Data Quality Management

Intermediate Report

Aravind RK (220200845)

Kavya Sasikumar(220202446)

Deepthy Paulose(221202854)

Institute for Web Science and Technologies (WeST), University Koblenz, Germany

Abstract. In today's data-centric world, the value of organizations greatly relies on the quality of their data. Data Quality Management (DQM) is a crucial discipline for ensuring reliable and accurate information for effective decision-making. This research explores DQM and innovative strategies to enhance data quality across various domains. It addresses challenges like data integration, cleansing, validation, and governance, aiming to enable organizations to fully utilize their data, make informed choices, optimize operations, and gain a competitive edge in the digital landscape. The study focuses on the impact of data bias and skewness on analytical outcomes by employing exploratory data analysis techniques. It seeks to identify and measure biases and skewness in datasets to enhance data integrity and reduce errors or misinterpretations. Through statistical analysis and visualization, the research proposes a comprehensive framework to accurately assess and address data bias and skewness.

Additionally, the research highlights the evaluation of temporal data changes in time series datasets. Given the prevalence of dynamic data, understanding the reliability of data transformations is crucial for precise forecasting, anomaly detection, and trend analysis. By developing robust evaluation methodologies, the study provides organizations with reliable techniques to measure the quality of temporal data changes, ensuring the integrity and dependability of time series analysis.

Keywords: coefficient of variation (CV), key performance indicators (KPIs), interquartile range (IQR)

1 Introduction

One of the major challenges in data quality management is the presence of data bias and skewness, which can have a substantial influence on analytical results. To solve this issue effectively, the research uses exploratory data analysis approaches to uncover and quantify biases and skewness inside datasets. Organizations can improve data integrity, reduce errors, and eliminate misinterpretations by getting a full grasp of these biases and skewness. To precisely analyze and successfully resolve data bias and skewness, the study provides a complete methodology that blends statistical analysis, visualization, and outlier analytics.

A major part of data quality management is analyzing the quality of changes in time series datasets in addition to addressing data bias and skewness. As dynamic data becomes more ubiquitous, understanding the dependability of data transformations for precise forecasting, anomaly detection, and trend analysis becomes increasingly important. The development of effective evaluation tools for assessing the quality of temporal data changes is a major focus of this research. Organizations may make well-informed decisions based on accurate and reliable information by assuring the integrity and reliability of time series analysis.

Furthermore, the project implements outlier analytics and anomaly detection technologies to find anomalies and outliers in the data. Outliers are data points that differ significantly from the norm, whereas anomalies are patterns in the dataset that are unexpected or abnormal. Organizations can efficiently discover and examine these

outliers and anomalies by using sophisticated analytics approaches, allowing them to unearth important insights, identify data quality issues, and take relevant corrective steps. Integrating outlier analytics and anomaly detection technologies into the data quality management process improves the overall dependability and trustworthiness of the data, boosting organizational decision-making processes.

This research project intends to promote data quality management practices by using exploratory data analysis to address data bias and skewness, building robust evaluation procedures for temporal data changes, and utilizing outlier analytics and anomaly detection technologies. The study's findings have the potential to enable organizations to make educated decisions based on accurate and reliable data, resulting in increased operational efficiency, improved decision-making processes, and a competitive advantage in the digital world.

2 Related work

In the related work section of this research lab, a thorough review of relevant literature was conducted to gain a comprehensive understanding of the existing research in the field of data quality assessment. This involved reading and analyzing various papers authored by experts in the field. The methodology adopted for this section primarily involved a literature review, where a systematic examination and synthesis of previously published studies were conducted. The following papers were chosen based on their relevance to the research topic and their contributions to the understanding and assessment of data quality.

Data Quality Assessment" paper authored by Leo L. Pipino, Yang W. Lee, and Richard Y. Wang discusses the importance of data quality and presents a comprehensive framework for assessing data quality in organizations. The authors highlight the significance of high-quality data for effective decision-making and emphasize the potential negative consequences of poor data quality. The authors also propose a systematic approach that involves defining data quality dimensions, establishing measurement criteria, and conducting data quality assessments. The framework includes both subjective and objective evaluation techniques, considering factors including correctness, completeness, consistency, and timeliness. The document is an excellent resource for organizations looking to assess and enhance the quality of their data assets.

The paper "Data measurement in research information systems: metrics for the evaluation of data quality" by O. Azeroual, G. Saake, and J. Wastl focuses on the evaluation of data quality in research information systems. The authors address the importance of reliable and accurate data in these systems and propose a set of metrics for assessing data quality. The paper further highlights the challenges in ensuring data quality in research information systems, where data comes from various sources and undergoes complex transformations. The authors argue that traditional data quality assessment methods are not sufficient for these systems and propose a tailored approach. The proposed metrics for data quality assessment in research information systems include completeness, correctness, consistency, currency, and conformity. The authors discuss the

relevance of each metric and provide insights into their measurement and interpretation. The paper emphasizes the need for ongoing data quality monitoring and improvement in research information systems. It concludes by suggesting that the proposed metrics can contribute to a comprehensive evaluation of data quality and facilitate the identification of areas requiring improvement.

The paper "Requirements for Data Quality Metrics" by Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz, focuses on identifying the requirements for data quality metrics. The authors explore the characteristics that data quality metrics should possess in order to effectively assess and monitor data quality. The study begins by emphasizing the significance of data quality and the necessity for trustworthy measures to assess it. According to the authors, existing data quality measures frequently lack crucial properties such as expressiveness, comparability, and traceability, limiting their utility. The authors propose a set of requirements for data quality metrics based on an extensive literature review and expert interviews and discuss these requirements in detail, providing insights into their significance They also highlight potential challenges and limitations associated with each requirement. By identifying these requirements, the paper contributes to the development and selection of more effective data quality metrics. It provides guidance for researchers and practitioners in designing and evaluating metrics that can accurately and meaningfully assess the quality of their data.

In the publication titled "Recommendations for Evaluating Mass Spectrometry Data Quality Metrics in Open Access Data (Supplementary to the Amsterdam Principles)," authored by CR Kinsinger, J Apffel, M Baker, and X Bian, the authors discussed the outcomes of the International Workshop on Proteomic Data Quality Metrics which aimed to identify fundamental principles for assessing data quality in mass spectrometry. It emphasizes the significance of addressing data quality through policies, metrics, and collaboration among various proteomics community organizations. It focuses on the initiatives taken by the proteomics community to guaranty data quality and accessibility through the creation of tools for calculating and recording data quality parameters. They emphasized the requirement for comprehensive quality standards and metrics, along with software analytics. Also, the importance of education and training programs to establish reliable protocols in proteomics. As a whole, the paper explores the historical context, key discussions, and future steps required to enhance the quality of open access proteomic data.

The paper “A formal definition of data quality problems” by P Oliveira, F Rodrigues, PR Henriques presents a comprehensive study on Data Quality (DQ) problems, offering a taxonomy that classifies them based on different levels of data organization. The authors distinguish their work by providing precise definitions for each DQ problem and a formal framework that ensures clarity. These definitions include metadata knowledge, mathematical expressions for detection, and potential transformation functions. The authors emphasize the value of this formal framework for automating DQ problem detection. The authors consider this paper as a crucial initial step towards developing an automated tool for detecting DQ problems, as it provides a thorough understanding of the entire range of DQ issues and the necessary computational methods. The goal is to

create a DQ tool architecture based on this knowledge, aiming to improve the limited detection capabilities of existing commercial data profiling tools.

The paper “Visual interactive creation, customization, and analysis of data quality metrics” by C Bors, T Gschwandtner, S Kriglstein introduces MetricDoc, an interactive visual exploration environment for assessing and customizing data quality metrics in tabular datasets. MetricDoc gives analysts the ability to develop and edit metrics, investigate quality problems, and communicate with them. Prototyping, expert evaluations, and a focus group with data quality specialists were all part of the development process. Long-term research will be done in the future to pinpoint essential components and improve scalability. The research advocates including statistical evaluation techniques and connected open data validation to broaden the scope of quality measures. The prototype shows promise in cutting the time required for data profiling across several sources. Future work will focus on creating metrics and tools for multisource data as well as incorporating provenance data to track dataset changes and enable well-informed data quality decisions.

The publication "A Novel Missing Data Imputation Approach for Time Series Air Quality Data Based on Logistic Regression" describes a novel approach dubbed FTLRI for imputing missing values in time series air quality datasets. The authors emphasize that present imputation algorithms do not sufficiently consider data timeliness and frequently favor low-missing-rate datasets over those with larger missing rates. To circumvent these constraints, FTLRI use logistic regression and a "first Five & Last Three" model to capture correlations between characteristics and extract highly relevant data based on both time and attributes linked to missing values. Using hourly concentration data from three distinct stations in Lanzhou in 2019 and varied missing rates, the performance of FTLRI is compared to five conventional baselines and a new dynamic imputation approach. The findings show that FTLRI outperforms the other techniques, with substantial advantages in both short-term and long-term time series air quality data. Notably, FTLRI outperforms other approaches even with datasets with very large missing rates since it concentrates on picking data particularly linked to the missing values rather than depending on all available data. FTLRI, the suggested approach, helps to solve the issues of missing data imputation in time series air quality datasets by considering the timeliness and qualities associated with the missing values.

In the context of data quality management, TC Redman's work "The Impact of Poor Data Quality on the Typical Enterprise" investigates the effects of poor data quality inside organizations. The study identifies several negative consequences of low data quality, such as customer discontent, increased operational expenses, ineffective decision-making, and a diminished capacity to design and implement plans. According to the findings, most organizations have data accuracy levels ranging from 1% to 5% at the field level. These errors have serious operational consequences, resulting in consumer unhappiness, increased expenditure, and lower staff work satisfaction. Furthermore, at the tactical level, low data quality jeopardizes decision-making processes, complicates reengineering efforts, and encourages mistrust among various departments.

The study shows that poor data quality impedes strategy creation, execution, and alignment at the strategic level. It also adds to political tensions and inefficiency in

organizations. These findings highlight the far-reaching effects of poor data quality, which range from operational concerns to tactical obstacles and, ultimately, have an influence on strategic endeavors' findings of this study emphasize the critical necessity of data quality management in organizations. It emphasizes the need to prioritize data correctness, addressing data quality concerns, and putting in place effective data quality management practices. Organizations may improve customer happiness, decrease operational costs, simplify good decision-making, and create a favorable climate for strategy creation and implementation by enhancing data quality.

"Data Quality: A Statistical Perspective" by AF Karr, AP Sanil, provides statistical insights into data quality management. It defines data quality as difficulties with data creation, assembly, and attributes. The study emphasizes the consequences of low data quality, such as difficulties in avoiding terrorist attacks, finding broken car components, financial losses, and diminishing survey response quality. It emphasizes the significance of high-quality data for decision-making, regulatory results, and precise measurements. The article goes through data quality problems, the data quality process, data quality components, the role of human factors and domain expertise, and the application of statistical approaches in data quality management.

Conducting a thorough literature review benefited in understanding collective knowledge and insights generated by previous researchers in the field of data quality assessment. The findings and insights obtained from the selected papers served as valuable input for the development of the research approach as well as in understanding data quality assessment techniques.

3 Approaches

In this research, various approaches are employed to address the challenges related to data quality management and enhance the integrity and reliability of the datasets. The following approaches are utilized:

3.1 Exploratory Data Analysis (EDA)

Exploratory Data Analysis is an important technique for extracting insights and knowledge from data. It entails visually and statistically reviewing the dataset to detect patterns, trends, outliers, and data quality concerns. The research tries to detect biases, skewness, missing values, inconsistencies, and other abnormalities within the dataset using EDA techniques such as data visualization, summary statistics, and correlation analysis. This method allowed us to run a full assessment on data quality and serves as a foundation for future data management efforts.

3.2 Statistical Analysis

Statistical analysis is essential for assessing data quality and detecting biases and skewness in the dataset. Statistical procedures such as hypothesis testing, regression analysis, and descriptive statistics are used in the study to quantify the amount of bias and skewness in the data. These studies gave useful insights into data distribution, highlight substantial departures from predicted patterns, and quantify the influence of biases on analytical findings.

3.3 Temporal Data Analysis

Temporal data analysis is used to assess the quality of temporal data changes in time series datasets. This helped in evaluating the consistency and dependability of data transformations across time, which is critical for accurate forecasting, trend analysis, and anomaly identification. We created comprehensive evaluation procedures for temporal data changes, considering criteria like data timeliness, completeness, consistency, and relevance.

3.4 Outlier Analytics and Anomaly Detection

Outliers and anomalies can have a substantial impact on data quality and analytical results. Outlier analytics and anomaly detection techniques are used in the study to uncover and analyze odd patterns or data items that depart from the norm. The research enabled the discovery and investigation of outliers and anomalies by employing sophisticated analytics approaches such as classification, and anomaly detection algorithms.

When these techniques were combined, we were able to build a complete framework for data quality management. The project intends to improve data integrity, decrease mistakes, and assure trustworthy and accurate information for effective decision-making by using exploratory data analysis, statistical analysis, temporal data analysis, and outlier analytics.

4 Experiments

In this study, experiments were conducted on three distinct time series datasets, namely online retail, demand forecasting, and air quality. Each dataset provides unique insights into different domains and poses specific challenges for time series analysis. The online retail dataset focuses on sales patterns and customer behavior, the demand forecasting dataset involves predicting future product demand, and the air quality dataset aims to forecast pollution levels for environmental monitoring.

4.1 "Online Retail Data Set."

This dataset contains transactional data from an online retail store between 01/12/2010 and 09/12/2011. The dataset consists of approximately 541,909 instances (rows) and eight attributes (columns). The attributes include:

- InvoiceNo: The unique identifier for each transaction.
- StockCode: The unique identifier for each product.
- Description: A brief description of the product.
- Quantity: The number of items purchased in each transaction.
- InvoiceDate: The date and time of the transaction.
- UnitPrice: The unit price of each item.
- CustomerID: The unique identifier for each customer.
- Country: The country where the customer resides.

The dataset can be used for market basket analysis, customer segmentation, and other retail-related analyses. Hence the research question decided to purse was analyzing sales behavior over a time.

To generate insights about the company's performance as well as to predict future sales, a time series analysis using prophet model was done on the online retail sales data. The data preprocessing step includes extracting the date and hour columns from the datetime field, as well as creating a new feature called Total Order Price, which can be calculated by multiplying the number of units sold against the price per unit. Then a time series was plotted which provides an overview of the Total Order Price over time. It can be concluded that the data exhibits variations and is non-stationary, and there appears to be an increasing trend in the data. Hence it can be noted that there are enough missing values, anomalies and other data quality issues that need to be addressed. To gain further understanding the correlation and skewness of the dataset was plotted which showed the attributes having anomalies. The consistency and relevancy score have also been calculated for the same.

4.2 "Daily Demand Forecasting Orders"

This dataset contains daily demand of orders, which was collected during 60 days from a Brazilian logistics company. The dataset consists of 60 instances, twelve predictive attributes and a target that is the total of orders for daily treatment. The attributes include:

- Week_of_the_month: Indicates the week number within the month (ranging from 1 to 5).
- Day_of_the_week_(Monday_to_Friday): Represents the day of the week (from Monday to Friday)
- Non_urgent_order: Number of non-urgent orders received on a particular day.
- Urgent_order: Number of urgent orders received on a particular day.
- Order_type_A: Quantity of orders classified as type A received on a particular day.
- Order_type_B: Quantity of orders classified as type B received on a particular day.

- Order_type_C: Quantity of orders classified as type C received on a particular day.
- Fiscal_sector_orders: Number of orders received from the fiscal sector on a particular day.
- Orders_from_the_traffic_controller_sector: Number of orders received from the traffic controller sector on a particular day.
- Banking_orders_(1): Number of orders received from banking sector category 1 on a particular day.
- Banking_orders_(2): Number of orders received from banking sector category 2 on a particular day.
- Banking_orders_(3): Number of orders received from banking sector category 3 on a particular day.
- Target_(Total_orders): The target attribute, representing the total number of orders received on a particular day for daily treatment.

The dataset can be used to forecast the total number of daily orders for the Brazilian logistics company, based on the given set of predictive attributes. Hence, the research question is how accurately can the total number of daily orders for the Brazilian logistics company be forecasted using the provided dataset and the predictive attributes, and which combination of attributes and forecasting techniques yields the highest level of accuracy in predicting the daily order totals.

The DQ metric used here is consistency, since it ensures that the predictive attributes consistently offer reliable information, enabling the generation of accurate and dependable forecasts for the total number of daily orders in the Brazilian logistics company. For calculating the consistency score, compute the information gain of all predictive attributes. The attribute with highest information gain is taken into account for the calculation of consistency score.Consistency score is calculated with the ratio of attribute with highest information gain (non-urgent order) to the target (total orders).

4.3 Air quality dataset UCI

Data Acquisition: The program downloads the Air Quality UCI dataset as a ZIP file from a specified URL and extracts the CSV file from it.

Data Loading: The program reads the CSV file into a Data Frame, allowing for easy data modification and analysis.

Data Exploration: By presenting the first few rows and producing summary statistics, it offers an overview of the dataset. This stage aids in understanding the structure and initial quality of the data.

Missing Data Detection: The program scans the dataset for missing values, allowing any gaps or discrepancies in the data to be identified.

Data Visualization: The program creates histograms to visualize the distribution of each characteristic, as well as a correlation heatmap to investigate the correlations between various variables. These visualizations help in the identification of trends, relationships, and potential data quality problems.

Skewness Assessment: The program computes the skewness coefficient to analyze the skewness of the data, assisting in the identification of deviations from a normal distribution that may compromise data quality.

Time Series Analysis: The program analyses the dataset using time series analysis, especially visualizing the air quality values over time. This study enables the detection of trends, patterns, and abnormalities in data.

Data Outlier Identification: To visualize the distribution of values, the code picks certain columns of interest and generates box plots. The program then computes z-scores for each data point and finds outliers using a predetermined threshold. The code returns both the number of outliers identified in each column and the total number of outliers discovered.

Data Consistency Assessment: The approach computes the coefficient of variation (CV) for each column of interest to examine data consistency. It calculates the consistency score by subtracting the CV from one, with lower CV values indicating greater consistency. The code computes and displays a consistency score for each column.

Data Bias and Skewed Data Analysis: For each column of interest, the code computes key performance indicators (KPIs) such as mean, median, mode, skewness, range, interquartile range (IQR), variance, and standard deviation. These KPIs reveal information on data bias and the existence of skewed data.

Anomaly Detection in Time Series Data: To detect abnormalities in the given time series data, the code applies the ARIMA (Autoregressive Integrated Moving Average) model. It fits the ARIMA model, produces predictions and residuals, computes the mean and standard deviation of the residuals, and specifies an anomaly detection threshold. The absolute residuals are compared to the threshold to identify anomalies. The code shows the original time series data with the identified anomalies and displays the discovered abnormalities.

Skewed Data Point Detection: The skewness of the residuals derived from the ARIMA model is calculated using the code. It establishes a skewness threshold and discovers skewed data points by comparing the absolute skewness to the threshold. The program populates the anomalies in Data Frame with the identified skewed data points and presents the findings.

5 References

1. Sidi, F., Shariat Panahy, P. H., Affendey, L. S., Jabar, M. A., Ibrahim, H., & Mustapha, A. (2012). Data Quality: A Survey of Data Quality Dimensions. *2012 International Conference on Information Retrieval & Knowledge Management*. https://doi.org/10.1109/infrkm.2012.6204995
2. Strong, D. M., Lee, Y. W., & Wang, R. Y. (1997). Data quality in context. *Communications of the ACM*, *40*(5), 103–110. https://doi.org/10.1145/253769.253804
3. Redman, T. C. (1998). The impact of poor data quality on the typical enterprise. *Communications of the ACM*, *41*(2), 79–82. https://doi.org/10.1145/269012.269025
4. Karr, A. F., Sanil, A. P., & Banks, D. L. (2006). Data Quality: A statistical perspective. *Statistical Methodology*, *3*(2), 137–173. https://doi.org/10.1016/j.stamet.2005.08.005

5. Heinrich, B., Hristova, D., Klier, M., Schiller, A., & Szubartowicz, M. (2017). Requirements for Data Quality Metrics. *Journal of Data and Information Quality*, *9*(2), 1–32. https://doi.org/10.1145/3148238
6. Liu, J., Li, J., Li, W., & Wu, J. (2016). Rethinking big data: A review on the data quality and usage issues. *ISPRS Journal of Photogrammetry and Remote Sensing*, *115*, 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006
7. Azeroual, O., Saake, G., & Wastl, J. (2018). Data Measurement in research information systems: Metrics for the evaluation of data quality. *Scientometrics*, *115*(3), 1271–1290. https://doi.org/10.1007/s11192-018-2735-5
8. Pipino, L. L., Lee, Y. W., & Wang, R. Y. (2002). Data Quality Assessment. *Communications of the ACM*, *45*(4ve). https://doi.org/10.1145/505999.506010
9. Kinsinger, C. R., Apffel, J., Baker, M., Bian, X., Borchers, C. H., Bradshaw, R., Brusniak, M.-Y., Chan, D. W., Deutsch, E. W., Domon, B., Gorman, J., Grimm, R., Hancock, W., Hermjakob, H., Horn, D., Hunter, C., Kolar, P., Kraus, H.-J., Langen, H., … Rodriguez, H. (2011). Recommendations for mass spectrometry data quality metrics for open access data (corollary to the Amsterdam principles). *Molecular & Cellular Proteomics*, *10*(12). https://doi.org/10.1074/mcp.o111.015446
10. Oliveira, P., Rodrigues, F., & Henriques, P. (2009). Smartclean: An incremental data cleaning tool. *2009 Ninth International Conference on Quality Software*. https://doi.org/10.1109/qsic.2009.67
11. Pipino, L. L., Wang, R. Y., Funk, J. D., & Lee, Y. W. (2006). *Journey to Data Quality*. https://doi.org/10.7551/mitpress/4037.001.0001
12. Bors, C., Gschwandtner, T., Kriglstein, S., Miksch, S., & Pohl, M. (2018). Visual interactive creation, customization, and analysis of Data Quality Metrics. *Journal of Data and Information Quality*, *10*(1), 1–26. https://doi.org/10.1145/3190578