ETC5521: Exploratory Data Analysis

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

Lecturer: *Emi Tanaka*
Department of Econometrics and Business Statistics
ETC5521.Clayton-x@monash.edu
Week 4 - Session 1

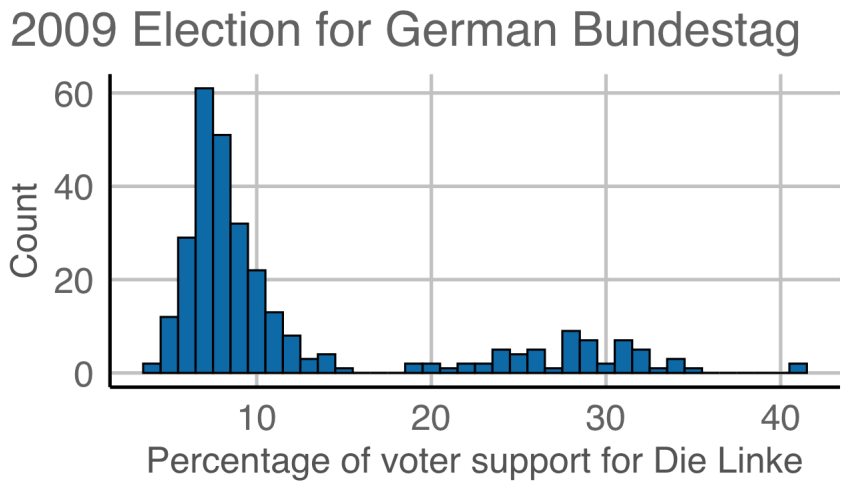
Continuous variables

This lecture is based on Chapter 3 of

Unwin (2015) Graphical Data Analysis with R

Case study 1 German Bundestag Election 2009 Part 1/2

data R



About

- Bundestag is the first chamber of the German parliament.
- Germany has 299 constituencies.
- Die Linke is a party on the left.

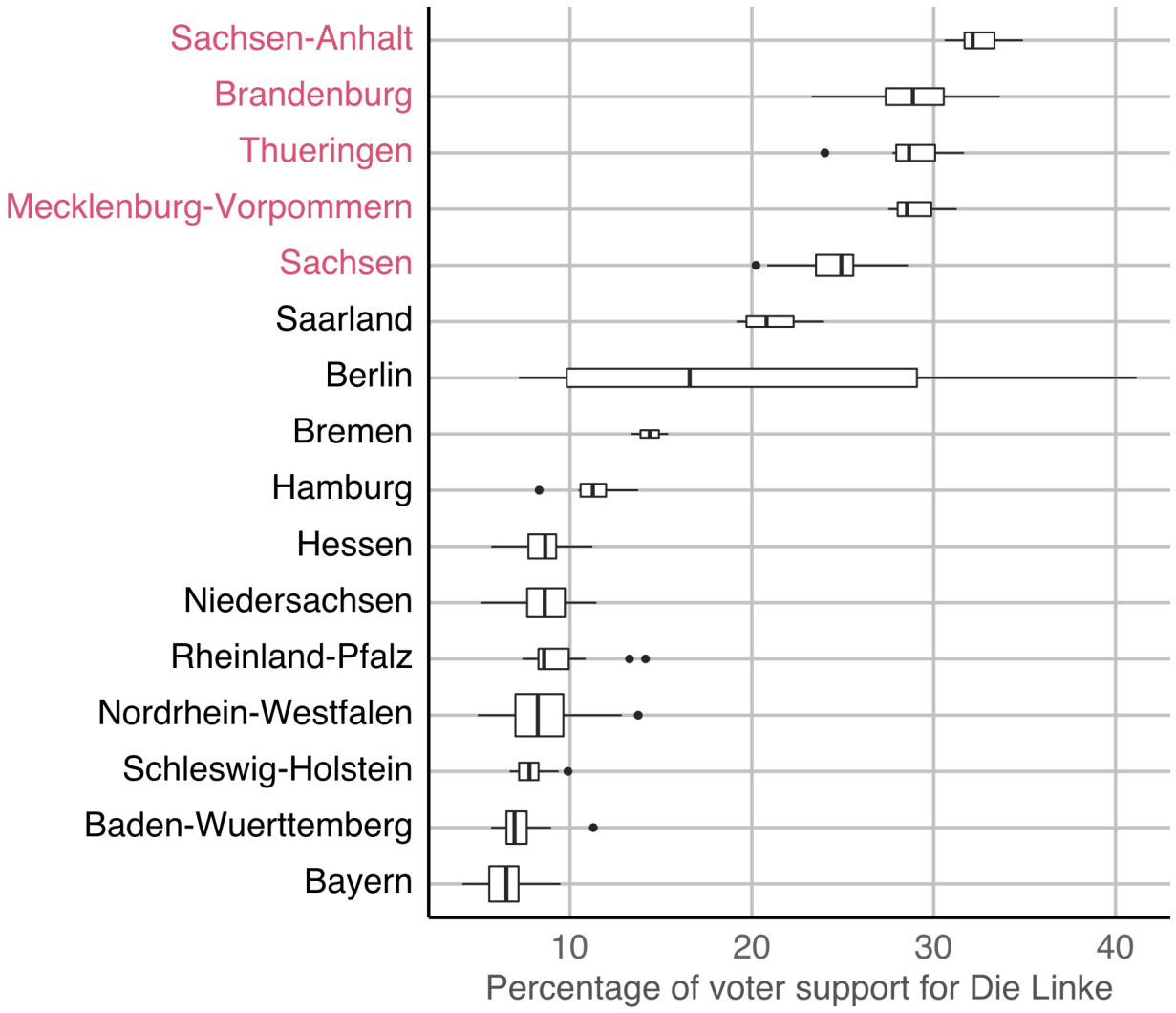
What does this graph tell you?

- Majority of the country does not support Die Linke.
- The country appear divided in support of Die Linke with some constituents more supportive than the others.

Leisch, Friedrich (2006) A Toolbox for K-Centroids Cluster Analysis. *Computational Statistics and Data Analysis* **51** (2) 526-544
Data source: https://www.bundestag.de/

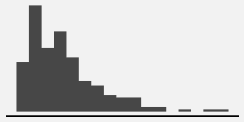
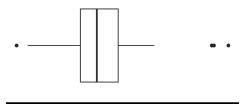
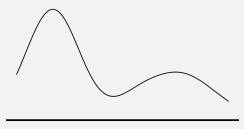
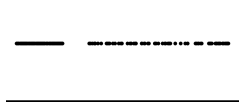
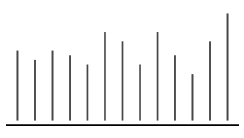
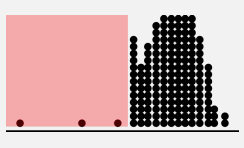
Case study 1 German Bundestag Election 2009 Part 2/2

data R

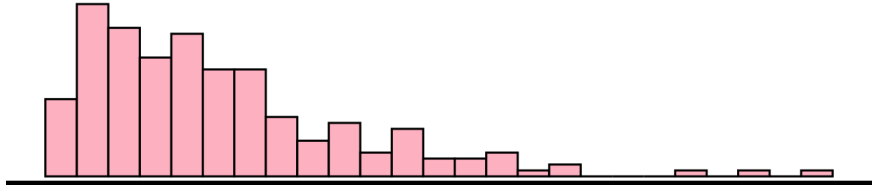


- The width of the boxplot is proportional to the Bundesland size.
- The pink colour indicates the old-East Germany.
- Die-Linke had more support in the old-East.

Possible features of continuous variables

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 contiuous variables



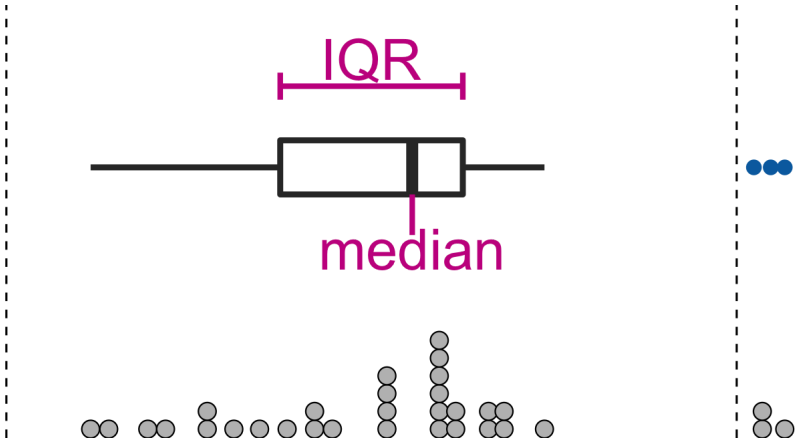
- A measure of ***central tendency***, e.g. mean, median and mode.
- A measure of ***dispersion*** (also called variability or spread), e.g. variance, standard deviation and interquartile range.
- There are other measures, e.g. ***skewness*** and ***kurtosis*** that measures "tailedness", but these are 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*.
- **Signifiance tests** or **hypothesis tests**: When testing for $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$ (often $\mu_0 = 0$), the t-test is commonly used if the underlying data are believed to be normally distributed.

Outliers

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

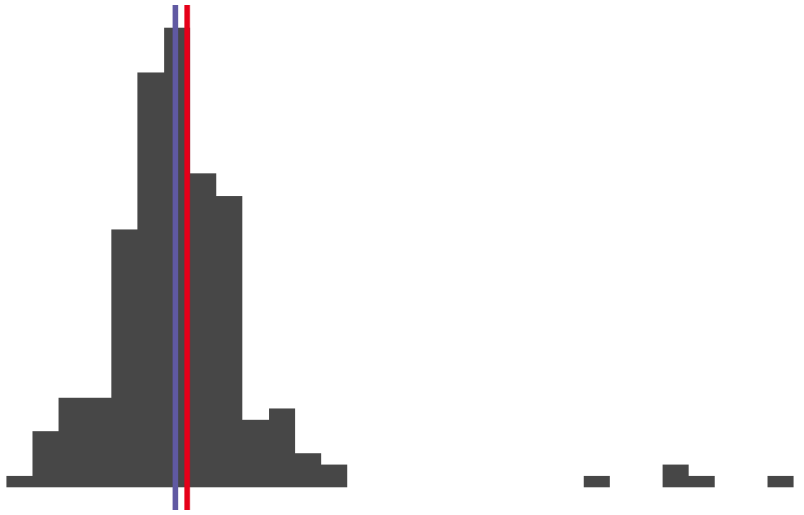
- Outliers can occur by chance in almost all distributions, but could be indicative of a measurement error, a different population, or an issue with the sampling process.
- Outlying values of independent variables are referred to as **high-leverage points**, although this distinction is not particularly important when analysing a single continuous variable.

Closer look at the *boxplot*



- Observations that are outside the range of lower to upper thresholds 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.

Robust statistics: measure of central tendency



- **Mean** is a non-robust measure of location.
- Some robust measures of locations are:
  - **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

Robust statistics: 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 calculated as

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

- Interquartile range is the difference between 1st and 3rd quartile and is more robust measure of spread than above.
- Median absolute deviance (MAD) is also more robust and defined as

$$\text{median}(|x_i - \bar{x}|).$$

Transformations

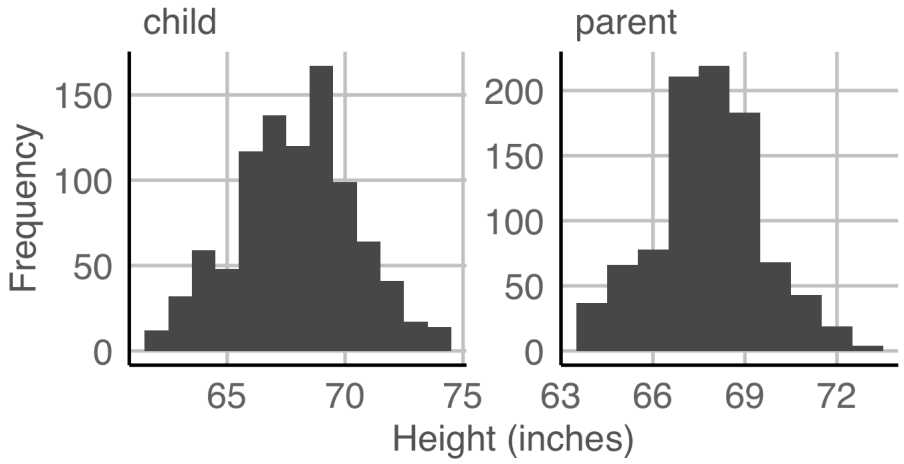
- Transformations to data are used as an attempt to stabilize the variance or make the data symmetrical.
- Log and square root transformations are popular.
- A range of λ values for (one-parameter) Box-Cox transformation is sometimes used to test for optimal transformation:

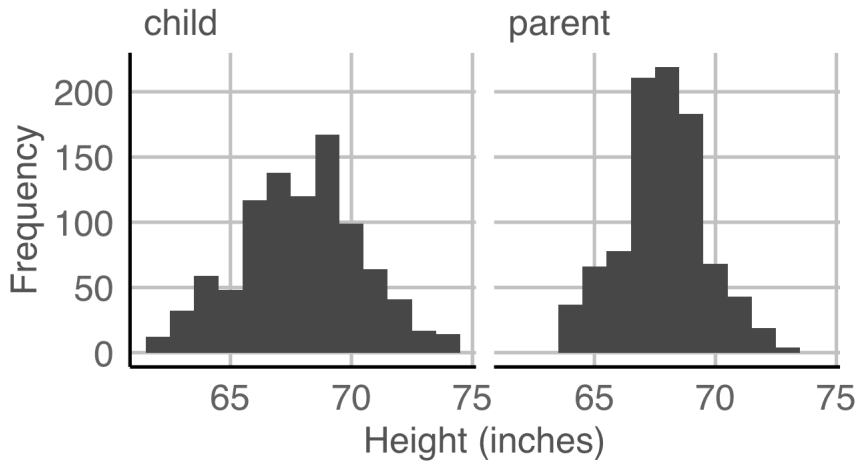
$$y(\lambda) = \begin{cases} \dfrac{(y^{\lambda} - 1)}{\lambda} \\ \log(y) \end{cases}$$

Box, G. E. P. and Cox, D. R. (1964). An analysis of transformations, Journal of the Royal Statistical Society, Series B, 26, 211-252.

Case study 2 Children and midparents heights Part 1/3

data R





- Midparent's height is an average of the father's height and 1.08 times the mother's height.
- The data contains 205 parents and 928 children.

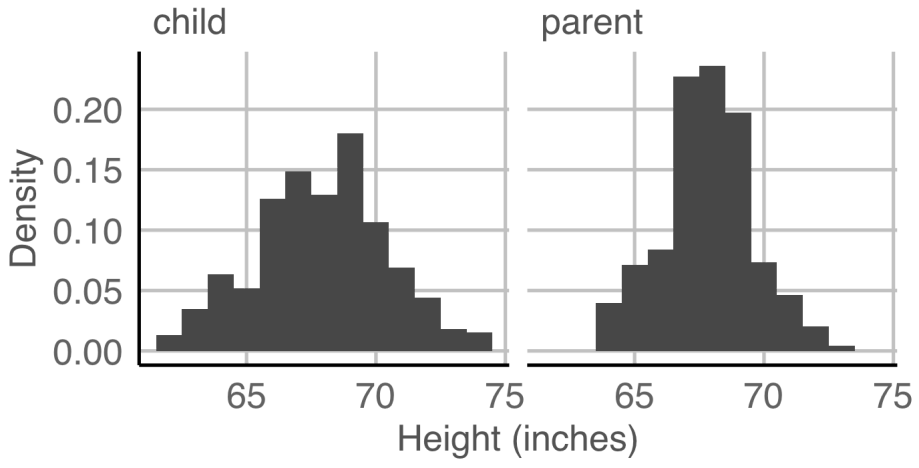
```
## # A tibble: 928 x 2
##   child parent
##   <dbl>  <dbl>
## 1  61.7     64
## 2  63.2     64
## 3  63.2     64
## # … with 925 more rows
```

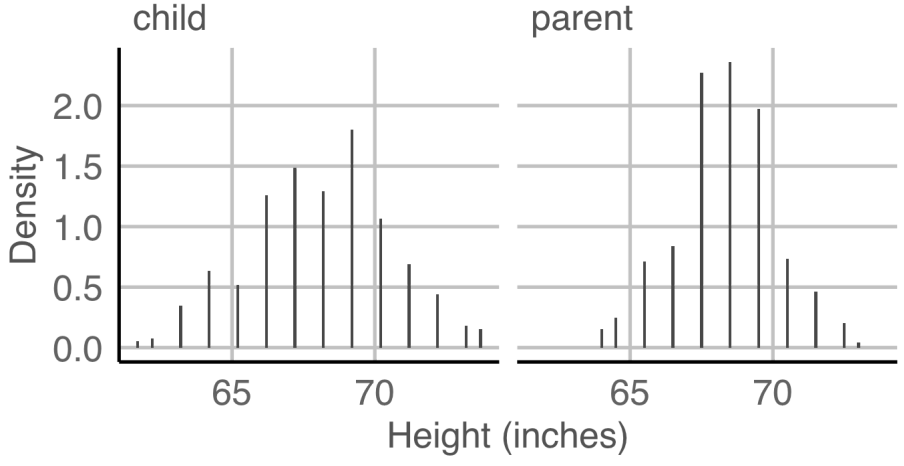
- The data included families of 1 to 15 children, so in the extreme case, one midparent data point is repeated 15 times in the data.
- The frequency of midparents heights therefore are over-represented with parents with large family size.

Verzani, John (2018). UsingR: Data Sets, Etc. for the Text "Using R for Introductory Statistics", Second Edition. R package version 2.0-6.
Galton, Franics (1886). Regression Towards Mediocrity in Hereditary Stature. *Journal of the Anthropological Institute* **15** 246-263

Case study 2 Children and midparents heights Part 2/3

data R





- Changing the bin width of histogram from 1 to 0.1, we can see the data have been rounded and hence some precision is lost in the data.
- The data confirms this with most children's height recorded with ".2" at the end and most midparents heights recorded with ".5" at the end.

`child`

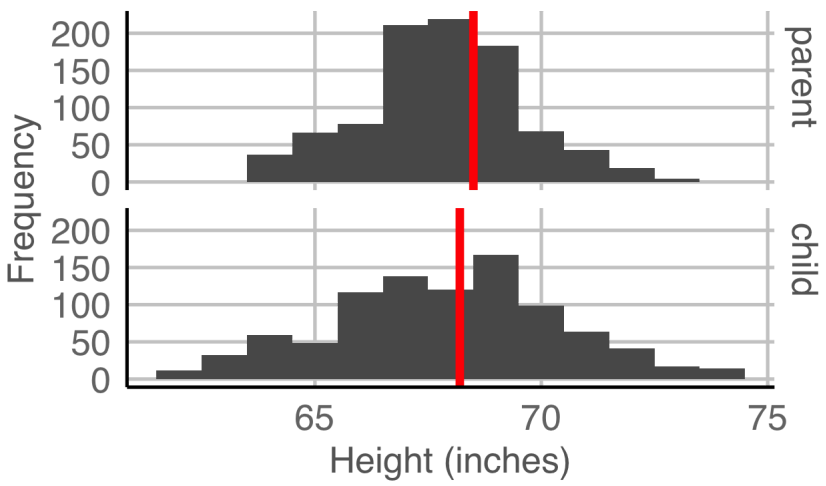
```
##  [1] 61.7 62.2 63.2 64.2 65.2 66.2 67.2 68.2 69.2
## [10] 70.2 71.2 72.2 73.2 73.7
```

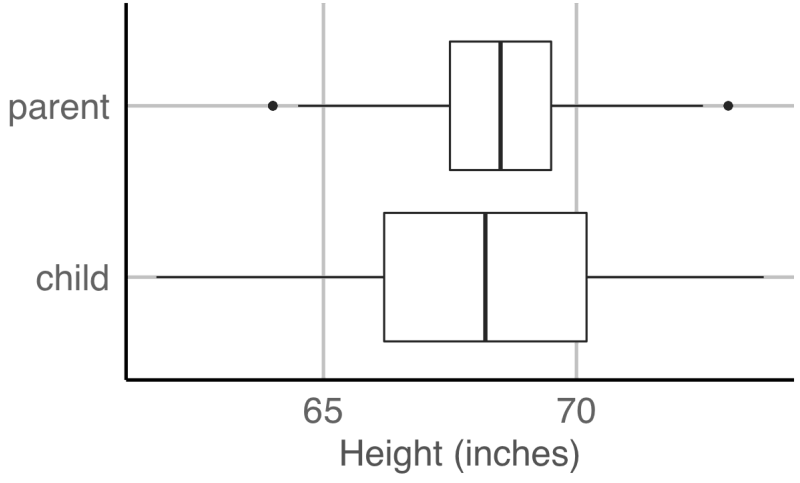
`parent`

```
##  [1] 70.5 68.5 65.5 64.5 64.0 67.5 66.5 69.5 71.5
## [10] 72.5 73.0
```

Case study 2 Children and midparents heights Part 3/3

data R

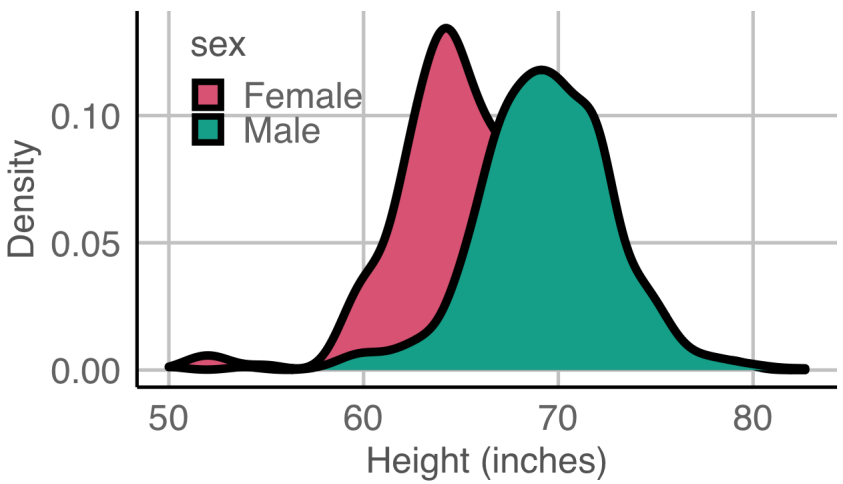


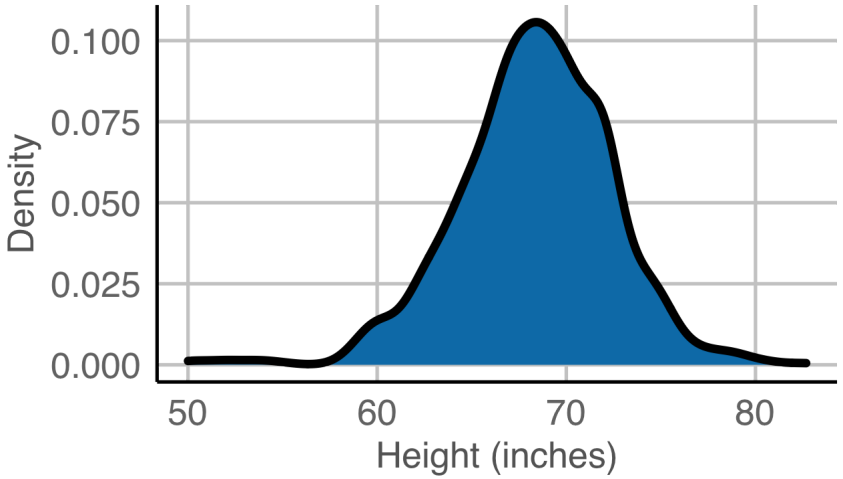


- Aligning the histogram vertically makes it easier to compare the heights of children and midparents.
- The side-by-side boxplots makes it easy to see the variability of the heights of the midparents are smaller than the children.
- The smaller variability is expected because the midparents heights are average of two values.
- We can also see that the median height is larger for the midparents than children.
- You may think that the heights of children should be ***bimodal***, *one peak for* ***male*** and the *other peak for* ***female***. But that is not necessary the case as we'll see next.

Case study 3 Self-reported heights

data R



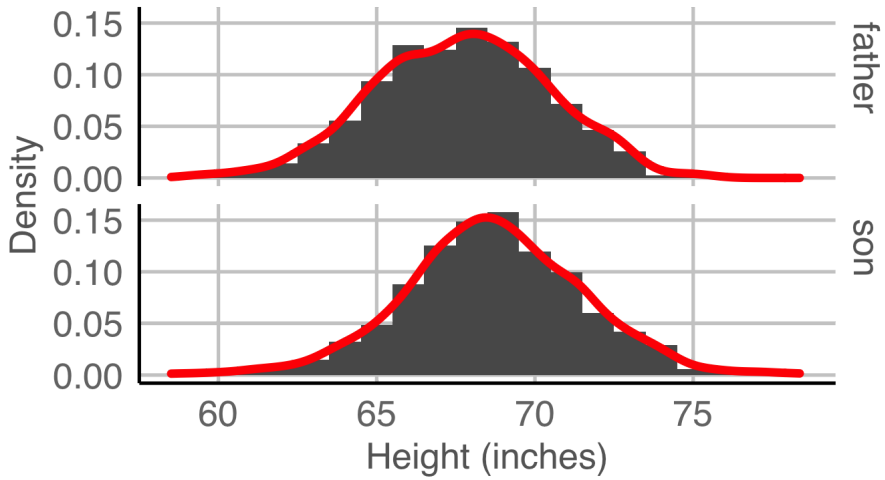


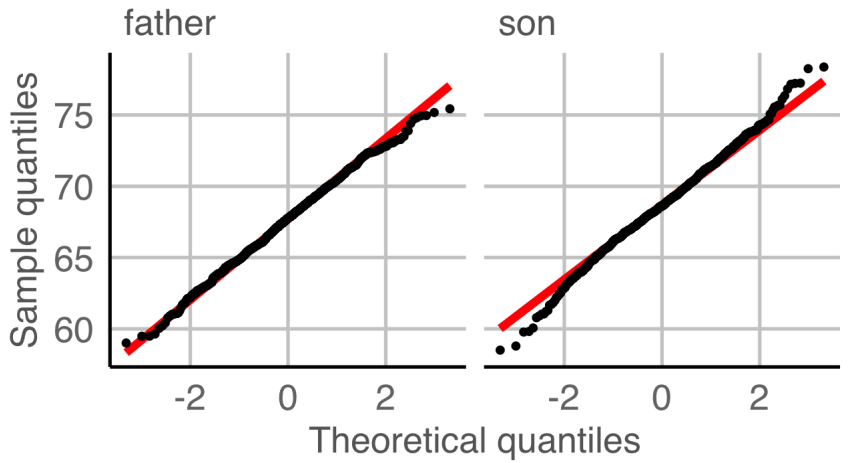
- You can see that drawing separate density plots for each sexes shows that the women are on average shorter than men.
- The bimodality is however not visible when the data are combined.

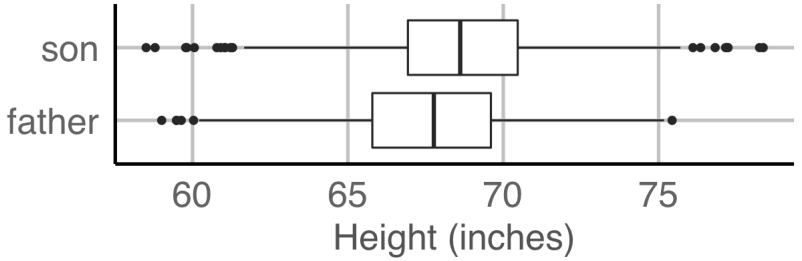
Irizarry, Rafael A. and Gill, Amy (2019). dslabs: Data Science Labs. R package version 0.7.3.

Case study 4 Men's heights

data R



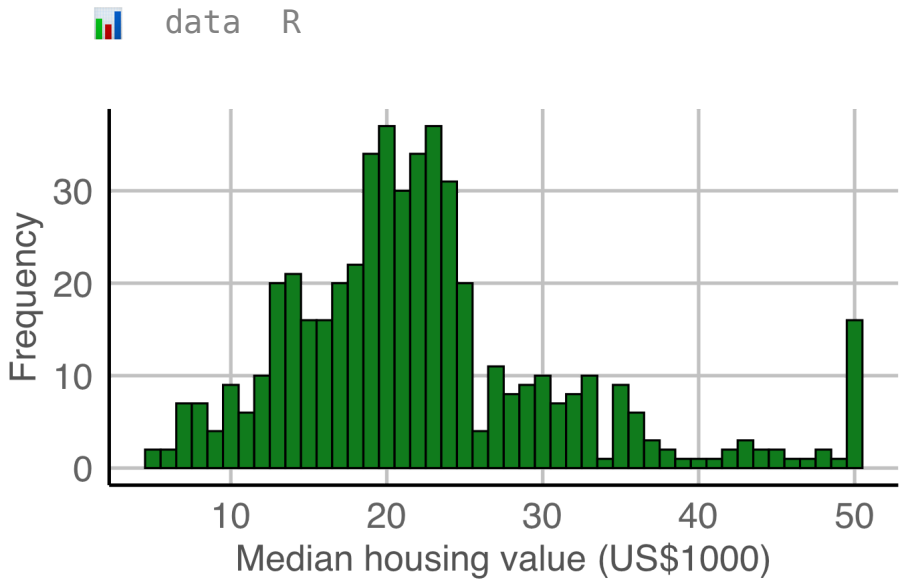




- The height distribution of the son appears more normal than of the father looking at the density plot.
- Looking at the Q-Q plots, this however looks the other way around.
- The heights are recorded to five decimal places (e.g. `65.04851`).
- It's unlikely that the heights were measured to such high precision and rather that someone must have "jittered" the data (i.e. added some small random perturbation to the observation).

Verzani, John (2018). UsingR: Data Sets, Etc. for the Text "Using R for Introductory Statistics", Second Edition. R package version 2.0-6.
Pearson, Karl and Lee, Alison (1903) On the laws of inheritance in man: I. Inheritance of physical characters. *Biometrika* **2** (4) 357-462

Case study 5 Boston housing data Part 1/4

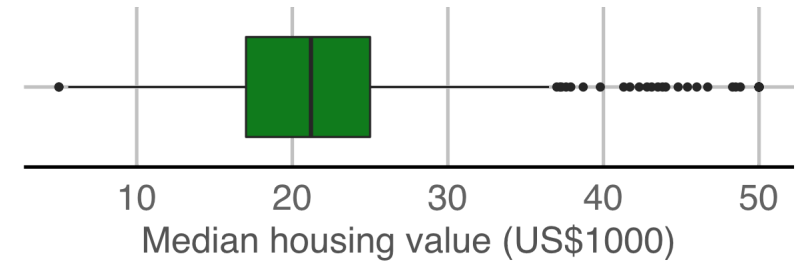


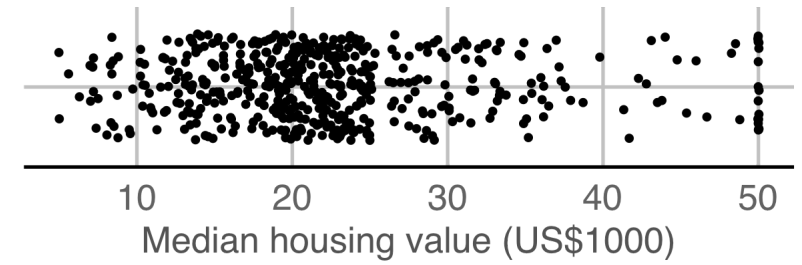
- Thre is a large frequency in the final bin.
- There is a decline in observations in the $40-49K range as well as dip in observations around $26K and $34K.
- The histogram is using a bin width of 1 unit and is **left-open** (or **right-closed**): (4.5, 5.5], (5.5, 6.5] ... (49.5, 50.5].
- Occasionally,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

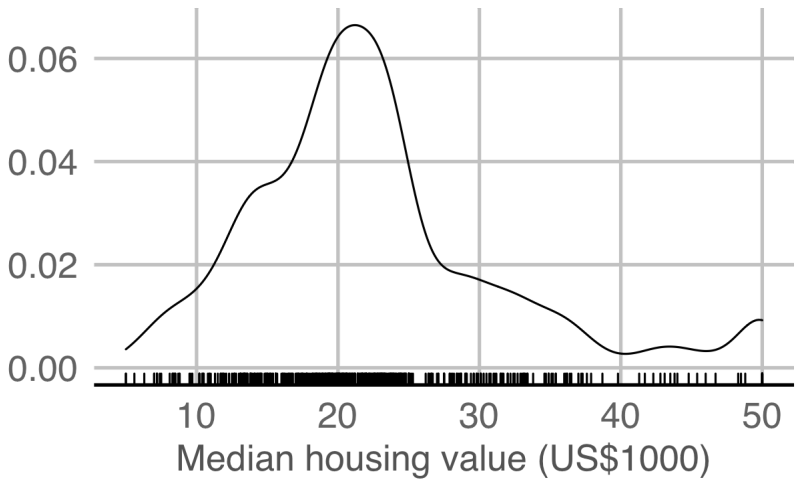
Harrison, David, and Daniel L. Rubinfeld (1978) Hedonic Housing Prices and the Demand for Clean Air, *Journal of Environmental Economics and Management* **5** 81-102. Original data.
Gilley, O.W. and R. Kelley Pace (1996) On the Harrison and Rubinfeld Data. *Journal of Environmental Economics and Management* **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 5 Boston housing data Part 2/4

data R





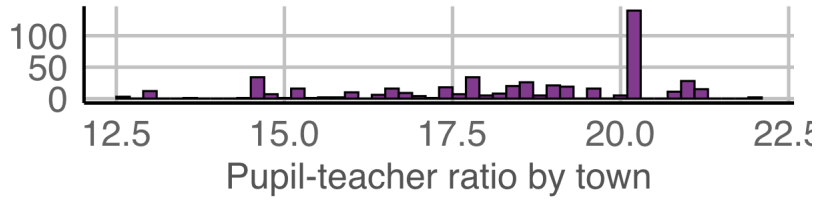


- Density plots depend on the bandwidth chosen and more than often do not estimate well at boundary cases.
- There are various way to present features of the data using a plot and what works for one person, may not be as straightforward for another.
- Be prepared to do multiple plots.

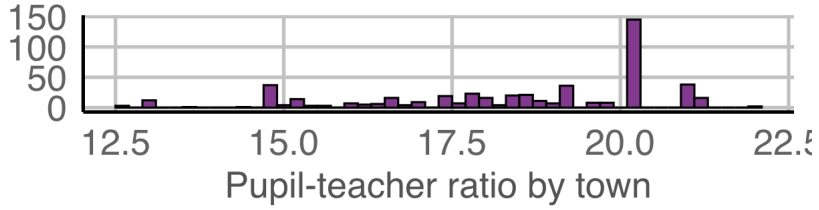
Case study 5 Boston housing data Part 3/4

data R

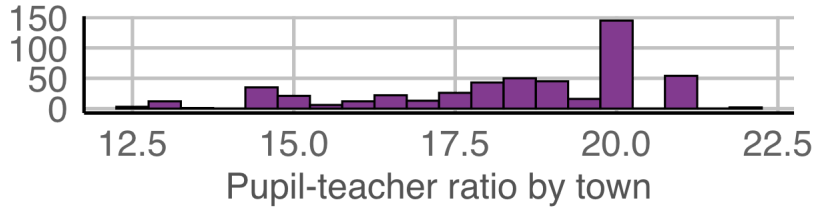
Bin width = 0.2, Left-open



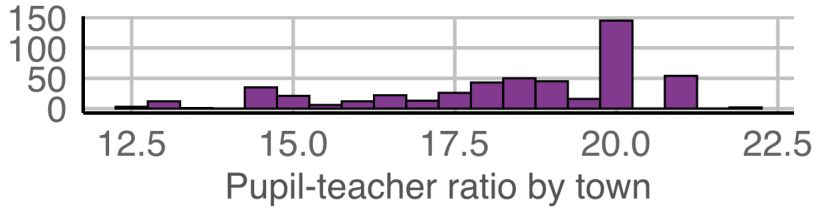
Bin width = 0.2, Right-open



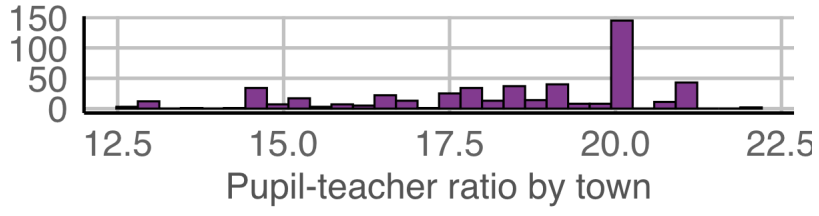
Bin width = 0.5, Left-open



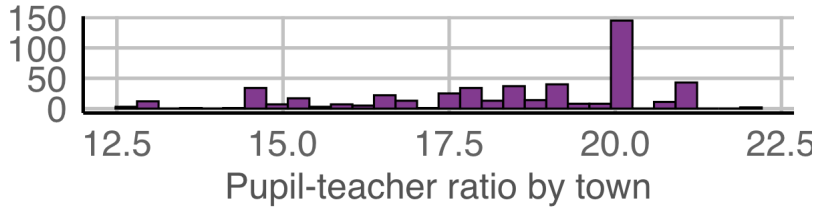
Bin width = 0.5, Right-open



Bin number = 30, Left-open

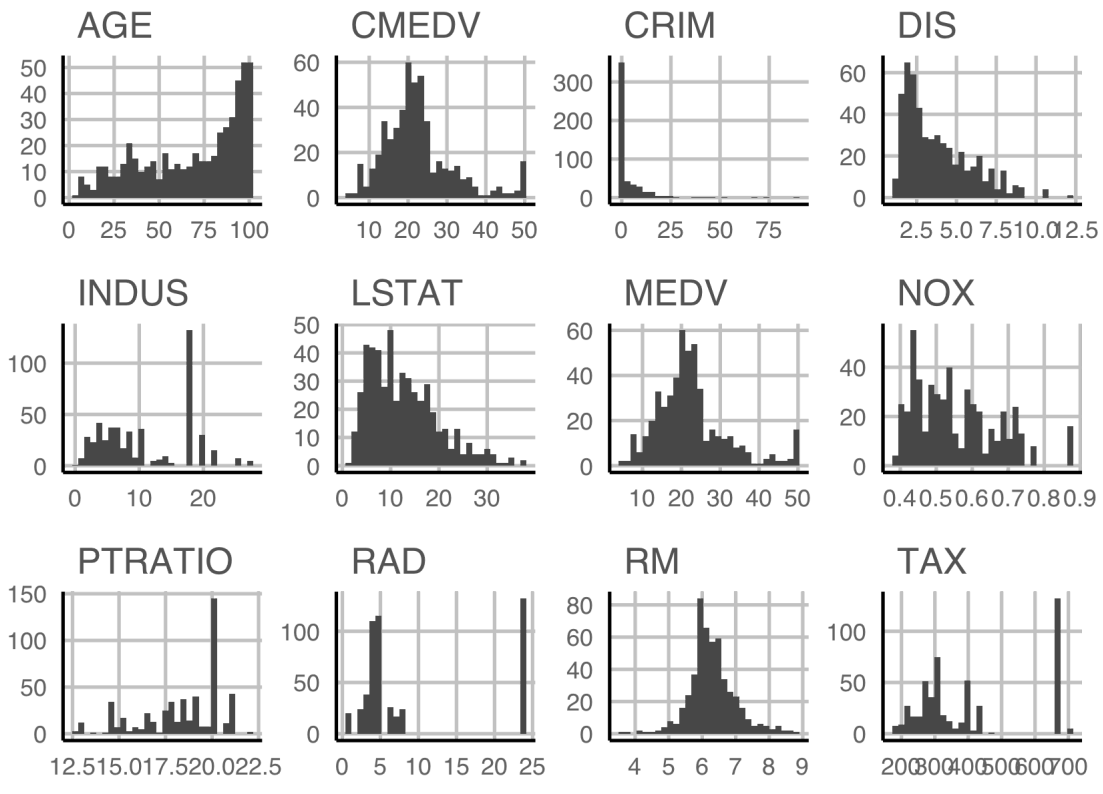


Bin number = 30, Right-open



Case study 5 Boston housing data Part 4/4

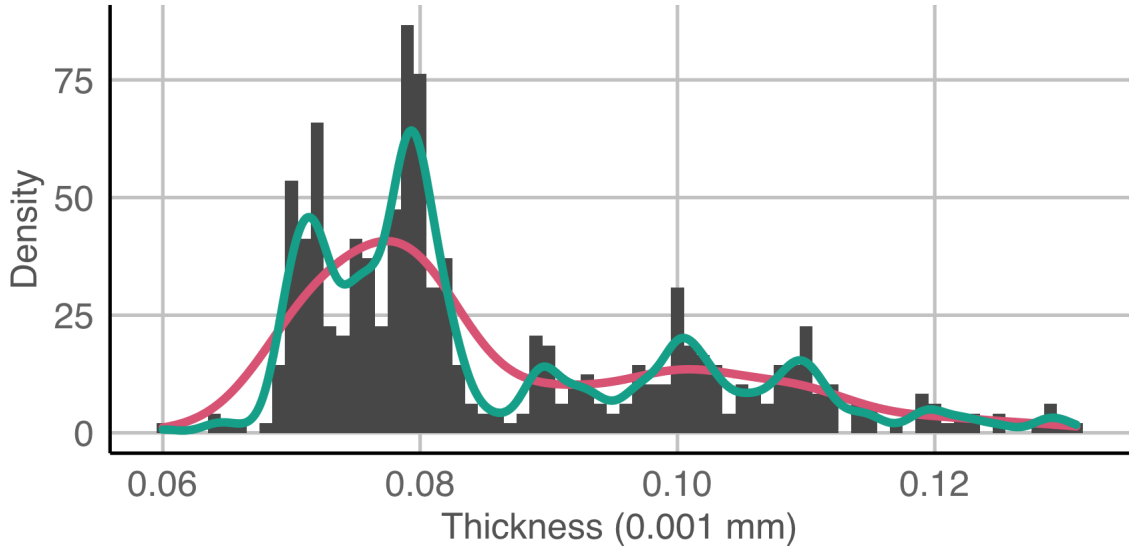
data R



- 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 6 Hidalgo stamps thickness

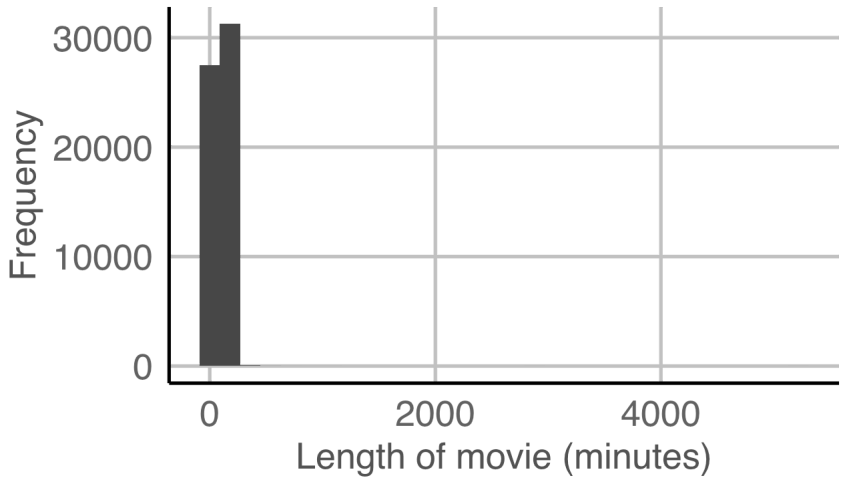
data R

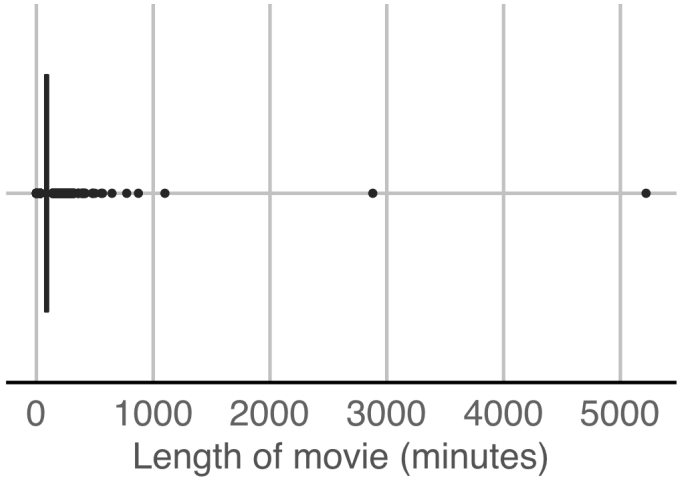


- A stamp collector, Walton von Winkle, bought several collections of Mexican stamps from 1872-1874 and measured the thickness of all of them.
- The different bandwidth for the density plot suggest either that there are two or seven modes.

Case study 7 Movie length

data R

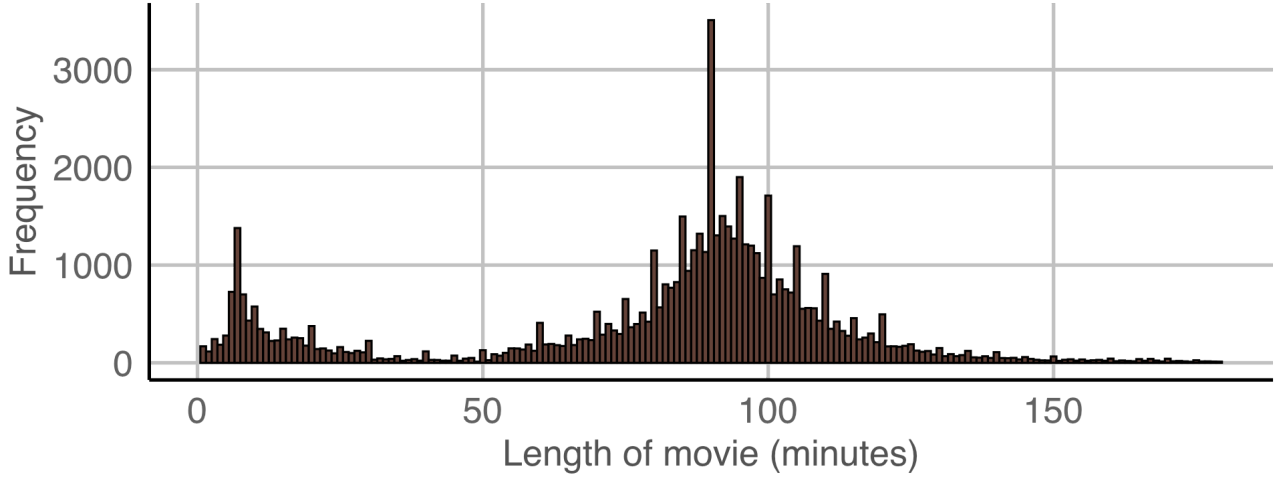




- Upon further exploration, you can find the two movies that are well over 16 hours long are:

```
## Cure for Insomnia, The
## Four Stars
## Longest Most Meaningless Movie in the World, The
```

- We can restrict our attention to films under 3 hours:



That's it, for this lecture!

Lecturer: Emi Tanaka

Department of Econometrics and Business Statistics

ETC5521.Clayton-x@monash.edu