# A Practical Guide to Characterising Data and Investigating Data Quality

Roy Ruddle, James Cheshire and Sara Johansson Fernstad

## Contents

List of Tables ..... 1
Introduction ..... 1
Workflow ("what and in which order?") ..... 2
Profiling tasks ("questions to answer about your data") ..... 2
Characterisation tasks ..... 3
Data quality tasks ..... 4
How? ("guidance about how to answer key questions") ..... 7
Further reading ..... 7
About the authors ..... 8

## List of Tables

Table 1: Characterisation tasks for each data source or file. ..... 3
Table 2: Characterisation tasks for each variable/field/column. ..... 3
Table 3: Characterisation tasks for distributions of each variable. ..... 4
Table 4: Characterisation tasks for combinations of variable. ..... 4
Table 5: Data quality tasks for each data source or file. ..... 5
Table 6: Completeness tasks for each variable. ..... 5
Table 7: Accuracy tasks for each variable. ..... 6
Table 8: Consistency tasks for each variable. ..... 6
Table 9: Data quality tasks for combinations of variable. ..... 7

## Introduction

Data science projects generally start with efforts to prepare the required data, then proceed to its analysis before, finally, reporting the results. The data preparation step often takes at least half the time (and sometimes as much as 90%) and may itself be subdivided into three stages: discovery, wrangling and profiling.

The focus of this guide is the data profiling stage, which can approached from two complementary perspectives: establishing the characteristics of the data and assessing its quality. For example, you may wish to determine how the number of records varies with time (data characterisation) or check whether records are missing from a given time-period (data quality). Similarly, you may calculate the distribution of data (characterisation) or determine if outlying values are implausibly high/low (quality).

This guide is designed for data scientists to use in their day-to-day work and describes:


Ruddle, R. A., Cheshire, J., and Johansson Fernstad, S. (2024):
A Practical Guide to Characterising Data and Investigating Data Quality. https://doi.org/10.5518/1481

a) A recommended workflow (what tasks to perform and in which order, to avoid rework or nasty surprises late on in your project).
b) A comprehensive list of tasks (each articulated as questions to answer about your data).
c) How (guidance about how to answer key questions).

The material in this guide is informed by a survey and interviews with public and private sector data scientists/analysts working on a diverse range of projects [3] and combined with our first-hand experiences as researchers interested in the data profiling process.

Currently, most data scientists take an ad-hoc approach to data profiling, making it more an art than a science. Many data scientists' profiling is rather superficial, so the quickest win for them is simply to perform a more comprehensive set of tasks.

In keeping with that, one strong wish that emerged from the interviews was to formalise the process of data profiling in some way to make it more rigorous, consistent, and reproducible, while at the same time avoiding a "one size fits all" mentality that may inhibit creativity. The result is this practical guide, which you can use as a checklist, rulebook or just to compare your approach with other practitioners.

# Workflow ("what and in which order?")

The tasks are divided into a six-step workflow. That will help you to identify whether there are any issues with your data, and remedy them as early as possible to avoid rework. The steps are:

1. Look at your data (is anything obviously wrong or missing?).
2. Watch out for special values (e.g., indicating a missing or invalid value; you will need to flag them as missing, etc. before proceeding to the next step).
3. Is any data missing (is it safe to proceed or do you need more data?)?
4. Check each variable (is it what you expect?).
5. Check combinations of variables (do they violate any of your assumptions or business rules?).
6. Characterise the cleaned data.

With this guide, we have also provided a checklist (a spreadsheet) that lists the tasks and questions that we recommend answering in each step. The mapping between each task/workflow step is also indicated below (see Tables 1 – 9). Our workflow includes 40 of the 62 characterisation and data quality tasks, spanning 66 of the 91 questions. As the tables indicate, a few tasks/questions are repeated in Step 6, but on the cleaned data (e.g., C-1 Example values). The other 22 tasks are not in the workflow, but they might me important for your work – use your judgement!

The spreadsheet also provides two other lists of tasks: Minimal and Intermediate. The Minimal list contains the nine profiling tasks that data scientists typically do today, but which cannot be considered a rigorous approach. The Intermediate list contains seven additional profiling tasks (making 16 in total) that were performed by a subset of the data scientists who we surveyed/interviewed (the "comprehensive profilers" in [3]). That list also lacks breadth.

# Profiling tasks ("questions to answer about your data")

This section describes a comprehensive set of tasks that will enable you to characterise your data and investigate data quality. The tasks are articulated as questions for you to answer about your data.

## *Characterisation tasks*

These tasks are divided into groups about whole data sources or files (see Table 1), each variable (see Table 2), the distribution of each variable (see Table 3) and combinations of variables (see Table 4). Within each table, the tasks are listed in alphabetical order and accompanied by one or more questions for you to answer. The answers to these questions can then inform how you might wish to proceed, for example if the range of values falls outside expectations, then you might need to revisit the original source of the data to establish where the error occurred.

| ID | Task | Questions to answer about each data source or file | Workflow Step(s) |
|---|---|---|---|
| C-1 | Example values | What does the data look like (print the first/last few rows)? | 1, 6 |
| C-2 | File format | What text encoding format do the files use? | 1 |
| C-3 | Number of columns | How many columns are there in the data? | 6 |
| C-4 | Number of rows | How large is each file (e.g., in gigabytes)?<br>How many rows of data are there? | 6 |

*Table 1: Characterisation tasks for each data source or file.*

| ID | Task | Questions to answer about each variable/field/column | Workflow Step(s) |
|---|---|---|---|
| C-5 | Characters & symbols | What characters and symbols occur in any value? | |
| C-6 | Data format | In what format are the values stored?<br>What are the patterns of the values?<br>Which characters are used in the values? | 6 |
| C-7 | Data type | What is the data type?<br>Is the analysis software using the correct data type? | 6 |
| C-8 | Number of infinite values | How many values are flagged as infinite? | |
| C-9 | Number of null values | How many null values are there? | |
| C-10 | Number of special values | Which values have a special meaning (e.g., not known or invalid) and how many are there? | 2 |
| C-11 | Number of unique values | How many unique (distinct) values are there?<br>What are those values?<br>Does the number of unique values equal the number of records (or the number of values, if some are missing)? | 1, 6 |
| C-12 | Number of zero values | How many values are equal to zero? | 2 |
| C-13 | Precision | How many significant figures or decimal places are there? | |
| C-14 | Units | What are the units? | |
| C-15 | Value lengths | How long is each value (how many characters are there)?<br>Is every value the same length?<br>If not, what is the minimum/maximum length? | 4 |

*Table 2: Characterisation tasks for each variable/field/column.*

| ID | Task | Questions to answer about variable distributions | Workflow Step(s) |
|---|---|---|---|
| C-16 | First digit | How many times does each digit (0 – 9) appear first? | |
| C-17 | Frequency measures | What is the distribution of the values?<br>How many times does each value occur?<br>What is the shape of the distribution (e.g., Gaussian, bimodal or skewed)? | 4, 6 |
| C-18 | Outliers | Are there any outliers? | 2, 6 |
| C-19 | Range and variability | What is the minimum/maximum value?<br>What is the mean, variance, etc?<br>What are certain percentiles (e.g., 25$^{th}$, median and 75$^{th}$)? | 2, 6 |

*Table 3: Characterisation tasks for distributions of each variable.*

| ID | Task | Questions to answer about combinations of variable | Workflow Step(s) |
|---|---|---|---|
| C-20 | Clusters | What clusters are there?<br>Are any groups of records similar? | |
| C-21 | Correlation | What is the correlation coefficient for each pair of numerical variables? | 5 |
| C-22 | Cross tabulation | How often does each combination of values occur? | |
| C-23 | Curve fitting | What type of curve (if any) fits each pair of numerical variables? | |
| C-24 | Primary features | What are the primary features in the data (e.g., principal components)? | |
| C-25 | Trends | Do any pairs of variables exhibit trends? | 5 |

*Table 4: Characterisation tasks for combinations of variable.*

### *Data quality tasks*

Data quality is divided in a similar manner to the characterisation tasks. Table 5 lists tasks for investigating data quality issues that most often occur when you combine data from different sources or files. The greatest number of issues are those that occur for individual variables, so those tasks are listed separately for completeness (see Table 6), accuracy (see Table 7) and consistency (see Table 8). Finally, Table 9 lists tasks for data quality tasks issues that involve combinations of variable.

Data is measured against three kinds of backdrop: time, space and population. Time may be at any granularity (centuries, years, seconds, etc.). Space includes places, geographic regions, other types of area or higher-dimensional spaces. Population may refer to people or any other type of entity (animals, plants, sensors, etc.)

| ID | Task | Questions to answer about missing data | Workflow Step(s) |
|---|---|---|---|
| DQ-1 | Missing column names | Are any variables missing their name (most often it is the last variable in a table)? | 1 |
| DQ-2 | Missing header | Are the names of the variables provided (e.g., in the first row)? | 1 |
| DQ-3 | Missing records | Is the number of records similar to the number you expect?<br>If there are multiple files, do any have notably fewer records (or a much smaller file size) than the others? | 1 |
| DQ-4 | Missing variables | Does the data contain all of the variables you expected?<br>If there are multiple files, do they all contain the same variables? | 1 |
| | **Task** | **Questions to answer about inconsistent data** | |
| DQ-5 | Different table structure | Do the files store certain variables in different ways (e.g., date vs. separate day, month and year)? | |
| DQ-6 | Different word orderings | Do the files store certain variables in different orders (e.g., first name then family name vs. the opposite way around)? | |

*Table 5: Data quality tasks for each data source or file.*

| ID | Task | Questions to answer about missing data | Workflow Step(s) |
|---|---|---|---|
| DQ-7 | Missing values | How many values are missing?<br>Are there no values at all?<br>Did you expect the variable to be complete, but some values are missing? | 3, 6 |
| | **Task** | **Questions to answer about coverage** | |
| DQ-8 | Coverage | Does the data cover the correct time period?<br>Does the data cover the correct areas (e.g., places or geographical regions)?<br>Are you missing data for any part of the population (people, sensor, etc.)?<br>Does the data contain enough samples of each group in the population (e.g., male/female)? | 3, 6 |
| DQ-9 | Granularity | Is the data too fine (what can you discard) or coarse? | |
| | **Task** | **Questions to answer about duplicates** | |
| DQ-10 | Duplicate header | Do the variables names appear in multiple rows? | |
| DQ-11 | Exact duplicates | Are any records identical? | 3 |
| DQ-12 | Uniqueness violation | Should every value be unique, and is that true? | 4 |

*Table 6: Completeness tasks for each variable.*

| ID | Task | Questions to answer about extreme values | Workflow Step(s) |
|---|---|---|---|
| DQ-13 | Extreme values | Are any values outliers (numerically, or in terms of time or space)? | 4 |
| DQ-14 | Unusual category name | Is the length of any categorical values much shorter/longer than the others, or otherwise different? | 4 |
| | **Task** | **Questions to answer about incorrect values** | |
| DQ-15 | Embedded values | Are multiple variables combined into one value (e.g., date and time, or multiple elements of an address)? | |
| DQ-16 | Misfielded | Do any values appear in the wrong field or row? | |
| DQ-17 | Misspelling | Are all words spelt correctly? | |
| DQ-18 | Noise | How noisy or how much variation is there in the values? | 4 |
| DQ-19 | Validity | Do any values contain characters or symbols that should not be present?<br>Do all of the values exist in the domain (e.g., names of categories or numbers from a specific set of choices)? | 4 |
| DQ-20 | Wrong data type | Is the data type correct for each variable?<br>Are any dates stored as strings?<br>Are any categorical variables stored as numbers? | 1 |
| DQ-21 | Wrong format | Are all the values stored in the correct format? | 1 |
| DQ-22 | Wrong values | Are any values impossible or definitely not correct (e.g., dates in the future)? | 4 |
| | **Task** | **Questions to answer about plausible values** | |
| DQ-23 | Implausible range | What is the difference between variables (e.g., start and finish time), and are those differences plausible? | 4 |
| DQ-24 | Implausibly low/high values | What are the minimum/maximum values, and are they plausible?<br>Do any values look suspicious when compared with the other values? | 4 |
| DQ-25 | Same value for too many records | Do any values occur many times, and is that plausible? | 4 |

*Table 7: Accuracy tasks for each variable.*

| ID | Task | Questions to answer about inconsistent semantics | Workflow Step(s) |
|---|---|---|---|
| DQ-26 | Inconsistent coverage | Does the data refer to different time periods?<br>Does the data refer to different areas (e.g., places or geographical regions)?<br>Does the data refer to different populations (people, sensor, etc.)? | 3 |
| DQ-27 | Inconsistent granularity | Is all of the data the same granularity (numbers, time, geography, etc? | |
| DQ-28 | Inconsistent units | Do variables of the same type all use the same units? | |
| DQ-29 | Naming conflicts | Are there any synonyms or homonyms? | |
| | **Task** | **Questions to answer about inconsistent syntax** | |
| DQ-30 | Different data formats | Are variables of the same type all in the same format?<br>Are some values decimals but most are integers?<br>Do some values have a different number of decimal places to the other values? | |

*Table 8: Consistency tasks for each variable.*

| ID | Task | Questions to answer about missing data | Workflow Step(s) |
|---|---|---|---|
| DQ-31 | Combinations of missing value | What combinations of variables are missing together? Are there any combinations of variables for which there should always be at least one record? | 5 |
| | **Task** | **Questions to answer about plausible values** | |
| DQ-32 | Implausible changes in values | By how much do values change (e.g., person's height from one date to another), and is that plausible? | |
| DQ-33 | Implausibly low/high combinations of value | If any variables are correlated (e.g., height and weight), are there any outliers? | 5 |
| | **Task** | **Questions to answer about inconsistent data** | |
| DQ-34 | Inconsistent duplicates | Do any records contradict each other (e.g., different date of birth or ethnicity for a person)? | 5 |
| DQ-35 | Violate assumptions | What assumptions have you made about variables (are they true)? | 5 |
| DQ-36 | Violate functional dependencies | Do any sets of values violate rules (e.g., age vs. date of birth)? | 5 |
| DQ-37 | Violate integrity | Is anything implied by one part of the data missing from another part (e.g., customers but no corresponding account details, or shop open at a given time but no sales data)? | 1 |

*Table 9: Data quality tasks for combinations of variable.*

## How? ("guidance about how to answer key questions")

The vizdataquality Python package (available from pyPI; https://pypi.org/project/vizdataquality/) implements our six-stage workflow in a Jupyter Notebook, with visual analytics functionality that is built on top of MatPlotLib and Pandas. The package also includes a Report class, which enables you to write a report about your data quality investigations and a dataset's profile, while you analyse it. You may output the report with the visualizations and tables you create as a webpage, in Latex or in plain text (e.g., for subsequent editing in a word processor).

Some profiling tasks only require calculations and textual output [1], but for others a visualization is very beneficial. Choosing an effective visualization is sometimes straightforward, but often not. To help. We have described a number of exemplars that data scientists talked about in interviews (see the "Uses of Visualization" section of [3]), and other sources are a YouTube film called Visualizing the Quality of Data (https://tinyurl.com/VizDataQuality) and visualizations of car park fines data [2].

## Further reading


[1] Abedjan, Z., Golab, L., & Naumann, F. 2015. Profiling relational data: a survey. The VLDB Journal, 24, 557–581.

[2] Ruddle, R. A. (2023). Using well-known techniques to visualize characteristics of data quality. Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – IVAPP, 89-100. Open access version: https://raruddle.files.wordpress.com/2023/01/ruddle-ivapp-2023.pdf.

[3] Ruddle, R. A., Cheshire, J., & Johansson Fernstad, S. (2023) Tasks and visualizations used for data profiling: A survey and interview study. IEEE Transactions on Visualization and

Computer Graphics. DOI: https://doi.org/10.1109/TVCG.2023.3234337. Open access version: https://eprints.whiterose.ac.uk/197083/1/ruddle-ieee-tvcg-2023-data-profiling.pdf.

## About the authors

Roy Ruddle is Professor of Computing at the University of Leeds (UK), Research Technology Director of the Leeds Institute for Data Analytics (LIDA), and was a Fellow of the Alan Turing Institute (2018-2023). He has a multidisciplinary background, and researches and teaches data visualization and data science methods. He co-founded Petriva to provide visual data analysis and data mining software for the petroleum industry, and his Leeds Virtual Microscope software was commercialised by the global healthcare company Roche.

James Cheshire is Professor of Geographic Information and Cartography in the UCL Department of Geography and Director of the UCL Social Data Institute. His research focuses on the use of new forms of data for the study of social science and he has published several data visualisation themed books, most recently *Atlas of the Invisible*.

Sara Johansson Fernstad is Senior Lecturer in Data Science at Newcastle University (UK), Director of Postgraduate Research in the School of Computing, and was a Fellow of the Alan Turing Institute (2021-2023). She is the founder of the NoVA lab at Newcastle University, and has a background as postdoctoral data visualization scientist with Unilever R&D. Sara's research focuses on visualization of incomplete and heterogeneous data, mainly applied to health and biosciences.