

Assessing Data Quality Inconsistencies in Brazilian Governmental Data

Gabriel P. Oliveira   [Universidade Federal de Minas Gerais | gabrielpoliveira@dcc.ufmg.br]

Bárbara M. A. Mendes   [Universidade Federal de Minas Gerais | barbaramit@ufmg.br]


Clara A. Bacha   [Universidade Federal de Minas Gerais | clarabacha@ufmg.br]

Lucas L. Costa   [Universidade Federal de Minas Gerais | lucas-lage@ufmg.br]


Larissa D. Gomide   [Universidade Federal de Minas Gerais | larissa.gomide@dcc.ufmg.br]

Mariana O. Silva  [Universidade Federal de Minas Gerais | mariana.santos@dcc.ufmg.br]

Michele A. Brandão  [Instituto Federal de Minas Gerais | michele.brandao@ifmg.edu.br]

Anísio Lacerda   [Universidade Federal de Minas Gerais | anisio@dcc.ufmg.br]

Gisele L. Pappa  [Universidade Federal de Minas Gerais | glpappa@dcc.ufmg.br]

 Computer Science Department, Universidade Federal de Minas Gerais, Av. Presidente Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, 31270-901, Brazil.

Received: 6 March 2023 • Published: 20 October 2023

Abstract In recent years, vast volumes of data are constantly being made available on the Web, and they have been increasingly used as decision support in different contexts. However, for these decisions to be more assertive and reliable, it is necessary to ensure data quality. Although there are several definitions for this area, it is a consensus that data quality is always associated with a specific context. This work aims to analyze data quality in a data warehouse with governmental information of the Brazilian state of Minas Gerais. We first present a brief comparison of eight open-source data quality tools and then choose the Great Expectations tool for analyzing such data in two real applications: public bids and public expenditure. Our analyses show that the chosen tool has relevant characteristics to generate good data quality indicators to reveal data quality issues that may directly impact the construction of final applications using such data.

Keywords: data quality, governmental data, great expectations, public bids, public expenditure

1 Introduction

The statement “Data is the new oil” made by the British mathematician Clive Humby¹ says a lot about the importance of building and maintaining data with quality. More and more decisions are being made based on data, especially in a reality where huge volumes of data are constantly available on the Web². Thus, data must be reliable for these decisions to be more assertive and precise [Medeiros *et al.*, 2020; Junior and Dorneles, 2021].

The area of data quality emerges in this context. Although there are several definitions for this area, it is a consensus that it is always associated with a specific context [Junior and Dorneles, 2021]. In other words, a given data set may be suitable for one scenario, but not another, or data has quality when it is “fitness for use” [Wang *et al.*, 2018]. Therefore, many works analyze quality in a specific domain [Cichy and Rass, 2019]. Another definition concerns multiple dimensions, identified by attributes, representing specific characteristics of the data [Scannapieco and Catarci, 2002; Medeiros *et al.*, 2020].

Therefore, this work aims to analyze data quality in a data warehouse with governmental spending information within the Brazilian state of Minas Gerais. The primary motivation is the identification of inconsistencies that may impact

the analyses carried out on public bids and expenditures. In this way, we use several quality indicators, *i.e.*, metrics that evaluate specific rules in the data. For example, considering a column that stores percentage values, an indicator can determine whether all records in that column range between 0% and 100%.

This work analyzes eight open-source tools that consider different data quality dimensions. The selection of these tools considers whether the tool is open-source and is easy to use in a way that allows to reproduce the methodology proposed here. After comparing their functionalities, we select the Great Expectations (GE) tool as the most appropriate for our context, as it verifies quality problems and reports them to the users in an automated way. This tool has several indicators implemented natively, in addition to the possibility of developing customized indicators, through which it is possible to implement business rules specific to the context of the analyzed data. GE also has a component for generating an interactive graphical interface with the results of the indicators. After selecting such a tool, we propose a novel methodology for assessing data quality analysis using GE.

This article extends a full paper from the 37th Brazilian Symposium on Databases [Oliveira *et al.*, 2022b]. As a new material, we introduce a new application of GE in public expenditure data and the existing application in bidding data. Furthermore, we propose a new quality metric that compares tables from different applications. Overall, the results show

¹Data is the new oil: <https://bit.ly/DataTheNewOil>

²A minute on the Internet: <http://bit.ly/3rdWUPf>

that using GE allows the identification of quality problems that would not be easily identified. Moreover, the proposed quality metric can help quality analysts determine the priority of the tables for further manual inspection. Thus, it could accelerate the resolution of problems in important tables.

The remainder of this article is organized as follows. Related work is presented in Section 2. Next, Section 3 describes the comparative analysis of data quality tools. Section 4 presents the methodology steps for data quality analysis using Great Expectations. Sections 5 and 6 presents the results from data quality analysis of public bidding and public expenditure data, respectively. Then, Section 7 introduces the new quality metric based on GE results. Finally, in Section 8, we present our conclusions and future work.

2 Related Work

The term “data quality” is related to a set of characteristics that data must have. These properties are called dimensions, which include, for example, precision, completeness, and consistency [Scannapieco and Catarci, 2002; Medeiros et al., 2020]. The process of managing this quality comprises four practices, namely: (i) *data profiling*, to create an overview of the data and identify how they are stored [Cichy and Rass, 2019]; (ii) *data quality measurement*, consisting, for example, of identifying missing data, outliers and corrupted information [Lee et al., 2002; Ehrlinger and Wöß, 2018]; (iii) *data cleaning*, to remove unwanted data [Elmagarmid et al., 2007]; and (iv) *data quality monitoring*, to maintain the data quality principles in a team, and to create/use tools and processes to be applied in the previous steps [Pipino et al., 2002; Laranjeiro et al., 2015].

Data quality must be defined in its context of use, as the same dataset may need different indicators depending on the needs of its users [Ballou and Pazer, 1985]. The importance of data quality has been noted in many different contexts, including cartography [Chrisman, 1983], biology [Etcheverry and Consens, 2011], and medicine [Goudar et al., 2015; Zöllner et al., 2016]. Other studies use data visualization techniques to support quality analyses on abstract and timeless data [Josko and Ferreira, 2021]. Furthermore, given the need for training artificial intelligence models, Sessions and Val-torta [2006] present an analysis of the effects of data quality on machine learning algorithms, demonstrating the importance of applying these concepts.

Thus, to analyze the quality of large volumes of data in different contexts, automated methods are required, resulting in a vast market of tools for this purpose. In this sense, previous works aim to compare data quality tools. For example, Pushkarev et al. [2010] evaluate seven tools *open-source* or with free trial periods using criteria such as connectivity, management, interface, and functionalities. Gao et al. [2016] analyze eight commercial tools considering their functionalities. Furthermore, Altendeitering and Tomczyk [2022] propose a taxonomy for data quality and analyze 18 tools in this context. Finally, a more extensive study is carried out by Ehrlinger and Wöß [2022], who analyze 667 different quality tools. The authors use a set of exclusion criteria to select 13 tools (eight commercial and five open-source) for further

comparison.

Although data quality is a research topic that has been extensively studied in different contexts, analyzing the quality of government data is still an area in constant expansion. In this sense, Wu et al. [2022] analyze the quality and applicability of open government data related to COVID-19 in the US, EU, and China. The results show that the data still lacks the necessary metadata.

To the best of our knowledge, existing work on government data does not perform quality analysis on public expenditure data. Thus, this work expands Oliveira et al. [2022b] to analyze open-source tools applied to this context and present two applications with real-world data. Ensuring data quality in this context is fundamental for further applications, such as detecting fraud in public bids and other prediction and recommendation tasks [Maia et al., 2020; Oliveira et al., 2022a].

3 Data Quality Tools

This section presents a comparative analysis of data quality tools selected from pre-defined criteria. Section 3.1 describes the tool selection criteria and all the considered tools. Then, Section 3.2 presents the comparison results of each tool’s functionalities. Finally, in Section 3.3, the *Great Expectations* tool is described in detail, as it is the tool that best meets our selection criteria.

3.1 Considered Quality Tools

We use two studies as a basis for selecting data quality tools. The first one presents a systematic review of 667 tools and applies exclusion criteria for reaching the final set of 13 tools considered in its comparative analyses [Ehrlinger and Wöß, 2022]. Such criteria mainly check if the tools are designed for specific tasks and domains and if they are publicly available or with a free trial period. The second work explores three additional tools, which are not considered in the first work [Foidl et al., 2022].

With the set of 16 tools pre-selected based on the two works mentioned above, we also include an extra criterion, which evaluates whether a tool is open-source and aims to guarantee the possibility of using the tool easily and reproducing the methodology proposed here. After considering all such criteria, our selection process resulted in eight data quality tools. Next, we describe each of them with reference to the source article in which the tool was presented.

Aggregate Profiler (AP)³. This tool includes an integrated data management platform that, in addition to features related to data preparation, also provides data cleansing, statistical analysis, pattern matching, and data profiling [Ehrlinger and Wöß, 2022].

Apache Griffin (AG)⁴. This tool focuses on big data and is dedicated to continuously measuring batch or streaming data quality. AG offers a set of well-defined data quality

³Aggregate Profiler: <https://sourceforge.net/projects/dataquality/>

⁴Apache Griffin: <https://griffin.apache.org/>

Table 1. Feature comparison of the considered data quality tools.

| # | Features | Aggregate Profiler | Apache Griffin | Great Expectations | MobyDQ | OpenRefine & Metric | PyDeequ | Talend Open Studio | Tensorflow Data Validation |
|----|--|--------------------|----------------|--------------------|----------|---------------------|----------|--------------------|----------------------------|
| 1 | Table formatting | <i>c</i> | <i>p</i> | ✓ | <i>c</i> | <i>p</i> | <i>p</i> | <i>p</i> | ✓ |
| 2 | Restrictions on values | <i>c</i> | <i>p</i> | <i>c</i> | <i>c</i> | <i>p</i> | <i>p</i> | <i>c</i> | <i>p</i> |
| 3 | Range of values | <i>p</i> | <i>p</i> | <i>c</i> | <i>c</i> | <i>p</i> | <i>p</i> | <i>c</i> | <i>p</i> |
| 4 | String matching | <i>c</i> | <i>p</i> | <i>p</i> | ✗ | <i>p</i> | <i>p</i> | <i>p</i> | <i>p</i> |
| 5 | Timestamp and JSON | ✗ | ✗ | ✓ | <i>p</i> | <i>p</i> | ✗ | <i>p</i> | ✗ |
| 6 | Aggregation functions | <i>p</i> | <i>p</i> | <i>p</i> | <i>p</i> | ✗ | <i>p</i> | <i>p</i> | <i>p</i> |
| 7 | Multi-column operations | <i>p</i> | ✗ | <i>p</i> | <i>p</i> | ✗ | <i>p</i> | ✗ | ✗ |
| 8 | Functions related to probability distributions | ✗ | ✗ | <i>p</i> | <i>p</i> | ✗ | ✗ | ✗ | <i>p</i> |
| 9 | Functions related to files | <i>p</i> | ✗ | ✓ | <i>c</i> | <i>p</i> | ✗ | ✗ | ✗ |
| 10 | Custom indicators | ✓ | ✗ | ✓ | ✓ | ✗ | ✗ | ✓ | ✗ |

domain models, which cover different data quality problems [Ehrlinger and Wöß, 2022].

Great Expectations (GE)⁵. GE is an open-source library for validating, documenting, and characterizing data. Its operation is based on the concept of test automation from Software Engineering, making it possible to attest to data quality based on what is expected [Foidl *et al.*, 2022].

MobyDQ⁶. A tool for automating data quality checks during data processing, capturing metric results, and triggering alerts in case of anomalies. MobyDQ was inspired by an internal project by Ubisoft Entertainment to measure and improve the data quality of its Enterprise Data Platform. However, its open-source version has been reformulated to improve its design and remove technical dependencies with commercial software [Ehrlinger and Wöß, 2022].

OpenRefine & Metric (ORM)⁷. A tool dedicated to cleaning and transforming data, operating on structured data (rows and columns), similar to how relational tables work. Specifically, ORM projects consist of a table whose rows can be filtered using defined criteria [Ehrlinger and Wöß, 2022].

PyDeequ⁸. A Python API for Amazon Deequ, a library which aims to perform “unit tests” on data, i.e., to measure data quality according to pre-established rules and conditions [Foidl *et al.*, 2022].

Talend Open Studio (TOS)⁹. The Talend company offers two products for data quality: Talend Data Management Platform and Talend Open Studio (TOS). The prior requires a paid subscription, while the latter is a free, open-source tool. Both products (Open Studio and Enterprise) offer good support for Big Data analytics (e.g., Spark or Hadoop) and a variety of profiling and data cleansing functionalities [Ehrlinger and Wöß, 2022].

Tensorflow Data Validation (TFDV)¹⁰. A library for exploring and validating machine learning data. TFDV is designed to be highly scalable and work well with TensorFlow and TensorFlow Extended (TFX) [Foidl *et al.*, 2022].

3.2 Feature Comparison

Here, we compare the eight open-source tools regarding their respective functionalities. Following a methodology similar to Ehrlinger and Wöß [2022], we define a catalog of evaluation requirements listed in the first column of Table 1. Our goal is to classify the fulfillment of each requirement into four categories: (✓) met, (✗) not met, (*p*) partially met, and (*c*) available after customization. In particular, the *c* category indicates the possibility of implementing custom indicators, making it possible to fulfill any requirement not available natively in the tool.

To perform the comparative analysis and define the requirements catalog, we evaluate only each tool’s documentation. Therefore, we do not include any functionality that is not mentioned in the documentation in the analysis. The final set of requirements contains functionalities related to the following categories: (#1) table formatting, such as size and existence of rows/columns; (#2) restrictions on values; (#3) range of values; (#4) pattern matching in strings; (#5) dates and JSON format; (#6) data aggregation functions; (#7) multi-column operations; (#8) functions related to probability distributions; and (#9) file-related functions. In addition to these nine categories, we also include one (#10) related to the possibility of creating custom indicators¹¹.

Table 1 shows that the most basic features (1–4) are covered by most tools, either entirely or partially. Most of the more sophisticated functionalities (5–9), such as functions related to probability distributions and files, are more unusual. As an exception, aggregation functions, despite also

⁵Great Expectations: <https://greatexpectations.io/>

⁶MobyDQ: <https://ubisoft.github.io/mobydq/>

⁷OpenRefine & Metric: <https://openrefine.org/>

⁸PyDeequ: <https://github.com/aws-labs/python-deequ>

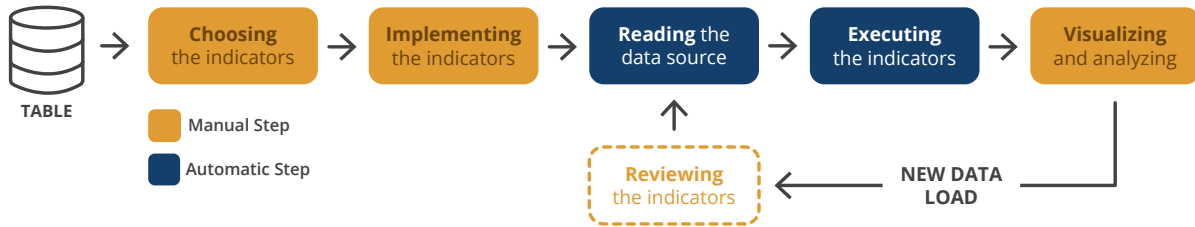
⁹Talend Open Studio: <https://www.talend.com/products/talend-open-studio/>

¹⁰Tensorflow Data Validation: <https://github.com/tensorflow/data-validation>

¹¹The details of each category of functionalities are in the Supplementary Material available at <https://doi.org/10.5281/zenodo.7007428>.

Table 2. Ranking of tools that best fulfill our requirements.

| # | Tool | ✓ | c | p | ✗ | Additional components |
|---|--------------------|-----|-----|-----|-----|-------------------------------|
| 1 | Great Expectations | 40% | 20% | 40% | 0% | Profiler, Graphical interface |
| 2 | MobyDQ | 10% | 40% | 40% | 10% | Graphical interface |
| 3 | Aggregate Profiler | 10% | 30% | 40% | 20% | Graphical interface |
| 4 | Talend Open Studio | 10% | 20% | 40% | 30% | – |

**Figure 1.** Methodology for data quality analysis using Great Expectations.

being a more complex feature, are partially covered by most tools. Finally, regarding the availability of creating customized indicators (#10), we observe that half of the tools meet such a requirement. Thus, even if such tools do not present specific indicators natively, it is possible to implement them in a customized way.

Overall, the tools that least meet the listed requirements are Apache Griffin, OpenRefine & Metric, PyDeequ, and Tensorflow Data Validation. In addition to having few features compared to other tools, they also do not provide customized indicators. In contrast, the tools that best meet the ten features are Great Expectations, MobyDQ, Aggregate Profiler, and Talend Open Studio. All four tools provide the application of business rules, as they provide customization of indicators. In particular, such functionality is essential for the domain analyzed in this study (*i.e.*, governmental data), given that such a context requires specific business rules.

We now rank the four tools mentioned above concerning the best fulfillment of the requirements and their additional components. Table 2 presents this ranking, with the percentage of fulfillment of each category and a list of the respective extra features, if any. Therefore, the tool that best fits the comparative analysis is *Great Expectations*, which in addition to having overcome the other tools in terms of functionality, provides additional relevant components, including *Profiler* and a graphical interface, called *Data Docs*, that shows the results of the executed indicators. Next, we describe the main components and functionalities of Great Expectations.

3.3 The *Great Expectations* Tool

Great Expectations (GE) is an open-source data quality tool that uses a mechanism similar to unit tests for data validation. Each validation is done by a module called *expectation* (here, we call them indicators). GE provides several native indicators that perform generic data validations, such as checking field types, value ranges, and null records. In addition, GE offers the possibility of creating custom indicators, allowing the implementation of specific business rules for each table. Such indicators are coded in Python and integrated into the tool's structure. Thus, they can be used in conjunction with native indicators. We now describe the main components and

functionalities available in GE.

Expectations. Correspond to a set of assertions expressed in declarative language and used for data validation. GE verifies such assertions in the desired table columns and returns the success or failure of the verification as a result. Thus, expectations are the indicators to evaluate the data quality, and they can run on Pandas, Spark, and SQLAlchemy data frames. Finally, the results are returned in a structured format (JSON), which facilitates post-processing tasks.

Profiler. This component performs a pre-analysis and returns a characterization of the data, as well as a collection of indicators that best fit the analyzed data. Such indicators serve as a recommendation of the best validations that should be made on this data.

Data Docs. This component displays the results of the indicators executed on the data. It provides an interactive graphical interface in HTML page format, where the user can navigate the results.

In summary, we choose GE as the best data quality tool for our context because: (i) the customized indicators allow the implementation of quality indicators that assess specific business problems; (ii) *Data Docs* generates a graphical interface containing the results of the executed indicators, facilitating the analysis of the results by final users; and (iii) *Profiler* does a pre-analysis of the structure of the stored data, and then shows an overview of the data with some recommended indicators to be implemented on them.

4 Methodology for Data Quality Analysis using Great Expectations

After choosing Great Expectations as our data quality tool, we propose a methodology for the data quality analysis task using it, as illustrated in Figure 1. This pipeline consists of five main steps and one optional step, from the choice of specific indicators for each table to the visualization and analysis of the results by specialists. Such steps are detailed below.

Choosing the indicators. This is the first step after selecting the table. It consists of a manual inspection of the table's structure and content to define the quality indicators that will

Table 3. Great Expectations indicators used in the analysis of public bidding data.

| | Indicator (expectation) | Description |
|--------|--|---|
| Native | expect_column_values_to_not_be_null | Column values must not be null |
| | expect_column_values_to_be_unique | There must be no duplicate values in the column |
| | expect_column_min_to_be_between | The smallest column value must be within the range [min, max] |
| | expect_column_values_to_be_in_type_list | Column values must be of the specified type |
| | expect_column_values_to_be_in_set | Column values must belong to a value set |
| | expect_column_values_to_be_between | Column values must be in the range [min, max] |
| | expect_column_values_to_match_regex | Column values must follow a given regular expression |
| Custom | expect_value_less_revenue | Bids must have a value less than or equal to the entity revenue |
| | expect_table_fato_licitacao_to_have_guests_if_invite | Bids in invitation mode must have invited bidders |
| | expect_dates_to_match_across_tables | Reported dates must be in valid chronological order |
| | expect_only_one_year_of_activity | Bids must have only a single year of activity |
| | expect_sum_of_item_values_to_match_fato_licitacao | Sum of bidding item values must match bid amount |

be implemented. The person conducting this stage must have technical and business knowledge to ensure that the chosen indicators are adequate. In this step, it is a good practice to use GE's *Profiler* component because it verifies which native indicators are best suited to the analyzed data.

Implementing the indicators. This step refers to the code implementation of the chosen indicators using Great Expectations.

Reading the data source. In this step, the selected table is loaded. It is necessary to read the entire content of the table for the indicators to be executed.

Executing the indicators. After reading the table, the implemented indicators are executed and the results are presented in an interactive graphical interface generated by the *Data Docs* component.

Visualizing and analyzing. The last step of the methodology corresponds to the visualization and analysis of the results of the indicators in the interactive graphic interface. From this analysis, it is possible to verify cases that indicate errors in the loading process and/or data format and take the necessary actions for correction.

Reviewing the indicators (optional). If necessary, this step can be performed after the new data loads in the evaluated tables. It comprises the reassessment and implementation of new indicators according to needs and demands that may arise.

The following sections present the application of the proposed methodology for data quality analysis using GE in real data from public bids and expenditures.

5 Application in Real Data of Public Bids

This section presents the application of a data quality tool in a big data environment with real data from public bids. As discussed in Section 3.2, we choose the Great Expectations (GE) tool because it is more appropriate to our context. Furthermore, we use the data quality methodology proposed in Section 4 to choose and generate the most suitable quality indicators for the data. Thus, this section is organized as follows: first, we describe the public bidding data on which the quality indicators are applied (Section 5.1). Then, we discuss

the main results generated by such indicators (Section 5.2).

5.1 Data Description and Choice of Indicators

We consider data from public biddings in the Brazilian state of Minas Gerais (both at the state and city levels). The municipal bids come from the portal of the Computerized System of Accounts of the Municipalities (SICOM) of the Court of Auditors of the State of Minas Gerais¹², and the state bids come from the Minas Gerais Government Transparency Portal¹³. Thus, the final dataset contains information on 378,137 bids comprising 12,522,661 bid items and 103,858 bidders (individuals or legal entities) from 2014 to 2021. The data are stored in a big data environment using the Apache Hive data warehouse¹⁴ version 2.0.0. This version of Hive does not support checking data integrity constraints. However, we use this version to show that GE manages to mitigate the lack of such restrictions by detecting inconsistent records.

Regarding the quality indicators of Great Expectations for this data source, we use native and custom expectations (according to Section 3.3). The first group comprises standard and generic rules implemented internally in the tool, such as validating the data domain and checking whether the data follows a regular expression or a range of values. Table 3 presents the list of native indicators used in the analysis of public bids data performed in this section.¹⁵

Custom indicators aim to validate a specific business rule in the data domain. For example, when manually analyzing the bid values, we observe very inconsistent numbers: a single bid had a value almost 200 times greater than the entire revenue of its city that year. Possible causes of this anomaly are a typing error by whoever entered this data in the source or a failure to extract these values from the bidding process documents. Thus, we implemented a custom indicator that compares the value of the bidding with the total revenue of the city or state in the bidding year. Overall, we implemented five custom indicators, described in Table 3. Such indicators were implemented following the naming and organization standards of native indicators.

¹²<https://portalsicom1.tce.mg.gov.br/>

¹³<https://www.transparencia.mg.gov.br/compras-e-patrimonio/compras-e-contratos>

¹⁴<https://hive.apache.org/>

¹⁵List of native expectations of GE: <https://greatexpectations.io/expectations>

Table 4. Overall statistics from GE indicators for bidding data.

| Table | Successes | Failures | Total |
|-------------------|-------------|-------------|-------|
| Bids | 88 (68.22%) | 41 (31.78%) | 129 |
| Qualified bidders | 26 (74.29%) | 9 (25.71%) | 35 |
| Winner bidders | 37 (50.00%) | 37 (50.00%) | 74 |
| Bidding items | 60 (73.17%) | 22 (26.83%) | 82 |
| Commissions | 47 (83.93%) | 9 (16.07%) | 56 |

Table 5. Number of failures captured by quality indicators for public bidding data.

| Error / Table | B | QB | WB | BI | BC |
|------------------------|-----------|----------|-----------|-----------|----------|
| Null values | 18 | 1 | 15 | 6 | 0 |
| Out of range values | 12 | 2 | 9 | 4 | 1 |
| Inconsistent data type | 8 | 2 | 3 | 8 | 7 |
| Duplicated values | 0 | 1 | 3 | 1 | 0 |
| Other | 3 | 3 | 7 | 3 | 1 |
| Total | 41 | 9 | 37 | 22 | 9 |

B: Bids **QB:** Qualified Bidders **WB:** Winner Bidders
BI: Bidding Items **BC:** Bidding Commission

5.2 Data Quality Analysis

In this section, we present the main results of the quality indicators of Great Expectations (GE) on real public bidding data. In this analysis, we consider the five main tables that gather bidding data: (i) general bidding information; (ii) bidders qualified to participate in bidding processes; (iii) bidders approved as winners in bids; (iv) bidding items; and (v) bidding commissions (committees established to act in bids).

For each table, we choose specific native indicators which make sense in the context of the table. In addition, we use our implemented custom indicators according to pre-defined business rules. Table 4 presents the number of successes and failures in the indicators for each analyzed table, as well as the total number of implemented indicators.

Table 5 presents the most common errors detected in the analyzed tables. One of the most frequent errors is the presence of null values in columns where they are not allowed according to business rules. For example, it is not expected that fields containing the year of the bidding exercise have null values. Other common errors include non-standard values and/or out-of-the-expected range and inconsistent data type. In addition, some tables have duplicated records, an error that can occur for two reasons: (i) data loading problems and (ii) the data warehouse used does not support integrity restrictions to avoid this duplicity. However, as GE detects these duplicate records, it is possible to mitigate this data warehouse limitation.

In addition, GE's custom indicators allow checking more complex business rules that native indicators cannot check. Table 6 presents a part of the result of the *expect_values_less_revenue* indicator in the table with bidding information. Using such an indicator, we can verify that three bids have discrepant values compared to the city's total revenue in that year (also obtained from the database). For example, bid A has a value more than 4,000 times higher than its city revenue. However, this is not the value in the price survey in the bidding notice, indicating a probable error in the data extraction and/or loading process.

Another business rule verified by a custom indicator is the chronological order of date fields in the bidding records, as

Table 6. Custom indicator that verifies bids whose value is greater than the city revenue in that year.

| Bid | Year | Entity name | Bid value | City revenue |
|-----|------|-------------|-----------------------|--------------------|
| A | 2014 | City X | R\$ 59,415,748,800.00 | R\$ 11,912,844.54 |
| B | 2015 | City Y | R\$ 16,880,000.00 | R\$ 13,124,280.52 |
| C | 2020 | City Z | R\$ 262,029,682.50 | R\$ 240,799,958.79 |

Table 7. Custom indicator for records that disrespect the chronological order of bidding dates (Date 1 \leq Date 2).

| Date 1 | Date 2 | Records | % |
|----------------------------|---|---------|------|
| Public notice date | Publication date of notice | 7,672 | 2.03 |
| Public notice date | Date of publication on the vehicle | 8,728 | 2.31 |
| Publication date of notice | Expected date of receipt of documentation | 2,186 | 0.58 |

the dates must respect the order of the bidding process. For example, the date of the bidding notice must be before its publication, as the preparation of the notice is the first stage of the process, and the receipt of the documentation only happens once it is published. Table 7 presents the number of cases that do not respect this order in the bidding table. Analyzing the number of records in this situation, it is possible that there was a problem with the imputation or loading of the data. This result reinforces the need for a thorough analysis of the data extraction, processing, and loading processes.

6 Application in Real Data of Public Expenditure

This section presents a second application of Great Expectations (GE) as a quality tool in government data. Here, we apply GE to five tables referring to purchases and public expenditures carried out by cities in the Brazilian state of Minas Gerais. As in the previous section, we apply the data quality methodology proposed in Section 4. Thus, in this section, we first present the description of the data and the choice of indicators in Section 6.1, and then we present and analyze the results of the indicators in Section 6.2.

6.1 Data Description and Choice of Indicators

For this application, we use public expenditure data from cities and the state of Minas Gerais that are not necessarily linked to public bids. According to Brazilian law No. 14,133 of 2021¹⁶, some purchases can be made without needing to start a bidding process for specific reasons. Thus, we consider five new tables with information on revenues obtained by federal entities, receipts issued (and their items), and signed contracts (and their items). As in Section 5.1, data also comes from the Computerized System of Accounts of the Municipalities (SICOM) of the Court of Auditors of the State of Minas Gerais and the Transparency Portal of the Government of Minas Gerais and is stored in a big data environment using Apache Hive. All tables contain information from 2014 to 2021, except for the revenue table, which has records from 2002.

¹⁶Law No. 14,133: https://www.planalto.gov.br/ccivil_03/_Ato2019-2022/2021/Lei/L14133.htm

Table 8. Great Expectations indicators used in analyzing public expenditure data.

| | Indicator (expectation) | Description |
|--------|--|---|
| Native | expect_column_values_to_not_be_null | Column values must not be null |
| | expect_column_values_to_be_unique | There must be no duplicate values in the column |
| | expect_column_min_to_be_between | The smallest column value must be within the range [min, max] |
| | expect_column_values_to_be_in_type_list | Column values must be of the specified type |
| | expect_column_values_to_be_in_set | Column values must belong to a value set |
| | expect_column_values_to_be_between | Column values must be in the range [min, max] |
| | expect_column_values_to_match_regex | Column values must follow a given regular expression |
| | expect_column_pair_values_a_to_be_greater_than_b | Values in column <i>a</i> must be greater than column <i>b</i> (pairwise) |
| Cus. | expect_value_less_revenue | Bids must have a value less than or equal to the entity revenue |
| | expect_dates_to_match_across_tables | Reported dates must be in valid chronological order |

Table 9. Overall statistics from GE indicators for public expenditure data.

| Table | Successes | Failures | Total |
|---------------|--------------|-------------|-------|
| Revenue | 69 (65.09%) | 37 (34.91%) | 106 |
| Receipt | 101 (78.29%) | 28 (21.71%) | 129 |
| Receipt item | 46 (20.20%) | 5 (9.80%) | 51 |
| Contract | 68 (83.95%) | 13 (16.05%) | 81 |
| Contract item | 12 (85.71%) | 2 (14.29%) | 14 |

Table 10. Number of failures captured by quality indicators for public expenditure data.

| Error / Table | RV | RC | RCI | C | CI |
|------------------------|-----------|-----------|----------|-----------|----------|
| Null values | 32 | 17 | 0 | 2 | 1 |
| Out of range values | 4 | 3 | 2 | 7 | 0 |
| Inconsistent data type | 0 | 5 | 3 | 2 | 0 |
| Duplicated values | 0 | 0 | 0 | 0 | 0 |
| Other | 1 | 3 | 0 | 2 | 1 |
| Total | 37 | 28 | 5 | 13 | 2 |

RV: Revenue RC: Receipt RCI: Receipt Item
C: Contract CI: Contract Item

Table 8 presents the indicators used in the tables considered in this application. We use both native and custom indicators that allow us to verify specific business rules in this context. In general, the native indicators are the same as used in the first application (see Section 5) since the source and structure of the data are the same. What is new is the indicator *expect column pair values a to be greater than b*, which we use to compare whether the values of a column are greater than another. For example, in the receipt table, we check if the gross amount is greater than or equal to the net amount. In addition, we use two custom indicators to check value fields (*expect value less revenue*) and the chronological order of dates (*expect dates to match across tables*).

6.2 Data Quality Analysis

In this section, we present and discuss the results of the quality indicators for the public expenditure tables considered in this application. As we described in the previous section, this application considers five distinct tables: (i) revenue from the municipalities and the State of Minas Gerais; (ii) invoices used for public purchases; (iii) items present in the invoices; (iv) contracts entered into by municipalities and the State; and (v) items present in such contracts.

Tables 9 and 10 present the general statistics and most common errors found in each table. Considering absolute and proportional values, the table with the highest number

Table 11. Custom indicator that verifies contract items whose value is greater than the city revenue in that year.

| Item | Year | Entity name | Item value | City revenue |
|------|------|-------------|--------------------|-------------------|
| X | 2017 | City A | R\$ 65,000,000.00 | R\$ 33,921,834.39 |
| Y | 2020 | City B | R\$ 954,750,000.00 | R\$ 30,853,996.41 |
| Z | 2021 | City C | R\$ 2,200.00 | R\$ 632.78 |

Table 12. Custom indicator for records that disrespect the chronological order of contract dates (Date 1 \leq Date 2).

| Date 1 | Date 2 | Records | % |
|---------------------|-------------------|---------|------|
| Signature date | Validity end date | 73,096 | 7.41 |
| Validity start date | Validity end date | 77,157 | 7.82 |
| Publication date | Validity end date | 43,856 | 4.45 |

of errors is the revenue one, in which 37 out of 106 expectations (34.91%) failed. When examining these failures in more depth, we note that the vast majority of them are related to the presence of null values in columns where they should not exist. Examples include fields that indicate the revenue source and its type. In addition, some records have a negative collected amount, which is out of the accepted range according to the pre-defined business rules. Both types of errors may have occurred due to a failure in the data extraction or loading process, which requires a detailed manual revision of this process by analysts.

The customized indicators also raise warning signals about data quality in the analyzed tables. For example, Table 11 presents a part of the results of the indicator that verifies the business rule in which the values of the contract items must be lower than the entity's revenue (city or state). We verify that the value of item Y contracted by City B is more than 30 times greater than the city's revenue. In this case, it is possible that there was an error in extracting or imputing the value of the contracted item in the database, requiring further analysis. On the other hand, item Z acquired by City C has a feasible value, but the sum of the city's revenues in that year registered in the database results in a very low value. Such a value is impossible and does not match the city's GDP, indicating a possible lack of revenue records in the database. It is important to note that such errors would not be noticed in a quick analysis without the Great Expectations indicators, reinforcing this tool's importance.

Finally, the indicator that verifies the chronological order of dates also brings relevant results for this application. Table 12 shows the results of this indicator for the contracts table. We observe that a small part of the records presents inconsistencies between the dates in the table. For example, 7.82%

Table 13. Overall statistics from GE indicators for public bidding and expenditure data. Tables are sorted by Table Error Score (TES).

| Application | Table | Records | Indicators | Failures | PF(t) | TES(t) |
|--------------------|-------------------|------------|------------|-------------|-------|--------|
| Public bids | Winner bidders | 12,298,683 | 74 | 37 (50.00%) | 0.666 | 4.724 |
| Public expenditure | Revenue | 1,155,372 | 106 | 37 (34.91%) | 0.510 | 3.092 |
| Public bids | Bidding items | 12,522,661 | 82 | 22 (26.83%) | 0.404 | 2.866 |
| Public expenditure | Receipt | 21,239,828 | 129 | 28 (21.71%) | 0.300 | 2.199 |
| Public bids | Qualified bidders | 833,777 | 35 | 9 (25.71%) | 0.354 | 2.093 |
| Public bids | Commissions | 1,505,358 | 56 | 9 (16.07%) | 0.332 | 2.052 |
| Public expenditure | Contract item | 6,218,470 | 14 | 2 (14.29%) | 0.264 | 1.793 |
| Public bids | Bids | 378,137 | 129 | 41 (31.78%) | 0.318 | 1.773 |
| Public expenditure | Contract | 986,524 | 81 | 13 (16.05%) | 0.229 | 1.373 |
| Public expenditure | Receipt item | 4,098,967 | 51 | 5 (9.80%) | 0.163 | 1.076 |

of the records have a validity start date later than a validity end date, which is impossible according to business rules. Likewise, some records have a signature date and publication date after the end of the contract validity. Again, a detailed analysis of the records is essential to detect the source of inconsistencies and correct this quality problem since such inconsistencies can impact other final applications, including trails for detecting fraud in public expenditure.

7 Data Quality Score Based on Expectations

In this section, we further analyze the data quality by presenting an error metric for each table using the Great Expectations (GE) results. Such a deeper analysis is necessary because simply aggregating the success rates of indicators by a simple average may not fully represent reality since the table size also influence its quality. In this work, validation failures mean GE indicators that returned an error, regardless of the number of records impacted. That is, n validation failures do not mean n records with a problem but n indicators that failed. In addition, the greater the number of validation failures, the greater the error score for the table in question should be since each failure requires a specific action by the analysts to correct it. For example, a table with 1,000,000 records and two validation failures of the GE indicators should have a higher error score than a table with 1,000 records and the same two failures.

To further analyze the data quality, we first calculate a penalty factor $PF(t)$ for each table t in our dataset (Equation 1). Each table has a set of N indicators implemented, of which n indicators fail. Furthermore, each indicator i presents a percentage of unexpected values found $e_i \in [0, 1]$ (for successful indicators, $e_i = 0$). Thus, PF is the product of each indicator's average percentage of validation failure by the proportion of failed indicators. Both factors are increased by one so that the penalty increases according to the number of errors. After the multiplication, the value is subtracted from one so that $PF(t)$ is zero when no indicator fails. The values of $PF(t)$ start from zero (when no indicators fail) and go up to 3 (when all of them fail with 100% of unexpected values). Thus, the greater the number of errors, the greater the score for a given table.

$$PF(t) = \left(1 + \frac{\sum_{i=0}^{N_t} e_i}{N_t} \right) \left(1 + \frac{n_t}{N_t} \right) - 1 \quad (1)$$

Therefore, we propose the Table Error Score (TES) to quantify the quality of each table in our application and to compare the tables more fairly. For each table t , it is calculated as the product of the penalty factor $PF(t)$ and the order of magnitude of the given table, represented by the logarithm of its number of records (Equation 2). Hence, the quality score of the table also considers its size since tables with millions of records must present a higher alert level than smaller tables with the same level of indicator failures.

$$TES(t) = \log_{10}(r_t) \cdot PF(t) \quad (2)$$

As an example of the application of TES , consider table A with 100 records and three implemented indicators, of which two fail with percentages of unexpected values 90% and 1%, respectively. The PF value for this table is $PF(A) = \left(1 + \frac{0+0.9+0+0.1}{3} \right) \left(1 + \frac{2}{3} \right) - 1 = 1.172$. Then, the TES value for such a table is $TES(A) = \log_{10}(100) \cdot 1.172 = 2.344$. Moreover, table B with 1000 records and the same indicators and faults, has the same PF value (1.172), but its TES value is $TES(B) = \log_{10}(1000) \cdot 1.172 = 3.516$. Such values follow the premise mentioned above since tables with more records represent a higher alert for quality analysts.

Table 13 presents general statistics for each table of both applications, including the number of failures in the indicators, as well as their Table Error Score (TES). The importance of calculating the TES is evident based on such results. For example, the winner bidders table has the highest TES value, but it is not the table with the highest absolute number of indicator failures. When considering such a metric, the first table in the ranking would be the bid table (41 failures). However, in our analysis, it makes more sense for the winner bidders table to have the highest error score since it is among the tables with the highest order of magnitude of size ($\log_{10}(r_i) \approx 7$). Thus, such a table would be a great candidate to start a manual analysis to correct quality problems.

Overall, the TES metric offers a great benefit for quality analysis as it is a metric that allows comparing the data quality of tables from different applications. This metric goes beyond the simple percentage of failures offered by Great Expectations because it also considers the number of unexpected values for each indicator and the total number of

records in the table. Thus, the metric is in accordance with the premise that tables with more records and indicators with a higher percentage of failure represent a higher level of alert, allowing people responsible for quality control of the data to determine the order of the tables to be examined.

8 Conclusion

This article presents a comparative analysis of eight open-source data quality assessment tools and the results of two applications in a big data environment with real data from governmental data using the Great Expectations (GE) tool. This tool was chosen because it has graphical interface components that help in the visualization of results and because it allows the creation of customized indicators that allow verifying complex business rules not verified by native indicators. In other words, such indicators allow the implementation of specific validations for the data analysis. It is worth noting that GE assists in identifying records that have problems caused by the impossibility of implementing data integrity restrictions by the data warehouse in question.

The results of GE's indicators in applications to real data from public bids and expenditures bring up quality problems that would not be easily identified, including inconsistency in the values in the tables and the chronological order of dates. Furthermore, we propose a new quality metric to allow the comparison of the GE results in tables from different applications. This metric goes beyond the percentage of failed indicators and considers the number of records in the tables and the percentage of unexpected values for each indicator. Its use can help quality analysts determine the priority of the tables for manual inspection. Thus, analyzing the quality of governmental data is a necessary step to ensure the reliability of records, which directly impacts the construction of future applications that use this data (e.g., fraud detection and analysis of overpricing).

As future work, we plan to expand the usage of Great Expectations (GE) for other tables with governmental data in both public bids and expenditure applications. We also intend to analyze the quality of other data domains, including expenditure data on electoral processes. Since the best quality tool depends on usage dynamics and data context, such analyses may require applying other data quality tools.

Acknowledgements

The authors thank this work's collaborators, Arthur P. G. Reis, Gabriel L. Canguçu, and Victor Caetano.

Funding

This work was funded by the Prosecution Service of State of Minas Gerais (in Portuguese, *Ministério Público do Estado de Minas Gerais*, or simply MPMG) through the Analytical Capabilities Project (in Portuguese, *Programa de Capacidades Analíticas*) and by CNPq, CAPES, and FAPEMIG.

Competing interests

The authors declare that they have no competing interests.

References

- Altendeitering, M. and Tomczyk, M. (2022). A functional taxonomy of data quality tools: Insights from science and practice. In *Wirtschaftsinformatik*.
- Ballou, D. P. and Pazer, H. L. (1985). Modeling data and process quality in multi-input, multi-output information systems. *Management Science*, 31(2):150–162.
- Chrisman, N. R. (1983). The role of quality information in the long-term functioning of a geographic information system. In *Auto-Carto*, pages 303–312.
- Cichy, C. and Rass, S. (2019). An overview of data quality frameworks. *IEEE Access*, 7:24634–24648.
- Ehrlinger, L. and Wöß, W. (2018). A novel data quality metric for minimality. *QUAT*, 1:1 – 15. DOI: 10.1007/978-3-030-19143-6_1.
- Ehrlinger, L. and Wöß, W. (2022). A survey of data quality measurement and monitoring tools. *Front. Big Data*, 5. DOI: 10.3389/fdata.2022.850611.
- Elmagarmid, A. K., Ipeirotis, P. G., and Verykios, V. S. (2007). Duplicate record detection: A survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16. DOI: 10.1109/TKDE.2007.250581.
- Etcheverry, L. and Consens, M. P. (2011). Summary-based comparison of data quality across public MAGE-ML genomic datasets. *J. Inf. Data Manag.*, 2(1):3–10.
- Foidl, H., Felderer, M., and Ramler, R. (2022). Data smells: Categories, causes and consequences, and detection of suspicious data in ai-based systems. In *arXiv*. DOI: 10.48550/ARXIV.2203.10384.
- Gao, J. Z., Xie, C., and Tao, C. (2016). Big data validation and quality assurance - issues, challenges, and needs. In *SOSE*, pages 433–441. IEEE Computer Society. DOI: 10.1109/SOSE.2016.63.
- Goudar, S. S., Stolka, K. B., Koso-Thomas, M., Honnunar, N. V., Mastiholi, S. C., Ramadurg, U. Y., Dhaded, S. M., Pasha, O., Patel, A., Esamai, F., et al. (2015). Data quality monitoring and performance metrics of a prospective, population-based observational study of maternal and newborn health in low resource settings. *Reproductive Health*, 12(2):1–10. DOI: 10.1186/1742-4755-12-S2-S2.
- Josko, J. M. B. and Ferreira, J. E. (2021). Using visual-interactive properties to support data quality visual assessment on abstract and timeless data. *J. Inf. Data Manag.*, 12(2).
- Junior, C. S. and Dorneles, C. F. (2021). Avaliação de dimensões de qualidade de dados para o agronegócio. In *SBBD*, pages 283–288. SBC.
- Laranjeiro, N., Soydemir, S. N., and Bernardino, J. (2015). A survey on data quality: Classifying poor data. *PRDC*, pages 179 – 188. DOI: 10.1109/PRDC.2015.41.
- Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. (2002). Aimq: a methodology for information quality assessment. *Information & Management*, 40(2):133 – 146.

- Maia, P., Meira Jr., W., Cerqueira, B., and Cruz, G. (2020). Auditing government purchases with a multicriteria anomaly detection strategy. *J. Inf. Data Manag.*, 11(1).
- Medeiros, G. F. d., Degrossi, L. C., and Holanda, M. (2020). Qualiosm: Melhorando a qualidade dos dados na ferramenta de mapeamento colaborativo openstreetmap. In *SBBD*, pages 77–82. SBC.
- Oliveira, G. P., Reis, A. P. G., Freitas, F. A. N., Costa, L. L., Silva, M. O., Brum, P. P. V., Oliveira, S. E. L., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022a). Detecting inconsistencies in public bids: An automated and data-based approach. In *Proceedings of the 28th Brazilian Symposium on Multimedia and Web*, pages 182–190, New York, NY, USA. ACM. DOI: 10.1145/3539637.3558230.
- Oliveira, G. P., Reis, A. P. G., Mendes, B. M. A., Bacha, C. A., Costa, L. L., Canguçu, G. L., Silva, M. O., Caetano, V., Brandão, M. A., Lacerda, A., and Pappa, G. L. (2022b). Ferramentas open-source de qualidade de dados para licitações públicas: Uma análise comparativa. In *Proceedings of the 37th Brazilian Symposium on Databases*, pages 116–127, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2022.224351.
- Pipino, L. L., Lee, Y. W., and Wang, R. Y. (2002). Data quality assessment. *Commun. ACM*, 45(4):211 – 218. DOI: 10.1145/505248.506010.
- Pushkarev, V., Neumann, H., Varol, C., and Talburt, J. R. (2010). An overview of open source data quality tools. In *IKE*, pages 370–376. CSREA Press.
- Scannapieco, M. and Catarci, T. (2002). Data quality under a computer science perspective. *Journal of The ACM - JACM*, 2:1–12.
- Sessions, V. and Valtorta, M. (2006). The effects of data quality on machine learning algorithms. In *ICIQ*, pages 485–498. MIT.
- Wang, R. Y., Strong, D. M., and Guarascio, L. M. (2018). Beyond accuracy: What data quality means to data consumers. 1996. *Total Data Quality Management Programme*.
- Wu, D., Xu, H., Wang, Y., and Zhu, H. (2022). Quality of government health data in COVID-19: definition and testing of an open government health data quality evaluation framework. *Libr. Hi Tech*, 40(2):516–534. DOI: 10.1108/LHT-04-2021-0126.
- Zöllner, F. G., Daab, M., Sourbron, S. P., Schad, L. R., Schoenberg, S. O., and Weisser, G. (2016). An open source software for analysis of dynamic contrast enhanced magnetic resonance images: Ummperfusion revisited. *BMC Med Imaging*, 16(7):1–13. DOI: 10.1186/s12880-016-0109-0.