Data\_Scientist\_R\_script.R

Xiawen Wang

2021-09-04

## set working directory  
# the path can be changed accordingly as user wish  
# the path need to contain the exercise file 'Data\_Scientist\_Exercise\_File.xlsx'  
setwd("~/Desktop/Data\_scientist\_R/instruction\_materials")  
  
## load library "openxlsx" (v4.2.4) and read in the Excel file  
library("openxlsx")  
# read the excel file, use the first row as header and do not skip empty rows  
exercise\_file <- read.xlsx(xlsxFile = "Data\_Scientist\_Exercise\_File.xlsx", sheet = 1, skipEmptyRows = FALSE)  
  
## normalise the contents in the DISEASE column  
exercise\_file$DISEASE <- replace(exercise\_file$DISEASE, exercise\_file$DISEASE == 'TRYPs', 'Trypanosomosis')  
exercise\_file$DISEASE <- replace(exercise\_file$DISEASE, exercise\_file$DISEASE == 'PPR', 'Peste des petits ruminants')  
  
## filter the dataframe, only entry with publication year later than 2010 (inclusive)  
exercise\_file <- exercise\_file[exercise\_file$YEAR\_PUBLICATION >= 2010,]  
  
## remove some columns   
exercise\_file <- subset(exercise\_file, select = -c(START\_DATE\_DATA, END\_DATE\_DATA))  
  
## reorder the dataframe   
exercise\_file <- exercise\_file[ , c(1, 2, 5, 6, 7, 3, 4)]   
# or  
new\_order <- c('IDENTIFIER', 'YEAR\_PUBLICATION', 'NUMBER\_POSITIVE', 'NUMBER\_TESTED', 'PERCENTAGE', 'STATE', 'DISEASE')  
exercise\_file <- exercise\_file[ , new\_order]   
  
## load library "tidyr" (v1.1.3) and split the IDENTIFIER column   
library("tidyr")  
# use comma that precedening with space and a number (regex: ',\\s[0-9]') as separator   
# make a new column AUTHOR to hold the new separated information  
# keep the IDENTIFIER column

# output into a sheet called exercise\_output\_file  
exercise\_output\_file <- separate(exercise\_file, IDENTIFIER, into = c('AUTHOR'), sep = ',\\s[0-9]', remove = FALSE, extra = "drop")  
  
## To check the result table  
print(exercise\_output\_file)

## IDENTIFIER AUTHOR  
## 1 Bekele et al, 2010a Bekele et al  
## 2 Tafese, Melaku and Fentahun, 2012 Tafese, Melaku and Fentahun  
## 3 Tafese, Melaku and Fentahun, 2012 Tafese, Melaku and Fentahun  
## 4 Tafese, Melaku and Fentahun, 2012 Tafese, Melaku and Fentahun  
## 5 Faris et al, 2012 Faris et al  
## 6 Faris et al, 2012 Faris et al  
## 7 Faris et al, 2012 Faris et al  
## 8 Faris et al, 2012 Faris et al  
## 9 Alemayehu et al, 2012 Alemayehu et al  
## 10 Alemayehu et al, 2012 Alemayehu et al  
## 11 Alemayehu et al, 2012 Alemayehu et al  
## 12 Ayele et al, 2012 Ayele et al  
## 13 Ayele et al, 2012 Ayele et al  
## 15 Ayele et al, 2012 Ayele et al  
## 16 Ayele et al, 2012 Ayele et al  
## 17 Tekle, 2013 Tekle  
## 18 Tekle, 2013 Tekle  
## 19 Tekle, 2013 Tekle  
## 20 Tekle, 2013 Tekle  
## 21 Bacha et al, 2013 Bacha et al  
## 22 Bacha et al, 2013 Bacha et al  
## 23 Bacha et al, 2013 Bacha et al  
## 24 Bacha et al, 2013 Bacha et al  
## 25 Bacha et al, 2013 Bacha et al  
## YEAR\_PUBLICATION NUMBER\_POSITIVE NUMBER\_TESTED PERCENTAGE STATE  
## 1 2010 71 323 22.0000000 SNNPR  
## 2 2012 10 114 8.7700000 Oromia  
## 3 2012 23 272 8.4600000 Oromia  
## 4 2012 NA 386 8.5490000 Oromia  
## 5 2012 0 74 0.0000000 Afar  
## 6 2012 1 271 0.3690037 Afar  
## 7 2012 1 360 0.2777778 Afar  
## 8 2012 21 1239 1.6949153 Afar  
## 9 2012 3 89 3.3700000 SNNPR  
## 10 2012 24 302 7.9400000 SNNPR  
## 11 2012 27 391 6.9000000 SNNPR  
## 12 2012 3 23 13.0434783 SNNPR  
## 13 2012 11 62 17.7419355 SNNPR  
## 15 2012 43 163 26.3803681 SNNPR  
## 16 2012 57 248 22.9838710 SNNPR  
## 17 2013 3 22 13.6400000 Oromia  
## 18 2013 16 152 10.5300000 Oromia  
## 19 2013 47 247 19.0300000 Oromia  
## 20 2013 66 424 15.5700000 Oromia  
## 21 2013 NA 410 6.8000000 Benishangul-Gumuz  
## 22 2013 13 200 6.5000000 Benishangul-Gumuz  
## 23 2013 15 210 7.1000000 Benishangul-Gumuz  
## 24 2013 23 276 8.3333333 Benishangul-Gumuz  
## 25 2013 5 134 3.7313433 Benishangul-Gumuz  
## DISEASE  
## 1 Trypanosomosis  
## 2 Trypanosomosis  
## 3 Trypanosomosis  
## 4 Trypanosomosis  
## 5 Peste des petits ruminants  
## 6 Peste des petits ruminants  
## 7 Peste des petits ruminants  
## 8 Peste des petits ruminants  
## 9 Peste des petits ruminants  
## 10 Peste des petits ruminants  
## 11 Peste des petits ruminants  
## 12 Trypanosomosis  
## 13 Trypanosomosis  
## 15 Trypanosomosis  
## 16 Trypanosomosis  
## 17 Trypanosomosis  
## 18 Trypanosomosis  
## 19 Trypanosomosis  
## 20 Trypanosomosis  
## 21 Trypanosomosis  
## 22 Trypanosomosis  
## 23 Trypanosomosis  
## 24 Trypanosomosis  
## 25 Trypanosomosis

## load library "writexl" (v1.4.0) and write the dataframe into an excel file  
library("writexl")  
# column names are kept, centered and bold by default in the output excel file   
write\_xlsx(exercise\_output\_file, "~/Desktop/Data\_scientist\_R/Data\_Scientist\_Exercise\_Output\_File.xlsx")  
  
## Summary numbers

# to print the IDENTIFIER column value with the highest value in the PERCENTAGE column  
exercise\_output\_file$IDENTIFIER[exercise\_output\_file$PERCENTAGE == max(exercise\_output\_file$PERCENTAGE)]

## [1] "Ayele et al, 2012"

# to print the sum of the values in the column NUMBER\_TESTED  
sum(exercise\_output\_file$NUMBER\_TESTED)

## [1] 6392

*###### EXTRA CODE ######*

*## to match to the example output file provided*

*#exercise\_output\_file$PERCENTAGE <- round(exercise\_output\_file$PERCENTAGE, digit = 0)*

*#exercise\_output\_file <- subset(exercise\_output\_file, select = -YEAR\_PUBLICATION)*