

HW 3 - Ethics + recap- Suggested Answers

! Important

This homework is due Friday, Oct 14 at 11:59pm ET.

! Important

Homeworks are to be turned in individually as usual (different from labs)

Getting started

- Go to the [sta199-f22-2](#) organization on GitHub. Click on the repo with the prefix **hw-03**. It contains the starter documents you need to complete the homework assignment.
- Clone the repo and start a new project in RStudio. See the [Lab 0 instructions](#) for details on cloning a repo and starting a new R project.

Workflow + formatting

Make sure to

- Update author name on your document.
- Label all code chunks informatively and concisely.
- Follow the [Tidyverse style guide](#).
- Make at least 3 commits.
- Resize figures where needed, avoid tiny or huge plots.
- Use informative labels for plot axes, titles, etc.
- Turn in an organized, well formatted document.

Packages

We'll use the **tidyverse** package for much of the data wrangling and visualization, though you're welcomed to also load other packages as needed.

```
library(tidyverse)
```

Exercises

Exercises 1 and 2 are review exercises based on common questions that came up during the exam.

Exercise 1

All about Quarto:

- a. For each of the character strings below, determine if the string is an proper code chunk label to use in a document **when rendering to PDF**. If not, explain why. You're welcomed to try them out to check.

- Chunk 1:

```
#| label: label with spaces
```

- Chunk 2:

```
#| label: reaaaaaaaaaalllllllllyyyyy-long-label  
#|           with-line-breaks
```

- Chunk 3:

```
#| label: 1-label-starting-with-number
```

- Chunk 4:

```
#| label: label-with-dashes
```

- b. What values does each of the following chunk options take and what do they do?
 - `eval`
 - `error`
 - `warning`

- echo

- What do the chunk options `fig-height` and `fig-width` do – what do they do when they're set in a single code chunk and what do they do when they're set in the document YAML on top?

Exercise 2

All about `group_by()`:

Suppose we have the following tiny data frame:

```
df <- tibble(
  x = 1:5,
  y = c("a", "b", "a", "a", "b"),
  z = c("K", "K", "L", "L", "K")
)
df
```

```
# A tibble: 5 x 3
      x y      z
<int> <chr> <chr>
1     1 a      K
2     2 b      K
3     3 a      L
4     4 a      L
5     5 b      K
```

- What does the following code chunk do? Run it and analyze the result and articulate in words what `group_by()` does.

```
df |>
  group_by(y)
```

- What does the following code chunk do? Run it and analyze the result and articulate in words what `arrange()` does. Also comment on how it's different from the `group_by()` in part (a)?

```
df |>
  arrange(y)
```

c. What does the following code chunk do? Run it and analyze the result and articulate in words what the pipeline does.

```
df |>
  group_by(y) |>
  summarize(mean_x = mean(x))
```

d. What does the following code chunk do? Run it and analyze the result and articulate in words what the pipeline does. Then, comment on what the message says.

```
df |>
  group_by(y, z) |>
  summarize(mean_x = mean(x))
```

e. What does the following code chunk do? Run it and analyze the result and articulate in words what the pipeline does. How is the output different from the one in part (d).

```
df |>
  group_by(y, z) |>
  summarize(mean_x = mean(x), .groups = "drop")
```

f. What do the following pipelines do? Run both and analyze their results and articulate in words what each pipeline does. How are the outputs of the two pipelines different?

```
df |>
  group_by(y, z) |>
  summarize(mean_x = mean(x))

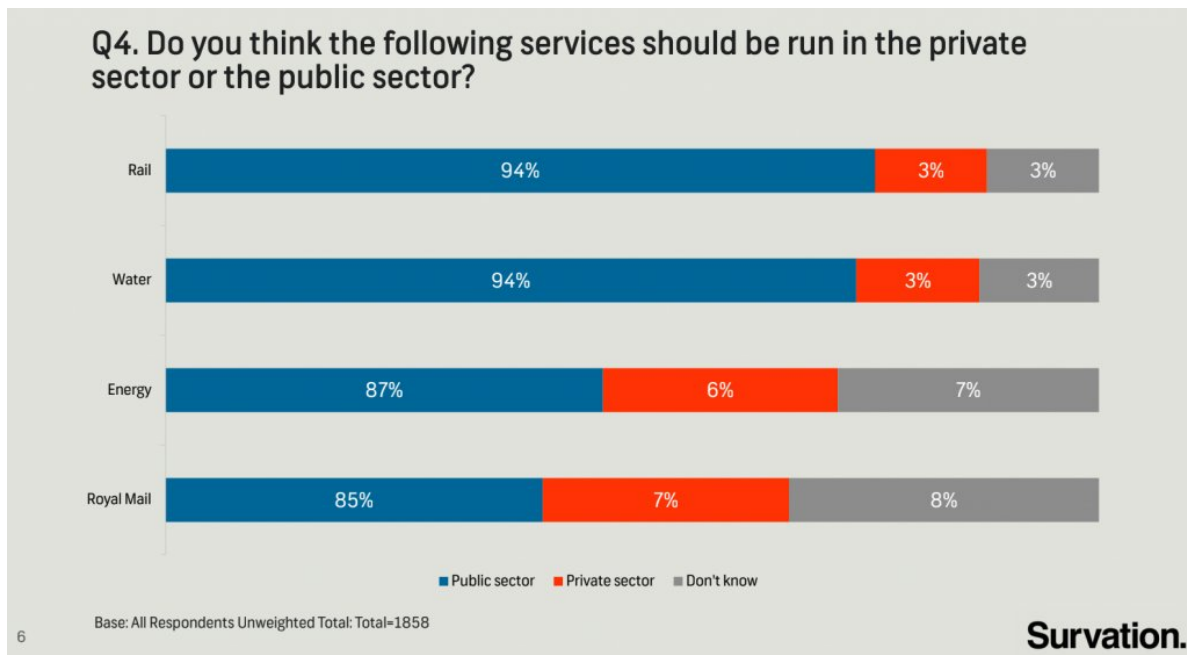
df |>
  group_by(y, z) |>
  mutate(mean_x = mean(x))
```

Render, commit (with a descriptive and concise commit message), and push. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceeding.

Exercise 3

The following chart was [shared](#) by @GraphCrimes on Twitter on September 3, 2022.

a. What is misleading about this graph?



- Suppose you wanted to recreate this plot, with improvements to avoid its misleading pitfalls from part (a). You would obviously need the data from the survey in order to be able to do that. How many observations would this data have? How many variables (at least) should it have, and what should those variables be?
- Load the data for this survey from `data/survation.csv`. Confirm that the data match the percentages from the visualization. That is, calculate the percentages of public sector, private sector, don't know for each of the services and check that they match the percentages from the plot.
- Recreate the visualization, and improve it. You only need to submit the improved version, not a recreation of the misleading graph exactly. Does the improved visualization look different than the original? Does it send a different message at a first glance?

Exercise 4

A data scientist compiled data from several public sources (voter registration, political contributions, tax records) that were used to predict sexual orientation of individuals in a community. What ethical considerations arise that should guide use of such data sets?¹

Once again, render, commit, and push. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceeding.

¹This exercise is from [MDSR, Chp 8](#).

Exercise 5

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 (internet protocol) address, demographic profiles, and preferences for relationships. Why might it be problematic to post a deidentified form of this data set where name and email address were removed?²

Exercise 6

To complete this exercise you will first need to watch the documentary Coded Bias. To do so, you either need to be on the Duke network or connected to the Duke VPN. Then go to <https://find.library.duke.edu/catalog/DUKE009834953> and click on “View Online”. Once you watch the video, write a one paragraph reflection highlighting at least one thing that you already knew about (from the course prep materials) and at least one thing you learned from the movie as well as any other aspects of the documentary that you found interesting / enlightening.

Render, commit, and push one last time. Make sure that you commit and push all changed documents and your Git pane is completely empty before proceeding.

Wrap up

Submission

- Go to <http://www.gradescope.com> and click Log in in the top right corner.
- Click School Credentials Duke Net ID and log in using your Net ID credentials.
- Click on your STA 199 course.
- Click on the assignment, and you’ll be prompted to submit it.
- Mark all the pages associated with exercise. All the pages of your homework should be associated with at least one question (i.e., should be “checked”). If you do not do this, you will be subject to lose points on the assignment.
- Select the first page of your PDF submission to be associated with the “Workflow & formatting” question.

²This exercise is from [MDSR, Chp 8](#).

Grading

- Exercise 1: 10 points
- Exercise 2: 10 points
- Exercise 3: 10 points
- Exercise 4: 3 points
- Exercise 5: 3 points
- Exercise 6: 10 points
- Workflow + formatting: 4 points
- Total: 50 points