Homework 10

Peyton Hall

03/28/2024

```
rm(list=ls())
```

```
library(rvest)
```

```
# 1. This website https://en.wikipedia.org/wiki/2016_Summer_Olympics recorded
#    the medal counts for the 2016 Summer Olympics. Scrape the table 11 to find
#    the medal counts

url <- "https://en.wikipedia.org/wiki/2016_Summer_Olympics"
page <- read_html(url) # Read the HTML content
# Scrape the table with the caption "2016 Summer Olympics medal table"
table_11 <- page %>%
  html_nodes("table.wikitable:contains('2016 Summer Olympics medal table')") %>%
  html_table()
table_11
```

```
## [[1]]
## # A tibble: 12 x 6
##    Rank                NOC                  Gold Silver Bronze Total
##    <chr>               <chr>               <int>  <int>  <int> <int>
##  1 1                   United States          46     37     38   121
##  2 2                   Great Britain          27     23     17    67
##  3 3                   China                  26     18     26    70
##  4 4                   Russia                 19     17     20    56
##  5 5                   Germany                17     10     15    42
##  6 6                   Japan                  12      8     21    41
##  7 7                   France                 10     18     14    42
##  8 8                   South Korea             9      3      9    21
##  9 9                   Italy                   8     12      8    28
## 10 10                  Australia               8     11     10    29
## 11 11-86               Remaining NOCs        124    150    181   455
## 12 Totals (86 entries) Totals (86 entries)   306    307    359   972
```

```
# 2. A data analyst received permission to post a data set that was scraped from
#    a social media site. The full data set included name, screen name, email
#    address, geographic location, IP address, demographic profiles, and
#    preferences for relationships.
# a. Why might this be a possible ethical issue?
# This could be a possible ethical issue because the data was scraped without
# consent from the users of the social media site. Users typically expect their
# information to be used only within each platform for specific purposes.
# b. Can the de-identified data be re-identified?
# The de-identified data might still be re-identified. Even if the original data
# was stripped of direct identifiers like names or e-mail addresses, other
# identifiers like geographic location or IP address, when combined with
# external data sources, could potentially lead to the re-identification. In
# summary, while de-identification can reduce the risk, it may not completely
# eliminate the possibility of re-identification.
```

```
# 3. A company uses a machine learning algorithm to determine which job
#    advertisement to display for users searching for technology jobs. Based on
#    past results, the algorithm tends to display lower-paying jobs for women
#    than for men.
# c. What was the ethical issue?
# The ethical issue is that the opposing genders are being shown non-equivalent
# job salaries on average. For jobs which require the same task, it would be
# most logical for the same pay to be provided regardless of gender.
# d. What can we do to avoid any ethical issues?
# Ensure the algorithm is evaluated and adjusted properly. Ensure that the
# training data used for the algorithm is diverse and representative of the
# population it serves. Implement regular monitoring and auditing processes to
# assess the algorithm's performance and detect any biases or discriminatory
# patterns.
```

```
# 4. A reporter carried out a clinical trial of chocolate where a small number
#    of overweight subjects (16 subjects) who had received medical clearance and
#    consented were randomized to either eat dark chocolate or milk chocolate.
#    They were followed for a period, and their change in weight was recorded
#    from the bass line until the end of the study. They found that there is a
#    significant weight gain difference between the dark chocolate group and the
#    milk chocolate group. This study was publicized and received coverage from
#    a number of magazines and television programs.
# e. What can be the potential problems with this study?
# The potential problems could include a small sample size, a lack of diversity,
# randomization issues (i.e. in terms of the subjects being from the same
# institution), publication bias, and generalizations due to the subjects
# having already had issues prior to receiving medical clearance.
# f. How could we implement this study to avoid this issue?
# Increase the sample size, ensure there is diversity (for the sake of data),
# randomize each subject to a different treatment group (i.e. dark or milk
# chocolate), and potentially extend the duration of the study.
```

```
# 5. Use the data Donation Data on D2L to complete a) to c). The variable
#    Donation ID shows the ID of each donation. The variable Type shows the type
#    of the donation. The variable Status shows whether a donation is complete
#    or failed, and the last variable shows the donation amount.
# Complete the following and submit your Excel file to the drop box.
# a) Create a pivot table to show the average donation amount of each type.
# b) Create a pivot graph to show the average donation amount of each type.
# c) Create a pivot table to keep only the completed donations and show the
#    average donation of each type.
# 6. Use the USTrading2021 data on D2L. The data recorded the trading between
#    the U.S. and other countries during 2021. The variables, trade flow, and
#    trade flow code, represent if the trading is an import or export or a
#    re-export. The variable Partner is the country traded with the United
#    States, and the Trade Value is the amount of trade between that country and
#    the U.S. Use the data to do the following:
# a) Create a pivot table to show the mean import trading values for each
#    Partner country, using the filter to keep only the imports. Filter to
#    remove world and blank as Partners.
# b) Sort the trade values from the largest to the smallest using the pivot
#    table in a) to only show the top 3 countries.
# c) Create a pivot graph to show the mean import trading values for each
#    Partner country. Do not include world or blank.
```