Lesson 4: Fundemantal of Visual Analytics

Dr. Kam Tin Seong
Assoc. Professor of Information Systems

School of Computing and Information Systems,
Singapore Management University

2020-01-20 (updated: 2021-05-18)

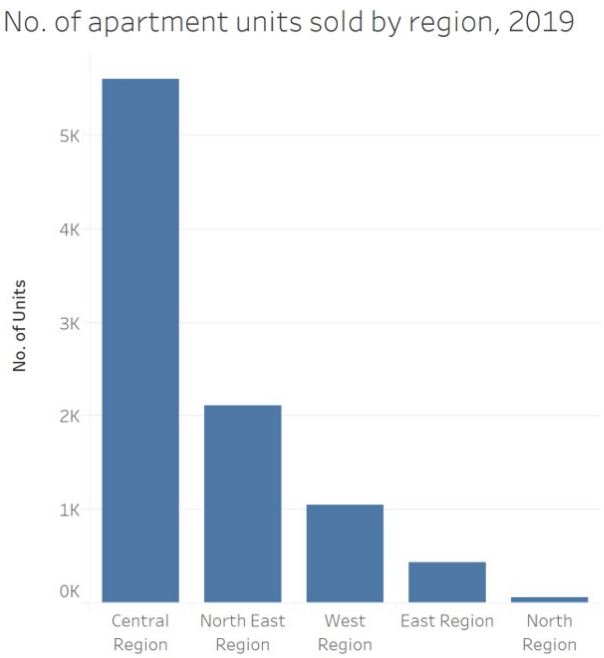
What will you learn from this lesson?

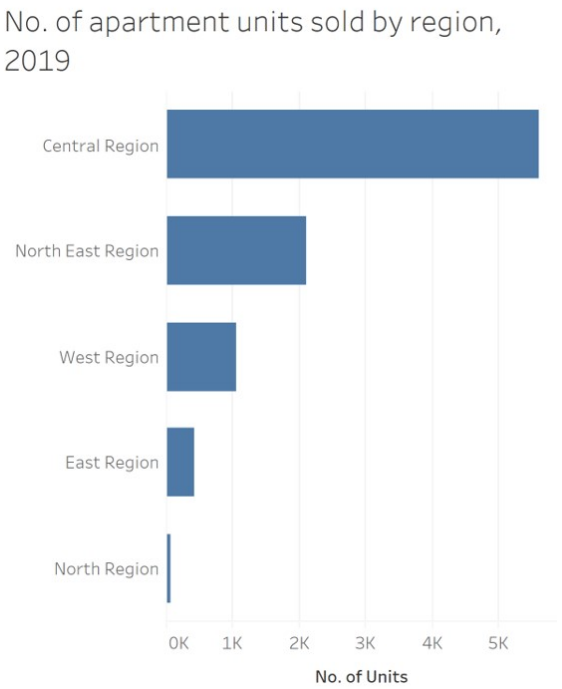
- Visualising count
- Visualising proportion
  - Part-whole and ranking analysis
- Visualising distribution
- Visualising deviation
- Visualising relationships
  - between two continuous variables
  - between two categorical variables
- Visualising relationship between sub-categories
- Visualising uncertainty

Visualising Count

Bar Chart

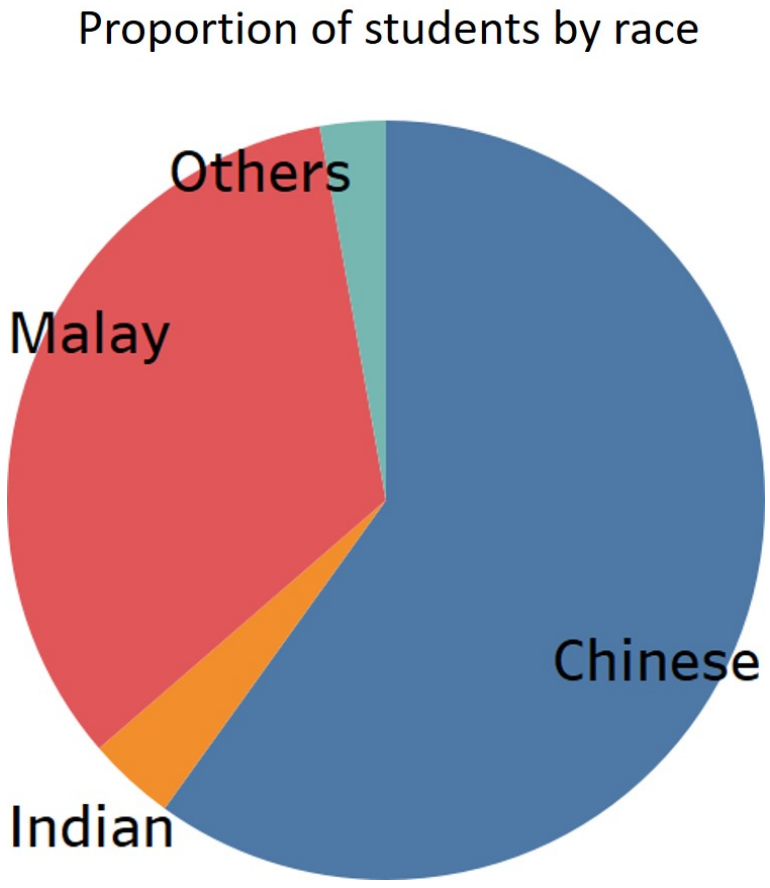
- A bar chart is used for plotting categorical data.
- It can be mapped horizontally or vertically.
- When displaying data using bar chart, it is a good practice to sort the count or frequency ascendingly or descendingly.





Visualising Proportion

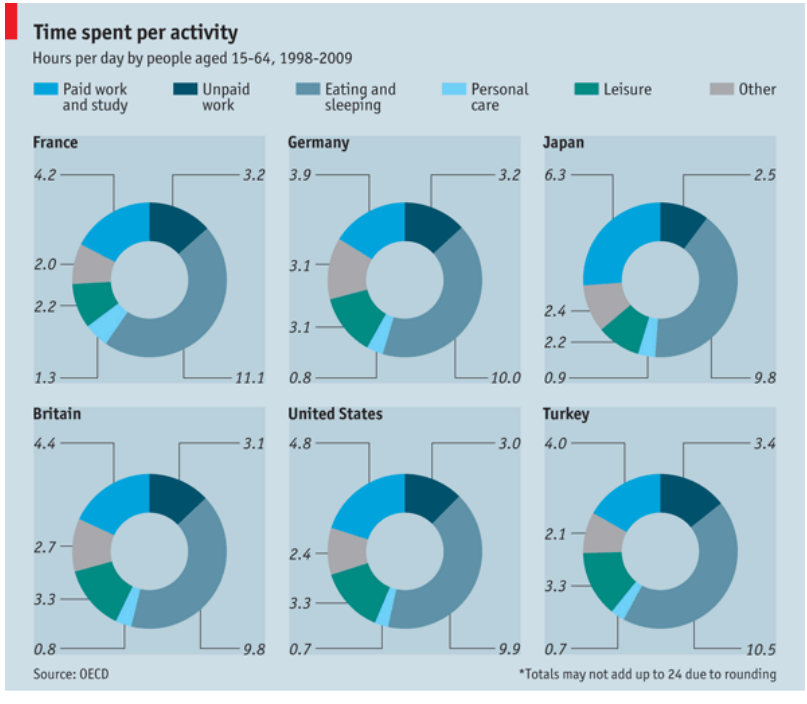
A case for pie chart



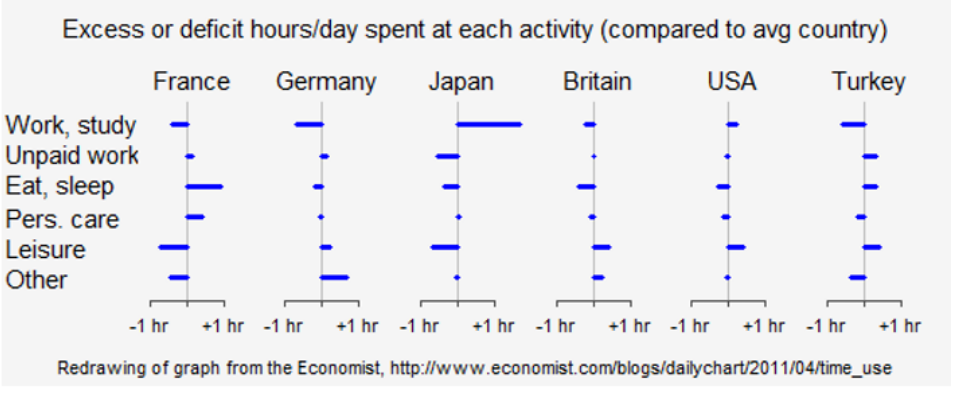
Visualising Proportion

A case against pie chart

- Avoid pie chart if the sub-groups are very similar because our eyes are not good in reading areas



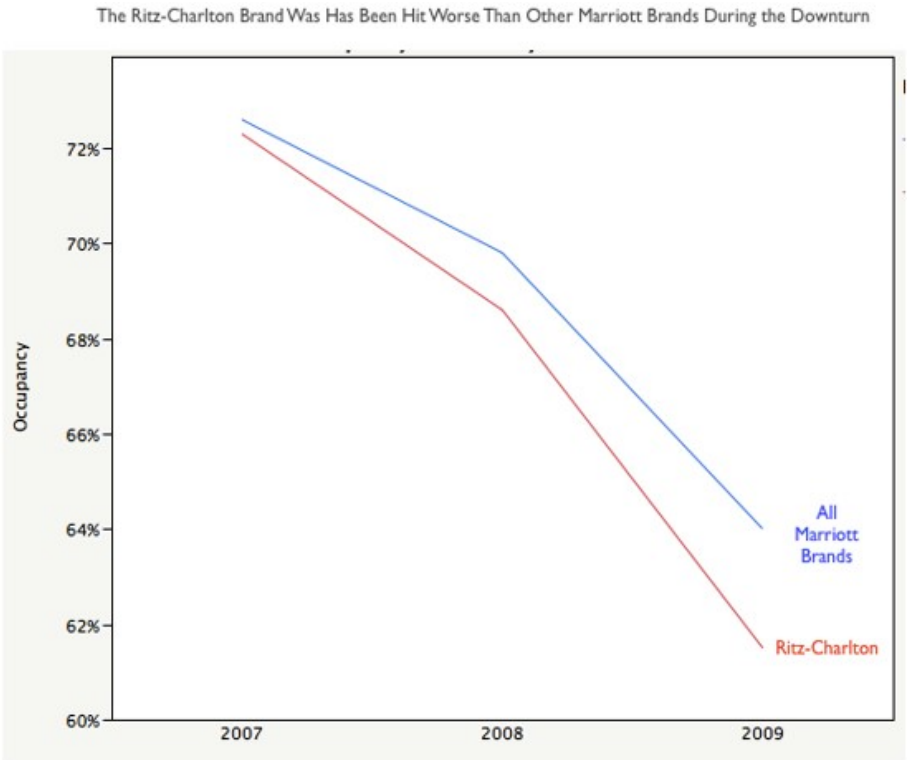
Source: Time use: A day in the life,
Apr 19th 2011, 15:00 by The Economist online



Visualising Proportion

A case against pie chart

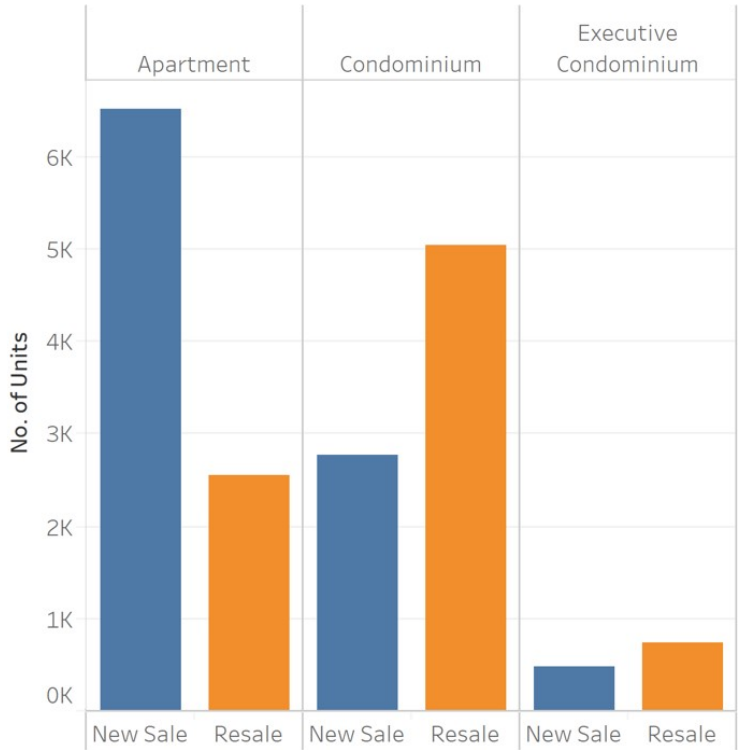
- Avoid pie chart if you are comparing changes over time



Visualising Proportion

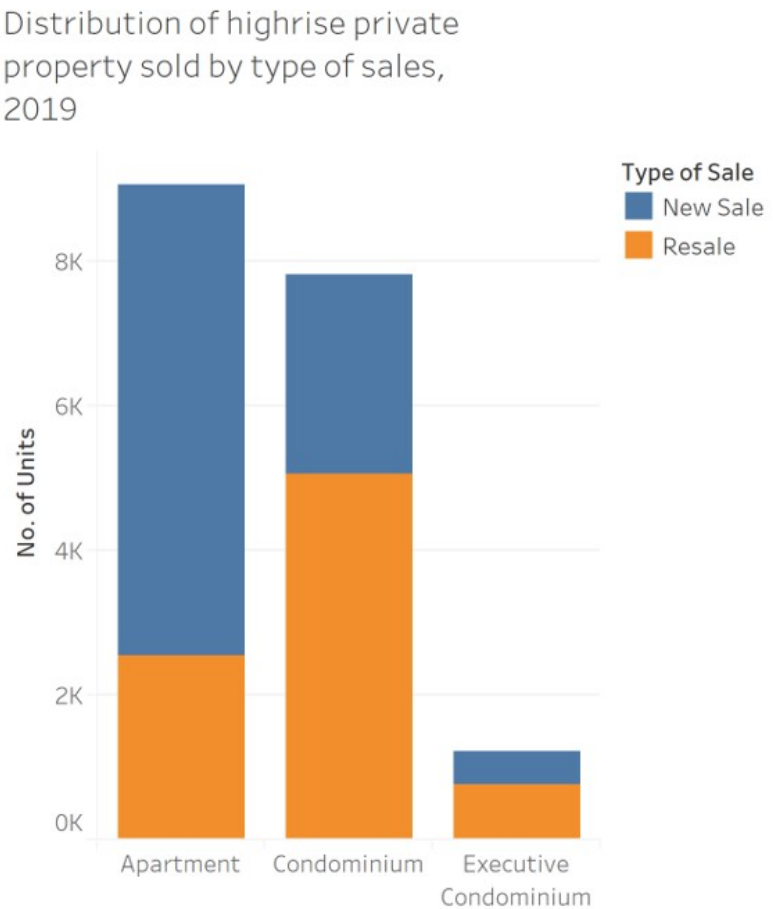
Side-by-side bar chart

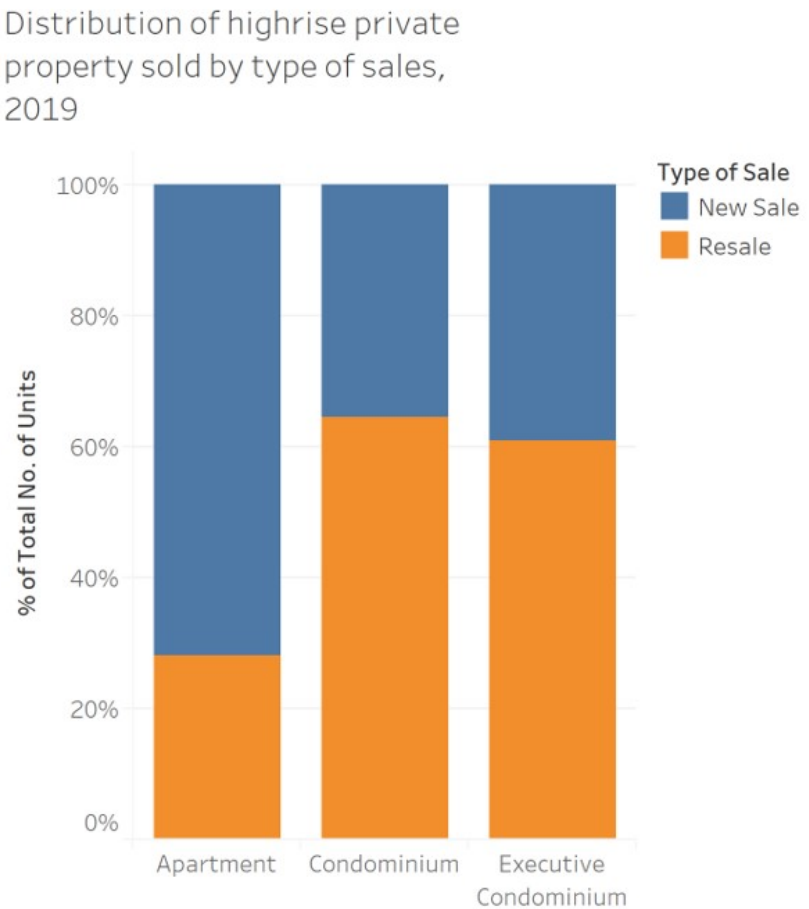
Distribution of highrise private property sold by type of sales, 2019



Visualising Proportion

Stacked bar chart



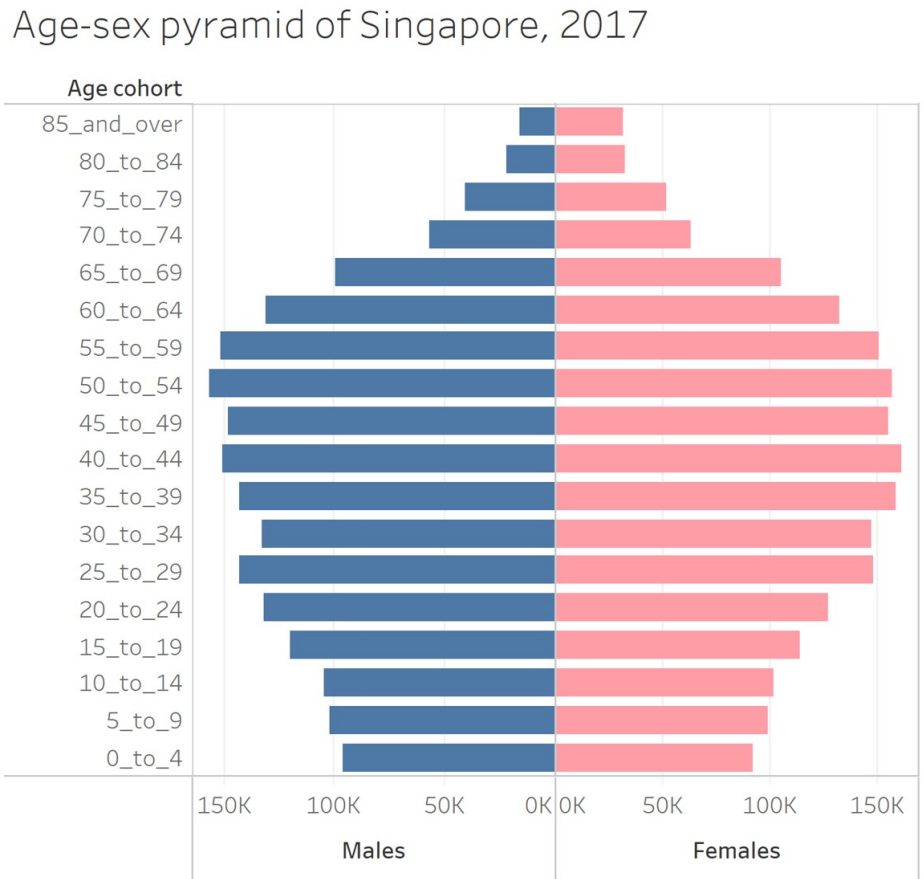


Comparing Proportion

Age-sex pyramid

- An age-sex pyramid, also popularly know as population pyramid, breaks down a country's or location's population into male and female genders and age cohorts.

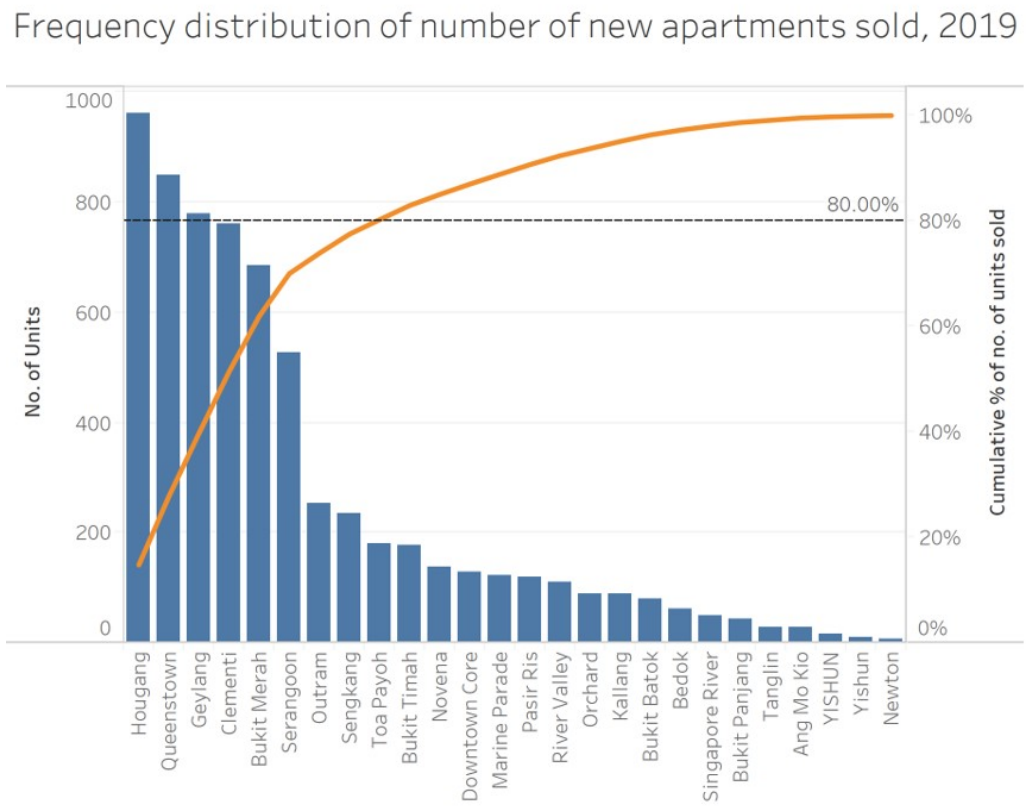
- Usually, you'll find the left side of the pyramid graphing the male population and the right side of the pyramid displaying the female population.



Part-to-Whole and Ranking Analysis

Pareto Chart

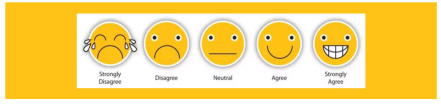
- A Pareto chart is a special type of bar chart where the values being plotted are arranged in descending order.
- Pareto chart was developed to illustrate the 80-20 Rule - that 80 percent of the problems stem from 20 percent of the various causes.
- In Pareto chart there are two y-axises. The primary y-axis is used to display the frequency counts of the sub-types and the secondary axis is used to display the cumulative frequency of the subtype.
- The frequency count usually is represented as bar chart and the cumulative frequency is represented as line chart.

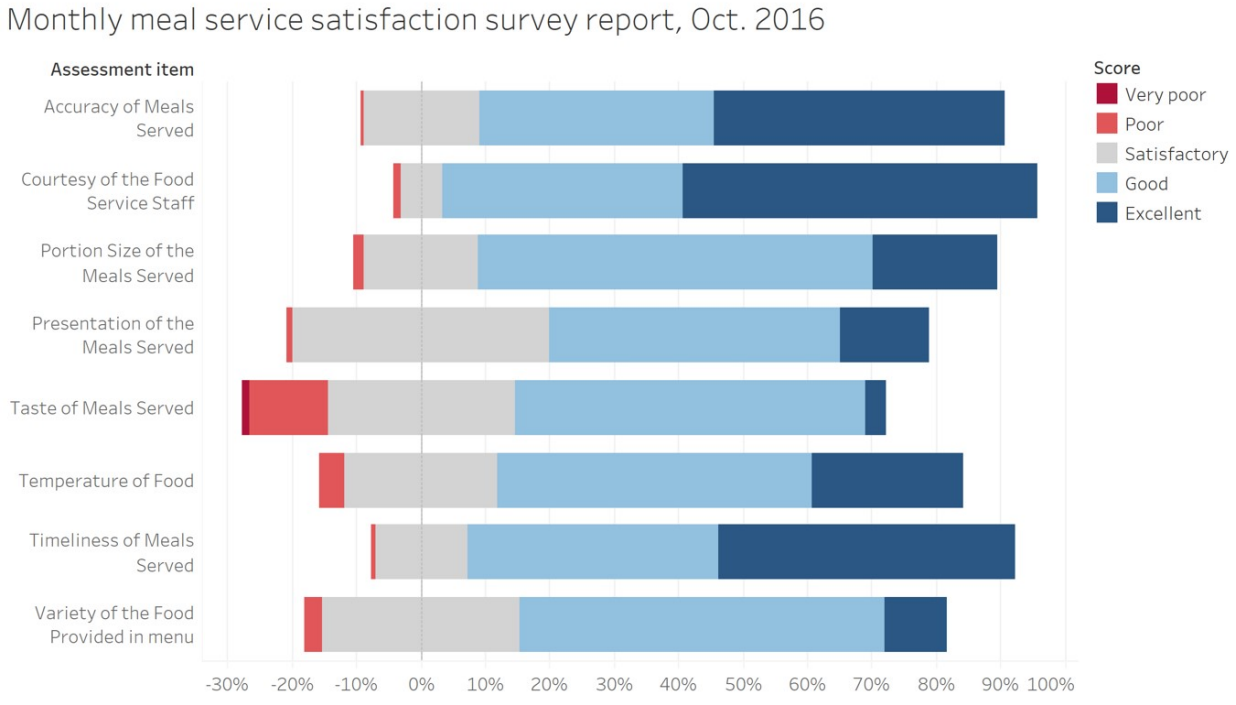


Visualising Likert Scale Data

Diverging stacked bar chart

- What is likert scale?





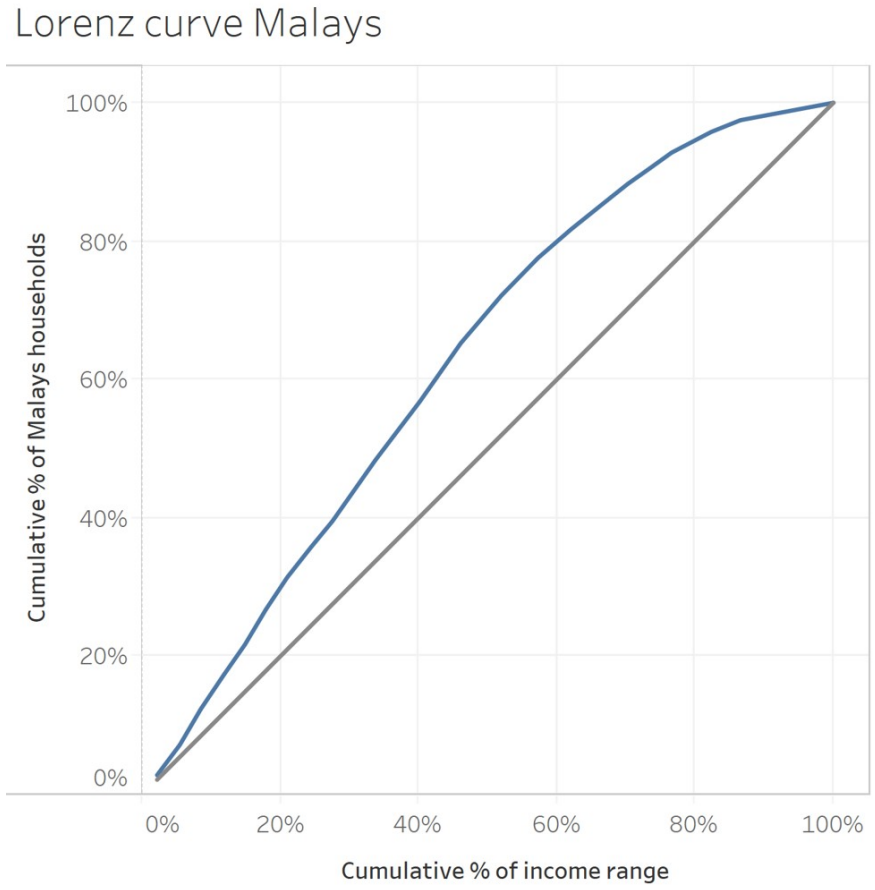
Reference:

- https://en.wikipedia.org/wiki/Likert_scale
- Heiberger RM, Robbins NB. Design of diverging stacked bar charts for Likert scales and other applications. Journal of Statistical Software. 2014;57(5): 1-32.

What about line graph?

Lorenz curve

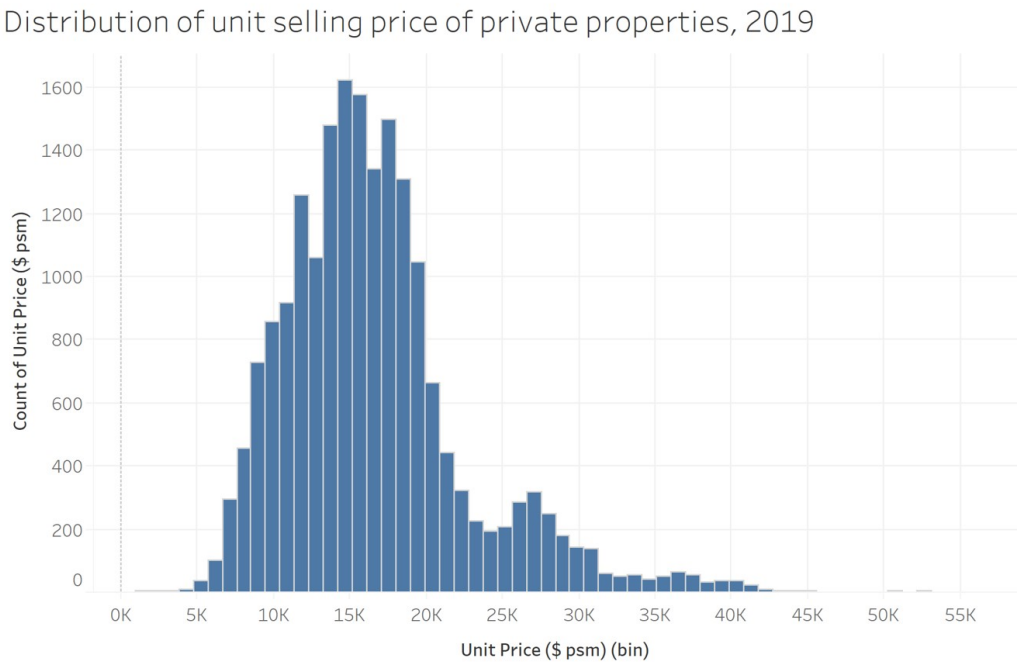
- A graphical representation of the distribution of income or of wealth.



Visualising Distribution

Histogram

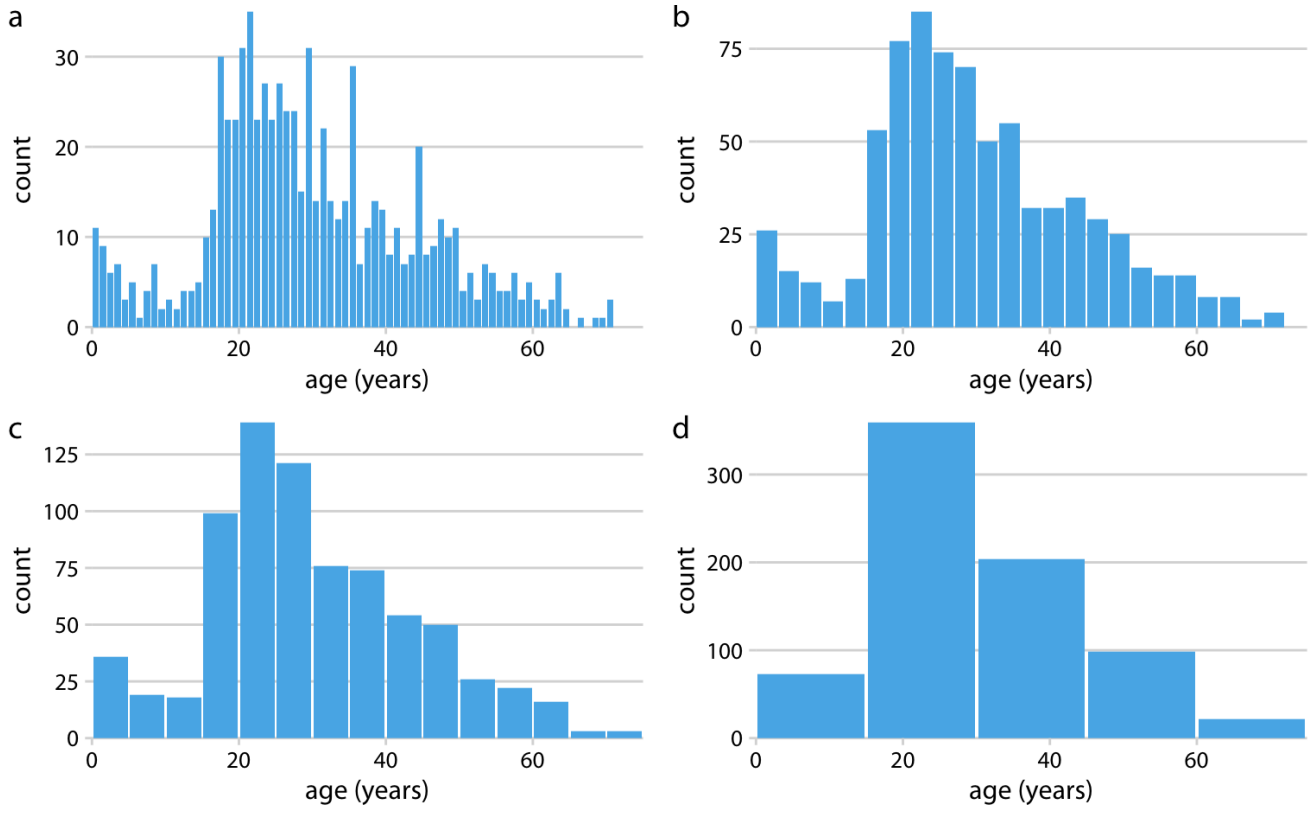
- A histogram is a graphical display of tabular frequencies, shown as adjacent rectangles.
- Each rectangle is erected over an interval, with an area equal to the frequency of the interval.
- The height of a rectangle is also equal to the frequency density of the interval, i.e. the frequency divided by the width of the interval.
- The total area of the histogram is equal to the number of data.



Visualising Distribution

Histogram

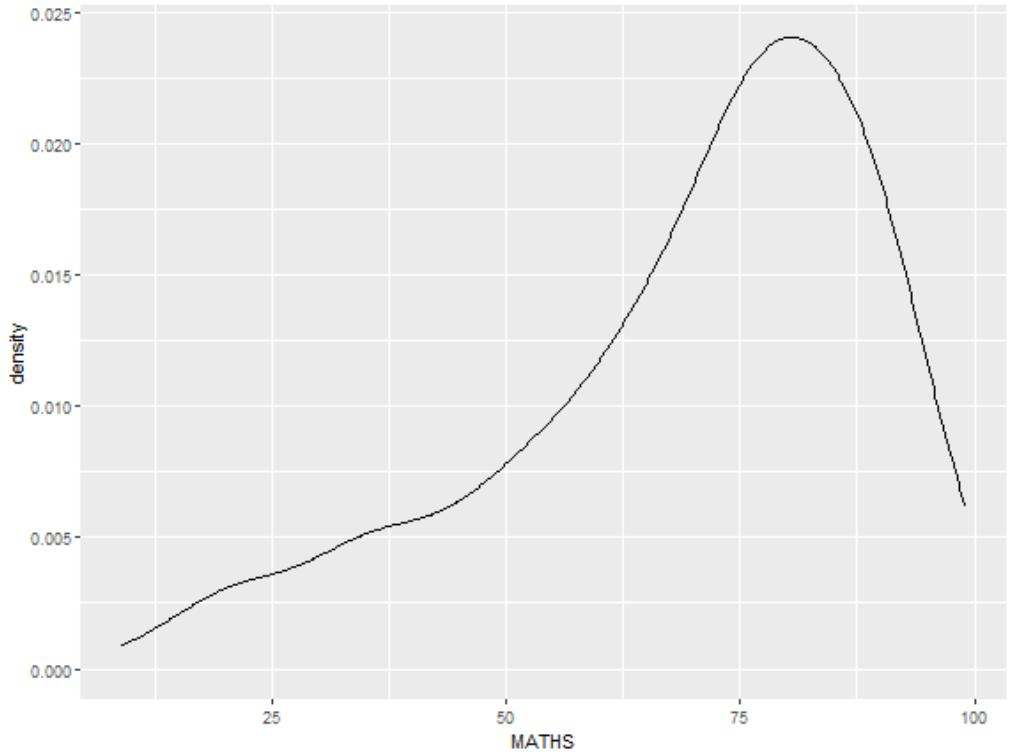
- It is important to note that the shape of a histogram can be affected by the number of bins or/and classification methods used.



Visualising Distribution

Density plot

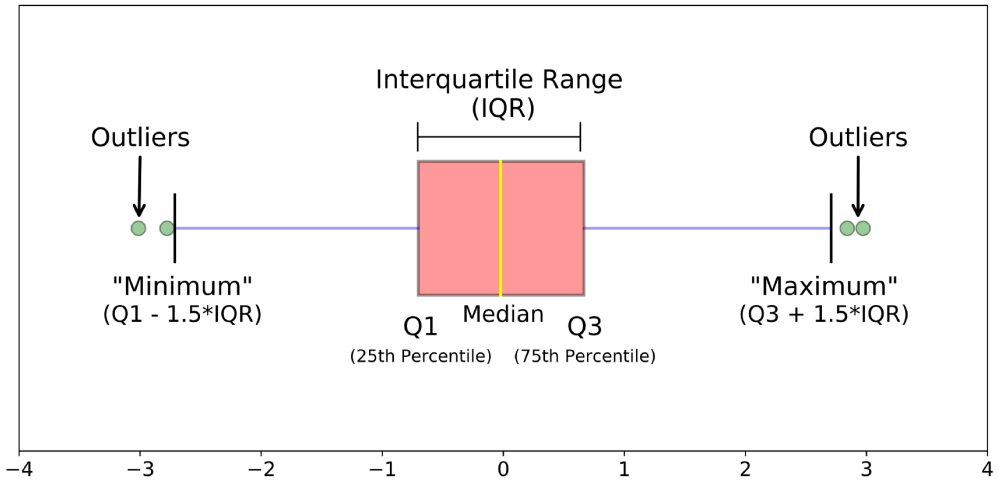
- To visualise the underlying probability distribution of the data by drawing an appropriate continuous curve.



Deviation Analysis

Boxplot

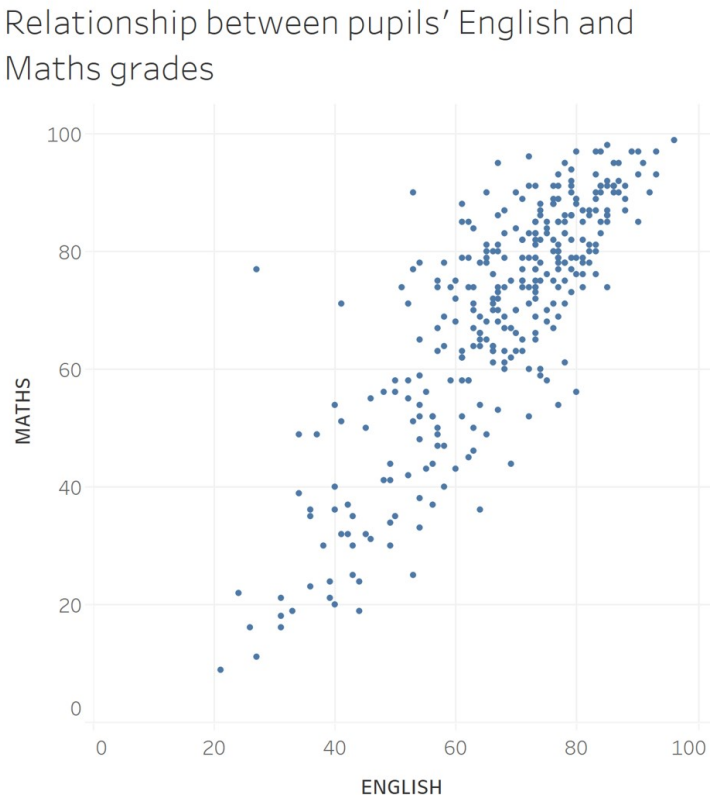
- A convenient way of graphically depicting groups of numerical data through their five-number summaries (the smallest observation, lower quartile (Q1), median (Q2), upper quartile (Q3), and largest observation).
- A box plot may also indicate which observations, if any, might be considered outliers.



Visualising Relationship Between Two Continuous Variables

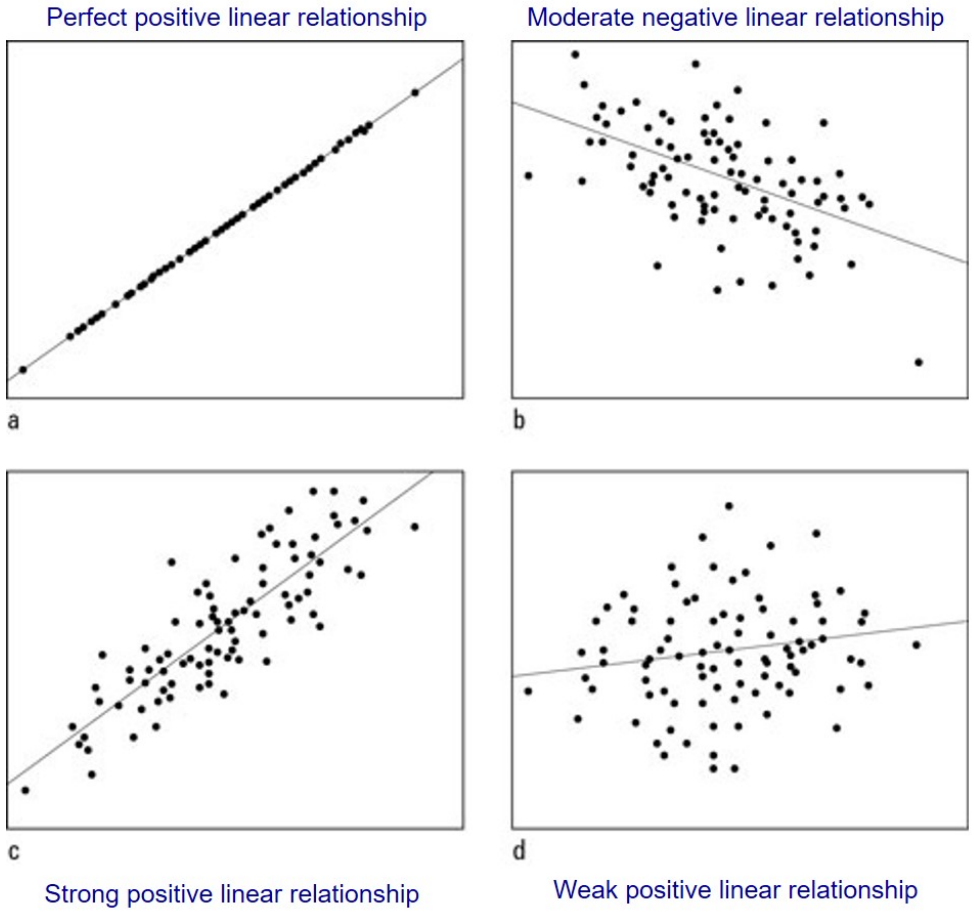
Scatterplot

- A scatter plot or scattergraph is a type of mathematical diagram using Cartesian coordinates to display values for two variables for a set of data.
- The data is displayed as a collection of points, each having the value of one variable determining the position on the horizontal axis and the value of the other variable determining the position on the vertical axis.
- Also known as scatter chart, scattergram, scatter diagram or scatter graph.



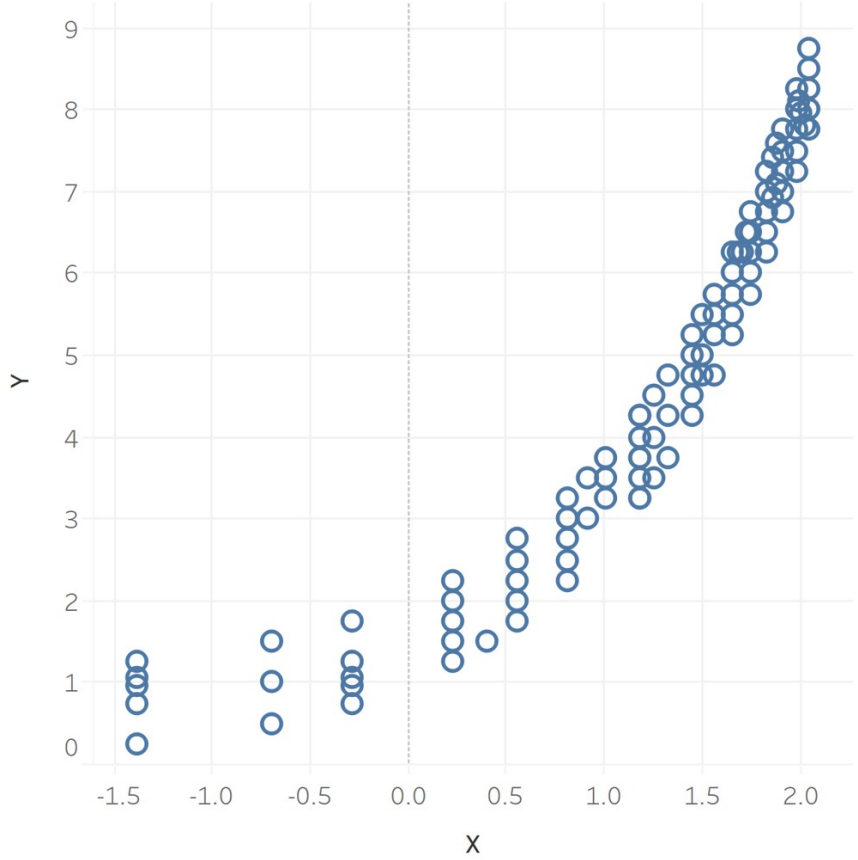
Visualising Relationship Between Two Continuous Variables

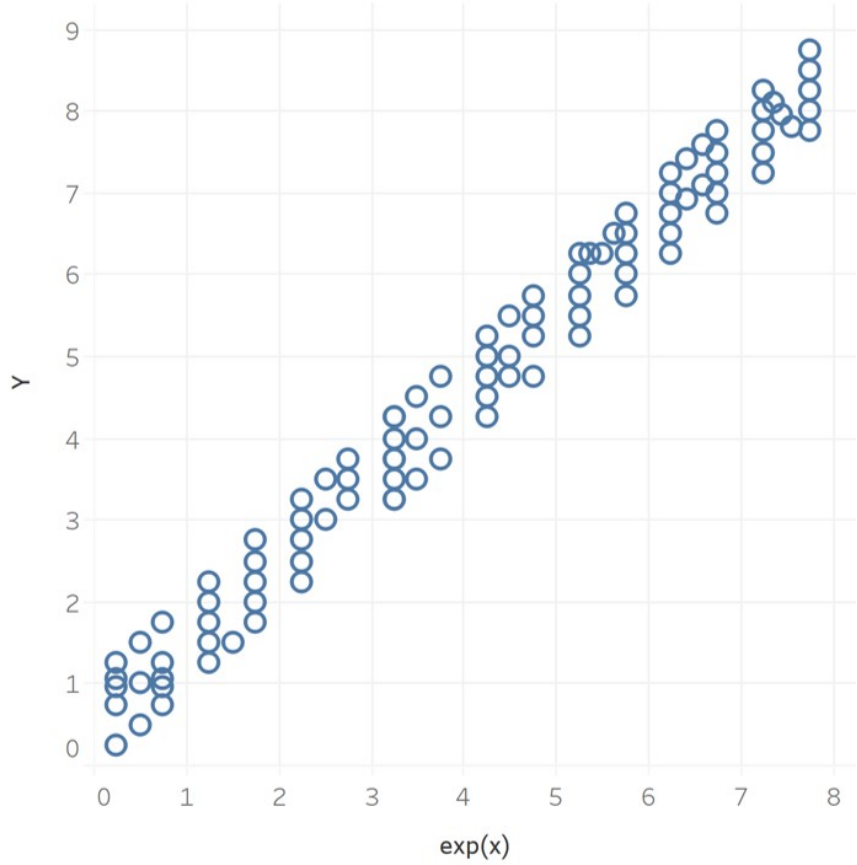
Interpreting scatterplot



Visualising Relatinships

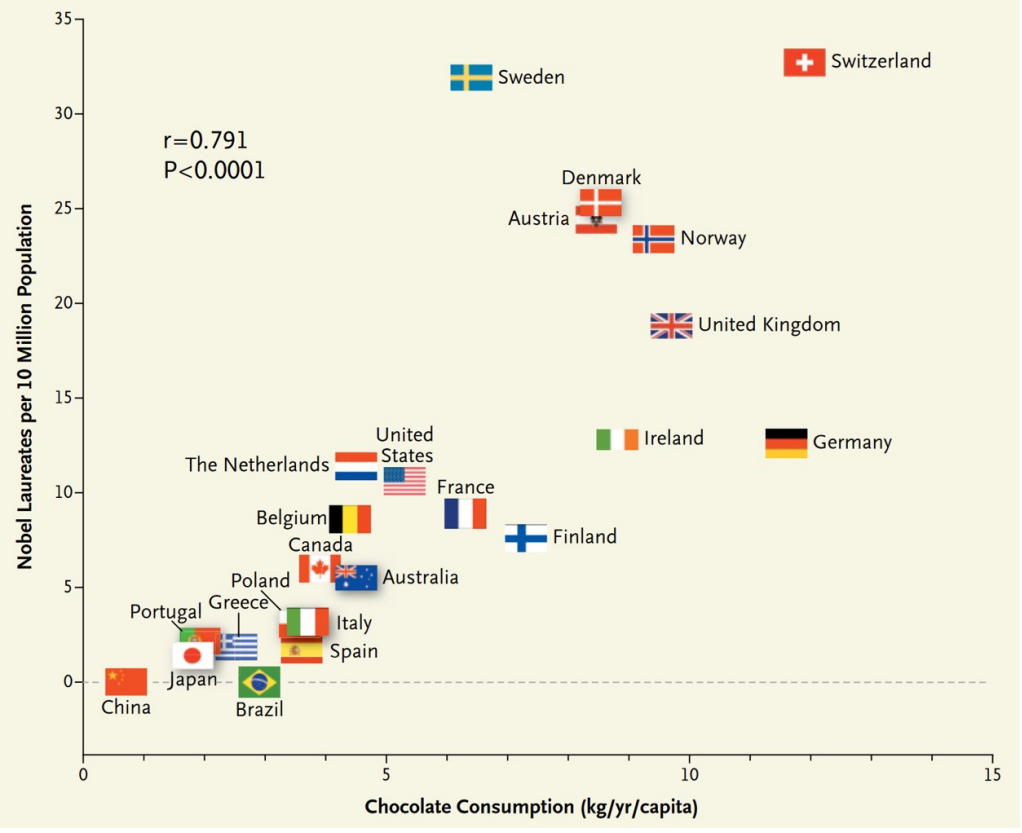
Caution: Not all relationships are linear





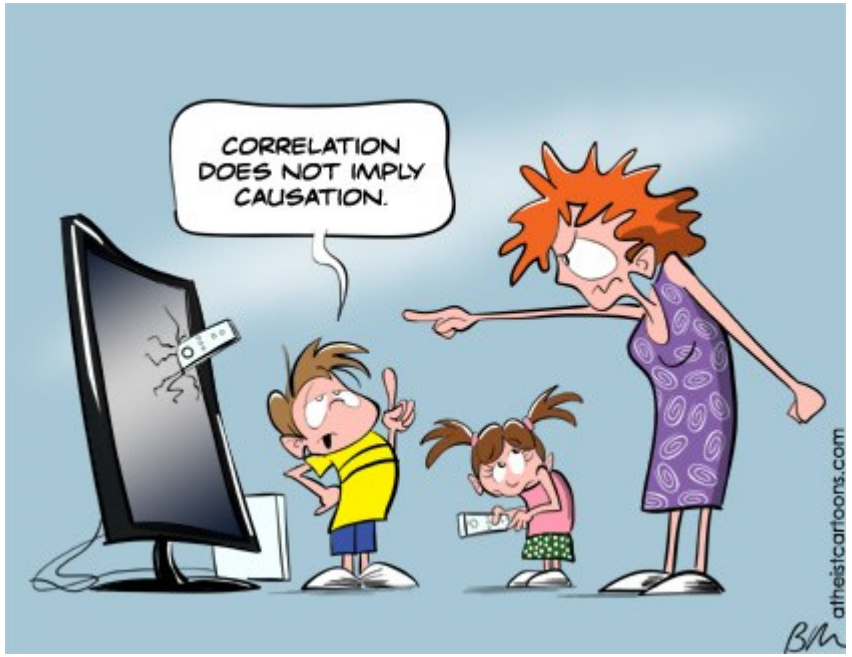
Visualising Relationships

Warning!



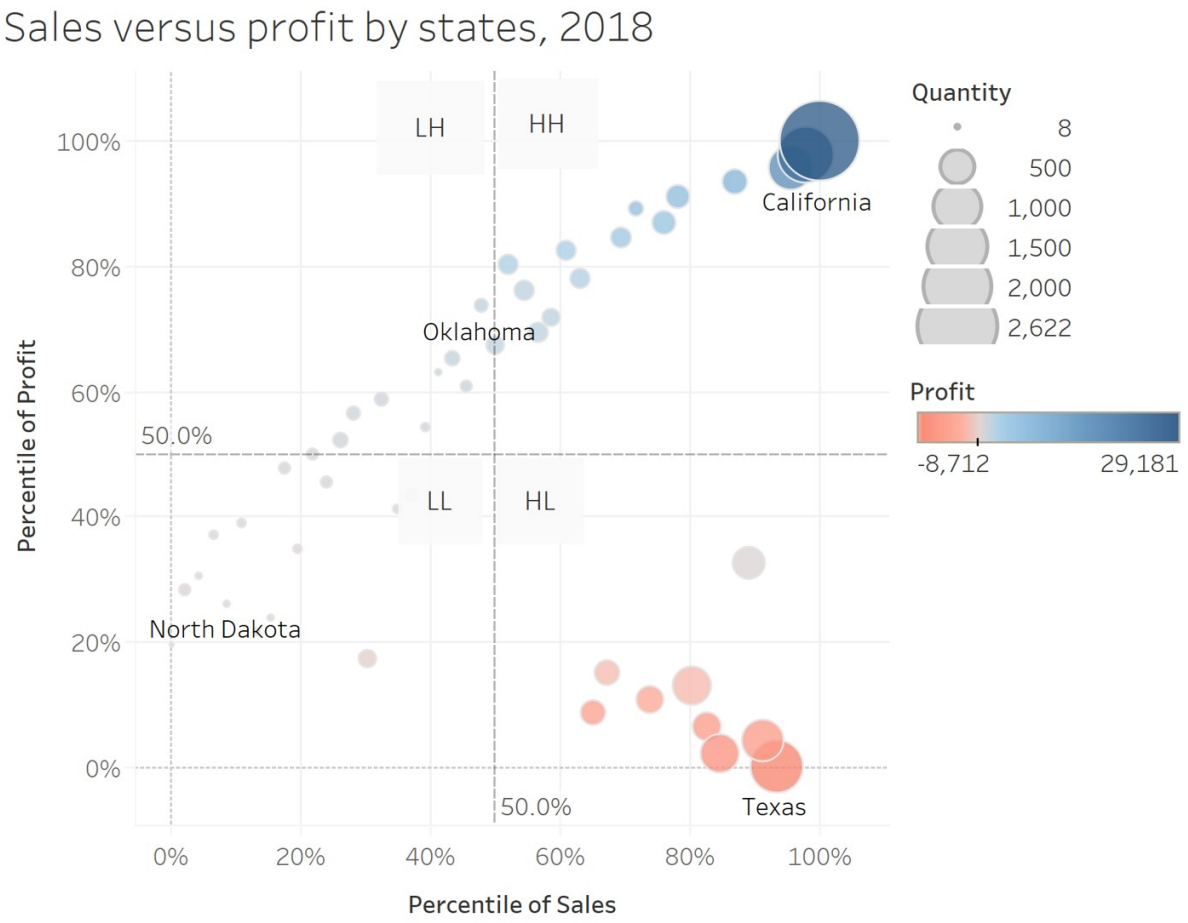
Source: Messerli (2012) "Chocolate Consumption, Cognitive Function, and Nobel Laureates", *The New England Journal of Medicine*.pdf)

Correlation does not imply causation



Visualising Relationship

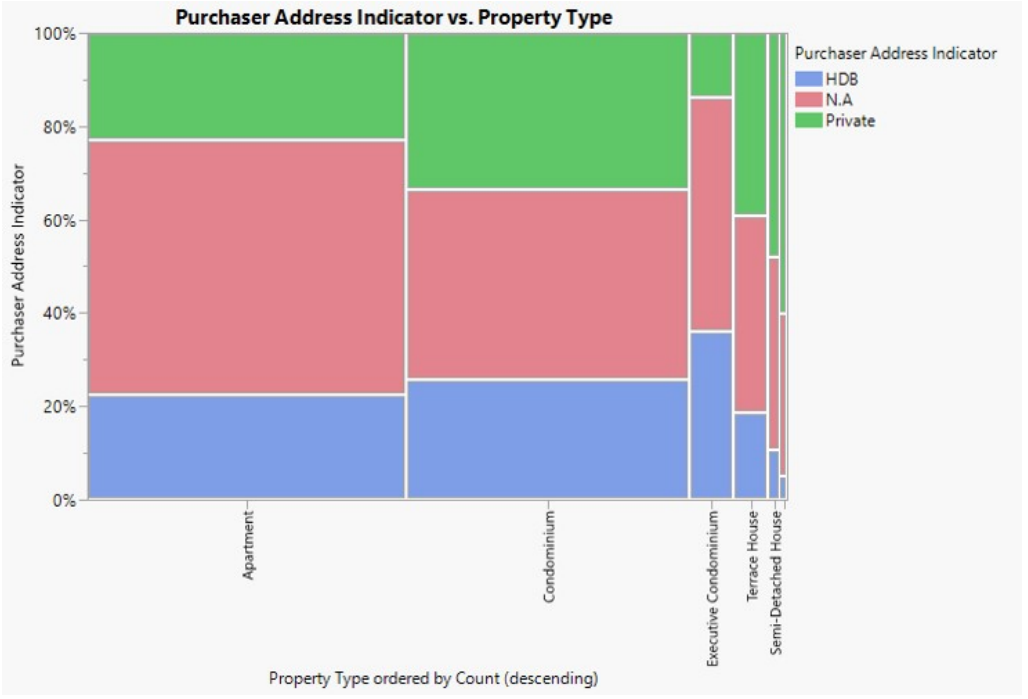
Quadrat analysis



Visualising Relationship Between Two Categorical Variables

Mosaic Plot

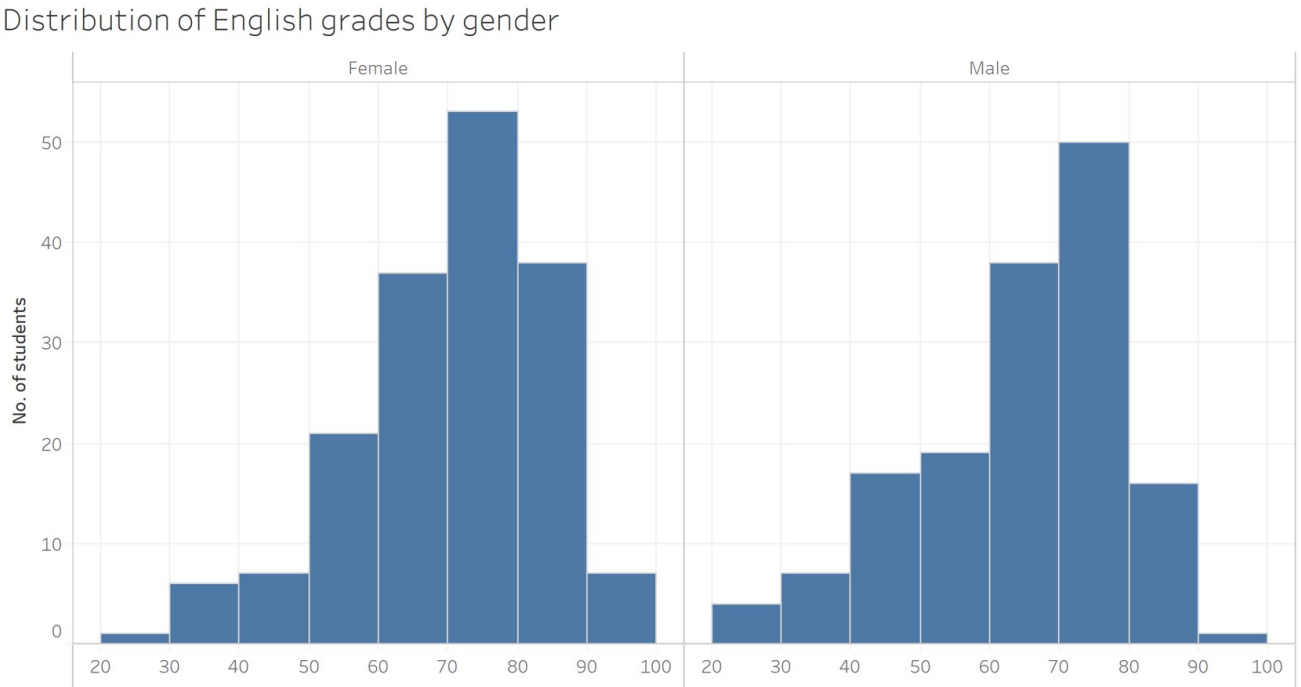
- A mosaic plot is a graphical display that allows you to examine the relationship among two or more categorical variables.



Visualising Relationship Between Sub-groups

Trellis

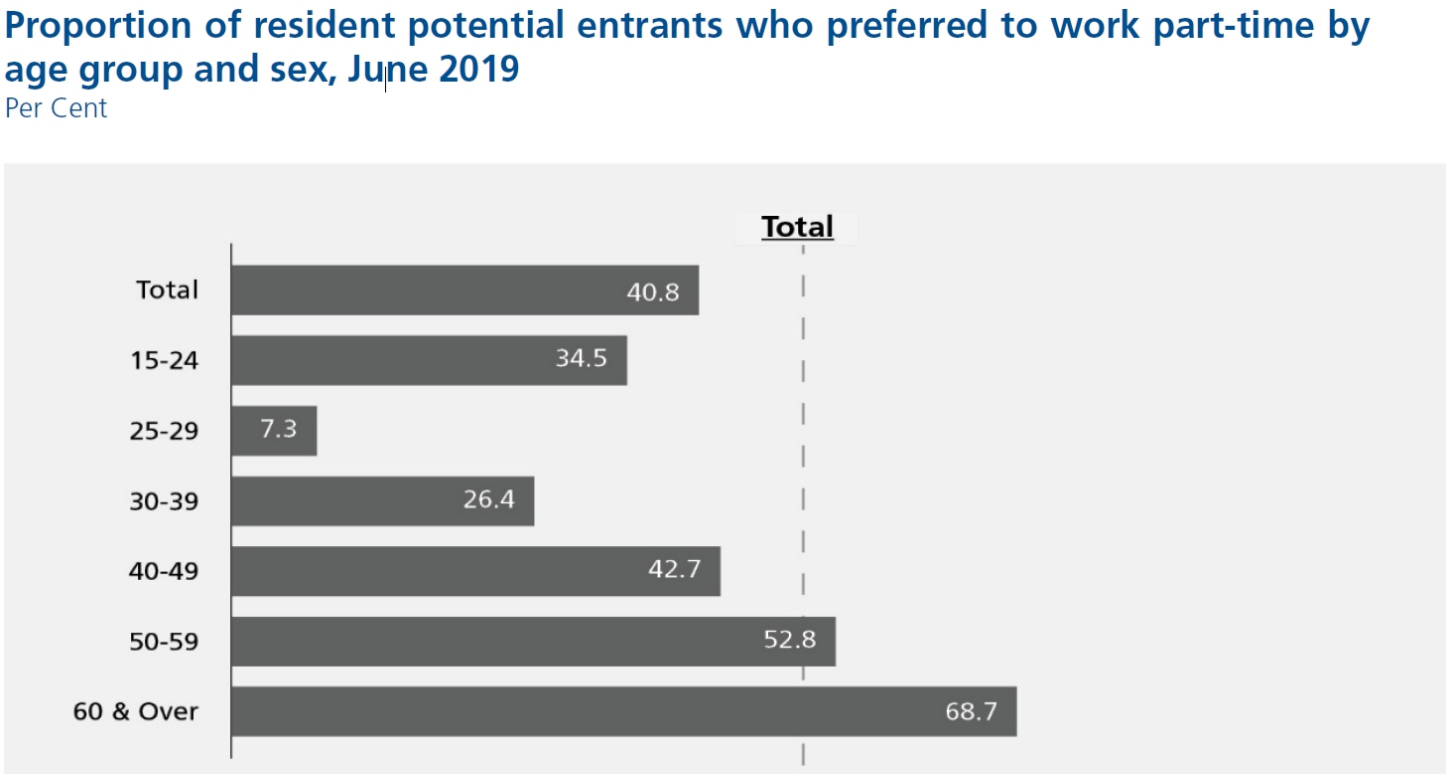
- Trellised visualizations enable you to quickly recognize similarities or differences between different categories in the data.
- Each individual panel in a trellis visualization displays a subset of the original data table, where the subsets are defined by the categories available in a column or hierarchy.



Visualising Uncertainty

Why it is important?

- One of the most challenging aspects of data visualization is the visualization of uncertainty.

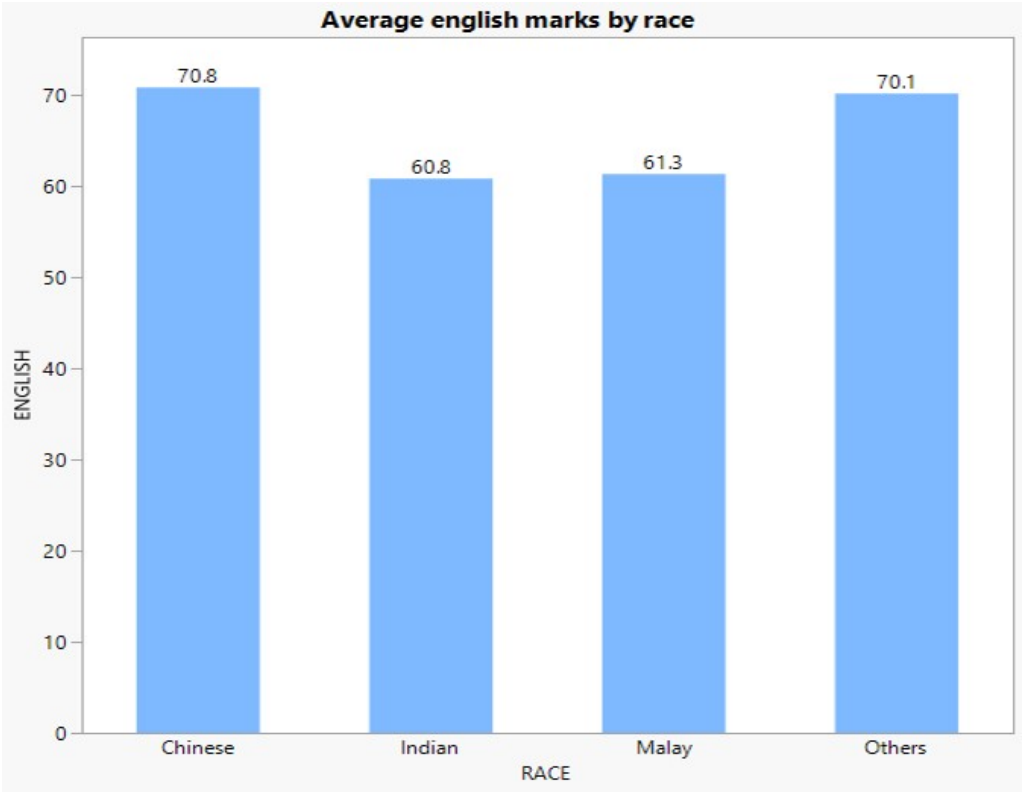


Source: Chart 61, LABOUR FORCE IN SINGAPORE 2019, pg. 52

Visualising Uncertainty

Commend mistake

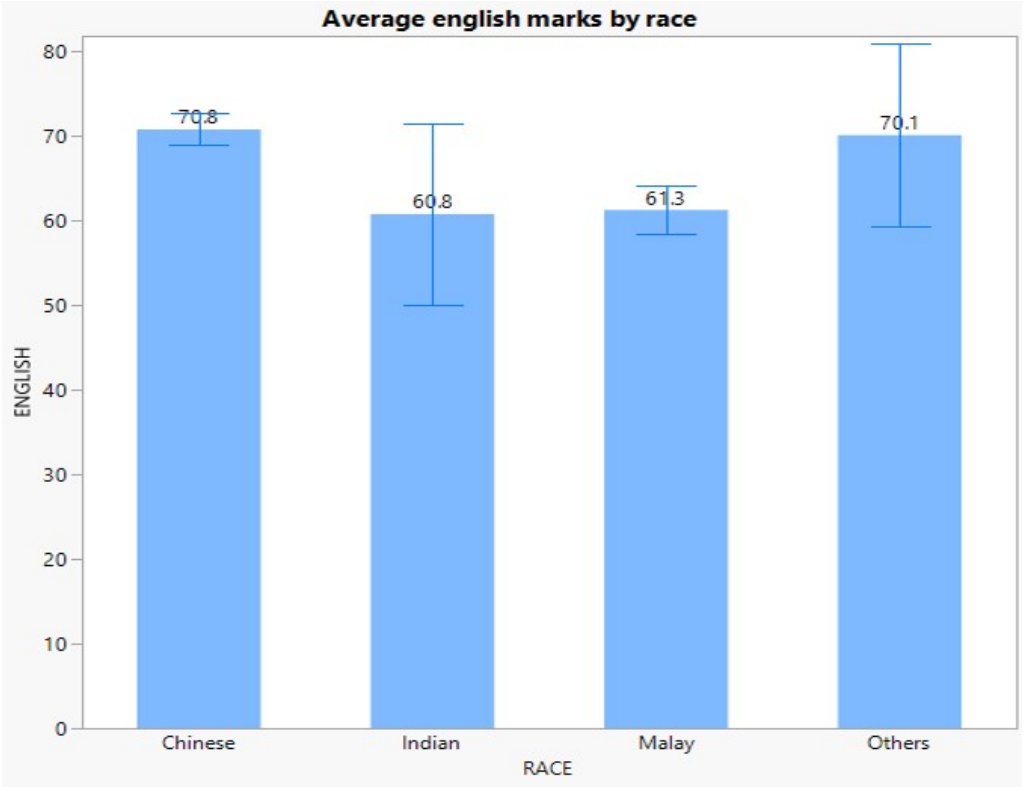
- It is not appropriate to displace average values on bars



1-d statistical graphic methods for visualising uncertainty

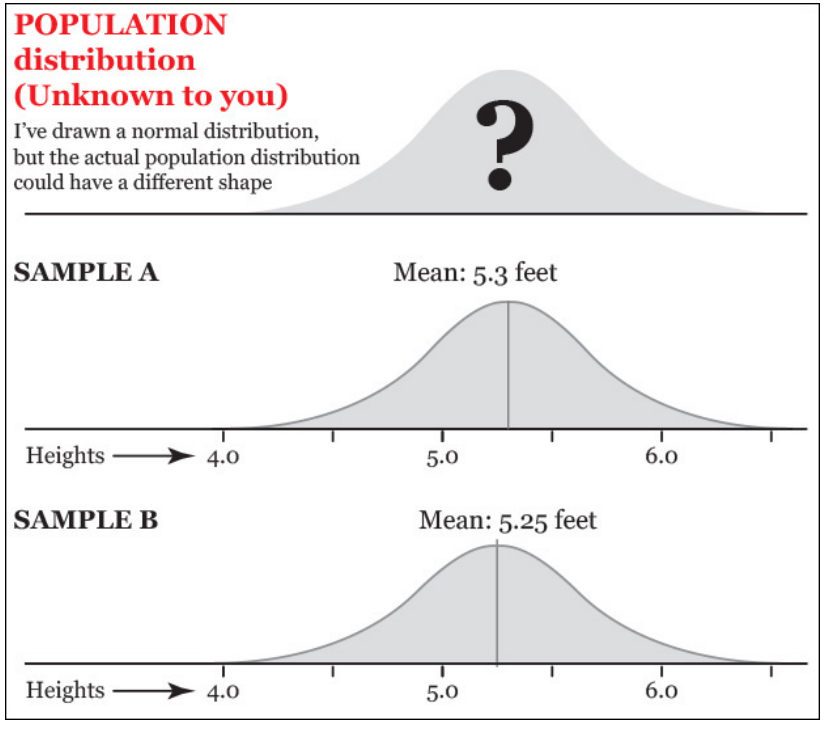
Error bar on a bar chart

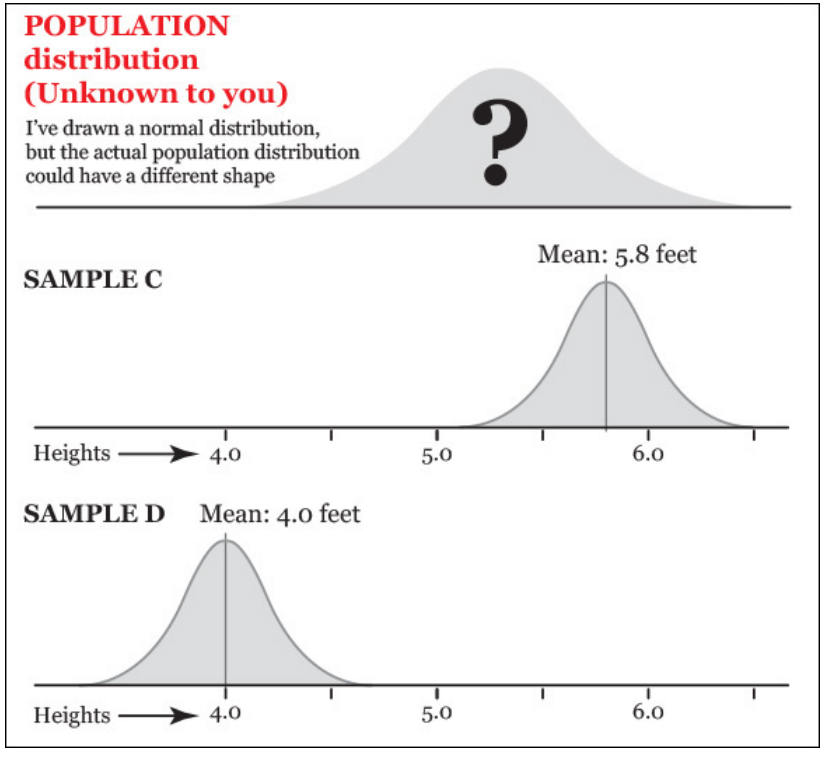
- Each error bar is constructed using a 95% confidence interval of the mean.



Back to Statistics 101

Population and samples

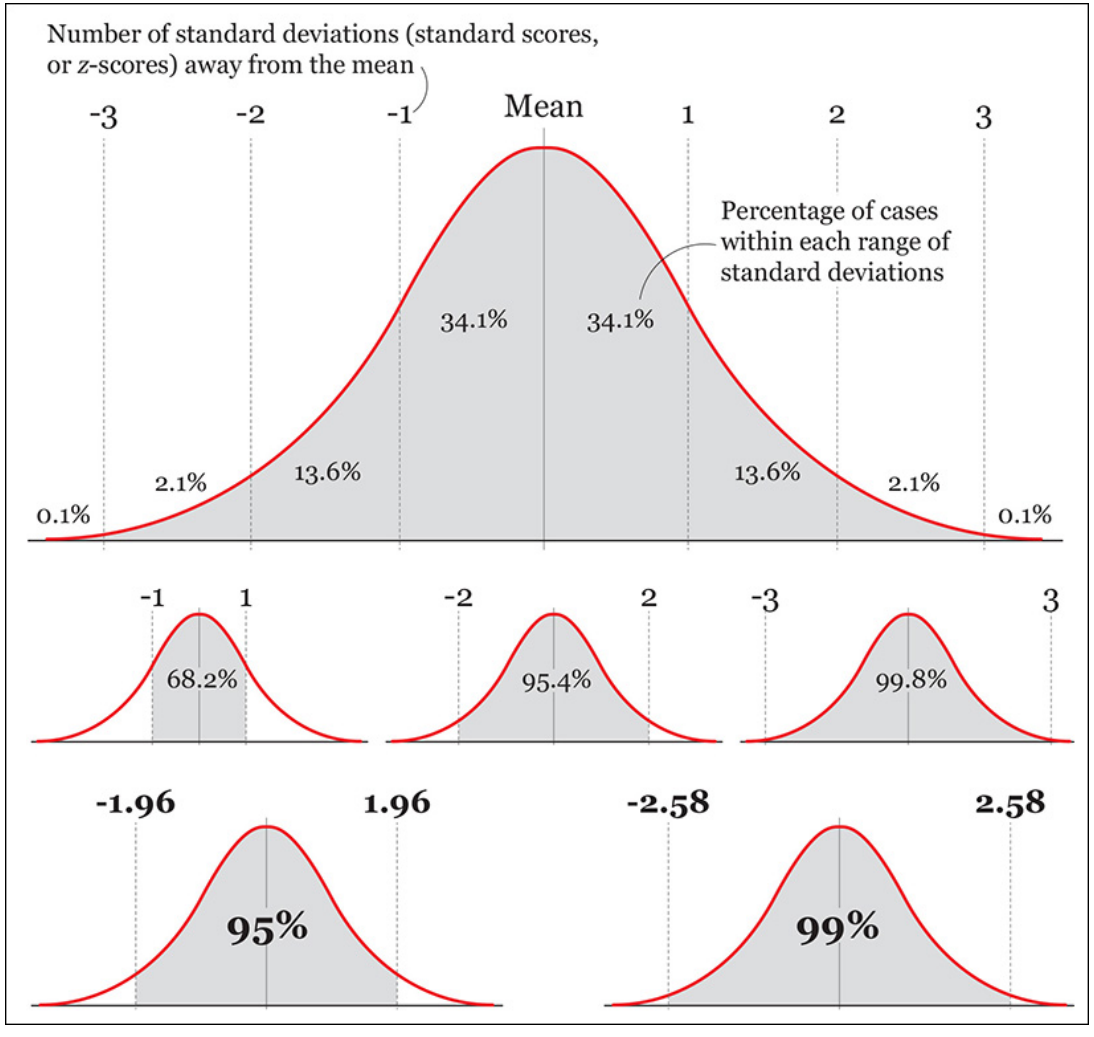




- When drawing many samples from a population, it is possible to obtain a few with means that greatly differ from the population.

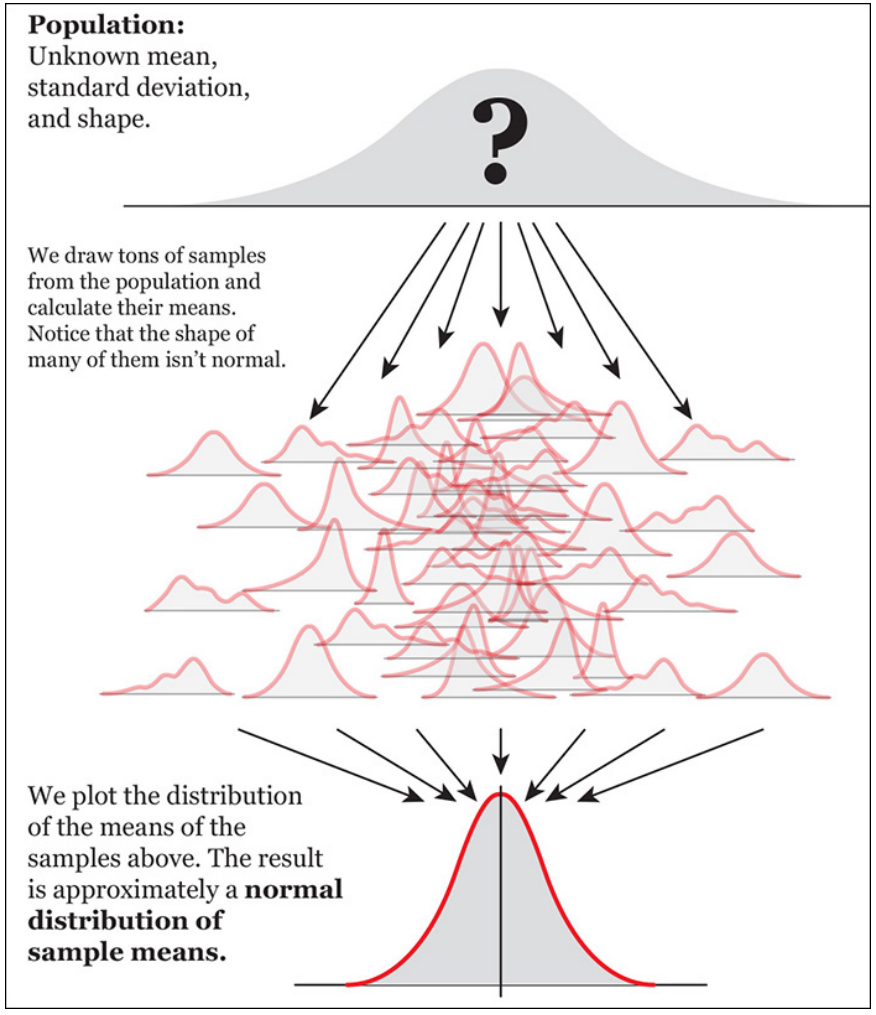
Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

A reminder of the standard normal distribution



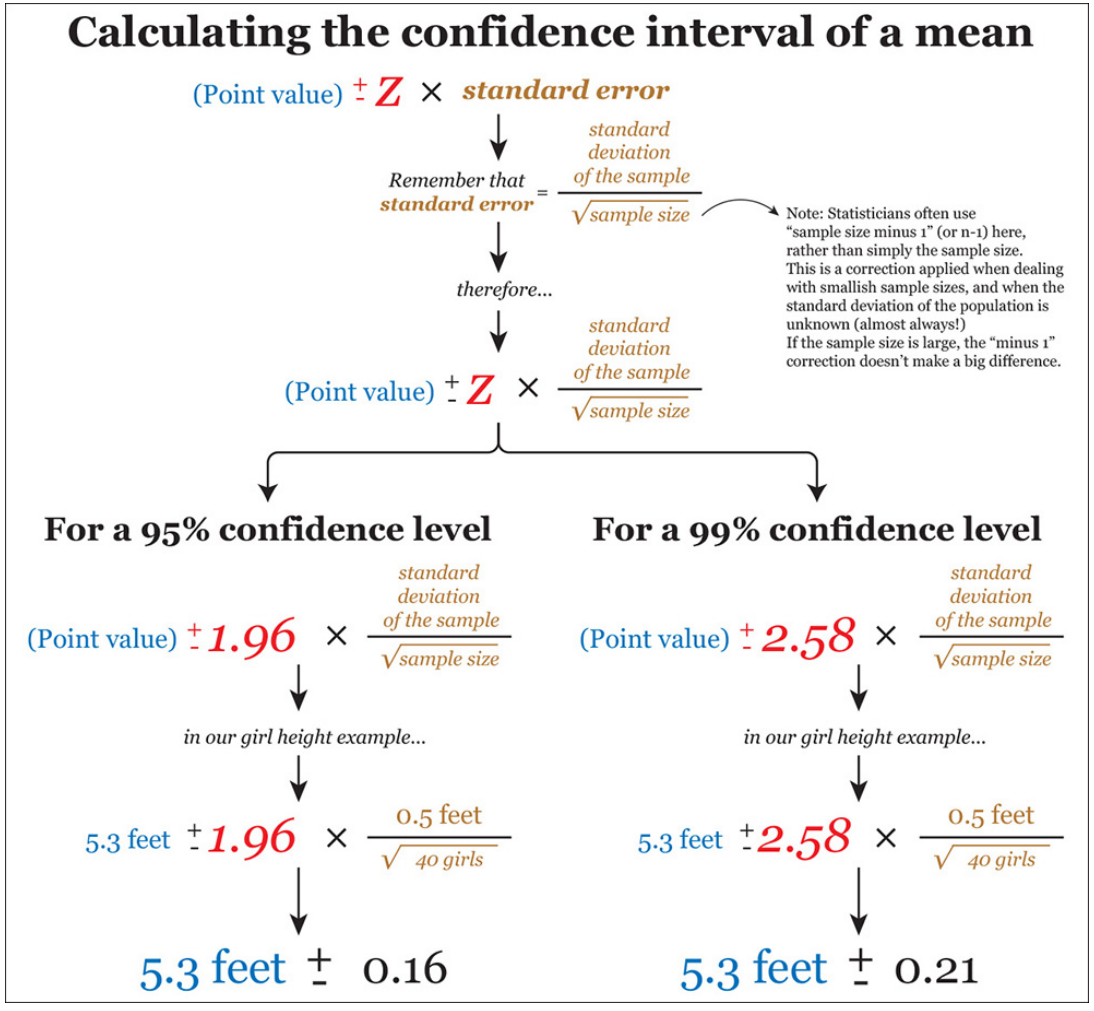
Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

The standard error



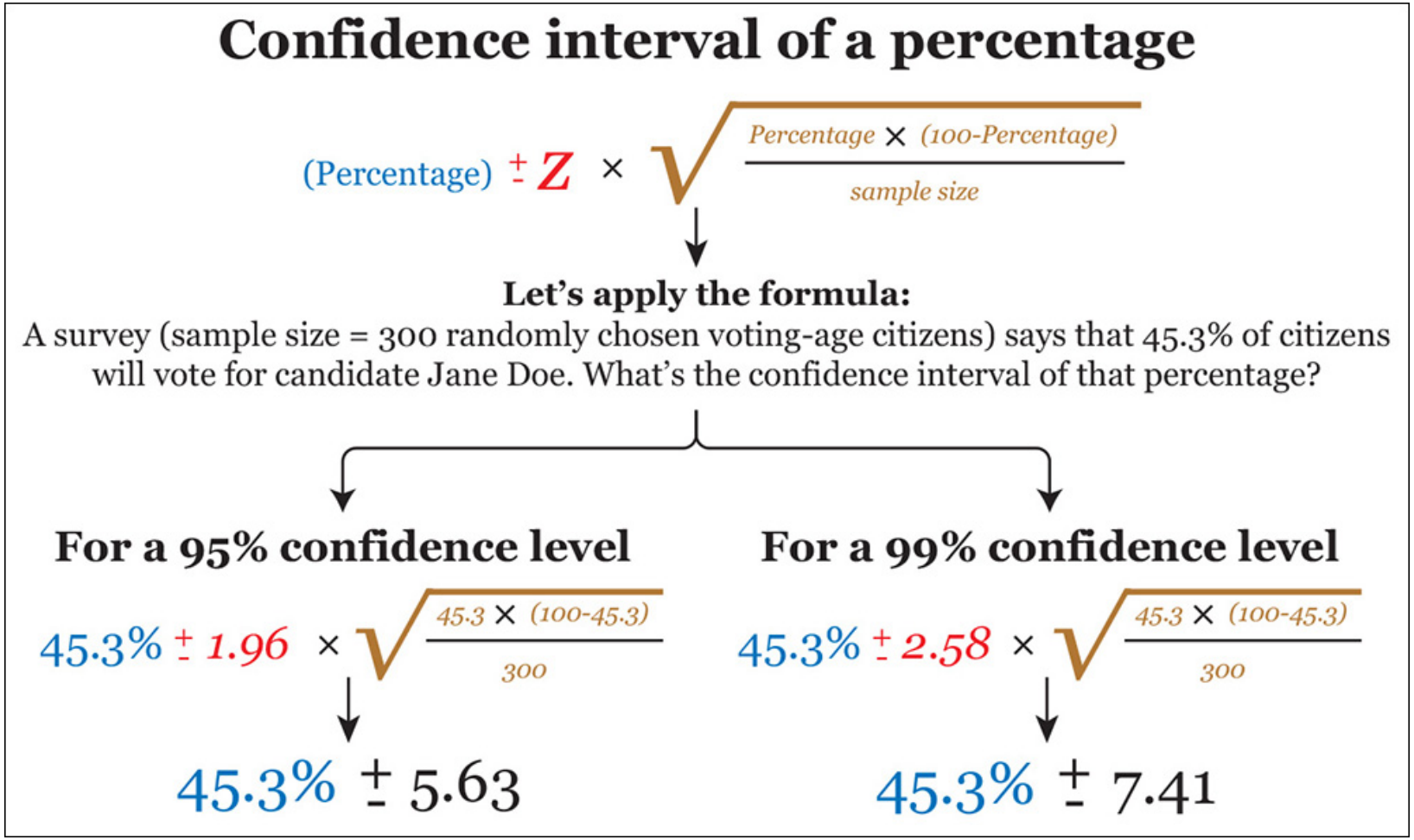
Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

Calculating the confidence interval of a mean



Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

Calculating the confidence interval of a percentage

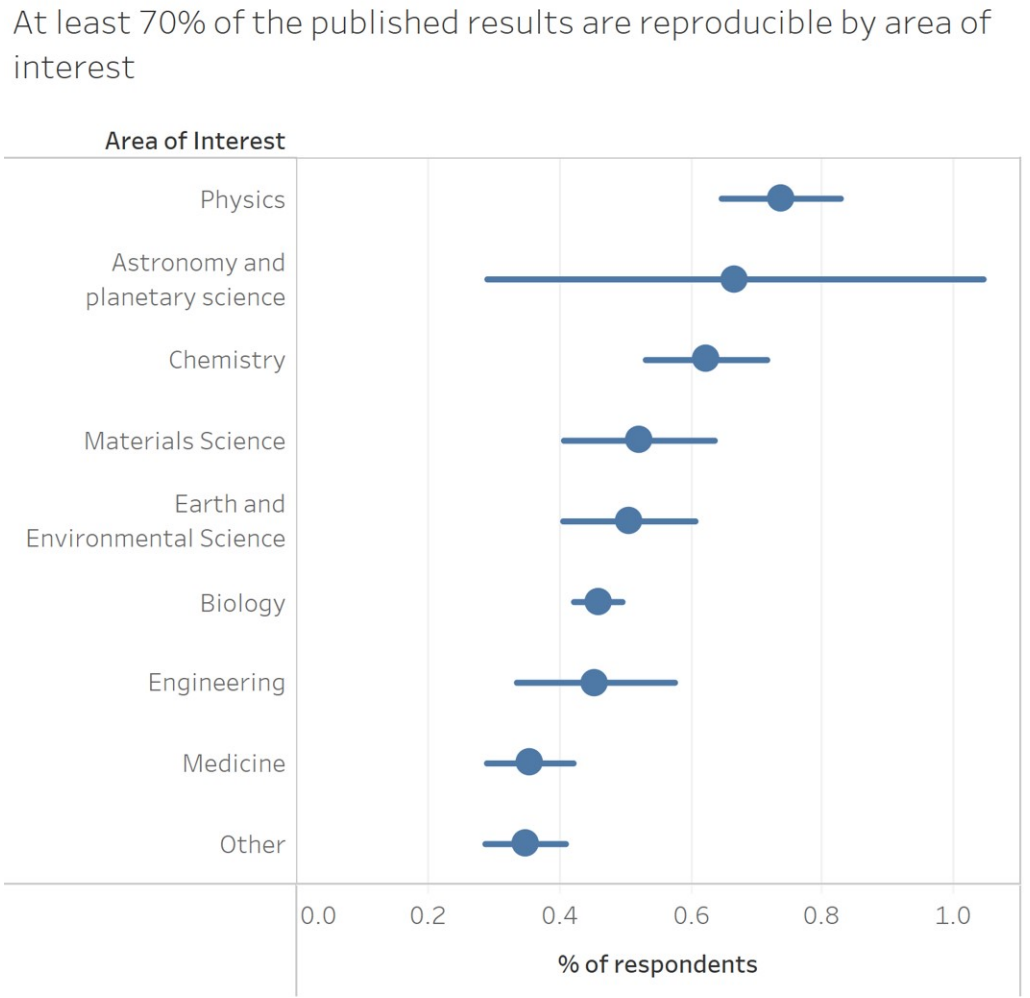


Reference: Cairo, A. (2016) The Truthful Art, Chapter 11, New Riders

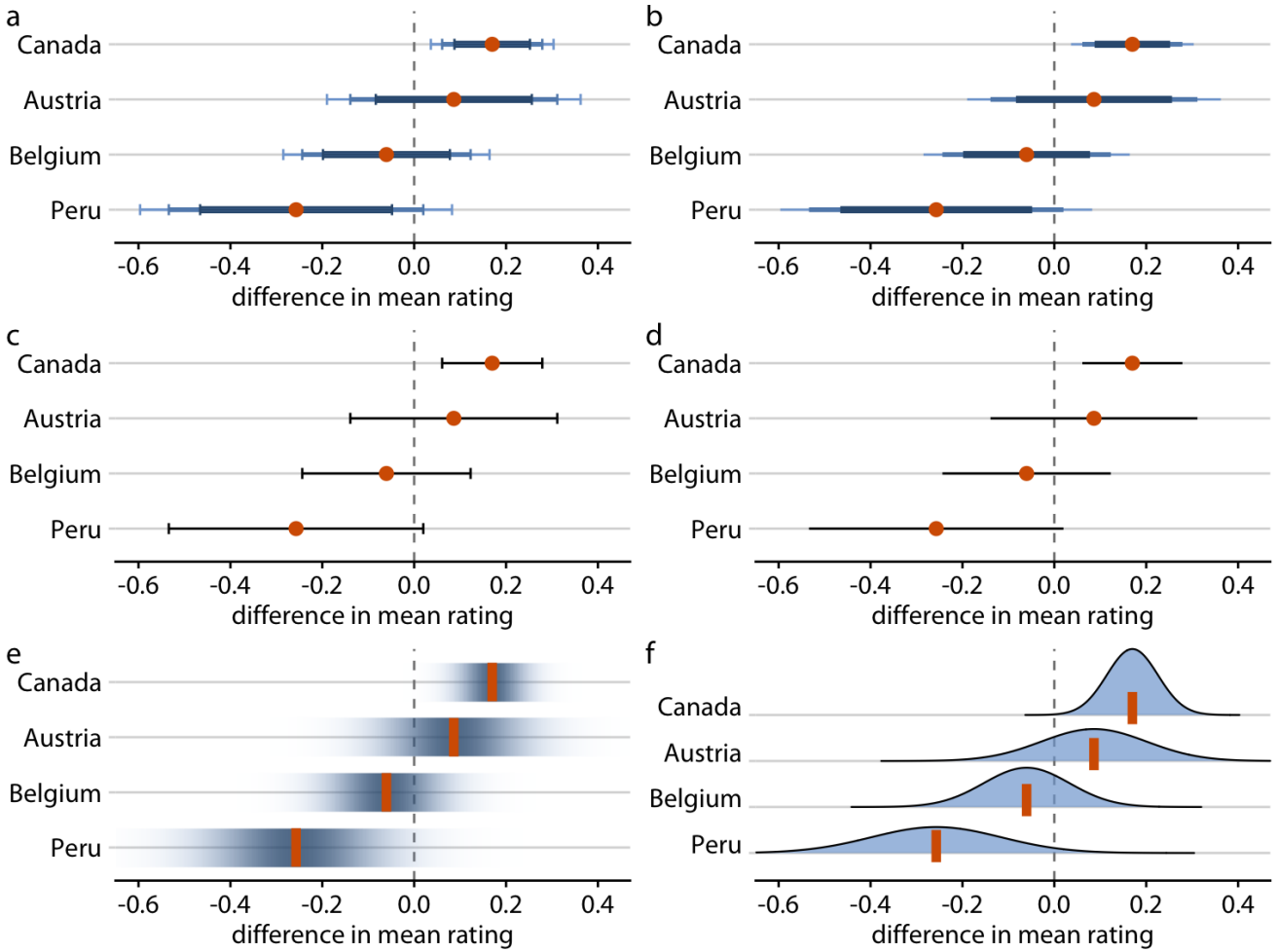
1-d graphical methods for visualising uncertainty

Error bar on a dot plot

- Each error bar is constructed using a 95% confidence interval of the percentage.



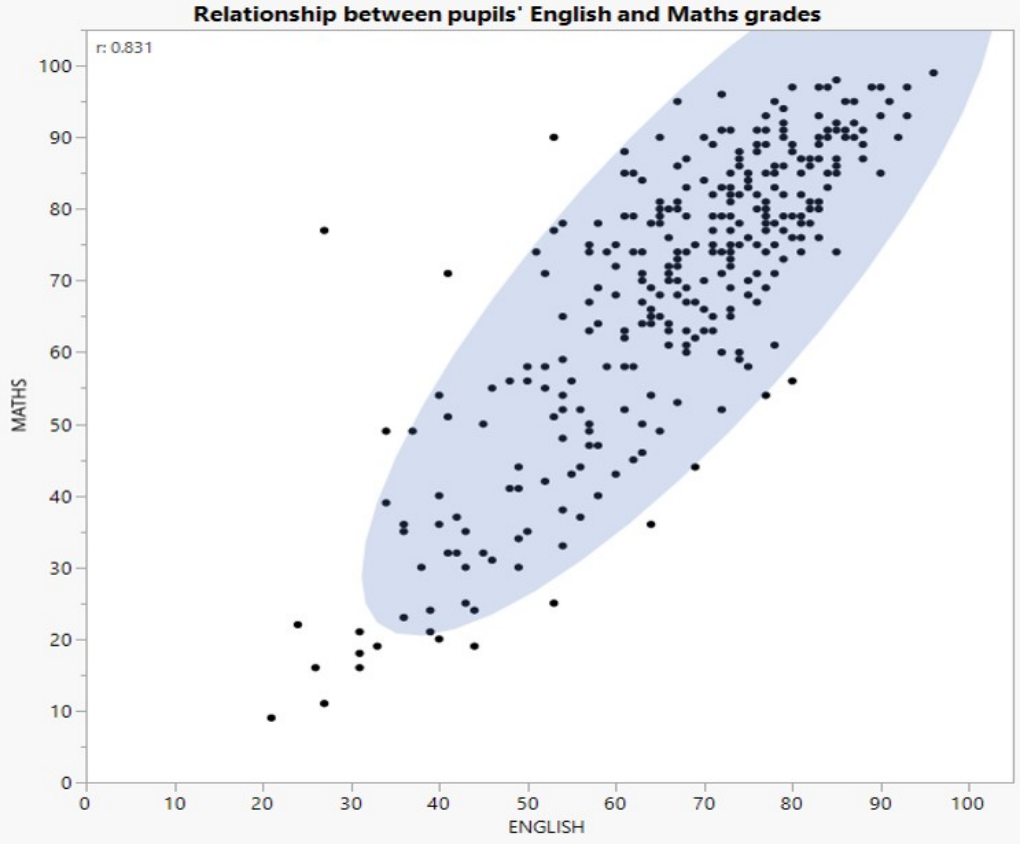
Graphical methods for visualising uncertainty



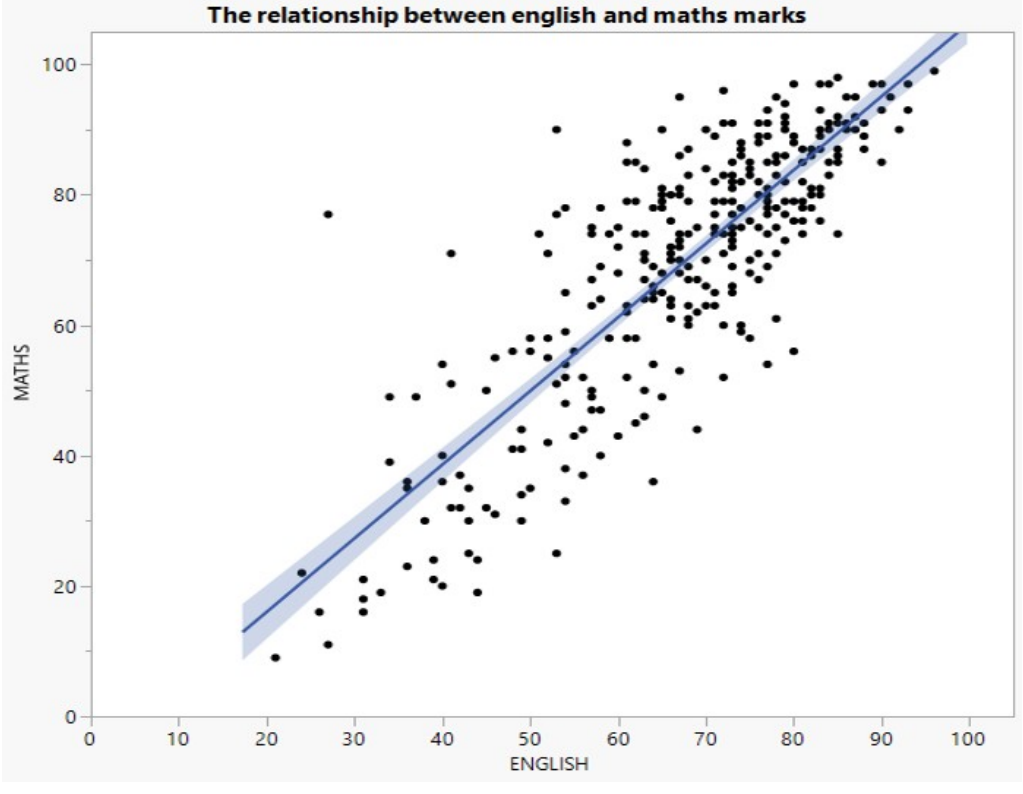
Reference: https://serialmentor.com/dataviz/visualizing-uncertainty.html#visualizing-the-uncertainty-of-point-estimates

2-d graphical methods for visualising uncertainty

Scatter plot with 95% confidence ellipse



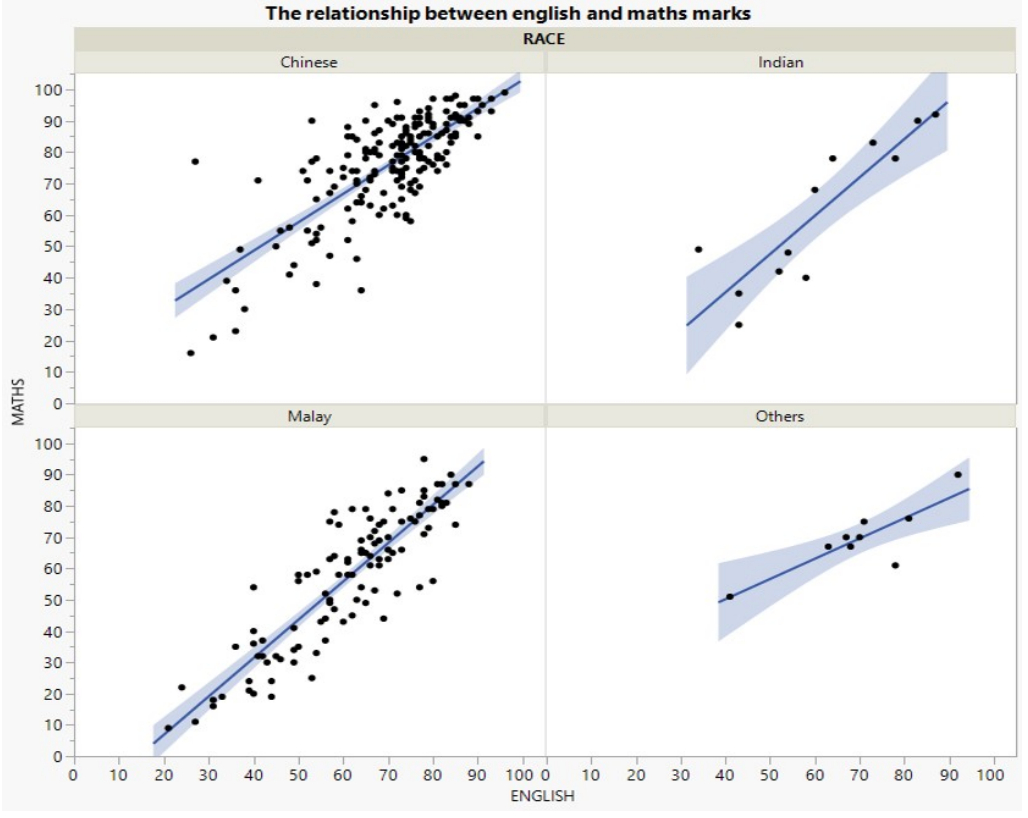
Scatter plot with 95% confidence intervals



2-d graphical methods for visualising uncertainty

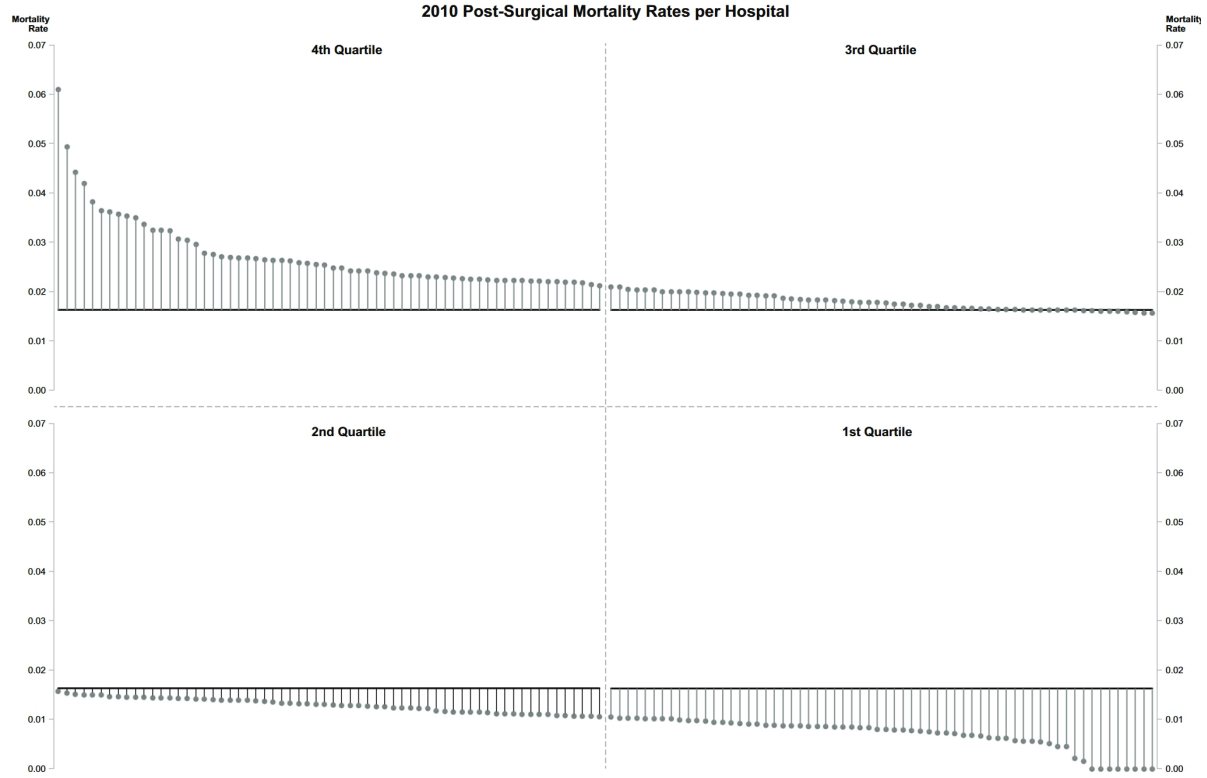
Trellis scatter plot with 95% confidence intervals

Figure on the right reveals that the relationships between english and maths grades for Indians and Other races are relatively less certain than Chinese and Malays.



Variation and Its Discontents

Random and unfair comparisons



Reference: https://www.perceptualedge.com/articles/visual_business_intelligence/variation_and_its_discontents.pdf

Variation and Its Discontents

Funnel plots to the rescue

Statistical details

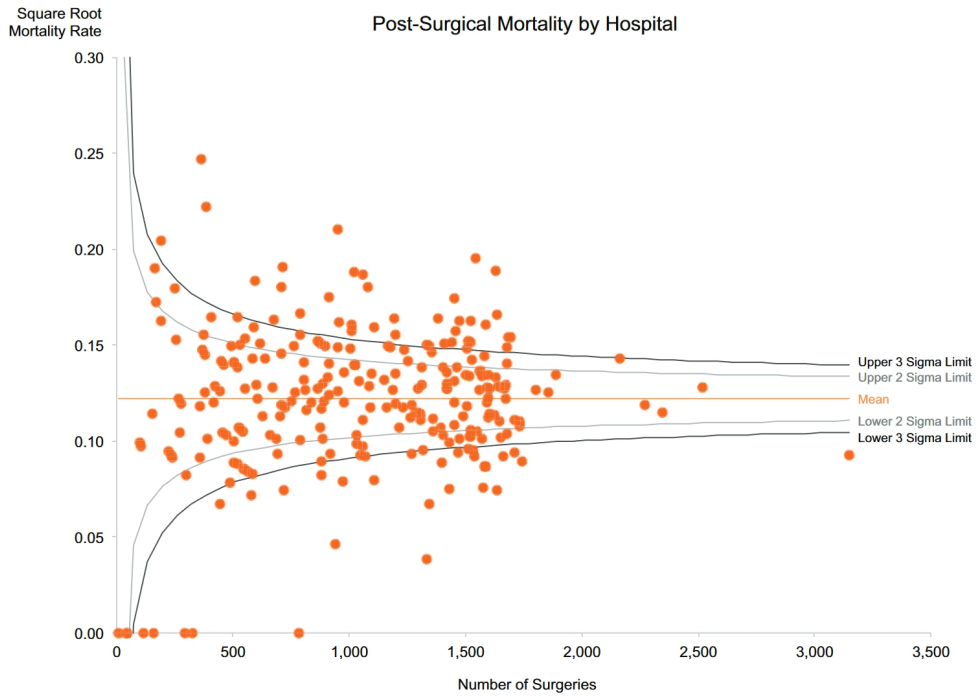
Calculations:

Fit Mean = Sum of Occurrences / Sum of Sample Sizes

$$\text{Two Sigmas (95\% Limit)} = \text{Fit Mean} \pm 1.96 * \sqrt{\frac{\text{Fit Mean} * (1 - \text{Fit Mean})}{x}}$$

$$\text{Three Sigmas (99.8\% Limit)} = \text{Fit Mean} \pm 3.0 * \sqrt{\frac{\text{Fit Mean} * (1 - \text{Fit Mean})}{x}}$$

$$\text{Histogram Values} = \frac{\left(\frac{\text{Occurrences}}{\text{Sample Sizes}}\right) - \text{Fit Mean}}{\sqrt{\frac{\text{Fit Mean} * (1 - \text{Fit Mean})}{\text{Occurrences}}}}$$



Normality assumption

- Before plotting the graph, it is important to check if the values are conformed to normal distribution assumption.

- If the raw values are not conformed to normality assumption, they have to be transformed.