**GV900 Fall 2023 – HW3 Reassessment**

**Guidelines:**

* This reassessment assignment is worth a total of **50 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 “GV900 HW3 Reassessment.” Not doing this correctly will lead to a deduction of **4 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 (the file is called “GV900\_HW3\_Reassessment\_evs.csv”. The “GV900\_HW3\_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. How many observations does the dataset have and how many variables? In a sentence, also explain how you figured this out. **[2 marks]**
2. Look at the summary statistics for the variable called *v64*. Based *just* on this and on looking at the questionnaire to see what this question is about, what *type* of variable is this most likely? Explain your answer in a sentence. (Hint: Remember that this question is about the variable type based on the types discussed in class.) Next, based just on this summary, explain if the variable’s distribution is likely skewed and, if so, in what direction? How can you tell if it’s skewed or not and whether it’s skewed a little or a lot? Finally, in a single sentence, what does this variable measure? **[3 marks]**
3. Create a new dataset called “important” that contains only the six variables from the dataset that ask about how important various features are in the respondent’s life (v1 to v6). **[2 marks]**
4. Using the new dataframe you just created in the previous question, change the name of all 6 variables to match the actual features that are asked about in the questionnaire. **[2 marks]**
5. In this new dataframe, create a new variable called *friends\_family* that is attached to the dataframe. To create this variable, first convert the two relevant variables (*Family* and *Friends and Acquaintances*) into numeric variables such that the *more* important family/friends are, the *higher* the person scores. Then, create *friends\_family* such that it represents the total importance of both these factors to the respondent in a single variable. In 3-4 sentences, explain how you went about creating this variable, i.e., explain your coding choices *in your own words* to someone who isn’t very familiar with R. (This should not sound like a computer wrote the explanation!) Finally, display the new variable with a histogram (making sure to use everything you’ve learned about them so far in terms of making them look visually nice and informative etc), and in 2-3 sentences explain the distribution of this variable and comment on what it means for how important people feel their friends and family tend to be. **[8 marks]**
6. Save this new dataframe as a .csv file somewhere on your computer (where you keep your GV900 materials) using the write.csv() function. (Hint: If your Markdown file does not knit properly, you can comment out the code for this after running it once, but you must provide the full line of code in order to get full marks.) **[2 marks]**
7. For this question, you will use Q11 (v66). Make sure you think carefully about the variable from the questionnaire before you begin, i.e., make sure to check whether you need to ‘clean’ the variable at all before analyzing it. If you do clean it in any way, make sure you show the code for it and explain briefly (in 1-2 sentences) what you changed and why:
8. I want to use this variable to test whether the population of people in Great Britain have a life satisfaction greater than 7.5. Write out (in words) the null hypothesis *and* the alternative hypothesis that I am going to be testing. **[2 marks]**
9. Conduct the relevant significance test using R (Hint: this is a quantitative continuous variable for our purposes so you should use the t-test. Do not interpret the output yet, that comes in a later part). Make sure you specify *all* the relevant parameters that we discussed for this function in class in your code. **[2 marks]**
10. From the output, explain what each of the following parameters represents/tells us: t, df, p-value. To do so, you need to identify the value of each parameter in this test *and* explain in your own words in a simple way *what* the relevance/purpose of each parameter is here. **[6 marks]**
11. Now, using an alpha level of 0.05, interpret the significance test. That is, identify what the test conclusion will be in terms of the hypothesis and explain (in 1-2 sentences) how you have reached this conclusion. **[2 marks]**
12. This question requires you to analyze an outcome/dependent variable based on Q16 (v90), which you should look up from the questionnaire before starting. The goal is to understand what explains a person’s satisfaction from their job. For our purposes, we will treat this as a quantitative, continuous variable.
13. First, select one other question from the questionnaire that you will use as your independent variable. The question you select should be ordinal, i.e., the values it has should be given in words rather than numbers. Identify which question you have selected and briefly (in 1-2 sentences) justify *why* it is a reasonable/good independent variable to explain this dependent variable, i.e., explain the mechanism that you think links X to Y. **[3 marks]**
14. Construct a null hypothesis and an alternative hypothesis that link this chosen independent variable with the given dependent variable. **[2 marks]**
15. Reconstruct the chosen independent variable (to make a new variable that is also attached to the dataset) such that it becomes a numeric variable that can be treated as a continuous variable for a regression. In doing so, make sure you also think carefully about respondents who didn’t reply to this question. In 2-3 sentences, explain/justify your choice for which numbers you assigned to which value of this variable and why. **[4 marks]**
16. Make a scatterplot of the dependent variable against the independent variable (using the new version you have created). As always, make sure the figure is informative, visually nice, and that you define the breaks in a meaningful way. Based *just* on this scatterplot, what does the relationship between X and Y look like? Describe it in 2-3 sentences, thinking about both the direction and the strength of the relationship. **[4 marks]**
17. Now, run a linear bivariate regression of the given dependent variable on this new independent variable you have created to test the hypothesis. Store the regression and then output its summary. **[2 marks]**
18. Using the regression output, interpret the effect of X on Y. Think about the direction, size, and significance of the effect and interpret it precisely. Overall, did you find support for your hypothesis or not, and how can you tell? **[4 marks]**