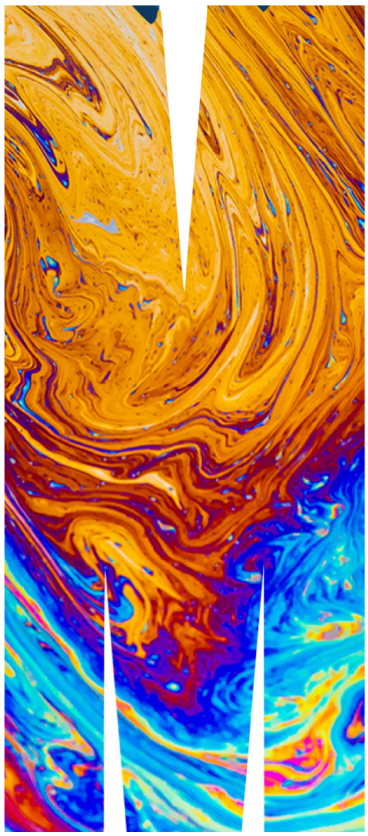
MONASH University

ETC5521: Exploratory Data Analysis

Learning from history

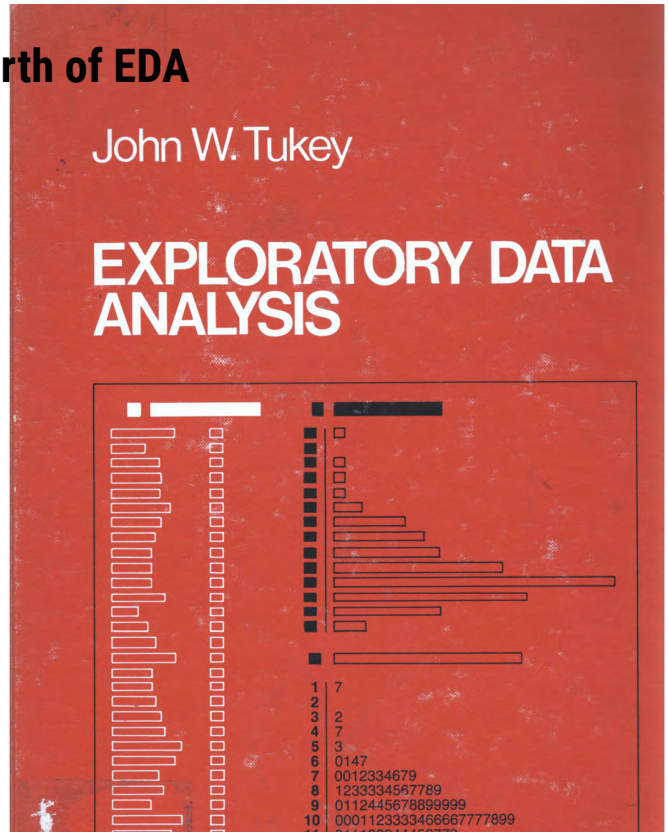


Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 2 - Session 1

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.

Taking a glimpse back in time

is possible with the American Statistical Association video lending library.

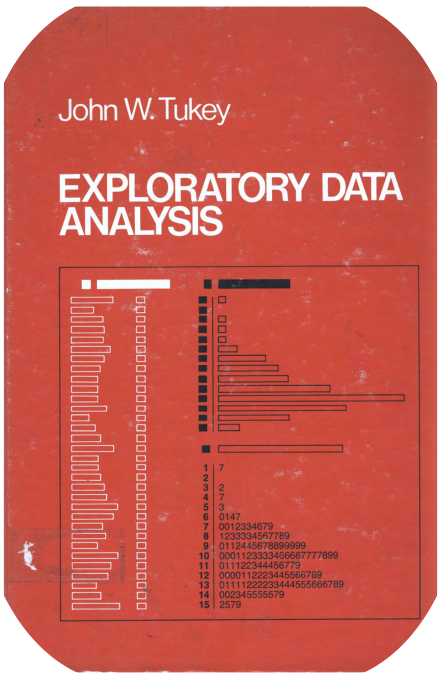
We're going to watch John Tukey talking about exploring high-dimensional data with an amazing new computer in 1973, four years before the EDA book.

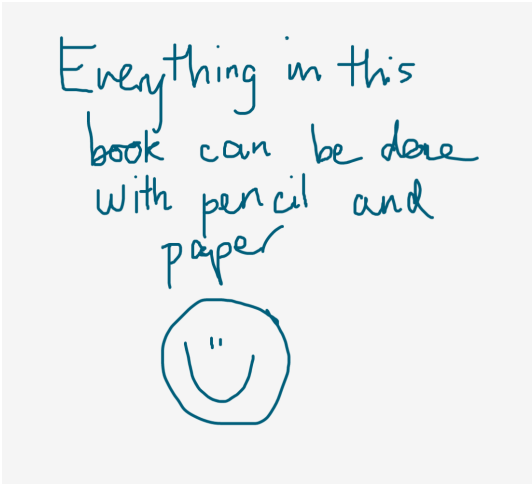
💡 Look out for these things:

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

First 4.25 minutes of https://www.youtube.com/embed/B7XoW2qiFUA)

prim9


John W. Tukey
EXPLORATORY DATA ANALYSIS
1 | 7
2 |
3 | 2
4 | 7
5 | 3
6 | 0147
7 | 0012334679
8 | 123333457789
9 | 01124456788999999
10 | 000112333346666777789
11 | 011122344456779
12 | 00001122234455667789
13 | 0111122223344455566789
14 | 002345555579
15 | 2579


Everything in this book can be done with pencil and paper

Setting the frame of mind

To learn about data analysis, it is right that each of us try many things that do not work--that we tackle more problems than we make expert analyses of. We often learn less from an expertly done analysis than from one where, by not trying something, we missed--at least until we were told about it--an opportunity to learn more. Each teacher needs to recognize this in grading and commenting on problems.

Precision

The teacher who heeds these words and admits that there need be *no one correct approach* may, I regret to contemplate, still want whatever is done to be digit perfect. (Under such a requirement, the write should still be able to pass the course, but it is not clear whether she would get an "A".) One does, from time to time, have to produce digit-perfect, carefully checked results, but forgiving techniques that are not too distributed by unusual data are also, usually, *little disturbed by SMALL arithmetic errors*. The techniques we discuss here have been chosen to be forgiving. It is hoped, then, that small arithmetic errors will take little off the problem's grades, leaving severe penalties for larger errors, either of arithmetic or concept.

Outline

Scratching down numbers

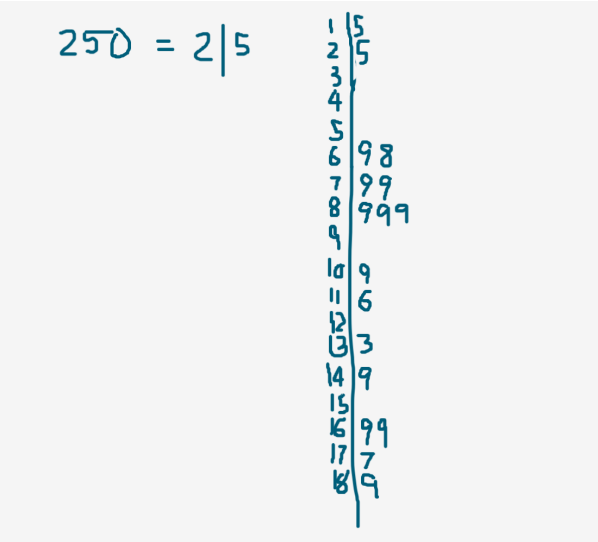
Prices of Chevrolet in the local used car newspaper ads of 1968.

Stem-and-leaf plot: still seen introductory statistics books

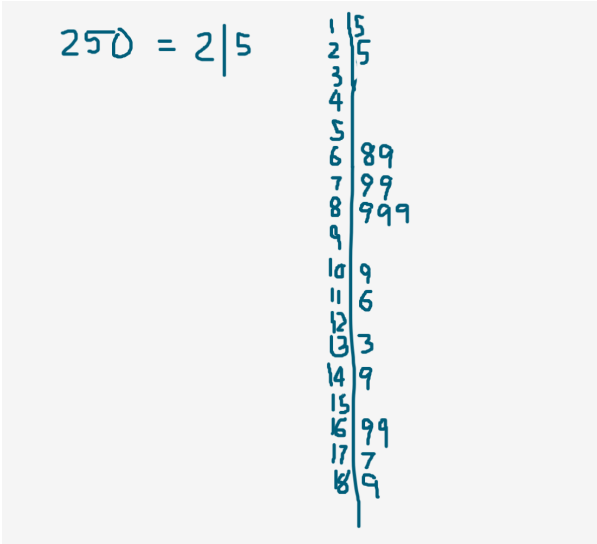
Export

```
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

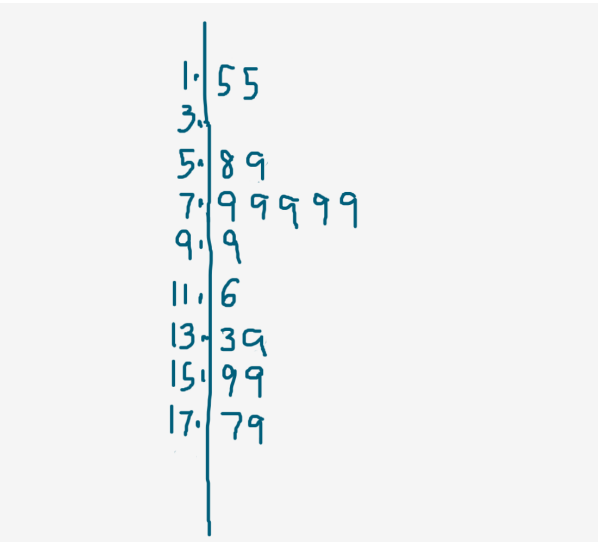


Order any leaves which need it, eg stem 6

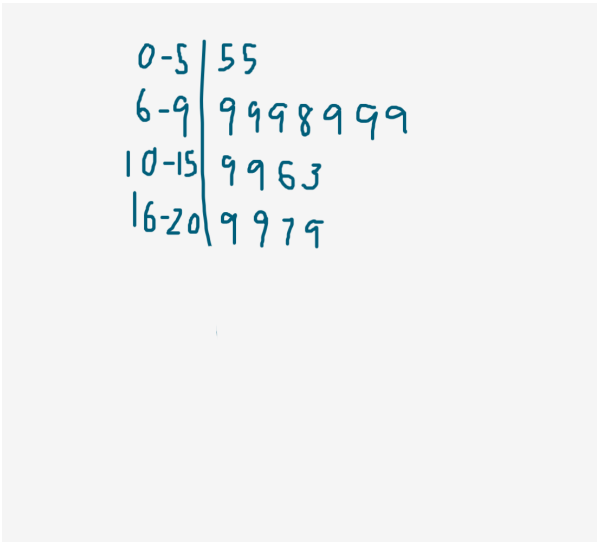


A benefit is that the numbers can be read off the plot, but the focus is still on the pattern. Also quantiles like the median, can be computed easily.

Shrink the stem



Shrink the stem more



And, in R ...

```
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

```
tips <- read_csv("http://ggobi.org/book/data/tips.csv")
stem(tips$tip, scale=0.5, width=120)
```

```
##
##   The decimal point is at the |
##
##    1 | 000001233334445555555555556666667777788889
##    2 | 00000000000000000000000000000000000000001122222223333555555555555556
##    3 | 000000000000000000000001111111222222233334444555555555555666778889
##    4 | 00000000000011122333357777
##    5 | 00000000011222267799
##    6 | 05577
##    7 | 6
##    8 |
##    9 | 0
##   10 | 0
```

Refining the size

			(#)
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)

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		


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

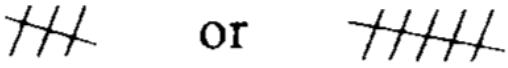
```
##
##   The decimal point is 1 digit(s) to the left of the |
##
##    10 | 0000107
##    12 | 55526
##    14 | 44578000000000678
##    16 | 1346781356
##    18 | 032678
##    20 | 00000000000000000000000000000011233598
##    22 | 0033440114
##    24 | 5700000000002456
##    26 | 01412455
##    28 | 382
##    30 | 00000000000000000000002678912456 88
##    32 | 133557159
```

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. However, its really designed for small data sets and for pencil and paper.

a different style of number scratching

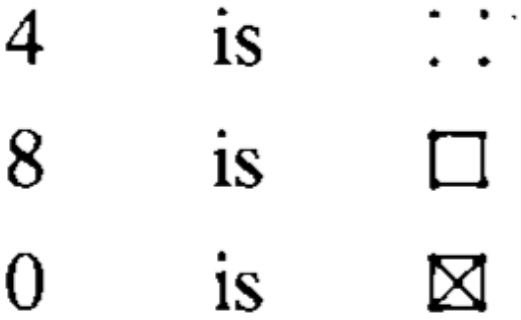
We know about

but its too easy to



make a mistake

Try this instead



Count this data using the squares approach.

```
##  [1] "Sun" "Sun" "Sun" "Sun"
##  [5] "Sun" "Sun" "Sun" "Sun"
##  [9] "Sun" "Sun" "Sun" "Sun"
## [13] "Sun" "Sun" "Sun" "Sun"
## [17] "Sun" "Sun" "Sun" "Sat"
## [21] "Sat" "Sat" "Sat" "Sat"
## [25] "Sat" "Sat" "Sat" "Sat"
## [29] "Sat" "Sat" "Sat" "Sat"
## [33] "Sat" "Sat" "Sat" "Sat"
## [37] "Sat" "Sat" "Sat" "Sat"
## [41] "Sat" "Sun" "Sun" "Sun"
## [45] "Sun" "Sun" "Sun" "Sun"
## [49] "Sun" "Sun" "Sun" "Sun"
## [53] "Sun" "Sun" "Sun" "Sun"
## [57] "Sat" "Sat" "Sat" "Sat"
## [61] "Sat" "Sat" "Sat" "Sat"
## [65] "Sat" "Sat" "Sat" "Sat"
```

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.

Resources

wikipedia

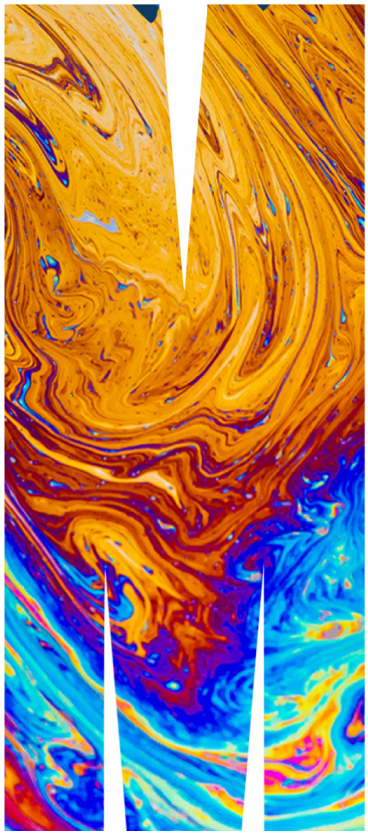
John W. Tukey (1977) Exploratory data analysis

Data coding using `tidyverse` suite of R packages

Sketching canvases made using `fabricerin`

Slides constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

MONASH University



Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 2 - Session 1