STAT 215A Lab 1

9/13/2018

1 Redwood Data Lab

1.1 Introduction

This report is based on the Redwood Data obtained in 2004 where researchers developed a wireless sensor network around a 70-meter tall redwood tree in great temporal and spatial detail. The journal paper published in 2005 examined the performance of the sensor network and provided guidance for future deployment.

1.2 Exploration of Data

1.2.1 Measurement of selected variables

In this section, I will discuss the measurement of the two PAR (photosynthetically active radiation) variables and temperature, as they relate to the questions I want to explore and the outlier detection process.

PAR was chosen over TSR (total solar radiation) for its biological relevance, and its measurement was straightforward. Incident PAR is related to direct sunlight, while reflected PAR is a measure of ambience radiation. As demonstrated in Figure 3 (journal paper) and explained in the paper, incident PAR readings show a bimodal distribution while the reflected PAR readings were less sensitive to the availability of direct sunlight. This is a remarkable finding as it reveals that variables similar in nature could indeed have very different distributions.

Temperature, on the other hand, was carefully calibrated in a controllable weather chamber with conditions "between 5 and 30 °C and between 20 and 90 %RH". This variable was helpful in data quality control; its range was easily determined because temperature was measured to a high precision and it was possible to corroborate the data with external data such as weather report. As a result, any anomaly could be easily detected and used as a basis for discarding problematic data points. This was indeed the case as described in Section 5.6.

1.2.2 Data quality

We first examine `sonoma-data-all.csv` with boxplots in Figure 1. Clearly many outliers exist, as the boxes are barely visible.

1.2.3 Data cleaning

We start by removing all entries with missing data. According to the Section 5.6 of the paper, outliers were detected based on temperature readings due to battery voltage problems (which should range between 2.4 V and 3 V) and humidity readings above 100 %RH. While the ones in `sonoma-data-log.csv` ("log" dataset) are around the order of magnitude 10^0 , voltage values in `sonoma-data-net.csv` ("net" dataset) seem to be multiplied by 100 (unless all of them were outliers). The two datasets were therefore treated using different voltage value criteria.

Upon closer investigation, certain fields contain values that were too large compared to their definitions in the paper and appeared to be zero-based powers of 2, suggesting mechanical failure. For example, one entry has a `nodeid` of 65535, 10,359 entries have `parent` node ID of 65535, and 10,662 entries have `depth` of 255. In the "net" dataset, a `humid_temp` value of 122.1530 is often associated with these failures and needs to be

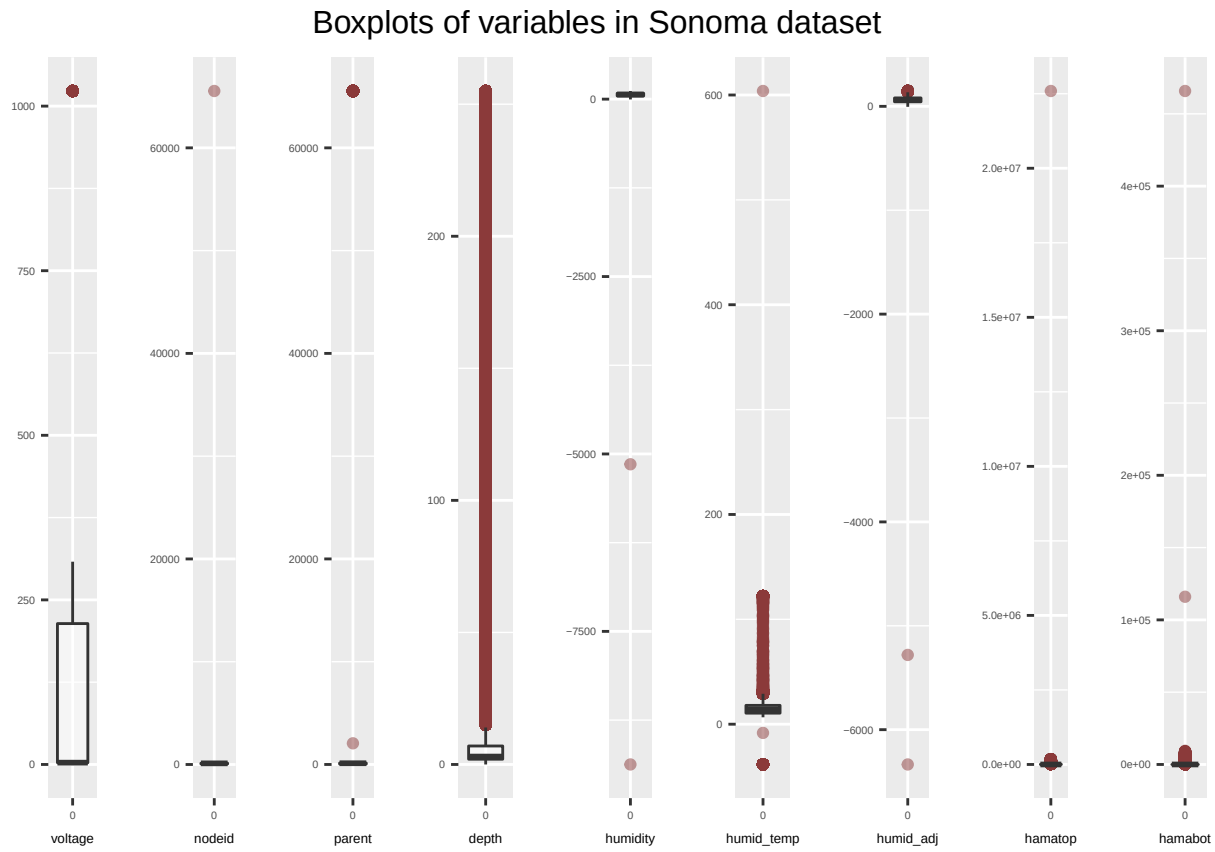

Figure 1: Boxplots of key variables, before cleaning.

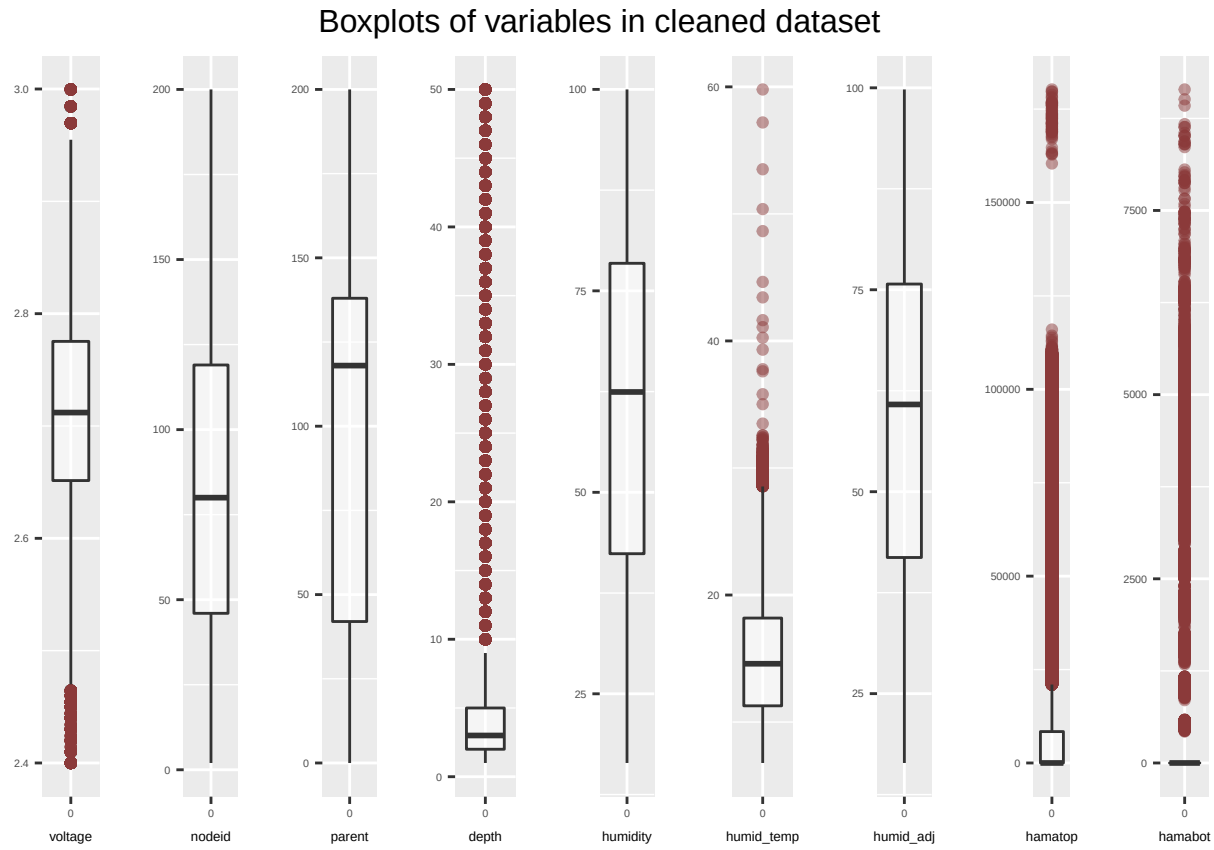

Figure 2: Boxplots of key variables, after cleaning.

removed; this reading most frequently happens to nodes with ID 3, 78 and 123. Since there is not enough information to determine whether these sensors entirely malfunctioned, the remaining data points of these nodes were kept.

Temperature above 60 degrees Celsius would also be problematic, so these points were removed. Further, we don't have any information on the `depth` variable, so a conservative approach was taken to only eliminate points with `depth` greater than 50.

It is also noticed that the ranges of PAR values could vary up to several orders of magnitude and do not fall in the ranges described in Table 1 in the paper. Unfortunately, Table 1 in the paper also did not provide a unit for PAR values, preventing readers from discerning abnormal values.

Duplicate values based on `nodeid` and `epoch`, the latter of which is a proxy of record time, were found in both datasets and removed.

The result of data cleaning is demonstrated below in Figure 2. Except for the variables on which the paper did not provide sufficient information, outliers were successfully removed.

1.2.4 Proposed findings

Without knowing a great deal about plant biology, I'm interested in the relationship between incident PAR and height. Breaking down data points by other variable values should provide more than three interesting findings for the purpose of this report. This also requires joining the sensor data with location data based on node ID.

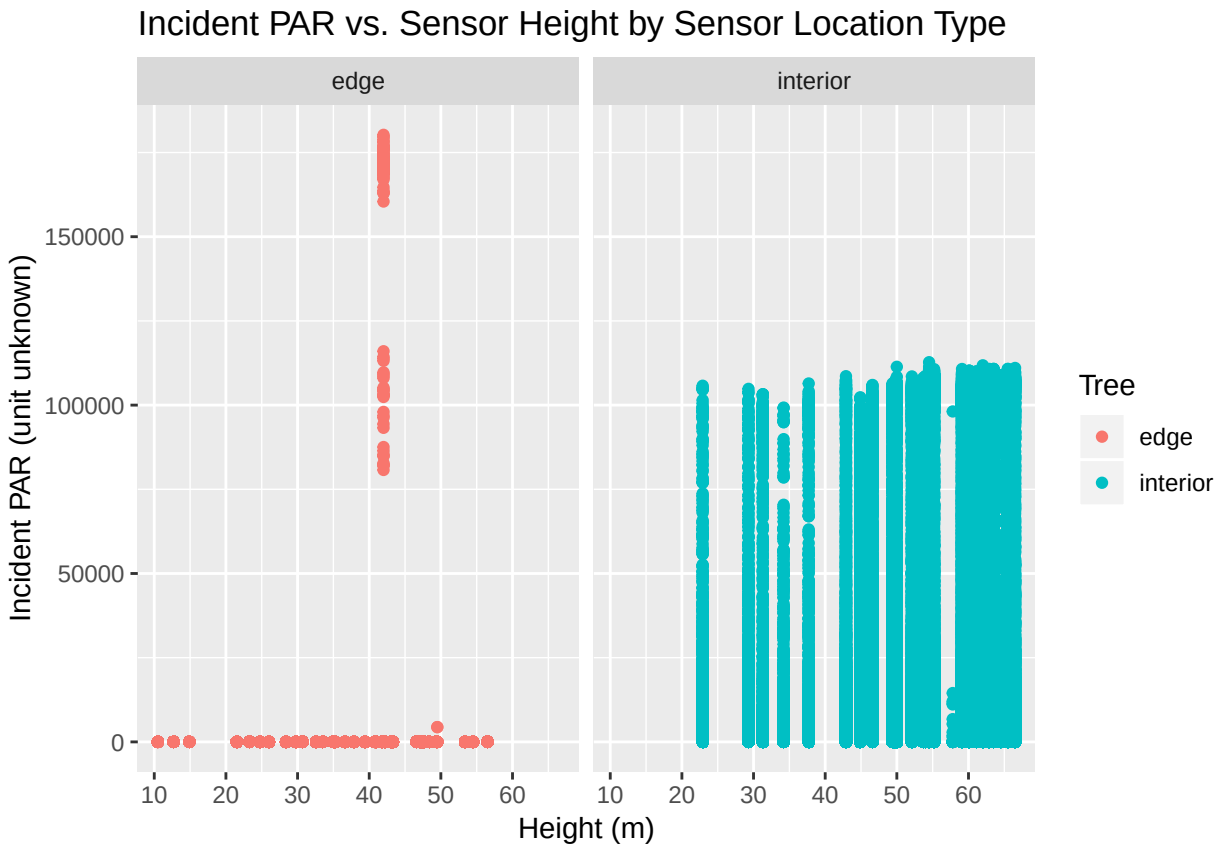

Figure 3: Incident PAR vs. Sensor Height by Sensor Location Type

1.3 Graphical critique

In the paper, Figure 3 was trying to provide a summary of distributions of key variables, while Figure 4 presented examples of temporal and spatial variations of key variables. The figures successfully addressed these basic questions, but the command was slightly problematic. For example, Figure 3 was confusing as the "value" dimension was not always on the same axis. It was also peculiar to see a histogram described as a projection onto the *value* dimension. In Figure 4, coloring was extremely confusing. It was not clear what pink and blue meant on the scatter plots; the blue vertical line in the line plots, meant to indicate the moment in time in which the spatial snapshot was made, was colored identically with the average value line, and readers might be misled into thinking that the blue vertical line was a sensor malfunction. I would change the axis orientation, wording, and color scheme for Figure 3 and Figure 4.

1.4 Presenting findings

We first look at how incident PAR changes with height and where sensors are located relative to the tree in Figure 3. "Edge" sensors are not particularly interesting because almost all the incident PAR readings were zero except for a few height values. For the "interior" sensors, the incident PAR could take almost every value between 0 and 100,000 at each height value. This is a more interesting regime than the "edge" sensors.

Next, we further break down the data points of "interior" sensors into groups based on times of day in Figure 4. It seems that hour of day is more indicative of incident PAR levels.

Next, we attempt to perform OLS on incident PAR and height grouped by four artificially created "time of day" labels. This approach has its limitations as incident sunlight varies continuously over time rather than

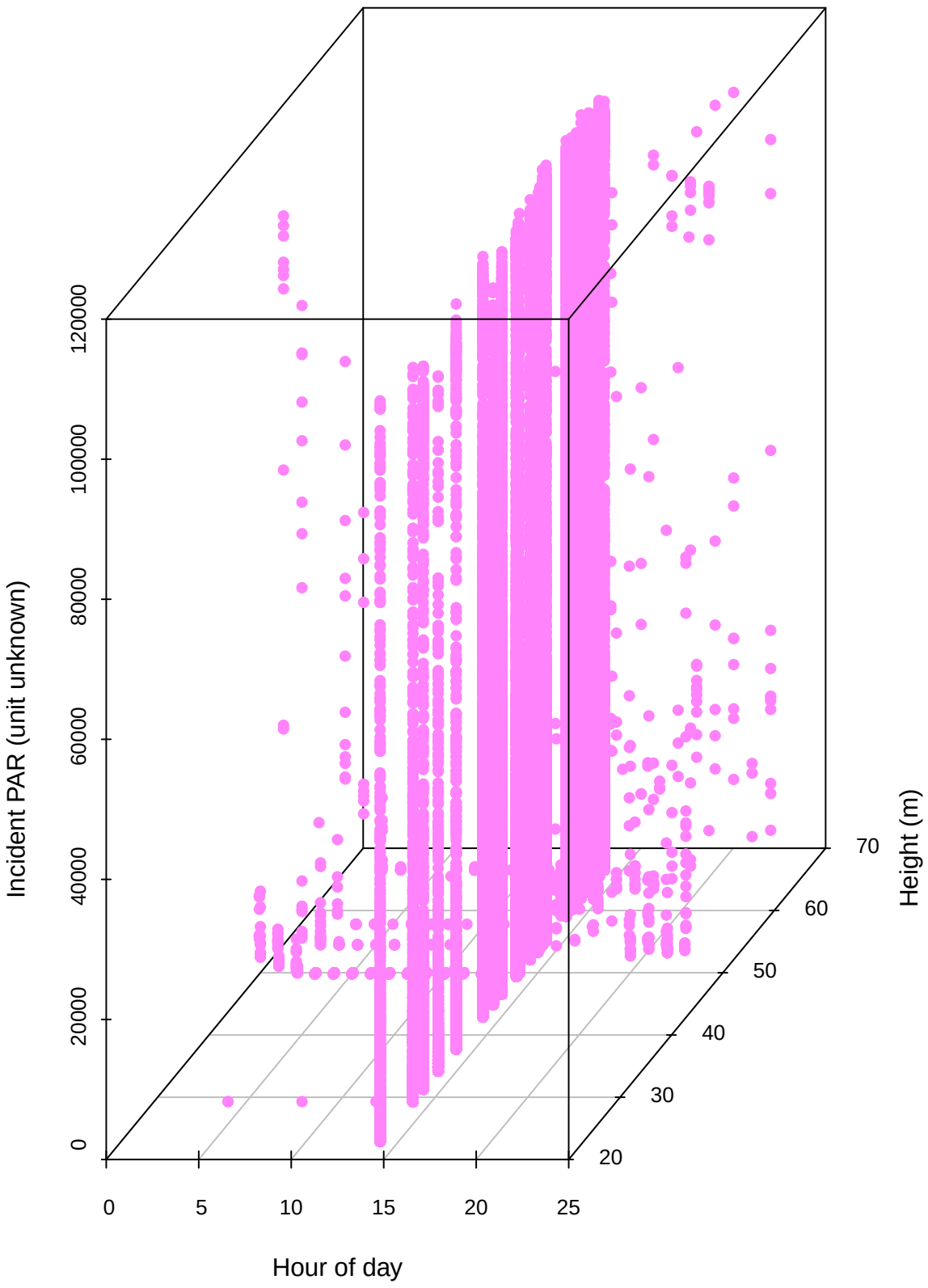

Figure 4: Incident PAR vs. Sensor Height and Hour of Day for Interior Sensor

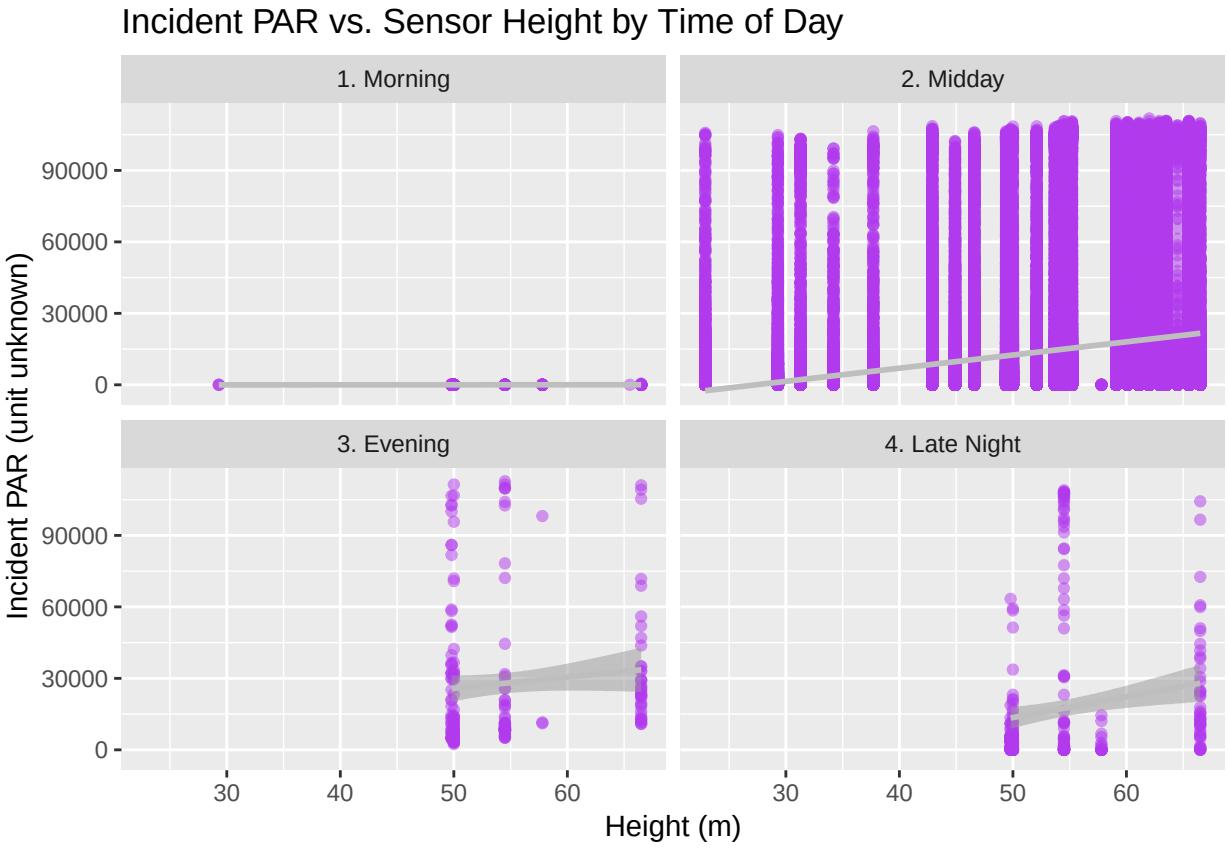

Figure 5: Incident PAR vs. Sensor Height by Time of Day

in a discrete manner. However, it is simplistic enough that many engineering areas still take such approaches in modeling.

The four times of day are defined as: morning (4 AM - 10 AM), midday (10 AM - 4 PM), evening (4 PM-10 PM), and late night (10 PM - 4 AM). The results as shown in Figure 5 are fairly bizarre in a few ways, suggesting that the sensor time was not in local timezone and required additional adjustments:

- There was no incident PAR in the morning.
- Incident PAR was often non-zero in late night.

Therefore, we might not be able to get more useful information out of this analysis until the issue has been corrected.

1.5 Discussion

Data size was not a particular challenge for aggregated summaries, but it was problematic for visualization purposes, which affects data cleaning and subsequent exploration. Plotting in R was especially challenging for me as I am used to working in Python, and I feel that the time costs associated with making mistakes prevented me from testing features in R extensively and making the best-looking plots.