DATA SCIENCE & MACHINE LEARNING WORKSHOP

Week 3 - Intro to Statistical Analysis

INDUSTRY APPLICATIONS MASTERCLASS

ARE YOU LOOKING FOR AN INTERNSHIP OR PLACEMENT?

WONDERING HOW TO START APPLYING OR LOOKING FOR SOME TIPS THAT COULD ULTIMATELY BOOST YOUR APPLICATIONS?

THIS EVENT IS GOING TO HAPPEN ON WEDNESDAY 29TH OCTOBER AT THE MURRAY LEARNING CENTRE (THE BUILDING OPPOSITE THE COMPUTER SCIENCE BUILDING) IN UG09.

LIMITED AVAILIBILITY!!!
SCAN THE QR CODE TO SECURE YOUR SPOT!!!

https://luma.com/sx1nfjll

TALK FROM PREVIOUS INTERNS AT:

NOMURA
BARCLAYS
CONFLUENT
AUTODESK
BlueCrest
BlueCrest Capital Management
NEXT

AGENDA

Week 3 topic:

- Recap of Week 2
- Descriptive statistics
- Correlations
- Data visualisation principles
- Data Visualisation types, common plots
- Multivariate correlations with visualisation
- Intro to matplot and Seaborn

Full agenda this semester:
https://bit.ly/DataScienceAlgosoc

WEEK 2 RECAP

ALGOSOC UOB

Types of data:

- Categorical (ordinal and nominal)
- Numerical (continuous and discrete)
- non-structured (images, texts, signal)

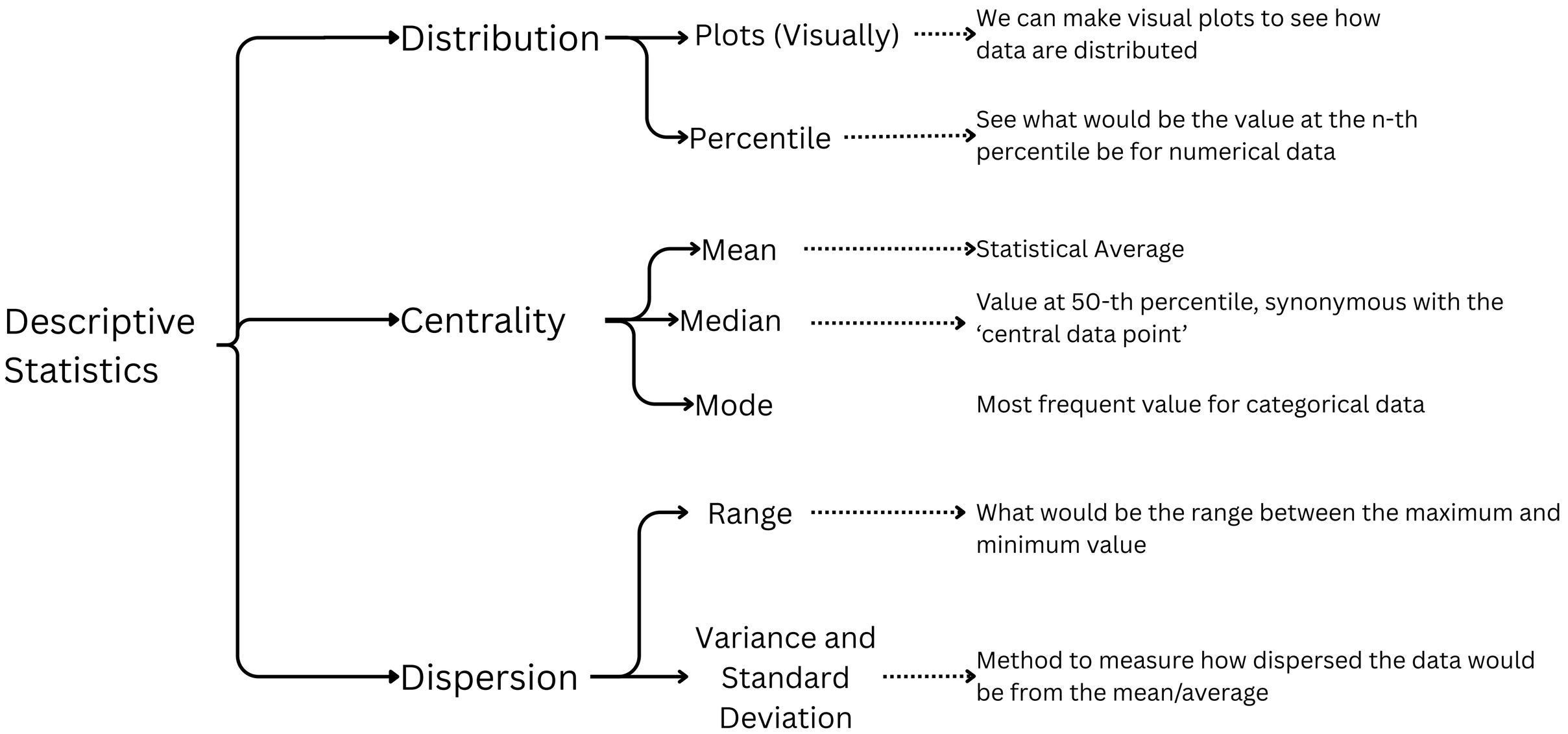
We have structured and unstructured data (most of data used in enterprise are unstructured) → we need knowledge on how to process them

Data sources, for practice and research, we can use UCI, Kaggle, and HuggingFace repositories. However, these are not enough for production and deployment in business

Problems: missing values, outliers, and inconsistencies

How to solve those problems: data wrangling methods

DESCRIPTIVE STATISTICS



DESCRIPTIVE STATISTICS

Distributions

We can use **visualisations** such as histogram to see how data are distributed

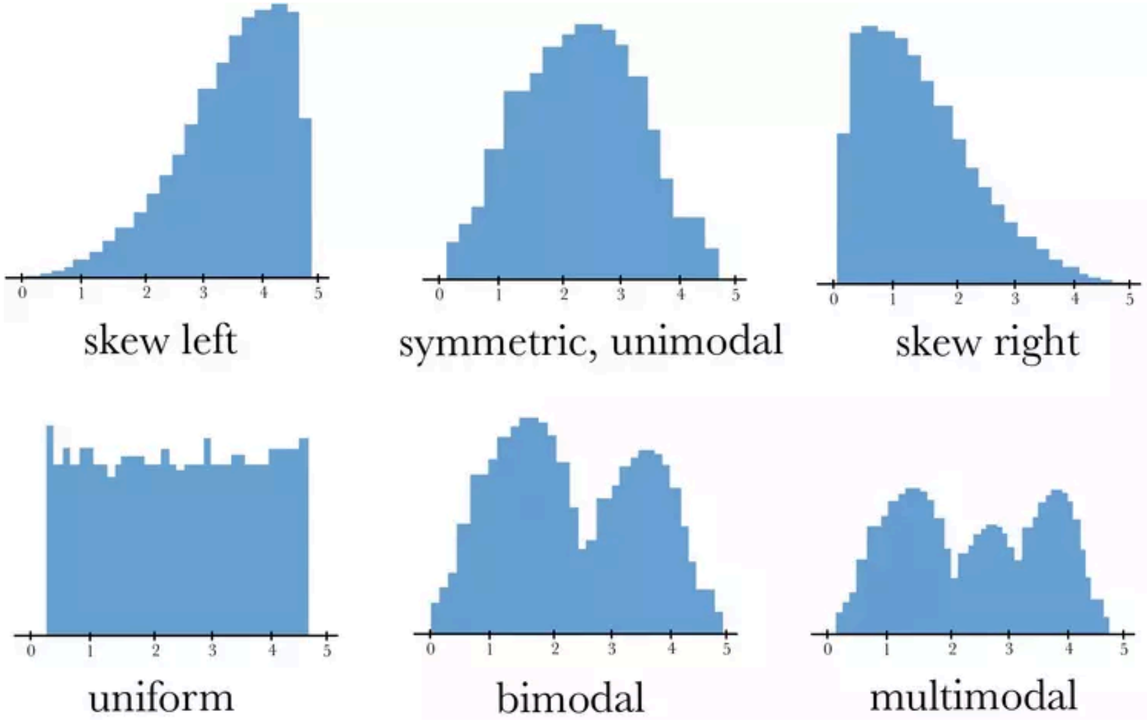


image source: https://serc.carleton.edu/mathyouneed/geomajors/histograms/index.html

A good way to think of **percentiles,** if your data looks like this: [1,2,...99, 100]. Then, the 95-th percentile value of your data would be 95

Boxplots are also good visualisation tool to see how data are distributed in their **percentiles.**

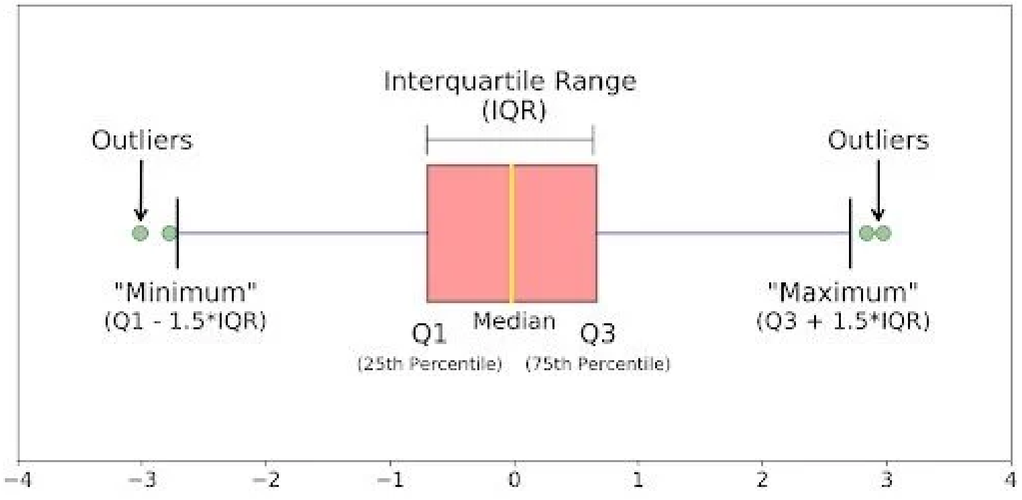


image source: https://builtin.com/data-science/boxplot

DESCRIPTIVE STATISTICS

Centrality

If all values of the data are **equally likely** to come up when randomly chosen, **mean** would be the **sum of all values of the data divided by the number of values the data have**

For medians, if your data look like this: **[1,2,3,4,5], 3 would be the median**

Why normal (i.e., bell curve) distribution matters? Because we would not want the average to be very different from the median.

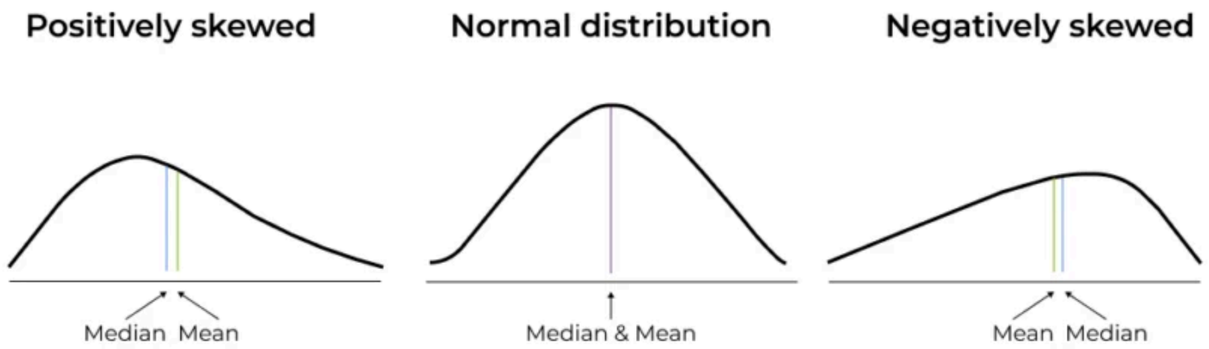
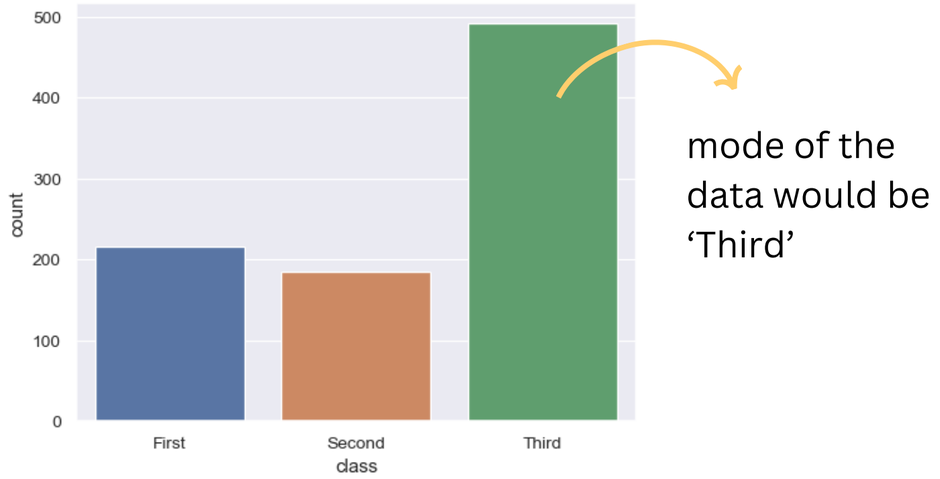


image source: https://gopractice.io/data/arithmetic-mean-and-median-for-product-managers/

For numerical data, **modes** are not used as much as the median (including percentiles) and means. Going back to our previous session, numerical data in today's case would most likely be continuous.

There's only a small chance that two data points from a continuous domain would have the exact same value. For example, two temperature measurements showing 25.5 degrees and 25.6 degrees; they are more or less the same but of different numbers, and **mode would not work.**



mode of the data would be 'Third'

image source: https://seaborn.pydata.org/archive/0.11/generated/seaborn.countplot.html

DESCRIPTIVE STATISTICS

Dispersion

Range of numerical data would be **amount of space** between the maximum and minimum values.

For example, your data might be: [1,2,3,4,5], then the range would be: 5-1 = 4

On the other hand, **variance and standard deviation** measure how the data are dispersed around the mean.

If we have:

- Mean = 5.0
- Standard Deviation = 2

That would mean that roughly **68.26% of the data are between 3.0 and 7.0** (if normally distributed, i.e., bell-shaped).

While standard deviation measures the dispersion in the same unit as the data (e.g., cm), variance measures the dispersion in the squared unit (e.g., cm^2)

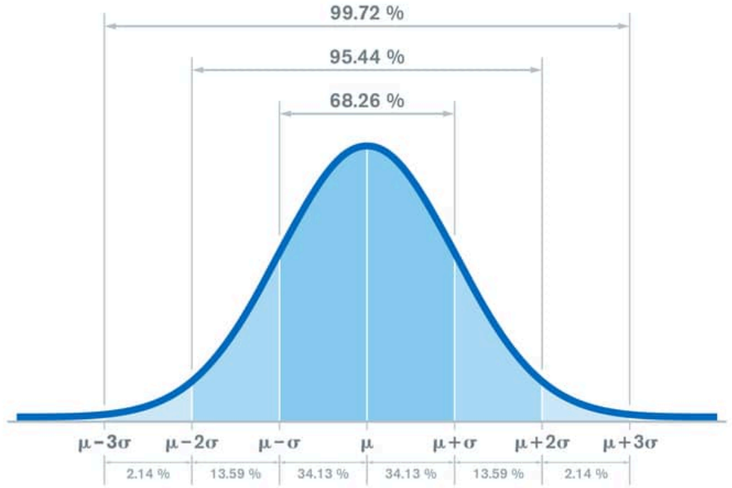
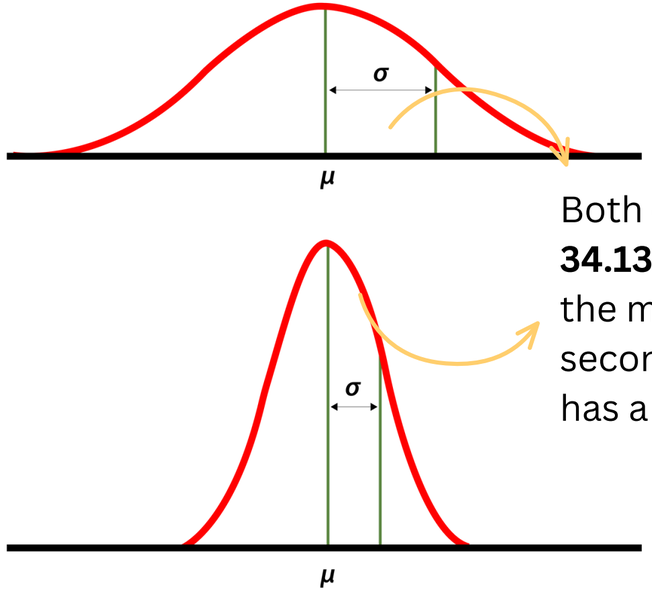


image source: https://www.sixsigmadaily.com/standard-deviation-6-steps-to-calculation/



Both contain roughly **34.13% data** around the mean but the second distribution has a lower STD

CORRELATION

Correlation would be a **statistical measure** of relationships between two or variables.

Values would usually in the range of -**1 to 1**.

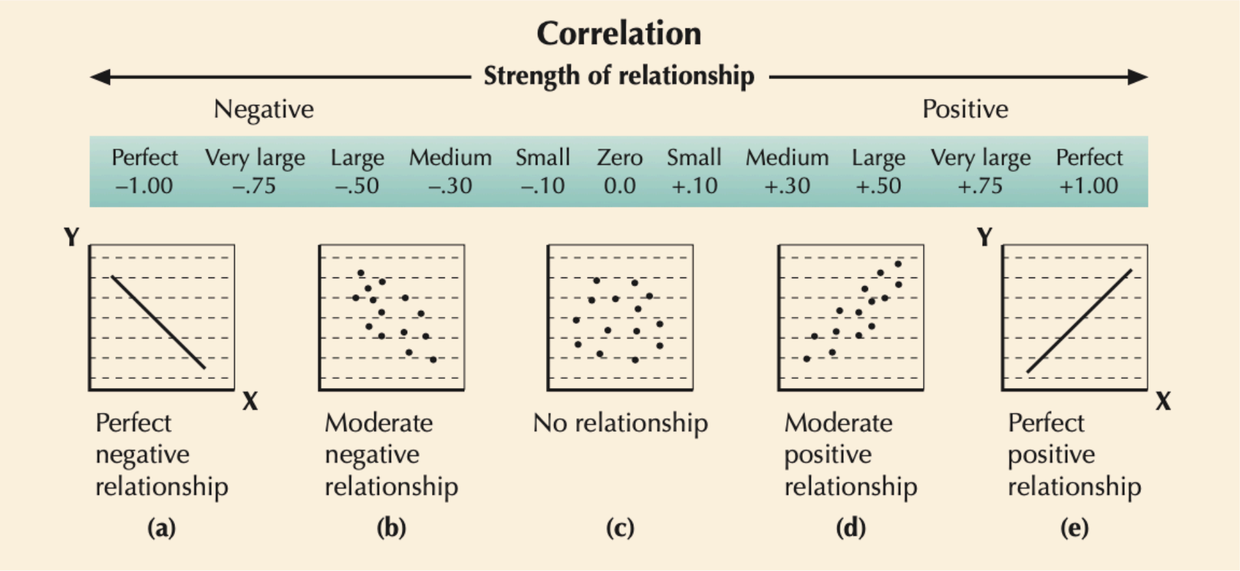


image source: https://www.glassthought.com/notes/olp57h1yrcuc42tw7fg0b6a/

An example case would be the rising of ice cream sales during rising temperatures.

However, does this mean "**Temperature increase caused people to buy more ice cream?**"

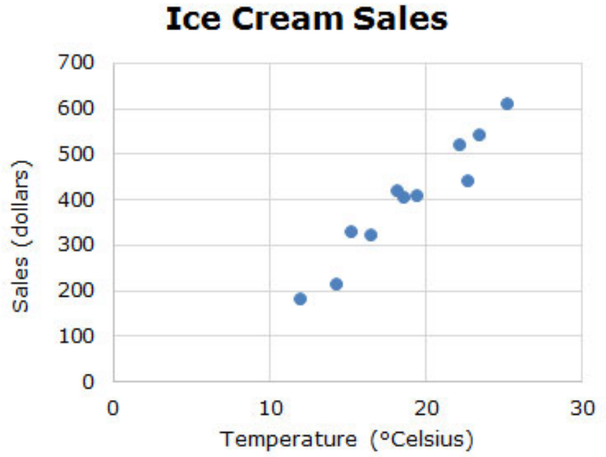


image source: https://texasgateway.org/resource/analyzing-scatterplots

CORRELATION

Correlation does not imply causation.

Example: EU Referendum (‘Brexit’) Votes in 2016 vs Mad Cow Disease in 1992

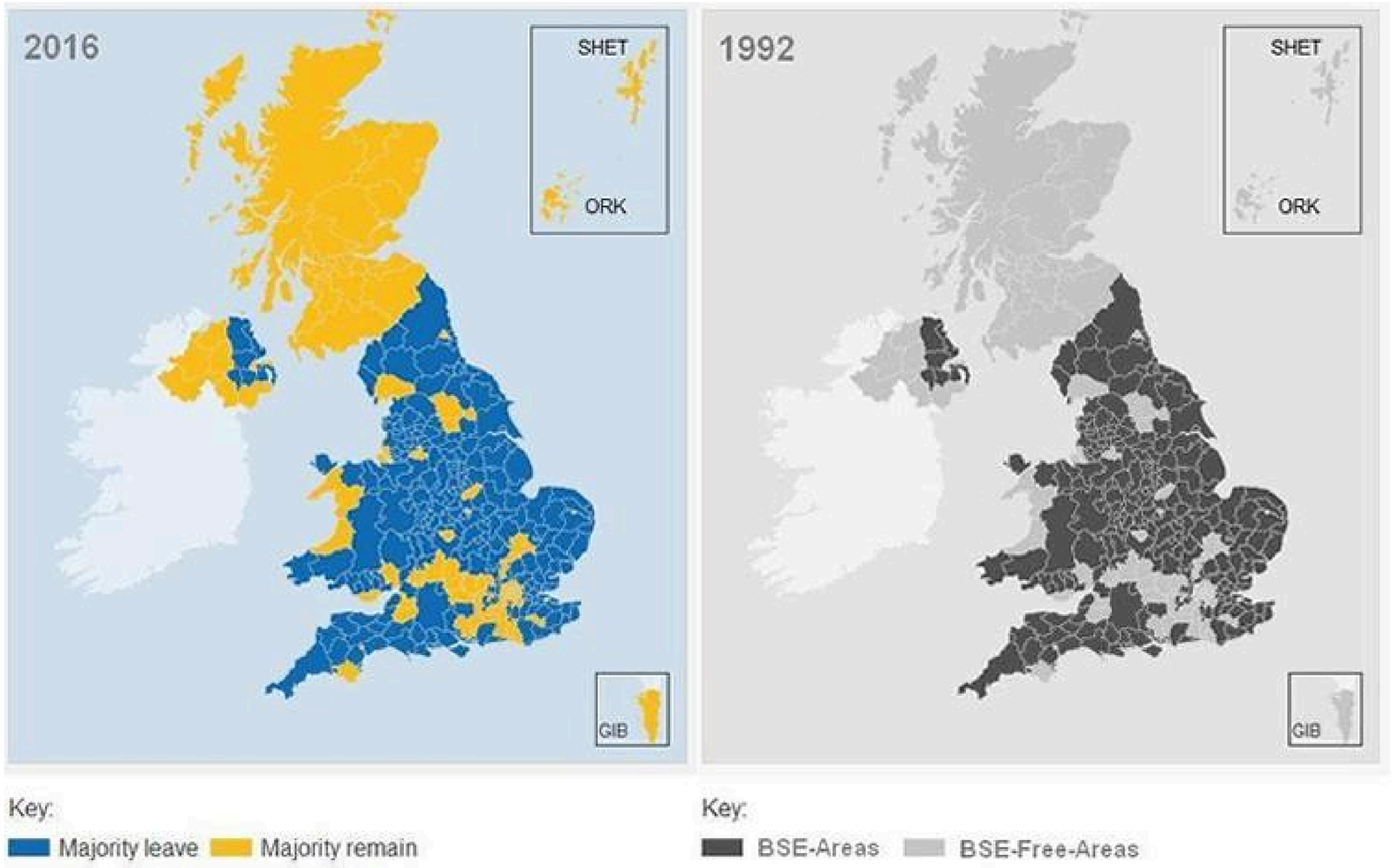


image source: https://tdwi.org/articles/2016/08/31/data-visualization-correlation.aspx

DATA VISUALISATION PRINCIPLES

ALGOSOC UOB

Why do we need data visualisation?

Imagine a **messy spreadsheet**, you're asked to give insights from the rows and columns → **visualising with charts and plots** would make our job of interpreting easier

How should we make visualisation?

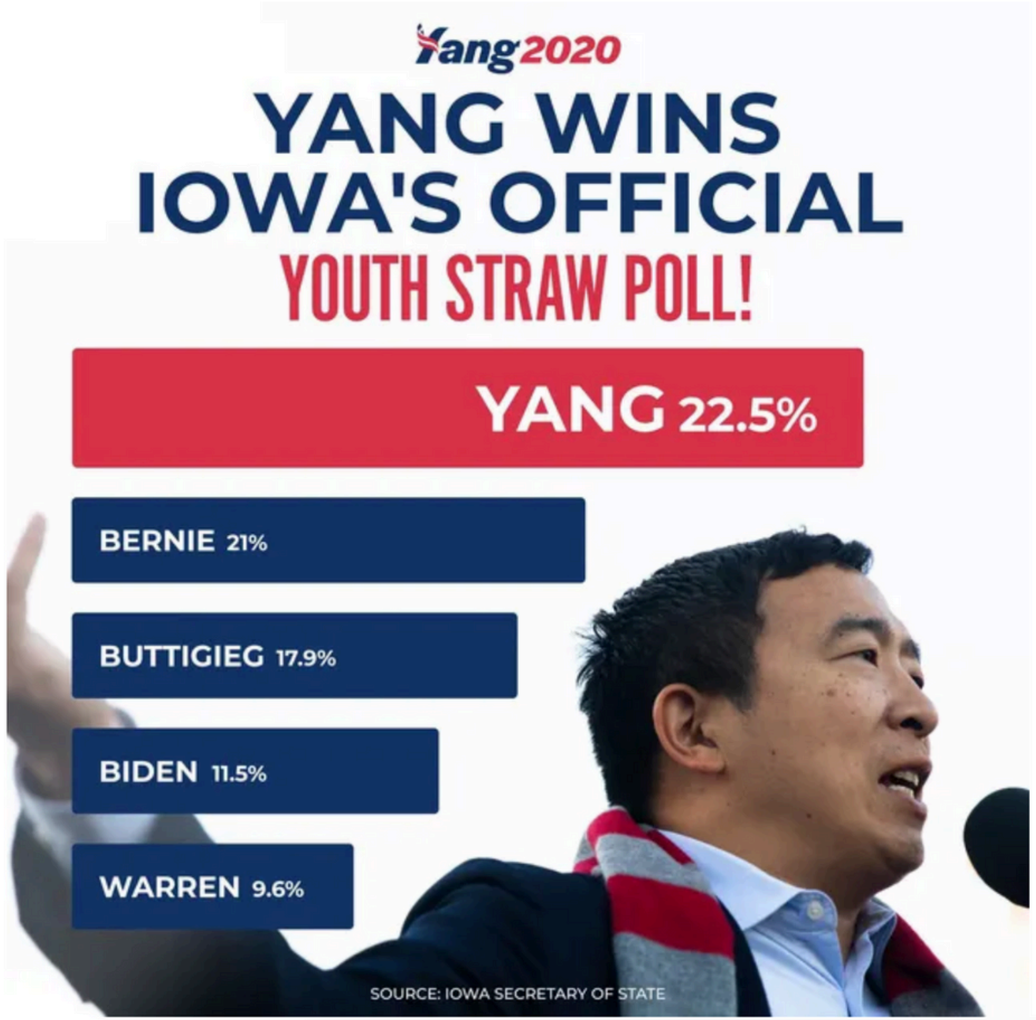
Several non-exhaustive principles that can be taken into consideration:

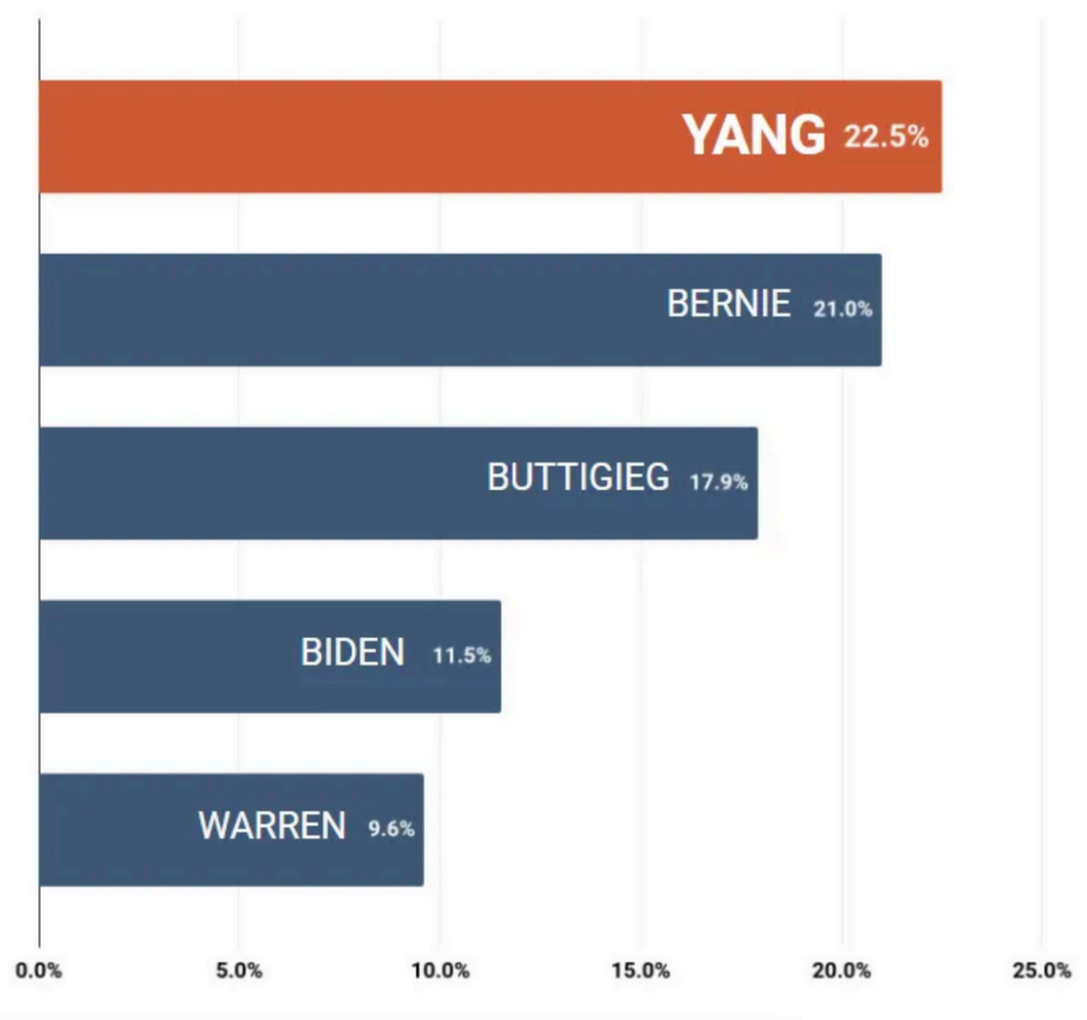
- Clarity: Know who you are talking to (make it at max 5 seconds for the audience to understand it)
- Simplicity: The less the better (make sure the visualisation is not cluttered with unwanted details)
- Accuracy: **DO NOT** present data in a biased manner
- Consistency: Make sure all of your visualisation have coherrent stylings
- Focus: Make it to the point.

note: It is easy to make visualisations, it is easier to make bad ones

DATA VISUALISATION PRINCIPLES

Bad Visualisation Examples



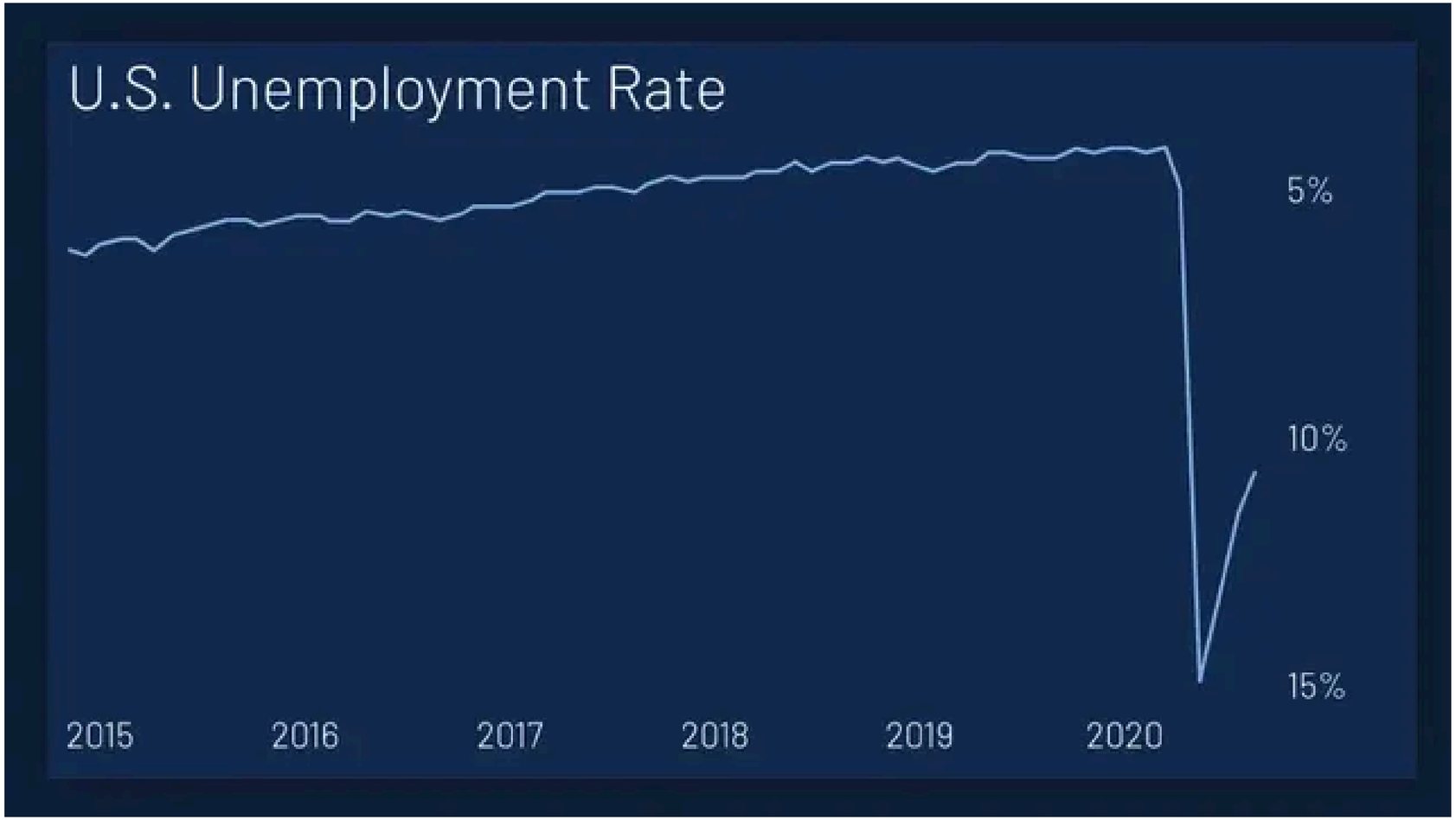


Source: https://medium.com/@thomas.ellyatt/bad-data-visualisation-real-life-examples-out-there-in-the-wild-eb5032329aeb

DATA VISUALISATION PRINCIPLES

ALGOSOC UOB

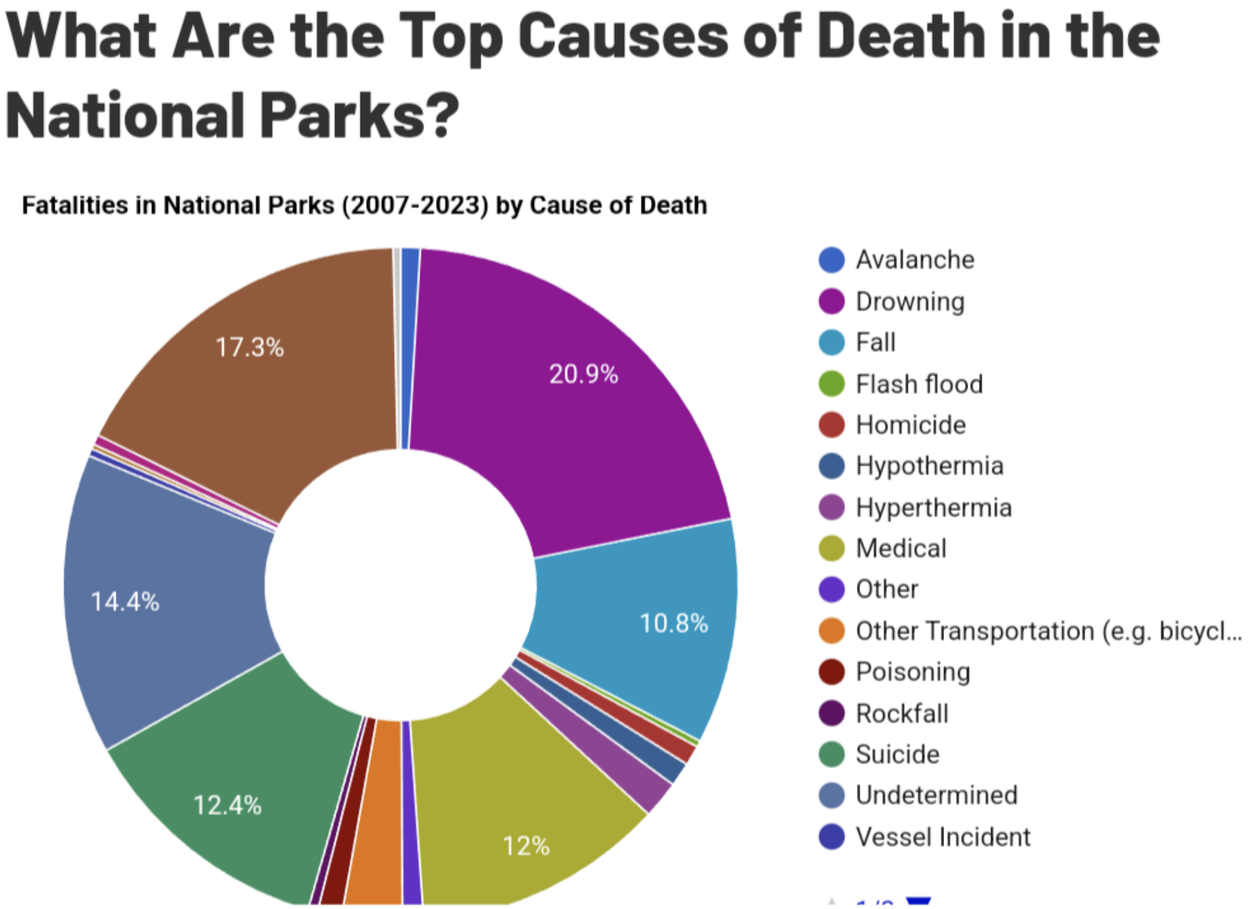
Bad Visualisation Examples



Source: https://medium.com/@thomas.ellyatt/bad-data-visualisation-real-life-examples-out-there-in-the-wild-eb5032329aeb

DATA VISUALISATION PRINCIPLES

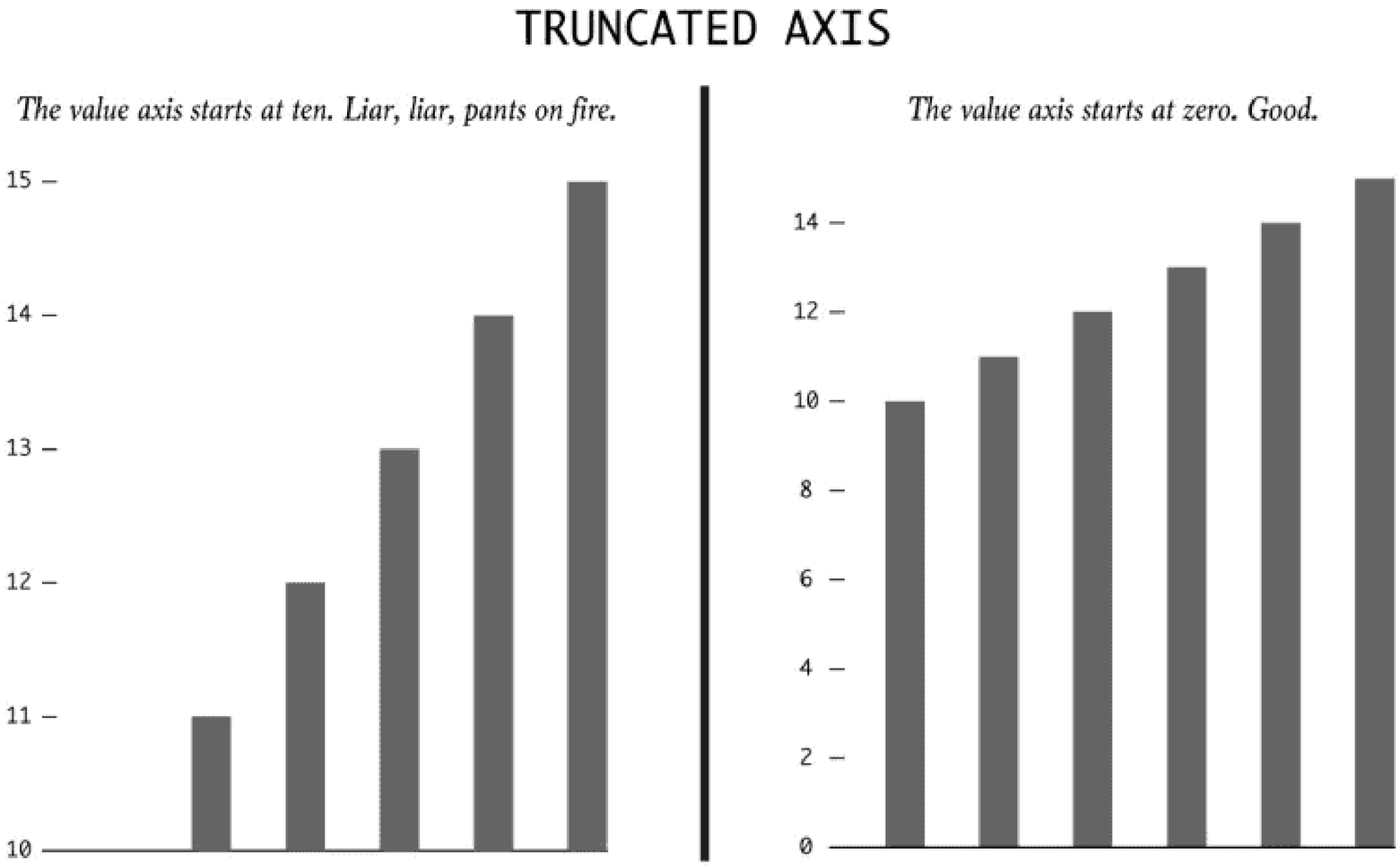
Bad Visualisation Examples



Source: https://www.luzmo.com/blog/bad-data-visualization

DATA VISUALISATION PRINCIPLES

Bad Visualisation Examples

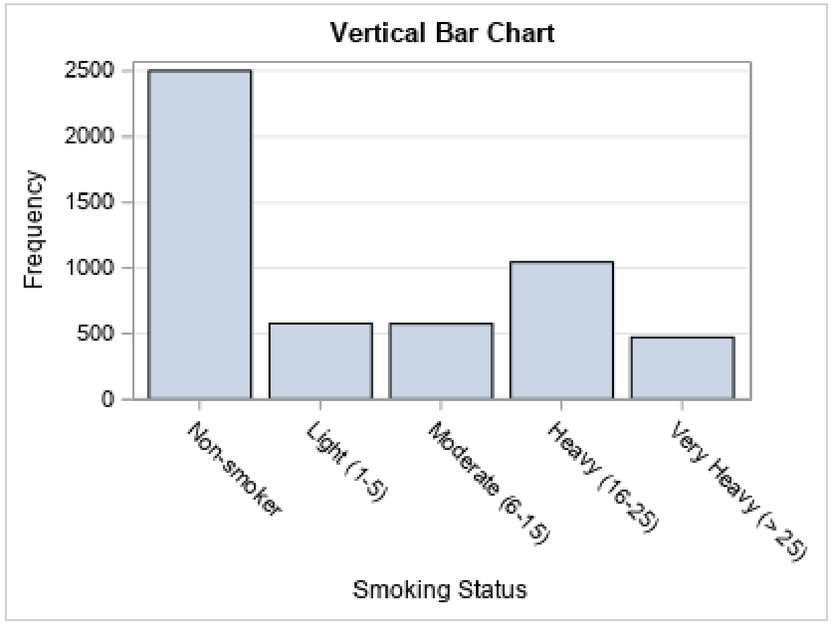


Source: https://www.tempo.io/blog/good-and-bad-data-visualization

DATA VISUALISATION TYPES

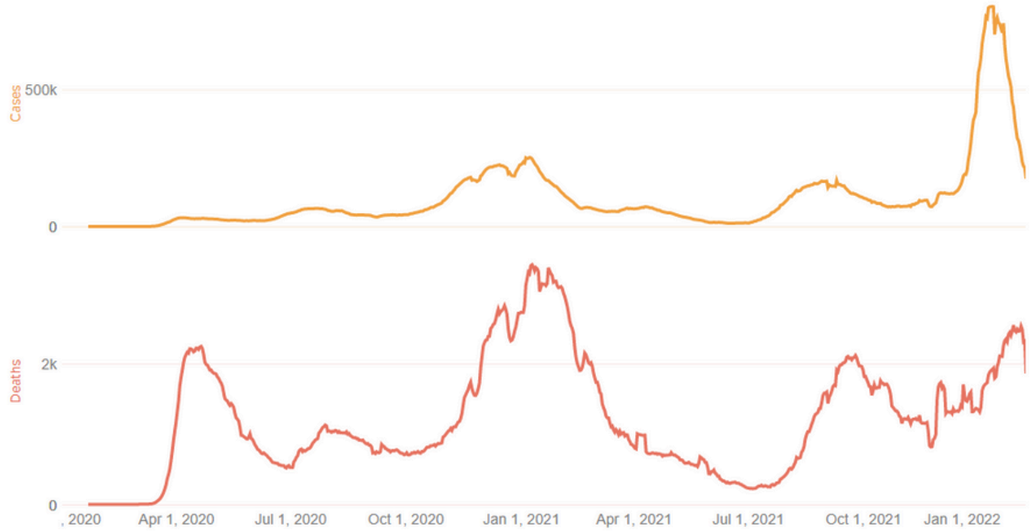
ALGOSOC UOB

Comparison: Bar Chart



Source: https://blogs.sas.com/content/iml/2021/04/12/horizontal-bar-chart.html

Trend over time: Line Chart

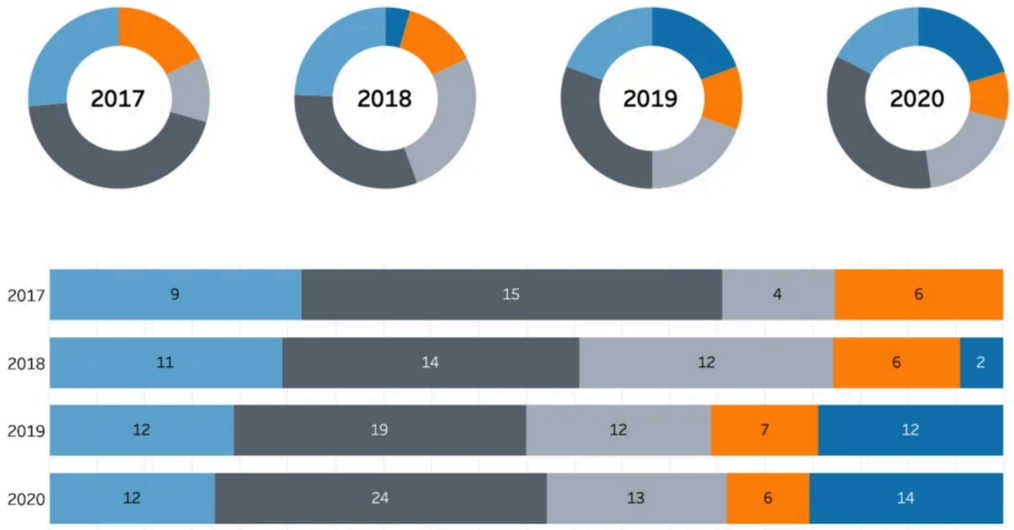


Source: https://coronavirus.jhu.edu/data

DATA VISUALISATION TYPES

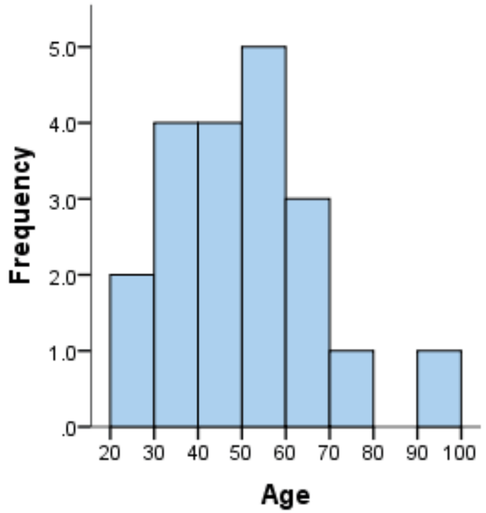
ALGOSOC UOB

Proportion: Pie or Stacked Bar Chart



Source:https://www.theinformationlab.nl/2021/09/23/part-to-whole-pie-charts-and-why-you-shouldnt/

Distribution: Histogram



Source: https://statistics.laerd.com/statistical-guides/understanding-histograms.php

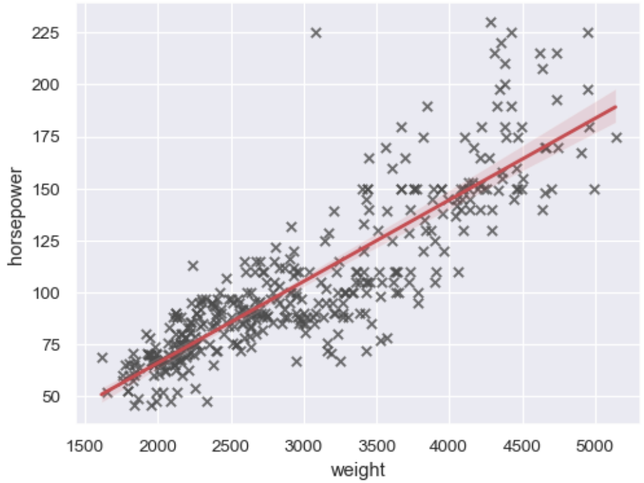
*Histogram is an example of ‘discretising continuous data’

VISUALISING CORRELATION

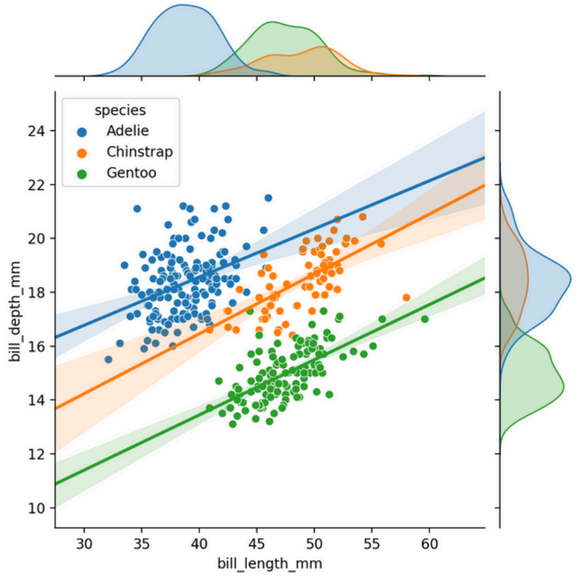
We can try to visualise correlation between two or more variables.

Between two variables, the easiest way is to use scatter (with regression line) chart

You can also add more variables into the plot to compare the correlation.



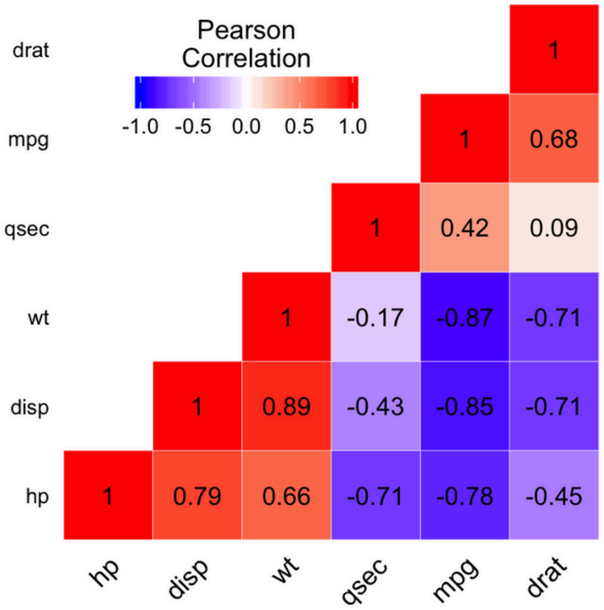
Source: https://seaborn.pydata.org/generated/seaborn.regplot.html



Source: https://stackoverflow.com/questions/64520426/possible-to-get-a-lmplot-in-a-seaborn-jointgrid-using-hue-parameter

VISUALISING CORRELATION

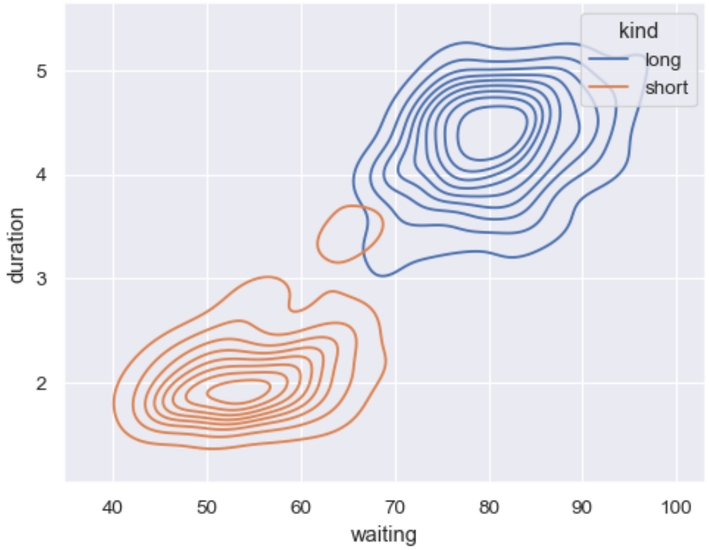
Mathematically, you can also provide a visualisation of the correlation values with a heatmap



Source: https://www.sthda.com/english/wiki/ggplot2-quick-correlation-matrix-heatmap-r-software-and-data-visualization

Another type of chart that we can use, if the data is not linearly-formed, is to use a contour plot.

You'll love this if you're a geologist



Source: https://seaborn.pydata.org/generated/seaborn.kdeplot.html

HANDS ON

ALGOSOC
UOB

We will be using Notebook

https://bit.ly/AlgoSocWK3Notebook