Week-01 (The importance of integrity)

✳ Data integrity:

The accuracy, completeness, consistency, and trustworthiness of data throughout its lifecycles.

✳ How to deals with insufficient data:

1. Identify trends with the available data.
2. Waits for more data if time allows.
3. Talk with stakeholders and adjusts objective.
4. Look for a new dataset.

Data issue 1: no data

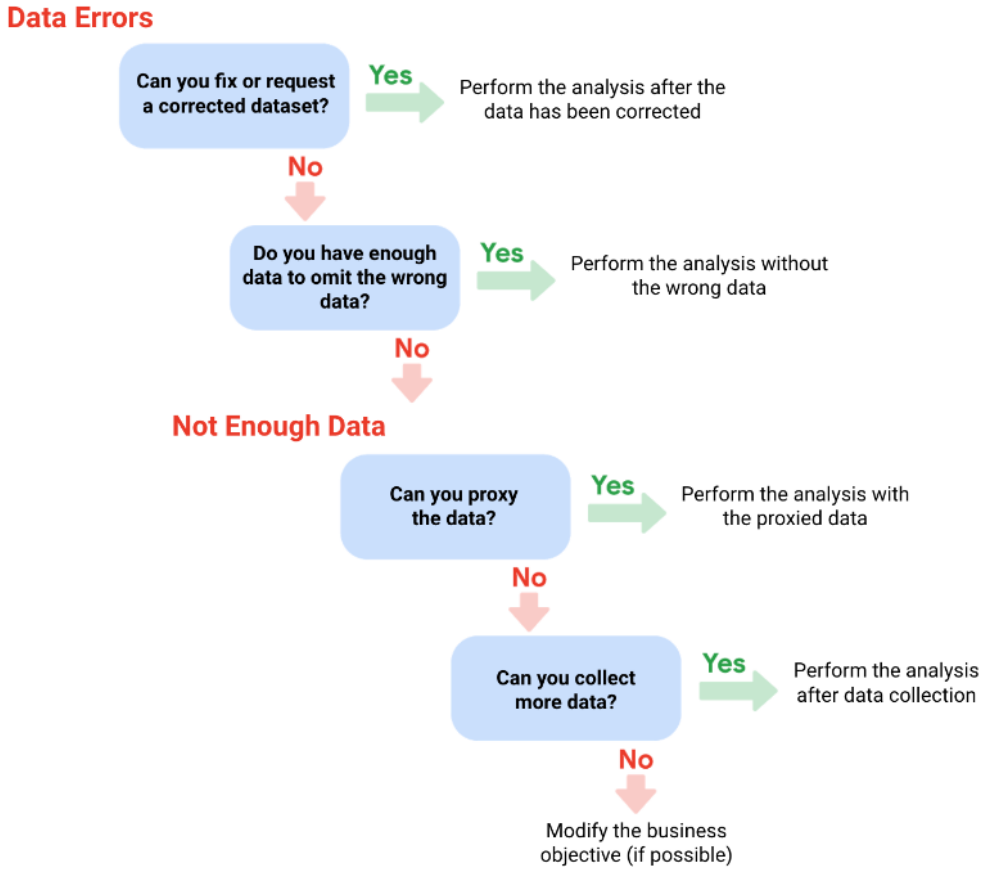
Possible Solutions	Examples of solutions in real life
Gather the data on a small scale to perform a preliminary analysis and then request additional time to complete the analysis after you have collected more data.	If you are surveying employees about what they think about a new performance and bonus plan, use a sample for a preliminary analysis. Then, ask for another 3 weeks to collect the data from all employees.
If there isn't time to collect data, perform the analysis using proxy data from other datasets. *This is the most common workaround.*	If you are analyzing peak travel times for commuters but don't have the data for a particular city, use the data from another city with a similar size and demographic.

Data issue 2: too little data

Possible Solutions	Examples of solutions in real life
Do the analysis using proxy data along with actual data.	If you are analyzing trends for owners of golden retrievers, make your dataset larger by including the data from owners of labradors.
Adjust your analysis to align with the data you already have.	If you are missing data for 18- to 24-year-olds, do the analysis but note the following limitation in your report: *this conclusion applies to adults 25 years and older only.*

Data issue 3: wrong data, including data with errors*

Possible Solutions	Examples of solutions in real life
If you have the wrong data because requirements were misunderstood, communicate the requirements again.	If you need the data for female voters and received the data for male voters, restate your needs.
Identify errors in the data and, if possible, correct them at the source by looking for a pattern in the errors.	If your data is in a spreadsheet and there is a conditional statement or boolean causing calculations to be wrong, change the conditional statement instead of just fixing the calculated values.
If you can't correct data errors yourself, you can ignore the wrong data and go ahead with the analysis if your sample size is still large enough and ignoring the data won't cause systematic bias.	If your dataset was translated from a different language and some of the translations don't make sense, ignore the data with bad translation and go ahead with the analysis of the other data.



✳ Sample:

A part of population that is representative of the population.

→ Sampling bias:

A sample is not representative of the population as a whole.

→ Random sampling:

A way of selecting a sample from a population so that every possible type of the sample has an equal chance of being chosen.

→ Margin of error:

~ is the difference between the sample statistics and population statistics. The smaller the margin of error, the closer the results of the sample are to what the results would have been if we used the population.

→ Confidance level:

How confident you are in the survey result. For example, a 95% confidence level means that if you were to run a survey 100 time, you would similar results 95/100 times.

→ Confident Interval:

The range of possible values that the population results would be at the confidence level of the study. The range is the sample result +/− the margin of error.

→ statistical significant:

The determination of whether your result could be due to random chance or not. The greater the significance, the less due to chance.

✳ Sample size constraints:

1. Don't use a sample size less than 30.
2. The confidence level most commonly used is 95%, but 90% can work in some case.

→ If we increase the sample size then:

i) Confidance level will increase.
ii) Margin of error will decrease.
iii) statistical sigificance will increase.

Week-02 (sparkling-Clean data)

✳ Dirty data:

Data that is incomplete, incorrect, or irrelevant to the problem that we want to solve.

→ Types of dirty data:

1. Duplicate data: Any data point occurs more than once.
2. Outdated data: Any data point that is old, which should be replace with newer data.
3. Incomplete data: Any data that has missing fields.
4. Incorrect data: Any data that is complete but incorrect
5. Inconsistant data: Any data that uses different format to represent the same things.

✳ Data validation:

A tools for checking the accuracy and quality of data before adding or importing it.

✳ Data cleaning process:

1. Removing unwanted data.
2. Removing extra spaces and blanks.
3. Fixing misspellings.
4. Inconsistance capitalization.
5. Incorrect punction and other typos.
6. Removing formatting.

✳ Merger:

An agreement that unites two organizations into a single new one.

✳ Data merging:

The process of combining two or more dataset into a single dataset.

✳ some spreadsheet function:

1. counif → for condition.
2. Len → for getting length of a character.
3. Left → for getting left portion of the string.

4. Right → for getting right portion of the string.

5. Mid → for getting mid portion of the string.

6. Concatenate → for concatenate two or more string.

7. Trim → for trimmed extra spaces.

* Data Mapping:

The process of matching fields from one data source to another.

→ Data mapping is used in :-

i) Data migration.

ii) Data integration.

Week-03 (Cleaning Data with SQL)

* SQL syntax :

1. DISTINCT : return only non-duplicate values.
2. LENGTH : return length of the string.
3. SUBSTR : return substring from any string.
4. TRIM : remove extra spaces.
5. CAST : return casted data type from a given type.
6. CONCATE : concatenate two or more strings to one.
7. COALESCE : used to return non-null values.

Week-04 (Varify and Report on your cleaning results)

✳ **Changelog:** A file containing a cronologically ordered list of modifications made to a project.

✳ **Verification:** A process to confirm that a data-cleaning effort was well executed and the resulting data is accurate and reliable.

Correct the most common problems

Make sure you identified the most common problems and corrected them, including:

- **Sources of errors:** Did you use the right tools and functions to find the source of the errors in your dataset?
- **Null data:** Did you search for NULLs using conditional formatting and filters?
- **Misspelled words:** Did you locate all misspellings?
- **Mistyped numbers:** Did you double-check that your numeric data has been entered correctly?
- **Extra spaces and characters:** Did you remove any extra spaces or characters using the **TRIM** function?
- **Duplicates:** Did you remove duplicates in spreadsheets using the **Remove Duplicates** function or **DISTINCT** in SQL?
- **Mismatched data types:** Did you check that numeric, date, and string data are typecast correctly?
- **Messy (inconsistent) strings:** Did you make sure that all of your strings are consistent and meaningful?
- **Messy (inconsistent) date formats:** Did you format the dates consistently throughout your dataset?
- **Misleading variable labels (columns):** Did you name your columns meaningfully?
- **Truncated data:** Did you check for truncated or missing data that needs correction?
- **Business Logic:** Did you check that the data makes sense given your knowledge of the business?

Week-05 (optional: Adding data to your resume)

✳ Resume building:

1. Contract info.: Name, Address, Mobile, Email.
2. Summary (optional).
3. Experiences / work history.
4. Skills
5. Education
6. Technical skills. / language.