
Curating Data: Resolving the Issues

Dr Sharon Bolton
Data Curation Manager

BD013 – Managing, Curating and Publishing Data
IADS Big Data and Analytics Summer School
University of Essex, 25 August 2015

UK Data Service



Curating Data

- Look at some basic quantitative social science data problems and their solutions
- What kind of problems might be encountered with big data? Are they similar?
- Real-life examples, from the UK Data Service and elsewhere.

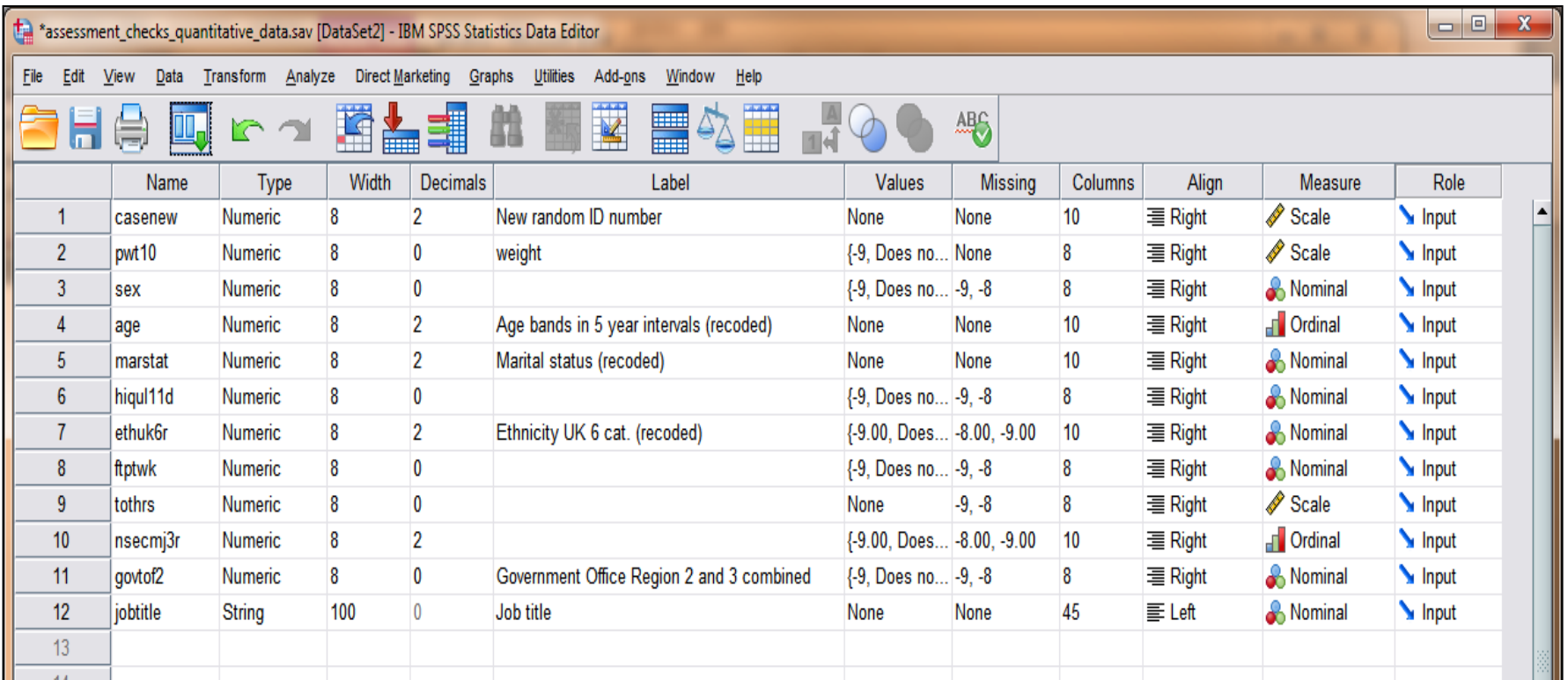


Curating data: finding the problems

- Metadata issues
 - Are all the variables labelled?
 - Do the data match the documentation?
 - Do the categorical variables have value labels?
 - Derived data, weights – enough information?
- Data issues
 - Data cleaning: duplicate records? Accurate values?
 - Data integrity: lots of missing data?
 - How are missing values handled?
 - Are the data disclosive?



Metadata: are all variables labelled?



*assessment_checks_quantitative_data.sav [DataSet2] - IBM SPSS Statistics Data Editor

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	casenew	Numeric	8	2	New random ID number	None	None	10	Right	Scale	Input
2	pwt10	Numeric	8	0	weight	{-9, Does no...	None	8	Right	Scale	Input
3	sex	Numeric	8	0		{-9, Does no...	-9, -8	8	Right	Nominal	Input
4	age	Numeric	8	2	Age bands in 5 year intervals (recoded)	None	None	10	Right	Ordinal	Input
5	marstat	Numeric	8	2	Marital status (recoded)	None	None	10	Right	Nominal	Input
6	hiql11d	Numeric	8	0		{-9, Does no...	-9, -8	8	Right	Nominal	Input
7	ethuk6r	Numeric	8	2	Ethnicity UK 6 cat. (recoded)	{-9.00, Does...	-8.00, -9.00	10	Right	Nominal	Input
8	ftptwk	Numeric	8	0		{-9, Does no...	-9, -8	8	Right	Nominal	Input
9	tothrs	Numeric	8	0		None	-9, -8	8	Right	Scale	Input
10	nsecmj3r	Numeric	8	2		{-9.00, Does...	-8.00, -9.00	10	Right	Ordinal	Input
11	govtof2	Numeric	8	0	Government Office Region 2 and 3 combined	{-9, Does no...	-9, -8	8	Right	Nominal	Input
12	jobtitle	String	100	0	Job title	None	None	45	Left	Nominal	Input
13											

- If not, are they listed in documentation?

Metadata: data matches documentation?

List of Variables

Observations: 25,162

Variables: 13

Variable name	Variable label
<u>casenew</u>	New random ID number
<u>pwt10</u>	Integer weight 2010
<u>sex</u>	Sex of respondent
<u>age</u>	Age in years
<u>marstat</u>	Marital status – 3 categories (recoded)
<u>hiqul11d</u>	Highest qualification – 6 categories
<u>ethuk6r</u>	Ethnicity –6 categories
<u>ilodefr</u>	Economic activity (reported) – 3 categories
<u>stat3r</u>	Employment status – 3 categories
<u>ftptwk</u>	Full-time or part-time in main job
<u>tothrs</u>	Total hours worked in reference week
<u>nsecmj3r</u>	NS-SEC 3 (main job) (SOC2010)
<u>govtof2</u>	Government Office Region 2 and 3 combined

assessing_data.sav [DataSet1] - IBM SPSS Statistics Data Editor								
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help								
	Name	Type	Width	Decimals	Label	Values	Missing	Col
1	casenew	Numeric	8	2	New random ID ...	None	None	10
2	pwt10	Numeric	8	0	weight	{-9, Does no...	None	8
3	sex	Numeric	8	0	Sex	{-9, Does no...	-9, -8	8
4	age	Numeric	8	2	Age bands in 5 ...	None	None	10
5	marsta3r	Numeric	8	2	Marital status (r...	None	None	10
6	hiqul11d	Numeric	8	0		{-9, Does no...	-9, -8	8
7	ethuk6r	Numeric	8	2	Ethnicity UK 6 c...	{-9.00, Does...	-8.00, -9.00	10
8	ftptwk	Numeric	8	0		{-9, Does no...	-9, -8	8
9	tothrs	Numeric	8	0	Total hours wor...	None	-9, -8	8
10	nsecmj3r	Numeric	8	2	NS-SEC 3 clas...	{-9.00, Does...	-8.00, -9.00	10
11	govtof2	Numeric	8	0	Government Off...	{-9, Does no...	-9, -8	8
12	jobtitle	String	100	0	Job title	None	None	45
13								
14								
15								

- variables **ilodefr** and **stat3r** in document but not in data file
- variable **jobtitle** included in data file but not in document

Metadata: sufficient value labels?

marstat Marital status (recoded)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00	69	69.7	69.7	69.7
	2.00	28	28.3	28.3	98.0
	3.00	2	2.0	2.0	100.0
Total		99	100.0	100.0	

- What do the categorical values 1, 2 and 3 mean? Single, Cohabiting, Married, Divorced or Widowed?
- If not in data, are the values given in codebook, questionnaire or other documentation?

Metadata: derived/weighting variables

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
pwt10 Integer weight 2010	25162	0	11787	595.95	381.404
Valid N (listwise)	25162				

- There is a weighting variable in the dataset, but the documentation is incomplete

variables were recoded is in Appendix II. In addition, a new ID variable was created in order to avoid linking the Unrestricted Access Teaching Dataset with the original dataset.

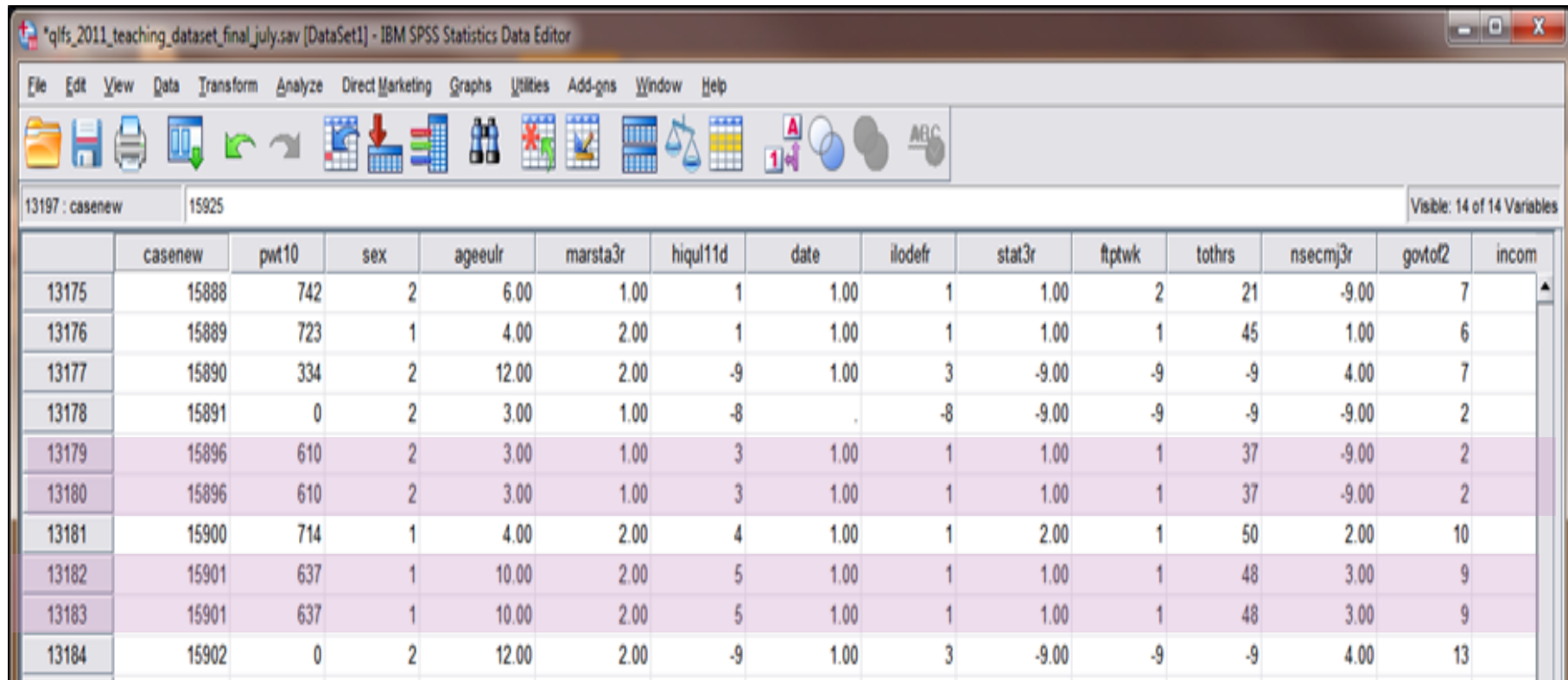
Weighting the Dataset

(To be completed.)

Comment [SEB1]: Must add some text here before deposit!



Data cleaning: duplicate cases



*qlfs_2011_teaching_dataset_final_july.sav [DataSet1] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

13197 : casenew 15925 Visible: 14 of 14 Variables

	casenew	pw10	sex	ageeulr	marsta3r	hiql11d	date	ilodefr	stat3r	ftptwk	tothrs	nsecmj3r	govtof2	incom
13175	15888	742	2	6.00	1.00	1	1.00	1	1.00	2	21	-9.00	7	
13176	15889	723	1	4.00	2.00	1	1.00	1	1.00	1	45	1.00	6	
13177	15890	334	2	12.00	2.00	-9	1.00	3	-9.00	-9	-9	4.00	7	
13178	15891	0	2	3.00	1.00	-8	.	-8	-9.00	-9	-9	-9.00	2	
13179	15896	610	2	3.00	1.00	3	1.00	1	1.00	1	37	-9.00	2	
13180	15896	610	2	3.00	1.00	3	1.00	1	1.00	1	37	-9.00	2	
13181	15900	714	1	4.00	2.00	4	1.00	1	2.00	1	50	2.00	10	
13182	15901	637	1	10.00	2.00	5	1.00	1	1.00	1	48	3.00	9	
13183	15901	637	1	10.00	2.00	5	1.00	1	1.00	1	48	3.00	9	
13184	15902	0	2	12.00	2.00	-9	1.00	3	-9.00	-9	-9	4.00	13	

- Also identify duplicate cases with syntax or menu command rather than by visual inspection

Data cleaning: unfeasible values

- Age variable with many people aged 140!

		Age of respondent			
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	1	.1	.1	.1
	16	903	64.8	64.8	64.8
	20	260	18.7	18.7	83.5
	44	25	1.8	1.8	85.3
	54	35	2.5	2.5	87.8
	72	27	1.9	1.9	89.7
	80	22	1.6	1.6	91.3
	95	7	.5	.5	91.8
	98	1	.1	.1	91.9
	99	1	.1	.1	92.0
	100	1	.1	.1	92.0
	110	6	.4	.4	92.5
	115	5	.4	.4	92.8
	140	100	7.2	7.2	100.0
Total		1394	100.0	100.0	



Data integrity: missing data

income Household income (banded, 14 categories)

		Frequency	Percent
Missing	System	25162	100.0

- Important variable completely missing from dataset

Data integrity: missing values

stat3r Employment status (recoded)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.00 Employee	15659	62.2	86.3	86.3
	2.00 Self-employed	2411	9.6	13.3	99.6
	3.00 Government scheme or unpaid family worker	71	.3	.4	100.0
	Total	18141	72.1	100.0	
Missing	-9.00 Does not apply	7021	27.9		
Total		25162	100.0		

stat3r Employment status (recoded)

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	-9.00	7021	27.9	27.9	27.9
	1.00 Employee	15659	62.2	62.2	90.1
	2.00 Self-employed	2411	9.6	9.6	99.7
	3.00 Government scheme or unpaid family worker	71	.3	.3	100.0
	Total	25162	100.0	100.0	

- Analysis can be affected if missing values are not properly defined

Data: are the data disclosive?

jobtitle	Job title	Frequency	Cumulative Percent
		74	74.7
carer		1	75.8
cleaner		1	76.8
dentist		2	78.8
don't know		2	80.8
Financial Director, Sunnyview Care Home, Colchester		1	81.8
general manager		1	82.8
Manager, Wivenhoe Village Stores		1	84.8
on benefits		1	86.9
professor of anthropological developmental osteology		1	88.9
Data Curation Manager		1	89.0
retired		1	90.9
Retired		1	91.9
retirement		1	92.9
teacher		2	94.9
Teacher		2	97.0
teeacher		1	98.0
unemployed		4	100.0
Total		99	

Specific job title
+ organisation
name

Job titles without
names may still be
disclosive when
combined with other
variables

Same concept,
different spellings -
data need cleaning

Curation: data vs big data

- Previously, we looked at problems you might find in survey research and similar data:
 - Poor metadata
 - Mismatched data and documentation
 - Erroneous data
 - Unfeasible values
 - Undocumented data
 - Incomplete documentation
 - Disclosive data
- What problems might exist in big data?



Curation: data vs big data

- Big data analysis techniques very different but data problems can be similar
 - Noisy data
 - Missing data
 - Unlabelled and undocumented – poor metadata
 - Disclosive data
- Need meaningful metadata to explain data contents, provenance and context regardless of size
- Analysing multiple sources – can they really be matched? Need metadata to find out!



Curation: data vs big data

- UK Data Service: early days development
- Energy Demand Research Project: Early Smart Meter Trials, 2007-2010 (Study number 7591)
 - DECC; how do domestic consumers react to improved information about their energy consumption?
 - 14,621 households
 - 246 million gas meter readings
 - 413 million electricity meter readings
 - readings collected every 30 mins
- Available to all registered UK Data Service users



Curation: data vs big data

- What do we plan to do with it?
 - Apache Hadoop cluster
 - Data visualisation
 - Socio-demographic analysis
 - Matching with weather data?
- How can we do it?
 - Sufficient metadata: advised on documentation
 - Includes geography and segmentation file to match data by ID variable to Acorn group, LA and NUTS areas
 - Recognised coding schemes, easy to match across datasets



Curation: data vs big data

“Metadata needs to be organized for data sharing to be productive, for the simple reason that users need to know what the data is. The need for this becomes increasingly pressing as the number of useful and potentially useful data sources increases ... potential will only be realized if the metadata is well-documented and well-managed.”

Robin Bloor, *Does Big Data Mean Big Metadata?*

<http://www.information-management.com/news/does-big-data-mean-big-metadata-10025760-1.html>



Curation: questions

- What's your experience with data, big or small?
- What problems did you find?
- How did you resolve them?
- Can the UK Data Service help?



Curating data

Dr Sharon Bolton, Data Curation Manager
sharonb@essex.ac.uk

