MONASH University

ETC5521: Exploratory Data Analysis

Working with a single variable, making transformations, detecting outliers, using robust statistics

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

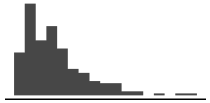
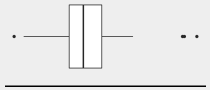
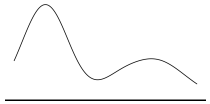
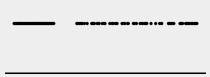
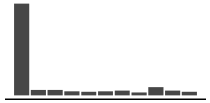
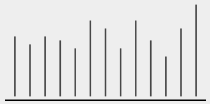
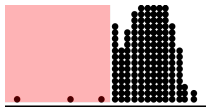
Week 5 - Session 1

Continuous variables

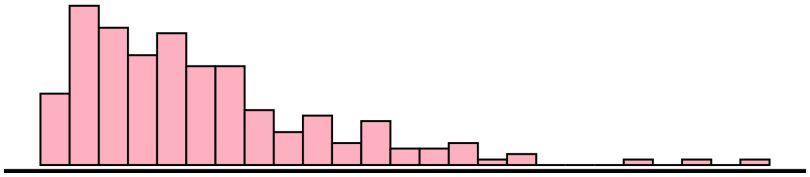
This lecture is partly based on Chapter 3 of

Unwin (2015) Graphical Data Analysis with R

Possible features of a single continuous variable

Feature	Example	Description
Asymmetry		The distribution is not symmetrical.
Outliers		Some observations are that are far from the rest.
Multimodality		There are more than one "peak" in the observations.
Gaps		Some continuous interval that are contained within the range but no observations exists.
Heaping		Some values occur unexpectedly often.
Discretized		Only certain values are found, e.g. due to rounding.
Implausible		Values outside of plausible or likely range.

Numerical features of a single continuous variables



- A measure of ***central tendency***, e.g. mean, median and mode
- A measure of ***dispersion*** (also called variability or spread), e.g. variance, standard deviation and interquartile range
- There are other measures, e.g. ***skewness*** and ***kurtosis*** that measures "tailedness", but these are not as common as the measures of first two
- The mean is also the *first moment* and variance, skewness and kurtosis are *second, third, and fourth central moments*

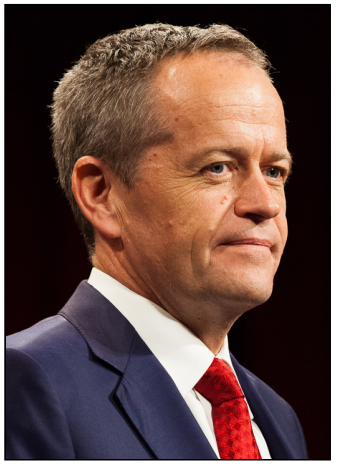
Significance tests or **hypothesis tests**

- Testing for $\mathrm{H}_0 : \mu = \mu_0$ vs. $\mathrm{H}_1 : \mu \neq \mu_0$ (often $\mu_0 = 0$)
- The t-test is commonly used if the underlying data are believed to be normally distributed

Case study 1 2019 Australian Federal Election Part 1/8

Context

- There are 151 seats in the House of Representative for the 2019 Australian federal election
- The major parties in Australia are:
  - the **Coalition**, comprising of the:
    - **Liberal**,
    - **Liberal National** (Qld),
    - **National**, and
    - **Country Liberal** (NT) parties, and
  - the Australian **Labor** party
- The **Greens** party is a small but notable party



Case study 1 2019 Australian Federal Election Part 2/8

https://results.aec.gov.au/24310/Website/Downloads/HouseFirstPrefsByCandidateByVoteTypeDownload-24310.csv

Copy CSV

Search:

StateAb	DivisionID	DivisionNm	CandidateID	Surname	GivenNm	BallotPosition	Elected	HistoricElected	PartyAb	PartyNm
ACT	318	Bean	33426	FAULKNER	Therese	1	N	N	AUP	Australian Progressives
ACT	318	Bean	32130	CHRISTIE	Jamie	2	N	N	IND	Independent
ACT	318	Bean	33391	RUSHTON	Ben	3	N	N	GAP	The Great A Party
ACT	318	Bean	32921	DONNELLY	Matt	4	N	N	LDP	Liberal Dem
ACT	318	Bean	32261	HANLEY	Tony	5	N	N	UAPP	United Austr
ACT	318	Bean	33397	COCKS	Ed	6	N	N	LP	Liberal
ACT	318	Bean	32253	SMITH	David	7	Y	N	ALP	Australian La Party
ACT	318	Bean	32405	DAVIS	Johnathan	8	N	N	GRN	The Greens

Case study 1 2019 Australian Federal Election Part 3/8

What is the number of the seats won in the House of Representatives by parties?

data R

Party	# of seats
Coalition	77
Liberal	44
Liberal National Party Of Queensland	23
The Nationals	10
Australian Labor Party	68
The Greens	1
Centre Alliance	1
Katter's Australian Party (Kap)	1
Independent	3

What does this table tell you?

- The Coalition won the government
- Labor and Coalition hold majority of the seats in the House of Representatives (lower house)
- Parties such as The Greens, Centre Alliance and Katter's Australian Party (KAP) won *only* a single seat

Only? Wait... **Did the parties compete in all electoral districts?**

Case study 1 2019 Australian Federal Election Part 4/8

data R

Copy CSV Search:

Party	# of electorates
Australian Labor Party	151
Informal	151
The Greens	151
United Australia Party	151
Liberal	107
Independent	95
Pauline Hanson's One Nation	59
Fraser Anning's Conservative National Party	48
Animal Justice Party	46
Christian Democratic Party (Fred Nile Group)	42

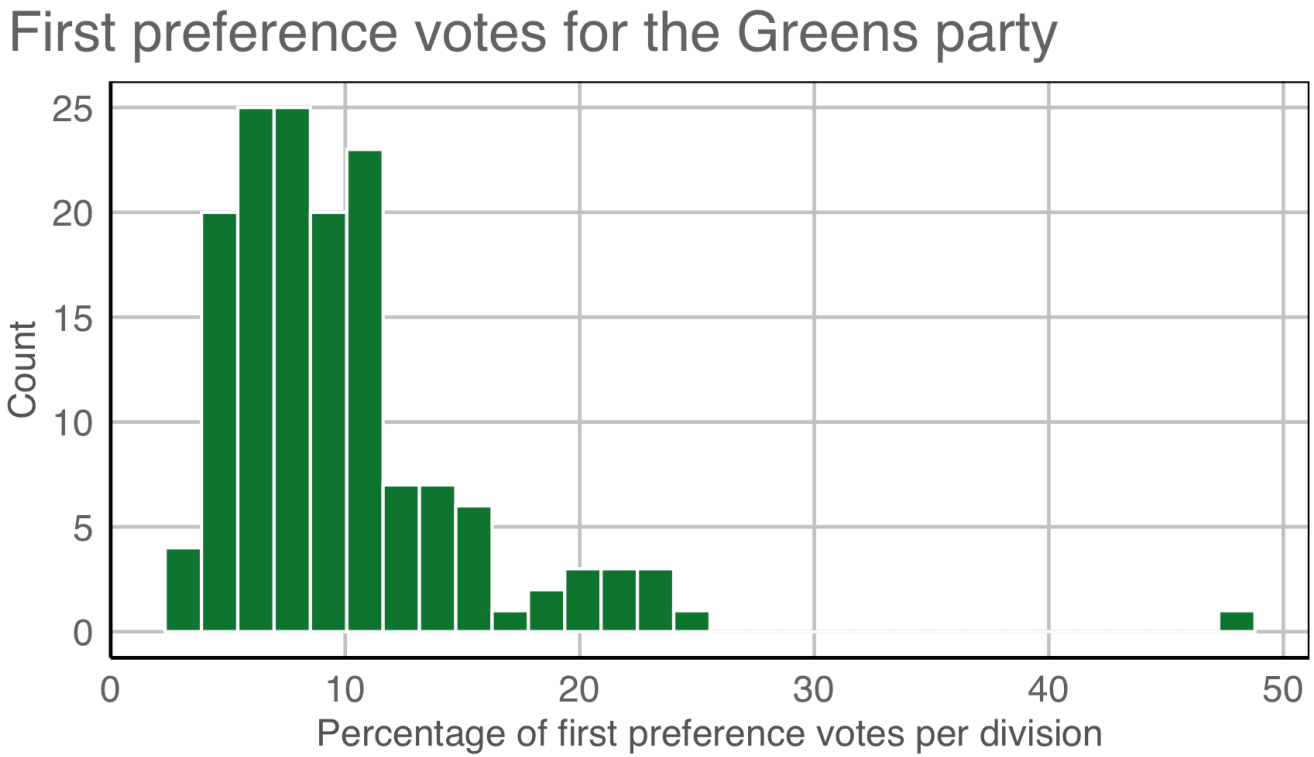
What do you notice from this table?

- The Greens are represented in every electoral districts
- United Australia Party is the only other non-major party to be represented in every electoral district
- KAP is represented in 7 electoral districts
- Centre Alliance is only represented in 3 electoral districts!

Let's have a closer look at the Greens party...

Case study 1 2019 Australian Federal Election Part 5/8

data R



What does this graph tell you?

- Majority of the country does not have first preference for the Greens
- Some constituents are slightly more supportive than the others

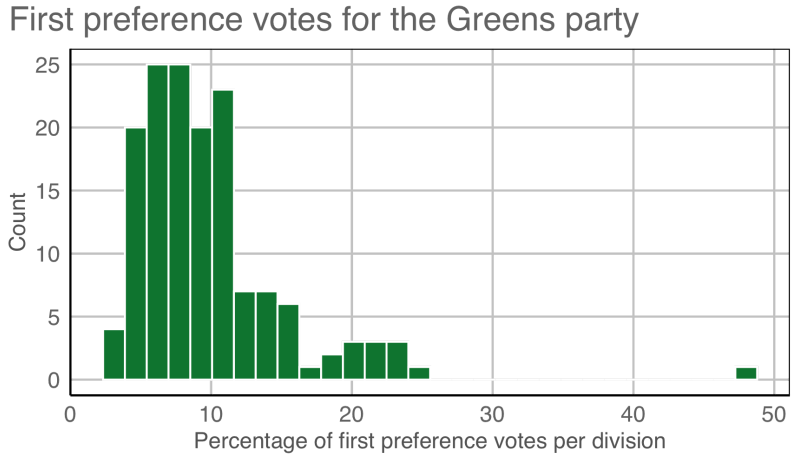
What further questions does it raise?

Formulating questions for EDA vs making observations from a plot

- BEFORE plotting or making summaries think **broad (open-ended) questions** that promotes discussion and divergent thinking
- Questions with simple answers (i.e. yes or no) less helpful in encouraging exploration
- For example,

> What is the distribution of the first preference vote percentages for the Labor party across Australia? Is it evenly spread across electorates or are there clusters of popularity?

- AFTER plotting or making summaries think **was this what you expected, are there any surprises**. Detail what you learn, and how you should follow up on these observations.



> Is the outlying observation the electoral district that won the seat?

Visual inference

Typical plot description:

```
ggplot(data, aes(x=var1)) +
  geom_histogram()
```

Is the distribution consistent with a sample from a particular statistical distribution?

Potential simulation methods from specific distributions

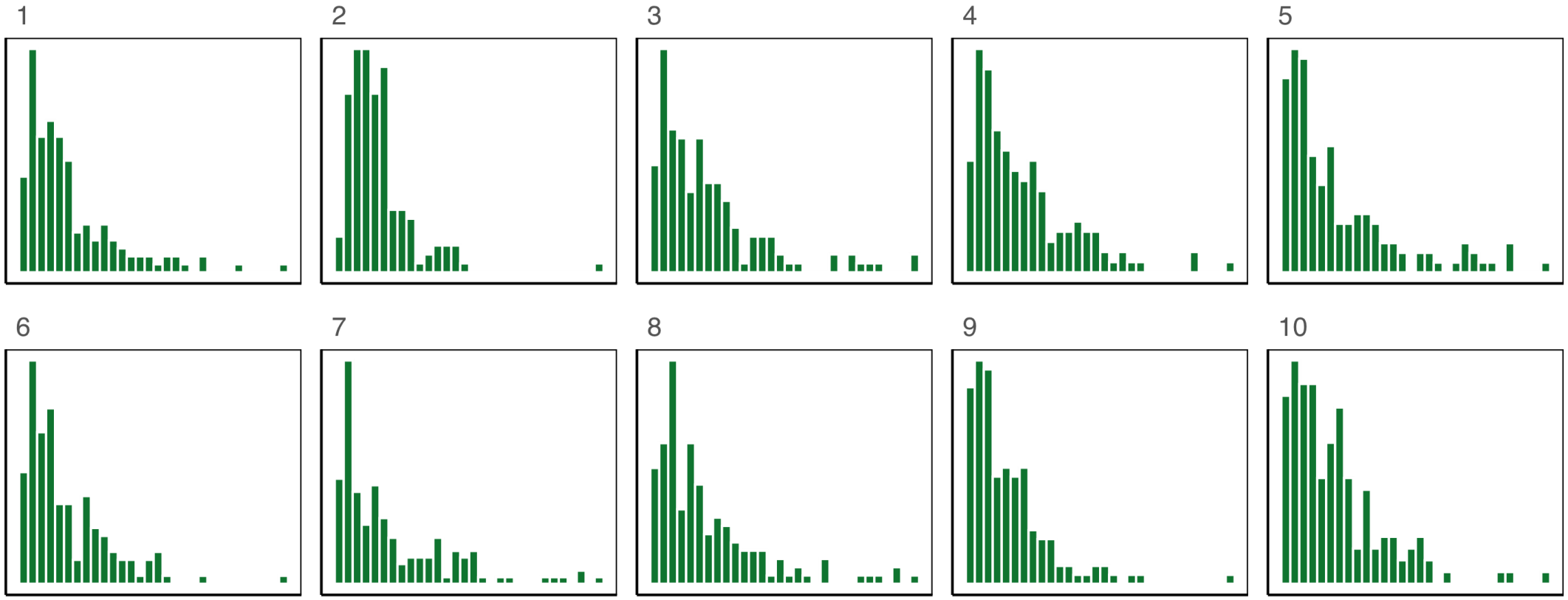
```
# Symmetric, unimodal, bell-shaped
null_dist("var1", "norm")
null_dist("var1", "cauchy")
null_dist("var1", "t")

# Skewed right
null_dist("var1", "exp")
null_dist("var1", "chisq")
null_dist("var1", "gamma")

# Constant
null_dist("var1", "uniform")
```

Lineup of Greens first preference percentages

R



Case study 1 2019 Australian Federal Election Part 6/8

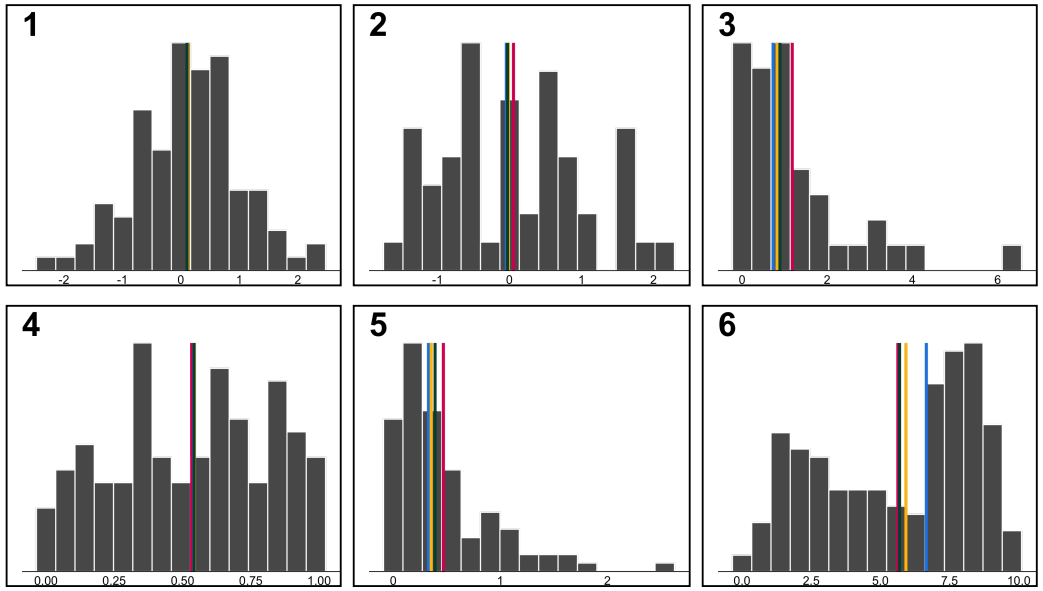
data R

State	% of first preference for the Greens					
	Mean	Median	SD	IQR	Skewness	Kurtosis
ACT	16.406	13.988	5.602	5.196	0.645	1.500
VIC	11.400	8.570	8.210	6.717	2.603	11.360
WA	10.993	10.756	3.018	3.116	0.802	3.026
QLD	9.764	8.808	5.096	4.753	1.092	3.886
TAS	9.721	9.339	4.009	0.985	0.326	2.493
NT	9.572	9.572	2.473	1.748	0.000	1.000
SA	9.120	8.903	3.024	3.412	0.384	2.920
NSW	8.101	6.635	4.087	3.948	1.502	4.859
National	9.874	8.547	5.632	5.001	2.671	15.798

- Why are the means and the medians different?
- How are the standard deviations and the interquartile ranges similar or different?
- Are there some other numerical statistics we should show?

Robust measure of central tendency

- **Mean** is a non-robust measure of location.
- **Median** is the 50% quantile of the observations
- **Trimmed mean** is the sample mean after discarding observations at the tails.
- **Winsorized mean** is the sample mean after replacing observations at the tails with the minimum or maximum of the observations that remain.



Plot	Mean	Median	Trimmed Mean*	Winsorized Mean*
1	0.109	0.114	0.120	0.103
2	0.054	-0.045	-0.016	-0.029
3	1.177	0.729	0.820	0.888
4	0.533	0.541	0.543	0.542
5	0.468	0.329	0.355	0.390
6	5.626	6.656	5.918	5.688

* Both trimmed and Winsorized mean trimmed 20% of the tails.

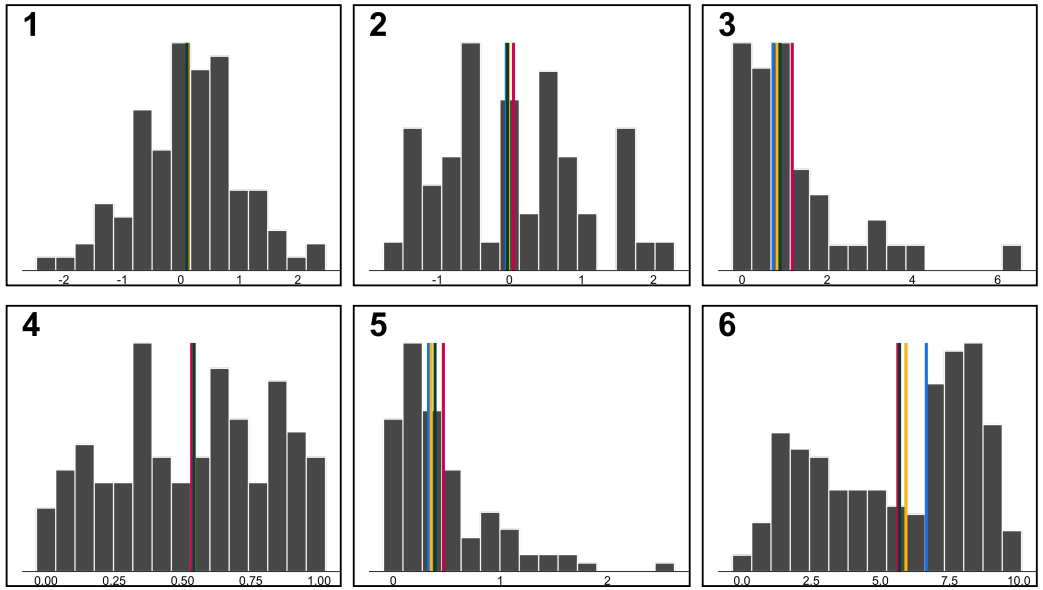
Robust measure of dispersion

- **Standard deviation** or its square, **variance**, is a popular choice of measure of dispersion but is not robust to outliers
- Standard deviation for sample $x_1, \ldots, x_n$ is

$$\sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n-1}}$$

- **Interquartile range** difference between 1st and 3rd quartile, more robust measure of spread
- **Median absolute deviance** (MAD) is even more robust

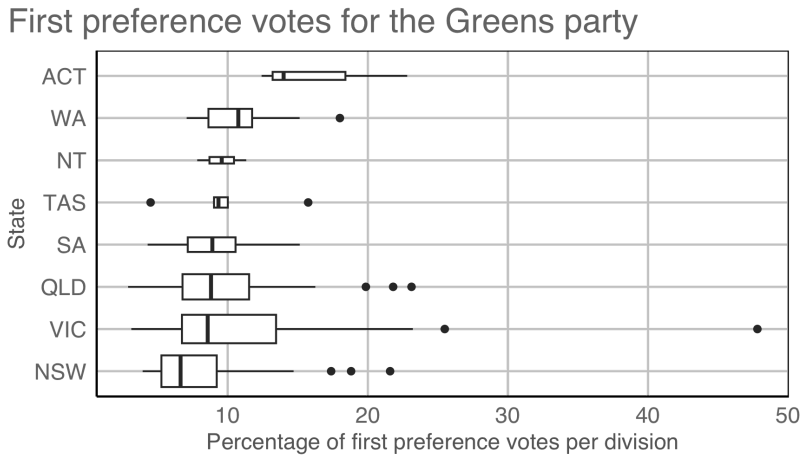
$$\text{median}(|x_i - \text{median}(x_i)|)$$



	Measure of dispersion				
Plot	SD	IQR	MAD	Skewness	Kurtosis
1	0.898	1.186	0.870	-0.072	3.008
2	0.986	1.411	1.077	0.358	2.212
3	1.326	1.176	0.793	1.944	7.184
4	0.288	0.450	0.335	-0.126	1.837
5	0.468	0.499	0.343	1.691	6.372
6	2.784	5.362	2.984	-0.351	1.678

Case study 1 2019 Australian Federal Election Part 7/8

data R



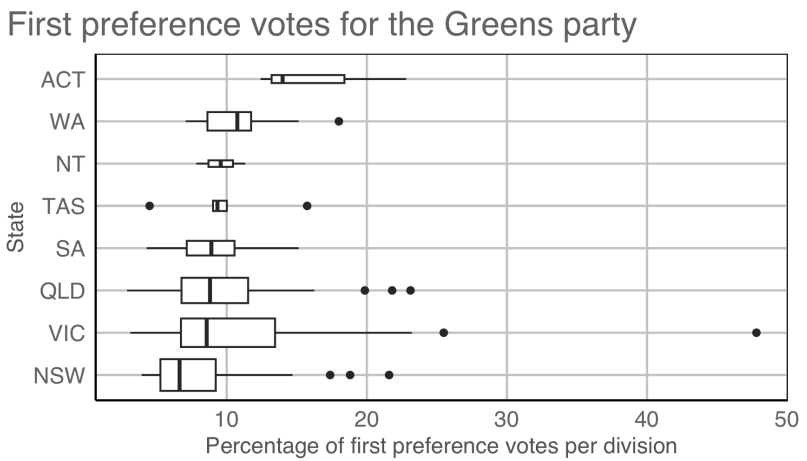
We should plot the data!

- The width of the boxplot is proportional to the number of electoral districts in the corresponding state (which is roughly proportional to the population)

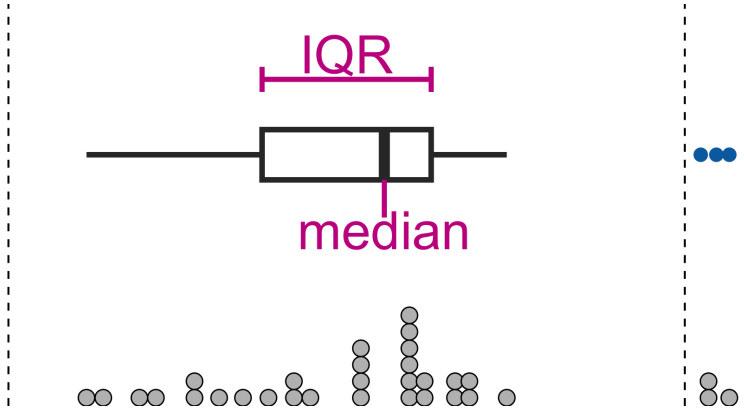
Outliers

Outliers are *observations* that are significantly different from the majority.

- Outliers can ***occur by chance in almost all distributions***, but could be indicative of:
  - a measurement error,
  - a different population, or
  - an issue with the sampling process.



Closer look at the *boxplot*

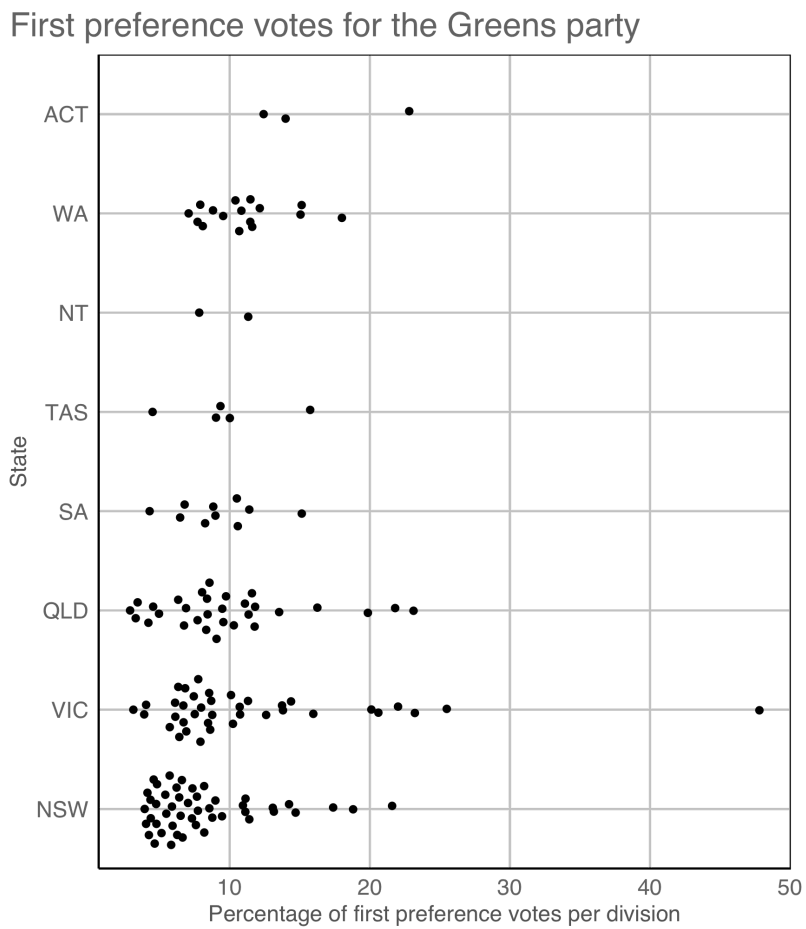


- Observations that are outside the range of lower to upper fence (1.5 times the box length) are referred at times as outliers
- Plotting boxplots for data from a skewed distribution will almost always show these "outliers" but these are not necessary outliers
- Some definitions of outliers assume a symmetrical population distribution (e.g. in boxplots or observations a certain standard deviations away from the mean) and these definitions are ill-suited for asymmetrical distributions

But are there some things we *cannot* see from boxplots?

Case study 1 2019 Australian Federal Election Part 8/8

data R



Now what do you notice from this graph that you didn't notice before?

- There are only two electoral districts in NT!
- And only 3 and 5 electoral districts in ACT and TAS, respectively!
- We have *not* computed the number of electoral districts for each state so far!

Both numerical and graphical summaries can either *reveal* and/or *hide* aspects of the data

Transformations

Case study 2 Melbourne Housing Prices Part 1/5

Suburb	Rooms	Type	Price ($)	Date
Abbotsford	3	Home	1,490,000	2017-04-01
Abbotsford	3	Home	1,220,000	2017-04-01
Abbotsford	3	Home	1,420,000	2017-04-01
Aberfeldie	3	Home	1,515,000	2017-04-01
Airport West	2	Home	670,000	2017-04-01
Airport West	2	Townhouse	530,000	2017-04-01
Airport West	2	Unit	540,000	2017-04-01
Airport West	3	Home	715,000	2017-04-01
Albanvale	6	Home	NA	2017-04-01
Albert Park	3	Home	1,925,000	2017-04-01
Albion	3	Unit	515,000	2017-04-01
Albion	4	Home	717,000	2017-04-01
Alphington	2	Home	1,675,000	2017-04-01
Alphington	4	Home	2,008,000	2017-04-01
Altona	2	Home	860,000	2017-04-01
Altona Meadows	4	Home	NA	2017-04-01
Altona North	3	Home	720,000	2017-04-01
Armadale	2	Unit	836,000	2017-04-01
Armadale	2	Home	2,110,000	2017-04-01

- This data was scrapped each week from domain.com.au from 2016-01-28 to 2018-10-13
- In total there are **63,023** observations
- All variables shown (there are more variables not shown here), except price, have complete records
- The are **48,433** property prices across Melbourne (roughly 23% missing)

How would you explore this data first?

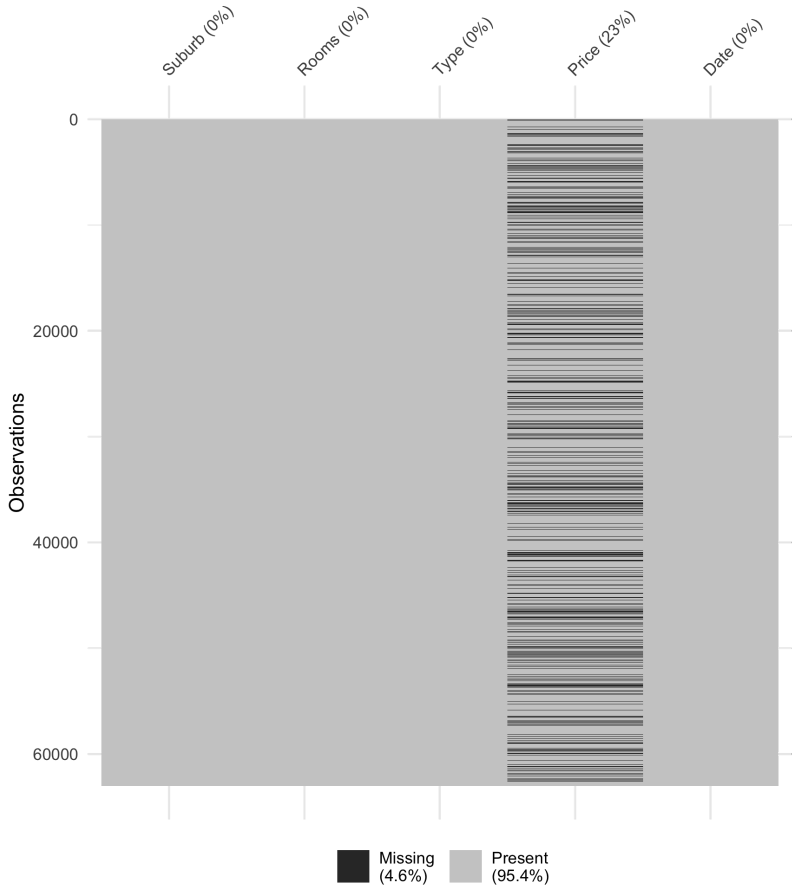
01:00

Data source: Tony Pio (2018) Melbourne Housing Market, Version 27. Retrieved August 2021 from https://www.kaggle.com/anthonypino/melbourne-housing-market.

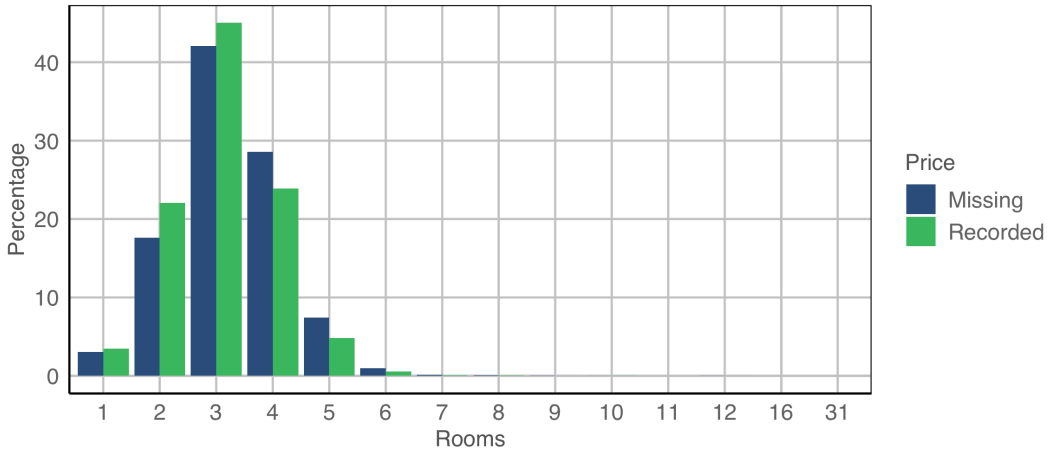
Case study 2 Melbourne Housing Prices Part 2/5

data R lineup R

Observations arranged by Suburb and Date:



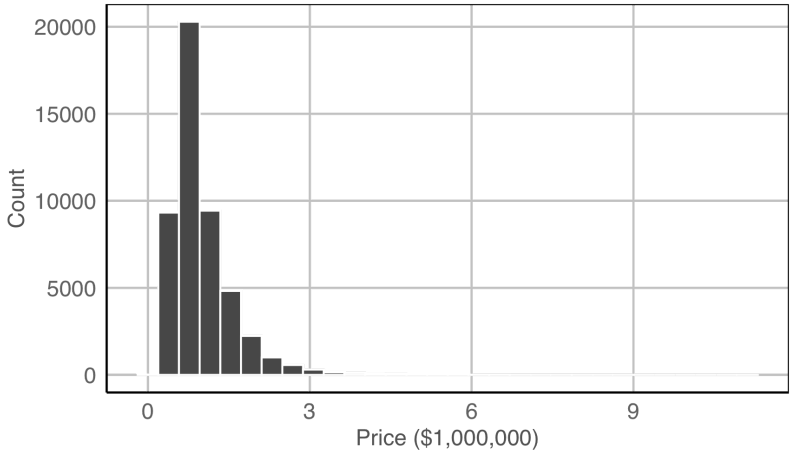
Comparing distribution of room number for observations with missing and non-missing price records:



- Seems to be okay nothing very notable - but check with a lineup
- What next?

Case study 2 Melbourne Housing Prices Part 3/5

data R

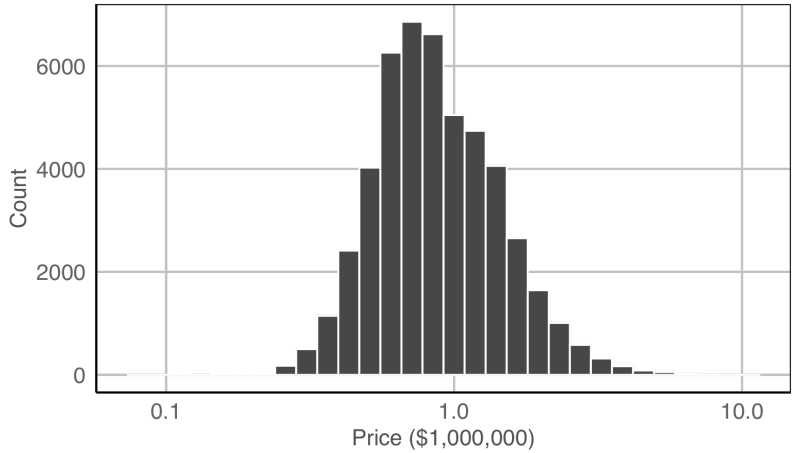


What can we say from this plot?

- The housing prices are right-skewed
- There appears to be a lot of outlying housing prices (how can we tell?)

Case study 2 Melbourne Housing Prices Part 4/5

data R



- The x-axis has been $\log_{10}$-transformed in this plot
- The plot appears more symmetrical now
- What is a measure of central tendancy here?

With no transformation:

Mean	Median	Trimmed Mean	Winsorised Mean
$997,898	$830,000	$871,375	$903,823

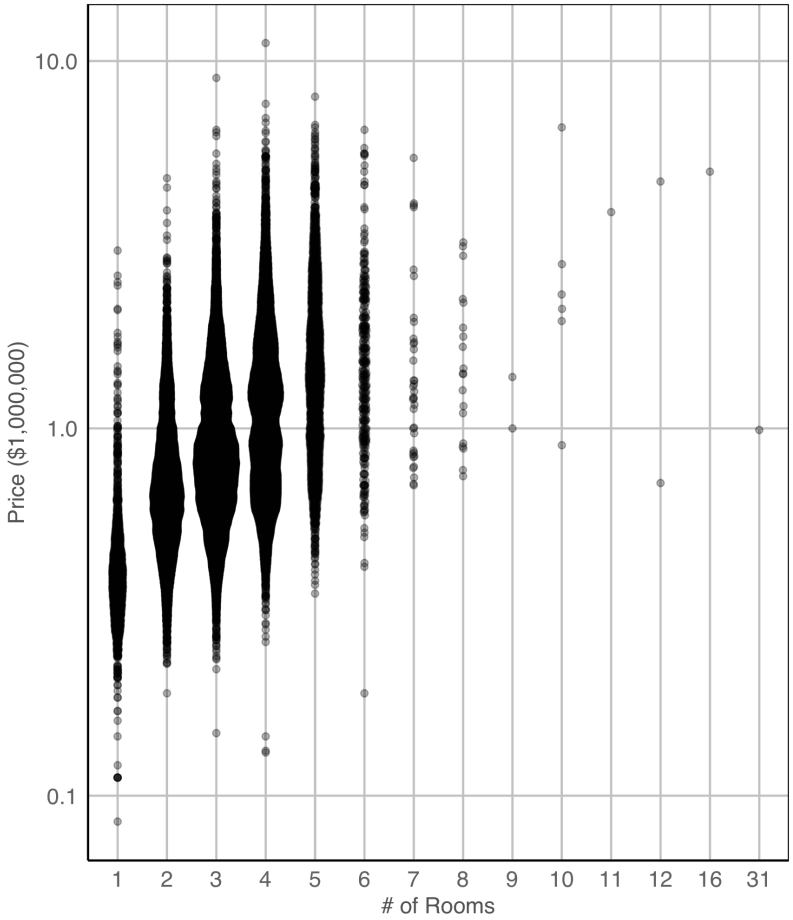
With log transformation (and back transformed to original scale):

Mean	Median	Trimmed Mean	Winsorised Mean
$874,166	$830,000	$847,973	$859,325

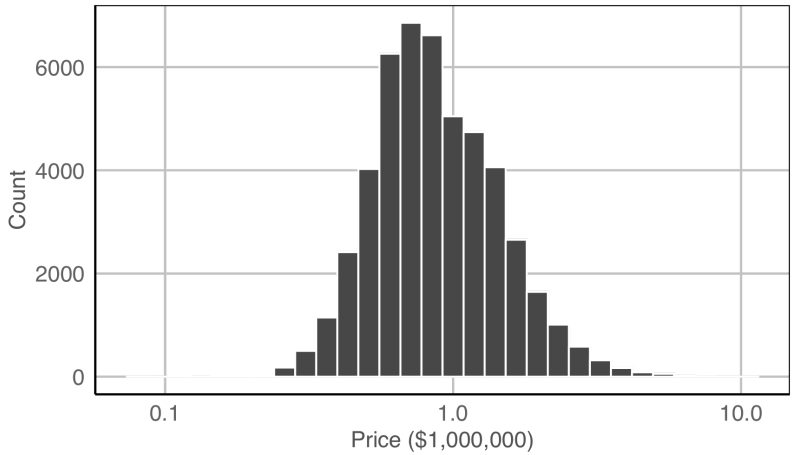
Multi-modality

Case study 2 Melbourne Housing Prices Part 5/5

data R



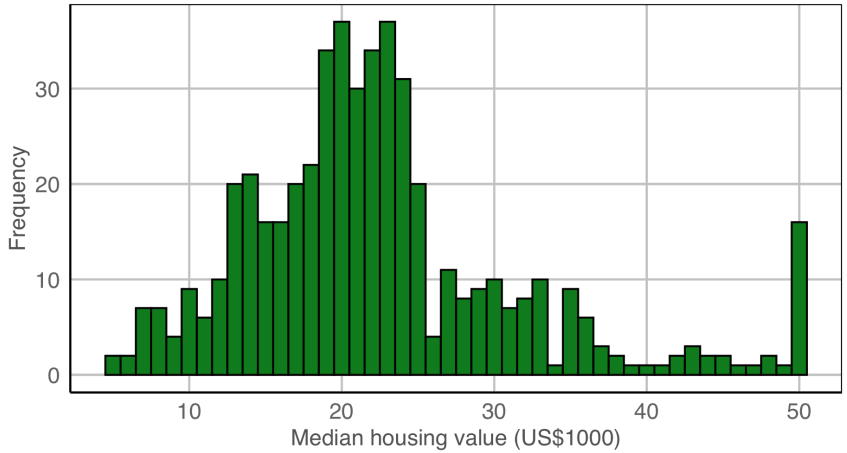
- You can see that drawing separate univariate plots for each room number show that higher number of rooms generally are pricier
- You could not see this, however, when the data are combined



Bins and Bindwidths: More details

Case study 3 Boston housing data Part 1/4

data R



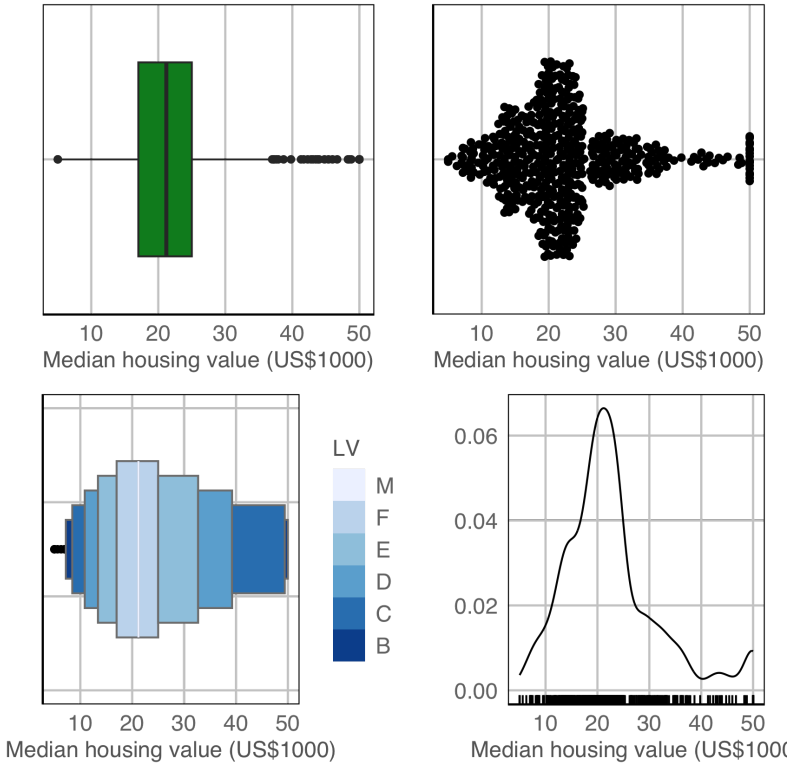
- There is a large frequency in the final bin.
- There is a decline in observations in the $40-49K range as well as dip in observations around $26K and $34K.

- The histogram is using a bin width of 1 unit and is **left-open** (or **right-closed**): (4.5, 5.5], (5.5, 6.5] ... (49.5, 50.5], so that 5.5 is in the smaller bin, where as right-open would place it in the larger bin.
- Occasionally, whether it is **left-** or **right-open** can make a difference. Or, you might also **set the breaks** controlling the min value where binning starts.

Harrison, David, and Daniel L. Rubinfeld (1978) Hedonic Housing Prices and the Demand for Clean Air, *Journal of Environmental Economics and Management* **5** 81-102. Original data.
Gilley, O.W. and R. Kelley Pace (1996) On the Harrison and Rubinfeld Data. *Journal of Environmental Economics and Management* **31** 403-405. Provided corrections and examined censoring.
Maindonald, John H. and Braun, W. John (2020). DAAG: Data Analysis and Graphics Data and Functions. R package version 1.24

Case study 3 Boston housing data Part 2/4

data R

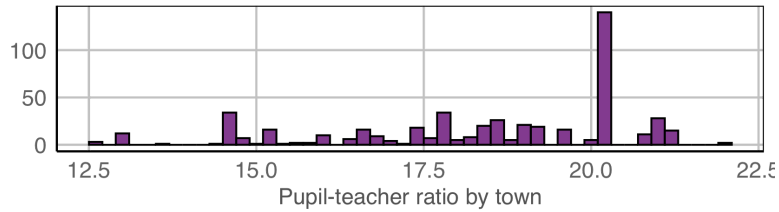


- Density plots depend on the **bandwidth** (binwidth) chosen and more than often do not estimate well at boundary cases
- There are various way to present features of the data using a plot and what works for one person, may not be as straightforward for another
- Be prepared to do multiple plots!

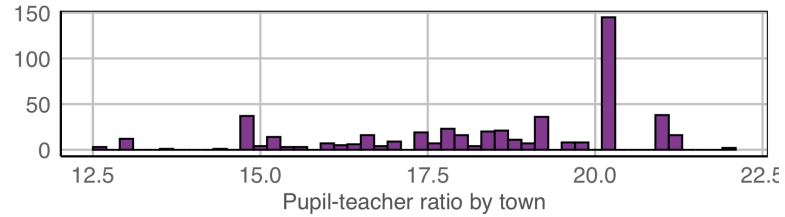
Case study 3 Boston housing data Part 3/4

data R

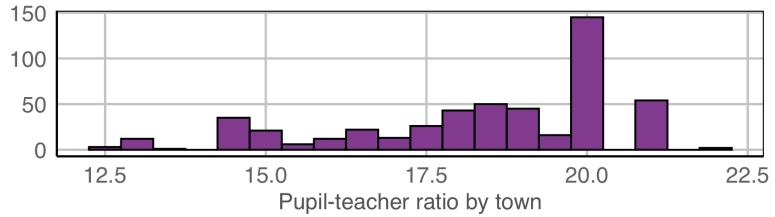
Bin width = 0.2, Left-open



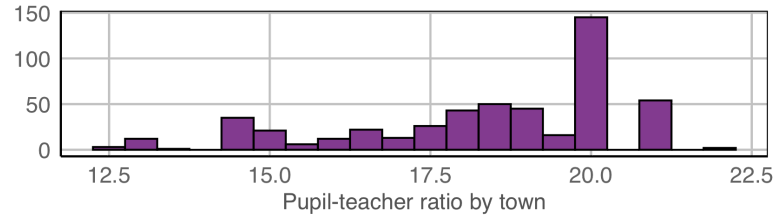
Bin width = 0.2, Right-open



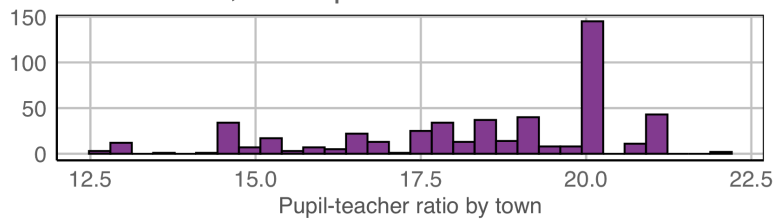
Bin width = 0.5, Left-open



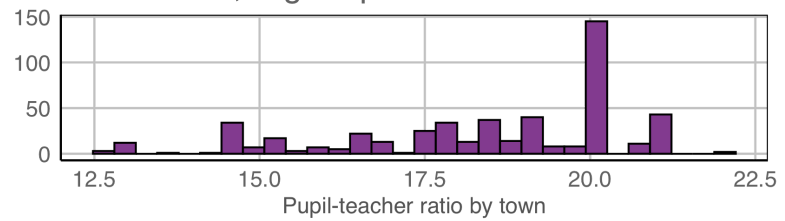
Bin width = 0.5, Right-open



Bin number = 30, Left-open

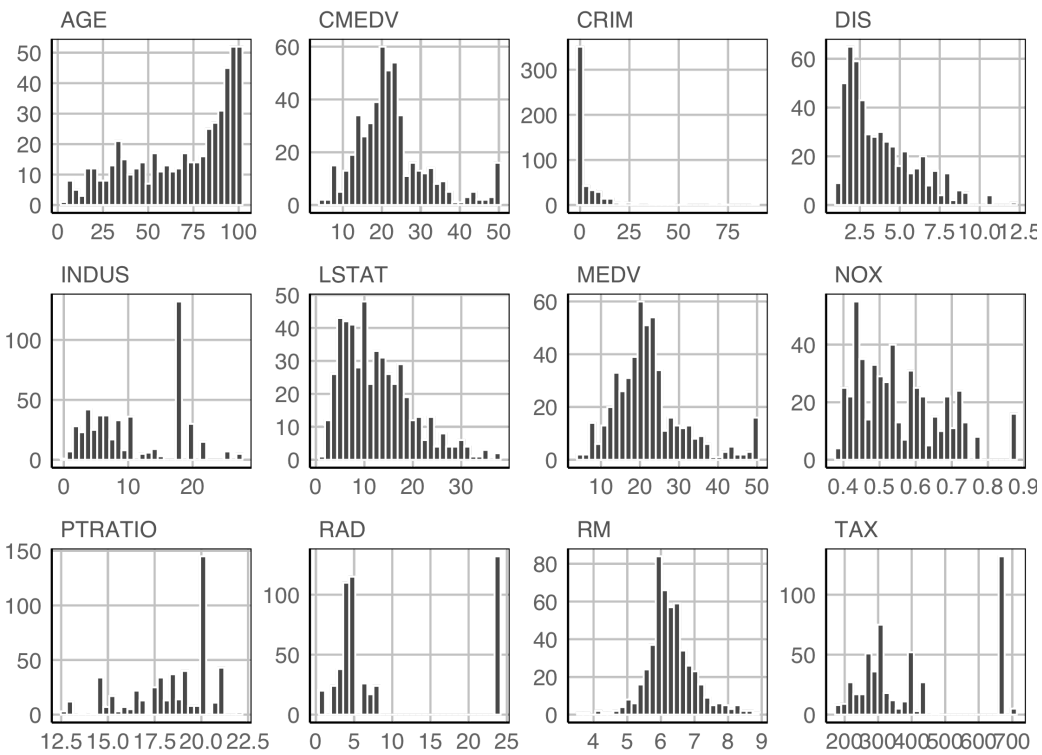


Bin number = 30, Right-open



Case study 3 Boston housing data Part 4/4

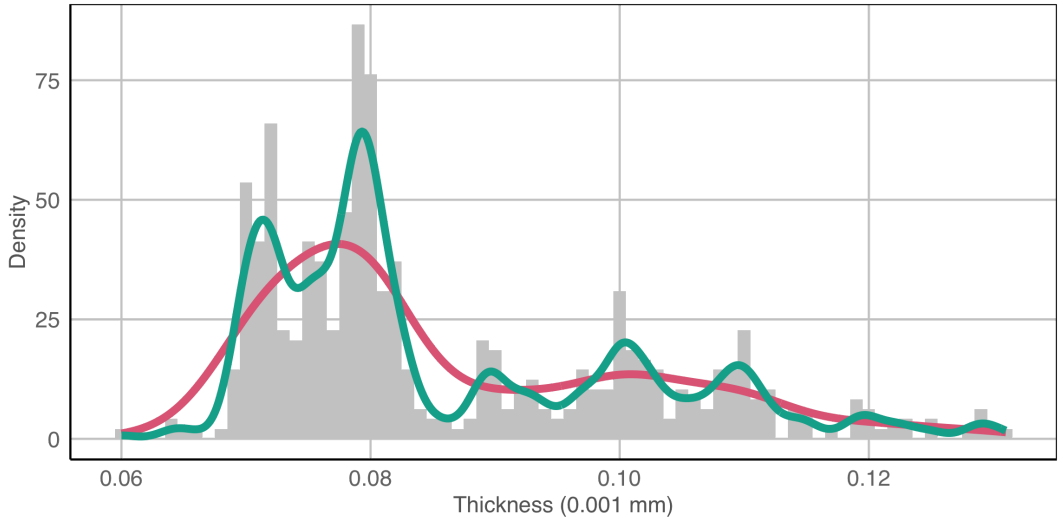
data R



- CRIM: per capita crime rate by town
- INDUS: proportion of non-retail business acres per town
- NOX: nitrogen oxides concentration (parts per 10 million)
- RM: average number of room per dwelling
- AGE: proportion of owner-occupied units built prior to 1940
- DIS: weighted mean of distances to 5 Boston employment centres
- RAD: index of accessibility to radial highways
- TAX: full-value property tax rate per $10K
- PTRATIO: pupil-teacher ratio by town
- LSTAT: lower status of the population (%)
- MEDV: median value of owner-occupied homes in $1000s

Case study 4 Hidalgo stamps thickness

data R

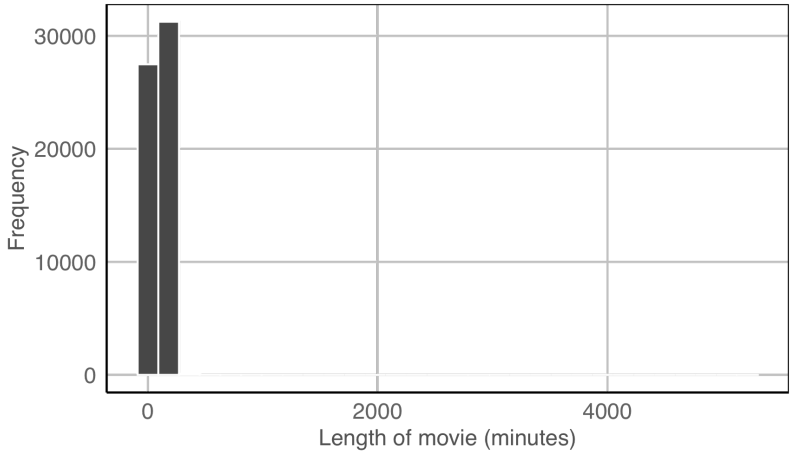


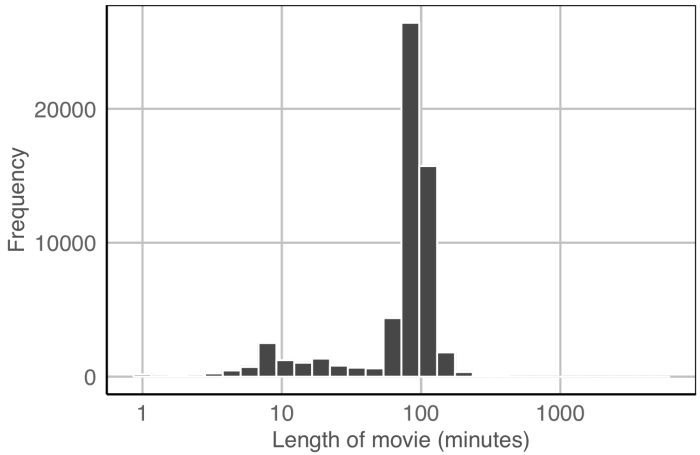
- A stamp collector, Walton von Winkle, bought several collections of Mexican stamps from 1872-1874 and measured the thickness of all of them.
- The different **bandwidth** for the density plot suggest either that there are two or seven modes.

Focus

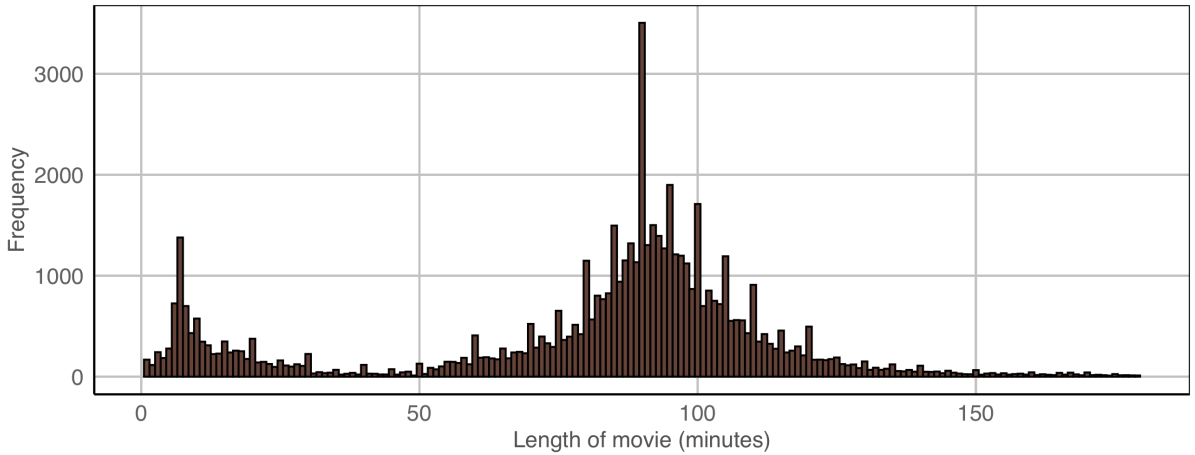
Case study 5 Movie length

data R





- Upon further exploration, you can find the two movies that are well over 16 hours long are "*Cure for Insomnia*", "*Four Stars*", and "*Longest Most Meaningless Movie in the World*"
- We can restrict our attention to films under 3 hours:



- Notice that there is a peak at particular times. Why do you think so?

Take away messages

- Numerical and graphical summaries can reveal, but also hide, aspects of data
- **Do many numerical and graphical summaries of the data!**

Resources and Acknowledgement

- Slides originally written by Emi Tanaka and constructed with `xaringan`, remark.js, `knitr`, and R Markdown.

Lecturer: *Di Cook*

ETC5521.Clayton-x@monash.edu

Week 5 - Session 1