RefDataCleaner: A Usable Data Cleaning Tool

Juan Carlos Leon Medina and Ixent Galpin

Universidad Jorge Tadeo Lozano, Bogotá, Colombia

ICAI 2019
Second International Conference on Applied Informatics
6-9 November 2019, Madrid, Spain

UTADEO
UNIVERSIDAD DE BOGOTA JORGE TADEO LOZANO

Introduction

What is the problem?

- Numerous attempts have been made to automate steps in the data wrangling pipeline
  - source selection, mapping generation, entity recognition, error detection, data cleaning
  - However, in practice, these steps are mostly done manually by experts
- This is costly for the organizations involved, given that anomalies are present in around 5% of data[1]
- A data scientist spends 80% time preparing data, and 20% analysing data, once it has been cleaned and integrated[2].

[1] Ken Orr. "Data Quality and Systems Theory". In: *Communications of the ACM* 41.2 (1998), pp. 66–71, Tomas C. Redman. "The impact of Poor Data Quality on the Typical Enterprise". In: *Communications of the ACM* 41.2 (1998), pp. 79–82.

[2] Steve Lohr. *For Big-Data Scientists, 'Janitor Work' Is Key Hurdle to Insights*. Online; Accessed 15 May 2019. URL: https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html.

Introduction

Why is it interesting and important?

- Numerous tools in the market purport to democratize data science, e.g., Tableau or Exploratory
- Furthermore, recently usability workshops have emerged associated with conferences in the data management research community, e.g., HILDA[3] and IDEA[4] co-located with SIGMOD and KDD respectively.
- *Usability* is becoming an ever more important consideration by tool designers.

[3] http://hilda.io/2019/

[4] http://poloclub.gatech.edu/idea2018/

Introduction

Why is it hard?

- Data cleaning requires the understanding of various issues
    - Functional dependencies, integrity constraints...
- Such concepts are not easy to grasp by non-expert users.
- It is a challenge to design tools that are easy-to-use and prevent users from applying the tools incorrectly.

Introduction

Why hasn't it been solved before?

- There has been relatively little research into the usability of tools used for data wrangling.
- Galpin *et al.*[5] carry out a usability study of source selection approaches.
- This work differs from previous work in that it proposes and evaluates the usability of a data cleaning tool.

[5] Ixent Galpin, Edward Abel, and Norman W Paton. "Source Selection Languages: A Usability Evaluation". In: *Proceedings of the Workshop on Human-In-the-Loop Data Analytics.* ACM. 2018, p. 8.

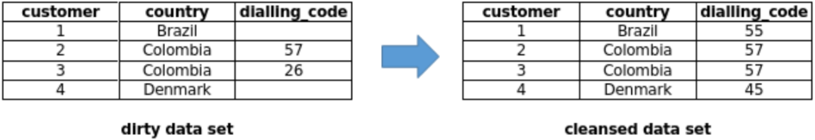
Application Description

How does it work?

- RefDataCleaner is a web-based tool developed using Shiny R that detects and repairs errors in structured and semi-structured data files
- RefDataCleaner checks, row-by-row, the input file, detecting inconsistencies in the data previously defined by conditions created by the user
- Once an anomaly is detected, RefDataCleaner repairs these problems by replacing them with
  - user-defined values, for Substitution rules
  - relationship-defined values between input file and a reference file, for Reference rules

Application Description

Substitution rules



customer	country	dialling_code
1	Brazil	
2	Colombia	57
3	Colombia	26
4	Denmark	

dirty data set

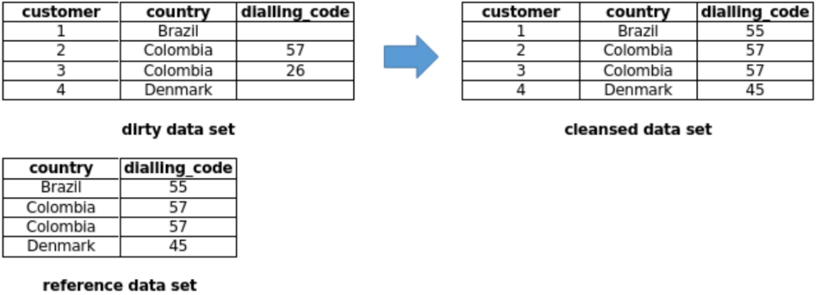
customer	country	dialling_code
1	Brazil	55
2	Colombia	57
3	Colombia	57
4	Denmark	45

cleansed data set

- If country is equal to ’Brazil’ assign 55 to dialling_code
- If country is equal to ’Colombia’ assign 57 to dialling_code
- If country is equal to ’Denmark’ assign 45 to dialling_code

Application Description

Reference rules



customer	country	dialling_code
1	Brazil	
2	Colombia	57
3	Colombia	26
4	Denmark	

dirty data set

customer	country	dialling_code
1	Brazil	55
2	Colombia	57
3	Colombia	57
4	Denmark	45

cleansed data set

country	dialling_code
Brazil	55
Colombia	57
Colombia	57
Denmark	45

reference data set

- Apply dialling_code from reference data set using country as the join key

Experiment Design

Overall Methodology

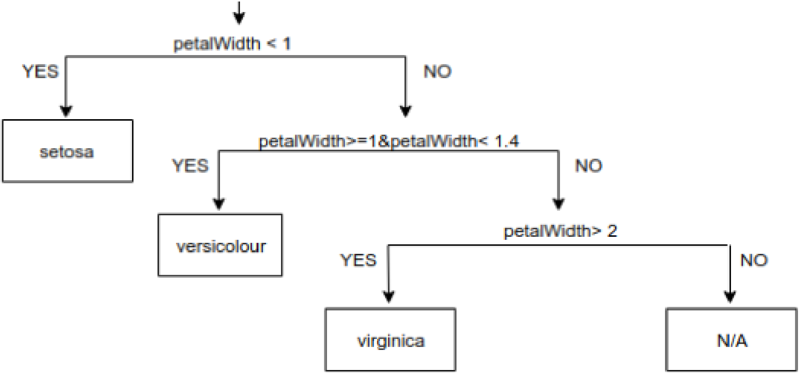
We carried out a usability study to compare RefDataCleaner and Microsoft Excel (our baseline)

- We recruited volunteers familiar with Microsoft Excel
- We presented a short tutorial in RefDataCleaner and Microsoft Excel for each tool
- Two data repair tasks were given to the participants. To prevent any variability in the results,
  - Group A used RefDataCleaner then Microsoft Excel
  - Group B used Microsoft Excel then RefDataCleaner
- Finally, participants answered a usability questionnaire about the tools used

Experiment Design

First task: Used substitution rules for repairing

- Iris data set from Ronald Fisher with five attributes and randomly deleted 27 data values for the species attribute.
- Using decision tree as guide.



Experiment Design

Second task: Used reference rules for repairing

- Movies data set from Wikipedia, a list of highest-grossing movies with six attributes: rank, title, worldwide gross, year, director and distributor .
- Randomly introduced 92 data errors into the last three.
- Using two reference data sets to fix data issues:
  - Companies data set with attributes: rank, title, worldwide gross, year, distributor code and distributor name.
  - Directors data set with attributes: title, year, and director name.

Experiment Design

Usability Questionnaire

Taken from System Usability Scale (SUS)[6]

- I found the system unnecessarily complex.
- I thought the system was easy to use.
- I would imagine that most people would learn to use this system very quickly.
- I needed to learn a lot of things before I could get going with this system.

[6] Jeff Sauro. *Measuring Usability with the System Usability Scale (SUS)*. Online; Accessed 10 May 2019. URL: https://measuringu.com/sus/.

Experiment Design

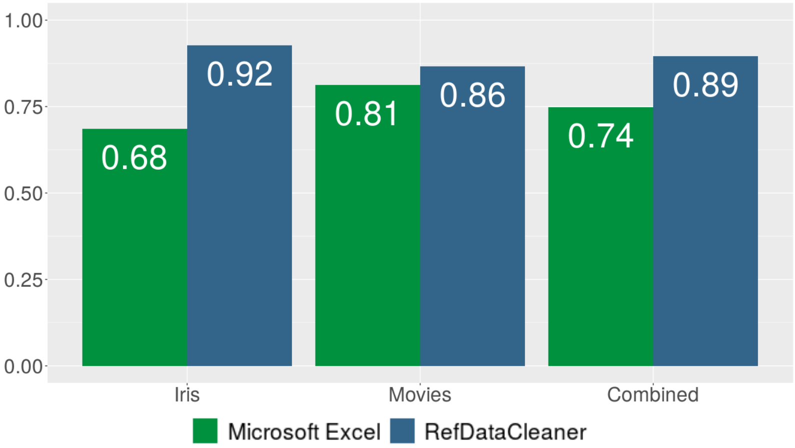
Usability Questionnaire

Three comparative usability questions

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

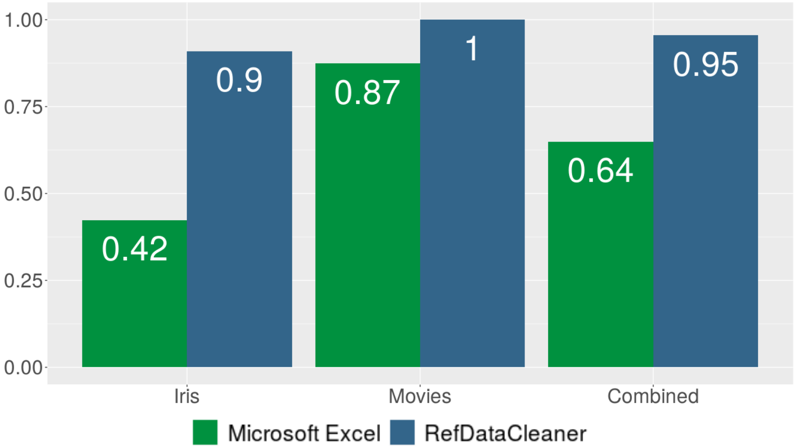
Evaluation Results

Error detection accuracy



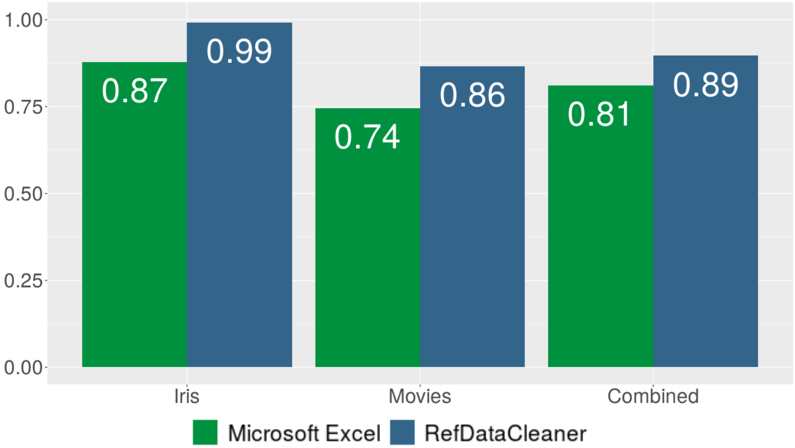
Evaluation Results

Error detection precision



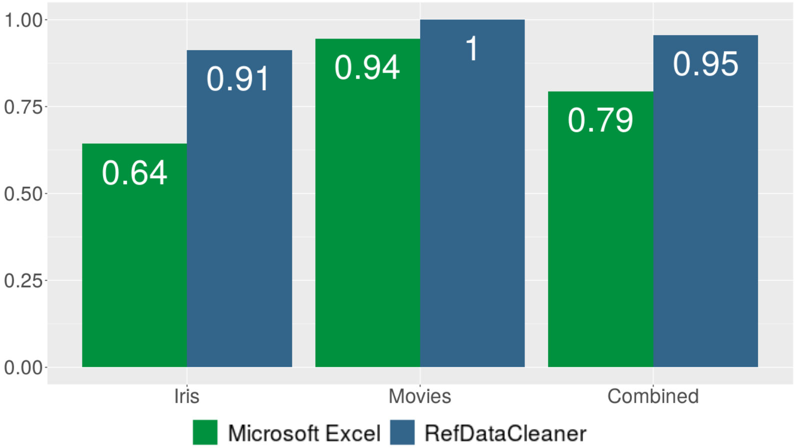
Evaluation Results

Error detection recall



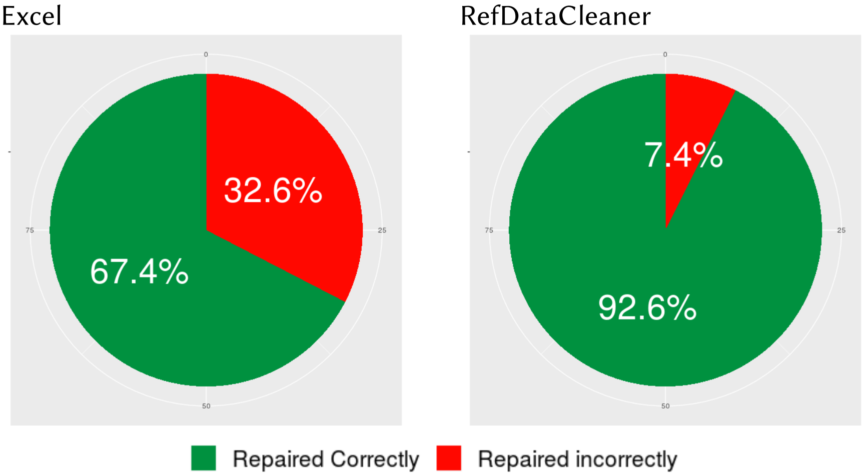
Evaluation Results

Error detection specificity



Evaluation Results

Data repair accuracy iris file

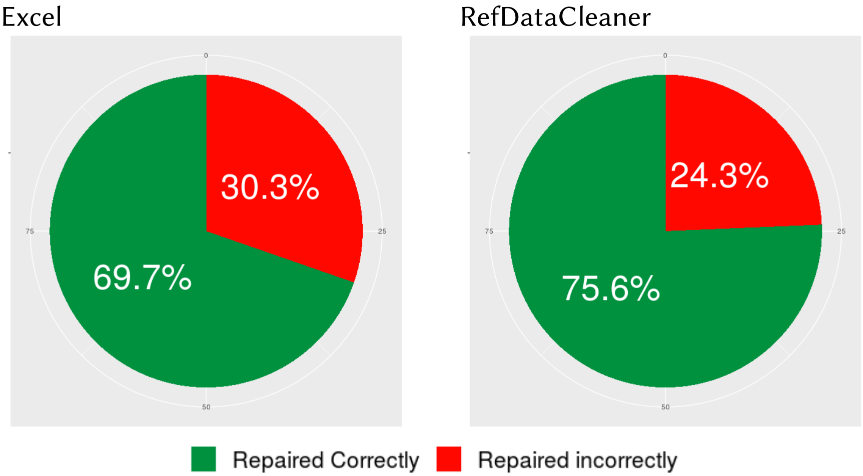


Excel

RefDataCleaner

Evaluation Results

Data repair accuracy movies file

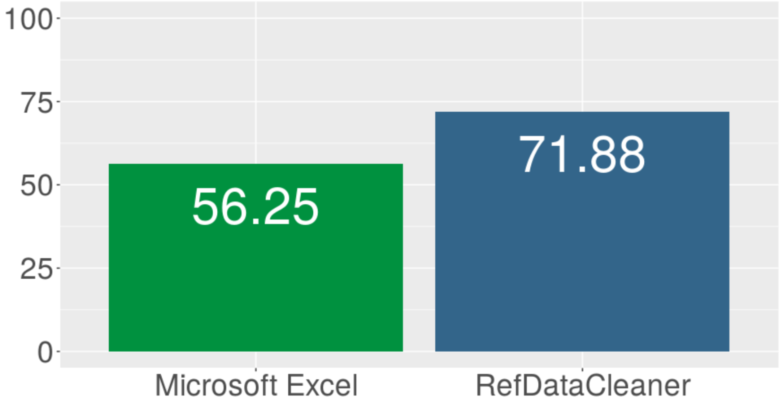


Excel

RefDataCleaner

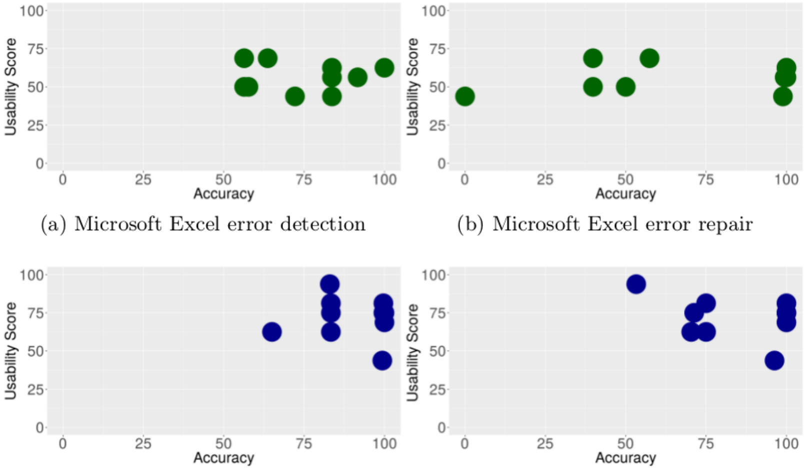
Evaluation Results

Usability score



Evaluation Results

Usability score vs. Error detection and data repaired accuracy



(a) Microsoft Excel error detection

(b) Microsoft Excel error repair

(c) RefDataCleaner error detection

(d) RefDataCleaner error repair

Conclusions

Finally...

- Higher error detection performance was obtained for RefDataCleaner in terms of accuracy, precision and specificity.
- The difference in error repair performance between the tools is not significant.
- The preferred tool by users was RefDataCleaner.
- 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.

Thank You

Any Questions?

Ixent Galpin ixent@utadeo.edu.co
Juan Leon-Medina juan.leonm@utadeo.edu.co

You can try the software at
https://refdatacleaner.shinyapps.io/version_1_0/
Download the source code at
https://github.com/refdatacleaner/version_1_0/.