Project : CRYPTO INSPECTION
DATA ANALYSIS

(JUNE - JULY 2021)

INTERNITY
'TOGETHER WE CAN'

INTERNITY FOUNDATION

Submitted By :

1. Aaryaman Thakur
2. Mohit Gola
3. Nancy Tayal
4. Taran Sonkar

TABLE OF CONTENT :

1. Motivation

(a) Introduction :

Crypto currencies, or virtual currencies, are digital means of exchange that uses cryptography for security. The word ‘crypto’ comes from the ancient greek word, ‘kryptós’, which means hidden or private. A digital currency that is created and used by private individuals or groups has multiple benefits.

Cryptocurrency is a new type of currency circulating in the market. With Bitcoin invented in 2009, now there emerges over 2,000 types of cryptocurrencies and has a total market value more than 110 billion. The price movement is exciting like riding with roller coaster, but people really do not know much about this virtual asset.

In this project, we are analyzing the historical crypto trading dataset, to portrait its dynamic landscape and dig into features of crypto currencies to analyze if any patterns exist in their price movement, to find which currency is more volatile, effect of one or more feature on others and much more.

The project is sub-divided following section. These are:

1. Loading necessary libraries.
2. Loading Dataset from a CSV file or from a Table.
3. Summarization of Data to understand Dataset (Descriptive Statistics).
4. Visualization of Data to understand Dataset (Plots, Graphs etc.).
5. Data pre-processing and Data transformation.
6. Applying different learning algorithms on the training dataset.
7. Evaluating the performance of the fitted model using evaluation metrics like confusion matrix, precision recall curves.

Crypto Inspection

(b) Research Questions:

1.) Can you predict the prices of these coins?

2.) Can you find significant patterns and trends in the prices of these coins?

3.) Compare different coins and their prices and behavior.

4.) Which currencies are more volatile and which ones are more stable?

5.) How does the price fluctuations of currencies correlate with each other?

6.) Seasonal trend in the price fluctuations.

7.) How do Bitcoin markets behave? What are the causes of the sudden spikes and dips in cryptocurrency values?

8.) Are the markets for different altcoins inseparably linked or largely Independent? How can we predict what will happen next?

9.) Predicting the future values of the currency using ARIMA.

10.) Predicting the alpha and beta values of the altcoins using CAPM model.

2. Dataset Description

(a)Source:

- From kaggle website, download dataset crypto-markets.csv, data Updated till July 7, 2021.
- URL: https://www.kaggle.com/sudalairajkumar/cryptocurrencypricehistory
- Files : 23 CSV files, Columns : 10 columns in each file

(b)Variables and their description:

The dataset has one CSV file for each currency. Price history is available on a daily basis from Jan 2019 – July 2021. This dataset has the historical price information of the top % crypto currencies by market capitalization.

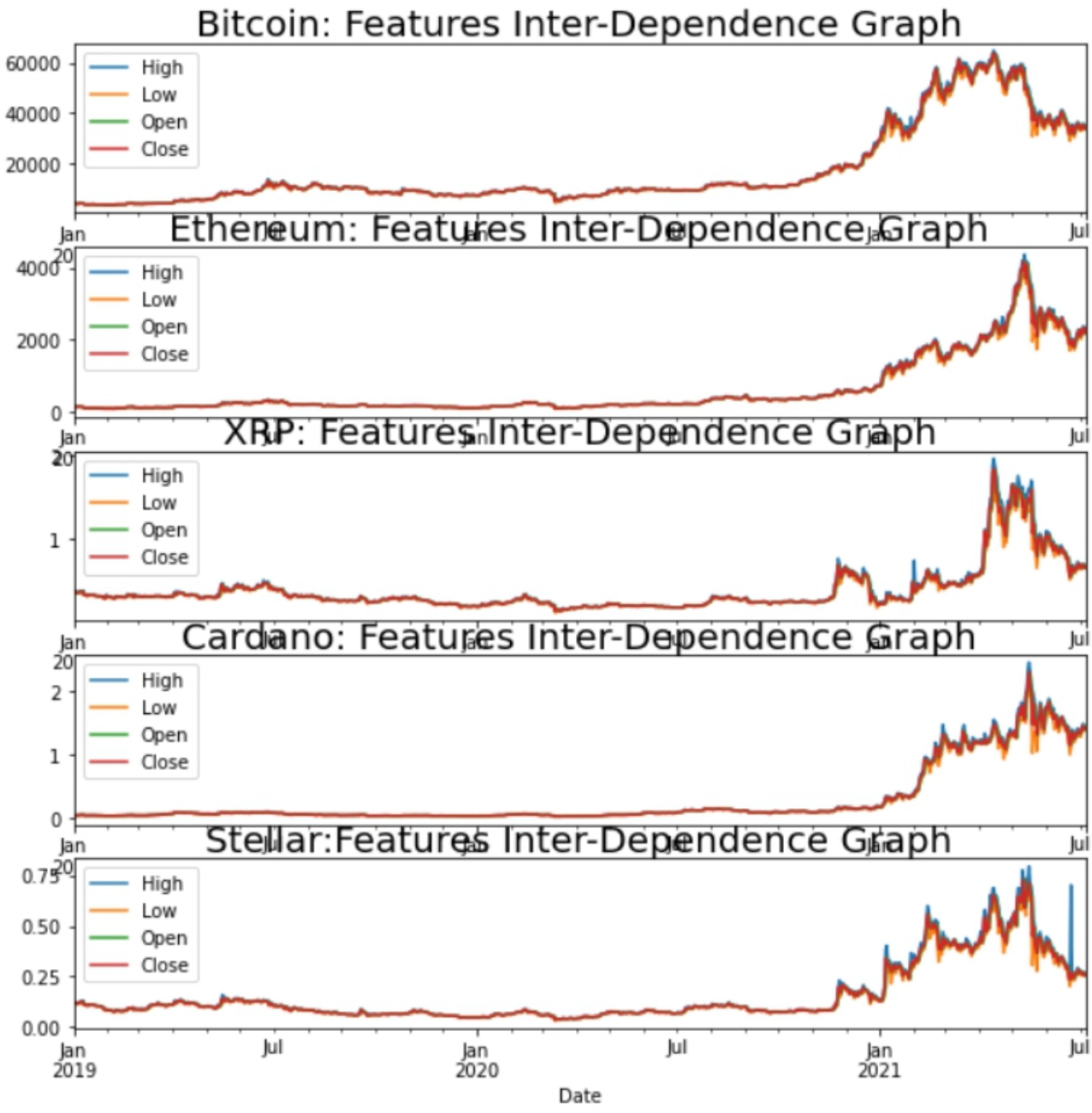
- Name : name of the currency
- Symbol : symbol of the currency
- Date : date of observation
- Open : Opening price on the given day
- High : Highest price on the given day
- Low : Lowest price on the given day
- Close : Closing price on the given day
- Volume : Volume of transactions on the given day
- Market Cap : Market capitalization in USD

(c) Data Manipulation and addition of new variables:

- From the downloaded data only the data of top 5 currencies from Jan 1, 2019 till July 7, 2021 was extracted in separate CSV files to be used in this project. From those files a combined CSV file containing the combined data of all the 5 currencies from Jan 1, 2019 till July 7, 2021 was prepared and the column SNo was dropped from all the files.
- The original variables are raw daily trading data for cryptocurrencies. To analyze its dynamic over time or compare one currency over another, we will be performed scaling of data in order to get accurate and precise results.
- Handling missing data: missing data, such as N/A, inf, are handled either by filling in the mean value or dropping the rows.

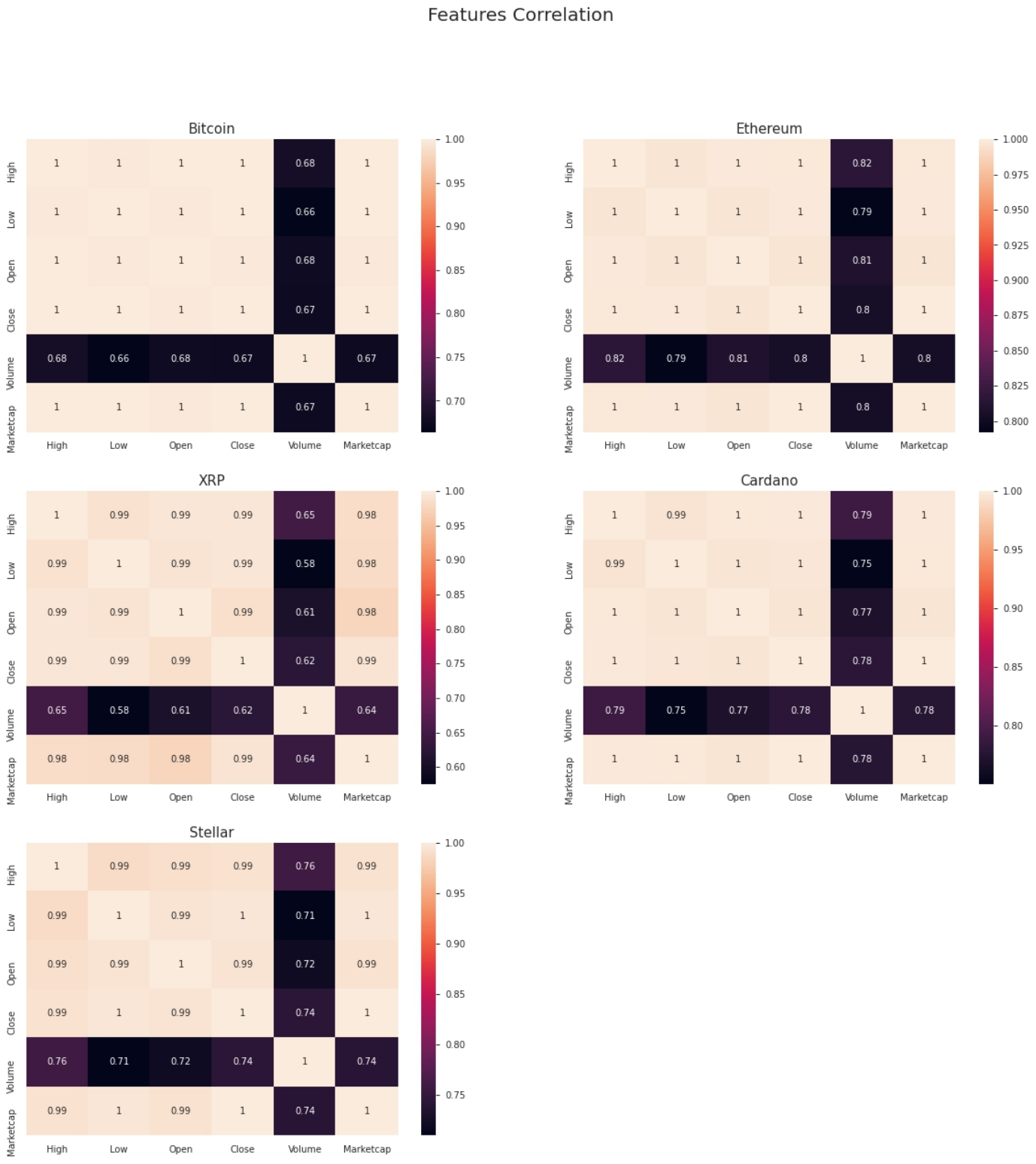
3. Methodology, Analysis and Results

Interdependence among features of each currency: To show the dependence among the features of the crypto currencies, we have plotted a line graph for each currency and we have found that the features are highly correlated. This is further assured by the heatmap we have plotted lately in the visualizations.



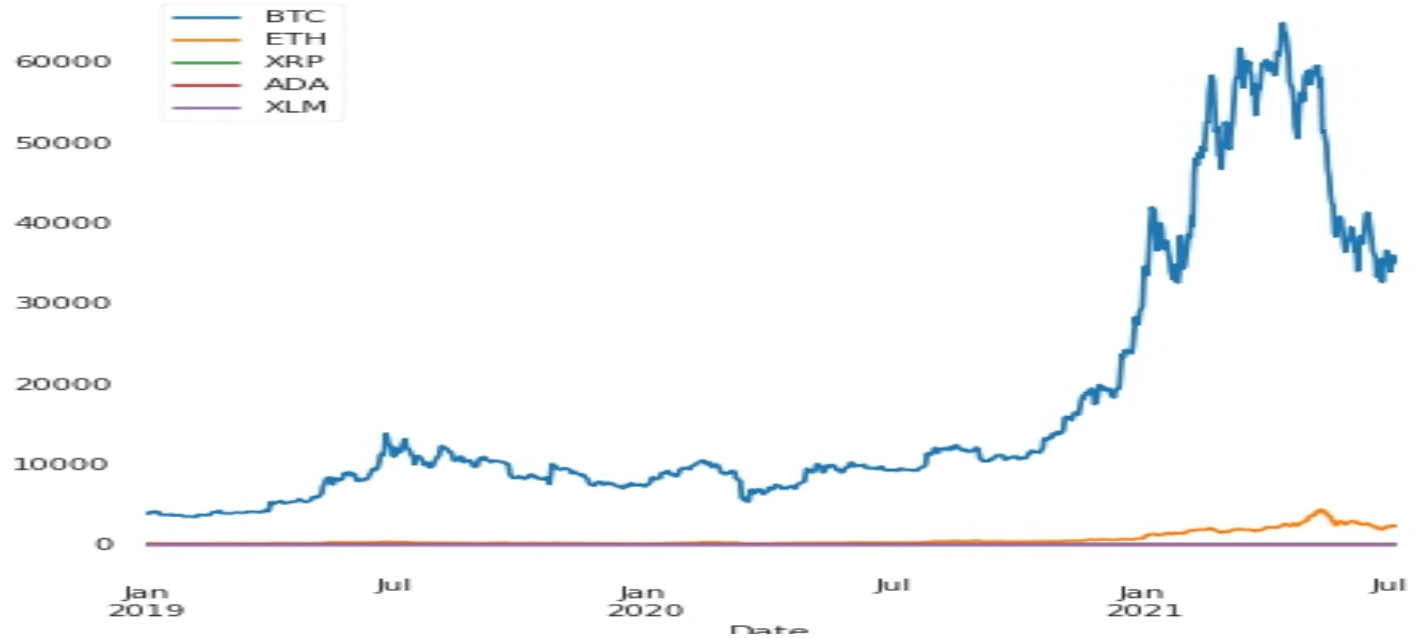
Crypto Inspection

Correlation Among the features of each currency : As show earlier the features of each currency are highly correlated and here is another proof in support of our prediction. The color of heatmap and the annotated values shows the extent of correlation.

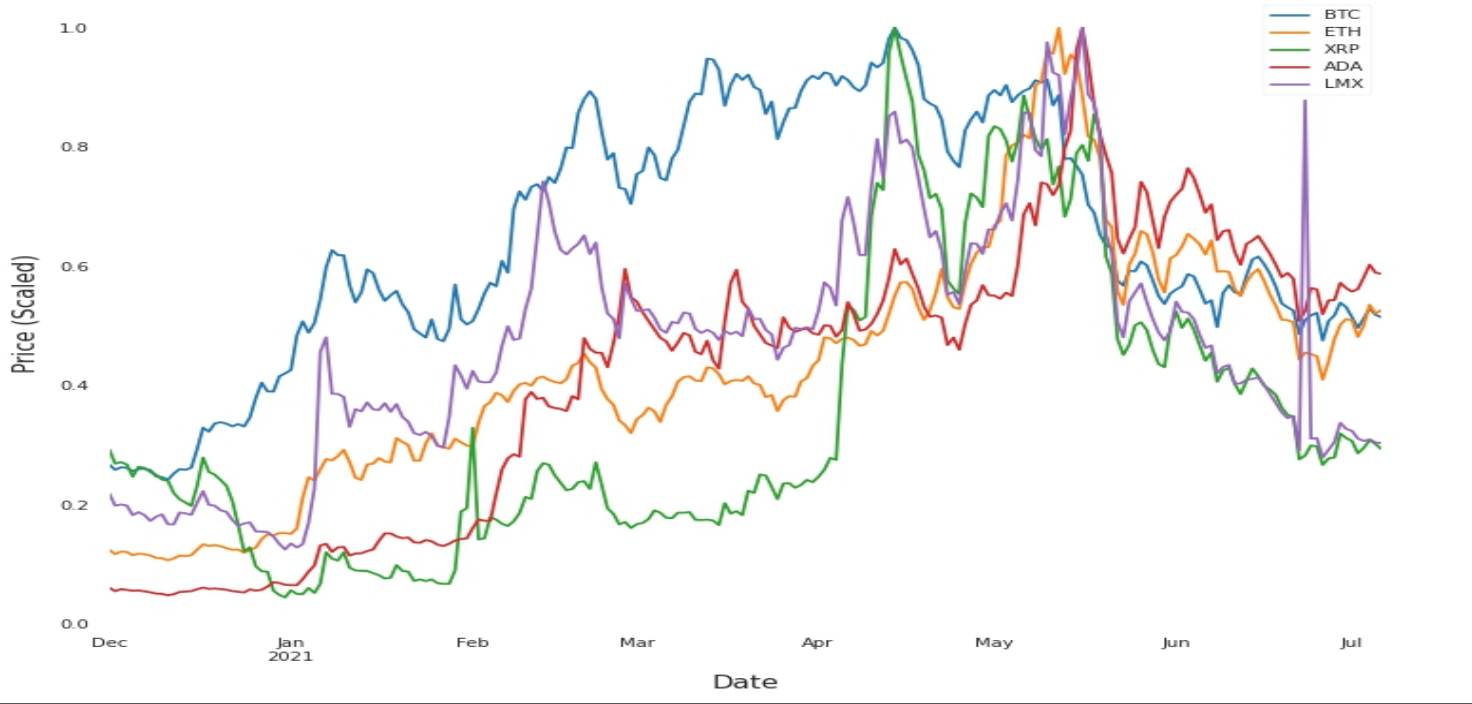


Crypto Inspection

Interdependence among the currencies: To predict what kind of relation occurs among the top 5 currencies, we have first tried to plot a line graph for the complete dataset for the 'high' value of currency, but since the currencies are having varying ranges so it was required to scale the data first (see the first plot, it is of the unscaled data). After scaling and plotting we have got that the currencies are having a nearly linear relation i.e. when the price of a currency goes up or down the same would happen with the price of other currencies in most cases. Also the overall performance of Stellar has improved a lot.

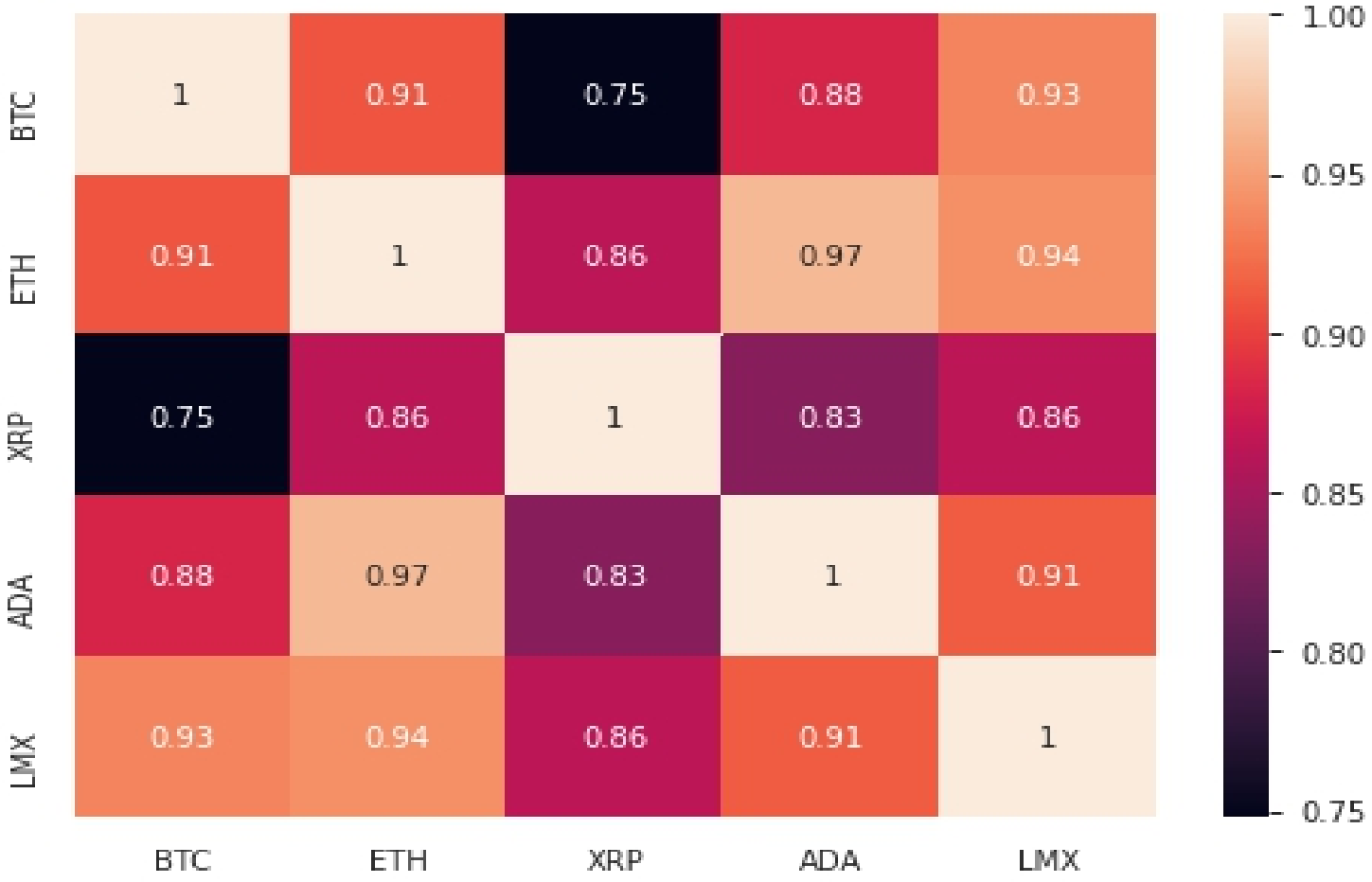


Graph to show the effect of spike or dip in price of a crypto on other crypto



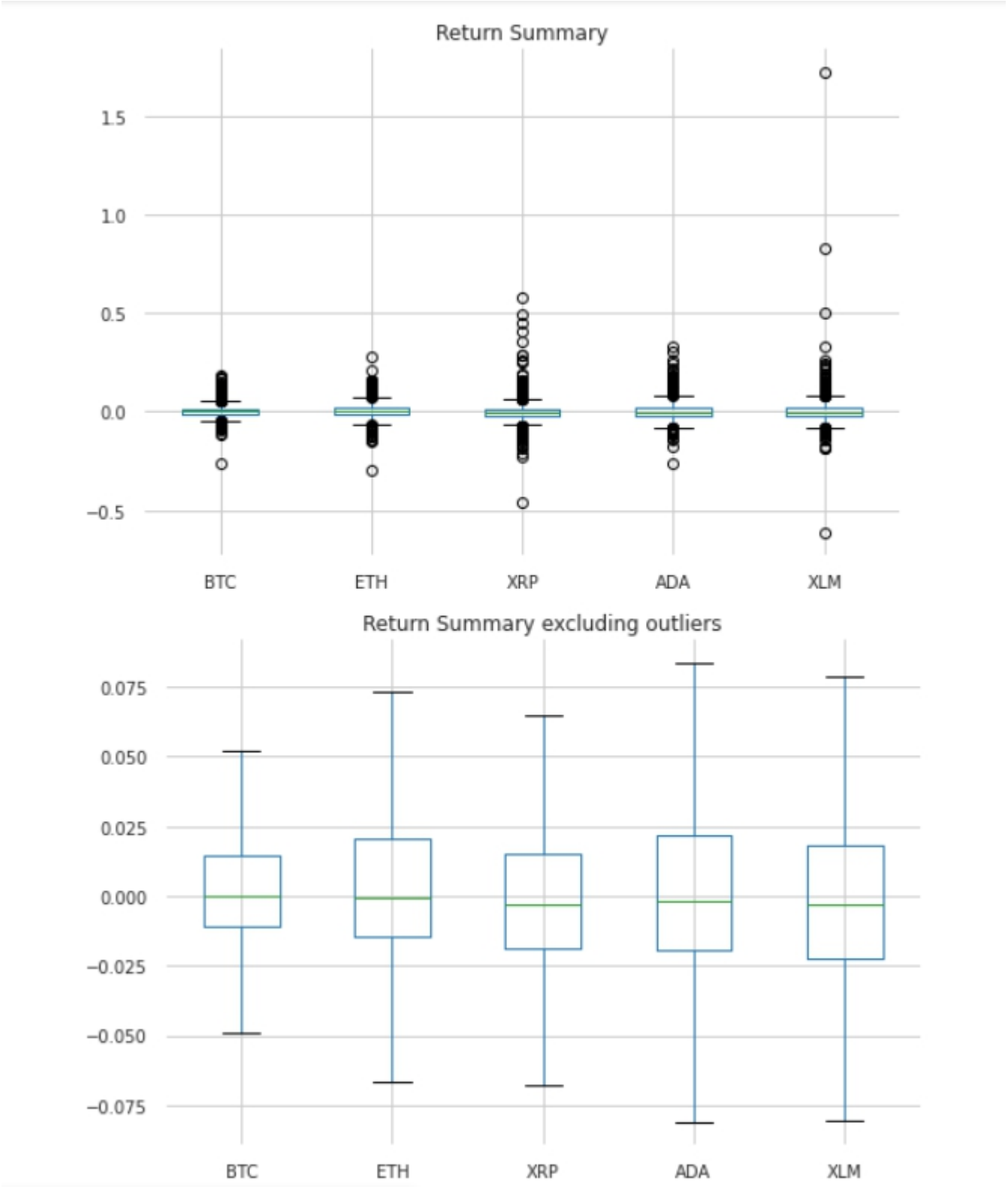
Crypto Inspection

Correlation Between The Cryptocurrencies : Here, we have shown extent of the correlation present between various currencies using a heatmap, the annotated values shows the measure of correlation.



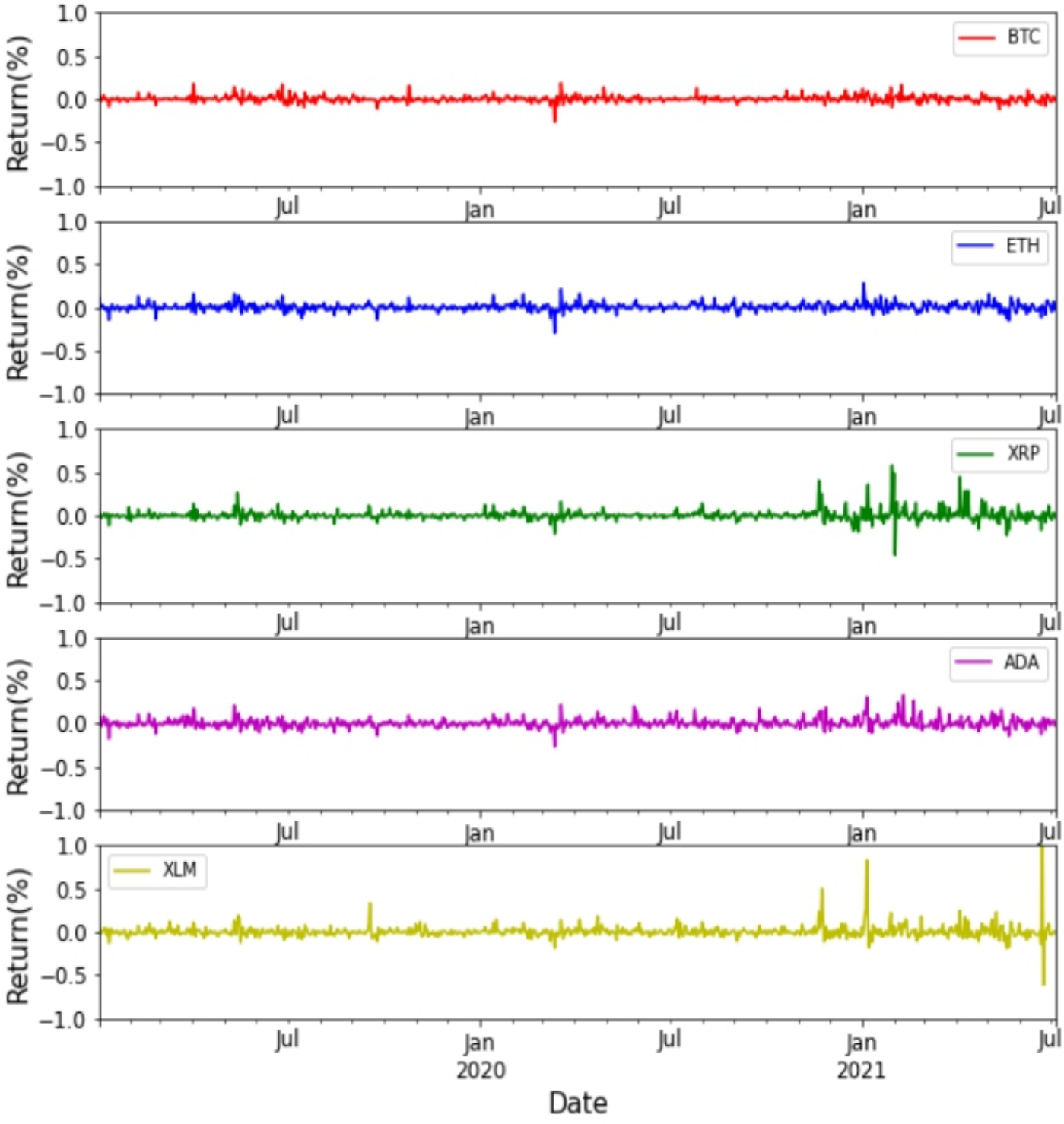
Crypto Inspection

Return Summary of each currency : We have first calculated the percentage daily return to find out which currency is more volatile and have shown the summary of the return percentages of each currency using a box plot.



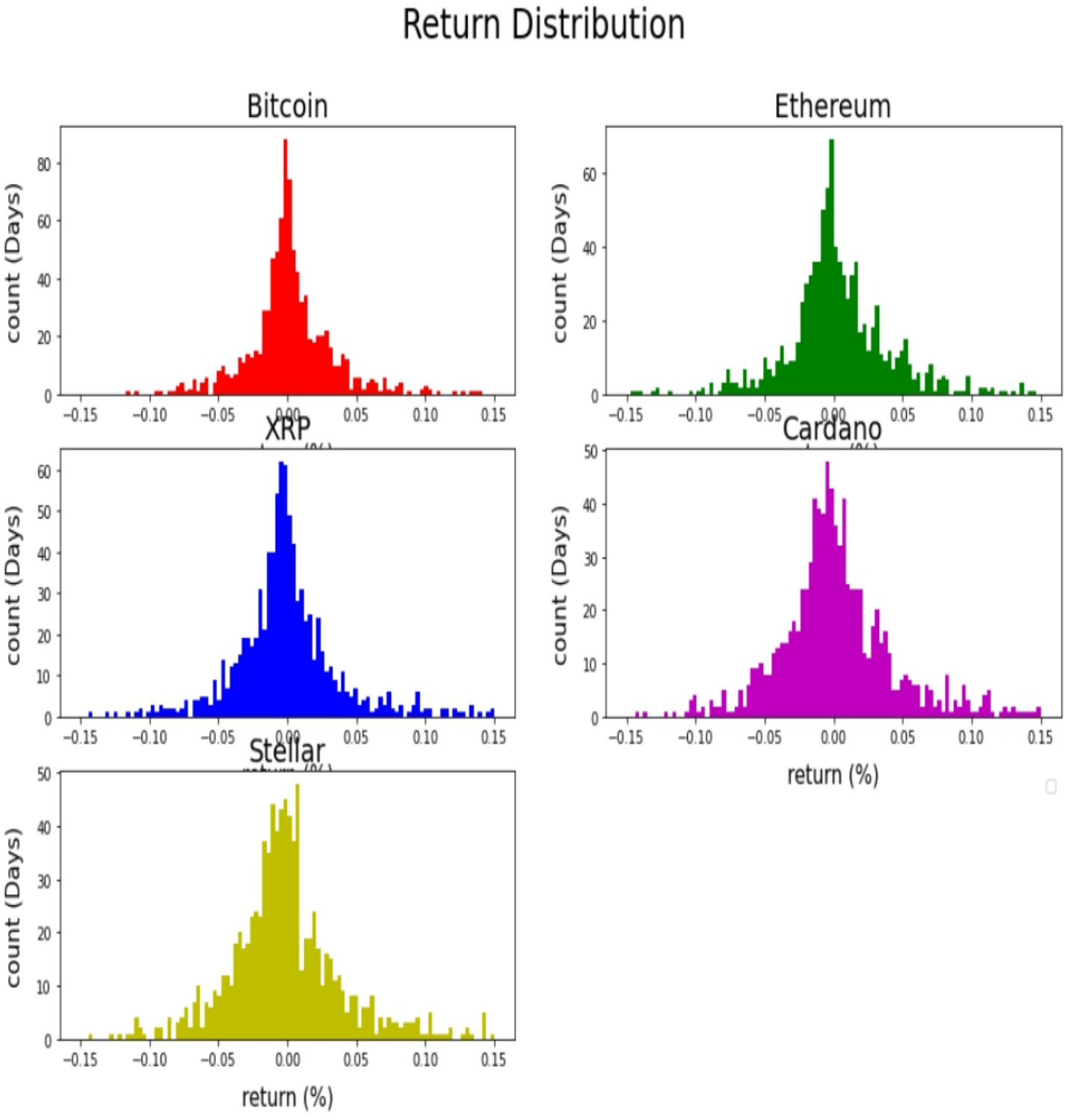
Crypto Inspection

Volatility of currencies : On the basis of the daily returns calculated earlier we have plotted line plots of each country and have found that all the currencies are volatile but Stellar has a high volatility as compared to other crypto. There is a sudden rise and dip in the plot for stellar(XLM).



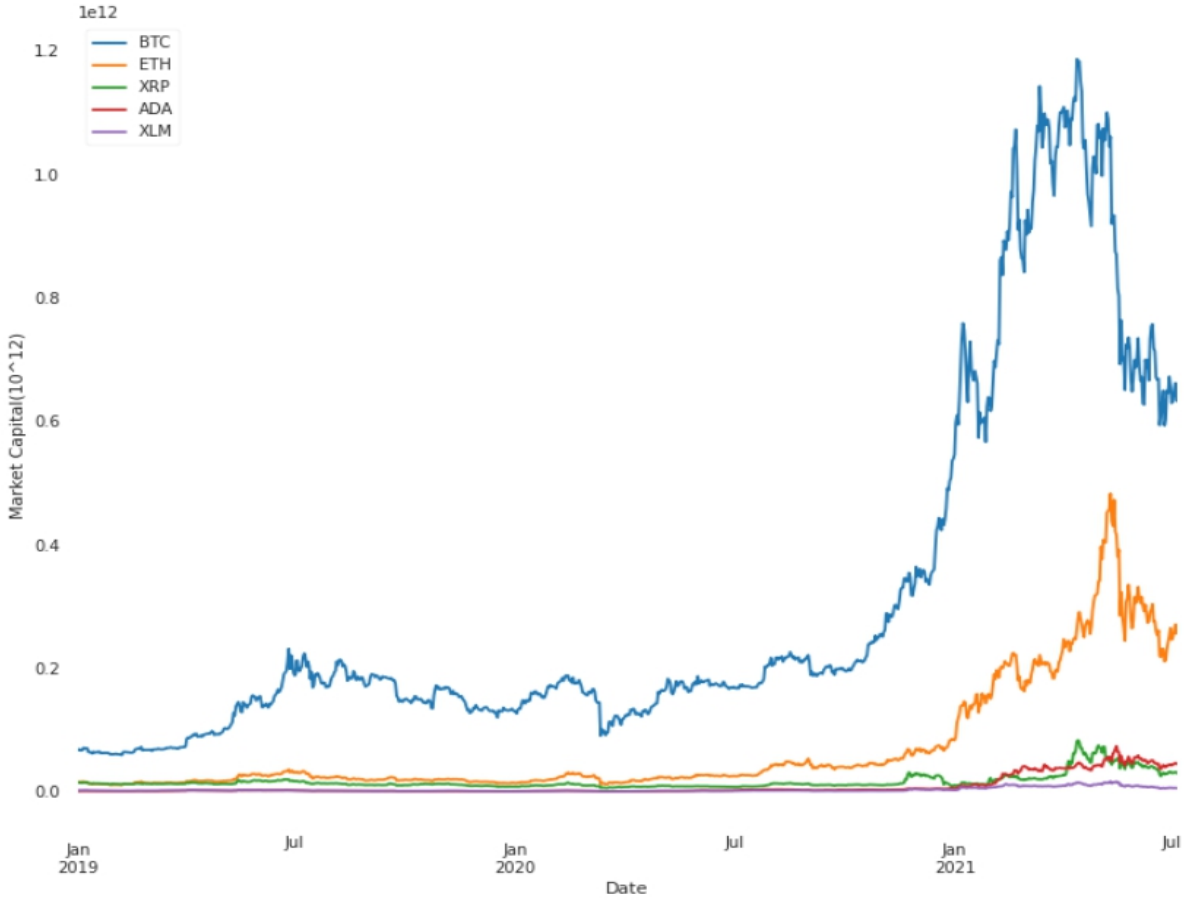
Crypto Inspection

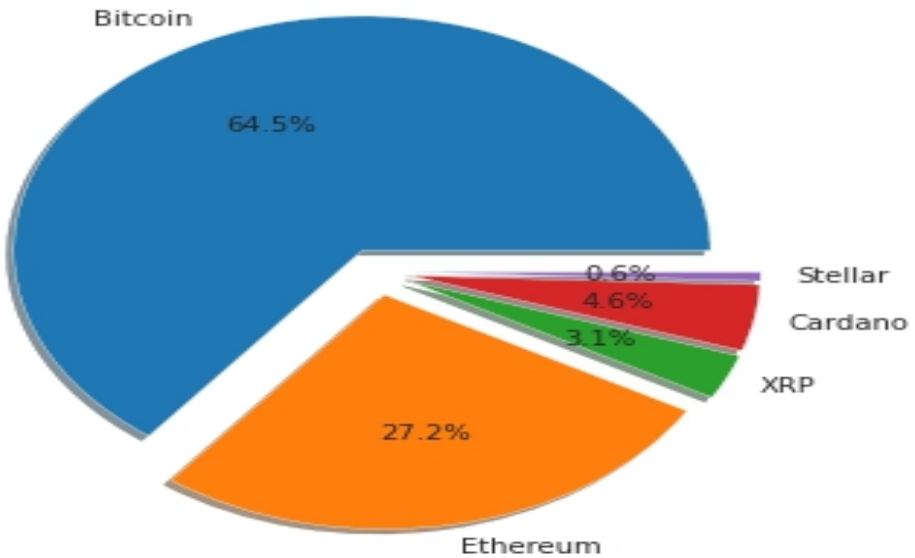
Return Distribution: We have used bar plots to depict the return distribution of each currency within the range (-0.15, 0.15) and for all the currencies the maximum return values are distributed in and near 0.00 and it can be said that the returns are normally distributed.



Crypto Inspection

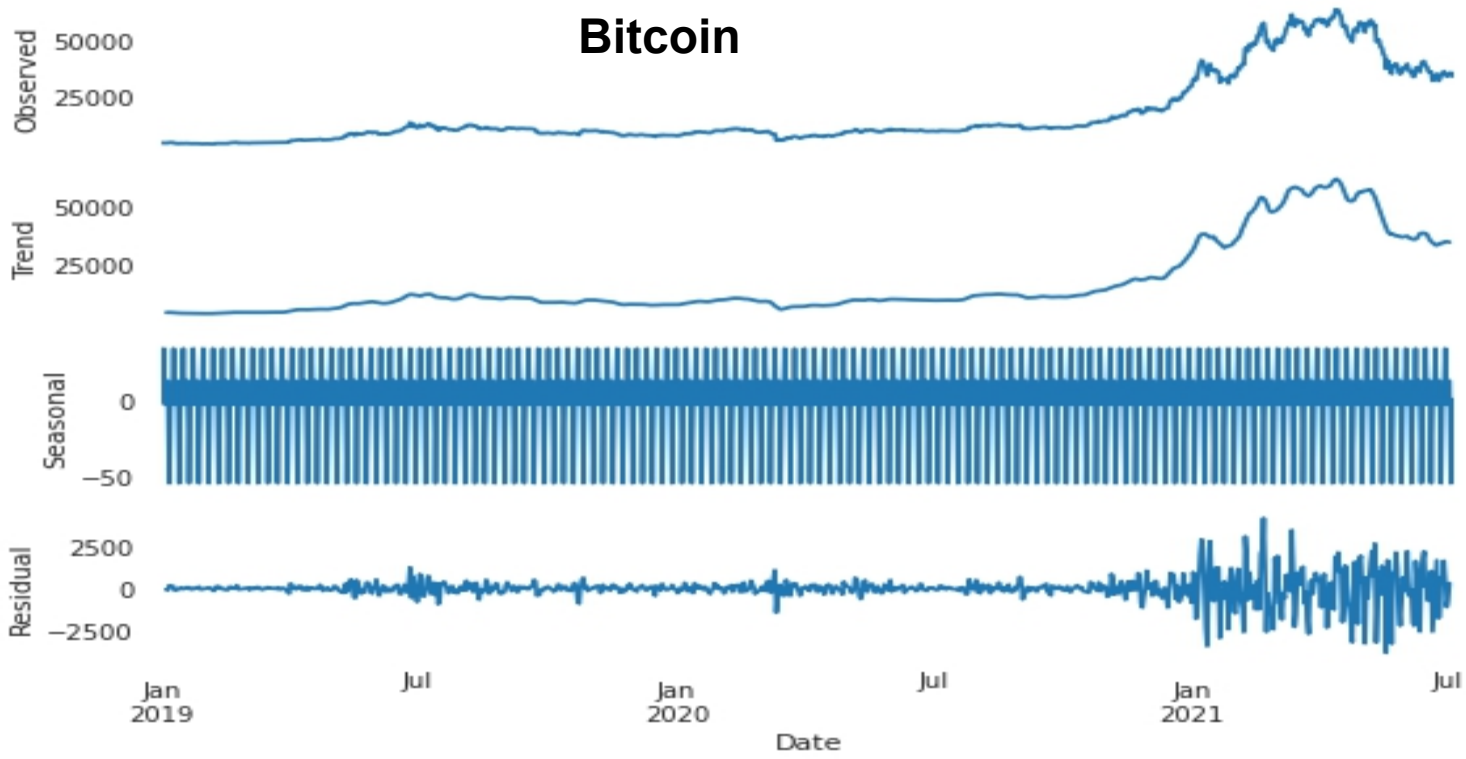
Market Acquired By Each Currency : To find out which currency is leading the market and the percentage share occupied in the market, we have used a line plot and a pie plot. We can clearly see that bitcoin
Is leading the market occupying 64.5% market capital. Also from the line plot it can be said that the overall performance of each currency has improved since the beginning of 2021 and there is a sudden demand for each of these currencies, thereby increasing their market capital.

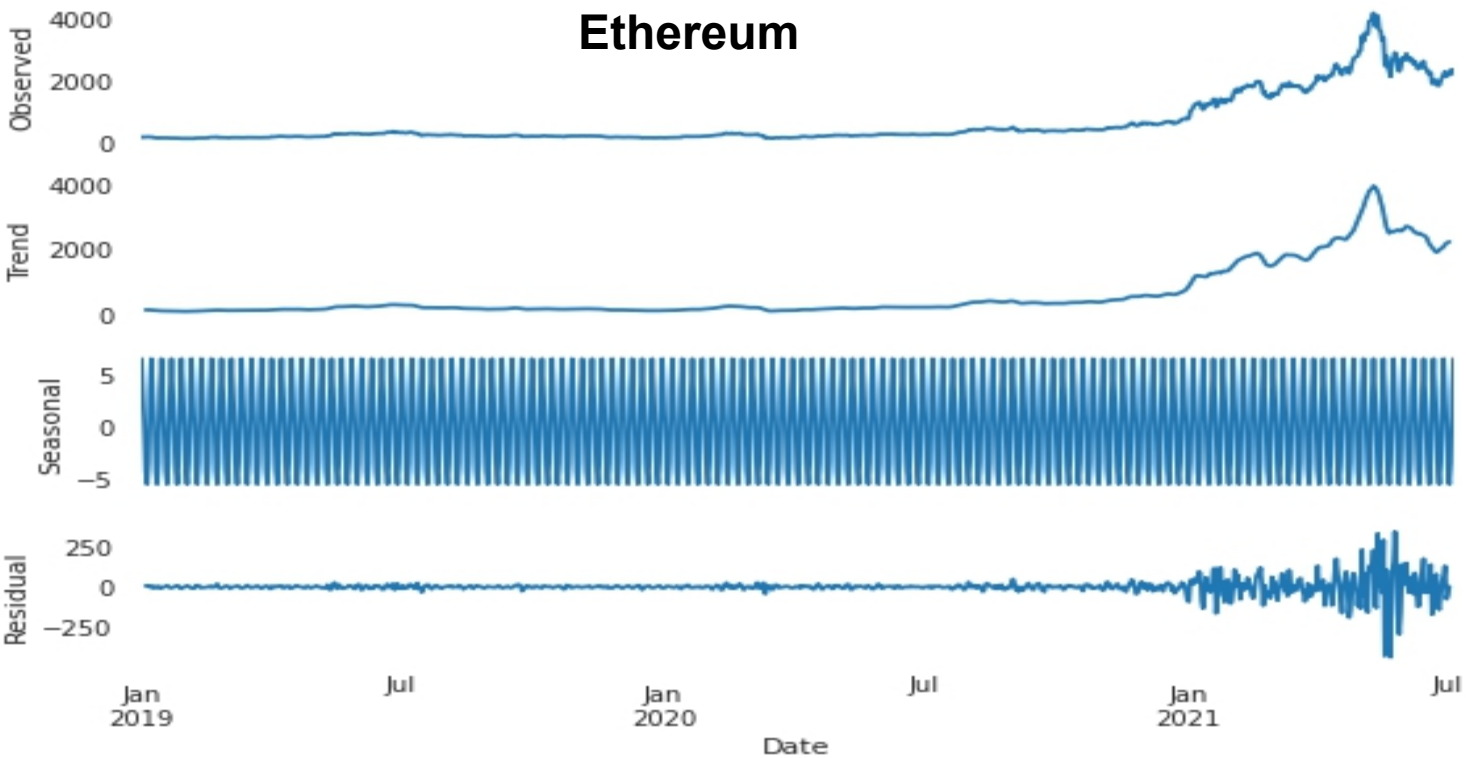




Crypto Inspection

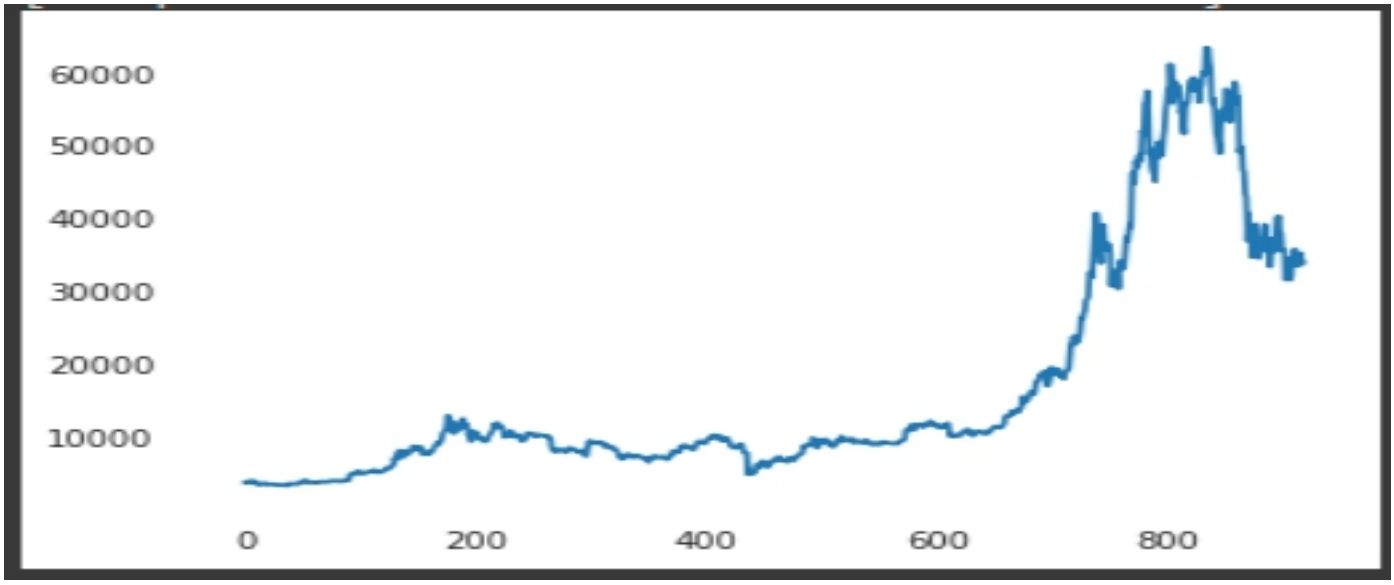
ETS Plot : Time series data can exhibit a variety of patterns, and it is often helpful to split a time series into several components, each representing an underlying pattern category. Using the seasonal_decompose() function of sklearn library we have decomposed the data for each currency to get Error, Trend and Seasonal patterns.

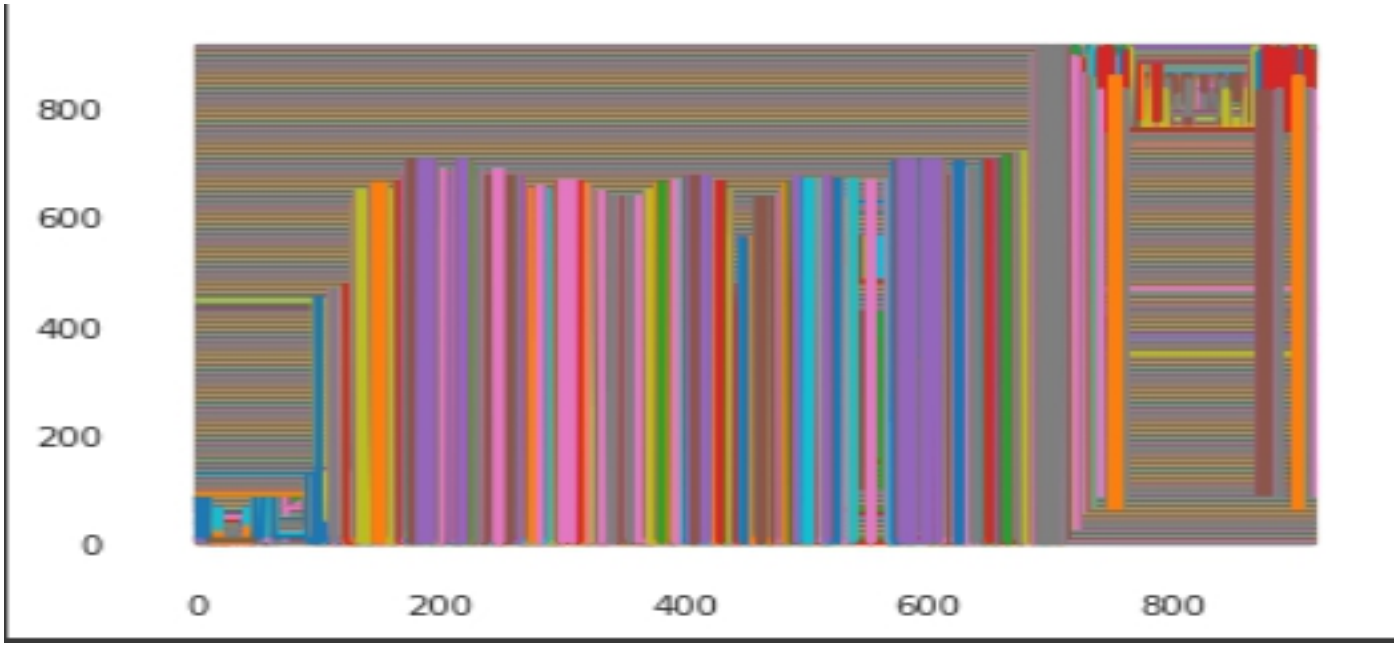




Crypto Inspection

KNN Model : The k-nearest neighbors (**KNN**) algorithm is a simple to - Find the nearest neighbors of each point. We will calculate the distance Each pair of point. KNN is one of the most basic yet essential classification algorithms in Machine Learning. We have found the distance of bitcoin closing in this, how near is every point. We have used simple line plot to return the nearest line.





Crypto Inspection

Linear Regression Model : Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task.

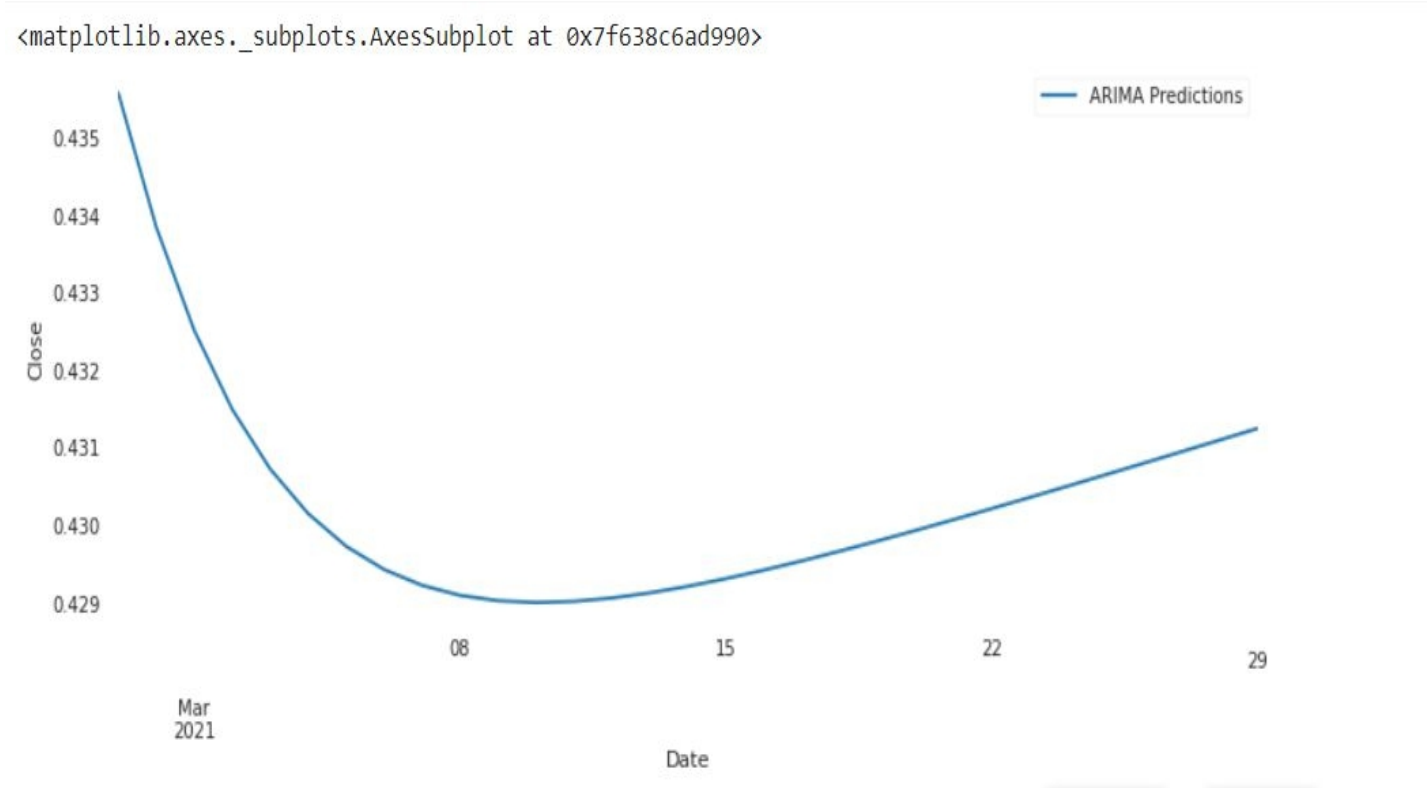
Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting. Different regression models differ based on – the kind of relationship between dependent and independent variables, they are considering and the number of independent variables being used. Linear regression performs the task to predict a dependent variable value (y) based on a given independent variable (x). So, this regression technique finds out a linear relationship between x (input) and y(output). Hence, the name is Linear Regression.

We have used linear regression algorithm in our project to forecast the values of a currency based on the previous values. Also we have used it to predict the values of a currency based on the values of another currency as we have seen earlier that the currencies are correlated and hence knowing the values for one currency we can predict the values for other currencies as well.

For example: We have used the bitcoin('High', 'Low', 'Open', 'Close') columns to predict the values of the ethereum('High') column. The R-squared value which we have got is around 0.8509 which is pretty good and hence our model is making right predictions.

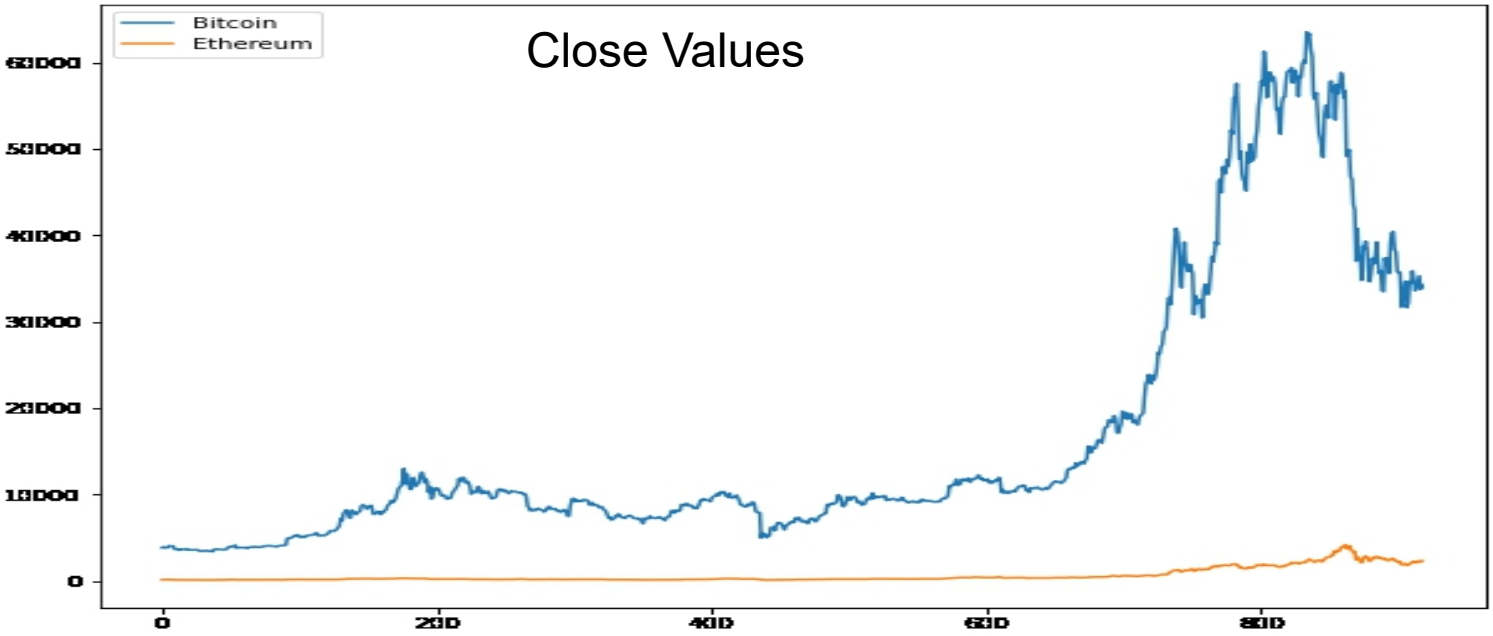
Crypto Inspection

ARIMA Model : An ARIMA model is a class of statistical models for analyzing and forecasting time series data.It explicitly caters to a suite of standard structures in time series data, and as such provides a simple yet powerful method for making skillful time series forecasts.
ARIMA is an acronym that stands for AutoRegressive Integrated Moving Average. It is a generalization of the simpler AutoRegressive Moving Average and adds the notion of integration.
This acronym is descriptive, capturing the key aspects of the model itself. Briefly, they are:
AR: Autoregression. A model that uses the dependent relationship between an observation and some number of lagged observations.
I: Integrated. The use of differencing of raw observations (e.g. subtracting an observation from an observation at the previous time step) in order to make the time series stationary.
MA: Moving Average. A model that uses the dependency between an observation and a residual error from a moving average model applied to lagged observations.
In our project we used to predict the close value for next 30 days with the root mean squared error of 0.12 which means our project is making right predictions.

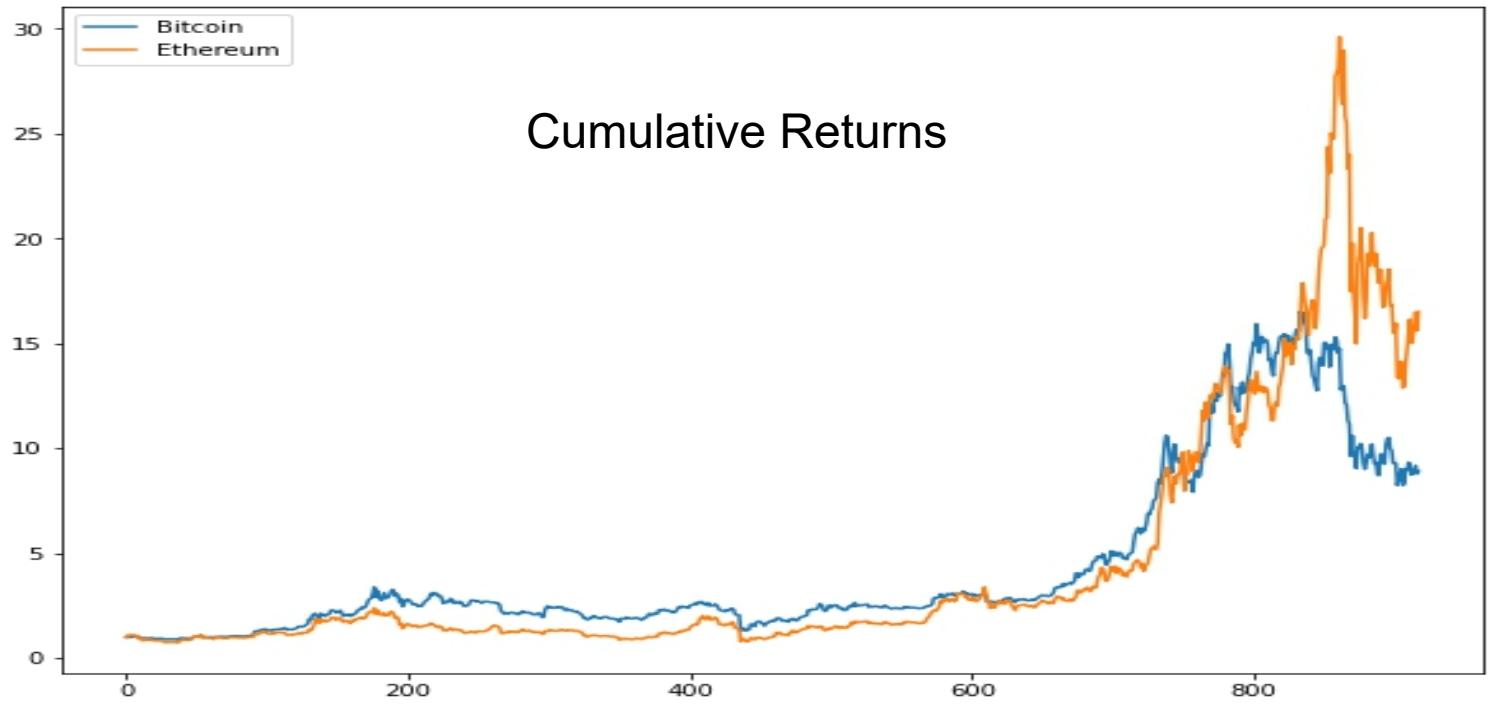


Crypto Inspection

CAPM Model : The capital asset pricing model (CAPM) is very widely used and is considered to be a very fundamental concept in investing. It determines the link between the risk and expected return of assets, in particular stocks.



As from the above plot, it seems like the crypto's performance is mimicking the market performance.So statistically they can be compared and cumulative returns are found.



4. Modules :

1. **Pandas :** Pandas allows importing data from various file formats such as comma-separated values, JSON, SQL, Microsoft Excel. Pandas allows various data manipulation operations such as merging, reshaping, selecting, as well as data cleaning, and data wrangling features.

2. **NumPy :** NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more**.**

3. **Matplotlib :** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK**.**

4. **Linear Regression :** Regression models describe the relationship between variables by fitting a line to the observed data. Linear regression models use a straight line, while logistic and nonlinear regression models use a curved line. Regression allows you to estimate how a dependent variable changes as the independent variable(s) change.

5. **KNN Model :** The k-nearest neighbors (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems.It is one of the simplest machine learning algorithms and is used in a wide array of institutions. KNN is a non-parametric, lazy learning algorithm

6. **Seaborn :** Seaborn is a library for making statistical graphics in Python. It builds on top of matplotlib and integrates closely with pandas data structures. Seaborn helps you explore and understand your data. Its plotting functions.
   Operate on dataframes and arrays containing whole datasets and internally perform the necessary semantic mapping and statistical aggregation to produce informative plots. Its dataset-oriented, declarative API lets you focus on what the different elements of your plots mean, rather than on the details of how to draw them.

7. **Scikit-Learn:** Scikit-learn (Sklearn) is the most useful and robust library for machine learning in Python. It provides a selection of efficient tools for machine learning and statistical modeling including classification, regression, clustering and dimensionality reduction via a consistence interface in Python. This library, which is largely written in Python, is built upon **NumPy, SciPy** and **Matplotlib**.

5. LIMITATIONS :

1. Certain limitations of our project are that one should know the dataset and all the components in full details as it is the major part.

2. Secondly one should have the knowledge of python and various models of machine learning to apply them.

3. The knowledge of working of these models and the output or results given by these models should be known.

4. The scope of the project is limited to the models and the relative dataset on which these models are applied.

5. Moreover we are not providing any user interface so you have to search and load your own dataset.

6. Summary and Conclusions :

1. First cryptocurrency was launced in 2009 by Nakamoto since then so many cryptocurrencies are launced .In which most popular cryptocurrencies are Bitcoin, ripple, Ethereum ,Cardano and Stellar. They aquire the major portion of the crypto market.

2. We know that trading in cryptocurrencies requires high skills and practice so we made a crypto Inspector. that will analyse the behaviour of these crytocurrencies on the basis of their past behaviour and help us to analyse their fluctuations in market.

3. With the help of this predictor we will be able to find whole analysis of any of these currencies from past few years where we will be able to find any Minimum / Maximum Crypto value of any currency, and we can also compare the growth of a crypto currency with any other and we can see the impact of growth of any currency on other one .

4. We use some Python modules like NumPy , Pandas and Matplotlib to perform some statical and algebric routines , manipulate and analyse data using pandas DataFrame , and using object-oriented API for plot embedding etc. respectively .

5. By using different algorithms like Linear Regression algorithm we studied the influence of one crypto currency on other or how they are related to each other and we also used KNN algorithm that will give new data points accordingly to the k number or the closest data points.

6. By using Autoregressive integrated moving average (ARIMA) models predicts future values based on past values and visualize the Time Series Data.

7. The Capital Asset Pricing Model (CAMP) desvcribes the relationship between systematic risk and expected return for assets, particularly stocks and given the alpha beta value.

7. References:

1. https://www.geeksforgeeks.org/k-nearest-neighbor-algorithm-in-python/
2. https://www.youtube.com/watch?v=8FCDpFhd1zk&t=1s&ab_channel=NachiketaHebbar
3. https://www.youtube.com/watch?v=sGQfiyXOvF0&t=427s&ab_channel=CodeWithHarry
4. https://towardsai.net/p/latest/capital-assets-pricing-model-capm%E2%80%8A-%E2%80%8Ausing-python
5. https://www.geeksforgeeks.org/k-nearest-neighbors-with-python-ml/
6. https://www.geeksforgeeks.org/ml-sklearn-linear_model-linearregression-in-python/
7. https://medium.com/@yrnigam/how-to-write-a-data-science-report-181bd49d8f4d

8. Appendix :

1. Google colab
2. Dataset: crypto-markets.csv (kaggle)