Final version availabe at SpringerLink : http://dx.doi.org/10.1007/978-3-319-69459-7_32

Usability of Visual Data Profiling in Data Cleaning and Transformation

Bjørn Marius von Zernichow and Dumitru Roman

SINTEF, Pb. 124 Blindern, 0314 Oslo, Norway
BjornMarius.vonZernichow@sintef.no,
Dumitru.Roman@sintef.no

Abstract. This paper proposes an approach for using visual data profiling in tabular data cleaning and transformation processes. Visual data profiling is the statistical assessment of datasets to identify and visualize potential quality issues. The proposed approach was implemented in a software prototype and empirically validated in a usability study to determine to what extent visual data profiling is useful and how easy it is to use by data scientists. The study involved 24 users in a comparative usability test and 4 expert reviewers in cognitive walkthroughs. The evaluation results show that users find visual data profiling capabilities to be useful and easy to use in the process of data cleaning and transformation.

Keywords: Data preparation, visual data profiling, usability testing, interactive data cleaning and transformation

1 Introduction

Data collection has become a necessary function in most large organizations both for record keeping and in support of different data analysis activities that are strategically and operationally critical [1]. In this context, proper data quality is a crucial aspect of extracting accurate information from data sources. Hence, incorrect or inconsistent data may distort analysis and compromise the benefits of any data-driven approaches. Examples of data quality issues, also labelled anomalies, include occurrences of missing, extreme, erroneous or duplicate values [2].

To illustrate the impact of poor quality data, IBM has estimated the yearly cost of inadequate data quality to be $3.1 trillion in US in 2016 [3]. Further, a recent survey [4] shows that data scientists spend 60% of their time on cleaning and organizing data, and 57% ranked this as a repetitive and tedious activity.

Considering the potential negative impact of poor data quality, there has been considerable research during the last decades, and different methods and tools have been proposed to cope with data cleaning [1]. Data cleaning is the process and technique of identifying and resolving missing values, outliers, inconsistencies, and noisy data, to improve data quality [5]. Closely related to data cleaning processes, additional data

transformation procedures, i.e. changing the data format while preserving the original meaning, are often required to improve data quality [5].

In the context of data cleaning, data profiling is the statistical assessment of datasets to identify quality issues such as potential outliers or missing values with the goal of achieving improved data quality [2]. Since determining what defines an error is context-dependent, human judgment is usually involved to determine whether the issues are actual errors and how the issues should be treated. The data quality assessment can be facilitated by a data profiling tool that performs statistical analysis [2], [5].

Visual data profiling is an extension of data profiling, achieved by supplementing statistical assessment of datasets with adequate data visualizations [2]. The integration of statistical analysis and visual analysis can reduce the time users spend on exploring and assessing data quality issues by providing constant real-time feedback on content and structure of datasets. Considering that data scientists use more than half their time cleaning and organizing data, and often find this activity tedious, visual data profiling approaches should be considered more often as it reduces the time and cost data scientists spend when addressing data quality issues.

The basic principle behind visual data profiling is to let the visual data profiling system perform the review of data quality and identification of data quality issues. The system collects statistics and information about the data, and then returns metadata that describes the quality of the data. Based on this information, the data scientist can make an informed decision about how any issues should be treated.

A recent example of a data cleaning and transformation framework is Grafterizer [6], [7], part of the cloud-based DataGraft[1] [8]–[10] platform. Grafterizer represents the state-of-the-art within data preparation research, supporting a wide range of cleaning and transformation operations. The framework provides an interactive user interface, and detailed specification and customization of transformation steps. Still, Grafterizer does not yet support visual data profiling that can ease the process of data cleaning, transformation, and improving data quality, for data scientists. Grafterizer provides research opportunities for evaluating usability of visual data profiling since the existing version serves as a benchmark in a comparison with the proposed prototype.

To address the problems with data quality, and time/cost consuming data preparation activities, we propose an approach that simplifies the data cleaning and transformation processes in Grafterizer, and reduces the effort spent on preparing data for analysis. We present a software prototype of the visual data profiling approach that features an interactive spreadsheet table view, suggestions for relevant data cleaning and transformation operations, and data quality feedback from a visual data profiling system. The goal was not only to create a prototype featuring the enumerated capabilities, but also to extensively evaluate it. To evaluate the usability of the approach and the prototype, a study was carried out that involved 24 users in a comparative usability test, and 4 expert reviewers in streamlined cognitive walkthrough sessions.

[1] https://datagraft.io

Key contributions of this paper include:

1. An *approach* to using visual data profiling in tabular data cleaning and transformation processes to improve data quality. The visual data profiling approach is realized by means of a prototype that includes features for identifying and visualizing data quality issues, i.e. missing values and outliers.
2. An *evaluation of the usability* of the visual data profiling approach by empirical validation of the prototype. A comparative usability study and expert reviews have been conducted to evaluate the usefulness and ease of use.

The remainder of this paper is organized as follows. Related work is presented in Section 2. Section 3 introduces the proposed visual data profiling approach. The implementation of the approach in a software prototype is presented in Section 4, and the evaluation of the approach and prototype is discussed in Section 5. Finally, Section 6 summarizes this paper and outlines avenues for future work.

2 Related Work

The development of the visual data profiling approach draws upon current research, and is inspired by existing solutions within the areas of data profiling technologies, visual analysis systems, and tabular data preparation approaches.

Profiler [2] is an example of a system for data quality analysis that includes data mining and anomaly detection techniques in addition to visualizations of relevant data summaries that can be used to evaluate data quality issues and possible causes. Profiler integrates statistical and visual analysis to reduce the time spent on data cleaning activities. The Profiler architecture and framework were developed by the former Stanford Visualization Group, now UW Interactive Data Lab. This team also developed Polaris [11] that evolved into the commercialized business and analytics software Tableau[2], and Data Wrangler [12] that together with Profiler merged into the commercialized data preparation solution, Trifacta[3].

The above-mentioned profiling solutions all originated in research environments, are well documented in research literature, and represent effective and user-friendly approaches to data profiling. Moreover, Talend[4] uses similar visual profiling techniques as Trifacta to automatically explore data characteristics and data quality issues. Talend focuses on ease of use and an intuitive user-interface.

In terms of usability testing of our visual data profiling approach, it would be challenging to use Trifacta or Talend as the system under test. First, it is difficult to isolate the data profiling capabilities from the data cleaning and transformation functionality. Hence, it would be problematic to know what is really evaluated. Second, the solutions are not open-source, and cannot be further developed to extend the existing version of Grafterizer.

---

[2] https://www.tableau.com
[3] https://www.trifacta.com
[4] https://www.talend.com/products/data-preparation

Generating visualizations from large data sets requires an understanding of users' needs and preferences along with knowledge of visual encoding rules and perception guidelines [13]. There are two general approaches to building a visual analysis system. First, considering visual encoding only will generate all possible valid visualizations without acknowledging the specific needs and preferences of users [14]–[16]. Second, introducing a visualization recommender system in a visualization pipeline [14]–[16] will potentially reduce the information overload of presenting all available visualizations. Tracking and storing information provided by the recommender system enables adaptation of the visualization system due to an evolving knowledge about which visualizations are valid and preferred by users [14].

Voyager [16] is an exploratory data analysis tool that is open-source, originated in research, and provides state of the art within open source data exploration. Voyager specifies visualizations through Vega-Lite [17], a high-level declarative JSON specification language based on Wilkinson's Grammar of Graphics [18], ggplot2 [19] and Tableau VizQL [11], [20]. Vega [21] is the underlying formal model for rendering Vega-Lite specifications. Our visual data profiling approach is inspired by Voyager, Vega, and Tableau, and implements a high-level declarative language to specify visualizations.

Microsoft Excel is a widely-used tool to prepare data for analysis and gaining insight into data. A central feature of Excel is the direct manipulation interface [22] where users can interact with the table to manipulate the dataset (e.g. selecting columns and/ or rows, right-clicking for options). The advantage of a direct manipulation interface, is that many users are already familiar with this interface, and less time is required to learn to use the tool. The proposed visual data profiling approach relies on the implementation of a spreadsheet-like table view for direct manipulation of data.

The proposed visual data profiling approach draws upon existing research to include capabilities for statistical profiling, suggestions for data cleaning and transformation operations, and a direct manipulation table. The approach differs from existing solutions by expanding the profiling capabilities with more relevant data quality feedback and visualizations for missing values and data distribution.

3 Proposed Approach

The requirements of the proposed approach emerge from needs of existing users of Grafterizer (Fig. 1), and as a research opportunity to propose an approach that will contribute to improving data quality in this context. Grafterizer provides state of the art functionality within data cleaning and transformation capabilities, but there is still a need for improving user experience by providing approaches that assist the users in achieving their goals of cleaning and transforming data. User feedback shows that Grafterizer has a steep learning curve, and is rather complex to use. Hence, novel approaches should be considered to provide useful functionality for improving data quality, and a user interface that is easy to learn and easy to use. Based on this exiting situation, our visual data profiling approach should provide the necessary statistical profiling capabilities that are needed to assist the user in identifying data quality is-

sues, and ease the process of improving quality. The visual data profiling capabilities should be integrated with a table view interface that lets the user manipulate columns and rows directly. Furthermore, the user interface should provide data cleaning and transformation functionality that is relevant to the user, and appropriately addresses the goals that the user tries to achieve. The applied data cleaning and transformation sequences should finally be reflected in a stepwise pipeline.

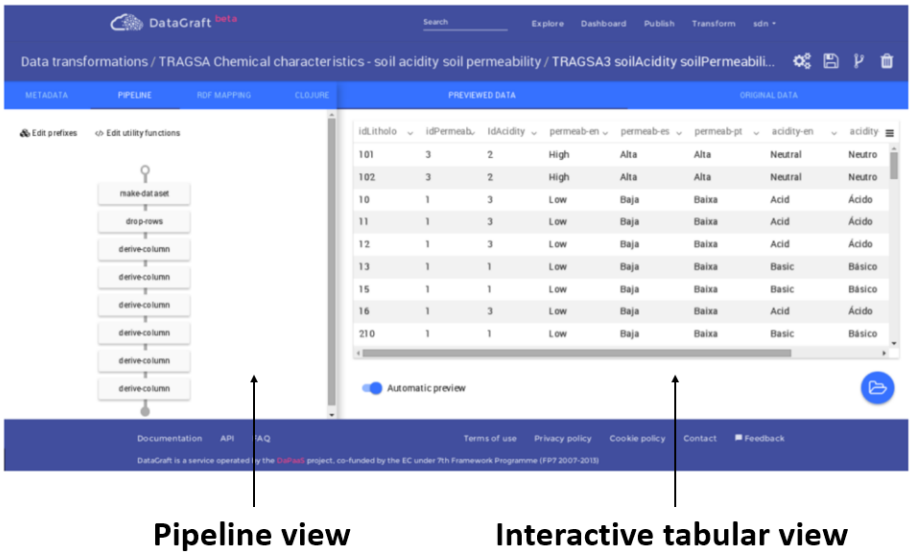

Fig. 1. Grafterizer user interface without visual data profiling capabilities

To facilitate the requirements process, a wireframe was created to describe the user interface and functionality, and the needs of users that led to a set of requirements. The wireframe outlined the basic graphical user interface components and functionality to resemble the final version of the application [23]. The wireframe was the first step to realizing the visual data profiling approach. Wireframes can be directly used in the implementation of the user interface of a prototype that supports the visual data profiling approach.

The user interface of the visual data profiling approach illustrated in the wireframe in Fig. 2, consists of the following main components and capabilities:

- A visual data profiling component (Fig. 2, component 1).
- A tabular table view that provides data cleaning and transformation functionality (Fig. 2, component 2).
- A sidebar that suggests relevant data cleaning and transformation actions to the user (Fig. 2, component 3).
- A steps pipeline that reflects applied data cleaning and transformation steps (Fig. 2, component 4).

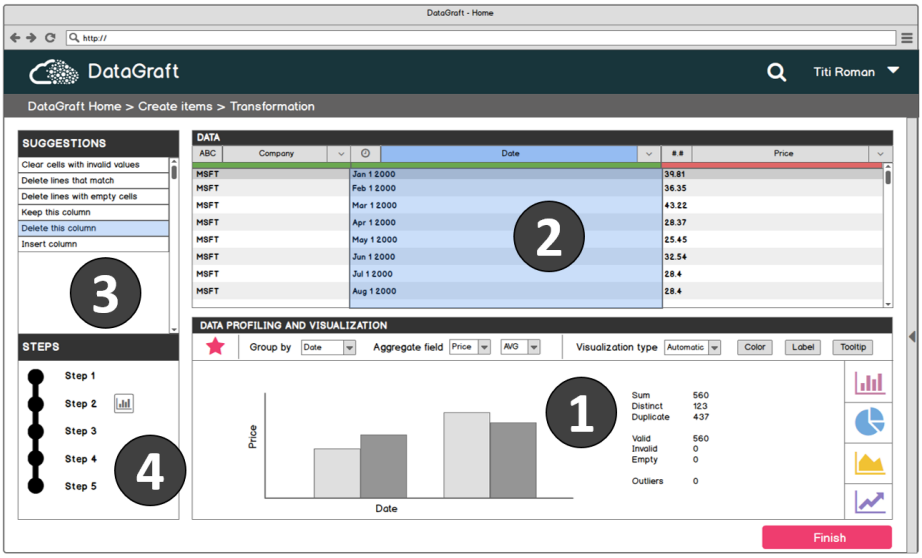

Fig. 2. Visual data profiling approach wireframe

The profiling assisted data cleaning and transformation process involves the following sequence of steps [2], [24], [25]:

1. **Discovery**: The user starts the data cleaning and transformation process by discovering the content, structure, and quality of the dataset. The visual data profiling system performs statistical assessment of data quality and returns the summarized feedback to the user.
2. **Cleaning and transformation:** Based on the statistical assessment of data quality, the user applies the appropriate procedures to clean the dataset, e.g. by correcting missing values. The dataset is further transformed to change shape into a desired format, e.g. by deleting a column.
3. **Validation**: Assisted by the data profiling system, the user validates the result of the applied cleaning and transformation procedures to ensure the output dataset has the intended content and structure.

4 Realization of the Proposed Approach in a Software Prototype

Prototyping is applied as an iterative design and development process to realize the concepts and requirements that are defined in the proposed visual data profiling approach [26], [27]. By prototyping, we always had something functional to test with users, collect feedback, implement changes, and then iterate. The prototype adds interactivity to the user interface wireframe, and provides functionality needed to demonstrate and validate the visual data profiling approach.

4.1 System Architecture

The high-level system architecture (Fig. 3) is based on a microservice architecture, and implements the design principles of Separation of Concerns (SoC) [28]. SoC is traditionally achieved in layered architectures, e.g. in a 3-Tier architecture, by defining interfaces and encapsulating information. A 3-Tier architecture would separate concerns into a presentation layer, an application tier, and a data layer.

A microservice architecture would take the SoC one step further by dividing the application tier and data layer into separate, domain-driven services that would operate autonomously from other services. A network-protocol would provide secure end point access to the services. While the SoC in a layered architecture is horizontal, the SoC in a microservice architecture would be both horizontal and vertical.

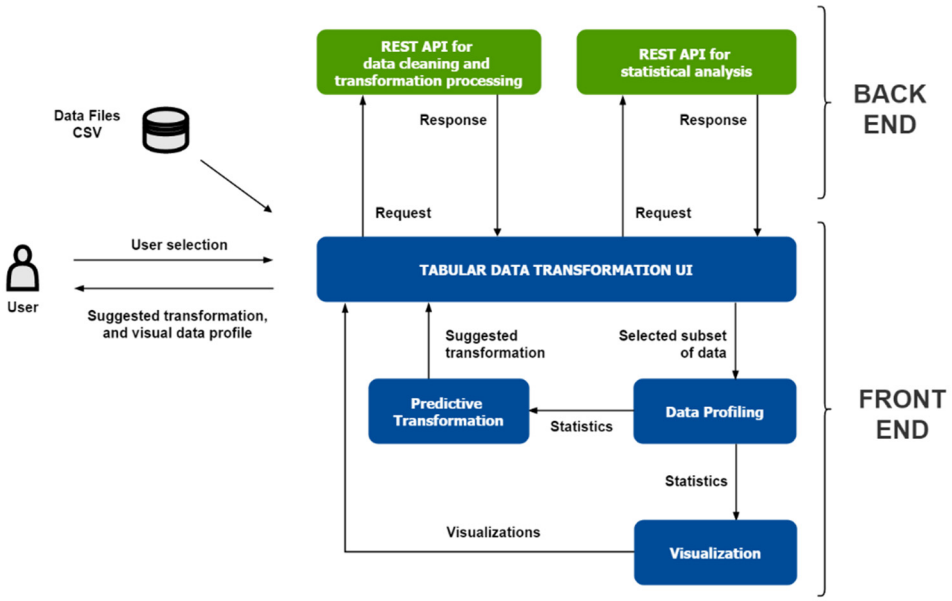

Fig. 3. Visual data profiling microservice architecture

4.2 Data Cleaning and Transformation Functions

The key functionality that was needed to evaluate usability of the visual data profiling approach is implemented in the prototype. The functionality is based on which data cleaning and transformation steps are needed to demonstrate and validate the visual data profiling approach in a user scenario developed by Statsbygg[5] and SINTEF[6]. The user scenario is named 'State of Estate', and is based on a dataset (reporting state-owned properties in Norway) that is cleaned and transformed by utilizing Grafterizer

[5] http://www.statsbygg.no
[6] https://www.sintef.no

[8]. Statsbygg is the Norwegian government's advisor in construction and property affairs, and serves as a building commissioner, property manager and developer. One of the purposes of cleaning and transforming the State of Estate dataset is to integrate information about public buildings in Norway with for example accessibility in buildings [8].

In total 14 data cleaning functions were defined and implemented in the prototype. Examples of functions include setting first row as header, replacing values, setting text to uppercase, concatenating values, and filling empty cells with a given value.

4.3 Implementation of the Software Prototype

The visual data profiling approach was implemented in a software prototype in four iterative stages. The final iteration of the prototype reflects the proposed functionality of the initial wireframe, and desired functionality to evaluate the usability of visual data profiling.

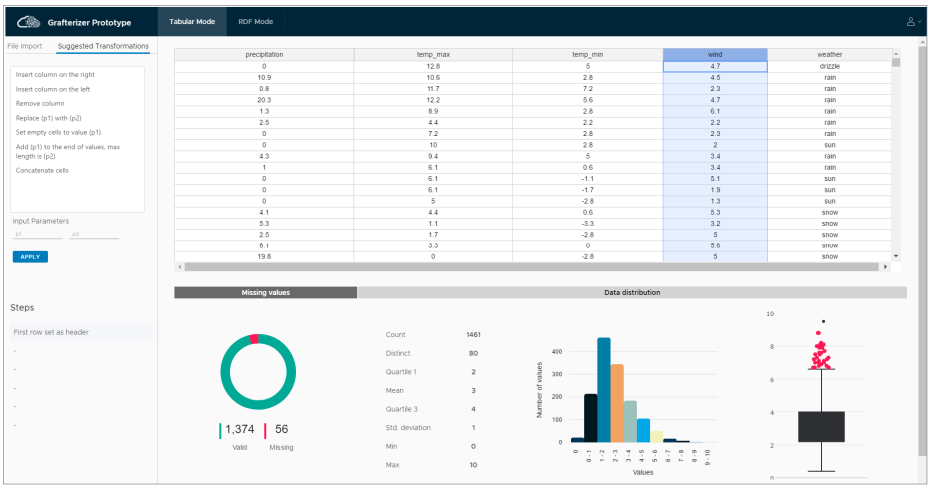

Fig. 4. Implementation of prototype, final iteration

The UI of the prototype depicted in Fig. 4 implements basic functionality of the following components: *Component 1*, the file import, is implemented for prototype development purposes only; *Component 2*, the table view, is a direct-manipulation table with Excel-like features such as right-clicking functionality (e.g. copy/paste, insert column/row); *Component 3*, the transformations sidebar, implements a rule-based system that suggests relevant data cleaning and transformation procedures; *Component 4*, the steps pipeline, displays a functioning pipeline that reflects all steps applied; *Component 5*, the visual data profiling service, features (from left to right) a data distribution chart, a chart that displays number of missing values, and basic measures of central tendency. The leftmost data profiling chart represents missing values and valid (non-null) values for the currently selected column. The three re-

maining charts (from left to right) represent the distribution of the currently selected column.

5 Evaluation: Usability Testing of the Software Prototype

5.1 Evaluation Methodology and Setup

In terms of usability and user acceptance of a system, it is essential that users believe that the system is useful and easy to use in order to adopt the technology [29], [30]. A user will consider a system to be useful if it enhances his or her work performance, and a system is easy to use if a user thinks that learning and using the system requires an acceptable amount of effort in terms of time and cost. Hence, a visual data profiling extension should not only provide the capabilities that the user needs, but the solution should also be considered useful in data scientists' work activities, and be easy to use [29]. We refer to these qualities as the usability of the visual data profiling system. The usability study addressed the following research questions:

- RQ1: How *useful* is the visual data profiling approach for users of tabular data cleaning and transformation tools?
- RQ2: How *easy to use* is the visual data profiling approach for users of tabular data cleaning and transformation tools?
- RQ3: Will the visual data profiling approach introduce usability issues in tabular data cleaning and transformation applications, and if so; which types of usability issues occur and how can they be corrected?

To understand users' experience with visual data profiling approaches, we have defined the typical users as data consumers, more specifically data scientists, that use data for data-driven decision making. The data scientist is an analytical expert that explores and analyses large volumes of data to solve complex problems and reveal business insights. Dedicated solutions for cleaning and transforming tabular data, e.g. Grafterizer, are often part of data scientist's toolbox.

We used two complementary methods of usability testing to evaluate whether users find the visual data profiling approach to be useful and easy to use.

A *comparative usability test*, survey based, was used to collect statistics and attitudinal data from users through an online questionnaire [31] which contains Likert-type rating scales. The test compared the prototype against the existing version of Grafterizer in terms of usefulness and ease of use. The survey was anonymized and voluntary, and only non-sensitive information was collected. A representative group of users was selected to participate in the survey. Voluntary participants from project meetings in current research initiatives with SINTEF were invited to participate in the comparative usability test, respond to the survey questionnaire, and provide qualitative feedback on the visual data profiling approach:

- EW-Shopp[7] (project meeting February 2017)

[7] http://www.ew-shopp.eu

- proDataMarket[8] (project meeting March 2017)
- euBusinessGraph[9] (project meeting May 2017)

We also used *streamlined cognitive walkthrough* as a usability inspection method where evaluators inspected the user interface by completing a set of tasks to simulate users' problem solving approaches [32]–[35]. The aim of this process was to identify usability issues introduced by the visual data profiling approach in data cleaning and transformation processes. In total four expert reviewers were selected to participate in the sessions. Users were divided in two subgroups and two corresponding sessions:

- Session 1: Two Human-Computer Interaction (HCI) experts from SINTEF Digital[10];
- Session 2: Two expert reviewers from the Logic and Intelligent Data (LogID) group at University of Oslo[11].

5.2 Analysis of Findings from the Comparative Usability Test

In total 24 participants responded to the survey questionnaire. The same users evaluated both the existing version of Grafterizer and the visual data profiling prototype, which defines the test setup as a within-subjects design [31]. The advantage of using this type of test design is that it removes some sources of variation in the datasets, as compared to between-subjects design where different users test each version of the application.

The online survey questionnaire[12] asked respondents to rate each application on the dimensions of *usefulness* and *ease of use*, respectively.

The summarized results from all respondents are illustrated in Fig. 5 and 6 below. The figures indicate the mean value of each question asked, e.g. the rating score of question Q1 in Fig. 5 shows the average of all 24 respondents' rating score on that specific question. High rating scores indicate that users perceive the application to be highly useful and easy to use, while the opposite is true for low scores.

8 https://prodatamarket.eu
9 http://eubusinessgraph.eu
10 http://www.sintef.no/en/information-and-communication-technology-ict/departments/networked-systems-and-services/human-computer-interaction-hci
11 http://www.mn.uio.no/ifi/english/research/groups/logid
12 https://goo.gl/forms/P3pD8zVPOj3uOSLT2

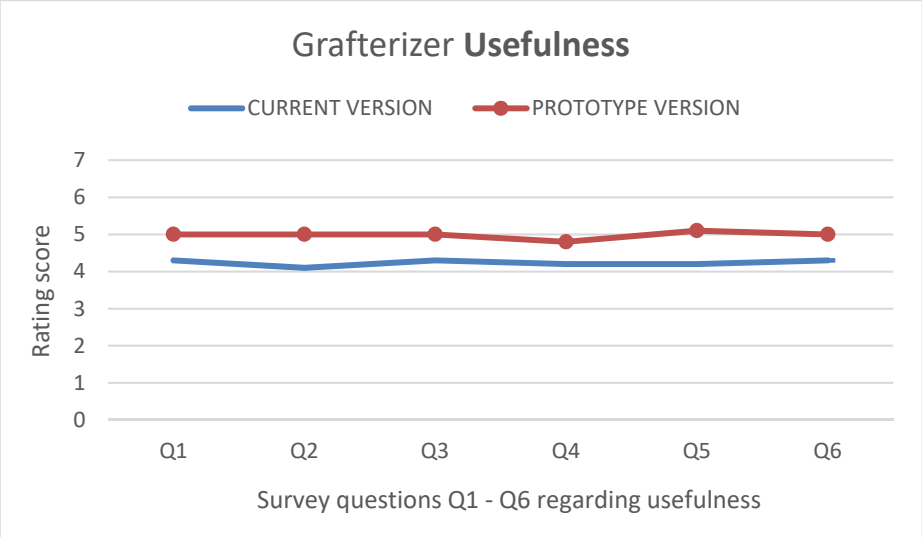

Fig. 5. Comparative usability test results (usefulness)

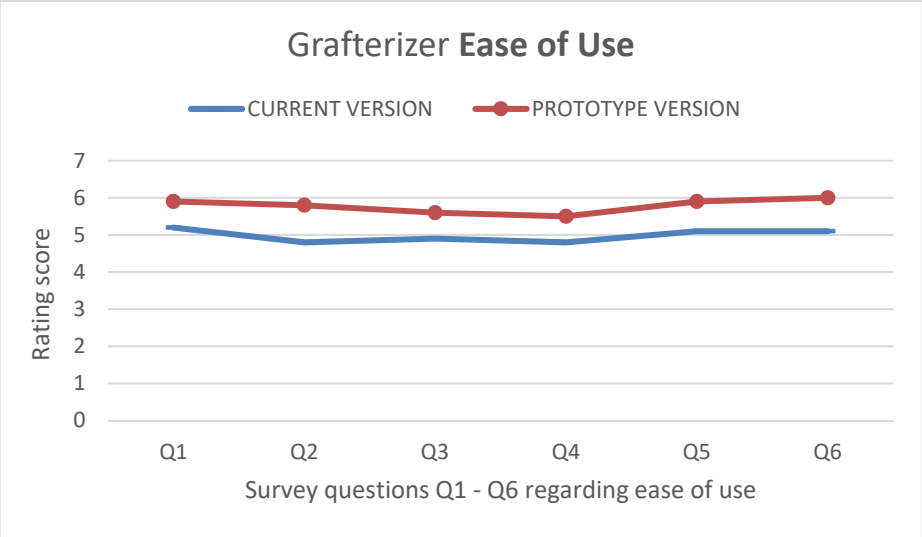

Fig. 6. Comparative usability test results (ease of use)

The results that are illustrated in Fig. 5 and 6 indicate that the visual data profiling approach consistently is rated higher than the existing version of Grafterizer on both dimensions of usefulness and ease of use. Still, it is insufficient to draw such conclusions based only on the kind of descriptive statistics [31] we find in Fig. 5 and 6. We need to determine if this difference between the applications is statistically significant, and if it is larger than we would expect from pure chance.

Since the usability test is a within-subject comparison of two applications, and the survey test results are continuous values, a paired t-test can be applied to appropriately determine if there is a significant difference between the mean ratings of the two applications [31]. The approach suggested by Sauro and Lewis [31] has been applied to compare the mean rating between the prototype and the existing version of Grafterizer.

We used a paired t-test to determine statistical significance of survey results for the *usefulness* dimension:

$$t = \frac{\widehat{D}}{\frac{s_D}{\sqrt{n}}}$$

Equation 1. Paired t-test

where

$\widehat{D}$	is the mean of the difference between the scores
s_D	is the standard deviation of the difference between the scores
n	is the sample size, i.e. the number of survey respondents
t	is the test statistic

Using the *t*-test to calculate the test statistic *t* of the values in Table 1 below, we get the following *t* value:

$$t = \frac{4.63}{\frac{4.48}{\sqrt{24}}} = 5.09$$

To determine whether the *t* value is significant, we use the TDIST function in Excel:

```
TDIST(t value, degrees of freedom, one-sided = 1 / two-sided = 2)
```

Equation 2. TDIST function

The degrees of freedom are equal to n – 1, and we use a two-sided test in the comparison. n = 24, which leads to the following calculation:

```
TDIST(5.09, 23, 2) = 0,000037
```

The calculations of statistical significance indicate that we can be approximately 99.999% sure that the prototype and the existing version have different scores, i.e. the difference is not due to chance. Hence, the prototype's rating score of 30 is statistically significantly higher than the existing version's score of 25.4. We can conclude that the users perceive that the prototype is more useful than the existing version of Grafterizer.

In terms of the *ease of use* dimension, the mean rating score for the prototype is 34.6, while the rating score of the existing version of Grafterizer is 30. The difference in rating scores is then 4.58, and applying the paired t-test leads to the conclusion that the rating score of the prototype is significantly higher than the existing version's score. The calculations of statistical significance indicate that we can be approximately 99.999% sure that the prototype and the existing version have different scores.

Based on the above analysis, we can conclude that the users perceive that the prototype is both more useful and easier to use compared to the existing version of Grafterizer.

Table 1. Survey rating scores, and difference, in terms of usefulness

Respondent	Prototype	Existing version	Difference
1	6	6	0
2	10	11	-1
3	24	24	0
4	34	35	-1
5	31	24	7
6	26	26	0
7	32	24	8
8	38	36	2
9	19	18	1
10	34	27	7
11	34	32	2
12	26	14	12
13	28	25	3
14	11	10	1
15	30	19	11
16	39	24	15
17	36	30	6
18	35	29	6
19	38	37	1
20	40	32	8
21	42	34	8
22	33	31	2
23	36	32	4
24	38	29	9
Mean	**30**	**25.4**	**4.63**

5.3 Analysis of Findings from the Cognitive Walkthrough

The two groups of expert reviewers went through user scenarios that were divided into tasks of the following format:

Task 1
I want to set first row as header.
Expert evaluation (questions answered by reviewers):
a. Will the user know what to do next?
b. Will the user get appropriate feedback if the correct action is taken?

The sessions resulted in an eight pages long document that describes the responses from the reviewers, and includes a discussion of the findings. To categorize and analyse the findings from the streamlined cognitive walkthrough sessions, a bottom-up approach [30] was used to organize and analyse the findings from the sessions. By using this method, we emphasize the advantage it provides by keeping the researcher open to the results the process will reveal. The method requires more time to organize and analyse than would a top-down approach that starts with predefined concepts, but this disadvantage is outweighed by the potential of identifying more usability issues.

The main findings from the reviews are summarized and categorized in Table 2 below. With each type of usability issue follows a suggestion on how the issue could be corrected.

Table 2. Identified usability issues and suggestions for further research

CATEGORY	USABILITY ISSUES	SUGGESTIONS FOR FURTHER RESEARCH
Visual data profiling	• Some of the charts are not domain specific enough. • The functionality and purpose of each visual data profiling chart are not clear. • Outlier detection and correction of missing values are too generic.	• Explore visual recommender system approaches to suggest relevant and domain specific charts to the user. • Explore approaches that include multivariate data profiling (i.e. by profiling two or more columns to reveal relevant information related to data cleaning and transformation).
'Excel' table view	• Missing information about data type of selected values. Lack of possibility to specify parameters directly in the table view.	• Explore direct table manipulation approaches to data cleaning and transformation to extend capabilities of the tabular table view.

'Suggested transformations' sidebar	• The sidebar is overlooked/ ignored in several cases because the suggested transformations are too generic and not specifically aimed at the current dataset. • Users also prefer to use the right-clicking functionality of the Excel-like table view.	• Explore approaches within predictive data cleaning and transformation, based on machine learning techniques, to provide more intelligent and relevant suggestions.

In terms of *learnability* of the visual data profiling approach, the expert reviews show that the system needs to recommend charts that are domain specific and relevant to the user. This improvement will probably increase the speed, and ease of use, of learning new and basic functionality to perform the specific data cleaning and transformation tasks. Advanced capabilities (i.e. clicking and zooming charts to display detailed information) are not intuitive, and should be considered moved up one level in the user interface hierarchy to be visible always (e.g. by providing access to detailed information in a drop-down menu). The expert reviews also identified a need for a more consistent pattern of visual data profiling sequences (e.g. every time a user clicks a table column, he or she would know what happens next in the visual data profiling view).

Furthermore, the table view and 'Suggested transformation sidebar' need to be consistent by displaying the exact same range of data cleaning and transformation options. Users were confused when only a subset of options were available when right-clicking the table view. The approach should also consider including a mode where the sidebar 'Suggested transformations' can be hidden on demand by the user to free up more space for the table view.

In general, the expert reviews indicate that users were satisfied with the immediate feedback that the visual data profiling approach provided. Feedback included information such as status of missing values, potential extreme values, and number of distinct values. Still, the partial lack of explicit feedback after clicking columns and rows of the table view, resulted in uncertainty about which parts of the dataset had been profiled. Hence, the visual data profiling approach should provide immediate feedback to the user by indicating which columns or rows have been selected, and indicate the data type of the values.

6 Summary and Outlook

With the increasing amounts of data in today's organizations and businesses, proper data quality has become essential to extract and analyse content from large volume data sources. Incorrect or inconsistent data may distort the results of analysis processes, and reduce the potential benefits of applying data-driven approaches in organiza-

tions. Furthermore, data scientists spend more than half their time on preparing data for analysis. Hence, there are considerable research opportunities to ease the process of data cleaning and transformation, and improve data quality.

As a response to the demand for solutions that improve data quality and reduce time spent on cleaning and transforming data, this paper proposes a visual data profiling approach that implements powerful visual data profiling capabilities. The visual data profiling approach has been evaluated in terms of usability, and found to be perceived useful and easy to use by users. Furthermore, critical usability issues have been identified and proposed as further work in future iterations of the prototype. We have also contributed to proposing a visual data profiling approach that can be further researched and implemented on the DataGraft platform to extend, or replace, the existing version of Grafterizer.

Future work includes the implementation of a visual recommender system for data profiling that can recommend relevant, personalized and domain specific visualizations to the user. Furthermore, the visual data profiling approach would benefit from combining a visual recommender system and an intelligent approach to the domain-specific data cleaning and transformation problem. Such a framework could relieve the burden of technical specification in a domain specific language, and guide the user through an incremental process of cleaning and transforming data.

Acknowledgements. The work in this paper is partly supported by the EC funded projects proDataMarket (Grant number: 644497), euBusinessGraph (Grant number: 732003), and EW-Shopp (Grant number: 732590).

References

[1] J. M. Hellerstein, "Quantitative Data Cleaning for Large Databases," *United Nations Economic Commission for Europe (UNECE)*, Feb. 2008.

[2] 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*, New York, NY, USA, 2012, pp. 547–554.

[3] T. C. Redman, "Bad Data Costs the U.S. $3 Trillion Per Year," *Harvard Business Review*, 22-Sep-2016. [Online]. Available: https://hbr.org/2016/09/bad-data-costs-the-u-s-3-trillion-per-year. [Accessed: 18-Mar-2017].

[4] "CrowdFlower | 2016 Data Science Report." [Online]. Available: //visit.crowdflower.com/data-science-report. [Accessed: 19-Mar-2017].

[5] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[6] D. Sukhobok *et al.*, "Tabular Data Cleaning and Linked Data Generation with Grafterizer," *ESWC (Satellite Events)*, pp. 134–139, 2016.

[7] D. Sukhobok, N. Nikolov, and D. Roman, "Tabular Data Anomaly Patterns," *To appear in the proceedings of The 3rd International Conference on Big Data In-*

novations and Applications (Innovate-Data 2017) 21-23 August 2017, Aug. 2017.
[8] D. Roman *et al.*, "DataGraft: One-Stop-Shop for Open Data Management," *To appear in the Semantic Web Journal (SWJ) – Interoperability, Usability, Applicability (published and printed by IOS Press, ISSN: 1570-0844)*, 2017.
[9] D. Roman *et al.*, "Datagraft: Simplifying open data publishing," in *ESWC (Satellite Events)*, 2016, pp. 101–106.
[10] D. Roman *et al.*, "DataGraft: A Platform for Open Data Publishing," *In the Joint Proceedings of the 4th International Workshop on Linked Media and the 3rd Developers Hackshop. (LIME/SemDev@ESWC 2016)*, 2016.
[11] C. Stolte, D. Tang, and P. Hanrahan, "Polaris: a system for query, analysis, and visualization of multidimensional relational databases," *IEEE Transactions on Visualization and Computer Graphics*, vol. 8, no. 1, pp. 52–65, Jan. 2002.
[12] S. Kandel, A. Paepcke, J. Hellerstein, and J. Heer, "Wrangler: Interactive visual specification of data transformation scripts," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 3363–3372.
[13] B. Mutlu, E. Veas, C. Trattner, and V. Sabol, "VizRec: A Two-Stage Recommender System for Personalized Visualizations," in *Proceedings of the 20th International Conference on Intelligent User Interfaces Companion*, New York, NY, USA, 2015, pp. 49–52.
[14] M. Voigt, M. Franke, and K. Meissner, "Using expert and empirical knowledge for context-aware recommendation of visualization components," *Int. J. Adv. Life Sci*, vol. 5, pp. 27–41, 2013.
[15] B. Mutlu, E. Veas, C. Trattner, and V. Sabol, "Towards a Recommender Engine for Personalized Visualizations," in *International Conference on User Modeling, Adaptation, and Personalization*, 2015, pp. 169–182.
[16] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer, "Voyager: Exploratory Analysis via Faceted Browsing of Visualization Recommendations," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 649–658, Jan. 2016.
[17] "Vega-Lite." [Online]. Available: https://vega.github.io/vega-lite/. [Accessed: 19-Mar-2017].
[18] L. Wilkinson, *The grammar of graphics*. Springer Science & Business Media, 2006.
[19] H. Wickham, *ggplot2: elegant graphics for data analysis*. Springer, 2016.
[20] J. Mackinlay, P. Hanrahan, and C. Stolte, "Show Me: Automatic Presentation for Visual Analysis," *IEEE Transactions on Visualization and Computer Graphics*, vol. 13, no. 6, pp. 1137–1144, Nov. 2007.
[21] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer, "Reactive Vega: A Streaming Dataflow Architecture for Declarative Interactive Visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 1, pp. 659–668, Jan. 2016.
[22] E. Bakke and D. R. Karger, "Expressive query construction through direct manipulation of nested relational results," in *Proceedings of the 2016 International Conference on Management of Data*, 2016, pp. 1377–1392.

[23] “The Guide to Prototyping Process & Fidelity,” *Studio by UXPin*. [Online]. Available: https://www.uxpin.com/studio/ebooks/prototyping-process-fidelity-guide/. [Accessed: 13-Apr-2017].
[24] J. Heer, J. M. Hellerstein, and S. Kandel, “Predictive Interaction for Data Transformation.,” in *CIDR*, 2015.
[25] S. Chen, “Six Core Data Wrangling Activities eBook,” *Trifacta*, 23-Nov-2015. .
[26] B. Hanington and B. Martin, *Universal methods of design: 100 ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport Publishers, 2012.
[27] “The Ultimate Guide to Prototyping,” *Studio by UXPin*. [Online]. Available: https://www.uxpin.com/studio/ebooks/guide-to-prototyping/. [Accessed: 13-Apr-2017].
[28] B. Familiar, *Microservices, IoT and Azure: Leveraging DevOps and Microservice Architecture to deliver SaaS Solutions*. Apress, 2015.
[29] F. D. Davis, “Perceived usefulness, perceived ease of use, and user acceptance of information technology,” *MIS quarterly*, pp. 319–340, 1989.
[30] C. M. Barnum, *Usability testing essentials: ready, set... test!* Elsevier, 2010.
[31] J. Sauro and J. R. Lewis, *Quantifying the user experience: Practical statistics for user research*. Morgan Kaufmann, 2016.
[32] J. Nielsen, “Usability inspection methods,” in *Conference companion on Human factors in computing systems*, 1994, pp. 413–414.
[33] R. Spencer, “The streamlined cognitive walkthrough method, working around social constraints encountered in a software development company,” 2000, pp. 353–359.
[34] T. Mahatody, M. Sagar, and C. Kolski, “State of the art on the cognitive walkthrough method, its variants and evolutions,” *Intl. Journal of Human–Computer Interaction*, vol. 26, no. 8, pp. 741–785, 2010.
[35] “Cognitive Walkthrough | Usability Body of Knowledge.” [Online]. Available: http://www.usabilitybok.org/cognitive-walkthrough. [Accessed: 10-May-2017].