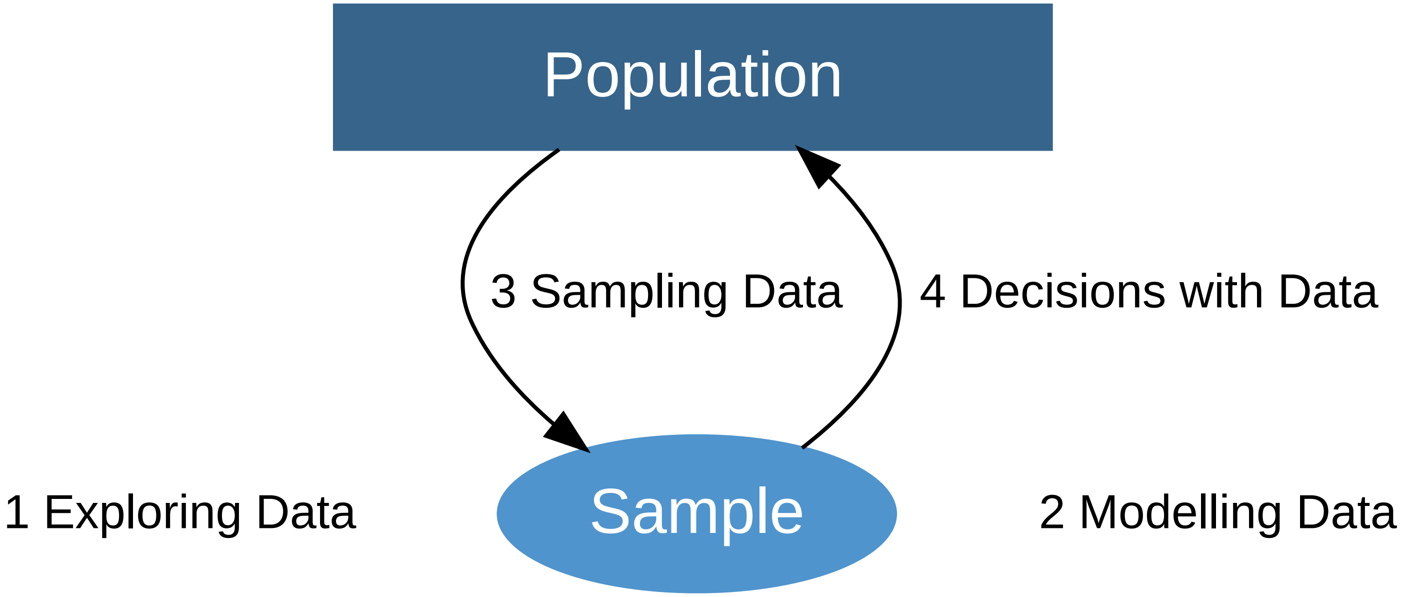
More on Correlation

Modelling Data | Linear Model

Unit Overview



Population

3 Sampling Data

4 Decisions with Data

1 Exploring Data

Sample

2 Modelling Data

Module2 Modelling Data

Normal Model

What is the Normal Curve? How can we use it to model data?

Linear Model

How can we describe the relationship between 2 variables? When is a linear model appropriate?

More on Correlation

Data Story | How is the air quality in North-West Sydney related to Central-East Sydney?

Properties of the Correlation Coefficient

Misleading correlations

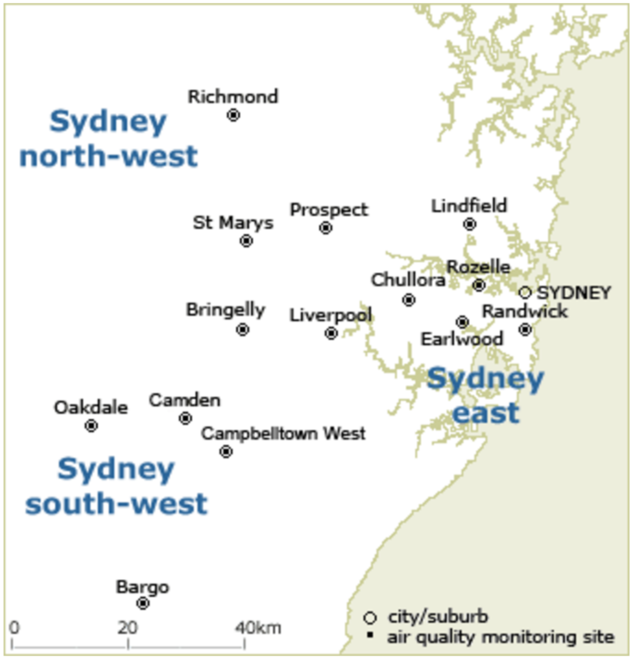
Summary

Data Story

How is the air quality in North-West Sydney related to Central-East Sydney?

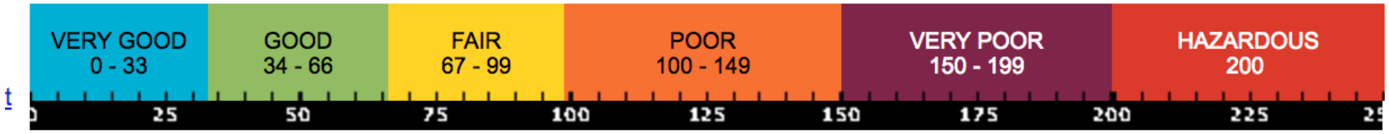
AQI data

- How do scientists monitor the air quality of Sydney?
- The Office of Environment and Heritage (OEH) has 14 active monitoring sites.



- At each site, data readings are taken for 6 pollutants:
  - Ozone
  - Nitrogen dioxide
  - Visibility
  - Carbon monoxide
  - Sulfur dioxide
  - Particles
- There are combined into the air quality index (AQI).

Why?



Statistical Thinking

Who is the AQI index useful for?

- People with sensitive respiratory conditions (e.g. people with asthma, older adults and children) should consider either cutting back or rescheduling strenuous outdoor activities when air quality is ‘poor’ or worse.
- Environmental scientists studying changes in air quality.
- Potential home-buyers.

- We will consider the data for July 2015 for two regions:
    - Sydney’s central-east (CE)
    - Sydney’s north-west (NW)

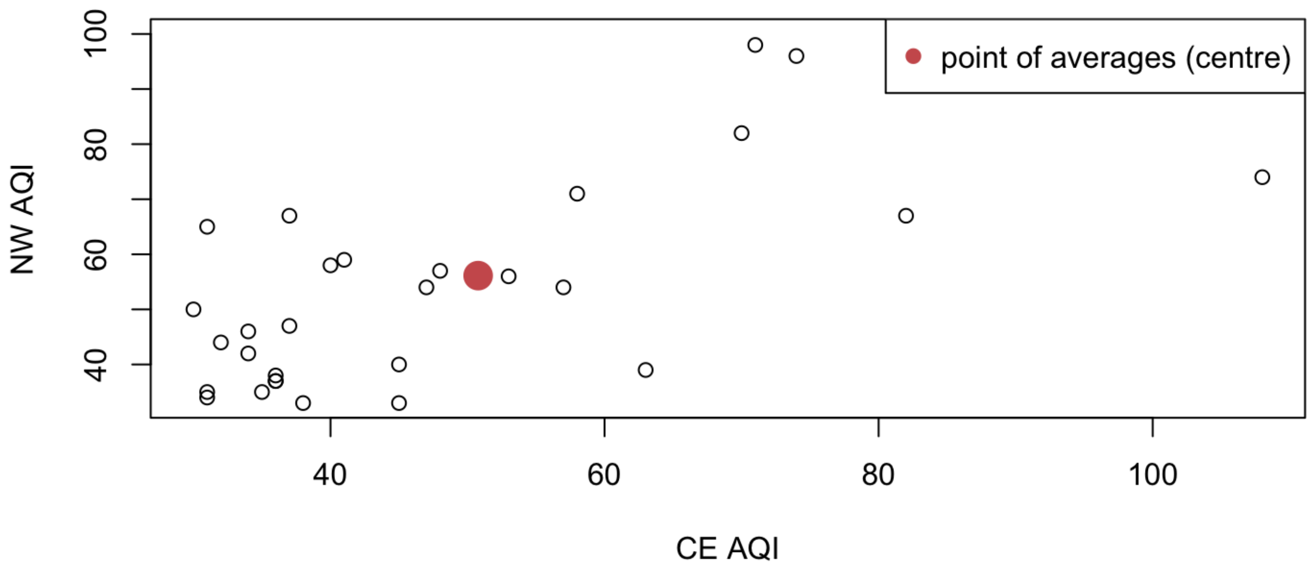
```r
library(readxl)
data = read_excel("data/AQI_July2015.xls")
```

```r
head(data)
```

```
## # A tibble: 6 x 3
##   Date       SydneyCEAQI SydneyNWAQI
##   <chr>            <dbl>       <dbl>
## 1 01/07/2015          99          92
## 2 02/07/2015          32          44
## 3 03/07/2015          70          82
## 4 04/07/2015          74          96
## 5 05/07/2015          95         100
## 6 06/07/2015          71          98
```

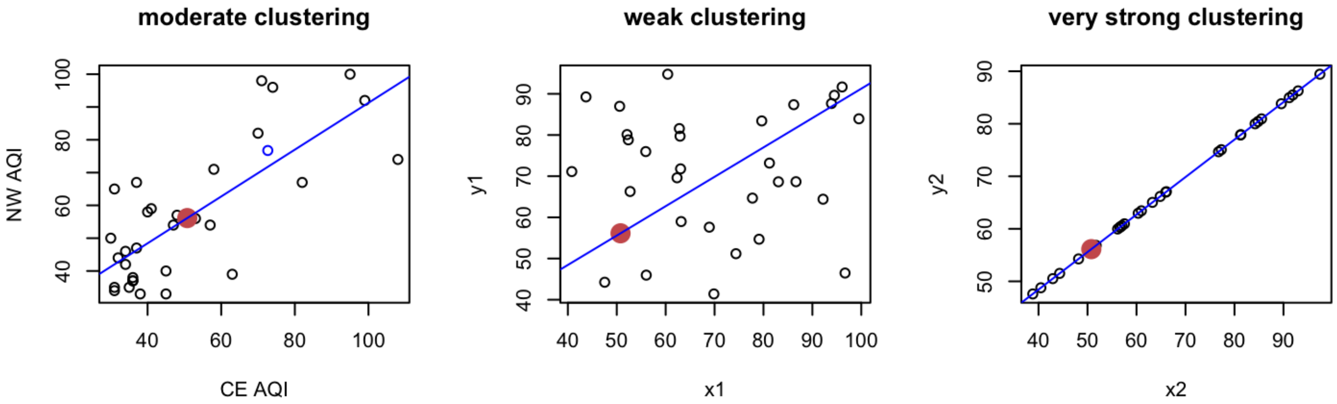
Scatter plot

```
CE = data$SydneyCEAQI
NW = data$SydneyNWAQI
plot(CE, NW, xlab="CE AQI", ylab="NW AQI")
points(mean(CE),mean(NW), col = "indianred",pch=19,cex = 2)  # point of averages (centre)
legend("topright",c("point of averages (centre)"),col="indianred",pch=19)
```



Guessing the correlation coefficient

We could **compare** our scatterplot to other data.



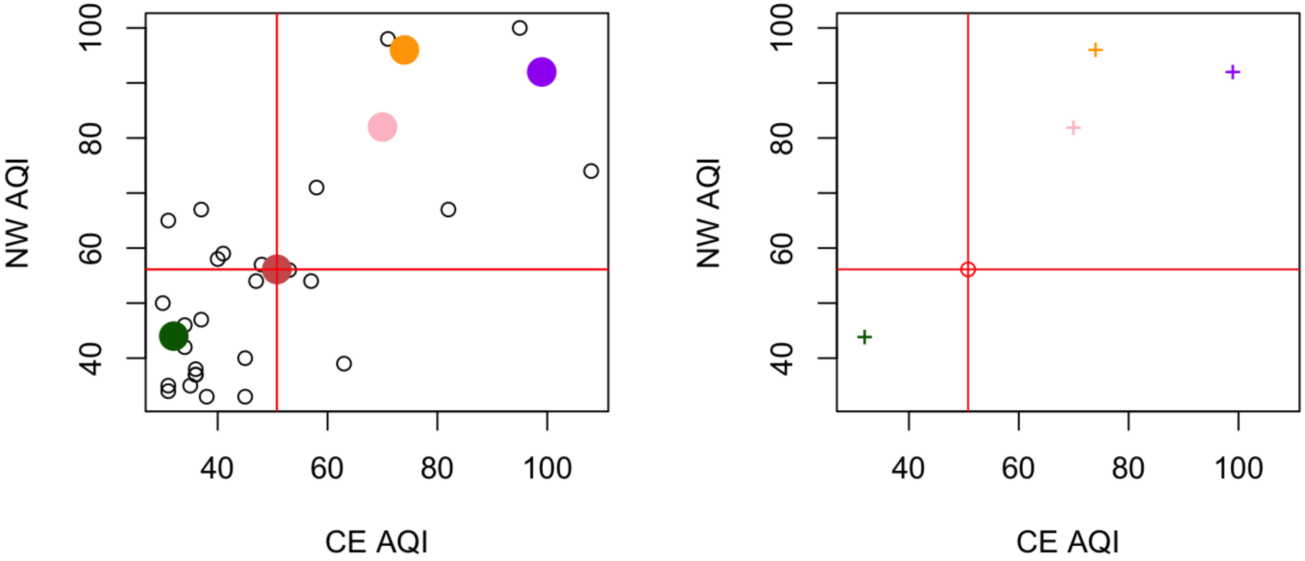
Quick calculation using R

```
cor(CE,NW)
```

```
## [1] 0.757917
```

Calculation by hand (revision)

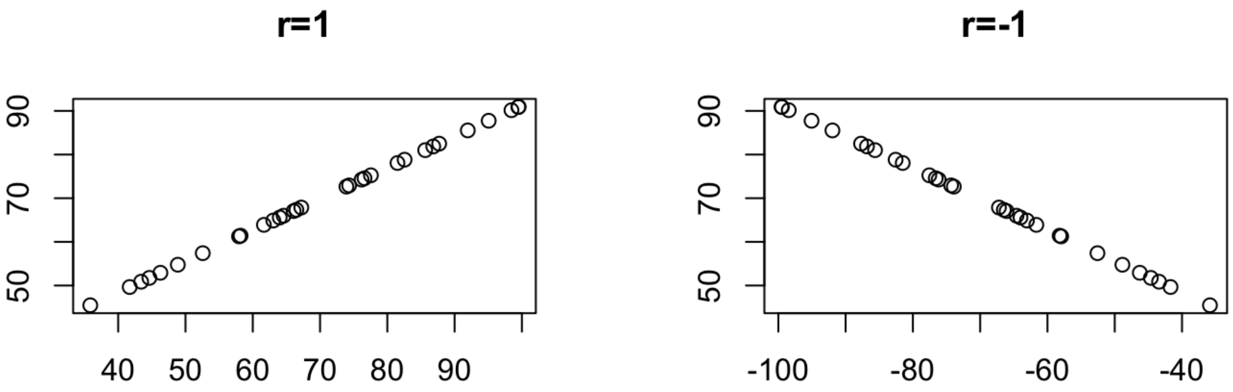
x	y	standard units	standard units	product	quadrant
		$\frac{x-50.77}{21.88}$	$\frac{y-56.13}{20.61}$	$\left(\frac{x-50.77}{21.88}\right)\left(\frac{y-56.13}{20.61}\right)$	
99	92	2.20	1.74	3.84	upper right
32	44	-0.86	-0.59	0.51	lower left
70	82	0.88	1.26	1.10	upper right
74	96	1.06	1.93	2.05	upper right
⋮	⋮	⋮	⋮	⋮	
				mean=+0.73	


NW AQI
100
80
60
40
40
60
80
100
CE AQI
NW AQI
100
80
60
40
40
60
80
100
CE AQI

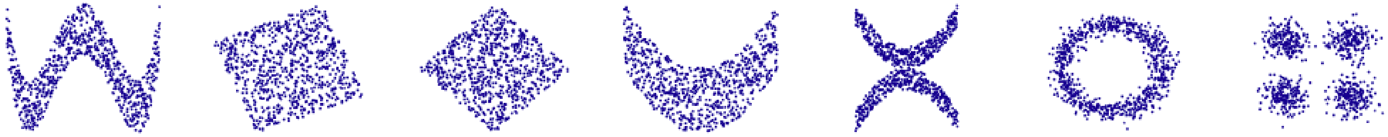
Properties of the Correlation Coefficient

Values

- The correlation coefficient (r) is a pure number (no units).
- It lies between -1 and 1 (inclusive).
- When $r = \pm 1$, all the points lie on a line (no cloud; perfect correlation)



- $r = 0$ occurs when the points don’t fit around a line.
  - But beware: this can happen in many **different** ways!



Statistical Thinking

Try the games:

🔗 1 2 3 4

Symmetry

The correlation coefficient is not affected by interchanging the variables.

```
cor(CE,NW)
```

```
## [1] 0.757917
```

```
cor(NW,CE)
```

```
## [1] 0.757917
```

Scaling

The correlation coefficient is shift and scale invariant.

```
cor(2*CE+2,3*NW-3)
```

```
## [1] 0.757917
```

Misleading correlations

Mistake1: Outliers can overly influence the correlation coefficient

Suppose there was an extra unusual reading of (100,20).

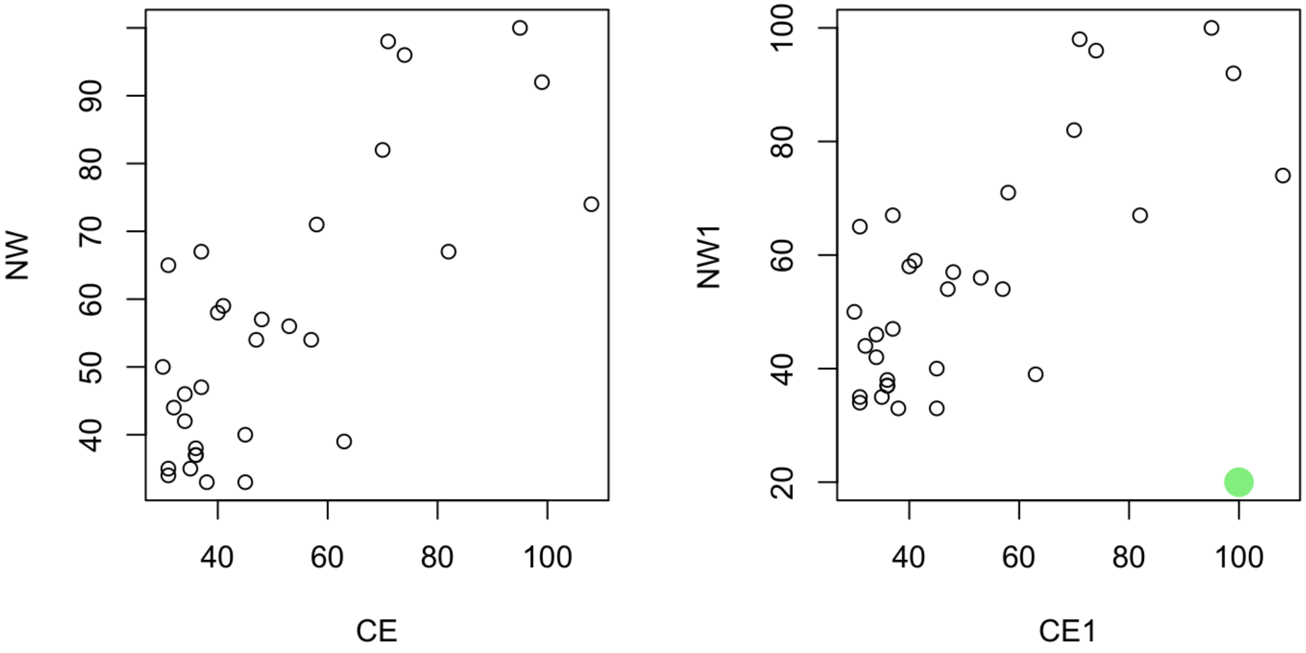
```
CE1 = c(CE,100)  # Add an extra point to data
NW1 = c(NW,20)
CE1
```

```
##  [1]  99  32  70  74  95  71  31  58 108  82  57  35  31  31  38  45  30  34  37
## [20]  37  47  63  45  53  36  36  36  40  41  48  34 100
```

```
NW1
```

```
##  [1]  92  44  82  96 100  98  65  71  74  67  54  35  34  35  33  33  50  42  47
## [20]  67  54  39  40  56  38  37  37  58  59  57  46  20
```

```
par(mfrow=c(1,2))
plot(CE,NW)
plot(CE1,NW1)
points(100,20,col="lightgreen",pch=19,cex = 2)
```



```
cor(CE,NW)
```

```
## [1] 0.757917
```

```
cor(CE1,NW1)
```

```
## [1] 0.5575432
```

💬 Which correlation coefficient best reflects the data? What are possible reasons for the outlier?

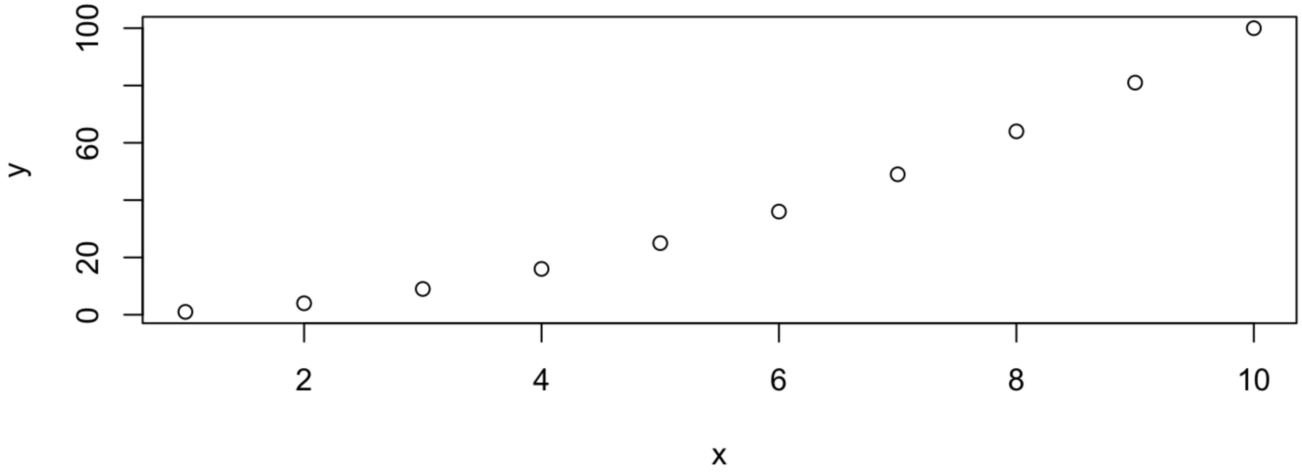
Mistake 2: Nonlinear association can’t be detected by the correlation coefficient

```
x=c(1:10)
y=x^2
cor(x,y)
```

```
## [1] 0.9745586
```

💬 What interpretation mistake could be made here?

```
plot(x,y)
```



Woops - this data should be modelled by a quadratic or even exponential curve, not a line.

Mistake3: The same correlation coefficient can arise from very different data

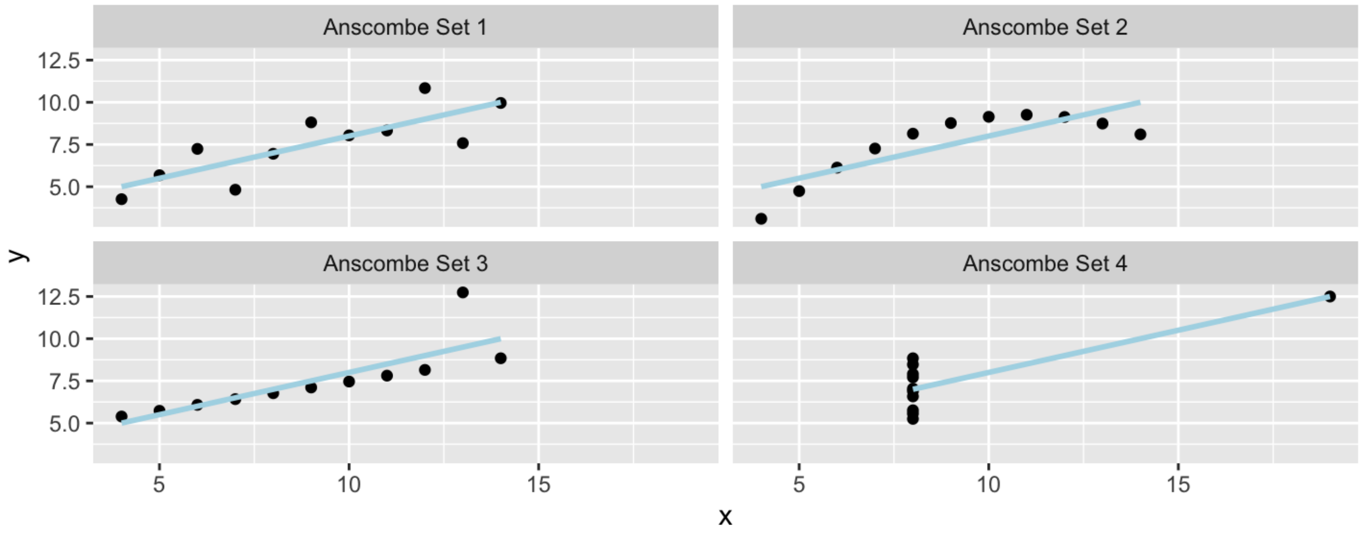
The following 4 data sets (Anscombes Quartet) have the **same** 5 numerical summaries, and hence the same value of r.

```
##        x1             x2             x3             x4           y1
##  Min.   : 4.0   Min.   : 4.0   Min.   : 4.0   Min.   : 8   Min.   : 4.260
##  1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 6.5   1st Qu.: 8   1st Qu.: 6.315
##  Median : 9.0   Median : 9.0   Median : 9.0   Median : 8   Median : 7.580
##  Mean   : 9.0   Mean   : 9.0   Mean   : 9.0   Mean   : 9   Mean   : 7.501
##  3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.:11.5   3rd Qu.: 8   3rd Qu.: 8.570
##  Max.   :14.0   Max.   :14.0   Max.   :14.0   Max.   :19   Max.   :10.840
##        y2              y3              y4
##  Min.   :3.100   Min.   : 5.39   Min.   : 5.250
##  1st Qu.:6.695   1st Qu.: 6.25   1st Qu.: 6.170
##  Median :8.140   Median : 7.11   Median : 7.040
##  Mean   :7.501   Mean   : 7.50   Mean   : 7.501
##  3rd Qu.:8.950   3rd Qu.: 7.98   3rd Qu.: 8.190
##  Max.   :9.260   Max.   :12.74   Max.   :12.500
```

```
## [1] 0.8164205 0.8162365 0.8162867 0.8165214
```

But look at the scatter plots!

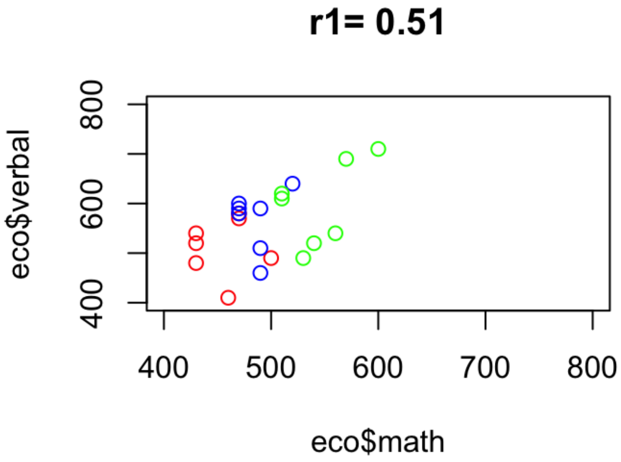
```
##                Set x        y
## 1 Anscombe Set 1 9 7.500909
## 2 Anscombe Set 2 9 7.500909
## 3 Anscombe Set 3 9 7.500000
## 4 Anscombe Set 4 9 7.500909
```

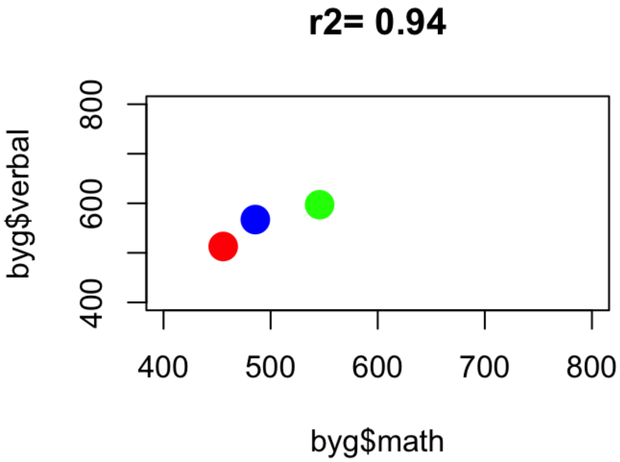


Mistake 4: Rates of averages tend to inflate the correlation coefficient

- An **ecological correlation** (or spatial correlation) is the correlation between two variables that are group means or rates.
- For example, if we recorded the AQI at many stations across NW Sydney and CE Sydney,and then calcuated the average for the 2 areas.
- Ecological correlations tend to overestimate the strength of association between the two variables.
- See Freedman et al, Statistics p148-149.

Example





- The 1st plot has all 3 sets of data combined: correlation = 0.51 (not very strong).
- The 2nd plot has the averages of the 3 data sets: correlation = 0.94 (very strong).

Mistake 5: Association is not causation

- Correlation measures association.
- But as discussed in Design of Experiments, association does not necessarily mean causation.
- Both variables may be simultanesouly influenced by a 3rd variable (confounder).

💬 Invent your own example of a Spurious Correlation.

Mistake 6: Small SDs can make the correlation look bigger

- The appearance of a scatter diagram depends on the SDs.
- The correlation coefficient measures clustering, not in absolute terms, but relative to the SDs.
- See Freedman et al, Statistics p145.

Summary

The correlation coefficient has special properties. Care needs to be taken to avoid interpretation mistakes.

Key Words

pure number, symmetry, shift and scale invariant, outliers, nonlinear association, ecological (spatial) correlation,