Problem Set 1

Due January 23, 10:00 AM (Before Class)

Instructions

1. The following questions should each be answered within an R script. Be sure to provide many comments in the script to facilitate grading. Undocumented code will not be graded. Once your script is finished, please email Dominique at dlockett@wustl.edu.
2. You may work in teams, but each student should develop their own R script. To be clear, there should be no copy and paste. Each keystroke in the assignment should be your own.
3. If you have any questions regarding the Problem Set, contact the TA or use her office hours.
4. For students new to programming, this may take a while. Get started.

Working with data in R

For this assignment, I have subsetted the expenditures data for all campaigns and PACs available from Open Secrets. The reduced dataset is available at:

https://www.dropbox.com/s/z6gw9lvve6jogi5/Expends2002.txt

Before you begin, you should get familiar with the variables. The codebook for this dataset is available at:

http://www.opensecrets.org/resources/datadictionary/Data%20Dictionary%20Expenditures.htm

Below is a detailed listing of the data management tasks that you will have to complete for this assignment. You should provide the R script needed to execute each task with clear documentation.

1. Open the dataset as a dataframe. This dataframe should have the following properties: a) The column names should match the column names in the original dataset. b) The row names should correspond to the variable `ID` in the original dataset.
2. Change the variable name `TransID` to `Useless`.
3. Remove the variables `Useless`, and `Source` from the dataframe.
4. Change the variable `EntType` to a factor. How many levels does this variable have?
5. The variable `State` contains several obvious errors, as it includes non-existent state codes.
   - Identify observations that have non-existent state codes.
   - Write a script to recode these observations. Use the additional information in the dataset (candidate name, city, zip code) to correctly identify each state.
6. Remove all observations from the dataset where the variable `State` is missing. Report the number of observations after removing missing values.
7. Change the variable `Zip` into a numeric. Be sure to document what you do with missing cases. What is the mean of this variable?
8. Create new variables that contain the following information (you will be making several variables), and answer the questions:
   - The number of words in the `Descrip` variable. What is the median value of this new variable?
   - A variable containing the numeric portion of `CRPFilerid`. This variable should be of length 8 for all observations. What is the number of unique values of this variable?
   - A vector containing the first four digits of `Zip`. What is the most frequent value of this vector?

- A boolean indicating whether the `Descrip` variable contains the word "Communications" REGARDLESS OF CAPITALIZATION. Report the number of `TRUE` values in this boolean.
- A variable indicating that either `CRPFilerid` is "N" or that BOTH `Amount` is greater than 500 and `Descrip` is non-missing. Report the number of `TRUE` values.
- EXTRA CREDIT: A variable that provides the most common letter in the `Descrip` variable.

9. Write a script that subsets the data by state, and writes out a unique CSV file for each subset, where each file has a unique (and meaningful) name (hint: look at by() function).