

# EDA\_v2

March 29, 2021

```
[2]: library(lubridate)
library(data.table)
library(dplyr)
library(Polychrome)
library(ggplot2)
library(grid)
library(gridExtra)
require(invgamma)
require(logitnorm)
```

## 1 Data preparation and cleaning

```
[3]: datetimediff = function(t){
  t = supply(strsplit(t, " "), "[", 2)
  #tt = as.Date(t, format='%d %b %Y %H:%M:%S')
  tt = as.POSIXct(t,format='%d %b %Y %H:%M:%S')
  tt = supply(tt, function(d) as.numeric(difftime(d , tt[1],units = 'mins'))
  ↪)
  return(tt)
}
```

```
[4]: header = c("RetweetCount", "UserId", "ScreenName", "FollowerCount",
  ↪ "DistanceFromRoot", "Time", "ParentScreenName", "Text")
file_list <- list.files(path="../data/retweet_data/",full.names = T)
dataset = data.frame()
counter = 1
all.data = list()
for (f in file_list){
  d = read.table(f,sep = '\t',col.names = header,header=FALSE, quote="")
  #d$FollowerCount = as.numeric(d$FollowerCount)
  d$TimeDiff = datetimediff(d$Time)
  d$TimeDiff = log(d$TimeDiff+1) # scale to log(x+1)
  d$Time = NULL
  d$Text = NULL
  #d$response = NA
  #d$rootFollow = NA
  #d$rootFollow[1] = head(d$FollowerCount,n=1)
```

```

    d$retweets = sapply(d$ScreenName, function(x) d %>%
↳filter(ParentScreenName==x) %>% nrow )
    #d$retweets[1] = d$retweets[1] - 1
    d$retweets[1] = tail(d$RetweetCount,n=1)
    d$TimeRatio = d$TimeDiff/tail(d$TimeDiff,n=1)
    d$RetweetRatio = d$RetweetCount/tail(d$RetweetCount,n=1)
    d$case = counter
    counter = counter + 1
    dataset <- rbindlist(list(dataset, d), use.names = T)
}
dataset %>%
    mutate(FollowerCount = as.
↳numeric(replace(FollowerCount,FollowerCount=="None",0))) -> dataset
dataset %>% select(c(RetweetCount, case, FollowerCount, DistanceFromRoot,
↳TimeDiff, TimeRatio, retweets, RetweetRatio, case)) -> dataset
C52 = createPalette(52, c("#ff0000", "#00ff00", "#0000ff"))
dataset$color = rep(C52, times=c(table(dataset$case)))
#-----
head(dataset,10)
dim(dataset)

```

A data.table: 10 × 9

	RetweetCount <int>	case <dbl>	FollowerCount <dbl>	DistanceFromRoot <int>	TimeDiff <dbl>	TimeRatio <dbl>	re <
	0	1	87039	0	0.0000000	0.00000000	2
	1	1	367	1	0.3364722	0.05749686	0
	2	1	444	1	0.4054651	0.06928646	0
	3	1	42	1	1.1205912	0.19148823	0
	4	1	235	1	1.1579289	0.19786854	0
	5	1	898	1	1.2575136	0.21488572	0
	6	1	52	1	1.3523928	0.23109882	0
	7	1	49	1	2.4009210	0.41027282	0
	8	1	415	1	2.4918271	0.42580698	0
	9	1	90	1	2.5533438	0.43631905	0

1. 12248 2. 9

```

[5]: # Saving model data to disk
model_input = list(M=dataset$retweets, S = dataset$TimeDiff, f =
↳dataset$FollowerCount,
    d=dataset$DistanceFromRoot, X=52, N=nrow(dataset),
    J = c(dataset %>% filter(RetweetCount==0) %>%
↳select(retweets))[[1]]+1)
model_input$JCUM = c(0,cumsum(model_input$J))
model_input$StoX = rep(1:52,model_input$J)
saveRDS(model_input, '../data/model_input.rds')

```

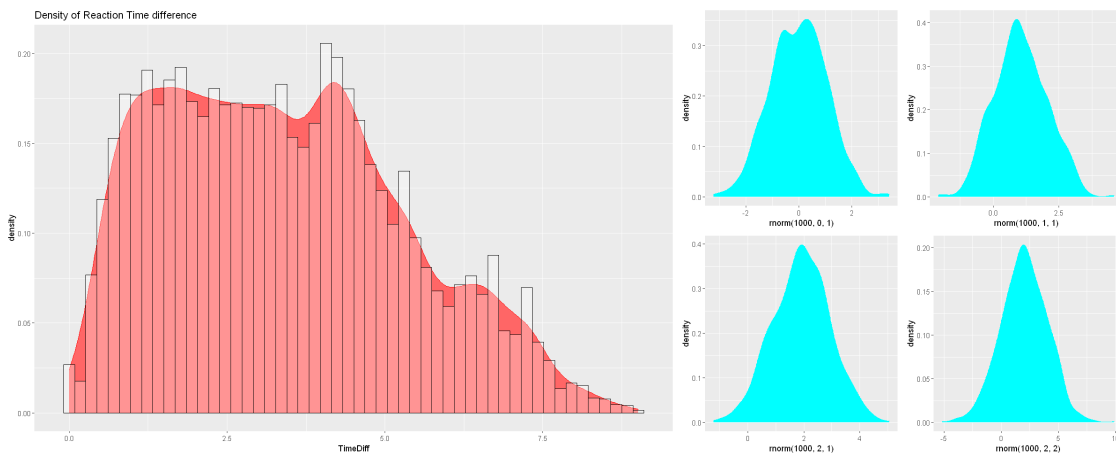
## 2 EDA

### 3 1. Response time

```
[6]: options(repr.plot.width=20, repr.plot.height=8)
#jpeg("rplot.jpg", width = 850, height = 350)
q1 <- dataset %>%
  ggplot(aes(x=TimeDiff)) +
  geom_density(color='red',fill="#FF6666") +
  geom_histogram(aes(y=..density..),bins=52,alpha=.3,
  position="identity",colour="black", fill="white")+
  #coord_cartesian(xlim = c(0, 4000)) +
  labs(title = "Density of Reaction Time difference")

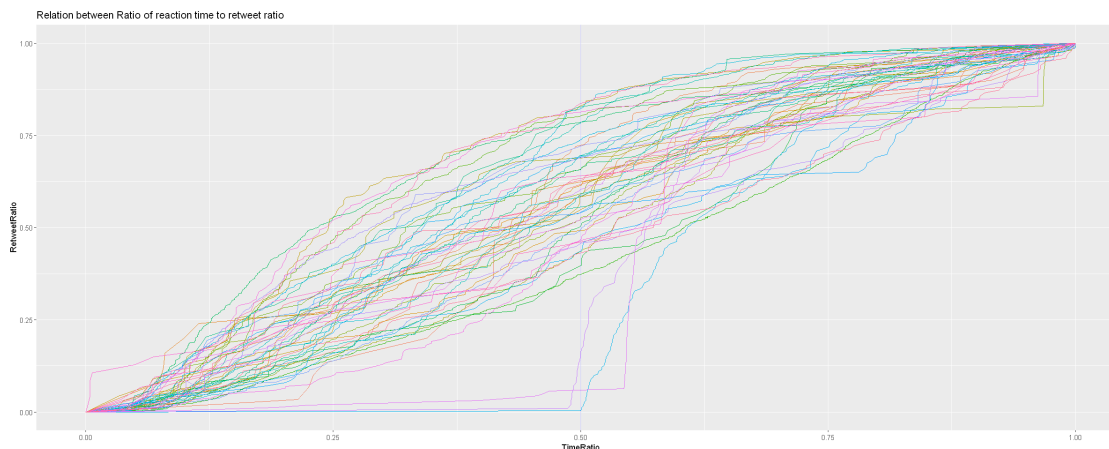
q2 <- ggplot() + geom_density(aes(x=rnorm(1000,0,1)), color='cyan',fill='cyan')
q3 <- ggplot() + geom_density(aes(x=rnorm(1000,1,1)), color='cyan',fill='cyan')
q4 <- ggplot() + geom_density(aes(x=rnorm(1000,2,1)), color='cyan',fill='cyan')
q5 <- ggplot() + geom_density(aes(x=rnorm(1000,2,2)), color='cyan',fill='cyan')

grid.arrange(q1,q2,q3,q4,q5,
  widths = c(2, 1, 1,1),
  layout_matrix = rbind(c(1, 1, 2,3),
                        c(1, 1, 4,5))
)
#dev.off()
```



```
[7]: options(repr.plot.width=20, repr.plot.height=8)
dataset %>% group_by(case) %>%
  ggplot(aes(y= RetweetRatio, x=TimeRatio,col=color)) +
  geom_line() +
  theme(legend.position = "none") +
```

```
geom_vline(xintercept = .5,col='blue',alpha=.2) +
labs(title="Relation between Ratio of reaction time to retweet ratio")
```



## 4 2. Retweet Count

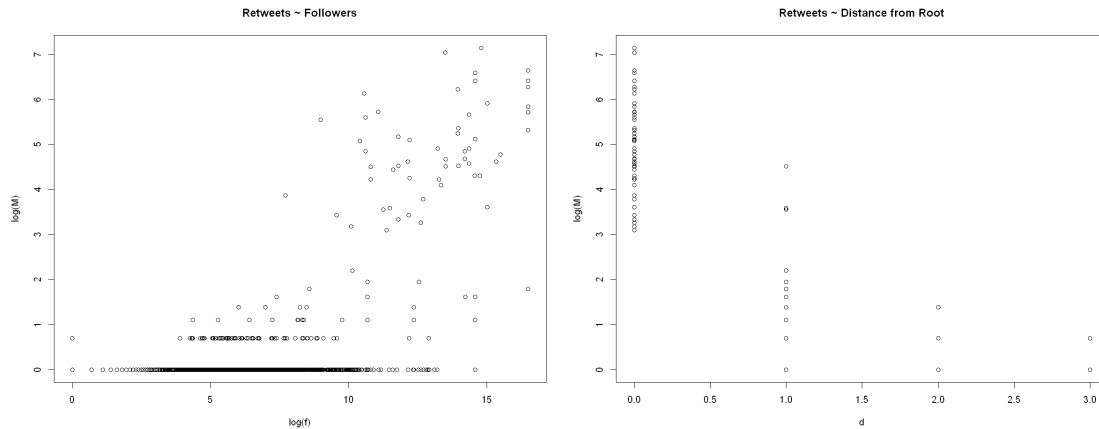
```
[8]: table(dataset$DistanceFromRoot)
```

```
0      1      2      3
52 11882   307      7
```

```
[7]: #jpeg("rplot.jpg", width = 850, height = 350)
par(mfrow=c(1,2))
plot(log(dataset$FollowerCount+1), log(dataset$retweets+1), main="Retweets ~ FollowerCount", xlab = "log(f)", ylab="log(M)")
plot(dataset$DistanceFromRoot, log(dataset$retweets+1), main = "Retweets ~ Distance from Root", xlab="d", ylab="log(M)")
print(summary(aov(log(retweets+1) ~ log(FollowerCount+1) + DistanceFromRoot + TimeDiff, dataset)))
#dev.off()
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
log(FollowerCount + 1)	1	162.8	162.83	1756.73	< 2e-16 ***
DistanceFromRoot	1	150.7	150.74	1626.27	< 2e-16 ***
TimeDiff	1	1.4	1.42	15.33	9.06e-05 ***
Residuals	12244	1134.9	0.09		

---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

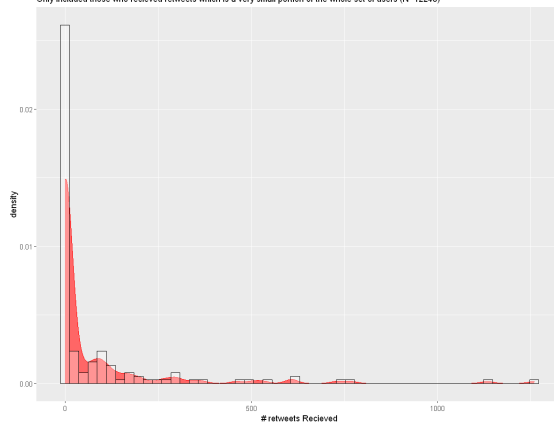


```
[8]: options(repr.plot.width=20, repr.plot.height=8)
q1 <- dataset %>% filter(retweets!=0) %>%
  ggplot(aes(x=retweets)) +
  geom_density(color='red',fill="#FF6666") +
  geom_histogram(aes(y=..density..),bins=52,alpha=.3,
  ↪position="identity",colour="black", fill="white")+
  #coord_cartesian(ylim = c(0, .06), xlim=c(0,100)) +
  labs(title = "Distribution of the number of retweets for all users",
  ↪(n=144)", x="# retweets Recieved",
  subtitle = "Only included those who recieved retweets which is a very
  ↪small portion of the whole set of users (N=12248)")

q2 <- dataset %>% filter(RetweetCount==0) %>%
  ggplot(aes(x=retweets)) +
  geom_density(color='red',fill="#FF6666") +
  geom_histogram(aes(y=..density..),bins=52,alpha=.3,
  ↪position="identity",colour="black", fill="white")+
  labs(title = "Distribution of the number of retweets for roots only",
  ↪(n=52)", x="# retweets Recieved")

grid.arrange(q1,q2,
  widths = c(1,1),
  layout_matrix = rbind(c(1, 2),
  ↪c(1, 2))
)
```

Distribution of the number of retweets for all users (n=144)  
 Only included those who recieved retweets which is a very small portion of the whole set of users (N=12248)



Distribution of the number of retweets for roots only (n=52)

