

## Week 1: Data Integrity

e.g. change all of the dates to the same format

However, in case that the data only partially aligns with an objective,  
↳ Alignment to business objectives + newly discovered variables + constraints  
= "Accurate Conclusions"

### Data Integrity

→ the accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycle.

these prevent data integrity

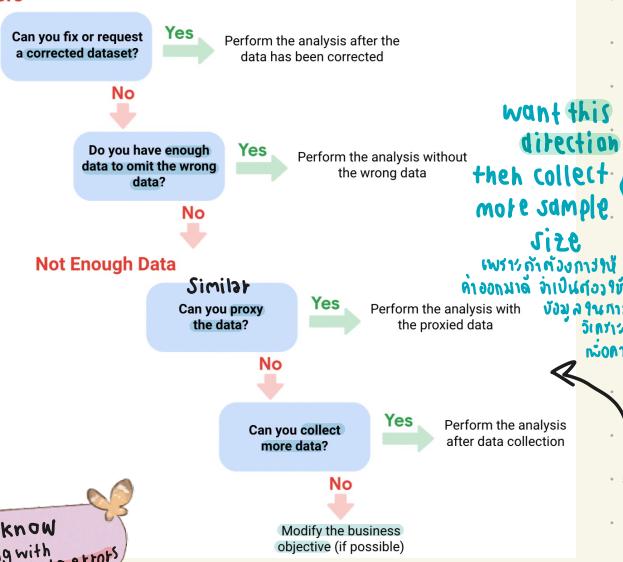
Data can be compromised when:

- Data Replication → the process of **storing** data in multiple locations  
↳ lack data integrity, not consistency ✗
- Data Transfer → the process of **copying** data from a storage device → memory, or from one comp → another
- Data Manipulation → the process of **changing** data to make it more organized & easier to read

### Data Issues that may occur:

- Duplicate data
- Not enough data / Insufficient → wait / find more other sources
- \* types of Insufficient data
  - Data from only 1 source.
  - Data that keeps updating.
  - Outdated data
  - Geographically-limited data

### Data Errors



### Testing your data

**statistical power** → the probability of getting meaningful results from a test.

- **Hypothesis testing** → a way to see if a survey or experiment has meaningful results.

e.g. test whether a milkshake campaign satisfies it

- Want statistically significant result (can believe) about 0.8 or 80%  
this means that 80% → results are reliable

want this direction  
then collect more sample.  
size  
maximum amount that the  
sample size can be  
commonly used is 95%  
significance level



→

The most common workaround is when there isn't time to collect data, you can perform **proxy data** from other datasets

solved by

- ✓ Identify trends with the available data
- ✓ Wait for more data (if time allows)
- ✓ Talk with stakeholders and adjust your objective
- ✓ Look for a new dataset

Use data from another city with a similar size and demographic in case that you don't have the data for a particular city

The smallest sample size for which CLT (Central Limit Theorem) is still valid is 30 (no lower than that)

**Sample Size terms:** Good when  $\frac{S.E.}{\text{margin error}} \geq 5$  (i.e., maximum amount that the sample size can be)

1. **Margin error** → result of sample is expected to differ from result of entire population, the smaller, the closer result differences.

2. **Confidence level** → is how confident you are in the survey results. (How much you can rely on your sample size)

3. **Statistical Significance** → is the determination of whether your result could be done due to random chance or not. The greater, the less due to chance.

### Pre-cleaning steps

1. Determine data integrity by assessing the overall accuracy, consistency, and completeness of the data.
2. Connect objectives to data by investigating how your business objectives can be served into the data.
3. Know when to stop collecting data.



## Next, let's clean the data

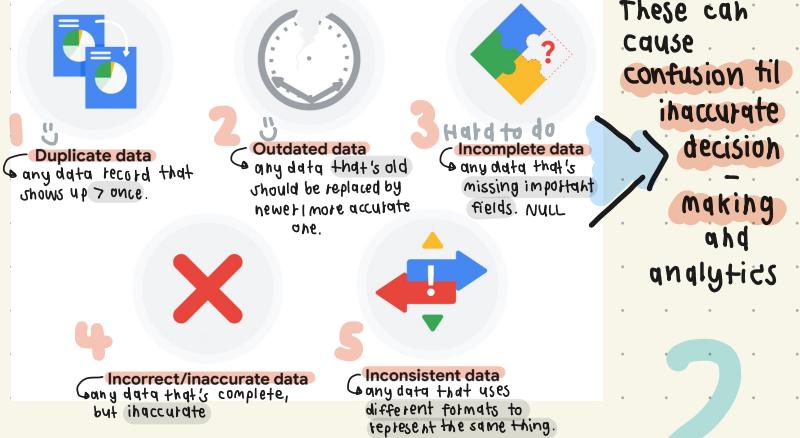


## Week 2: Clean the data in Spreadsheets

want to solve

**Dirty Data** → is incomplete, incorrect, or irrelevant to the problem you're trying to solve.

→ Types of Dirty Data:



Roles working with data:

Data Engineers

transform data into a useful format for analysis and give it a reliable infrastructure.  
(Internal data)

Data Warehousing Specialists

→ develop processes and procedures to effectively store and organize data.  
→ make sure data is secure, available, and not lost.

→ Spreadsheet Workshop Commands → to clean up your data

ep. 1

Used:

- conditional formatting
- clear formatting
- change format
- Data > Split text to columns
- Data > Data cleanup >
  - Remove Duplicates / Trim whitespace (e.g. extra spaces).  
(`=TRIM()`)
  - PivotTable
  - chart plotting to see something unusual

Function:

- COUNTIF()
- LEN() count string length
- LEFT(), RIGHT(), MID()  
to extract substring from left, right, and middle in order.
- CONCATENATE() concat 2 strings  
vertical
- VLOOKUP() like PK, FK in the DB, link between 2 tables.

\* **Data Mapping** → the process of matching fields from one source to another.

- o Primary unique value in column.
- o Foreign key link to PK in another table.

Workflow Automation → the process of automating parts of your work.

Some parts can't be automated such as **Communicating with your team & stakeholders**,

**Present your findings** because there is no replacement of person-to-person communications,

Some can be partially automated: **Preparing data/Cleaning data**, **Data Exploration**,

But **Modeling data** can be automated.

validity (conform to defined business rules / constraints)  
accuracy (conform to true value)  
completeness (all is known)  
consistency (all is equivalent)

\* **Clean Data** → is complete, correct, and relevant to the problem you're trying to solve.

→ very important for external data (however internal data is also required to clean data too)

o **Data validation**: a tool for checking the accuracy and quality of data before adding / importing it such as provide a fixed length / format when receiving inputs.

→ tends to make data misaligned, inconsistent

**Data merging** → the process of combining two or more datasets into a single dataset.

\* **Want compatibility** → how well two/more datasets are able to work together

Some errors you might come across while cleaning your data include:

- Not checking for spelling errors "John" → "Jon"
- Forgetting to document errors write errors you solve, can fix later when it doesn't work.
- Not checking for misaligned values wrong cell
- Overlooking missing values
- Since collecting data
- Not analyzing the system prior to data cleaning understand where this bad data comes from (e.g. not getting a spell check, lack of formats)
- Looking at a subset of data and not the whole picture e.g. using diff sources, you need to look over all data, in case some might be repeated.
- Losing track of the business objectives losing track on what you want to solve.
- Not fixing the source of the error Root cause of error
- Not backing up your data prior to data cleansing na spate
- Not accounting for data cleaning in your deadlines/process not forget deadlines



Next, clean your data on SQL ❤



## Week 3: Clean the data in SQL (SEQUEL)

What tool to use, depends on where the data lives

Spreadsheets	SQL Databases
Smaller datasets	Larger datasets
Create graphs & visualizations in the same program.	Prepare data for further analysis in another software.
Best when working solo on a project.	Great for collaborative work and tracking queries run by all users.
Store locally	Stored across a database
Built-in functionality	Useful across multiple programs

## → Advanced Cleaning functions →

## ① CAST()

\* Beware of number that's string, when searching data it isn't the same, you need to do TYPE CASTING to convert data to another type e.g. × CAST(purchase\_price AS FLOAT64)

BigQuery stores number in bits 64 system ↑

× CAST(date AS date) // change datetime to date format

## ② CONCAT()

to concatenate between text strings

e.g. CONCAT(product\_code, product\_color)

## ③ COALESCE()

to return non-null values in a list.

e.g. COALESCE(product, product\_code)

↑ look for this first, if it's null then look & return from this instead

values (all that are not blank)

\* COUNT() and COUNTA() not the same \*

provided reason of changes  
would be easier to revert back like version control

→ Spreadsheet Workshop Commands →  
to clean up your data ep.2

- Misspelling solved by using → Find and Replace

→ Pivot Table (see COUNTA of each group)  
→ UNIQUE()

\* ปุ่ม กดลัดใน การแก้ไข → it also helps team get on the same page

Documentation is the process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort.

- It allows us to
- Recover data-cleaning errors (ตรวจสอบให้ถูกต้อง errors)
  - Inform other users of changes (แจ้งผู้อื่น เกี่ยวกับการเปลี่ยนแปลง)
  - Determine quality of data (ประเมินคุณภาพของข้อมูล)

Function	Syntax (Google Sheets)	Menu Options (Microsoft Excel)	Primary Use
IMPORTRANGE	=IMPORTRANGE(spreadsheet_url, range_string)	Paste Link (copy the data first)	Imports (copies) data from one sheet to another and keeps it automatically updated.
QUERY	=QUERY(Sheet and Range, "Select *")	Data > From Other Sources > From Microsoft Query	Enables pseudo SQL (SQL-like) statements or a wizard to import the data.
FILTER	=FILTER(range, condition1, [condition2,...])	Filter (conditions per column)	Displays only the data that meets the specified conditions.

Advanced functions for Speedy Data Cleaning

## → → SQL Workshop Commands →

- SELECT data
- INSERT INTO customer\_data.customer\_addr (customer\_id, address, country) VALUES (2645, '333 SQL Road', 'USA');
- UPDATE customer\_data.customer\_addr SET country = "US" new value WHERE customer\_id = 2645;
- CREATE TABLE IF NOT EXISTS
- DROP TABLE IF EXISTS
- SELECT DISTINCT customer\_id FROM \_\_\_\_\_;
- IS NULL
- LENGTH(~) → return length of text string
- TRIM(~) → trim in case it has whitespaces
- SUBSTR(~, # start at, length)
- ASC, DESC when deleting records
- CASE → to detect and replace misspellings first-name="chan"  
WHEN first\_name = "Cahn" THEN END

3

no duplicate customer\_id  
(instead of remove duplicates)

ตอนนี้เรามาท่องเที่ยวใน document Week 4: Verifying and Reporting Clean Data

## ✓ stamp of Approval

Verification is a process to confirm that a

data-cleaning effort was well-executed and the resulting data is accurate and reliable.

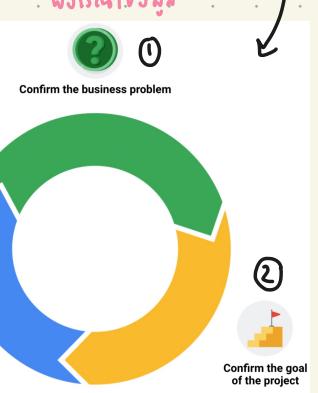
\* See version history on spreadsheet

chronologically ordered list of modifications made to a project. (In Reporting Step)

4

On Module 1: Data Everywhere, we have talked about the five Analytical Thinking skills, there is one of them, Big picture and Detail-oriented thinking, here is the "Big Picture when verifying data-cleaning"

- ① Consider the business problem
- ② Consider goal
- ③ Consider data



(optional)

## Weeks: Adding data to your resume

### Notes:

- Be direct & coherent



# 5

- \* \* - PAR (Problem, Action, Result) to describe your summary statement concisely.
- Experiences that you have done e.g. Spreadsheet, SQL, PivotTable
- Data-related skills and technical skills are things that you should prioritize on the top.
- Explain that allows everyone to understand (Easy to understand)
- Soft skills → Problem-Solving (PAR) \* \*
- (non-tech) → Detail-Oriented (gives attention to detail of work)
  - Persistence (being hardworking & finishing what's started even the obstacles arise)
  - Teamwork (not just team, other people in org. included)

### Experiences

\* o Accomplishments (using PAA format)

→ Describe job that highlights Transferable skills (in case it's not exactly the data analyst role, but it

o jobs that highlights your soft skills

might have some similar things)

### Common Soft Skills:

1	Presentation Skills	
2	Collaboration	
3	Communication	
4	Research	
5	Problem-solving skills	
6	Adaptability	
7	Attention to detail	

See you  
in  
Analyze data  
course!

