Data Cleaning: Redwood Lab

February 3, 2025

Today’s plan

1. **Continuation of redwood data cleaning**

Lab 1 Announcement

New due date: Lab 1 is now due this **Sunday, Feb 9 at 5pm ET.**

- Originally, Lab 1 asks for 2 exploratory data analysis plots.
- Given the timing, Lab 1 now only **requires 1 exploratory data analysis plot.** The second plot is extra credit.

Clarification: Think of the lab assignment as a data analysis report, rather than answering questions in a problem set

- Output should be cohesive paragraphs, rather than disjoint responses to the each of the questions.

Last time

1. **Load in the data**
   a. Epoch/dates and redwood datasets have already been filled out for you.
   b. You need to fill out the `load_mote_location_data()` in the load.R/load.py file.
      - Uploaded "a solution" to `dsip-s25` GitHub repository
      - **Main challenge:** header row uses different *delimiter* than other rows

Last time

1. **Load in the data**
   a. Epoch/dates and redwood datasets have already been filled out for you.
   b. You need to fill out the `load_mote_location_data()` in the load.R/load.py file.

Before proceeding, read all available documentation!

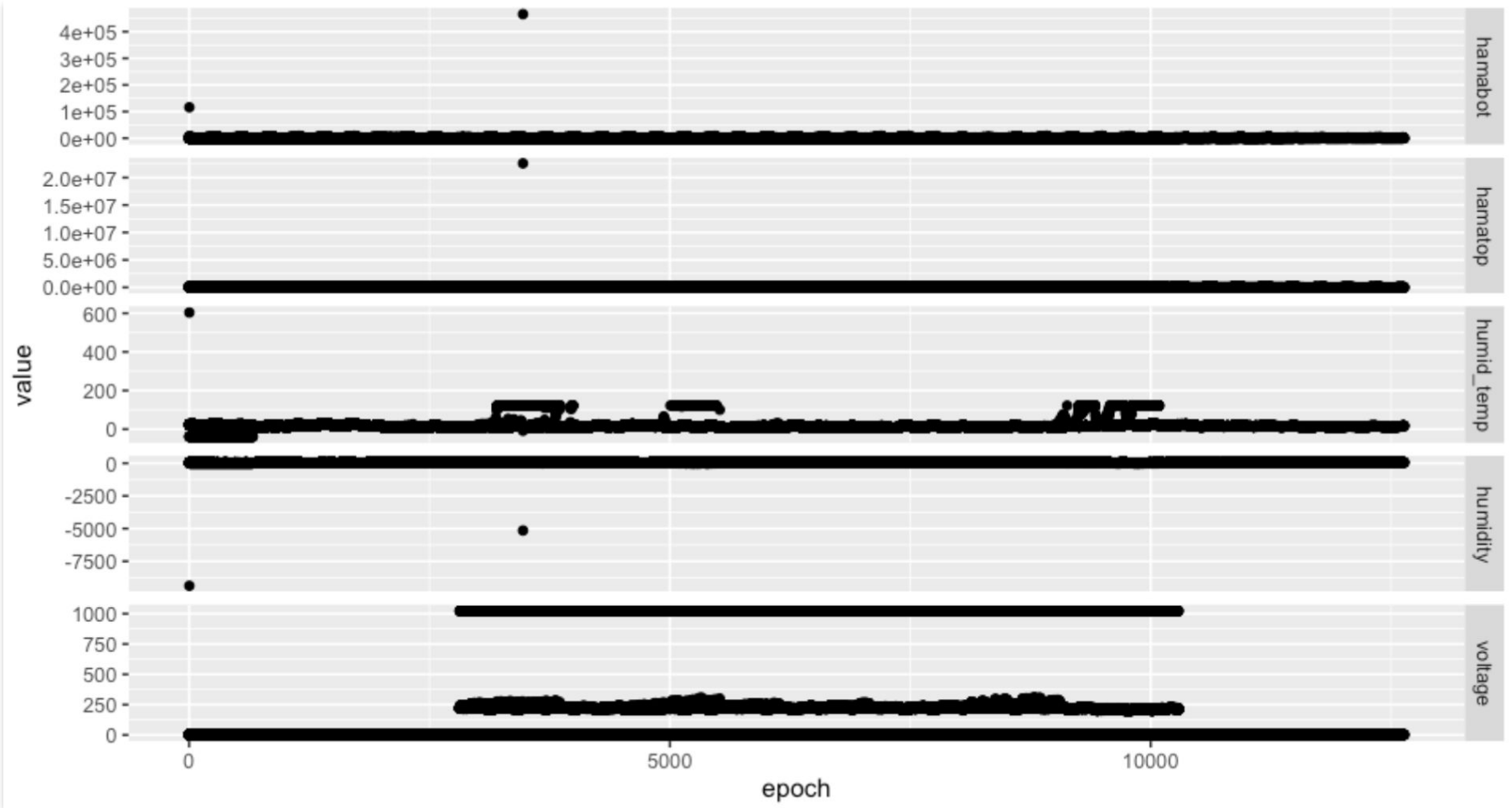
2. Look and "play" around with the data in order to:
   a. Try to **identify as many issues or oddities** with the data as you can. *Hint:* there are many!!
   b. Also **think** about how you might address these issues and clean the data. Jot down these ideas, but no need to take the time to implement it *yet*.
      - *Time permitting*: you can start implementing your ideas, but prioritize identifying the issues over fixing them.

Redwood Data Issues: Our First Look

- No units! Lack of documentation
  - Please do not blame the messenger...
- Lots of NAs
- Strange humidity values
- Strange voltages
- Depth, parent, voltage: all variables relating to the data collection process (i.e., wireless sensor network)
- ... to be continued as you work through lab 1

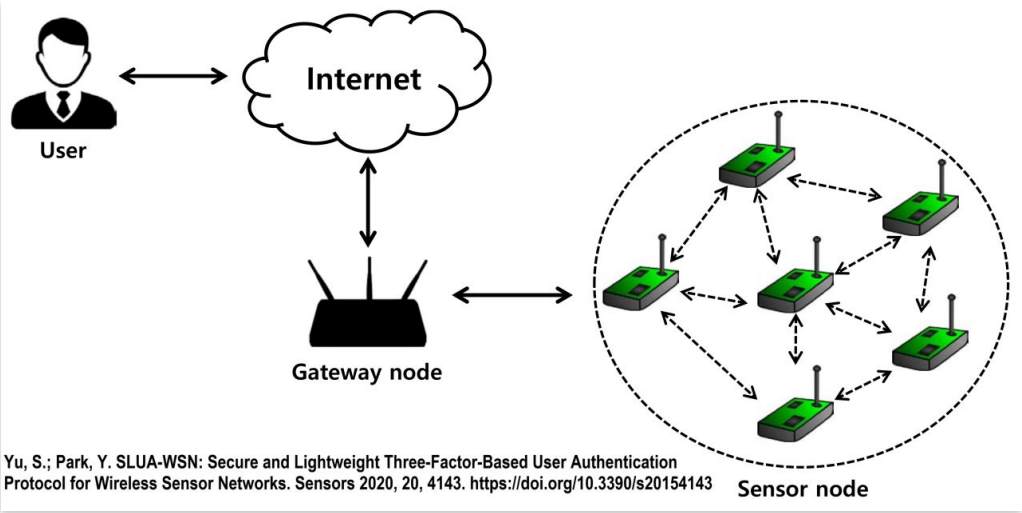
What are some techniques to identify these data issues?

- Read all documentation
- Look at summary statistics, NAs, duplicates
  - `skimr::skim()` in R
- Use domain knowledge/common sense
  - Ask lots of questions to the data? Do the answers make sense?
- **Plot, plot, plot, and iterate**
  - Scatter plots, histograms, density plots, heatmaps, ...


hamabot
hamatop
humid_temp
humidity
voltage
value
epoch
4e+05
3e+05
2e+05
1e+05
0e+00
2.0e+07
1.5e+07
1.0e+07
5.0e+06
0.0e+00
600
400
200
0
-2500
-5000
-7500
1000
750
500
250
0
5000
10000

What to do about the redwood "all", "log", and "net" datasets?

- Initial reaction: look at the "all" dataset
- **Q: What is the difference between the "all", "log", and "net" datasets?**
  - *Hint:* Related to the data collection process



How the data was collected should inform how we clean/preprocess the data

Your task for today

Goal: create a single merged, clean-ish dataset (to be our primary data source for EDA)

Note: some *very basic* data cleaning has already been done for you in `clean_redwood_data()` (can find in `dsip-s25/`)

Your task: Create a `merge_redwood_data()` function in clean.R/clean.py:

1. *Inputs:* the cleaned network, log, mote, and dates data
   (i.e., the output of the `clean_*()` functions)
2. Concatenate network and log sensor data, and add column named "source" indicating whether each observation came from the network or the logged dataset
3. Remove duplicate observations (i.e., copies in both the local log and network data)
4. Merge redwood sensor data with mote and dates data
5. *Output:* a single merged, clean-ish dataset

Code solutions will not be uploaded, but the resulting data frame is on Canvas.

+ This is not the only way to merge the data appropriately
+ You are not allowed to simply load this data frame and use it in your lab
  (Remember: assume that I only have access to the *raw* data)

Your next steps

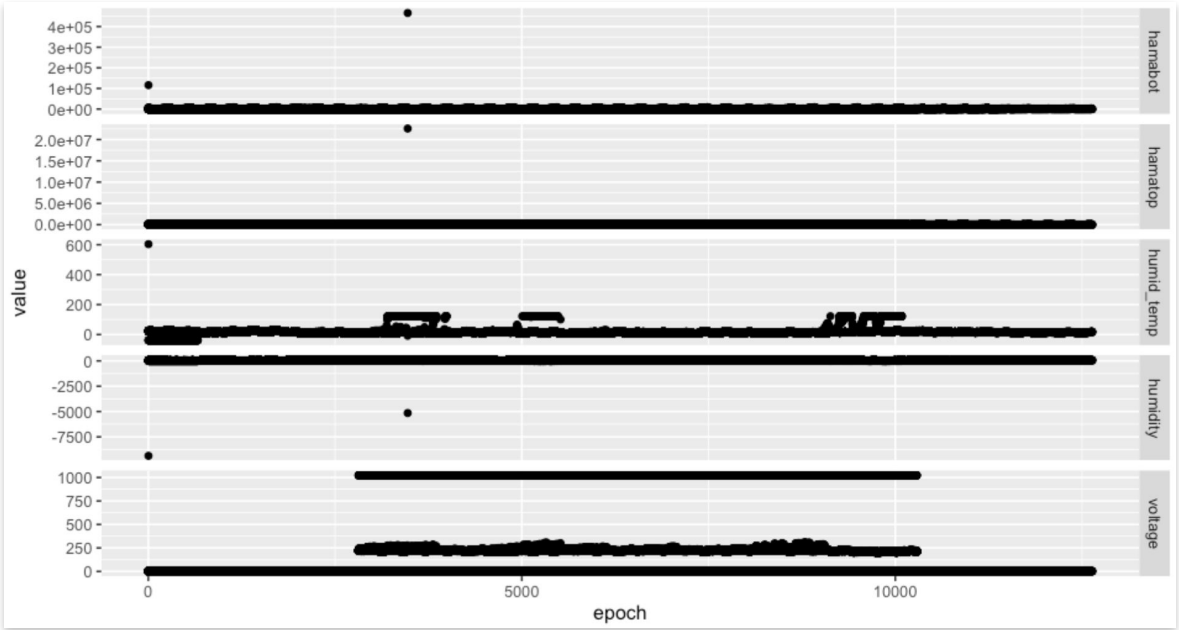
Once you have created your merged dataset, there is probably more data cleaning that needs to be done, e.g., deal with

- Lots of NAs
- Strange humidity values
- Strange voltages
- ...

What plots would you like to see to inform your data cleaning?

See snippet of the merged redwood data here:
https://tinyurl.com/dsip-redwoods-merged

Example:



Data cleaning to be continued...

The data cleaning journey is to be continued as you work through lab 1 on your own.

Remember: there is more than one right way to clean the dataset

Some of your to-dos:

- Deal with voltages (e.g., convert them to the same scale/unit)
- Identify and remove erroneous measurements or "outliers"
- Other data cleaning steps that you think are appropriate – you have free reign!
  - There are more issues than what has been discussed in class

Document your data cleaning steps and explain **why** you chose to do it that way

- This includes documenting and justifying data cleaning steps that we've done in class
- If you don't like how we cleaned it in class, that's also totally fine.
  Just document and justify.

Next Time: Exploratory data analysis [chapter 5 from VDS textbook]

Lab 1 Announcement

New due date: Lab 1 is now due this **Sunday, Feb 9 at 5pm ET.**

- Originally, Lab 1 asks for 2 exploratory data analysis plots.
- Given the timing, Lab 1 now only requires 1 exploratory data analysis plot. The second plot is extra credit.