Homework 5, My Name.

Please submit the solution on Canvas into the corresponding assignment (e.g. "Homework #1") in the form of R Markdown report, knitted into either of the available formats (HTML, pdf or Word). Provide only code and output. NO NEED TO COPY THE PROBLEM FORMULATION (!)

Problem #1

For the *fl_crime.csv* data, proceed to:

1. ~~Subdivide *urbanization* variable into three groups: ≤ 33, $(34, 66]$, $(66, 100]$.~~
2. For each urbanization group:
   a. ~~Provide a plot of crime rate (response) against education (explanatory). Fit simple linear regression of crime rate (response) onto education (explanatory), and add the fitted line to the plot.~~
   b. ~~Calculate correlation between education and crime rate.~~
3. ~~Compare the results from 2(a) with the line fitted for the whole data set: did the direction of the **overall** *crime-education* linear relationship change after conditioning on *urbanization* (at least for some *urbanization* levels)? Compare the **overall** correlation between *crime* and *education*, with the ones calculated in 2(b): did the directions change?~~
4. ~~If you witnessed any changes as a result of breaking the data down by groups according to *urbanization*, what is the name for that phenomena? What do we call the *urbanization* variable with respect to the *crime-education* relationship then?~~
5. ~~For the highest urbanization level, proceed to write down the fitted linear regression equation, interpret the intercept (does it make sense to interpret it?), the slope and the R^2.~~
6. ~~For any of the urbanization levels, does there appear to be an influential observation? If yes - what is it? In either case, proceed to outline the ways to deal with influential observations introduced in class.~~

Problem #2 (Continuation to Problem #2 from HW #4)

Broadband.csv contains data on each country's GDP (measured in billions USD) and the number of broadband subscribers.

1. ~~Plot the broadband subscribers against GDP, along with a fitted regression line. Provide the correlation between the two variables. Does there appear to be a regression outlier? Which country is it? Proceed to dispose of it and:~~
   - ~~Re-calculate the correlation. Compare it with the correlation you got prior to removing the outlier. What do you witness?~~
   - ~~Re-fit the linear regression: Would you trust the new slope value more than before the outlier was removed? Why?~~

2. ~~Fit a least absolute deviations (LAD) regression (for the whole data set, don't exclude the outlier). Comment on its slope value: how does it compare to the least squares regression slope values we had before (with outlier included & excluded)? Why do you think that is?~~
3. ~~For confirmation of part 2, provide all three fitted regression lines (least squares regression with & without outlier, least absolute deviation with outlier) on the same plot. Make sure to apply different coloring to those three lines.~~

Problem #3

1. ~~Extending on the example of Belarus president Alexander Lukashenko from slide #46, outline what kind of data we would need to collect in order to actually verify the claim about "deaths attributable to other health issues the patients already had". What kind of variables we would need to measure?~~

2. ~~Joe Rogan loves his weed. In fact, Joe loves it so much that he instantly became proficient in statistical reasoning whilst trying to defend marijuana legalization. Watch him drop this sweet (and actually pretty legitimate) "correlation is not causation" argument in the video below, between 3m:00ss and 4m:00ss~~

   ~~https://www.youtube.com/watch?v=Z8QlfsKWEYY&t=180s~~

   ~~Analogously to Lukashenko example from class, identify the main response and explanatory variables of interest, the likely lurking variable in play, and exactly how it affects the main relationship.~~

Problem #4

Presume that a professor at the University of Florida wants to compare the effectiveness of two teaching methods. She posts an announcement and gets 60 students to volunteer (*fl_student_survey.csv*).

~~1. How should she assign students to the treatment groups? Why? Provide the code executing this assignment, and demonstrate the balance between groups on gender, age and college GPA (similar to *heart* example in class).~~
~~2. If you were her, how would you conduct the exposure of students to teaching methods, and how would you measure the effectiveness of the method? It is an open-ended question.~~
~~3. Given the sampling design, can she generalize the measured effects of teaching methods onto a wider student population? Why?~~

Problem #5

~~3.91, 3.92, 4.2, 4.3, 4.9, 4.22, 4.26~~