hw_04

Samantha Rutledge

10/26/2021

```
tinytex::install_tinytex()
```

1. Use the rvest R package to scrape the schedule and materials table into R from the course webpage (https://introdatasci.dlilab.com/schedule_materials/). Read the documentation of rvest so you get a better idea about the functions provided by rvest and their usages

```
library(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()
```

```
library(rvest)
```

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

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

```
url_data <- "https://introdatasci.dlilab.com/schedule_materials/"
url_data %>%
  read_html()
```

```
## {html_document}
## <html lang="en">
## [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
## [2] <body>\n    <a href="#main">skip to content</a>\n    <noscript>\n  <style ...
```

```
css_selector <- "#main > table"
```

```
x<- url_data %>%
  read_html() %>%
  html_element(css = css_selector) %>%
  html_table()
x
```

```
## # 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
```

2. With the extracted data frame, create two new columns based on the Date column: month and day. month would be the month abbrevations from the Date column; day would be the numeric numbers from the Date column. Although you can use whatever approach to get this done (do not enter them by hand...), I suggest you try to practice regular expression here (sub() or stringr::str_extract()).

```
library(stringr)
x$day <- str_extract(x$Date, "\\d{2}")
x$month <- str_extract(x$Date, "\\D{3}")
```

```
x$day <- as.numeric(as.character(x$day))
x
```

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

3. With the data frame generated from Q2, use group_by() and summarise() to find out the number of lectures for each month, order the results by the number of lectures (high to low).

```
y <- x %>% group_by(month) %>% summarise(lecture_number = n()) %>% arrange(desc(lecture_number))
y
```

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

4. For the Topic column, split all values into words (hint: stringr::str_split()). Observe the values in the Topic column and use regular expression to specify the pattern in the stringr::str_split() or strsplit() function. Once this is done, you should get a list of list, you can use unlist() to convert it into a vector and name it as words. Use table() and sort() to find the top 5 most frequent words.

```
w <- strsplit(x$Topic, split = " ")
w
```

```
## [[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"
```

```
w1 <- unlist(w)
w1
```

```
##   [1] "About"         "the"           "course"        "Data"
##   [5] "science"       "project"       "cycle"         "Class"
##   [9] "cancelled"     "because"       "of"            "Hurricane"
##  [13] "Ida"           "Class"         "cancelled"     "because"
##  [17] "of"            "Hurricane"     "Ida"           "Introduction"
##  [21] "and"           "install"       "tools"         "Version"
##  [25] "control"       "with"          "Git"           "Introduction"
##  [29] "to"            "GitHub"        "RStudio"       "project"
##  [33] "and"           "dynamic"       "documents"     "with"
##  [37] "R"             "Markdown"      "The"           "file"
##  [41] "system"        "and"           "basic"         "unix"
##  [45] "shell"         "R"             "basics:"       "data"
##  [49] "types,"        "vectors,"      "matrix,"       "data"
##  [53] "frame,"        "etc."          "More"          "R"
##  [57] "basics:"       "lists,"        "dates,"        "etc."
##  [61] "R"             "programming"   "basics:"       "conditional"
##  [65] "statements"    "R"             "programming"   "basics:"
##  [69] "loops,"        "apply"         "Strings"       "and"
##  [73] "Regular"       "expressions"   "API"           "and"
##  [77] "data"          "scraping"      "Data"          "input"
##  [81] "and"           "output"        "Data"          "manipulation"
##  [85] "with"          "R"             "More"          "data"
##  [89] "manipulation"  "with"          "R"             "Data"
##  [93] "visualization" "with"          "R"             "Exploratory"
##  [97] "data"          "analysis"      "Regression"    "methods"
## [101] "More"          "on"            "Regression"    "methods"
## [105] "Write"         "your"          "own"           "functions"
## [109] "Write"         "your"          "own"           "R"
## [113] "package"       "Open"          "Science"       "and"
## [117] "automating"    "things"        "with"          "Makefile"
## [121] "Ethics"        "in"            "data"          "science"
## [125] "(virtual)"     "Thanksgiving," "no"            "class"
## [129] "Final"         "project"       "presentation"  "Final"
## [133] "project"       "presentation"  "and"           "wrap"
## [137] "up"            "Final"         "grades"        "due"
```

```
sort(table(w1),decreasing=TRUE)[1:5]
```

```
## w1
##       R     and    data    with basics:
##       9       8       6       6       4
```