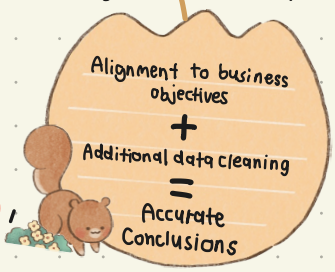
# Week 1: Data Integrity

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 (ตัวแปรใหม่ที่ค้นพบระหว่างการทำงาน) = "Accurate Conclusions"

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




these prevent data integrity

Data can be compromised (harmful) 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 *

what makes data insufficient:
- → Data from only 1 source.
- → Data that keeps updating.
- → Outdated data
- → Geographically-limited data

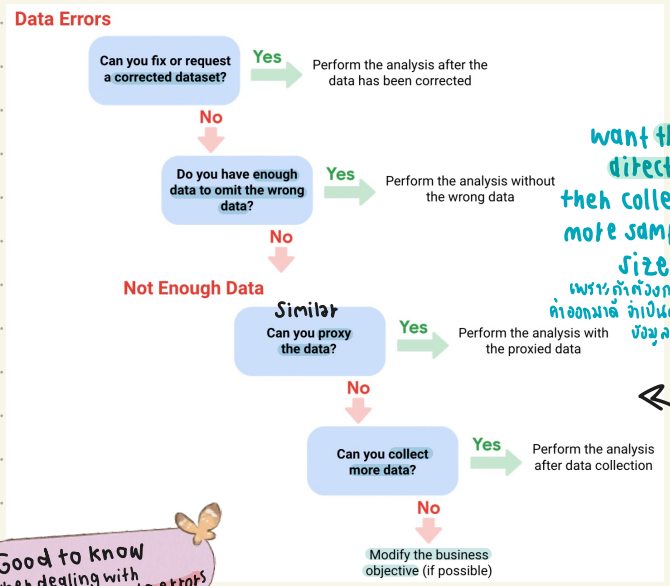
The most common workaround is when there isn't time to collect data, you can perform proxy data from other datasets

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

## Way to address Insufficient data

solved by

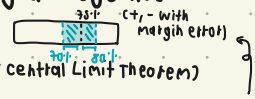
- ✓ Identify trends with the available data
- ✓ Wait for more data (if time allows)
- ✓ Talk with Stakeholders and adjust your objective
- ✓ Look for a new dataset




Good to know when dealing with data errors

Testing your data

## Sample Size terms:




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

Good when ↓ 1. Margin error (+/-) → result of sample is expected maximum amount that the 68% 75% of respondents report they would buy the product again, margin of error is 5%, then the population is true is 70-80%. 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) (commonly used is 95%)

↑ 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.

want this direction then collect more sample size (เพราะถ้าต้องการให้ค่าออกมาดี ก็เป็นต้องใช้ข้อมูลในการวิเคราะห์มากขึ้น เพื่อความแม่นยำ)

* * not have to add up to 100%

**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 mean that 80% → results are reliable

## Pre-cleaning steps

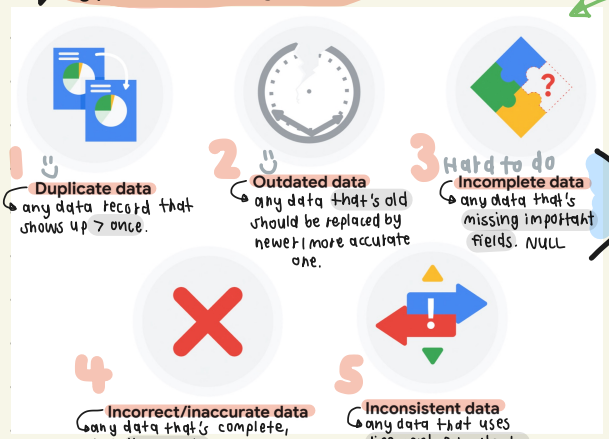
1. Determine data integrity by assessing the overall accuracy, consistency, and completeness of the data.
2. Connect objectives to data by investigating how your business objectives can be served into the data.
3. Know when to stop collect data.

# 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:




solve these

These can cause confusion til inaccurate decision-making and analytics

Important!
- validity (conform to defined business rules / constraints)
- accuracy (conform to a 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)

- **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

## Roles working with data:

**Data Engineers** (ผู้สร้างข้อมูล)
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

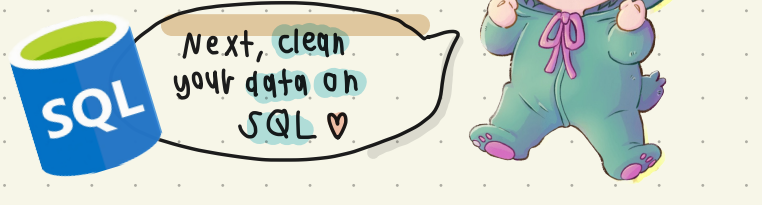
**Used:**
- Conditional formatting
- clear formatting
- change format
- Data > split text to columns
- Data > Data cleanup > Remove Duplicates / Trim whitespace (e.g. extra spaces) or =TRIM()
- Pivot Table ผู้ช่วยสรุป
- chart plotting to see something unusual

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

* **Data Mapping** → the process of matching fields from one source to another.
- Primary unique value in column.
- Foreign key link to PK in another table.

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


**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.



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

→ used to interact with database programs

what tool to use, depend 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 |

## SQL Workshop Commands

- SELECT data
- INSERT INTO customer_data.customer_addr (customer_id, address, country) VALUES (2645, '333 SQL Road', 'USA');
  - dataset name, table name, column name
- UPDATE customer_data.customer_addr SET country = "US" WHERE customer_id = 2645;
  - "US" ← new value
- CREATE TABLE IF NOT EXISTS
- DROP TABLE IF EXISTS
- SELECT DISTINCT customer_id FROM ~~~;
  - → no duplicate customer_id (instead of remove duplicates)
- IS NULL
- LENGTH(~) → return length of text string
- TRIM(~) → trim in case it has whitespaces
- SUBSTR(~, # start at, length)
- ASC, DESC when ordering records

3

## 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
- ถ้าหาค่า product ของ record = null จะให้ return product_code แทน