ETC5521: Exploratory Data Analysis

Exploring bivariate dependencies

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 6 - Session 2

Numerical measures of association

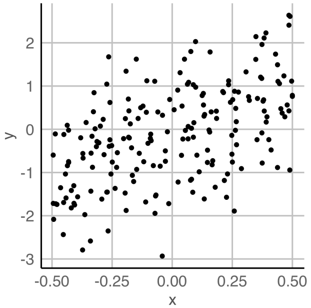
Correlation

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

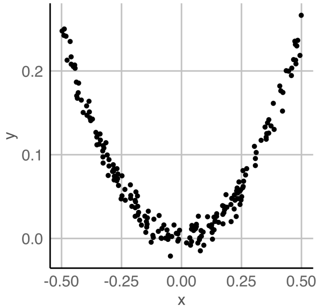
$$r = \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2}} = \frac{\text{covariance}(x_1, x_2)}{(n-1)s_{x_1} s_{x_2}}$$

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

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




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

## [1] 0.5228401

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

##
##      Pearson's product-moment correlation
##
## data:  d1$x and d1$y
## t = 8.6306, df = 198, p-value = 1.993e-15
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.4141406 0.6168362
## sample estimates:
##       cor
## 0.5228401
```

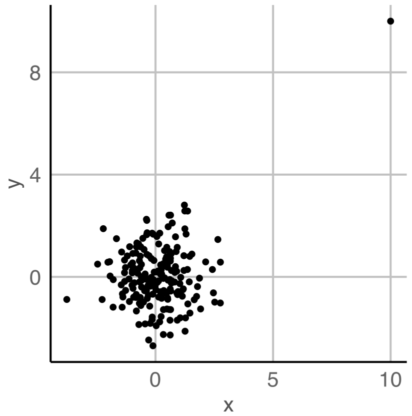



```r
cor(d2$x, d2$y)
```

```
## [1] -0.04993755
```

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

```
##
##      Pearson's product-moment correlation
##
## data:  d2$x and d2$y
## t = -0.70356, df = 198, p-value = 0.4825
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.18738032  0.08942303
## sample estimates:
##         cor
## -0.04993755
```



All observations

```
## $estimate
##       cor
## 0.2994041
##
## $statistic
##        t
## 4.426682
##
## $p.value
## [1] 1.576086e-05
```

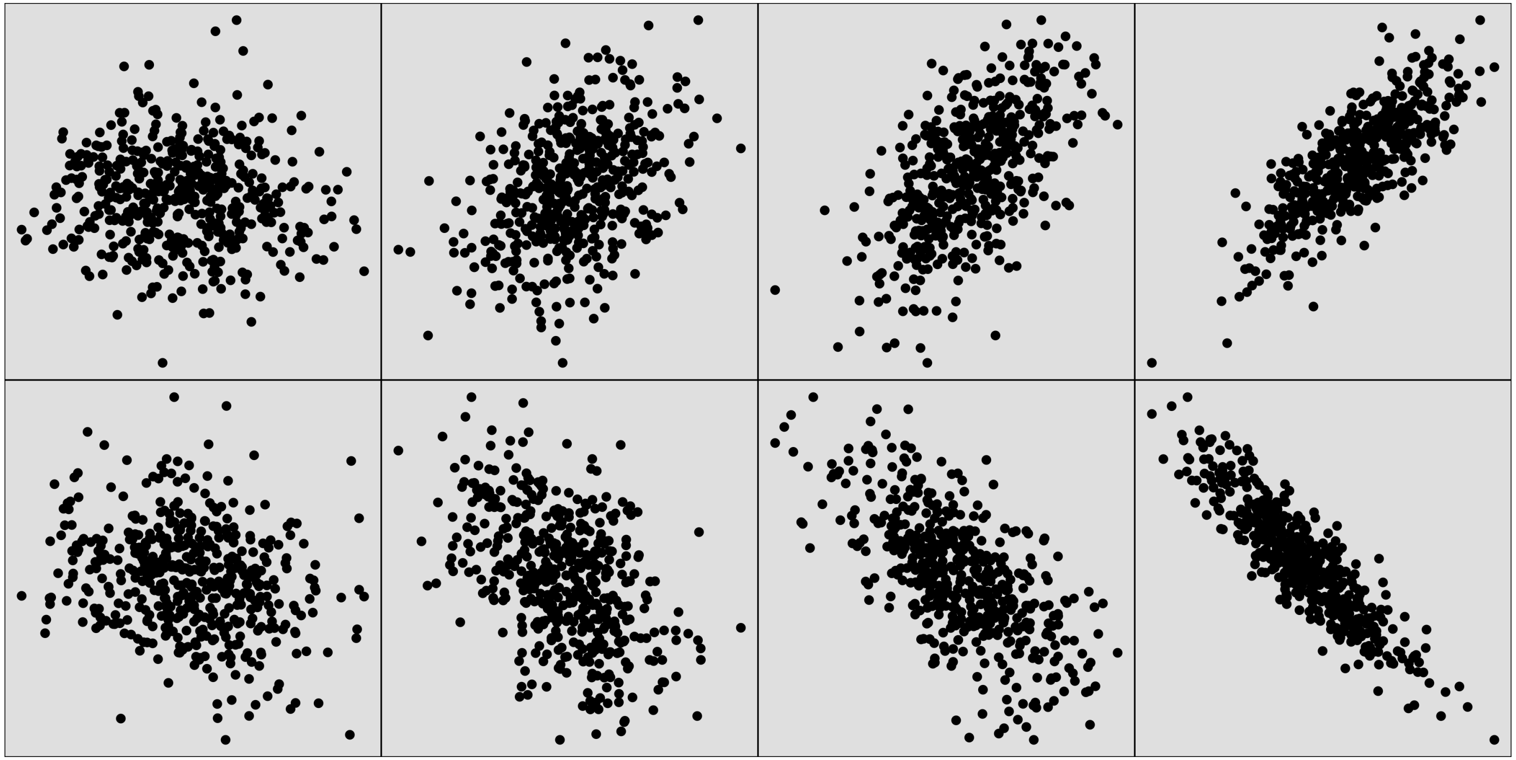
Without outlier

```
## $estimate
##         cor
## -0.01173776
##
## $statistic
##          t
## -0.1651764
##
## $p.value
## [1] 0.8689737
```

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

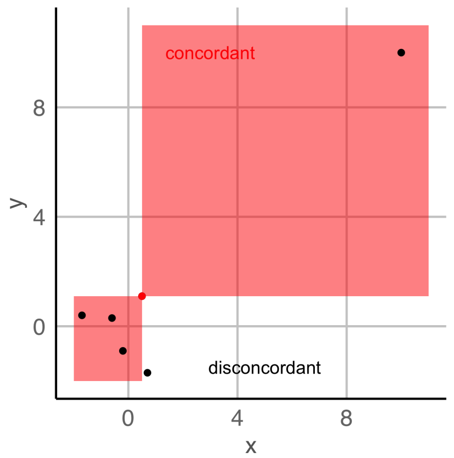
```
## # A tibble: 6 × 4
##         x      y     xr     yr
##     <dbl>  <dbl>  <dbl>  <dbl>
## 1     0.7   -1.7      5      1
## 2     0.5    1.1      4      5
## 3    -0.6    0.3      2      3
## 4    -0.2   -0.9      3      2
## 5    -1.7    0.4      1      4
## 6    10     10        6      6
```

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

## [1] 0.935397

cor(df$xr, df$yr)

## [1] 0.2

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

## [1] 0.2
```

Robust correlation measures 2/2

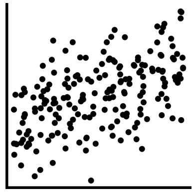
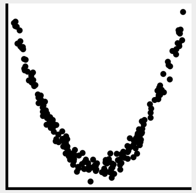
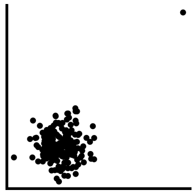
- Kendall τ (based on comparing pairs of observations)
  - Sort each variable, and return rank (of actual value)
  - For all pairs of observations $(x_i, y_i), (x_j, y_j)$, determine if **concordant**, $x_i < x_j, y_i < y_j$ or $x_i > x_j, y_i > y_j$, or **discordant**, $x_i < x_j, y_i > y_j$ or $x_i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

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




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

## [1] 0.935397

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

## [1] 0.06666667
```

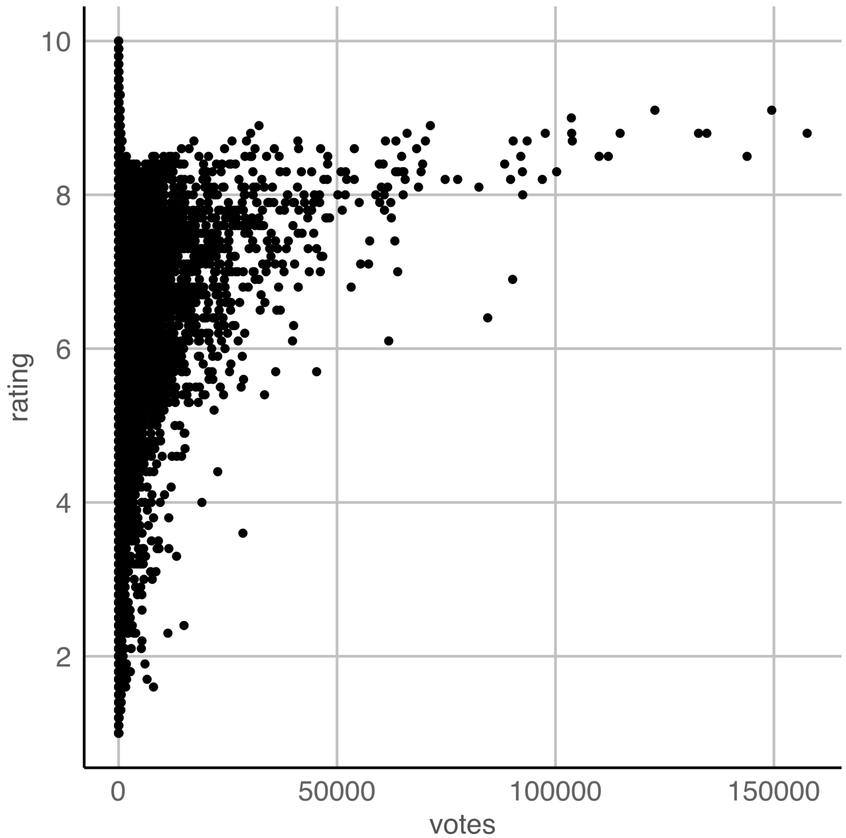
Comparison of correlation measures

sample	corr	spearman	kendall
	0.523	0.512	0.355
	-0.050	-0.087	-0.073
	0.299	-0.023	-0.014

Scatterplot case studies

Case study 2 Movies

learn R



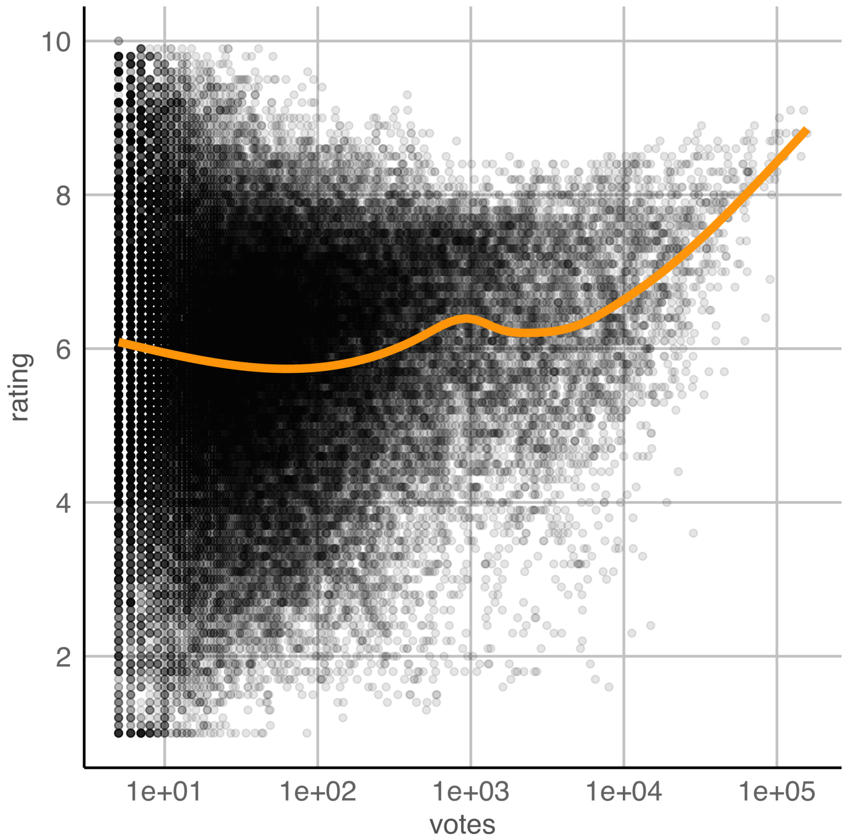
- `votes`: Number of IMDB users who rated this movie
- `rating`: Average IMDB user rating

Describe the relationship between rating and votes.

01:00

Case study 2 Movies

R

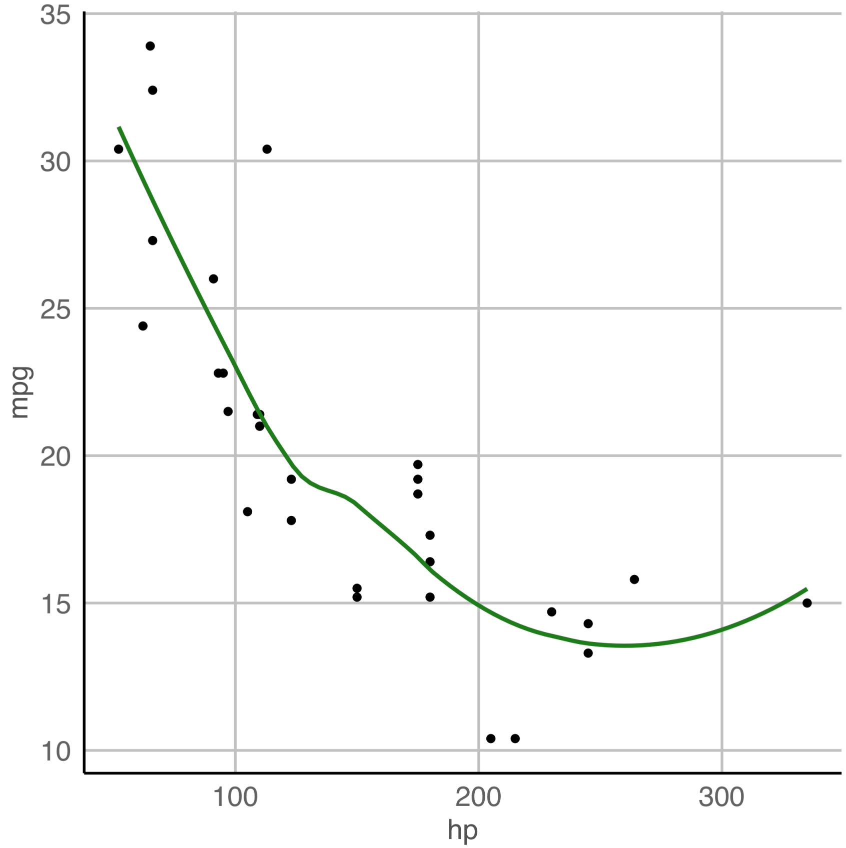


🤔 Something funny happens, right at 1000 votes

Some positive association between two variables only for large number of votes.

Case study 3 Cars

learn R

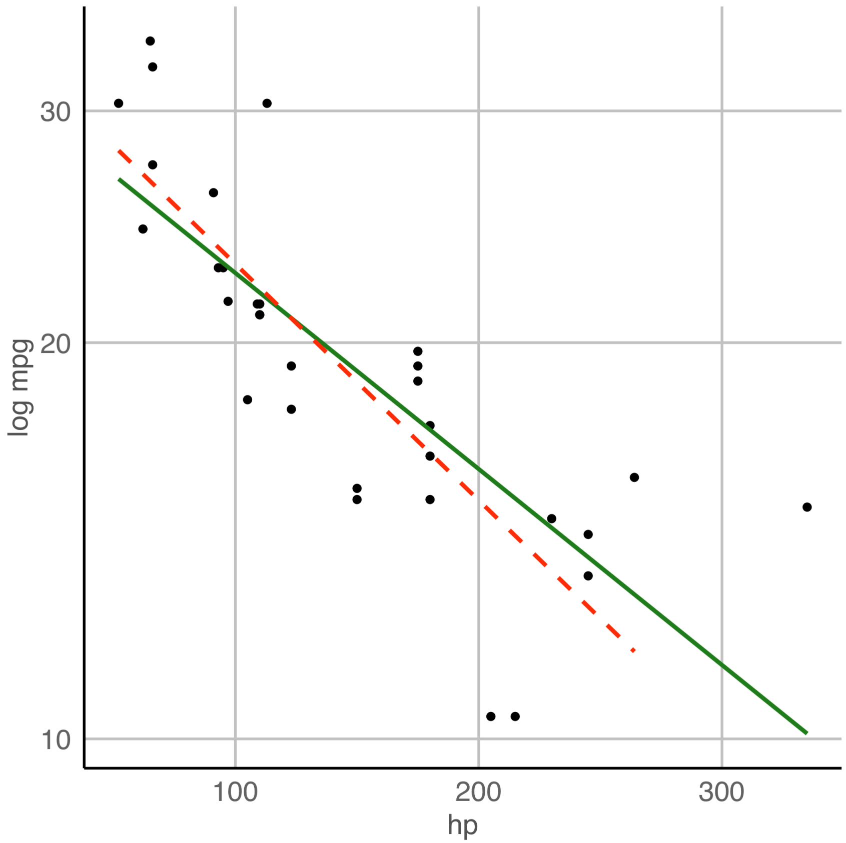


- `mpg`: Miles/(US) gallon
- `hp`: Gross horsepower

Describe the relationship between horsepower and mpg.

Case study 3 Cars

🖼️ R



- `mpg`: Miles/(US) gallon
- `hp`: Gross horsepower

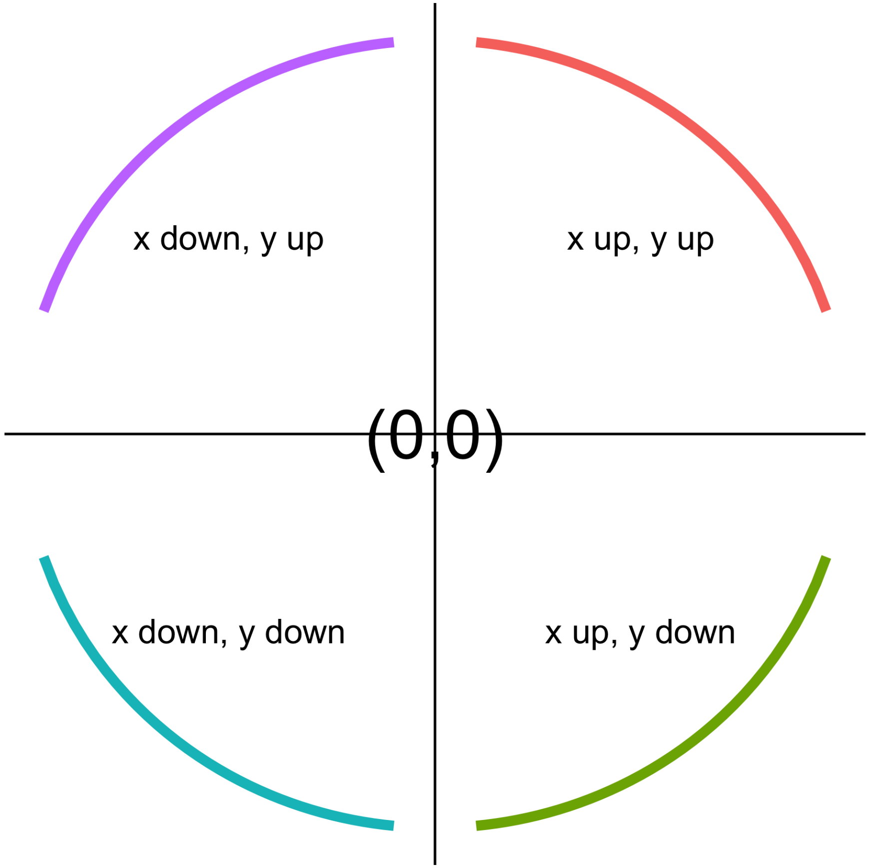
Log transforming `mpg` linearised the relationship between horsepower and mpg.

Need to also remove the outlier, because it is a little influential (swinging the line towards it).

Transformations

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

Circle of transformations for linearising



Remember the power ladder:

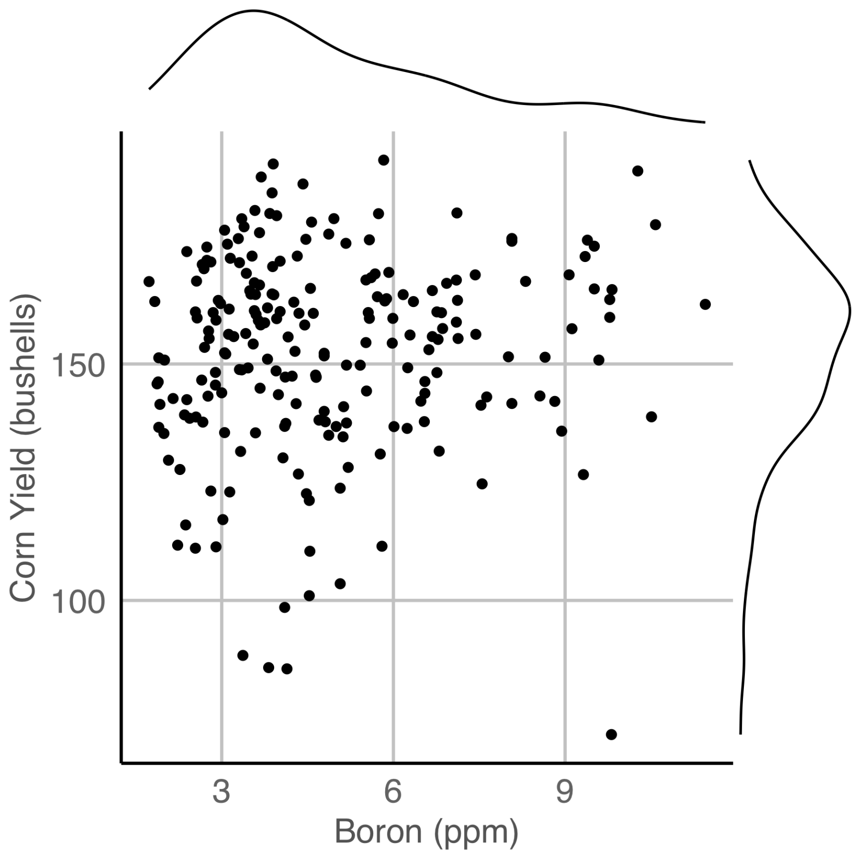
-1, 0, 1/3, 1/2, 1, 2, 3, 4

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 4 Soils



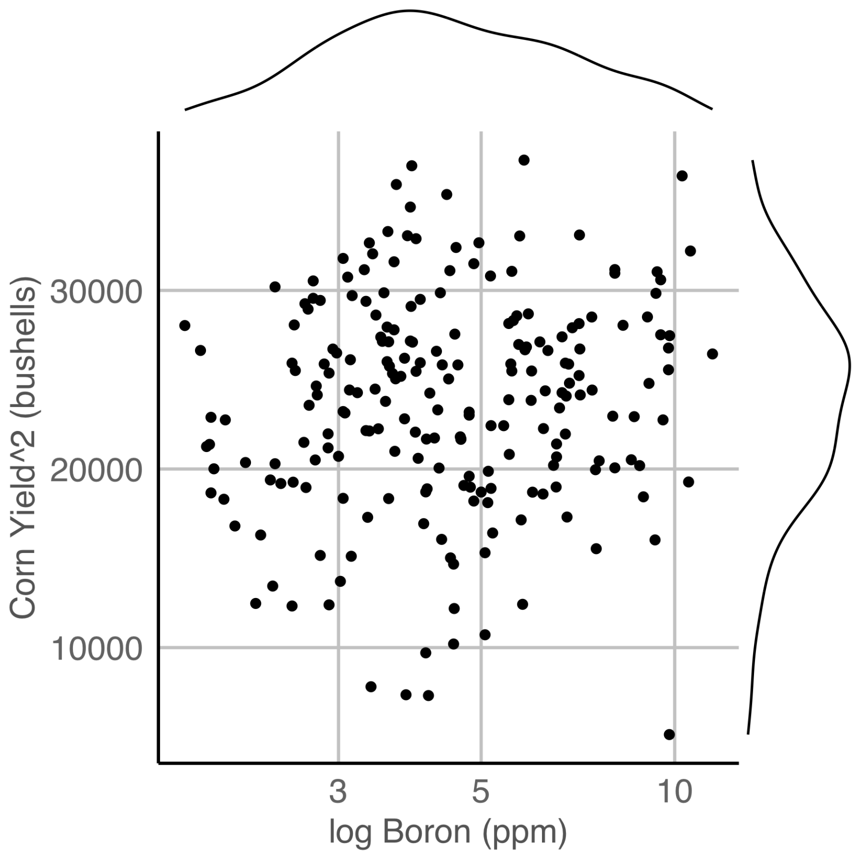
Interplay between skewness and association

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.

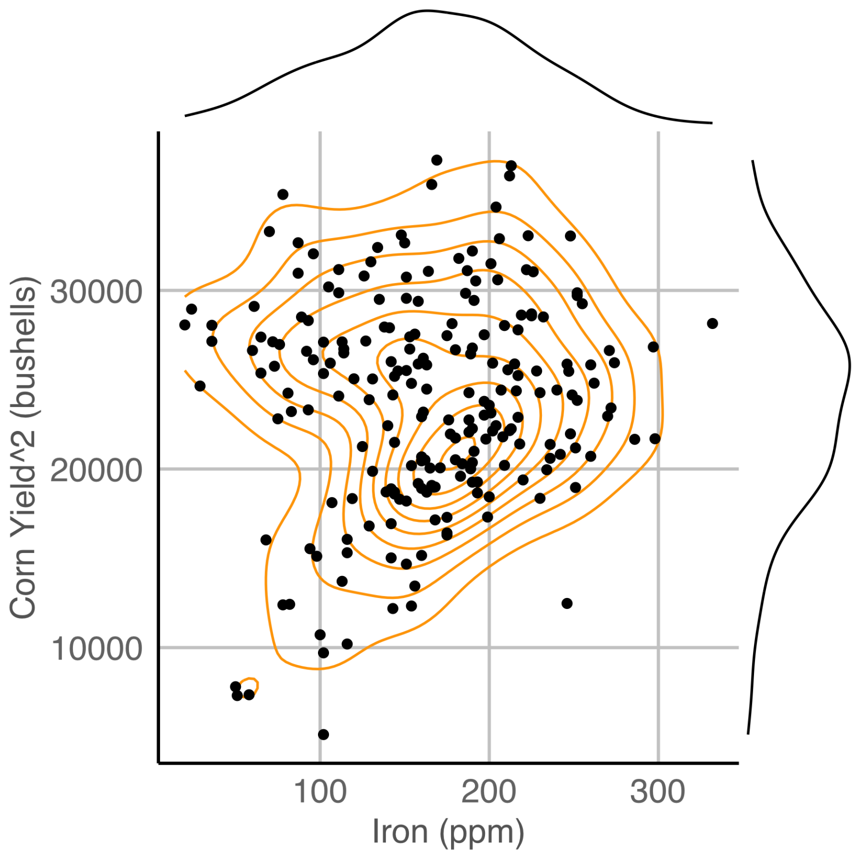
Boron is right-skewed Yield is left-skewed. With skewed distributions in marginal variables it is hard to assess the relationship between the two. Make a transformation to fix, first.

Case study 4 Soils




```
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")
```

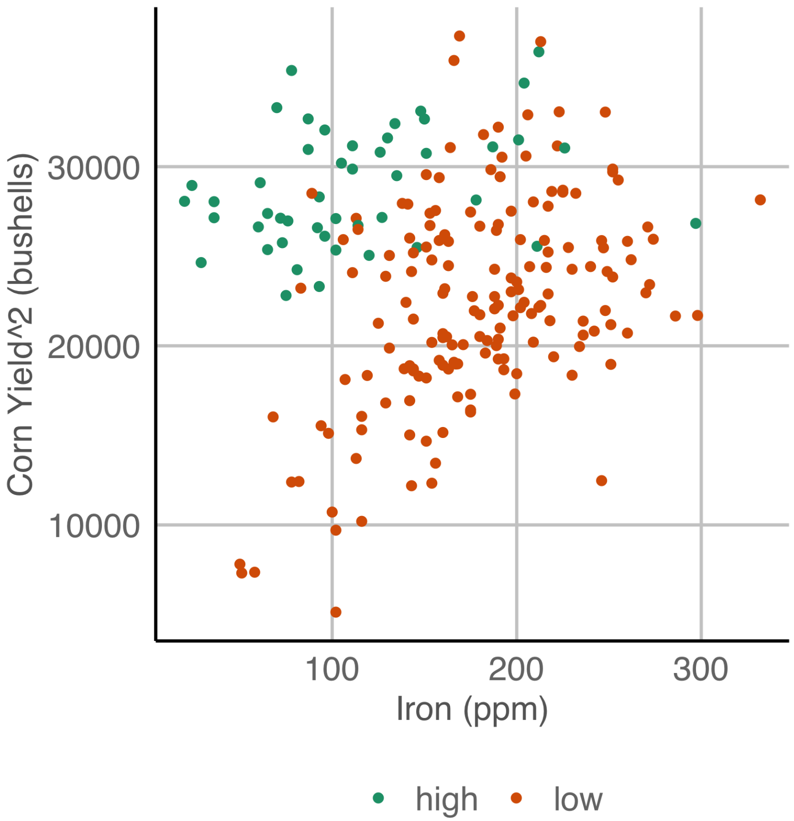
Case study 4 Soils



Lurking variable?

```r
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 4 Soils

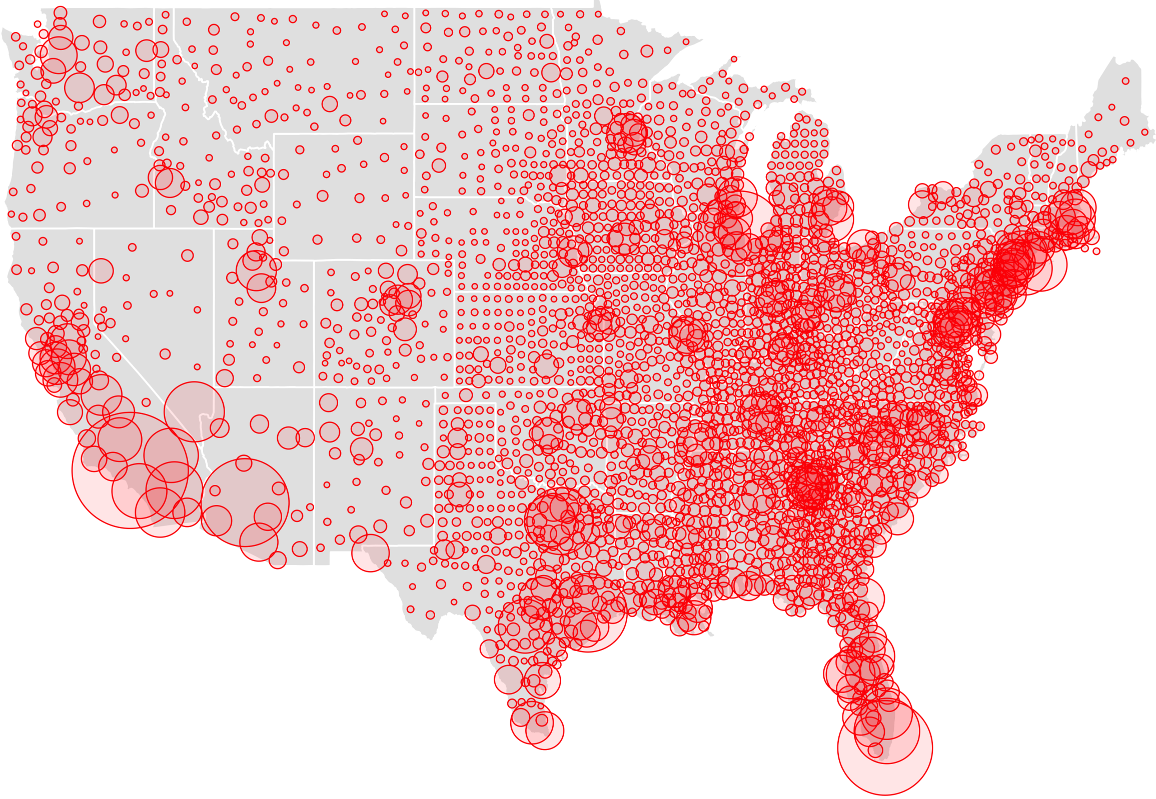


Colour high calcium (>5200ppm) calcium values

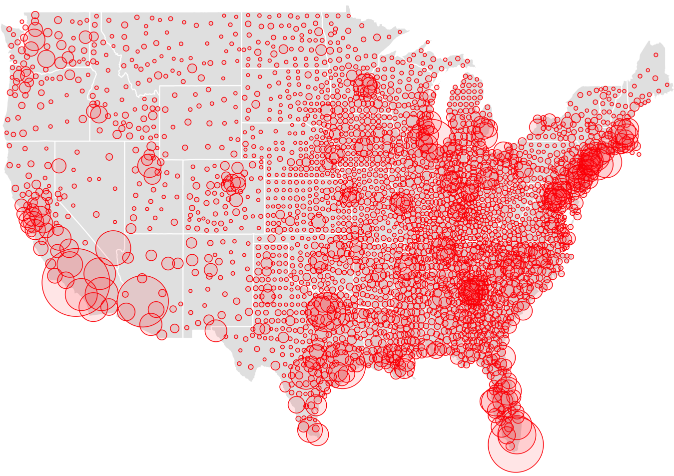
```
ggplot(baker, aes(
  x = Fe, y = Corn97BU^2,
  colour = ifelse(Ca > 5200,
    "high", "low"
  )
)) +
  geom_point() +
  xlab("Iron (ppm)") +
  ylab("Corn Yield^2 (bushells)") +
  scale_colour_brewer("", palette = "Dark2") +
  theme(
    aspect.ratio = 1,
    legend.position = "bottom",
    legend.direction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

Case study 5 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?

The type of variable determines the choice of mapping.

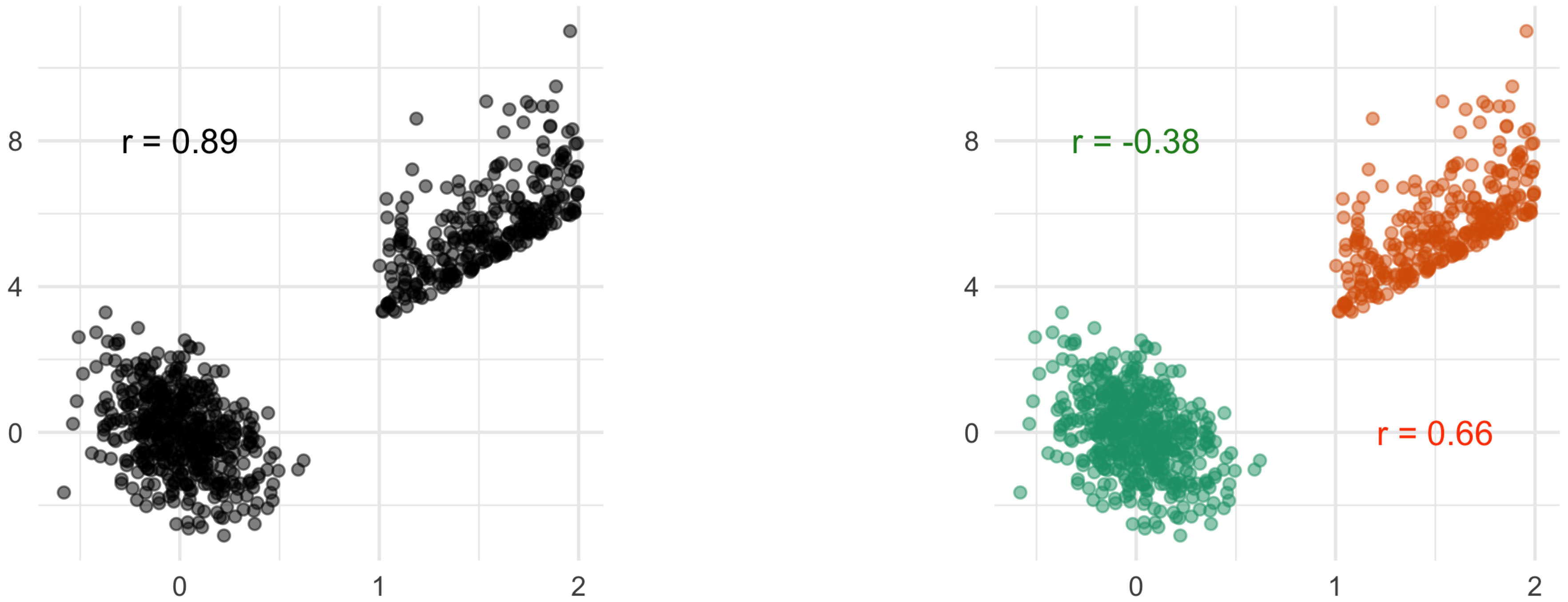
- Continuous and categorical $\longrightarrow$ side-by-side boxplots, side-by-side density plots
- Both categorical $\longrightarrow$ faceted bar charts, stacked bar charts, mosaic plots, double decker plots

We'll see more examples soon.

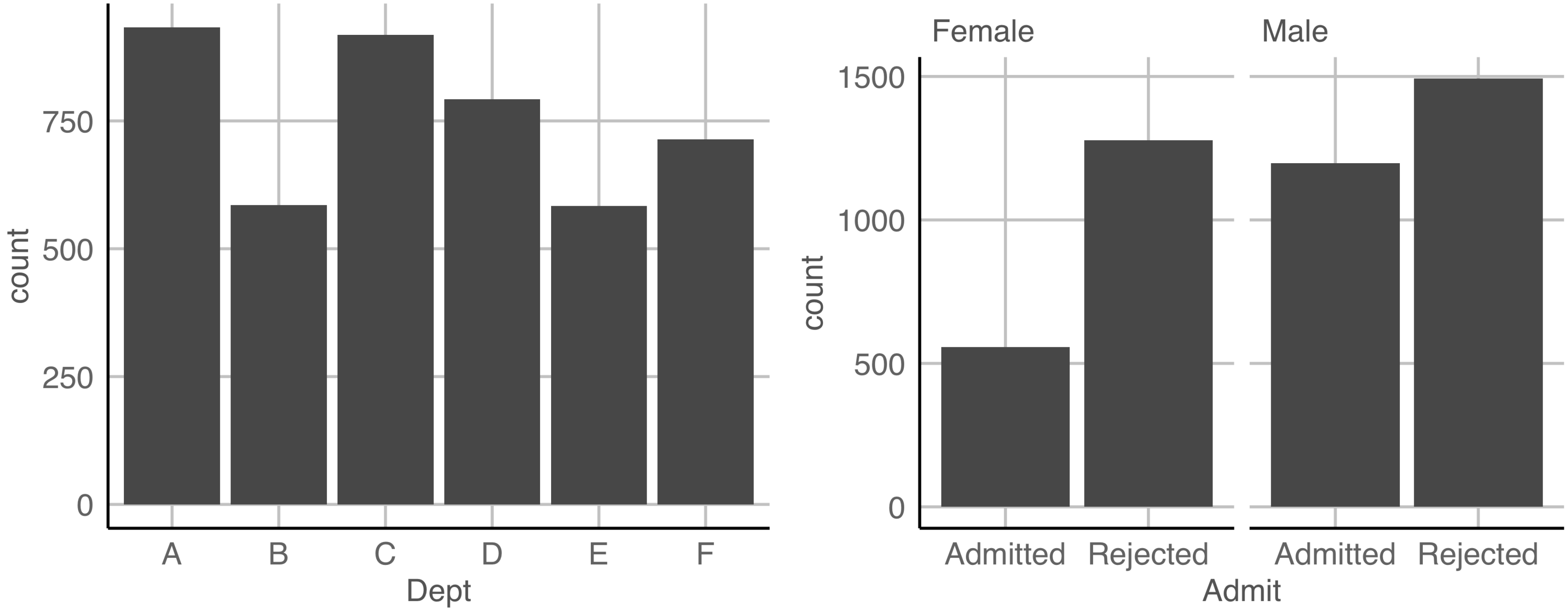
Paradoxes

Simpsons paradox

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

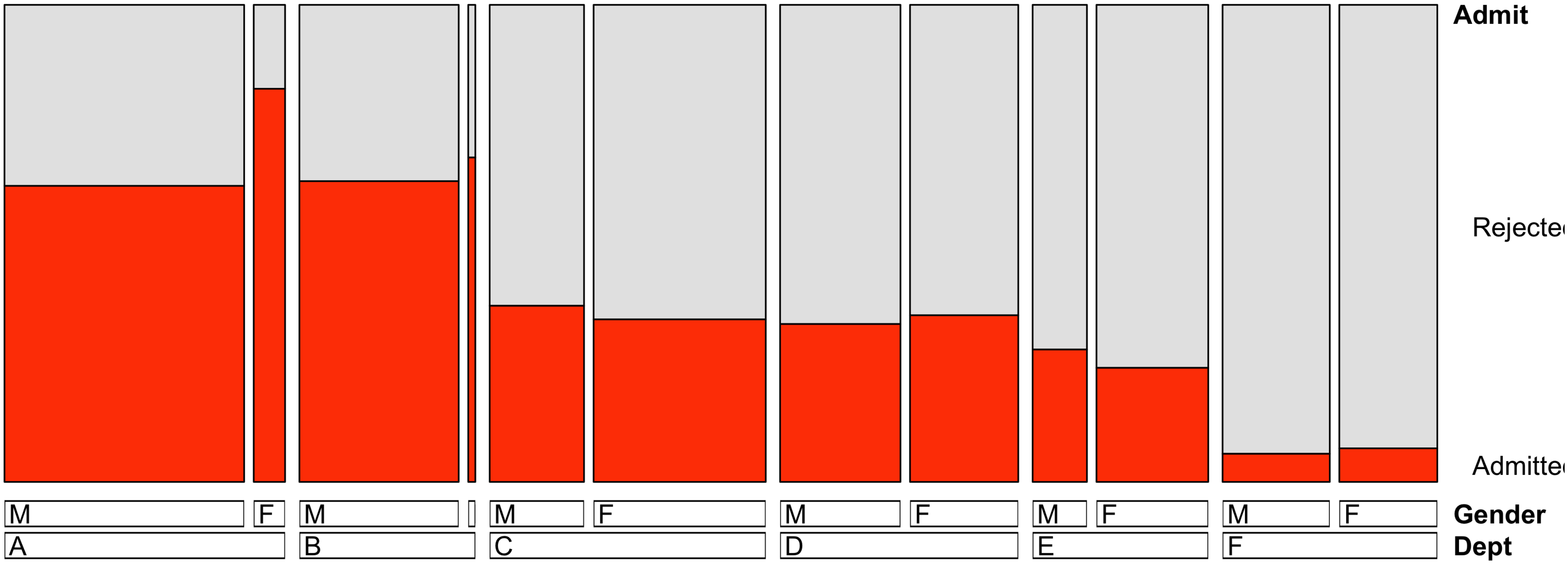


Simpsons paradox: famous example



Did Berkeley discriminate against female applicants?

Simpsons paradox: famous example



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

Example from Unwin (2015)

Is what you see really association?

Checking association with visual inference

Soils R Olympics R

```
ggplot(
  lineup(null_permute("Corn97BU"), baker, n = 12),
  aes(x = B, y = Corn97BU)
) +
  geom_point() +
  facet_wrap(~.sample, ncol = 4)
```

11 of the panels have had the association broken by permuting one variable. There is no association in these data sets, and hence plots. Does the data plot stand out as being different from the null (no association) plots?

Resources

- Friendly and Denis "Milestones in History of Thematic Cartography, Statistical Graphics and Data Visualisation" available at http://www.datavis.ca/milestones/
- Unwin (2015) Graphical Data Analysis with R
- Graphics using ggplot2
- Wilke (2019) Fundamentals of Data Visualization https://clauswilke.com/dataviz/

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 6 - Session 2