BINANCE TOKEN DATA ANALYSIS

SUBMITTED
BY

MANOHAR KATAM (MXK164930)

SAIPRAVEEN VABBILISETTY (SXV165130)

DECEMBER 1, 2018

UNIVERSITY OF TEXAS AT DALLAS

Table of Contents

1. INTRODUCTION

1.1 About Ethereum and ERC20 Tokens

The main motto of the blockchain technologies is to eliminate the centralized system of transactions. The biggest disadvantage of current financial system which follows a centralized system to authorize transactions is the single point of failure. This has become the bottleneck to the global finance system. Blockchain, with no central point of failure and secured using cryptography, applications are well protected against hacking attacks and fraudulent activities. Immutability, Security, Shield against Corruption and tamper are its salient features. Ethereum enables the development of potentially thousands of different applications all on one platform. Blockchain is going to be the electricity of the future financial systems.

In Wikipedia, Ethereum is defined as "An open-source, public, blockchain-based distributed computing platform and operating system featuring smart contract (scripting) functionality. It supports a modified version of Nakamoto consensus via transaction-based state transitions".

Reference: *https://en.wikipedia.org/wiki/Ethereum*

Ethereum is one of the most popular cryptocurrencies after Bitcoin. ERC-20 is the benchmark/technical standard used for smart contracts on the blockchain platform for implementation. ERC stands for Ethereum Request for Comment and 20 is the number assigned for that request. These rules include how the tokens are transferred between addresses and how data is accessed with in each token.

Reference: *https://en.wikipedia.org/wiki/ERC-20*

1.2 About Binance (BNB) Coin

The network token we chose is Binance Coin (BNB). Binanace is one of the most popular cryptocurrency exchange. The total supply is nearly 200 million. With this cryptocurrency, one can pay a commission for transactions on the exchange. The company offers wide variety of discounts for the customers who opt to pay commission. As per the stats on 28th October,2018, Market cap of this coin is 1.26 billion dollars and volume are $24 million with a current price of 9.63 USD.

2. MOTIVE OF THE PROJECT

The main motive of this project is to find the distribution model of token amounts and understand the pricing pattern of the Ethereum token. The second part of the project deals with extracting features such as layers of transactions, median of all the transactions for a given day which are highly correlated with price data. These features could be used as independent variables to predict the dependent variable i.e., closing price of the network token.

3. PREPROCESSING THE DATA

3.1 Data Description

Data files contain two primary groups: token network edge files, and token price files. The Ethereum project is a blockchain platform, and our data comes from there. Although Ethereum started in 2015, most tokens have been created since 2016. As such, tokens have different starting dates, and their data starts from that initial date.

3.1.1 Token edge data

Token edge files have this row structure: **fromNodeID\ttoNodeID\tunixTime\ttokenAmount\r\n**

This row implies that fromNodeID sold tokenAmount of the token to toNodeID at time unixTime. fromNodeID and toNodeID are people who invest in the token in real life; each investor can also use multiple addresses. Two addresses can sell/buy tokens multiple times with multiple amounts. For this reason, the network is considered a weighted, directed multi(edge) graph. Each token has a maximum token count maxt; you can think of maxt as the total circulating token amount.

3.1.2 Price data

Price files have no extensions, but they are text based. If you open them with a text editor (use notepad++ or similar), you will see this row structure: Date\tOpen\tHigh\tLow\tClose\tVolume\tMarketCap\r

The price data is taken from https://coinmarketcap.com/. Open and close are the prices of the specific token at the given date. Volume and MarketCap give total bought/sold tokens and market valuation at the date.

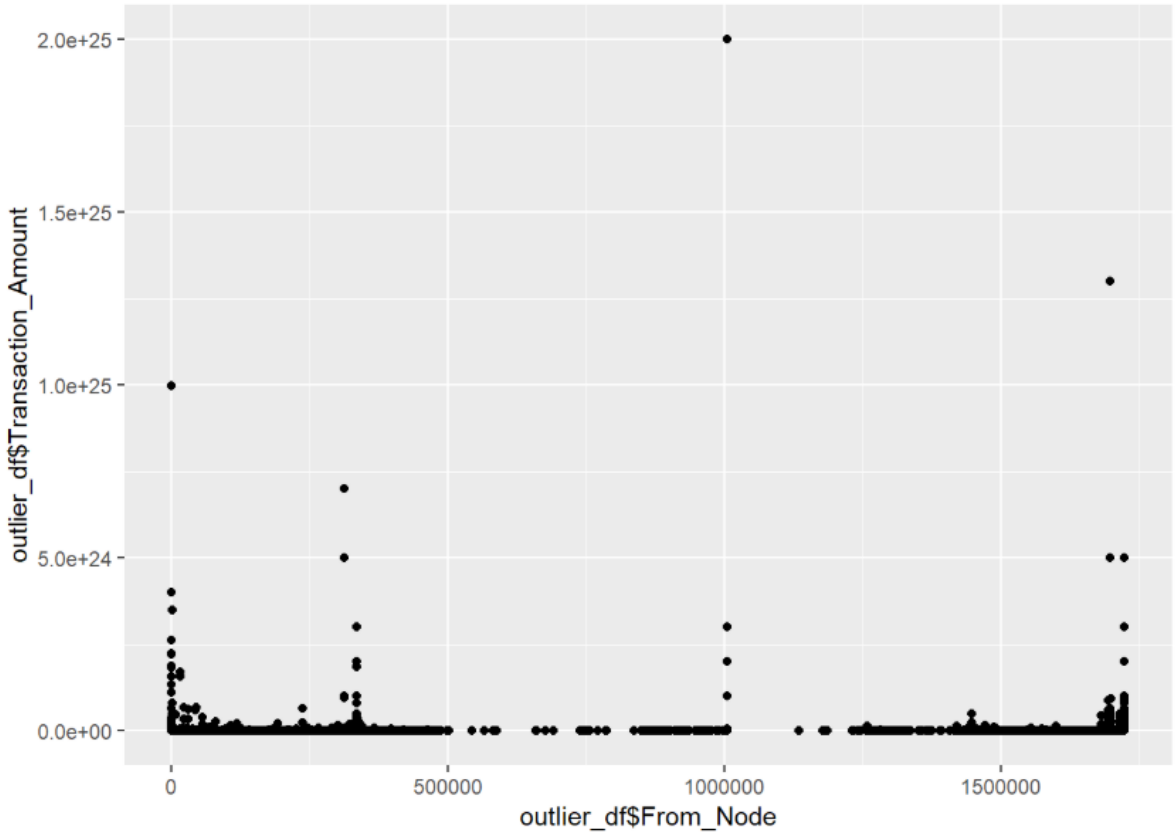
3.1.3 Token supply and sub-unit definitions

Token has a limited supply (i.e., token count, which can be found on coinmarketcap.com as circulating amount). Then each token may have sub-units. This idea comes from Bitcoin where subunits are called Satoshis, 1 Bitcoin $=10^8$ satoshis. Coin market cap gives the total supply, but not sub-units, which differ from token to token. Some tokens have 10^{18} sub-units. That means there can be numbers as big as $\text{totalAmount} * 10^{18}$.

3.2 Checking for Outliers

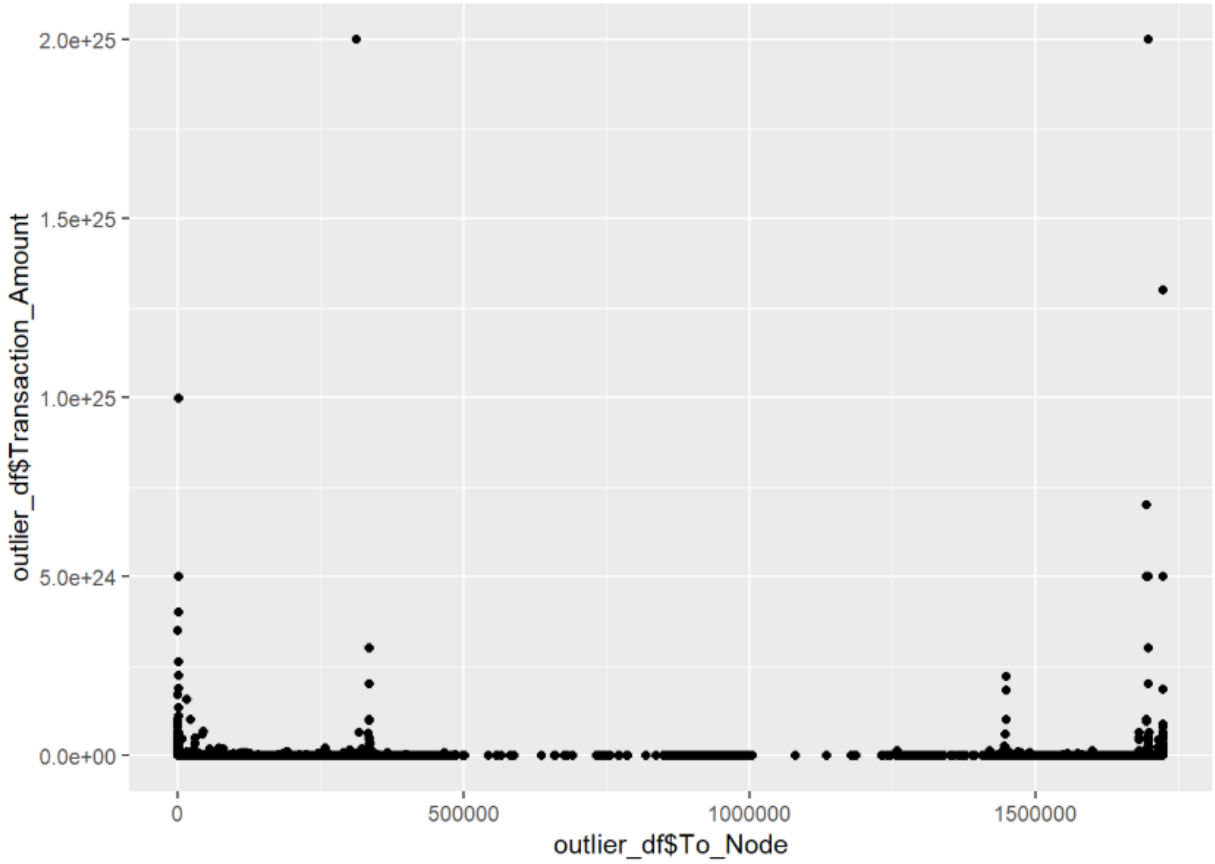
It is found in most of the articles that there can be some noise values in the dataset which would be later removed by the block, but this still has an entry in the dataset.

So, the initial steps of building our project involved in looking for outliers, to be precise looking for transactions whose token amount values are greater than 10^{26}.



Above is the graph for seller to transaction amount. We see that none of the points are outside the range of $1.92x10^{26}$. So, that we conclude here that there are no outliers for seller's vs the transaction amount.

Similarly, in the below graph of buyer's vs transaction amount too there are no outliers.



3.3 Checking for NA values

As another step in preprocessing we must check for NA or N/A values and remove them if we have enough data or we can impute them mean of our data or average of its previous and next value. This step-in preprocessing is called Imputation (i.e., imputing the missing or NA data with meaningful values)

To do this in R we have na.remove = TRUE.

3.4 R Packages Used

Before going any further, here we will describe the R packages we used.

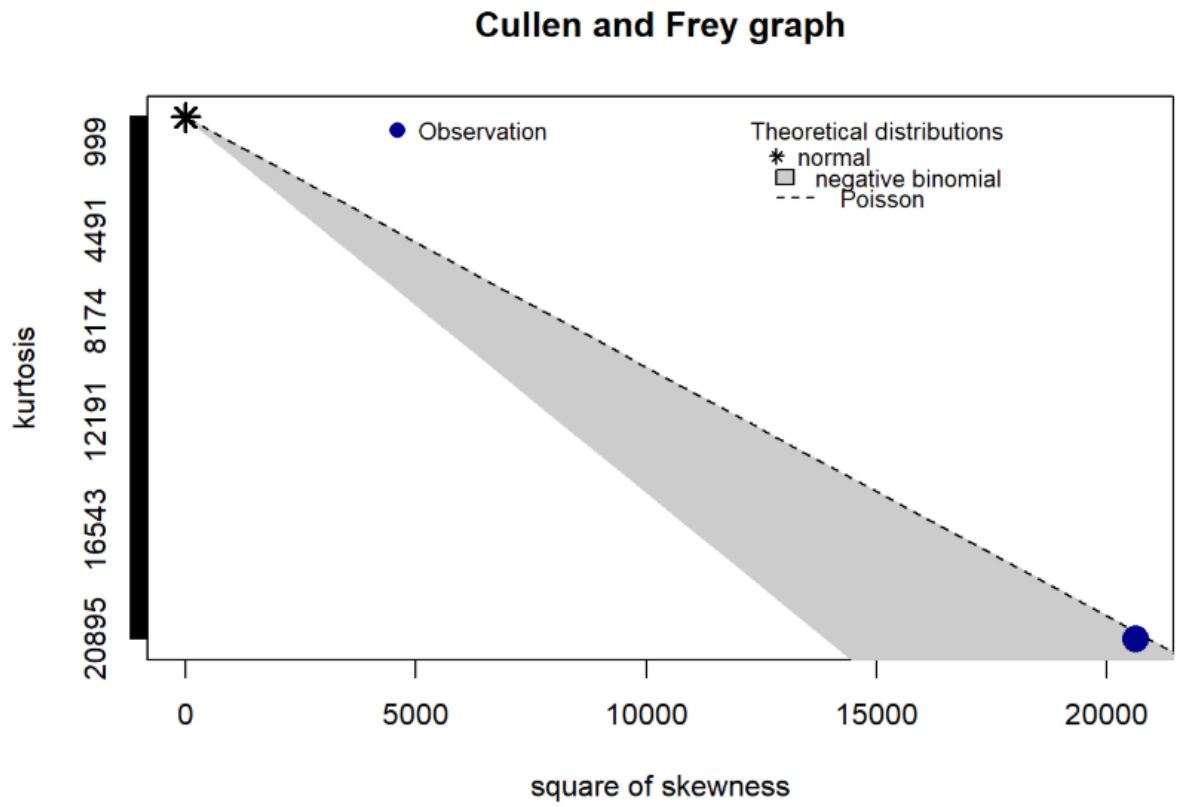
Package Name	Description
ggplot2	package to plot data visualizations
imputeTS	package to replace the "NA" values
plyr	package to implement count
dplyr	package for filtering the data
fitdistrplus	package to fit distribution

4. ANALYSIS AND CONCLUSIONS

4.1 Buyer and Seller Frequency Distributions

4.1.1 Estimating best distribution for selling a token

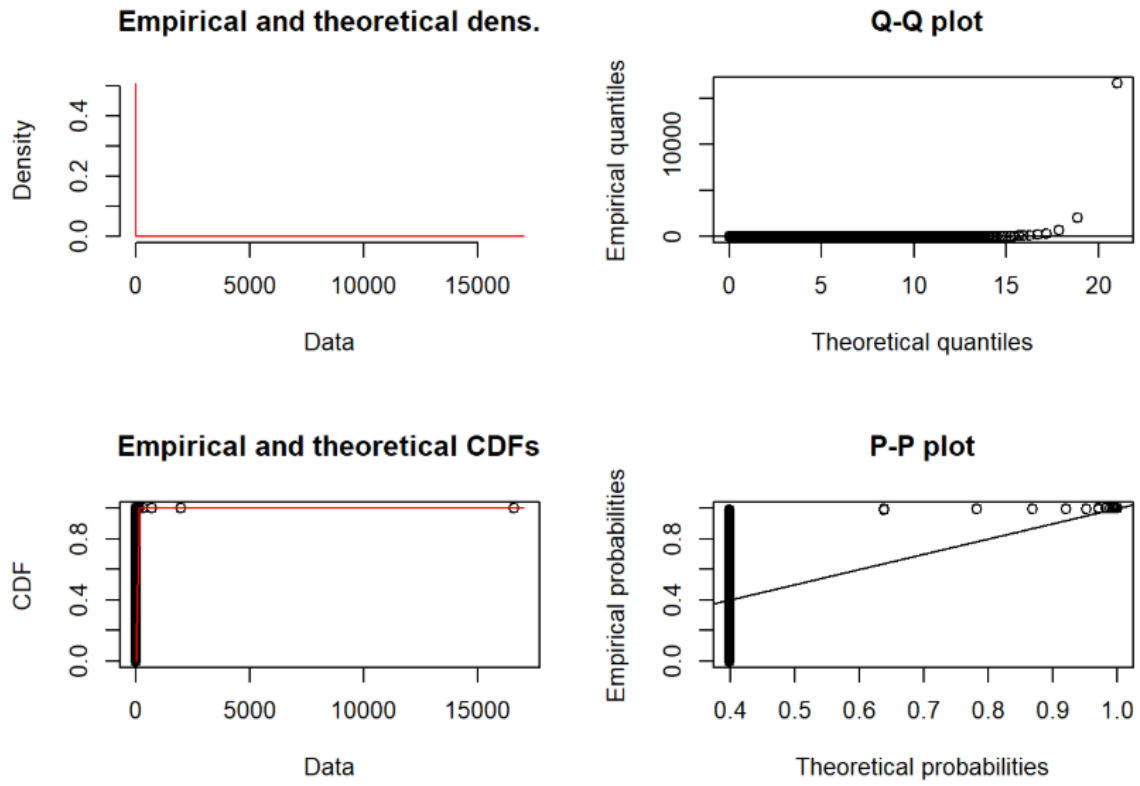
Best distribution for frequency of selling a token:



The above graph gives us an intuition that distribution of data is close to exponential, but Poisson distribution doesn't apply to our case. So the data is fit to Exponential distribution.

Estimated parameters:

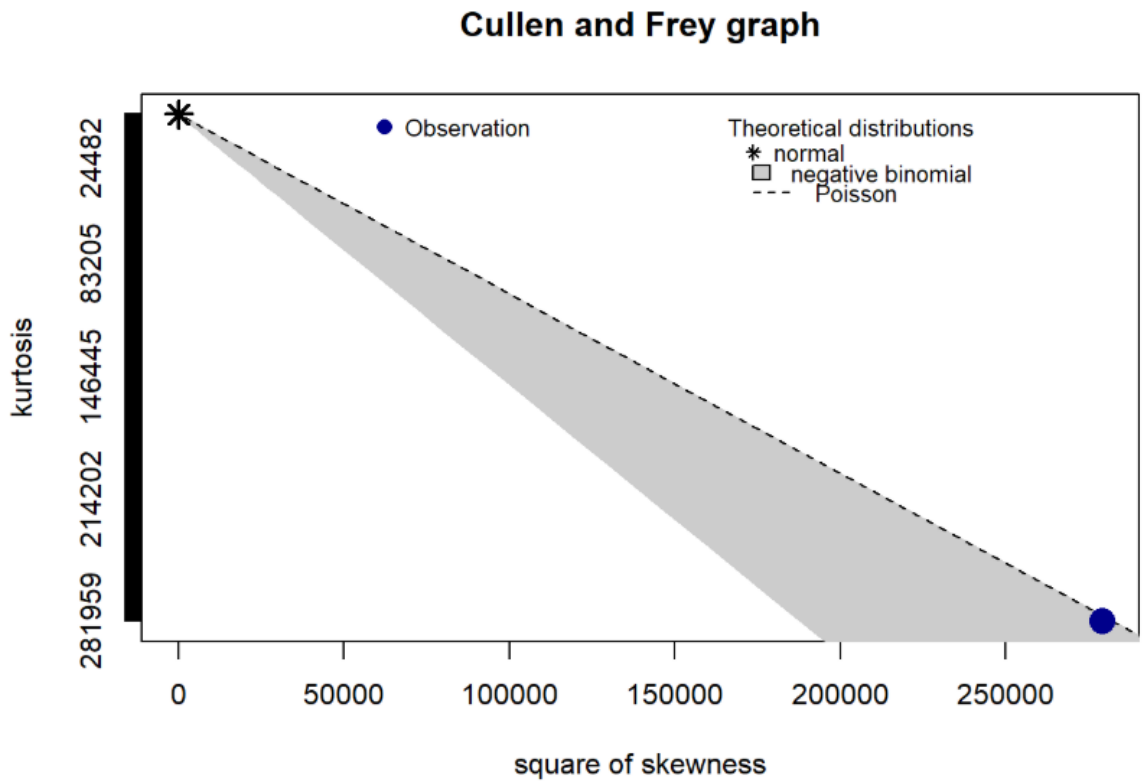
```
## summary statistics
## ------
## min:  1   max:  16575
## median:  1
## mean:  1.968041
## estimated sd:  113.7337
## estimated skewness:  143.5891
## estimated kurtosis:  20894.34
```



If we observe the QQ Plot, the theoretical and Empirical limits are almost the same and gives us an intuition how the data is distributed (exponential)

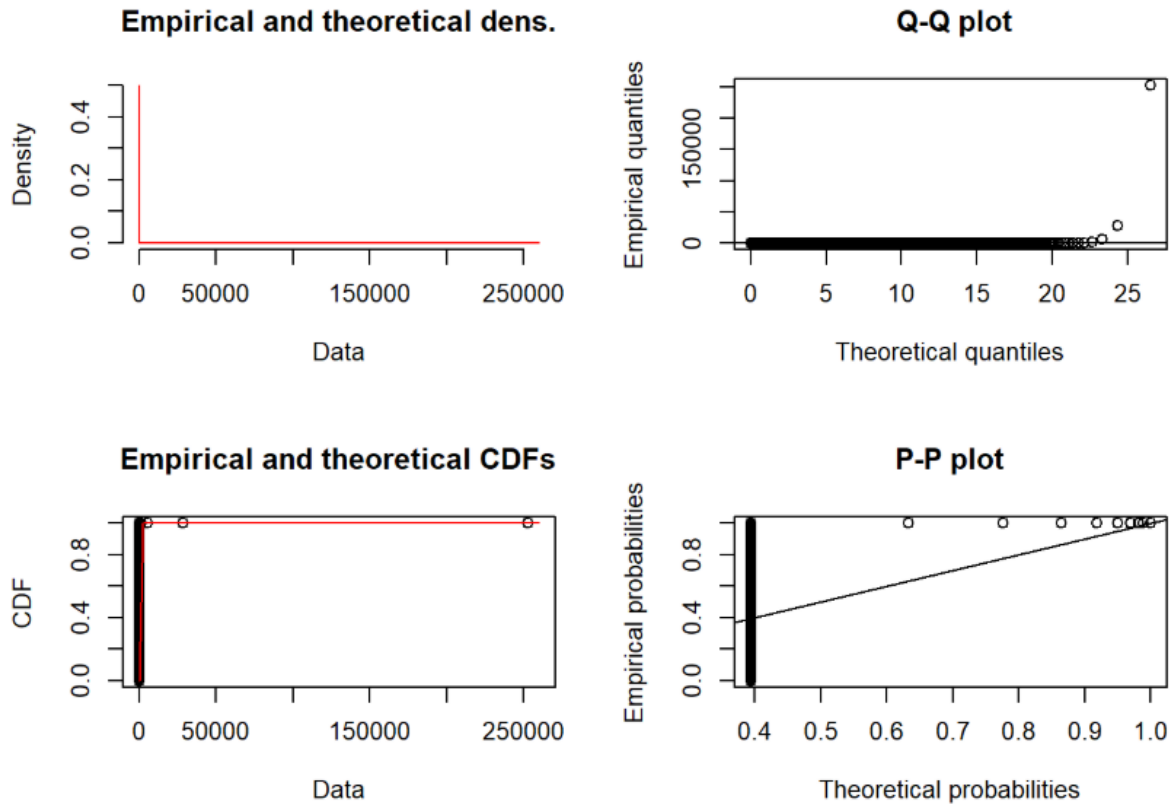
4.1.2 Estimating best distribution for buying a token

Best distribution for frequency of buying a token:



Estimated Parameters:

```
## summary statistics
## ------
## min:  1   max:  252994
## median:  1
## mean:  1.999661
## estimated sd:  473.4125
## estimated skewness:  528.2734
## estimated kurtosis:  281958.7
```



4.2 Number of layers selection and Finding Correlation between the token and price data

The main motive behind dividing into layers of transactions is to approximate the exponential distribution in each of the layers.

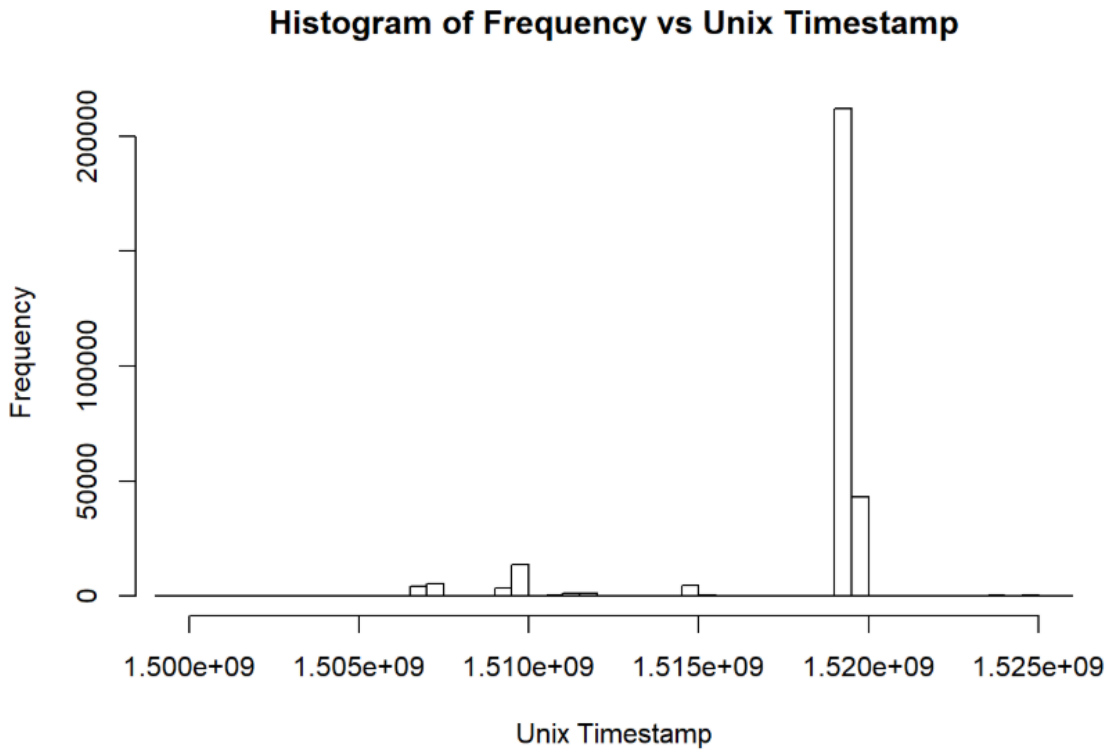
4.2.1 Number of Layers Estimation

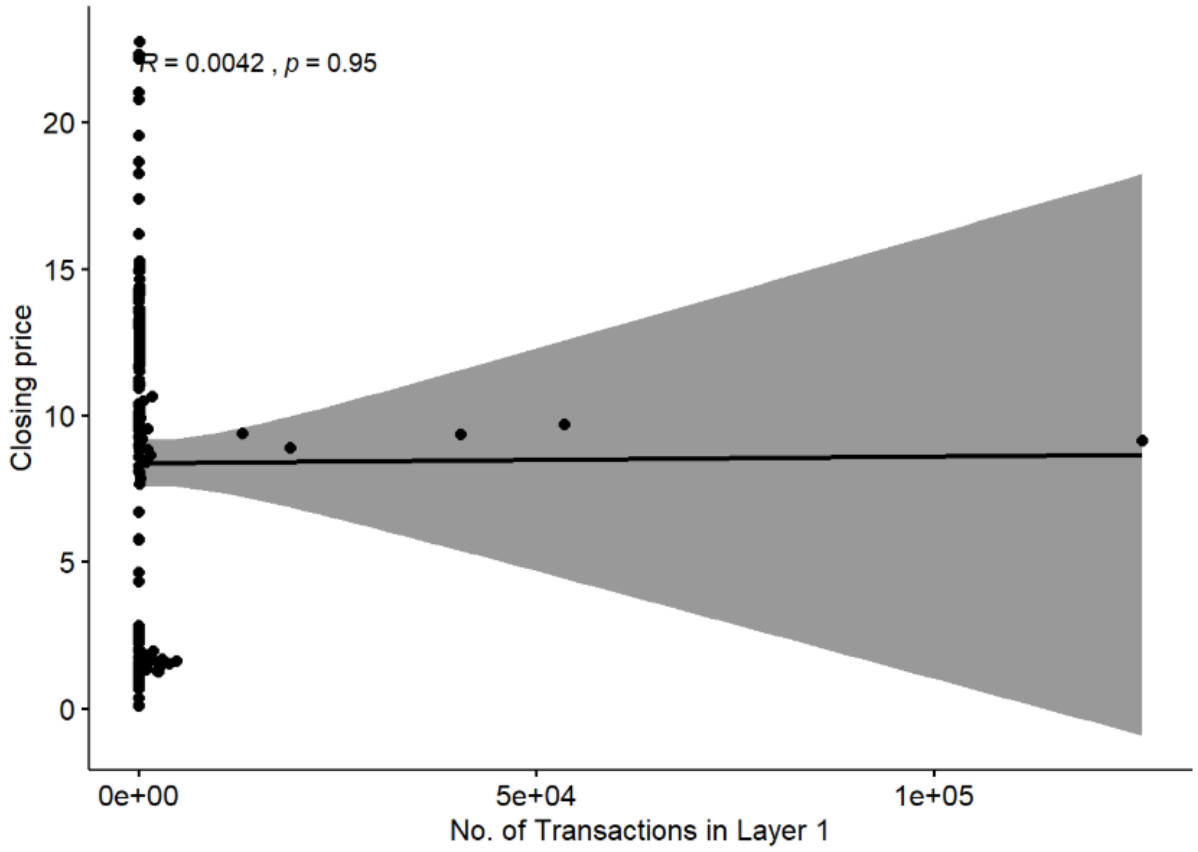
The total token data is divided into three layers based upon density of transactions. The first layer consists of transactions/token amounts less than 1 binance coin, second layer consists of transactions between 1 and 1000 binance coins and the last layer consists of transactions which are greater than 1000 binance coins. The model found that the closing price of the token depends upon the number of second layer transactions on that day.

4.2.2 Finding Pearson Correlation between Features and Price Data

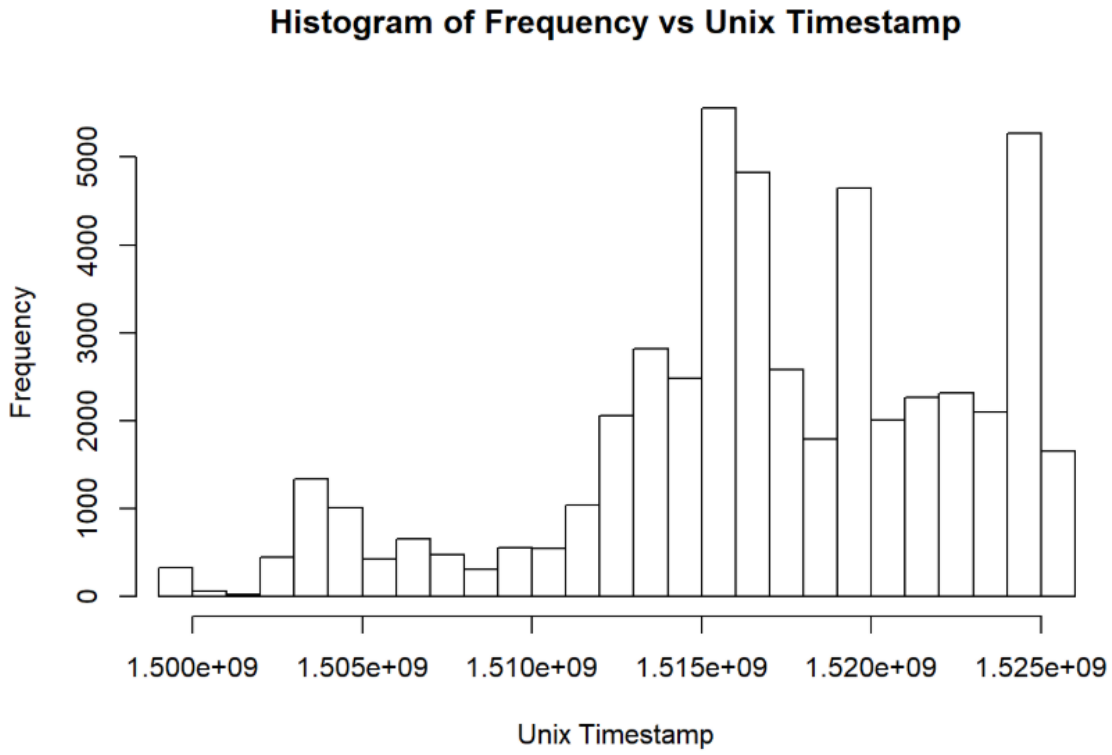
The important findings that are found on implementing this project are the token amount of Binance network token follow an exponential distribution, the closing price of the token for a given day is inversely correlated (negative correlated) to median of token amounts (normalized) and positively correlated (0.6) to number of layer two transactions for a given day.

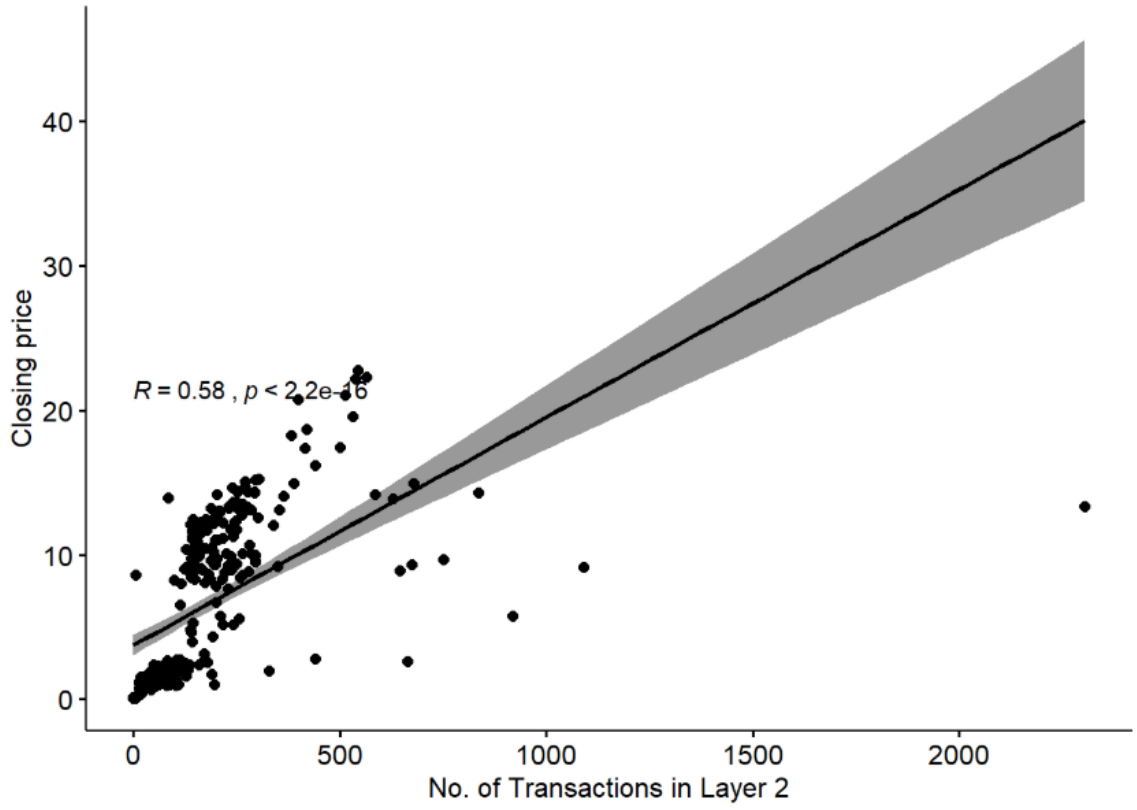
Layer1: Less than 0.817xtotal number of transactions



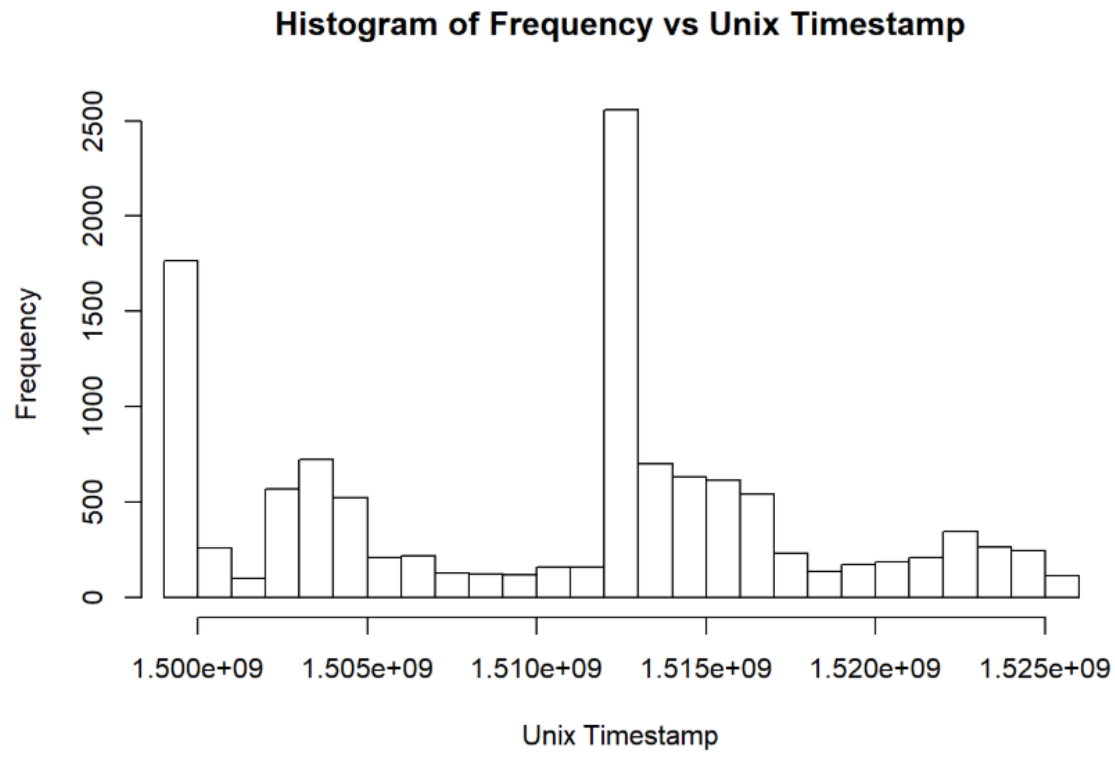


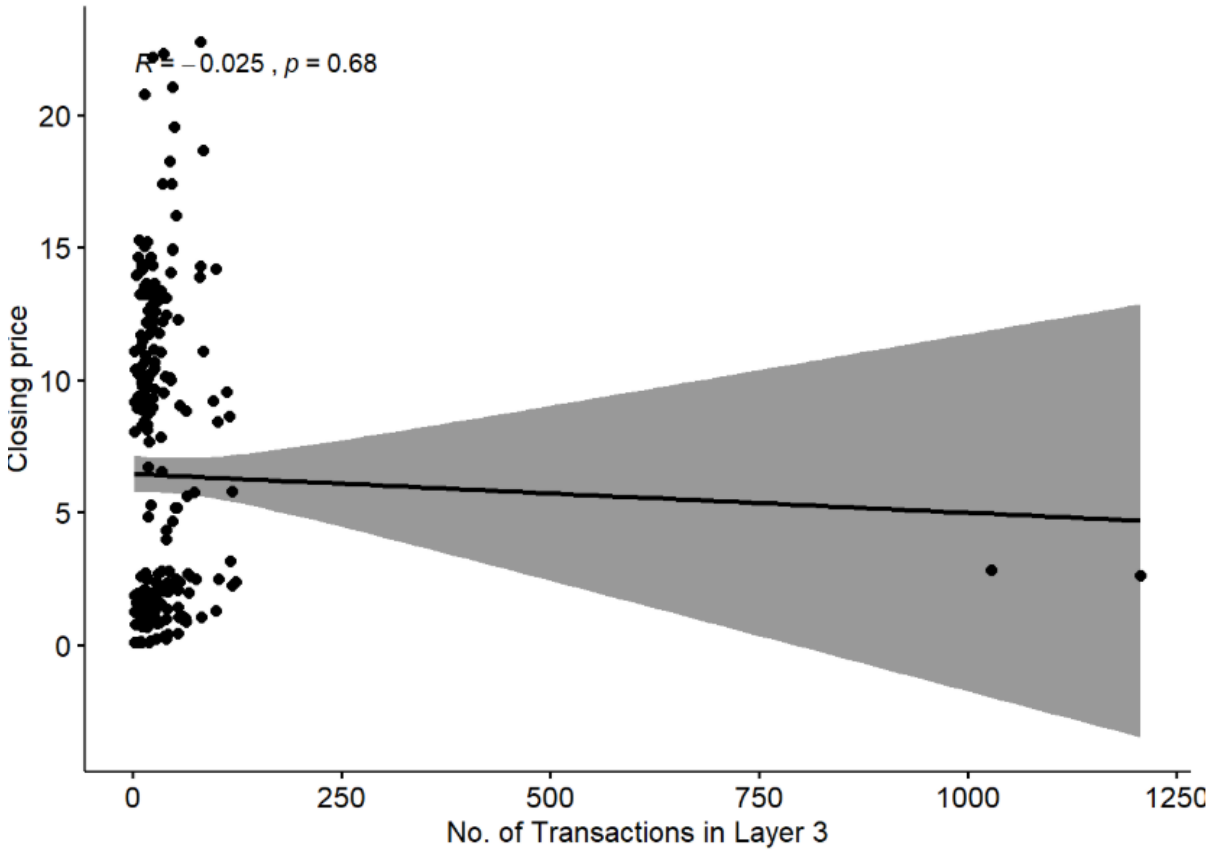
Layer2: 0.817xtotal number of transactions < n < 0.96xtotal number of transactions





Layer3: Greater than 0.96xnumber of transactions


Histogram of Frequency vs Unix Timestamp
Frequency
0
500
1000
1500
2000
2500
1.500e+09
1.505e+09
1.510e+09
1.515e+09
1.520e+09
1.525e+09
Unix Timestamp


R = −0.025 , p = 0.68
Closing price
0
5
10
15
20
0
250
500
750
1000
1250
No. of Transactions in Layer 3

5. Feature Engineering

From the above experiment, although Layer 2 transactions play a key role it cannot be concluded that Layer 2 transactions occur every day. Taking the layer 2 transactions and fitting the model gives poor result. The other features like median and standard deviation of token amounts when fitted to a model the r squared values for each of these features are non-significant.

After going through the data and understanding it thoroughly we came across these interesting questions.

When you look at the closing price reported per day in the data, can we say that the closing price of current day is affected by conditions on previous day?

What could be the relationship between the reported closing price and the data provided for the day?

What we observed here is that closing price of the day is not a direct cause-effect of the conditions recorded on that same day. This is because stock market's value depends on many other factors like what happened on the previous day or the days before. Example like issuance of a government's new policy which has effect on the company or the quarterly results which company publishes are the ones which directly effect the prices of coin.

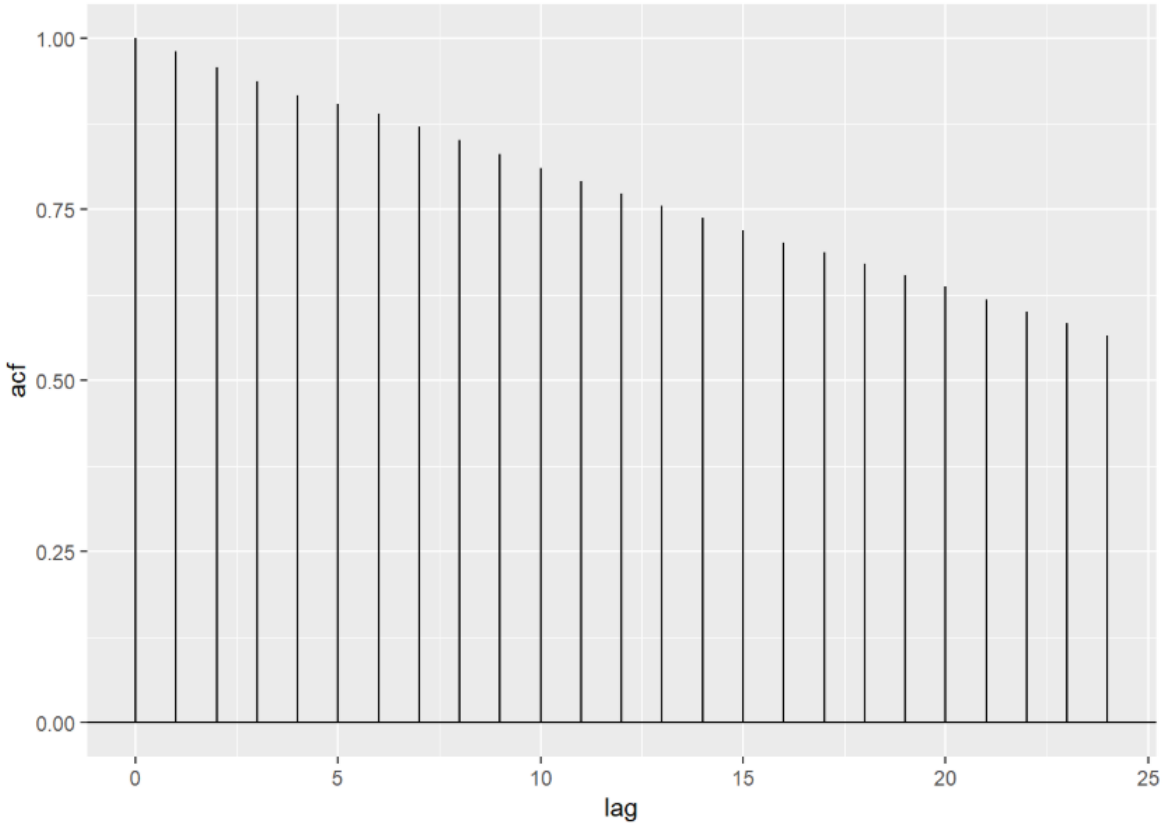
The other problem here is that we should also account for the outbreaks and in breaks. Sometimes the price of the coin rigorously increases or decreases. To catch up with this we have to build a model that progressively predicts a new value while taking into account the previous prediction.

Since we are essentially dealing with time series here, we thought it would be good to add variables on past time lags for each observation.

6. Model Fitting and Validation

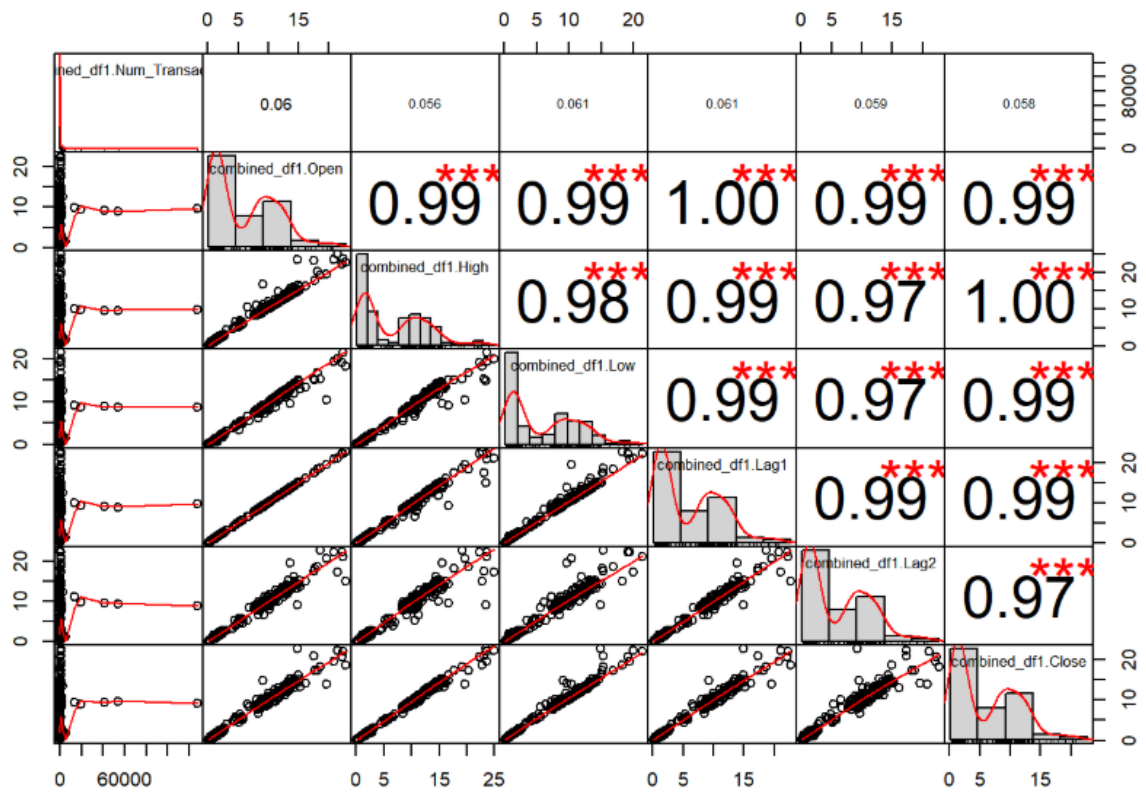
Performance Metrics Used: Mean Square Error (MSE), R Squared and Predicted vs Actual plots

Auto correlation plot of lag features:



The autocorrelation plot above confirms the correlation of the closing price with the previous days. If we consider 0.5 as a good correlation, the closing price is correlated with lags up to 30.

Correlation plot:



From the above correlation plot we see that the Lags as explained are highly correlated.

7. CONCLUSION

Frequency of Buys and Sells: We could see that from the above plots the token amounts are distributed exponentially. This is a feasible distribution because as the feature importance grows on increasing day by day, the token amounts being increased in the same way. For Example, a country 'X' has allowed crypto exchange as their official exchange there is a high chance the token amounts would increase.

Number layers and correlation: From the above experiments, results conclude that the number of layer-2 transactions, Median of token amounts, Standard deviation of token amounts for a given day could be the most significant features which contributes the maximum variance to the distribution. These results could be used in building a multi/polynomial regression model and could be used in predicting the price of Binance Closing price for a given day.

Findings From above Layers:

The most significant features which contributes to prediction of closing token price from our experiments are

1.**The number of Layer 2 transaction (1 to 1000 binance coins) with a positive correlation of 0.58.**

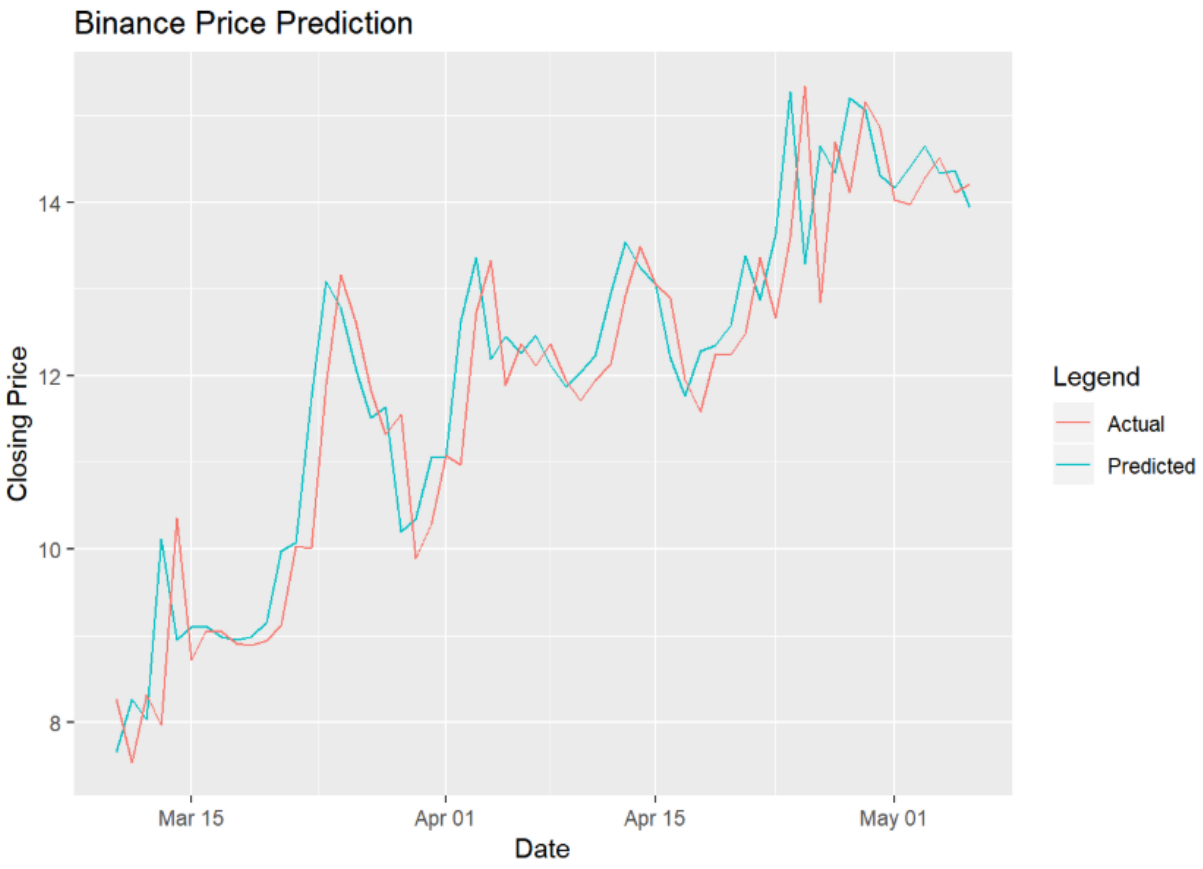
2.The Median of token amounts for a given day with a negative correlation of -0.29

3.The Standard Deviation of token amounts for a given day with a negative correlation of -0.23

Actual vs Predicted values plot and Multi-Linear Regressor model fitting:

Multiple linear regression model for predicting today's closing price:

Number of Lags: 2 (With Lag 1 and Lag 2 features)

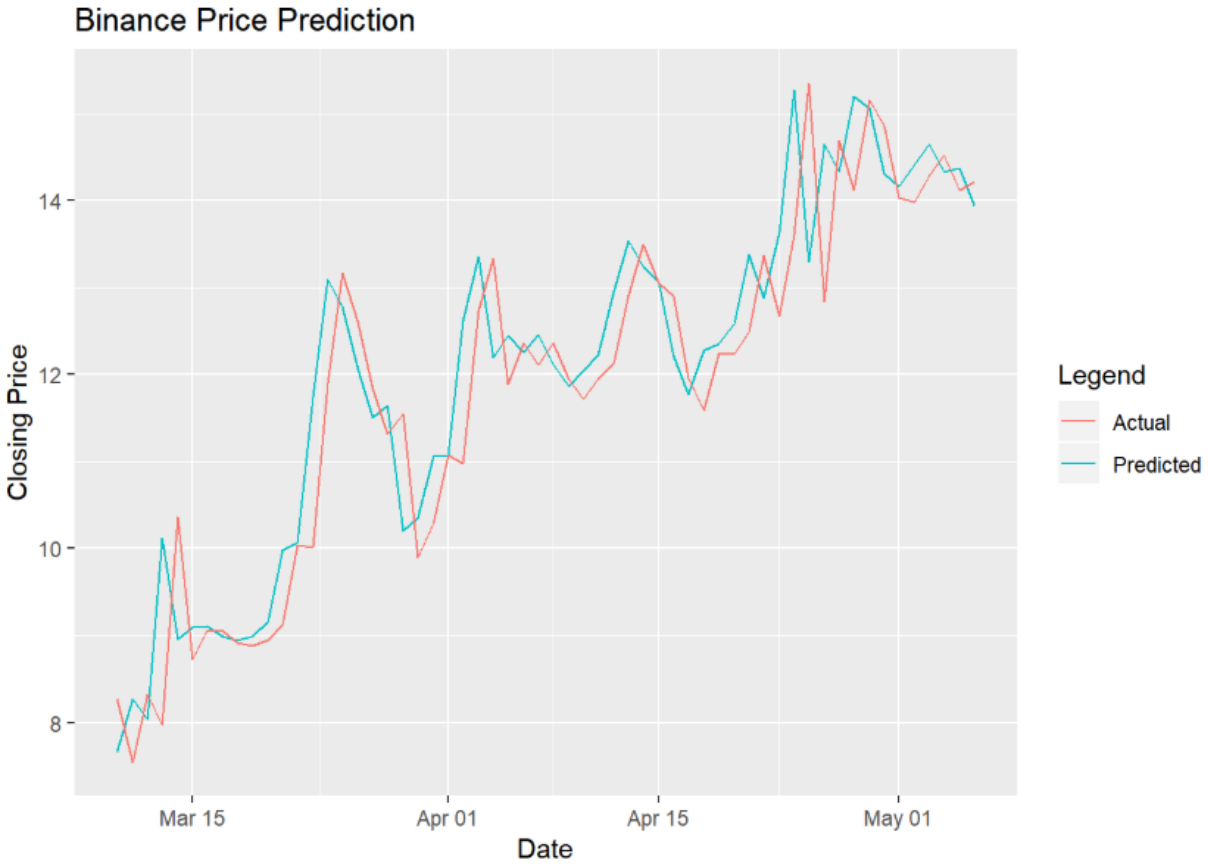


Coefficients of estimation:

```
## lm(formula = Close ~ Lag1 + Lag2, data = training_data)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2099 -0.2186 -0.0952  0.1530  7.1299
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13043    0.08788   1.484   0.1392
## Lag1         1.13461    0.06606  17.175   <2e-16 ***
## Lag2        -0.15513    0.06600  -2.351   0.0196 *
## Residual standard error: 0.964 on 224 degrees of freedom
```

```
## Multiple R-squared:  0.9676, Adjusted R-squared:  0.9673
## F-statistic:  3340 on 2 and 224 DF,  p-value: < 2.2e-16
```

Number of Lags 2 (With open price and lag features)




```
## lm(formula = Close ~ Lag1 + Lag2 + Open, data = training_data)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.2070 -0.2209 -0.0963  0.1518  7.1199
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.13034    0.08809   1.480   0.1404
## Lag1         1.19643    1.22218   0.979   0.3287
## Lag2        -0.15601    0.06841  -2.280   0.0235 *
## Open        -0.06080    1.20038  -0.051   0.9596
## Residual standard error: 0.9662 on 223 degrees of freedom
```

```
## Multiple R-squared:  0.9676, Adjusted R-squared:  0.9671
## F-statistic:  2217 on 3 and 223 DF,  p-value: < 2.2e-16
```

REFERENCES

1. *https://en.wikipedia.org/wiki/Ethereum*
2. *https://en.wikipedia.org/wiki/ERC-20*
3. https://coinmarketcap.com/currencies/binance-coin/historical-data/
4. https://www.binance.com/en
5. https://www.rdocumentation.org/packages/ggplot2/versions/3.1.0
6. http://www.di.fc.ul.pt/~jpn/r/distributions/fitting.html
7. https://cran.r-project.org/web/packages/dplyr/vignettes/dplyr.html

APPENDIX

We also open sourced our source code on GitHub and published our results. Below are the links.

GitHub code base: https://github.com/saipraveen1994/Binance-Price-Predictions

Web page for documentation:

https://saipraveen1994.github.io/Binance-Price-Predictions/Project1.html

https://saipraveen1994.github.io/Binance-Price-Predictions/Project2.html

END OF PROJECT REPORT