Exploratory Data Analysis (EDA)

February 3, 2026

Today's plan

1. **Review: Data Cleaning**
2. **Exploratory Data Analysis**

Review: Our Data Cleaning Journey

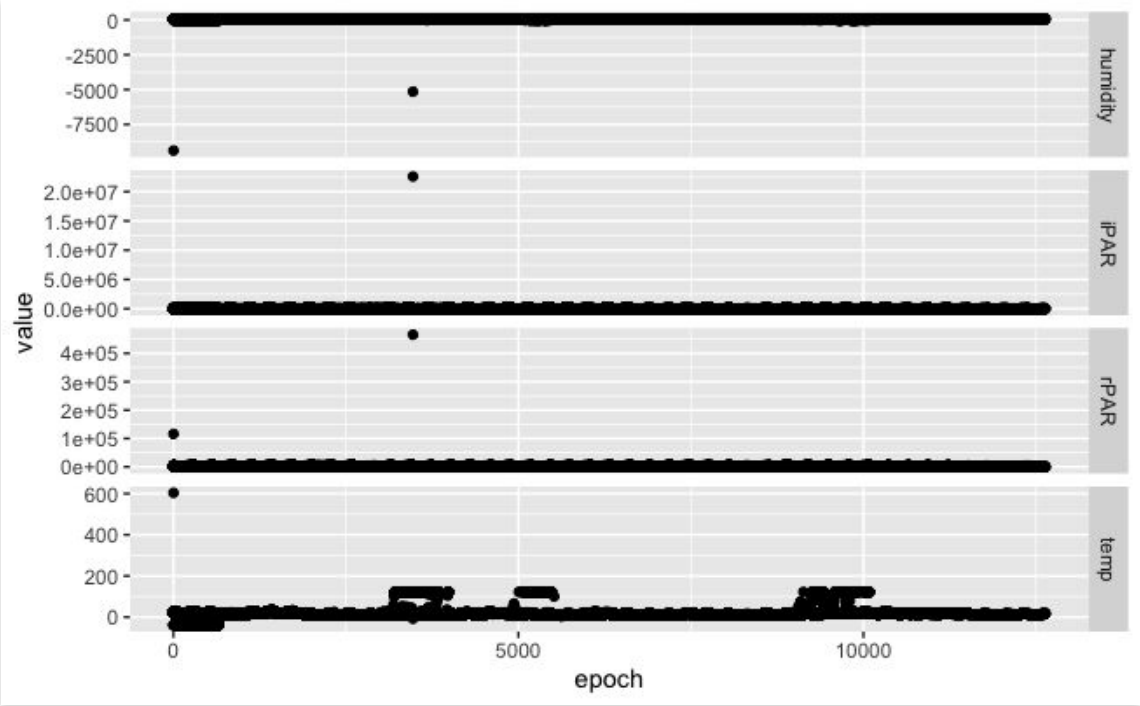
Our data cleaning journey

Where we started:

- Original data
  - Dates table
  - Mote location information
  - Three redwood sensor datasets (all, log, net)
    - Temperature
    - Humidity
    - Incident PAR
    - Reflected PAR

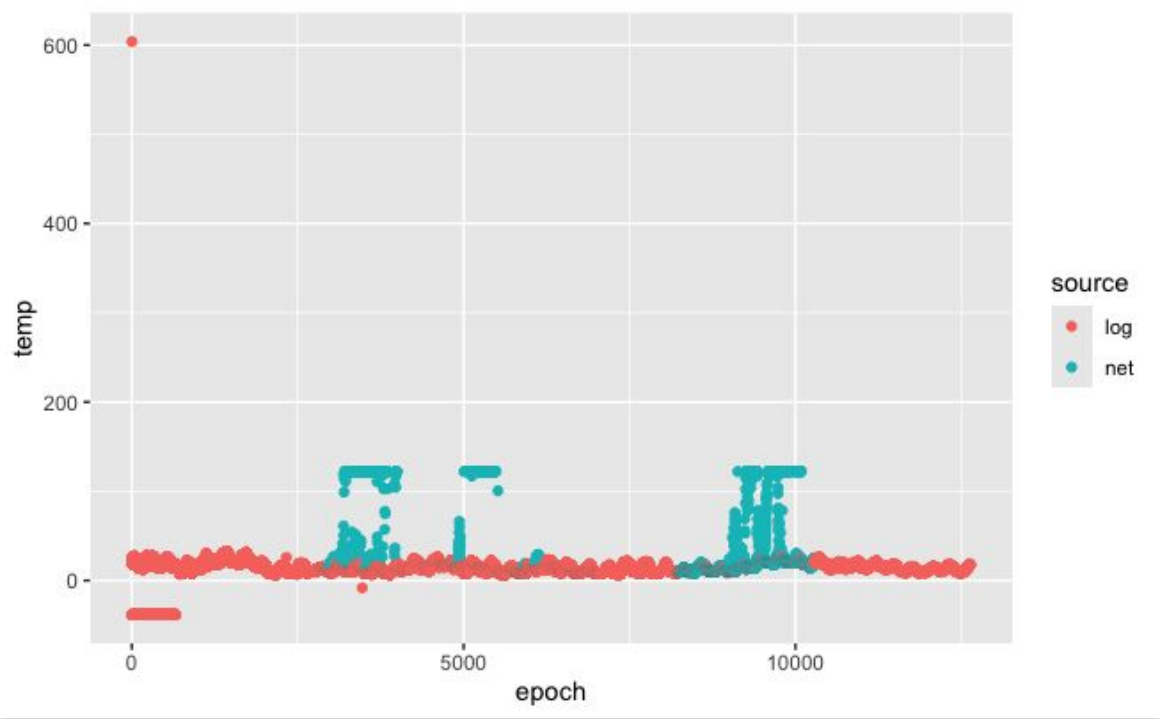
Our data cleaning journey

First glance:



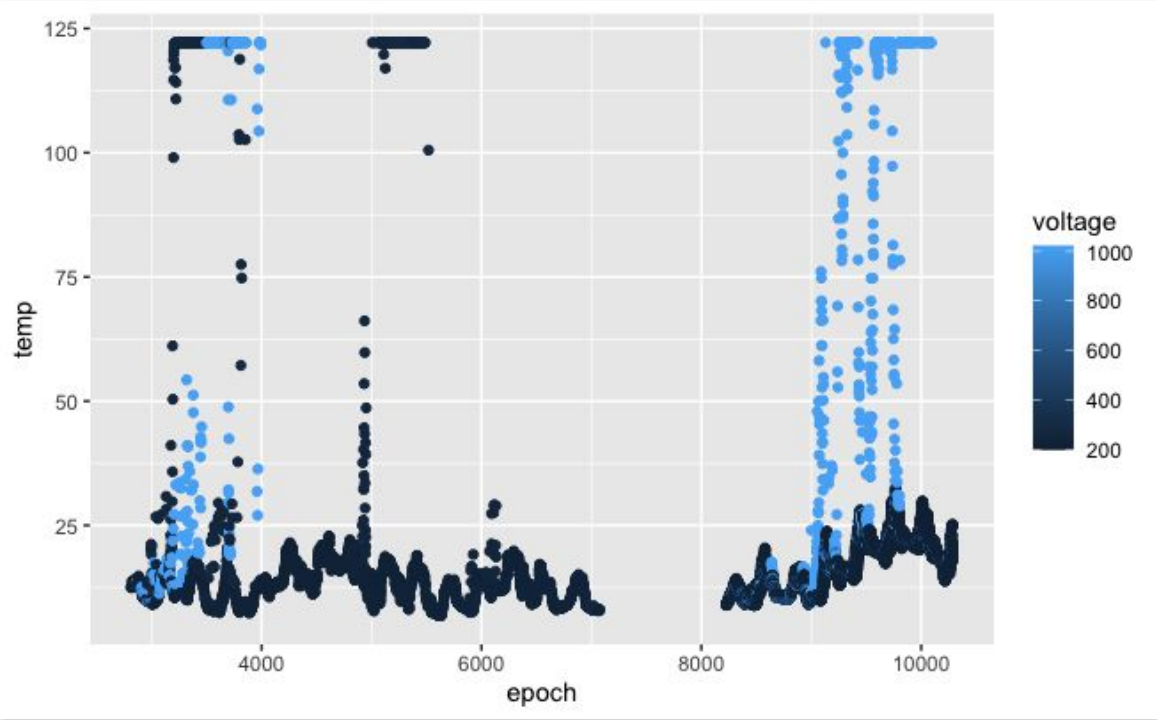
Our data cleaning journey

Taking a closer look at temperature (by source):



Our data cleaning journey

Why do the temperatures trail off like that?



Our data cleaning journey

Issues:

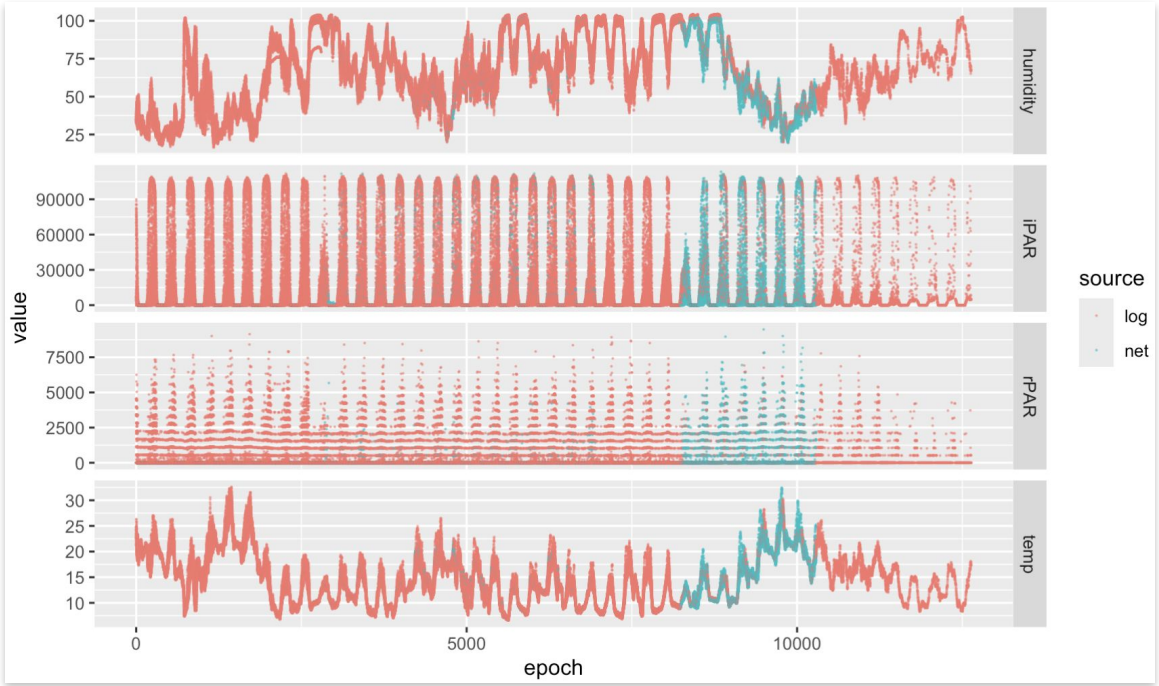
- Measurement error (e.g., absurd temperature and humidity values)
  - Likely a result of network or battery failure
- NAs
- Duplicates
- Missing motes and/or mote location info
- **Many** other issues…

When possible, the data collection process should guide our data cleaning!

Our data cleaning journey

What's next? Use this new information to **iteratively refine** your data cleaning, e.g.,

- Identify one issue
- Fix the issue
- Identify another issue
- Fix the issue
- Do some EDA
- Find another issue
- Fix the issue
- …



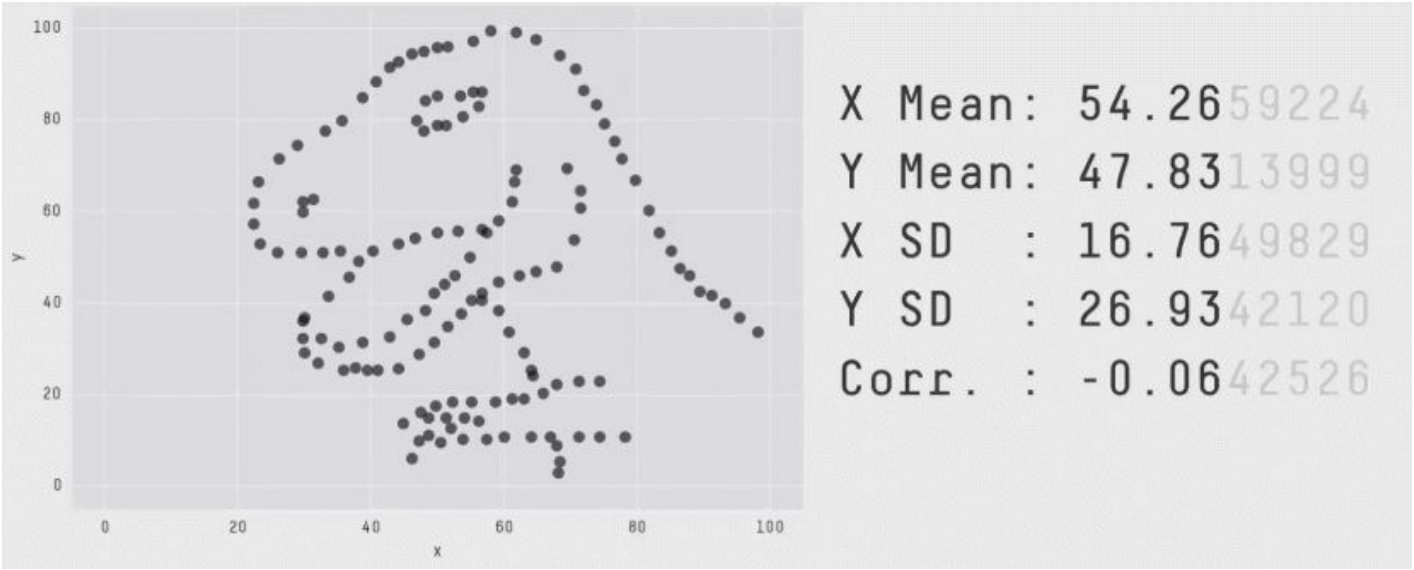
Exploratory Data Analysis (EDA)

Why do we need EDA/visualizations?

Visualizations can tell a more detailed story than numeric summaries

Why do we need EDA/visualizations?

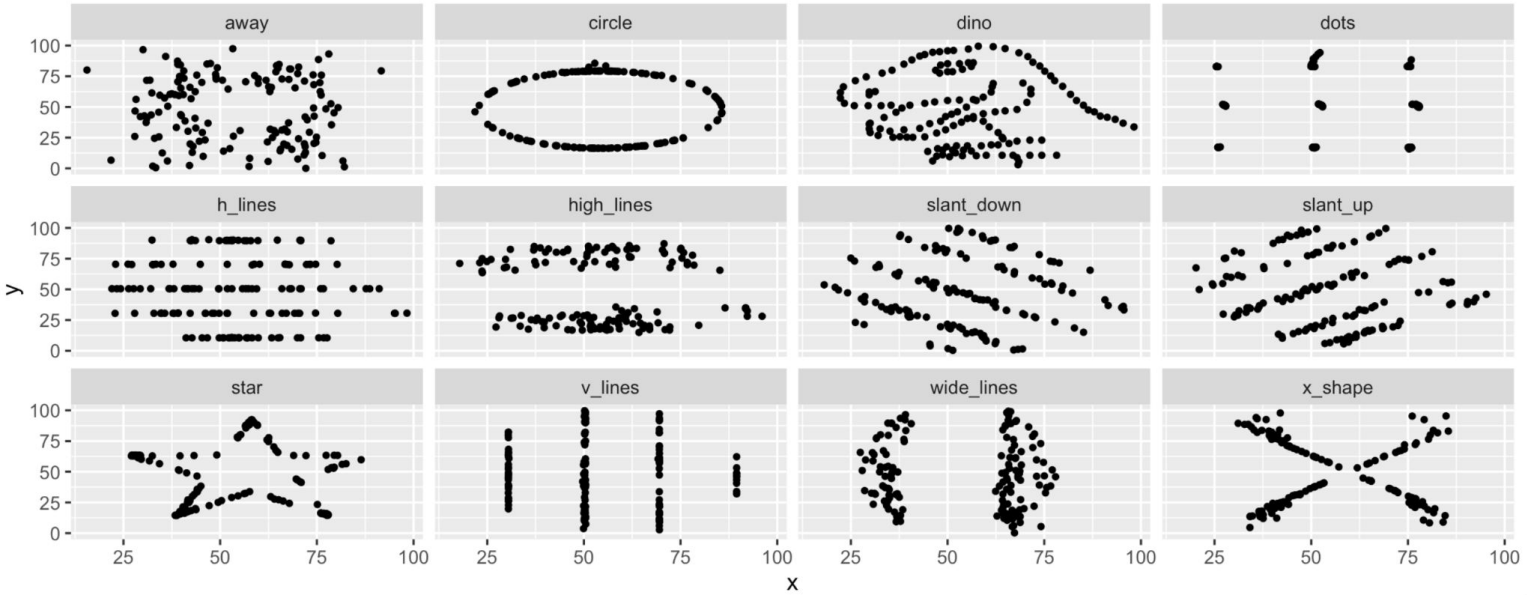
Visualizations can tell a more detailed story than numeric summaries



"The Datasaurus Dozen" [Matejka and Fitzmaurice (2017)]

Why do we need EDA/visualizations?

Visualizations can tell a more detailed story than numeric summaries



"The Datasaurus Dozen" [Matejka and Fitzmaurice (2017)]

Same story holds with p-values! Pictures > p-values

Exploratory Data Analysis (EDA): Purpose

What can we use EDA for?

- To illuminate oddities with the data and **inform data cleaning**
- To provide insights on the inherent data structure that can **guide modeling**
- To **discover substantively-meaningful patterns** (i.e., data-driven discoveries)
  - Your two EDA plots for Lab 1 should be for this purpose

Two modes of EDA plots

- “Scratchwork”: for internal use
- “Publication-quality”: for public use
  - Your two EDA plots for Lab 1 should be of publication-quality

"Scratchwork" Plots (for internal use)

"Scratchwork" Mode: Quantity over quality

What are some helpful visualizations when digging into a dataset for the first time?

- Data distribution
- Relationship between variables/features (X)
- Relationship between variables (X) and response (y)
- Others...
  - Structured data: time series, images, text, geospatial data, ...

For each, remember to plot the information in multiple different ways

"Scratchwork" Mode: Quantity over quality

What are some helpful visualizations when digging into a dataset for the first time?

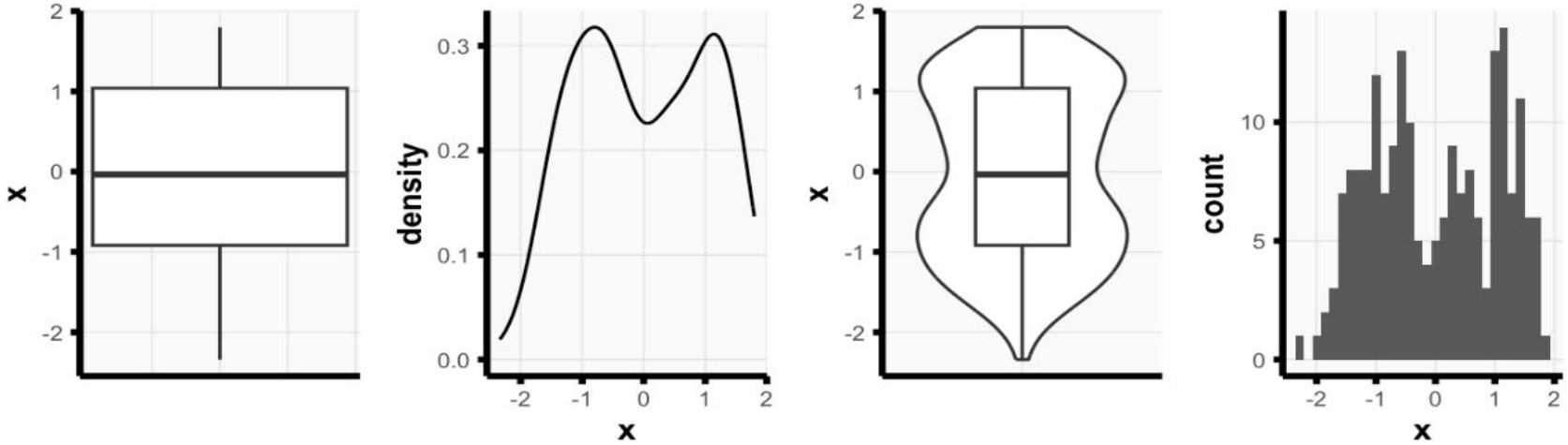
- Data distribution
- Relationship between variables/features (X)
- Relationship between variables (X) and response (y)
- Others...
  - Structured data: time series, images, text, geospatial data, ...

For each, remember to plot the information in multiple different ways

Different plots of the same data reveal different information

Quantity over quality: Plot the same data in multiple different ways

Example: Visualizing the distribution of your data



+ different kernel bandwidth, number of histogram bins, etc

"Scratchwork" Mode: Quantity over quality

What are some helpful visualizations when digging into a dataset for the first time?

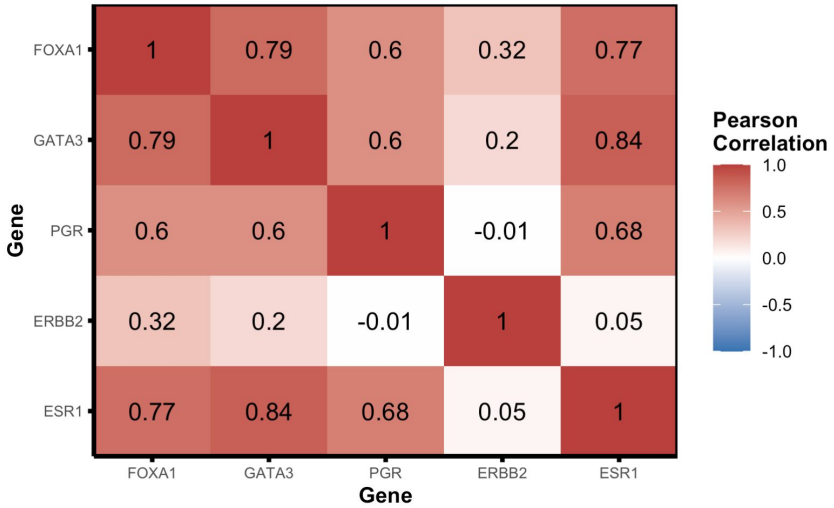
- Data distribution
- Relationship between variables/features (X)
- Relationship between variables (X) and response (y)
- Others...
  - Structured data: time series, images, text, geospatial data, ...

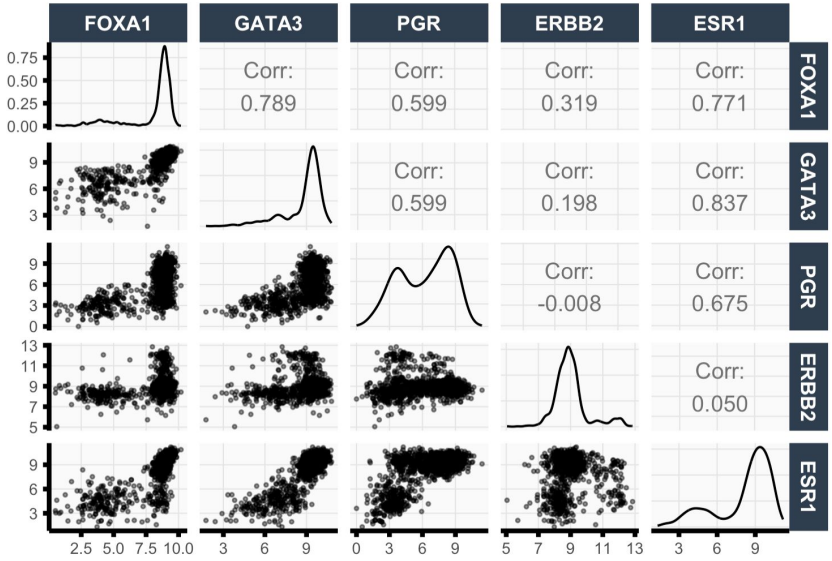
For each, remember to plot the information in multiple different ways

Different plots of the same data reveal different information

Quantity over quality: Plot the same data in multiple different ways

Example: Visualizing the relationship between variables/features (X)



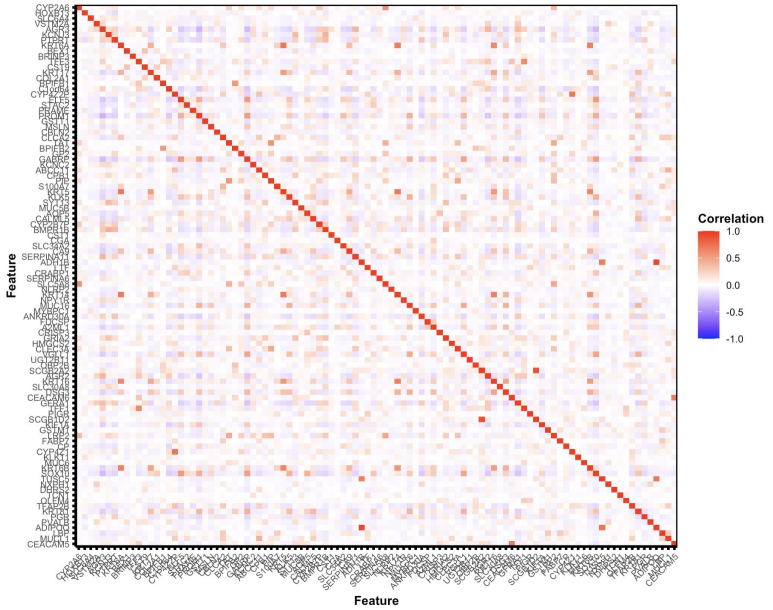


Different plots of the same data reveal different information

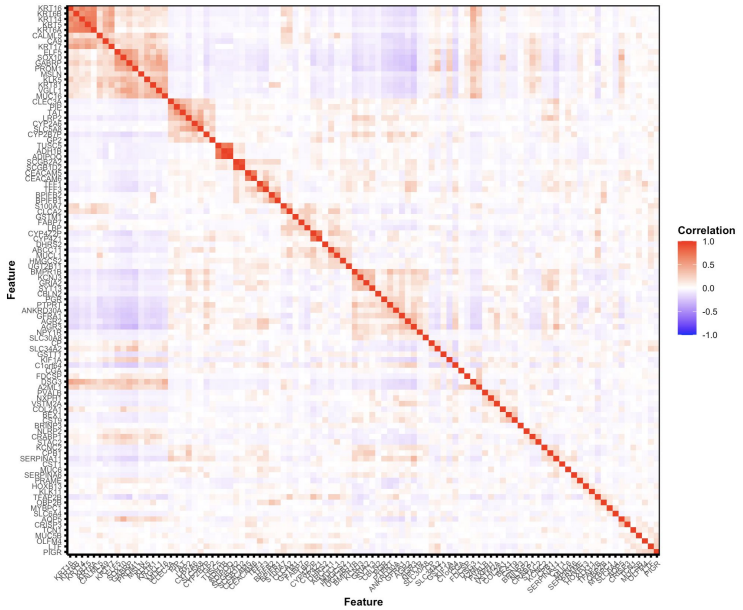
Quantity over quality: Plot the same data in multiple different ways

Example: Visualizing the relationship between variables/features (X)

Unclustered Heatmap



Clustered Heatmap



"Scratchwork" Plotting Tools

Main plotting libraries:

- **R:** ggplot2
- **Python:** matplotlib, seaborn, plotnine (aka ggplot2)

Some little-known but useful R/Python functions to know:

- **Heatmaps:** `ggplot::geom_tile` and `seaborn.heatmap` (Python)
- **Clustered heatmaps:** `superheat::superheat` (R) and `seaborn.clustermap` (Python)
- **Pair plots:** `GGally::ggpairs` (R) and `seaborn.pairplot` (Python)
- **3D plots:** `plotly` (R/Python)

For a ggplot2 tutorial, see course website.

"Publication-quality" Plots (for public use)

When presenting EDA visualizations to the public...

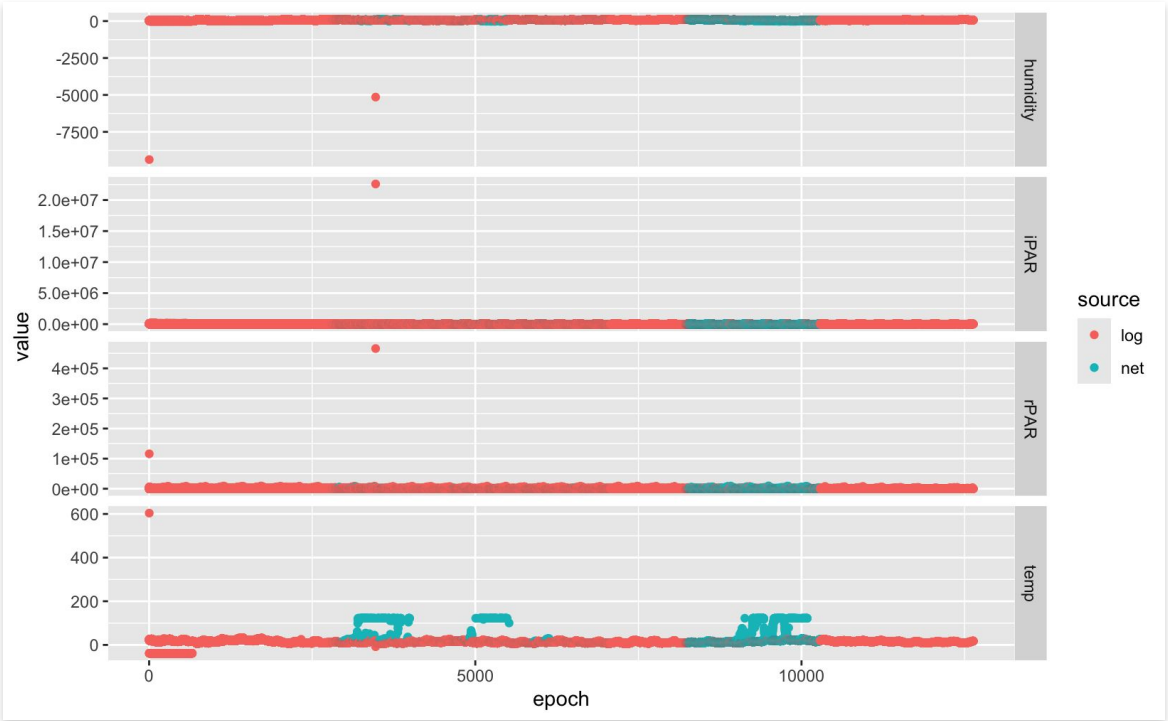
First think about your main message.

Then craft the plot to clearly communicate this (singular) message.

Remember the #1 rule: The simpler, the better.

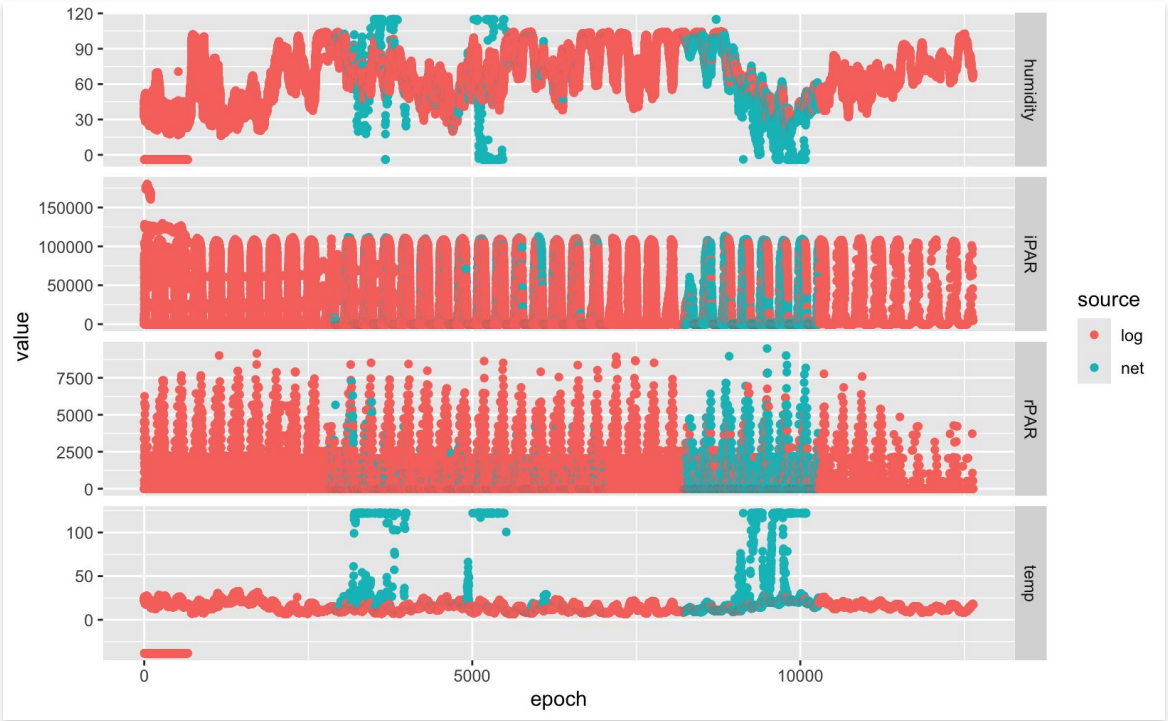
EDA Example: Before

Main message: Outliers generally come from the network data



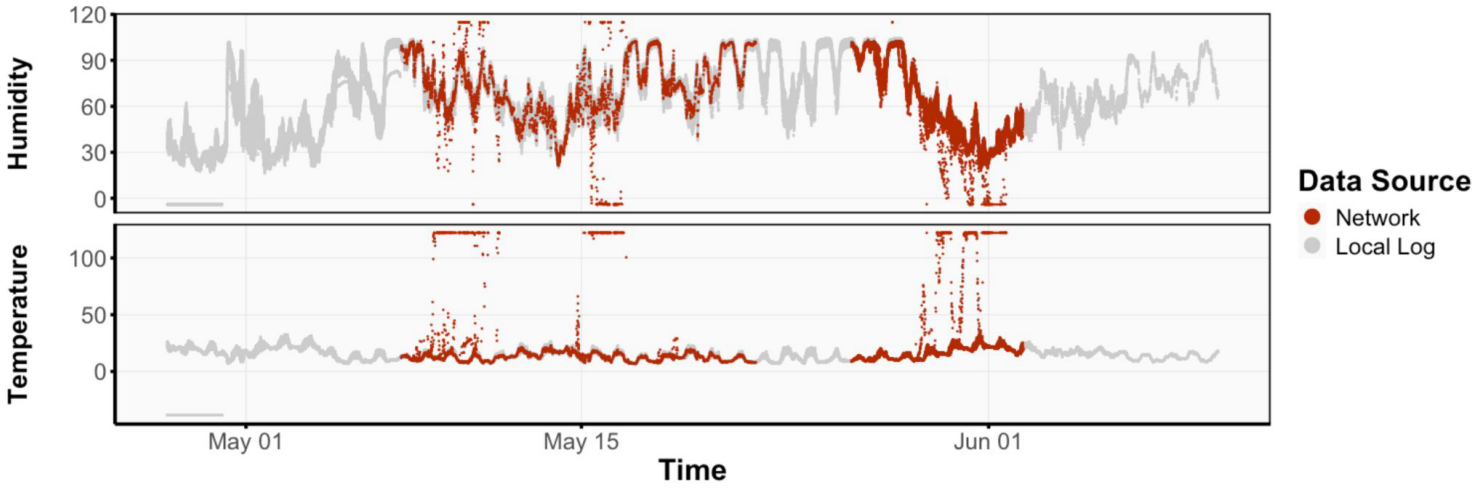
EDA Example: Before

Main message: Outliers generally come from the network data



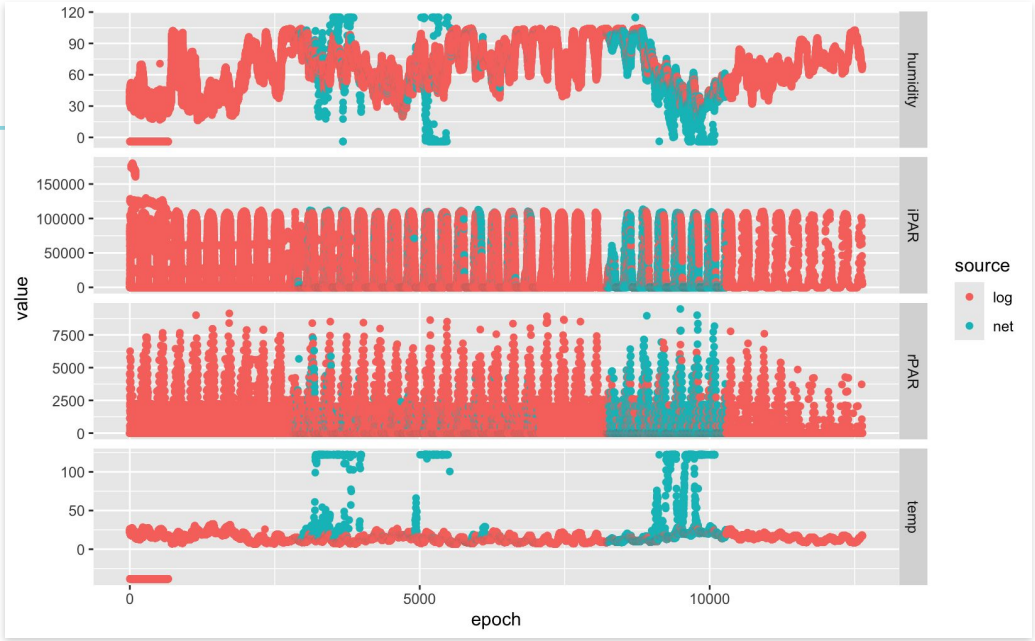
EDA Example: After

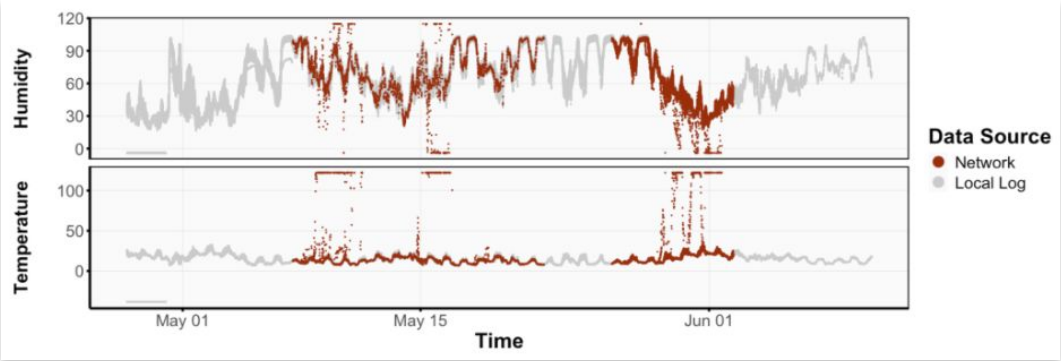
Main message: Outliers generally come from the network data



EDA Example

Spot the differences





Basic Aesthetics Checklist

- Labels should be meaningful, human-readable names (not variable names)
  - E.g., plot date/time instead of epochs in redwood lab
- Add labels and capitalize them appropriately
- Text size should be large enough and legible (e.g., in reports and on slides)
- Legend order matters
- Change the (ggplot) theme
- Add an informative caption
- Did I overplot?
- Choose colors thoughtfully

````
To add a figure caption in .qmd:

```{r/python}
#| label: fig-label
#| fig-cap: "An example caption."

<insert plotting code here>
```
````

Basic Aesthetics Checklist

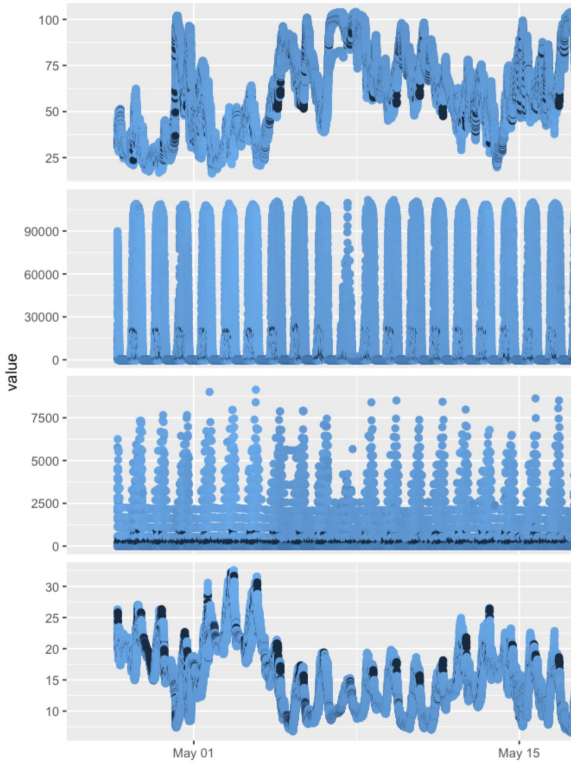
- Labels should be meaningful, human-readable names (not variable names)
  - E.g., plot date/time instead of epochs in redwood lab
- Add labels and capitalize them appropriately
- Text size should be large enough and legible (e.g., in reports and on slides)
- Legend order matters
- Change the (ggplot) theme
- Add an informative caption
- **Did I overplot?**
- **Choose colors thoughtfully**

To add a figure caption in .qmd:

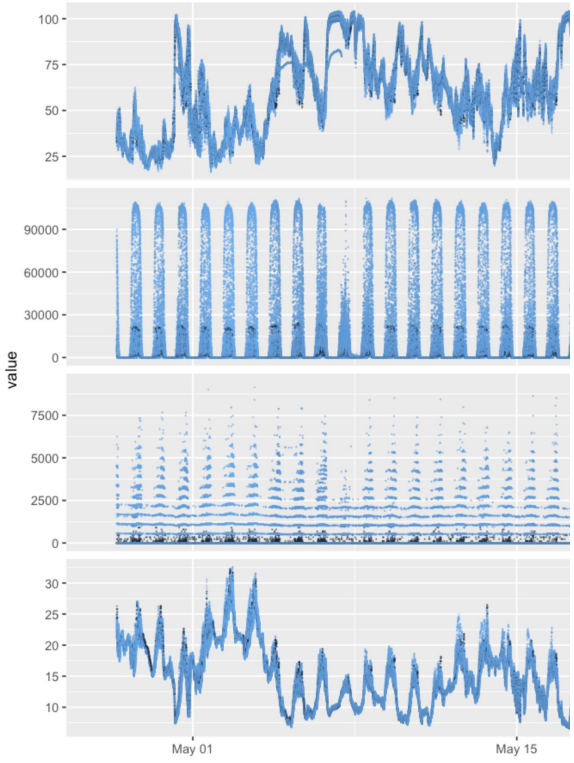
````
```{r/python}
#| label: fig-label
#| fig-cap: "An example caption."

<insert plotting code here>
```
````

The biggest pitfall in EDA/visualizations: **Overplotting**

Bad



Better



Strategies to avoid overplotting

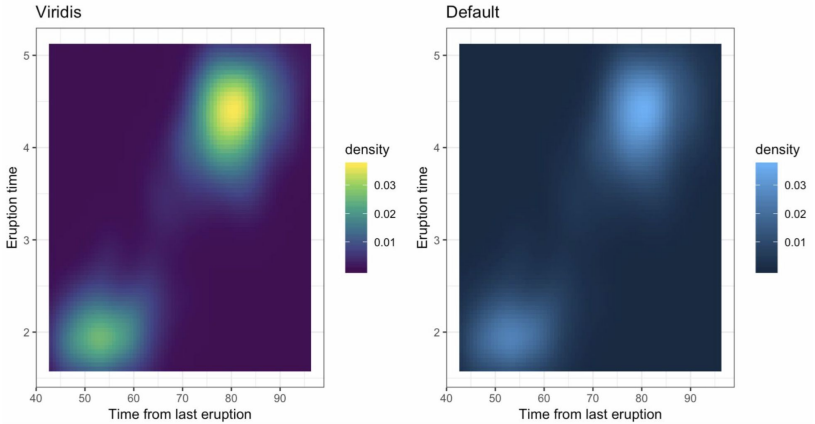
+ Use smaller point sizes

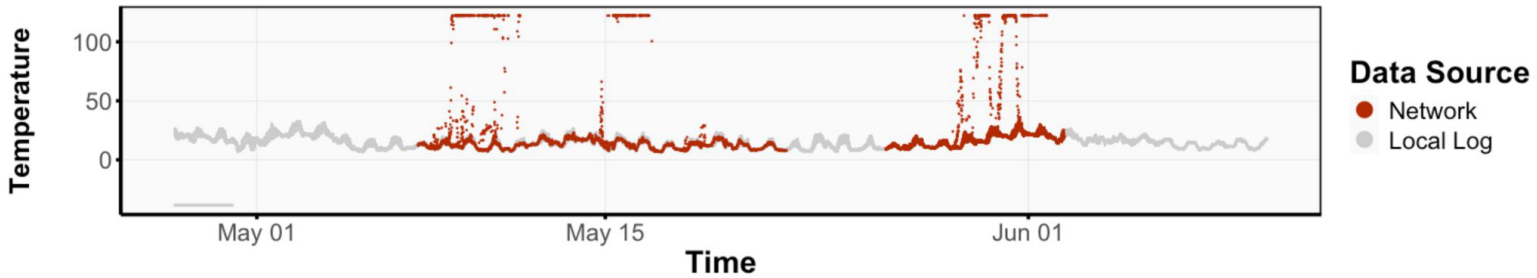
  ggplot2: geom_point(size = ...)
  matplotlib: plot.plot(markersize = ...)

+ Use transparency (alpha)
+ Subsample data points
+ Remember to focus on a specific message

Color matters!

- Color choices can affect the way we perceive the plot





Color matters!

When choosing colors, be considerate of...

- Colors have inherent connotations
  - Red = bad or temperature hot
  - Green = good
  - Gray = ignored
  - Black = bold/draws attention to
- Discrete versus continuous color scales
  - Different shades of the same color suggest relatedness
- Color-blind friendly
  - ~10% of all men are red-green colorblind

Resources for choosing colors

Color scheme generator: https://coolors.co/

HTML color codes: https://htmlcolorcodes.com/

Encycolorpedia: https://encycolorpedia.com/

Viridis color palette

If you need inspiration for visualizations...

NY Times Data Visualizations: https://www.nytimes.com/column/whats-going-on-in-this-graph

+ Great for finding new color schemes

Storytelling with Data: https://www.storytellingwithdata.com/

Sprucing up your visualizations with interactivity

- Shiny: https://shiny.posit.co/
  - R Tutorial: https://shiny.posit.co/r/getstarted/shiny-basics/lesson1/
  - Python Tutorial: https://shiny.posit.co/py/docs/overview.html
- Plotly
  - R: https://plotly.com/r/ (also see `plotly::ggplotly()`)
  - Python: https://plotly.com/python/

Exploratory Data Analysis (EDA) tips in a nutshell

- Start with your domain problem
- Explore with "scratchwork" EDA: quantity over quality
- Once you have identified your main finding, think before you plot
  - Your plot should clearly communicate your **message**
  - Your main EDA plot should not be a "data cleaning" plot
- Plot type should be an intentional choice
  - Line, scatter, bar, heatmap, ...
- Aesthetics matter
  - Color
  - Point size
  - Transparency
  - Labels
  - Theme
  - Be wary of overplotting
- **Take your time**

Recap + Next Time

Recap

+ **Exploratory data analysis** is a great way to get a feel for the data. [chapter 5 from VDS textbook]
  - "Scratchwork" EDA (internal): quantity over quality
  - "Publication-quality" EDA (public): quality over quantity
    - Think then plot. The simpler, the better.

Don't forget

+ Lab 1 due **Friday 6pm** submitted to GitHub

Next Time

+ Beginning of unsupervised learning unit

https://forms.gle/h4qPkY4DFcpv7nL6A