Lab 1 - Redwood Data, Stat 215A, Fall 2018

Blind

September 13, 2018

1 Introduction

This report analyzes the Redwoods dataset collected in 'A Macroscope in the Redwoods' project, which is a case study of a wireless sensor network that recorded 44 days in the life of a 70-meter tall redwood tree. Inspired by the original paper, this report presents a renovated way to clean, explore and analyze the dataset.

2 The Data

2.1 Data Collection

Using two systems, the wireless sensor network TASK and the local data logging system, researchers recorded 44 days (April 27 2004 at 5:10 PM to June 10 2004 at 2 PM) of microclimate information of a 70-meter tall redwood tree, at a density of every 5 minutes in time and every 2 meters in space. A total of 72 nodes were deployed, which measured air temperature, relative humidity, incident photosynthetically active solar radiation (PAR) and reflected PAR. Readings from the TASK framework were taken every 5 minutes, while the local data logging system, as a backup of TASK, recorded every reading taken by every query until its 512 kB flash chip was full. Data from TASK and local logging system were named as 'net' and 'log' respectively. Because of the existence of missing values in certain time periods, we will use both datasets to re-generate the complete picture of the microclimatic dynamics around the tree.

2.2 Data Cleaning

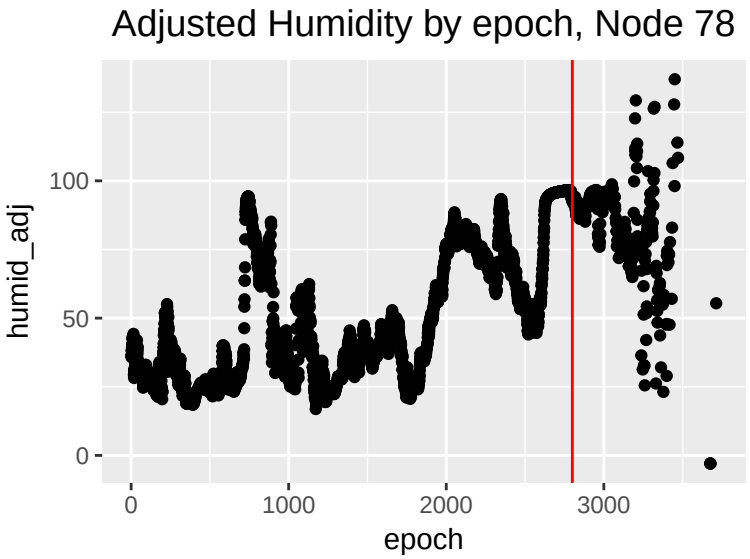
Data Cleaning is a two-part process. First, we use 'all', the simple binding of log and net, to perform variable checking, making sure all variabels of interest are in the right range, converting epoch and voltage to the right scale and removing outliers when necessary. Second, we separate out 'net' and 'log', and combine them together on a node-epoch basis. Additional pre-processing is then performed on the combined and cleaned dataset.

2.2.1 Variable Checking, by epoch and node

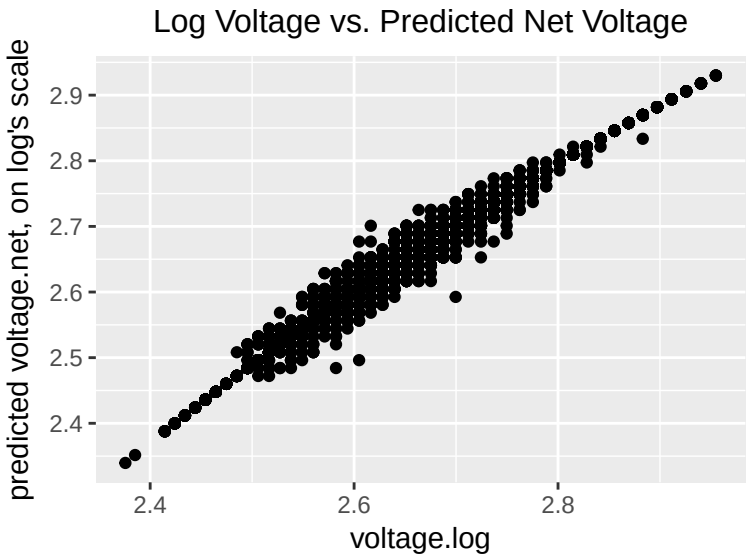
The variables of interest are adjusted humidity(humid_adj), temperature(humid_temp), incident and reflected PAR(hamatop and hamabot). The general assumption of variable checking is that these measurements should be within their normal ranges (i.e., one obvious example, in early summer, North California's temperature should not be negative) and move smoothly over time.

- Adjusted humidity: after plotting adjusted humidity over time(epoch), we notice two groups of values deviated from the common crowd: the negative ones and those > 100. For each of the two groups, we find the problematic nodes behind these values and investigate the nodes one by one. For example, all observations from node 29 have the same values, which is clearly a recording error. We thus drop node 29 values in its entirety. In other cases, nodes started to record wrong values after several epochs. For example, as the graph below shows, node 78's recording for adjusted humidity and temperature went wrong when epoch > 2800, 2950, respectively. In these cases, we identify the 'epoch' when values

went wrong and set the outliers to NA. We repeat this process on node 123, 141, 198, 42, 145, 3, 118 to identify and remove all the outlier points.



- Temperature: while investigating on adjusted humidity, we find that usually when humidity went wrong, temperature values went wrong as well. So along side adjusted humidity outliers, we removed temperature outliers as well. After the variable checking of adjusted humidity is done, we plot temperature over time, and gladly see that all values are within normal range and the patterns seem normal. No more cleaning needed for temperature.
- Incident PAR: after plotting incident PAR over time, we find a group of suspicious outliers when hamatop > 150000. All of them come from node 40. Further investigation shows that the values went wrong in node 40 when epoch > 38. We thus removed all the outliers in node 40.
- Reflected PAR: the reflected PAR vs. epoch plot shows that, all values are within normal range with reasonable pattern. No more cleaning needed.
- **Epoch Conversion**: code from clean.R and load.R is used to create epoch_df, a mapping table between epoch and the actual date time. We then merge epoch_df with our data to add exact date and time information for each recording. Datetime will be useful in later exploration.
- **Voltage**: after checking the distribution of voltage, we find two suspicious group of values: > 1000 and < 2. Further investigation shows that they come from node 134, 141, 145, 128, 142, 143. Node by node checking shows that albeit their voltage values are strange, their humidity, temp, hamatop and hamabot seem normal. We thus decide to remove the abnormal voltage values and leave other values as they are. A more prominent problem with voltage is that, the voltage in net and log is not on the same scale. The ones in log are mainly between 2~3 while those in net are in the $200 \sim 300$ range. A quick scatterplot shows that their relationship is linear. So we use linear regression to convert the net voltage to the same scale as log. As the plot below shows, though the model is very basic, it does a great job at **voltage conversion**.



2.2.2 Dataset Combination and other Pre-Processing

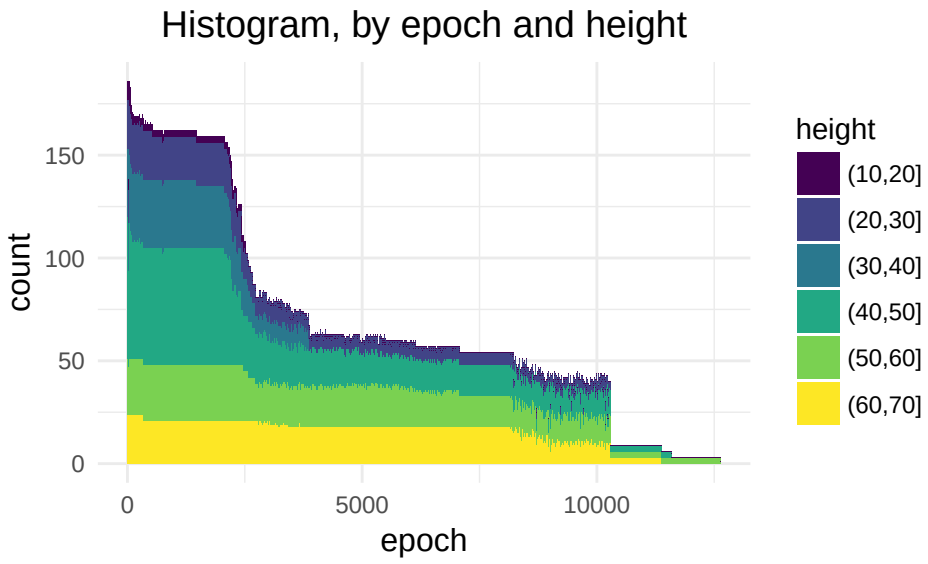
Next, using their differences in voltage (the original values instead of the converted ones), we separate the dataset into net and lot. Because the two systems differ in availability during the 44-day experiment, both of them have missing values for certain nodes in certain epochs. Using nodeid-epoch as the unique identifier, we join the two datasets, hoping to fill in as many missing slots as possible. In the case of overlapping, that is, both net and log have values for a given nodeid-epoch pair, we discover that the values are only different on a negligible level (i.e., < 0.1 level), we thus decide to use net as the value for this given pair, for the sole reason that the TASK framework was the core of the researchers' study so we have more faith in their recordings. Duplicates are also dropped in this process.

We then join the combined dataset with loc, to get the spatial information of node. At the very last step, we perform additional pre-processing, such as extracting 'hour' from time for later use.

2.3 Data Exploration

We start with the histogram of epoch, colored by height. From the plot, we can generally divide our data into three phases:

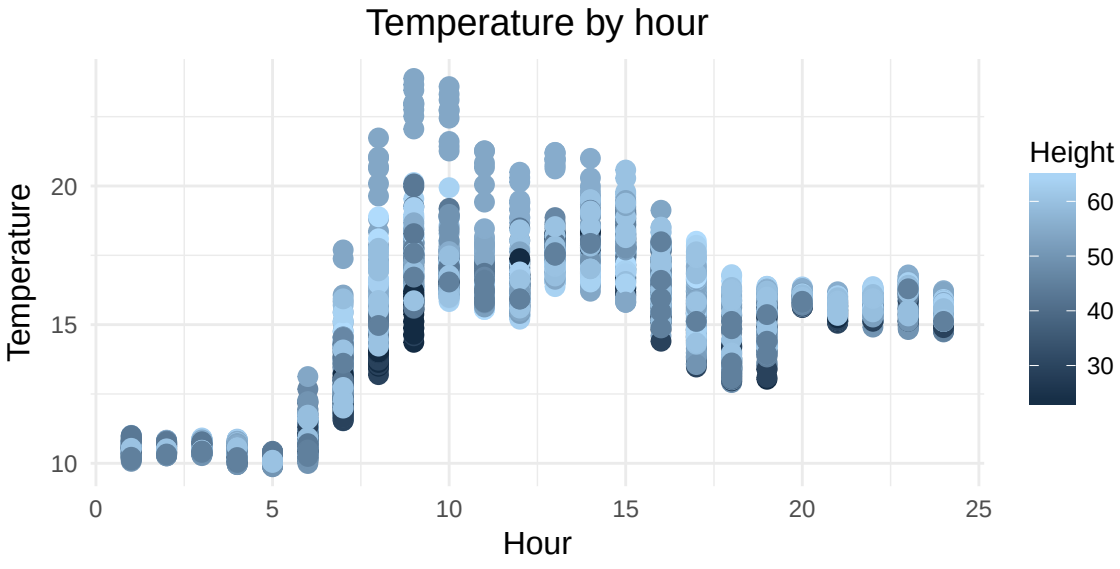
- Phase 1: epoch < 2500. From the very start to roughly epoch $= 2500$, we have a relatively large and stable amount of observations at each time period and height bin. As indicated by the color, most observations are in the 40 m ~ 50 m range, while 30 ~ 40, 50 ~ 60 and 60 ~ 70 come as close seconds, each has approximately 50 observations at each epoch. We have very few observations at the 10m ~ 20m range, the bottom part of the tree, as shown by the thin purple layer on the very top.
- Phase 2: $2500 <$ epoch < 10000. The quantity suddenly drops by half, when epoch moves past 2500 (approximately). This is largely due to the plummet in 30 m ~ 40 m range, which almost disappears as epoch moves towards 4000. Apparently nodes located in this range ran into some problems. The 40 m ~ 50 m also drops by half, as does the 20 m ~ 30 m range. The higher levels, 50 m ~ 60 m and 60 m ~ 70 m, seem unaffected, as their number of observations remain largely stable.
- Phase 3: epoch > 11000. The number of observations plummets, even hardly. Once we past epoch $=$ 11000, observations in all height bin drop dramatically, the bottom levels disappear from the plot and only very thin layers of 40 m ~ 50 m, 50 m ~ 60 m, 60 m ~ 70 m remain.



2.3.1 Deep Dive: Measurements on May 20

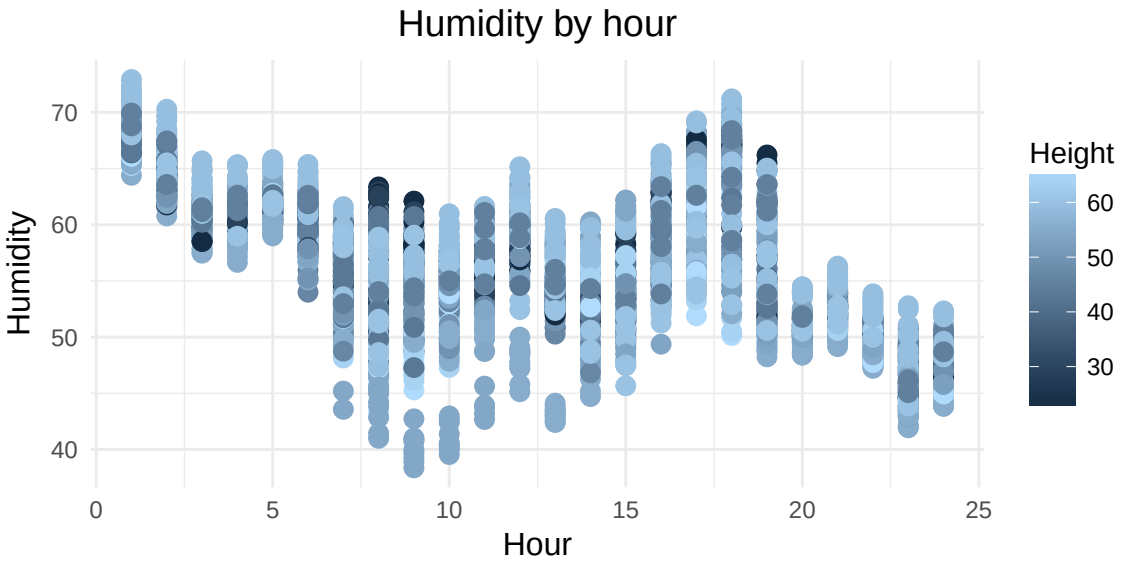
We now take one day, May 20, to get a closer look of the four measurements, as well as their changes across time and height. The epoch of May 20 is in the 9000 range, which is in Phase 2.The reason why we did not pick 1 day from Phase 1 is that, Phase 1 mostly includes days in early June, where TASK framework ran into some problems and its data became unavailable. Phase 2 includes the great majority of May dates, where both system were available. Thus, out of reliability concern, we decide to pick a day in Phase 2, May 20.

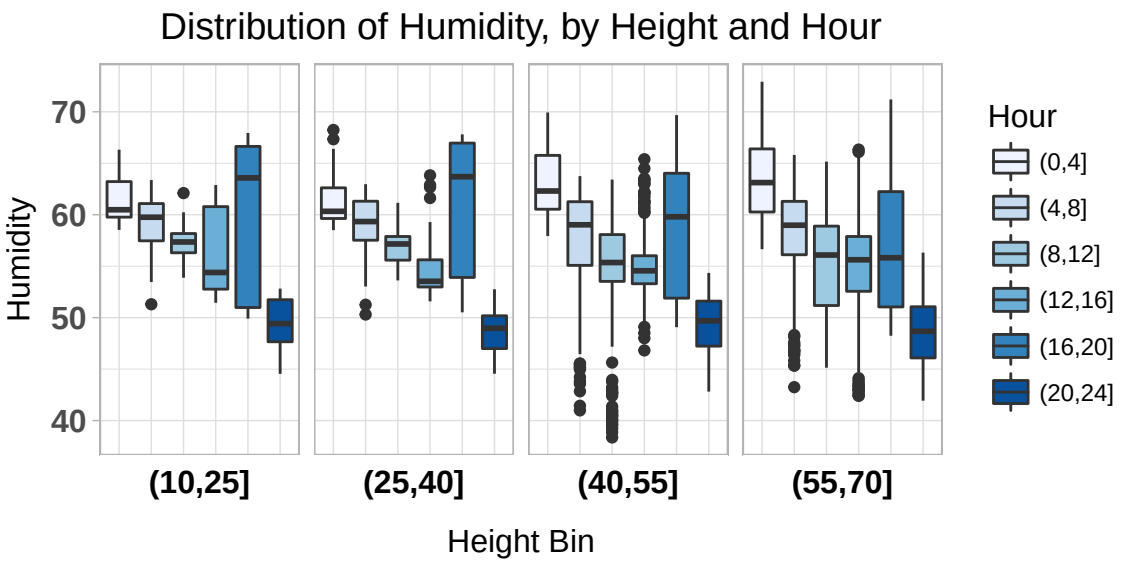
First, we want to see how the measurements progress over time. Since the microclimate varies by height, we color the plot by height, to distinguish measurement changes over varying heights. The plot below shows how temperature progresses. We can see a clear inverted-U shape: at late night, from 0 am to 5 am, the temperature is relatively low, around 10 degrees. Then the sun comes out, and the temperature gradually increases and reaches its peak at noon. In the afternoon, it starts to go down and end at around 15 degrees by midnight, which is warmer than the late night. Now we take height into consideration. As the color scale shows, the darker the color, the lower the height. We can see that the lower levels are generally colder than the higher levels, especially in the morning, when the sun first comes out.



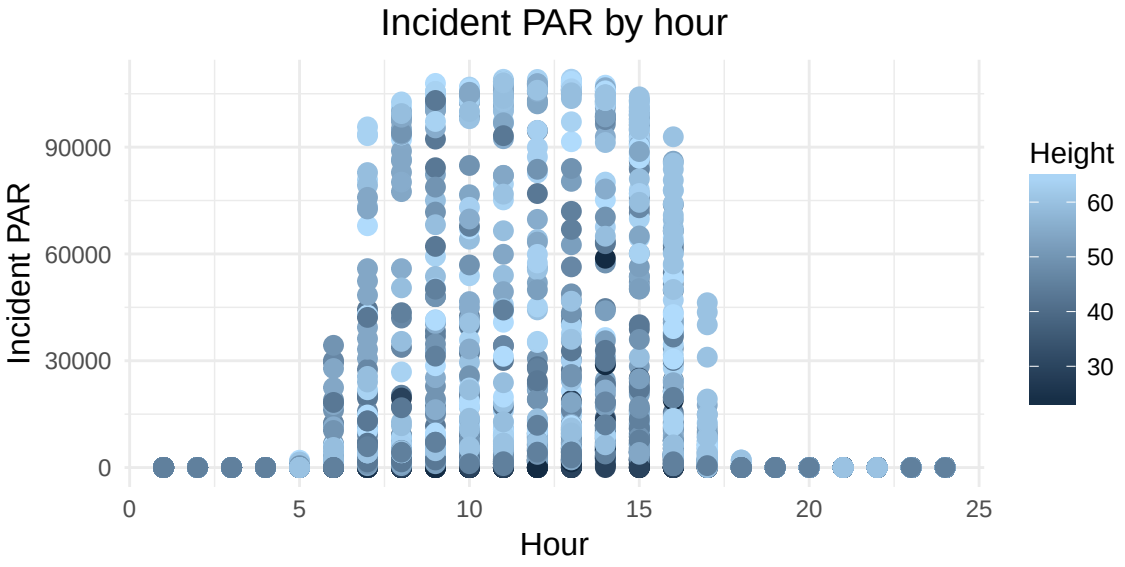
Next, we explore adjusted humidity. The plot belows shows how humidity progresses over time, colored by height. In late night, the humidity is at its hightest, than it gradually decreases as sun comes out, climbs up again in the late afternoon and drops again at night. The bottom levels appear to be the most humid in the morning, no other height-related pattern is discernable as the colors all mixed together.

We thus decide to do a box plot of adjusted humidity, group by height bin and colored by hour bin, in order to take a closer look at the role of height. As the box plot below shows, in all four height bins, the pattern largely follows the trend we discovered in the previous plot. One noticeable thing is that, the bottom level, 10 m ~ 25 m, seems to have a wider distribution in the afternoon and early night, which indicates the unpredictability of their humidity.





Lastly, we explore incident PAR. The plot shows a clear inverted-U shape pattern. At night, incident PAR = 0 because there is no direct sunlight. At 6 am, when the sun comes out, incident PAR increases, reaches to its peak at noon, and drops to 0 again at 6 pm, when the sun goes down. And as indicated by the plot, the higher values have lighter colors while the lower values are darker. It makes perfect sense because the higher levels are exposed to direct sunlight while it is harder for sunlight to reach the lower parts.



3 Graphical Critique

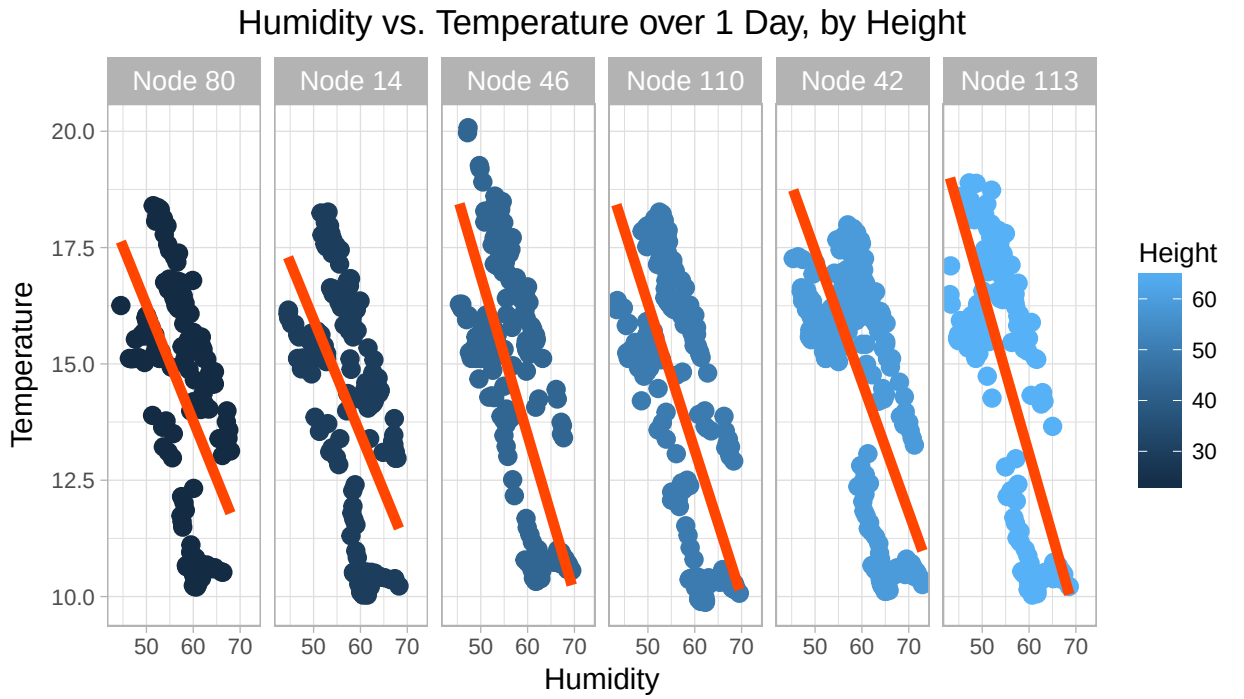
Figure 3a and 3b could have provided more information with additional elements, such as color. For example, it would be great if 3a is colored by height, in this way we can directly tell how sensor readings are distributed across heights, which provides valuable information in both checking data validity and learning from the data. The same logic applies to 3b. Instead of dissecting the readings on value, time/height * value dimensions, the researchers should make use of more plotting elements to make their plots more informative and intuitive.

A good plot should allow the readers understand the key message it wants to convey within several seconds. However, apparently Figure 4 fails to do it. Because of the vagueness of its caption and (more importantly) the lack of plot title, it is almost impossible to understand the plots without reading the text. It is especially true for the spatial location plots on the right. Also, so many lines are mingled together without any legend, or explanation. Moreover, 4d can be re-scaled to be more informative. For both left and right plots of 4d, the great majority of the space is blank while the great majority of the values squeezed together in one corner. We can re-scale the x/y axis and remove extreme outliers, to move the great majority of the values back into the middle, so that we can have a much clearer view of their relationship, and deeper understanding of the point the researchers try to convey.

4 Findings

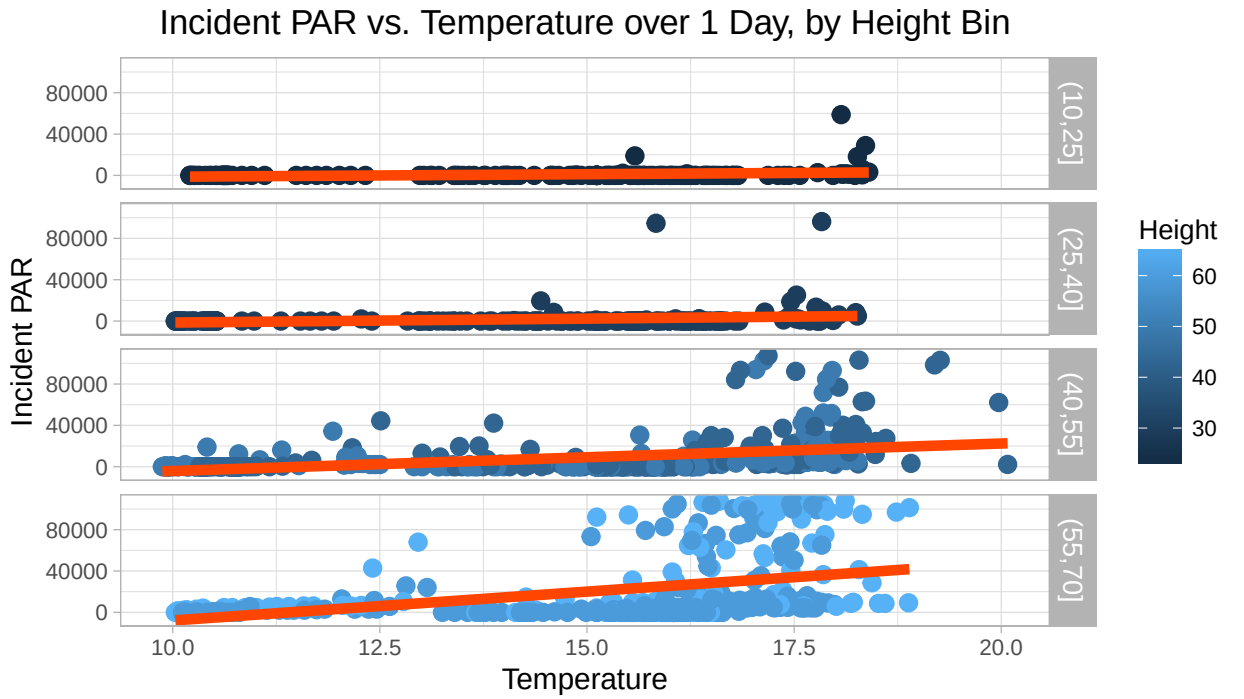
4.1 First finding

First, we explore the relation between humidity and temperature. Still using the May 20 Data as an illustration, we choose 6 nodes of varying heights (heights denoted by the darkness of color: the lighter, the higher) and plot their temperature vs. humidity, respectively. Trendlines are added to better present the correlation. As the plot shows, there exists **a negative correlation between Temperature and Humidity**. Moreover, the trendlines of the six nodes are almost in parallel, which indicates that the negative correlation exists, and has largely the same magnitude across different levels of height.



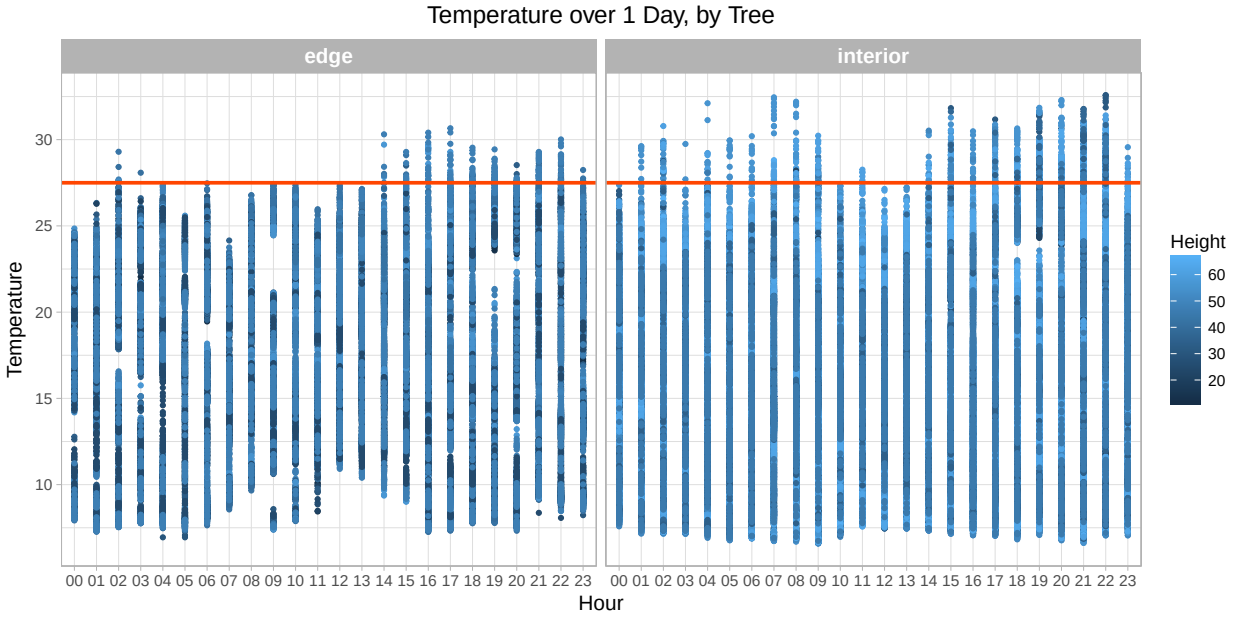
4.2 Second finding

Second, we explore the relation between incident PAR and temperature. Still using the May 20 Data as an illustration, we break heights into 4 equal intervals (heights denoted by the darkness of color: the lighter, the higher) and plot their incident PAR vs. temperature, respectively. Trendlines are added to better present the correlation. As the plot shows, the slope of the trendlines become increasingly negative as we move to higher levels (the lighter color). Incident PAR and temperature are basicly uncorrelated at lower levels, but at higher levels, **there exists a clearly negative relation between them**.



4.3 Third finding

Third, we explore the difference in temperature between edge and interior of the tree, over time. Using the all the data, for edge and interior, respectively, we plot how temperature progresses over time, colored by height. Horizontal lines at temperature = 27.5 degrees are added to better illustrate the point. As the plot shows, at almost all times, interior has much more points over the trendline than edge, indicating that **its temperature is higher than that of edge**. Moreover, in the morning, edge's above-trendline temperatures are generally from higher levels (as indicated by the lighter color) and in late afternoon/early evening, the high temperatures are from bottom levels. It seems like the interior's higher level does a good job at absorbing sunlight in the morning and its lower levels at preserving heat. On the contrast, edge's color is more mixed and we could not discern any meaningful patterns.



5 Discussion

Did the data size restric you in any way? On some level, yes. Large data size brings about computational constraints to the analysis process. A simple temperature vs. epoch scatterplot will take a while to finish, not to mention something fancier. The large data size slows the whole process in the beginning, as we need to start with manipulation on the whole data set. It only gets better when we shrink the dataset of interest down.

Moreover, due to the large data size, for almost all variables of interest, their distributions are very messy, with points all over the place. At the start, it may be very challenging for understanding and cleaning the data, because the patterns are not very discernable and it's hard to know which way to go. Reading the paper and every bit of documentation about the dataset have become very important in this case.

6 Conclusion

This report presents a complete pre-processing and exploratory analysis of a relatively raw and large dataset. Due to time and space constraints, there is still a lot about the dataset worth exploring, i.e., correlation matrix of all measurements, trend analysis of a longer period, etc.