RefDataCleaner: A Usable Data Cleaning Tool

Juan Carlos Leon-Medina and Ixent Galpin(✉)

Dpto. de Ingeniería, Universidad Jorge Tadeo Lozano, Bogotá, Colombia
{juan.leonm,ixent}@utadeo.edu.co

Abstract. While the democratization of data science may still be some way off, several vendors of tools for data wrangling and analytics have recently emphasized the usability of their products with the aim of attracting an ever broader range of users. In this paper, we carry out an experiment to compare user performance when cleaning data using two contrasting tools: RefDataCleaner, a bespoke web-based tool that we created specifically for detecting and fixing errors in structured and semi-structured data files, and Microsoft Excel, a spreadsheet application in widespread use in organizations throughout the world which is used for diverse types of tasks, including data cleaning. With RefDataCleaner, a user specifies rules to detect and fix data errors, using hard-coded values or by retrieving values from a reference data file. In contrast, with Microsoft Excel, a non-expert user may clean data by specifying formulae and applying find/replace functions. The results of this initial study, carried out using a focus group of volunteers, show that users were able clean dirty data-sets more accurately using RefDataCleaner, and moreover, that this tool was generally preferred for this purpose.

Keywords: Usability · Data wrangling · Data cleaning · Reference data

1 Introduction

While attempts have been made to automate, as much as possible, the data wrangling pipeline (e.g., [8,12]), in practice, these steps are most often done manually by experts. This is costly for the organizations involved, given that authors such as [17] and [19] indicate that anomalies are present in around 5% of data, and that an analyst spends 80% of his or her time in the preparation of data, and 20% in the analysis of this data once it has been cleaned and integrated [13]. Given the exponentially increasing volumes of data in the world, it is reasonable to conjecture that organizations may achieve significant savings if tools in a data scientist's pipeline may be effectively used by a broader range of people.

Indeed, the vendors of several tools in a data scientist's data processing pipeline now purport to further the democratization of data science. For example, Tableau conveys this vision through its advertising materials on its website [3], and Exploratory has the marketing slogan *Data Science is not just for*

H. Florez et al. (Eds.): ICAI 2019, CCIS 1051, pp. 102–115, 2019.
https://doi.org/10.1007/978-3-030-32475-9_8

Engineers and Statisticians. Exploratory makes it for Everyone [1]. Furthermore, recently usability workshops have emerged associated with conferences in the data management research community, e.g., HILDA[1] and IDEA[2] co-located with SIGMOD and KDD respectively. This reflects how *usability*, defined by the International Organization for Standardization (ISO) as "the ability of the software product to be understood, learned, used and attractive to the user, when used under certain conditions" [21], is now becoming an ever more important consideration by tool designers.

The problems related to data cleaning and integration encountered during the data wrangling process are varied and require domain expertise, as well as an understanding of issues such as functional dependencies and integrity constraints. Such concepts are not easy to grasp by non-expert users and, as such, there is the risk that tools may be applied incorrectly during this process. Thus, it is a challenge to design tools that are easy-to-use and prevent users from applying the tools incorrectly.

There has been relatively little research into the usability of tools used for data wrangling. In [9], a usability study is carried out of source selection approaches. This work differs from previous work in that it proposes and evaluates the usability of a data cleaning tool.

This paper describes RefDataCleaner, a usable tool to clean dirty data using reference data sets. We design and carry out an experiment in which users are asked to perform various data cleaning tasks using RefDataCleaner and the Microsoft Excel spreadsheet application. We chose Microsoft Excel as a baseline, given that it is a widely-used software tool by organizations throughout the world for a range of purposes, including for tasks for which it was not originally envisioned, such as data cleaning. The results of our experiment show that users perform better with RefDataCleaner for the purposes of diagnosing and repairing data errors. Moreover, we find that RefDataCleaner is preferred by users over Microsoft Excel, despite their increased familiarity with the latter.

This paper is structured as follows. Section 2 presents a brief background. Section 3 describes the RefDataCleaner application. Section 4 presents the experiment design. Section 5 reports the results obtained in the experiments. Finally, Sect. 6 concludes.

2 Background

Data errors may be classified in different ways [4,18] and several taxonomies have been proposed [10,15]. Müller *et al.* [14] classify errors into three groups: syntactic, semantic and contextual. Fan *et al.* [6] define categories of errors pertaining to consistency, duplication, accuracy, existence, conformance and integrity. The focus of this paper is on errors which can be fixed by using reference datasets. A *reference dataset* is a collection of correct and complete data items which make up a subset of the attributes in the dataset being repaired [11]. One such

[1] http://hilda.io/2019/.
[2] http://poloclub.gatech.edu/idea2018/.

example would be a lookup table with country names and the respective dialling codes, as used in the illustrated example in Sect. 3. Reference data is used extensively in organizations for data repair during data wrangling. The Colombian tax authority (DIAN) is one such organization, and the tool of choice for this is the Microsoft Excel spreadsheet application.

3 RefDataCleaner Application Description

This section describes RefDataCleaner, a web-based application that we developed using Shiny R which enables error detection and repair rules to be defined and applied to dirty data files. It supports both semi-structured and structured data sets, and operates over diverse file types, including Microsoft Excel, CSV, HTML tables, XML and JSON. We have made our source-code available on GitHub[3]. Furthermore, we have a demo version for readers to try at ShinyApps[4].

RefDataCleaner supports the application of two different types of rules, viz., substitution rules and reference rules. With a *substitution rule*, a user specifies one or more conditions that must hold for a data repair action to be triggered. A condition is a predicate involving an attribute data name, operand, and value, e.g., *country* = '*Colombia*'. The data repair action involves one or more assignments of attributes which are required for the data repair action, e.g., *dialling_code* ← 57. In essence, with this option every possible repair value needs to be hard-coded explicitly by the user, and is illustrated in Fig. 1.

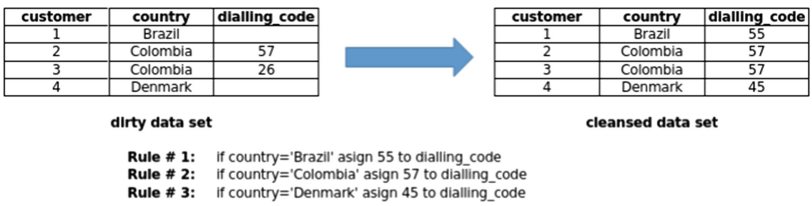


Fig. 1. Using substitution rules.

In contrast, in the case of *reference rules*, a repair is carried out using a reference dataset. For this example, the reference dataset comprises a complete set of records with a country attribute and the respective country code. Thus, for this type of rule, a user specifies a reference data set which can be used for data repair, one or more attributes to be used for an equi-join between the input data set and the reference data set, and one or more assignments of attributes from the reference data set to the input data set. Figure 2 shows an example whereby the *country_code* file is corrected based on reference data.

[3] https://github.com/refdatacleaner/version_1_0/.
[4] https://refdatacleaner.shinyapps.io/version_1_0/.

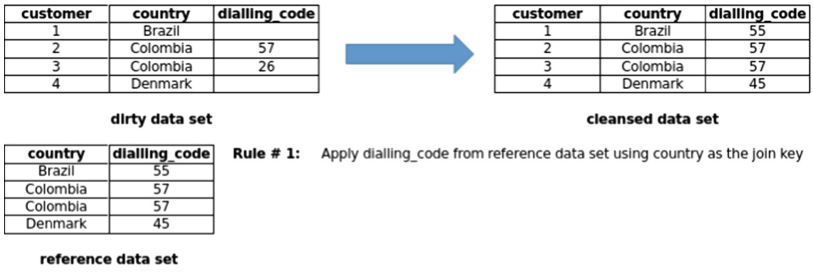


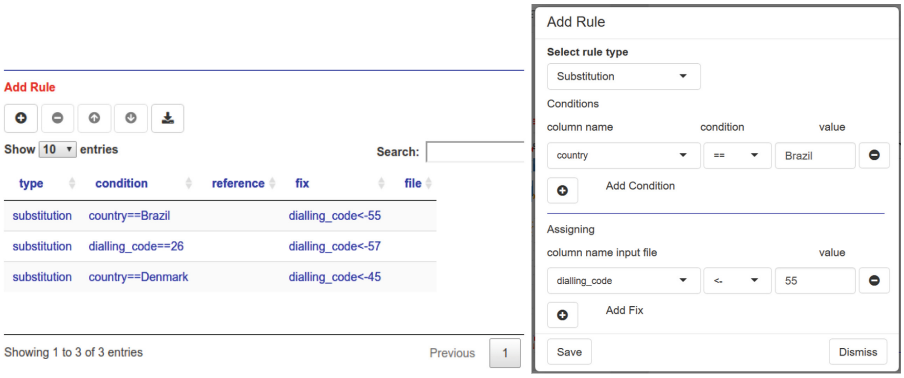
Fig. 2. Using a reference rule.

RefDataCleaner consists of four steps, illustrated by the screenshots in Figs. 3, 4, 5, which may be performed in an iterative manner until the user obtains a result that he or she is satisfied with:

1. **Input File Selection.** The user selects the input file with the data set to be repaired. This is uploaded and displayed to the user (see Fig. 3a).
2. **Reference File Selection.** In this optional step, shown in Fig. 3b, the user can add files with reference data. This is only required if the user intends to add reference rules. If several reference files are added, the drop-down menu shown on the top-right enables the user to select the reference data set to view.
3. **Rule Management.** The next step is for the user to manage data repair rules. The plus and minus icons enable rules to be added and removed respectively, and the up and down arrow icons enable rule order to be changed. Figures 4 and 5 show the running example in this section with substitution and reference rules respectively. Note that although not shown in these screenshots, it is possible to mix both types of rules interchangeably.

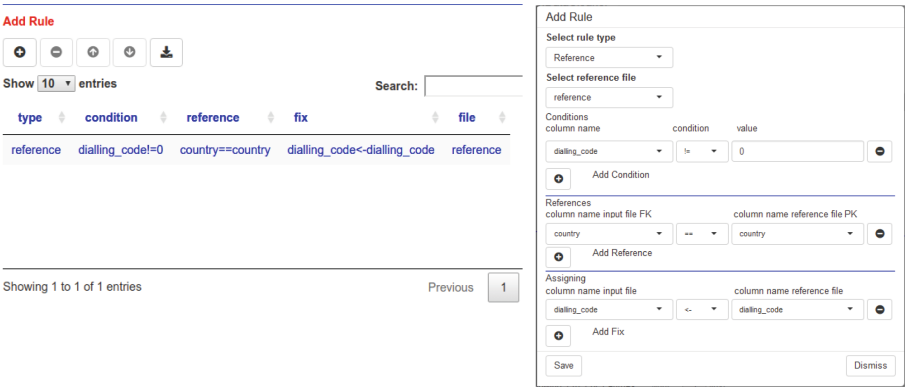
(a) The dirty data file (b) A reference file

Fig. 3. Uploading the input files to RefDataCleaner.



(a) Set of created rules (b) Adding a new rule

Fig. 4. Substitution rule example.



(a) Set of created rules (b) Adding a new rule

Fig. 5. Reference rule example.

4. **Result Generation.** By clicking on the "Run" icon, the rules are applied to the input file one record at a time, in the order that they have been specified in the Rule Management step. The user may then download the repaired data set in the desired file format (Fig. 6).

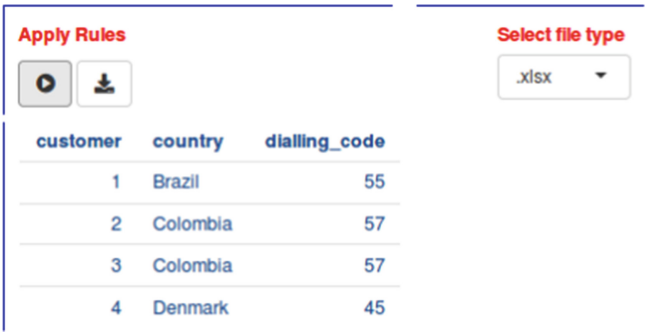


Fig. 6. Apply, look, select and download result file.

4 Experiment Design

In order to evaluate the effectiveness of RefDataCleaner, we carried out an experiment to compare user performance and subjective user preference with the Microsoft Excel 2016 spreadsheet application as a baseline due to its widespread use in organizations throughout the world.

Overview. For the experiment, a focus group comprising an hour-long session was devised as follows:

- Initially, participants are presented with a tutorial on data cleaning for both tools using a practical example [15 min][5].
- Then, using the first tool [20 min]:
  - Users carry out two data cleaning tasks, which involve using tool functionality to correct errors in a data file as explained in the natural language task description (see task descriptions ahead);
  - Users answer a usability questionnaire about the tool.
- The same process, with the same data cleaning tasks, is repeated for the second tool [20 min].
- Finally, a comparative questionnaire is presented to participants in which they give free text answers comparing both tools [5 min].

Participants were divided into two groups. Group A used Microsoft Excel first, and RefDataCleaner second. Group B used the tools in the reverse order. This was done to mitigate any variability which tool order and increased user familiarity with the tasks may cause to the results. The files repaired by the users, and the answers to the usability questions, were all recorded on a Google Form. Users were prompted when the time allocated for each step was reached,

[5] In the case of Microsoft Excel, participants are shown how substitution rules may be mimicked using find/replace/copy/paste functionality, and reference rules using `VLOOKUP` formulae. However, participants are free to use any functionality available in Excel for the data cleaning process.

and asked to upload the repaired data files as they were (even if they were not entirely satisfied with the outcome). This ensured that an equal amount of time was spent using both tools, to enable a fairer comparison.

Task 1: *Repairing the Iris data set.* The first task involves repairing the Iris data set, a well-known multivariate data set introduced by Ronald Fisher in his 1936 paper [7]. This data set comprises 150 records, and five attributes: `sepalLength`, `sepalWidth`, `petalLength`, `petalWidth` and `species`. We randomly deleted 27 data values for the `species` attribute, which the participants were subsequently requested to fix using the decision tree shown in Fig. 7 as a guide. Users were expected to use substitution rules to fix this data set, as this task does not involve a reference data set.

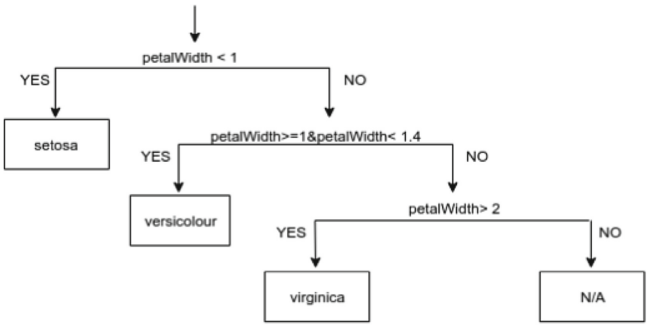


Fig. 7. Iris data set decision tree used to inform data cleaning.

Task 2: *Repairing the Movies Data set.* The second task involves repairing a data set taken from Wikipedia with a list of highest-grossing movies [2]. To make it more manageable, a subset of 46 records are taken from this data set. This data set contains six attributes: `rank`, `title`, `worldwide_gross`, `year`, `director`, and `distributor`. We randomly introduced 92 data errors into the `year`, `director`, and `distributor` attributes.

Furthermore, two reference data sets were made available to participants: (1) *Company*, which contains 8 records with the attributes `distributor_code` and `distributor`, and (2) *Directors*, which contains 56 records with the attributes `title`, `year`, and `director`. Users were expected to use reference rules to fix this data set, by using the reference data sets provided.

Usability Questions. For each tool, we adapted four questions from the System Usability Scale (SUS) [20] to evaluate subjective user preference. We adapted two positive and two negative questions for this purpose, which users answered according to a five-point Likert scale. At the end of the session, we also posed the following three comparative usability questions, which users answered using free text:

- What tool seemed easier to use? Why?
- What tool would you use to clean your data? Why?
- What tool offered you the simplest functionality more simple to clean the data? Why?

4.1 Evaluation Metrics

We have two types of user performance measures, viz., error detection and data repair performance measures. *Error detection* performance measures evaluate how effectively users were able to identify erroneous data using the tools, and *Data repair* performance measures whether erroneous data items were repaired correctly.

Error Detection User Performance. For these measures, we define the following concepts:

- the true positives (TP), i.e., the items of data that are erroneous and were identified as being erroneous.
- the false positives (FP), i.e., the items of data that were not erroneous but were identified as being erroneous.
- the true negatives (TN), i.e., the items of data that are not erroneous and were correctly identified as not being erroneous; and
- the false negatives (FN), i.e., the items of data that are erroneous but were not identified as being erroneous.

For the purposes of these measures, we deem an item of data to have been identified as erroneous when it has been modified. Conversely, if an item of data is not modified, we deem it as having been identified as being correct. Based on this, we define error detection accuracy, precision, recall and specificity as follows [16]:

$$\text{Error detection accuracy} = \frac{TP + FP}{TP + TN + FP + FN} \tag{1}$$

$$\text{Error detection precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Error detection recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{Error detection specificity} = \frac{TN}{TN + FP} \tag{4}$$

Data Repair User Performance. Taking into account only the erroneous data, we examine the fraction of records which were correctly repaired. We define the *data repair accuracy* as:

$$\text{Data repair accuracy} = \frac{\text{Records repaired correctly}}{\text{Number of erroneous records}} \tag{5}$$

Usability Score. Apart from user performance, we also compute a usability score to measure subjective user preference. Our approach is based by taking four questions from the System Usability Scale (SUS) [20]. Each question is scored using the Likert scale, where 1 indicates total disagreement and 5 total agreement. The score for each individual question is computed as follows:

- The positive questions take the value assigned by the user minus one.
- The negative questions are 5 minus the value assigned by the user.

The individual scores for the questions are summed, and the total is scaled to give a number between 0 and 100 as follows:

$$\text{Usability Score} = \sum_{i=1}^{N}(score_i) \times \frac{100}{4N} \tag{6}$$

where N is the number of questions (four in this case), $score_i$ is the score awarded to the ith question, and the constant 4 represents the maximum score for any given question.

5 Evaluation Results

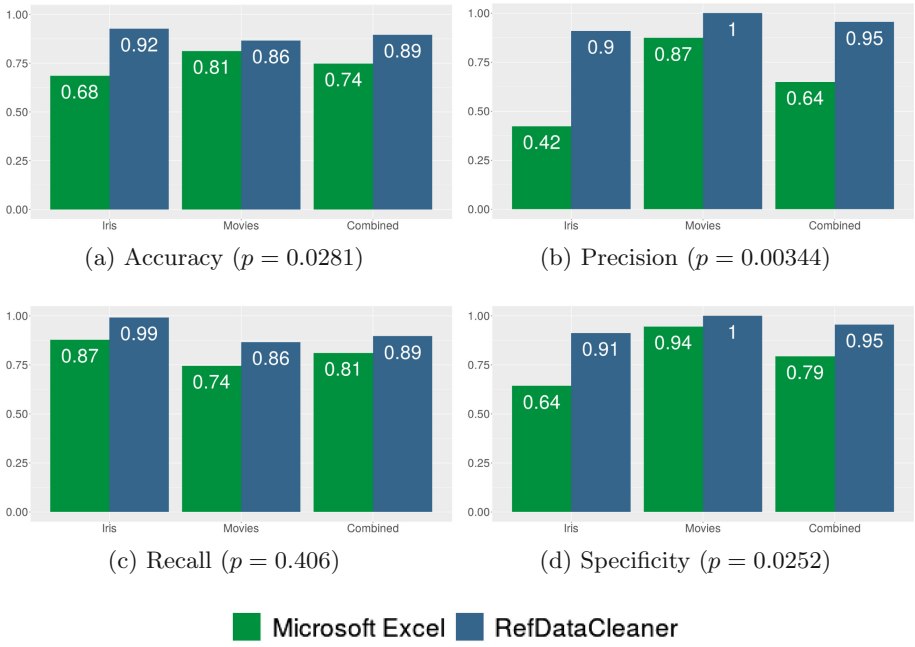
We recruited 11 student volunteers familiar with data analysis with experience in Microsoft Excel providing a bonus grade as an incentive for participation to ensure participant engagement. 6 students conformed group A and 5 students group B, resp. In order to ensure equal participation in both groups in our results, we randomly discarded the results obtained from one of the participants in group A. All participants were familiar with Microsoft Excel, and were new to RefDataCleaner. This section reports the results obtained.

5.1 Error Detection Performance

In the first instance we evaluate whether data is correctly diagnosed as being erroneous or correct. Figure 8 presents the error detection performance results obtained, with the results of a paired two-tailed t-test used to determine statistical significance.

For accuracy, shown in Fig. 8a, we observe that more accurate results are obtained for RefDataCleaner than Microsoft Excel for both tasks combined: on average, the accuracy measure is 0.148 higher for RefDataCleaner. This difference is starkest for the Iris Task, where RefDataCleaner average accuracy is 0.241 higher than Microsoft Excel. Moreover, the result is extremely statistically significant, as the p-value obtained is under the commonly-used 0.05 threshold. This result tells us that, overall, the diagnoses made are more likely to be correct with RefDataCleaner than with Microsoft Excel.

The results for error detection precision are shown in Fig. 8b. For both tasks combined, the average precision is 0.306 higher for RefDataCleaner than Microsoft Excel. Once again, this is particularly stark for the Iris task, where the average



(a) Accuracy ($p = 0.0281$) (b) Precision ($p = 0.00344$)

(c) Recall ($p = 0.406$) (d) Specificity ($p = 0.0252$)

Fig. 8. Error detection performance.

precision is 0.486 higher for RefDataCleaner, and the results are statistically significant. This result indicates that, overall, users were more effective at correctly detecting erroneous data with RefDataCleaner than with Microsoft Excel.

Figure 8c shows the error detection recall. For both tasks combined, the error detection recall is 0.085 greater for RefDataCleaner than for Microsoft Excel. This result suggests that participants were less likely to miss erroneous data items with RefDataCleaner than with Microsoft Excel, although the p-value obtained indicates that this result is not statistically significant.

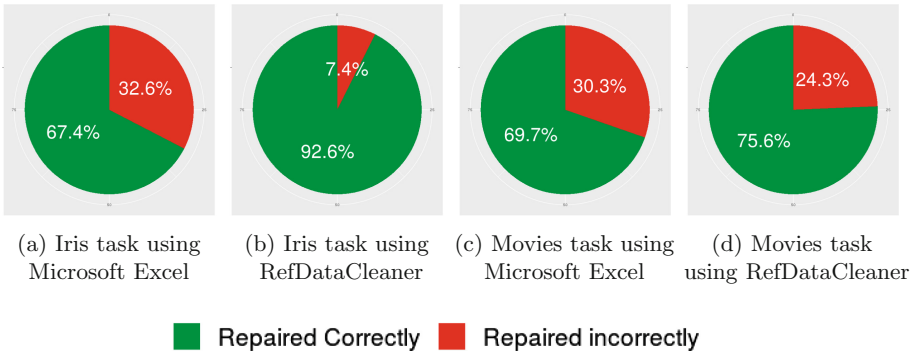
The results for error detection specificity are similar to the overall trend. The specificity obtained for RefDataCleaner is 0.162 higher for RefDataCleaner than Microsoft Excel, with the difference being starker for the Iris task (0.269). This statistically significant result tells us that participants were more effective at identifying non-erroneous records with RefDataCleaner compared to Microsoft Excel.

5.2 Data Repair

For the second part of the evaluation, we consider the issue of data repair. The results obtained for data repair accuracy are shown in Fig. 9.

The results show that, for both tasks, erroneous data was repaired correctly more often for RefDataCleaner than for Microsoft Excel, equally stark for the Iris task, where the average of data repair accuracy was better. However, the

p-value obtained in the exercises iris and movies (0.151 and 0.753 respectively) indicates that these results are not statistically significant.



(a) Iris task using Microsoft Excel (b) Iris task using RefDataCleaner (c) Movies task using Microsoft Excel (d) Movies task using RefDataCleaner

Fig. 9. Data repair accuracy.

5.3 Usability

Figure 10 presents the usability scores obtained for each tool. For Microsoft Excel the usability score was 56.3, whereas for RefDataCleaner the usability score was 71.9, approximately 15.6% higher. This result shows that, overall, RefDataCleaner scored higher and was preferred by users. This matches the results obtained for the comparative questions given to participants at the end of the session: 90% percent of participants considered that RefDataCleaner was easier to use, compared to 10% who preferred Microsoft Excel on the basis that it is a familiar tool used in their daily work. Similarly, 90% of participants rated RefDataCleaner as being the more intuitive tool. However, a lower 70% expressed that they would use RefDataCleaner for data cleaning, the justification being that Microsoft Excel provides a broader range of functionality for data cleaning.

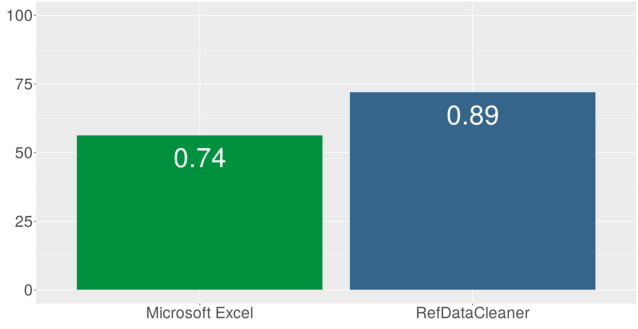


Fig. 10. Usability score ($p = 0.039$)

5.4 Usability Score vs. Error Detection and Data Repaired Accuracy

Finally, when comparing usability score against accuracy, we can observe with RefDataCleaner in Figs. 11c and d a tendency towards the upper right for both error detection and repair.

(a) Microsoft Excel error detection

(b) Microsoft Excel error repair

(c) RefDataCleaner error detection

(d) RefDataCleaner error repair

Fig. 11. Accuracy vs. Usability score

This is indicative of the greater performance and subjective user preference exhibited by RefDataCleaner. On the other hand, in Figs. 11a and b we can see that the points are more spread out, an indication that both performance and subjective user preference varied more greatly for Microsoft Excel.

5.5 Qualitative Analysis

We can glean further comparisons between Microsoft Excel and RefDataCleaner from qualitative analysis of the user answers regarding its usability. Some of the user responses indicating the software application they preferred, and the reasons, were:

- "[I preferred RefDataCleaner as] it is more intuitive."
- "[I preferred RefDataCleaner as] it is optimized to carry out two very useful functions."
- "[I preferred RefDataCleaner as it enables a] faster cleaning processes to be performed."
- "[I preferred Microsoft Excel as it is] more familiar to my daily work.".

6 Conclusions

With the growing development of humanity and the expansion for the need for data management in most everyday fields, it is estimated that some organizations invest up to 40% of their budget in integrating information reliably [5]. Cleaning data is one of the main challenges in this process. This paper reports the results of a preliminary study that shows significant differences on performance and subjective user preference of two data cleaning approaches, RefDataCleaner which is a bespoke application to carry out data cleaning tasks using reference data, and Microsoft Excel, a generic tool with a broad range of functionality.

The main findings were that (1) higher error detection performance was obtained for RefDataCleaner in terms of accuracy, precision and specificity; (2) the difference in error repair performance between the tools is not significant; (3) the preferred tool by users was RefDataCleaner; and (4) usability and performance are more highly correlated for RefDataCleaner than for Microsoft Excel, indicating that performance and usability was much more diverse for Microsoft Excel.

To gain further insights on this issue, further work may usefully investigate the trade-off between specific and generic data cleaning tools in more detail, as well as other types of data wrangling tasks which may usefully lend themselves to these types of tools.

References

1. Exploratory home page. https://exploratory.io/. Accessed 17 June 2019
2. List of highest-grossing films. https://en.wikipedia.org/wiki/List_of_highest-grossing_films. Accessed 14 Apr 2019
3. Tableau website. https://www.tableau.com/learn/whitepapers/make-everyone-your-organization-data-scientist. Accessed 17 June 2019
4. Abedjan, Z., et al.: Detecting data errors: where are we and what needs to be done? Proc. VLDB Endow. **9**(12), 993–1004 (2016)
5. Bernstein, P.A., Haas, L.M.: Information integration in the enterprise. Commun. ACM **51**(9), 72–79 (2008)
6. Fan, W., Geerts, F.: Foundations of Data Quality Management (2012)
7. Fisher, R.A.: The use of multiple measurements in taxonomic problems. Ann. Eugen. **7**(2), 179–188 (1936)
8. Furche, T., Gottlob, G., Libkin, L., Orsi, G., Paton, N.W.: Data wrangling for big data: challenges and opportunities. In: EDBT, pp. 473–478 (2016)
9. Galpin, I., Abel, E., Paton, N.W.: Source selection languages: a usability evaluation. In: Proceedings of the Workshop on Human-In-the-Loop Data Analytics, p. 8. ACM (2018)
10. Kim, W., Choi, B.J., Hong, E., Kim, S.K., Lee, D.: A taxonomy of dirty data. Data Min. Knowl. Discov. **7**(1), 81–99 (2003)
11. Koehler, M., et al.: Data context informed data wrangling. In: 2017 IEEE International Conference on Big Data (Big Data), pp. 956–963. IEEE (2017)
12. Konstantinou, N., et al.: The VADA architecture for cost-effective data wrangling. In: Proceedings of the 2017 ACM International Conference on Management of Data, pp. 1599–1602. ACM (2017)

13. Lohr, S.: For big-data scientists, 'janitor work' is key hurdle to insights. https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html. Accessed 15 May 2019
14. Müller, H., Freytag, J.C.: Problems, Methods, and Challenges in Comprehensive Data Cleansing, pp. 1–23. Humboldt-Universität zu, Berlin (2003)
15. Oliveira, P., Rodrigues, F., Rangel Henriques, P., Galhardas, H.: A taxonomy of data quality problems. J. Data Inf. Qual. JDIQ (2005)
16. Olson, D., Dursun, D.: Advanced Data Mining Techniques, 1st edn. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-76917-0
17. Orr, K.: Data quality and systems theory. Commun. ACM **41**(2), 66–71 (1998)
18. Rahm, E., Do, H.H.: Data cleaning: problems and current approaches. IEEE Data Eng. Bull. **23**(4), 3–13 (2000)
19. Redman, T.C.: The impact of poor data quality on the typical enterprise. Commun. ACM **41**(2), 79–82 (1998)
20. Sauro, J.: Measuring usability with the system usability scale (SUS). https://measuringu.com/sus/. Accessed 10 May 2019
21. International Organization for Standardization: Software product quality. https://iso25000.com/index.php/en/iso-25000-standards/iso-25010. Accessed 21 May 2019