ETHEREUM PROJECT - BAT TOKEN

Praveen Ramani, Pawan Dasharath Patil, Navaneetha Krishnan Muralidaran
30 November 2018

ETHEREUM

- Ethereum is an open-source, public, blockchain-based distributed computing platform and operating system featuring smart contract (scripting) functionality.(https://en.wikipedia.org/wiki/Ethereum)
- It enables applications (DAPPs) to be built on blockchain technology.
- Blockchain is used to determine the ownership of an item without using a third-party.
- The data is stored in a distributed ledger which provides consensus without a central authority.

Ether

Ether is the cryptocurrency that runs on the ethereum platform. To run applications or view resources on the ethereum platform, a certain amount of fees should be paid. The payment is done using "gas" which is a subunit of ether. Ether can also be transferred to other individuals.

ETC-20

- ERC stands for Ethereum Request for Comment and this request is assigned the number 20.
- It defines a set of rules that an Ethereum token must follow.
- These rules define how the token data can be viewed and how tokens can be tranferred.

Tokens

Ethereum tokens are used to represent digital assets on the ethereum platform.

Two Types

Useage Tokens

These act like currency for their corresponding DAPP. e.g Golem

Work Tokens

These act like stocks of the DAPP.

Primary Token

The primary token of our project is networkbatTX.

BAT

- BAT stands for Basic Attention Token.
- This token is used to obtain advertising services.
- The publisher of an advertisement gets paid in BATs based on the amount of user attention it gets.

Primary goal

The primary goal of this project is to 1. find the distribution of how many times a user buys and sells the token. 2. form layers on the number of transactions and find the correlation between the number of transactions and price of the token, for each layer. 3. track the activity the most active buyers and sellers of BAT in other tokens.

Eliminating Outliers of BatchOverflow Vulnerability

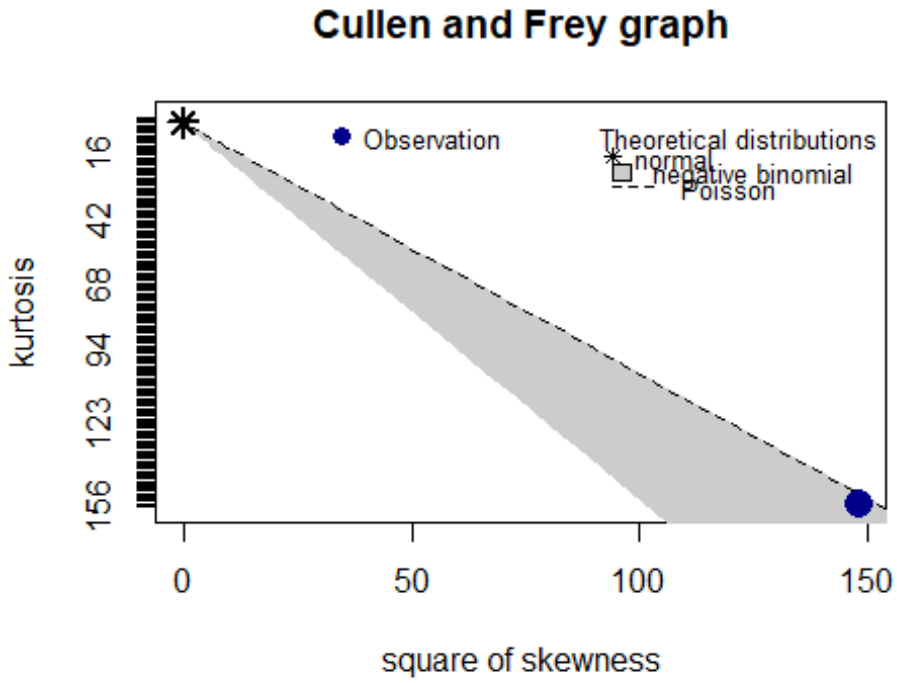
- Batchoverflow vulnerability enabled hackers to create counterfeit tokens by causing an integer overflow.
- Therefore, we should eliminate such transactions from the token data.
- Total supply: 1.5e+9
- Digits: e+18
- Therefore, we must eliminate transactions with amounts more than 1.5e+27 Total number of records were 327854 and after eliminating 6 outliers we have total of 327848 records.

Question 1

Buyer Data

We have taken the number of buys and their frequencies.

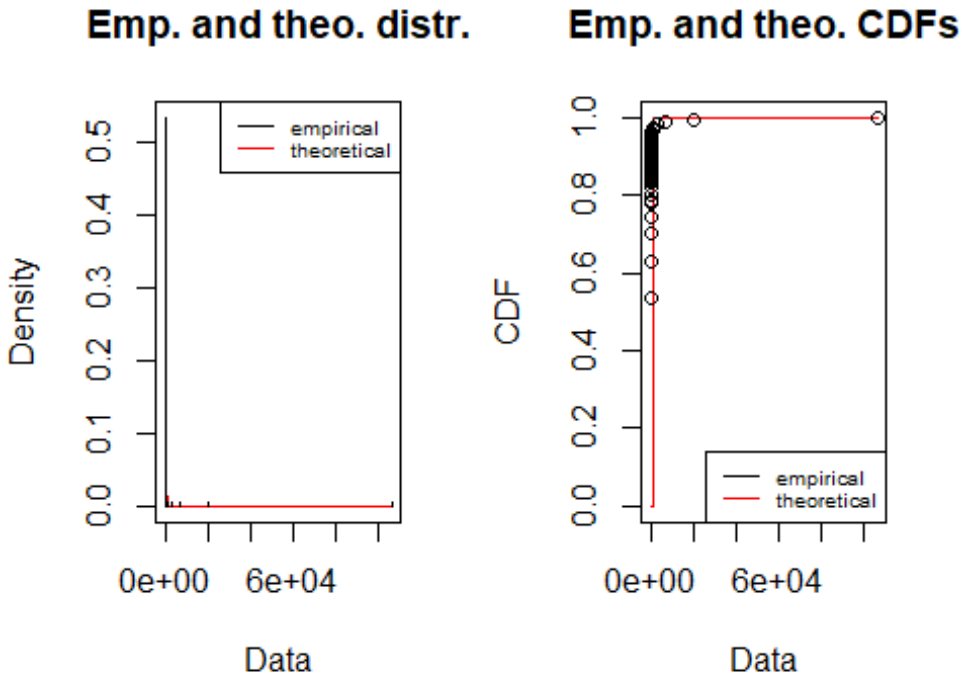
Estimating as a discrete distribution.



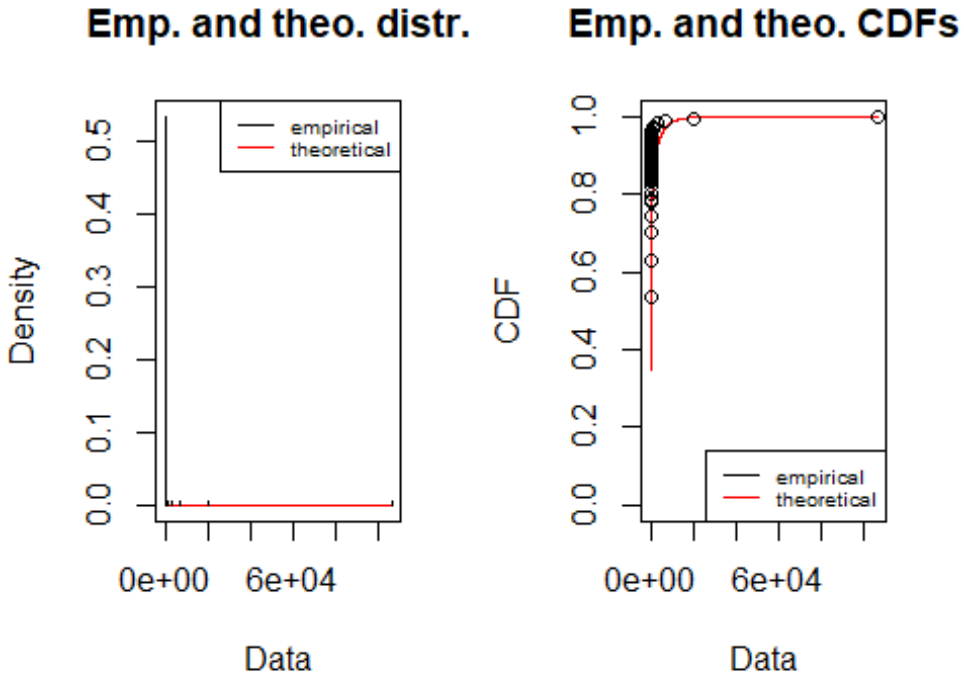

```
## summary statistics
## ------
## min:  1   max:  106056
## median:  1
## mean:  858.1152
## estimated sd:  8409.252
## estimated skewness:  12.16062
## estimated kurtosis:  155.0217
```

On observation from the Cullen and Frey Graph the distribution is closer to poisson distribution than negative binomial distribution.

Fitting with Poisson distribution.



Fitting with Negative binomial distribution.

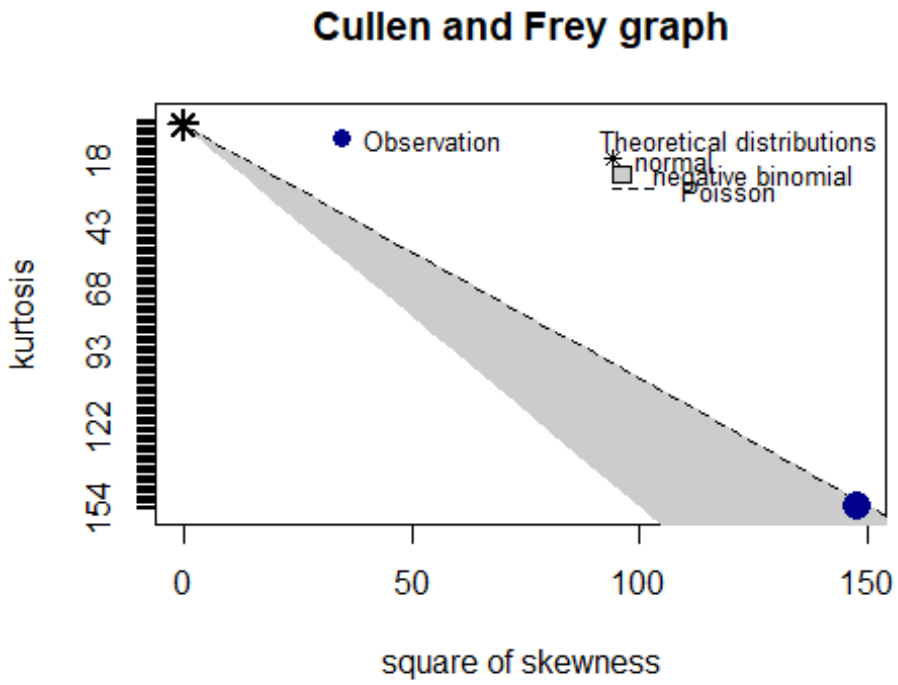


Comparing the Emp. and theo. CDFs graphs of the two distributions, we can say that the dataset fits better with poisson distribution when compared to negative binomial distribution.

Seller data

We have taken the number of sells and their frequencies.

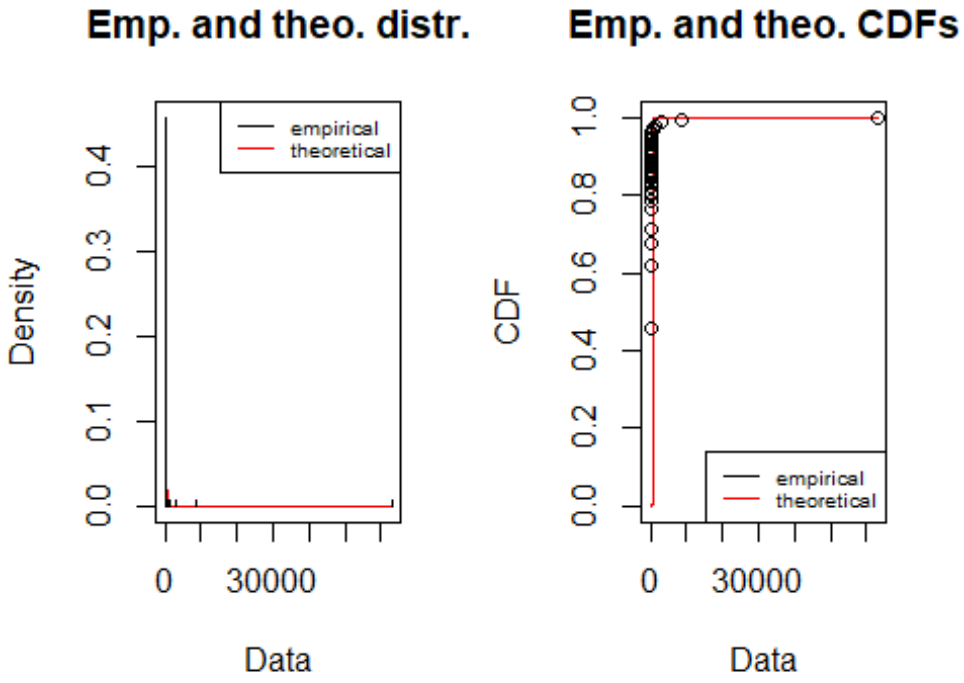
Estimating as a discrete distribution.



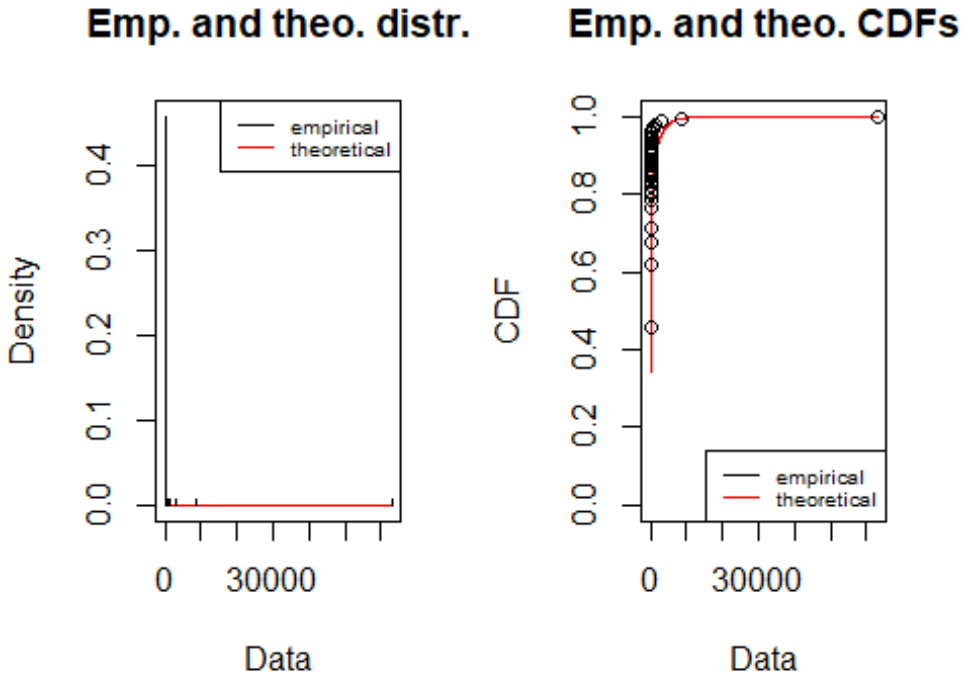

```
## summary statistics
## ------
## min:  1   max:  63039
## median:  2
## mean:  506.2675
## estimated sd:  5077.273
## estimated skewness:  12.1571
## estimated kurtosis:  153.2665
```

The observation is found to be closer to poisson distribution than negative binomial distribution

Fitting with poisson distribution



Fitting with negative binomial distribution



Comparing the Emp. and theo. CDFs graphs of the two distributions, we can say that the dataset fits better with poisson distribution when compared to negative binomial distribution.

Question 2

Layers

We create 4 layers based on the quartile values of the 'amount' column.

Layer1:

Date and number of transactions on that day with amounts between min and q1.

Layer2:

Date and number of transactions on that day with amounts between q1 and q2(median).

Layer3:

Date and number of transactions on that day with amounts between q2 and q3.

Layer4:

Date and number of transactions on that day with amounts between q3 and max.

Price Values:

We take the average of high and low values of price from the price graph, as that days's price measure.

Correlation values for each layers

We find the correlation between number of transactions and price for each layer.

```
cor(cor_lay1$count,cor_lay1$avg, method = "pearson")
```

```
## [1] 0.4264666
```

```
cor(cor_lay2$count,cor_lay2$avg, method = "pearson")
```

```
## [1] 0.4612127
```

```
cor(cor_lay3$count,cor_lay3$avg, method = "pearson")
```

```
## [1] 0.5064532
```

```
cor(cor_lay4$count,cor_lay4$avg, method = "pearson")
```

```
## [1] 0.3717038
```

Correlation for entire Dataset

```
## [1] 0.505124
```

The correlation seems to be around 0.5. We can say that this token has a moderate positive correlation.

Question 3:

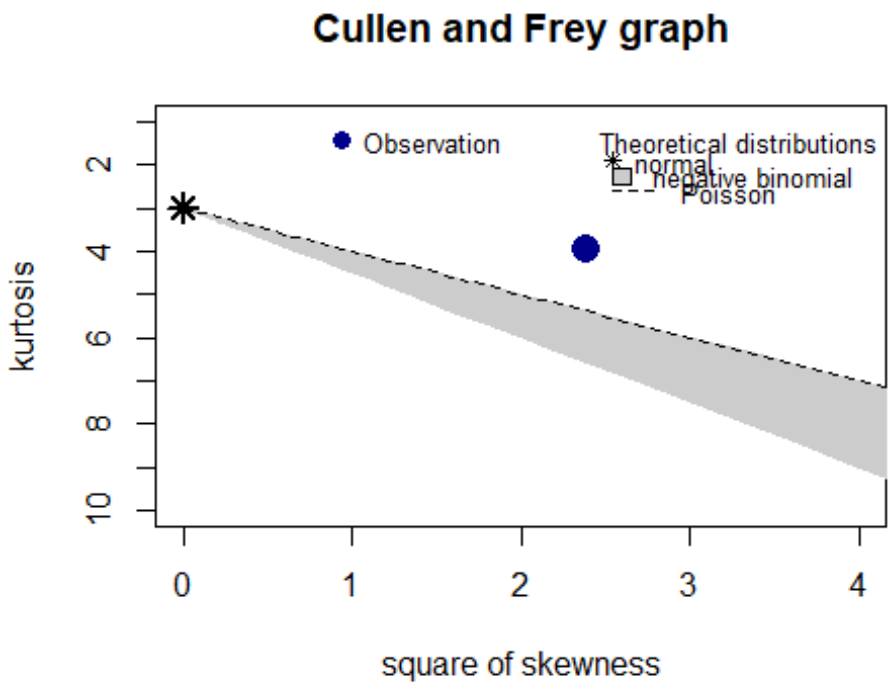
For Buyer

Acc.no of top investors and unique tokens they invest in

```
##     buyer uniq_token_count
## 1  351141                1
## 2      17               12
## 3      49               12
## 4  351248                1
## 5       5               14
## 6  296381               10
## 7      82               12
## 8      44               12
## 9       6               12
## 10  48315               14
```

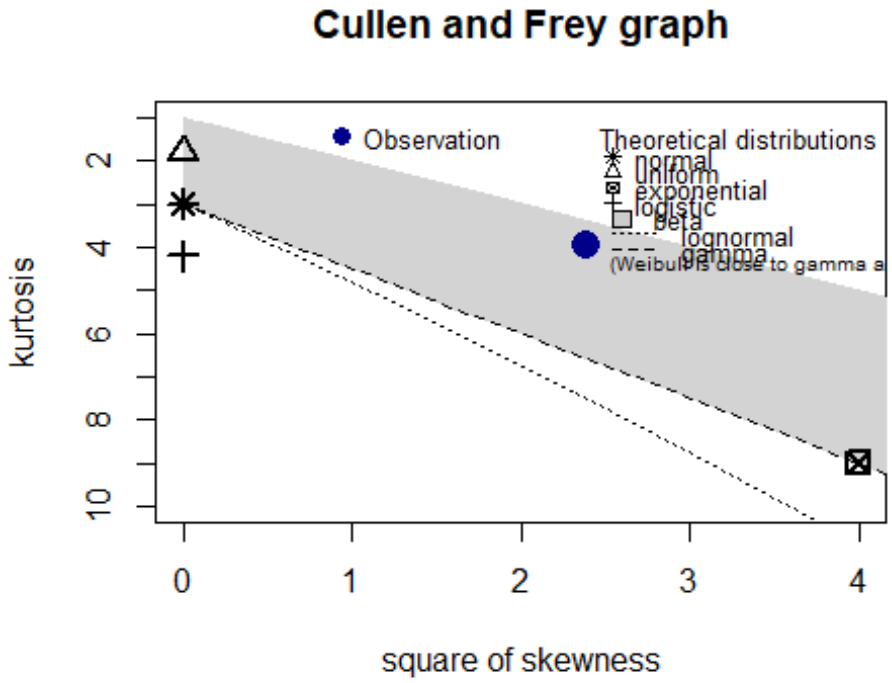
Fitting Distributions

Discrete Distribution




```
## summary statistics
## ------
## min:  1   max:  14
## median:  12
## mean:  10
## estimated sd:  4.876246
## estimated skewness:  -1.545256
## estimated kurtosis:  3.954712
```

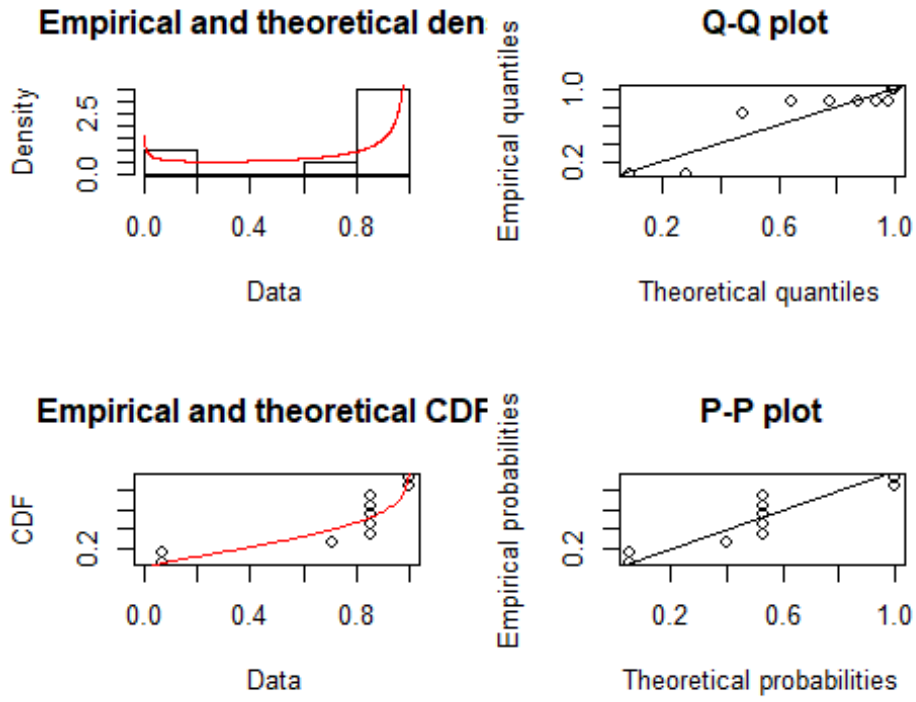
Continuous Distribution




```
## summary statistics
## ------
## min:  1   max:  14
## median:  12
## mean:  10
## estimated sd:  4.876246
## estimated skewness:  -1.545256
## estimated kurtosis:  3.954712

## [1] 14

## Warning in fitdist(temp, distr = "beta", method = "mge"): maximum
GOF
## estimation has a default 'gof' argument set to 'CvM'
```



The observation point is no where near any of the discrete distributions and is closer to the beta (continuous) distribution. So we tried to fit the data in beta distribution.

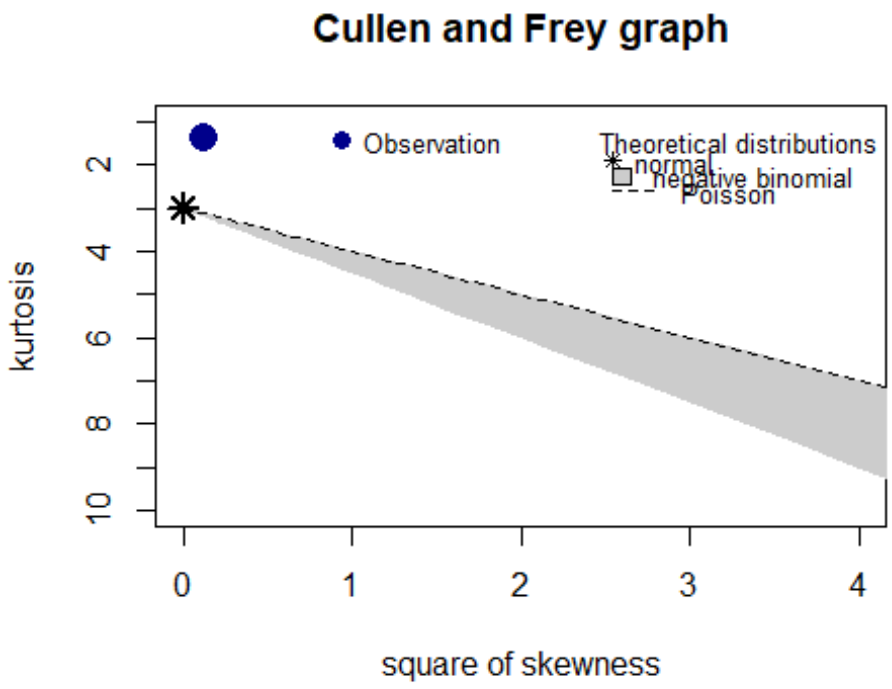
For seller

Acc.no of top investors and unique tokens they invest in

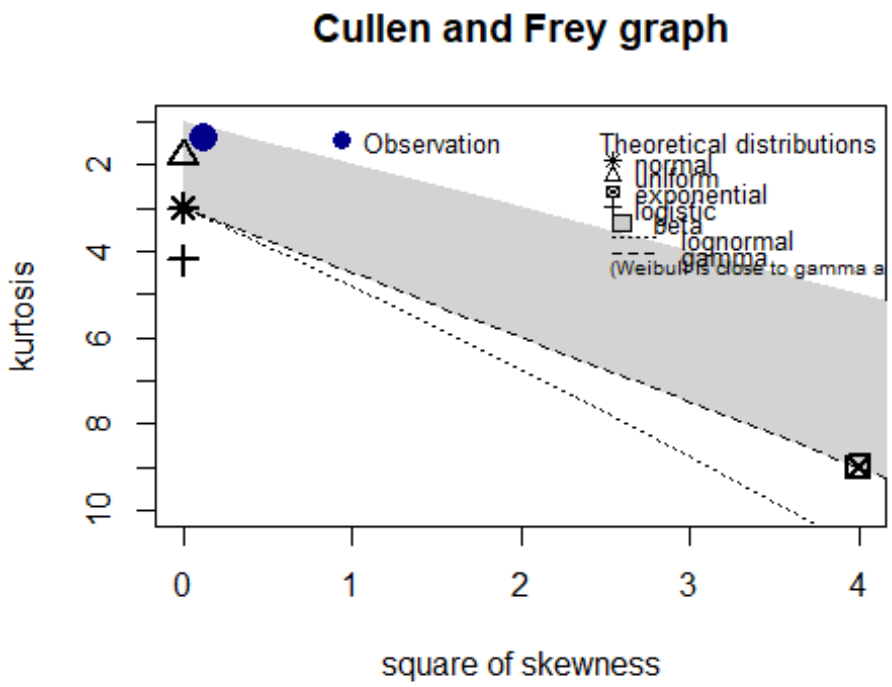
```
##    seller uniq_token_count
## 1  351141                1
## 2       5               15
## 3  351248                1
## 4      49               12
## 5  296381               10
## 6   75994               11
## 7  352681                1
## 8  310645                1
## 9  142341                7
## 10 352820                4
```

Fitting Distributions

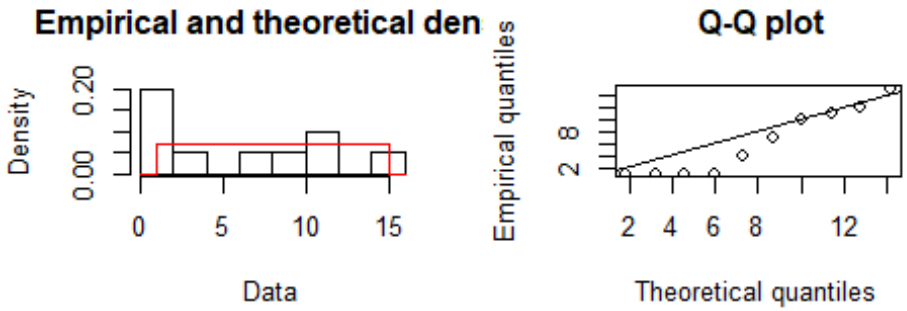
Discrete Distribution

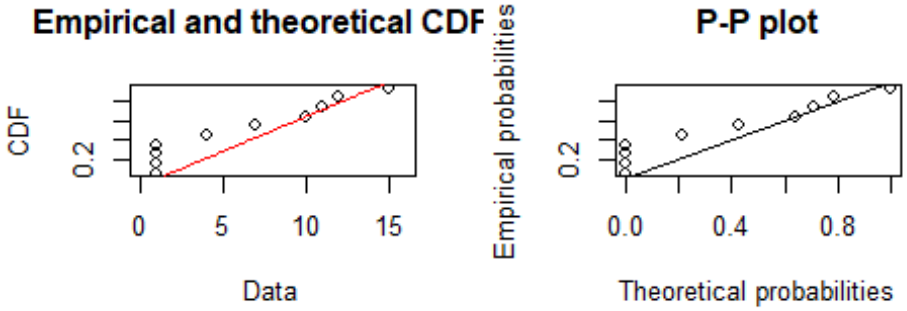



```
## summary statistics
## ------
## min:  1   max:  15
## median:  5.5
## mean:  6.3
## estimated sd:  5.396501
## estimated skewness:  0.3454059
## estimated kurtosis:  1.399975
```

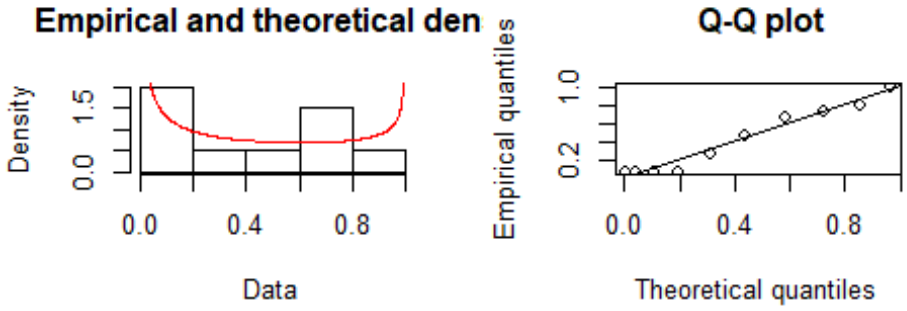


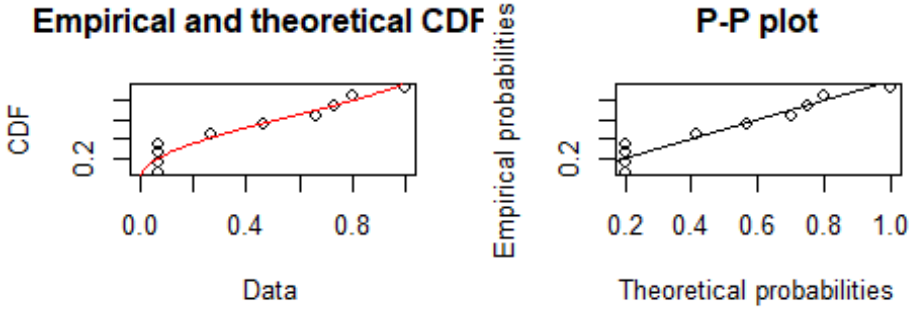

```
## summary statistics
## ------
## min:  1   max:  15
## median:  5.5
## mean:  6.3
## estimated sd:  5.396501
## estimated skewness:  0.3454059
## estimated kurtosis:  1.399975
```






```
## Warning in fitdist(temp, distr = "beta", method = "mge"): maximum
GOF
## estimation has a default 'gof' argument set to 'CvM'
```





The observation point is no where near any of the discrete distribution and is closer to the beta and uniform continuous distribution. So we tried to fit the data in both beta and uniform distribution.

Comparing the P-P plots we can see that beta distribution fits the data better than uniform distribution.

Conclusions

- The number of transactions people do follows an approximate poisson distribution.
- BAT has a moderate positive correlation with price.
- The most active buyers and sellers of BAT are also quite active in other tokens.
- The number of unique tokens that the most active traders invest in is approximately a beta distribution.

Project 2

Candidates

[1] We have taken the number of transactions of the last 3 days as 3 features, to predict the day's average price.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = data2b)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.31543 -0.09130 -0.03264  0.05827  0.46680
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.637e-01  1.101e-02  14.866  < 2e-16 ***
## x1          7.204e-05  2.320e-05   3.105  0.00206 **
## x2          3.875e-05  2.832e-05   1.368  0.17211
## x3          1.049e-04  2.320e-05   4.523 8.49e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1259 on 333 degrees of freedom
## Multiple R-squared:  0.3065, Adjusted R-squared:  0.3003
## F-statistic: 49.07 on 3 and 333 DF,  p-value: < 2.2e-16
```

[2] We have taken the number of transactions of the past 3 days and the price of the previous day, as 4 features, to predict the day's average price.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4, data = data2c)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.116366 -0.012229 -0.000541  0.009883  0.146476
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.990e-03  3.119e-03   1.600   0.1105
## x1          -1.299e-05  5.214e-06  -2.492   0.0132 *
## x2          -1.055e-05  6.264e-06  -1.685   0.0929 .
## x3           2.475e-05  5.202e-06   4.758 2.92e-06 ***
## x4           9.820e-01  1.214e-02  80.865  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02771 on 332 degrees of freedom
## Multiple R-squared:  0.9665, Adjusted R-squared:  0.9661
## F-statistic:  2394 on 4 and 332 DF,  p-value: < 2.2e-16
```

The r square value is very high. This is because the correlation between average price of two consecutive days is very high as shown below.

```
## [1] 0.9816524
```

[3] We have taken the number of transactions of the past 3 days, the percentage change in price and percentage change in number of transactions of the past 2 days, as 5 features, to predict the day's average price.

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data2d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.32497 -0.09011 -0.02665  0.05538  0.43912
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.624e-01  1.102e-02  14.736  < 2e-16 ***
## x1           7.696e-05  2.327e-05   3.308  0.00104 **
## x2           3.482e-05  4.484e-05   0.776  0.43803
## x3           1.054e-04  4.685e-05   2.250  0.02511 *
## x4           1.588e-01  8.019e-02   1.981  0.04845 *
## x5          -5.834e-03  2.445e-02  -0.239  0.81155
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.1255 on 331 degrees of freedom
## Multiple R-squared:  0.3147, Adjusted R-squared:  0.3043
## F-statistic:  30.4 on 5 and 331 DF,  p-value: < 2.2e-16
```

[4] We have taken the number of transactions of the past 3 days, difference in price and difference in the number of transactions of the past 2 days, as 5 features, to predict the day's average price.

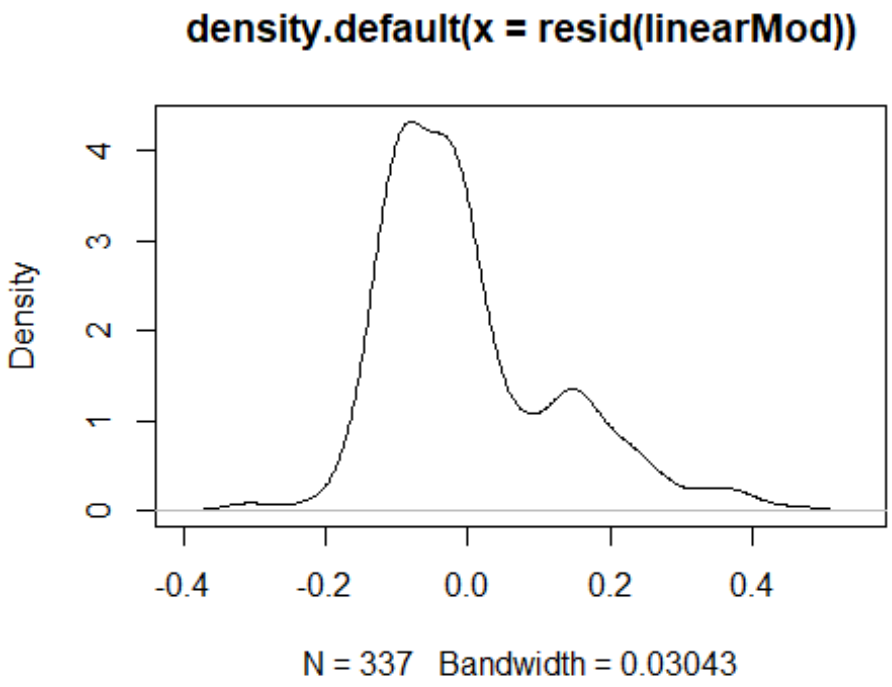
```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5, data = data2da)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.30848 -0.08953 -0.02662  0.05556  0.45722
##
## Coefficients: (1 not defined because of singularities)
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.624e-01  1.093e-02  14.854  < 2e-16 ***
## x1          8.128e-05  2.328e-05   3.491 0.000546 ***
## x2          4.603e-05  2.823e-05   1.631 0.103896
## x3          9.014e-05  2.371e-05   3.802 0.000171 ***
## x4                 NA         NA      NA       NA
## x5          6.348e-01  2.464e-01   2.576 0.010417 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1248 on 332 degrees of freedom
## Multiple R-squared:  0.3201, Adjusted R-squared:  0.3119
## F-statistic: 39.08 on 4 and 332 DF,  p-value: < 2.2e-16
```

Comparison

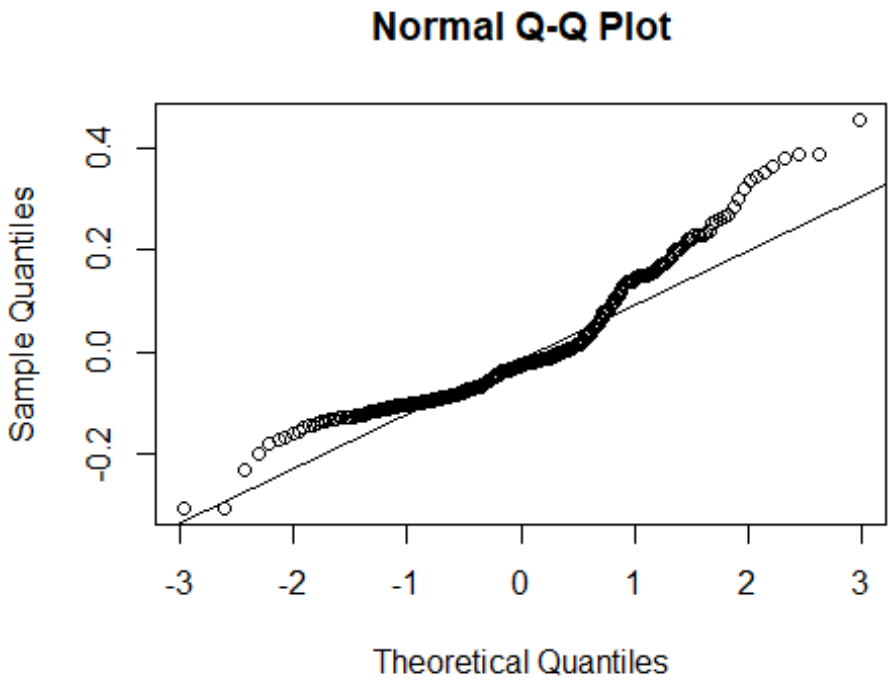
Comparing the r square value of the 4 candidates, the best adjusted r square value is 0.3119 and is got from [4].

Residuals For [4]

```
plot(density(resid(linearMod)))
```




```
qqnorm(resid(linearMod))
qqline(resid(linearMod))
```



References

[1] http://r-statistics.co/Linear-Regression.html

[2] https://stats.stackexchange.com/questions/53254/how-to-find-residuals-and-plot-them

[3] https://www.r-bloggers.com/make-r-speak-sql-with-sqldf/

[4] https://stats.stackexchange.com/questions/236118/fitting-distribution-for-data-in-r

Full Project

https://github.com/PravGitHub/Stats_Project