Lab 1 - Stat 215A, Fall 2021

Submission: push a folder called *lab1/* to your stat-215-a GitHub repository by **11:59 PM on Thursday, September 16th**. I will run a script that pull from each of your GitHub repositories promptly at 12am the 17th so take care not to be late as late labs will not be accepted. The lab1 folder you push will contain *lab1.Rmd, lab1.pdf, lab1_ blind.Rmd, lab1_ blind.pdf*, and *R/*. The blinded files have your name removed but are otherwise the same as the un-blinded files. The *.Rmd* 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. Your output must be a pdf document (maximum of 12 pages). Note: do not push the *data/* folder (as this may be too large for some labs).

You are welcome to discuss ideas with me or other students, but your report must be written up and completed individually. If you do consult with other students, please acknowledge these students in your lab report.

1 Redwood Data Lab

The data for this lab is taken from Tolle et al., which can be found in the *lab1/data/* folder of the *stat-215-a-gsi* GitHub repo. You should read this paper before doing the lab and understand the source of the data. I have provided a template 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 gsi repository. 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). Be sure to discuss all inconsistencies, problems, and oddities that you find with the data. Please also 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. Don't blindly follow their data cleaning method. You will need to make your own choices on how to most appropriately clean the data. Record in your report the steps you take to clean the data, and when necessary, explain why you cleaned the data in that way. Here, you may choose to bring in domain knowledge (from Tolle et al.) or common knowledge to support your data cleaning decisions.

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 Reality check

At this point, you've done some extensive data cleaning and made many judgment calls along the way.

Using common sense, check your cleaned data against some external reality. This could be from your prior understanding of the world or you could cite an external source of information. Either way, be sure to clearly state your assumptions. Does your cleaned data pass the reality check or are there issues? Discuss.

1.3 Graphical Critique

Your third task is to 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?

1.4 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.

1.5 Stability check

Next, select one of your judgment calls in your data cleaning or presentation of findings and perturb it somehow. By this, I mean to modify the judgment call in some way that seems reasonable to you. Clearly explain the call and your reasoning behind it, and explain the change you intend to make. Choose one of your findings from above. How does this change affect your finding? Create a before-and-after comparison visualization to show what – if anything – changes in the presentation of the finding and discuss.

1.6 Discussion

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

Recall the three realms of data science: data / reality, algorithms / models, and future data / reality. Where do the different parts of this lab fit into those three realms? Do you think there is a one-to-one correspondence of the data and reality? What about reality and data visualization?

1.7 Academic honesty

Finally, I ask you to draft a personal academic integrity pledge, addressed to Bin, that you will include with all of your assignments throughout the semester. This should be a short statement, in your own words, that the work in this report is your own and that all sources you used are properly cited, including your classmates. Please answer the following question: Why is academic research honesty necessary? If you feel it is not, make a clear argument why not.

1.8 Reproducibility

I will use the scf cluster to test the reproducibility of your code, by cloning into your repository and attempting to compile your reports. For your convenience, I provide you with the script I will be using to test your code. If you wish, you can use the script to test your code in the following way:

1. Go into your lab1 directory, for example using my directory:

```
cd /Users/omerronen/Documents/Phd/215a2021/stat-215-a-gsi/lab1
```

2. Run

```
bash test.sh .
```

To perform testing on scf, you may need to transfer the data file there which can be done using *SCP* (see detailed instructions on this link).

Lastly if you wish to use any R packages that require installation, please ask for approval here - Link. I will try to install it on scf, and approve if possible.

1.9 Additional Remarks on Grading

Due to the importance of good communication, readability and grammar of your write-up will also be part of your grade. Moreover, we emphasize that the domain problem and context matters greatly in practice. While we do not expect you to be an expert on redwood trees a priori, we do expect you to learn a little about redwood trees through reading Tolle et al and to incorporate some bits of this domain information throughout your report. Ideally, in every (sub)section, you should try to ground your discussions of your findings/analyses in the domain context. A great report will also tell a story, where the writing flows from one section to the next and each plot has a reason for being included.