# Tasks and Visualizations used for Data Profiling: A Survey and Interview Study

Roy A. Ruddle, James Cheshire, and Sara Johansson Fernstad

**Abstract**—The use of good-quality data to inform decision making is entirely dependent on robust processes to ensure it is fit for purpose. Such processes vary between organisations, and between those tasked with designing and following them. In this paper we report on a survey of 53 data analysts from many industry sectors, 24 of whom also participated in in-depth interviews, about computational and visual methods for characterizing data and investigating data quality. The paper makes contributions in two key areas. The first is to data science fundamentals, because our lists of data profiling tasks and visualization techniques are more comprehensive than those published elsewhere. The second concerns the application question "what does good profiling look like to those who routinely perform it?", which we answer by highlighting the diversity of profiling tasks, unusual practice and exemplars of visualization, and recommendations about formalizing processes and creating rulebooks.

**Index Terms**—Data profiling, data quality, survey, interview.

## 1 Introduction

Good-quality data has become an essential part of decision making, and is entirely dependent on robust processes to ensure it is fit for purpose. Data analysts therefore spend huge amounts of time performing the exploratory analysis required to characterize data (e.g., distributions) and assess its quality (e.g., missing values), which are known collectively as "data profiling" [1], [2], before it is used in detailed analysis and decision making.

The motivating hypotheses for our research are twofold. First, most analysts perform profiling in an ad-hoc manner, following an undocumented process that makes data profiling more an art than a science that adopts a rigorous and reproducible method. Second, visualization techniques are underused in profiling, perhaps due to a lack of formal education/training in visualization and knowledge of how to apply visualization for complex/large-scale data.

This paper reports on a survey with 53 data analysts, and follow-up interviews with 24 of them, about the computational and visual methods they use to characterize data and investigate data quality. The respondents worked in a range of industry sectors, on a wide variety of projects. The paper makes contributions across two broad areas. The first is to the fundamentals of data science, by providing lists of the tasks and visualization techniques used for data profiling, via input from a diverse set of practitioners. The second is in terms of applications and answering the question "what does good profiling look like to those who routinely perform it?" We found that data analysts are aware of their strengths and limitations, articulate the breadth of profiling tasks that experienced analysts adopt, and highlight exemplars and unusual practice in visualization. We make recommendations about formalizing profiling processes and creating rulebooks.

- *R.A. Ruddle is with the University of Leeds, Leeds, UK.*
- *J. Cheshire is with University College London, UK.*
- *S. Johansson Fernstad is with Newcastle University, UK.*

*Manuscript received Month xx, 20xx; revised Month xx, 20xx.*

## 2 Related Work

Data analysis may be subdivided into a five-stage process (discovery, wrangling, profiling, analysis and reporting) [1], [3]. Analysts may approach the profiling stage from two complementary perspectives – characterizing data and investigating data quality. For example, they may determine how the number of records varies with time or check whether records are missing from a given time-period. Similarly, they may calculate the distribution of data or determine if outlying values are implausibly high/low.

Previous research has classified the tasks that analysts perform, drawing on personal knowledge [4], literature reviews [2], [5], surveys & interviews [3], [6], and recording analysts' work [7]. In this section we summarize previous studies about three topics that are central to the present paper, namely surveys and interviews that investigated the work of data analysts, the types of task that are used in the two approaches to data profiling, and visualization techniques that are used during profiling.

### 2.1 Surveys and interviews

Researchers commonly use surveys and interviews to gather information about data analysis and usages of visualization. Surveys allow information to be gathered from more people, whereas interviews allow researchers to gather evidence first-hand and can explore topics in depth depending on interviewees' answers.

Kandel et al's landmark study [3] interviewed 35 people to understand difficulties they encounter during each stage of data analysis. The study had a broad scope, including data quality issues as part of profiling, but provided little information about how analysts investigated those issues. Other studies had a more specific focus, investigating how people describe data during wrangling [6], perform exploratory analysis after data has been profiled [1], use alternatives in their workflows [8] or how visualization usage differs between analysts vs. decision makers [9]. A study that, like ours, focussed on data profiling interviewed 13 analysts to produce a 10-item wish list for future tools, finding that interviewees were generally concerned about data quality issues but made little use of visualization [10].

Overall, the above studies provide rich descriptive framework of working practices across the stages of data analysis, and sometimes also the usage of tools (Python, D3, etc.) [1], [3], [8], [9]. However, with one exception [9] there is a lack of quantifiable detail about the tasks analysts perform and visualization techniques they use.

### 2.2 Profiling tasks

Data profiling is one of the first steps in the analysis pipeline. The scale of this task is highly dependent on the source of the data, its size, and its complexity. An analyst might for example expend significantly less effort profiling a well formatted dataset comprising 150 rows and 5 columns of official statistics in comparison to millions of records obtained from web scraping. It is impossible to manually check the latter so larger datasets require a suite of heuristics [1], [11], [12]. The results from this invariably surface some of "the many sources of data problems" referred to by Kandel et al. [3], which the analyst then needs to decide how to handle before the more substantive analysis can proceed.

The nature of the steps undertaken during profiling should also be determined by the objectives of the analysis. How these translate into specific tasks is captured by work that distinguishes between profiling tasks (characterized as "single-column", "multi-column" and "dependencies") and their primary use-cases (e.g., data management, integration, cleansing and analytics) [11]. However, these are not always clear to the analyst themselves since they may have been given the task without clarity and precision [13], or are working independently on a purely exploratory basis with no clear end in mind [14]. The diversity in both the characteristics of the data and the objectives of those analyzing it therefore makes data profiling a potentially rich seam of research, not least because there is no settled definition of the term "data profiling" itself [12] or its reach into the analysis pipeline and the composition of activities undertaken.

In their review of previous work Weiskopf and Weng [15] showed that tasks associated with data quality can be broken down into issues of completeness and correctness as well as concordance, plausibility and currency. For the purpose of this study we have selected the two most widely referred to: completeness and correctness. For the former tasks include counts of rows/columns, identifying missing values, and cardinalities whilst the latter might take the form of validation against "gold standard" datasets or establishing any bias.

### 2.3 Visualization

The ability of visualization to reveal the characteristics of a dataset is well known, and demonstrated by the likes of Anscombe's quartet [16] and the Datasaurus Dozen [17]. Exploratory visual analysis (EVA) is a subset of exploratory data analysis where visualization is the primary interface [18]. Several EVA tasks (e.g., characterizing distributions and understanding correctness) overlap with data profiling tasks. Nonetheless, visualization is often seen as a tool for communication rather than exploration among data analysts [19], and is prevented from becoming an integral part of exploratory data analysis due to visualization tools being separate from common data analysis tools, requiring substantial data wrangling, and lacking functionality for exporting visual findings. A range of tools for visual investigation of aspects of data quality and characteristics have been presented by visualization researchers [2], and commercial data analysis tools support a range of data profiling tasks and visualization [20]. Some of the most popular include Microsoft Power BI, Tableau, Alteryx, Trifacta and Qlik [21]. Some analytics tools (e.g., Alteryx and Trifacta) enhance their capability through integration of the wide variety of visualization provided by the likes of Tableau, Power BI and Qlik. This paper provides an overview of the main techniques, rather than a comprehensive review of them all.

The majority of visualizations for data quality analysis focus on tabular data [22]. For example, Profiler [23] combines data mining and summary visualizations, including histograms, bar charts, area charts, choropleth maps, binned scatter plots and small multiples. A related approach uses donut charts, bar charts and box plots to discover and correct outliers and missing values [24]. To broaden the utility of visual analysis in data profiling, Liu et al. [22] propose a framework for visual quality analysis for a range of data types, and suggest two types of visualization designs: summaries to display overview, patterns, distributions and constraints; and visualization for data error correction.

Several tools address quality issues in time series data using a variety of visualization techniques. TimeCleanser [25] provides semi-automated quality checks using line charts, bar charts and heatmaps. Visplause [26] utilises line charts, bar charts, histograms and tabular representations for hierarchical and summary visualization. "Know Your Enemy" (KYE) [27] supports quality assessment in time series data using heatmaps, histograms and tabular views. Gschwandtner et al. [25] also conclude that while analytical methods are preferred for easily defined quality issues, visualization makes it easier to identify more complex issues.

Missing values are a commonly mentioned quality issue. Several studies emphasise the importance of showing missing values with dedicated visual attributes and highlight the impact the choice of visual representation can have on the identification of missing value patterns and interpretation of the underlying data [28], [29], [30].

Combining the research above with a recent overview [31], a set of visualization techniques commonly used for data quality investigation can be extracted. These include area charts, bar charts, box plots, choropleth maps, histograms, line charts, pie charts, scatterplots, tree maps, heatmaps, small multiples and tabular representations. Furthermore, many tools make use of interactive dashboards with multiple views to facilitate analysis [23], [25], [26], [27].

In summary, while a large range of tools support visual data profiling, only a small number provide any guidance to the user as to what types of visualization to use and when. This may become an issue for data analysts with limited visualization experience when analyzing multivariate patterns and large-scale data. A first step in providing visualization guidance is therefore to understand its current use. Thus, this paper aims to provide empirical insight into

TABLE 1
The number of survey participants with each combination of job and experience as data scientist.

| Participant's job | Experience (years) | | | |
|---|---|---|---|---|
| | 0-1 | 2-5 | 6-10 | >10 |
| Academic faculty | | 2 | | 3 |
| Consultant | 1 | 2 | 1 | 1 |
| Data manager/architect | 2 | 1 | | 1 |
| Data scientist | 5 | 10 | 1 | |
| Data visualization | | 1 | 1 | |
| Management | 1 | 1 | 1 | |
| PhD student | 3 | 1 | | |
| Research software engineer | | 1 | 1 | |
| Researcher (academic) | | 3 | 2 | 4 |
| Researcher (industry) | | 2 | | 1 |
| TOTAL | 12 | 24 | 7 | 10 |

the tasks that data analysts perform to profile data and how they use visualization for those tasks. Doing so will have important implications for the development of profiling software, and also in developing a sense of common practice for those encountering a dataset for the first time [15].

## 3 METHODS

This section starts by providing information about the study's participants, and the method used for the survey and interviews. Then we describe how those two sources of data were analyzed.

### 3.1 Participants

Participants were recruited via news bulletins in our organizations, advertising at workshops and sending emails to professional contacts. The recruitment messages explained that we were "investigating how data scientists and analysts perform data profiling" and that "some participants will be asked to take part in a follow-up interview at a later date." The survey was completed by 53 people, who had a variety of jobs, range of data analysis experience and worked for 32 different organizations (see Table 1). Due to data protection the survey did not ask people for their age or gender. Participants were not paid. Instead, the authors undertook to share with them a practitioner's resource that is being written.

In the survey, people were asked "to consider your data profiling activities within a specific project that you are working on or recently worked on." The projects came from 16 industry sectors, ranged from less than a week to more than a year in duration (see Table 2), and involved widely differing numbers of records and fields (see Table 3).

We interviewed 24 of the survey respondents. They worked for 17 organizations, on projects that spanned a 10 industry sectors and involved a variety of scales of data.

Neither survey respondents nor interviewees were paid for their time. The study was approved by the Ethics Committee at the first author's university.

### 3.2 Survey

An online survey was created to gain a breadth of responses from those undertaking data profiling tasks. The responses also formed the basis to the follow-up interviews. The survey was launched in September 2019, with the majority of responses received in its first seven months, but other responses were received up until May 2021 due to delays recruiting the final interviewees.

The objective was to create a series of questions that were comprehensive, interpretable to those with a broad range of expertise and experience and that could be answered in 15-20 minutes. The survey (see supplementary material for a blank copy) started by providing information about the study, consent and confidentiality. That was followed by pages that asked for general information about the respondent (see Table 1) and their project (see Tables 2 and 3).

Page 5 used checkboxes to gather information about the tasks the respondent performed to characterize data in their project, subdivided into cardinalities, value distributions and patterns that each also had a free text 'other' option (see Table 4). This balanced the need for efficiency in how the survey is filled out, facilitating comparison between respondents and also offering useful prompts against the potential for being overly prescriptive in the answers we were seeking. Page 6 then gathered information about the visualization techniques (if any) the respondent used for data characterization, again with checkboxes and a free text option (see Table 5).

Page 7 used checkboxes and a free text option to gather information about the tasks the respondent performed to investigate data quality, subdivided into completeness and correctness (see Table 6). Page 8 gathered information about the visualization techniques (if any) the respondent used for data quality, providing identical options to Page 6.

For Pages 5-8, the lists of options were determined following a literature review (see Section 2), extensive discussions between the authors and also after reflecting on results from a pilot survey and initial interviews.

### 3.3 Interviews

Follow-up semi-structured interviews took place with 24 of the survey respondents, with the aim of gaining deeper understanding of the methods, tasks and challenges faced as part of data profiling. In particular, the interviews focused on detailing the tasks that analysts perform during the data profiling stage, and how they investigate data profiling issues related to data quality and characterization, which has not been the main focus of earlier studies [1], [3].

The interviews took place between September 2019 and July 2021, and were conducted by the three authors. The initial nine interviews took place in person, while the remaining fifteen were carried out online using Microsoft Teams. All interviews were recorded and sent for transcription prior to analysis. The participants were asked to sign or email a consent form before the interviews and received information about the aim of the project and interview, as well as the collection and use of data from the interviews. The length of the interviews varied between 15 and 66 minutes, with an average of 36 minutes.

The interviews were structured around the survey responses, with most questions directly referring to the project described by the respondent in the survey. They were designed as a series of open-ended questions that were separated into six main sections, as detailed below:

TABLE 2
The number of projects in the survey, for each combination of industry sector (from the UK Standard Industrial Classification) and duration.

| Industry sector of project | Project duration ≤ 1 week | 1 month | 6 months | 1 year | > 1 year |
|---|---|---|---|---|---|
| Accommodation and Food Service Activities | | | | | 1 |
| Activities of Extraterritorial Organisations and Bodies | | | 1 | | |
| Construction | | | 1 | 1 | |
| Education | | 1 | | | 4 |
| Electricity, Gas, Steam and Air Conditioning Supply | | | 2 | | |
| Financial and Insurance Activities | | 1 | | | 3 |
| Human Health and Social Work Activities | | 1 | 2 | | 5 |
| Information and Communication | 2 | 2 | | 1 | |
| Manufacturing | | 1 | | | 1 |
| Other (multiple sectors) | | 1 | | | |
| Other Service Activities | | | | | 1 |
| Professional, Scientific and Technical Activities | | | 1 | 1 | 6 |
| Public Administration and Defence; Compulsory Social Security | | | 1 | 1 | 3 |
| Real Estate Activities | | 1 | | | |
| Transportation and Storage | | 2 | 1 | | |
| Wholesale and Retail Trade; Repair of Motor Vehicles and Motorcycles | | 3 | | 1 | |
| TOTAL | 2 | 13 | 9 | 5 | 24 |

TABLE 3
The number of projects with each combination of number of records and fields.

| Number of records | Number of fields ≤ 25 | 50 | 100 | > 100 | I cannot say |
|---|---|---|---|---|---|
| 100 | 1 | | | 2 | |
| 10,000 | 3 | | 3 | 4 | |
| 1 million | 2 | 2 | 1 | 3 | 1 |
| > 1 million | 3 | 4 | 4 | 10 | 1 |
| I cannot say | 1 | 1 | | 3 | 4 |
| TOTAL | 10 | 7 | 8 | 22 | 6 |

1) **Constraints**: Included questions aiming to identify constraints that affect profiling ability, such as timescales, availability of expertise and the use and re-use of standard procedures.
2) **Profiling workflow**: This section aimed not only to check the correctness of the tasks and methods included in the survey response, but mainly to detail how profiling tasks are approached, in which order they take place, which methods are used for what task, and how repeatable the workflows are.
3) **Purpose of profiling steps**: Here the output, consequences, and issues of the different steps of the profiling workflow were discussed. This also included the definition of audience and stakeholders, and how the profiling outputs were shared with them.
4) **Workflow time**: This section addressed the overall time taken to carry out the workflow as well as the individual steps, aiming to identify the most time-consuming parts of profiling.
5) **Tools and techniques**: Covered questions related to which techniques were used for data characterization and data quality tasks, and what software that was used for visualization.
6) **Bottlenecks and pain points**: This section aimed to identify challenges in data profiling as well as potential solutions to these. It included discussion about recurring bottlenecks and pain points that interviewees had come across, such as issues with hardware and computer systems, automation, re-usability of code, as well as assumptions and simplifications made. Additionally, the interviewees were asked to reflect on features, tools and techniques that would help them solve these difficulties.

As part of the investigation of tasks and issues addressed by analysts, the interviews also aimed to identify exemplars of good and innovative practice in data profiling, in order to identify opportunities and make recommendations.

### 3.4 Analysis of the Survey Data

The survey captured factual information about each participant, the project they chose to discuss, the project's data, the data profiling tasks they performed and visualization techniques they used. Apart from questions such as name and experience, participants responded by selecting options from drop-down menus, radio buttons or check boxes. One participant responded that their project's clients were large (FTSE 100) companies, so that project is classified as "Other (multiple sectors)" in Table 2. Seventeen participants selected "other" for the task and/or visualization questions, and typed a free text response. The authors discussed those responses to agree how to incorporate them in the results.

Sixteen participants included "other" responses for the profiling tasks. Some of those responses mentioned tasks that the participant subsequently selected in later part of the survey (e.g. mentioning "missing values" in the cardinalities responses). Some responses described detailed tasks that were already encompassed by the survey's list of characterization and quality tasks, so we hand-edited the survey output to select appropriate tasks from the survey's list. Specifically, those detailed tasks were natural ranges of the variables as implied by data semantics, natural zero of variables, long tail distributions for log-scaling, etc. (all encompassed by the frequency measures task), artificial upper/lower limit (encompassed by ranges), discrete vs.

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3234337



TABLE 4
The data characterization tasks that were included in the survey.

| Cardinalities | Distribution | Patterns |
|---|---|---|
| I don't examine cardinalities | I don't examine distribution | I don't examine patterns |
| Number of distinct values | First digit | Character types |
| Number of rows in file (dataset) | Frequency measures (count, percent etc.) and/or histograms | Clusters |
| Value lengths | Mean, median and/or mode | Correlation |
| | Outliers | Cross tabulation |
| | Ranges (percentile, quartile etc.) | Curve fitting |
| | Variance, standard deviation, skewness and/or kurtosis | Data format |
| | | Data type (e.g. numerical, categorical, ordinal, sets, text etc.) |
| | | Example values |
| | | Precision (e.g. no. decimal places, no. significant figures etc.) |
| | | Principal Components Analysis |
| | | Trends |
| | | Value patterns |

TABLE 5
The visualization techniques that were included in the survey.

| | | |
|---|---|---|
| I do not use visualization | Geographical map | Network diagram |
| Area chart | Heat map | Pie chart |
| Bar chart | Histogram | Scatter plot |
| Box plot | Line chart | Tree map |
| Dashboard | | |

TABLE 6
The data quality tasks that were included in the survey.

| Completeness | Correctness |
|---|---|
| I don't examine completeness | I don't examine correctness |
| Coverage (e.g. temporal or geographic) | Accuracy |
| Duplicates | Bias |
| Missing records | Consistency |
| Missing values | Integrity |
| Rate of recording | Misleadingness |
| Recency | Noise |
| | Outlier |
| | Plausibility |
| | Use of default values |
| | Validity |
| | Variation |

continuous data (encompassed by precision), differences in distributions between clusters (encompassed by clusters), pseudo missing data, distribution of missing values, relationships between missing values (encompassed by missing values), relationships between missing values and clusters (encompassed by clusters/missing values), and miscoding (encompassed by accuracy). Three interviewees said they checked the number of columns/variables/dimensions but had not indicated that in their surveys, so we added the number of columns to their responses for our analysis.

Tasks that were listed but not encompassed by the survey were added manually to the outputs. Some of those were types of cardinalities (number of columns; units; number of zero values; number of infinite values; number of special values). The survey included principal components analysis (PCA) but some participants mentioned other related tasks (e.g., t-Distributed Stochastic Neighbor Embedding), so we grouped them all in a new task called primary features. The others (data structure; direct mapping) were types of dependencies [11], which is an aspect of data profiling that we omitted from the survey because it is at a higher-level than the other tasks and, therefore, is not included in Section 4.

Four participants included "other" responses for the visualizations they used to characterize data. After discussing the responses we added bubble chart, chord chart, matrix plot, parallel plots, Sankey diagram, sparklines and volcano plot as distinct visualization techniques. One participant responded phylogenetic trees which is a type of network visualization, and another responded time series graphs which are usually displayed as line charts. Those participants had also selected network visualization and line chart in their respective responses, so we did not expand the list of visualization techniques that we used to analyze the data. One participant responded map mashups, which is a method for integrating data and is therefore outside the scope of the research. Finally, a violin plot and a funnel plot were mention by one interviewee each, so we added those plots as visualization techniques for our analysis.

### 3.5 Analysis of the Interview Data

The interviews were professionally transcribed. Each author analyzed their transcripts by highlighting explicit references to tasks, visualization techniques and workflow stages in three colors (that made subsequent analysis easier), and cross-referencing the transcript with the interviewee's survey responses. The cross-referencing involved completing a spreadsheet for each interviewee to state the workflow stage(s) in which the interviewee performed each task and note the tasks(s) that were used as examples of each visualization. The stage was left blank if the interviewee had selected a given task/visualization in the survey, but not mentioned it in the interview. The opposite sometimes occurred, so if a task/visualization was mentioned in the interview but not the survey response then the task/visualization was added to the spreadsheet. Individual interviewees are referred to as **Ix** in the results.

We used affinity diagramming [8] to analyze the challenges and exemplars that interviewees described. Affinity diagramming involves: (1) generating sticky notes (small documented facts), and (2) organizing the notes into groups. We divided (1) into two parts. First, working with their transcripts, each author extracted or paraphrased text that described each challenge/exemplar and entered that text and a unique identifier into a spreadsheet. Then, in an online group working session the authors worked together to write and agree a short caption for each bottleneck/pain point that was suitable for a sticky note.

We divided (2) into four parts that also took place during the group working session, which lasted three hours. Using Google's Jamboard software we iteratively arranged the 105 stickies into 10 categories. Once we had reached a consensus about those categories we divided them into subcategories to produce the final diagram (see supplementary material), exported the categories/subcategories into a table, added a written explanation for each and then followed that up with a verbal explanation. Finally, we discussed and documented links between the subcategories.

## 4 Results

This section reports the tasks that participants performed to characterize data and investigate data quality, combining results from the survey and interviews. It then details an important theme - formal processes - that emerged from the interviews before describing how visualization was used.

### 4.1 Profiling tasks

All 53 survey respondents checked data quality with at least one completeness and one correctness task, and all except six respondents also performed at least one task for each type of characterization (cardinalities, distributions and patterns). The number of characterization vs. data quality tasks that respondents performed was significantly correlated, $r(51) = .68$ ($p < .01$). However, there was wide variation in the total number of tasks that were performed, with one respondent only performing five whereas, at the other extreme, another respondent selected all 38 that were provided as checkboxes (see Figure 1). That illustrates many data analysts lack a rigorous approach to characterizing data and investigating quality. On average, there was a slight increase in the number of tasks with respondents' increasing experience and the length of their projects, but that increase was small when compared with the differences between individual respondents (see supplementary material).

An in-depth analysis showed the pattern across tasks and between respondents (see Figure 2). Some tasks were performed by most respondents. At the other extreme, the first digit, curve fitting and misleadingness tasks were only performed by a small minority of respondents, in addition of course to the six tasks that were added during our analysis of the free-text responses about other tasks. The remaining tasks account for the greatest difference between respondents who performed a fairly comprehensive set of tasks (26 or more) vs. respondents who only performed a small set (12 or fewer tasks). Ten of the differentiating tasks (Value lengths; Ranges; Variance, skewness, etc; Correlation; Data format; Trends; Coverage; Noise; Outlier; Variation) were performed by 75+% of the comprehensive set respondents, but no more than a quarter of the small set respondents.

#### 4.1.1 *Characterizing data*

The interviews provided further insight into the contexts that different tasks were used for. Characterization tasks (cardinalities, distribution and patterns) were often used to get a first overview of the data. Interviewee **I22** checked "number of rows, numbers of unique sensors to give me an idea of the number of datasets I'm pulling from", and **I23** described that "you start off with just getting a feel for, depending on the type of data it is, you would basically look at individual statistics of each column … Look at means and spreads to get a feel for what's going on".

**I23** described the general aim of data characterization as "try to answer two questions: are there any problems in here, and can I ask the questions I think I need to ask as part of doing the analysis of this data". The connection between characterization tasks and data quality was also indicated by exemplars describing cardinality and distribution tasks in context of quality checking, with **I11** stating that "I would probably check the cardinalities first, because if those aren't right then I can't see anything else being right". These types of tasks were also used to ensure that the data fit what was expected, with **I10** stating that "having a look at, yes, the number of rows that we have is a very good first indication of just getting an idea of the size of the dataset and whether that seems realistic based on what we think should be in the data", and an exemplar from **I11** on the analysis of airport data "checking whether that … matches what you're expecting, because you're always expecting certain airports to be bigger than other". **I20** said that the "number of distinct values, number of rows and value lengths tends to be something that I look at pretty much up front because they can sort of determine the methods that we use."

Several exemplars combined frequency measures with outlier analysis to identify errors and examine the need of data cleaning: **I3** stated that "I would start with the frequencies or histograms to look actually for the outliers and then I would look specifically at those outliers to see, okay, why is this, is this a missing dataset or are these missing data or something else is wrong", and **I21** said that "when I do, for example, statistical analysis, there are some standard things I would do ... to understand ... if there is any outlier, for example, if there's something that we need to some data cleaning". **I13** does a lot of range checking because "most of our values are numerated". **I23** starts by looking at distributions of data to see whether it's normally distributed or skewed and "get a feel for what's going on."

The most common pattern task was data type, with **I7** emphasising "we definitely look into whether there's continuous values or category co-values. Do they have flag values? Do they have an order in the dataset? Is it ordinally valued or is it nominal values?". Some interviewees routinely calculate or visualize correlations to determine how columns of data are related to each other. Curve fitting was one of the least common pattern tasks, with exemplars indicating that it was mainly used at later stages of profiling. **I20** said that "I'd be certainly looking more at sort of general trends before I would be trying to formalize correlation and [curve] fittings".

#### 4.1.2 *Data quality*

Checking for duplicates, missing values and missing records was by far the most common data quality task (see Figure 2). **I4** mentioned duplicates in context of joining multiple data sources "when I was trying to link one table with another table by [case], I would check whether there are duplicates with joining tables". Several exemplars also highlight the use of duplicate checks as part of revealing errors in data

This article has been accepted for publication in IEEE Transactions on Visualization and Computer Graphics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2023.3234337



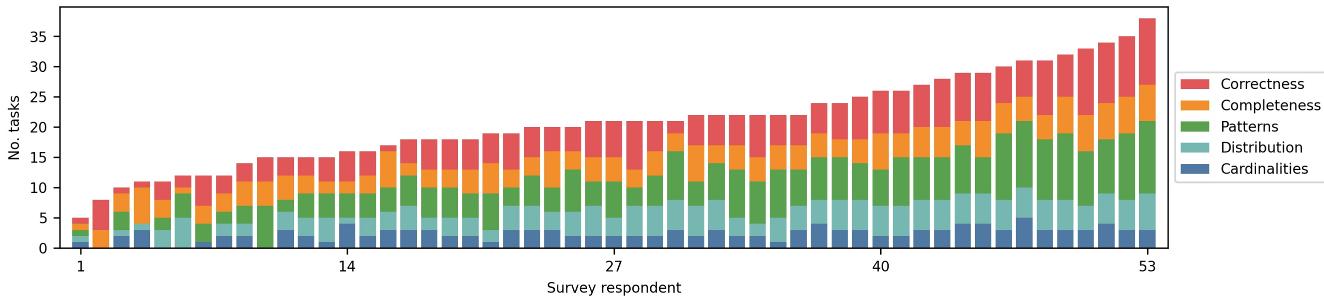



Fig. 1. The number of tasks from each group that survey respondents performed, ordered according to the total number of tasks they performed.

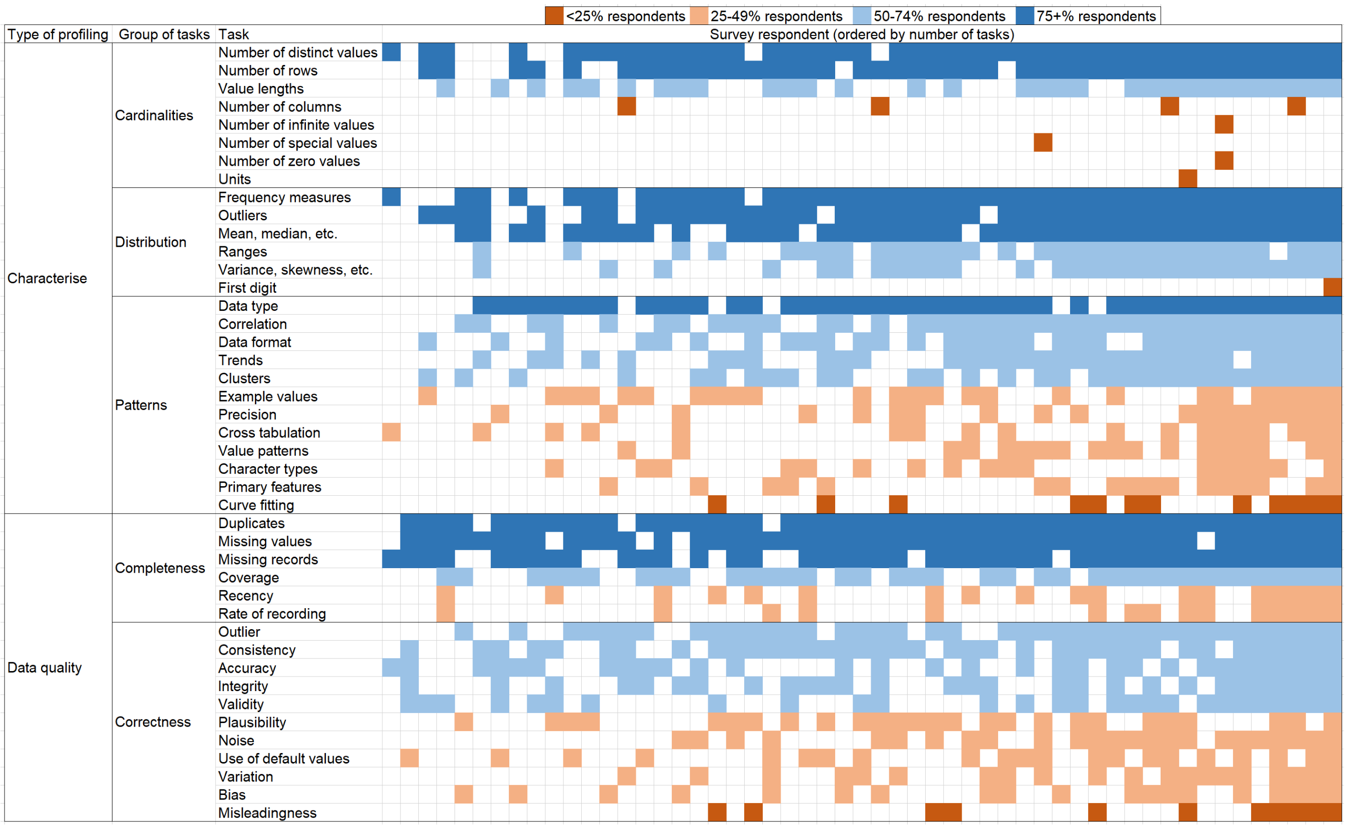



Fig. 2. The tasks that each survey respondent used to characterize data and investigate data quality. Task frequency is in color-coded four bands.

collection or recording: e.g. **I10** stated that "Often that means that there was a sensor with one name existed for some time period. Then they changed the sensor so they gave it a different name but they forgot to note that one had closed and the other one had opened or something so then they show up as duplicates."

When it came to missing values and records some exemplars emphasized looking for patterns and reasons for the missingness as an initial part of profiling. **I23** described "is there any obvious missing data? So are there rows that are missing? Are there columns that are missing? What fraction of a particular row is missing? What fraction of a particular column is missing? ... Does there seem to be, at least by eye, a pattern to the missingness?", and **I2** said "I really want to see, like, what are the reasons for the missing values?". This indicates the importance of exploring missingness patterns – beyond counts of missing values – in profiling pipelines, with previous work highlighting the benefit of visualization to understand such patterns [28], [30], [31]. Checking spatial and/or temporal coverage was also important, with **I6** saying that it is sometimes possible to insist that the data provider documents the coverage, or get one's money back if the data is not fit for purpose. **I8** flagged the importance of checking for records that had incorrectly been added to a dataset, thereby causing noise in the analysis.

**I6**, who deals with COVID-19 data, mentioned the challenge of missing data affecting which dataset they had to base analysis on: "imputation or interpolation of this data is something almost impossible, so we basically decided to, in a sense, base the analysis on the most, let's say reliable of these datasets, and provide also additional analysis for the remaining two that were less. . . that were identified as less reliable". In terms of ensuring data quality, **I15** pointed out the preference of removal rather than imputation of missing values, but there are risks of serious bias and loss of information if removal is applied to data where values are not randomly missing [32]. **I17** mentioned examining missing values in the context of adapting future data gathering: ”in fact one of the tests was to examine data in missing fields, and whether they’re key fields that should be mandatory as part of the process going forward”.

### 4.2 Formalization of Processes

The need to formalize processes in some way was seen as desirable by many of the interviewees who either expressed this as an exemplar aspect of the way they – or their team – perform profiling, or as challenge they see as essential but lacking. A range of rationales were given that coalesced around working more efficiently and/or gaining reassurance that an individual’s own efforts were on the right track. On the point of working more effectively it was clear that some felt there was too much “back and forth” between team members as initial efforts may have missed a crucial detail that a colleague had previously discovered in another iteration of the profiling process or that a full set of checks were not completed. The related point, therefore, is the degree to which an analyst felt confident in the profiling they had undertaken. Some felt limited by their own knowledge of a dataset and so see a formal process as a support, others felt frustrated by a lack of documentation around the data/databases they were working with and therefore had a sense that they were working inefficiently through problems already solved. The final aspect of this was that formalized processes – such as checklists – were cited as exemplars and could offer a sense of progression from the basic to the more complex visualisations associated with profiling, rather than going immediately to the most advanced aspects, which is an approach that may result in something being missed. There was also a sense that those more experienced in a specific dataset – or in data profiling more generally – were seen as having accumulated tacit knowledge that should be formalized as much as possible in order that others could then benefit from this experience within their own workflow.

The desire for formalization, however, was also tempered by concerns raised in the interviews about imposing too many restrictions in case they inhibit creativity or confine the approach to a “one size fits all” mentality especially amongst teams or analysts who work across a range of datasets. Related to the tacit knowledge, therefore, the formalization process may be something that evolves over time.

### 4.3 Uses of Visualization

Overall, 45 out of all 53 survey respondents used visualization to both characterize data and investigate data quality. There was no general pattern between the number of visualization techniques respondents used and their experience, the length of their projects (see supplementary material).

A more in-depth analysis (see Figure 3) showed the visualization techniques that each respondent used for the two types of profiling (characterizing data and investigating data quality). Respondents used from two to 14 different techniques, and the only one that was used by a large majority of respondents was a scatter plot for data characterization. Most of the visualization techniques were only used by a minority of respondents. Twenty-five respondents used interactive visualization for both types of profiling, whereas 11 respondents only used static visualizations.

One question that we asked interviewees was ”what tools and techniques do you use to characterize data and investigate data quality?”, with particular emphasis on visualization techniques. Time precluded an exhaustive discussion with each interviewee. However, we did discuss one or more profiling tasks for an average of 62% of the visualization techniques that each interviewee used (see Figure 4), and that provides some rich insights.

The three workhorses of visualization are scatter plots, bar charts and histograms, which were each used by 50+% users of survey respondents to characterize data and to investigate data quality (see Figure 3). Those visualizations were used for a wide variety of profiling tasks (see Figure 4) with **I5** commenting that “the initial stages are definitely much more about bar charts and . . . distributions”, scatter plots “to just validate relationships”, and **I7** visualizes correlations “because clients are very interested to see visuals rather than just numbers”.

Amongst the more unusual uses of histograms were “to check the completeness of data between different years” (**I18**), compare metrics such as the number of links and number of words for a full vs. sampled data that was being used to train a deep learning model, and **I10** found “big spikes” because “different types of sensors . . . sometimes use particular default values”.

Geographical maps were mentioned in the interviews for the greatest range of profiling tasks (see Figure 4). Some interviewees worked in global businesses; maps are important for **I17** to “characterize [data] into geographical location, or by their offices”, and **I11** uses “a map with . . . coloured circles as to whether the year on year was up or down for certain destinations” to filter data prior to more detailed analysis. Other interviewees used maps to investigate aspects of data quality such as outliers (**I5**: “looking at the context in which [a store] is sitting. For example . . . there is loads of competition around or something like that”), coverage (**I10**: “where are all of the sensors? . . . see that they’re covering the area we think they’re covering”), check origins and destinations (**I8**: “I had routes all over the UK which were not supposed to be there”), and “see if the [geographic] shapes look reasonable” (**I10**). A map was also important for showing context to **I2**, because they were analyzing data “about mobile networks and jamming in different cell sites.” **I10** used a map to check their analysis code (“find me all of the things of this type within, say, 500 metres of this point”), noting that “it’s nice to have a visual way to see ... the output”.

Line charts and box plots were workhorses for characterizing data, but less so for investigating data quality (see Figure 3). Line charts were used to show trends or time-series (**I2**: “we can just plot the number of events or number of measurements, or number of anything really”). Commenting how line charts and bar charts complement each other, **I11** said they used line charts to investigate general temporal patterns (e.g., to check the consistency of

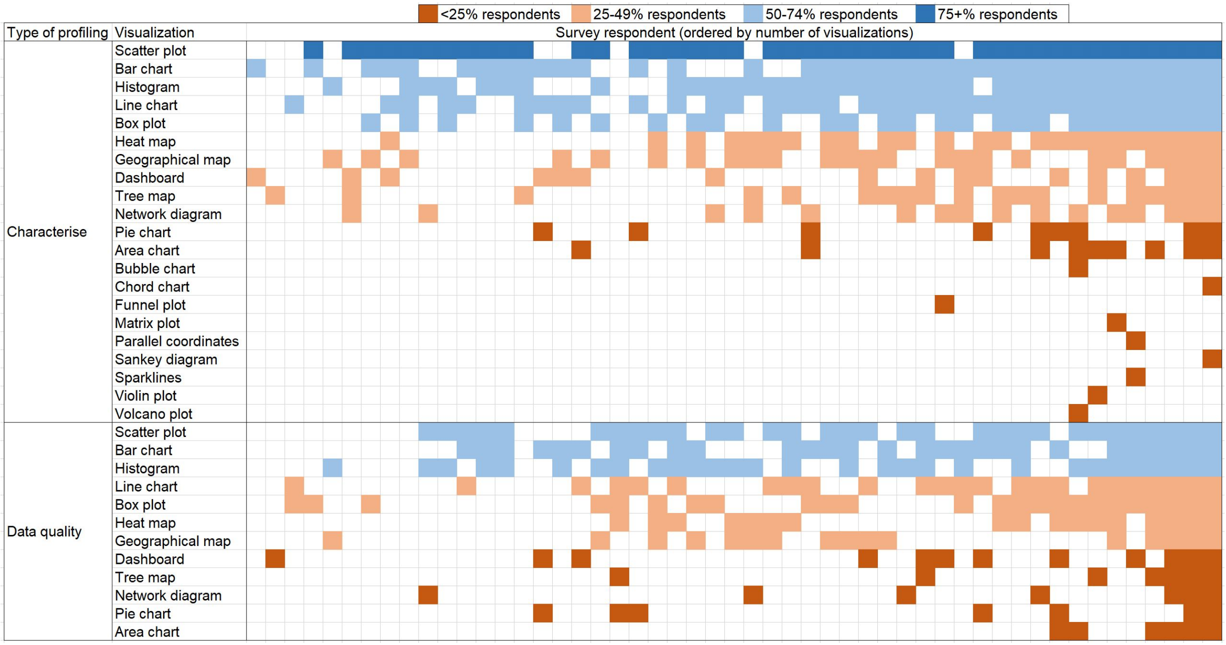



Fig. 3. The visualization techniques that each survey respondent used to characterize data and investigate data quality. Task frequency is in color-coded four bands.

the number of searches for flights over time), whereas "bar charts would be ones where you can't really do a line chart, so one example for that was, we have searches by the actual particular day that people are interested in flying."

Like histograms, interviewees used box plots and violin plots to show distributions (see Figure 4), and each technique has strengths and weaknesses. Histograms highlight unusually common values for specific intervals (e.g., a spike for default values; see above), but infrequent values have short bars, so are not salient and may even be invisible. Box plots explicitly highlight outliers, as well as the median, quartiles and range of a set of values, but hide detailed information about the values' distribution. Box plot limitations were captured by interviewees who use violin plots, **I20** saying "up until about six months or a year ago, a box plot would've been the go-to and I wouldn't have thought about anything else", and that they have replaced box plots by "violin plots just because a lot of the data that I end up dealing with is bimodal."

Heat maps were used by **I23** to "get missing values popping out", by **I21** to highlight "any out of range values", show correlations and more general associations (**I9**: "visualize topic vectors ... the colour would be how strongly each publication was associated with each topic"), show distributions (e.g., population in places vs. age ranges), and "to get a really quick visual" (**I17**) of numerical properties. An unusual use of heat maps was to visualize the location of fast- and slow-moving products in a warehouse, to recommend how the warehouse's layout should be changed to reduce congestion and speed-up picking activity. However, it should be noted that the term "heat map", which is in common usage by visualization researchers, is sometimes misunderstood – two interviewees thought it was where colour was used in a geographic map (e.g., **I8**: "to see what routes can get congested"; **I19**: "you want to see the load magnitude on each [mobile phone] tower, so you just colour code them").

Interviewees used dashboards both for themselves as analysts and for clients. One purpose was to provide overviews, automatically summarizing data streams (e.g., **I22**: "I would first go and have a look at the dashboard because that has the last 28 days from every single sensor on there") or reveal data that was odd (e.g., a very specific tumor type) and be able to select it. Client-facing dashboards were used to provide key performance indicators (KPIs) that "the user would like to view . . . daily or weekly" (**I12**) or, more generally, "when I realise actually there's a discussion to be had with stakeholders" (**I20**). Also, interactive dashboards helped make models understandable, without users having to read programming code.

Network diagrams come in many shapes and forms, but were rarely used by interviewees and survey respondents. Three exceptions were as a phylogenetic tree that provided structure and sorting functionality for cancer data, to show the flow of people in the context of a map, and to cluster data. A fourth was to compare the appearance of full vs. sampled datasets of websites "as a measure of how well we were doing on the way with the sampling'' (**I14**). That is not a safe approach, because the comparison could be affected by all manner of perceptual distortions.

Pie charts are much maligned in the visualization research community but, as with all techniques, can be appropriate and effective. Examples were interviewees checking whether data were distributed equally across categories, and comparing the number of patients who came from each year in longitudinal data analysis (using pie charts, **I18** "found lack of data in certain years").

Tree maps were discussed in five interviews and the overriding finding was that, like heat maps, the term is sometimes misunderstood by data analysts. **I5** (a geogra-

| Type of profiling | Group of tasks | Task | Area chart | Bar chart | Box plot | Bubble chart | Chord chart | Dashboard | Funnel plot | Geographical map | Heat map | Histogram | Line chart | Matrix plot | Network diagram | Parallel coordinates | Pie chart | Sankey diagram | Scatter plot | Sparklines | Tree map | Violin plot | Volcano plot | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Characterize | Cardinalities | Number of distinct values | | | | | | | | | | 2 | | | | | 1 | | | | | | | 3 |
| | | Number of rows | | 2 | 1 | | | 2 | | 3 | 1 | 1 | 2 | | | | | | 1 | | | | | 13 |
| | | Value lengths | | | | | | | | | | 1 | | | | | | | | | | | | 1 |
| | Distribution | Frequency measures | | 6 | 5 | | | 2 | | 4 | 3 | 13 | 3 | | 1 | | 1 | | 1 | | | 1 | | 40 |
| | | Mean, median, etc. | | | 3 | | | 1 | | 1 | 1 | | | | | | | | | | | | | 6 |
| | | Outliers | | 2 | 5 | | | 2 | 1 | 3 | | 3 | | | | | | | | | | | | 16 |
| | | Ranges | | | 4 | | | | | | 1 | | | | | | | | | | | | | 5 |
| | | Variance, skewness, etc. | | 1 | | | | | 1 | | | 1 | | | | | 1 | | 1 | | | | | 5 |
| | Patterns | Clusters | | | | | | | | 5 | 2 | | | | 1 | 1 | | | 3 | | | | | 12 |
| | | Correlation | | | | | | | | | 3 | | 1 | | 1 | 1 | | | 5 | | | | | 11 |
| | | Curve fitting | | | | | | | | | | | 1 | | | | | | 1 | | | | | 2 |
| | | Primary features | | | | | | | | 1 | | | | | | | | | 1 | | | | | 2 |
| | | Trends | | 1 | | | | | | 1 | | | 2 | | | | | | 1 | | | | | 5 |
| | | Value patterns | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| Data quality | Completeness | Coverage | | 1 | | | | | | 5 | | 1 | | | 1 | | | | | | | | | 8 |
| | | Missing records | | 1 | | | | 1 | | 1 | | | | | | | | | | | | | | 3 |
| | | Missing values | | 2 | | | | | | 1 | 1 | 1 | | | | | | | | | | | | 5 |
| | Correctness | Accuracy | | | | | | | | | | | | | 1 | | | | | | | | | 1 |
| | | Bias | | | | | | | | | | 1 | | | | | | | | | | | | 1 |
| | | Consistency | | 2 | | | | | | | | | 1 | | 1 | | 1 | | 1 | | | | | 6 |
| | | Integrity | | | | | | | | | | | | | 1 | | | | | | | | | 1 |
| | | Noise | | | | | | | | 1 | | | | | | | | | | | | | | 1 |
| | | Outlier | | | | | | | | | | 1 | 1 | | | | | | | | | | | 2 |
| | | Plausibility | | | | | | | | 2 | | | | | | | | | 2 | | | | | 4 |
| | | Use of default values | | | | | | | | | | 1 | | | | | | | | | | | | 1 |
| | | Validity | | | | | | | | 1 | | | | | | | | | 2 | | | | | 3 |
| TOTAL | | | 0 | 18 | 18 | 0 | 0 | 8 | 2 | 30 | 12 | 26 | 11 | 0 | 7 | 2 | 4 | 0 | 19 | 0 | 0 | 1 | 0 | |

Fig. 4. The number of interviewees that mentioned each combination of data profiling tasks and visualization techniques in our discussions. 16 tasks were not mentioned in the interviews, so have been omitted. Of the 8 visualization techniques that were not linked to specific profiling tasks, only one (area chart) was included in the survey whereas the others were provided in participants' free-text responses.

pher by training) seemed to assume that "tree map" referred to a type of geographic map, whereas **I8** talked about hierarchical clustering and, therefore, seemed to think the term tree map referred to a tree network like a dendrogram. Other interviewees said they used tree maps to provide "quick summaries" of data (**I23**), or that they only used tree maps with certain audiences (**I1**). The most interesting usage was **I18** making "a tree map of the tables and the dependencies", because they had inherited a set of data tables and processing scripts and needed "to actually figure out in what order I should run those scripts".

Parallel coordinates, spark lines and funnel plots were each only mentioned in one survey response, but in each case the respondent was amongst the interviewees. Parallel coordinates were used when there were a lot of variables, with **I23** saying that they often use "dimension reduction" but "I don't use parallel plots [sic] as much as I probably ought to." **I23** also sometimes used spark lines to automatically provide "quick summaries" of data, but did not describe the profiling tasks that those summaries supported. **I13** used funnel plots to visualize infant mortality vs. the number of admissions for about 50 hospitals, and to alert analysts to hospitals where the data needs detailed investigation because the mortality measure is higher than expected, after allowing for a confidence interval.

## 5 Discussion

Previous studies [1], [3] provided the conceptual framework for data analysis and uses of visualization. By taking a narrower focus our study provides much greater depth about the data profiling aspect of analysis. The tasks in our survey were refined and slightly expanded by responses from 53 practitioners, to provide a final list of 44 profiling tasks (see Figure 2) that is more comprehensive than those published elsewhere [11], [12], [15], [23]. The practitioners worked on projects spanning 15 industry sectors, but we did not detect differences between sectors in terms of the survey responses and interviews. That said, it is possible that some sectors (e.g., health) tend to apply more advanced profiling methods, which future work could investigate by focusing on specific domains, each with a large number of participants to mitigate individual differences.

The survey responses combined with important nuance from the interviews helped us identify common practice that also answers the question "what does good profiling look like to those who routinely perform it?" The survey made it clear that there are a few tasks (e.g., distinct values and missing value checks; see Figure 2) that are almost universally performed and a larger number of perhaps more complex tasks undertaken by fewer respondents. Interestingly other work captures well the nature of the tasks but recommends "strong co-operation between the database community ... and the visualization community" [11], implying that those creating the metadata and data visualizers are different. Many of our interviewees take care of the entire profiling process, and the fact that they are generalists helps to further inform our suggestions below. Most (30 out of 53) respondents performed fewer than half of that list of 44 tasks, so the quickest win for analysts who aspire to good profiling is simply to perform a more comprehensive set of tasks. Examples include ranges to sense-check numerical data (e.g., the weight of a child) and correlations to determine how columns are related.

We hypothesized that visualization is underused in profiling, and our findings showed that there was notably less usage for investigating data quality than to characterize data. On average 32% fewer respondents used each of the 12 visualization techniques listed in the survey for data quality than characterization, and respondents used fewer different visualizations (an average of 4.1 for data quality vs. 6.0 for characterization). Interviewees gave three times more examples of applying visualization to characterization than to data quality. Of course, visualization brings little added benefit for some profiling tasks (e.g., data type), and small datasets can be checked by hand but our respondents typically used quite large ones. However, despite the range of tools with advanced visualization features for profiling, interviewees stated concerns about their own skills and abilities suggest much more can be done to enable usage of such tools. This finding is also supported by an industry-wide survey of data visualisation practitioners conducted by the Data Visualization Society (DVS), who identify the need for skills and training as a theme in their 2021 report [33]. A difference between our survey and that report is that the latter is slightly finer-grained (e.g., network visualization techniques are separated into flow charts, force-directed graphs, dendrograms and network diagrams).

We also hypothesised that profiling is typically somewhat ad-hoc. This was supported by the often limited range of tasks utilised by survey respondents out of the wide range deployed overall, and also by interviewees highlighting the formalized (or automated) approaches they had in place whilst expressing regret about not doing more. Thus, formalizing as much as possible is the final element of good profiling. Routine tasks (e.g., basic descriptive statistics) can be easily automated so they may be recalculated from data updates. So too can more advanced analytical and plotting steps that are regularly used and have proven their worth with a particular dataset. In this latter case a few interviewees felt comfortable with a "black box" approach, where they may not have the skillset to perform the analysis but did feel confident in the interpretation and reporting of it. This has the dual advantage of ensuring consistency but also efficiency to liberate time for the more advanced – and creative – aspects of the profiling pipeline.

Interviewees had a desire for more guidance on what constitutes common practice for their workflows, which dovetails with previous work into frameworks/exemplars for checking common issues in data [34], [35] and narrative schema for the visualisation process [36]. That said, the need for flexibility – or rather not too much rigidity – was something our respondents cited as a reason not to over-formalise and this too was echoed by the 2021 DVS survey where 38% respondents cited "lack of customisation, flexibility, or versatility" as the biggest challenge when working with tools selected by others in their organisation [33]. Thus, a general challenge is facilitating creativity while also responding to sentiments in the interviews around needing more scaffolding to support day-to-day profiling activities. That is where checklists promote reproducibility and help to ensure that nothing is missed in the profiling process. Items on the list can be tightly defined or relatively open ended where appropriate, and might be best articulated as questions to answer about the data, rather than descriptive steps. This might mean "check for outliers" is better articulated as "does the dataset contain outliers?" and accompanied with a rulebook or tips for how to establish the answer. This encourages an active rather than passive engagement with the process and enables space for creativity whilst ensuring consistency. We see the rulebook as something that analysts can consult to ensure their practice aligns with those of their team or industry standards more broadly. The rulebook should provide exemplars (e.g., see Section 4.3) to be a useful reference to what is and is not appropriate whilst encouraging innovation. Crucially this should extend to the interpretation of the graphical/statistical outputs since some interviewees expressed concerns about how best to do this, not just how to create a visual output in the first place.

## 6 Conclusions and Future Work

This paper details the results from both a comprehensive survey and as series of follow-up interviews, which sought to establish the current practices and desires of analysts engaged in data profiling. Whilst respondents were drawn from a range of industries, backgrounds and experience levels, consistent themes emerged that both inform the current state of profiling and visualization practice as well as offer guidance.

The elements of that recommended practice are: (1) perform a more comprehensive set of profiling tasks, (2) greater and more varied use of visualization, and (3) formalize profiling via increased automation and the creation of rulebooks with tips and exemplars for guidance. The first could be adopted immediately with practitioners using our list of 44 tasks as the basis, whereas the other two require further work to develop suitable materials and stimulate adoption.

Finally, the combination of survey and follow-up interviews have proved crucial to garner the breadth of approaches but also the depth of insights required to determine the motivations for their use. The thematic areas that emerged through the affinity diagramming of interview themes were consistent with the wider survey responses, giving us confidence that our findings and conclusions were relevant to both research and practitioner communities.

## Acknowledgments

This research was supported by the Alan Turing Institute.

## References


[1] S. Alspaugh, N. Zokaei, A. Liu, C. Jin, and M. A. Hearst, "Futzing and moseying: Interviews with professional data analysts on exploration practices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 22–31, 2018.

[2] A. Crisan, B. Fiore-Gartland, and M. Tory, "Passing the data baton: A retrospective analysis on data science work and workers," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1860–1870, 2020.

[3] S. Kandel, A. Paepcke, J. M. Hellerstein, and J. Heer, "Enterprise data analysis and visualization: An interview study," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2917–2926, 2012.

[4] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations," in *The craft of information visualization*. Elsevier, 2003, pp. 364–371.

[5] M. Brehmer and T. Munzner, "A multi-level typology of abstract visualization tasks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2376–2385, 2013.

[6] A. Bigelow, K. Williams, and K. E. Isaacs, "Guidelines for pursuing and revealing data abstractions," *IEEE Transactions on Visualization and Computer Graphics*, vol. 27, no. 2, pp. 1503–1513, 2020.

[7] D. Cashman, S. R. Humayoun, F. Heimerl, K. Park, S. Das, J. Thompson, B. Saket, A. Mosca, J. Stasko, A. Endert *et al.*, "A user-based visual analytics workflow for exploratory model analysis," in *Computer Graphics Forum*, vol. 38, no. 3. Wiley Online Library, 2019, pp. 185–199.

[8] J. Liu, N. Boukhelifa, and J. R. Eagan, "Understanding the role of alternatives in data analysis practices," *IEEE Transactions on Visualization and Computer Graphics*, vol. 26, no. 1, pp. 66–76, 2019.

[9] E. Dimara, H. Zhang, M. Tory, and S. Franconeri, "The unmet data visualization needs of decision makers within organizations," *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[10] A. M. P. Milani, F. V. Paulovich, and I. H. Manssour, "Visualization in the preprocessing phase: Getting insights from enterprise professionals," *Information Visualization*, vol. 19, no. 4, pp. 273–287, 2020.

[11] Z. Abedjan, L. Golab, and F. Naumann, "Profiling relational data: a survey," *The VLDB Journal*, vol. 24, no. 4, pp. 557–581, 2015.

[12] F. Naumann, "Data profiling revisited," *ACM SIGMOD Record*, vol. 42, no. 4, pp. 40–49, 2014.

[13] A. Mosca, S. Robinson, M. Clarke, R. Redelmeier, S. Coates, D. Cashman, and R. Chang, "Defining an analysis: A study of client-facing data scientists." in *EuroVis (Short Papers)*, 2019, pp. 73–77.

[14] H. Estiri, T. Lovins, N. Afzalan, and K. A. Stephens, "Applying a participatory design approach to define objectives and properties of a "data profiling" tool for electronic health data," *AMIA Summits on Translational Science Proceedings*, vol. 2016, p. 60, 2016.

[15] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.

[16] F. J. Anscombe, "Graphs in statistical analysis," *The American Statistician*, vol. 27, no. 1, pp. 17–21, 1973.

[17] J. Matejka and G. Fitzmaurice, "Same stats, different graphs: generating datasets with varied appearance and identical statistics through simulated annealing," in *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 2017, pp. 1290–1294.

[18] L. Battle and J. Heer, "Characterizing exploratory visual analysis: A literature review and evaluation of analytic provenance in tableau," in *Computer Graphics Forum*, vol. 38, no. 3. Wiley Online Library, 2019, pp. 145–159.

[19] A. Batch and N. Elmqvist, "The interactive visualization gap in initial exploratory data analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 24, no. 1, pp. 278–287, 2017.

[20] S. Menon and E. Zaidi, "Market guide for data preparation tools." 2019, [Online]. Available: https://www.gartner.com/en/documents/3906957/market-guide-for-data-preparation-tools.

[21] "Data preparation tools reviews and ratings," 2022, [Online]. Available: https://www.gartner.com/reviews/market/data-preparation-tools, accessed 15 Aug 2022.

[22] S. Liu, G. Andrienko, Y. Wu, N. Cao, L. Jiang, C. Shi, Y.-S. Wang, and S. Hong, "Steering data quality with visual analytics: The complexity challenge," *Visual Informatics*, vol. 2, no. 4, pp. 191–197, 2018.

[23] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer, "Profiler: Integrated statistical analysis and visualization for data quality assessment," in *Proceedings of the International Working Conference on Advanced Visual Interfaces*, 2012, pp. 547–554.

[24] B. M. von Zernichow and D. Roman, "Usability of visual data profiling in data cleaning and transformation," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*. Springer, 2017, pp. 480–496.

[25] T. Gschwandtner, W. Aigner, S. Miksch, J. Gärtner, S. Kriglstein, M. Pohl, and N. Suchy, "Timecleanser: A visual analytics approach for data cleansing of time-oriented data," in *Proceedings of the 14th International Conference on Knowledge Technologies and Data-driven Business*, 2014, pp. 1–8.

[26] C. Arbesser, F. Spechtenhauser, T. Mühlbacher, and H. Piringer, "Visplause: Visual data quality assessment of many time series using plausibility checks," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 641–650, 2016.

[27] T. Gschwandtner and O. Erhart, "Know your enemy: Identifying quality problems of time series data," in *2018 IEEE Pacific Visualization Symposium (PacificVis)*. IEEE, 2018, pp. 205–214.

[28] C. Eaton, C. Plaisant, and T. Drizd, "Visualizing missing data: graph interpretation user study," in *Human-Computer Interaction-INTERACT 2005*. Springer, 2005, pp. 861–872.

[29] H. Song and D. A. Szafir, "Where's my data? evaluating visualizations with missing data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 914–924, 2018.

[30] S. J. Fernstad, "To identify what is not there: A definition of missingness patterns and evaluation of missing value visualization," *Information Visualization*, vol. 18, no. 2, pp. 230–250, 2019.

[31] R. Ruddle and M. Hall, "Using miniature visualizations of descriptive statistics to investigate the quality of electronic health records," in *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 5: HEALTHINF*. SciTePress, 2019, pp. 230–238.

[32] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

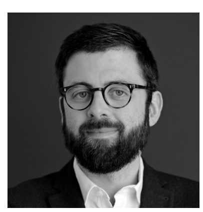
[33] D. V. Society, "Data visualization state of the industry survey," 2021, [Online]. Available: https://www.datavisualizationsociety.org/report-2021.

[34] N. Hynes, D. Sculley, and M. Terry, "The data linter: Lightweight automated sanity checking for ml data sets," in *NIPS MLSys Workshop*, 2017. [Online]. Available: http://learningsys.org/nips17/assets/papers/paper_19.pdf

[35] A. Shome, L. Cruz, and A. v. Deursen, "Data smells in public datasets," in *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, 2022, pp. 205–216.

[36] J. Wood, A. Kachkaev, and J. Dykes, "Design exposition with literate visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 25, no. 1, pp. 759–768, 2019.

**Roy Ruddle** is Professor of Computing at the University of Leeds. He has a multidisciplinary background, combining research and development in the software industry with a PhD in psychology. He conducts basic and applied research at the interface of computer graphics and human-computer interaction. His current research focuses on ultra-high-definition displays and visual analytic tools.



**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.

**Sara Johansson Fernstad** is Lecturer in Data Science at Newcastle University (UK). She has a PhD from Linköping University (Sweden) and was a PostDoc at Unilever R&D and Cambridge University (UK). Her research focus on visualization of incomplete, high dimensional and heterogeneous data, mainly applied to health and biosciences.