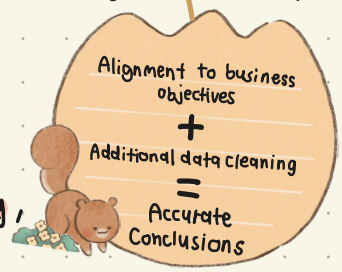
Week 1: Data Integrity

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- → Outdated data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 ☺

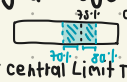
Way to address Insufficient data

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

Sample Size terms:

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



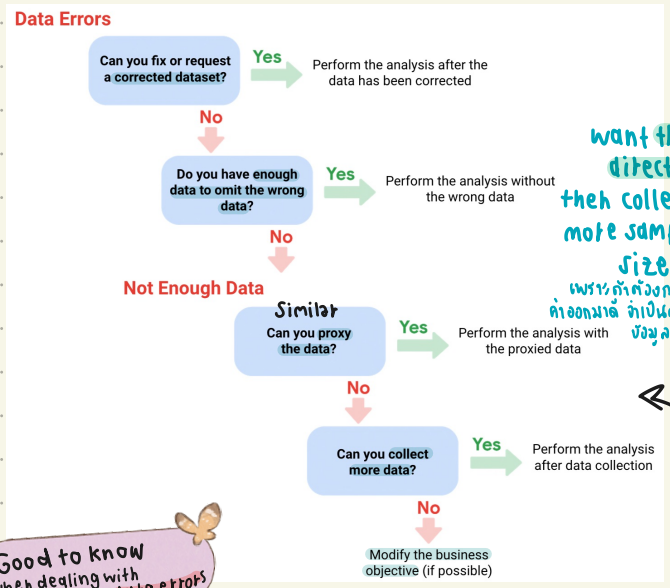
e.g. 95% of respondents report they would buy the product again, margin of error is 5%, then the population is true is 90-80%

Good when ↓ 5% (+/-)

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

** not have to add up to 100%.

want this direction then collect more sample size (เพราะถ้าต้องการให้ค่าออกมาดี ก็เป็นต้องใช้ข้อมูลในการวิเคราะห์มากขึ้นเพื่อความแม่นยำ)



Good to know when dealing with data errors

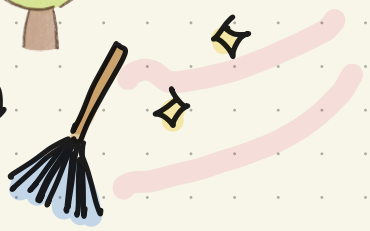
Testing your data

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 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