Coding Challenge 2: Tidyverse in Action

Public Health 460

Due: 6pm, Friday February 17th, 2023

The functions in tidyverse are powerful tools for exploring, modifying, and arranging data. Please turn in an HTML file and your .Rmd file.

Setup Install the packages `heplots` and `palmerpenguins`, using, e.g. `install.packages("heplots")`. (Note: you only need to run this line of code once in your console, you should not include it in your Rmd file! Once you install the package you will not need to reinstall it –unless you want to update it– until you upgrade R. But you will need to run the `library()` command every time you want to access the functions in the package.)

Then, load the `Diabetes` dataset using the following code

```
library(heplots)
data(Diabetes)
```

1) Information on this data set and the meaning of its variables can be found by entering `?Diabetes` in the console after the data set is loaded. What does one row in this dataset represent? What does the variable `group` measure and what are its possible values? (1 pt)
2) Next, load the tidyverse package.
3) Use the `head()` function to see the first 6 lines of the Diabetes data set. Then use the `glimpse()` function to see a preview, or get a glimpse of, the dataset. Both of these functions are good ways to get a first look at a data set. Write down one difference between the output of these two functions. Which one do you find more helpful for this dataset and why? (1 pt)
4) What type of R object is `Diabetes`? (Hint:Use the `class()` function.) Create a new version of the Diabetes dataset called `diabetes_tbl` that is a `tibble` R object. (Hint: use the `tibble()` function.) Run a line of code that just has the word `diabetes_tbl` on it. How is this output similar or different to the output from `head()` above? Write your answers to these questions in the text of the RMarkdown document, and also show your code that supports your answers. (2 pts)
5) Use the `select()` function to choose the columns `glufast`, `glutest`, and `group` from `diabetes_tbl`. Save the new data set as `diabetes_select`. (2 pts)
6) On `diabetes_select`, use `filter()` to extract only the rows with `glufast` greater than 105 and also a `glutest` greater than 450. What do the people in this new data set have in common? (3 pts)
7) On `diabetes_select`, use the "pipe operator" (`%>%`) to string together the following data-wrangling verbs/operations: a) `mutate` a new column "gluratio" that is equal to `glufast/glutest`, b) `arrange` the data with the lowest `gluratio` values at the top, c) `print` out the top 13 rows. Note that no data needs to be saved here, you should run one command that prints output at the end. It is standard tidyverse style to start a newline after every "pipe". (3 pts)
8) Load the `penguins` data set from the `palmerpenguins` package. What year were these data originally published? On what continent were they collected? (2 pts)
9) Create a data set that `filter`s to only include the data of female penguins of the Gentoo species. Name this data set `gentoo_female`. How many female Gentoo penguins are in this data set? (3 pts)
10) In one set of piped operations that ends up printing out a single number, a) `mutate` a column `body_mass_kg` that gives the body mass in kg of each penguin in `gentoo_female`, b) `pull` out that column, c) compute the `mean` of the observations on female Gentoo penguins. Show your code and write one sentence that says what the mean body mass, in kg, of female Gentoo penguins is. (3 pts)