IRCLASS
Indian Register of Shipping

GUIDELINES ON
DATA QUALITY ASSESSMENT

SEPTEMBER 2025

Guidelines

Data Quality Assessment

September 2025

Contents

Sections

Abbreviations

DQ	:	Data Quality
ISO	:	International Organization for Standardization
DAMA	:	Data Management Association
IoT	:	Internet of things
DSM	:	Data Stewardship Manager
BSG	:	Business Semantic Glossary
RDA	:	Reference Data Accelerator

Section 1

General

1.1 Scope

1.1.1 These Guidelines are intended to provide an outline of data quality standards, general guidance and recommendation on assessment of data quality for marine data.

1.2 Application

1.2.1 This document outlines the process for assessing and improving data quality. It introduces a framework for evaluating data quality and provides practical recommendations for ensuring high-quality data based on established metrics and criteria.

1.2.2 These Guidelines focus on marine applications and focusses primarily on how to apply data quality standards to time-series data generated by onboard sensors, measurement instruments, and automation systems.

1.2.3 These guidelines are applicable to the following:

(a) Systems (software, hardware and associated equipment) requiring IoT data to develop data-driven applications with diverse functionalities, such as machinery health monitoring, hull structural monitoring, and optimizing operational performance.

(b) Condition monitoring software which is used to provide a great understanding of equipment condition, enabling condition-based replacement and maintenance strategies, and enhancing maintenance efficiency.

1.3 Definitions

1.3.1 ***Data***: Representation of information in a formalized manner which should be suitable for interpretation and decision making.

1.3.2 ***Data life cycle:*** The stages through which data progresses, including generation, acquisition, processing, storage, usage in a designed data utilization process to create value, and eventual disposal.

1.3.3 ***Data Quality***: The extent to which a set of characteristics of data fulfils requirements in ISO 8000 series of standards.

1.3.4 ***Data Quality dimensions***: The dimensions represent the views, criteria, or measurement attributes for data quality problems that can be assessed, interpreted, and possibly improved individually. By assigning scores to these dimensions, the overall data quality can be determined as an aggregated value of individual dimensions relevant in the given application context.

1.3.5 ***Dataset***: A collection or grouping of records.

1.3.6 ***Internet of things (IoT):*** The Internet of Things (IoT) describes the network of physical objects—"things"—that are embedded with sensors, software, and other technologies for the purpose of connecting and exchanging data with other devices and systems over the internet. These devices range from ordinary household objects to sophisticated industrial tools.

1.3.7 ***Metadata***: Data that describes about other data

1.3.8 ***Time series data***: Sequence of data in the order of time

1.3.9 ***Organization:*** person or group of people that has its own functions with responsibilities, authorities and relationships to achieve its objectives.

Section 2

Overview of Data Quality

2.1 General

2.1.1 In the modern digitized world, shipping operations are increasingly dependent on information systems for control and analysis. Organizations, whether traditionally operational or digitally driven, are shifting towards data and analytics-centric models. This transition blurs the lines between conventional and digital business operations.

2.1.2 Data is now considered as an asset and significant costs are involved in collecting, storing, and performing actions based on the same. Data value chains are prevalent in various industries, where data is created, collected, refined, and utilized for diverse tasks. However, users or systems may not always be aware of the origin of the data, its quality, its limitations, legal or contractual obligations, or the context in which the data was collected.

2.1.3 To ensure reliable operations and valid analytics, it is imperative that data quality is assessed and continuously monitored across all critical systems and services. High-quality data enables the reuse of data and supports analytics, particularly when utilizing historical data.

2.1.4 Organizations should establish data quality policies and implement processes to support these policies. Clearly defined data quality requirements and metrics are essential for verifying compliance and optimizing data management practices. These measures should be integrated throughout the Organization to ensure comprehensive data quality management.

2.1.5 The framework outlined in this guidelines consists of three main parts: data quality measurement, data quality management, and data quality risk assessment. Each part is described in separate subsections below.

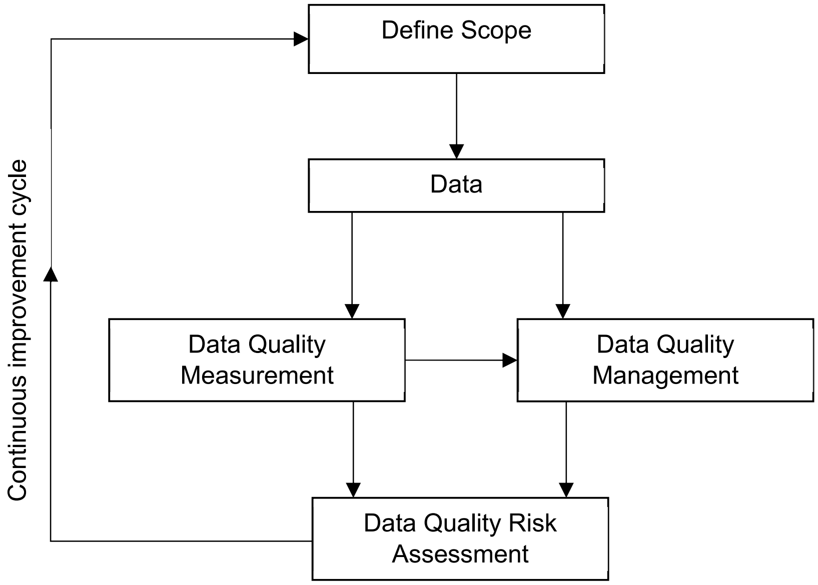


Figure 2.1: Data Quality assessment and improvement process

2.2 Processes and Dimensions

This sub-section describes the processes and data quality dimensions involved with each process. These dimensions serve as a comprehensive framework for evaluating various aspects of data to ensure it aligns with operational needs. A lot of dimensions have been proposed in various fields but there are no unified set of dimensions. Appropriate dimensions should be selected to control or improve the data quality, considering the following:

- Relevance to Data
- Alignment with Objectives
- Cost-effectiveness
- Measurability and Monitoring

In 2013, DAMA UK identified six primary dimensions: Accuracy, Completeness, Consistency, Timeliness, Uniqueness, and Validity. The number of dimensions an organization uses can vary, typically ranging between 5 and 20.

The key dimensions which are to be considered for data quality assessment are indicated in Fig. 2.2. These dimensions are integral to ensuring data quality throughout its lifecycle.

Below, we highlighted the key dimensions relevant to the maritime domain. (See Appendix 1 for details.)

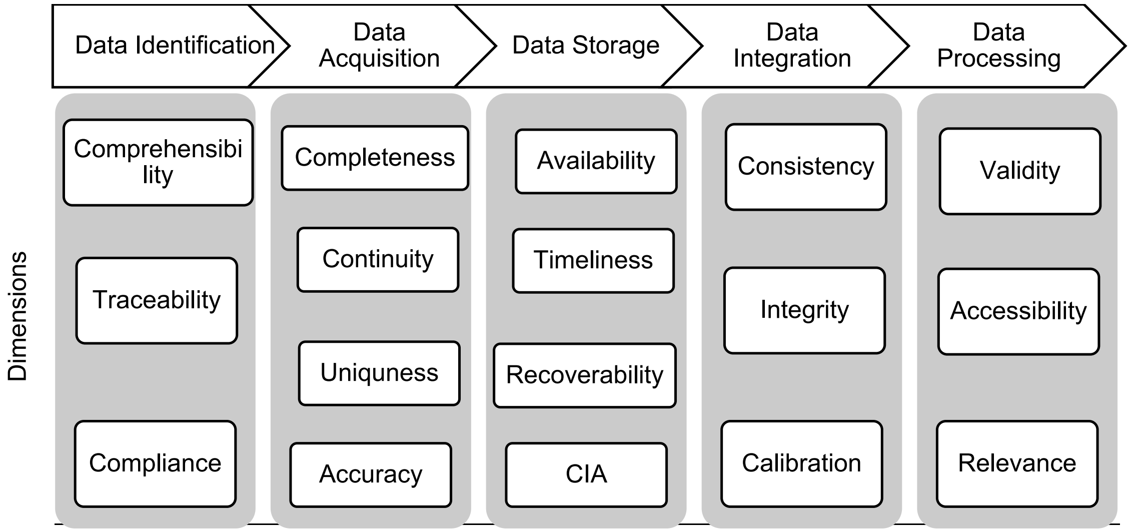


Figure 2.2: Data Quality Assessment – Key elements

2.2.1 Data Identification

2.2.1.1 Data identification refers to the process of recognizing, labelling, and classifying data elements based on their properties and purpose/application within a system. This process ensures that each data element is identifiable, traceable back to its origin, and can be categorized appropriately for effective management and security. Proper identification of data ensures that it can be effectively managed, integrated, and utilized for analysis or operational purposes. A critical aspect of data identification is locating where the Organization’s sensitive data resides, including cloud repositories or physical hard drives. Key dimensions for measuring effectiveness of data identification include the following:

- **Comprehensibility**: Data should be clear and interpretable, ensuring it is understandable by all relevant users.
- **Traceability**: Data should be traceable to its origin, allowing users to verify its source and authenticity.
- **Compliance**: Adherence to standards, conventions, or rules, ensuring correct usage of identifiers and codes.

2.2.2 Data Acquisition

2.2.2.1 Data acquisition refers to the process of gathering and capturing data from various sources into a system. Ensuring completeness, uniqueness, continuity and accuracy of the data is crucial for its effective use. The following can be used to measure effectiveness of data acquisition:

- **Completeness**: It is crucial that all expected attributes and entities are captured during data acquisition. Missing attributes in the data can hinder decision-making processes.
- **Uniqueness**: Each data element should be distinct, preventing duplication and ensuring accurate identification
- **Continuity**: Data should be collected at a frequency that meets the operational and regulatory requirements, ensuring consistency over time.
- **Accuracy**: Accuracy measures how well the data reflects the real-world objects or events it represents. It is crucial during data acquisition to ensure that the collected data is accurate.

2.2.3 Data Storage

2.2.3.1 Effective data storage practices will ensure that data remains accessible to authorized users, retained in accordance with policy, and properly archived for long-term preservation while maintaining its quality. Data storage systems should ensure that data is available, secure, and easily retrievable. The effectiveness of data storage can be measured using the following-:

- **Availability**: Availability refers to the ease of access to stored data when it's needed.
- **Timeliness**: Timeliness refers to the availability of data when needed. Data should be collected and made available promptly to support timely decision-making.
- **Recoverability**: It ensures that data is securely backed up and capable of being recovered in case of data loss.
- **Confidentiality, Integrity, and Availability (CIA)**: Ensures the security of data throughout its lifecycle, protecting it from unauthorized access, corruption, or loss.

2.2.4 Data Integration

2.2.4.1 Data integration involves the process of combining data from multiple sources into a unified dataset, ensuring consistency and comparability across these sources. This process also includes the elimination of errors, inconsistencies, and irrelevant information to uphold data quality. The effectiveness of data integration can be measured using the following:

- **Consistency**: Consistency refers to the uniformity of data across different systems or datasets. When integrating data from multiple sources, consistency ensures that the data does not conflict with each other.
- **Integrity**: Integrity focuses on reliability of the data throughout the integration process ensuring that it remains intact and uncorrupted while adhering to defined rules and constraints.
- **Calibration**: Calibration involves alignment and adjustment of data from various sources to ensure consistency and comparability. Calibration may include standardizing measurements and formatting to achieve uniformity and reliability across different datasets.

2.2.5 Data Processing

2.2.5.1 Once data is acquired and integrated, it is essential to process it accurately to generate meaningful insights. The goal of data processing is to ensure that raw data is transformed into a format that is suitable for decision-making and operational efficiency. Proper management of this stage is essential to maintain its relevance and applicability. The effectiveness of data processing can be measured using the following:

- **Validity**: Validity refers to whether the data conforms to defined business rules or constraints. During processing, data should be validated to ensure it complies with predefined criteria.
- **Accessibility**: Processed data should be readily accessible to users and systems without additional rework.
- **Relevance**: Ensures that the data processed is meaningful and serves its intended purpose.

Section 3

Data Quality Measurement

3.1 Introduction

3.1.1 The metrics and dimensions used for data quality measurement should provide a coherent and complete evaluation framework for data quality.

3.1.2 Data quality measurement involves a systematic approach identifying relevant data quality dimensions, establishing metrics for assessment, and applying both qualitative and quantitative methods to prepare data for assessment.

3.1.3 The process emphasizes continuous monitoring and improvement, enabling organizations to address inconsistencies, anomalies, and gaps effectively. A structured framework ensures data remains fit for purpose and aligned with organizational objectives.

3.2 Data Quality Measurement

3.2.1 ISO 8000-8:2015 provides fundamental concepts to plan and perform data quality measurements. Its application is independent of status of organization, type of information or data, hardware storage medium, software, information security and information life cycle stage.

3.2.2 The main purpose of ISO 8000-8 is to provide a foundation for measuring data quality according to the following categories:

a) *Syntactic quality* – the degree to which data conforms to its specified syntax, i.e. requirements stated by the metadata;

b) *Semantic quality* – the degree to which data corresponds to what it represents;

c) *Pragmatic quality* – the degree to which data is found suitable and worthwhile for a particular purpose.

3.2.3 Measuring syntactic and semantic quality is performed through a verification process, while measuring pragmatic quality is performed through a validation process. (See section 3.3)

a) Verification is the evaluation of whether or not a product, service, or system complies with a regulation, requirement, specification, or imposed condition.

b) Validation is the assurance that a product, service, or system meets the needs of the customer and other identified stakeholders. It often involves acceptance and suitability with external customers.

3.2.4 ISO 8000-8 also provides prerequisites for measuring data quality, and it specifically describes requirements along with a set of data quality rules/dimensions for data quality verification and validation relevant to each of the three data quality categories.

a) Syntactic quality is measured as conformance with the schema, metadata, and any defined business requirement definitions, rules, and vocabulary. Syntactic verification is normally fully automated and the entire data set is considered for assessment.

b) Semantic data quality measurements verify conformance between the data and the entity or object that the data represents. Semantic verification is normally performed by statistical sampling methods in order to achieve the desired percentile.

c) Pragmatic data quality validation represents users' perceptions of whether the data is fit for use and is normally measured by user groups, questionnaires, and feedback.

3.3 Verification and Validation

3.3.1 Verification and validation are essential processes in ensuring the quality, accuracy, and reliability of data. These methods help confirm that the data meets the required standards and is suitable for its intended purpose.

3.3.2 Adopting robust verification and validation techniques can mitigate risks associated with poor data quality, thereby enhancing the overall efficiency and safety of maritime operations. (Refer Appendix 2 for verification and validation)

3.4 Data Profiling

3.4.1 Data profiling is a technique of data analysis used to inspect data and assess quality. Data profiling uses statistical techniques to detect the true structure, content, and data quality issues in data sets with little or no metadata available. It produces data statistics that enable the data analysts to understand the data patterns and data quality issues such as illegal values, outliers, and other anomalies. Data profiling is often used in the early stages of data activities to assess the current state of data that is targeted for improvement. As data quality requirements develop and the project progresses to higher capability levels, the data quality assessments often shift to more formal framework, such as those outlined in ISO 8000-8.

3.4.2 The insights gained from data profiling highlights opportunities to enhance the quality of both data and metadata. However, the revealed data quality issues require data consumers and subject matter experts to interpret/identify the root cause and evaluate the potential impact.

3.4.3 Data profiling techniques

3.4.3.1 Several statistical and counting techniques are used to profile datasets, both for single datasets and for connected datasets e.g. cross-column analysis for single datasets and inter-table analysis for connected datasets.

3.4.3.2 Cross-column analysis

3.4.3.2.1 Cross-column analysis is applied within a single dataset to identify overlapping or duplicate columns and uncover embedded value dependencies. This type of analysis utilizes various statistical and counting techniques to provide detailed insights into the structure and relationships between columns in the dataset. Table 3.4.4.2 lists five main types of analyses performed by data profiling tools commonly employed for single datasets (Refer Appendix 3 for data profiling tools).

Table 3.4.3.2.1 : Cross-column analysis	
Uniqueness Analysis and Completeness Analysis	
Uniqueness	Identifies the percentage of the unique values for individual columns
Completeness	Identifies the percentage of rows that contain actual values for individual columns
Counts of Null	Identifies the number of rows that do not contain values (null) for individual columns
Pattern Analysis	
Data Type	Identifies the syntactic patterns of the data (e.g., the code “W” means a word and “N” means a number) and the total number of each pattern counts for individual columns based on the value contents
Data Format	Identifies the syntactic formats of the data (e.g., the code “L” means a letter and “D” means a digit) and the total number of each format counts for individual columns based on the value contents
Precision	Identifies the largest number of digits in a number for individual columns, which includes digits on both sides of the decimal point (e.g., the precision for “DDD.DD” is 5)
Scale	Identifies the greatest number of digits to the right of the decimal point for individual columns (e.g., the scale for “DDD.DD” is 2)
Min/Max length	Identifies the overall minimum and maximum lengths of values for individual columns and detects outliers based on significant deviations in length.
Range analysis	
Min/Max Value	Identifies the overall Min / Max value for individual columns across all data types
Value distribution analysis	
Frequency Distribution	Identifies how many times each value in the selected column occurs
Quantile Distribution	Identifies the data values in the selected column that occur at designated intervals in the ordered data set. The first value in the list is at 0% and the last value is at 100%. The median value is at 50%

3.4.3.3 Inter-table analysis

3.4.3.3.1 Inter-table analysis explores overlapping values sets and helps identify foreign key relationships between tables. When key definitions, such as primary and foreign keys, are available, inter-table validations can be performed to ensure reference integrity between datasets. These validations allow analysts to compare records across tables, ensuring that data in one table matches appropriately with data in another.
The criteria in Table 3.4.4.3 can be used for evaluation of referential integrity across tables/ datasets A and B. This helps in assessing the consistency and completeness of the data across both datasets.

Table 3.4.3.3: Inter-table analysis		
Number of records in A	Percentage of records in A also found in B	Number of records in B found in A
Number of records in A not in B	Number of records in B	Percentage of records in B found in A
Percentage of records in A not in B	Number of records in B, but not in A	Number of records matchings
Number of records in A also found in B	Percentage of records in B but not in A	

A practical illustration using Jupyter Notebook for both cross-column analysis and inter-table analysis are provided in Appendix 4.

Section 4

Data Quality Management

4.1 General

4.1.1 Data quality management is a set of practices aimed at improving and maintaining the quality of data. It involves the planning, implementation, and control of activities that apply quality management techniques to data, in order to assure it is fit for consumption and meets the needs of data consumers.

4.1.2 A common approach to improving data quality is based on the Shewhart/Deming cycle, also known as the 'plan-do-check-act' model. This scientific method-driven cycle provides a structured process for continuous improvement.

4.1.3 Data conditions are measured against defined standards, and if discrepancies are found, the root causes—whether technical or non-technical—are identified and remedied. Once the issues are addressed, ongoing monitoring ensures that data continues to comply with the required standards.

4.1.4 ISO 8000-61 outlines a foundational framework for organizational data quality management, comprising three main components: the Implementation Component, the Data-Related Support Component, and the Resource Provision Component, as illustrated in Figure 4.1.

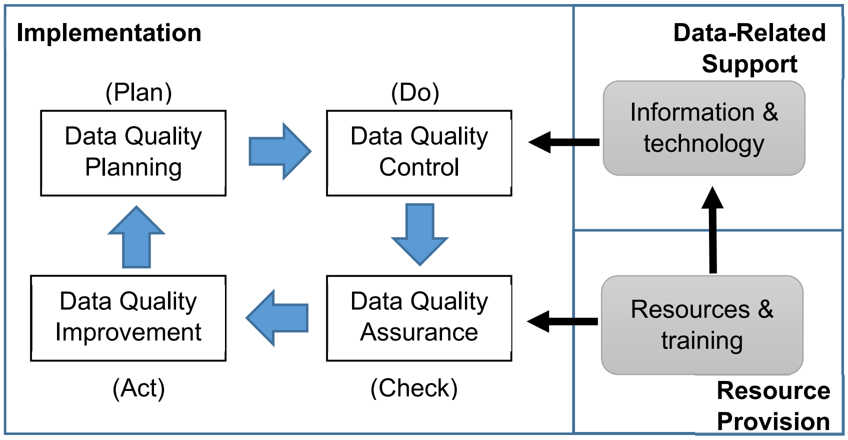


Figure 4.1: Data Quality Structure - ISO 8000-61

4.2 Key Steps in Data Quality Management

4.2.1 To maintain data quality throughout the data lifecycle, organizations should integrate the following practices into their operations:

Step 1: Plan (Data Quality Planning)

This step is very crucial in determining the scope, requirements, strategy, policies, standards, procedures and the steps involved in implementing these within the organization. It is essential that we take these into consideration and improve them as we proceed further in the cycle.
In the "Plan" step, data quality requirements are defined based on identified data quality issues through data profiling. A basic data quality improvement plan should be prepared to address these issues. The main tasks include:

A. Preliminary Data Quality Assessment

I. Select a data set for initial assessment (e.g., a small or a full-scale dataset from historical data sets generated by a similar data application) by using statistical analysis approaches (e.g., data profiling and similar techniques) to identity the data characteristics (e.g., data type, format, precision, range, etc.), discover potential data quality issues and define the data quality validation rules.

II. Identify potential data quality issues or anomalies through data profiling of the identified data set, as well as the potential data quality issues like data collection, transmission and storage, with the help of subject matter experts.

III. Evaluate potential impacts of the identified data quality issues on the potential data use, which requires input from stakeholders along the data chain from both technical and business aspects.

B. Define Data Quality Requirements

Two approaches for identifying data quality requirements are commonly adopted in this step, which are described in detail in section 4.3

I. Define data quality requirements based on the identified data quality issues and their potential impacts. The requirement definition includes data quality validation rules, measurable metrics and dimensions. (Refer to Appendix 2)

II. Define the thresholds, weighting factors and acceptance level for each data quality assessment level: metrics, dimension and overall consolidation.

C. Assess Data Quality

I. Test and validate the defined data quality validation rules by applying them against a test dataset other than the full-scale data sets used for the initial assessment in preliminary data quality assessment.

II. Review the defined data quality validation rules with the users to make sure that they understand them

III. Evaluate data quality levels based on the data quality validation rules and the defined acceptability and thresholds.

D. Prioritize Improvements

I. Prioritize data quality issues based on their impact on data use and evaluate improvement alternatives to address data quality issues.

II. Prioritize the remediation and improvement efforts, which requires a combination of a full-scale data profiling and inputs from relevant stakeholders

III. Identify data quality issues requiring in-depth analysis for root cause determination and potential improvement alternatives to overcome the issues.

Step 2: Do (Data Quality Control)

In the "Deploy" step, data quality requirements defined in Step 1 are configured and managed during operational data quality measurements. After a monitoring period, an operational remediation plan should be developed.

It includes provision of well-documented data specifications and work instructions. The actual data processing, and the monitoring and control of these processes. These activities contribute to ensuring the information product being produced is of high quality.

A. Manage Validation Rules and standards

I. Rules should be documented with a consistent format with a clear description.

II. Rules should be defined in terms of measurable data quality metrics and dimensions.

III. Rules should be created with consideration of data use. The defined data quality validation rules need to be tested against actual data sets as identified in Subsection 4.2/Step1/C. Continuous refinement of the validation rules is recommended throughout the data quality improvement lifecycle.

IV. Data consumers and subject matter experts should be involved in defining the data quality validation rules. The defined rules should be confirmed by the data consumers and subject matter experts

B. Develop procedure for Handling Data Issues

I. Diagnose data quality issues: Review the data quality problem as identified in subsection 4.2/Step 1 and discover the potential root causes of the problem with assistance from data consumers and subject matter experts.

II. Identify options for addressing data quality issues

Address non-technical root causes:

a) possibly provide proper training to data handlers;
b) improve the data handling procedure;
c) enhance leadership support;
d) establish clear accountability and ownership

Address technical root causes:

a) Need to correct flawed data directly;
b) Improve the performance of data collection (e.g. sensors), transmission and storage;
c) Modify systems and technical processes to prevent the issue from recurring;
d) Continuous monitoring and taking no immediate actions after balancing the impact of the data quality issues versus the cost of the corrective/ improvement actions

III. Resolve data quality issues.

a) Perform cost-benefit analysis to compare the potential correction options as identified in Subsection 4.2/Step2). Positive return on investment (ROI) for improvements should be achieved.

b) Give advice from the data consumers and subject matter experts to select the best option to resolve the issue:

- Simple remediation: Fixing and correcting the data directly in records (e.g., data cleansing/ data parsing and formatting)
- Remediation of root causes: Formulating a long-term improvement plan for strategic changes (e.g., modification of the systems). It focuses on modifying the systems to resolve root causes and putting in place mechanisms to prevent issues in the first place. Prevention is generally more cost saving than correction.

c) Develop and implement a remediation plan which intends to re-evaluate the quality level of the remediated data set and to ensure the applied changes do not introduce additional errors, and perform as expected.

Step 3: Check (Data Quality Assurance)

With the aim of process improvement, this includes a review of data quality issues, provisioning the measurement criteria, the actual measurement of data quality and process performance, and evaluating the meaning of measurement results. Following steps can be used for continuous monitoring of data quality:

I. Set a time interval for periodically assessing the data quality against the defined rules

II. Visualize and monitor the data quality results/scores in hierarchical means (e.g., metrics, dimensions and overall levels)

III. Apply a threshold(s) or acceptance criterion for each measurement. The data quality results often reflect the percentage of correct data (passing the validation rule) or the percentage of exceptions (failing the validation rule) depending on the formula used.
   a) Confirm that the data is fit for its intended application (e.g., data analytics) if the data conforms to the defined data quality validation rules
   b) Notify and alarm data quality issues timely and recommend potential actions according to the developed remediation plan when the data does not conform to the defined data quality validation rules

IV. Confirming data fitness for intended applications and promptly addressing emerging issues.

V. Monitor and trend the data's ongoing conformance with validation rules, and report on all data quality assessment levels
   a) Data quality scorecard provides data quality scores in metrics, dimensions, overall level and dashboards related to the execution of data quality validation rules
   b) Data quality trend shows the data quality changes over time
   c) Data quality issue management tracks the data quality issue handling according to the remediation plan

Step 4: Act (Data Quality Improvement)

This step focuses on the improvement of processes and data quality. Continuous improvement is achieved by root cause analysis (to prevent future errors), and data cleansing (to repair existing errors) which may result in restarting the data quality improvement cycle. Following situations can be considered for data quality improvement whenever:

a) Existing data quality results fall below acceptance criteria.
b) New data sets come under investigation.
c) New data quality requirements are identified for existing datasets.

4.3 Approaches for Data Quality Requirements Identification

4.3.1 Data quality requirements analysis typically involves gathering input from data users and subject matter experts through surveys to uncover data quality issues. This process helps identify critical data sets, establish relevant data quality rules, metrics, and dimensions, and set quality targets accordingly.

4.3.2 The two approaches that can be used to define data quality requirements are top-down and bottom-up.

a) Top-down

   In the top-down approach, the pre-defined requirements of the data quality are initialized by the data consumers who are knowledgeable about the subject data issues. The requirements are converted into data quality validation rules and organized into metrics and dimensions. The data quality assessment and monitoring are executed based on the defined requirements.

   This is a data consumer-driven approach, requiring the data consumer to be familiar with the subject data sets and knowledgeable of the features of the intended data application, the critical data dependencies, and the potential data issues that are significant to the success of such applications.

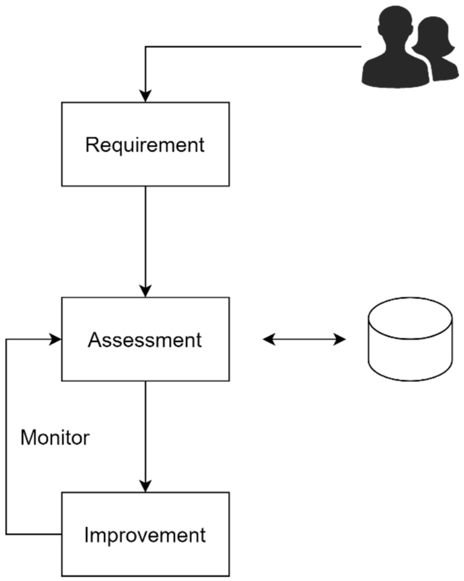


Figure 4.3.2: Top-down approach

b) Bottom-Up Approach
In the bottom-up approach, the data is initially assessed through data exploration or data profiling to uncover any data anomalies, the requirements on data quality are continuously reviewed and updated according to the potential issues revealed. The data quality is then measured and monitored against the requirements identified from the previous step. If the quality level is below acceptable levels, it should trigger actions to improve the data. The defined requirements should be reviewed and refined accordingly when new data sets come under investigation.

This is a data-driven approach, which allows requirements to be dynamically discovered and adapted as the use and understanding of the data set expands, which is suitable for assessing large and unfamiliar data sets.

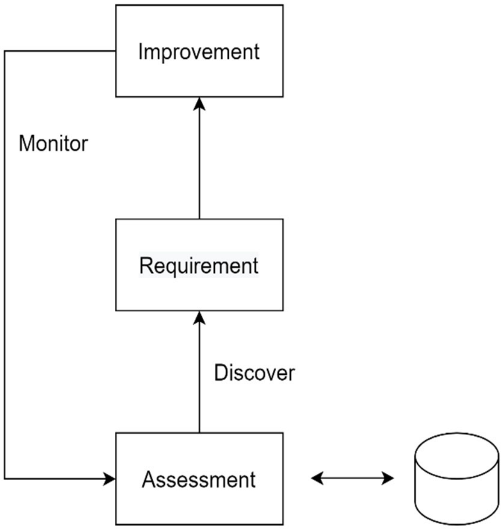


Figure 4.3.2: Bottom-down approach

Section 5

Data Quality Risk Assessment

5.1. General

5.1.1 This guideline provides a general framework for assessing data quality risk assessments. The specific steps and methods should be adjusted to suit the nature of the data and its application. The risk assessment process may vary depending on factors such as real-time data integrity, network reliability for IoT-generated data streams, or data accuracy and consistency in digital models.

5.1.2 The consequences and business impacts of data quality issues on different uses and contexts of data are evaluated by common risk management frameworks, such as those detailed in ISO 31000:2018.

5.1.3 The risks associated with data quality are to be identified, evaluated, and communicated to stakeholders.

5.1.4 Data quality risk assessment can help the organization save unnecessary costs, increase customer retention, and improve work efficiency. All these factors are critical for building a sustainable and growing business model.

5.1.5 The organization should monitor and mitigate data risks through data acquisition, storage, integration, and processing. Figure 5.1 illustrates the process of risk assessment and contains the following sequence of tasks:

a) establish the context and scales for risk assessment
b) perform risk identification
c) analyse risk and figure out potential threats
d) evaluate the risk and rank risks according to the criticality of consequences
e) identify, evaluate, and prioritize treatment and mitigating actions
f) monitor the risk scope, changes, incidents, and assessment process continuously
g) review to improve the efficiency of the process and repeat based on needs or review criteria
h) Communicate with all stakeholders and consult experts at regular intervals.

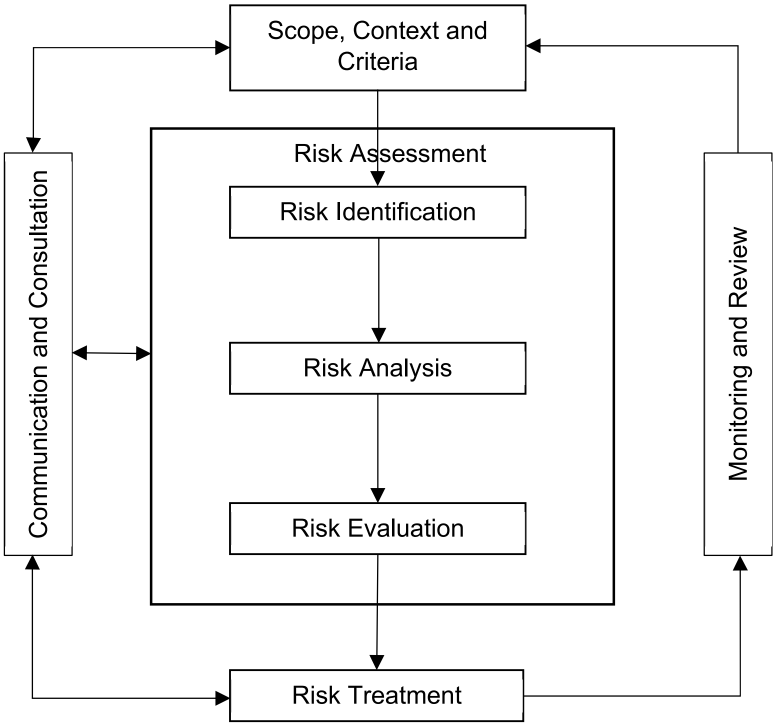


Figure 5.1: Risk management process based on ISO 31000:2018

5.2 Scope, Context and Criteria

5.2.1 Establishing the context and scope design is very important. A narrow scope can result in assessment fragmentation and major threats or vulnerabilities can be outside the defined scope. This could result in a situation where major risks fall outside the scope, overlapping or similar risks could occur in disparate scopes with slightly different wordings, or a major risk could be listed as several non-critical risks.

5.3 Risk Assessment

Risk assessment is the overall process of risk identification, risk analysis and risk evaluation. Risk assessment should be conducted systematically, iteratively and collaboratively, drawing on the knowledge and views of stakeholders.
Follow these steps for a comprehensive data quality risk assessment:

5.3.1 Risk Identification

5.3.1.1 The first step in managing data quality risks involves identifying potential risks that may impact the integrity, accuracy, and availability of data.

5.3.1.2 The purpose of risk identification is to find, recognize and describe risks that might help or prevent an organization achieving its objectives. Relevant, appropriate and up-to-date information is important in identifying risks.

5.3.1.3 The organization can use a range of techniques for identifying uncertainties that may affect one or more objectives. Common risks associated with data quality include metadata inconsistency, missing data, inaccurate data, and duplicate records.

5.3.2 Risk Analysis

5.3.2.1 Once risks have been identified, risk analysis is conducted to estimate the likelihood and severity of these risks. Risk analysis helps organizations understand which risks are most likely to materialize and how severe their consequences might be. It provides a basis for prioritizing risks and deciding which ones require immediate attention.

5.3.2.2 Figure 5.3.2 presents an example of data quality risk, illustrating the impact of risks. For instance, if a risk is evaluated as high, mitigating actions can be implemented to control the data quality over time.

5.3.2.3 The scale used for evaluating consequences or business impacts will vary based on the organization and is typically evaluated from low to high.

5.3.2.4 Analysis techniques can be qualitative or quantitative, depending on the circumstances and intended use. Risk analysis should consider factors such as:

a) the likelihood of events and consequences;
b) the nature and magnitude of consequences;
c) complexity and connectivity;
d) time-related factors and volatility;
e) the effectiveness of existing controls;
f) Sensitivity and confidence levels.

Qualitative and Quantitative Risk Analysis:

a) **Qualitative Analysis:** Risks are categorized into levels (e.g., high, medium, low) based on their likelihood and impact. This method is useful when precise numerical data is unavailable but requires careful interpretation of expert input.
b) **Quantitative Analysis**: Risks are analyzed using numerical data to calculate the likelihood of occurrence and the severity of impact. Quantitative risk estimation is often more objective and data-driven but may require more detailed data sets.

Examples:

a) **Transmission Errors:** Evaluate the likelihood of data corruption during transmission and its potential impact on operational decision-making processes.
b) **Sensor Inaccuracies:** Estimate the probability that sensor malfunctions will lead to erroneous data, potentially affecting critical systems.

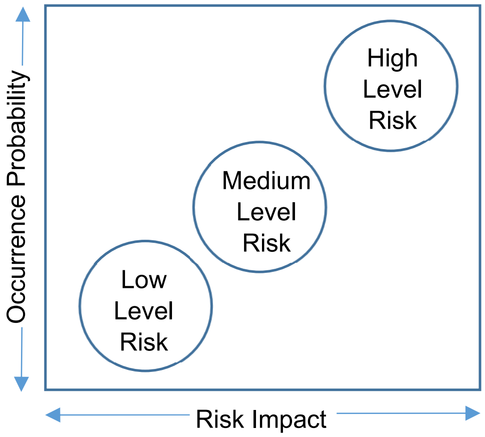


Figure 5.3.2: Risk impact

5.3.3 Risk Evaluation

5.3.3.1 The purpose of risk evaluation is to support decision-making by comparing the results of the risk analysis with the established risk criteria to determine where further action is required. Based on this, decision can be made to:

a) Take no further action;
b) consider risk treatment options;
c) undertake further analysis to better understand the risk;
d) maintain existing controls;
e) Reconsider objectives

5.3.3.2 Decisions should take account of the wider context and the actual and perceived consequences to external and internal stakeholders. The outcome of risk evaluation should be recorded, communicated and then validated at appropriate levels of the organization.

Key Aspects of Risk Evaluation:

a) **Comparison with Risk Criteria:** Risks are evaluated based on factors like severity, likelihood, and detectability, enabling effective prioritization.
b) **Risk Ranking:** Risks are ranked using a "risk score" to determine which risks should be addressed first, based on their criticality.

Examples:

a) Determine whether the risk of faulty sensor data affecting operations is within acceptable levels or needs corrective action.
b) Rank risks such as unauthorized access to data versus operational risks from incomplete data sets.

5.4 Risk Treatment

Once risks have been identified, analyzed, and evaluated, mitigation strategies are formulated and implemented to reduce their impact on data quality. These strategies should align with the organization’s data quality objectives and available resources.

5.4.1 Selection of Risk Treatment Options

5.4.1.1 Choosing the most suitable risk treatment options requires balancing the potential benefits of achieving objectives against the associated costs, effort, or downsides of implementation. These options may not be mutually exclusive and can vary depending on the situation. Options for treating risk may involve one or more of the following:

a) **Risk Avoidance**: Discontinue or avoid activities that generate the risk.
b) **Risk Acceptance**: Intentionally take or increase the risk to pursue an opportunity.
c) **Risk Removal**: Eliminate the source of the risk.
d) **Likelihood Modification:** Reduce or alter the likelihood of the risk occurring.
e) **Consequence Modification**: Minimize or change the potential consequences of the risk.
f) **Risk Sharing:** Distribute the risk through contracts, or other mechanisms.
g) **Risk Retention:** Retain the risk after making an informed decision.

5.4.1.2 The decision to adopt a particular treatment method should go beyond economic factors and should consider the organization’s commitments, obligations, and the views of stakeholders. This decision-making process should align with the organization's objectives, risk criteria, and available resources.

5.4.1.3 When selecting risk treatment options, the organization should consider the perceptions and involvement of stakeholders and ensure proper communication. Even well-planned risk treatments may not yield the expected results and could lead to unintended consequences. Therefore, continuous monitoring and review are essential to ensure that treatments remain effective.

5.4.2 Preparing and Implementing Risk Treatment Plans

5.4.2.1 The purpose of a risk treatment plan is to outline how the selected treatment strategies will be implemented. This ensures that everyone involved understands their roles and responsibilities and allows progress to be monitored effectively. The treatment plan should clearly prioritize the implementation of each treatment option.

5.4.2.2 Risk treatment plans should be incorporated into the organization's overall management processes and developed in consultation with relevant stakeholders. The plan should include the following information:

a) Justification for the chosen treatment options, including anticipated benefits.
b) Identification of those accountable and responsible for approving and implementing the plan.
c) Detailed actions required to execute the plan.
d) Resources needed, including contingency measures.
e) Performance measures to track progress.
f) Constraints that may affect the treatment process.
g) Reporting and monitoring requirements.
h) Timelines for the implementation and completion of actions.

5.5 Monitoring and Review

5.5.1 The primary goal of monitoring and review is to ensure the quality, effectiveness, and continuous improvement of the risk management process, its design, implementation, and outcomes. Ongoing monitoring and periodic review of the risk management process and its outcomes should be a planned part of the risk management process, with responsibilities clearly defined.

5.5.2 Monitoring and review includes activities such as planning, collecting, and analysing relevant data, documenting the outcomes, and providing feedback.

5.5.3 The results of monitoring and review should be incorporated throughout the organization's performance management, measurement and reporting activities.
Examples:

a) Conduct regular audits of data quality processes and controls to identify any weaknesses or new risks that may have arisen since the last assessment.
b) Adjust risk management strategies in response to evolving business needs or newly identified risks.

5.5.4 This iterative process of monitoring, reviewing, and adjusting ensures that data quality risks are effectively managed over time, leading to continuous improvement in data quality practices.

5.6 Communication and Consultation

5.6.1 Communication and consultation are integral to the entire risk management process, as they facilitate the involvement of stakeholders and ensure that all parties are aligned with the organization's risk management strategies.

5.6.2 Communication should occur at every stage, ensuring that risks, mitigation plans, and outcomes are clearly conveyed to all relevant stakeholders.

Key Elements of Communication and Consultation:

a) **Stakeholder Engagement**: Engage stakeholders, including IT teams, data owners, risk managers, and other relevant parties, throughout the risk management process. Their input and feedback are vital in identifying risks and evaluating treatment options.
b) **Effective Communication Channels**: Establish clear communication channels that allow for the transparent flow of information regarding risk management activities.
c) **Continuous Feedback:** Ensure that stakeholders provide ongoing feedback, which can be used to refine risk management strategies and treatment plans.

5.6.3 By fostering open communication and consultation, the organization ensures that risk management efforts are aligned with both internal and external expectations, and that everyone involved understands the actions being taken to manage data quality risks.

Appendix 1

Common data quality dimensions for marine data

Data quality dimensions provide a vocabulary for defining data quality requirements and can be used to define results of initial data quality assessments as well as ongoing measurements. Regardless of the terminologies used, dimensions focus on whether there is enough data (completeness), whether the data is right (accuracy, validity), how well the data fits together (consistency, integrity, uniqueness), whether the data is up to date (timeliness), accessible, usable and secure. Table A.1 contains definitions of a set of common data quality dimensions mapped to their measuring methods, as well as the data quality categories as defined in ISO 8000-8. This table can be used in the creation of an initial set of metrics to be used by an organization in measuring data quality dimensions.

Table A.1: Common data quality dimensions in marine data

Dimension	Description	Measurement	Unit of measure	Maritime Example	ISO 8000-8 Data Quality Category
Consistency	The extent to which data content and format are consistently represented within a data set and between data sets	i) Analysis of pattern consistency by using the presented structure and format within a column. – Structural Consistency ii) Analysis of the sampling rates in the same time series dataset. – Semantic Consistency	Percentage of consistent data (or inconsistent data)	Engine sensor data (e.g., temperature, pressure) from multiple generators must use the same format and sampling rate—e.g., temperature in °C with one decimal precision at 10-second intervals.	Syntactic Semantic
Validity	The degree to which data conforms to the syntax (format, type, range) of its definition	i) Comparison between the data and the metadata or documentation for the data item, such as the allowable types (string, integer, floating point etc.), the format (length, number of digits, etc.). – Structural Validity ii) Validating the data value by comparing it to the defined domain of value (minimum, maximum or contained within a set of allowable values). – Semantic Validity	Percentage of data items deemed valid (or invalid)	Sensor data from an engine monitoring system must adhere to defined data types (e.g., temperature as a floating-point number) and fall within the allowable range specified for safe operation.	Syntactic Semantic

Calibration	The process of aligning and adjusting data from various sources to ensure consistency and comparability, including standardizing measurements and formatting for uniformity and reliability across different datasets.	i) Comparison of measured values against known standards to determine deviations – Structural Calibration ii) Regular recalibration of systems based on usage or time intervals – Temporal Calibration	Calibration accuracy percentage	Recalibrating wind speed sensors on ships to ensure they provide consistent data compared to known meteorological standards.	Syntactic
Completeness	The degree to which all required data is present in a dataset, measured by the absence of null values and the validity of missing data.	i) A measure of the absence of blank (null) values or the presence of non-blank values. ii) Analysis of the conditions under which the missing value may/may not be valid for a certain data element.	Percentage of records with valid data (or invalid data)	Fuel consumption data logs from a voyage should have no missing entries for any time interval to support accurate performance and efficiency analysis.	Semantic
Continuity	The extent to which data is consistently available over time, without interruptions or gaps in coverage.	i) Assessment of gaps in time series or historical data – Structural Continuity ii) Evaluation of consistent data collection over time periods – Temporal Continuity	Percentage of data with no gaps over time	Continuous weather data collected from a ship's onboard systems during a voyage must have no time gaps to ensure accurate route optimization.	Pragmatic
Uniqueness	The degree to which data has no duplicate records within a data set.	Analysis of the number of the given data element as assessed in the "real-world" compared to the number of records of the data element in the dataset.	Percentage of unique data (or duplicated data)	Vessel traffic data in a port should contain unique vessel identifiers (e.g., IMO number) without duplicate records for the same entry time.	Semantic
Accuracy	The degree to which data correctly describes the physical parameter or event.	Accuracy is difficult to measure since the true physical value is typically unknown. Most measures of accuracy rely on comparison to a data source that	Percentage of data entries that pass the data accuracy rules (or fail the data rules)	A ship's position data must accurately represent its real-world GPS location to ensure collision avoidance and efficient navigation.	Semantic

		has been verified as accurate.			
Integrity	The degree to which the intended relationship exists between the data in one column and the data in another column of the same or different data sets	Assessment of the intended relationship to ensure that the data in one column of a table can be traced and connected to data in another column of the same or different table.	Percentage of data missing important relationship	Connecting ship voyage data with fuel usage data ensures that fuel consumption is linked to specific voyages for operational analysis.	Semantic
Comprehensibility	The degree to which data can be easily understood by its intended users.	i) Assess the clarity and simplicity of data labels, values, and metadata. – Structural Comprehensibility ii) Determine how well the data aligns with business context. – Semantic Comprehensibility	Percentage of users who correctly understand the data	Labels on ship maintenance records should clearly indicate components serviced, such as "engine oil replacement," to enable easy interpretation by engineers and management.	Semantic
Timeliness	The degree to which data represents reality from the required point in time.	Comparison of the time between when information is expected (standard timestamps) and when it is readily available for use (actual timestamps). Time difference	Percentage of time difference	Real-time data from collision-avoidance systems must be available instantly to ensure the safety of vessels during navigation.	Pragmatic
Compliance	The degree to which data adheres to regulatory standards and industry guidelines.	i) Compliance with regulatory requirements and industry standards. ii) Audit of data against compliance checklists.	Percentage of compliant data	Emission data from a ship must comply with MARPOL Annex VI requirements to meet international environmental standards.	Semantic
Recoverability	The ability to restore data after an incident, failure, or loss, ensuring minimal disruption to operations.	i) Time taken to restore data after a failure – Structural Recoverability ii) Evaluation of backup processes and their effectiveness – Procedural Recoverability	Percentage of data successfully restored	Recovery of voyage data logs after system failure to continue performance analysis.	Pragmatic
Traceability	The ability to track the history, usage,	i) Tracking data changes and	Percentage of data that can be	Tracking the origin of cargo information from	Semantic

	and location of data throughout its lifecycle.	updates – Structural Traceability ii) Monitoring the movement of data through processes – Semantic Traceability	traced effectively	port of loading to final destination ensures supply chain visibility.	
Availability	The extent to which data is accessible and retrievable when needed by authorized users.	i) Measure of data accessibility when requested – *Structural Availability* ii) Assessment of data retention period – *Temporal Availability*	Percentage of time data is available	Ship maintenance records should be accessible during inspections.	Pragmatic
Relevance	The degree to which data is applicable and helpful for the purpose it was collected.	i) Alignment of data with business objectives or user requirements – *Structural Relevance* ii) Evaluation of data usability – *Semantic Relevance*	Percentage of data deemed relevant	Weather forecast data must align with a ship's route planning requirements to avoid adverse conditions and optimize performance.	Semantic
Accessibility	The extent to which information is available, or easily and quickly retrievable.	Assessment of how easy it is to acquire data when needed, how long it is retained, how access is controlled.	User surveys through questionnaires or interviews. (e.g. If a SQL database is provided by the 3rd-party data supplier)	Accessing real-time Automatic Identification System (AIS) data to monitor vessel positions and ensure authenticity by cross-referencing with radar systems.	Pragmatic
CIA (Confidentiality, Integrity, Availability)	Refers to ensuring data is protected from unauthorized access (Confidentiality), remains accurate (Integrity), and is accessible when needed (Availability).	i) Confidentiality: Measure of data protection from unauthorized access ii) Integrity: Percentage of data free from unauthorized changes iii) Availability: Percentage of data available on time	Percentage of secure, accurate, and available data	Securing engine performance data from unauthorized access, ensuring no unauthorized modifications, and making it accessible to engineers when needed.	Syntactic/ Pragmatic

Appendix 2

Common data quality rules, metric and dimensions for data quality verification and validation

Data quality rules provide the foundation for operational management of data quality. Below Tables give samples of the common data quality rules, measurable metrics and dimensions in accordance with the specific data quality problems. Each table provides a structured way to verify and validate data quality in alignment with the three categories of data quality described in ISO 8000-8:2015.

Table A.2.1: Sample of Syntactic Data Quality Rules and Dimension

Data Quality Issue	Problem Description	Syntactic Rule	Data Quality Metrics	Data quality dimension
Invalid Data Type	Invalid data type for the same data element, e.g. a text string found in a list of floating numbers.	Data type for the same data element must conform to metadata of its definition (e.g. text, number or date/ time).	Percentage of invalid/ inconsistent data values	Structural Validity
Different Data Format/Pattern	Different data format/ pattern for the same data element, e.g. different format in a timestamp that makes data manipulation/ comparation difficult.	Data format/pattern for the same data element must conform to metadata of its definition (e.g. DD.DD, DDDD-DD-DD DD:DD, where D represents a digit).		Structural Consistency
Values out of Range	Data values are out of range for the domain under observation, e.g. value spikes or sudden changes which are implausible (99999, -99999) for the domain.	Data values for the same data element must conform to metadata of its definition (e.g. the assigned value is within a defined numeric, lexicographic, or time range).		Validity
Different Data Accuracy	Different level of data accuracy for the same data element (i.e. significant digits: number of digits on the right of the decimal point).	The level of data accuracy for the same data element must be same.		Precision

Table A.2.2 Sample of Semantic Data Quality Rules and Dimensions

Data Quality Issue	Problem Description	Semantic Rule	Data Quality Metrics	Data quality dimension
Missing Data Values	There are gaps in the time series data.	The missing values under a specific condition are unacceptable.	Percentage of invalid/ inconsistent data values	Structural Validity
Duplicated Data Records	The data element is recorded with the same values more than once within the dataset that is unacceptable to what it represents in the real-world, e.g., duplicated timestamps.	The data element must have a unique representation in the real-world.		Uniqueness
Inaccurate Measurement	The value is slightly wrong which might result in the detection of a wrong trend etc. e.g., signal noise.	The measured data value shall match the property value for the real-world object it represents.		Accuracy
Diverging Sampling Rate	Different sampling rates in the same time series (same data source) can lead to problems (e.g., irregular timestamps).	The sampling rates in the same time series (same data source) need to be consistent.		Semantic Consistency
Divergent Despite High Correlation	Values which are normally correlated behave unexpectedly	The intended data relationship linkage must be retained within the dataset.		Integrity
Forced / Calculated Value	Compensated values are used instead of real measurements.	Minimize the corrected (e.g., interpolation) data values which may reflect the assumptions made and no longer represent reality.		Plausibility
Wrong Timestamps or Wrong Timestamp Order	Timestamps are mismatched to the expected time (e.g., there are gaps/redundancies in timestamps); Timestamps are not in chronological order.	The recorded time shall match to the expected (standard) time.	Percentage of mismatched timestamps	Timeliness
Data not updated (stuck values)	Data is not up to date. Sensor might still display old values (e.g., defective sensors for which the data value is out of calibration).	Data must be updated with time.	Percentage of stuck values	Currency

Table A.2.3: Sample of Pragmatic Data Quality Rules and Dimensions

Data Quality Issue	Problem Description	Pragmatic Rule	Data Quality Metrics	Data quality dimension
Missing Foreign Keys	Foreign Keys are missing, which are used to identify the referential relation between tables in a SQL database.	The data is easy and quick to retrieve from a database (e.g., foreign keys need to maintain referential integrity in SQL database).	User surveys (e.g., questionnaires)	Accessibility

Appendix 3

Data Profiling Tools

Data profiling tools play a crucial role in evaluating and enhancing the quality of data, which is fundamental to effective data governance and analytics. These tools assist organizations in understanding data structures, identifying inconsistencies, and enhance the overall quality of datasets. This section categorizes data profiling tools into two primary types: Open Source/Free Tools and Commercial Tools, providing an overview of the key features, and capabilities of popular tools in each categories, allowing readers to make informed decisions based on their specific needs and resources.

Disclaimer: The tools listed in this section are included solely for informational purposes. We do not endorse or promote any specific tools.

A.3.1 Open Source/ Free Tools

A.3.1.1 Some tools are free software and open source; however, many, but not all free data profiling tools are open source projects. In general, their functionality is limited compared to that of commercial products, and they may not offer free support (telephonic/online). Further, their documentation is not always thorough. However, some small organizations still use these free tools instead of expensive commercial software, considering the benefits that free tools provide.

(a) Aggregate Profiler Tool:
Aggregate Profiler (AP) is an open source project developed in Java. AP supports both traditional database and big data, such as Hadoop or Hive, and it offers statistical analysis, pattern matching, distribution chat, basket analysis, etc. AP also supports data generation, data preparation, data masking features, and address correction for data quality projects. Moreover, this tool offers data validation (metadata profiling, analytical profiling, and structural profiling), and data quality (removing duplicate data, null values, and dirty data).

(b) Talend Open Studio for Data Quality:
Talend Open Studio for Data Quality (TOSDQ) is also based on Java and is a mature open source tool. TOSDQ offers navigator interface to access databases and data files. This tool supports catalogue analysis, time correlation analysis, column analysis, table analysis, column correlation analysis, and schema analysis; it also supports column functional dependency, redundancy analysis, numerical correlation analysis, nominal correlation analysis, connection analysis, column set analysis, and match analysis. Furthermore, TOSDQ reports several different types of statistics indicators, including simple statistics, text statistics, summary statistics, pattern frequency statistics, Soundex frequency statistics, phone number statistics, and Fraud detection (Ben-ford's law frequency).

(c) DataCleaner
DataCleaner is a commercial tool for data profiling and data cleaning, but it has a free version which offers multiple data profiling functions, including pattern matching, boolean analysis, weekday distribution, completeness analysis, value matcher, character set distribution, value distribution, date gap analysis, unique key check, date/time analysis, string analysis, number analysis, referential integrity, and reference data matching.

(d) Jupyter Notebook
Jupyter Notebook is an open-source tool that allows users to write and execute code in an interactive environment, primarily using Python. It offers a high level of flexibility for data profiling tasks by enabling users to write customized scripts based on specific project needs. Jupyter supports tasks such as Uniqueness checks, Completeness analysis, identifying Null values, Pattern Analysis, checking Data Types, Data Formats, and more. Additional features include the ability to analyze Precision, Scale, Min/Max Length, Range, Min/Max Value, Value Distribution, Frequency Distribution, and Quantile

Distribution. This tool is widely used in data science and machine learning for its versatility and integration with popular data manipulation libraries like pandas and numpy.

A.3.2 Commercial Tools

A.3.2.1 Commercial data profiling products usually come packaged in data governance suites. These products have multiple functions, high performance, and strong capabilities; they can connect to other suites to provide comprehensive solutions for customers. Moreover, such software is not only powerful, but end-users also can find online services and telephone support.

(a) IBM InfoSphere Information Analyzer
IBM InfoSphere Information Analyzer is part of IBM's data governance suite that includes InfoSphere Blueprint Director, Metadata Workbench, DataStage, QualityStage, Data Click, Business Glossary, and Information Services Director. It supports column analysis (statistics, distribution, cardinality, and value analysis.), identifying keys and relationships, discovering redundant data, comparing data and structures through history baselines, analyzing data via data rules, and importing and exporting data rules.

(b) Informatic Data Profiling
This profiling software supports aggregate functions (count null values, calculate averages, get maximum or minimum values, and get lengths of strings), candidate key evaluation (unique or non-unique), distinct value count, domain inference, functional dependency analysis, redundancy evaluation, and row count. In addition, users can add business rules (verbose mode) or configure profile functions in this tool.

(c) Oracle Enterprise Data Quality
Oracle Enterprise Data Quality permits address verification, profiling data (files, databases, and spreadsheets), standardization, audit reviews (incorrect values, missing data, inconsistencies, duplicate records, and key quality metrics), matching and merging columns (duplicate prevention, de-duplication, consolidation, and integration), and case management (data reviewing). Furthermore, the tool can utilize pre-built templates or user-defined rules to profile data. It also can connect to other Oracle data governance products, including Oracle Data Integrator and Oracle Master Data Management.

(d) SAP Information Steward
SAP Information Steward can improve information quality and governance via the Data Insight module (data profiling and data quality monitoring), Metadata Management module (metadata analysis), Metapedia Module (business term taxonomy), and cleansing package builder (cleansing rules). The data insight module can define validation rules, determine profiling (column, address, uniqueness, dependency, and redundancy), import and export metadata, and create views.

(e) SAS DataFlux Data Management Studio
SAS DataFlux Data Management Studio is a data governance suite that consists data profiling, master data management, and data integration. This data pro-filing tool covers key analysis (primary and foreign keys), pattern frequency distribution analysis, redundant data analysis, and data profiling reports.

(f) Collibra Data Stewardship Manager
Collibra Data Stewardship Manager (DSM) module is part of Collibra's Data Governance Center that also includes Business Semantic Glossary (BSG) and Reference Data Accelerator (RDA) module. DSM also provides historical data quality re-ports around trend analysis and reports to understand the impact of resolved data is-sues. In addition, DSM provides fully configurable data quality reporting dashboard (see figure below) by bringing data quality rules and metrics calculated in one or multiple sources (data quality tools, databases, and big data).

Appendix 4

Data Profiling Techniques and Implementation Codes

This section provides a detailed explanation of various data profiling techniques and their corresponding implementations, helping analysts assess and evaluate datasets effectively.

A.4.1 Cross-column analysis

Cross-column analysis examines the relationships and dependencies between columns within a single table. This section includes code examples and visualizations, created using Matplotlib, to evaluate aspects such as uniqueness, completeness, null value counts and other such matrices. These analyses are presented through practical implementation and accompanying diagrams for better clarity.

A.4.1.1 Uniqueness Analysis

Uniqueness analysis helps determine the percentage of unique values within individual columns. It provides insights into the diversity of data in a column. This can be calculated using the nunique() method in Python.

```
Python Copy code:
uniqueness = df.nunique() / len(df) * 100
```

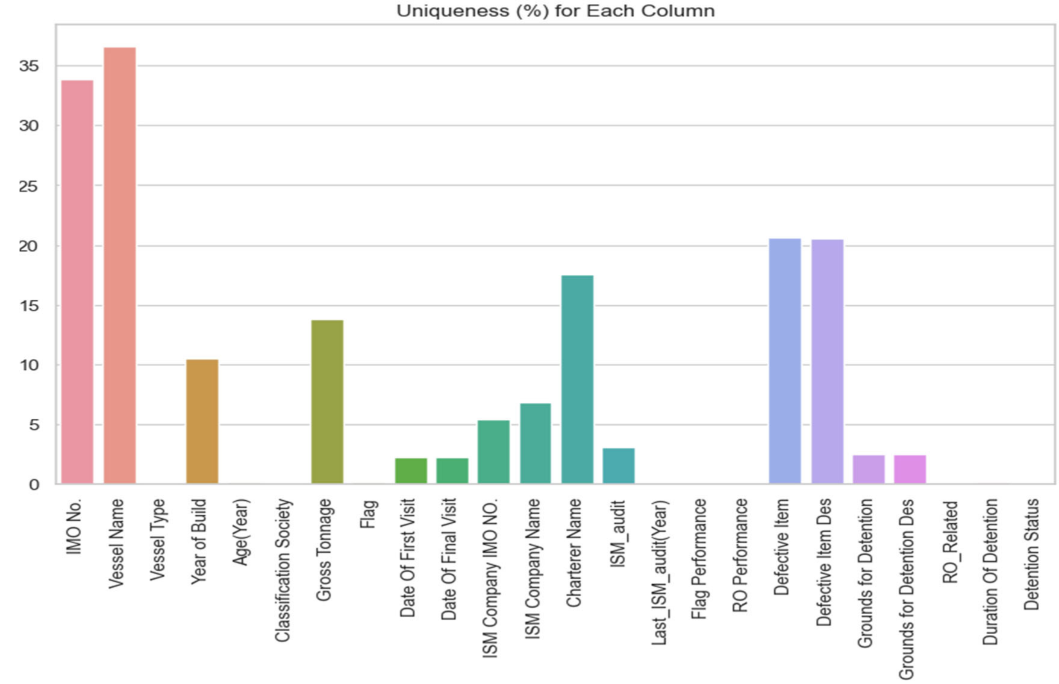


Figure A.4.1.1: Uniqueness Analysis

A.4.1.2 Completeness Analysis:

Completeness analysis assesses the proportion of non-null values in each column, which indicates how complete the dataset is. It is calculated by comparing the count of non-null values (count()) to the total number of rows in the dataset.

```
Python Copy code
completeness = df.count() / len(df) * 100
```

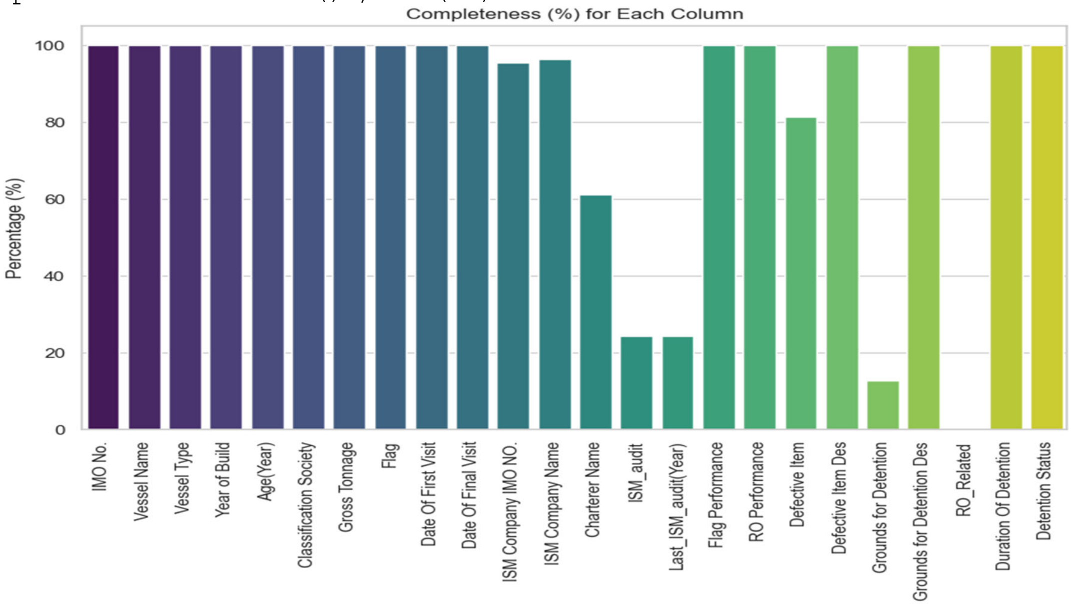


Figure A.4.1.2: Completeness Analysis

A.4.1.3 Counts of Null:

This analysis identifies the number of null or missing values in each column, which is essential for understanding the gaps in the data. It can be calculated using isnull().sum().

```
Python Copy code
null_counts = df.isnull().sum()
```

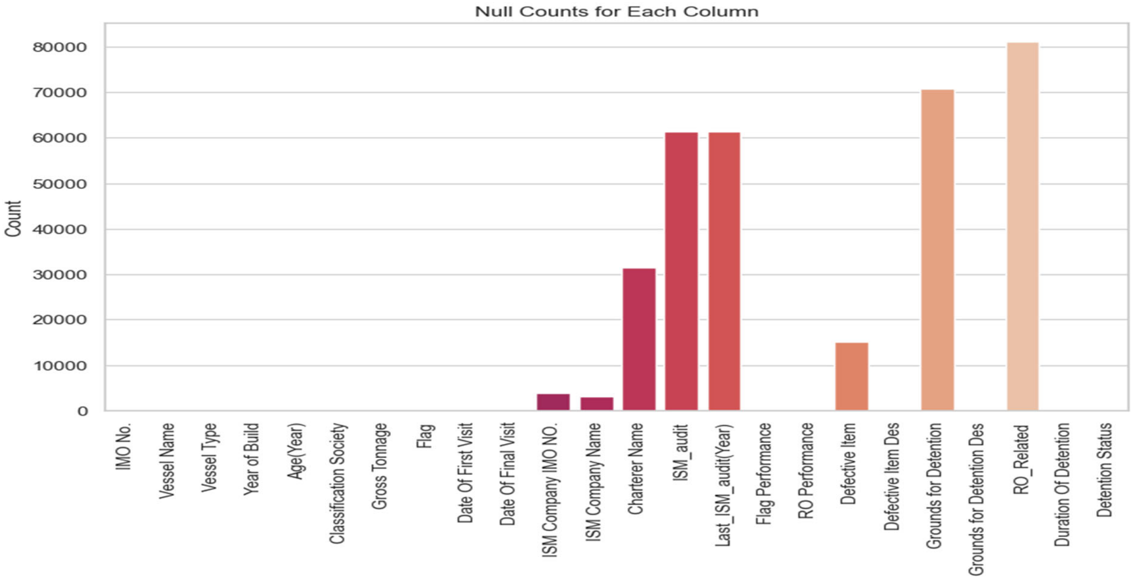


Figure A.4.1.3: Count of null

A.4.1.4 Pattern Analysis

Pattern analysis focuses on identifying patterns in column values, helping to ensure consistency and detect anomalies.

A.4.1.4.1 Data type pattern

Patterns such as words (e.g., "W") and numbers (e.g., "N") can be identified using regular expressions. This helps detect structural patterns in string columns. For example, regex can check if values in a column follow a specific structure like phone numbers or email addresses.

```
Python Copy code
import re

# Define a pattern function for each column
def detect_pattern(val):
    if pd.isnull(val):
        return 'NaN'
    elif re.match(r'^[0-9]+$', str(val)):
        return 'N'  # Numbers
    elif re.match(r'^[A-Za-z]+$', str(val)):
        return 'W'  # Words
    else:
        return 'Mix'  # Mixed or other

# Apply the function to columns
df['pattern'] = df['column_name'].apply(detect_pattern)
```

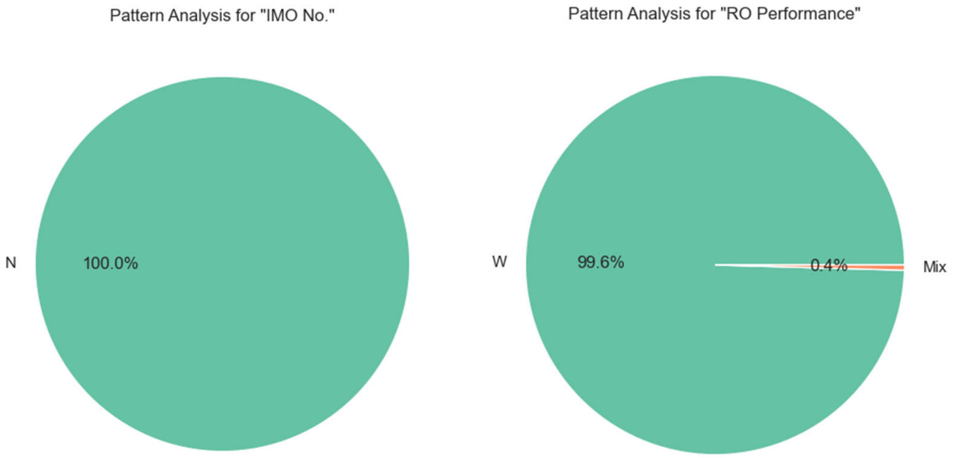
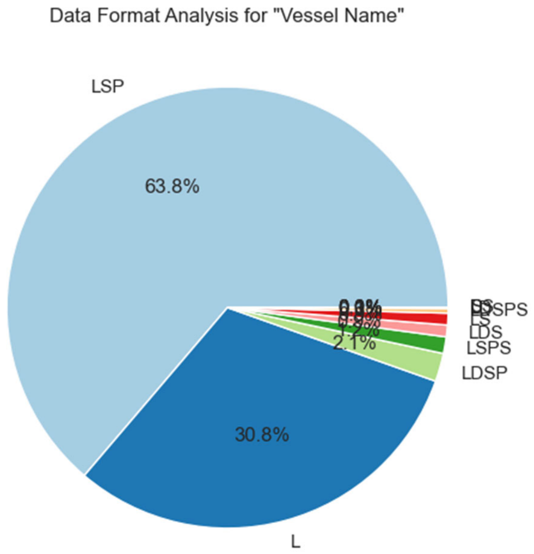


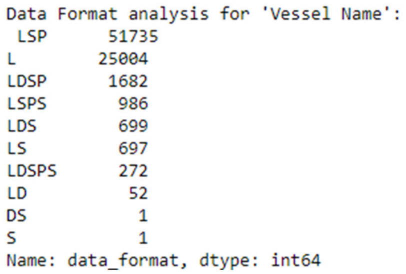
Figure A.4.1.4.1: Data type pattern

A.4.1.4.2 Data Format

This is a specialized form of pattern analysis used to verify specific formats, such as combinations of letters and digits. Custom functions can be used to validate formats like "ABC123" or "dd-mm-yyyy".

```
Python Copy code
def detect_format(val):
    if pd.isnull(val):
        return 'NaN'
    format_pattern = ''
    if re.search(r'[A-Za-z]', str(val)):
        format_pattern += 'L'  # Letters
    if re.search(r'[0-9]', str(val)):
        format_pattern += 'D'  # Digits
    if re.search(r'\s', str(val)):
        format_pattern += 'SP'  # Spaces
    if re.search(r'[^A-Za-z0-9\s]', str(val)):
        format_pattern += 'S'  # Symbols

    return format_pattern if format_pattern else 'Other'

df['format'] = df['column_name'].apply(detect_format)
```





```
Data Format analysis for 'Vessel Name':
 LSP       51735
L         25004
LDSP       1682
LSPS        986
LDS         699
LS          697
LDSPS       272
LD           52
DS            1
S             1
Name: data_format, dtype: int64
```

Figure A.4.1.4.2: Data format

A.4.1.4.3 Precision and Scale

For numeric columns, precision (total number of digits) and scale (number of digits after the decimal point) can be analyzed by parsing each value. This ensures numeric data conforms to expected accuracy levels.

```
Python Copy code
# Precision: Total digits
df['precision'] = df['numeric_column'].apply(lambda x:
len(str(x).replace('.', '')) if pd.notnull(x) else 0)

# Scale: Digits to the right of the decimal
df['scale'] = df['numeric_column'].apply(lambda x: len(str(x).split('.')[1])
if pd.notnull(x) and '.' in str(x) else 0)
```

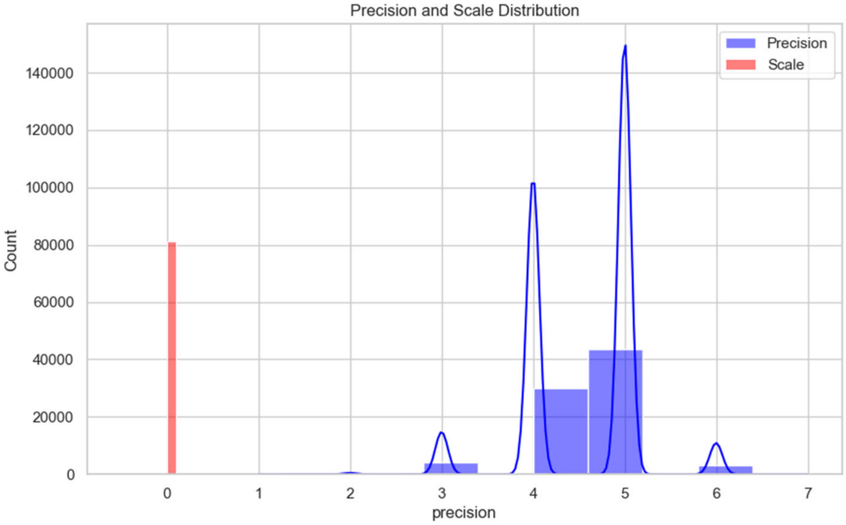


Figure A.4.1.4.3: Precision and Scale

A.4.1.4.4 Min/Max Length

The minimum and maximum lengths of values in each column can be computed using the len() function. For example, df['column_name'].apply(len).min() returns the shortest value length, and df['column_name'].apply(len).max() returns the longest value length.

```
Python Copy code
df['min_length'] = df['column_name'].apply(lambda x: len(str(x)) if
pd.notnull(x) else 0)
df['max_length'] = df['column_name'].apply(lambda x: len(str(x)) if
pd.notnull(x) else 0)
```

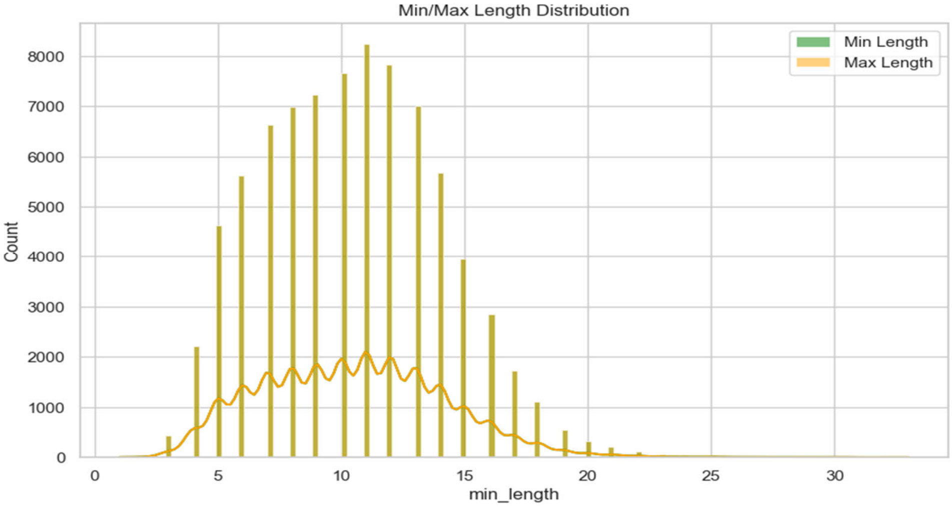
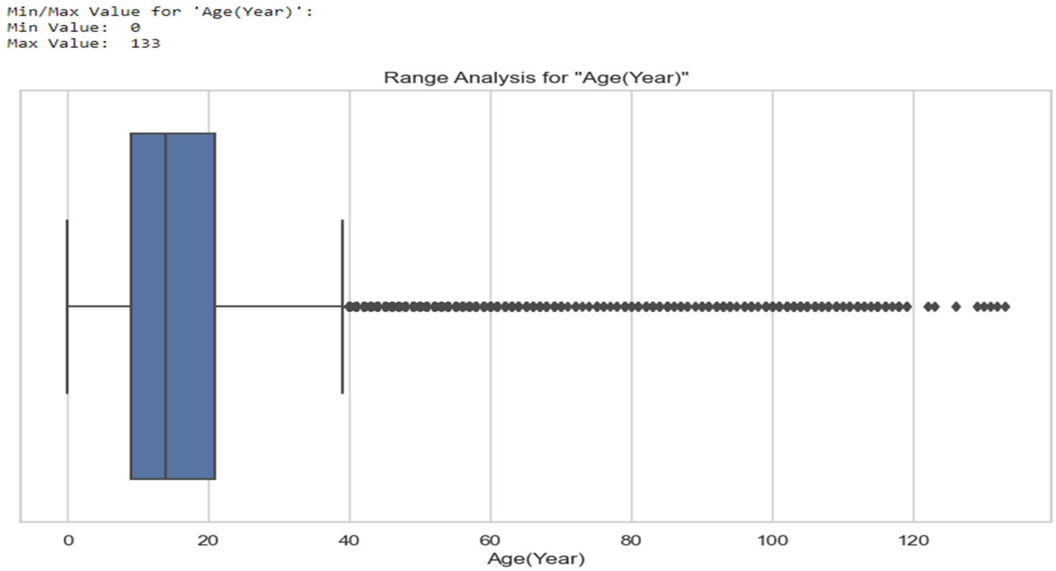


Figure A.4.1.4.4: Min/Max Length

A.4.1.5 Range Analysis (Min/Max Value)

Range analysis identifies the smallest (minimum) and largest (maximum) values in numeric columns, helping to detect outliers or validate data ranges. This can be performed using the min() and max() functions. For example, df['column_name'].min() returns the smallest value in the column.

```
Python Copy code
min_value = df['numeric_column'].min()
max_value = df['numeric_column'].max()
```



```
Min/Max Value for 'Age(Year)':
Min Value:  0
Max Value:  133
```

Figure A.4.1.5: Range Analysis (Min/Max Value)

A.4.1.6 Value Distribution Analysis

Value distribution analysis helps understand the spread and concentration of data within a column.

A.4.1.6.1 Frequency Distribution

This analysis determines the frequency of each unique value in a column, which is useful for categorical data. It can be performed using the value_counts() function.
For example, df['column_name'].value_counts() shows the frequency of each unique value.

```
Python Copy code
frequency_distribution = df['column_name'].value_counts()
```

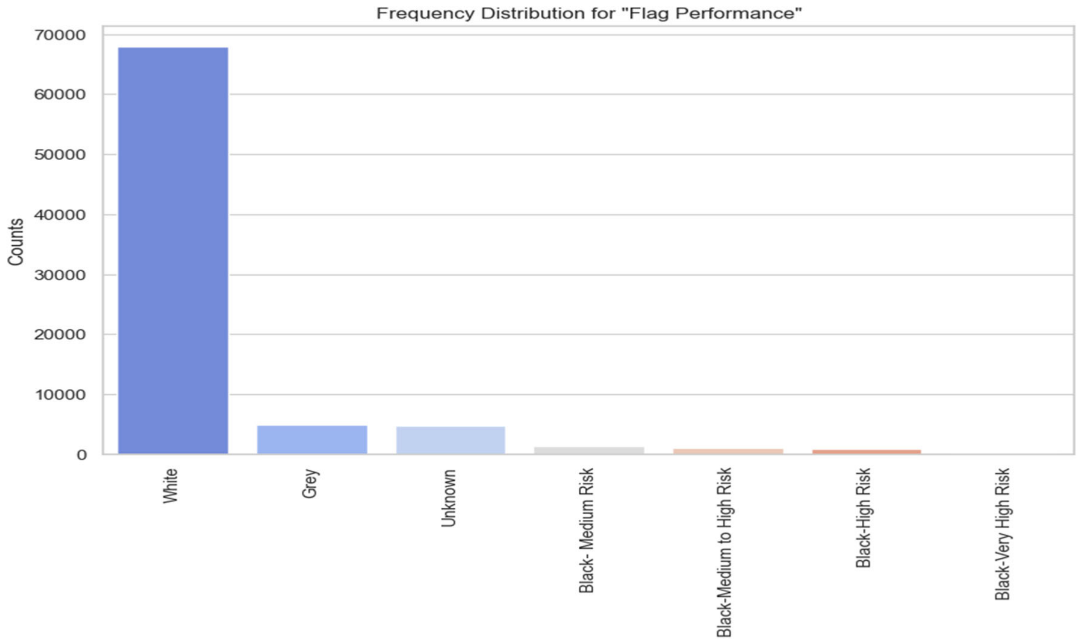
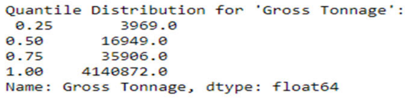


Figure A.4.1.6.1: Frequency Distribution

A.4.1.6.2 Quantile Distribution

Quantile distribution divides data into intervals to understand its spread. It can be calculated using the quantile() function for different intervals (e.g., 0%, 25%, 50%, etc.).

```
Python Copy code
quantiles = df['column_name'].quantile([0.25, 0.5, 0.75, 1.0])
```



```
Quantile Distribution for 'Gross Tonnage':
 0.25       3969.0
0.50      16949.0
0.75      35906.0
1.00    4140872.0
Name: Gross Tonnage, dtype: float64
```

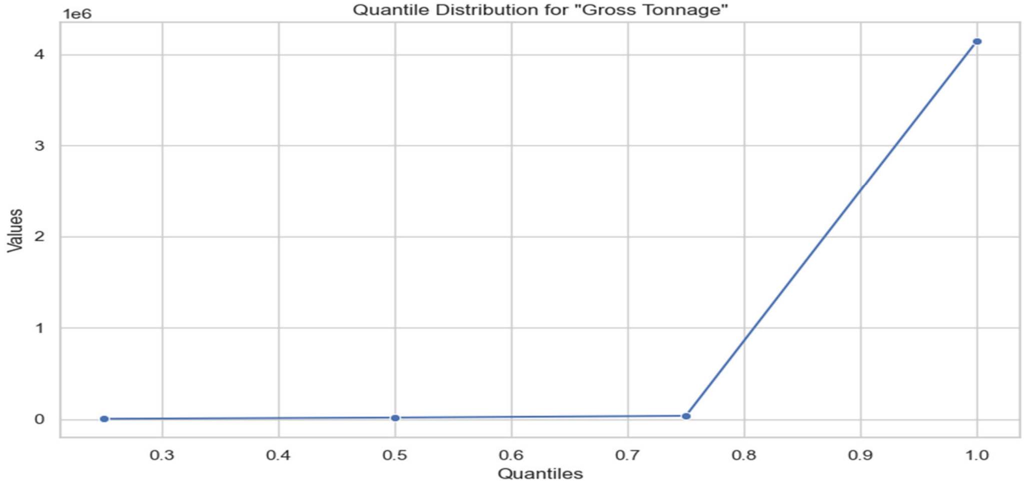
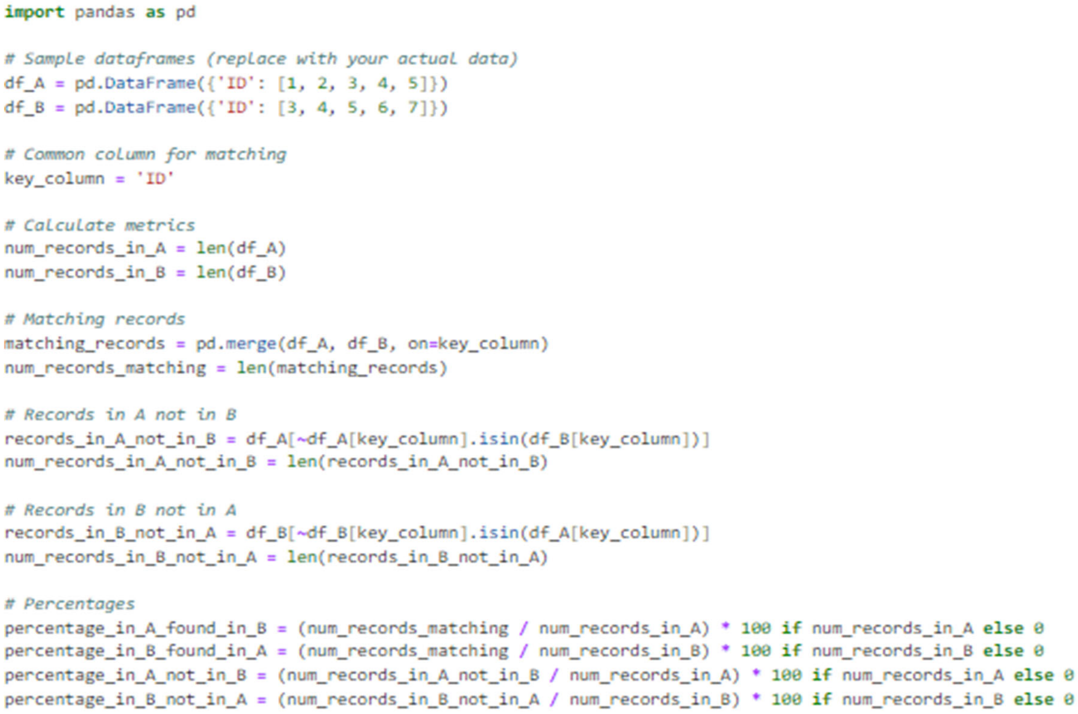


Figure A.4.1.6.2: Quantile Distribution

A.4.2 Inter-table analysis

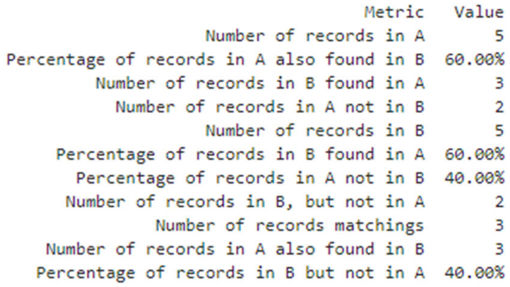
A.4.2.1 Inter-table analysis focuses on identifying and evaluating the relationships between tables and assessing how effectively these relationships are maintained. The code example assumes that the data frames being analyzed share a common column (key), often a unique identifier, used for joining the tables.

Code:



```
import pandas as pd

# Sample dataframes (replace with your actual data)
df_A = pd.DataFrame({'ID': [1, 2, 3, 4, 5]})
df_B = pd.DataFrame({'ID': [3, 4, 5, 6, 7]})

# Common column for matching
key_column = 'ID'

# Calculate metrics
num_records_in_A = len(df_A)
num_records_in_B = len(df_B)

# Matching records
matching_records = pd.merge(df_A, df_B, on=key_column)
num_records_matching = len(matching_records)

# Records in A not in B
records_in_A_not_in_B = df_A[~df_A[key_column].isin(df_B[key_column])]
num_records_in_A_not_in_B = len(records_in_A_not_in_B)

# Records in B not in A
records_in_B_not_in_A = df_B[~df_B[key_column].isin(df_A[key_column])]
num_records_in_B_not_in_A = len(records_in_B_not_in_A)

# Percentages
percentage_in_A_found_in_B = (num_records_matching / num_records_in_A) * 100 if num_records_in_A else 0
percentage_in_B_found_in_A = (num_records_matching / num_records_in_B) * 100 if num_records_in_B else 0
percentage_in_A_not_in_B = (num_records_in_A_not_in_B / num_records_in_A) * 100 if num_records_in_A else 0
percentage_in_B_not_in_A = (num_records_in_B_not_in_A / num_records_in_B) * 100 if num_records_in_B else 0
```

Output:



```
                                  Metric   Value
                    Number of records in A       5
 Percentage of records in A also found in B  60.00%
          Number of records in B found in A       3
            Number of records in A not in B       2
                    Number of records in B       5
      Percentage of records in B found in A  60.00%
        Percentage of records in A not in B  40.00%
       Number of records in B, but not in A       2
                Number of records matchings       3
     Number of records in A also found in B       3
    Percentage of records in B but not in A  40.00%
```

Figure A.4.2: Inter-table analysis code demonstration

References

1. ISO 8000-1:2022. Data Quality-Part 1: Overview.
2. ISO 8000-8:2015. Data Quality-Part 8: Information and data quality: Concepts and measuring.
3. IS/ISO 8000-60:2017 Data quality management: Overview
4. ISO 8000-61:2016 Data quality management: Process reference model
5. ISO 8000-81:2021 Data quality assessment: Profiling
6. ISO 31000:2018. Risk management
7. ISO 19848:2018 Ships and marine technology - Standard data for shipboard machinery and equipment
8. International Association of Classification Societies (IACS). (2024). Recommendation 183: Ship Data Quality. Retrieved from https://www.iacs.org.uk
9. Abdul Ghani Mohammed, Aziz Eram (2017, October): ISO 8000-61 Data Quality Management Standard, TDQM Compliance, IQ Principles. MIT International Conference on Information Quality, UA Little Rock.
10. Ian Michael Prilop (2014, July): Continuous Data Quality Assessment in Information Systems (Diploma Thesis). Natural Language Processing Group, Institute of Computer Science, Faculty of Mathematics and Computer Science, University of Leipzig, Germany.
11. Wei Dai, Isaac Wardlaw (2016, April): Data Profiling Technology of Data Governance Regarding Big Data: Review and Rethinking. Information Technology: New Generations: 13th International Conference on Information Technology (pp.439-450)
12. Ziawasch Abedjan, Lukasz Golab (2016, May): Data Profiling. IEEE 32nd International Conference on Data Engineering, Finland.
13. DAMA International (2017) - The DAMA Guide to the Data Management Body of Knowledge (DAMA-DMBOK) -Second Edition. USA: Technics Publications, LLC.
14. AltexSoft (2019, October): Data Quality Management: Roles, Process, And Tools. Retrieved from https://www.altexsoft.com/blog/data-quality-management-and-tools/.