Forum 2
Data Skills for Scientists Data: Summarising and Visualising

2025-09-24

Course Structure

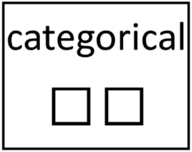
Data

The 3 topics in the first module, ‘Data’ are:

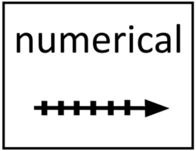
- Navigating Uncertainty;
- Summarising and Visualising; and
- Probabilities and Distributions.

Summarising and Visualising

In this forum will cover the second topic, how to summarise and visualise different types of variables. Recall that variables are either



or



Summarising and Visualising

In this forum we'll start with a general discussion and introduction to data visualisation, and then structure the remainder of the forum into sections depending on which kinds of variables we are looking at:

- One categorical variable,
- Two categorical variables,
- One numeric variable,
- One numeric and one categorical variable,
- Two numeric variables.

Summarising and Visualising

In each section, for each combination of variable types, we'll introduce:

- Data visualisations including those we will use in this course, and examples of others that we won't but you might see,
- Methods to summarise data, and
- Discussion on how to interpret visualisations and summaries.

Practicals

This forum is special in how it relates to the practicals in that how to visualise and summarise data yourself in R will be split across two practicals:

- the first on summarising data and
- the second on visualising data.

In the remainder of the course, each forum will correspond to one practical in the week after the forum.

Intro to Data Visualisation

Causes of Mortality in the Army in the East

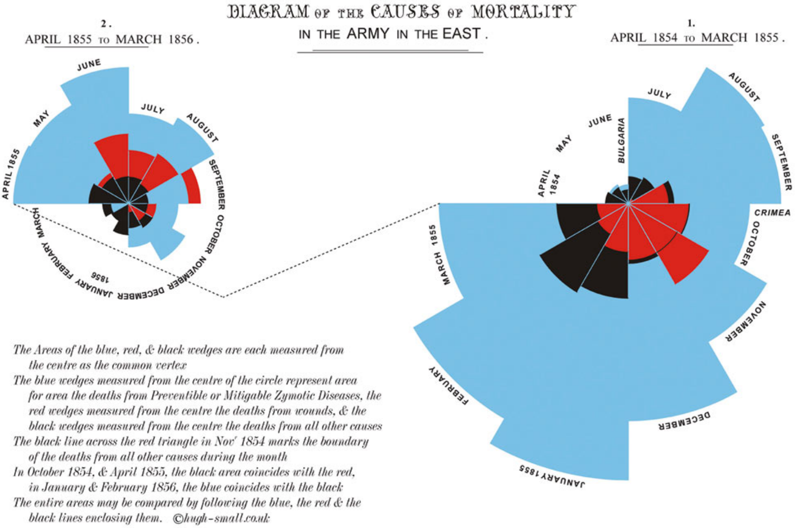


DIAGRAM OF THE CAUSES OF MORTALITY
IN THE ARMY IN THE EAST.

The Areas of the blue, red, & black wedges are each measured from the centre as the common vertex

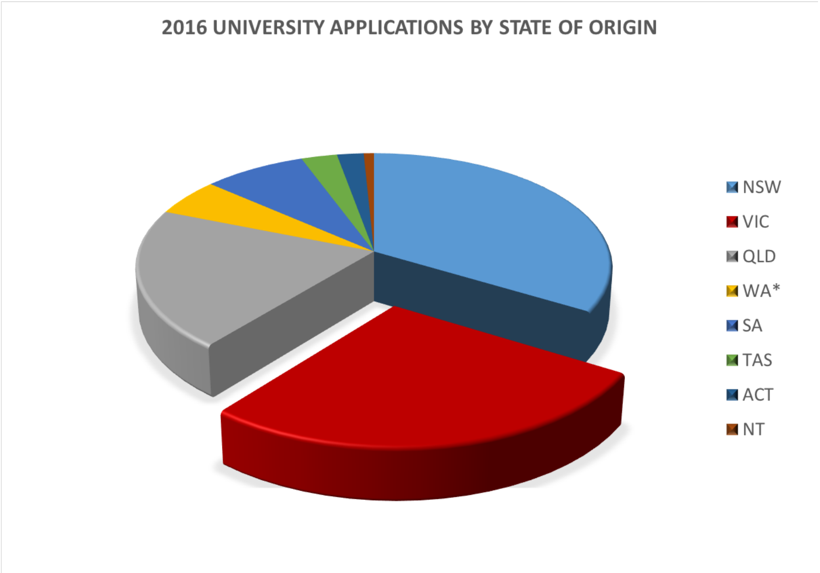
The blue wedges measured from the centre of the circle represent area for area the deaths from Preventible or Mitigable Zymotic Diseases, the red wedges measured from the centre the deaths from wounds, & the black wedges measured from the centre the deaths from all other causes

The black line across the red triangle in Novr 1854 marks the boundary of the deaths from all other causes during the month

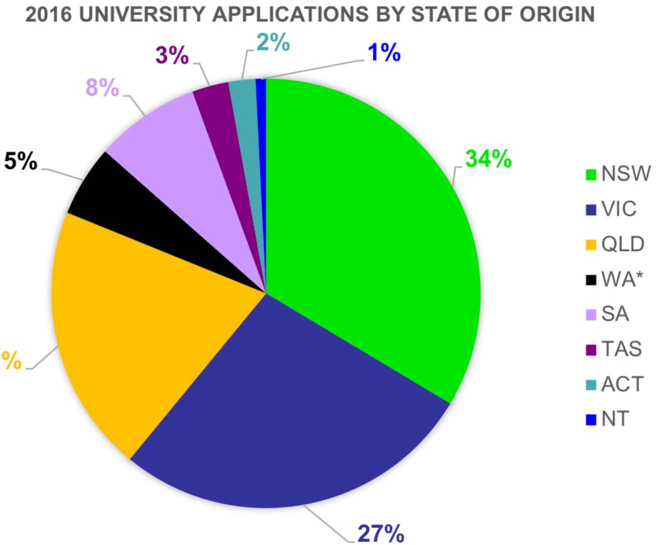
In October 1854, & April 1855, the black area coincides with the red, in January & February 1856, the blue coincides with the black

The entire areas may be compared by following the blue, the red & the black lines enclosing them. ©hugh-small.co.uk

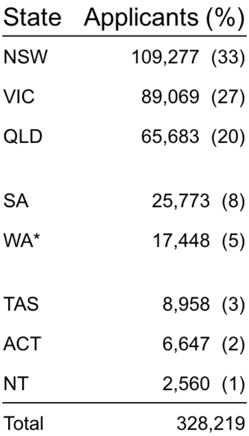
Why 3D effects are bad



Why pie charts are bad

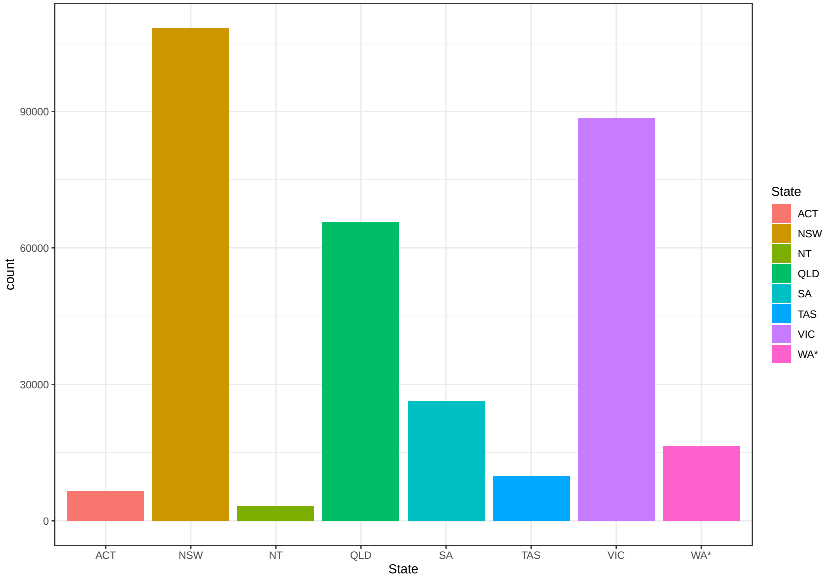


What's the alternative?

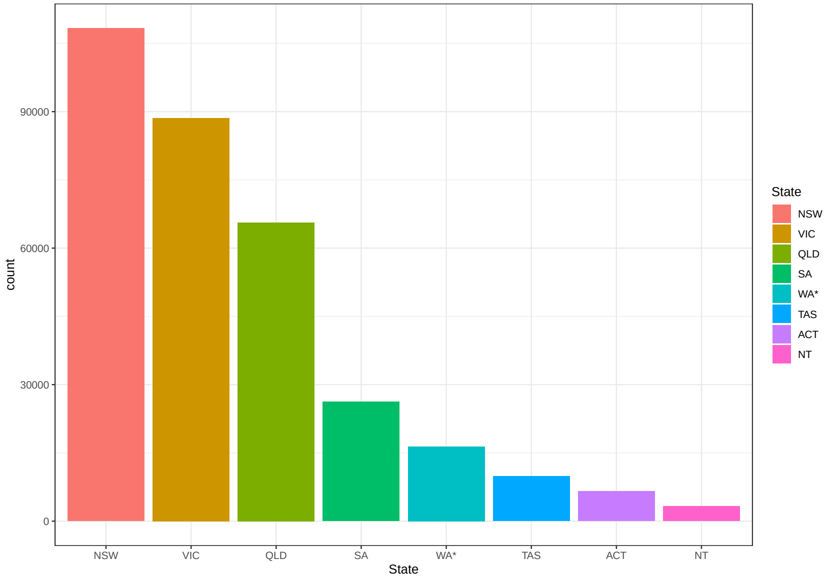


State	Applicants (%)
NSW	109,277 (33)
VIC	89,069 (27)
QLD	65,683 (20)
SA	25,773 (8)
WA*	17,448 (5)
TAS	8,958 (3)
ACT	6,647 (2)
NT	2,560 (1)
Total	328,219

Bar Chart Alphabetical



Bar Chart Sorted Decreasing



Data Visualisation

Is a powerful tool to explore and understand data.

But it can be easy to make bad visualisations that:

- Introduce perception biases,
- Obscure important information,
- Are misleading or inaccurate.

! Important

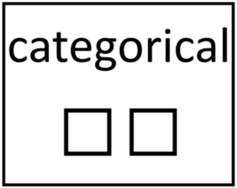
Being mindful when producing visualisations, and being careful when interpreting them is important!

Data Visualisation

There are entire courses on data visualisation.

In this course, we'll just touch on it briefly and show you some useful visualisations for different types of variables.

One Categorical Variable



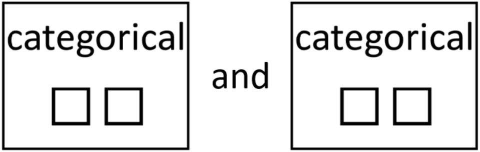
A single categorical variable is often visualised with either

- Bar charts (which we'll use in this course) or
- Pie charts (which we won't).

and summarised in tables of counts and/or proportions (or percentages).

Two Categorical Variables

Summarising Two Categorical Variables



categorical □□ and categorical □□

Will usually be summarised in a **two-way table** also called a contingency table, in which counts are organised by assigning one categorical variable to rows, and the other to columns.

Carbon Tax Repeal Bill

In 2014, the Australian Senate under Prime Minister Tony Abbott at the time voted to repeal the carbon tax bill that was at the time in place.

Lets take a look at how the senators voted!

Two-Way Table of Counts

Party	Absent	Against	For
Coalition	1	0	32
Labor	2	23	0
Greens	1	9	0
Palmer United	0	0	3
Xenophon	1	0	0
Motoring Enthusiast	0	0	1
Liberal Democratic	0	0	1
Family First	0	0	1
Democratic Labour	0	0	1

- What are the variables described by this two-way table?
- How many senators are there?
- How many were in the greens?
- How many were absent?

Marginal distribution

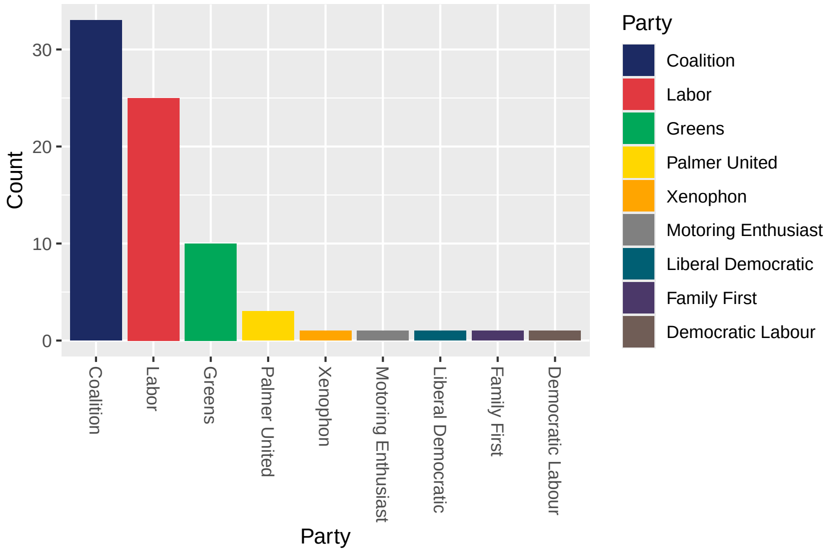
In a two-way table of counts, the **marginal distribution** or marginal sums of one of the categorical variables is the table of counts for just that variable alone, and is also the sum of values in the corresponding rows or columns of the two-way table.

> ℹ Note
>
> It's called *marginal* because it's traditionally included in the *margins* of the two-way table.

Marginal Distribution of Party

Party	Absent	Against	For	Total
Coalition	1	0	32	33
Labor	2	23	0	25
Greens	1	9	0	10
Palmer United	0	0	3	3
Xenophon	1	0	0	1
Motoring Enthusiast	0	0	1	1
Liberal Democratic	0	0	1	1
Family First	0	0	1	1
Democratic Labour	0	0	1	1

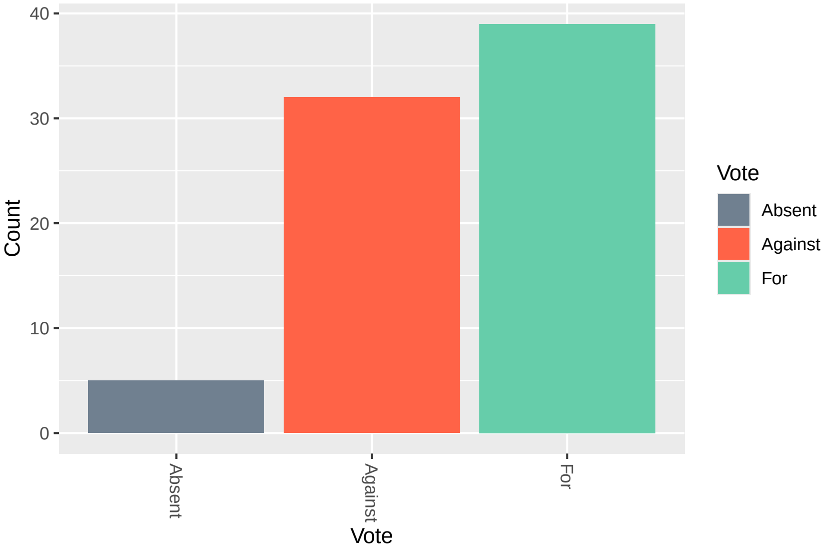
Marginal Distribution of Party



Marginal Distribution of Votes

Party	Absent	Against	For
Coalition	1	0	32
Labor	2	23	0
Greens	1	9	0
Palmer United	0	0	3
Xenophon	1	0	0
Motoring Enthusiast	0	0	1
Liberal Democratic	0	0	1
Family First	0	0	1
Democratic Labour	0	0	1
Total	5	32	39

Marginal Distribution of Votes



Both Marginal Distributions

Party	Absent	Against	For	Total
Coalition	1	0	32	33
Labor	2	23	0	25
Greens	1	9	0	10
Palmer United	0	0	3	3
Xenophon	1	0	0	1
Motoring Enthusiast	0	0	1	1
Liberal Democratic	0	0	1	1
Family First	0	0	1	1
Democratic Labour	0	0	1	1
Total	5	32	39	76

Proportions

Proportions (or equivalently percentages) can are often more informative than counts, especially when comparing groups of different sizes.

> Tip
>
> Whenever calculating proportions or percentages, always ask yourself the question — ‘of what?’.

Proportions and percentages are not exactly the same! What is the difference?

Both Marginal Distributions (Percent)

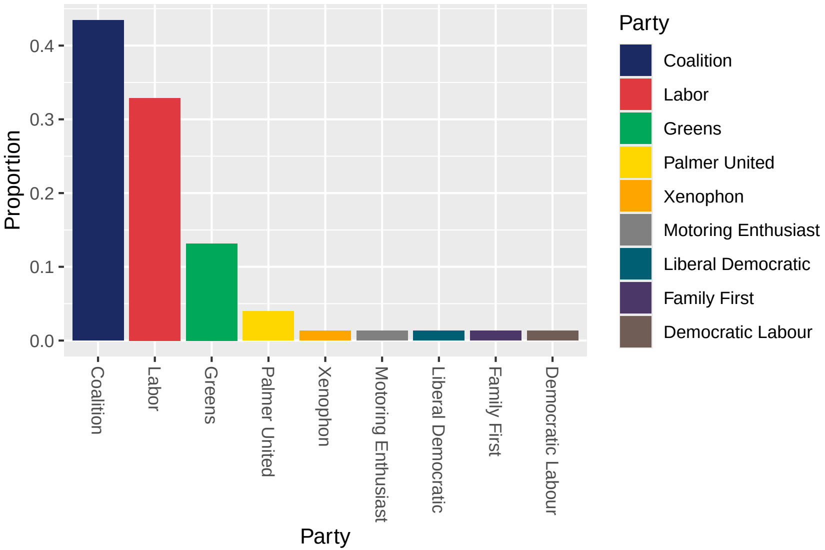
Party	Absent	Against	For	Total
Coalition	1	0	42	43
Labor	3	30	0	33
Greens	1	12	0	13
Palmer United	0	0	4	4
Xenophon	1	0	0	1
Motoring Enthusiast	0	0	1	1
Liberal Democratic	0	0	1	1
Family First	0	0	1	1
Democratic Labour	0	0	1	1
Total	7	42	51	100

What's wrong with this?

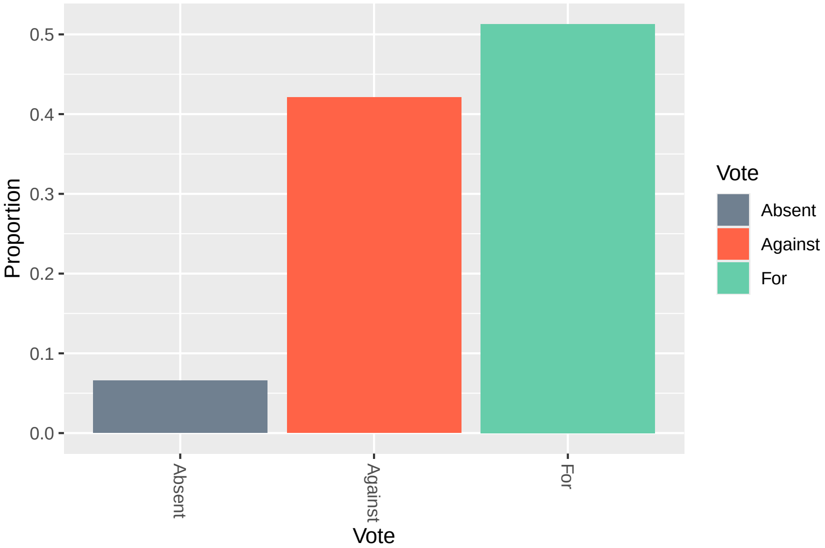
Overall Proportions or Percentages

When calculating overall percentages, in the table we get both marginals as percentages, but the corresponding bar charts will only change by relabeling of their vertical axes to differnt units:

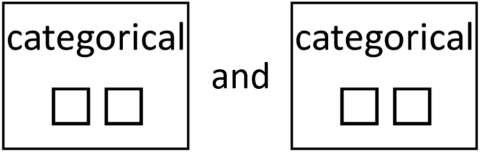
Marginal Distribution of Party (Proportion)



Marginal Distribution of Votes (Proportion)



Visualising Two Categorical Variables



categorical □□ and categorical □□

Can be visualised with variants of bar charts:

- Stacked bar charts, and
- Side-by-side bar charts

Which are variants of the regular bar charts we've shown so far for one categorical variable in which we split the bars using a second categorical variable.

Conditional Distribution

Marginal distributions tell us nothing about the relationship between two variables. For that, we need to explore the conditional distributions of the variables.

A **conditional distribution** of a variable describes the values of that variable among individuals who have a specific value of another variable.

ℹ Note

It's called *conditional* because it is the distribution of one variable when *conditioned* on another variable having a specific value.

Distribution of Votes Conditioned on Party (Percent)

Party	Absent	Against	For
Coalition	3	0	97
Labor	8	92	0
Greens	10	90	0
Palmer United	0	0	100
Xenophon	100	0	0
Motoring Enthusiast	0	0	100
Liberal Democratic	0	0	100
Family First	0	0	100
Democratic Labour	0	0	100

Distribution of Votes Conditioned on Party

If we consider the conditional distribution in counts, these are the same counts as earlier, but as percentages they are very different!

This is because the ‘of what?’ is different!

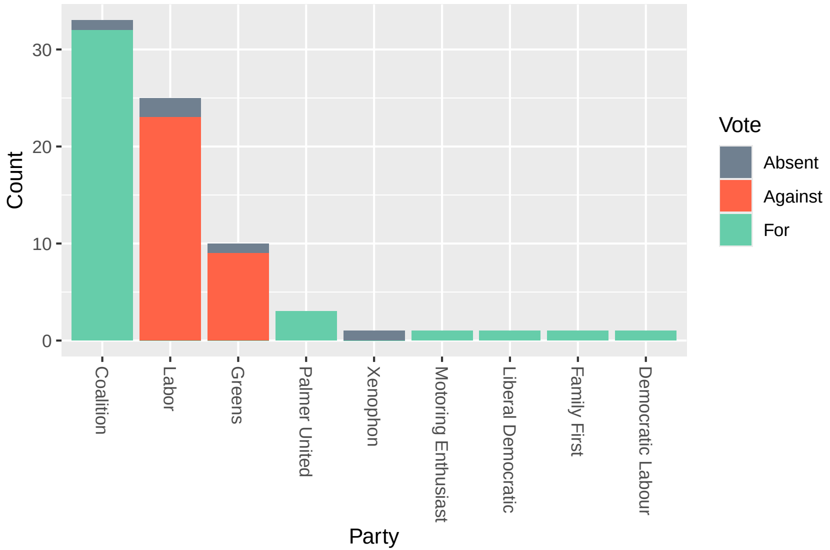
Earlier, the overall percentages were of all senators, now the percentages are of the senators of each party.

Stacked bar charts

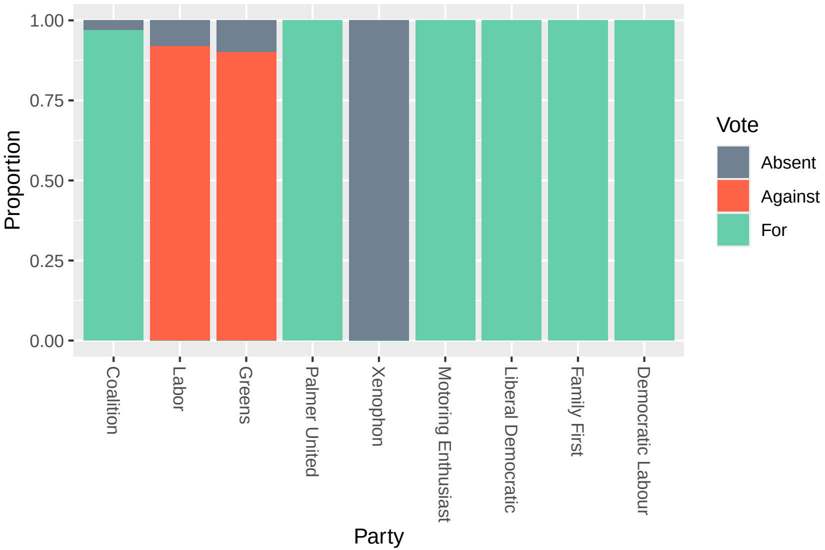
Converting between counts and proportions *overall* doesn't change the bar charts much at all.

However, converting between counts and *conditional* proportions makes a **huge** difference.

Distribution of Votes Conditioned on Party (Count)



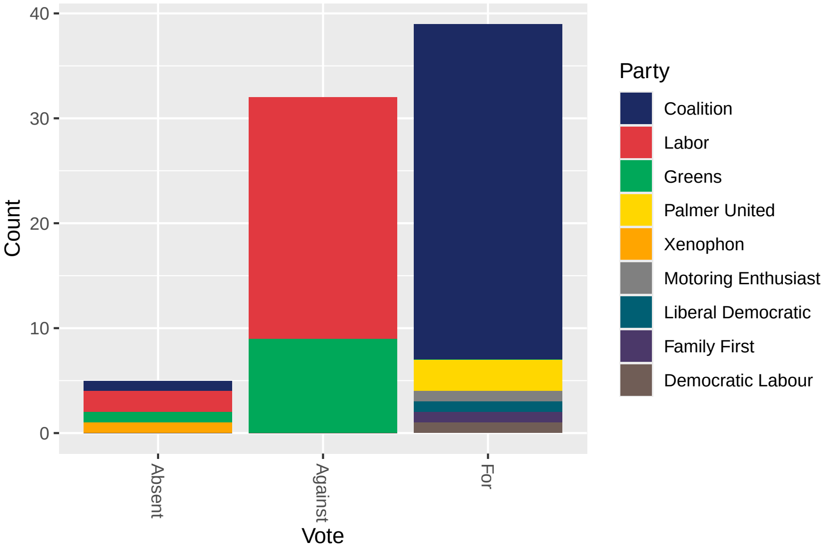
Distribution of Votes Conditioned on Party (Proportion)



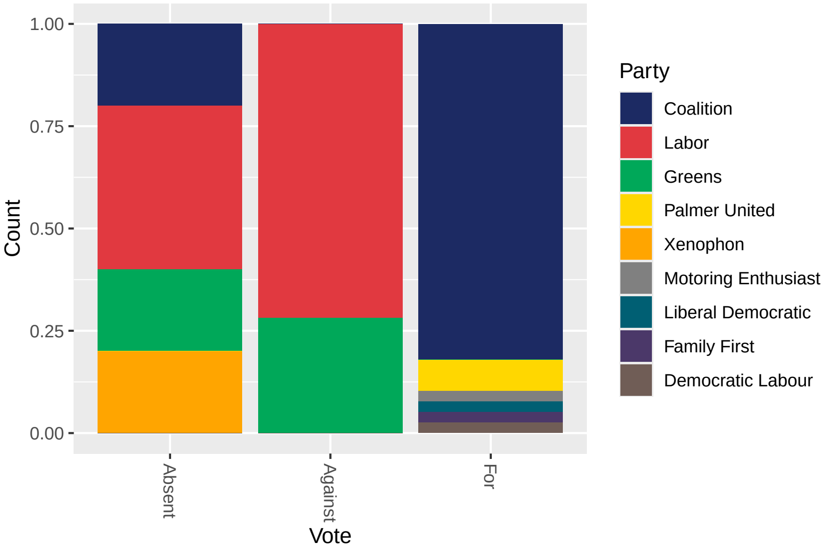
Distribution of Parties Conditioned on Vote (Percent)

Party	Absent	Against	For
Coalition	20	0	82
Labor	40	72	0
Greens	20	28	0
Palmer United	0	0	8
Xenophon	20	0	0
Motoring Enthusiast	0	0	3
Liberal Democratic	0	0	3
Family First	0	0	3
Democratic Labour	0	0	3

Distribution of Parties Conditioned on Vote (Count)



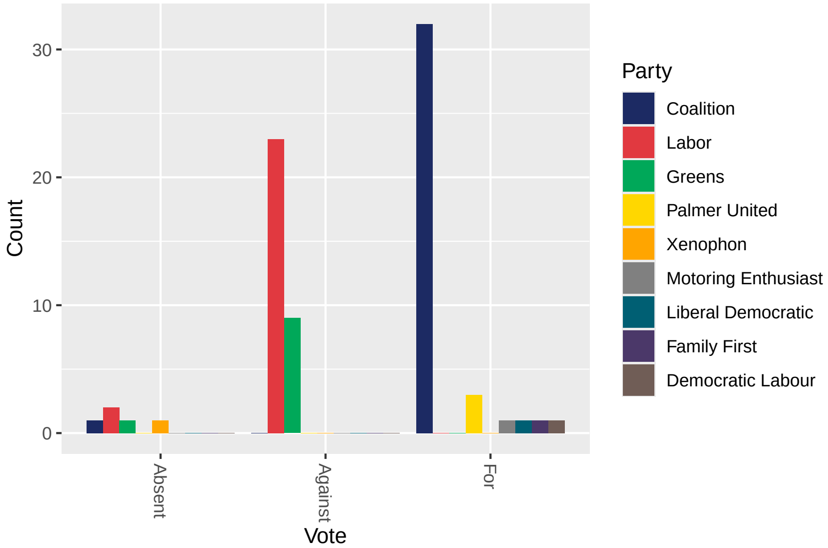
Distribution of Parties Conditioned on Vote (Proportion)



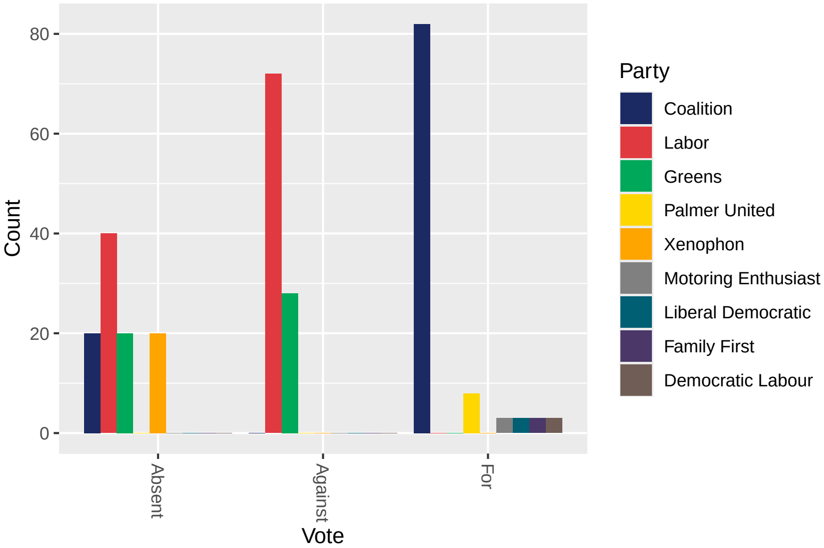
Side-by-side bar charts

Aside from Stacked bar charts as above, it is sometimes also desirable to make side-by-side bar charts instead, where the sub-divisions of each bar are stacked next to each other instead of on top of one another.

Distribution of Parties Conditioned on Vote (Count)

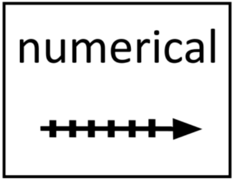


Distribution of Parties Conditioned on Vote (Percent)



One Numeric Variable

Visualising One Numeric Variable



A single numerical variable is often visualised with either

- Stem and Leaf plots.
- Boxplots, or
- Histograms,

We will mostly use histograms.

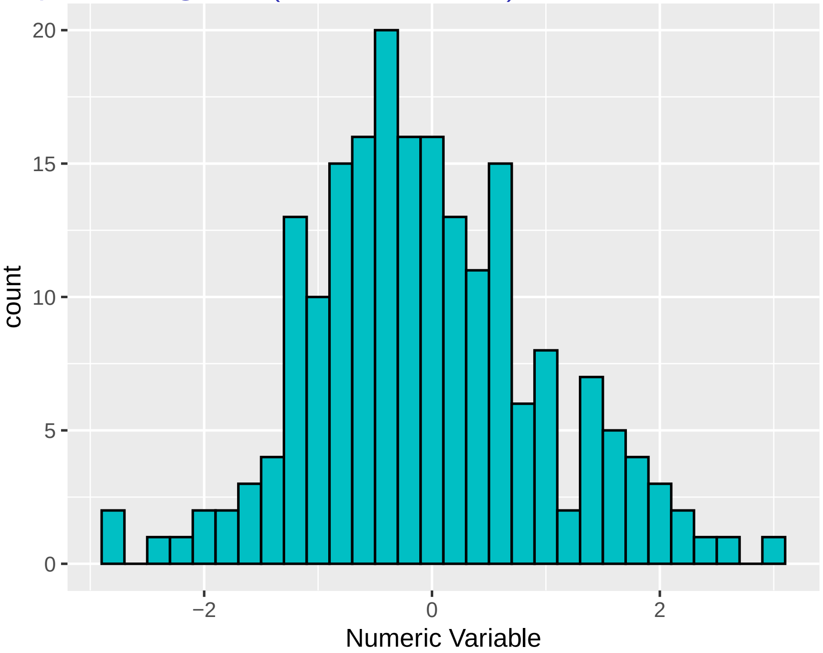
Histograms

In a histogram, the values of a numeric variable are grouped into equal-width intervals called **bins**.

A histogram looks a little like a bar chart, because by constructing intervals and counting the values in them we make categories.

Histograms are a useful way to visualise the **distribution** of data, i.e. how the measurements are distributed across different values.

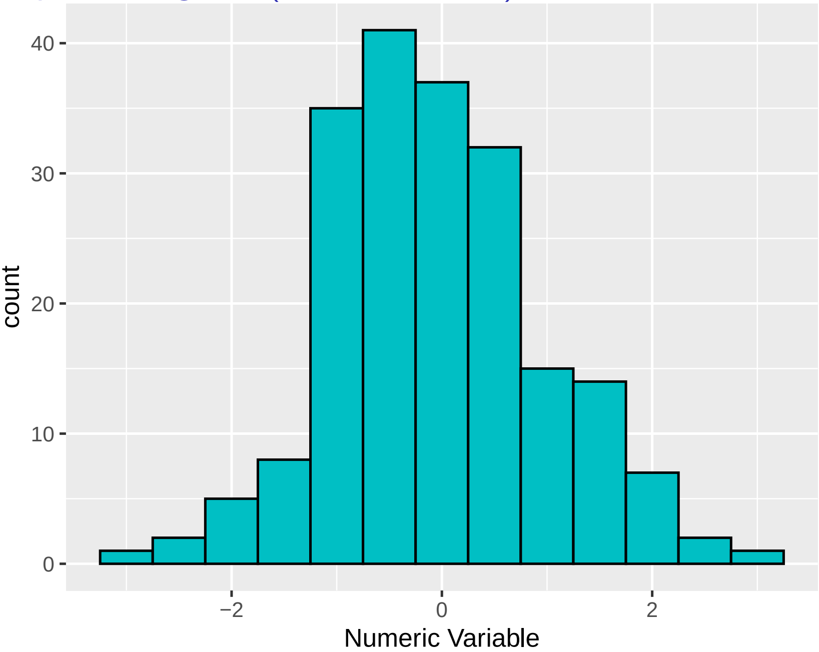
Example Histogram (Bin Width 0.2)



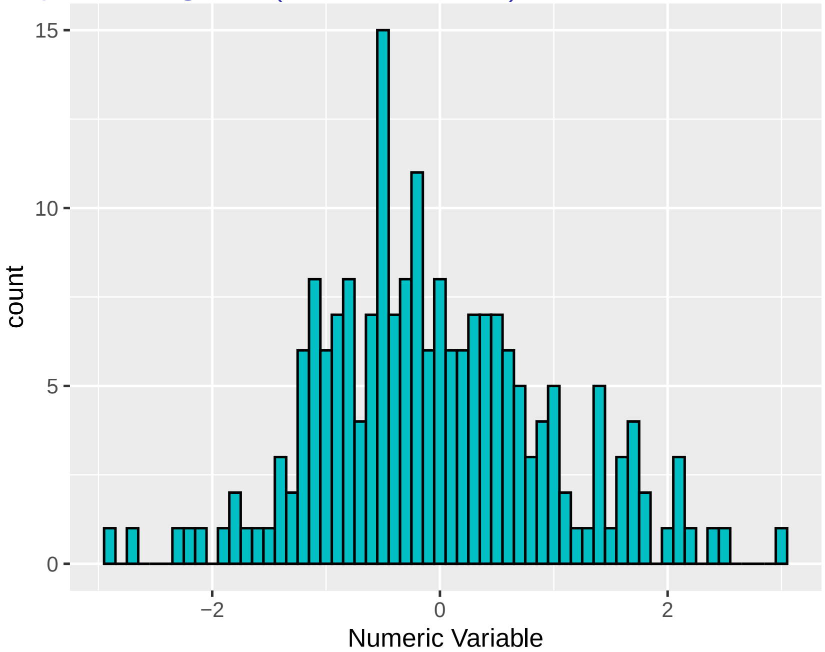
Varying the Bin Width

When constructing histograms, it is often important to try out a variety of different bin widths, as sometimes small changes to the bin width can make the visualisation look quite different and this can affect the impression you get about the distribution of the data.

Example Histogram (Bin Width 0.5)



Example Histogram (Bin Width 0.1)



Perspective

None of these are wrong in the sense that they all correctly represent the same data, but they show very different things.

A good analogy for this is to think about what you would see in a photo if you were looking at it from accross the room, compared to if you were looking at it close up — you would see different things, because you are observing features of the same picture at different levels of detail

Histogram Variants

Although they look very different to a histogram visually,

- Stem and leaf plots, and
- Dot plots

can be thought of as variants of histograms, and are used for the same thing — to describe the distribution of a numeric variable.

We won't use this much in this course, but they are worth being familiar with in case you run into them in the future.

Stem and Leaf Plots

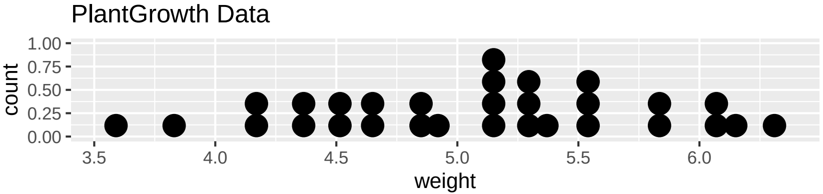
Are sometimes used to efficiently record data by hand while in the field. This example of a stem and leaf plot of the `PlantGrowth` `weight` data, and effectively uses intervals of width 1.

```
The decimal point is at the |

3 | 68
4 | 2234
4 | 5567899
5 | 11223334
5 | 55689
6 | 0123
```

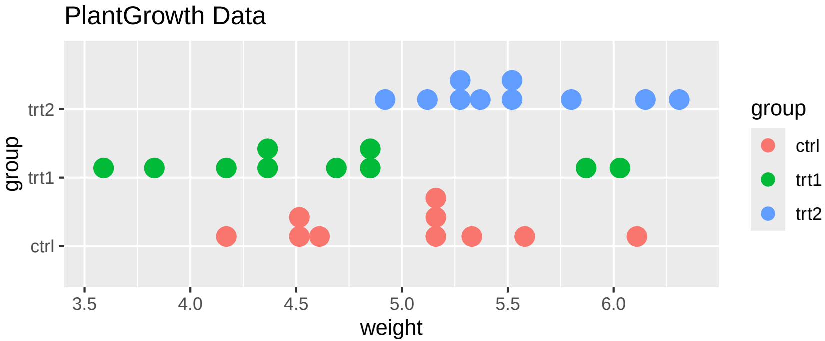
Dot Plots

Are sometimes a useful variant of a histogram for very small samples.



Dot Plots

While histograms can sometimes be difficult to compare, one of the advantages of dot plots is that they can be more easily used to compare groups.



Describing Distributions

Often we will be concerned with the overall distribution of the data rather than fine details.

There are a number of characteristics of an overall distribution that you can describe:

- Centre
- Spread
- Shape
- Outliers

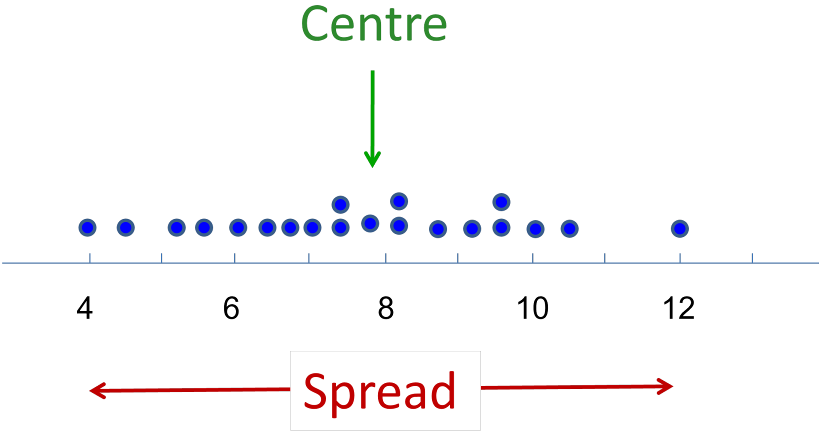
Centre and Spread

Centre describes the middle of the data, where the most common values sit.

Spread describes how far from the middle of the data do the data distribute themselves, are they all close to the middle, or is it common for data to be far from the middle?

We'll introduce different measures for centre and spread later on in this topic, but for now we just need a general sense for these ideas.

Centre and Spread

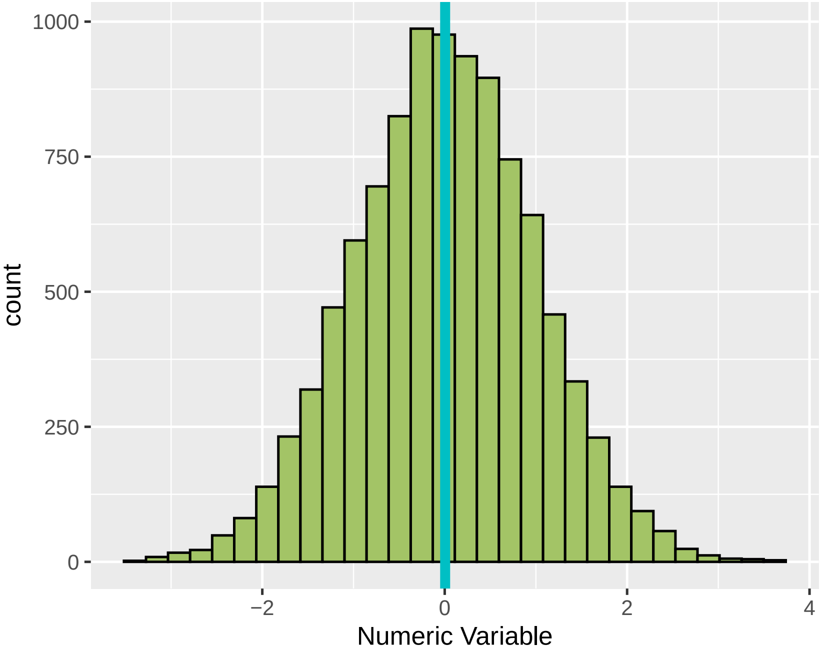


Shape

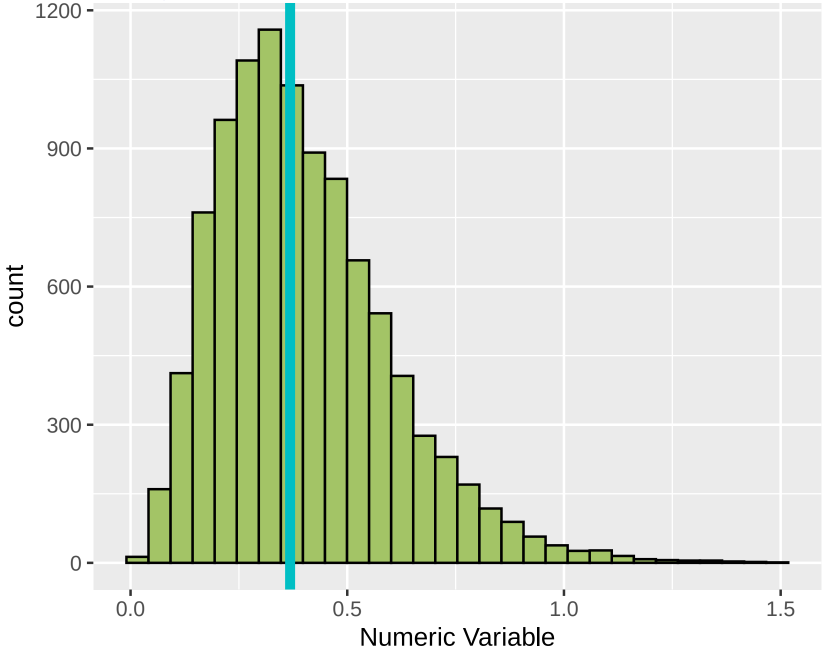
When describing the shape of a distribution, one of the things we look for is **symmetry**, if you reflect the histogram through a vertical line drawn in the middle of the data, do the two sides look similar?

When a distribution is not symmetric, it will often be **skewed**, which is where one side is more spread out than the other. If the left is more spread out than the right, then it is called **left-skewed** and if it's the other way around it's called **right-skewed**.

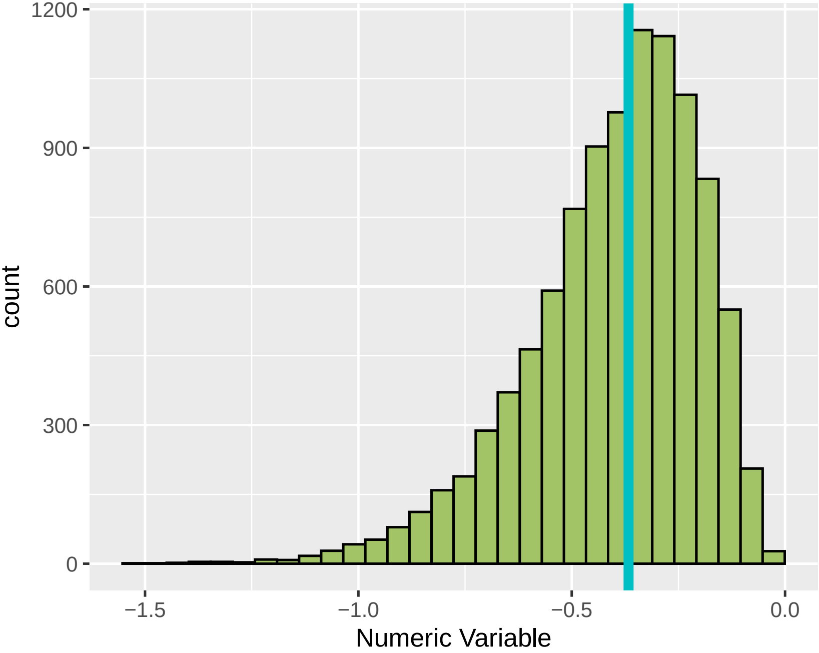
Example - Symmetric



Example - Right-Skewed



Example - Left-Skewed

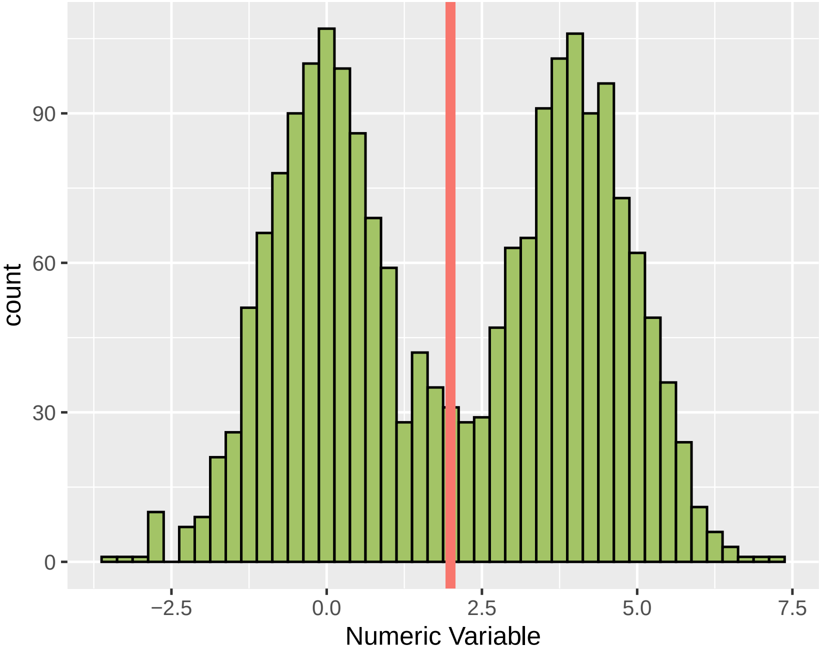


Bimodality

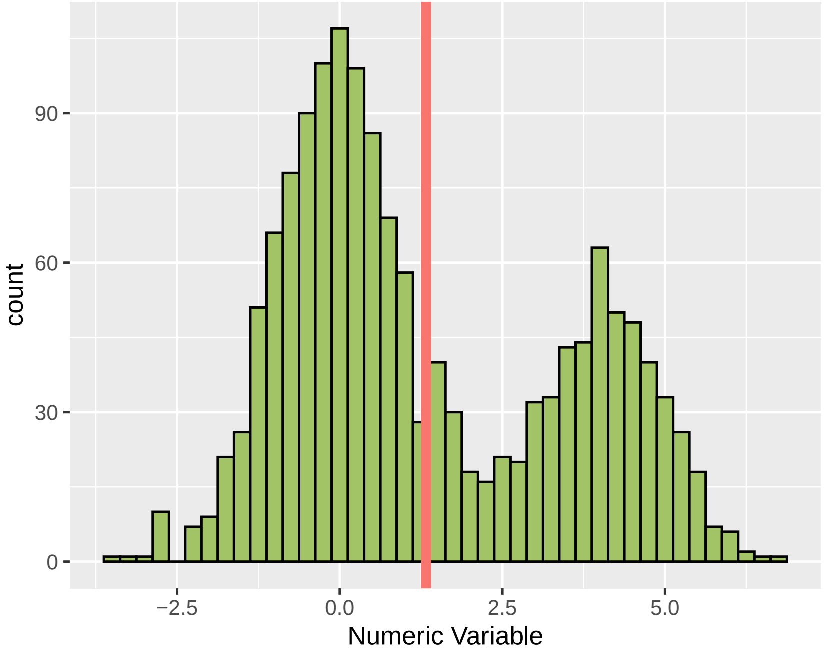
Aside from symmetry and skewness, **bimodality** is another feature we can describe the shape of a distribution by.

A **mode** is a “peak” in the histogram, and “bi” refers to two, sometimes there can be even more than two modes, but be careful sometimes a distribution can just look bimodal because we’ve used a small binwidth, so keep that in mind.

Example - Symmetric Bimodality



Example - Skewed Bimodality



Outliers

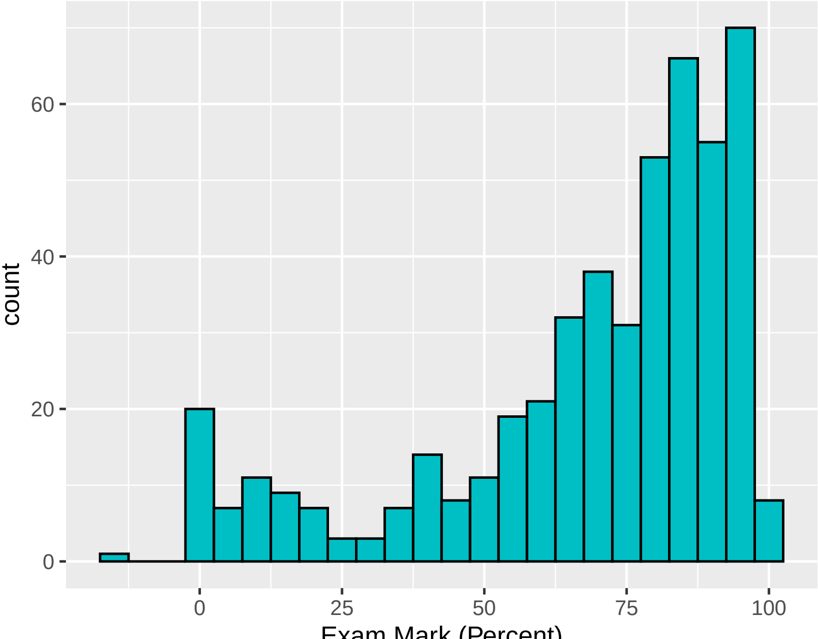
An important thing to note when describing data is any potential **outliers**. Outliers are observations are unusual, usually by being much further away from the centre than the rest of the data.

Always be on the lookout for outliers, and try to explain them, but

Do not remove them or correct them unless you are first able to confirm that they are a typo or an error.

If you do, you will be potentially ommitting important information and biasing your analysis!

Example - Exam Marks



Summarising One Numeric Variable

Can be summarised by calculating numbers called summary statistics that give a numeric indication of the different aspects of a distribution including

- its centre,
- its spread,
- its shape,

Centre

There are a number of statistics that can be used to measure the centre of a numeric variable. We will use the two most common:

- the mean, and
- the median.

Centre - Mean

The arithmetic **mean**, also called the average, is perhaps the most common measure of centre.

To find the mean $\bar{x}$ (pronounced “x-bar”) of a set of observations, add their values, and divide by the number of observations. If the n observations are $x_1, x_2, x_3, \dots, x_n$, their mean is:

$$\bar{x} = \frac{\text{sum of observations}}{n} = \frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

Example - Mean

Calculate the average for the following data set:

$$x_1 = 1, \quad x_2 = 6, \quad x_3 = 4, \quad x_4 = 3, \quad x_5 = 4, \quad x_6 = 9$$

Example - Mean - Solution

$$\bar{x} = \frac{1+6+4+3+4+9}{6} = 4.5$$

Centre - Median

One of the limitations of the mean is that it is heavily influenced by outliers, since they contribute disproportionately to the sum of values.

The **median** is another common measure of centre that is much less affected by outliers.

The median is the number for which half the data are less, and half more, than the median.

It can be calculated by sorting the data in ascending order and finding the middle value.

Example - Median

Using the same example, calculate the median for the data set

$$1, \quad 6, \quad 4, \quad 3, \quad 4, \quad 9$$

Example - Median - Solution

First we sort the data in ascending order:

$$1, \quad 3, \quad 4, \quad 4, \quad 6, \quad 9$$

In this case there is no middle value, as the middle of the data lies between the third (4) and fourth (4) values, we the median will be halfway between them: 4.

The Mean, Median, and Outliers

To see the influence outliers can have on the mean, first consider the data consisting of four values:

[1,2,3,4]

- What is the mean?
- What is the median?
- Now change 4 to 400.
- What is the mean? What is the median?

Exercise - Solution

The mean is

$$\bar{x} = \frac{1+2+3+4}{4} = \frac{10}{4} = 2.5$$

The median is halfway between 2 and 3, so also 2.5.

However if we replace 4 with 400, the median is unchanged — it is still 2.5, and the mean becomes

$$\bar{x} = \frac{1+2+3+400}{4} = \frac{406}{4} = 101.5$$

The Mean and Media

The mean and median measure centre in different ways and so can be used to answer different kinds of quesitons about the centre of data.

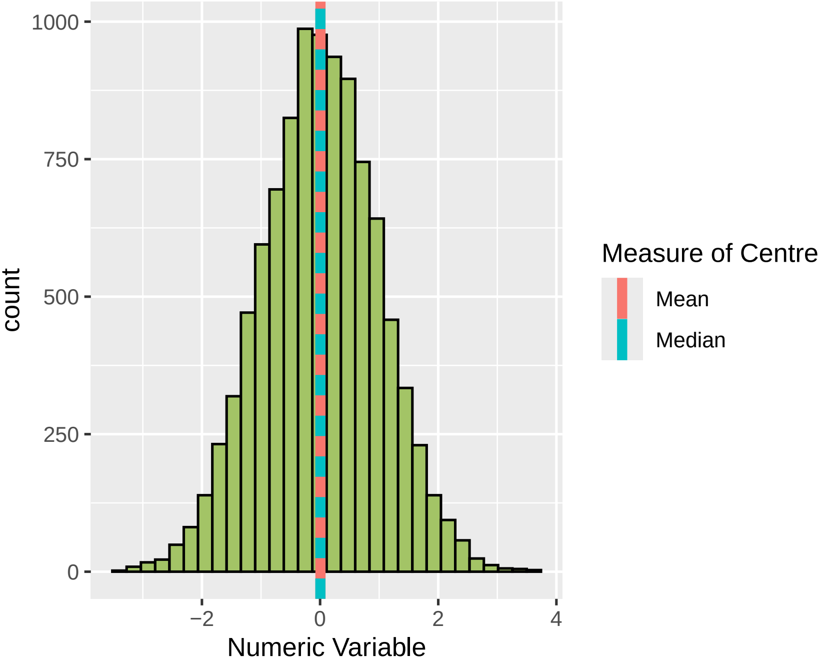
The Mean, Median, and Symmetry

Together the mean and the median can also give an indication of symmetry or skewness:

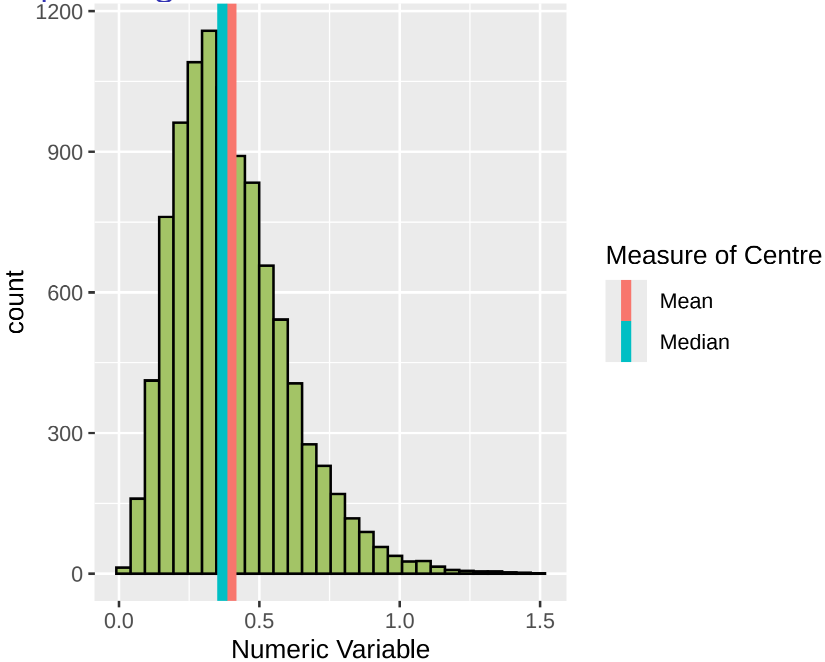
- If a distribution is fairly symmetric, the mean and median will be similar.
- If the distribution is perfectly symmetric, the mean and median will be exactly the same.
- In a skewed distribution, the mean will be closer to the side that is more spread out, so in a left-skewed distribution the mean will be less than the median and in a right-skewed distribution the mean will be greater than the median.

Earlier we showed a line on the example histograms to indicate the center of the data, but for some we used the mean and for others we used the median!

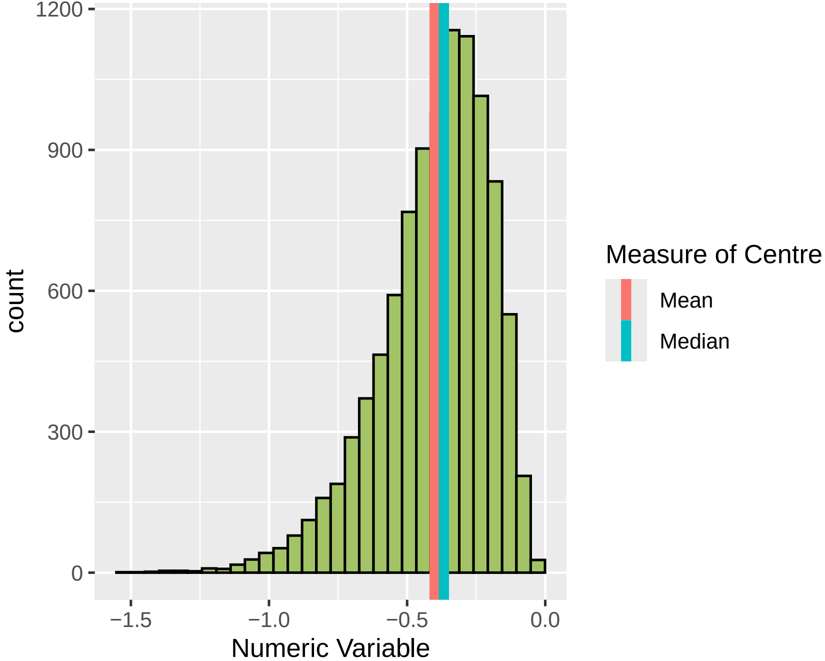
Example - Symmetric



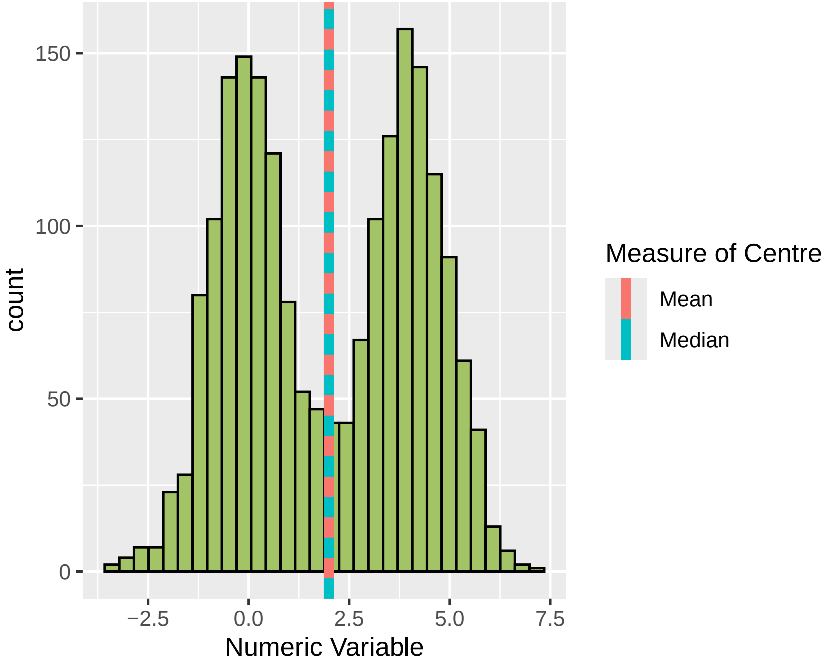
Example - Right-Skewed



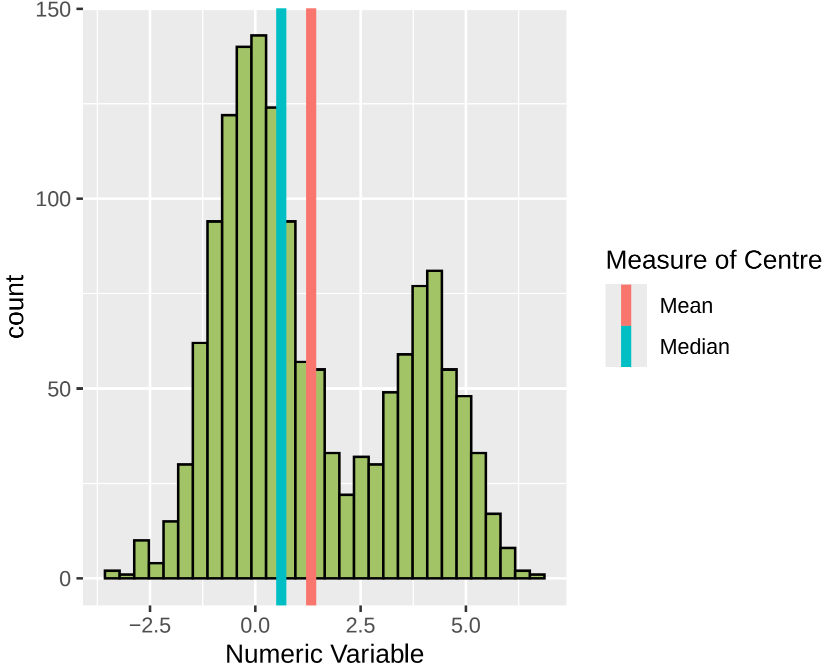
Example - Left-Skewed



Example - Symmetric Bimodality



Example - Skewed Bimodality



Spread

Centre alone can be misleading since it gives no indication of how close to that value the data are. A useful numerical description of a distribution requires both a measure of centre and a measure of spread.

There are a number of statistics that can be used to measure the spread of a numeric variable, some common ones include:

- Variance,
- Standard deviation,
- Range, and
- Interquartile range or IQR.

We will mostly use standard deviation and IQR.

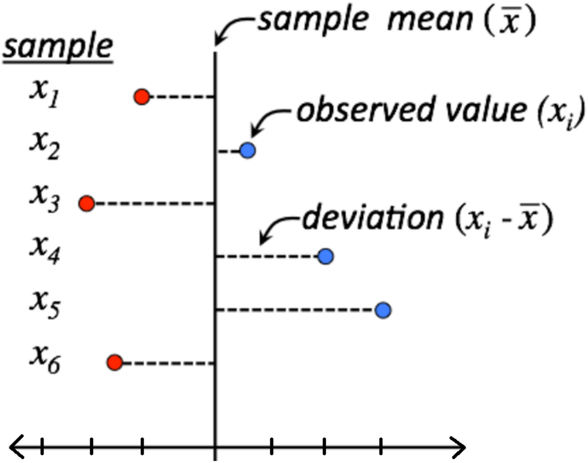
Variance and Standard Deviation

The most common measure of spread is the **standard deviation**.

The standard deviation attempts to measure what a typical distance from the mean is amongst the data, and is calculated from the **variance**.

The variance can be loosely thought of as the *average squared deviation from the mean*, and the standard deviation is the square root of the variance.

Variance and Standard deviation


sample mean ($\overline{x}$)
sample
x_1
x_2
x_3
x_4
x_5
x_6
observed value (x_i)
deviation (x_i - $\overline{x}$)

Variance and Standard Deviation

We won't go any further into the detail of how variance and standard deviation are calculated, instead we will rely on using software to calculate them for us.

However, the general concept is important in order to be able to interpret these values appropriately.

Quantiles

The median is an example of a **quantile**. Quantiles are values such that a given proportion of the data is less than them. Quantiles have different names depending on how they are measured. In particular

- Quartiles are quantiles measured in quarters,
- Percentiles are quantiles measured in percentages.

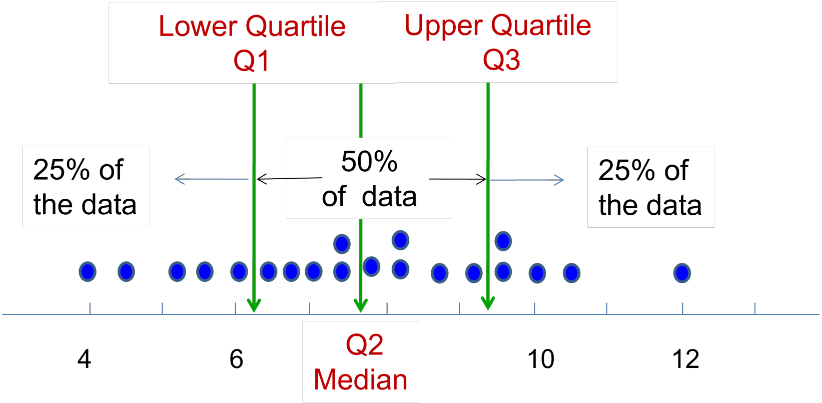
Quantiles - Examples

For example, the 20th percentile of the data is the value that 20% of the data is less and 80% is more than it.

The first quartile, also called Q1, is the value such that one quarter of the data is less and three quarters are more than it, and could also be called the $25th$th percentile.

The median could also be called the second quartile, or the 50th percentile.

Quartiles

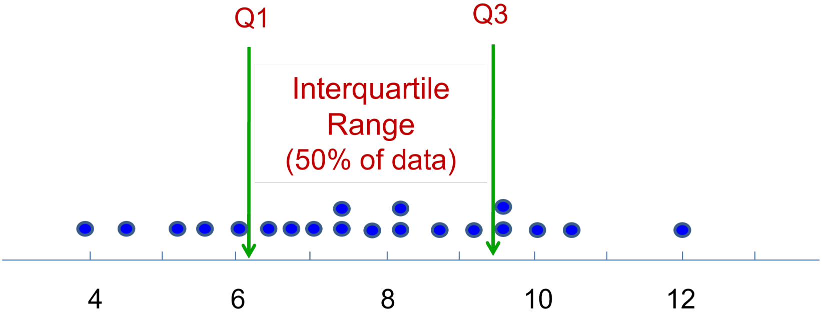

Lower Quartile
Q1
Upper Quartile
Q3
25% of
the data
50%
of data
25% of
the data
4
6
Q2
Median
10
12

Interquartile Range (IQR)

The **interquartile range** or IQR is the distance between the first (Q1) and third (Q3) quartiles of the data, and so can be calculated by subtracting Q1 from Q3:

$$IQR = Q3 - Q1$$

Interquartile Range (IQR)



Calculating IQR

Similar to standard deviation, while there are further details on how to calculate quantiles and IQR, we will rely on using R to do these calculations and are more concerned with the interpretation.

Range

Another commonly reported measure of spread is the **range**, which is the distance between the smallest value (called the **minimum**) and the largest value (called the **maximum**).

Between the IQR and the range, which do you think would be more affected by outliers?

The Five-Number Summary

Often a numeric variable is quickly summarised with five-numbers called the **five-number summary**. These five numbers are the: minimum, Q1, median, Q3, and maximum.

The `summary()` function in R will give us the five-number summary plus the mean:

```r
summary(PlantGrowth$weight)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.590   4.550   5.155   5.073   5.530   6.310 
```

The $1.5 \times IQR$ Rule

In addition to serving as a measure of spread, the interquartile range (IQR) is used as part of a rule of thumb for identifying outliers creatively named the 1.5 **times IQR** rule.

According to this rule of thumb, an observation is categorised as an outlier if it is more than $1.5 \times IQR$ above the third quartile or below the first quartile.

Example - New York Commute

Commuting times (in minutes) of 20 randomly selected New York workers were measured and sorted in ascending order:

10 30 5 25 40 20 10 15 30 20 15 20 85 15 65 15 60 60 40 45

The output of `summary()` on this data is

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   5.00   15.00   22.50   31.25   41.25   85.00
```

Are there any outliers in these data?

Example - New York Commute - Solution

The $Q_1 = 15$ minutes and the $Q_3 = 42.5$ minutes, and so the $IQR = 27.5$ minutes.

$1.5 \times IQR = 1.5 \times 27.5 = 41.25$, so:

- to be $1.5 \times IQR$ below the Q1, a worker would need to have a commute less than $Q_1 - 1.5 \times IQR = 15 - 41.25 = -26.25$ minutes. Impossible!
- to be $1.5 \times IQR$ above the Q3, a worker would need to have a commute more than $Q_3 + 1.5 \times IQR = 42.5 + 41.25 = 83.75$!

According to the The $1.5 \times IQR$ rule there is one outlier, with a commute of 85 minutes, but it is only just barely large enough to be categorised as an outlier by this rule of thumb.

The $1.5 \times IQR$ Rule - Disclaimer

There is nothing special about the number 1.5, and like all such arbitrary cut-off points for decision making, the fact that it is arbitrary should be kept in mind when making judgements and interpretations.

If a value is $1.4 \times IQR$ or $1.6 \times IQR$ above the Q3 or below the Q1 it will change if it is categorised as an outlier according to the rule, but it will *not* change the significance of that value very much.

Guidelines for Measures of Centre and Spread

Speaking of rules of thumb,

- The median and IQR are usually better than the mean and standard deviation for describing a skewed distribution or a distribution with outliers.
- Use mean and standard deviation only for reasonably symmetric distributions that do not have outliers.

Changing Units of Measurement

Multiplying each observation by a positive number multiplies both measures of centre (mean, median) and spread (IQR, standard deviation) by that same number.

Adding a number to each observation adds the same number to measures of centre and quantiles, but it does not change measures of spread (IQR, standard deviation).

Example

Height of 100 randomly selected students in inches has $\bar{x} = 60$, median 61, standard deviation 3, and $IQR = 5$.

If we considered the measurements in centimetres, what would the mean, median, standard deviation and IQR be? (Hint: there are 2.54 centimetres in an inch).

If every student is sensibly made to wear 2 inch platform shoes, assuming the original measurements were all made without wearing shoes, what would the mean, median, standard deviation and IQR be?

Example - Solution

To convert to centimeters, we would multiply the mean, median, standard deviation and IQR by 2.54. So the mean would be 152.4, median 154.94, standard deviation 7.62 and IQR 12.7 centimeters.

Adding two inches will change the measures of centre but not spread. So the mean would become 62, the median 63, but the standard deviation and IQR would remain unchanged at 3 and 5 inches respectively.

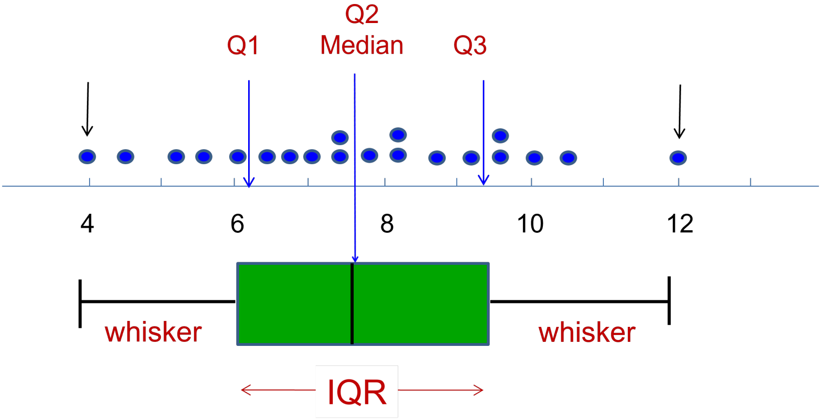
Boxplots

Aside from histograms, boxplots are another common way to visualise a numeric variable that combines the five-number summary, IQR and $1.5 \times IQR$ rule.

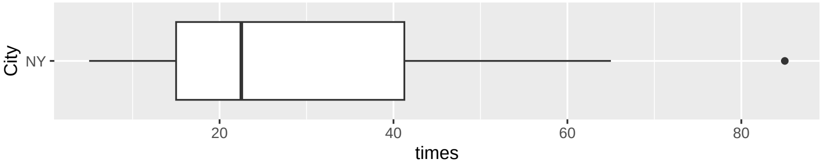
In a boxplot or a “box and whisker” plot:

- a “box” is constructed from the Q1 to the Q3, so it has a width equal to the IQR and will contain a line indicating the median in it, and
- “whiskers” are constructed going out to either side from the box that end at the mimumum and maximum of the data *excluding any outliers* as classified by the $1.5 \times IQR$ rule.

Boxplot



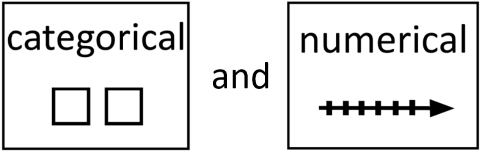
Example - New York Commute



Although a boxplot shows less information than a histogram, it also highlights and visually summarises some key summary statistics, and displays them much more concisely.

One Numeric and One Categorical Variable

One Numerical and One Categorical Variable



Can be summarised and visualised together by splitting the data into groups according to the categorical variable, and then summarising and visualising the numeric variable *within each group*.

This is often what we are most interested in doing since the **method of comparison**[1] is such a fundamental part of the scientific method.

[1] see the forum from the previous topic to recap!

Summarising One Numerical and One Categorical Variable.

When summarising one categorical and one numeric variable together, we can use all the summary statistics we've already introduced to summarise the numeric variable, but seperately for each category as defined by the categorical variable!

Example - Penguins

For example, in the penguins data a number of variables are measured on each penguin including its species and body mass (in grams). We can seperate the data into groups by species, and summarise body mass:

species	Count	Mean	Min	Med	Max	SD	IQR
Adelie	146	3706	2850	3700	4775	459	638
Chinstrap	68	3733	2700	3700	4800	384	462
Gentoo	119	5092	3950	5050	6300	501	800

what do you notice?

Example - Penguins - Solutions

Some observations:

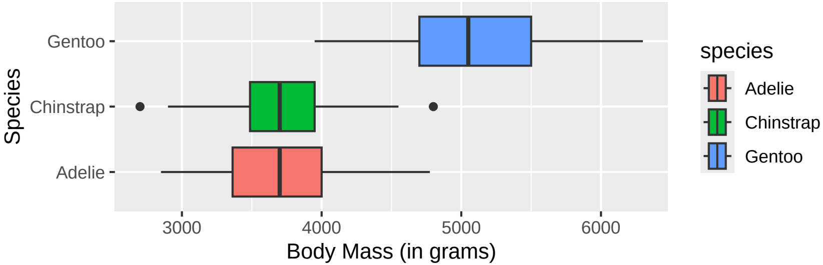
- The distributions for each species are fairly symmetric since the means and medians are similar.
- Adelie and Chinstrap have similar typical body mass by mean or median, but Adelie seem to have higher spread by both standard deviation and IQR, so it is more common to see much smaller or much larger Adelie than Chinstrap.
- Gentoo have much larger body mass both by mean and median compared to the other two species, and also have larger spread on average, but even smaller Gentoo tend to be larger than the typical penguins of the other two species despite this.

Visualising One Numerical and One Categorical Variable.

Since a boxplot summarises the distribution of a numeric variable concisely, it is perfect for visualising the comparison of groups.

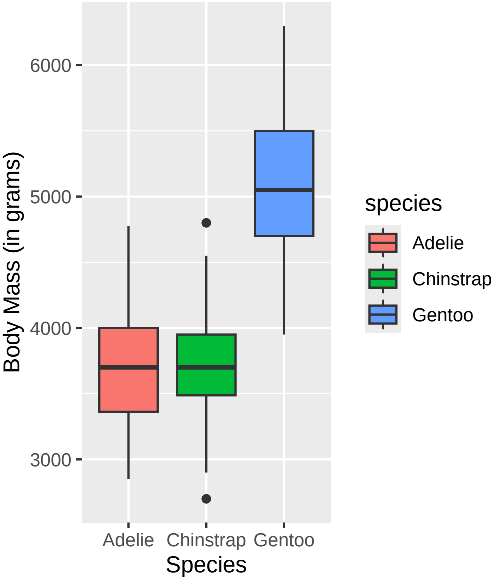
This is called a **side-by-side boxplot** and is made by producing a boxplot for each group and arranging them on a single numeric axes.

Example - Side-by-side Boxplot - Penguins



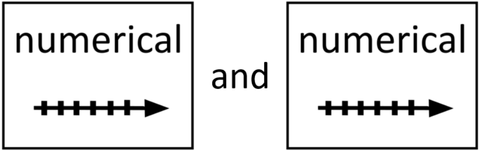
Boxplots can be arranged either horizontally or vertically, although be aware that certain scientific disciplines are very particular about arranging boxplots a certain way.

Example - Vertical Boxplots - Penguins



Two Numeric Variables

Two Numerical Variables



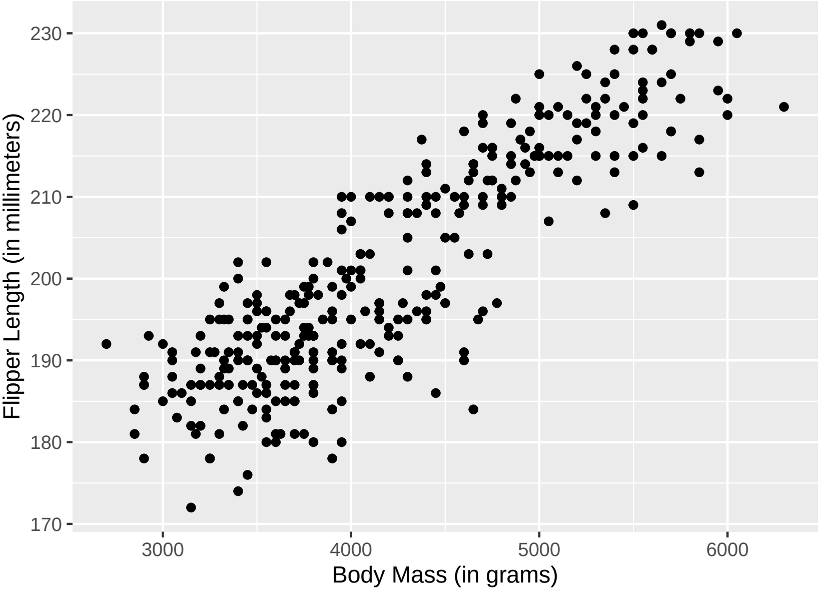
Can be visualised together with a scatterplot.

Scatterplot

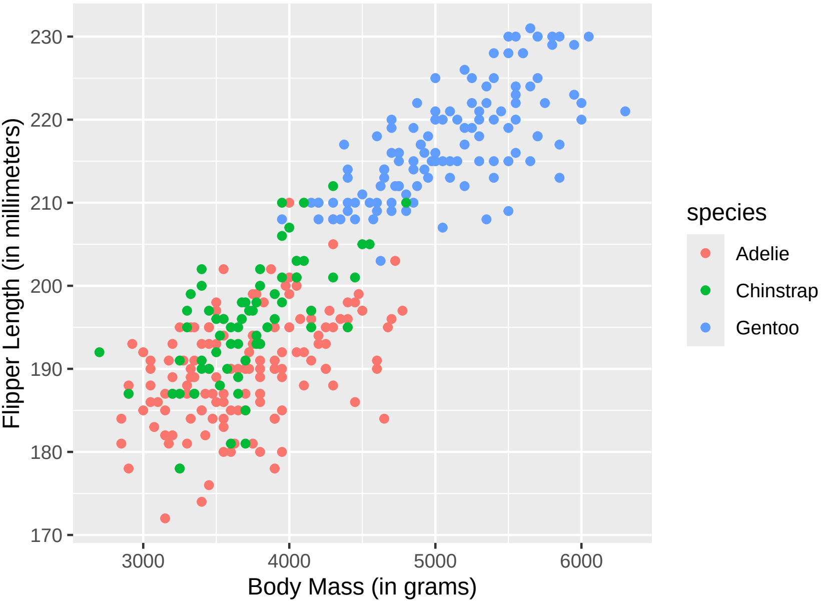
In a scatterplot, two numeric variables are used as the two axes and subjects are represented with points at the coordinates corresponding to the values they have for the two corresponding variables.

For example, in the penguins data, there are several other numeric variables in addition to body mass, such as flipper length for example.

Example - Penguins



Adding a Categorical Variable with Colour



Example - Penguins

Which highlights what we saw earlier, than the Gentoo penguins are much bigger than the other two species!

We can see from the scatterplot that as you might expect regardless of species penguins with a larger body mass tend to also have longer flippers, but there is also some individual variation with some individual penguins having a higher mass but shorter flippers, and vice versa.

This relationship between body mass and flipper length, where one tends to increase alongside the other, is an example of an association between these two numerical variables.

Association Between Variables

Two variables measured on the same subjects are **associated** if knowing the value of one of the variables tells you something that you would not otherwise know about the value of the other variable.

Describing Associations Between Numeric Variables

Much like when describing the distribution of data in a histogram, look for an overall pattern and for notable deviations from that pattern.

There are a number of characteristics of an overall pattern shown in a scatterplot that you can describe:

- Direction
- Shape
- Strength
- Outliers

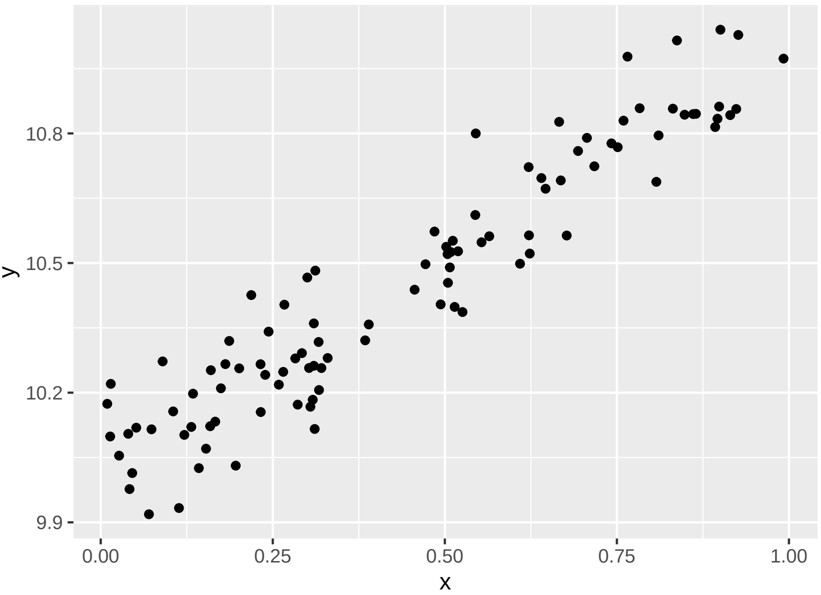
Shape and outliers are also characteristics that we describe about the distribution in a histogram, and while we use the same word because they are related concepts, they describe quite different characteristics when describing a scatterplot.

Direction

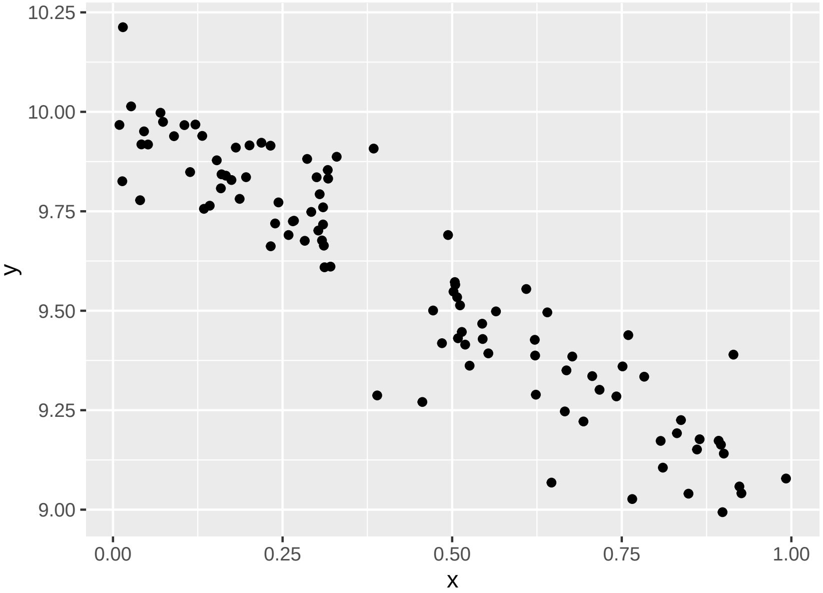
The direction of an association will either be

- **Positive:** as one variable increases, so does the other variable, or
- **Negative:** as one variable increases, the other variable decreases.

Example - Positive Association



Example - Negative Association



Shape

There are many different shapes than an association can take, but or now we will put them into two categories:

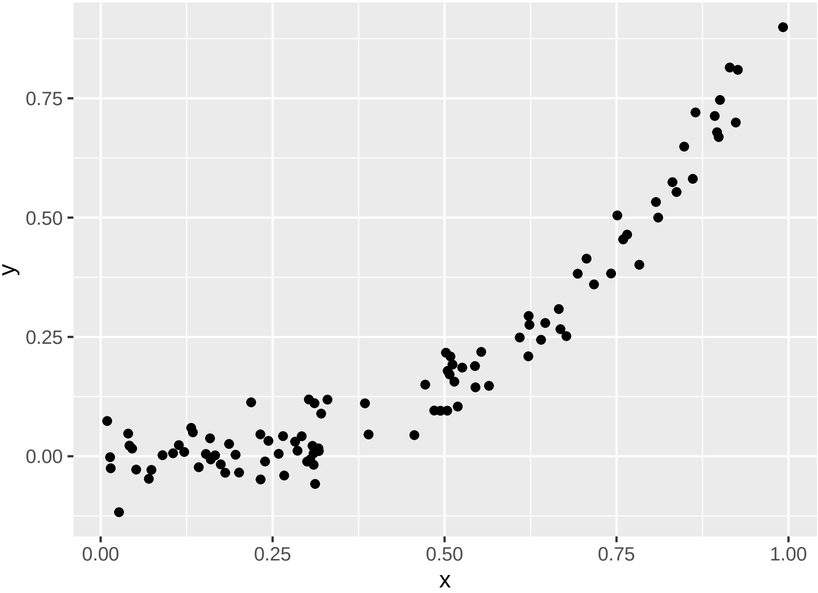
- Linear: straight-line associations, and
- Non-linear: everything else.

Non-linear Associations

The examples we've seen so far have all shown linear associations. So what do non-linear associations look like?

We'll show a few examples, but bear in mind that non-linear associations can look like many different things, not just the examples we'll show. The only thing they will all have in common is that they don't follow a single straight line.

Example Non-linear association



Strength

The strength of an association is related to how the data are spread around the line that describes the overall pattern, regardless of if that line is straight (in a linear association) or not (in a non-linear association).

- A **strong** association is one where the data is tightly clustered around the line, and
- A **weak** association is one where the data is spread widely around the line.

These are subjective terms, and there is a spectrum of strength between them.

The previous example illustrates a moderate to high strength non-linear association, due to how clustered along the non-linear line the data are.

Summarising Two Numeric Variables

We won't try to quantify the strength of non-linear associations beyond describing their scatterplot using words.

However our eyes and intuition are not always good judges of how strong a relationship is.

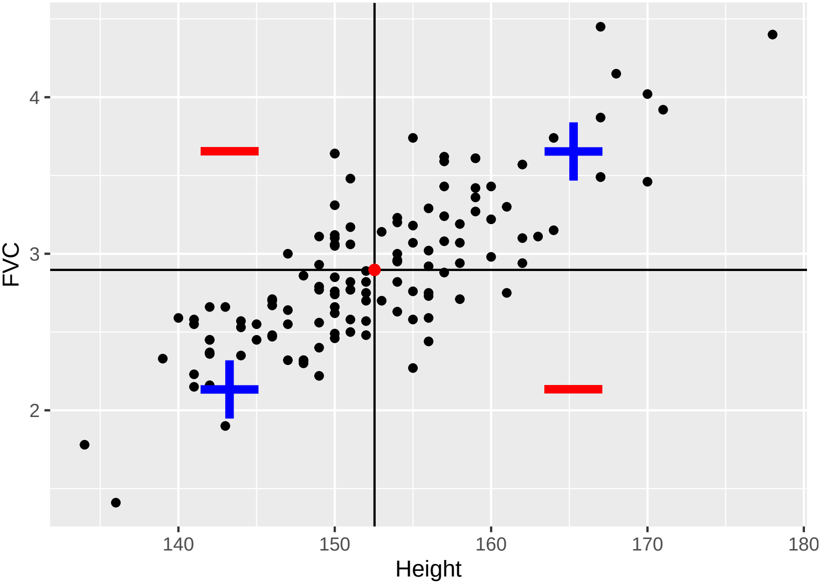
So for linear associations, we use correlation to summarise the strength (and direction) of a linear association between two numeric variables.

Correlation

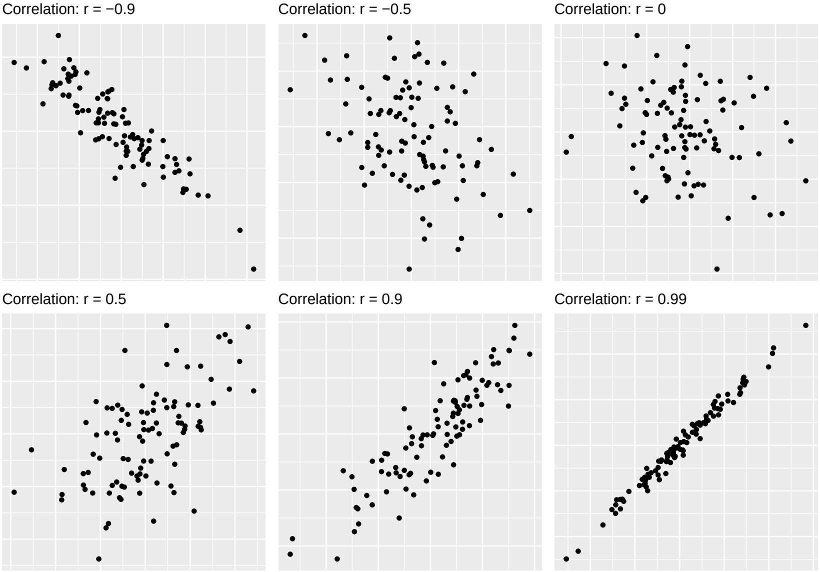
The **correlation**, typically written as r measures the strength (and direction) of a linear relationship between two numeric variables and has the following properties:

- r is always a number between -1 and 1.
- $r > 0$ indicates a positive association.
- $r < 0$ indicates a negative association.
- Values of r near 0 indicate a very weak linear relationship.
- The strength of the linear relationship increases as r moves away from 0 toward -1 or 1.
- The extreme values $r = -1$ and $r = 1$ occur only in the case of a perfect linear relationship, where the data lie precisely on a straight line.

Correlation - How does it work



Examples - Correlation



Correlation: r = −0.9

Correlation: r = −0.5

Correlation: r = 0

Correlation: r = 0.5

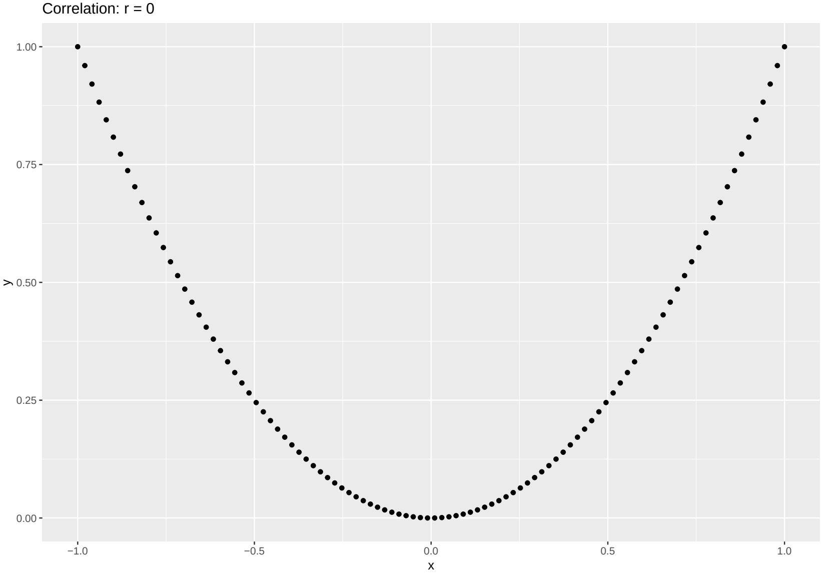
Correlation: r = 0.9

Correlation: r = 0.99

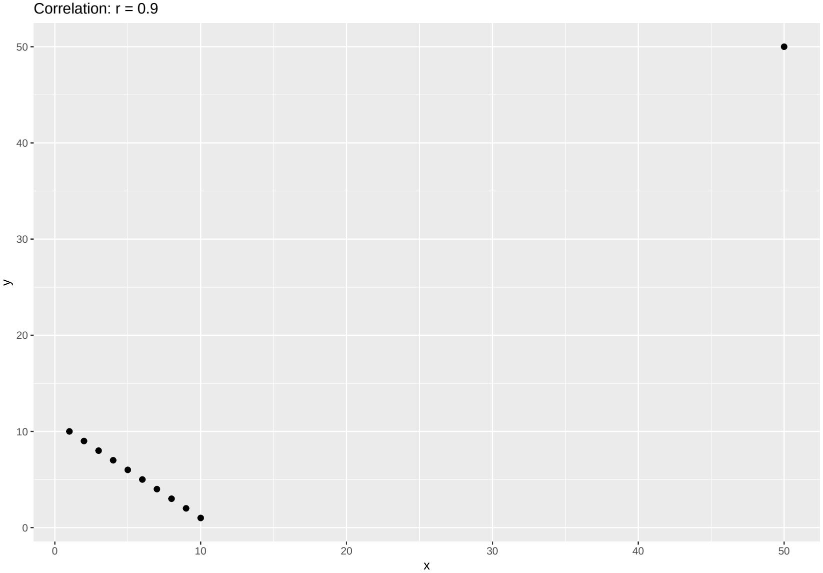
Correlation - Warnings

- Correlation requires that both variables be numeric.
- Correlation does not describe the strength of curved relationships between variables, no matter how strong the relationship is.
- The correlation r, like the mean and standard deviation, can be strongly affected by a outliers.
- Correlation is not a complete summary of two-variable data, looking at a scatterplot alongside the correlation is always important, as the scatterplot will allow you to see:
    - if the association is linear or not,
    - the presence of outliers, and
    - other features of the association the correlation will not reflect.

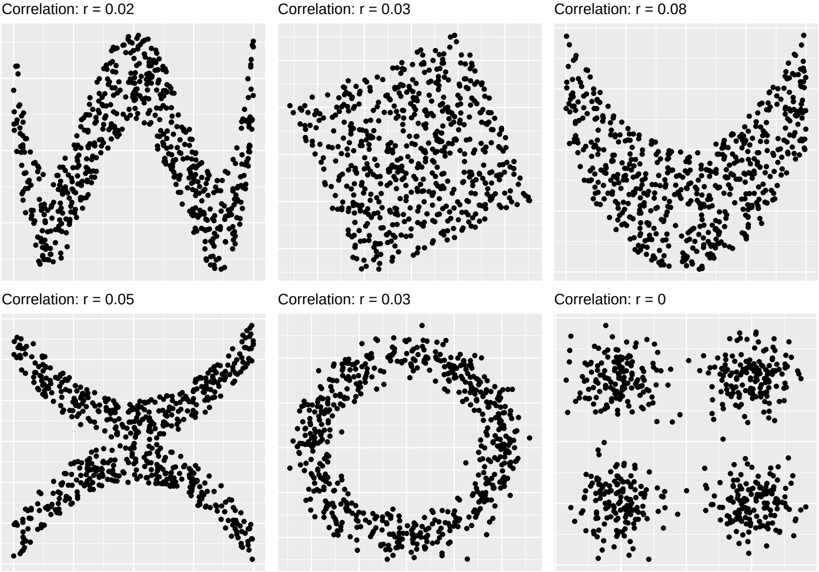
Example - Pitfalls of Correlation



Example - Pitfalls of Correlation



More Examples with Nearly Zero Correlation[2]



Correlation: r = 0.02

Correlation: r = 0.03

Correlation: r = 0.08

Correlation: r = 0.05

Correlation: r = 0.03

Correlation: r = 0

[2] Attribution DenisBoigelot Wikipedia released to public domain

Multiple Variables

One of the strengths of data visualisation is the ability to show multiple variables together easily

This can be done by introducing aesthetic elements like

- Colour by using distinct colours for categories, or a scale for numerical variables,
- Shape by using different shapes to indicate categories
- Size by scaling visual elements according to a numeric variable (like in the Causes of Mortality figure earlier!)
- and more.

But there are also different geometries of visualisation that can be used to visualise particular combinations of variable types.

More than Two Variables

We won't discuss this too much in this course, but some cases are straightforward, for example

- Adding coloured categories to a scatterplot of two numerical variables, or
- Visualising a third numerical variable represented by the size of dots in a scatterplot

But in general visualising too many variables at once can quickly make a visualisation confusing and difficult to interpret.