

# RedditAnalysis

Frederik Mann

29 6 2021

## Posts

```
library(pacman)
p_load(RColorBrewer, # color pallets
       ggplot2, # reportable graphs
       cowplot, # arranges ggplot graphs nicely
       stargazer,
       MASS,
       DescTools,
       plyr)
```

## Import data set

```
df_berlin <- read.csv("posts_berlin_2020.csv")
df_germany <- read.csv("posts_de_2020.csv")
df_europe <- read.csv("posts_europe_2020_partial.csv")
df <- rbind(df_berlin, df_germany, df_europe)
```

## Inspect

```
str(df)

## 'data.frame': 150733 obs. of 23 variables:
## $ id : chr "eianf4" "eib7eb" "eib7f1" "eibze1" ...
## $ permalink : chr "/r/berlin/comments/eianf4/berlin_changed_my_opinion_on_fireworks/" ...
## $ author : chr "ziozxxzioz" "oyeahmann" "" "" ...
## $ author_fullname : chr "t2_6vajm" "t2_hryiqix" "NULL" "NULL" ...
## $ title : chr "Berlin changed my opinion on fireworks" "Alexanderplatz" "We need th
## $ url : chr "https://www.reddit.com/r/berlin/comments/eianf4/berlin_changed_my_op
## $ subreddit : chr "berlin" "berlin" "berlin" "berlin" ...
## $ stickied : chr "False" "False" "False" "False" ...
## $ created_utc : num 1.58e+09 1.58e+09 1.58e+09 1.58e+09 1.58e+09 ...
## $ is_original_content : chr "False" "False" "False" "False" ...
## $ author_flair_text : chr "Mitte" "" "" "" ...
## $ is_video : chr "False" "False" "False" "False" ...
## $ locked : chr "False" "False" "False" "False" ...
## $ selftext : chr "So, first NYE here since coming from Argentina. In the past few year
## $ link_flair_richtext : chr "[]" "[]" "[]" "[]" ...
## $ domain : chr "self.berlin" "i.redd.it" "bbc.com" "i.redd.it" ...
## $ over_18 : chr "False" "False" "False" "False" ...
## $ score : int 261 121 18 2 1 2 209 15 4 18 ...
## $ total_awards_received: int 0 0 0 0 0 0 0 0 0 0 ...
```

```
## $ upvote_ratio      : num  0.97 0.92 0.67 0.55 0.56 0.57 0.97 0.94 0.75 0.65 ...
## $ num_comments      : int   142 3 16 11 0 1 7 6 1 6 ...
## $ epoch             : int    1 1 1 1 1 1 1 1 1 1 ...
## $ datetime          : chr   "2020-01-01 00:31:48" "2020-01-01 01:16:07" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10" "2020-01-01 01:16:10"
```

```
df$over_18 <- as.logical(df$over_18)
df$locked <- as.logical(df$locked)
df$is_video <- as.logical(df$is_video)
df$is_original_content <- as.logical(df$is_original_content)
df$stickied <- as.logical(df$stickied)
df$subreddit <- as.factor(df$subreddit)

dim(df)
```

```
## [1] 150733      23
```

Preprocess: Treat missing values, if applicable

```
df$author_fullname[df$author_fullname == "NULL"] <- NA
df$author[df$author == ""] <- NA
nrow(df)
```

```
## [1] 150733
```

```
df <- df[!duplicated(df$id), ]
nrow(df)
```

```
## [1] 149097
```

```
# Track down variables with missing values
sum(is.na(df))
```

```
## [1] 95306
```

```
colSums(is.na(df))
```

```
##           id           permalink           author
##           0              0          47653
##  author_fullname           title           url
##           47653              0              0
##      subreddit           stickied  created_utc
##           0              0              0
##  is_original_content  author_flair_text  is_video
##           0              0              0
##           locked           selftext  link_flair_richtext
##           0              0              0
##           domain           over_18           score
##           0              0              0
## total_awards_received  upvote_ratio  num_comments
##           0              0              0
##           epoch           datetime
##           0              0
```

```
# Check the percentage of missing values in the data set
(nrow(df) - nrow(na.omit(df))) / nrow(df)
```

```
## [1] 0.3196107
```

```
df$date <- as.Date(df$datetime)
```

```
to_interval <- function(anchor.date, future.date, interval.days){
  round(as.integer(future.date - anchor.date) / interval.days, 0)
}
```

```
df$week_interval <- to_interval(as.Date('2020-01-01'),
                                df$date, 7 )
```

```
df$month <- format(df$date, "%m")
df$month <- factor(df$month)
```

```
df <- df[!(df$stickied == TRUE),]
```

```
str(df)
```

```
## 'data.frame': 149097 obs. of 26 variables:
## $ id : chr "eianf4" "eib7eb" "eib7f1" "eibze1" ...
## $ permalink : chr "/r/berlin/comments/eianf4/berlin_changed_my_opinion_on_fireworks/" ...
## $ author : chr "ziozxxioz" "oyeahmann" NA NA ...
## $ author_fullname : chr "t2_6vajm" "t2_hryiqix" NA NA ...
## $ title : chr "Berlin changed my opinion on fireworks" "Alexanderplatz" "We need th
## $ url : chr "https://www.reddit.com/r/berlin/comments/eianf4/berlin_changed_my_op
## $ subreddit : Factor w/ 3 levels "berlin","de",...: 1 1 1 1 1 1 1 1 1 1 ...
## $ stickied : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ created_utc : num 1.58e+09 1.58e+09 1.58e+09 1.58e+09 1.58e+09 ...
## $ is_original_content : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ author_flair_text : chr "Mitte" "" "" "" ...
## $ is_video : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ locked : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ selftext : chr "So, first NYE here since coming from Argentina. In the past few year
## $ link_flair_richtext : chr "[]" "[]" "[]" "[]" ...
## $ domain : chr "self.berlin" "i.redd.it" "bbc.com" "i.redd.it" ...
## $ over_18 : logi FALSE FALSE FALSE FALSE FALSE FALSE ...
## $ score : int 261 121 18 2 1 2 209 15 4 18 ...
## $ total_awards_received: int 0 0 0 0 0 0 0 0 0 0 ...
## $ upvote_ratio : num 0.97 0.92 0.67 0.55 0.56 0.57 0.97 0.94 0.75 0.65 ...
## $ num_comments : int 142 3 16 11 0 1 7 6 1 6 ...
## $ epoch : int 1 1 1 1 1 1 1 1 1 1 ...
## $ datetime : chr "2020-01-01 00:31:48" "2020-01-01 01:16:07" "2020-01-01 01:16:10" "20
## $ date : Date, format: "2020-01-01" "2020-01-01" ...
## $ week_interval : num 0 0 0 0 0 0 0 0 0 0 ...
## $ month : Factor w/ 12 levels "01","02","03",...: 1 1 1 1 1 1 1 1 1 1 ...
```

## Data Visualisation

```
data.frame(table(df$month))
```

```
##   Var1 Freq
## 1    01 13342
## 2    02 13734
## 3    03 20170
```

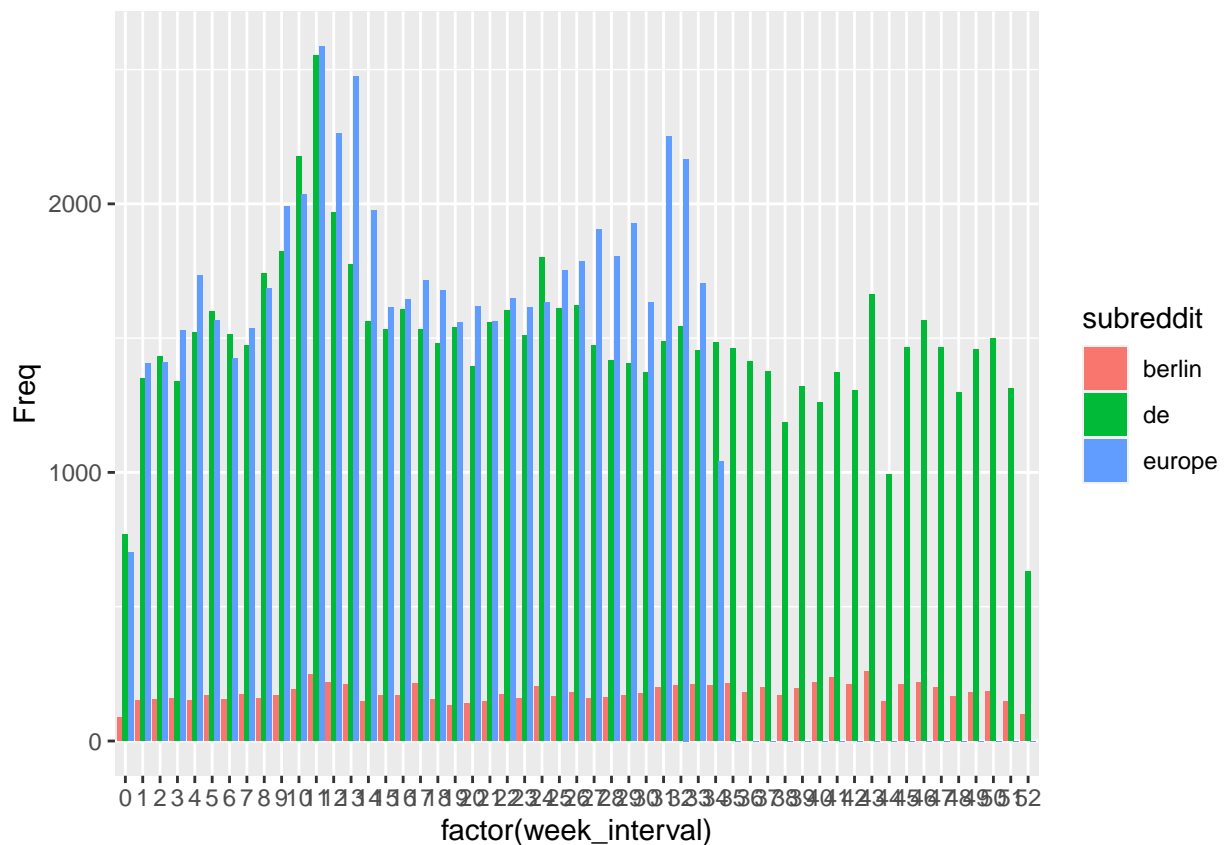
```
## 4    04 15449
## 5    05 14353
## 6    06 14851
## 7    07 15301
## 8    08 14883
## 9    09 6557
## 10   10 7218
## 11   11 6668
## 12   12 6571
```

```
dfwi <- data.frame(table(df$week_interval))
dfwi
```

```
##      Var1 Freq
## 1         0 1557
## 2         1 2904
## 3         2 2996
## 4         3 3025
## 5         4 3405
## 6         5 3336
## 7         6 3089
## 8         7 3180
## 9         8 3584
## 10        9 3982
## 11       10 4403
## 12       11 5389
## 13       12 4446
## 14       13 4459
## 15       14 3685
## 16       15 3316
## 17       16 3421
## 18       17 3460
## 19       18 3312
## 20       19 3232
## 21       20 3154
## 22       21 3266
## 23       22 3423
## 24       23 3285
## 25       24 3635
## 26       25 3531
## 27       26 3588
## 28       27 3539
## 29       28 3384
## 30       29 3498
## 31       30 3183
## 32       31 3936
## 33       32 3917
## 34       33 3367
## 35       34 2731
## 36       35 1673
## 37       36 1593
## 38       37 1577
## 39       38 1356
## 40       39 1514
## 41       40 1481
```

```
## 42 41 1609
## 43 42 1518
## 44 43 1919
## 45 44 1140
## 46 45 1676
## 47 46 1785
## 48 47 1666
## 49 48 1461
## 50 49 1637
## 51 50 1684
## 52 51 1460
## 53 52 730
```

```
tbl <- with(df, table(subreddit, week_interval))
ggplot(as.data.frame(tbl), aes(factor(week_interval), Freq, fill = subreddit)) +
  geom_col(position = 'dodge')
```



```
Gini(dfwi$Freq)
```

```
## [1] 0.2115102
```

```
nrow(df)
```

```
## [1] 149097
```

```
df_with_acc <- na.omit(df)
```

```
nrow(df_with_acc)
```

```
## [1] 101444
```

```

#df_with_acc <- df_with_acc[df_with_acc$score > 10,]

gini_by_7days <- data.frame(
  interval=character(),
  gini=double(),
  subreddit=factor(levels = levels(df_with_acc$subreddit)),
  stringsAsFactors=TRUE
)

for (subreddit in levels(df_with_acc$subreddit)){
  df_subreddit <- df_with_acc[df_with_acc$subreddit == subreddit, ]
  for (interval in unique(df_subreddit$week_interval)){
    df_7day <- df_subreddit[df_subreddit$week_interval == interval, ]

    df_author_posts <- count(df_7day, vars = "author")

    df_author_score <- aggregate(df_7day$score, by=list(author=df_7day$author), FUN=sum)

    df_author <- merge(df_author_posts, df_author_score, by="author")
    df_author$score_per_post <- df_author$x / df_author$freq

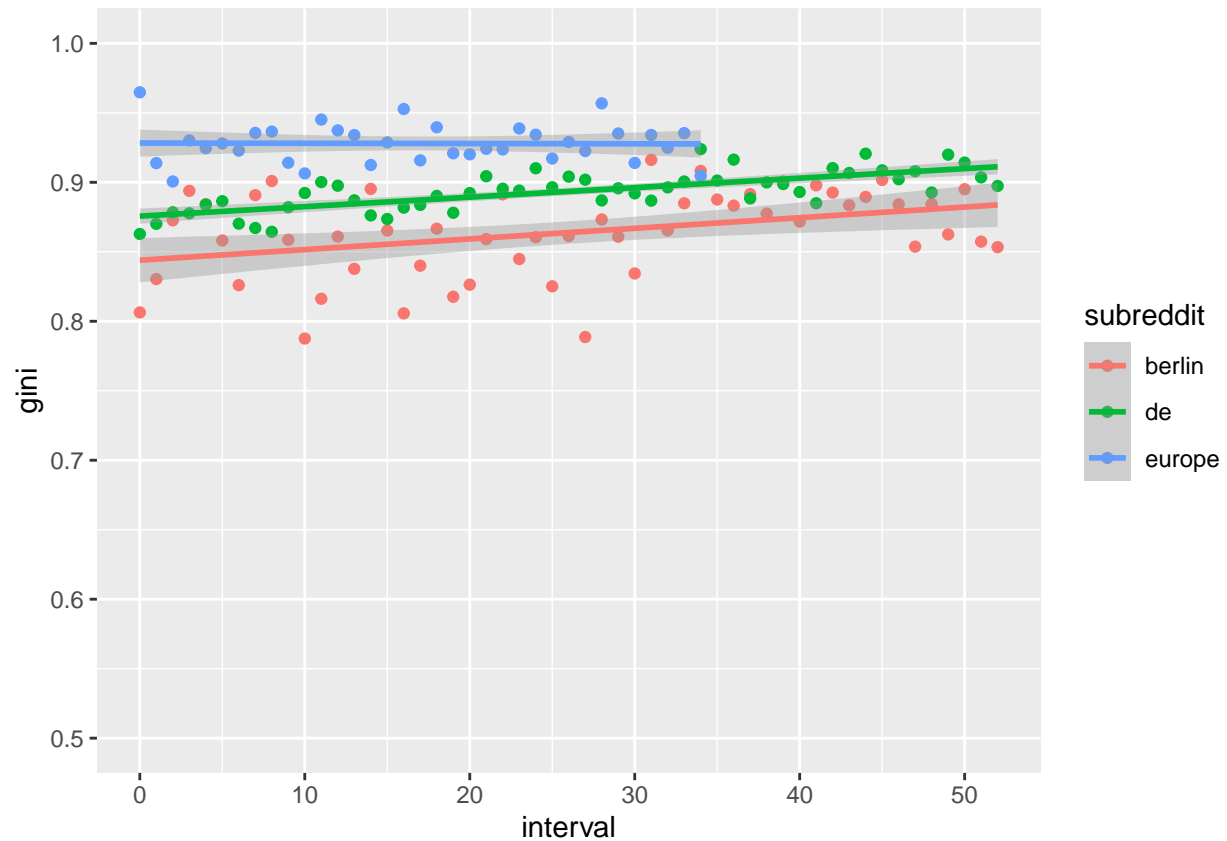
    gini <- Gini(df_author$score_per_post)
    gini_by_7day <- data.frame(
      interval=interval,
      gini=gini,
      subreddit=subreddit,
      stringsAsFactors=TRUE
    )

    gini_by_7days <- rbind(gini_by_7days, gini_by_7day)
  }
}

gini_by_7days$interval <-as.numeric(as.character(gini_by_7days$interval))

ggplot(gini_by_7days, aes(x=interval, y=gini, group = subreddit, color = subreddit)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x) +
  ylim(0.5,1)

```



```
gini_by_months <- data.frame(
  month=character(),
  gini=double(),
  stringsAsFactors=TRUE
)

for (month in unique(df_with_acc$month)){
  mnth_df <- df_with_acc[df_with_acc$month == month, ]

  df_author_posts <- count(mnth_df, vars = "author")

  df_author_score <- aggregate(mnth_df$score, by=list(author=mnth_df$author), FUN=sum)

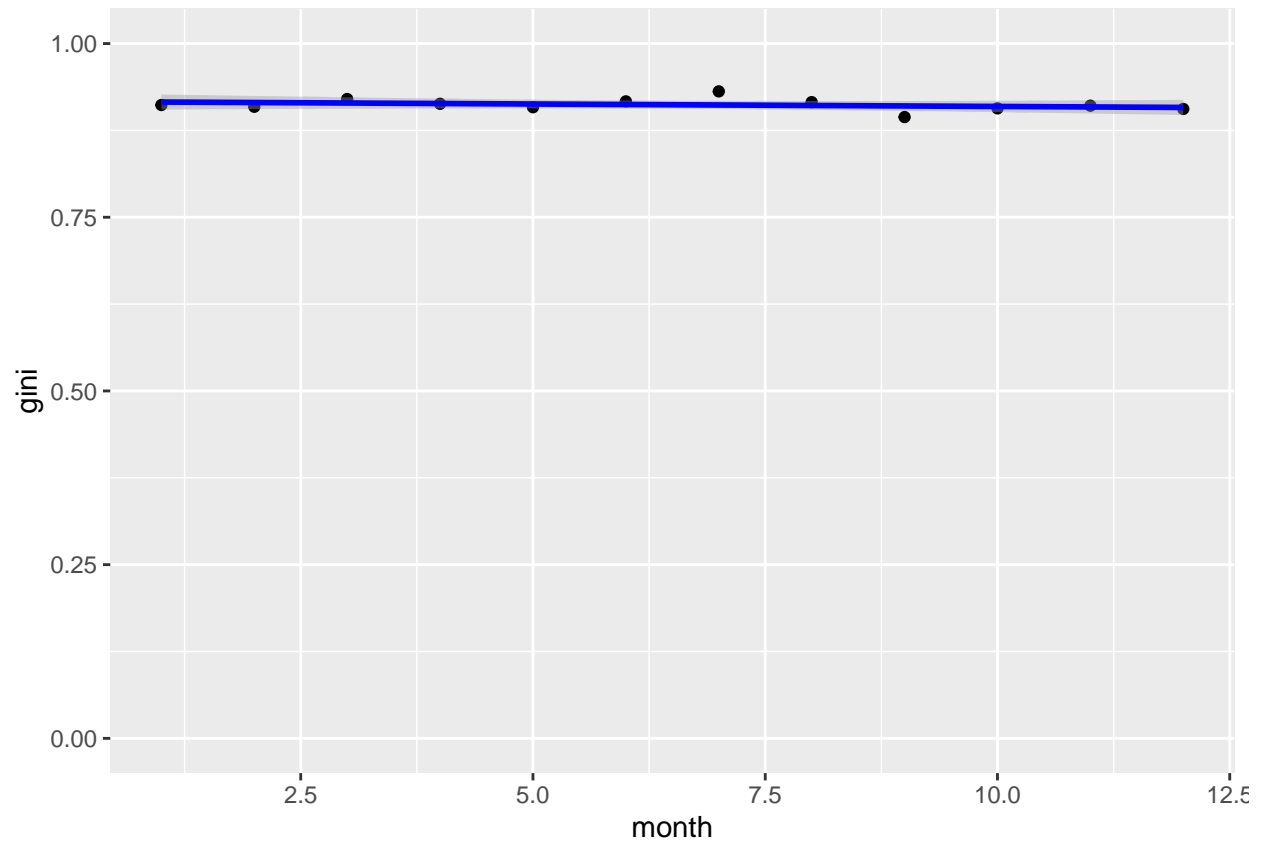
  df_author <- merge(df_author_posts, df_author_score, by="author")
  df_author$score_per_post <- df_author$x / df_author$freq

  gini <- Gini(df_author$score_per_post)
  gini_by_month <- data.frame(
    month=month,
    gini=gini,
    stringsAsFactors=TRUE
  )

  gini_by_months <- rbind(gini_by_months, gini_by_month)
}
```

```
gini_by_months$month <-as.numeric(as.character(gini_by_months$month))

ggplot(gini_by_months, aes(x=month, y=gini)) +
  geom_point() +
  geom_smooth(method='lm', formula= y~x, color = "blue") +
  ylim(0,1)
```



```
agg <- aggregate(df_with_acc$score, by=list(author=df_with_acc$author), FUN=sum)
#agg
Gini(agg$x)
```

```
## [1] 0.9347126
```

```
ggplot(agg, aes(x=author, y=x)) + geom_point()
```



