**DS 710**

**Homework 9**

**R assignment**

1. In this problem, you will do further cleaning and analysis of the data from the 1995 US News and World Report on colleges and universities in the US.
2. In the Python portion of homework 9, you created a modified version of the data set usnews.csv.  Read the modified data into R.  Check the first few values of each vector to ensure that they were read accurately.
3. Examine the summary of each variable.  Identify any unrealistic values and set them to missing.  Write a sentence describing what you did, naming the colleges or universities affected.  (For example, “Listed ages less than zero (ABC University, XYZ College) were converted to missing data.”)
   * It may be helpful to use control flow or functions to help organize your work.
4. Find the mean percentage of alumni who donate, for private and public schools.
5. Test whether there is evidence that a higher percentage of alumni from private schools donate to their schools, compared to alumni from public schools.  State your conclusion in context.
   * Hint: In part c, we took the mean of this variable in each group. What does this tell you about what type of hypothesis test to use?
6. Use write.csv() to save your updated data set.  Consult the R documentation to set the arguments for the write.csv function.  Your output file should not have row names or row numbers, and it should not have quotation marks around the entries.
   * If you attached the data, be sure that the version you are saving to a file contains the updates you made in part c.
   * After you save the file, open it in Notepad++ or another plain text editor (*not* Excel) to verify that it does not have quotation marks around the entries.

2.  The data set cps.csv contains data from the 1985 Current Population Survey.

Dataset:  “Wages from the Current Population Survey,” <http://www.macalester.edu/~kaplan/ism/>, from Daniel Kaplan, *Statistical Modeling:  A Fresh Approach*.  Original source:  Berndt, ER.  *The Practice of Econometrics* 1991.  Addison-Wesley.

Metadata:  cps\_metadata.pdf, from p. 418 of *Statistical Modeling:  A Fresh Approach* by Daniel Kaplan.

1. Read the data into R and plot wages versus education.  Comment on the appropriateness of linear regression.
2. Perform the linear regression and examine the diagnostic plots.  Explain why transforming the wages variable is a good idea in this case.
3. The variable **wage** has units of dollars/hour.  Create a new variable, **time**, equal to 1/wage.  (So **time** has units of hours/dollar, or the length of time a person must work to earn $1.00.)
4. Plot time versus education.  Comment on the appropriateness of linear regression.
5. Perform the linear regression.  Based on these results, are you happy with your decision to pursue a master’s degree?  Explain.
6. Examine the diagnostic plots.  Which individuals appear to be outliers on the residual vs. predicted plot?  Re-do the regression without these individuals.  Does your conclusion change?

Submit a .docx or .pdf file to GitHub, containing your R code, R output (including graphs), and written interpretations and explanations. (Include your responses for problems 1 and 2 in the same file.)