Stat 222: Redwood Project

1 Data Description

For this project, your primary data source will be from a wireless sensor network, which captured spatial and temporal information (e.g., temperature, humidity) from the microclimate around a coastal redwood tree [1]. Wireless sensor networks are becoming increasingly prevalent in a growing number of areas including health care, environmental and industrial monitoring, as well as home monitoring and automation.

An important goal for this project is to give you more hands-on experience working with the types of messy, incomplete, and inconsistent data that you will encounter in real world applications. It is also designed to give you more time to develop your ability to understand and critique statistical graphics as well as to provide you an opportunity to practice designing good graphics to convey information and reveal patterns. Finally, the project will provide you with the opportunity to practice reading and writing about applied statistical data analysis.

2 Your Assignment

This is an individual project. While you may talk to other students in the class about the assignment, you will be responsible for producing your own work. This includes all code, figures, and text.

I've created a Git repository for you with the following structure:

```
redwood
|-- data
|   |-- mote-location-data.txt
|   |-- README.md
|   |-- sonoma-data-all.csv
|   |-- sonoma-data-log.csv
|   |-- sonoma-data-net.csv
|   `-- sonoma-dates.Rda
`-- redwoods-sensys05.pdf

1 directory, 7 files
```

The files of interest are `sonoma-data-all.csv` and `mote-location-data.txt`.

Exploration of Data

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.

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 do **not** 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 or Python to answer them graphically. Try to show what interesting findings can be gained from the data. You may show general patterns or anecdotal events. Experiment with linked plots. 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.

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?

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.

Discussion

Did the data size restrict you in any way? Discuss a new aspect of large data sets from the lab.

Timeline and logistics

Here is the tentative schedule:

Monday	Wednesday
(2/8) Start Redwood project	(2/10) Workflow I
No class	(2/17) Workflow II
(2/22) Poster presentations	(2/24) Workflow III
(2/29) Redwood report	

You should spend

3 Report Details

I have provided a template in your individual class repositories for the writeup. Since the template is intended, in part, to make grading easier, please do not deviate from it without good reason. Please restrict your writeup to twelve pages, including figures. This is a strict limit.

References

[1] Gilman Tolle, Joseph Polastre, Robert Szewczyk, David Culler, Neil Turner, Kevin Tu, Stephen Burgess, Todd Dawson, Phil Buonadonna, David Gay, and Wei Hong. A Macroscope in the Redwoods. In *Proceedings of the 3rd International Conference on Embedded Networked Sensor Systems*, pages 51–63. ACM, 2005.