ETC5521: Diving Deeply into Data Exploration

Exploring bivariate dependencies

Professor Di Cook

Department of Econometrics and Business Statistics

The world is full of obvious things which nobody by any chance observes

Quote from Arthur Conan Doyle, The Hound of the Baskervilles

The story of the galloping horse

Galloping horses throughout history were drawn with all four legs out.

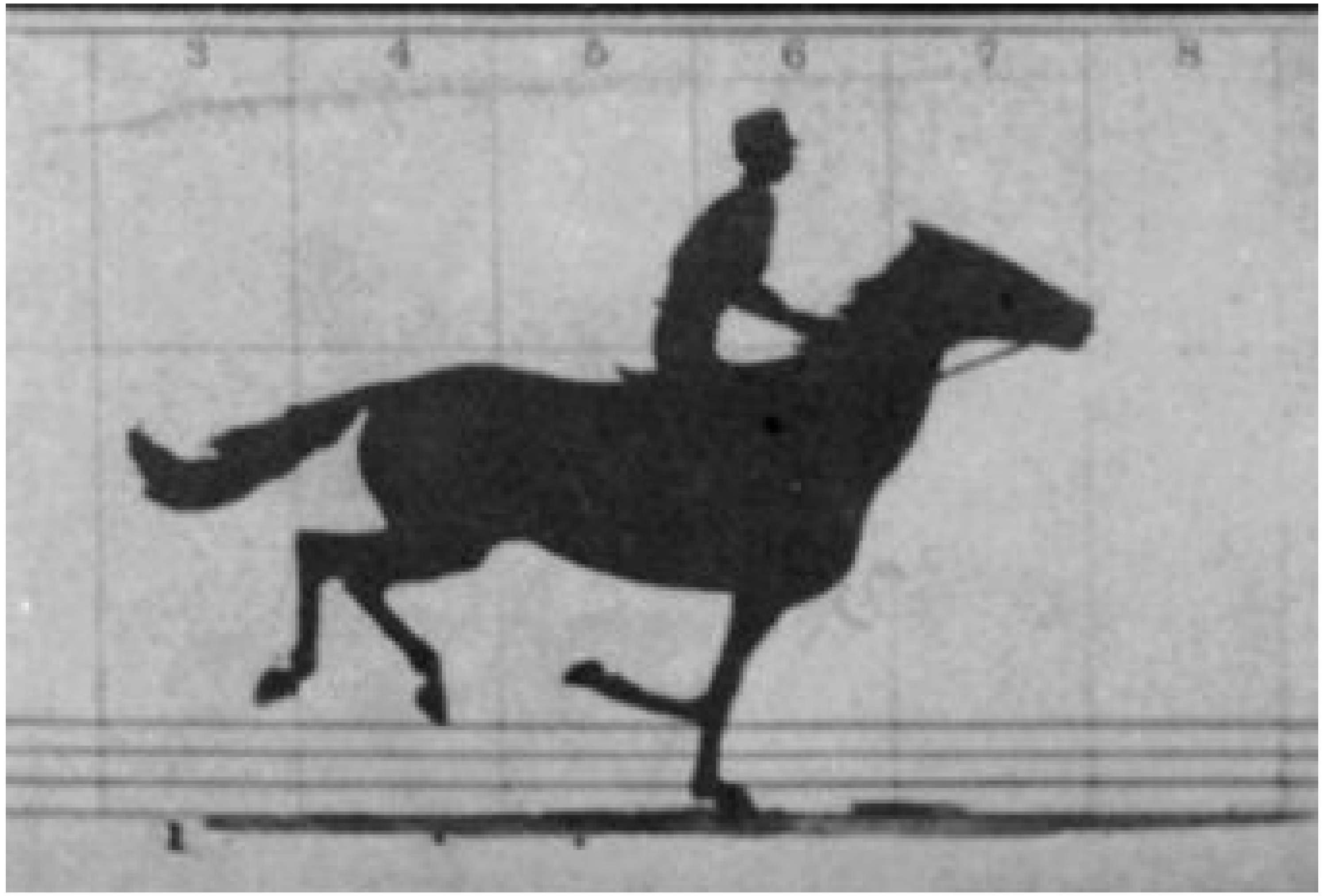
Baronet, 1794

Derby D'Epsom 1821

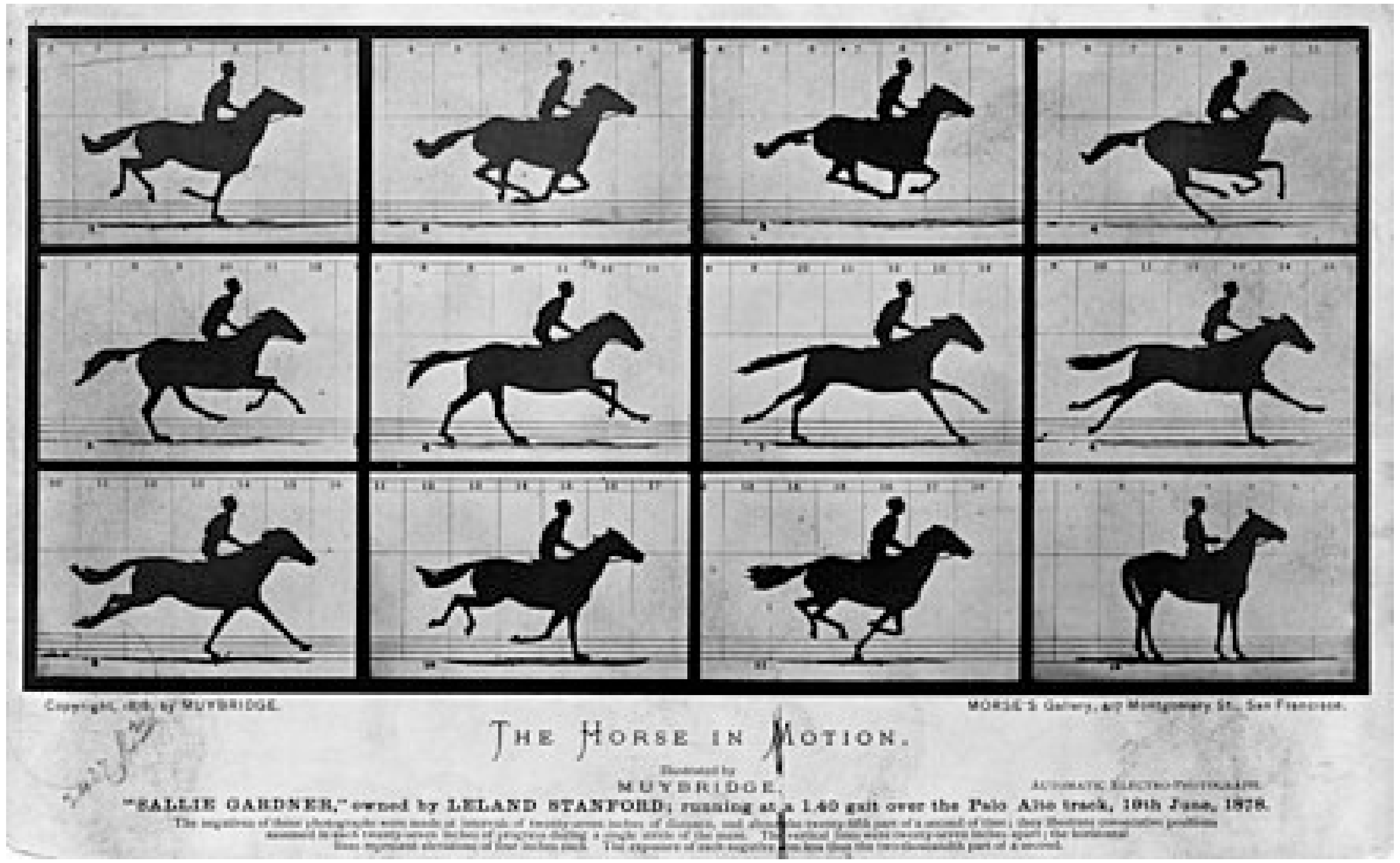
Read more in Lankester: The Problem of the Galloping Horse

The story of the galloping horse

With the birth of photography, and particular motion photography, Muybridge was able to illustrate that all four legs are never extended simultaneously.



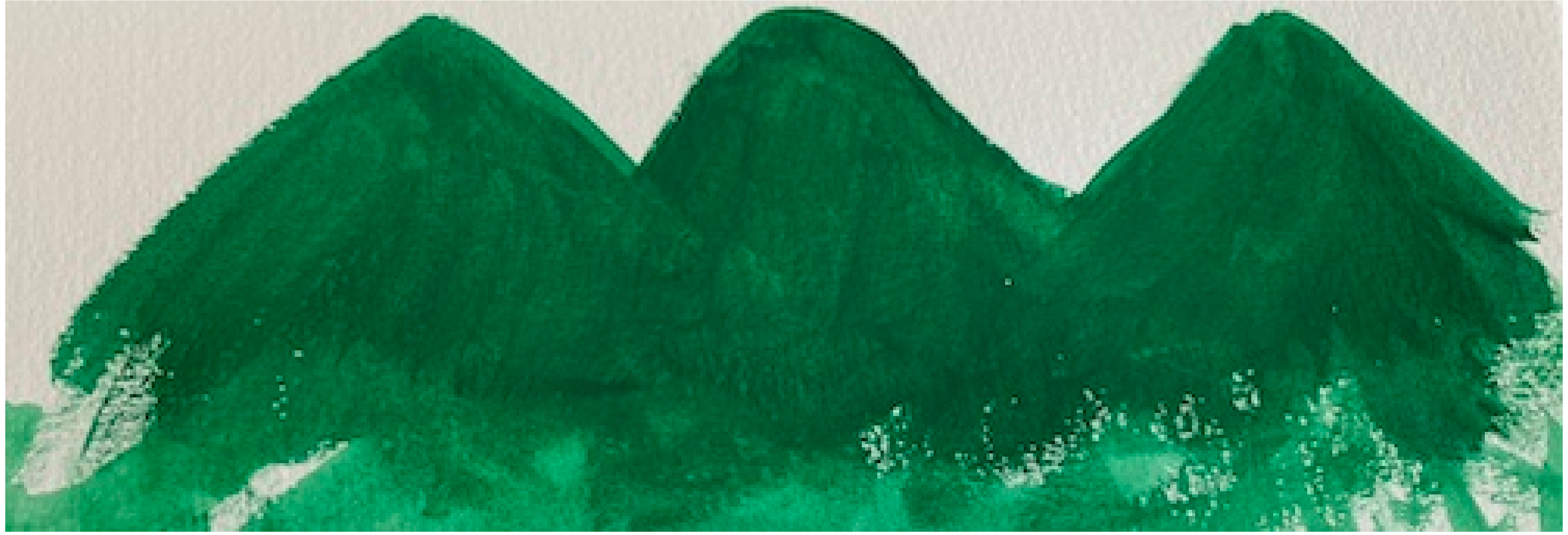
Source: wikimedia



Source: wikimedia

An evolution in seeing the world (1/3)

hills - beginner | hills - terrible fix | hills - what do you see

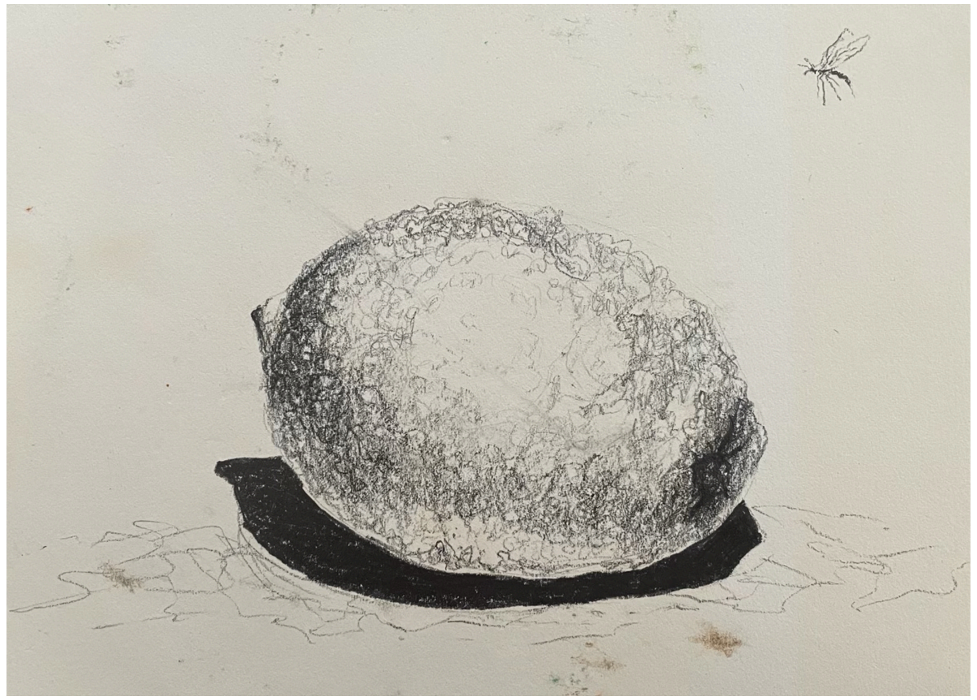


Mrs Robinson says *Hills are more interesting than that. Usually you can see valleys and shadows.*

An evolution in seeing the world (2/3)

Drawing lemons

Is something is missing?



My sketch

Notice the yellow reflection(s) and shine on the skin?

An evolution in seeing the world (3/3)

Drawing trees

Does it look like a tree? What is not realistic?

Tree foliage has lots of different colours

Source: https://toppng.com

Philosophical reflection

You, me, we humans have a tendency to

- only see what other people have done or say,
- not what we can see,
- or impose beliefs, like trees are green.

When you look at data, you might discover that there is a different story, or many different stories.

Try to see with fresh eyes

Outline

- The humble but powerful scatterplot
- Additions and variations
- Transformations to linearity
- (Robust) numerical measures of association
- Simpson’s paradox
- Making null samples to test for association
- Imputing missings

The scatterplot

> Scatterplots are the natural plot to make to explore association between two **continuous** (quantitative or numeric) variables.

They are not just for linear relationships but are useful for examining nonlinear patterns, clustering and outliers.

We also can think about scatterplots in terms of statistical distributions: if a histogram shows a marginal distribution, a scatterplot allows us to examine the bivariate distribution of a sample.

History

- Descartes provided the Cartesian coordinate system in the 17th century, with perpendicular lines indicating two axes.

- It wasn't until **1832** that the scatterplot appeared, when John Frederick Herschel plotted position and time of double stars.

- This is 200 years after the Cartesian coordinate system, and 50 years after bar charts and line charts appeared, used in the work of William Playfair to examine economic data.

- Kopf argues that *The scatter plot, by contrast, proved more useful for scientists*, but it clearly is useful for economics today.

http://www.datavis.ca/milestones/

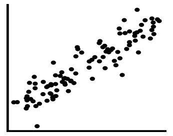
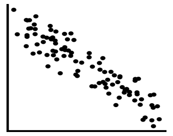
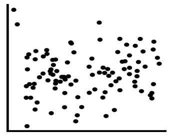
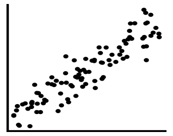
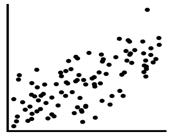
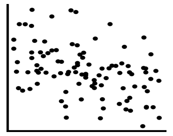
Language and terminology

Are the words "correlation" and "association" interchangeable?

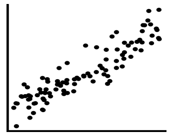
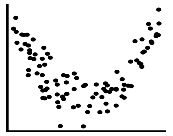
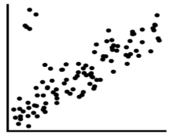
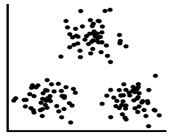
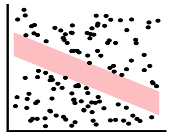
> In the broadest sense **correlation** is any statistical association, though it commonly refers to the degree to which a pair of variables are **linearly** related. Wikipedia

If the relationship is not linear, call it association, and avoid correlated.

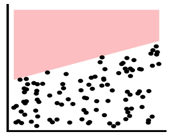
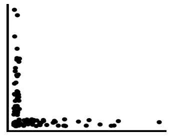
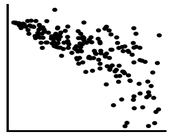
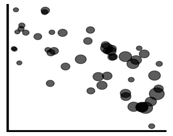
Features of a pair of continuous variables (1/3)

Feature	Example	Description
positive trend		Low value corresponds to low value, and high to high.
negative trend		Low value corresponds to high value, and high to low.
no trend		No relationship
strong		Very little variation around the trend
moderate		Variation around the trend is almost as much as the trend
weak		A lot of variation making it hard to see any trend

Features of a pair of continuous variables (2/3)

Feature	Example	Description
linear form		The shape is linear
nonlinear form		The shape is more of a curve
nonlinear form		The shape is more of a curve
outliers		There are one or more points that do not fit the pattern on the others
clusters		The observations group into multiple clumps
gaps		There is a gap, or gaps, but its not clumped

Features of a pair of continuous variables (3/3)

Feature	Example	Description
barrier		There is combination of the variables which appears impossible
l-shape		When one variable changes the other is approximately constant
discreteness		Relationship between two variables is different from the overall, and observations are in a striped pattern
heteroskedastic		Variation is different in different areas, maybe depends on value of x variable
weighted		If observations have an associated weight, reflect in scatterplot, e.g. bubble chart

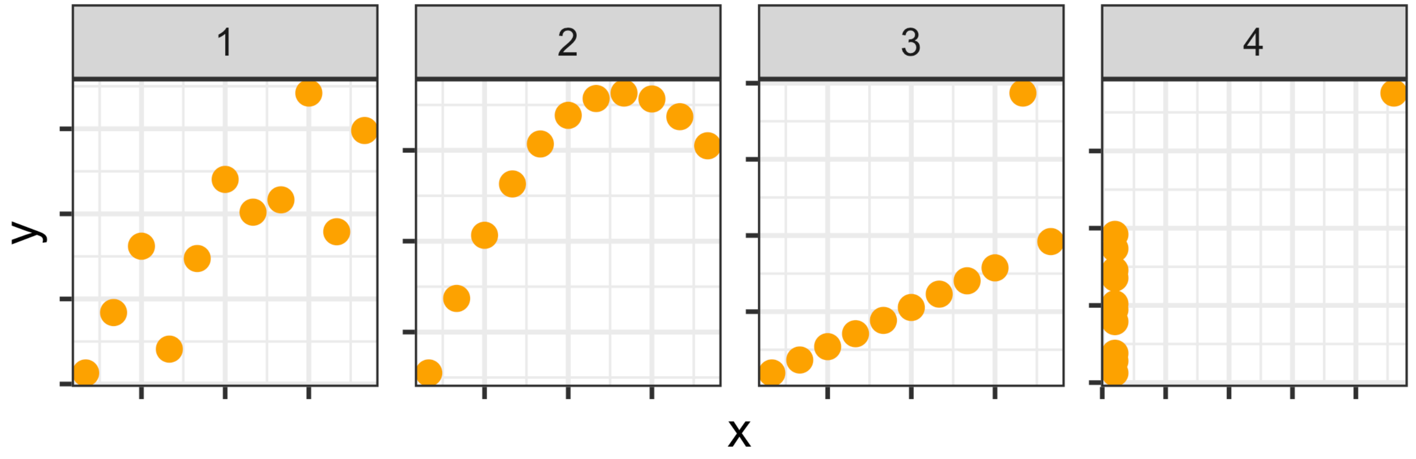
Additional considerations (Unwin, 2015):

- **causation**: one variable has a direct influence on the other variable, in some way. For example, people who are taller tend to weigh more. The dependent variable is conventionally on the y axis. *It's not generally possible to tell from the plot that the relationship is causal, which typically needs to be argued from other sources of information.*
- **association**: variables may be related to one another, but through a different variable, eg ice cream sales are positively correlated with beach drownings, is most likely a temperature relationship.
- **conditional relationships**: the relationship between variables is conditionally dependent on another, such as income against age likely has a different relationship depending on retired or not.

Famous data examples

Famous scatterplot examples

Anscombe's quartet

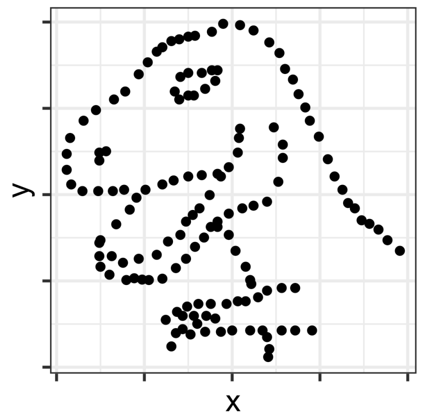


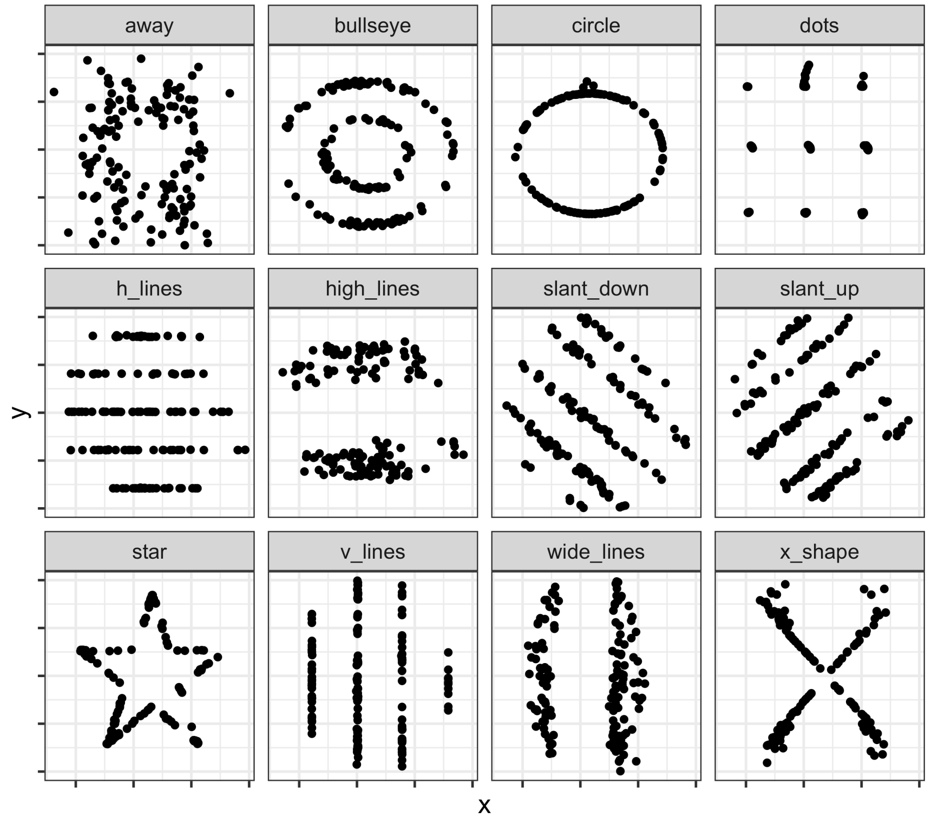
All four sets of Anscombe has same means, standard deviations and correlations, $\bar{x} = 9$, $\bar{y} = 7.5$, $s_x = 3.3$, $s_y = 2$, $r = 0.82$.

> Numerical statistics are the same, for very different association.

Datasaurus dozen

And similarly all 13 sets of the datasaurus dozen have same means, standard deviations and correlations, $\bar{x} = 54$, $\bar{y} = 48$, $s_x = 17$, $s_y = 27$, $r = -0.06$.

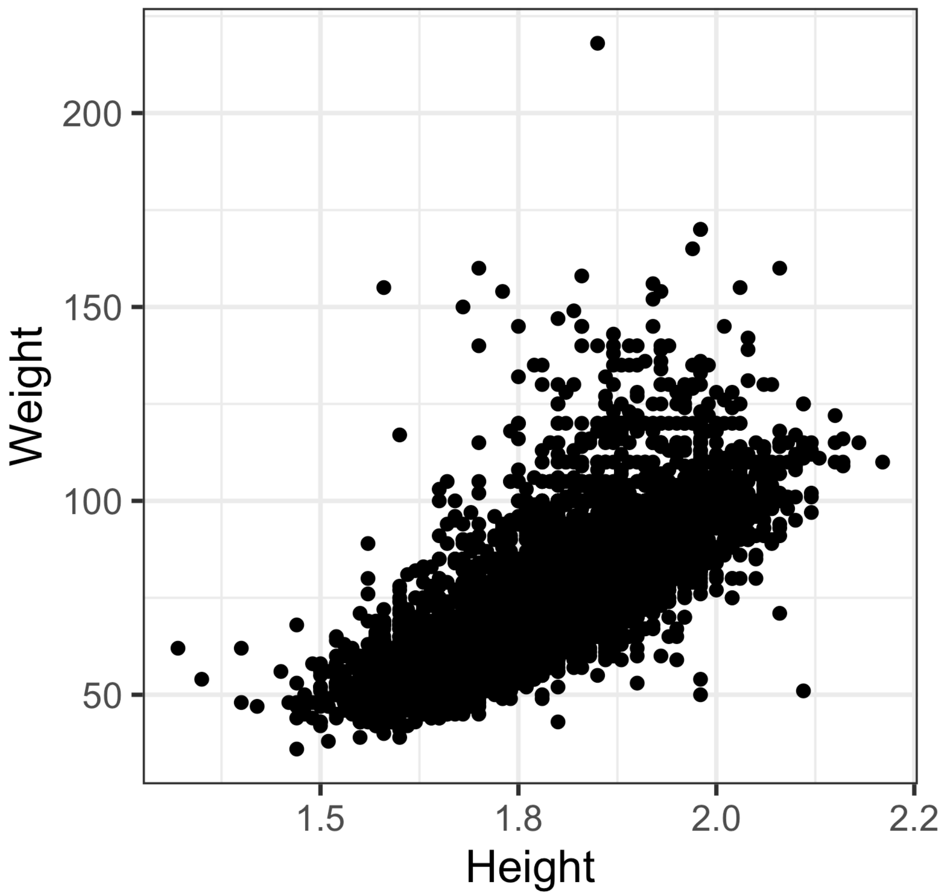




Scatterplot case studies

Case study: Olympics

🖼️ data R



- Note: `Warning message: Removed 1346 rows containing missing values (geom_point)`
- Features:
  - linear relationship (expected, more than?)
  - outliers
  - discretization
- Substantial overplotting, >10000 athletes.
- What is interesting? Are there some sport(s) where you would expect specific relationships?

Try this

> Interactivity can be a useful tool for exploring relationships.

Cut and paste the code into your R console, and the resulting plot to examine the sport of the athlete.

```
library(tidyverse)
library(plotly)
data(oly12, package = "VGAMdata")
p <- ggplot(oly12, aes(x = Height, y = Weight, label = Sport)) +
  geom_point()
ggplotly(p)
```

How many athletes in the different sports?

Search:

Sport	n
Athletics	2119
Swimming	907
Football	596
Rowing	524
Hockey	416
Judo	368
Shooting	368
Sailing	360
Wrestling	324

Categories need re-working:

- so many different events grouped into athletics
- cycling split among many categories

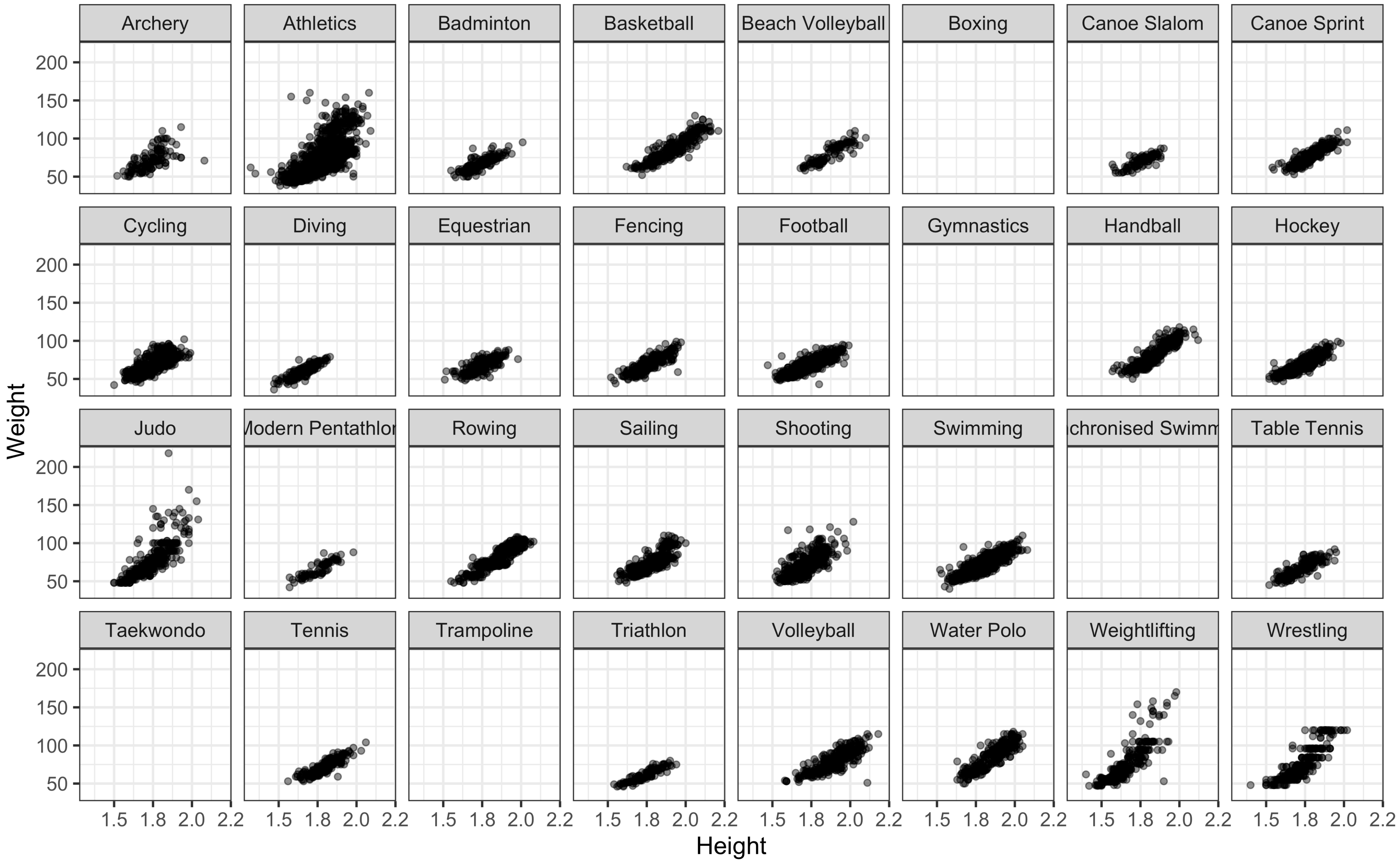
Consolidate factor levels

There are several cycling events that are reasonable to combine into one category. Similarly for gymnastics and athletics.

```r
oly12 <- oly12 |>
  mutate(Sport = as.character(Sport)) |>
  mutate(Sport = ifelse(grepl("Cycling", Sport),
    "Cycling", Sport
  )) |>
  mutate(Sport = ifelse(grepl("Gymnastics", Sport),
    "Gymnastics", Sport
  )) |>
  mutate(Sport = ifelse(grepl("Athletics", Sport),
    "Athletics", Sport
  )) |>
  mutate(Sport = as.factor(Sport))
```

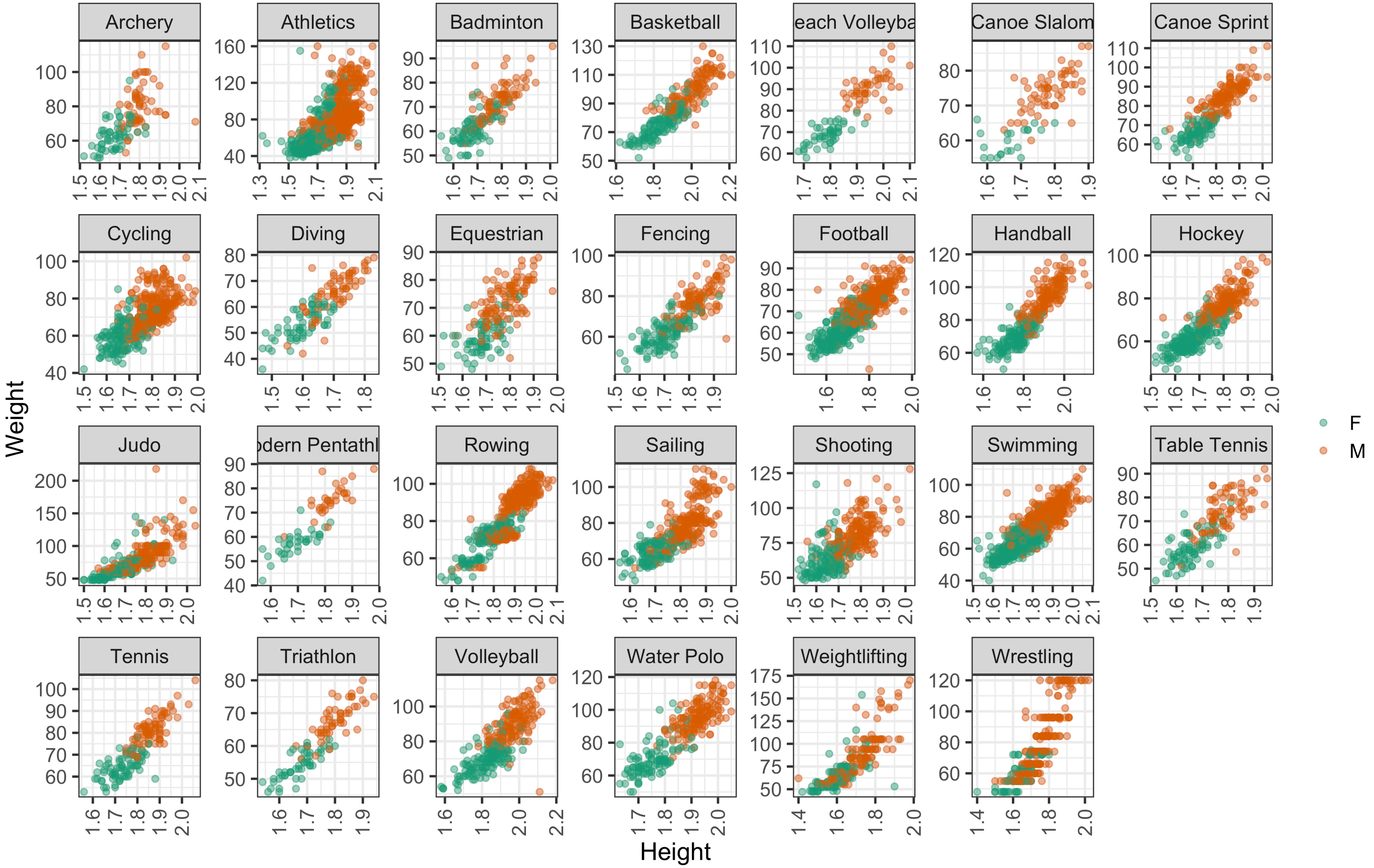
Drill down the by sport

🖼️ learn R

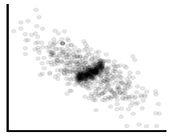
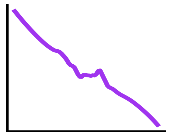
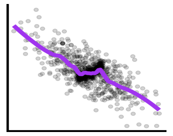
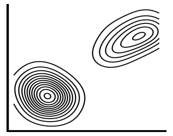
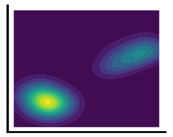
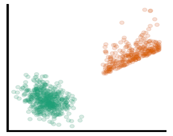


Remove missings, explore difference by sex

learn R

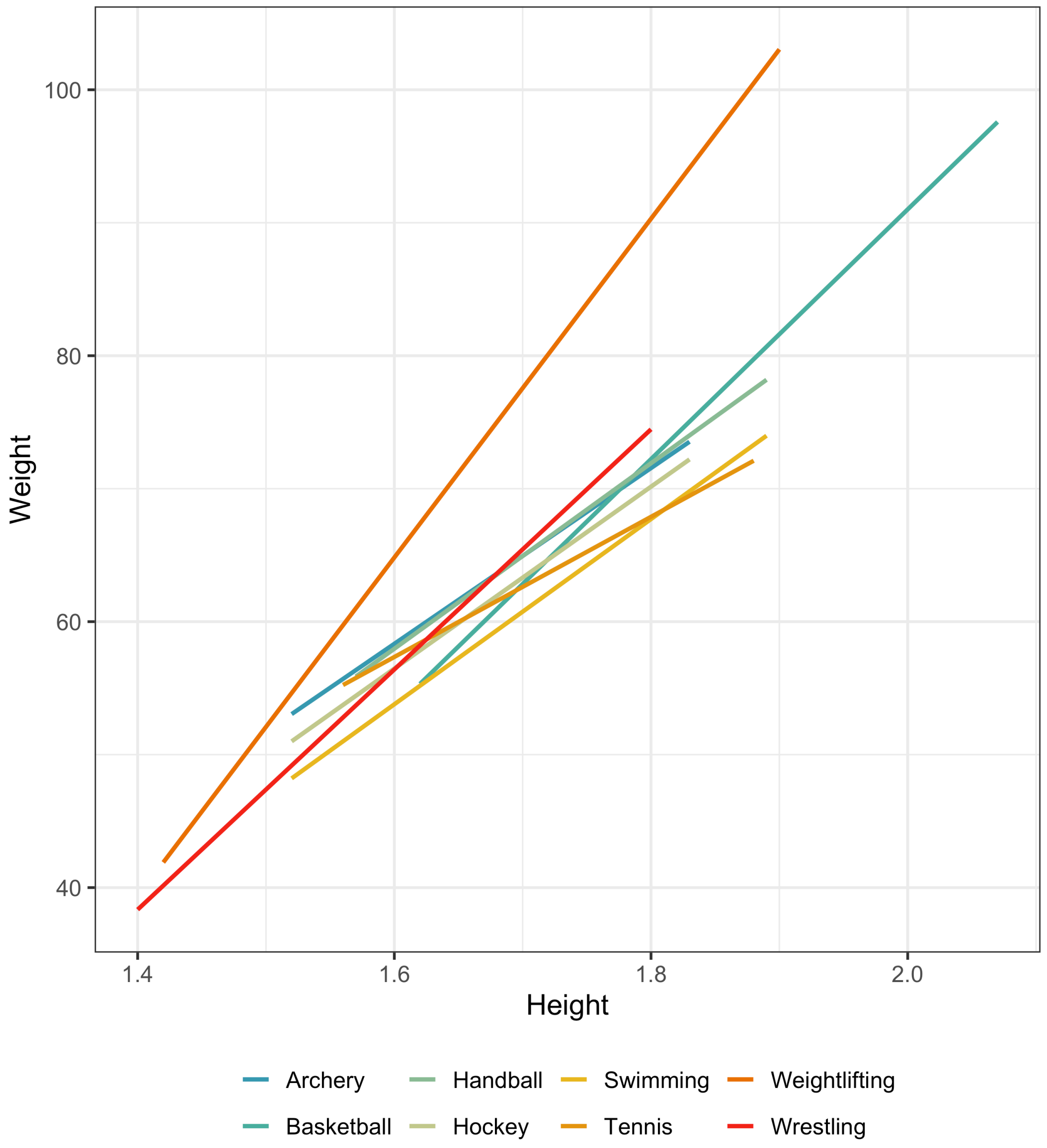


Common ways to augment scatterplots

Modification	Example	Purpose
alpha-blend		alleviate overplotting to examine density at centre
model overlay		focus on the trend
model + data		trend plus variation
density		overall distribution, variation and clustering
filled density		high density locations in distribution (modes), variation and clustering
colour		relationship with conditioning and lurking variables

Comparing association

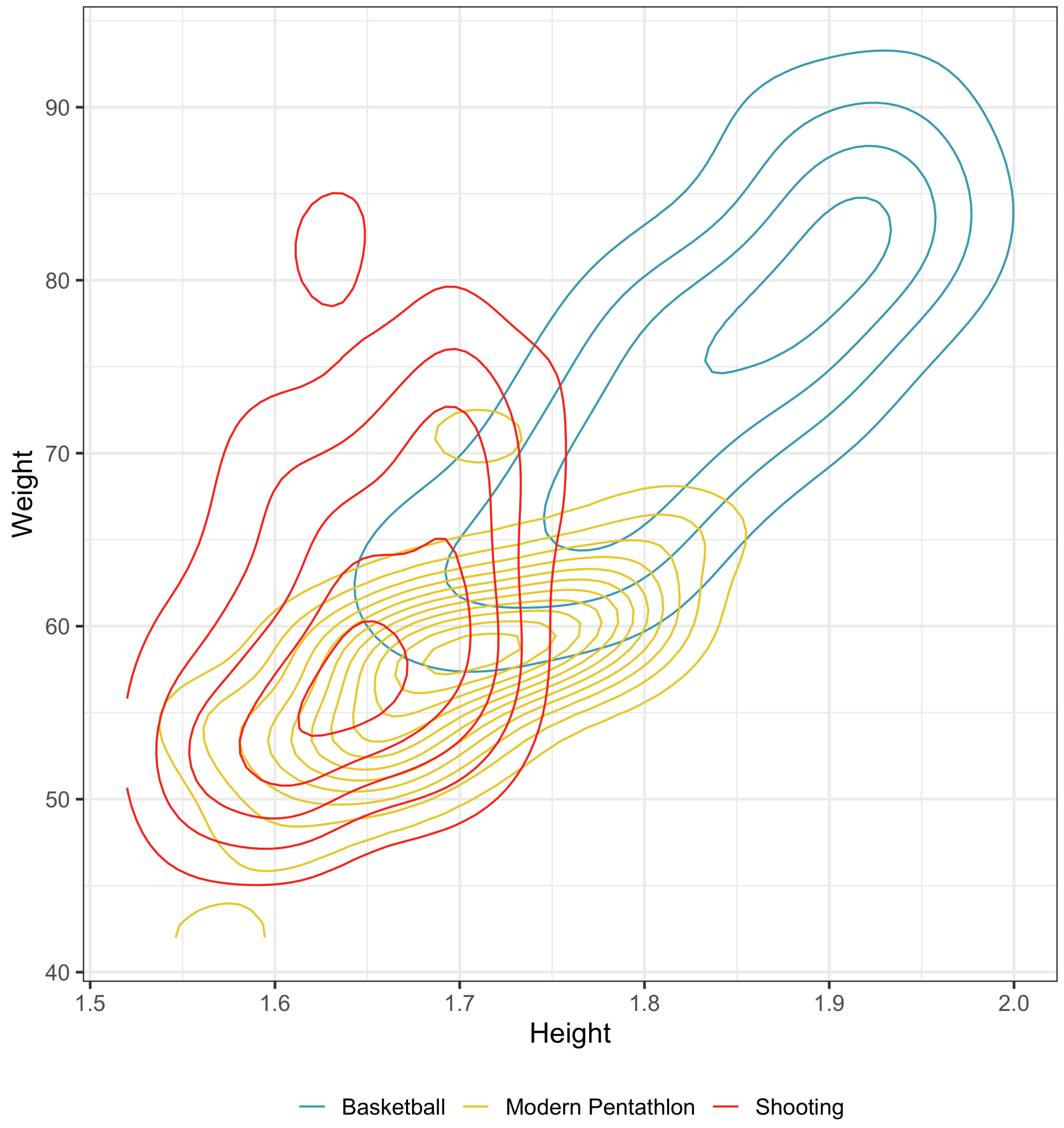
🖼️ R



- Weightlifters are much heavier relative to height
- Swimmers are leaner relative to height
- Tennis players are a bit mixed, shorter tend to be heavier, taller tend to be lighter

Comparing spread

🖼 R



- Modern pentathlon athletes are uniformly height and weight related
- Shooters are quite varied in body type

Case study: Olympics

We learned that association between height and weight is different strata, defined by categorical variables: sport, gender, and possibly country and age, too.

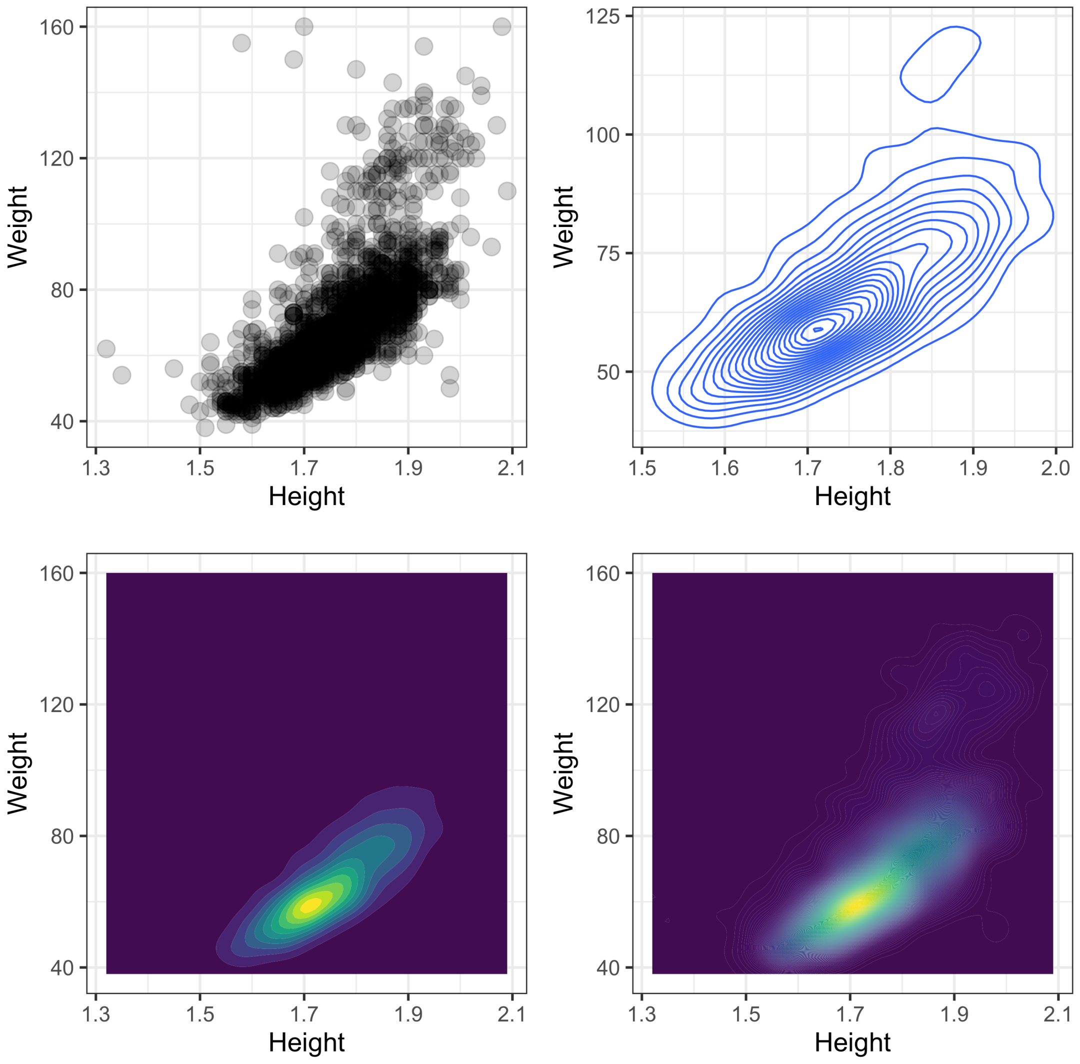
Some of the association may be due to unmeasured variables, for example, “Athletics” is masking different body types in throwing vs running. This is a **lurking** variable.

If you were just given the `Height` and `Weight` in this data could you have detected the presence of conditional relationships?

It may appear as multimodality.

Can you see conditional dependencies?

R



There is a barely hint of multimodality.

It's not easy to detect the presence of the additional variable, and thus accurately describe the relationship between height and weight among Olympic athletes.

Numerical measures of association

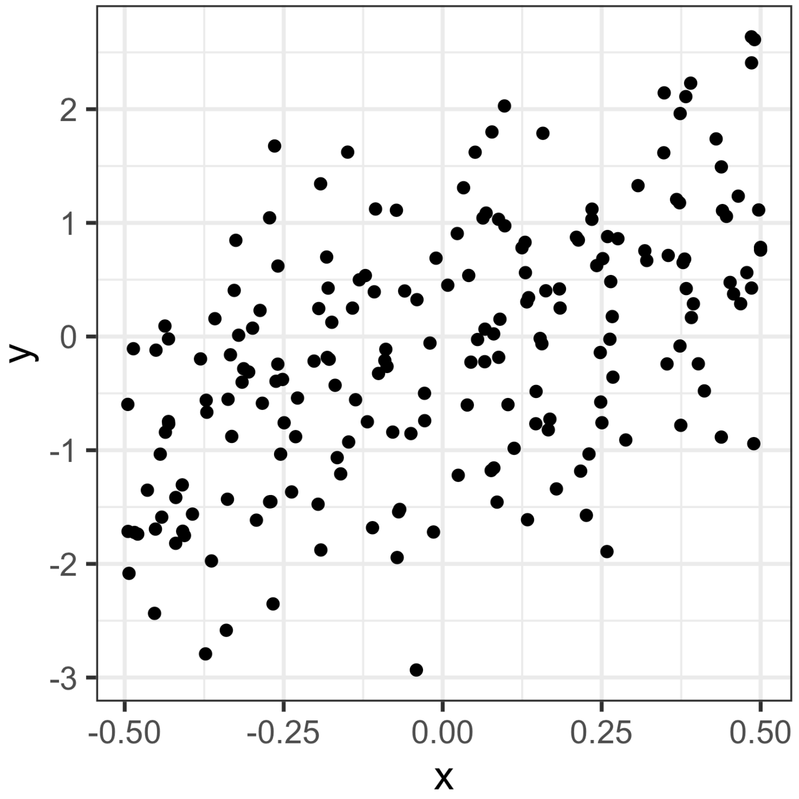
Correlation

- Correlation between variables x_1 and x_2, with n observations in each.

$$r = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} = \frac{\text{covariance}(x_1, x_2)}{(n-1)s_{x_1}s_{x_2}}$$

- Test for statistical significance, whether population correlation could be 0 based on observed r, using a t_{n-2} distribution:

$$t = \frac{r}{\sqrt{1 - r^2}}\sqrt{n-2}$$



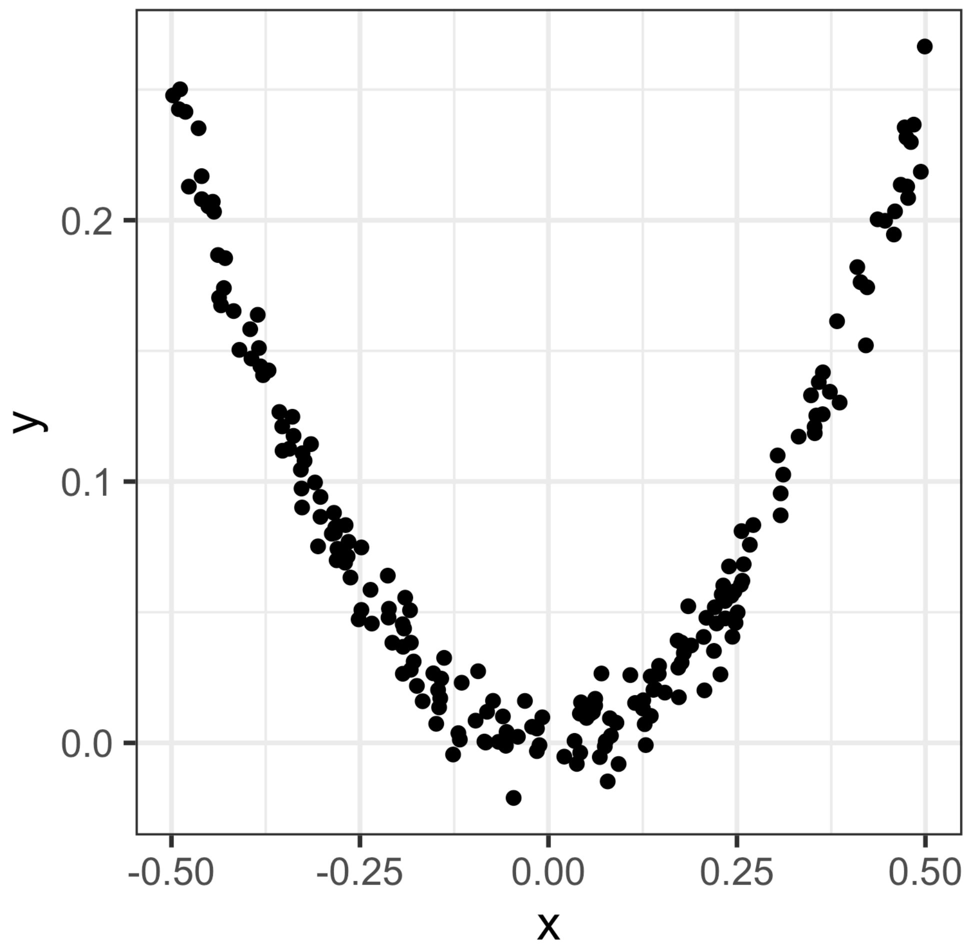

```r
cor(d1$x, d1$y)
```

```
[1] 0.52
```

```r
cor.test(d1$x, d1$y)
```

```

	Pearson's product-moment correlation

data:  d1$x and d1$y
t = 9, df = 198, p-value = 2e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.41 0.62
sample estimates:
 cor
0.52
```

Problems with correlation (1/2)




```r
cor(d2$x, d2$y)
```

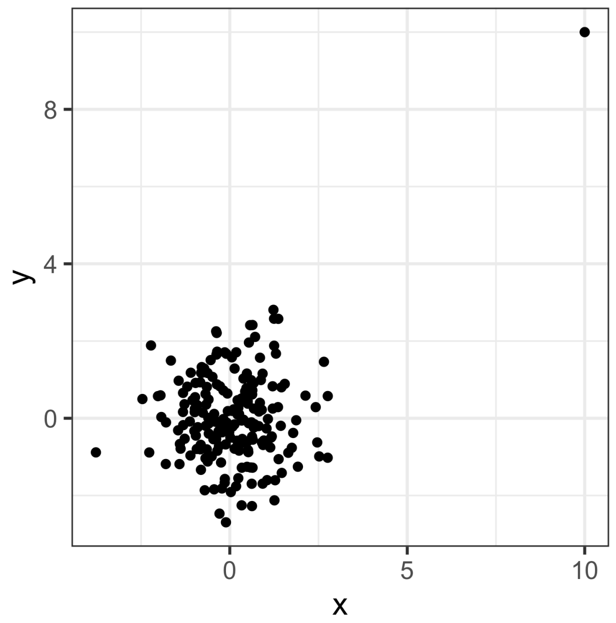
```
[1] -0.05
```

```r
cor.test(d2$x, d2$y)
```

```

	Pearson's product-moment correlation

data:  d2$x and d2$y
t = -0.7, df = 198, p-value = 0.5
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.187  0.089
sample estimates:
  cor 
-0.05 
```

It does not summarise non-linear associations.

Problems with correlation (2/2)



All observations

```
$estimate
cor
0.3

$statistic
  t
4.4

$p.value
[1] 1.6e-05
```

Without outlier

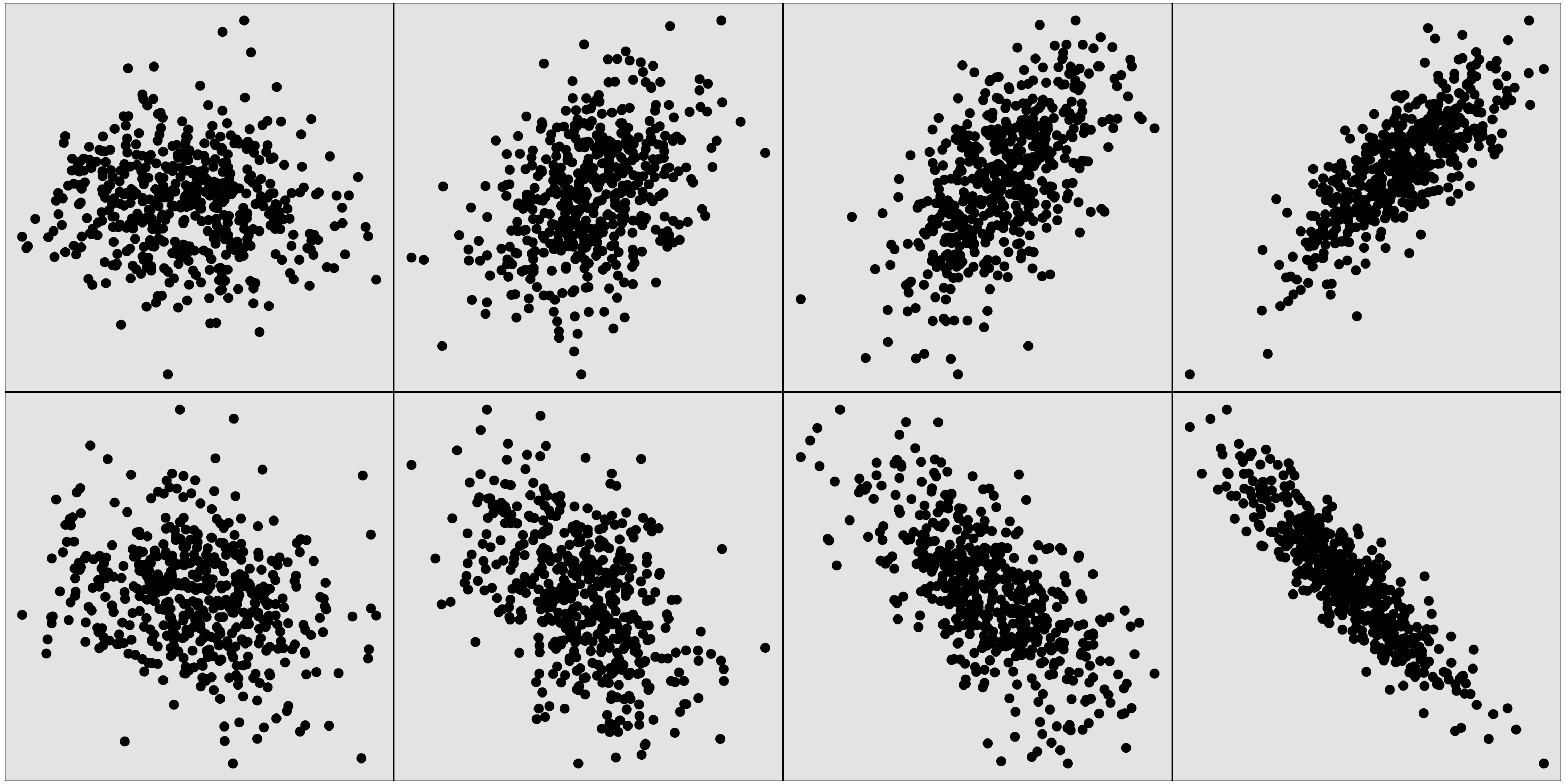
```
$estimate
   cor
-0.012

$statistic
    t
-0.17

$p.value
[1] 0.87
```

It is affected by extreme values.

Perceiving correlation

🖼 answers R

Let's play a game: Guess the correlation!



Robust correlation measures (1/2)

- Spearman (based on ranks)
  - Sort each variable, and return rank (of actual value)
  - Compute correlation between ranks of each variable

```
set.seed(60)
df <- tibble(
  x = c(round(rnorm(5), 1), 10),
  y = c(round(rnorm(5), 1), 10)
) |>
  mutate(xr = rank(x), yr = rank(y))
df
```

```
# A tibble: 6 × 4
      x     y    xr    yr
  <dbl> <dbl> <dbl> <dbl>
1   0.7  -1.7     5     1
2   0.5   1.1     4     5
3  -0.6   0.3     2     3
4  -0.2  -0.9     3     2
5  -1.7   0.4     1     4
6  10    10       6     6
```

```
cor(df$x, df$y)
```

```
[1] 0.94
```

```
cor(df$xr, df$yr)
```

```
[1] 0.2
```

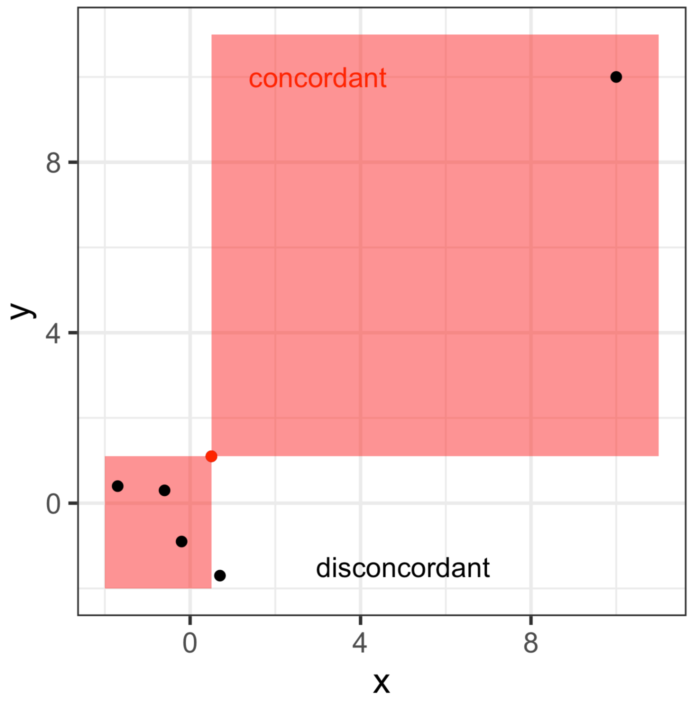
```
cor(df$x, df$y, method = "spearman")
```

```
[1] 0.2
```

Robust correlation measures (2/2)

- Kendall τ (based on comparing pairs of observations)
  - Sort each variable, and return rank (of actual value)
  - For all pairs of observations $(x_i, y_i), (x_j, y_j)$, determine if **concordant**, $x_i < x_j, y_i < y_j$ or $x_i > x_j, y_i > y_j$, or **discordant**, $x_i < x_j, y_i > y_j$ or $x_i > x_j, y_i < y_j$.

$$\tau = \frac{n_c - n_d}{\frac{1}{2}n(n-1)}$$



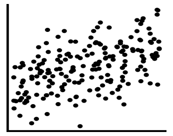
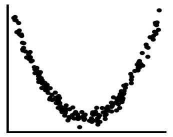
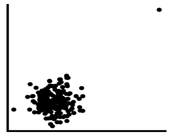

```r
cor(df$x, df$y)
```

```
[1] 0.94
```

```r
cor(df$x, df$y, method = "kendall")
```

```
[1] 0.067
```

Comparison of correlation measures

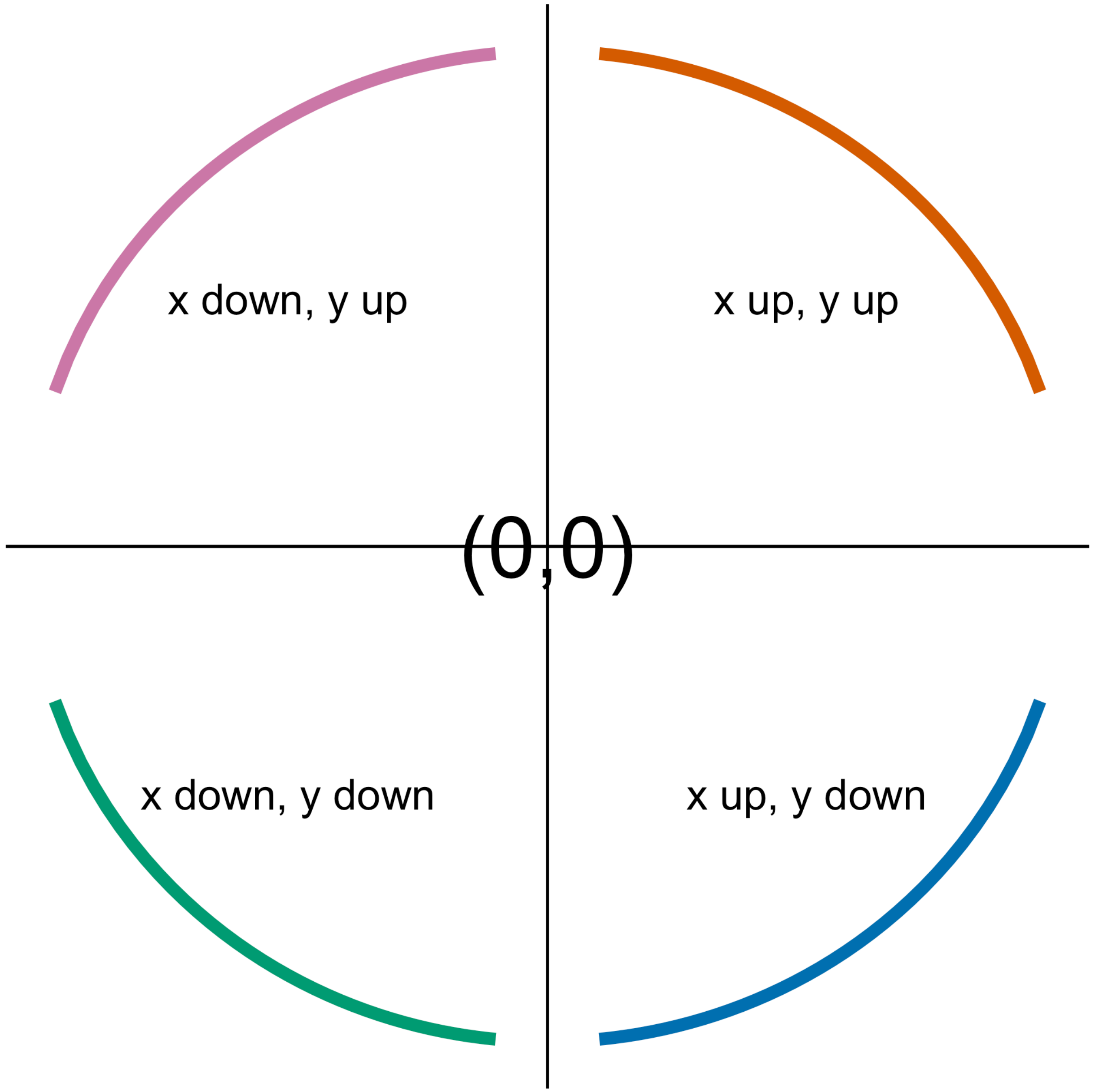
sample	corr	spearman	kendall
	0.52	0.512	0.355
	-0.05	-0.087	-0.073
	0.30	-0.023	-0.014

Robust calculation corrects outlier problems, but nothing measures the non-linear association.

Transformations

for skewness, heteroskedasticity and linearising relationships, and to emphasize association

Circle of transformations for linearising



Remember the power ladder:

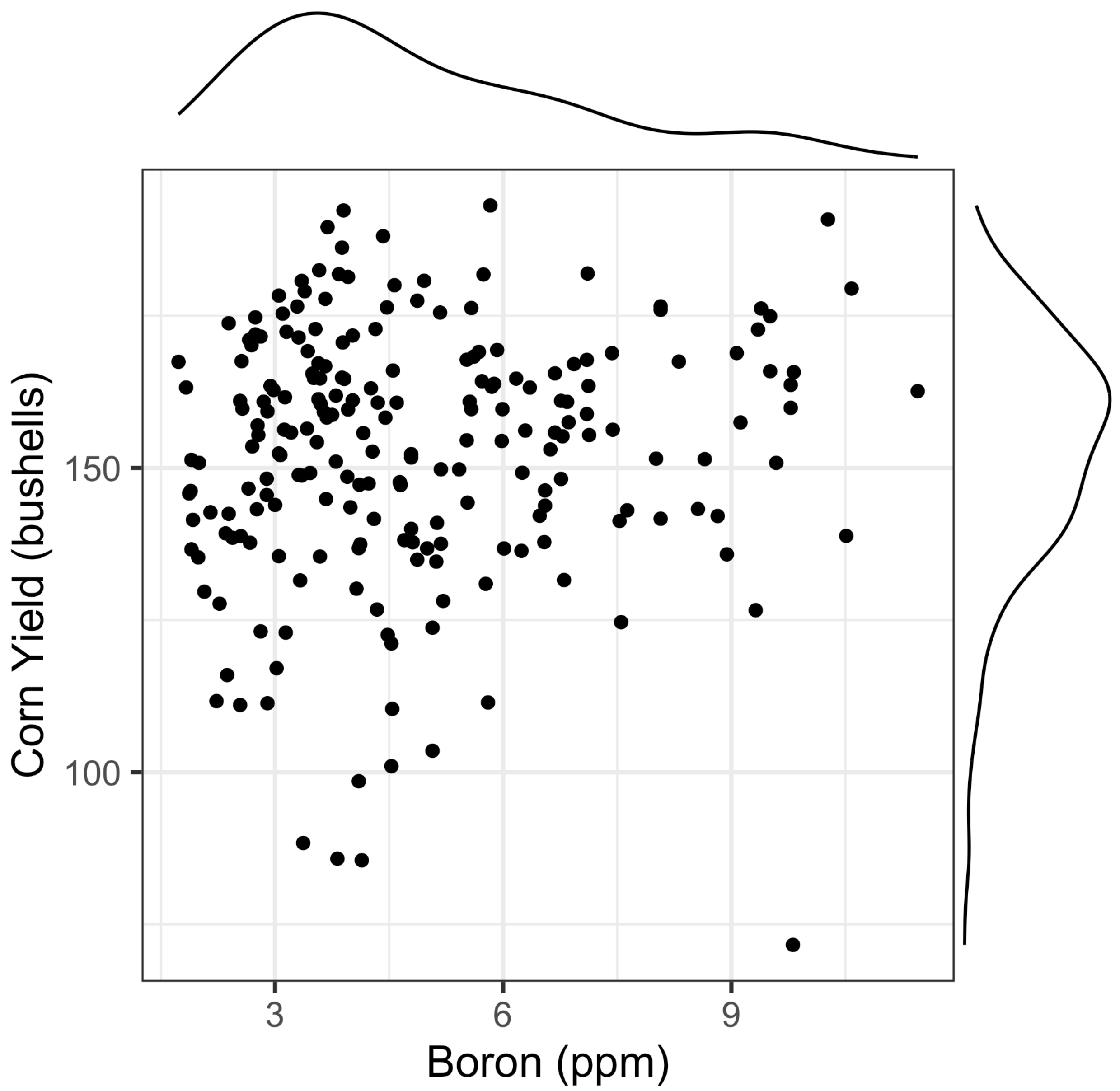
-1, 0, 1/3, 1/2, 1, 2, 3, 4

1. Look at the shape of the relationship.
2. Imagine this to be a number plane, and depending on which quadrant the shape falls in, you either transform x or y, up or down the ladder: +,+ both up; +,− x up, y down; −,− both down; −,+ x down, y up

If there is heteroskedasticity, try transforming y, may or may not help

Scatterplot case studies

Case study: Soils (1/4)



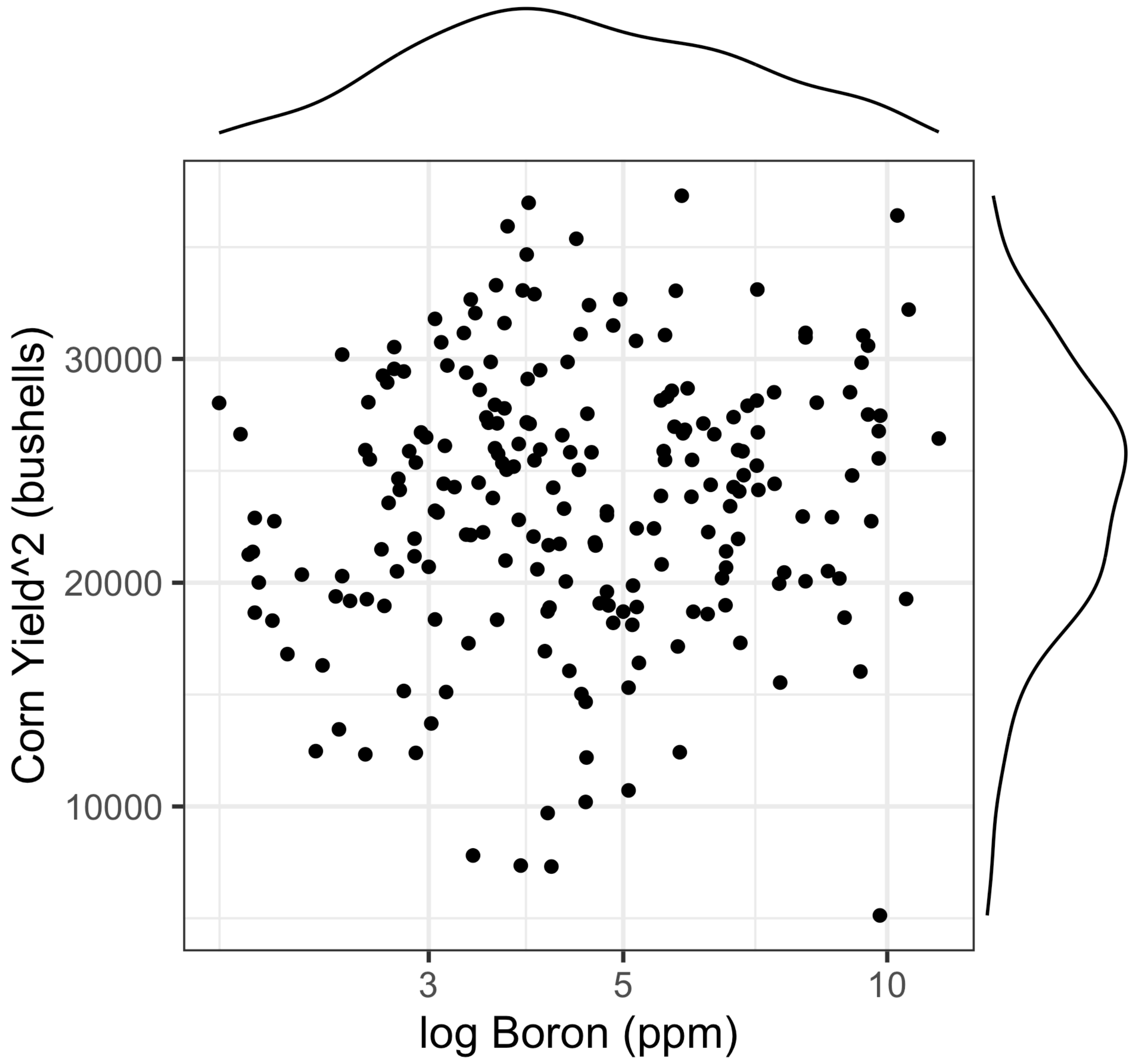
Interplay between skewness and association

Data is from a soil chemical analysis of a farm field in Iowa. Is there a relationship between Yield and Boron?

You can get a marginal plot of each variable added to the scatterplot using `ggMarginal`. This is useful for assessing the skewness in each variable.

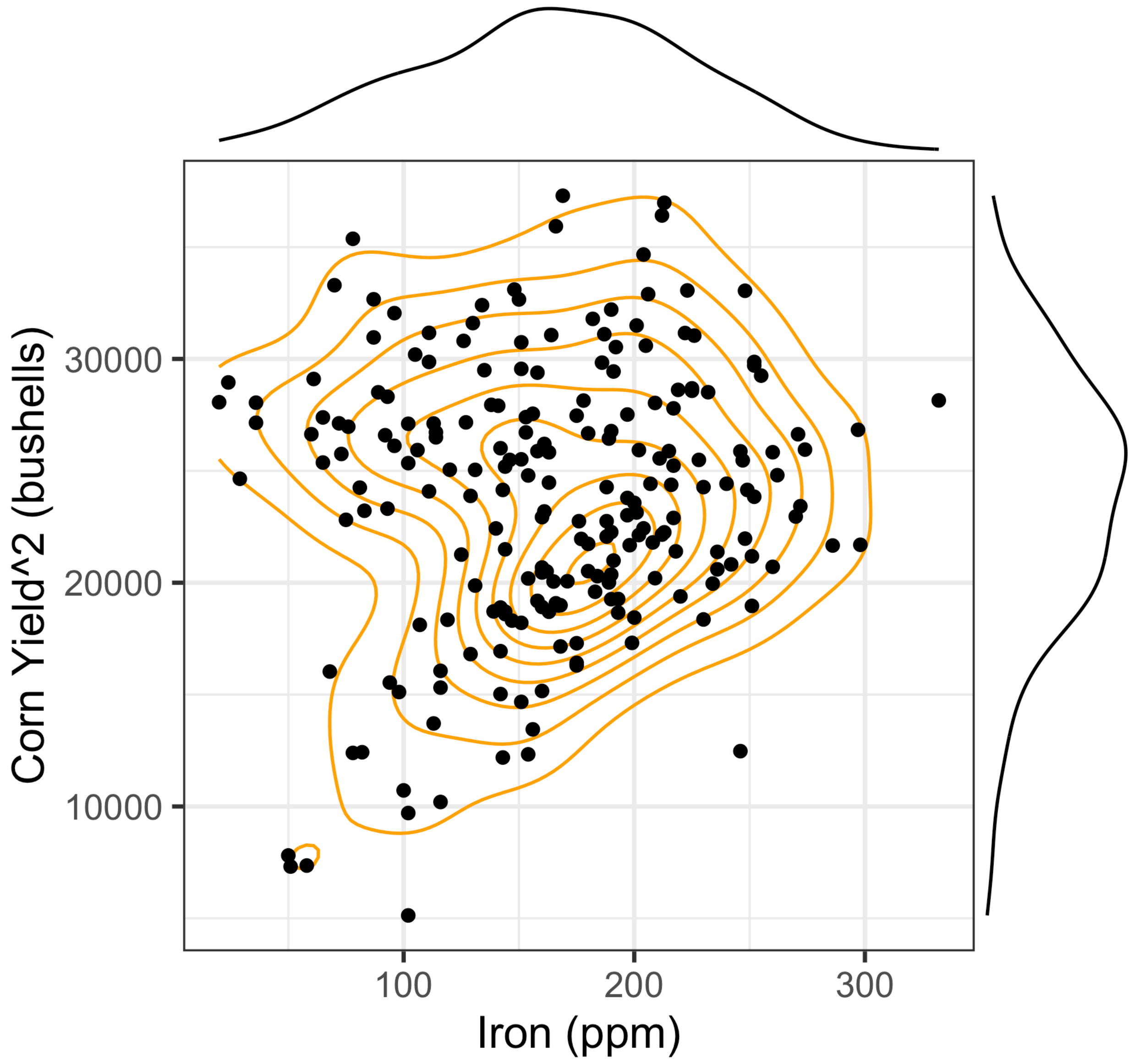
Boron is right-skewed Yield is left-skewed. With skewed distributions in marginal variables it is hard to assess the relationship between the two. Make a transformation to fix, first.

Case study: Soils (2/4)




```r
p <- ggplot(
  baker,
  aes(x = B, y = Corn97BU^2)
) +
  geom_point() +
  xlab("log Boron (ppm)") +
  ylab("Corn Yield^2 (bushells)") +
  scale_x_log10()
ggMarginal(p, type = "density")
```

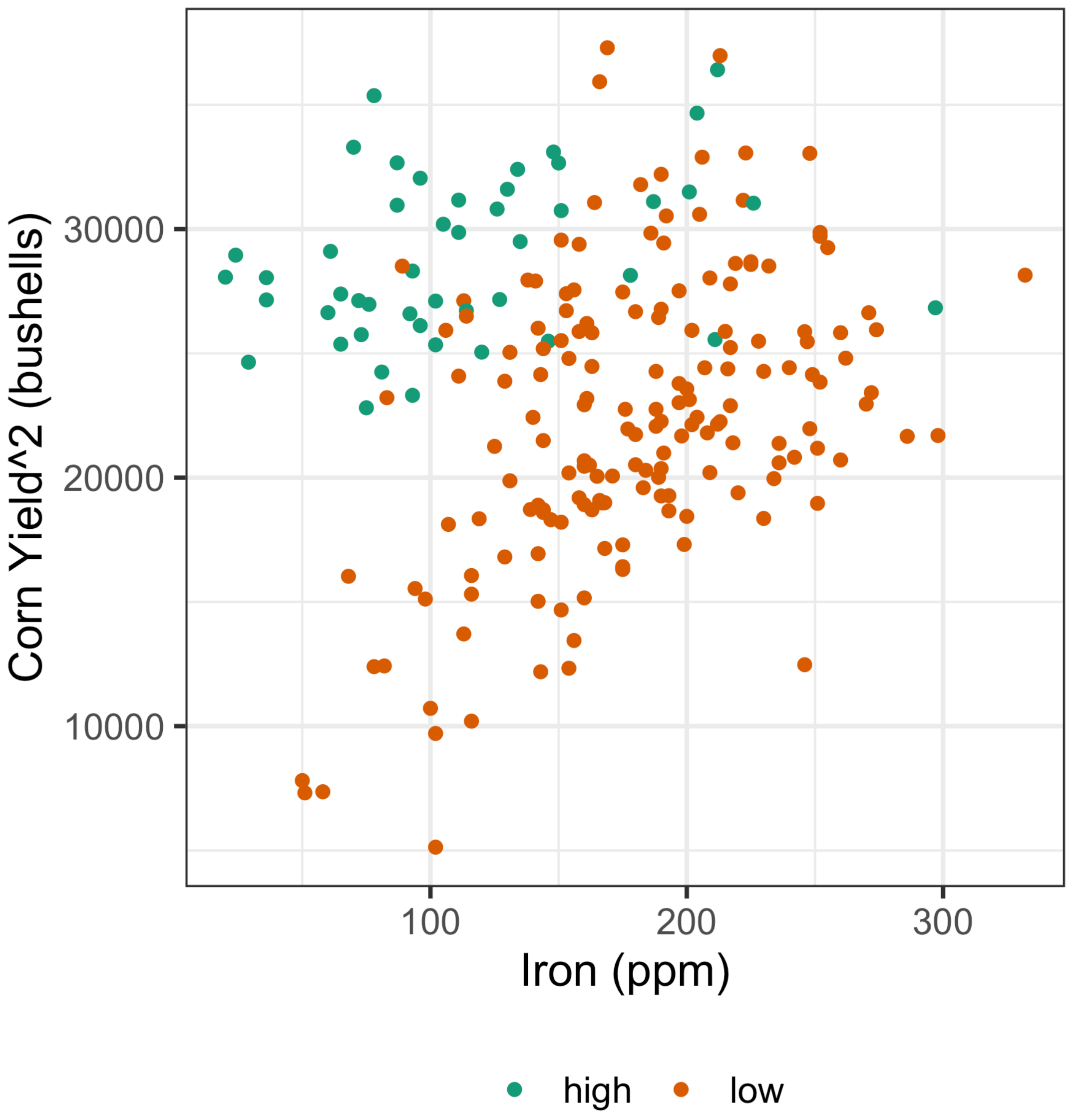
Case study: Soils (3/4)



Lurking variable?

```
p <- ggplot(
  baker,
  aes(x = Fe, y = Corn97BU^2)
) +
  geom_density2d(colour = "orange") +
  geom_point() +
  xlab("Iron (ppm)") +
  ylab("Corn Yield^2 (bushells)")
ggMarginal(p, type = "density")
```

Case study: Soils (4/4)



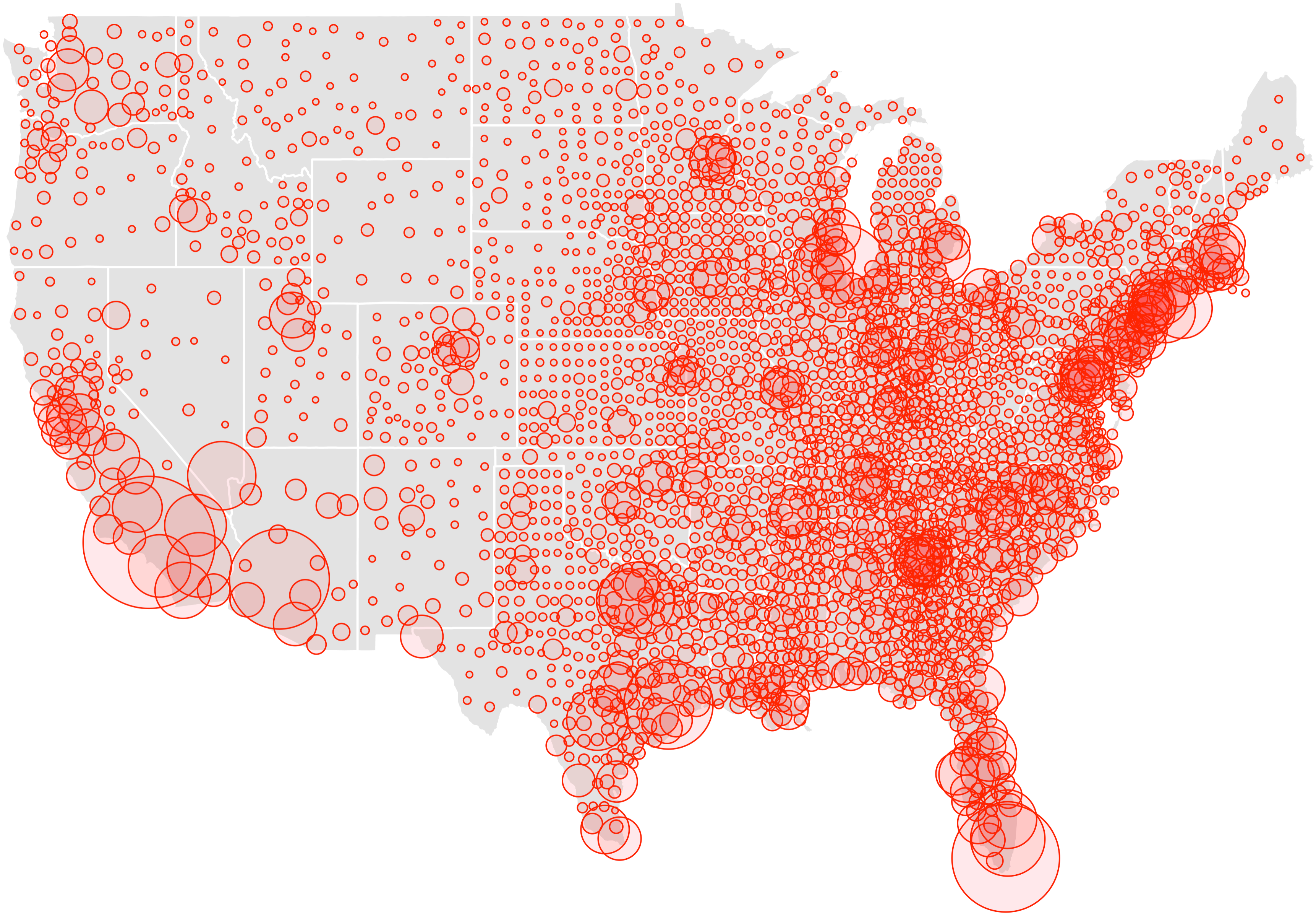
Colour high calcium (>5200ppm) calcium values

```
ggplot(baker, aes(
  x = Fe, y = Corn97BU^2,
  colour = ifelse(Ca > 5200,
    "high", "low"
  )
)) +
  geom_point() +
  xlab("Iron (ppm)") +
  ylab("Corn Yield^2 (bushells)") +
  scale_colour_brewer("", palette = "Dark2") +
  theme(
    aspect.ratio = 1,
    legend.position = "bottom",
    legend.direction = "horizontal"
  )
```

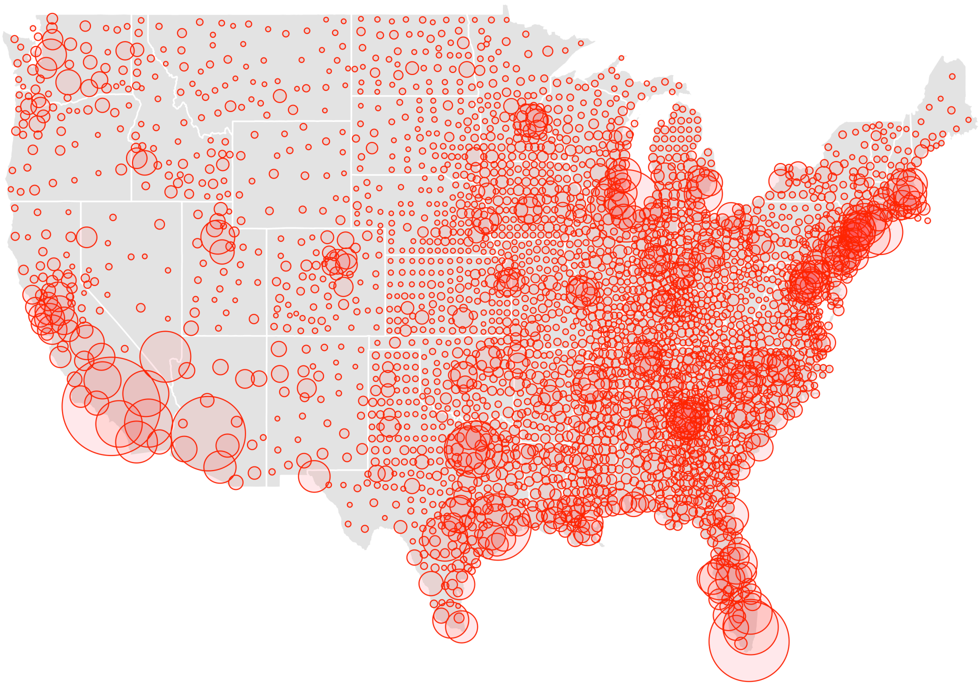
If calcium levels in the soil are high, yield is consistently high. If calcium levels are low, then there is a positive relationship between yield and iron, with higher iron leading to higher yields.

Case study: COVID-19

🖼️ info R



Scales matter



Where has COVID-19 hit the hardest?

Where are there more people?

This plot tells you NOTHING except where the population centres are in the USA.

To understand relative incidence/risk, report COVID numbers relative the population. For example, number of cases per 100,000 people.

Beyond quantitative variables

When variables are not quantitative

What do you do if the variables are not continuous/quantitative?

Type of variable determines the appropriate mapping.

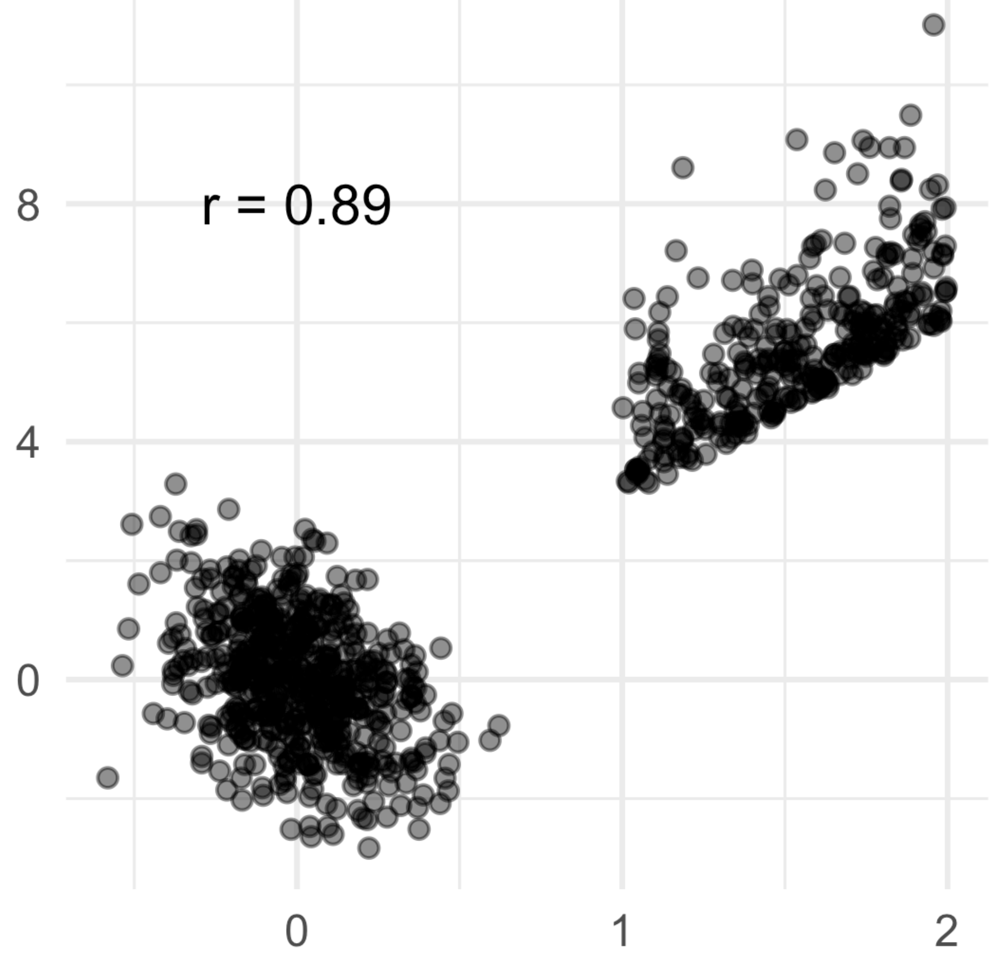
- Continuous and categorical: side-by-side boxplots, side-by-side density plots
- Both categorical: faceted bar charts, stacked bar charts, mosaic plots, double decker plots

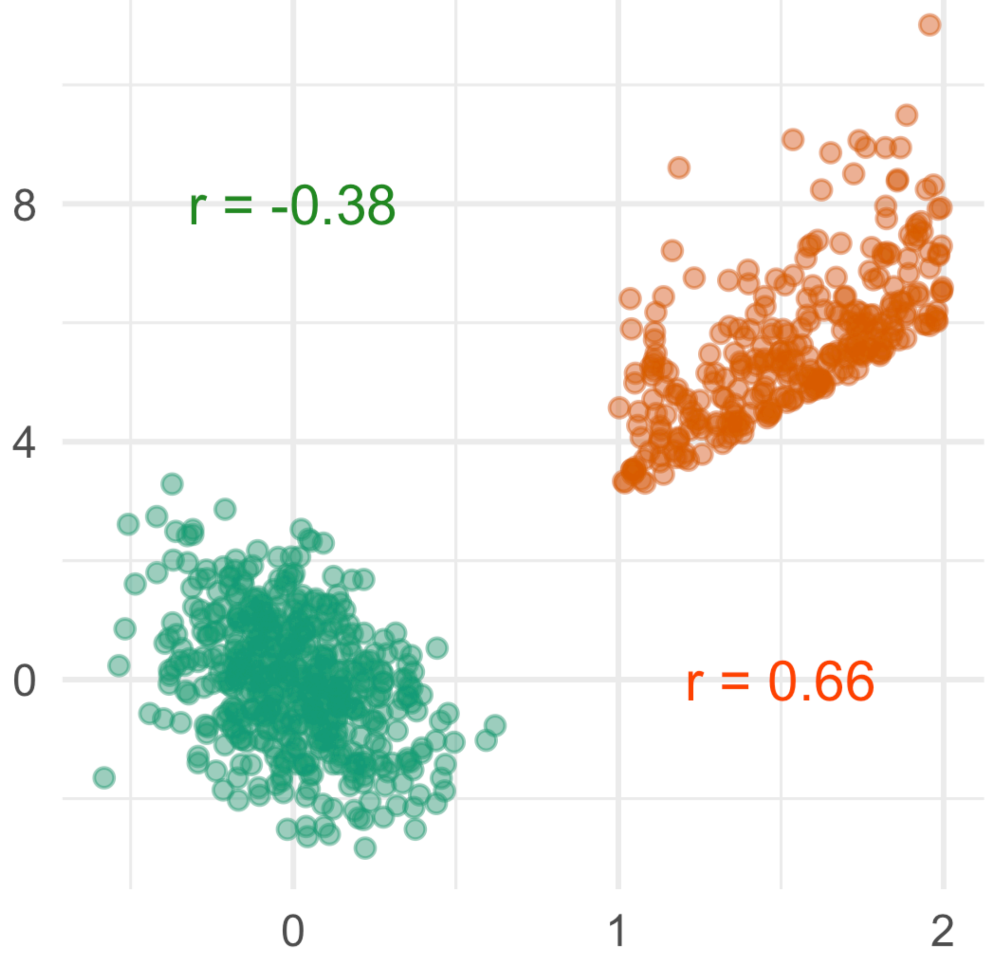
Stay tuned!

Paradoxes

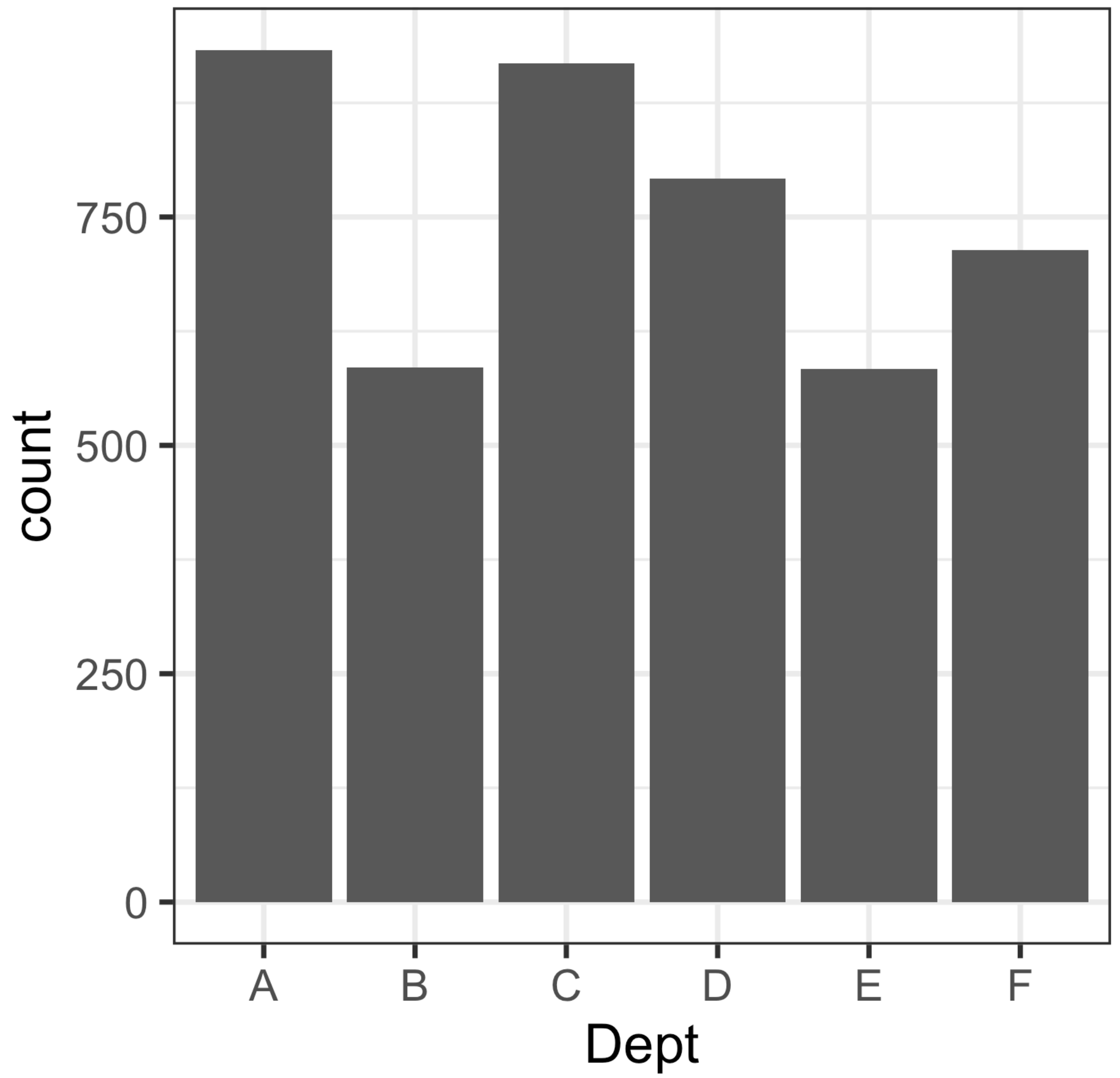
Simpsons paradox

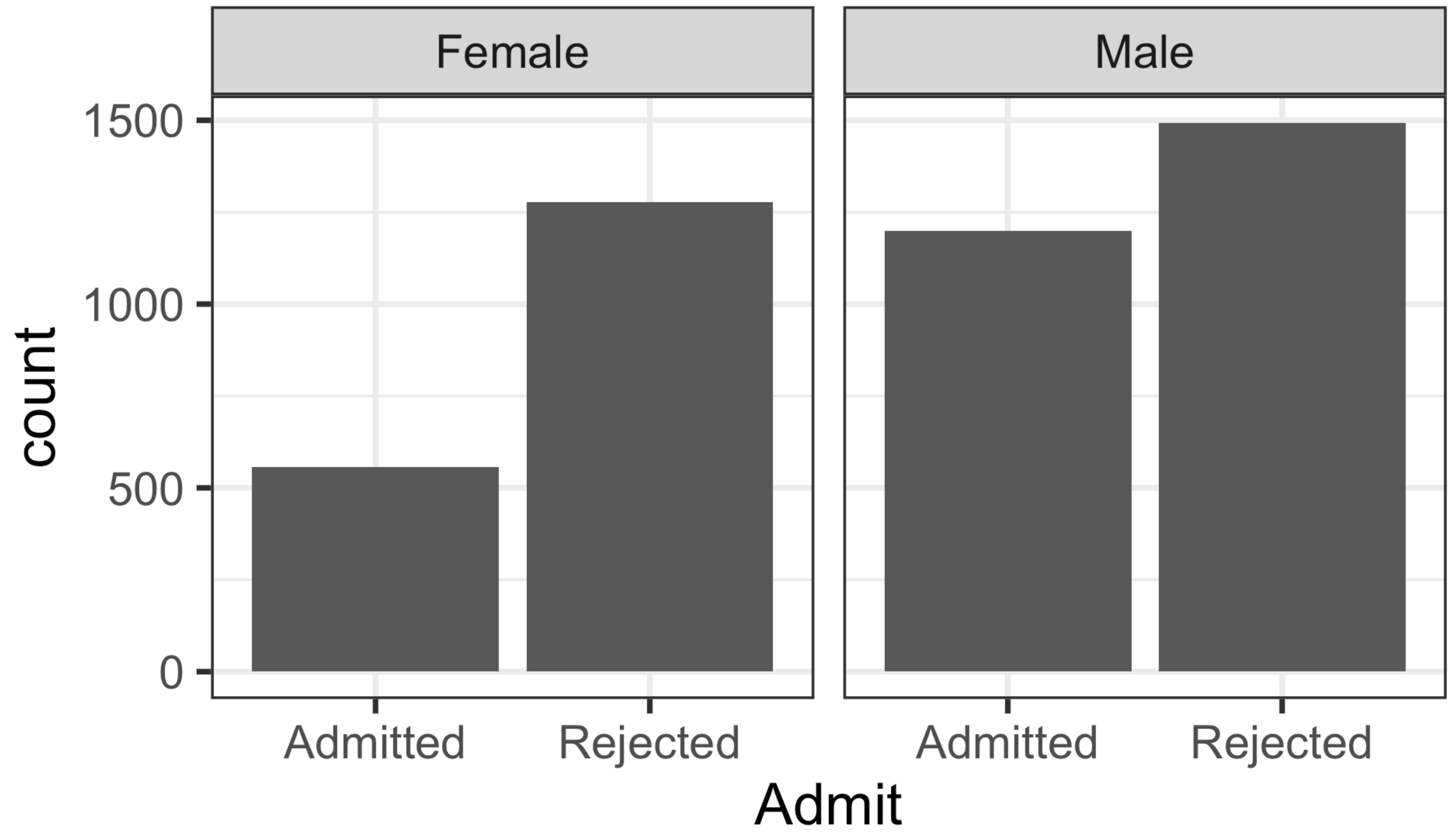
There is an additional variable, which if used for conditioning, changes the association between the variables, you have a paradox.





Simpsons paradox: famous example

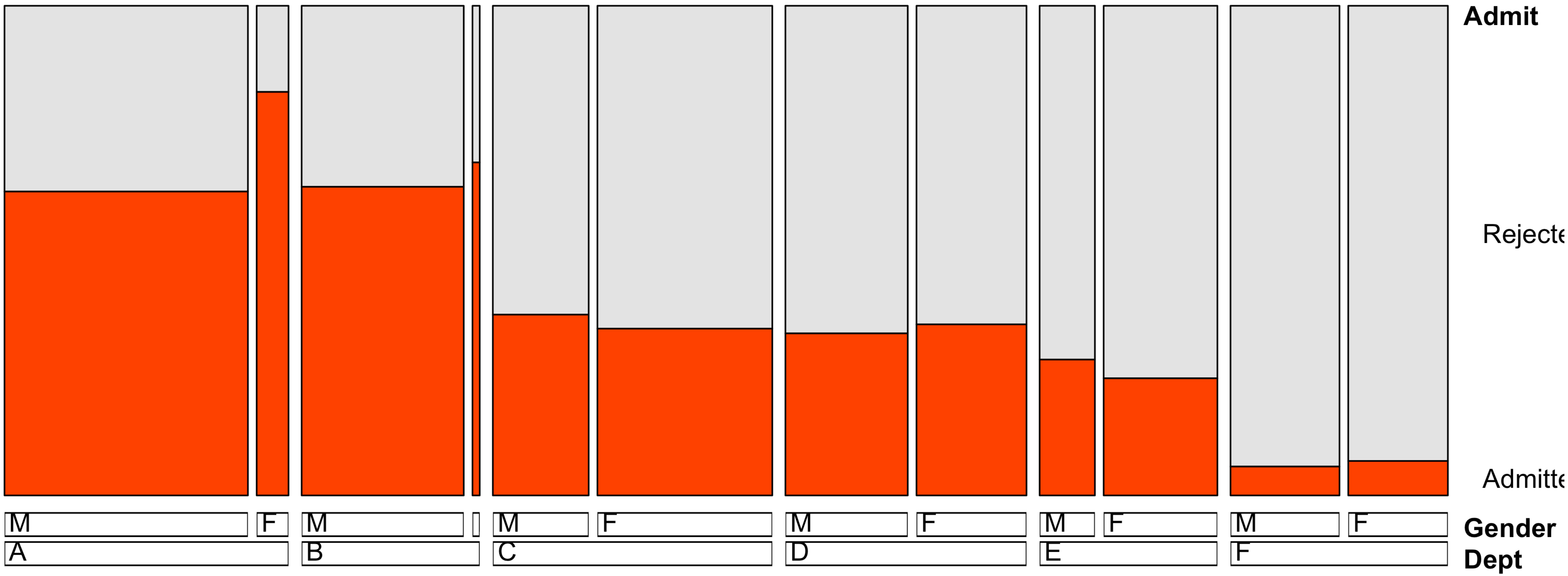




Did Berkeley discriminate against female applicants?

Example from Unwin (2015)

Simpsons paradox: famous example



Based on separately examining each department, there is no evidence of discrimination against female applicants.

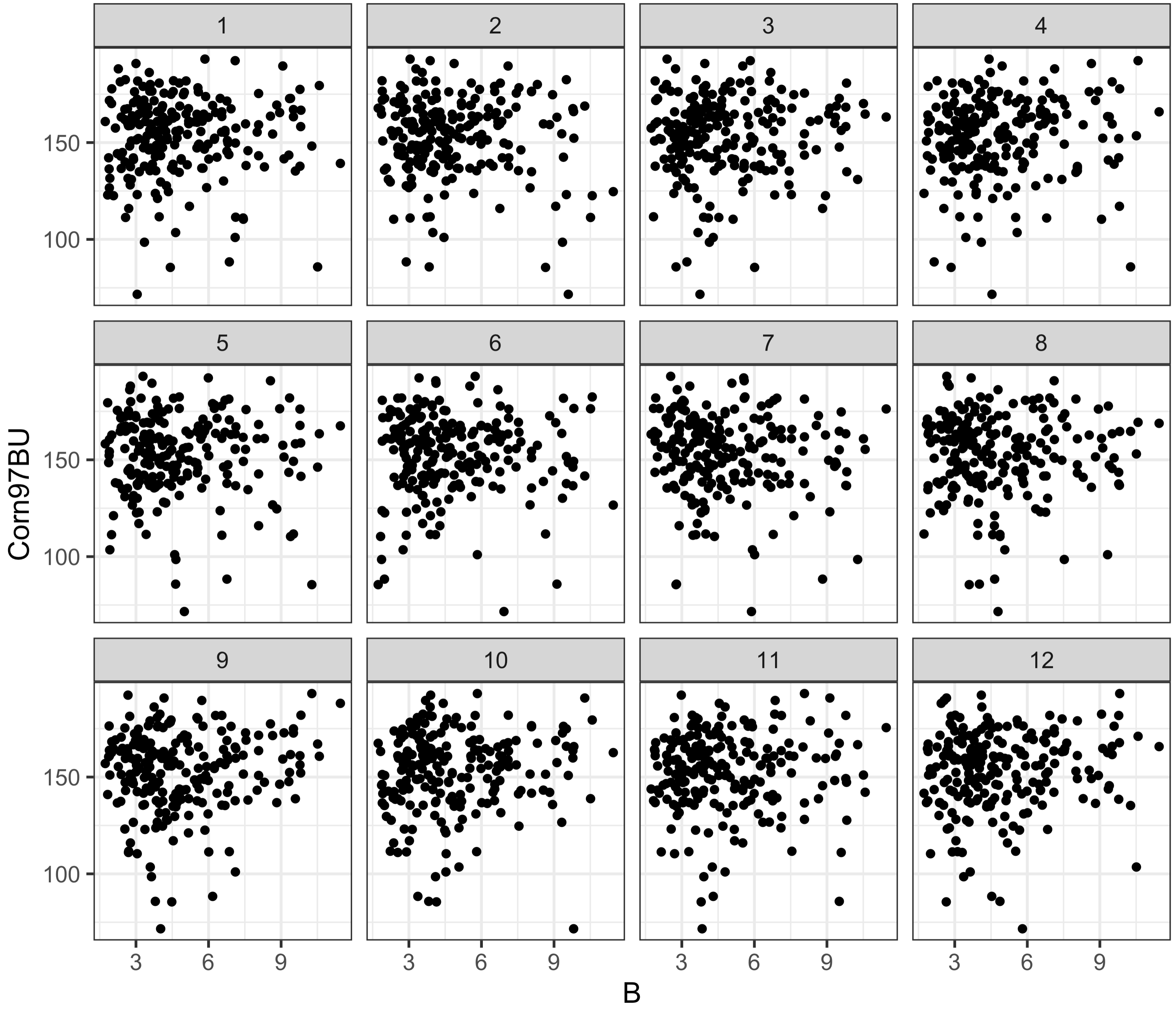
Example from Unwin (2015)

Always examine the associations in each strata

Is what you see really association?

Checking association with visual inference

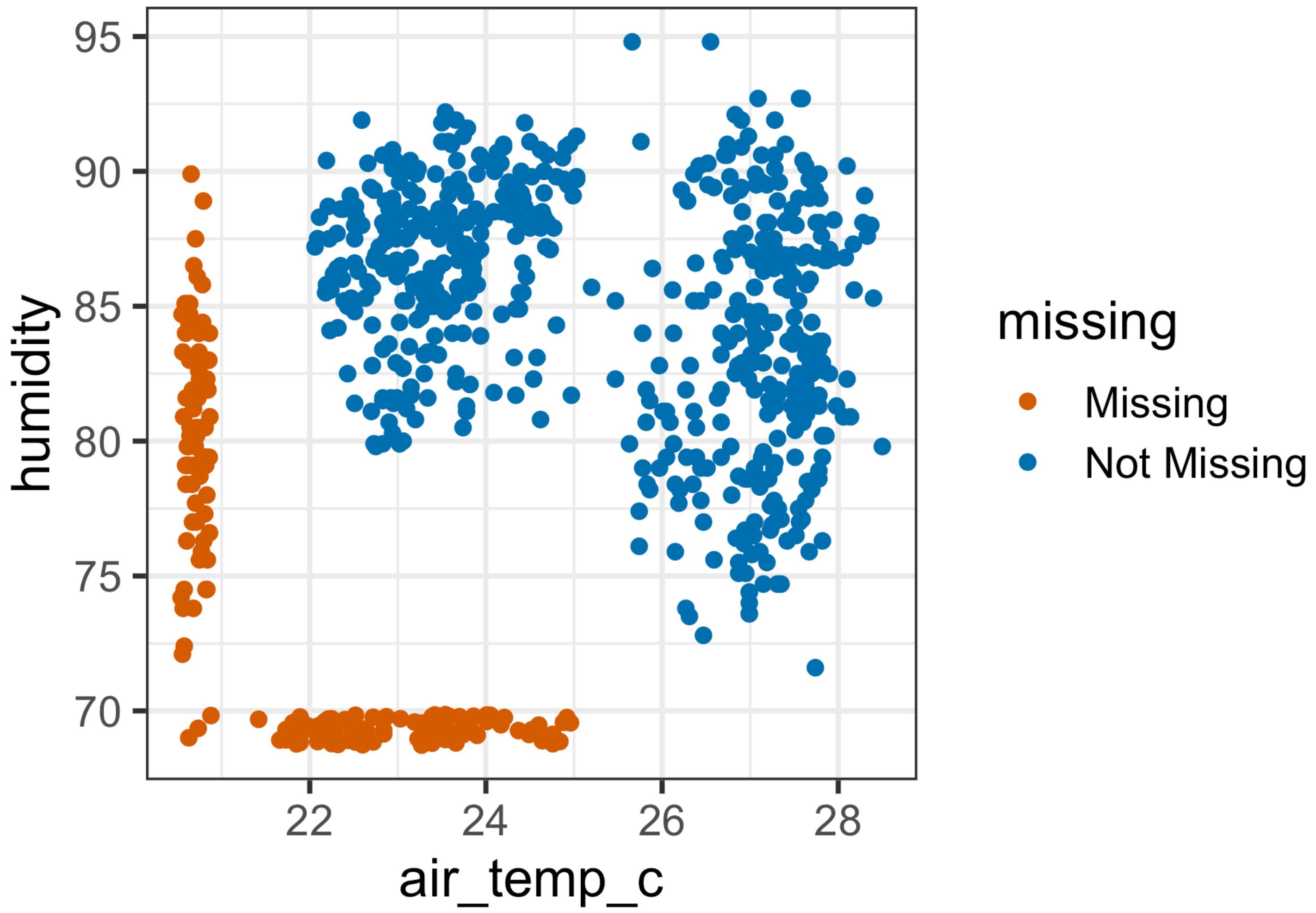
Soils | R | Olympics | R



Handling and imputing missings (1/2)

Check if missings on one variable are related to distribution of the other variable.

▶ Code

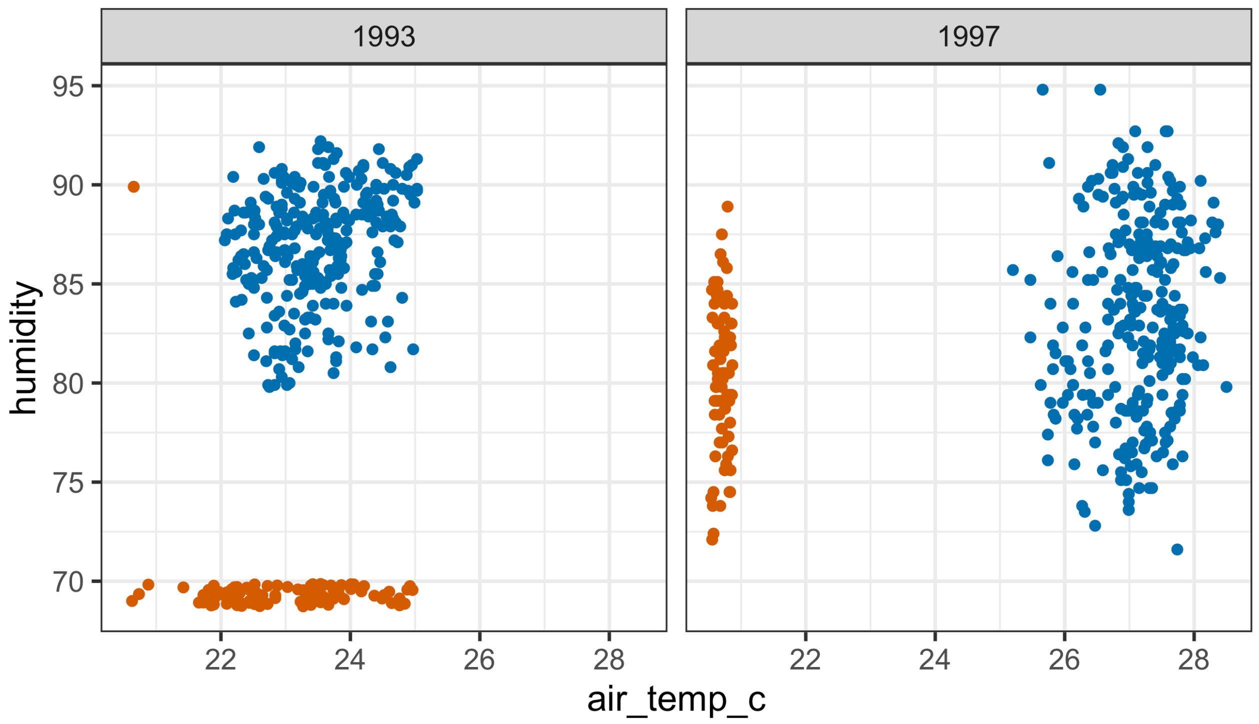


- Missings plotted in the margins.
- Missings on humidity only occur for lower values of air

Imputing missings, at least for humidity requires using air temperature values.

But the clustering is due to year

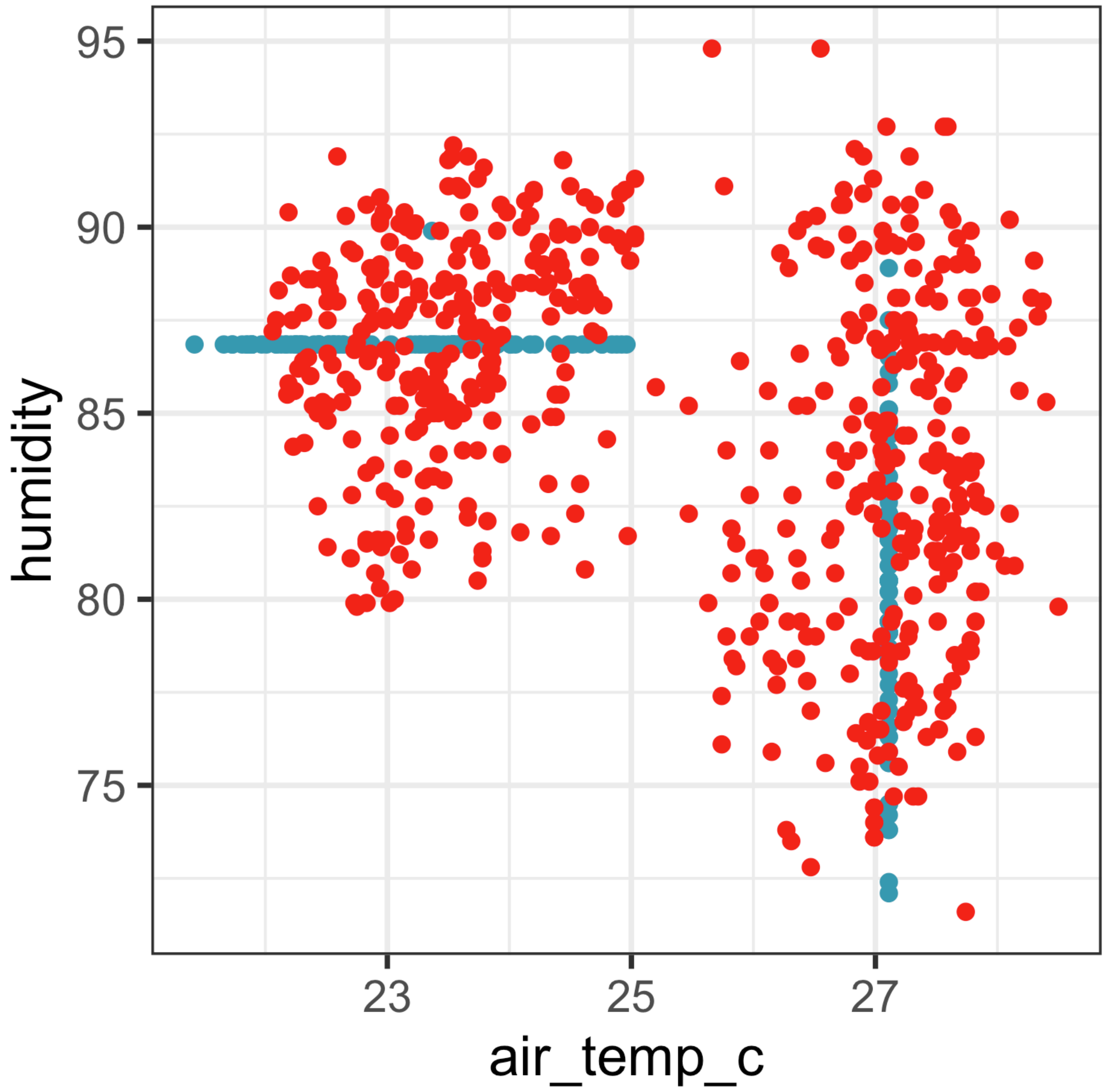
▶ Code



Handling and imputing missings (2/2)

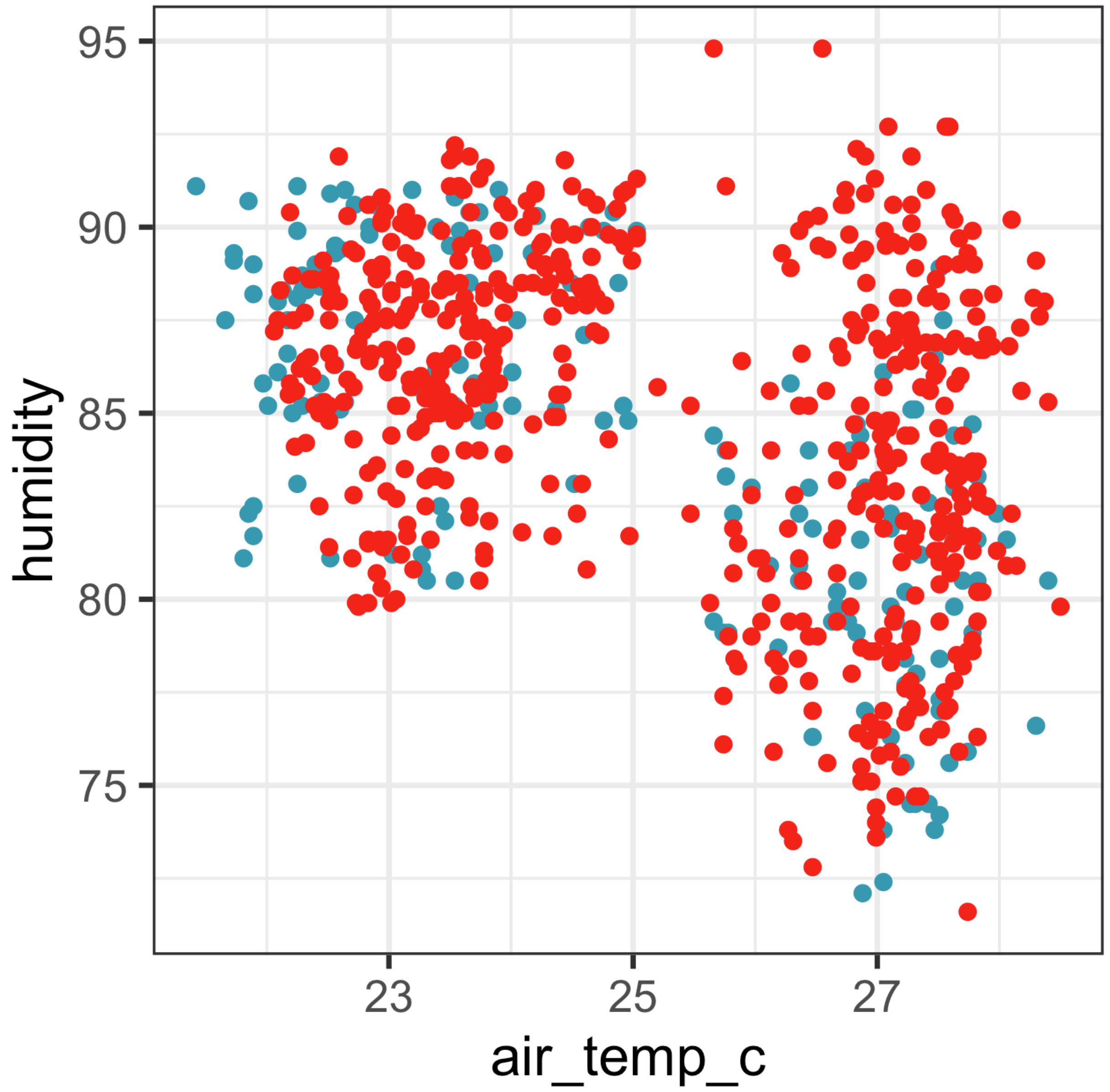
Use the mean of complete cases to impute the missings

▶ Code



Use simulation from a bivariate normal distribution, for each year.

▶ Code



Resources

- Unwin (2015) Graphical Data Analysis with R
- Graphics using ggplot2
- Wilke (2019) Fundamentals of Data Visualization https://clauswilke.com/dataviz/
- Friendly and Denis “Milestones in History of Thematic Cartography, Statistical Graphics and Data Visualisation” available at http://www.datavis.ca/milestones/
- Tierney et al (2023) Expanding Tidy Data Principles to Facilitate Missing Data Exploration, Visualization and Assessment of Imputations.