

DOI: [10.21105/joss.03680](https://doi.org/10.21105/joss.03680)

Software

- [Review](#) ↗
- [Repository](#) ↗
- [Archive](#) ↗

Editor: [Arfon Smith](#) ↗

Reviewers:

- [@hakanfriden](#)

Submitted: 18 January 2021

Published: 01 September 2021

License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License ([CC BY 4.0](#)).

DataCleaningTool: An open-source MATLAB app for cooperative data cleaning

Devosmita Chatterjee¹, Sven Ahlinder², and Anton Johansson¹

¹ Department of Mathematical Sciences, Chalmers University of Technology, Gothenburg, Sweden

² Powertrain Department, Volvo Group Trucks Technology, Gothenburg, Sweden

Summary

Nowadays, organizations rely on data driven decision making. Data driven decision making is a process which involves collecting large amount of data, analyzing patterns and facts from these insights and using them to develop strategies that benefit organizations. Successful analysis depends on accurate and well structured data. Thus, high quality data is of utmost priority to draw good and valid business decisions and strategies. However, real world data is messy and almost never comes in a clean way. Data cleaning is the ultimate way to solve the data problems. But, data cleaning is really a time consuming task. This demands data cleaning tools for systematically examining data for errors and automatically cleaning them using algorithms. These data cleaning tools help organizations save time and increase their efficiency.

In order to achieve the task of data cleaning, we develop a cooperative, free and open source data cleaning standalone MATLAB application named 'DataCleaningTool.' This application is able to identify the potential data problems and report results such that the users can make informed decisions to clean data effectively.

Statement of Need

The primary ideas behind developing DataCleaningTool are the following.

1. **Time effectivity** - Data cleaning is a time consuming task to go through manually such large number of datasets for identifying the errors. DataCleaningTool is a tool which can help with the task.
2. **Cooperativeness** - DataCleaningTool is not a black box which implies that it does not produce any result which cannot be easily understandable by the user. It motivates and illustrates its suggestions. However, the user is always in control who takes the final action at every stage of the data cleaning process.
3. **Addressing reasonable number of data problems which cause erroneous conclusion and failing algorithms** - DataCleaningTool aims to clean data by resolving inconsistencies, smoothing noisy data, removing outliers or imputing missing observations using model based imputation method.

Method

DataCleaningTool provides various assistances to users to clean data quickly and efficiently. DataCleaningTool is shown in figure [Figure 1](#).

37 Automated Assistances

- 38 1. **Automated Statistical Information of Data** - DataCleaningTool displays input data
39 in tabular format. It also shows data properties.
- 40 2. **Automated Data Type Discovery** - DataCleaningTool discovers basic statistical data
41 types such as numerical, text and datetime. Moreover, it shows descriptive statistics of
42 the different data types.

43 User Assistances

- 44 1. **Removal of Unwanted Data** - DataCleaningTool can identify irrelevant observations
45 which do not fit the specific problem that the user is trying to solve. It can either
46 replace an irrelevant observation with a missing observation or delete any row containing
47 irrelevant observations.
- 48 2. **Outlier Detection** - DataCleaningTool can illustrate possible outliers. It can either
49 replace an outlier with a missing observation or delete any row containing outliers.
- 50 3. **Missing Data Handling** - DataCleaningTool can illustrate missing observations. It can
51 delete rows or features containing large number of missing observations. It can impute
52 missing observations using missForest method (Stekhoven & Buhlmann, 2011). It shows
53 expected error of imputation for both numerical and categorical features.
- 54 4. **Encoding** - DataCleaningTool can encode categorical features into numerical equiva-
55 lents. Both label encoding and one-hot encoding approaches are implemented in the
56 application.
- 57 5. **Data Transformation** - DataCleaningTool can standardize, normalize, logarithm, expo-
58 nential, squareroot or inverse transform numerical features. It can illustrate transformed
59 numerical features.
- 60 6. **Interactive Data Visualization** - DataCleaningTool can explore and manipulate various
61 aspects of graphical representation of data by clicking on a button or moving a slider.
62 It provides different visual methods such as histogram, bar graph, box plot, missingness
63 plot and bar chart.

64 Related Research and Software

- 65 Similar research projects that are associated with data cleaning include development of the
66 following software packages.
- 67 ■ datacleaner (Olson & McGinnis, 2016) is a Python library which only works with data
68 in pandas DataFrames. It does the following tasks: deletes features with missing ob-
69 servations, replaces missing observations with mean for continuous features and median
70 for categorical features, and encodes categorical features into numerical ones.
 - 71 ■ dataMaid (Petersen & Ekstrøm, 2019) is a R package. It summarizes each feature in
72 the data and encounters several data problems such as extra whitespace, missing data
73 and outliers, depending on the datatype of the feature.

74 Figures

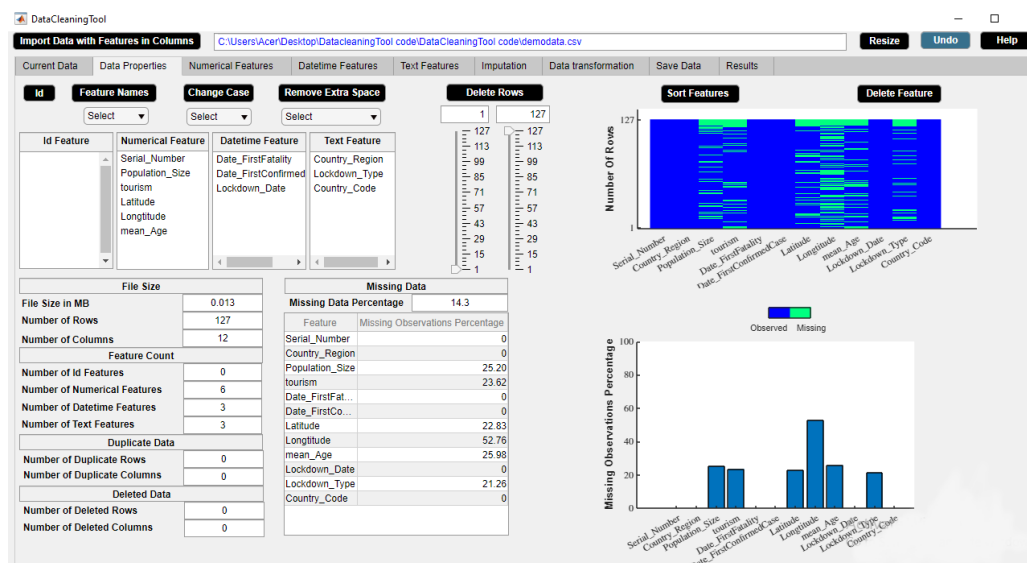


Figure 1: Data Properties Widget. DataCleaningTool consists of multiple widgets and buttons. Each widget illustrates its corresponding data cleaning mechanism and each button aims to deal with a specific data problem.

75 Acknowledgements

76 The authors wish to thank Klara Jansson, Volvo Group Trucks Technology for the helpful
77 discussions during the development of the application.

78 References

- 79 Olson, R., & McGinnis, W. (2016). *Datacleaner: Basic functionality* (Version v0.1.4) [Com-
80 puter software]. Zenodo. <https://doi.org/10.5281/zenodo.47063>
- 81 Petersen, A. H., & Ekstrøm, C. T. (2019). dataMaid: Your assistant for documenting
82 supervised data quality screening in R. *Journal of Statistical Software*, 90(6), 1–38.
83 <https://doi.org/10.18637/jss.v090.i06>
- 84 Stekhoven, D. J., & Buhlmann, P. (2011). MissForest–non-parametric missing value im-
85 putation for mixed-type data. *Bioinformatics*, 28(1), 112–118. [https://doi.org/10.1093/](https://doi.org/10.1093/bioinformatics/btr597)
86 [bioinformatics/btr597](https://doi.org/10.1093/bioinformatics/btr597)