KONVENSYEN
QA KEBANGSAAN
KALI KE-12
"Compassionate Co-production"
8 – 10 OKTOBER 2024 | KLANA RESORT, SEREMBAN

MSQH

SESI SIMPOSIUM (8 OKTOBER 2024 /8.30 - 9.30 MALAM)

SILA IMBAS KOD QR ATAU AKSES PAUTAN UNTUK PENGESAHAN KEHADIRAN

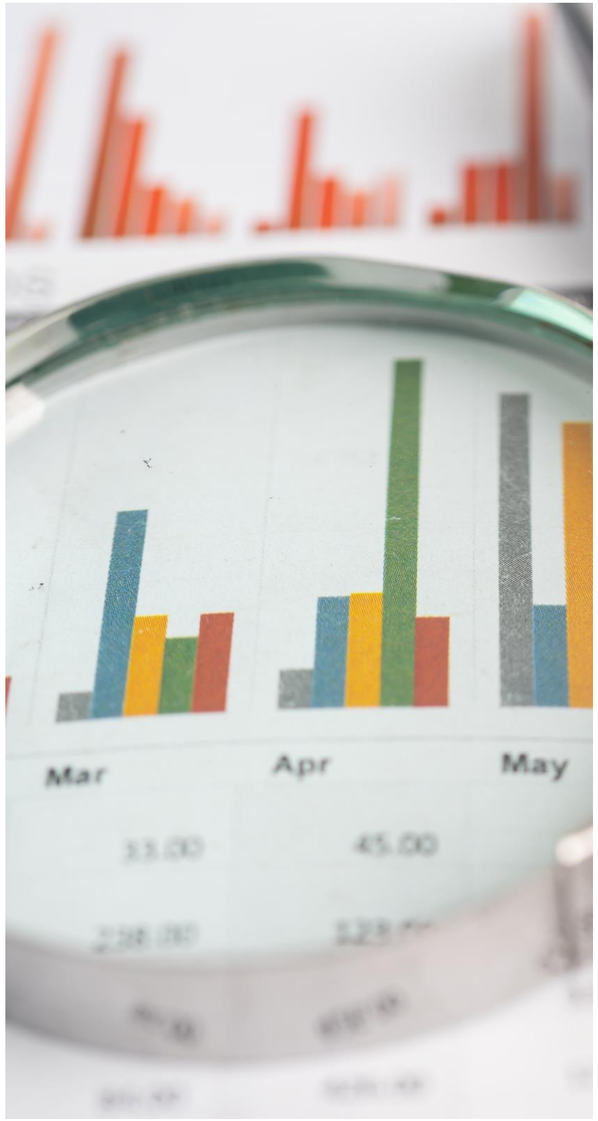
QA CONVENTION 2024 Volume 101 Tuesday,October 8,2024

POLISHING DATA

Clean data, better decisions: A hands-on approach to data cleaning

by: IHSR

Data cleaning is the process by which raw data are transformed into data that are of an appropriate quality for statistical analysis.



This process involves two key steps:

- Identifying errors and inconsistencies in the data
- Correcting and managing these data issues to ensure accuracy and reliability

Welcome to our data cleaning challenge!

Session objectives

- Identify and correct various types of data issues
- Understand the critical role of documenting data cleaning processes

In this session,
You will get to engage in a hands-on activity designed to simulate the process of data cleaning.

Possible data errors

1. Identical Records

Identical ID & identical values in all variables

Identical ID & identical values in some variables

Identical ID but different values for all other variables

2. Inconsistencies between variables

Related variables in a dataset show conflicting information

Example:

Age = 8 years old, with 5 pregnancies

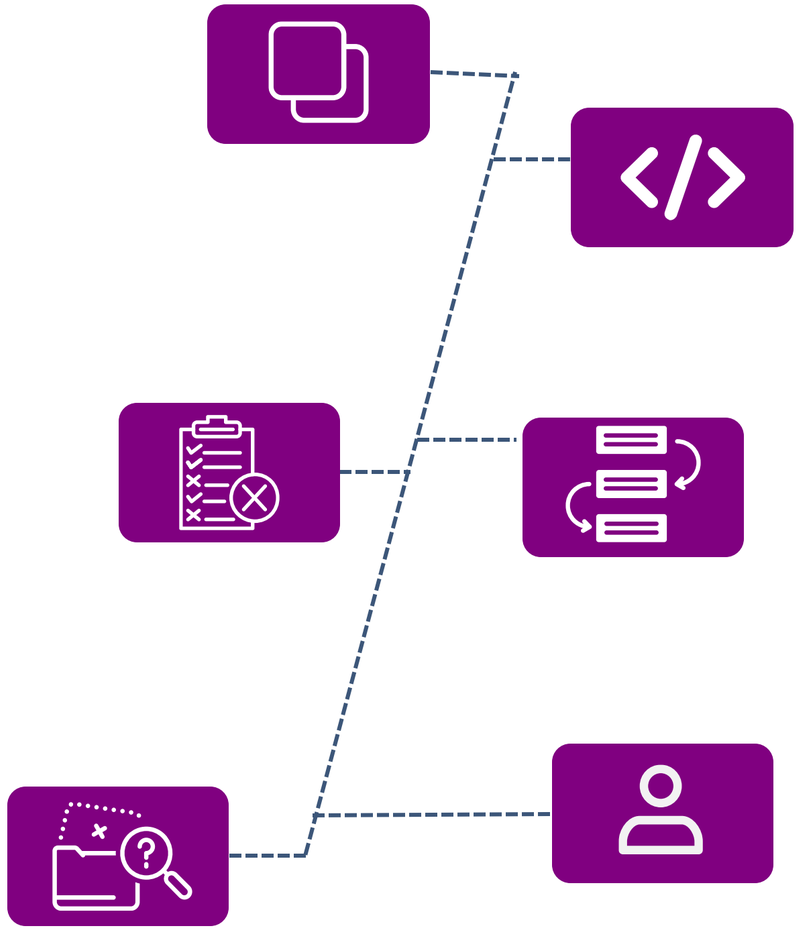
3. Extreme values

Data points that are much larger or smaller than the rest of the data.

Example of outliers:

Height 275 cm

Weight 5 kg (in study amongst elderly)



4. Code range

Occurs when an input falls outside the predefined value range of values

Example:

1 – male

2 – female

3 – means???

5. Logical sequence error

Issue in the chronological order of events

Example:

Dates are out of order.

If an end date precedes a start date

6. Data entry errors

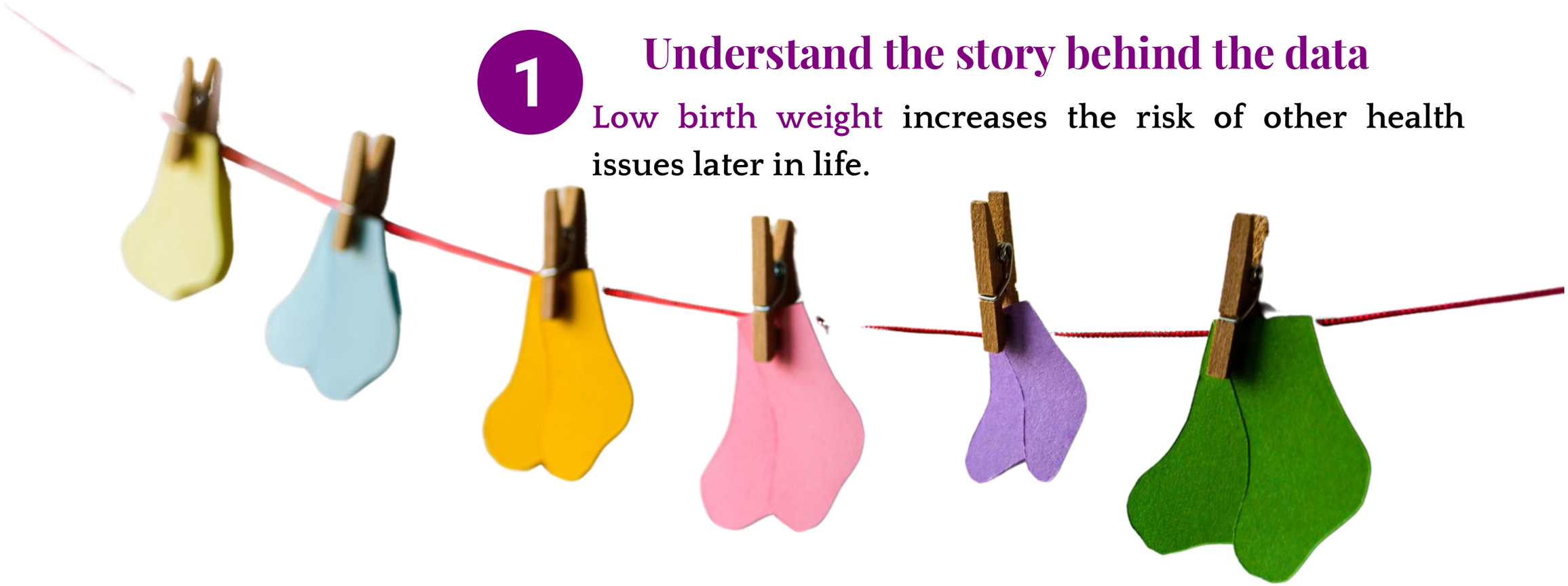
Mistakes made during the process of inputting data into a system/ database

Example:

Misspelled words

43 instead of 34

How to Play



1 Understand the story behind the data

Low birth weight increases the risk of other health issues later in life.

To address this, it is essential to understand its risk factors. This study aimed to identify those risk factors.

How to Play

2 Familiarise yourself with your BINGO box

Your Bingo card is your guide for this activity.

Each box represents a data cleaning task related to common data errors.

You will also receive a corresponding list of data errors (no. 1-25). Think of it as your checklist for spotting and fixing these issues.

How to Play

3 Examine the dataset and identify data issues

Review the dataset of 50 babies' birthweights and their mothers' information. Check for any mistakes.

Use the data dictionary. The data dictionary explains what each variable represents and how it should be formatted.

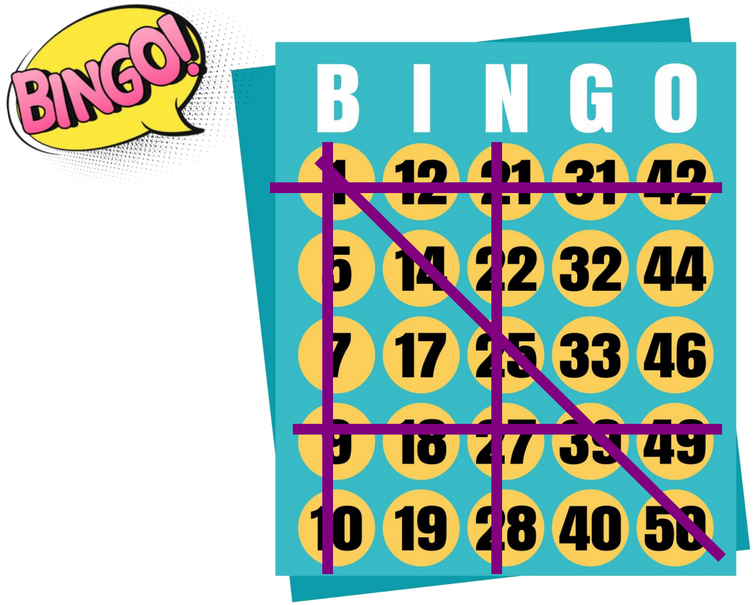
Understand the variables, their definitions, and the types of data involved.

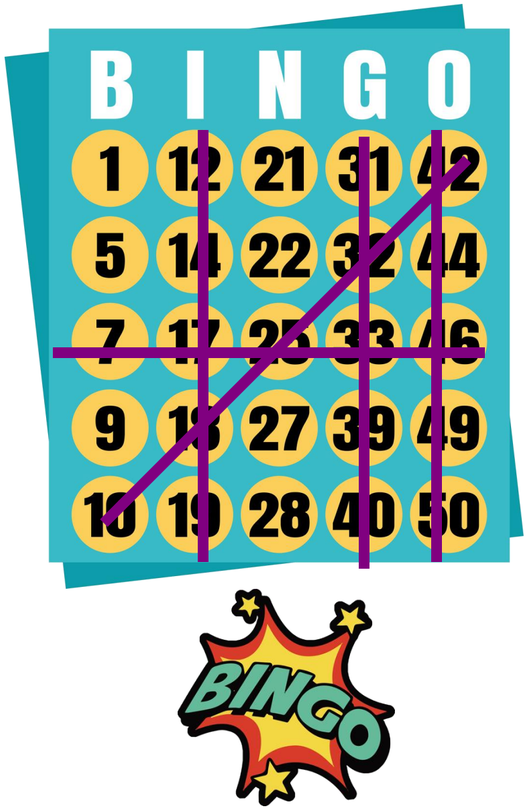
Familiarising yourself with these details will make it easier to identify inconsistencies or errors.

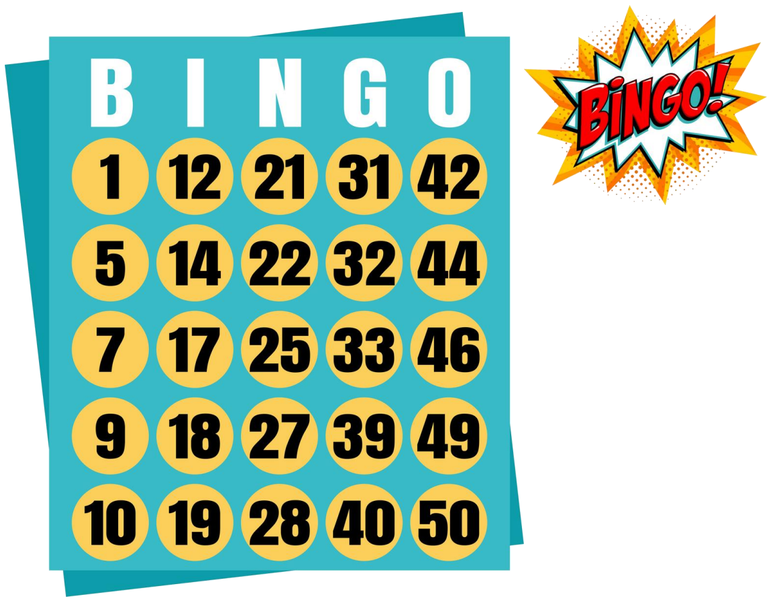
How to Play

4 Play and Call Out “BINGO!”

Your goal is to complete any FIVE lines on your Bingo card. These could be rows, columns, or diagonals.





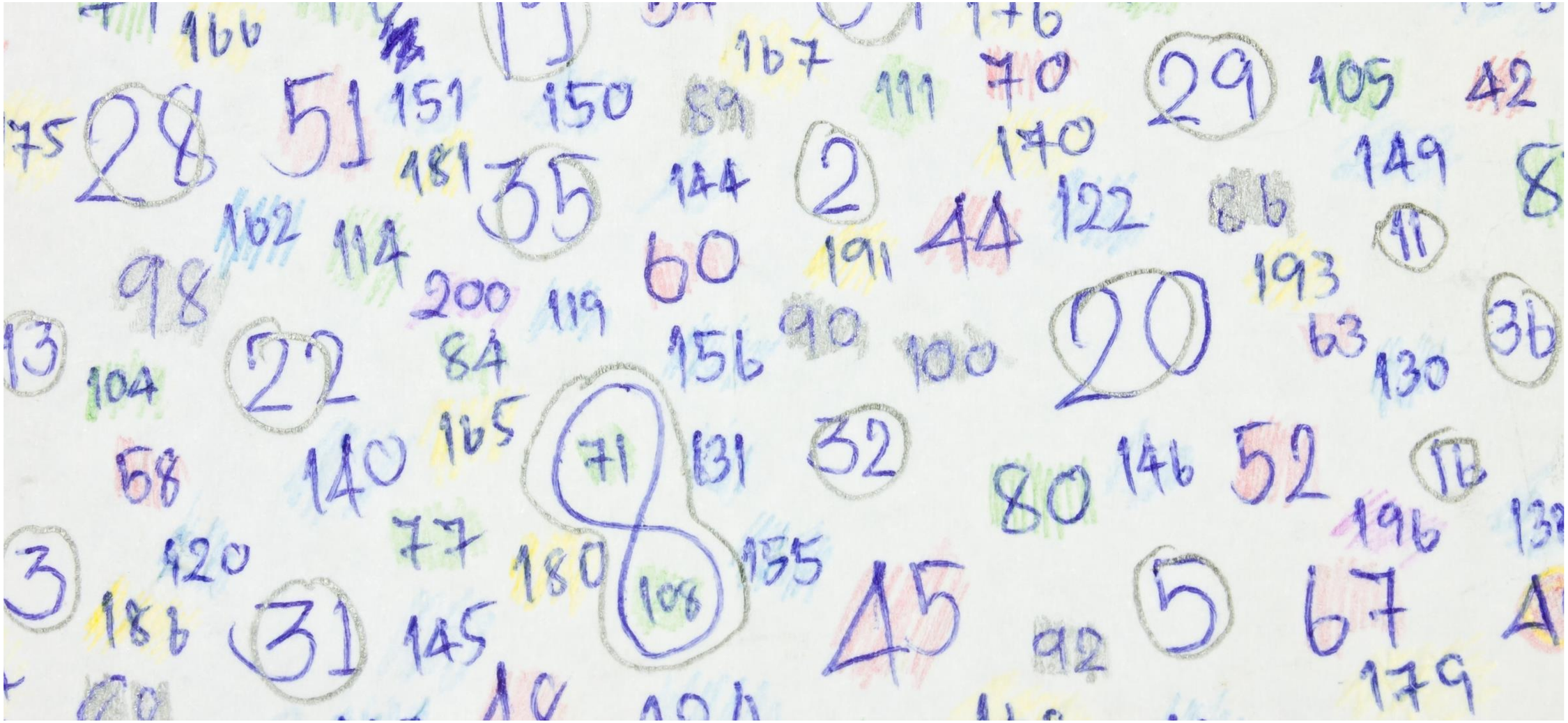


Example: Starting at Item 1... continue strategically to choose which number to select next

Let’s Begin!

Time starts now

Discussion


166
167
176
28
51
151
150
89
111
70
29
105
42
75
181
35
144
2
170
149
8
162
114
60
191
44
122
86
91
98
200
119
193
20
3b
13
104
84
156
90
100
63
130
22
165
71
131
32
80
146
52
15
68
140
77
180
108
155
196
3
120
186
31
145
45
92
5
67
179

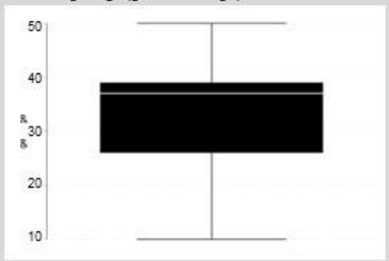
Data Examination & Error Identification

Visualisation approach

Figure 1: Box plot distribution of key variables (Panels 1-4)

Panel 1: weight_kg (infant weight)

Panel 2: gesAge (gestational age)



Panel 3: mom_age (mother's age)

Panel 4: no_preg (Number of pregnancies)

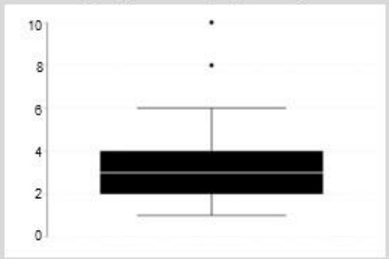
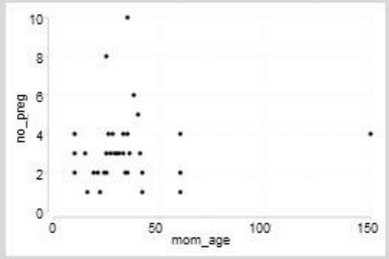
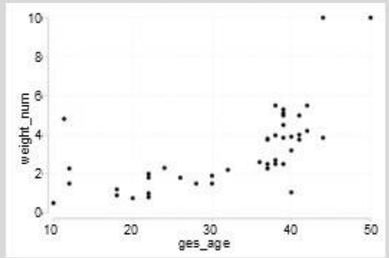


Figure 2: Scatter plots of key variables (Panels 5-6) and box plot by infant outcome (Panel 7)

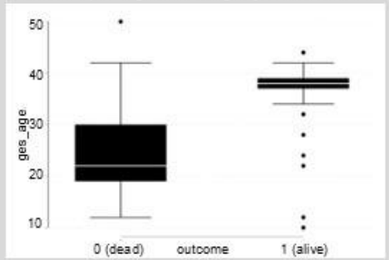
Panel 5: Scatter plot of maternal age & no. of pregnancies



Panel 6: Scatter plot of gestational age & infant weight



Panel 7: Box plot of gestational age by infant outcome



Descriptive approach

Table 1: Descriptive statistics for baby's sex, race, and residential area (n=50)

	Freq.	Percent
baby_sex		
1	26	50.98
2	21	41.18
3	2	3.92
4	1	1.96
missing	1	1.96
race		
bknwarga	1	1.96
chinese	14	27.45
indian	13	25.49
malay	19	37.25
non-citizen	2	3.92
other	1	1.96
missing	1	1.96
resident_area		
1	30	58.82
2	12	23.53
3	1	1.96
kapit	1	1.96
kl	2	3.92
kuching	1	1.96
pendang	1	1.96
missing	1	1.96

Table 2: Summary statistics for key variables

	Obs	Mean	Std. dev.	Min	Max
weight_kg	48	3.15	2.05	0.5	10
gesAge	50	33.26	9.41	10	50
momAge	50	33.04	21.81	10	150
no_pregnant	48	2.98	1.68	1	10

Table 3: Cross tabulation of ever smoked by current smoking status

	mCurrentSmoke		
mEverSmoke	0	1	missing
0	36	3	0
1	1	5	0
missing	0	0	6

Table 4: Inconsistent smoking status records

id	mEver_smoke	mCurrentSmoke	noCig_day
004	0	1	missing
013	1	1	missing
043	0	1	missing

Solutions for Data Cleaning Challenge (Part 2)

17. Check for any categories that are unexpected or wrongly labelled (e.g., malay, chinese, indian, others).
    - Row 11 has "bknwarga: instead of a coded race **(code range;** see Table 1).

18. Identify instances where free text has been used for observations that should have specific options (e.g., 1 for urban and 2 for rural in residential area data).
    - Rows 46- 50 free text entries like "kapit" and "kl" in column O, which should follow the predefined codes for residential areas **(data entry errors;** see Table 1).

19. Check for **logical consistency** between ever smoked and current smoking status.
    - if "ever smoked" is marked as 0 in row 7, "current smoke" should not be 1.

20. Check that the number of cigarettes field is appropriately filled or left blank when it should be.
    - For rows where "current smoke" is marked as 0 (e.g., row 3), the number of cigarettes in column P should be blank (see Table 3 and 4).

21. Check that smoking status **matches** the number of cigarettes reported.
    - If "current smoke" in column N is marked as 1, there should be a non-zero number of cigarettes reported in column P (e.g., row 5 where the smoking status and cigarette number may be inconsistent; Table 3 and 4).

22. Review **logical consistency** between all smoking-related variables.
    - For example, row 12 has a smoking status of "current smoke" as 1, but the number of cigarettes is 0, which needs correction (see Table 3 and 4).

23. Identify any **unnecessary spaces or special characters** in the dataset.
    - Row 18 contains unnecessary characters, such as '!', and row 15 includes unnecessary spaces (e.g., ' alive').

24. Check that all key variables are **complete** (e.g., study outcome and other main variables).
    - Rows 49 and 50 are missing outcome data, and rows 40 and 41 are missing gestational age data. All missing data in these columns should be reviewed and addressed.

25. Locate any rows in the dataset that are completely empty or only have a few observations filled.
    - Row 51 is completely empty, and row 50 has many **missing** fields. These rows should be flagged and investigated for potential removal or further clarification

Solutions for Data Cleaning Challenge (Part 1)

1. Identify rows where all the data is exactly the same.
    - Review rows 1 and 2, where the ID in column A is different, but all other data are **identical.** These rows should be flagged for further investigation and potential correction.

2. Check if the same ID has different information.
    - In row 41 and row 42, ID 040 is **duplicated** but contains different information. Ensure the data consistency for this ID and correct as needed.

3. Identify dates that are in the wrong format.
    - Rows 3, 4, and 5 contain dates that do not follow the expected DD/MM/YYYY **date format.** For example, row 3 has "21-Jan-22"

4. Locate any dates that do not exist in the calendar.
    - Row 6: "31/4/2023" would be **invalid date** and should be corrected.

5. Check if the delivery dates are within the data collection period (2022–2023).
    - Review rows like 38 and 39, which contain dates in the year "2025," **outside of range** from the expected data collection period.

6. Check if all birth weights are recorded as numerical values.
    - **Non-numerical values** such as "abc" in row 4 should be corrected to ensure all birth weights in column F are numerical.

7. Look for birth weights that are too high or too low.
    - Review entries like row 3, where a birth weight of "10 kg" is entered, which seems excessively high. These **outliers** should be verified or corrected. See also Panel 1 (page 6) and Table 2 (page 7).

8. Check that the birth weight **category** (low birth weight, normal weight, macrosomia) aligns with the weight in kilograms.
    - in row 18, a birth weight of 2.27kg is marked as "normoweight", which should be corrected.

9. Locate gestational ages that fall outside of the expected range (i.e., less than 20 weeks or more than 44 weeks).
    - Entries like row 3, with a gestational age of 50 weeks, should be flagged as **extreme value.** See also Panel 2 (page 6) and Table 2 (page 7).

10. Check that the gestational **category** (premature or not premature) **matches** the number of weeks of pregnancy.
    - in row 5, a gestational age of 18 weeks is marked as "not premature," which should be corrected.

11. Check if the baby's weight is appropriate for the number of weeks of pregnancy
    - Row 15 reports a baby at 11 weeks gestation with a weight of 4.99 kg (see Panel 6)

12. Check for **consistency** between birth weight, gestational age, and outcome.
    - Row 47 reports a birth weight of 0.5 kg at 10 weeks gestation with the outcome "alive", which is biologically improbable (see Panel 7).

13. Identify any maternal age values that are **biologically implausible**.
    - Row 3 has a maternal age of 150 years, which is biologically impossible (see also Panel 3 and Table 2)

14. Check that maternal age and number of pregnancies are consistently treated as continuous variables (numbers).
    - Row 12 lists "ten" as the number of pregnancies, which should be corrected to a number **(data entry errors).**

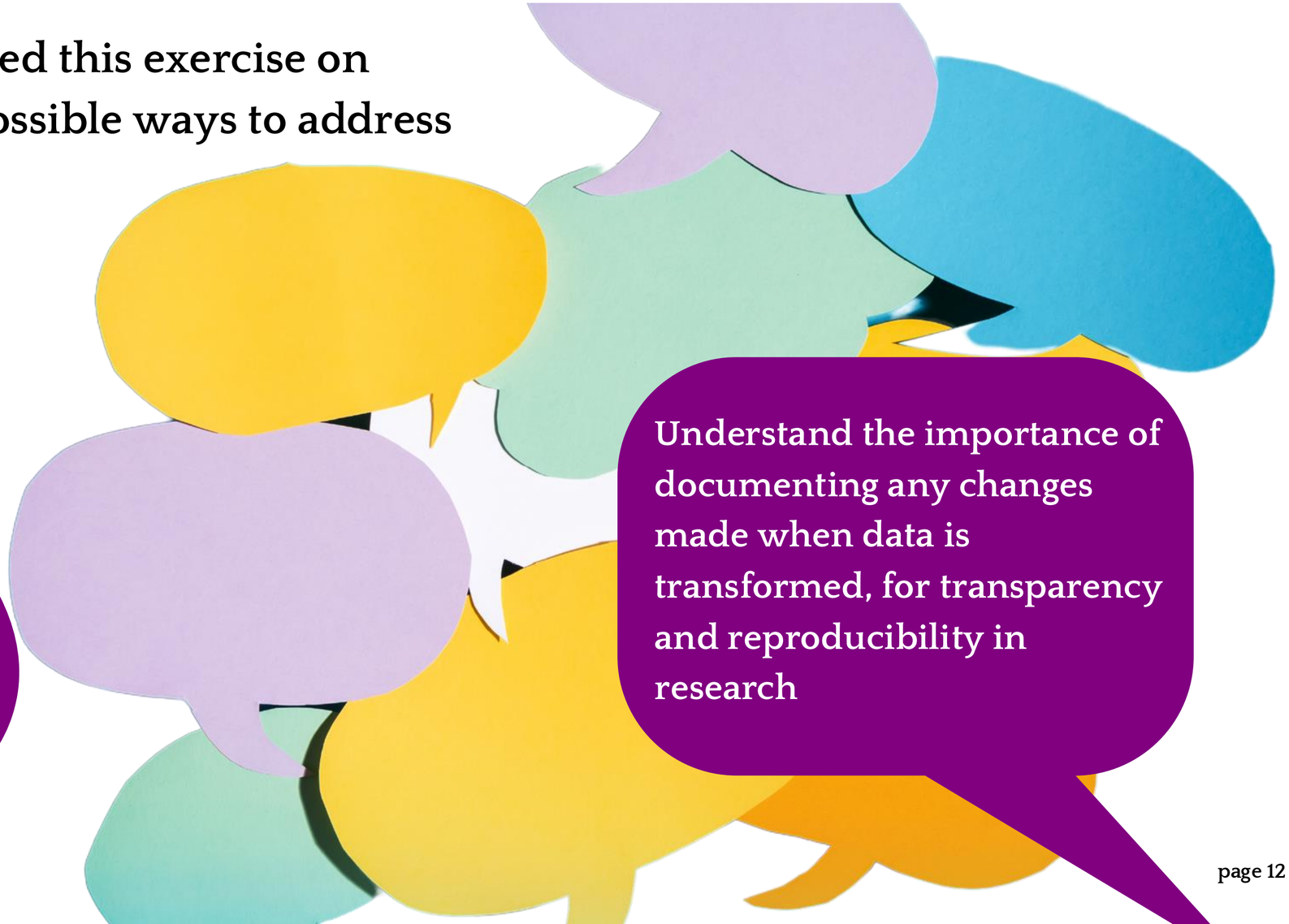
15. Verify that the number of pregnancies is **consistent** with the mother's age and biologically plausible.
    - Row 9 reports a mother's age as 10 years with 4 pregnancies (see Panel 4 and 5).

16. Review observations that fall outside the expected range for categorical variables (e.g., expected codes are 1-2 for sex of baby).
    - Row 11 reports the baby's sex as 3, and Row 27 as 4, both of which are outside the expected range **(code range).** See also Table 1 (Page 7)

Session Outcomes

We hope you have enjoyed this exercise on identifying data issues and possible ways to address them

&



Understand the potential impact of data errors on a study's outcomes

Understand the importance of documenting any changes made when data is transformed, for transparency and reproducibility in research

KEMENTERIAN KESIHATAN MALAYSIA
INSTITUT KESIHATAN NEGARA

Vision

To position NIH as a leading health research organisation towards enhancing the health and wellbeing of the nation.

Mission

- To conduct effective and high impact health related research, training and consultancy to improve quality of life.
- Govern and manage research in the country to address national health priorities.

INSTITUTE FOR HEALTH SYSTEMS RESEARCH

Mission

To provide high quality evidence and expertise in health policy and systems research.

Goal

Advancing nation's health through health policy and systems research.

Get in touch with us!

Dr Awatef Amer Nordin
awatef.an@moh.gov.my
Centre for Health Equity Research, IHSR

Dr Diane Chone Woei Quan
chong.dwq@moh.gov.my
Centre for Health Services Research, IHSR

KONVENSYEN
QA KEBANGSAAN
KALI KE-12
"Compassionate Co-production"
8 - 10 OKTOBER 2024 | KLANA RESORT, SEREMBAN

BORANG PENILAIAN SIMPOSIUM

SILA IMBAS KOD QR UNTUK BORANG PENILAIAN SIMPOSIUM

Your feedback matters greatly to us!

Some information resources