Assignment 1: Solution

Karsten Donnay

14.06.2021

Contents

Assignment 1: RMarkdown **1**

Assignment 2: Import dataset **1**

Assignment 3: Insights **2**

```
library(knitr)

### Global options
options(max.print="150")
opts_chunk$set(echo=FALSE,
                 cache=FALSE,
               prompt=FALSE,
               tidy=TRUE,
               comment=NA,
               message=FALSE,
               warning=FALSE)
opts_knit$set(width=75)
rm(list = ls())
```

Assignment 1: RMarkdown

- Change the author of this file to your name.
  - Change the third line to your name
- Change the settings of this file so that the code is shown by default.
  - Change the parameter "code_folding:" in the 9th line to "show".

Assignment 2: Import dataset

In the data folder of this the first exercise session ("data/") you will find two datasets (twitter.Rda, guardian.Rda). These two datasets contain information retrieved from the respective APIs of the two platforms in 2020:

- **guardian.Rda**: articles of The Guardian obtained via its API. Contains information about an article, its author and its content.
- **twitter.Rda**: Tweets published by the account 'guardian' on Twitter. These Tweets often represent links to articles of The Guardian and contain information about the reactions to those articles (such as how many people favorited these statuses).

Assignment: Find a suitable variable to merge these two datasets on and then create a combined dataset that contains information about both the article characteristics (from the Guardian API) and the Twitter characteristics. Then use the **merge()** command to merge the two datasets. Report how many observations you lose from each original dataset.

Merge them by the 'link' variable

```
# Load
load("data/twitter.Rda")
load("data/guardian.Rda")

# Merge them by the link variable
combined <- merge(guardian, twitter, by = "link")
```

Twitter: 61.2% included, 38.7% lost (1579)

```
# Twitter
table(twitter$status_id %in% combined$status_id)
```

```
FALSE  TRUE
 1579  2498
```

```
table(twitter$status_id %in% combined$status_id)/sum(table(twitter$status_id %in%
    combined$status_id))
```

```
    FALSE      TRUE
0.3872946 0.6127054
```

```
# 61.2% included, 38.7% lost (1579)
```

Guardian: 67.3% included, 32.7% lost (1142)

```
# Guardian
table(guardian$id %in% combined$id)
```

```
FALSE  TRUE
 1142  2351
```

```
table(guardian$id %in% combined$id)/sum(table(guardian$id %in% combined$id))
```

```
    FALSE      TRUE
0.3269396 0.6730604
```

```
# 67.3% included, 32.7 lost (1142)
```

Assignment 3: Insights

In the combined data you just generated in the previous question:

- Does the page number where the article occurred in the Guardian newspaper have a positive or negative correlation to the number of retweets an article received?
- Do articles about music in our sample data get more or less frequently liked ("favorited") than sport articles?

We work with the "combined" data from the previous exercise or load the (identical) combined dataset (combined.Rda).

```
rm(list = ls())
load("data/combined.Rda")
```

Correlation between the Page number and retweets (negative)

```r
cor(combined$newspaperPageNumber, combined$retweet_count, use = "pairwise.complete.obs")
```

```
[1] -0.06685956
```

Average favorite count for music articles

```r
mean(combined[combined$sectionId == "music", ]$favorite_count, na.rm = T)
```

```
[1] 139.1321
```

Average favorite count for sports articles

```r
mean(combined[combined$sectionId == "sport", ]$favorite_count, na.rm = T)
```

```
[1] 38.70192
```

```r
# dplyr::top_n(aggregate(combined$favorite_count, list(Section =
# combined$sectionName), mean, na.rm=T), 10)
# dplyr::top_n(aggregate(combined$favorite_count, list(Section =
# combined$sectionName), mean, na.rm=T), -10)
```