Data Curation: Towards a Tool for All*

José Dias[1], Jácome Cunha[1,2], and Rui Pereira[2]

[1] University of Minho, Portugal
a78494@alunos.uminho.pt jacome@di.uminho.pt
[2] HASLab/INESC Tec, Portugal
ruipereira@di.uminho.pt

Abstract. Data science has started to become one of the most important skills one can have in the modern world, due to data taking an increasingly meaningful role in our lives. The accessibility of data science is however limited, requiring complicated software or programming knowledge. Both can be challenging and hard to master, even for the simple tasks.
With this in mind, we have approached this issue by providing a new data science platform, termed *DS4All.Curation*, that attempts to reduce the necessary knowledge to perform data science tasks, in particular for data cleaning and curation. By combining HCI concepts, this platform is: *simple* to use through direct manipulation and showing transformation previews; allows users to *save time* by eliminate repetitive tasks and automatically calculating many of the common analyses data scientists must perform; and suggests data transformations based on the contents of the data, allowing for a *smarter* environment.

Keywords: Human-Centered Data Science · Data Cleaning · Data Curation.

1 Introduction

The use of data cannot be dissociated from our daily lives - data supports, e.g., social media, is fundamental to guide us in traffic and is being used in precision medicine by promising health-care avenues. In order to support all these data-based services, the amount of data which are produced these days are tremendous, and are still expected to increase significantly within the near future. For example, Facebook experiences about 2.5 billion likes and 300 million photo uploads on a regular day [22]. Of course data by itself, even if in massive amounts, has very little value. Indeed, it is the information extracted from data which has the potential to change and improve our lives. However, the information extraction process is complex, requiring cleaning, transforming, understanding, analyzing and interpreting data [21]. This is what is currently called Data Science (DS) [3], and one incorrect or inaccurate decision in any step of the

* This work is financed by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project UIDB/50014/2020.

process is sufficient enough to compromise the extracted information [7]. However, the challenge for any data scientist is that performing these steps requires a variety of skills including mathematics, statistics, machine learning (ML), data structures, algorithms, and correlation or causation [9]. Nevertheless, there is a worldwide movement towards pushing everyone to have DS skills. For instance, a study by IBM advocates that academia must ensure data literacy for any student in any field of education [5]. Similarly, the Portuguese Government has also defined that until the end of 2023 all students with higher education must have the opportunity to learn DS [15]. In fact, many other countries have defined national strategies for DS [3]. However, to teach advanced techniques and tools to an entire academic community is challenging, tedious, and difficult to entirely fulfil. Indeed, a study by Kaggle, with more than 16.000 answers from DS practitioners, shows that textual programming languages (PLs) such as Python or R are the most used tools (76.3% and 59.2%, respectively) [6]. Unfortunately, programming is a very challenging task, taking years to train and master. While there are other tools targeting inexperienced users, such as Tableau or Excel, these are much less used (20.4% and 13.7%, respectively [6]). Moreover, there is no empirical evidence of their efficiency and efficacy amongst non-expert users.

Human-computer interaction (HCI) related communities have been proposing several methodologies to aid users in developing their own software. These users are usually termed end users, i.e. computer users with no (or little) software development background, yet still need to develop software, i.e. end-user programming [10]. The proposed methodologies include visual programming [2], programming by example [4] or direct manipulation [19].

In this work we build on such works to further design methodologies and a tool (termed DS4All.Curation) that can be productively used by any end user for performing DS, particularly focusing on data cleaning and curation. The curation and transformation of data is generally a very complex and time consuming process for an experienced data scientist [14, 7]. Oftentimes, several tools or programming languages (a PL can also be seen as a tool) are used for this. But to do so, data scientists must properly learn to use these. This is a larger issue for end user data scientists, with their limited (or inexistent) computer science background.

Thus, we believe that a visual development environment for data science (DS) direct manipulation will help diminish such difficulties and limitations. Naturally, data should be represented in a way that (end) users can actually see and manipulate it using some tabular format, e.g., resembling Excel. Whenever a user wishes to apply a certain transformation, they should also be able to see a preview of how their data will be altered. Such a side-by-side look at the dataset, prior and post changes, aims to help remove a level of abstraction of how data will be changed. Additionally, a user should be able to, at any point, directly manipulate the data within such a dataset previewer, such as updating cell values, or through a drop down menu to allow changes or filtering data on a specific column. For many operations related to data curation [13] this should be sufficient. In essence, this environment must be *simple*.

Such a visual environment must also help guide the user to more efficiently perform their work. Indeed, prior studies suggest DS environments should guide their users [21]. For example, it is very common to calculate the statistical information (average, min, max, etc.) or grouping/clustering of data prior to manipulating the data [18]. Such statistics help data scientists summarize the contents of their data, understanding if there are any outliers present, or if something appears to be incorrect. For such operations, data scientists have to repeatedly turn to using programming or complex tools to perform such common tasks each time and every time they tackle a new dataset. We propose that such common tasks should be automatically performed within our visual environment, in order to facilitate the end user data scientists' work, and in turn *save time*.

We propose to go one step further and use such information to automatically present suggestions of common (or uncommon) transformations to the user, which can be automatically applied by the system. An example would be, in a column representing gender, when detecting similar values such as `FEMALE` and `female`, to suggest replacing one entry by the other or by a new value. Another example would be for columns inferred as numerical, where a suggestion to remove data entries based on minimum and maximum bounds may be presented. The system should also learn with the user, by understanding what operations they repeatedly need and/or use, and intelligently offer suggestions. Offering both statistical information on the data and suggested operations to be performed will lower the amount of time taken to perform such tasks, reduce errors, and also reduce the possibility of incorrectly programming the tasks. As such, the final requirement of a data science environment for any use is that it must be *smart*.

In summary, we propose that a visual data science development environment must be *simple*, *saves time*, and is *smart*. Section 2 presents our initial steps in providing data science end users with an environment adhering to these three principals. In Section 3 we discuss related work and in Section 4 we summarize out contribution and discuss future work.

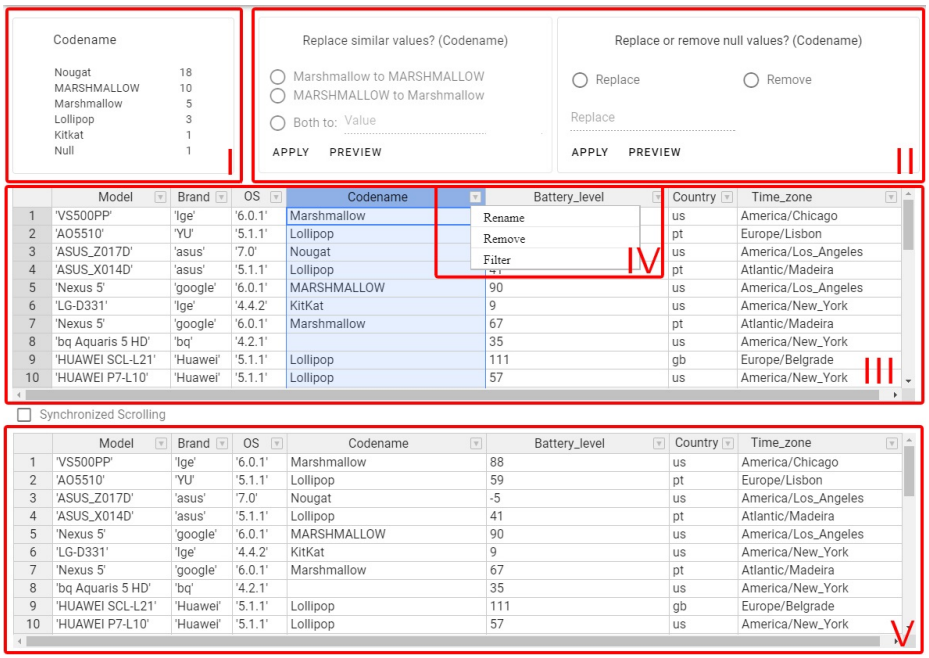
2 DS4All.Curation: A Data Curation Tool for All

In accordance to what we have previously discussed, we believe there are several paths one may take when developing a visual environment for the direct manipulation of data. We have developed a prototype of a humanized data cleaning tool, termed DS4All.Curation[3], shown in Figure 1, that we now describe. The dataset represents Android smartphone usage information [12].

Since we are proposing methodologies and tools for data science, it seems natural that data should be represented in a way users can actually see and manipulate it using some tabular format, e.g., resembling Excel. Indeed, shown in Figure 1 - V (*Original dataset*), we have the original and unaltered dataset shown at all times, allowing the end user to better accompany their transformations. All

[3] DS4All.Curation can be found at https://github.com/Zamreg/HDC.

such transformations would be shown and previewed in Figure 1 - III (*Preview dataset*). This side-by-side look at the dataset before and after applying changes aims to help remove a level of abstraction of how data will be changed, and directly present such actions. At any point, the user may directly manipulate the data within the *Preview dataset*, such as updating cell values, or through a drop down menu (as shown in Figure 1 - IV) to allow changes or filtering data on a specific column. For many operations related to data cleaning/curation [13] this should be sufficient.



	Model	Brand	OS	Codename	Battery_level	Country	Time_zone
1	'VS500PP'	'lge'	'6.0.1'	Marshmallow		us	America/Chicago
2	'AO5510'	'YU'	'5.1.1'	Lollipop		pt	Europe/Lisbon
3	'ASUS_Z017D'	'asus'	'7.0'	Nougat		us	America/Los_Angeles
4	'ASUS_X014D'	'asus'	'5.1.1'	Lollipop		pt	Atlantic/Madeira
5	'Nexus 5'	'google'	'6.0.1'	MARSHMALLOW	90	us	America/Los_Angeles
6	'LG-D331'	'lge'	'4.4.2'	KitKat	9	us	America/New_York
7	'Nexus 5'	'google'	'6.0.1'	Marshmallow	67	pt	Atlantic/Madeira
8	'bq Aquaris 5 HD'	'bq'	'4.2.1'		35	us	America/New_York
9	'HUAWEI SCL-L21'	'Huawei'	'5.1.1'	Lollipop	111	gb	Europe/Belgrade
10	'HUAWEI P7-L10'	'Huawei'	'5.1.1'	Lollipop	57	us	America/New_York

	Model	Brand	OS	Codename	Battery_level	Country	Time_zone
1	'VS500PP'	'lge'	'6.0.1'	Marshmallow	88	us	America/Chicago
2	'AO5510'	'YU'	'5.1.1'	Lollipop	59	pt	Europe/Lisbon
3	'ASUS_Z017D'	'asus'	'7.0'	Nougat	-5	us	America/Los_Angeles
4	'ASUS_X014D'	'asus'	'5.1.1'	Lollipop	41	pt	Atlantic/Madeira
5	'Nexus 5'	'google'	'6.0.1'	MARSHMALLOW	90	us	America/Los_Angeles
6	'LG-D331'	'lge'	'4.4.2'	KitKat	9	us	America/New_York
7	'Nexus 5'	'google'	'6.0.1'	Marshmallow	67	pt	Atlantic/Madeira
8	'bq Aquaris 5 HD'	'bq'	'4.2.1'		35	us	America/New_York
9	'HUAWEI SCL-L21'	'Huawei'	'5.1.1'	Lollipop	111	gb	Europe/Belgrade
10	'HUAWEI P7-L10'	'Huawei'	'5.1.1'	Lollipop	57	us	America/New_York

Fig. 1. Humanized Data Cleaning Example Interface

When the user selects one specific column, a *statistics card* is displayed in order to help summarize the contents of the chosen column. An example is shown in Figure 1 - I, where the `Codename` column is selected and a *statistics card* detailing the different data entries (and their quantification) is shown. In addition to displaying a *statistics card*, a collection of *suggestion cards* are automatically displayed (shown in Figure 1 - II), where each presents a data transformation action, based on the statistics and data inference. Following our example, the system detects two very similar values: `Marshmallow` and `MARSHMALLOW`, and thus suggests replacing one data value by the other or by a new value. In the same example, it also detected the presence of `null` or empty values, and suggests either replacing them with a new value or removing such data entries. Shown in Figure 2, is another example of such cards if one would choose the

`Battery_level` column. In this case, as the column is inferred to be numerical, a set of common numerical metrics are shown, followed by a suggestion to remove data entries based on minimum and maximum bounds. Knowing that a smartphone's battery level could not be higher than 100% nor lower than 1%, such data entries might present themselves as dirty data and could accordingly be removed through the *suggestion card*.

Such *statistic cards* and *suggestion cards* aim to remove another layer of complexity in data cleaning by automatically presenting common statistical information which users otherwise have to calculate, and by suggesting transformations based on their data. In both cases, the user would have to resort to either programming or using complex tools to gather the statistical information and apply their transformation.

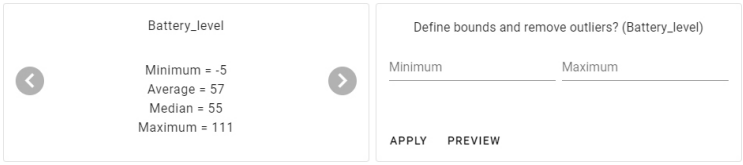


Fig. 2. Numerical statistics and suggestion card example

3 Related Work

Several authors have proposed related approaches to make DS more accessible. Potter's Wheel provides an interactive data transformation and cleaning system that allows users to define transforms through graphical operations or examples and see the effects instantly, making it easy to experiment with different transformations [16]. Unfortunately, the project ended about 20 years ago and does not seem to have been evaluated with users.

Milo [17] and BlockPy [1] are tools that offer a block-based language for users, but focus on different aspects. While Milo aims to help users with no computer science background to only perform machine learning techniques, we propose a tool for data cleaning. BlockPy is a visual interface for the Python programming language to motivate students to start learning how to program. In our case, our visual environment is designed for data cleaning, and not a programming language interface.

Wallace et al. propose a tool to allow users with less statistical skills to make use of advanced models written using the R language [20]. Their motivation is similar to ours although their goal is to provide a graphical user interface for a given R model whilst we provide a tool specific for data cleaning tasks.

DataScience4NP is a web platform aiming to provide an intuitive user interface for users to build sequential DS workflows [11]. This system intends to perform all the steps of extracting knowledge from data, which includes data

insertion, pre-processing, transformation, mining and interpretation/evaluation of results, without requiring users to program. However, similarly to Milo, this platform is focused on data mining techniques whilst ours focus on data cleaning.

Industry and open-source communities have also proposed several tools for DS. Popular tools include Microsoft PowerBi[4], Tableau (Prep)[5], Jupyter (notebooks)[6], and RapidMiner[7]. These tools allow their users to make data exploration, data mining, visualization and reporting tasks through visual interactive dashboards. However, there does not seem to exist any scientific evidence of their effectiveness amongst end user data scientists. In fact, Jupyter notebooks have been found to be messy by some users [8].

4 Conclusions

In this work we propose a platform for data cleaning/curation intended for end user data scientists. We achieve this by relying on suggestions and direct data manipulation. Currently as we're still improving upon what we have we plan to explore suggestions further and explore programming by example as a way to transform data where the user can specify input and output examples. This has the potential to easily allow users to normalize data, map it to other representations, and further remove a layer of abstraction of data and mental work for our end user data scientists. We also intend to empirically evaluate our tool comparing its usability (effectiveness, efficiency and satisfaction) against other popular tools.

References

1. Bart, A.C., Tibau, J., Tilevich, E., Shaffer, C.A., Kafura, D.: BlockPy: An Open Access Data-Science Environment for Introductory Programmers. Computer **50**(5), 18–26 (may 2017). https://doi.org/10.1109/MC.2017.132
2. Burnett, M.M.: Visual Programming. In: Wiley Encyclopedia of Electrical and Electronics Engineering. John Wiley & Sons, Inc. (dec 1999). https://doi.org/10.1002/047134608x.w1707
3. Cao, L.: Data science: A comprehensive overview. ACM Comput. Surv **50**(43) (2017). https://doi.org/10.1145/3076253
4. Gulwani, S.: Programming by examples (and its applications in data wrangling). In: Dependable Software Systems Engineering, vol. 45, pp. 137–158. IOS Press (apr 2016). https://doi.org/10.3233/978-1-61499-627-9-137
5. IBM and Business-Higher Education Forum and Burning Glass: The Quant Crunch: How the Demand for Data Science Skills Is Disrupting the Job Market (2017), https://www.ibm.com/downloads/cas/3RL3VXGA
6. Kaggle Inc.: The State of Data Science & Machine Learning (2017), https://www.kaggle.com/surveys/2017

[4] https://powerbi.microsoft.com
[5] http://tableau.com
[6] https://jupyter.org
[7] https://rapidminer.com

7. Kandel, S., Paepcke, A., Hellerstein, J.M., Heer, J.: Enterprise Data Analysis and Visualization: An Interview Study. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2917–2926 (dec 2012). https://doi.org/10.1109/TVCG.2012.219
8. Kery, M.B., Radensky, M., Arya, M., John, B.E., Myers, B.A.: The Story in the Notebook. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18. vol. 2018-April, pp. 1–11. ACM Press, New York, New York, USA (apr 2018). https://doi.org/10.1145/3173574.3173748
9. Kim, M., Zimmermann, T., DeLine, R., Begel, A.: The emerging role of data scientists on software development teams. Proceedings - International Conference on Software Engineering pp. 96–107 (2016). https://doi.org/10.1145/2884781.2884783
10. Ko, A.J., Abraham, R., Beckwith, L., Blackwell, A., Burnett, M., Erwig, M., Scaffidi, C., Lawrance, J., Lieberman, H., Myers, B., Rosson, M.B., Rothermel, G., Shaw, M., Wiedenbeck, S.: The state of the art in end-user software engineering. ACM Computing Surveys **43**(3) (apr 2011). https://doi.org/10.1145/1922649.1922658
11. Lopes, B., Pedroso, A., Correia, J., Araujo, F., Cardoso, J., Paiva, R.P.: DataScience4NP -A Data Science Service for Non-Programmers. In: 10º Simpósio de Informática – INForum 2018 (2018)
12. Matalonga, H., Cabral, B., Castor, F., Couto, M., Pereira, R., de Sousa, S.M., Fernandes, J.P.: Greenhub farmer: real-world data for android energy mining. In: 2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR). pp. 171–175. IEEE (2019)
13. Muller, M., Lange, I., Wang, D., Piorkowski, D., Tsay, J., Liao, Q.V., Dugan, C., Erickson, T.: How data science workers work with data: Discovery, capture, curation, design, creation. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. CHI '19, Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3290605.3300356
14. Pereira, P., Cunha, J., Fernandes, J.P.: On Understanding Data Scientists. In: IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC) (2020), to appear
15. Portuguese Government: Contrato para a Legislatura com o Ensino Superior para 2020–2023 (2019), https://www.portugal.gov.pt/download-ficheiros/ficheiro.aspx?v=d2607a18-51c9-489c-a61c-1ff420dab2f0
16. Raman, V., Hellerstein, J.M.: Potter's wheel: An interactive data cleaning system. VLDB 2001 - Proceedings of 27th International Conference on Very Large Data Bases pp. 381–390 (2001)
17. Rao, A., Bihani, A., Nair, M.: Milo: A visual programming environment for Data Science Education. In: 2018 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC). vol. 2018-Octob, pp. 211–215. IEEE (oct 2018). https://doi.org/10.1109/VLHCC.2018.8506504
18. Refaat, M.: Data Preparation for Data Mining Using SAS. Elsevier (2007). https://doi.org/10.1016/B978-0-12-373577-5.X5000-5
19. Shneiderman, B.: Direct Manipulation: A Step Beyond Programming Languages. Computer **16**(8), 57–69 (aug 1983). https://doi.org/10.1109/MC.1983.1654471
20. Wallace, B.C., Dahabreh, I.J., Trikalinos, T.A., Lau, J., Trow, P., Schmid, C.H.: Closing the Gap between Methodologists and End-Users: R as a Computational Back-End. Journal of Statistical Software **49**(5), 1–15 (jun 2012). https://doi.org/10.18637/jss.v049.i05

21. Wongsuphasawat, K., Liu, Y., Heer, J.: Goals, Process, and Challenges of Exploratory Data Analysis: An Interview Study. arXiv preprint arXiv:1911.00568 (nov 2019)
22. Zikopoulos, P.C., DeRoos, D., Parasuraman, K., Deutsch, T., Corrigan, D., Giles, J.: Harness the power of Big Data : the IBM Big Data platform. McGraw-Hill (2013)