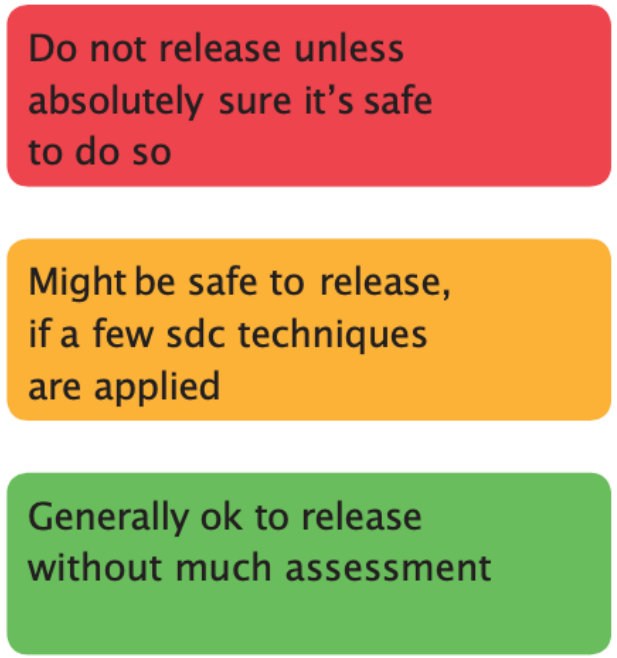
Data, Politics and Society

W7 Seminar – Safe Research with Sensitive data

Guidelines followed by output checkers and researchers

A Statistical Disclosure Control booklet for outputs is made available by the safe data access professionals working group, with guidelines to follow

A traffic light system can be followed for assessing whether an output is safe to release



Descriptive Statistics tables

Table 3:
Table of frequencies

ETHNIC BACKGROUND/AGE	16	17	18	19	20
African_Asian	0	0	0	0	0
Bangladeshi	3	4	6	5	4
Black African	3	5	7	6	8
Caribbean_West Indian	0	4	2	4	3
Chinese	0	2	1	1	0
Far Eastern	1	0	1	2	0
Indian	5	9	4	4	4
Middle Eastern	1	1	0	0	1
Mixed Caribbean_West Indian	0	0	1	0	2
Mixed Indian	0	0	0	0	0
North African	1	0	1	0	1
Pakistani	6	10	5	4	3
Sri Lankan	0	0	2	0	0
Turkish	1	1	1	0	1
White	54	135	141	146	130

Consider grouping columns or rows?

Understandable labels

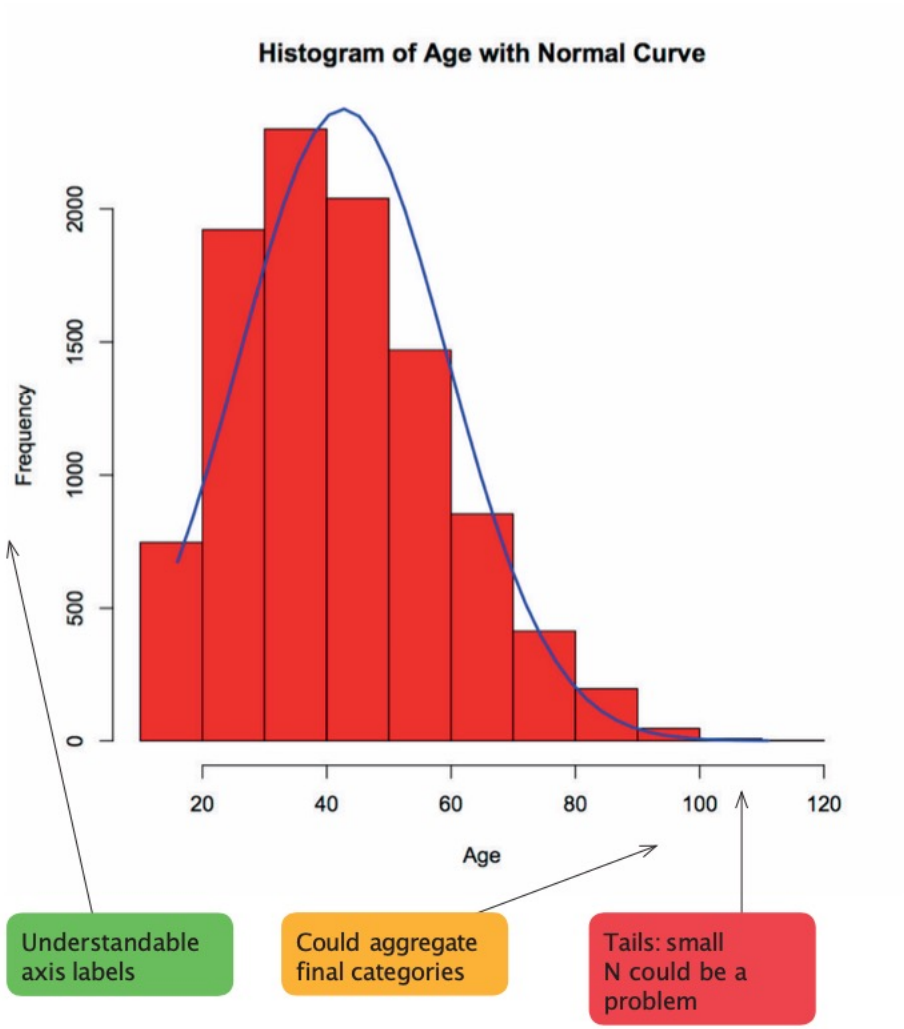
Small cell frequencies

Practices to reduce disclosure:

- Banding/combining columns or rows
- Averaging values e.g. for 10 observations in a bracket
- Rounding values
- Suppressing cells (be mindful that usually two cells in a row/ column should be suppressed, otherwise the original values could be deduced).

Histograms

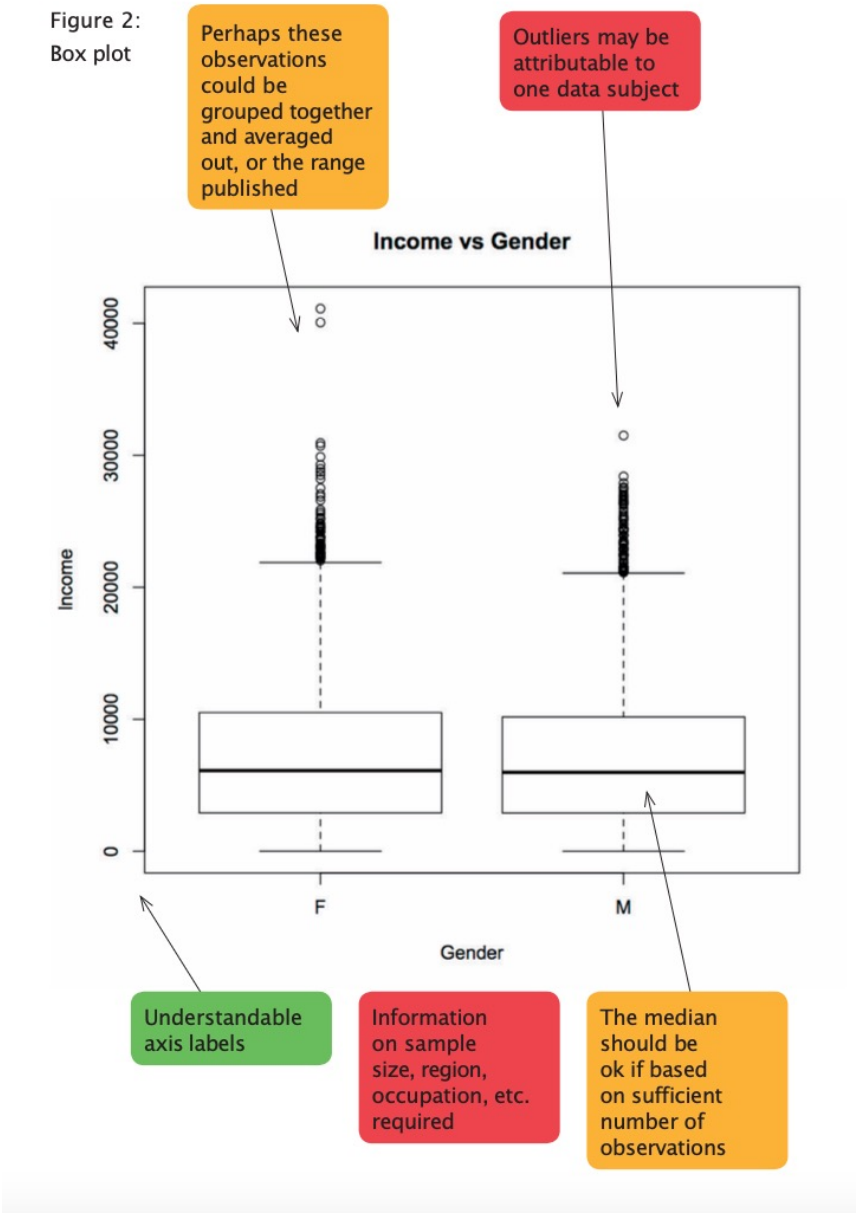
Figure 1:
Histogram



Practices to reduce disclosure:

- If the graph is intended to show that the distribution has a long tail (i.e. there are many outliers) then analysts should cap all these values in one class.
- This approach can mask the maximum or minimum values.

Boxplots

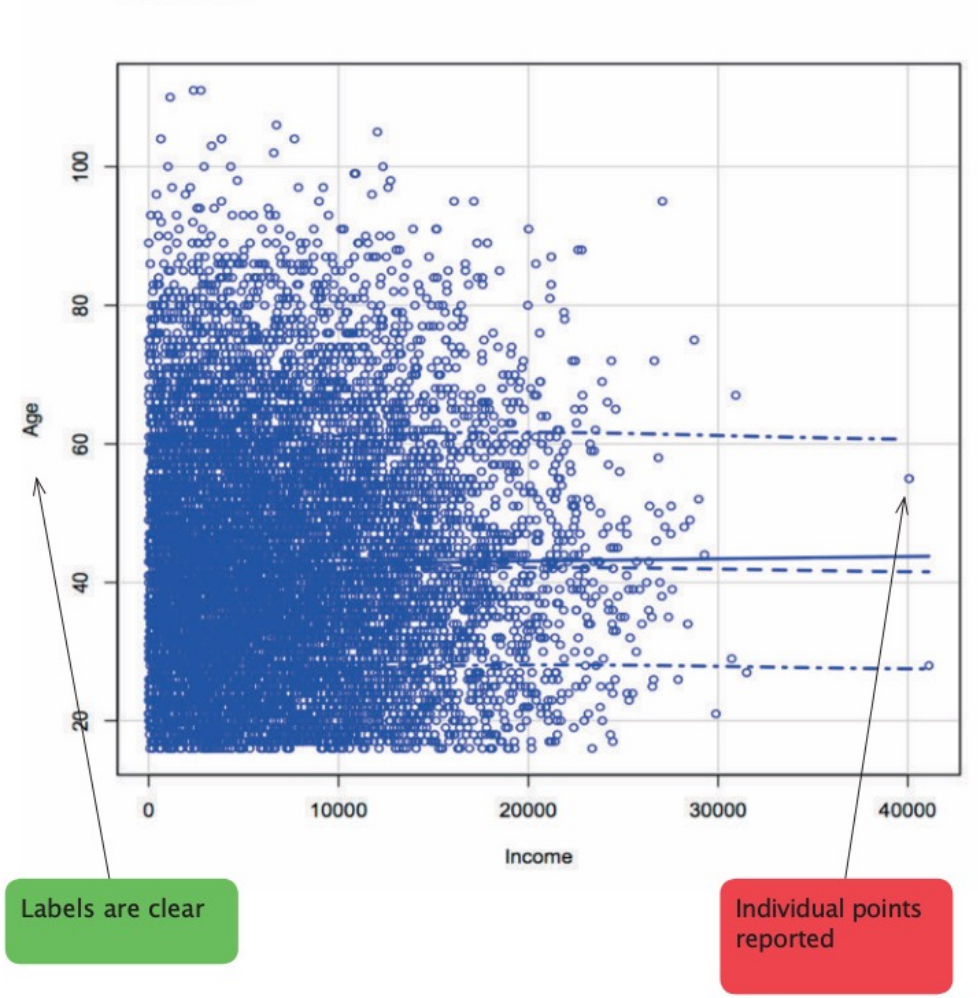


Practices to reduce disclosure:

- The minima and maxima values could be grouped or averaged.
- For example, instead of displaying a tail as a minimum value of hourly earnings of £7.40,band into £7 – 8 per hour (providing this met the threshold number of observations, e.g.10).

Scatterplots

FIGURE 4:
Scatter plot



Practices to reduce disclosure:

- Group data subjects together, to ensure that the statistics presented are based on a sufficient number of observations.

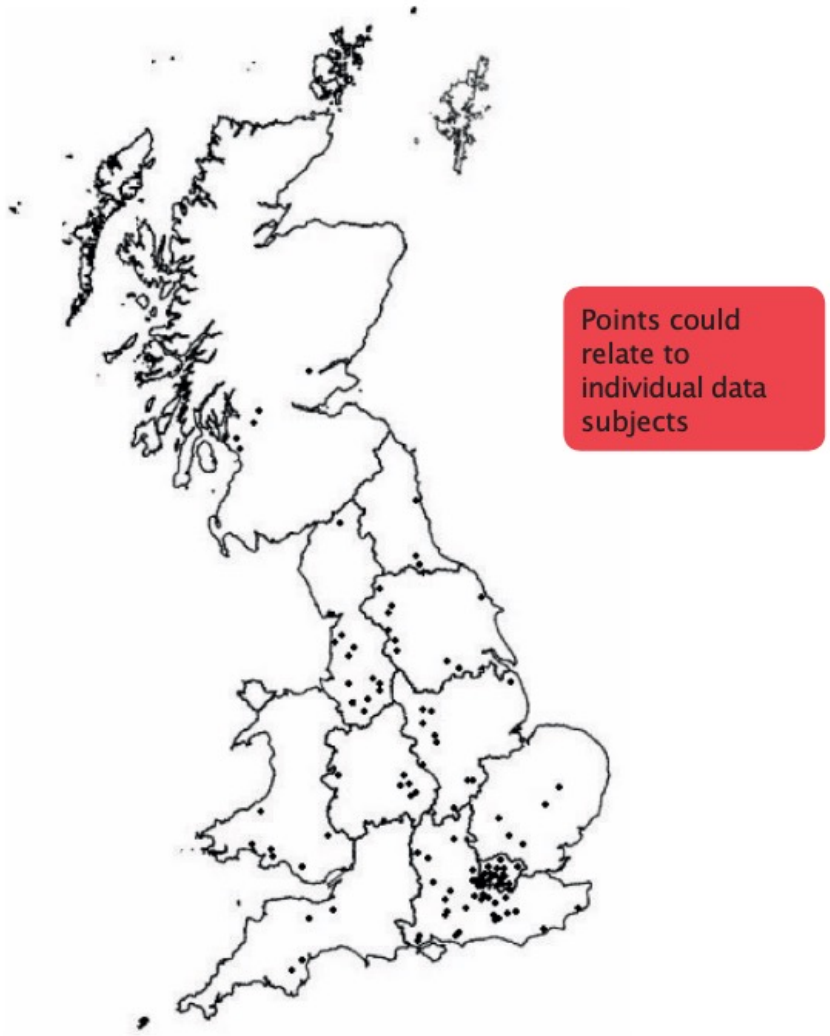
TABLE 7:
Data perturbation for new scatter plot

Could perturb the original data and create a scatter plot

AGE GROUP	EARNINGS PER HOUR (£)	N
16-24	6.10	24
25–34	8.20	20
35–44	8.90	18
45–55	8.80	22
55 or over	10.20	14
Total		98

Spatial analysis (maps)

FIGURE 9:
Point map

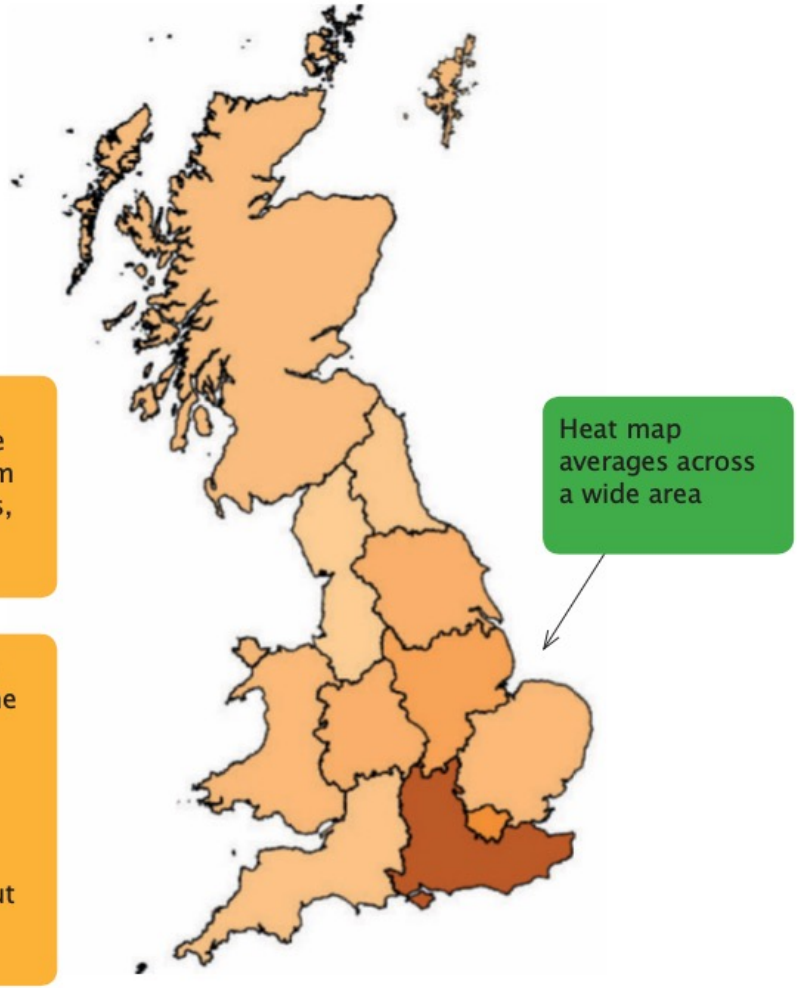


Problems with disaggregate location data maps:

- Each dot may represent a single observation
- Are the observations ‘unusual characteristics’ (e.g.rare condition X)?
- Dots may be very precisely positioned. Could be coupled with other information such as gender, age, or ethnicity, which would increase the risk of reidentification.

Spatial analysis (maps)

FIGURE 10:
Heat map



Could information be combined from public sources, e.g. Google Maps?

Depending on the scale of the map, and the nature of the data subject, it could be possible to release without disclosure occurring.

Heat map averages across a wide area

Practices to reduce disclosure:

- Uses a colour scheme to indicate levels of activity, intensity, concentration etc.
- Darker colours indicate high activity whilst lighter colours indicate lower activity.
- Aggregating point data by administrative spatial units → regions in this case, reduces disclosure risk.
- Note: we still need to be aware of low counts e.g. less than 10

Exercise: De-identification of quantitative data

In groups, discuss the actions you would take to tackle the various disclosure risks associated with each variable

Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables).	
Age	Low counts of older respondents over 75 years old	'
Main occupation during last 12 months	Low counts of very specific occupations.	
Ethnicity of the Household Head	Low counts of specific ethnicities.	
Household's primary type or energy/fuel used for cooking: Firewood, Electricity-based, Charcoal, Electricity-solar panel, Gas/LPG	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	
Main material of the wall of the house, e.g. Cane/palm/trunks, Dirt/mud, Wood/bamboo with mud, Stone with mud, Uncovered adobe , Cardboard , Cement blocks, Stone with lime/cement, Bricks, ...	A number of low-frequency responses; exterior features (households/buildings easily identifiable).	
Crops grown on plots	A number of low-frequency specific responses for each variable.	

Exercise: De-identification of quantitative data

Millennium Village Impact Evaluation in Northern Ghana, 2012-2016: Special Licence Access

Details | Documentation | Resources | Access data

Details

Title:	Millennium Village Impact Evaluation in Northern Ghana, 2012-2016: Special Licence Access
Alternative title:	MVP; SADA-North Ghana Household Survey
Study number (SN):	7734
Access:	These data are safeguarded
Persistent identifier (DOI):	10.5255/UKDA-SN-7734-4
Data creator(s):	University of Sussex, Institute of Development Studies Columbia University (New York), Earth Institute

Exercise: De-identification of quantitative data

Variables	Disclosure risk	Action
Community	Low frequency counts for all named communities, respondents who gave answers very easily identifiable (especially in combination with other variables).	Exclude variable from dataset
Age	Low counts of older respondents over 75 years old	Top-code age >= 75 as '75 and over'
Main occupation during last 12 months	Low counts of very specific occupations.	Occupations aggregated into standard occupation codes
Ethnicity of the Household Head	Low counts of specific ethnicities.	Recode the low-frequency responses (all responses but 'Mamprusi' and 'Builsa') into 'Other'.

Exercise: De-identification of quantitative data

Variables	Disclosure risk	Action
Household's primary type or energy/fuel used for cooking	Very low counts for 'Gas/LPG' and 'Electricity-solar panel' responses may lead to household identification (especially if combined with other datasets)	Recode all responses into the following main categories: 1 - 'Firewood'; 2 - 'Electricity-based'; 3 - 'Charcoal'; 4 - 'Other', 5 - 'Don't know'; 6 - 'NA/missing'.
Main material of the wall of the house	A number of low-frequency responses; exterior features (households/buildings easily identifiable).	As the main material of the wall refers to the exterior of a building, it may be advisable to recode the low-frequency and 'Other' variables into 'Other (incl. wood-based and stone-based') and retain the remaining groups
Crops grown on plots	A number of low-frequency specific responses for each variable.	Variables are recoded into crop categories

Exercise: De-identification of qualitative data

1. In this example interview transcript, where would you have concerns for the risk to disclose the identity of the interviewee? What direct and indirect identifiable information do you note in the text that might concern you? Highlight any words, phrases or sections that you think need to dealt with.

2. How might you de-identify the text to reduce the risk to disclose the identity of the interviewee?

Exercise: FAIRness of a dataset

Assess how Findable, Accessible, Interoperable and Reusable (FAIR) the below datasets are:

A. The lived experiences of migration 1996-2017.
B. Facility ownership and mortality among older adults residing in care homes

Go to

www.menti.com

Enter the code

4266 0921

Or use QR code