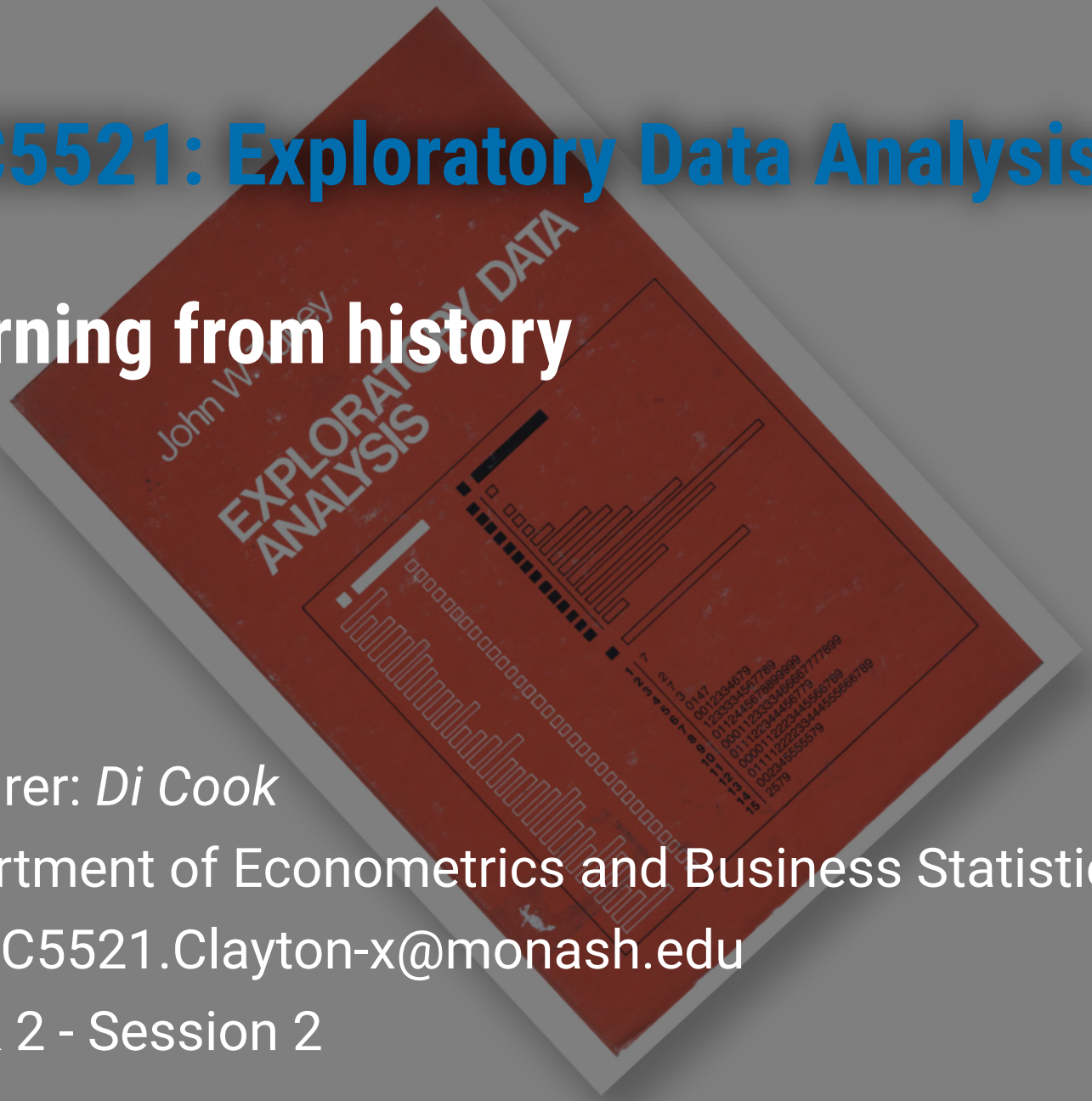
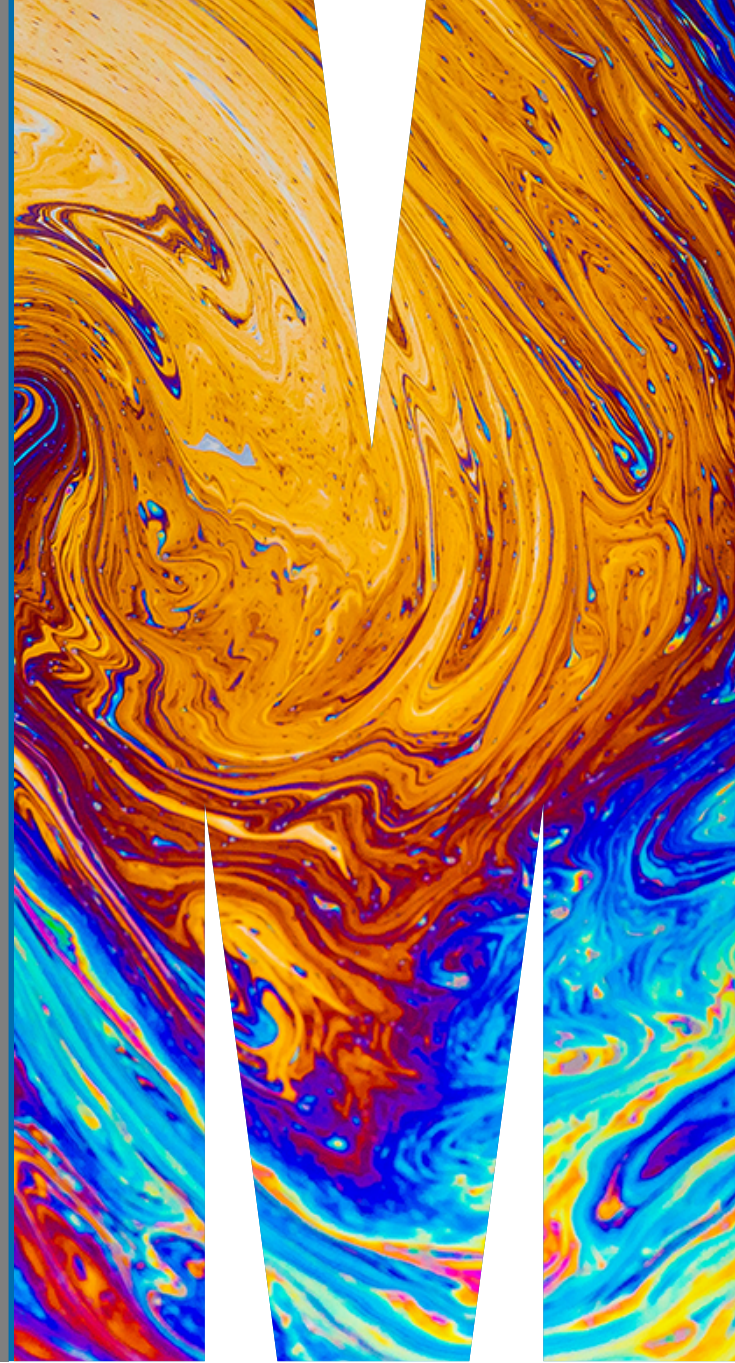
ETC5521: Exploratory Data Analysis



Learning from history

Lecturer: *Di Cook*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 2 - Session 2



Easy summaries -- numerical and graphical

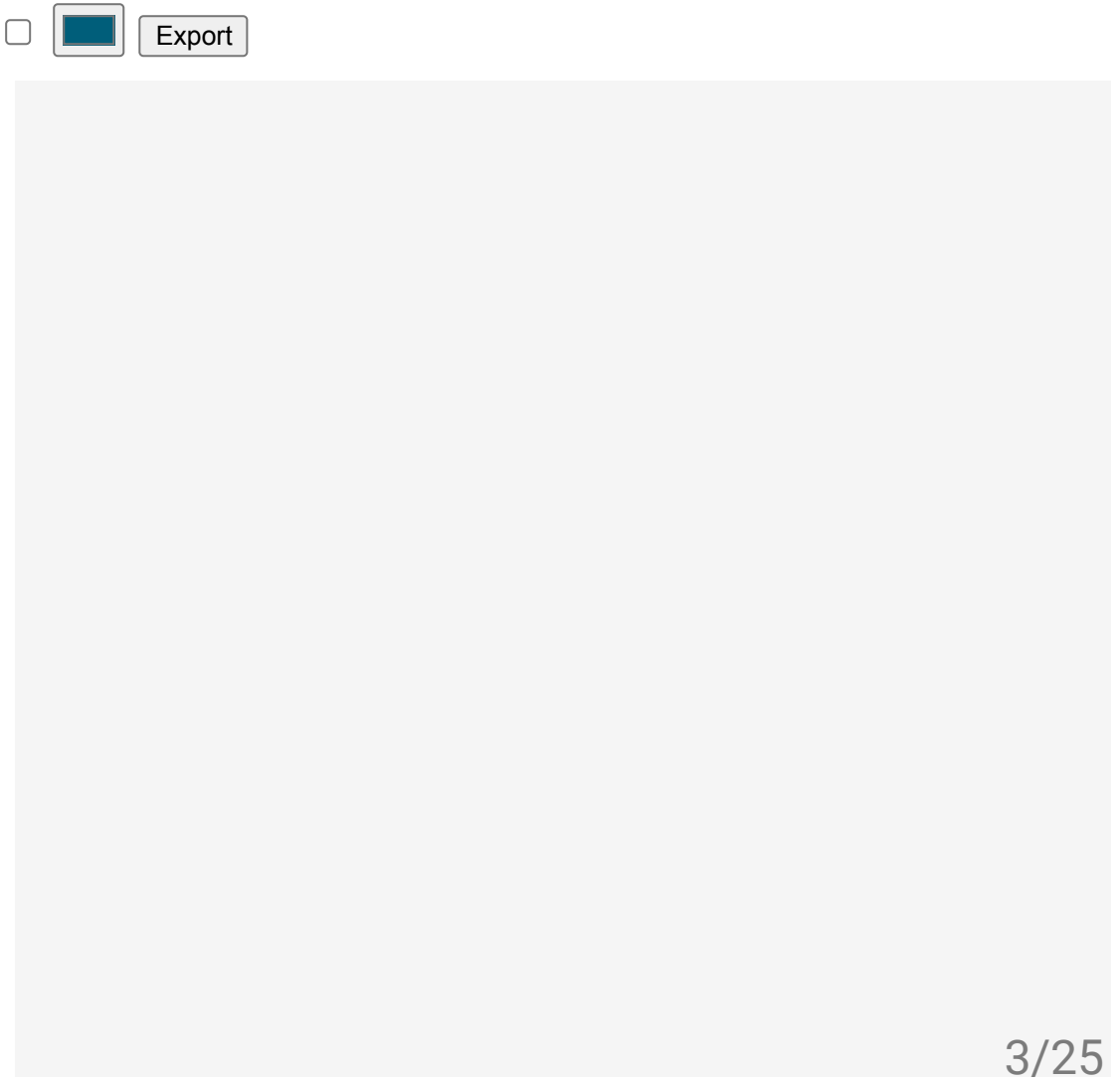
Hinges and 5-number summaries

```
##  [1] -3.2 -1.7 -0.4  0.1
##  [5]  0.3  1.2  1.5  1.8
##  [9]  2.4  3.0  4.3  6.4
## [13]  9.8
```

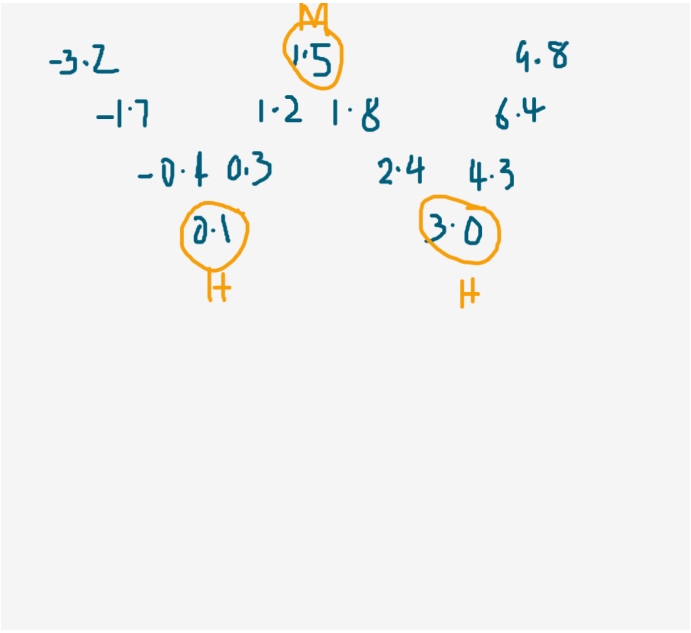
You know the median is the middle number. What's a hinge?

There are 13 data values here, provided already sorted. Write them into a down-up-down-up pattern, evenly.

Median will be 7th, hinge will be 4th from each end.



Hinges and 5-number summary



Hinges illustrated

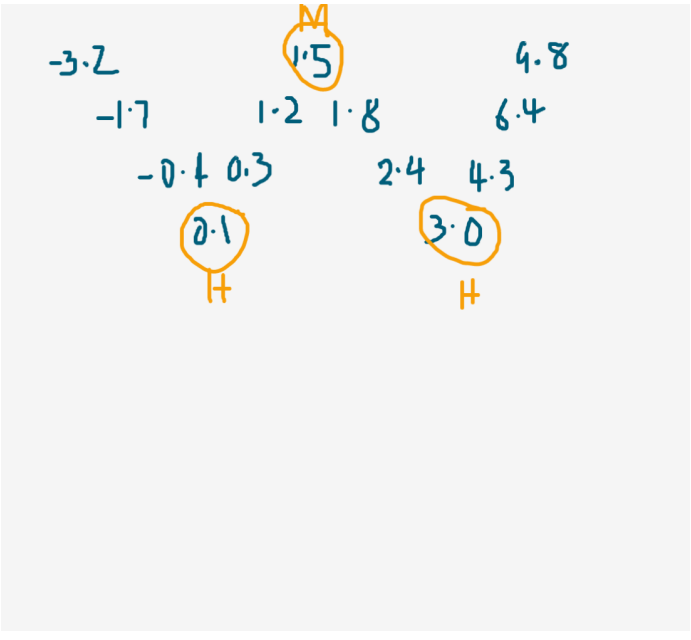
A) The 17 AUTO PRICES of EXHIBIT 1––in folded form

(1) (M) (1)
150 895 1895
250 895 1099 1775
688 895 1166 1699
695 795 1333 1693
795 1499
(H) (H)

[17 prices, 1HMH1: 150, 795, 895, 1499, 1895 dollars]

hinges are alternatively known as Q1 and Q3.

box-and-whisker display



Starting with a 5-number summary

#13

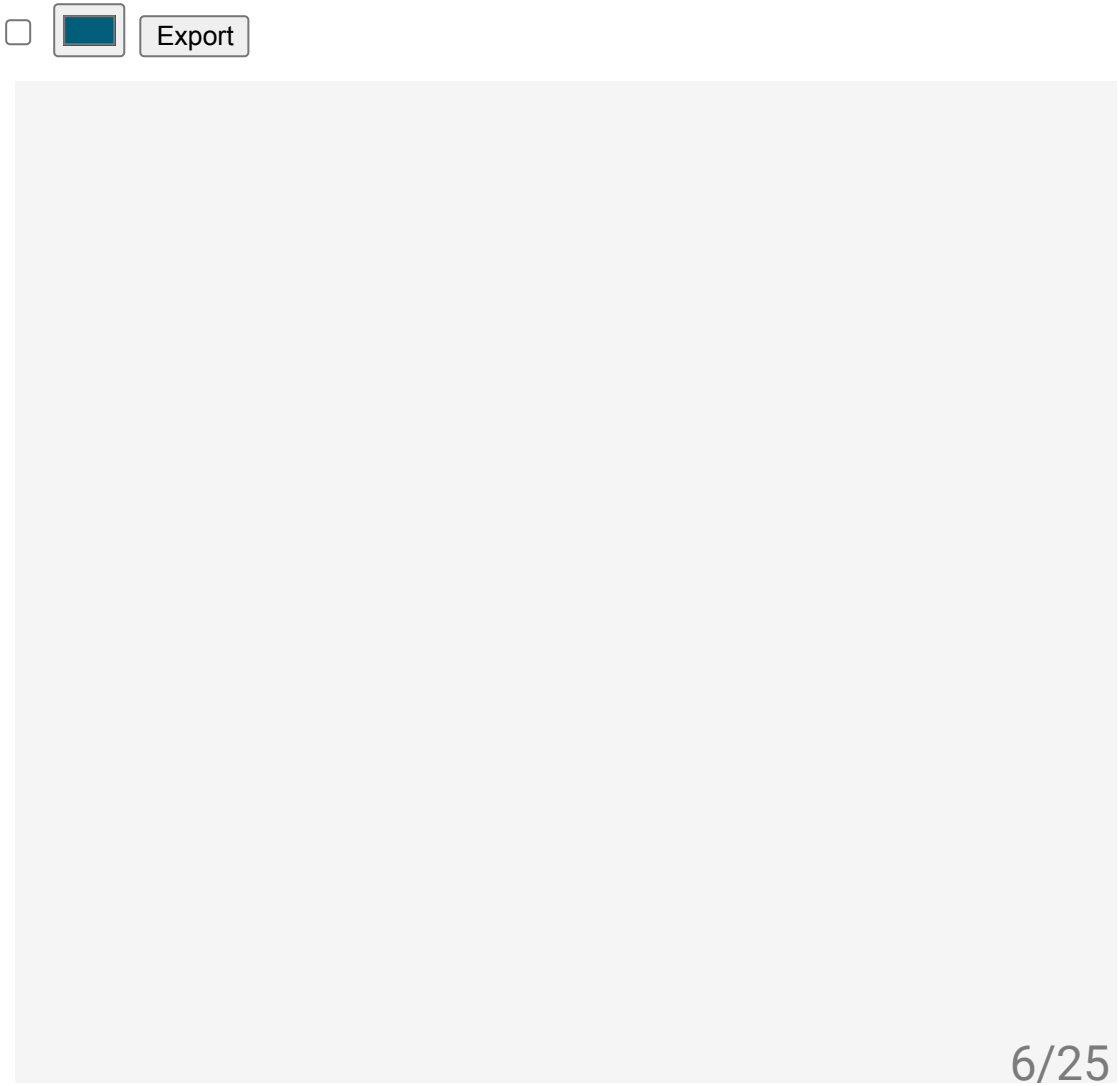
M7	1.5	
H4	0.1	3.0
1	−3.2	9.8

box-and-whisker display

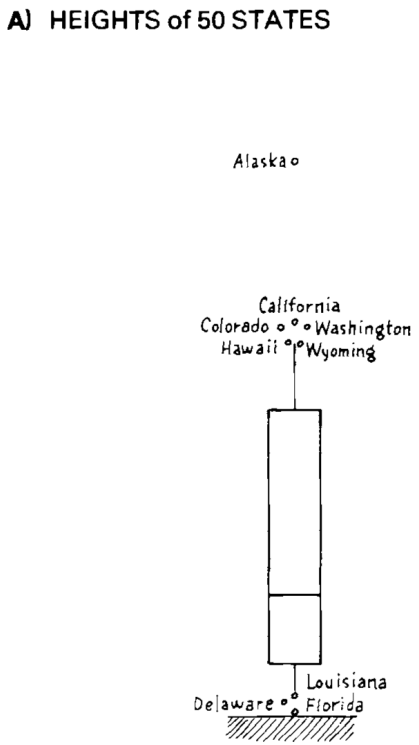
Starting with a 5-number summary

#13

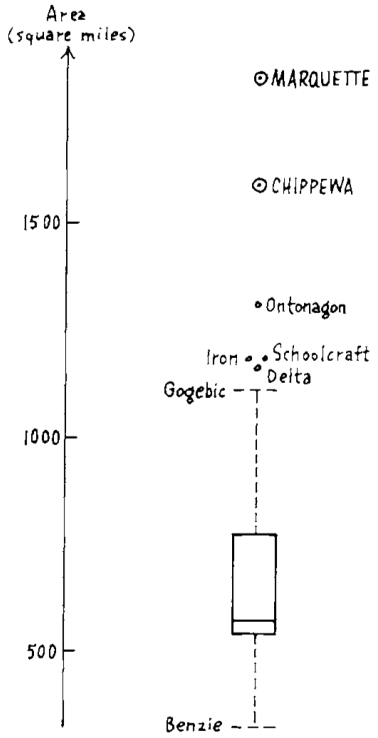
M7		1.5	
H4	0.1		3.0
1	−3.2		9.8



Identified end values



Why are some individual points singled out?



Rules for this one may be clearer?

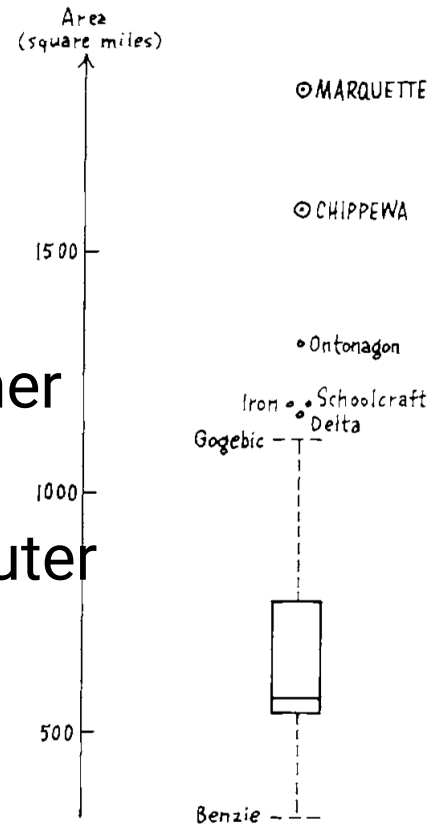
🙀 Isn't this imposing a belief?

There is no excuse for failing to plot and look

Another Tukey wisdom drop

Fences and outside values

- H-spread: difference between the hinges (we would call this Inter-Quartile Range)
- step: 1.5 times H-spread
- inner fences: 1 step outside the hinges
- outer fences: 2 steps outside the hinges
- the value at each end closest to, but still inside the inner fence are "adjacent"
- values between an inner fence and its neighbouring outer fence are "outside"
- values beyond outer fences are "far out"
- these rules produce a SCHEMATIC PLOT



New statistics: trimeans

The number that comes closest to

$$\frac{\text{lower hinge} + 2 \times \text{median} + \text{upper hinge}}{4}$$

is the **trimean**.

Think about trimmed means, where we might drop the highest and lowest 5% of observations.

Letter value plots

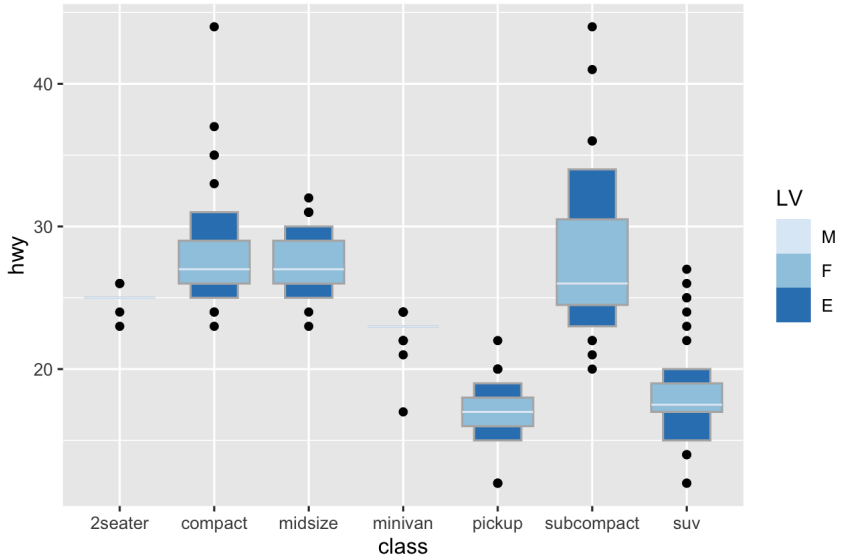
Why break the data into quarters? Why not eighths, sixteenths? k-number summaries?

What does a 7-number summary look like?

	50	populations	
(Seven-number summary)	M25h	246	
	H13	89 432	343 (H-spread)
	E7	63 782	719 (E-spread)
	1	23 1678	1655 (range)

How would you make an 11-number summary?

```
library(lvplot)
p <- ggplot(mpg,
            aes(class, hwy))
p + geom_lv(aes(fill=..LV..)) +
  scale_fill_brewer()
```



Box plots are ubiquitous in use today.

🐶 🐱 Mostly used to compare distributions, multiple subsets of the data.

Puts the emphasis on the middle 50% of observations, although variations can put emphasis on other aspects.

Easy re-expression

Logs, square roots, reciprocals

What you need to know about logs?

- 🚀 how to find good enough logs fast and easily
- 🚀 that equal differences in logs correspond to equal ratios of raw values. (This means that wherever you find people using products or ratios-- even in such things as price indexes--using logs--thus converting producers to sums and ratios to differences--is likely to help.)

The most common transformations are logs, sqrt root, reciprocals, reciprocals of square roots

-1, -1/2, +1/2, +1

What happened to ZERO?

It turns out that the role of a zero power, is for the purposes of re-expression, neatly filles by the logarithm.

Re-express to symmetrize the distribution

Deaths for 59 selected causes, 1964 (total 1,798,051, less "all other diseases 54,000")

A) SMALL COUNTS

Raw	Log	Cause
17	1.23	Polio
42	1.62	Diphtheria
93	1.97	Whooping cough (WC below)
95	1.98	Scarlet fever and strep throat (SFST below)

B) RAW VALUES--in 100's

0*	7,3,8,4,2
1*	7,8,3,1
2	6,5
3	5,8
4	9,4,6,4
5*	9,4
6	5
7	6
8	2
9*	9,9,8
1**	35,35,67,59,22,11,10
2	62,77,57,34,32,03,52,53,06
3	28,23,72,07
4	92,00,69
5**	32,74,78,69
0***	932,
1***	982,
2	
3	
4***	454,

C) LOGS--in 0.1's

·		
s	7	Polio
f		
t	3	Diph
−0*	00	WC,SFST
0*		
t	3	Abortion
f	4	Dysentery
s	6	Measles
·	89	
1*	01	
t	22	
f	4455	
s	666677	
·	8899	
2*	00000011	
t	23333	
f	444444555	
s	6667777	
·	9†	
3*		
t	3†	
f		
s	7†	
·		

Power ladder

→ fix RIGHT-skewed values

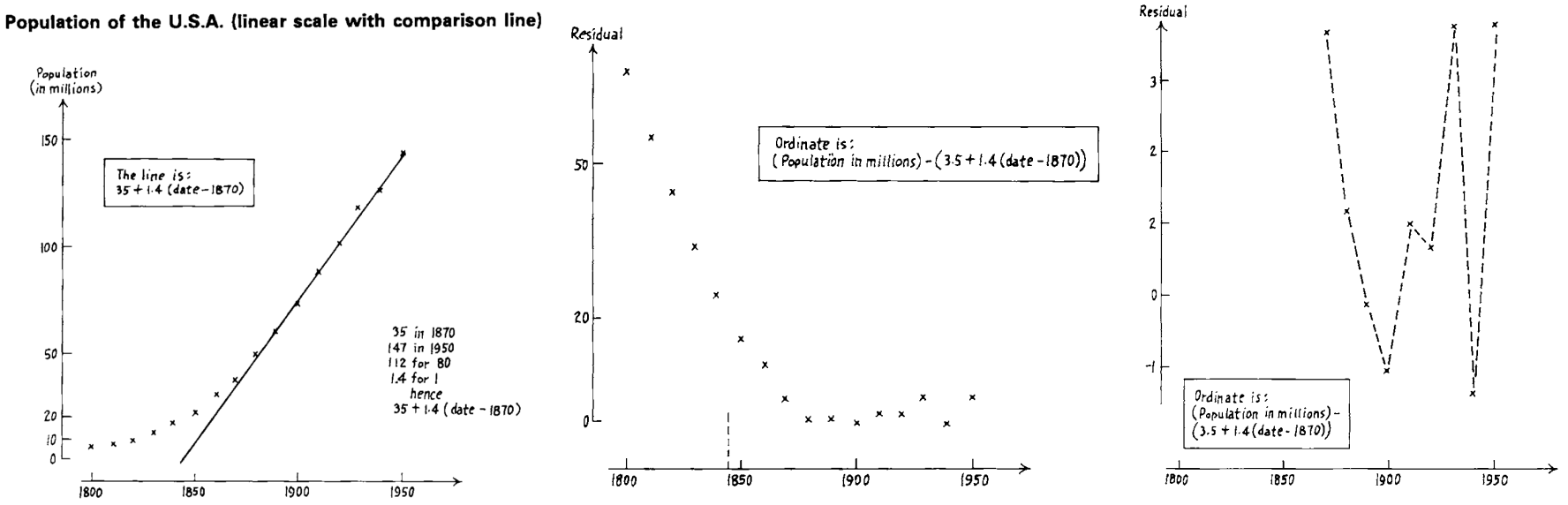
-2, -1, -1/2, 0 (log), 1/3, 1/2, 1, 2, 3, 4

← fix LEFT-skewed values

We now regard re-expression as a tool, something to let us do a better job of grasping. The grasping is done with the eye and the better job is through a more symmetric appearance.

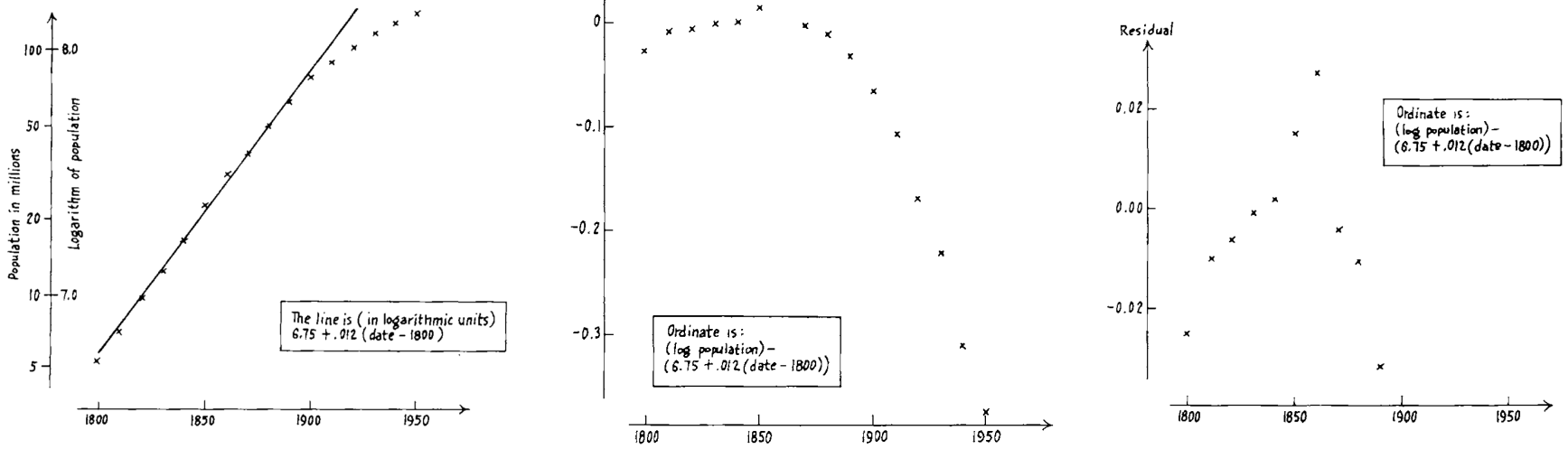
Another Tukey wisdom drop

Linearising bivariate relationships



The small fluctuations in later years are apparently errors or problems with data collection in the Censuses.

Linearising bivariate relationships



No explanation for these fluctuations. (Log didn't linearise, btw.)

Whatever the data, we can try to gain by straightening or by flattening. When we succeed in doing one or both, we almost always see more clearly what is going on.

Rules and advice

1.Graphics are friendly.
2.Arithmetic often exists to make graphs possible.
3.Graphs force us to note the unexpected; nothing could be more important.
4.Different graphs show us quite different aspects of the same data.
5.There is no more reason to expect one graph to "tell all" than to expect one number to do the same.
6."Plotting y against x" involves significant choices--how we express one or both variables can be crucial.
7.The first step in penetrating plotting is to straighten out the dependence or point scatter as much as reasonable.
8.Plotting y^2, $\sqrt{y}$, $log(y)$, $-1/y$ or the like instead of y is one plausible step to take in search of straightness.
9.Plotting x^2, $\sqrt{x}$, $log(x)$, $-1/x$ or the like instead of x is another.
10.Once the plot is straightened, we can usually gain much by flattening it, usually by plotting residuals.
11.When plotting scatters, we may need to be careful about how we express x and y in order to avoid concealment by crowding.

Making two-way analyses

- Stay tuned for median polish, coming in a later session on spatial data

The book is a digest of ⭐ tricks and treats ⭐ of massaging numbers and drafting displays.

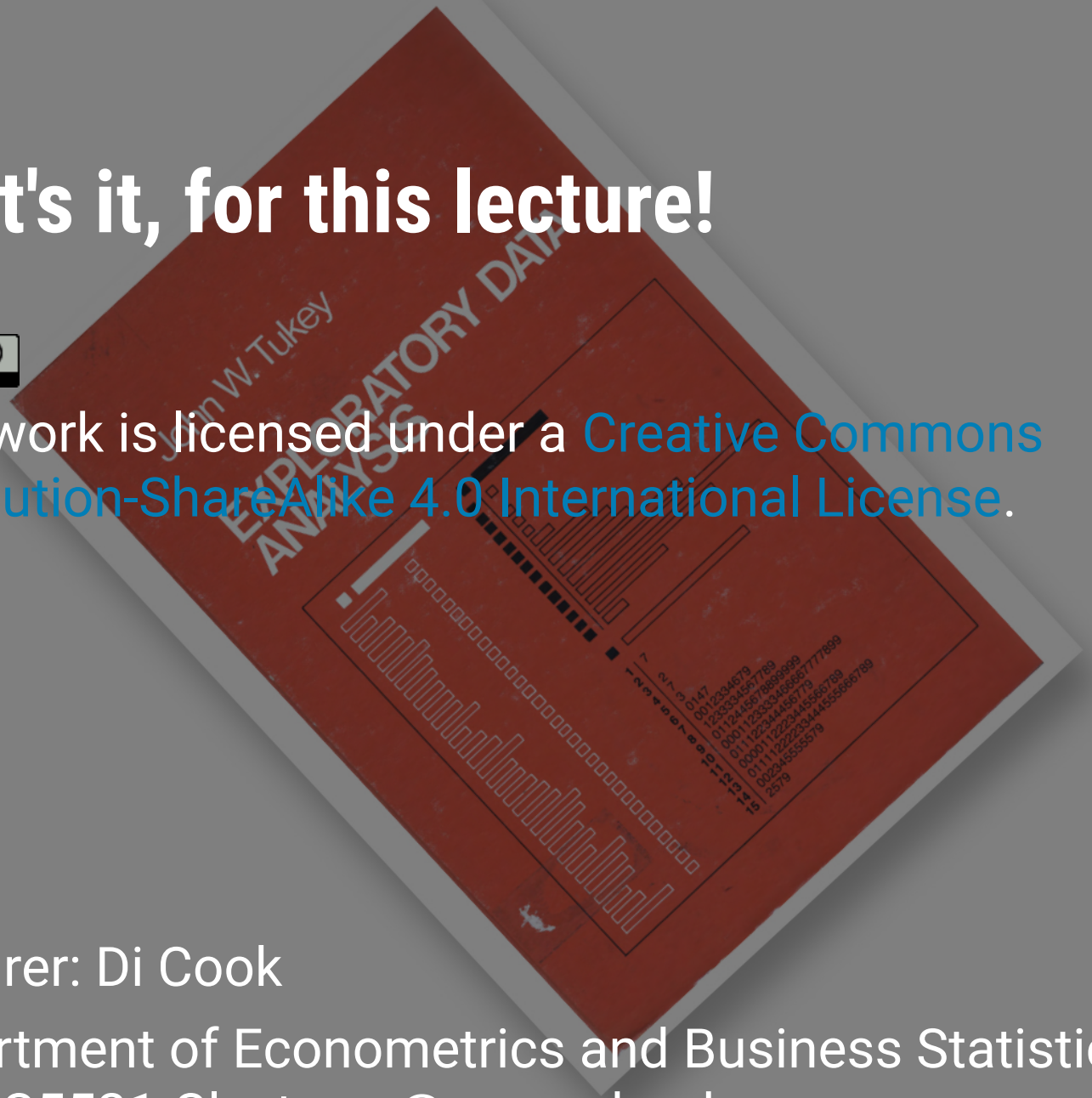
Many of the tools have made it into today's analyses in various ways. Many have not.

Notice the word developments too: froots, fences. Tukey brought you the word "software"!

The temperament of the book is an inspiration for the mind-set for this unit. There is such delight in working with numbers!

We love data!

That's it, for this lecture!



Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu