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 1

Dependency Relationships and Associations

This lecture is based on Chapter 5 of

Unwin (2015) Graphical Data Analysis with R

"The world is full of obvious things which nobody by any chance observes." Sherlock Holmes

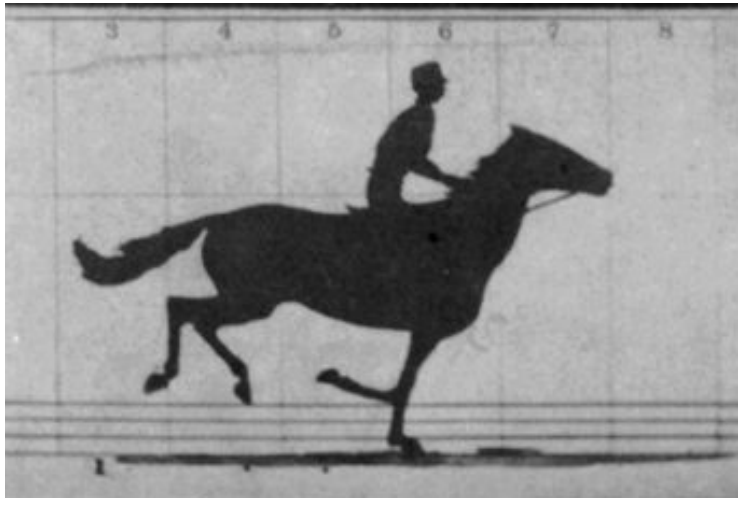
The story of the galloping horse

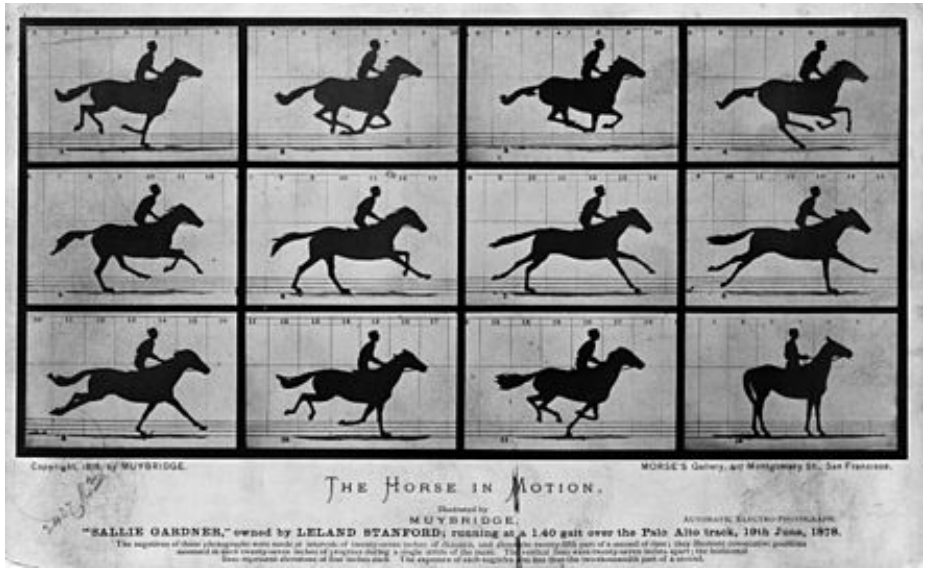
Baronet, 1794

Derby D'Epsom 1821

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

The story of the galloping horse





With the birth of photography, and particular motion photography, Muybridge illustrated that this leg position was impossible.

My painting story(s)

- "Take another look at the hills"
- Reflection from lemons
- Green trees
- Tendency to
  - paint what other people have drawn, not what we see.
  - Or what we impose, like trees are green.

> *Try to see with fresh eyes*

The scatterplot

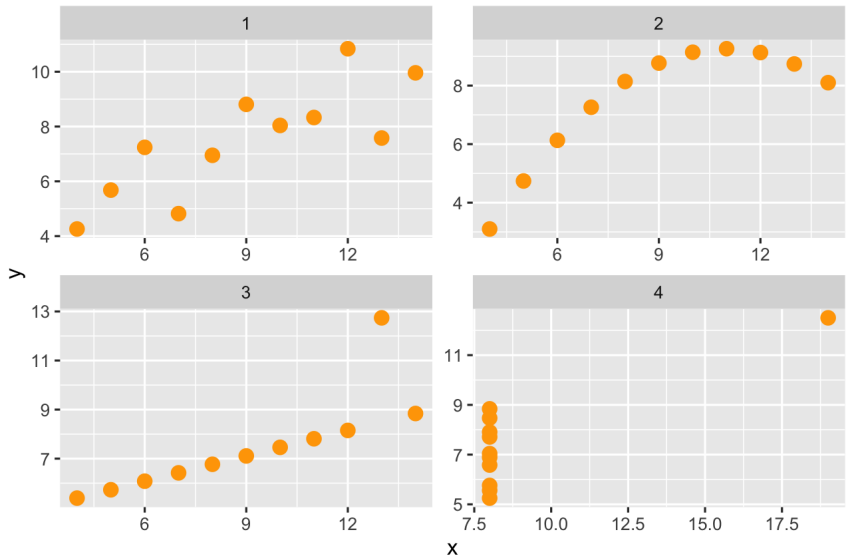
Scatterplots are the natural plot to make to explore association between two **continuous** (quantitative) variables.

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.

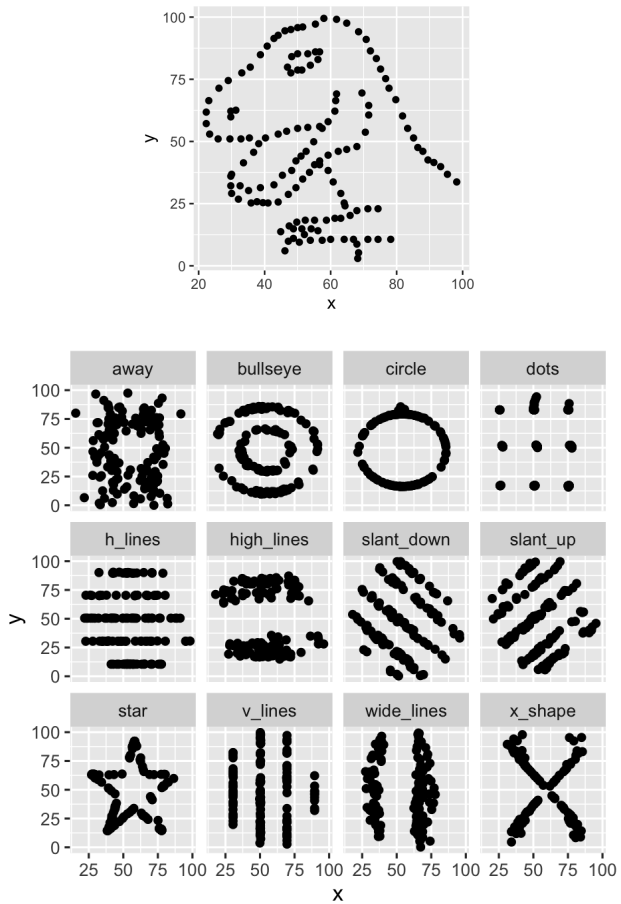
Famous scatterplot examples

Anscomb's quartet



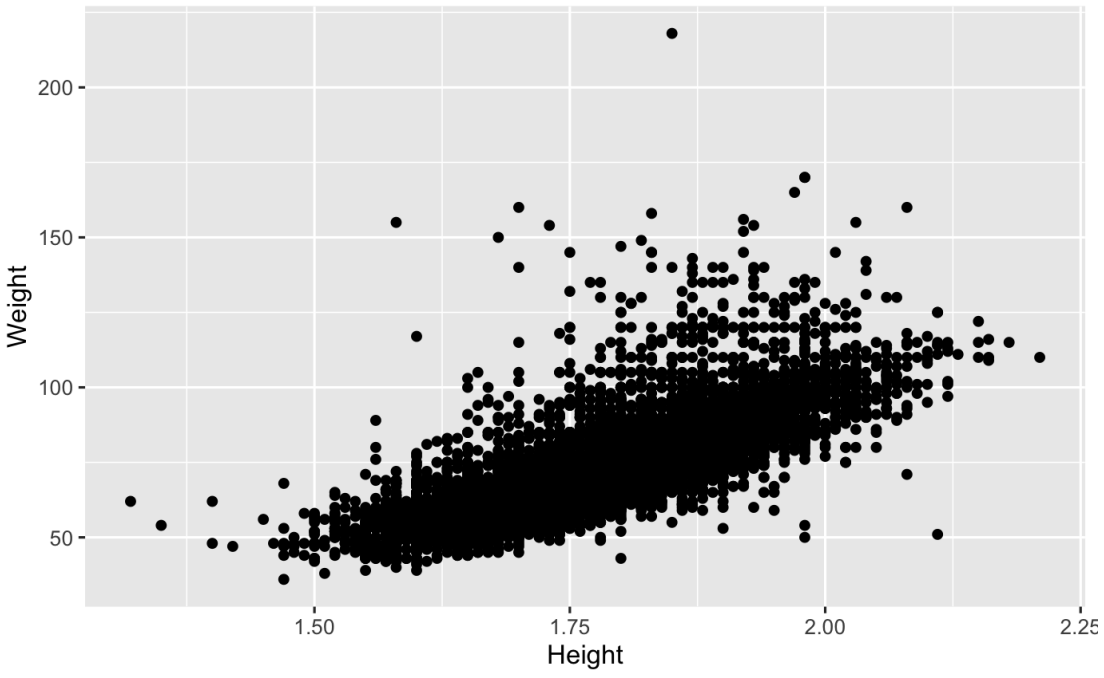
> *All data has same means, standard deviations and correlation*

Datasaurus dozen



Case study 1 Olympics

data R



Warning message: Removed 1346 rows containing missing values (geom_point)

The expected linear relationship between height and weight is visible, although obscured by outliers.

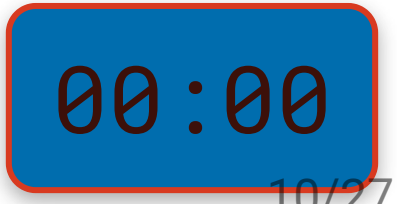
Some discretization of heights, and higher weight values.

Likely to be substantial overplotting (57 athletes 1.7m, 60kg can't tell this from this plot).

Note the unusual height-weight combinations. What sport(s) would you expect some of these athletes might be participating in?

Your turn, cut and paste the code into your R console, and mouse over 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)
```



Sports summary

Synchronised Swimming	101
Beach Volleyball	93
Gymnastics - Rhythmic	92
Canoe Slalom	80
Cycling - Mountain Bike	72
Modern Pentathlon	69
Cycling - BMX	43
Trampoline	31
Cycling - Road, Cycling - Track	16
Cycling - Mountain Bike, Cycling - Road, Cycling - Track	3
Cycling - Mountain Bike, Cycling - Track	3

Consolidate some factor levels

```
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))
```

Split the scatterplots by sport

 `learn R`

What do we learn?

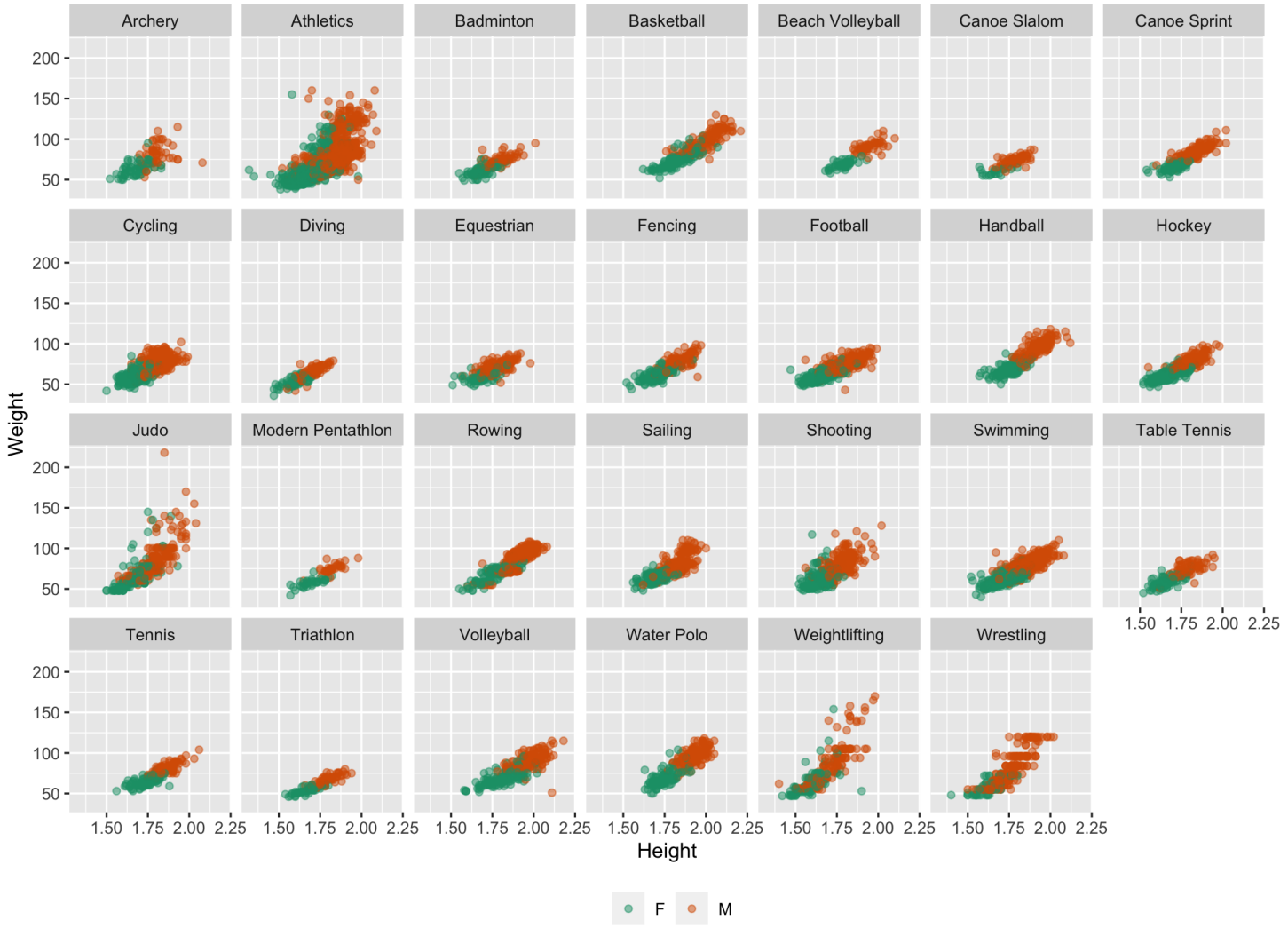
- Missing values for some sports
- The positive association between height and weight is visible across sports
- Maybe nonlinear in wrestling
- An outlier in judo, and football, and archery
- Maybe flatter among swimmers
- Taller athletes in basketball, volleyball and handball
- Shorter athletes in athletics, weightlifting and wrestling
- Little variance in tennis players
- *Its still messy, and hard to digest*

What would you do to make comparisons easier?

- Remove sports with missings
- Make regression lines for remaining sports on one plot
- Separately examine male/female athletes
- Compare just one group against the rest

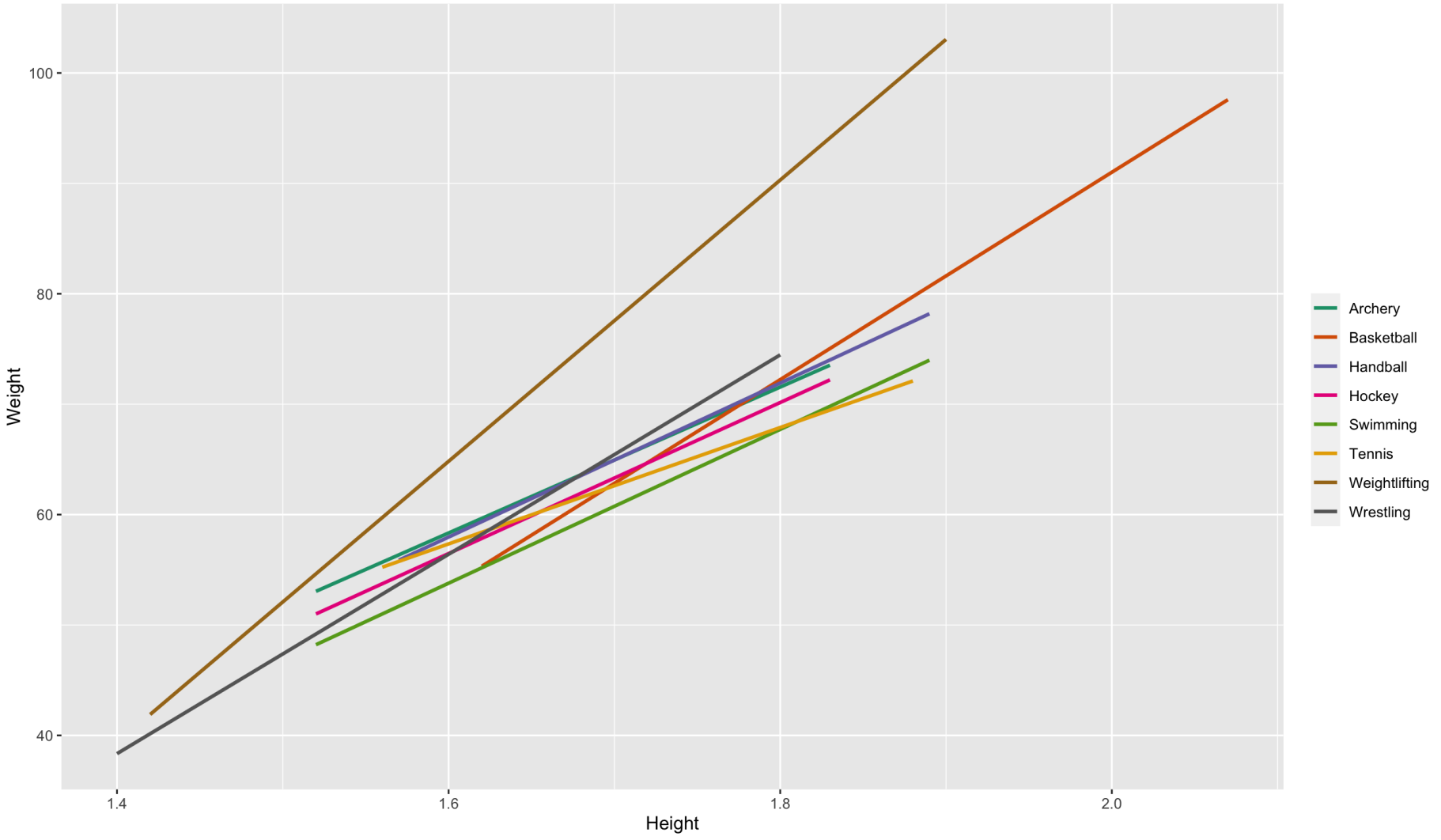
Remove missings, add colour for sex

learn R



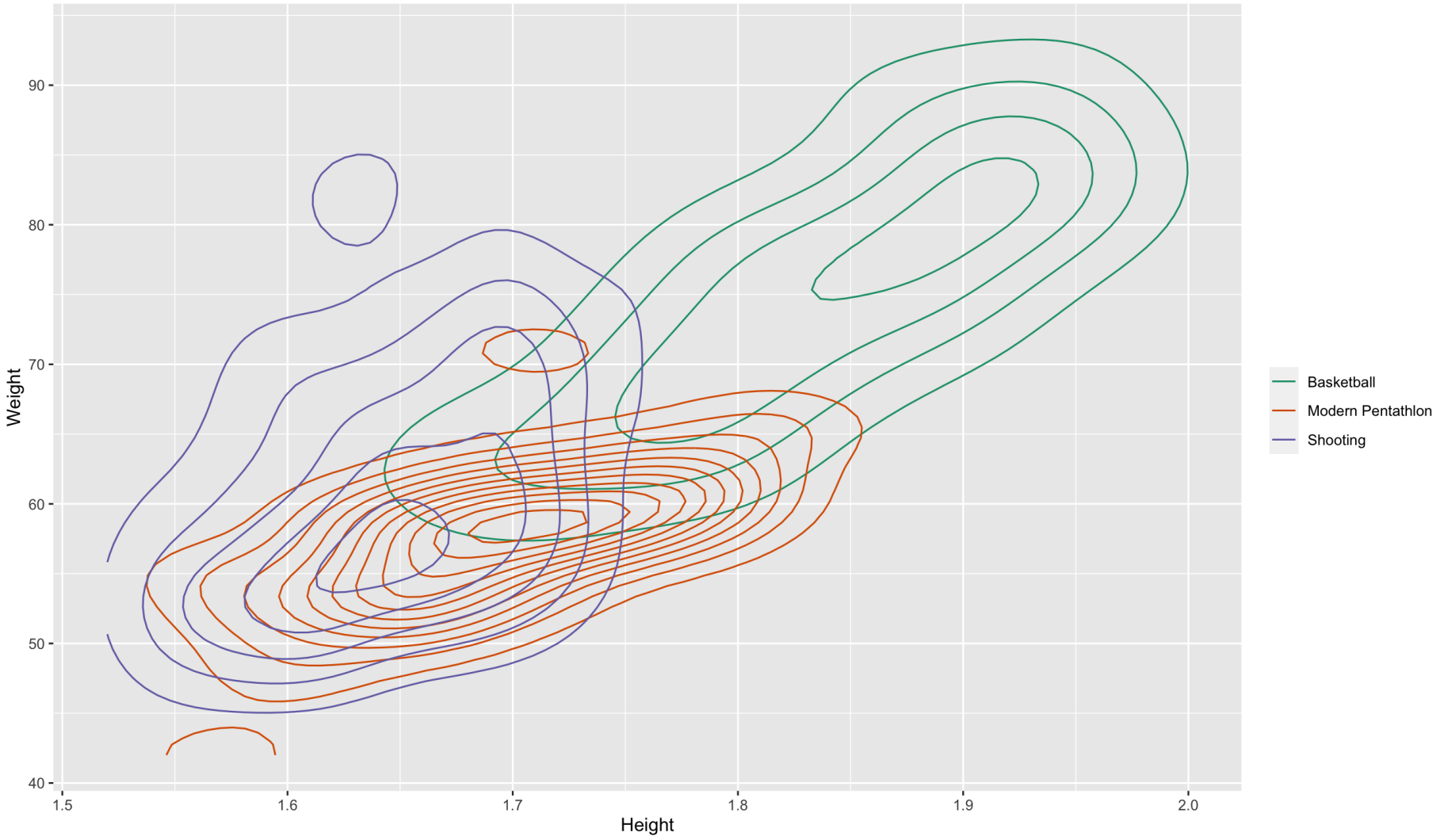
Comparing association

learn R



Comparing variability

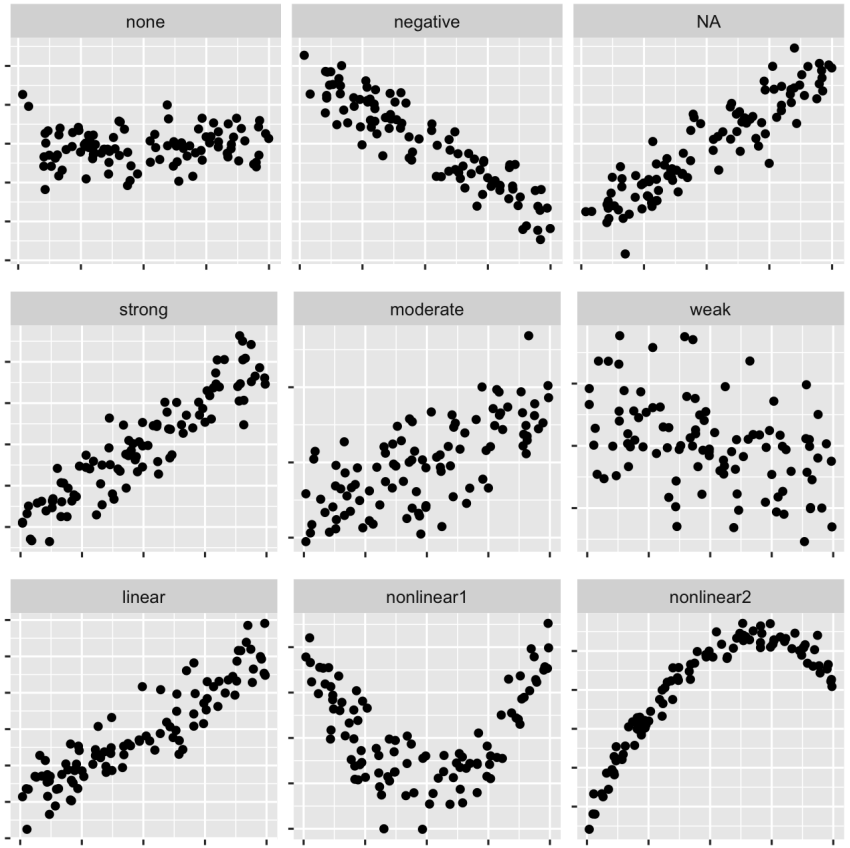
learn R



What does it mean to say associated?

A traditional summary of a scatterplot would include these elements:

- **trend**: positive or negative?
- **strength**: what is the variation around the trend, are points close or scattered
- **form**: linear or nonlinear/curvilinear/clustered



What features might be visible in scatterplots?

I rather prefer Unwin's taxonomy of features that might be visible in a scatterplot:

- **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. *Its 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.
- **outliers or groups of outliers**: observations can be outliers in two dimensions without being an outlier in either of the single variables, particularly if there is a strong association between the variables.

What features might be visible in scatterplots?

I rather prefer Unwin's taxonomy of features that might be visible in a scatterplot:

- **clusters**: some observations separate from others
- **gaps**: sometimes a particular combination of values does not occuer together.
- **barriers**: some combinations are impossible, for example, being younger than the years of experience in the workforce
- **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.

Case study 1 Olympics

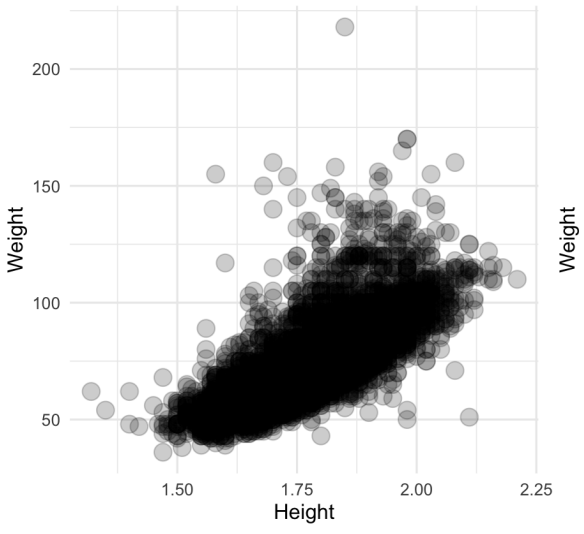
We have seen that the association between height and weight is "contaminated" by different variables, sport, gender, and possibly country and age, too.

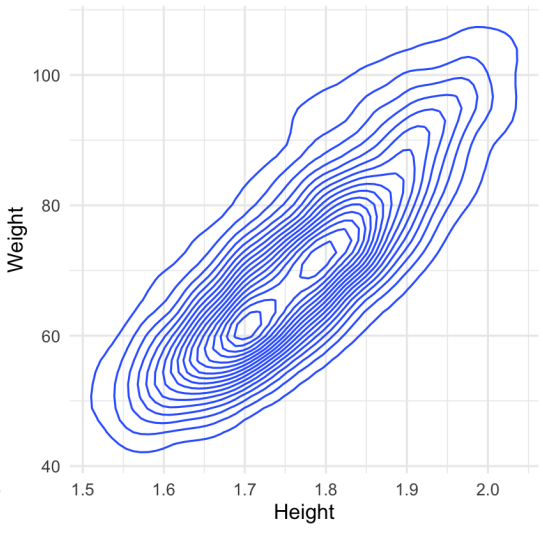
Some of the categories also are "contaminated", for example, "Athletics" is masking many different types of events. This **lurking** variable probably contributes to different relationships depending on the event. There is another variable in the data set called `Event`. Athletics cuold be further divided based on key words in this variable.

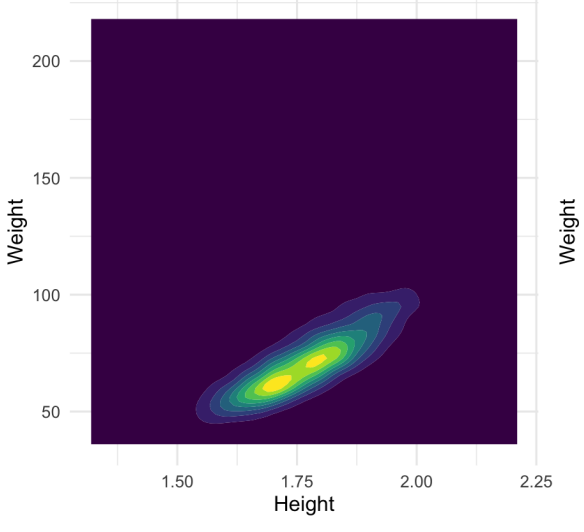
> *If you were just given the Height and Weight in this data could you have detected the presence of conditional relationships?*

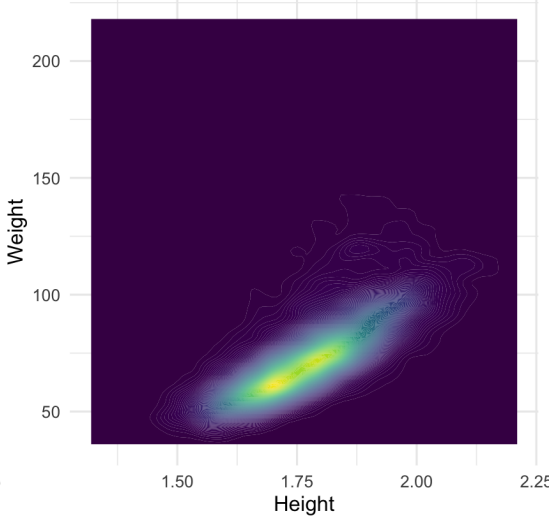
Can you see conditional dependencies?

R









There is a hint of multimodality, just a hint.

> *Its not easy!*

Focus on just women's tennis

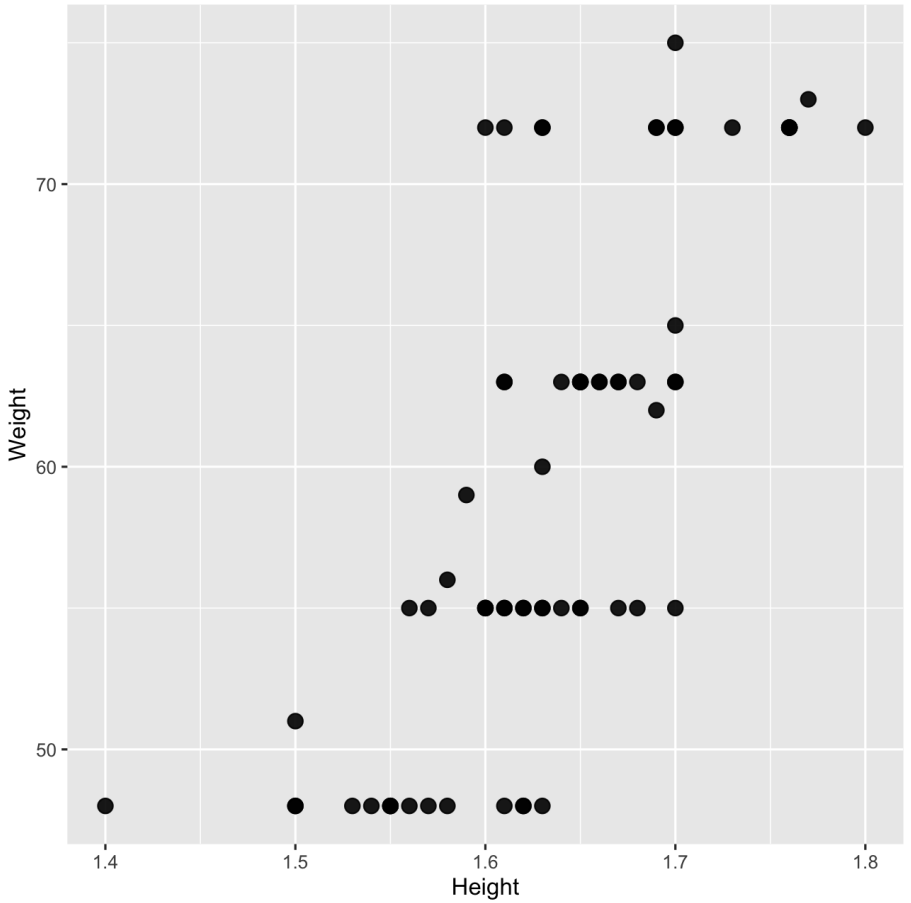
```
learn R
```

- positive
- linear
- moderate

- causation
- outliers: one outlier, maybe two: one really short and light, and one tall but skinny

Focus on just women's wrestling

learn R

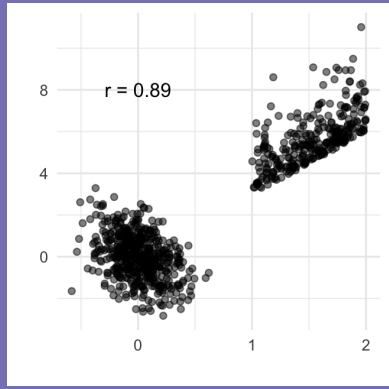


Thinking about the Olympics 2012 data

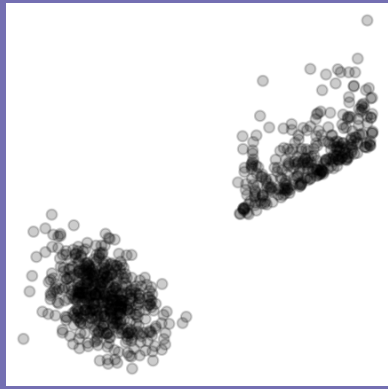
- What can this data be used for?
- What's the population?
- What could be informed by what is learned from this sample?

Re-cap on scatterplots and modifications and purpose

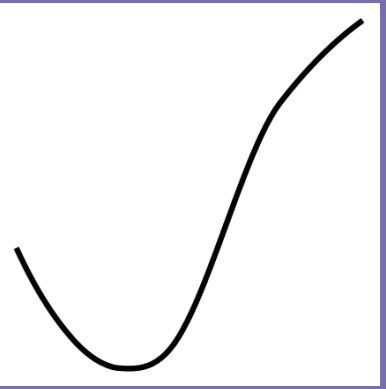
Scatterplot: raw information



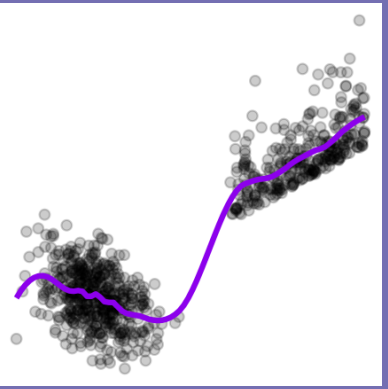
Alpha-blending: overplotting



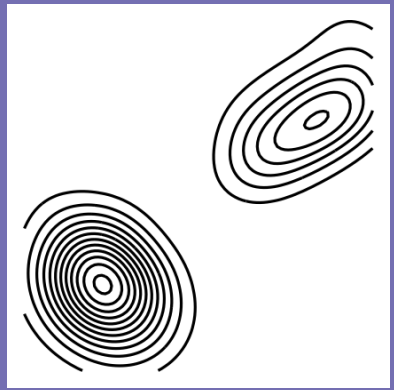
Model overlay: check the trend



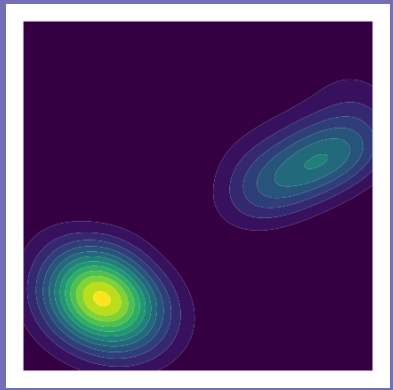
Model+data: trend/var



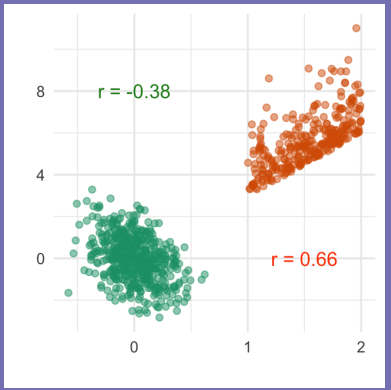
Density contours: variance, clusters



Density fill: variance, clusters



Colour: conditioning vars



Colour/density: lurking vars

That's it, for this lecture!

Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu