Data Science and Statistics in Research: unlocking the power of your data

Session 1.6:

Visualising data

Outline

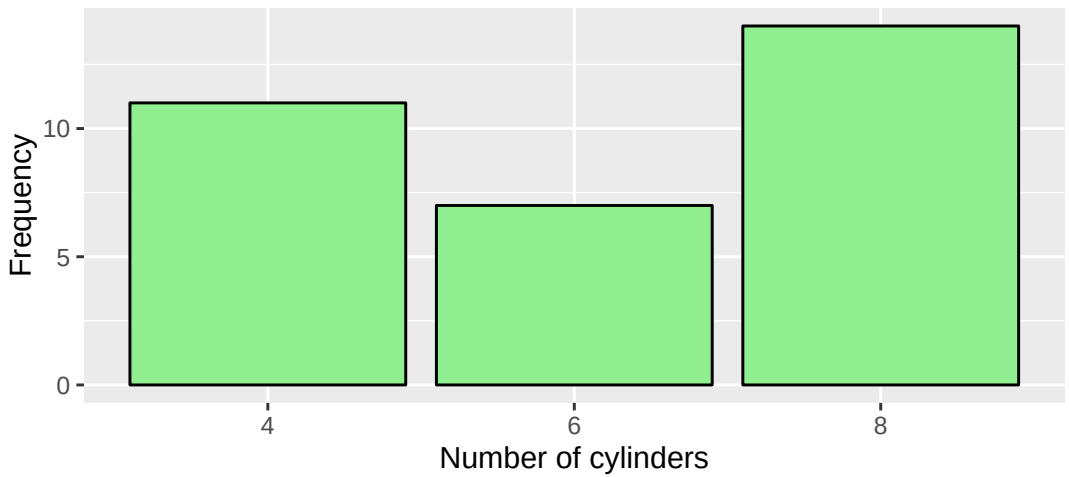
Visualisation

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!

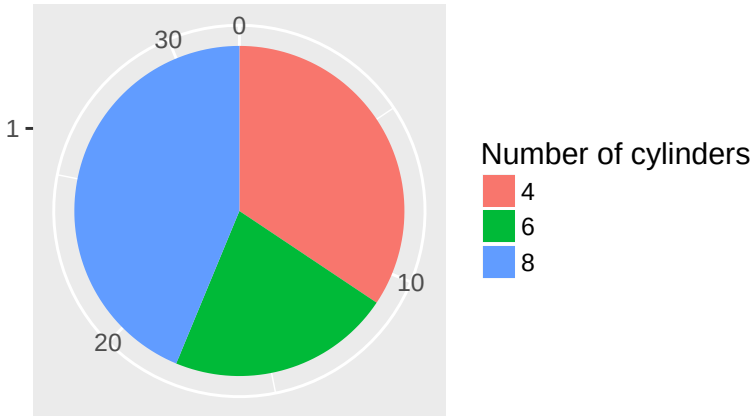
BAR CHARTS

- 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`.



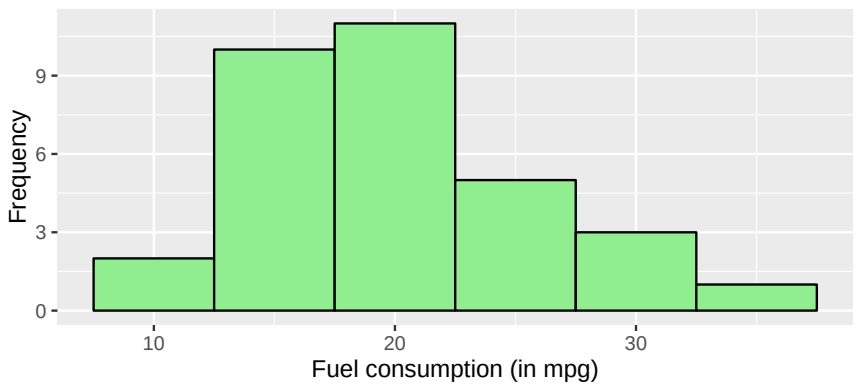
Pie charts

- 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`.



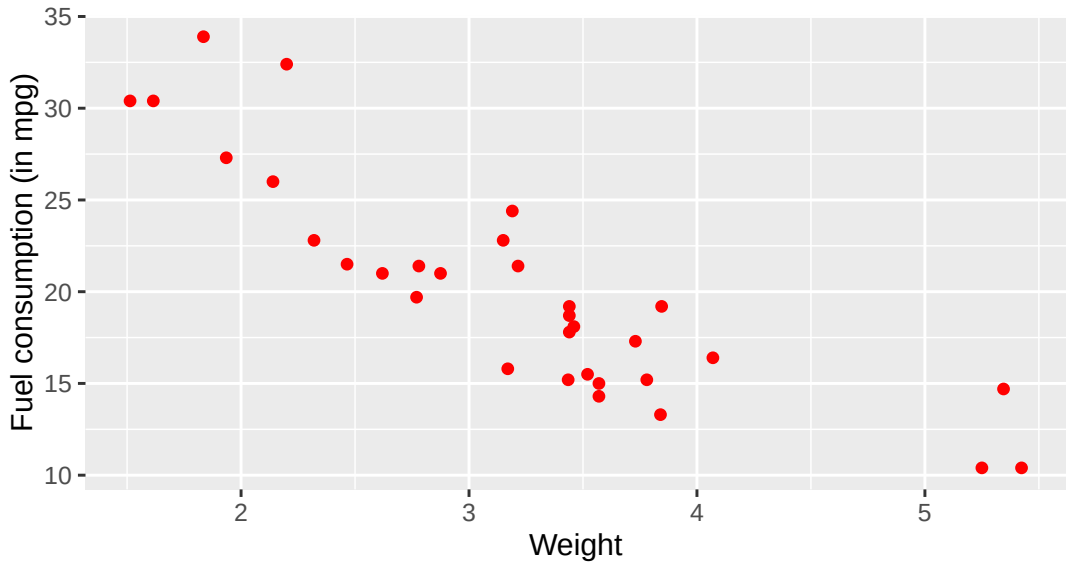
HISTOGRAMS

- 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`.



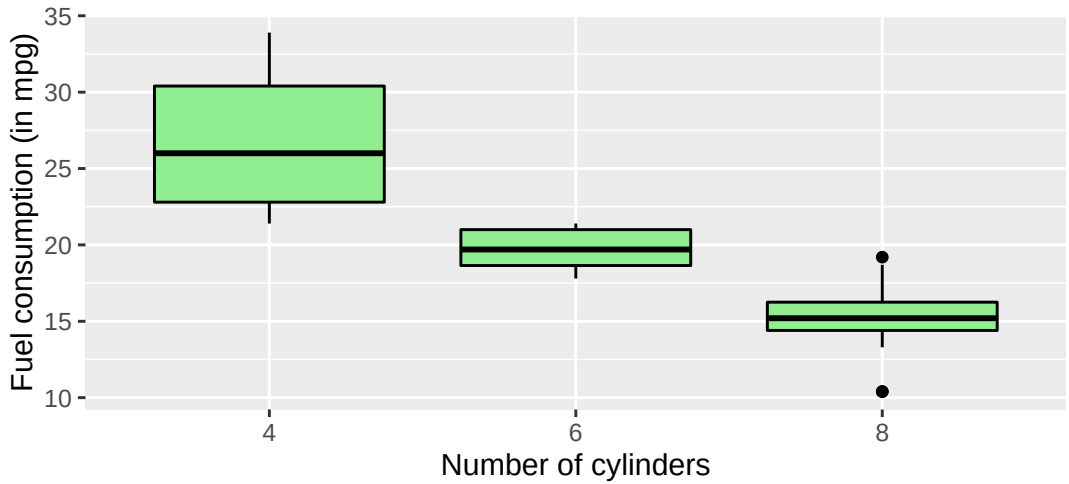
SCATTER PLOTS

- 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`.



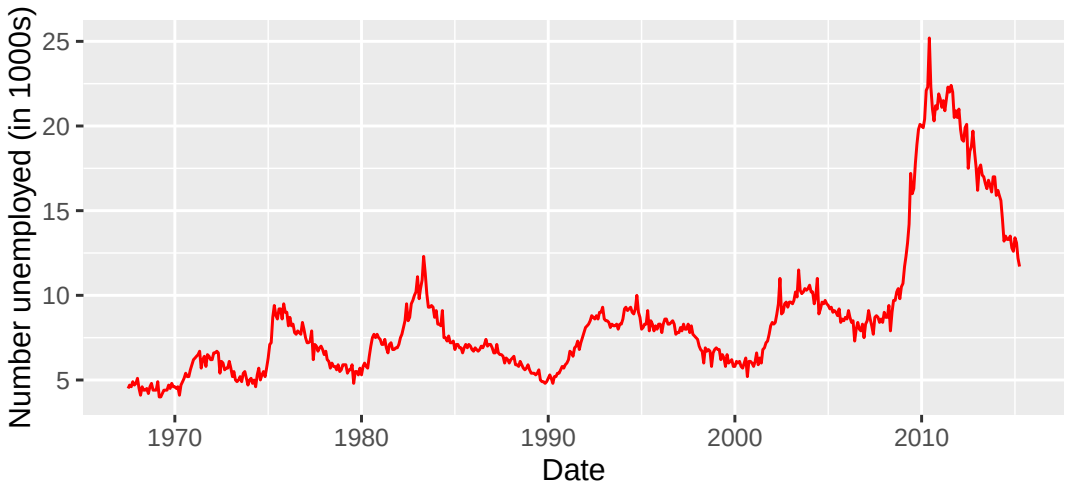
Box plots

- 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`.



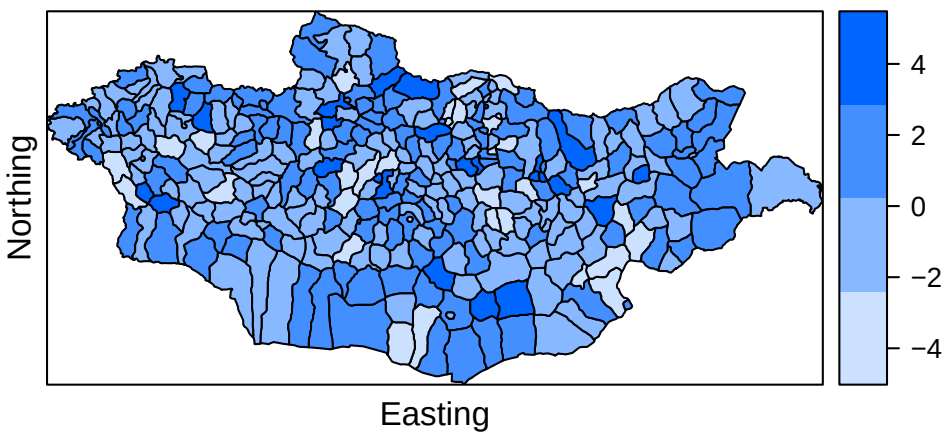
Line plots

- 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

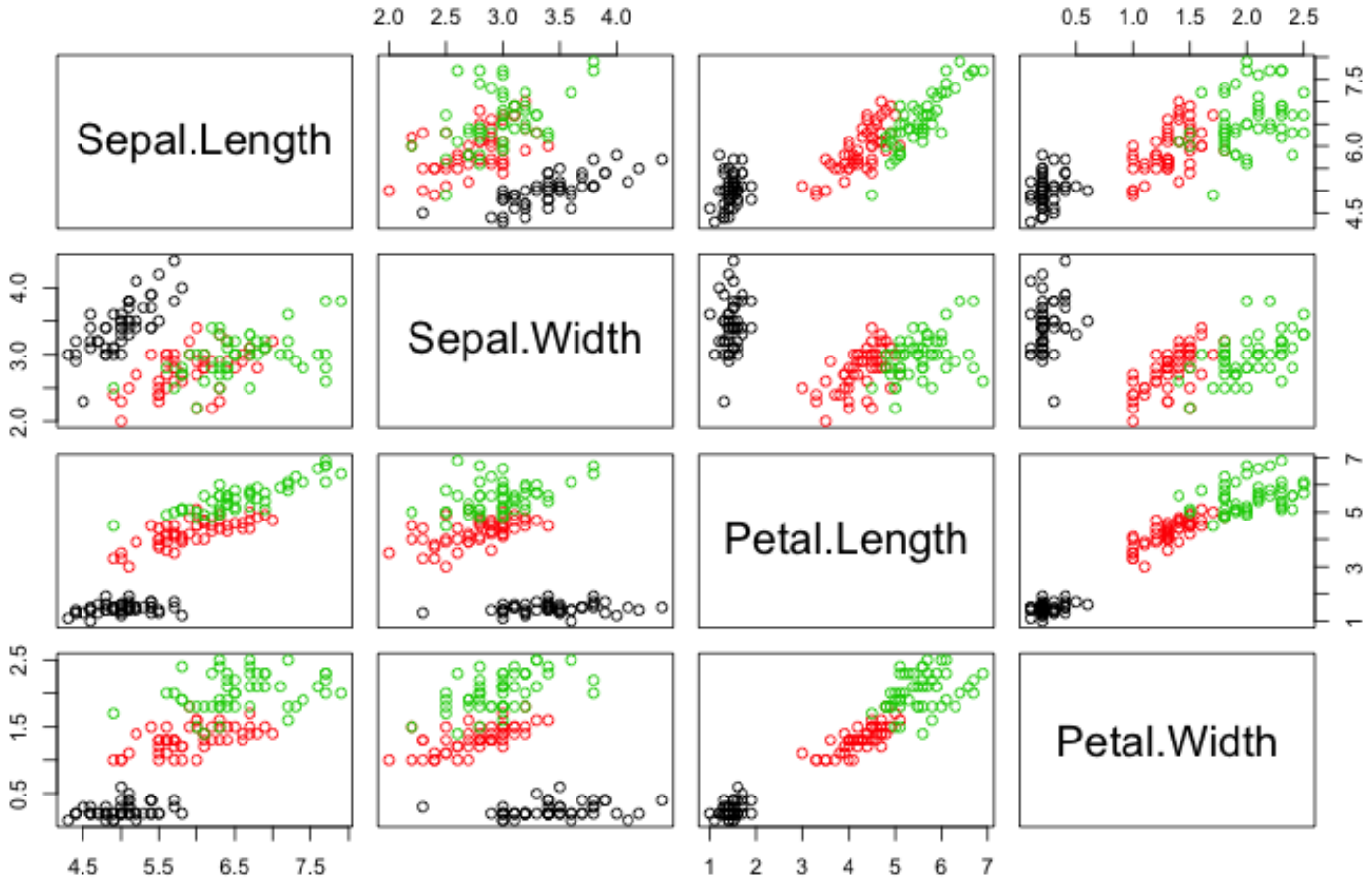
- The most appropriate type of visualisation depends on the type (qualitative/quantitative, explanatory/response) and number of variables being presented.
- Good visualisation consists of complex ideas communicated with clarity, precision, and efficiency.
- They give the viewer the greatest information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

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

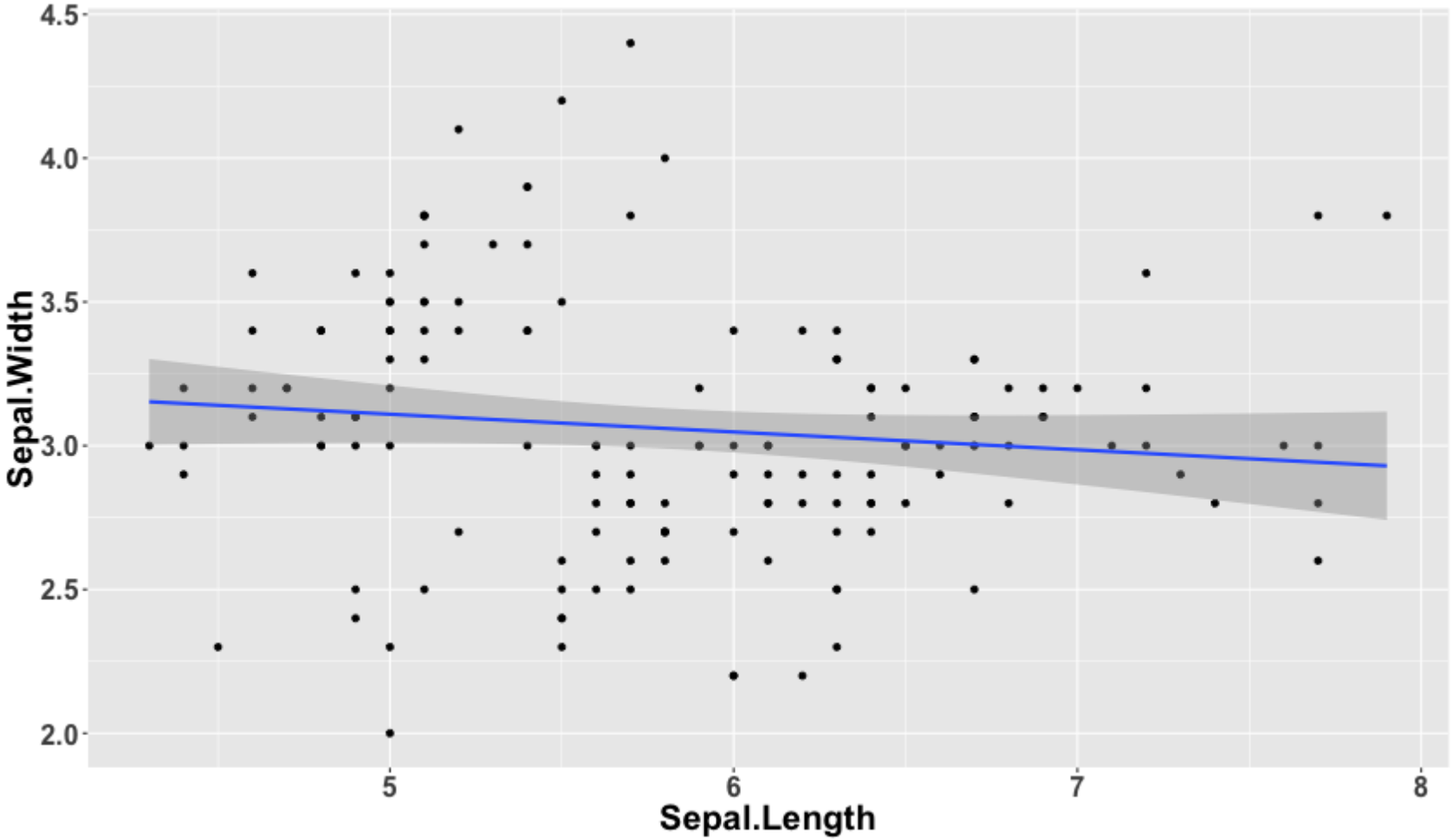
- For simple data sets, you can often present everything at once.



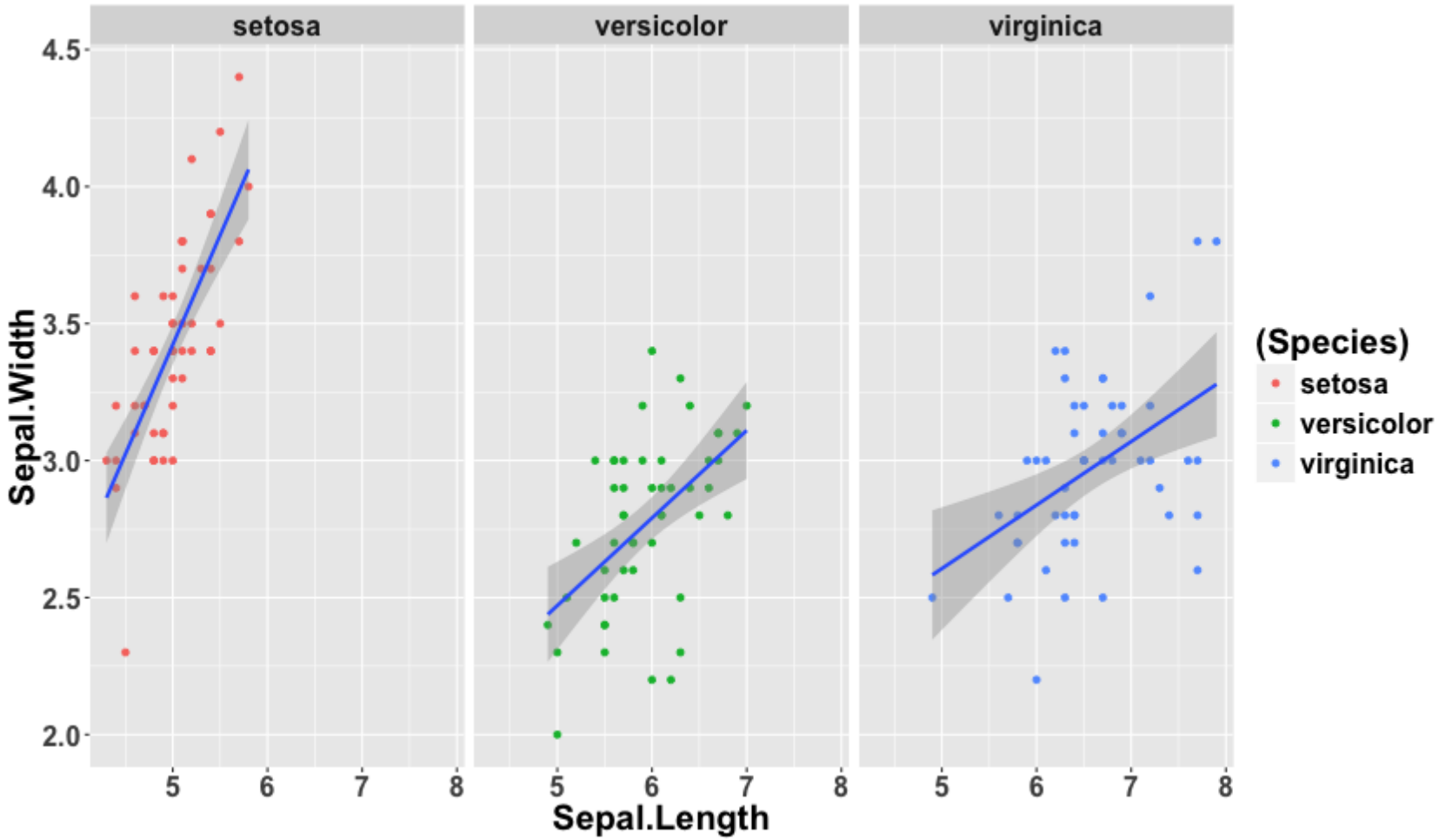
Using Graphics to Present Data

- 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 that 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.

EXAMPLE: FISHER'S IRIS DATA



EXAMPLE: FISHER'S IRIS DATA



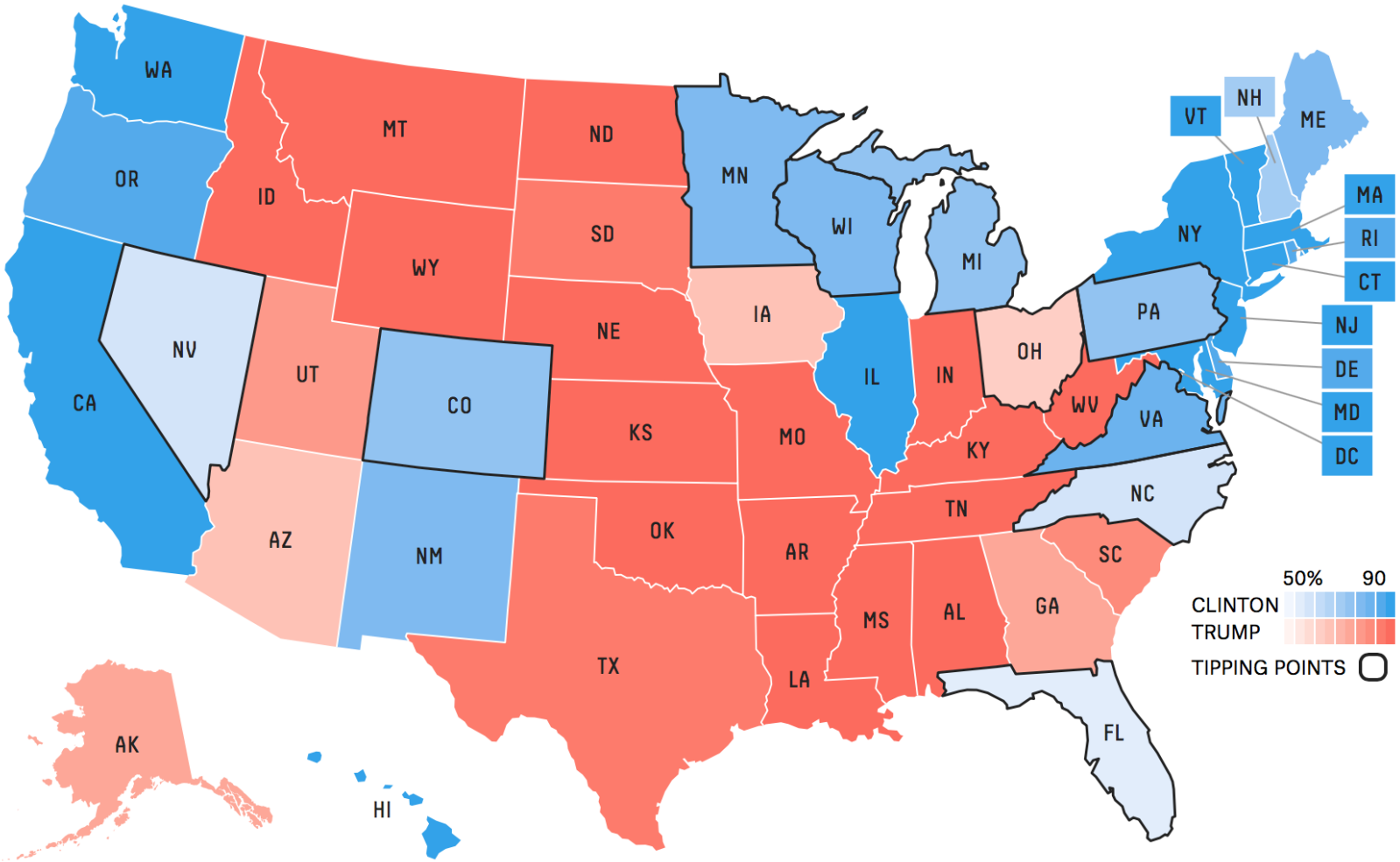
Simpson's paradox

- 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.

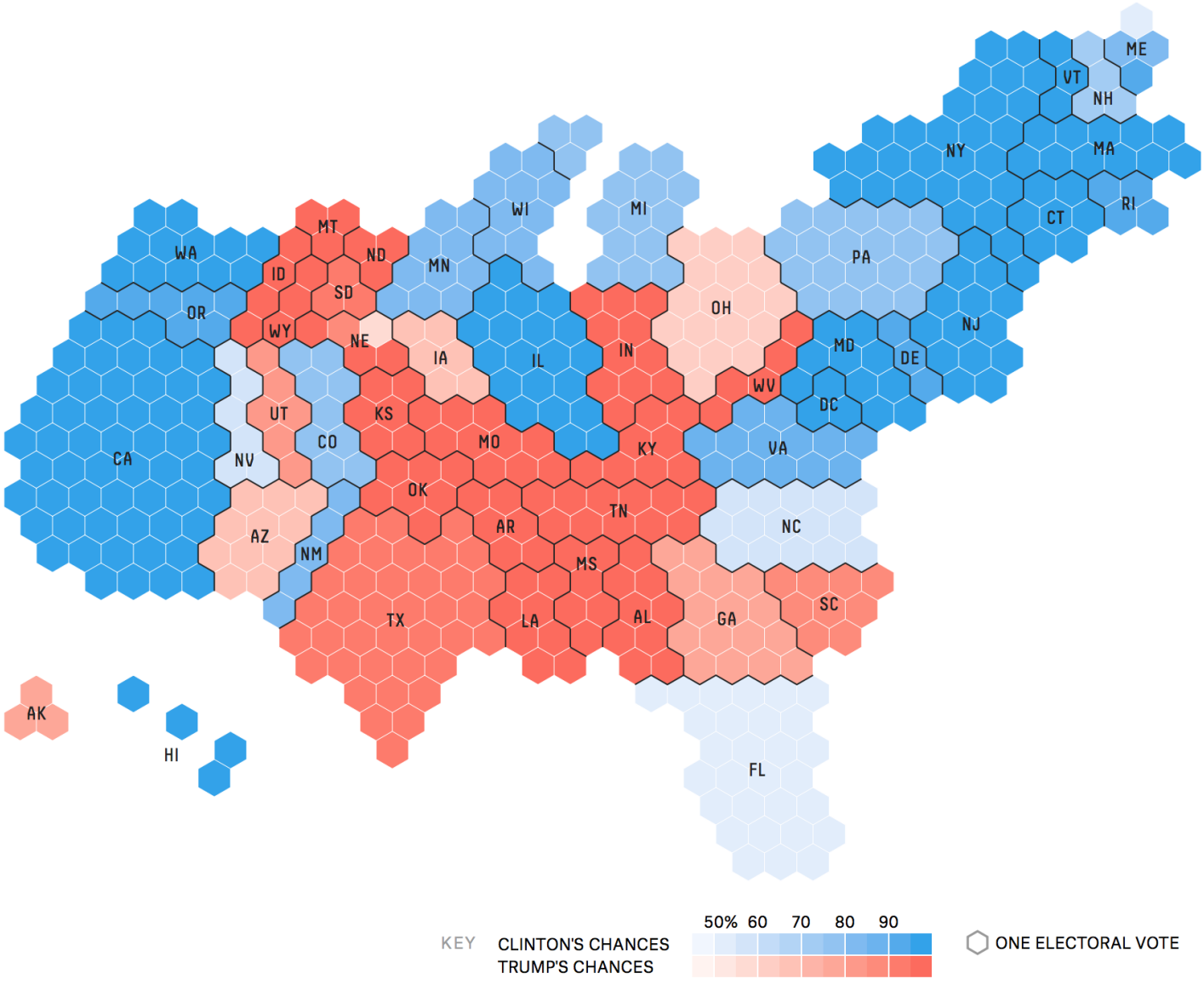
Examples of visualisation

One graphic can say a lot

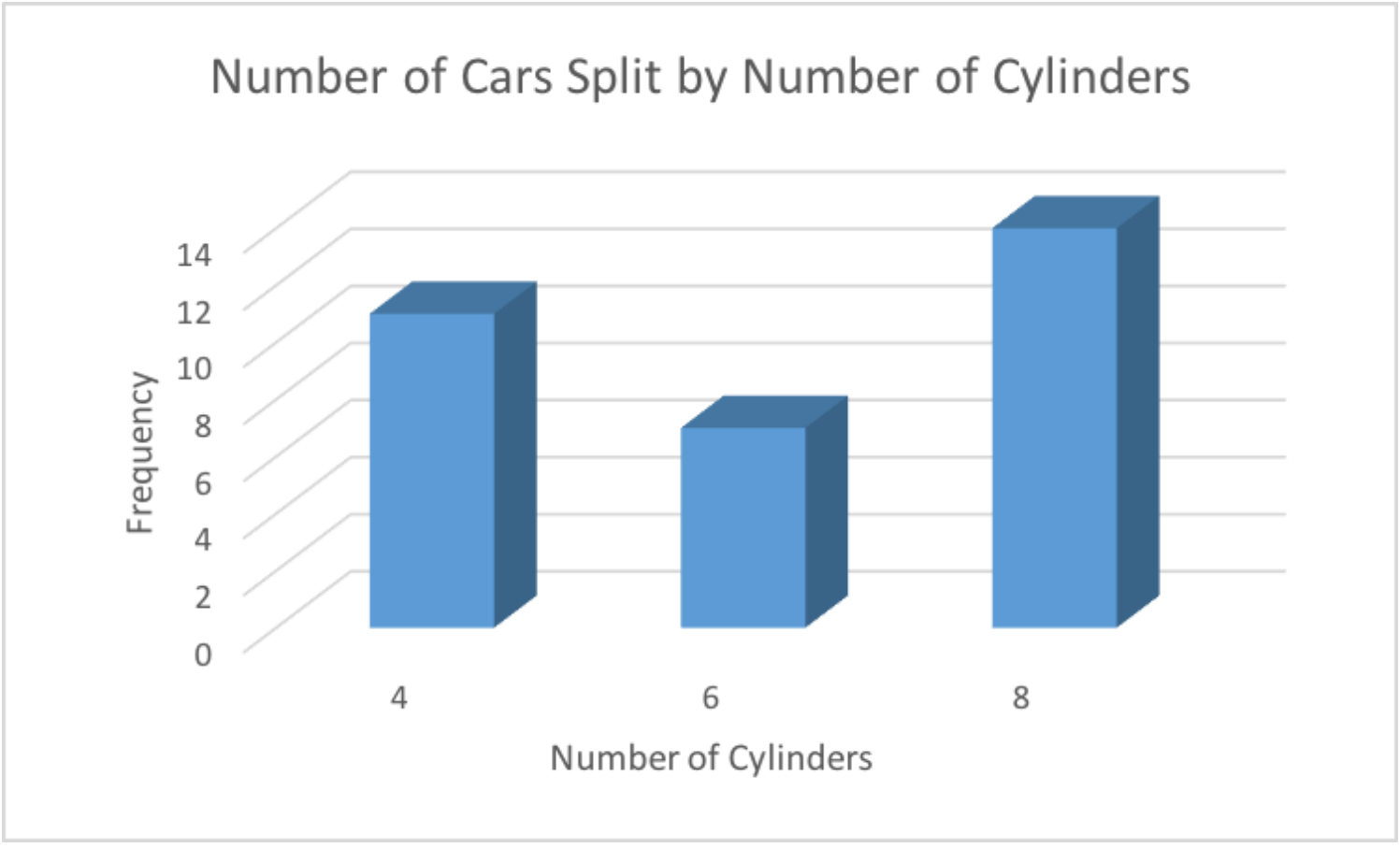
How you display the information is important



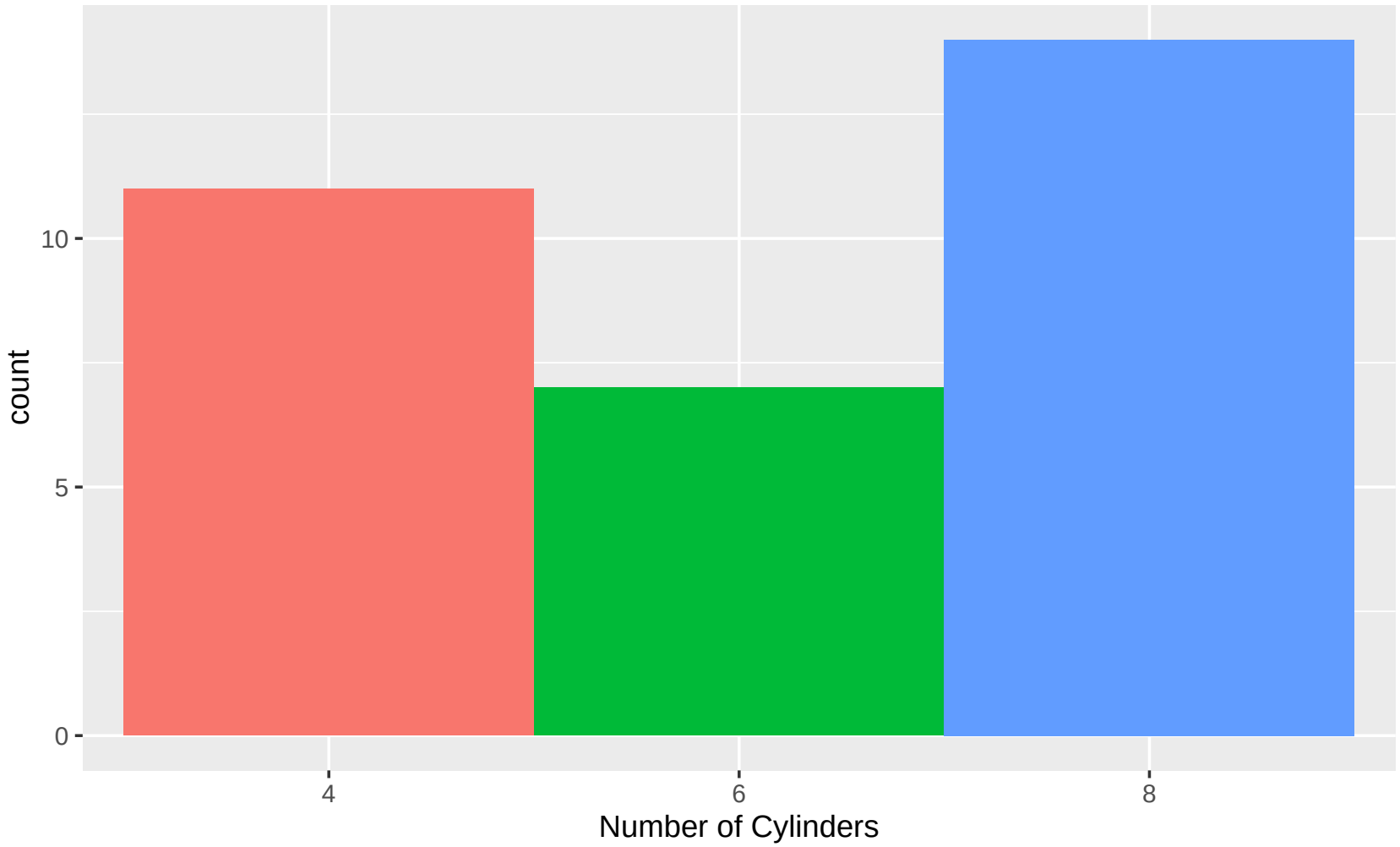
WITH A SIMPLE CHANGE, YOU SAY MORE



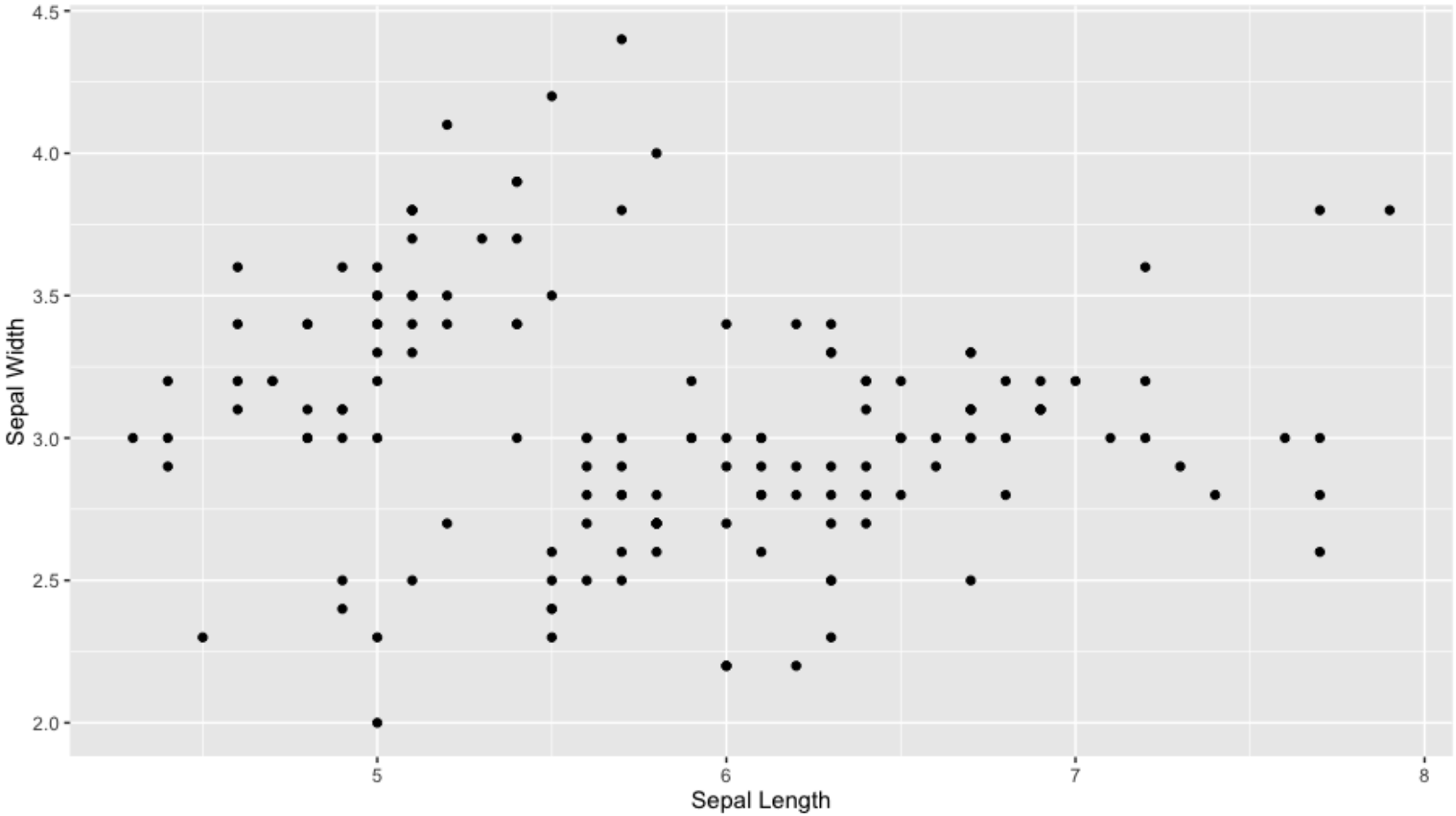
3D Graphics can be very misleading



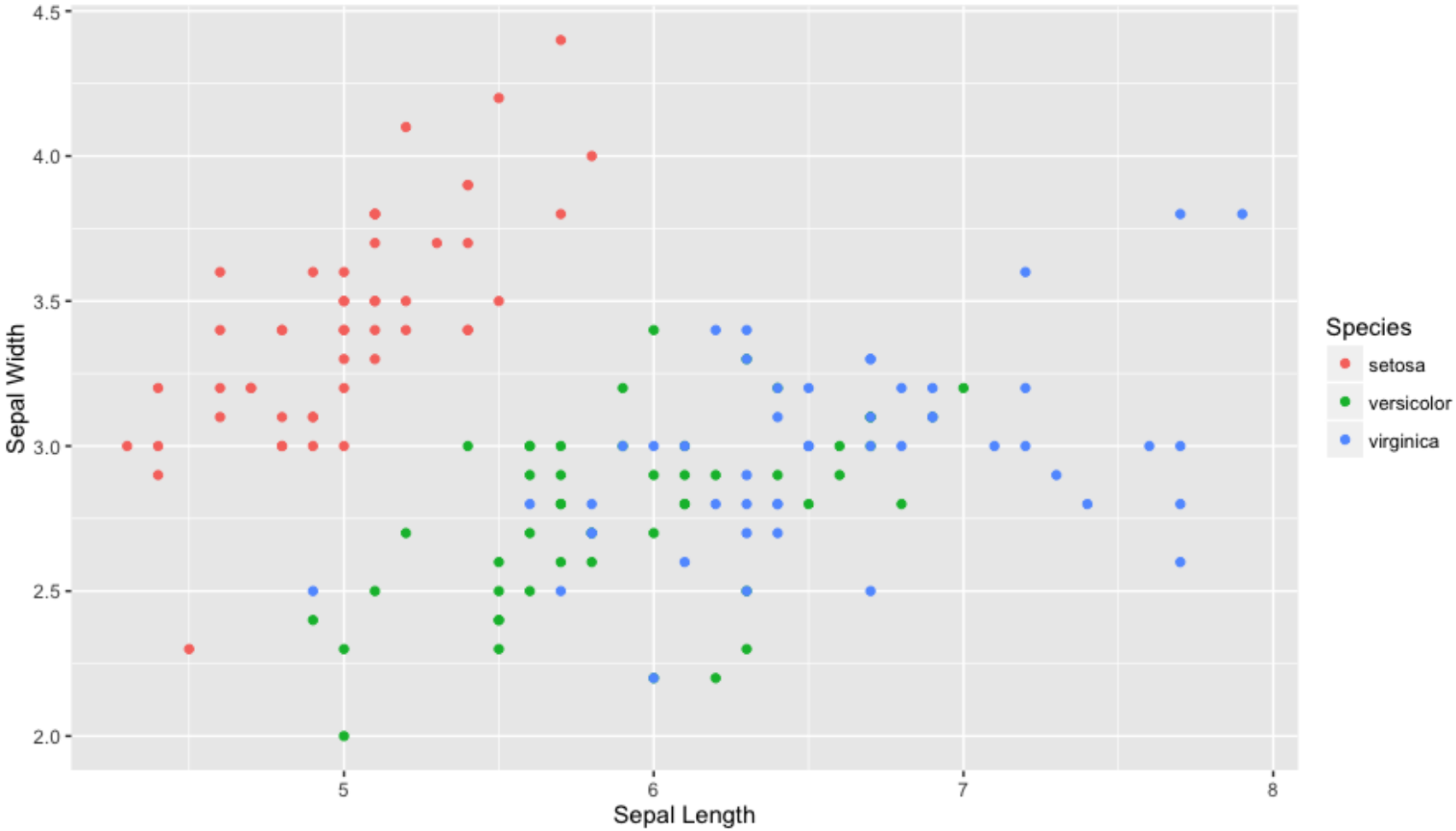
Stick to 2D where possible



USE COLOUR



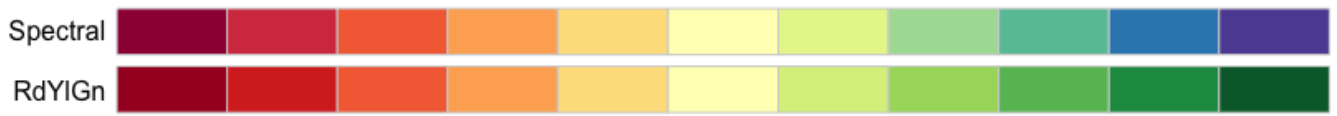
USE COLOUR


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

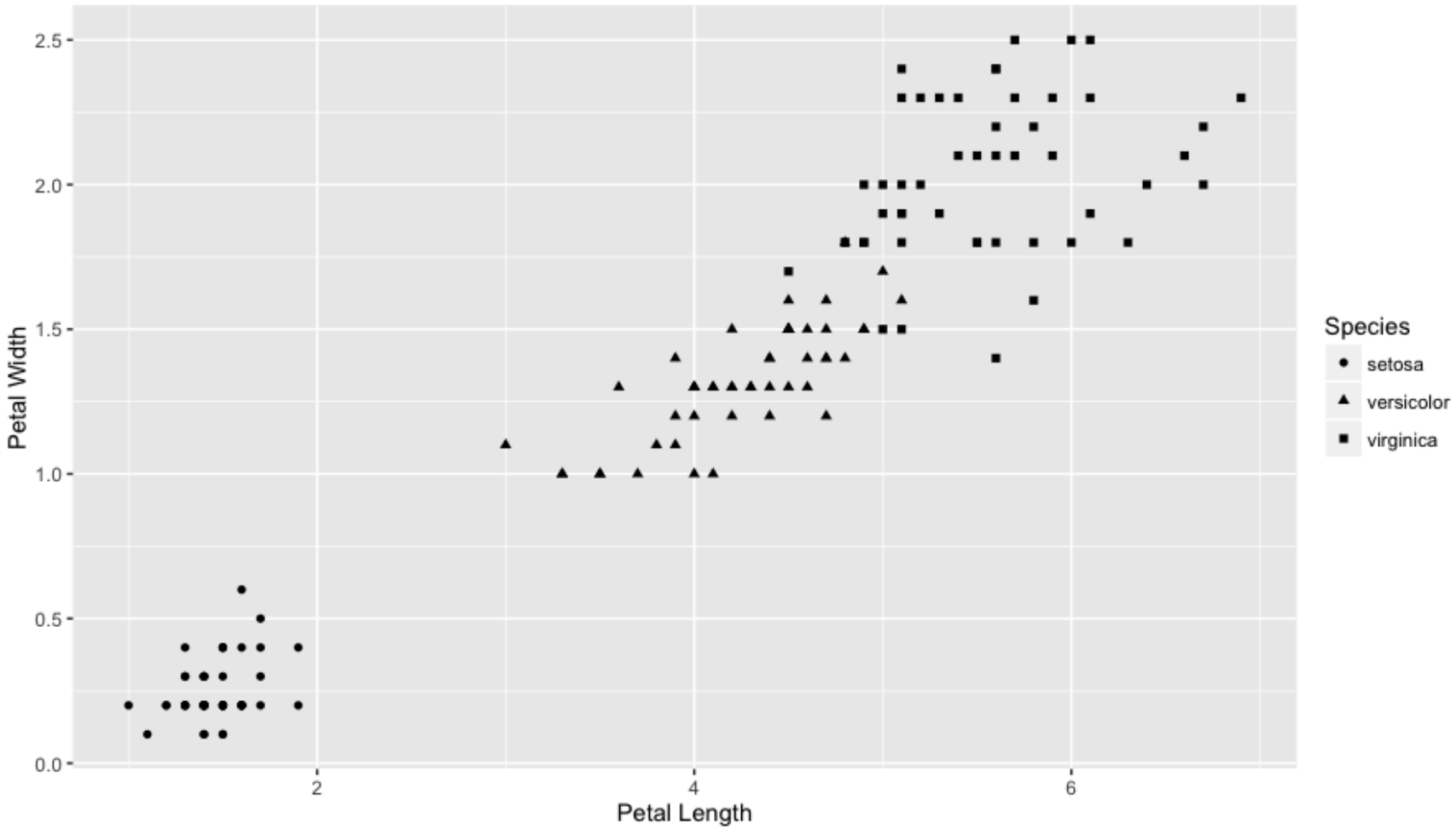
- Colour can be very helpful but there are practical issues.
- Colour scheme must be meaningful.
- Sequential colour schemes are good for ordered data, for example, population density.

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

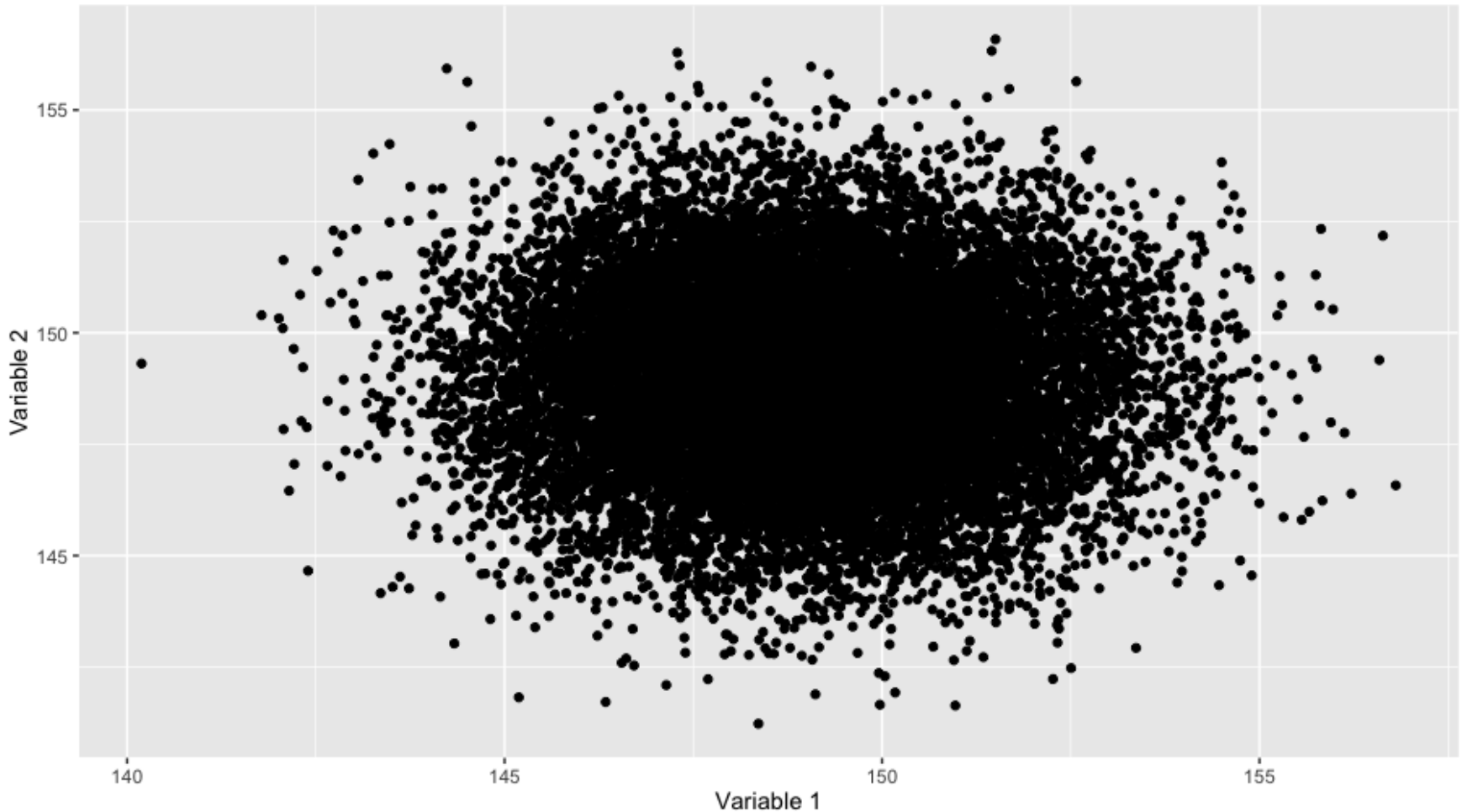


- A good choice of colour scheme are available from `http://colorbrewer2.org` and `RColorBrewer` R package.

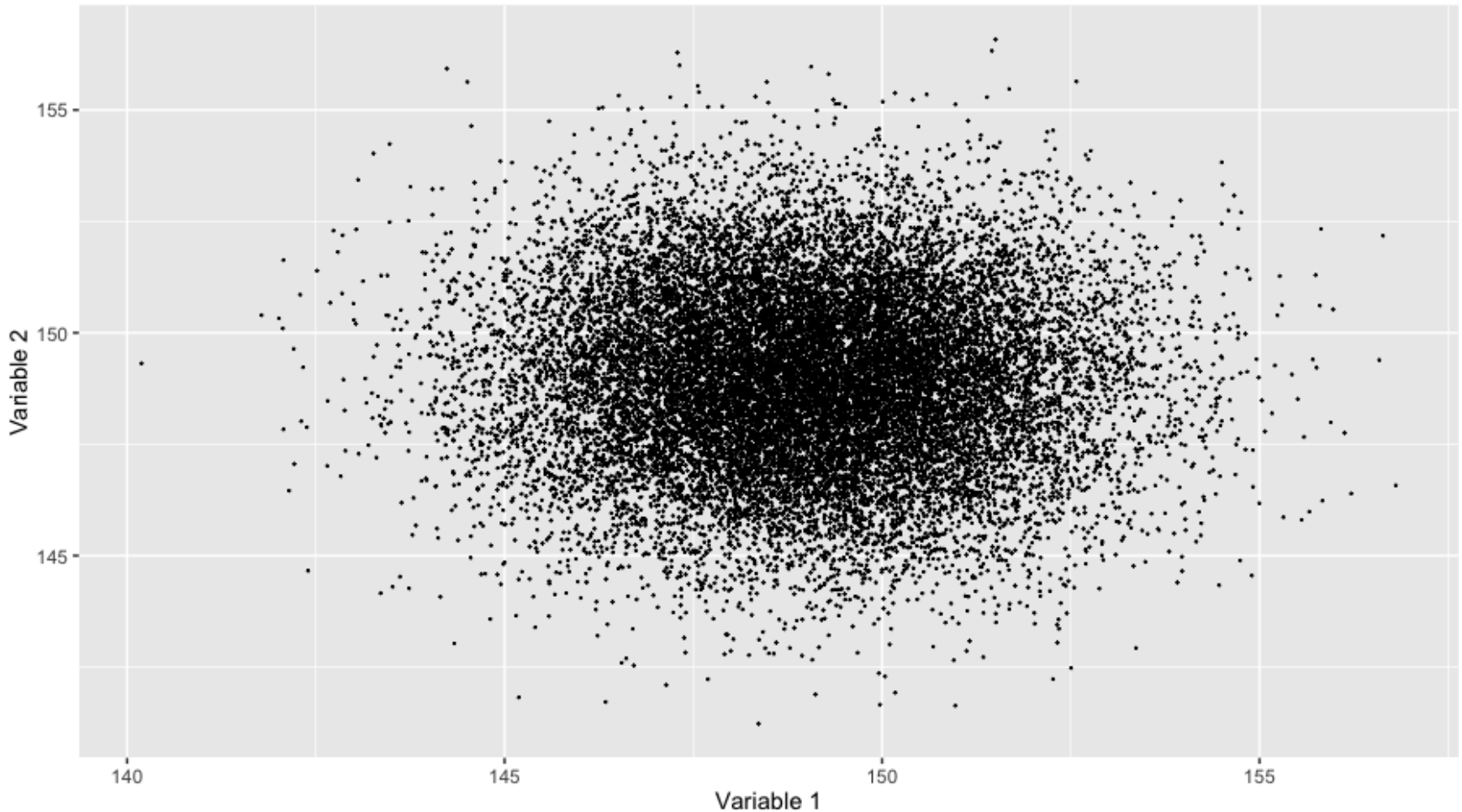
CHANGE THE STYLE



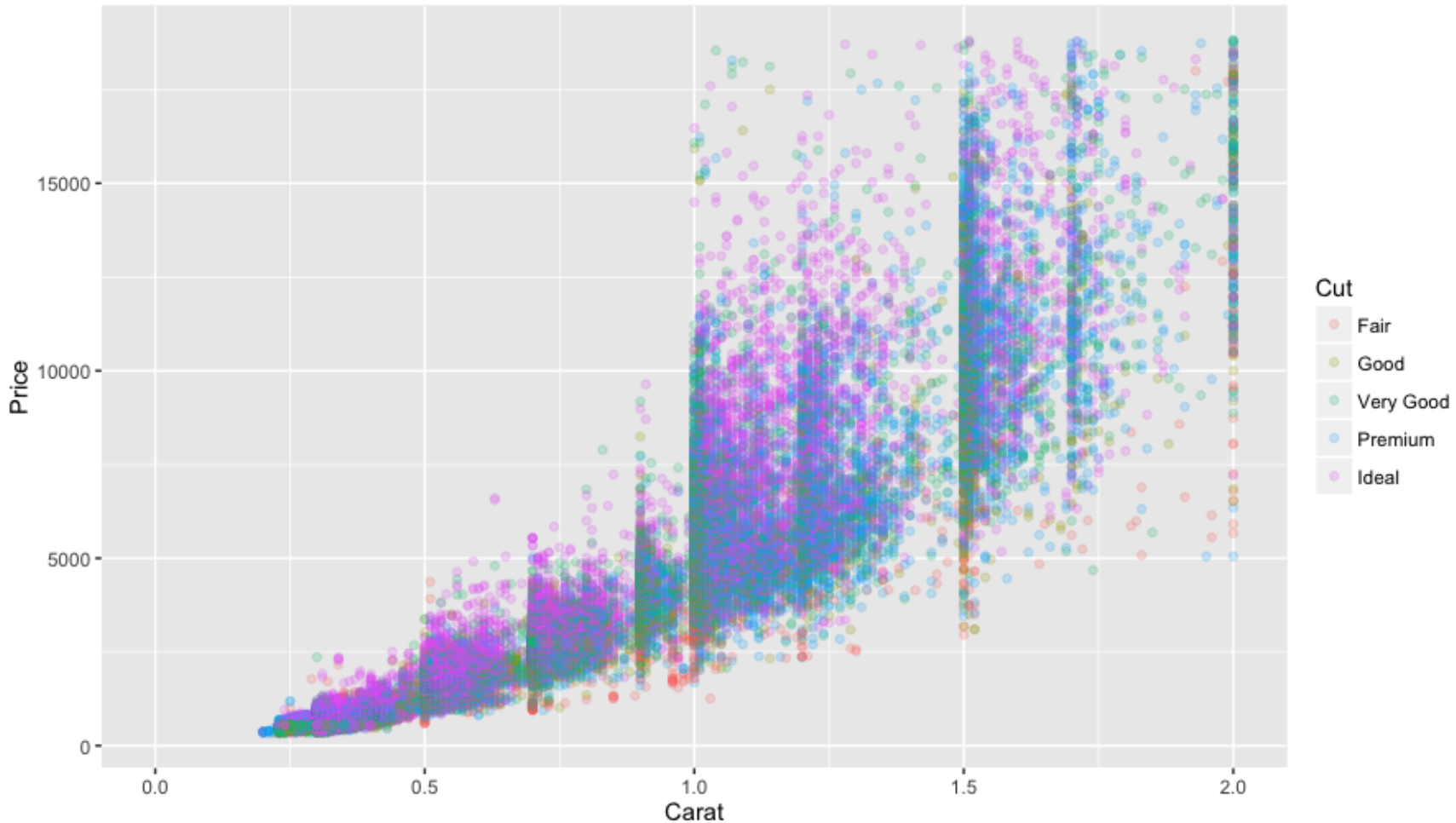
DISPLAYING BIG DATA CAN BE DIFFICULT



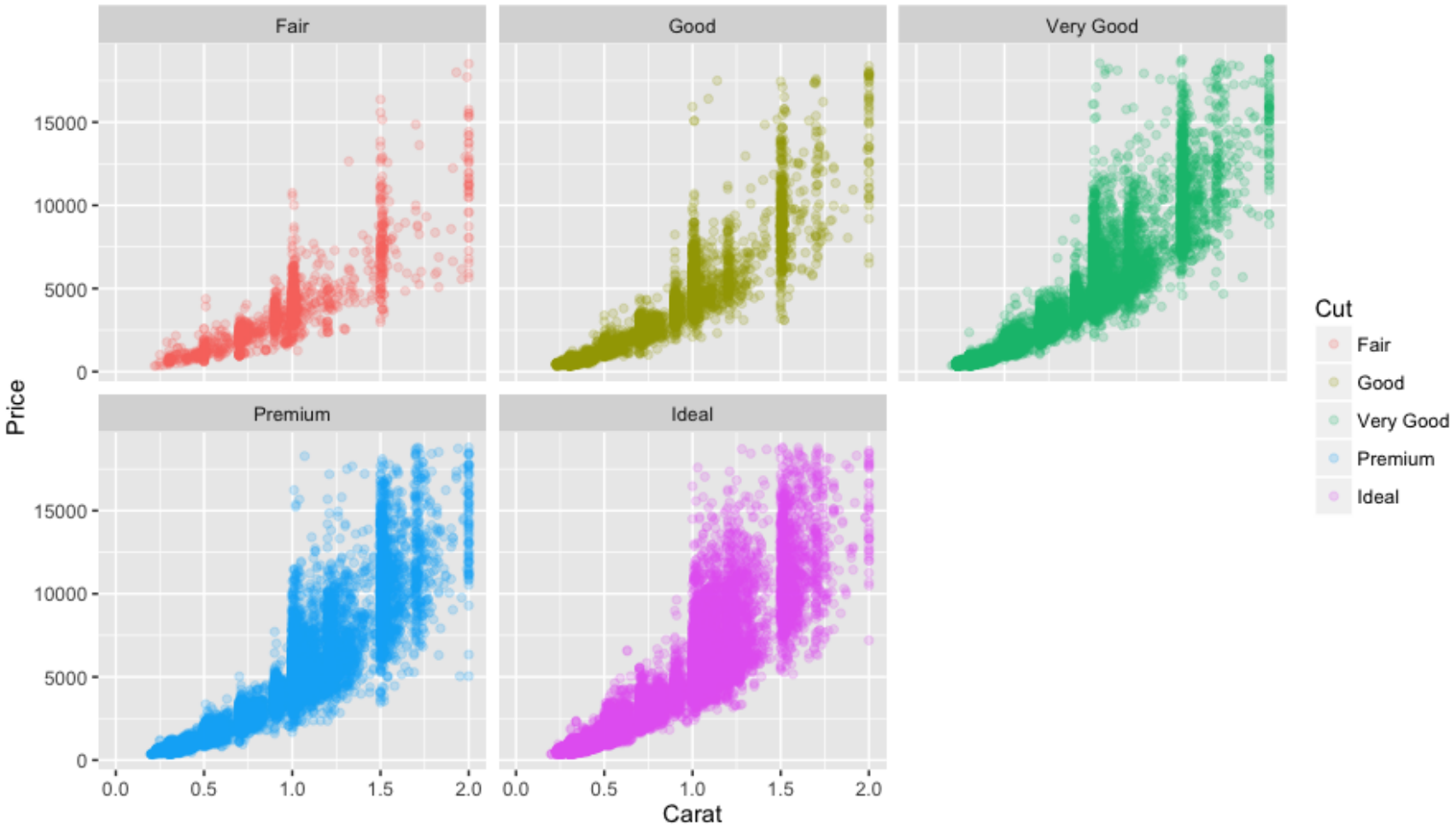
Change the scale



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

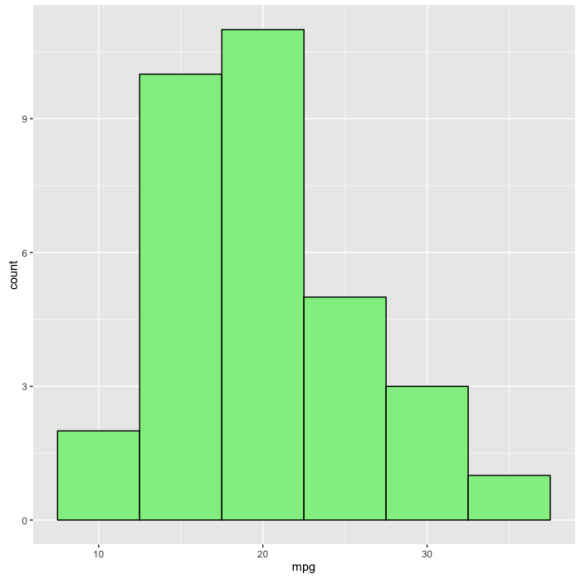
Professional graphics in R

- 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`.

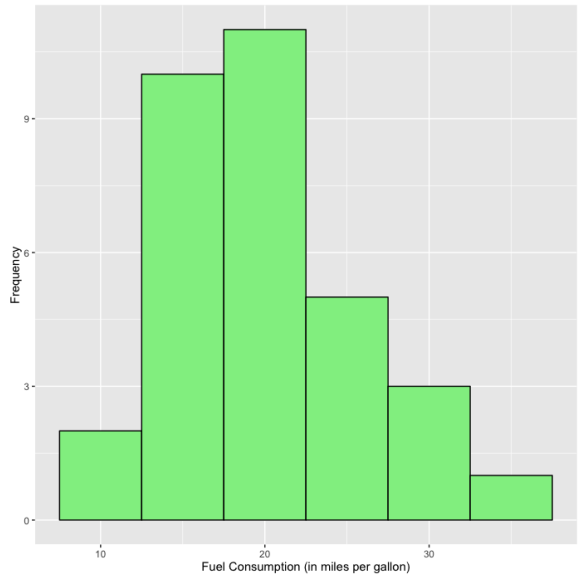
```
p1 <- ggplot(mtcars, aes(x=mpg)) + geom_histogram(binwidth=5,
colour='black',fill='lightgreen')
p1
```



PROFESSIONAL GRAPHICS IN R

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

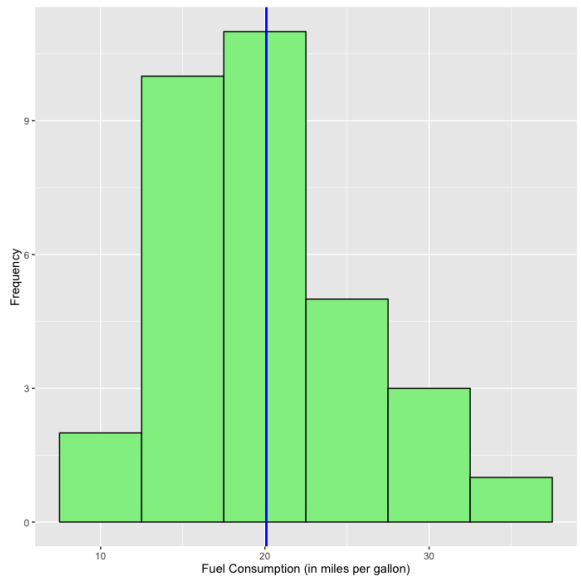
```
p1 <- p1 + labs(x = 'Fuel Consumption (in miles per gallon)',
                y='Frequency')
p1
```



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
```



Any Questions?