



ENVIRONMENTAL DATA ANALYTICS: M5 – DATA VISUALIZATION

Spring 2024

Nicholas School of the Environment - Duke University

Catch up

□ Debugging:

- ▣ Goal: find where the error might occur...
- ▣ Start simple, add complexity in increments...
- ▣ Check outputs for logical consistency...

□ Knit issues:

```
Calls: <Anonymous> ... eval_with_user_handlers -> eval -> eval -> install.packages -> contrib.url
```

- ▣ Check paths (knit directory = project working directory)
- ▣ Restart R (clear environment) and run entire Rmd file...

M5.1 - Data Visualization

- Approaches to visualizations ([link](#))
- The `ggplot2` package
- `ggplot` structure: layers = `geoms`
- Aesthetics, axes, colors, shapes, facets, axis limits, reference lines
- Plot types...

geoms



- geom_bar
- geom_histogram
- geom_freqpoly
- geom_boxplot
- geom_violin
- geom_dotplot
- geom_density_ridges
- geom_point
- geom_errorbar
- geom_smooth
- geom_line
- geom_area
- geom_abline (plus geom_hline and geom_vline)
- geom_text

aesthetics

- color
- fill
- shape
- size
- transparency

scales

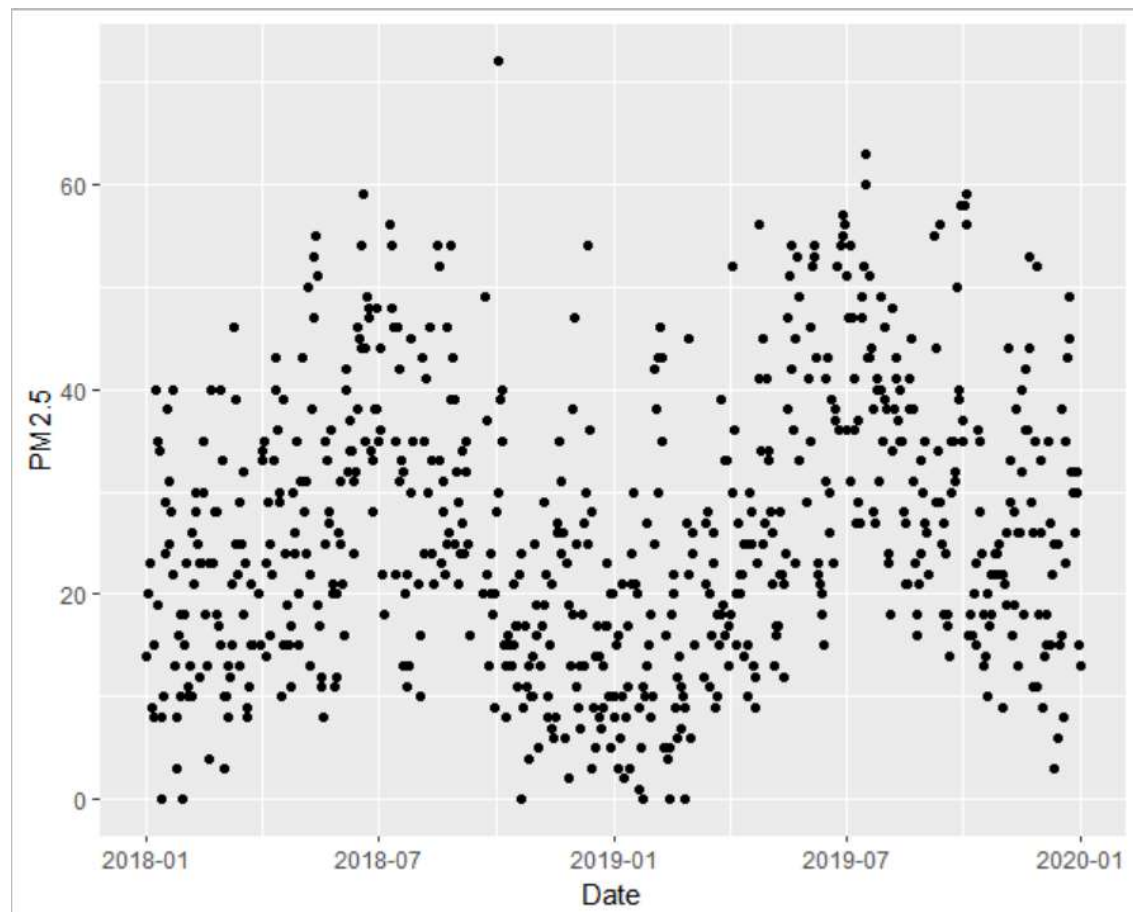
A **scale** is a function from a region in data space (the domain of the scale) to a region in aesthetic space (the range of the scale).

<https://ggplot2-book.org/scales-guides.html>

`scale_<aesthetic>_<name>`

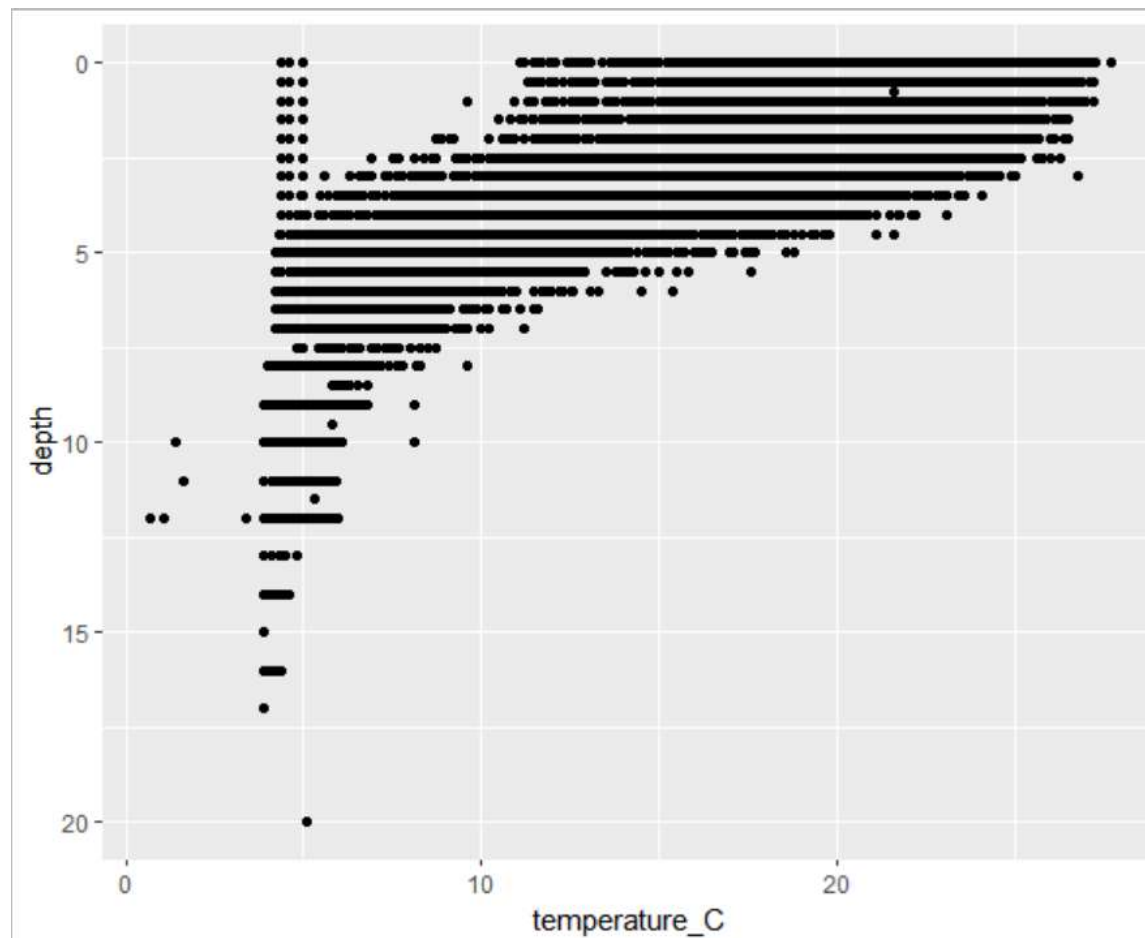
Plot types: point or line plots

- Continuous over time (or distance, or other ordered...)



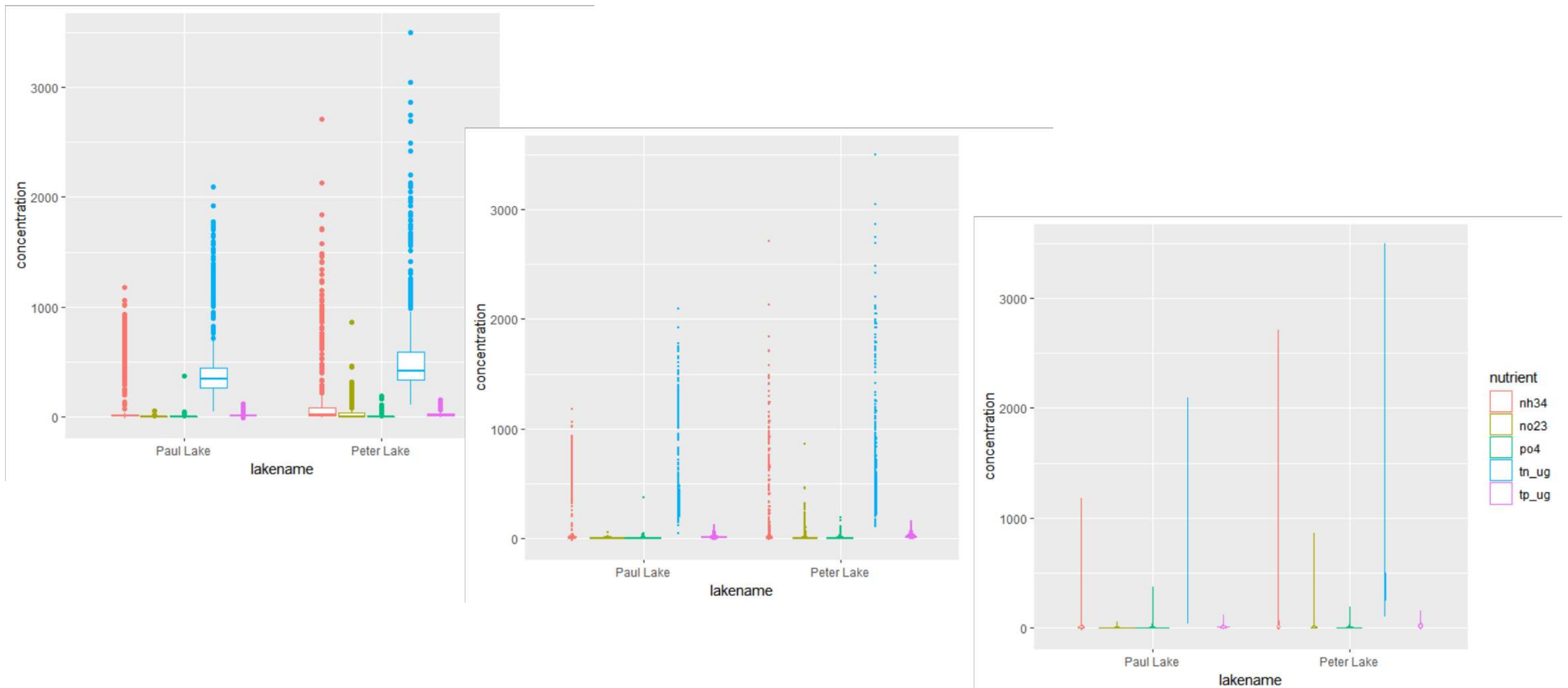
Plot types: scatterplots

- Relationship between two variables: Scatterplot



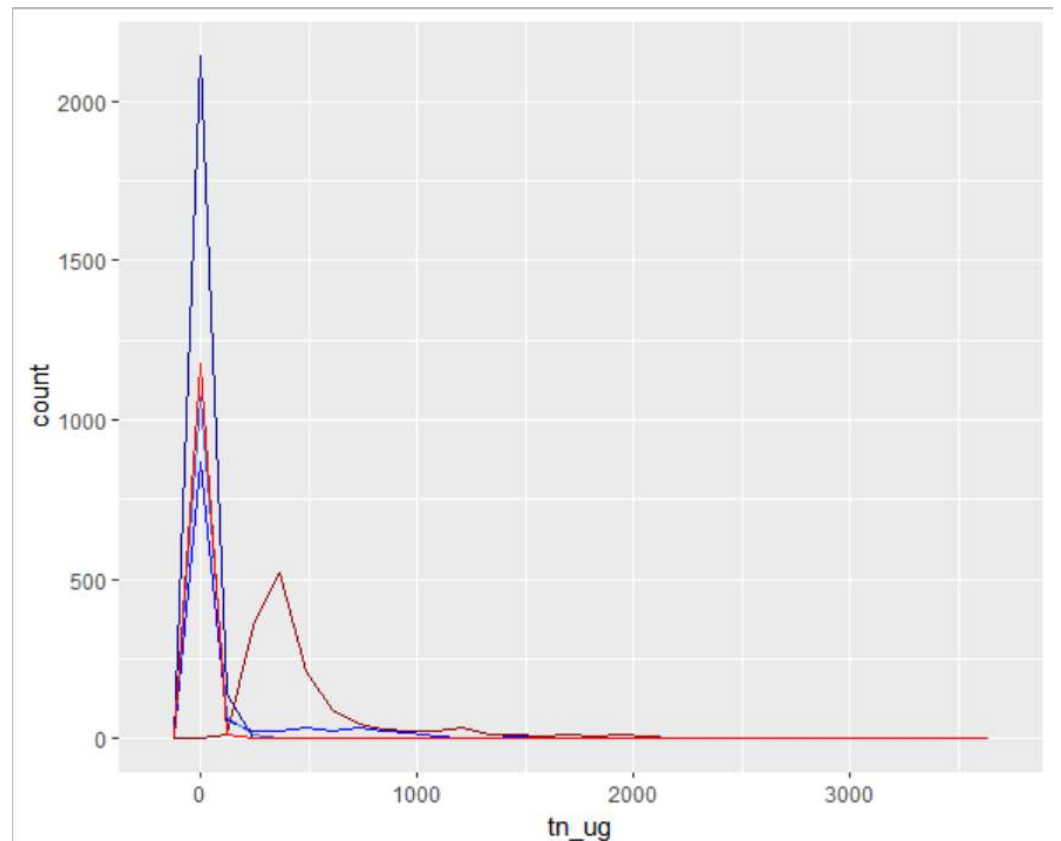
Plot types: box, dot, & violin plots

- Continuous vs categorical
 - ▣ Issue with error bars on box plots...



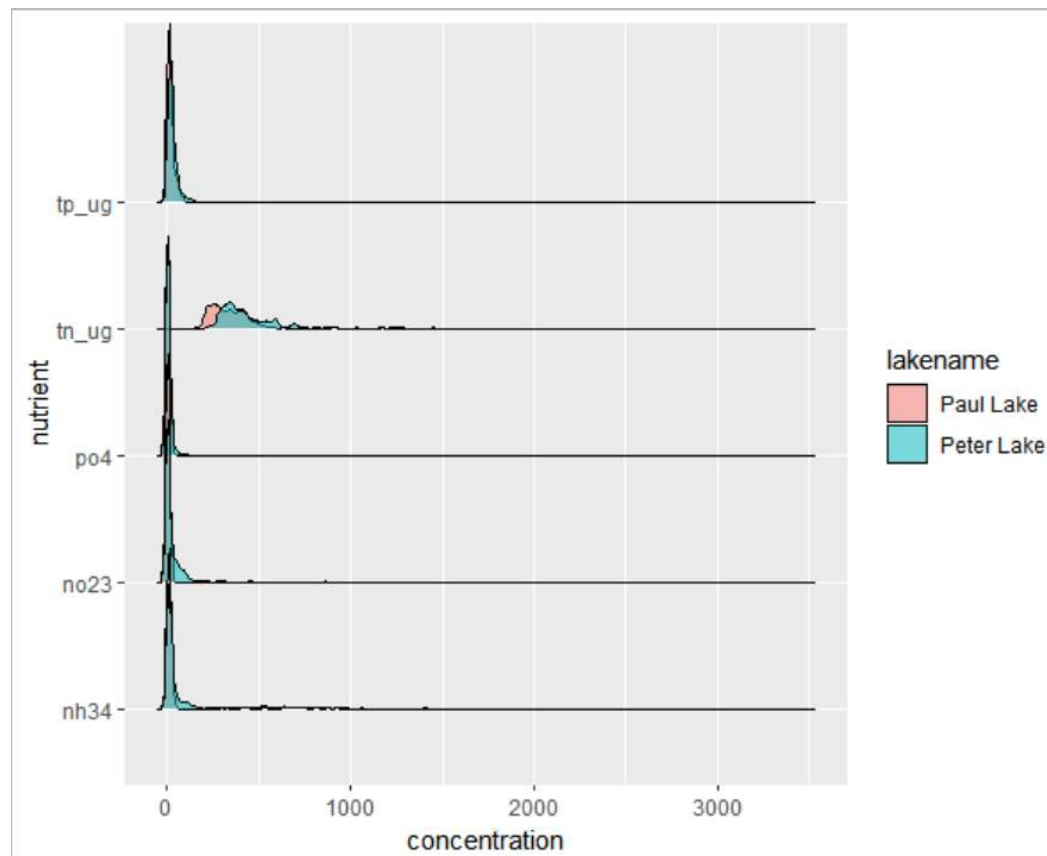
Plot types: Frequency polygons

- Continuous vs categorical
 - ▣ Tidy vs gathered
 - ▣ “Spaghetti” problem



Plot types: Ridgeline plots

- Continuous vs categorical



M5.2 – Formatting plots

- Themes
- Custom layers
- Color palettes
- Cowplots package
- Saving plots

ggplot “themes”



- Built-in themes (`> ?theme_grey`)
- What is controlled in a theme?
- How is a theme called?
- How to create a custom theme?
- How to set a default theme for all plots?

- Theme elements: `> ?theme`

Other plot components



- `geom_abline()` | `geom_vline()` | `geom_hline()`
- `geom_point()`
- `geom_text()`
- `scale_x_date()` | `scale_y_date()`
- `xlab()` | `ylab()` | `labs(...)`

Color palettes



- Why, when, how...
- Color packages: **RcolorBrewer viridis colormap**
 - ▣ Exploring
- Plot “scale...” options
 - ▣ **scale_shape_manual()**
 - ▣ **scale_color_...()**
- Types of color ramps
 - ▣ Continuous, categorical, divergent

Arranging plots: **cowplot**



- `plot_grid()`
 - ▣ Set plots & # of rows & relative heights

Saving plots: ggsave

- filename and relative path, with file extension and in quotes (required)
- plot object (required)
- width, height, units
- resolution (dpi)

M4.3 – Data Visualization III (lab)

Expressions



- `geom_text()`
 - The `paste` and `paste0` commands
 - `MathJax`

A note on **factors**...



- *Factors..*
 - ...are useful for analyzing/visualizing categorical data
 - ...have *levels*
 - ...can have *labels* too

- *Plot the number of lake measurements by month...*
 - What kind of variable is `month` in the dataframe?
 - How many unique values in this column?
 - Why might this pose a problem?
 - What can we do?

Solution

```
#what kind of values are months?
class(PeterPaul.chem.nutrients$month)

#List unique values
unique(PeterPaul.chem.nutrients$month)

#Plot how many Lake observations occur each month
ggplot(PeterPaul.chem.nutrients, aes(x=month)) +
  geom_bar()

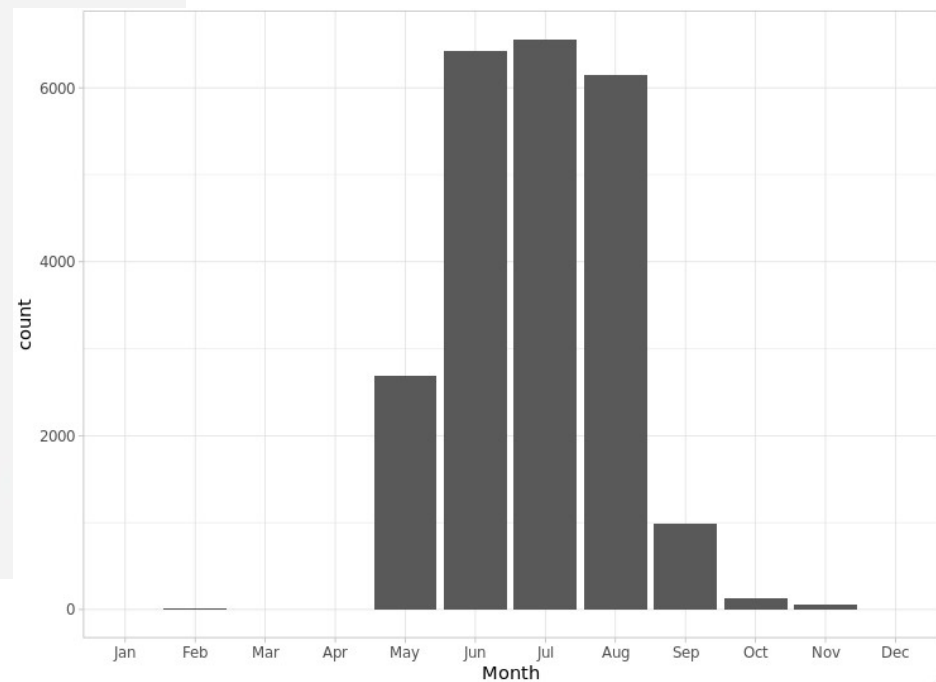
#Convert to a factor -- with 12 levels, labelled with month names
factor(PeterPaul.chem.nutrients$month,
       levels = 1:12,
       labels = month.abb)

#Try again: Plot how many Lake observations occur each month
ggplot(PeterPaul.chem.nutrients, aes(x=factor(month, levels=1:12, labels=month.abb))) +
  geom_bar()

#Fix missing months
ggplot(PeterPaul.chem.nutrients, aes(x=factor(month, levels=1:12, labels=month.abb))) +
  geom_bar() +
  scale_x_discrete(drop=FALSE)
```

A note on factors... *Solution*

```
#Tidy up the code
the_plot <- PeterPaul.chem.nutrients %>%
  ggplot(
    aes(x=factor(
      month,
      levels=1:12,
      labels=month.abb)
    )
  ) +
  geom_bar() +
  scale_x_discrete(
    name="Month",
    drop=FALSE
  )
#Show the plot, in the light theme
the_plot + theme_light()
```



More on *themes*...

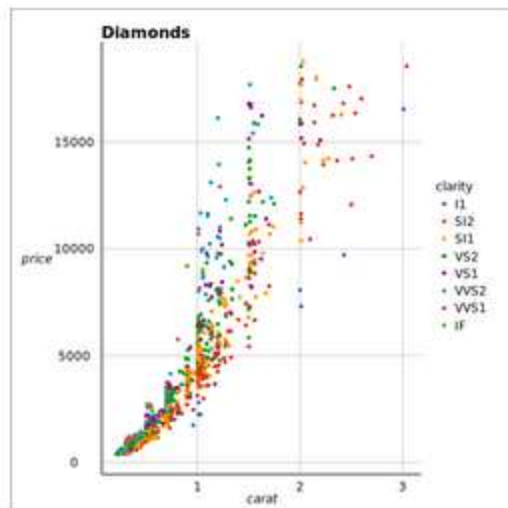
Themes control the following elements

- **Plot background:**
The background color or fill pattern of the plot area.
- **Plot title:**
The size, font, and position of the plot title.
- **Axis labels:**
The font, size, and position of the x-axis and y-axis labels.
- **Axis ticks and grid lines:**
Color, size, & position of the tick marks & grid lines on the axes.
- **Legend:**
The font, size, and position of the legend.

ggthemes()

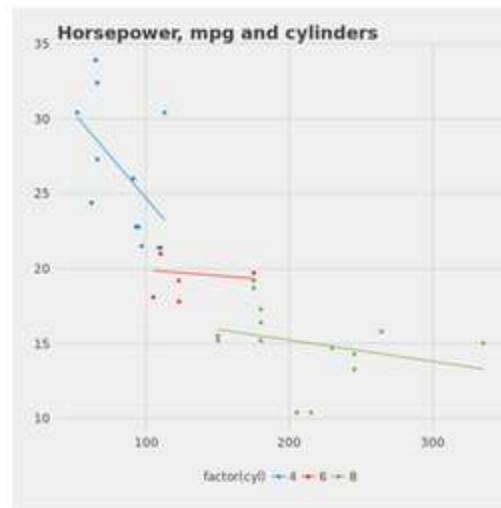
Adds custom themes and scales

□ [Link to examples](#)



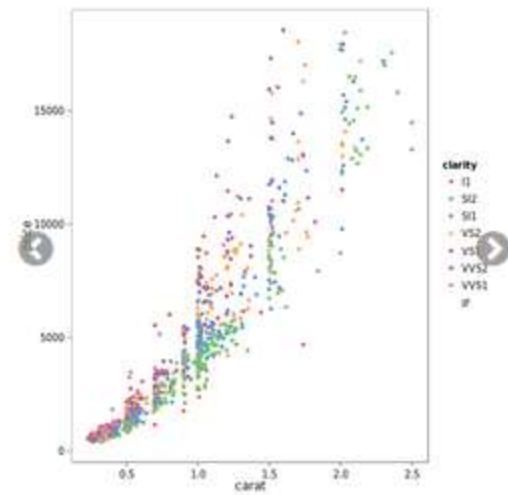
theme_gdocs

Theme with Google Docs Chart defaults



theme_fivethirtyeight

Theme inspired by fivethirtyeight.com plots

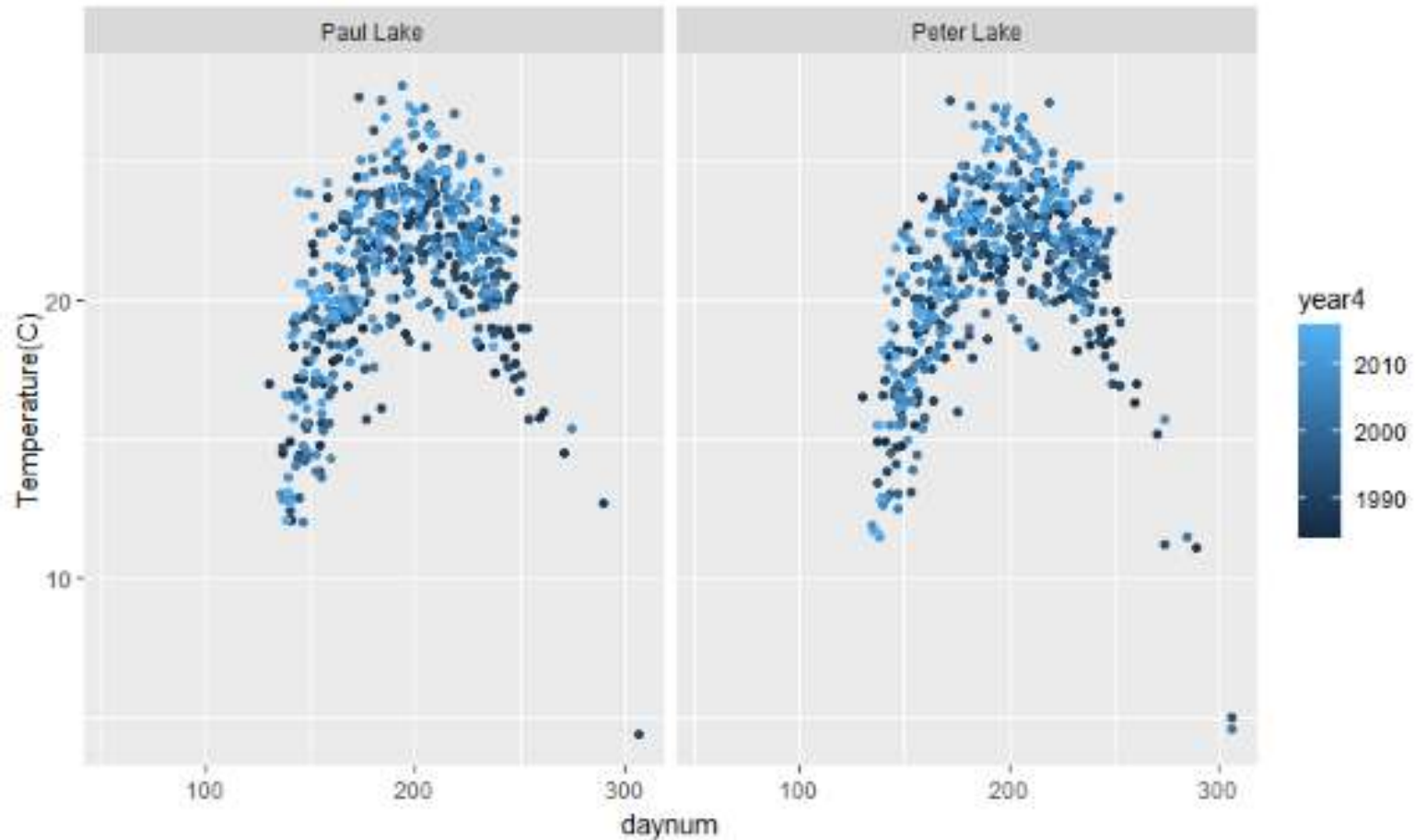


theme_few

Theme based on Few's "Practical Rules for Using Color in Charts"

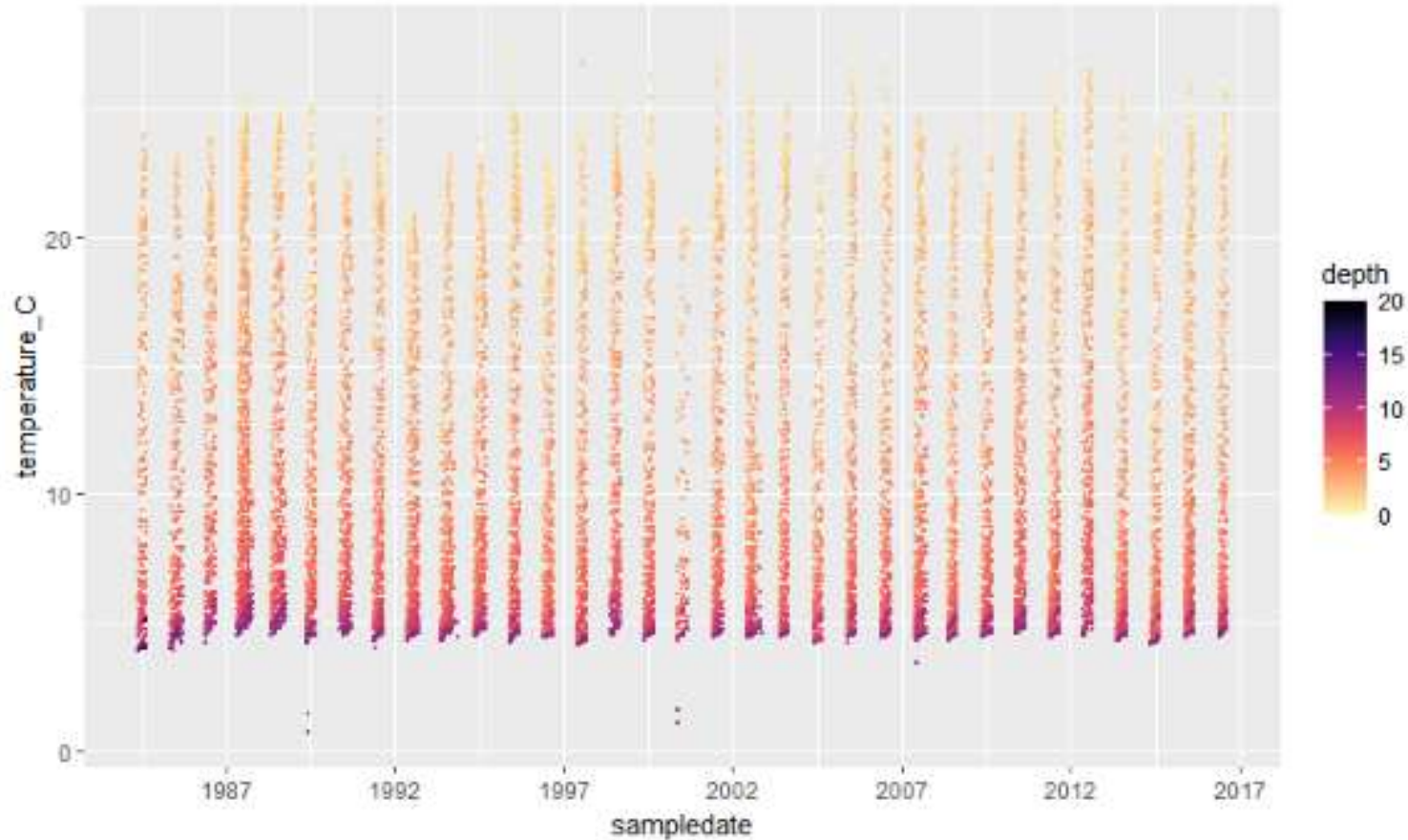
Exercise 1

```
# 1.  
# Plot surface temperatures by day of year.  
# Color your points by year, and facet by lake in two rows.  
# Change the ylab name
```



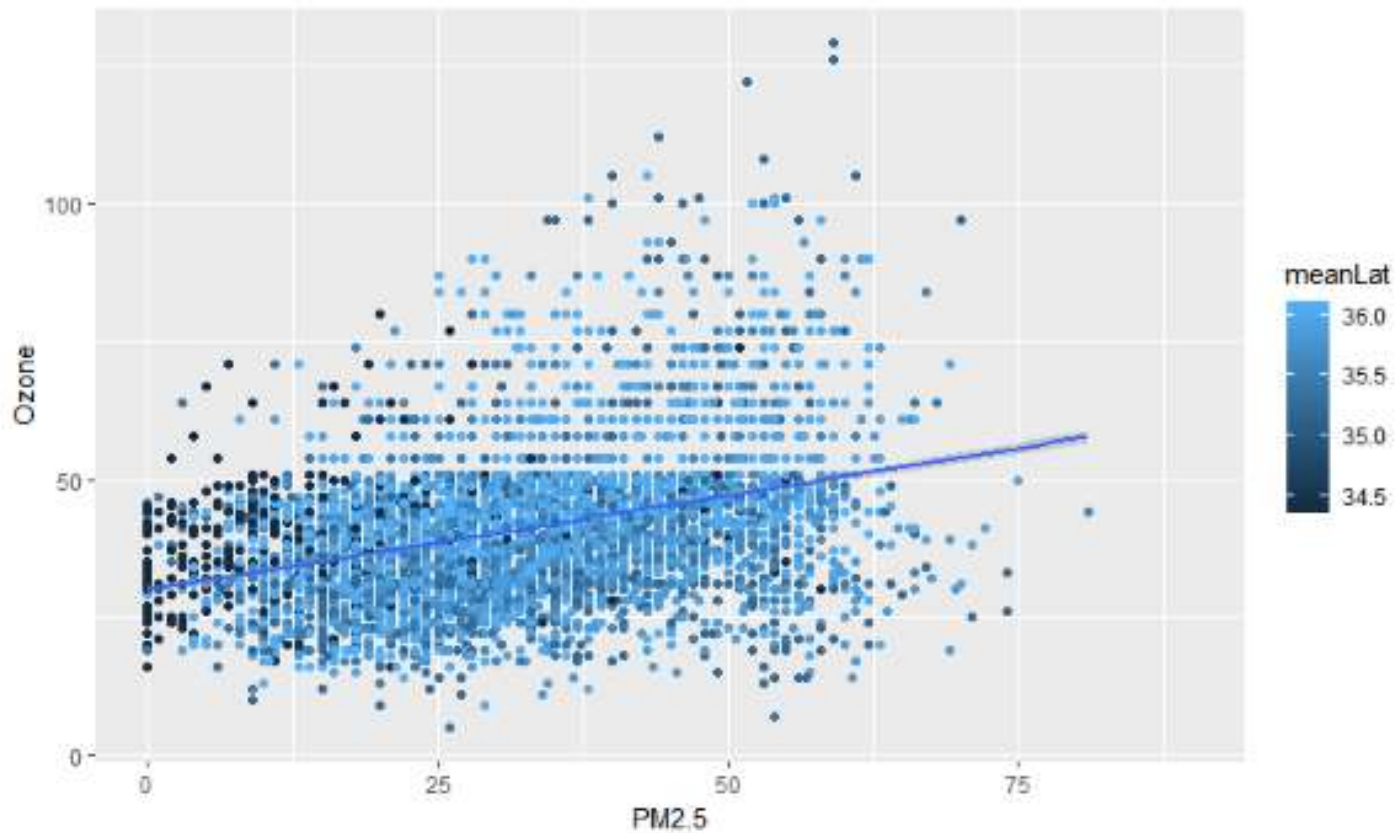
Exercise 2

```
#2.  
# Plot temperature by date. Color your points by depth.  
# Change the size of your point to 0.5  
# Change the color palette to magma and play with direction (+- 1), which one makes more sense?  
# Change x axis to include marker/labels every 5 years
```



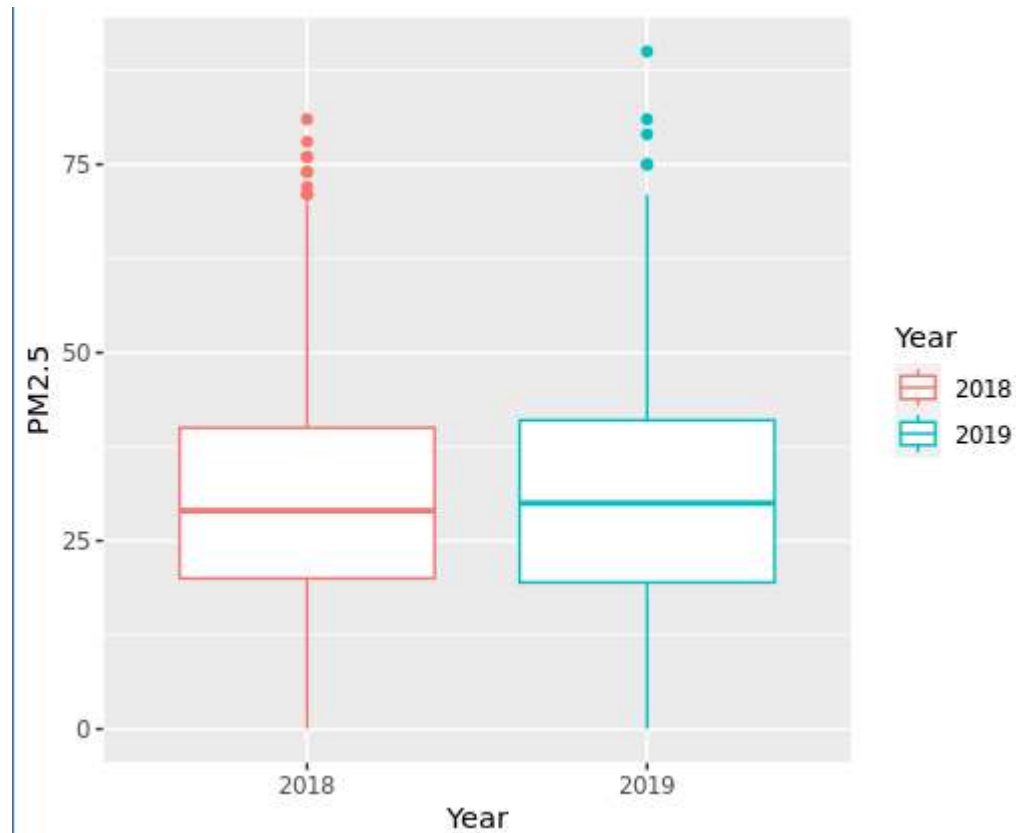
Exercise 3

```
# 3.  
# Plot AQI values for ozone by PM2.5, colored by latitude  
# Make the points 50 % transparent  
# Add a line of best fit for the linear regression of these variables.
```



Exercise 4

```
# 4.  
# Create several types of plots depicting PM2.5, divided by year.  
# Choose which plot displays the data best and justify your choice.
```



Viz challenge?

