BIOL 7800 - Nevyn Neal - Homework 4

N.N.

10/21/2021

Question 1

```
library(tidyverse)
```

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

## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     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
```

```
library(stringr)

page = read_html("https://introdatasci.dlilab.com/schedule_materials/")

table = page %>%
        html_nodes("table") %>%
        html_table(fill=TRUE)
table_df = data.frame(table)
table_df
```

```
##      Date                                                  Topic        Notes
## 1  Aug 24                                       About the course <U+0001F4D9>
## 2  Aug 26                             Data science project cycle <U+0001F4D9>
```

```
## 3  Aug 31                Class cancelled because of Hurricane Ida
## 4   Sep 2                Class cancelled because of Hurricane Ida
## 5   Sep 7                          Introduction and install tools <U+0001F4D9>
## 6   Sep 9                               Version control with Git <U+0001F4D9>
## 7  Sep 14                                 Introduction to GitHub <U+0001F4D9>
## 8  Sep 16   RStudio project and dynamic documents with R Markdown <U+0001F4D9>
## 9  Sep 21                    The file system and basic unix shell <U+0001F4D9>
## 10 Sep 23 R basics: data types, vectors, matrix, data frame, etc. <U+0001F4D9>
## 11 Sep 28                         More R basics: lists, dates, etc. <U+0001F4D9>
## 12 Sep 30            R programming basics: conditional statements <U+0001F4D9>
## 13  Oct 5                      R programming basics: loops, apply <U+0001F4D9>
## 14  Oct 7                         Strings and Regular expressions <U+0001F4D9>
## 15 Oct 12                                   API and data scraping <U+0001F4D9>
## 16 Oct 14                                   Data input and output <U+0001F4D9>
## 17 Oct 19                                Data manipulation with R <U+0001F4D9>
## 18 Oct 26                           More data manipulation with R <U+0001F4D9>
## 19 Oct 28                               Data visualization with R
## 20  Nov 2                               Exploratory data analysis
## 21  Nov 4                                      Regression methods
## 22  Nov 9                              More on Regression methods
## 23 Nov 11                                Write your own functions
## 24 Nov 16                                Write your own R package
## 25 Nov 18        Open Science and automating things with Makefile
## 26 Nov 23                          Ethics in data science (virtual)
## 27 Nov 25                                    Thanksgiving, no class
## 28 Nov 30                               Final project presentation
## 29  Dec 2                   Final project presentation and wrap up
## 30 Dec 14                                        Final grades due
##    HW                       Reading
## 1   -              Leek & Peng 2015
## 2            Mason and Wiggins 2010
## 3
## 4
## 5               Cooper & Hsing 2017
## 6              Blischak et al. 2016
## 7
## 8  01          Xie et al, Chapter 2
## 9     Allesina & Wilmes, Chapter 1
## 10
## 11                Hadley, Chapter 4
## 12 02
## 13
## 14 03              Peng, Chapter 17
## 15
## 16               Hadley, Chapter 11
## 17 04             Hadley, Chapter 5
## 18                Hadley, Chapter 5
## 19 05
## 20
## 21 06
## 22
## 23 07
## 24
## 25
```

```
## 26
## 27
## 28
## 29
## 30
```

Question 2

```
table_df$month = sapply(table_df$Date, str_extract, "[:alpha:]+")
table_df$day = sapply(table_df$Date, str_extract, "[:digit:]+")
table_df
```

```
##      Date                                                  Topic        Notes
## 1  Aug 24                                       About the course <U+0001F4D9>
## 2  Aug 26                             Data science project cycle <U+0001F4D9>
## 3  Aug 31            Class cancelled because of Hurricane Ida
## 4   Sep 2            Class cancelled because of Hurricane Ida
## 5   Sep 7                       Introduction and install tools <U+0001F4D9>
## 6   Sep 9                            Version control with Git <U+0001F4D9>
## 7  Sep 14                              Introduction to GitHub <U+0001F4D9>
## 8  Sep 16   RStudio project and dynamic documents with R Markdown <U+0001F4D9>
## 9  Sep 21                The file system and basic unix shell <U+0001F4D9>
## 10 Sep 23 R basics: data types, vectors, matrix, data frame, etc. <U+0001F4D9>
## 11 Sep 28                    More R basics: lists, dates, etc. <U+0001F4D9>
## 12 Sep 30         R programming basics: conditional statements <U+0001F4D9>
## 13  Oct 5                   R programming basics: loops, apply <U+0001F4D9>
## 14  Oct 7                      Strings and Regular expressions <U+0001F4D9>
## 15 Oct 12                                API and data scraping <U+0001F4D9>
## 16 Oct 14                                Data input and output <U+0001F4D9>
## 17 Oct 19                             Data manipulation with R <U+0001F4D9>
## 18 Oct 26                       More data manipulation with R <U+0001F4D9>
## 19 Oct 28                           Data visualization with R
## 20  Nov 2                           Exploratory data analysis
## 21  Nov 4                                  Regression methods
## 22  Nov 9                          More on Regression methods
## 23 Nov 11                            Write your own functions
## 24 Nov 16                            Write your own R package
## 25 Nov 18       Open Science and automating things with Makefile
## 26 Nov 23                     Ethics in data science (virtual)
## 27 Nov 25                               Thanksgiving, no class
## 28 Nov 30                          Final project presentation
## 29  Dec 2              Final project presentation and wrap up
## 30 Dec 14                                   Final grades due
##    HW                    Reading month day
## 1   -          Leek & Peng 2015   Aug  24
## 2        Mason and Wiggins 2010   Aug  26
## 3                                 Aug  31
## 4                                 Sep   2
## 5           Cooper & Hsing 2017   Sep   7
## 6           Blischak et al. 2016  Sep   9
## 7                                 Sep  14
## 8  01       Xie et al, Chapter 2  Sep  16
```

```
## 9     Allesina & Wilmes, Chapter 1   Sep  21
## 10                                   Sep  23
## 11               Hadley, Chapter 4   Sep  28
## 12 02                                Sep  30
## 13                                   Oct   5
## 14 03             Peng, Chapter 17   Oct   7
## 15                                   Oct  12
## 16              Hadley, Chapter 11   Oct  14
## 17 04            Hadley, Chapter 5   Oct  19
## 18               Hadley, Chapter 5   Oct  26
## 19 05                                Oct  28
## 20                                   Nov   2
## 21 06                                Nov   4
## 22                                   Nov   9
## 23 07                                Nov  11
## 24                                   Nov  16
## 25                                   Nov  18
## 26                                   Nov  23
## 27                                   Nov  25
## 28                                   Nov  30
## 29                                   Dec   2
## 30                                   Dec  14
```

Question 3

```
df1 = group_by(table_df, month)
df2 = summarise(df1, freq = n())
df2 = df2[order(-df2$freq),]
df2
```

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

Question 4

```
topic_lists = sapply(table_df$Topic, str_split, " ")
topics = sapply(topic_lists, str_extract, "[:alpha:]+")
words = unlist(topics)
top5 = tail(sort(table(words)), 5)
top5
```

```
## words
## project    data    with     and       R
##       4       6       6       8       9
```