Data Quality

Mingjun Zhong

Department of Computing Science

University of Aberdeen

Content

- What is data quality?
- Data quality issues

What is data quality?

The following was taken from Wikipedia

- Data quality (Wikipedia): refers to the state of qualitative and quantitative pieces of information.
- Data is generally considered high quality if it is “fit for intended uses in operations, decision making and planning.”
- Data is deemed of high quality if it correctly represents the real-world construct to which it refers.

What is data quality?

The following was taken from IBM

- Data quality measures how well a dataset meets criteria for **accuracy, completeness, validity, consistency, uniqueness, timeliness and fitness for purpose**, and it is critical to all data governance initiatives within an organisation.
- When data quality meets the standard for its intended use, data consumers can trust the data and leverage it to improve decision-making, leading to the development of new business strategies or optimisation of existing ones.
- When a standard isn't met, **data quality tools** provide value by helping businesses to diagnose underlying data issues.

Data quality metrics

- **Completeness:** This represents the amount of data that is usable or complete. If there is a high percentage of ***missing values***, it may lead to a biased or misleading analysis if the data is not representative of a typical data sample.
  - Missing values
- **Uniqueness:** This accounts for the amount of ***duplicate data*** in a dataset. For example, when reviewing customer data, you should expect that each customer has a unique customer ID.
  - Duplication records
- **Validity:** This dimension measures how much data matches the required format for any business rules. ***Formatting*** usually includes metadata, such as valid data ***types, ranges, patterns***, and more.
  - Data types, ranges, patterns match the requirements?
- **Timeliness:** This dimension refers to the readiness of the data within an expected time frame. For example, customers expect to receive an order number immediately after they have made a purchase, and that data needs to be generated in ***real-time***.
  - Data received within the required timeline.

Data quality metrics

- **Accuracy:** This dimension refers to the ***correctness of the data values*** based on the agreed upon "source of truth." Since there can be multiple sources which report on the same metric, it's important to designate a primary data source; other data sources can be used to confirm the accuracy of the primary one. For example, tools can check to see that each data source is trending in the same direction to bolster confidence in data accuracy.
  - Patients recorded addresses match their physical addresses.
  - Data recorded has high precision for analysis (e.g., stock price recorded every hour vs. week.)
- **Consistency:** This dimension evaluates data records from two different datasets. As mentioned earlier, ***multiple sources*** can be identified to report on a single metric. Using different sources to check for consistent data trends and behavior allows organizations to trust the any actionable insights from their analyses. This logic can also be applied around relationships between data. For example, the number of employees in a department should not exceed the total number of employees in a company.
  - A person's address is updated, this should also be updated in patient record and bank statement, etc.
- **Fitness for purpose:** Finally, fitness of purpose helps to ensure that the data asset meets a business need. This dimension can be difficult to evaluate, particularly with new, emerging datasets. These metrics help teams conduct data quality assessments across their organizations to evaluate how ***informative and useful data*** is for a given purpose.
  - To study the energy consumption of Scottish houses on average, can we survey Aberdeen homes only?

Why is Data Quality important?

- Data is the foundation for the application of Artificial Intelligence, Internet of Things (IoT), Edge Computing, etc
- Accuracy and reliability of information (data) used for data analysis, decision-making
- Crucial for business decisions – high-quality data provides accurate key important factors in business, and so the growth of the business: improved business processes – good data helps the team to make the right operations; increased customer satisfaction – good data provides marketing and sales better ideas to target buyers.

Data quality issues

Corresponding to the data quality metric, we will look at:

- Canonicalization, standardization, or normalization
- Missing data
- Anomalies
- Non-stationarity
- Data duplication

Canonicalization, standardization, or normalization

- Normalization – to convert *entities* or *column names* to a standard/canonical format
- **Scenario 1**: identifying which values of a given feature correspond to the same entities – termed ***cell entity normalization***
  - For example, **U.K., UK, and United Kingdom** refer to the same entity, but have different representations
- **Scenario 2**: obtaining a standard representation for every value under a given feature
  - E.g., specific formats for dates, addresses, etc
  - E.g., Standardising the units of physical measurements for a given feature
- **Scenario 3:** Standardize column names

Cell entity normalization

- Identifying those entities of a given feature which refer to the same entity: `U.K.' vs `United Kingdom'
- Identifying feature names which refer to the same feature: `Country' vs `country'
- Variability: different names for the same entity; abbreviations, syntactic mismatches (i.e., typos, double spaces, capitalization, etc.) - `Country' vs `County'
- Example: `Highly Active' vs `Highly_Active'; `UK' vs `United Kingdom'; `Highly Active' vs `Highly active'.

Canonicalization of features

- Refers to any process that involves representing a specific feature type with a standard format
- Decision made by data engineering/scientist people or domain knowledge people to select the standard format
- Standard format is decided by domain expert
- For example: dates can be represented as `16/06/2025' or `16 June 2025' or '2025-06-16'

Canonicalization of units

- Refers to any process that involves ***transforming*** the numerical values and units of a feature into a standard representation
- This problem commonly arises from physical measurements being recorded with different units.
- For example: Height can be recorded in ***metres*** or ***centimetres***, temperature in ***Celsius*** or ***Fahrenheit***
- Different units can lead to different results; scientists need to normalise them to a single unit
- In home energy data, refrigerators were recorded with different units: litres, cubic feet or no unit, also written in different formats, e.g., `Liters', `Ltr', `L', or no unit.
  - Need to remove units from feature values and standardize them to a common unit

Canonicalization of column names

- Refers to any process that involves representing the feature names with a standard format
- Decision made by data engineering/scientist people or domain knowledge people to select the standard format
- Sometime, standard format must be decided by domain expert
- For example, in this table: `package_for_weighting' -> `Package_for_Weighting'; `ISP'?

	A	B	C	D	E	F	G	H	I	J	K	L	M	
1	unit_id	ISP	Technolog	PACKAGE (download by upload)	MarketCla	Distance from exchange - DSL only	package_for_weighting	Region	Country	Geography	Download - 24 (ave)	Download - 24 min(ave)	Download	Dow
2	940006	BT	ADSL1	8	A	3698.29	BT 8 ADSL1	Scotland	Scotland	Rural	1.748510753	1.434534194	1.858125	1.7
3	33806	BT	ADSL1	8	B	3102.02	BT 8 ADSL1	East	England	Rural	3.558533778	3.0890032	3.694687	3.3
4	943154	BT	ADSL1	8	B	3034.22	BT 8 ADSL1	East Midlands	England	Rural	3.557283673	2.925570065	3.765915	3.4
5	942412	BT	ADSL1	8	B	3934.59	BT 8 ADSL1	North West	England	Rural	1.313005372	1.268160516	1.342373	1.3
6	947276	BT	ADSL1	8	B	3404.29	BT 8 ADSL1	Northern Ireland	Northern Irelan	Rural	0.972343345	0.888351226	1.014567	0.9
7	947260	BT	ADSL1	8	B	5068.77	BT 8 ADSL1	Scotland	Scotland	Rural	1.66331283	0.828585032	2.027452	1.6

Missing data

- Missing data is common in data wrangling process
- Downstream tasks like Machine Learning assume data is complete
- Data scientists need to detect missing values, and operate on them
- Need to detecting, understanding, and imputing missing values
- Often you may not need to impute missing values

Detection

- It is common that data have missing values
- Missing values are noted differently: `Null', `NaN', `NA', or blank
- But sometime need to understand the data: missing values of numerical features may be noted as`?', `-99'

Understanding

- To understand the missing values, and classify missing values:
- 1) Missing Completely At Random (MCAR) – missingness does not depend on the data.
  - Missing values occur randomly and don’t systematically relate to any other variables in the dataset
  - The analysis performed on the data will be unbiased
  - But there might be a loss of statistical power as data sample reduced

Understanding

- To understand the missing values, and classify missing values:
- 2) Missing At Random (MAR) – missingness only depends on observed values, but not to the missing values themselves
  - Missingness is random, if you account for all the observed data
  - MAR is more realistic assumption than MCAR
  - The missing information can be predicted from other information
  - Analysis on data may or may not result in bias
  - If other information, which can predict the missing information, are accounted properly, unbiased results in analysis can be made

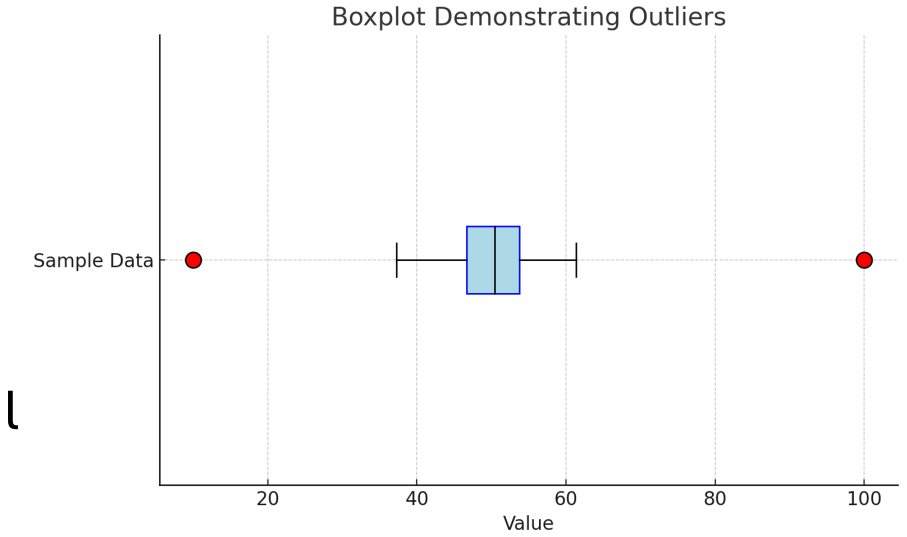
Understanding

- To understand the missing values, and classify missing values:
- 3) Missing Not At Random (MNAR) – missingness can depend on both observed and missing values
  - The data are missing systematically related to the unobserved data
  - Missingness is related to events or factors, which are not measured by researchers
  - Data analysis of a data with MNAR also may or may not result in bias

Repair

- Involves any operation performed on the missing data
- For example, removing any rows with missing entries – this is called “complete case analysis”; or substituting those missing entries with other values according to different rules – imputation
- Replacing missing values with mean or mode of the observed data
- But be careful with imputation, as you never know that ground-truth values of the missing values

Anomalies (or outliers)



- A pattern that does not conform to expected normal behaviour
- They are normally outliers in the data
- Anomalies lead to biased models, and so wrong conclusions from data
- Various causes: errors in measurements, handling errors (e.g., labelling microwave as shower), malicious activity (malicious emails), fraud (scam phone calls)
- Example:
  - In home energy data, electricity consumption of microwave (0.8kw) is labelled as shower (9.5kW)
  - Months range: 1-12, but the value 13 is anomalous
  - Tumor: malignant vs Benign

- Detecting anomalies is important, but a challenge

Detecting Anomalies

- Detecting anomalies is harder than missing values
- Approaches for anomaly detection:
    1. Supervised anomaly detection: need labels for training a detection model (classification: breast cancer – normal tissues vs abnormal)
    2. Unsupervised anomaly detection: no labels exists for anomalous data – clustering methods can be used
    3. Semi-supervised anomaly detection – a mixture of supervised and unsupervised
- Univariate vs multivariate features:
    - Outliers could be anomalies: univariate entry or multivariate feature (a row)
    - Statistical methods could be used: anomalies have low-probability events

Repair

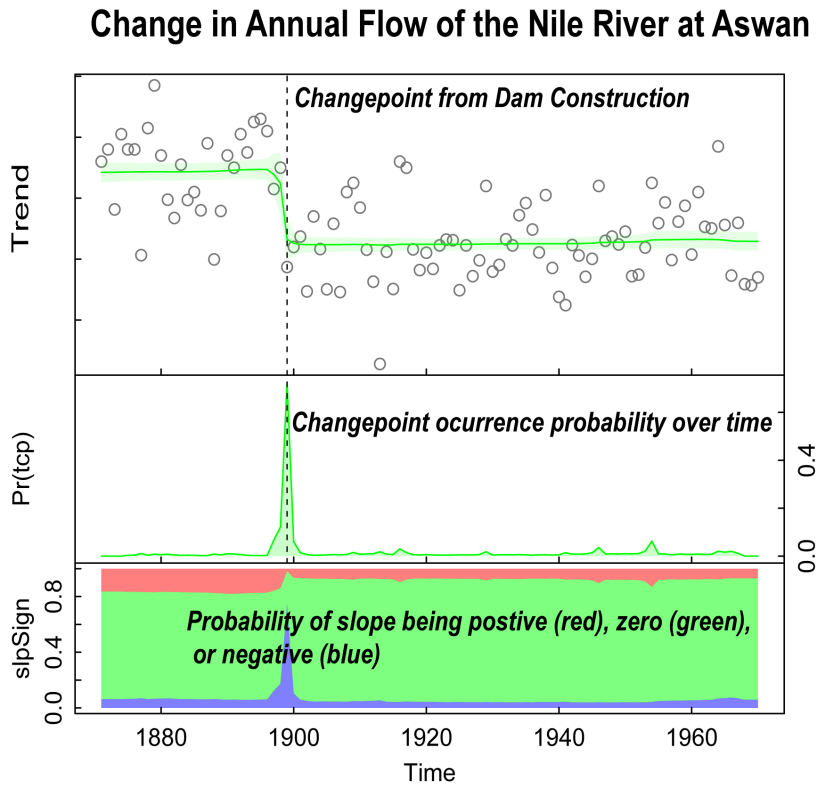
- Remove the detected anomalies completely
- Repair the data by inserting *sensible values*

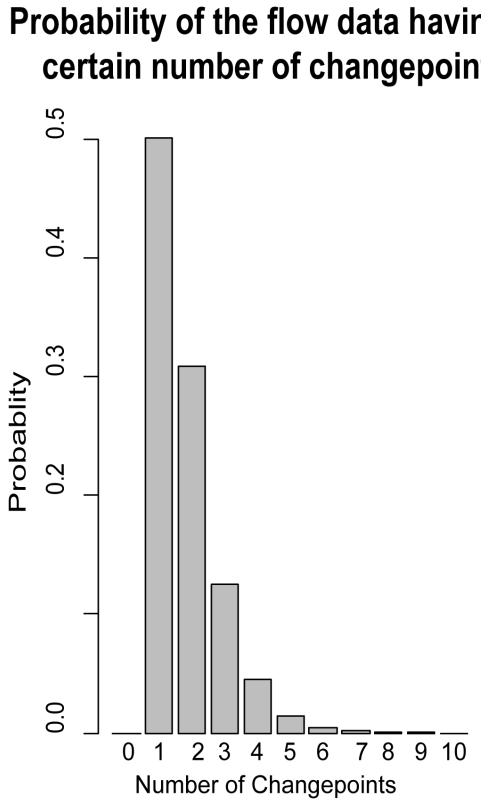
Non-Stationarity

- Change point detection (time series): Detecting any changes occurring in data distribution over time
  - Change point can be detected by using the distribution of the data: the distribution is changed after a time point
- Tabular data: non-stationary data could also happen in tabular data
  - Dataset shift: in machine learning, distributions of training data and test data are different
  - In data wrangling, data may change due to the change of measurements, e.g., the protocol of data collection can change over time (units changed, labels changed)
  - Data collection protocol is not established, for example, different people measured the data with different units

Non-Stationarity: time series

- Streaming data can change distributions over time

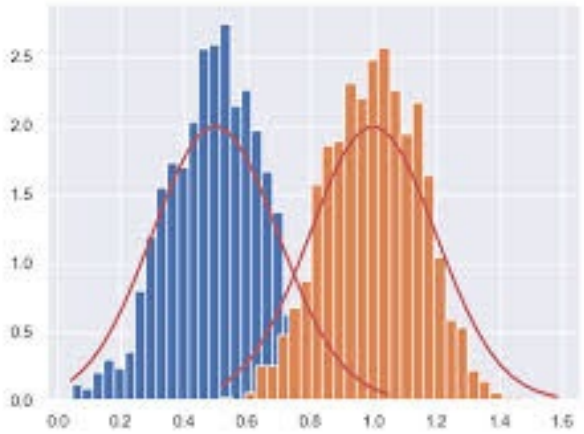




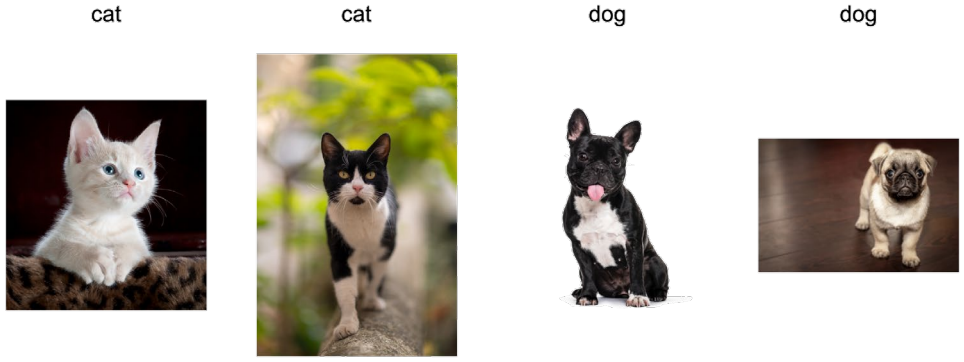
https://github.com/zhaokg/Rbeast

Non-Stationarity: tabular data

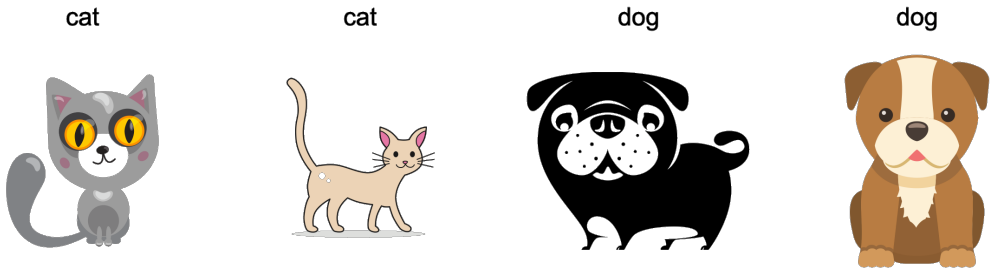
- Dataset shift
- Training data vs test data
- Distributions change



Training data



Test data



https://d2l.ai/chapter_linear-classification/environment-and-distribution-shift.html

Data duplication

- Data duplication is the situation where any variant of a piece of data duplicates somewhere in the same data infrastructure

- Data deduplication is a technique for eliminating duplicate copies of repeating data. (Wikipedia)

Types of data duplication (source: Oracle)

- Exact duplicates: Exact copies of some data records in the same data set; all fields and values are exactly same
- Near duplicates: Some records in the data set have high similarities; for example, different spellings of a word
- Data integration duplicates: combing data from different sources may lead to duplicate records – introducing exact duplicates
- Transformation duplicates: one record is transformed from other records – e.g., unit changes (m -> cm); numerical transformation (x -> exp(x))

Examples

- Exact duplicates
- Near duplicates

id	First name	Last name	email
1	Carine	Schmitt	cschmitt@gmail.com
2	Janine	Labrune	jlabrune@gmail.com
3	Janine	Labrune	jlabrune@gmail.com
4	Jonas	King	jking@gmail.com
5	Kwai	Lee	klee@gmail.com
6	Janie	Labrune	jlabrune@gmail.com

	A	B	C	D	E	F	G
1		*CustomerName*	*ProductName*	*Date*	*Qty*	*Price*	
2	Unique	BOBS UNCLE	Archies Red Chillies	2003/07/01	4	144.1290323	
3	Unique	BOTTLE SHOP	Crystal Puddle Ranberry Juice	2003/05/17	3	180	
4	Duplicate	BOTTLE SHOP	Crystal Puddle Ranberry Juice	2003/04/06	3	180	
5	Unique	BOTTLE SHOP	Crystal Puddle Ranberry Juice	2003/04/06	3	180	
6	Unique	BOTTLE SHOP	Crystal Puddle Ranberry Juice	2003/04/05	3	180	
7	Duplicate	BRONNIES CC	Captain Vital Signs Cane	2003/10/25	3	105.4064516	
8	Unique	BRONNIES CC	Captain Vital Signs Cane	2003/10/25	3	105.4064516	
9	Unique	CINDYS DELI	Paddys Rock Spice	2003/03/23	3	1755.825806	
10	Unique	FRUITY JOES	Archies Red Apples	2003/10/30	3	1516.129032	
11	Unique	JEFFS SNACKS	Bluff Finger Snack - Sausages	2003/03/31	6	972.5032258	
12	Unique	JENS JAMS	B & X Wholenuts	2002/11/09	3	1097.393548	
13	Unique	KENS TAVERN	B & X Wholenuts	2003/10/28	3	1077.096774	
14	Unique	SPICE POT	Crystal Puddle Grape Juice	2003/03/22	3	522.5677419	
15	Unique	SPICE POT	Crystal Puddle Ranberry Juice	2003/03/22	4	194.3483871	
16	Unique	TEA 4 U (DBN NRTH)	Bohemian Grapey Tea	2003/03/23	5	149.8	
17	Unique	TEA 4 U (DBN NRTH)	Crystal Puddle Ranberry Juice	2003/03/23	3	150.5032258	
18	Unique	THIRSTY KIRSTIES	B & X Wholenuts	2003/07/29	3	1077.096774	
19	Duplicate	TROTTERS CC	Bohemian Grapey Tea	2003/11/03	3	145.1612903	
20	Unique	TROTTERS CC	Bohemian Grapey Tea	2003/11/03	3	145.1612903	
21							

Examples

- Exact duplicates
- Near duplicates

Customer_ID	Name	Email	Phone	Source System
001	John Smith	john.smith@gmail.com	123-456-7890	CRM_A
002	J. Smith	john.smith@gmail.com	123-456-7890	CRM_B
003	Emily Davis	emily.davis@outlook.com	321-654-0987	CRM_A
004	Emily D.	emily.davis@outlook.com	321-654-0987	CRM_B
005	Michael Chen	michaelc@example.com	555-111-2222	CRM_A
006	Michael Chen	michael.chen@abc.com	555-111-2222	CRM_B

Examples

- Transformation duplicates

Person_ID	Height_cm	Height_m	Source
101	180	1.80	Dataset_A
102	165	1.65	Dataset_B
103	170		Dataset_A
104		1.75	Dataset_B
105	160	1.60	Merged_View