For the best hands-on experience, it’s ideal to print it on A3 paper, but A4 works fine if your eyesight is sharp.

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

Session objectives

By the end of the session, the participants will be able to:

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

Contents

Data cleaning

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 & inconsistencies in the data
- Correcting & managing these issues to ensure accuracy & reliability

Learning activities

- Hands-on practice
- Discussion

Duration

- One hour

To access additional resources, scan the QR code:

We value your feedback. Kindly scan the QR code to share your thoughts:

For any inquiries, feel free to reach out to:
Awatef awatef.an@moh.gov.my
Diane chong.dwq@moh.gov.my

Possible data errors

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

Code range

- Occurs when an input falls outside the predefined value range of values
- Example:
  1 – male
  2 – female
  3 – means???

Inconsistencies between variables

- Related variables in a dataset show conflicting information
- Example:
  - Age = 8 years old, with 5 pregnancies

Logical sequence error

- Issue in the chronological order of events
- Example:
  - Dates are out of order.
  - If an end date precedes a start date

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)

Data entry errors

- Mistakes made during the process of inputting data into a system/ database
- Example:
  - Misspelled words
  - 43 instead of 34

DATA CLEANING CHALLENGE

"When you have marked off five squares in a row, call out "Bingo!" and get ready to share what you found."

Ensuring Data Integrity for QA/QI Initiatives

By Awatef Amer Nordin & Diane Chong Woei Quan

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.

2 Familiarise with the Bingo box & error list

Your Bingo box 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.

3 Examine dataset & identify data issues

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

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.

Example:

Item 1 on your Bingo box is "Duplicate entry".

As you eyeball the dataset, if you notice that a particular entry appears more than once with the exact same information, you've identified a duplicate.

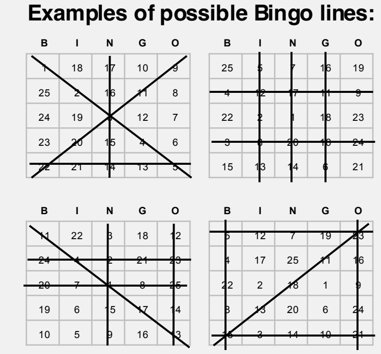
Mark off this square on your Bingo card.

Then, move on to another square of your choice.

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.

Examples of possible Bingo lines:



DATA DICTIONARY for Baby Birth Statistics (2022-2023)

Key principles

- Defines variables for **clear understanding**
- **Consistency** in data handling & structure
- **Reduces mistakes** by outlining acceptable values, ranges & categories for each variable
- **Reference guide** for current & future data users

Variable name	Variable label	Variable definition	Variable type
id	ID of case	Unique identifier	Categorical
baby_sex	Baby sex	1= male 2= female	Categorical
race	Race group	malay chinese indian others	Categorical
outcome	Baby alive or dead	1= alive 0= dead	Categorical

Variable name	Variable label	Variable definition	Variable type
deliverydate	Date of delivery. Data was collected 2022-2023	Date in DD/MM/YYYY	Date
weight_kg	Birth weight of baby in kilogram	Weight of baby at birth in kg	Continuous
weight_gp2	Birth weight of baby in groups. Low birth weight (LBW) refers to weight of <2500 grams regardless of gestational age; macrosomia refers to the body weight reaching or exceeding 4000 grams	lbw normoweight macrosomia	Categorical
gesAge	Gestational age (number of weeks of pregnancy)	Number of weeks	Continuous
gesCat	Gestational age in category. Gestational age of 37 weeks or more is classified as not premature	premature not premature	Categorical

Variable name	Variable label	Variable definition	Variable type
momAge	Mother's age at delivery	Age of mother in years	Continuous
no_pregnant	Pregnancy number	Number of pregnancies	Continuous
resident_area	Residential area	1 = urban 2 = rural	Categorical
mEver_smoke	Mother's smoking status (ever smoked)	1= yes 0= no	Categorical
mCurrentSmoke	Mother's current smoking status	1= yes 0= no	Categorical
noCig_day	Number of cigarette smoked in a day (current)	Number of cigarettes currently smoked per day	Continuous

List of data errors

1. Identify rows where all the data is exactly the same.
2. Check if the same ID has different information.

Note: Errors 3–5 may be reviewed together.

3. Identify **dates** that are in the wrong format.
4. Locate any dates that do not exist in the calendar.
5. Check if the delivery dates are within the data collection period (2022–2023).

Note: Errors 6–12 may be reviewed together.

6. Check if all **birth weights** are recorded as numerical values.
7. Look for birth weights that are too high or too low.
8. Check that the birth weight category (low birth weight, normal weight, macrosomia) aligns with the weight in kilograms
9. Locate **gestational ages** that fall outside of the expected range (i.e., less than 20 weeks or more than 44 weeks).
10. Check that the **gestational category** (premature or not premature) matches the number of weeks of pregnancy.
11. Check if the baby's weight is appropriate for the **number of weeks of pregnancy**
12. Check for consistency between birth weight, gestational age, and outcome.

Note: Errors 13–15 may be reviewed together.

13. Identify any **maternal age** values that are biologically implausible.
14. Check that maternal age and number of pregnancies are consistently treated as continuous variables (numbers).
15. Verify that the number of pregnancies is consistent with the mother's age and biologically plausible.
16. Review observations that fall outside the expected range for categorical variables (e.g., expected codes are 1-2 for **baby sex**).

B I N G O

B	I	N	G	O
5	12	7	19	23
4	17	25	11	16
22	2	18	1	9
8	13	20	6	24
15	3	14	10	21

17. Check for any categories that are unexpected or wrongly labelled (e.g., **race** malay, chinese, indian, others).
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).

Note: Errors 19–22 may be reviewed together.

19. Check for logical consistency between **ever smoked** and **current smoking** status.
20. Check that the **number of cigarettes** field is appropriately filled or left blank when it should be.
21. Check that smoking status matches the number of cigarettes reported.
22. Review logical consistency between all smoking-related variables.
23. Identify any unnecessary spaces or special characters in the dataset.
24. Check that all key variables are complete (e.g., study outcome and other main variables).
25. Locate any rows in the dataset that are completely empty or only have a few observations filled.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
	id	baby_ sex	race	deliverydate	weight_kg	weight_gp2	gesAge	gesCat	outcome	momAge	no_ pregnant	resident_ area	mEver_ smoke	mCurrent smoke	noCig_ day
1	001	1	malay	27/7/2022	2.27	lbw	37	not premature	1	25	2	1	0	0	6
2	002	1	malay	27/7/2022	2.27	lbw	37	not premature	1	25	2	1	0	0	6
3	003	2	chinese	21-Jan-22	10	normo-weight	50	not premature	0	150	4	2	0	0	
4	004	2	indian	31-Jan-22	abc	lbw	34	not premature	1	60	ten	1	0	1	
5	005	1	non-citizen	23-Jan-22	1.2	lbw	18	not premature	0	42	1		1	1	12
6	006	1	malay	31/4/2023	5.5	macrosomia	38	not premature	1	15	3	1	0	0	
7	007	1	malay	1/2/2023	1.05	normoweight	40	not premature	1	35	2	1	0	0	
8	008	2	indian	20/6/2022	3.75	normoweight	37	Premature	1	25	3		0	0	
9	009	1	chinese	15/1/2022	2.27	lbw	37	not premature	1	10	4	1	0	0	
10	010	1	malay	22/5/2023	2.5	normoweight	39	not premature	1	25	2	2	0	0	
11	011	3	bknwarga	12/8/2022	3.97	normoweight	38	not premature	1	60	4	1	0	0	
12	012	1	malay	14/9/2023	0.75	lbw	20	premature	0	18	ten	2	1	0	6
13	013	2	chinese	31/12/2022	1.5	lbw	30	premature	0	10	3	1	1	1	
14	014	1	indian	29/10/2022	3.85	normoweight	39	not premature	!	25	2	2	0	0	
15	015	2	malay	4/11/2023	4.99	normoweight	11	not premature	alive	35	2	1	0	0	
16	016	1	indian	15/1/2023	10	macrosomia	44	not premature	dead	60	1	2	0	0	
17	017	2	chinese	5/12/2022	0.8	lbw	22	not premature	o	25	8	1	0	0	
18	018	2	malay	25/3/2023	2.27	normoweight	37	not premature	!	10	4	2	0	0	
19	019	2	other	14/6/2022	5.5	macrosomia	42	not premature	0	42	2	1	0	0	
20	020	2	malay	7/8/2022	1.5	lbw	12	premature	0	25	3	2	0	0	
21	021	2	chinese	17/9/2022	2.5	normoweight	37	not premature	1	60	2	1	0	0	
22	022	1	indian	20/10/2023	3.75	normoweight	41	not premature	1	35	4	2	0	0	
23	023	1	malay	25/11/2023	2.27	lbw	12	premature	1	10	3	1	1	1	20
24	024	1	chinese	30/7/2022	4.99	normoweight	39	not premature	1	25	2	1	1	1	1
25	025	2	indian	29/12/2022	3.85	macrosomia	44	not premature	1	60	2	1	0	0	25
26	026	3	non-citizen	15/3/2023	0.8	lbw	22	not premature	0	35	10	1	0	0	
27	027	4	malay	25/6/2023	2.5	normoweight	38	not premature	1	10	2	2	0	0	
28	028	2	chinese	1/12/2022	3.75	normoweight	37	not premature	1	25	3	2	0	0	
29	029	1	indian	11/7/2023	1.5	lbw	28	premature	1	35	4	1	0	0	
30	030	1	malay	15/9/2023	1.8	lbw	22	premature	0	60	2	1	0	0	
31	031	2	malay	25/10/2023	2.3	lbw	24	premature	1	22	1	1	0	0	
32	032	1	chinese	5/12/2022	5.5	macrosomia	38	not premature	1	33	3	1			
33	033	2	indian	15/3/2023	1.9	lbw	30	premature	0	38	6	2			
34	034	1	malay	29/8/2023	2	lbw	22	premature	1	19	2	3	0	0	
35	035	1	chinese	22/7/2022	3.8	normoweight	37	not premature	1	27	3	1	0	0	
36	036	2	indian	14/2/2023	0.9	lbw	18	premature	0	16	1	2	0	0	
37	037	1	malay	11/11/2022	4.2	macrosomia	42	not premature	1	28	4	1	0	1	12
38	038	1	chinese	21/9/2025	2.6	normoweight	36	not premature	1	24	2	1	0	0	
39	039	2	indian	17/4/2025	3.2	normoweight	40	not premature	1	31	3	1	0	0	
40	040	2	malay	13/6/2025	2.2	lbw		premature	1	21	2	1	0	0	
41	040	1	chinese	10/3/2023	5.1	macrosomia		not premature	1	29	3	1	0	0	
42	041	1	indian	1/12/2022	1	lbw	22	premature	0	34	2	1	0	0	
43	042	2	malay	28/9/2022	2.7	normoweight	38	not premature	1	26	4	1	0	1	
44	043	2	chinese	6/7/2023	4	normoweight	41	not premature	1	30	3	1	0	0	
45	044	1	indian	19/5/2022	1.8	lbw	26	premature	0	40	5	1			
46	045	1	malay	15/8/2023	5.3	macrosomia	39	not premature	1	33	4	kl			
47	046	2	chinese	18/2/2023	0.5	lbw	10	premature	1	41	3	kapit			
48	047	2	indian	23/7/2023	4.5	normoweight	39	not premature	1	22	1	pendang	0	0	
49	048	1	malay	7/1/2023	3.9	normoweight	40	not premature		36	3	kl	0	0	
50	049	1	chinese	14/4/2023			38	not premature		25	2	kuching	1	1	15
51	050														

QA Convention 2024 Polishing Data

Date
Tuesday, October 8, 2024

Appendix 1

Dataset
Risk factors for low birth weight

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.
    - Row 41-42: 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.
    - Row 5: 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).

Data Examination & Error Identification

Visualisation approach

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

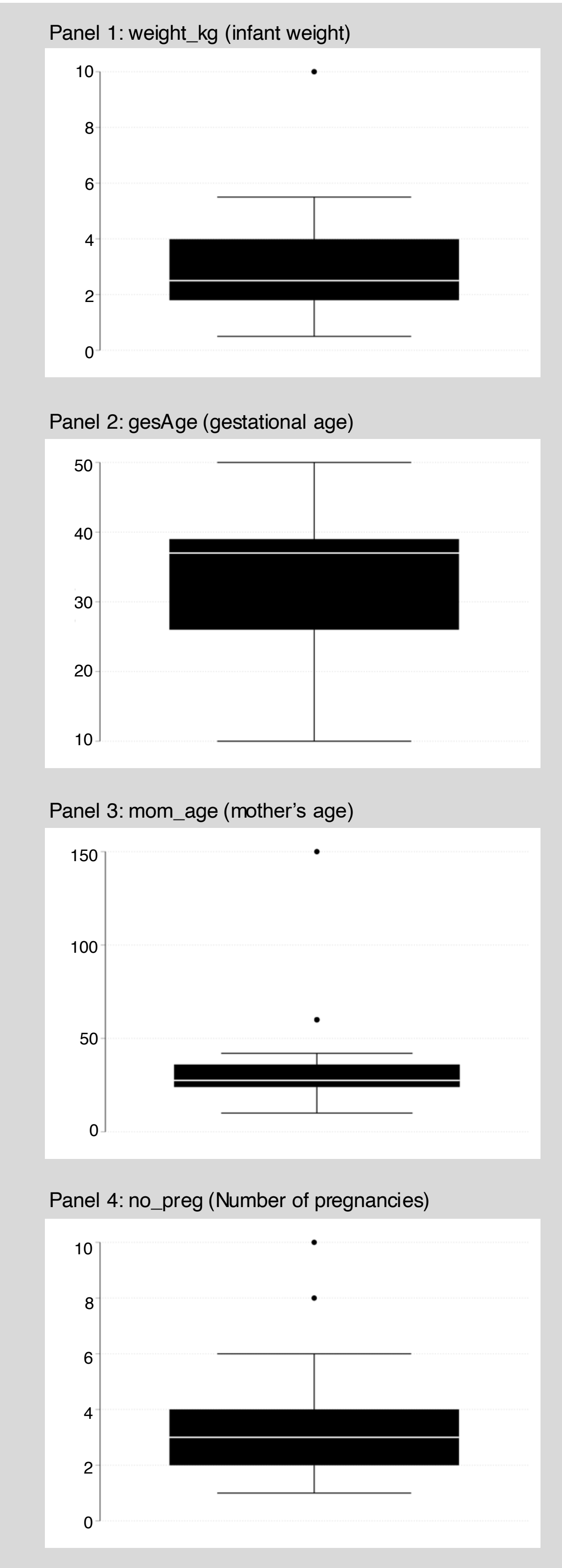
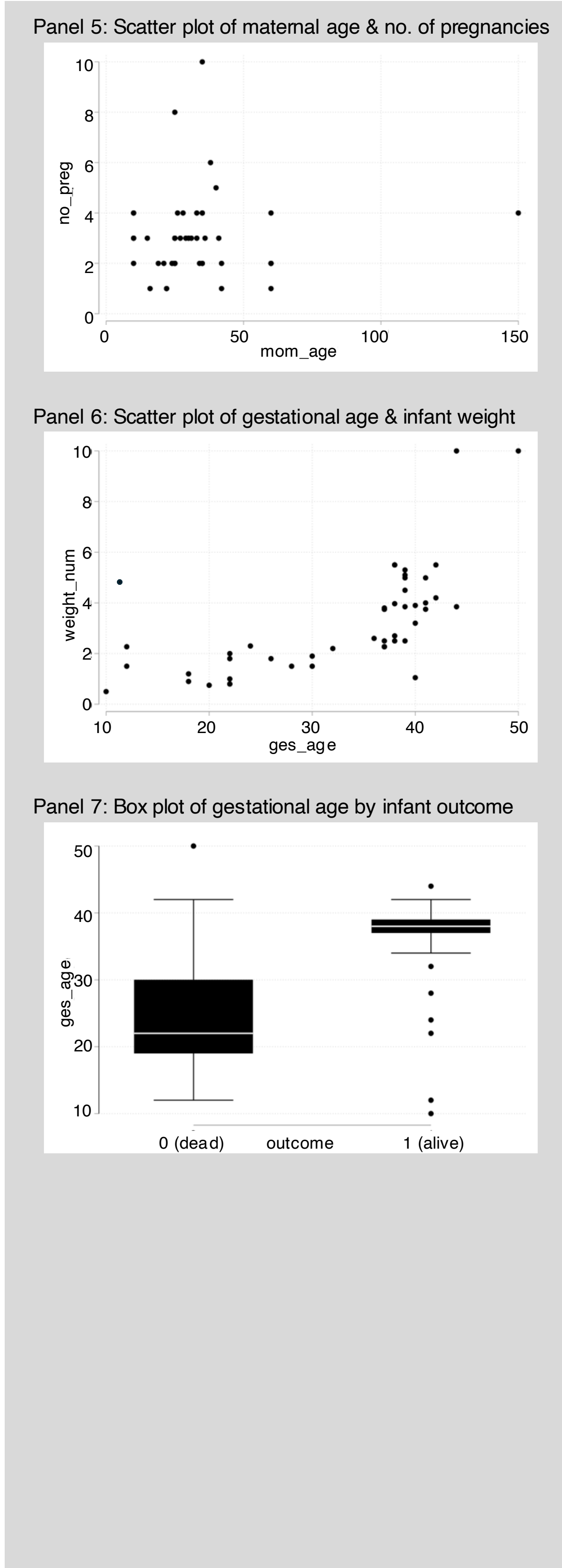


Figure 2: Scatter plots of key variables (Panels 5-6) and box plot by infant outcome (Panel 7)



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