Week 3 Practice

Willa Voorhies & Elena Leib

Try doing the following questions first on your own, and then check your answers with your neighbors.

Remember you can get help on any function by typing ? followed by the function into the console. e.g., `?filter`. You can also find help by searching on Google e.g., "tidyverse filter"

Open an R script and load the covid_attitudes data (and don't forget to set options(stringsAsFactors = FALSE) before loading the data)! Then perform the following data processing steps. After you've performed each step, use the functions you learned last week to check if it worked.

1) Remove the variable `Q6.consent` from your dataframe (we only have data for consented participants) and `Q13.trust_none_of_above` (we saw that this column had a lot of NAs and bad/not useful data)

2) Make a new data frame that only has observations from large cities.

3) Make a different data frame that has people who either live in a large city or a small city

4) Make a third data frame that only has people below the age of 50 that have earned a 4 year degree

5) Take the data frame from step #3. Remove participants from the dataframe that have NAs for any cells. Then, use what you learned last week to see if all the NAs are removed.

6a) Take the data frame from step #3 and only remove participants that have NA in the age column

6b) Compare the size of the data frames resulting from questions 5 and 6a. What do you think about using drop_na() across all columns? Write a sentence or two explaining to a friend what drop_na() is good for and when to be cautious with it.

7) Take your pipe from step #5. Add a new variable to your dataframe called "apprehension_score" that is a composite score of Q18, Q20, and Q21.
8) Now, make one of the likert-scale columns into a factor. Be sure to specify the levels of the factor so that they are in the right order and make sense

9) Save your data frame as a csv file using `write_csv()` using the name `covid_attitudes_w3_clean`. ***Hint:*** Your code should look something like this:

```
write_csv(your_df, file ="add-your-file-name-here.csv")
```

Note: You can also use write.csv() [Same difference as read.csv() vs read_csv()]

```
write_csv(your_df, file ="add-your-file-name-here.csv",
row.names=FALSE)
```

Try it out! What happens if you don't include row.names=FALSE? Check the help file (?write.csv) to help you make your hypothesis and then check the output file to confirm.

***If you have extra time*… explore some additional tidyverse functionality with the bonus tasks below. We'll talk more about these functions next week, too!**

Bonus:

We can use `summarise()` to get summary statistics for our variables.
Run the following code to see how it works:

```
apprehension_summary <- covid_attitudes_clean %>%
                summarise(apprehension_mean =
                mean(apprehension_score))
```

Try this yourself but now get the standard deviation instead of the mean (***hint:*** you can use `?summarise` if you're stuck, and/or look up how to do standard deviation in R).

Now let's look at the average apprehension score for each age group using another powerful tidyverse command: `group_by()`. Run the following code:

```
apprehension_summary <- covid_attitudes_clean %>%
                # group by age
                group_by(Q84.community) %>%
                # get mean apprehension score for each community group
                summarise(apprehension_mean=mean(apprehension_score))%>%
                # it's important to ungroup!
                ungroup()
```

i) View the new `apprehension_summary` variable. What did it give you?
ii) Try grouping by another variable of interest (or more than one). Don't forget to ungroup() after you're done !

Grouping doesn't alter your data frame, it just changes how it's listed and how it interacts with the other commands.

Check out the tidy cheat sheet for more tidyverse and data wrangling functionality!
https://www.rstudio.com/wp-content/uploads/2015/02/data-wrangling-cheatsheet.pdf