Project 1 Redwood Data Report

William Tirone (2774025) and Natalie Smith (2819547)

Paper here: https://sakai.duke.edu/access/content/group/cca7dc7c-9906-43db-af3b-91d09e2e9748/proj1/tolle2005.pdf

```
library(tidyverse)
```

```
-- Attaching packages --------------------------------------- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

"The file sonoma-data-all.csv is a simple concatenation of the two files sonoma-datalog.csv and sonoma-data-net.csv"

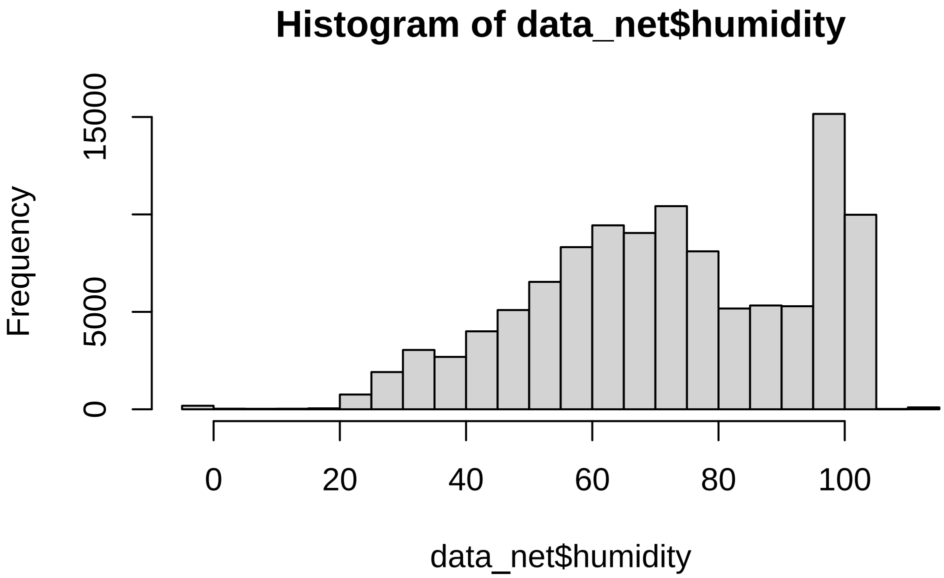
```
#loading in data
data_net = read.csv('data/sonoma-data-net.csv')
data_log = read.csv('data/sonoma-data-log.csv')
data_all = read.csv('data/sonoma-data-all.csv')
```

```
names(data_all)
```

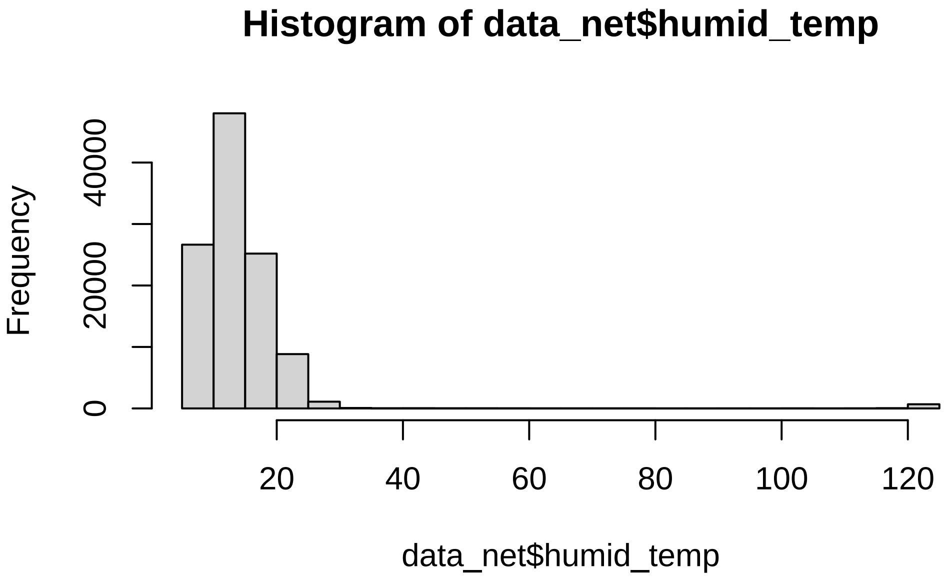
```
 [1] "result_time" "epoch"       "nodeid"      "parent"      "voltage"
 [6] "depth"       "humidity"    "humid_temp"  "humid_adj"   "hamatop"
[11] "hamabot"
```

2 Data Cleaning

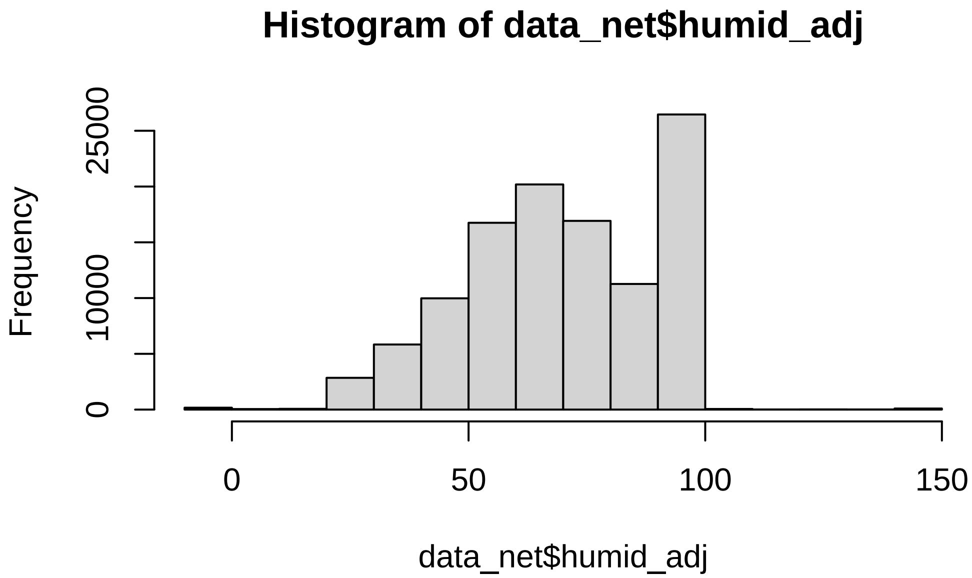
```
#data_net file
hist(data_net$humidity)
```



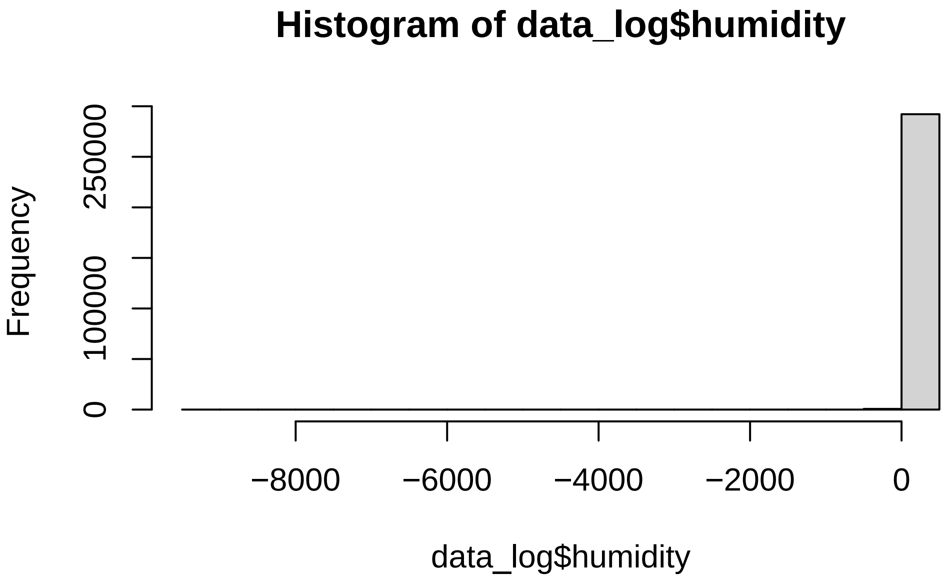
```
hist(data_net$humid_temp)
```



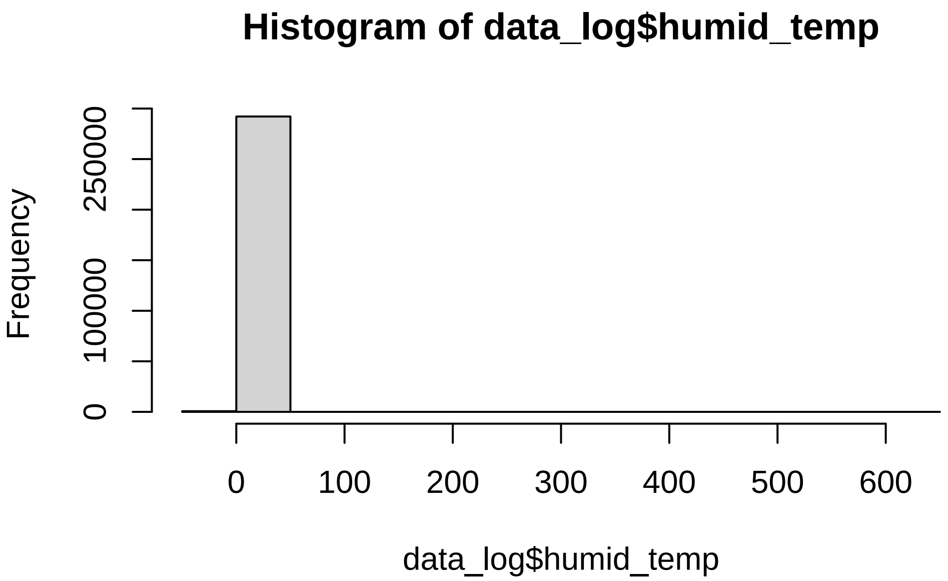
```
hist(data_net$humid_adj)
```



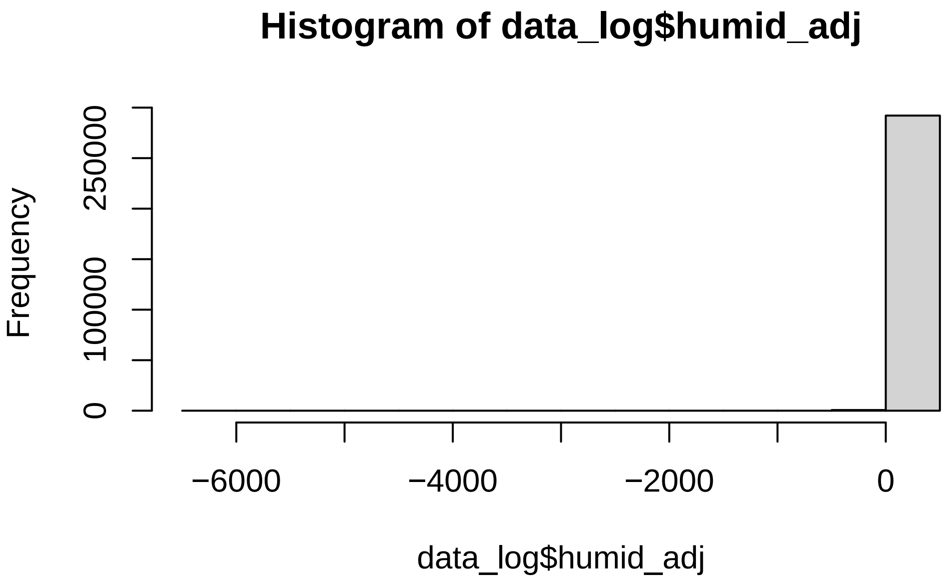
```
#data_log
hist(data_log$humidity)
```



```
hist(data_log$humid_temp)
```



```
hist(data_log$humid_adj)
```



(b) Remove missing data. Comment on the number of missing measurements and the corresponding date and time period.

(c) We left joined the data on the noideid / ID column to add 4 new variables to our total data: Height, Direc, Dist, and Tree. However, comparing the number of unique ID's in the location data to the total data, we have 73 unique IDs in the total data with 80 in the location data. This represents the sensors that broke or did not transmit data during the study.

```r
location_data = read.table('data/mote-location-data.txt',header=TRUE)
data_all = dplyr::left_join(data_all,location_data,by=c('nodeid' = 'ID'))
dim(data_all)[2]
```

```
[1] 15
```

d)

(e) (Bonus) Discuss other possible outliers and explain your reason why it is better to remove them than to keep them.

3 Data Exploration

(a) Make some pairwise scatterplots of some variables. Pick a reasonable time period. Explain your choice and describe your findings.

(b) Are any of the predictors associated with Incident PAR? If so, explain the relationship.

(c) Each variable of our data basically have three dimensions: value, height and time. Consider each variable as a time series and look at its temporal trend. Generate such plots (value vs time) with height as color cue for at least four variables (Temperature, Relative Humidity, Incident PAR and Reflected PAR). You can do it for different time scales (during an hour, during a day or during the entire experiment). However, at least the plots with days as x-axis are required. Comment on the range, continuity and strange behaviors in these variables.

(d) After PCA analysis, generate scree plot of the data. Can this data be approximated by some low-dimensional representation?

4 Interesting Findings

Describe two/three interesting findings from exploratory analysis of the data. Try to use the techniques that you have learned, such as histograms, PCA, K-means, GMM and hierachical clustering etc. Note that even though you got a dataframe with only a few columns, you may reshape the dataframe before doing any EDA, such as reorganizing such that aggregated information in each day is a column, or a particular hour in each day is a column. Comment on your interesting findings. Different bonuses are given based on how interesing your result is.

(a) Finding 1.

(b) Finding 2.

(c) (Bonus) Finding 3. Bonus is given only if we find all three findings interesting.

5 Graph Critique

The overall quality of the paper by Tolle et al. is good. However, some plots are not perfect from a statistician's point of view.

(a) Figure 3[a] shows the distributions of sensor readings projected onto the value dimension, using a histogram. It turns out that both the incident and reflected PAR have long tail. We could not read full information from this histogram. Try to make a better plot with log transform of the data.

(b) What message do the boxplots in Figure 3[c] and 3[d] try to convey? Do you think the plots convey the right messages? If not generate a new plot with the same data. Hint: compare to some plots in Figure 4.

(c) Any suggestions for improving the first two plots in Figure 4? Can you distinguish all the colors in these two plots?

(d) Comment on Figure 7. Is it possible to generate a better visualization to highlight the difference between network and log data?