ECON 386

Dr. Levkoff

Data Cleaning Project

The purpose of this project is to demonstrate your ability to work with and clean data. The goal is to produce a tidy data dataset that can be implemented for exploratory and predictive analysis. You will be randomly assigned to a group of 4 other data scientists (your classmates). You will need to submit a link to your group's Github repository (this can be hosted by one group member's account) which contains the necessary files for submission (described below). For each task folder, you should submit:

- Tidy data set(s) for each of the tasks described below
- An R script that implements the cleaning process as well as documents each of the steps taken (with careful comments indicating which group members are responsible for which steps) including the steps used to import the data into R.
- A code book that describes each of the variables, the data, and any transformations or work that you performed to clean up the data called **CodeBook.md**. Here is a link describing how to write a doc in markdown (.md): https://guides.github.com/features/mastering-markdown/
- You should also include a **README.md** in the repo with your scrips explain how all of the files work in each folder and how they are related.

Task 1 – Harmonizing, Merging, and Cleaning

One of the most exciting areas in all of data science right now is wearable computing. Companies like Fitbit, Nike, and many mobile device hardware developers are racing to develop the most advanced algorithms to attract new users. The data linked to from the course website represent data collected from the accelerometers from the Samsung Galaxy S smartphone. A full description of the data including how it was obtained can be found at the UCI machine learning repository using the link below:

http://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

The data (many files you will have to sort through) for this task can be found in the Task 1 folder within the Cleaning Data project folder in the ECON386REPO: https://github.com/slevkoff/ECON386REPO

After importing the data to your machine locally, you should create one R Markdown file called **Cleaning1.rmd** that does the following.

- Merges the training and the test sets to create one data set.
- Extracts only the measurements on the mean and standard deviation for each measurement.
- Uses descriptive activity names to name the activities in the data set
- Appropriately labels the data set with descriptive variable names.

- From the data set in the previous bullet point, creates a second, independent tidy data set with the average of each variable for each activity and each subject.
- Adequately documents all steps / transformations used in the cleaning process so that it is clear which group members conducted which cleaning processes (ie: which people wrote which code chunks).
- Use the **write.table()** function to save your results in a file called **tidy1.txt.**

Task 2 – Cleaning, Transforming, and Parsing/Partitioning

The next task will involve using the **Panel8589.txt** file found in the Task 2 folder. The first set of steps involves cleaning up the data set and transforming some variables. Create an R script called **Cleaning2.R** which will:

- Import **Panel8589.tx**t into R using the **read.table()** function (should make the process seamless without needing Excel as an intermediary assuming that you're able to at least get the file to your current working directory). Note that the column headers are labeled, but it is not clear what they are measuring (Y, X1, X2…).
- Then take a look at the summary statistics *carefully* from the research paper **Pasurka 2006** (Table 1, page 37) located in the Task 2 folder to figure out what is being measured in each column (and in what units) and read the description of the data in the paper. Remove any superfluous variables not to be used in the analysis, which you may assume will attempt to relate energy produced (assume this is the output) to resources consumed and pollution produced (assume these are the inputs).
- Convert all energy measurements (energy produced and heat contents) into *daily averages*, measured in Mwh.
- Convert all pollutants quantities, measured in annualized short tons, into *daily averages*.
- Convert all dollars (measured in 1973 $'s) into 2017 dollars.
- Add a factor variable indicating whether or not Phase I of the Clean Air Act had already been announced or not (the CAA Phase I restrictions were announced in 1990).
- Clearly document each of the above steps by appropriately commenting within the R script and save your cleaned data set using the **write.table()** function as **tidy2.txt**.

Using the **tidy2.txt** file created above, you will now create a few more tidy data sets:

- Create another data set called **tidy2_a.txt** that *averages* all variables across all years for each plant for the 11 year period so that the tidy data set has 92 rows of observations for all of the relevant variables.
- Create another data set called **tidy2_b.txt** that *aggregated* (adds) all variables within a particular year across all 92 plants so that the tidy data set has 11 rows of observations for all of the relevant variables.

The files you should push and include for submission in your group's repo before midnight of 4/17/2018 should be:

CodeBook.md, README.md, Cleaning1.R, Cleaining2.R, tidy1.txt, tidy2.txt, tidy2_a.txt, tidy2_b.txt