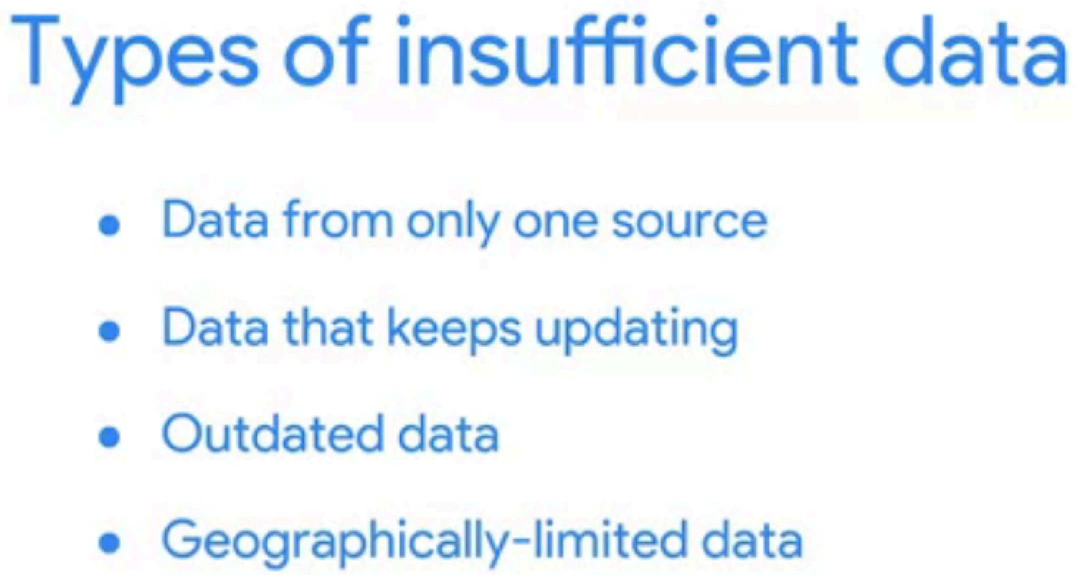
Process Data from Dirty to Clean

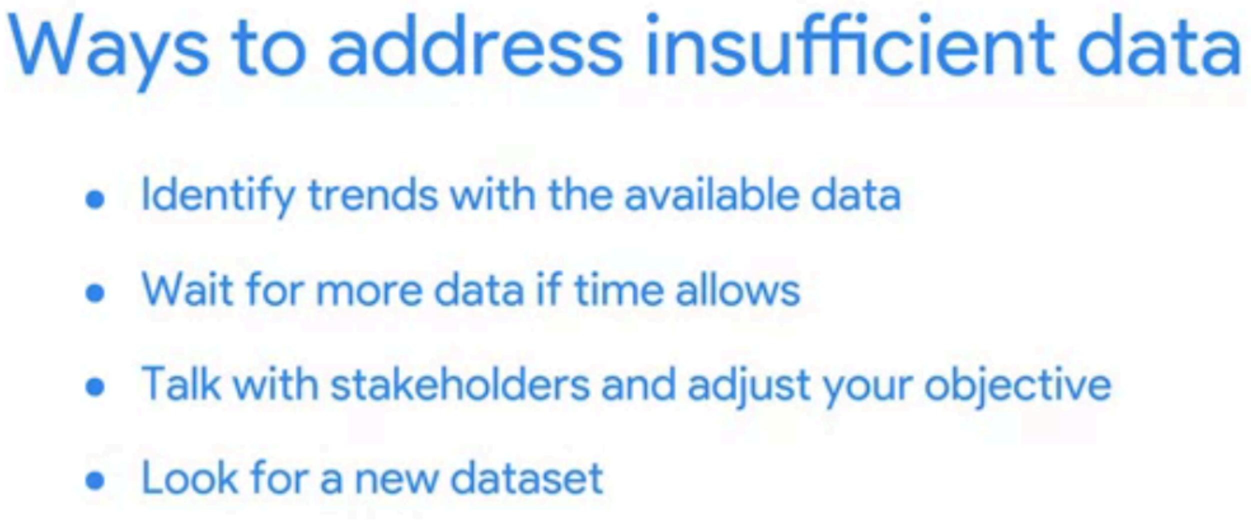
★ **Data Integrity**

Clean data + alignment to business objective = accurate conclusions



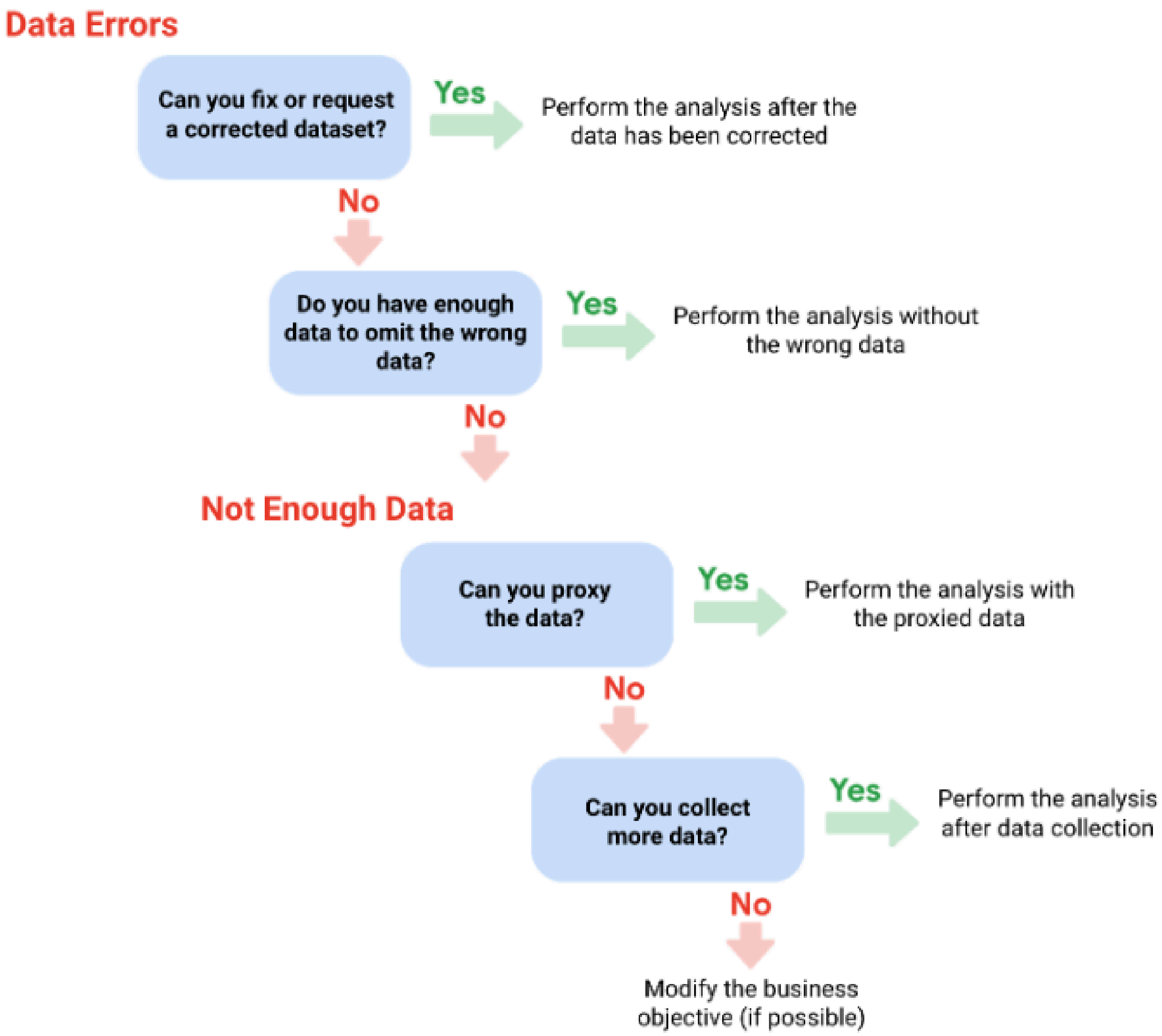
Types of insufficient data

- Data from only one source
- Data that keeps updating
- Outdated data
- Geographically-limited data

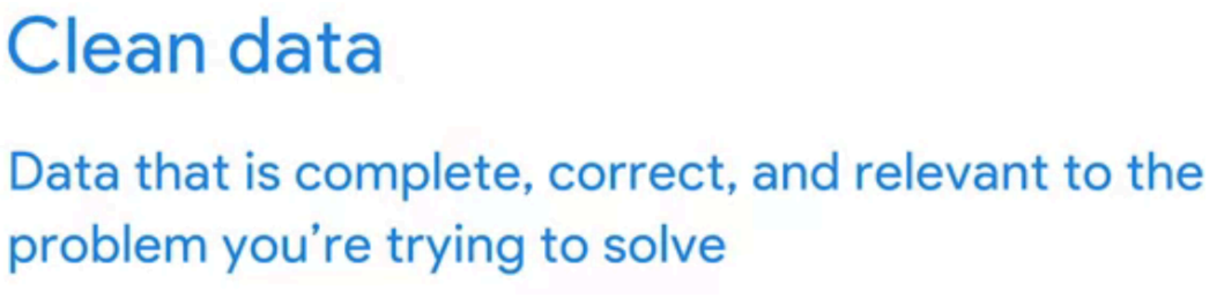


Ways to address insufficient data

- 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


Data Errors
Can you fix or request a corrected dataset?
Yes
Perform the analysis after the data has been corrected
No
Do you have enough data to omit the wrong data?
Yes
Perform the analysis without the wrong data
No
Not Enough Data
Can you proxy the data?
Yes
Perform the analysis with the proxied data
No
Can you collect more data?
Yes
Perform the analysis after data collection
No
Modify the business objective (if possible)

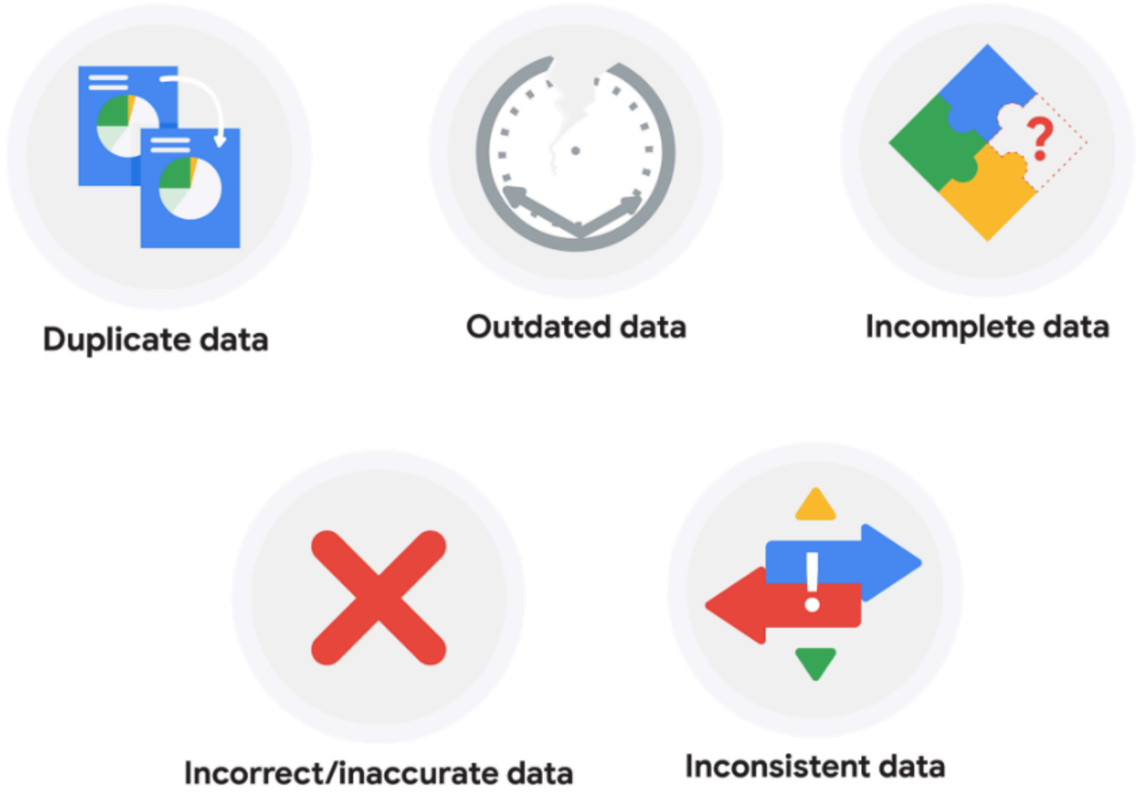
★ **Clean data for more accurate insights**

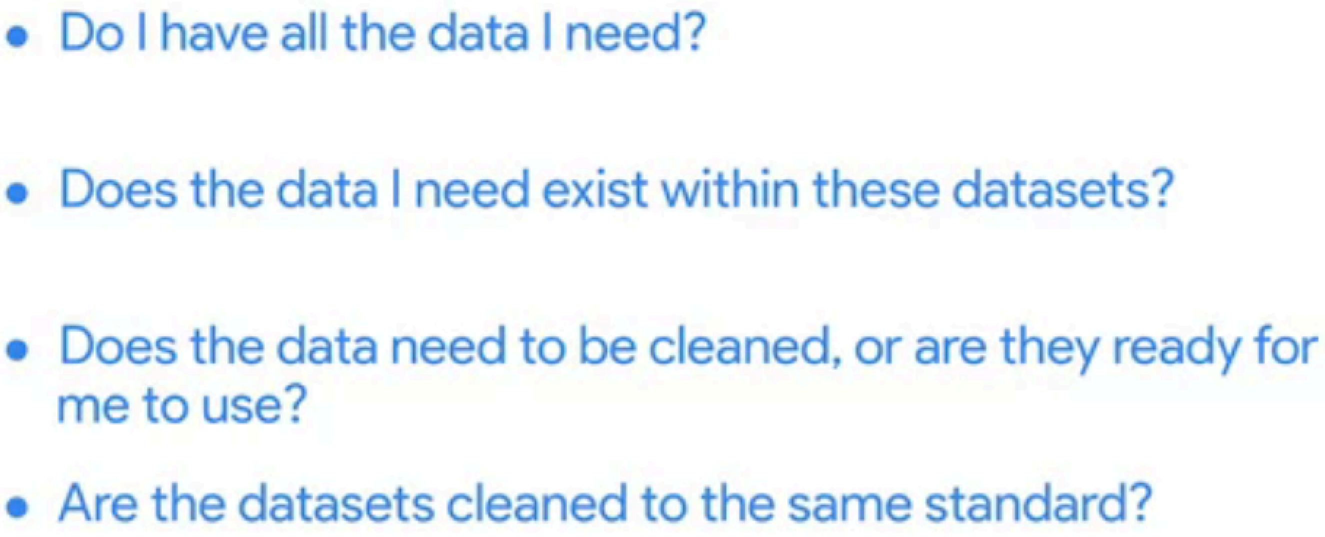


Clean data

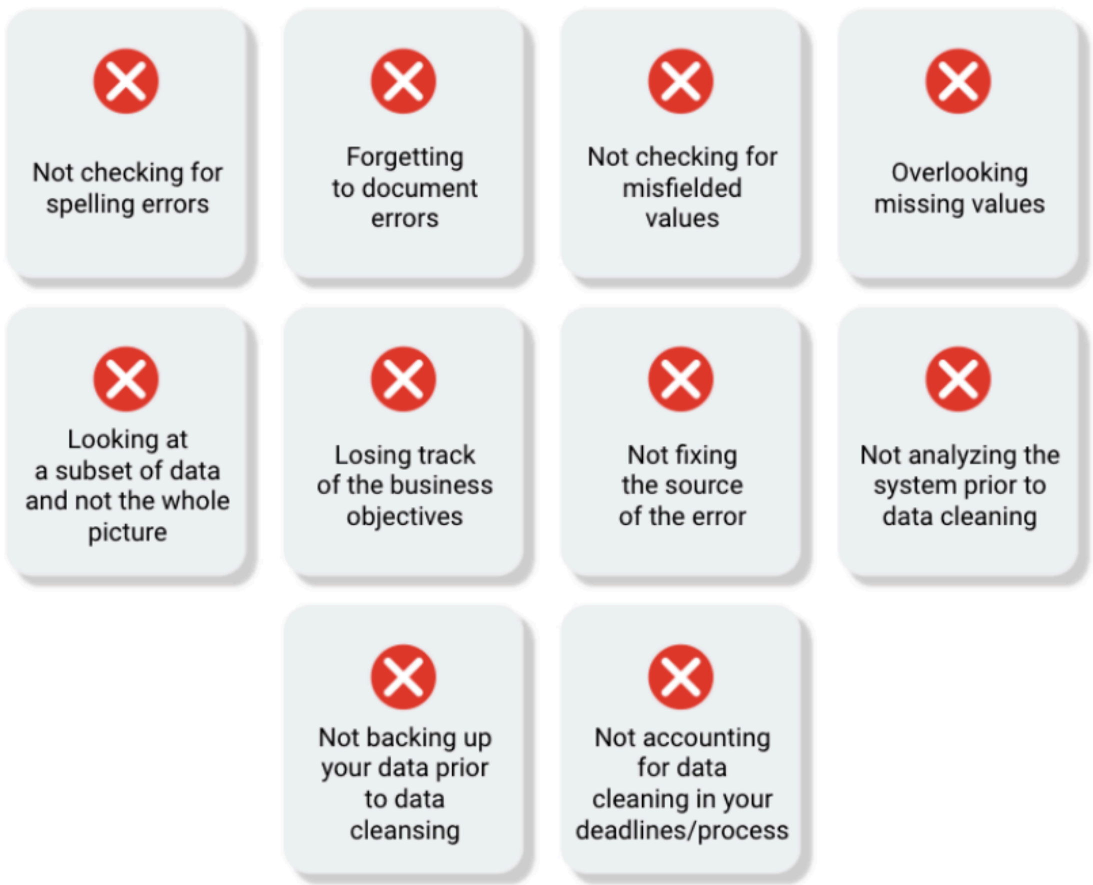
Data that is complete, correct, and relevant to the problem you're trying to solve

Types of dirty data



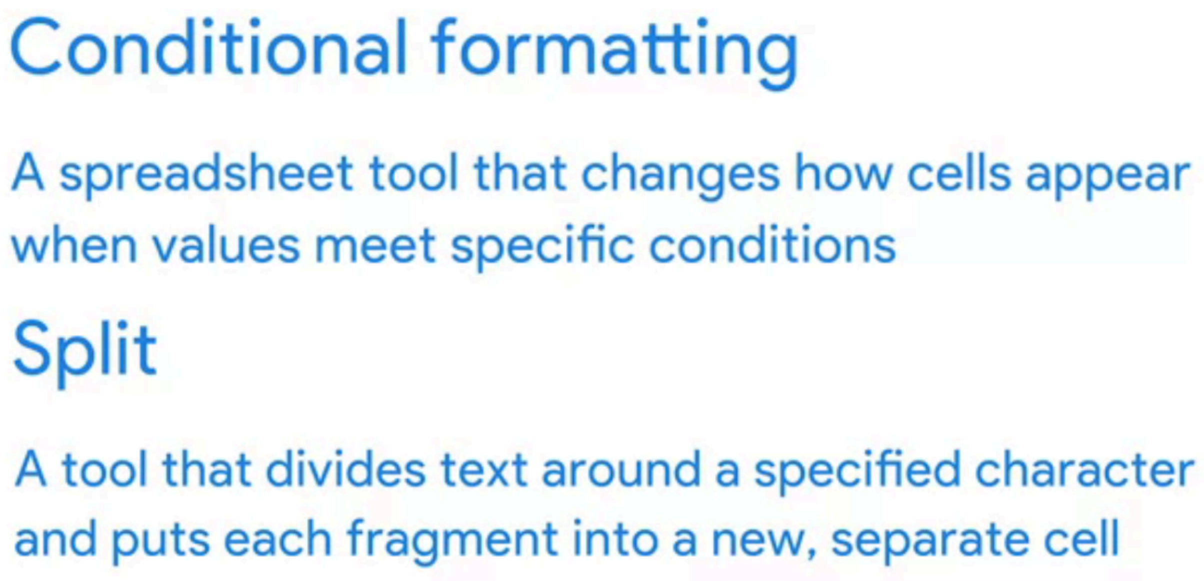


- Do I have all the data I need?
- Does the data I need exist within these datasets?
- Does the data need to be cleaned, or are they ready for me to use?
- Are the datasets cleaned to the same standard?



- Not checking for spelling errors
- Forgetting to document errors
- Not checking for misfielded values
- Overlooking missing values
- Looking at a subset of data and not the whole picture
- Losing track of the business objectives
- Not fixing the source of the error
- Not analyzing the system prior to data cleaning
- Not backing up your data prior to data cleansing
- Not accounting for data cleaning in your deadlines/process

★ Data cleaning with Spreadsheet(Google Sheet)

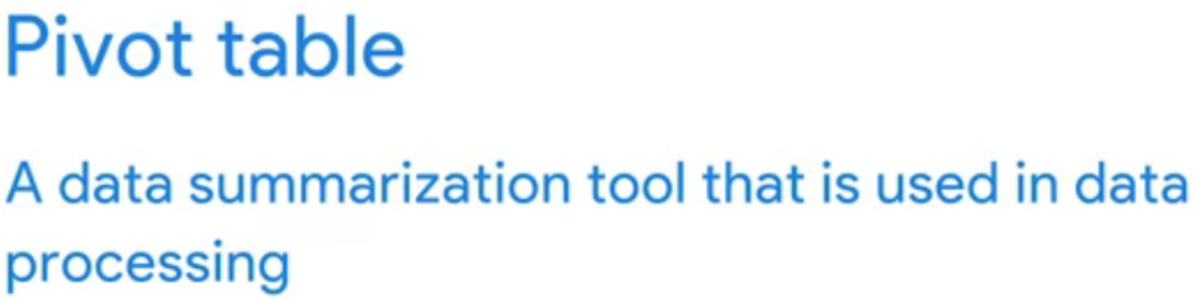


Conditional formatting

A spreadsheet tool that changes how cells appear when values meet specific conditions

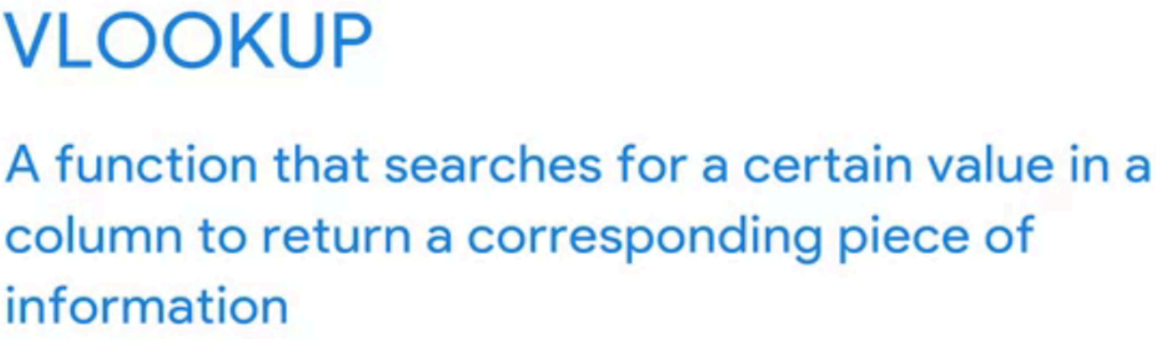
Split

A tool that divides text around a specified character and puts each fragment into a new, separate cell



Pivot table

A data summarization tool that is used in data processing



VLOOKUP

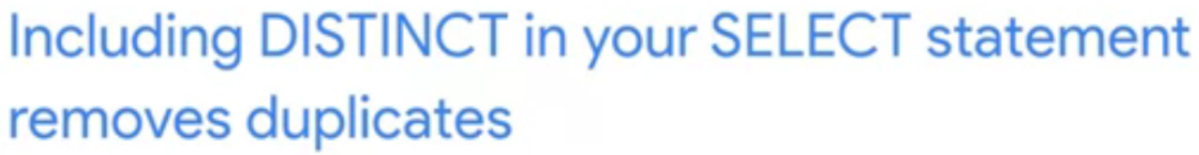
A function that searches for a certain value in a column to return a corresponding piece of information

`VLOOKUP` searches for matches to a specified value in one column, returning a corresponding piece of information from another location.

Develop your approach to cleaning data

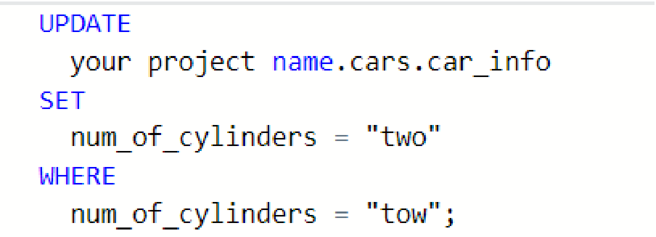
- **Determine the size of the dataset:** Large datasets may have more data quality issues and take longer to process. This may impact your choice of data cleaning techniques and how much time to allocate to the project.
- **Determine the number of categories or labels:** By understanding the number and nature of categories and labels in a dataset, you can better understand the diversity of the dataset. This understanding also helps inform data merging and migration strategies.
- **Identify missing data:** Recognizing missing data helps you understand data quality so you can take appropriate steps to remediate the problem. Data integrity is important for accurate and unbiased analysis.
- **Identify unformatted data:** Identifying improperly or inconsistently formatted data helps analysts ensure data uniformity. This is essential for accurate analysis and visualization.
- **Explore the different data types:** Understanding the types of data in your dataset (for instance, numerical, categorical, text) helps you select appropriate cleaning methods and apply relevant data analysis techniques.

★ Data Cleaning with SQL

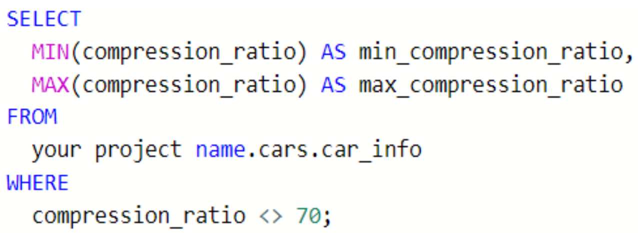


Including DISTINCT in your SELECT statement removes duplicates

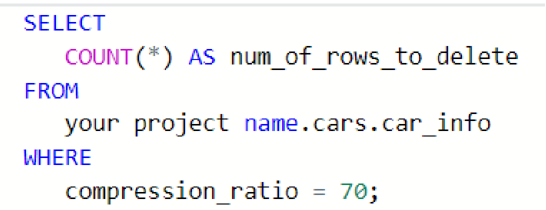
LENGTH
SUBSTR
TRIM



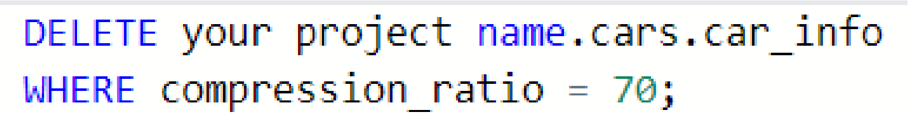
```
UPDATE
  your project name.cars.car_info
SET
  num_of_cylinders = "two"
WHERE
  num_of_cylinders = "tow";
```



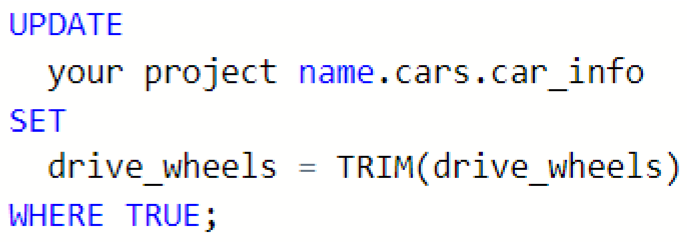
```
SELECT
  MIN(compression_ratio) AS min_compression_ratio,
  MAX(compression_ratio) AS max_compression_ratio
FROM
  your project name.cars.car_info
WHERE
  compression_ratio <> 70;
```



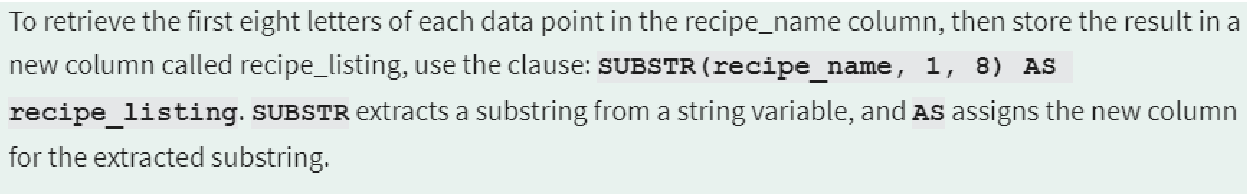
```
SELECT
  COUNT(*) AS num_of_rows_to_delete
FROM
  your project name.cars.car_info
WHERE
  compression_ratio = 70;
```



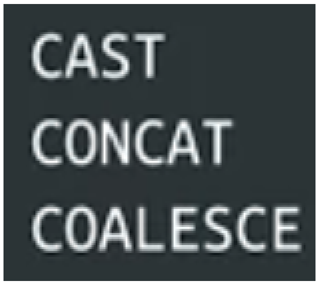
```
DELETE your project name.cars.car_info
WHERE compression_ratio = 70;
```



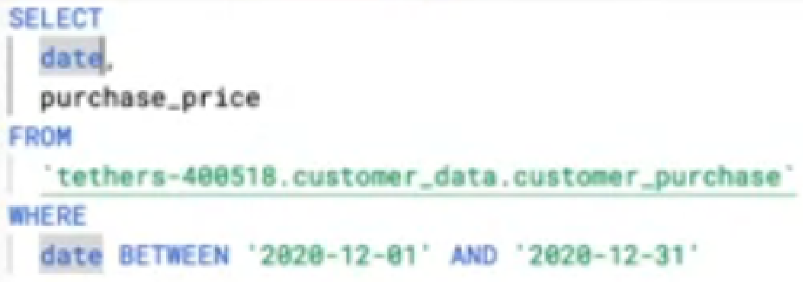
```
UPDATE
  your project name.cars.car_info
SET
  drive_wheels = TRIM(drive_wheels)
WHERE TRUE;
```



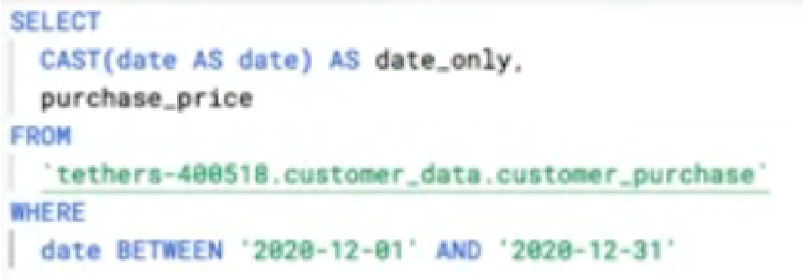
To retrieve the first eight letters of each data point in the recipe_name column, then store the result in a new column called recipe_listing, use the clause: `SUBSTR(recipe_name, 1, 8) AS recipe_listing`. `SUBSTR` extracts a substring from a string variable, and `AS` assigns the new column for the extracted substring.



CAST
CONCAT
COALESCE



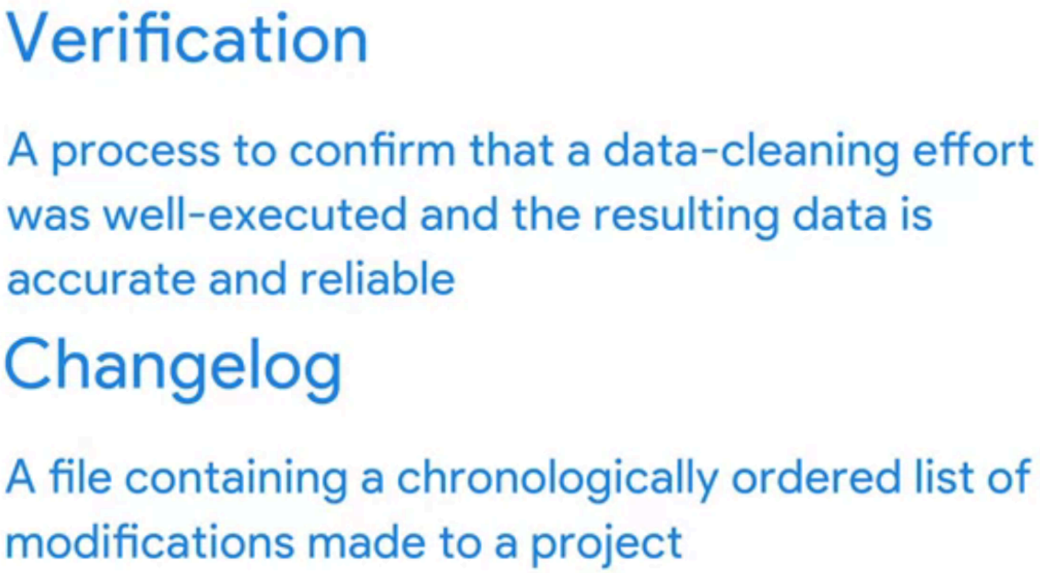
```
SELECT
  date,
  purchase_price
FROM
  `tethers-400518.customer_data.customer_purchase`
WHERE
  date BETWEEN '2020-12-01' AND '2020-12-31'
```



```
SELECT
  CAST(date AS date) AS date_only,
  purchase_price
FROM
  `tethers-400518.customer_data.customer_purchase`
WHERE
  date BETWEEN '2020-12-01' AND '2020-12-31'
```

- Getting data from a table using `SELECT` statements.
- De-duplicating data using commands like `DISTINCT` and `COUNT` + `WHERE`.
- Manipulating string data with `TRIM()` and `SUBSTR`.
- Creating/dropping tables with `CREATE TABLE` and `DROP TABLE`.
- Changing data types with `CAST`.

★ Verify and report cleaning results

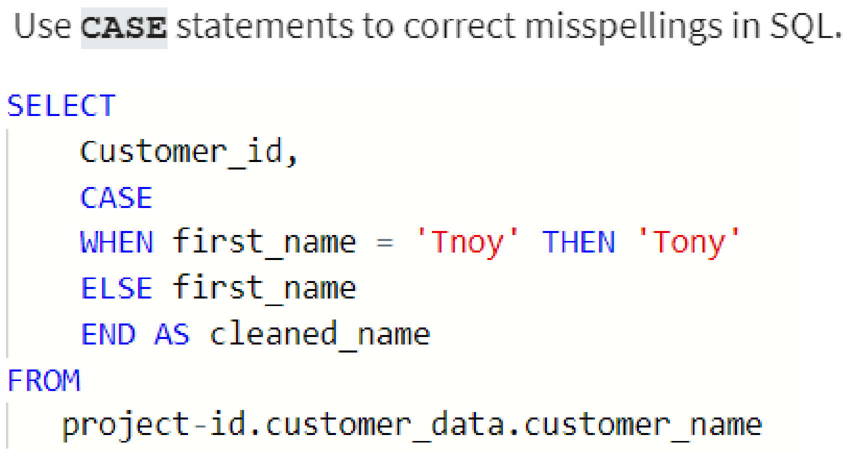


Verification

A process to confirm that a data-cleaning effort was well-executed and the resulting data is accurate and reliable

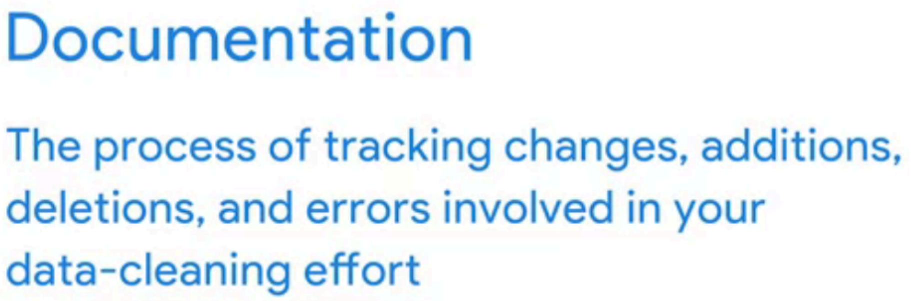
Changelog

A file containing a chronologically ordered list of modifications made to a project



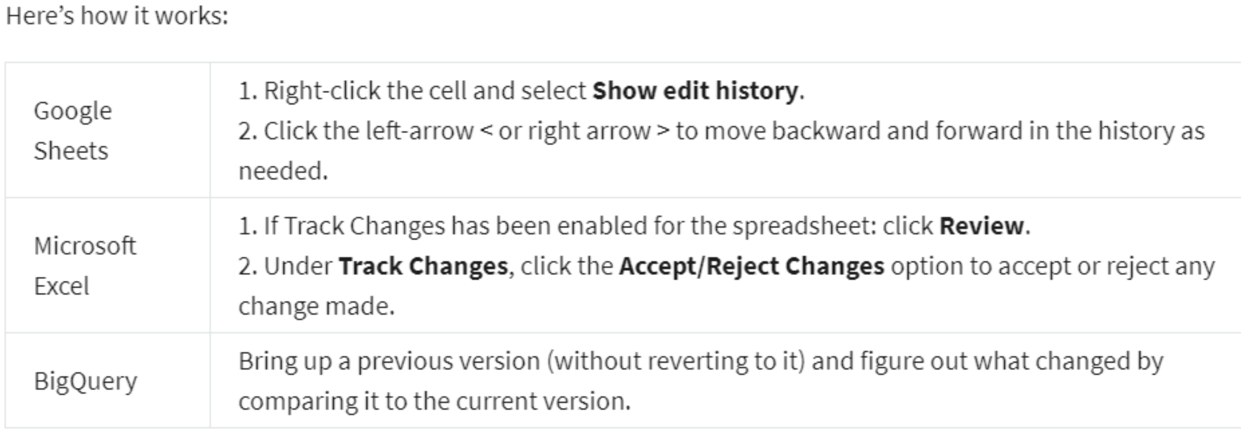
Use `CASE` statements to correct misspellings in SQL.

```
SELECT
    Customer_id,
    CASE
    WHEN first_name = 'Tnoy' THEN 'Tony'
    ELSE first_name
    END AS cleaned_name
FROM
   project-id.customer_data.customer_name
```



Documentation

The process of tracking changes, additions, deletions, and errors involved in your data-cleaning effort

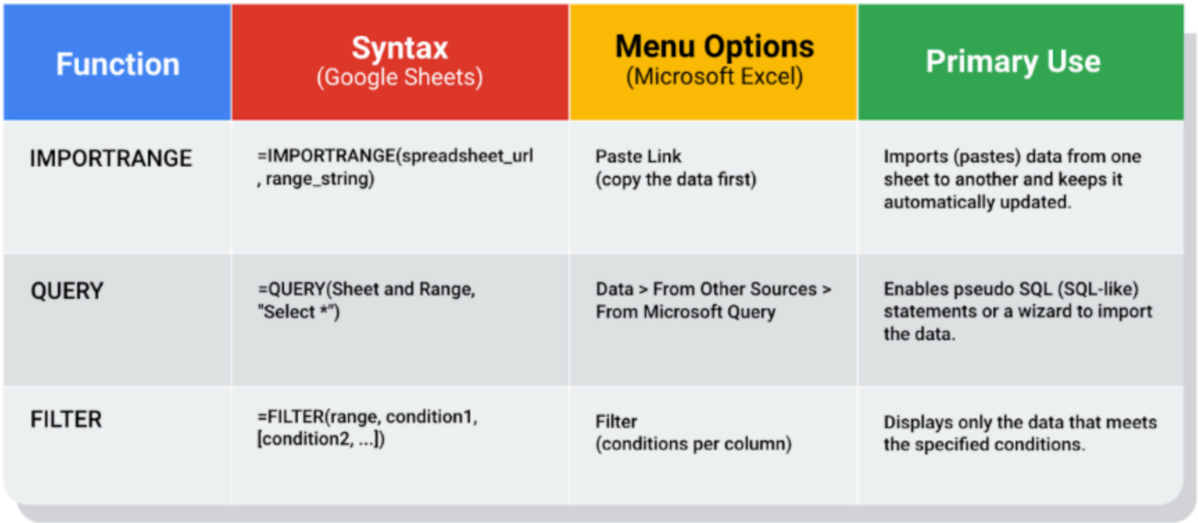


Here's how it works:

Google Sheets	1. Right-click the cell and select **Show edit history**. 2. Click the left-arrow < or right arrow > to move backward and forward in the history as needed.
Microsoft Excel	1. If Track Changes has been enabled for the spreadsheet: click **Review**. 2. Under **Track Changes**, click the **Accept/Reject Changes** option to accept or reject any change made.
BigQuery	Bring up a previous version (without reverting to it) and figure out what changed by comparing it to the current version.

Advanced functions for speedy data cleaning

In this reading, you will learn about some advanced functions that can help you speed up the data cleaning process in spreadsheets. Below is a table summarizing three functions and what they do:



Function	Syntax (Google Sheets)	Menu Options (Microsoft Excel)	Primary Use
IMPORTRANGE	=IMPORTRANGE(spreadsheet_url , range_string)	Paste Link (copy the data first)	Imports (pastes) 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.

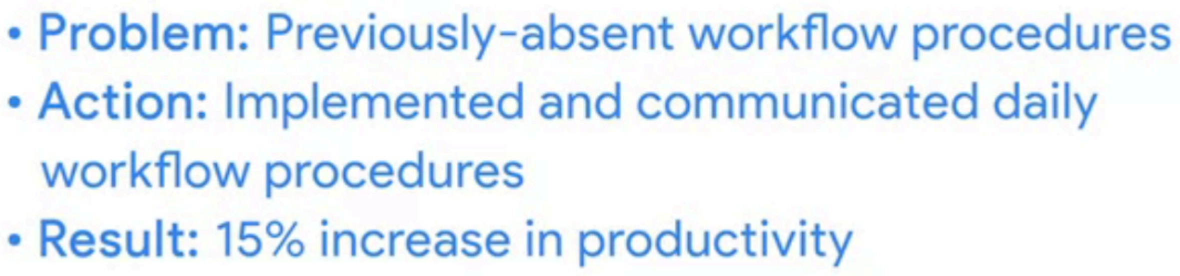
★ Data Analyst Profile and Hiring Process

★ *Skill section of the Resume(Template)*

Step 4: Identify skills to add to your resume

The skills section on your resume likely only has room for 2-4 bullet points, so be sure to use this space effectively. You might want to prioritize technical skills over professional skills. This is a great chance for you to highlight some of the skills you've picked up in these courses, such as:

- Strong analytical skills
- Pattern recognition
- Relational databases and SQL
- Strong data visualization skills
- Proficiency with spreadsheets, SQL, R, and Tableau



- **Problem:** Previously-absent workflow procedures
- **Action:** Implemented and communicated daily workflow procedures
- **Result:** 15% increase in productivity

Add professional skills to your resume

There is more than just data when it comes to being a data analyst—there are plenty of professional skills that can set you apart from other candidates so that potential employers will notice you and know that you have the ability to succeed in this role. Here are some of the most common professional skills you will find in an entry-level data analyst resume.

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