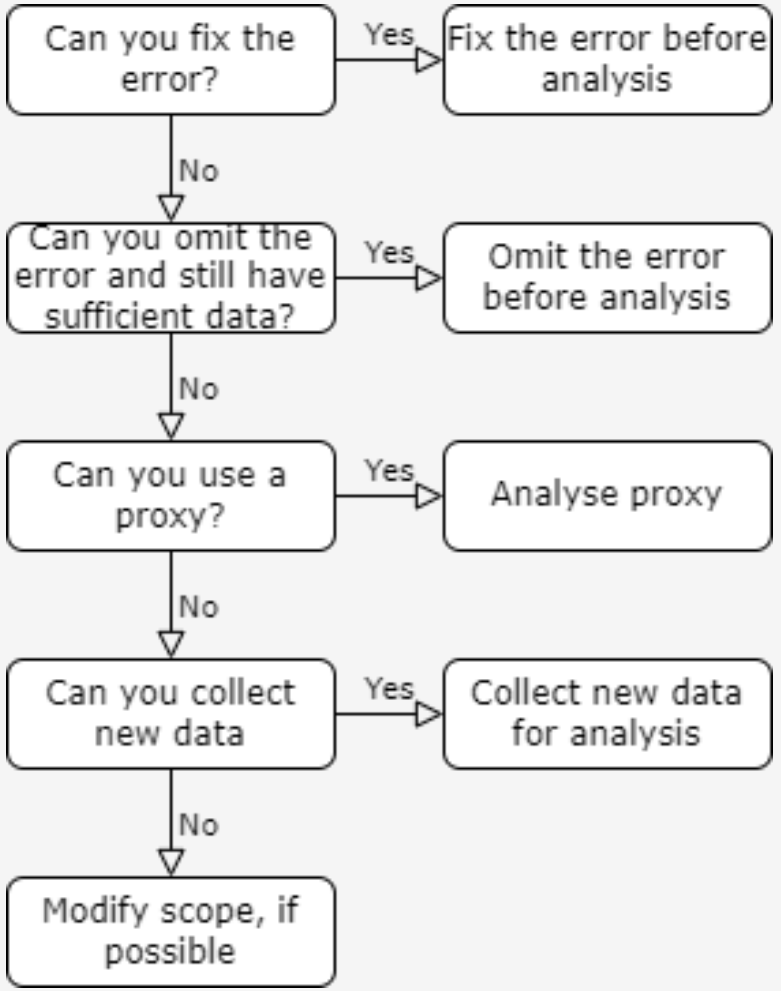
Assessing Data

How to treat errors
Types of assessment
Forms of assessment
Documentation

What to do with errors



Types of data assessment

Visual Assessment

- Opening the data and looking through it randomly.
- The data can be opened in a spreadsheet, database, or text editor

Systematic Assessment

- View specific parts of the data consecutively using functions or methods
- This can be done through spreadsheet functions, SQL queries, or code.

Documenting Assessment

A systematic framework for documenting issues identified by assessing the data

1. Detecting the issue (quality or tidiness) and writing it down
2. Writing the solution to the issue
3. Solving the issue (data cleaning)

Forms of data assessment

Data Quality

- Poor quality data is usually called **dirty data**.
- It refers to content issues such as missing, duplicate, or incorrect data.
- Clean data should be complete, unique, and correct
- Could be caused by errors in data entry, transformation, merging or transmission

Data Tidiness (Structure)

- Data with poor structure is usually called **messy data**.
- It refers to structure issues (rows, columns and tables).
- Tidy data should have unique columns and relational tables.
- It is usually caused by poor data management standards

Data Quality

- Completeness: Are there missing records or fields?
- Validity: Is the data trustworthy? For example, data about people's ages should not be negative numbers.
- Accuracy: Is the data correct? For example, finding 2023 data in a table that should contain 2020 data
- Consistency: Are the data types consistent? For example, a date column should have a constant format such as date/month/year but not a mixture of different formats. Also across tables, the data should be the same
- A primary key is a unique column in a table and doesn't contain duplicates in itself
- A foreign key connects tables to one another and is a primary key in another table

Constraints to help data quality

Constraint	Meaning	Example
Data type	Values should be a certain type such as date, number, text, Boolean etc.	Dates should not be stored as text
Data range	There is a predefined minimum and maximum value	Ages cannot be negative
Mandatory	They should not be blank or empty	People's names in biodata
Unique	Should not be duplicated	Email addresses are unique
Regex patterns	There is a pattern that must be followed	Phone numbers should match 234-##-####-####
Cross validation	Certain conditions must be satisfied	Percentage columns should add up to 100%
Data validation	Values must be from a set of values	Attendance should either be Present or Absent
Primary key	Must be unique	Every row should have a unique primary key
Foreign key	To relate tables to one another	Example shown in practical