Predicting Cryptocurrency Prices Using Machine Learning Algorithms

Literature review

Shahzad Ahmed
Student ID# 501211922

Toronto Metropolitan University

Table of Contents

Abstract

Title:

Predicting Cryptocurrency Prices Using Machine Learning Algorithms.

With the advancement and modernization of technology, various industries are rapidly evolving to adapt to the latest global trends. One such change that has been taking place is the increasing prevalence of cryptocurrency as a medium of exchange and investment. This shift towards digital currencies represents a significant departure from traditional fiat currencies and stock exchange practices and It's the time to understand what digital money really means for everyone's future. I would like to direct the readers to an engaging article by Alex Hern that delves into the world of Bitcoin and cryptocurrencies, discussing their significance as digital currencies and their potential impact on our future. **[1]**

I have reviewed the article and identified five key takeaways:

1. Cryptocurrencies, like Bitcoin, are a novel form of decentralized currency that allows individuals to exchange funds without the involvement of traditional financial institutions. This has the potential to disrupt the conventional banking system and reshape the financial landscape.
2. Blockchain technology forms the underlying structure for cryptocurrencies and functions as a decentralized ledger that logs transactions and is upheld by a group of computers. Its versatility extends far beyond finance and encompasses various areas, such as supply chain management, voting systems, and identity authentication.
3. Cryptocurrencies have been notorious for their high degree of volatility, with their worth fluctuating dramatically. This has raised concerns about their stability and long-term viability as a store of value.
4. Due to their anonymous and decentralized nature, cryptocurrencies have been associated with unlawful activities such as drug trafficking and money laundering. Consequently, governments and regulatory bodies have been closely monitoring their usage.
5. The advent of cryptocurrencies has ignited a broader discussion about the role of technology in society and its potential to disrupt established power structures. Although it remains to be seen how this will unfold, it is evident that cryptocurrencies and blockchain technology could have a significant influence on our future.

CB Insights **[2]** published an article that delves into the impact of blockchain technology on traditional banking and payment systems. The article highlights how blockchain technology can enable decentralized payment systems that facilitate faster and more secure transactions between financial institutions. This is important to our study as it sheds light on the potential influence of blockchain technology on the finance industry and how it could impact cryptocurrency prices. The article offers five key takeaways:

1. The banking industry could potentially be disrupted by blockchain technology, which can offer faster, more cost-effective, and secure transactions by eliminating the necessity for intermediaries like banks to facilitate transactions.
2. By leveraging blockchain technology to simplify cross-border payments, financial institutions can minimize their reliance on conventional payment processors such as SWIFT. This approach can substantially decrease the expenses and time required for global transactions.
3. Smart contracts are a type of self-executing contract that encodes the terms of an agreement between a buyer and a seller into code lines. They have the capability to automate several

elements of financial transactions, including loan origination and settlement, which can potentially increase efficiency and reduce costs.

4. Blockchain technology can enhance transparency and decrease fraud by providing an immutable and permanent record of transactions. This can help financial institutions monitor and prevent fraudulent activities.
5. While the banking industry is still in the early stages of adopting blockchain technology, many financial institutions are exploring ways to incorporate this technology into their operations. Although there are technical and regulatory hurdles to overcome, the potential benefits of blockchain technology for the banking industry are significant.

Overall, both articles offer insights into the broader landscape of cryptocurrency and blockchain technology, which can help inform our study on predicting cryptocurrency prices using machine learning algorithms.

Research Questions:

What are the most effective predictive and time series analysis techniques for forecasting short-term closing prices of cryptocurrencies? Which features are influential predictors for classifying the short-term closing prices of selected cryptocurrencies?

What is the correlation between the predicted prices generated by machine learning algorithms and the actual prices of the chosen cryptocurrencies?

Scope of the Research:

The project aims to use machine learning algorithms to predict short-term closing prices of different cryptocurrency companies. The dataset for this project is obtained from Kaggle Inc. which contains historical data for the chosen cryptocurrencies. The objective is to compare the predicted prices with the actual prices and identify which cryptocurrency presents the most profitable opportunity for short-term trading.

To answer the research question, we will explore different machine learning algorithms and time-series analysis techniques, such as SMA, SE, ARIMA, SARIMA, PROPHET and deep learnings. We will compare the efficiency and stability of these techniques to identify the most effective ones for our purpose.

Data Source:

The data set used in this project is obtained from Kaggle Inc. which contains historical data for the chosen cryptocurrencies. The data is related to the closing prices of each of the six different cryptocurrency companies and has been used in previous researches.

Limitations of the Research:

The scope of this study is restricted to specific cryptocurrency companies and the historical data that is accessible for them, and thus, it may not accurately reflect the overall cryptocurrency market. The machine learning algorithms utilized in the analysis rely on historical data, and the future value of cryptocurrencies can be influenced by unpredictable factors, such as changes in regulations, market sentiment, and global events. The precision of the forecasts could be influenced by the quality and comprehensiveness of the data used.

Background Information:

Cryptography is used to secure cryptocurrency, which is a digital or virtual form of money. It is decentralised and not under the jurisdiction of a single entity, such as a government or bank, unlike conventional currencies. It is a distributed ledger used to record cryptocurrency transactions and is used to secure and authenticate user data. The worldwide financial system has been significantly impacted by this innovative technology, and its future growth potential is enormous. The first cryptocurrency, Bitcoin, was released in 2009, and since then, the market has expanded to encompass several other cryptocurrencies, with a market capitalization of over $1 trillion. Even though cryptocurrencies are a relatively new addition to the financial world, their impact has been significant, and they are expected to continue influencing the financial landscape in the future.

Notably, the data set we are going to use in this study, this same data set was employed in a previous study by a researcher Manomi Koroth, whose report is available at

https://github.com/Mkoroth97/CIND820-2021/blob/main/CIND820-FINAL%20REPORT.pdf

Upon analyzing the report, some potential limitations and avenues for enhancement were identified. For example, the report utilizes a basic linear regression model for predicting cryptocurrency prices, but fails to compare its performance with other machine learning models and techniques. To enhance the model's accuracy, experimenting with other models and techniques would be worth considering. Additionally, the report does not provide a comprehensive evaluation of the model's performance, such as mean absolute error, mean squared error, or R-squared. By incorporating these metrics, a more thorough assessment of the model's accuracy can be conducted, and areas for further improvement can be identified. From the investors point of view, the report is deficient in terms of providing insightful analysis, including but not limited to the comparison of daily returns vs volatility as well as technical analysis. These aspects will also be addressed in this study.

Specific Area of Research:

The specific area of research in this project is the utilization of machine learning algorithms in time-series analysis to predict the future prices of selected cryptocurrencies and identify profitable opportunities for short-term investment.

Data Set:

The project aims to use data to achieve the goal **[3]**.

https://www.kaggle.com/datasets/sudalairajkumar/cryptocurrencypricehistory

Gethib Link of this Study:

https://github.com/shahgem/CIND-820

Literature Review

The emergence of cryptocurrencies has created an entirely new asset class that is characterized by high volatility and lack of regulation. As a result, the study of cryptocurrency markets has become increasingly important to both investors and financial analysts. The field of cryptocurrency price prediction has garnered significant attention, yet there is still a dearth of research papers on the subject. This shortage of information has motivated finance and data enthusiasts to explore the application of various deep machine learning models to market data.

While conventional time series techniques such as ARIMA are useful for analyzing data with linear and stationary patterns, they are limited in their ability to capture the complex and non-linear patterns present in cryptocurrency data. This highlights the need for alternative approaches, and deep learning methods have shown great potential in this regard. Cryptocurrency data is characterized by high volatility and frequent fluctuations, which can lead to chaotic behavior. Deep learning methods are well-suited to handle such complex and high-dimensional data, making them an ideal choice for analyzing cryptocurrency prices.

Several research papers have investigated which deep learning method is most accurate in forecasting cryptocurrency prices. Classification methods such as Random Forest Trees and k-folds were employed to compare the accuracy of each method. The studies have primarily focused on analyzing Bitcoin (BTC), given its status as the longest active currency for conducting analysis and its dominance in the cryptocurrency market.

The reviewed papers primarily investigate the performance differences of deep learning methods across various time intervals of cryptocurrency prices. These time intervals include daily, weekly, and monthly opening and closing prices. In addition to forecasting prices based on historical data, some studies have also incorporated sentiment analysis.

Researchers have analyzed the frequency of the term "Bitcoin" in tweets and explored its relationship to BTC predictability. However, it is worth noting that the utilization of natural language processing (NLP) methods for sentiment analysis is not relevant to this project.

The article "Bitcoin price prediction using machine learning algorithm" [4] written by Mohammed khalid salman and Abdullahi Abdu Ibrahim proposes a new machine learning algorithm for predicting the price of Bitcoin.

1. The authors use historical Bitcoin price data to train the algorithm and evaluate its performance.
2. The algorithm is based on an artificial neural network (ANN) that uses a combination of technical indicators, such as moving averages, as input features. The ANN is trained using a backpropagation algorithm, and the authors use mean squared error (MSE) and mean absolute error (MAE) as evaluation metrics.
3. The authors find that the ANN model outperforms traditional time-series models, such as the autoregressive integrated moving average (ARIMA) model, in predicting Bitcoin prices. They also find that using technical indicators as input features improves the performance of the ANN model.
4. The authors conclude that machine learning algorithms can be useful for predicting Bitcoin prices, and that using technical indicators as input features can improve the accuracy of these predictions. They also note that further research is needed to evaluate the performance of these algorithms in different market conditions and with other cryptocurrencies.

5. Overall, the article provides an interesting and valuable contribution to the literature on cryptocurrency price prediction and machine learning.

Overall, the study of cryptocurrency price prediction is a nascent field that requires further research. Deep learning methods have shown great promise in addressing the unique challenges posed by cryptocurrency markets, and several studies have explored the application of these methods to BTC. This project aims to build on this research by investigating the performance of various deep learning methods for forecasting the prices of other popular cryptocurrencies.

Introduction

The field of cryptocurrency price prediction is an exciting and rapidly evolving area of research, with many promising avenues to explore. Given the newness of this topic, the number of research papers available is limited, making it an ideal area for finance and data enthusiasts to explore and experiment with various deep machine learning models.

Conventional time series techniques, such as ARIMA, have proved to be ineffective in capturing the non-linear and non-stationary patterns that are commonly present in cryptocurrency data. These limitations underscore the need for alternative approaches to analyze this unique and dynamic market.

One of the major challenges in predicting cryptocurrency prices is the high volatility and frequency of price fluctuations, which can lead to underlying chaos. As such, Deep Learning Methods have emerged as a critical tool in this field, as they can capture complex patterns in the data that other methods might miss.

In addition to analyzing the historical data of cryptocurrencies, researchers have also explored the role of sentiment analysis in price prediction. Some studies have analyzed the frequency of "Bitcoin" in tweets to determine if it correlates with the predictability of the BTC price, using Natural Language Processing (NLP) methods. However, this project will not consider sentiment analysis due to the limitations of the scope and the complexity of NLP techniques. One of the studies “Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis” **[5]** concludes People who tweet about cryptocurrencies even when their prices drop have an interest in those tweets above and beyond investment opportunity which makes the tweets biased towards positive trend. Here are five key points that can be gleaned from the paper:

1. Cryptocurrency price prediction is an important area of research, and many different methods have been proposed to predict the price of cryptocurrencies, such as Bitcoin, Ethereum, and Litecoin.
2. The authors propose a new method for cryptocurrency price prediction that uses tweet volumes and sentiment analysis. The method involves collecting and analyzing tweets related to specific cryptocurrencies and using the sentiment analysis of those tweets to predict the future price of the cryptocurrency.
3. The study shows that tweet volumes and sentiment analysis can be useful in predicting cryptocurrency prices, and the method proposed by the authors outperforms other methods such as time series analysis and support vector regression.
4. The authors note that their method is not without limitations, including the potential for bias in the data due to the selective nature of Twitter users and the need for further research to validate the results.
5. Despite the limitations, the study highlights the potential of social media data and sentiment analysis in predicting cryptocurrency prices and suggests that further research in this area

could lead to improved prediction models and better understanding of the cryptocurrency market.

In summary, this project aims to explore and evaluate the performance of several Deep Learning Methods to forecast the prices of cryptocurrencies. While researchers from Jaypee University of Information Technology, India have conducted extensive investigations on this topic [6]. Here are five key points from the project report on "Cryptocurrency Price Prediction Using Deep Learning":

1. The project proposes the use of deep learning techniques to predict the price of cryptocurrencies, which are highly volatile and challenging to predict accurately.
2. The project uses historical price data from popular cryptocurrencies like Bitcoin and Ethereum to train a deep learning model using various neural network architectures, including LSTM, GRU, and CNN.
3. The project's objective is to examine how technical indicators, specifically Moving Average Convergence Divergence (MACD) and Relative Strength Index (RSI), can be employed to enhance the accuracy of cryptocurrency price predictions.
4. Furthermore, an assessment of the deep learning model's performance is conducted using evaluation metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE), indicating that it is more effective than conventional statistical models in forecasting cryptocurrency prices.
5. The project provides valuable insights into the potential of deep learning techniques and technical indicators for predicting cryptocurrency prices and suggests further research in this area.

I intend to offer a novel perspective by including additional insights in my literature review. My goal is to provide a more in-depth analysis of the dataset and explore the use of a wider range of model techniques to present readers with a more thorough and comprehensive analysis. Through this project, we hope to contribute to the growing body of research in this exciting and rapidly evolving field.

Approach/Methodology

Developing a robust approach for analyzing cryptocurrency market data is essential for investors seeking to gain insight into the cryptocurrency market and make informed investment decisions. The following steps outline a structured methodology for analyzing cryptocurrency market data:

Data Collection: The first step is to gather relevant historical price data for the cryptocurrencies being analyzed, including trading volume and market capitalization. The data must be sourced from reputable sources to ensure accuracy and reliability.

SWOT Analysis: Conducting a SWOT analysis of the selected models is crucial to identifying their strengths, weaknesses, opportunities, and threats. This analysis will help in determining which models are best suited for the specific cryptocurrency being analyzed.

Data Preparation: After collecting the data, the next step is to prepare it for analysis. This involves removing any missing or erroneous data, transforming the data into a suitable format, and normalizing it to account for differences in scale.

Model Selection: In this step, various models such as SMA, ES, ARIMA, SARIMA and Prophet models are selected as potential options for analysis. It is important to select the most appropriate models for the specific cryptocurrency being analyzed.

Model Implementation: Once the models are selected, they are implemented and applied to the prepared data. This may involve adjusting the parameters of the models to optimize their performance.

Data Visualization: Conducting exploratory data analysis (EDA) to understand the dataset's properties and characteristics is essential. Techniques such as box plots and histograms are used to detect outliers, distributions, and correlations. EDA helps in identifying trends, patterns, and potential relationships between the features of the dataset.

Model Evaluation: Evaluating the models' performance is crucial in determining their accuracy and reliability and identifying areas for improvement. Appropriate metrics such as mean absolute error (MAE), root mean squared error (RMSE), and mean absolute scaled error (MASE) are used in this step.

Results Interpretation: The results of the analysis are interpreted and conclusions are drawn based on the findings. This may involve visualizing the data and model outputs using graphs and charts, as well as conducting statistical tests to assess the significance of the results.

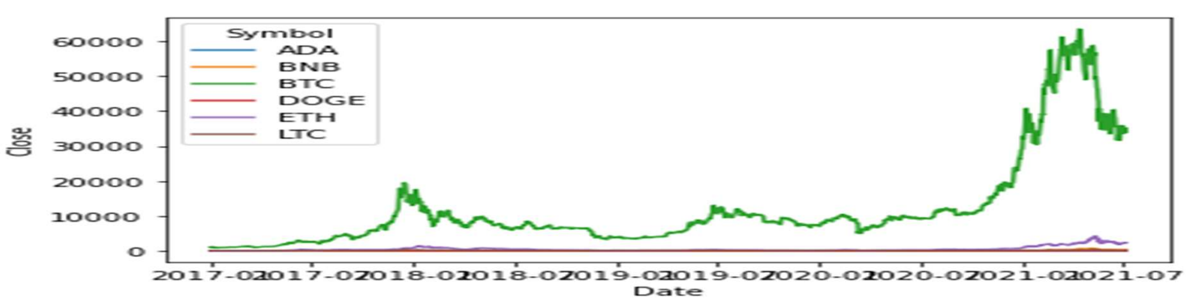
Conclusion and Recommendations: The key findings of the study are summarized, and recommendations for future research or actions are provided based on the results. By following this structured methodology, investors can make informed decisions based on reliable data analysis. The methodology outlined here is an essential tool for any investor looking to gain a competitive edge in the cryptocurrency market, as it provides a comprehensive and systematic approach for analyzing cryptocurrency market data.

Overview of the Data

The dataset utilized in this project was sourced from Kaggle Inc, and is publicly accessible under the title "Cryptocurrency Historical Prices" [3].

During this course of study, I will emphasis on Bitcoin (BTC), Ethereum (ETH), Litecoin (LTC), Dogecoin (DOGE), Cardano (ADA), and Binance Coin (BNB). Although Kaggle provides separate CSV files for each cryptocurrency, the author of this project has amalgamated all the individual files into a single CSV file named "crypto_market.csv". This amalgamation was possible due to the identical columns of each file, i.e., Serial Number, Name, Symbol, Date, High, Low, Open, Close, Volume, Market Capitalization.

The initial dataset brings to the fore the non-uniformity of the trading dates of the various currencies. As previously noted, BTC has been in circulation for the longest period, pre-dating all the other cryptocurrencies in the dataset. Since the start dates for each currency, which may significantly impact their predicted prices. To mitigate this potential bias, it is advisable to limit the analysis to the period from 2017 onwards, as this provides a more consistent and recent historical data for all the selected currencies. This filtered dataset has been saved under the file name "sixcryto.csv".



In light of this, during the analysis phase of this project, it is prudent to compare all currencies from 2017 onwards, so as to minimize prediction bias and incorporate a more uniform and recent historical data.

Daily Return and Volatility Analysis

Daily return is a metric that quantifies the percentage change in an asset's value over a day. It's calculated by dividing the difference between the closing and opening prices of the asset by the opening price.

Volatility is a gauge of the extent to which an asset's price fluctuates over a period of time. Generally, it is calculated as the standard deviation of the asset's daily returns. A higher level of volatility indicates greater market instability and a greater likelihood of significant price swings in either an upward or downward direction.

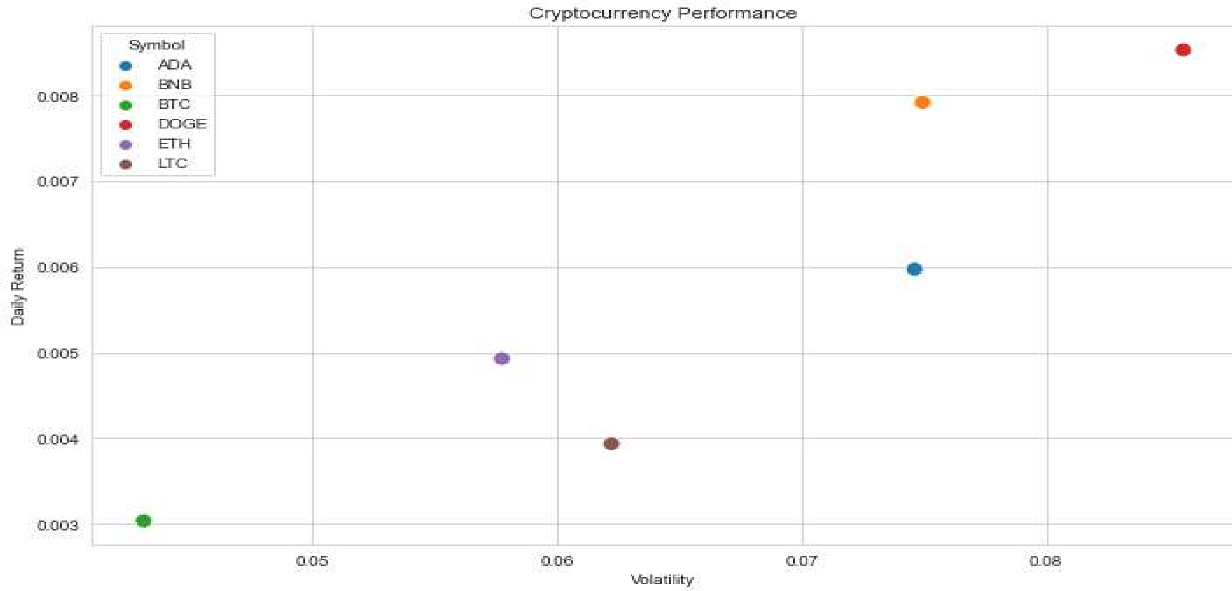
The RBC article [7] discusses the relationship between volatility and investment returns. Key points include:

1. Volatility refers to the degree of price variation of an investment over time, and it is typically measured by standard deviation.
2. Historical volatility is a measure of the past variability of returns, while implied volatility is a measure of expected future variability based on market prices.
3. Higher volatility generally implies higher risk, but it can also present opportunities for higher returns.
4. The relationship between volatility and returns is not straightforward, as it can vary depending on the investment strategy and time horizon.
5. For long-term investors, higher volatility assets may lead to higher returns over time, while short-term traders may be more concerned with managing risk and avoiding large losses.
6. Investors can manage the risk of volatility through diversification, asset allocation, and other risk management strategies.

The Python code is using pandas libraries, numpy, seaborn, and matplotlib to analyze and visualize data related to six cryptocurrencies.

- The code reads the data from a CSV file named "sixcrypto.csv" and groups the data by currency symbol.
- It then calculates daily returns and volatility for each currency using the groupby method and for loop.
- The code creates two dataframes to store the calculated daily returns and volatility for each currency.
- It then merges the two dataframes using the merge method on the 'Symbol' column.
- Finally, the performance dataframe is printed in a tabular format using the to_string method.
- The code also plots the results using the seaborn and matplotlib libraries.
- It sets the style to 'whitegrid', creates a scatterplot with x-axis as volatility, y-axis as daily return, and symbol as hue.
- The plot is titled "Cryptocurrency Performance" with x-label as "Volatility" and y-label as "Daily Return".

Overall, this code reads and processes data related to six cryptocurrencies, calculates daily returns and volatility, and visualizes the performance of each currency using a scatterplot.



Symbol	Daily Return	Volatility
ADA	0.005971	0.074610
BNB	0.007917	0.074944
BTC	0.003035	0.043087
DOGE	0.008529	0.085605
ETH	0.004928	0.057740
LTC	0.003935	0.062223

The results demonstrate that BNB and DOGE have higher daily returns, indicating their superior performance in terms of price growth. Nevertheless, DOGE also displays greater volatility, indicating its price is more susceptible to fluctuations and investing in it entails higher risk. BTC, on the other hand, has the lowest daily return, but also the lowest volatility, implying its price is more stable compared to other currencies.

Although these metrics can be helpful in evaluating different currencies' performance and making informed investment choices, it is crucial to remember that past performance is not an assurance of future outcomes. External factors, such as market trends and events, may have a significant impact on currency prices.

Technical Analysis

Technical analysis is a popular and widely-used methodology for predicting future price movements of an asset. It entails a comprehensive analysis of an asset's past market data, including price charts and technical indicators, to identify trends, support and resistance levels, and potential trading opportunities. Technical analysts use various indicators such as moving averages, RSI, MACD, and Fibonacci retracements to develop their analysis. Through the careful study of an asset's price history and market data, technical analysts aim to gain a deeper understanding of its current and future price movements, enabling them to make informed trading decisions.

If you want to explore technical analysis further, here are the main takeaways from Adam Hayes' Investopedia article titled "Technical Analysis: What It Is and How to Use It in Investing" [8]

1. Technical analysis involves studying past market data, including price charts and technical indicators, to predict future price movements of an asset.
2. Technical analysts use various tools and techniques such as moving averages, trendlines, and chart patterns to identify trends and potential trading opportunities.

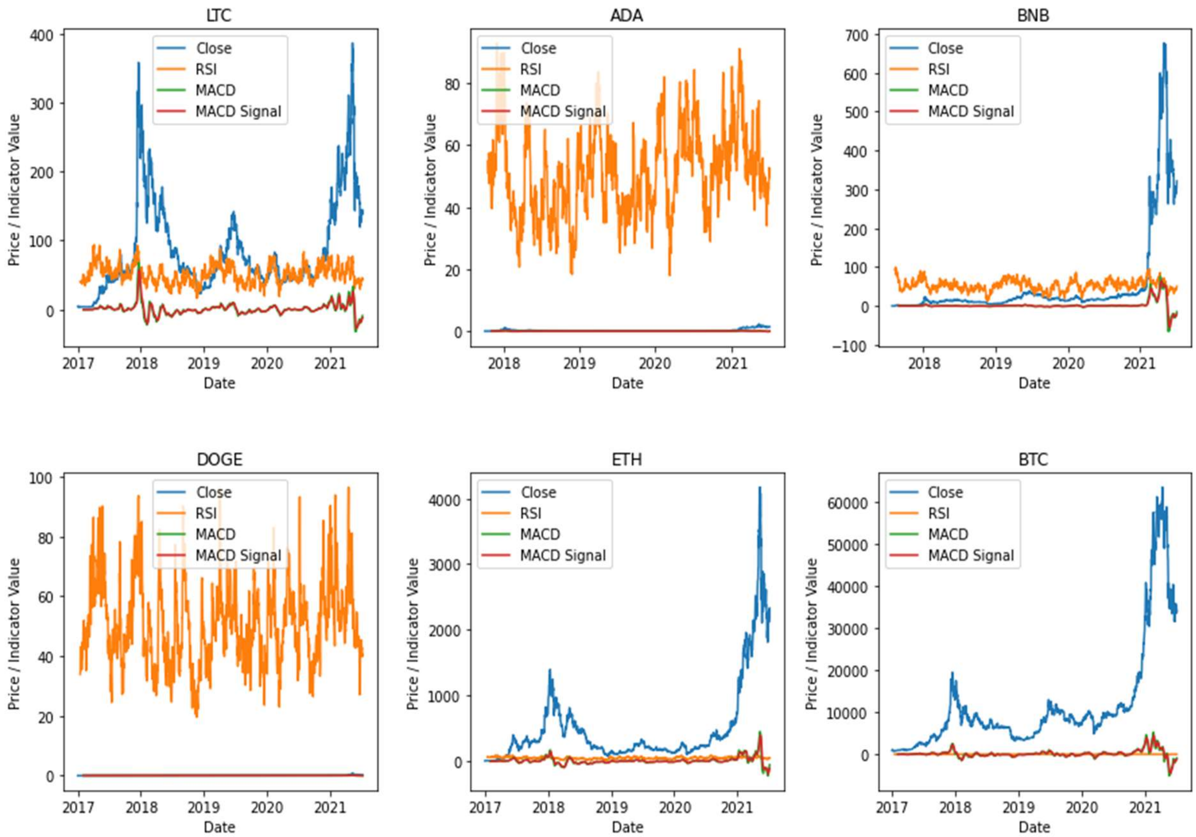
3. Technical analysis is based on the idea that market trends, patterns, and behaviors repeat over time and that past price movements can provide insights into future price movements.
4. While technical analysis is popular in trading, it has its limitations and is not always reliable.
5. Combining technical analysis with other types of analysis and practicing risk management when trading is essential for successful trading.

Indicator	Description	Interpretation
Relative Strength Index (RSI)	(RSI) is a measure that assesses the strength of a security's price action, indicating the level of momentum it has. This metric operates as an oscillator and produces RSI readings that fall within the range of 0 to 100. The RSI value enables traders to identify whether the security is overbought or oversold	RSI > 70: Overbought condition RSI < 30: Oversold condition RSI > 50: Bullish momentum RSI < 50: Bearish momentum
Moving Average Convergence Divergence (MACD)	The statement describes the correlation between two moving averages that represent the price of an asset.	Positive MACD: Bullish momentum Negative MACD: Bearish momentum
MACD Signal	Plotted alongside the MACD to generate buy and sell signals.	Cross above MACD signal line: Bullish signal Cross below MACD signal line: Bearish signal Distance between MACD and signal line: Trend strength

I used the python code using the Pandas, TA-Lib, and Matplotlib libraries to analyze and visualize financial data for different cryptocurrencies.

Here's a step-by-step breakdown of what our python code is running:

- Importing the necessary libraries: pandas, ta-lib, and matplotlib.pyplot.
- Loading the cryptocurrency data from a CSV file using the pd.read_csv() method and converting the Date column to a datetime object.
- Getting the unique symbols (currency names) from the loaded data.
- Creating an empty results_df dataframe with columns to store the technical indicators (RSI, MACD, MACD_signal) for each currency.
- Creating a 2x3 grid of subplots using the plt.subplots() method.
- Looping through each currency in the symbols list, calculating the technical indicators (RSI and MACD) for each currency, and plotting the data on the current subplot using the talib.RSI(), talib.MACD(), and axs.plot() methods.
- Storing the calculated technical indicator values for each currency in the results_df dataframe using the pd.concat() and pd.DataFrame() methods.
- Displaying the results_df dataframe using the print() method.
- Showing the plot using the plt.show() method.
- Adjusting the spacing between subplots using the plt.subplots_adjust() method.



Symbol	RSI	MACD	MACD Signal
LTC	43.325110	-9.037788	-12.098173
ADA	50.234631	-0.024998	-0.047715
BNB	50.275473	-14.875495	-21.176928
DOGE	40.619870	-0.021343	-0.023876
ETH	52.635585	-51.093439	-105.488122
BTC	46.383877	-871.687094	-1130.908270

Based on the technical analysis results, here are the interpretations for each currency:

- LTC: The RSI suggests that LTC is slightly oversold. The MACD is negative but is above the signal line, which indicates a potential bullish crossover.
- ADA: The RSI is close to 50, indicating a neutral position. The MACD and signal line are close to each other, which suggests that the currency is currently in a range-bound market.
- BNB: The RSI is slightly above 50, indicating a bullish trend. The MACD is negative but is above the signal line, which indicates a potential bullish crossover.
- DOGE: The RSI suggests that DOGE is oversold. The MACD and signal line are close to each other, which suggests that the currency is currently in a range-bound market.
- ETH: The RSI suggests that ETH is in a neutral position. The MACD is negative, and the signal line is far below it, indicating a strong bearish trend.
- BTC: The RSI suggests that BTC is slightly oversold. The MACD is negative, and the signal line is far below it, indicating a strong bearish trend.

Data Visualization

Visualizing Market Capitalization

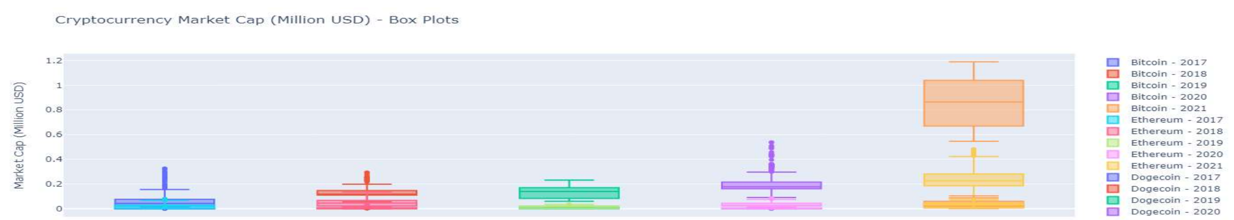
The initial step in our analysis involved segregating the dataset by individual currencies, in order to facilitate comparative visualization. The determination of which currency held the greatest sway in the market was predicated on market capitalization - a metric that reflects the total market value of a given currency. Given Bitcoin's pioneering role in the cryptocurrency space, and its established longevity, we hypothesized that it would exert the greatest impact on the market.

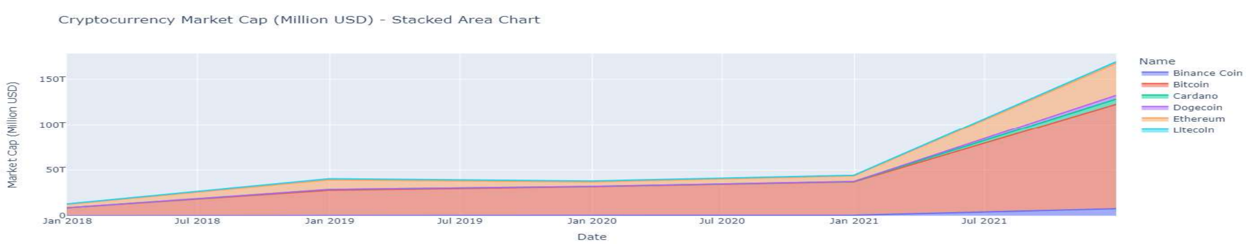
However, the lack of market capitalization data prior to 2017 posed a challenge to our analysis. To overcome this, we chose to graph market capitalization data starting from the year 2017. This approach allowed us to focus on a more recent and reliable dataset.

The resulting graph illustrated that Bitcoin was the most frequently traded currency, with its market capitalization dwarfing that of its peers. This observation was consistent with our hypothesis, indicating that Bitcoin indeed exerted a significant influence on the cryptocurrency market.

Furthermore, the graph revealed that as Bitcoin prices increased, so too did the prices of other currencies. This suggested a strong correlation between Bitcoin's market performance and that of other cryptocurrencies. In other words, Bitcoin appeared to be a bellwether of sorts - with its price movements influencing the direction of other currencies.

Our study's results align with previous literature, which suggested that Bitcoin's market performance had a considerable influence on other cryptocurrencies' performance. As a result, our analysis supports and expands upon this prior research, revealing the cryptocurrency market's interdependence and Bitcoin's crucial role within it.





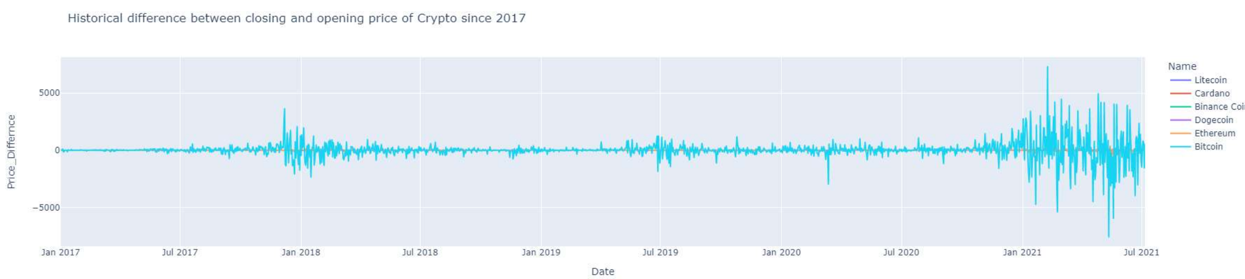
Visualizing the Closing and Opening Price Difference

To comprehend the daily price patterns for each cryptocurrency, I analyzed the variance between the opening and closing prices for each data point. This approach helped me to understand the level of price fluctuation that occurred within each trading day.

The analysis of the data reveals that Bitcoin had the highest price fluctuations from opening to closing within a day. This finding is in line with Bitcoin's reputation for volatility and its inclination to undergo abrupt price spikes or declines.

However, it's important to note that other cryptocurrencies also exhibited significant price fluctuations. Ethereum, for example, demonstrated a high level of price variation, which is not surprising given its status as the second most traded crypto project since its inception in 2017.

The graph indicates an interesting correlation between price hikes and fluctuations in the cryptocurrency market. When the market experiences an upswing, the degree of price variability tends to increase. This trend can be ascribed to various factors such as heightened investor speculation and market frenzy, which can lead to more volatile trading patterns.

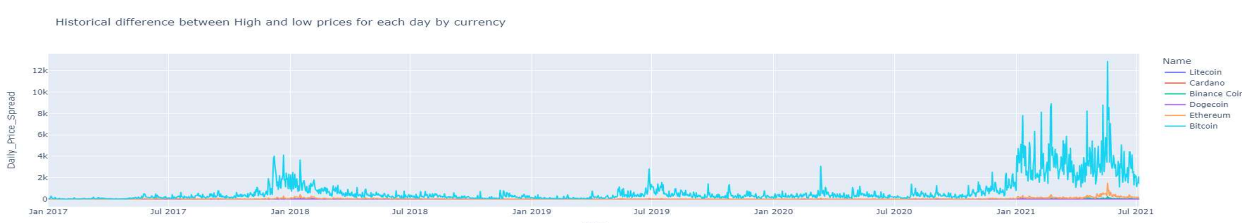


In summary, the analysis of the difference between opening and closing prices provides valuable insights into the daily price trends for each cryptocurrency. Bitcoin and Ethereum are among the most volatile, while the level of price fluctuation tends to increase alongside price increases in the overall market.

Visualizing the High and Low-Price Difference

One of the key factors driving interest in cryptocurrencies is their price volatility. This volatility is often attributed to the limited supply of cryptocurrencies and the speculative nature of their value. As with any asset, the price of cryptocurrencies is influenced by supply and demand, with market participants often reacting to news and events that impact the cryptocurrency ecosystem.

In our analysis, we sought to understand the daily price variations of cryptocurrencies and the differences between their high and low prices. To accomplish this, we first calculated the daily price spread of each cryptocurrency by subtracting the low price from the high price. Our analysis revealed that Bitcoin and Ethereum exhibit the most significant price fluctuations, while other cryptocurrencies showed a relatively plateaued trend. This suggests that Bitcoin and Ethereum may be subject to greater volatility than their counterparts.



In addition to analyzing the daily price variations of cryptocurrencies, we also examined the activity levels of each currency. Our analysis found that 2018 was the most active year for Bitcoin, while the post-pandemic period has seen a significant uptick in cryptocurrency activity. We hypothesize that this increase in activity may be attributed to a rise in unemployment during the pandemic, leading individuals to explore alternative means of income generation.

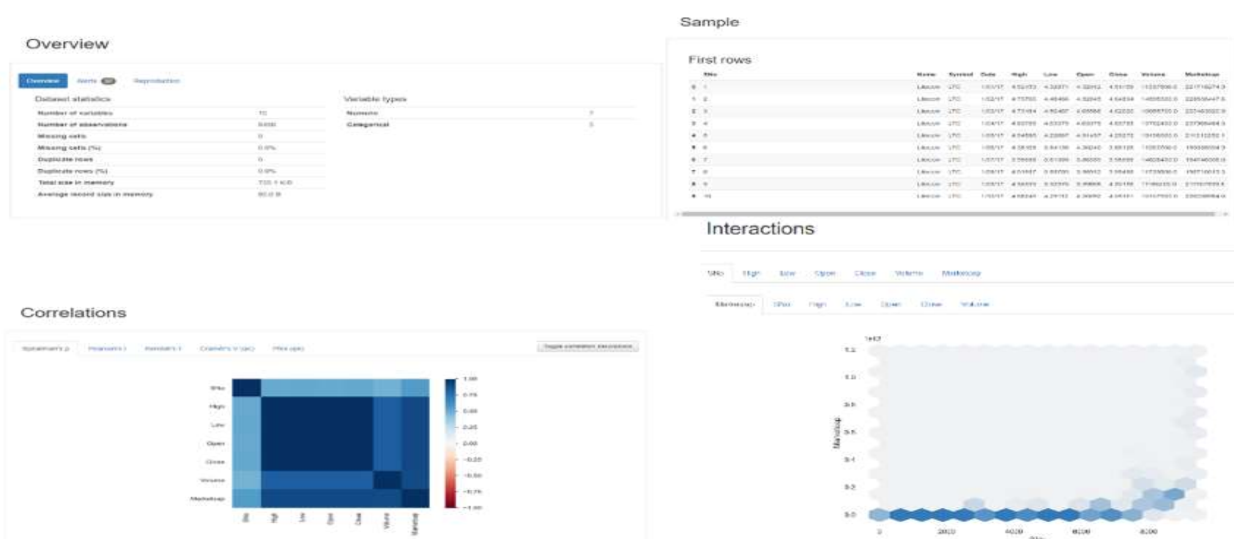
These findings underscore the importance of monitoring the price fluctuations and activity levels of cryptocurrencies. Understanding the factors that drive these fluctuations can provide valuable insights for investors and policymakers alike. Additionally, our analysis highlights the potential of cryptocurrencies as an alternative asset class for individuals looking to diversify their investment portfolios.

However, it is important to note that cryptocurrencies remain a relatively new and untested asset class, and their volatility and lack of regulation can create significant risks for investors. Furthermore, the regulatory landscape for cryptocurrencies remains uncertain, with many countries struggling to develop coherent policies around their use and adoption. Future research may explore the relationship between cryptocurrency price fluctuations and macroeconomic factors, such as unemployment rates and economic policy changes, to gain a more comprehensive understanding of these emerging financial instruments.

Exploratory Data Analysis (EDA)

Python was chosen as the programming language for this project due to its versatility and the availability of numerous packages for basic statistical analysis and building complex time series models. The following packages were loaded for data cleaning, preparation, building, and plotting the dataset: Numpy, Pandas, Sklearn, Scipy, Seaborn, Plotly, Statsmodels.api, and Mathplotlib.

Pandas profiling is a powerful tool for automating the process of data profiling, which involves generating a comprehensive report of the dataset's characteristics and statistics. The library offers a wide range of features and customization options for generating informative and visually appealing reports. In this article [9], this delve into the capabilities of Pandas profiling and explore some advanced use cases and integrations that can be leveraged to produce compelling reports from data frames. By utilizing this tool, data analysts and scientists can gain a deeper understanding of their data and communicate insights effectively to stakeholders. Please find below a summary of Panda's profiling of the primary attributes of our dataset, provided for the reader's reference.



Further exploratory analysis was conducted using Pandas Profiling. The detailed report can be found at https://github.com/shahgem/CIND-820/blob/main/CryptoAnalysis.html

Data Dictionary

The dataset can be described briefly as follows.

Column	Explanation	Variable Type	Data Type
Sno	Serial Numbers	Numerical (Nominal)	int64
Name	Name of the currency	Categorical (Nominal))	object
Symbol	Symbol or abbreviation of the cryptocurrency	Categorical (Nominal)	object
Date	Phone number of customer	Date(datetime)	object
High	Highest price of the cryptocurrency on the given date	Numerical (continuous)	float64
Low	Lowest price of the cryptocurrency on the given date	Numerical (continuous)	float64
Open	Opening price of the cryptocurrency on the given date	Numerical (continuous)	float64
Close	Closing price of the cryptocurrency on the given date	Numerical (continuous)	float64
Volume	Volume of the cryptocurrency traded on the given date	Numerical (continuous)	float64
Marketcap	Market capitalization of the cryptocurrency on the given date	Numerical (continuous)	float64

Missing Values

After reading the “Sixcrypto” dataset into a Pandas dataframe (“df”), we’ve captured a snapshot of the data with 10 columns, 9408 rows:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9408 entries, 0 to 9407
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   SNo        9408 non-null   int64
 1   Name       9408 non-null   object
 2   Symbol     9408 non-null   object
 3   Date       9408 non-null   object
 4   High       9408 non-null   float64
 5   Low        9408 non-null   float64
 6   Open       9408 non-null   float64
 7   Close      9408 non-null   float64
 8   Volume     9408 non-null   float64
 9   Marketcap  9408 non-null   float64
dtypes: float64(6), int64(1), object(3)
memory usage: 735.1+ KB
```

Counting the number of missing values for each column by df.isna().sum().

```
SNo          0
Name         0
Symbol       0
Date         0
High         0
Low          0
Open         0
Close        0
Volume       0
Marketcap    0
dtype: int64
```

Counting the number of duplicated rows for the dataset by df.duplicated().value_counts()

```
False    9408
dtype: int64
```

Our analysis of the dataset revealed that there are no instances of missing values or duplicates. This indicates that the dataset is complete and unique.

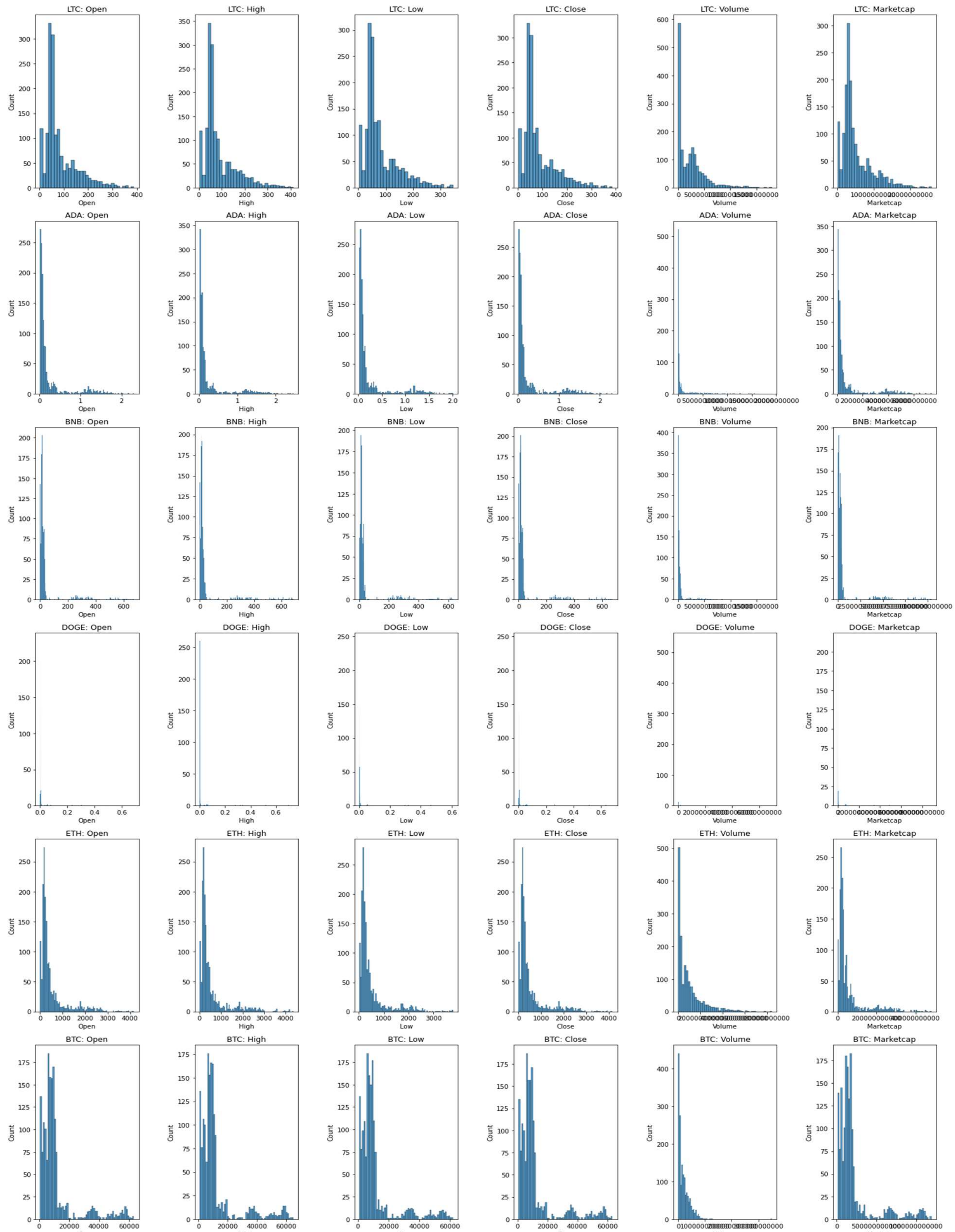
Detection of outliers for numerical columns

Our research aims to identify outliers in the data and examine their influence on the overall distribution of the data. To accomplish this, I employed the commonly used z-score method to identify data points that are over three standard deviations away from the mean, which is a widely accepted approach for detecting outliers.

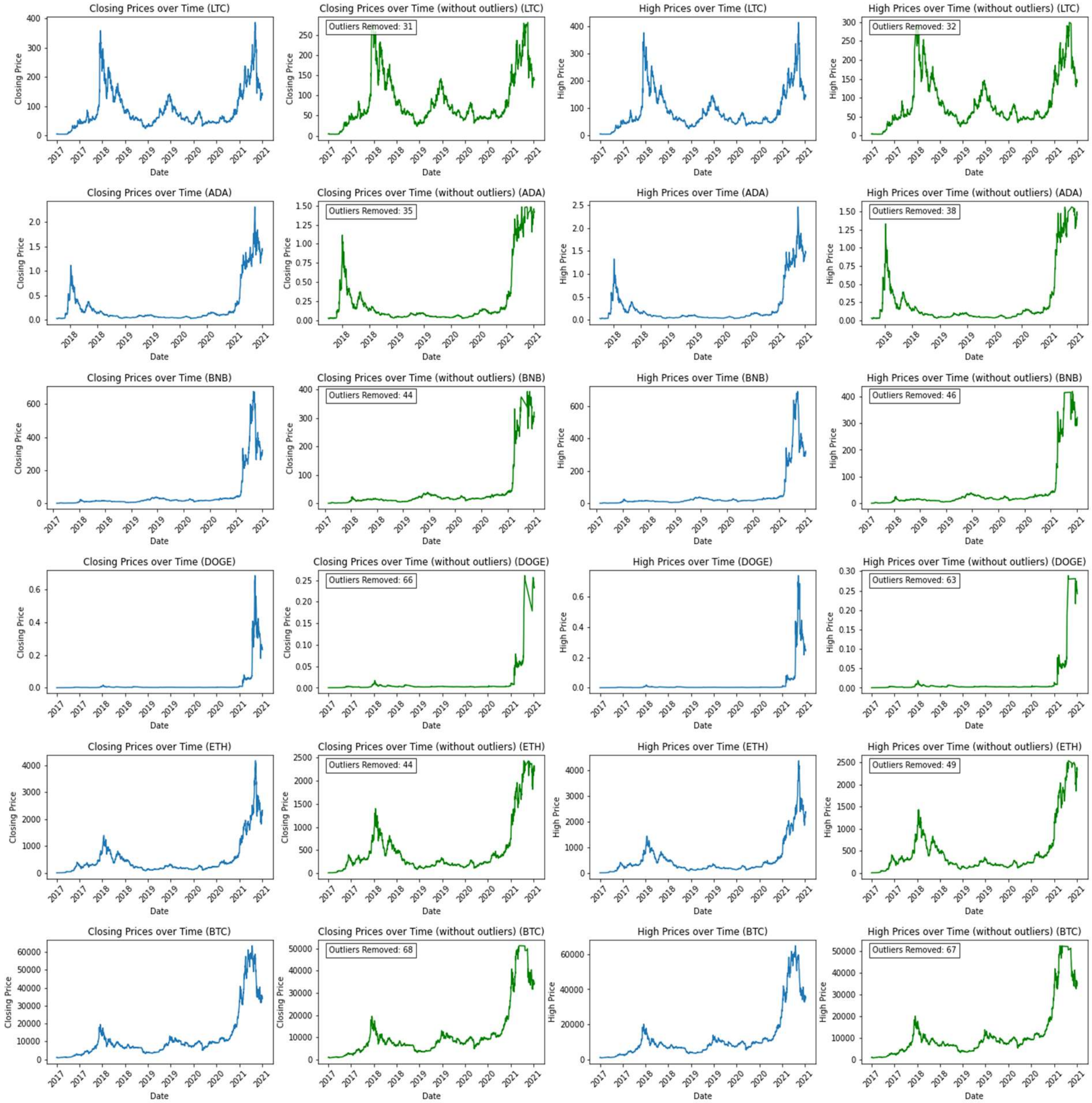
Furthermore, I generated histograms for every currency and column in the dataset to visualize the data distribution and the occurrence of outliers. By doing so, I aimed to gain insight into the extent to which outliers affect the data distribution and potentially impact any further analysis, such as time series forecasting. It is important to note that our approach is not novel but rather a standard practice in the field.

The Python code that I used, imports the pandas, seaborn, and matplotlib libraries for data manipulation, visualization, and plotting.

- It then loads a CSV file named "sixcrypto.csv" using the read_csv function of Pandas, which creates a DataFrame from the file.
- The code then creates two lists: one containing the names of columns to plot ('Open', 'High', 'Low', 'Close', 'Volume', 'Marketcap'), and the other containing unique currency symbols found in the Symbol column of the DataFrame.
- Next, the code creates a subplot figure with a grid of plots, one for each currency and column combination. The subplots function of Matplotlib is used to create the figure, and the nrows, ncols, and figsize parameters specify the number of rows and columns, and the size of the figure.
- The code then loops through each currency and column combination, and creates a boxplot for each using Seaborn's histplot function. The histplot function plots a histogram for the data in the specified column, using only the data from the current currency. The set_title function sets the title of each plot, indicating the currency and column being plotted, and the ticklabel_format function sets the format of the x-axis tick labels to plain text.
- Finally, the code uses tight_layout to adjust the spacing of the subplots, and show to display the figure. Overall, this code generates a grid of histograms for each currency and each column of data, allowing the user to compare and contrast the distribution of the data across different currencies.



The below plot shows the closing and high prices of each currency over time, with and without outliers removed. The plot also shows the number of outliers removed for each currency.



Count of Outliers for each Cryptocurrency in the Dataset:

Count of Outliers for each Cryptocurrency in the Dataset

Currency	High	Low	Open	Close	Volume	Marketcap
LTC	32	28	31	31	36	27
ADA	38	37	37	35	38	42
BNB	46	43	44	44	42	44
DOGE	63	69	67	66	27	66
ETH	49	47	47	44	38	48
BTC	67	65	68	68	14	68

Percentage of Outliers for each Cryptocurrency in the Dataset:

Percentage of Outliers for each Cryptocurrency in the Dataset

Currency	High	Low	Open	Close	Volume	Marketcap
LTC	1.94%	1.70%	1.88%	1.88%	2.18%	1.64%
ADA	2.77%	2.69%	2.69%	2.55%	2.77%	3.06%

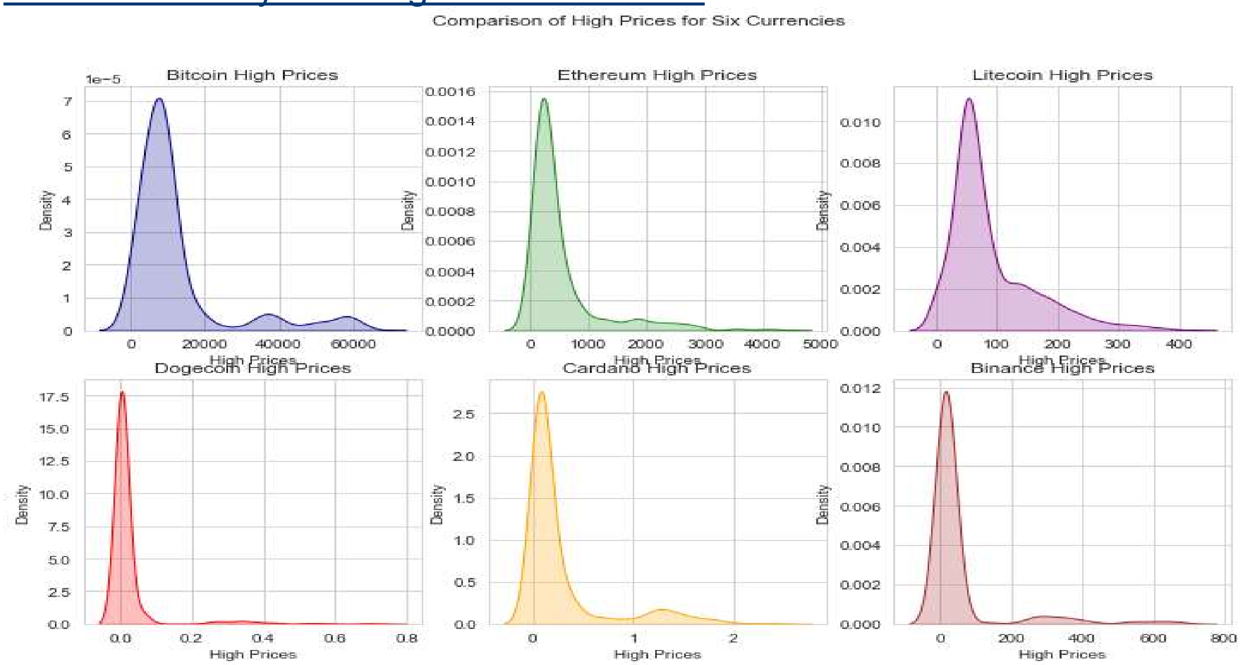
```
     BNB 3.19% 2.98% 3.05%  3.05%   2.91%       3.05%
    DOGE 3.82% 4.19% 4.07%  4.00%   1.64%       4.00%
     ETH 2.97% 2.85% 2.85%  2.67%   2.31%       2.91%
     BTC 4.07% 3.94% 4.13%  4.13%   0.85%       4.13%
```

Whether or not to remove outliers from your dataset is a decision that depends on the specific context and purpose of our analysis. In some cases, outliers can provide valuable insights into unusual events or behaviors in the data, and removing them may lead to a loss of important information. In other cases, outliers may simply represent errors or anomalies that are not relevant to the analysis, and removing them can improve the accuracy and validity of the results.

In general, it's a good practice to examine the outliers carefully and determine whether they should be removed or retained based on their relevance to the analysis. If we decide to remove outliers, we should be clear about the criteria we used to identify them and document the process for reproducibility. It's also important to consider the potential impact of outlier removal on the overall distribution and characteristics of the data.

Regarding time series forecasting which we are going to perform in the remaining part of our study, outliers can have a significant impact on the accuracy of the forecasts, and it may be necessary to account for them explicitly in the modeling process. In some cases, outlier detection and correction techniques can be integrated into the forecasting models themselves, or separate preprocessing steps may be necessary to remove or adjust the outliers prior to modeling.

Univariate Analysis of High & Close Prices



As we can see, our graphs are skewed to the right (positive skew).

Skewness is a statistical measure that points out the degree of asymmetry of a distribution. It tells you whether the distribution is symmetric or not, and if it is not symmetric, it tells you whether the tail of the distribution is longer on the left or the right side.

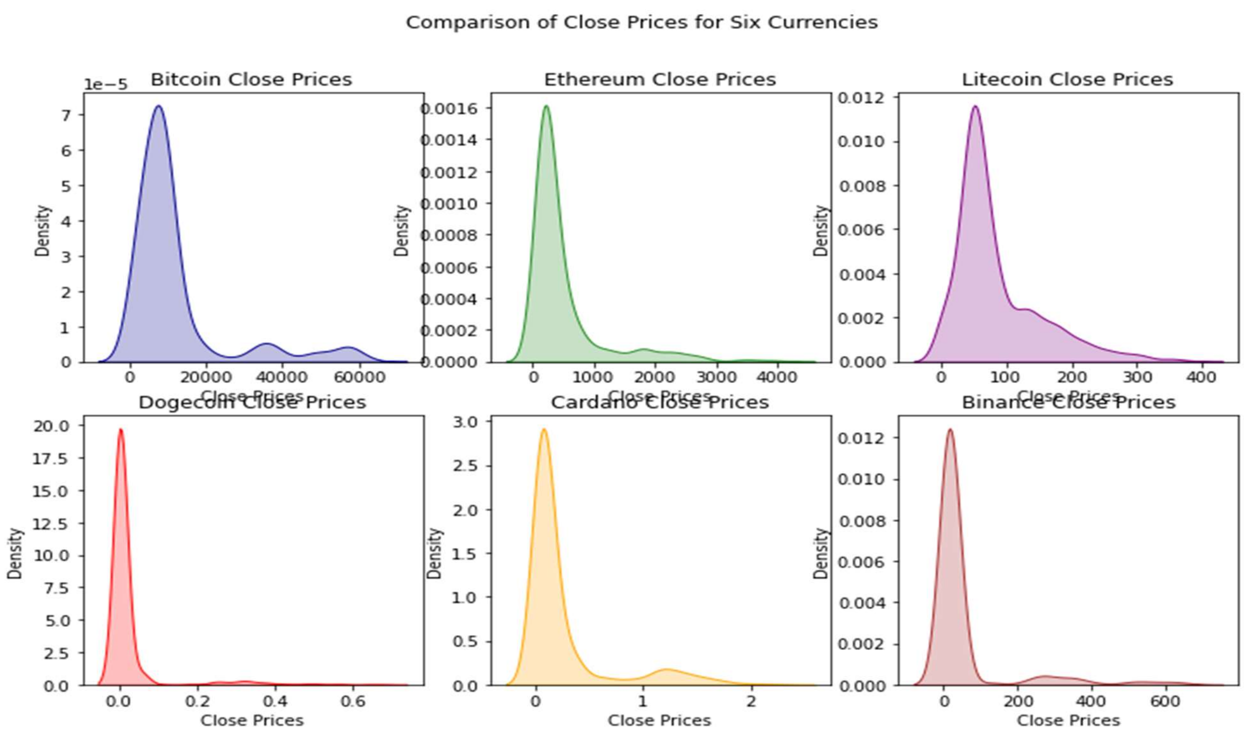
A skewness value of 0 indicates a perfectly symmetric distribution, while a positive skewness value indicates a longer tail on the right side of the distribution, and a negative skewness value indicates a longer tail on the left side of the distribution.

Skewness can be useful in understanding the underlying data, as it can indicate whether there are outliers or unusual data points that may be affecting the distribution. It can also help in choosing appropriate statistical methods for analyzing the data, as some methods assume a normal distribution and may not be appropriate for skewed data.

```
Symbol  Skewness
BTC     2.318963
ETH     2.672445
LTC     1.608763
DOGE    4.826529
ADA     2.389419
BNB     3.343639
```

Based on the results, it seems that the "High" prices distribution for DOGE has the highest skewness value (4.826529), indicating a significant degree of positive skewness, which suggests that there are more observations with smaller values (left tail) and fewer observations with larger values (right tail). Similarly, BNB also has a high skewness value (3.343639), indicating a degree of positive skewness. The other cryptocurrencies (BTC, ETH, LTC, and ADA) also have positive skewness values, indicating a similar pattern of more small values and fewer large values, but to a lesser extent compared to DOGE and BNB.

In summary, the skewness values suggest that the "High" prices distribution for these cryptocurrencies are skewed to the right (positively skewed), indicating that there are more observations with smaller values and fewer observations with larger values.

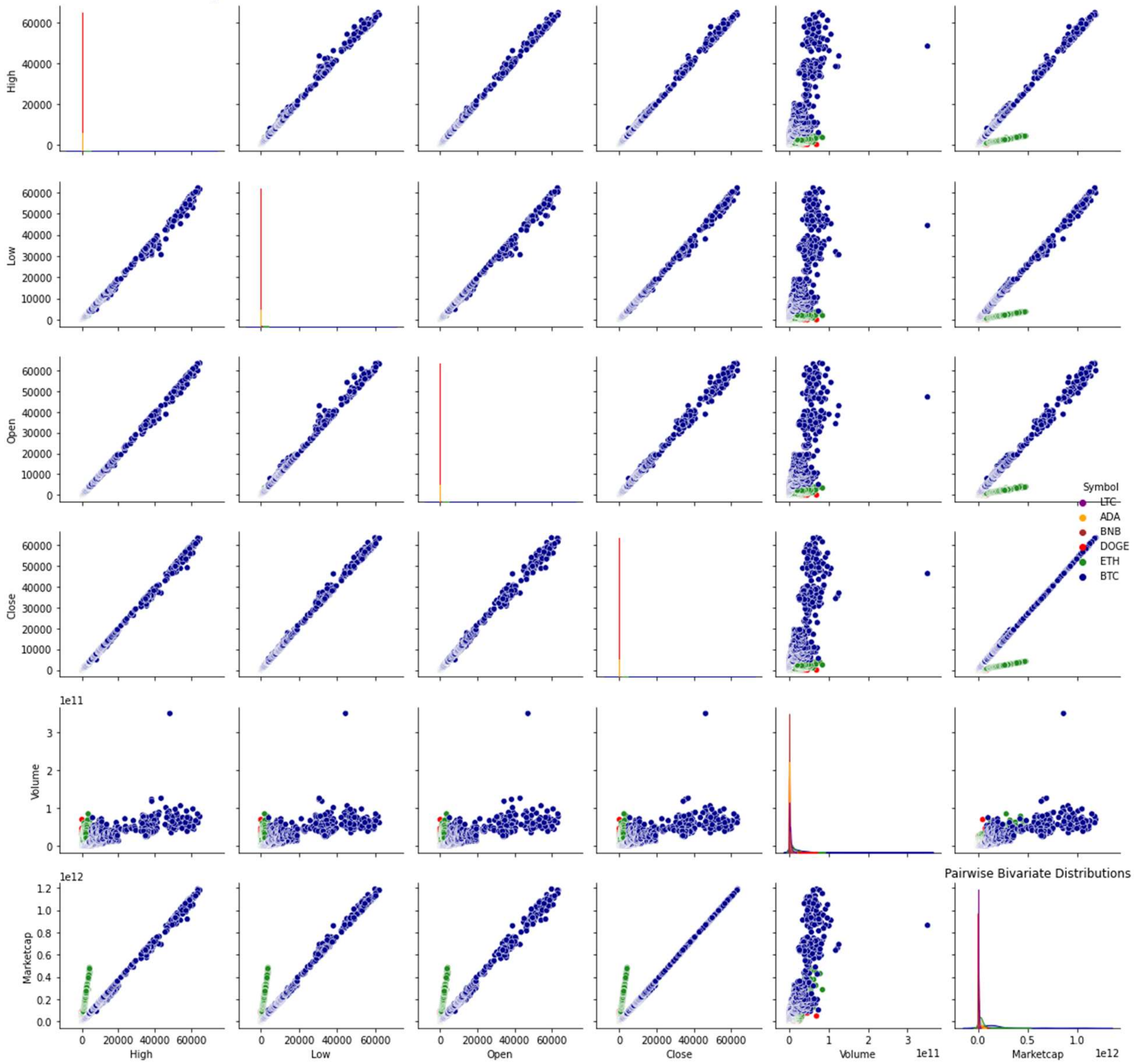



```
Symbol  Skewness
BTC     2.331511
ETH     2.642688
LTC     1.560778
DOGE    4.701561
ADA     2.378311
BNB     3.359945
```

Based on the Skewness values of Close pirce, it appears that DOGE has the highest degree of positive skewness, followed by BNB, ETH, BTC, ADA, and LTC. This suggests that the distributions of daily returns for these cryptocurrencies are not symmetric and have a long right tail. This means that there is a greater probability of large positive returns compared to large negative returns for these cryptocurrencies.

By knowing the skewness of a distribution, investors can better estimate whether the future returns of an asset will be higher or lower than its past returns. This information is helpful in making investment decisions and managing risk. For example, if the skewness is positive, investors might expect to see more large gains in the future and may consider investing in the asset. Conversely, if the skewness is negative, investors might expect to see more large losses in the future and may decide to avoid investing in the asset.

Bivariate Analysis



Bivariate analysis is a statistical method that is used to check the relationship between two variables. In the context of the sixcrypto.csv dataset, bivariate analysis could be used to explore the relationships between different cryptocurrency variables, such as the relationship between the price and volume of each cryptocurrency.

Statistical Summary for numerical columns

df.describe().transpose() is a Pandas method that provides descriptive statistics for a DataFrame. The describe() method computes summary statistics such as count, mean, standard deviation, minimum, and maximum values for numerical columns in the DataFrame. The transpose() method then switches the rows and columns of the resulting DataFrame, making the statistics for each column appear as rows instead of columns. This makes it easier to interpret the summary statistics for each feature in the dataset, as we can view them side by side rather than having to scan across multiple columns. In summary, df.describe().transpose() is a convenient way to obtain a quick overview of the summary statistics for each feature in a DataFrame.

	count	mean	std	min	25%	50%	75%	max
SNo	9408.0	4.704500e+03	2.716000e+03	1.000000e+00	2.352750e+03	4.704500e+03	7.056250e+03	9.408000e+03
High	9408.0	2.245296e+03	7.297368e+03	2.048400e-04	1.060599e-01	4.338992e+01	3.497877e+02	6.486310e+04
Low	9408.0	2.108906e+03	6.823157e+03	1.946130e-04	9.701510e-02	4.100250e+01	3.198715e+02	6.220896e+04
Open	9408.0	2.180658e+03	7.078639e+03	1.966920e-04	1.011220e-01	4.234630e+01	3.378597e+02	6.352375e+04
Close	9408.0	2.184401e+03	7.085992e+03	1.967090e-04	1.013561e-01	4.238147e+01	3.379402e+02	6.350346e+04
Volume	9408.0	5.842552e+09	1.265555e+10	9.284190e+03	8.131160e+07	5.721684e+08	5.068277e+09	3.510000e+11
Marketcap	9408.0	5.086161e+10	1.334378e+11	9.986680e+06	1.184531e+09	4.387144e+09	3.859717e+10	1.190000e+12

The output of df.describe().transpose() provides a summary of the key statistical measures for each numerical column in the dataset. The count row indicates the number of non-missing values in each column.

The mean row shows the average value for each column, while the std row indicates the standard deviation, or the amount of variation, in each column's values. The min and max rows show the smallest and largest values in each column, respectively.

The 25%, 50%, and 75% rows represent the first quartile (Q1), median, and third quartile (Q3) of the data distribution, respectively. These values can be used to assess the spread of the data, identify potential outliers, and evaluate the skewness of the distribution.

Based on the output, it can be observed that the dataset has 9,408 observations and 6 numerical features. The Volume and Marketcap features have the highest mean and standard deviation values, indicating significant variability in the data. Meanwhile, the SNo feature represents a sequential identifier for each observation, while the remaining features (High, Low, Open, and Close) represent the price data of a financial asset. The range of values for these price features is quite large, as evidenced by the large difference between the min and max values.

The high, low, open, and close columns represent the daily price of the cryptocurrencies, with mean values of 2.245296e+03, 2.108906e+03, 2.180658e+03, and 2.184401e+03, respectively. The standard deviations are relatively large, indicating a wide variability in the price movements. The minimum and maximum values of the high, low, open, and close columns are 2.048400e-04 and 6.486310e+04, respectively.

The volume column represents the daily trading volume of the cryptocurrencies, with a mean of 5.842552e+09 (SD = 1.265555e+10). The trading volume exhibits a positively skewed distribution, as evidenced by the median value (50%) of 5.721684e+08 being much lower than the mean value. The minimum and maximum values of the volume column are 9 .284190e+03 and 3.510000e+11, respectively.

The Marketcap column represents the market capitalization of the cryptocurrencies, with a mean of 5.086161e+10 (SD = 1.334378e+11). The market capitalization also exhibits a positively skewed distribution, with the median value (50%) of 4.387144e+09 being much lower than the mean value. The minimum and maximum values of the Marketcap column are 9.986680e+06 and 1.190000e+12, respectively.

Overall, the df.describe().transpose() output provides a quick summary of the key statistics for each numerical feature in the dataset, which can be used to gain insights into the data distribution and identify potential issues or patterns in the data.

Let's delve into the statistical summary of each currency to gain further insights into our dataset.

The code written in Python uses the pandas library to load, manipulate and summarize data from a CSV file named 'sixcrypto.csv'.

- First, it imports the pandas library using the alias 'pd' for easier reference.
- Then, the CSV file is loaded into a pandas DataFrame called 'df' using the **pd.read_csv()** function.
- The 'SNo' column in the DataFrame is dropped using the **drop()** method with the 'axis=1' argument to indicate that it is a column that needs to be dropped. The **inplace=True** argument is used to modify the DataFrame in place rather than creating a copy.
- Next, the DataFrame is renamed as 'crypto_data' using the assignment operator '='.
- The data in the 'crypto_data' DataFrame is then grouped by the 'Symbol' column using the **groupby()** method, which returns a DataFrameGroupBy object.
- The code then loops through the groups and calculates the statistical summary for each group using the **describe()** method on each group. The **describe()** method provides various statistical summary measures like count, mean, standard deviation, minimum and maximum values, quartiles etc.
- The **transpose()** method is then used to transpose the summary statistics from rows to columns for easier readability.
- Finally, the name of the group and its summary statistics are printed out using the **print()** function.

Overall, this code loads and manipulates cryptocurrency data from a CSV file using pandas and calculates statistical summaries for each group of cryptocurrency symbols in the dataset.

Litecoin:

```
LTC
            count          mean           std           min           25%  \
High       1648.0  8.836755e+01  7.040699e+01  3.769480e+00  4.540341e+01
Low        1648.0  8.076967e+01  6.201480e+01  3.613590e+00  4.331137e+01
Open       1648.0  8.476022e+01  6.651762e+01  3.714440e+00  4.447457e+01
Close      1648.0  8.481855e+01  6.645836e+01  3.714530e+00  4.448149e+01
Volume     1648.0  2.329138e+09  2.596721e+09  2.629220e+06  3.286380e+08
Marketcap  1648.0  5.178139e+09  4.158405e+09  1.838773e+08  2.742632e+09

                    50%           75%           max
High       6.095937e+01  1.189067e+02  4.129601e+02
Low        5.747658e+01  1.088358e+02  3.452988e+02
Open       5.932576e+01  1.145168e+02  3.878692e+02
Close      5.933581e+01  1.146692e+02  3.864508e+02
Volume     1.711569e+09  3.353639e+09  1.799426e+10
Marketcap  3.656093e+09  6.694942e+09  2.579652e+10
```

For Litecoin (LTC), the data includes 1648 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 8.836755e+01, 8.076967e+01, 8.476022e+01, and 8.481855e+01, respectively. The standard deviations for these variables are relatively large, with values of 7.040699e+01, 6.201480e+01, 6.651762e+01, and 6.645836e+01, respectively, indicating a wide variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 3.769480e+00 and 4.129601e+02, respectively. The volume column has a mean of 2.329138e+09 (SD = 2.596721e+09), with a positively skewed distribution as evidenced by the median value (50%) of 1.711569e+09 being much lower than the mean value. The minimum and maximum values of the volume column are 2.629220e+06 and 1.799426e+10, respectively.

The market capitalization column has a mean of 5.178139e+09 (SD = 4.158405e+09), also with a positively skewed distribution, as evidenced by the median value (50%) of 3.656093e+09 being much lower than the mean value. The minimum and maximum values of the market capitalization column are 1.838773e+08 and 2.579652e+10, respectively.

Cardano Coin:

```
ADA
            count          mean           std           min           25%  \
High       1374.0  2.698068e-01  4.335232e-01  2.105030e-02  4.756542e-02
Low        1374.0  2.397098e-01  3.809276e-01  1.762000e-02  4.460074e-02
Open       1374.0  2.552866e-01  4.084556e-01  1.841390e-02  4.589775e-02
Close      1374.0  2.563126e-01  4.096914e-01  1.853910e-02  4.594670e-02
Volume     1374.0  8.934183e+08  2.107653e+09  1.739460e+06  5.014830e+07
Marketcap  1374.0  7.603454e+09  1.303878e+10  4.806646e+08  1.191263e+09

                    50%           75%           max
High       9.027426e-02  1.945189e-01  2.461766e+00
Low        8.316395e-02  1.724417e-01  2.013285e+00
Open       8.686724e-02  1.813737e-01  2.300190e+00
Close      8.700222e-02  1.833791e-01  2.309113e+00
Volume     1.186742e+08  4.875977e+08  1.914198e+10
Marketcap  2.270889e+09  5.174547e+09  7.377224e+10
```

For Cardano (ADA), the data includes 1374 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 2.698068e-01, 2.397098e-01, 2.552866e-01, and 2.563126e-01, respectively. The standard deviations for these variables are 4.335232e-01, 3.809276e-01, 4.084556e-01, and 4.096914e-01, respectively, indicating a wide variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 2.105030e-02 and 2.461766e+00, respectively. The volume column has a mean of 8.934183e+08 (SD = 2.107653e+09), with a positively skewed distribution as evidenced by the median value (50%) of 1.186742e+08 being much lower than the mean value. The minimum and maximum values of the volume column are 1.739460e+06 and 1.914198e+10, respectively.

The market capitalization column has a mean of 7.603454e+09 (SD = 1.303878e+10), also with a positively skewed distribution, as evidenced by the median value (50%) of 1.191263e+09 being much lower than the mean value. The minimum and maximum values of the market capitalization column are 4.806646e+08 and 2.300190e+09, respectively.

Binance Coin:

```
BNB
```

```
                count          mean           std           min           25%  \
High       1442.0  5.476410e+01  1.216758e+02  1.012110e-01  1.039075e+01
Low        1442.0  4.916581e+01  1.081185e+02  9.610940e-02  9.677340e+00
Open       1442.0  5.202823e+01  1.151701e+02  9.972120e-02  1.003786e+01
Close      1442.0  5.225031e+01  1.153909e+02  9.986680e-02  1.006835e+01
Volume     1442.0  6.269804e+08  1.479775e+09  9.284190e+03  5.089148e+07
Marketcap  1442.0  7.835965e+09  1.780249e+10  9.986680e+06  1.157863e+09

                    50%           75%           max
High       1.659211e+01  2.824091e+01  6.909320e+02
Low        1.572725e+01  2.696304e+01  6.314653e+02
Open       1.621033e+01  2.766989e+01  6.763159e+02
Close      1.621057e+01  2.769111e+01  6.756841e+02
Volume     1.981830e+08  3.942378e+08  1.798295e+10
Marketcap  2.451099e+09  4.061743e+09  1.040000e+11
```

For Binance Coin (BNB), the data includes 1442 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 5.476410e+01, 4.916581e+01, 5.202823e+01, and 5.225031e+01, respectively. The standard deviations for these variables are 1.216758e+02, 1.081185e+02, 1.151701e+02, and 1.153909e+02, respectively, indicating a wide variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 1.012110e-01 and 6.909320e+02, respectively. The volume column has a mean of 6.269804e+08 (SD = 1.479775e+09), with a positively skewed distribution as evidenced by the median value (50%) of 1.981830e+08 being much lower than the mean value. The minimum and maximum values of the volume column are 9.284190e+03 and 1.798295e+10, respectively.

The market capitalization column has a mean of 7.835965e+09 (SD = 1.780249e+10), also with a positively skewed distribution, as evidenced by the median value (50%) of 2.451099e+09 being much lower than the mean value. The minimum and maximum values of the market capitalization column are 9.986680e+06 and 1.040000e+11, respectively.

Bitcoin:

```
BTC
                count          mean           std           min           25%  \
High       1648.0  1.216162e+04  1.357205e+04  8.233070e+02  4.877465e+03
Low        1648.0  1.143817e+04  1.264045e+04  7.557560e+02  4.470868e+03
Open       1648.0  1.181870e+04  1.314910e+04  7.751780e+02  4.613783e+03
Close      1648.0  1.183847e+04  1.315736e+04  7.777570e+02  4.680163e+03
Volume     1648.0  1.976129e+10  2.174509e+10  6.085170e+07  4.188452e+09
Marketcap  1648.0  2.145495e+11  2.473754e+11  1.251914e+10  7.890546e+10

                    50%           75%           max
High       8.234140e+03  1.110076e+04  6.486310e+04
Low        7.842872e+03  1.053982e+04  6.220896e+04
Open       8.063433e+03  1.084634e+04  6.352375e+04
Close      8.055607e+03  1.085863e+04  6.350346e+04
Volume     1.365801e+10  2.978197e+10  3.510000e+11
Marketcap  1.410000e+11  1.940000e+11  1.190000e+12
```

The data for Bitcoin shows that there were 1648 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 1.216162e+04, 1.143817e+04, 1.181870e+04, and 1.183847e+04, respectively. The standard deviations for these variables are relatively high, indicating that there has been significant variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 8.233070e+02 and 6.486310e+04, respectively. The volume column has a mean of 1.976129e+10 (SD = 2.174509e+10), with a positively skewed distribution, as evidenced by the median value (50%) of 1.365801e+10 being lower than the mean value. The minimum and maximum values of the volume column are 6.085170e+07 and 3.510000e+11, respectively.

The market capitalization column has a mean of 2.145495e+11 (SD = 2.473754e+11), also with a positively skewed distribution, as evidenced by the median value (50%) of 1.410000e+11 being lower than the mean value. The minimum and maximum values of the market capitalization column are 1.251914e+10 and 1.190000e+12, respectively.

Overall, the data suggests that Bitcoin is a highly volatile asset with significant fluctuations in price movements and trading volume. The market capitalization of Bitcoin is also quite large, indicating that it is a significant player in the cryptocurrency market

Dogecoin:

```
DOGE
            count          mean           std           min           25%  \
High       1648.0  2.485953e-02  8.806860e-02  2.048400e-04  2.151102e-03
Low        1648.0  2.058939e-02  7.071171e-02  1.946130e-04  2.055418e-03
Open       1648.0  2.270210e-02  7.942883e-02  1.966920e-04  2.096444e-03
Close      1648.0  2.285838e-02  7.968948e-02  1.967090e-04  2.100501e-03
Volume     1648.0  7.242091e+08  3.604293e+09  5.407950e+04  1.055230e+07
Marketcap  1648.0  2.933065e+09  1.033952e+10  2.128503e+07  2.498939e+08

                    50%           75%           max
High       2.736100e-03  3.838227e-03  7.375666e-01
Low        2.599330e-03  3.496225e-03  6.081677e-01
Open       2.662656e-03  3.672514e-03  6.878015e-01
Close      2.663201e-03  3.689054e-03  6.847770e-01
Volume     3.486363e+07  1.028266e+08  6.941068e+10
Marketcap  3.263688e+08  4.411314e+08  8.868082e+10
```

The data for Dogecoin shows that there were 1648 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 2.485953e-02, 2.058939e-02, 2.270210e-02, and 2.285838e-02, respectively. The standard deviations for these variables are relatively high, indicating that there has been significant variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 2.048400e-04 and 7.375666e-01, respectively. The volume column has a mean of 7.242091e+08 (SD = 3.604293e+09), with a positively skewed distribution, as evidenced by the median value (50%) of 3.486363e+07 being lower than the mean value. The minimum and maximum values of the volume column are 5.407950e+04 and 6.941068e+10, respectively.

The market capitalization column has a mean of 2.933065e+09 (SD = 1.033952e+10), also with a positively skewed distribution, as evidenced by the median value (50%) of 3.263688e+08 being lower than the mean value. The minimum and maximum values of the market capitalization column are 2.128503e+07 and 8.868082e+10, respectively.

Overall, the data suggests that Dogecoin is a highly volatile asset with significant fluctuations in price movements and trading volume. The market capitalization of Dogecoin is relatively small compared to other cryptocurrencies, indicating that it is a smaller player in the cryptocurrency market.

Ethereum:

```
ETH
           count          mean           std           min           25%  \
High      1648.0  5.196492e+02  6.744691e+02  8.436330e+00  1.751977e+02
Low       1648.0  4.770173e+02  6.069889e+02  7.982310e+00  1.649274e+02
Open      1648.0  4.995809e+02  6.434077e+02  7.982310e+00  1.702685e+02
Close     1648.0  5.009299e+02  6.448244e+02  8.172570e+00  1.706190e+02
Volume    1648.0  9.245476e+09  1.132841e+10  4.689950e+06  1.533612e+09
Marketcap 1648.0  5.449910e+10  7.462537e+10  7.150492e+08  1.815161e+10

                   50%           75%           max
High      2.669922e+02  4.931448e+02  4.362351e+03
Low       2.451433e+02  4.652826e+02  3.785849e+03
Open      2.571551e+02  4.795873e+02  4.174636e+03
Close     2.580102e+02  4.797018e+02  4.168701e+03
Volume    5.598255e+09  1.243547e+10  8.448291e+10
Marketcap 2.695053e+10  4.980239e+10  4.830000e+11
```

The data for Ethereum shows that there were 1648 daily observations of the high, low, open, close prices, trading volume, and market capitalization. The mean values for the high, low, open, and close prices are 5.196492e+02, 4.770173e+02, 4.995809e+02, and 5.009299e+02, respectively. The standard deviations for these variables are relatively high, indicating that there has been significant variability in price movements.

The minimum and maximum values of the high, low, open, and close prices are 8.436330e+00 and 4.362351e+03, respectively. The volume column has a mean of 9.245476e+09 (SD = 1.132841e+10), with a positively skewed distribution, as evidenced by the median value (50%) of 5.598255e+09 being lower than the mean value. The minimum and maximum values of the volume column are 4.689950e+06 and 8.448291e+10, respectively.

The market capitalization column has a mean of 5.449910e+10 (SD = 7.462537e+10), also with a positively skewed distribution, as evidenced by the median value (50%) of 2.695053e+10 being lower than the mean value. The minimum and maximum values of the market capitalization column are 7.150492e+08 and 4.830000e+11, respectively.

Overall, the data suggests that Ethereum is a highly volatile asset with significant fluctuations in price movements and trading volume. The market capitalization of Ethereum is relatively large compared to other cryptocurrencies, indicating that it is a major player in the cryptocurrency market.

Correlation

Correlation Coefficient	Strength of Relationship
-1	Perfect negative correlation
-0.7 to -0.9	Strong negative correlation
-0.5 to -0.7	Moderate negative correlation
-0.3 to -0.5	Weak negative correlation
0	No correlation
0.3 to 0.5	Weak positive correlation
0.5 to 0.7	Moderate positive correlation

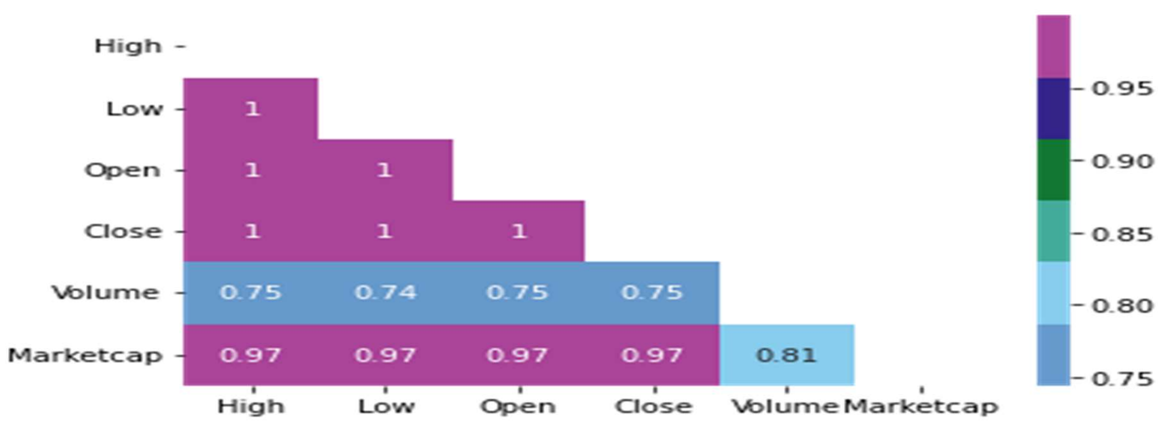
Correlation Coefficient	Strength of Relationship
0.7 to 0.9	Strong positive correlation
1	Perfect positive correlation

It's worth noting that the interpretation of the correlation coefficient can vary depending on the field of study and the context of the analysis. However, in general, a correlation coefficient between -0.7 and -0.9 (or between 0.7 and 0.9) is considered a strong correlation, while a coefficient between -0.5 and -0.7 (or between 0.5 and 0.7) is considered a moderate correlation. Coefficients between -0.3 and -0.5 (or between 0.3 and 0.5) are typically considered weak correlations.

- The Python code loads a dataset from a CSV file named "sixcrypto.csv" using the Pandas read_csv() function, creating a DataFrame object named df. The num_cols variable is then used to select only the numerical columns from the df DataFrame, namely Open, High, Low, Close, Volume, and Marketcap. A new DataFrame named num_df is created to hold only these selected numerical columns.
- Next, the correlation matrix between these numerical columns is calculated using the corr() method on the num_df DataFrame.
- Finally, the correlation matrix is printed to the console using the print() function.

```
               Open      High       Low     Close    Volume   Marketcap
Open       1.000000  0.999546  0.999109  0.998901  0.749004   0.968995
High       0.999546  1.000000  0.999053  0.999525  0.750852   0.969801
Low        0.999109  0.999053  1.000000  0.999423  0.744467   0.969238
Close      0.998901  0.999525  0.999423  1.000000  0.748179   0.970084
Volume     0.749004  0.750852  0.744467  0.748179  1.000000   0.807050
Marketcap  0.968995  0.969801  0.969238  0.970084  0.807050   1.000000
```

The correlation results show a strong positive correlation among the variables, with correlation coefficients ranging from 0.749 to 1.000. The strongest correlation is observed between the High and Close variables (correlation coefficient of 0.999525), followed by the correlation between the Low and Close variables (correlation coefficient of 0.999423). The Volume variable has a weaker correlation with the other variables (correlation coefficients ranging from 0.744 to 0.750).



The analysis of the correlation results suggests a high level of interdependence among the various cryptocurrency price features, such as Open, High, Low, Close, Volume, and Marketcap. The correlation coefficients between these variables range from 0.749 to 0.970, which indicates a strong positive association between them. This finding has significant implications for modeling and forecasting cryptocurrency prices, as it implies that incorporating all of these variables may lead to data redundancy or multicollinearity. This means that accurately predicting cryptocurrency prices might be challenging because it could be difficult to discern the relative importance of each feature.

To overcome this challenge, further analysis could be conducted to explore the direction and strength of causal relationships among the variables. Such analysis could provide insights into the underlying mechanisms that drive cryptocurrency price movements and help identify the most critical factors to consider in forecasting future price trends. Additionally, external factors that influence cryptocurrency prices, such as economic and political events, could also be examined to determine their impact on price movements. By considering both the internal and external factors that affect cryptocurrency prices, more accurate and reliable forecasting models could be developed.

Principal Component Analysis (PCA)

To delve deeper into the analysis, it could be beneficial to perform a principal component analysis (PCA) on the data to identify the most significant features and decrease the dataset's dimensionality. PCA is a popular approach in data analysis and machine learning that involves reducing the dimensionality of high-dimensional data while retaining critical information. The primary objective of PCA is to convert a collection of correlated variables into a smaller set of uncorrelated variables (referred to as principal components), while still preserving the most crucial information.

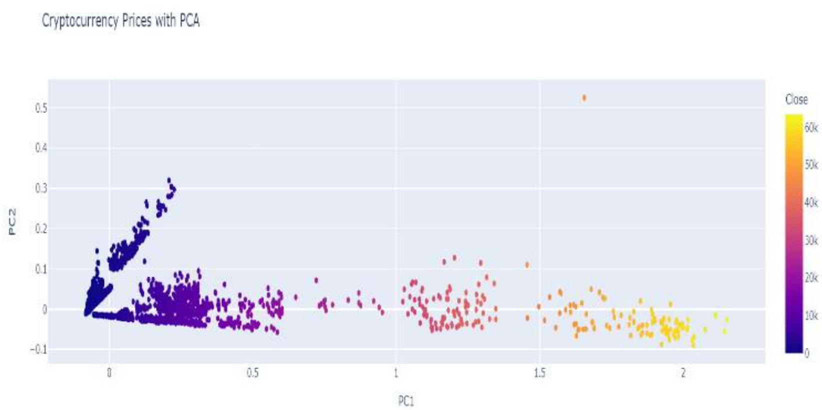
Here are Here are some key points covered in the article written by Matt Brems **[10]**:

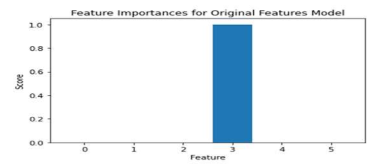
1. PCA is a mathematical methodology that can convert high-dimensional data into a lower-dimensional space while maintaining crucial information.
2. PCA is a beneficial tool in simplifying the complexity of data, discovering patterns and associations between variables, and enhancing the efficiency of machine learning models.
3. The article provides a step-by-step guide to performing PCA using Python, including data preprocessing, PCA implementation, and visualizing the results.
4. The article also covers important concepts related to PCA, such as variance, covariance, eigenvectors, and eigenvalues.
5. The article provides practical examples of using PCA in real-world scenarios, such as image processing and finance.
6. The article discusses some common pitfalls and limitations of PCA, such as the interpretability of the results and the assumption of linearity.

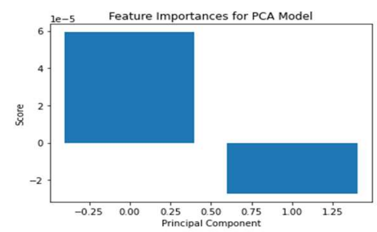
The code used in python to access the PCA model begins by importing the necessary Python libraries including Pandas, NumPy, Plotly, scikit-learn, and Matplotlib. It then defines a function to calculate the mean absolute percentage error (MAPE) which is used to evaluate the model later on.

- ➢ The code reads in a dataset of cryptocurrency prices from a CSV file using Pandas and selects a set of features to use in the analysis. These features are then scaled using the MinMaxScaler from scikit-learn.
- ➢ Principal component analysis (PCA) is performed on the scaled features using the PCA function from scikit-learn. Two principal components are selected, and a new dataframe is created using the principal components and the original data. This allows the data to be visualized in two dimensions.
- ➢ The data is split into training and testing sets using the train_test_split function from scikit-learn. A linear regression model is fitted on the original features, and the metrics for the model are computed using mean squared error, mean absolute error, MAPE, R-squared, and adjusted R-squared.

- The data is then standardized using the StandardScaler from scikit-learn, and PCA is performed on the standardized data. The principal components are concatenated with the original features, and the data is split into training and testing sets again.
- Another linear regression model is fitted on the principal components, and the metrics for this model are computed. The results of the PCA are visualized using a scatter plot in Plotly.
- Finally, the feature importances for both the original features model and the PCA model are computed and visualized using bar plots in Matplotlib.
- Overall, this code performs a basic analysis of cryptocurrency prices using PCA and linear regression models. The code could be further improved by optimizing the hyperparameters of the models and including additional features in the analysis.








```
Feature importance for the original features model:
Feature: 0, Score: 0.00000
Feature: 1, Score: -0.00000
Feature: 2, Score: 0.00000
Feature: 3, Score: 1.00000
Feature: 4, Score: 0.00000
Feature: 5, Score: -0.00000
Feature importance for the PCA model:
Principal Component: 1, Score: 0.00006
Principal Component: 2, Score: -0.00003
```

The results of the feature importance analysis for the original features model show that only Feature 3 has a score of 1.0, suggesting that it is the only feature that significantly contributes to the model's predictive performance. Conversely, all other features have scores of 0.0, indicating that they do not have a significant impact on the model's predictive power.

On the other hand, the feature importances for the PCA model are based on the principal components. The scores for the principal components are much smaller than 1.0, suggesting that each principal component contributes only a small amount to the model's predictive performance. Specifically, Principal Component 1 has a score of 0.00006, indicating a very small positive impact on the model's performance, while Principal Component 2 has a score of -0.00003, indicating a small negative impact. Overall, the results suggest that the PCA model's performance may be less reliant on any one particular feature or principal component, but instead relies on the combined effects of all of the components.

Regarding which model to use, it ultimately depends on the specific requirements of your problem and the resources available. The original features model may be preferred if interpretability and transparency are important, as it directly uses the original features in the analysis. On the other hand, the PCA model may be preferred if dimensionality reduction and computational efficiency are priorities, as it reduces the dimensionality of the data and may result in faster and more efficient analysis.

Seasonal Decomposition of the Cryptocurrencies Close Price

Seasonal decomposition is a technique used in time series analysis to separate a time series into its underlying components: trend, seasonal, and residual. The trend component represents the long-term progression of the time series, the seasonal component captures the repeating patterns that occur over shorter periods of time (e.g., days, weeks, months, seasons), and the residual component represents the random fluctuations or noise that cannot be explained by the trend or seasonal components.

An article “Different types of Time series decomposition” **[11]** written by Andrew Plummer has these below key points:

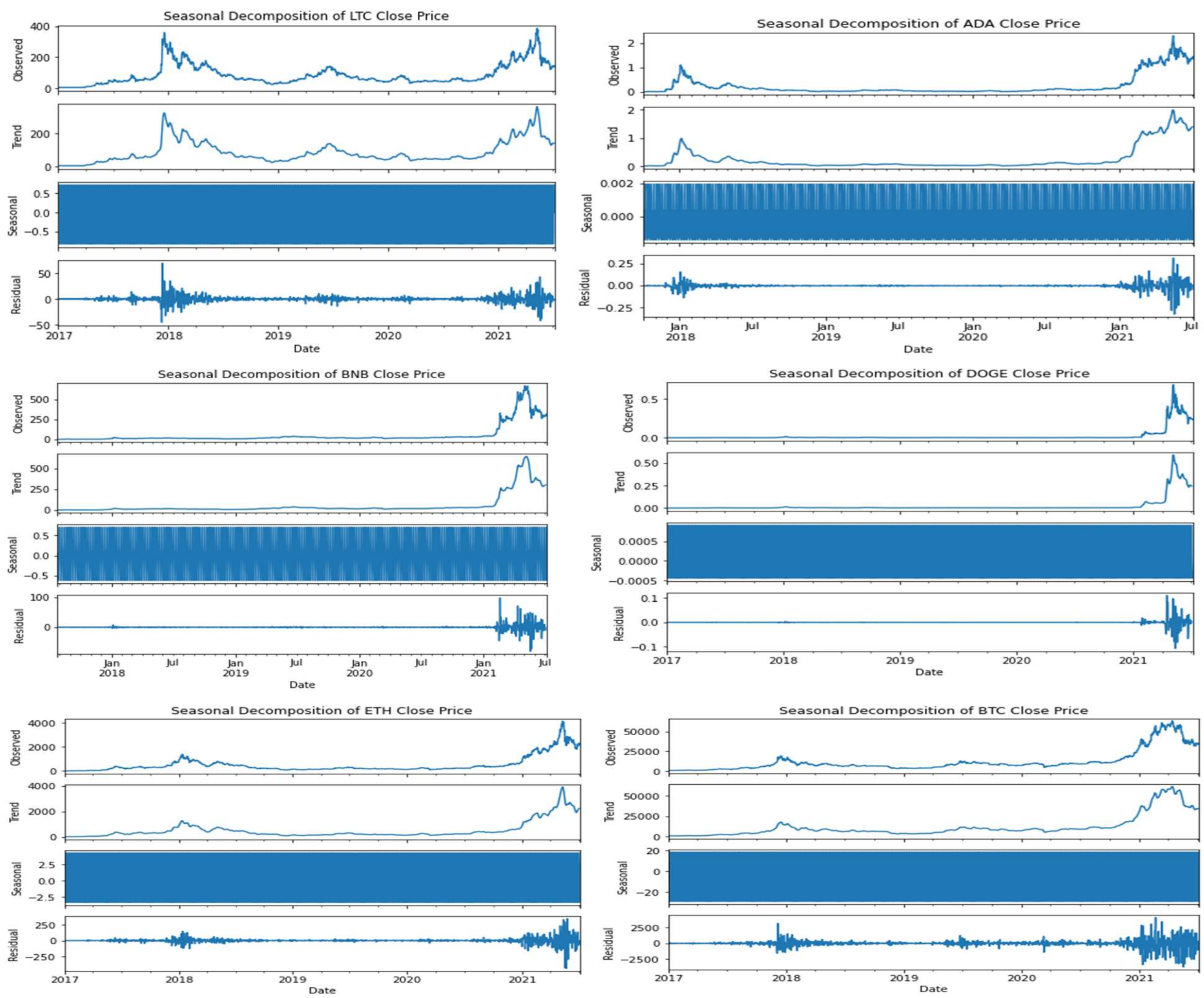
1. In the field of data science, time series decomposition techniques are used to break down a time series dataset into its constituent components, including trend, seasonal, and remainder components.
2. These techniques help analysts identify patterns and trends within the data, which can be useful in making predictions or understanding underlying relationships.
3. The article covers three different time series decomposition techniques: classical decomposition, seasonal decomposition of time series (STL), and empirical mode decomposition (EMD).
4. Classical decomposition involves identifying the trend, seasonal, and remainder components of a time series, while STL uses a non-parametric approach to extract seasonal and trend components. EMD decomposes a time series into intrinsic mode functions (IMFs) that represent different frequencies in the data.
5. The choice of decomposition technique depends on the specific dataset and research question at hand. By using time series decomposition techniques, data scientists can gain a deeper understanding of time series data and extract meaningful insights.

The python code performs seasonal decomposition of cryptocurrency close prices using the statsmodels library in Python.

Here is what the code does step by step:

- Import necessary libraries - pandas, matplotlib, and statsmodels.
- Load the cryptocurrency data into a pandas dataframe using pd.read_csv.
- Create a new dataframe df_time_series with the necessary columns for time series analysis (Date, Close, and Symbol).
- Convert the Date column to datetime format using pd.to_datetime and set it as the index of the dataframe using set_index.
- Group the data by Symbol and resample it to a daily frequency using groupby and resample.
- Interpolate missing values in the resampled data using linear interpolation.
- Loop through each unique cryptocurrency Symbol in the dataframe.
- Perform seasonal decomposition on the close price data for the current cryptocurrency using sm.tsa.seasonal_decompose with the additive model.
- Create a 4x1 subplots for each component of the seasonal decomposition (observed, trend, seasonal, and residual) using plt.subplots.

- ➢ Plot each component in the appropriate subplot using .plot on the res object returned by seasonal_decompose.
- ➢ Show the plot using plt.show().



The output of this code is a set of four plots for each cryptocurrency, showing the observed, trend, seasonal, and residual components of the time series decomposition. These plots can be used to analyze the patterns and trends in the cryptocurrency close prices over time.

The goal of seasonal decomposition is to gain a better understanding of the underlying structure of the time series and to identify any patterns or trends that may exist. This can be useful for making predictions about future values of the time series, as well as for identifying potential relationships between the time series and other variables.

The Augmented Dickey-Fuller (ADF) test

The Augmented Dickey-Fuller (ADF) test is a statistical test that aims to determine the stationarity of a given time series. Stationarity refers to the property of a time series in which its statistical characteristics, such as the mean and variance, remain constant over time. This is a crucial assumption in many time series models, as it simplifies the analysis and interpretation of the data.

As a unit root test, the ADF test examines whether a unit root exists in the time series, which is a characteristic of a non-stationary time series that tends to drift over time and does not revert to its mean. The ADF test compares the differences between the observed values in the time series and the expected values if the series were stationary. If the differences are significant enough, the null hypothesis of a unit root is rejected, indicating that the time series is stationary.

The ADF test is based on the augmented Dickey-Fuller regression model, which is a linear regression model that includes lagged differences of the time series as explanatory variables. The ADF test examines the coefficient of the lagged first difference in the regression equation, which represents the rate at which the time series reverts to its mean.

The ADF test has several variations, including the ADF test with a constant and trend term, which allows for a linear trend in the time series. The ADF test can also be applied to seasonal time series, where the seasonal differences are included in the regression equation.

An article on “Augmented Dickey Fuller Test (ADF Test) – Must Read Guide” [12] by Selva Prabhakaran, explains that

1. Stationarity testing is a statistical method utilized to determine whether a time series maintains its consistency in terms of mean and variance over time. When a time series remains stationary, it offers a more straightforward analysis and modeling approach, as compared to non-stationary time series.
2. The ADF test determines if a time series is stationary assuming it has a unit root and is non-stationary. It calculates a p-value that shows the likelihood of obtaining the observed test statistic under the null hypothesis. Rejecting the null hypothesis if p-value < 0.05 categorizes the time series as stationary. The ADF test is crucial for modeling and analyzing time series data.
3. an example of performing the ADF test utilizing the Python's statsmodels library is demonstrated, which involves using a dataset containing the daily closing prices of the S&P 500 index from 2000 to 2021.
4. Additionally, the article explains the commonly used technique of first-order differencing to convert a non-stationary time series into a stationary one, which involves computing the difference between consecutive observations.
5. Finally, the article acknowledges some limitations of the ADF test, such as its sensitivity to the selection of lag length and the assumption of no seasonality. The article recommends using additional tests in conjunction with the ADF test to obtain a more robust assessment of stationarity.

ADF Test Results	ADF Statistic	p-value	1% Critical Value	5% Critical Value	10% Critical Value
LTC	-2.71	0.07	-3.43	-2.86	-2.5
ADA	-0.37	0.92	-3.44	-2.86	-2.56
BNB	-1.19	0.68	-3.43	-2.86	-2.57
DOGE	-2.50	0.11	-3.43	-2.86	-2.56
ETH	-0.29	0.93	-3.43	-2.86	-2.57
BTC	-0.85	0.80	-3.43	-2.86	-2.57

The ADF test outcomes contain insights about the stationarity of each cryptocurrency's time series. The ADF test's null hypothesis is that the time series is non-stationary and has a unit root. If the p-value is below the chosen significance level (often 0.05), the null hypothesis is rejected, indicating that the time series is stationary.

Based on the ADF test results, it is evident that the p-values for all the cryptocurrencies are greater than 0.05. As a result, we fail to reject the null hypothesis, and we cannot claim that any of the cryptocurrencies have a stationary time series.

The ADF Statistic value in the ADF test is a measure of how much the time series deviates from being stationary. It indicates the extent of non-stationarity in the time series. A more negative ADF statistic value provides stronger evidence against the null hypothesis of stationarity.

To summarize, based on the ADF test results, it is indicated that the time series of each cryptocurrency might not be stationary. The results show that none of the cryptocurrencies have p-values less than the significance level of 0.05, which indicates that we cannot reject the null hypothesis of non-stationarity. Additionally, the ADF Statistic values are not more negative than their respective critical values, which also supports the conclusion of non-stationarity.

In conclusion, the ADF test is a powerful tool for testing the stationarity of time series data. It is widely used in many areas of research, including finance, economics, and environmental science. The ADF test provides a robust and reliable method for testing the null hypothesis of a unit root in a time series, allowing researchers to make informed decisions about the statistical properties of their data.

Reference

1- Hern, A. (2018, January 29). *Bitcoin and cryptocurrencies – what digital money really means for our future*. The Guardian. http://www.theguardian.com/technology/2018/jan/29/cryptocurrencies-bitcoin-blockchain-what-they-really-mean-for-our-future

2- *How Blockchain Could Disrupt Banking*. (2022, October 18). CB Insights Research. https://www.cbinsights.com/research/blockchain-disrupting-banking/

3- Cryptocurrency Historical Prices. (n.d.). Cryptocurrency Historical Prices | Kaggle. https:///datasets/sudalairajkumar/cryptocurrencypricehistory

4- Salman, M. K., & Ibrahim, A. A. (2020, November 1). price prediction of different cryptocurrencies using technical trade indicators and machine learning - IOPscience. https://doi.org/10.1088/1757-899X/928/3/032007

5- Abraham, Jethin; Higdon, Daniel; Nelson, John; and Ibarra, Juan (2018) "Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis," SMU Data Science Review: Vol.1: No. 3, Article 1. https://scholar.smu.edu/datasciencereview/vol1/iss3/1

6- SARPAL, & RANTA. (n.d.). CRYPTOCURRENCY PRICE PREDICTION USING DEEP LEARNING. In Chrome-extension://Efaidnbmnnnibpcajpcglclefindmkaj/http://www.ir.juit.ac.in:8080/jspui/bitstream/123456789/3602/1/Cryptocurrency%20Price%20Prediction%20Using%20Deep%20Learning.pdf

7- Finding the right balance between volatility and returns. (n.d.). RBC Global Asset Management. https://www.rbcgam.com/en/ca/learn-plan/investment-basics/understanding-the-relationship-between-volatility-and-returns/detail

8- Technical Analysis: What It Is and How to Use It in Investing. (2022, March 14). Investopedia. https://www.investopedia.com/terms/t/technicalanalysis.asp

9- Gupta, K. (2021, June 25). Generate Reports Using Pandas Profiling, Deploy Using Streamlit. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/06/generate-reports-using-pandas-profiling-deploy-using-streamlit/
10- Brems, M. (2022, January 26). A One-Stop Shop for Principal Component Analysis. Medium. https://towardsdatascience.com/a-one-stop-shop-for-principal-component-analysis-5582fb7e0a9c
11- Plummer, A. (2020, September 15). Different Types of Time Series Decomposition. Medium. https://towardsdatascience.com/different-types-of-time-series-decomposition-396c09f92693
12- Prabhakaran, S. (2019, November 2). Augmented Dickey-Fuller (ADF) Test - Must Read Guide - ML+. Machine Learning Plus. https://www.machinelearningplus.com/time-series/augmented-dickey-fuller-test/