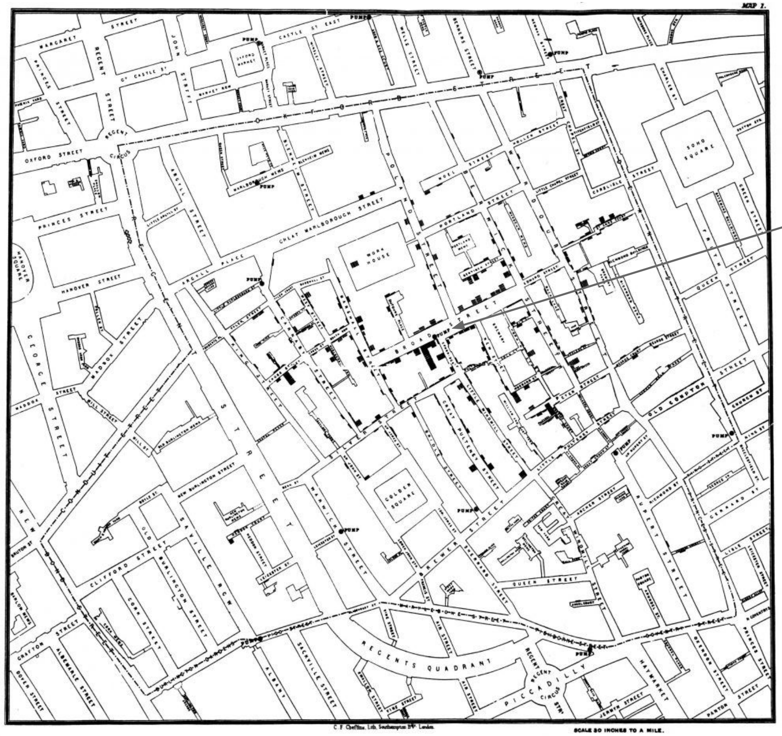
Data Visualisation

Dr Andrew J. Stewart

E: drandrewjstewart@gmail.com
T: @ajstewart_lang
G: ajstewartlang

Some Classic Visualisations



John Snow’s map of the cholera outbreak in 1854 in London showed that the outbreak was centred around a contaminated water pump in Broad Street.

Some Classic Visualisations

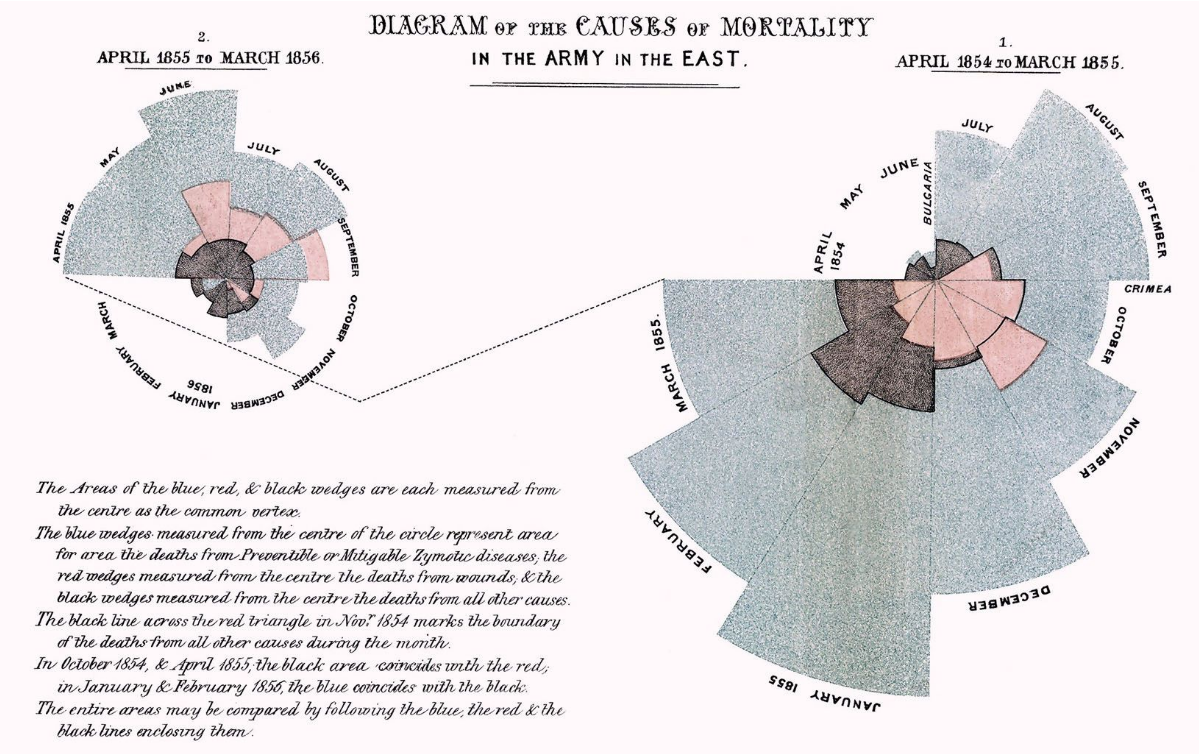


DIAGRAM OF THE CAUSES OF MORTALITY IN THE ARMY IN THE EAST.

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.

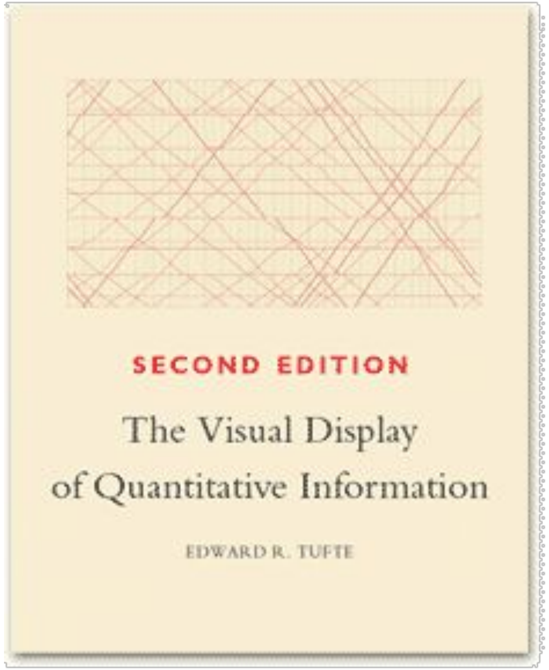
The black line across the red triangle in Nov.r 1854 marks the boundary of the deaths from all other causes during the month.

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.

This Rose Chart or Coxcomb by Florence Nightingale (yes, that Florence Nightingale!) was used to capture the causes of death of “The Army in the East”.

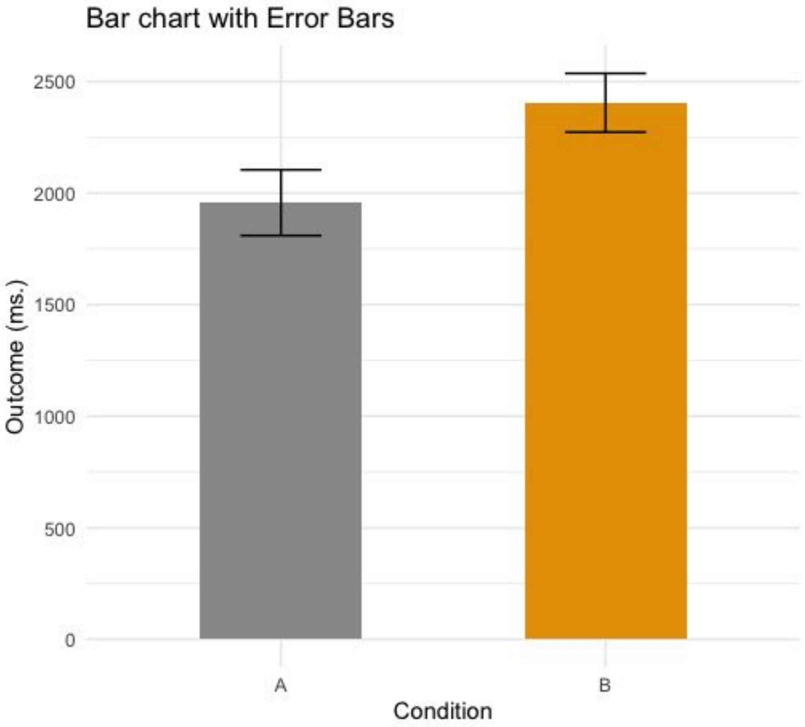
The Classic Book



“The Visual Display of Quantitative Information” by Edward Tufte is the classic book on data visualisation - hugely influential and contains lots of examples of the best (and worst kinds) of data visualisations.

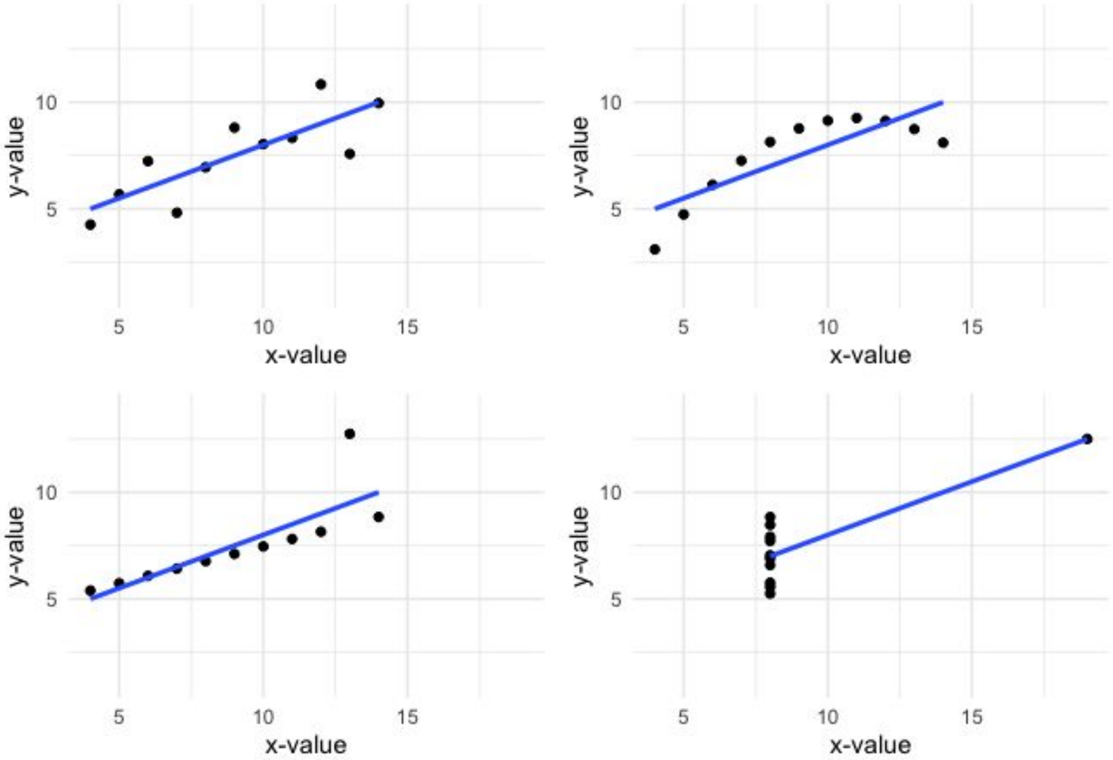
And speaking of one of the worst kinds...

Visualisations that Can Mislead



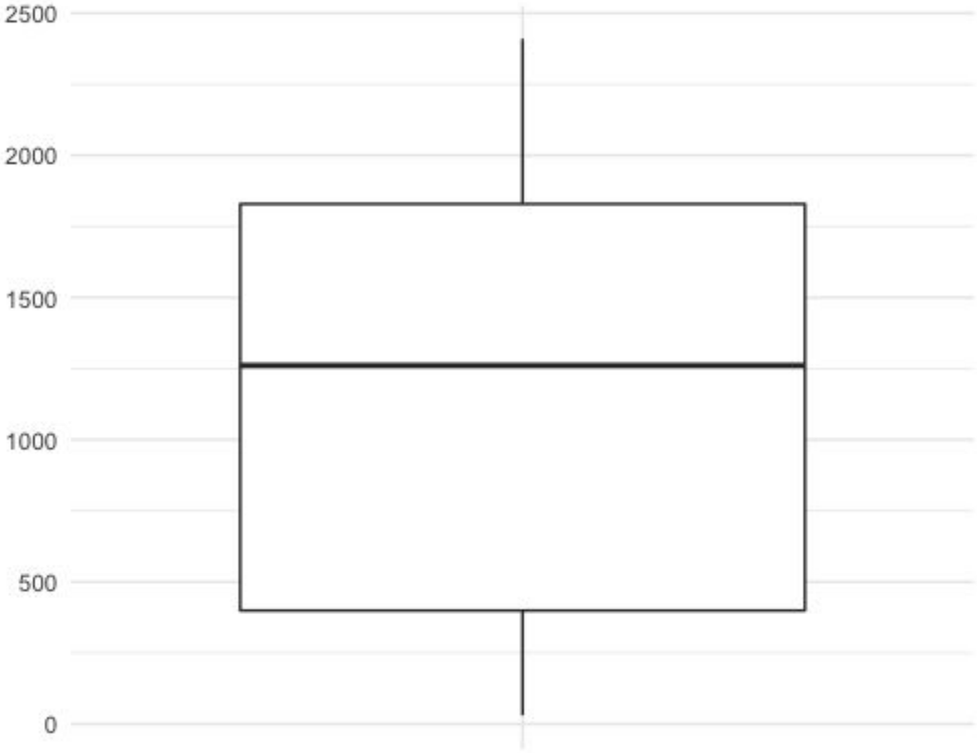
Bar graphs tend to be quite limited in terms of what they communicate. Here they communicate the means for two conditions and information about variability around the means. But they don’t tell us anything about the **distribution** of the data.

Anscombe's Quartet



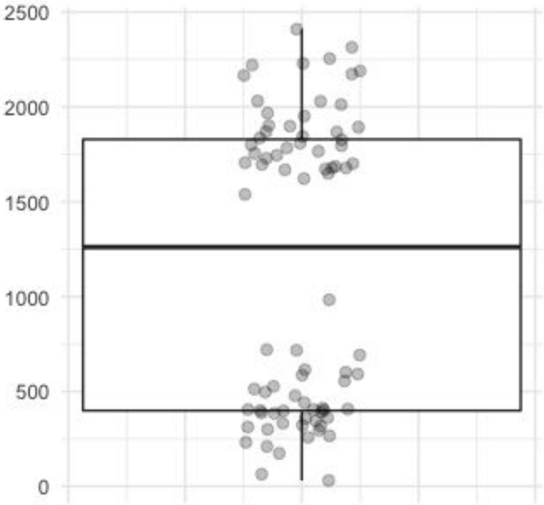
The data underlying each of the four plots on the left are all different, but each has the same mean and standard deviation for their x-values, the same mean and standard deviation for their y-values, the same correlation coefficient, and the same regression line. If we reported only these things, we'd think the four datasets were identical (whereas they are clearly not!)

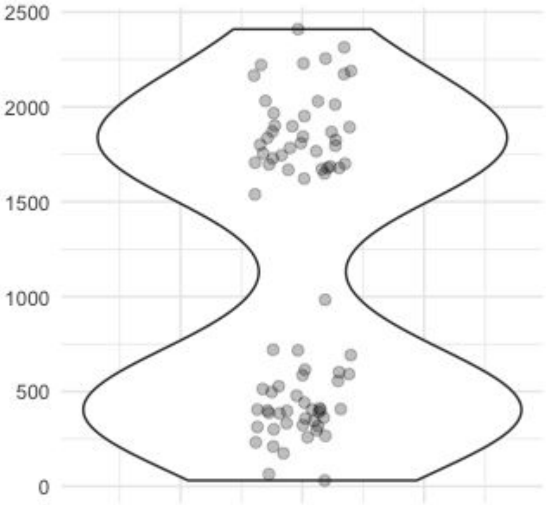
Plots Based on Aggregated Data Can Mislead



You might make one set of inferences based on this boxplot - maybe a median around 1,250 with the 25th and 75th percentiles being ~480 to ~1,980

But look more closely at the actual data...





The data are clearly bimodal.

Distribution shape matters and we need to capture that in our data visualisations.

The ggplot2 Package

The ggplot2 package is part of the Tidyverse and the function `ggplot()` forms the basis for data visualisations using Tidyverse packages and verbs.

There are three basic components when using the `ggplot()` function to build a visualisation:

- Data
    - The raw data that you want to plot
- Geometries (e.g., `geom_point()` and `geom_jitter()`)
    - The geometric shapes that will represent the data.
- Aethetics (`aes()`)
    - Aesthetics of the geometric and statistical objects, such as color, size, shape and position.

Onto the R script below…