Operationalizing the Data Quality Framework

Sezal Chug[1], Priya Kaushal[1], Dr. Ponnurangam Kumaraguru[1], Dr. Tavpritesh Sethi[1]

1 Department of Computer Science, Indraprastha Institute of Information Technology, New Delhi, Delhi, India, 110020

¤Current Address: Department of Computer Science, Indraprastha Institute of Information Technology, New Delhi, Delhi, India, 110020

* sezal17101@iiitd.ac.in
* priya17081@iiitd.ac.in
* pk@iiitd.ac.in
* tavpriteshsethi@iiitd.ac.in

Abstract

Background

Data is expanding at an unimaginable rate, and with this development comes the responsibility of the quality of data. Data Quality refers to the relevance of the information present and helps in various operations like decision making and planning in a particular organization. Mostly data quality is measured on an ad-hoc basis, and hence none of the developed concepts provided any practical application.

Methods and Findings

The current empirical study was undertaken to formulate a concrete automated data quality platform to assess the quality of incoming dataset and generate a quality label, score and comprehensive report. The proposed system quantifies and qualifies the provided data and evaluates them at subjective and objective levels. We utilize various datasets from healthdata.gov, opendata.nhs and Demographics and Health Surveys (DHS) Program to observe the variations in the quality score and formulate a label using Principal Component Analysis. The results of the current empirical study revealed a metric that encompasses nine quality 'ingredients', namely provenance, dataset characteristics, uniformity, metadata coupling, percentage of missing cells and duplicate rows, skewness of data, the ratio of categorical columns, and correlation between these columns.

Conclusion

Judging various datasets based on this metric proved that due to the growing technology upgradations in data collection and processing, there is a constant gradient increase in the

quality of data. The study also validates the metric by using modified Mutation Testing approaches to show that the metric completely captures the essence of any incoming dataset. The value of the quality label instill confidence in the user in deploying the data for his/her respective practical application.

1 Introduction

With advancements in technology, use of data has become immensely influential and hence the importance of its quality. Several sources predict exponential data growth by 2022 and anticipate that data would become an integral part of our lives. While human-generated data is experiencing an exponential growth rate, machine data is increasing even more rapidly [1]. This ever-increasing speed of data growth and its exploding volume has introduced several challenges encompassing high operational costs for irrelevant data at data processing stations for storage and processing [2]. Setting up an empirical metric to evaluate data quality leads to increased profitability, which helps make better decisions for organizations by improving data usability and reducing storage and processing costs.

Data quality refers to how relevant the information is for use in a particular application. Low data quality has been curbing the growth of various organizations by preventing them from performing to their full potential [3]. Analyzing the data quality levels can help organizations identify the pitfalls that need to be resolved to enhance their clarity. Furthermore, inaccurate data can be identified and fixed to ensure that executives, data analysts, and other end users work with accurate and efficient information. Some other essential elements of good data quality include **Completeness, Consistency, Concordance, and Conformity** to the standard data formats created by a particular organization [4]. Meeting all of these factors is necessary to ensure that data sets are reliable, trustworthy, and suitable for use.

With the advent of the machine learning/artificial intelligence realm, quality of data has mainly been neglected under the assumption that the data feeding these algorithms is of high quality. There is always more focus on learning algorithms and models instead of ensuring data quality. A study conducted by Wand et. al. [5] stated that the actual use of the data is outside of a researcher's control, however, it is essential to provide data conforming to a particular level to ensure its proper usage. Poor data quality can have a severe impact on the overall effectiveness of data in an organization. The "new normal" of massive generation, utilization and elimination of data has urged researchers to consider its data quality aspect. It is difficult to define quality for data because unlike manufactured products, data do not have physical characteristics that allow quality to be easily assessed [6]. The actual use of the data is outside of the designer's control, hence a design-oriented definition of data quality is necessary which will reflect the intended use of the information [5]. Currently, most of the data quality measures are being developed on a need basis to solve a specific problem, but creating a fundamental principle or a metric to measure data quality is lacking [7]. Hence, it is essential to conduct research on how to operationalize these formally derived dimensions of data quality that would quantitatively measure a quality score. As famously said by Willcocks and Lester [8], "what gets measured gets managed", hence quality label measuring the dimensions of quality of data signify a crucial management element in the domain of data quality.

Pipino et al. [7] proposed a new data management paradigm to help unify the diverse efforts using a flexible schema that pursued data integration and unification. This article

defines certain factors which can be used on any given dataset to measure its quality. They include traditional data quality metrics, such as **free-of-error, completeness, consistency, concise representation, relevancy, and ease of manipulation.**

Starting with a pretty high-level description of the components of data quality by Veregin [6], researchers have gone into intricate aspects of datasets. Wand et. al. [5] in 1996 introduced us to data quality dimensions in ontological foundations, which have now become advanced and more detailed. Jayawardene et al. [9] in 2015 provided comprehensive classifications of data quality dimensions, which helped develop a streamlined and unified set of quality dimensions. They provided a basis for modelling data quality which has been expanded in the current empirical study. Initially, researchers gave an idea about quality dimensions which encompass the **accuracy, precision, consistency, and completeness** of an incoming dataset. Further, a few researchers elaborated on these principles and established measurable quantities of these DQ dimensions, which help quantify data quality scores. **Reliability, Null Values, Size of data, Correctness, Accuracy, Conformance and Duplication**, were a few core dimensions central to the practical analysis of formal data quality requirements. However, the growing number and evolution of data quality dimensions and the emergence of new classifications and definitions point to a lack of shared understanding amongst various organizations for a widely accepted practice of calculating data quality scores.

- Provenance provides useful information on the history of a dataset, such as when it was last updated, the source data and the authority that certifies the dataset, if any [10].
- Uniformity refers to whether instances of data are either stored, exchanged, or presented in a format that is consistent with the domain of values and consistent with other similar attribute values [9].
- Accuracy is the first and foremost requirement that many users expect from data. The accuracy of an entire database can be measured by finding the fraction of incorrect tuples in the database [11].
- Missing values refer to a null value as a missing value [9].
- Duplication of rows has been defined as a measure of unwanted duplication existing within or across systems for a particular field, record, or data set [12].

Currently, most data quality measures provide an ad hoc basis to solve specific problems, as suggested by Huang et al. [13] and Laudon [14]. They insinuate fundamental principles which are necessary for developing usable metrics but are lacking in practice. They concluded that assessments of data could be both objective and subjective. Several factors or processes generated bad data; which include human data entry, sensor devices readings, social media, unstructured data, and missing values [15].

In another research by Sun et al. [16], metadata matching with the purpose of information integration was considered to be an essential aspect of the assessment of data quality. Data quality documentation plays a key role in many standards due to the realisation that an understanding of quality is essential to the effective use. Pantulkar and Srinivas [17] insisted that semantic similarity has a vital role in natural language processing and application. He proposed three different semantic similarity approaches in their research, i.e. cosine similarity, path-based approach and feature-based approach. The feature-based approach to perform the tagging and lemmatization helped calculate the data quality score.

Research needs to be conducted on how to operationalize these formally derived dimensions in terms usable in systems design practice. These data quality dimensions can be used to develop data quality audit guidelines and procedures and metrics can be developed for use in specification and audit of information systems. Hence, these Data quality dimensions identified can be used to guide data collection in field studies of data quality problems and practices [9]. A comprehensive description of data quality dimensions is instrumental in the pursuit of developing a streamlined automated platform that provides a basis for quality modelling. To bridge this gap of quality, this paper undertakes a study by defining quality indicators called 'data quality ingredients' that incorporates a semantic perspective on data quality [9]. On the similar lines of Holland et al. [18], the proposed method utilizes existing data quality dimensions, metadata concordance techniques and set Nutrition Label Approach for quality judgment.

The ibid empirical study aims to formalize these data quality dimensions and suggests a Nutrition Label approach towards building a quality label that captures the data quality of any incoming dataset and evaluates it on these quality ingredients to generate a DQ score. This widely accepted quality metric would quantify data and measure the degree to which it fits our purpose. It presents a comprehensive report which gives an overview of the "ingredients" of the dataset and suggests ways and means to improve this quality score.

Many previous works of literature have given a brief idea about these dimensions that define the quality of data; however, none of them partakes the importance of metadata into their quality judgement. The current empirical study not only evaluates the concordance of metadata with respect to column descriptions but also takes into account the dataset characteristic values for each column. These studies didn't include aspects of datasets like metadata matching, correlation, skewness and categorical nature between these columns variables.

After a thorough analysis of the previous literature on Data Quality and considerate efforts and discussions, the current research came to conclude the "ingredients" of a dataset that will further adjudge data quality. These nine quality "ingredients" are **provenance, dataset characteristics, uniformity, metadata coupling, percentage of missing cells and duplicate rows, skewness of data, the fraction of categorical columns, and correlation between all columns**. These "ingredients" would be guiding factors that help calculate the final score that would help achieve better data quality.

Our model would observe a new incoming dataset and evaluate the "ingredients" of this dataset. Further, using the proposed metric, the system would formulate a quality label, DQ score and generate a comprehensive report that analyzes the characteristics of the incoming datasets and helps users weigh the data against other possibilities. In the further sections of this study we will describe these dimensions of data quality in detail, elaborating upon the platform that generates the quality label, DQ score and the comprehensive report. It would also include the validation and application of the proposed model which helps instill faith towards our approach.

2 Dataset retrieval

The Demographics and Health Surveys (DHS) Program dataset of India contains restricted survey data files for legitimate academic research. We collected healthcare

survey data from the DHS website for countries Myanmar, Ethiopia, Zimbabwe, Maldives, Nepal, Nigeria, Afghanistan, Bangladesh and Cambodia for the year 2015-16 for training purposes. Along with these countries, healthcare datasets from HealthData.gov(https://healthdata.gov/) and Scottish Health and Social Care Open Data platform(opendata.nhs) were also combined to finalize 200 Training datasets. DHS Indian datasets over the years 1992-93, 1998-99, 2005-06 and 2015-16 were collected for testing the formulated metric. Data quality assessment of the collected data is conducted by dividing the dataset into sections based on the DHS Recode Manual.

The above approach helped the model understand the incoming dataset based on various parameters and individually analyze each section. The numerous takeaways of the proposed model are discussed in further sections, which suggest a data quality metric to adjudge the quality of data and provide empirical validation to the proposed metric. The formulated in the attached link [19].

3 Methodology

The research illustrates the formulation of quality label "ingredients" and data quality metric using Training Datasets. In the current study, we calculate the value of the ingredients of our model for the training dataset to formulate a sheet containing 'ingredient' scores for all 200 training datasets. We further use two techniques to propose a metric and then compare trends retrieved from the said two metrics.

After considerate efforts on reading previous works on data quality, the study finalized **nine 'ingredients'**, namely provenance, dataset characteristics, uniformity, metadata coupling, percentage of non-missing cells and non-duplicate rows, skewness of data, number of continuous and categorical columns and the correlation between columns of a dataset. Every incoming dataset is judged on these nine parameters to generate a nutrition label, data quality score and comprehensive report detailing these factors.

3.1 Data Quality Ingredients

1. Reliability is a measure of between the expectations and capability of the dataset. It indicates whether is the information can be counted on to convey the right information to a researcher [5]. **Provenance** refers to the record trail, which accounts for the origin of the dataset and details about the latest version. It provides baseline information for assessing authenticity, integrity and helps in enabling the trust of the researcher for using the dataset. The provenance can be part of the local repository, present on the website or may be present in a separate file [20]. The parameters used for calculating the provenance of a dataset are origin/source, author, latest updated date and data accessibility. All the parameters are given equal weightage to calculate the provenance percentage of the dataset. The provenance of data helps to answer questions like "who", "how", "where", "when", and "by whom" was data produced. This is also referred to as data lineage, which includes either authentic government sources which are given full score for origin or private sources like kaggle or opendata where this score is decided on the basis of the number of usages of the dataset. The number of years between last updated date and present give us the percentage of weightage for correctness. Data accessibility which refers to the properties related to accessing

Fig 1. Data Quality Ingredients

the data for secondary use, this includes permissions or intellectual property,the format of the data or the file, the kind of pre-processing done to the data before publishing [21].

2. After the data is imported from the original path, preliminary checks are conducted, and the values are cross-checked with the data provided on the author's website. Accuracy is the first and foremost requirement that many users expect from data. Hence, data is correct if it conveys the same meaning lexically, syntactically and semantically [9]. The **Dataset Characteristics** ingredient includes the comparison of mean, median, mode, standard deviation, range of values (Min-Max values) and the total number of observations in a dataset. These values are then matched with the values scrapped of the source website to generate a percentage of correctness in these characteristic quantifications [10]. All parameters are given equal weightage to calculate the final percentage, which helps judge the quality of data at a discrete level and provide the user with a high level of information about the dataset.

3. **Uniformity** in a datasets highlights instances within a dataset which are consistent with values present on the sources of the dataset [9]. It is calculated by verifying the data type of each column to its values. To measure uniformity, we count the number of cells where the data type does not match the column and divide the result by the total number of cells in the dataset. We take the mean of these incorrect matchings over the number of columns in a dataset to generate the final percentage of uniformity.

4. The current empirical study includes a unique highlighted novel approach for including metadata information along with the dataset. Metadata refers to structured information provided along with the dataset describing its columns and respective values. It helps the user gain more insight into the dataset and understand the relationship between data and columns. **Metadata coupling** checks the metadata matching score of data by comparing the column name to the column description provided by the metadata codebook for any incoming dataset [22]. Byrne et.al [9]

described that "Adherence to metadata standards is an aspect of Data Quality". Metadata should comply with the dataset columns and clearly define its purpose. We approached this problem with a classical view of natural language processing and combined character based, token based, feature based and phonetic based similarity algorithms. Firstly, we convert all letters to lower case to avoid any case sensitive inconsistencies. Further, we remove the special characters, integers and stop words, followed by stemming and lemmatization [23]. The remaining data is converted into feature vectors which ultimately calculate the similarity score.

(a) **Character based similarity** also known as edit distance measure takes two strings and calculates the edit distance that is the minimum number of edits required to transform one string into the other. The character based algorithms included in our algorithm are Hamming distance [24], Levenshtein distance [25], Jaro-winkler distance [26], Needleman wunsch [27], Smith waterman [28] and longest common subsequence [29]. Character based similarity is useful in recognizing typographical errors and existing data inconsistencies [30].

(b) **Token based similarity** models encompass situations where each string in the sentence is a set of tokens and similarity is calculated by manipulation of these tokens [31]. The similarity is greater if there is an overlap of tokens in the two matching sentences. The token based similarity algorithms used in our metric include Jaccard similarity [32], Cosine similarity [33], Manhattan distance [34],Tanimoto similarity [35].

(c) **Feature based similarity** uses set theory operations between features to calculate sentence similarity. We included the Tversky similarity [36] and the overlap algorithm in our model [30] to capture the essence of the similarity with respect to the sets defined by matching algorithms.

(d) **Phonetic based similarity** approach uses variation of sound to recognize misspelled data. In our model, we used the match rating approach to calculate sentence similarity [37].

We combined the above mentioned algorithms to create a hybrid approach to convert the abstract term of metadata matching to a measurable quantity. The process of calculation of the matching score consisted of three steps. Firstly, Data Pre-processing and Vectorisation, which includes conversion into lower case, removal of stop words and symbols, stemming and lemmatization. Further these sentences are converted into feature vectors which aid in the calculation of similarity score of all the algorithms separately. Vectors generated in the first step are fed into the algorithm for similarity calculation. The similarity scores from these string comparison algorithms are normalized to values between zero and one to make them comparable amongst the 13 defined algorithms. After normalization, the score zero represents low similarity, and one represents high similarity. The final step is calculation of metadata matching score. With equal weightage to each algorithm, this score averaged over thirteen helps generate Metadata Coupling percentage allows to adjudge the concordance of the incoming dataset with the descriptive codebook and ultimately provides a measurable compliance score for the same.

5. **Statistics** modelling of a dataset is performed by calculating and studying the percentage of missing cells, duplicate rows, and skewness. The total percentage of non-missing cells and non-duplicate rows are calculated to aid in its quality judgement.

- **Missing Cells**: As stated by Jayawardene et. al. [9] "Data is complete if no piece of information is missing". Large portions of missing cells therefore, reflect poor data quality and render the dataset useless.
- **Duplication of Values**: Sidi et. al. [12] stated that a measure of unwanted duplication within dataset indicates that the respective data is inappropriate for usage among end users as it contains redundant data.
- **Skewness**: Skewness is a measure of the lack of symmetry in a dataset distribution. Any symmetric data should have a skewness nearing to zero, whereas, data that does not include objectivity would be highly skewed. The skewness of the dataset helps the user understand if any bias is present in the data. Highly skewed data would represent unfair statistics and would not yield good results when applying that dataset. A symmetric or unbiased dataset/survey would always have zero skewness, and hence the lower value of skewness would indicate a higher percentage of data quality [38].

6. **Correlations** between columns of a dataset is a way of understanding the relationship between its multiple dimensions or features. We use Pearson's correlation coefficient to generate a percentage of correlation of the dataset wherein the high value would indicate low data quality and vice versa. The system finds the highly correlated columns and presents them to the user in the detailed report. High correlation means a noisy dataset, which can either be helpful in certain situations or harmful in some. Hence, it's up to the user to decide and generate the score for the same. We provide a comprehensive report on interrelationships between variables used to guide the user if these variables need to be removed or kept.

3.2 Metric Formulation

Our system uses the above mentioned "ingredients" on the training dataset to formulate a metric that further carefully analyses and compares data quality trends on the testing dataset by an alternative approach.

Principal Component Analysis or PCA can be used to assign weights to input variables and generate innovative indices. We create data-driven indices by aggregating input variables from our training data by using principal component analysis (PCA) loadings [39]. In this approach, we formulated coefficients of the linear combination of the original variables from which the principal loadings are constructed. These loadings can be both positive and negative; wherein, positive loadings indicate a positive correlation between the variable and the principal component, and negative loadings indicate a negative correlation. Large (either positive or negative) loadings suggest that a variable has a strong effect on that principal component. While formulating the metric from these loadings, we shifted the offset of these values from [-1,1] to [0,2] by adding 1 to all principal component loadings to make all values positive. Furthermore, these loadings are normalized to retrieve the percentage of each "ingredient" over a total of 100 as shown in Figure 2. This approach formulated the metric (figure 5) that calculated data quality scores of the Demographics and Health Surveys (DHS) Program Indian dataset over the years 1992-93, 1998-99, 2005-06 and 2015-16 with the help of values of the proposed ingredients.

Labels	Principal Component Loadings	Positive Prinicpal Component Loadings	Normalization	Percentage
Provenance	0.068	1.068	0.097409704	9.7409704
Uniformity	0.867	1.867	0.170284568	17.028457
DatasetCharacterstics	0.871	1.871	0.170649398	17.06494
MetadataCoupling	-0.084	0.916	0.083546151	8.3546151
NonDuplicateRows	-0.202	0.798	0.072783656	7.2783656
NonMissingRows	0.101	1.101	0.100419555	10.041955
Unskewness	0.703	1.703	0.155326523	15.532652
CategoricalColumns	-0.082	0.918	0.083728566	8.3728566
Uncorrelation	-0.278	0.722	0.065851879	6.5851879
TOTAL		10.964		100

Fig 2. Principal Component Loadings

4 Results

The underlying study proposes a metric that encompasses the ingredients of data quality and provides a quantitative metric to measure it. This empirical study highlights ingredients that can be individually measured and finally used to calculate the overall quality of data. In this section, we showcase a few easy to validate the proposed metric and empirically observe the highs and lows of data quality successfully identified by the metric. We aim to use a case study involving DHS datasets of India for the years 1998-99, 2005-06 and 2015-16 and various synthetically modified datasets from Kaggle. We have also created a data quality platform that enables users to utilize the benefits of the proposed metric and retrieve a comprehensive report containing values of these defined ingredients. The report also provides ways and means to improve the quality of data to the researcher.

4.1 Case Study: DHS Indian Datasets

Demographic and Health Surveys (DHS) are nationally representative household surveys that provide data for a wide range of monitoring and impact evaluation indicators in population, health, and nutrition [40]. The DHS Program encourages the use of the DHS data in journalistic reporting across the world. Print, television, radio, and digital media use DHS data to support stories about HIV/AIDS, malaria, maternal and child health, and nutrition. DHS Program staff also provide technical assistance to in-country press conferences and press training workshops [41]. It supports a range of data collection options that can be tailored to fit specific monitoring and evaluation needs of host countries while maintaining strict standards for protecting the privacy of respondents and household members in all DHS surveys [42].

As an application to the metric proposed, we conducted an in-depth study on the DHS India Dataset (Individual Recode) for 1998-99, 2005-06 and 2015-16. The data collected from surveys conducted in these three years was divided into sections based on the division on the DHS Recode Manual, commonly called the codebook. All these

divided sections were considered independent datasets and analyzed for data quality using the proposed quality metric.

There were a total of 17 sections, namely, **Respondent's basic data, Reproduction and Birth History, Reproduction, Contraceptive Table +Contraceptive Use, Maternity, Maternity and Feeding, Health History along with Height and Weight, Marriage, Fertility Preferences, Partner's Characteristics and Women's Work, AIDS and Condom Use, AIDS, STIs and Condom Use continuation, Calendar, Maternal Mortality, Malaria, and Domestic Violence**. Each section was analyzed on the nine data quality dimensions to calculate the data quality scores.

For a long time, The Demographic Health Survey (DHS) project is considered the gold standard for nationally represented data collection [43]. Heavy emphasis on data quality is a hallmark of DHS surveys, and hence Computer-assisted personal interviewing (CAPI) is used by DHS officials to improve data quality [44]. Data quality is also enhanced because functions built into the data entry program do not allow inconsistent data to be entered, so the interviewers can probe to avoid inconsistencies during the interview [44]. The data collection process is supervised on every level. The team supervisors and field editors provide the first level of supervision. The supervisors are responsible for closely monitoring the teams' work to ensure that all sampled households are visited, and all eligible respondents are contacted. The second level of supervision consists of central office staff visits to the field. It is expected that the survey director, field coordinators, trainers, DHS staff and possibly other qualified survey staff members visit teams regularly to check on their fieldwork. Finally, a set of field control tables are produced periodically during fieldwork to check. All the factors mentioned above ensure the high data quality of the DHS dataset [44].

The DHS data cites the provenance details, including the source country, origin on its website, and data processing details have been specified in the survey org manual [43]. Hence, according to the metric, the provenance score was calculated and was set to 100 for all three datasets. All the datasets follow the rules of uniformity and dataset characteristics according to the proposed metric.

After conducting statistical analysis on all the datasets, we observed that all the values match perfectly with the data provided by DHS recode manual. Hence, the value of provenance, uniformity and data characteristics are 100 for all sections of the datasets from the three years. The metadata codebook supplied with the dataset was used to calculate the text-similarity scores with the column descriptions in the recode manual to calculate the metadata concordance scores. The recode manual for all the years is similar except for new sections and changes in columns of existing sections. After in-depth analysis, we observe that the metadata matching score ranges from 85-95% for each section, with a few exceptions. While observing the results, we see a constant increase in the metadata matching scores over the years in almost all sections, which support the hypothesis that individual variables of metadata coupling improve with time and technology.

The comprehensive report provides the details of correlations between all the columns from highest to most minor correlation. In the DHS dataset, we observe varying correlations ranging from 2% to 40%. An in-depth analysis of missing cells, duplicate rows and skewness were carried out. The results were recorded and displayed on the platform developed and in the comprehensive report. We observe the frequency of these three parameters ranging from 2-15%, indicating the high standard of data quality of

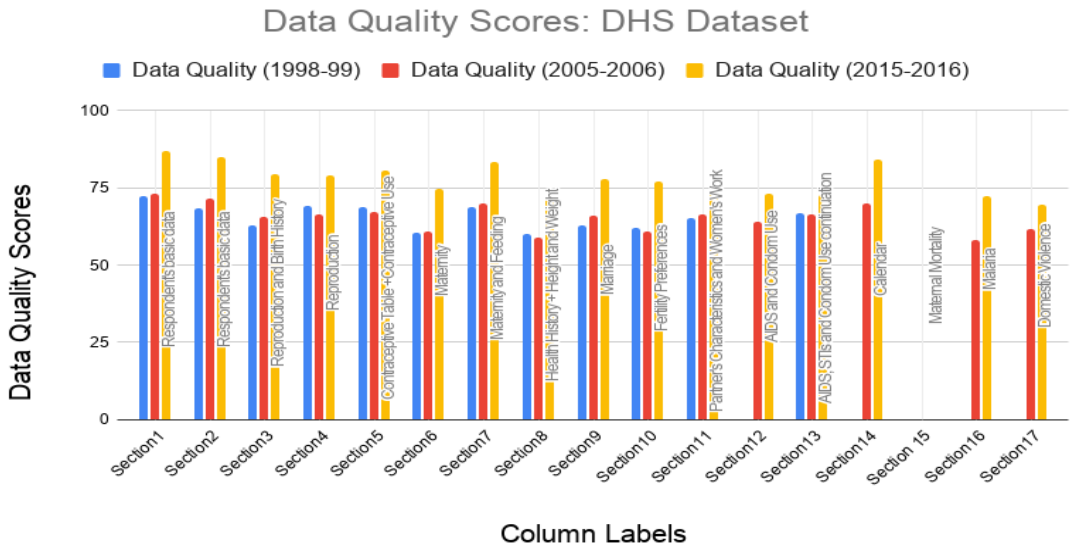

Fig 3. Data Quality Scores: DHS India Dataset

the DHS dataset. For any researcher, categorical datasets are easier to read and have fewer chances of being misinterpreted. Further, we calculated the number of cases where the columns were defined as categorical in the recode manual but were found to be continuous in the dataset and vice-versa to get the values of inconsistencies in their description and provided a score for the study. All the datasets showed a high ratio, again indicating high data quality practices.

After calculating the scores for all data quality ingredients and calculating the data quality for each section for the years 1998-99, 2005-06 and 2015-16, we observe an increase in data quality as seen in figure 3. The graph shows the difference of the Data Quality scores between 2015-16 and 2005-06 alongside 2005-06 and 1998-99. It provides us with better visualization of the increase and decrease of data quality over two consecutive surveys. Data quality increases over the years in all the sections with a few outliers where the decrease is about 2-3%. Our analysis shows that the DHS dataset has a high data quality which is in concordance that data quality is the primary factor for creating survey DHS datasets.

4.2 Mutation Testing on Synthetic Datasets

Mutation testing is an error-based testing technique involving the construction of test data designed to uncover specific errors in metrics [45]. Inspired by this idea, the current empirical study implemented a modified version of mutation testing by forming test data after many mutated versions from the original version. We formed the idea of Mutation Testing wherein some aspects of data are changed, and noise is added to check if our metric can identify these errors. The goal of Mutation testing in data quality is to ensure the authenticity of the defined metric and prove whether it can capture the essence of data quality. Synthetic Datasets are noise-induced datasets generated through computer programs having some form of corruptness. If the proposed metric in this study successfully detects the percentage of noise included by the researcher, then we

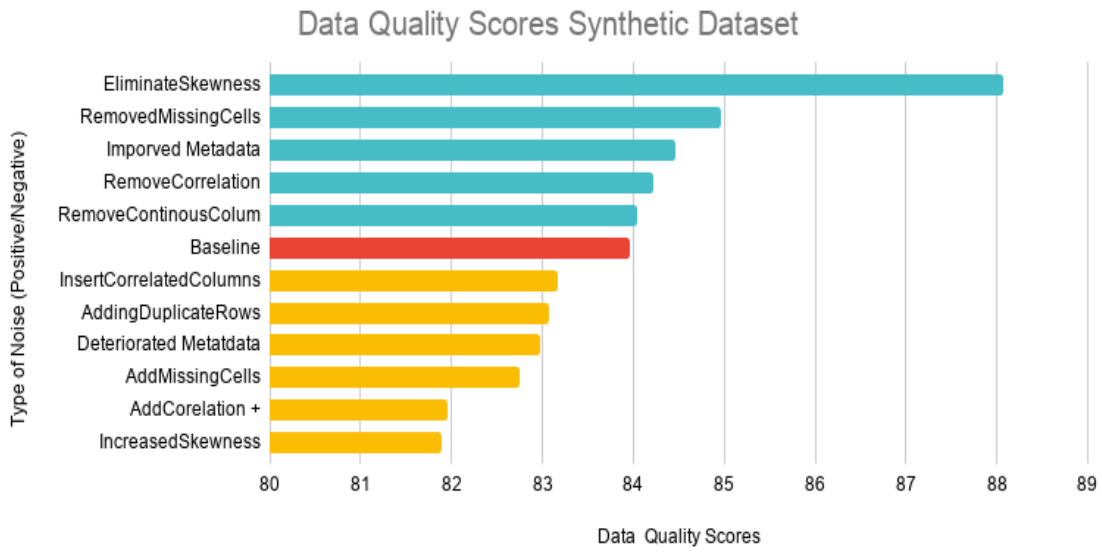

Fig 4. Data Quality Scores Synthetic Dataset

can say that the metric ultimately measures the quality of data.

In this research, we generated synthetic datasets from existing healthcare datasets from Kaggle by introducing and removing impurities.

1. Removing Impurity here refers to factors that would increase the data quality of a dataset. Examples include removing missing cells and duplicate rows, eliminating columns with high correlation, or making the metadata/codebook more descriptive.
2. Introducing Noise highlights the factors that would reduce data quality by making it unfit for machine learning/artificial intelligence algorithms. Examples include adding missing cells or duplicated rows, changing the format of a few cell components making it less uniform/ machine-readable or deteriorating the metadata. It does not convey the meaning of a column heading accurately.

Synthetic datasets were created by adding a combination of positive and negative noise to original data. For every existing dataset, ten additional datasets were analyzed to formulate results for our study. After calculating the data quality for all the datasets, we see an increase in data quality where impurities were removed and a decrease in data quality value when noise was introduced. The trends are shown in figure 4 wherein if the noise is added, i.e., adding missing cells and duplicate rows and adding columns that increase correlation, the data quality gets reduced. However, when the study eliminates the missing cells and duplicate rows, removes all skewness in the dataset and improves the metadata description for all columns, the data quality is improved.

This Mutation Testing of datasets shows that the metric proposed in the underlying study correctly captures the quality of data which would help the researcher gain insights before using it further for research purposes. It would eliminate any instances where a machine learning/artificial intelligence model fails and provides a foundational concept in the study of data quality.

Data Quality Nutrition Label	
INPUT DATASET	
	% value
Provenance	9.74%
Dataset Characteristics	17.06%
Uniformity	17.03%
Number of Categorical and Continuous Columns	8.37%
Statistics	32.85%
% of missing cells	10.04%
% of duplicated rows	7.28%
Percentage of Skewness	15.53%
Metadata Coupling	8.36%
Correlation	6.59%

Fig 5. Data Quality Label

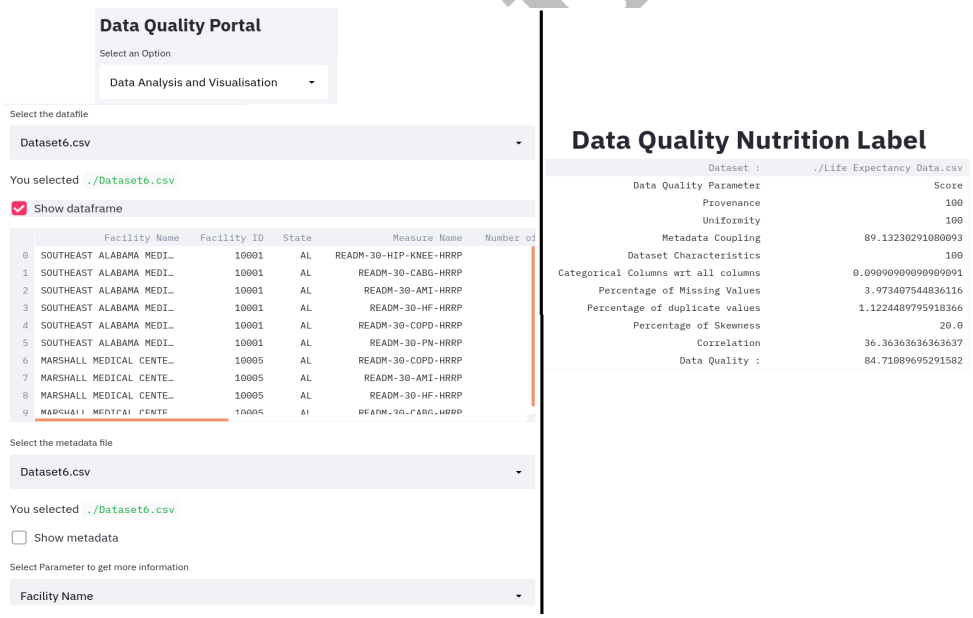

Fig 6. Data Quality Platform : Label

4.3 Data Quality Platform

Machine learning and Artificial intelligence models are severely dependent on the quality of data. Erroneous decisions and results resulting from bad data are inconvenient and time-consuming and, many times, costly. In this research, we present a data quality metric to accurately measure data quality and a data quality platform that provides integrated statistical and visual analysis to summarise the data quality of a dataset. It enables the users to generate a comprehensive report highlighting the problems with their dataset and presents ways and means to improve this quality index.

The platforms take an incoming dataset and the metadata file in an SPSS or CSV format as input. Further, it feeds the pre-processed data into our quality model, which analyses all data quality parameters and evaluates all data quality score as shown in figure 6. Our dynamic platform enables the user to select any variable/columns and see all its descriptive characteristics, including the data type, mean, median, maximum value and minimum value. The correlation graph of the dataset can also be viewed using a simple check box. We present the user with the names of the columns that are highly correlated for future analysis. The values of the data quality parameters, i.e., uniformity, dataset characteristics, percentage of missing values, duplicate rows, skewness, the ratio of categorical variables to all variables, metadata matching score, and correlation value, are displayed on the dashboard. The model uses these measured values to fill the quality label, further added to the comprehensive report. This comprehensive report includes details of all the ingredients and mentions columns where skewness and correlation can be decreased, if any. It also shows columns where the model received low metadata coupling and suggests improving the description of those columns. It provides rows where the model found missing or duplicate values and offers continuous columns that should be converted to categorical to enhance data quality.

If the user wants to learn about the metric and its parameters, they can navigate the "About the metric" section from the drop-down menu and select the parameter to learn more. On choosing the data quality label section, the platform displays the value of all the parameters in the form of a label, as shown in figure 5.

5 Conclusion and Future Work

> You can't control what you can't measure
> - Tom DeMarco

Increased use of data has urged the need for quality data for decision making. Hence data quality checks and their interpretation has become the need of the hour. This can be achieved when the state of the art technologies come into existence to improve the data quality. This includes the coupling of carefully analyzed and discussed **Data Quality "ingredients"** to further improve upon the quality of a dataset. Following the words of Tom DeMarco, we aimed to quantify data quality and formulate an approach to measure the same with and aim to improve it further.

In an effort to improve the current state of practice of data analysis, in this research study, we created the Dataset Nutrition Label, a diagnostic framework that provides a concise yet robust and standardized view of the core components of a dataset. Assessing data quality is an on-going effort that requires awareness of the fundamental principles

underlying the development of subjective and objective data quality metrics. In our research, we represent subjective and objective assessments of data quality in terms of scores generated that check the quality of data. We have developed illustrative metrics for important data quality dimensions.

Finally, we have presented an approach that combines the **subjective and objective assessments** of data quality and demonstrated how the approach can be used effectively in practice. Together, this provides **flexibility, scalability, and adaptability**. With this approach, data specialists can efficiently compare, select, and interrogate datasets. They can provide qualitative and quantitative modules that leverage different statistical and probabilistic models. As a result, data specialists have a better, more efficient process of data interrogation, which will produce efficient Artificial Intelligence models. This research could be the first step in a broader effort toward improving the outcomes of Artificial Intelligence systems that play an increasingly central role in our lives.

Quality of data is an every growing aspect, we can never stop increasing data quality. In our research, we formulated a metric and a data quality platform which can we used by any user to formulate a score of their dataset and utilize it in the best possible way.In the future, we plan to improve our metadata matching algorithm by including sentimental word importance and other corpus, knowledge and hybrid based text similarity algorithms.We also aim to improve the platform by incorporating datasets in forms other than CSV or SPSS and a feature can be added to read metadata directly from the website or from the code book which is in the form of a PDF.The platform will be made more user friendly and more visualization techniques can be added to help the researcher study data in a better way.In addition to this there is also a scope of improving the metric in order to give the researcher solutions on how to improve data quality before using in machine learning/artificial intelligence applications.Additional information can be provided to the owner of the dataset/survey members on why the data quality of the whole dataset or a particular dataset is less and ways to improve data quality in future. We feel that after all these improvements, our project will be well enough for deployment.

6 Acknowledgments

References

1. Yaqoob I, Hashem I, Gani A, Mokhtar S, Ahmed E, Anuar N, et al. Big Data: From Beginning to Future. International Journal of Information Management. 2016;36. doi:10.1016/j.ijinfomgt.2016.07.009.

2. Kaisler S, Armour F, Espinosa J, Money W. Big Data: Issues and Challenges Moving Forward; 2013. p. 995–1004.

3. Fürber C. Data Quality Management with Semantic Technologies. Springer; 2015. Available from: `https://books.google.co.in/books?id=nLQvCwAAQBAJ&pg=PA20&redir_esc=y#v=onepage&q&f=false`.

4. Thatipamula S. Data Analytics; 2020. Available from: `https://smartbridge.com/data-done-right-6-dimensions-of-data-quality/`.

5. Wand Y, Wang RY. Anchoring Data Quality Dimensions in Ontological Foundations. Commun ACM. 1996;39(11):86–95. doi:10.1145/240455.240479.

6. Veregin H. Data quality parameters. Geographical Information Systems. 1999; p. 177–189.

7. Pipino L, Lee Y, Wang R. Data Quality Assessment. Communications of the ACM. 2003;45. doi:10.1145/505248.506010.

8. Willcocks L, Lester S. Beyond the IT productivity paradox. European Management Journal. 1996;14(3):279–290. doi:https://doi.org/10.1016/0263-2373(96)00007-2.

9. Jayawardene V, Sadiq S, Indulska M. An analysis of data quality dimensions; 2015.

10. Missier P, Lalk G, Verykios V, Grillo F, Lorusso T, Angeletti P. Improving Data Quality in Practice: A Case Study in the Italian Public Administration. Distributed and Parallel Databases. 2003;13:135–160. doi:10.1023/A:1021548024224.

11. Fox C, Levitin A, Redman T. The notion of data and its quality dimensions. Information Processing Management. 1994;30:9–19. doi:10.1016/0306-4573(94)90020-5.

12. Sidi F, Hassany Shariat Panahy P, Affendey L, A Jabar M, Ibrahim H, Mustapha A. Data quality: A survey of data quality dimensions. 2013;doi:10.1109/InfRKM.2012.6204995.

13. Huang KT, Lee YW, Wang RY. Quality Information and Knowledge. Proceedings of the Sixth International Conference on Information Quality. 1998;.

14. Laudon KC. Data Quality and Due Process in Large Interorganizational Record Systems. Commun ACM. 1986;29(1):4–11. doi:10.1145/5465.5466.

15. Taleb I, Kassabi HTE, Serhani MA, Dssouli R, Bouhaddioui C. Big Data Quality: A Quality Dimensions Evaluation. In: 2016 Intl IEEE Conferences on Ubiquitous Intelligence Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld); 2016. p. 759–765.

16. Daming Sun BZKG Anxiang Ma, Zhang Y. Metadata matching based on Bayesian network in DataSpace. International Conference On Computer Design and Applications. 2010;5:358–362.

17. Sravanthi P, SRINIVASU DB. SEMANTIC SIMILARITY BETWEEN SENTENCES. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING. 2017; p. 1–14.

18. Holland S, Hosny A, Newman S, Joseph J, Chmielinski K. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. CoRR. 2018;abs/1805.03677.

19. Chug S, Kaushal P. Data Quality Analysis; 2021. Available from: `https://docs.google.com/spreadsheets/d/13jXGvzRNS3YbxiMeodIhZBkNQiKG3cwwhDrEoSvGlqI/edit#gid=1140568609`.

20. Zhao J, Simmhan Y, Gomadam K, Prasanna V. Integration of Distributed, Semantic Provenance Information for EnergyInformatics Workflows. 2021;.

21. Simmhan Y, Plale B. Using Provenance for Personalized Quality Ranking of Scientific Datasets. In: INTL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS: SPECIAL ISSUE ON SCIENTIFIC WORKFLOWS, PROVENANCE AND THEIR APPLICATIONS; 2011. p. 180–196.

22. Deppenwiese N, Duhm-Harbeck P, Ingenerf J, Ulrich H. MDRCupid: A Configurable Metadata Matching Toolbox. Studies in health technology and informatics. 2019;264:88–92. doi:10.3233/SHTI190189.

23. Pawar A, Mago V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. 2018;.

24. Hamming RW. Error detecting and error correcting codes. The Bell System Technical Journal. 1950;29(2):147–160. doi:10.1002/j.1538-7305.1950.tb00463.x.

25. Jürgensen H, Konstantinidis S. Error correction for channels with substitutions, insertions, and deletions. In: Chouinard JY, Fortier P, Gulliver TA, editors. Information Theory and Applications II. Berlin, Heidelberg: Springer Berlin Heidelberg; 1996. p. 149–163.

26. Winkler W. String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. Proceedings of the Section on Survey Research Methods. 1990;.

27. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology. 1970;48(3):443–453. doi:https://doi.org/10.1016/0022-2836(70)90057-4.

28. Smith TF, Waterman MS. Identification of common molecular subsequences. Journal of Molecular Biology. 1981;147(1):195–197. doi:https://doi.org/10.1016/0022-2836(81)90087-5.

29. Wagner RA, Fischer MJ. The String-to-String Correction Problem. J ACM. 1974;21(1):168–173. doi:10.1145/321796.321811.

30. Prasetya DD, Wibawa AP, Hirashima T. The performance of text similarity algorithms. International Journal of Advances in Intelligent Informatics. 2018;4(1):63–69. doi:10.26555/ijain.v4i1.152.

31. Yu M, Li G, Deng D, Feng J. String similarity search and join: a survey. Frontiers of Computer Science. 2015;10. doi:10.1007/s11704-015-5900-5.

32. Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard Coefficient for Keywords Similarity; 2013.

33. Bhattacharyya A. On a measure of divergence of two multinomial populations. Sankhyā Indian J Stat. 1945;7.

34. Oghbaie M, Mohammadi Zanjireh M. Pairwise document similarity measure based on present term set. Journal Of Big Data. 2018;5:1–23. doi:10.1186/s40537-018-0163-2.

35. Jingling Z, Huiyun Z, Baojiang C. Sentence Similarity Based on Semantic Vector Model. Proceedings - 2014 9th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2014. 2015; p. 499–503. doi:10.1109/3PGCIC.2014.101.

36. Siegel P, McCord D, Crawford A. An experimental note on Tversky' s "features of similarity". Bulletin of the Psychonomic Society. 2013;19:141–142. doi:10.3758/BF03330212.

37. Koneru K, Pulla VSV, Varol C. Performance Evaluation of Phonetic Matching Algorithms on English Words and Street Names. In: Proceedings of the 5th International Conference on Data Management Technologies and Applications. DATA 2016. Setubal, PRT: SCITEPRESS - Science and Technology Publications, Lda; 2016. p. 57–64. Available from: https://doi.org/10.5220/0005926300570064.

38. Wang RY, Strong DM. Beyond Accuracy: What Data Quality Means to Data Consumers. Journal of Management Information Systems. 1996;12(4):5–33. doi:10.1080/07421222.1996.11518099.

39. Chao YS, Wu CJ. Principal component-based weighted indices and a framework to evaluate indices: Results from the Medical Expenditure Panel Survey 1996 to 2011. PLOS ONE. 2017;12:e0183997. doi:10.1371/journal.pone.0183997.

40. Program D. DHS Program Methodology Survery Types;. Available from: https://dhsprogram.com/Methodology/Survey-Types/DHS.cfm.

41. Program D. DHS Program Brochure;. Available from: https://dhsprogram.com/Who-We-Are/News-Room/index.cfm.

42. Program D. DHS Program Methodology;. Available from: https://dhsprogram.com/methodology/.

43. Program D. DHS Program Publications Summary;. Available from: https://dhsprogram.com/publications/publication-as19-analytical-studies.cfm.

44. Program D. DHS SURVEY ORGANIZATION MANUAL;. Available from: https://dhsprogram.com/pubs/pdf/DHSM10/DHS6_Survey_Org_Manual_7Dec2012_DHSM10.pdf.

45. Woodward MR. Mutation testing—its origin and evolution. Information and Software Technology. 1993;35(3):163–169. doi:https://doi.org/10.1016/0950-5849(93)90053-6.