Homework 1

Please copy and paste your code and answers into this document and submit on Canvas.

1. Using the NLSY data, make a scatter plot of the relationship between hours of sleep on weekends and weekdays. Color it according to region (where 1 = northeast, 2 = north central, 3 = south, and 4 = west).
2. Replace geom\_point() with geom\_jitter(). What does this do? Why might this be a good choice for this graph? Play with the width = and height = options. This site may help: https://ggplot2.tidyverse.org/reference/geom\_jitter.html
3. Use the shape = argument to map the sex variable to different shapes. Change the shapes to squares and diamonds. (Hint: how did we manually change colors to certain values? This might help: https://ggplot2.tidyverse.org/articles/ggplot2-specs.html)
4. When we’re comparing distributions with very different numbers of observations, instead of scaling the y-axis like we did with the facet\_grid() function, we might want to make density histograms. Use google to figure out how to make a density histogram of income. Facet it by region.
5. Make each of the regions in your histogram from part 1 a different color. (Hint: compare what col = and fill = do to histograms).
6. Instead of a log-transformed x-axis, make a square-root transformed x-axis.
7. Doing part 3 squishes the labels on the x-axis. Using the breaks = argument that all the scale\_x\_() functions have, make labels at 1000, 10000, 25000, and 50000.
8. The order of the different levels of a factor variable is important. Describe a scenario where different orders of the levels of a factor variable may lead to different results. (Hint: Think about how factor variables are treated a regression analysis.) Comment SF: This may be too advanced. But if the students already know how to run regression analysis in STATA, they might know this. In STATA, the same as in R the lowest/first level is used as reference level for categorical variables in a regression model as default.
9. The “race\_eth”-variable in the NLSY dataset relates to the following level: hispanic (1), black (2), other (3). When loading the data it is a numeric variable. Turn it into a factor variable making sure it is labeled correctly.
10. Create a new variable called “race\_reor”, reordering the levels such that “other” would automatically be used as reference level in analyses. Use the count()-function to check the order of these two variables.
11. Age of individuals contained in the NLSY dataset ranges from 13 to 52 years (variable: age\_bir). Use case\_when() to create the following age categories: teenager (13-19), young adult (20-29), adult (30-52). Make sure it is a correctly labeled factor variable (hint: this can be a two step process).
12. Using the age group factor variable, show the mean age and total number of individuals in each age group.