Homework #8

Sang Doan

10/18/2020

Problem 1

(a)

```
dat <- all_processed_data()
dat %>%
  select(title, country) %>%
  head(n = 5)
```

```
##                                                     title country
## 1:            Eminem - Walk On Water (Audio) ft. Beyoncé      CA
## 2:                         PLUSH - Bad Unboxing Fan Mail      CA
## 3: Racist Superman | Rudy Mancuso, King Bach & Lele Pons      CA
## 4:                              I Dare You: GOING BALD!?      CA
## 5:           Ed Sheeran - Perfect (Official Music Video)      CA
```

(b)

```
pop_titles(dat) %>%
  tail(n = 3)
```

```
##      country                                            popTitle busiestDay
## 2178      US                          Suicide: Be Here Tomorrow.   18.31.01
## 2179      US                Boomerang Trick Shots | Dude Perfect   18.31.03
## 2180      US Childish Gambino - This Is America (Official Video)   18.31.05
##                                                              uniquepopTitle
## 2178 Did Alexa Lose Her Voice? - Teaser - Amazon Super Bowl Commercial LII
## 2179                                             Stray Kids District 9 M/V
## 2180                         The Spider and The Butterfly - Animated Short
```

(c)

```
countriesPop <- countries_pop(dat)
countriesPop %>%
  select(video_id, nCountry) %>%
  head(n = 3)
```

```
##      video_id nCountry
## 1 __-22AJoFxY        1
## 2 __-RHlreaec        1
## 3 __01xyhgG6M        1
```

(d)

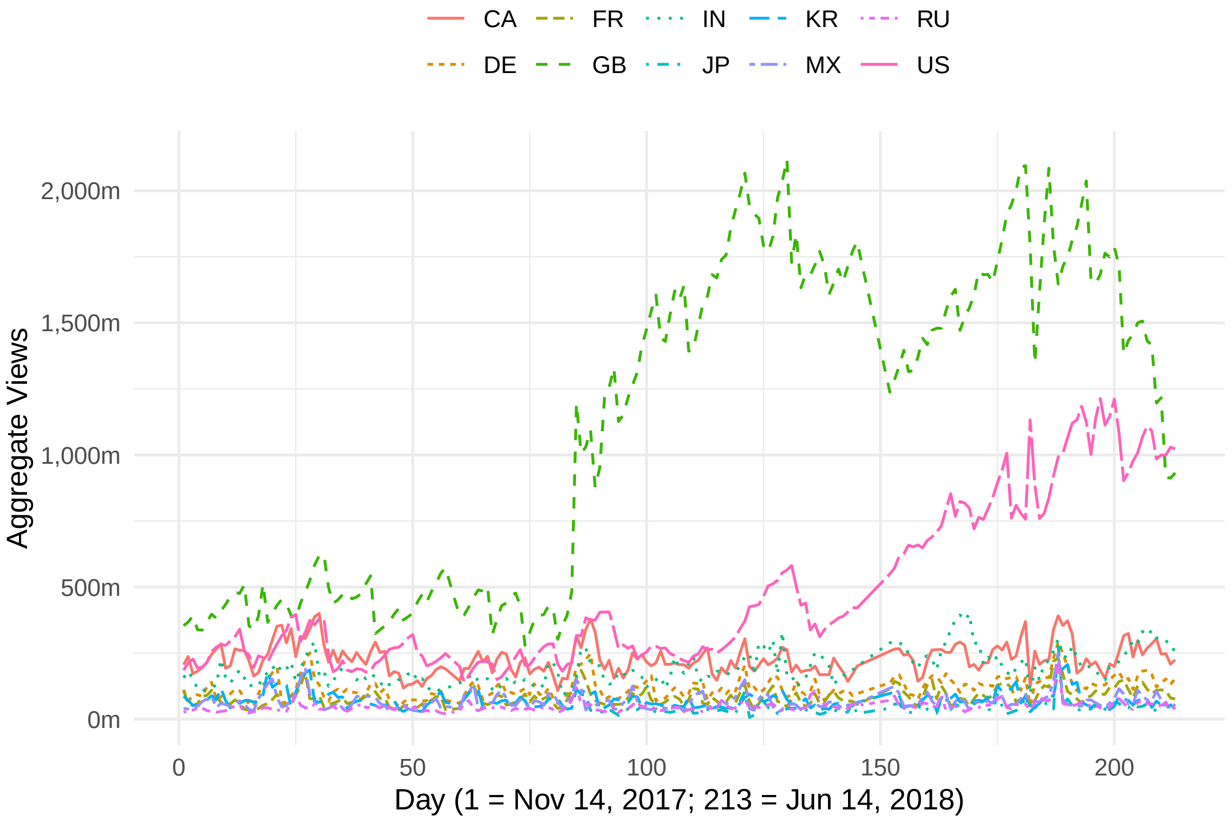
```
all_countries_pop(countriesPop) %>%
  head(n = 2)
```

```
##                                                                       title
## 1                           Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial
## 2 Nicky Jam x J. Balvin - X (EQUIS) | Video Oficial | Prod. Afro Bros & Jeon
```

Problem 2

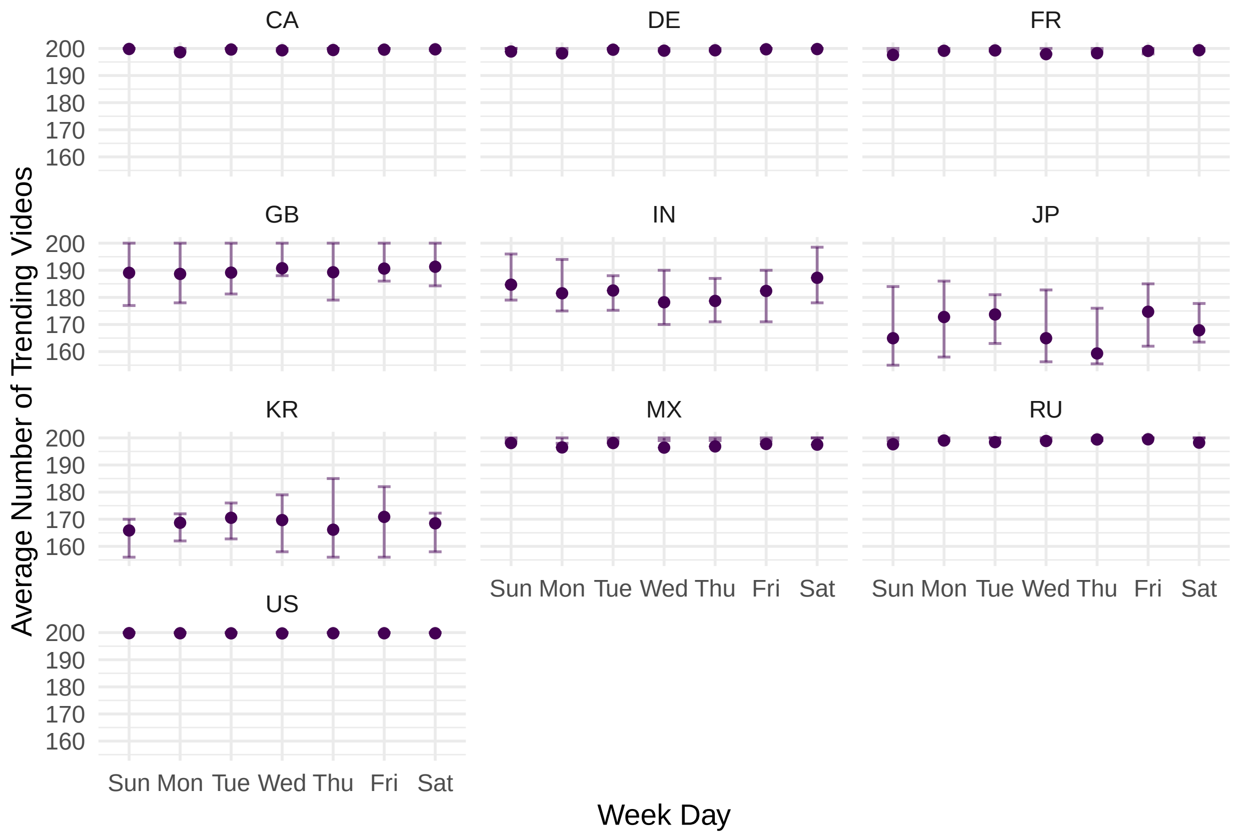
(a)

```
ggarrange(
  total_views_plot(dat),
  legend = 'top'
)
```



(b)

```
total_views_wday_plot(dat)
```



Code

data.R

```
# PROBLEM 1a

all_processed_data <- function() {
  dat <- list.files('../raw_data', full.names = T) %>%
    map_df(function(.thisFile) {
      thisDF <- fread(.thisFile)
      thisDF$country <- str_sub(.thisFile, -12, -11)
      return(thisDF)
    })

  return(dat)
}
```

analysis.R

```
# PROBLEM 1b

pop_titles <- function(dat) {
  dat <- unique_marking(dat)

  dat_out <- dat %>%
    ddply(.(country, trending_date), function(sdat) {
      popTitle <- sdat %>%
        slice_max(views) %>%
        select(country,
               popTitle = title,
               busiestDay = trending_date)

      uniqueTitle <- sdat %>%
        filter(unique == TRUE) %>%
        slice_max(views) %>%
        select(uniquepopTitle = title)

      sdat_out <- cbind(popTitle, uniqueTitle)

      return(sdat_out)
    }) %>%
    select(-trending_date)

  return(dat_out)
}

unique_marking <- function(dat) {
  #  Creates a logical column 'unique':
  #  TRUE if a title is unique to a country regardless of # of its ocurrences;
  #  FALSE otherwise.

  dat$unique <- FALSE
  titleRowNums <- dat[, .I[length(unique(country)) == 1], by = title]$V1
        #  This returns a dataframe with titles unique to one country.
        #  V1 is an auto-named column, containing row #s of such titles in the original data table.
        #  I came to choose data table because dplyr took forever to do the same task.
  dat[titleRowNums]$unique <- TRUE

  return(dat)
}

# PROBLEM 1c

countries_pop <- function(dat) {
  dat_out <- dat %>%
    select(video_id, title, country) %>%
    distinct() %>%
    ddply(.(video_id), function(sdat) {
      sdat_out <- data.frame(
        video_id = sdat$video_id,
        nCountry = length(sdat$country),
        title = sdat$title,
        stringsAsFactors = FALSE)

      return(sdat_out)
    })

  return(distinct(dat_out))
}

# PROBLEM 1d
```

```r
all_countries_pop <- function(countries_pop_dat) {
  titles <- countries_pop_dat %>%
    filter(nCountry == 10) %>%
    select(title)

  return(titles)
}

# PROBLEM 2a

day_count <- function(dat) {
  dat$dcount <- dat$trending_date %>% ydm %>% as.numeric - 17483
  return(dat)
}

total_views_plot <- function(dat) {
  dat <- dat %>% day_count
  ggplot(data = dat, aes(x = dcount, y = views, color = country, linetype = country)) +
    stat_summary(fun.y = 'sum', geom = 'line', position = 'identity') +
    theme_minimal() +
    labs(
      x = 'Day (1 = Nov 14, 2017; 213 = Jun 14, 2018)',
      y = 'Aggregate Views',
      color = '',
      linetype = ''
    ) +
    scale_y_continuous(labels = scales::number_format(
      accuracy = 1,
      scale = (1/(1000000)),
      big.mark = ',',
      suffix = 'm'
    ))
}

# PROBLEM 2b

total_views_wday_plot <- function(dat) {
  dat <- dat %>% day_count
  dat$weekday <- dat$trending_date %>% ydm %>% wday(label = TRUE)
  dat2 <- dat[, .(country, weekday, dcount)] %>%
    group_by(country, weekday) %>%
    mutate(upper =  quantile(table(dcount), 0.75),
           lower = quantile(table(dcount), 0.25),
           avgNum = mean(table(dcount))) %>%
    distinct(country, weekday, .keep_all = TRUE)

  dat2 %>%
    ggplot(aes(x = weekday, y = avgNum)) +
    geom_errorbar(
      aes(min = lower, max = upper),
      color = viridis(1, alpha = .5),
      width = .25
    ) +
    geom_point(
      size = 1.5,
      color = viridis(1)
    ) +
    labs(
      x = 'Week Day',
      y = 'Average Number of Trending Videos'
    ) +
    theme_minimal() +
    scale_color_viridis(discrete = T) +
    facet_wrap('country', ncol = 3, nrow = 4)
}
```

config.R

```
source('analysis.R')
source('data.R')

#Data Manipulation
library(plyr) #Load plyr before dyplr
library(tidyverse)
library(magrittr)
library(data.table)
library(lubridate)

#Data Communication
library(viridis)
library(ggpubr)
library(ggplot2)
```