Week 8 assignment Part 2 (R/RMarkdown)

Due Nov 8, 2022 by 10am **Points** 5 **Submitting** a file upload
Available Oct 30, 2022 at 12am - Nov 8, 2022 at 10am

This assignment was locked Nov 8, 2022 at 10am.

This assignment builds on your Week 5 and 6 assignments and uses the same LFS data.

As usual, follow the instructions below in an RMarkdown document. Don't forget to include your name, student number, and an informative title. Please make sure that your knitted document includes the R code chunks (so don't use the "echo=FALSE" option). We want to see your code and the results! You should suppress unnecessary messages and warnings using the appropriate code chunk options. When you've answered the questions below, save your .Rmd file and knit the RMarkdown document to produce a nicely formatted html document. To complete the assignment, submit BOTH your RMarkdown (.Rmd) file AND the knitted .html file via canvas.

Your instructions are:

- Starting from the raw LFS .csv file, create a data object that includes only individuals who are "employed, at work" and who work Full-time in their main job.
- Create a new character variable `EduCat` from `EDUC`, that has proper text labels ("0 to 8 years", "Some high school", etc.) for each of the categories.
- Create a new numeric variable `Wage` equal to `HRLYEARN` divided by one hundred. Recall that `HRLYEARN` has two implied decimal places, i.e., it is measured in cents. Hence `Wage` is measured in dollars.
- Use the `lm()` function to regress `Wage` on job tenure with the current employer and `EduCat`.
- Assign the output of `lm()` to an object. Use the `summary()` function to display a summary of your regression results.
- You'll notice that R reports estimated coefficients for all of the education categories except one. That's called the "reference category." Coefficients for the other education categories measure *the predicted difference in earnings between that category and the reference category*, holding job tenure constant. Report which category is the reference category.
- Do your best to interpret the estimated regression coefficients (e.g. "I find that a one month increase in job tenure is associated with a $XXXX increase in wages, holding education constant" and "... on average, individuals who have completed *Some high school* earn $XXX more than the reference category, holding job tenure constant". Are any of the results surprising?

- Use broom's `tidy()` to plot the estimated coefficients (and their 95% confidence intervals) for each of the education categories. Do any of the categories' confidence intervals include zero? If so, which ones? What does this tell you?
- Create a new character variable `ImmigCat` from `IMMIG`, that has proper text labels ("Immigrant, landed 10 or less years earlier", etc.) for each of the categories.
- Use the `lm()` function to regress `Wage` on job tenure with the current employer, `EduCat`, and `ImmigCat`. Assign the output of `lm()` to an object. Use the `summary()` function to display a summary of your regression results. Interpret the estimated regression coefficient on the job tenure variable. Has it changed from what you found in the first regression? If so, can you explain why?
- Do your best to visualize your regression model's predictions on a scatter plot of earnings against the other variables in your regression model. Be sure to include a prediction interval for your regression model, and don't forget that you can use colors, groups, and facets in your plot!

As always, make sure your report and everything in it look professional. Make sure labels on plots are clear and easy to interpret.