Assignment 3

Nick Climaco

2023-02-08

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.1
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Exercise 1

Using the 173 majors listed in fivethirtyeight.com's College Majors dataset [https://fivethirtyeight.com/features/the-economic-guide-to-picking-a-college-major/], provide code that identifies the majors that contain either "DATA" or "STATISTICS"

```
data <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/college-majors/majors-li
head(data)
```

```
##   FOD1P                                Major                 Major_Category
## 1  1100                  GENERAL AGRICULTURE Agriculture & Natural Resources
## 2  1101 AGRICULTURE PRODUCTION AND MANAGEMENT Agriculture & Natural Resources
## 3  1102               AGRICULTURAL ECONOMICS Agriculture & Natural Resources
## 4  1103                     ANIMAL SCIENCES Agriculture & Natural Resources
## 5  1104                        FOOD SCIENCE Agriculture & Natural Resources
## 6  1105           PLANT SCIENCE AND AGRONOMY Agriculture & Natural Resources
```

```
# creating a subset of majors containig the desired key words
subset_data <- data %>%
    filter(grepl("DATA", Major) | grepl("STATISTICS", Major))
head(subset_data)
```

```
##   FOD1P                                        Major          Major_Category
## 1  6212 MANAGEMENT INFORMATION SYSTEMS AND STATISTICS                Business
## 2  2101      COMPUTER PROGRAMMING AND DATA PROCESSING Computers & Mathematics
## 3  3702               STATISTICS AND DECISION SCIENCE Computers & Mathematics
```

Excercise 2

Write code that transforms the data below: [1] “bell pepper” “bilberry” “blackberry” “blood orange” [5] “blueberry” “cantaloupe” “chili pepper” “cloudberry”
[9] “elderberry” “lime” “lychee” “mulberry”
[13] “olive” “salal berry” Into a format like this: c(“bell pepper”, “bilberry”, “blackberry”, “blood orange”, “blueberry”, “cantaloupe”, “chili pepper”, “cloudberry”, “elderberry”, “lime”, “lychee”, “mulberry”, “olive”, “salal berry”)

```
strStart <- '[1] "bell pepper" "bilberry"   "blackberry"  "blood orange"[5] "blueberry"  "cantaloupe"
```

```
create_fruits_vector <- function(string) {
    # using stringr package to get words inside double quotes
    fruits <- str_extract_all(string, "\"[^\"]+\"")
    # convert list to vector using unlist()
    fruits <- unlist(fruits)
    # replace double quotes with empty string
    fruits <- str_remove_all(fruits, "\"")
    # create a character vector for fruits
    fruit_vector <- c(fruits)

    return(fruit_vector)
}
```

```
# calling the function create_fruits_vector
create_fruits_vector(strStart)
```

```
##  [1] "bell pepper"  "bilberry"     "blackberry"   "blood orange" "blueberry"
##  [6] "cantaloupe"   "chili pepper" "cloudberry"   "elderberry"   "lime"
## [11] "lychee"       "mulberry"     "olive"        "salal berry"
```

Exercise 3

Describe, in words, what these expressions will match:

```
(.)\1\1
"(.)(.)\\2\\1"
(..)\1
"(.).\\1.\\1"
"(.)(.)(.).*\\3\\2\\1"
```