Lab 1 - Stat 215A, Fall 2017

Submission: push a folder called *lab1/* to your stat215a GitHub repository by **9:00 PM on Thursday September 14th**. I will run a script that pull from each of your GitHub repositories promptly at 9pm so take care not to be late as late labs will not be accepted. The lab1 folder will contain *lab1.Rnw, lab1.pdf, lab1_ blind.Rnw, lab1_ blind.pdf*, and *R/*. The blinded files have your name removed but are otherwise the same. The *.Rnw* files will contain text and code but no code should appear in the .pdf output. The *R/* folder will contain scripts such as *clean.R* and *load.R* that help keep your workflow neat. Note that you may use .Rmd files instead of .Rnw, but your output must be a pdf document. Note: do not push the *data/* folder (as this may be too large for some labs).

1 Redwood Data Lab

The data for this lab is taken from Tolle et al., which can be found in the bCourses lab1 folder. You should read this paper before doing the lab and understand the source of the data. I have provided a template on BCourses as a guideline for the writeup. Since the template is intended to make grading easier, please do not deviate from it significantly without good reason. Please restrict your writeup to twelve pages, including figures. This is a strict limit: I will crop anything that appears beyond the twelfth page. In your lab1/ folder, please provide everything (except the data) that is needed for someone else to be able to compile your report and receive the exact same pdf. Your peers will be reviewing your code and attempting to recompile your report (they will manually add the data folder).

1.1 Exploration of Data

The original data can be found in the lab1 folder in bCourses. The files of interest are `sonoma-data-all.csv` and `mote-location-data.txt`. The goal of this task is to simulate receiving data in a collaboration. Your first goal is to explore the data on your own. Try to understand how variables behave, and what their relationships are. This also involves carefully cleaning the data set. Do not take data consistency or correctness for granted. The following is a suggestion on how you might proceed.

Your first task will be to check the data quality and explicitly address the issues we discussed in class, such as the data collection method and data entry issues (e.g. missing values, errors in data, etc). Please read the paper to understand how the sensor works, and write a paragraph to discuss the measurement of each variable you find interesting in the data. Please have at least 3 variables in your report, and those variables should be related to your findings in 1.3.

Bearing the data quality in mind, your second task will be data cleaning. This data set is quite raw - it contains some gross outliers, inconsistencies, and lots of missing values. Read the "Outlier rejection" section in the paper carefully and critically. You will need to do some cleaning of the data but don't blindly follow their method. Record in your report the steps you take and any evidence you use to support them.

Next, think of some questions you would like to ask of the data and use R to answer them graphically. Try to show what interesting findings can be gained from the data. You may show general patterns or anecdotal events. Using the entire dataset may be challenging. Try just a subset of sensor nodes or a day's worth of data. Again record in your report your process - include plots you make. Don't be afraid to try methods that are new to you and be critical of your own graphics.

1.2 Graphical Critique

Critique the plots in Figures 3 & 4. What questions did they try to answer? Did they answer them successfully? Did they raise any questions not addressed in the text? Would you change them at all?

1.3 Presenting findings

Choose three of your interesting findings and produce a publication quality graphic for each along with a short caption of what each shows. This is where I expect to see very polished graphics. Think carefully about use of color, labeling, shading, transparency, etc. This is your chance to do something innovative. If you are feeling bored or ambitious consider doing something dynamic or interactive (show a static version in the pdf) and provide either an additional .html document with the interactive graphic or a web link to where the interactive graphic is hosted.

1.4 Discussion

Did the data size restrict you in any way? Discuss some challenged that you faced as a result of the data size.