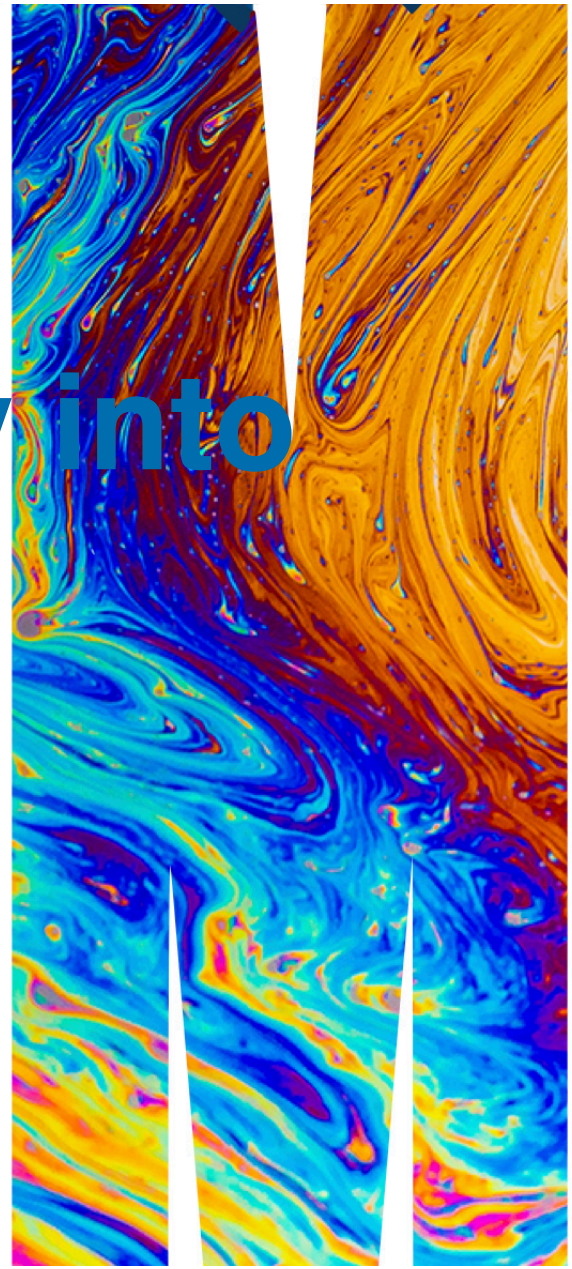
MONASH University

ETC5521: Diving Deeply into Data Exploration

Week 2: Learning from history

Professor Di Cook

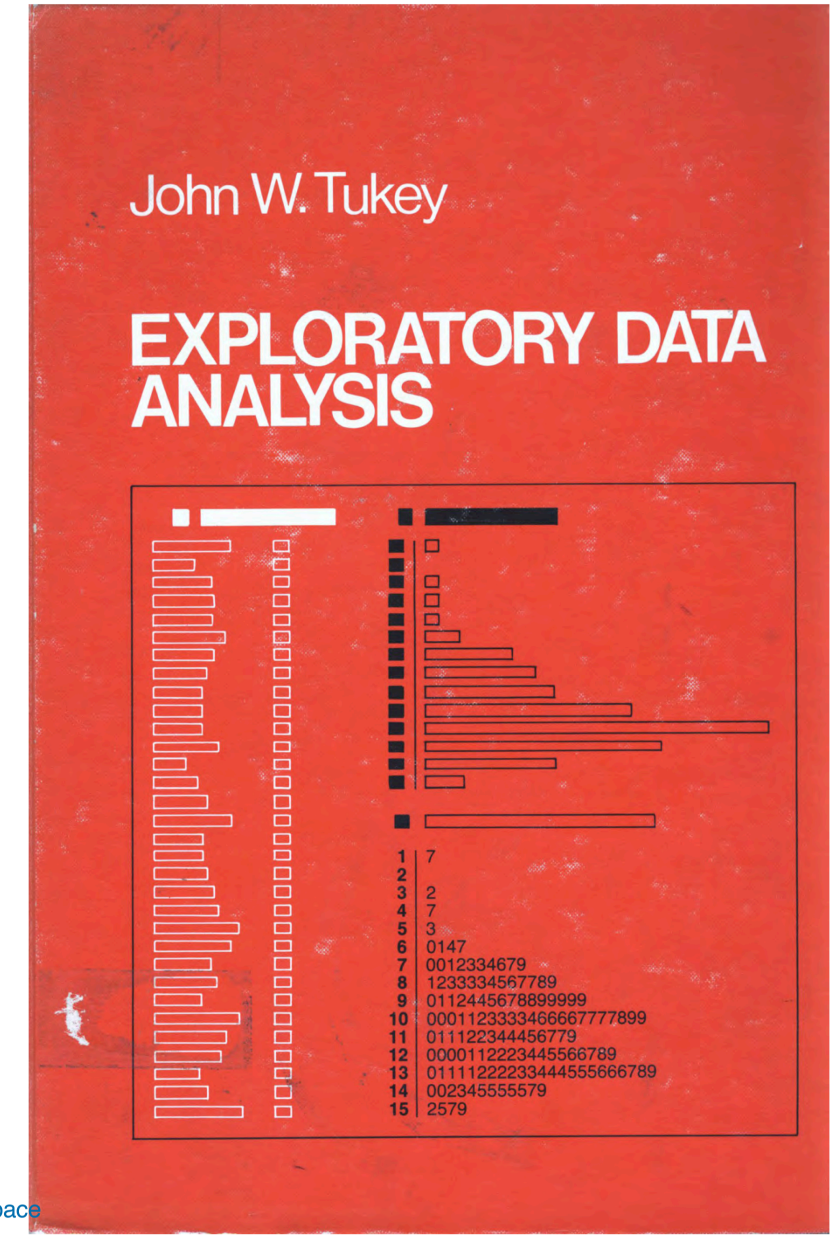
Department of Econometrics and Business Statistics



Birth of EDA

The field of exploratory data analysis came of age when this book appeared in 1977.

Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test.



John W. Tukey

- Born in 1915, in New Bedford, Massachusetts.
- Mum was a private tutor who home-schooled John. Dad was a Latin teacher.
- BA and MSc in Chemistry, and PhD in Mathematics
- Awarded the National Medal of Science in 1973, by President Nixon
- By some reports, his home-schooling was unorthodox and contributed to his thinking and working differently.

Image source: wikimedia.org

Taking a glimpse back in time

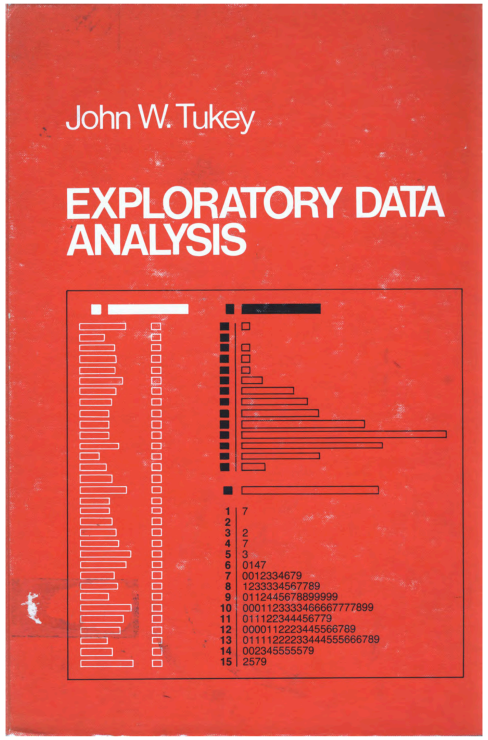
is possible with the American Statistical Association video lending library.

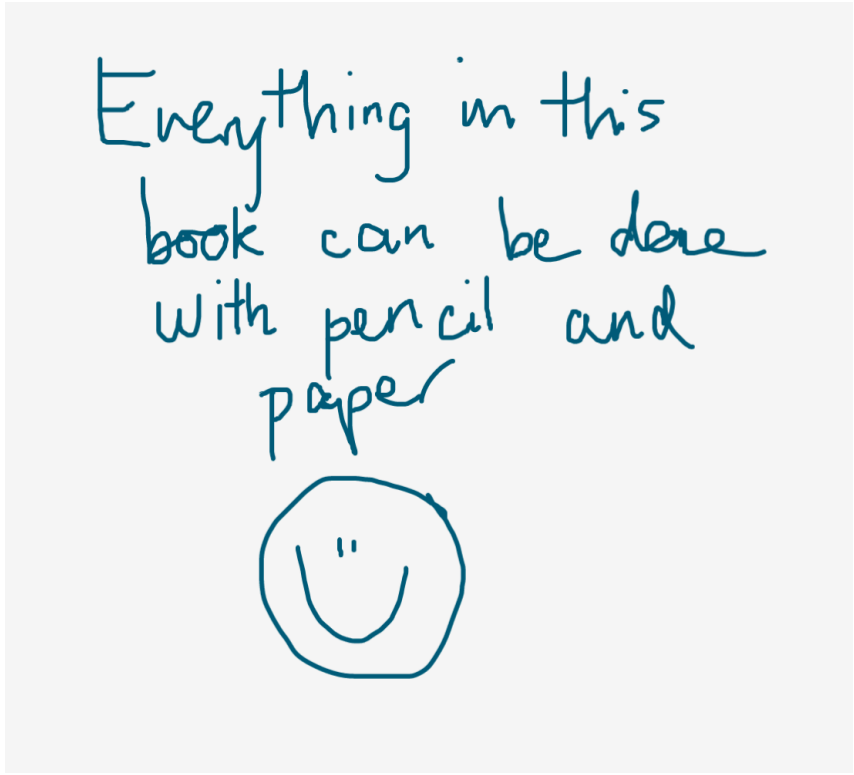
We're going to watch John Tukey talking about exploring high-dimensional data with an amazing new computer in 1973, four years before the EDA book.

Look out for these things:

Tukey's expertise is described as for trial and error learning and the computing equipment.

First 4.25 minutes


John W. Tukey
EXPLORATORY DATA ANALYSIS
1 | 7
2 |
3 | 2
4 | 7
5 | 3
6 | 0147
7 | 0012334679
8 | 1233334567789
9 | 01124456788999999
10 | 000112333346666777789
11 | 0111223444567779
12 | 000011222344556789
13 | 011112222334445556678
14 | 002345555579
15 | 2579


Everything in this book can be done with pencil and paper

Setting the frame of mind

Excerpt from the introduction

This book is based on an important principle.

It is important to understand what you CAN DO before you learn to measure how WELL you seem to have DONE it.

Learning first what you can do will help you to work more easily and effectively.

This book is about exploratory data analysis, about looking at data to see what it seems to say. It concentrates on simple arithmetic and easy-to-draw pictures. It regards whatever

Outline

1. Scratching down numbers
2. Schematic summary
3. Easy re-expression
4. Effective comparison
5. Plots of relationship
6. Straightening out plots (using three points)
7. Smoothing sequences
8. Parallel and wandering schematic plots
9. Delineations of batches of points
10. Using two-way analyses
11. Making two-way analyses
12. Advanced fits
13. Three way fits
14. Looking in two or more ways at batched of points
15. Counted fractions
16. Better smoothing
17. Counts in bin after bin
18. Product-ratio plots
19. Shapes of distributions
20. Mathematical distributions

Looking at numbers with Tukey

Scratching down numbers

Prices of Chevrolet in the local used car newspaper ads of 1968.

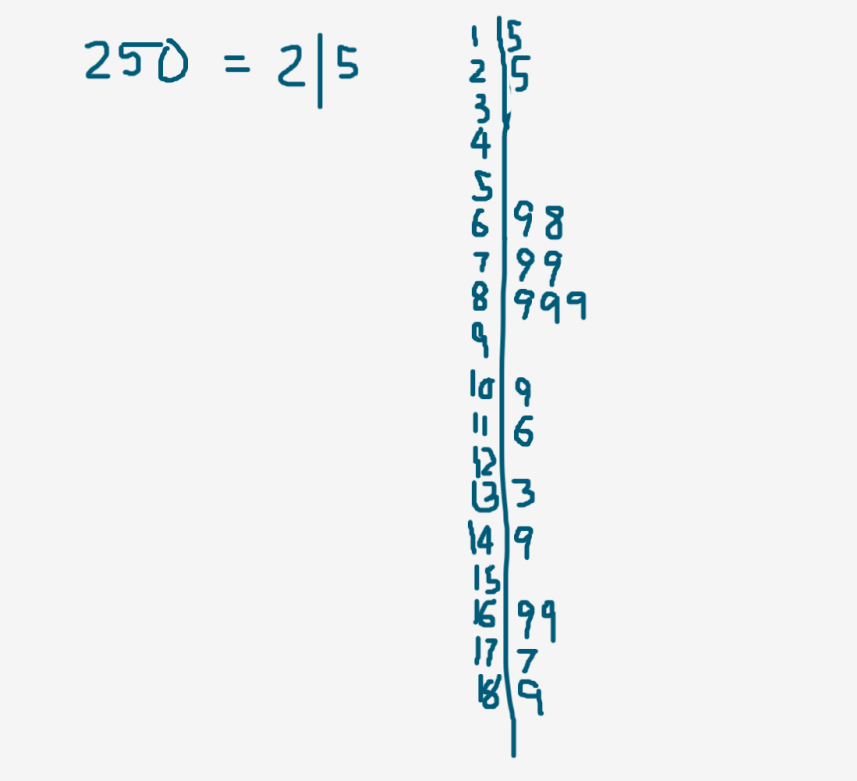
```
options(width=20)
chevrolets <- tibble(
  prices = c(250, 150, 795, 895, 695,
             1699, 1499, 1099, 1693, 1166,
             688, 1333, 895, 1775, 895,
             1895, 795))
#chevrolets$prices
```

Stem-and-leaf plot: still seen in introductory statistics texts

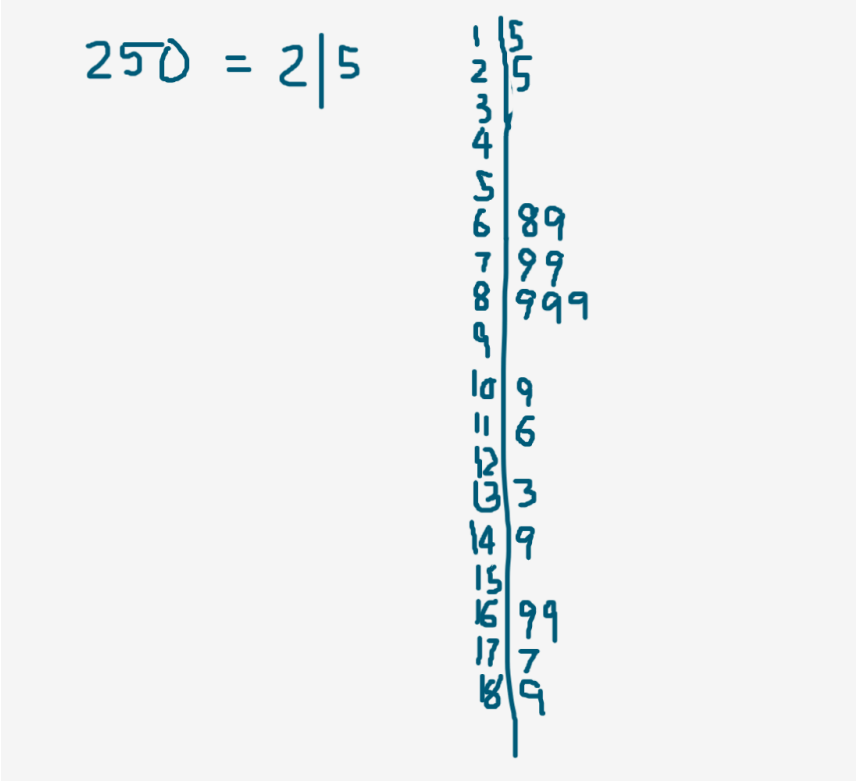
Export

```
250, 150, 795, 895, 695,
1699, 1499, 1099, 1693,
1166, 688, 1333, 895,
1775, 895, 1895, 795
```

First stem-and-leaf, first digit on stem, second digit on leaf

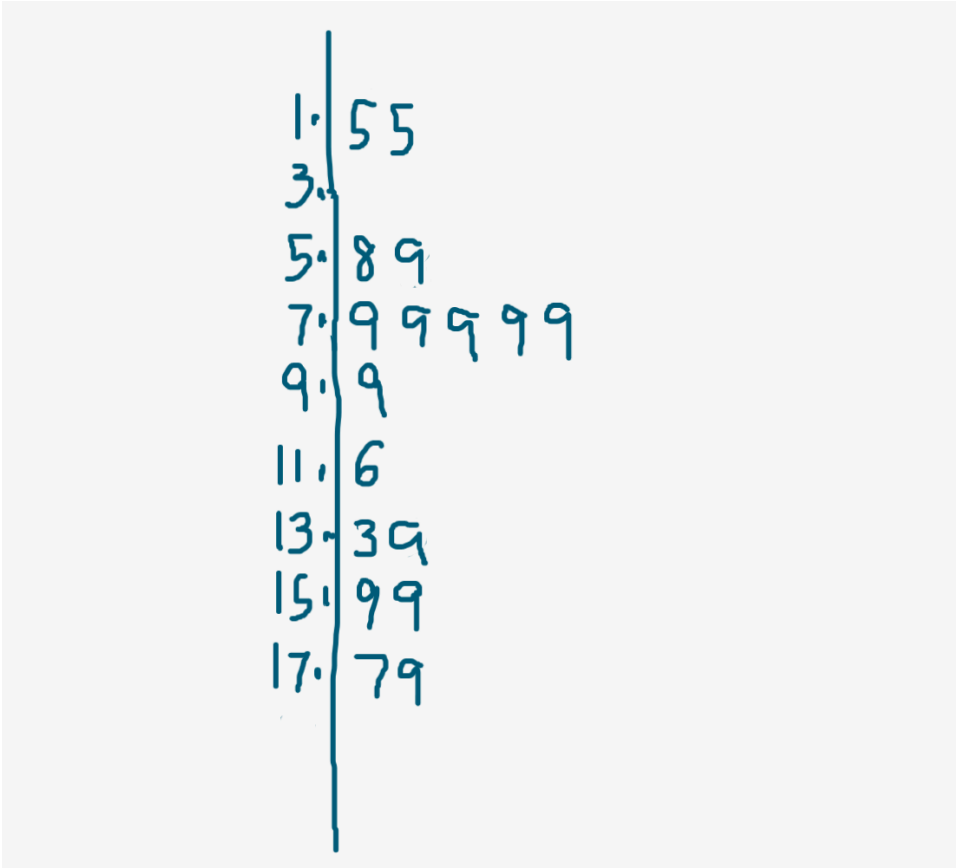


Order any leaves which need it, eg stem 6

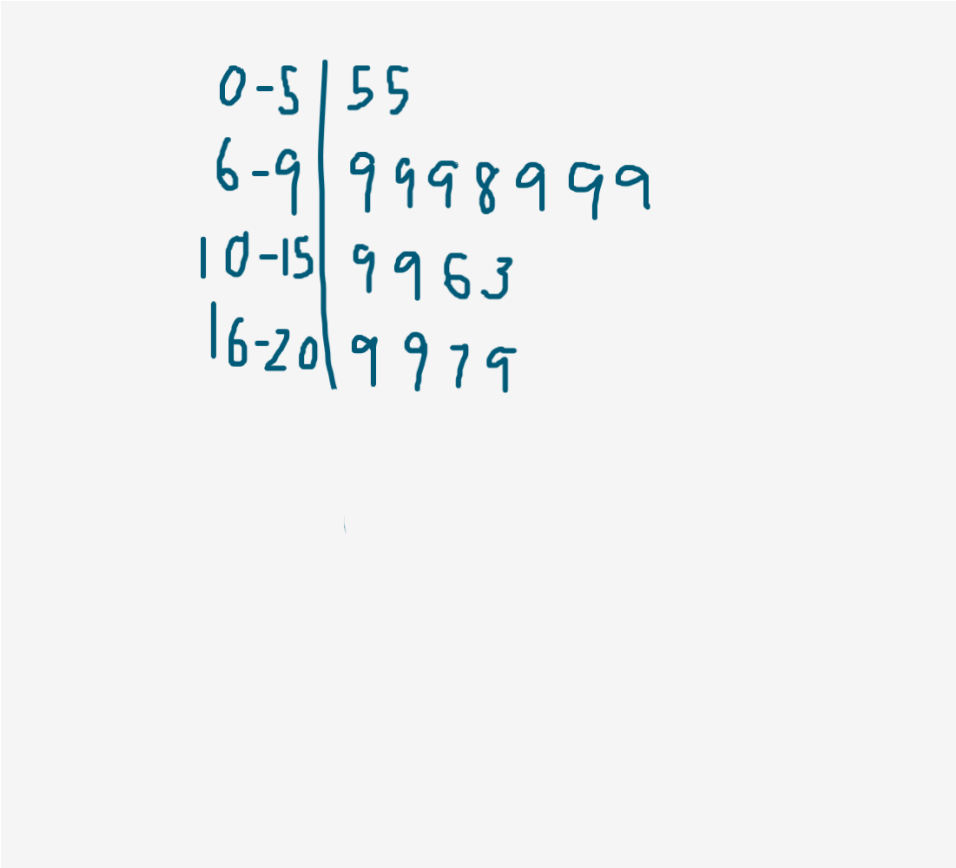


A benefit is that the numbers can be read off the plot, but the focus is still on the pattern. Also quantiles like the median, can be computed easily.

Shrink the stem



Shrink the stem more



And, in R ...

```
1 chevrolets$prices
```

```
 [1]  250  150  795
 [4]  895  695 1699
 [7] 1499 1099 1693
[10] 1166  688 1333
[13]  895 1775  895
[16] 1895  795
```

```
1 stem(chevrolets$prices)
```

```
The decimal point is 3 digit(s) to the right of the |

0 | 23
0 | 7788999
1 | 123
1 | 57789
```

Remember the tips data

```
 [1] 1.01 1.66 3.50 3.31 3.61 4.71 2.00 3.12 1.96 3.23 1.71 5.00 1.57 3.00 3.02
[16] 3.92 1.67 3.71 3.50 3.35 4.08 2.75 2.23 7.58 3.18 2.34 2.00 2.00 4.30 3.00
[31] 1.45 2.50 3.00 2.45 3.27 3.60 2.00 3.07 2.31 5.00 2.24 2.54 3.06 1.32 5.60
[46] 3.00 5.00 6.00 2.05 3.00
```

```
1 stem(tips$tip, scale=0.5, width=120)
```

```
The decimal point is at the |

 1 | 000001233334445555555555556666667777788889
 2 | 00000000000000000000000000000000000011222222233335555555555555556666677788899
 3 | 00000000000000000000000011111112222222333344445555555555555666778889
 4 | 000000000000111223333577 7
 5 | 00000000001122226799
 6 | 05577
 7 | 6
 8 |
 9 | 0
10 | 0
```

Refining the size

Five digits per stem

			(#)
3·	8	Tate	(1)
4*	0121243121300214202		(19)
4·	597886556569		(12)
5*	142010		(6)
5·	977899958797		(12)
6*	412441		(6)
6·	898598		(6)
7*	320341203		(9)
7·	86657		(5)
8*	303		(3)
8·	8	Hinds	(1)
9*	24	Bolivar, Yazoo	(2)

What is the number in parentheses? And why might this be useful?

Two digits per stem

A) FIVE-LINE VERSION

		(#)
1*	1	(1)
t	2333	(4)
f	445555	(6)
s	66677	(5)
·	88	(2)
2*	0000011	(7)
t	23	(2)
f	445	(3)
s	6	(1)
·	9	(1)
3*	1	(1)
t	3	(1)
f		
s		
·		(34√)

```
1 stem(tips$tip, scale=2)
```

```
  The decimal point is 1 digit(s) to the left
of the |

   10 | 0000107
   12 | 55526
   14 | 44578000000000678
   16 | 1346781356
   18 | 032678
   20 |
00000000000000000000000000000000011233598
   22 | 0033440114
   24 | 5700000000002456
   26 | 01412455
```

Why no number in parentheses?

```
1 median(tips$tip)
```

```
[1] 2.9
```

Summary

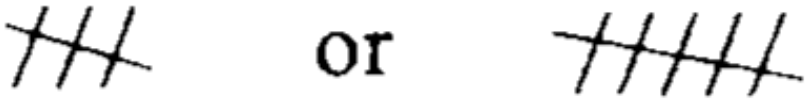
- Stem-and-leaf plots are similar information to the histogram.
- Generally it is possible to also read off the numbers, and to then easily calculate median or Q1 or Q3.
- It's great for small data sets, when you only have pencil and paper.
- Alternatives are a histogram, (jittered) dotplot, density plot, box plot, violin plot, letter value plot.

a different style of number scratching

for categorical variables

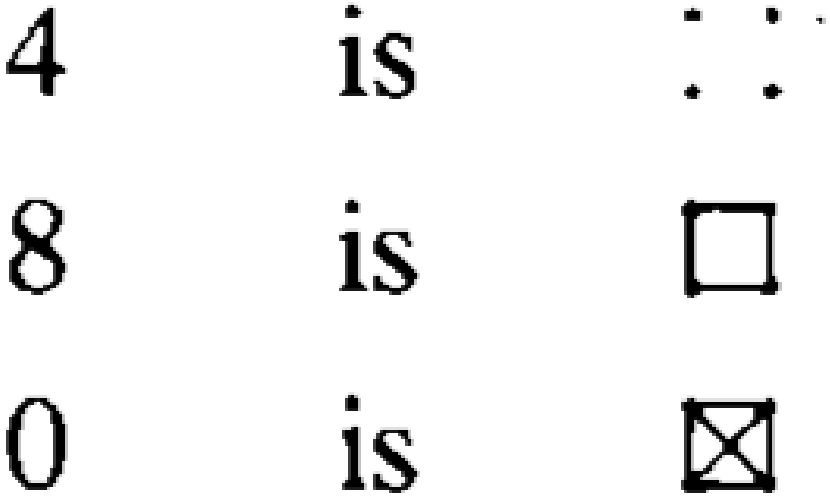
We know about

but its too easy to



make a mistake

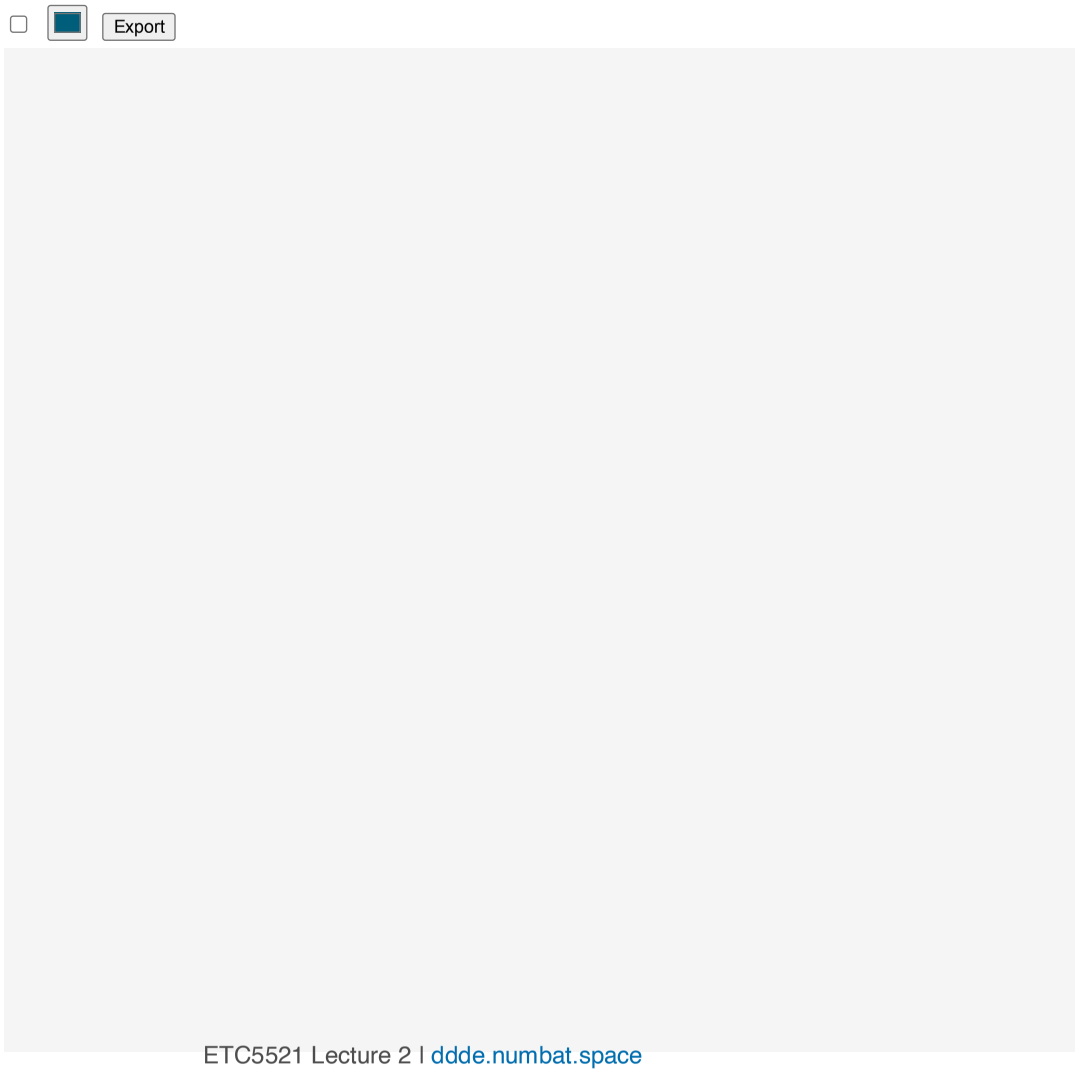
Is this easier?



or harder

Count this data using the squares approach.

```
 [1] "F" "M" "M" "M" "F" "M"
 [7] "M" "M" "M" "M" "M" "F"
[13] "M" "M" "F" "M" "F" "M"
[19] "F" "M" "M" "F" "F" "M"
[25] "M" "M" "M" "M" "M" "F"
[31] "M" "M" "F" "F" "M" "M"
[37] "M" "F" "M" "M" "M" "M"
[43] "M" "M" "M" "M" "M" "M"
[49] "M" "M" "M" "F" "F" "M"
[55] "M" "M" "M" "F" "M" "M"
[61] "M" "M" "M" "M" "M" "M"
[67] "F" "F" "M" "M" "M" "F"
```



What does it mean to “feel what the data are like?”

exhibit **10** of chapter 1: state heights

The heights of the highest points in each state

A) STEM-and-LEAF--unit 100 feet

			(#)
0*	43588	Del, Fla, La, Miss, RI	(5)
1	237886		(6)
2	484030		(6)
3	45526		(5)
4*	80149		(5)
5	34307		(5)
6	376		(3)
7	2	S. Dak	(1)
8*	8	Texas	(1)
9			
10			
11	2	Oregon	(1)
12*	768		(3)
13	81258		(5)
14	544	Calif, Colo, Wash	(3)
15			
16*			
17			
18			
19			
20*	3	Alaska	(1)

(50 ✓)

This is a stem and leaf of the height of the highest peak in each of the 50 US states.

The states roughly fall into three groups.

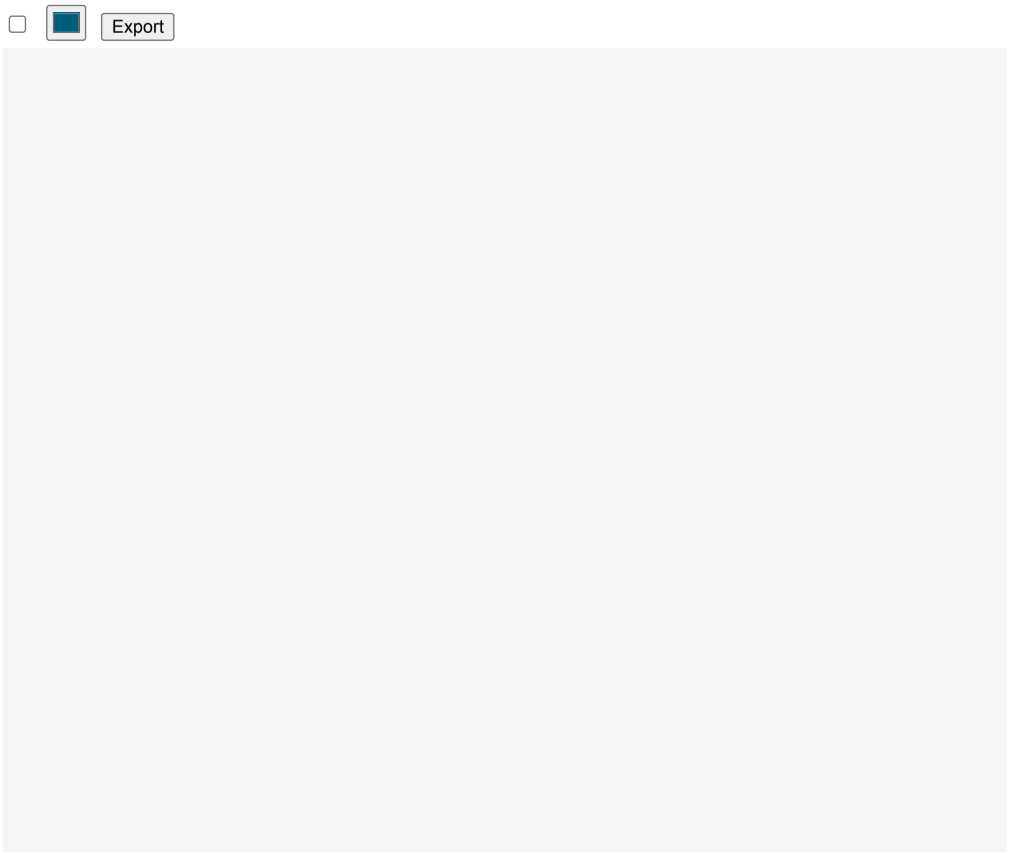
It's not really surprising, but we can imagine this grouping. Alaska is in a group of its own, with a much higher high peak. Then the Rocky Mountain states, California, Washington and Hawaii also have high peaks, and the rest of the states lump together.

Exploratory data analysis is detective work – in the purest sense – finding and revealing the clues.

More summaries of numerical values

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
```

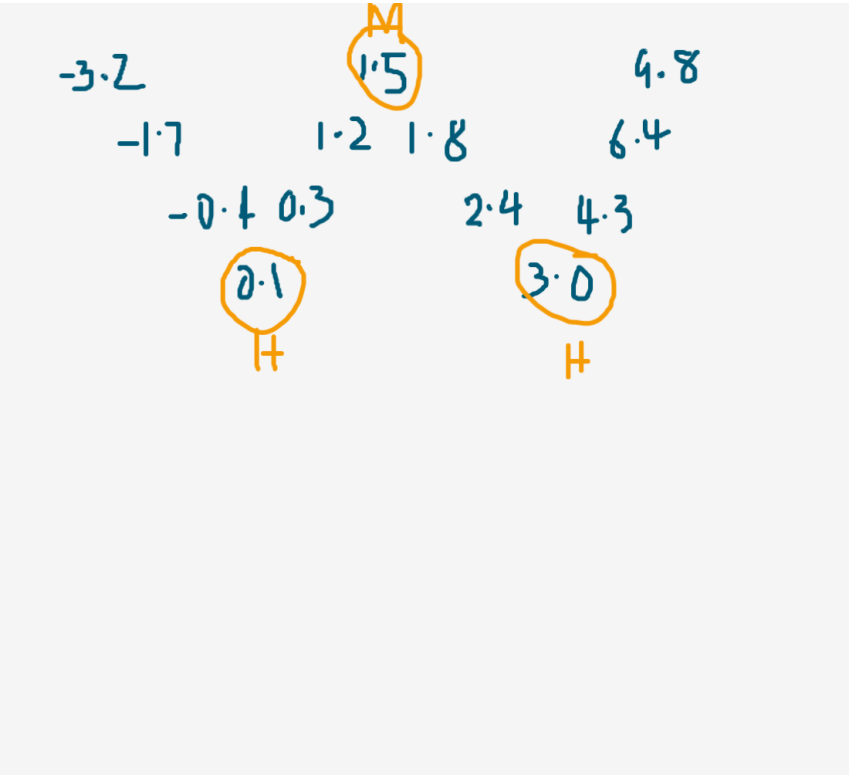


You know the median is the middle number. What's a hinge?

There are 13 data values here, provided already sorted. We are going to write them into a Tukey named down-up-down-up pattern, evenly.

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

Hinges and 5-number summary



Hinges illustrated

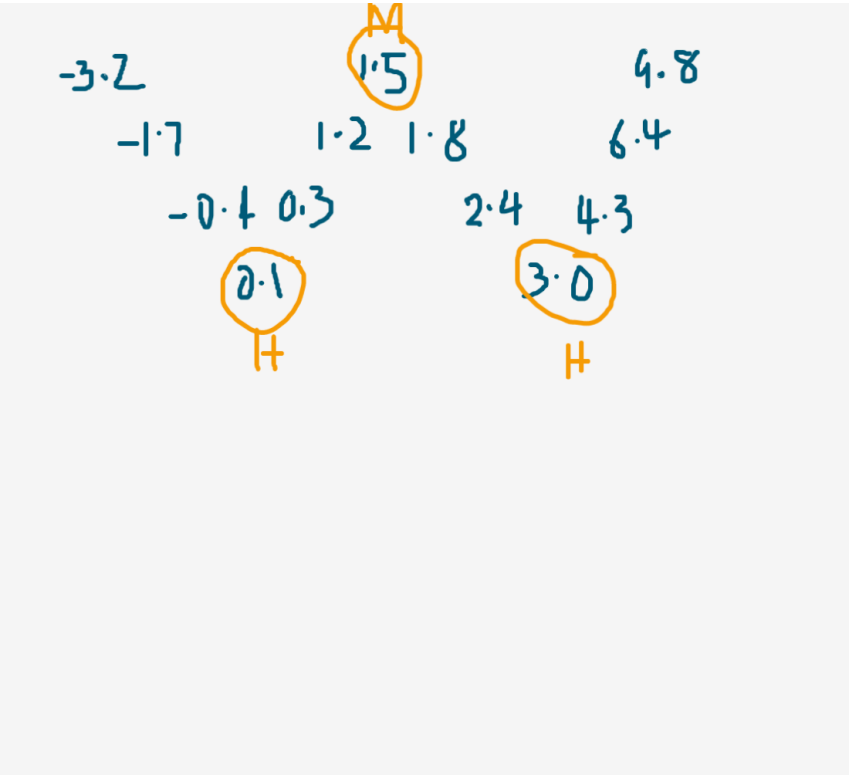
A) The 17 AUTO PRICES of EXHIBIT 1--in folded form

```
(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 almost always the same as Q1 and Q3

box-and-whisker display



Starting with a 5-number summary

#13

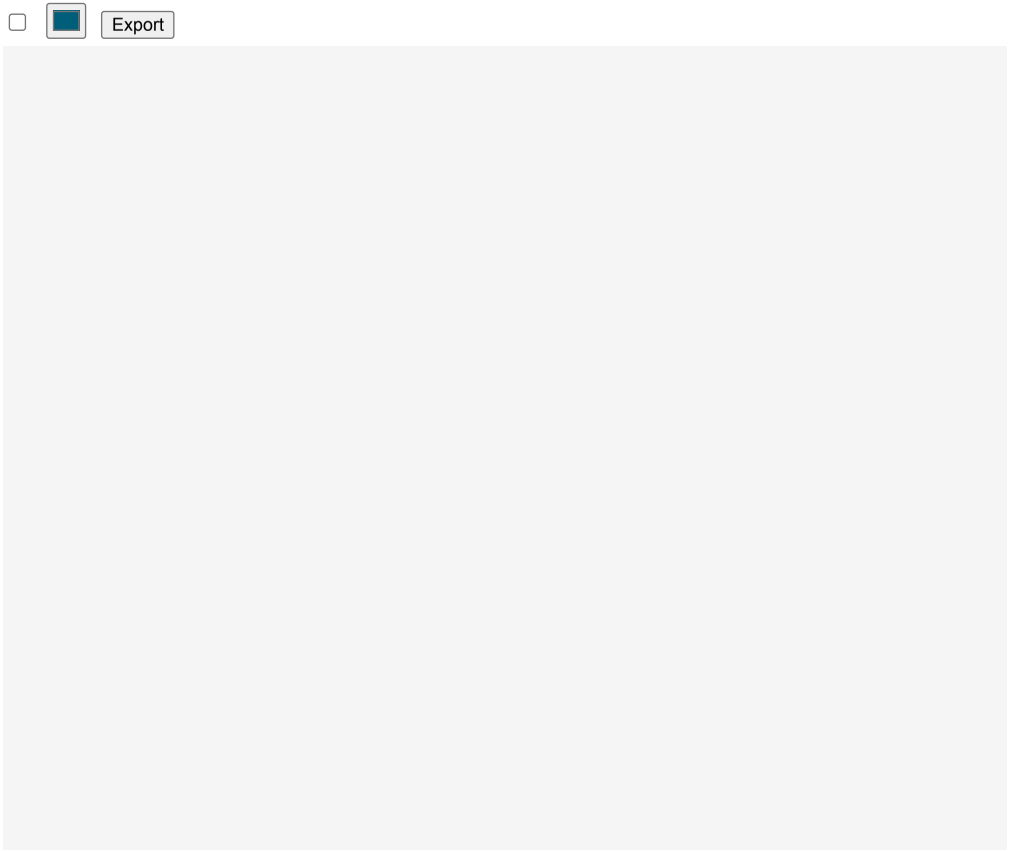
M7	1.5	
H4	0.1	3.0
1	−3.2	9.8

box-and-whisker display

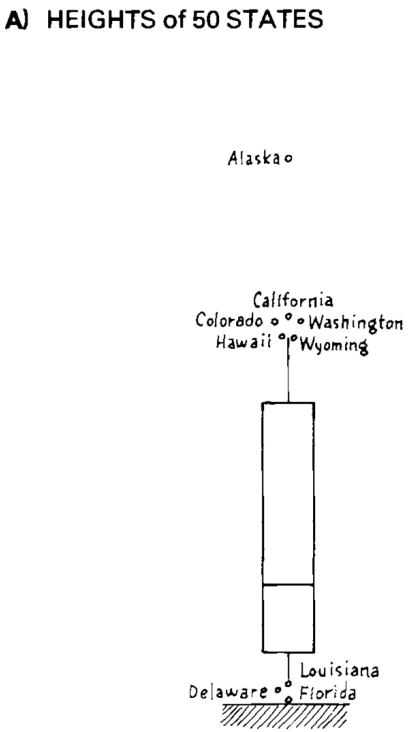
Starting with a 5-number summary

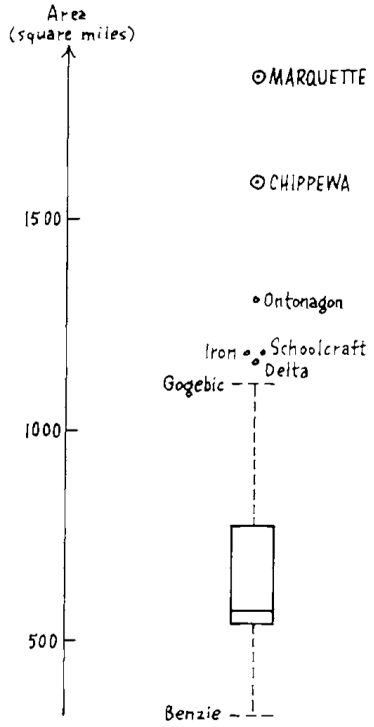
#13

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?

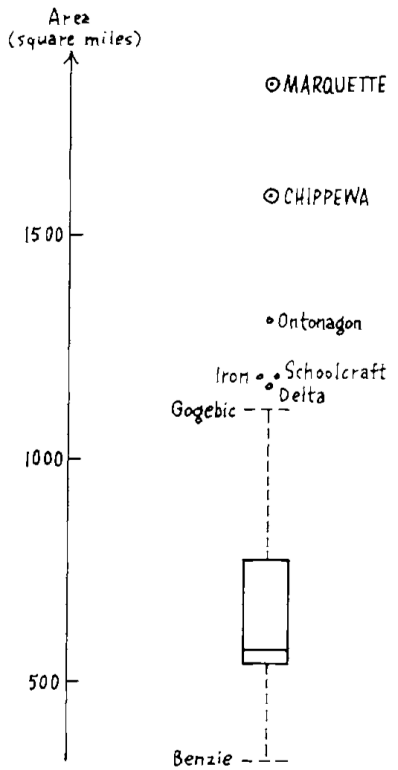
Isn’t this imposing a belief?

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: today's solution

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

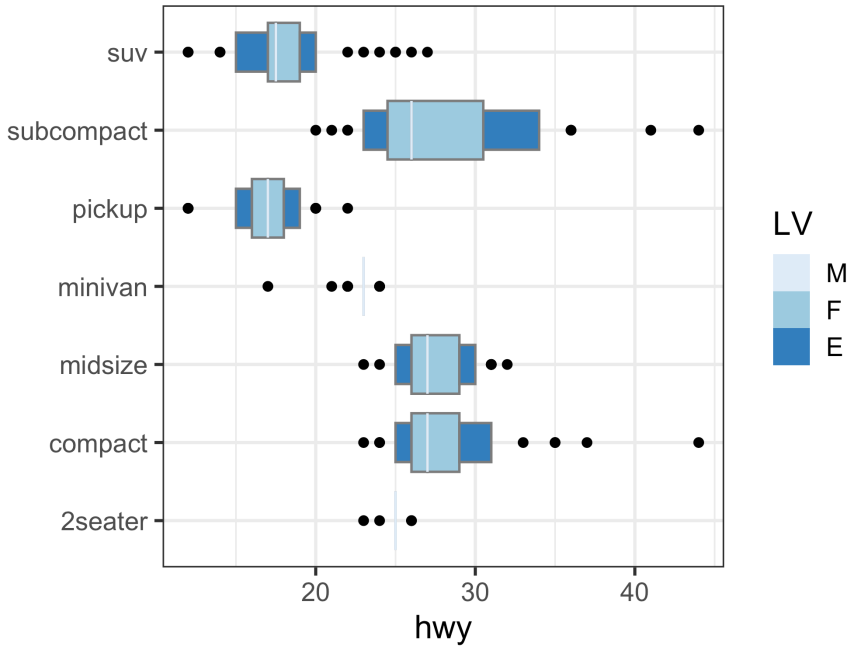
What does a 7-number summary look like?

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?

```
1 library(lvplot)
2 p <- ggplot(mpg,
3             aes(class, hwy))
4 p + geom_lv(aes(fill=..LV..)) +
5   scale_fill_brewer() +
6   coord_flip() +
7   xlab("")
```



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 solved 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

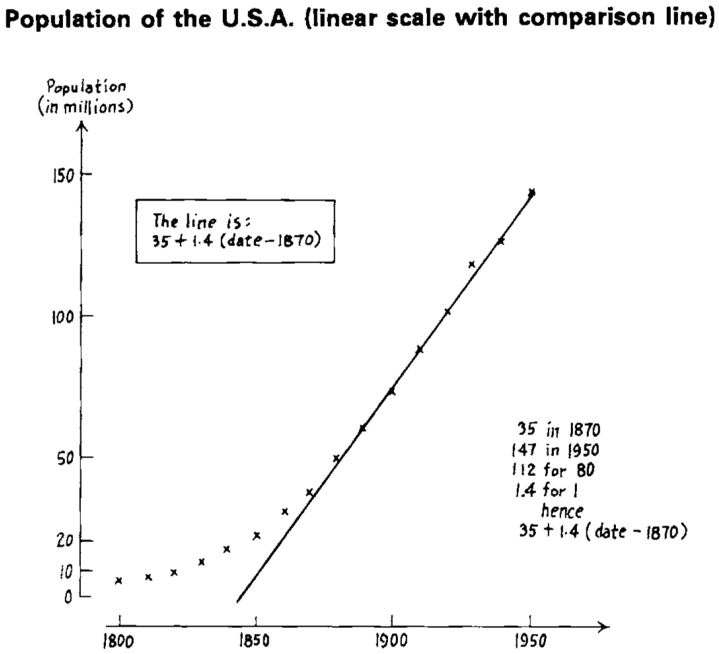
-2, -1, -1/2, 0 (log), 1/3, 1/2, 1, 2, 3, 4

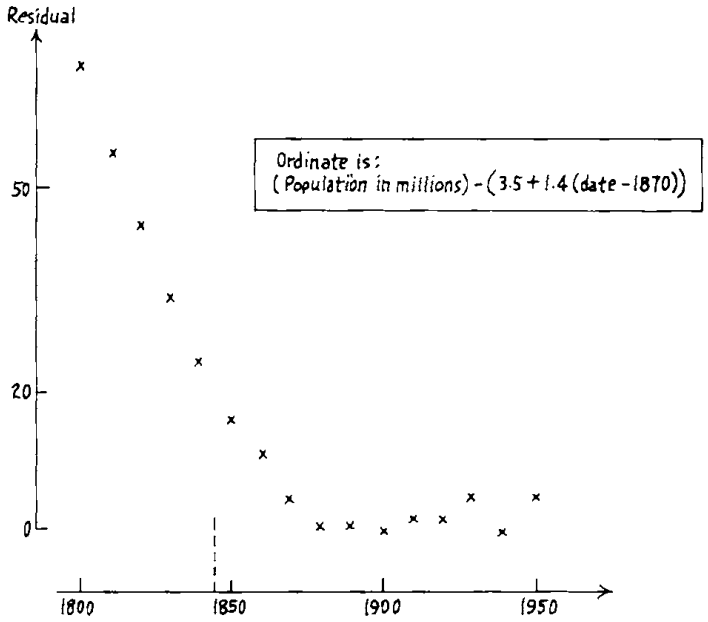
fix LEFT-skewed values ➡️

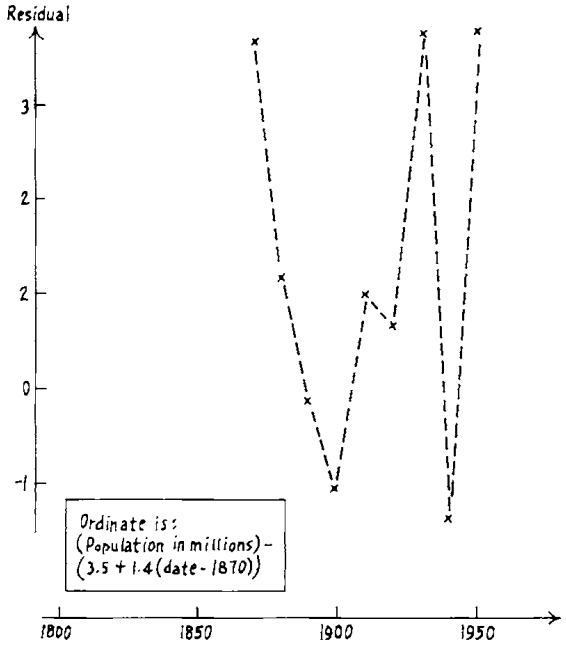
We now regard re-expression as a tool, something to let us do a better job of grasping. The grasping is done with the eye and the better job is through a more symmetric appearance.

Another Tukey wisdom drop

Linearising bivariate relationships



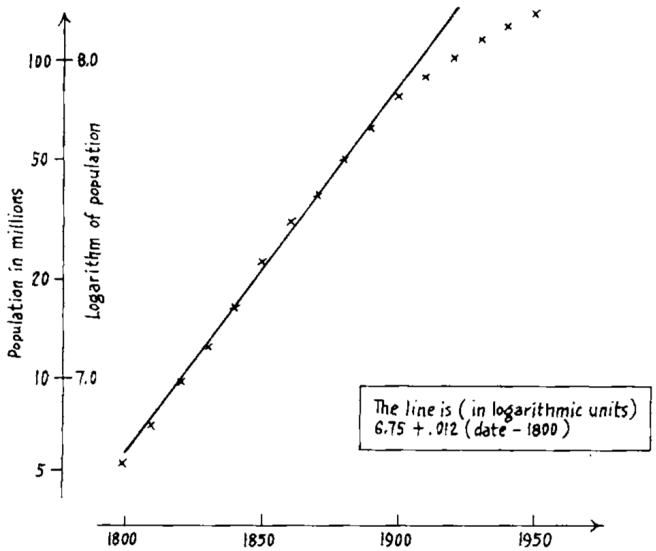


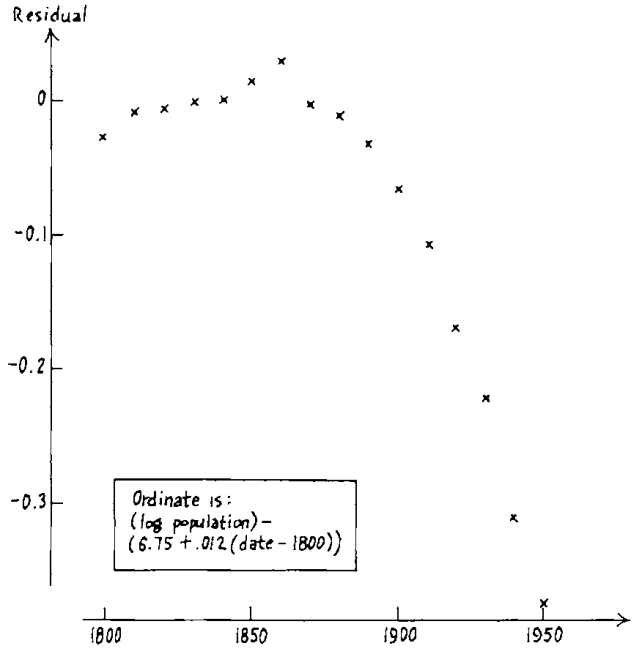


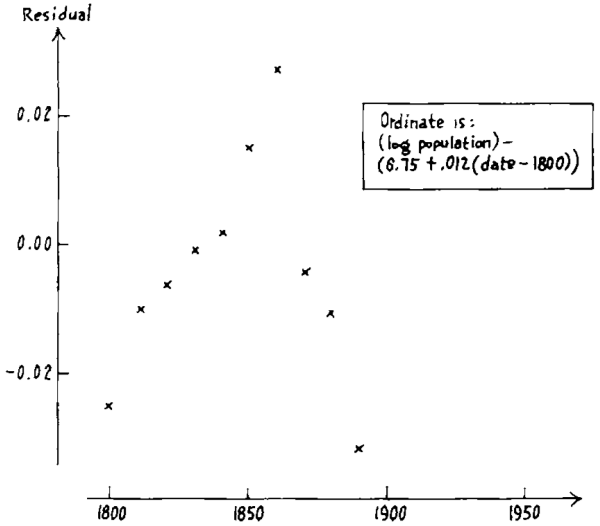
Surprising observation: The small fluctuations in later years.

What might be possible reasons?

Linearising bivariate relationships







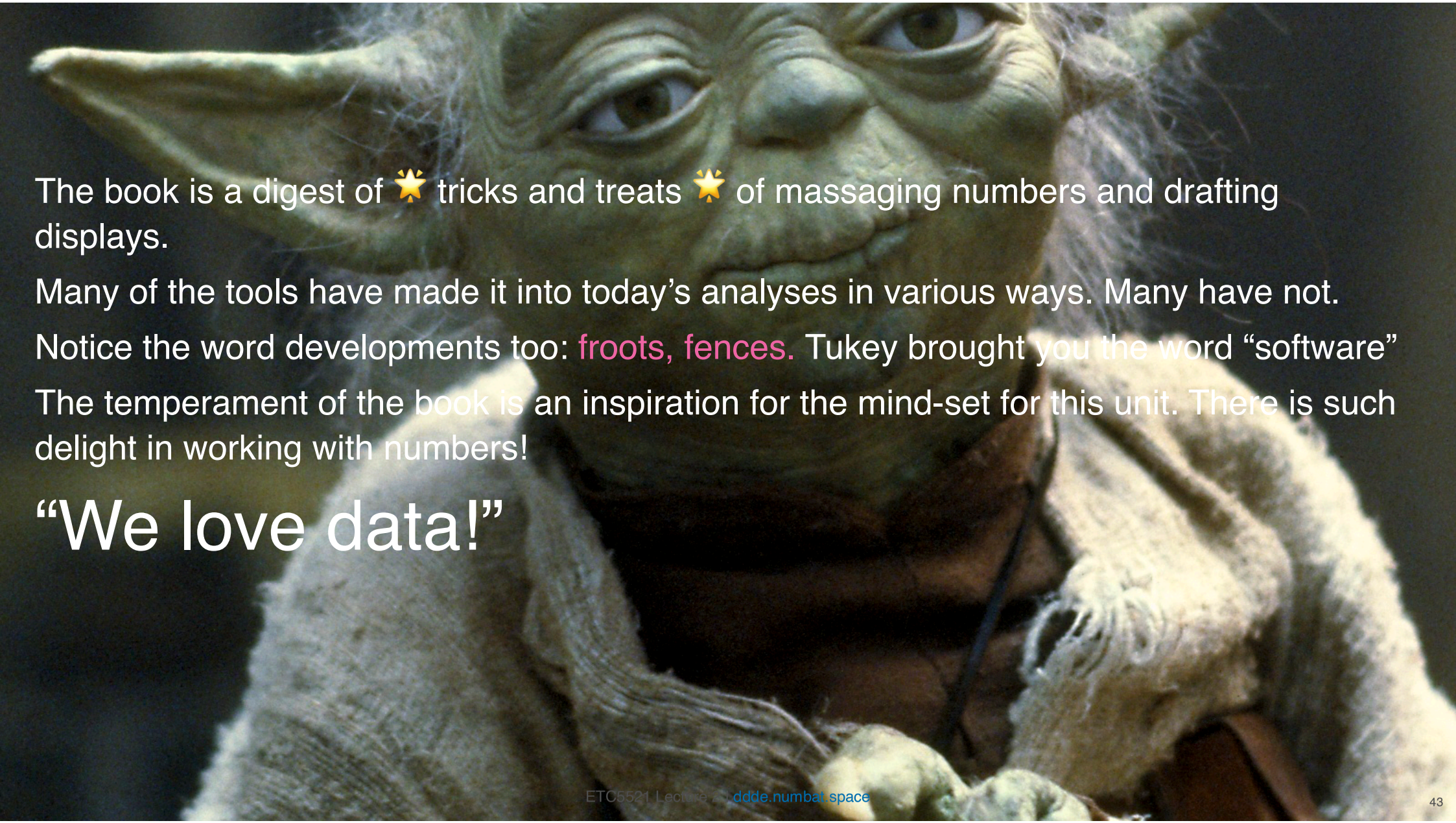
See some fluctuations in the early years, too. Note that the log transformation couldn't linearise.

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 notice 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.



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!"

Take-aways

- Tukey's approach was a reaction to many years of formalising data analysis using statistical hypothesis testing.
- Methodology development in statistical testing was a reaction to the ad-hoc nature of data analysis.
- Complex machine learning models like neural networks are in reaction to the inability of statistical models to capture highly non-linear relationships, and depend heavily on the data provided.
- Exploring data today is in reaction to the need to explain complex models, to support organisations against legal challenges to decisions made from the model
- It is much easier to accomplish computers.
- "Exploratory data analysis" as commonly used today term is unfortunately synonymous with "descriptive statistics", but it is truly much more. Understanding its history from Tukey's advocation helps you see it is the tooling to discover what you don't know.

Resources

- wikipedia
- John W. Tukey (1977) Exploratory data analysis
- Data coding using `tidyverse` suite of R packages
- Sketching canvases made using `fabricerin`