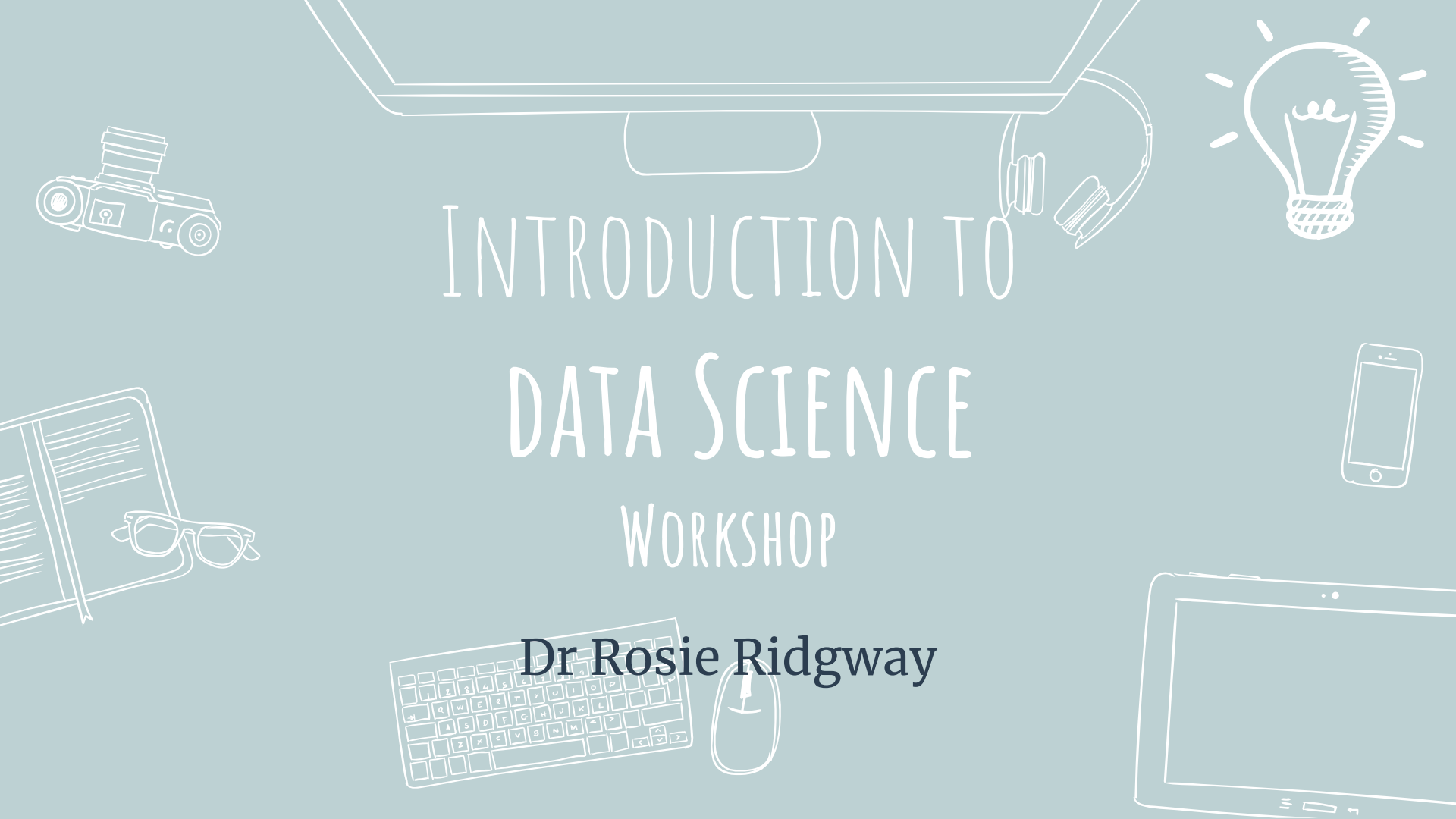


Introduction to Data Science

Workshop

Dr Rosie Ridgway

How this workshop will work

Teacher talk

Workshop is led by Rosie. There will be some direct instruction in the workshop to guide the session.

Small Group discussion

Group work will encourage discussion about ideas- the idea is to explore and understand more- we want you to participate and work together on this!

Break out Tasks

Break out tasks will be based on exploring interactive data visualisations online, you will need a good web connection and will be expected to explore these and be able to talk about what you find out.

Tutors as Facilitators

Tutors are here to help facilitate discussion, they may ask and answer questions and join in with discussions. They are there as 'guides on the side' to help point the way!

We hope you enjoy the workshop!

1.

Statistics, Data and Information

Data are the lifeblood of decision-making and the raw material for accountability. Without high-quality data providing the right information on the right things at the right time; designing, monitoring and evaluating effective policies becomes almost impossible

United Nations – mobilizing the data revolution for sustainable development (2014, p2)

What Are Data, Information and Statistics?

Data

Raw (unprocessed)

Disorganised

Multi-source

Multi-format

Information

Processed

Organised (structured)

Contextualised

Can still be multi- source but structured

Can still be multiformat (as long as we can interpret it)

Information

Data that is structured so that we can use it:
to make decisions, predictions, model and explain things

Census Data

Raw data:

Extensive,

Disorganised,

Hand written forms,

Digital entries,

Survey items structured (multi-choice questions)

and unstructured (multi-choice questions to open ended responses)

Processed information

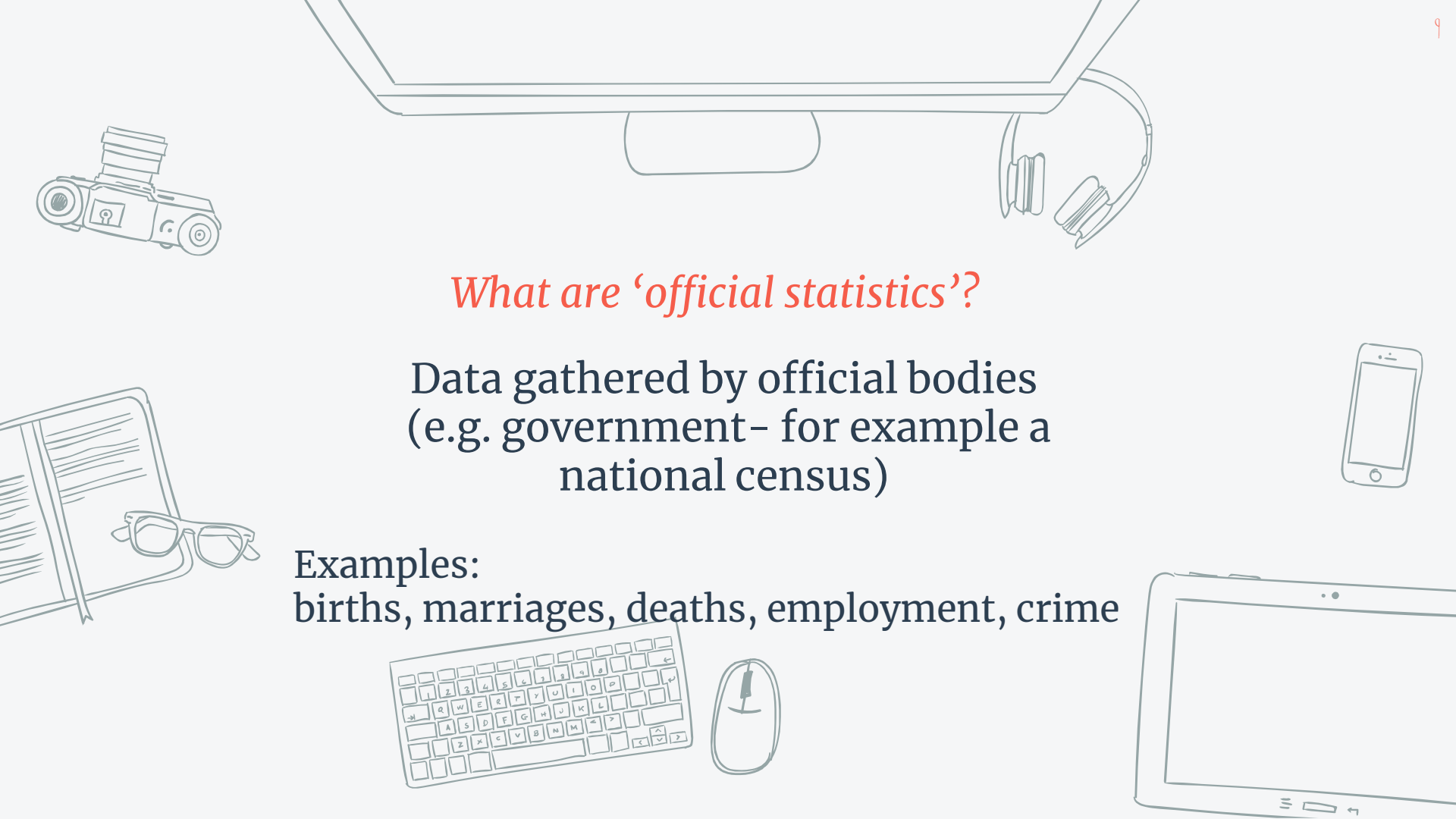
Clean: Remove inconsistencies and errors *e.g. duplicate records; incomplete entries.*

Organise: Sort data into a structured format *e.g. database where each row represents a respondent and columns represent survey questions.*

Analyse: Use statistical methods to analyse and identify patterns *e.g. response rates by age.*

Present: present information as official statistics— *e.g. average household size, employment rate, demographic distributions*

Interpret: interpret information presented to make decisions and predictions *e.g. for policy making and resource allocation*



What are 'official statistics'?

Data gathered by official bodies (e.g. government- for example a national census)

Examples:
births, marriages, deaths, employment, crime

What might make official statistics more or less reliable?

Open Data as a Form of Democracy (Citizen Empowerment)

Examples:

India Census

https://censusindia.gov.in/census.website/data/census-tables

USA Census

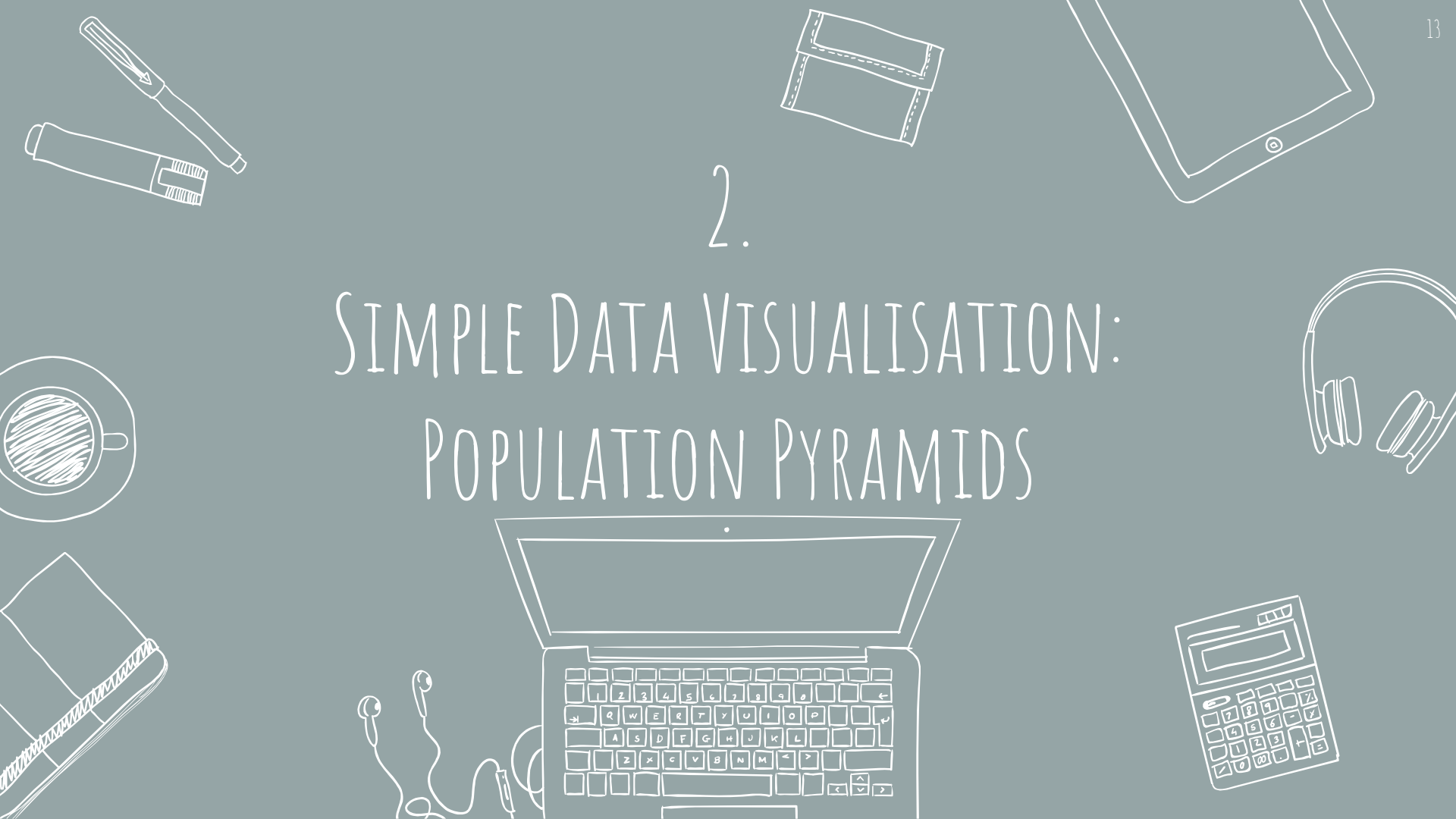
https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html

UK Census

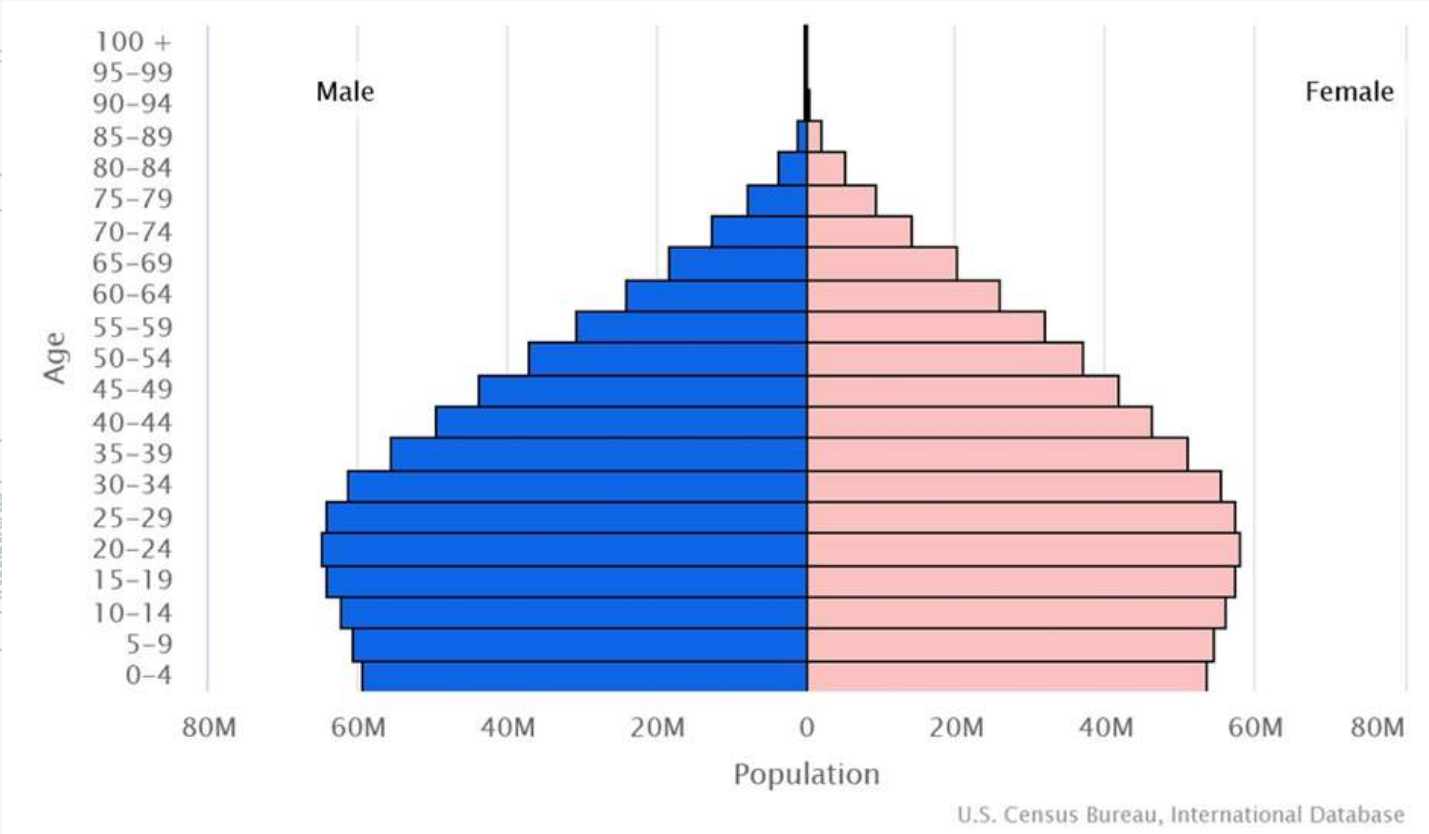
https://www.ons.gov.uk/census

2.

Simple Data Visualisation: Population Pyramids

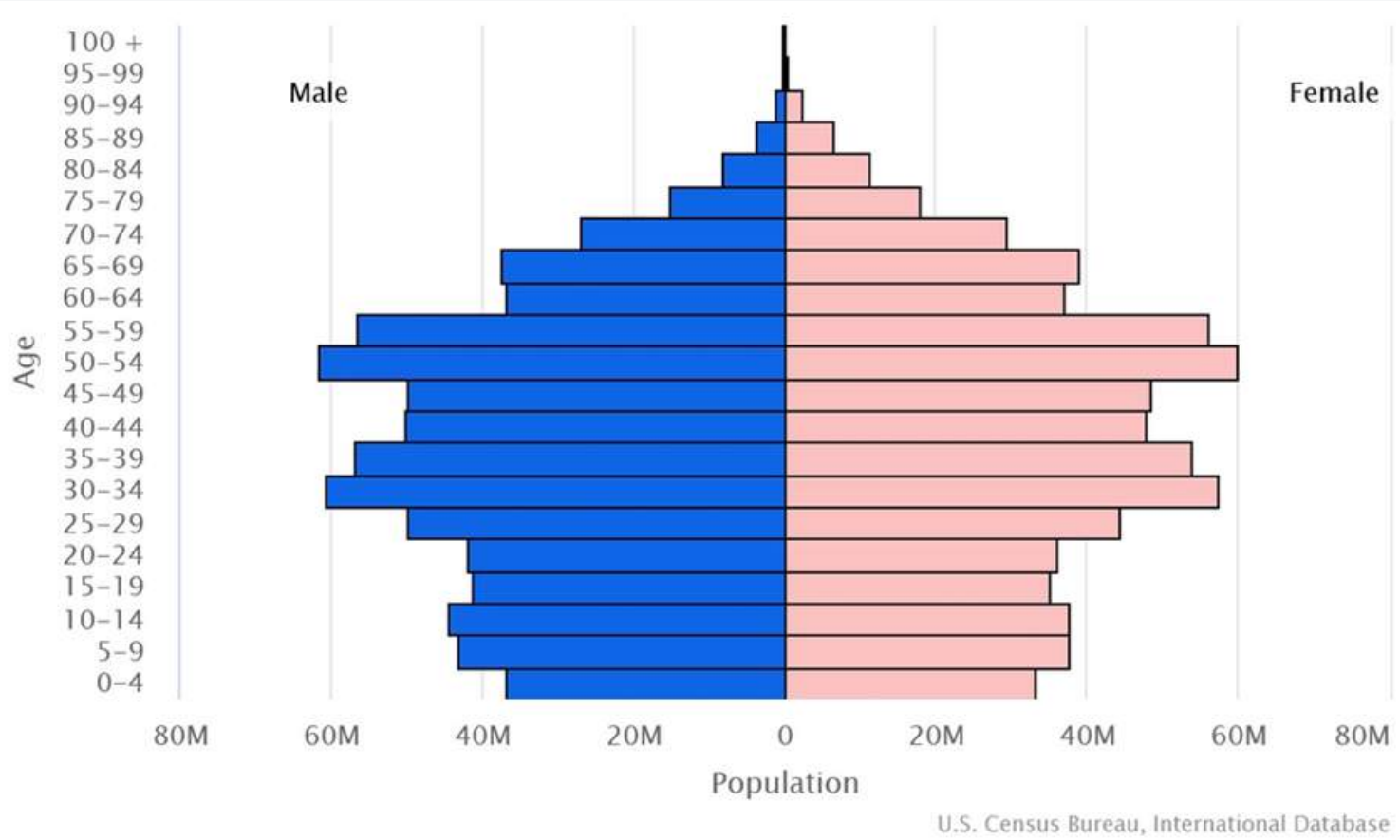


Official STATS: Population Pyramid



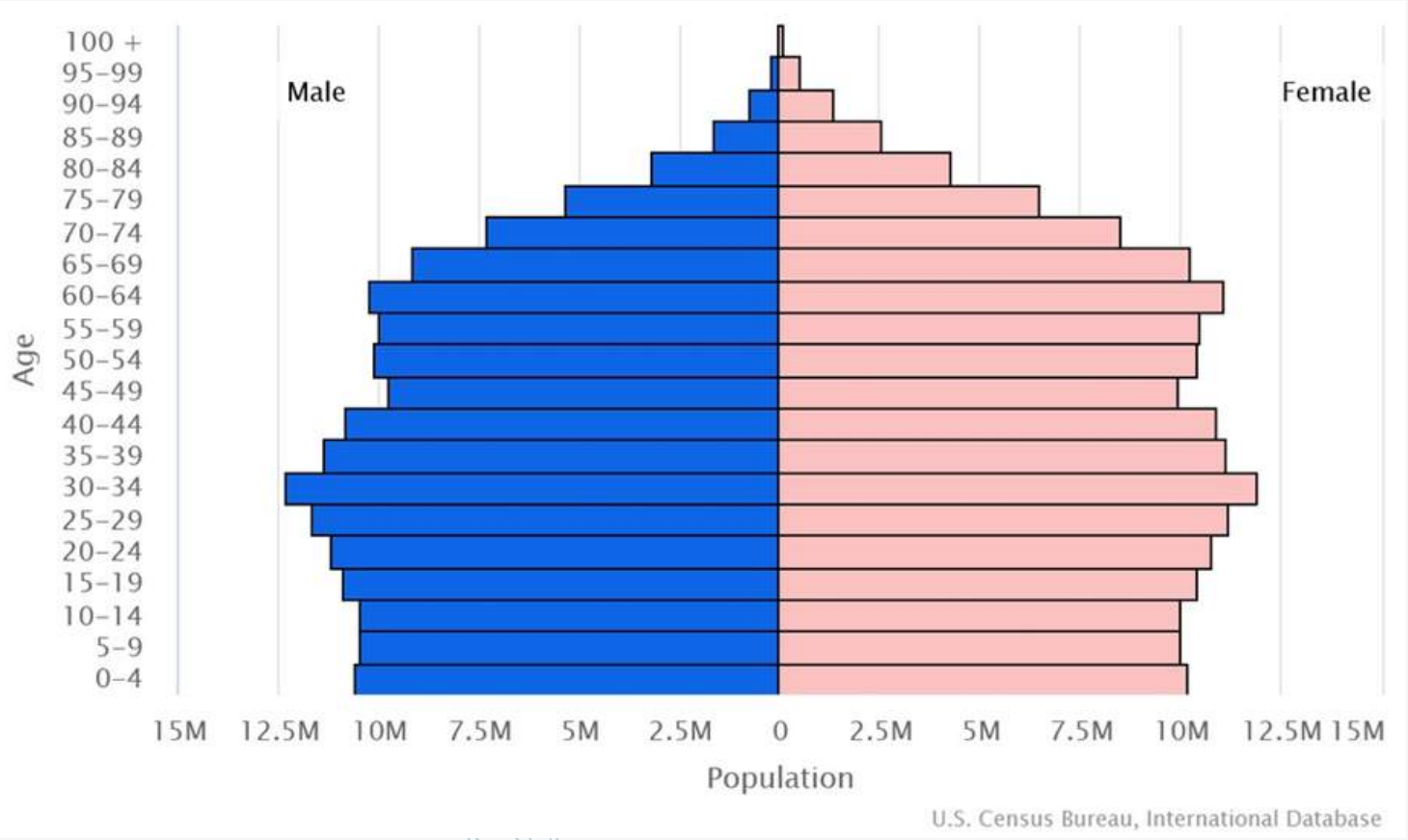
India
2023

Official STATS: Population Pyramid



China 2023

Official STATs: Population Pyramid



USA
2023

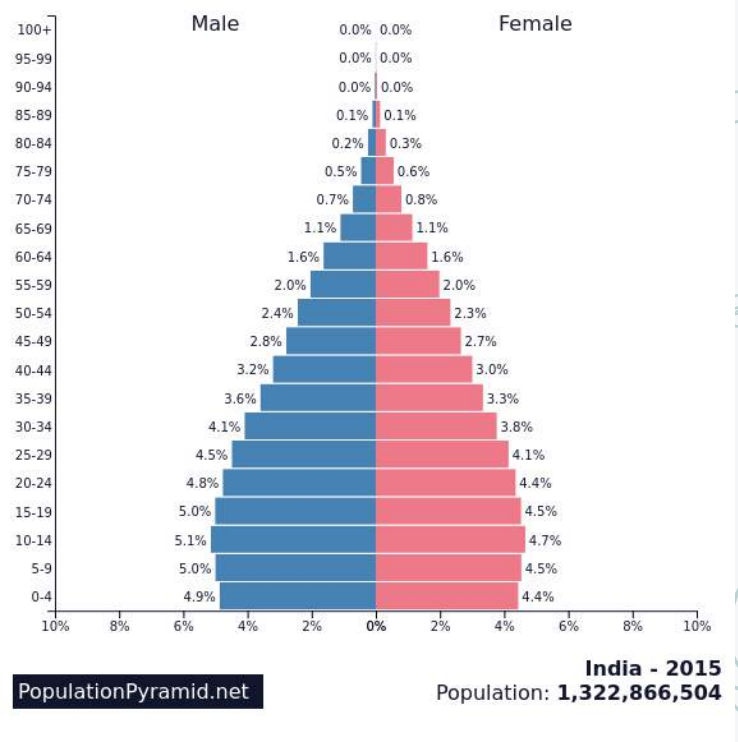
Break Out Task

Go to populationpyramid.net

Explore the tool- population pyramids over time (different years), different countries

What can we tell from this data vis?

Pick two pyramids (e.g. from different countries or years) to compare- tell a story about this with your group



What stories did We see in the population pyramids?

Demographic stories

Sex

Age

Fertility

Life expectancy

Mortality

Policy impact

War/ famine/ disease

Migration

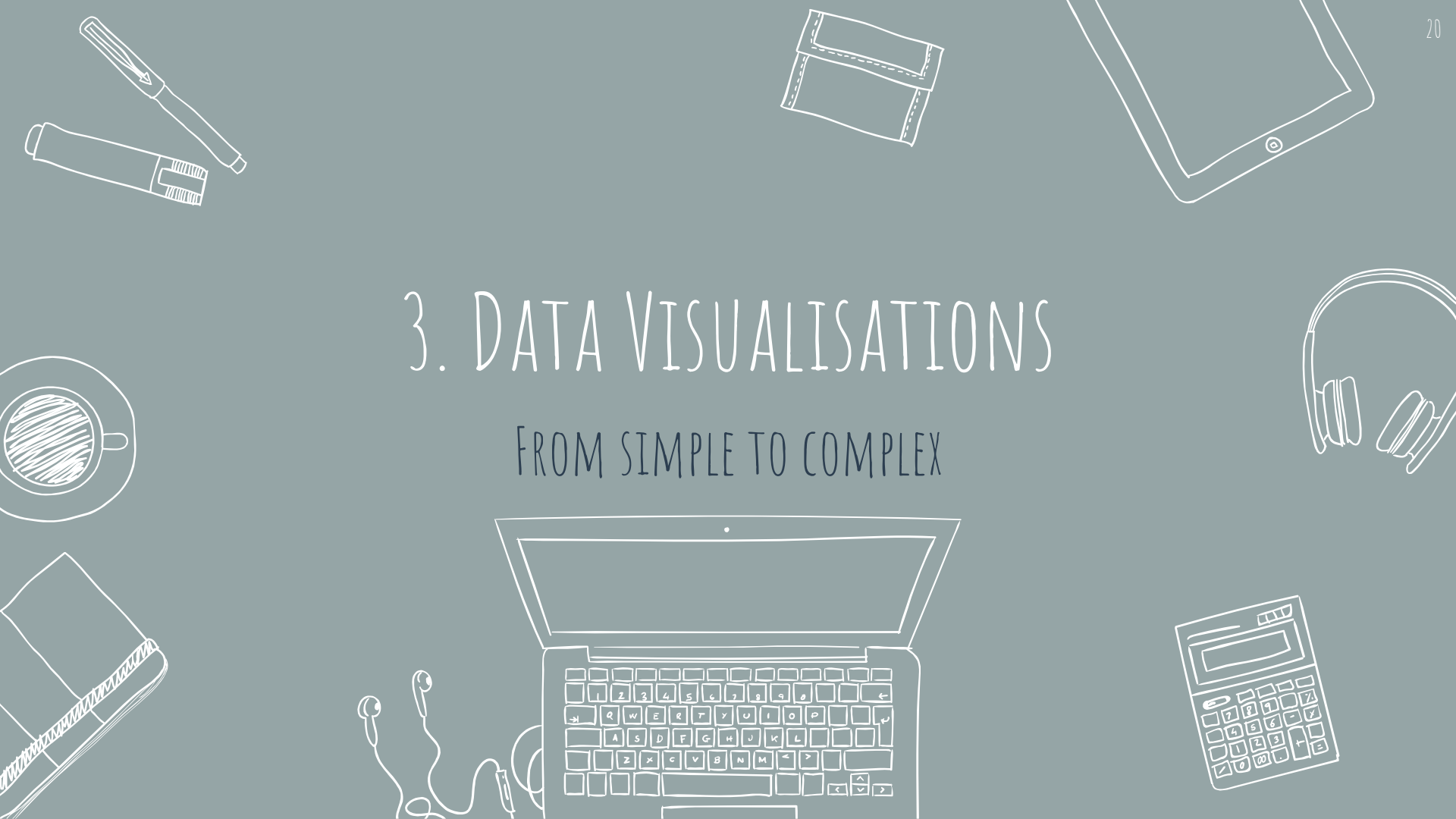
Anything else...

Data Visualisation

Present information so that it can be easily understood, and used to support interpretation

3. Data Visualisations

From simple to complex



Principles for Data Visualisation

Useful

Does it present relevant information?

Will people be interested in it?

Does it help answer or explore a question(s)?

Usable

Does the visualisation help to make sense of the data for non-experts?

Does it help make it tractable?

Intuitive

Simplicity*

Data visuals should be simple to understand for non experts

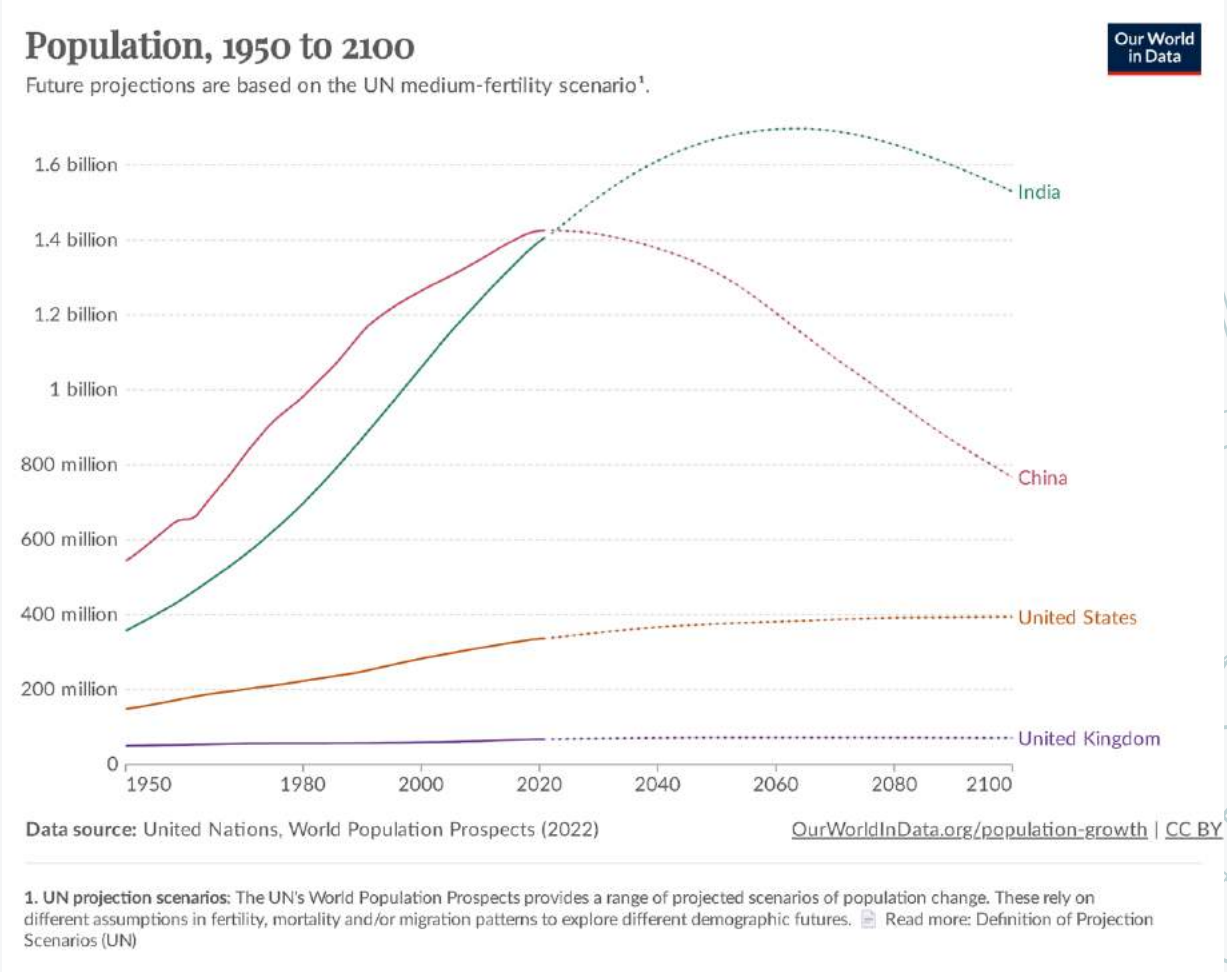
Interfaces should be simple (even if the programming and underlying data isn't)

Global Population Data/year/country

BIG dataset
multivariate
Tractable
Interpretable

Interactive
Choice making for users

Explore and engage with interface without damaging underlying data

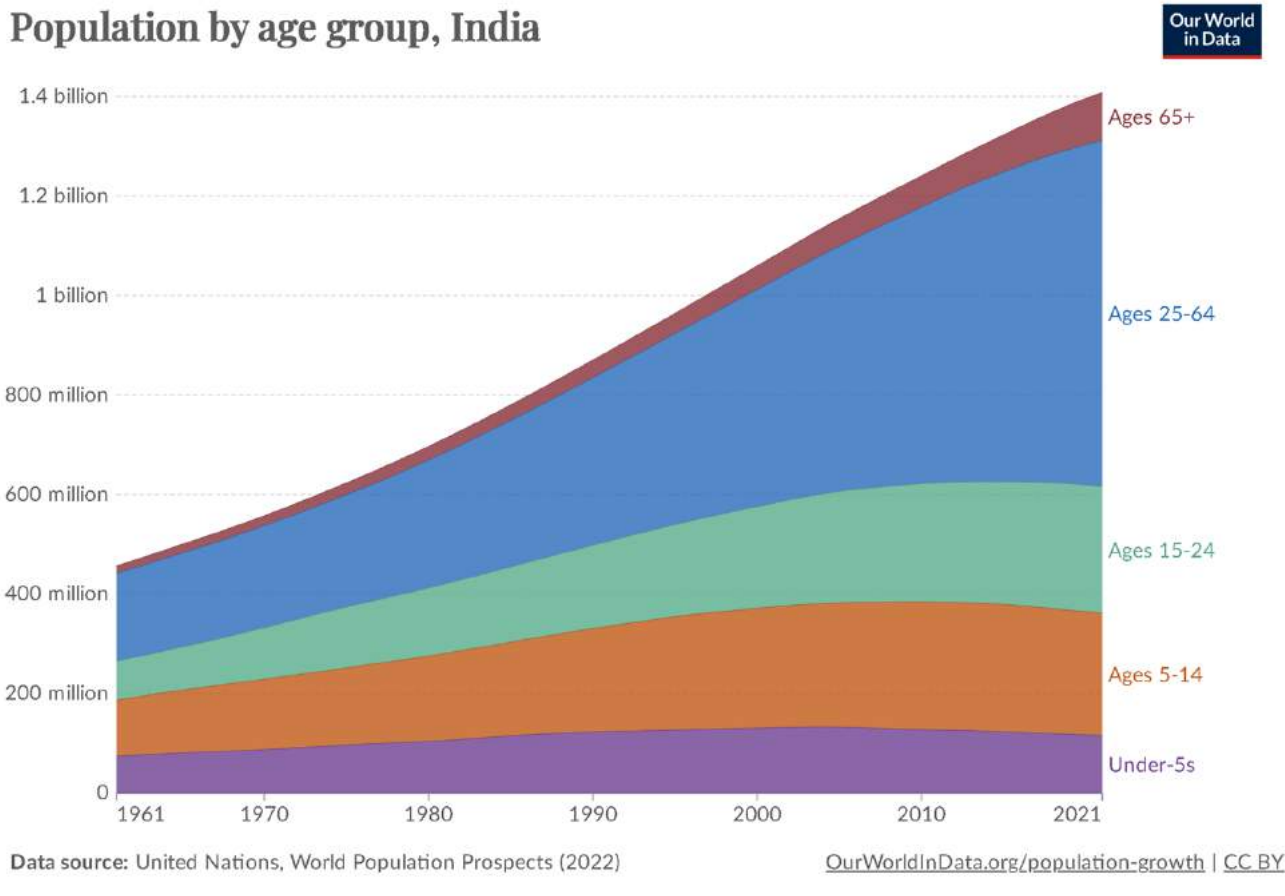


India Population data by age group/ year

BIG dataset
multivariate
Tractable
Interpretable

Interactive
Choice making for users

Explore and engage with interface without damaging underlying data



Break out Task

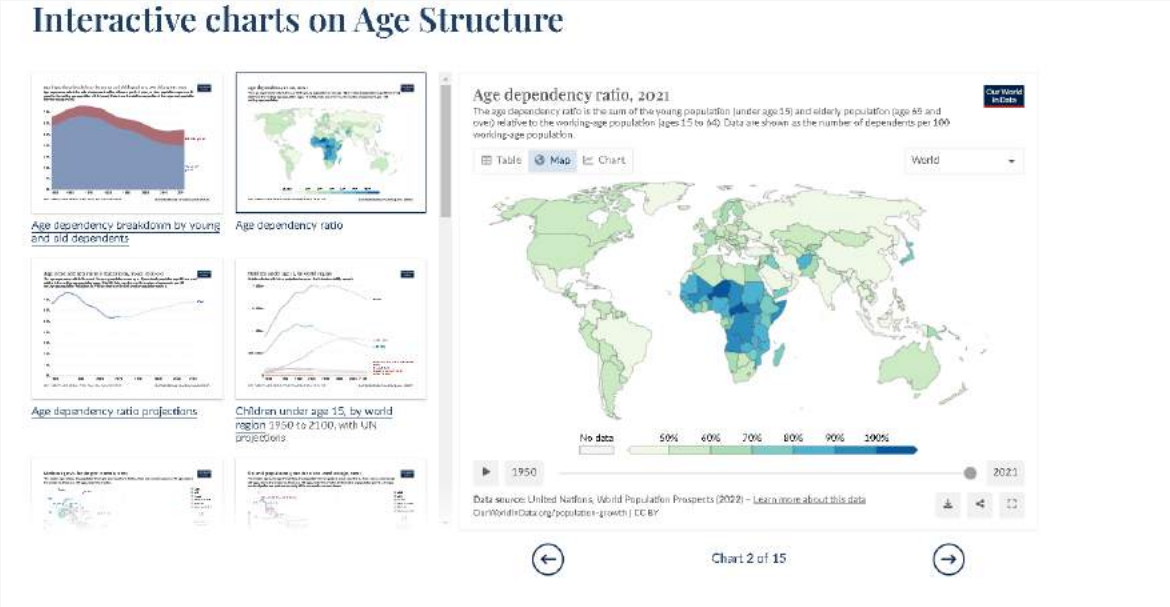
Go to our world in data

(Look at the page on population Age structure data)

There are lots of different data visualisation examples

Pick 1 of the visualisations we *haven't* talked about to explore - is it useful? usable? intuitive?

Tell a story about the data using the data visualisation



Feedback from Task

Which Visualisations did you choose?

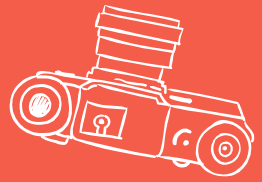
Was it useful? Was it usable? Was it intuitive?

What were the strengths and limitations of the visualisation?

Could you tell a story about the data based on the data visualisation?

What stories did you tell?

Introduction to Data Science...



Data, information, official statistics

Simple Data Visualisations

Complex multivariate data displayed visually

Open Data
Trust and Reliability

Evaluating Data Visualisations

Data Science

Converting the deluge of data into usable, useful information and presenting it so that we can tell data based stories that help us understand the world better.

Thanks!

Thank you for listening and participating in this workshop.

It is my honour to teach you

Reference list

- Engel, J., Campos, P., Nicholson, J., Ridgway, J. & Teixeira, S. (2020*) Visualizing Multivariate Data: Graphs that tell stories*. IASE Conference: New Skills in the Changing World of Statistics Education: https://iase-web.org/documents/papers/rt2020/IASE2020%20Roundtable%2022_ENGEL.pdf?1610923749
- Sutherland, S., & Ridgway, J. (2017) *Interactive visualisations and statistical literacy.* Statistics Education Research Journal , 16 (1), 26-30.
- Ridgway, J., Ridgway, R., & Nicholson, J. (2018). *Data Science for all: A stroll in the foothills*. Paper presented at the Looking back, looking forward. Proceedings of the 10th International Conference on Teaching Statistics (ICOTS 10, July 2018), Kyoto, Japan. http://icots.info/10/proceedings/pdfs/ICOTS10_3A1.pdf?1531364253
- Ridgway, R., & Ridgway, J. (2022) Ch 23:Civic Statistics in context: mapping the global evidence ecosystem. In J Ridgway (ed) Statistics for empowerment and social engagement: teaching. Civic Statistics to develop informed citizens. Springer; London.
- United Nations *(2014) A World that counts: mobilising the data revolution for sustainable development*. https://www.undatarevolution.org/

Reading List and Resources for more...

On Open Data:

Governance lab at NYU thegovlab.org open data project odimpact.org/periodic-table.html

India Census

https://censusindia.gov.in/census.website/data/census-tables

USA Census

https://www.census.gov/programs-surveys/decennial-census/decade/2020/2020-census-main.html

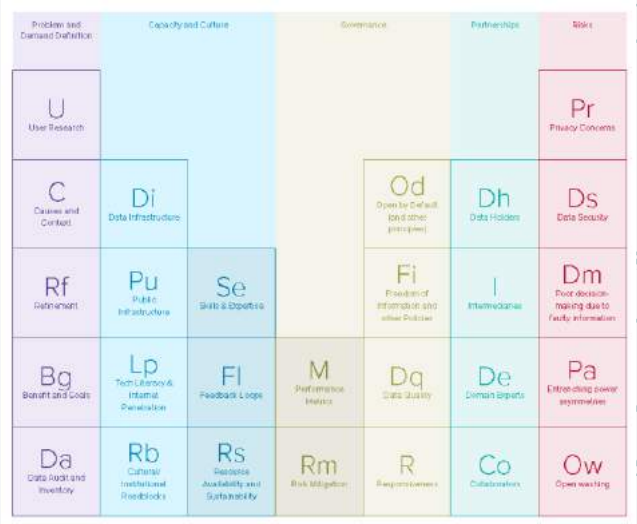
UK Census

https://www.ons.gov.uk/census

Data Visualisations in this workshop:

https://www.populationpyramid.net/india/2015/

https://ourworldindata.org/age-structure



Where can I learn to build Data Visualisations?

CODAP – extremely simple to use, intuitive – suitable for absolute beginners- Excellent start point- fun free datasets to play with

Tableau- has a free 'public' version which is free and includes some excellent tools- this is a great place for intermediate and developing skills- good demos and community

Kaggle- great tutorials and community here, you'll probably really enjoy Kaggle challenges- https://www.kaggle.com/learn

Microsoft Excel- *very widely used- has recently made improvements because it was clunky, it's not exciting to use but it's alright (expensive if you buy a Microsoft account)*

PowerBI *used a lot in industry- it's powerful, and useful for business purposes but…it's a bit boring (expensive if you buy a Microsoft account)*

☺ Just Rosie's opinion- please don't be offended!☺