Homework # 8

MATH 110

The dataset for this homework involves statistics for trending youTube vidoes across different countries. Read about the dataset here.

`https://www.kaggle.com/datasnaek/youtube-new`

The dataset is contained in 10 csv files, corresponding to 10 countries, found in the `youtubeData/` folder. Open up one of the csv files in Excel to see the different columns.

As in previous homeworks, all R scripts must contain only functions. All lines of code must be in an .Rmd and the code in the .Rmd must be simple and concise. Any involved manipulations of the data should be written as a function in the R script. For your homework, submit only a pdf file, but be sure to include all code.

1. (a) Write a function that reads in each of the csv files into a data.frame, adds a column `country` with value specifying the country of the csv file (e.g. "US"), rbinds all the data.frames into one large data.frame, which is then returned by the function. Do not repeat lines of code 10 times to do this.

   (b) Use `plyr::ddply` to construct a data.frame with the following columns: `country`, `popTitle`, `busiestDay`, `uniquepopTitle`. Each row corresponds to a particular country, as specified in the `country` column. For that country, the other columns give the title of the video with the most views in a single day (`popTitle`), the date of the day with the most views over all videos (`busiestDay`), and the title of the video with the most views in a single day that did not trend (i.e. is not in the dataset) in any other country (`uniquepopTitle`)

   (c) Use `plyr::ddply` to construct a data.frame with the columns `video_id`, `nCountry`, that gives the number of countries in which the video has trended (`nCountry`) for each video in the dataset.

   (d) Find the titles of the videos that trended in every country.

2. For these questions, use the R package `lubridate`. There are many good introductions to `lubridate` on the web.

   (a) The dataset starts with the date "17.14.11" (November 14, 2017) and ends with the date "18.14.06". Let the first date be day 1 and number the subsequent days $2, 3, \ldots$. You can use lubridate to convert the dates to numbers. See

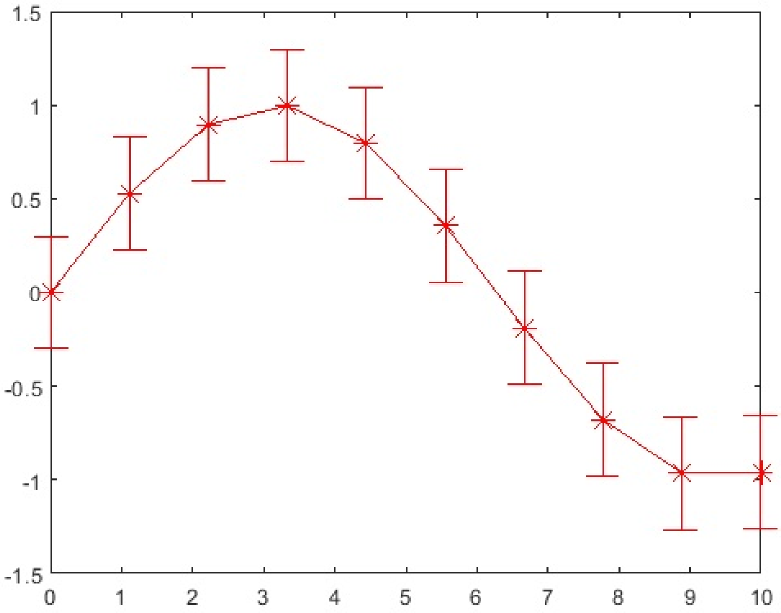
   `https://data.library.virginia.edu/working-with-dates-and-time-in-r-using-the-lubridate-package/`

   Use ggplot to produce a plot showing the total number of view over all videos on each day for each country. Your x-axis should be the day number (e.g. 1, 2,...), your y-axis should be the total views for a particular country.. The graph should be a line plot. Each country should have a separate line distinguished by both color and line type.

   (b) Produce a plot describing the number of videos viewed in each country on each day of the week (e.g. Sunday, Monday, etc.). You can use lubridate to determine the day of the week from the date. See second answer in

   `https://stackoverflow.com/questions/9216138/find-the-day-of-a-week`

   For each country, the form of the plot should resemble the plot shown below, except that the x-axis will give days of the week. The points of the graph should give the average number of videos viewed on the given day of the week (use `geom_point`). The whiskers above and below the points should give the 25% and 75% quantiles for the number of videos viewed on the given day of the week (use `geom_errorbar` along with the function `quantile`). Use `facet_wrap` to produce a separate plot for each country.


1.5
1
0.5
0
-0.5
-1
-1.5
0
1
2
3
4
5
6
7
8
9
10