

Introduction to QAMyData 'health-check' tool for numeric data

Louise Corti
Cristina Magder
Myles Offord,
UK Data Service

Webinar
2 December 2019



Introductions

Louise Corti
Service Director,
Data Publishing
and Access

Cristina Magder
Senior Data
Curation Officer

Myles Offord
Hadoop Systems
Engineer

What we will cover today

1. The origin of the tool
2. What is useful to check for numeric data
3. The development of our QAMyData tool
4. The tests that are available
5. A demo
6. Quick techie overview
7. Future plans

We will be sharing our slides on our events page:

<https://www.ukdataservice.ac.uk/news-and-events/eventsitem/?id=5548>

When does data need quality assessing?

Data publishing

- A researcher submitting data to a repository
- A repository for checking quality
- Peer review of published analysis for a journal
- Supporting FAIR Principles

Using data

- A researcher using a new data source
- Preparing data for students learning about quality



When we want **data to be healthy and safe**

QAMyData Tool

- UK Data Service project to develop a **light weight, open-source tool for quality assessment of research data**
- A '**data health check**' tool that identifies the most common problems in data submitted in disciplines that utilise quantitative methods
- Helps set **Data Quality Profiles** for data publishers using own default settings



Question

How do you apprise numeric data

Data review

- As data publishers we are aiming to:
 - ✓ Share clean and well documented data
 - ✓ Share data under the right conditions
- Quality issues arise in the **data description** and **data itself**
- Datasets to be shared require a **Privacy Impact Assessment** (e.g. GDPR)
 - Privacy level defines the **legal gateway** for access

UK Data Service Data Access Policy

Open

- Download/online access under open licence without registration

Safeguarded

- Download/online access: authorised, authenticated and audited users

Controlled

- Remote or safe room access: authorised, authenticated and audited users; projects have been approved users receive specialist training; research outputs are checked

Current ways of checking data

- Getting to know your data:
 - Check structure
 - Look for incorrect, missing, inconsistent values
 - Check for unanticipated/accidental disclosure risk
 - Locate issues and decide how to treat them
 - Treatment: clean data or flag errors
- Data creators/publishers largely use manual methods:
 - Integrity rules used in data collection
 - Statistical software commands
 - Mostly eyeballing data

How can a tool help?

- Flag issues: enable a machine or human to resolve the problems
- Be deployed as a service for self deposit repository, e.g. DataVerse for a submission health check
- Be deployed into data publishing pipelines

Question

Biggest problems encountered with data?

What do you check?

- Do the data values seem correct?
- Do they make sense?
- Check frequency distributions for **erroneous outliers**
- Check **formats of the values** entered for a variable
- Check **low frequency counts/possible disclosive outliers**
- Irritating **missing data!**

Checking data: example 1

pregnant	sex	age in years	occupation
y	female	29	2018-07-16: 12:37
n	female	15	2018-07-16: 12:38
y	female	32	2018-07-16: 12:39
n	male	37	2018-07-16: 12:40
n	female	-10	2018-07-16: 12:41
n	female	51	2018-07-16: 12:42
n	male	22	2018-07-16: 12:43
n	male	126	2018-07-16: 12:44
n	male	28	2018-07-16: 12:45
y	male	31	2018-07-16: 12:46
n	male	42	2018-07-16: 12:47
y	female	37	2018-07-16: 12:48

Checking data: example 2

timestamp
2017-01-22: 10:15
201701241016
2017 01 22 10-17
2017-01-22-1018
2017 01 22 10 19

Checking data: example 3

age	income (£)	education_level	occupation type
40	52,000	postgraduate	professional
34	35,000	A-level	non-professional
25		GCSE	non-professional
30		GCSE	non-professional
28	22,000	GCSE	non-professional
55	37,000	graduate	professional
45		postgraduate	professional
33		A-level	non-professional
41	53,000	postgraduate	professional
43		postgraduate	professional
27	22,500	GCSE	non-professional

Data review: what to look for

- Basic file checks
- Metadata issues
- Data integrity Issues
- Disclosure review and control

Scoping the tests

- In-house procedures for data checking
- Prepared 'dirty' test datasets for evaluation
- Reached out to other archives/data publishers to gather information on their own QA checks
- Feedback from tool use in early training

Basic file checks

File opens

Checks whether acceptable format

Bad filename check, regular expression pattern - RegEx

Regex requires quotes "[a-z]". To use a special characters, e.g. a backslash (\) a backslash before is required e.g. \\

Metadata checks

Report on number of cases and variables

Always run

Count on grouping variables

Missing variable labels

Must be set to true

No label for user defined missing values e.g. - 9
not labelled

SPSS only

Odd characters in variable names and labels

User specifies the characters

Odd characters in value names and labels

User specifies the characters

Maximum length of variable labels e.g. > 79

User specifies the characters

Maximum length of value labels e.g. > 39

User specifies the characters

Spelling mistakes (non-dictionary words) in
variable labels using a dictionary file

User specifies a dictionary file

Spelling mistakes (non-dictionary words) in value
labels using a dictionary file

User specifies a dictionary file

Data integrity checks

Report number of numeric and string variables

Check for duplicate IDs

Odd characters in string data

% of values missing ('Sys miss' and undefined missing)

Spelling mistakes (non-dictionary words) in string data using a dictionary file (can check if date format set correctly!)

User specifies the variables.
Multiple variables can be added on new lines e.g.
- Caseno
- AnotherVariableHere

User specifies the characters

User sets the threshold, e.g. more than 25%

User specifies a dictionary file

Disclosure checks

Unique values or low thresholds (freqs of categorical vars or minimum values)

User sets the threshold value, e.g. 5

Direct identifiers using RegEx pattern search

User runs separately for postcodes, telephone numbers etc.
Advise tests are separately as may be resource intensive

Direct identifiers/named entities in string data using a dictionary file

Specify a dictionary file containing lists of stop words or named entities e.g. for places, names etc.
Advise tests are separately as may be resource intensive

Key-identifiers and non-identifying variables (combinations of key variables) – we use and train on **sdcMicro**

Formats and test selection

- Accepts formats: **SPSS, Stata, SAS, CSV**
- Modular design: user configures for: **data type, threshold**

Checks can be commented out or omitted to exclude them from the checks to be run

```
# Variable Configuration

[variable_config.odd_characters]
setting = ["!", "#", " ", "@", "ë", "ç", "ô", "ü"]
desc = "Variable names and labels cannot contain certain 'odd' characters."

[variable_config.missing_variable_labels]
setting = true
desc = "Variables should have a label."

[variable_config.label_max_length]
setting = 79
desc = "Variable labels cannot exceed a max length"
```

```
qamd run ./data/test/mtcars.sav --metadata-only --output-format json --
output results_mtcars.json --config ./config/default.yaml
```

RegEx

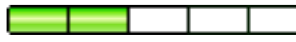
▣ **Title** **Email - Overly Simple** [Details](#) [Test](#)

▣ **Expression** `^\w+@[a-zA-Z_]+?\.[a-zA-Z]{2,3}$`

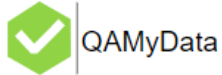
Description Simple email expression. Doesn't allow numbers in the domain name and doesn't allow for top level domains that are less than 2 or more than 3 letters (which is fine until they allow more). Doesn't handle multiple ";"; in the domain (joe@abc.co.uk).

Matches `joe@aol.com` | `ssmith@aspalliance.com` | `a@b.cc`

Non-Matches `joe@123aspx.com` | `joe@web.info` | `joe@company.co.uk`

▣ **Author** **Steven Smith** **Rating:** 

Reporting



teaching-data%set.sav

Raw Case Count: 10210

Aggregated Case Count: 0

Total Variables: 188

Data Type Occurrences: Numeric: 186, String: 2

Created At: 2019-02-18 13:37:39

Last modified at: 2019-02-18 13:37:39

File Label:

File Format Version: 2

File Encoding: WINDOWS-1252

Compression type: Rows

Basic File Checks

Name	Status (N)	Description
Bad file name	failed (1)	File name should match the user specified pattern

Metadata Checks

Name	Variable odd characters		
Missing variable labels	# (limited to 1000)	Variable	Row number
Variable odd characters	1	OwnTV	-
Variable label max length	2	V137	-

Demo



Software decisions

Use **existing open source libraries** where possible

- SPSS Java libraries – IBM is licenced and difficult to deploy, Dext project
- Stata Java libraries – cross-version implementation not good
- R libraries – Haven based on Readstat – difficult to deploy, server licence, performance
- Python libraries – no good stable libraries for statistical packages
- Enter [Readstat](#)

Using the ReadStat library

- A command-line tool and C library for reading files from popular stats packages
- Originally developed for Wizard for free stat data analysis on a Mac
- Supports SPSS, Stata and SAS files – new & old formats
- In active development since 2012, continually receiving security patches and bug fixes from the open source community
- Currently used by the R library Haven, part of the Tidyverse collection of R packages for data science

RUST programming language

- Tried the following wrappers: [Java](#), [Clojure](#), [R](#), [Python](#)
- RUST from the Mozilla Foundation for improving the Firefox browser: [an environment that demands things just 'work'](#)
- **Performance**: Rust generates executables that run very fast, without needing to write low level C code; easily integrates with other languages
- **Reliability**: enables the developer to eliminate many classes of bugs at compile-time
- **Productivity**: has great documentation, a friendly compiler with useful error messages, and excellent tooling

Deployment

- Downloadable to run on Linux, Windows, Mac
- Simple to install and deploy from our Github
- Lightweight to run and set tests - edit config file
- Wiki with documentation
- Space for suggesting new tests
- Released under a MIT licence
- 2020: Will deploy as a service, prior to ReShare upload

GitHub space

ukdataservice / qamd

Watch 3

Star 10

Fork 3

Code

Issues 3

Pull requests 0

Projects 0

Wiki

Security

Insights

Join GitHub today

Dismiss

GitHub is home to over 40 million developers working together to host and review code, manage projects, and build software together.

Sign up

QAMyData, a data quality assurance tool for SPSS, STATA and SAS files

spss

stata

data-quality

qa

readstat

quality

assurance

151 commits

2 branches

0 packages

2 releases

1 contributor

MIT

Branch: master

New pull request

Find file

Clone or download



Lyrain added string value stopword

Latest commit 4bf0647 on Sep 30

src

added string value stopword

2 months ago

test

added string value stopword

2 months ago

Try it out

Evaluation has been conducted with:

- UKDS data curation staff
- Peer data repositories in our international network
- Data owners, researchers, data managers, quant. methods lecturers, journal publishers for data peer review
- **More testing by you!**

Advocate data publishers to develop a Data Quality Profile with stated thresholds

Resources

- Table of Available Tests
- Download and Run Guide, including step by step for editing config. file, setting own thresholds
- Teaching resources, slides, exercises and test data
- A blog

<https://www.ukdataservice.ac.uk/about-us/our-rd/qamydata.aspx>

Question

How might you use QAMyData in your own work?

Acknowledgements

Thank you to our colleagues on the **QAMyData** team:

- Jon Johnson: lead specs and development
- Anca Vlad: input into tests and testing

And to Australian Data Archive for adding an Open Source front end!

Keep connected

QAMyData@ukdataservice.ac.uk
UK Data Service
University of Essex, Colchester, UK

Subscribe to UK Data Service list:

www.jiscmail.ac.uk/cgi-bin/webadmin?A0=UKDATASERVICE

Follow UK Data Service on Twitter: @UKDataService, @UKDSRDM

Youtube: www.youtube.com/user/UKDATASERVICE

Thankyou

Thank you
Any questions?