

# BRFSS2015

For this assignment, name your R file BRFSS2015.R

- You may use the tidyverse and any other package, but indicate any additional packages you will use at the top of your file
- 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...)
```

## **IMPORTANT:**

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

*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.

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.

- Q1: How many people have any kind of health care coverage?
- Q2: What is the average "Number of Days Mental Health Not Good" for those in Pennsylvania who have numeric data? Make sure to change the response corresponding to none to 0.
- Q3: Compare only those who have and have not had some form of arthritis, rheumatoid arthritis, gout, etc. For those groupings, convert reported weight in kilograms to pounds. Then, compute the mean and standard deviation of weight in pounds. Use the conversion 1KG = 2.20462 LBS. Make sure the units are in pounds, not two decimals implied. The names of the variables should be mean\_weight and sd\_weight. mean\_weight should equal 183.04.

For the next questions you'll be exploring the relationship between marital status and minutes of total physical activity per week. Ignore those who refused to answer, weren't asked, or had missing data. You'll need to convert the marital status variable to a factor. Then:

- Q4: Remove outliers from minutes of total physical activity per week using 0.997 and 0.003 as criteria. What percentage of observations remain? Assign that value to Q4.

Answer the following questions using the dataset without outliers.

- Q5: Group by marital status and calculate the mean, standard deviation, minimum, and maximum of total exercise, to two decimals.
- Q6: Create a boxplot for total exercise by marital status.
- Q7: Run a regression predicting exercise by marital status. Assign the model summary to Q7.
- Q8: Run an ANOVA comparing exercise across marital status, and assign the TukeyHSD post-hoc test to Q8.
- Q9: Run a regression as in Q7, but add total fruits consumed per day. Based on the R-squared and AIC, what is the better model? Assign the better AIC value to Q9.

For the final section, you will choose four variables to explore we previously have not. Complete the following;

- Q10: Remove any outliers. Briefly explain why you chose the method you used. Make sure to comment it out.
- Q11: Address the values of any variables. For instance, is “none” equal to a value other than 0? Are there extra decimals implied?
- 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.

It should be easy for me to identify each portion of the final section.