

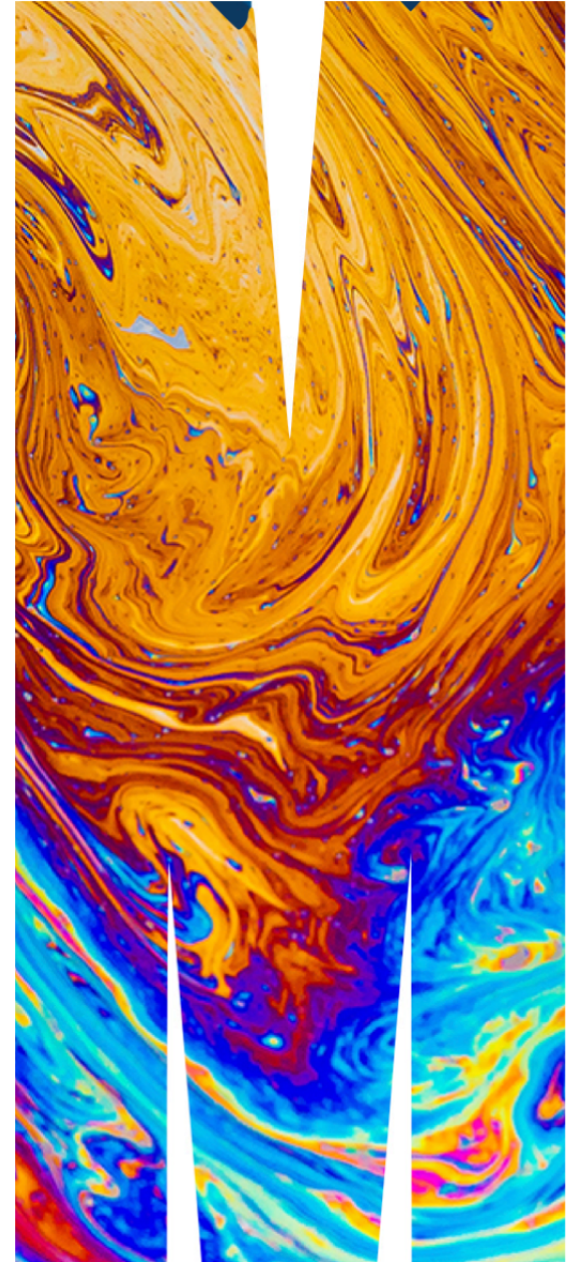
# ETC5521: Exploratory Data Analysis

## Exploring bivariate dependencies

Lecturer: *Di Cook*

✉ ETC5521.Clayton-x@monash.edu

📅 Week 6 - Session 1



“

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

*— Quote from Sherlock Holmes*

Some parts of this lecture are based on Chapter 5 of Unwin (2015) Graphical Data Analysis with R

# The story of the galloping horse

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



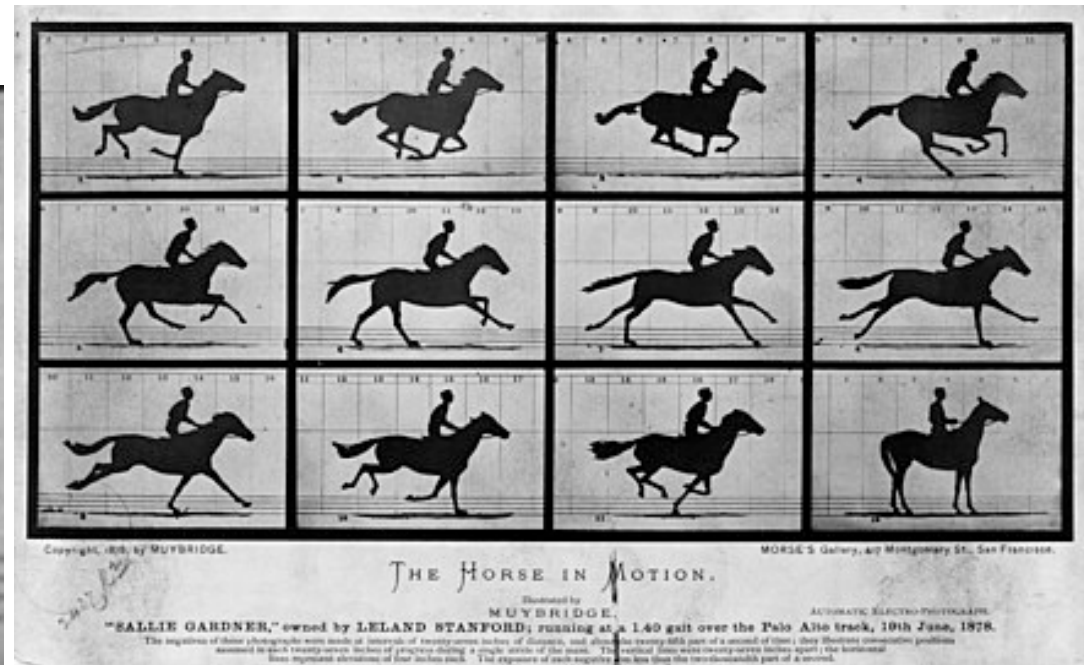
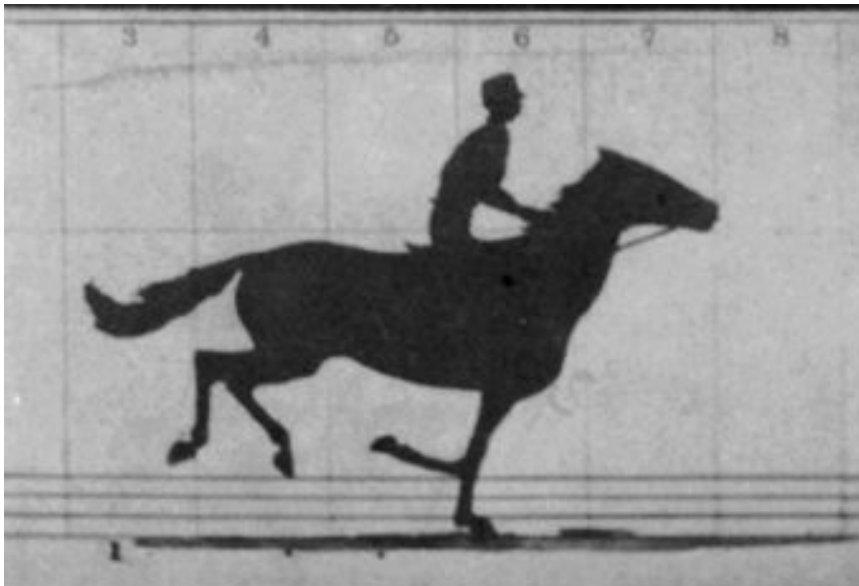
Baronet, 1794



Derby D'Epsom 1821

# The story of the galloping horse

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





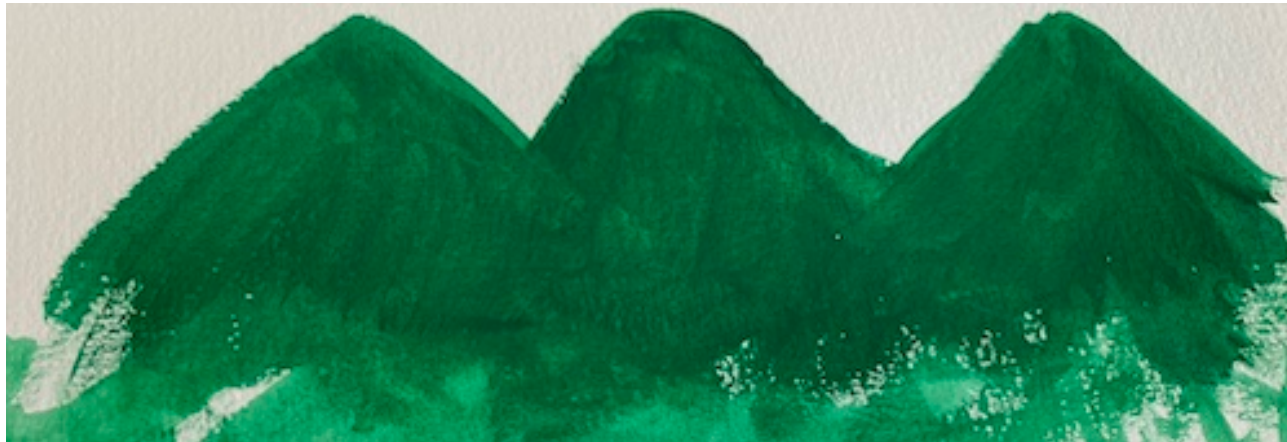
# My painting stories

hills first try   hills second try   hill photo   lemons   trees

“


*Hills are more interesting than that*


— *Mrs Robinson, my high school art teacher*



## Take-away message

We have a tendency to

 only see what other people have done or say, not what we can see, e.g. paint based on what other people have painted.

 Or impose beliefs, like trees are green.

You might discover that there is a different story.

**Try to see with fresh eyes**

# 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

*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 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.*

## Language and terminology

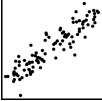
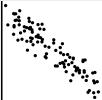
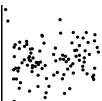
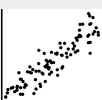
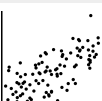
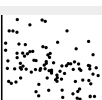
Are the words "correlation" and "association" interchangeable?

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](#)

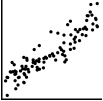

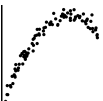
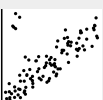
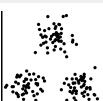



If the **relationship is not linear**, call it **association**, and avoid correlated.


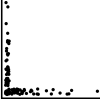

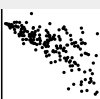
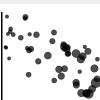
## Possible features of a pair of continuous variables 1/3

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

## Possible 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


## Possible 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.*

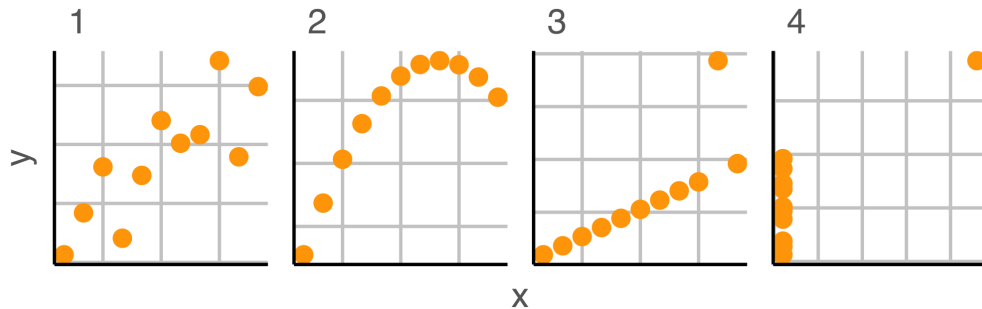
 **association:** variables may be related to one another, but through a different variable, eg ice cream sales are positively correlated with beach drownings, is most likely a temperature relationship.

 **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

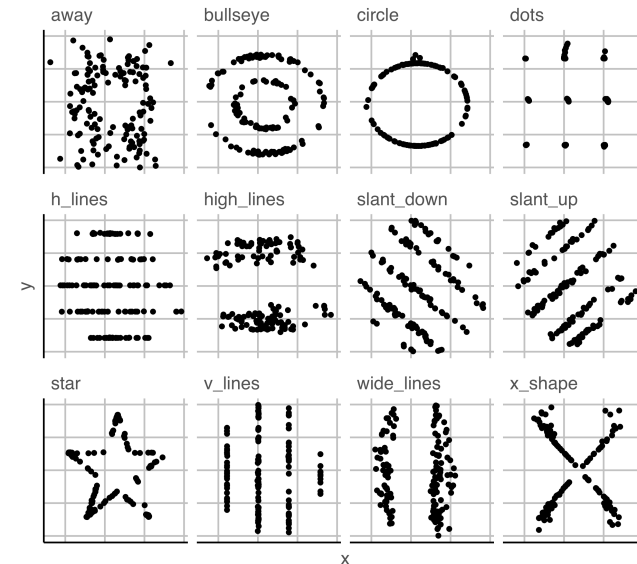
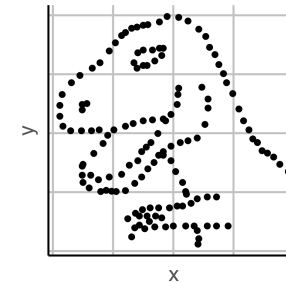
## Anscombe's quartet



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

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

## Datasaurus dozen

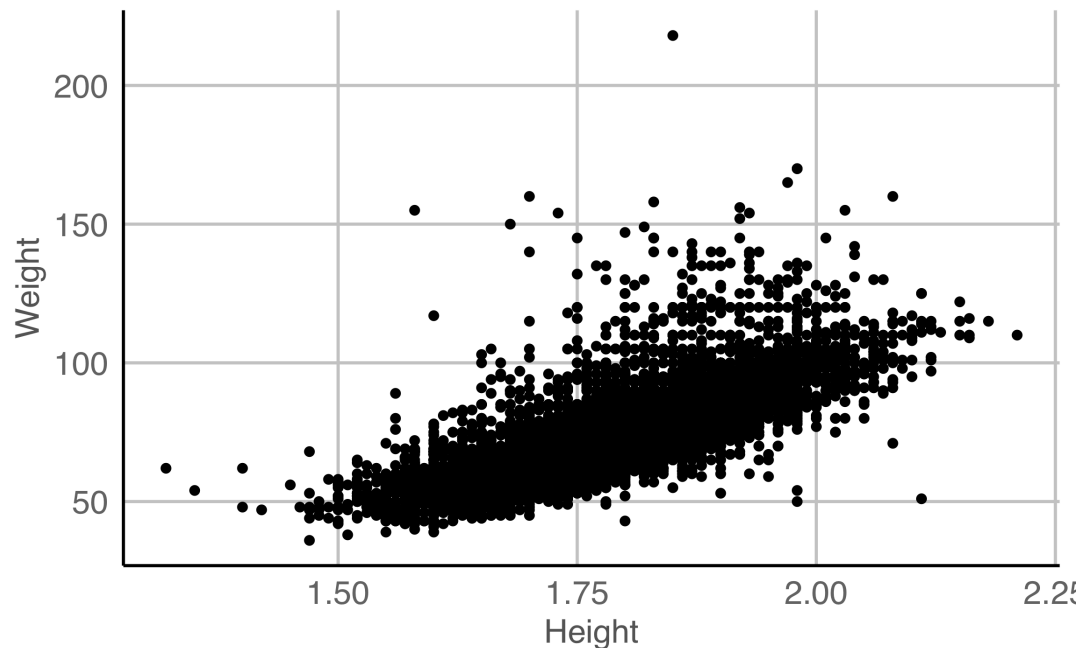


# Scatterplot case studies

# 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)
```

## How many athletes in the different sports?

Sport	n
Athletics	2119
Swimming	907
Football	596
Rowing	524
Hockey	416
Judo	368
Shooting	368
Sailing	360
Wrestling	324
Handball	319

scroll ↓

## Consolidate factor levels

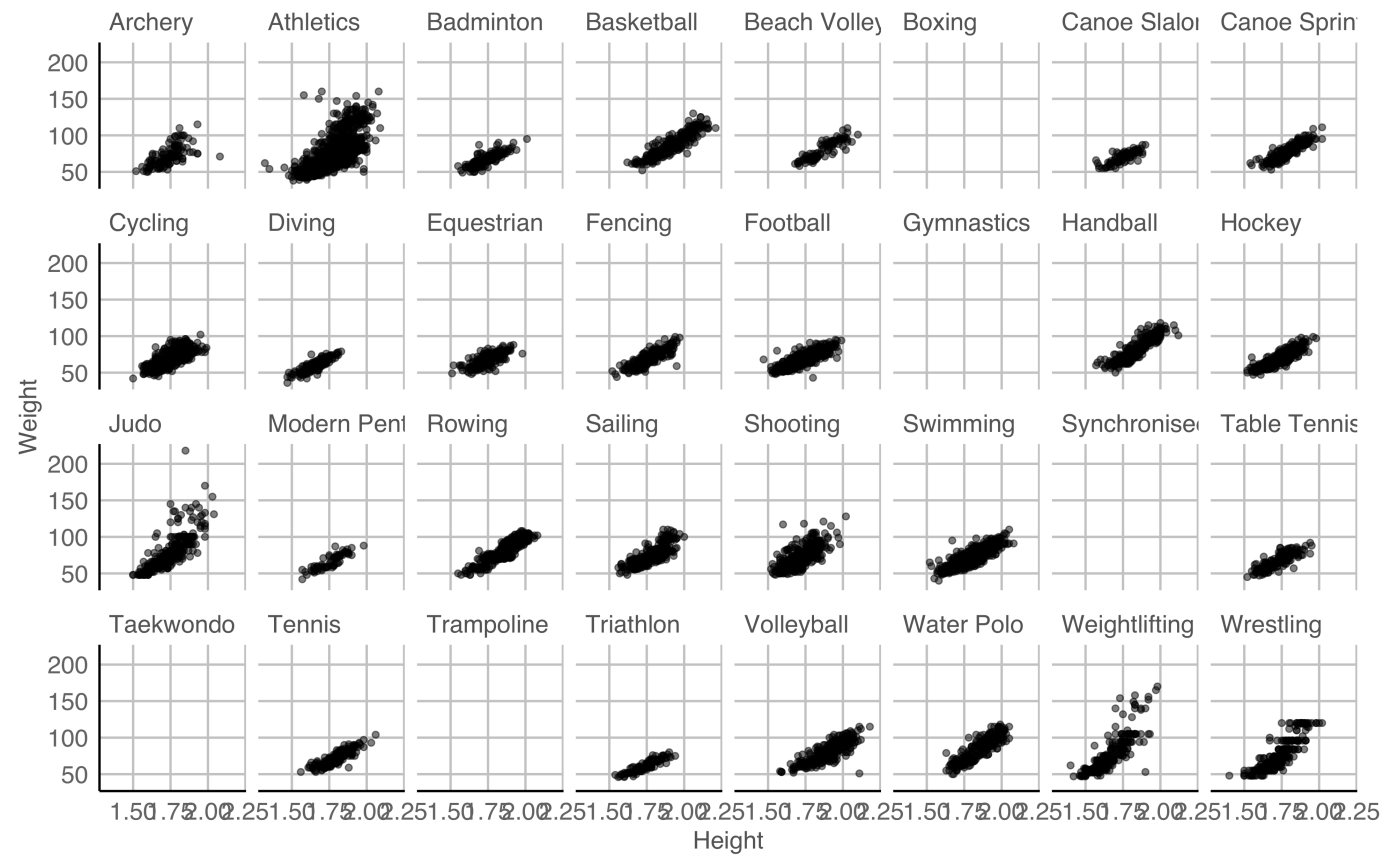
There are several cycling events that are reasonable to combine into one category. Similarly for gymnastics and athletics.

```
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



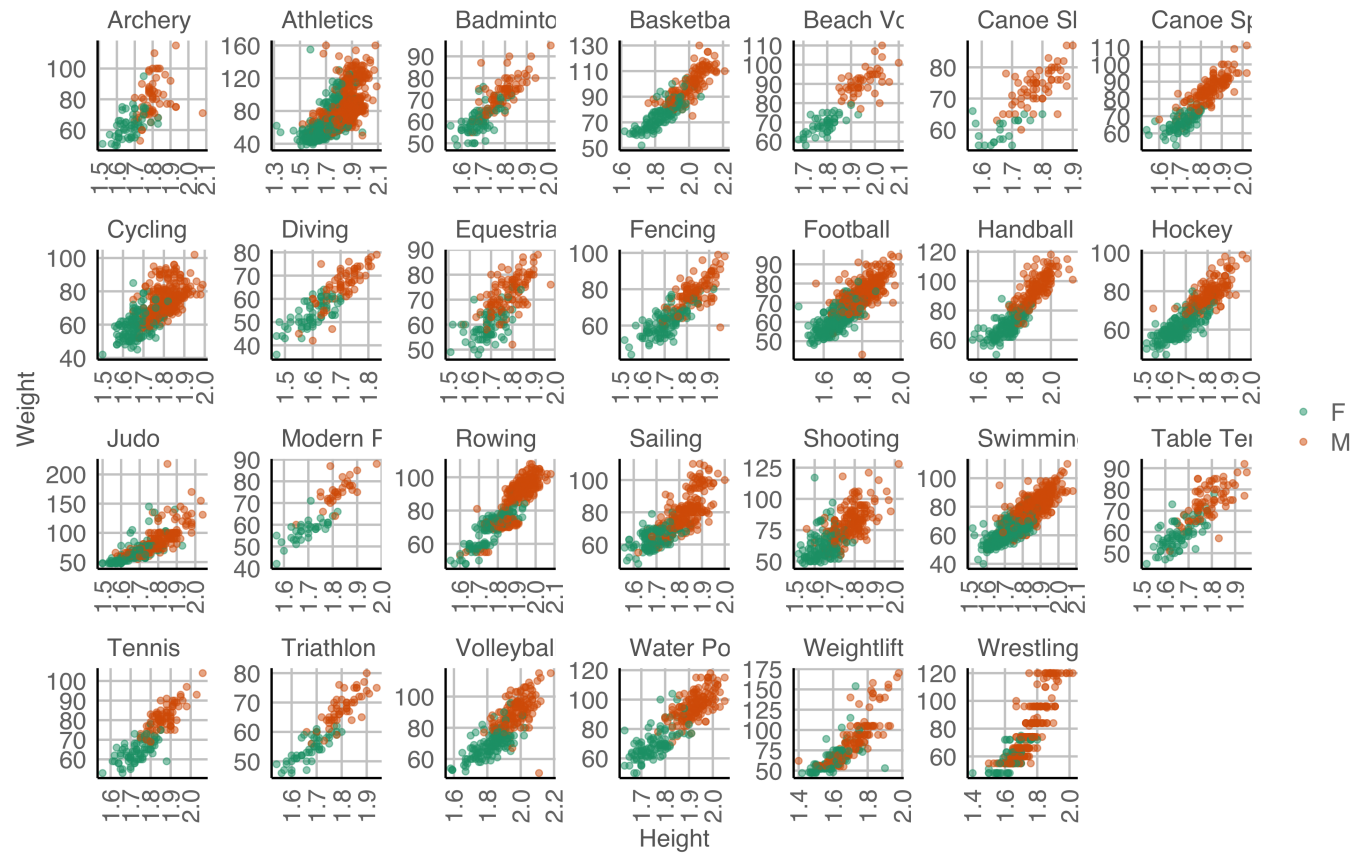
learn R



# Remove missings, add colour for sex



learn R

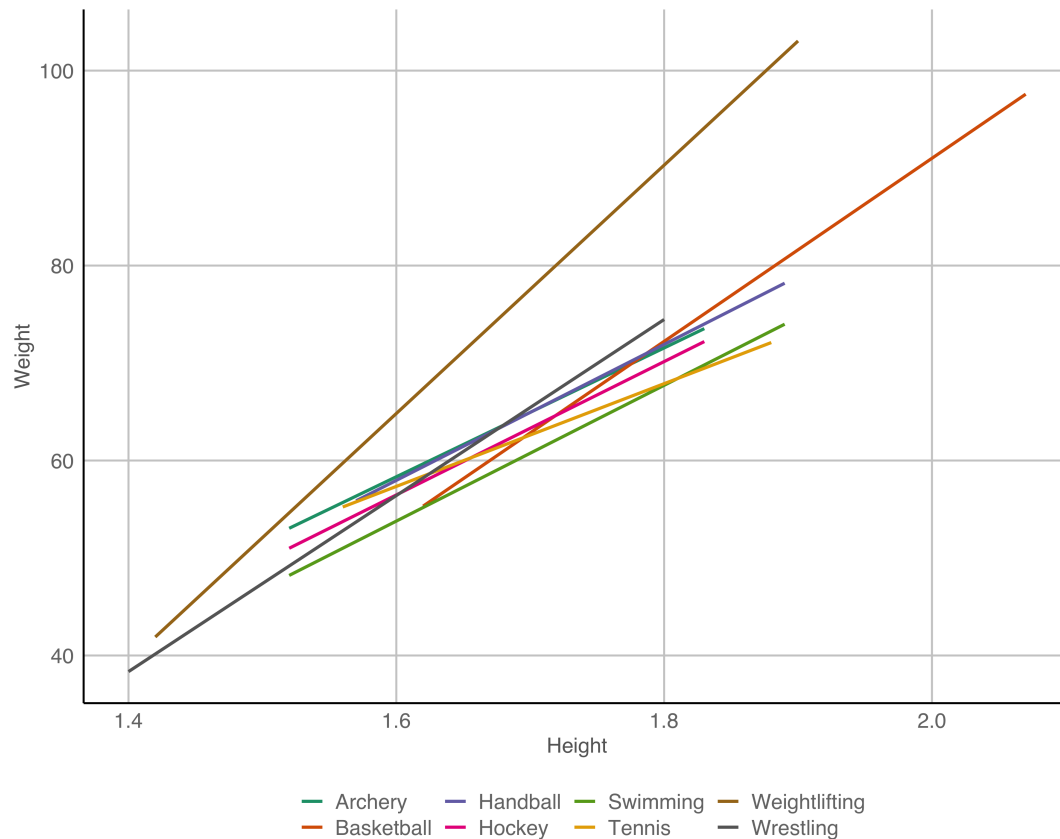




# Comparing association



R

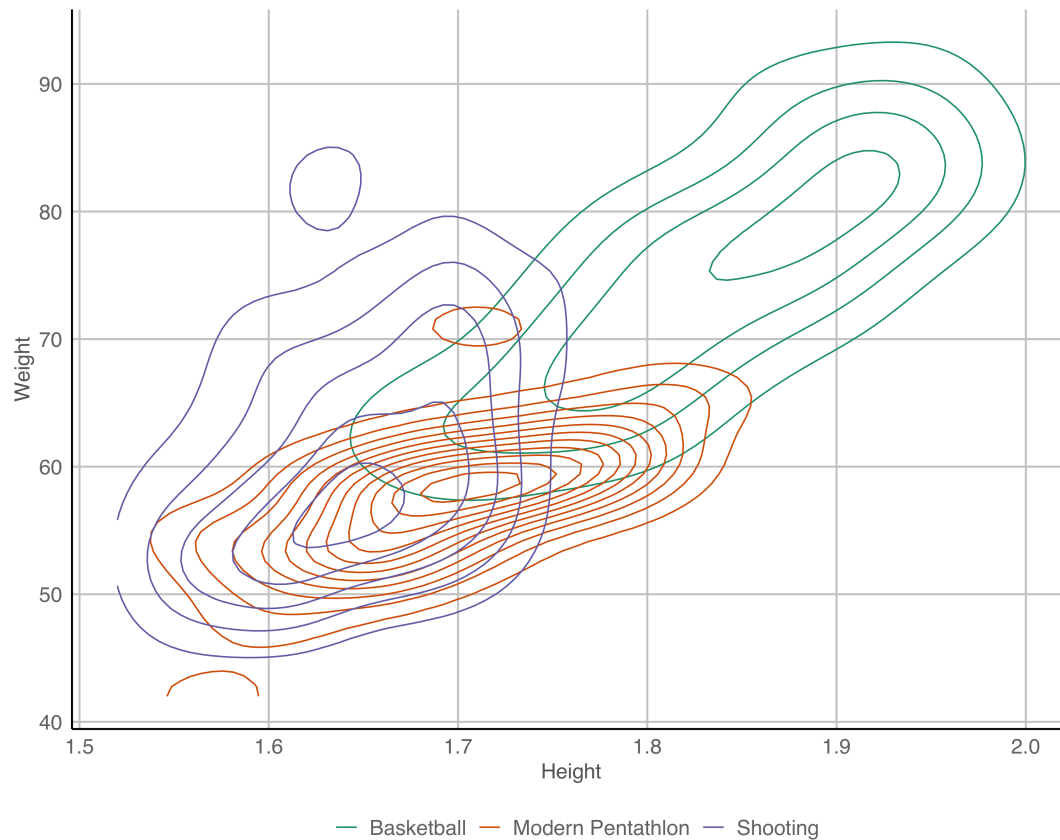


- Weightlifters are much heavier relative to height
- Swimmers are leaner relative to height
- Tennis players are a bit mixed, shorter tend to be heavier, taller tend to be lighter

# Comparing variability



R



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

## 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 could be further divided based on key words in this variable.

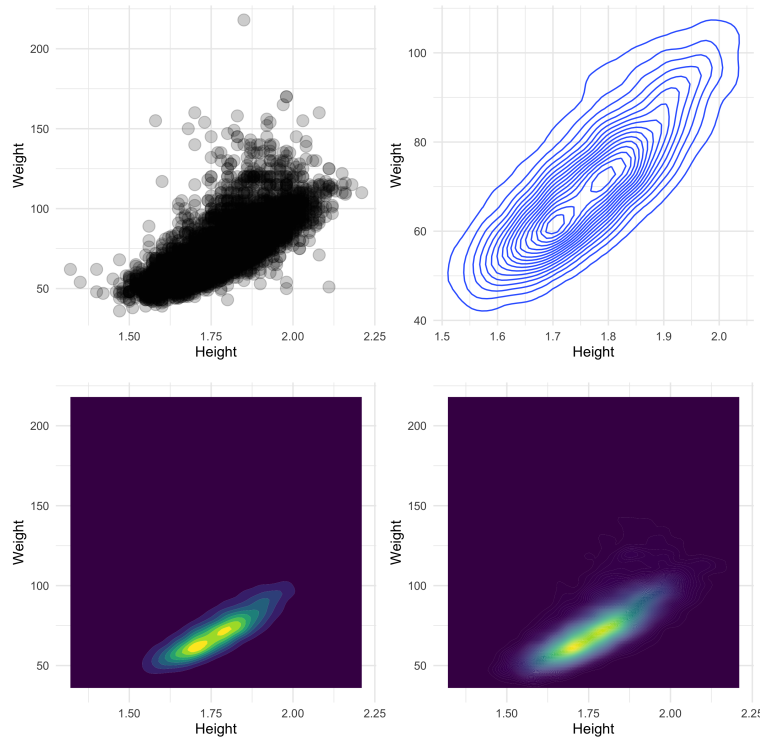


If you were just given the Height and Weight in this data could you have detected the presence of conditional relationships?

# Can you see conditional dependencies?



R



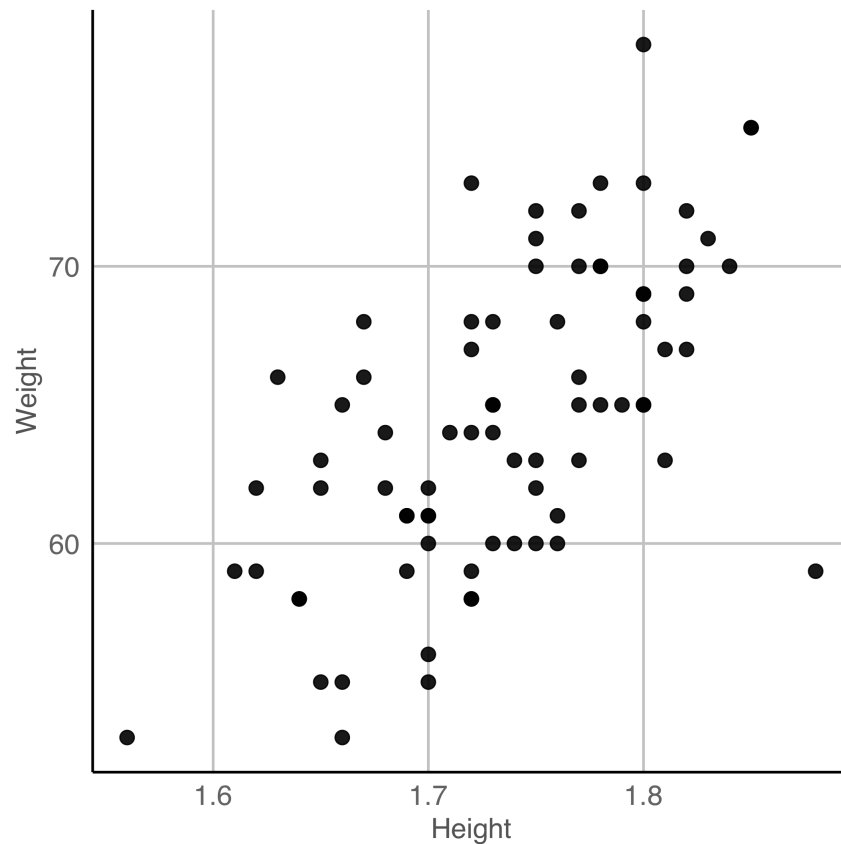
There is a hint of multimodality, barely a hint.

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.

# Focus on just women's tennis




R




 positive

 linear

 moderate

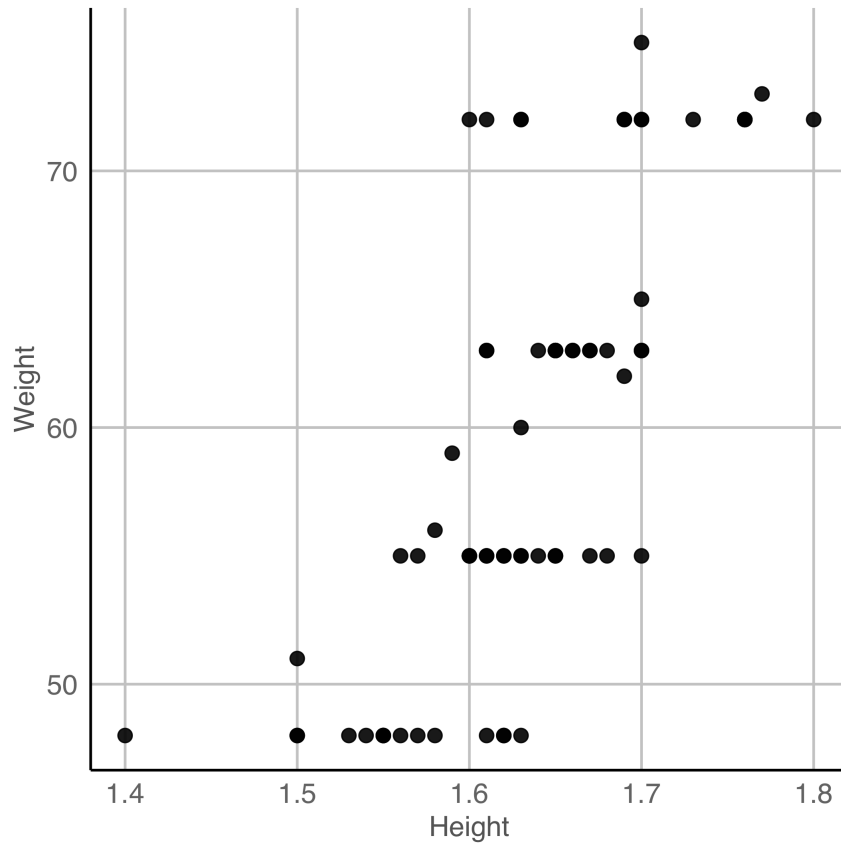
 relationship could be considered to be causation rather than association

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




What is surprising here?


## Focus on just women's wrestling



 positive

 non-linear

 moderate

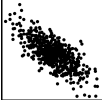
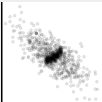
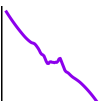
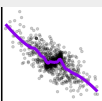

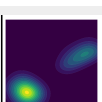
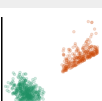

 relationship could be considered to be causation rather than association

 gaps: discreteness







## What is surprising here?

# **Review: Modifications of scatterplots for particular purposes**

Modification	Example	Purpose
none		raw information
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
colour + density		relationship with conditioning and lurking variables



# 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/>



This work is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International License](https://creativecommons.org/licenses/by-sa/4.0/).

Lecturer: *Di Cook*

✉ [ETC5521.Clayton-x@monash.edu](mailto:ETC5521.Clayton-x@monash.edu)

📅 Week 6 - Session 1

