STAT215A - Redwood Data Lab

Timothy Meyers

September 21, 2014

Exploration of Data

Please read the paper to understand how the sensor works, and write a paragraph to discuss the measurement of each variable you nd interesting in the data. Please have at least 3 variables in your report, and those variables should be related to your ndings in 1.3.

1) Variables

The data is measured from a series of sensors which collect temperature, humidity, and luminosity readings. The Sensirion SHT11 is a digital sensor that records both temperature and humidity. The measurement error for each of these variables is 0.5C and 3.5 percent, respectively. Humidity is greater than zero, but can be greater than 100 percent in foggy conditions. Temperature readings, even if they are within normal ranges, can deviate significantly from historical values depending on the sensor's battery. Voltage is recorded and provides an indication of the reliability of temperature recordings, and it's advised to focus only on temperature readings generated from a sensor with voltage between 2.4 and 3. Luminosity is recorded as Photosynthetically active radiation (PAR), and both incident (direct) and reflected (ambient) PAR are recorded. These measurements are collected by two Mamamatsu S1087 photodiodes, and are referred to as hamatop and hamabot in the data. The units of measurement are unclear.

2) Data Cleaning

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 don't blindly follow their method. Record in your report the steps you take and any evidence you use to support them.

After loading the data and making a few plots, I saw that some rows had values missing. Also, looking at the plots helped me understand the ranges where values were expected to lie, allowing me to remove obvious outliers.

1.
2.
3.
4.
5

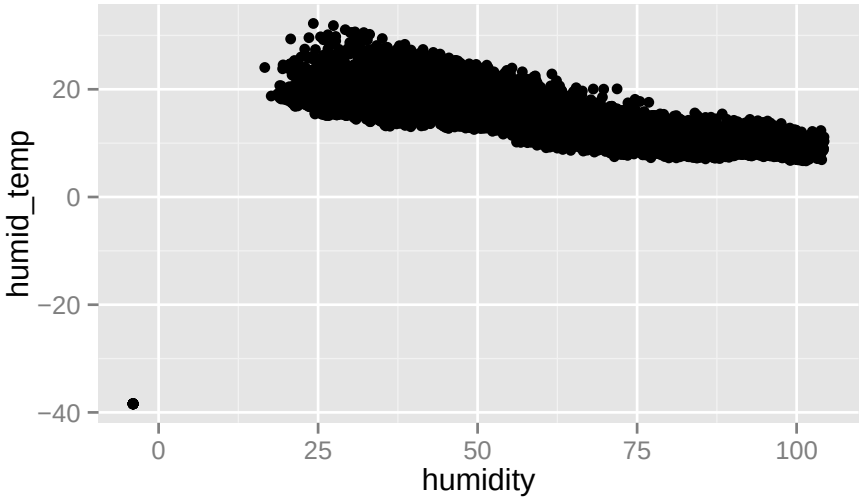

Figure 1: Observe expected range for humidity and humidtemp

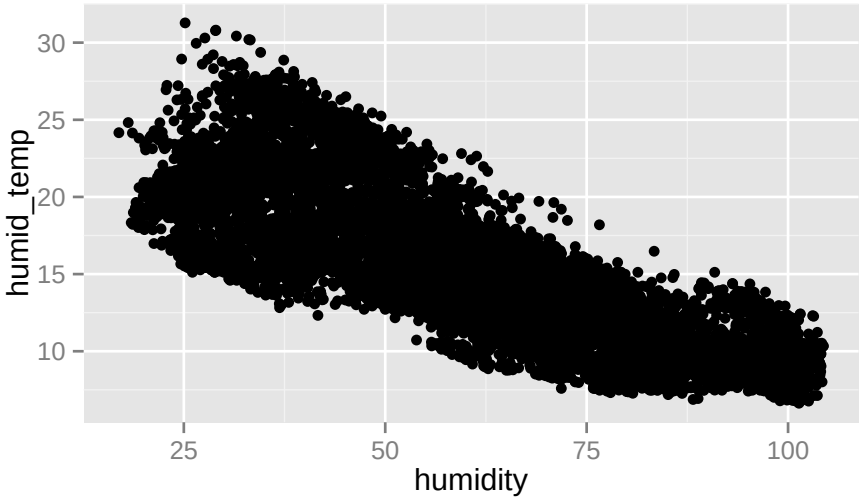

Figure 2: Remove empty values and outliers (humidtemp and humidity)

6

After applying basic data cleansing to the net and log data, I ran humidity and temperature versus epoch on the net data and found that there were quite a few outliers in the temperature measurements. These became more appar-

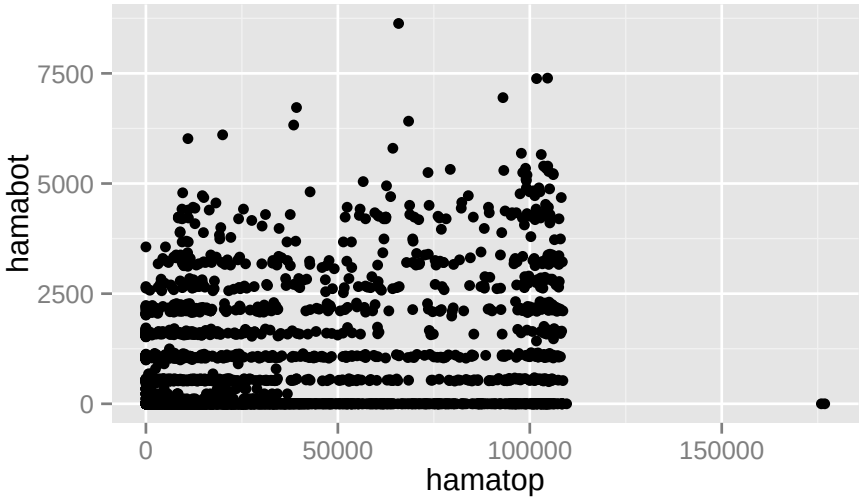

Figure 3: Observe expected range for hamatop and hamabot

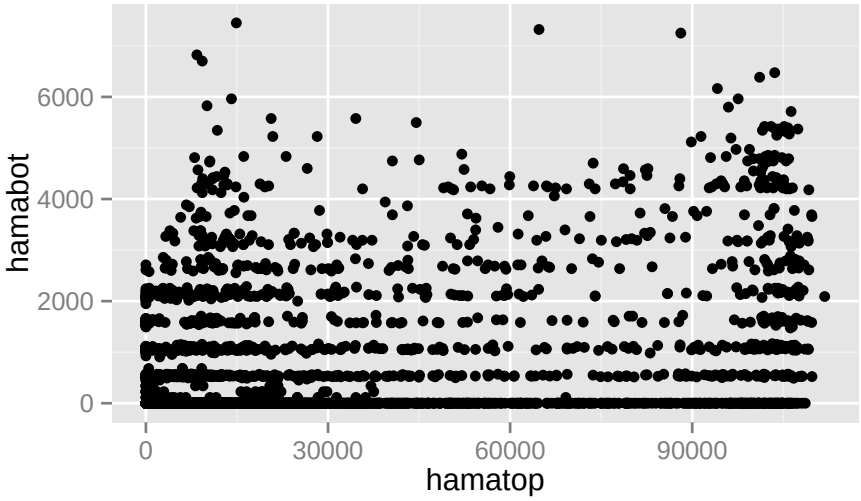

Figure 4: Remove empty values and outliers (Hamatop and Hamabot)

ent when plotting against voltage, and I decided that 240 was the appropriate voltage cutoff to remove these outliers.

7

8

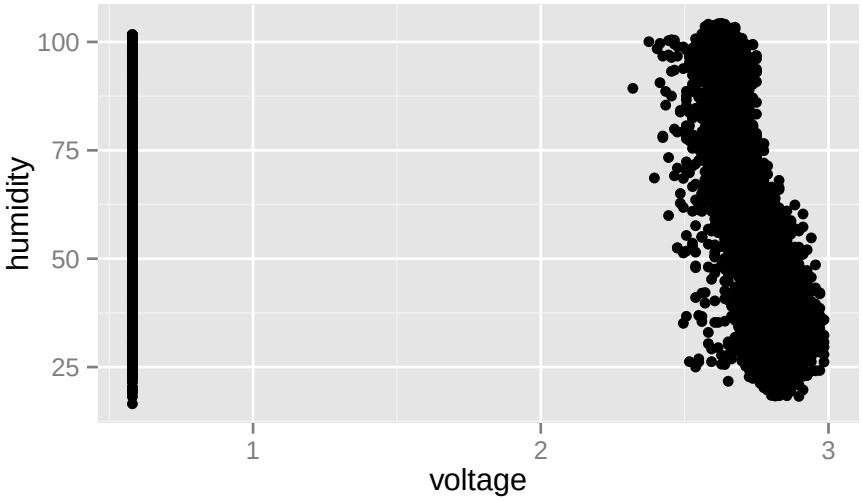

Figure 5: Observe expected range for voltage and humidity

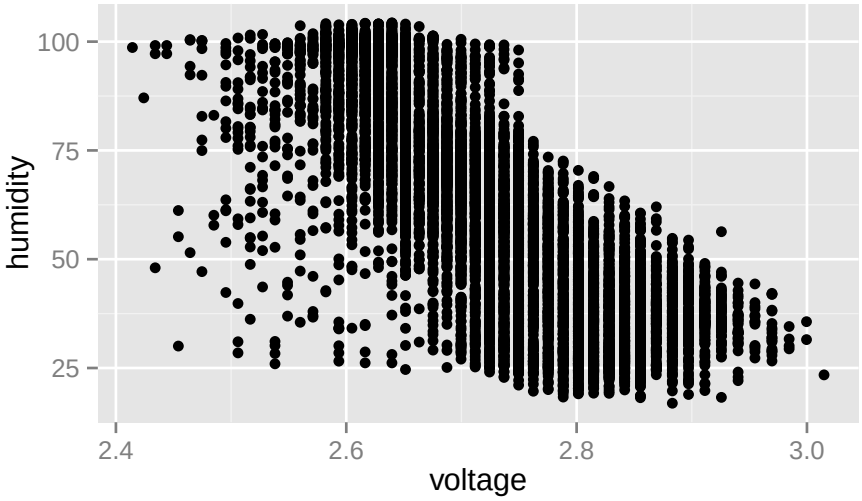

Figure 6: Remove empty values and outliers (voltage)

9
10
11
12

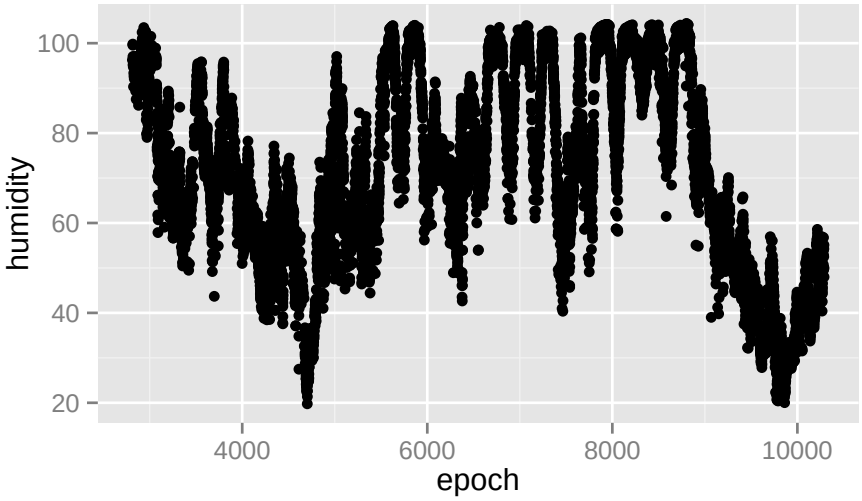

Figure 7: Observe humidity over time

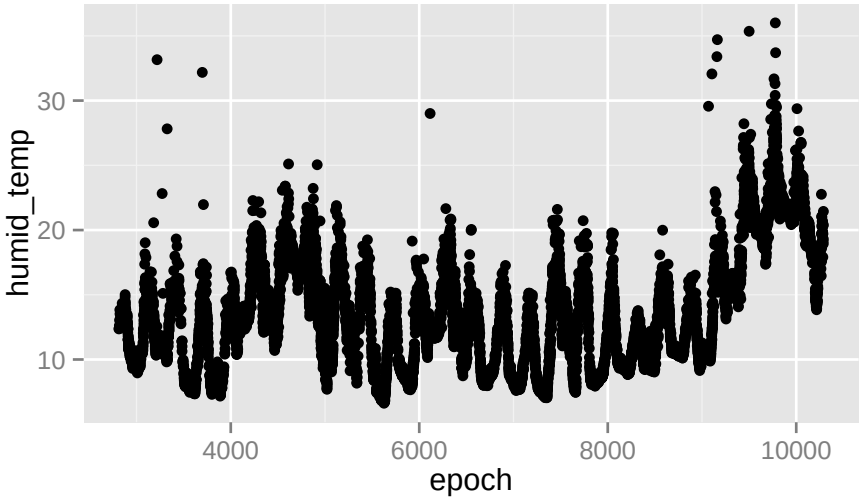

Figure 8: Observe temperature over time

With sensible values, we can now anaylze whether these values make sense over time. Several plots can reveal pattern outliers.

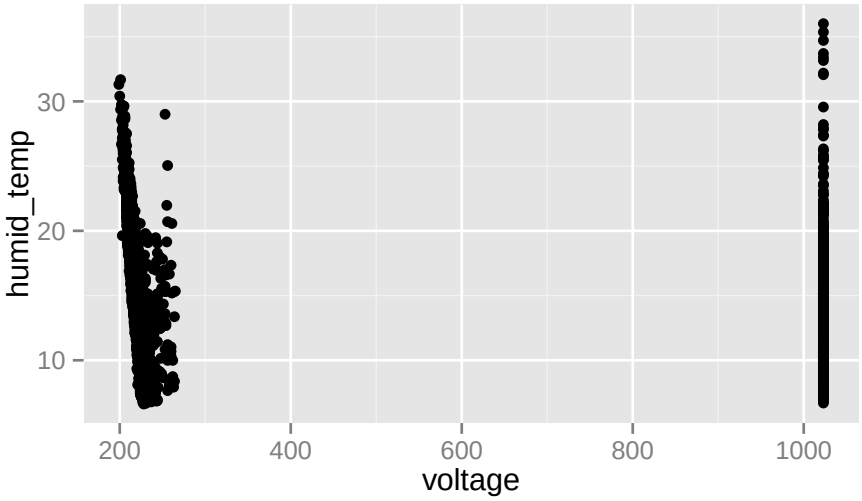

Figure 9: Observe temperature over voltage

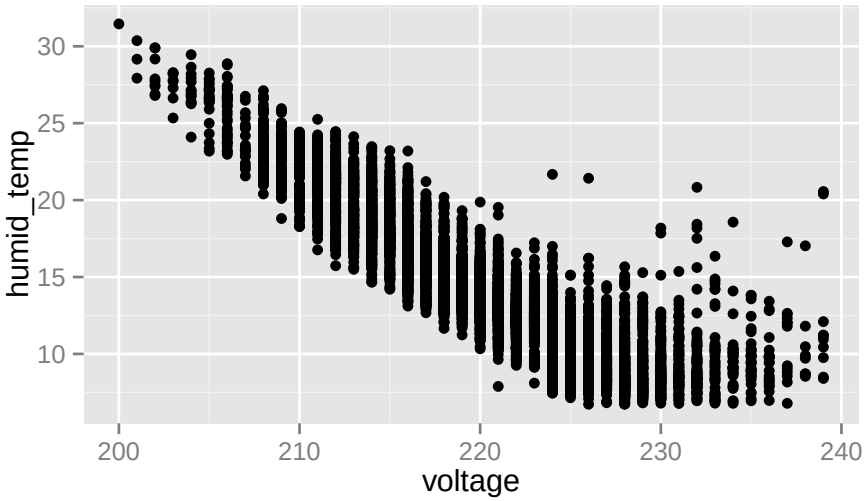

Figure 10: Remove voltage outliers

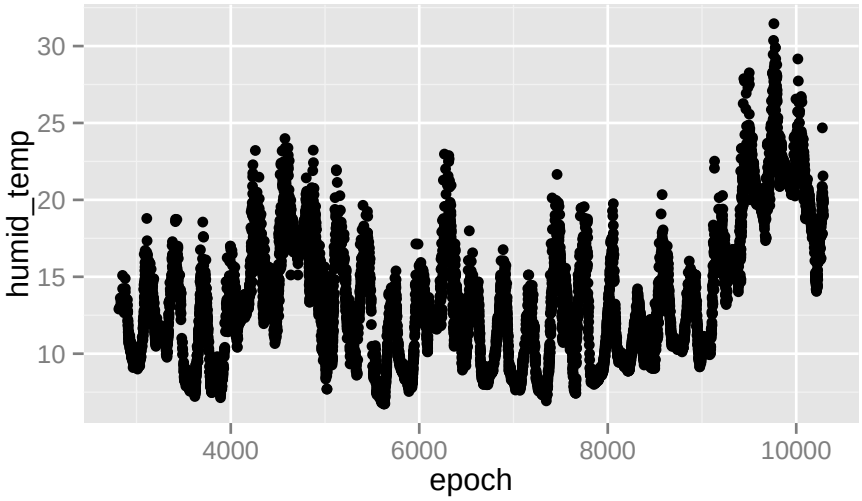

Figure 11: Cleansed temperature over time

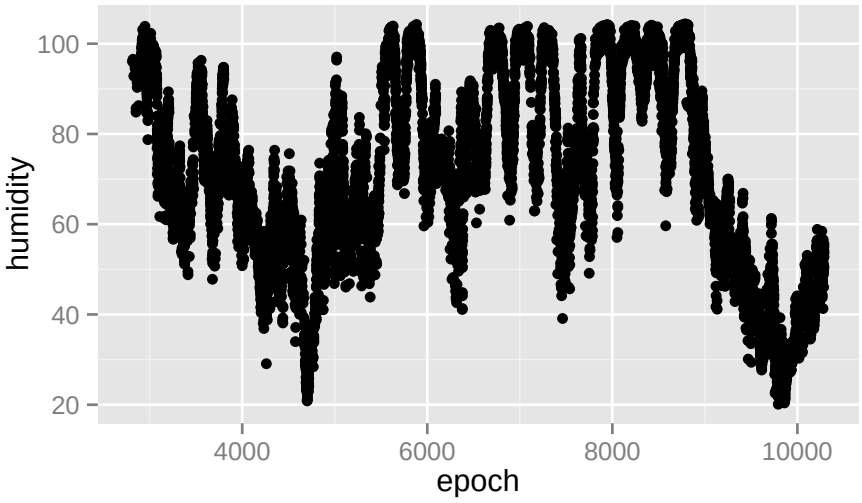

Figure 12: Cleansed humidity over time