Data cleaning

Practical data integrity

Data cleaning

Overview

01. Data integrity

02. Data validation in Google Sheets

03. Data access control in Google Sheets

04. Tracking changes in Google Sheets

05. Data backup in Google Sheets

Data cleaning

Data integrity

Data integrity

> Data integrity refers to a **set of rules and processes** that are implemented to govern how data are entered, stored, processed, and transferred.

The importance of data integrity

- **Effective decision-making:** Organizations that regularly base their decisions on data can be misled by inaccurate and unreliable data brought about by a lack of data integrity.
- **Regulatory compliance:** Certain industries have regulatory requirements in terms of data accuracy, consistency, and reliability. Failure to meet these requirements can result in legal consequences.
- **Improved reusability:** With strong data integrity, data can be tracked and reused more easily in the future.
- **Minimized data risks:** Maintaining data integrity can help reduce common data risks such as loss or alteration.

Data integrity ensures that **data remain accurate, complete, consistent, and reliable** over its entire lifecycle, from creation to deletion.

Through data integrity, we can **rely on** and **trust our data** to be used **for its intended purpose.** It gives us some confidence that the data have not been corrupted, altered, or lost during storage, transmission, or processing.

Factors affecting data integrity

Factors that pose a **risk to data integrity** include:

Human errors

Human error, whether intentional or unintentional, such as data entry mistakes and failure to follow appropriate data protocols, is one of the top causes of data integrity issues.

System failures

Issues can occur as a result of system failures such as hardware or software crashes or power outages, which can corrupt and limit or eliminate access to data.

Bugs and viruses

These are malicious software that can infiltrate the systems and compromise data integrity by altering or deleting data.

Cyberattacks

Security breaches with the intention of damaging or stealing data can minimize the integrity of our data.

Transfer errors

While data is being moved from one system to another, errors can occur, leading to data inconsistency and inaccuracy.

Maintaining data integrity

Below are some **best practices** that can be used to preserve data integrity. We will look at how we can implement some of them in Google Sheets.

 **Data access control**	Issues of unauthorized access can be minimized when we control who has access to what data in a way that ensures a person only has enough data to perform their required tasks.
 **Data backup**	To prevent data loss brought about by system failures, cyberattacks, etc., a reliable data backup is essential.
 **Audit trail**	It is necessary to keep track of data changes involving events such as creation, deletion, and updating, along with details such as what, when, and by whom the changes were made.
 **Security**	Security measures such as authentication, monitoring, setting up firewalls, and physical security can help protect data systems from breaches.
 **Validate input data**	It is important to verify or validate input data, regardless of the source, to ensure that it is correct and accurate before letting it into your storage system.

Data cleaning

Data validation in Google Sheets

Data validation in Google Sheets

Google Sheets has a data **validation feature** that allows us to **add rules** to control the data that can be added to a particular row, column, or cell. This helps to keep **data consistent** and prevent misspelled or unwanted values.

Ways to validate data

- Add a list of **predefined values** to choose from.
- **Limit input** to a specific range of numbers, value, or data type.

When a user tries to enter data that **doesn't meet the validation criteria**, Google Sheets will display an **error message** or prevent the user from entering the incorrect data.

Why validate data?

Setting perimeters around data entered:

- **Prevents errors** in the spreadsheet by notifying users in case of incorrect or wrongly formatted data.
- Makes **collaboration more seamless** as there is control over data that can be entered.
- **Ensures consistency** in the data as rows and columns get to follow the same rules.
- **Saves on time** spent to clean the data.

Task overview

We have a dataset containing the maize prices in some major towns in Kenya from January to March 2022. Suppose we have gathered additional data on maize prices for the month of April 2022.

Action:

We want to **append** this data to our already existing dataset.

We will **set up some data validation rules** in the current dataset to ensure that the new data entered meet certain requirements, failure to which an error message will be displayed in the cell.

	A	B	C	D	E	F	G
1	**date**	**market**	**code**	**unit**	**kes_price**	**usd_price**	**ex_chrate**
2	1/15/2022	Nairobi	047122	KG	38.90	0.34	113.3448873
3	1/15/2022	Kisumu	042122	KG	36.54	0.32	113.3448873
4	1/15/2022	Nakuru	032122	KG	35.88	0.32	113.3448873
5	2/15/2022	Nairobi	047222	KG	40.04	0.36	112.6617895
6	2/15/2022	Kisumu	042222	KG	35.59	0.32	112.6622349
7	2/15/2022	Nakuru	032222	KG	34.15	0.30	112.6319261
8	3/15/2022	Nairobi	047322	KG	37.88	0.33	113.2456703
9	3/15/2022	Kisumu	042322	KG	34.57	0.31	113.2456703
10	3/15/2022	Nakuru	032322	KG	35.31	0.31	113.2456703
11							
12							
13							

By setting up these validation rules, we can identify some of the data errors as soon as the data are entered.

Data validation options

There are various data validation options in Sheets, grouped into categories as follows:

01. Dropdown

This option allows us to create a dropdown menu with a list of choices that can be input into a cell. There are two variations:

- **Dropdown:** The choices are entered manually. A text box will appear below where we add the values to be included in the list of choices.
- **Dropdown (from a range):** The choices are generated by cell reference. A text box will appear below where we specify the range of cells from which a list of choices will be created.

Example:

We can add a **dropdown** on the `market` column such that it only accepts the three city names we have in our dataset. This would also prevent other name variations other than the full name.

Data validation options

02. Text

This option provides conditions that enable us to limit text that can be input into a cell.

The following text validation options are available:

- Text **contains**
- Text **does not contain**
- Text **is exactly**
- Text **is valid email**
- Text **is valid URL**

Depending on the text validation option that we choose, a text box may appear requiring us to provide a specific text string.

Example:

We will use the **Text is exactly** validation option on the `unit` column and set the text to `'KG'`. This way, we can ensure that the unit of each observation entered is the standard 1 KG. We will also ensure that all the entries are uniform, i.e. `'KG'` vs `'1 KG'`.

Data validation options

03. Date

This option allows us to enforce validation rules on cells that contain dates.

The following date validation options are available:

- **Is valid** date
- Date **is**
- Date **is before**
- Date **is on or before**
- Date **is after**
- Date **is on or after**
- Date **is between**
- Date **is not between**

Depending on the date validation option that we choose, a text box may appear requiring us to provide a specific date(s).

Example:

We can use the **Is valid date** validation on the `date` column. An error will therefore be returned if the value entered is not a correct date.

Data validation options

04. Number

This option provides conditions that enable us to set up restrictions on the numerical values or formulas that can be input into a cell. The following number validation options are available:

- **Greater than**
- **Greater than or equal to**
- **Less than**
- **Less than or equal to**
- **Is equal to**
- **Is not equal to**
- **Is between**
- **Is not between**

A text box will appear requiring you to provide a specific numeric value(s).

Example:

To ensure only positive numbers that are greater than 0 are entered, we can use the **greater than** number validation option on the `price` and `ex_chrate` columns and set the number to 0.

Data validation options

05. Other

Additional validation options include:

- **Custom formula is:** This option allows us to insert a custom formula to be used as the validation rule.
- **Tick box:** With this option, a checkbox will appear in the cell, which we can check and uncheck to represent a specific value, e.g., checked for TRUE and unchecked for FALSE.

Example:

The code column, which is a unique identification code for each observation, is derived from the area code of the market, month and year. It can contain 6 digits only.

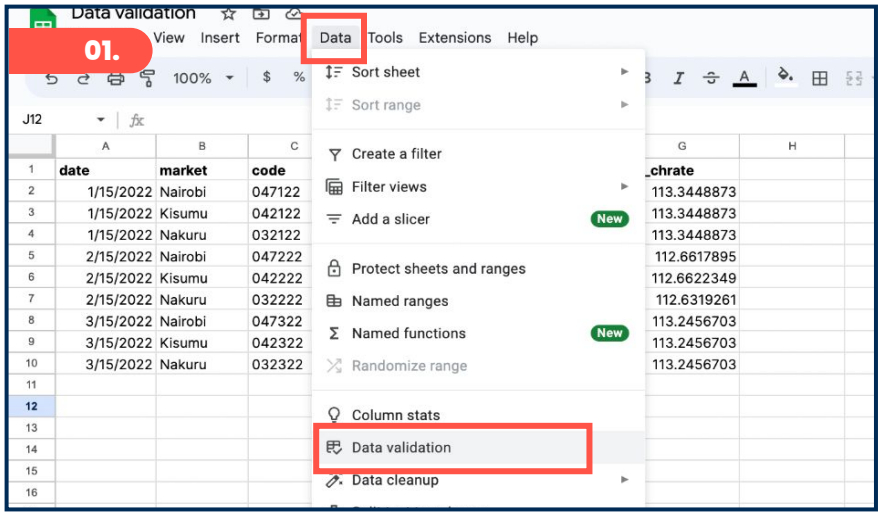
We can enforce this as a validation rule using the **Custom formula is** validation option. The formula we will use is `=LEN(C2:C25)=6` which ensures that the length of characters in a cell is exactly 6.

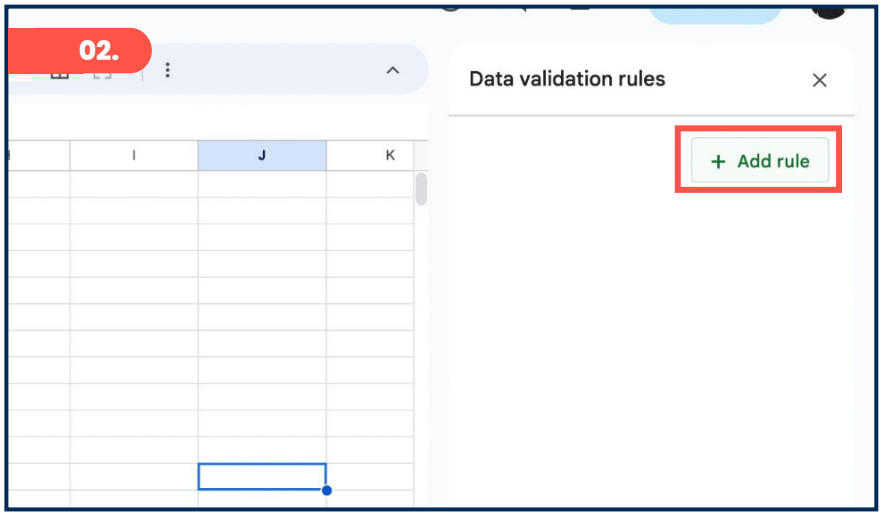
Data validation in Google Sheets

Steps:

01. Click on **Data** > **Data validation**.

02. A dialogue box appears on the right side of the Sheet. Click on '**+ Add rule**'.

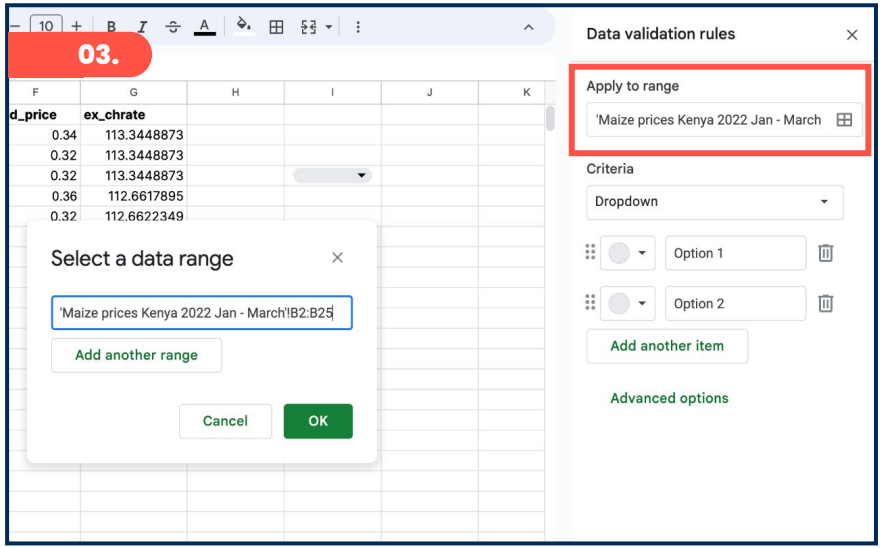


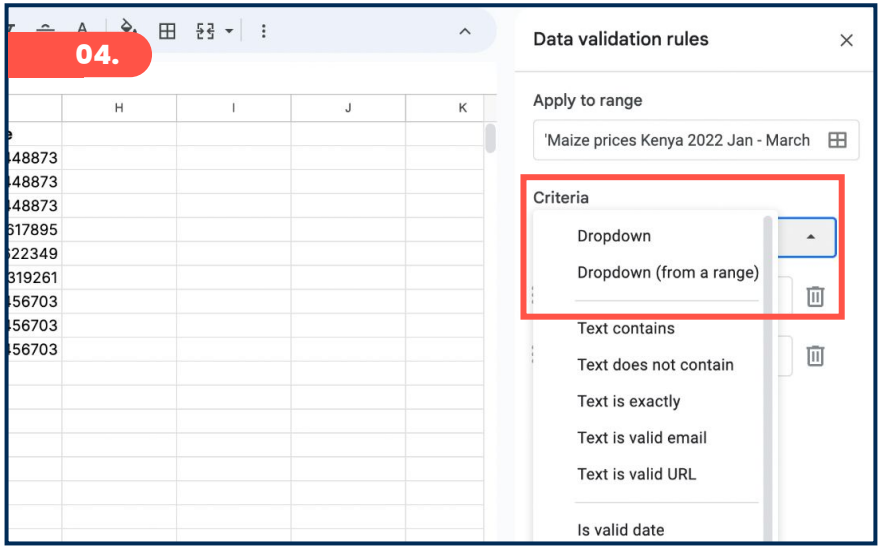


Data validation in Google Sheets

03. We are directed to the **Data validation rules** pane. We fill in the **data range** on which we want to apply data validation. Note to include some empty cells where the new data will be entered.

04. Under **Criteria**, we choose the data validation option we want to apply. Other text boxes may appear depending on the criterion picked.





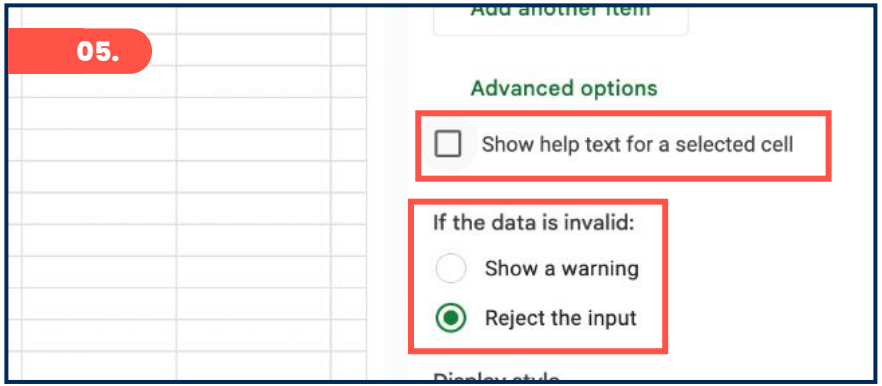
Data validation in Google Sheets

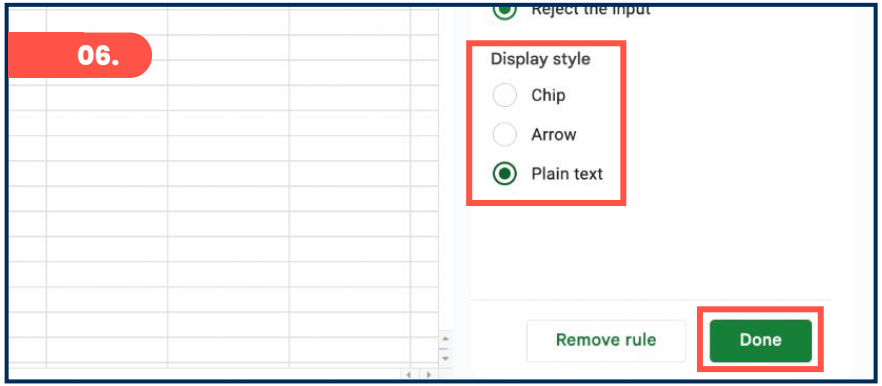
05. Under **Advanced options**, there is a '**Show help text for a selected cell**' checkbox which we select if we want to leave a note for the editor explaining how to make their input valid.

Next, we choose an action in the case where **invalid data** is entered. One is to allow the input of the invalid data but give a warning message and the other is to reject the invalid input altogether.

06. Additional settings may appear here depending on the criterion chosen. e.g. the **display style** in the case of a dropdown.

Click on **Done** for the rule to be added.



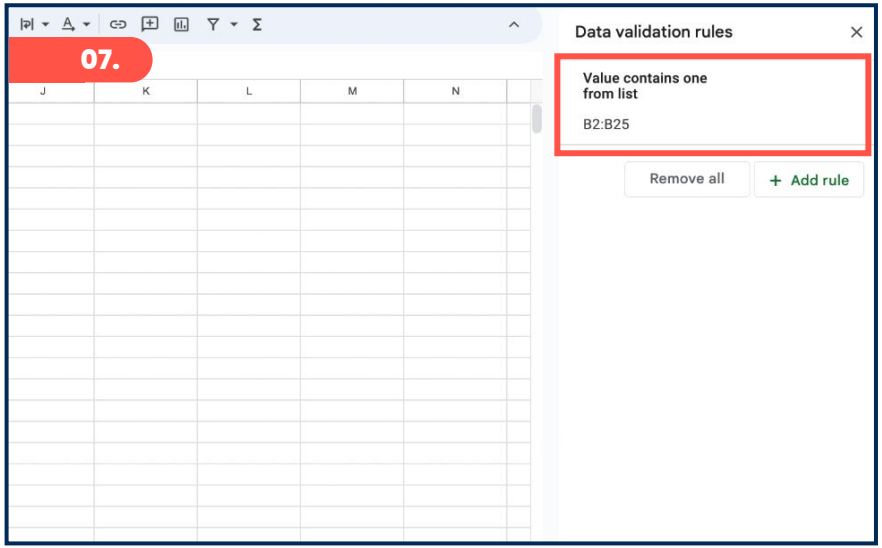


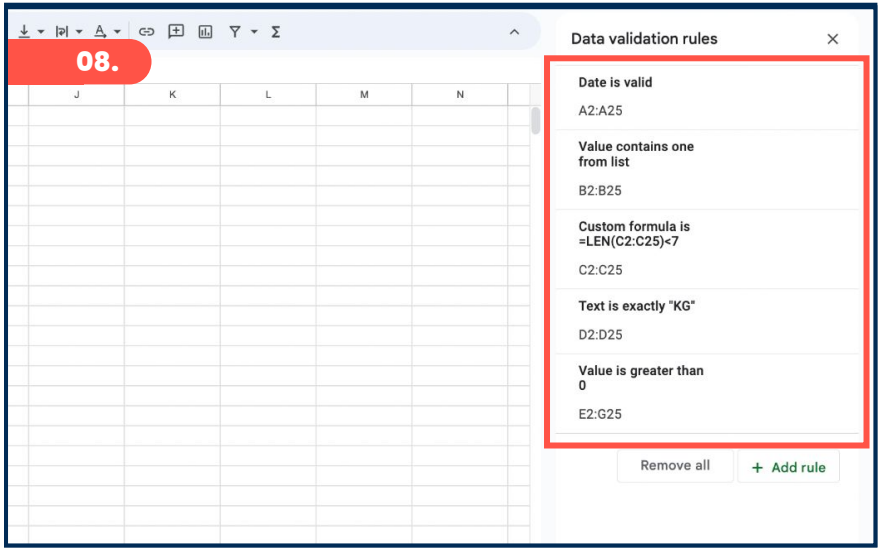
Data validation in Google Sheets

07. The rule will appear here on the **Data validation rules** pane.

To add another rule, we click on **+ Add rule** once again and repeat the steps above.

08. Eventually, a list of all the rules we have added will be displayed on the **Data validation rules** pane.





Data validation in Google Sheets

09. We append the new data to our dataset.

If an input violates the validation rules, a **small red triangle appears** in the top-right corner of the cell. Hovering over it displays the full error message.

We can now diagnose the problem and fix it to promote the integrity of our data.

Depending on the type of file our data are stored in, we can import and append the data directly into the current dataset, or import into a new sheet and then copy the data into the original dataset.

	A	B	C	D	E	F	G
		market	code	unit	kes_price	usd_price	ex_chrate
3	1/15/2022	Kisumu	042122	KG	36.54	0.32	113.3448873
4	1/15/2022	Nakuru	032122	KG	35.88	0.32	113.3448873
5	2/15/2022	Nairobi	047222	KG	40.04	0.36	112.6617895
6	2/15/2022	Kisumu	042222	KG	35.59	0.32	112.6622349
7	2/15/2022	Nakuru	032222	KG	34.15	0.30	112.6319261
8	3/15/2022	Nairobi	047322	KG	37.88	0.33	113.2456703
9	3/15/2022	Kisumu	042322	KG	34.57	0.31	113.2456703
10	3/15/2022	Nakuru	032322	KG	35.31	0.31	113.2456703
11	4152022	NRB	047422	KG	37.88	0.33	0
12	4-15-2022	Kisumu	042422	5 KG			3.2456703
13	4-15-2022	Nakuru	0032422	KG			3.2456703
14							
15							
16							
17							
18							

Data cleaning

Data access control in Google Sheets

Data access control in Google Sheets

The **permissions feature** in Google Sheets allows **control over the access of the data** contained within a Google Sheet document. We can restrict access to specific individuals and define the level of permissions they will have.

Different access levels in Sheets:

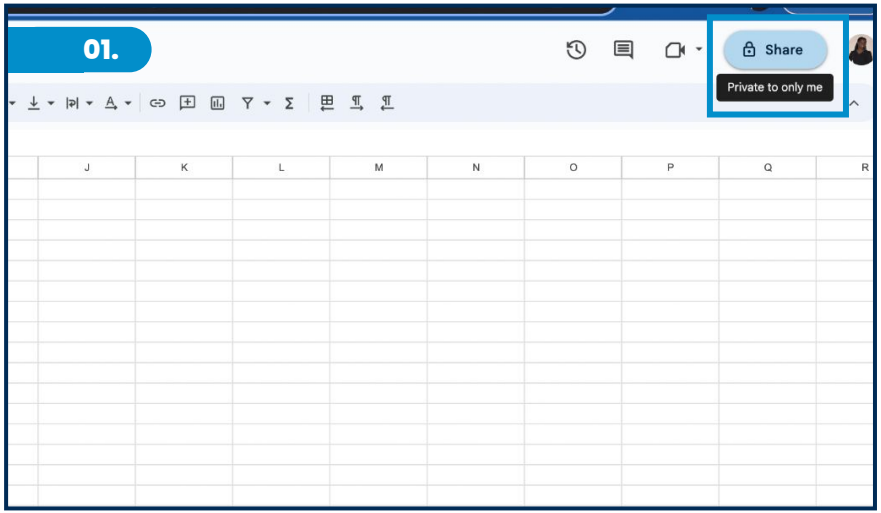
Owner	Editor	Commenter	Viewer
This level is by **default,** given to the person who **created** the file. This is the **highest permission level,** allowing the user to view, comment, edit, share, and delete the spreadsheet.	This permission allows the user to **view, comment, and edit** the data in the spreadsheet. They can also **share** the spreadsheet with others.	This level offers less freedom as the user is **only** able to **view and add comments** to the spreadsheet. To share the spreadsheet with others, they must send a request to the owner.	This is the lowest permission level, allowing the user to **only view** the spreadsheet's content.

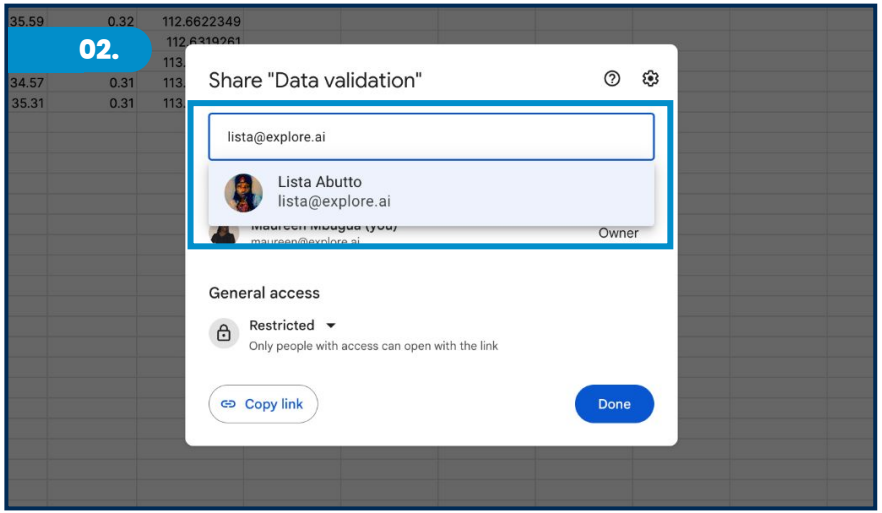
Setting permissions in Google Sheets

Steps:

01. In the top-right corner of the spreadsheet, click on **Share.**

02. Enter the email address of the person(s) you want to give access to.





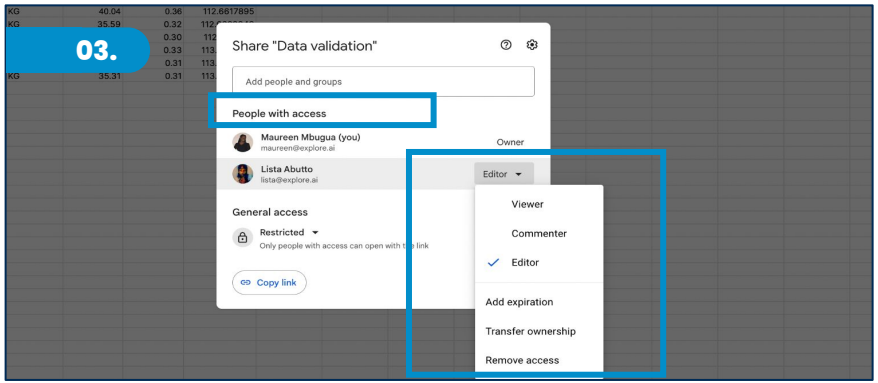
Setting permissions in Google Sheets

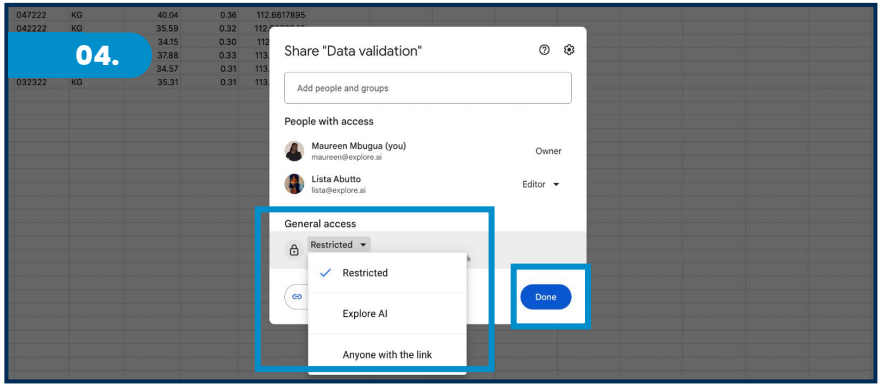
03. The person is added under '**People with access'**. From the dropdown next to their name, we **choose a permission level.** We can also add an expiration date for temporary access, transfer ownership, or remove access. You can change this anytime as you wish.

04. Under '**General access**', we set the sharing setting as either:

- **Restricted:** This restricts access to specific individuals, that is, those we have specified above under 'People with access'.
- **Anyone with the link:** This allows access to anyone as long as they have the link.

Click on **Done.** The permission will be granted and an email will be sent to them.





Locking cells in Google Sheets

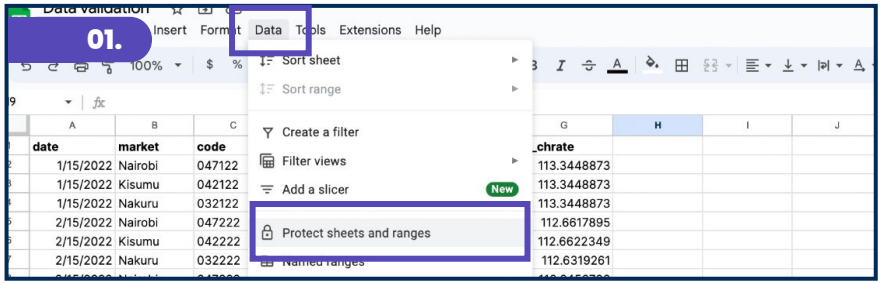
> Another way we can control data access in Sheets is by **locking cells**. This refers to the process of **protecting a cell, range, or entire sheet from editing**. This feature allows us to control who can edit specific data on a spreadsheet.

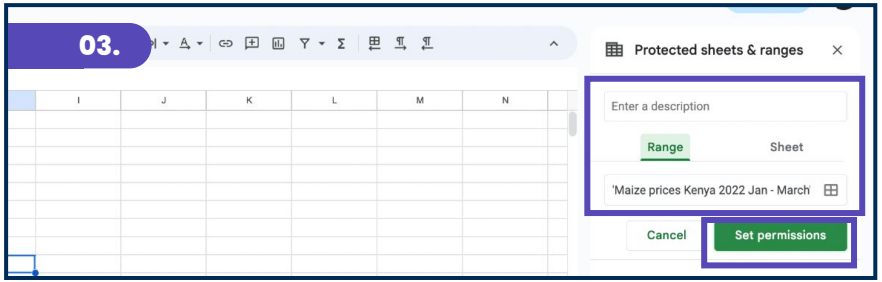
Steps:

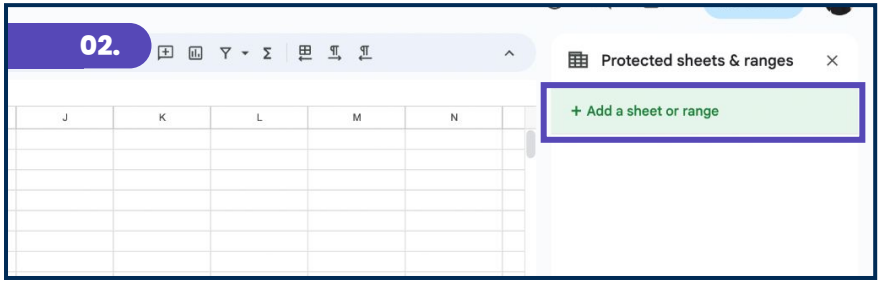
01. Click on **Data** on the menu bar. Select **Protect sheets and ranges** from the dropdown menu

02. A **Protected sheets & Ranges** dialogue appears on the right. Click on **+ Add a sheet or range**.

03. We specify either a **range or sheet** which we want to protect. We can also optionally add a **description**. Click **Set permissions**.







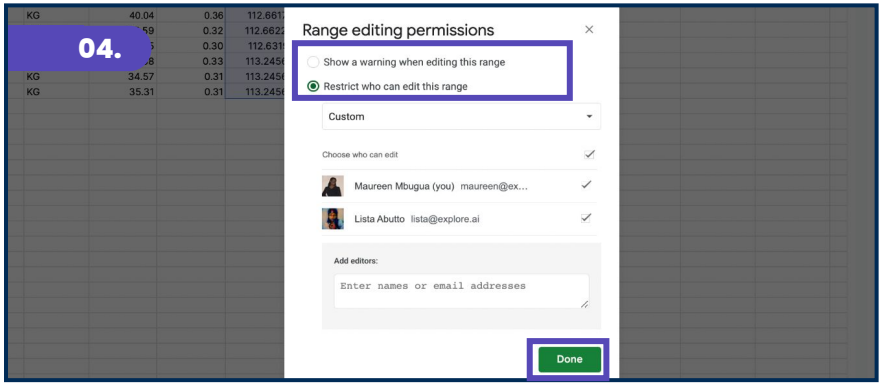
Locking cells in Google Sheets

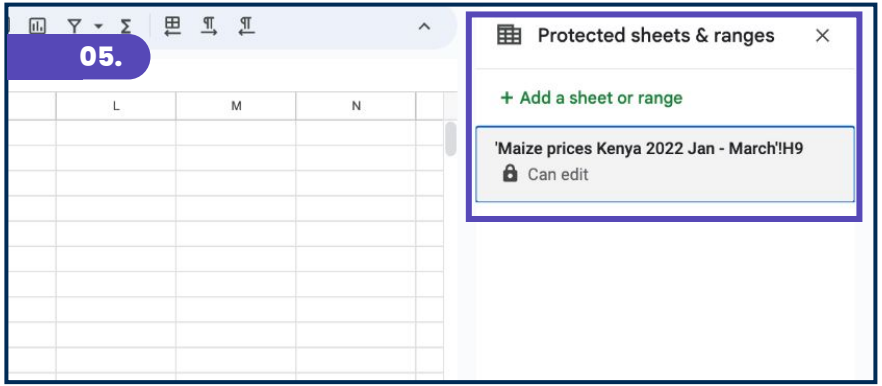
04. The **Range editing permissions** window appears. We can choose to either:

- Show a **warning** when this range is edited.
- Specify **who can edit** the range.

Click on **Done**.

05. The setting will be added on the **Protect sheets & ranges** pane. You can add another protected sheet or range.





Data cleaning

Tracking changes in Google Sheets

Tracking changes in Google Sheets

There are certain methods that we can use to **track changes** made in Google Sheets. This way, we implement an audit trail that enables us to know what changes were made, when, and by who.

Why is it important to track changes?

- **Get notified** – know that a change has been made.
- **Follow the history of change** – help understand why a change was made.
- **Know when a change was made and by who** – follow up on changes and get clarification.
- **Avoid losing data** – roll back to a previous version of the spreadsheet if necessary.

We can track changes in Google Sheets in the following ways:

- **Notification rules**
- **Version history**
- **Cell edit history**

Notification rules

> This feature allows us to be notified, through email, of any changes made within the spreadsheet.

Steps:

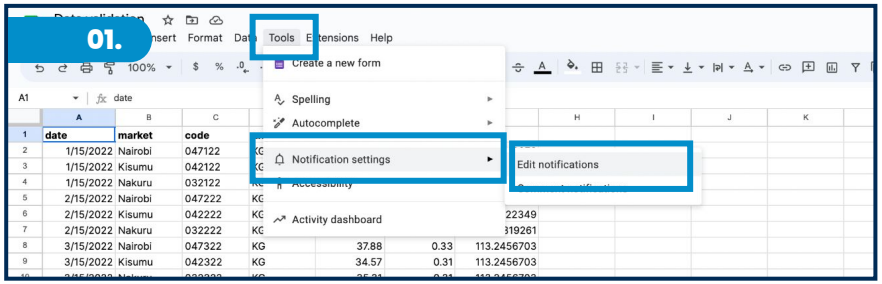
01. Click on **Tools** > **Notification settings** > **Edit notifications**

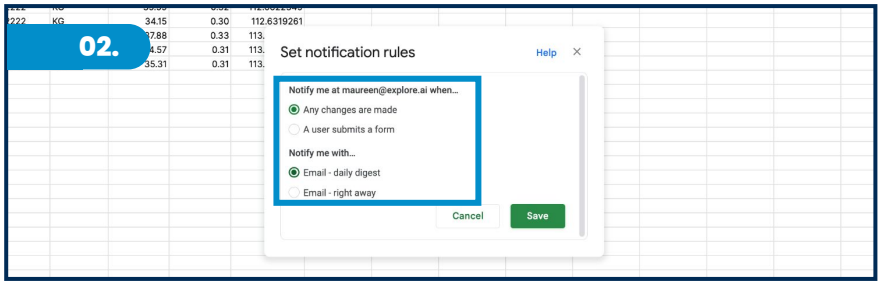
02. On the **Set Notification Rules** window, we choose the **type of change** we want to be notified of:

- Any changes made to the spreadsheet.
- When a user submits a form – this is useful if you are using Google Forms to collect information for your spreadsheet.

We also specify the **frequency** of receiving notifications:

- As soon as a change is made.
- A single daily digest with a list of changes made in a day.



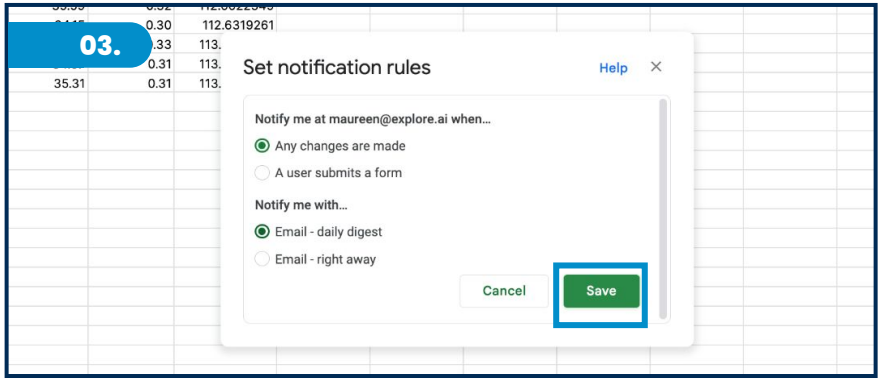


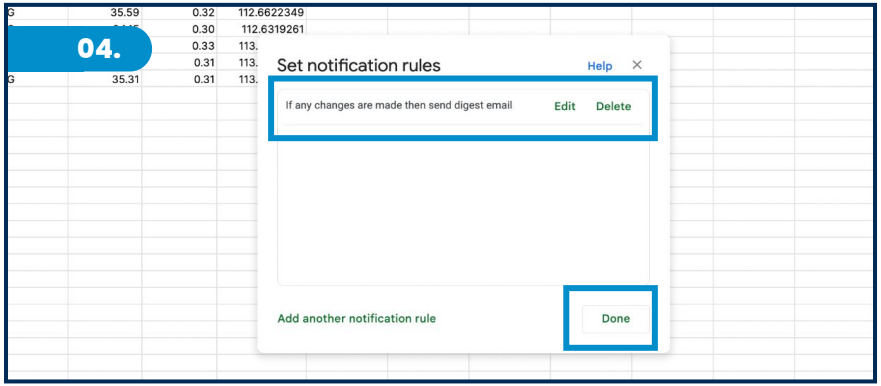
Notification rules

03. Click on **Save**.

04. The rule will be **added** to the **Set notification rules** window.

Here, you can edit, delete, or create another notification rule, then click on **Done**.





Data cleaning

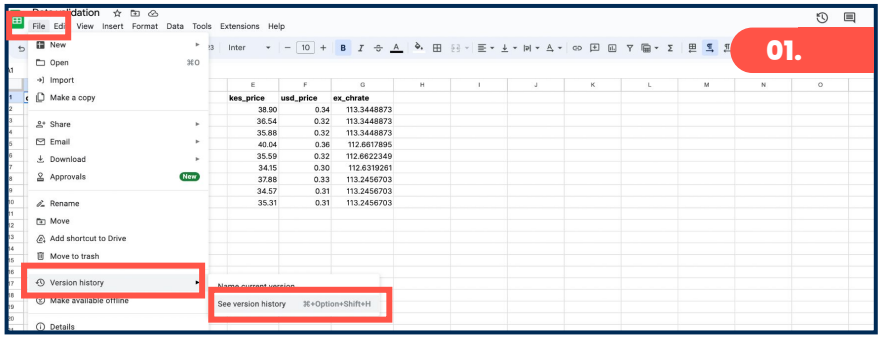
Version history

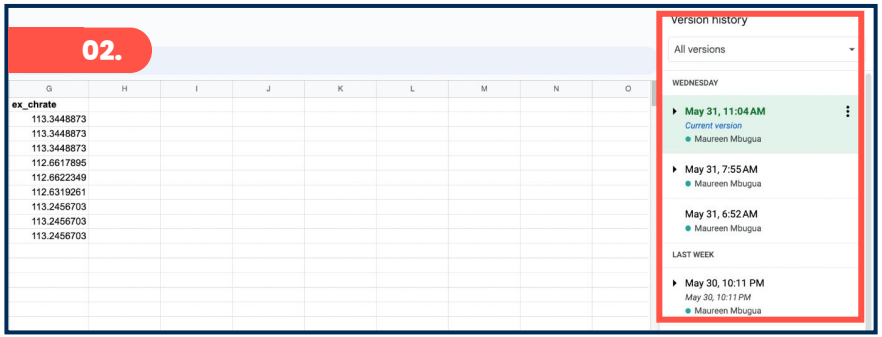
> Google Sheets has a version history feature that **tracks and keeps a record of all the modifications** made to the document over time. This allows us to view and revert back to earlier versions as necessary.

Steps:

01. Click on **File** > **Version history** > **See version history**

02. The **Version history** sidebar that appears on the right contains a list of **current and previous versions**. The **date, time,** and **user** responsible for the changes are also displayed for each version.

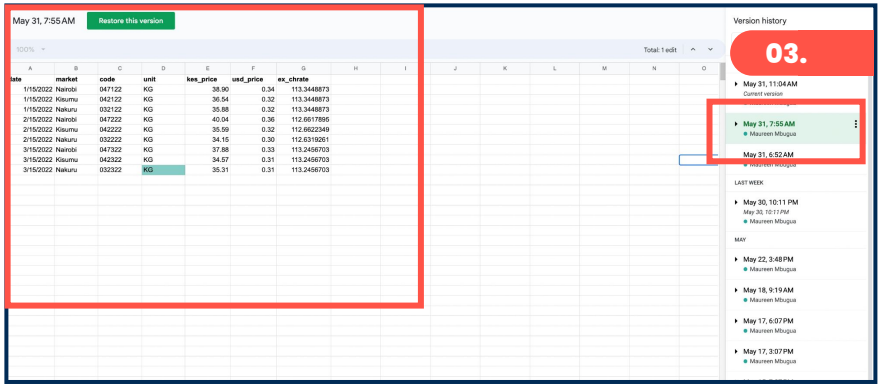


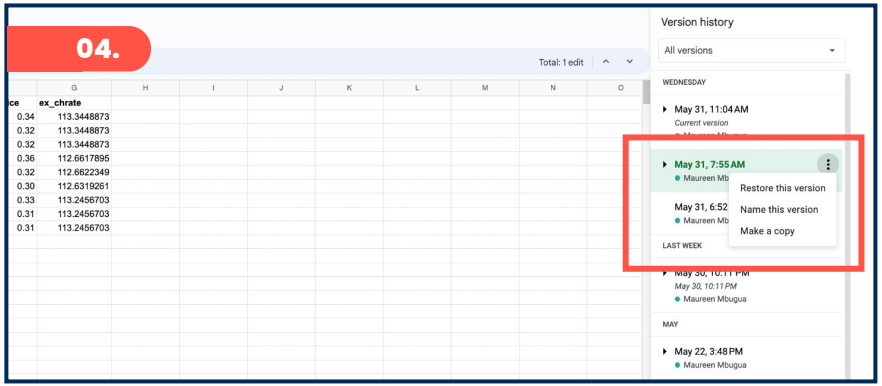


Version history

03. When we select a particular version, on the left we get a preview of the state of the spreadsheet during that version. We can restore that version by clicking on the **Restore this version** tab at the top.

04. Or we can use the **'More actions'** menu for that version which gives us the option to **restore** it, **name** it, or make a **copy** of it.





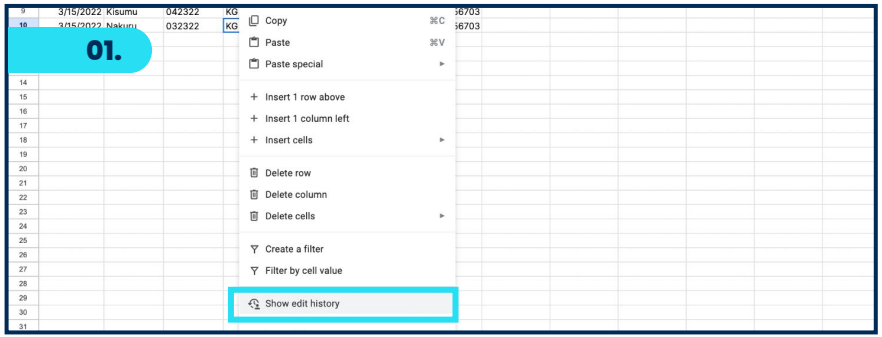
Cell edit history

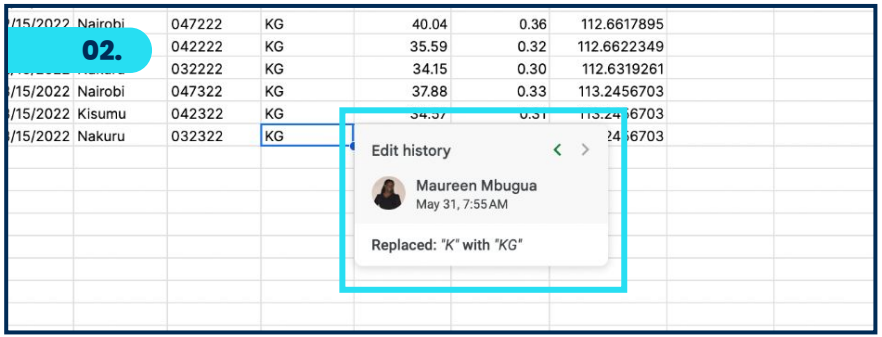
> Google Sheets offers a feature that allows us to view the **edit history of a specific cell** rather than the entire spreadsheet. These are changes made to the cell values, and not the formatting.

Steps:

01. **Select** the **cell** > Right-click on it to open the **context menu** > **Show edit history**.

02. A tiny **Edit history** window appears next to the cell. We scroll back and forth using the arrows to view the different versions. For each version, we can see when the edit was made, by who, and the specific change that was made.





Data cleaning

Data backup in Google Sheets

Data backup in Google Sheets

> Data backup in Google Sheets refers to the process of **creating copies** of your spreadsheet file to **ensure its availability** in the case of accidental data loss.

There are several methods we can use to backup our data in Google Sheets:

01. **Export the spreadsheet file**

We can follow the exporting process to **manually download** our Google Sheet onto our **local machine** in a preferred file format. This provides an offline backup which we can keep on our computer or any other external storage device.

This should be done frequently to keep the downloaded file up to date with the online one.

Data backup in Google Sheets

02. Make a copy of the spreadsheet file

We can make a copy of our Google Sheet in a preferred destination in our Google Drive.

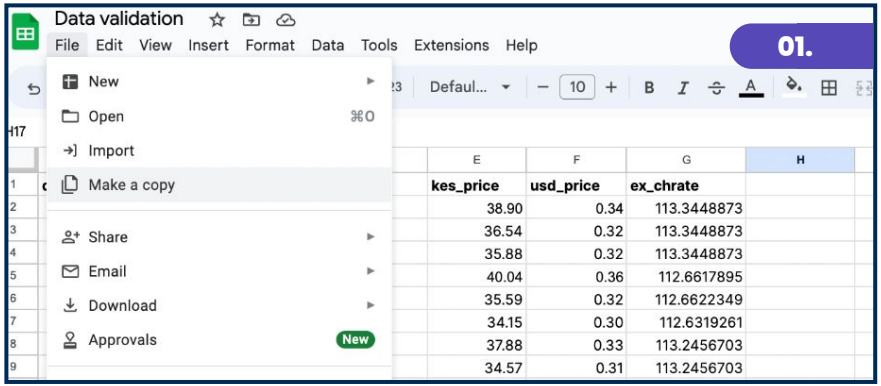
The copy made is separate from the original file. Any changes made to the original will not be updated on the copy.

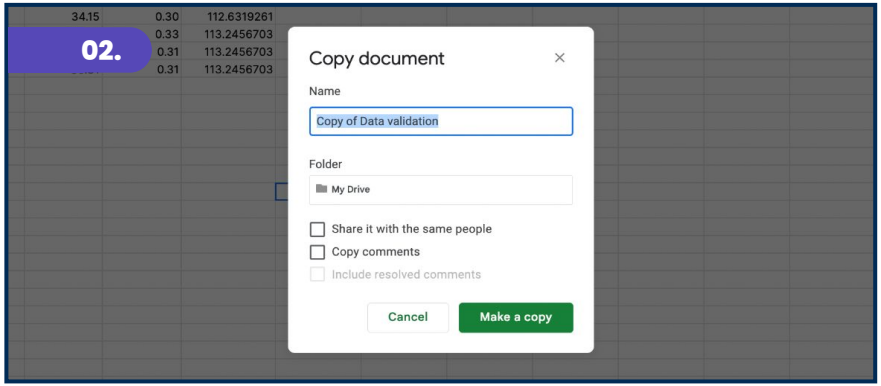
Steps:

01. Go to **File** > **Make a copy**.

02. A **Copy document** dialogue box appears. We specify a **name** and a **folder** within our Google Drive where the copied file will be stored.

We also choose whether we want to share it with the same people and include comments made in the original.





Data backup in Google Sheets

03. Use Google Drive for desktop

Google Drive for desktop enables us to **sync** our Google Sheets files stored in the cloud to a copy on our local computer. Any changes made to the synced files will be reflected both online and on our computer.

04. Use a third-party tool

There are also **third-party tools and services** available that specialize in backing up Google Sheets data. These tools provide features like scheduled and selective backups, among others.