Homework 4

Nathaniel Haulk

10/28/2021

Question 1

```r
sched = read_html("https://introdatasci.dlilab.com/schedule_materials")

sched = sched %>%
  html_elements("table") %>%
  html_table()

sched= as.data.frame(sched)
```

Question 2

```r
sched$Month = gsub("\\d","",sched[,1])

sched$Day = gsub("\\D","",sched[,1])

## Prints out schedule in a more readable dataframe containing just the month, day, and topic.
print(sched[c(6,7,2)])
```

```
##    Month Day                                                   Topic
## 1    Aug  24                                        About the course
## 2    Aug  26                              Data science project cycle
## 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
## 6    Sep   9                                Version control with Git
## 7    Sep  14                                  Introduction to GitHub
## 8    Sep  16   RStudio project and dynamic documents with R Markdown
## 9    Sep  21                    The file system and basic unix shell
## 10   Sep  23 R basics: data types, vectors, matrix, data frame, etc.
## 11   Sep  28                       More R basics: lists, dates, etc.
## 12   Sep  30            R programming basics: conditional statements
## 13   Oct   5                      R programming basics: loops, apply
## 14   Oct   7                         Strings and Regular expressions
## 15   Oct  12                                   API and data scraping
## 16   Oct  14                                   Data input and output
## 17  Oct   19                                Data manipulation with R
## 18  Oct   26                           More data manipulation with R
## 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
```

Question 3

```
lec.num = sched %>%
  group_by(Month) %>%
  summarise(num = n())

lec.num = lec.num[order(lec.num$num, decreasing = TRUE),]

print(lec.num)
```

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

Question 4

```
words = unlist(stringr::str_split(sched$Topic," "))

words = tolower(sub("[[:punct:]]","",words))


t.words = table(words)

t.words = sort(t.words, decreasing = TRUE)

top5 = as.data.frame(t.words[1:5])
```