MONASH University

ETC5521: Exploratory Data Analysis

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

Lecturer: *Emi Tanaka*

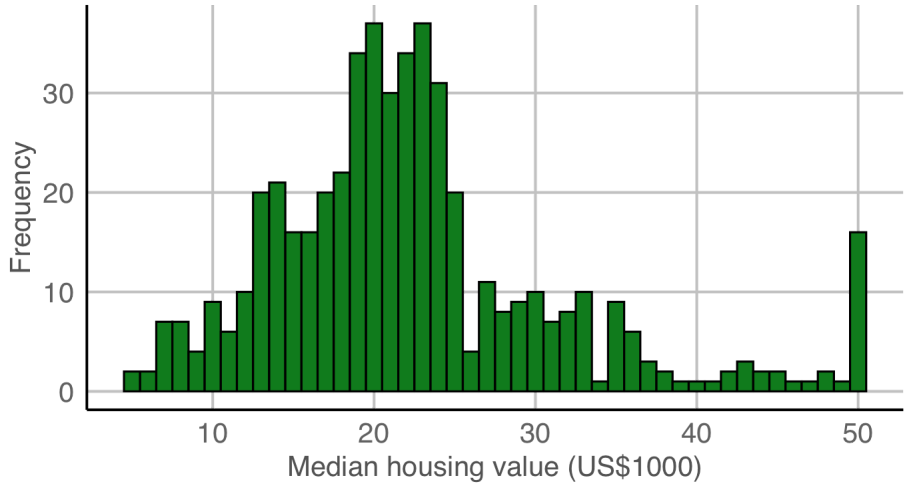
ETC5521.Clayton-x@monash.edu

Week 4 - Session 2

Bins and Bandwidths

Case study 3 Boston housing data Part 1/4

data R

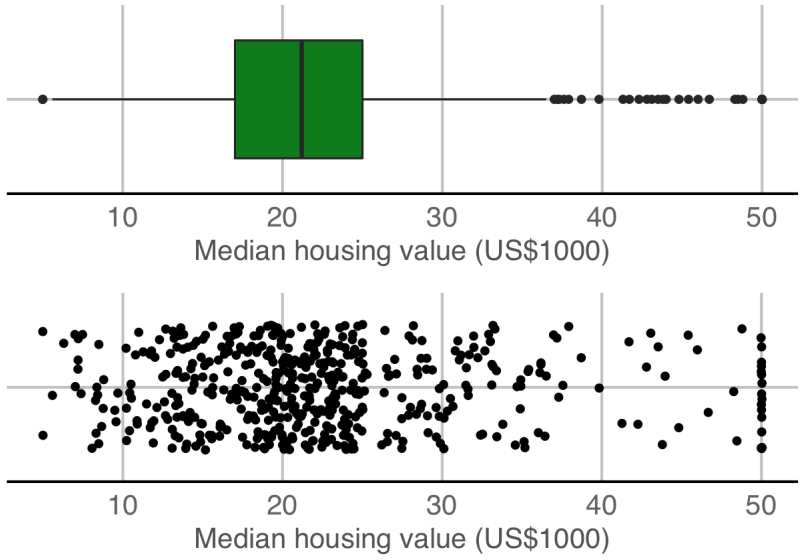


- Thre is a large frequency in the final bin.
- There is a decline in observations in the $40-49K range as well as dip in observations around $26K and $34K.
- The histogram is using a bin width of 1 unit and is **left-open** (or **right-closed**): (4.5, 5.5], (5.5, 6.5] ... (49.5, 50.5].
- Occasionally, whether it is left- or right-open can make a difference.

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

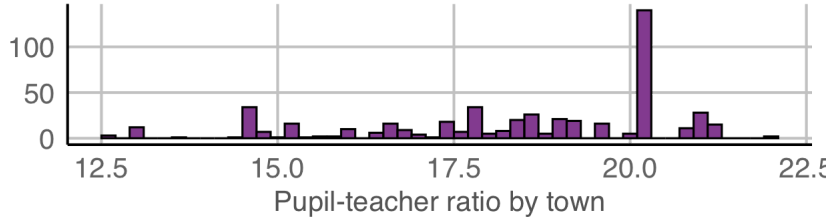


- Density plots depend on the bandwidth chosen and more than often do not estimate well at boundary cases
- There are various way to present features of the data using a plot and what works for one person, may not be as straightforward for another
- Be prepared to do multiple plots!

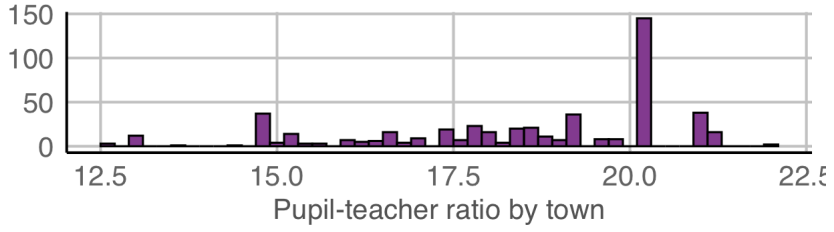
Case study 3 Boston housing data Part 3/4

data R

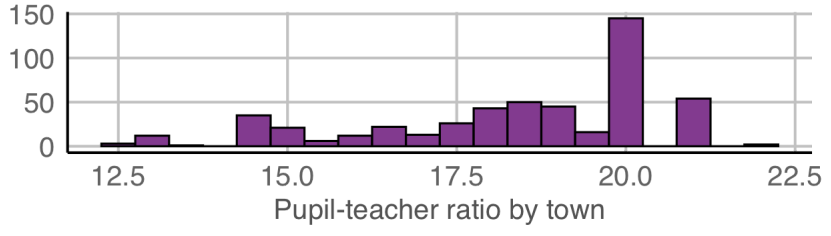
Bin width = 0.2, Left-open



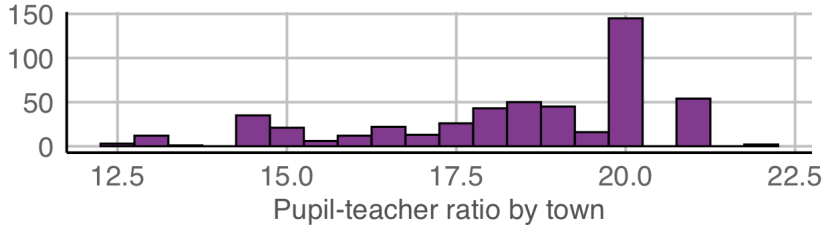
Bin width = 0.2, Right-open



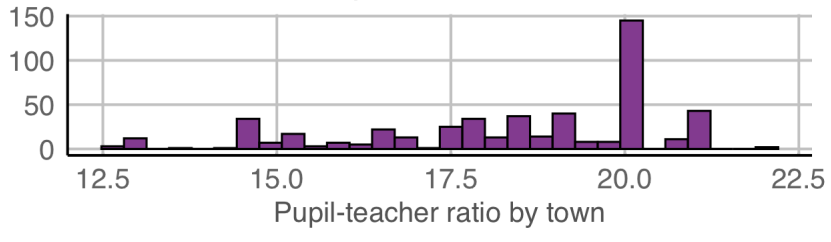
Bin width = 0.5, Left-open



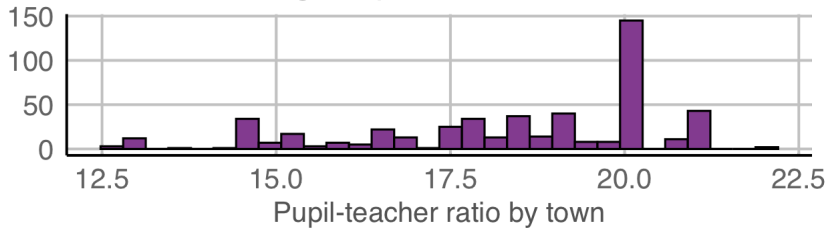
Bin width = 0.5, Right-open



Bin number = 30, Left-open

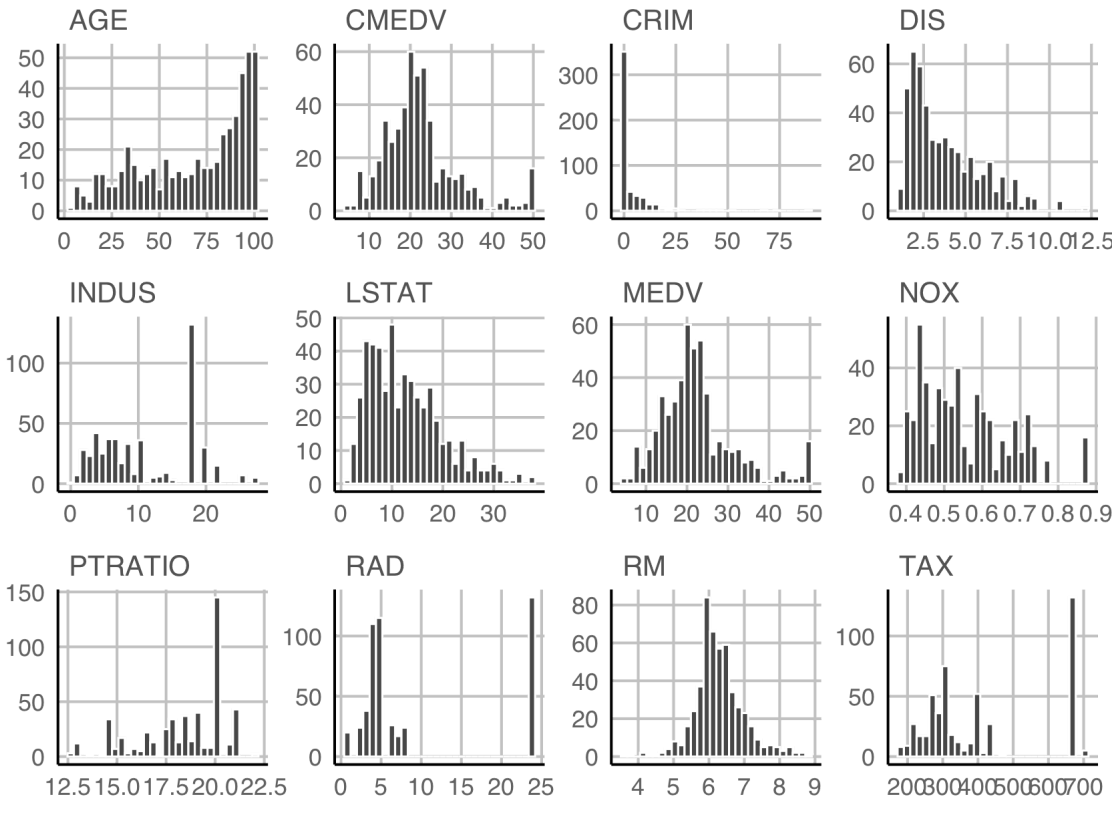


Bin number = 30, Right-open



Case study 3 Boston housing data Part 4/4

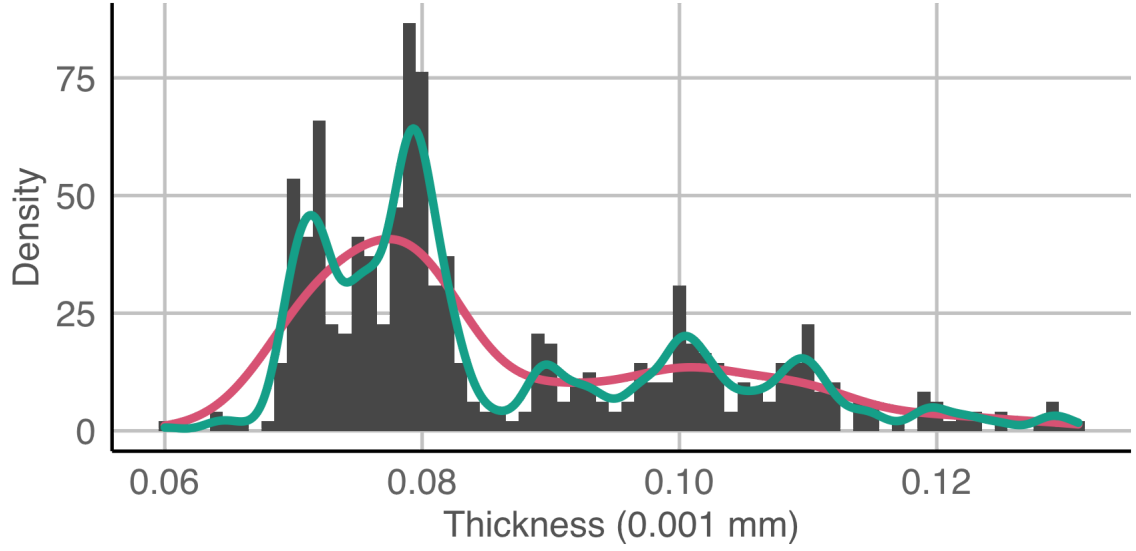
data R



- 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

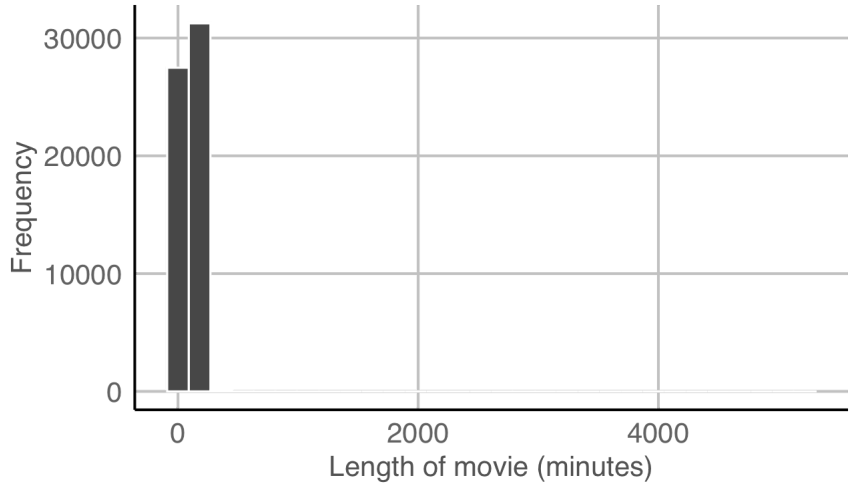


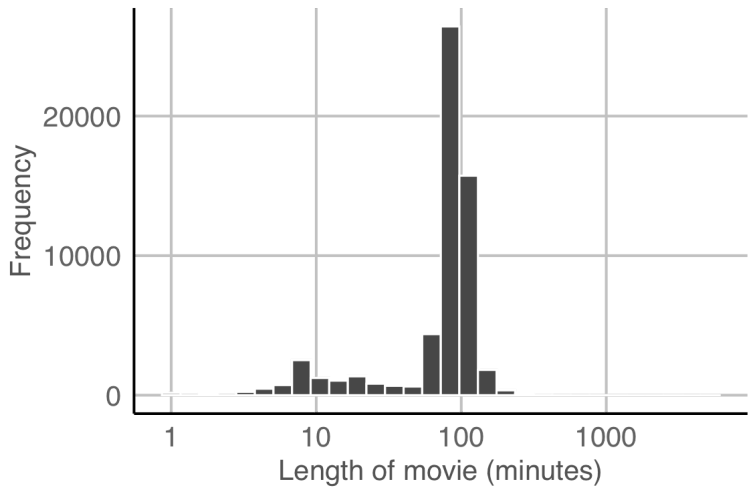
- A stamp collector, Walton von Winkle, bought several collections of Mexican stamps from 1872-1874 and measured the thickness of all of them.
- The different bandwidth for the density plot suggest either that there are two or seven modes.

Focus

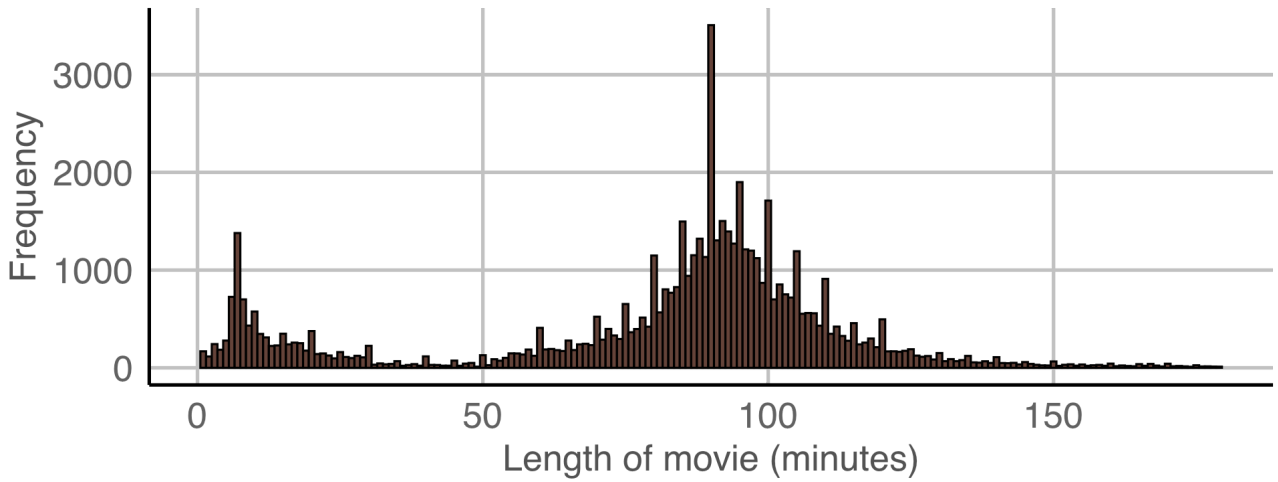
Case study 5 Movie length

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?

Categorical variables

This lecture is based on Chapter 4 of

Unwin (2015) Graphical Data Analysis with R

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**.

```
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:

```
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:

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

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

```
# character input are ordered alphabetically
factor(c("1", "3", "10"))
```

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

```
# 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 factors in R

```
x <- factor(c(10, 20, 30, 10, 20))
mean(x)
```

```
## Warning in mean.default(x): argument is not numeric or logical: returning NA
```

```
## [1] NA
```

⚠ `as.numeric` function returns the internal integer values of the factor

```
mean(as.numeric(x))
```

```
## [1] 1.8
```

You probably want to use:

```
mean(as.numeric(levels(x)[x]))
```

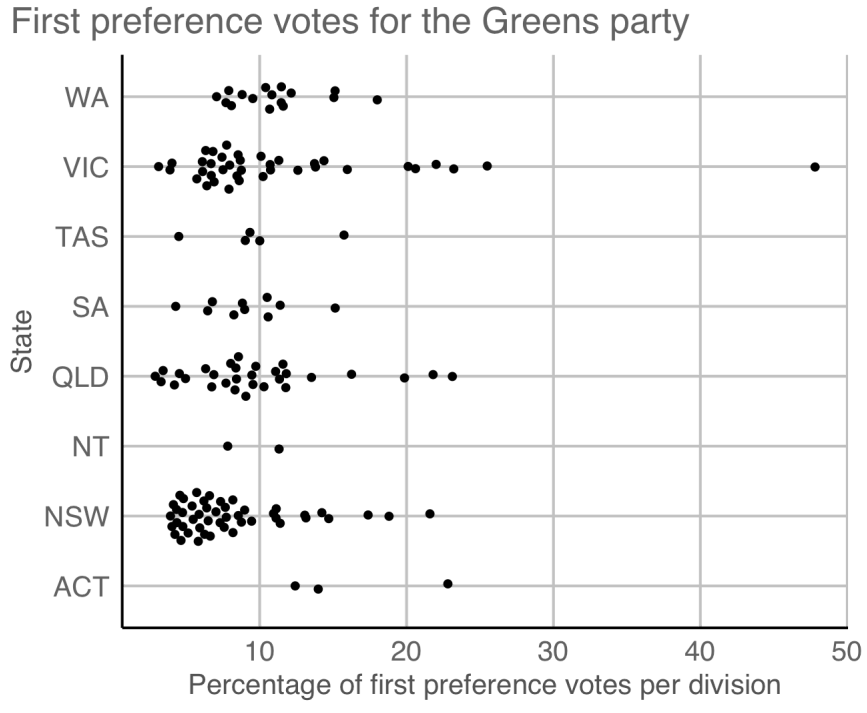
```
## [1] 18
```

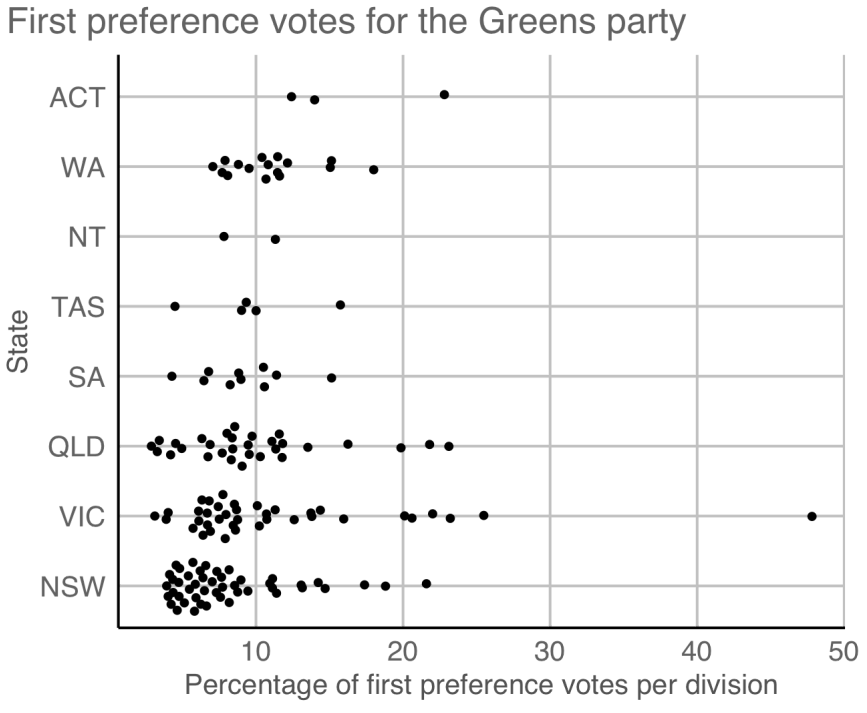
```
mean(as.numeric(as.character(x)))
```

```
## [1] 18
```

Revisiting Case study 1 2019 Australian Federal Election

data R





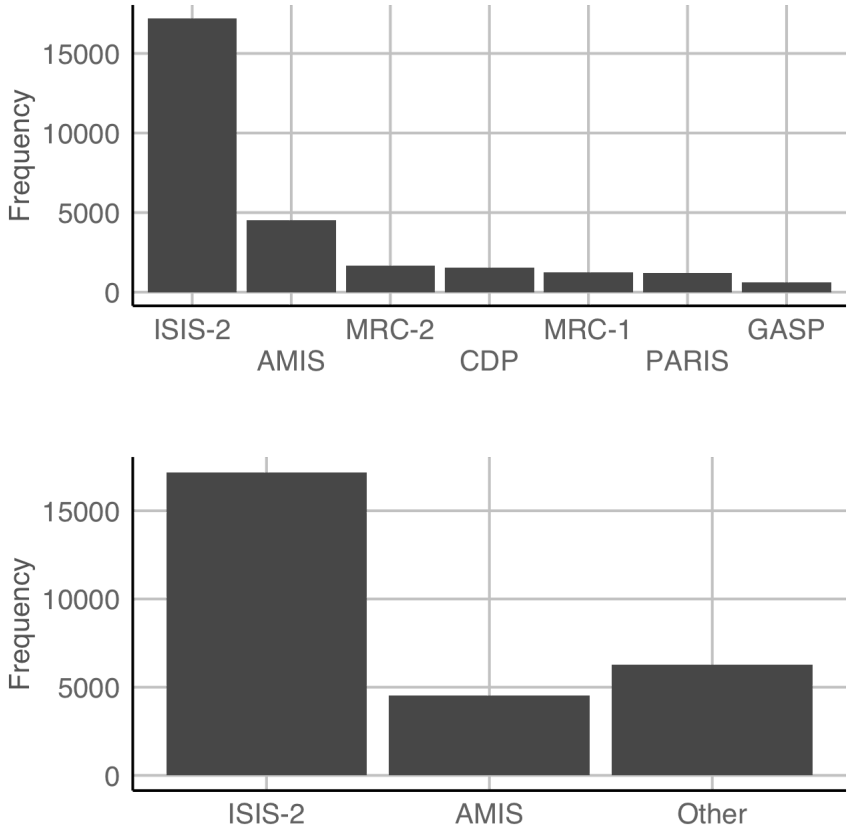
Order nominal variables meaningfully

</> 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)
```

Case study 6 Aspirin use after heart attack

data R



- Meta-analysis is a statistical analysis that combines the results of multiple scientific studies.
- This data studies the use of aspirin for death prevention after myocardial infarction, or in plain terms, a heart attack.
- The ISIS-2 study has more patients than all other studies combined.
- You could consider lumping the categories with low frequencies together.

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.

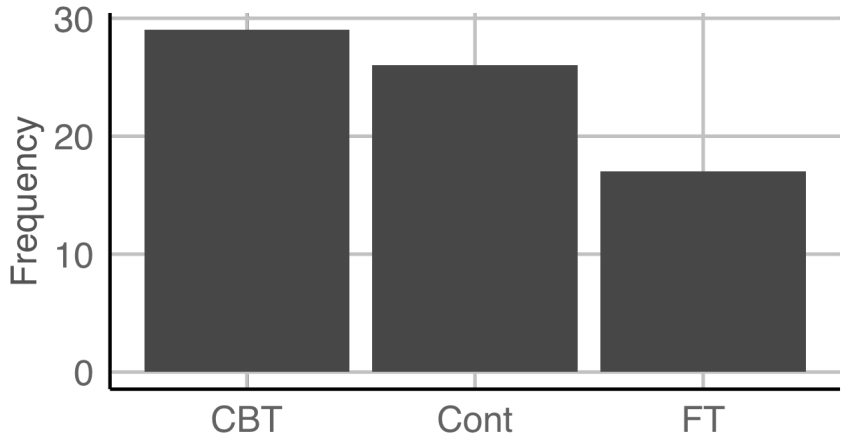
Consider combining factor levels with low frequencies

</> Coding tip: the following family of functions help to easily lump factor levels together:

```
forcats::fct_lump()
forcats::fct_lump_lowfreq()
forcats::fct_lump_min()
forcats::fct_lump_n()
forcats::fct_lump_prop()
# if conditioned on another variable
ifelse(cond, "Other", factor)
dplyr::case_when(cond1 ~ "level1",
                 cond2 ~ "level2",
                 TRUE ~ "Other")
```

Case study 7 Anorexia

data R

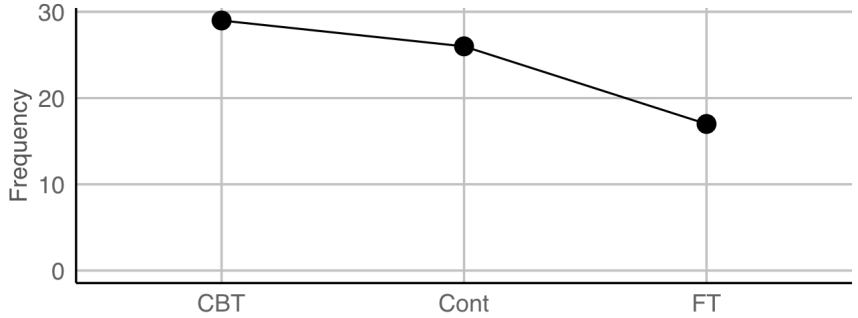


Treatment	Frequency
CBT	29
Cont	26
FT	17

Table or Plot?

- Table for accuracy, plot for visual communication

Why not a point or line?

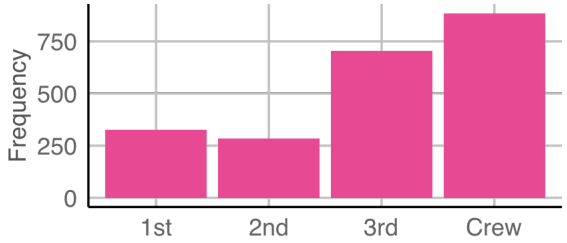


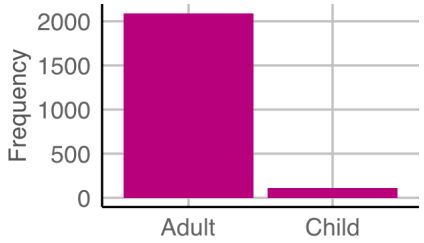
- This can be appropriate depending on what you want to communicate
- A barplot occupies more area compared to a point and the area does a better job of communicating size
- A line is suggestive of a trend

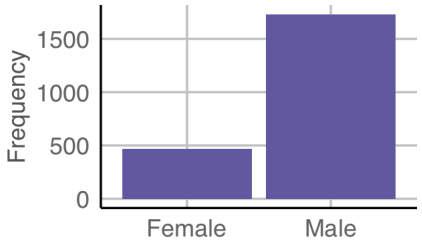
Hand, D. J., Daly, F., McConway, K., Lunn, D. and Ostrowski, E. eds (1993) A Handbook of Small Data Sets. Chapman & Hall, Data set 285 (p. 229)
Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

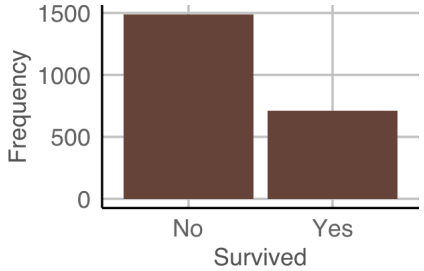
Case study 8 Titanic

data R







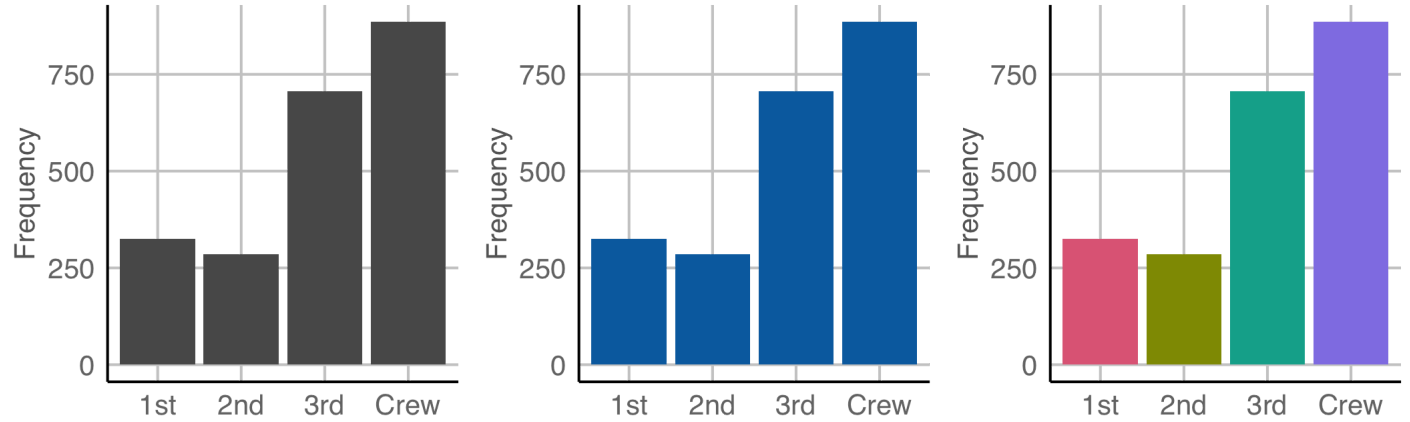


What does the graphs for each categorical variable tell us?

- There were more crews than 1st to 3rd class passengers
- There were far more males on ship; possibly because majority of crew members were male. You can further explore this by constructing two-way tables or graphs that consider both variables.
- Most passengers were adults.
- More than two-thirds of passengers died.

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

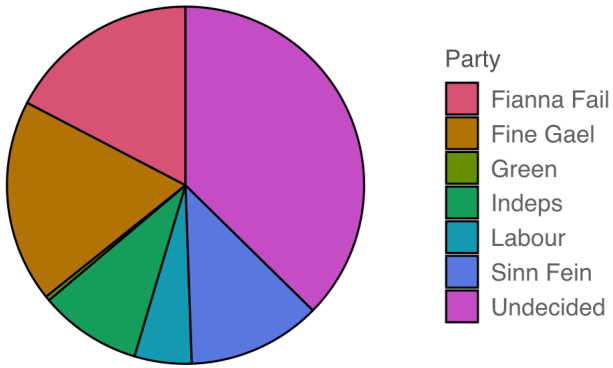
Coloring bars

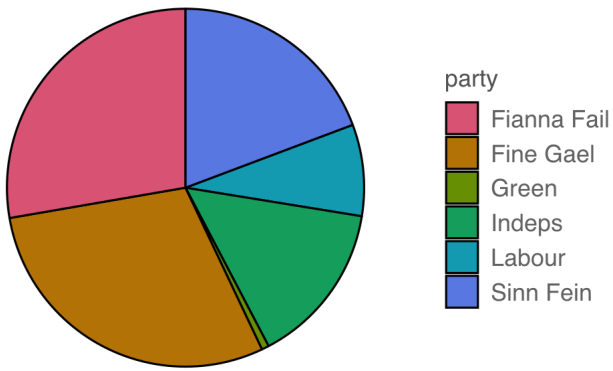


- Colour here doesn't add information as the x-axis already tells us about the categories, but colouring bars can make it more visually appealing.
- If you have too many categories colour won't work well to differentiate the categories.

Case study 9 Opinion poll in Ireland Aug 2013

data R

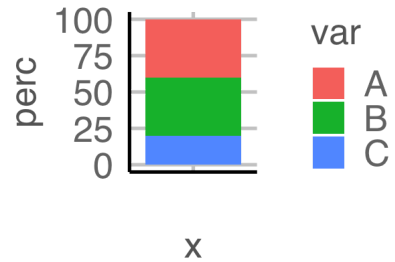




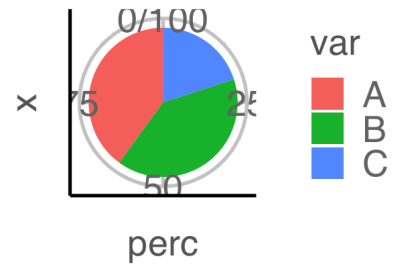
- Pie chart is popular in mainstream media but are not generally recommended as people are generally poor at comparing angles.
- 3D pie charts should definitely be avoided!
- Here you can see that there are many people that are "Undecided" for which political party to support and failing to account for this paints a different picture.

Piechart is a stacked barplot just with a transformed coordinate system

```
df <- data.frame(var = c("A", "B", "C"), perc = c(40, 40, 20))
g <- ggplot(df, aes("", perc, fill = var)) +
  geom_col()
g
```

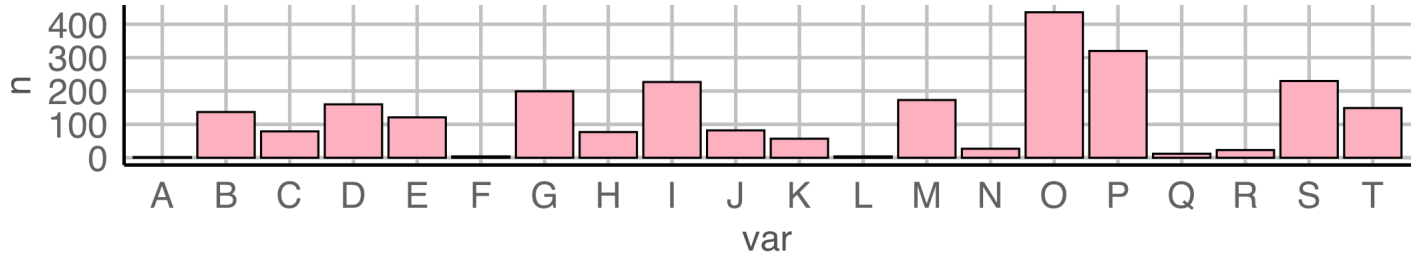



```
g + coord_polar("y")
```

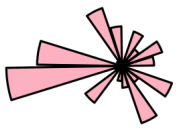


Roseplot is a barplot just with a transformed coordinate system

```
dummy <- data.frame(var = LETTERS[1:20],
                    n = round(rexp(20, 1/100)))
g <- ggplot(dummy, aes(var, n)) + geom_col(fill = "pink", color = "black")
g
```




```
g + coord_polar("x") + theme_void()
```



Take away messages

- Again, be prepared to do multiple plots
- Changing bins or 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, missing categories)

Lecturer: *Emi Tanaka*

ETC5521.Clayton-x@monash.edu

Week 4 - Session 2