Problem Set 1

SOCI 385 - Fall 2023

Matt Lawrence

Opening header

In the YAML section above (visible in the qmd document but not the rendered PDF), note that I am including my name in the author field and adding a title and subtitle. I am stating PDF as the format type and - this is new - adding "fig-pos: H" as a PDF option. This forces figures to be "hard positioned" where they are located in the document rather than being pushed to a separate page.

Set up

Start by loading the tidyverse package and loading the data. Note that you only have to load tidyverse once per quarto document, not for every function it uses. Also, tidyverse includes ggplot2, dplyr, and other "data wrangling" packages that some students are loading separately. Just do tidyverse and that will cover all of those. The packages we sometimes use that are not included in tidyverse and that will need to be separately loaded are kableExtra, DT, and ggplotly.

To render properly, there cannot be a hashtag in front of a library line and there must be a hashtag in front of an install.packages line (though you shouldn't need to be installing packages during assignments).

```
library(tidyverse)

ps1 <- read_csv("https://raw.githubusercontent.com/mjclawrence/soci385_f23/main/data/ps1.c
ps1_long <- read_csv("https://raw.githubusercontent.com/mjclawrence/soci385_f23/main/data/
```

1. In a few sentences, describe any gender differences in who uses Facebook vs who uses Twitter. Consider the proportion self-identifying with each gender who report using the sites as well as the proportion of each site's users self-identifying with each gender.

This question is asking for multiple proportion tables. First let's look at the proportion of respondents in each gender category who use Facebook and Twitter. The gender, use_facebook, and use_twitter variables are all you need for this question; you do not have to create any new variables.

In the chunk below, the ",1" option connected to the prop.table parentheses asks for row proportions (since variable number 1 in your table code is the row variable). The ",3" option connected to the round parentheses asks to round the decimal points in the table to three places.

```
round(prop.table(table(ps1$gender, ps1$use_facebook),1),3)
```

```
          0     1
Man   0.346 0.654
Other 0.417 0.583
Woman 0.194 0.806
```

Remember to use the kableExtra package to produce better looking tables. This package has to be loaded separately from tidyverse.

```
library(kableExtra)
```

Set up the table exactly as above but chain it to the kable() function. The "booktabs = TRUE" option sets up the basic table styling that works well for a PDF output. The "align = rep("c", 2)" option centers (that's what the c is for) the columns. The number should equal the number of categories in the column variable in your table. The "caption =" option adds a table title. Kable automatically numbers tables with captions.

There are two other things in the chunk below that we have not seen in class yet. I add the kable_paper() function to make the output easier to read in R Studio. And I add the kable_styling() line to force the table to be placed where it is in the notebook, otherwise kable likes to move tables to the top of pages.

```
round(prop.table(table(ps1$gender, ps1$use_facebook),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Facebook use by gender") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Let's make the Twitter table too:

Table 1: Facebook use by gender

	0	1
Man	0.346	0.654
Other	0.417	0.583
Woman	0.194	0.806

```
round(prop.table(table(ps1$gender, ps1$use_twitter),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Twitter use by gender") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 2: Twitter use by gender

	0	1
Man	0.671	0.329
Other	0.667	0.333
Woman	0.736	0.264

The key takeaways from these tables are that a higher proportion of women than men or self-identified "other" respondents use Facebook while a lower proportion of women than men or self-identified "other" respondents use Twitter. Overall, Facebook use is more common in this sample. More than half of respondents in any gender category report using Facebook (with 80% of women and two-thirds of men using that site). No more than one-third of respondents in any gender category report using Twitter.

The second part of this question asks for the proportions of users of each site who are in each gender category. We will use the same table setup for this part of the question but will get column proportions rather than row proportions (by changing the prop.table option to ",2").

```
round(prop.table(table(ps1$gender, ps1$use_facebook),2),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Gender composition of Facebook users") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 3: Gender composition of Facebook users

	0	1
Man	0.574	0.386
Other	0.014	0.007
Woman	0.412	0.607

```
round(prop.table(table(ps1$gender, ps1$use_twitter),2),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Gender composition of Twitter users") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 4: Gender composition of Twitter users

	0	1
Man	0.414	0.49
Other	0.009	0.01
Woman	0.578	0.50

Not surprisingly, most Facebook users (sixty percent or three-fifths of the sample) are women. Men and women represent equal shares of Twitter users, but that is probably more a result of the sample having more women than men. If we look at the proportions of respondents who do not use Twitter, the fact that a majority of them are women is more consistent with the earlier table showing that higher proportions of men than women use Twitter.

2. In a paragraph, describe the responses to the get digital news question. What strikes you as sociologically interesting about the responses? What would you want to know more about? Your paragraph should include:

- The overall proportion of respondents reporting each answer
- The proportion of respondents in each age category reporting each answer
- The proportion of respondents in each education category reporting each answer

Let's start by looking at the distribution of responses to the news_social_media question:

```
round(prop.table(table(ps1$news_social_media)),3)
```

```
  Never     Often    Rarely Sometimes
  0.263     0.160     0.240     0.337
```

These categories should be ordered from most negative to most positive. Let's do that once and save the results in the ps1 data frame so we don't have to do it for each table:

```
ps1 <- ps1 |>
  mutate(news_social_media = factor(news_social_media,
                                    levels = c("Never",
                                               "Rarely",
                                               "Sometimes",
                                               "Often")))
```

Now the table will show the categories in that order:

```
round(prop.table(table(ps1$news_social_media)),3)
```

```
  Never    Rarely Sometimes     Often
  0.263     0.240     0.337     0.160
```

The age categories are already in order so we can simply use them in a cross table. Age should be in the rows and news_social_media should be in the columns because news_social_media is our outcome. We think age will predict differences in answers to the news question. Since the question asks for "the proportion of respondents in each age category reporting each answer" the age categories (or rows) should add up to 1 so we want the ",1" option with the prop.table function. Note that the align option for kable should be set to 4 now since there are 4 categories in our outcome variable.

```
round(prop.table(table(ps1$age_cat, ps1$news_social_media),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 4),
        caption = "Age differences in getting news from social media") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

This table suggests a strong relationship between age and the use of social media to get news. Higher proportions of younger respondents report getting their news from social media: more than seventy-five percent say they do so sometimes or often. That contrasts with the one-third of respondents in the oldest age group who report doing so. Additionally, more than forty percent of respondents in the oldest age group say they never do so.

Table 5: Age differences in getting news from social media

	Never	Rarely	Sometimes	Often
18-29	0.059	0.175	0.409	0.358
30-49	0.156	0.248	0.397	0.199
50-64	0.296	0.257	0.326	0.121
65+	0.428	0.234	0.251	0.087

The education categories do need to be ordered before we use them. Watch that your levels option refers to the categories with exactly the same spelling, capitalization, and characters as recorded in the data. If there are any differences, that category will not show up correctly when you use it later.

```
ps1 <- ps1 |>
  mutate(educ_cat = factor(educ_cat,
                           levels = c("HS Grad or Less",
                                      "Some College",
                                      "College Grad+")))
```

```
round(prop.table(table(ps1$educ_cat, ps1$news_social_media),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 4),
        caption = "Education differences in getting news from social media") |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 6: Education differences in getting news from social media

	Never	Rarely	Sometimes	Often
HS Grad or Less	0.200	0.204	0.390	0.206
Some College	0.260	0.241	0.346	0.153
College Grad+	0.286	0.251	0.313	0.150

The education differences are less obvious than the age differences. Since these education categories are broad, the most common additional variables students suggested including - class, income, region, etc. - could be helpful for identifying differences within education categories.

3. In a few sentences, describe racial/ethnic differences in responses to the news_trust question. What is an additional variable (not in the dataset for this assignment) you would want to have in order to explain the relationship between these variables? Why might this additional variable be important? How would you phrase a survey question

and possible responses to gather data about this additional variable?

This is another proportion table with ordered categories for the group_trust variable. Re-ordering the race variable is not necessary (but you can do so using the same code if you think there is a reason to do so).

```
ps1 <- ps1 |>
  mutate(news_trust = factor(news_trust,
         levels = c("Not at all",
                    "Not too much",
                    "Some",
                    "A lot")))
```

```
round(prop.table(table(ps1$racethn, ps1$news_trust),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 4),
        caption = "Racial and ethnic differences in trusting the information from social m
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 7: Racial and ethnic differences in trusting the information from social media

	Not at all	Not too much	Some	A lot
Asian	0.147	0.450	0.319	0.084
Black	0.154	0.413	0.317	0.116
Hispanic	0.181	0.410	0.344	0.065
Other	0.268	0.383	0.268	0.082
White	0.352	0.418	0.198	0.032

Most students were surprised by these results, especially that more than seventy-five percent of white respondents reported negative views of trust in social media. Breaking out these responses by variables like how closely one follows the news, other news sites respondents use, voting behavior, or trust in other institutions could all be good ways to get more information about the relationship between race and trust in social media.

4. Collapse the political ideology responses into three categories: any liberal, moderate, any conservative. In a few sentences, compare differences across these three new categories in responses to the misinformation and government question.

Let's start by getting the three categories. There are multiple ways to do this, all of which would use ifelse in some way. The most efficient approach would be to use str_detect to take advantage of the two liberal categories sharing that term (or "string") and the two conservative

categories sharing that term. Also remember to give your new variable a new name or the original ideology values will be replaced!

I'll also change the order of the categories so Moderate is in the middle.

```
ps1 <- ps1 |>
  mutate(ideology_cat = ifelse(str_detect(ideology, "Liberal"),
                               "Liberal",
                               ifelse(str_detect(ideology,
                                                 "Conservative"),
                                      "Conservative",
                                      "Moderate"))) |>
  mutate(ideology_cat = factor(ideology_cat,
                               levels = c("Conservative",
                                          "Moderate",
                                          "Liberal")))
```

You can check your work by setting up a cross tab with the original ideology values and the new ideology_cat values:

```
table(ps1$ideology, ps1$ideology_cat)
```

	Conservative	Moderate	Liberal
Conservative	1362	0	0
Liberal	0	0	1033
Moderate	0	2033	0
Very Conservative	449	0	0
Very Liberal	0	0	479

Looks like it worked: the two original conservative categories are in the new conservative category, the two original liberal categories are in the new liberal category, and moderates are still moderates.

Final step is a proportion table using the new ideology_cat variable and the misinformation variable:

```
round(prop.table(table(ps1$ideology_cat, ps1$misinformation),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Political ideology differences in how to address misinformation") |>
  kable_paper() |>
```

```
  kable_styling(latex_options = "hold_position")
```

Table 8: Political ideology differences in how to address misinformation

	1	2
Conservative	0.251	0.749
Moderate	0.541	0.459
Liberal	0.627	0.373

We haven't changed column names in any of our previous tables. This is a good example of a table that would benefit from real column names. You can change them with the "col.names" option inside the kable function. Note that these names change the values of the column categories, so the first column (above the row categories) does not change. In this case, using the full statements as your column names would be too long; simplify to column names that give enough information for the reader to connect the table to the description of results in your text.

```
round(prop.table(table(ps1$ideology_cat, ps1$misinformation),1),3) |>
  kable(booktabs = TRUE,
        align = rep("c", 2),
        caption = "Political ideology differences in how to address misinformation",
        col.names = c("Favors Government Restriction", "Favors Freedom To Publish")) |>
  kable_paper() |>
  kable_styling(latex_options = "hold_position")
```

Table 9: Political ideology differences in how to address misinformation

	Favors Government Restriction	Favors Freedom To Publish
Conservative	0.251	0.749
Moderate	0.541	0.459
Liberal	0.627	0.373

5. Use the ps1_long.csv file to create two figures. In the first figure, show the proportion of respondents using each social media site. In the second figure, show the proportion of users of each site who regularly get news from it. In a few sentences, describe the similarities and differences across these two figures. What do these figures suggest to you about social media as a source of news and information?

The key to answer this question is remembering that - drumroll, please - the mean of a binary variable is the proportion with a 1. We have binary variables for use (0 = does not use, 1 = does use) and news (0 = does not get news, 1 = does get news) for each respondent and each

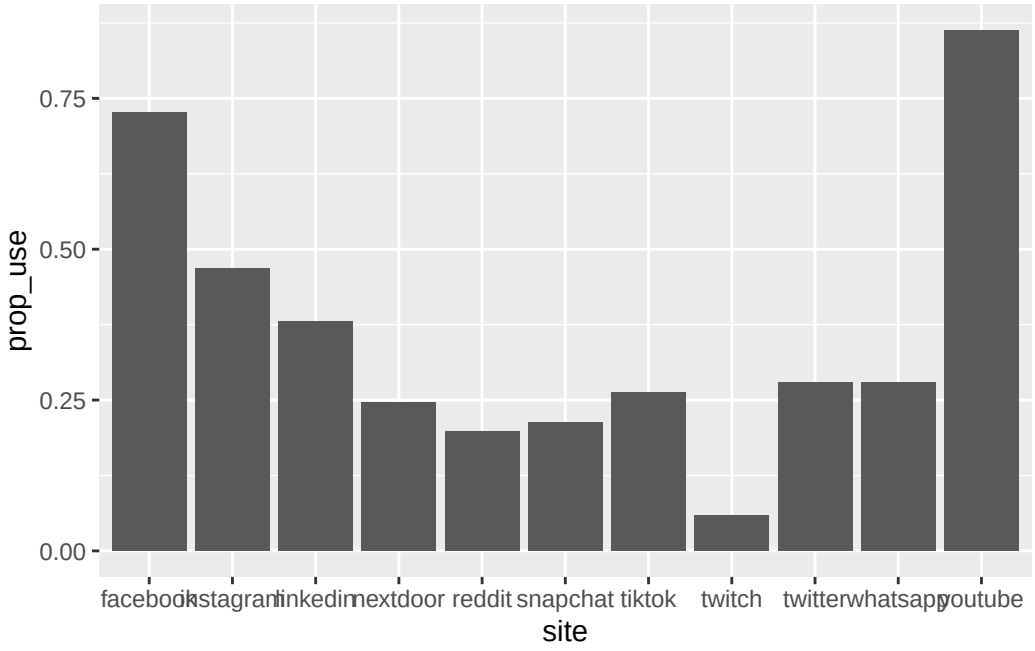
site. Since we want the proportion that does use each site, we can combine group_by (that gives us something for each site) with summarise (to get the means).

```
ps1_long |>
  group_by(site) |>
  summarise(prop_use = mean(use))
```

```
# A tibble: 11 x 2
   site      prop_use
   <chr>        <dbl>
 1 facebook    0.727
 2 instagram   0.468
 3 linkedin    0.380
 4 nextdoor    0.246
 5 reddit      0.198
 6 snapchat    0.212
 7 tiktok      0.263
 8 twitch      0.0591
 9 twitter     0.280
10 whatsapp    0.279
11 youtube     0.862
```

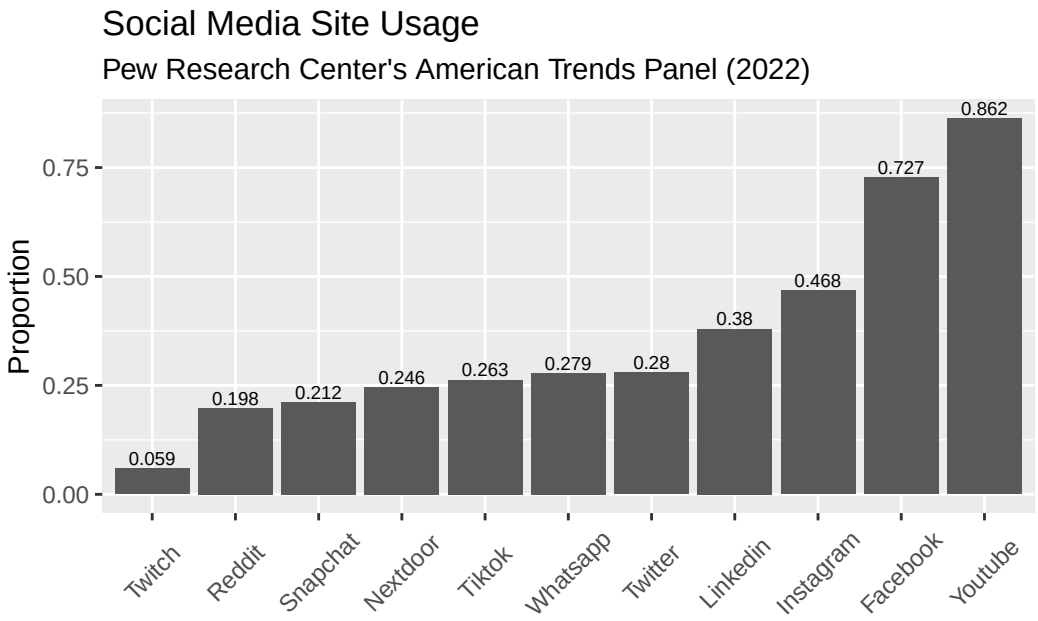
That gives us the x (site) and y (prop_use) for a nice figure. Piping the ggplot function onto the code we already have simplifies the process. Bar plots were popular choices for this figure. Recall that since we ahve an x and a y, we would use geom_col rather than geom_bar if we want a bar plot.

```
ps1_long |>
  group_by(site) |>
  summarise(prop_use = mean(use)) |>
  ggplot(aes(x = site, y = prop_use)) + geom_col()
```



That's the basic setup. Now let's use reorder() to order the categories by value of prop_use, change the site names to title case, and add labels to the bars, axes, and plot. You can also shrink and rotate the axis labels using the theme options.

```
ps1_long |>
  group_by(site) |>
  summarise(prop_use = mean(use)) |>
  ggplot(aes(x = reorder(str_to_title(site), prop_use),
             y = prop_use)) +
  geom_col() +
  geom_text(aes(label = round(prop_use,3), vjust = -.25), size = 2.5) +
  labs(x = "", # the values are self explanatory, so leaving title blank is okay
       y = "Proportion",
       title = "Social Media Site Usage",
       subtitle = "Pew Research Center's American Trends Panel (2022)") +
  theme(axis.text.x = element_text(size = 9, angle = 45, vjust = .5))
```



The difference for the second plot is that we only want to include users of the sites. To get them, add a filter before summarizing to get the mean of the news variable. From the previous figure, we know that we can filter for use == 1 to restrict our sample to users of each site.

```
ps1_long |>
  filter(use == 1) |>
  group_by(site) |>
  summarise(prop_news = mean(news))
```

```
# A tibble: 11 x 2
   site      prop_news
   <chr>         <dbl>
 1 facebook     0.410
 2 instagram    0.264
 3 linkedin     0.138
 4 nextdoor     0.220
 5 reddit       0.322
 6 snapchat     0.126
 7 tiktok       0.278
 8 twitch       0.0948
 9 twitter      0.496
10 whatsapp     0.115
```

```
11 youtube      0.259
```

The code for the figure will be the same setup and another column plot works well. To add some excitement and show you an alternative, I'll use geom_point in the example below and flip the x and y axes (using coord_flip) to make the figure more visually appealing.

```
ps1_long |>
  filter(use == 1) |>
  group_by(site) |>
  summarise(prop_news = mean(news)) |>
  ggplot(aes(x = reorder(str_to_title(site), prop_news),
             y = prop_news)) +
  geom_point() +
  coord_flip() +
  geom_text(aes(label = round(prop_news,2), hjust = -.25), size = 3) +
  labs(x = "", # the values are self explanatory, so leaving title blank is okay
       y = "Proportion",
       title = "Use Of Social Media Site For News",
       subtitle = "Pew Research Center's American Trends Panel (2022)")
```