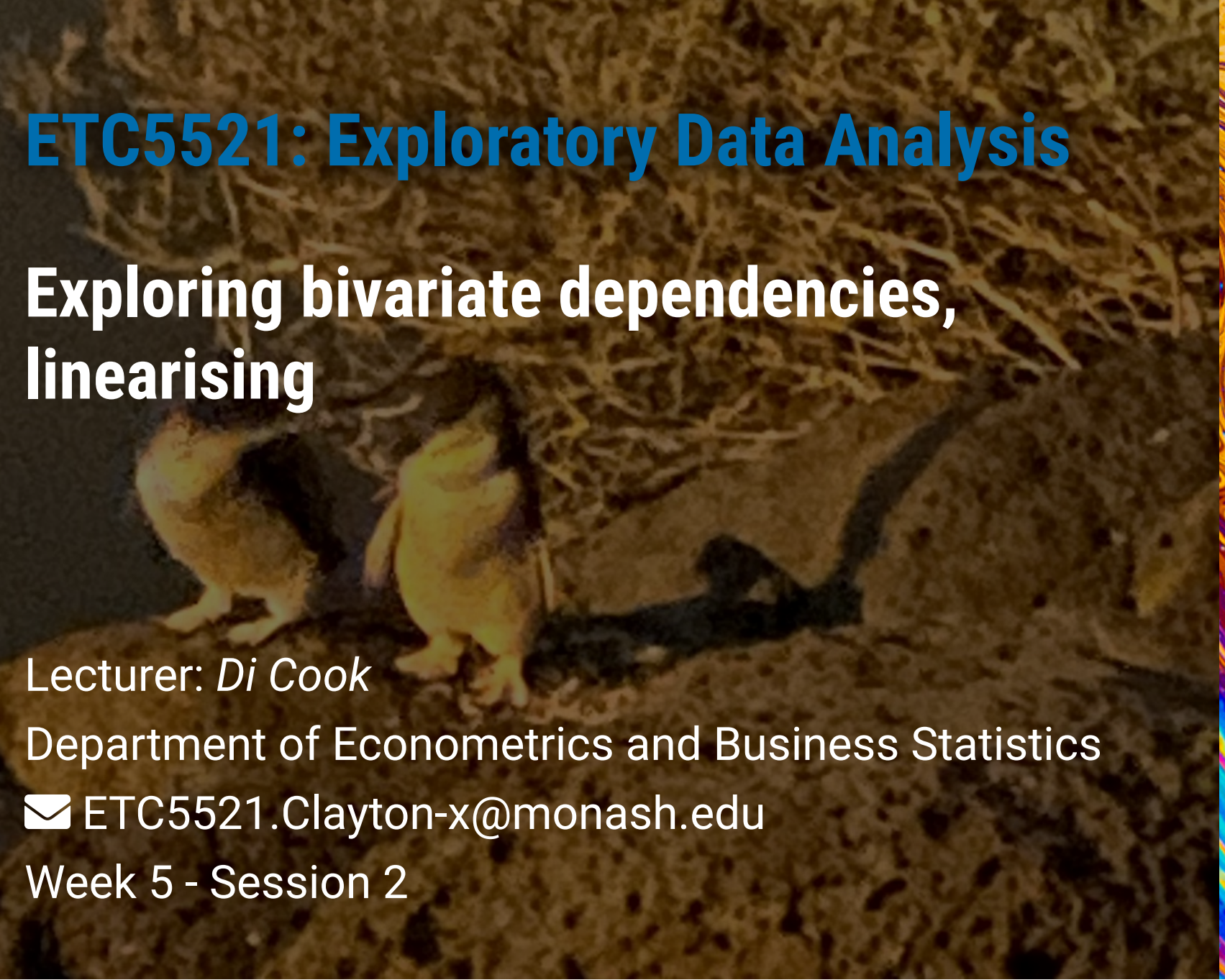
ETC5521: Exploratory Data Analysis

Exploring bivariate dependencies, linearising



Lecturer: *Di Cook*

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu

Week 5 - Session 2

History

> *Scatter plots are glorious. Of all the major chart types, they are by far the most powerful. They allow us to quickly understand relationships that would be nearly impossible to recognize in a table or a different type of chart. ... Michael Friendly and Daniel Denis, psychologists and historians of graphics, call the scatter plot the most "generally useful invention in the history of statistical graphics." Dan Kopf*

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 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.

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.

Peceiving correlation

answers R

$\rho = 0$ $\rho = 0.4$ $\rho = 0.6$ $\rho = 0.8$

$\rho = -0.2$ $\rho = -0.5$ $\rho = -0.7$ $\rho = -0.9$

My guess is that you didn't do very well. You likely under-stated r, particularly around 0.4-0.7. The variation in scatter is not linear with r.

When someone says *correlation is 0.5* it sounds impressive. BUT when someone shows you a scatterplot of data that has correlation 0.5, you will say that's a weak relationship.

Transformations

for skewness, heteroskedasticity and linearising relationships

Case study 2 Movies

`learn R`

- Odd pattern, almost looks like an "r"
- No films with lots of votes and low rating
- No film with lots of votes has rating close to maximum possible: **barrier?**
- Films with very high ratings only have a few votes
- Generally, rating appears to increase as votes increases (its hard to really read this with so few points though)
- A few films with really large number of votes: **outliers?** or just **skewness?**
- Films with few votes have ratings that span the range of the scale.

Would you say this is positive, linear, moderate?

Or positive, non-linear, and moderate? Or weak?

In some sense, these descriptions are meaningless, here.

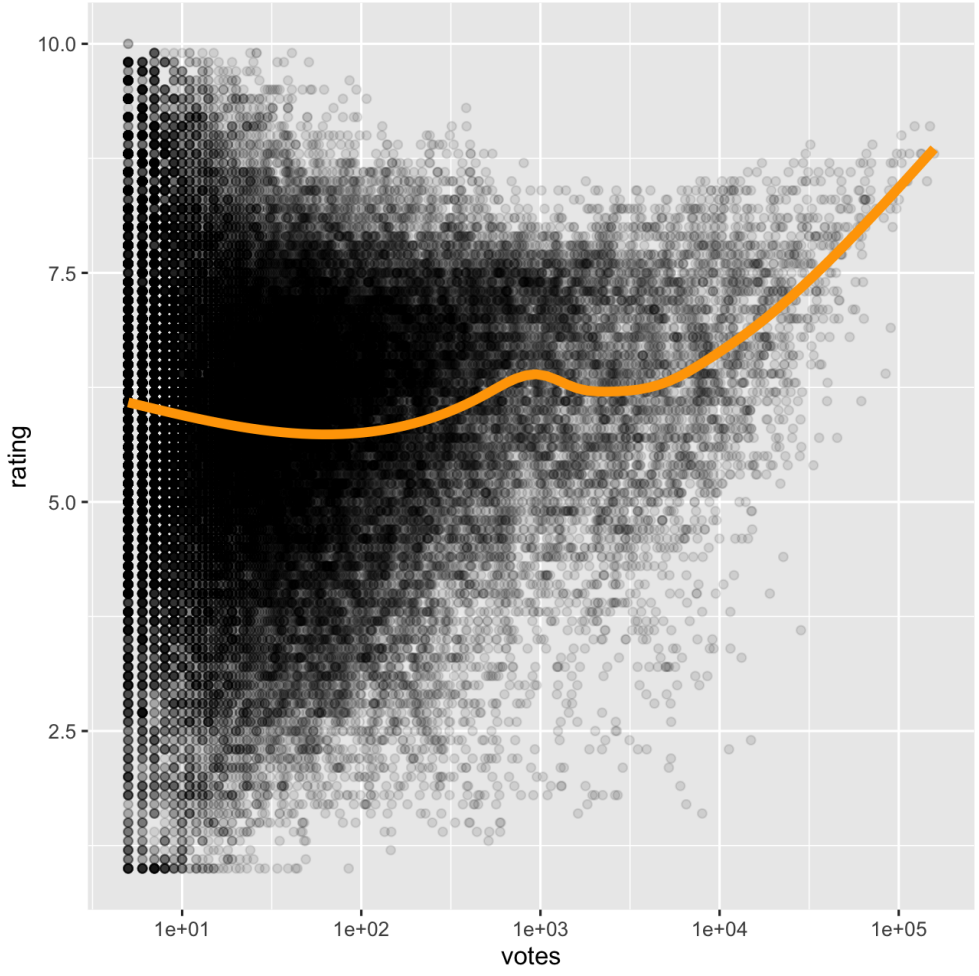
What about causation? association? outliers? clusters? gaps? barrier? conditional relationships?

These descriptive help to describe relationships generally, but it is important to convert them into the context of the (variables in the) data.

BUT, BUT there is a skewness in votes that needs fixing before assessing the relationship.

Case study 2 Movies

R

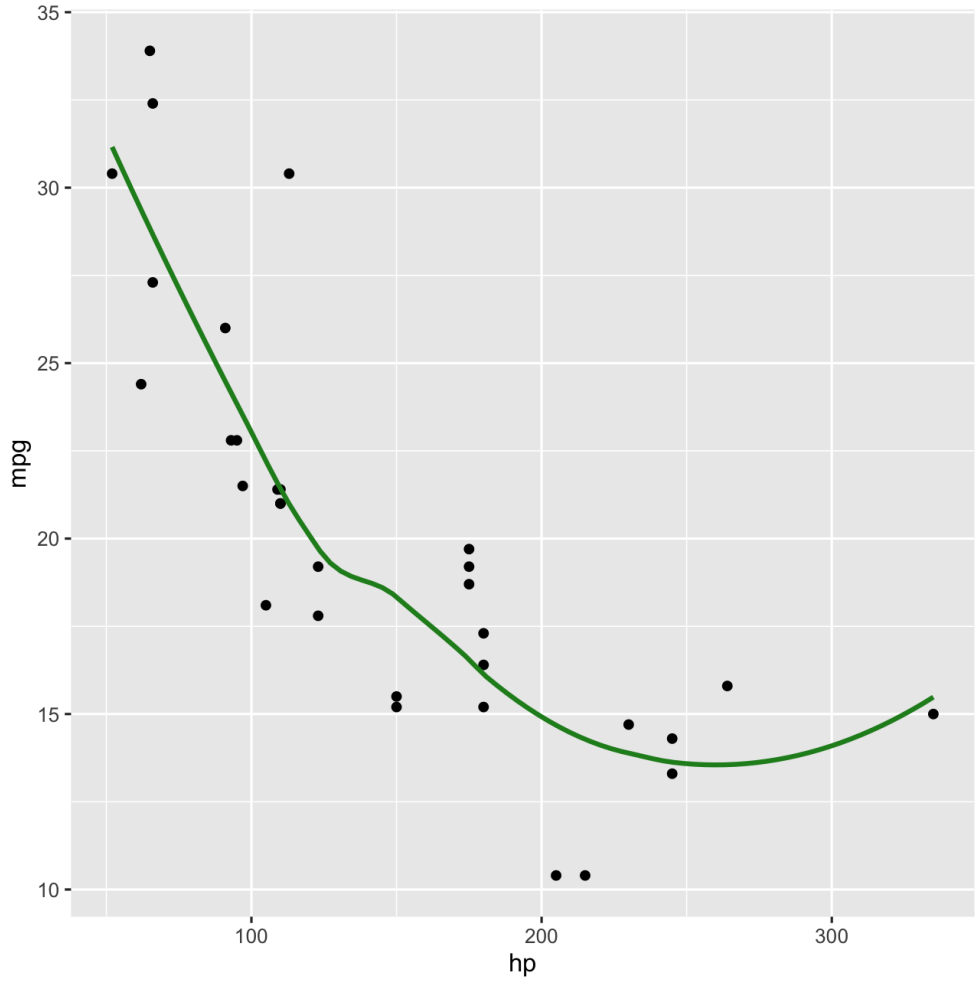


🤔 Something funny happens, right at 1000 votes

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

Case study 3 Cars

learn R

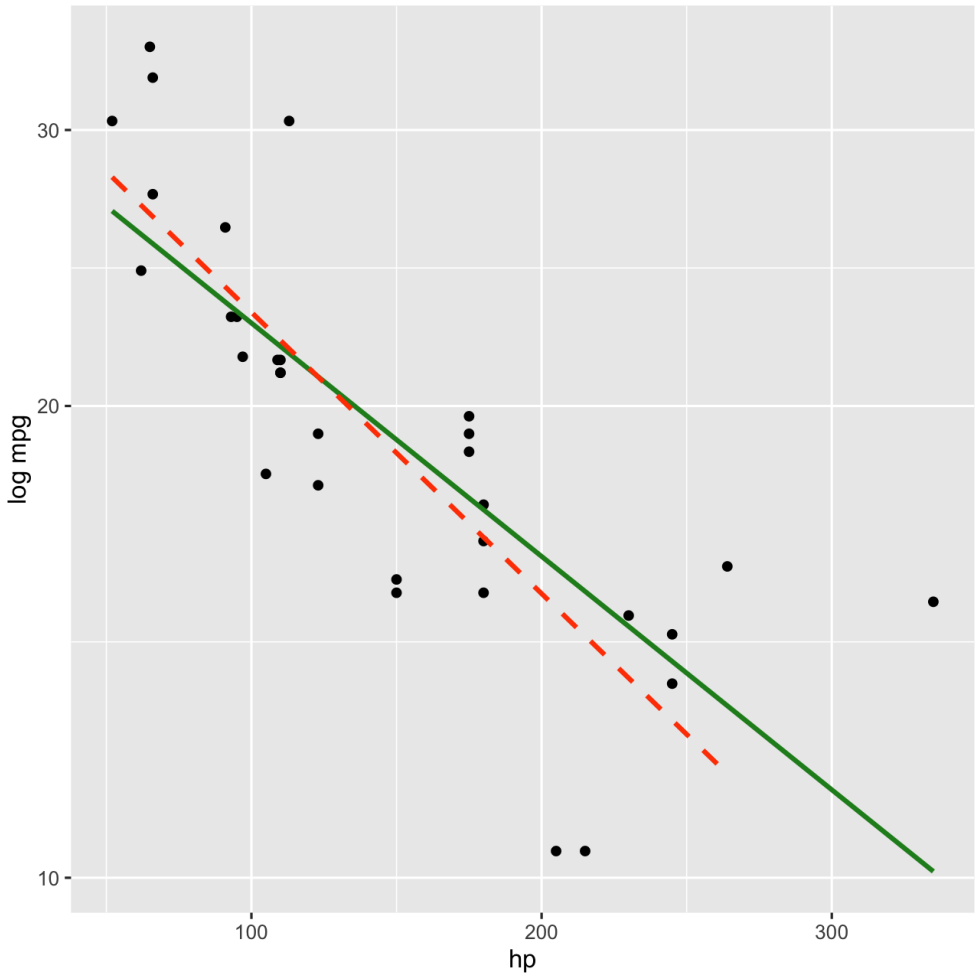


- mpg: Miles/(US) gallon
- hp: Gross horsepower

Describe the relationship between horsepower and mpg.

Case study 3 Cars

R

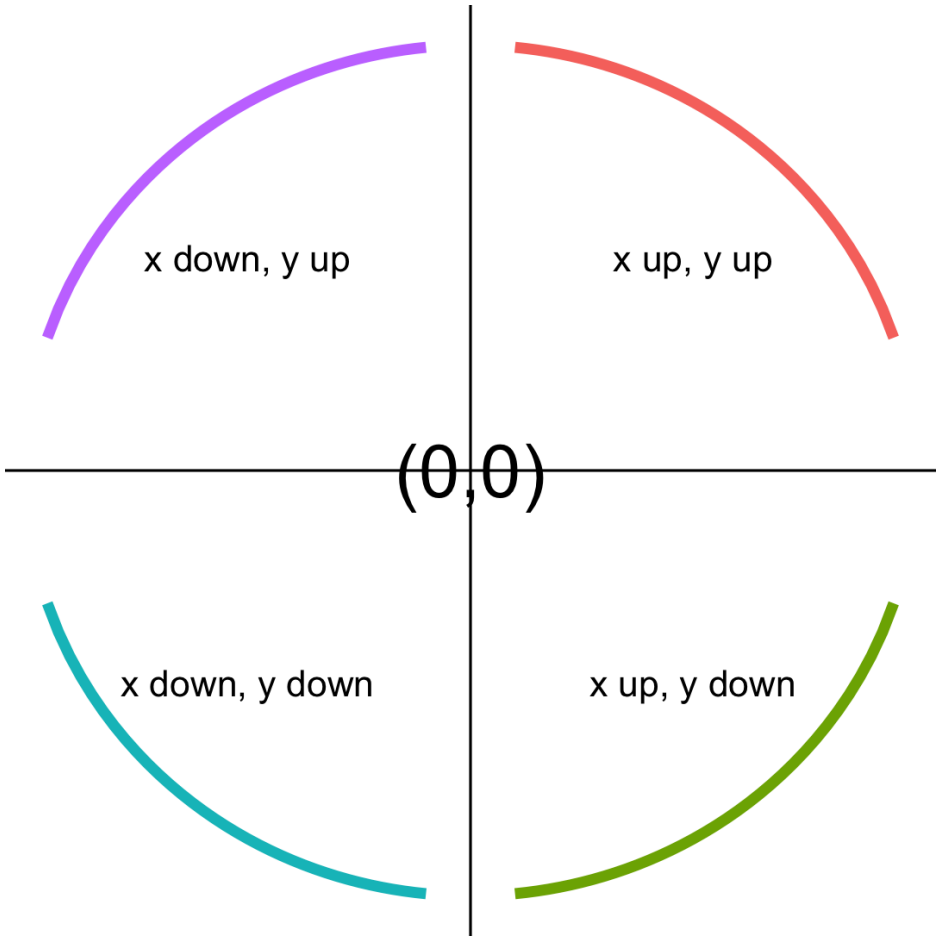


- `mpg`: Miles/(US) gallon
- hp: Gross horsepower

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

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

Circle of transformations for linearising



Remember the power ladder:

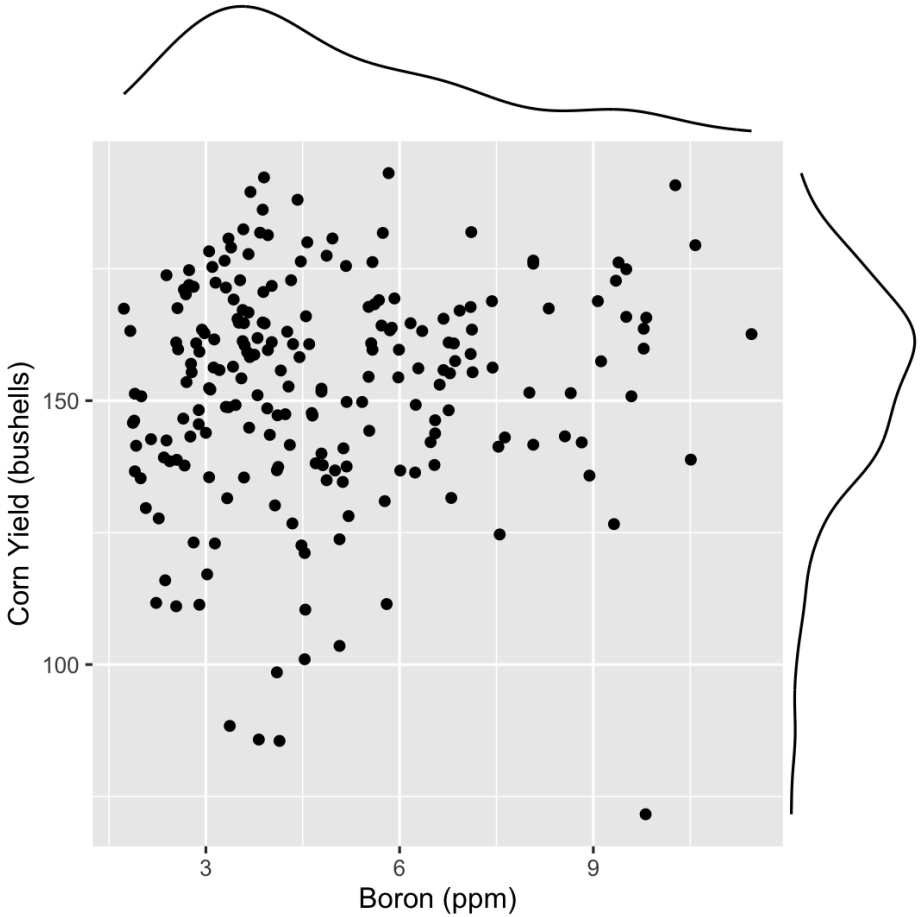
-1, 0, 1/3, 1/2, 1, 2, 3, 4 ?

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

Case study 4 Soils

Interplay between skewness and association



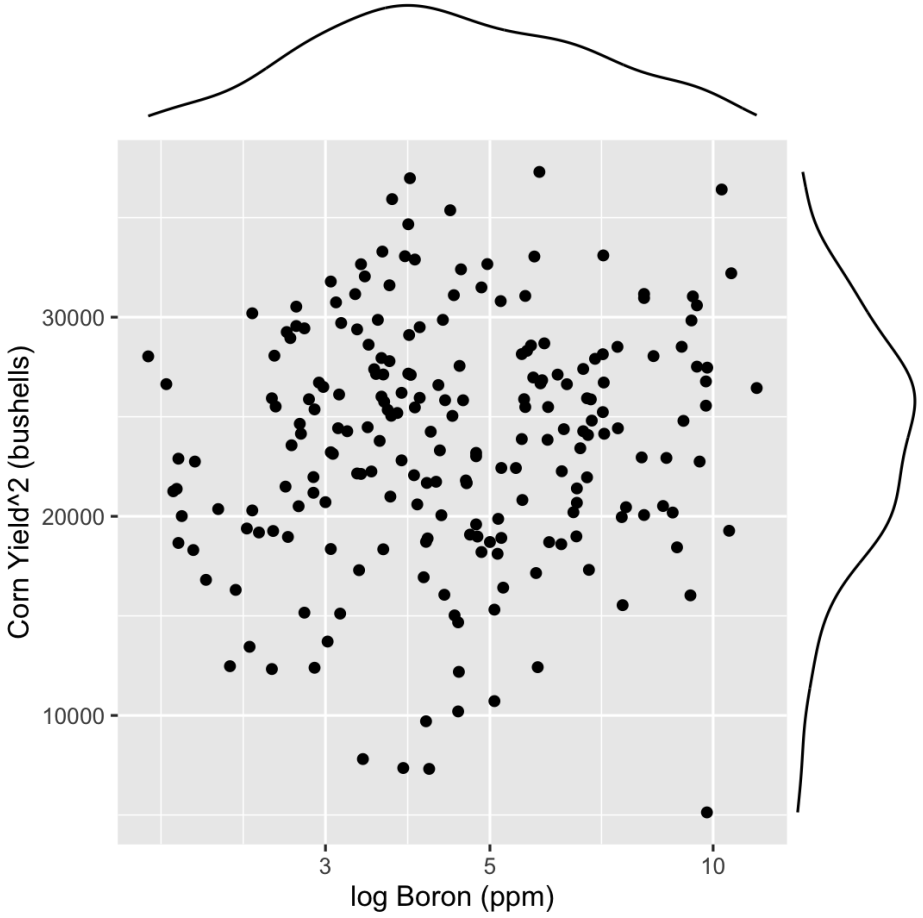
Soil chemical analysis of a farm field in Iowa

Is there a relationship between Yield and Boron?

`ggMarginal` adds density plots for each variable to the scatterplot

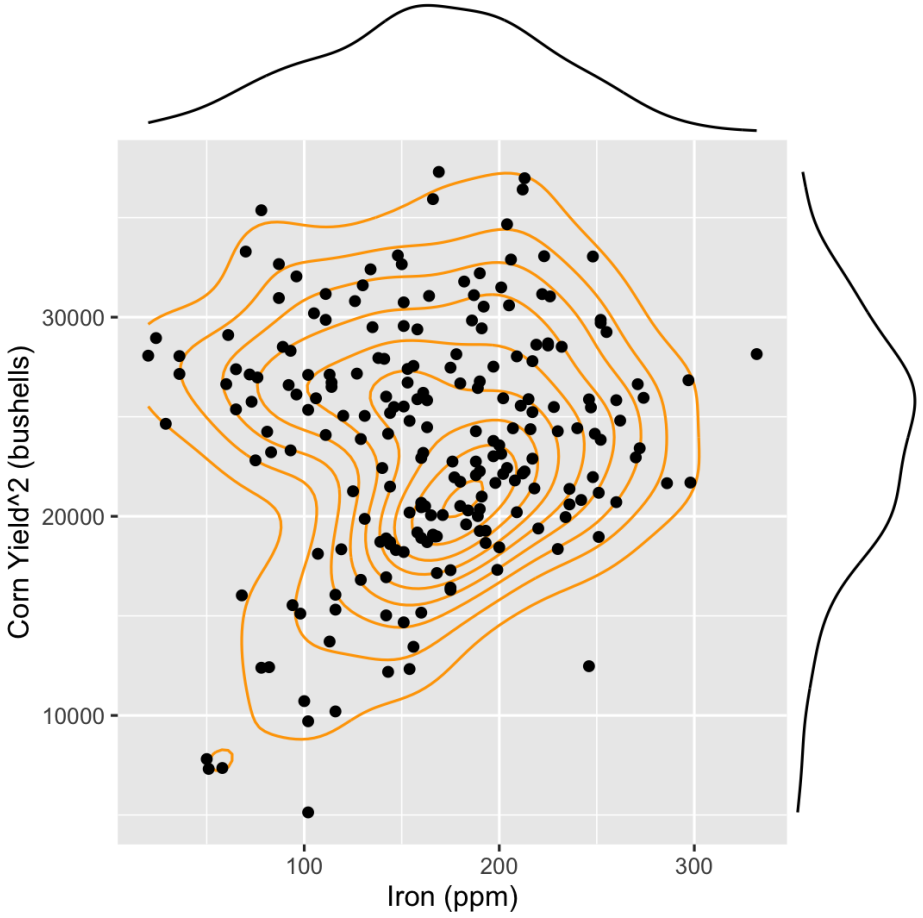
Is Boron skewed? Is Yield skewed? Then it is hard to assess the relationship.

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")
```

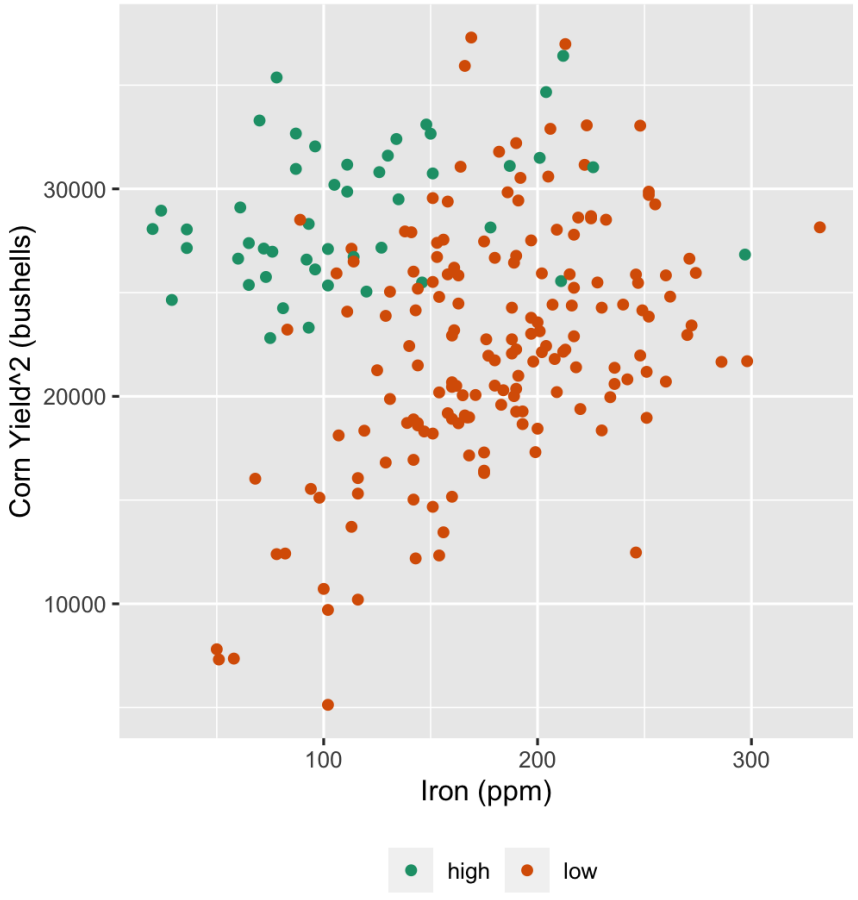
Case study 4 Soils



Lurking variable?

```
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



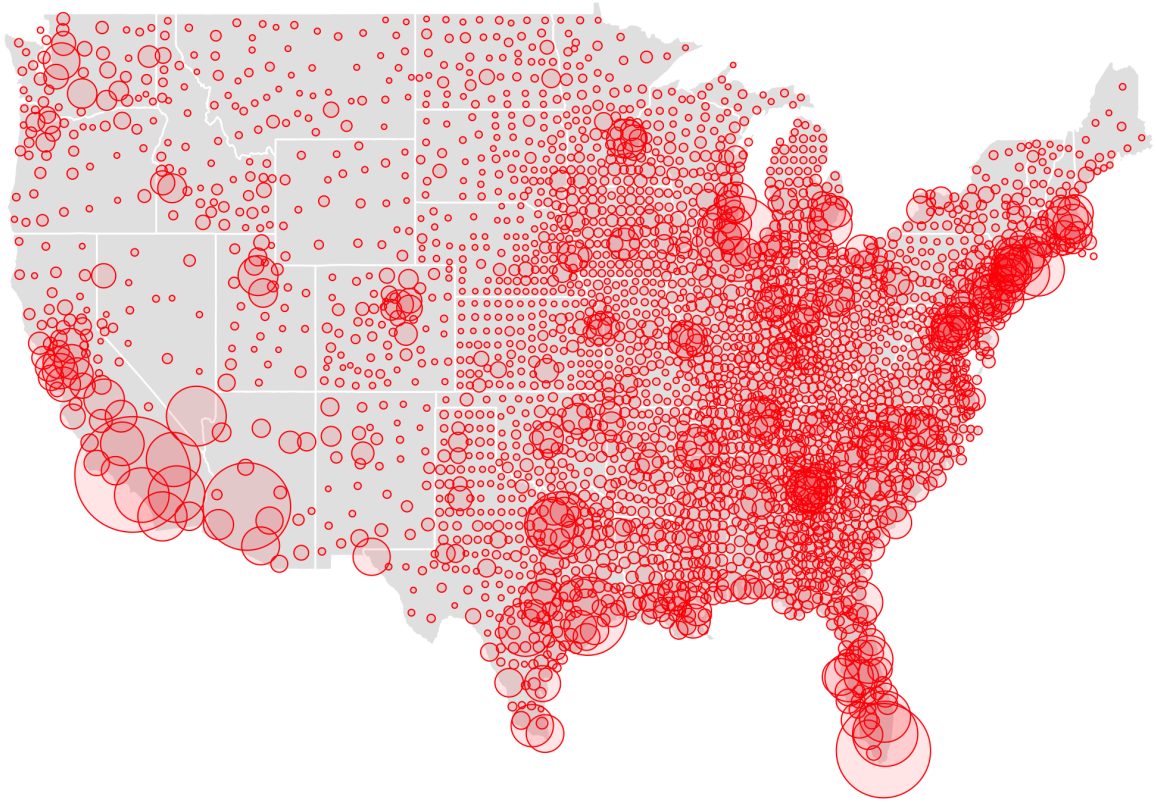
Colour high (>5200ppm) calcium values

```
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")
```

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 5 COVID-19

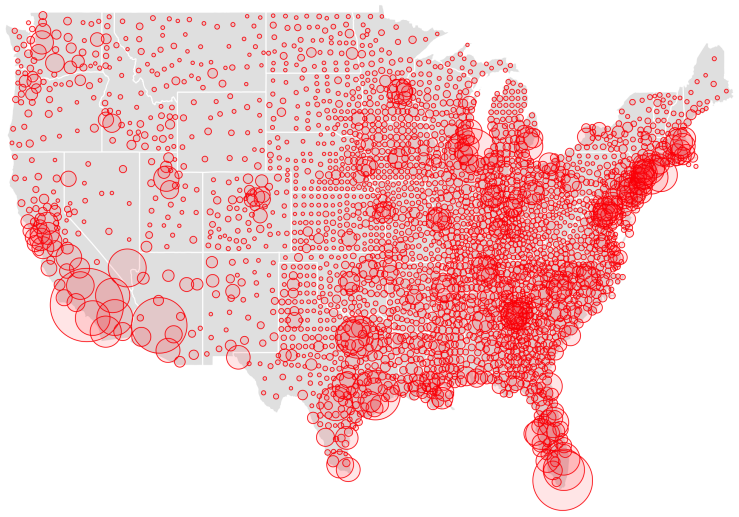
R



Bubble plots, size of point is mapped to another variable.

This bubble plot here shows total count of COVID-19 incidence (as of Aug 30, 2020) for every county in the USA, inspired by the New York Times coverage.

Scales matter



Where has COVID-19 hit the hardest?

Where are there more people?

Is it only a problem in population centres? Should we calibrate the counts by population?

Generalised

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

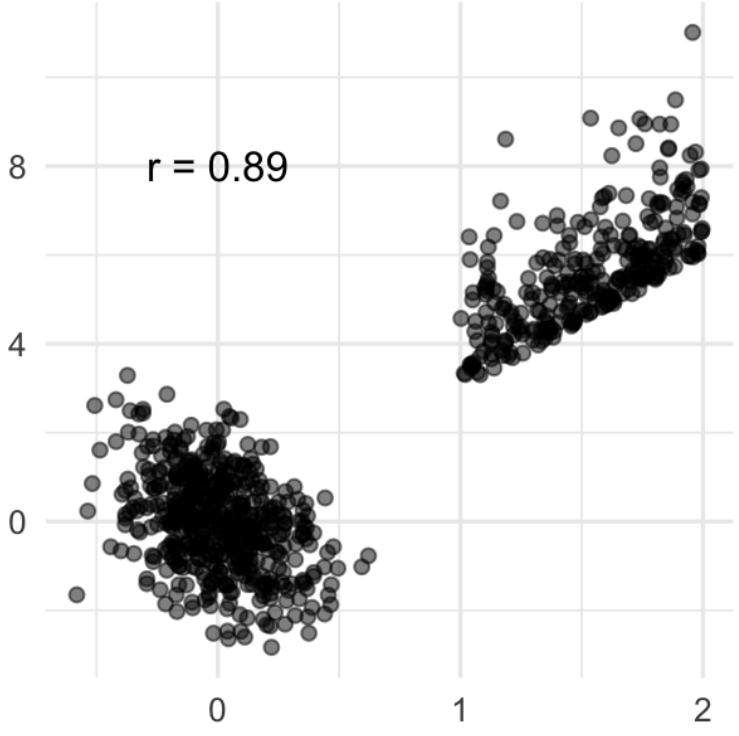
The type of variable determines the choice of mapping.

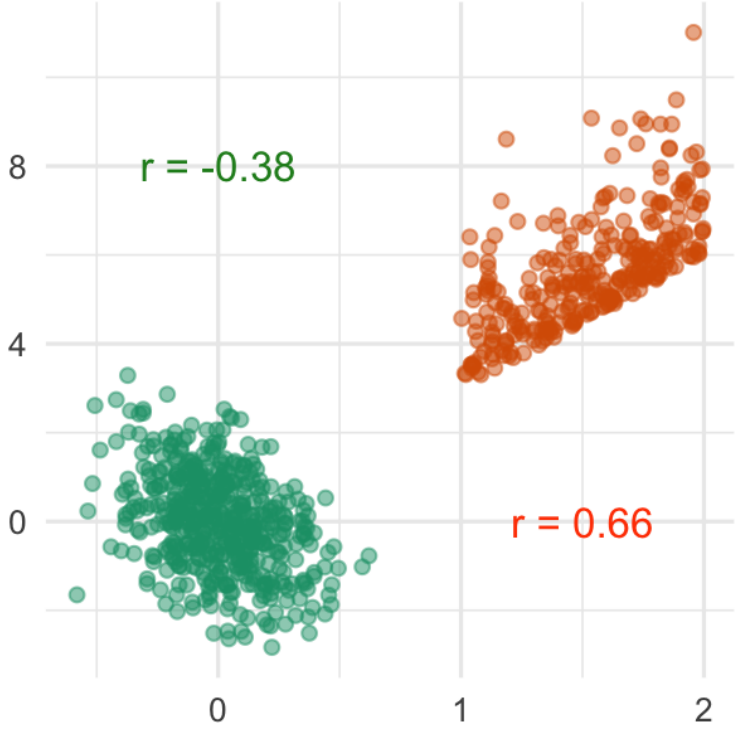
- Continuous and categorical ⟶ side-by-side boxplots, side-by-side density plots
- Both categorical ⟶ faceted bar charts, stacked bar charts, mosaic plots, double decker plots

> *We'll see more examples soon.*

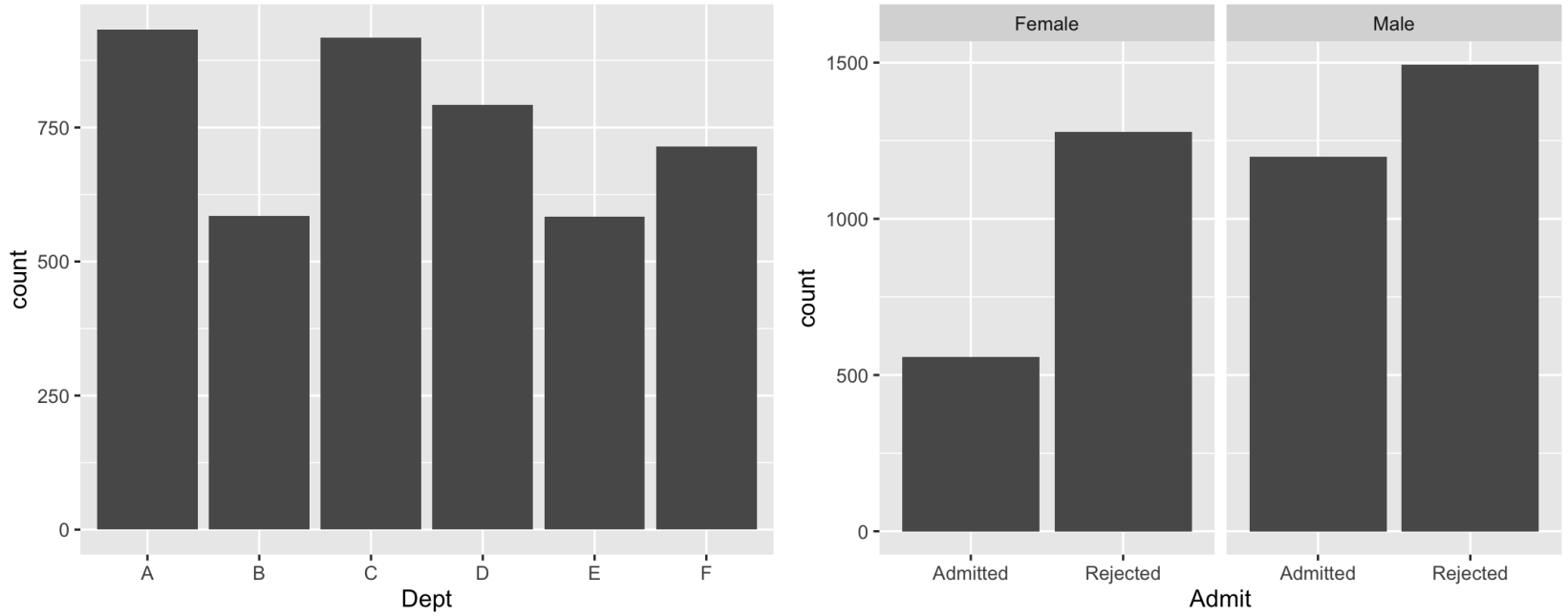
Simpsons paradox

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





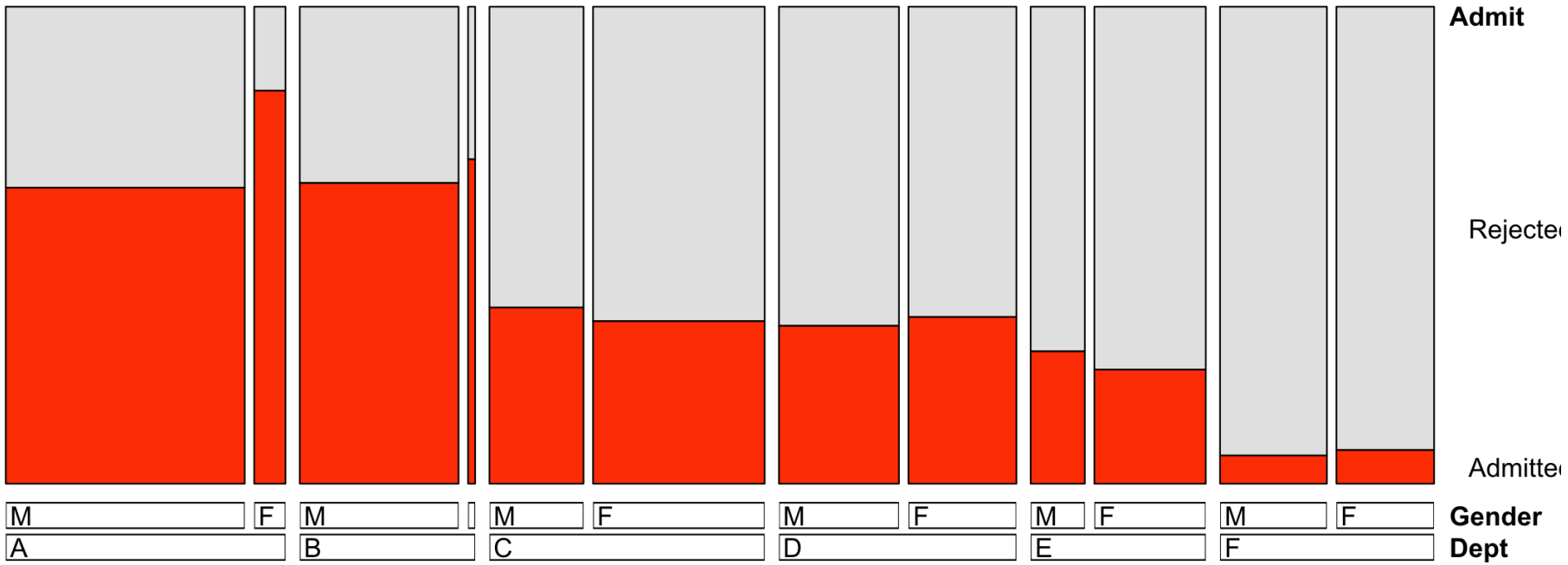
Simpsons paradox: famous example



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.

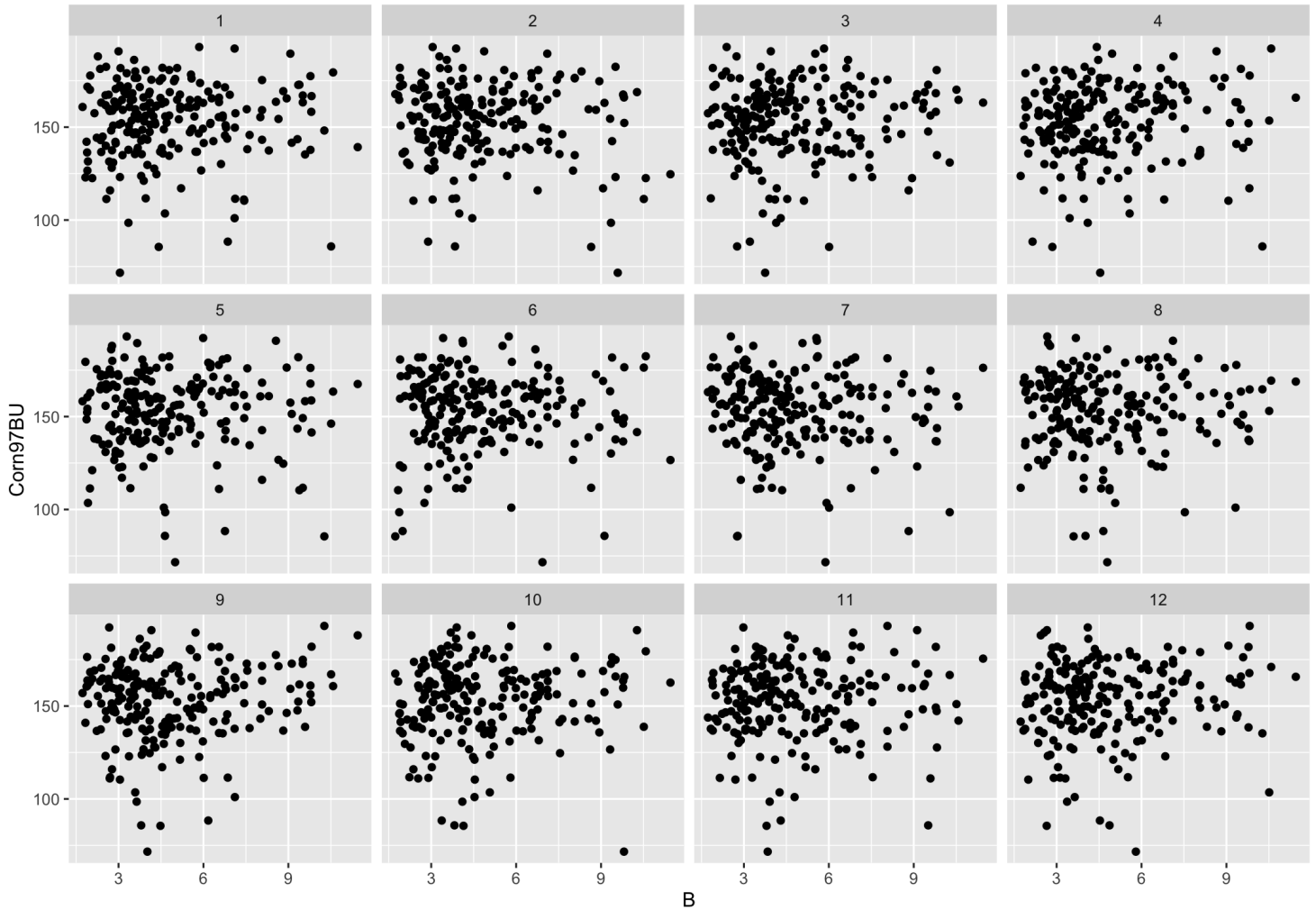
Example from Unwin (2015)

Spurious association

Be careful of double time series, especially if the data is not being shown.

Checking association with randomisation

`Soils` R `Olympics` R



That's it, for this lecture!

Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu