Data Profiling Terminology

# Introduction

Data profiling is the exploratory analysis required to characterise data (e.g., distributions) and assess its quality (e.g., missing values). Abedjan et al. (2015) provide an overview of characterisation tasks, and definitions of less common ones are:

|  |  |
| --- | --- |
| **Characterisation task** | **Description** |
| Value lengths | The number of characters in a value |
| First digit | The distribution of the first digit of a numerical variable (cf. Benford’s law) |
| Character types | The distinct characters used in any value of a variable (see Ruddle & Hall, 2019) |
| Value patterns | The sequences of types of value (e.g., a1234; lowercase letter then 4 digits) |

BS ISO/IEC 25012:2008 lists 15 high-level types of data quality (below). Accuracy is “the degree to which data has attributes that correctly represent the true value of the intended attributes of a concept or event in a specific context of use.” Completeness is “the degree to which subject data associated with an entity has values for all expected attributes and related entity instances in a specific context of use.” Consistency is “the degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use.”

|  |  |  |
| --- | --- | --- |
| Accessibility | Confidentiality | Portability |
| Accuracy | Consistency | Precision |
| Availability | Credibility | Recoverability |
| Completeness | Currentness | Traceability |
| Compliance | Efficiency | Understandability |

## Data quality tasks

Each of high-level type of data quality may be subdivided into a variety of tasks, including:

|  |  |
| --- | --- |
| **Completeness task** | **Description (where not obvious)** |
| Missing variable name | - |
| Missing column | - |
| Missing record | - |
| Missing value | - |
| Zero value | Where a zero indicates a value was missing |
| Use of default value | A default value was incorrectly used |
| Duplicate record | - |
| Coverage | E.g., geographic or temporal |
| Granularity | E.g., too coarse in terms of space or time |

|  |  |
| --- | --- |
| **Accuracy task** | **Description (where not obvious)** |
| Numeric outlier | An extreme numerical value |
| Time-series outlier | An extreme time value |
| Unusual category name | A categorical value that looks out of place |
| Special value | A value that has a particular meaning (e.g., a code for “unknown”) |
| Measurement or recording error | A value that is clearly wrong |
| Misspelling | - |
| Invalid value | - |
| Implausibly low/high value | Domain knowledge is needed to assess whether a value is plausible |
| Implausible change in value | Ditto |
| Implausible range | Ditto |
| Wrong data format | - |
| Wrong data type | - |

|  |  |
| --- | --- |
| **Consistency task** | **Description (where not obvious)** |
| Different data format | - |
| Different granularities | E.g., levels of aggregation |
| Different times | E.g., information refers to different points in time |
| Different word orderings | E.g., for first name and family name |
| Inconsistent spelling | - |
| Inconsistent units | - |
| Inconsistent value | - |
| Incorrect reference | E.g., employee associated with the wrong department |
| Naming conflicts | Synonyms; homonyms |
| Referential integrity violation | One entity has no correspondence in another entity |
| Violation of functional dependency | E.g., age vs. birth date |

# References

Z. Abedjan, L. Golab, and F. Naumann, “Profiling relational data: a survey,” The VLDB Journal, vol. 24, no. 4, pp. 557–581, 2015.

International Organization for Standardization/International Electrotechnical Commission et al. 2008. Software engineering-Software product Quality Requirements and Evaluation (SQuaRe) Data quality model. ISO/IEC 25012 (2008), 1–13.

R. Ruddle and M. Hall, “Using miniature visualizations of descriptive statistics to investigate the quality of electronic health records,” in Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies-Volume 5: HEALTHINF. SciTePress, 2019, pp. 230–238.