Operationalizing the Data Quality Framework

Tindering Datasets

Sezal Chug
2017101
sezal17101@iiitd.ac.in

Priya Kaushal
2017081
priya17081@iiitd.ac.in

BTech Project report for complete fulfillment of the requirements
for the Degree of B.Tech. in Computer Science & Engineering
on 7th May, 2021

BTP Track : Research

BTP Advisors

Dr. Ponnurangam Kumaraguru

Dr. TavPritesh Sethi

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Indraprastha Institute of Information Technology
New Delhi

Student's Declaration

We hereby declare that the work presented in the report entitled **"Tindering Datasets: Data Quality in Action"** submitted by us for the partial fulfillment of the requirements for the degree of *Bachelor of Technology in Computer Science & Engineering* at *Indraprastha Institute of Information Technology, Delhi*, is an authentic record of our work carried out under the guidance of **Dr. Tavpritesh Sethi** and **Prof. Ponnurangam Kumaraguru**.
Due acknowledgments have been given in the report to all material used. This work has not been submitted anywhere else for the reward of any other degree.

..... **Place & Date: .....**

Sezal Chug

Priya Kaushal

Certificate

This is to certify that the above statement made by the candidate is correct to the best of my knowledge.

..... **Place & Date: .....**

Dr. Tavpritesh Sethi

Dr. Ponnurangam Kumaraguru

Abstract

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 gives a specific practical application for the same. The current investigation was undertaken with a purpose to formulate a concrete platform where one can assess the quality of data and get a nutrition label for the same. The proposed system quantifies and qualifies the provided data and assesses them at subjective as well as objective levels. In our research, we have proposed a metric which generates a **Data Quality Label Approach**, **Data Quality Score** and a **Comprehensive Report** for its quality judgment. In this empirical study, the **Demographics and Health Surveys (DHS) Program dataset** is used to judge the quality of data and assign a nutrition label using statistical modelling approaches. The value of the nutrition label would instil confidence in the user in deploying the data for his/her respective application. The results of the current empirical study revealed that due to the growing technology upgradations in data collection and processing, there is a constant gradient increase in the nutrition label score over the years in the DHS dataset. The nutrition label would successfully define the quality of the dataset using **nine "ingredients"**, namely provenance, dataset characteristics, uniformity, metadata coupling, statistics encompassing percentage of missing cells and duplicate rows, skewness of data, number of continuous and categorical columns, the correlation between columns of a dataset and inconsistencies between the highly correlated columns.

The output of ibid research generates data quality metric that helps our model to formulate a comprehensive report which gives an overview of the "ingredients" of the dataset and predicts data quality score that helps the end-users to adjudge the overall quality of data.

Keywords: Data Quality, Demographics and Health Surveys (DHS) Program, Dataset Nutrition Label, MetaData Matching, Pearson Correlation, Data Quality Metric

Acknowledgments

On the occasion of submitting our BTP project report, We would take this opportunity to express our immeasurable appreciation and deepest gratitude for the help and support who in one way or the other, have contributed in making this study possible.

We would like to thank our advisors **Dr. Ponnurangam Kumaraguru and Dr. TavPritesh Sethi** for providing us the opportunity to work on this project and constantly guiding us throughout this project. We would remain obliged to them for being patient, precise and motivating us at all times. We also appreciate the consistent advice of the **Research Assistants** for providing valuable feedbacks during the course of the project.

Last but not the least, we would like to thank our family for supporting us throughout the research and keeping us in good spirits during this pandemic.

Sezal Chug 2017101
Priya Kaushal 2017101

Work Distribution

The reports describes the work that we have done this semester. The distribution of work amongst the sections is given as follows:
Section 3.2 and 4.3: Sezal and Priya
Section 3.5 and 4.3.4 : Priya
Section 3.3, 3.4 and 4.2: Sezal

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

With advancements in the 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. Technologists believe that by 2025, 75 percent of the world population will not only be connected but will also be responsible with continuous generation of data. Researchers have further shown that by the end of 2020, every human would create up to 1.7MB of information every second. While human-generated data is experiencing an exponential growth rate, machine data is increasing even more rapidly. This ever-increasing speed of data growth and its exploding volume has introduced several challenges, some of which are as follows:

- Data Clusters are becoming more complex.
- Additional floor space to store the growing data is needed.
- Data processing stations need excessive power to remain stable and functional.
- High operational costs involved in maintaining a perfect cooling for these data centres is a big challenge.
- Data transactions are becoming hard to manage and insecure.
- Normalization of data from various sources before storing and further processing has become an ever-increasing issue.

- Specific skills are required to integrate and manage big data centres which are still in shortage.
- Standard processing algorithms do not work on high-volume, high-velocity, and high-variety data.

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. Analyzing the data quality levels can help organizations in identifying the pitfalls that need to be resolved in order to enhance its quality. Furthermore, it also helps to assess whether the data in their IT systems are fit to serve its intended purpose. Inaccurate data needs to be identified, documented, and fixed to ensure that executives, data analysts, and other end users are working with accurate and efficient information. Some other essential elements of good data quality include**C**ompleteness, **C**onsistency, **C**oncordance, and **C**onformity to the standard data formats created by a particular organization. These we refer to as the 4C's of dataset norms. Meeting all of these factors is necessary to ensure that data sets are reliable, trustworthy, and suitable for use.

Therefore, data quality has made great strides in gaining respect in the business community. However, with the advent of the machine learning/artificial intelligence realm, it has mainly been neglected under the assumption that the data feeding these algorithms are of high quality. There is always more focus on learning algorithms and models instead of ensuring data quality. Currently, most of the data quality measures are being developed on a need basis to solve a specific problem, but developing a fundamental principle or a metric to measure data quality is lacking.

In our BTech project thesis, we aim to formulate and prove an empirical data quality metric in the form of a nutrition label which quantifies data and measures the degree to which it is fit to serve our purpose. The research illustrates the implementation of our proposed methodology using **Demographics and Health Surveys (DHS)** Program dataset of India as the test dataset and 200+ healthcare datasets from various websites as training data. We propose to include **seven "ingredients", namely provenance, dataset characteristics, uniformity, metadata coupling, statistics encompassing percentage of missing cells and duplicate rows, skewness of data, number of continuous and categorical columns, the correlation between columns of a dataset and inconsistencies between the highly correlated**

columns. Using the train data we train our models to get the value of the weight of each parameter and compare it with the model proposed by us in order to validate our metric.

1.2 Problem Statement

The problems we hope to solve are as follows:

1. Develop a preliminary empirical metric which defines a **nutrition label of dataset quality** to help provide a score to adjudge and formulate a report which describes the intricacies of the dataset.

2. To validate the defined metric, using 200+ healthcare datasets and measure the degree to which it is fit to serve our purpose.

3. Check the proposed ingredients that define the parameters of the dataset to generate a nutrition label for each of the given sections in the dataset. These include **seven "ingredients"**, namely provenance, dataset characteristics, uniformity, metadata coupling, statistics encompassing percentage of missing cells and duplicate rows, skewness of data, number of continuous and categorical columns, the correlation between columns of a dataset and inconsistencies between the highly correlated columns.

4. Create an **automated platform** using state of the art AI/ML techniques to automatically formulate a comprehensive report that encompasses the nutrition label defining the data quality ingredients of the incoming dataset.

Chapter 2

Literature Survey

2.1 Related Work

With the advent of growing technology, the use of data has immensely increased. This "new normal" has not only brought the need for more and more data but also urged researchers to consider the data quality aspect at the same time. Answering this question required the establishment of an empirical data quality metric which defined the parameters for quantifying and qualifying the given data. 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.

Currently, most data quality measures provide an ad hoc basis to solve specific problems, as suggested by Huang et al. [5] and Laudon [6]. 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. In another research by Sun et al [1], metadata matching with the purpose of information integration was considered to be an essential aspect of the assessment of data quality.

In order to address this problem, various researchers have used a variety of techniques. Li and Clifton [8] proposed a classifier which categorizes various attributes based on their speci-

fications and data values. Further, a neural network model used to train and identify similar attributes proved to be efficient. Pantulkar and Srinivas [12] 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. The study observed that the similarity score using a feature-based approach generates a better semantic score than the one generated based on nouns and verbs.

Pawar and Mago [11] proposed an edge-based approach in which a lexical database and corpus statistics helped calculate the similarity between the given two words. They tested their methodology on several domains of the benchmark standard and concluded that word-order and sentence similarity approaches were crucial.

Palopoli et al. [9] proposed an automatic and probabilistic approach for detecting type conflicts if present in the database schemes. Object neighbourhoods were analyzed using graph structures, and only those pairs of objects were identified, which belonged to different schemes but had a similar meaning. This technique helped effectively detect type conflicts using many instances.

Fowler et al. [3] have developed a distributed agent architecture which addresses the need for semantic interoperability. They provided semantic interchange among users by allowing an application developer to express their concepts and relationships in high-level terms that were further translated into the low-level database schemas. To apply such concepts in the business environment, it urges the user to create or discover an appropriate ontology for that domain as well as identify the kinds of data that will be suitable to the application.

Embley et al. [2] proposed using decision tree models in solving the metadata matching problem by investigating additional facets, mainly including terminological relationships, data-value characteristics and expected data values.

In another article by Ken Orr [10], the author says "One certain way to improve the quality of

data: improve its use!". He clarifies by stating that the problem of data quality is fundamentally intertwined in our system and fits into the real world. In other words, how users utilize the data in the system should be given importance. He set up a few general rules which use the feedback control system. These general rules defined the basic norms which any dataset should follow.

- Unused data cannot remain correct for a very long time
- Data quality problems tend to become worse as the system ages;
- The less likely some data attribute (element) is to change, the more traumatic it will be when it finally does change;
- Laws of data quality apply equally to data and metadata (the data about the data) i.e. improve the semantic relationships between the two.

Although most researchers define general rules as the fundamental norms which any dataset should follow, however, in our research, we plan to use empirical data quality metric in establishing decision-making models for matching. On the similar lines of Holland et al. [4], the proposed method utilizes existing data to match the metadata and set **Data Nutrition Label Approach** for quality judgment. We evaluate the incoming dataset based on the chosen "ingredients" and generate a comprehensive report summarizing the quality of the dataset.

Chapter 3

Research Methodology

This section provides a detailed description of the research methodology adopted to carry out this empirical study. The various takeaways and applications of the proposed model are also discussed in section 4.3, such calculation of data quality scores, formation of a comprehensive report with detailed information about all data quality ingredients and the data quality label underlying the respective values of parameters and improvements for any incoming dataset.

In the given research, we calculate the value of the ingredients of our datasets and then the model generates the score to judge the quality of the dataset. The figure 3.1 describes the process in systemic manner easily understandable to the viewer.

- Firstly we check the provenance of the dataset provided and assign a score based on the authenticity and reliability of the dataset.
- Secondly, we divide the dataset into sections to generate analysis reports based on these individual sections.
- It considers the metadata given alongside the section dataset values to check for concordance between the two and generates a metadata score for the same.
- The model calculates the percentage of Uniformity and accuracy of Dataset characteristics of the dataset.
- Further, our system takes into account different statistical attributes like percentage of

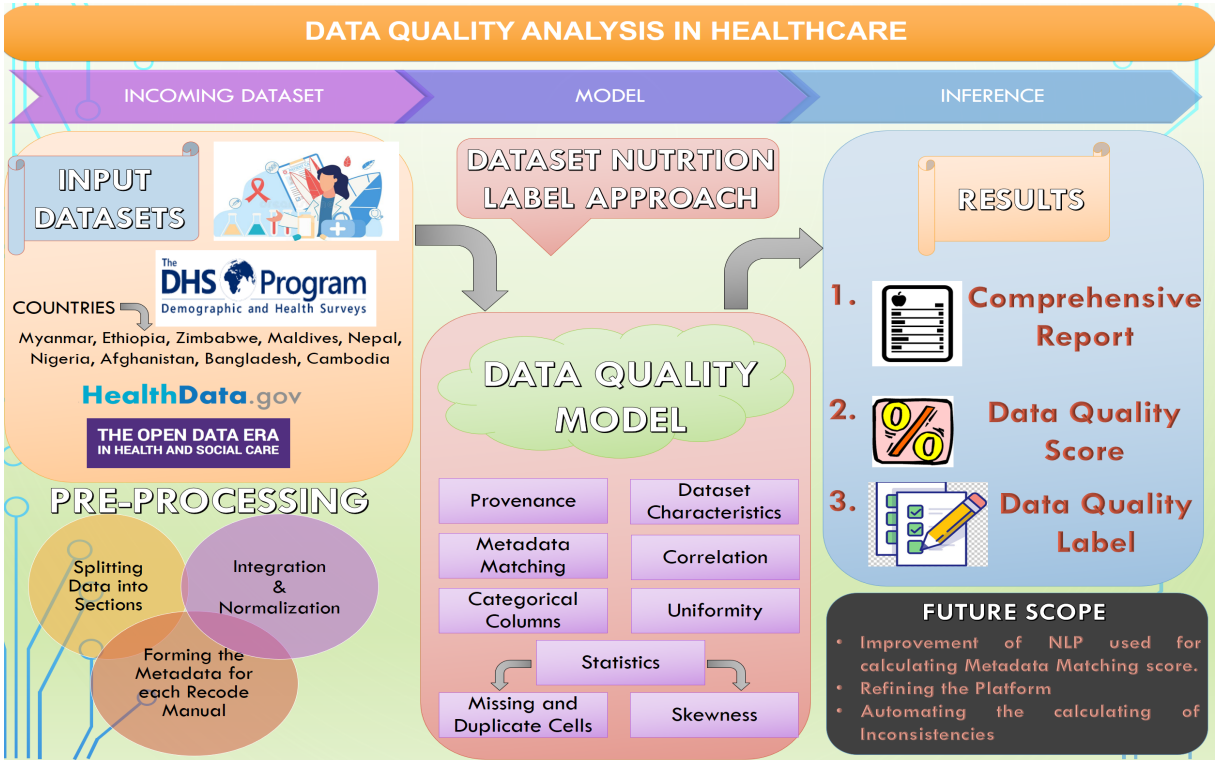

Figure 3.1: Research Methodology

missing and duplicate cells in the section, its value of skewness of the columns and categorical over continuous columns.

- Lastly, it considers correlation matrices for all the given columns of a section of the dataset to judge the dimensionality of our dataset.

3.1 Selection of Quality Parameters

While developing the empirical metric, various parameters available from the literature were studied. However, nine ingredients were carefully selected for the data quality measurement, as shown in Figure 3.2. These nine "ingredients", namely provenance, dataset characteristics, uniformity, metadata coupling, statistics encompassing percentage of missing cells and duplicate rows, skewness of data, number of continuous and categorical columns, and the correlation between columns of a dataset were found to be more relevant for data quality quantification. The brief definition of these data quality expectations' parameters are specified in Table 3.1.

Name of the Attribute	Detailed Definition
Provenance	Information regarding the origin, version, author, and last updated should be correctly mentioned.
Uniformity	The parameters indicates the specific characteristics of the dataset which should match the details given in the metadata such as Number of observations, Number of variables and Size of the dataset.
Dataset Characteristics	All columns should have all the data values similar in datatypes. The values of mean, median, mode, max and min should match the description given in the metadata and the one calculated from the dataset.
Metadata coupling	One of our main focuses was that the documentation of the dataset should be neatly and properly defined. The Metadata description present in the DHS Recode Manual and the Column Description present in the .spss file should correspond to the same data definition. These datasets should have itemized definitions of columns and their values.
Statistics	Statistical Modelling in order to adjudge the data quality are : • Percentage of Missing Values • Percentage of duplicate values • Skewness of the data.
Correlations	Our study uses the Pearson Correlation model to generate the correlation heatmap plots for the data. The columns resulting in high correlations should be taken into account and their interpretation should be handled separately. • Our system includes these highly correlated columns in the comprehensive report published to the end-user. • Pearson Correlation plots are also provided in the comprehensive report
Inconsistency	The highly correlated columns are analyzed individually and inconsistencies are taken into account by the following ways : • Comparing the distribution of data in the dataset with the census data. • If two or more columns are interdependent they must correspond to the similar information. • If two columns are highly correlated they must have a strong relation and should also match when checked. For Instance: Age of a person calculated using Date of Birth column should match the age noted in the categorically divided columns of Age Cluster. If not matched, it is considered as an inconsistency in data.

Table 3.1: List of parameters to decide the score of the dataset on the basis of quality

Figure 3.2: Parameters for Data quality

3.2 Description of Data Quality Parameters

As discussed in section 3.1, in total nine parameters were selected to measure the data quality for DHS India dataset for the following years :

- 1998-1999
- 2005-2006
- 2015-2016

This section includes a detailed description of the nine ingredients and ways on how to calculate the values for each.

3.2.1 Provenance

Provenance is the chronology of the ownership or location of a particular object. It refers to the personal information related to the dataset which specifies the origin, author, version and date uploaded of that particular dataset. While extracting the dataset, we cross-referenced the information regarding these parameters of the dataset and verified them with the ones retrieved from the metadata.

3.2.2 Dataset Characterization

In our research, after ensuring that the data is imported through the original mentioned path, preliminary checks were conducted about the type of data, number of unique values, size of the dataset, and the total number of observations. Finally a report is generated using different python libraries that consist of the basic information about the dataset in consideration.

3.2.3 Uniformity

Quantile statistics like mean median, mode, minimum value, and maximum value are calculated and analyzed for the input dataset as they help us to find real-time information. The descriptive statistics like standard deviation and coefficient of variation are also calculated as it helped in efficient analysis and classification of data. This helps judge the quality of data at a discrete level and provides the user with a high level information about the dataset in question. It also matched these values with the metadata information present in the Recode Manual.

3.2.4 Metadata coupling

Metadata is the description of the dataset that helps a reader understand the relationship between the data and its columns. To check the correctness of data and measure data quality for the DHS Indian dataset, the column description provided by the recode manual must correspond to the same column description as the dataset's metadata. Approaching this problem with a classical view of natural language processing we followed the following steps to calculate metadata matching score.

- **Preprocessing data**: Both the column descriptions from metadata and the recode manual are preprocessed using features of the NLTK library in python. In this process, the data is converted to lowercase, all the special characters and integers are removed. Further, stemming and lemmatization is carried out to get a complete understanding of the description and generate keywords and stopwords.
- **Similarity score calculation**: Thirteen text similarity algorithms help calculate the similarity score, which is further normalized to values between 0 and 1. This normalisation makes the different algorithms at par with each other which facilitates analysis. After this

normalization, we classify each result as similar or not similar based on the algorithm and generate thirteen values of 0 and 1 where 0 represents low similarly and 1 represents high similarity. These scores are finally averaged to generate the metadata matching score.

- **Results**: For each algorithm, the results 1 and 0 values being averaged help us understand the similarity of the metadata in a more generic manner and removes all biases. The highest similarity value is 13 (when all 13 algorithms show high similarity), and the lowest is 0 (when all 13 show low similarity). The final percentage of metadata matching is calculated based on this score.

The state of being similar or a Similarity measure is a quantitative estimate of how two objects are alike. In data mining context, similarity measures are the distance with dimensions representing the features of these word objects. If this distance is small, it will be the high degree of similarity where large distance will be the low degree of similarity. This similarity measure is highly subjective and highly dependent on the domain of the dataset and its application. For example, two fruits are similar because of color or size or their taste. However, while calculating distances for similarity estimate these dimensions/features are unrelated. The relative values of each element must be normalized, or one feature could end up dominating the distance calculation.

Similarity estimates are measured in the range 0 to 1 [0,1] considering:

Similarity = 1 if X = Y (Where X, Y are two objects)

Similarity = 0 if $X \neq Y$

In this study, the features of the TextDistance python library are utilized to get the desired results.

TextDistance : It is a python library used for comparing distance between two or more sequences of data. The thirteen algorithms used are described below:

1. **Hamming distance** :Hamming Distance between two equal length strings is the number of positions of character at which the corresponding sentences differ. The range is [0,length of sequence] where 0 is the maximum similarity and length of sequence is minimum similarity.

2. **Levenshtein distance**: This algorithm gives the minimum number of single-character

$\mathrm{lev}(a, b)$ where

$$\mathrm{lev}(a,b) = \begin{cases} |a| & \text{if } |b| = 0, \\ |b| & \text{if } |a| = 0, \\ \mathrm{lev}(\mathrm{tail}(a), \mathrm{tail}(b)) & \text{if } a[0] = b[0] \\ 1 + \min \begin{cases} \mathrm{lev}(\mathrm{tail}(a), b) \\ \mathrm{lev}(a, \mathrm{tail}(b)) \\ \mathrm{lev}(\mathrm{tail}(a), \mathrm{tail}(b)) \end{cases} & \text{otherwise.} \end{cases}$$

Figure 3.3: Levenshtein distance

The Jaro Similarity sim_j of two given strings s_1 and s_2 is

$$sim_j = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m}\right) & \text{otherwise} \end{cases}$$

Where:

- $|s_i|$ is the length of the string s_i;
- m is the number of *matching characters* (see below);
- t is half the number of *transpositions* (see below).

Two characters from s_1 and s_2 respectively, are considered *matching* only if they are the same and not farther than $\left\lfloor \frac{\max(|s_1|, |s_2|)}{2} \right\rfloor - 1$ characters apart.

Figure 3.4: Jaro winkler distance

edits i.e. insertions, deletions or substitutions which are required to change one word into the other. The range is [0,length of sequence] where 0 is the maximum similarity and length of sequence is minimum similarity. Figure 3.3 shows the formula to calculate Levenshtein distance.

3. **Jaro winkler distance**: The Jaro–Winkler distance is a string metric measuring the edit distance between two incoming sequences. The lower the Jaro–Winkler distance for two strings is, the more similar the incoming strings are. The score is normalized such that 0 means an exact match and 1 means there is no similarity. Figure 3.4 shows the formula to calculate Levenshtein distance.

4. **Strcmp95**:The strcmp95 function returns a double precision value from 0.0 (total disagreement) to 1.0 (character-by-character agreement). The returned value is a measure of the similarity of the two strings.

5. **Needleman wunsch**:The Needleman-Wunsch (N-W) algorithm is a dynamic programming algorithm for optimal sequence alignment formulated by Needleman and Wunsch in 1970. The N-W algorithm uses four parameters for computing the similarity score of two strings. These parameters are:

    - The first input string

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Figure 3.5: Jaccard Similarity

For sets X and Y the Tversky index is a number between 0 and 1 given by: $tversky_index(X,Y) = \frac{|X \cap Y|}{|X \cap Y| + lpha|X - Y| + eta|Y - X|}$ where, :math: lpha, eta >=0

Figure 3.6: Tversky Index

- The second input string
- Similarity matrix (which has "number of nodes" rows and "number of nodes" columns)
- A penalty gap value for the unmatched character

The higher score obtained from the N-W algorithm shows the better similarity. In order to make scores comparable with each other, the scores should be normalized.

6. **Smith waterman** :The Smith–Waterman algorithm is a dynamic programming algorithm which performs local sequence alignment. This helps the algorithm determine similar regions between two strings. The higher score signifies higher similarity between the incoming strings.

7. **Jaccard Similarity**:The Jaccard coefficient measures similarity between finite sample sets, and is defined as the size of the intersection divided by the size of the union of the sample sets. The range is from [0, 1] where 0 is the least similar and 1 signifies highly similar. Figure 3.5 shows the formula to calculate Jaccard Similarity.

8. **Tversky Index :** The Tversky index is an asymmetric similarity measure on sets that compares a variant to a prototype. The Tversky index can be seen as a generalization of Dice's coefficient and Tanimoto coefficient. The range is from [0, 1] where 0 is the least similar and 1 signifies highly similar. Figure 3.6 shows the formula to calculate Tversky Index.

9. **Overlap** : It is a similarity measure that calculates the overlap between two finite sets where 1 refers to total overlap (high similarity) and 0 signifies no overlap (low similarity). Figure 3.7 shows the formula to calculate Overlap.

$$\text{overlap}(X, Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)}$$

If set X is a subset of Y or the converse then the overlap coefficient is equal to 1.

Figure 3.7: Overlap

Presented in mathematical terms, if samples X and Y are bitmaps, X_i is the ith bit of X, and $\wedge, \vee$ are bitwise *and*, *or* operators respectively, then the similarity ratio T_s is

$$T_s(X, Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

Figure 3.8: Tanimoto similarity

10. **Tanimoto similarity** This similarity algorithm is defined as the ratio between the number of common bits, divided by the number of bits set (i.e. nonzero) in either sample. The range is from [0, 1] where 0 is the least similar and 1 signifies highly similar. Figure 3.8 shows the formula to calculate Tanimoto similarity.

11. **Cosine similarity** This similarity algorithm helps calculate the normalized dot product of the two attributes. By determining the cosine similarity, we would effectively try to find the cosine of the angle between the two objects. Two vectors having same orientation have a cosine similarity of 1, whereas, incoming vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0,1]. The smaller the angle, higher the cosine similarity. The mathematical representation of the index is written as the shown figure 3.9.

12. **Match rating approach**: This algorithm calculates the number of unmatched characters by comparing the strings from left to right and then from right to left, simultaneously removing identical characters. This value is subtracted from 6 and then compared to a minimum threshold.The minimum threshold are shown in the figure 3.10.

13. **Editex distance measure**: This similarity algorithm gives the Editex distance between

$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Figure 3.9: Cosine similarity

Sum of Lengths	Minimum Rating
≤ 4	5
4 < sum ≤ 7	4
7 < sum ≤ 11	3
= 12	2

Figure 3.10: Match rating approach

two strings. High match score refers to high similarity and low score represents low similarity.

3.2.5 Statistics

Statistical Modeling of the dataset while calculating the percentage of missing and duplicate cells and skewness of data helped us understand if any kind of bias is present in the data. This helped understand the dataset at an intricate level to get a clear picture how divided the information is in the dataset.

3.2.6 Correlations

Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. The formulas return a value between -1 and 1, where:

- A correlation coefficient of 1 means that for every positive increase in one variable, there is a positive increase of a fixed proportion in the other. For example, shoe sizes go up in (almost) perfect correlation with foot length.
- A correlation coefficient of -1 means that for every positive increase in one variable, there is a negative decrease of a fixed proportion in the other. For example, the amount of gas in a tank decreases in (almost) perfect correlation with speed.
- Zero means that for every increase, there isn't a positive or negative increase. The two just aren't related.

Mainly Pearson's correlation coefficient is used to calculate the values and get highly correlated columns. These correlation coefficients define if there exists some interrelationship between independent variables and hence, gives us a metric for the same. In case a high correlation

between two different independent variables is found, these variables may or may not be removed depending on the need of the user.

Pearson's correlation coefficient

It is used to summarise the strength of linear relationship between two sets of data. In simple terms, define if we can draw a line graph to represent the dataset to be judged. It is represented using the formula shown in equation 3.1.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{3.1}$$

$$N = Numberofpairsofscores$$

$\Sigma xy = sumofproductsofpairedscores$

$\Sigma x = sumofxscores$

$\Sigma y = sumofyscores$

$\Sigma x2 = sumofsquaredxscores$

$\Sigma y2 = sumofsquaredyscores$(3.2)

3.2.7 Inconsistencies

Correlation is a term that refers to the strength of a relationship between two variables where a strong, or high, correlation means that two or more variables have a strong relationship with each other while a weak or low correlation means that the variables are hardly related. The highly correlated columns are columns containing similar data are analyzed individually and inconsistencies are taken into account by the following ways :

- Comparing the distribution of data in the dataset with the census data of that particular year and generating the percentage of inconsistencies.
- If two or more columns are interdependent they must correspond to the similar information else calculate the difference and add the result.

- If two columns are highly correlated they must have a strong relation and should also match when checked.
  - This implies that if two columns present similar data, then they should contain the same values of that data. The values from Column A should match the values corresponding to Column B.
  - For Instance: Age of a person calculated using Date of Birth column should match the age noted in the categorically divided columns of Age Cluster. If not matched, it is considered as an inconsistency in data.

3.3 Formulation of metric

Heath care datasets from various websites and platforms were calculated to formulate the training data which was further used to formulate the metrics given below. This section includes the approaches of retrieving loadings of each ingredients to get the final data quality score for any incoming dataset.

3.3.1 Principal Component Analysis (PCA)

This technique to calculate the principal components loadings of dataset which then decrease dimensionality of large datasets by considering only a top few variables, at the same time reducing the loss. It gives the coefficients of linear combination of the original columns of the dataset. These coefficients help to determine the correlation of the variable with the principal loading. The values can be positive or negative depending on the variable indicating positive or negative correlation. Positive loadings indicate a positive correlation between the variable and the principal component, and negative loadings indicate a negative correlation. Higher the value of loading, the stronger is the effect of the variable.

After normalising the values of the nine ingredients, PCA analysis was carried out to find the loading for each ingredient. To formulate a final metric 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. These loadings are normalized and the percentage of each “ingredient” over a total of 100 was calculated. The results of the above approach is given in the next chapter 4 for the training data.

3.3.2 Linear Regression

Another approach was taken to get trends of data quality to find concordance and validate the above metric. In addition to PCA we used **Linear regression** to get the coefficients of all the data quality "ingredients". The ingredients were treated as independent features and the data quality calculated using cosine similarity as the dependent feature. Linear regression gave us another metric which was used to validate the results achieved on testing data. This approach was also used to calculate the 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 and trends over both metrics were observed in chapter 4.

An ideal dataset would have full provenance description, correct uniformity, 100 percent concordance between metadata and data columns, no missing cells, no duplicate values, and unskewed data with no correlation between the columns. That is the value of all the ingredients would be full and hence leading to perfect data quality score i.e 100. The value of the nine data quality ingredients were considered as vectors and similarity score of each dataset with ideal standard dataset was calculated. The cosine similarity between ideal dataset and any incoming dataset is calculated using the formulae given above in section 11. We used this score to calculate data quality value of all the datasets in our training data and then further feed this dependent and independent dataset into the Linear Regression model which further retrieved the values of coefficients for each ingredient.

3.4 Report Generation

In this research study, the provided comprehensive report outlines the existing details about the parameters of a dataset and improvements that the user can take to get better data quality results. The below mentioned points show the improvements suggested by the model.

- The variables which 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 to improve the description of those columns.

- It provides rows where the model found missing or duplicate values.
- It also suggests continuous columns which can be converted to categorical to improve data quality.

3.5 Data Quality Platform

Streamlit is an open-source app framework for machine learning and data science. We used streamlit to make a data quality dashboard application.

The user can enter data in the form of a CSV or SPSS file along with the metadata in the form of a CSV file. Our application will return the following.

- Detailed view of statistics and Datatype for data variables
- Provenance score
- Uniformity score
- Metadata Matching score
- Missing cells
- Duplicate rows
- Skewness
- Number of categorical and continuous columns
- Name of highly correlated columns and percentage of correlation
- Correlation Graph
- Final Data Quality score with label

Along with this, the researcher can also see a detailed view of the data quality parameters and what they signify. The researcher then has the freedom to use this dataset or not.

Chapter 4

Results

4.1 Data Retrieval

This empirical study involved the **Demographics and Health Surveys (DHS) Program datasets** which contains unrestricted survey data files for legitimate academic research. Approved registration of this dataset over the years of **1998-1999, 2005-2006 and 2015-2016** gave us the ability to access the DHS program and formulate data quality metric to get score and a comprehensive report to analyse the retrieved ingredient values.

In this semester we collected healthcare survey datasets from the DHS portal for the year 2015-16 of different countries like **Myanmar, Ethiopia, Zimbabwe, Maldives, Nepal, Nigeria, Afghanistan, Bangladesh and Cambodia** for training data. In addition to this we also studied various datasets from **Government Health Data** and **Health and Social Care Open Data platform**. All the datasets were combined to form our training data which had around two hundred entries.

- In our study, data quality assessment of the collected DHS data was performed by dividing the dataset into sections based on the **DHS Recode Manual** using the IBM SPSS tool.
- Web scrapping was performed to retrieve the contents on the webpages of these websites to get the column descriptions.
- This data was then fed to our data quality platform to get the values of the data quality ingredients.

This provided us with an opportunity to get results of different attributes and their values over the different years to compare them respectively. In the given research, we calculate the value of the proposed "ingredients" and generate a nutrition label to quantify the quality of the dataset.

4.2 Metric Formulation

In our thesis, we came upon certain parameters that we considered could define the way the quality of data is judged. These parameters can be individually monitored or effectively observed. All these parameters are mentioned already in figure 3.2 and defined in Table 3.1.

The DHS Recode Manual contains a description for each column which would help in metadata matching of the column name and its description. The Manual was available in the form of a PDF with no structure, hence to use this comparison metric of column names and description, we manually added to the metadata to generate a CSV file for easy analysis. This CSV includes the name of the column along with the description available in Recode Manual. The metadata coupling values for each columns are then averaged over all the columns to get a metadata matching score for each incoming dataset.

To begin the analysis, the subjective and objective assessments of a specific dimension were compared. After careful analysis of the literature on Data Quality and considerable efforts and discussions, we came to conclude a metric which encompassed "ingredients" of a dataset that helped adjudge its quality.

4.2.1 PCA

The weights of the data quality ingredients and their percentage over the overall data quality using the Principal Component Analysis approach is shown in figure 4.1. Using the training data, the research first calculated the first principal loadings of the dataset using PCA. After shifting the offset of these values from [-1,1] to [0,2] by adding 1 to all principal component loadings to make all values positive, the research conducted normalisation of the values of the nine ingredients. This helped retrieved the percentage of each "ingredient" over a total of 100 which further provided the weightages for each ingredient to get Data Quality Scores.

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

Figure 4.1: PCA Component Loadings

4.2.2 Linear Regression

The weights of the data quality ingredients and their percentage over the overall data quality using the Cosine Similarity and Linear Regression is shown in figure 4.2. After calculating the cosine similarity values of the each dataset with the ideal dataset, we subtracted this value from 100 to get an approximate data quality score for each dataset. This approximate Data Quality now became our dependent variable where the ingredient values for all dataset sin training data became our independent values. Further our model performed Linear Regression to retrieve coefficients for each ingredient as shown in figure 4.2.

In our research we compared the results retrieved using these two metrics and analyze the data quality trends int the DHS India dataset over the years 1998-99, 2005-06 and 2015-16. The weights retrieved by PCA and Linear Regression are shown in figure 4.1 and figure 4.2 respectively.

4.3 Applications of Metric

The study helped us propose a metric which then led to calculation of data quality scores for any incoming dataset. However, to prove viability of any metric, we need to show applications where the proposed metric shows validity and usability. In the below sections, the study shows few instances where using the proposed metric, we retrieved good results that not only prove its

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients
		B	Std. Error	Beta
1	(Constant)	37.747	9.746	
	Provenance	.065	.004	.473
	Uniformity	.142	.098	.068
	DatasetCharacterstics	.215	.135	.073
	MetadataCoupling	.056	.005	.342
	DuplicateRows	.047	.002	.747
	MissingRows	.043	.004	.371
	Unskewness	.030	.004	.249
	CategoricalContinous	-.380	.159	-.084
	Uncorrelation	.042	.004	.417

a. Dependent Variable: DataQuality

Figure 4.2: Linear Regression Weights

validity but also show its usability.

4.3.1 Data Quality Trends

After calculating the data quality values for the Indian DHS dataset over the years using the metric from PCA and Linear Regression we observe a similar trend in the change in data quality. Figure 4.4 shows the trends in data quality with respect to the Years of publishing of datasets which is 2015-16, 2005-06, and 1998-99 over all the 17 sections of the DHS dataset. Figure 4.3 shows the trends in data quality with respect to all the 17 Sections the DHS dataset ordered by publishing years of datasets which is 2015-16, 2005-06, and 1998-99.

In the graphs, we observe that the increase and decrease of data quality with respect each dataset and observe similar trends. The similarity in trends prove that the **data quality is independent of the metric and can be accurately captured by the ingredients proposed by the study**.

- PCA Metric: Coefficients as viewed in data quality label shown in figure 4.5.
- Equal Weightage: Equal percentage i.e. 100% over all the nine data quality ingredients gives 11.11%.
- Linear Regression: Weightages retrieved by calculating data quality using cosine similarity

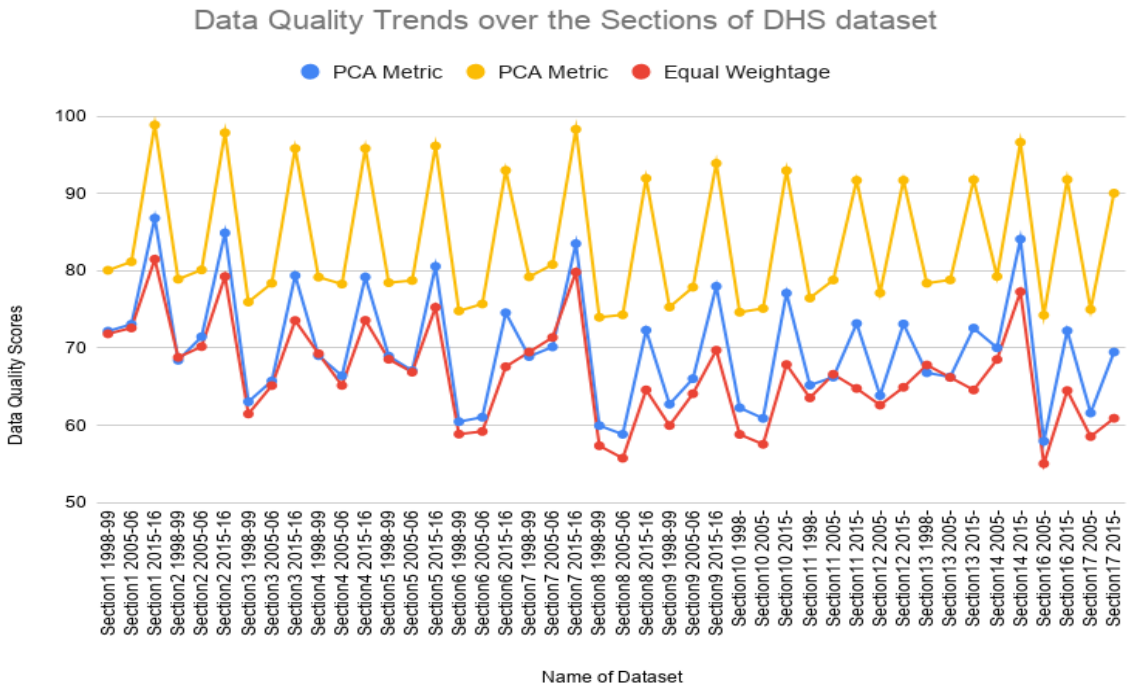

Figure 4.3: Data Quality Trends over the Sections of DHS dataset

between ideal and ith dataset and further performing Linear Regression between dependent and independent columns.

In order to finalise one metric for getting the final data quality scores, we used the weights calculated after PCA analysis of the dataset. This is because **PCA is considered a standard and reliable metric for calculating the weights of each parameter**. Linear regression is not as reliable as we used the approach of cosine similarity to get the dependent variable, the correctness of this approach is not yet confirmed. Depending on the approach to calculate the dependent variable for all datasets, the linear regression coefficients may vary making this approach less reliable.

Hence the weights from PCA are considered to formulate the proposed metric that the study uses. The data quality label can be seen in figure 4.5.

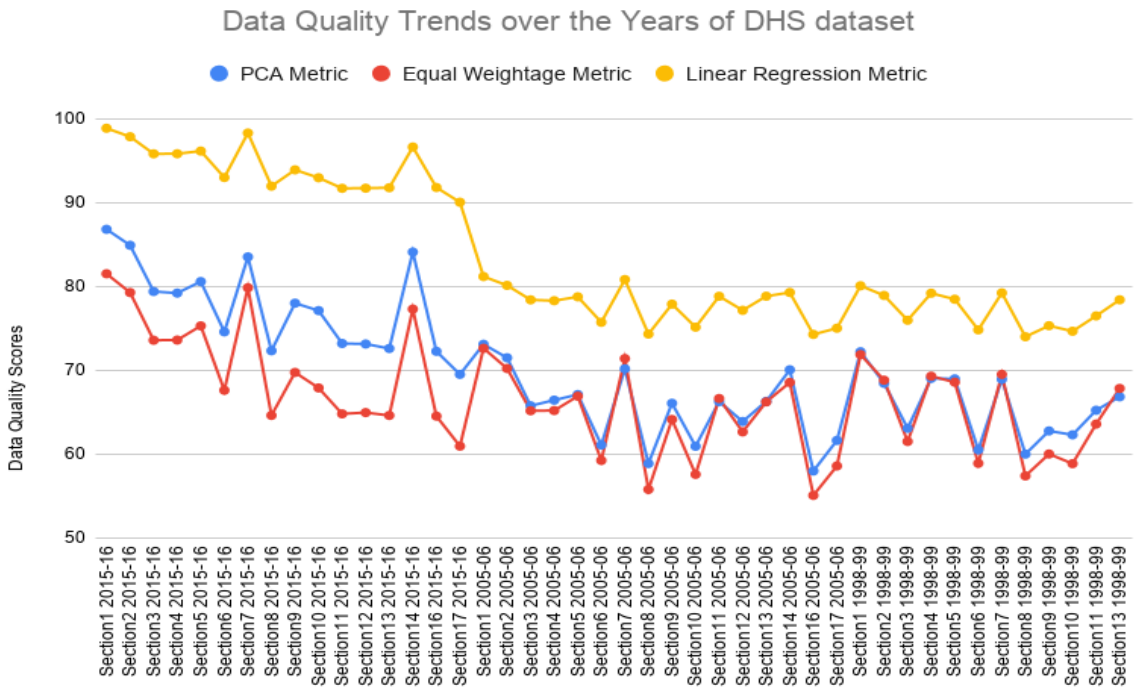

Figure 4.4: Data Quality Trends over the Years of DHS dataset

4.3.2 Case Study

The Demographics and Health Surveys (DHS) Program dataset over the years 1998-99, 2005-06 and 2015-16 when analyzed using the **newly proposed nutrition label metric showed an increase in the data quality score which successfully authenticated the metric**. The metric defined in the research included two aspects, i.e. the nutrition label and a calculated score shown figure 4.5 shows the Nutrition Label and its ingredients discussed above.

While observing the results, we see a **constant increase in the data quality scores over the years** in almost all sections. This validates our defined data quality metric and justifies the hypothesis that Data Quality is increasing with the advancement of technology. Figure 4.6 gives us a visual representation of the above data wherein we observe a constant increase of the size of the bar in the chart.

To better understand this increase, we use the graph in Figure 4.7 which shows the difference of the Data Quality scores between 2015-16 and 2005-06 alongside 2005-06 and 1998-99. This gives us a better visual representation of the increase or decrease of data quality scores between the two consecutive survey reports.

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%

Figure 4.5: Data Quality Label

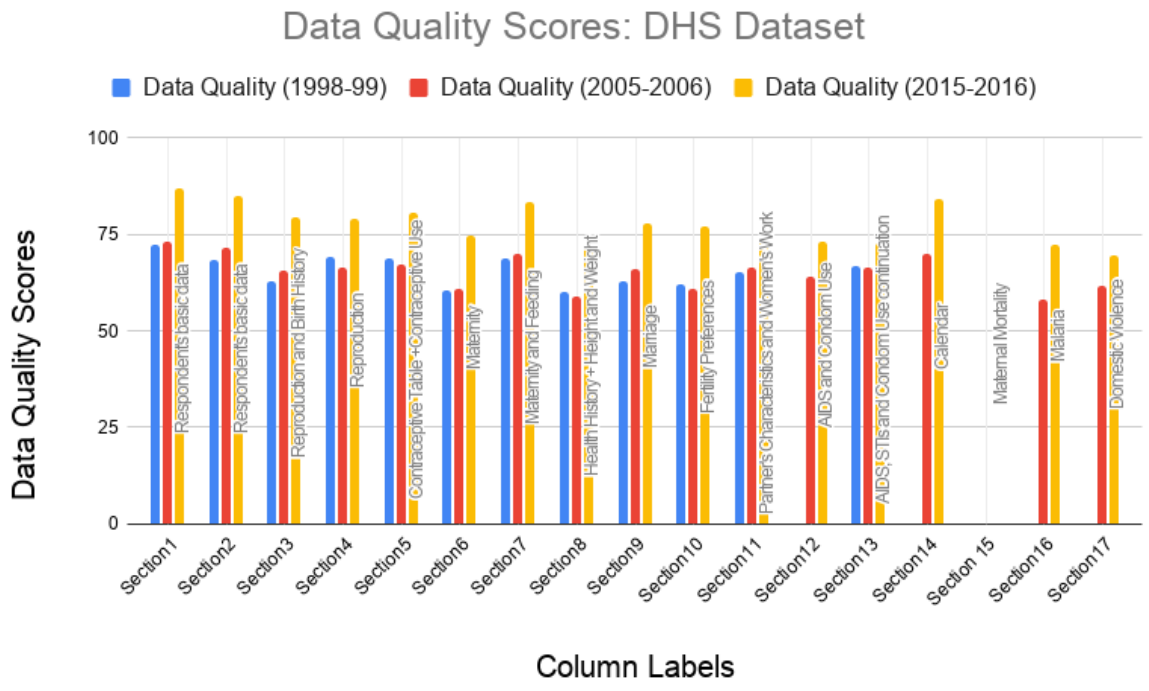

Figure 4.6: Data Quality Scores of DHS Dataset over the years 2015-16 and 2005-06 alongside 2005-06 and 1998-99

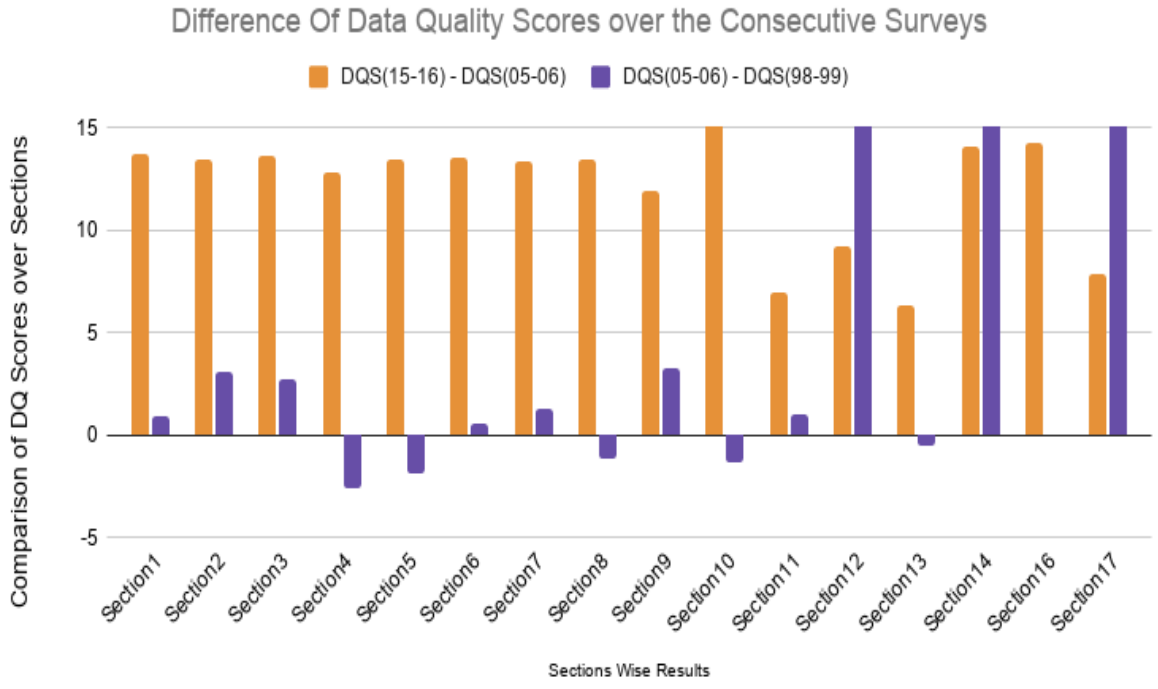

Figure 4.7: Difference of Data Quality Scores over the Consecutive Survey

The results of the current empirical study revealed that due to the growing technology upgradations in data collection and processing, there is a constant gradient increase in the nutrition label score over the years in the DHS dataset. The nutrition label would successfully define the quality of the dataset using the proposed **nine "ingredients", namely provenance, dataset characteristics, uniformity, metadata coupling, statistics encompassing of missing cells and duplicate rows, skewness of data, number of continuous and categorical columns, the correlation between columns of a dataset and inconsistencies between the highly correlated columns.**

4.3.3 Synthetic Datasets

To analyse the effectiveness of the proposed metric, the study suggested to observe the change in values of data quality with respect to change in values of ingredients. This helped to check if our model captures the difference between a dataset with good data quality and a dataset with less data quality. Taking a few datasets from Kaggle, we performed the above experiment and observed values of data quality score changing over the type of noise that was added to the data.

- Negative Noise was inserted in the form of adding missing cells, duplicate row samples, highly correlated columns, skewness to the dataset by removing some category in a column, and making the metadata worse.
- Positive noise was added in the form of removing missing cells, duplicate rows, making the metadata better, removing correlation between columns, and making continous columns categorical.
- For every dataset, 10 other datasets were created by adding one type or multiple noise sources. The data quality for all these baseline and modified datasets was calculated and compared as shown in figure 4.8.

We observed that the data quality of the datasets created by adding negative noise decreased as compared to the original baseline dataset. The data quality increased after adding positive noise. These observations prove that our metric can capture the data quality trend for all kinds of datasets correctly and in concordance with the weightage of ingredients. We can observe that the parameters vary with the Correlation, Percentage of Non Duplicate Cells, Metatdata Coupling, Fraction of Categorical Columns, Provenance, Percentage of Non Missing Cells, Skewness of

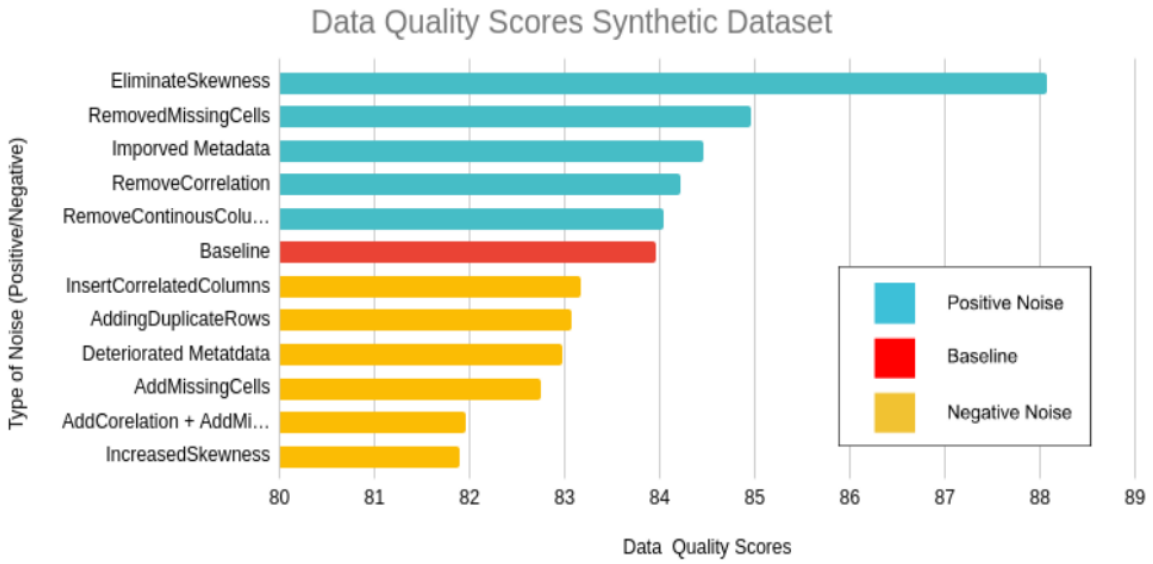

Figure 4.8: Data Quality of Synthetic Datasets

data, Uniformity and Dataset Characteristics which is same as they vary with the coefficients of the PCA Metric.

4.3.4 Data Quality Platform

Data Quality Platform built using streamlit, which is a python library. It is a dashboard where the user can add any dataset along with metdata in the form of a SPSS or a CSV file and get data descriptive characteristics of all variables along with value of data quality ingredients. Using this dashboard the researcher can decide on how to utilize the data and make the most of it.

1. After opening the Data Quality Portal the user can upload a CSV or SPSS file as shown in figure 4.9. After uploading the file, the user had the option to view the dataset uploaded using a checkbox. After clicking on the checkbox the top five columns of the data will be visible.

2. Further, the user is required to upload a metadata file in the form of a CSV file as shown in the figure 4.10 with the same option to view the uploaded file using a checkbox. The user can view descriptive characteristics of each column in the dataset by selecting the name of the column using the drop down menu, these include if the variable is numeric or

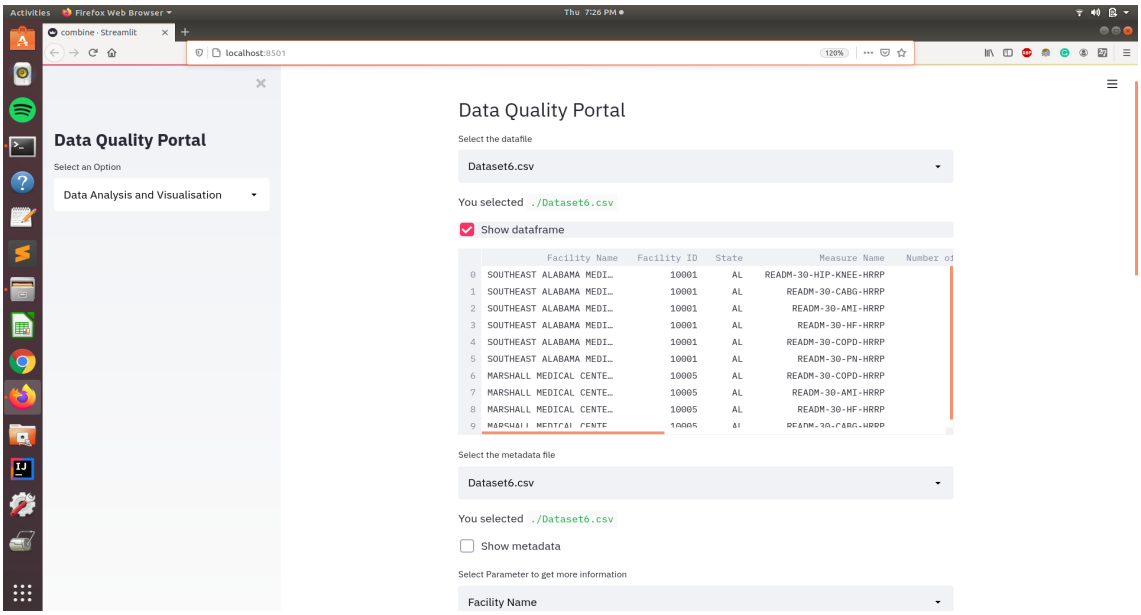

Figure 4.9: Data Quality Platform: Uploading Dataset Files

categorical, the mean, standard deviation, minimum and maximum value as shown in the figure 4.11.

3. After both the uploads are complete the platform automatically calculates the value of data quality "ingredients" and displays it on the dashboard as seen in figure 4.12. The user also has an option to view the Pearson correlation graph for the dataset using a checkbox.

If the user wishes to learn more about the data quality metric being used, he can use the drop down menu on the left and select "About the metric". Here the user can select any data quality ingredient from the drop down menu and see the details.This can be seen in the figure 4.13.

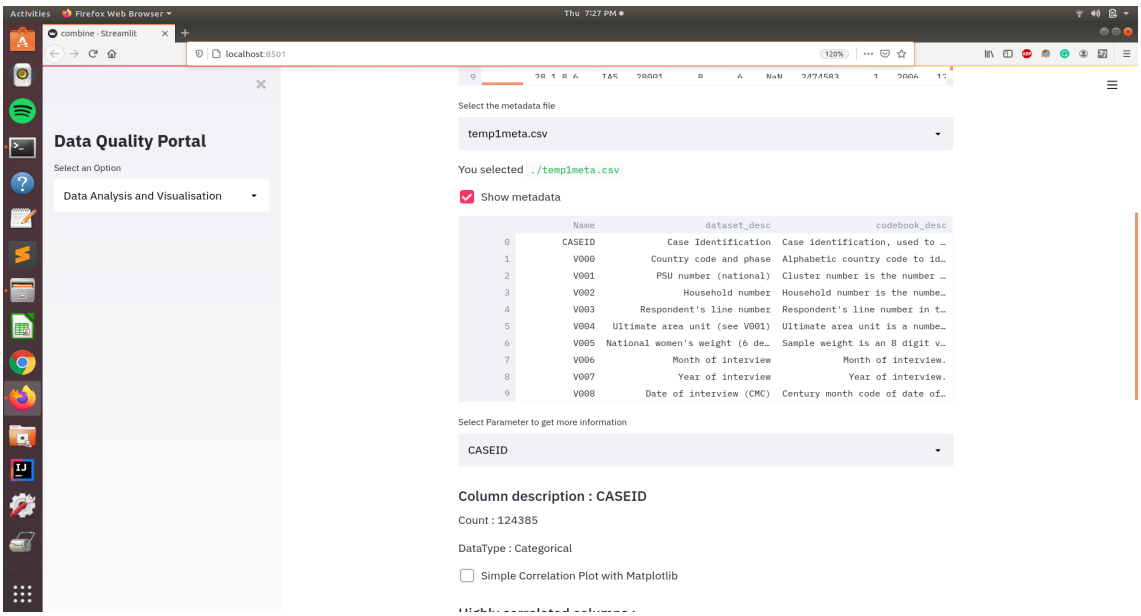

Figure 4.10: Data Quality Platform: Uploading Metadata files

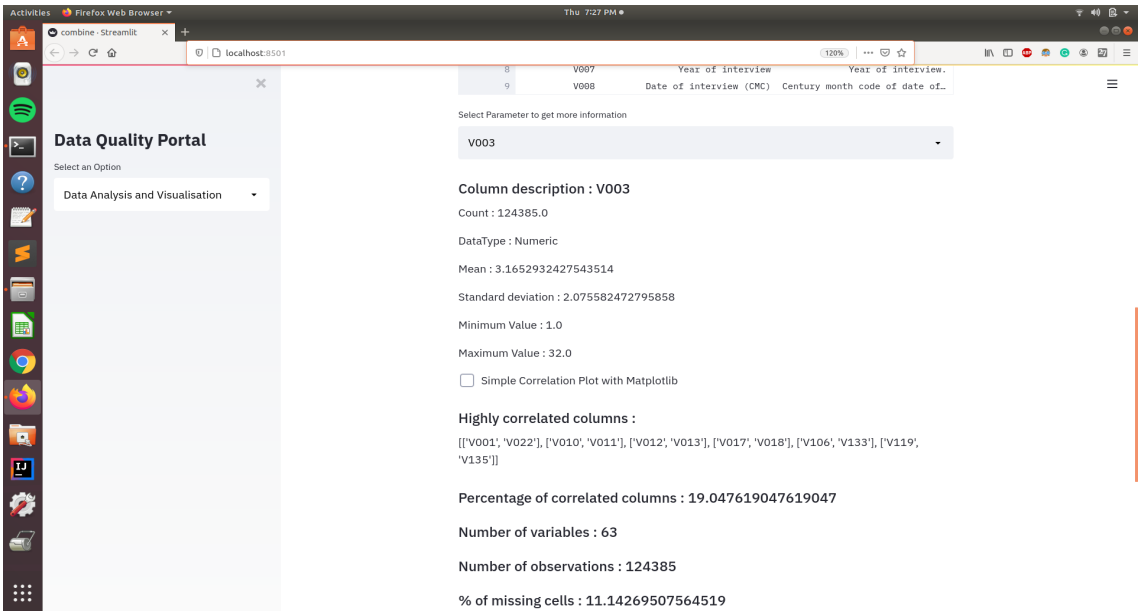

Figure 4.11: Data Quality Platform: Dataset Characteristics

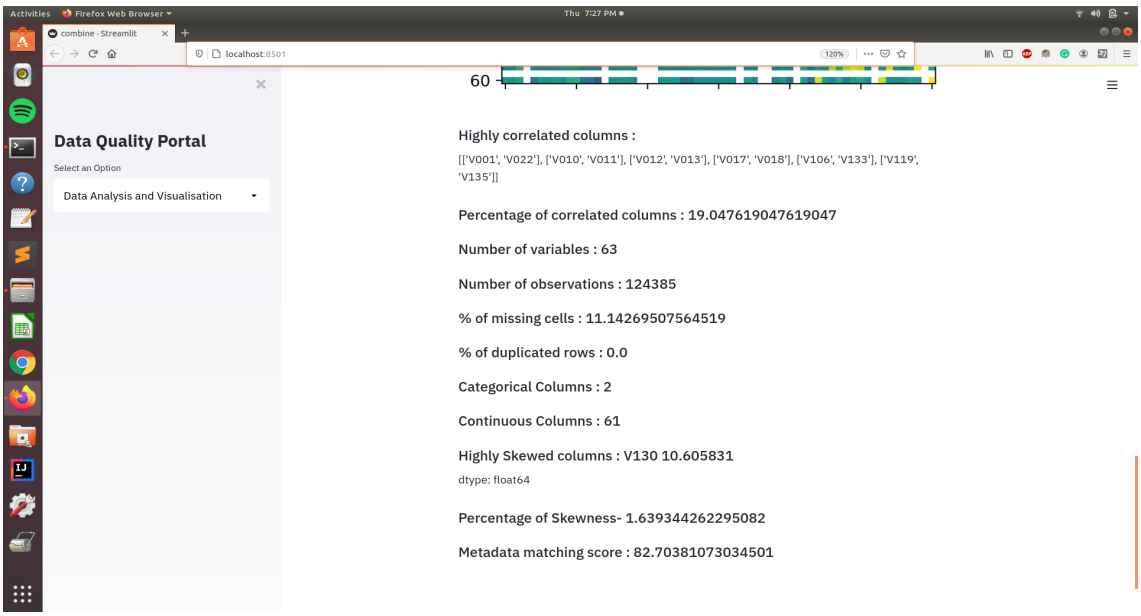

Figure 4.12: Data Quality Platform: Values of Data Quality Ingredient

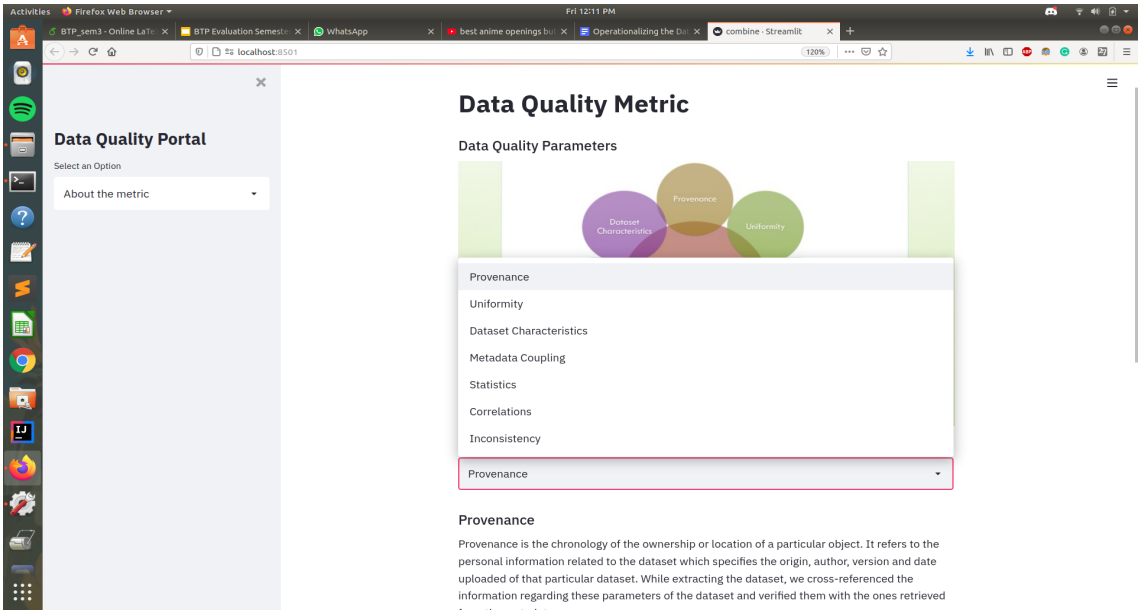

Figure 4.13: Data Quality Platform: About the Metric

Chapter 5

Conclusion

> 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.

Chapter 6

Future Plan

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.

1. In the future, we plan to improve our metadata matching algorithm by including **sentimental word importance and Bag of words algorithm** containing more words related to the healthcare industry.

2. 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.

3. 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.

4. 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.

5. 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.

Bibliography

[1] Daming Sun, Anxiang Ma, B. Z. K. G., and Zhang, Y. Metadata matching based on bayesian network in dataspace. *International Conference On Computer Design and Applications 5* (2010), 358–362.

[2] Embley, D., Jackman, D., and Xu, L. Multifaceted exploitation of metadata for attribute match discovery in information integration. *Workshop on Information Integration on the Web* (01 2001).

[3] Fowler, B.Perry, M., and Bargmeyer, B. Agent-based semantic interoperability in infosleuth. *ACM SIGMOD* (1999), 60–67.

[4] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR abs/1805.03677* (2018).

[5] Huang, K.-T., Lee, Y. W., and Wang, R. Y. Quality information and knowledge. *Proceedings of the Sixth International Conference on Information Quality* (1998).

[6] Laudon, K. C. Data quality and due process in large interorganizational record systems. *Commun. ACM 29*, 1 (Jan. 1986), 4–11.

[7] Leo L. Pipino, Y. W. L., and Wang, R. Y. Data Quality Assessment. *COMMUNICATIONS OF THE ACM 45* (2002).

[8] Li, W., and Clifton, C. Semantic integration in heterogeneous databases using neural networks.

[9] L.Palopoli, D., and D.Ursino. An automatic technique for detecting type conflicts in database schemes. *ACM CIKM International Conference on Information and Knowledge Management* (1998), 306–313.

[10] Orr, K. Data quality and systems theory. *Commun. ACM 41*, 2 (Feb. 1998), 66–71.

[11] Pawar, A., and Mago, V. Calculating the similarity between words and sentences using a lexical database and corpus statistics. *CoRR abs/1802.05667* (2018).

[12] Sravanthi, P., and SRINIVASU, D. B. Semantic similarity between sentences. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (2017), 1–14.