## **Summary:**

**Sorting and Filtering :**

Sorting is the process of rearranging rows of data in a spreadsheet based on the values in one or more columns. Filtering is the process of hiding rows of data that do not meet certain criteria.

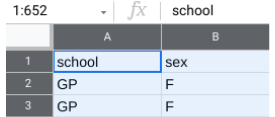
Sorting is often used to organize data in a more meaningful way, such as alphabetically by name or numerically by price. Filtering is often used to focus on a specific subset of data, such as only showing data for a specific time period or only showing data for a specific product.

Both sorting and filtering can be useful tools for working with large datasets in spreadsheets. They allow you to quickly find and analyze specific pieces of information, and to present your data in a clear and organized way.

**ACTIVITY:**

### **Sorting data**

1. To start, rename your spreadsheet. In the upper left corner, click **Untitled Spreadsheet** and enter a new name. You can use the name **student\_performance\_data** or a similar name that describes the data your spreadsheet contains.
2. Now, sort by school. Because you want to sort on multiple columns, you need to select all the data in your spreadsheet. Click the **blank rectangle** above row 1 and to the left of column A. This lets you select all the data on your sheet.



3. Next, from the menu bar, select **Data,** then **Sort range**. (Note: For some versions of Google Sheets, the selection **Advanced range sorting options** may appear on the **Data** drop-down menu instead of **Sort range**).

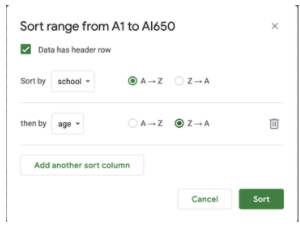
4. In the pop-up window, select **Data has a header row**. Now you can choose specific column headers to sort by.

5. In the **Sort by** dropdown, choose the header **school.** Then, click **A → Z** to sort in ascending order.

6. You also want to sort for age. Before you can sort by age, you need to click **Add another sort column** to choose a second column header.

7. In the **Sort by** dropdown, choose the header **age.** This time, click **Z → A** to sort in descending order. This way, the oldest students will be listed first.

Your popup window should appear like this:



8. Once both selections have been made, click **Sort**.

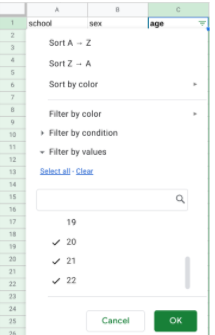
### **Removing incorrect data**

****

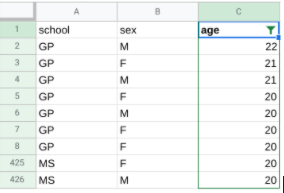
The superintendent tells you that the maximum age limit for which public education is provided is 19 years old and that the age range should be 15-19 for both schools. Any student outside this age range should be deleted from the dataset.

To clean your data, you need to remove the ages **20**, **21**, and **22** from your dataset. You can start by applying a filter to the **age** column. **Filtering** is the process of showing only the data that meets a specified criteria while hiding the rest. Filtering makes it easier to find data that you need.

1. First, apply a filter to the **age** column. Select the **age** column by clicking the letter at the top of the column (C).
2. Then, from the menu bar, select **Data,** then **Create a filter**.
3. You can now inspect the values in the **age** column by going to the top of the column and clicking the **Filter icon** ().
4. In Google Sheets, there are nine possible values for the field (**15, 16, 17, 18, 19, 20, 21,** and **22**). You may notice that all the values have check marks. Filter this column for the values you want to select by unchecking all the other values (**15, 16, 17, 18,** and **19**).



5. Then, click **OK**. This will single out the rows that contain the ages **20, 21,** and **22**. After you apply the filter, there should be nine such rows (seven for the GP school and two for the MS school).



6. To delete the nine rows, first select them by clicking their row numbers.

7. Then, from the menu bar, select **Edit** and **Delete selected rows**.

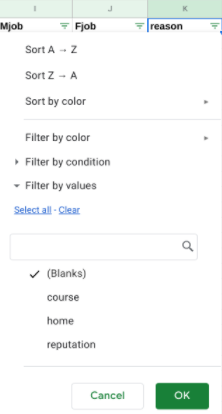
8. Click the **Filter icon** at the top of the **age** column to inspect the values once again. Now that you’ve removed the three incorrect ages (**20, 21,** and **22**), there are five ages remaining (**15, 16, 17, 18,** and **19**). The remaining ages are legitimate and can be used for analysis.

9. Finally, turn off the filter. From the menu bar, select **Data** and **Turn off filter**.

### **Filling in missing data**

Filling in missing data is an important part of data cleaning. It’s your job to fill in these blank spaces in your data with accurate values.

1. Start by applying a filter across the entire spreadsheet. Click on any cell in the sheet. Then, from the menu bar, select **Data** and **Create a filter**.
2. All the cells are now highlighted, and there are filters at the top of every column containing data. Click the **Filter icon** on the **reason** column (K).
3. You may notice that the data values in the **reason** column include blanks. Filter this column for **blanks** by unchecking all the other values (**course, home, reputation**).



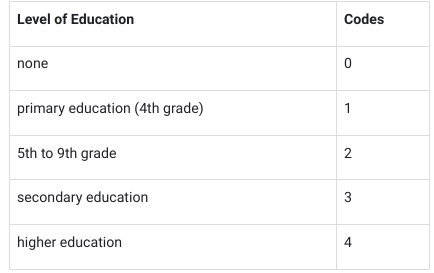
4. Then, click **OK**. Now, your sheet shows all the blank rows in the **reason** column.

5. To clean your data, you need to find a good way to fill in these missing values. In this case, you cannot know what each missing value should be (that is, without a new survey, you can’t discover each student’s reason for choosing a specific school). So, you can replace the missing values with the value **none\_given**. To do this while the column is still filtered for blanks, type **none\_given** in the first empty cell (K38). Then, press **Enter**.

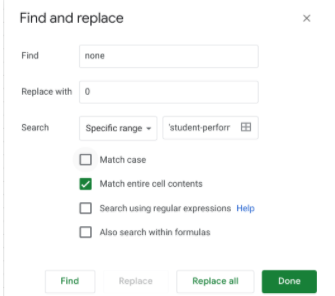
6. Select cell K38 again. A small blue square, known as the fill handle, appears in the bottom-right corner of the cell. Double click the fill handle to fill all the other blank cells with the value **none\_given**.

7. Finally, turn off the filter. From the menu bar, select **Data** and **Turn off filter**. If you scroll down the **reason** column, you should find that the value **none\_given** has replaced all the blanks in the **reason** column.

### **Converting data**



1. To start, remove the filter from the **Medu** column.
2. Next, select the unfiltered **Medu** column data by clicking its column letter (G).
3. Then, from the menu bar, select **Edit,** then **Find and replace**.
4. Fill in the popup window for the **none** value. Next to **Find**, type **none.** Next to **Replace with**, type **0**. Check the box next to **Match entire cell contents**.



5. Then, click **Replace all**.

6. While still in the popup window, repeat this process (**steps 4-5**) for the other four educational levels: **primary education (4th grade), 5th to 9th grade, secondary education,** and **higher education**.

7. After replacing all five educational levels with numeric values, click **Done** to close the pop-up window.

8. Check out your spreadsheet. All the cells in the **Medu** column now display numeric values.

9. Change the text data in the **Fedu** column (H) in the same way.

* Why is cleaning data such an important part of the data analysis process?
* How can sorting and filtering help you clean data more effectively?

The dataset could have biased, missing, null values which could have a adverse effect on our results. So cleaning the data is important to land on a correct conclusion. Sorting and filtering let's us arrange the data in some fashion which could highlight the defected data making it easy to remove it or make appropriate changes.

| **Question** | **Spreadsheet** | **Database** |
| --- | --- | --- |
| How do they store data? | Stores data in cells. | Stores data in tables. |
| How are they used to interact with data? | Calculation, graph and reports (GUI) | Store, search and sort (CLI) |
| How powerful is each? | Less powerful, cannot handle large dataset with ease | Powerful, can handle large datasets with ease |
| What are their pros and cons when sorting? | Ability to easily sort small datasets with inbuilt functions, poor handling or large datasets | Can sort large datasets with ease but requires advance knowledge |
| What are their pros and cons when filtering? | Simple and easy to use inbuilt functions, dataset size | Can handle complex filter operation with ease but may require advanced knowledge |