MTHM501:Working with Data

Mark Kelson, m.j.kelson@exeter.ac.uk

Overview

- Online
- 2 practicals per week - each 2 hours
- Self-directed learning outside of that
  - 10 hours of lectures
  - 20 hours of practicals
  - 36 hours of background reading
  - 84 hours guided independent study - assessments
- Student hour: Wednesday 9-10 (changed since the video!)- Book in advance
- Assessment:
  - 4 formative courseworks
  - 1 project (10 pages)
  - 100% weighting for the project

Aims

- to bring everyone up to speed on the fundamentals of analysis and statistical thinking
- to establish core competencies in RStudio
- to instil confidence in your ability to handle data
- to encourage rigour in your analytical approaches
- to exercise your self-directed learning skills

Topics

- Data- types, wrangling, coding
- RStudio and Markdown
- Visualisation
- Maps
- Missing data
- Uncertainty in data
- P-values, communication

VISUALISATION

- Data visualisation is the presentation of data in a graphical format.
- It can provide a valuable insight into your data and help in identifying patterns.
- Numerous methods are available to visualise your data
  - bar charts
  - pie charts
  - scatterplots
  - histograms
  - box plots
  - line plots
  - maps
  - ... and many more!

VISUALISATION

Deviation

Emphasise variations (+/-) from a fixed reference point. Typically the reference point is zero but it can also be a target or a long-term average. Can also be used to show sentiment (positive/neutral/negative).

Example FT uses
Trade surplus/deficit, climate change

Diverging bar

A simple standard bar chart that can handle both negative and positive magnitude values.

Diverging stacked bar

Perfect for presenting survey results which involve sentiment (eg disagree/neutral/agree).

Spine

Splits a single value into two contrasting components (eg male/female).

Surplus/deficit filled line

The shaded area of these charts allows a balance to be shown – either against a baseline or between two series.

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (ie. one causes the other).

Example FT uses
Inflation & unemployment, income & life expectancy

Scatterplot

The standard way to show the relationship between two continuous variables, each of which has its own axis.

Line + Column

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.

Ranking

Use where an item's position in an ordered list is more important than its absolute or relative value. Don't be afraid to highlight the points of interest.

Example FT uses
Wealth, deprivation, league tables, constituency election results

Ordered bar

Standard bar charts display the ranks of values much more easily when sorted into order.

Ordered column

See above.

Ordered proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Dot strip plot

Dots placed in order on a strip are a space-efficient method of laying out ranks across multiple categories.

Slope

Perfect for showing how ranks have changed over time or vary between categories.

Lollipop

Lollipops draw more attention to the data value than standard bar/column and can also show rank and value effectively.

Bump

Effective for showing changing rankings across multiple dates. For large datasets, consider grouping lines using colour.

Distribution

Show values in a dataset and how often they occur. The shape (or 'skew') of a distribution can be a memorable way of highlighting the lack of uniformity or equality in the data.

Example FT uses
Income distribution, population (age/sex) distribution

Histogram

The standard way to show a statistical distribution – keep the gaps between columns small to highlight the 'shape' of the data.

Boxplot

Summarise multiple distributions by showing the median (centre) and range of the data

Violin plot

Similar to a box plot but more effective with complex distributions (data that cannot be summarised with simple average).

Population pyramid

A standard way for showing the age and sex breakdown of a population distribution; effectively, back to back histograms.

Dot strip plot

Good for showing individual values in a distribution, can be a problem when too many dots have the same value.

Dot plot

A simple way of showing the change or range (min/max) of data across multiple categories.

Barcode plot

Like dot strip plots, good for displaying all the data in a table, they work best when highlighting individual values.

Cumulative curve

A good way of showing how unequal a distribution is: y axis is always cumulative frequency, x axis is always a measure.

Frequency polygons

For displaying multiple distributions of data. Like a regular line chart, best limited to a maximum of 3 or 4 datasets.

Change over Time

Give emphasis to changing trends. These can be short (intra-day) movements or extended series traversing decades or centuries. Choosing the correct time period is important to provide suitable context for the reader.

Example FT uses
Share price movements, economic time series

Line

The standard way to show a changing time series. If data are irregular, consider markers to represent data points.

Column

Columns work well for showing change over time – but usually best with only one series of data at a time.

Line + column

A good way of showing the relationship over time between an amount (columns) and a rate (line).

Stock price

Usually focused on day-to-day activity, these charts show opening/closing and high/low points of each day.

Slope

Good for showing changing data as long as the data can be simplified into 2 or 3 points without missing a key part of story.

Area chart

Use with care – these are good at showing changes to total, but seeing change in components can be very difficult.

Fan chart (projections)

Use to show the uncertainty in future projections - usually this grows the further forward to projection.

Connected scatterplot

A good way of showing changing data for two variables whenever there is a relatively clear pattern of progression.

Calendar heatmap

A great way of showing temporal patterns (daily, weekly, monthly) – at the expense of showing precision in quantity.

Priestley timeline

Great when date and duration are key elements of the story in the data.

Circle timeline

Good for showing discrete values of varying size across multiple categories (eg earthquakes by continent).

Vertical timeline

Presents time on the Y axis. Good for displaying detailed time series that work especially well when scrolling on mobile.

Seismogram

Another alternative to the circle timeline for showing series where there are big variations in the data.

Magnitude

Show size comparisons. These can be relative (just being able to see larger/bigger) or absolute (need to see fine differences). Usually these show a 'counted' number (for example, barrels, dollars or people) rather than a calculated rate or per cent.

Example FT uses
Commodity production, market capitalisation

Column

The standard way to compare the size of things. Must always start at 0 on the axis.

Bar

See above. Good when the data are not time series and labels have long category names.

Paired column

As per standard column but allows for multiple series. Can become tricky to read with more than 2 series.

Paired bar

See above.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Proportional symbol

Use when there are big variations between values and/or seeing fine differences between data is not so important.

Isotype (pictogram)

Excellent solution in some instances – use only with whole numbers (do not slice off an arm to represent a decimal).

Lollipop

Lollipop charts draw more attention to the data value than standard bar/column – does not have to start at zero (but preferable).

Radar

A space-efficient way of showing value of multiple variables- but make sure they are organised in a way that makes sense to reader.

Parallel coordinates

An alternative to radar charts – again, the arrangement of the variables is important. Usually benefits from highlighting values.

Bullet

Good for showing a measurement against the context of a target or performance range.

Grouped symbol

An alternative to bar/column charts when being able to count data or highlight individual elements is useful.

Part-to-whole

Show how a single entity can be broken down into its component elements. If the reader's interest is solely in the size of the components, consider a magnitude-type chart instead.

Example FT uses
Fiscal budgets, company structures, national election results

Stacked column

A simple way of showing part-to-whole relationships but can be difficult to read with more than a few components.

Marimekko

A good way of showing the size and proportion of data at the same time – as long as the data are not too complicated.

Pie

A common way of showing part-to-whole data – but be aware that it's difficult to accurately compare the size of the segments.

Donut

Similar to a pie chart – but the centre can be a good way of making space to include more information about the data (eg total).

Treemap

Use for hierarchical part-to-whole relationships; can be difficult to read when there are many small segments.

Voronoi

A way of turning points into areas – any point within each area is closer to the central point than any other centroid.

Arc

A hemicycle, often used for visualising political results in parliaments.

Gridplot

Good for showing % information, they work best when used on whole numbers and work well in multiple layout form.

Venn

Generally only used for schematic representation.

Waterfall

Can be useful for showing part-to-whole relationships where some of the components are negative.

Spatial

Aside from locator maps only used when precise locations or geographical patterns in data are more important to the reader than anything else.

Example FT uses
Population density, natural resource locations, natural disaster risk/impact, catchment areas, variation in election results

Basic choropleth (rate/ratio)

The standard approach for putting data on a map – should always be rates rather than totals and use a sensible base geography.

Proportional symbol (count/magnitude)

Use for totals rather than rates – be wary that small differences in data will be hard to see.

Flow map

For showing unambiguous movement across a map.

Contour map

For showing areas of equal value on a map. Can use deviation colour schemes for showing +/- values

Equalised cartogram

Converting each unit on a map to a regular and equally-sized shape – good for representing voting regions with equal value.

Scaled cartogram (value)

Stretching and shrinking a map so that each area is sized according to a particular value.

Dot density

Used to show the location of individual events/locations – make sure to annotate any patterns the reader should see.

Heat map

Grid-based data values mapped with an intensity colour scale. As choropleth map – but not snapped to an admin/political unit.

Flow

Show the reader volumes or intensity of movement between two or more states or conditions. These might be logical sequences or geographical locations.

Example FT uses
Movement of funds, trade, migrants, lawsuits, information; relationship graphs.

Sankey

Shows changes in flows from one condition to at least one other; good for tracing the eventual outcome of a complex process.

Waterfall

Designed to show the sequencing of data through a flow process, typically budgets. Can include +/- components.

Chord

A complex but powerful diagram which can illustrate 2-way flows (and net winner) in a matrix.

Network

Used for showing the strength and inter-connectedness of relationships of varying types.

Visual vocabulary

Designing with data

There are so many ways to visualise data - how do we know which one to pick? Use the categories across the top to decide which data relationship is most important in your story, then look at the different types of chart within the category to form some initial ideas about what might work best. This list is not meant to be exhaustive, nor a wizard, but is a useful starting point for making informative and meaningful data visualisations.

FT graphic: Alan Smith; Chris Campbell; Ian Bott; Liz Faunce; Graham Parrish; Billy Ehrenberg; Paul McCallum; Martin Stabe
Inspired by the Graphic Continuum by Jon Schwabish and Severino Ribecca

ft.com/vocabulary

VISUALISATION

Correlation

Show the relationship between two or more variables. Be mindful that, unless you tell them otherwise, many readers will assume the relationships you show them to be causal (i.e. one causes the other).

Example FT uses
Inflation & unemployment, income & life expectancy

Scatterplot

The standard way to show the relationship between two continuou variables, each of whic has its own axis.

Line + Column

A good way of showing the relationship between an amount (columns) and a rate (line).

Connected scatterplot

Usually used to show how the relationship between 2 variables has changed over time.

Bubble

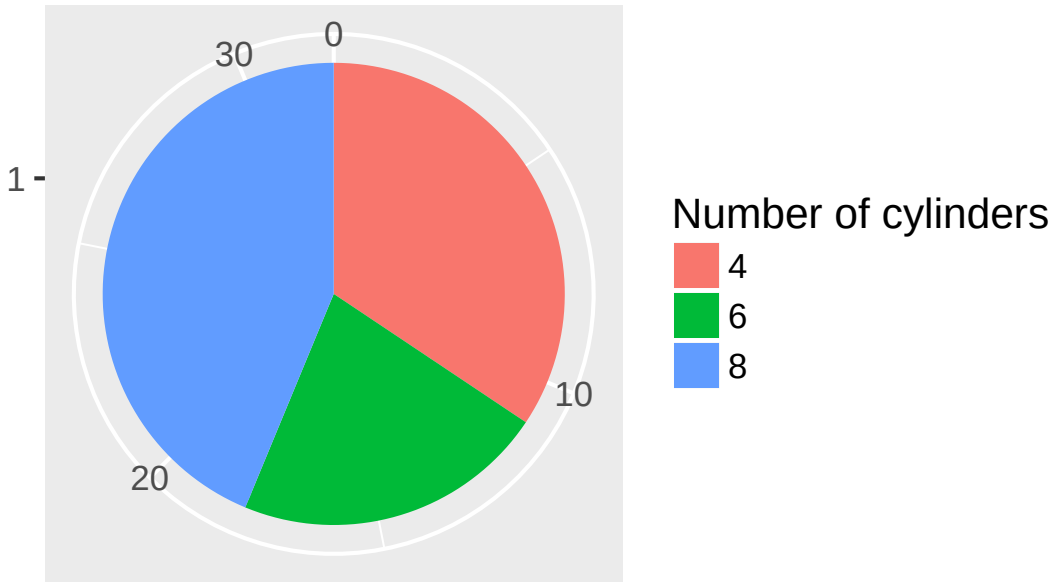
Like a scatterplot, but adds additional detail by sizing the circles according to a third variable.

XY heatmap

A good way of showing the patterns between 2 categories of data, less good at showing fine differences in amounts.

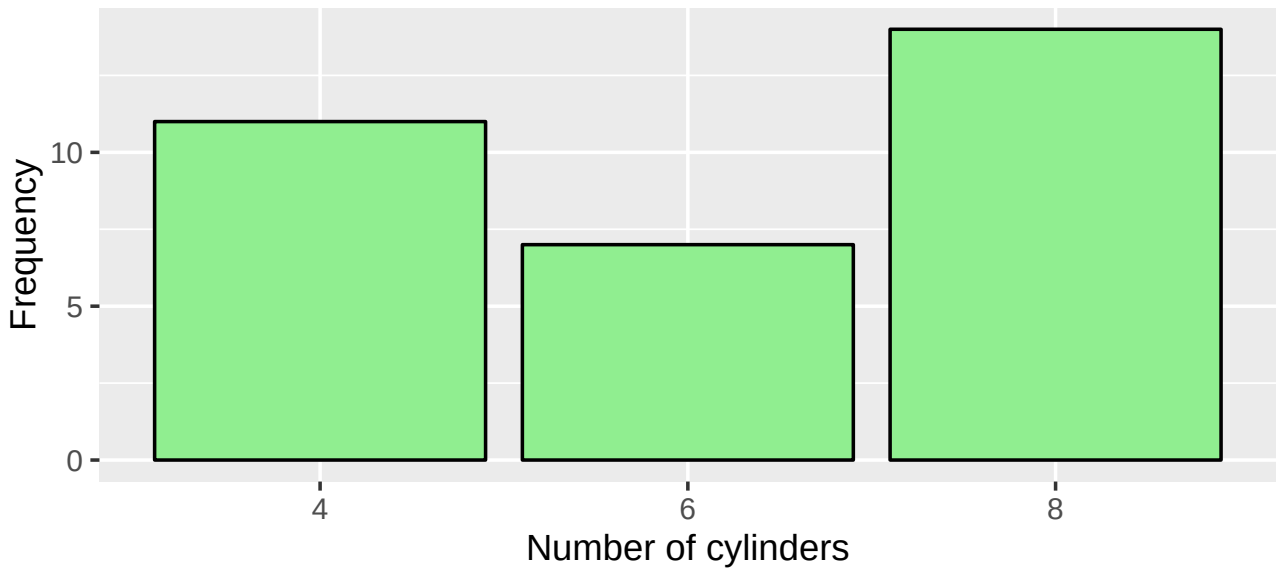
Pie charts

- You can use pie charts to display data where **proportions are important**.
- For example, the proportion of cars with 4, 6 and 8 cylinders tested in the `mtcars` dataset in `R`.
- However, they can be difficult to read and interpret



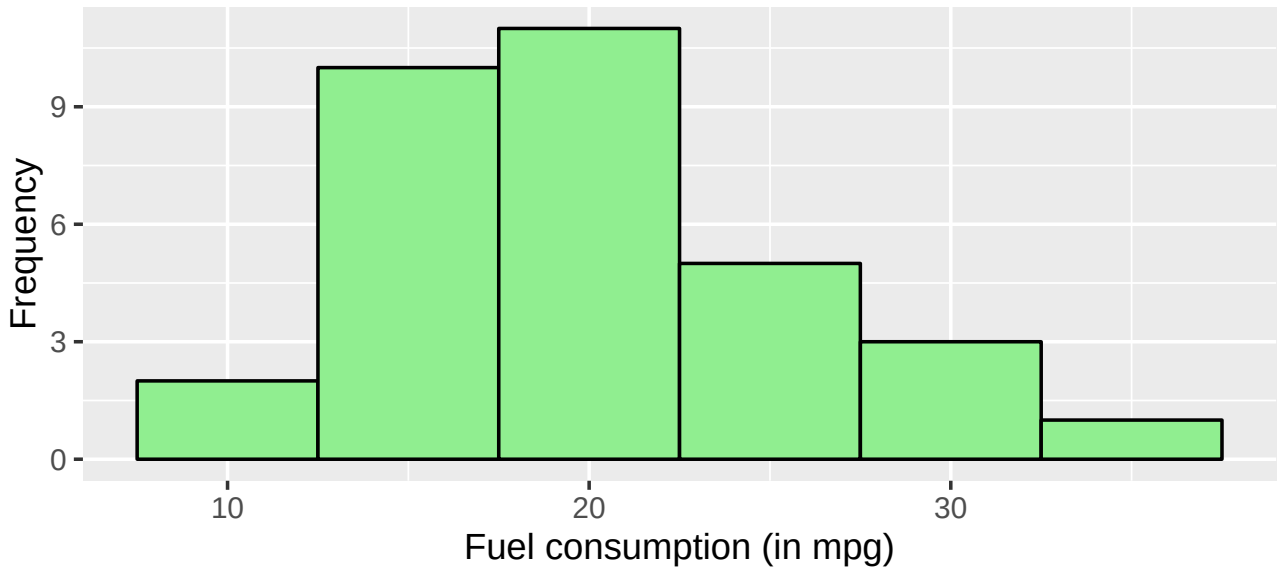
Bar charts

- You can use bar charts to display **frequencies for qualitative variables**.
- The value of a qualitative variable is represented by a bar.
- For example, the number of cars with 4, 6 and 8 cylinders tested in the `mtcars` dataset in `R`.



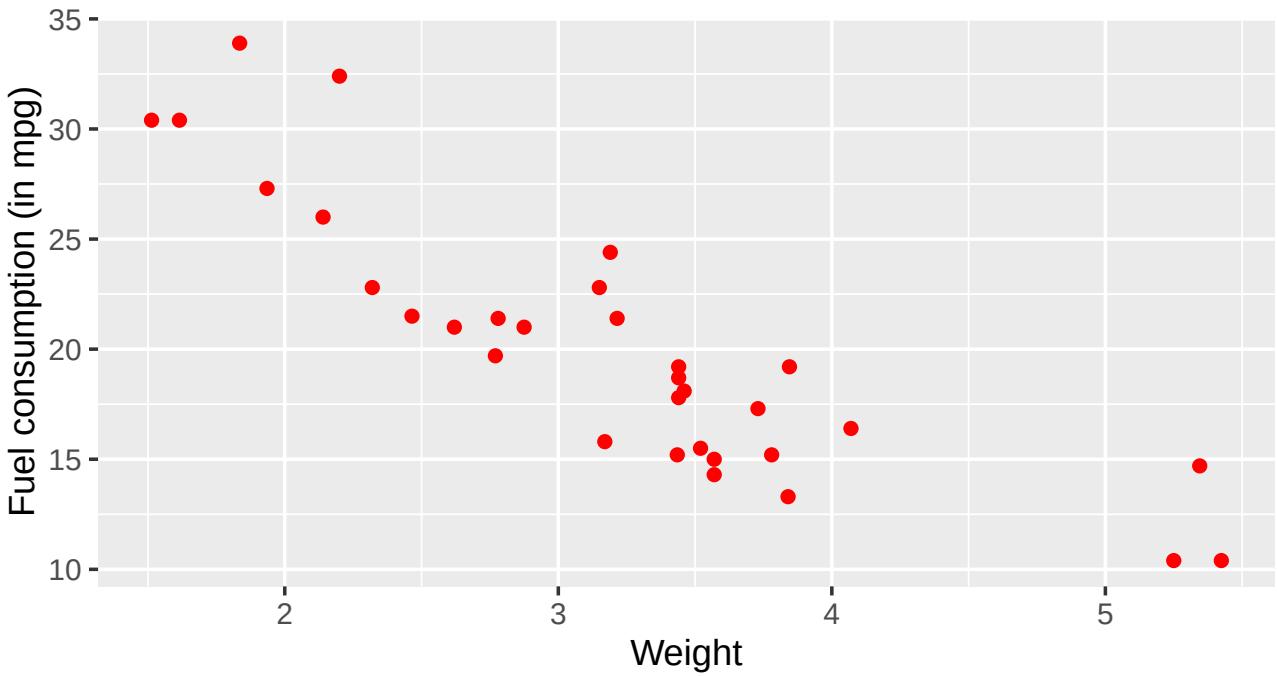
Histograms

- You can use histograms to display the **distribution of a quantitative variable using relative frequencies**.
- The area of each bar has a natural interpretation as a proportion of the total area of all the bars displayed
- There is no space between the bars, and only one variable can be displayed on a single graph.
- For example, histogram of fuel consumption (in miles per gallon) from the `mtcars` dataset in `R`.



Scatter plots

- You can use scatter plots to display **pairs of values of two quantitative variables**, often to check for correlation and association.
- For example, fuel consumption against weight of cars from the `mtcars` dataset in `R`.

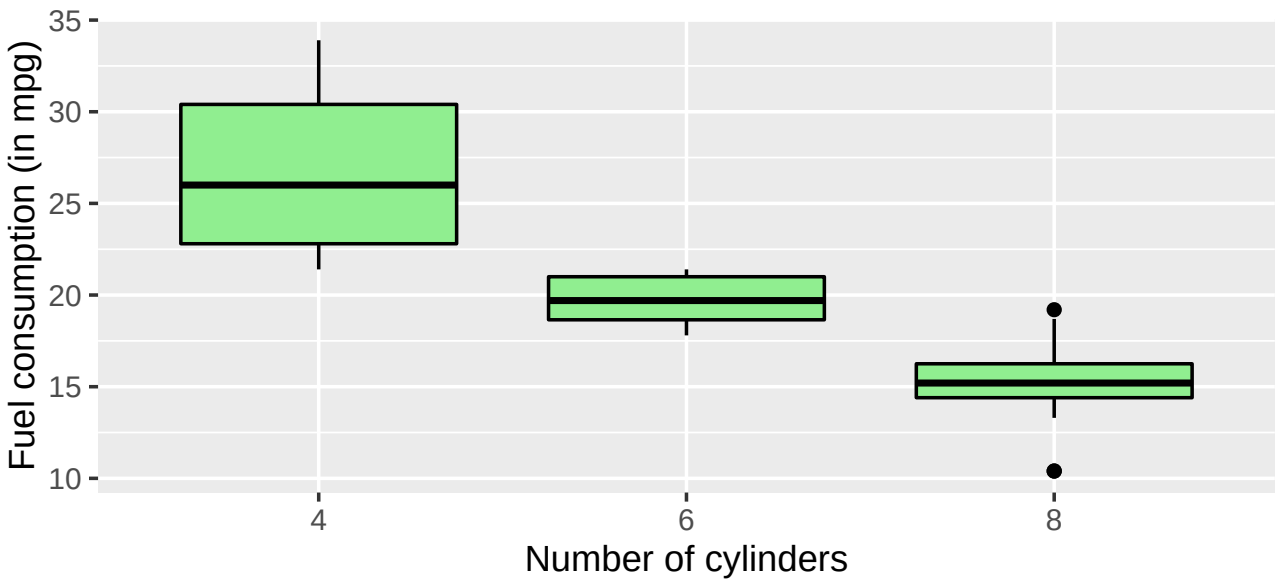


Scatter plots

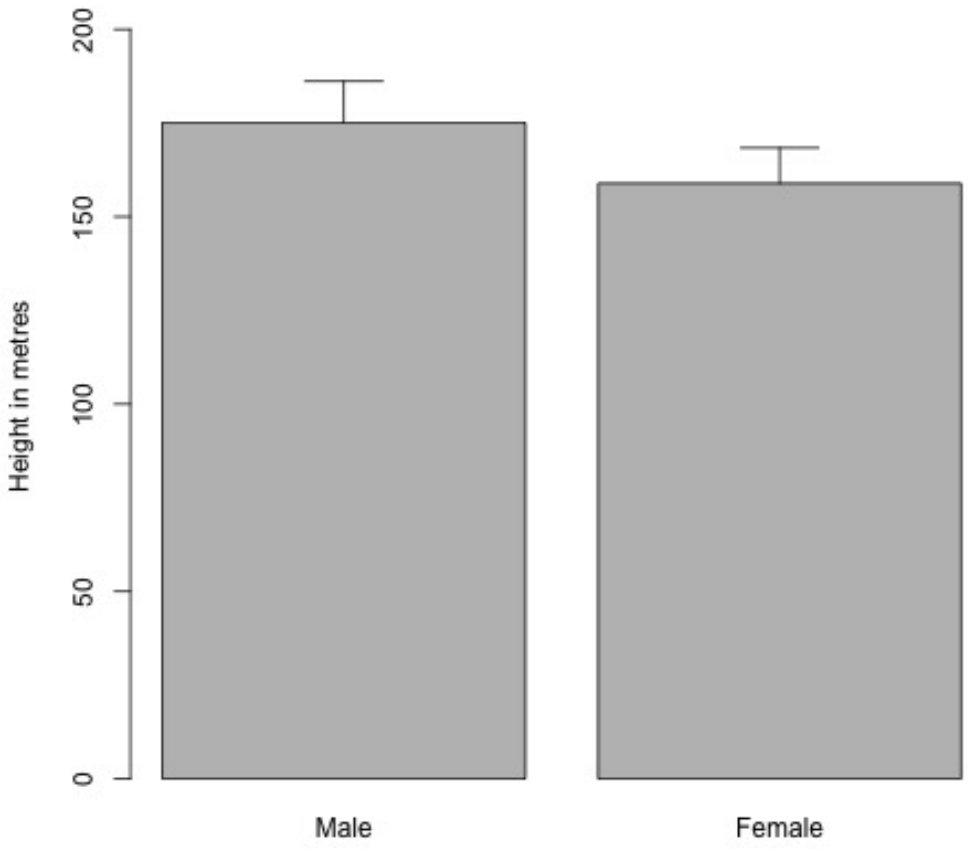
https://twitter.com/DinaPomeranz/status/859198869061677056

Box plots

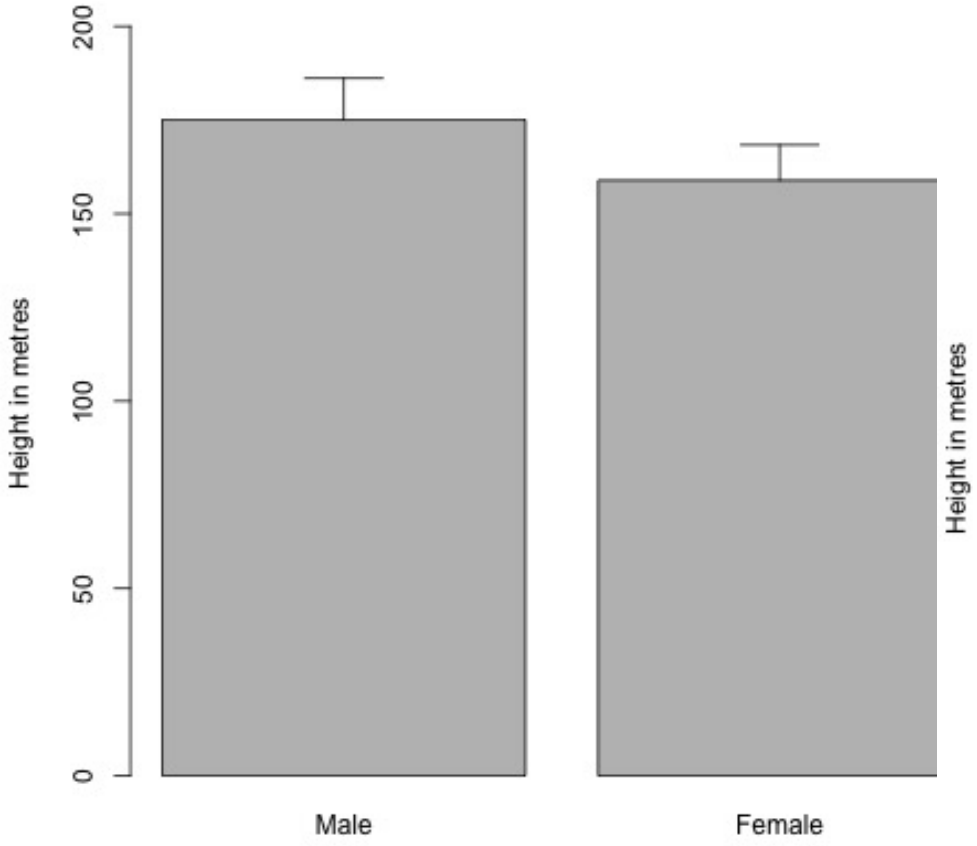
- You can use box plots to display the **median and variability between several sets of observations**.
- The central line is drawn at the median, and the box extends from the lower quartile to the upper quartile.
- For example, box plots of the fuel consumption (in miles per gallon) for cars with 4, 6 and 8 cylinders from the `mtcars` dataset in `R`.

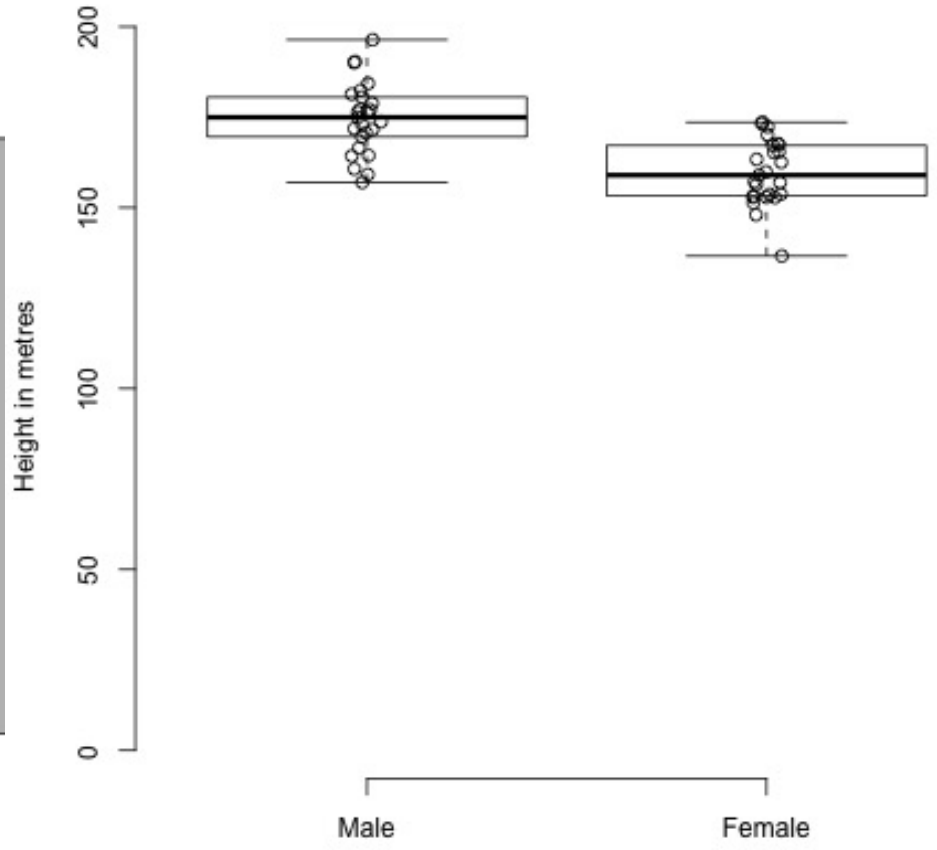


Dynamite plots



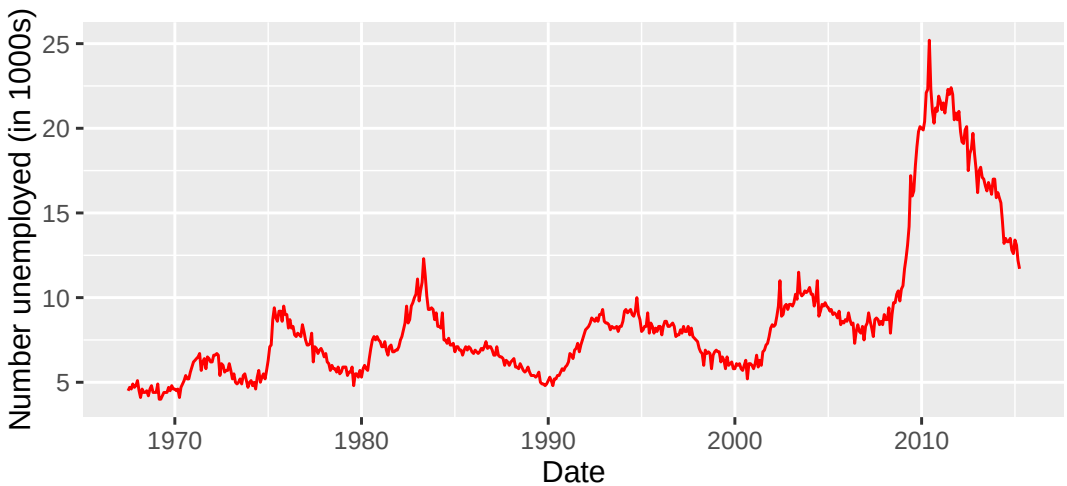
Dynamite plots





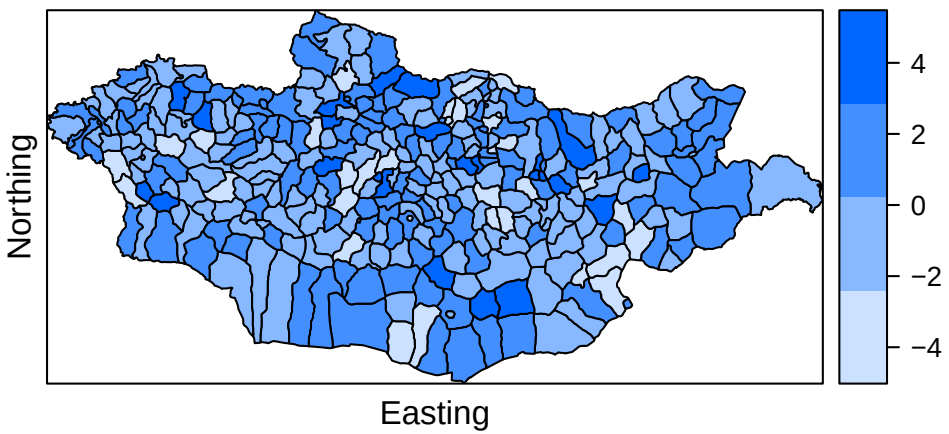
Line plots

- You can use line plots to show values of **one or more variables measured at different times**, connected by a curve.
- For example, the number of unemployed people in the US in thousands over time.



Maps

- You can use maps to display information and variation over space.
- For example, here is a map of Mongolia.



Choice of Visualisation

- The most appropriate type of visualisation depends on the type (qualitative/quantitative, explanatory/response) and number of variables being presented.
- Good visualisation consists of complex ideas communicated with clarity, precision, and efficiency.
- They give the viewer the greatest information in a small amount of space.

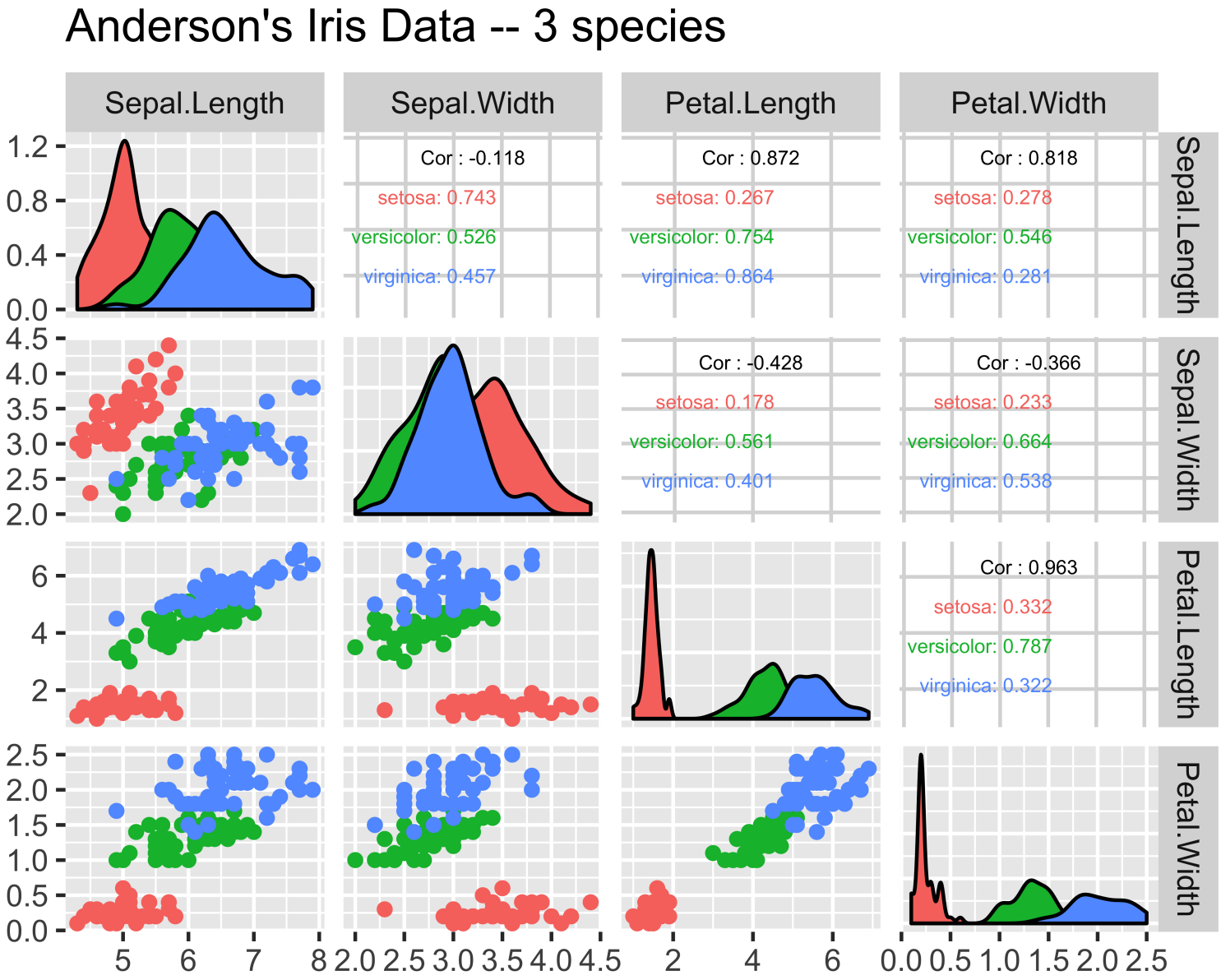
"There are no routine statistical questions, only questionable statistical routines."
D. R. Cox

Choice of Visualisation

- Visualisation can be done throughout an analysis
- Working
  - detect data errors and outliers
  - suggests models
  - may solve the problem alone.
- Presentation
  - effective communication (especially to non-technical audiences)
  - best and perhaps the only chance to get your message across.

Choice of Visualisation

- For simple data sets, you can often present everything at once.



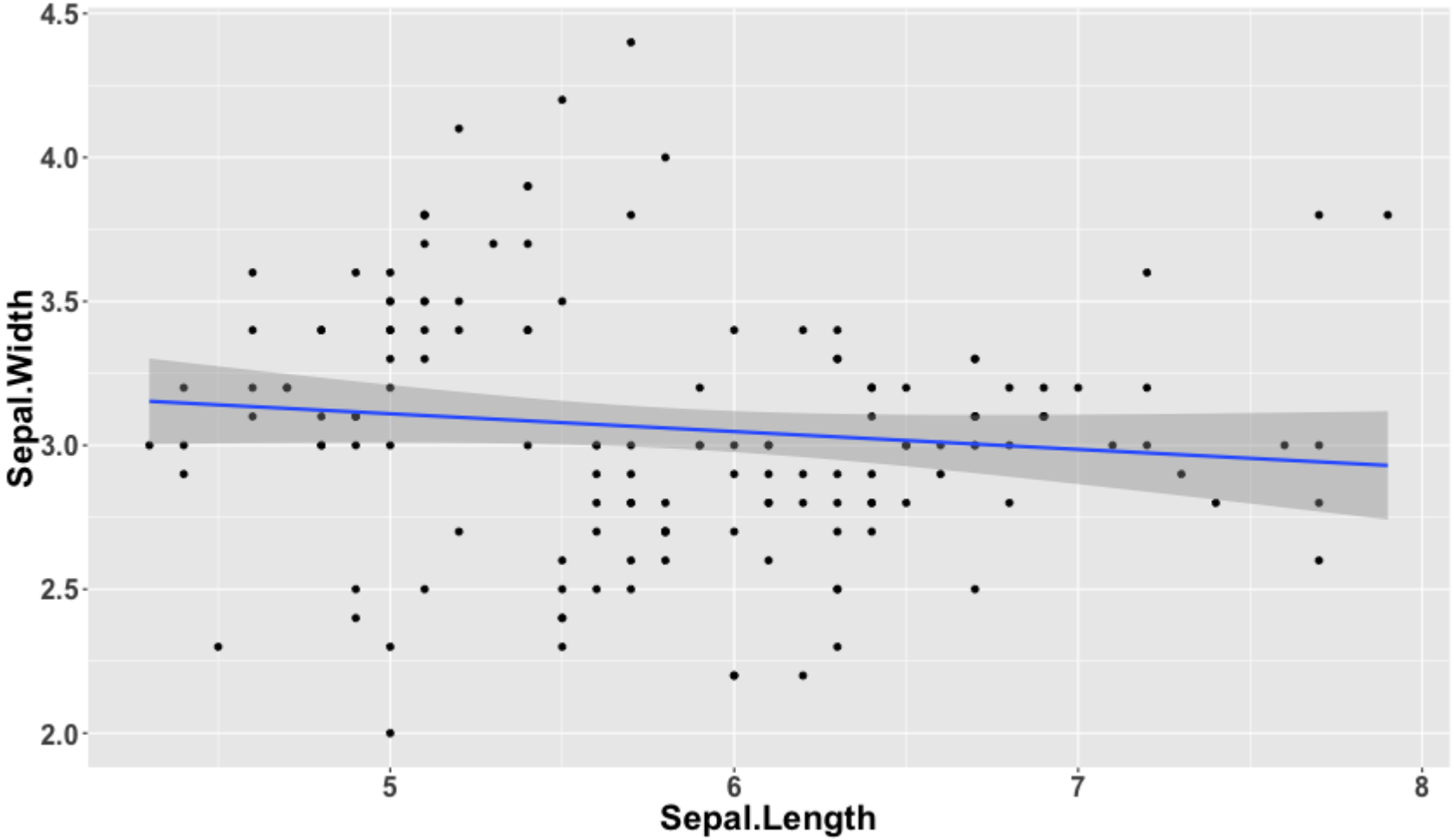
Using Graphics to Present Data

- However, it is much more difficult to view bigger datasets.
- It is important that you choose the right information to display.
- The general guidelines for visualisation are the similar to those for tables.
  - ensure that figures are self-explanatory
  - be consistent in the way that you display information
  - give clear, informative captions and titles
  - make sure your figures only contains information that adds value to your analysis and aids interpretation
  - no space is wasted.
- Always review as if you are a non-expert.

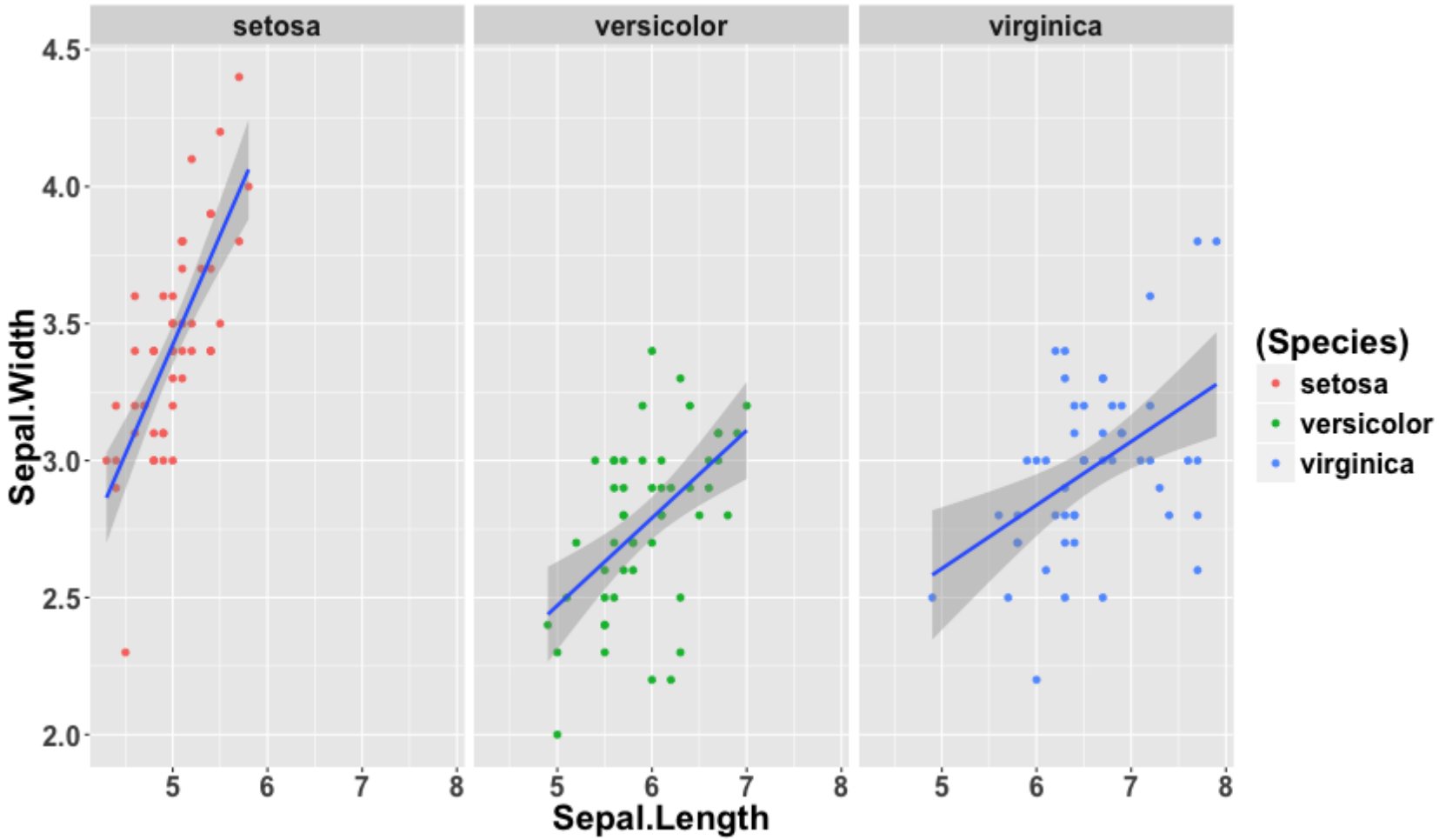
TUFTE'S RULES

http://www.sealthreinhold.com/school/tuftes-rules/rule_one.php

Example: Fisher's iris data



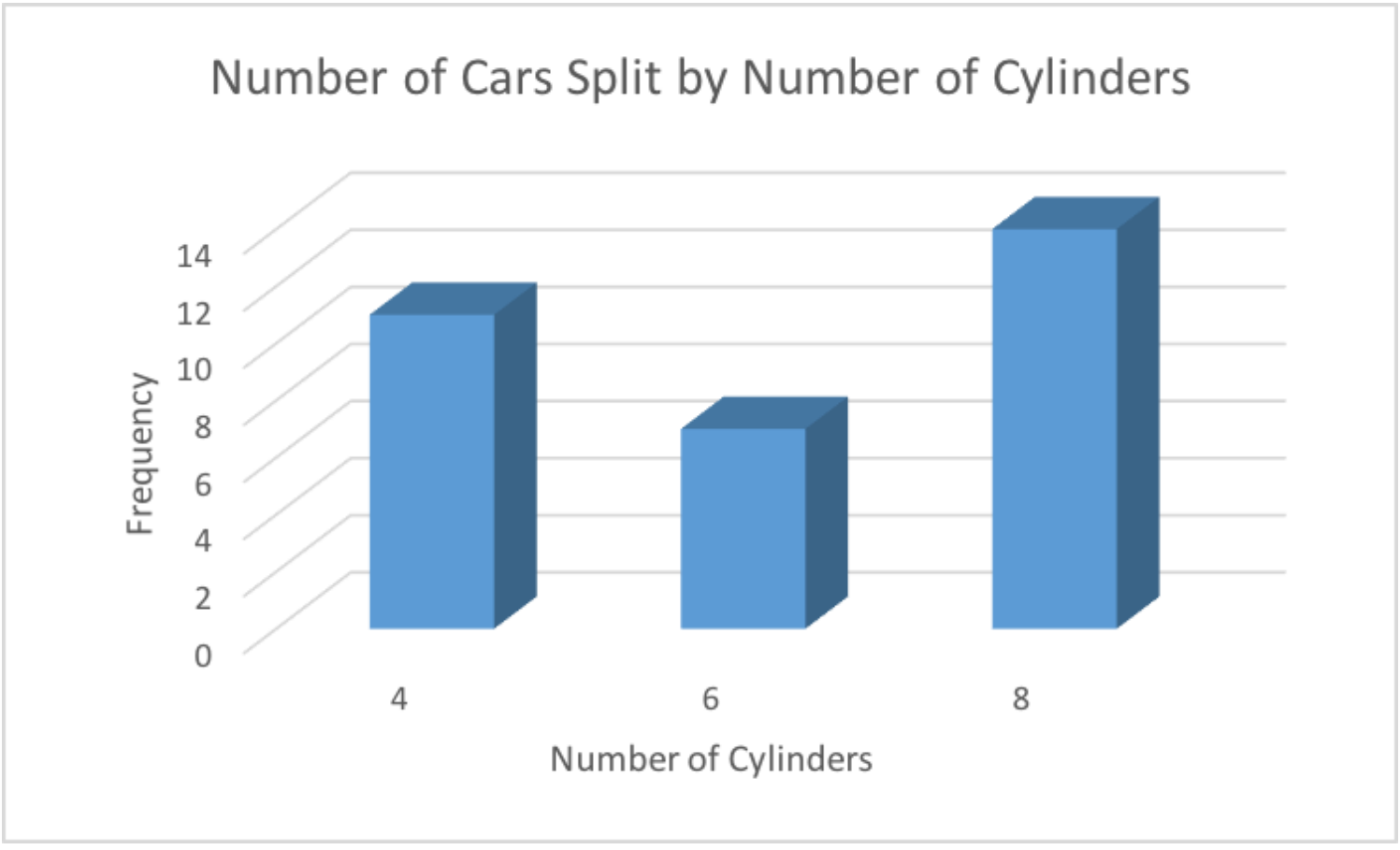
EXAMPLE: FISHER'S IRIS DATA



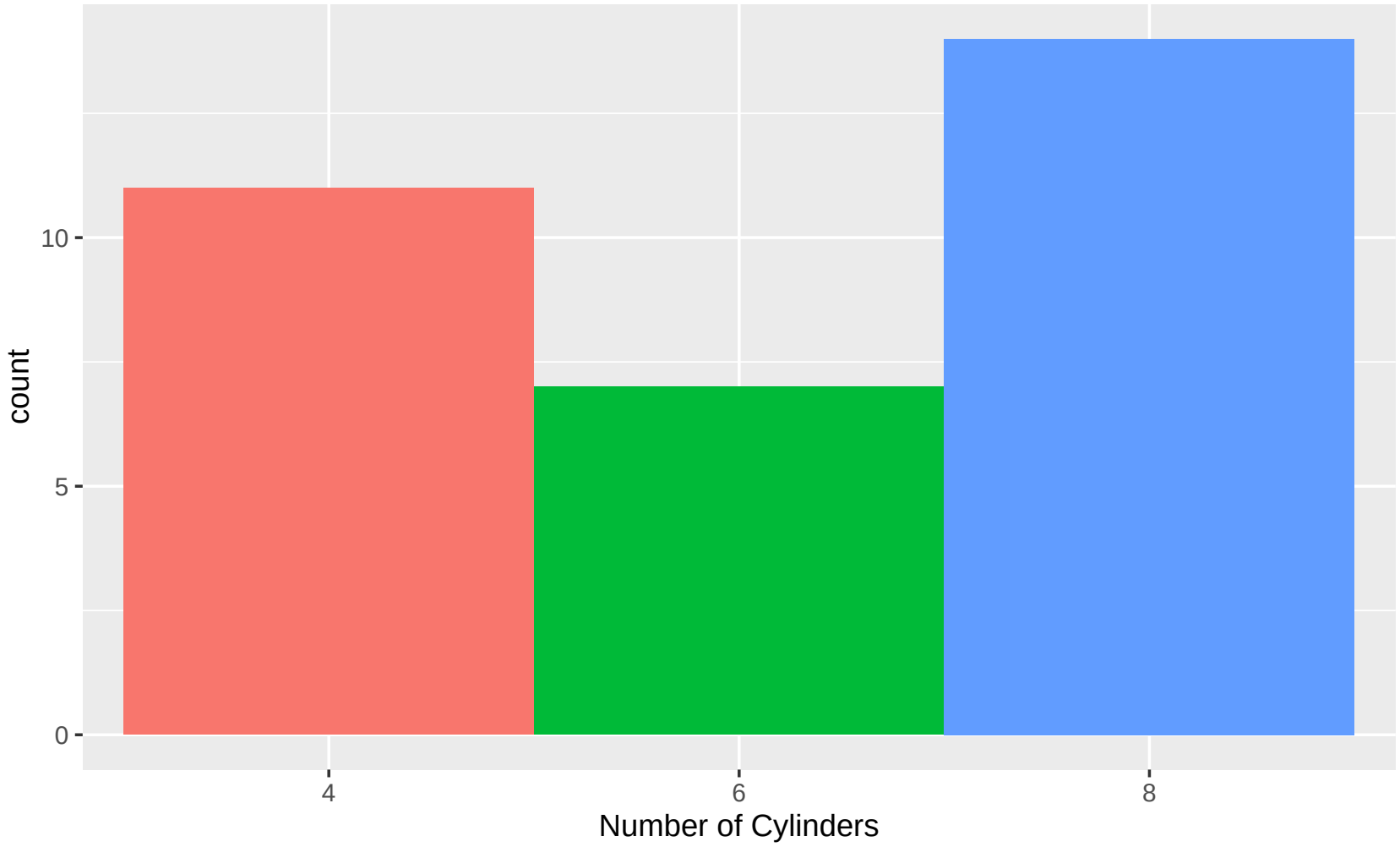
Simpson's paradox

- Trends within groups of data can disappear or reverse when that data is aggregated
  - patterns from aggregated data do not carry over to individual-level data.
  - this means we need to explore our data very carefully to identify patterns
  - as a rule: always engage with a subject specific expert.

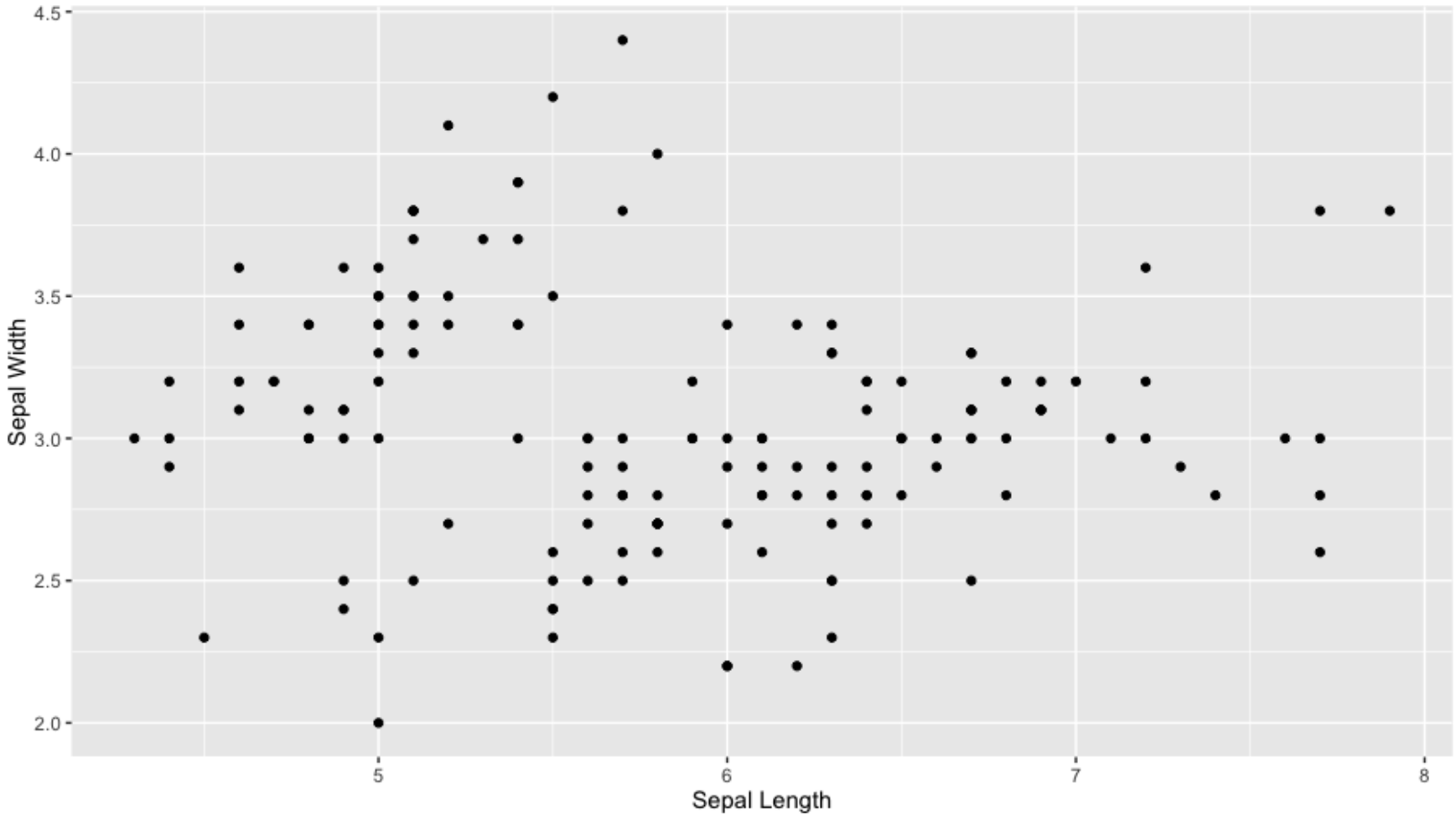
3D Graphics can be very misleading



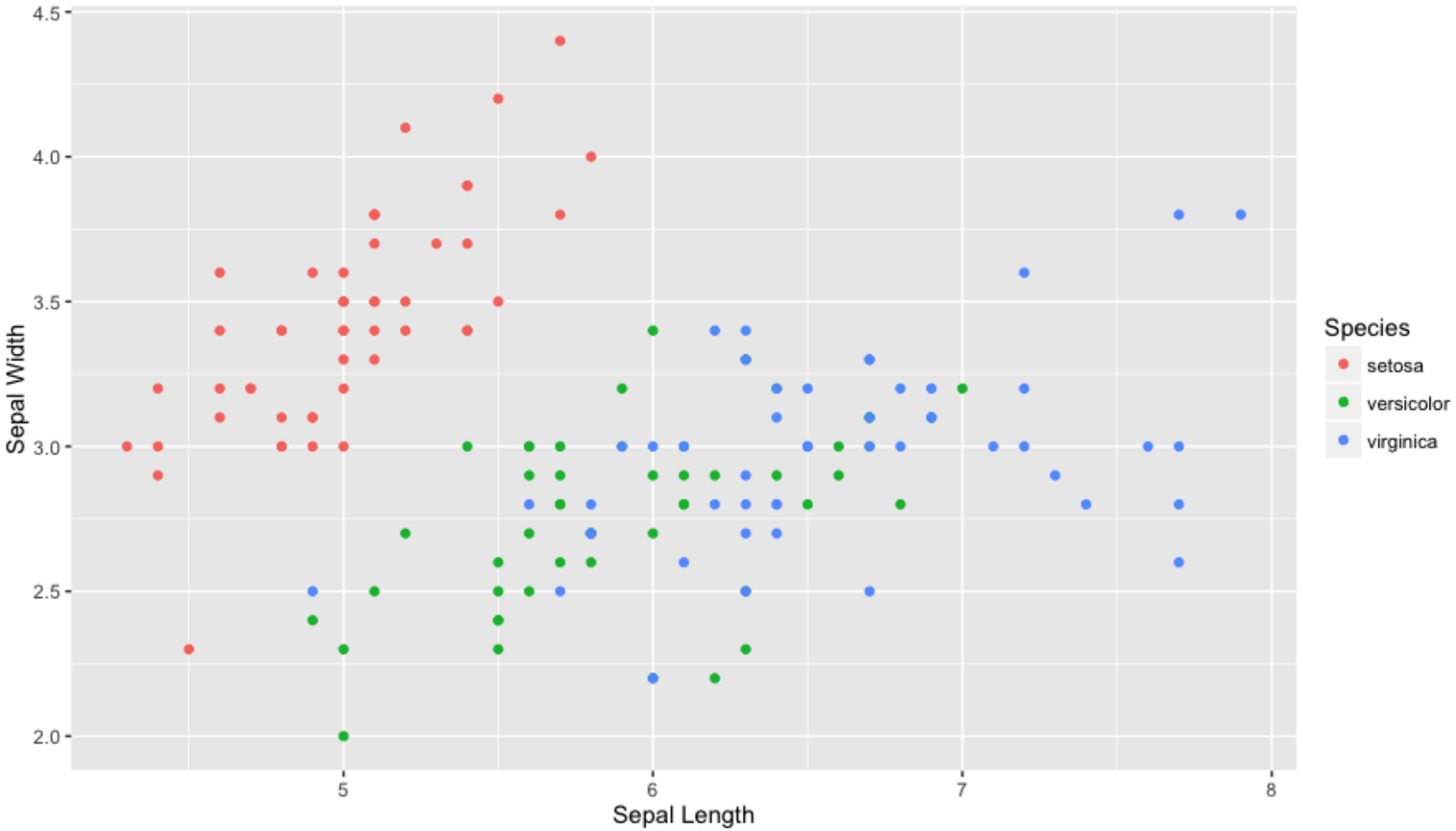
Stick to 2D where possible



Use colour



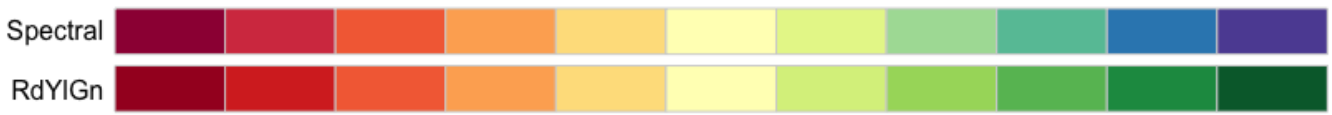
USE COLOUR


4.5
4.0
3.5
3.0
2.5
2.0
Sepal Width
5
6
7
8
Sepal Length
Species
setosa
versicolor
virginica

Colour schemes

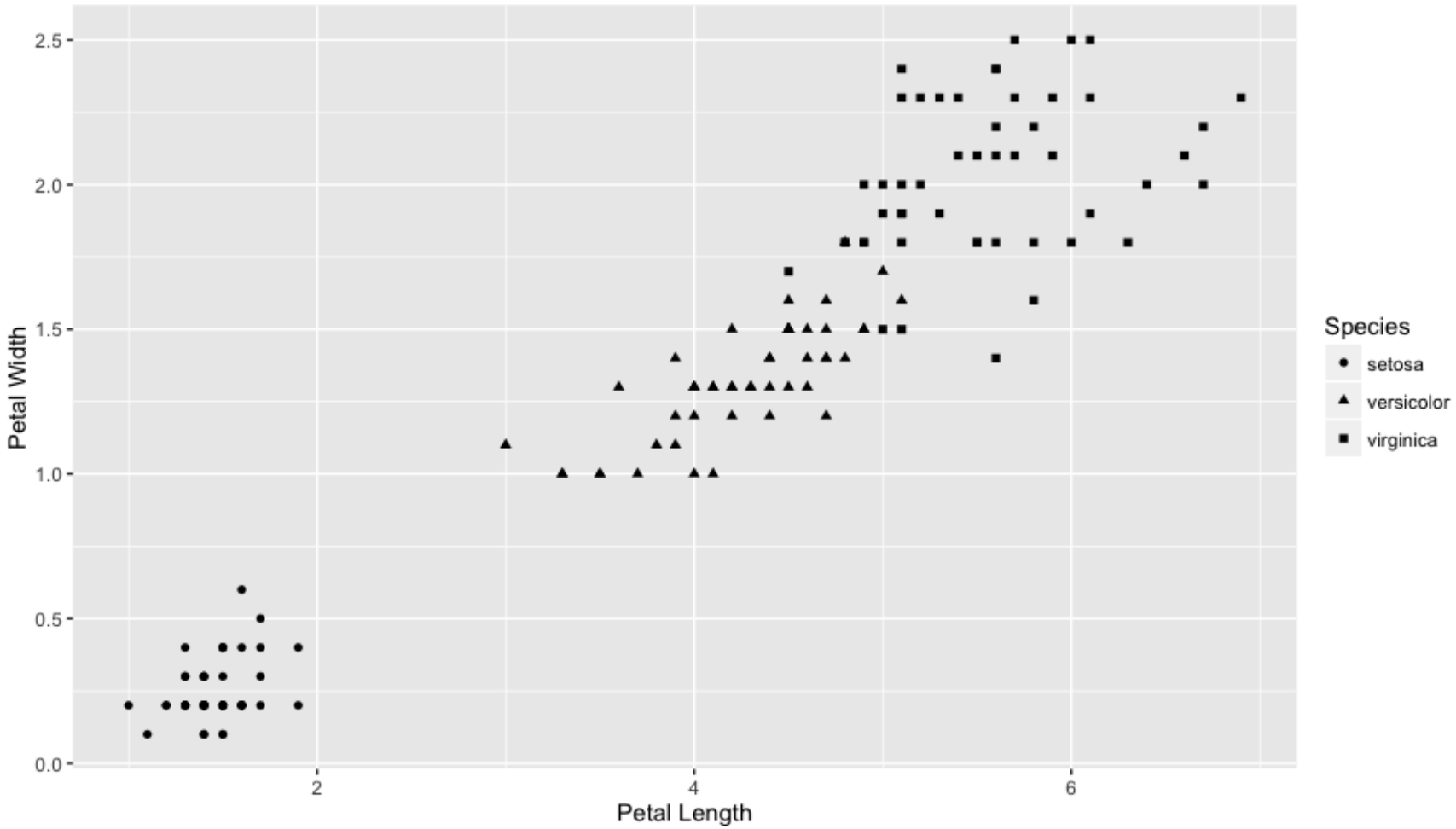
- Colour can be very helpful but there are practical issues.
- Colour scheme must be meaningful.
- Sequential colour schemes are good for ordered data, for example, population density.

- Divergent colour schemes are good for ordered data where you want to focus on deviation from a mean level, deviance for average temperature

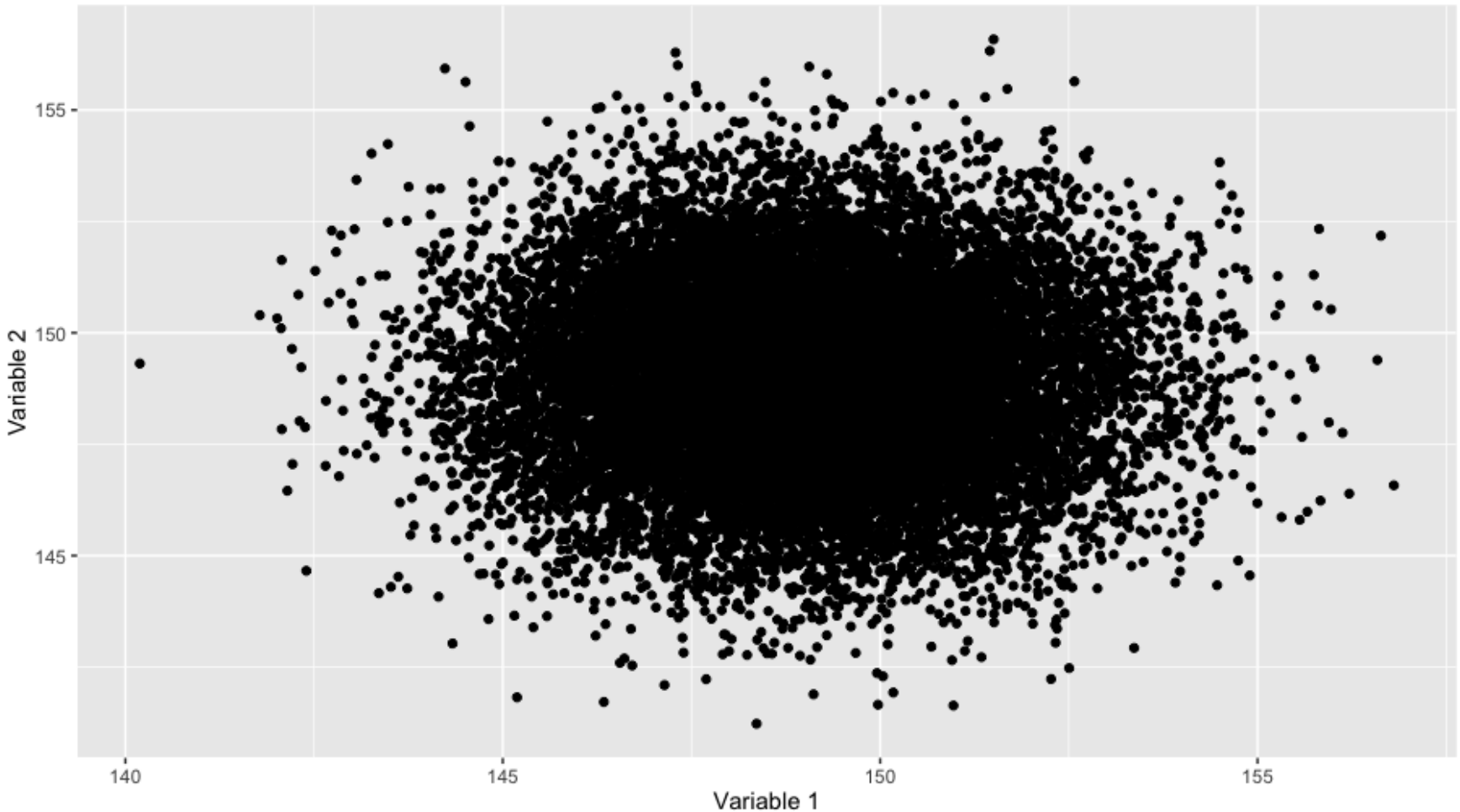


- A good choice of colour scheme are available from `http://colorbrewer2.org` and `RColorBrewer` R package.

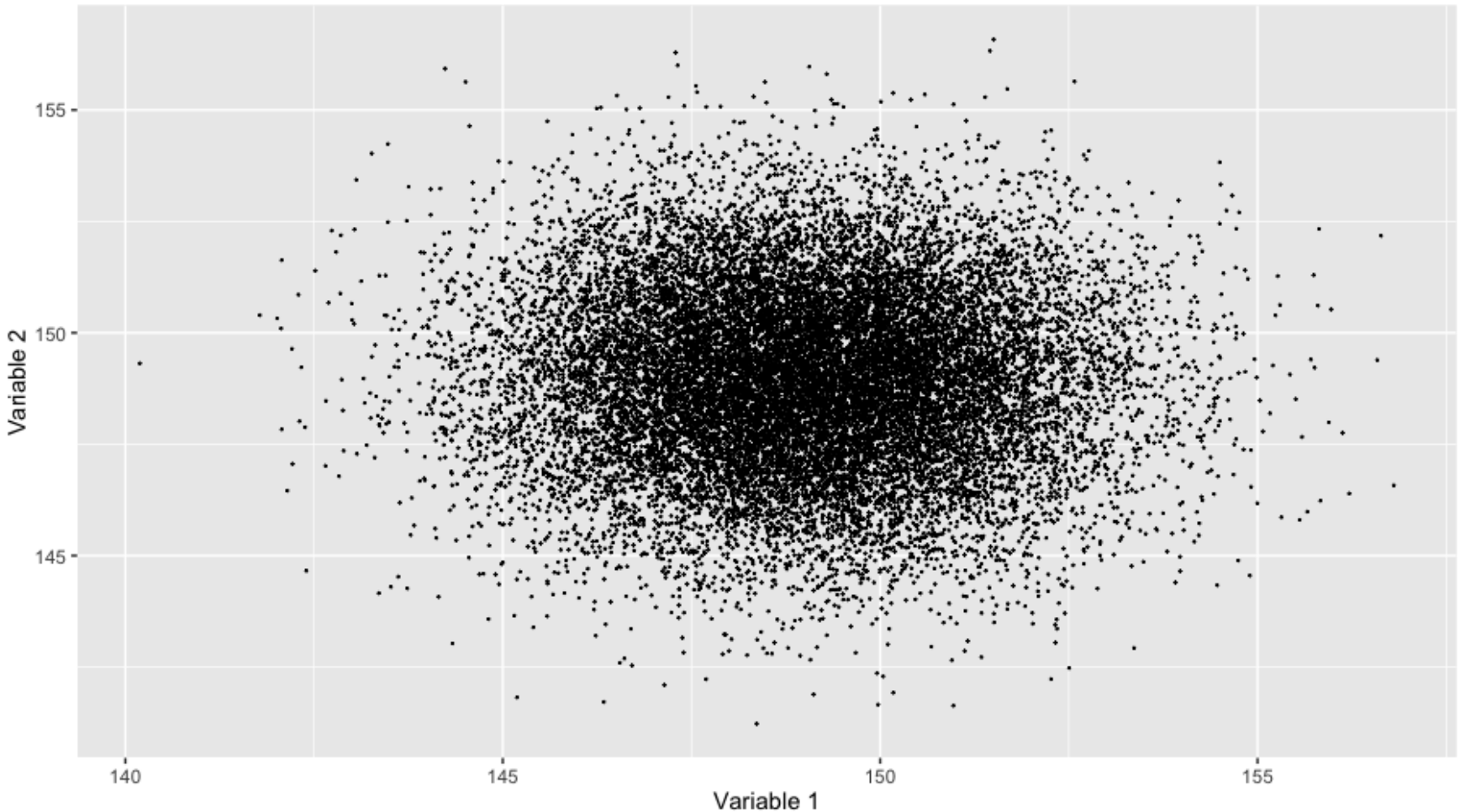
Change the style



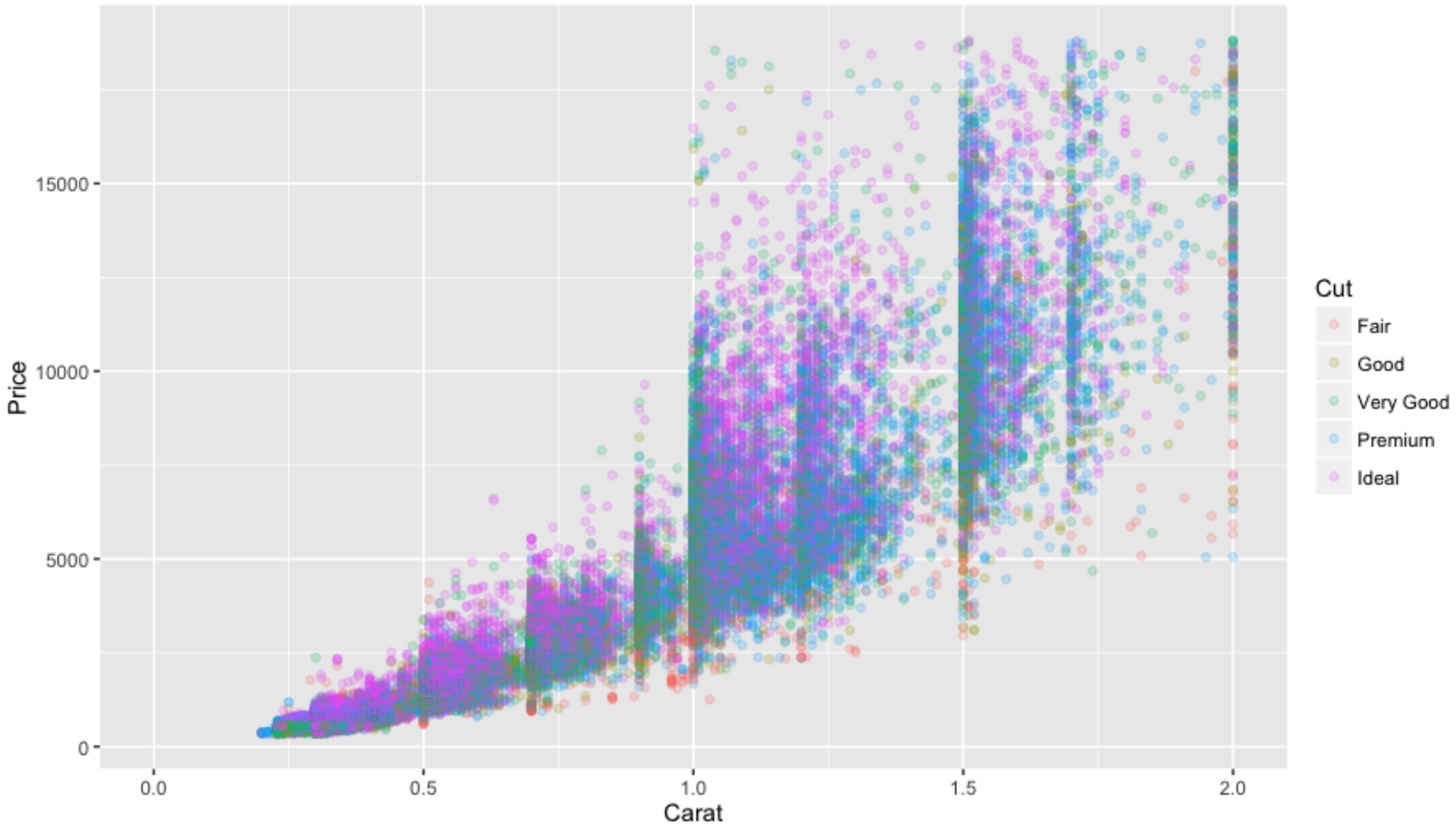
Displaying big data can be difficult



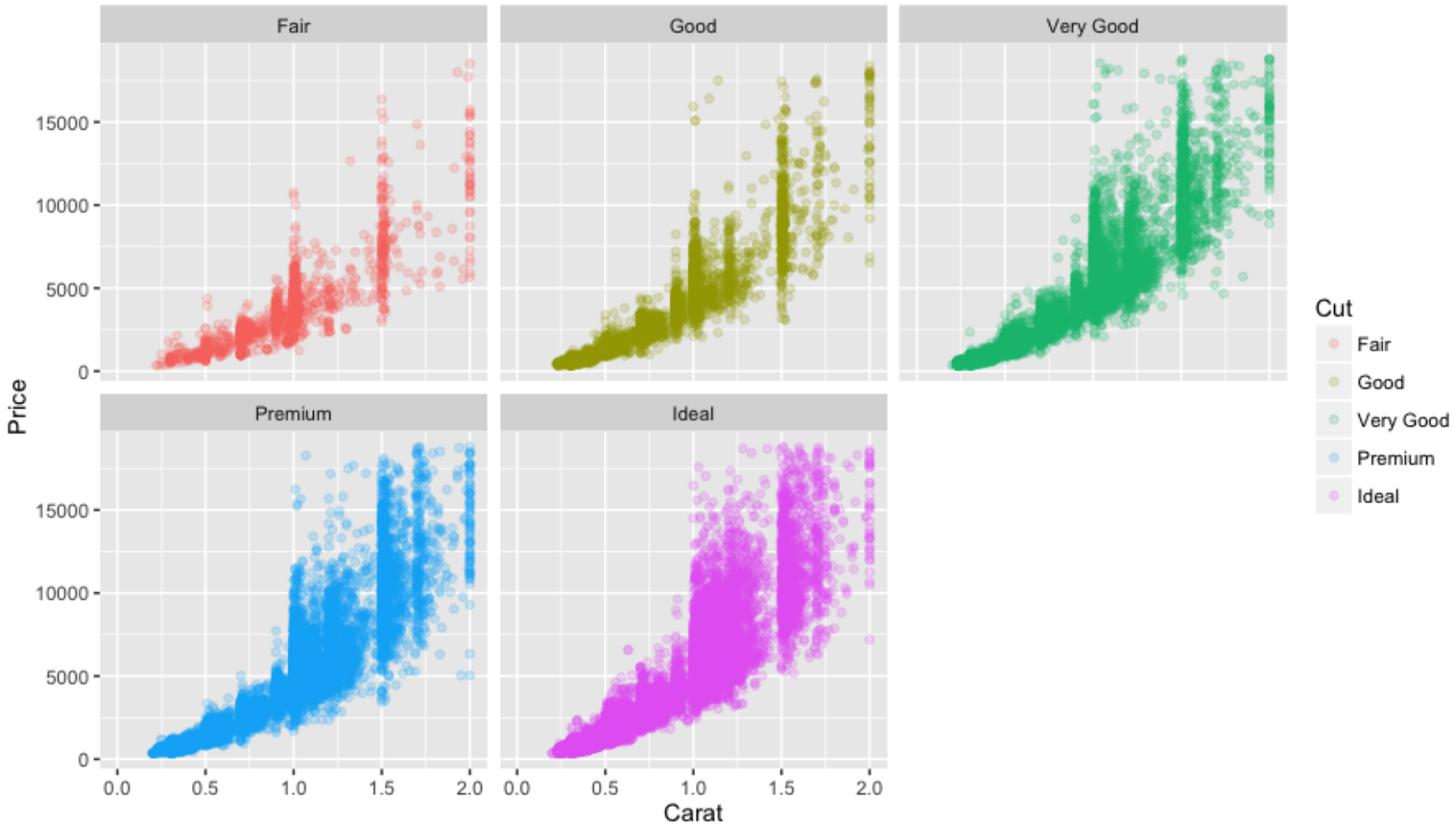
Change the scale



There are some limitations to this



Use Facets


Fair
Good
Very Good
Premium
Ideal
Price
15000
10000
5000
0
0.0
0.5
1.0
1.5
2.0
Carat
Cut
Fair
Good
Very Good
Premium
Ideal

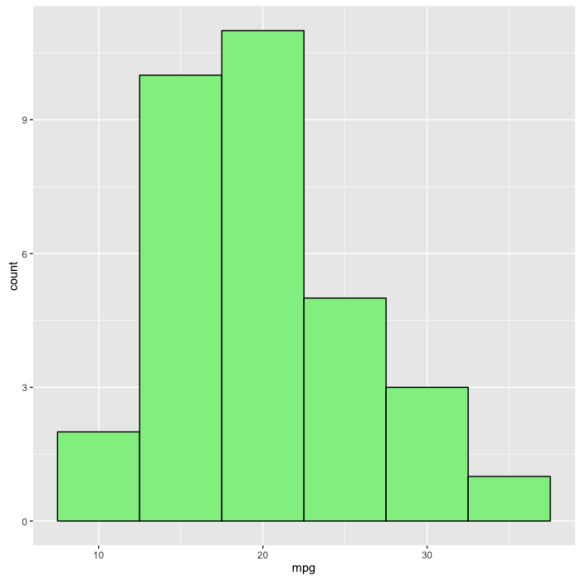
Professional graphics in R

- The `ggplot2` package is a powerful graphics package in `R`.
- You build a `ggplot` up piece by piece, combining the pieces with the "+" operator.
- Graphics using `ggplot2` can be tailored to your analysis.

PROFESSIONAL GRAPHICS IN R

- For example we can create a histogram and store it in `p1`.

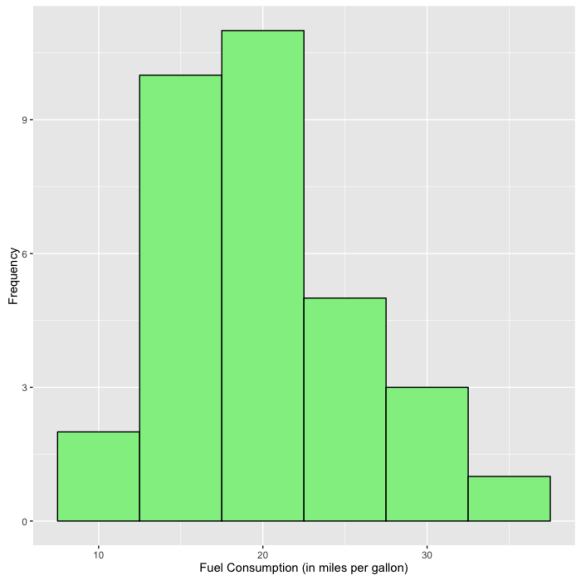
```
p1 <- ggplot(mtcars, aes(x=mpg)) + geom_histogram(binwidth=5,
colour='black',fill='lightgreen')
p1
```



PROFESSIONAL GRAPHICS IN R

- We can change the labels of the axes by adding to `p1`.

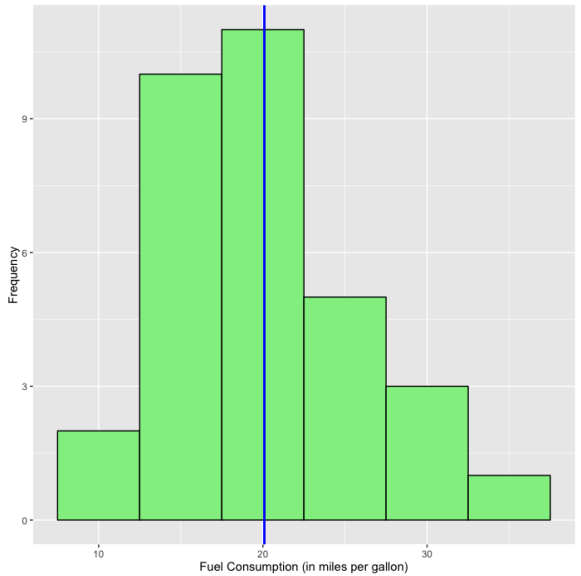
```
p1 <- p1 + labs(x = 'Fuel Consumption (in miles per gallon)',
                y='Frequency')
p1
```



PROFESSIONAL GRAPHICS IN R

- We can add a line to `p1` to indicate the location of the mean.

```
p1 <- p1 + vline(xintercept=mean(mtcars$mpg),col='blue',size=1)
p1
```



Other resources

Data Visualization: A practical introduction, Kieran Healy
`https://socviz.co/`
Introduction to data visualisation
`https://rafalab.github.io/dsbook/introduction-to-data-visualization.html`

Any Questions?