comments

Frederik Mann

15 7 2021

Comments

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

Import data set

```
df_europe <- read.csv("comments_europe_0.csv")
df <- rbind(df_europe)
```

Inspect

```
str(df)
```

```
## 'data.frame':    2425587 obs. of  25 variables:
##  $ all_awardings                  : chr  "[]" "[]" "[]" "[]" ...
##  $ associated_award               : chr  "" "" "" "" ...
##  $ author                         : chr  "rhinemanner" "easy_pie" "RVFullTime" "powellmd" ...
##  $ author_flair_text              : chr  "" "" "" "" ...
##  $ awarders                       : chr  "[]" "[]" "[]" "[]" ...
##  $ body                           : chr  "&gt; Her choice of words were pretty bad \\n\\nIndeed, my i
##  $ collapsed_because_crowd_control: logi  NA NA NA NA NA NA ...
##  $ created_utc                    : int  1580995430 1580995426 1580995417 1580995389 1580995347 1580
##  $ gildings                       : chr  "{}" "{}" "{}" "{}" ...
##  $ id                             : chr  "fgp2yyl" "fgp2ys7" "fgp2yd6" "fgp2x3m" ...
##  $ is_submitter                   : chr  "False" "True" "False" "True" ...
##  $ link_id                        : chr  "t3_ezr3cb" "t3_ezrq2w" "t3_ezc3jv" "t3_ezrmwv" ...
##  $ locked                         : chr  "False" "False" "False" "False" ...
##  $ no_follow                      : chr  "True" "True" "True" "True" ...
##  $ parent_id                      : chr  "t1_fgow2mj" "t1_fgp0szs" "t1_fgoqrzu" "t1_fgp2on5" ...
##  $ permalink                      : chr  "/r/europe/comments/ezr3cb/merkel_demands_reversal_of_far_ri
##  $ retrieved_on                   : int  1580995432 1580995428 1580995418 1580995390 1580995347 1580
##  $ score                          : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ send_replies                   : chr  "True" "True" "True" "True" ...
##  $ steward_reports                : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ stickied                       : chr  "False" "False" "False" "False" ...
```

```
##  $ subreddit                     : chr  "europe" "europe" "europe" "europe" ...
##  $ subreddit_id                  : chr  "t5_2qh4j" "t5_2qh4j" "t5_2qh4j" "t5_2qh4j" ...
##  $ total_awards_received         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ datetime                      : chr  "2020-02-06 14:23:50" "2020-02-06 14:23:46" "2020-02-06 14:
```

```
df$is_submitter <- as.logical(df$is_submitter)
df$send_replies <- as.logical(df$send_replies)
df$subreddit <- as.factor(df$subreddit)
df$no_follow <- as.logical(df$no_follow)
df$stickied <- as.logical(df$stickied)
df$subreddit <- as.factor(df$subreddit)
df$date <- as.Date(df$datetime)
```

Preprocess: Treat missing values, if applicable

```
df$author[df$author == ""] <- NA
df <-df[!duplicated(df$id), ]
nrow(df)
```

```
## [1] 2425587
```

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

```
## [1] 2425588
```

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

```
##                   all_awardings               associated_award
##                               0                              0
##                          author             author_flair_text
##                               0                              0
##                        awarders                           body
##                               0                              1
## collapsed_because_crowd_control                     created_utc
##                         2425587                              0
##                        gildings                             id
##                               0                              0
##                     is_submitter                        link_id
##                               0                              0
##                          locked                     no_follow
##                               0                              0
##                       parent_id                     permalink
##                               0                              0
##                    retrieved_on                         score
##                               0                              0
##                    send_replies                 steward_reports
##                               0                              0
##                        stickied                     subreddit
##                               0                              0
##                    subreddit_id          total_awards_received
##                               0                              0
##                        datetime                          date
##                               0                              0
```

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

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

```
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':    2419244 obs. of  28 variables:
##  $ all_awardings                 : chr  "[]" "[]" "[]" "[]" ...
##  $ associated_award              : chr  "" "" "" "" ...
##  $ author                        : chr  "rhinemanner" "easy_pie" "RVFullTime" "powellmd" ...
##  $ author_flair_text             : chr  "" "" "" "" ...
##  $ awarders                      : chr  "[]" "[]" "[]" "[]" ...
##  $ body                          : chr  "&gt; Her choice of words were pretty bad \\n\\nIndeed, my i
##  $ collapsed_because_crowd_control: logi  NA NA NA NA NA NA ...
##  $ created_utc                   : int  1580995430 1580995426 1580995417 1580995389 1580995347 1580
##  $ gildings                      : chr  "{}" "{}" "{}" "{}" ...
##  $ id                            : chr  "fgp2yyl" "fgp2ys7" "fgp2yd6" "fgp2x3m" ...
##  $ is_submitter                  : logi  FALSE TRUE FALSE TRUE FALSE FALSE ...
##  $ link_id                       : chr  "t3_ezr3cb" "t3_ezrq2w" "t3_ezc3jv" "t3_ezrmwv" ...
##  $ locked                        : chr  "False" "False" "False" "False" ...
##  $ no_follow                     : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ parent_id                     : chr  "t1_fgow2mj" "t1_fgp0szs" "t1_fgoqrzu" "t1_fgp2on5" ...
##  $ permalink                     : chr  "/r/europe/comments/ezr3cb/merkel_demands_reversal_of_far_ri
##  $ retrieved_on                  : int  1580995432 1580995428 1580995418 1580995390 1580995347 1580
##  $ score                         : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ send_replies                  : logi  TRUE TRUE TRUE TRUE TRUE TRUE ...
##  $ steward_reports               : chr  "NULL" "NULL" "NULL" "NULL" ...
##  $ stickied                      : logi  FALSE FALSE FALSE FALSE FALSE FALSE ...
##  $ subreddit                     : Factor w/ 1 level "europe": 1 1 1 1 1 1 1 1 1 1 ...
##  $ subreddit_id                  : chr  "t5_2qh4j" "t5_2qh4j" "t5_2qh4j" "t5_2qh4j" ...
##  $ total_awards_received         : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ datetime                      : chr  "2020-02-06 14:23:50" "2020-02-06 14:23:46" "2020-02-06 14:1
##  $ date                          : Date, format: "2020-02-06" "2020-02-06" ...
##  $ week_interval                 : num  5 5 5 5 5 5 5 5 5 5 ...
##  $ month                         : Factor w/ 12 levels "01","02","03",..: 2 2 2 2 2 2 2 2 2 2 ...
```

Data Visualisation

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

```
##    Var1   Freq
## 1    01 177590
## 2    02 161197
## 3    03 227739
## 4    04 209423
```

```
## 5    05 199603
## 6    06 226922
## 7    07 224272
## 8    08 198048
## 9    09 181809
## 10   10 210499
## 11   11 203962
## 12   12 198180
```

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

```
##    Var1  Freq
## 1     0 23482
## 2     1 35423
## 3     2 35599
## 4     3 37589
## 5     4 55925
## 6     5 33148
## 7     6 37426
## 8     7 38715
## 9     8 41480
## 10    9 54632
## 11   10 42274
## 12   11 55582
## 13   12 48634
## 14   13 59663
## 15   14 54770
## 16   15 43233
## 17   16 46422
## 18   17 43290
## 19   18 56229
## 20   19 41604
## 21   20 42559
## 22   21 42877
## 23   22 47879
## 24   23 60641
## 25   24 47727
## 26   25 51220
## 27   26 58177
## 28   27 47047
## 29   28 53233
## 30   29 49399
## 31   30 47036
## 32   31 44923
## 33   32 49435
## 34   33 41367
## 35   34 42253
## 36   35 48987
## 37   36 43482
## 38   37 40542
## 39   38 34203
## 40   39 46264
## 41   40 39584
## 42   41 45217
```

```
## 43   42 48562
## 44   43 59368
## 45   44 42572
## 46   45 50208
## 47   46 49062
## 48   47 49871
## 49   48 44145
## 50   49 46243
## 51   50 42599
## 52   51 45111
## 53   52 32331
```

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