ETC5521: Diving Deeply into Data Exploration

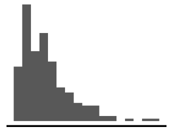
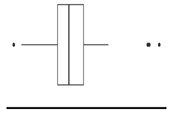
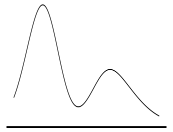
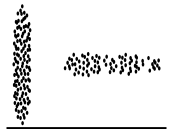
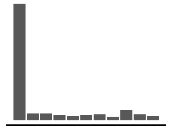
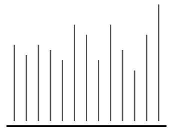
Working with a single variable, making transformations, detecting outliers, using robust statistics

Professor Di Cook

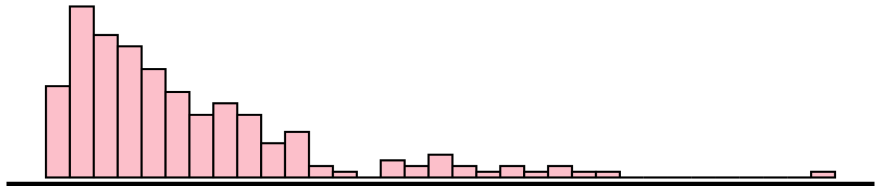
Department of Econometrics and Business Statistics

Quantitative variables

Features of a single quantitative 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.

Numerical features of a single quantitative 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 vs. (often)

$H_1 : \mu\mu \neq \mu_0 0$

- The -test is commonly used if the underlying data are believed to be normally distributed

t

2019 Australian Federal Election (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

Source: PRObono

2019 Australian Federal Election (2/8)

Copy CSV

Search:

StateAb	DivisionID	DivisionNm	CandidateID	Surname	GivenNm	BallotPosition
ACT	318	Bean	33426	FAULKNER	Therese	1
ACT	318	Bean	32130	CHRISTIE	Jamie	2
ACT	318	Bean	33391	RUSHTON	Ben	3
ACT	318	Bean	32921	DONNELLY	Matt	4
ACT	318	Bean	32261	HANLEY	Tony	5
ACT	318	Bean	33397	COCKS	Ed	6
ACT	318	Bean	32253	SMITH	David	7

Data source: Australian Electoral Commission. (2019)

2019 Australian Federal Election (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?**

2019 Australian Federal Election (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

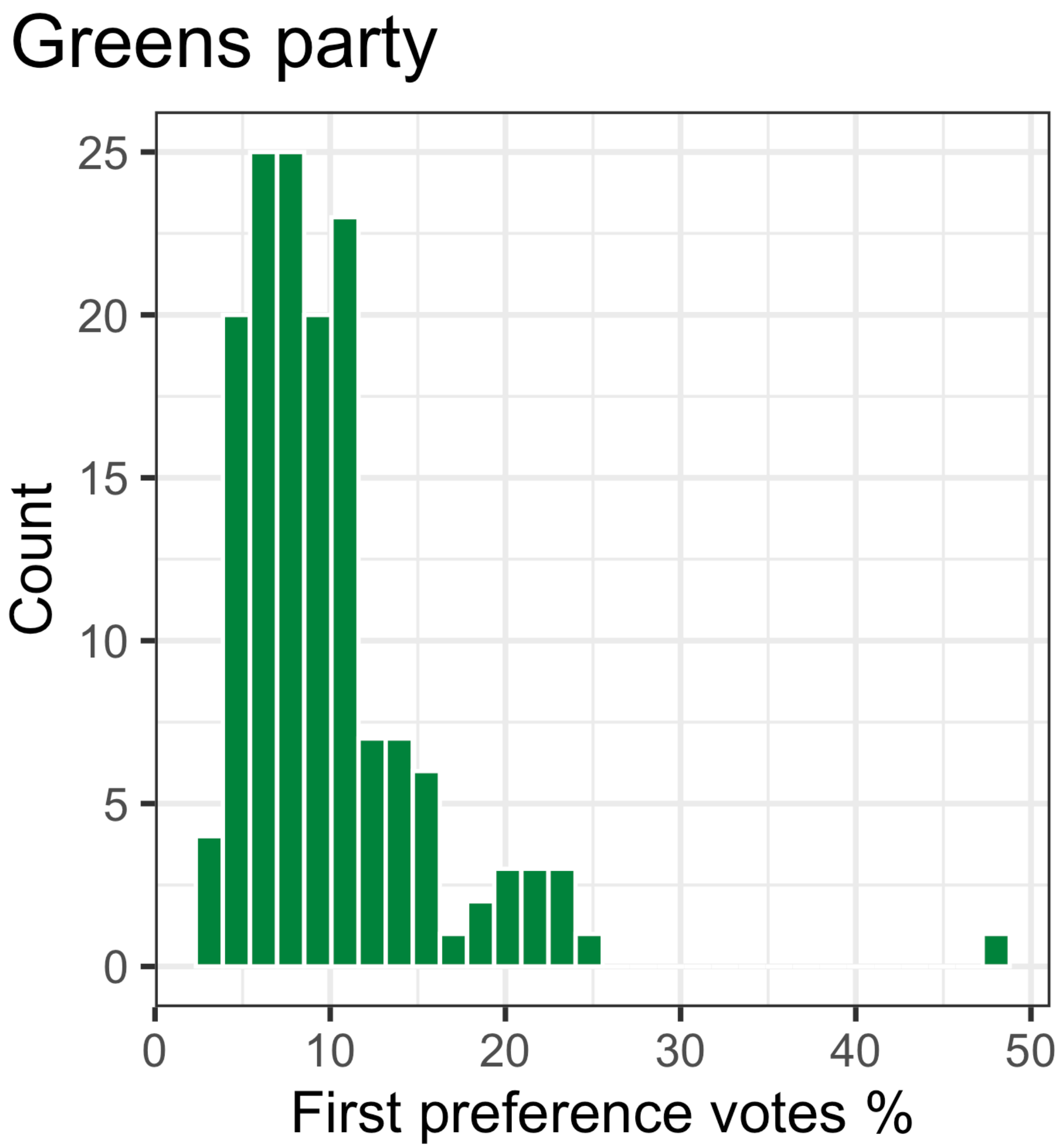
What do you notice from this table?

- 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…

2019 Australian Federal Election (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?

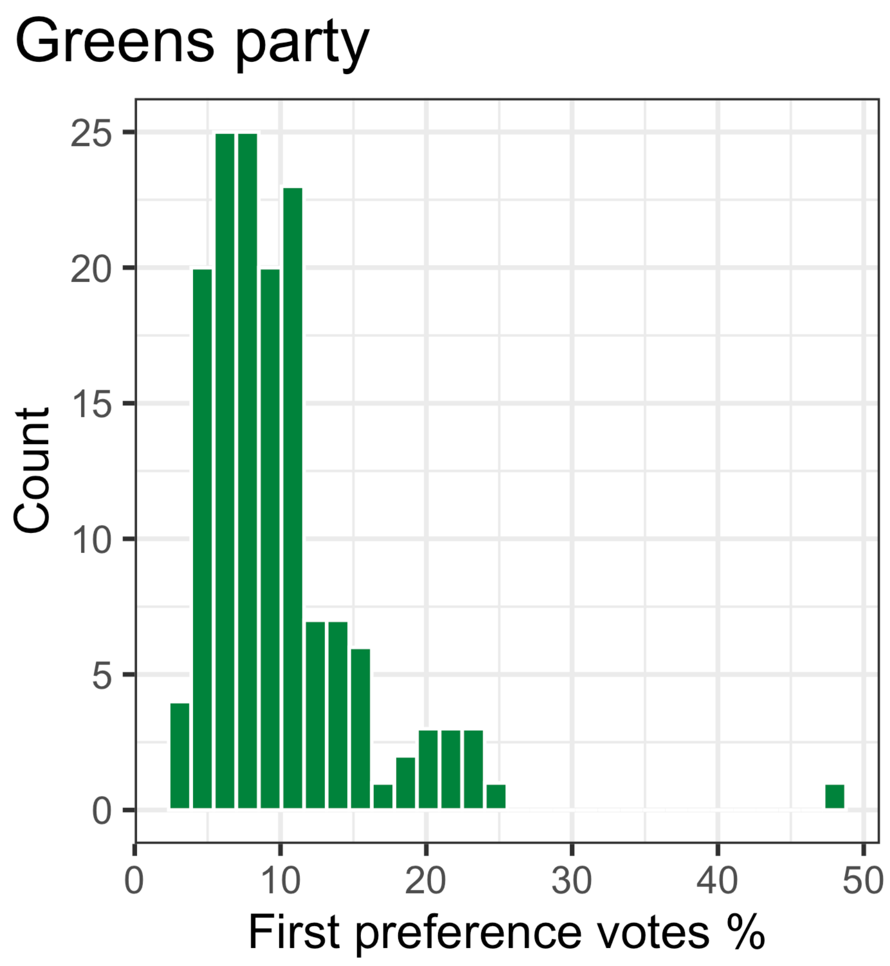
Notes:

- Australia uses full-preference instant-runoff voting in single member seats
- Following the full allocation of preferences, it is possible to derive a two-party-preferred figure, where the votes have been allocated between the two main candidates in the election.
- In Australia, this is usually between the candidates from the Coalition parties and the Australian Labor Party.

Formulating questions for EDA vs making observations from a plot

- BEFORE plotting or making summaries think **broad (open-ended) questions** about the distribution of values
- Questions with simple answers (i.e. yes or no) less helpful in encouraging exploration using graphics
- 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.

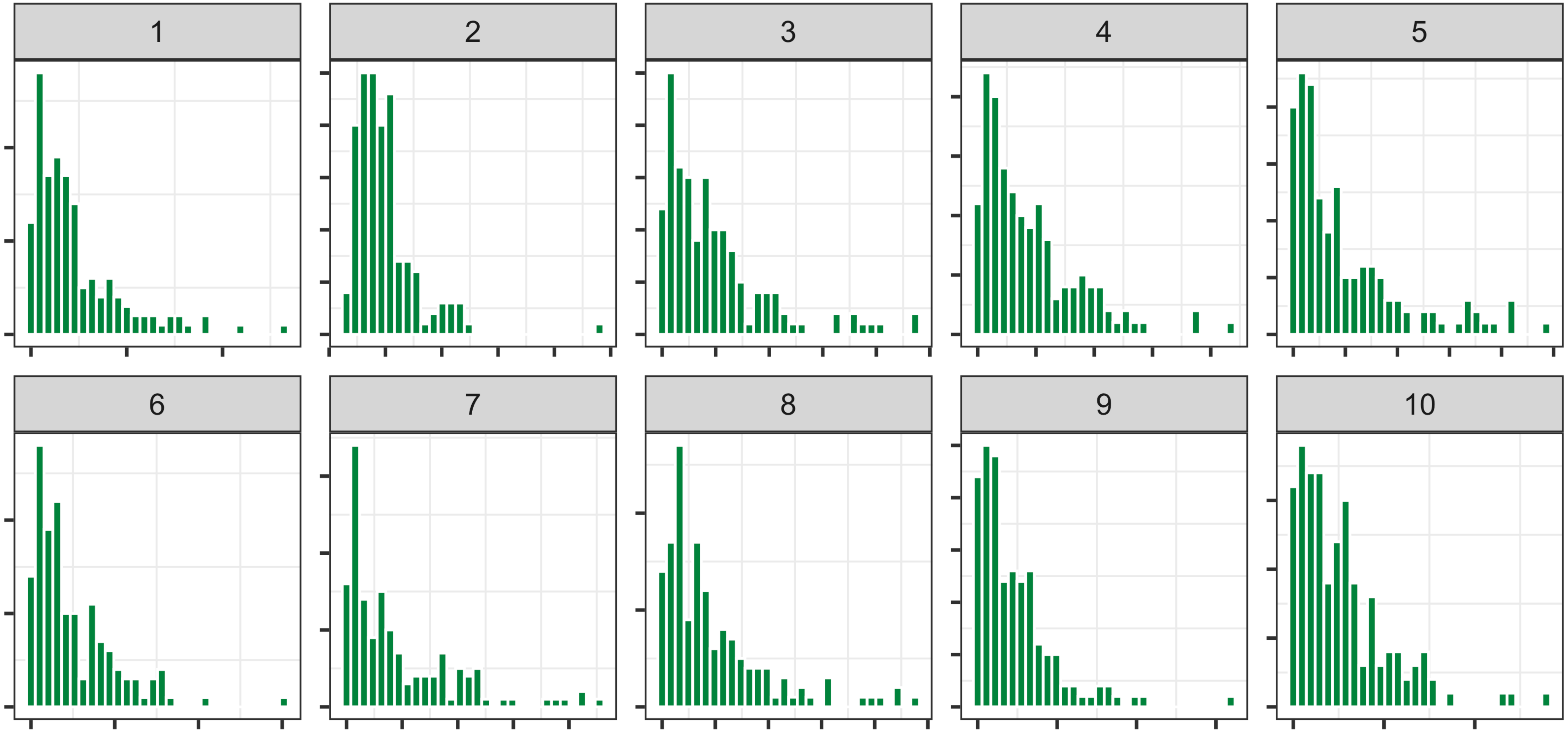


Is the outlying observation the electoral

Visual inference

Lineup of Greens first preference percentages

Explanation R



2019 Australian Federal Election (6/8)

data R

	% of first preference for the Greens					
State	Mean	Median	SD	IQR	Skewness	Kurtosis
ACT	16.4	14.0	5.6	5.20	0.65	1.5
VIC	11.4	8.6	8.2	6.72	2.60	11.4
WA	11.0	10.8	3.0	3.12	0.80	3.0
QLD	9.8	8.8	5.1	4.75	1.09	3.9
TAS	9.7	9.3	4.0	0.98	0.33	2.5
NT	9.6	9.6	2.5	1.75	0.00	1.0
SA	9.1	8.9	3.0	3.41	0.38	2.9
NSW	8.1	6.6	4.1	3.95	1.50	4.9
National	9.9	8.5	5.6	5.00	2.67	15.8

- 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

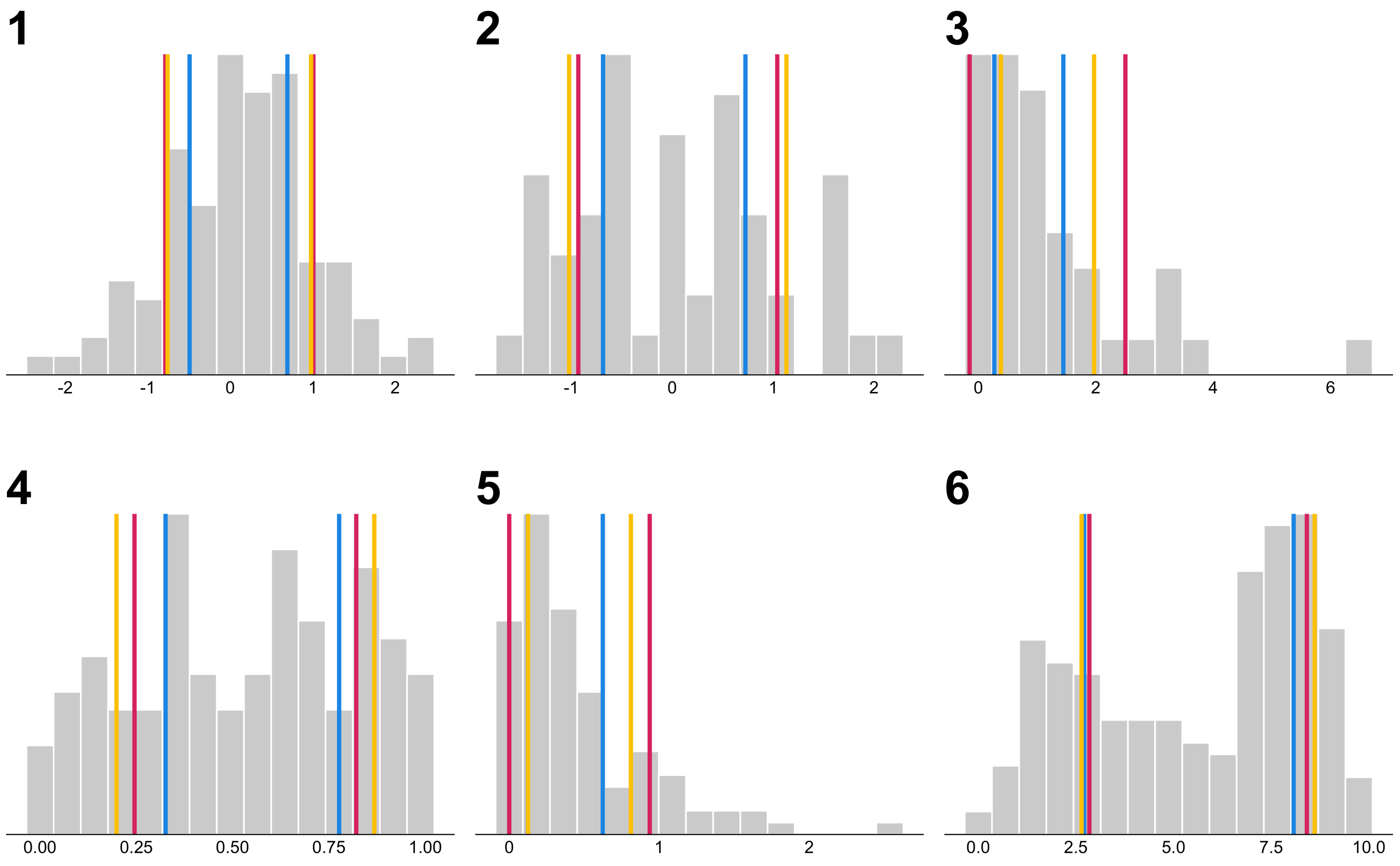
Robust measure of dispersion

- **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

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

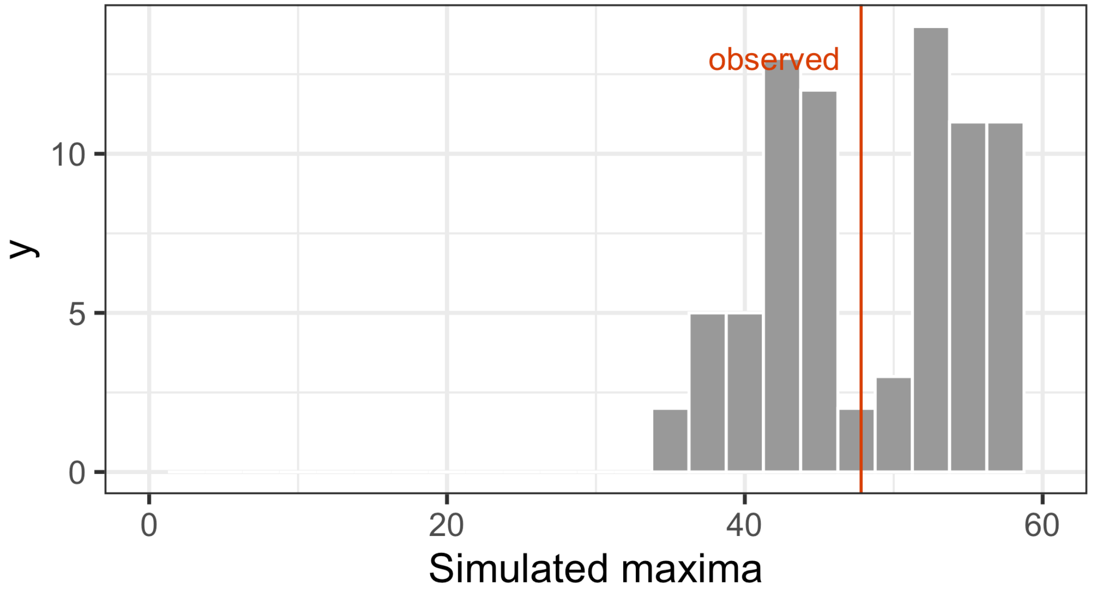


	Measure of dispersion				
Plot	SD	IQR	MAD	Skewness	Kurtosis
1	0.90	1.19	0.87	-0.072	3.0
2	0.99	1.41	1.08	0.358	2.2
3	1.33	1.18	0.79	1.944	7.2
4	0.29	0.45	0.34	-0.126	1.8
5	0.47	0.50	0.34	1.691	6.4
6	2.78	5.36	2.98	-0.351	1.7

Inference for robust statistics

We have seen the re-sampling methods simulation and permutation used for generating null plots in a lineup. Re-sampling methods can be used with numeric statistics also.

Simulation from distribution, can be used to to check for outliers.



We can also compute how many simulated values are more than the observed which gives a simulation p-value: 0.61.

For sample means, conventional tests provide a means for assessing what might be observed if different samples were taken.

Bootstrapping the current sample, can be used for robust statistics. If we have a sample of values:

```
[1] 2 2 3 6 7 7 8 8
```

to bootstrap sample with replacement:

```
sort(sample(x, replace=TRUE))
```

```
[1] 2 2 3 3 7 7 7 7
```

```
sort(sample(x, replace=TRUE))
```

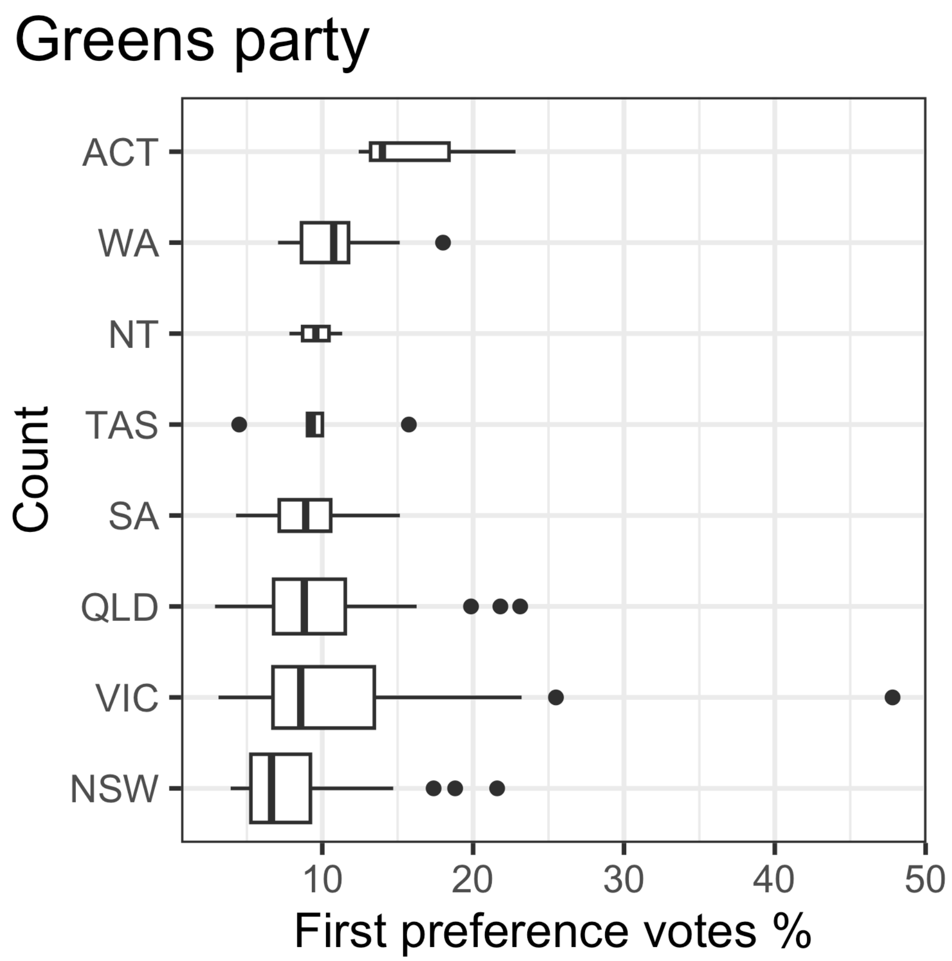
```
[1] 2 3 6 6 6 6 8 8
```

Here's an example of bootstrapping to get a confidence interval for a median.

```
[1] "Median: 6.34"
[1] "95% CI: ( 4.99 , 9.16 )"
```

2019 Australian Federal Election (7/8)

data R



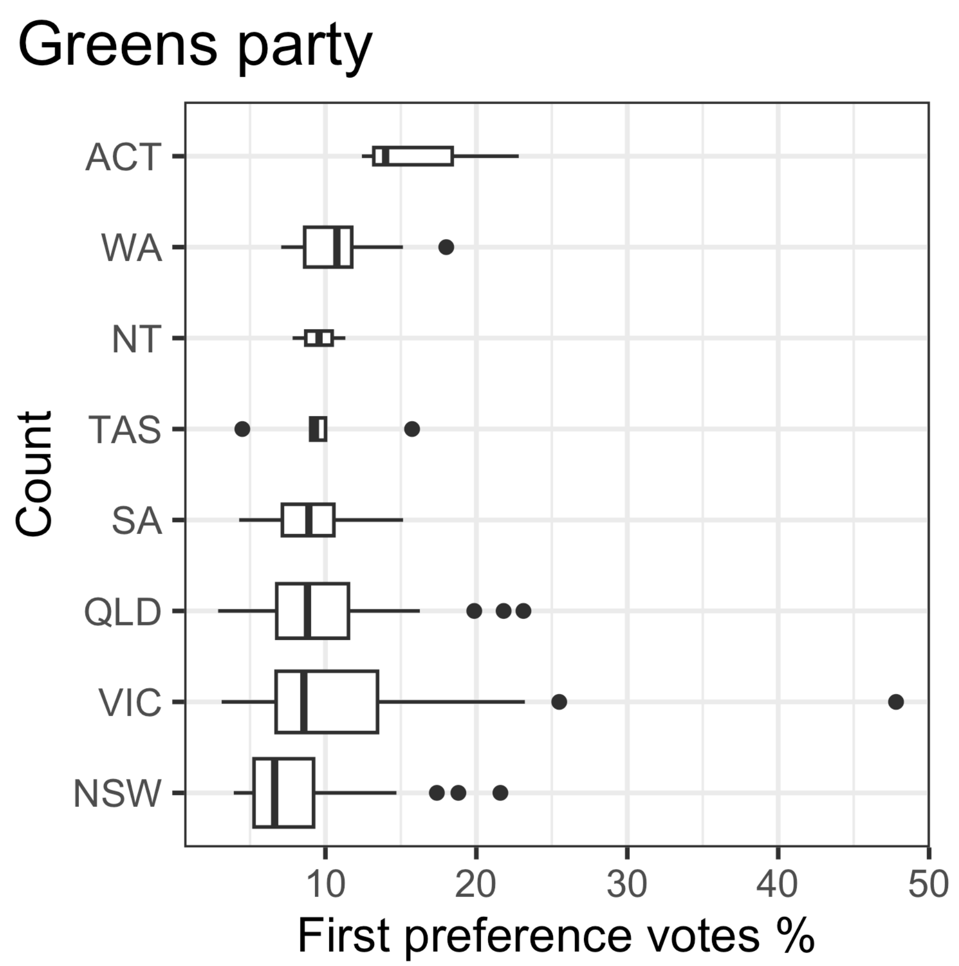
Where are these electorates?

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

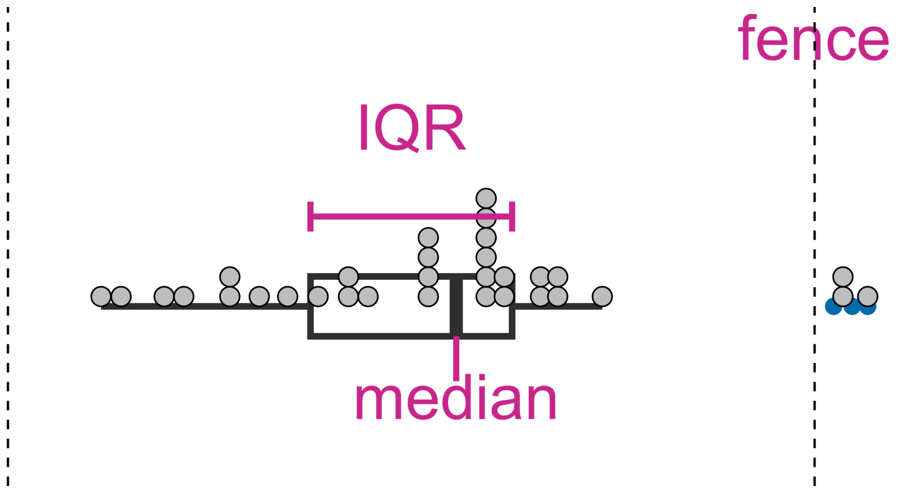
Outliers

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*

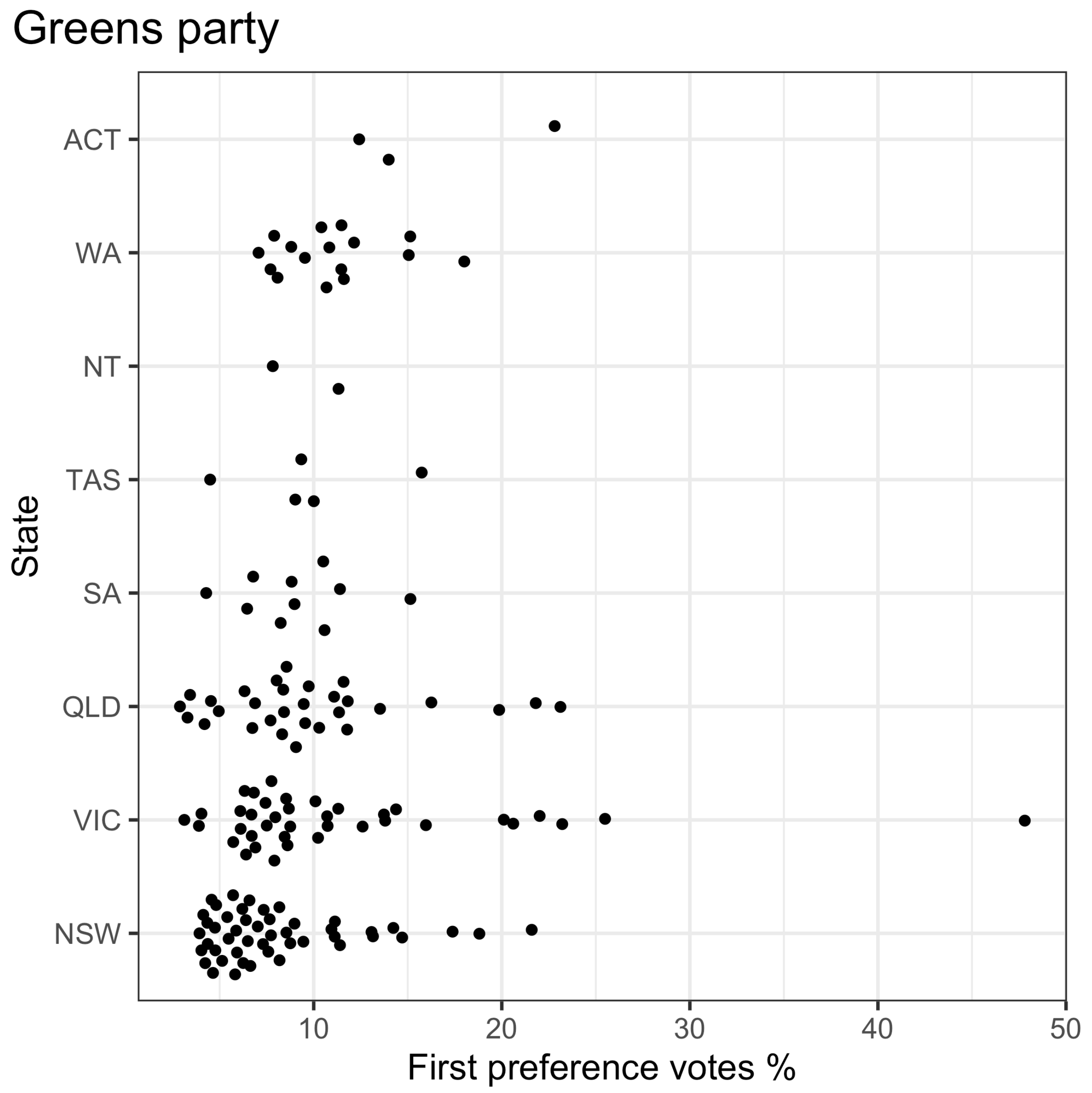


- Observations that are outside the range of lower to upper fence (1.5 times the box length) are often referred to as outliers.
- Plotting boxplots for data from a skewed distribution will almost always show these "outliers" but these are not necessarily outliers.
- Some definitions of outliers assume 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.
- Declaring observations outliers typically requires additional data context.

What *cannot* be seen from boxplots?

2019 Australian Federal Election (8/8)

📊 data R



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

- Only two electoral districts in NT.
- And only 3 and 5 electoral districts in ACT and TAS, respectively!
- Boxplots requires 5 points!
- We should have summarised the number of electoral districts for each state with numerical statistics as a first step.
- Also the outlier (yes, safe to call this an outlier!) and the cluster in the Victoria electorates.

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

Transformations

Melbourne Housing Prices (1/6)

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
Armadale	3	Home	1,386,000	2017-04-01

- This data was scraped 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)

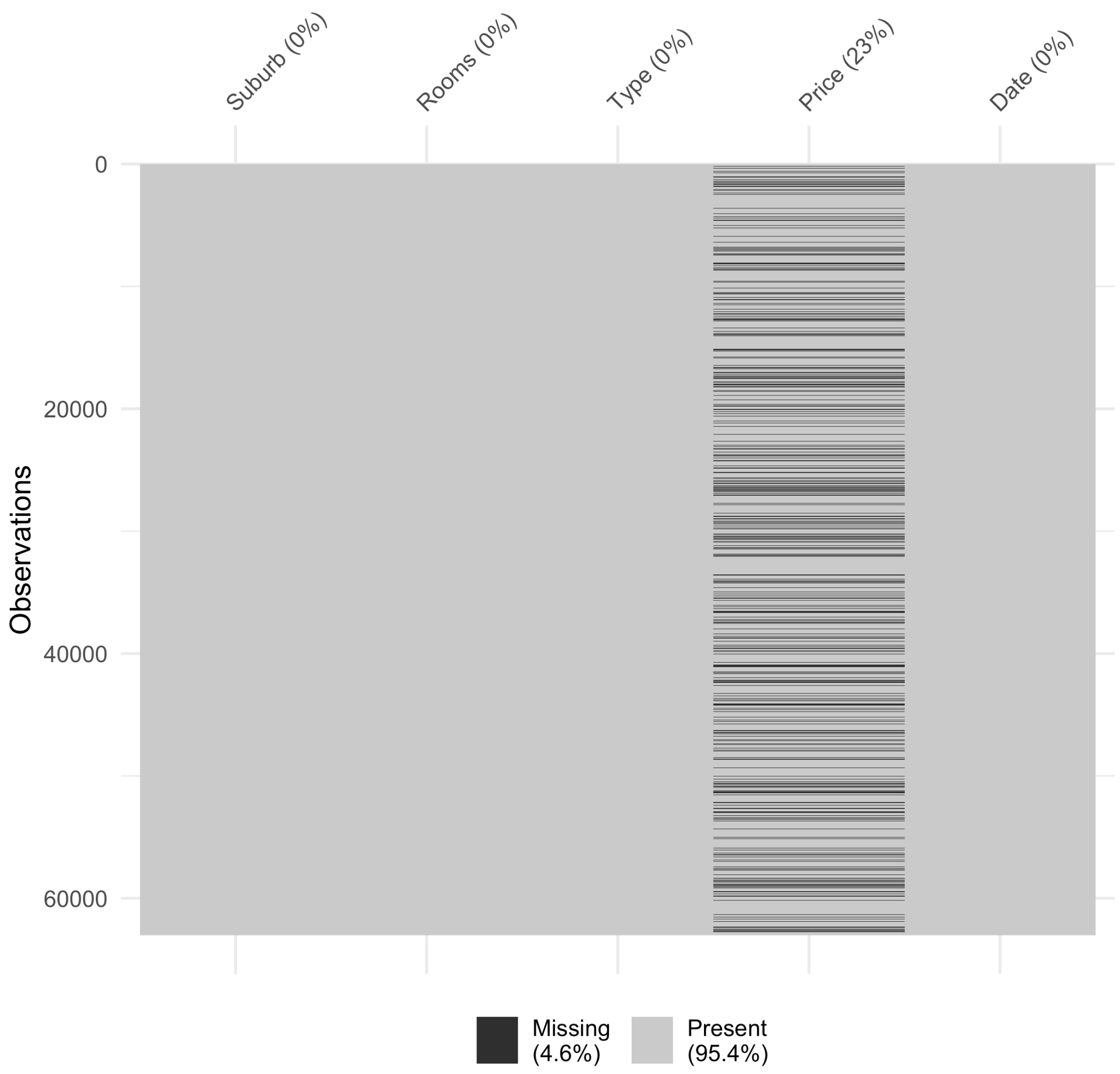
Data source: Tony Pio (2018) Melbourne Housing Market

How would you explore this data first?

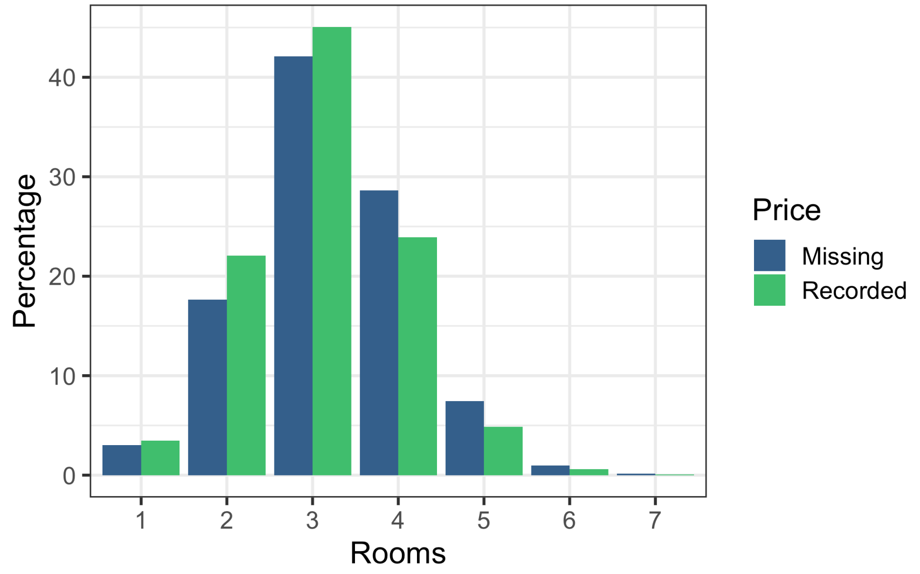
Yes, with an overview plot.

Melbourne Housing Prices (2/6)

data R lineup R



Is missingness more likely for expensive houses?



- Check with a lineup
- To impute missings other variables will need to be used.

Note: Houses with more than 8 rooms removed. Why?

🧩 Your turn

What might be alternative plots? Especially to reveal the relationship more clearly.

Check the support of your data

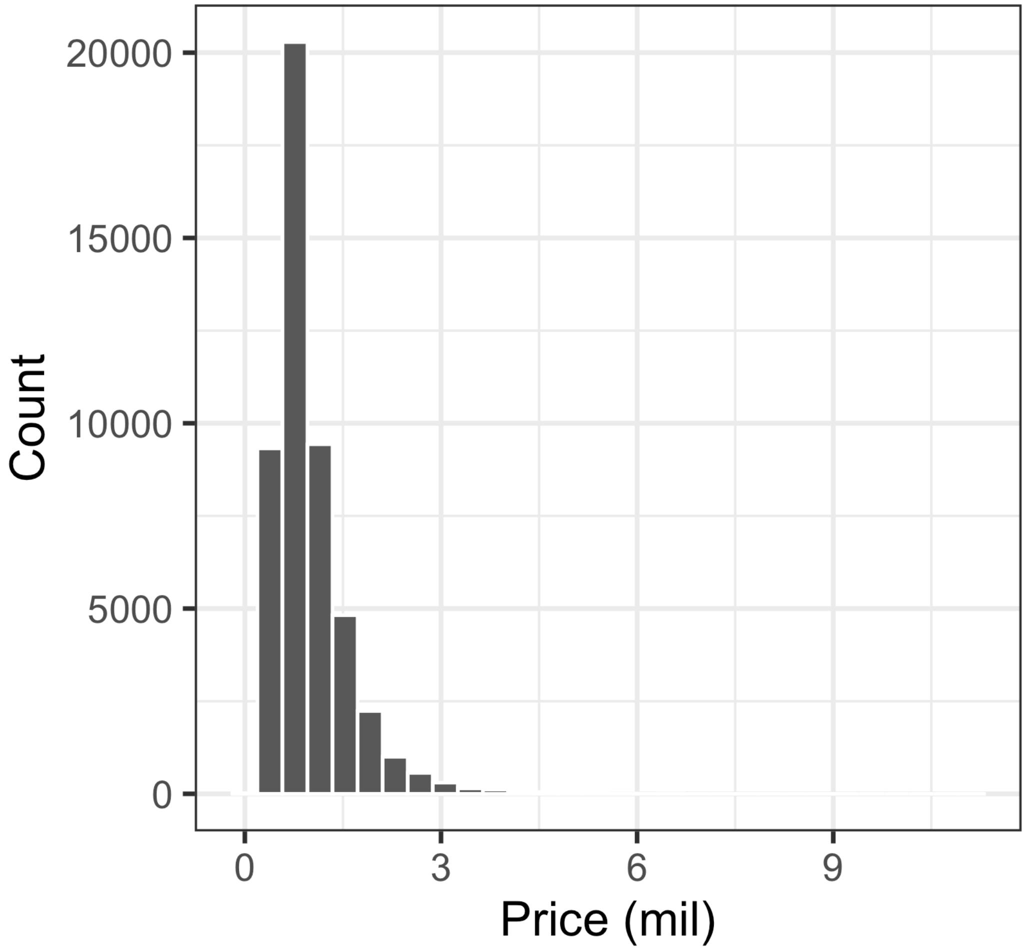
> If you have too few measurements in any region (extreme), summaries for these regions will be unreliable.

- For quantitative variables, it may be necessary to remove extremes.
- If the variable is categorical it might be best to combine levels.
- It is important to script so decisions can be reversed or rare events are not ignored.

We removed houses with 8 or more rooms. What other way might we have handled these houses?

Melbourne Housing Prices (3/6)

data R



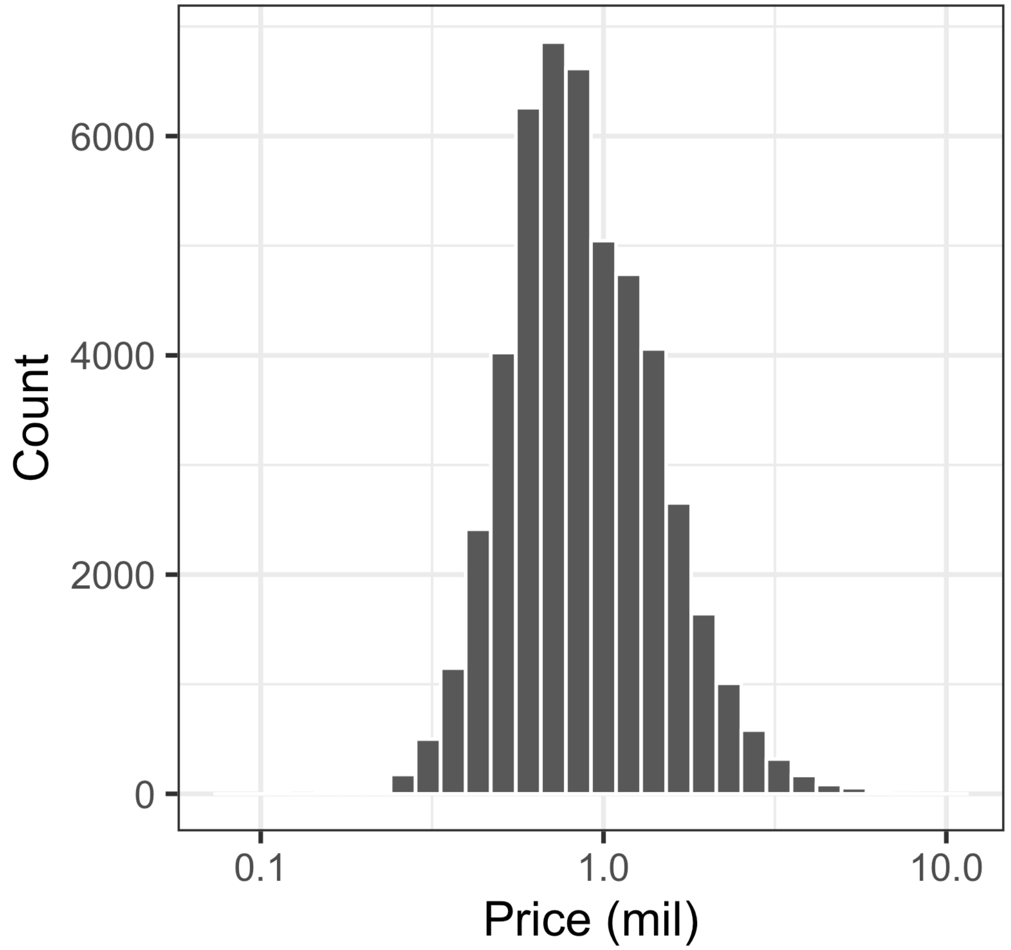
What can we say from this plot?

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

Note: We determined that it is likely that more higher price houses have not disclosed the sale price. The distribution of price will need to be checked again after imputation.

Melbourne Housing Prices (4/6)

data R



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

Melbourne Housing Prices (5/6)

Central tendency | R

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):

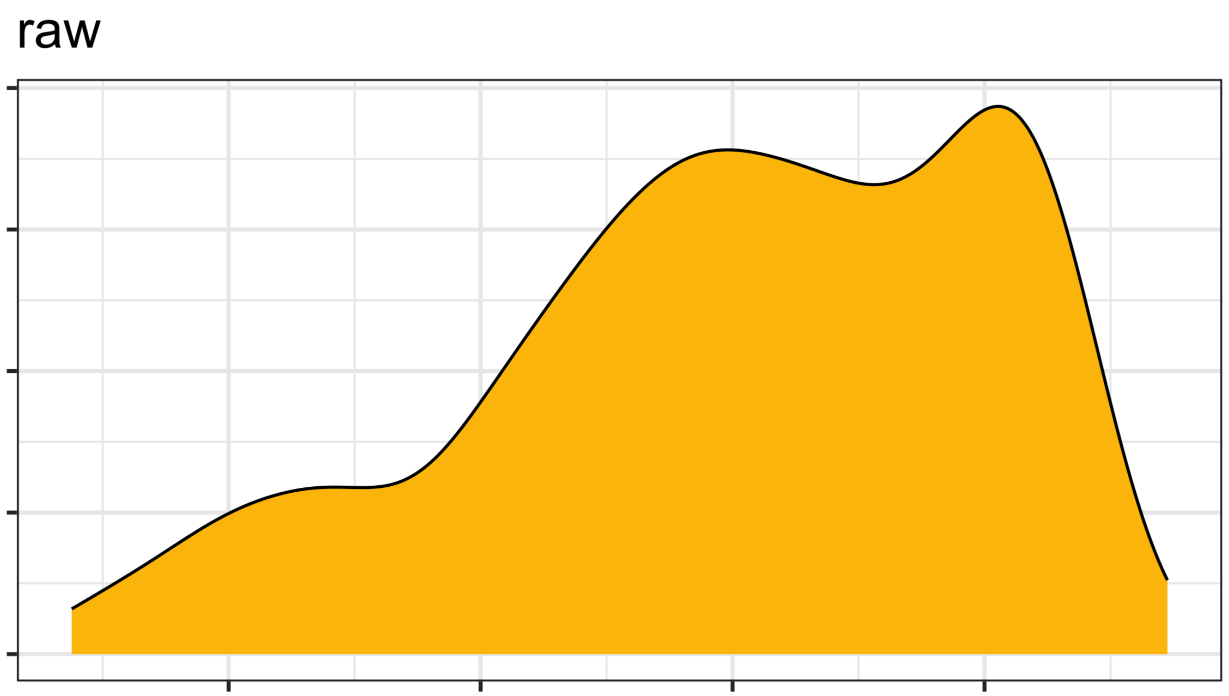
Mean	Median	Trimmed Mean	Winsorised Mean
$874,166	$830,000	$847,973	$859,325

General rules for transform quantitative variables

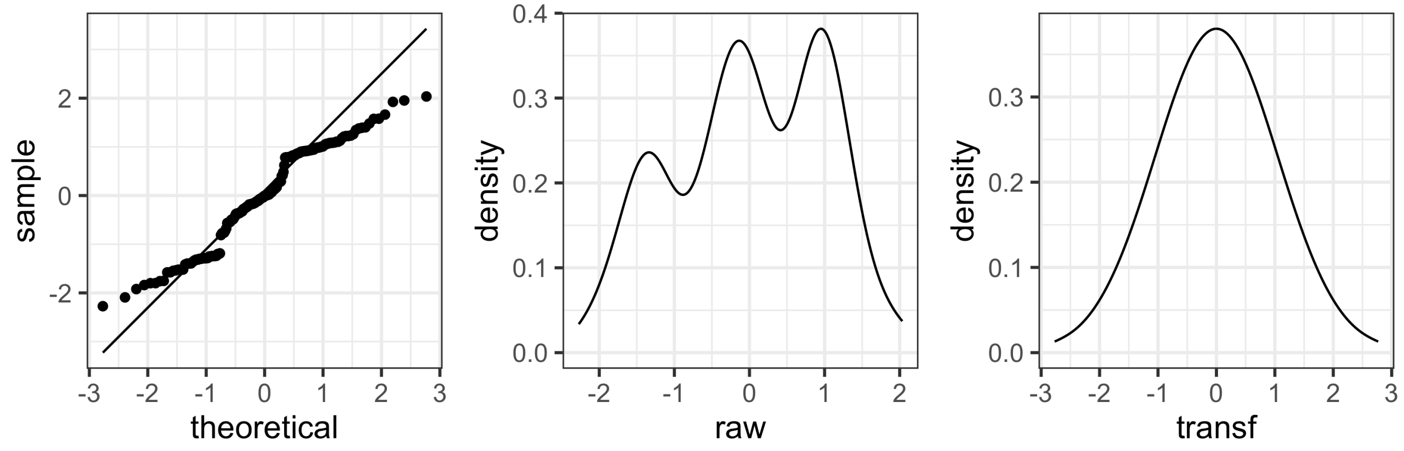
Non-shape changing, scaling:

- standardise to mean 0, sd 1
- standardise to min 0, max 1
- z-score

Shape changing, transformations: Remember the ladder of power transformations. (eg transforming left-skewed to more uniform using $)_{2}$



Distribution changing: quantile



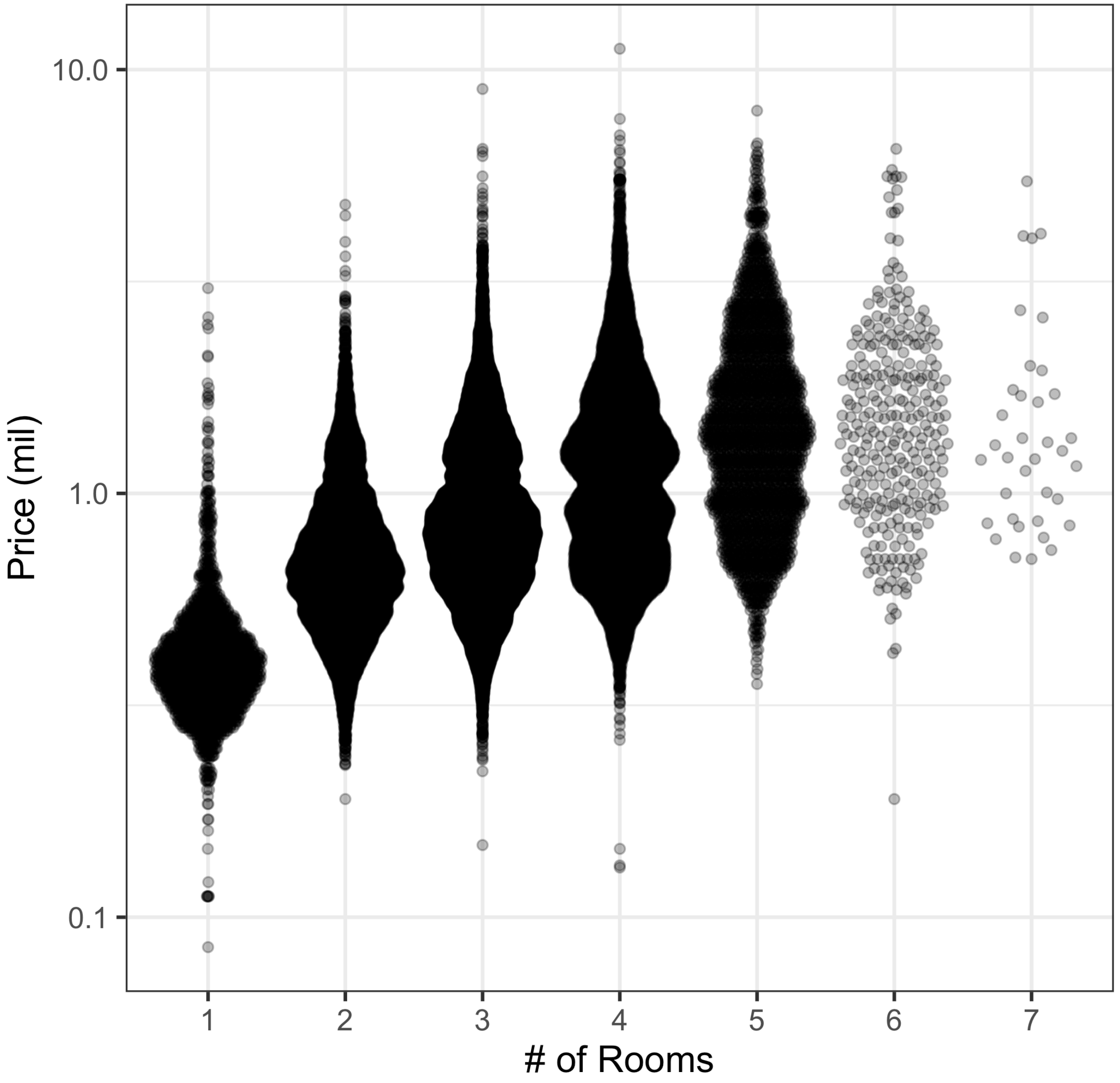
Some features **cannot be fixed**: *gaps, multimodality, heaping*. You need to find some explaining variable.

Some features can be **artificially fixed**: discreteness. If regularly discretized, add random uniform noise to spread equally between gaps.

Multi-modality

Melbourne Housing Prices (6/6)

data R

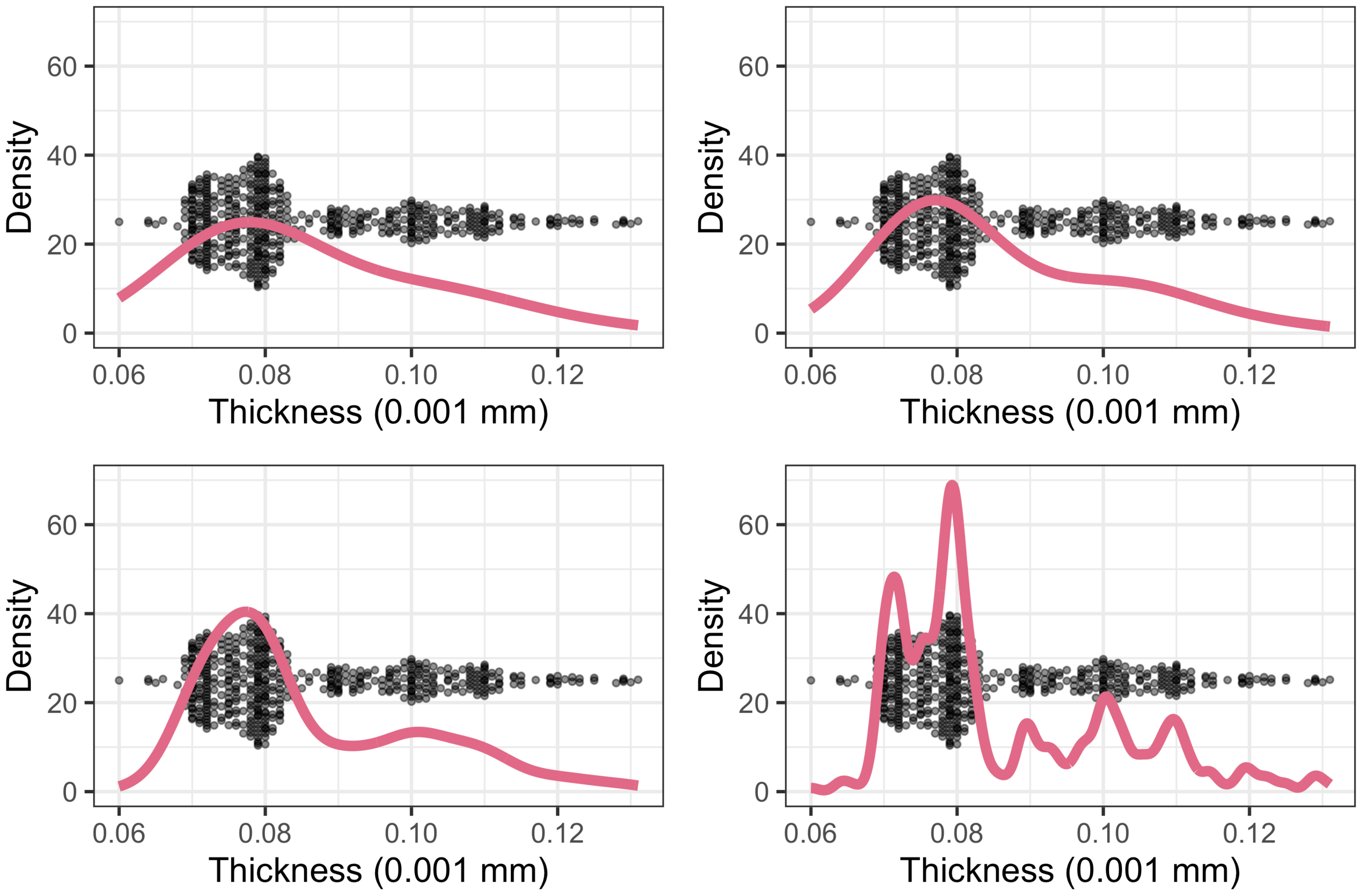


- Distribution from side-by-side univariate plots shows that higher number of rooms generally are pricier.
- This strata could be responsible for multimodality in price distribution, even though it is not visible in the histogram.
- Accounting for rooms is important.

Bins and Bandwidths: More details

Hidalgo stamps thickness

📊 data R

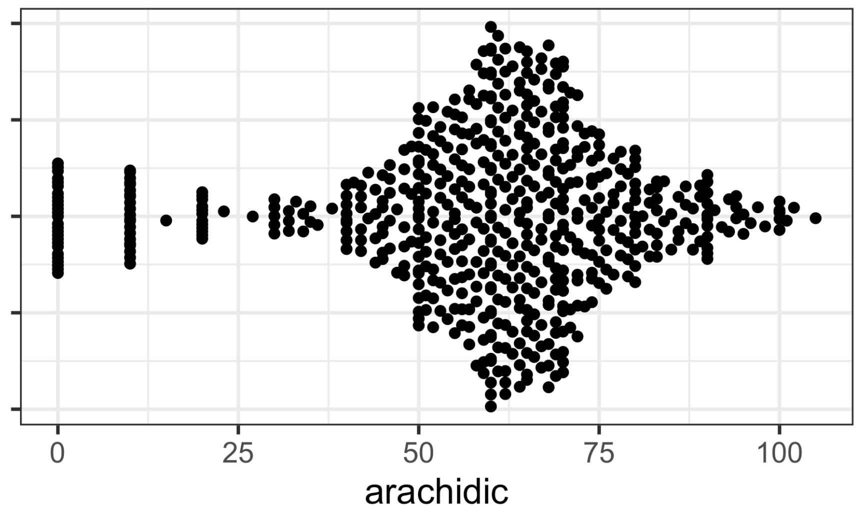


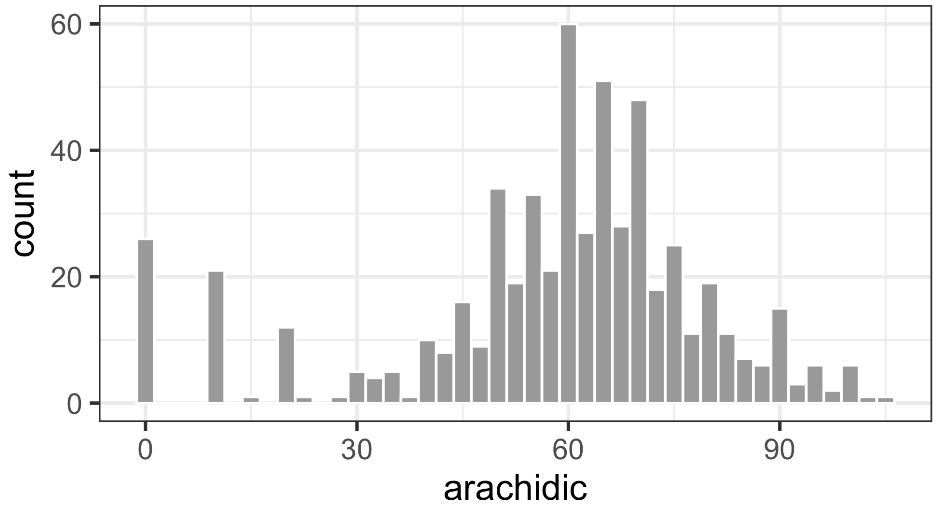
Famous historical example

- 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 suggests different possibilities for number of modes.

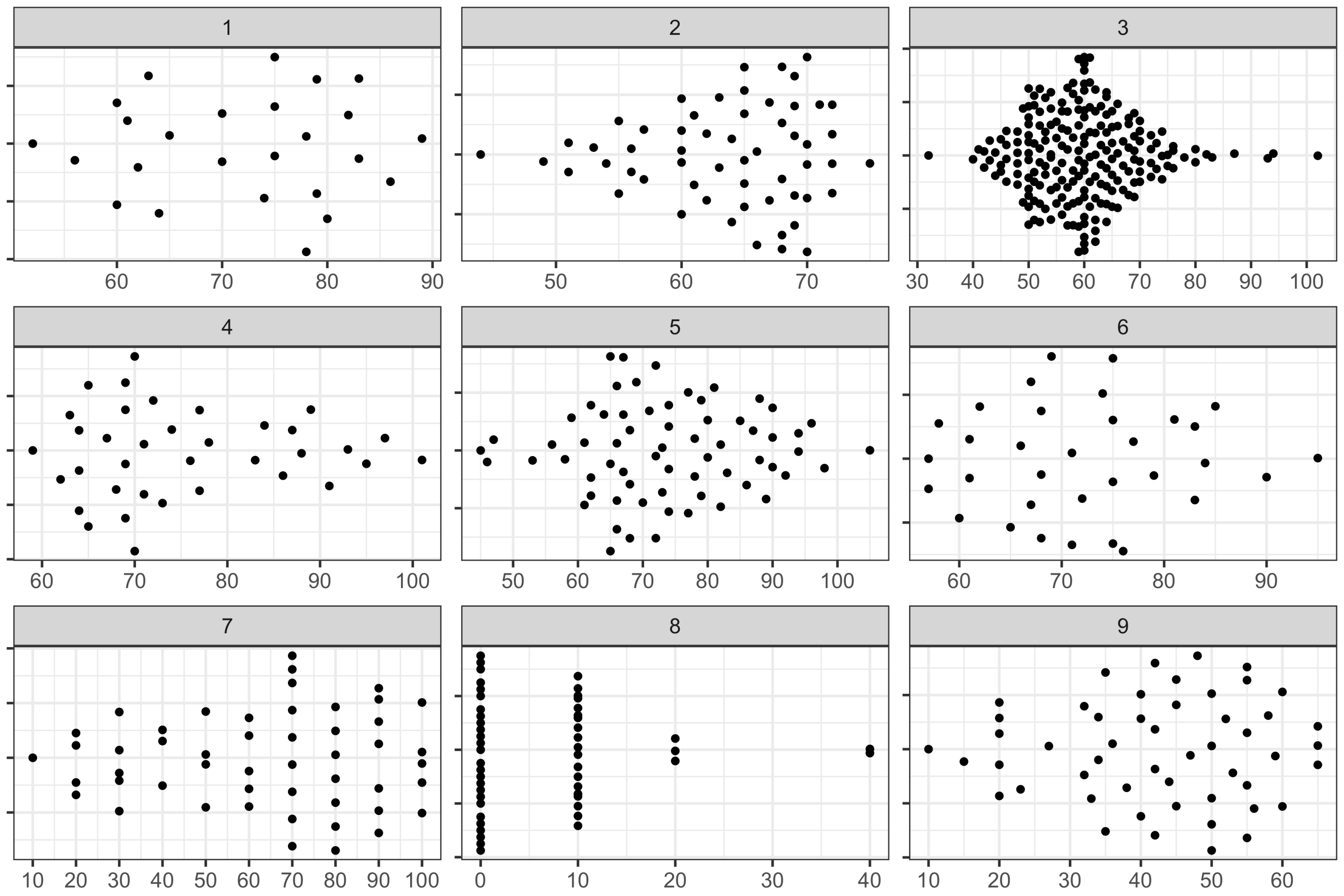
Which do you think most accurately reflects what's in the data?

Olive oil content





Check if there is a difference in the strata (here 1 thru 9), implying measurement policy differences.



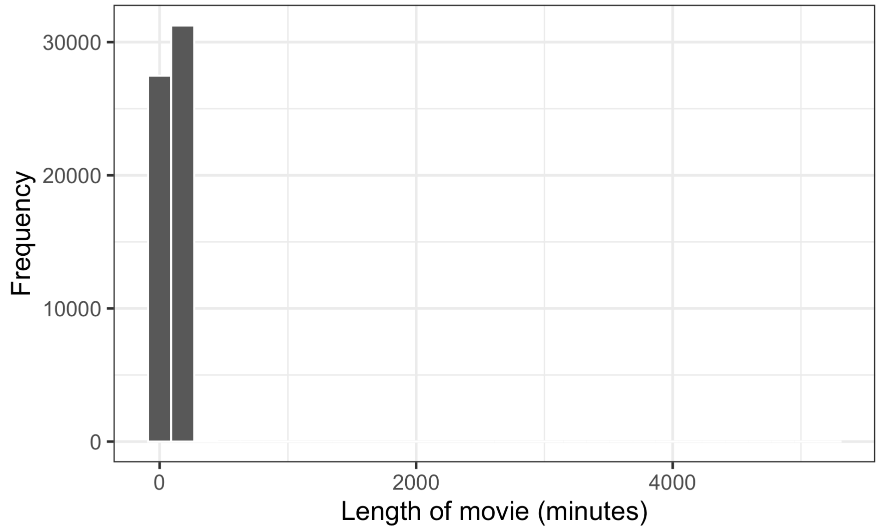
What do you see?

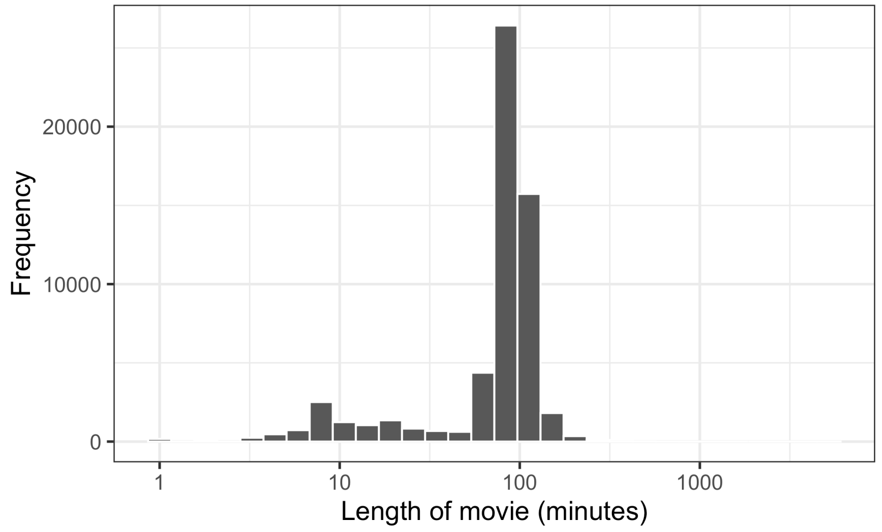
Mixture of discreteness and normal shape of continuous values. *Why might this happen?*

Re-focus

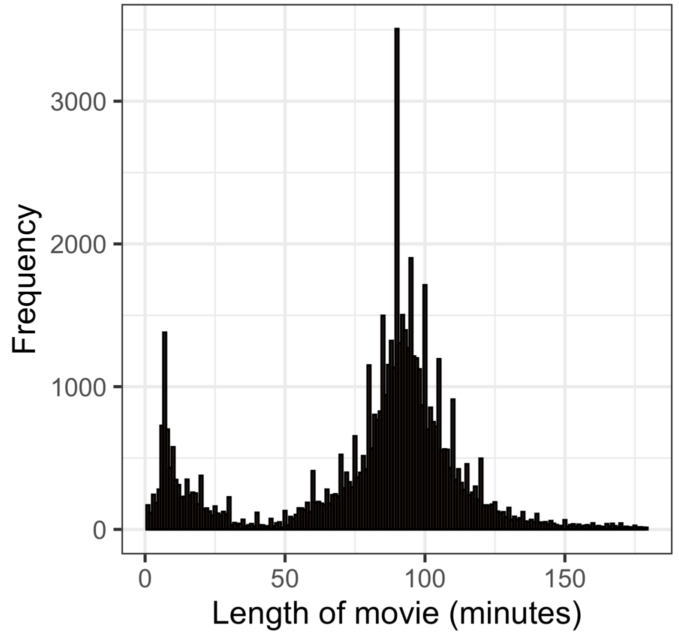
Movie length

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?

Categorical variables

There are two types of categorical variables

Nominal where there is no intrinsic ordering to the categories
E.g. blue, grey, black, white.

Ordinal where there is a clear order to the categories.
E.g. Strongly disagree, disagree, neutral, agree, strongly agree.

Categorical variables in R

- In R, categorical variables may be encoded as **factors**.

```r
data <- c(2, 2, 1, 1, 3, 3, 3, 1)
factor(data)
```

```
[1] 2 2 1 1 3 3 3 1
Levels: 1 2 3
```

- You can easily change the labels of the variables:

```r
factor(data, labels = c("I", "II", "III"))
```

```
[1] II  II  I   I   III III III I
Levels: I II III
```

- Order of the factors are determined by the input:

```r
# numerical input are ordered in increasing or
factor(c(1, 3, 10))
```

```
[1] 1  3  10
Levels: 1 3 10
```

```r
# character input are ordered by first char, a
factor(c("1", "3", "10"))
```

```
[1] 1  3  10
Levels: 1 10 3
```

```r
# you can specify order of levels explicitly
factor(c("1", "3", "10"),
  levels = c("1", "3", "10")
)
```

```
[1] 1  3  10
Levels: 1 3 10
```

Numerical summaries: counts, proportions, percentages and odds

Tuberculosis counts in Australia

```
# A tibble: 22 × 7
   country   iso3   year count      p   pct  odds
   <chr>     <chr> <dbl> <dbl>  <dbl> <dbl> <dbl>
 1 Australia AUS    2000   982 0.0522  5.22 1
 2 Australia AUS    2001   953 0.0507  5.07 0.970
 3 Australia AUS    2002  1008 0.0536  5.36 1.03
 4 Australia AUS    2003   926 0.0493  4.93 0.943
 5 Australia AUS    2004  1036 0.0551  5.51 1.05
 6 Australia AUS    2005  1030 0.0548  5.48 1.05
 7 Australia AUS    2006  1127 0.0600  6.00 1.15
 8 Australia AUS    2007  1081 0.0575  5.75 1.10
 9 Australia AUS    2008  1182 0.0629  6.29 1.20
10 Australia AUS    2009  1176 0.0626  6.26 1.20
11 Australia AUS    2010  1146 0.0610  6.10 1.17
12 Australia AUS    2011  1202 0.0640  6.40 1.22
```

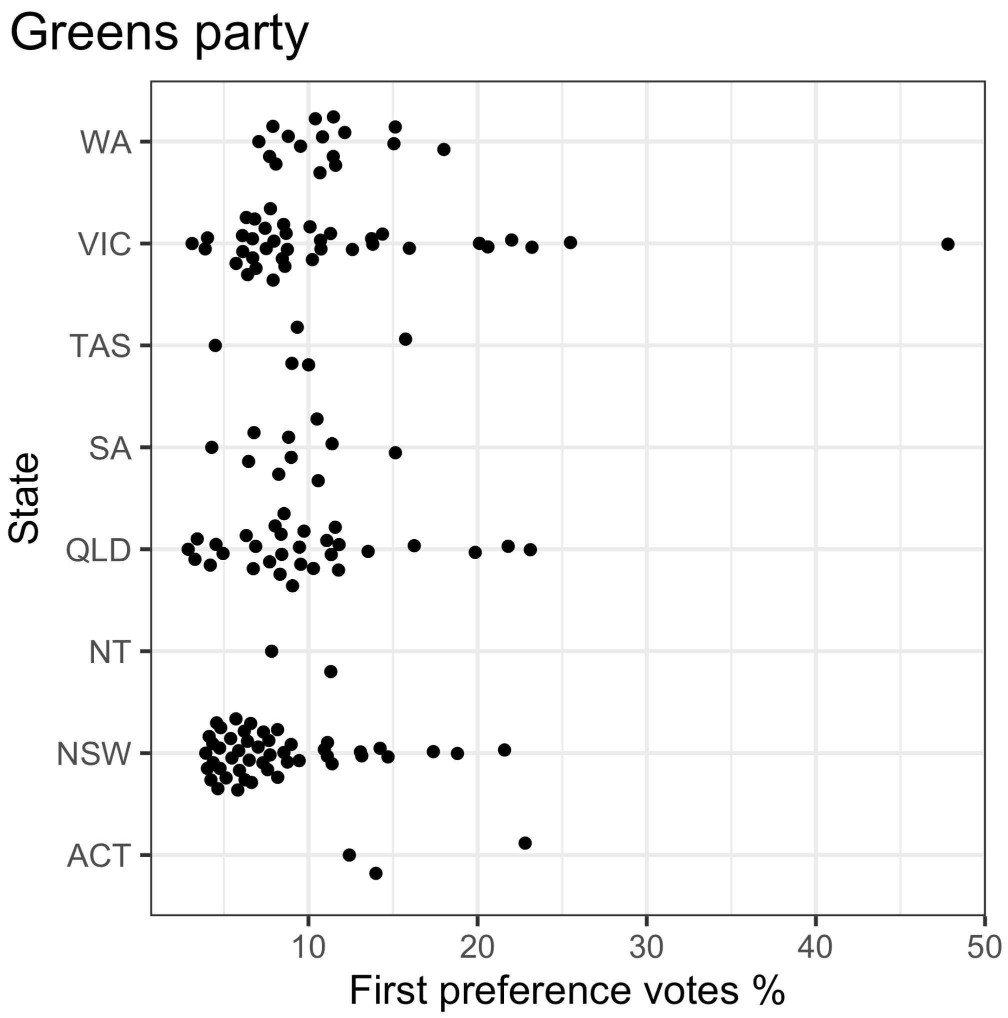
For qualitative data, compute

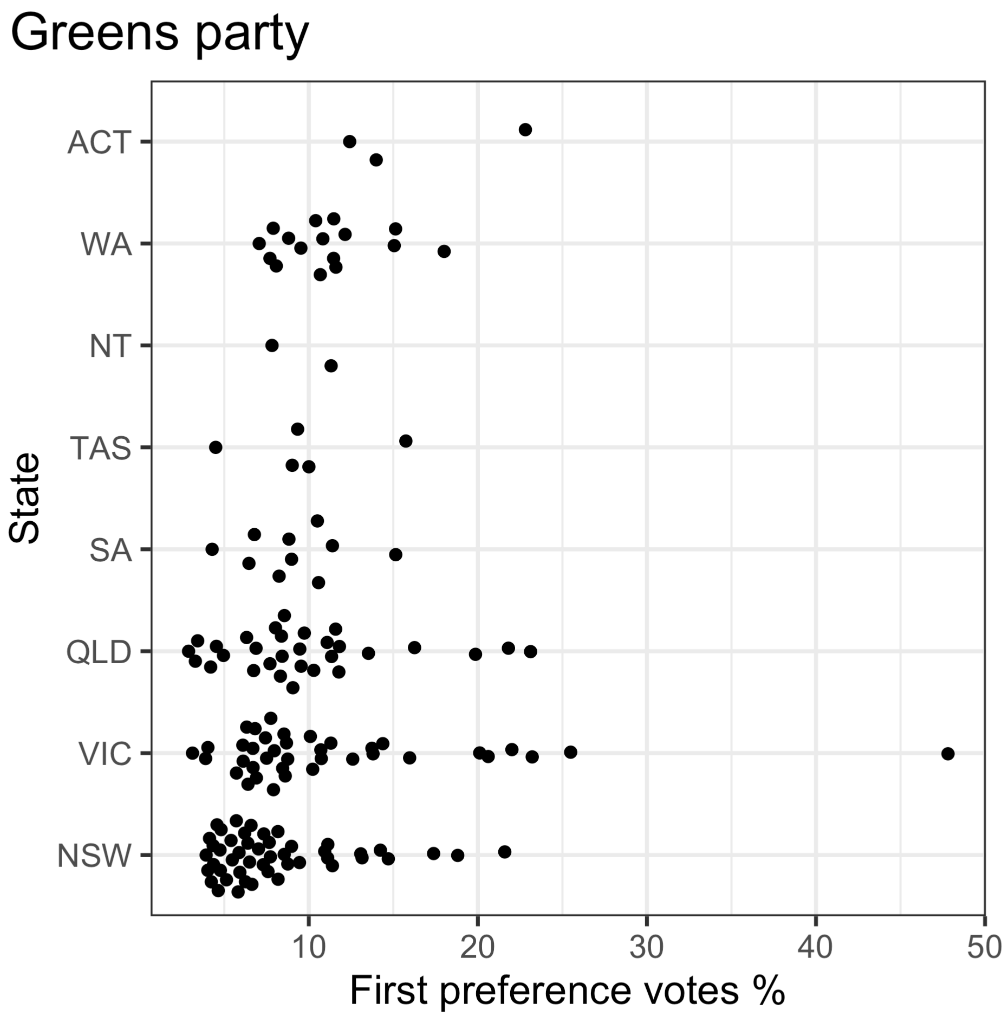
- count/frequency,
- proportion/percentage
- and sometimes, an **odds ratio**. Here we have used ratio relative to the count in year 2000.

Note: For exploration, no rounding of digits was done, but to report you would need to make the numbers pretty.

2019 Australian Federal Election

 data R





Sorting levels sensibly is (almost) always better when plotting

Order nominal variables meaningfully

Coding tip: use below functions to easily change the order of factor levels

```
stats::reorder(factor, value, mean)
forcats::fct_reorder(factor, value, median)
forcats::fct_reorder2(factor, value1, value2, func)
```

Visual inference

Typical plot description:

```r
ggplot(data, aes(x=var1)) +
  geom_col()

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

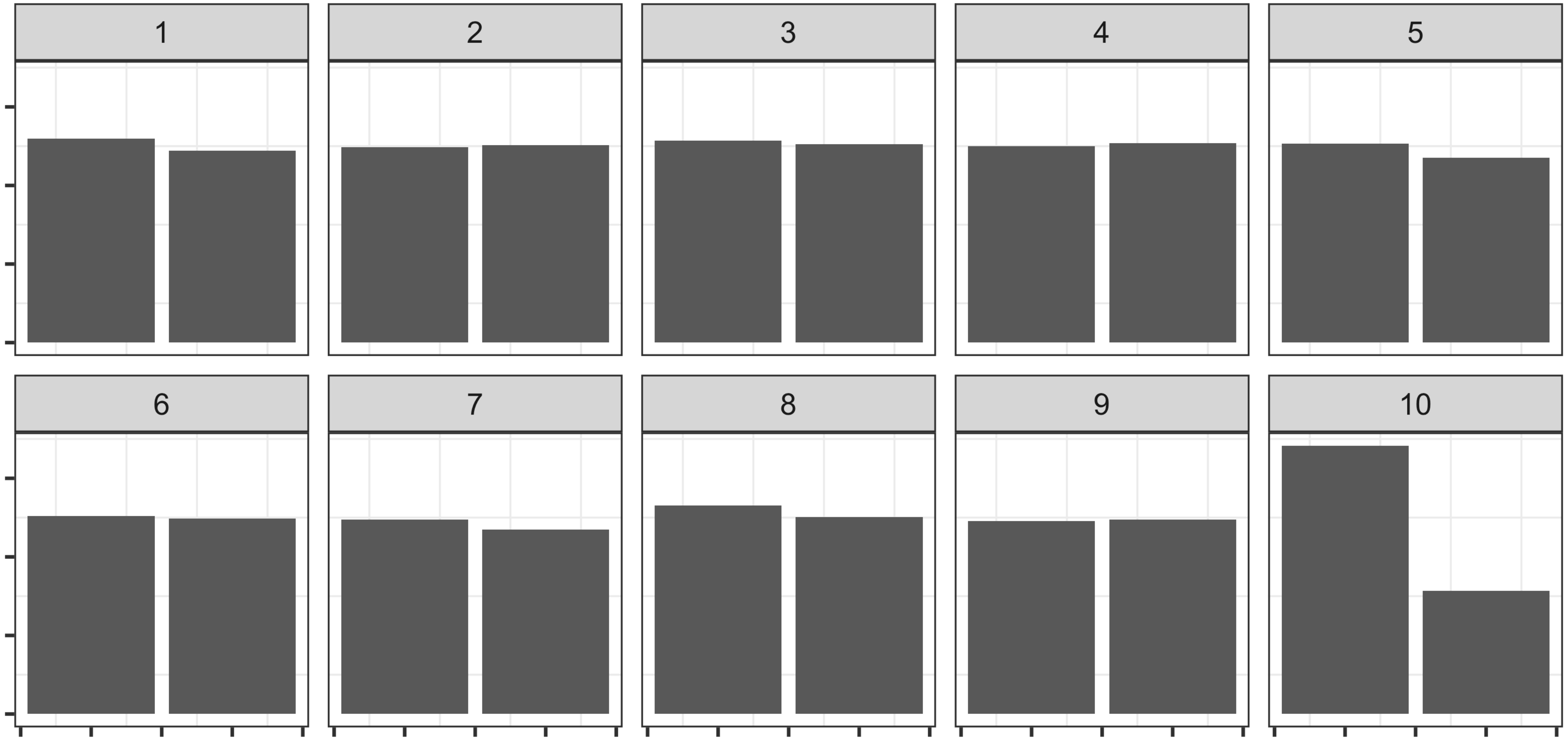
Potential simulation method from binomial

```r
# Only one option
null_dist("var1", "binom",
  list(size=n, p=phat))
```

Is the distribution consistent with a sample from a binomial distribution with a given p?

Lineup of tuberculosis count between sexes

Use conventional test R



Key points

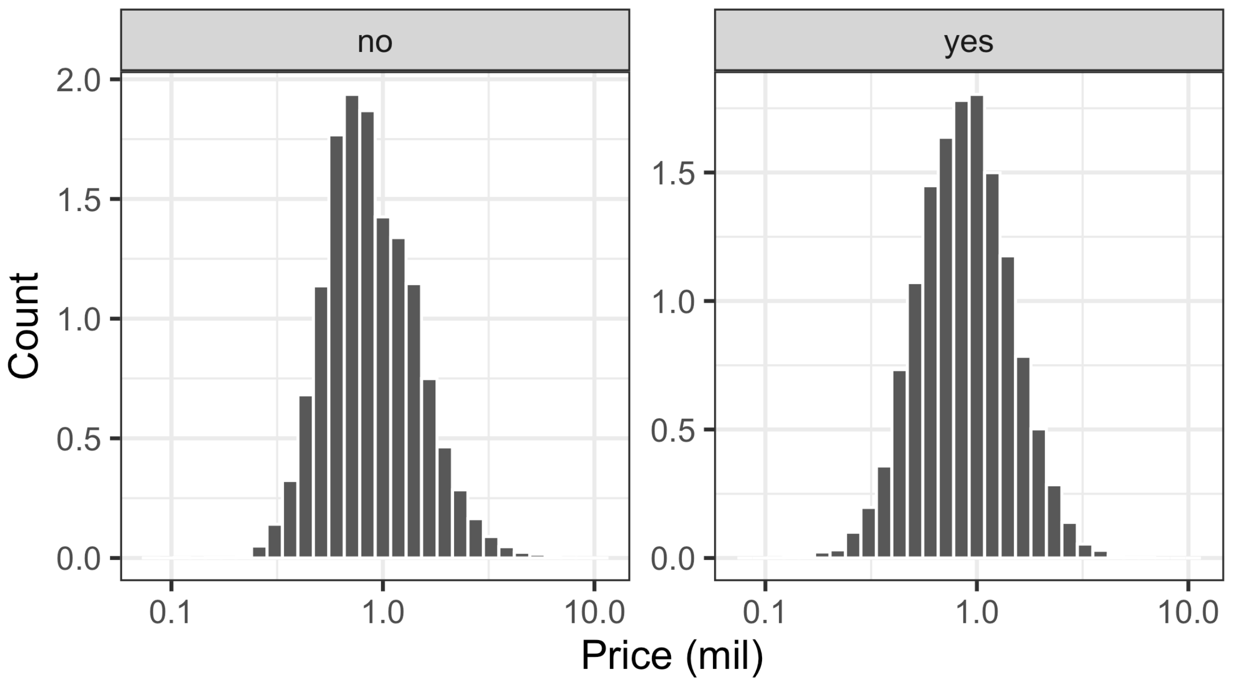
- Be prepared to do multiple plots
- Changing bins or binwidth/bandwidth in histogram, violin or density plots can paint a different picture
- Consider different representations of categorical variables
  - reordering meaningfully,
  - lumping low frequencies together,
  - plot or table, pie or barplot,
  - include a missing category

Imputing missings for univariate distributions

Quantitative variable: Simulate from a fitted distribution.

```
df2 <- df2 |>
  mutate(lPrice = log10(Price),
         price_miss = ifelse(is.na(Price), "yes", "no"))

df2_smry <- df2 |>
  summarise(m = mean(lPrice, na.rm=TRUE),
            s = sd(lPrice, na.rm=TRUE))
set.seed(1003)
df2 <- df2 |>
  rowwise() |>
  mutate(lPrice = ifelse(price_miss == "yes",
    rnorm(1, df2_smry$m, df2_smry$s), lPrice)) |>
  mutate(Price = ifelse(price_miss == "yes", 10^lPrice, Price
```



Categorical variable: Simulate from multinomial.

```
# A tibble: 7 × 2
  age   count
  <chr> <dbl>
1 1524     27
2 2534     48
3 3544     15
4 4554     11
5 5564      9
6 65       15
7 u        12

 [1] 5 3 1 1 2 1 1 2 1 1 1 1

# A tibble: 7 × 2
  age   count
  <chr> <dbl>
1 1524     35
2 2534     50
3 3544     16
4 4554     11
5 5564     10
6 65       15
7 u        12
```

`imputeMulti` library can automate for multiple variables.

Resources

- Unwin (2015) Graphical Data Analysis with R
- 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
- British Board of Trade (1990), Report on the Loss of the 'Titanic' (S.S.). British Board of Trade Inquiry Report (reprint). Gloucester, UK: Allan Sutton Publishing
- Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) A Handbook of Small Data Sets. Chapman & Hall, Data set 285 (p. 229)
  Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0
- Fleiss JL (1993): The statistical basis of meta-analysis. *Statistical Methods in Medical Research* **2** 121–145
  Balduzzi S, Rücker G, Schwarzer G (2019), How to perform a meta-analysis with R: a practical tutorial, Evidence-Based Mental Health.
- Josse et al (2022) R-miss-tastic, https://rmisstastic.netlify.app