

Data Preparation

Cleaning data before analysis

การเตรียมข้อมูลก่อนการวิเคราะห์



Raw data is
often messy.

ข้อมูลดิบมักไม่สมบูรณ์

Before analysis, data should be:



checked



cleaned



organized



ready to use



Good analysis starts with clean data.

การวิเคราะห์ที่ดีเริ่มจากข้อมูลที่สะอาด



Collect Data



Prepare Data



Analyze Data



Make Decision

From Data Collection to Data Preparation

Collected data is not always ready.

ข้อมูลที่เก็บมาอาจยังไม่พร้อมใช้งาน

Previous lessons:

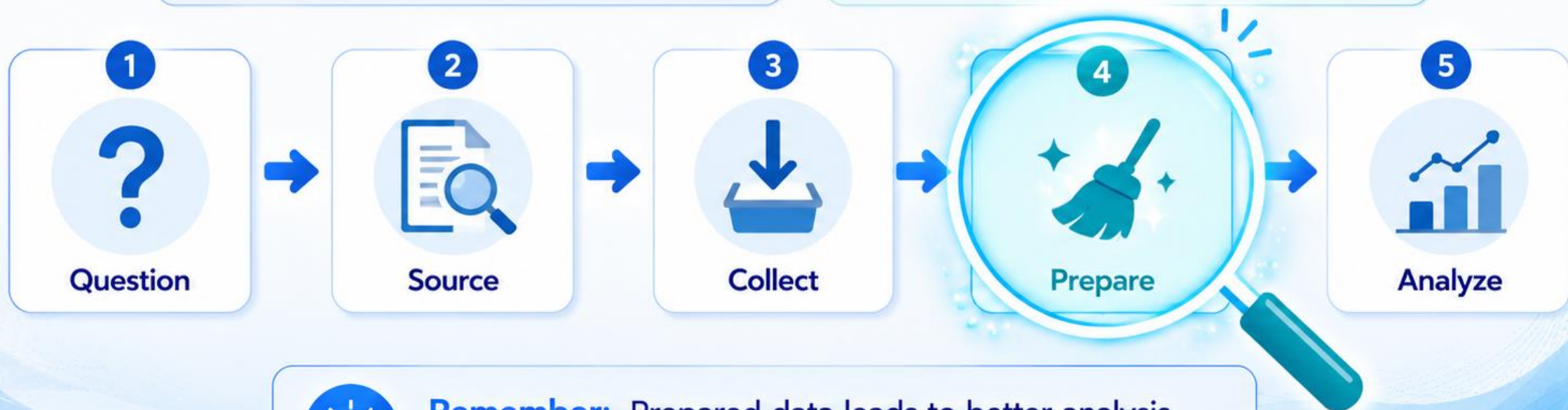


- collect data
- find data sources
- check source correctness

Today:



Prepare the data
before using it.
เตรียมข้อมูลก่อนนำไปใช้



Remember: Prepared data leads to better analysis.
อย่าลืม: ข้อมูลที่เตรียมอย่างดีช่วยให้การวิเคราะห์แม่นยำขึ้น

Why Do We Prepare Data?

Messy data can create wrong results.

ข้อมูลที่ไม่เรียบร้อยอาจทำให้ผลลัพธ์ผิด

We prepare data to:

1



remove mistakes

ลบข้อผิดพลาด

2



fill or handle
missing values

จัดการข้อมูลที่หายไป

3



make formats
consistent

ทำรูปแบบให้เหมือนกัน

4



make data easier
to analyze

ทำให้วิเคราะห์ง่ายขึ้น



Reminder:

GIGO = Garbage In, Garbage Out

Raw Data vs Prepared Data

Before and after cleaning

ก่อนและหลังการทำความสะอาดข้อมูล



Raw Data

Messy, incomplete, inconsistent

ID	Name	Age	Score	Date
1	Alice	20	85	5/1/2024
2	bob		Seveny	01-05-24
3	Charlie	19 years	90	2024/5/1
4	Alice	20	85	5/1/2024
5	Eve			5-2-24
six	Frank	21	88	02/05/2024
7	Grace	18	eighty	5/3/24
8	Heidi	twenty two	92	

- collected first
- may have mistakes
- may be incomplete
- may be hard to analyze



Prepare



Prepared Data

Clean, complete, consistent

ID	Name	Age	Score	Date
1	Alice	20	85	2024-05-01
2	Bob	19	70	2024-05-01
3	Charlie	19	90	2024-05-01
4	Eve	21	82	2024-05-02
5	Frank	21	88	2024-05-02
6	Grace	18	80	2024-05-03
7	Heidi	22	92	2024-05-03

- checked
- cleaned
- organized
- ready for analysis



Common Data Problems

What can be wrong with raw data?

ข้อมูลดิบอาจมีปัญหอะไรบ้าง?

1



Missing Data

ข้อมูลหายไป

Some values are blank.

2



Duplicate Data

ข้อมูลซ้ำ

The same record appears more than once.

3



Wrong Values

ค่าผิดปกติ

Values are not possible or do not make sense.

4



Inconsistent Formats

รูปแบบไม่สอดคล้อง

Same information in different formats.

5



Spelling / Category Problems

ปัญหาการสะกด / หมวดหมู่

Spelling mistakes or categories are not consistent.

6



Outliers

ค่าผิดปกติสูงหรือต่ำ

Values are far from the rest of the data.



Which problem do you think happens most often?

ปัญหาใดที่คุณคิดว่าเกิดขึ้นบ่อยที่สุด?



Problem 1: Missing Data

Some values are not recorded.

มีข้อมูลบางส่วนหายไป

Example: Student Survey

ID	Name	Age	Date	Favorite Subject	Satisfaction (1–5)	PM2.5 ($\mu\text{g}/\text{m}^3$)
1	Alice	15	2024-05-01	Math	4	23
2	Bob		2024-05-01	Science	5	28
3	Charlie	16		English	3	
4	David	15	2024-05-02		4	
5	Eva	16	2024-05-02	Thai		
6	Frank	15	2024-05-03	Math		
7	Grace		2024-05-03	Science	4	22
8	Hana	16	2024-05-03	English	5	19



What is missing data?

Missing data means a value is empty or not recorded.

ข้อมูลบางช่องว่างหรือไม่ได้ถูกบันทึก

Examples



missing age
อายุหายไป



missing date
วันที่หายไป



missing answer in a survey
คำตอบในแบบสำรวจหายไป



missing PM2.5 reading
ค่าฝุ่น PM2.5 หายไป



Possible action:

Mark it, check the source, or collect it again.

ทำเครื่องหมายไว้ ตรวจสอบแหล่งที่มา หรือเก็บข้อมูลใหม่อีกครั้ง



Caution: Do not guess the value.

Always check why it is missing first.

อย่าเดาค่าที่หายไป ควรตรวจสอบเหตุผลก่อนเสมอ

Problem 2: Duplicate Data

The same record appears more than once.

ข้อมูลซ้ำกันมากกว่าหนึ่งครั้ง



Example: Student Survey

ID	Name	Grade	Favorite Subject	Score (1–5)	Date
1	Alice	M4	Math	4	2024-05-01
2	Bob	M4	Science	5	2024-05-01
3	Charlie	M4	English	3	2024-05-02
3	Charlie	M4	English	3	2024-05-02
4	David	M4	Thai	4	2024-05-02
5	Eva	M4	Math	5	2024-05-03
6	Frank	M4	Science	4	2024-05-03

Same record!



What is duplicate data?

Duplicate data means the same record appears again.

ข้อมูลซ้ำคือข้อมูลชุดเดิมที่ปรากฏซ้ำ



Why it matters

It can make totals or averages wrong.

อาจทำให้ผลรวมหรือค่าเฉลี่ยผิด



Possible action:

Check and remove repeated records carefully.

ตรวจสอบและลบข้อมูลที่ซ้ำออกอย่างระมัดระวัง



Check



Remove



Keep clean data

Problem 3: Wrong or Impossible Values

Some values do not make sense.

ข้อมูลบางค่าไม่สมเหตุผล

Wrong values may be:



typing mistakes
พิมพ์ผิด



impossible numbers
ตัวเลขที่เป็นไปไม่ได้



wrong units
หน่วยผิด



values in the wrong column
อยู่ผิดคอลัมน์

Examples



Age = 250



Temperature =
-90°C in Thailand



Sleep time =
30
hours per day



Example: Daily Student Data

ID	Name	Age (years)	Temperature (°C)	Sleep Time (hours/day)	Weight (kg)
1	Alice	15	28	8	52
2	Bob	250	29	7	60
3	Charlie	16	-90	8	55
4	David	17	31	30	58
5	Eva	16	27	7	53
6	Frank	15	30	8	1290

Wrong or impossible values ค่าที่ผิดหรือเป็นไปไม่ได้



Action:

Check the original source before fixing.

ตรวจสอบแหล่งข้อมูลต้นทางก่อนแก้ไข



1. Review the
original data



2. Confirm what
is correct



3. Then correct
or remove

Problem 4: Inconsistent Formats

The same type of data is written in different ways.

ข้อมูลชนิดเดียวกันถูกเขียนหลายรูปแบบ

Inconsistent formats can include:



different date formats
รูปแบบวันที่ต่างกัน



different units
หน่วยวัดต่างกัน



different spelling styles
การสะกดต่างกัน



different number formats
รูปแบบตัวเลขต่างกัน

Examples



Dates
รูปแบบวันที่

Different formats

05/07/2026

July 5, 2026

Standard format

2026-07-05
(YYYY-MM-DD)



Units
หน่วยวัด

Different formats

170 cm

1.70 m

Standard format

1.70 m
(meters)



Categories / Spelling
ประเภท / การสะกด

Different formats

Male

M

male

Standard format

Male
(Use one style)



Action:
Use one standard format.
ใช้รูปแบบมาตรฐานเดียวกัน



1. Different formats
หลายรูปแบบ



2. Standardize
ทำให้เป็นมาตรฐาน



3. One standard format
รูปแบบเดียวกัน

Problem 5: Spelling and Category Problems

Small text differences can create wrong groups.

ความแตกต่างของข้อความเล็ก ๆ อาจทำให้จัดกลุ่มผิด

Raw Data (Messy Categories)

ข้อมูลดิบ (หมวดหมู่ไม่สม่ำเสมอ)

City เมือง	Chiang Mai	Chiangmai	Chieng Mai
City เมือง	Bangkok	BKK	
Satisfaction ความพึงพอใจ	very good	good	Good
Gender เพศ	Male	male	M

Standardize
ทำให้เป็นมาตรฐาน



Prepared Data (Clean Categories)

ข้อมูลที่เตรียมแล้ว (หมวดหมู่ที่สะอาด)

City เมือง	Chiang Mai
City เมือง	Bangkok
Satisfaction ความพึงพอใจ	Good
Gender เพศ	Male



Problem

ปัญหา

The computer may count them as different categories.

คอมพิวเตอร์อาจนับเป็นคนละกลุ่ม



Why it matters

ทำไมจึงสำคัญ

It can make counts, totals, and analysis results wrong.

อาจทำให้จำนวนรวมและผลวิเคราะห์ผิดพลาด



Action:

Choose one correct category name.

เลือกใช้ชื่อหมวดหมู่ที่ถูกต้องเพียงแบบเดียว



1. Check all variations



2. Choose the correct standard name



3. Use it consistently in all data

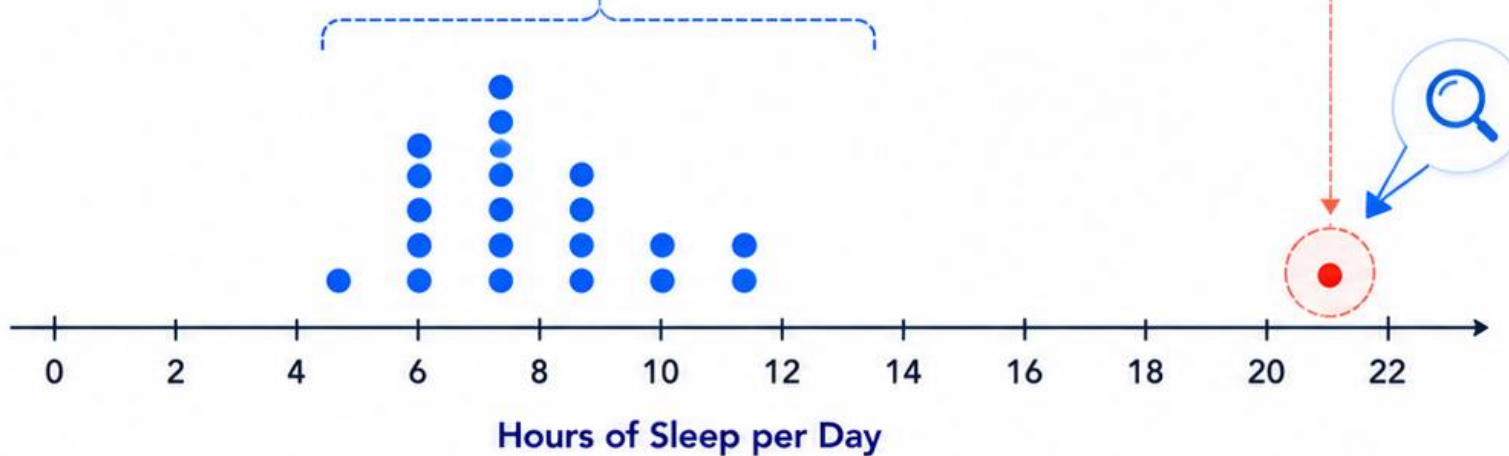
Problem 6: Outliers

A value that is very different from the others.

ค่าที่แตกต่างจากข้อมูลอื่นมากผิดปกติ

Example: Hours of Sleep per Day (Survey of 20 students)

Most answers are between 6–8 hours.
คำตอบส่วนใหญ่อยู่ระหว่าง 6–8 ชั่วโมง



What is an outlier?

An outlier is a value that is far away from most values.
ค่าผิดปกติคือค่าที่อยู่ห่างจากค่าส่วนใหญ่มาก



Examples

- Most students sleep 6–8 hours, but one answer is 20 hours.
นักเรียนส่วนใหญ่นอน 6–8 ชั่วโมง แต่มีคำตอบหนึ่งคือ 20 ชั่วโมง
- Most expenses are 50–200 baht, but one value is 5,000 baht.
ค่าใช้จ่ายส่วนใหญ่อยู่ที่ 50–200 บาท แต่มีค่าหนึ่งคือ 5,000 บาท



Important

Check if it is a real value or a mistake.
ตรวจสอบว่าเป็นข้อมูลจริงหรือข้อผิดพลาด



Do not remove outliers too quickly.

Outliers are not always wrong.
They can be mistakes, but they can also show something important.

อย่าลบค่าผิดปกติออกอย่างรวดเร็ว ค่าผิดปกติอาจเป็นข้อผิดพลาด แต่ก็อาจบอกข้อมูลสำคัญได้เช่นกัน



1. Check the value
ตรวจสอบค่า



2. Check the source
ตรวจสอบแหล่งข้อมูล



3. Decide: Keep or Fix
ตัดสินใจ: เก็บไว้หรือแก้ไข

Data Cleaning Workflow

A simple process for preparing data

ขั้นตอนง่าย ๆ ในการเตรียมข้อมูล



Clean data is ready for analysis.

ข้อมูลที่สะอาดพร้อมสำหรับการวิเคราะห์



Step 1: Inspect the Dataset

Look before you clean.

ตรวจสอบก่อนทำความสะอาดข้อมูล

Dataset Preview: Student Survey

ID	Name	Age (years)	Gender	Study Hours (hours/day)	Sleep Hours (hours/day)	Favorite Subject	Pocket Money (baht/week)
1	Alice	15	Female	2.5	7	Math	150
2	Bob	16	Male	3	6	Science	200
3	Charlie		Male	4	8	Math	180
4	David	250	Male	5	20	History	5000
5	Eve	14	Female	-1	6	Thai	120
6	Frank	15	M	2,5	7	math	150
7	Grace	16	Female	3	6	English	180
8	Henry	15	Male	4	-90	Science	195
9	Ivy	15	female	3.5	7	Arts	-190
10	Jack	14	Male	3	6	Sci.	160



Before cleaning, check:



Column names

ชื่อคอลัมน์



Number of rows

จำนวนแถวข้อมูล



Missing values

ค่าที่หายไป



Unusual values

ค่าที่ผิดปกติ



Data types

ชนิดของข้อมูล



Source notes

บันทึกแหล่งที่มา



Question:

What problems can you see first?

คุณเห็นปัญหาอะไรบ้างเป็นอันดับแรก?



Step 1: Inspect the Dataset

Look before you clean.

ตรวจสอบก่อนทำความสะอาดข้อมูล

Dataset Preview: Student Survey

ID	Name	Age (years)	Gender	Study Hours (hours/day)	Sleep Hours (hours/day)	Favorite Subject	Pocket Money (baht/week)
1	Alice	15	Female	2.5	7	Math	150
2	Bob	16	Male	3	6	Science	200
3	Charlie		Male	4	8	Math	180
4	David	250	Male	5	20	History	5000
5	Eve	14	Female	-1	6	Thai	120
6	Frank	15	M	2,5	7	math	150
7	Grace	16	Female	3	6	English	180
8	Henry	15	Male	4	-90	Science	195
9	Ivy	15	female	3.5	7	Arts	-190
10	Jack	14	Male	3	6	Sci.	160



Before cleaning, check:



Column names

ชื่อคอลัมน์



Number of rows

จำนวนแถวข้อมูล



Missing values

ค่าที่หายไป



Unusual values

ค่าที่ผิดปกติ



Data types

ชนิดของข้อมูล



Source notes

บันทึกแหล่งที่มา



Question:

What problems can you see first?

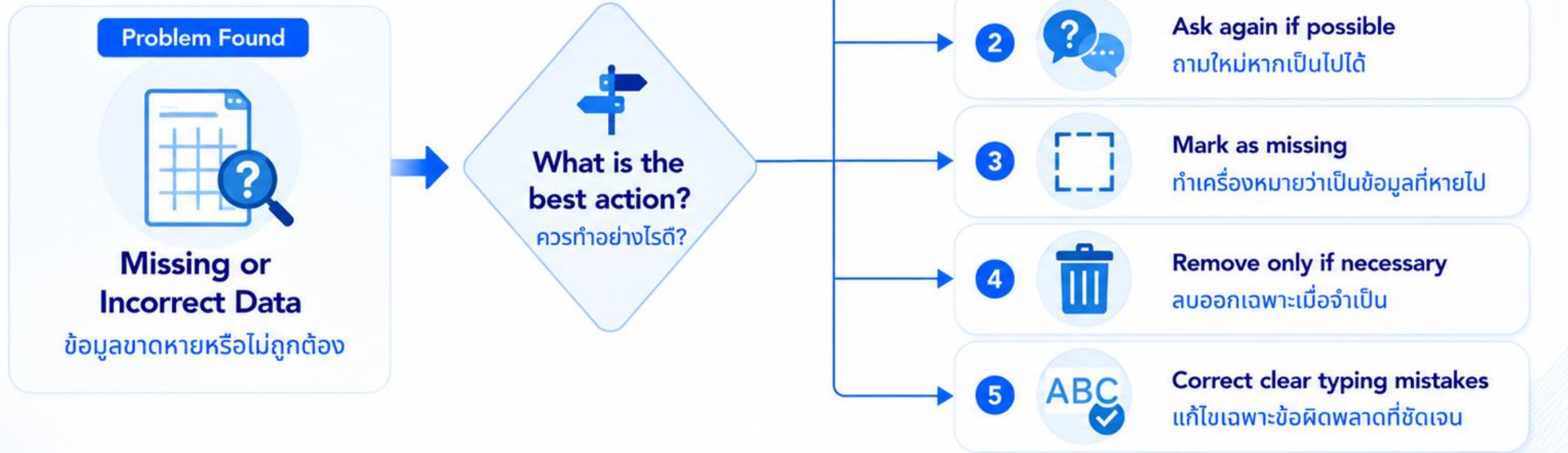
คุณเห็นปัญหาอะไรบ้างเป็นอันดับแรก?



Step 2: Fix Missing or Incorrect Data

Choose the safest action.

เลือกวิธีแก้ที่ปลอดภัยที่สุด



Important!

Do not change data without a reason.

อย่าแก้ไขข้อมูลโดยไม่มีเหตุผล



Be careful
ระวังให้ดี



Record what you did
บันทึกสิ่งที่你做



Keep data trustworthy
รักษาความน่าเชื่อถือของข้อมูล

Step 3: Standardize Formats and Categories

Make similar data look the same.

ทำข้อมูลที่เหมือนกันให้มีรูปแบบเดียวกัน

Messy Formats (Before)

Date	Units (Weight)	Gender	Score (Number)
1/5/2024	5 kg	M	1,200
05-01-24	5000 g	male	1200.0
May 1, 2024	5KG	Male	1.2e3
2024/05/01	5.0 kg	F	1200
1 May 24	5 kilograms	female	1200
		Female	1 200

Standardize

ทำให้เป็น
รูปแบบเดียวกัน

Clean Standard Formats (After)

Date (YYYY-MM-DD)	Units (kg)	Gender (Category)	Score (Number)
2024-05-01	5.00	Male	1200
2024-05-01	5.00	Male	1200
2024-05-01	5.00	Male	1200
2024-05-01	5.00	Female	1200
2024-05-01	5.00	Female	1200

Standardize:



Dates

วันที่



Units

หน่วย



Spelling

การสะกดคำ



Categories

หมวดหมู่



Number Formats

รูปแบบตัวเลข

Example:

M

,

male

,

Male



Male

F

,

female

,

Female



Female



Key idea:

Same meaning = same format.

ความหมายเดียวกัน = รูปแบบเดียวกัน



Consistent data
ข้อมูลสม่ำเสมอ



Accurate analysis
วิเคราะห์ได้ถูกต้อง



Reliable results
ผลลัพธ์ที่เชื่อถือได้

Step 4: Check Again and Document Changes

Good cleaning should be explainable.

การทำความสะอาดข้อมูลควรอธิบายได้

Cleaned Dataset (After)

ID	Name	Date (YYYY-MM-DD)	City	Sleep Hours (hours/day)	Gender (Category)
1	Alice	2024-05-01	Bangkok	7.0	Female
2	Bob	2024-05-01	Chiang Mai	6.0	Male
3	Charlie	2024-05-02	Phuket	8.0	Male
4	David	2024-05-02	Bangkok	7.0	Male
5	Eve	2024-05-02	Khon Kaen	6.5	Female
6	Frank	2024-05-03	Bangkok	6.0	Male
7	Grace	2024-05-03	Chiang Mai	8.0	Female
8	Henry	2024-05-03	Phuket	7.0	Male
9	Ivy	2024-05-04	Bangkok	6.0	Female
10	Jack	2024-05-04	Khon Kaen	7.0	Male

After cleaning, check:

- ✓ Did we fix the problem?
เราแก้ปัญหามาได้หรือไม่?
- ✓ Did we create a new problem?
เราได้สร้างปัญหาใหม่หรือไม่?
- ✓ Can we explain what we changed?
เราสามารถอธิบายสิ่งที่เราเปลี่ยนได้หรือไม่?
- ✓ Is the dataset ready for analysis?
ข้อมูลพร้อมสำหรับการวิเคราะห์หรือไม่?



Change Log (Record what we did)

- Changed "BKK" to "Bangkok"
- Marked missing sleep time as blank



Ready for Analysis!

พร้อมสำหรับการวิเคราะห์!



Clean, consistent, and well-documented data leads to reliable analysis.

ข้อมูลที่สะอาด สมบูรณ์ และมีการบันทึกที่ดี
นำไปสู่การวิเคราะห์ที่เชื่อถือได้



Mini Activity: Clean the Dataset

Find the problems before analysis.

หาปัญหาก่อนวิเคราะห์ข้อมูล

Dataset to Check

ID	Name	Gender	Age	City	Study Hours (hours/day)	Sleep Hours (hours/day)	Pocket Money (baht/week)	Date
1	Nattapol	M	14	Bangkok	2.5	7	150	1/5/24
2	Pimchanok	Female	14	Chiang mai	3	6	200	02-05-2024
3	Krit	male	13	Phuket	4	8	180	2024/05/02
4	Nattapol	M	14	Bangkok	2.5	7	150	1/5/24
5	Benjamas	F	fifteen	Khon Kaen	3.5		100	05-03-2024
6	Thanakorn	Male	16	chiangmai	2	6	1600	3/5/2024
7	Siriporn	Female	14	Bangkokk	4	8.5	220	2024-05-03
8	Weerayut	M	13	Hat yai	three	7	180	04/05/24
9	Kanokwan	female	14	Phuket	2.5	6	130	5-4-24
10	Pongsakorn	M	12	Chiang Rai	3.5	9	150	

Your Task



Work with a partner.
ทำงานกับเพื่อน



Find and mark:

- 1 missing values
ค่าที่หายไป
- 2 duplicate records
ข้อมูลซ้ำ
- 3 wrong values
ข้อมูลที่ผิด
- 4 inconsistent formats
รูปแบบไม่สอดคล้องกัน
- 5 spelling/category problems
การสะกดหรือหมวดหมู่
- 6 possible outliers
ค่าที่ผิดปกติ



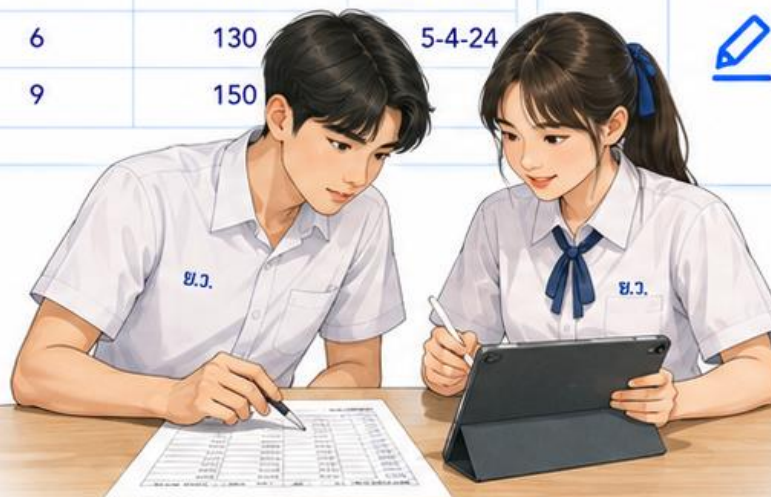
Then write:

One change we would make is...
การเปลี่ยนแปลงหนึ่งอย่างที่เรากำลังจะทำคือ...



Pair Work: 8 minutes

ทำงานเป็นคู่: 8 นาที



Before and After Clean Data

Prepared data is easier to analyze.

ข้อมูลที่เตรียมแล้ววิเคราะห์ได้ง่ายขึ้น

Before Cleaning



Missing values
ข้อมูลหายไป



Repeated rows
ข้อมูลซ้ำกัน



Inconsistent labels
ป้ายกำกับไม่ตรงกัน



Wrong values
ค่าผิดพลาด

ID	Name	Gender	Age	City	Study Hours (hours/day)	Sleep Hours (hours/day)
1	Alice	F	15	Bangkok	2.5	7
2	Bob	male	16	Chiangmai	3	6
3	Charlie	Male		Phuket	4	8
4	David	M	16	Bangkok	5	7
5	Eve	female	14	Khon Kaen	3.5	
6	Frank	Male	15	Bangkok	2	6
7	Frank	Male	15	Bangkok	2	6
8	Grace	F	thirteen	Hatyai	4	8.5
9	Henry	M	15	BKK	3	20
10	Ivy	Female	15	Chiang Mai	2.5	6.0

Clean
ทำความสะอาด

After Cleaning



Clearer table
ตารางชัดเจน



Consistent labels
ป้ายกำกับตรงกัน



Fewer errors
ข้อผิดพลาดน้อยลง



Ready for graphs
พร้อมทำกราฟ

ID	Name	Gender	Age	City	Study Hours (hours/day)	Sleep Hours (hours/day)
1	Alice	Female	15	Bangkok	2.5	7.0
2	Bob	Male	16	Chiang Mai	3.0	6.0
3	Charlie	Male	15	Phuket	4.0	8.0
4	David	Male	16	Bangkok	5.0	7.0
5	Eve	Female	14	Khon Kaen	3.5	7.0
6	Frank	Male	15	Bangkok	2.0	6.0
7	Grace	Female	13	Hat Yai	4.0	8.5
8	Henry	Male	15	Bangkok	3.0	7.0
9	Ivy	Female	15	Chiang Mai	2.5	6.0



Question:

Which version would you trust more?

คุณเชื่อถือข้อมูลชุดไหนมากกว่ากัน?



What Did We Learn Today?

Clean data supports better analysis.

ข้อมูลที่สะอาดช่วยให้วิเคราะห์ได้ดีขึ้น

Today we learned:

1



Raw data is often messy.

ข้อมูลดิบมักจะยุ่งเหยิง

2



Data should be checked before analysis.

ควรตรวจสอบข้อมูลก่อนวิเคราะห์

3



- Missing data
- Duplicates
- Wrong values
- Inconsistent formats
- Spelling problems
- Outliers

Common problems include missing data, duplicates, wrong values, inconsistent formats, spelling problems, and outliers.

ปัญหาทั่วไป ได้แก่ ข้อมูลหาย ข้อมูลซ้ำ ค่าผิด รูปแบบไม่สอดคล้อง การสะกดผิด และค่าผิดปกติ

4



Cleaning decisions should be careful and explainable.

การแก้ไขข้อมูลควรทำอย่างรอบคอบและอธิบายได้

5



Prepared data is easier to analyze.

ข้อมูลที่เตรียมแล้ววิเคราะห์ได้ง่ายขึ้น



Inspect
ตรวจสอบ



Clean
ทำความสะอาด



Standardize
ทำให้เป็นรูปแบบเดียวกัน



Check Again
ตรวจสอบอีกครั้ง



Analyze
วิเคราะห์

Exit Ticket

Before you leave...

ก่อนจบคาบเรียน...



Answer in your notebook:

- 1 What is data preparation?
- 2 Name two common data problems.
- 3 Why should we not change data without a reason?
- 4 What does it mean to standardize data?
- 5 Why is clean data important before analysis?



Good analysis starts with clean data.

การวิเคราะห์ที่ดีเริ่มจากข้อมูลที่สะอาด

