Zeru-Zhou-project11

November 18, 2021

1 Project 11 – Zeru Zhou

TA Help: NA

Collaboration: NA

- Get help from Dr. Ward's video

1.1 Question 1

[1]:
```
library(lubridate)

countries <- c('US', 'DE', 'CA', 'FR')
```

```
Attaching package: 'lubridate'


The following objects are masked from 'package:base':

    date, intersect, setdiff, union
```

[2]:
```
# EITHER use a for loop to create the data frame `yt`
yt <- data.frame()
for (c in countries) {
    filename <- paste0("/depot/datamine/data/youtube/", c, "videos.csv")
    dat <- read.csv(filename)
    dat$country_code <- c
    yt <- rbind(yt, dat)
}
```

[3]:
```
dim(yt)
```

1. 163394 2. 17

[4]:
```
# OR use an sapply function to create the data frame `yt`
myDFlist <- lapply( countries, function(c) {
```

```
                    dat <- read.csv(paste0("/depot/datamine/data/youtube/", c,␣
↪"videos.csv"))
                    dat$country_code <- c
                    return(dat)} )
yt <- do.call(rbind, myDFlist)
```

[5]:
```
dim(yt)
```

1. 163394 2. 17

[3]:
```
# convert columns to date formats
yt$trending_date <- ydm(yt$trending_date)
yt$publish_time <- ymd_hms(yt$publish_time)
```

[4]:
```
# extract the trending_year and publish_year
yt$trending_year <- year(yt$trending_date)
yt$publish_year <- year(yt$publish_time)
```

[5]:
```
count_tags <- function(tag_vector){
    length(strsplit(tag_vector, "|", fixed=TRUE)[[1]])
    }
```

[6]:
```
tag_test <- yt$tags[2]
tag_test
count_tags(tag_test)
```

'last week tonight trump presidency|last week tonight donald trump|john oliver trump|donald trump'

4

The function count_tags is created and the example has 4 unique tags.

1.2 Question 2

[7]:
```
yt$n_tags <- sapply(yt$tags, count_tags)
```

[18]:
```
head(yt)
```

A data.frame: 6 x 20

video_id <chr>	trending_date <date>	title <chr>
2kyS6SvSYSE	2017-11-14	WE WANT TO TALK ABOUT OUR MARRIAGE
1ZAPwfrtAFY	2017-11-14	The Trump Presidency: Last Week Tonight with John (
5qpjK5DgCt4	2017-11-14	Racist Superman | Rudy Mancuso, King Bach & Lele P
puqaWrEC7tY	2017-11-14	Nickelback Lyrics: Real or Fake?
d380meD0W0M	2017-11-14	I Dare You: GOING BALD!?
gHZ1Qz0KiKM	2017-11-14	2 Weeks with iPhone X

[16]:
```
US_DE <- subset(yt, (country_code=='US')|(country_code=='DE'))
```

[17]:
```
dim(US_DE)
```

1. 81789 2. 20

[19]:
```
US_DE$video_id[which.max(US_DE$n_tags)]
```

'4AelFaljd7k'

[21]:
```
US_DE$title[which.max(US_DE$n_tags)]
```

'TOP 20 SINGLE CHARTS 27. Dezember 2017 [FullHD]'

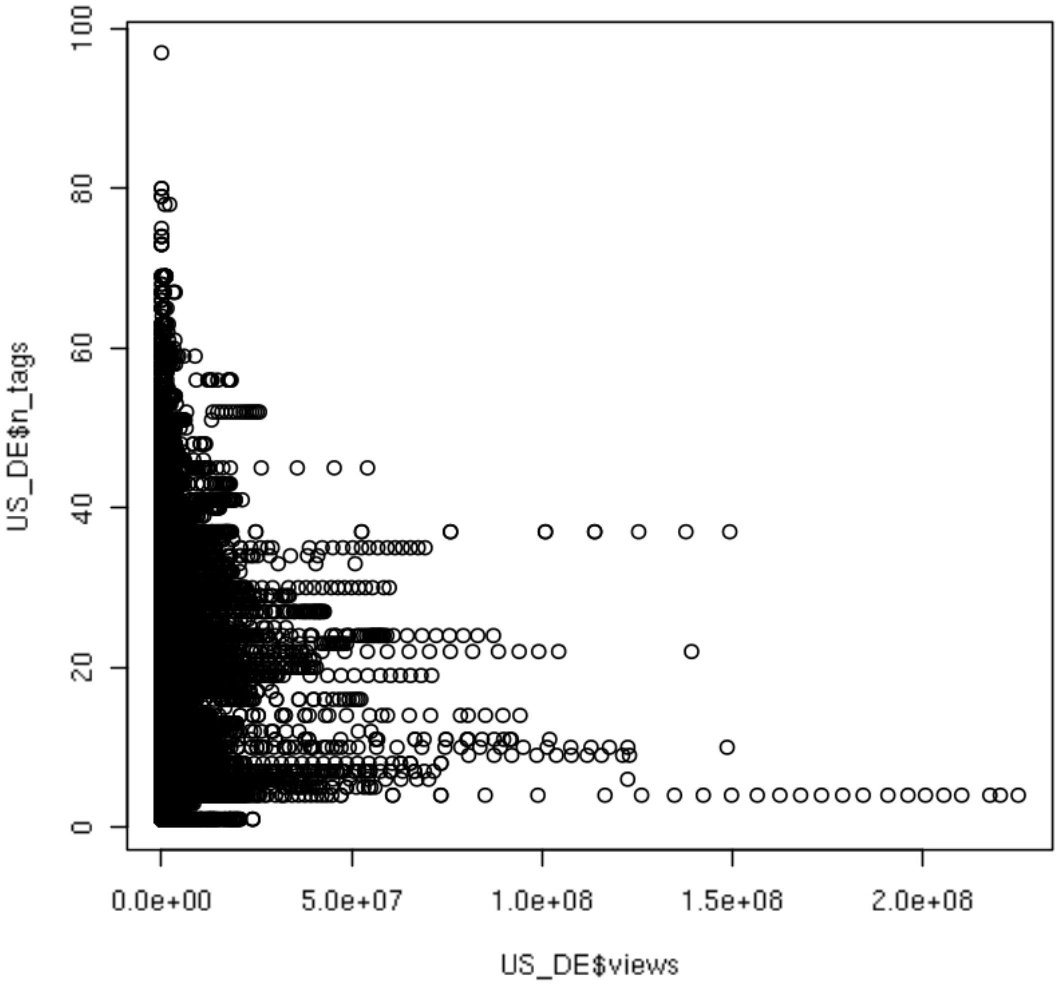
[20]:
```
US_DE$n_tags[which.max(US_DE$n_tags)]
```

97

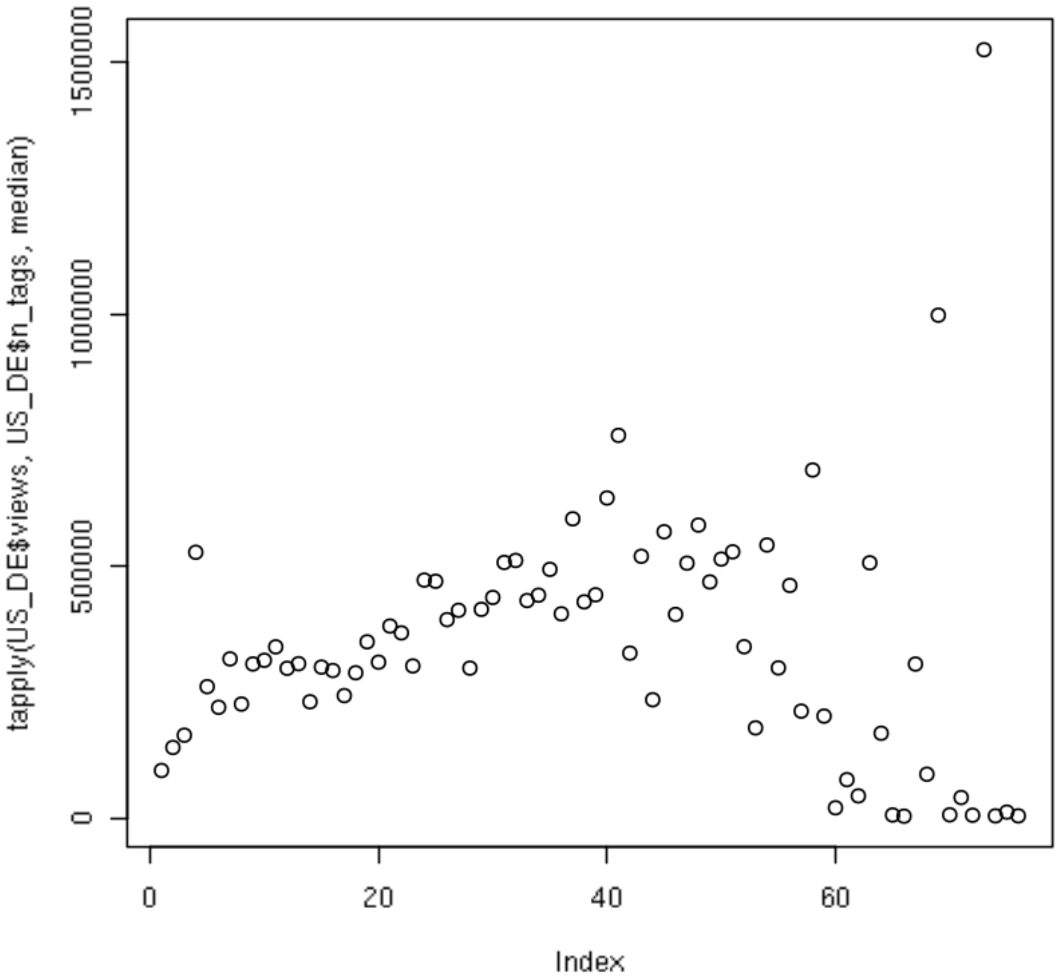
The title is 'TOP 20 SINGLE CHARTS 27. Dezember 2017 [FullHD]', and the number of tags it contains is 97.

1.3 Question 3

[22]:
```
plot(US_DE$views, US_DE$n_tags)
```



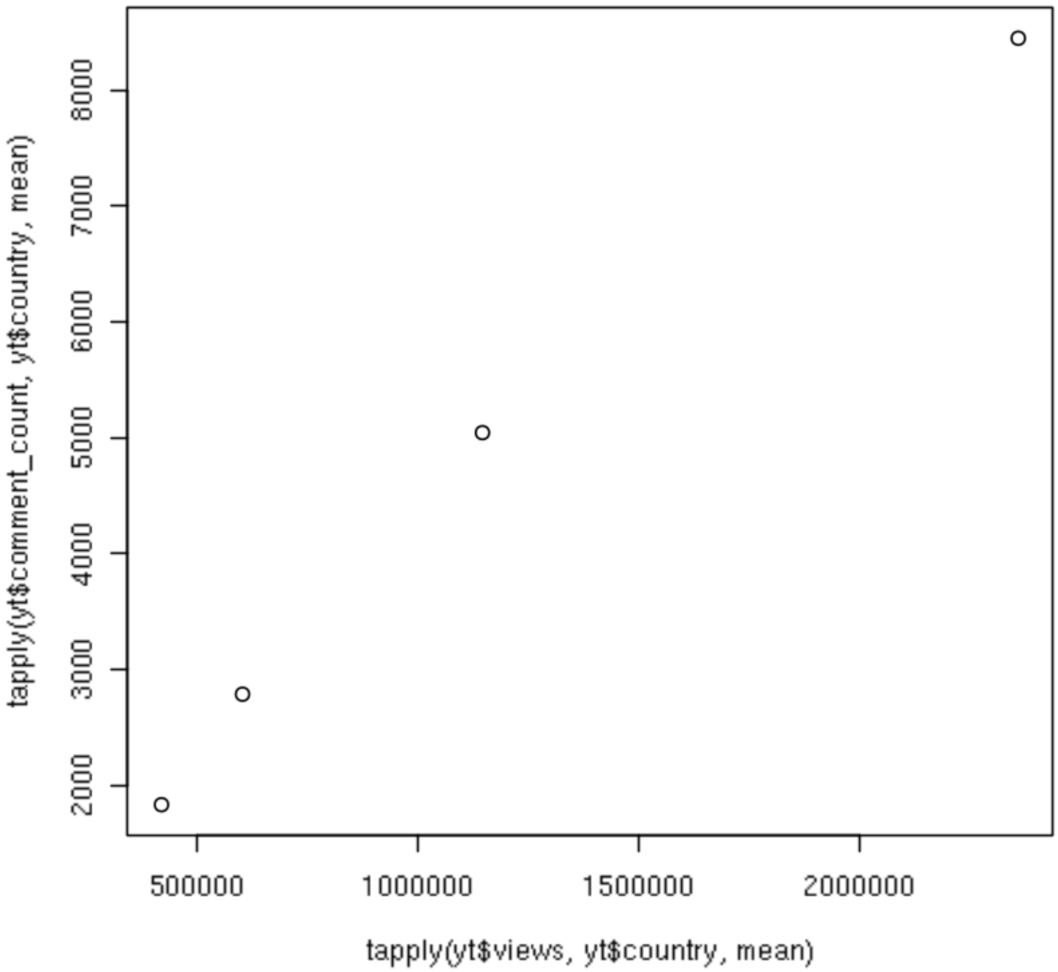
[24]:
```
plot(tapply(US_DE$views, US_DE$n_tags, median))
```



Scatter plot is a little bit messed up. I can only know that it is not a fully positive correlation. By using tapply function and use the median to evaluate the number of views under different number of tags, we can clearly see that when the number of tags is around 40, the median of the number of views is the highest.

1.4 Question 4

[8]:
```
plot(tapply(yt$views, yt$country, mean), tapply(yt$comment_count, yt$country,␣
↪mean))
```




```
[9]: tapply(yt$views, yt$country, mean)
```

CA 1147035.91078985 **DE** 603455.318437806 **FR** 419921.850604066 **US** 2360784.63825734

```
[10]: tapply(yt$comment_count, yt$country, mean)
```

CA 5042.97470707664 **DE** 2785.85651322233 **FR** 1832.45270602102 **US** 8446.80368262961

```
[11]: table(yt$country)
```

```
   CA    DE    FR    US 
40881 40840 40724 40949 
```

Here we compared the mean value of the views and number of comment with respect to different countries. We can see that the more the average views, the more the average number of comments. It is fair because we are comparing the mean value, and also the samples we collect for each country are approximately the same (almost 40900).

1.5 Question 5

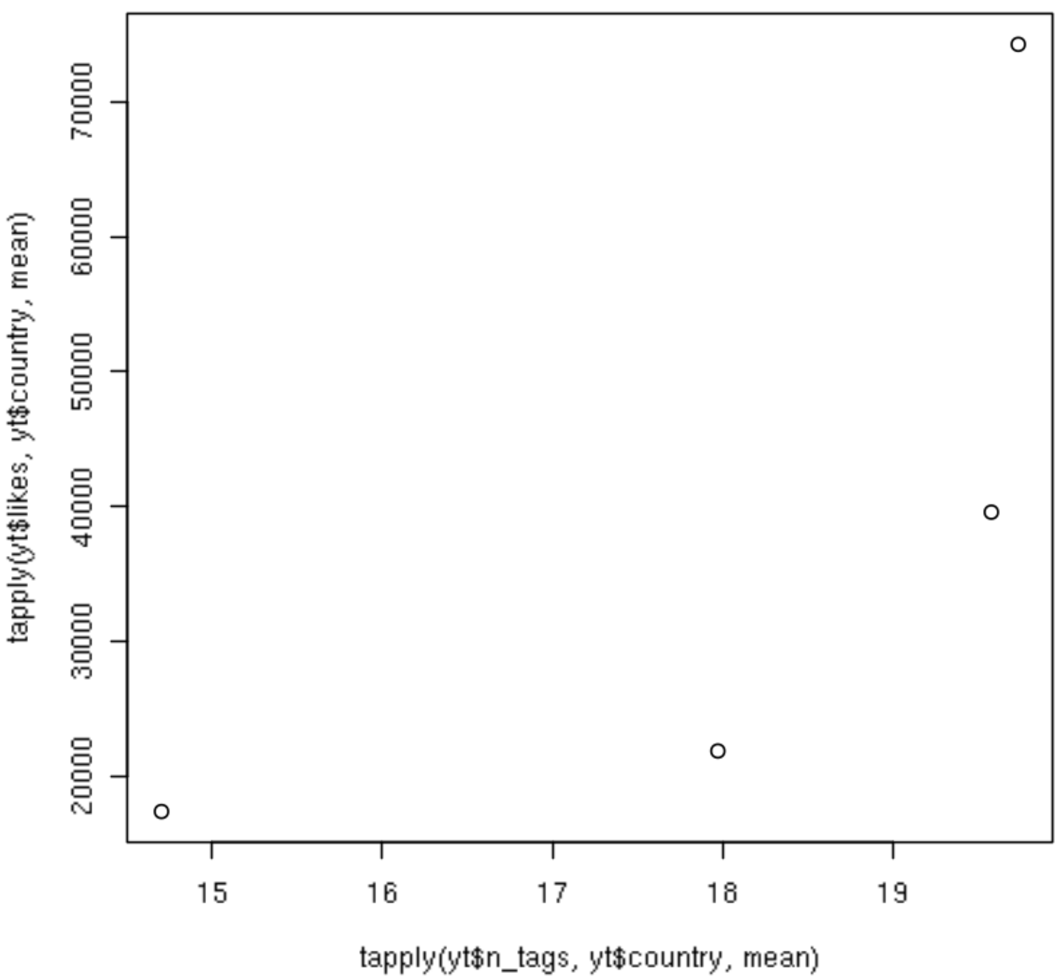
[12]:
```
tapply(yt$n_tags, yt$country, mean)
```

CA 19.5780925124141 **DE** 17.9715719882468 **FR** 14.7029024653767 **US** 19.7363305575228

[13]:
```
tapply(yt$likes, yt$country, mean)
```

CA 39582.6882414814 **DE** 21875.5028893242 **FR** 17388.8638149494 **US** 74266.7024347359

[14]:
```
plot(tapply(yt$n_tags, yt$country, mean), tapply(yt$likes, yt$country, mean))
```



My logic is that the more number of tags, there should be more "likes" because it is more conclusive with many tags. Let's compare those four countries: as we can see, this trend is perfectly applied on this dataset. The country with more average number of tags has more average number of likes. Also, to compare these four countries horizontally, the US has the largest average number of tags and likes, and France has the least average number of tags and likes.

1.6 Pledge

By submitting this work I hereby pledge that this is my own, personal work. I've acknowledged in the designated place at the top of this file all sources that I used to complete said work, including but not limited to: online resources, books, and electronic communications. I've noted all collaboration with fellow students and/or TA's. I did not copy or plagiarize another's work.

> As a Boilermaker pursuing academic excellence, I pledge to be honest and true in all that I do. Accountable together – We are Purdue.