hw 4 add

justyn koenig

10/24/2021

#Question 1

```
if(!require("tidyverse")) install.packages("tidyverse")
```

```
## Loading required package: tidyverse
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.5     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.0.2     v forcats 0.5.1
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
if(!require("rvest")) install.packages("rvest")
```

```
## Loading required package: rvest
```

```
##
## Attaching package: 'rvest'
```

```
## The following object is masked from 'package:readr':
##
##     guess_encoding
```

```
library(dplyr)
library(rvest)
html <- read_html("https://introdatasci.dlilab.com/schedule_materials/")
x = html %>%
  html_table()
x
```

```
## [[1]]
## # A tibble: 30 x 5
##    Date   Topic                              Notes     HW    Reading
```

```
##    <chr>  <chr>                                <chr>     <chr> <chr>
##  1 Aug 24 About the course                     "\U0001f~ "-"   "Leek & Peng 2015"
##  2 Aug 26 Data science project cycle           "\U0001f~ ""    "Mason and Wiggins~
##  3 Aug 31 Class cancelled because of Hurric~   ""        ""    ""
##  4 Sep 2  Class cancelled because of Hurric~   ""        ""    ""
##  5 Sep 7  Introduction and install tools       "\U0001f~ ""    "Cooper & Hsing 20~
##  6 Sep 9  Version control with Git             "\U0001f~ ""    "Blischak et al. 2~
##  7 Sep 14 Introduction to GitHub               "\U0001f~ ""    ""
##  8 Sep 16 RStudio project and dynamic docum~   "\U0001f~ "01"  "Xie et al, Chapte~
##  9 Sep 21 The file system and basic unix sh~   "\U0001f~ ""    "Allesina & Wilmes~
## 10 Sep 23 R basics: data types, vectors, ma~   "\U0001f~ ""    ""
## # ... with 20 more rows
```

Question 2

```
Datasci <- x[[1]]

Datasci %>%
  separate(Date, into = c("Month", "Day"), sep = ' ') %>%
  select(Month, Day)
```

```
## # A tibble: 30 x 2
##    Month Day
##    <chr> <chr>
##  1 Aug   24
##  2 Aug   26
##  3 Aug   31
##  4 Sep   2
##  5 Sep   7
##  6 Sep   9
##  7 Sep   14
##  8 Sep   16
##  9 Sep   21
## 10 Sep   23
## # ... with 20 more rows
```

```
Datasci <- separate(Datasci, Date, into = c("Month", "Day"), sep = " ", remove = FALSE)
Datasci
```

```
## # A tibble: 30 x 7
##    Date   Month Day   Topic                      Notes    HW    Reading
##    <chr>  <chr> <chr> <chr>                      <chr>    <chr> <chr>
##  1 Aug 24 Aug   24    About the course           "\U0001~ "-"   "Leek & Peng 20~
##  2 Aug 26 Aug   26    Data science project cycle "\U0001~ ""    "Mason and Wigg~
##  3 Aug 31 Aug   31    Class cancelled because o~ ""       ""    ""
##  4 Sep 2  Sep   2     Class cancelled because o~ ""       ""    ""
##  5 Sep 7  Sep   7     Introduction and install ~ "\U0001~ ""    "Cooper & Hsing~
##  6 Sep 9  Sep   9     Version control with Git   "\U0001~ ""    "Blischak et al~
##  7 Sep 14 Sep   14    Introduction to GitHub     "\U0001~ ""    ""
##  8 Sep 16 Sep   16    RStudio project and dynam~ "\U0001~ "01"  "Xie et al, Cha~
```

```
##  9 Sep 21 Sep   21    The file system and basic~ "\U0001~ ""    "Allesina & Wil~
## 10 Sep 23 Sep   23    R basics: data types, vec~ "\U0001~ ""    ""
## # ... with 20 more rows
```

Code from lines 30 to 38 separated the date column into Month and Day, but not on the actual data table. To get the separation into month and day on the actual data table I used code on line 40 to generate the new columns.

Question 3

```
Datasci
```

```
## # A tibble: 30 x 7
##    Date   Month Day   Topic                      Notes    HW    Reading
##    <chr>  <chr> <chr> <chr>                      <chr>    <chr> <chr>
##  1 Aug 24 Aug   24    About the course           "\U0001~ "-"   "Leek & Peng 20~
##  2 Aug 26 Aug   26    Data science project cycle "\U0001~ ""    "Mason and Wigg~
##  3 Aug 31 Aug   31    Class cancelled because o~ ""       ""    ""
##  4 Sep 2  Sep   2     Class cancelled because o~ ""       ""    ""
##  5 Sep 7  Sep   7     Introduction and install ~ "\U0001~ ""    "Cooper & Hsing~
##  6 Sep 9  Sep   9     Version control with Git   "\U0001~ ""    "Blischak et al~
##  7 Sep 14 Sep   14    Introduction to GitHub     "\U0001~ ""    ""
##  8 Sep 16 Sep   16    RStudio project and dynam~ "\U0001~ "01"  "Xie et al, Cha~
##  9 Sep 21 Sep   21    The file system and basic~ "\U0001~ ""    "Allesina & Wil~
## 10 Sep 23 Sep   23    R basics: data types, vec~ "\U0001~ ""    ""
## # ... with 20 more rows
```

```
new_lec = Datasci %>%
  group_by(Month) %>%
  dplyr::summarise(lectures = n())
new_lec
```

```
## # A tibble: 5 x 2
##   Month lectures
##   <chr>    <int>
## 1 Aug          3
## 2 Dec          2
## 3 Nov          9
## 4 Oct          7
## 5 Sep          9
```

```
arrange(new_lec, desc(lectures))
```

```
## # A tibble: 5 x 2
##   Month lectures
##   <chr>    <int>
## 1 Nov          9
## 2 Sep          9
## 3 Oct          7
## 4 Aug          3
## 5 Dec          2
```

Question 4

```
strsplit(Datasci$Topic, " ", fixed = TRUE)
```

```
## [[1]]
## [1] "About"  "the"    "course"
##
## [[2]]
## [1] "Data"    "science" "project" "cycle"
##
## [[3]]
## [1] "Class"     "cancelled" "because"   "of"        "Hurricane" "Ida"
##
## [[4]]
## [1] "Class"     "cancelled" "because"   "of"        "Hurricane" "Ida"
##
## [[5]]
## [1] "Introduction" "and"          "install"      "tools"
##
## [[6]]
## [1] "Version" "control" "with"    "Git"
##
## [[7]]
## [1] "Introduction" "to"           "GitHub"
##
## [[8]]
## [1] "RStudio"   "project"   "and"       "dynamic"   "documents" "with"
## [7] "R"         "Markdown"
##
## [[9]]
## [1] "The"    "file"   "system" "and"    "basic"  "unix"   "shell"
##
## [[10]]
## [1] "R"        "basics:"  "data"     "types,"   "vectors," "matrix,"  "data"
## [8] "frame,"   "etc."
##
## [[11]]
## [1] "More"    "R"       "basics:" "lists,"  "dates,"  "etc."
##
## [[12]]
## [1] "R"           "programming" "basics:"     "conditional" "statements"
##
## [[13]]
## [1] "R"           "programming" "basics:"     "loops,"      "apply"
##
## [[14]]
## [1] "Strings"     "and"         "Regular"     "expressions"
##
## [[15]]
## [1] "API"      "and"      "data"     "scraping"
##
## [[16]]
```

```
## [1] "Data"   "input"  "and"    "output"
##
## [[17]]
## [1] "Data"         "manipulation" "with"         "R"
##
## [[18]]
## [1] "More"         "data"         "manipulation" "with"         "R"
##
## [[19]]
## [1] "Data"          "visualization" "with"          "R"
##
## [[20]]
## [1] "Exploratory" "data"        "analysis"
##
## [[21]]
## [1] "Regression" "methods"
##
## [[22]]
## [1] "More"       "on"         "Regression" "methods"
##
## [[23]]
## [1] "Write"     "your"      "own"       "functions"
##
## [[24]]
## [1] "Write"   "your"    "own"     "R"       "package"
##
## [[25]]
## [1] "Open"       "Science"    "and"        "automating" "things"
## [6] "with"       "Makefile"
##
## [[26]]
## [1] "Ethics"    "in"        "data"      "science"   "(virtual)"
##
## [[27]]
## [1] "Thanksgiving," "no"            "class"
##
## [[28]]
## [1] "Final"        "project"      "presentation"
##
## [[29]]
## [1] "Final"        "project"      "presentation" "and"          "wrap"
## [6] "up"
##
## [[30]]
## [1] "Final"  "grades" "due"
```

```
grep(" ", Datasci$Topic)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25
## [26] 26 27 28 29 30
```

```
grep(", ", Datasci$Topic)
```

```
## [1] 10 11 13 27
```

```
unlist(Datasci$Topic)
```

```
##  [1] "About the course"
##  [2] "Data science project cycle"
##  [3] "Class cancelled because of Hurricane Ida"
##  [4] "Class cancelled because of Hurricane Ida"
##  [5] "Introduction and install tools"
##  [6] "Version control with Git"
##  [7] "Introduction to GitHub"
##  [8] "RStudio project and dynamic documents with R Markdown"
##  [9] "The file system and basic unix shell"
## [10] "R basics: data types, vectors, matrix, data frame, etc."
## [11] "More R basics: lists, dates, etc."
## [12] "R programming basics: conditional statements"
## [13] "R programming basics: loops, apply"
## [14] "Strings and Regular expressions"
## [15] "API and data scraping"
## [16] "Data input and output"
## [17] "Data manipulation with R"
## [18] "More data manipulation with R"
## [19] "Data visualization with R"
## [20] "Exploratory data analysis"
## [21] "Regression methods"
## [22] "More on Regression methods"
## [23] "Write your own functions"
## [24] "Write your own R package"
## [25] "Open Science and automating things with Makefile"
## [26] "Ethics in data science (virtual)"
## [27] "Thanksgiving, no class"
## [28] "Final project presentation"
## [29] "Final project presentation and wrap up"
## [30] "Final grades due"
```

```
words <- unlist(Datasci$Topic)
words
```

```
##  [1] "About the course"
##  [2] "Data science project cycle"
##  [3] "Class cancelled because of Hurricane Ida"
##  [4] "Class cancelled because of Hurricane Ida"
##  [5] "Introduction and install tools"
##  [6] "Version control with Git"
##  [7] "Introduction to GitHub"
##  [8] "RStudio project and dynamic documents with R Markdown"
##  [9] "The file system and basic unix shell"
## [10] "R basics: data types, vectors, matrix, data frame, etc."
## [11] "More R basics: lists, dates, etc."
## [12] "R programming basics: conditional statements"
## [13] "R programming basics: loops, apply"
## [14] "Strings and Regular expressions"
## [15] "API and data scraping"
## [16] "Data input and output"
## [17] "Data manipulation with R"
```

```
## [18] "More data manipulation with R"
## [19] "Data visualization with R"
## [20] "Exploratory data analysis"
## [21] "Regression methods"
## [22] "More on Regression methods"
## [23] "Write your own functions"
## [24] "Write your own R package"
## [25] "Open Science and automating things with Makefile"
## [26] "Ethics in data science (virtual)"
## [27] "Thanksgiving, no class"
## [28] "Final project presentation"
## [29] "Final project presentation and wrap up"
## [30] "Final grades due"
```

```
freqtable <- sort(table(unlist(strsplit(Datasci$Topic, " "))), decreasing = T)
top5 <- paste0("(", paste0("\\b", names(freqtable)[1:5], "\\b", collapse = ""), ")", collapse = "|")
freqtable[1:5]
```

```
##
##       R     and    data    with basics:
##       9       8       6       6       4
```