Statistical Methods for Data Science
Final report submission

By Kavya Jampani, Beryl Mario Shairu Joseph Antony Bose

Analysis of the Tronix token data

Website: http://www.utdallas.edu/~kxj170004/Stats_project_Kavya/

Introduction

Ethereum and ERC-20 tokens:
Ethereum was proposed by Vitalik Buterin, a cryptocurrency researcher and programmer. Ethereum is a distributed public blockchain framework which uses the block chain technology to create smart contracts to provide a static and consistent transactions record. Although commonly associated with bitcoin, there are significant technical differences between the two. There is a substantial difference in both in terms of purpose and capability. Ethereum enables developers to build and deploy any decentralized application whereas Bitcoin offers one particular application of blockchain technology i.e, a peer-to-peer electronic cash system that enables online bitcoin payments. In Ethereum blockchain, miners work to earn Ether, a type of crypto token that fuels the network. Ethereum is also used a platform to launch other cryptocurrencies. The benefits of the Ethereum decentralized platform are as follows:
Immutability - A third party cannot make any changes to the data.
Corruption and Tamper proof - Apps are based on a network formed around the principle of consensus, making censorship impossible.
Secure- With no central point of failure and secured with cryprography, applications on this framework are well protected against hacking attacks and fraudulent activities.
Zero downtime - Applications can never go down and can never be switched off.

Token:
In general sense, a token is used to describe any digital asset. They serve as the transaction units on the blockchains that are created using the standard templates like that of Ethereum network, where user can create his own tokens.

ERC-20 Tokens:
These are the tokens designed solely for the usage on Ethereum platform. ERC-20 is a technical standard used for smart contracts on the Ethereum blockchain for implementing tokens. It defines a common list of rules that an Ethereum token has to implement, giving developers the ability to program how new tokens will function within Ethereum ecosystem. According to Etherscan.io there are a total of 103621 of ERC-20 compatible tokens found on Ethereum main network as of 07-26-2018. Because of ERC-20 token standard definition other developers can issue their own versions of the token and raise funds with an initial coin offering (ICO). Depending on its purpose, decentralized applications might use ERC-20 tokens to function as a currency, a share in a company or even proof of ownership. These tokens are created using smart contracts. ERC-20 makes the creation of new tokens extremely easy, and that is why Ethereum has become most popular platform for ICO's in 2017.

Primary Token
In our project we have considered Tronix token which is based on ERC-20 Ethereum standard.
About the token:
TRON Foundation was established in September 2017 by Justin Sun. TRX is the cryptographic currency developed by TRON, which aims to be a decentralized entertainment content sharing platform using blockchain and peer-to-peer network technology. In 2018 TRON launched its own proprietary blockchain, Mainnet to which it migrated all the TRX (ERC-20) tokens that previously circulated on the Ethereum blockchain. In February 2018, TRX was ranked 15th on the list of largest cryptocurrencies by market capitalization. TRON Foundation seeks to tackle the

global entertainment industry and is currently valued at $1 trillion. Bit-Z, Liqui, and Gatecoin are the next three biggest TRX exchanges. Major features of TRON are:
-Uncontrolled and free data
-Using content spreading, to enable content ecosystem, where users ca obtain digital assets.
-Initial Coin Offering to distribute the digital assets
-Framework to allow distributed digital assets exchange (such as games) and market forecasting.

Scope

In this project, we examined the Ethereum dataset: Tronix(TRX). We have preprocessed the data removing the outliers where the transactions are impossible and then identified the distributions of number of user transactions by buyer and number of user transactions by the seller. Then we tried to find the correlation of the number of users made transactions on a specific date with the closing price on the same date. Also, we have modelled multiple regression to predict simple price return.

Dataset

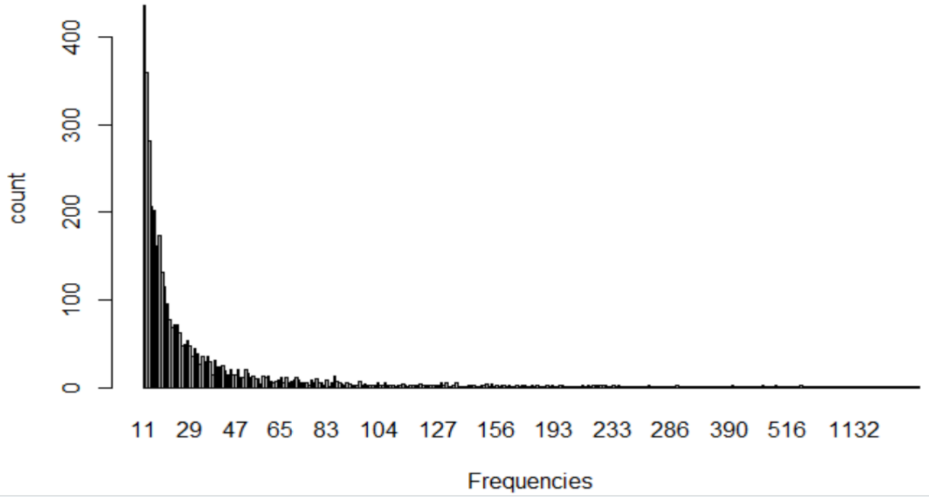
The dataset we have considered has two files:
Token edge file which has 1,51,8537 transactions. This set has the row structure of "fromNodeID NodeID unixTime tokenAmount" which implies that fromNodeId sold tokenAmount of the token to toNodeId at time unixTime. Each token has a total circulating amount and subunits. For our token the value is 100000000000 (according to coinmarketcap.com) and has 10e+06 subunits. Token price file which has 254 rows and the row structure is "Date Open High Low Close Volume MarketCap"" which provides the price data on a specific date. Open and Close are the prices of the token at the given date. Volume and MarketCap give total bought/sold tokens and market valuation at that date.

Preprocessing

Based on the total supply value of the tronix token, we have removed the outliers which are greater than totalAmount*subunits. There are 65 transactions where the token amount is greater than the totalsupply*subunits and 41 unique users are involved in these transactions.

Distribution of how many times user buys a token

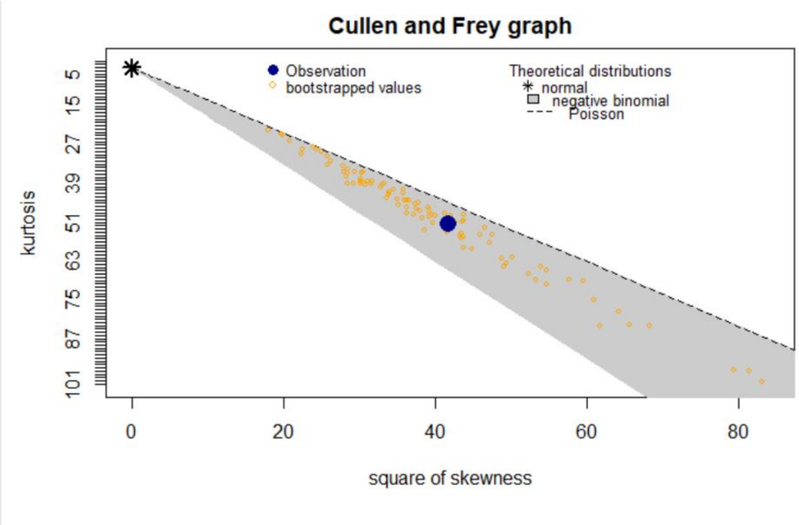
To observe the behavior of number of times user buys a token, we have derived frequencies for each user and then plotted the distribution of frequencies against user counts. As most of the data belongs to smaller frequencies, we have limited the set to see the behavior of the dataset. For this we have considered the subset of data where user counts<= 600. This results in a set of 97% data excluding 10 outliers. We have used ggplot2 package of R for visualization.

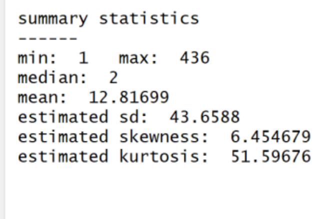


The parameters of the data are

Mean: 12.816

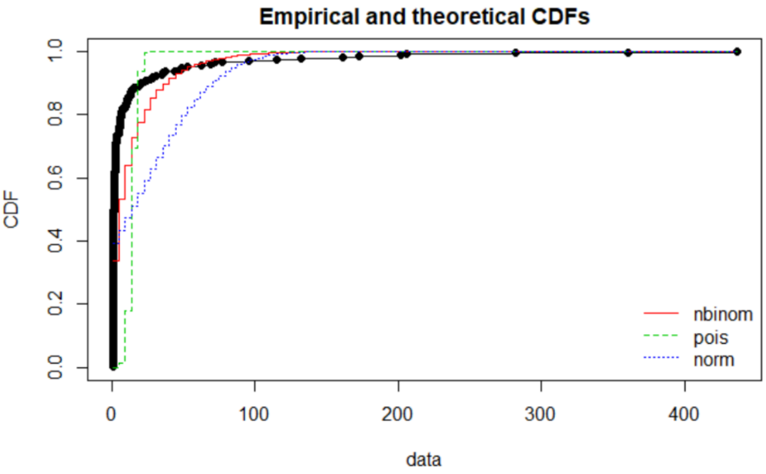
Median: 2
Variance: 1906.09
Standard deviation: 43.66
'Cullen and Frey Graph' gives the summary of best distributions that can fit the data From the below graph, we can see that normal, negative binomial and poisson distributions fit the data better. We have used 'fitdistrplus' package of R for the data visualization. We have used the descdist() function of ‘fitsdistr’ package for visualization.



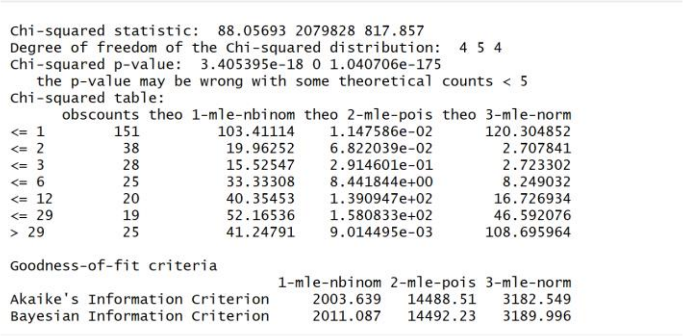


```
summary statistics
------
min:  1   max:  436
median:  2
mean:  12.81699
estimated sd:  43.6588
estimated skewness:  6.454679
estimated kurtosis:  51.59676
```

Cumulative Distribution Function (CDFs) have most meaning for visualizing the discrete fits. We have fitted the above distributions using fitdist() function of the ‘fitdistr’ package and compared the CDFs of the above ditributions. The CDFs of the above distributions are as follows:

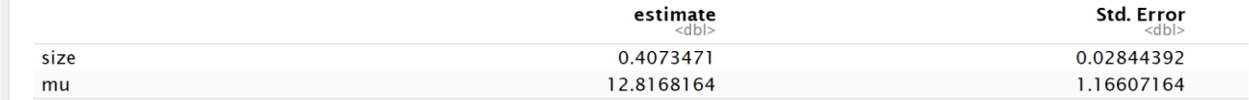


The Goodness-of-fit criteria is calculated by using the gofstat() function



```
Chi-squared statistic:  88.05693 2079828 817.857
Degree of freedom of the Chi-squared distribution:  4 5 4
Chi-squared p-value:  3.405395e-18 0 1.040706e-175
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
       obscounts theo 1-mle-nbinom theo 2-mle-pois theo 3-mle-norm
<= 1         151         103.41114    1.147586e-02      120.304852
<= 2          38          19.96252    6.822039e-02        2.707841
<= 3          28          15.52547    2.914601e-01        2.723302
<= 6          25          33.33308    8.441844e+00        8.249032
<= 12         20          40.35453    1.390947e+02       16.726934
<= 29         19          52.16536    1.580833e+02       46.592076
> 29          25          41.24791    9.014495e-03      108.695964

Goodness-of-fit criteria
                               1-mle-nbinom 2-mle-pois 3-mle-norm
Akaike's Information Criterion     2003.639   14488.51   3182.549
Bayesian Information Criterion     2011.087   14492.23   3189.996
```

From the above summary, we can observe that the AIC and BIC counts are minimum for the negative binomial distribution. So, we can say that the negative binomial distribution fits the data better. The estimated parameter of the distribution is as follows:



	estimate <dbl>	Std. Error <dbl>
size	0.4073471	0.02844392
mu	12.8168164	1.16607164

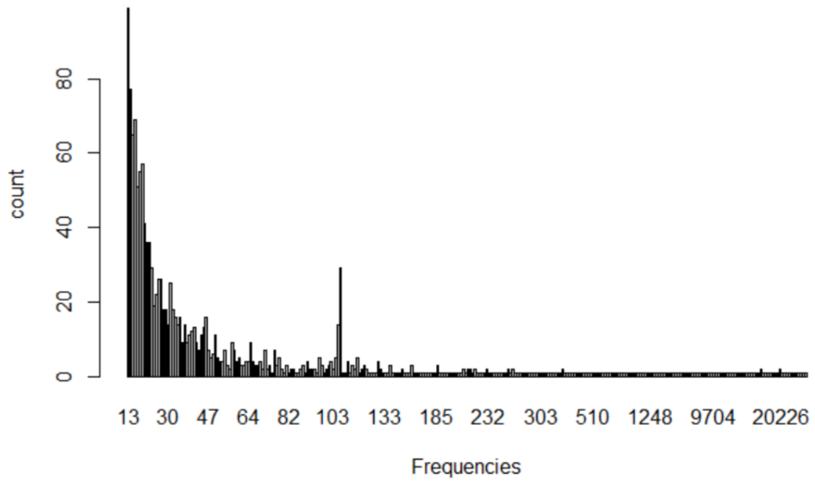
The estimated mean of the distribution is same as the mean of the sample. The negative binomial distribution can be modelled with mean = 12.82

and p = size/mu = 0.032

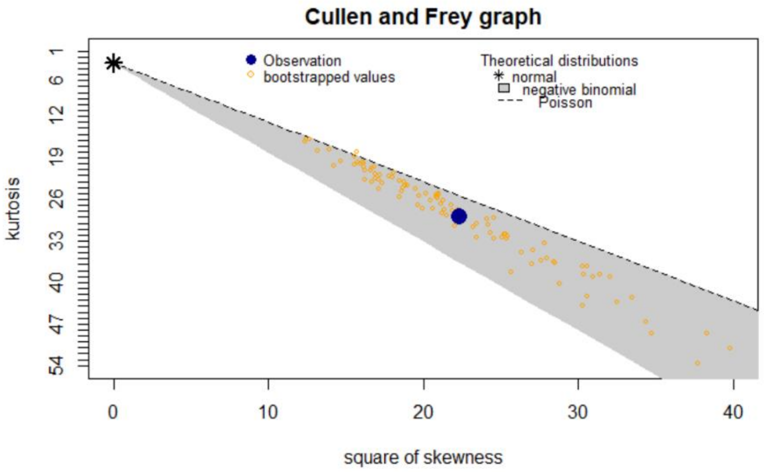
and standard deviation = $\frac{size(1-p)}{p^2}$ = 19.76

Distribution of how many times user sells the token

To observe the behavior of number of times user sells a token, we have derived frequencies for each user and then plotted the distribution of frequencies against user counts. As most of the data belongs to smaller frequencies, we have limited the set to see the behavior of the dataset. For this we have considered the subset of data where user counts<= 100. This results in a set of 99.4% data excluding 2 outliers.

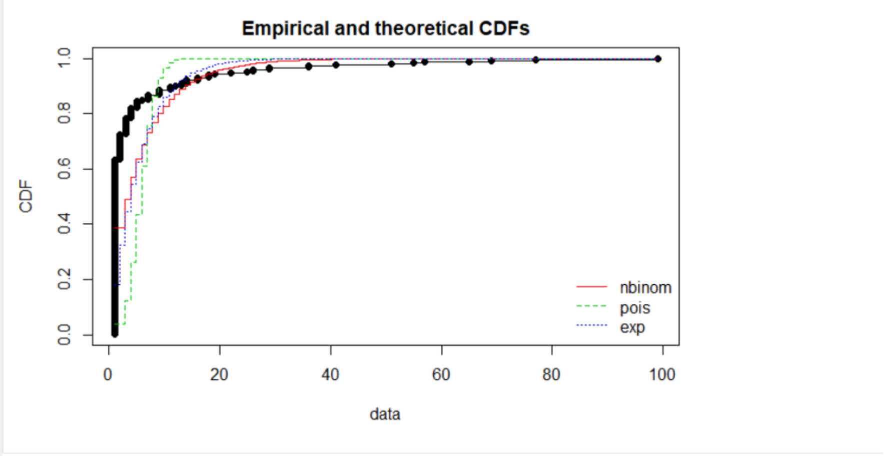


The parameters of the data are as follows:
Mean: 5.0243056
Median: 1
Variance: 139.3269333
Standard Deviation: 11.803683
We have plotted Cullen and Frey graph to see the distributions which fit the data better.

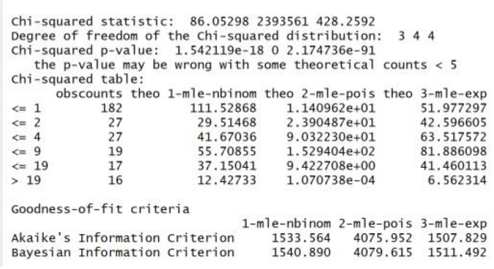


```
summary statistics
------
min:  1   max:  99
median:  1
mean:  5.024306
estimated sd:  11.80368
estimated skewness:  4.722361
estimated kurtosis:  28.85259
```

From above graph, we see that negative binomial and poisson fits better than the normal distribution. However, from the Histogram we can see that there is a possibility for exponential distribution as well. So, we have plotted the CDFs of negative binomial, poisson and normal distributions.



The goodness-of-criteria results are calculated using the gofstat() function and are as follows:



```
Chi-squared statistic:  86.05298 2393561 428.2592
Degree of freedom of the Chi-squared distribution:  3 4 4
Chi-squared p-value:  1.542119e-18 0 2.174736e-91
   the p-value may be wrong with some theoretical counts < 5
Chi-squared table:
      obscounts theo 1-mle-nbinom theo 2-mle-pois theo 3-mle-exp
<= 1        182         111.52868    1.140962e+01      51.977297
<= 2         27          29.51468    2.390487e+01      42.596605
<= 4         27          41.67036    9.032230e+01      63.517572
<= 9         19          55.70855    1.529404e+02      81.886098
<= 19        17          37.15041    9.422708e+00      41.460113
> 19         16          12.42733    1.070738e-04       6.562314

Goodness-of-fit criteria
                               1-mle-nbinom 2-mle-pois 3-mle-exp
Akaike's Information Criterion     1533.564   4075.952  1507.829
Bayesian Information Criterion     1540.890   4079.615  1511.492
```

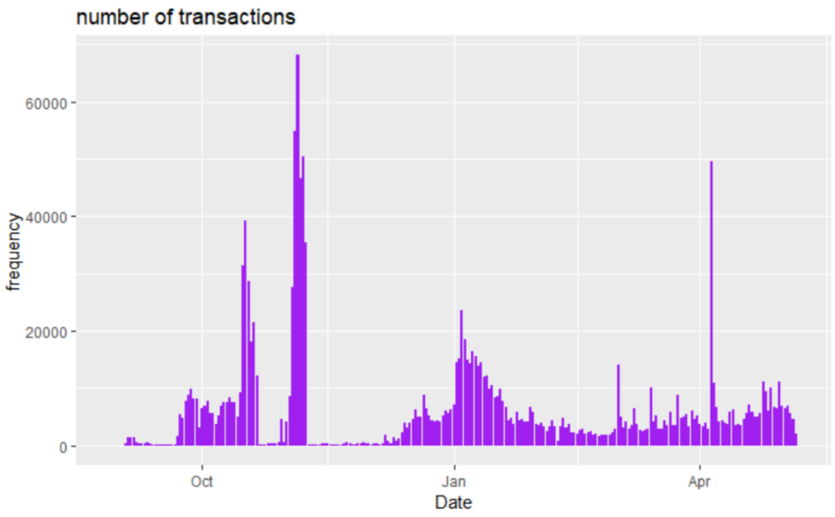
The AIC and BIC count are minimum for the exponential distribution.So, we can say that exponential distribution fits the data better. The mean and standard deviation of the estimated exponential distribution are

Mean = 5.024305

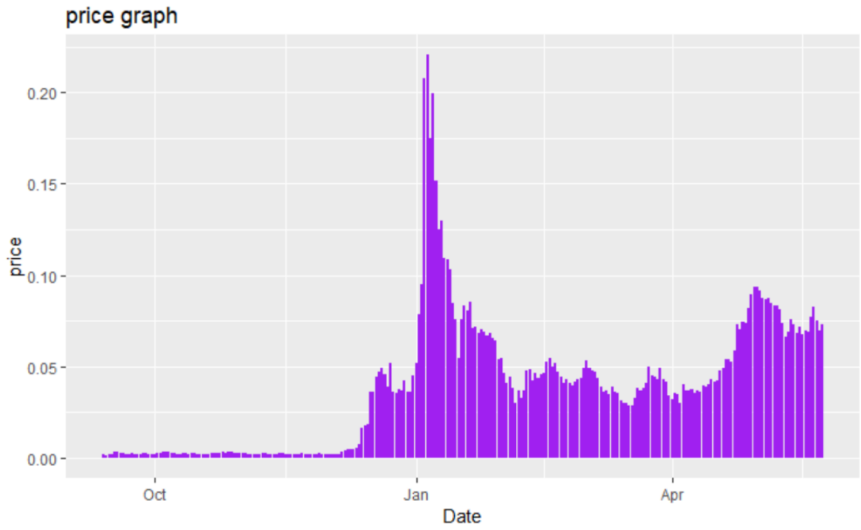
Variance = 5.024305

Correlation of number of users and price on a specific date

The behavior of number of users doing transactions on a particular date can be studied by plotting the graph of number of users against date. For the plot visualization we have used the 'ggplot2' library

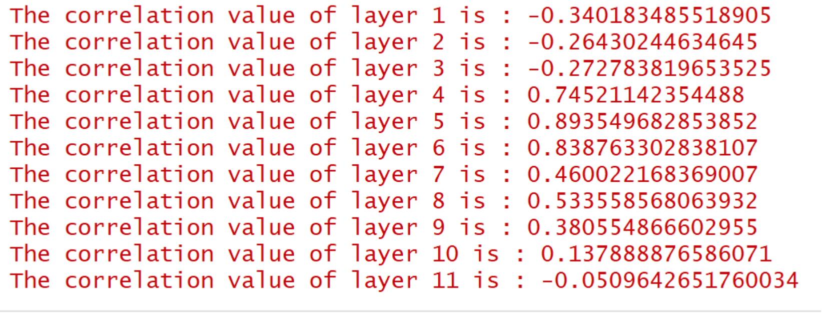


The following is the price trend of the tronix token



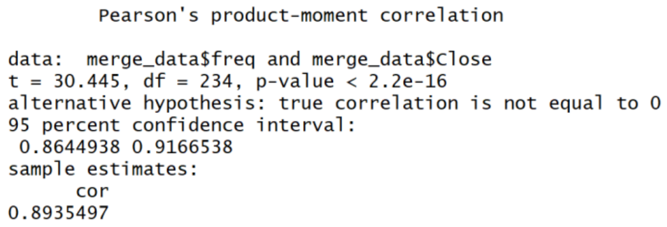
We have grouped the data to get the number of transactions on a particular date and merged the data with the price data and obtained the correlation value of 0.14 for the number of transactions and closing price of a particular date. The correlation value is computed using the cor() function of the 'lubridate' package.

As the correlation value is very less for the complete date, we have investigated by layering the data and obtained the correlation for different layers. We have divided the dataset into 11 layers and computed the correlation value with price data for each of the layer. The values are as follows:



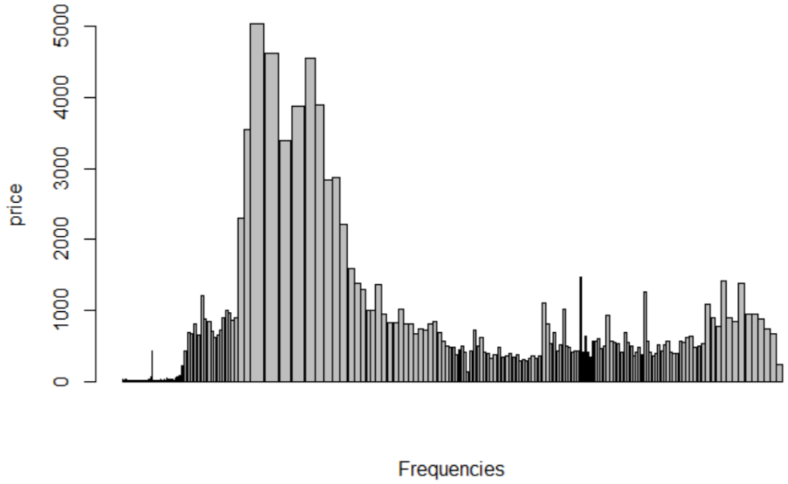
```
The correlation value of layer 1 is : -0.340183485518905
The correlation value of layer 2 is : -0.26430244634645
The correlation value of layer 3 is : -0.272783819653525
The correlation value of layer 4 is : 0.74521142354488
The correlation value of layer 5 is : 0.893549682853852
The correlation value of layer 6 is : 0.838763302838107
The correlation value of layer 7 is : 0.460022168369007
The correlation value of layer 8 is : 0.533558568063932
The correlation value of layer 9 is : 0.380554866602955
The correlation value of layer 10 is : 0.137888876586071
The correlation value of layer 11 is : -0.0509642651760034
```

We see that layer 5 has maximum correlation with value 0.894(approximate). In this layer there are 126504 transactions and the p-value is 2.2e-16 which is very less indicating that correlation value is different from 0. The maximum correlation value is 0.894(approximate) which is higher than the Standard Statistic Correlation co-efficient Value of 0.5. Inspite of the time span of the tokens being 9 months (2017-08-29 to 2018-05-06), due to the transactions being spread out, it gives a relatively high correlation, hence denotes a strong positive Linear relationship between the variables being - the number of users and price on a specific date. Since the relationship is known to be linear, or the observed pattern between the two variables appears to be linear, then the correlation coefficient of 0.894 provides a reliable measure of the strength of the linear relationship. This justifies the number (11) we have chosen to obtain layers, which is to maximize correlation, and have achieved the purpose of layering the data in that pattern. The results of using the Pearson correlation are as follows:



```
        Pearson's product-moment correlation

data:  merge_data$freq and merge_data$Close
t = 30.445, df = 234, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8644938 0.9166538
sample estimates:
      cor
0.8935497
```

The following plot of number of transactions and price follows normal distribution and the mean and standard deviation parameters obtained from fitting the distribution are same as sample. So, the choice of “Pearson” test for the correlation is also justified and there is not a very notable deviation from the other method “Spearman”



Multiple Regression Model

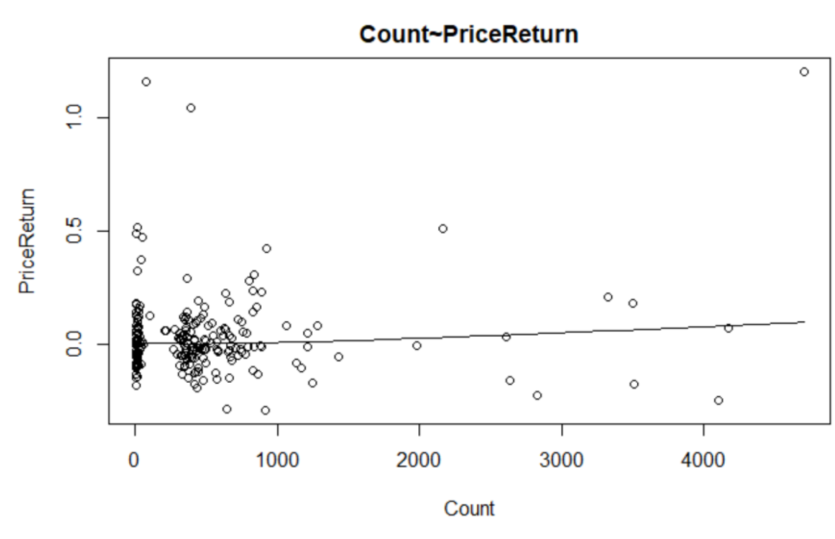
The simple price return is defined as $(P_t - P_{t-1})/P_{t-1}$ where P_t is the token price for day t. After careful analysis we have considered a layer which contains (126504) transactions and build the model on it. We have considered the following three features which gives maximum correlation with the price return compared to others for our data.

1. The number of unique users selling or buying the tokens on the previous day.
2. The Market Cap Value of the previous day/High price of the previous day.
3. The growth of the price on the previous day.

However, only the Market Cap Value/ High Price have the significant correlation value with the data.

Unique users selling or buying the tokens on the previous day

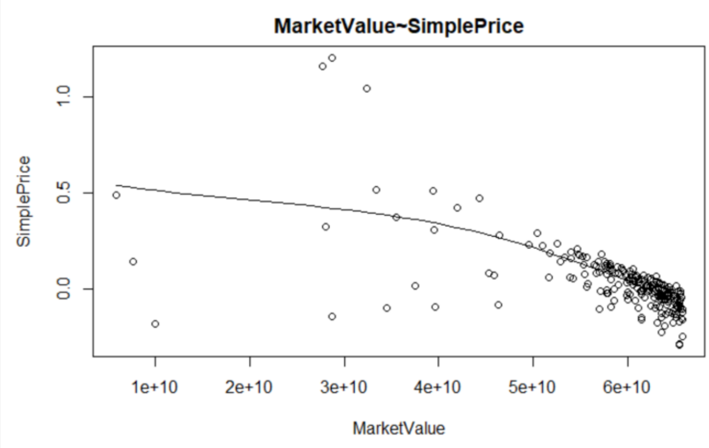
We have plotted the graph between the number of users buying or selling the tokens on the previous day and the price return value and the plot is shown below:



The correlation value is calculated using the cor() function and the value is 0.125.

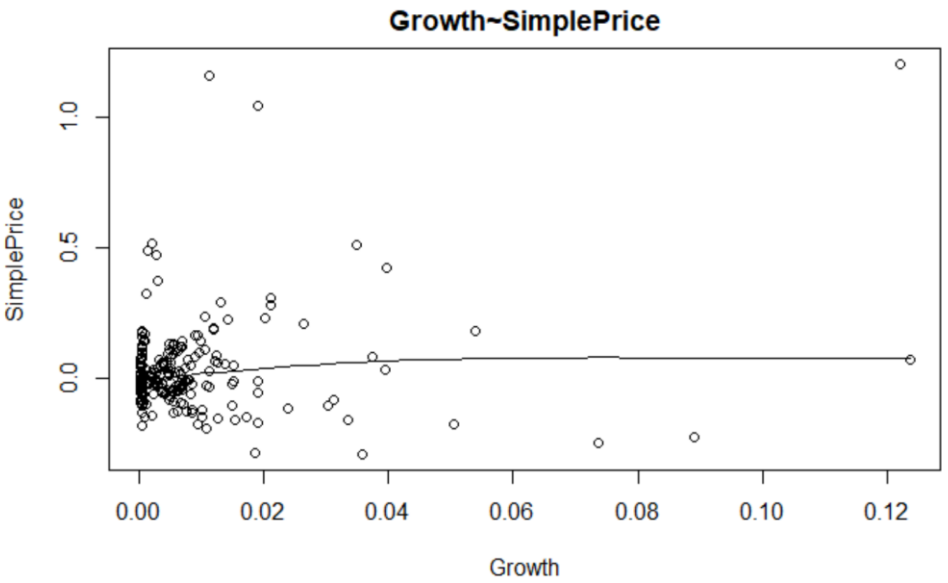
Market Value of the previous day

The following is the plot of the market value of the token on previous day and the simple price return.



The correlation value is calculated using the cor() function and the value is -0.605.

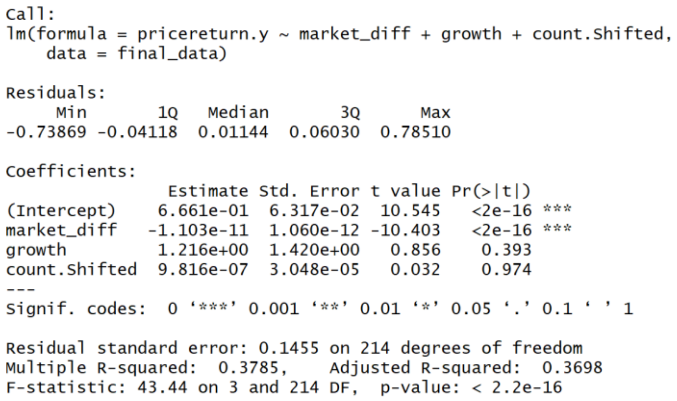
Growth of the price on the previous day



The correlation value between simple price and growth is 0.212

Modelling the multiple regression

The multiple regression is modelled using lm() function in R. The summary of the linear model is as follows:

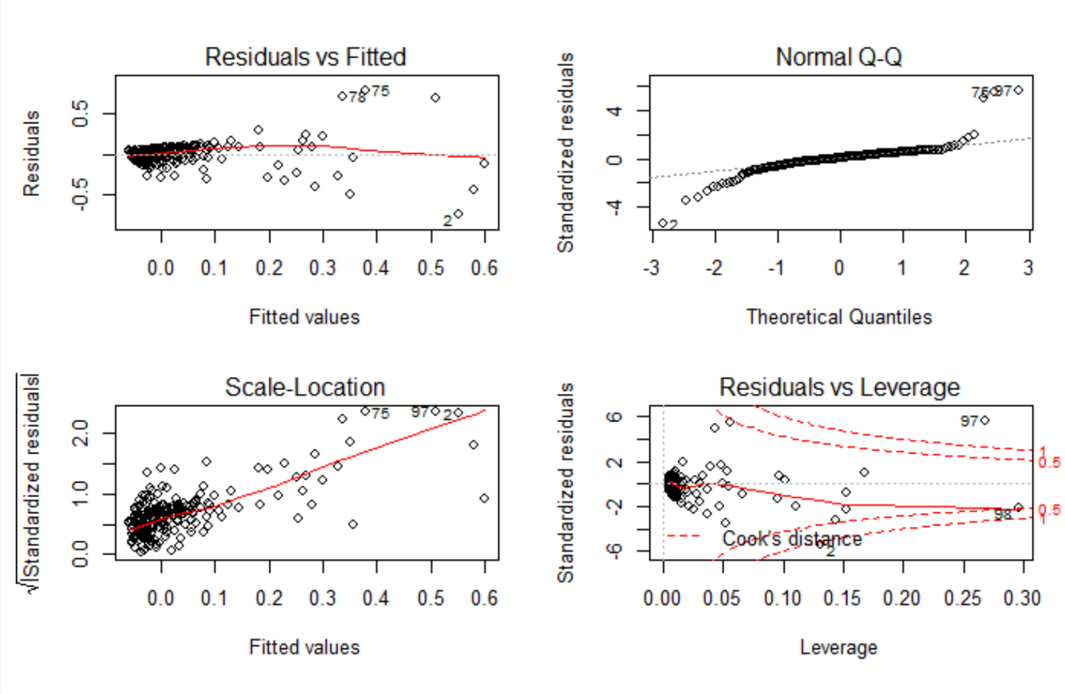


```
Call:
lm(formula = pricereturn.y ~ market_diff + growth + count.Shifted,
    data = final_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.73869 -0.04118  0.01144  0.06030  0.78510

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    6.661e-01  6.317e-02  10.545   <2e-16 ***
market_diff   -1.103e-11  1.060e-12 -10.403   <2e-16 ***
growth         1.216e+00  1.420e+00   0.856    0.393
count.Shifted  9.816e-07  3.048e-05   0.032    0.974
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1455 on 214 degrees of freedom
Multiple R-squared:  0.3785,	Adjusted R-squared:  0.3698
F-statistic: 43.44 on 3 and 214 DF,  p-value: < 2.2e-16
```

From the P values we can see that market_diff is significant. The null hypotheses is that the coefficients associated with the variables is equal to zero. The alternate hypothesis is that the coefficients are not equal to zero i.e, there exists a relationship between the independent variable in question and the dependent variable. The larger t value indicates that it is less likely that the coefficient is not equal to zero purely by chance. So, higher the t-value, the better. *Pr(>|t|)* or *p-value* is the probability that you get a t-value as high or higher than the observed value when the Null Hypothesis (the *β* coefficient is equal to zero or that there is no relationship) is true. So if the *Pr(>|t|)* is low, the coefficients are significant (significantly different from zero). If the *Pr(>|t|)* is high, the coefficients are not significant. When p Value is less than significance level (< 0.05), we can safely reject the null hypothesis that the co-efficient *β* of the predictor is zero. In our model, market_diff and count features have values < 0.05.

The R-squared value of the model is 0.379 and the adjusted R squared value is 0.369. These values can be greater for a better model. However, we can see that only two features are contributing to the model, the values are less. The Akaike's information criterion(AIC) and Bayesian information criterion(BIC) values are -215.944 and -199.039 respectively for the model. The model with lower values is preferred.

The following residual plots are plotted using the plot() function.



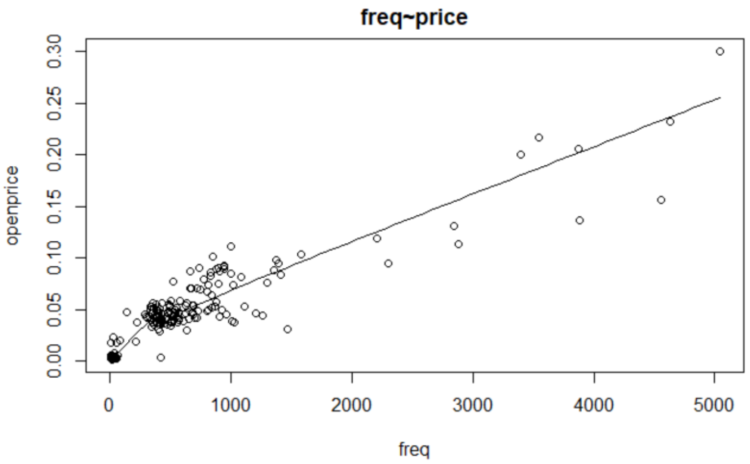
Multiple Regression Model For Open Price

We have further analyzed for the features which can be used to predict the open price instead of simple price return. The following features have maximum correlation with the open price

1. Number of transactions on a previous date.
2. Number of unique users/buyers doing transactions on a previous date.
3. Volume of the token on previous day

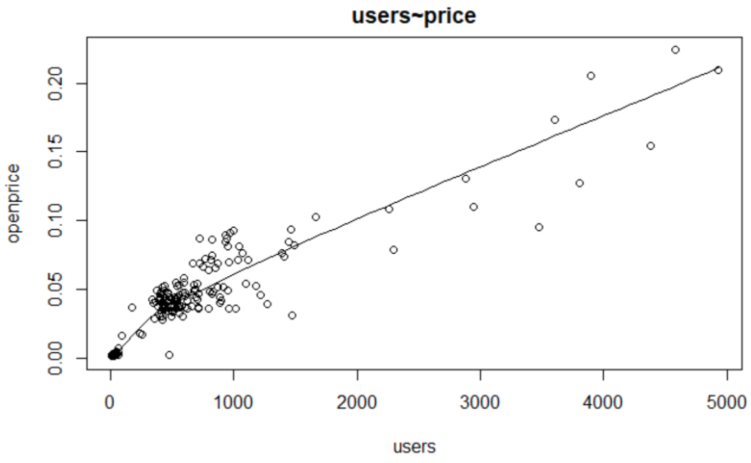
Number of transactions on a previous date

The following is the plot showing the relationship between number of transactions on a previous date and opening price. These two features have a correlation of 0.9



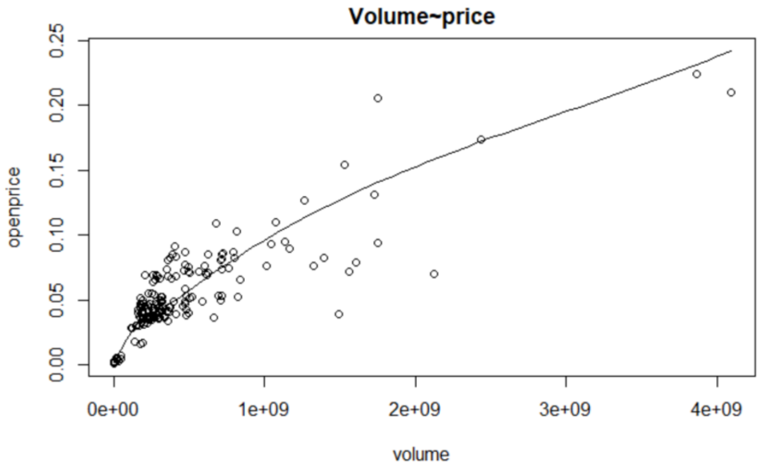
Number of unique buyers/sellers on a previous date

We have plotted the graph between the number of unique users and buyers on previous date and the correlation value obtained for them is 0.912



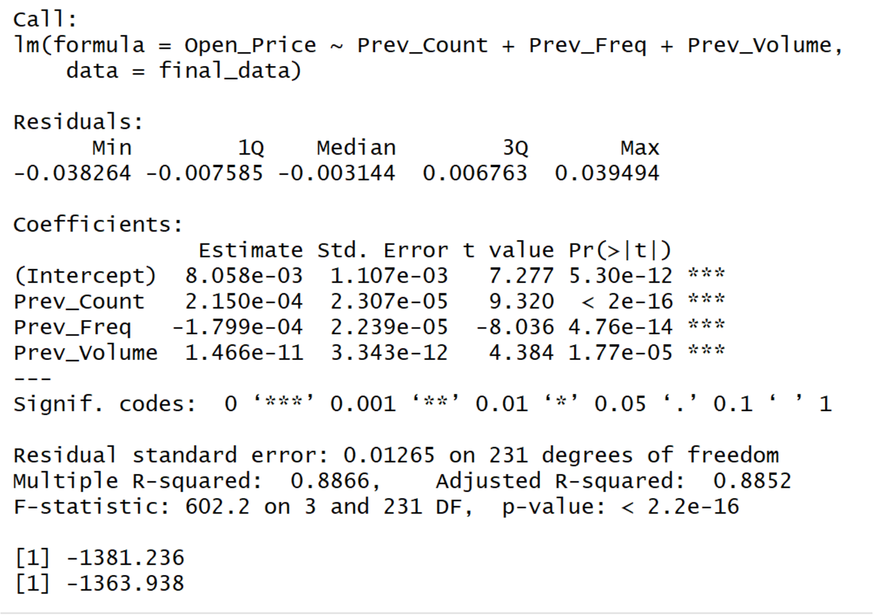
Volume of the token on previous date

The plot shows the relationship between volume of the token on previous date and the opening price. The correlation value of these two variables is 0.85



Modelling the multiple linear regression

The multiple regression is modelled using lm() function in R. The summary of the linear model is as follows:

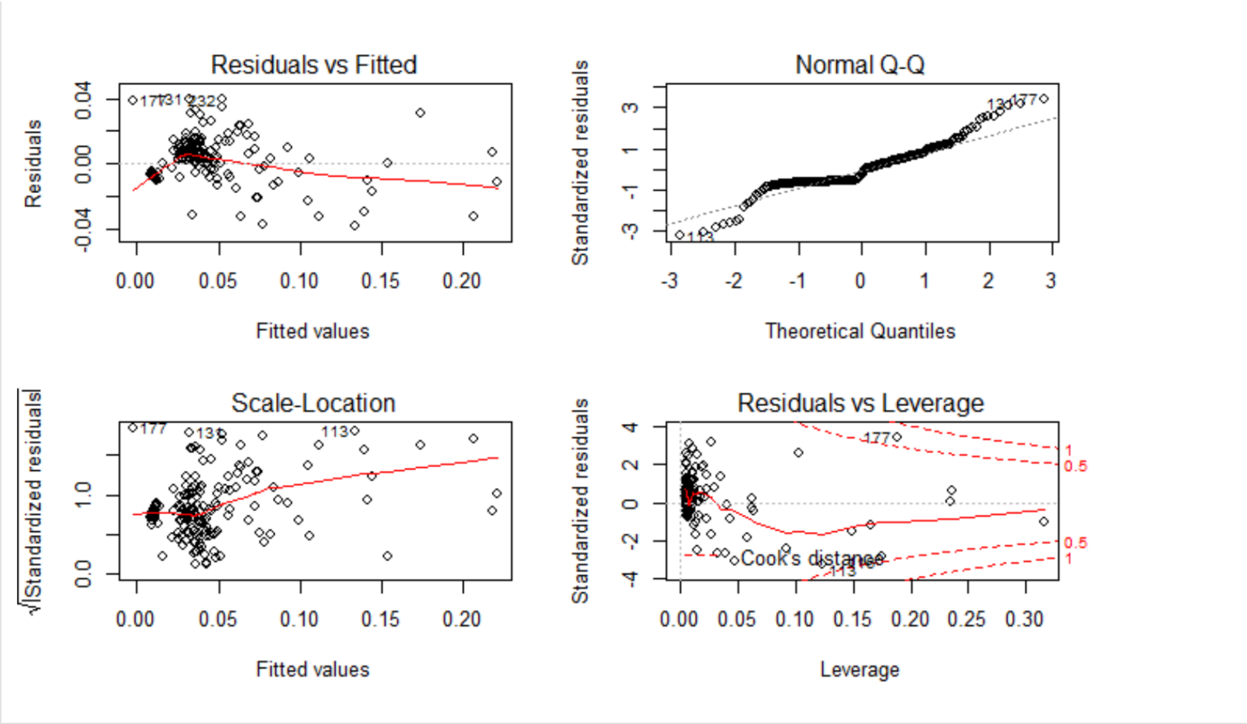


```
Call:
lm(formula = Open_Price ~ Prev_Count + Prev_Freq + Prev_Volume, 
    data = final_data)

Residuals:
      Min        1Q    Median        3Q       Max 
-0.038264 -0.007585 -0.003144  0.006763  0.039494 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  8.058e-03  1.107e-03   7.277 5.30e-12 ***
Prev_Count   2.150e-04  2.307e-05   9.320  < 2e-16 ***
Prev_Freq   -1.799e-04  2.239e-05  -8.036 4.76e-14 ***
Prev_Volume  1.466e-11  3.343e-12   4.384 1.77e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01265 on 231 degrees of freedom
Multiple R-squared:  0.8866,	Adjusted R-squared:  0.8852 
F-statistic: 602.2 on 3 and 231 DF,  p-value: < 2.2e-16

[1] -1381.236
[1] -1363.938
```

In contrast to the previous model, we can see that all the features in this model are significant with high |t| values and less pr(>|t|) values. The R-squared value is 0.8866 and Adjusted R squared value is 0.8852 which are significant.

The AIC and BIC values of this model are also very less. Therefore, this model can be used in predicting the open price.

The following are the residual plots for this model



Conclusion

Based on the data analysis performed, we found that distribution of number of user buys follows negative binomial distribution and number of times user sells follows exponential distribution. From the layers of the data created, we have found that in layer5 the number of users have strong correlation with price data. The value of 0.89 shows the high correlation value, where the value usually ranges from -1 to 1. We have created a multiple linear regression model on the layer which has 126504 transactions using the features which has maximum correlation with the simple price return and analyzed the results. We also created multiple linear regression model on the same layer for opening price and compared both the models.

References

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

https://blockgeeks.com/guides/what-is-ethereum/

https://blog.coindirect.com/coin-profile-tron-tronix-trx/

https://developers.tron.network/docs/getting-started

https://www.inverse.com/article/39961-tron-trx-cryptocurrency-ripple-bitcoin

https://blockonomi.com/tron-trx-guide/