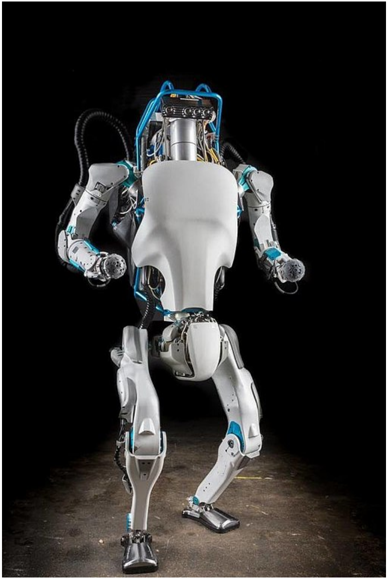
Exploring data

Rote analysis vs. snooping



Spurious correlations

There's a whole website about this

What can you do?

The best you can

- Identify scientific questions

What can you do?

The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis

What can you do?

The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis
- Pre-register studies when possible

What can you do?

The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis
- Pre-register studies when possible
- Keep an exploration and analysis journal

What can you do?

The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis
- Pre-register studies when possible
- Keep an exploration and analysis journal
- Explore predictors and responses separately at first

What can you do?

The best you can

- Identify scientific questions
- Distinguish between exploratory and confirmatory analysis
- Pre-register studies when possible
- Keep an exploration and analysis journal
- Explore predictors and responses separately at first

Outline

Individual variables

Bivariate data

Multiple dimensions

Multiple factors

Individual variables

- Look at location and shape

Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables

Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts

Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts
  - Parametric vs. non-parametric

Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts
  - Parametric vs. non-parametric
  - Exploratory vs. diagnostic

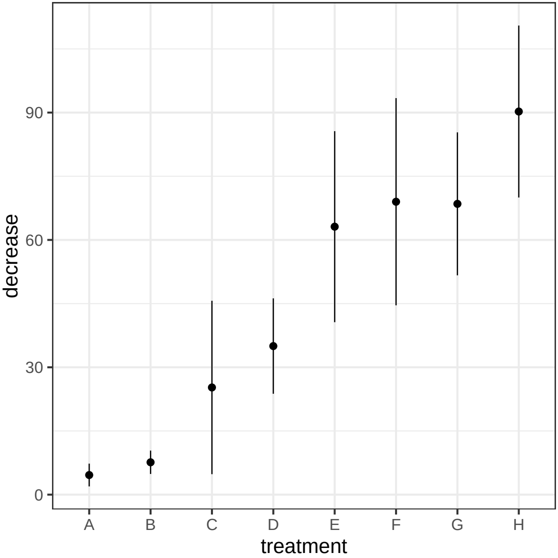
Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts
  - Parametric vs. non-parametric
  - Exploratory vs. diagnostic
  - Data vs. inference

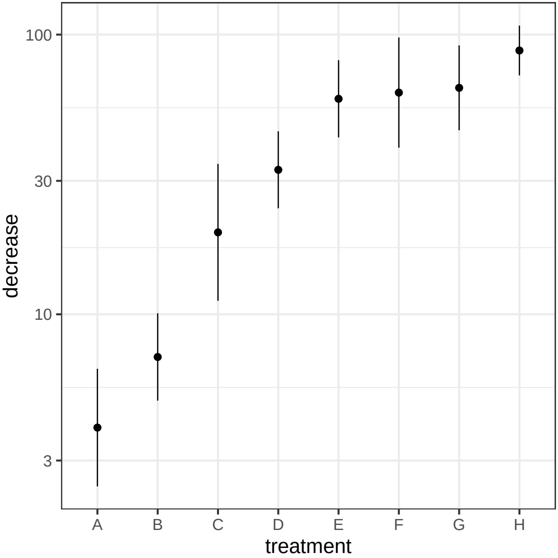
Individual variables

- Look at location and shape
- Maybe with different sets of grouping variables
- Contrasts
  - Parametric vs. non-parametric
  - Exploratory vs. diagnostic
  - Data vs. inference

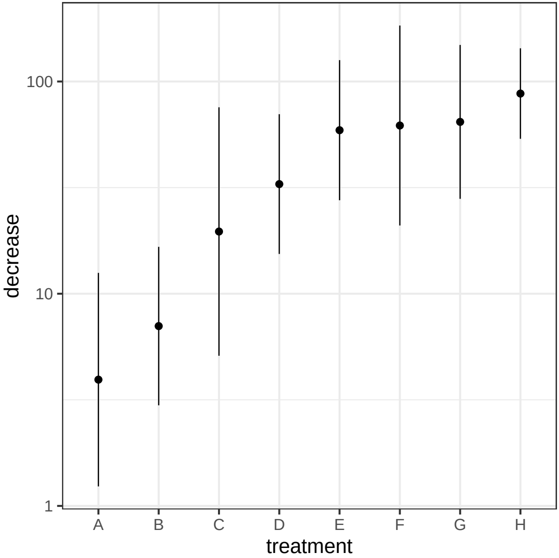
Means and standard errors



Means and standard deviations

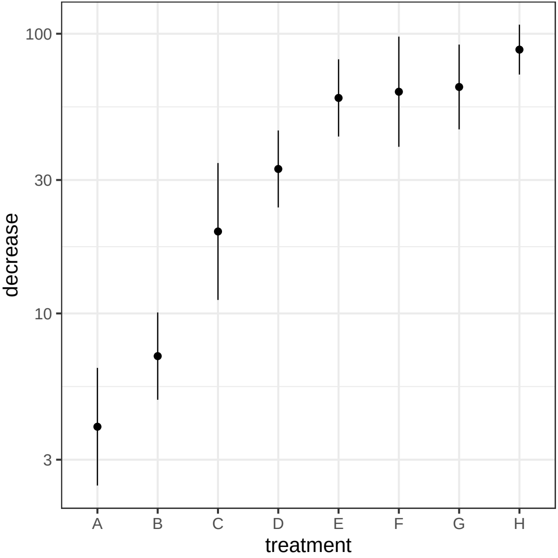


Means and standard deviations

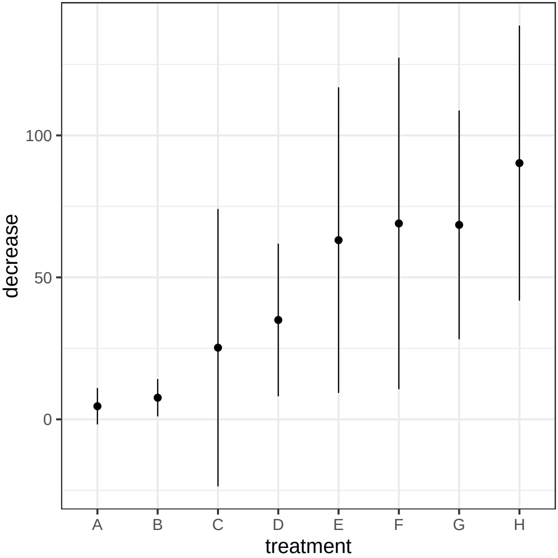


Means and standard deviations (repeat)

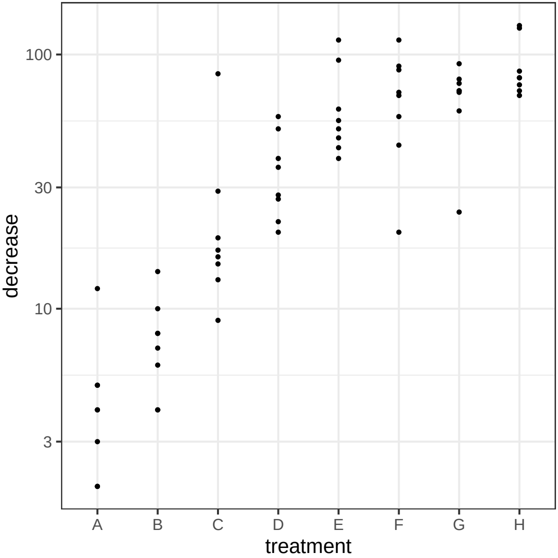
Next steps?



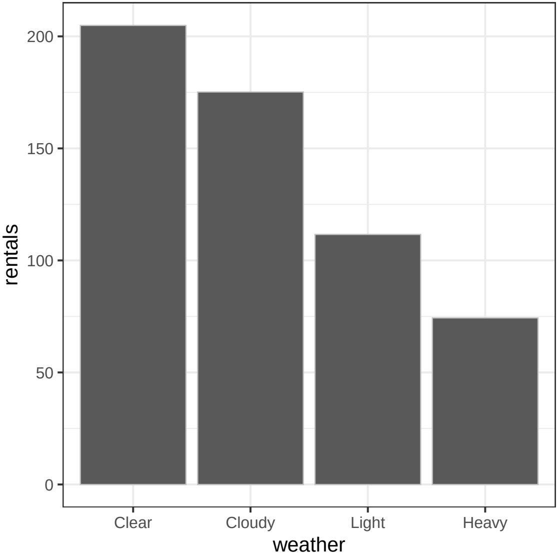
Non-parametric (repeat)



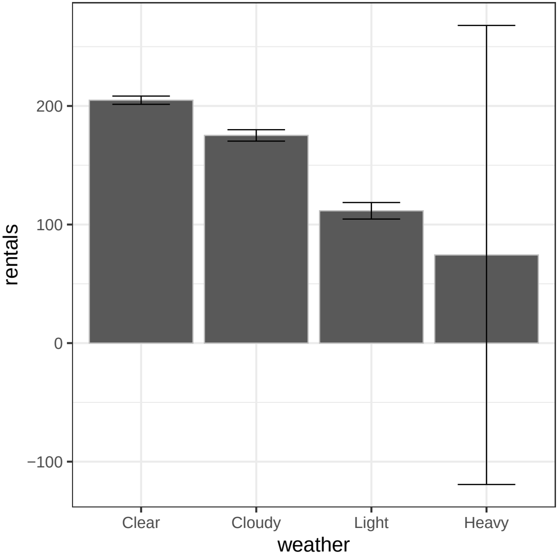
Non-parametric (repeat)



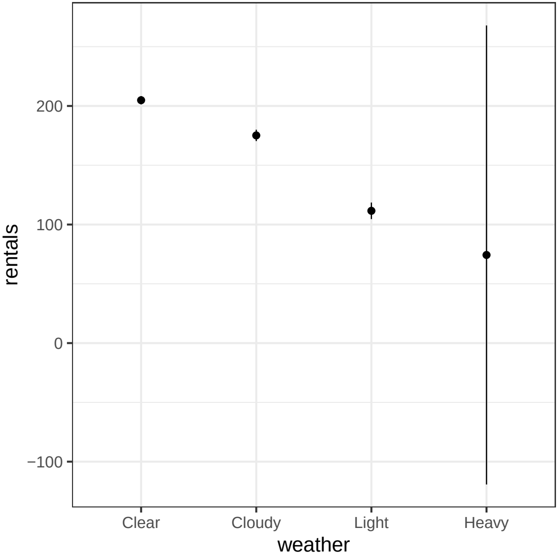
Bike example



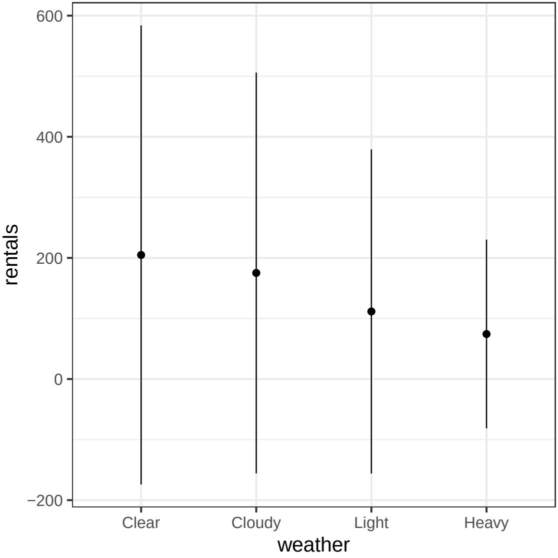
Standard errors



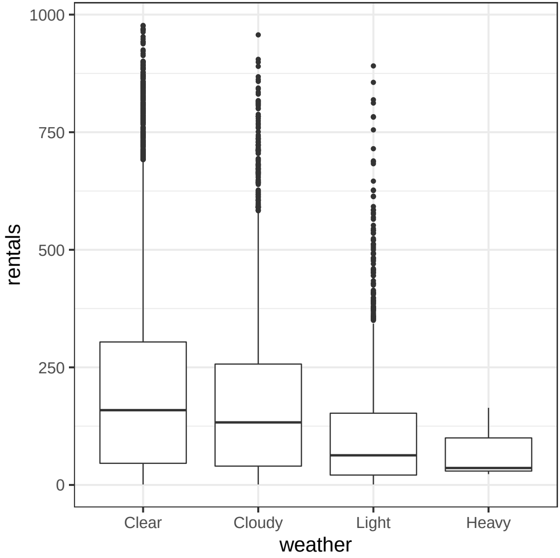
Standard errors



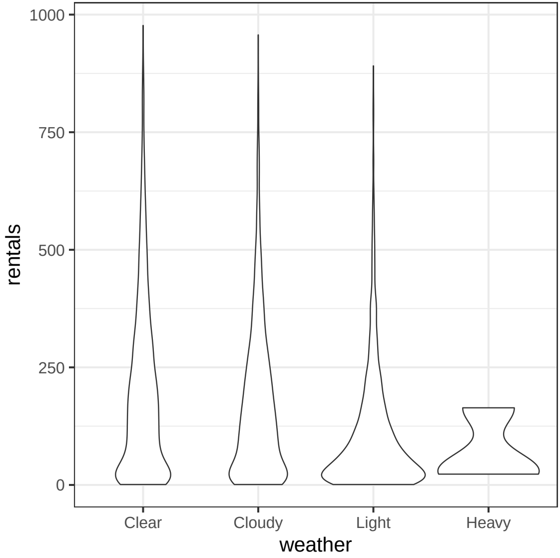
Standard deviations



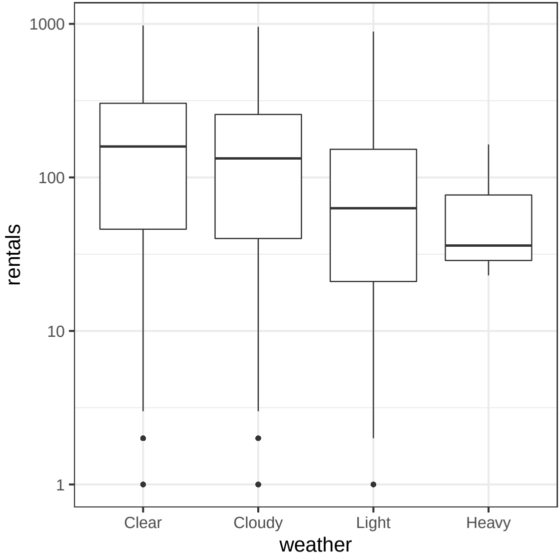
Data shape



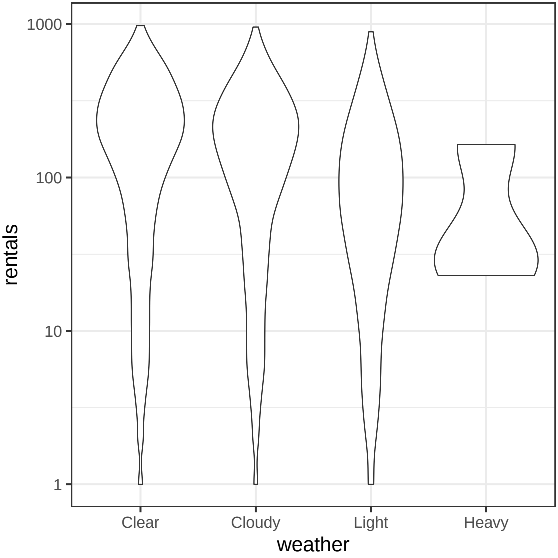
Data shape



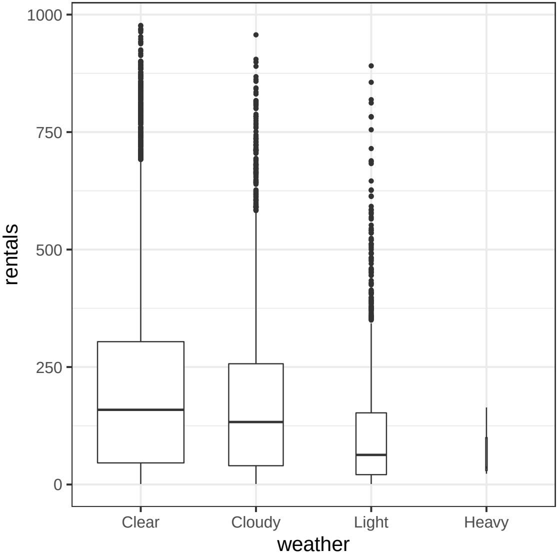
Data shape



Data shape



Data shape and weight



Log scales

- In general:

Log scales

- In general:
  - If your logged data span $<$ 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)

Log scales

- In general:
  - If your logged data span < 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
  - If not, just embrace "logs" (log10 particles per ul is from 3–8)

Log scales

- In general:
  - If your logged data span < 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
  - If not, just embrace "logs" (log10 particles per ul is from 3–8)
    - But remember these are not physical values

Log scales

- In general:
  - If your logged data span < 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
  - If not, just embrace “logs” (log10 particles per ul is from 3–8)
    - But remember these are not physical values
- I love natural logs, but not as axis values

Log scales

- In general:
  - If your logged data span < 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
  - If not, just embrace “logs” (log10 particles per ul is from 3–8)
    - But remember these are not physical values
- I love natural logs, but not as axis values
  - Except to represent proportional difference!

Log scales

- In general:
  - If your logged data span $<$ 3 decades, use human-readable numbers (e.g., 10-5000 kilotons per hectare)
  - If not, just embrace "logs" (log10 particles per ul is from 3–8)
    - But remember these are not physical values
- I love natural logs, but not as axis values
  - Except to represent proportional difference!

Outline

Banking

- Banking is a real thing

Banking

- Banking is a real thing
  - Even though many examples are bogus

Banking

- Banking is a real thing
  - Even though many examples are bogus
- Since the point is make patterns visually clear, trial-and-error is usually as good as algorithm

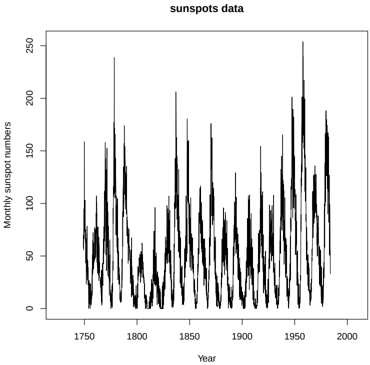
Banking

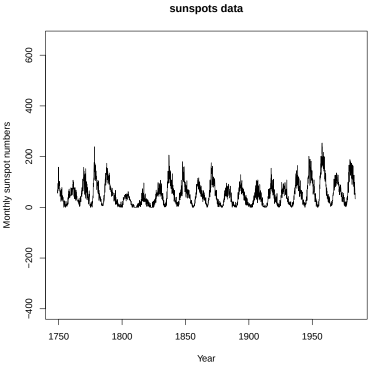
- Banking is a real thing
  - Even though many examples are bogus
- Since the point is make patterns visually clear, trial-and-error is usually as good as algorithm
  - But it is worth considering

Banking

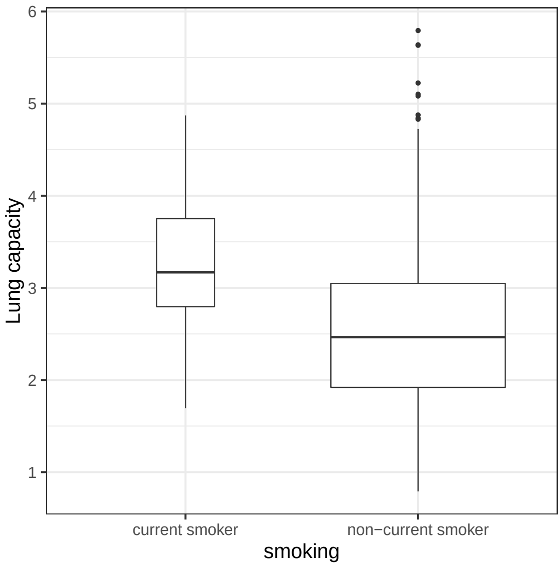
- Banking is a real thing
  - Even though many examples are bogus
- Since the point is make patterns visually clear, trial-and-error is usually as good as algorithm
  - But it is worth considering

Sunspots

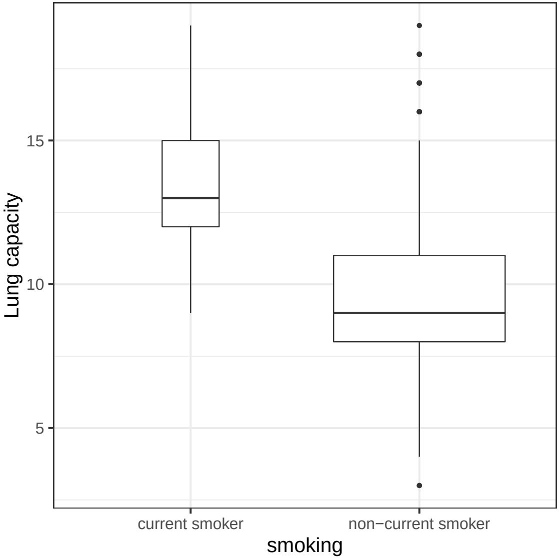

sunspots data
Monthly sunspot numbers
0
50
100
150
200
250
1750
1800
1850
1900
1950
2000
Year


sunspots data
Monthly sunspot numbers
−400
−200
0
200
400
600
1750
1800
1850
1900
1950
Year

Smoking data



Smoking data



Scatter plots

- Depending on how many data points you have, scatter plots may indicate relationships clearly

Scatter plots

- Depending on how many data points you have, scatter plots may indicate relationships clearly
- They can often be improved with trend interpolations

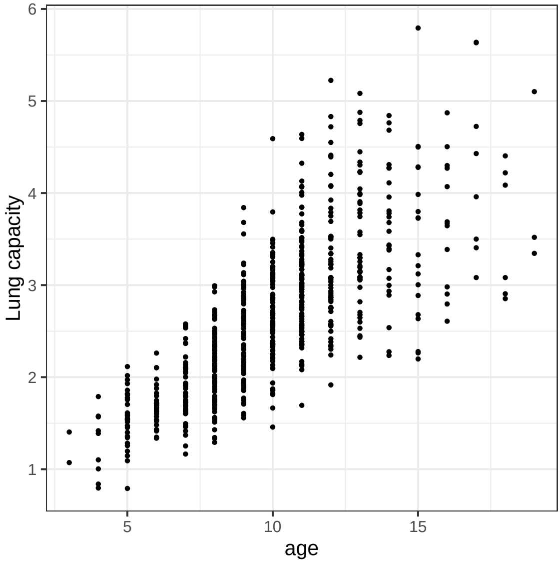
Scatter plots

- Depending on how many data points you have, scatter plots may indicate relationships clearly
- They can often be improved with trend interpolations
  - Interpolations may be particularly good for discrete responses (count or true-false)

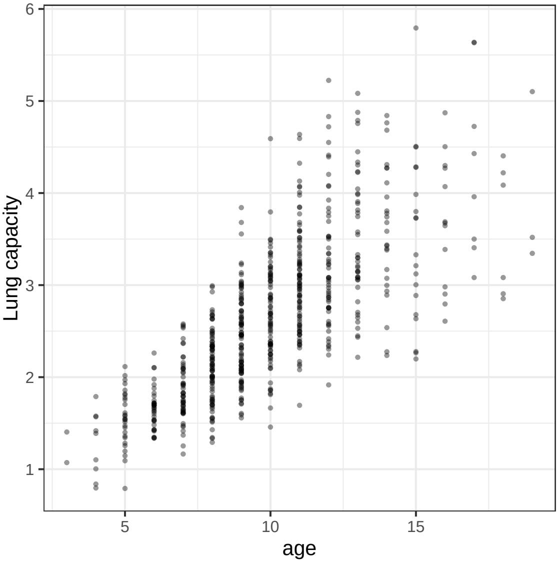
Scatter plots

- Depending on how many data points you have, scatter plots may indicate relationships clearly
- They can often be improved with trend interpolations
  - Interpolations may be particularly good for discrete responses (count or true-false)

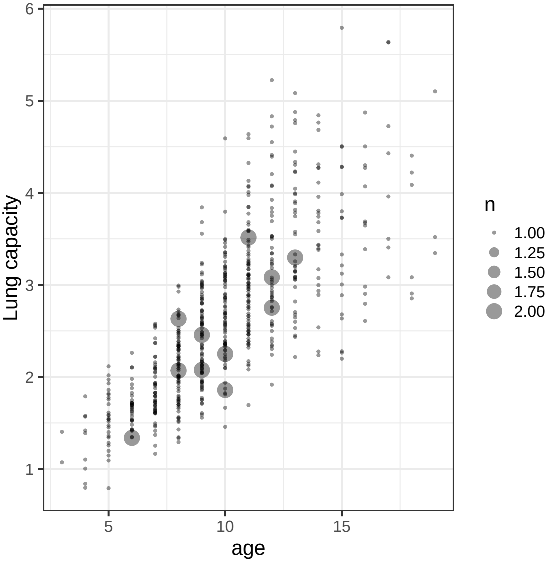
Scatter plot



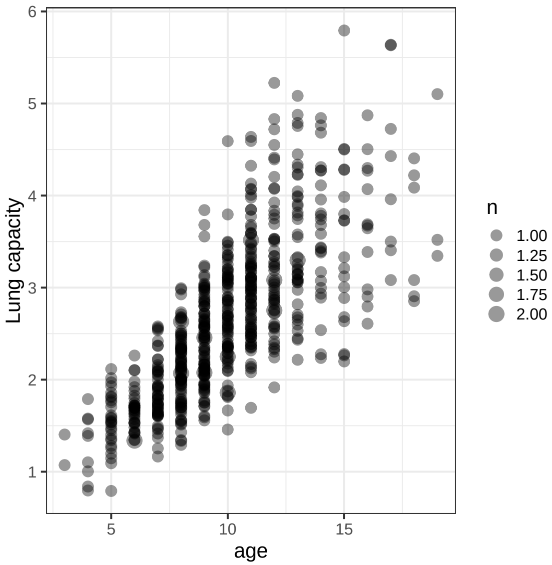
Seeing the density better



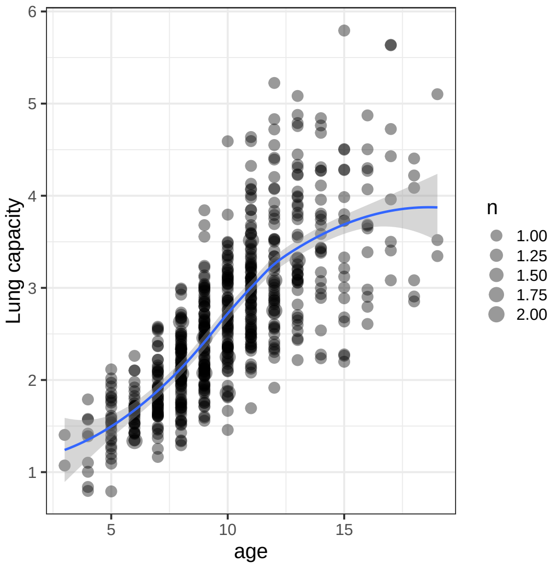
Seeing the density worse



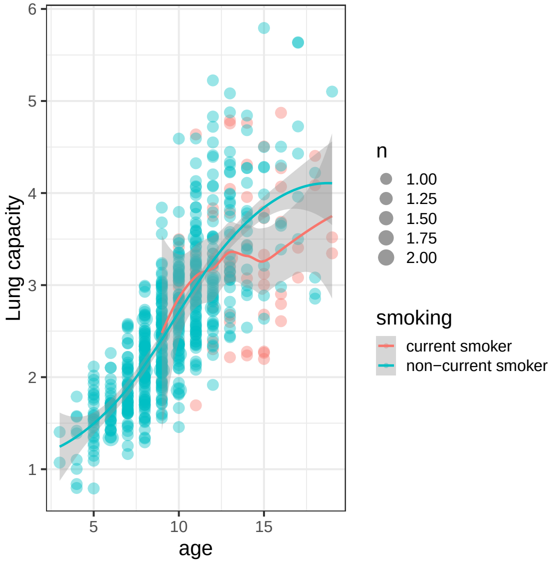
Maybe fixed



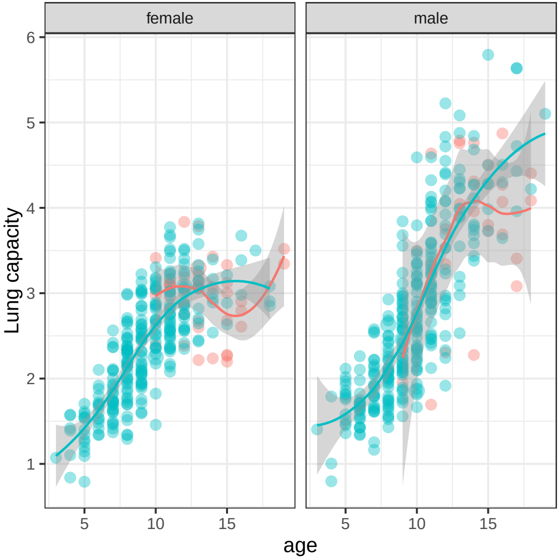
A loess trend line



Two loess trend lines



Many loess trend lines



Theory of loess

- Local smoother (locally flat, linear or **quadratic**)

Theory of loess

- Local smoother (locally flat, linear or **quadratic**)
- Neighborhood size given by alpha

Theory of loess

- Local smoother (locally flat, linear or **quadratic**)
- Neighborhood size given by alpha
  - Points in neighborhood are weighted by distance

Theory of loess

- Local smoother (locally flat, linear or **quadratic**)
- Neighborhood size given by alpha
  - Points in neighborhood are weighted by distance
- Check help function for loess

Theory of loess

- Local smoother (locally flat, linear or **quadratic**)
- Neighborhood size given by alpha
  - Points in neighborhood are weighted by distance
- Check help function for loess

Robust methods

- Loess is local, but not robust

Robust methods

- Loess is local, but not robust
  - Uses least squares, can respond strongly to outliers

Robust methods

- Loess is local, but not robust
  - Uses least squares, can respond strongly to outliers
- R is has a very flexible function called rlm to do robust fitting

Robust methods

- Loess is local, but not robust
  - Uses least squares, can respond strongly to outliers
- R is has a very flexible function called rlm to do robust fitting
  - *Not local*

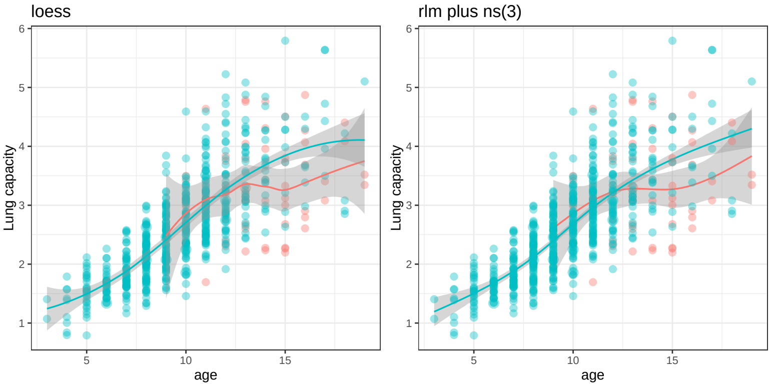
Robust methods

- Loess is local, but not robust
  - Uses least squares, can respond strongly to outliers
- R is has a very flexible function called rlm to do robust fitting
  - *Not local*
  - But can be combined with splines

Robust methods

- Loess is local, but not robust
  - Uses least squares, can respond strongly to outliers
- R is has a very flexible function called rlm to do robust fitting
  - *Not local*
  - But can be combined with splines

Fitting comparison



Density plots

- Contours

Density plots

- Contours
  - use `_density_2d()` to fit a two-dimensional kernel to the density

Density plots

- Contours
  - use `_density_2d()` to fit a two-dimensional kernel to the density
- hexes

Density plots

- Contours
  - use `_density_2d()` to fit a two-dimensional kernel to the density
- hexes
  - use `geom_hex` to plot densities using hexes

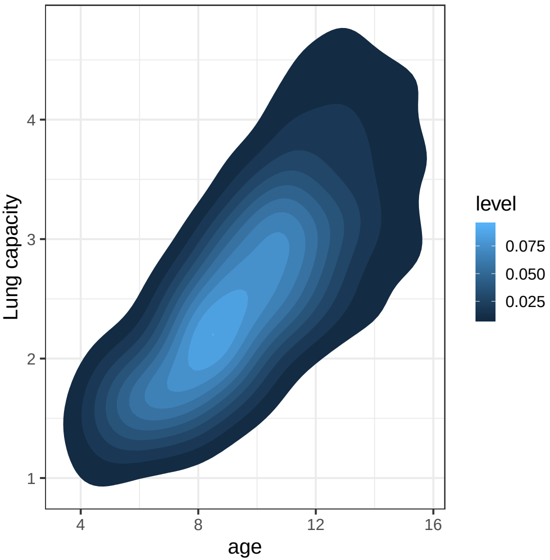
Density plots

- Contours
  - use `_density_2d()` to fit a two-dimensional kernel to the density
- hexes
  - use `geom_hex` to plot densities using hexes
  - this can also be done using rectangles for data with more discrete values

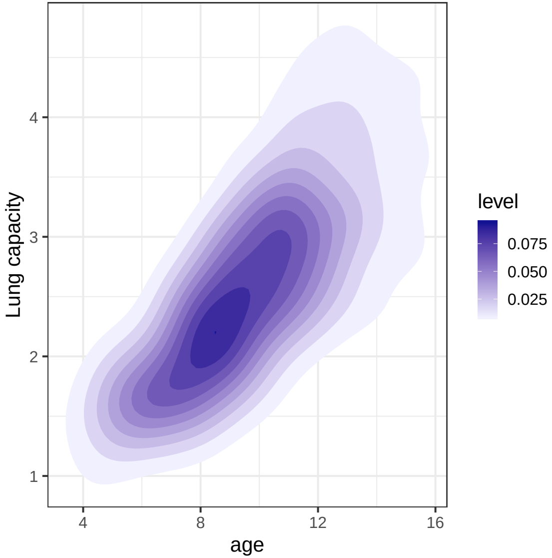
Density plots

- Contours
  - use `_density_2d()` to fit a two-dimensional kernel to the density
- hexes
  - use `geom_hex` to plot densities using hexes
  - this can also be done using rectangles for data with more discrete values

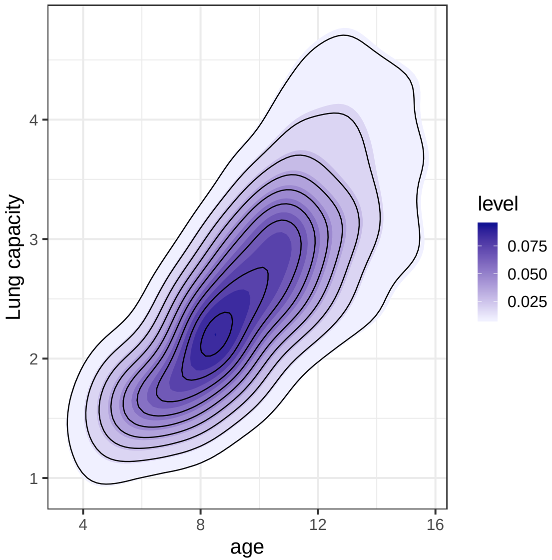
Contours



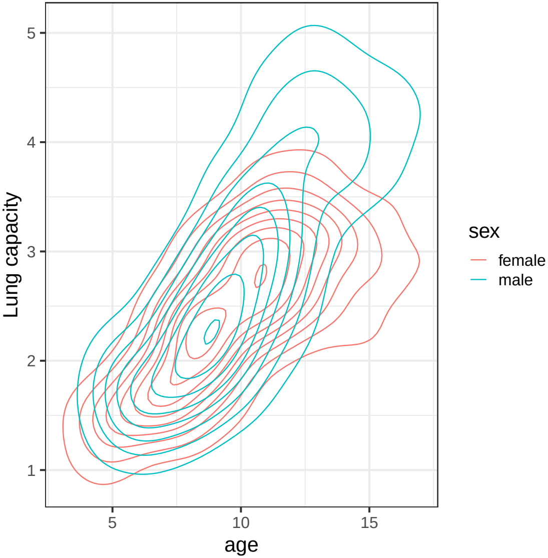
Contours



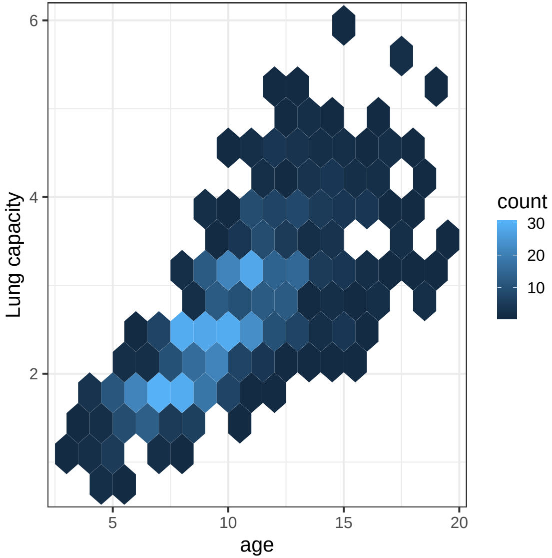
Contours



Hexes



Hexes



Color principles

- Use clear gradients

Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction

Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction
  - e.g., white to blue, white to red

Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction
  - e.g., white to blue, white to red
  - If the map contrasts with a background, zero should match the background

Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction
  - e.g., white to blue, white to red
  - If the map contrasts with a background, zero should match the background
- If there's a natural *middle*, you can use blue to white to red, or something similar

Color principles

- Use clear gradients
- If zero has a physical meaning (like density), go in just one direction
  - e.g., white to blue, white to red
  - If the map contrasts with a background, zero should match the background
- If there's a natural *middle*, you can use blue to white to red, or something similar

Outline

Multiple dimensions

- Three dimensional data is a lot like two-d with densities: contour plots are good

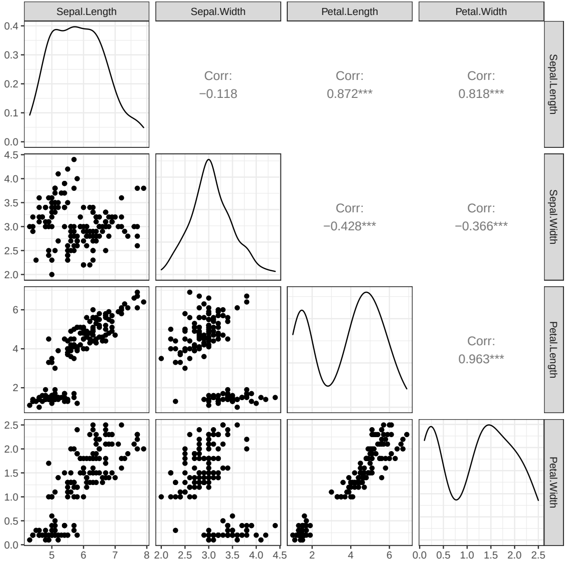
Multiple dimensions

- Three dimensional data is a lot like two-d with densities: contour plots are good
- Pairs plots: `pairs`, `ggpairs`

Multiple dimensions

- Three dimensional data is a lot like two-d with densities: contour plots are good
- Pairs plots: `pairs`, `ggpairs`

Pairs example



Outline

Multiple factors

- Use boxplots and violin plots

Multiple factors

- Use boxplots and violin plots
- Make use of `facet_wrap` and `facetgrid`

Multiple factors

- Use boxplots and violin plots
- Make use of `facet_wrap` and `facetgrid`
- Use different combinations (e.g., try plots with the same info, but different factors on the axes vs. in the colors or the facets)

Multiple factors

- Use boxplots and violin plots
- Make use of `facet_wrap` and `facetgrid`
- Use different combinations (e.g., try plots with the same info, but different factors on the axes vs. in the colors or the facets)