CS 6313.002 Statistical Methods of Data Science - Final Project

Shweta Rakesh Sahalot (sxs180145), Siddhi Chechani (sxc170042)

December 3, 2018

## Abstract

The aim of the project is to implement different statistical methods on the TENXPAY Token transaction data.The results of the methods are analyzed to derive inferences about the token data.In 4.1, we found how many times a user buys or sells the TENXPAY token, then fit a distribution and estimate the best distribution.In 4.2, we find the correlation between the TENXPAY Token price data and layer features of the TENXPAY Token.In 4.3, we find the most active buyer and seller of our TENXPAY token and then track them in all other tokens, plot how many unique tokens have they invested in the provided time frame then fit and estimate distributions as part 1. In 4.4, we created a multiple linear regression model to explain price return on day t.

**source code https://github.com/shwetasahalot95/6313\_project**

## Introduction

### 1.1 Key Concepts:

Let us first understand a few concepts, before we get into the detailing of the project

#### 1.1.1 Blockchain:

It is a public ledger formed of multiple blocks.Each block contains crytographic hash of the previous block, transaction data, timestamp and other metadata.The blockchain is a chain of these blocks linked together cryptographically and stored on a distributed peer-to-peer network.

#### 1.1.2 Ethereum

Ethereum is adecentralized platform that runs smart contracts: applications that run exactly as programmed without any possibility of downtime, censorship, fraud or third-party interference.

These apps run on a custom builtblockchain, an enormously powerful shared global infrastructure that can move value around and represent the ownership of property.

Smart Contract is a set of instructions of the format if-this-then-that, it is formed when someone needs to performing particular task involving one or more than one entities of the blockchain. The contract is a code which executes itself on occurence of a triggering event such as expiration date.The smart contracts can be written with different languages such as solidity

The EVM is a runtime environment for smart contracts in Ethereum. Every Ethereum node in the network runs an EVM implementation and executes the same instructions.

Ether is the digital currency(cryptocurrency) of Ethereum. Every individual transaction or step in a contract requires some computation. To perform any computation user has to pay a cost calulated in terms of ‘Gas’ and paid in terms of ‘Ether’. The Gas consist of two parts:

Gas limit: It is the amount of units of gas Gas price: It is the amount paid per unit of gas

#### 1.1.3 Ethereum ERC-20 Tokens:

If a user needs some service provided by the DAPPS, then he has to pay for that service in terms of ‘token’ associated with the DAPPS. These Ethereum tokens can be bought using Ether or other cryptocurrencies and can serve the following two purposes:

1. Usage Token: These tokens are used to pay for the services of the Dapp
2. Work Token: These tokens identify you as a shareholder in the DAPP

ERC-20 is a technical standard used for smart contracts on the Ethereum Blockchain for implementing tokens. It is a common list of rules for Ethereum token regarding interactions between tokens,transferring tokens between addresses and how data within th e token is accessed

### 1.2.1 Project Primary Token: Tenxpay

When we as the co-founders of TenX got together to start this company, it was our vision to have assets on the blockchain be not only available to industry insiders, but rather something that can be used by any individual user in the “real world”.

Additionally, with the emergence of more and more different tokens, a growing number of users and businesses truly struggle to leverage on the existing infrastructure to make this interconnectedness of physical and virtual platforms become a reality.

At TenX, we strive to offer the user access to as large as possible a range of blockchain assets at a maximum degree of convenience, while adhering to the highest security standards in the ecosystem

To the end-user, we offer the TenX Card, a debit (and, in time to come, credit) card, with an accompanying TenX Wallet, a mobile wallet that can be funded not only with Bitcoin (BTC),Ether (ETH), and Dash (DASH) as currently possible, but also with virtually any blockchain asset in time to come. TenX payment facilities which include the physical and virtual debit card can be used in almost 200 countries at over 36 million points of acceptance today.

This is possible as we have card issuance partnerships with major credit card companies.

Moreover, users and businesses can exchange their blockchain assets seamlessly from one user to another in a decentralized manner, removing any risk that is usually associated with current centralized solutions.

#### 1.2.1.1 Product Advantages:

1. Multi-asset (any blockchain asset compatible with and accepted by the TenX Wallet)
2. Assets stay in cryptocurrency
3. Best available foreign exchange and transaction fees (with no other charges)
4. Decentralized and trustless storage

### 1.2.2 Tenxpay Token:

TenX connects your blockchain assets for everyday use. TenX’s debit card and banking licence will allow us to be a hub for the blockchain ecosystem to connect for real-world use cases.

Details: 12/03/2018

1 PAY = $0.67 USD Market Cap = $29,302,619.00 USD  
Circulating Supply = 109,347,861.00 PAY  
Total Supply = ($54,993,597.93) 205,218,255.948577763364408207 PAY Subunits =

## 2. Data Description

The Data used for the project is divided in two parts:

1. Token network edge files:

There are 40 Token network edge files.Token edge files have 4 columns: from\_node, to\_node, unix-time, total-amount.For each row it implies that from\_node sold total-amount of the token to to\_nodeid at time unix-time.

1. from\_node : Id which sells the token in the transaction
2. to\_node : Id which buys the token in the transaction
3. unixtime : Unix time of the transaction
4. totalamount : Total amount of the tenxpay token involved in the transaction For Part 1,2 and 4 of the project we will only use the tenxpay token network edge file. For part 3 we will use the token network edge files of all 40 tokens.
5. Token price files: Price dataset for ten token it contains 334 rows and 7 columns as follows:
6. Date
7. Open : Opening price of the token on that day
8. High : Max price of the token on that day
9. Low : Min price of the token on that day
10. Close : Closing price of the token on that day
11. Volume : Volume of the token on that day
12. Market Cap: Market Cap of that token on that day We use price data in part 2 and 4.

## 3. Preprocessing

There could be some records in the transactions where total amount is very large. Some of this records could be due to some bug or glitch(integer overflow problem). These values can be separated from data using a threshold value.

Calculating this value for the tenxpay token:The value of transaction amount can’t be greater than the max value where, max value = total supply of tokens \* subunits.subsituting the values from above.

##Load Token   
data <- read.delim("networktenxpayTX.txt",header=FALSE,sep=" ")  
tokenFrame<-as.data.frame(data)  
colnames(tokenFrame)<- c(" fromNodeID", "toNodeID", "Time", "Amount")  
  
## fromNodeID toNodeID Time Amount  
## 560 1452 1524611450 1.728672e+20  
## 2011173 2011174 1524611865 4.556238e+20  
## 75989 1822217 1524612292 5.795000e+20  
## 40002 6382858 1524612655 4.481100e+20  
## 17 2029263 1524612851 4.998000e+21  
## 222770 4848204 1524612957 3.283584e+20  
  
#Claculating Outlier Threshold  
x <- 205218255.948577763364408207  
y <- 10^18  
threshold <- x\*y  
  
## Finding Outliers  
Outlierdata <-tokenFrame[which(tokenFrame$Amount> threshold),]  
  
## Total number of Outliers  
message("Total number of outliers: ", length(Outlierdata$tokenAmount))

## Total number of outliers: 0

## Data Without Outliers  
data1 <-tokenFrame[which(tokenFrame$Amount< threshold),]  
View(data1)  
  
## Distribution   
## fromNodeID toNodeID Time Amount  
## 560 1452 1524611450 1.728672e+20  
## 2011173 2011174 1524611865 4.556238e+20  
## 75989 1822217 1524612292 5.795000e+20  
## 40002 6382858 1524612655 4.481100e+20  
## 17 2029263 1524612851 4.998000e+21  
## 222770 4848204 1524612957 3.283584e+20

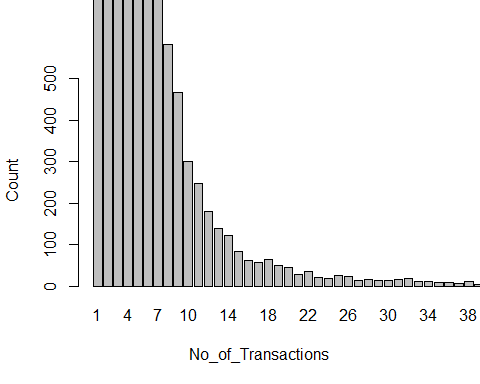
## 4.Token Data Analysis

### 4.1 Finad and Fit Distribution

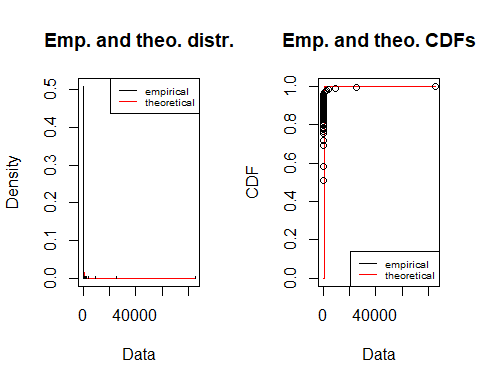
**The package we use to fit distribution:**  fitdistrplus, provide the function fitdist() we can use to fit distribution of our data. **The function we use to fit distribution:**  fitdist(). Fit of univariate distributions to different type of data with different estimate method we can choose: maximum likelihood estimation (mle), moment matching estimation (mme), quantile matching estimation (qme), maximizing goodness-of-fit estimation (mge), the default and we mostly used one is MLE. Output of this function is S3 object, we can use several methods like plot(), print(), summary() to visualize it or get more detailed information.We use plot in this project.

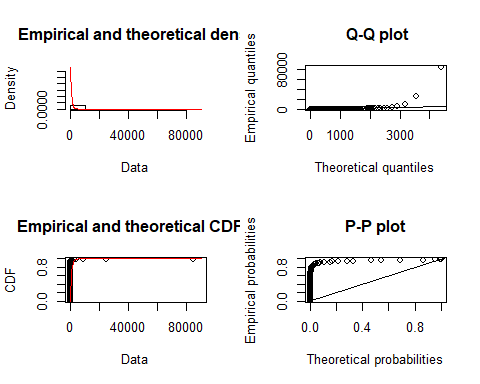
## Table 1: BUyer Transaction and its frequency   
BuyerData <- (table(data1[2]))  
DataFrequency <- as.data.frame(BuyerData)  
colnames(DataFrequency) <- c("UserID", "frequency")  
View(DataFrequency)

## Table 2: Frequency of No\_of\_Transactions  
FreqBuyers= table(DataFrequency$frequency)  
frenumbuy=as.data.frame(FreqBuyers)  
colnames(frenumbuy)<- c("No\_of\_Trasactions", "Count")  
View(frenumbuy)  
## Bar Plot  
barplot(frenumbuy$Count, names.arg = frenumbuy$No\_of\_Trasactions, ylab=" Count", xlab="No\_of\_Transactions", xlim=c(0,40),ylim=c(0,500))



##Fit Distribution  
library(MASS)  
library(lsei)  
library(npsurv)  
library(survival)  
library(fitdistrplus)  
fit<-fitdist(frenumbuy$Count, "pois", method='mle')  
plot(fit)# fit distribution  
# Poisson Distribution  
library(fitdistrplus)  
fit <- fitdist(frenumbuy$Count, "pois", method="mle")  
# Exponential Distribution  
fit1 <- fitdist(frenumbuy$Count, "exp",method = "mme")  
# Normal Distribution  
fit2 <- fitdist(frenumbuy$Count, "norm")  
plot(fit1)





## Fitting of the distribution ' pois ' by maximum likelihood   
## Parameters:  
  
## p1 = 0.033  
## p2 = 0.364  
## p3 = 0.603

The sells part is similar to buys part, so we didn’t show the code in this report, the follow is the plot of sells:

Correlations of Token Data and Price DataMax amount of the token data without outlier:

max(WithoutOutlierdata$totalAmount)

## [1] 7.108079e+24

First step: Ensure that the tenxpay token data is within the time frame of the Price Data

Following is the date interval of the price data:

Start Date: 06/27/2017 End Date: 07/14/2018

Lets check the date range for the token data

First convert the Unix time of the token data to Date format

newwithoutoutlierdata <- WithoutOutlierdata  
newwithoutoutlierdata$unixTime <- as.Date(as.POSIXct(WithoutOutlierdata$unixTime,origin="1970-01-01",tz="GMT"))  
class(newwithoutoutlierdata$unixTime)

## [1] "Date"

loading the price data :

priceData <- read.delim("tenxpay", header = TRUE, sep = "\t")  
head(priceData)

priceData$Date <- as.Date(priceData$Date, "%m/%d/%Y")  
nrow(priceData)

## [1] 384

adding the price return column to the pricedata dataframe

priceData$return <- return  
priceData$return[378]

## [1] -0.1317901

head(newwithoutoutlierdata$Date)

## NULL

colnames(newwithoutoutlierdata)<-c('fromNode','toNode','Date','totalAmount')  
newwithoutoutlierdata$Date[4]

## [1] "2018-04-24"

class(priceData$Date)

## [1] "Date"

merge both the dataframes according to dates

newdataset <- merge(newwithoutoutlierdata,priceData,by = "Date")  
head(newdataset$Date)

## [1] "2017-06-27" "2017-06-27" "2017-06-27" "2017-06-27" "2017-06-27"  
## [6] "2017-06-27"

frequencytrans <- table(newdataset$Date)  
freqtrans<- as.data.frame(frequencytrans)  
summary(freqtrans)

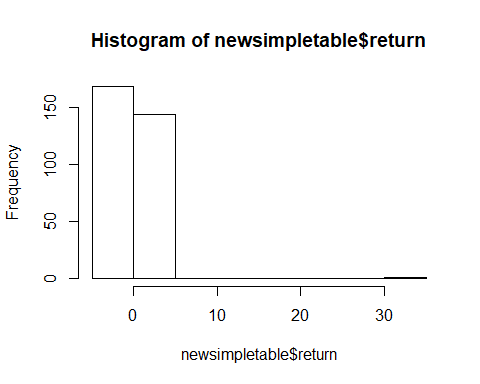
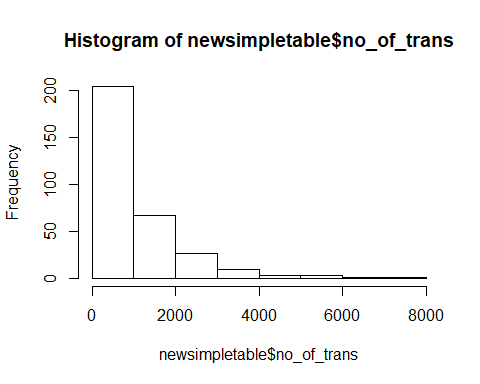
## Var1 Freq   
## 2017-06-27: 1 Min. : 9.0   
## 2017-06-28: 1 1st Qu.: 357.2   
## 2017-06-29: 1 Median : 790.0   
## 2017-06-30: 1 Mean :1049.8   
## 2017-07-01: 1 3rd Qu.:1272.2   
## 2017-07-02: 1 Max. :7088.0   
## (Other) :308

Data Preparation:

summary(newsimpletable)

## Date return no\_of\_trans   
## Min. :2017-06-27 Min. :-0.98481 Min. : 9.0   
## 1st Qu.:2017-09-13 1st Qu.:-0.05329 1st Qu.: 357.2   
## Median :2017-11-30 Median : 0.00000 Median : 790.0   
## Mean :2017-11-30 Mean : 0.10975 Mean :1049.8   
## 3rd Qu.:2018-02-16 3rd Qu.: 0.06177 3rd Qu.:1272.2   
## Max. :2018-05-06 Max. :32.40000 Max. :7088.0

hist(newsimpletable$no\_of\_trans)



Analysis:

par(mfrow = c(1,1))  
bx = boxplot(newsimpletable$no\_of\_trans,ylim = c(0,5000))

distribution of no of transactions:

quantile(newsimpletable$no\_of\_trans, seq(0,1,0.02))

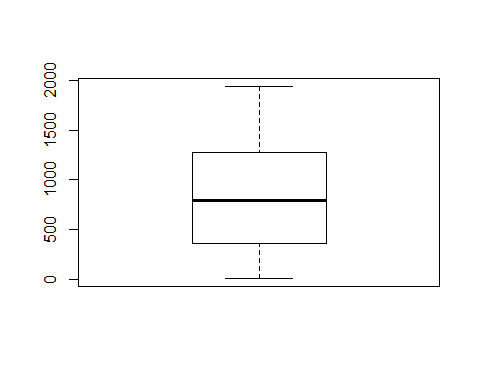
boxplot stats:

bx$stats

## [,1]  
## [1,] 9  
## [2,] 356  
## [3,] 790

capping at 96%

newsimpletable$no\_of\_trans<-ifelse(newsimpletable$no\_of\_trans>2000,1500,newsimpletable$no\_of\_trans)  
boxplot(newsimpletable$no\_of\_trans)



checking unique buyers:

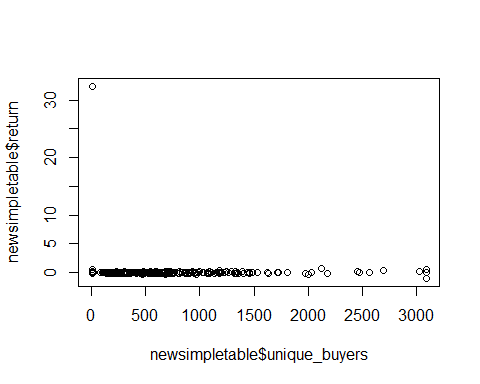
par(mfrow=c(1,2))  
bx = boxplot(newsimpletable$unique\_buyers)  
quantile(newsimpletable$unique\_buyers, seq(0,1,0.02))

## 0% 2% 4% 6% 8% 10% 12% 14% 16%   
## 7.00 11.26 115.52 135.34 155.00 164.50 174.00 180.28 190.16   
## 18% 20% 22% 24% 26% 28% 30% 32% 34%   
## 199.34 211.60 218.72 230.00 239.00 253.64 275.60 302.96 322.26   
## 36% 38% 40% 42% 44% 46% 48% 50% 52%   
## 353.04 388.62 415.20 444.84 475.88 489.94 517.44 532.00 559.32   
## 54% 56% 58% 60% 62% 64% 66% 68% 70%   
## 579.04 589.12 605.08 624.00 652.78 677.64 686.90 704.00 734.50   
## 72% 74% 76% 78% 80% 82% 84% 86% 88%   
## 768.00 787.62 856.44 886.50 955.60 1030.30 1108.56 1174.08 1214.44   
## 90% 92% 94% 96% 98% 100%   
## 1331.50 1401.96 1551.12 1889.12 2467.54 4708.00

newsimpletable$unique\_buyers<-ifelse(newsimpletable$unique\_buyers>3087,3088,newsimpletable$unique\_buyers)  
boxplot(newsimpletable$unique\_buyers)

lets check unique buyers vs return:

plot(newsimpletable$unique\_buyers,newsimpletable$return)

 lets check the correlation:

cor(newsimpletable[,c(2:5)])

## Warning in cor(newsimpletable[, c(2:5)]): the standard deviation is zero

## return no\_of\_trans total\_token unique\_buyers  
## return 1.00000000 -0.1337920 NA -0.06212329  
## no\_of\_trans -0.13379200 1.0000000 NA 0.75555525  
## total\_token NA NA 1 NA  
## unique\_buyers -0.06212329 0.7555553 NA 1.00000000

lets extract one more feature

no of investors who have bought more than ten token:

library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:MASS':  
##   
## select

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

summary(newdataset)

## Date fromNode toNode   
## Min. :2017-06-27 Min. : 5 Min. : 5   
## 1st Qu.:2017-08-15 1st Qu.: 17 1st Qu.: 431538   
## Median :2017-09-17 Median : 297031 Median :2062611   
## Mean :2017-10-16 Mean :1644441 Mean :3179799   
## 3rd Qu.:2017-12-15 3rd Qu.:1943406 3rd Qu.:6393477   
## Max. :2018-05-06 Max. :6438429 Max. :6438432   
##   
## totalAmount Open High Low   
## Min. :0.000e+00 Min. : 0.5685 Min. : 0.6111 Min. : 0.4787   
## 1st Qu.:5.445e+11 1st Qu.: 1.7300 1st Qu.: 1.9500 1st Qu.: 1.5800   
## Median :2.658e+12 Median : 2.3200 Median : 2.4900 Median : 2.0300   
## Mean :1.844e+13 Mean : 3.6169 Mean : 3.9031 Mean : 2.2656   
## 3rd Qu.:1.005e+13 3rd Qu.: 3.6000 3rd Qu.: 3.8600 3rd Qu.: 3.1600   
## Max. :7.108e+16 Max. :73.0600 Max. :86.2600 Max. :49.0500   
##   
## Close Volume Market.Cap   
## Min. : 0.570 144,769,000: 7088 - : 9809   
## 1st Qu.: 1.740 11,828,700 : 6226 471,783,000: 7088   
## Median : 2.220 20,158,400 : 5945 333,047,000: 6226   
## Mean : 2.545 72,243,200 : 5161 131,122,000: 5945   
## 3rd Qu.: 3.590 17,115,200 : 5073 300,417,000: 5161   
## Max. :63.830 9,172,650 : 4588 224,227,000: 4532   
## (Other) :295556 (Other) :290876   
## return   
## Min. :-0.98481   
## 1st Qu.:-0.06126   
## Median : 0.00000   
## Mean : 0.02160   
## 3rd Qu.: 0.07759   
## Max. :32.40000   
##

we will first need to find the total number of investors per day and then find the total tokens bought by each investor per day. then we have to find the number pf investors who bought more than 10 token that day.

investortable <- newdataset %>% group\_by(Date)  
totalinvestor\_table<- summarize(investortable , total\_investors = n\_distinct(toNode))  
head(totalinvestor\_table)

## # A tibble: 6 x 2  
## Date total\_investors  
## <date> <int>  
## 1 2017-06-27 21  
## 2 2017-06-28 11  
## 3 2017-06-29 10  
## 4 2017-06-30 7  
## 5 2017-07-01 11  
## 6 2017-07-02 10

newinvestortable <- newdataset %>% group\_by(Date,toNode)  
newinvestortable\_totaltoken <- summarise(newinvestortable,total\_token = sum(totalAmount))  
summary(newinvestortable\_totaltoken)

## Date toNode total\_token   
## Min. :2017-06-27 Min. : 5 Min. :0.000e+00   
## 1st Qu.:2017-08-21 1st Qu.: 460418 1st Qu.:6.031e+11   
## Median :2017-09-26 Median :3571979 Median :2.500e+12   
## Mean :2017-10-18 Mean :3617180 Mean :2.940e+13   
## 3rd Qu.:2017-12-15 3rd Qu.:6400372 3rd Qu.:1.000e+13   
## Max. :2018-05-06 Max. :6438432 Max. :7.108e+16

subsetnewinvestortable\_totaltoken <- newinvestortable\_totaltoken[ which(newinvestortable\_totaltoken$total\_token>100), ]  
finalinvestortable <- summarise(subsetnewinvestortable\_totaltoken, final\_investors = n\_distinct(toNode))  
newfinalinvestortable <- merge(totalinvestor\_table,finalinvestortable,by = "Date")  
newfinalinvestortable$percentage\_investors <- (newfinalinvestortable$final\_investors/newfinalinvestortable$total\_investors)\*100  
newsimpletable <- merge(newsimpletable,newfinalinvestortable,by = "Date")

writiting the previous day fix :

head(newsimpletable$Date)

## [1] "2017-06-27" "2017-06-28" "2017-06-29" "2017-06-30" "2017-07-01"  
## [6] "2017-07-02"

tail(newsimpletable$return)

## [1] 0.04666667 0.02547771 0.01863354 -0.01829268 0.01242236 -0.04294479

shift <- function(x, n){  
 c(x[-(seq(n))], rep(NA, n))  
}  
newsimpletable$return <- shift(newsimpletable$return, 1)  
newsimpletable <- newsimpletable[1:(nrow(newsimpletable)-1),]

### 4.4 Create Linear Regression Model

In this part we created a multiple linear regression model on the price data and our primary token data.Our response variable(y) is simple price return given by where, is the token price in dollar for day. The regressor variables(x1…xn) are as follows: 1)No of token transactions per day 2)No of total token bought or sold per day 3)No of unique buyers per day 4)Percentage of investors who bought more than token per day.

Calculate response variable from the price data using open price. The first few records of the response variable as follows:

return <- numeric(nrow(priceData))  
return[379] <- priceData$Open[379]  
for (i in 378:1) {  
 return[i] <- (priceData$Open[i]-priceData$Open[i+1])/priceData$Open[i+1]  
}  
priceData$return <- return  
head(priceData$return)

## [1] 0.068813980 0.163917541 0.001947088 0.010113908 -0.057572838  
## [6] -0.022667458

Merge both the datasets according to date which would help us in calculating the regressor variables.

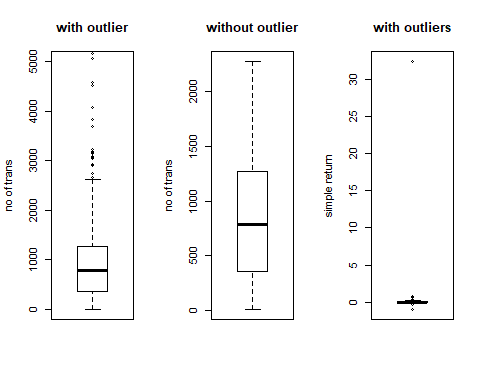
colnames(newwithoutoutlierdata)<-c('fromNode','toNode','Date','totalAmount')  
newdataset <- merge(newwithoutoutlierdata,priceData,by = "Date")

#### 4.4.1 Model 1:Simple Linear Regression

## [1] 25 13 15 10 12 15

Data preparation: univariate and bivariate analysis. **Univariate Analysis**:

Check for outliers and then remove them by capping our data to the max value.To remove majority of the outlier we are capping the data 92% quantile. Then we will alsocheck the response variable.



Boxplots of variables for Univariate analysis in simple regression

But for simple return response varaible case we won’t be removing any outliers as we don’t want to have any impact on the response variable.

**Bivariate Analysis**: Draw the scatterplot between the regressor and the response variable.Before we do that it is important to shift our response variable up by one row because we will be using yesterday to model today’s response.

shift <- function(x, n){  
 c(x[-(seq(n))], rep(NA, n))  
}  
newsimpletable$return <- shift(newsimpletable$return, 1)  
newsimpletable <- newsimpletable[1:(nrow(newsimpletable)-1),]  
head(newsimpletable$return)

## [1] 0.0000000 0.0000000 0.0000000 0.0000000 32.4000000 -0.1317901

#plot(newsimpletable$no\_of\_trans,newsimpletable$return)

We can see that there isn’t much relation between no of transactions each day and the simple return variable. We can confirm this by finding the correlation between the twocor(newsimpletable[,c(6,7)])

## return no\_of\_trans  
## return 1.00000000 -0.07630676  
## no\_of\_trans -0.07630676 1.00000000

After completing the analysis we move ahead to fit the linear model.

final\_data <- lm(return~no\_of\_trans,data=newsimpletable)  
summary(final\_data)

##   
## Call:  
## lm(formula = return ~ no\_of\_trans, data = newsimpletable)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.816 -0.231 -0.141 -0.014 32.099   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.3039009 0.1770723 1.716 0.0871 .  
## no\_of\_trans -0.0002076 0.0001538 -1.350 0.1781   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.833 on 311 degrees of freedom  
## Multiple R-squared: 0.005823, Adjusted R-squared: 0.002626   
## F-statistic: 1.821 on 1 and 311 DF, p-value: 0.1781

From the summary of the linear model we can clearly see that it is not a good model, the p-values of the intercept and the regressor are very high, also the r-square conveys only 2% of the response variable.

**Testing**

We can test this linear model using the plot function:

1. **Residuals vs fitted**:

Even though we see a horizontal line in this plot we can see that the residuals are not evenly spread across the line.It may not suggest a pattern but we can see most of the points in the left speculating a non-linear behaviour

1. **Normal Q-Q**:

The plot leads us to believe that the residuals follow normality as most of the points are seen to lie on the line or near it although few observations such as 42,2 and 156 are significantly away from the line.

1. **Scale-Location**:

The horizontal line in this plot suggests that homoscedasticity is observed in the residuals.

1. **Residuals vs leverage**:

This plots finds the observations which have a high influence on the regression model according to the cooks distance. We can see that observation 42 is out of the region thus is an outlier which could have an impact on the model if we do remove or decide to keep it in the model.

#### 4.4.2 Model 2:Multiple Linear Regression

Create three new features for our multiple regression model:

1. Total number of tokens involved per day:

First divide the total amount of each transaction with number of sub units of the tenxpay token(10^8) to get the number of tokens, then we will group by and sum each amount according to the date

First few records are as follows

## [1] 1.409702e+14 3.119692e+15 2.003250e+14 3.125740e+14 1.575796e+14  
## [6] 4.625641e+14

1. Number of Unique buyers per day:

Group the dataset according to the date and then count the number of unique buyers for each day.

First few records are as follows:

## [1] 21 11 10 7 11 10

1. Percentage of investors who have bought more than 10 tokens:

Find the total number of investors per day and then find the total tokens bought by each investor per day. Then find the number of investors who bought more than 10 token that day.

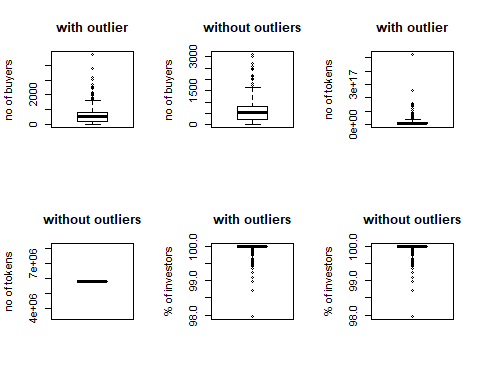
First few records are as follows:

## [1] 100 100 100 100 100 100

Now that we have all the required features lets do the analysis as we did for the simple linear regression:

**Univariate analysis**:

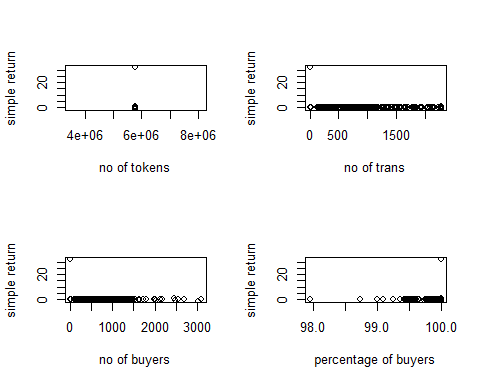
Create a boxplot of unique buys regressor variable and then we check for the number of tokens(and capping at 94%) and percentage of investors variable:



Boxplot if variables for univariate analysis in multiple regression

After the univariate analysis lets do the **bivariate analysis**:

plot number of transactions vs return:  
plot total token vs return:  
plot unique buyers vs return:  
plot percentage investors vs return:



Scatter plot of variables for Bivariate analysis in multiple regression

After analysing lets create an initial multiple regression linear model, Use the car library to calculate variable inflation factor and find the multicollinearity.

final\_data3 <- lm(return~no\_of\_trans+total\_token+unique\_buyers+percentage\_investors,data = newsimpletable)  
library(car)

Since the VIF for all the regressor variables is not much higher than 5 we will keep all of them in the initial model

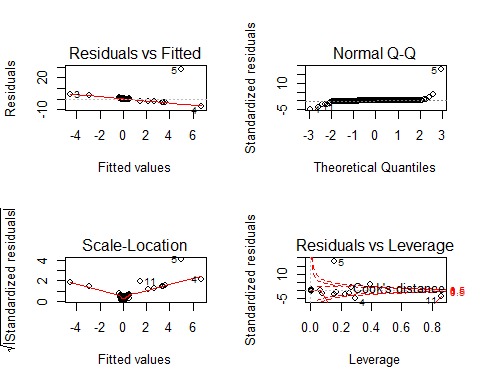
summary(final\_data3)

##   
## Call:  
## lm(formula = return ~ no\_of\_trans + total\_token + unique\_buyers +   
## percentage\_investors, data = newsimpletable)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.830 -0.233 -0.138 -0.002 32.091   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -1.449e+01 5.307e+01 -0.273 0.785  
## no\_of\_trans -2.048e-04 3.104e-04 -0.660 0.510  
## total\_token NA NA NA NA  
## unique\_buyers -2.541e-07 3.763e-04 -0.001 0.999  
## percentage\_investors 1.480e-01 5.310e-01 0.279 0.781  
##   
## Residual standard error: 1.839 on 309 degrees of freedom  
## Multiple R-squared: 0.006073, Adjusted R-squared: -0.003577   
## F-statistic: 0.6294 on 3 and 309 DF, p-value: 0.5965

From the summary, p-value for the variables is very high suggesting that there is no linear relation between the regressor variable and the response variables.But before giving up on this we haven’t yet used any features from the price data lets use the inherent features such as open, low, high etc to create a new linear model and see how we fare:

First find the multicollinearity and then use the step function to find the right combination of features:

Finally, Check the summary and plot of our final data model:

##   
## Call:  
## lm(formula = return ~ total\_token + Open + High + Low, data = newsimpletable)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.6956 -0.1706 -0.0263 0.1039 27.4404   
##   
## Coefficients: (1 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -0.32508 0.11549 -2.815 0.00519 \*\*   
## total\_token NA NA NA NA   
## Open 0.30760 0.05222 5.891 1.00e-08 \*\*\*  
## High -0.28557 0.05075 -5.627 4.11e-08 \*\*\*  
## Low 0.18978 0.04375 4.338 1.95e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.703 on 309 degrees of freedom  
## Multiple R-squared: 0.1476, Adjusted R-squared: 0.1393   
## F-statistic: 17.84 on 3 and 309 DF, p-value: 1.057e-10

Residual plot of the linear model for multiple regression

## 5.Conclusion

### 5.1 Q1

We totally tried 5 distributions while fiting distributions with our data: Poisson Distribution, Weibull Distribution, Exponential Distribution, Geometric Distribution and Normal Distribution. After compare with parameters and graphs of different distributions, we can find: The distribution of buyers and sellers are similiar. The estimate parameters of Normal Distribution are very close to the parameters of our token data, but the graphs don’t fit well. The graphs of Poisson Distribution and Exponential Distribution both fit our token data well, but for the parameters,we should not use Poisson Distribution.

Hence, we finally conclude that **Exponential Distribution** fit our distribution best.

### 5.2 Q2

The number of layers is 15. We tried several features of token data : (1) number of transactions(best correlation is 0.193) (2)Number of Unique buyers per day(best correlation is 0.14943) (3)Percentage of investors who have bought more than 10 tokens(best correlation is 0.1347)

Hence, the best correlation is 0.4943, when feature of token data is number of unique buyers per day.

### 5.4 Q4

When we use the price data columns and remove the features which have high p-value, we get value of **71%**. We have analysed how we can use simple and multiple linear regression to create linear models by doing feature extraction on the tenxpay token and price data. The model can be redeveloped and tested with different features to improve the R-squared value however th

**Following are the libraries**

Library MASS: This function is generic; there exist methods for classes lm and glm and the default method will work for many other classes.

Library LSEI:

Library NPSURV: Computes an estimate of a survival curve for censored data using either the Kaplan-Meier or the Fleming-Harrington method or computes the predicted survivor function. For competing risks data it computes the cumulative incidence curve. This calls the survival package's survfit.formula function. Attributes of the event time variable are saved (label and units of measurement).

For competing risks the second argument for Surv should be the event state variable, and it should be a factor variable with the first factor level denoting right-censored observations.

Library SURVIVAL: Returns an object of class "aareg" that represents an Aalen model

Other libraries include: dplyr, car, etc

References:

<https://cointelegraph.com/explained/erc-20-tokens-explained>

<https://medium.com/blockchannel/understanding-the-ethereum-ico-token-hype-429481278f45>

<https://tenx.tech/en/>

<https://etherscan.io/token/0xd850942ef8811f2a866692a623011bde52a462c1>

<https://downloads.ctfassets.net/xecblntwky6m/5sWzq3FOoMiWuiQ0i8gKCG/84d359acabe7bb1ef3d1c1252c82828a/tenx_whitepaper_final.pdf>