Final Project Instructions - BRFSS2015

For this assignment, name your R file BRFSS2015.R

- **Allowable libraries:** tidyverse, carat, Hmisc, lsr, olsrr, psych
- You **must use the** `read_csv` function when loading the .csv file.
- **Do not** rename the .csv file that you download from Brightspace.
- Round all float/dbl values to two decimal places.
- All statistics should be run with variables in the order I state
    - E.g., "Run a regression predicting mileage from mpg, make, and type" would be: `lm(mileage ~ mpg + make + type...)`
- Download the BRFSS2015_650.csv file from Brightspace and place it in the same folder/directory as your script file. In RStudio, set your Working Directory to your Source File location and then load the data file using `read_csv()`. Note: it may take a bit of time to load the data file - could be a minute or more depending on your computer. You most likely will see some warnings after it loads due to the fact that read_csv will try to guess the column type but because there are so many rows it won't read enough of them to accurately make a guess.
- There may be some new challenges (that we haven't covered in the lecture videos) and curveballs for this final assignment so be prepared to think outside the box and problem-solve with outside resources for some of these questions.

IMPORTANT:
When you turn this in to CodeGrade you will not see the result. This will be graded manually by the instructor after the term is finished. CodeGrade will be used only to allow the instructor to make comments directly in your code; all assessment of the assignment will be done by the instructor.

Additionally, partial credit will be given to responses that are not exactly what I'm looking for but are reasonable ways of interpreting a question, in my mind.

These data come from https://www.kaggle.com/cdc/behavioral-risk-factor-surveillance-system

To answer these questions you will need to use the codebook on Brightspace, called codebook15_llcp.pdf. **Please note that not all of the variables listed in the codebook are included in the .csv file to be downloaded from Brightspace.**

The answer to each question should be assigned to the value before the colon. For example, the answer to the first question should be assigned to Q1.

Use plenty of comments throughout your code so that the grader can follow your line of thinking.

- Q1: How many people reported their general health is excellent?

- Q2. What is the highest value for number of adult women in household where someone has ever had a stroke? Summarise the value in a variable called `max_numwomen`.
- Q3: Compute the means and standard deviations for `MENTHLTH` comparing caregivers who managed personal care such as giving medications, feeding, dressing, or bathing and those who did not. The output should be a dataframe. It should be formatted like this:

CRGVPERS	mean_health	sd_health
1	5.23	xxx.xx
2	xxx.xx	xxx.xx

- Q4: What is the median age when respondents were told they had diabetes for those living in Pennsylvania? Only calculate it for those who gave an age. The value should be called `med.diab.age` and be in a 1 x 1 dataframe.
- Q5: Predict number of days in the past 30 days mental health was not good from marital status. Assign the summary of the model to Q5.
- Q6: Use `summarise` to compare the mean number of days in the past 30 days mental health was not good by marital status and assign to Q6 as a dataframe. Round to two decimals. The mean should be called `mean_mental`. You should be able to confirm the results of Q5 after creating this.
- Q7: Calculate the means and standard deviations of `MENTHLTH` for those who have had a stroke and those who have not had a stroke only for those who do not have any kind of healthcare coverage. The output should be a dataframe assigned to Q7, and be formatted like this:

xxx	mean_mental	sd_mental
1	8.89	xxx.xx
2	xxx.xx	xxx.xx

- Q8: Run an ANOVA comparing how many times per week respondents took part in the exercise they spent the most time doing in the past month by marital status, and assign the TukeyHSD post-hoc test to Q8. (You may need to look up how to do this in R.)
- Q9: Calculate the variance in number of days a respondent drank alcohol in the past week for each type of physical activity or exercise spent in the past month for men. **Note: Pay special attention to how values are coded in the Codebook.**
    - Arrange in descending order, and include only the five with the highest variance in drinks.
    - The output should be a 6 X 2 dataframe, rounded to two decimals and look like:

EXRACT11	var_drinks
55	xxx.xx
xxx.xx	xxx.xx
xxx.xx	xxx.xx
xxx.xx	6.67
xxx.xx	xxx.xx
xxx.xx	xxx.xx

For the final section, you will choose four variables to explore in ways we have not. Complete the following:

- Q10: Address the values of any variables. For instance, is “none” equal to a value other than 0? Are there extra decimals implied?
- Q11: Remove any outliers. Briefly explain why you chose the method you used. Make sure to comment it out.
- Q12: Complete exploratory analyses doing appropriate visualizations with ggplot2.
- Q13: Run basic descriptive statistics.
- Q14: Finally, run an appropriate regression predicting one of those variables. Identify the best model.

Your answers must be clearly identifiable. Take time to tidy your code once you are finished. The easier it is for us to understand, the more partial credit you could receive.