STA303: M2 tutorial activity

Professional skills for data analysis

Instructions

To participate in this activity you will need have to two windows readily available to you:

1) Your Zoom window

2) The Team Up! activity linked from Quercus in a browser window for voting.

I would recommend that one member of the team shares their screen with the rest of the team and shows this activity where you can see the question and options.

Note 1: In the Team Up! activity you will just see the letters for the questions, not the options themselves.

Note 2: There are hints for some questions, but if you get really stuck, please use the 'Ask for Help' option in Zoom.

Question 1

These are the two packages I like to use for web scraping.

```
library(polite)
library(rvest)
```

We're interested in scraping some data from the University of Toronto website. Specifically information on social media accounts: https://www.utoronto.ca/social-media-directory/all.

In the web scraping video for this week, we talked about checking the robots.txt. We can use the `polite` package to check, from within R, whether it seems like this part of the site is allowed to be scraped, based on the robots.txt.

```
## <polite session> https://www.utoronto.ca/social-media-directory/all
##     User-agent: class activity for STA303, sta303@utoronto.ca
##     robots.txt: 68 rules are defined for 1 bots
##    Crawl delay: 10 sec
##   The path is scrapable for this user-agent
```

Note: I am setting the `user_agent` to have our contact details in case the web master wanted to get in touch to tell us we were causing issues.

Use the output from the data section and/or go directly to the robots.txt for U of T (https://www.utoronto.ca/robots.txt). Based on this, can you figure out what aspect of ethical practice U of T has explicitly asked of us?

A) Not to scrape any of the website at all.
B) Limit the rate at which we scrape to 10 seconds per call.
C) Not to crawl any pages that have more than a 10 second loading delay.
D) No specific aspects of ethical practice explicitly asked of us.

Question 2

In addition to checking the robots.txt, what else do we need to consider as ethical scrapers?

A) Check if there is an API available instead.
B) Credit our source.
C) Only take what we need.
D) All of the above.

Question 3

Below is the code I used to get the data. THIS CODE IS NOT BEING ASSESSED IN STA303. You aren't responsible for understanding it, but I thought some of you might be interested. With the crawl delay it takes a while to run, so I have just given you the data directly in a csv file.

```
pages <- map(1:24, ~scrape(session, query = list(page=.x)) )

social_data <- map_dfr(pages, ~html_node(.x, css = ".view-content") %>%
  html_text() %>%
  str_split(pattern = "\\n") %>%
  unlist() %>%
  as_tibble() %>%
  mutate(value = str_trim(value)) %>%
  filter(value != "") %>%
  mutate(type = if_else(grepl("http", value), "link", "group")) %>%
  mutate(group_name = if_else(type == "group", value, NULL)) %>%
  fill(group_name) %>%
  filter(type == "link") %>%
  select(group_name, value) %>%
  rename(link = "value")) %>%
  mutate(platform = str_remove(link, "https\\:\\/\\/ca\\.")) %>%
  mutate(platform = str_remove(platform, "https\\:\\/\\/")) %>%
  mutate(platform = str_remove(platform, "http\\:\\/\\/")) %>%
  mutate(platform = str_remove(platform, "www\\.")) %>%
  mutate(platform = str_split_fixed(platform, "\\.", 2)[,1])

write_csv(social_data, "scraped_data.csv")
social_data <- read_csv("scraped_data.csv")
```

Write code to determine what the most common social media platform used by U of T schools/departments/groups etc. is

```
## Rows: 484
## Columns: 3
## $ group_name <chr> "Centre for Diaspora and Transnational Studies", "Centre fo~
## $ link       <chr> "https://www.facebook.com/UofTCDTS/", "https://www.instagra~
## $ platform   <chr> "facebook", "instagram", "twitter", "facebook", "instagram"~
```

A) Facebook
B) Instagram
C) Twitter
D) YouTube

Question 4

Suppose when enrolling in STA303 you were randomly assigned to either a mandatory synchronous class or a totally asynchronous pre-recorded class and then final grades were compared between the two groups. Which ONE of the following is TRUE?

A) Any differences in grades between the groups must be due to random chance because participants were randomly assigned.
B) Random assignment hopefully means the two groups are comparable across potential confounding variables, like previous preparation or convenience of time zone.
C) The fact the students have been randomly assigned means we could use this data as an empirical version of a null distribution for a hypothesis test.
D) As we have observed both the final grades and the group students were in, this is an example of a case-control study.

Question 5

Suppose a company wanted to understand how remote working was affecting their employees. In one of their staff surveys they had asked employees to rate their current sleep quality. 30 employees with generally poor sleep quality and 30 employees with generally excellent sleep quality were then invited to be part of a further study where they were asked whether or not they usually worked on their computer within 2 hours of their bedtime. The goal of the study was to understand if being exposed to the computer close to bedtime was associated with poor sleep quality. What kind of study is this?

A) Randomized control trial.
B) Prospective cohort study.
C) Retrospective cohort study.
D) Case-control study.

Question 6

Suppose U of T is currently recruiting as study participants students graduating with one of the following degrees in 2021: Statistics Specialist, Data Science Specialist, Actuarial Science Specialist. The goal of the study is to identify what students with these degrees are earning 5 years after graduating (2026) and whether there were any differences in their incomes by program. What kind of study is this?

A) Randomized control trial.
B) Prospective cohort study.
C) Retrospective cohort study.
D) Case-control study.

Question 7

Which ONE of the following statements about p-values is TRUE?

A) Suppose the P-value for a one sample t-test we conduct is 0.67. This means we can accept the null hypothesis for this test.
B) The size of a P-value gives us an indication of how surprised we are to observe a test statistic like ours, under the assumption that our null and alternative hypotheses are correct.
C) A small p-value is calculated when the area under in the tails(s) of our test distribution's probability density curve (e.g. a t-distribution), is small. The size of the tails are determined by the difference between the population parameter and the test statistic.
D) Saying you 'fail to reject the null hypothesis' is not the same as saying 'the null hypothesis is true'.