Practice R

Belnikov Sergei

#Read in the data from `humansofnewyork.csv` into `R` and perform some sample tasks.

```
rm(list = ls())
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

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

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

```
library(tinytex)
```

Sample solutions

```
getwd()
```

```
## [1] "C:/Users/belni/Documents"
```

```
hony <- read.csv("C:\\Users\\belni\\Documents\\GitHub\\TAD_2021\\R lessons\\humansofnewyork.csv", string
str(hony) # examine its structure
```

```
## 'data.frame':    5835 obs. of  10 variables:
##  $ from_id       : num  1.02e+14 1.02e+14 1.02e+14 1.02e+14 1.02e+14 ...
##  $ from_name     : chr  "Humans of New York" "Humans of New York" "Humans of New York" "Humans of Ne
##  $ message       : chr  "Life settles you down." "All that peace and love will make you tense." "Boy
##  $ created_time  : chr  "2011-10-01T13:34:43+0000" "2011-10-01T16:49:23+0000" "2011-10-02T14:11:13+0
##  $ type          : chr  "photo" "photo" "photo" "photo" ...
##  $ link          : chr  "https://www.facebook.com/humansofnewyork/photos/a.102107073196735.4429.1020
##  $ id            : chr  "102099916530784_182302295177212" "102099916530784_182363265171115" "1020999
##  $ likes_count   : int  6977 550 1046 426 185 586 8441 1003 1159 364 ...
##  $ comments_count: int  27 13 7 13 3 12 50 22 14 7 ...
##  $ shares_count  : int  79 7 18 3 3 15 31 15 42 5 ...
```

```
glimpse(hony)
```

```
## Rows: 5,835
## Columns: 10
## $ from_id        <dbl> 1.020999e+14, 1.020999e+14, 1.020999e+14, 1.020999e+...
## $ from_name      <chr> "Humans of New York", "Humans of New York", "Humans ...
## $ message        <chr> "Life settles you down.", "All that peace and love w...
## $ created_time   <chr> "2011-10-01T13:34:43+0000", "2011-10-01T16:49:23+000...
## $ type           <chr> "photo", "photo", "photo", "photo", "photo", "photo"...
## $ link           <chr> "https://www.facebook.com/humansofnewyork/photos/a.1...
## $ id             <chr> "102099916530784_182302295177212", "102099916530784_...
## $ likes_count    <int> 6977, 550, 1046, 426, 185, 586, 8441, 1003, 1159, 36...
## $ comments_count <int> 27, 13, 7, 13, 3, 12, 50, 22, 14, 7, 26, 19, 11, 26,...
## $ shares_count   <int> 79, 7, 18, 3, 3, 15, 31, 15, 42, 5, 6, 25, 3, 57, 9,...
```

#1. How many status updates have been posted on this page?

```
table(hony$type) # type of facebook post
```

```
##
##   link  photo status  video
##     37   5672     83     43
```

```
sum(hony$type == "status")
```

```
## [1] 83
```

#2. What is the total number of likes, comments, and shares it received?

```
total.likes <- sum(hony$likes_count)
total.comm <- sum(hony$comments_count)
total.shares <- sum(hony$shares_count)
total.likes + total.comm + total.shares # wow!
```

```
## [1] 687316982
```

#3. What is the content of the post with the highest number of shares?

```
max(hony$shares_count) # maximum num shares
```

```
## [1] 363590
```

```
top.post <- which.max(hony$shares_count)
hony$message[top.post]
```

```
## [1] "Today I met an NYU student named Stella.  I took a photo of her.  Afterwards, she told me about
```

#4. What was the date in which the first photo was posted?

#The dates we have are characters, so we can't sort them. However, we can notice that the rows in our data frame are ordered by date. Thus let's find the first row that includes a photo.

```r
head(hony$created_time) # in order
```

```
## [1] "2011-10-01T13:34:43+0000" "2011-10-01T16:49:23+0000"
## [3] "2011-10-02T14:11:13+0000" "2011-10-02T23:41:38+0000"
## [5] "2011-10-03T13:14:46+0000" "2011-10-03T22:51:46+0000"
```

```r
tail(hony$created_time) # in order
```

```
## [1] "2015-11-29T19:50:39+0000" "2015-11-29T21:50:02+0000"
## [3] "2015-11-30T00:05:01+0000" "2015-11-30T16:03:27+0000"
## [5] "2015-11-30T19:15:01+0000" "2015-11-30T21:45:00+0000"
```

```r
first.photo <- min(which(hony$type == "photo"))
hony$created_time[first.photo] # October 1, 2011
```

```
## [1] "2011-10-01T13:34:43+0000"
```

#5. What is the total number of likes that the page has ever received, excluding its most popular post?

```r
max.likes <- max(hony$likes_count) # likes on most popular page
sum(hony$likes_count) - max.likes
```

```
## [1] 645303419
```

#6. How many posts have received more than 1,000,000 likes?

```r
sum(hony$likes_count > 1000000)
```

```
## [1] 15
```

#7. What was the total number of shares received by posts published each year?

```r
year <- substr(hony$created_time, 1, 4) # extracts year from date created variable
tapply(hony$shares_count, year, sum) # sum of shares by year
```

```
##     2011     2012     2013     2014     2015
##    88575   982605  3146146 10782456 10768177
```

#8. What was the total number of likes received by posts published each month?

```r
month <- substr(hony$created_time, 6, 7)# month from the date
?tapply
```

```
## starting httpd help server ... done
```

```r
tapply(hony$shares_count, month, sum)  # apply a sum function over range of array
```

```
##      01      02      03      04      05      06      07      08      09      10
## 1866757 1570954 1649598 2008699 1711300 2654737 2355755 2932212 3358422 2887256
##      11      12
## 1813703  958566
```