Homework 4

1.(3 points) 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(rvest)
```

```
## Warning: package 'rvest' was built under R version 4.1.1
```

```
webpage_data <- read_html("https://introdatasci.dlilab.com/schedule_materials/")
table <- webpage_data %>%
  html_nodes(xpath='//*[@id="main"]/table') %>%
  html_table()
table <- table[[1]]
print(table)
```

```
## # A tibble: 30 x 5
##    Date   Topic                             Notes    HW    Reading
##    <chr>  <chr>                             <chr>    <chr> <chr>
##  1 Aug 24 About the course                  "\U0001~ "-"   "Leek & Peng 2015"
##  2 Aug 26 Data science project cycle        "\U0001~ ""    "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    "\U0001~ ""    "Cooper & Hsing 201~
##  6 Sep 9  Version control with Git          "\U0001~ ""    "Blischak et al. 20~
##  7 Sep 14 Introduction to GitHub            "\U0001~ ""    ""
##  8 Sep 16 RStudio project and dynamic docum~ "\U0001~ "01"  "Xie et al, Chapter~
##  9 Sep 21 The file system and basic unix sh~ "\U0001~ ""    "Allesina & Wilmes,~
## 10 Sep 23 R basics: data types, vectors, ma~ "\U0001~ ""    ""
## # ... with 20 more rows
```

2. (2 points) 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)
```

```
## Warning: package 'stringr' was built under R version 4.1.1
```

```r
month <- str_extract(table$Date, boundary("word"))
day <- str_extract(table$Date, "\\d?\\d")
table <- data.frame(table, month, day)
table
```

```
##      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  Holmes and Huber, Chapter 3   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
```

3. (2 points) 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).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
table %>%
    group_by(month) %>%
    summarise(lecture_count = n())%>%
    arrange(desc(lecture_count))
```

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

4. (3 points) 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.

```
topic_words <- unlist(str_split(table$Topic, boundary("word")))
word_occurance <- sort(table(topic_words), decreasing = TRUE)
head(word_occurance, 5)
```

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