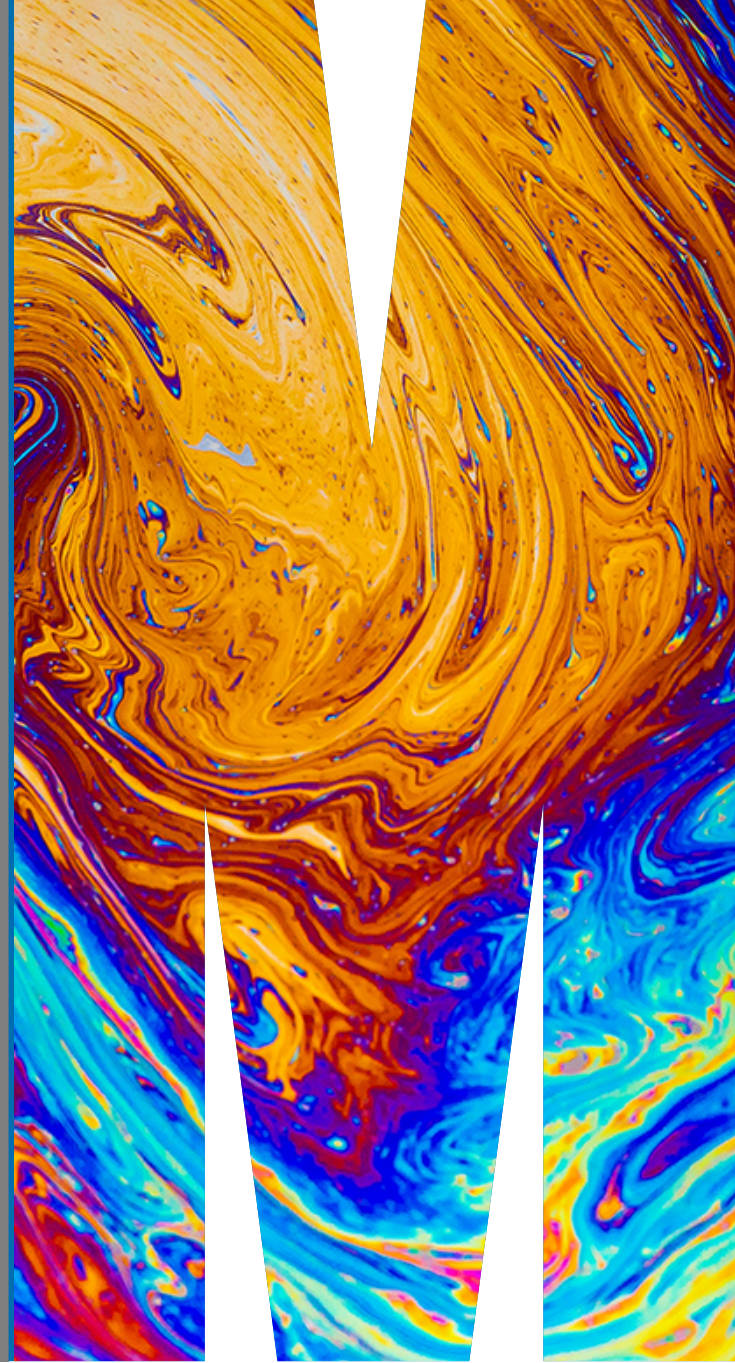
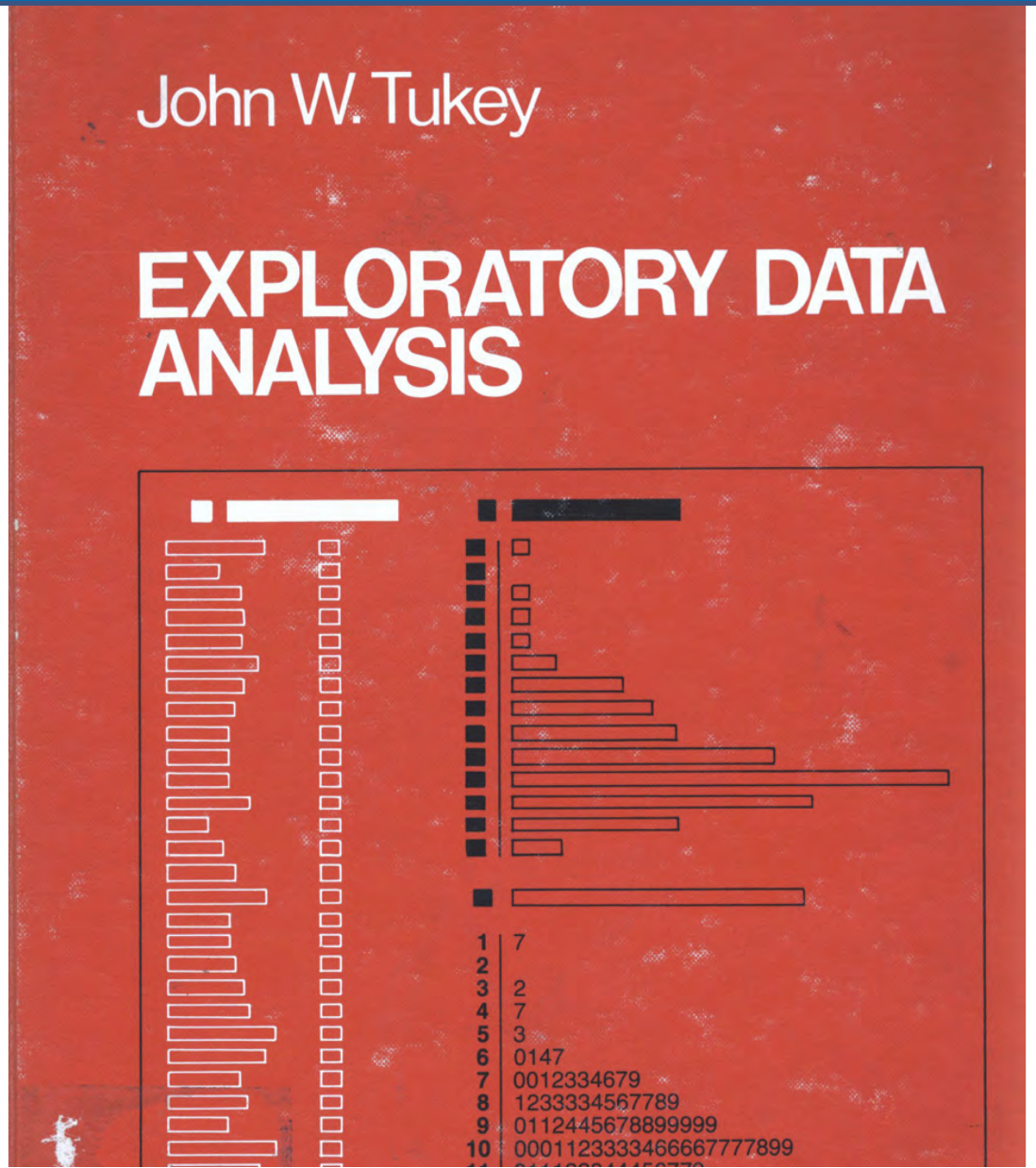
ETC5521: Exploratory Data Analysis

Learning from history

Lecturer: *Di Cook*
Department of Econometrics and Business Statistics
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.

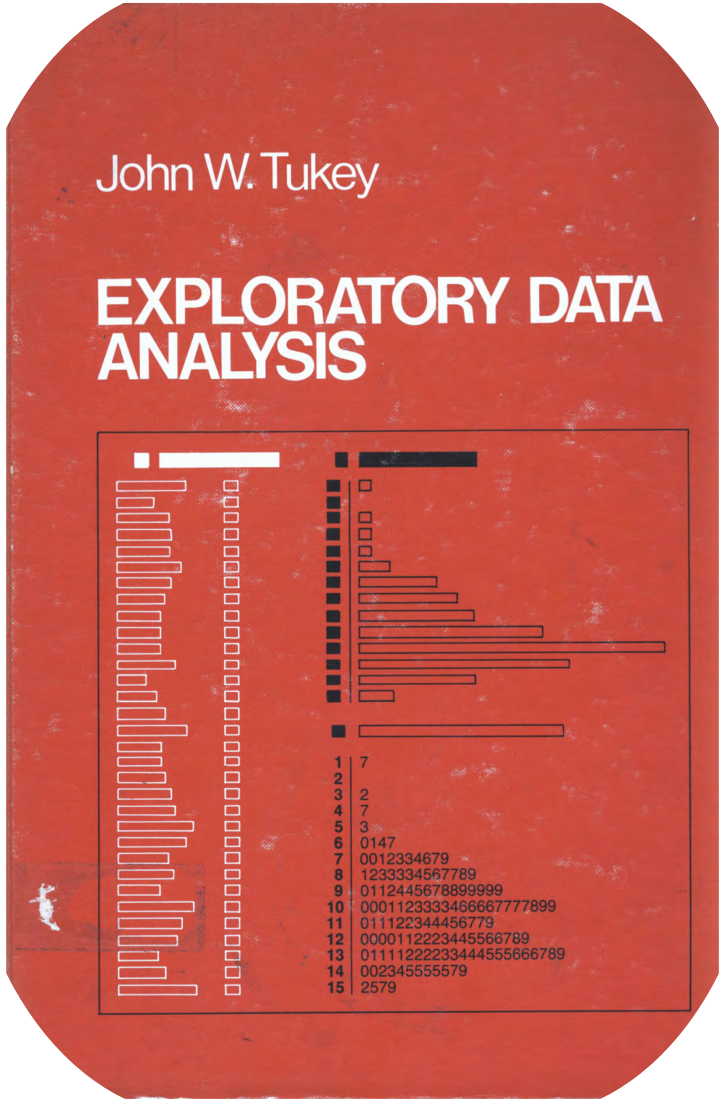
We're going to watch John Tukey talking about exploring high-dimensional data with an amazing new computer in 1973, four years before the EDA book.

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



Everything in this book can be done with pencil and paper

Setting the frame of mind

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 appearances we have recognized as partial descriptions, and tries to look beneather them for new insights. Its concern is with appearance, not with confirmation.

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

Here we go  

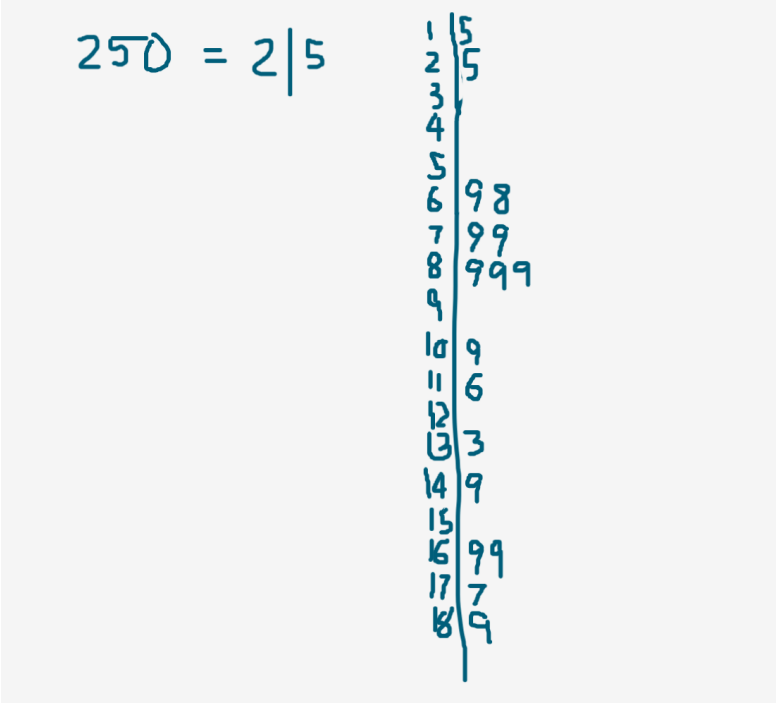
Scratching down numbers

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

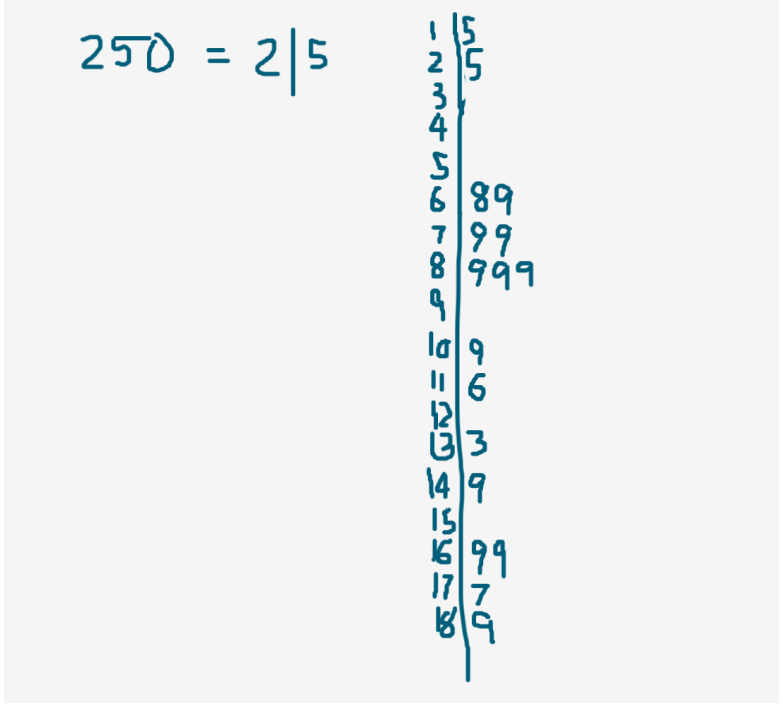
Stem-and-leaf plot: still seen introductory statistics books


```
 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

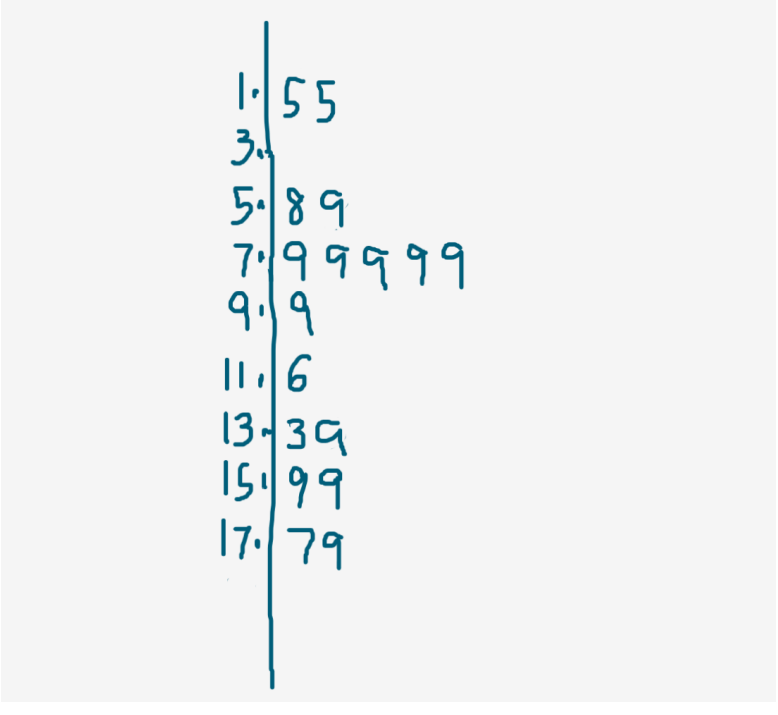


Order any leaves which need it, eg stem 6

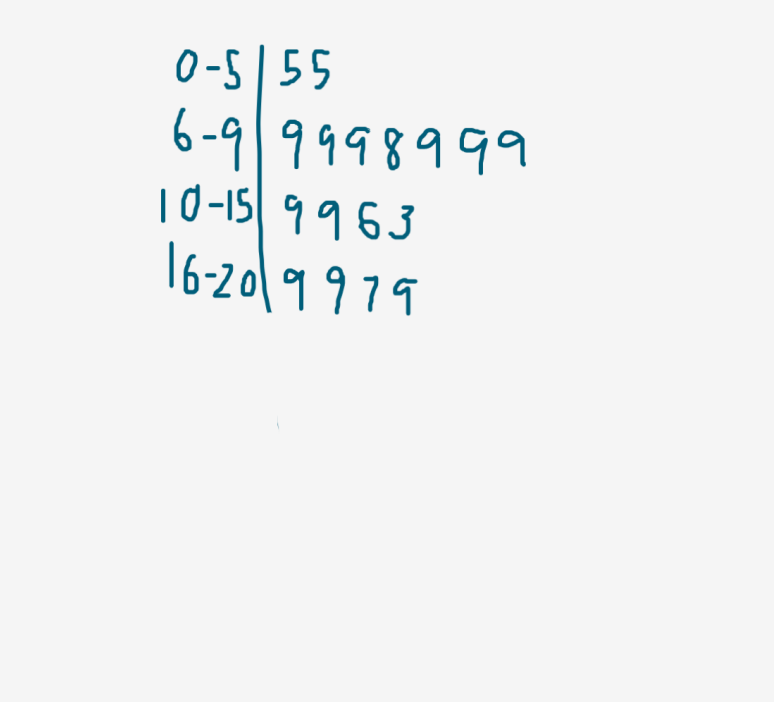


A benefit is that the numbers can be read off the plot, but the focus is still on the pattern. Also quantiles like the median, can be computed easily.

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 | 000001233334445555555555566666677777788889
##    2 | 00000000000000000000000000000000000011222222233335555555555555566666677788899
##    3 | 00000000000000000000001111111222222233334444555555555555566677888 9
##    4 | 000000000001112233335777
##    5 | 00000000011222267 99
##    6 | 05577
##    7 | 6
##    8 |
##    9 | 0
##   10 | 0
```

```
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 | 00000000000000000000000267891245688
##    32 | 133557159
##    34 | 0188800000000015
##    36 | 0181566
##    38 | 2
##    40 | 0000000000006889
##    42 | 09004
##    44 | 0
##    46 | 713
##    48 |
##    50 | 000000000074567
```

💭 Similar information to the histogram, but we can see the actual numbers too.

Refining the size

			(#)
3·	8	Tate	(1)
4*	01212431213002142O2		(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)

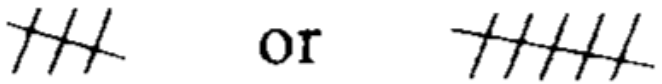
A) FIVE-LINE VERSION

		(#)
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		
·		(34√)

a different style of scratching

We know about

but its too easy to



make a mistake

Try this instead

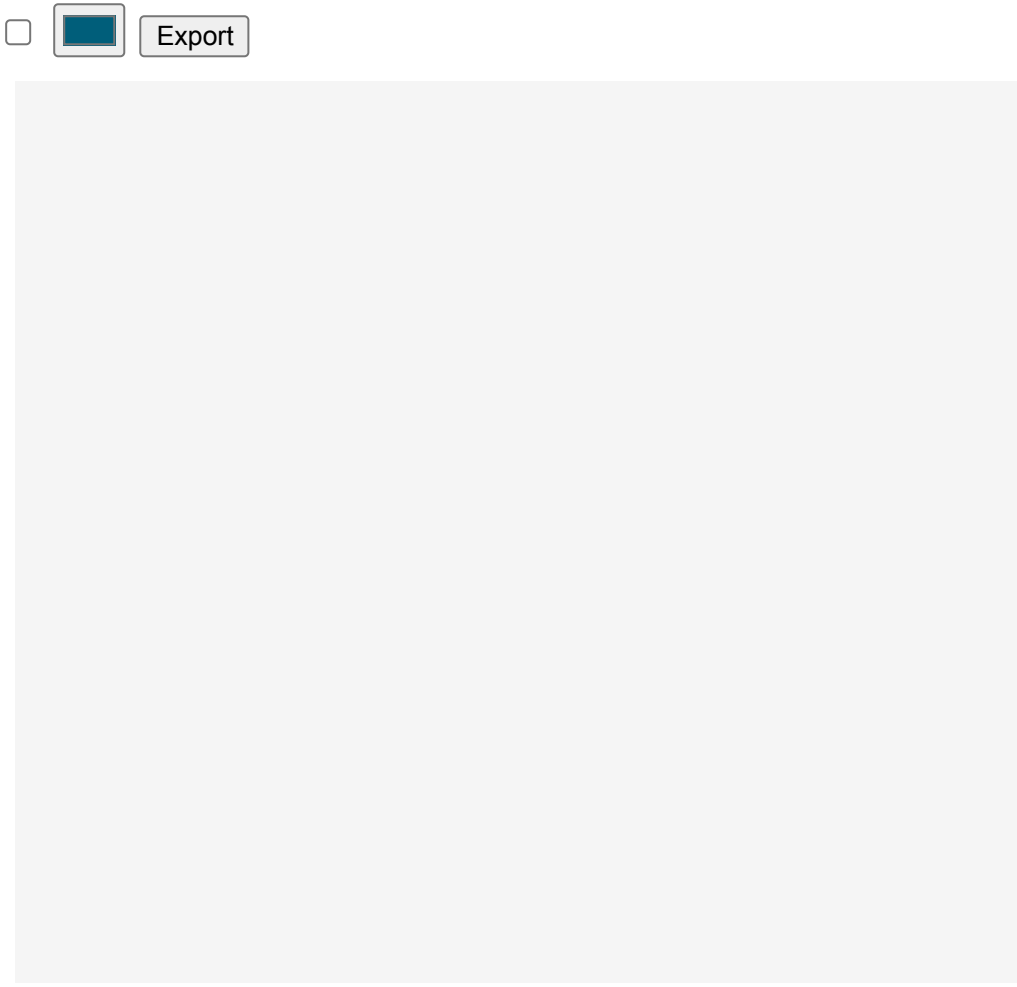
4 is 

8 is 

10 is 

Count this data using the squares approach.

```
## # A tibble: 25 x 2
##    smoker day
##    <chr>  <chr>
##  1 Yes    Sun
##  2 Yes    Sun
##  3 Yes    Sun
##  4 No     Sun
##  5 No     Sat
##  6 No     Sun
##  7 Yes    Sun
##  8 No     Sat
##  9 No     Sat
## 10 Yes    Sun
## 11 No     Sun
## 12 Yes    Sat
## 13 No     Sun
## 14 No     Sun
## 15 Yes    Sat
## 16 Yes    Sat
## 17 No     Sun
## 18 Yes    Sat
## 19 No     Sun
## 20 No     Sat
## 21 Yes    Sun
## 22 No     Sun
## 23 No     Sun
## 24 No     Sun
## 25 Yes    Sat
```



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, √)

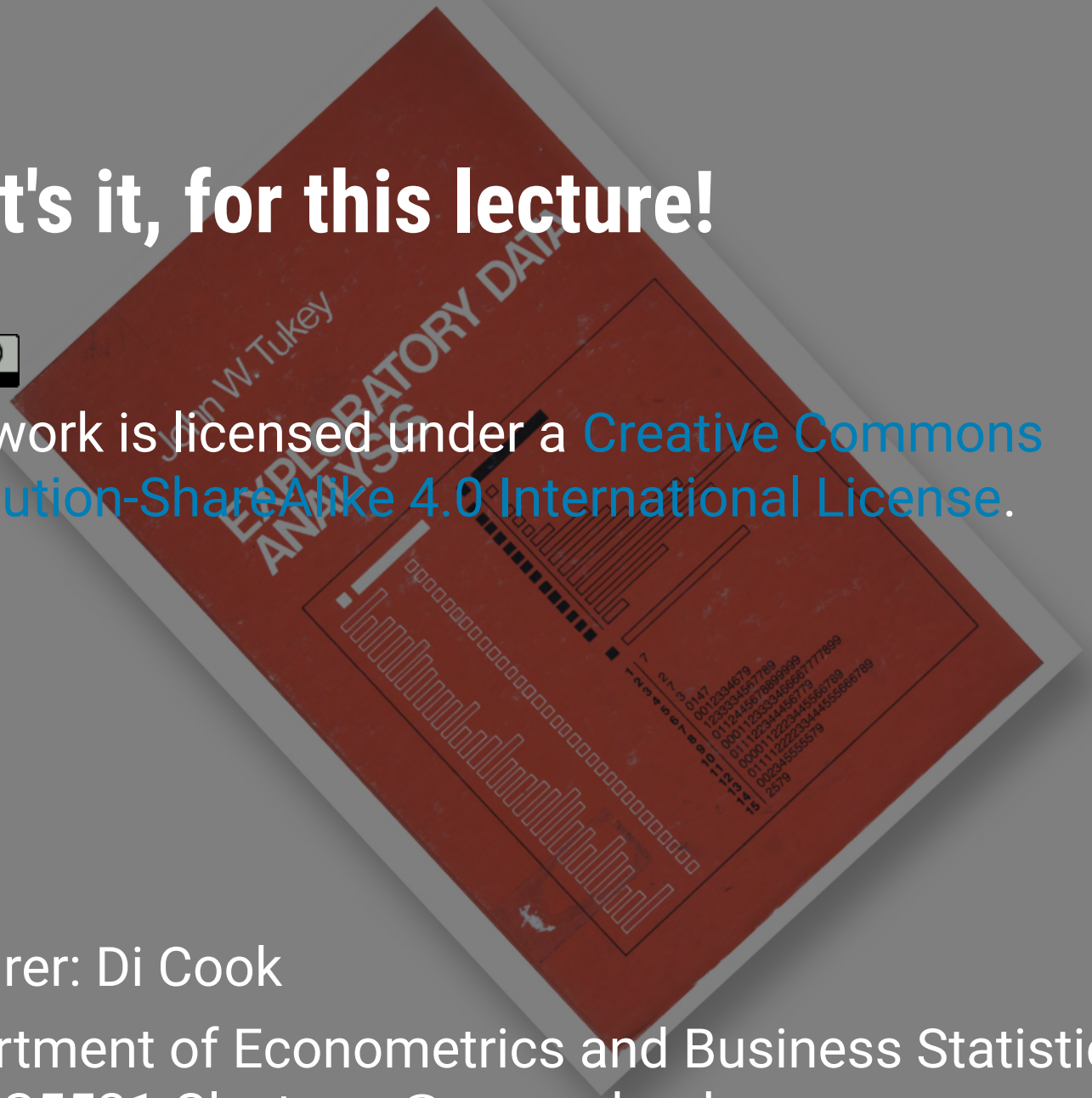
This is a stem and leaf of the height of the highest peak in each of the 50 US states.

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.

That's it, for this lecture!



Lecturer: Di Cook

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu