Project Report

GitHub URL

https://github.com/shivankgarg/UCDPA_ShivankGarg

Abstract

This project has been done as a part of project submission for UCD Specialist Certificate in Data Analytics. The objective of this project is to predict price of cryptocurrency Ripple (XRP) based on the prices of 4 cryptocurrencies (Bitcoin (BTC), Ethereum (ETH), Binance Coin (BNB) and Cardano (ADA)). For the project, we use averaged price action or an average of OHLC (open, high, low, close) values) on an hourly interval and use machine learning models namely Multiple Linear Regression, Random Forest Regression and AdaBoost Regression to determine the price of XRP. The best performing model with a high R^2 value of 0.88 on the test data was the AdaBooster Regression model.

Introduction

Stocks and Cryptocurrencies always seems lucrative for me as they are good source of secondary income. However, given the variability in cryptocurrency prices within a day (they can move as much as 10% in a day) and Unlike trading stocks and commodities, the cryptocurrency market is open 24/7. I wanted to explore the idea of a data-driven trading strategy, where I would base the price of a cryptocurrency on the prices of 4 other (larger) cryptocurrencies.

I wanted to base the prediction model on the prices of other currencies and no other factors such as DateTime or trading volume as the price of one cryptocurrency strongly affects the price of the other.

Dataset

Source - https://www.cryptodatadownload.com/data/binance

The datasets for different cryptocurrencies were downloaded from the website. It contains historical data from different exchanges across the world and I choose the prices from the BINANCE exchange which has the largest daily trading volume in the world.
To have large sample space I had choose hourly data. I chose this source since the data was free and easy to access (no sign-up required) and reliable (as per reviews).

Dataset downloaded from website:

1. Binance_BTCUSDT_1h.csv
2. Binance_ADAUSDT_1h.csv
3. Binance_XRPUSDT_1h.csv
4. Binance_ETHUSDT_1h.csv
5. Binance_BNBUSDT_1h.csv

And each dataset contains the following columns:

- Unix Timestamp - This is the unix timestamp or also known as "Epoch Time". Use this to convert to your local timezone
- Date - This timestamp is in UTC datetime
- Symbol - The symbol for which the timeseries data refers
- Open - This is the opening price of the time period
- High - This is the highest price of the time period
- Low - This is the lowest price of the time period
- Close - This is the closing price of the time period
- Volume (Crypto) - This is the volume in the transacted Ccy. Ie. For BTC/USDT, this is in BTC amount
- Volume Base Ccy - This is the volume in the base/converted ccy. Ie. For BTC/USDT, this is in USDT amount
- Trade Count - This is the unique number of trades for the given time period

The aim of the cleaning process would be to get 2 columns each from the 5 datasets, the timestamp and an averaged value of the ‘Open’, ‘High’, ‘Low’ and ‘Close’ values for each cryptocurrency and then merge them along the timestamp.
We would be plotting open-high-low-close chart which can be used to illustrate movements in the price of different cryptocurrency over the time.

Implementation Process

Five Major Task were carried out in this project as below:

1. **Data Importing**
   - Fetching the Data from the API.
   - Importing the Data from CSV file into Dataframe.

2. **Data cleaning and merging**
   - Using .head() and .info() method, to know more about dataset imported.
   - Remove multi-index: Datasets came through as a multi-index and only one column where the column name was the data source (https://www.CryptoDataDownload.com).
   - Function remove_columns(): Custom function were created to reset the multi-index, rename columns using a dictionary.
   - Function data_check(): Check and remove the missing/duplicate values from dataset Using methods like isnull(), duplicated(), sum(), drop_duplicates().
   - Get average of open, high, low, close column: Combine open, high, low, close into one column OHLC that can be used for analysis using get_OHLC() function. All columns are objects, ‘date’ will be converted into datetime object while ‘open’, ‘high’, ‘low’, and ‘close’ columns will be converted into floats.
   - Dropping null values: During conversion, 25,936 ‘date’ values did not convert to datetime object in dataFrame_btc (BTC) and dataframe_eth (ETH) datasets. These returned ‘NaT’ null value. The remaining values (17720) also happen to be the exact number of rows that are found in the dataFrame_bnb (BNB), dataFrame_ada (ADA) and dataFrame_xrp (XRP) datasets. This is likely due to the datasource changing the datatime format at the moment. Since all datasets would require equal values to merge and any NaT values would drop anyway, we drop these values using function dropna().
   - Merging datasets: Merge all 5 datasets into a single dataset namely df

3. **Exploratory Data Analysis**

- Plot datasets using matplotlib library: Plot the 5 datasets having datetime on the x-axis and average asset price (OHCL) on the y-axis.
- Comparative plot: Ploting all 5 datasets together with datetime on the x-axis and average asset price(OHCL) on the y-axis.
- Correlation matrix: Develop a correlation matrix for the 5 assets.

4. Data preparation and model training

- Creating feature and target variables: ‘df’ columns of 'OHLC_ada','OHLC_bnb','OHLC_btc','OHLC_eth' were added as feature variables ‘X’ and the target variable of ‘OHLC_xrp’ was assigned the target variable ‘y’.
- The feature and target variable ‘X’ and ‘y’ are split using a 70:30 ratio train-test split.
- Multiple Linear Regression: The multiple regression model is called and fit to the training data, before being used to predict the test ‘X’ dataset. Then, regression metrics are calculated to evaluate the model.
- Random Forest Regression: The Random Forest regression model is called and fit to the training data, before being used to predict the test ‘X’ dataset. Then, regression metrics are calculated to evaluate the model accuracy.
- AdaBoost Regression: The AdaBoost regression model is called and fit to the training data, before being used to predict the test ‘X’ dataset. Then, regression metrics are calculated to evaluate the model accuracy.

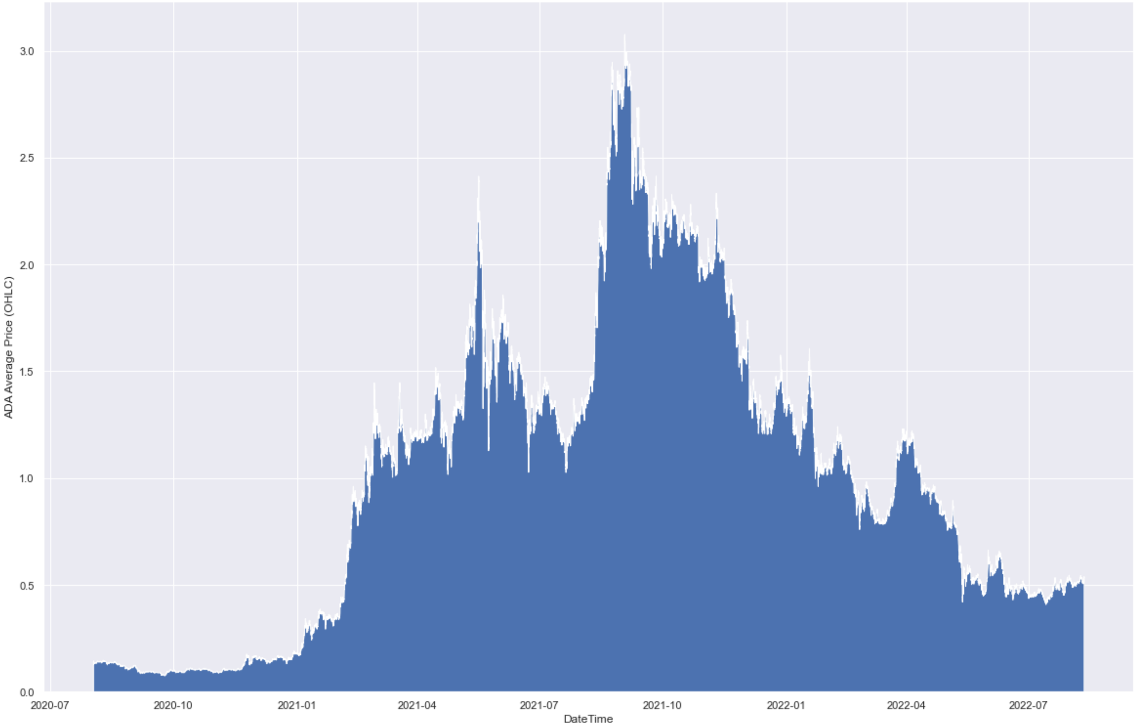
5. Developing Insights

- Correlation Matrix: Pearson correlation method was called on the dataframe to get an understanding of the correlation between the 5 cryptocurrencies and heatmap was plotted for the same.
- Multiple Regression Plot: The true values of the test data were plotted on the x-axis with the predicted values on the y-axis on a scatter plot to look at the differences between the two values. The regression metrics are added below the plot to have them available to develop insights over the model.
- Random Forest Regression Plot: The true and predicted values of the test data were charted on the scatter plot to look at the differences between the two values. Regression metrics were added as above.
- AdaBooster Regression Plot: The true and predicted values of the test data were charted to the scatter plot to look at the differences between the two values. Regression metrics were added as above.

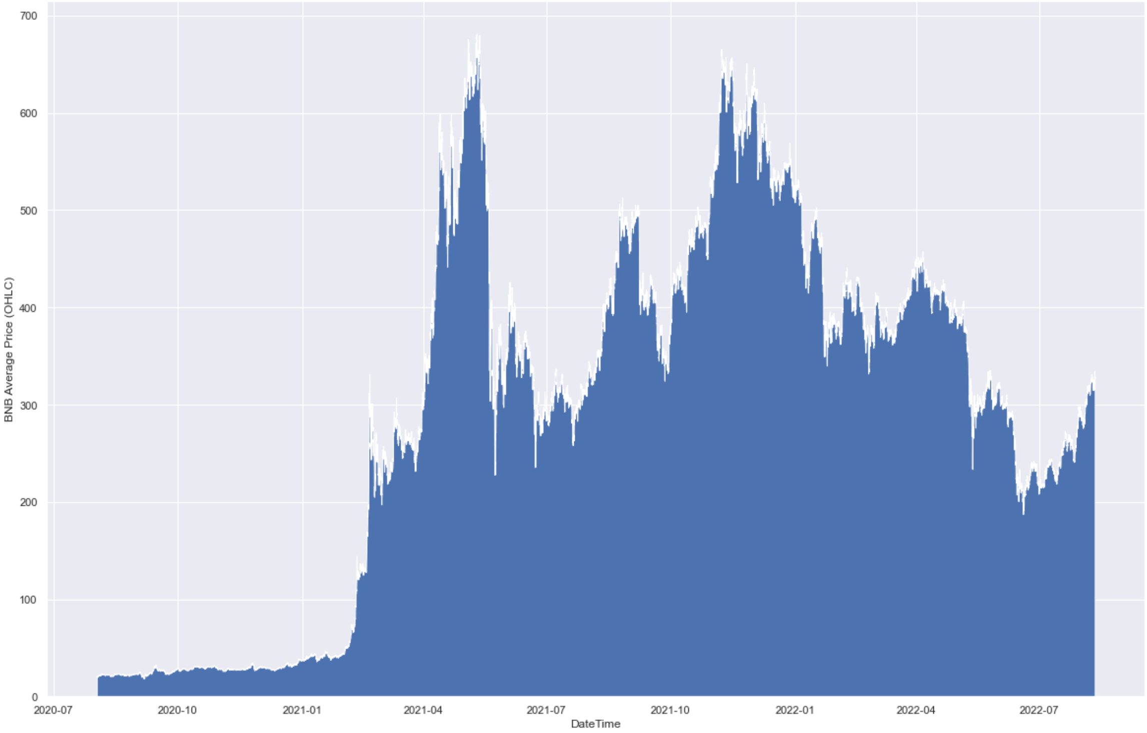
To conclude, data was loaded onto the Jupyter notebook, it was then cleaned to get one dataframe that contained the timestamp and the averaged OHLC values of the 5 assets before fitting 3 different machine learning models and then measuring them on various metrics. Finally, insights were developed on the correlation

Results

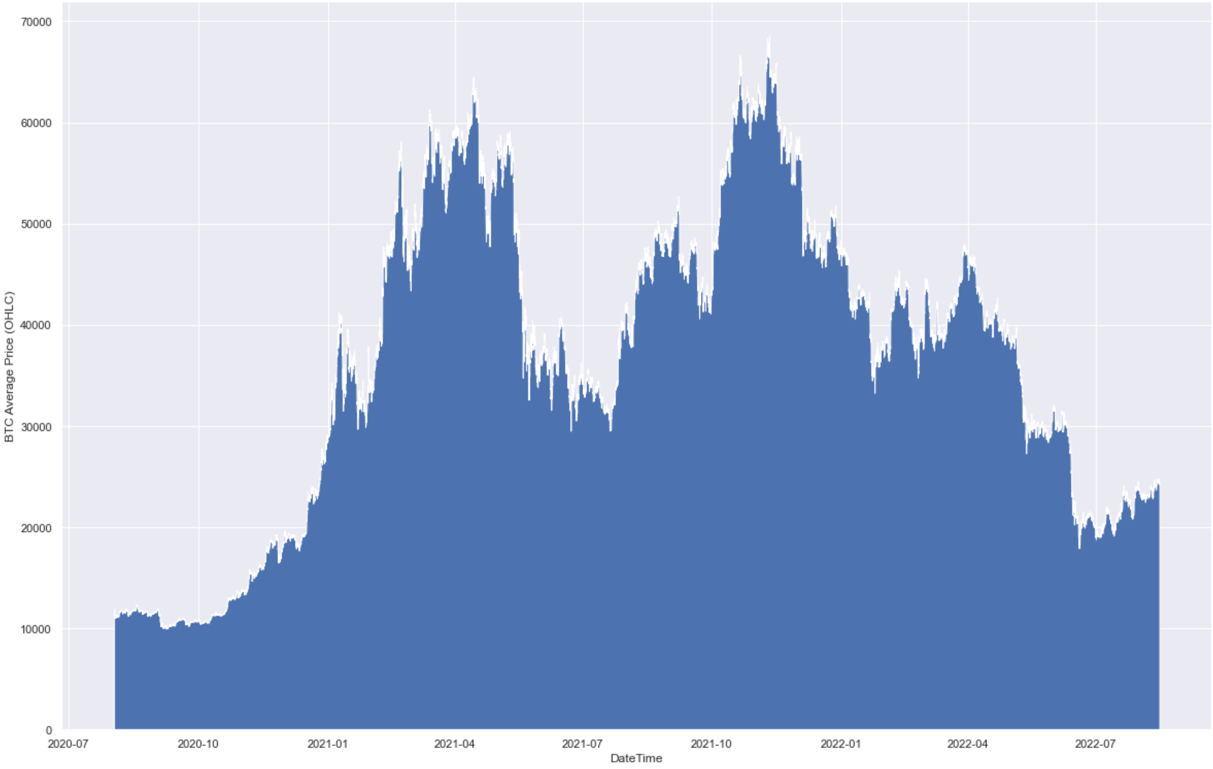
ADA Chart



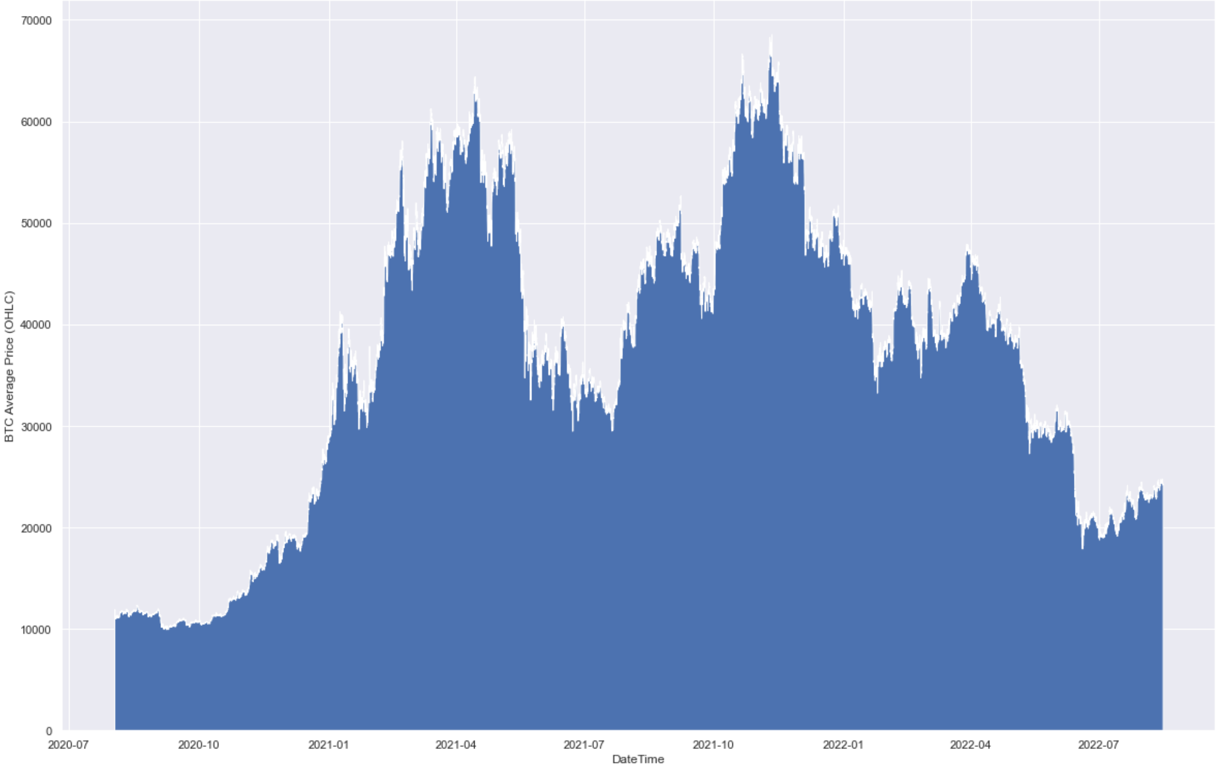
BNB Chart



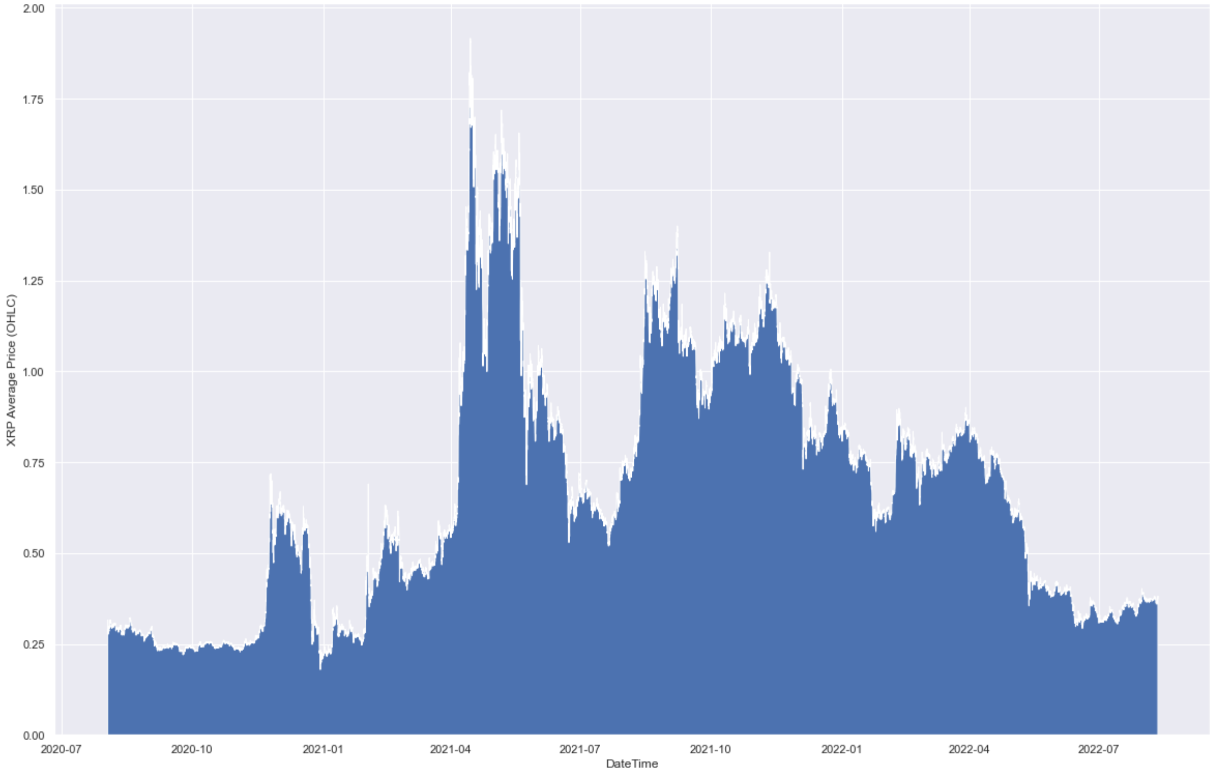
BTC Chart



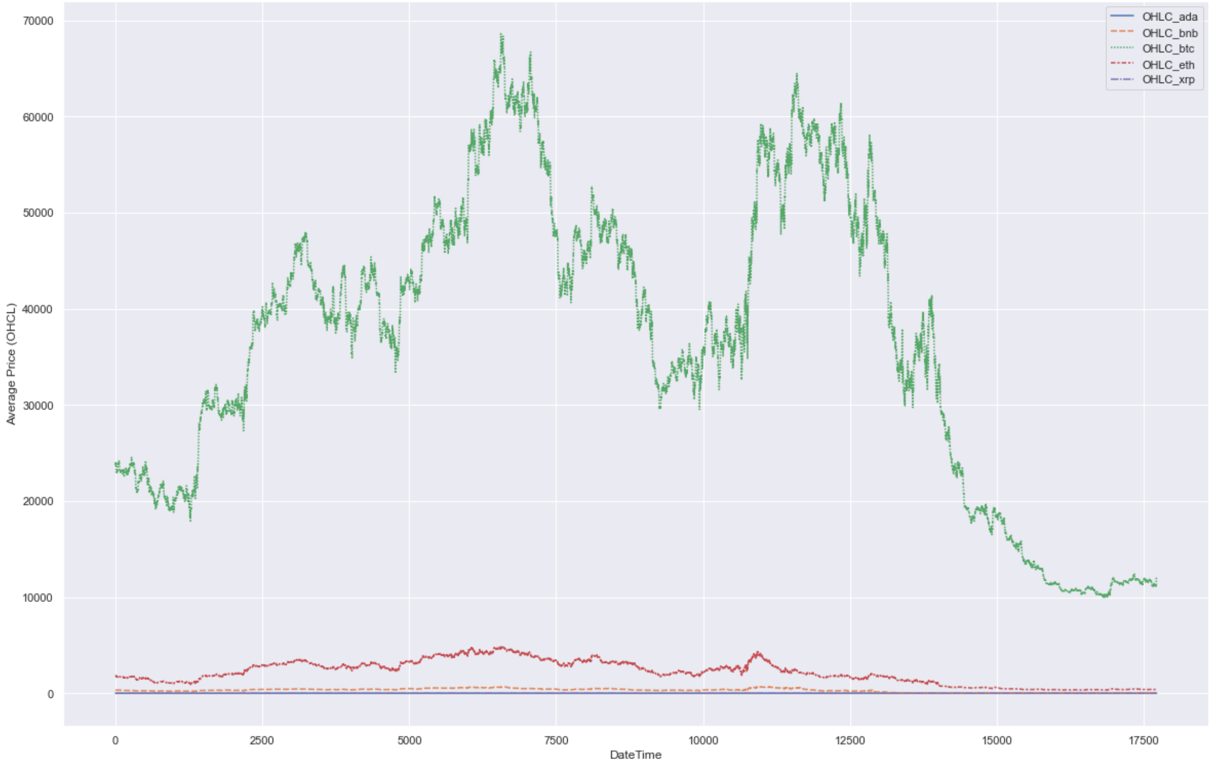
ETH Chart



XRP Chart

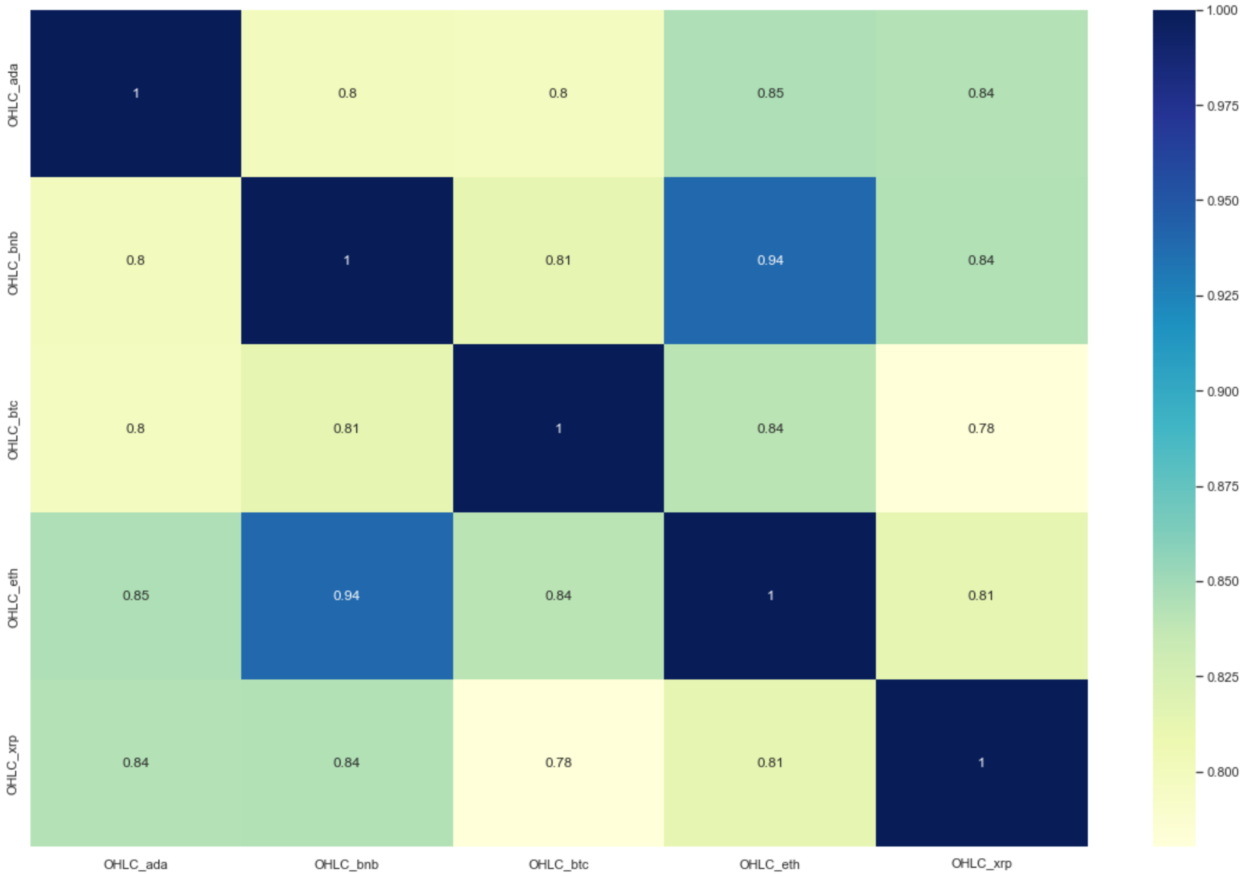


All Combined



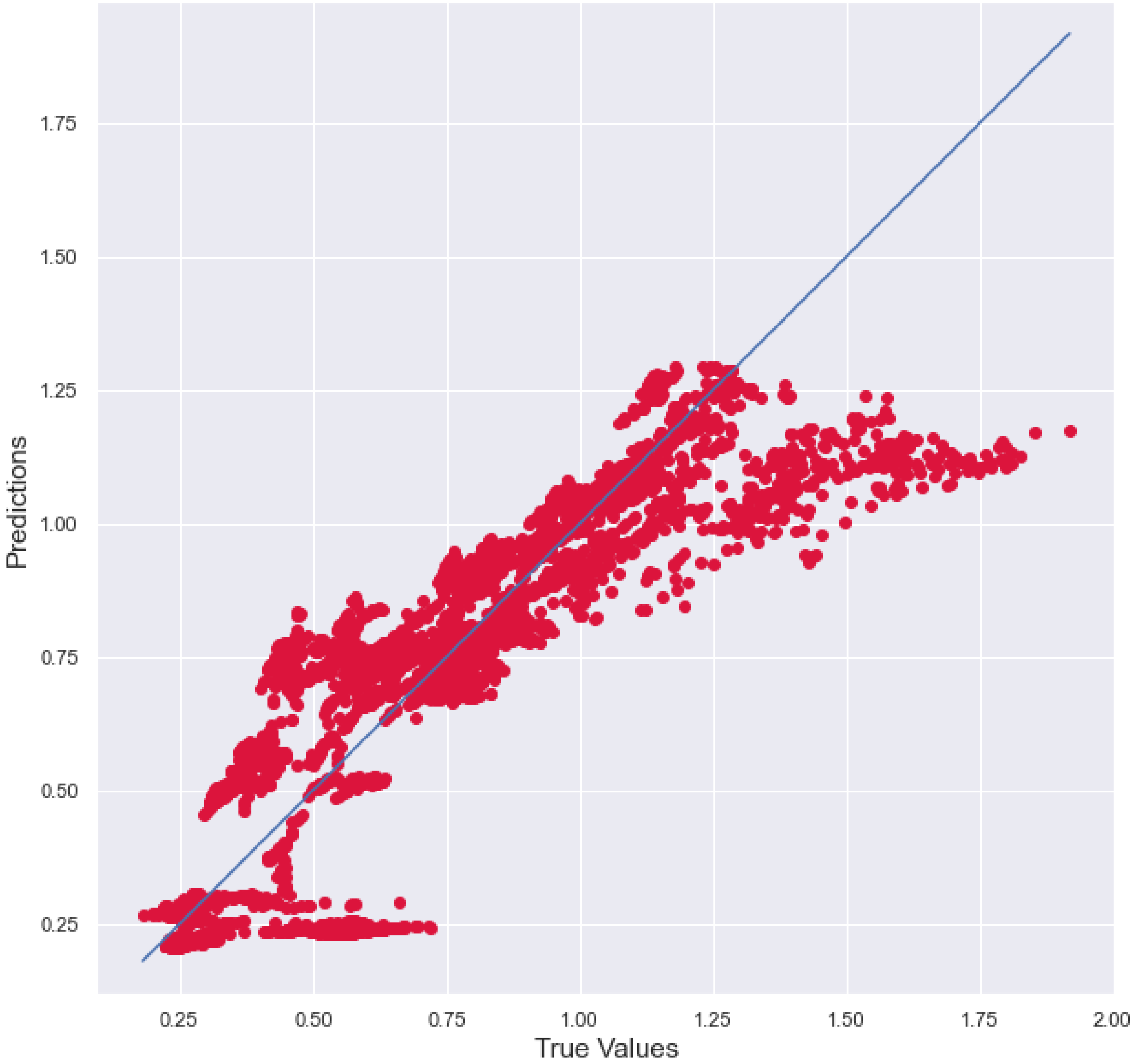
The results from the exploratory data analysis and prediction modelling are below:

1. Correlation Matrix



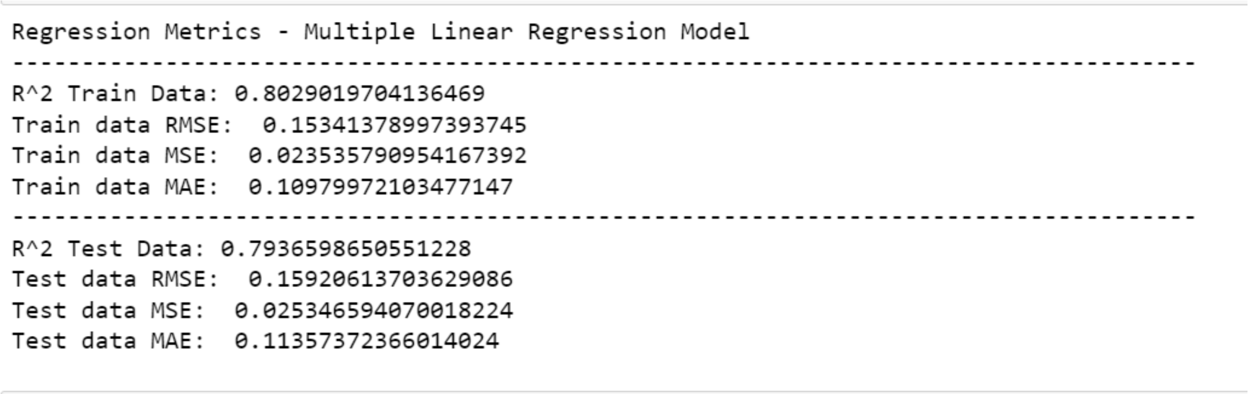
‘OHLC_xrp’ prices are positively correlated to all of the other cryptocurrencies, with it being most correlated to the ‘OHLC_ada’, ‘OHLC_bnb’ price (0.84) and least correlated to the ‘OHLC_btc’ price (0.78).

2. Multiple Linear Regression (MLR)
    a. Graph



The multiple linear regression scatter plot shows the True Values of the test on the x-axis and the predicted values on the y-axis. We can see that the model is poor at higher true values of 'OHLC_xrp' but overall, the model undervalues the 'OHLC_xrp' price consistently.

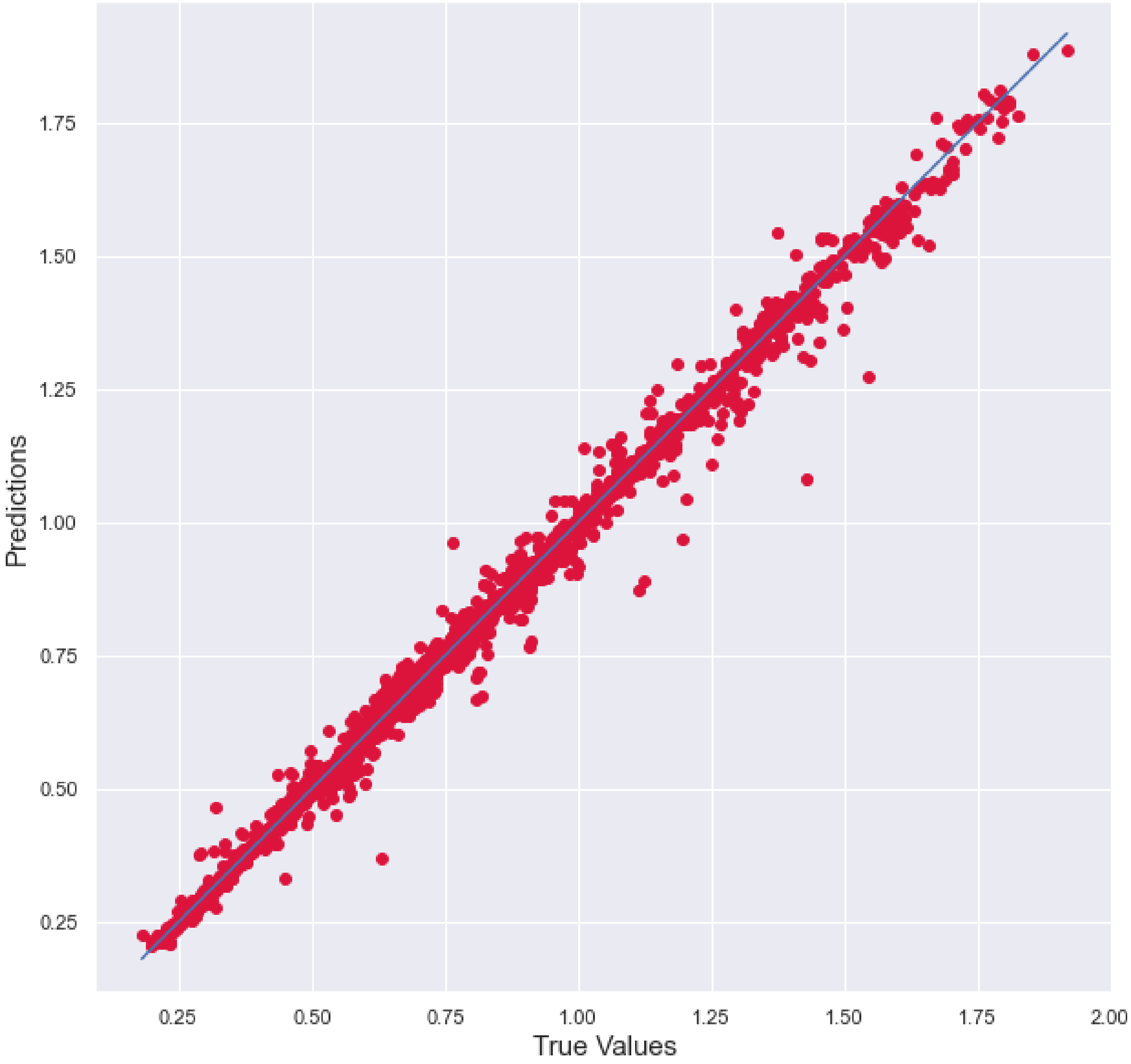
b. Metrics



```
Regression Metrics - Multiple Linear Regression Model
--------------------------------------------------------------------------------------
R^2 Train Data: 0.8029019704136469
Train data RMSE:  0.15341378997393745
Train data MSE:  0.023535790954167392
Train data MAE:  0.10979972103477147
--------------------------------------------------------------------------------------
R^2 Test Data: 0.7936598650551228
Test data RMSE:  0.15920613703629086
Test data MSE:  0.025346594070018224
Test data MAE:  0.11357372366014024
```

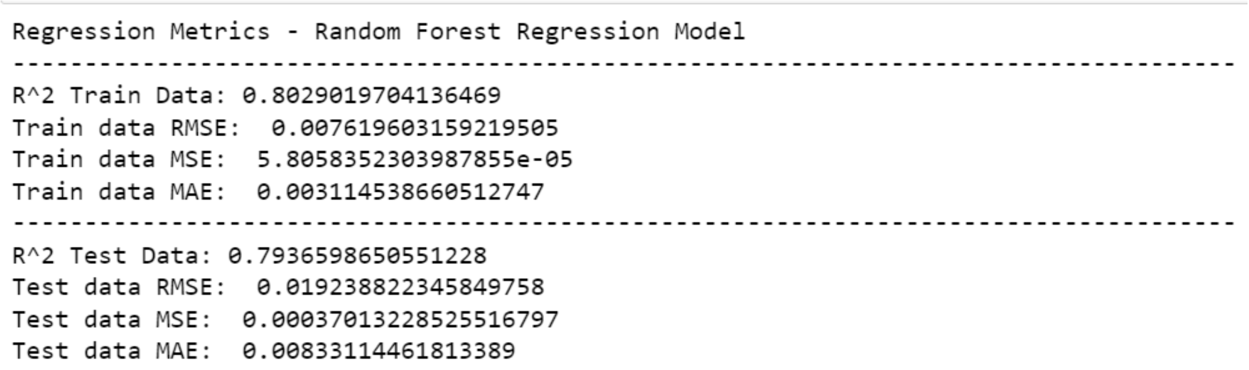
The multiple linear regression model gives has an R^2 score of 0.79 on the test data, a Root Mean Squared Error of 0.15, Mean Squared Error of 0.025 and Mean Absolute Error of 0.11.
Scores on the Training Data were marginally better with slightly higher R^2 and slightly lower RMSE, MSE and MAE values.

3. Random Forest Regression
    a. Graph



The Random Forest Regression scatterplot gives values that are more closely located around the line of fit, and the differences between the true values and the predicted values are less in comparison to the Multiple Linear Regression model.

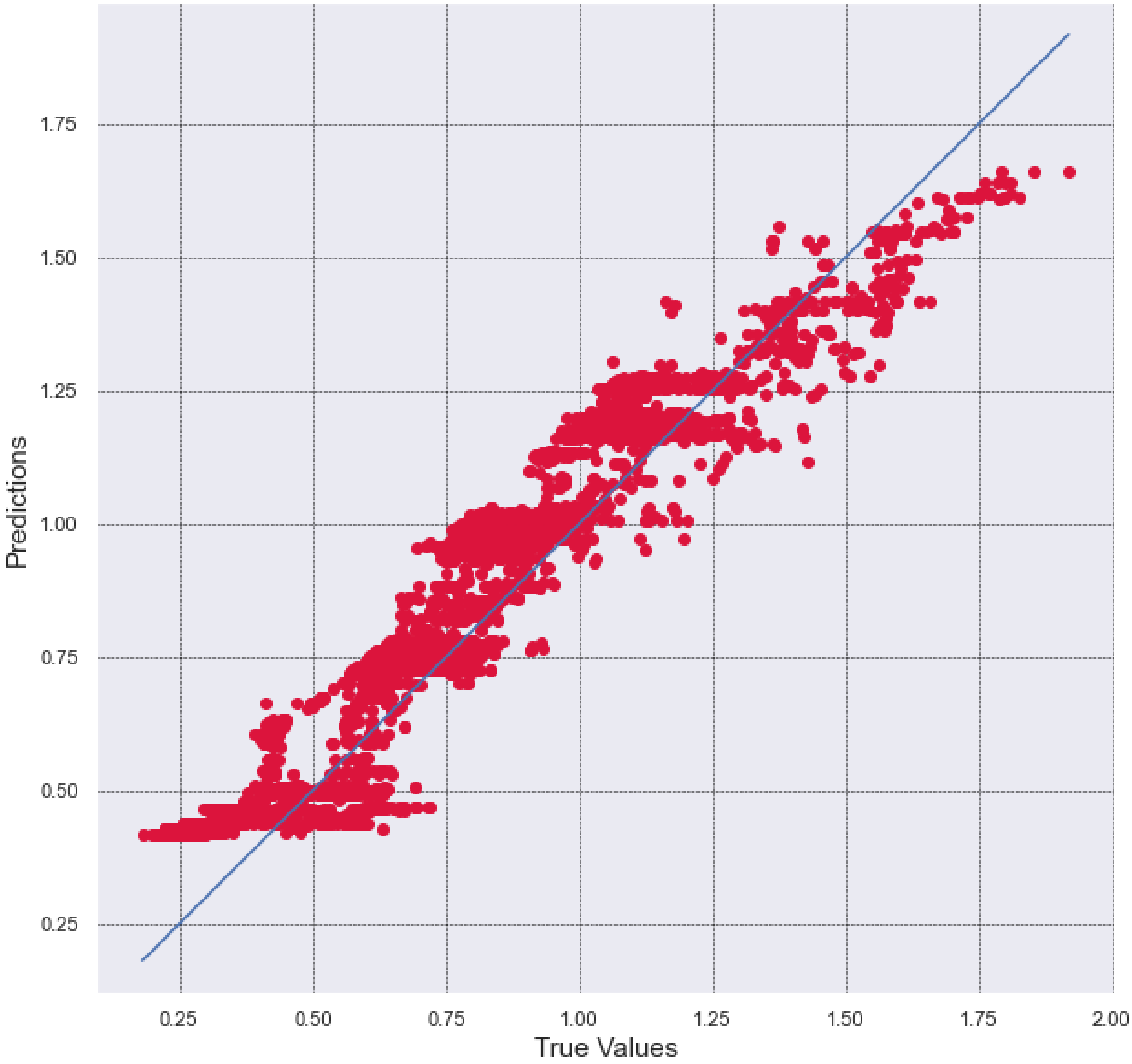
b. Metrics



```
Regression Metrics - Random Forest Regression Model
-------------------------------------------------------------------------------------
R^2 Train Data: 0.8029019704136469
Train data RMSE:  0.007619603159219505
Train data MSE:  5.8058352303987855e-05
Train data MAE:  0.003114538660512747
-------------------------------------------------------------------------------------
R^2 Test Data: 0.7936598650551228
Test data RMSE:  0.019238822345849758
Test data MSE:  0.00037013228525516797
Test data MAE:  0.00833114461813389
```

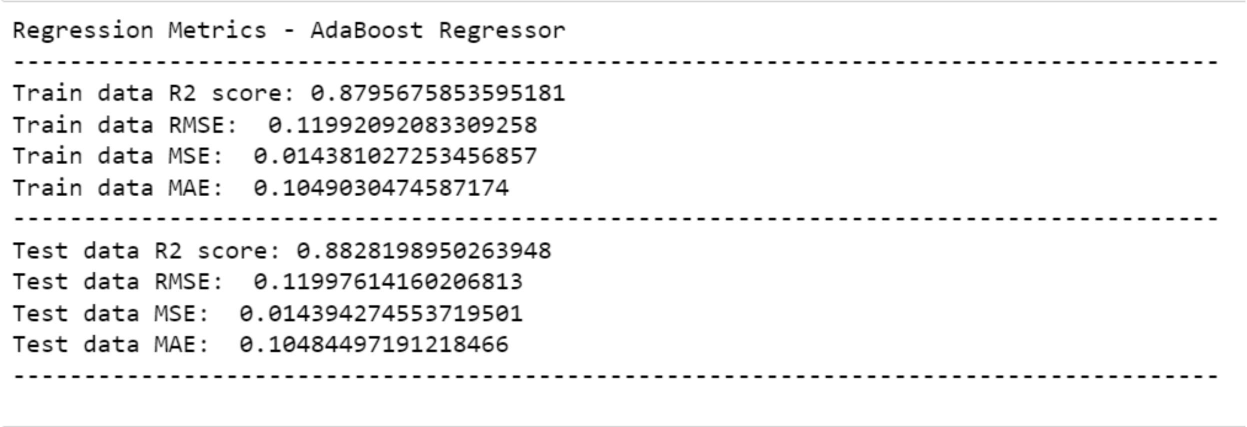
The model has a test data R^2 score of 0.79, with an RMSE of 0.019, MSE of 0.0003 and MAE of 0.008. Scores on the random forest regression model is similar in comparison to the baseline model. The training data scores marginally better.

4. Adaboost Linear Regression (ABR)
    a. Graph



The Adaboost Regressor scatterplot gives values that are closely located around the line of fit.

b. Metrics



```
Regression Metrics - AdaBoost Regressor
-------------------------------------------------------------------------------------
Train data R2 score: 0.8795675853595181
Train data RMSE:  0.11992092083309258
Train data MSE:  0.014381027253456857
Train data MAE:  0.1049030474587174
-------------------------------------------------------------------------------------
Test data R2 score: 0.8828198950263948
Test data RMSE:  0.11997614160206813
Test data MSE:  0.014394274553719501
Test data MAE:  0.10484497191218466
-------------------------------------------------------------------------------------
```

The model has a test data R^2 score of 0.88, with an RMSE of 0.11, MSE of 0.014 and MAE of 0.10. This scores the Adaboost Regression model was best among all 3 models.

Insights

- Overall, the AdaBooster Regression model performs the best against the accuracy metrics in comparison to the other two models. It provides the most accurate predictions between the three models.

- Linear Regression and Random Forest Regression have the same R2 score for test data.

- OHLC_xrp prices are positively correlated to all of the other cryptocurrencies, with it being most correlated to the ‘OHLC_ada’, ‘OHLC_bnb’ price (0.84) and least correlated to the ‘OHLC_btc’ price (0.78). Other cryptocurrencies share a greater correlation, with the ‘OHLC_eth’ showing the most positive correlation using the Pearson method.

- All cryptocurrencies are very volatile in nature. As we can see in the graph in a single day there is a change of more than 20% in price of crypto.

- In early 2021, we saw cryptocurrency market saw boom and matured. The same can be seen in the OHLC graph for all 5 crypto as the price increased exponentially.

- As seen in the combined chart of all crypto, Bitcoin has the most volatility among all 5 crypto used for visualization.

- After so many up and down over a period of age in average price of XRP currency, price in 2020 and 2022 are almost same.