Data Quality - Notes

Nils Rechberger

2026-02-23

Tasks for Exercise 01: Introduction

Task 01

Data quality describes data characteristics at the meta-level. High-quality data provides a reliable and useful structure for information. Conversely, poor data quality can lead to mistrust and inaccurate outcomes.

> Note: Personal definition.

Task 02

Scenario 2: E-Commerce – Inventory Management

How would incomplete data affect decision-making in your scenario?

Incomplete data can lead to a misunderstanding of consumer behavior and purchasing patterns. Without a full picture, businesses may fail to identify emerging trends or customer needs.

What could go wrong if the data in your scenario is not accurate?

Inaccurate data can result in supply chain bottlenecks or an oversupply of products. This leads to either lost sales due to stockouts or increased storage costs due to excess inventory.

How could the fact, that data is not available when required, affect your scenario?

A lack of real-time data availability undermines trust in data-driven processes. If stakeholders cannot access information when needed, they may revert to "gut-feeling" decisions, which are prone to error.

In what ways does the data need to be reliable and relevant?

To support effective inventory management, data must meet specific quality standards, such as:

- Availability: Data must be accessible to decision-makers at all times.
- Velocity: Data must be processed and updated at the speed of the business.
- Completeness: No critical data points (like regional demand) should be missing.
- Usability: Data must be in a format that is easy to interpret and act upon.

  Note: The answers are not disjunct.

Tasks for Exercise 02: Data Quality Dimensions

Task 01

id	last_name	first_name	age	department	function	salary	commision_rate
1	Smith	Bill	56	Sales	Head of Sales	120.000	15%
2	Muller	John	25	Social Media	Creative Director	100.000	N/A
3	Grey	Anna	37	SEO	Google Expert	90.000	N/A
4	Berger	Lia	22	NULL	Freelancer		NULL
5	?	Mike	46	Facility	Team Manager	75.000	N/A
6	Doe	Jane	30	IT	Dev	N/A	NaN

- Scenario 1: The value exists but is not known (that is, known unknown): `last_name` of Mike (`ID` 5). Every person has a last_name, but it is currently missing from the dataset.
- Scenario 2: The value does not exist at all: `department` of Lia (`ID` 4). As an external freelancer, she is not part of the internal organizational structure.
- Scenario 3: The existence of the value is not known (that is, unknown unknown): `salary` of Jane (`ID` 6). It is unclear if she is a paid employee or an unpaid volunteer/intern; thus, the existence of the attribute itself is in question.

- Scenario 4: The attribute is not applicable: `commision_rate` for non-sales employees. This metric is only defined for sales roles and is fundamentally inapplicable to other functions.
- Scenario 5: The value is only populated under specific conditions: `salary` of Lia (`ID` 4). As a freelancer, the field remains empty until a specific hourly-based invoice or contract condition triggers the entry.

Tasks for Exercise 03: Exploratory Data Analysis

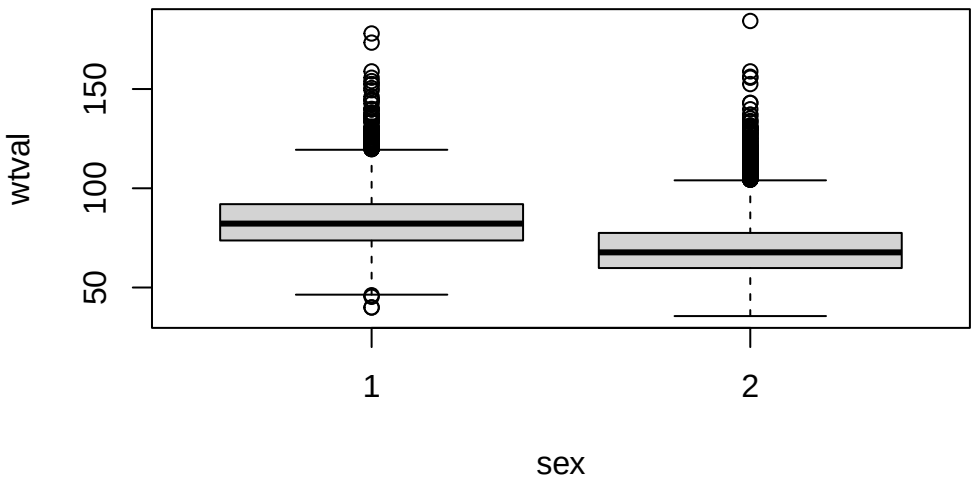
Task 01

Using R, generate a boxplot broken down by gender (variable sex). How can the boxplot be interpreted?

```r
library(readxl)

data <- read_excel("/home/nils/dev/mscids-notes/fs26/dq/data/HSE.xlsx")

boxplot(wtval ~ sex, data = data)
```



We can clearly se a difference between the two classes.

Task 02

What are the distributions of the following boxplots?

Answer

- Varriable A: Normal
- Varriable B: Right skewed
- Varriable C: Right skewed