Exploratory data analysis with ggplot2

2031 - Statistical Computing

Sam Clifford

Session 8

Why do we visualise?

- Want to tell a story in an engaging way
- Want to explore relationships in data with view to
    - checking our assumptions about the data
    - model formulation
    - designing further experiments
- Reader should be able to understand what the graph means and not be
    - misled into thinking something that is untrue
    - distracted from the main point

Tufte's principles

Tufte [1983] and Pantoliano [2012]

- Show the data
- Provide clarity
- Allow comparison where appropriate
    - use aesthetics to draw attention to important details
    - make clear that data has multiple levels of structure
- Produce graphs with high data density
    - make every drop of ink count
    - careful use of whitespace
- Avoid excessive and unnecessary use of graphical effects

Building plots

R package `ggplot2` uses a grammar of graphics [Wickham, 2010, RStudio, 2012]

- map variables in data frame to aesthetic options in the plot
- choose a geometry for how to display these variables
- adjustments to axis scales
- adjustments to colors, themes, etc.
- adding extra commands in a ‘do this, then do this’ manner

Building plots

How do we structure a call to `ggplot` to make a plot?

- load ggplot2 package
- Specify we want a `ggplot` object and which data frame we're going to use,
- Set **aesthetic options** to tell R which variables to map to the x and y axes of the plot
- State geometry we're using to show variables

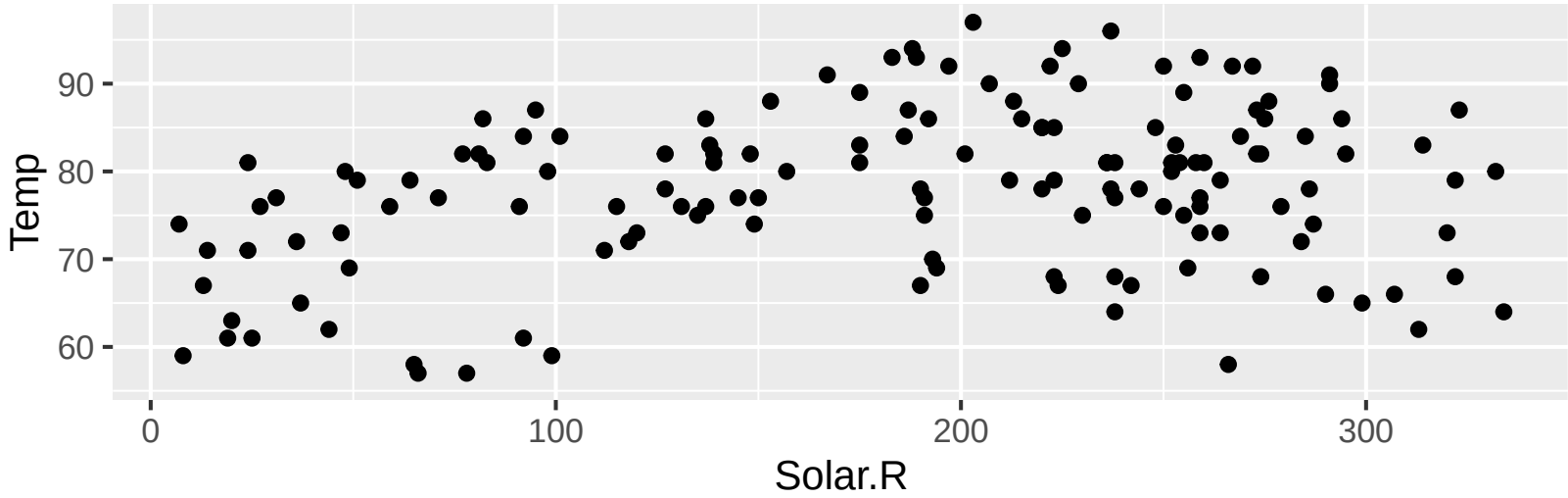
```
library(ggplot2)
ggplot(data = my.data.frame,
       aes(x = my.x.variable,
           y = my.y.variable)) +
  geom_point()
```

Building plots

- For example, consider a scatter plot of daily maximum temperature varying with solar radiation in New York City 1973
- Each row in data has a pair of values (x, y), shown as a point

```
data(airquality)
solar_temp_plot <- ggplot(data = airquality,
       aes(x = Solar.R, y = Temp)) +
  geom_point()

solar_temp_plot
```

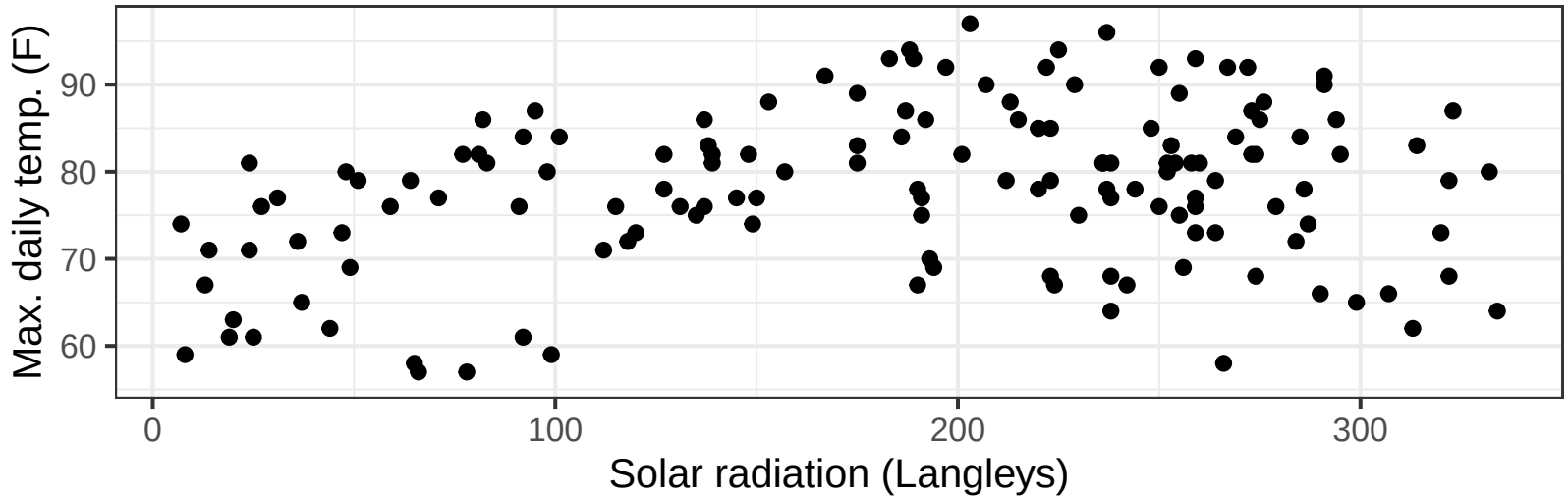


Scatter plot

- We can add some human-friendly labels and change the theme

```
solar_temp_plot <- solar_temp_plot +
  labs(x = 'Solar radiation (Langleys)',
       y = 'Max. daily temp. (F)') +
  theme_bw()

solar_temp_plot
```



Line plot

- Similar to scatter plot, but joins pairs of values
- Useful when showing how something changes over time
- Use only when (x, y) are ordered pairs of numeric values, e.g. x is time or date
- For this reason, often referred to as **time series plot**

Line plot

- Show the Ozone concentrations over time

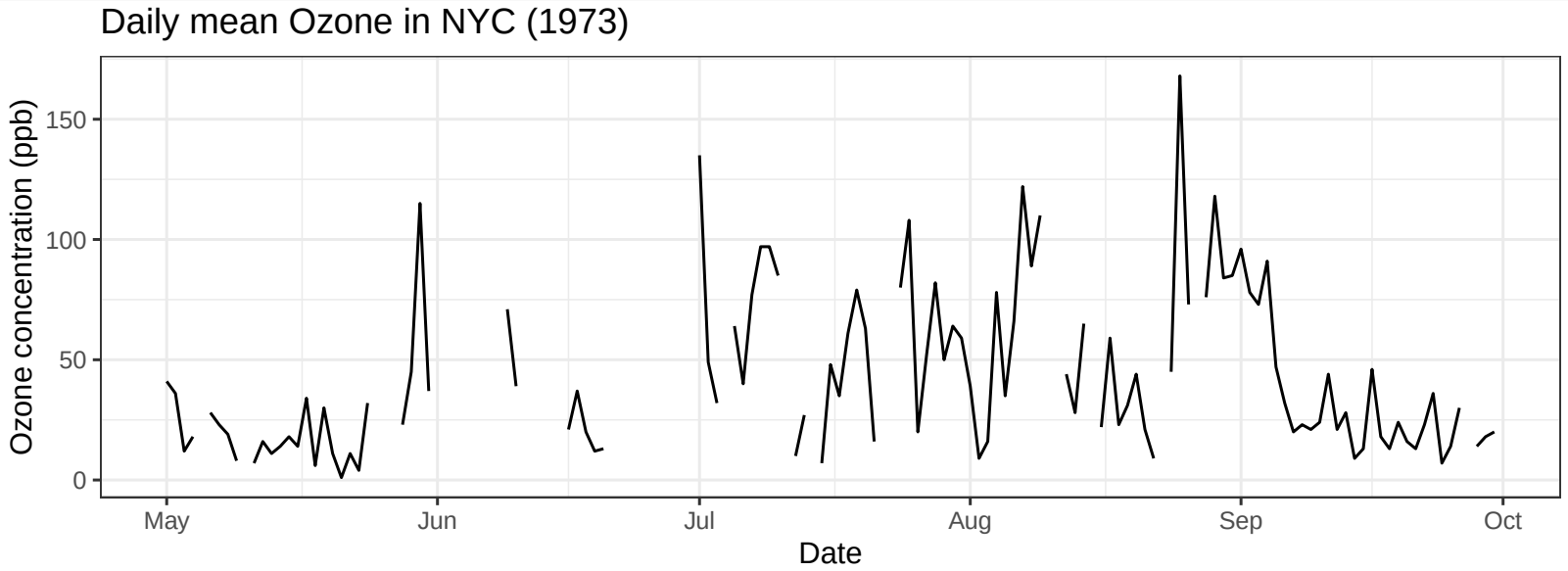
```
# make the date column

airquality <- mutate(airquality,
                     Date = as.Date(paste('1973',
                                          Month,
                                          Day, sep

airquality_plot <-
  ggplot(data=airquality,
         aes(x=Date, y=Ozone)) +
  geom_line() + theme_bw() +
  labs(y     = 'Ozone concentration (ppb)',
       title = 'Daily mean Ozone in NYC (1973)')
```

Line plot

```
airquality_plot
```

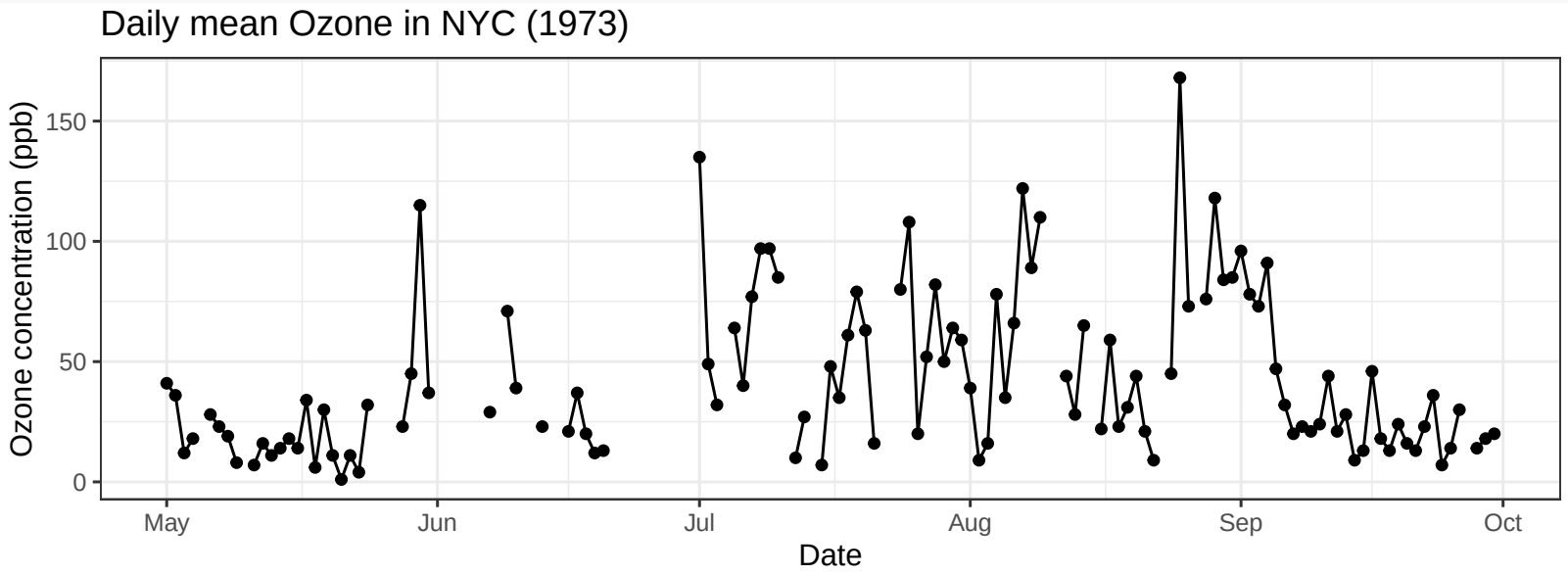


- We see here that there are gaps in the line due to missing data
- If we have an observation whose neighbours are both NA values it can't be plotted with a line

Line plot

- Can layer multiple geometries for same aesthetic mapping

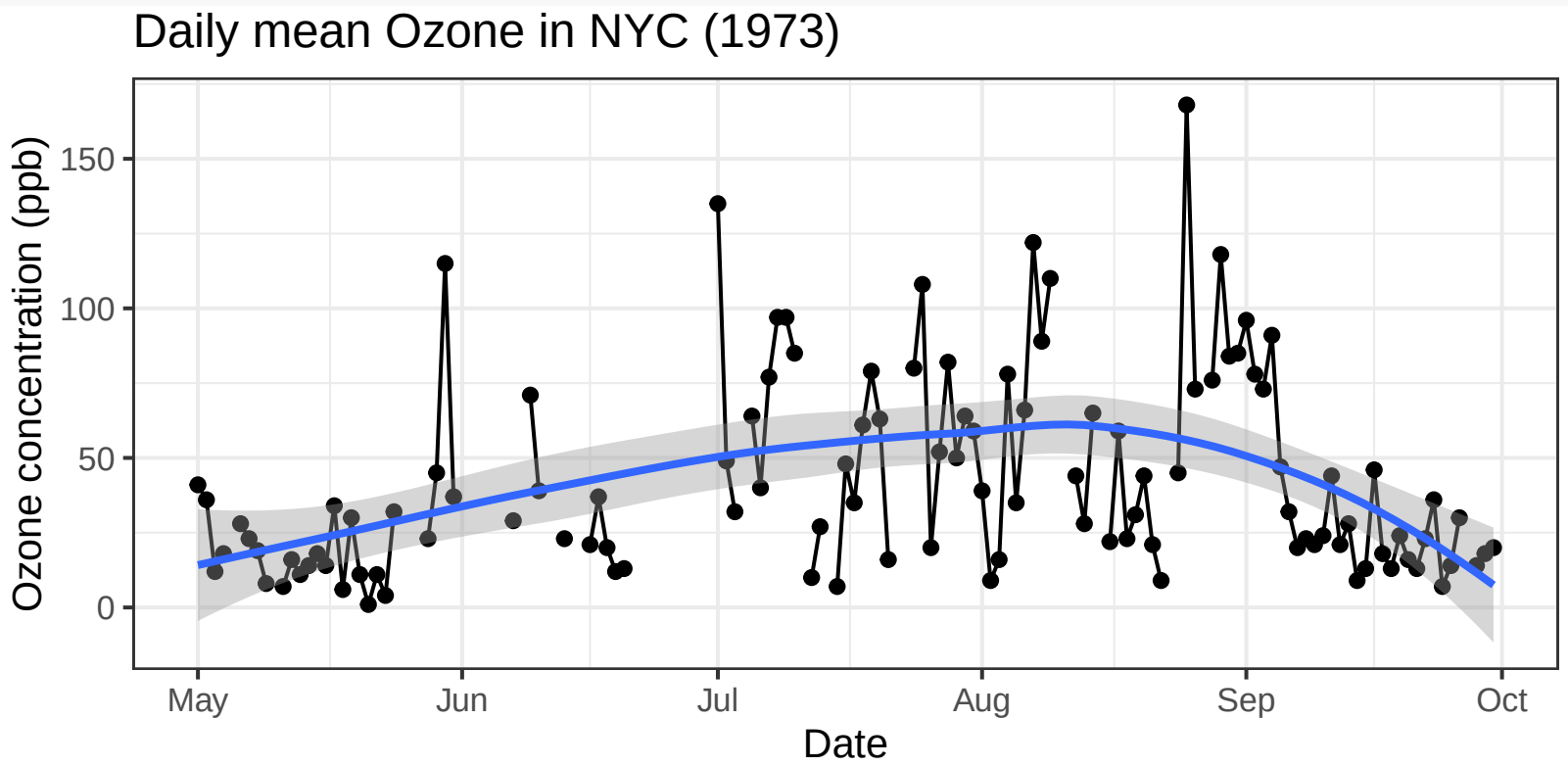
```
airquality_plot + geom_point()
```



Smooth plot

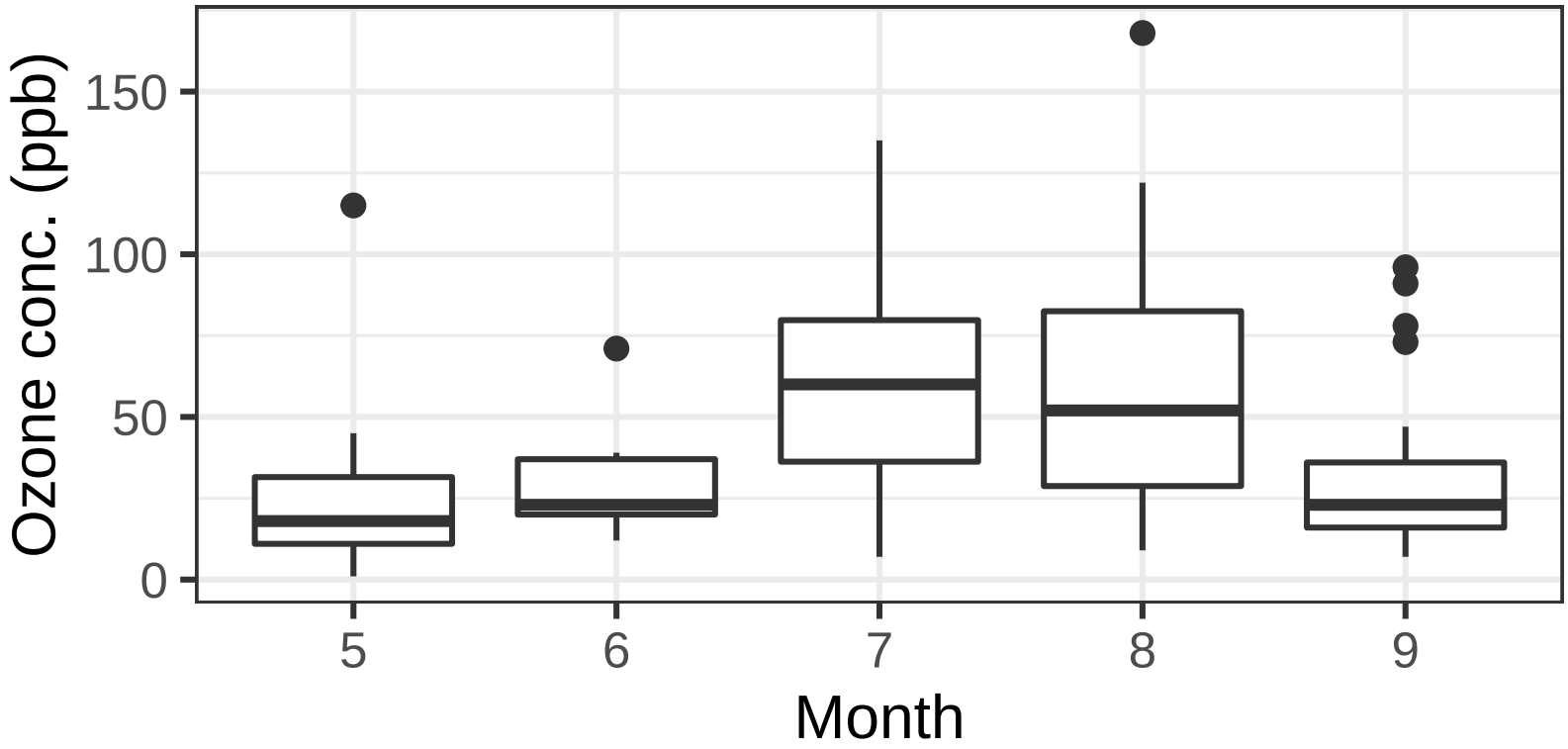
- Often too much data in a scatter plot to see pattern
- Maybe we want to show the reader the trend in the data
- `geom_smooth()` generates a **scatterplot smoother** that shows the overall relationship between y and x

```
airquality_plot + geom_point() + geom_smooth()
```



Boxplot

```
ggplot(data = airquality, aes(x = factor(Month), y = Ozone)) +
  geom_boxplot() + theme_bw() +
  labs(y = 'Ozone conc. (ppb)', x = 'Month')
```



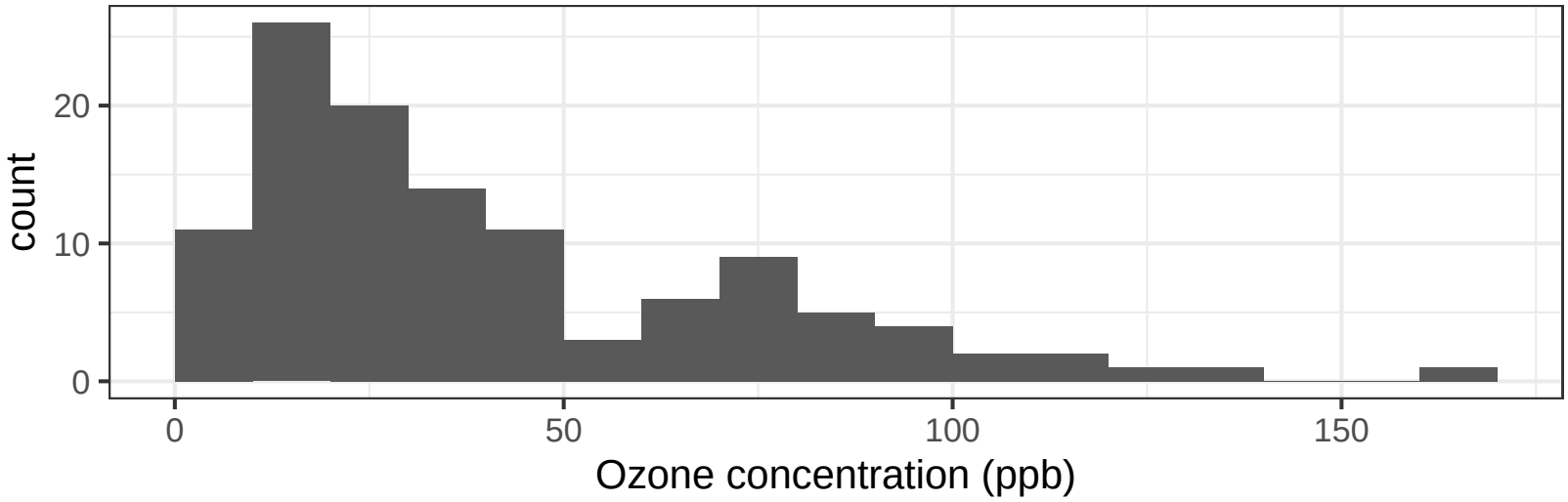
- **outliers** shown as dots ($> 1.5\,\mathrm{IQR}$)

Histograms

- univariate graphical summary needs only one aesthetic, x
- e.g. a histogram of Ozone concentrations

```
ozone_hist <-
  ggplot(data = airquality, aes(x = Ozone)) +
  geom_histogram(binwidth = 10, boundary = 0) +
  labs(x = 'Ozone concentration (ppb)') +
  theme_bw()

ozone_hist
```



Aesthetics

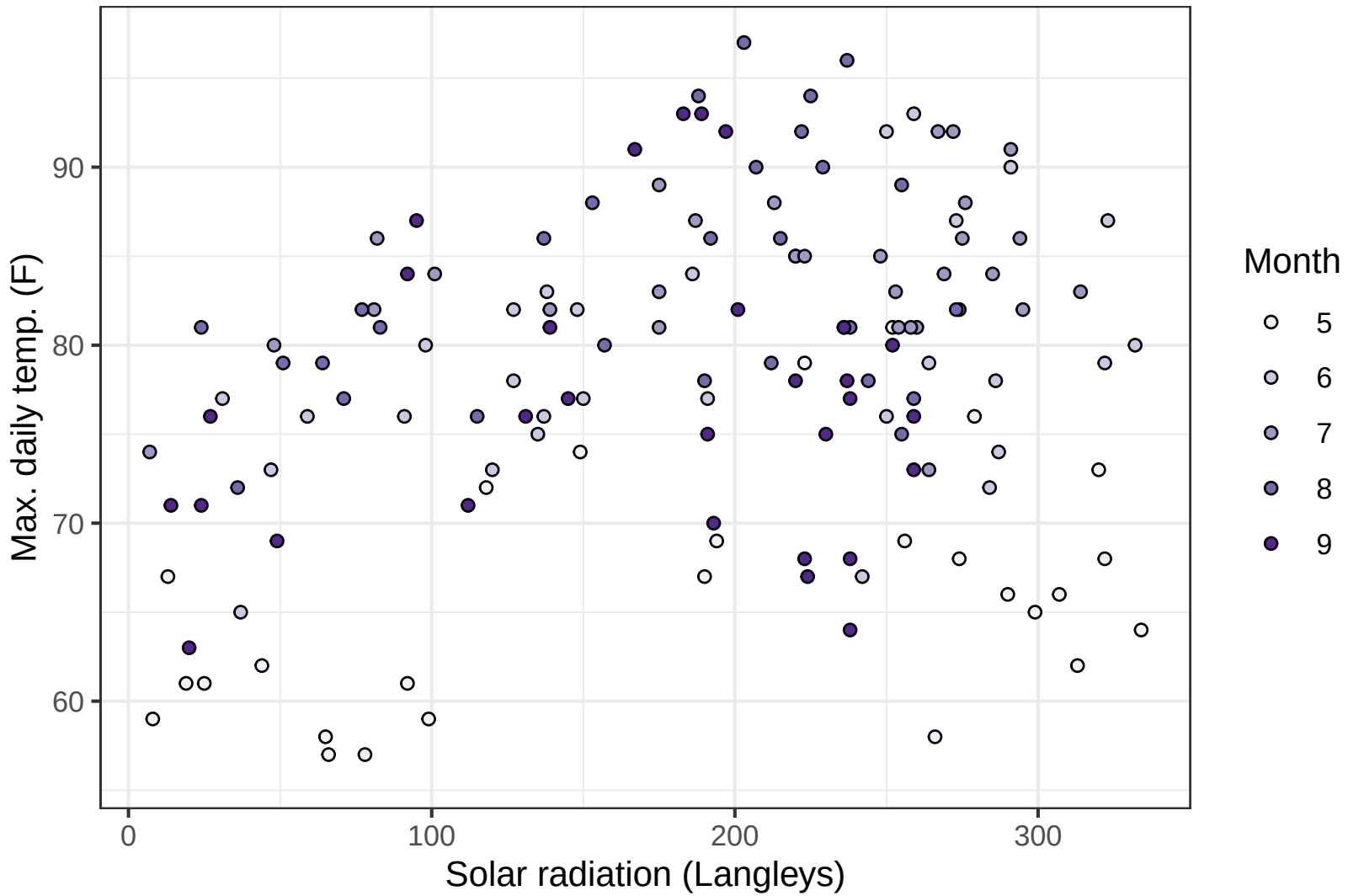
- We've seen the x and y positions so far
- We can also map the following options to variables

Size	of point or thickness of boundary
Shape	of points
Colour	of boundary
Alpha	transparency
Fill	colour of internals of geometry
Group	to repeat geometry for each level

- We can also put these (except group) *outside* `aes()` to fix the value for all parts of that geometry
- Any aesthetics specified inside `ggplot()` will be inherited by all geometries for that plot

Aesthetics

```
data(airquality)
solar_temp_plot_colored <-
  ggplot(data = airquality,
         aes(x = Solar.R, y = Temp)) +
  geom_point(aes(fill = factor(Month)),
             shape = 21,
             color = 'black')  +
  labs(x = 'Solar radiation (Langleys)',
       y = 'Max. daily temp. (F)') +
  theme_bw() +
  scale_fill_brewer(palette = "Purples",
                    name   = 'Month')
```


Max. daily temp. (F)
90
80
70
60
0
100
200
300
Solar radiation (Langleys)
Month
5
6
7
8
9

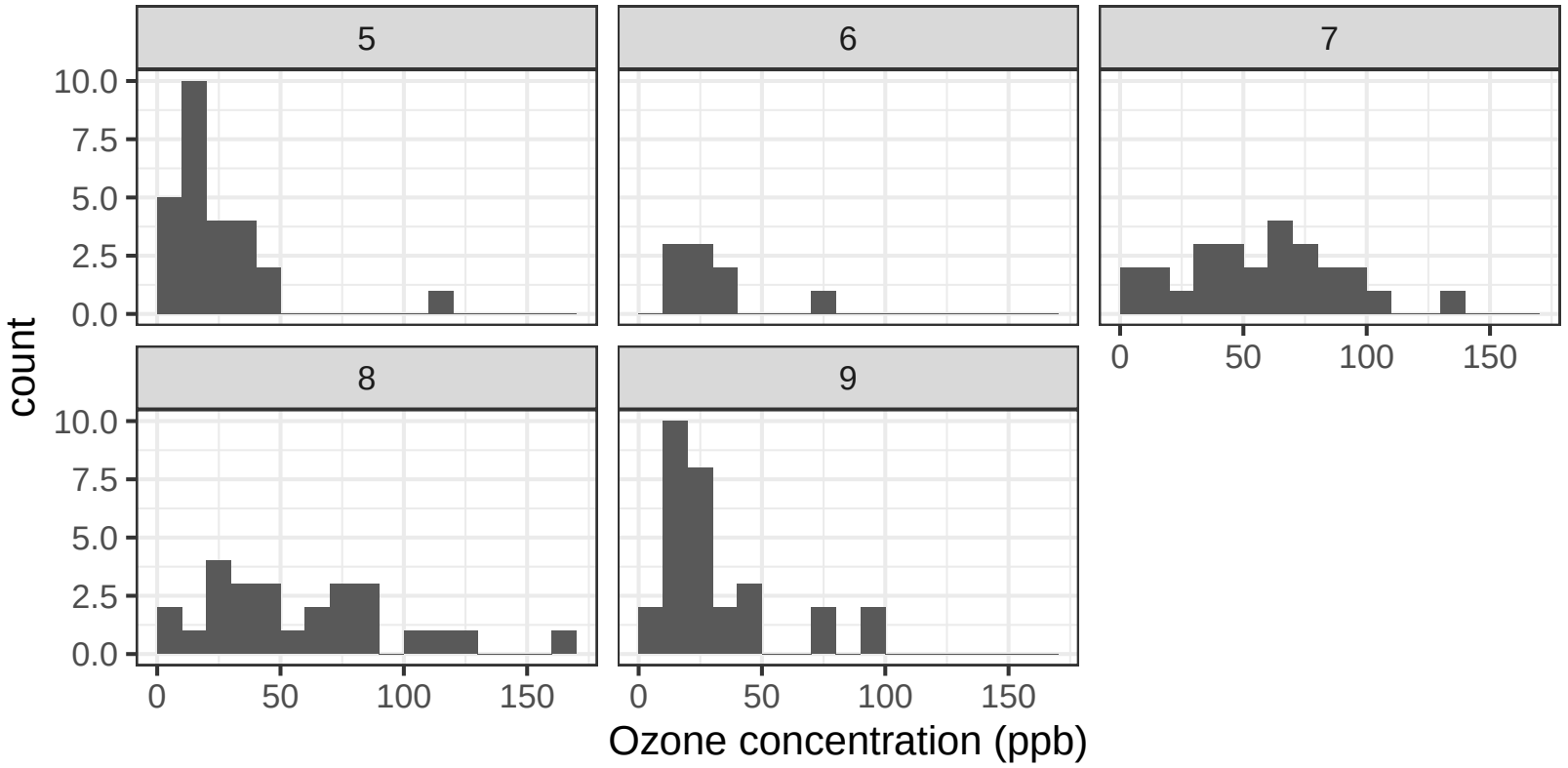
Small multiples

- Group a plot by some categorical variable
- Repeat a basic graph for groups in the data
    - air quality data has information about, e.g. months
- Can view 3-5 dimensions in the data on a 2D page
    - Often a better alternative to 3D, since it doesn't distort comparisons
    - Inner axes relate to the smallest X-Y plots
    - Outer axes relate to the grouping variables
- Avoids using loops

Small multiples

- We can repeat the histogram plot for each value of `Month`, one per facet

```
ozone_hist + facet_wrap( ~ Month)
```



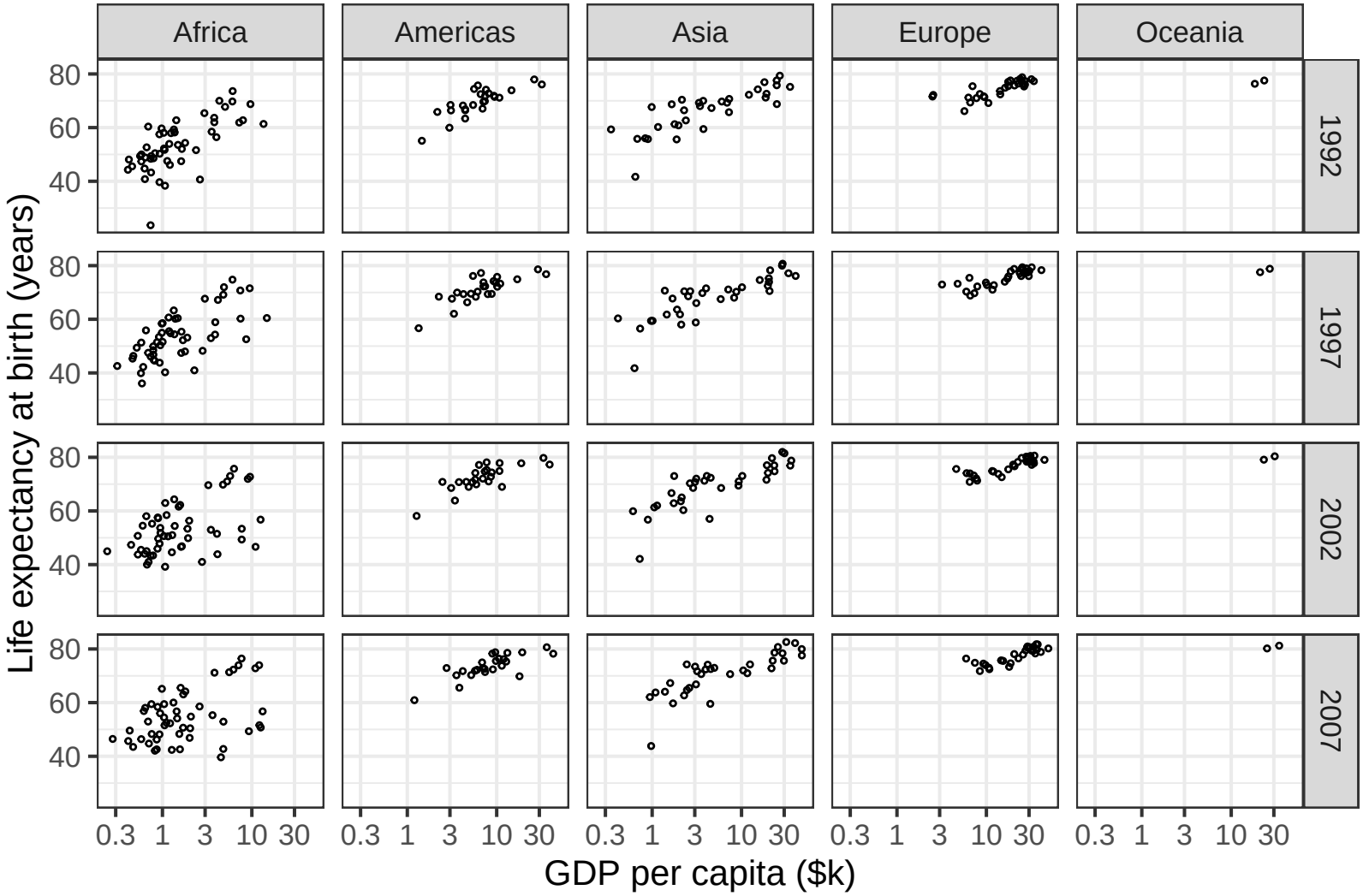
Small multiples

We can also use `facet_grid()` to repeat the aesthetic and geometries for specified `rows` and `cols` variables

```
library(gapminder)
data(gapminder)

gapminder_plot <-
  ggplot(data = subset(gapminder, year >= 1992),
         aes(x = gdpPercap/1e3,
             y = lifeExp)) +
    geom_point(shape = 1, size = 0.5) +
    facet_grid(rows = vars(year),
               cols = vars(continent)) +
    scale_x_log10(labels = ~sprintf("%g", .)) +
  xlab("GDP per capita ($k)") +
  ylab("Life expectancy at birth (years)") +
  theme_bw() +
  theme(panel.grid.minor.x = element_blank())
```

Small multiples



Summary

- We make graphs to tell a story with data
- Should draw reader in and explain what they're seeing
- Plots are built from
    - geometric objects
    - axis scales
    - coordinate systems (linear or logarithmic scale, 2D, 3D, etc.)
    - annotations (e.g. heading in small multiples)

Summary

- Successively building a plot with a grammar of graphics allows development of complex plots from simple elements and small changes [Wickham, 2010, RStudio, 2012]
- Choose a plotting geometry that helps tell the story
- Meaningful labels remove ambiguity and confusion
- Be careful not to put too much in

Further reading

- History of visualisation
  - Friendly [2005]
  - Friendly [2006]
- Visualisation to help decision making
  - Tufte [1997]
- ggplot2 resources
  - Wickham [2010]
  - RStudio [2012]
  - Chang [2017]

Winston Chang. *R Graphics Cookbook*. O'Reilly Media, 2nd edition, 2017. ISBN 1449316956. URL http://www.cookbook-r.com/Graphs/.

M. Friendly. Milestones in the history of data visualization: A case study in statistical historiography. In C. Weihs and W. Gaul, editors, *Classification: The Ubiquitous Challenge*, pages 34–52. Springer, New York, 2005. URL http://www.math.yorku.ca/SCS/Papers/gfkl.pdf.

M. Friendly. A brief history of data visualization. In C. Chen, W. Härdle, and A Unwin, editors, *Handbook of Computational Statistics: Data Visualization*, volume III. Springer-Verlag, Heidelberg, 2006. URL http://www.datavis.ca/papers/hbook.pdf.

Mike Pantoliano. Data visualization principles: Lessons from tufte, 2012. URL https://moz.com/blog/data-visualization-principles-lessons-from-tufte.

RStudio. Rstudio cheat sheets, 2012. URL https://www.rstudio.com/resources/cheatsheets/.

Edward R. Tufte. *The Visual Display of Quantitative Information*. Graphics Press, 1983.

Edward R. Tufte. *Visual and statistical thinking*. Graphics Press, 1997. ISBN 9781930824157. URL https://www.edwardtufte.com/tufte/books_textb.

Hadley Wickham. A layered grammar of graphics. *Journal of Computational and Graphical Statistics*, 19(1):3–28, 2010. doi: 10.1198/jcgs.2009.07098.