**GV950 Spring 2023 – HW1 Reassessment**

**Guidelines:**

* This reassessment assignment is worth a total of **45 marks**. The marks for each question are listed at the end of each question.
* You **must** submit **two** files for this reassessment to count as complete: an RMarkdown file (.Rmd) and a MS Word file (.docx), which is the output of the Rmd file. When I knit your Rmd file, I should get the *exact* same output as the Word file. If the two files do not match, **20-25** marks will be deducted (depending on how large the differences are).
* In your Rmd file, the author field must have your **7-digit student ID (not your name)** and the title field must state “GV950 HW1 Reassessment.” Not doing this correctly will lead to a deduction of **3** marks from the total.
* You **must** briefly comment all your R code to briefly explain what you are doing and how.
* All text that is part of the answer must be written as text in the Rmd file, *not* as R comments.
* All questions **must** be answered using packages and functions taught in the lectures. Using other functions will receive *no* credit for those questions even if the output is correct. Similarly, all figures *must* be made using *ggplot2* functions, *not* using base R functions. No credit will be given for using base R figures.
* When a question asks you to explain something, make sure to justify and explain your answer. E.g., if a question asks what direction a variable is skewed, say more than ‘left’ or ‘right.’ How can you tell? Is it only slightly skewed or very much skewed? And so on.
* Any answer you give must have the relevant code and its output given as well. E.g., if you look at a histogram to determine how a variable is skewed, you must also give the code and output of the histogram before sharing your conclusion.
* Please remember that the homework is to be done individually, not in consultation with anyone else.
* Please note that not following the instructions given in these guidelines *will* adversely affect your overall marks.

**Questions:**

Download and load the “evs.csv” dataset that has been made available to you for this reassessment (This file is titled “GV950\_HW1\_Reassessment\_evs.csv”) The “GV950\_HW1\_Reassessment\_evs\_questions.pdf” file lists the questions from this survey, which is the European Values Survey that was conducted for respondents in Great Britain. You can see that the questions in the PDF file give you the variable names that match the .csv file so you can use this PDF as the codebook, as needed.

1. Import the *evs.csv* dataset into R. What is the *unit of analysis* in this dataset, and how can you tell? **[1 mark]**
2. Say you have a friend who is interested in understanding what British people think are the primary reasons that some people in the country live in poverty/need, but this friend does not know much about data analysis or R, so they ask you to help them. For this, you will want to look at Q12 and Q13 in the codebook PDF and think about how to use the information given there in order to help them answer their question. You will answer it *descriptively* only (i.e., using no statistical tests) and should choose to summarize the information in a way that presents one *type* of figure along with a 5-6 line summary/description of what the figure depicts regarding what people’s opinions are on this issue. You must also briefly explain *why* you chose to summarize the information in the way that you did. Remember to provide all code and output that you use to describe the relevant patterns, and to make the answer and figure(s) as informative and accessible as possible. **[5 marks]**
3. You want to determine whether a significantly different proportion of respondents think that employers should give scarce jobs to British people over immigrants compared to the proportion that thinks that men should have more of a right to scarce jobs than women. For this, you should look at Q21 in the PDF. To answer this question, state (in words) the hypothesis you are testing and conduct an appropriate test. Make sure to display the output of your test, identify the test statistic and test conclusion; don’t forget to explain *how* you reached that conclusion. (Hint: you will have to think about how to use the relevant variables for the statistical test and whether they will need to be cleaned in any way first.) **[5 marks]**
4. Make a scatterplot that plots how often a respondent attended religious services at age 12 (see Q26 in PDF) against how important they consider God to be in their life now (see Q36 in PDF). To do so, think carefully about whether you should use all values of the answers or not. If you choose not to use all answers, make sure you explain *why* you didn’t use them all. Make sure that the figure is informative, as always. Based *just* on the figure, describe the correlation between the two variables, i.e., think about the direction, strength, and possible reasons for both of these features. **[5 marks]**
5. I want to try to understand people’s self-placement on the political ideology scale. For this you should refer to v193 (Q57). To help me do so, first choose a non-dummy/non-binary main independent variable that you think would help to explain this dependent variable. Ideally, your chosen X variable will be continuous but, as there are not many of those, you can justify using an ordinal variable that has enough variation and values to count as effectively continuous for this purpose. In 1-2 sentences, justify the choice of variable in terms of its continuity and make sure to mention which Q. this is on the PDF. Next, state a clear hypothesis *and* a credible mechanism (i.e., think about causal hurdle 1) to link X to Y. **[4 marks]**
6. Explain how your chosen X and the given Y fulfil causal hurdle 2. Make sure to clearly justify your answer. If they do not fulfil this hurdle, you should go back and reconsider the choice of X from the previous question because a good variable *will* fulfil this hurdle. **[3 marks]**
7. Show how your chosen X and the given Y fulfil causal hurdle 3. To do so, make an appropriate figure, describe what it shows, and also calculate the correlation. Once again, if the variable does not fulfil this hurdle, you should go back to Question 5 and rethink your choice of X. **[3 marks]**
8. For causal hurdle 4, think of at least *three* potential confounders that we should be concerned about. For each, state what the confounding variable is, *why* we should be concerned about it (meaning how might it correlate with *both* X and Y!), and what variable you will use specifically from the dataset (and the corresponding Q. number from the PDF) to control for it. Each confounding variable’s discussion will be 3-4 sentences long ideally. **[6 marks]**
9. Now, run a linear regression of Y on *just* your main independent variable, X. Display the model results and interpret the coefficient on X. Remember to comment on both the statistical significance *and* the substantive significance. Have you found support for your main hypothesis – why or why not? **[4 marks]**
10. Run a second linear regression, this time of Y on *all four variables* from above. Interpret *each* independent variable in 1-2 sentences each, focusing on both the substantive and the statistical significance. Finally, explain whether/how the coefficient on the main independent variable has changed. **[6 marks]**
11. Which of the two models above explains variation in Y better? Run an appropriate test to compare both models and then comment on which fits the data better, and why. Make sure you show the test output, briefly explain why you chose to use whichever test you conducted and explain how you reached your conclusion. **[3 marks]**