Graphs and data visualization

Topics

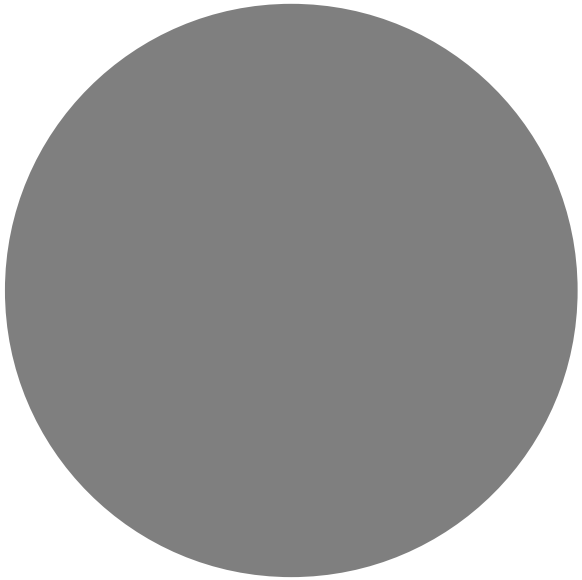
Why do you want to plot graphs?

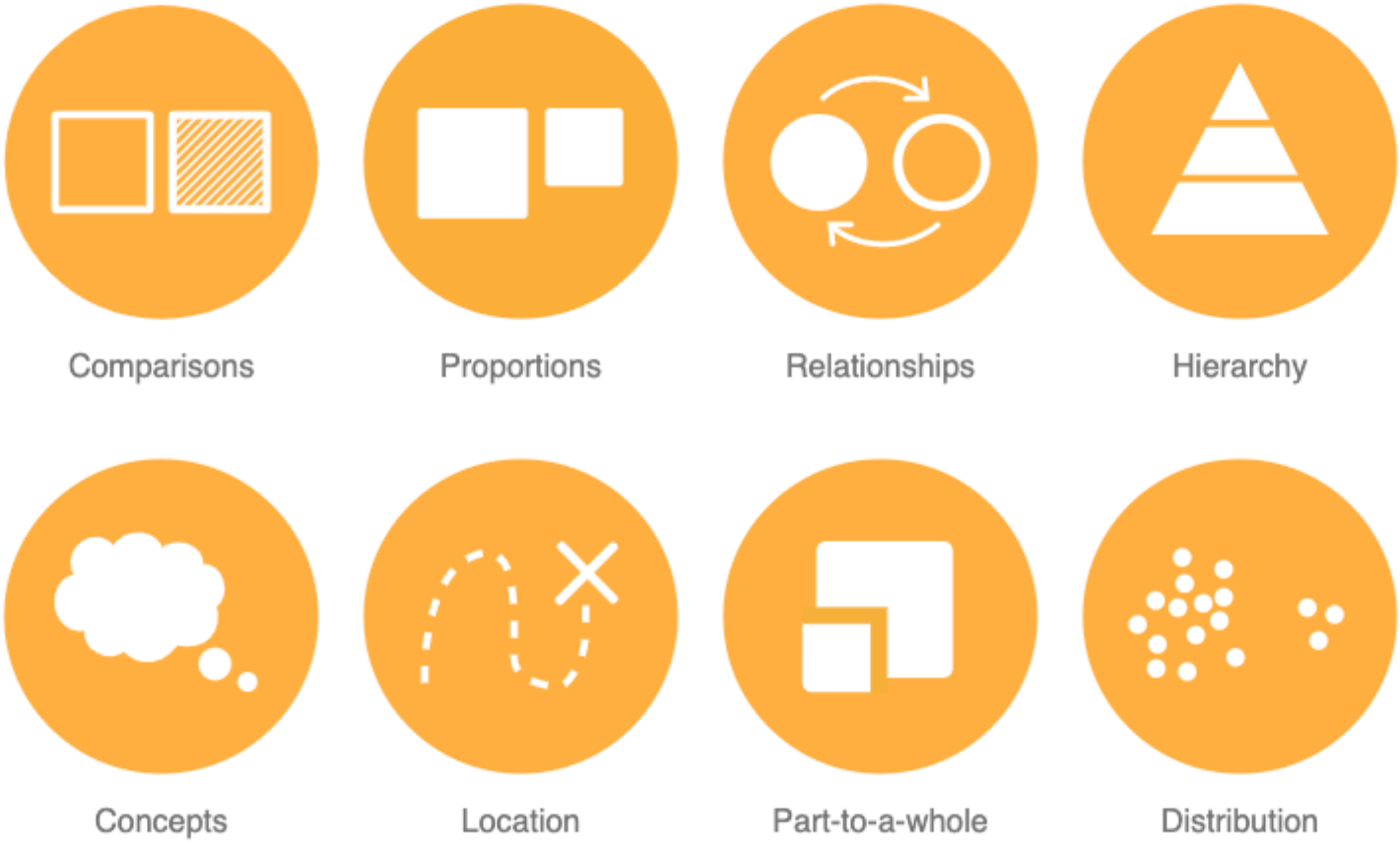
The importance of visualization

Good graphs, Bad graphs

Graphs in R

Why do you want to plot graphs?




Comparisons
Proportions
Relationships
Hierarchy
Concepts
Location
Part-to-a-whole
Distribution

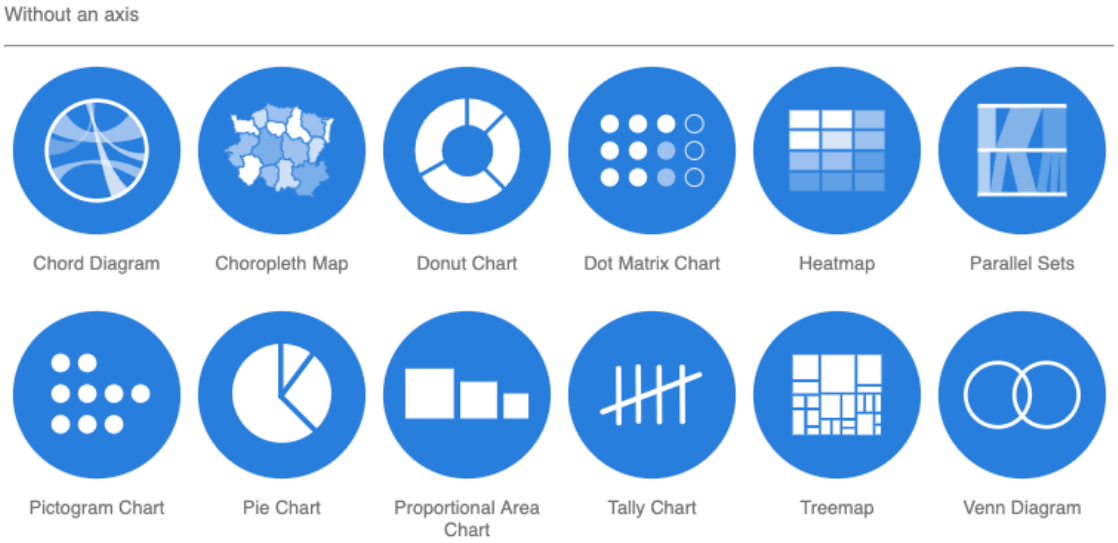

How things work
Processes & methods
Movement or flow
Patterns
Range
Data over time
Analysing text
Reference tool

Comparisons

Visualisation methods that help show the differences or similarities between values.

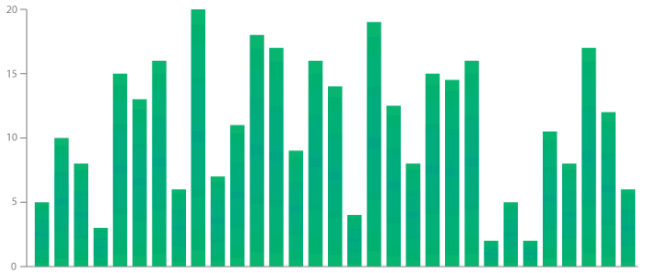
Comparisons

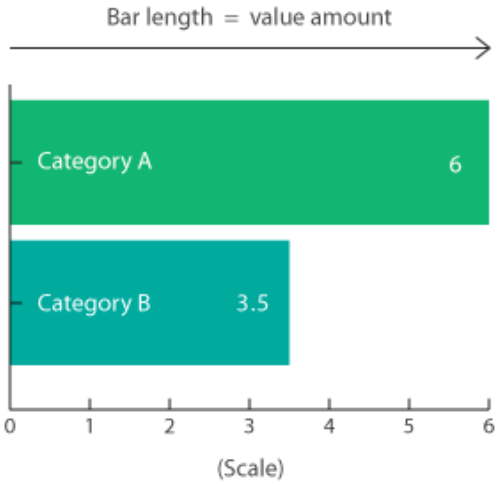
Visualisation methods that help show the differences or similarities between values.



Bar Chart

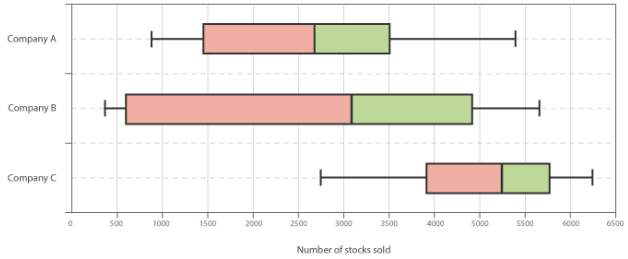
- As known as *Bar Graph* or *Column Graph*.
- The classic Bar Chart uses either horizontal or vertical bars (column chart) to show discrete, numerical comparisons across categories. One axis of the chart shows the specific categories being compared and the other axis represents a discrete value scale.

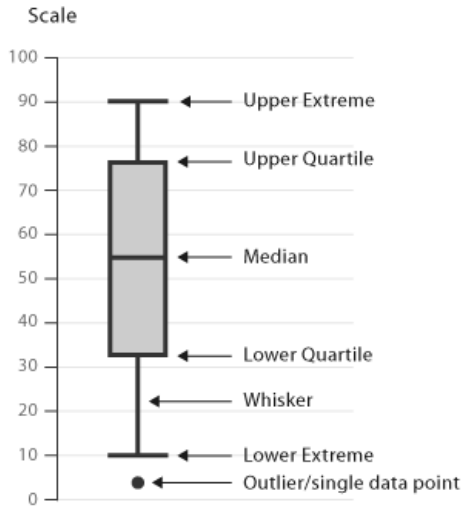




Box and Whisker Plot

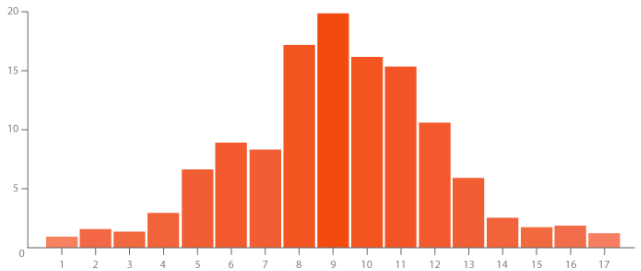
- A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles.
- The lines extending parallel from the boxes are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles. Outliers are sometimes plotted as individual dots that are in-line with whiskers. Box Plots can be drawn either vertically or horizontally.

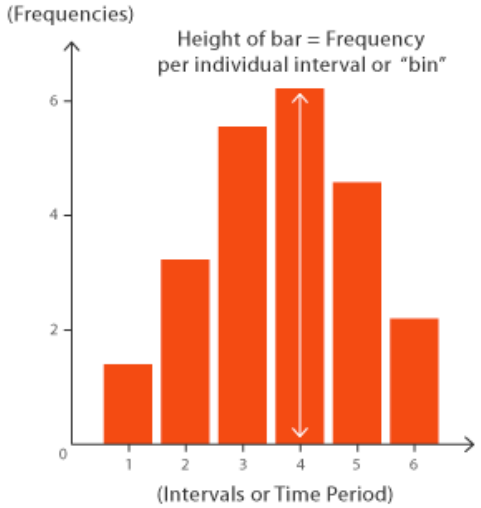




Histogram

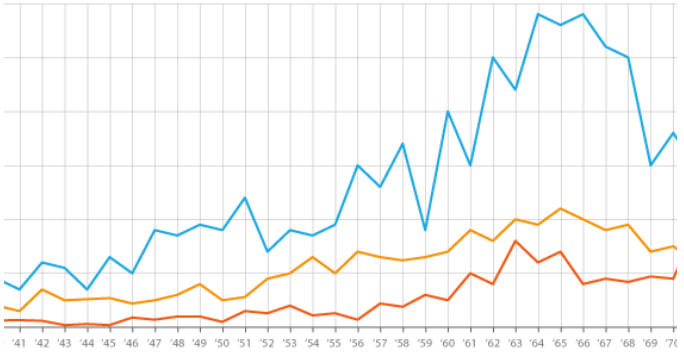
- A Histogram visualises the distribution of data over a continuous interval or certain time period. Each bar in a histogram represents the tabulated frequency at each interval/bin.

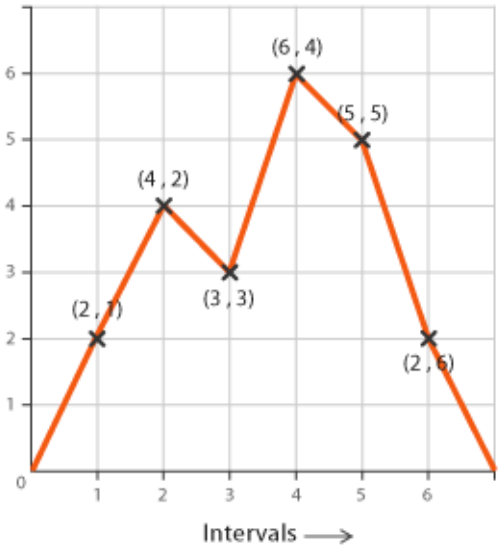




Line Graph

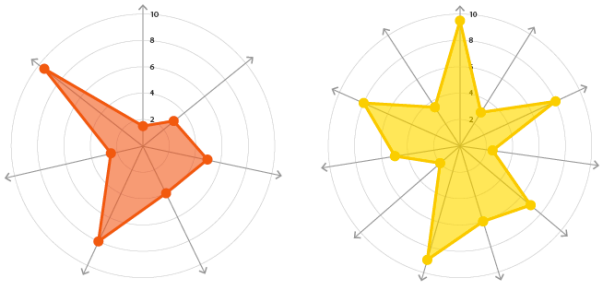
- Line Graphs are used to display quantitative values over a continuous interval or time period. A Line Graph is most frequently used to show trends and analyse how the data has changed over time.

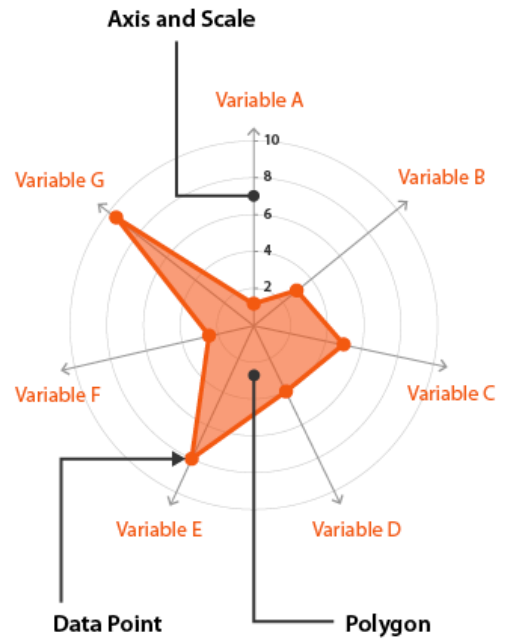




Radar Chart

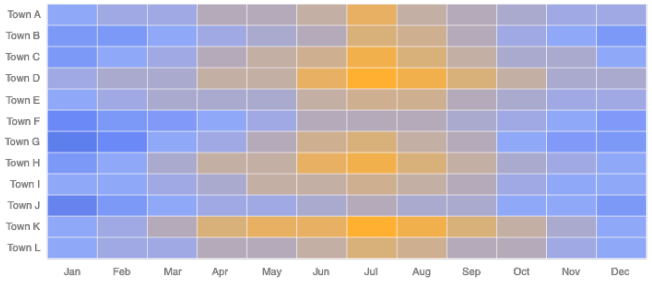
- As known as: *Spider Chart, Web Chart, Polar Chart, Star Plots*.
- Radar Charts are a way of comparing multiple quantitative variables. This makes them useful for seeing which variables have similar values or if there are any outliers amongst each variable. Radar Charts are also useful for seeing which variables are scoring high or low within a dataset, making them ideal for displaying performance.



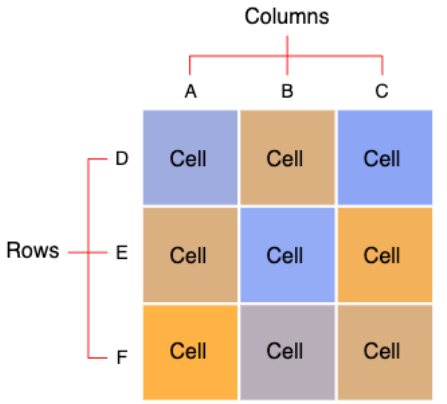


Heatmap

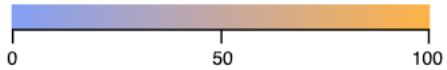
- Heatmaps visualise data through variations in colouring. When applied to a tabular format, Heatmaps are useful for cross-examining multivariate data, through placing variables in the rows and columns and colouring the cells within the table.
- Heatmaps are good for showing variance across multiple variables, revealing any patterns, displaying whether any variables are similar to each other, and for detecting if any correlations exist in-between them.



Heatmap using numerical data:

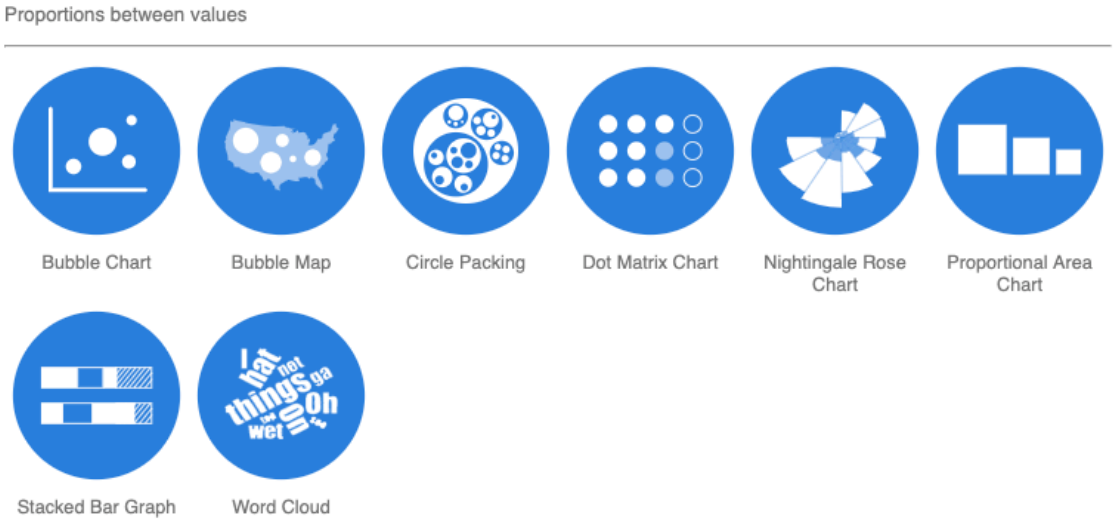


Value scale for determining cell colouring:

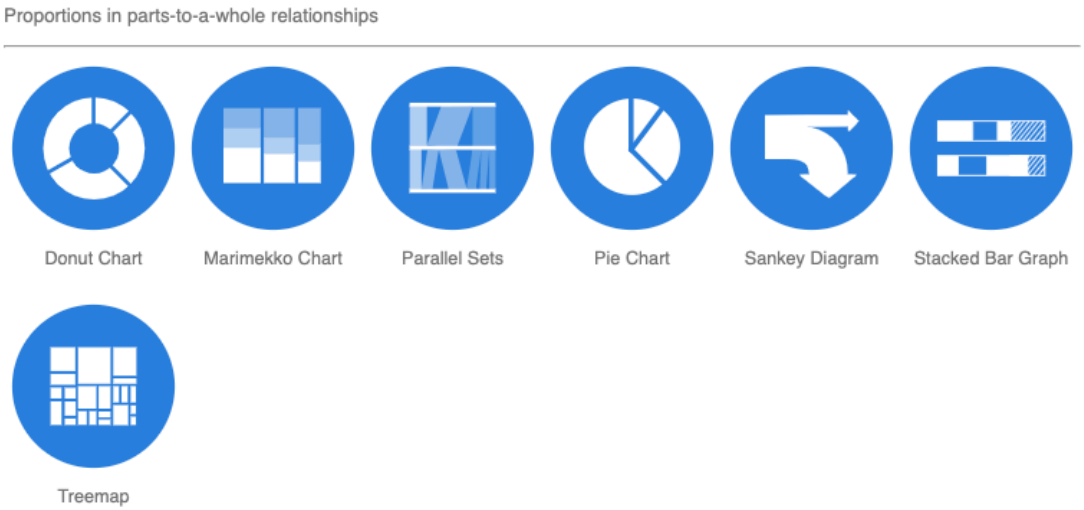


Alternative value scale broken into ranges:

Proportions
Visualization methods that use size or area to show differences or similarities between values or for parts to a whole.

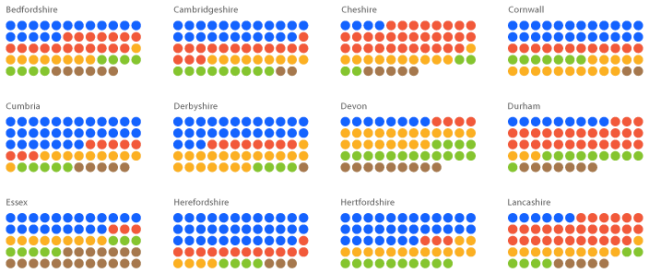


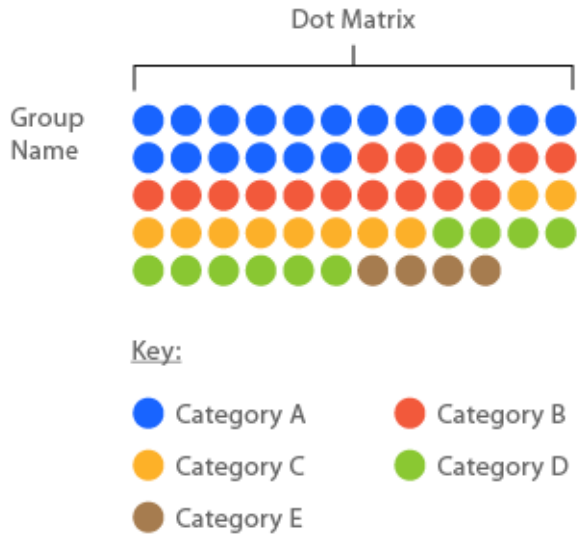
Proportions
Visualization methods that use size or area to show differences or similarities between values or for parts to a whole.



Dot Matrix Chart

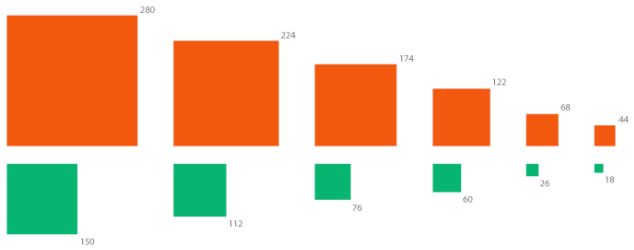
- Dot Matrix Charts display discreet data in units of dots, each coloured to represent a particular category and grouped together in a matrix. They are used to give a quick overview of the distribution and proportions of each category in a data set and also to compare distribution and proportion across other datasets, in order to discover patterns.

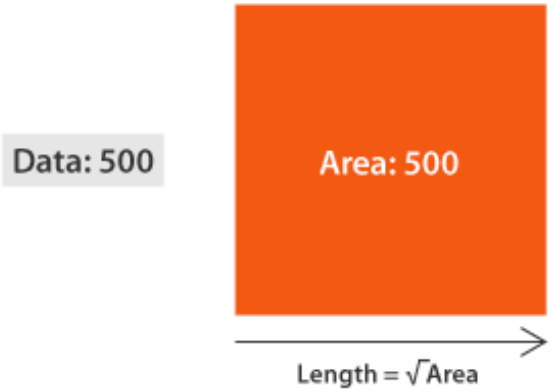




Proportional Area Chart

- Great for comparing values and showing proportions (in sizes, quantities etc) to give a quick, overall view of the relative sizes of the data, without the use of scales.
- The downside to this chart is that it's difficult to estimate values using Proportional Area Charts. This means they're almost exclusively used for communication purposes instead of analytical ones.



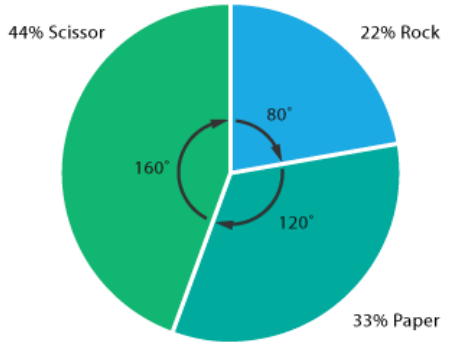


Word Cloud

- Also known as a *Tag Cloud*.
- A visualisation method that displays how frequently words appear in a given body of text, by making the size of each word proportional to its frequency. All the words are then arranged in a cluster or cloud of words. Alternatively, the words can also be arranged in any format: horizontal lines, columns or within a shape.

Pie Chart

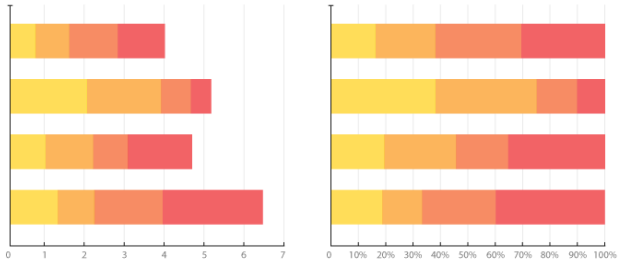
- Extensively used in presentations and offices, Pie Charts help show proportions and percentages between categories, by dividing a circle into proportional segments.
- Each arc length represents a proportion of each category, while the full circle represents the total sum of all the data, equal to 100%.



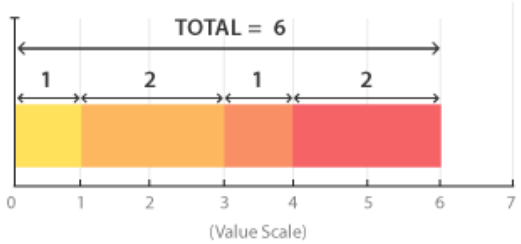
Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Stacked Bar Chart

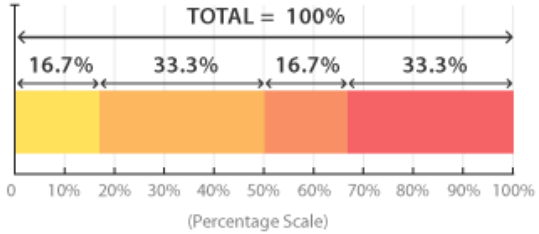
- Stacked Bar Graphs segment their bars of multiple datasets on top of each other. They are used to show how a larger category is divided into smaller categories and what the relationship of each part has on the total amount. There are two types of Stacked Bar Graphs:
  - **Simple Stacked Bar Graphs**
  - **100% Stack Bar Graphs**



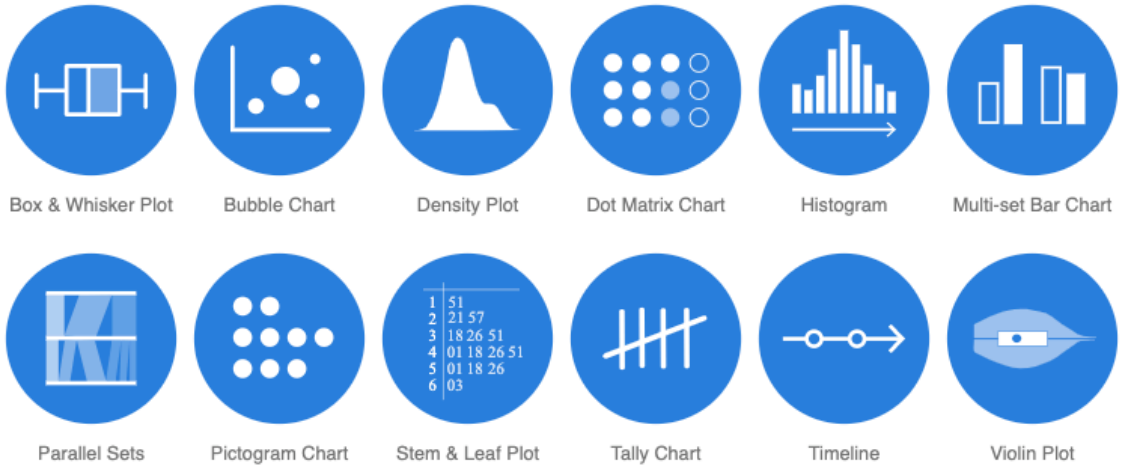
Simple



100%

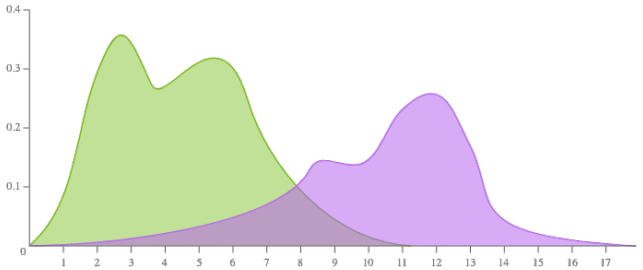


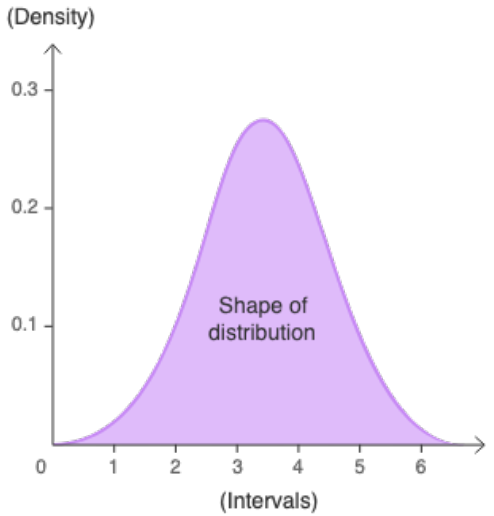
Distribution
Visualization methods that display frequency, how data spread out over an interval or is grouped.



Density Plot

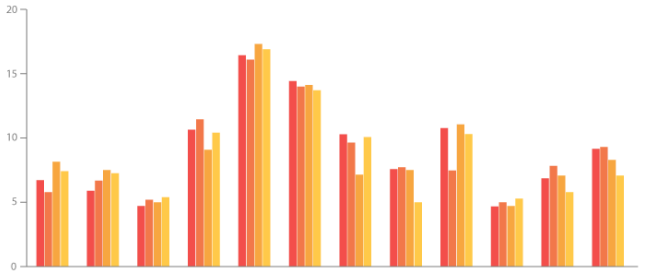
- As known as *Kernel Density Plots, Density Trace Graph*.
- A Density Plot visualises the distribution of data over a continuous interval or time period. This chart is a variation of a Histogram that uses kernel smoothing to plot values, allowing for smoother distributions by smoothing out the noise. The peaks of a Density Plot help display where values are concentrated over the interval.

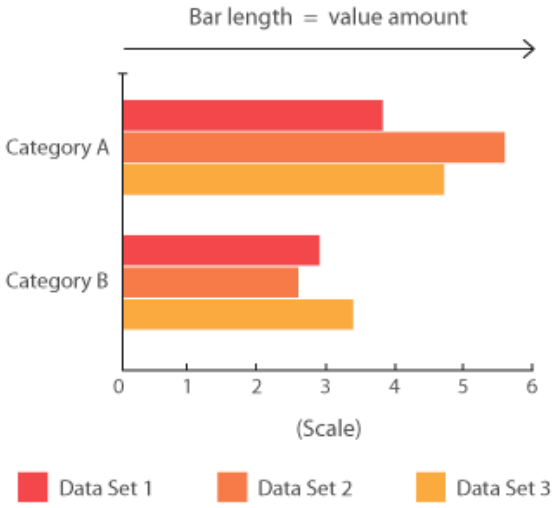




Multi-set Bar Chart

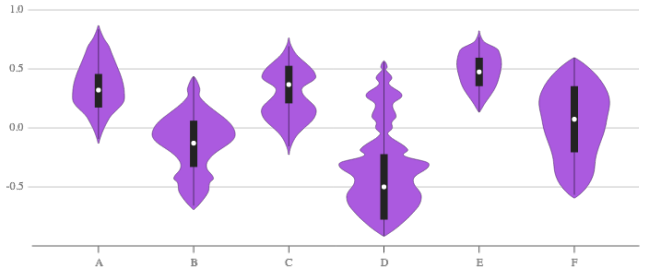
- Also known as a *Grouped Bar Chart* or *Clustered Bar Chart*.
- This variation of a Bar Chart is used when two or more data series are plotted side-by-side and grouped together under categories, all on the same axis.

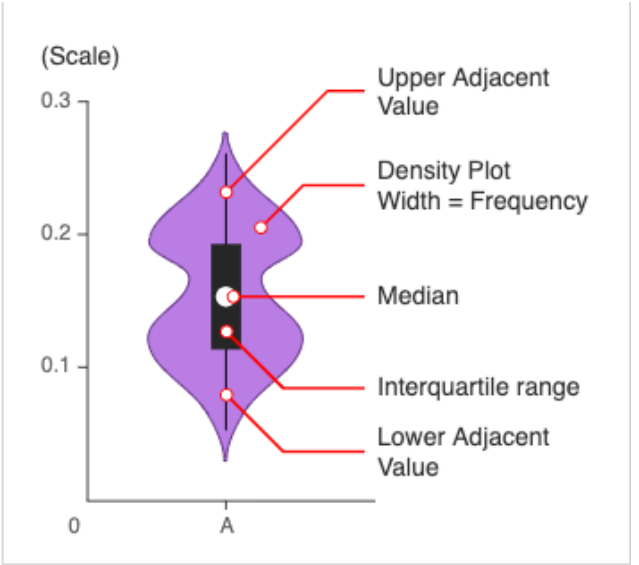




Violin Plot

- A Violin Plot is used to visualise the distribution of the data and its probability density.
- This chart is a combination of a Box Plot and a Density Plotthat is rotated and placed on each side, to show the distribution shape of the data. The white dot in the middle is the median value and the thick black bar in the centre represents the interquartile range. The thin black line extended from it represents the upper (max) and lower (min) adjacent values in the data. Sometimes the graph marker is clipped from the end of this line.

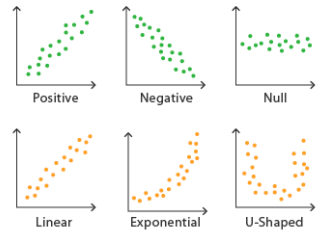


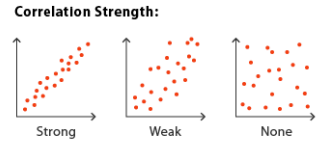


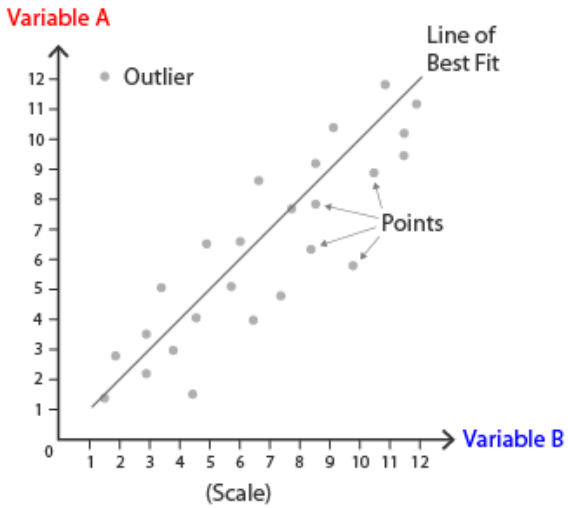
Pattern
Visualization methods that can reveal forms or patterns in the data to give it meaning.

Scatterplot

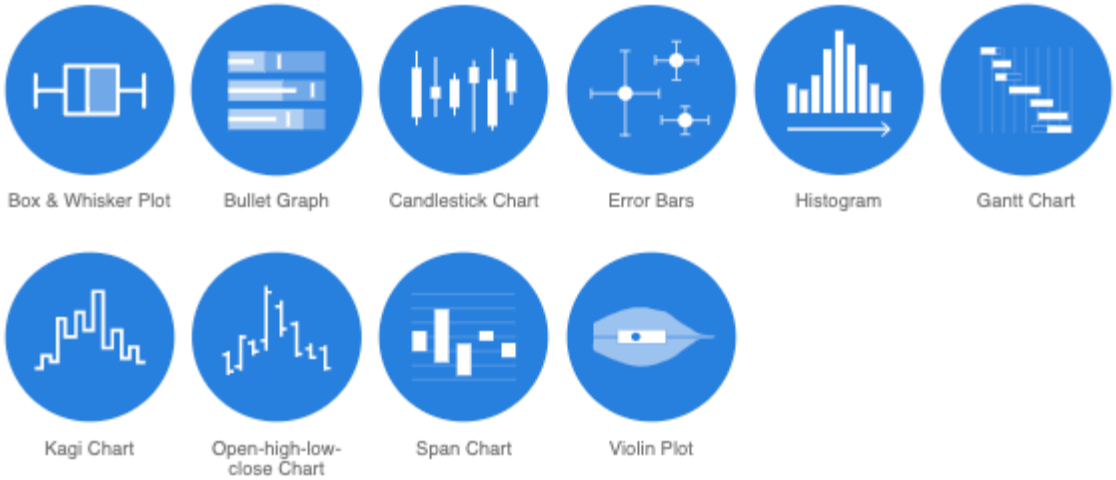
- Also known as a *Scatter Graph, Point Graph, X-Y Plot, Scatter Chart or Scattergram*.
- Scatterplots use a collection of points placed using Cartesian Coordinates to display values from two variables. By displaying a variable in each axis, you can detect if a relationship or correlation between the two variables exists.





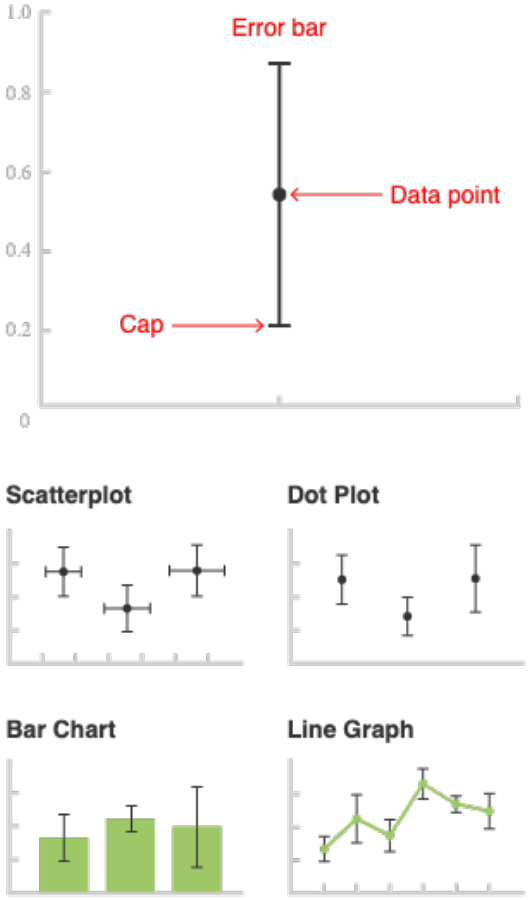


Range
Visualization methods that display the variations between upper and lower limits on a scale.



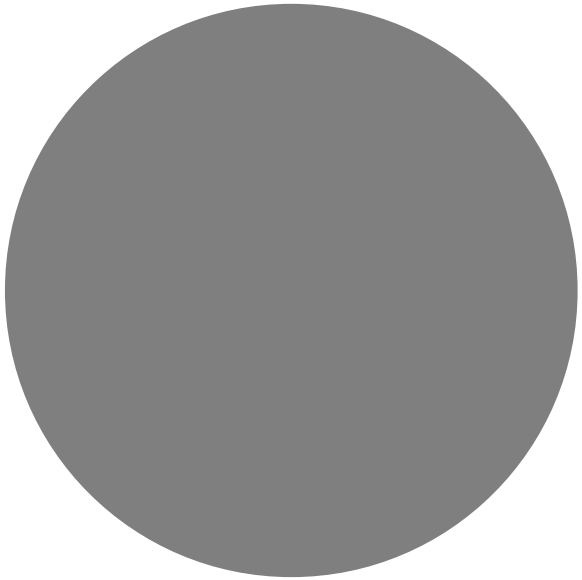
Error bars

- Error Bars function as a graphical enhancement that visualises the variability of the plotted data on a Cartesian graph.
- Error Bars can be applied to graphs such as Scatterplots, Dot Plots, Bar Charts or Line Graphs, to provide an additional layer of detail on the presented data.



End of Segment

Let’s take a break



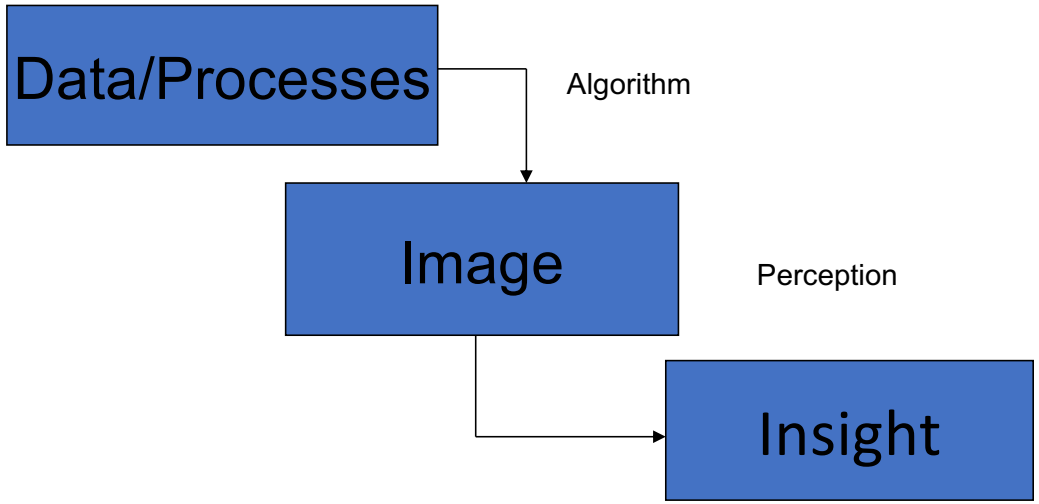
The importance of visualization

- Data visualization is the process of converting raw data into easily understood pictures of information that enable faster and effective exploration, discovery, insight, and decision-making

What is data visualization?

Data -> Visuals

The “transformation from numbers to insight requires two stages.”
Jacques Bertin



X marks the spot

- Much of our communication is done via words.
- The specific arrangement of words conveys meaning.
- Can meaning also be conveyed via pictorial means?

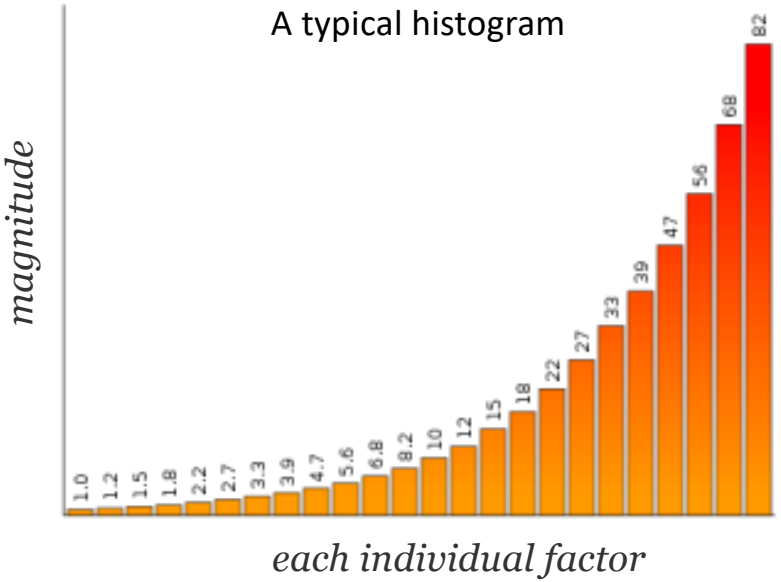
Abstraction without words (marks)

- A mark is made to represent some information other than itself. It is also referred to as a sign.
- Marks can be
  - **Points** are dimensionless locations on the plane, represented by signs that obviously need to have some size, shape or color for visualization.
  - **Lines** represent information with a certain length, but no area and therefore no width. Again lines are visualized by signs of some thickness.
  - **Areas** have a length and a width and therefore a two-dimensional size.
  - **Surfaces** are areas in a three-dimensional space, but with no thickness.
  - **Volumes** have a length, a width and a depth. They are thus truly three-dimensional.

Adding value to visual representation of data using perception

But after using up those two dimensions, what other attributes can you use?

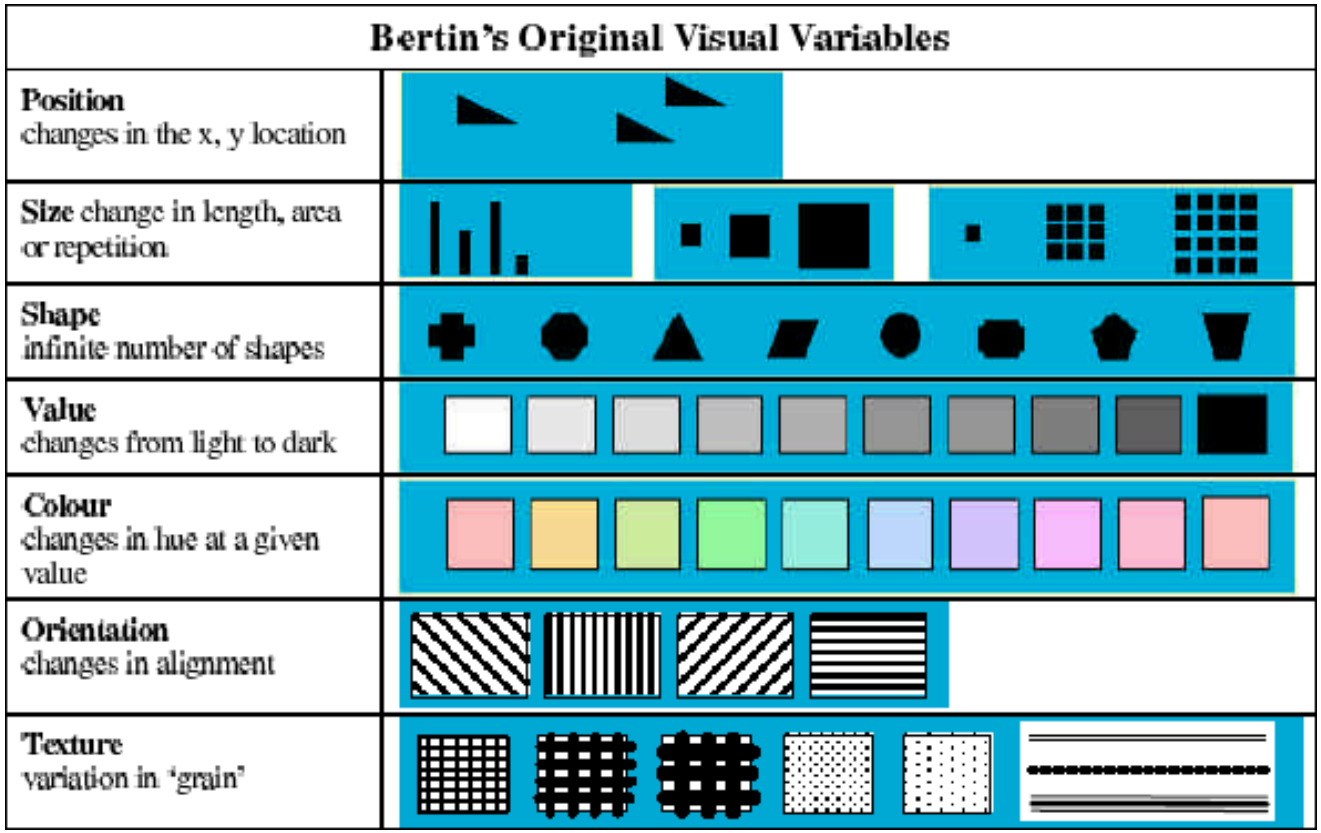
A typical histogram



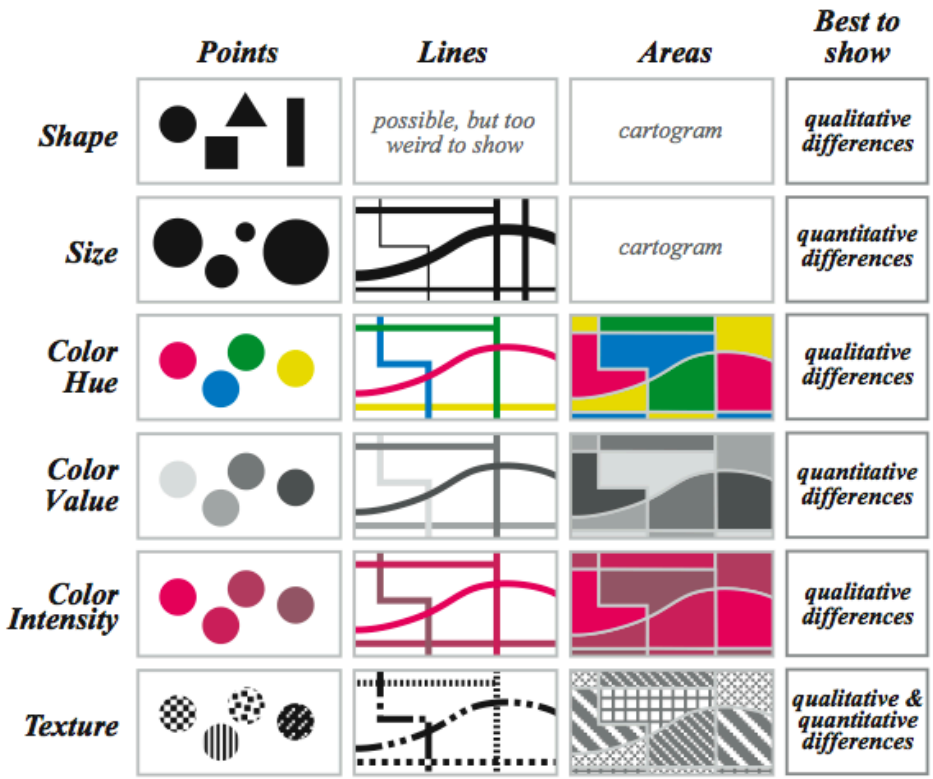
For depicting additional factors, you then have to choose between size, color, value, texture, line orientation or shape

Not all retinal variables are equally effective in their ability to represent information.

The 7 retinal variables



Idealizing Bertin's visual (retinal) variables



Making Maps: A Visual Guide to Map Design for GIS by John Krygier and Denis Wood

What are the benefits of Data Visualization?

Data visualization allows users see several different **perspectives** of the data.

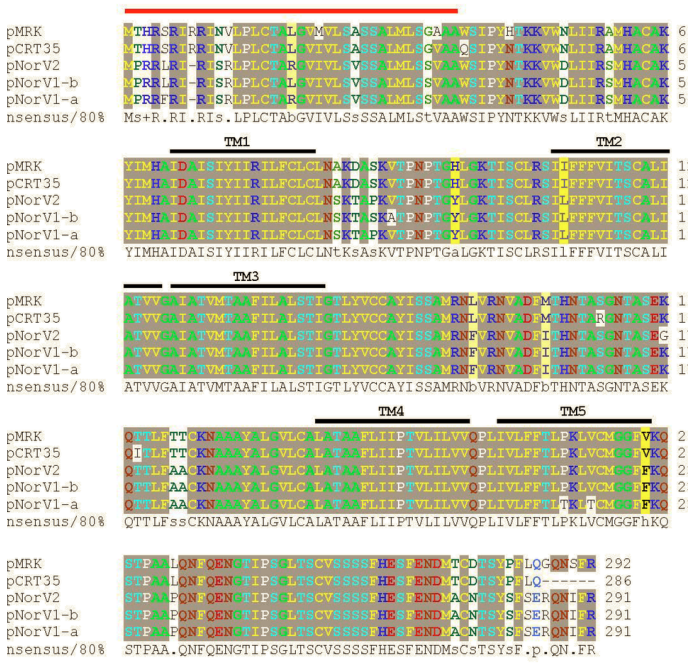
Data visualization makes it possible to **interpret vast amounts** of data

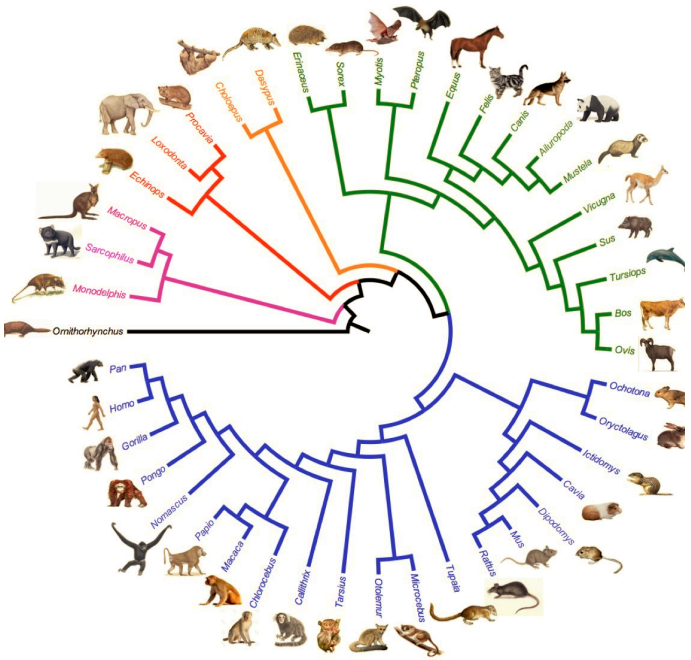
Data visualization offers the ability to note **exceptions** in the data.

Data visualization allows the user to **analyze visual patterns** in the data.

Exploring trends within a database through visualization by letting analysts **navigate through data** and visually orient themselves to the patterns in the data.

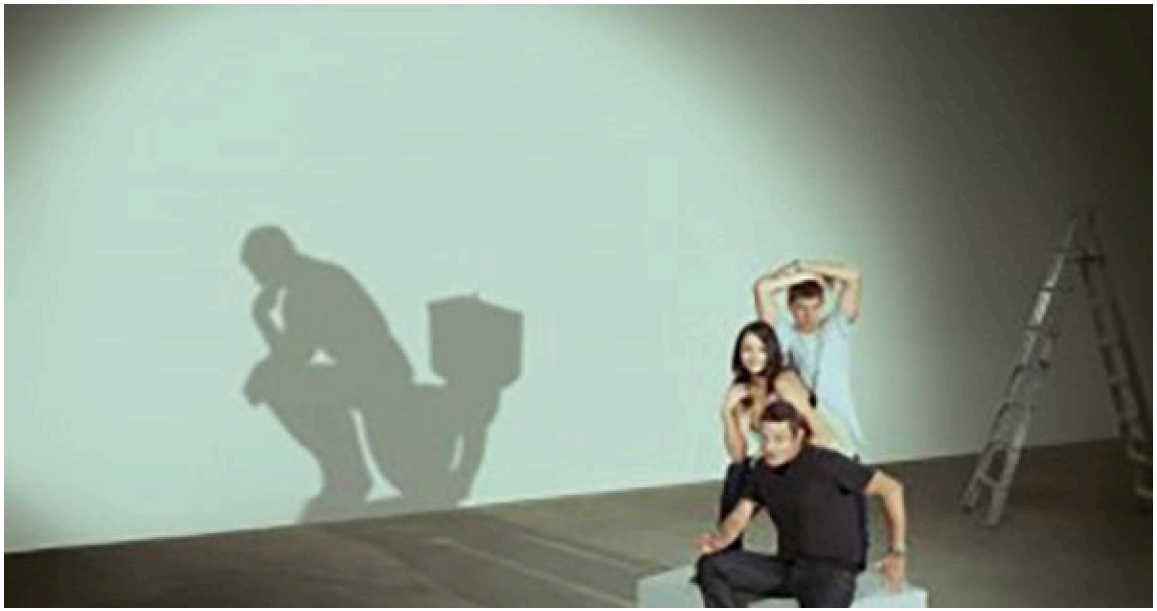
A picture paints a thousand words





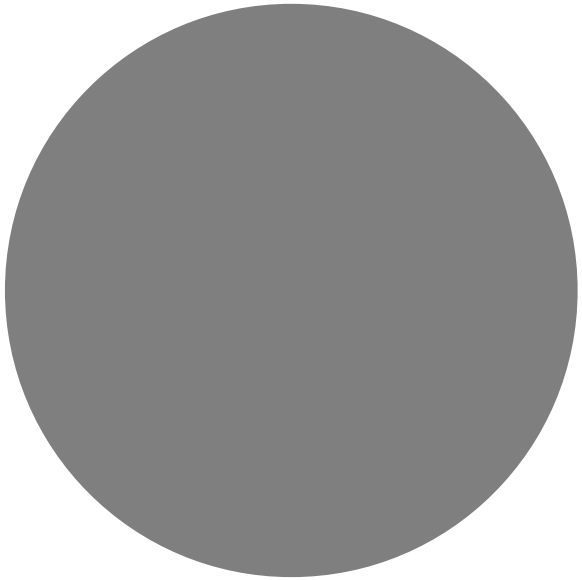
A picture paints a thousand words

- Data presentation forms the foundation of our collective scientific knowledge
- A picture may paint a thousand words, BUT a picture can also mislead.



End of Segment

Let’s take a break

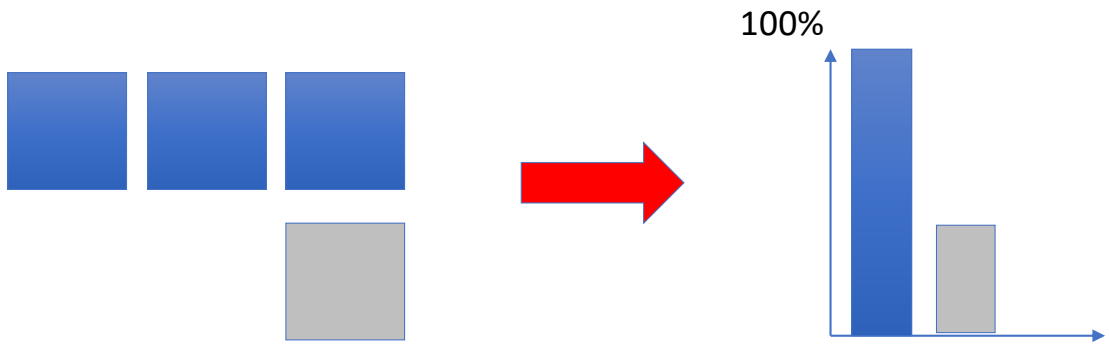


Good graphs/ Bad graphs

Some rules of thumb

Small Data size

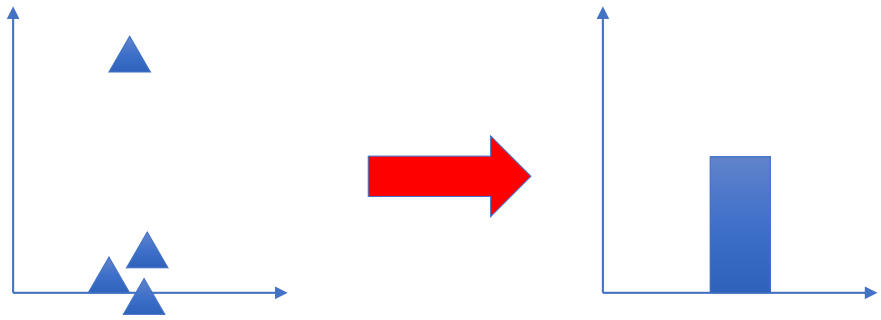
- It does not make sense to use graphs to display very small amounts of data.
- The human brain is quite capable of grasping one two, or even three values.



Data quality

- Graphs are only as good as the data they display
- No amount of creativity can produce a good graph from dubious data
- Anything less would be trying to lie via misleading representation

Data quality



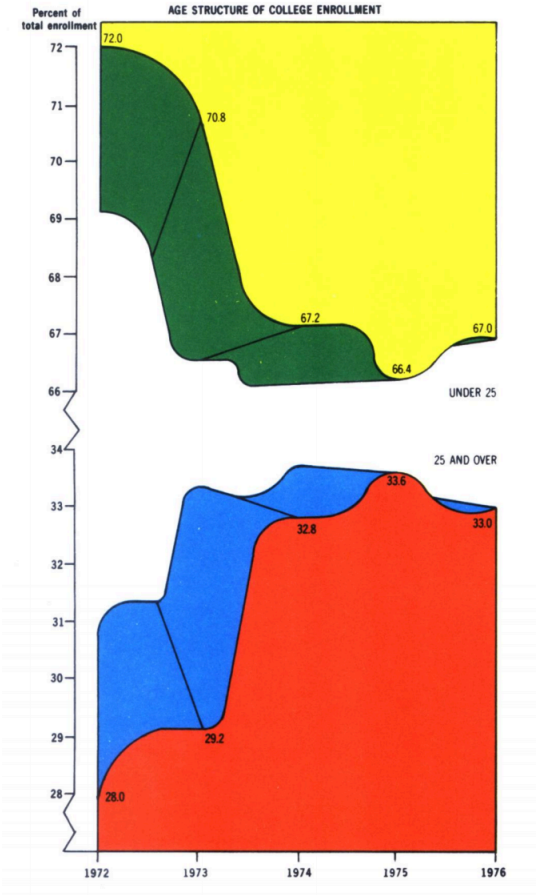
Data with no obvious signal
And 1 outlier

Summarization as a barchart
to hide that obvious fact

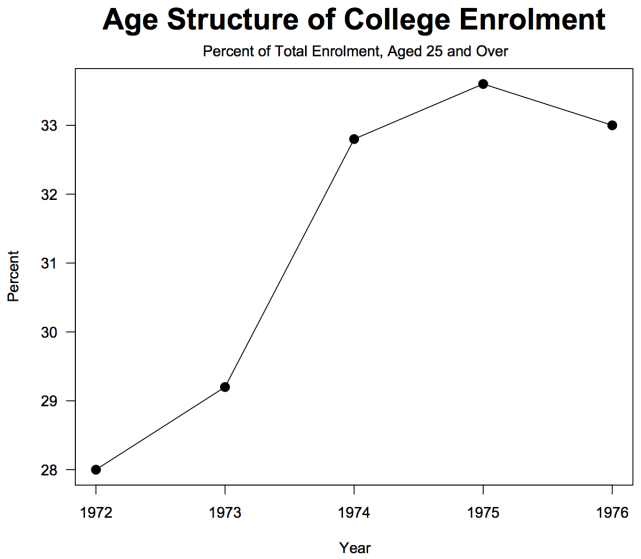
Complexity

- Graphs should be no more complex than the data which they portray
- Unnecessary complexity can be introduced by
  - irrelevant decoration
  - colour
  - 3d effects

Complexity



Age Structure of College Enrolment (1972-1976)



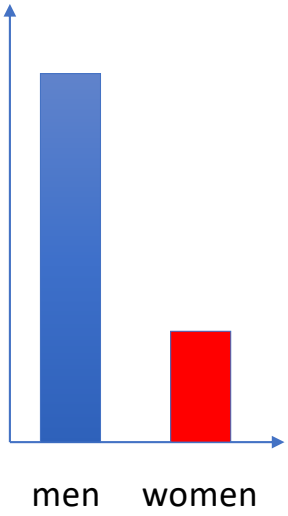
Both graphs present the same data

Distortions

- Graphs should not provide a distorted picture of the values they portray
- Distortion can be either deliberate or accidental (especially if one do not really understand the data)

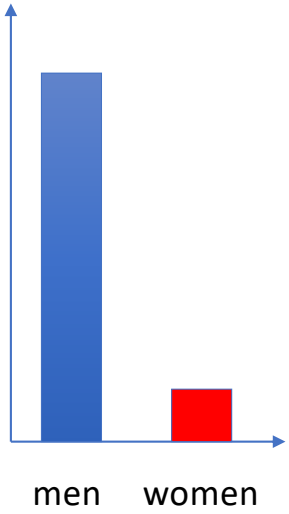
Distortions (uncovering hidden context)

Unequal gender representation in an office implies gender discrimination?



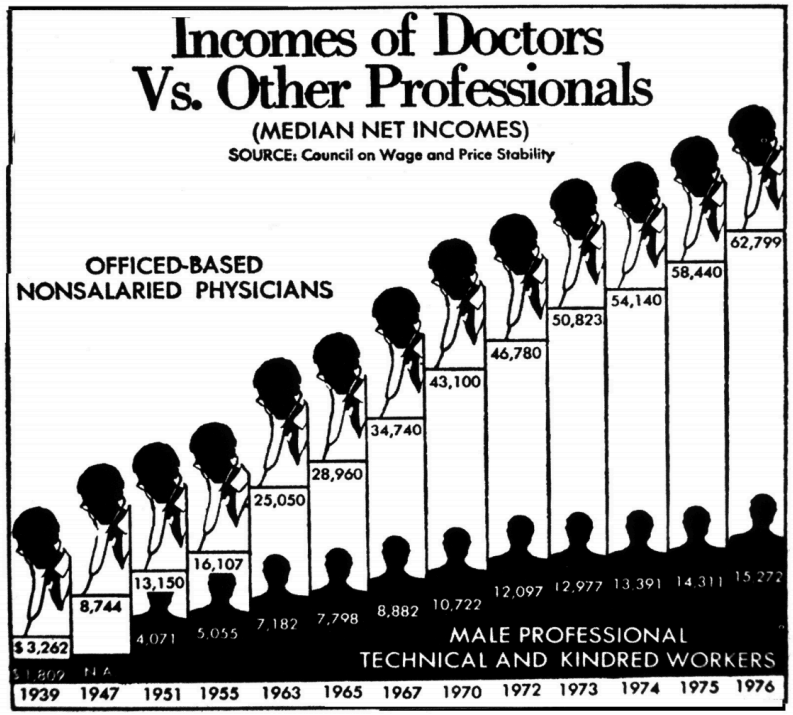
Seems like we need to do something!

Interviewee proportion (split by gender)

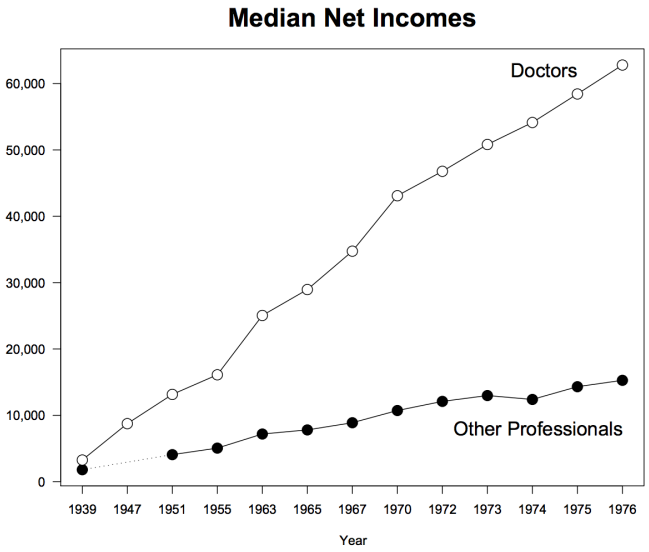


It seems that there is deliberate attempts to increase female representation despite the very low gender representation amongst interviewees

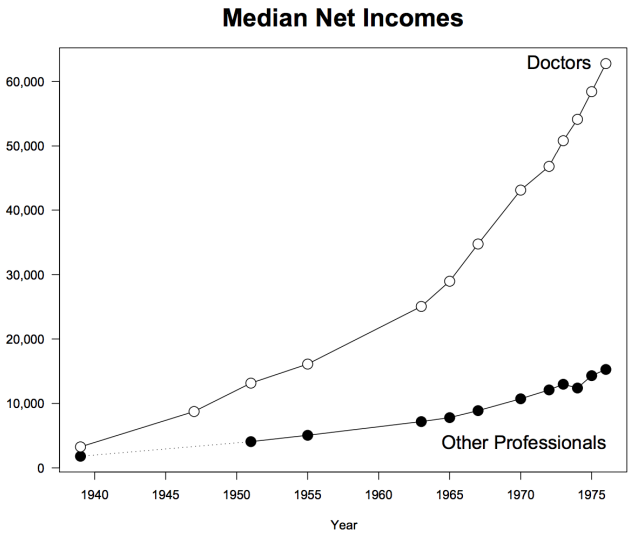
Distortions (inappropriate use of linear scaling)



Distortions (inappropriate use of linear scaling)



Seems pretty innocuous. But look carefully again at the x-axis intervals



Once the intervals are now pretty aligned. You can see a disturbing trend

Generic guides for good graphing

- Draw the graph with an aim to communicate
  - If the “story” is simple, keep it simple
  - Ensures that axes, legends, annotations are fully visible
- If the “story” is complex, make it look simple
  - The aim is to draw insight quickly and accurately. If the graph is as complex as the data, it is of limited use
- Avoid distorting the data
  - Don’t use aesthetic features unless it serves useful purpose
  - Understand the context of the date you are representing
  - Don’t “hide” or “lie” by using inconsistent intervals or other visual tricks

Some Generic Good Analytical Practices (GAPs)

- For small sample size (< 5), summary statistics are not meaningful -> Use scatterplots
- Check the actual distribution of individual data points (do not skip right to summary statistics)
- Use the median rather than the mean to identify the center of your data
- Always check for outliers, non-symmetry, hidden subpopulations, and handle them accordingly
- Never apply statistical tests before checking the data distribution

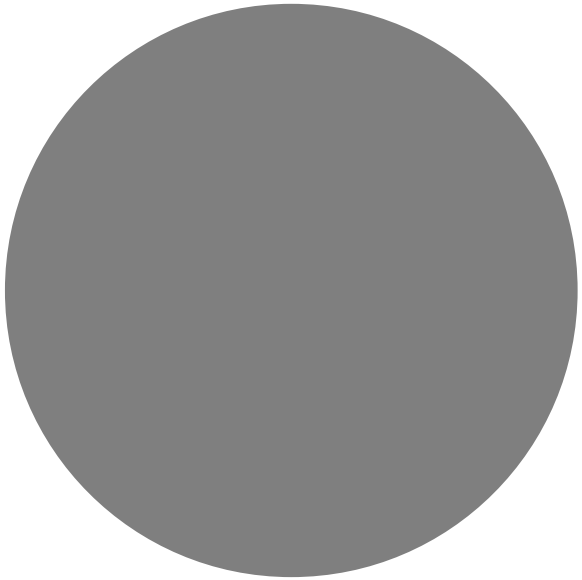
To learn how to draw univariate scatterplots in Excel go to
https://www.ctspedia.org/do/view/CTSpedia/TemplateTesting

Readings

- How NOT to Lie with Visualization (https://pdfs.semanticscholar.org/058e/2e38420b61d8d870590d971d4e7d1cd078c2.pdf)
- 14 Ways to Say Nothing with Scientific Visualization (http://crack.seismo.unr.edu/ftp/vis/14ways.pdf)

End of Segment

Let’s take a break



Graphs in R

Graphs in R

One of the main reasons data analysts turn to R is for its strong graphic capabilities.

Creating a Graph provides an overview of creating and saving graphs in R.

The remainder of the section describes how to create basic graph types. high density plots, and 3D plots).

The Advanced Graphs section describes how to customize and annotate graphs, and covers more statistically complex types of graphs.

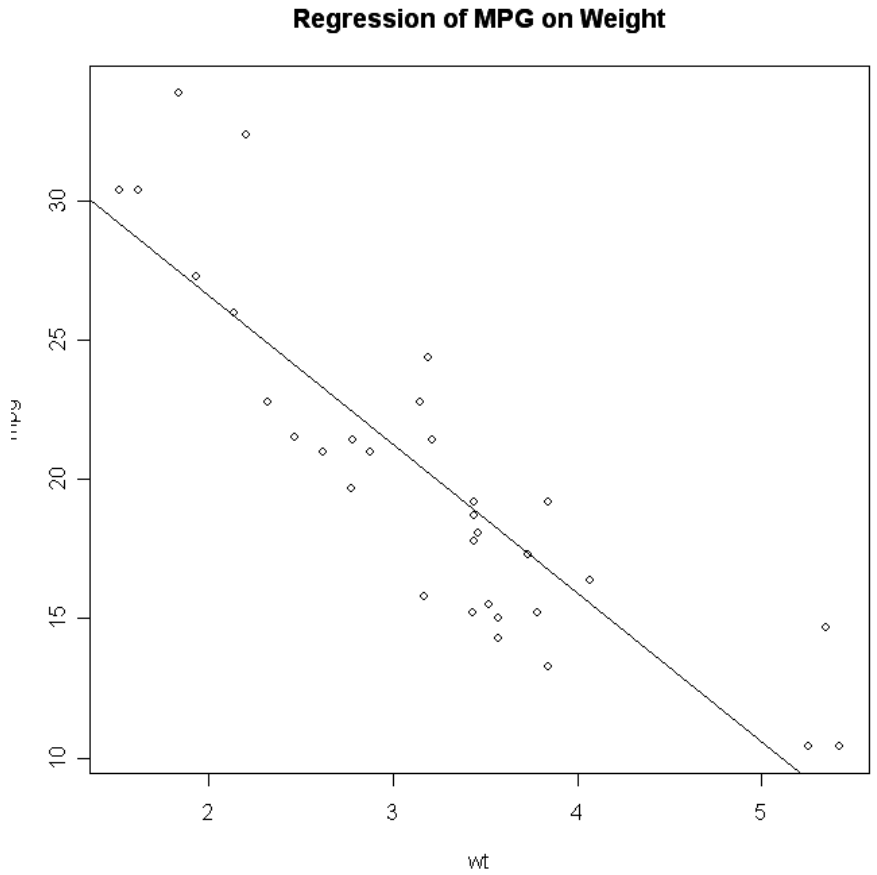
Creating a graph

In R, graphs are typically created interactively.

```
# Creating a Graph
attach(mtcars)
plot(wt, mpg)
abline(lm(mpg~wt))
title("Regression of MPG on Weight")
```

Creating a graph

- The **plot( )** function opens a graph window and plots weight vs. miles per gallon.
- The next line of code adds a regression line to this graph. The final line adds a title.



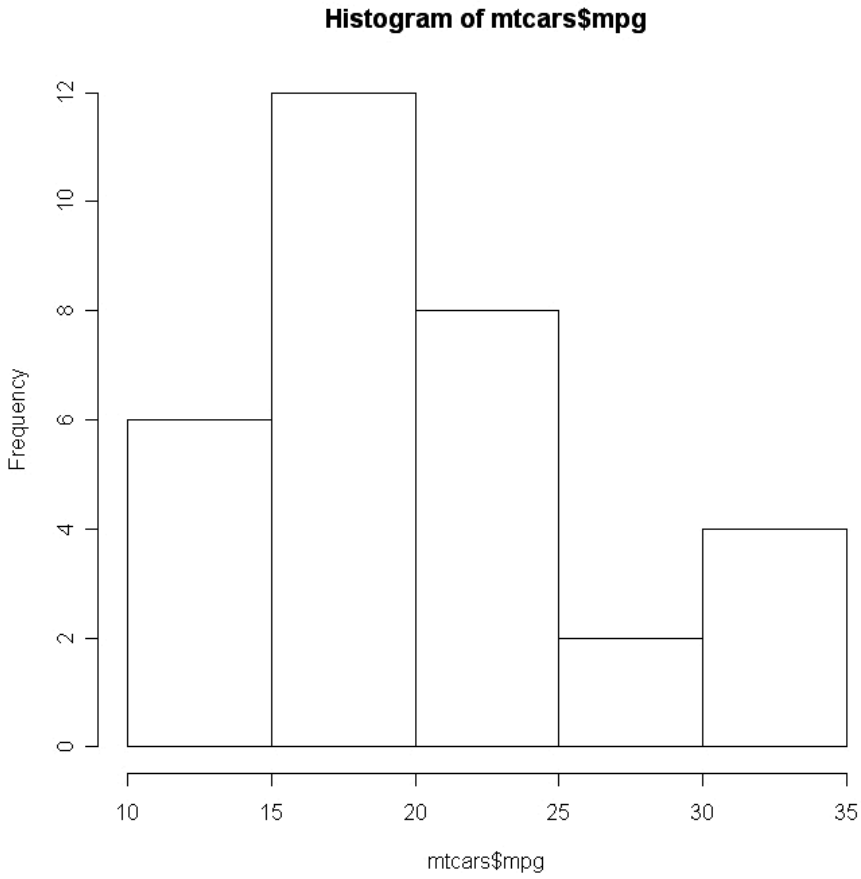
Saving graphs

- You can save the graph in a variety of formats from the menu
  File -> Save As.
- You can also save the graph via code using one of the following functions.
- # example - output graph to jpeg file
  - jpeg("c:/<working directory>/myplot.jpg")
  - plot(x)
  - dev.off()

Function	Output to
pdf("mygraph.pdf")	pdf file
win.metafile("mygraph.wmf")	windows metafile
png("mygraph.png")	png file
jpeg("mygraph.jpg")	jpeg file
bmp("mygraph.bmp")	bmp file
postscript("mygraph.ps")	postscript file

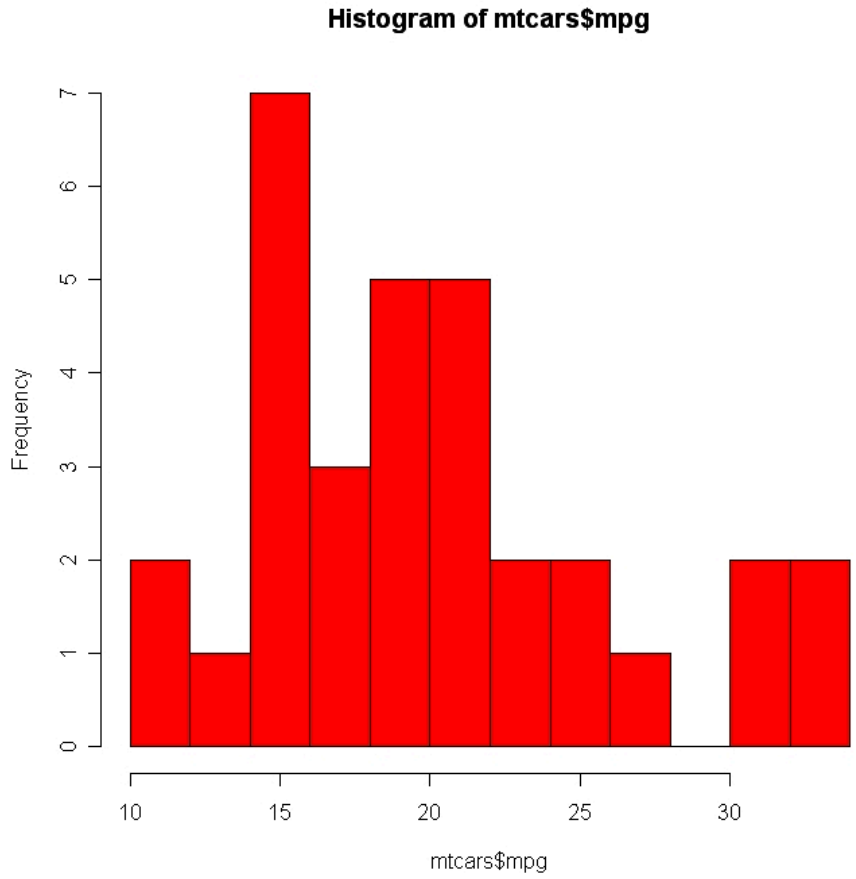
Histograms

- You can create histograms with the function **hist(*x*)** where *x* is a numeric vector of values to be plotted. The option **freq=FALSE** plots probability densities instead of frequencies. The option **breaks=** controls the number of bins.
- # Simple Histogram
  hist(mtcars$mpg)



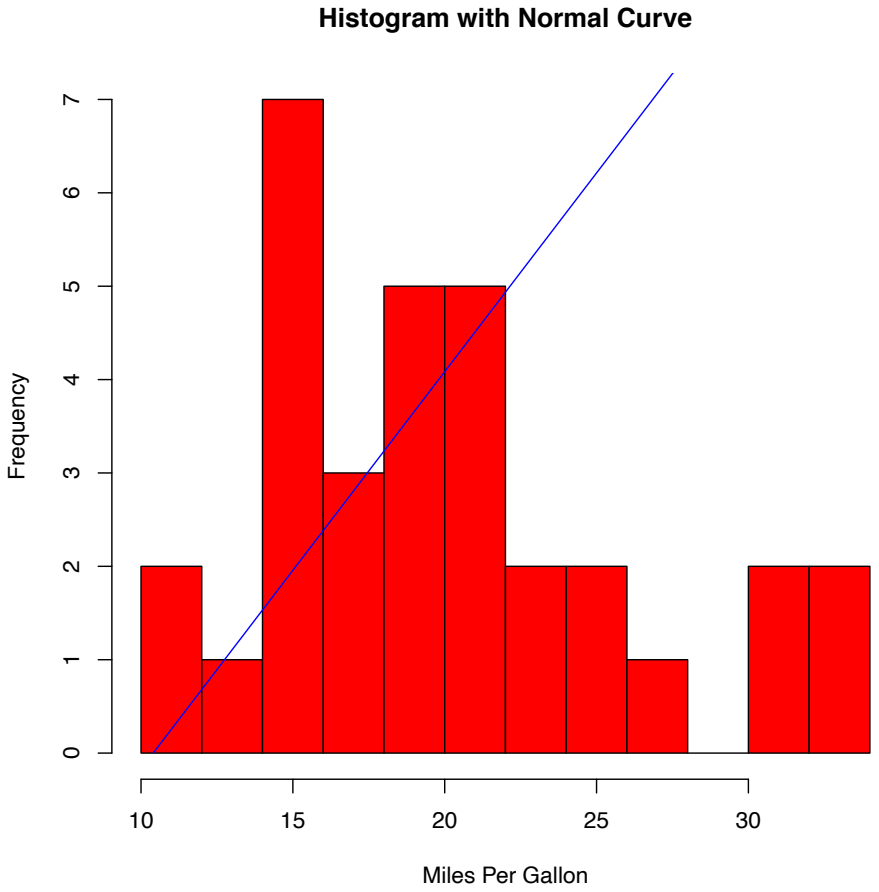
Histograms

- # Colored Histogram with Different Number of Bins hist(mtcars$mpg, breaks=12, col="red")



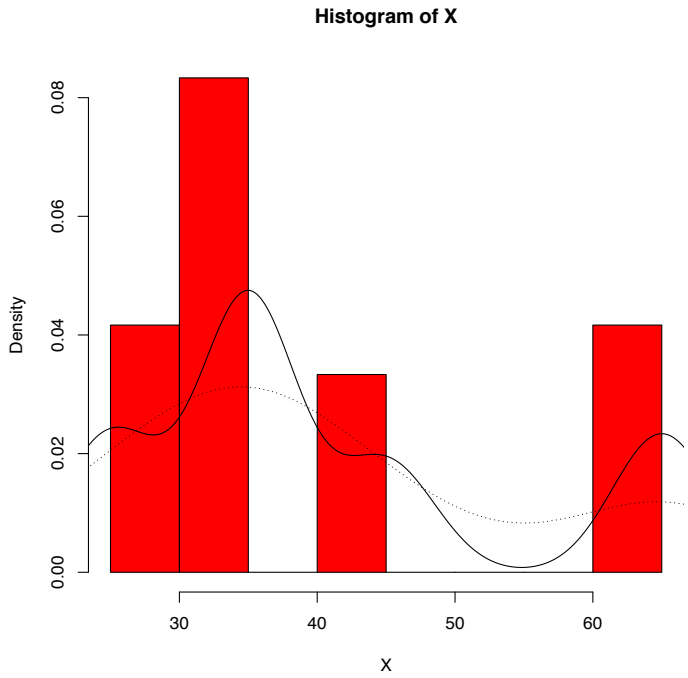
Histograms

- # Add a straight line
- x <- mtcars$mpg
  h<-hist(x, breaks=10, col="red",
  xlab="Miles Per Gallon",
  main="Histogram with Normal Curve")
  xfit<-seq(min(x),max(x),length=40)
  lines(xfit, seq(0, 10, length=40),
  col="blue")



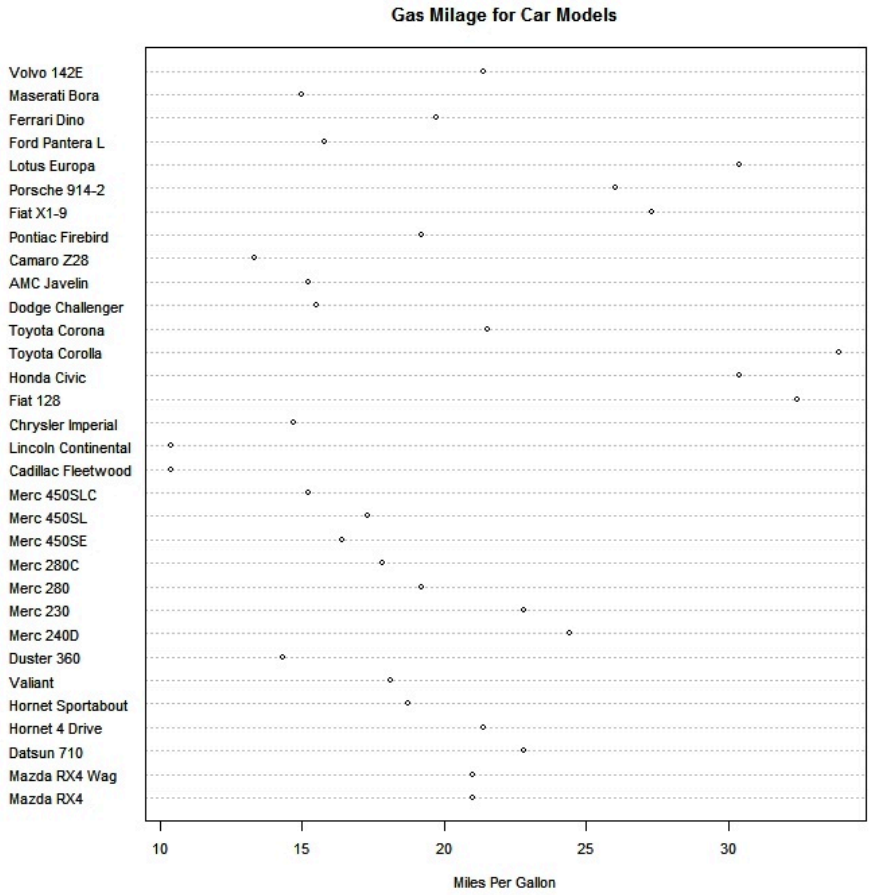
Histograms

- X <- c(rep(65, times=5), rep(25, times=5), rep(35, times=10), rep(45, times=4))
- hist(X, prob=TRUE, col="red") # prob=TRUE for probabilities not counts
- lines(density(X)) # add a density estimate with defaults
- lines(density(X, adjust=2), lty="dotted") # add another "smoother" density



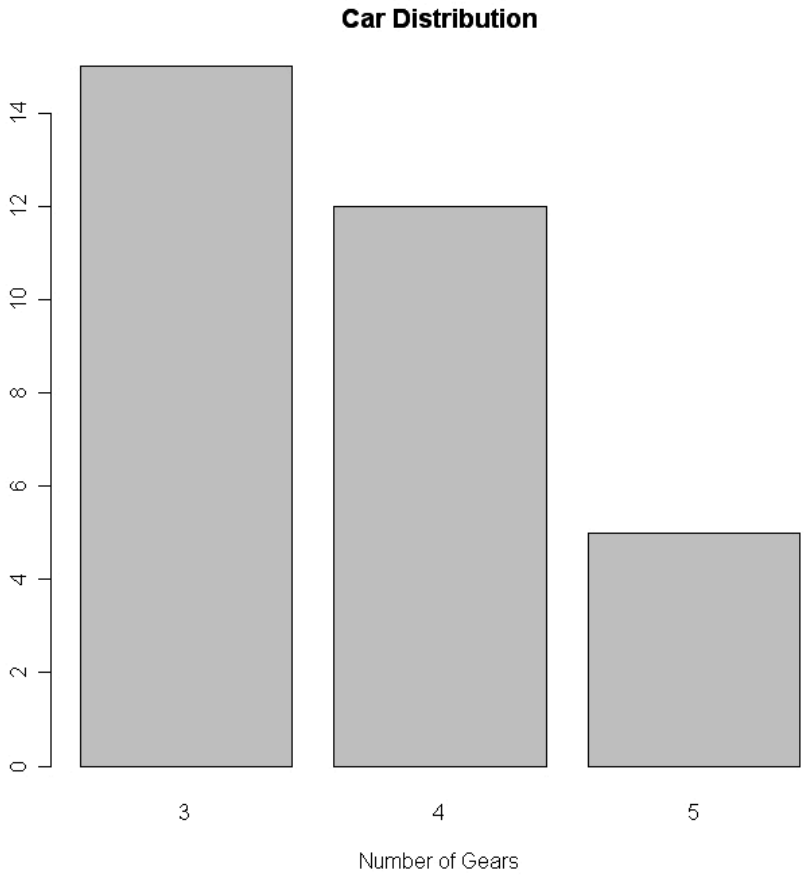
Dotplots

- Create dotplots with the **dotchart(*x*, labels=)** function, where *x* is a numeric vector and **labels** is a vector of labels for each point.
- You can add a **groups=** option to designate a factor specifying how the elements of *x* are grouped. If so, the option **gcolor=** controls the color of the groups label. **cex** controls the size of the labels.
- # Simple Dotplot
  dotchart(mtcars$mpg,labels=row.names(mtcars),cex=.7, main="Gas Milage for Car Models", xlab="Miles Per Gallon")



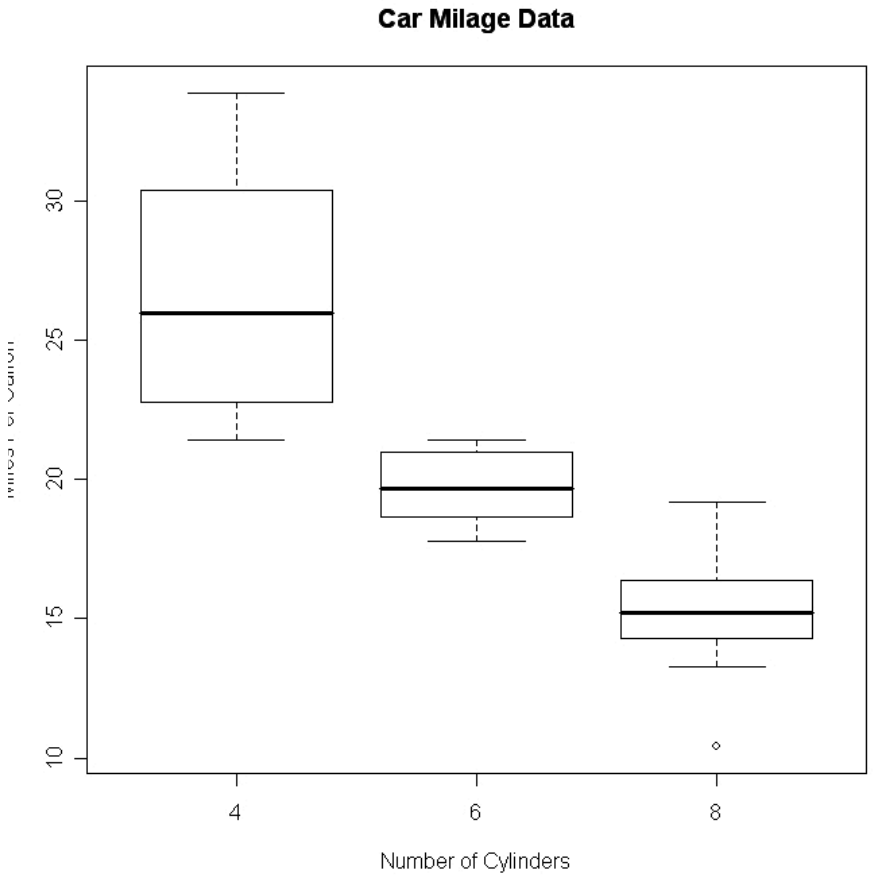
Barplots

- Barplots are created with the **barplot(***height***)** function, where *height* is a vector or matrix.
- If **height is a vector**, the values determine the heights of the bars in the plot.
- # Simple Bar Plot
  counts <- table(mtcars$gear)
  barplot(counts, main="Car Distribution",
    xlab="Number of Gears")



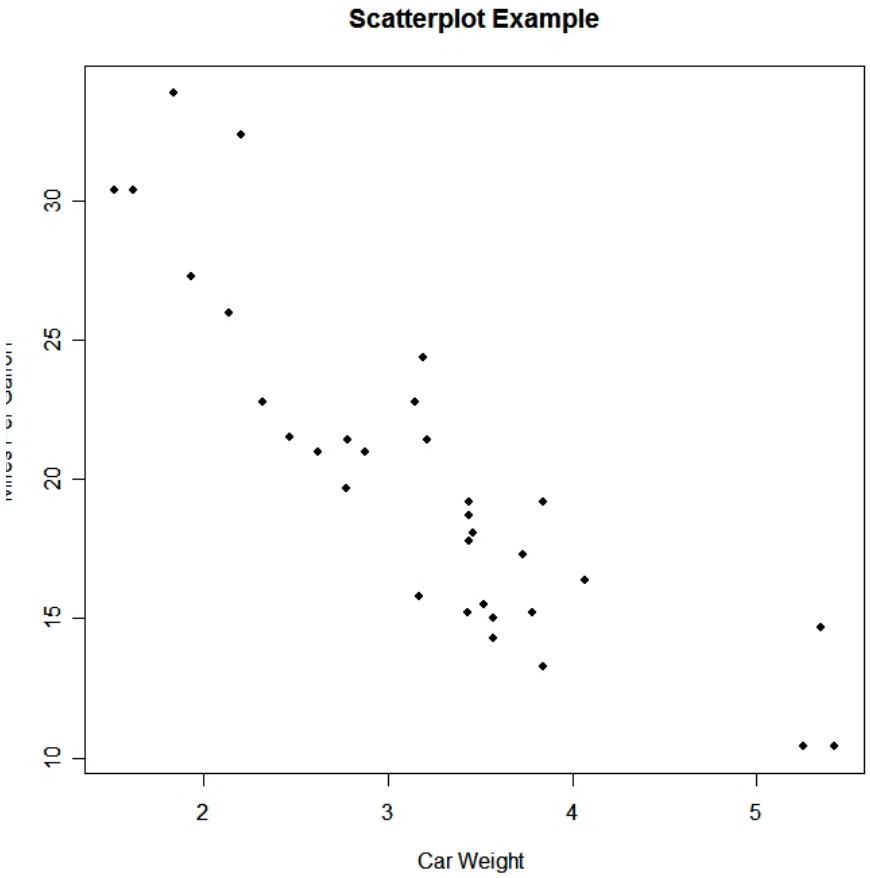
Boxplots

- Boxplots can be created for individual variables or for variables by group.
- The format is **boxplot(*x*, data=)**, where *x* is a formula and **data=** denotes the data frame providing the data.
- An example of a **formula** is y~group where a separate boxplot for numeric variable y is generated for each value of group.
- # Boxplot of MPG by Car Cylinders
  boxplot(mpg~cyl,data=mtcars, main="Car Milage Data",
  xlab="Number of Cylinders", ylab="Miles Per Gallon")



Scatterplots

- The basic function is **plot(*x*, *y*)**, where *x* and *y* are numeric vectors denoting the (x,y) points to plot.
- # Simple Scatterplot
  attach(mtcars)
  plot(wt, mpg, main="Scatterplot Example",
  xlab="Car Weight ", ylab="Miles Per Gallon ", pch=19)



End of Segment

Let’s take a break