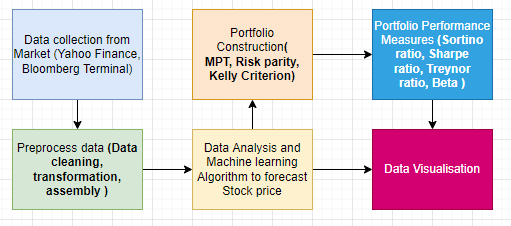
### **3. Methodology**

***[Here you bring in all the methods you know that you came across in your literature review. You compare them. Finally, at the end of this chapter, you must show that you have justified your choice of the methods you are using in Chapter 4 for experimentation/simulation and modeling]***

#### In this chapter, Methodology, Forecasting Algorithms, Portfolio construction, and performance measures will be discussed/compared that will be used to build a portfolio construction framework.

#### **3.1. Outline of the Methodology**

Precise methodology should be followed to build a machine learning model. There are six steps to follow for the development of models: Data Collection, Data pre-processing, Data analysis and Forecasting, Portfolio construction, and portfolio performance analysis. The below figure illustrates the order of six steps methodology:



*Figure 3.1 Methodology outline for Experiment*

1. **Data collections:** There are several cryptocurrency data gathering sources, including CoinMarketCap, Yahoo Finance, Binance, and Bloomberg terminal. Data will be gathered in two formats: daily data and intraday data. Daily data is easily available through Yahoo Finance, which provides data from CoinMarketCap, however intraday data is not free. The Bloomberg terminal at the University of Westminster will be used to collect data for intraday data.
2. **Pre-process Data:** It is a critical phase for every machine learning project since the data collected in step 1 is cleaned and transformed. It can be utilized effectively while developing an ML model.
3. **Stock Price Forecasting:** As per the literature review, various machine learning algorithms are used in different studies to predict the price of cryptocurrency/stocks for example Random forest, Linear regression, Support vector Regression and Neural networks (CNN, RNN, LSTM, etc). As part of this section, we will decide which algorithm should be employed to predict the price of an asset. An Exploratory Data Analysis (EDA) will be performed to fully comprehend the data. The purpose of exploratory data analysis is to uncover potential insights or patterns in the data.

#### **Portfolio construction**: In this stage, the optimal weight for each asset will be computed, which will maximize the portfolio's return while decreasing the risk. We shall employ approaches such as the Mean-variance portfolio method, the Equal-weight portfolio method, Kelly's criterion, and others.

#### **Portfolio performance Measure:** It directs investors to examine portfolio performance to enhance returns while decreasing risk. A few approaches will be employed to measure performance, as explained in section 3.3.

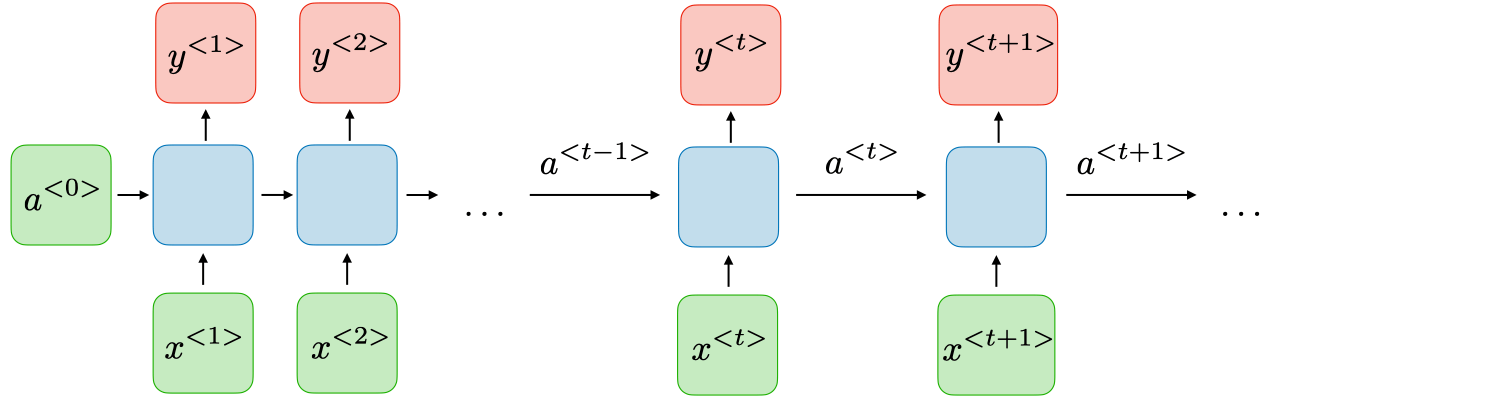
#### **3.2 Machine Learning Algorithms for Crypto Price Prediction**

As discussed in section 2.2 of the literature review, the Forecasting of the asset price is the core problem of portfolio management for cryptocurrency. Different supervised, unsupervised, and deep learning algorithms could be used to forecast pricing, according to many studies and research papers.

**Support vector Machine** is a nonlinear machine learning technique that supervises machine learning. It can be used for time-series forecasting, but it lacks the precision and accuracy of random forest. It can be enhanced by removing irrelevant and disorganized data points. **Random forest** is a decision-tree model that adheres to the tree structure. It is effective at removing network instabilities, and if the optimal feature with the maximum information efficiency is chosen, it outperforms SVM in price prediction. **K-mean** is an unsupervised machine-learning algorithm that can be used to form clusters of data with similar characteristics. The Elbow method can be used to derive the ideal number of clusters. It can be used to find a group of assets that have some similar characteristics, but It is not suitable for predicting cryptocurrency prices.

Deep learning methods are another common machine learning methodology used to tackle issues in a variety of domains, such as automated driving, aerospace and defense, medical research, and finance. Deep-learning methods use traditional neural networks that’s why it is called deep-learning neural network. There are a few deep learning methods that are very popular to forecast the price prediction of cryptocurrency i.e. A recurrent neural network (RNN) and Long short-term memory neural network (LSTM).

RNN is a forward and backward dynamic network that has input, output, and context layers. Typically, RNN traditional architectures are as follows:



*Figure-3.2 Architecture of the RRN network*

For each timestep t, the activation a<T> and the output y< t > are expressed as follows:

** **(4)**

Where *Wax, Waa, Wya* and *ba,by* are coefficients that are shared temporally and *g1* and *g2* activation functions.

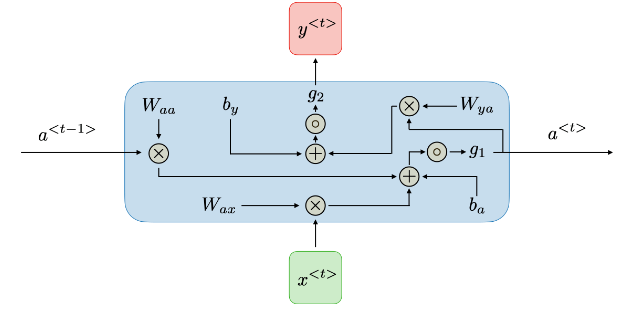


Figure-3.2.2

**Advantage of RNN as follows:**

* The ability to handle input of any duration
* Model size does not increase with input size
* Computation takes into account past data
* Weights are shared throughout time

**Disadvantage of RNN as follow:**

* Slow computation
* Difficulty obtaining information from a long time ago
* Inability to evaluate any future input for the current state

In the context of RNNs, the vanishing, and exploding gradient phenomena are frequently found. They occur because it is difficult to capture long-term relationships due to the multiplicative gradient, which might be exponentially decreasing/increasing with respect to the number of layers. Traditional RNNs suffer from the vanishing gradient problem, which is addressed by gated recurrent units (GRU) and long short-term memory units (LSTM), with LSTM being a generalization of GRU.

Long Short-Term Memory networks, sometimes known as "LSTMs," are a kind of RNN that can learn long-term dependencies. Hochreiter and Schmidhuber (1997) introduced them, and numerous individuals developed and popularised them in subsequent work. They operate extremely effectively on a wide range of issues and are now frequently employed.

The LSTM model converts a series of previous observations into an output observation. The split-sequence () function is used to divide an input series into output samples, each with input and output timesteps.

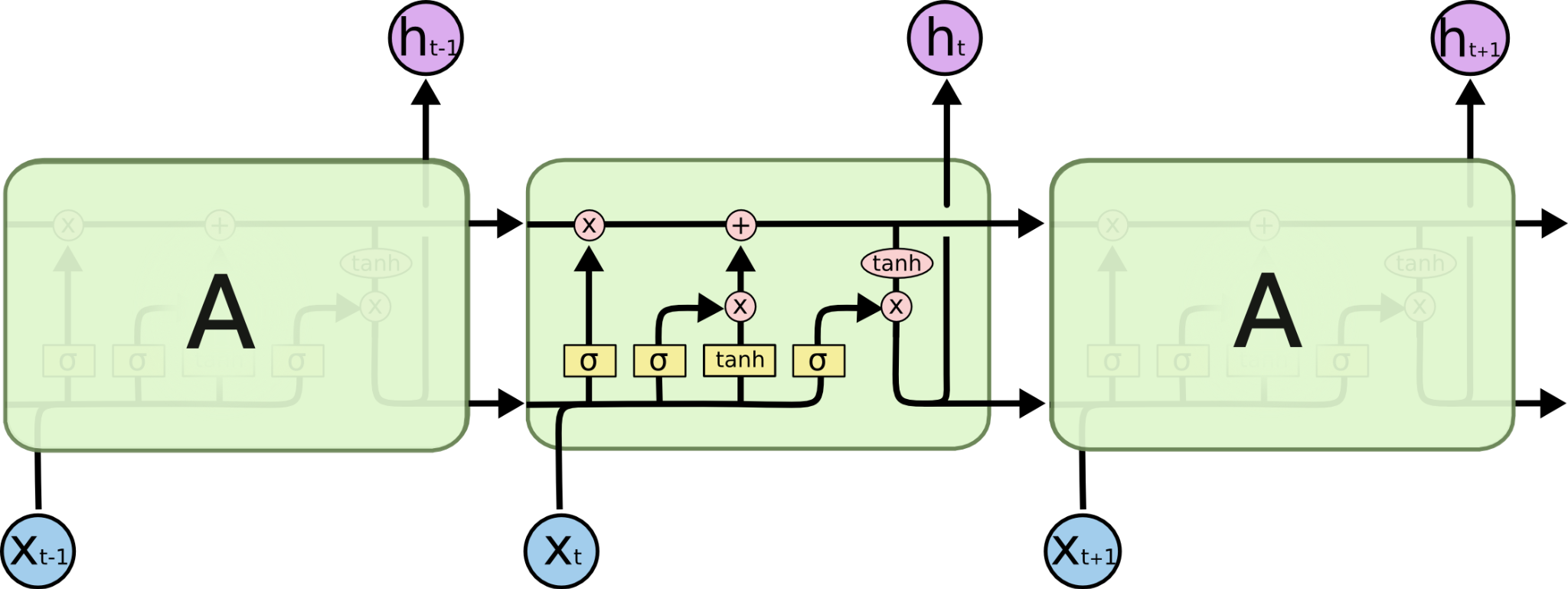


Figure-3.2.3

#### 

#### 

#### Figure-3.2.4

#### The cell state, represented by the horizontal line across the top of the picture, is crucial to LSTMs. The cell state is similar to a conveyor belt. It follows the whole chain, with only a few small linear interactions. It is quite simple for information to just pass over it unmodified. Because LSTMs handle both gradient and signal passing problems with long-term dependence, they are preferred to RNN.

***Multivariate v/s univariant model with research paper.***

<https://www.mdpi.com/1911-8074/14/10/486>

#### **3.3 Portfolio Construction and Performance Analysis Methods**

As part of the literature study, portfolio construction methods and portfolio performance analysis measures have been identified and details about those methods are given in this section.

**Equal weight method**

One of the fundamental portfolio construction methods is the equal weight technique, which assigns equal weightage to all assets. It is sometimes referred to as the 1/N approach or the naïve portfolio creation method.

Wi = 1/N, **(5)**

For ∀i = 1, ..., N

Where,

Wi is the weight allocated to the ith asset

N is the number of assets in the portfolio

**Kelly’s criteria:**

The Kelly criterion's delightfully simple formula estimates the ideal proportion of your bankroll to bet in order to maximize the geometric growth rate of wealth. However, it not only guarantees you maximum profit by successfully maximizing your possibilities; it also guarantees you safety from a gambler's disaster.

*f = (bp - q)/b* **(6)**

Where

*f = the fraction of the bankroll to bell*

*b = the odds*

*p = probability of winning*

*q = probability of losing i.e., 1-p*

Below are the prerequisites of Kelly’s criteria:

1. Odd and probability should be known
2. scale the Kelly output such that your bet equals your possible loss
3. If only one of them is in your favour, it must outweigh the other.

**Min-Variance Portfolio:**

Markowitz's mean-variance portfolio optimization methodology can be thought of as the selection of portfolio weights (x) that maximize the Sharpe ratio. The portfolio selection issue as defined by Markowitz [1952] can be written as

minimize *σr2* = xTVx **(7)**

subject to xT1 = 1, xTR = Rp.

Where

Rp: n-column vector of mean returns

T: the transpose of a vector

V: the (n \* n) covariance matrix with entries *σ*ij, i, j = 1, 2 . . . . n.

X: n -column vector whose components Xl, . . . Xn denote the weight of the investor's wealth

allocated to the ith asset in the portfolio with i = 1, 2 . . . . n.

To minimize the portfolio variance constraints: first, the portfolio weights must total to unity, indicating that all of the wealth is invested, and second, the portfolio must have an expected rate of return equal to Rp.

**Hierarchical Risk Parity:**

Lopez de Prado (2016) described how to create diversified portfolios using the Hierarchical Risk Parity method. The Hierarchical Risk Parity approach constructs a portfolio using data from the assets' covariance matrix. In comparison to conventional risk parity approaches, hierarchical risk parity methods also result in less risky portfolios outside of the sample. Using graph theory and machine learning, this risk parity technique attempts to overcome the flaw in the Critical Line Algorithm (CLA), which Markowitz developed in 1954.

Below are performance measures that can employ to the performance analyse of the crypto portfolio:

1. **Sharpe Ratio:** William F. Sharpe established the Sharpe ratio, which is the ratio of portfolio total return minus risk-free rate divided by portfolio standard deviation, which is a measure of risk.

***Sharpe Ratio*** = (*Rp*−*Rf)/σp* **(8)**

where:

*Rp =* return of portfolio

*Rf =* risk-free rate

*σp* = standard deviation of the portfolio’s excess return

1. **Sortino Ratio:** Sortino ratio is a variant of the Sharpe ratio that utilizes downside deviation rather than standard deviation to assess risk, i.e. only returns that fall below a user-specified goal or needed rate of return is deemed dangerous.

**Sortino Ratio** = *(Rp*−*rf)/σd***(9)**

where:

*Rp* = Actual or expected portfolio return

*Rf* = Risk-free rate

*σd* = Standard deviation of the downside

1. **Downside Standard deviation:** The portfolio return's downside standard deviation shows if it is or is not below the minimum allowed return. The Sortino ratio is also computed using it. Since upside deviation is not covered, downside deviation does not provide all the necessary information.
2. **Drawdown:** A drawdown is a financial word that refers to the drop in the value of a single investment or an investment portfolio from a relative peak to a comparable trough. A drawdown is a significant risk issue that investors must consider.
3. **Volatility:** A stock portfolio's value fluctuates from day to day. This is known as volatility, and it is a type of risk. The formula calculates the standard deviation from the variance of each stock's return in the portfolio by taking the square root of the sum.

***Portfolio variance*** = *w12σ12 + w22σ22 + 2w1w2Cov1,2* **(10)**

where:

w1 = the portfolio weight of the first asset

w2 = the portfolio weight of the second asset

σ1= the standard deviation of the first asset

σ2 = the standard deviation of the second asset

Cov1,2 = the covariance of the two assets, which can thus be expressed as p(1,2)σ1σ2,

where p(1,2) is the correlation coefficient between the two assets

1. **Annualized Return:** The geometric average return of portfolio for each year over given period of the time is call annualized return.Annualized total return gives details about portfolio’s performance does not provide information about price fluctuations or volatility.

***Annualized Return****=​((1+r1​) ×(1+r2​) ×(1+r3​) ×⋯×(1+rn​))1/n​−1​***(11)**

r = return over the period

n = number of years investment held

1. **Mean return:** The mean return of the portfolio can be used to determine the expected return or loss for the portfolio. Additionally, it aids in establishing a link between portfolio risk and return.

**3.4 Dataset Understanding and Collection Methods**

The next part will offer a detailed knowledge of the dataset gathering methods from various sources, as well as exploratory data analysis of the dataset to comprehend the cryptocurrency dataset.

**3.4.1 Data Source and Collections**

The data for this project has been gathered from reputable and trustworthy sources like Yahoo Finance, CoinMarketCap, and Bloomberg. Yahoo Finance uses CoinMarketCap APIs for daily and historical daily data on around 9089 distinct cryptocurrencies. For the previous 5 years, the most renowned top 20 coins statistics have been collected using Pandas' DataReader package. Intraday data is also required for this project, which is not available on Yahoo Finance, hence the Bloomberg terminal is utilized to obtain the Intraday data. Bloomberg has been providing crypto data since 2013, and after 2018, the scope of instruments increased from 10 to 50, as well as providing additional information like headlines, historical intraday data, and other fundamental information. Bloomberg terminal facility is provided by the University of Westminster at the Marylebone campus. Cryptocurrency data can be downloaded after obtaining credentials for the Bloomberg terminal.

The datasets give up-to-date information on the cryptocurrency's price. Details about the dataset's fields, such as column name, type, and description, are provided below:

|  |  |  |  |
| --- | --- | --- | --- |
| **SR no** | **Column name** | **Type** | **Description** |
| 1 | Date | Date/Datetime | Date/Datetime of cryptocurrency price recorded |
| 2 | Open | Float64 | Opening price of the day in USD |
| 3 | High | Float64 | Highest price of the day in USD |
| 4 | Low | Float64 | Lowest price of the day in USD |
| 5 | Close | Float64 | Closing price of the day in USD |
| 6 | Adj Close | Float64 | the closing price after adjustments for all applicable splits and dividend distributions |
| 7 | Volume | Float64 | Volume of the transactions |
| 8 | Market Cap | Float64 | Market capitalization in USD |

#### **Table**

#### **3.4.2 Exploratory Data Analysis**

An exploratory data analysis is the most efficient technique to comprehend the components of the data and its characteristics. The aim of exploratory data analysis (EDA) is to discover what the data can tell us apart from applying machine learning models to it.

The exploratory analysis has been carried out using Python's data visualization packages (matplotlib, plotly, seaborn, pandas, DateTime) and data from 20 cryptocurrencies. First, for all columns, use the pandas' library's isnull() and sum() functions to find null values and convert string values to suitable data types (date value to DateTime and currency to float datatype). There are no missing values in any row, although data for some coins before 2018 are not available because several cryptocurrencies were introduced after 2018 ***(<Example: Needs to be added later>).*** As a result of that further analysis has been performed on the top 5 coins by market capitalization. Market capitalization (or market cap) is the total worth of all coins created for a cryptocurrency. It is computed by multiplying the current market price of a single coin by the number of coins in circulation. To achieve this, horizontal bar charts have been used. Figure-3.4.2.1 represents the top five crypto coins by market capitalization which are bitcoin, Ethereum, Tether, Binance Coin, and USD Coin. Among the top coins, bitcoin has the highest, whereas USD coin has the lowest market capitalization.



Figure-3.4.2.1

Further exploring the end of the day price of all the assets, Bitcoin has the highest close price (refer to Figure-3.4.2.2), followed by etherium(refer to Figure-3.4.2.3), binance (refer to Figure-3.4.2.4). The close price of Tether and USD coins is lowest as compared to other coins (refer to Figure-3.4.2.4).

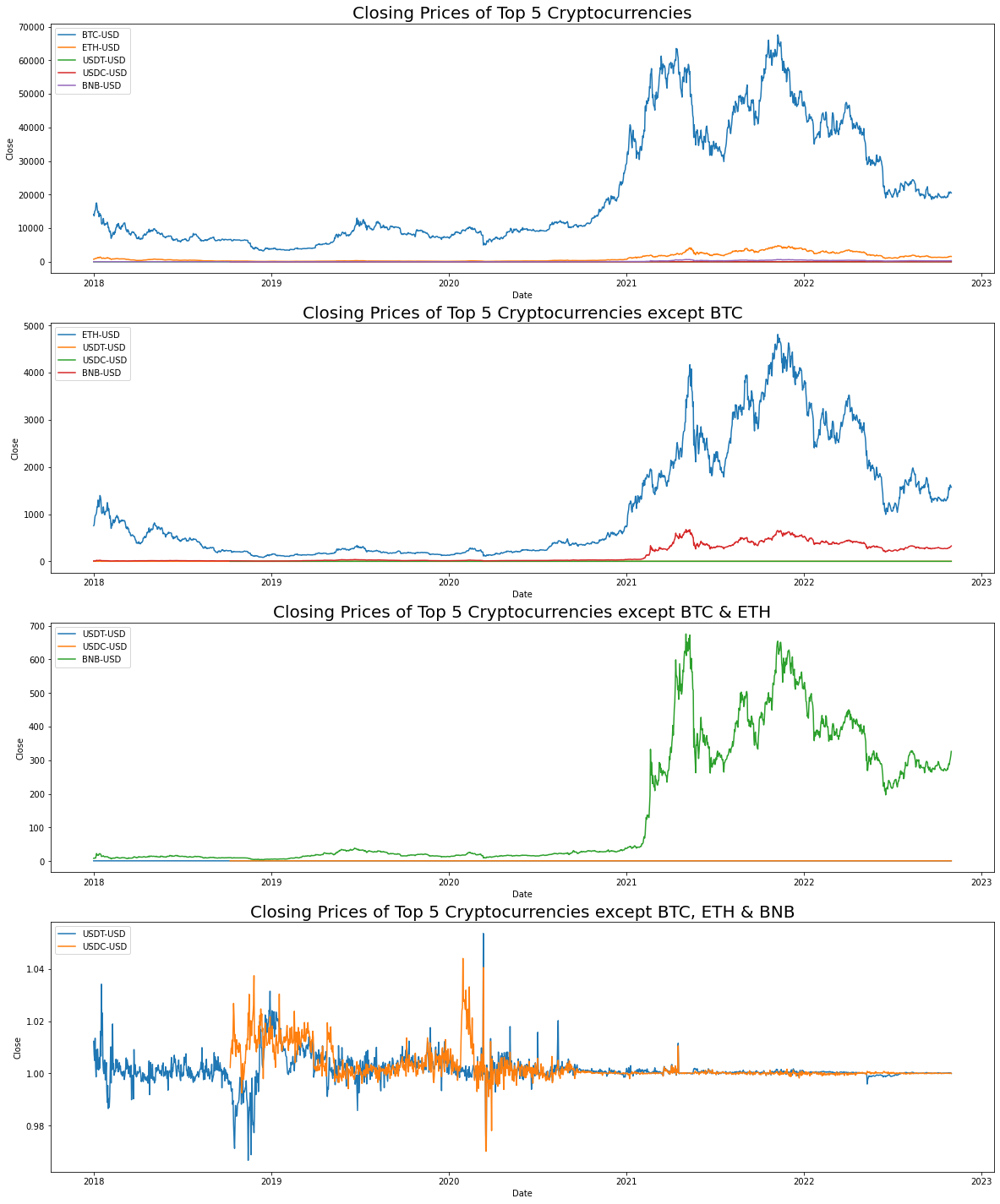


Figure-3.4.2.2

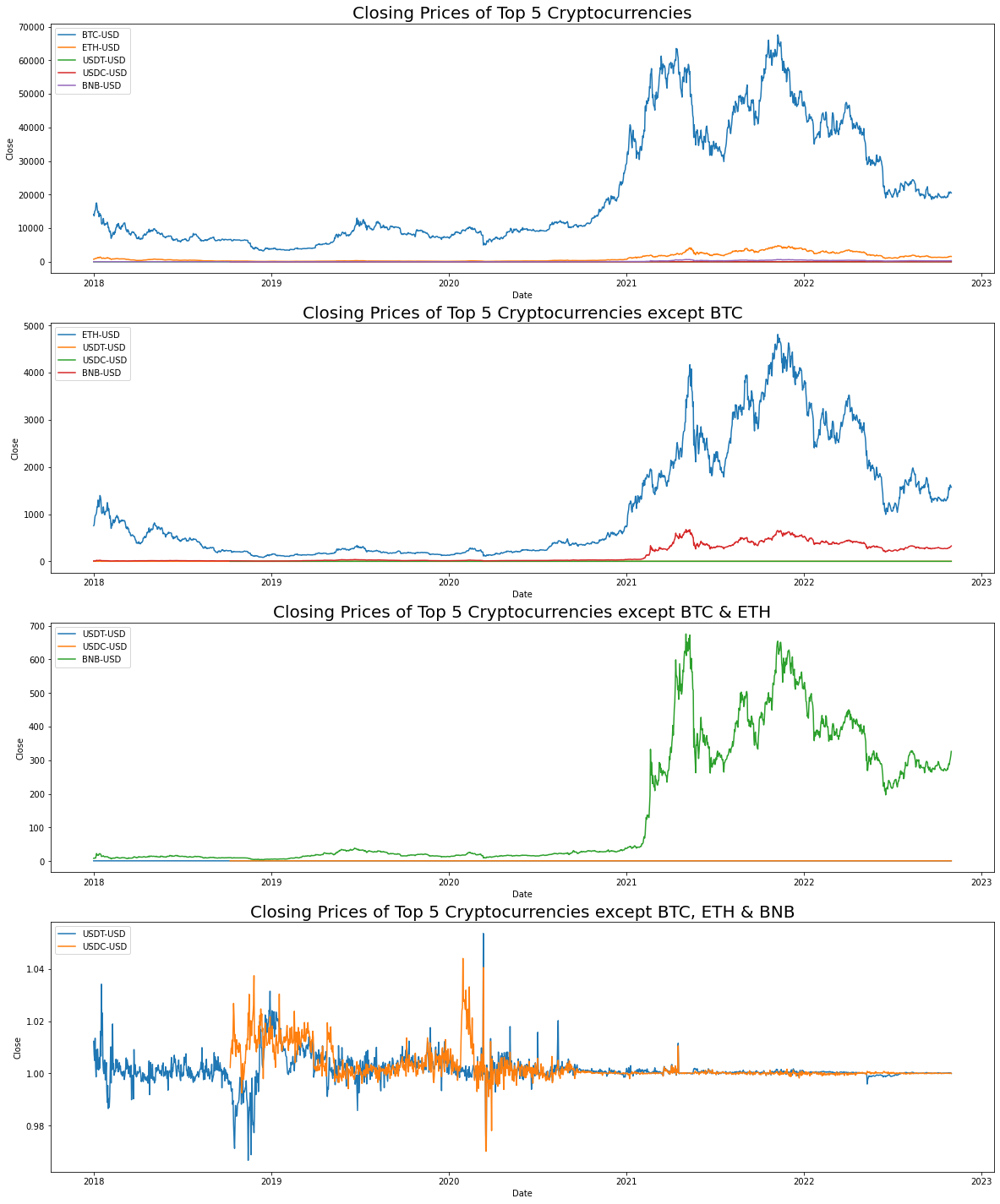


Figure-3.4.2.3

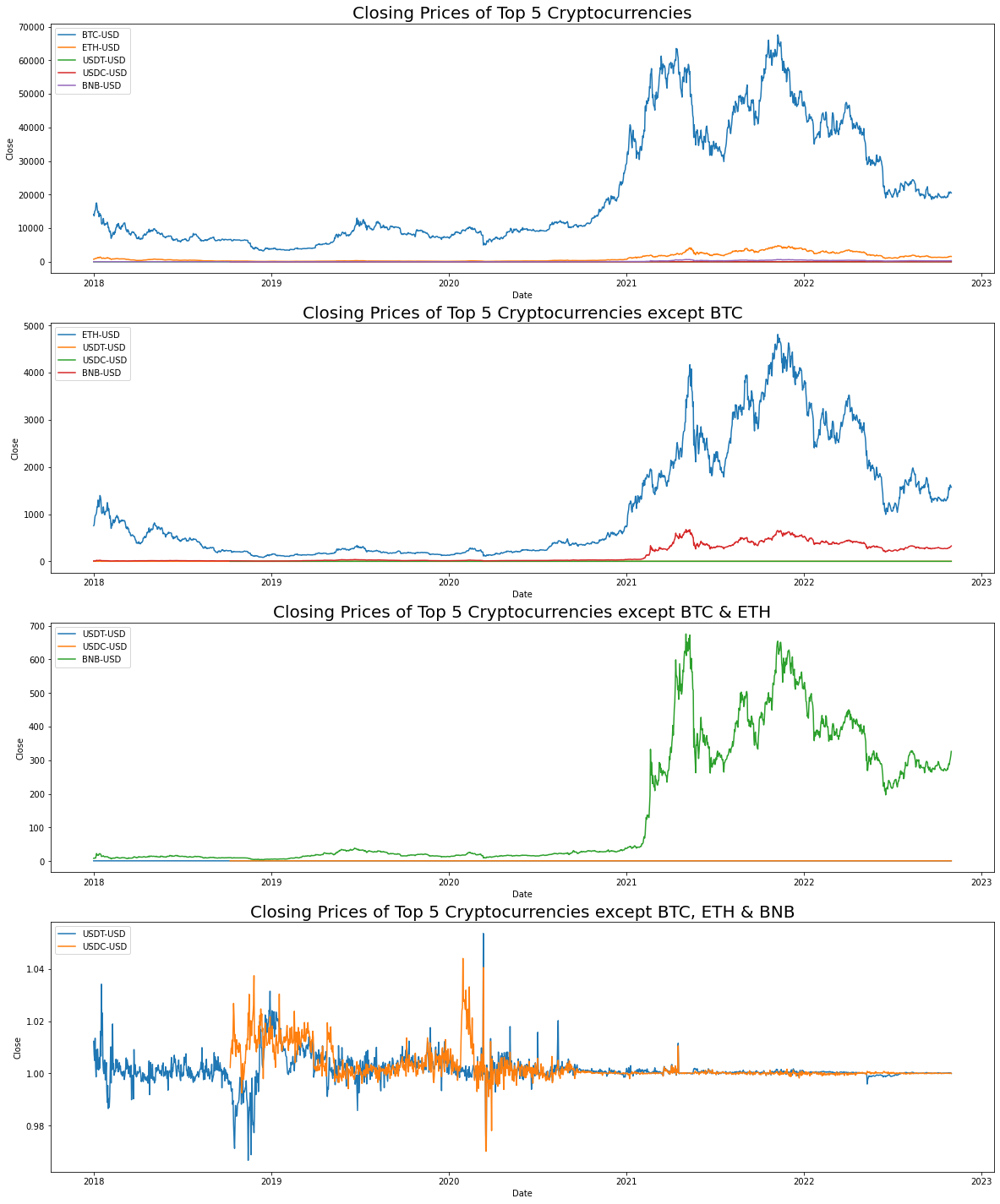


Figure-3.4.2.4

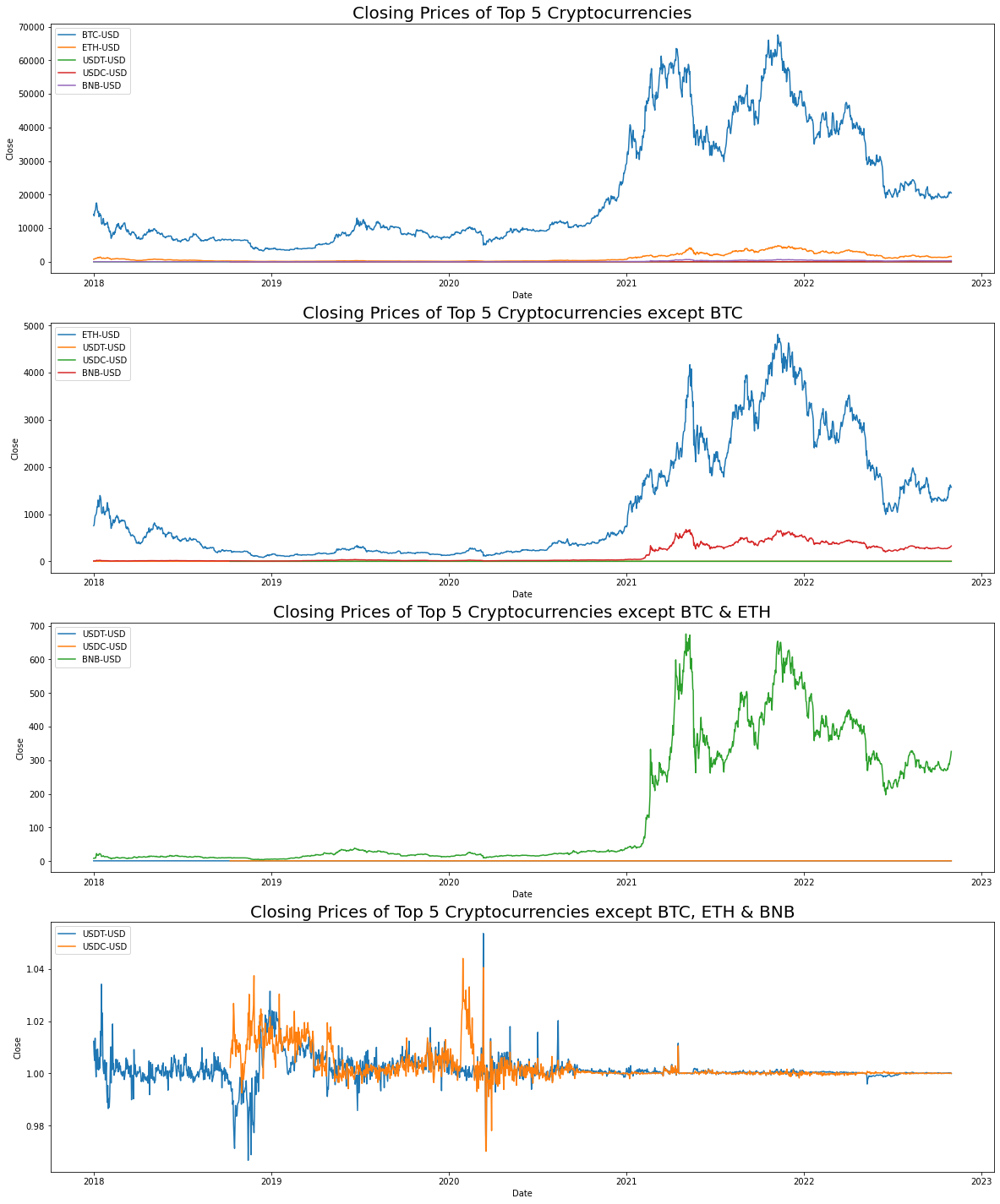


Figure-3.4.2.5

Moving Average is a common indicator for asset technical analysis. It aids in identifying bear signals (price drops) and bull signals (price rises). Traders keep an eye on the 50-day and 200-day moving averages, which are used to calculate the death cross (50-day moving average crossing below the 200-day moving average) and golden cross (50d moving average crossing above the 200d moving average). Both 50 days and 200 days Moving averages are derived using rolling(window=n).mean() function of the dataframe of the pandas' package and plot for moving averages versus the close price for all five coins are shown below (refer to Figures-3.4.2.6 through Figure-3.4.2.10):

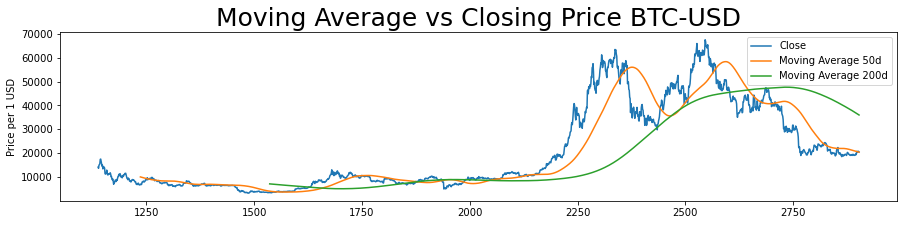


Figure-3.4.2.6

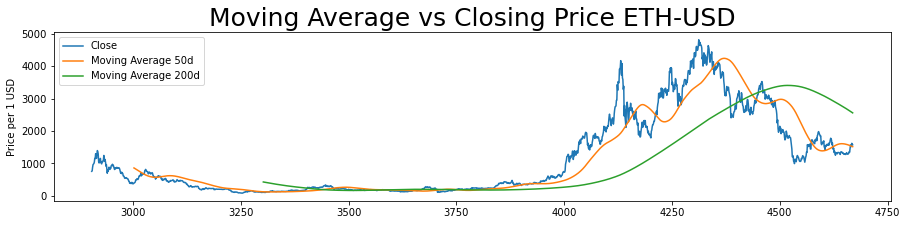


Figure-3.4.2.7

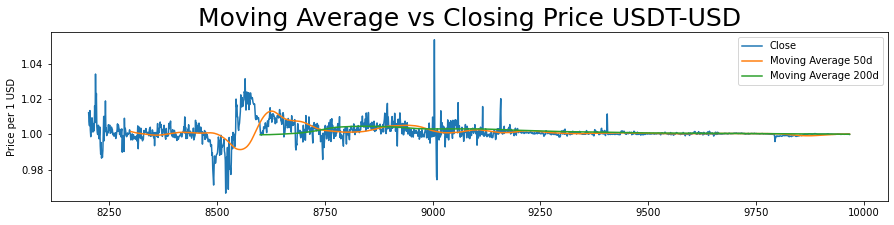


Figure-3.4.2.8

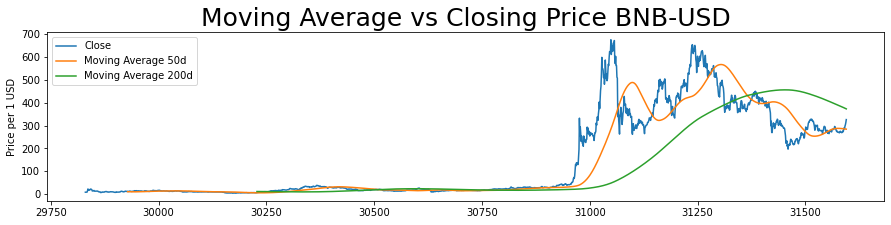


Figure-3.4.2.9

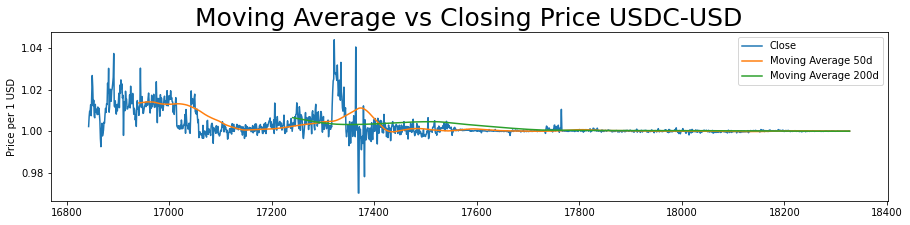
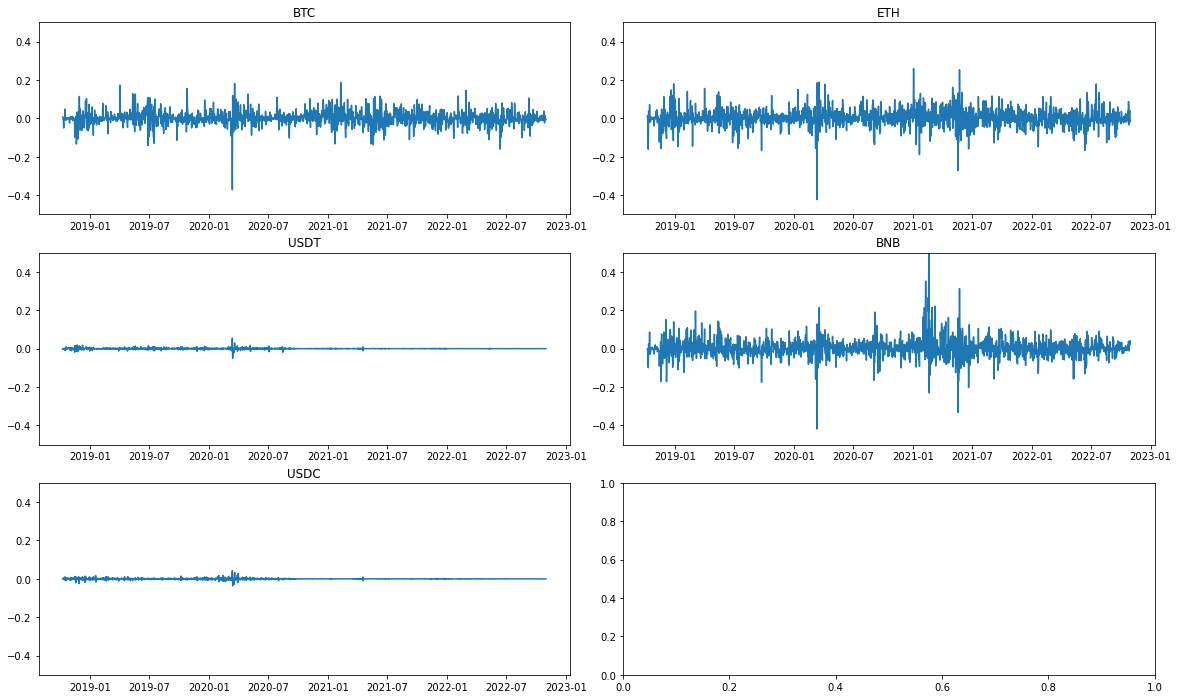


Figure-3.4.2.10

There are too many death crosses and golden crosses in USDC-USD and USDT-USD, indicating that those assets are more volatile than the other coins.

A return is a change in an asset's price over time. Positive returns indicate a profit, whereas negative returns indicate a loss. To compute the results, we will utilise the pandas pct change() method. Return of all coins are shown in Figure-3.4.2.11. BNB is the most volatile cryptocurrency, followed by ETH, whereas BTC is the least volatile. There was a massive fall in all cryptocurrencies around March 2020 during the initial COVID pandemic wave.



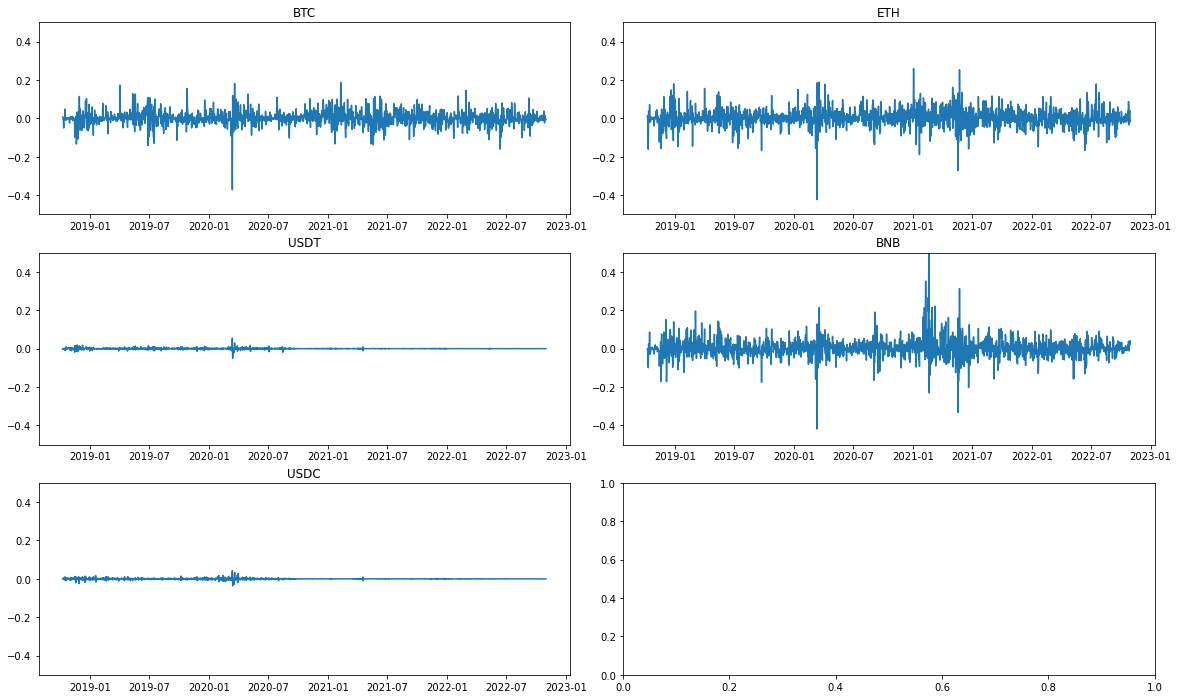


Figure-3.4.2.11

The corr() function in the pandas package is used to calculate the correlation of daily returns of the top five bitcoins. It depicts the correlation between daily returns for all coins. The following is a heatmap depicting the correlation of daily returns:

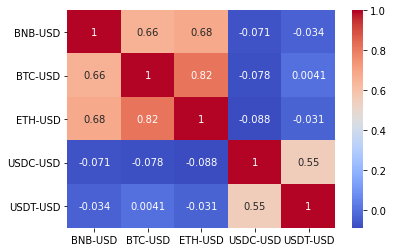
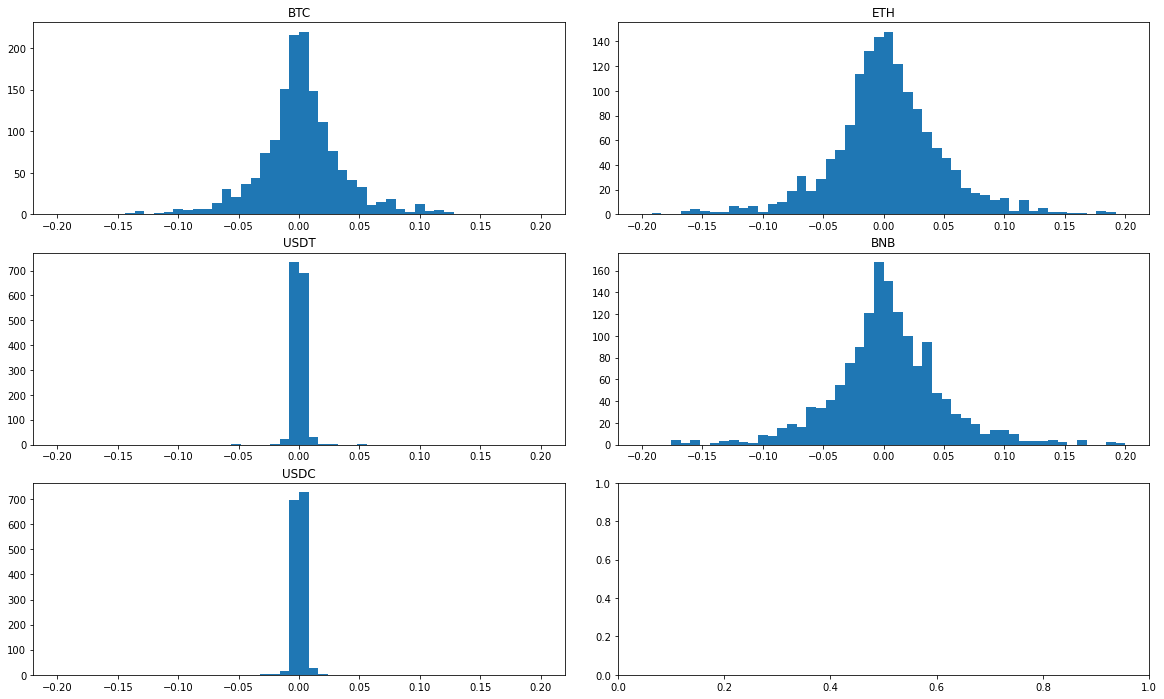


Figure-3.4.2.12

According to the heatmap, Bitcoin and Ethereum are significantly associated, whereas Ethereum and Tether have the least correlation.

Volatility(standard deviation of the returns) is a measure of how an asset's price changes over time. The greater an asset's standard deviation, the more volatile it is. The histogram in Figure 3.4.2.13 represents the volatility of all coins. USD and Tethra are more volatile, whereas Binance and Ethereum are the least volatile assets.



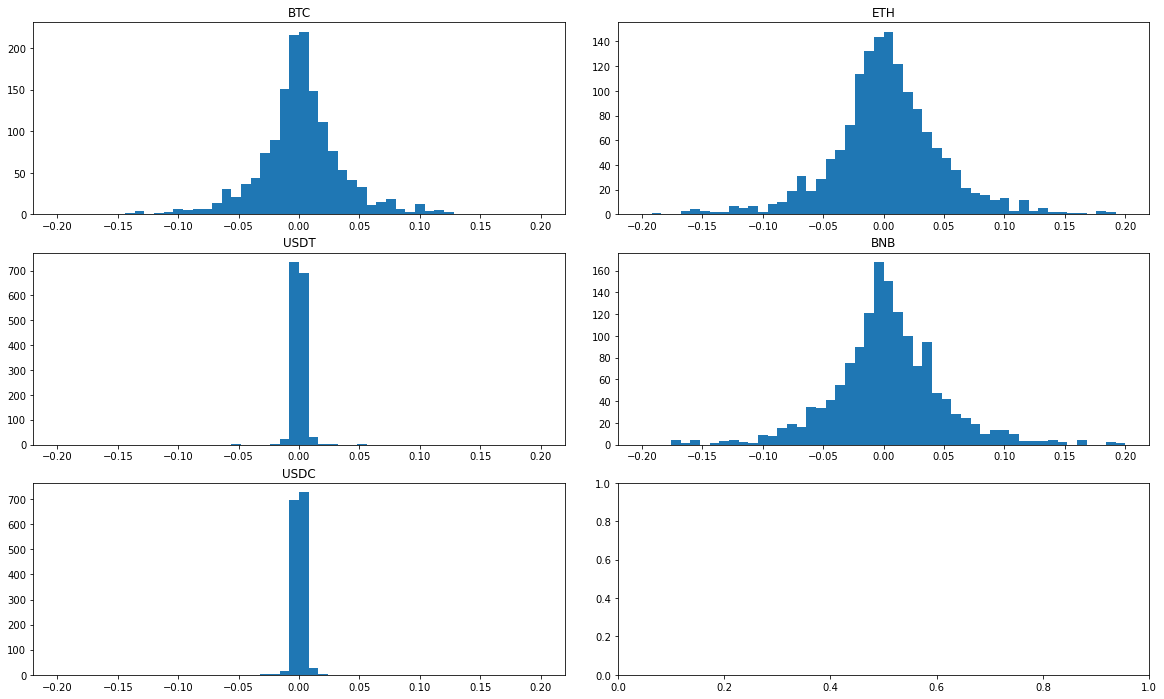


Figure-3.4.2.13

The overall change in the price of an asset over time is expressed by the cumulative return. To compute the daily cumulative simple returns, the pandas cumprod() method is utilized. The cumulative returns of all coins are represented using a line chart in Figure 3.4.2.1.

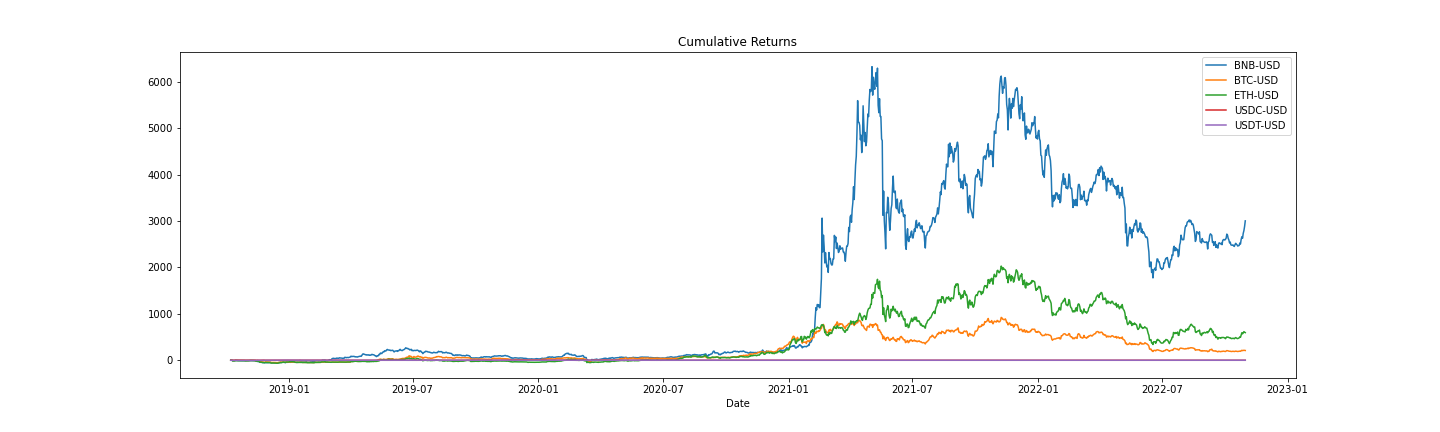


Figure-3.4.2.14

From March 2021, Binance beats Ethereum, Bitcoin, and XRP, while Ethereum outperforms Bitcoin and XRP.

**References: (word:211)**

[1] Stanford.edu. (2021). *CS 230 - Recurrent Neural Networks Cheatsheet*. [online] Available at: <https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-recurrent-neural-networks#architecture>

[2] Olah, C. (2015). *Understanding LSTM Networks -- colah’s blog*. [online] Github.io. Available at: <https://colah.github.io/posts/2015-08-Understanding-LSTMs/>.

[3] ‌Hochreiter, E., Rovelli, R. and Winckler, G. (1997). Centrální banky a využití ražebného? studie tří transformujících se ekonomik. *Politická ekonomie*, [online] 45(2), pp.193–206. doi:[10.18267/j.polek.274](http://www.bioinf.jku.at/publications/older/2604.pdf).

[4] ‌Stanford.edu. (2019). *The Sharpe Ratio*. [online] Available at: <https://web.stanford.edu/~wfsharpe/art/sr/sr.htm>.

[5] ‌Srivastava, P. and Mazhar, S.S., 2018. Comparative analysis of sharpe and sortino ratio with reference to top ten banking and finance sector mutual funds. *International journal of management studies*, *93*(10).

[6] Fernando, J. (2022). *Sharpe Ratio*. [online] Investopedia. Available at: <https://www.investopedia.com/terms/s/sharperatio.asp>

[7] ‌Investopedia. (2019). *Understanding the Sortino Ratio*. [online] Available at: <https://www.investopedia.com/terms/s/sortinoratio.asp>.

[8] ‌Platanakis, E. and Urquhart, A. (2019). Portfolio management with cryptocurrencies: The role of estimation risk. *Economics Letters*, [online] 177, pp.76–80. doi:10.1016/j.econlet.2019.01.019.‌

[9] Sung, O. (2021). The Kelly Criterion Applied to Long-Term Value Investing. *Junto*. [online] 17 Aug. Available at: <https://junto.investments/kelly-criterion/> [Accessed 21 Dec. 2022].

[10] Lopez de Prado, M. (2016). *Building Diversified Portfolios that Outperform Out-of-Sample*. [online] papers.ssrn.com. Available at: <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2708678>

‌

[11] Investopedia. (2019). Annualized Total Return. [online] Available at: <https://www.investopedia.com/terms/a/annualized-total-return.asp>.  In-text citation: (Investopedia, 2019)

[12] Chen, J. (n.d.). *Mean Return*. [online] Investopedia. Available at: https://www.investopedia.com/terms/m/meanreturn.asp. In-text citation: (Chen, n.d.)

‌

‌

<https://www.investopedia.com/terms/d/downside-deviation.asp>