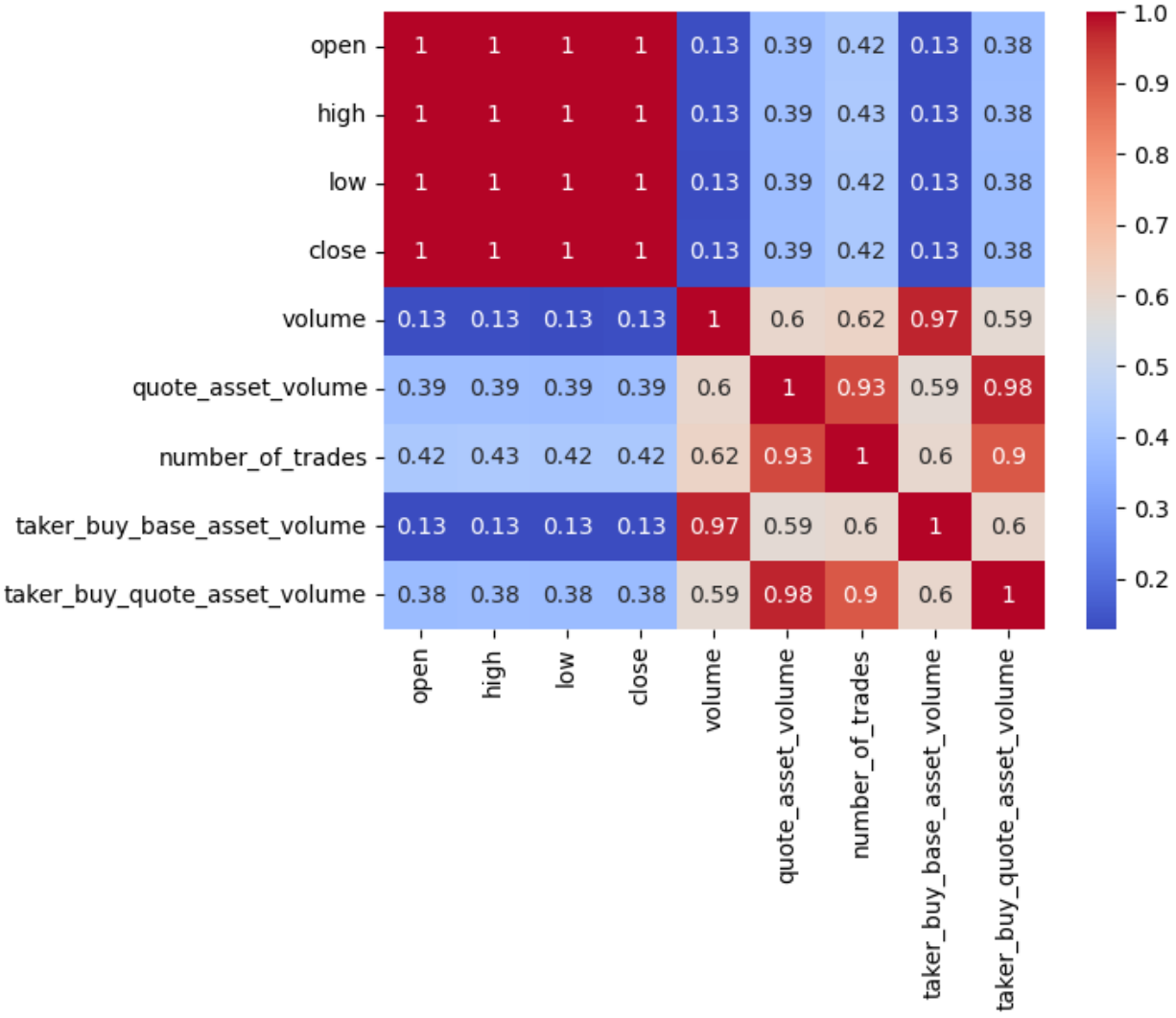
Appendix A

Appendix 1: Dogecoin 2021 Closing Price and Volume



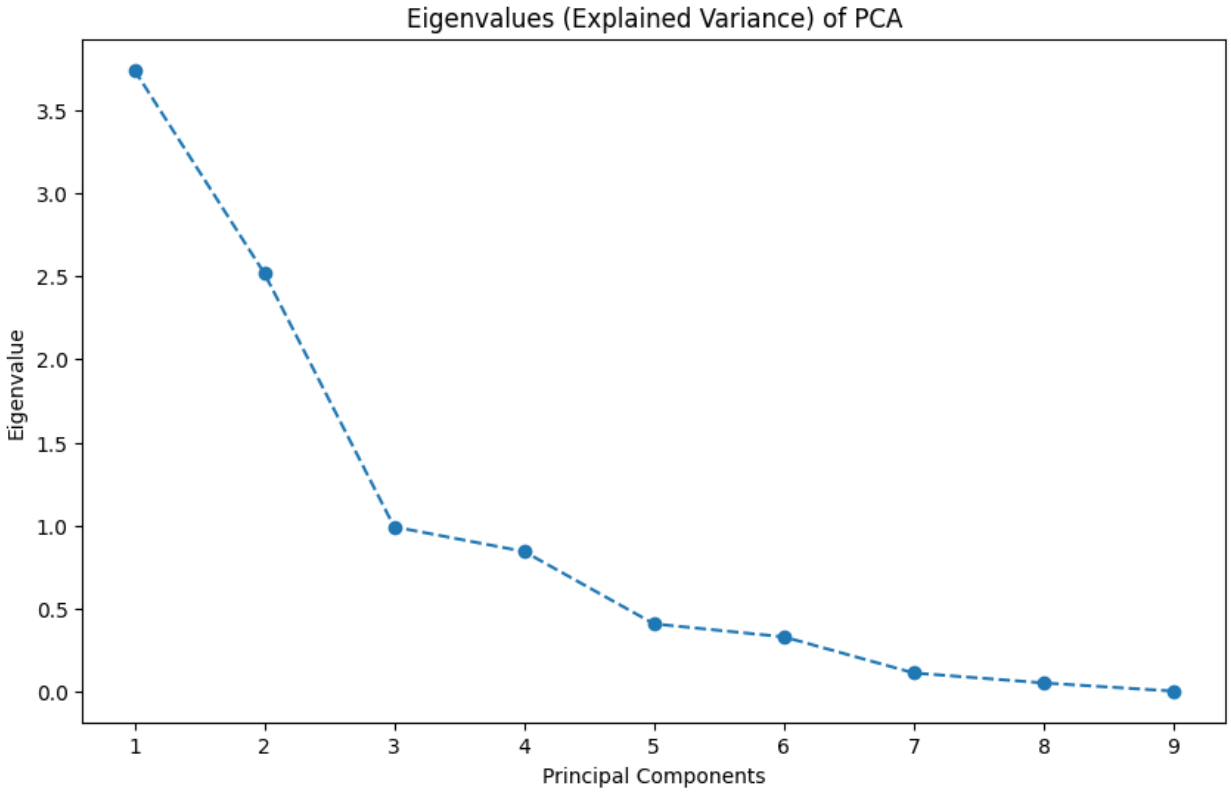
Appendix 2: Dogecoin Daily Price Change and Volatility

Appendix B



Appendix 3: Correlation Matrix for Features

Appendix C



Appendix 4: PCA Eigenvalues

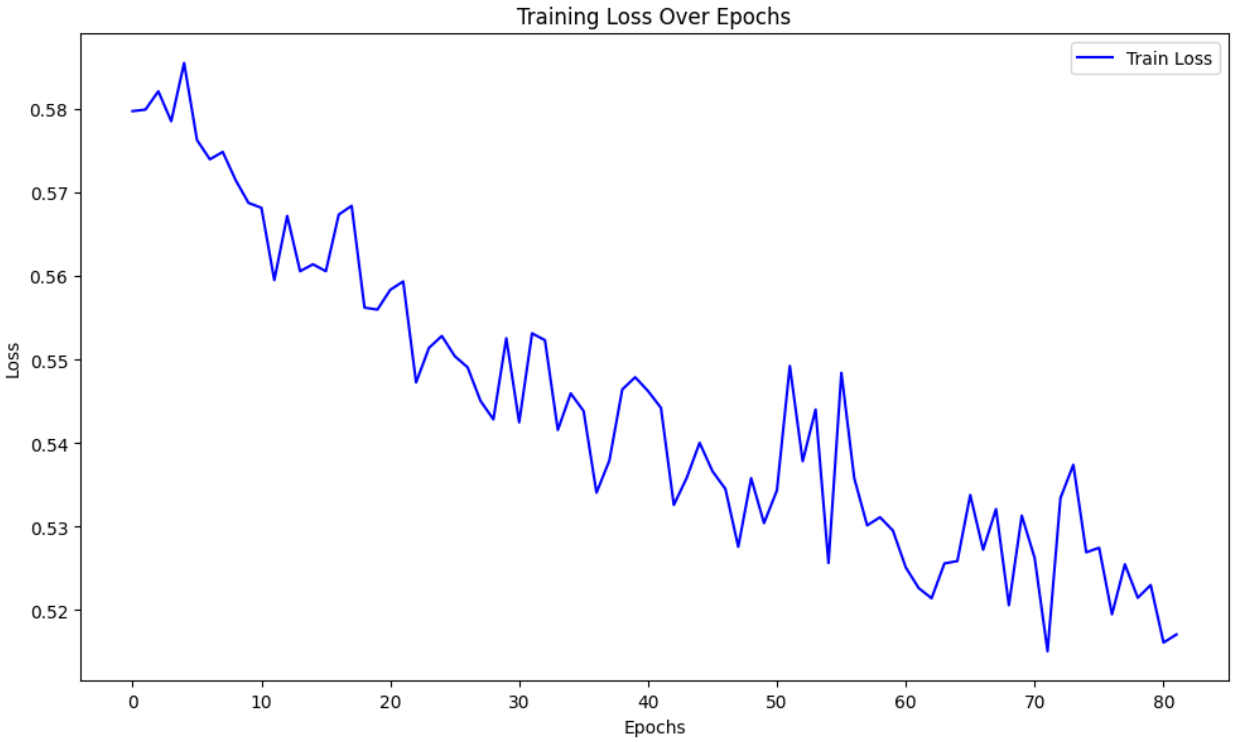
Principal Components (First 4):

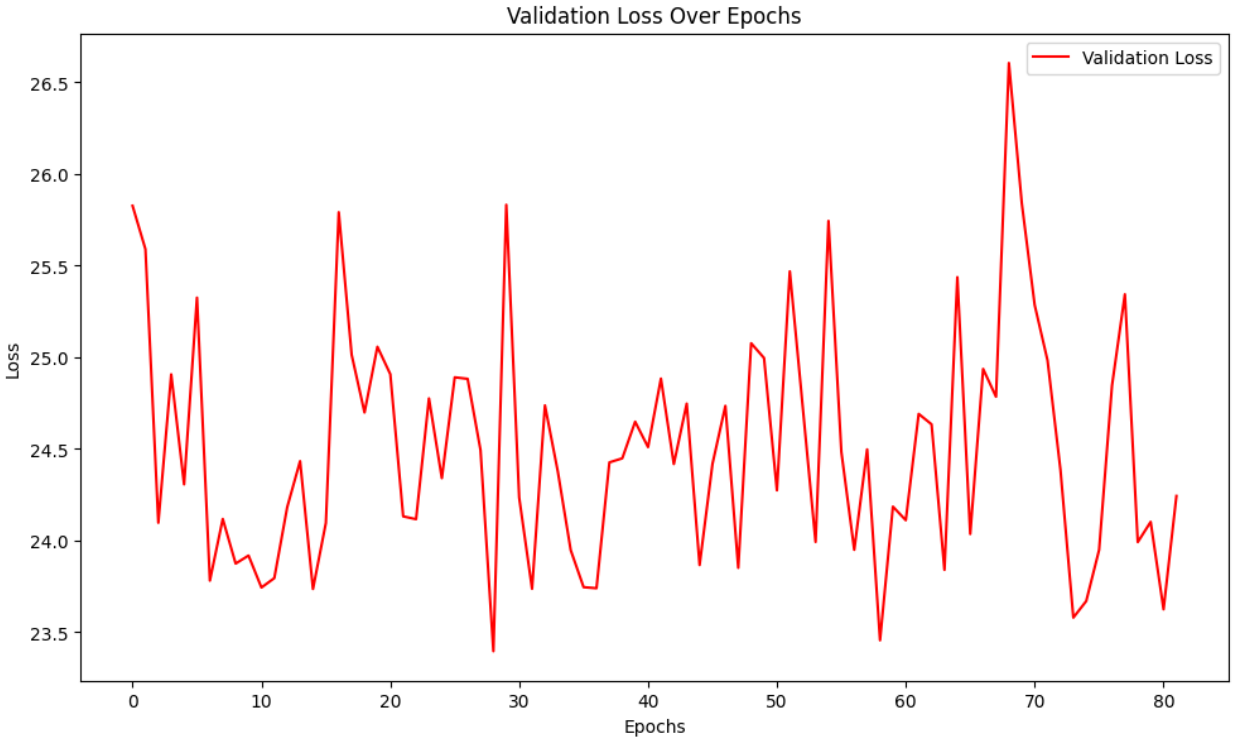
	volume	quote_asset_volume	number_of_trades	relative_volume
0	0.396193	0.467910	0.475684	0.146174
1	-0.049260	-0.129758	-0.125669	-0.000694
2	-0.015333	-0.015978	0.025469	0.929103
3	0.547070	-0.325104	-0.265265	0.229987

	log_return	volatility	range	change	close_open_ratio
0	0.154685	0.341737	0.437159	0.127841	0.168185
1	0.582891	-0.104122	-0.172922	0.503218	0.575508
2	-0.013805	-0.365234	0.000417	-0.043507	-0.012163
3	0.001134	0.572158	-0.355476	-0.131640	0.026542

Appendix 5: PCA Principal Components

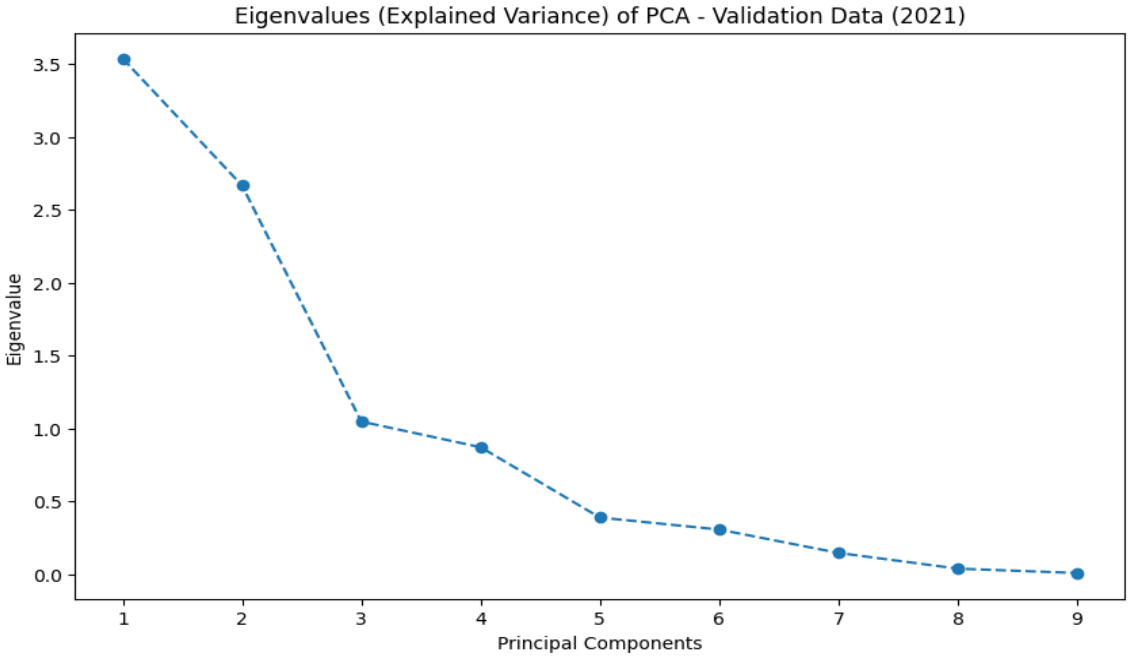
Appendix D



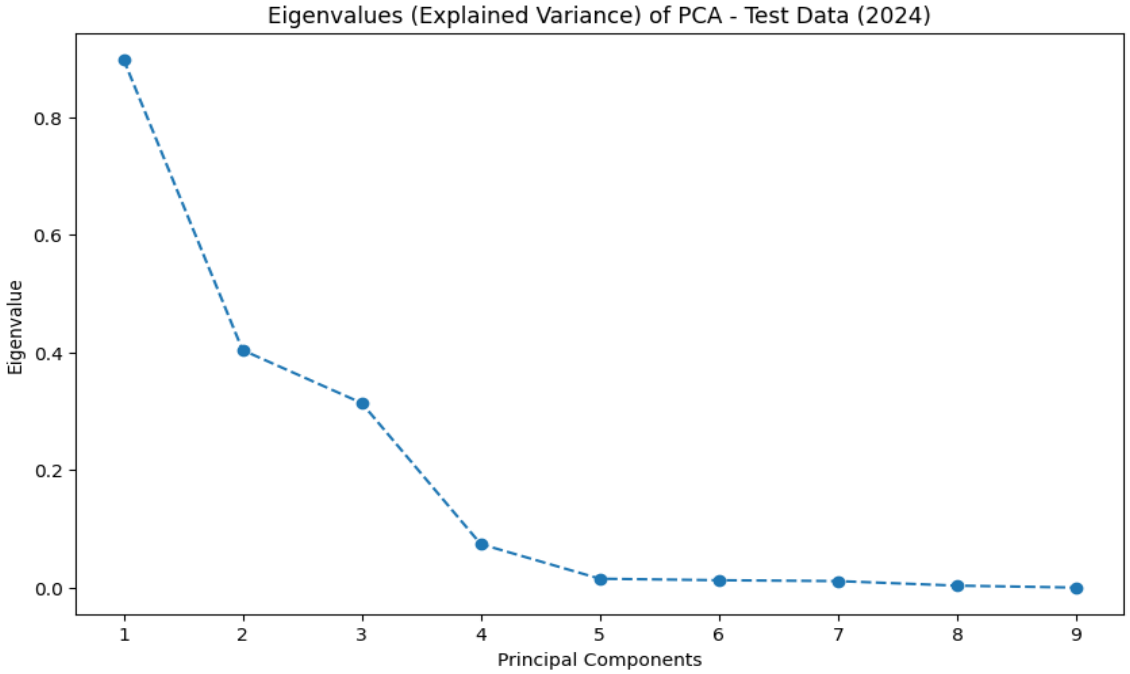


Appendix 6: Training and Validation Loss vs. Epochs

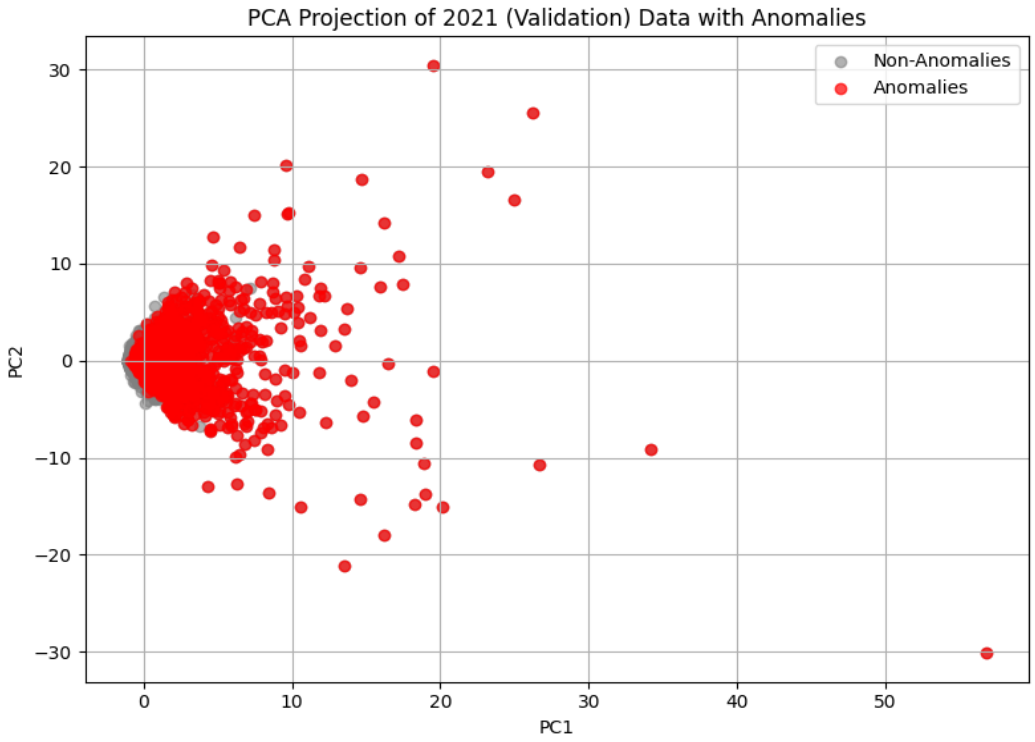
Appendix E



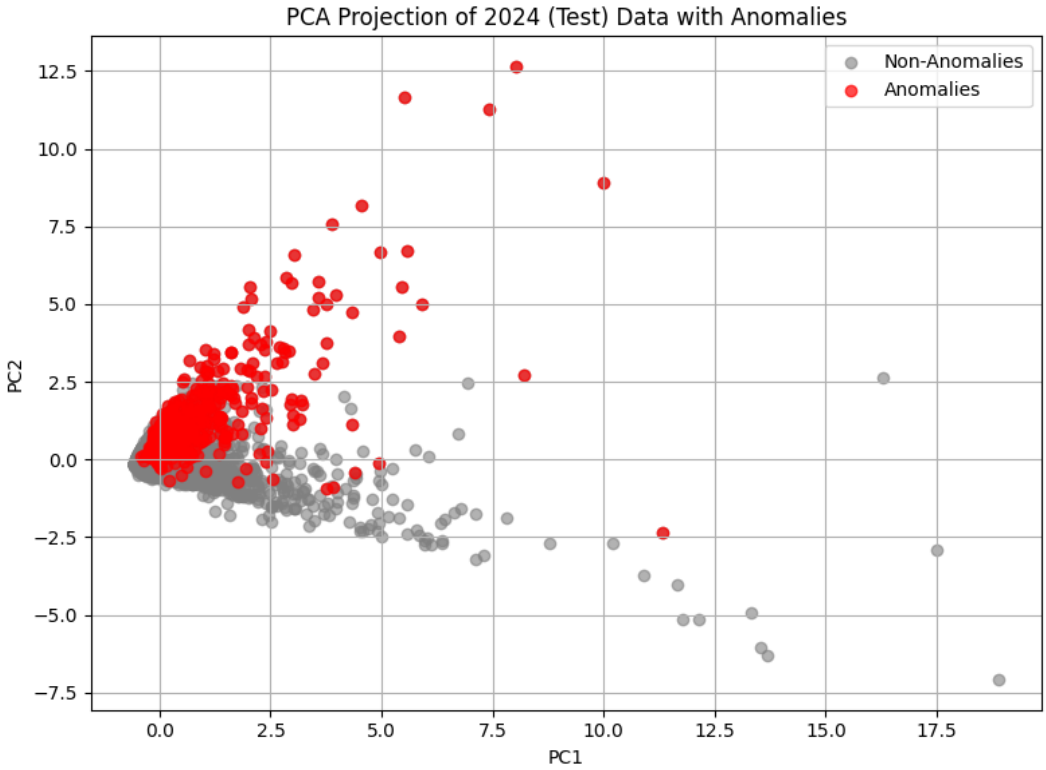
Appendix 7: Anomalies PCA Eigenvalues – Validation Data (2021)



Appendix 8: Anomalies PCA Eigenvalues - Test Data (2024)

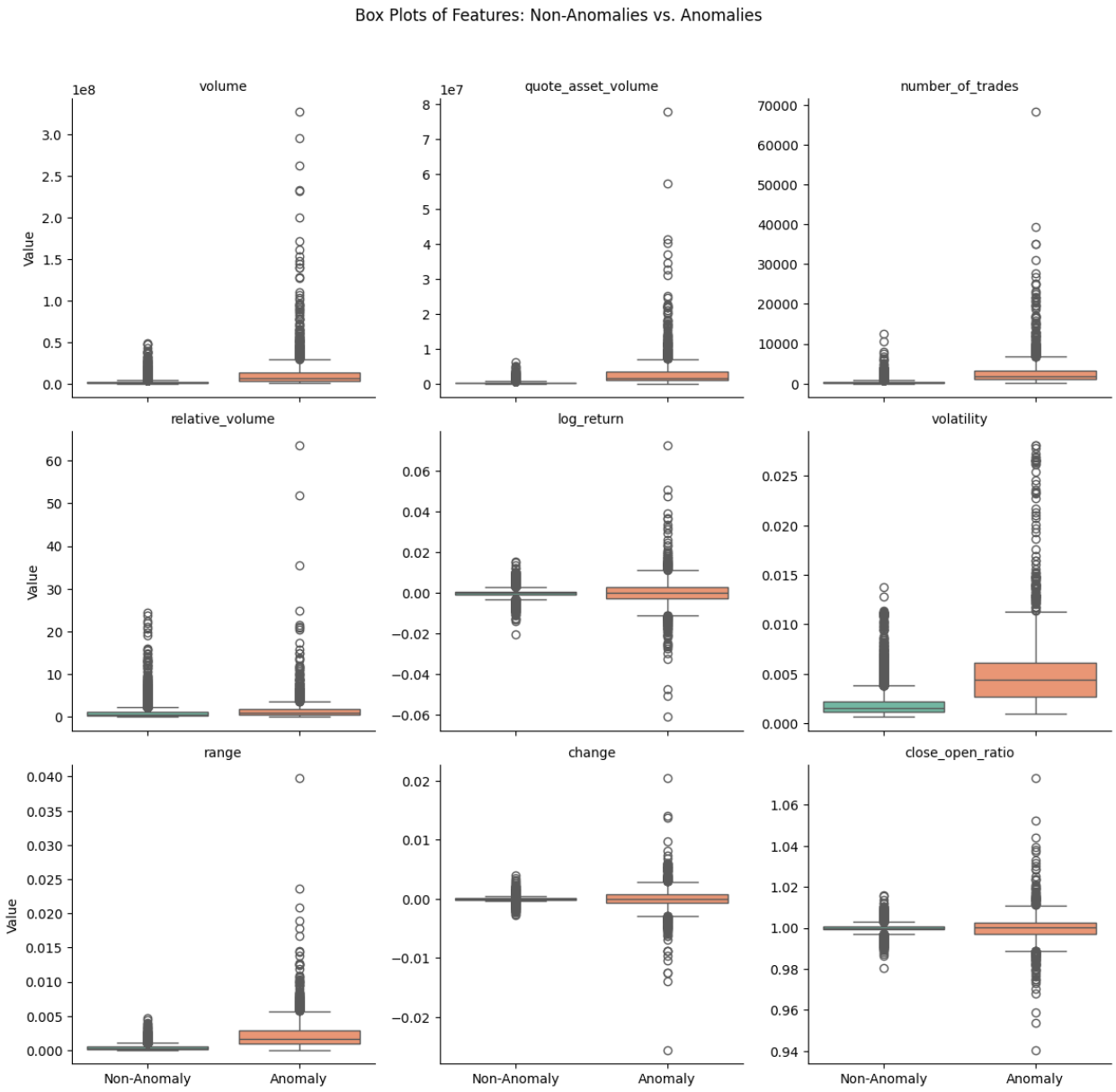


Appendix 9: Anomalies PCA Projection - Validation Data (2021)

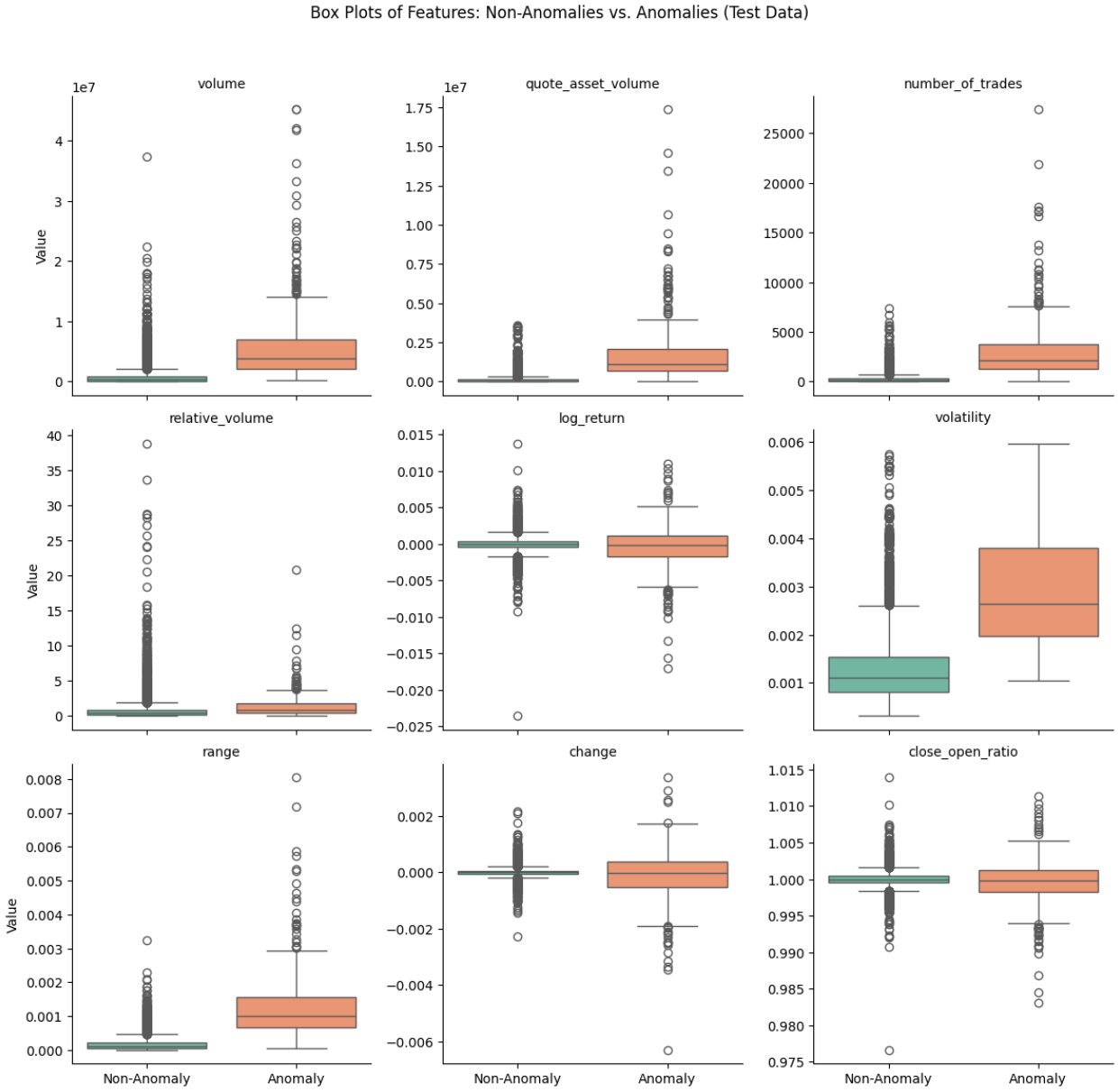


Appendix 10: Anomalies PCA Projection - Test Data (2024)

Appendix F



Appendix 11: Box Plot of Features - Validation Data (2021)



Appendix 12: Box Plot of Features - Test Data (2024)