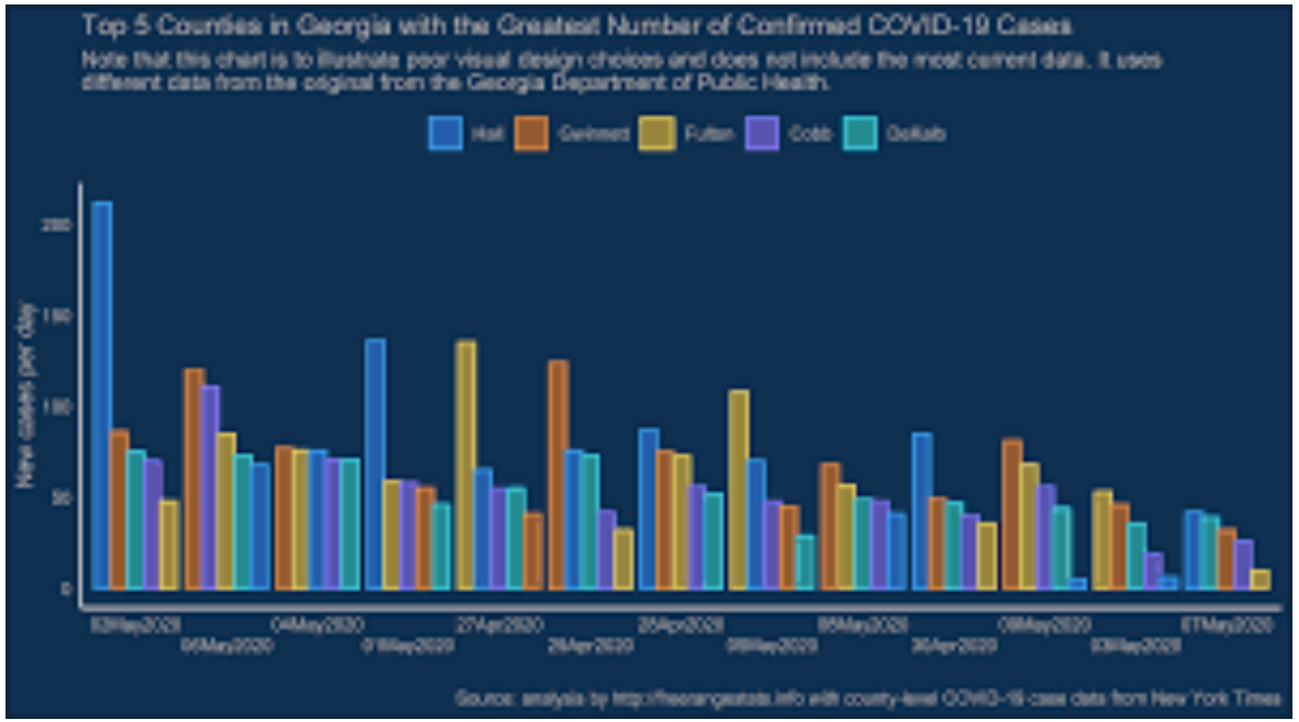
EDA and Statistical Modeling

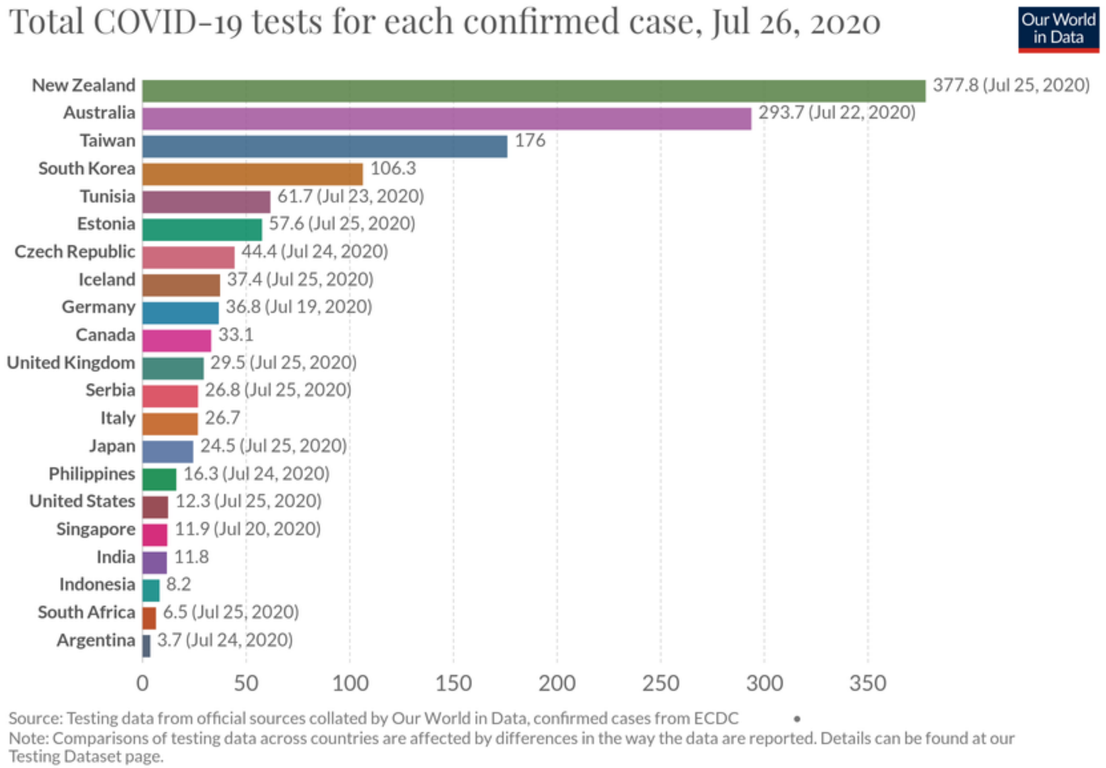
Mainak Thakur

20.08.2020

Visualisation

- Barchart:

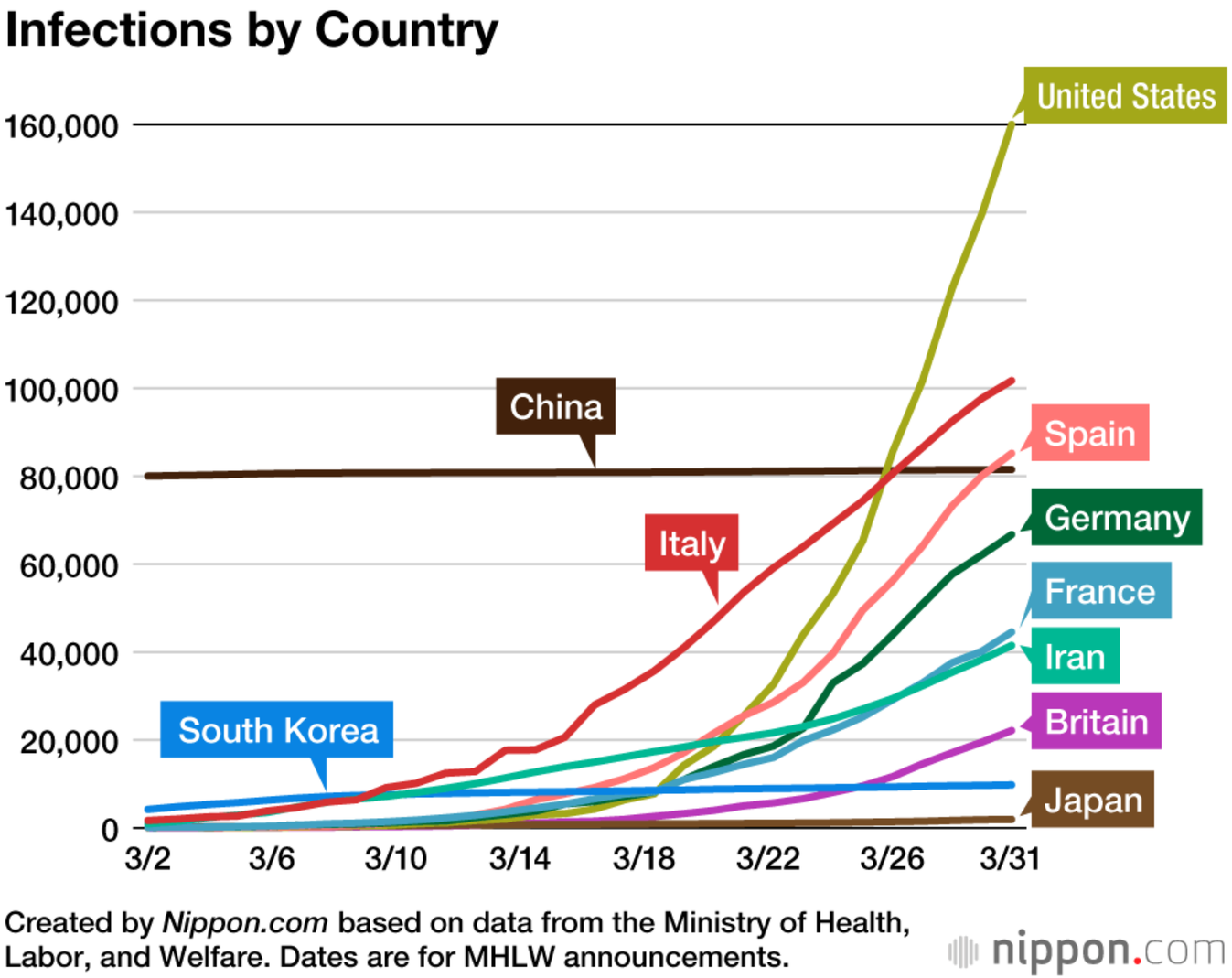




Ref: http://freerangestats.info/blog/2020/05/23/ordering-in-bar-charts
https://ourworldindata.org/grapher/number-of-covid-19-tests-per-confirmed-case-bar-chart

Visualisation

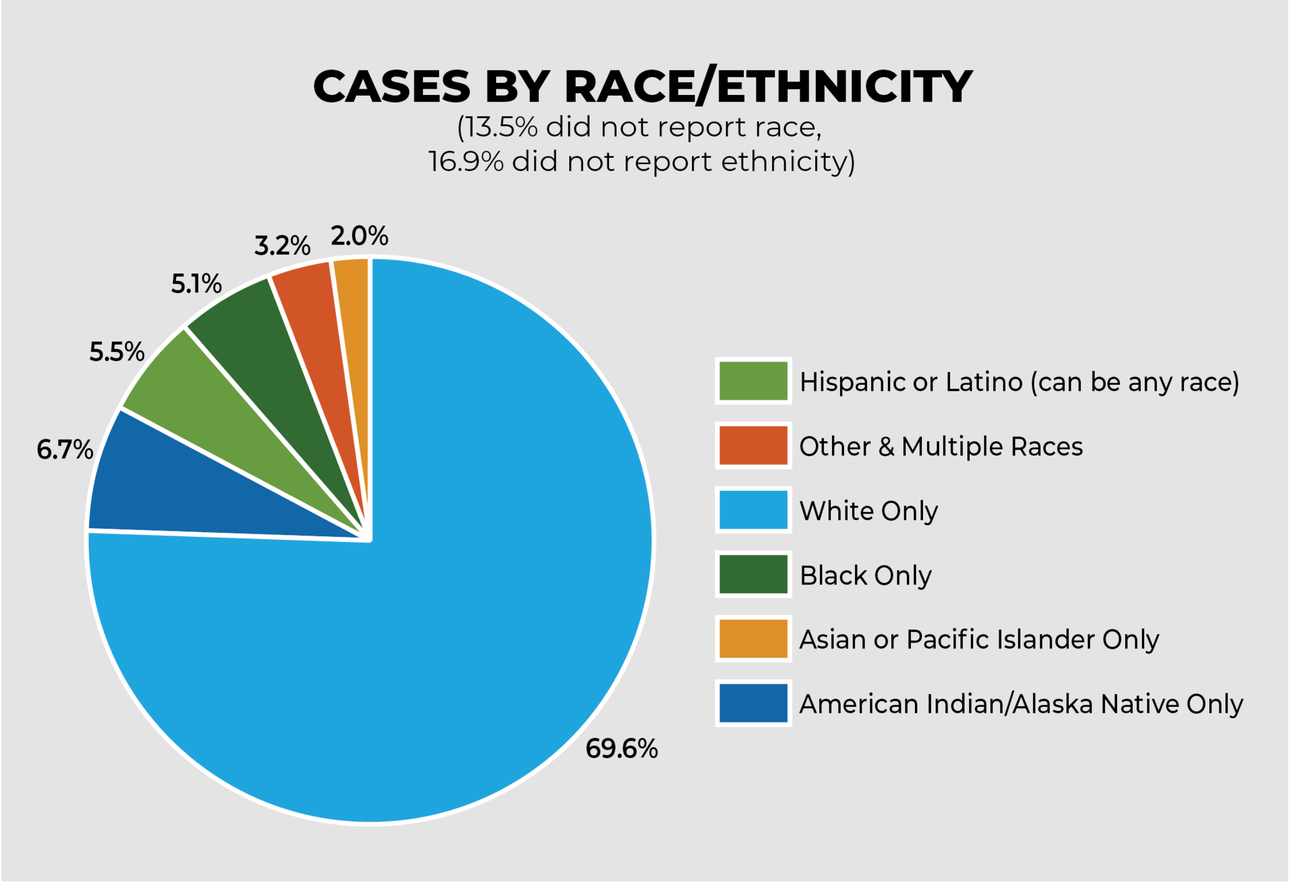
- Line Graph



Ref: https://www.nippon.com/en/japan-data/h00673/

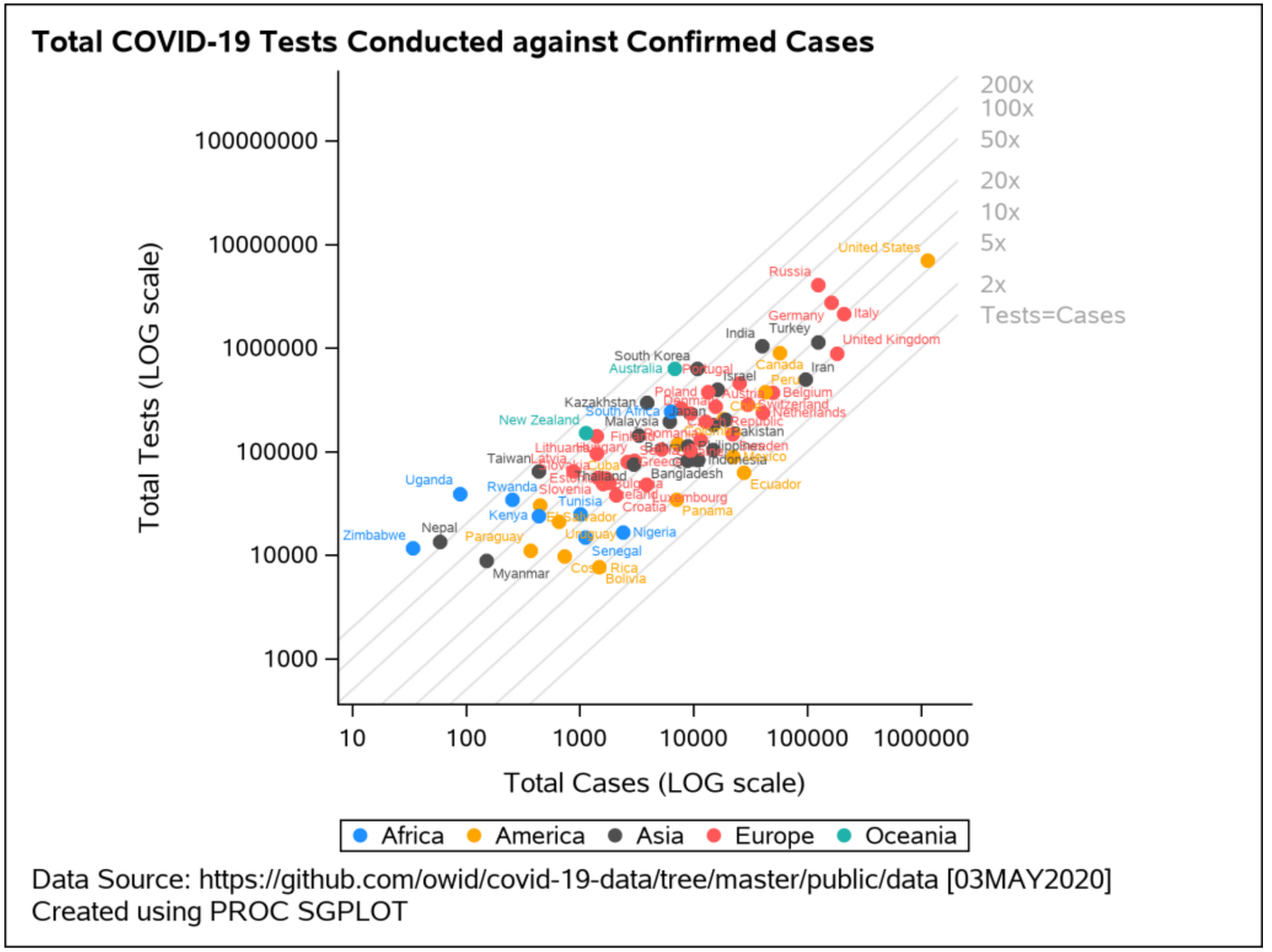
Visualisation

- Pie Chart



Ref: https://coronavirus.health.ok.gov/files/pie-chart-cases-race-ethenicity1png

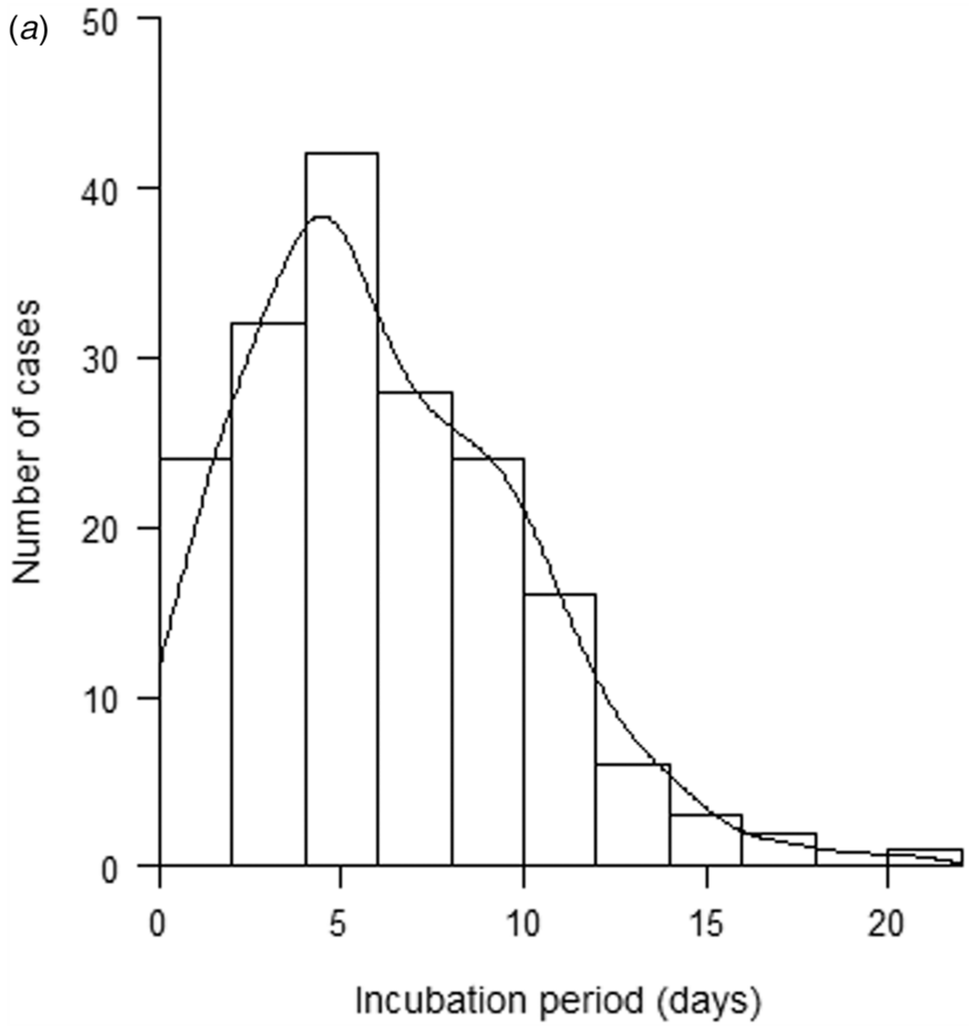
Scatter Plot



Data Source: https://github.com/owid/covid-19-data/tree/master/public/data [03MAY2020]
Created using PROC SGPLOT

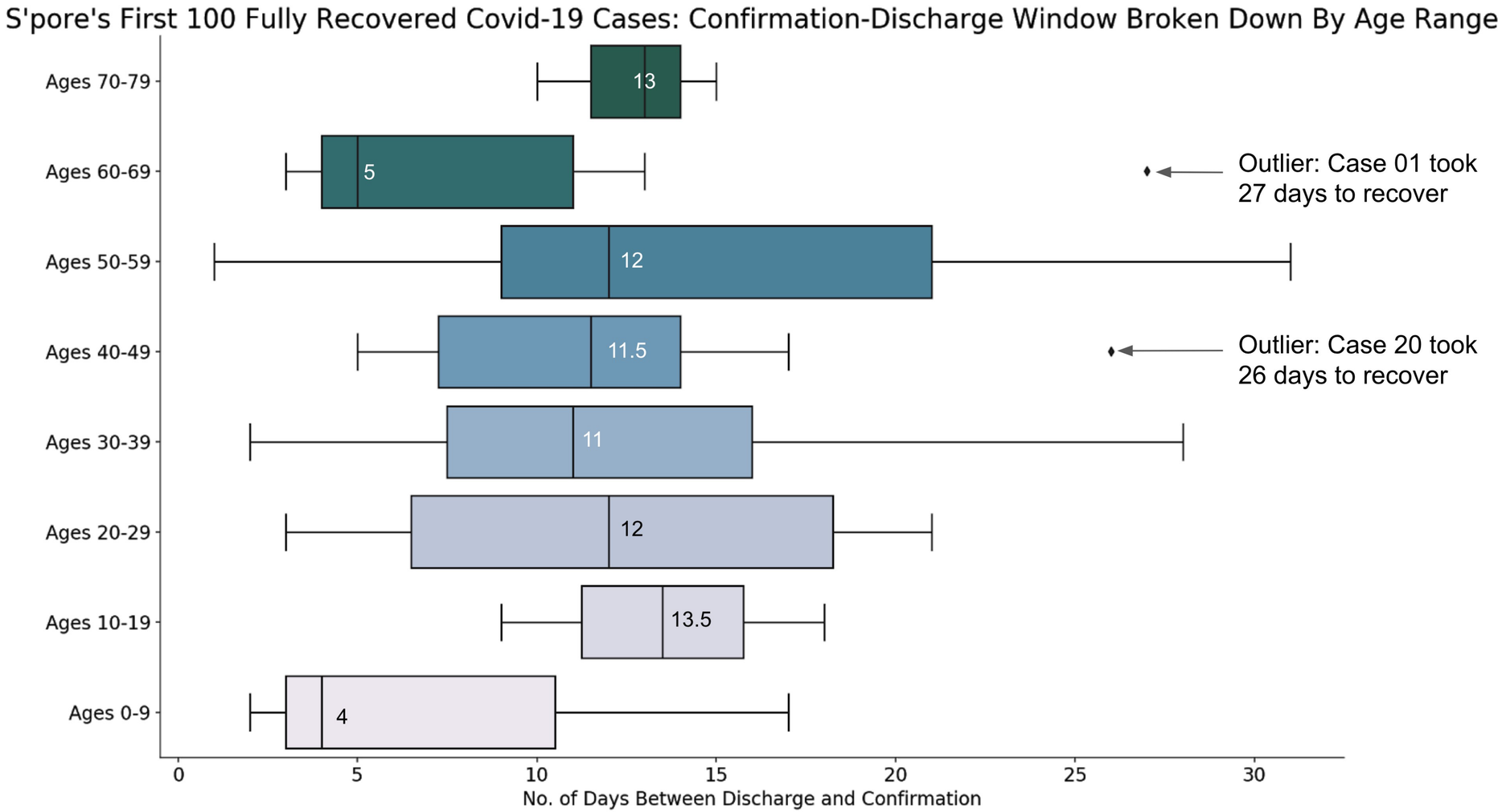
Ref: https://blogs.sas.com/content/graphicallyspeaking/2020/05/08/visualizing-covid-19-data-using-sgplot-and-sgpanel/

Histogram



Ref: Lin Yang et. al. 2020

Box Plot



Ref: Chua Chin Hon 2020

Summarisation

- Mean: $\bar{x}$ and μ; sample mean and population mean

$$\bar{x} := \frac{\sum_{i=1}^{n} x_i}{n}.$$

- Median: Data values ordered as $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$,

If n is odd, median= $x_{(n+1)/2}$

If n is even, median= $\frac{x_{n/2} + x_{n/2+1}}{2}$

Ref: https://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/lecture-notes/MIT15_075JF11_chpt04.pdf

Summarisation

- Mode: generally considered as the value that appears most often in the data set
- Quantiles: “In statistics and probability, quantiles are cut points dividing the range of a probability distribution into continuous intervals with equal probabilities, or dividing the observations in a sample in the same way. There is one fewer quantile than the number of groups created. Thus quartiles are the three cut points that will divide a dataset into four equal-sized groups. Common quantiles have special names: for instance quartile, decile (creating 10 groups).”
- “q-**quantiles** are values that partition a finite set of values into q subsets of (nearly) equal sizes.”
- Inter-quartile Range: A measure of dispersion. A quartile is a type of quantile which divides the number of data points into four more or less equal parts, or quarters. IQR = $Q_3 - Q_1$ (third quartile - first quartile).

Ref: Wikipedia

Summarisation

- Variance:

$$s^2 = \frac{1}{\underbrace{n-1}_{\text{see why later}}} \sum_{i=1}^{n} (x_i - \bar{x})^2.$$

- Coefficient of variation: standard deviation per unit mean, $s/\bar{x}$. Provides idea about degree of variation with respect to mean.

Ref: https://ocw.mit.edu/courses/sloan-school-of-management/15-075j-statistical-thinking-and-data-analysis-fall-2011/lecture-notes/MIT15_075JF11_chpt04.pdf
Wikipedia

Why Visualization is important?

Anscombe's Quartet

- Four groups each containing 11 x, y pairs

x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Ref: https://www.youtube.com/watch?v=_0Y7X4RGm68

Why Visualization is important?

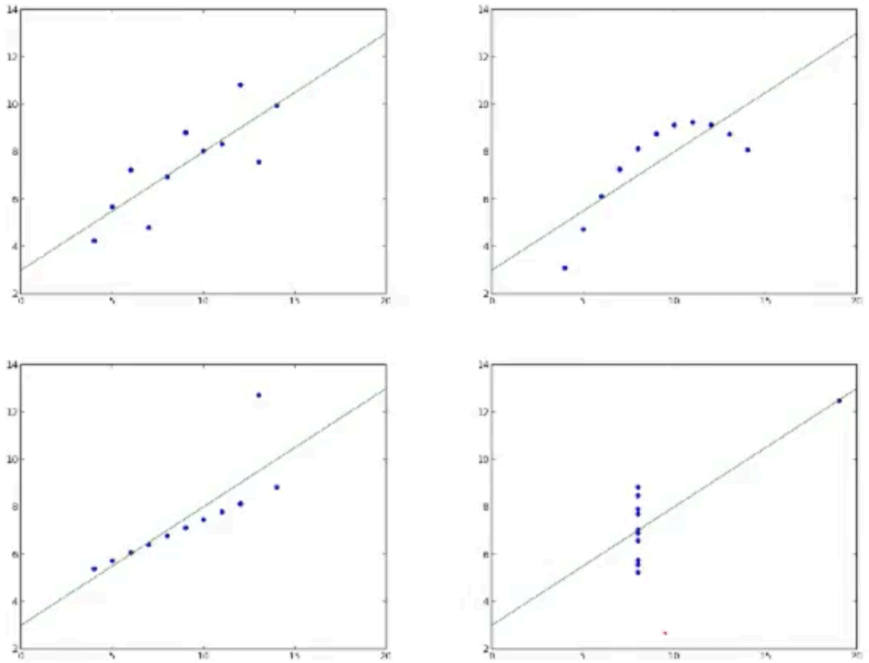
Summary Statistics

- Summary statistics for groups identical
  - Mean x = 9.0
  - Mean y = 7.5
  - Variance of x = 10.0
  - Variance of y = 3.75
  - Linear regression model: y = 0.5x + 3

Ref: https://www.youtube.com/watch?v=_0Y7X4RGm68

Why Visualization is important?

Let's Plot the Data



Moral: Statistics about the data is not the same as the data

Moral: Use visualization tools to look at the data itself

Ref: https://www.youtube.com/watch?v=_0Y7X4RGm68

Models and modeling

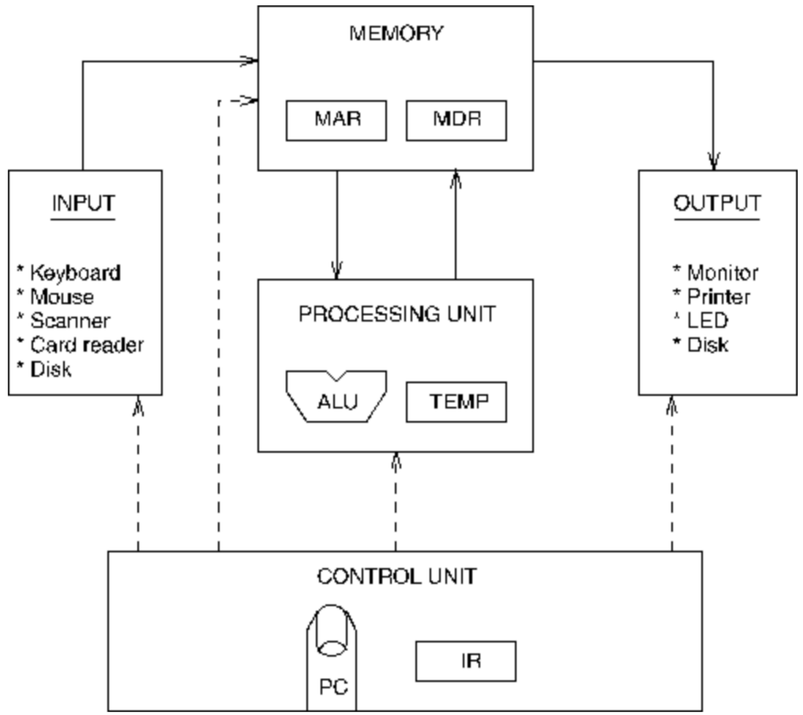
- Representation of some complicated phenomenon or process using something else which is simpler and known.
- “Generally, the process of representing a real-world object or phenomenon as a set of mathematical equations. More specifically, the term is often used to describe the process of representing 3-dimensional objects in a computer. All 3-D applications, including CAD/CAM and animation software, perform modeling.”
- “A mathematical model is a description of a system using mathematical concepts and language. The process of developing a mathematical model is termed mathematical modeling.
- Modeling and simulation is the use of models (e.g., physical, mathematical, or logical representation of a system, entity, phenomenon, or process) as a basis for simulations to develop data utilized for managerial or technical decision making.”

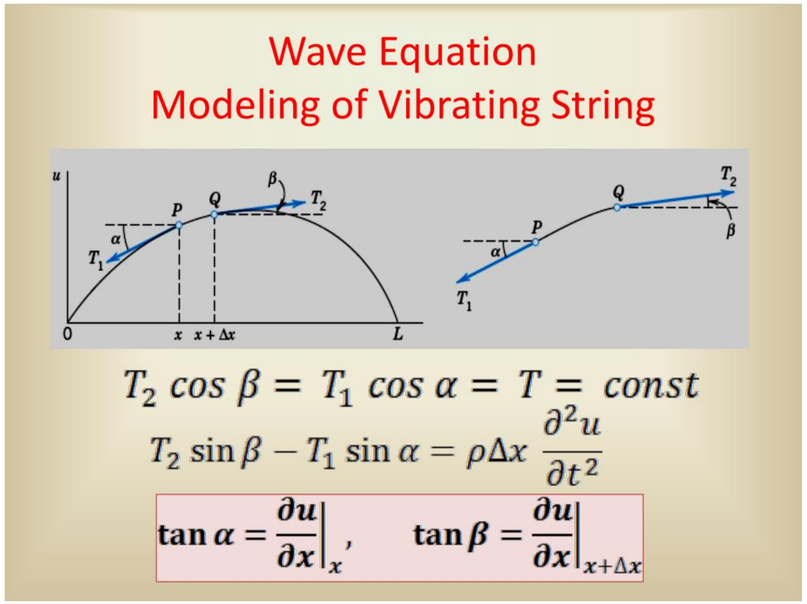
Ref: webopedia, wikipedia

What is modeling?

- Scientific Models: variety of the forms and functions
- diagrams, physical three-dimensional things, mathematical equations, computer simulations

The von Neumann Machine:





Ref: https://www.slideserve.com/lillith-walker/wave-equation-modeling-of-vibrating-string
http://gzszjmodelmaking.sell.everychina.com/p-108033213-3d-house-building-model-1-100-scale-physical-3d-model-for-sellling.html
http://www.c-jump.com/CIS77/CPU/VonNeumann/lecture.html

Why models?

- Explanation
- Prediction
- used for heuristic purposes and as a tool for theory construction
- employed to explore the implications, dynamics, or internal consistency of multiple theoretical assumptions
- To prove theories

Ref: https://www.encyclopedia.com/literature-and-arts/art-and-architecture/art-general/model-and-modeling

Deterministic vs Statistical Model

- Since statistics is a branch of mathematics, statistical models are a **subset** of mathematical models. Statistical models include randomness which is the characteristic of statistics.

- Statistical or data based models are enough flexible to change as per arrival of new data as they can incorporate new and emerging patterns and trends

- Estimation or forecast of a system based on data, extrapolation or interpolation, error estimates

Principles of Statistical Modeling

- Testing of Assumptions
- Adequacy
- Diagnostic
- Integrating
- Error Analysis
- Prediction