BIOMEDICAL DATA SCIENCE

Project Report
Data Quality Labelling

Submitted by:

Igor Czudy
Anna Panfil
Juras Lukaševičius

Professor:

Carlos Sáez Silvestre

ACADEMIC YEAR 2023/2024

Contents

1 Introduction

In the field of data analysis, we find that a valuable metric for designing accurate models and obtaining various statistics is the quality of the data itself. According to Gartner's report, 40% of businesses fail to achieve their business targets because of poor data quality issues. The importance of utilizing high-quality data for data analysis is realized by many data scientists, and so it is reported that they spend about 80% of their time on data cleaning and preparation. This means that they spend more time on pre-analysis processes rather than focusing on extracting meaningful insights (Ziad 2021).

Seeing how data quality has a direct and profound impact on the outcome of analysis, it is of the utmost importance to differentiate between low and high-quality data sets. One way of doing so is by checking various rules and finding inconsistencies in the data set, then providing the user with a score, based on how well the data passed each requirement. In this report, we will describe our created badges to check for data quality based on different principles: completeness, consistency, uniqueness, etc. We will display an app that automatically calculates the scores of a given data set and displays a general score of data quality. Also, we will describe the possible issues that arise when making such universal badges when applying them to all data sets.

It must be noted that the data quality metrics in this report are created based on our knowledge and found literature surrounding the topic. The app creates a general guideline of the data set's quality but shouldn't be used as the only tool for verifying a data set's quality. Even so, as highlighted in Haug's paper, the usefulness of such scores is not to find the perfect data but to find data of a quality level where the costs of the maintenance work do not exceed savings from the costs inflicted by poor quality data (Haug, Zachariassen, and Van Liempd 2011). Therefore these badges may give the user an important threshold of quality to measure the later required length of work in data wrangling[1].

[1] **Data wrangling** – the process of converting raw data into a usable form.

2 Data Quality Measurements

An effective methodology for data quality conceptualization is the **ALCOA** principles, which stand for *attributable, legible, contemporaneous, original* and *accurate*. The extended **ALCOA+** put additional emphasis on the attributes of being *complete, consistent, enduring* and *available* (Organization et al. 2019). The extended **ALCOA+** principles are what we aimed to display with our generated badges, seeing as the **ALCOA+** principles are a common data quality regulatory measure in highly regulated areas, such as the pharmaceutical industry (see Leal et al. 2021), but we found that it wasn't a good fit for basing the design of our data quality measures, considering that *attributable, legible, contemporaneous* and *original* data is rare to come by. There is little focus on the universal application of **ALCOA+** principles for all data sets in research, so we set out to create universal badges that would fit any data around five high-level dimensions, highlighted by Weiskopf and Weng: *completeness, correctness, concordance, plausibility* and *currency* (Weiskopf and Weng 2013). The workings of our created metrics and their results are further explained in this section.

2.1 Completeness

Completeness focuses on exemplifying whether there is a truth about each data point or, as described by Weng and Strong, the extent to which data are of sufficient breadth, depth, and scope for the task at hand (Wang and Strong 1996).

Many types of analysis require all of the values to be present for select variables. Therefore, a lack of data either leads to eliminating features or replacing them with other values. The first approach deprives the analysts of valuable data and the second creates a new problem of choosing the best method to replace these values, additionally distorting the data.

2.1.1 Missing values

Completeness can be measured by the percentage of missing values. In this work all `None`, `''`, `null`, `n/a`, `nan`, and `none` values are treated as missing, taking into account different capitalization. Completeness of both all data at once and each column separately is considered. That produces the metric of missing percentages, which includes the percentage of missing data in the whole data set, and the most missing column with the maximum percentage of missing values in one column.

Badges described: `missing percentage` , `most missing column`
Badge interpretation: between 0 (no missing values) and 1 (all of the values are missing).

2.2 Correctness

Correctness can be described as the degree of accuracy and precision of data records with respect to their real-world states. Using the correctness method, we can highlight disturbances in the existing data regarding their statistical extremes.

2.2.1 Outlier detection

The outliers percentage metric allows us to measure the correctness on the individual level. Based on the possible distribution of the data we can mark some values as suspicious. These values can be either incorrect or abnormal, leading to difficulties in deriving overall conclusions.

With this metric, we seek to detect outliers both in numerical and categorical features. For the numerical variables, outliers are determined using quantiles. If said point is greater than

$Q3 + 1.5\ IQR$ or less than $Q1 - 1.5\ IQR$, similarly as in the box plot, we mark it as an outlier. In categorical attributes, a category is marked as an outlier if it appears in less than 5% of the observations. This percentage was set empirically. Similarly to completeness, the column with the most outliers is also stated.

Badges described: `outliers percentage` , `most outliers column`
Badge interpretation: between 0 and 1, 0 being no outliers.

2.2.2 Mish-mashed cases

The max mish-mashed cases badge is used for categorical data. This badge is determined by comparing the number of unique values in a particular column to the number of *truly unique values*. *Truly unique values* are unique values that were converted to lowercase equivalents. The final result of the mish-mashed badge is the maximum found difference between unique and truly unique values divided by unique values for each column:

$$MMC = max\left(\frac{UD - TUD}{UD}\right),$$

Where,
- MMC are the max. mish-mashed cases,
- UD are the unique values,
- TUD are the truly unique values.

This metric provides information about human error in data caused by a lack of standardization.

Badge described: `max mishmashed case`
Badge interpretation: between 0 and 1, 0 meaning no case mish-mashing.

2.3 Concordance

Concordance describes the degree to which data semantics and statistics are concordant among multiple data sources. In general, concordance takes into account the data consistency and reliability, also noting its variation between the variables. In other terms, it answers whether there is agreement between elements in the data set.

2.3.1 Duplicated observations

Duplicated information brings nothing but trouble when it comes to working with data. Low mentions the adverse effect of duplicate data in his paper, stating that it negatively affects a manager's ability to make logical and well-informed decisions (Low, Lee, and Ling 2001). This data gives off the illusion of having many observations, yet ultimately leads to eliminating duplicated instances. That is also why the duplication percentage badge is one of the vital metrics included in this work.

Badge described: `duplication percentage`
Badge interpretation: between 0 and 1, 0 meaning no duplicates.

2.3.2 Correlation

Another type of duplicated information, this time between attributes, is correlation. Correlation is a statistical measure that quantifies the degree to which two variables are related or move together. It is often used to assess the strength and direction of a linear relationship between two variables.

It is known that a high correlation between features will not bring additional information (or just very little) but will increase the complexity of the algorithm, thus increasing the risk of errors. That is why it is proposed that the higher the correlation between features results in a worse data quality score.

In this work, two correlation badges were used, which are further described below.

2.3.2.1 Numerical correlation Numerical correlation badges use Pearson's correlation, which measures linear correlation. It is a ratio between the covariance of two variables and the product of their standard deviations, as shown in the formula below:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \cdot \sum (y_i - \bar{y})^2}} \tag{1}$$

Figure 1: Pearson correlation formula

This formula does not support empty values, which is why, for the purpose of this task, correlation is calculated only for the existing values.

Thereafter, a count of all sums of pairs of features whose correlation is bigger than 0.8 (without comparing attributes with themselves) is done. Finally, this sum is divided by the number of numerical columns and displayed as a metric.

Badge described: `correlation numerical`
Badge interpretation: between 0 and 1, 0 meaning no correlation.

2.3.2.2 Categorical correlation The categorical correlation badge uses a chi-square test to check for independence. This test is commonly used for checking if there is a significant association between two categorical variables.

The hypotheses for this test are as follows:

H_0 : Two variables are independent
H_1 : Two variables are **not** independent

After applying this test to specific pairs of categorical columns, we check whether the hypothesis is rejected or not. We examine whether the p-value is lower than the threshold, set to 0.01.

The bottom line of the categorical correlation badge is a division of independent features and the sum of independent and dependent features. To get value 1 as the worst value and 0 as the best, we subtract this result from 1.

Badge described: `correlation categorical`
Badge interpretation: between 0 and 1, 0 meaning no correlation.

2.3.3 Unique and dominated columns

Columns such as `id` can be very useful for distinguishing examples but do not bring a lot when it comes to conclusions. That is why a unique column percentage can be used as a metric of data quality.

A similar situation can be proposed with columns dominated by one value. Therefore a feature with more than 80 % of the same values is marked as such.

Badges described: `unique columns` , `dominated columns`
Badge interpretation: between 0 and 1, 0 meaning non-existent.

2.4 Plausibility & Currency

Plausibility focuses on the measure of whether each point in the set makes sense in the real-life context of the data, while currency points to the timeliness of the data – is the data point relevant at a given time? These measures require contextual information about the data set and understanding its purpose for further analysis.

2.4.1 Documentation

Since a direct approach to interpreting the plausibility and currency of a given data set ultimately requires the participation of the user, we can't solely use the data set to extrapolate a metric for these features. Therefore, we can look for such information elsewhere. Such sources can be relevant documentation or metadata. As noted by Satija, metadata is the modern equivalent of catalogued library resources, which effectively gives us a wider understanding of the analyzed material (Satija, Bagchi, and Martínez-Ávila 2020).

For this badge, key phrases are scanned in an additionally provided informational document. These key phrases include synonyms for variable definitions, formatting constraints, allowable ranges, the moment, place and author of data acquisition, rules, and information about significant derived variables. The number of unique information points is counted and divided by the number of possible information points. It must be noted that not all possibilities in the documentation are accounted for because of the limited size of word banks used.

Badges described: `missing documentation`
Badge interpretation: between 0 and 1, 0 meaning all necessary documentation was provided.

2.5 Overall score

To calculate the overall score badge, we use a weighted average. The formula for the label is provided below:

$$\mathrm{OC} = \frac{\sum_{i=1}^{n}(\mathrm{BS}_i \times \mathrm{w}_i)}{\sum_{i=1}^{n} \mathrm{w}_i}$$

Where,
- OC is the overall score,
- n is the number of badge,
- BS_i is the score of the i-th badge,
- w_i is the weight assigned to the i-th badge.

Using the weights allows the user to highlight the different importance of each badge, depending on the goal of analysis. Default badge weights are provided but can also be customized

on demand depending on the user's needs. The table of badge weight provided by the authors is given in table 1.

Badge	Weight
missing percentage	10
most missing column	2
duplication percentage	4
outliers percentage	2
most outliers column	1
unique columns	5
dominated columns	3
max mishmashed case	1
correlation numerical	4
correlation categorical	1
missing documentation	2

Table 1: Weights table

The authors have decided that the most important badge is the one about missing values. There are many possibilities to handle missing values, but, in general, they all have an unfavorable effect. Also, the unique columns badge has a high weight. This is due to the fact that unique columns do not allow for good generalization. Badge duplication percentage gets a weight equal to four. It is like that because many duplicates in the data set, in turn, lower its dimensionality (there are repeated rows in need of removal). Similarly for the numerical correlation badge, which has the same weight. More and more correlated features do not provide any new information. If some column in the data set has one dominated category, it is also not very useful. That is why the dominated columns badge has a weight equal to 3.

All other badges, according to the authors, are not as important and have weights equal to one or two. To be more precise, the most missing columns badge informs about one column with the highest number of missing values, but it concerns only one column, which is why it is not very important. The outliers percentage badge informs about outliers. However, it's not difficult to remove outliers or scale data. This is similar to the case of the max mish-mashed badge.

The presence of documentation is seen as a positive indicator for the data set; however, many data sets lack proper documentation and are formatted in many non-standardized ways.

As mentioned earlier, a correlation between columns doesn't necessarily reflect well on the data set. The low weight assigned to the correlation categorical badge is because, after several tests, it appears to underestimate the data set and provides information that certain features are overly correlated.

Badge described: `unique quality score`
Badge interpretation: between 0 and 1, 0 meaning poor quality.

3 Data set generator

For testing purposes, a data set generator was created. It is implemented using the builder pattern and facilitates the creation of a *.csv* file with certain characteristics. The generated data includes numerical, date, and categorical features, including unique attributes. Users have the flexibility to specify the size of the data set and the desired amounts of missing data, duplicates,

outliers, and mish-mashed case categorical values. Additionally, they can introduce a dominated column with a specified percentage of majority values.

4 Web application

The authors, alongside this document, also provide a web application allowing users to test the data labeler on any data set. The application accepts *.csv* files with the data of choice and, optionally, a documentation file. After that, scores of all badges and the final score with set weight values will be displayed. There is also a possibility to manipulate all weights of the final score and customize it using sliders. Below are screenshots of the described web application:

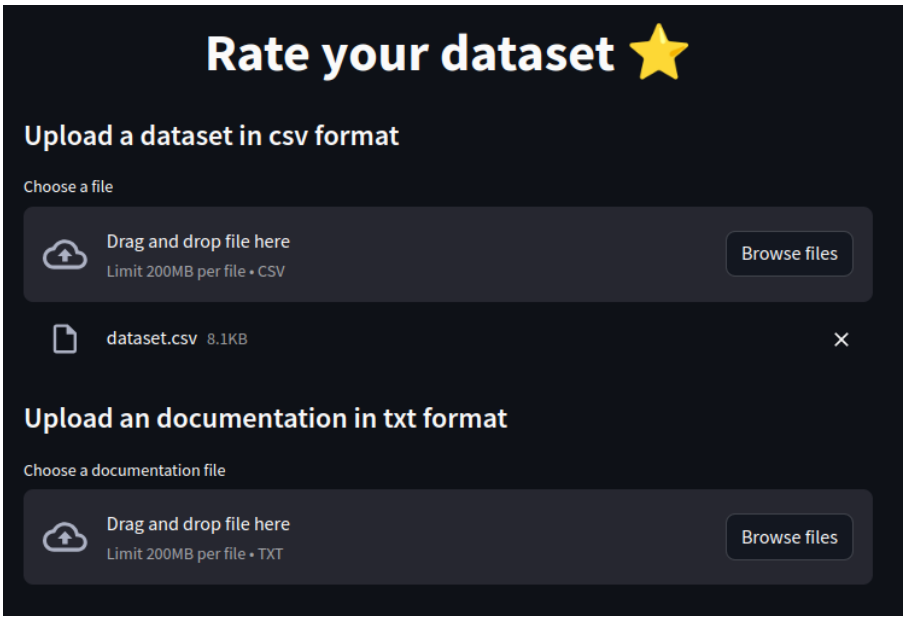

Figure 2: Web application – drop data set

Figure 3: Web application badge scores

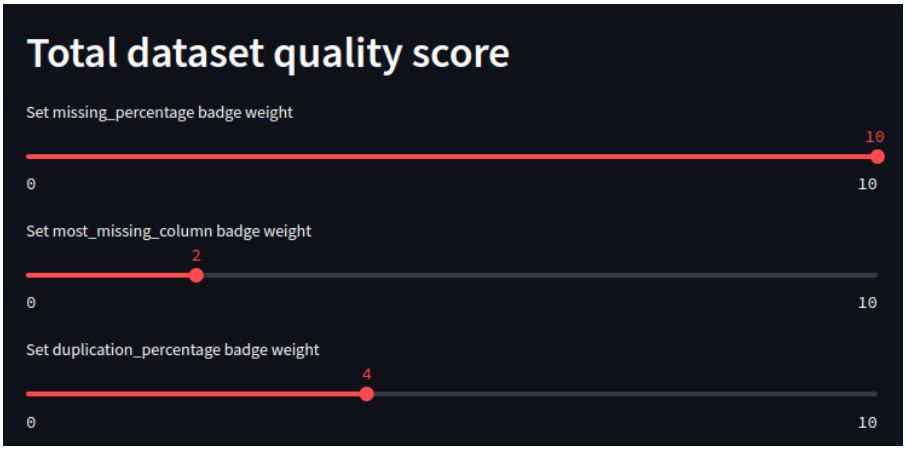

Figure 4: Web application sliders

5 Testing

The following part is a display of data quality results for a variety of data sets. A short summary of the data used:

- **Generated** data set using Python, consisting of numerical (`results1` , `results2`) and categorical (`name` , `surname` , `birthdate` , `category` , `email` , `gender`) variables; generated with a command:

```
FakeDataset(dataset_size = 100)\
    .add_dominated_string_column(dominated_percentage=0.9)\
    .add_mishmashed_case(mishmashed_percentage=0.1)\
    .add_outliers_above(outlier_percentage = 0.1)\
    .add_duplicates(duplicate_percentage = 0.15)\
    .add_missing(missing_percentage = 0.1)\
    .to_csv(filename)
```

- **iris** data set with info about plants three plant species [source];
- **Lottery Powerball Winning Numbers** data set with winning lottery numbers from 2011 with the number and its multiplier [source];
- **name_mapping** data set acquired from the radiomics lab assignment in Biomedical Data Analysis;
- **Titanic survival** data set, consisting of numerical (`Age` , `SibSp` , `Parch`) and categorical (`PassengerId` , `Survived` , `Pclass` , `Name` , `Sex` , `Ticket` , `Fare` , `Cabin` , `Embarked`) variables [source].

The badge weights are as assigned by the authors. All data sets are provided with the code.

Data	GDS+D[1]	iris	Lottery+D[2]	name_maping (radiomics)	titanic
Missing percentage	0.1	0	0.04	0.18	0.1
Most missing column	0.1	0	0.13	0.55	0.77
Duplication percentage	0.02	0.02	0	0	0
Outliers percentage	0.01	0.01	0.02	0	0.04
Most outliers column	0.1	0.03	0.07	0	0.24
Unique columns	0.75	0	0.67	0.67	0.25
Dominated columns	0.12	0	0	0	0
Max mish-mashed case	0.43	0	0	0	0
Correlation (numerical)	0	0.75	0	1	0
Correlation (categorical)	0.4	1	0	0	0.33
Missing documentation	0	1	0.86	1	1
Overall Score	0.82	0.83	0.83	0.65	0.82

[1] Generated data set with additionally provided documentation;
[2] Lottery Powerball Winnings Beginning from 2010 with converted provided documentation.

Table 2: Testing results

When applying the Data Quality Labeller application to a collection of randomly chosen data sets, we find the following results:

- all data sets do not have more than 20% missing values. The highest number of missing values can be found in the **name_maping** data set (18%);
- the **name_mapping** and **titanic** data sets had two of the least informative columns: one was missing 55% of values, the other - 77%;
- no data sets have shown high duplication or correlation percentages;
- **titanic** is the data set with the most outliers in a column - 24%. It must be noted, that **titanic** has an undefined key value, which is considered by the labeler as numerical. This might impact the outlier information.
- the **generated**, **lottery** and **name_mapping** data sets show a high percentage of unique columns, yet none of the data sets feature many dominating columns;
- only the **generated** data set features a column with mish-mashed cases, of which there are 43% of;
- the highest numerical correlation is found between the **iris** and **name_mapping** variables, and the highest categorical correlation is found in the **iris** data set. Overall, we can tell that many of the **iris** variables are highly correlated;
- out of the two data sets that have additional documentation, the **generated** data set has shown the best results, while the **lottery** data set only matched a small number of requirements and output a 0.86 score.
- according to our labeler, all data sets show a good score of over 80%, except for the **name_mapping** data set, which shows the overall score of 65%.

6 Conclusions

The designed Data Quality Labelling tool shows potential but needs a lot of polish until it can be used reliably for any applied data set. The tool is effective and useful when determining missing values, duplicates, and outliers, but it struggles to measure its quality in the context of the data. It is also highly explainable. Some suggestions for improvement could be an implementation of the ability to read other types of data and documentation in the web application. Since context is lost on a sole data set, an interesting proposal would be to cross-check the data online on acquisition – if the data is acquired online, check for keywords and phrases in the web page to link with the data frame and later create badges by association.

7 References

[1] Anders Haug, Frederik Zachariassen, and Dennis Van Liempd. "The costs of poor data quality". In: *Journal of Industrial Engineering and Management (JIEM)* 4.2 (2011), pp. 168–193.

[2] Fátima Leal et al. "Smart pharmaceutical manufacturing: Ensuring end-to-end traceability and data integrity in medicine production". In: *Big Data Research* 24 (2021), p. 100172.

[3] Wai Lup Low, Mong Li Lee, and Tok Wang Ling. "A knowledge-based approach for duplicate elimination in data cleaning". In: *Information Systems* 26.8 (2001), pp. 585–606.

[4] World Health Organization et al. "Guideline on data integrity". In: *WHO Drug Information* 33.4 (2019), pp. 773–793.

[5] MP Satija, Mayukh Bagchi, and Daniel Martínez-Ávila. "Metadata management and application". In: *Library Herald* 58.4 (2020), pp. 84–107.

[6] Richard Y Wang and Diane M Strong. "Beyond accuracy: What data quality means to data consumers". In: *Journal of management information systems* 12.4 (1996), pp. 5–33.

[7] Nicole Gray Weiskopf and Chunhua Weng. "Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research". In: *Journal of the American Medical Informatics Association* 20.1 (2013), pp. 144–151.

[8] Zara Ziad. *Using Machine Learning to Automate Data Cleansing*. Source. Published: 2021-03-08. 2021.