☰ Hide menu

Manually cleaning data

✓ **Video:** Verify and report results
3 min

✓ **Video:** Confirm data-cleaning meets business expectations
5 min

✓ **Reading:** Step-by-Step: Verification of data cleaning
20 min

Video: Verification of data cleaning
8 min

Reading: Data-cleaning verification checklist
10 min

Practice Quiz: Test your knowledge on manual data cleaning
8 min

Document the cleaning process

Video: Capture cleaning changes
5 min

Reading: Embrace changelogs
20 min

Practice Quiz: Self-Reflection: Creating a changelog
20 min

Video: Why documentation is important
3 min

Video: Feedback and cleaning
2 min

Reading: Advanced functions for speedy data cleaning
10 min

Practice Quiz: Test your knowledge on documenting the cleaning process
8 min

Module 4 challenge

Step-by-Step: Verification of data cleaning

This reading outlines the steps the instructor performs in the following video, Verification of data cleaning. The video demonstrates how to verify cleaned data in both spreadsheets and SQL.

Keep this step-by-step guide open as you watch the video. It can serve as a helpful reference if you need additional context or clarification while following the video steps. This is not a graded activity, but you can complete these steps to practice the skills demonstrated in the video.

What you'll need

If you'd like to follow along with the examples in this video, choose a spreadsheet tool. Google Sheets or Excel are recommended.

To access the spreadsheet the instructor uses in this video, click the link to the template to create a copy of the dataset. If you don't have a Google account, download the data directly from the attachments below.

Link to dataset: Jeff's Party Planet - Data for Cleaning

OR

Jeff's Party Planet - Data for Cleaning
XLSX File

Note: The SQL table used in this example is not available for this activity.

Example 1: Verify data with spreadsheets

Use spreadsheet tools such as Find and Replace and pivot tables to find, understand, and fix errors in your spreadsheet.

Use Find and Replace to replace all instances of a mistake

1. Use the Jeff's Party Planet - Data for Cleaning dataset.
2. From the **Edit** menu, choose **Find and Replace** to open the **Find and replace** dialog box.
3. In the **Find** field, enter the misspelled word in the supplier name, **Plos**.
4. In the **Replace with** field, enter **Plus**.
5. Click **Replace all** to replace all instances of "Plos" with "Plus". Click **Done** to close the **Find and replace** dialog box.
6. Select the **Undo** button to use a different method to correct this misspelling. This can also be done with **Ctrl** (Windows) or **Command** (Mac) **Z**.

Use a pivot table to understand errors in a spreadsheet

1. Select the **Suppliers** column.
2. Select **Insert > Pivot Table**. In the **Create pivot table** dialog box, choose **New Sheet** then **Create**.
3. This creates a new tab that is mostly blank.
4. Additionally, the **Pivot table editor** pane is in the window.
5. Next to **Rows**. Select **Add**, then the **Suppliers** column.
6. Next to **Values**, select **Add** then select **Suppliers**. This adds a value for the **Suppliers** column.
7. By default, Google Sheets sets the value to summarize by `COUNTA` (the total number of values in a range). This will show how many times each supplier name comes up. It's a great way to check for misspellings and other anomalies. **Note:** Don't use **COUNT**, because **COUNT** counts only numerical values.
8. When there is only one instance of the misspelled name, manually change it to the correct spelling.
9. To return to the original sheet, select the **Sheet1** tab.

Example 2: Use a CASE statement to verify data in SQL

Use `CASE` statements to correct misspellings in SQL.

1. The SQL table used in this example is not available for download, but if you were performing a similar query you'd first make sure to load the data in BigQuery.

2. Start your SQL query with the basic structure:

`SELECT`

`FROM`

`WHERE`

3. In the `FROM` clause, specify the table you're pulling data from after `FROM`. For example, `project-id.customer_data.customer_name`

4. In the `SELEC`T clause, specify the columns you want to return. In this example, you want `customer_id` and `first_name`.

5. However, there is a misspelling in a customer's first name.

i. To correct the misspelled name "Tnoy" to "Tony", use a `CASE` statement.

ii. Enter `CASE`. On the next line, enter `WHEN first_name = 'Tnoy'THEN 'Tony'`. This tells SQL to replace any instances of `Tnoy` in the `first_name` column with `Tony`.

iii. On the next line, add the `statement ELSE first_name` to keep other names as they are.

iv. End the statement with `END AS cleaned_name`.This creates a new column called `cleaned_name` that will contain the data cleaned with the `CASE` statement.

6. Delete the `WHERE` clause because you don't want to filter the query.

7. The final statement should be:

```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
```

8. This SQL query will correct the misspelled name and leave other names unchanged in a new column called `cleaned_name`. Note that this query corrects only the display of the name; it does not update the table's data.

Go to next item ✓ Completed