Homework 2

The homework is in this format to facilitate you showing your code and output. Render the document when you are finished and submit it on Canvas. If you have trouble rendering to Word, try format: html instead. If you are still having trouble, you can always copy and paste your code and output (or screenshots of the output) into a Word document. What is important is that we can see both your code and its output.

1. Start a new R session. Before you can read in the data, load the tidyverse set of packages. Then read in the data.
2. Use select() to select the first 5 variables in the dataset. Show how to do this in as many ways as you can (at least 2).
3. Create a vector of 3 variable names, and 1 that doesn’t exist: “id”, “region”, “income”, and “education”. Use select() twice: once with the any\_of() and next with the all\_of() helper. What is the difference between the two?
4. Did you know you can also use select() to rename variables? Use select() to rename the id variable to ID, and select everything else with its same name.
5. Use filter() to select only the rows where sex is equal to 1. Then use filter() to select only the rows where sex\_cat is equal to “Male”. Show that the two filters return the same number of rows.
6. Create a dataset that includes only individuals who have an income between 25,000 and 75,000. Next, look up the function between() and use it to create the same dataset.
7. Create a dataset that includes only individuals who either have more than 2 siblings and less than 6 hours of sleep on weekdays, or who have an income between 25,000 and 75,000.
8. Create a new variable total\_sleep that sums sleep\_wkdy and sleep\_wknd. Then, create a categorical variable sleep\_category based on total\_sleep: “Low” for less than 10 hours, “Medium” for 10-14 hours, and “High” for more than 14 hours. Select only the people with “High” sleep. Put this all in chain using pipes.
9. Using pipes, create a dataset that filters out individuals with an income below 30,000, and then creates a mean-centered version of the income variable. Then create another dataset reversing the order of the piped steps. What is the difference between the datasets?