Project 1

Sanskriti Purohit & Caleb Woo

2022-10-13

```
knitr::opts_chunk$set(message = F, warning = F)
```

```
library(tidyverse)
data_net <- read_csv("data/sonoma-data-net.csv")
data_log <- read_csv("data/sonoma-data-log.csv")
```

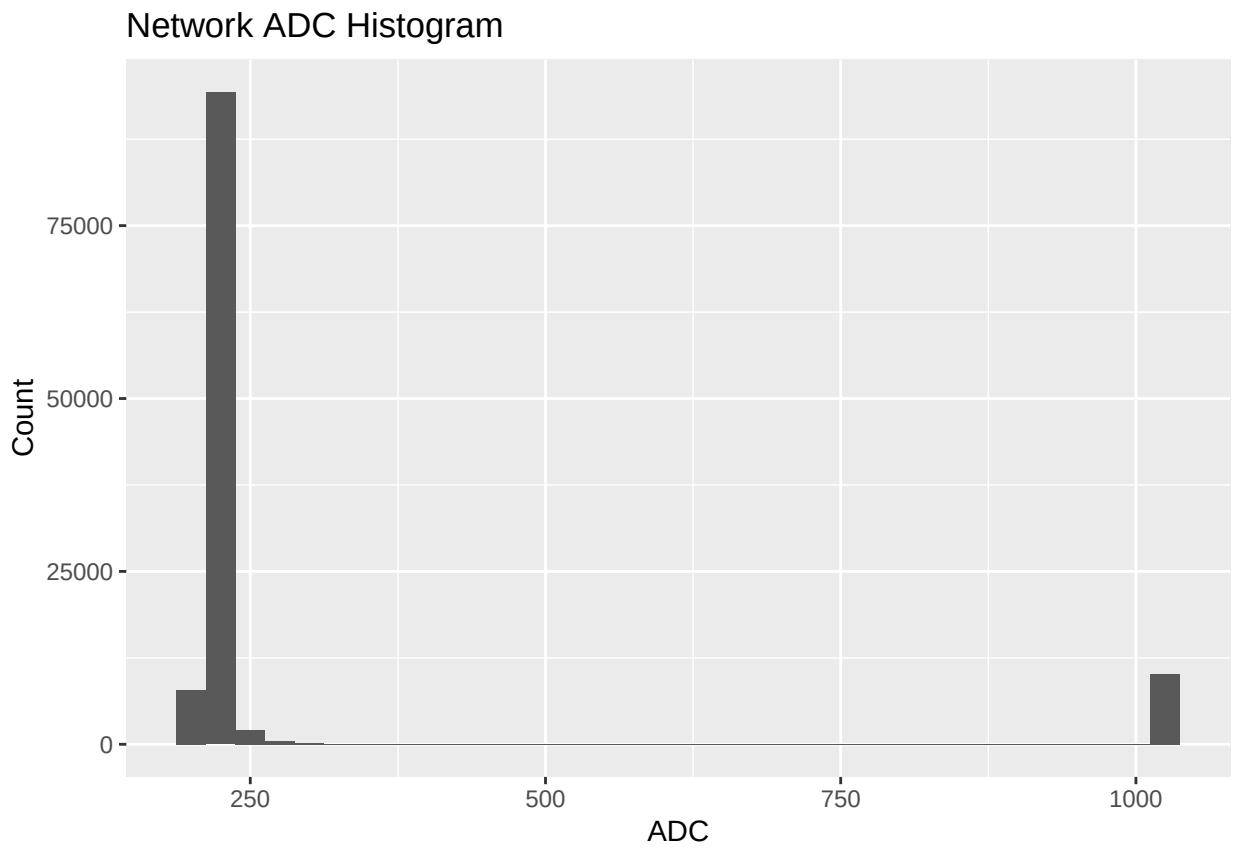
```
dates <- read.table("data/sonoma-dates", sep = " ")
dates <- data.frame(t(dates)[-2,], row.names = NULL)
names(dates) <- dates[1,]
dates <- dates[-1,]
dates <- dates[-13001,]
dates[1,] <- c("1", "Tue Apr 27 17:10:00 2004", "12536.0069444444")
row.names(dates) <- NULL

dates <- dates %>%
  mutate(epochNums = as.integer(epochNums),
         epochDates = as.POSIXct(epochDates, format = "%a %b %d %H:%M:%S %Y"),
         epochDays = as.numeric(epochDays))
names(dates)[1] <- "epoch"
```

2.

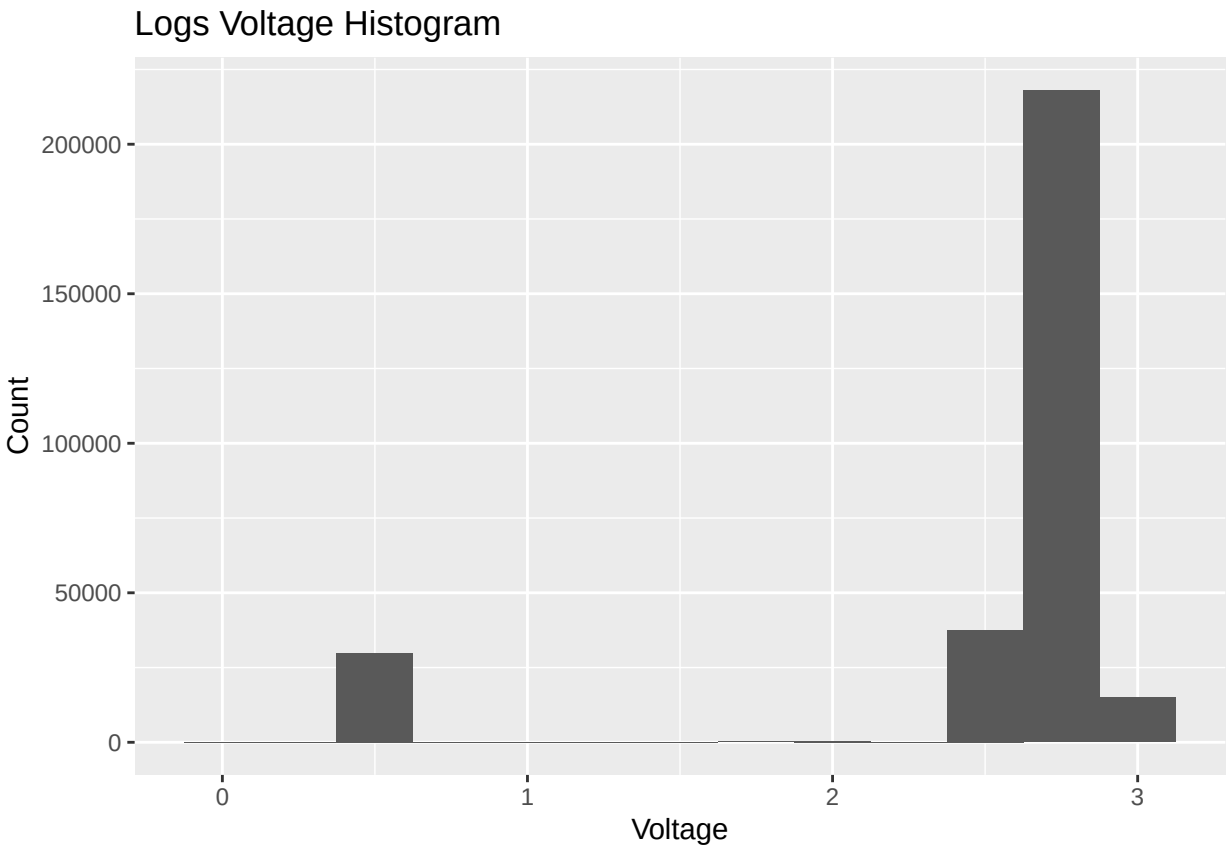
a)

Voltage

```
data_net %>%
  ggplot(data=., aes(x=voltage)) +
  geom_histogram(binwidth=25) +
  labs(title="Network ADC Histogram", x="ADC", y="Count")
```



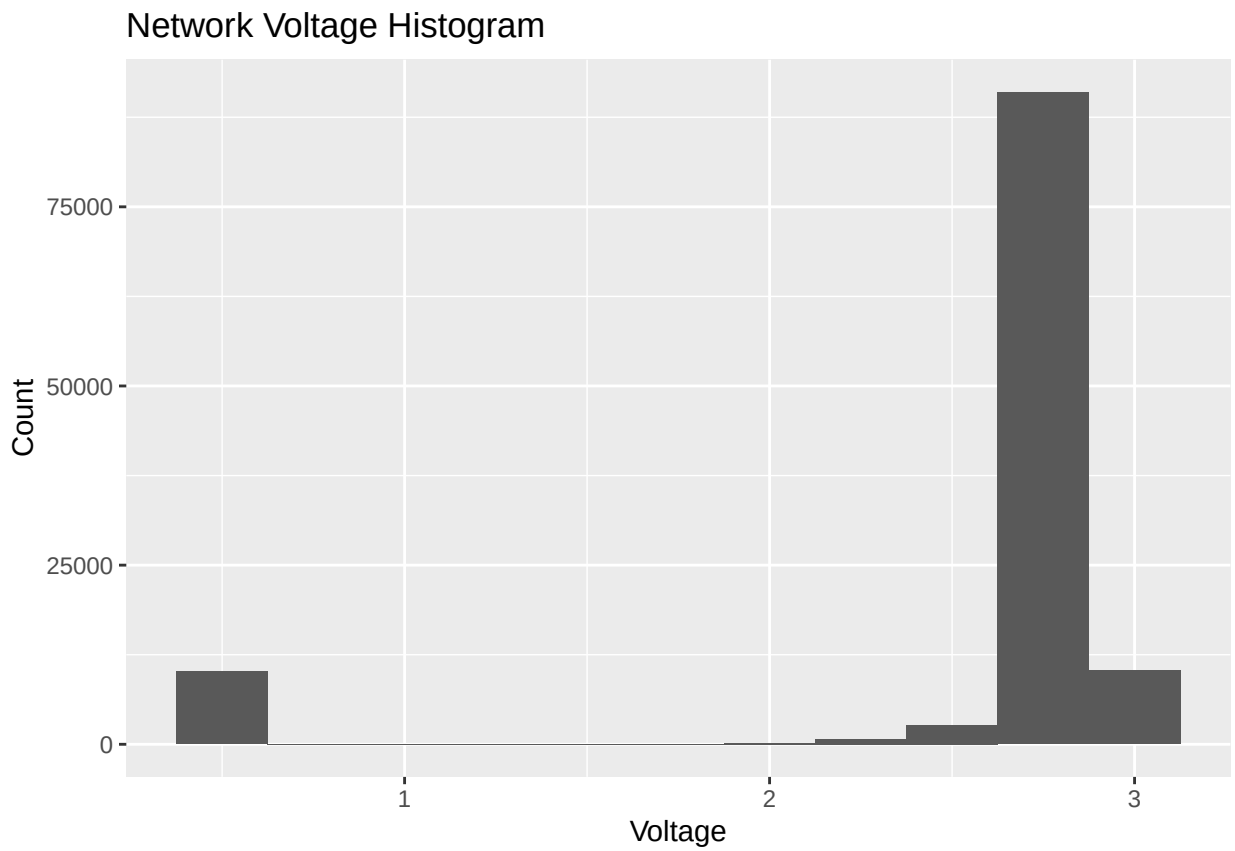

```
data_log %>%
  ggplot(data=., aes(x=voltage)) +
  geom_histogram(binwidth=0.25) +
  labs(title="Logs Voltage Histogram", x="Voltage", y="Count")
```



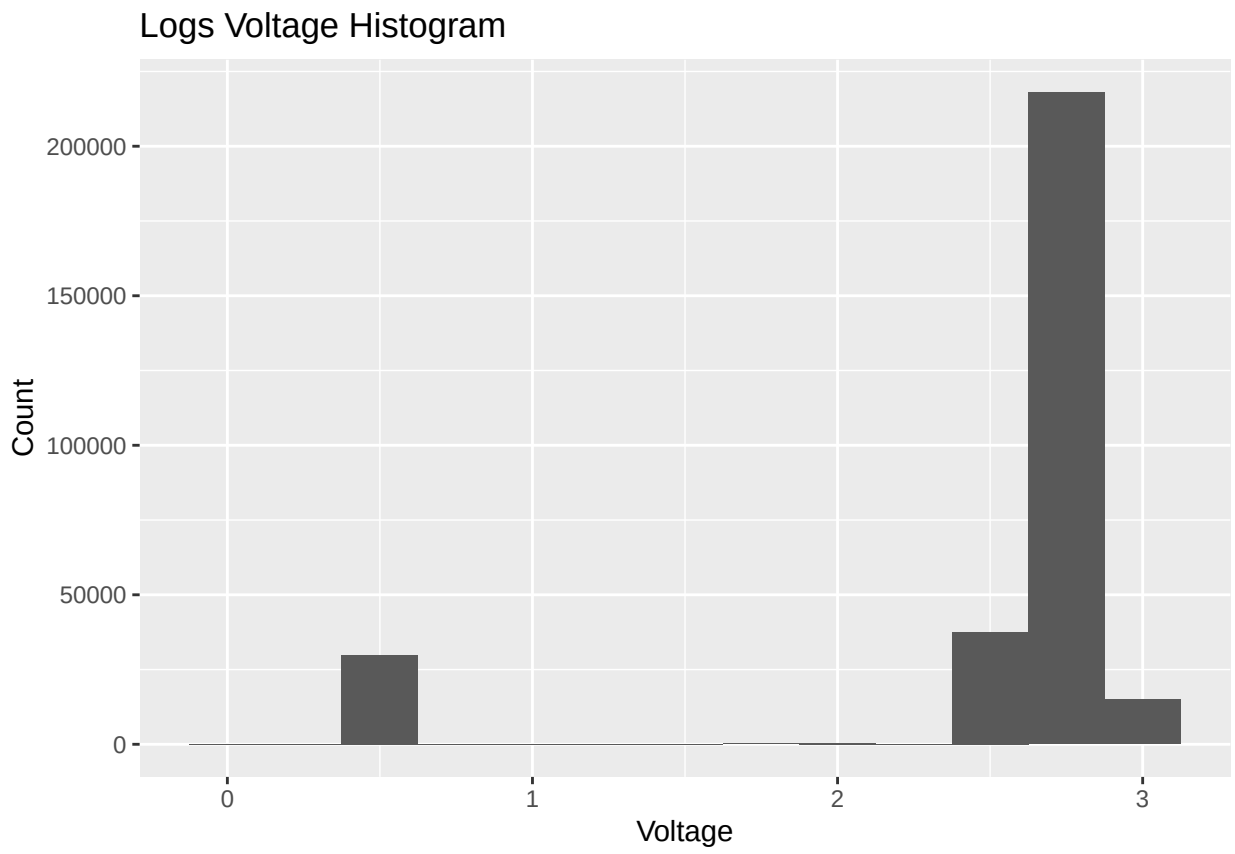
Voltage is inconsistent, link below for ADC to voltage conversion

- http://www-db.ics.uci.edu/pages/research/quasar/MPR-MIB%20Series%20User%20Manual%207430-0021-06_A.pdf
- page 23 (25 of pdf) has MICA2DOT conversion

```
data_net <- data_net %>%
  mutate(voltage = 0.6*(1024/voltage))

data_net %>%
  ggplot(data=., aes(x=voltage)) +
  geom_histogram(binwidth=0.25) +
  labs(title="Network Voltage Histogram", x="Voltage", y="Count")
```




```
data_log %>%
  ggplot(data=., aes(x=voltage)) +
  geom_histogram(binwidth=0.25) +
  labs(title="Logs Voltage Histogram", x="Voltage", y="Count")
```



Humidity

```
data_net %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4) %>%
  select(humidity) %>%
  pull() %>%
  min()
```

```
## [1] 19.5147
```

```
data_log %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4) %>%
  select(humidity) %>%
  pull() %>%
  min()
```

```
## [1] 16.2653
```

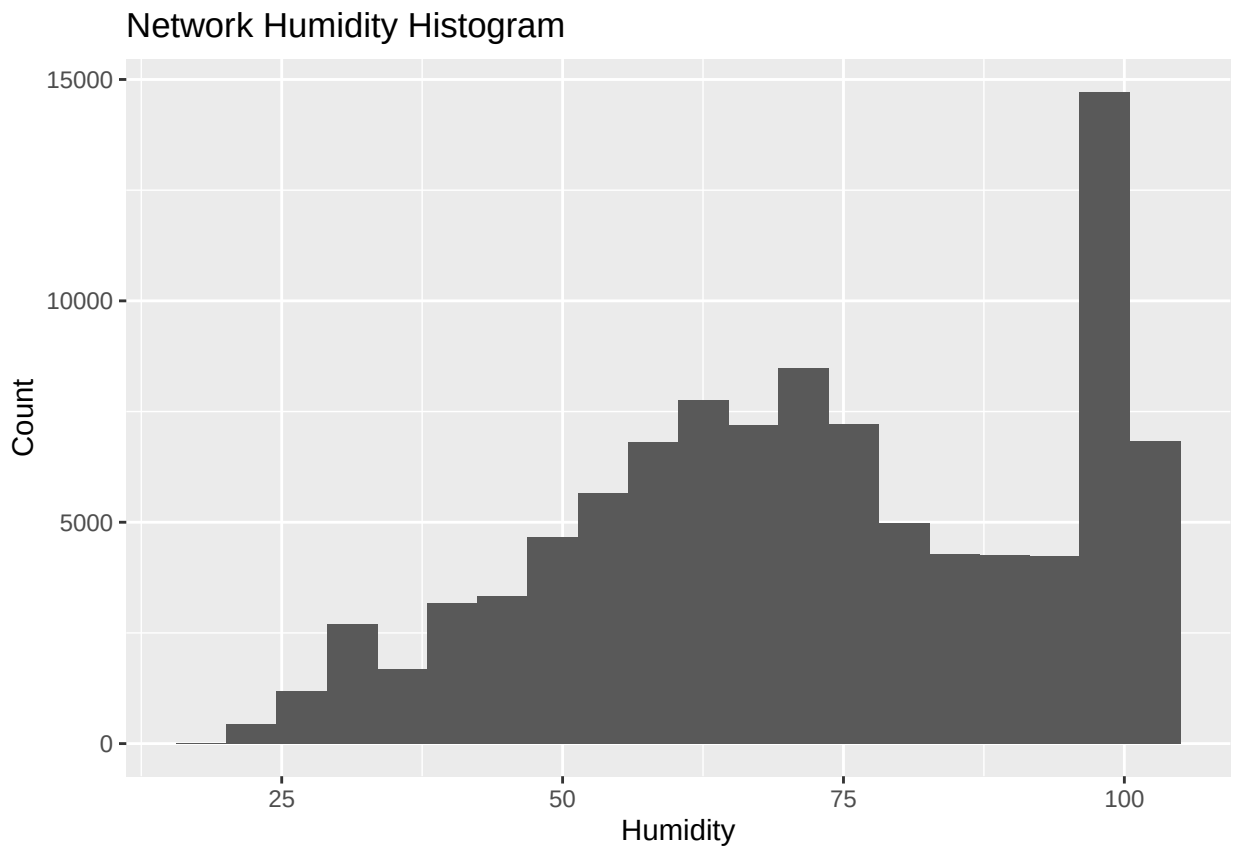
```
data_net %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4 & humidity < 105) %>%
  select(humidity) %>%
  pull() %>%
  max()
```

```
## [1] 104.385
```

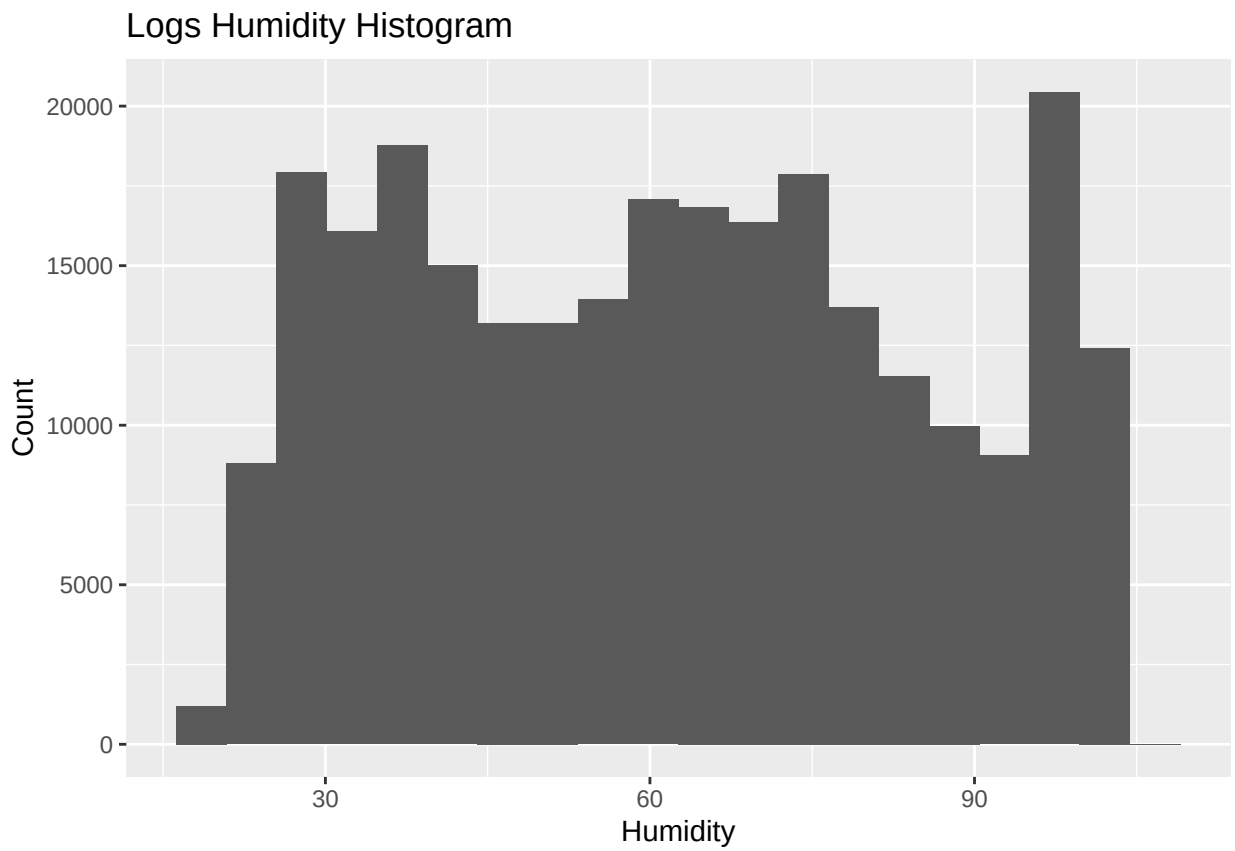
```
data_log %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4) %>%
  select(humidity) %>%
  pull() %>%
  max()
```

```
## [1] 104.405
```

```
data_net %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4 & humidity < 105) %>%
  ggplot(data=., aes(x=humidity)) +
  geom_histogram(bins=20) +
  labs(title="Network Humidity Histogram", x="Humidity", y="Count")
```




```
data_log %>%
  filter(!is.na(humidity) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=humidity)) +
  geom_histogram(bins=20) +
  labs(title="Logs Humidity Histogram", x="Humidity", y="Count")
```



Adjusted Humidity

```
data_net %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_adj) %>%
  pull() %>%
  min()
```

```
## [1] 19.3107
```

```
data_log %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_adj) %>%
  pull() %>%
  min()
```

```
## [1] 16.2282
```

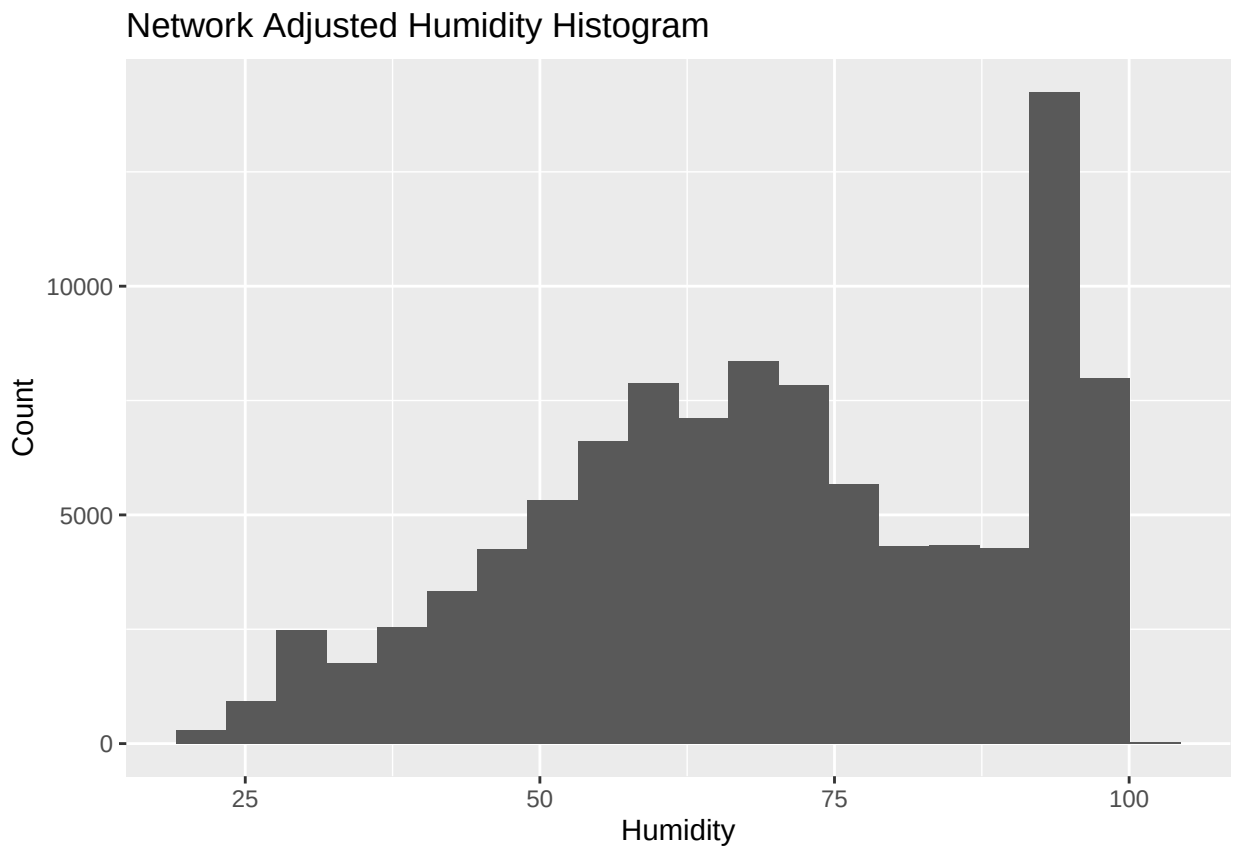
```
data_net %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4 & humid_adj < 105) %>%
  select(humid_adj) %>%
  pull() %>%
  max()
```

```
## [1] 100.223
```

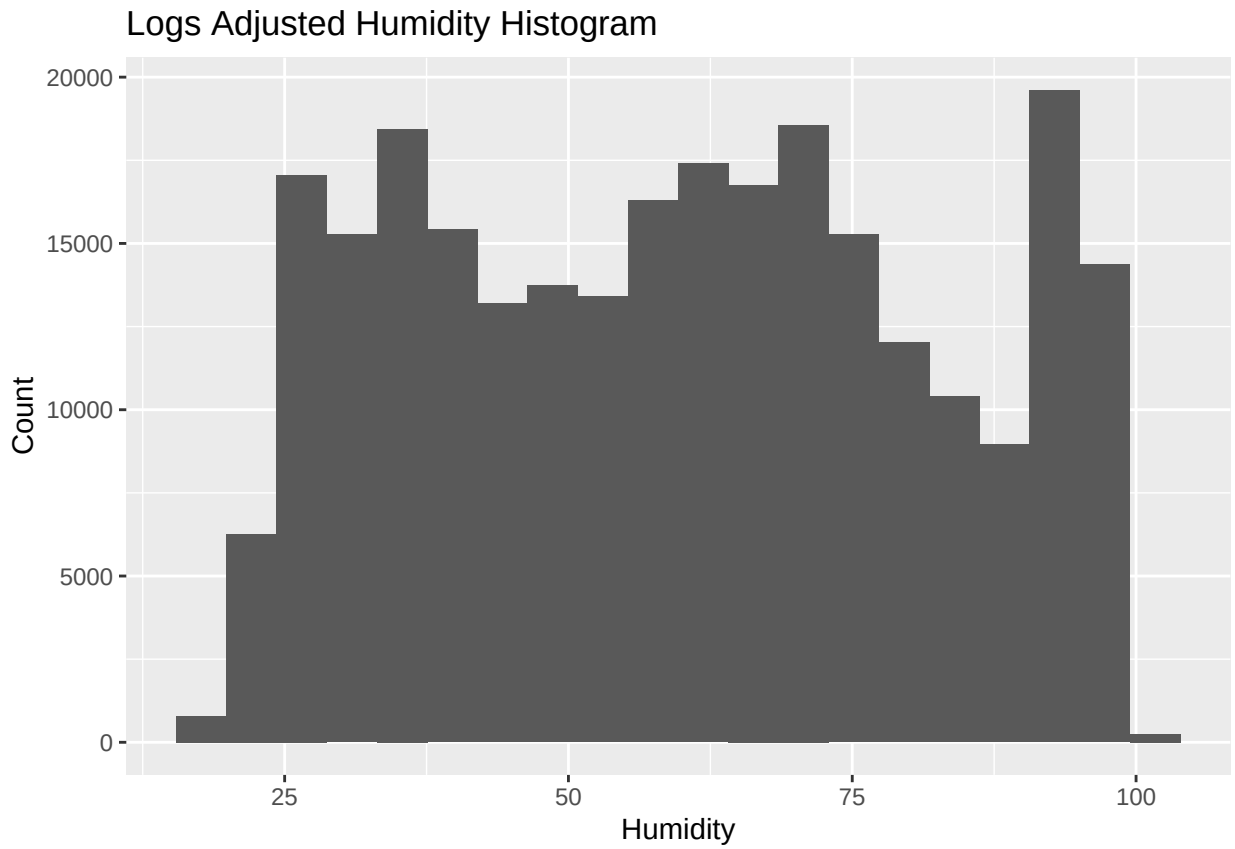
```
data_log %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_adj) %>%
  pull() %>%
  max()
```

```
## [1] 100.223
```

```
data_net %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4 & humid_adj < 105) %>%
  ggplot(data=., aes(x=humid_adj)) +
  geom_histogram(bins=20) +
  labs(title="Network Adjusted Humidity Histogram", x="Humidity", y="Count")
```




```
data_log %>%
  filter(!is.na(humid_adj) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=humid_adj)) +
  geom_histogram(bins=20) +
  labs(title="Logs Adjusted Humidity Histogram", x="Humidity", y="Count")
```



Temperature

```
data_net %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_temp) %>%
  pull() %>%
  min()
```

```
## [1] 6.582
```

```
data_log %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_temp) %>%
  pull() %>%
  min()
```

```
## [1] 6.582
```

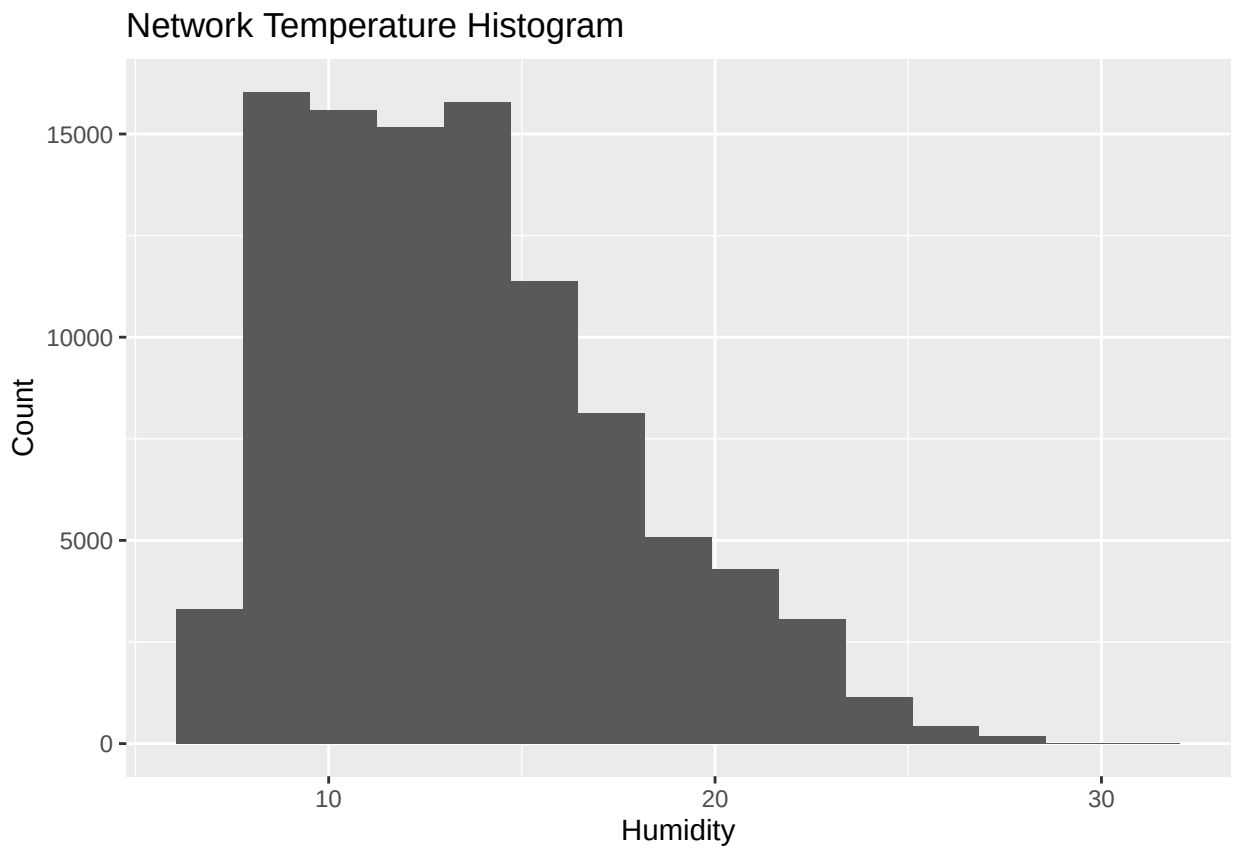
```
data_net %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4 & humid_temp < 35) %>%
  select(humid_temp) %>%
  pull() %>%
  max()
```

```
## [1] 30.8272
```

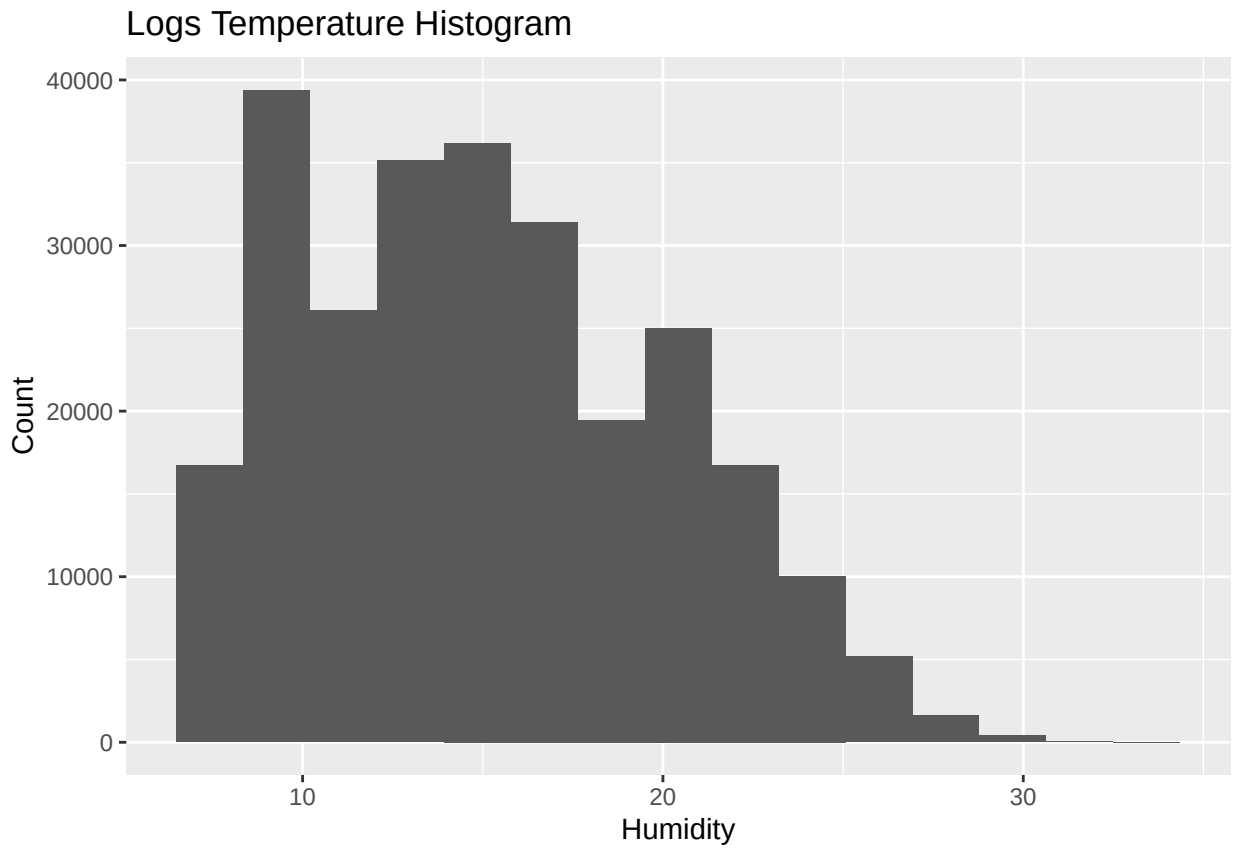
```
data_log %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4) %>%
  select(humid_temp) %>%
  pull() %>%
  max()
```

```
## [1] 32.5814
```

```
data_net %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4 & humid_temp < 35) %>%
  ggplot(data=., aes(x=humid_temp)) +
  geom_histogram(bins=15) +
  labs(title="Network Temperature Histogram", x="Humidity", y="Count")
```




```
data_log %>%
  filter(!is.na(humid_temp) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=humid_temp)) +
  geom_histogram(bins=15) +
  labs(title="Logs Temperature Histogram", x="Humidity", y="Count")
```



Incident PAR

```
data_net %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamatop) %>%
  pull() %>%
  min()
```

```
## [1] 0
```

```
data_log %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamatop) %>%
  pull() %>%
  min()
```

```
## [1] 0
```

```
data_net %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamatop) %>%
  pull() %>%
  max()
```

```
## [1] 113376
```

```
data_log %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamatop) %>%
  pull() %>%
  max()
```

```
## [1] 180255
```

Hamatop is inconsistent, link below for Lux to PPFD conversion

- https://www.apogeeinstruments.com/conversion-ppfd-to-lux/

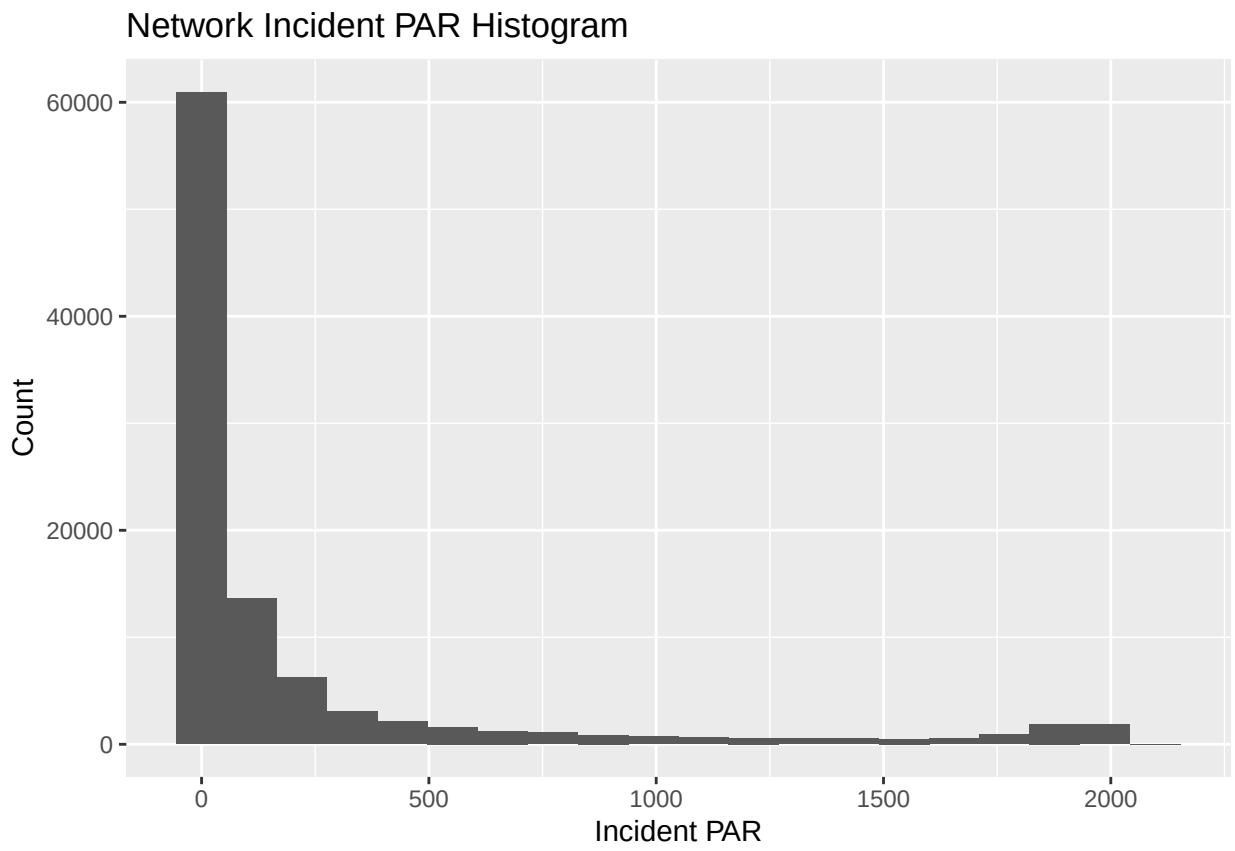
```
data_net <- data_net %>%
  mutate(hamatop=0.0185*hamatop)

data_log <- data_log %>%
  mutate(hamatop=0.0185*hamatop)

data_net %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamatop) %>%
  pull() %>%
  max()
```

```
## [1] 2097.456
```

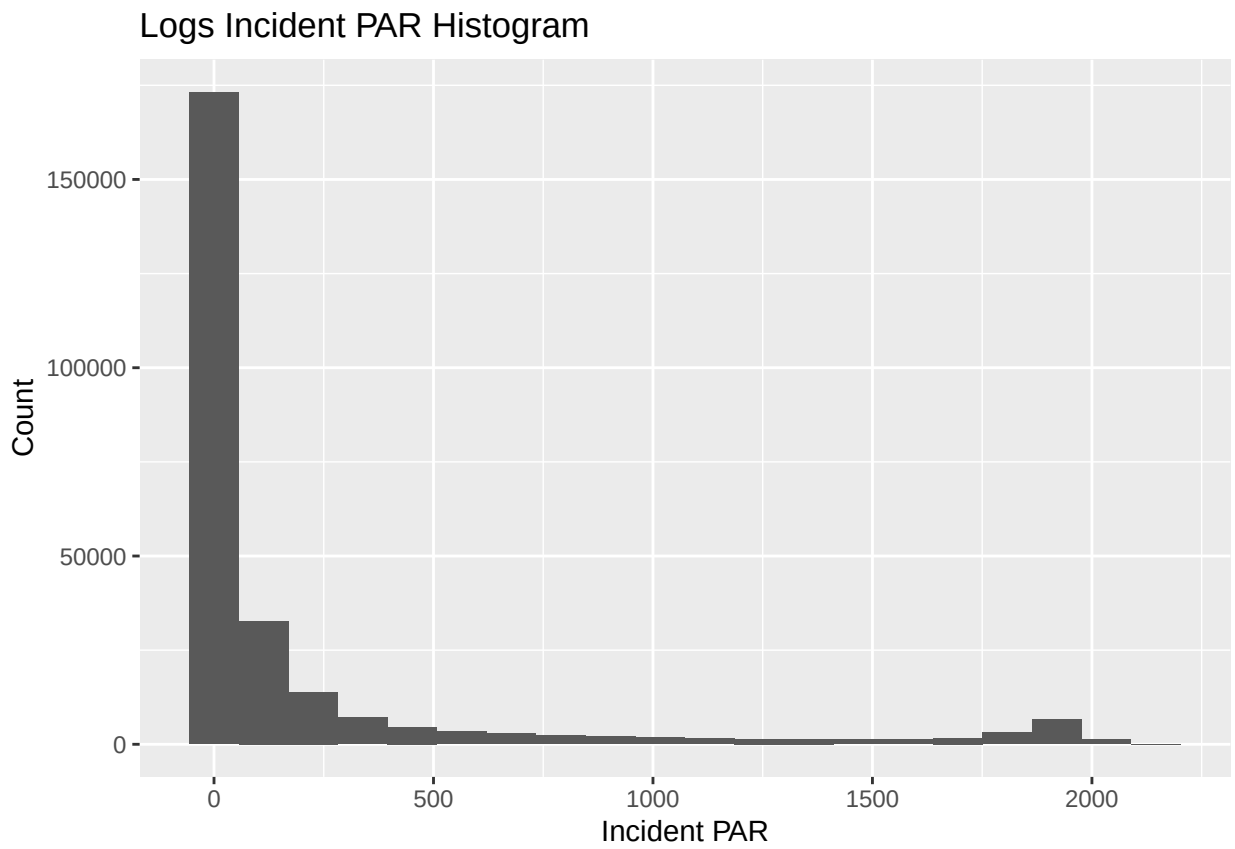
```
data_log %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4 & hamatop < 2200) %>%
  select(hamatop) %>%
  pull() %>%
  max()
```

```
## [1] 2146
```

```
data_net %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=hamatop)) +
  geom_histogram(bins=20) +
  labs(title="Network Incident PAR Histogram", x="Incident PAR", y="Count")
```




```
data_log %>%
  filter(!is.na(hamatop) & voltage <= 3 & voltage >= 2.4  & hamatop < 2200) %>%
  ggplot(data=., aes(x=hamatop)) +
  geom_histogram(bins=20) +
  labs(title="Logs Incident PAR Histogram", x="Incident PAR", y="Count")
```



Reflected PAR

```
data_net %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  min()
```

```
## [1] 0
```

```
data_log %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  min()
```

```
## [1] 0
```

```
data_net %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  max()
```

```
## [1] 9480.77
```

```
data_log %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  max()
```

```
## [1] 9142.86
```

Fix hamabot like hamatop

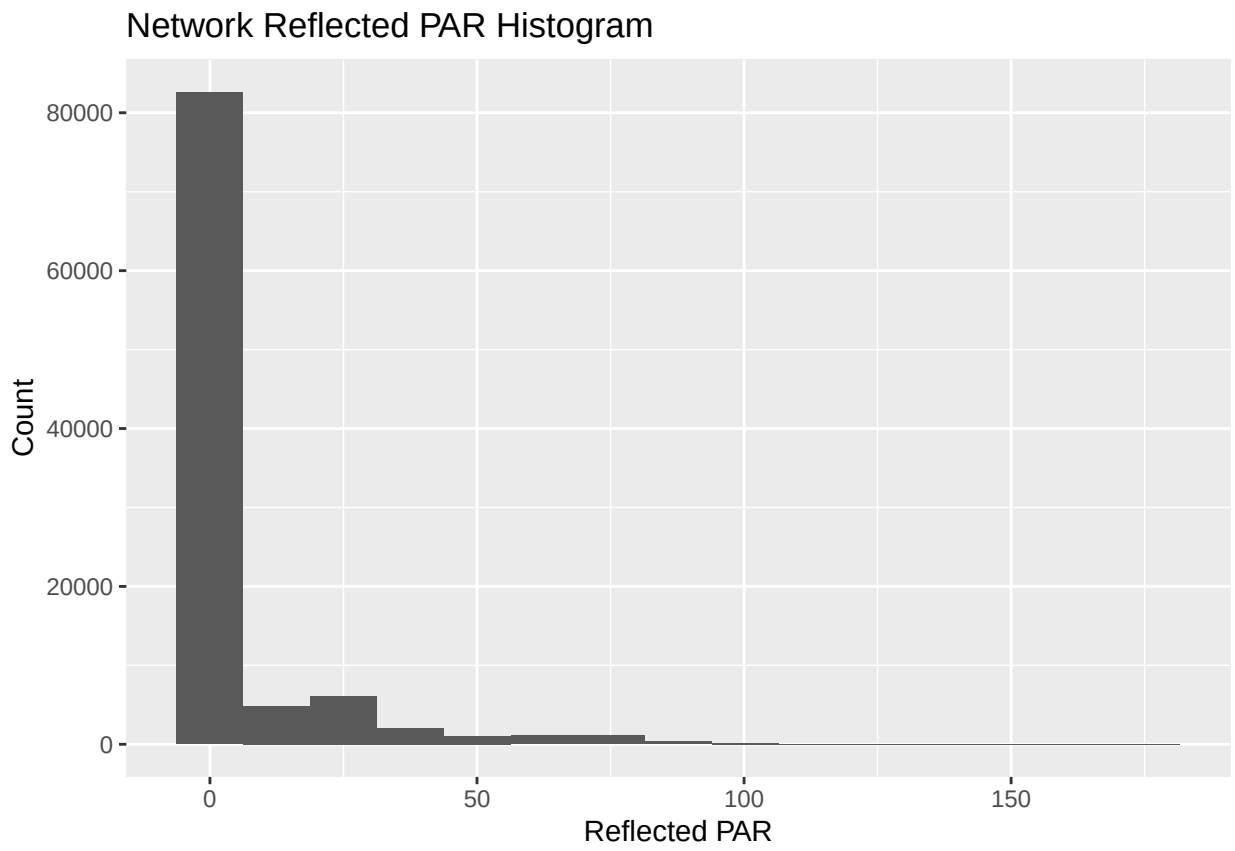
```
data_net <- data_net %>%
  mutate(hamabot=0.0185*hamabot)

data_log <- data_log %>%
  mutate(hamabot=0.0185*hamabot)

data_net %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  max()
```

```
## [1] 175.3942
```

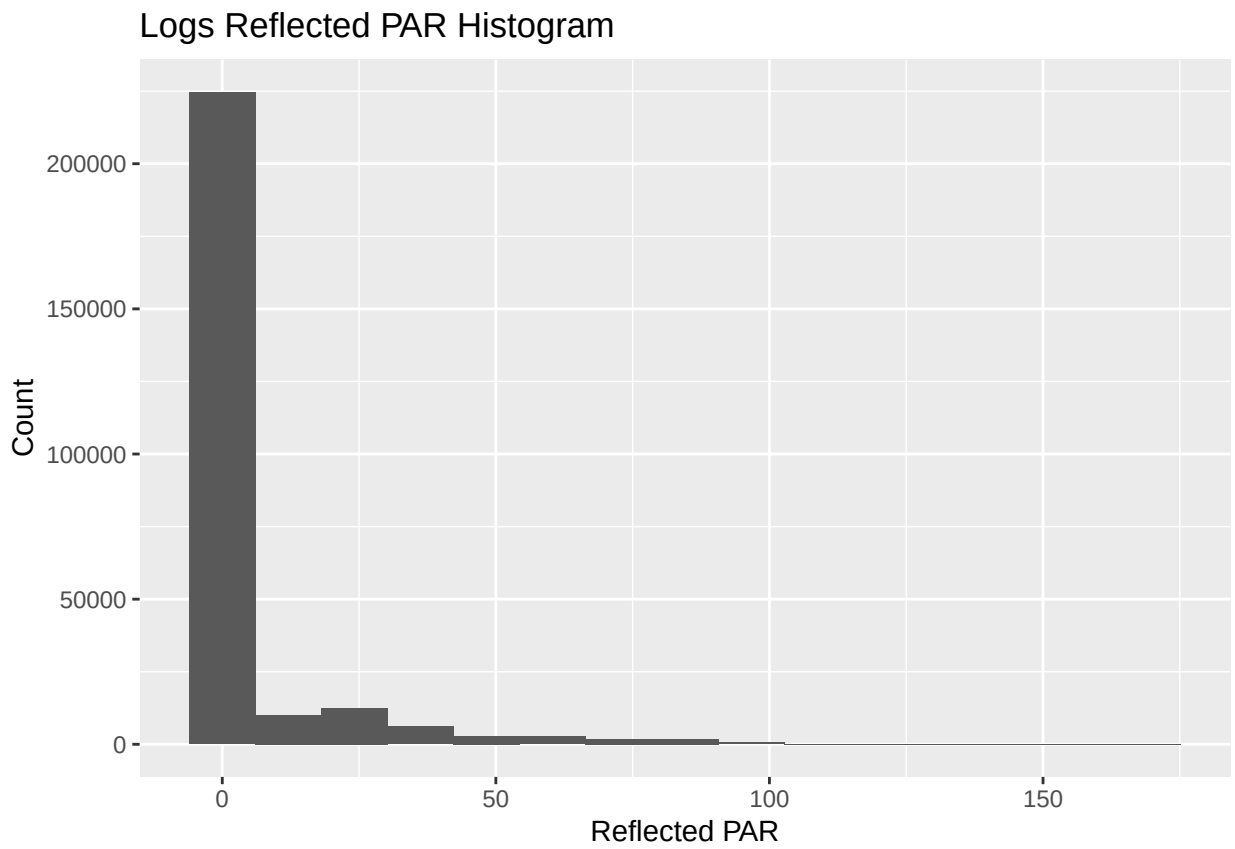
```
data_log %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  select(hamabot) %>%
  pull() %>%
  max()
```

```
## [1] 169.1429
```

```
data_net %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=hamabot)) +
  geom_histogram(bins=15) +
  labs(title="Network Reflected PAR Histogram", x="Reflected PAR", y="Count")
```




```
data_log %>%
  filter(!is.na(hamabot) & voltage <= 3 & voltage >= 2.4) %>%
  ggplot(data=., aes(x=hamabot)) +
  geom_histogram(bins=15) +
  labs(title="Logs Reflected PAR Histogram", x="Reflected PAR", y="Count")
```



b)

```
data_both <- merge(data_net, data_log, by = c("epoch", "nodeid"), all.y = T, suffixes = c(".net", ".log
data_all <- merge(data_both, dates, by = "epoch", all.x = T)
```

```
missing_data <-  data_all %>%
  select(ends_with(".log"), epochDates, epochDays) %>%
  filter_all(any_vars(is.na(.)))
  #filter(is.na(parent.log) | is.na(voltage.log) | is.na(depth.log) | is.na(humidity.log) | is.na(humid
  #       is.na(humid_adj.log) | is.na(hamatop.log) | is.na(hamabot.log))

paste0("# of missing measurements: ", nrow(missing_data))
```

```
## [1] "# of missing measurements: 8760"
```

```
missing_data %>%
  select(epochDates) %>%
  summarise(min = min(epochDates),
            max = max(epochDates))
```

```
##                  min                 max
## 1 2004-04-30 08:05:00 2004-05-25 21:15:00
```

```
data_all <- data_all %>%
  filter(!is.na(parent.log) & !is.na(voltage.log) & !is.na(depth.log) & !is.na(humidity.log) & !is.na(hu
         !is.na(humid_adj.log) & !is.na(hamatop.log) & !is.na(hamabot.log))
row.names(data_all) <- NULL
```

c)

```
locations <- read.table("data/mote-location-data.txt")
names(locations) <- locations[1,]
names(locations)[1] <- "nodeid"
locations <- data.frame(locations[-1,])
row.names(locations) <- NULL
```

```
data_all <- merge(data_all, locations, by = "nodeid", all.x = T)
```

The network data and logs data have identical columns in this merged data frame except for voltage. This must be because network voltage data had to be converted from ADC so the conversion may not be exact. Since network data is missing for some rows, logs data is not missing for any rows, and logs data has identical values as network data (other than voltage but logs voltage is more accurate anyways), we can remove network data columns from the merged data frame. In addition, result_time and epochDates represent the same information with epochDates from sonoma-dates being more accurate, so I will also remove the result_time column from the logs data.

```
data_all <- data_all %>%
  select(-ends_with(".net"), -result_time.log)
names(data_all)[3:11] <- str_replace(names(data_all)[3:11], ".log", "")
names(data_all)
```

```
##  [1] "nodeid"     "epoch"      "parent"     "voltage"    "depth"
##  [6] "humidity"   "humid_temp" "humid_adj"  "hamatop"    "hamabot"
## [11] "epochDates" "epochDays"  "Height"     "Direc"      "Dist"
## [16] "Tree"
```

There are 16 total variables in my complete merged data frame that includes network and logs data, dates data, and locations data.

d)

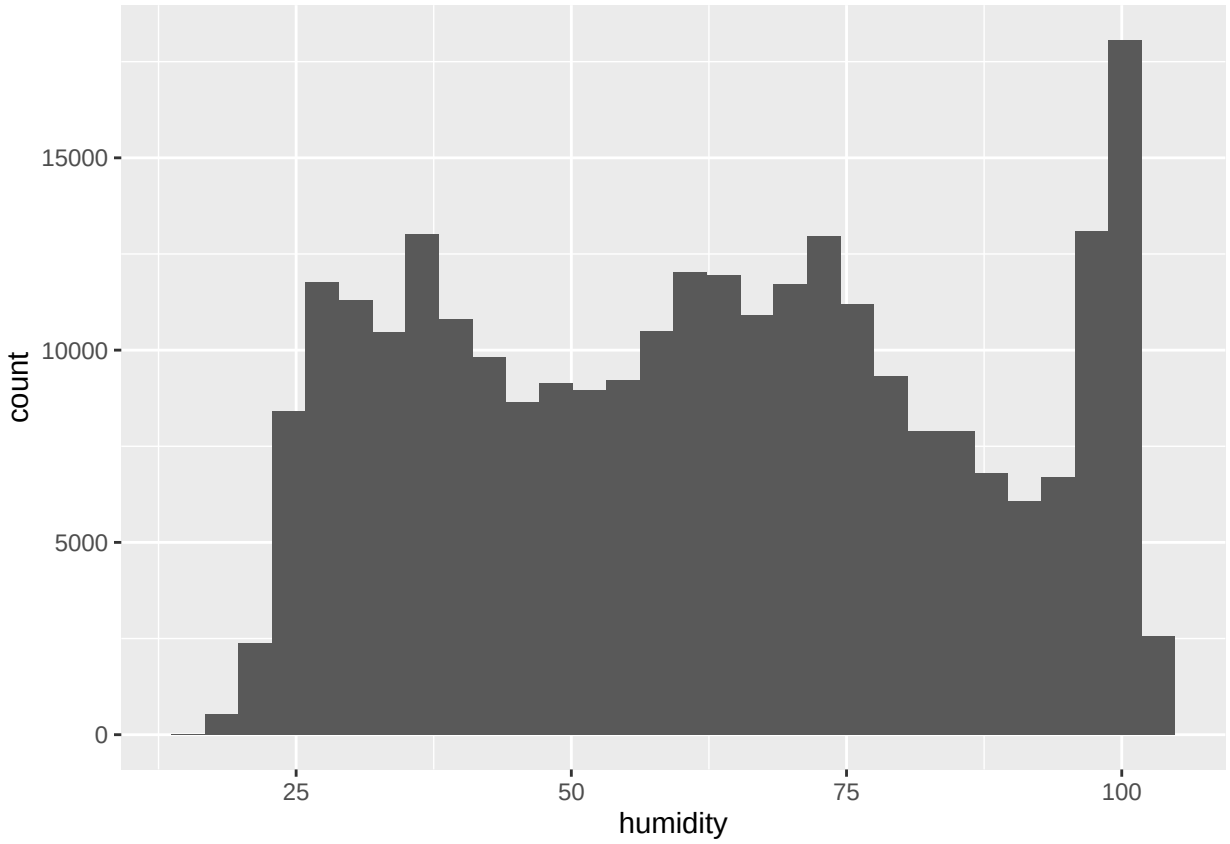
Since sensors with low batteries were not accurate, we removed all data with voltages greater than 3 and less than 2.4 like Tolle et al. Below are the quantiles and histograms of each variable after removing the voltages and before removing any other outliers.

```
data_all <- data_all %>%
  filter(voltage <= 3 & voltage >= 2.4)
```

```
quantile(data_all$humidity)
```

```
##       0%      25%      50%      75%     100%
##  16.2653  40.9730  62.3336  80.7267 104.4050
```

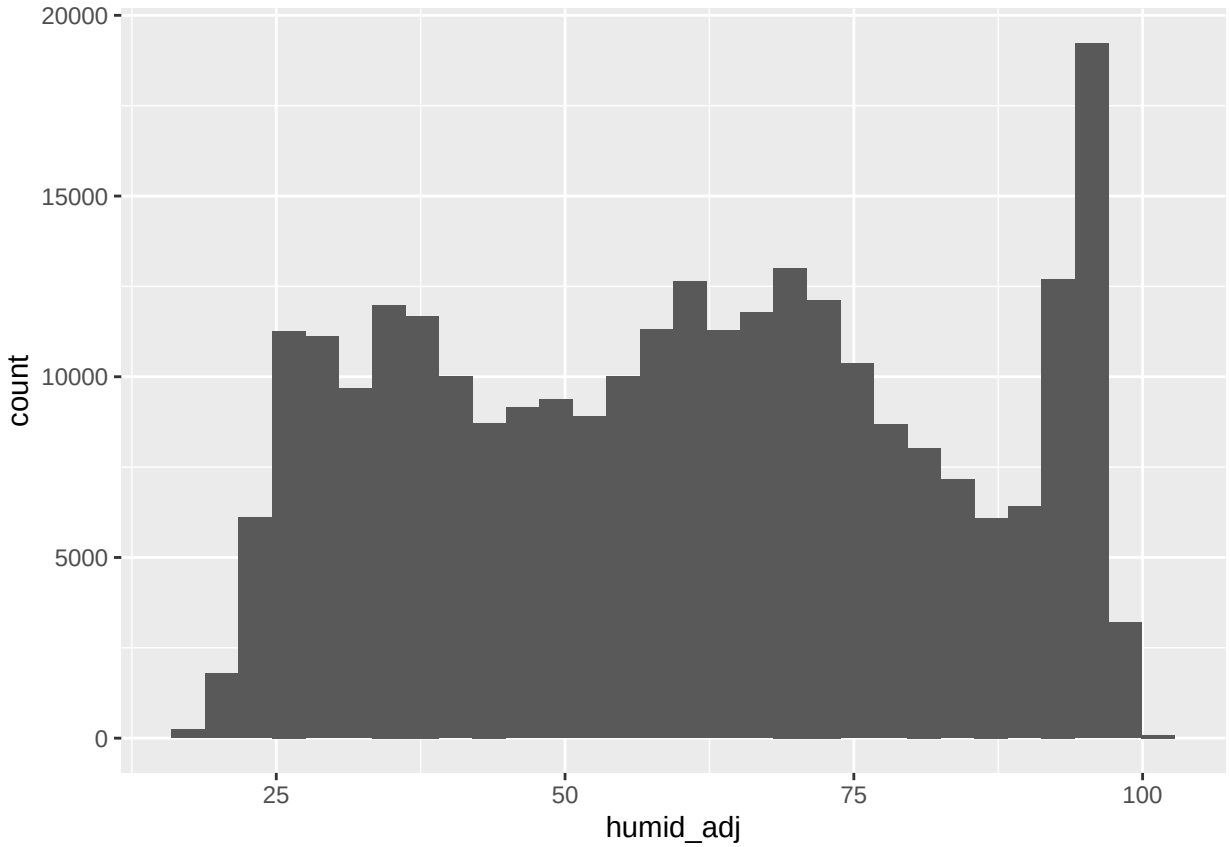
```
data_all %>%
  ggplot(., aes(x=humidity)) +
  geom_histogram()
```




```
quantile(data_all$humid_adj)
```

```
##        0%       25%       50%       75%      100%
##  16.22820  40.44530  60.67515  77.71650 100.22300
```

```
data_all %>%
  ggplot(., aes(x=humid_adj)) +
  geom_histogram()
```

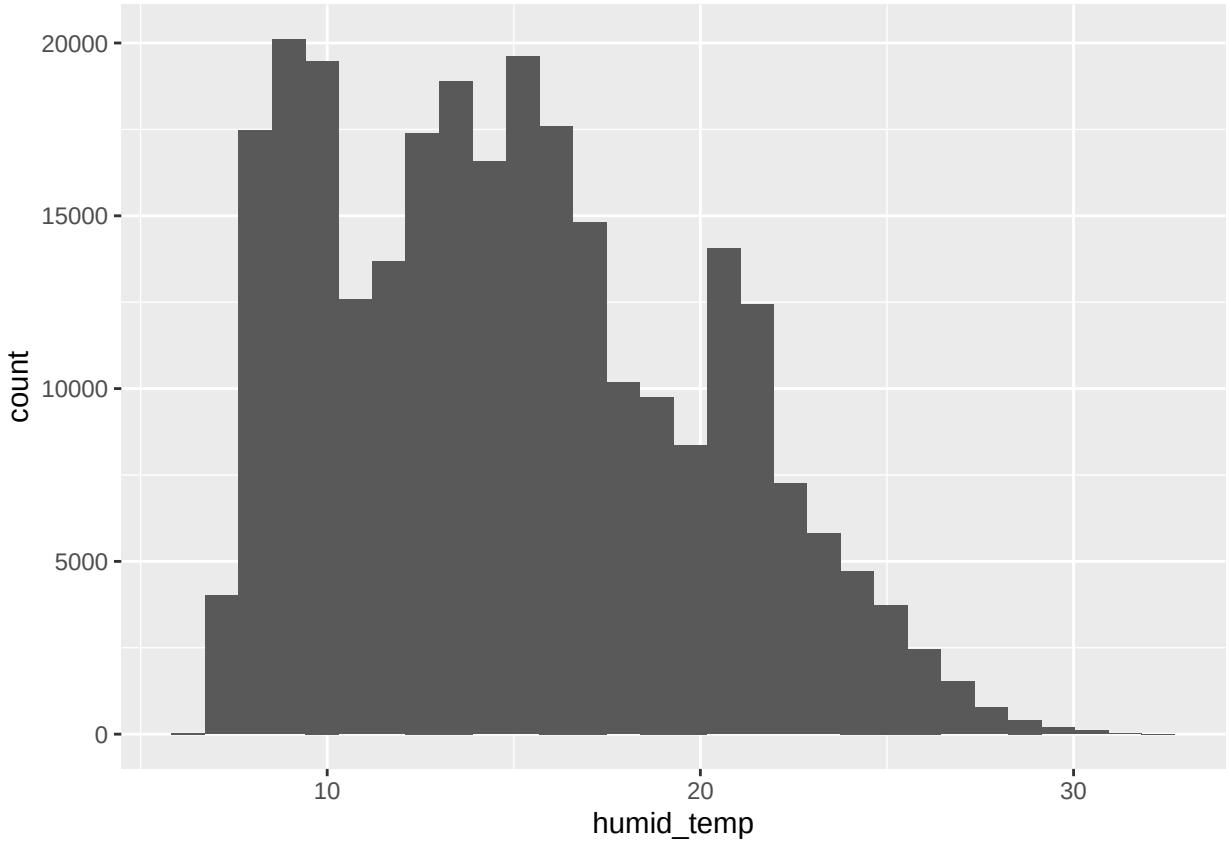


The quantiles in humid_adj more closely matches that of Tolle et al. compared to the quantiles of humidity so we should use humid_adj in place of humidity. This allows us to remove 1 more variable so we have a total of 15 variables. There do not seem to be any other outliers in humid_adj.

```
quantile(data_all$humid_temp)
```

```
##      0%     25%     50%     75%    100%
##  6.5820 10.8156 14.6082 18.6556 32.5814
```

```
data_all %>%
  ggplot(., aes(x=humid_temp)) +
  geom_histogram()
```

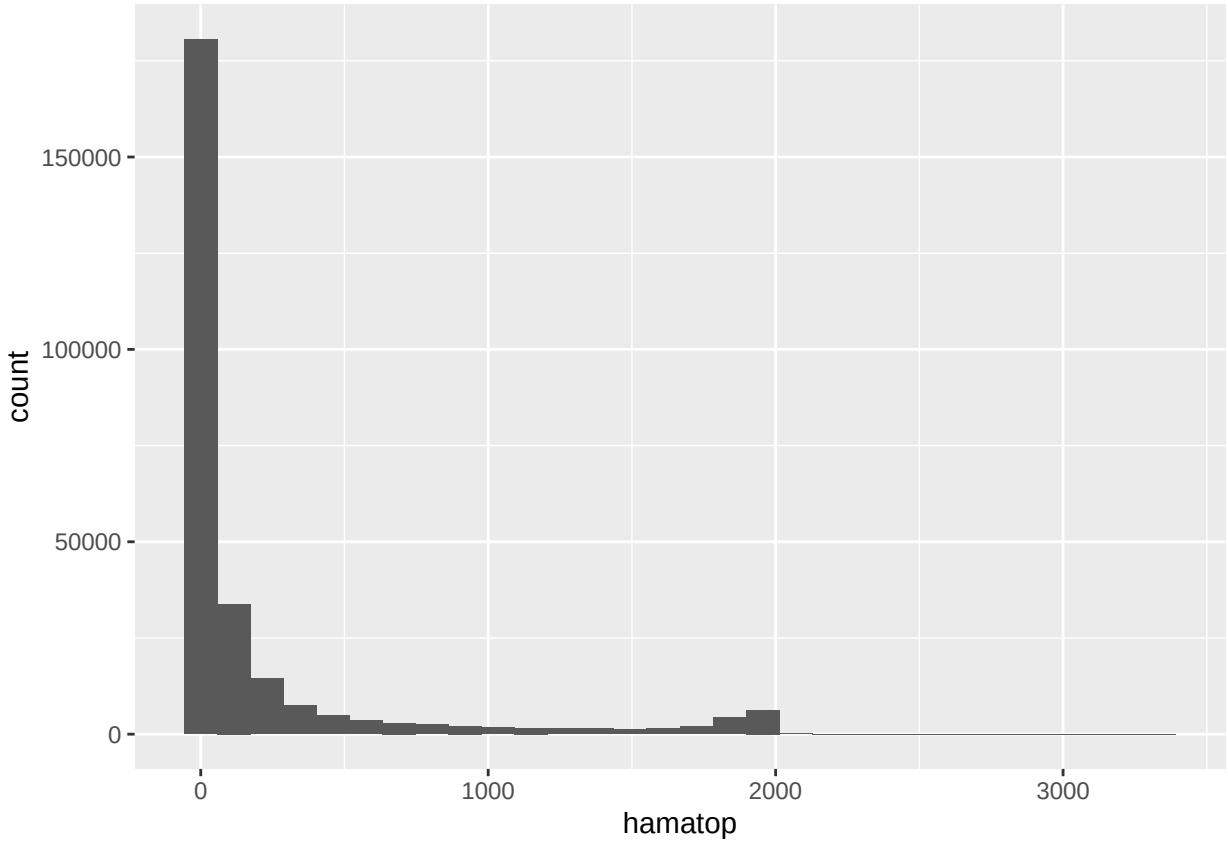


The quantiles of temperature match up well with that of Tolle et al. so there does not seem to be any other outliers in temperature.

```
quantile(data_all$hamatop)
```

```
##        0%       25%       50%       75%      100%
##    0.0000    0.0000    0.0000  129.1574 3334.7175
```

```
data_all %>%
  ggplot(., aes(x=hamatop)) +
  geom_histogram()
```




```
data_all %>%
  filter(hamatop > 2154) %>%
  select(nodeid) %>%
  unique()
```

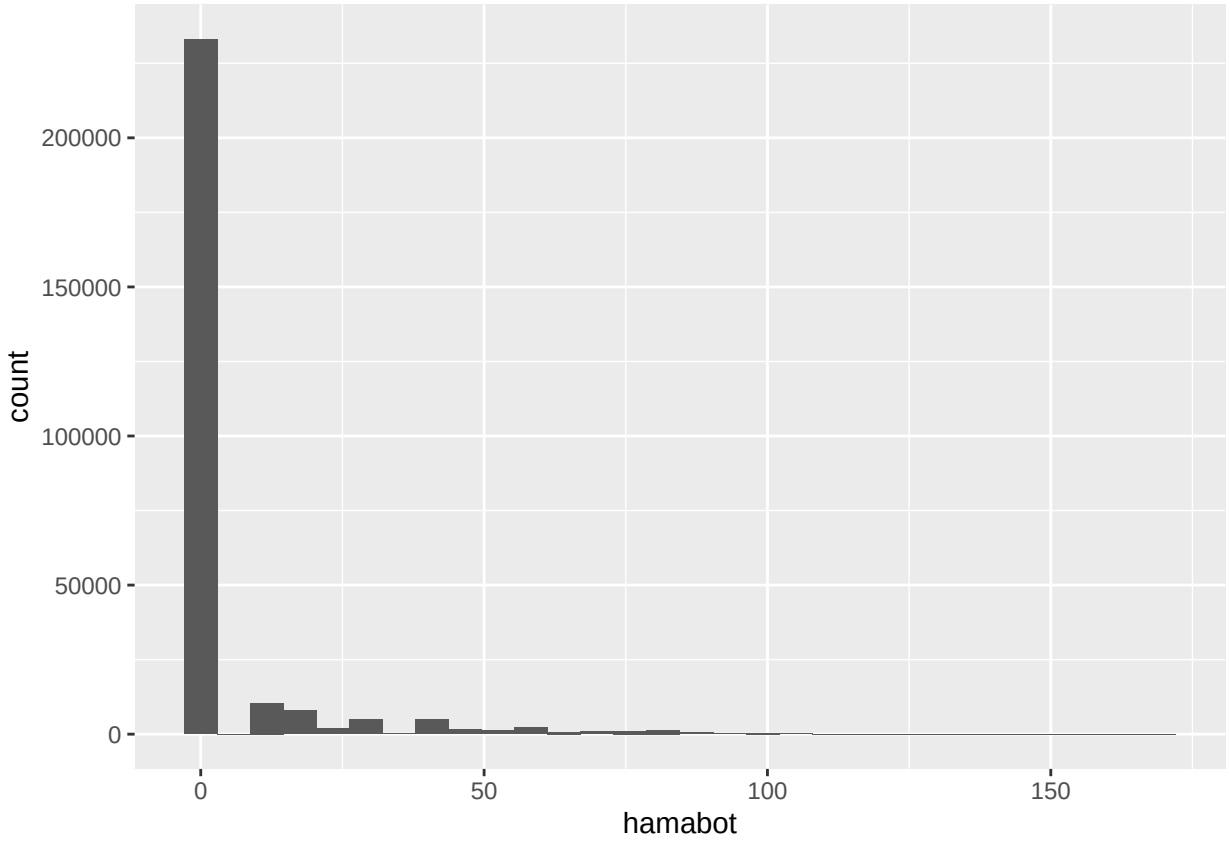
```
##   nodeid
## 1     40
```

There seem to be some abnormally high values of incident PAR and when we look for the nodes that gave a reading higher than the max incident PAR reading of 2154 given in the paper, we see that node 40 is solely responsible for all those readings. So we will remove all observations with node 40.

```
quantile(data_all$hamabot)
```

```
##       0%      25%      50%      75%     100%
##   0.0000   0.0000   0.0000   0.0000 169.1429
```

```
data_all %>%
  ggplot(., aes(x=hamabot)) +
  geom_histogram()
```



The quantiles of reflected PAR match up well with that of Tolle et al. so there does not seem to be any other outliers in reflected PAR.

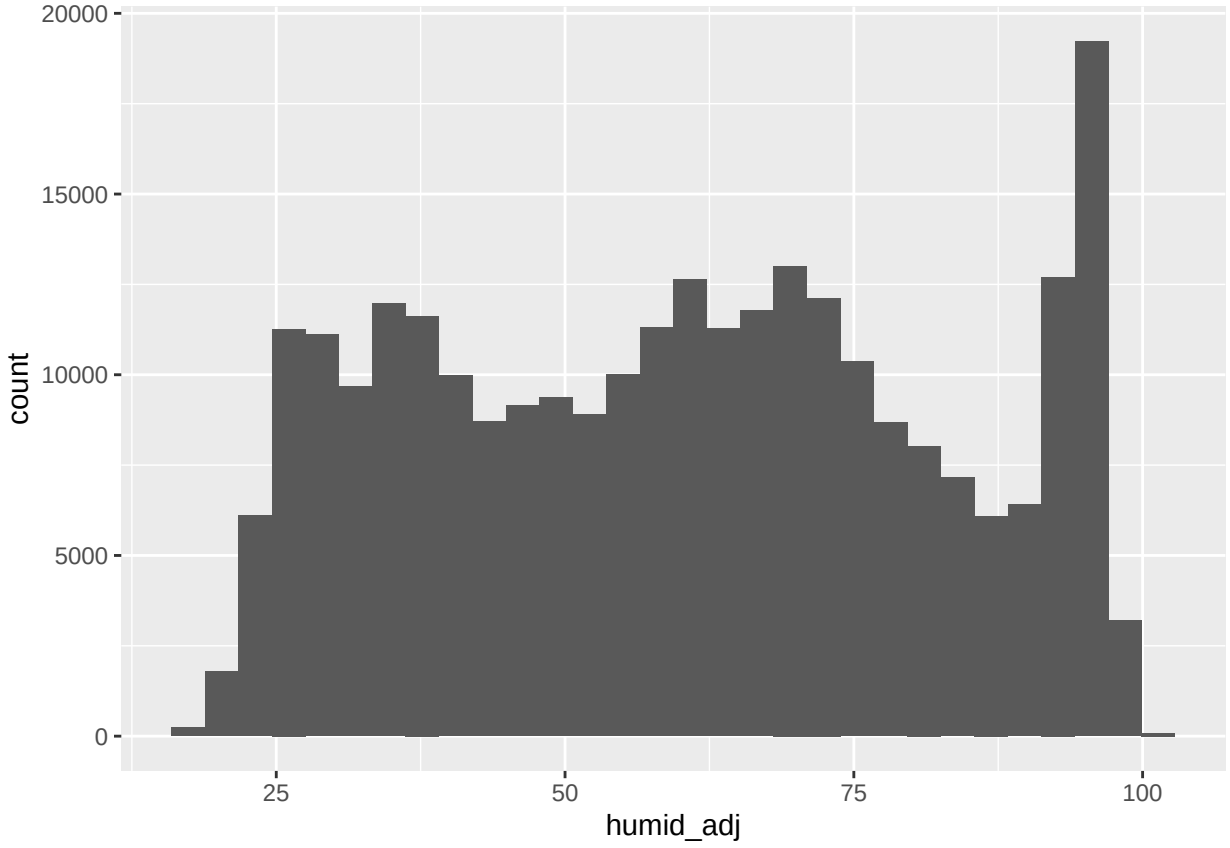
Below are the quantiles and histograms of each variable after removing both the voltages and observations with node 40.

```
data_all <- data_all %>%
  filter(nodeid != 40)
```

```
quantile(data_all$humid_adj)
```

```
##        0%       25%       50%       75%      100%
##  16.22820  40.46055  60.68580  77.72570 100.22300
```

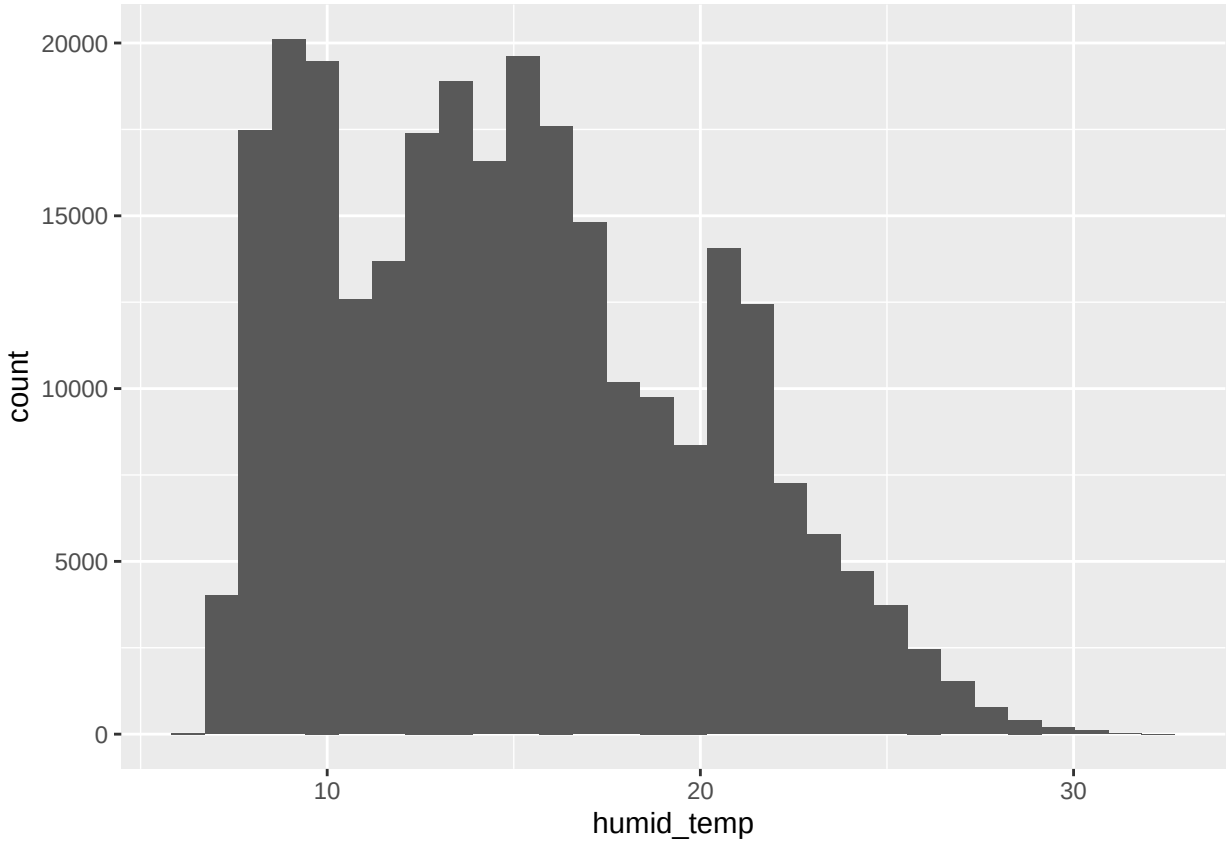
```
data_all %>%
  ggplot(., aes(x=humid_adj)) +
  geom_histogram()
```




```
quantile(data_all$humid_temp)
```

```
##      0%     25%     50%     75%    100%
##  6.5820 10.8156 14.6082 18.6458 32.5814
```

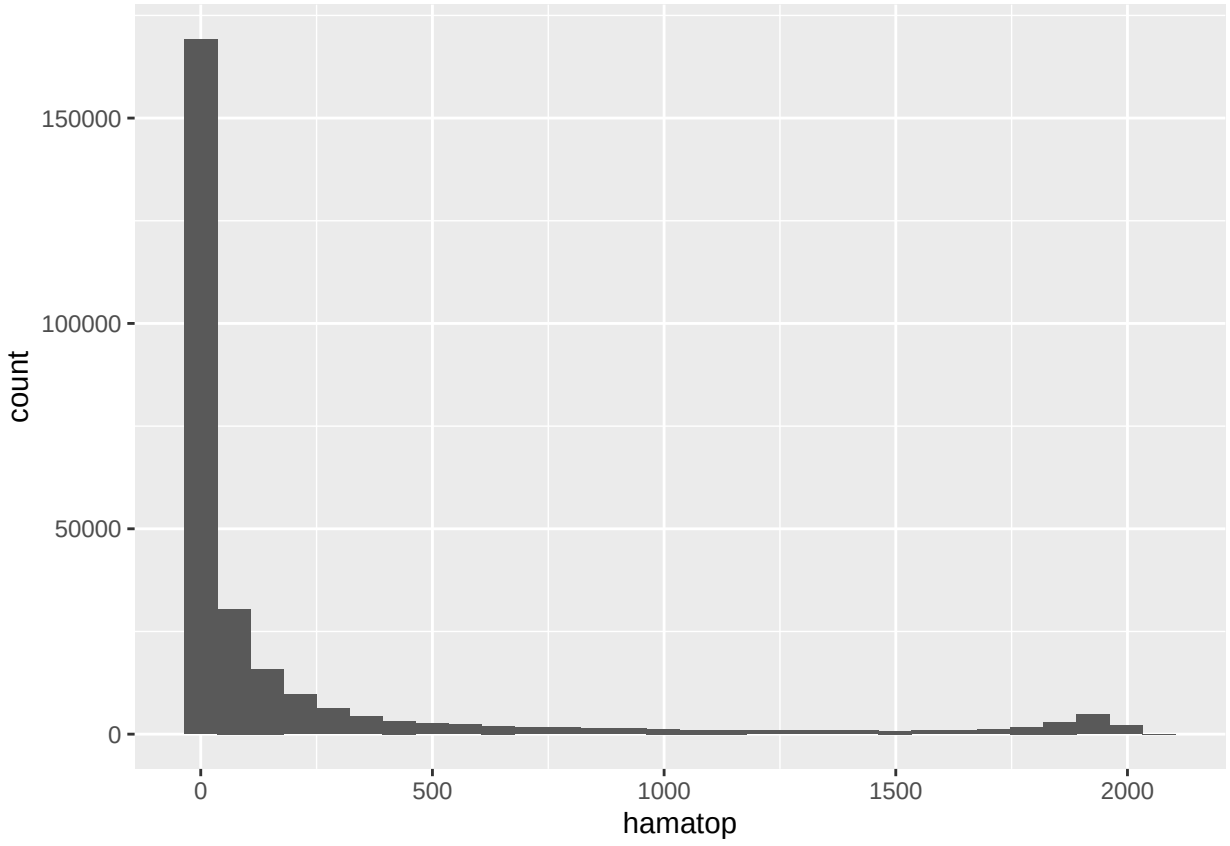
```
data_all %>%
  ggplot(., aes(x=humid_temp)) +
  geom_histogram()
```




```
quantile(data_all$hamatop)
```

```
##        0%       25%       50%       75%      100%
##    0.0000    0.0000    0.0000  128.5622 2068.6700
```

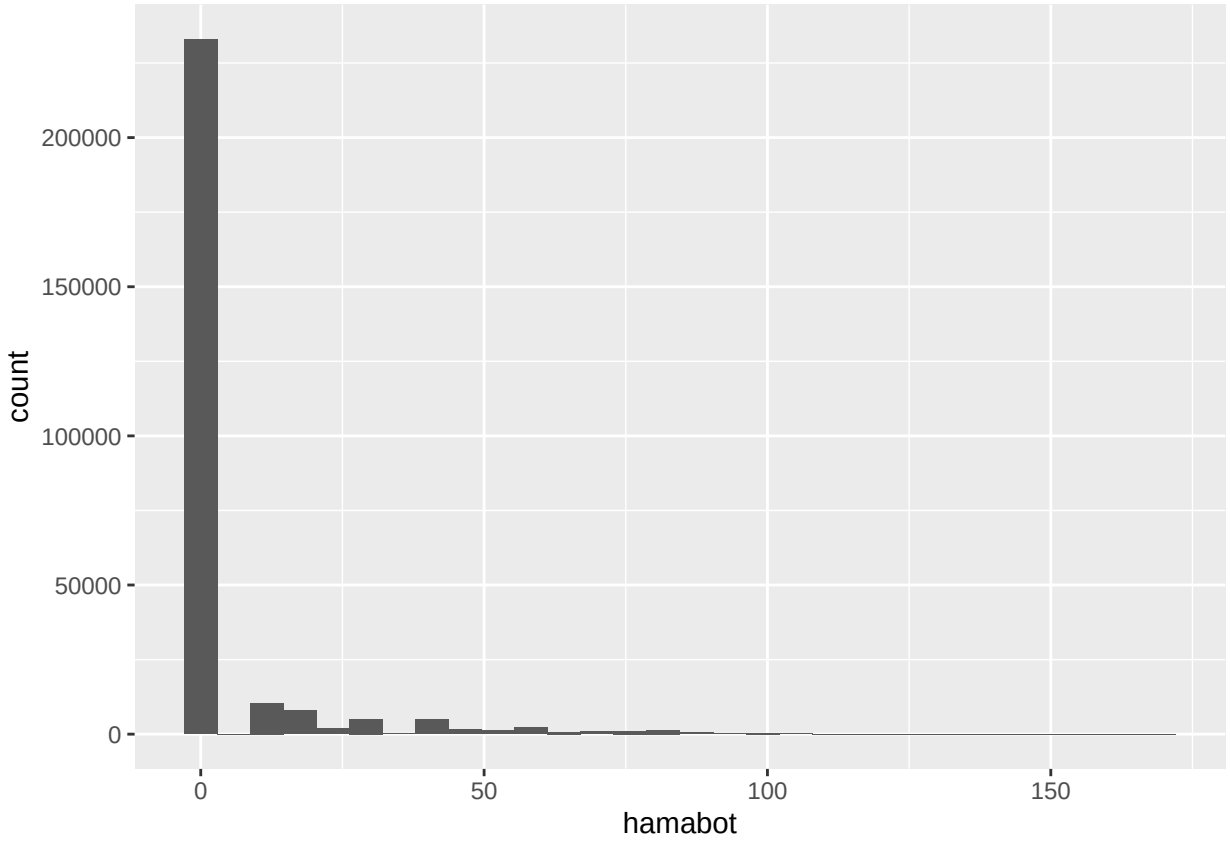
```
data_all %>%
  ggplot(., aes(x=hamatop)) +
  geom_histogram()
```




```
quantile(data_all$hamabot)
```

```
##       0%      25%      50%      75%     100%
##   0.0000   0.0000   0.0000   0.0000 169.1429
```

```
data_all %>%
  ggplot(., aes(x=hamabot)) +
  geom_histogram()
```



All quantiles and histograms of humidity, temperature, incident PAR, and reflected PAR match up well with the quantiles of Tolle et al. after removing observations with voltages greater than 3 and less than 2.4 and removing observations with node 40.

```
#write.csv(data_all, "data/data-all-dates-locations.csv")
```