**12-709 Data Analytics for Engineered Systems (CEE) - Fall 2017**

**Homework 4 (Due Nov 15th at the beginning of class - 20% penalty per day late)**

**Basic R Model Building**

**Learning Objectives:** This individual assignment is intended to:

* Practice basic R operations and experience programming-based data analytics;
* Revisit familiar data analysis approaches and explore how done in R;
* Manage imperfect data and perform analysis on it in R;
* Demonstrate your ability to do data cleaning and data manipulation in R;
* Practice exploratory data analysis, visualization and analytics tools in R.

**General Directions:**

As mentioned in lecture and syllabus, writing skills are a key component of your homework grades (and your future careers). Please take care in writing, editing, formatting, and printing your work before submitting it. Your deliverable should look professional. While I have asked you to ‘document’ your processes, the final result of this assignment can be submitted as a series of homework questions and answers and does NOT need to be organized as a project report. Use charts and significant figures appropriately.

Notes:

1. In this assignment, you’re free to use any of the tools used before, e.g., Excel (or any spreadsheet program) Access (or any SQL-like database program). However, all of the analysis tasks should be completed with code. Ideally R, but feel free to use another. Also realize that it is acceptable to use a tool like Excel in parallel to ensure you’re getting the right answers and check your work – but we want you to show and submit code work.

1. You can do any data cleaning/transformation/etc., as needed, again using code. Of course, document what you have done.

You can use word or any text editor (including the R text editor!) for writing the text answers. You **can not (and do not need to)** use any other software or code, or any other graphic-assistance program (e.g., InDesign, etc.) for any of your work.

1. Deliverable: You will submit online to canvas. I presume you will have a mix of text and tables/graphics. It is your choice on the format to submit – for example you can submit a single PDF file of questions and answers, with a .R file separately showing all code, or you could submit a single PDF file with questions, answers, and code chunks all in the same file. Submitting .R files for your code is good – you do NOT need to use R markdown (.Rmd) files for this assignment. The goal is for you to generate results from running code – you cannot just take screenshots of sorted data frames as your answers.

**Note**: There was originally an infographic question on this homework, but I will push it to the next homework to keep the focus here on R skill building.

**Introduction**

It is often said that 80% of data analysis is spent on the cleaning and preparing data. It’s not just a first step, but must be repeated many times over the course of analysis as new problems about the data come to light. Real-world data tend to have errors, be incomplete, noisy, and inconsistent. Data cleaning methods attempt to normalize the data, fill in missing values, smooth out noise while identifying outliers, and correct inconsistencies in the data.

In this assignment, you’ll work on a “messy” dataset and perform analysis on it. As mentioned above, much of your time will be spent cleaning and manipulating the data. When you have completed the assignment, you will become familiar with the most practical R operations that are used in data analytics. Note this is our first homework in R, and you do not need to use any statistical tools beyond summary statistics in R (e.g., calculating mean/variance, counts, etc.). That said, you are welcome to use higher-level statistical tools if you feel appropriate.

**Data Description**

In this assignment, you are provided with data extracted from the US Environmental Protection Agency’s Air Markets Program (AMP, https://ampd.epa.gov/ampd/). Two report modules are applied to create the spreadsheets: the emissions data and facility attribute modules. They provide information on power generation and emissions at selected power plants in the US. One incomplete Data Dictionary (dataDictionary.xlsx) is also attached (as downloaded from the EPA website that provided this data). The single ZIP file contains all data you need.

Note the ZIP file provided has already been significantly improved and cleaned from what EPA provides (in the folder ‘Data to USE’), but you will need to do some more. However, we do not expect you to make all possible improvements – focus on high level important things, and don't worry about different plant names, IDs, etc. If you are concerned about the level of cleaning needed, ask! Note we also provided the raw data from EPA (before initial preparation for you) in the folder ‘Source Data to SEE’. There is also a file of ‘caveats’ in that folder that came from EPA, which is just for information for you.

**Submission Requirements**

Please use the 3-step format to document what you have done for all questions. Although for most questions we don’t require you to make plots, we recommend you use plots to illustrate your results. When done, please submit all code and created files used on Canvas (e.g., R files, Excel data dictionary, etc.).

**Part A. Exploring Familiar R Features**

**Question 1a.** Before doing any data manipulation of your own, look at the files in ‘Data to SEE’ and ‘Data to USE’ to see how EPA provided the data (and problems) and also what we have improved – write a brief summary of what we have already modified (about half a page).

**Question 1b.** Improve DataDicionary.xlsx using the provided datasets, with an emphasis on missing data. Please upload your improved DataDictionary.xlsx with changes emphasized.

**Question 1c.** After looking at the whole homework and what variables we will be working on, find at least six relevant variables you feel interested to explore during this process. Then, after doing data manipulation, perform summary statistics in R for these variables of interest. Document your results and findings.

**Question 2a**. How many different plants, EPA regions, NERC regions are there in the dataset? Find the top 3 states that have the largest number of plants.

**Question 2b**. Then in the top (#1) state, find all the plant names whose average heat input and gross loads every year are both higher than their 75 percentiles, but at least one of the three average pollutants are lower than their 50 percentiles. Make a table to present the results.

**Question 3a.** For each year in the dataset, find the total heat input (in MMBTU), total gross load (generation), and total emissions of each of the three air pollutants for all power plants in the dataset.

**Question 3b.** For the top 5 power plants in terms of gross load in 2015, plot over time all of the ‘total’ values from question 3a.

**Question 3c.** Similarly, for 2016, find all of the same total values for each EPA region and for each NERC region and plot the three total values of the top 5 EPA regions and NERC regions.

**Part B. Explorative analysis in R**

**Question 4.** Draw the distributions of average plant gross loads, average heat input, average total emissions of the three pollutants for all plants in different states. Together with total gross loads and pollutants of the state, try to find emission patterns of different states. Note that you may produce large amounts of graphs during this process, but you should answer with a few graphs that transmit the key messages.

**Question 5.** Design one (or two) metric(s) for describing the spatial location patterns of plants for each state. Requirements of the metric(s) are: (1) When any two plants in the state become closer to each other, the metric(s) should be increased; (2) If there’re more plants in the state, the metric(s) should be increased. What does the metric(s) measure? Find the top 3 highest states described by your designed measures. Do you think the metrics are fair? Do you have any requirements for the metrics to add for improving the metrics? Hint: To ease the question, you’re allowed to use distm((lon1,lat1),(lon2,lat2)) in geosphere package to calculate the distance between two points.

**Part C: Summary**

**Question 6** [5 pts]:Write a short summary (about one page in length) that discusses generally what aspects of this assignment were easy and hard to do in coding software such as R (and think about how long it would have taken to do similar things in Excel or Tableau). From your experience, in which situations would you want to keep using these tools and in which would you seek different tools?

**Optional Extra Credit/Bonus Question 7** [3 pts]: Please suggest specific edits that could be made to this homework (e.g., edit the questions given, suggest different datasets or points of focus, etc.). Don’t just say a question is vague – suggest a specific ‘fix’ for problem text.