Data-Driven Insights: Real Estate, Crypto, and Stock Market Analysis

Yi Peng, Chenmeinian Guo, Xiao Lu, Weizhou Ding
New York University

Abstract
This report explores trends and insights across multiple domains using data analysis techniques to provide overall investment insights. For real estate, we analyzed a public dataset to investigate zero-sale transactions, category and neighborhood trends, and temporal patterns, identifying key growth areas and market behaviors through clustering. In the cryptocurrency and stocks domain, clustering and time series analysis were applied to uncover correlations between cryptocurrencies and industry stocks, aiding investment strategies. These analyses demonstrate the effectiveness of data-driven approaches in extracting actionable insights from diverse datasets and offer guidance for informed decision making in investment contexts.

Contents

1 Introduction

This analytic helps investors and analysts find links between cryptocurrencies and stocks, making it easier to diversify portfolios, manage risks, and boost returns. It also gives real estate market trends, outliers, and growth patterns for data-driven insights. It's all about giving clear insights for smarter investment decisions in today's fast-moving markets.

1.1 Data source

- AMEX, NYSE, NASDAQ stock histories
- NYC property sales
- 400+ crypto currency

- Yahoo Finance

1.2 Project Objectives

- Data Analysis: Collect and preprocess historical market data about the selected industries and cryptocurrencies.
- Technical Indicators: Calculate important technical indicators to better understand market behaviors.
- Cluster Analysis: Group similar stocks using clustering algorithms to identify patterns.
- Correlation Study: Compare assets with major market indices to assess correlations.
- Investment Advice: Formulate investment recommendations tailored to different risk profiles.

2 Technology Stocks

Using clustering to get correlations

2.1 Introduction

In this part, we cluster technology stocks to find correlations between different stocks. The data set comes from Kaggle and contains historical data of more than 8000 stocks on the market.

We extracted all stocks in the technology sector, integrated their data into the same dataframe, and removed illegal data. In order to overcome the impact of stock prices on clustering, we normalize the data and use the rate of change instead of stock prices. Finally, we use k-means[1] to cluster stocks and divide them into 5 categories, using the date as the dimension

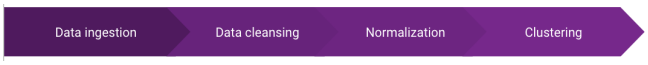


Figure 1: Clustering steps

2.2 Processed Data

After processing the data, we got the rate of change of technology stocks from 2015 to 2019. Considering that this is a long-term analysis, we used the rate of change every 5 days to reduce volatility

2.3 Clustering

The clustered stocks are shown here. We picked 3 stocks from each cluster for display. It is worth noting that cluster 4 only has one stock, the Research Solution company. This is an outlier. I think this is mainly due to It changes so much.

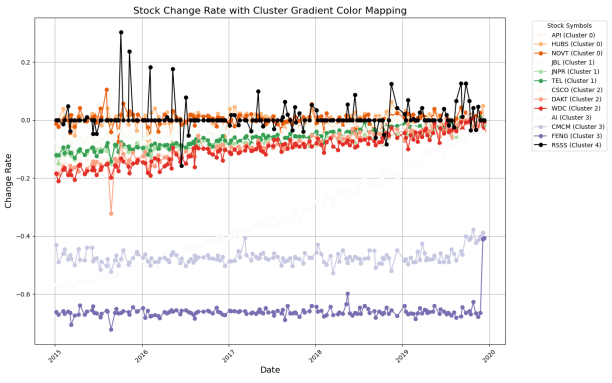


Figure 2: Clustering results

2.4 Correlations

Then we calculated the correlation between these stocks, and we can find that basically the changes in the same type of stocks show a high positive correlation. For example, JBL represents the Jabil; the TEL represents the TE Connectivity. Both companies are deeply

involved in electronics.
For those companies whose businesses are not as similar, they may also have similar trends. For example, Cisco and Western Digital also have a high positive correlation. Perhaps this means that Cisco and Western Digital have some cooperative relationships

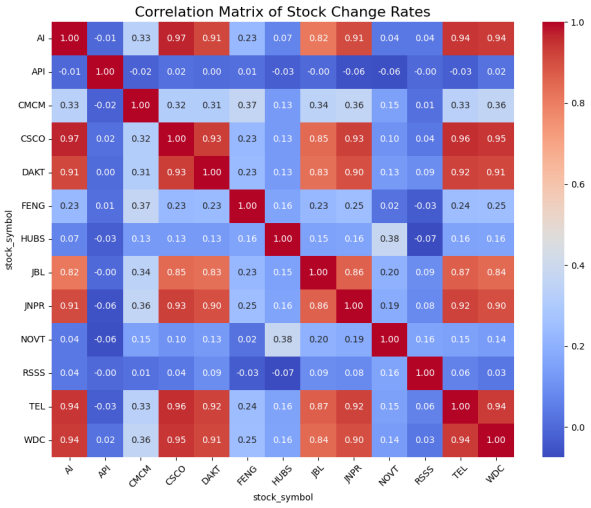


Figure 3: Correlation

3 Real Estate

Analyzing NYC Property Trends and Market Clusters

3.1 Introduction

This section analyzes NYC property sales using a public Kaggle dataset, focusing on key features such as sale price, property category, neighborhood, and transaction date. The dataset highlights market trends, outliers, and growth patterns for data-driven insights.

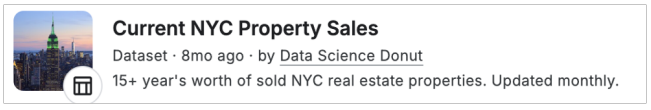


Figure 4: Dataset Source

3.2 Zero-Sale Analysis

Approximately 27.8% of property sales had a sale price of zero, representing non-standard transactions like inheritances or donations. Analysis showed consistent zero-sale cases across years, with Midtown West and Flushing-North having the highest counts.

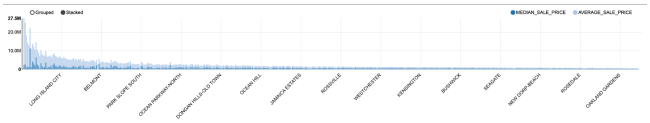


Figure 5: Zero-Sale Transactions Across Neighborhoods

3.3 Category and Neighborhood Insights

Analyzed property categories and neighborhoods revealed high-value markets such as Midtown West, Flushing-North, and premium categories like theaters and commercial condos. Insights identified areas with robust market activity and growth.

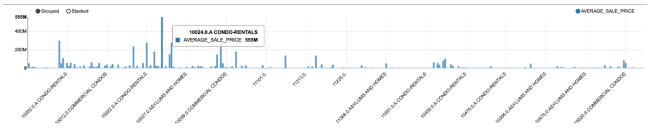


Figure 6: Category Analysis

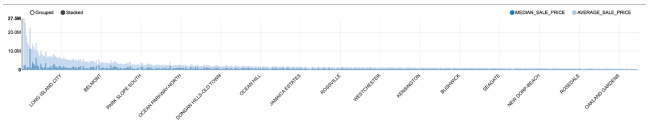


Figure 7: Neighborhood Analysis

3.4 Temporal Analysis

Monthly and quarterly trends highlighted seasonal market behaviors. December exhibited the highest average sale prices, which is counterintuitive as year-end discounts are typically expected; however, properties in December turned out to be the most expensive. June and August led in transaction volumes. The quarterly analysis showed that Q4 had the strongest average sales.

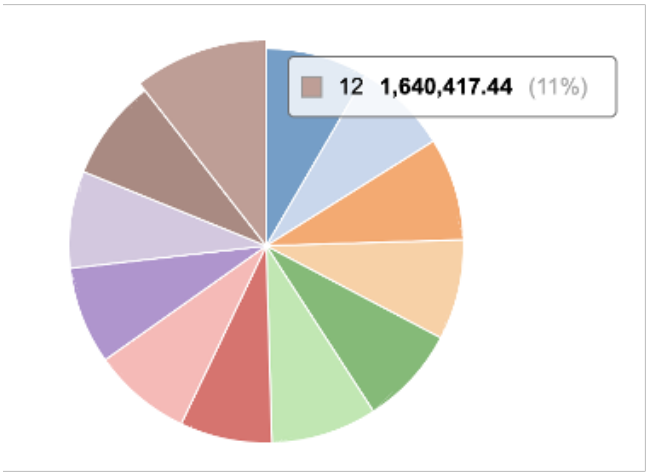


Figure 8: Monthly Trends

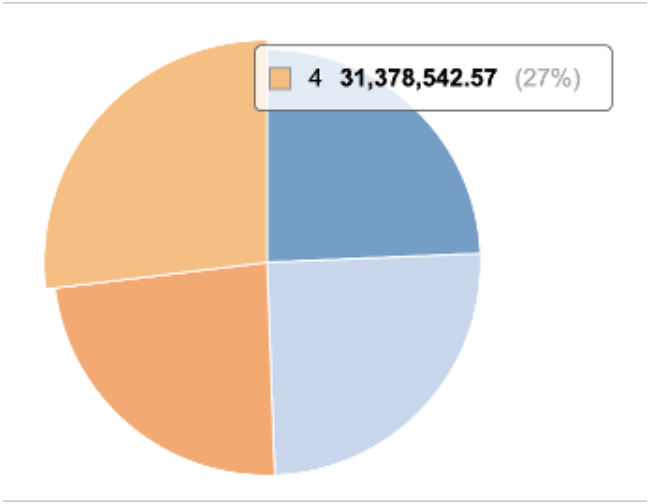


Figure 9: Quarterly Trends

3.5 Clustering Analysis

Grouped properties into five clusters based on sale price, growth rate, and total sales using the K-Means algorithm. High-growth clusters were identified, showcasing promising categories and neighborhoods. PCA-based visualization highlighted clear segmentation.

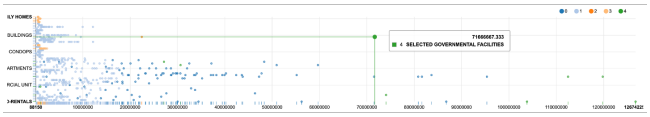


Figure 10: Clusters Based on Sale Price

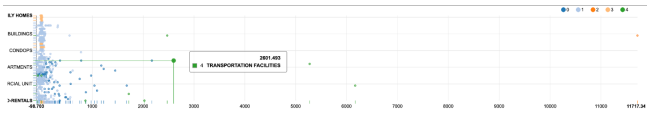


Figure 11: Clusters Based on Growth Rate

4 Cryptocurrency

Analyze and Prediction for Cryptocurrency Price

4.1 Introduction

Cryptocurrency price prediction is a challenging task due to the market's volatility. This part used two Kaggle datasets to forecast prices via data cleaning,exploratory data analysis and models like Random Forest, Gradient Boosted Trees, and a custom time series model. Their results are compared to choose the best model for the baseline, and the result highlights machine learning's potential for improving accuracy.

4.2 Related Work

Cryptocurrency prices are highly volatile, making prediction a complex task[2]. Machine learning models, utilizing historical price data (open, close, high, low), have proven effective for forecasting trends and supporting trading strategies.

Since our goal is to make a prediction model, we can consider this problem as a regression problem and due to the uncertainty of price movements, the data is considered as nonlinear. So the model we can use first is Random Forest and Gradient Boosted Trees.

Random Forest[3] is an ensemble learning method for regression and classification. By building multiple decision trees and averaging predictions, it minimizes overfitting and enhances accuracy. This algorithm handles non-linear data effectively, processes large datasets, and ranks feature importance, making it ideal for cryptocurrency price prediction.

Gradient Boosted Trees(GBT)[4] builds decision trees sequentially, with each tree correcting previous errors using gradient descent. It excels in capturing non-linear relationships and is highly accurate for regression tasks. GBT's flexibility and performance make it a popular choice for volatile cryptocurrency datasets.

ARIMA is a time series model. It is suitable for linear trends but struggles with non-linear, volatile datasets like cryptocurrencies. Despite limitations, it remains a benchmark in time series forecasting.

4.3 Experiment

I followed the step in figure12 to design the whole experiment process.

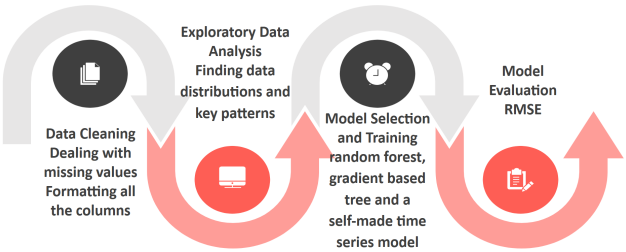


Figure 12: Dataset Source

In the data cleaning process, I filled missing values in each column with 0, and rows with null values and duplicates were removed. The UNIX timestamps were converted to a readable datetime format. For the other timestamps, I used a regular expression to do pattern match so that I can change to a standard version as well. I also calculated the standard deviation to remove outliners.

In the exploratory data analysis process, I calculated some key statistical metrics, including average, standard deviation to have a basic understanding of the cleaned data. I also identified high-volatility periods so that I can aggregate Minute-level trends to gain insights from average price range fluctuations and total trading volume. Additionally, I also used the correlation between BTC and ETH price ranges to try to figure out the interdependence of each Cryptocurrency.

Then I trained the Random Forest and GBT model by using the model provided by Spark.ml package and a self-made time series model. I use lagged features incorporate past prices as inputs, enabling the model to capture time series autocorrelation and predict future prices effectively.

In order to compare the result of the training process, I introduced Root Mean Square Error (RMSE) as the most important metric. RMSE is a commonly used metric to measure the differences between predicted and actual values. It is defined as:

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

where y_i is the actual value, $\hat{y}_i$ is the predicted value, and n is the number of observations.

4.4 Result

Because the training set is divided randomly, for the purpose of comparing the final results, I align the output results with the most recent dates of the actual data. So that I can show the results of the three models on a single graph.

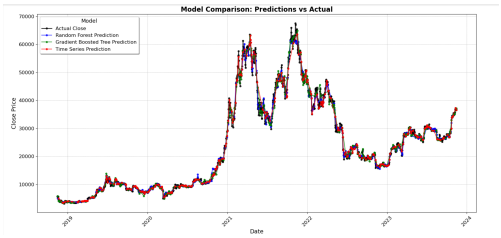


Figure 13: Enter Caption

Figure 13 compares the predictions of my models. From the chart we can see that all models capture the overall trends effectively. For example, there is a peak in 2021 with significant price fluctuations, our model captures that very well. However, our models show some differences for minor changes in stable periods. Overall, the Gradient Boosted Tree model shows the closest alignment with the actual values, while the Random Forest model is slightly less than GBT. The time series model also demonstrates robust performance over longer time periods while with some more differences.

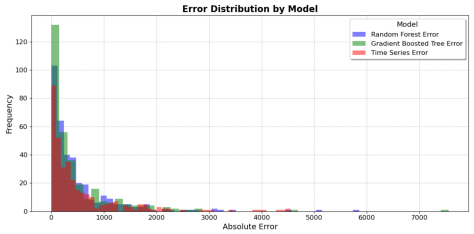


Figure 14: Enter Caption

Figure 14 shows the detailed absolute error distribution for each model. From it we can see that most of the errors are concentrated in a smaller range, indicating that the models fit the data well and generally provide reliable predictions. However, the Time Series model has a more concentrated error distribution with a higher peak. All in all, even though the GBT model compares better than two other models, there are still some higher errors in some cases, likely due to overfitting to local features. The Random Forest model has a more evenly spread error distribution.

However, we have to take potential overfitting into consideration since all models demonstrated similar prediction trends. For further improvement, I will try to explore deep learning models like LSTM.

5 Energy Stocks

Using Return and Volatility

5.1 introduction

Energy stocks play a crucial role in global financial markets due to their significant impact on both macroeconomic and microeconomic levels. These stocks represent companies involved in the production, distribution, and

commercialization of energy resources, such as oil, gas, and renewable energy. With the background that the Trump administration is going to reduce regulation of energy industries, which migh introduces a favorable environment for these companies to expand operations, innovate, and potentially yield higher returns for investors.

This mainly focuses on using data analysis techniques to evaluate historical performance, identify patterns, and propose an optimized investment strategy that maximize returns while minimizing risks.

5.2 Data Source, Cleaning and Pre-processing

The dataset initially consists of a list of all stocks currently traded on the American stock market. From this list, I filtered stocks classified under the energy sector and extracted their ticker symbols. After cleaning the data, I used these ticker symbols to download daily price data from Yahoo Finance. Finally, I created a new dataset where the rows represent dates, and each column corresponds to a different stock.

5.3 Annual Return and Volatility

To evaluate the performance and risk of energy stocks, we first calculated daily returns based on the percentage change in stock prices: $R(t) = \frac{P(t)-P(t-1)}{P(t-1)}$.

Next, we applied a 14-day sliding window to calculate the moving average of daily returns to smooth out short-term fluctuations and better capture trends: 14-Day Return at Day $t = \frac{1}{14}\sum_{i=t-13}^{t} R(i)$.

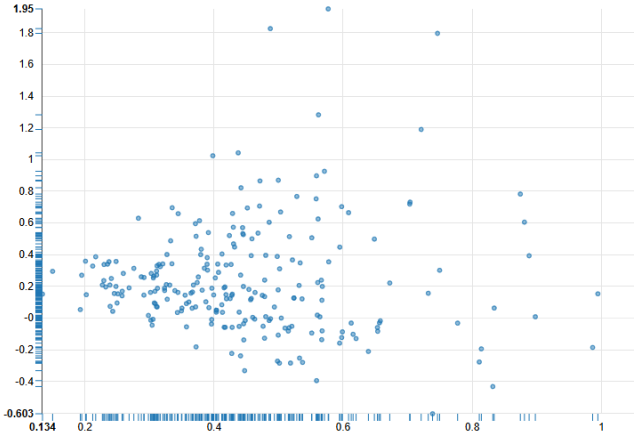
This method reduces noise and provides a more reliable measure of return. The smoothed daily return was then annualized to estimate long-term profit potential: Annual Return $= (1 +$ Average Daily Return$)^{252} - 1$.

Here, 252 represents the typical number of trading days in a year. Similarly, we calculated volatility using the 14-day standard deviation of daily returns: 14-Day Volatility at Day t $= \sqrt{\frac{1}{14}\sum_{i=t-13}^{t}(R(i) - \bar{R})^2}$.

This daily volatility was scaled to an annual level: Annual Volatility = Daily Volatility $\times \sqrt{252}$.

This approach ensures that both return and risk metrics are robust and reflective of the underlying trends in stock performance.

This following graph is a demonstration of part of stocks' return and volatility distribution.



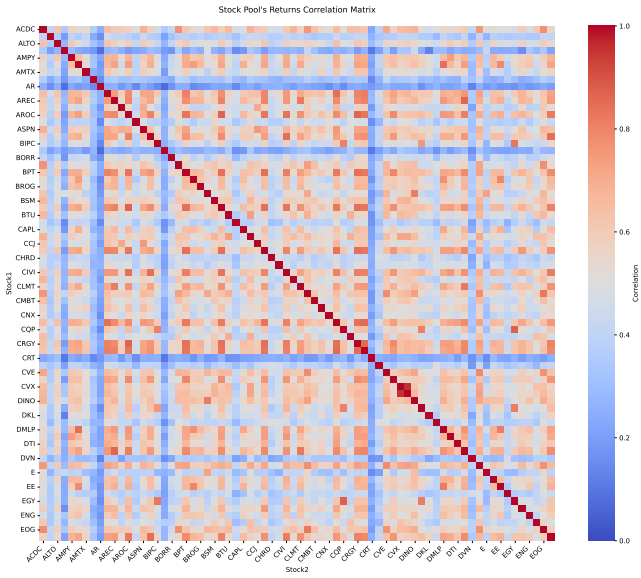
5.4 Threshold Filtering for stock pools

To enhance portfolio performance, I select stocks that fall within the top 20th percentile for returns while maintaining volatility below the median. This approach avoids the common pitfall of relying solely on historically best-performing stocks, as the stock market's inherent volatility means past performance is not always indicative of future results. By focusing on stocks outperforming the 20th percentile, we target fundamentally strong stocks with growth potential. Similarly, excluding stocks with the lowest volatility ensures that we account for the risk-return trade-off in financial markets. In essence, reasonable levels

of volatility are acceptable, as higher returns often accompany moderate risk.

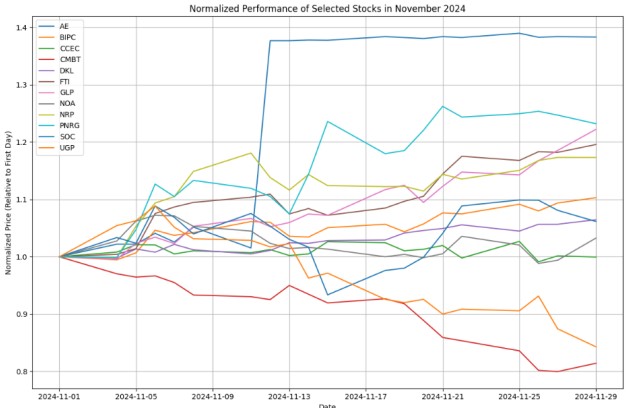
5.5 Greedy Algorithm and Covariance Analysis

To construct a well-diversified portfolio that balances maximizing returns and minimizing risks, we applied a greedy algorithm. The process begins by randomly selecting stocks from the predefined stock pool. Stocks are evaluated and sequentially added to the portfolio one by one. However, before including a new stock, we calculate its covariance with all previously added stocks in the portfolio. If the covariance of the candidate stock exceeds a threshold of 0.3 with any existing stock, the candidate is rejected. This step is crucial for maintaining diversification and avoiding over-concentration in correlated assets, which could amplify portfolio risk. By carefully managing the covariance between stocks, we aim to achieve a portfolio with a more robust risk-return profile. Since the order of how we add the stock might influence the final result, we redid the process 50 times and select the highest return one based on the historical data.



5.6 Results

The final portfolio consists of 12 carefully selected stocks. To evaluate its performance, I constructed the portfolio using historical data up to October 31, 2024. Then, I tested its real-world effectiveness using newly available data from November 2024. The figure below illustrates the normalized return of each stock in the portfolio over this period, providing a clear view of its performance under actual market conditions beyond the training phase.



We can observe from the graph that 10 out of the 12 selected stocks generated positive returns, indicating they contributed to the portfolio's overall profitability. Only two stocks experienced a decline in value during the testing period. This result highlights the effectiveness of the stock selection strategy, demonstrating that the effectiveness of the way of constructing the portfolio's components, even in a volatile market environment.

6 Conclusions

After analyzing technology stocks, real estate, cryptocurrencies, energy stocks, we also conducted a joint analysis of stocks and cryptocurrencies. Based on the results of clustering, we randomly selected some stocks and analyzed stocks and cryptocurrencies in 2018. Data between 2019 and 2019 were analyzed for correlation. And we came to the following conclusions.

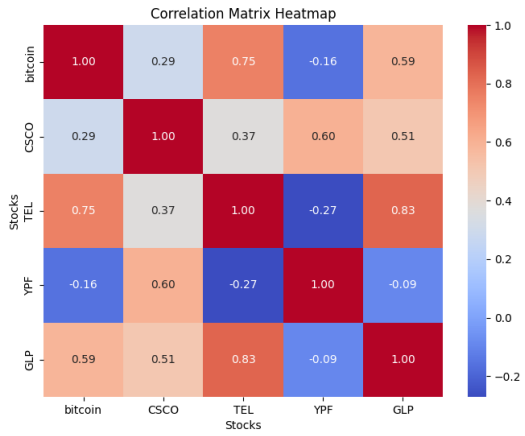


Figure 15: Correlations between stocks and bitcoin

- For technology stocks: After conducting correlation analysis, we found that the stocks of Cisco and Western Digital showed a high correlation, although they are involved in different fields, which may indicate that the two companies have deeper cooperation
- For stocks and cryptocurrency: Bitcoin and TEL stock prices exhibit high positive correlation

References

[1] Jiun Yen. Stock diversity analysis 1. https://www.kaggle.com/code/qks1lver/stock-diversity-analysis-1, 2018.

[2] Satoshi Nakamoto. Bitcoin: A peer-to-peer electronic cash system. https://bitcoin.org/bitcoin.pdf, 2008.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.