# Qwiklab: Sorting and Filtering (Cleaning Data)

# Cleaning Data

**Introduction**

Roll up your sleeves because we are going to get our hands "dirty" cleaning data. This lab is about cleaning data to get it ready to be analyzed. You are going to take a "dirty" dataset and clean it up using basic commands and key spreadsheet functions taught right here in this Qwiklab.

Data cleaning corrects or removes incorrect, missing, and faulty data. This is very important. "Bad" data can lead to wrong conclusions. The better your data, the better your results. Believe it or not, for most projects as a data analyst you will spend about 80% of your time cleaning data.

**Let's look at a real-world situation**. **A school district wants to know what affects student grades in core subjects.** To do this, they are going to analyze some student data. It is extremely important that whatever database they decide to use, it must have clean data. Analyzing bad data could cause the district to reach the wrong conclusions that could move instruction in a direction that does not improve student grades. Here is the dataset we are going to use [student-performance-data](https://drive.google.com/u/0/uc?id=1vMFJsvsIgzj1tXZge2QoDAqLql2zq63S&export=download). It contains information on high school student achievement in two Portuguese schools. **It was collected using school reports and questionnaires**. The data includes information such as:

* Student grades in Math and Portuguese.
* Student background information, also called demographics.
* Student study time.
* Student travel time from home to school.
* Student participation in extracurricular activities.

What you'll do

For this lab, we will use Google Sheets. You will receive a temporary login as part of this lab. To do this, you will need to set up a Google account. There are other options in the Data Analyst's toolbox as well, such as Microsoft Excel and Libra Office 3. No matter what tool you use, the steps we'll follow will be similar. You **will use some spreadsheet actions like**:

* Import data from a given file into a workbook
* Sort and organize data in different ways
* Apply filters to look at data in various ways
* Use filters to handle missing data
* Use conditional formatting to highlight specific categories of data
* Use formulas to summarize data

**You will have 90 minutes to complete this lab.**

**Setup**

**Before you click the Start Lab button** read these instructions. Labs are timed and you cannot pause them. The timer, which starts when you click **Start Lab** , shows how long Google Cloud resources will be made available to you.

This Qwiklab hands-on lab lets you do the lab activities yourself in a real cloud environment, not in a simulation or demo environment. It does so by giving you new, temporary credentials that you use to sign in and access Google Cloud for the duration of the lab.

What you need

To complete this lab, you need:

* Access to a standard internet browser (Chrome browser recommended).
* Time to complete the lab.

**Note:** If you already have your own personal Google Cloud account or project, do not use it for this lab.

Sign in to Google Workspace

1. Click the **Start Lab** button at the top of the screen. 
2. Click **Open Workspace Admin Console** in the upper-left panel and enter the **username** and **password** as prompted.

**Note:** Make sure you signed into Google Workspace using an **incognito window.**

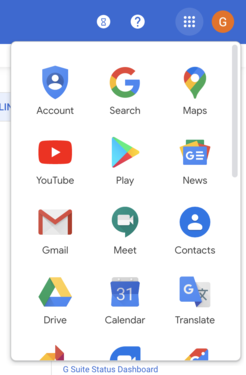
1. Copy the link and paste it into your incognito window. Then enter the **username** and **password** as prompted. You can check **time left** in the top left corner. 

Read and then **Accept** the terms of service and start the free Google workspace Trial.

The **Admin console** opens.

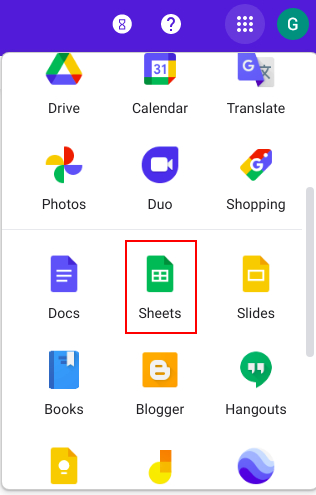
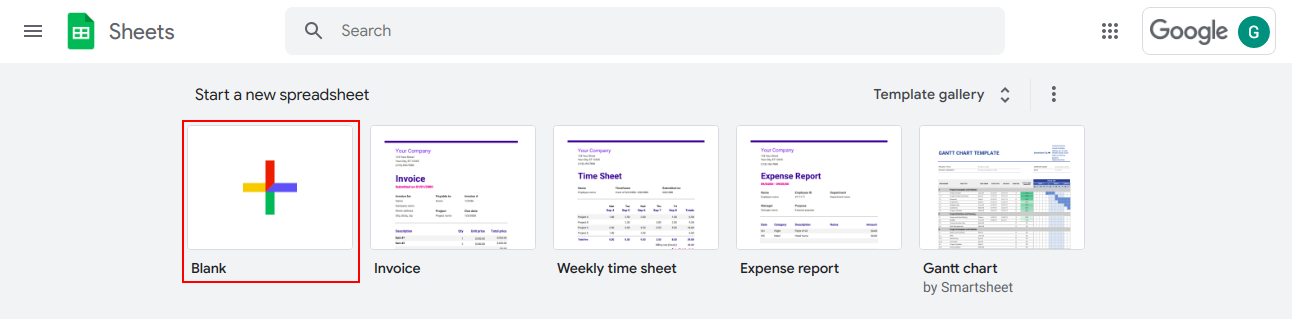
**Note:** Wait for 3-5 minutes before using any Google Apps. Meanwhile, try opening the required Google App after 3 minutes until it opens successfully.

You can navigate to any Google workspace application from this page by clicking the **Google apps** grid **icon** in the **top right**. If you do not see the icon, **widen the browser window**.



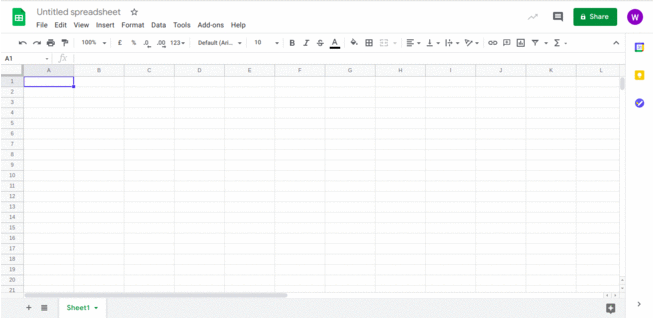
Do not click **End Lab** unless you are finished with the lab or want to restart it. This clears your work and removes the project.

1. Click on the Google apps grid icon in the top right and select the Sheets App. Then start a new blank spreadsheet by clicking on Blank.

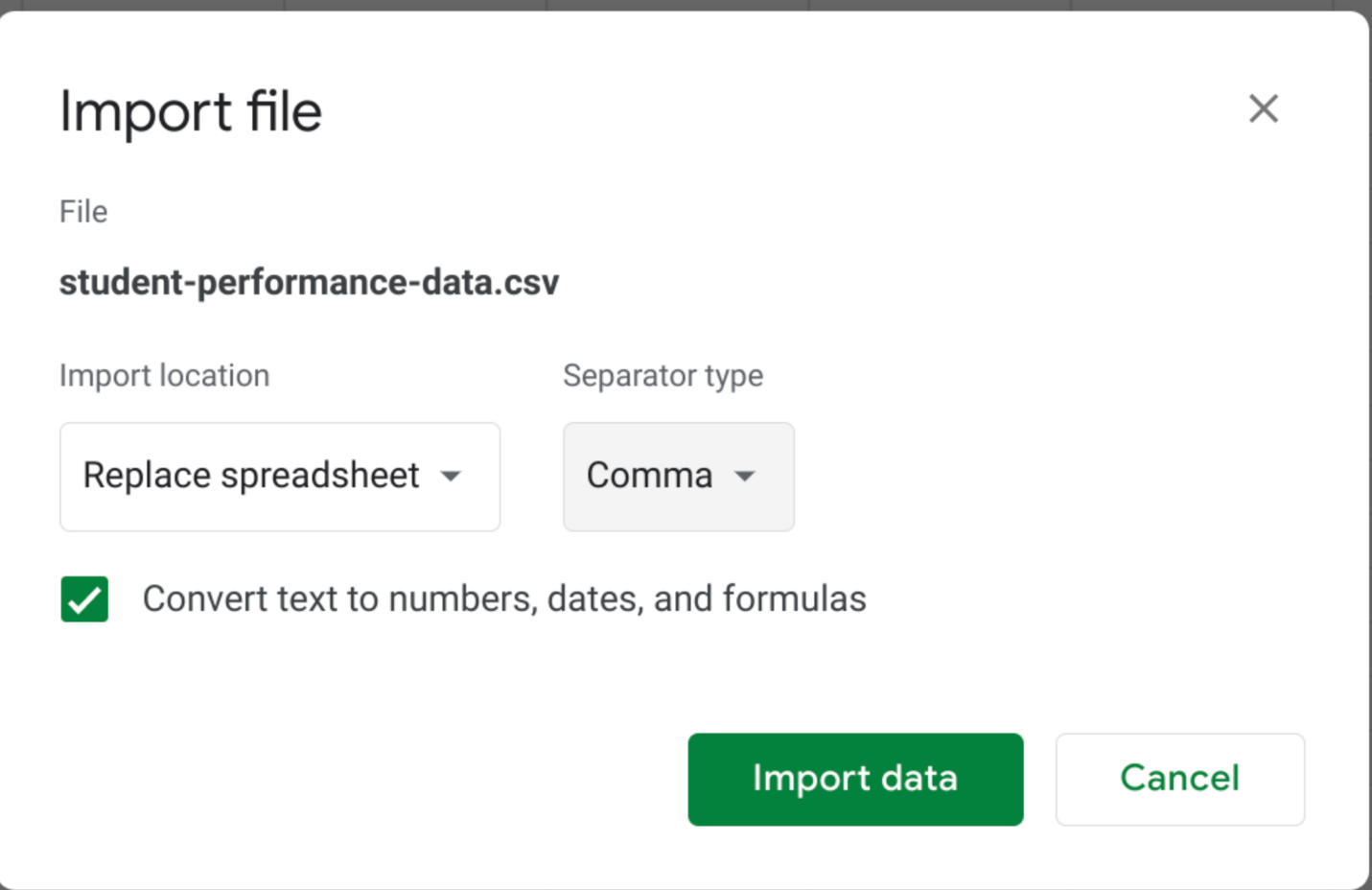
 

**Importing the data**

First, you need to import your data into a spreadsheet. If you've not already done so, click on the link to the dataset file given above to download it to your computer. In Google Sheets, you will go to the top menu bar and click **File → Import**.



You will be prompted to upload the file. Click upload and then click **Select a file from your device** and then navigate to your downloaded file. If you did not make any changes to the file name when you downloaded it, the name of the dataset file should be *student-performance-data.csv*. Next, on the popup window, make the following selections:



Click **Import data**. Now the dataset has been imported!

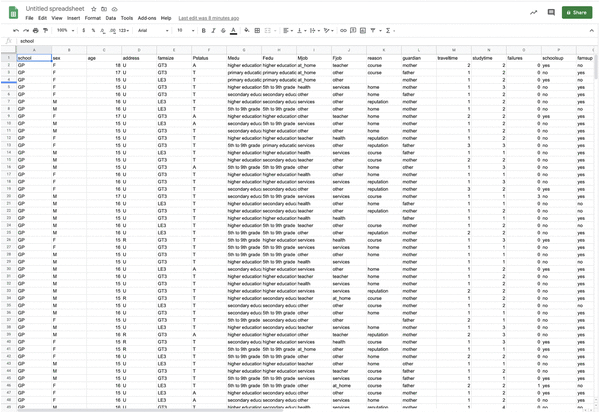
You can see if the import was a success by scrolling right and then down on your spreadsheet. You should see 33 columns and 650 rows of data. A quick way to check this is to select the row labeled 1. The status bar at the bottom should say **Count: 33.** Then select the column labeled A. The status bar should say **Count: 650**.

Row count Coloumn count

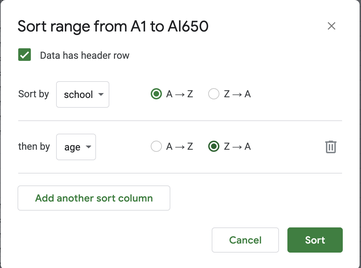
**Sorting the data**

Sorting is organizing data so that you can analyze it easier. Since we are cleaning data from two schools, Gabriel Pereira (GP) and Mousinho da Silveira (MS), it might be useful to sort the data by school. **Then we could also sort by age to see the age ranges of the students for each school. Let's try it!**

Because we want to sort on multiple columns, we will need to select all the data on the sheet. Click the blank rectangle above row number 1 and to the left of Column A.



The entire sheet's data should be selected. Now it can be sorted on multiple columns. To sort by school on the top menu bar, select **Data → Sort range**. In the popup window, make the following selections.



Before we can sort by age, we need to choose **Add another sort column** (see above).

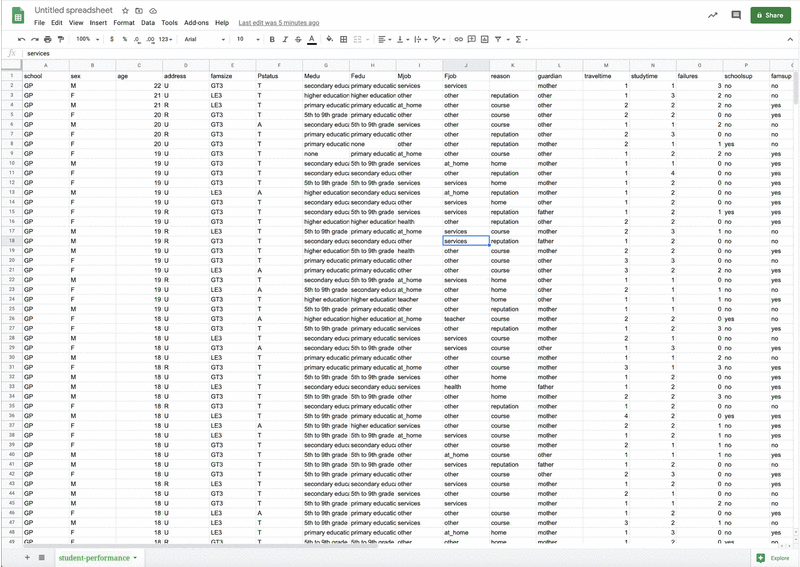
Once both selections have been made, click **Sort**.

**If you scroll through the data you will see that the age range of Gabriel Pereira (GP) is 15 - 22 years** and **the age range of Mousinho da Silveira (MS) is 15 - 20 years.** It appears that both schools have similar age ranges. The fact that the GP school has students that are a little older could be a problem, but we will deal with that later.

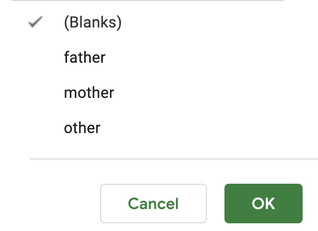
**Inspecting the data**

To really clean the data, you will need to know what data values and data types there are in each column. This is called inspecting the data. A good way to inspect the data in each column is by applying a filter. **The filter will make it easier to find data that needs to be cleaned.**

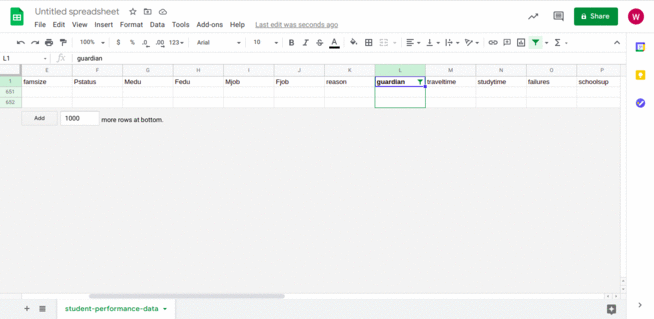
Let's first apply a filter to the *guardian* column (column L). Select the *guardian* column data by clicking its column letter. Then from the top menu bar select **Data → Create a filter**.



You can now inspect the values in the *guardian* column by going to the top of that column and clicking the Filter filter sign. In Google Sheets, there are four possible values for this field (mother, father, other, and blanks). Filter this column for the blanks by unchecking everything else. Then click **OK**.



**Doing this will single out the blanks (rows 651-652) so that they can be deleted.** To delete the two blank rows, select them by clicking the row numbers. From the top menu bar select **Edit → Delete selected rows**.

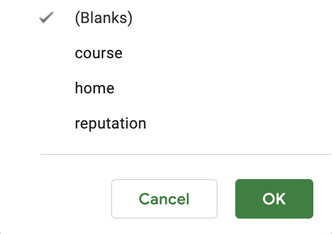


Click the filter to inspect the *guardian* column once again. Now there are only three possible values for this field (mother, father, and other) all of which provide useful information.

**Filling in missing data**

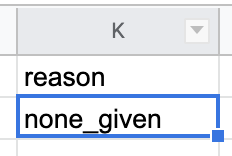
From the top menu bar, select **Data → Turn off filter**. Let's apply a filter across the entire spreadsheet. Click on any cell in the sheet. From the top menu bar select **Data → Create a filter**.

You should see filters at the top range of every column containing data. Click the filter on the *reason* column (column K). To understand what the values in this column mean, you can take a look at the [dataset documentation](http://archive.ics.uci.edu/ml/datasets/Student+Performance). You can see that this column has some blanks. Filter this column on only the blanks by unchecking everything else. Then click **OK**.



By scrolling right and then down, you can see that the blanks appear to be random. That is, looking at the data in the other columns for these missing values, there doesn't appear to be a pattern. For example, it doesn't appear to be the case that the values are missing for all students in GP school. It appears that the data is missing because the survey participant simply chose not to answer this question. **This is great news because these missing values should not produce distorted analysis results**. Although, we still need to find a good way to fill them in.

**Because we have no idea what these missing values should be, let's replace them with the value "*none\_given*".** **To do this while the column is still filtered on blanks, type *none\_given* in the first empty cell. Then hit ENTER. Now when you select that cell, a small blue square, known as the fill handle, will appear in the bottom-right corner.**



Double click it to fill the value *none\_given* to all other cells with missing values.

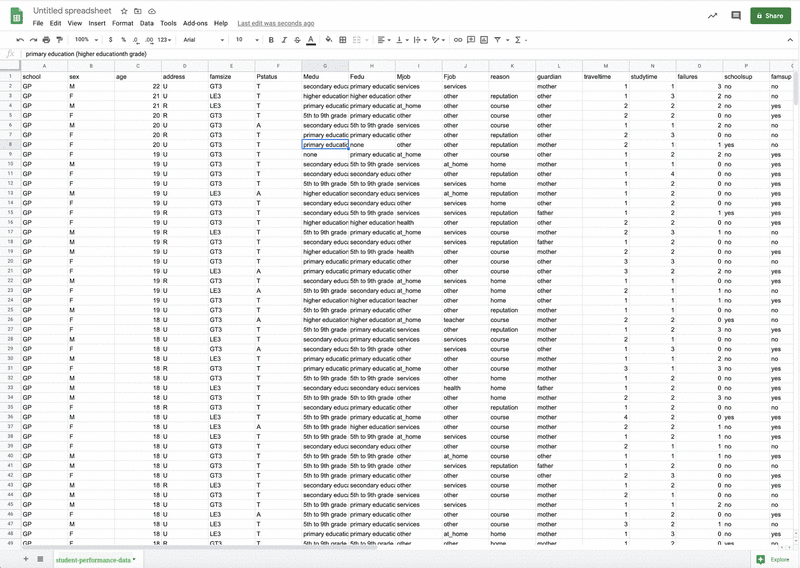
After turning off the filter, you can see that now the value *none\_given* is filled in for blank cells in the *reason* column.

**Converting text data format**

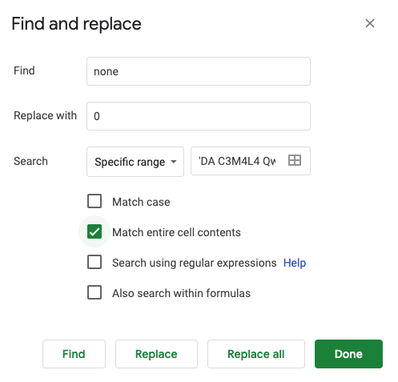
Continue to inspect the data by clicking the filter on the *Fedu* (Father's education) column(column H). This column has text data. For analysis, it's sometimes required to change text data (words) to numeric data (numbers). For example, some statistical packages like those used to perform machine learning will only accept numerical data values as input. To change the text to numbers for Father's education, we are going to use the following codes:

|  |  |
| --- | --- |
| **Education** | **Codes** |
| none | 0 |
| primary education (4th grade) | 1 |
| 5th to 9th grade | 2 |
| secondary education | 3 |
| higher education | 4 |

First, select the unfiltered *Fedu* column data by clicking its column letter. Then from the top menu bar, select **Edit → Find and replace**.



Fill in the popup window as follows for the *none* cell values. Then click **Replace all**.

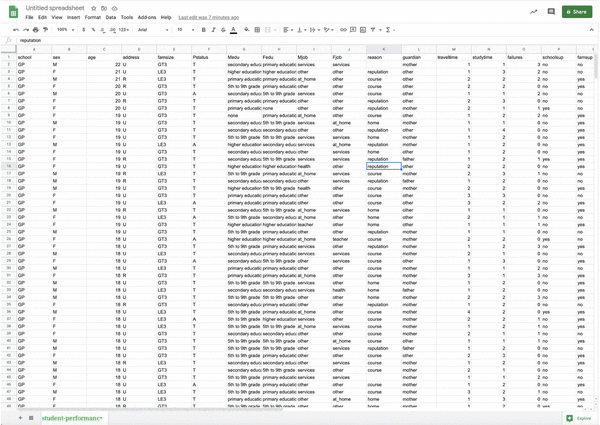


Repeat the same process for all the other educational levels. After replacing all the educational levels click **Done** to close the dialogue window. Change the text data in the *Medu* (Mother's education) column(column G) the same way.

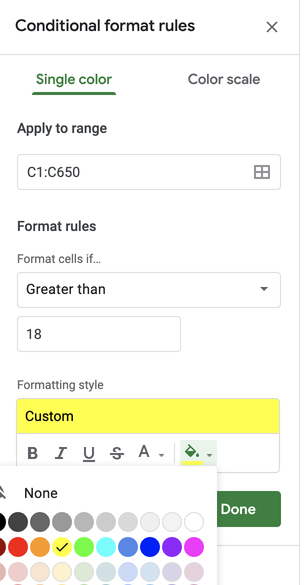
**Identifying potential errors in the data**

Now let's look at the *age* column. Because this dataset represents high school student achievement, it may be helpful to highlight students in the data who are over 18 years old. Any age older than 18 may mean a mistake was made when entering that student's age.

To do this, apply conditional formatting to the *age* column to highlight students older than 18. Select the age column, and then from the top menu bar, select **Format → Conditional formatting**.



On the popup window, click the **Format rules** dropdown menu and select **Greater than**. Then enter 18 in the **Value** textbox. Finally, in the Formatting Style section, choose yellow as the fill color.



Now that these cells are highlighted you know which ages may need to be researched and fixed.

**Using formulas to create new data**

It is sometimes simpler to analyze data by creating a new data column from one that is already on the spreadsheet. The newly created column can be used to synthesize data in your analysis. This is called data engineering.

Use *schoolsup*, *famsup*, *paid*, *activities*, *nursery*, *higher*, *internet*, and *romantic* to create a new column named "*positivity\_score"*. To understand what the values in these columns mean, you can take a look at the [dataset documentation](http://archive.ics.uci.edu/ml/datasets/Student+Performance). Create the new column by totaling all the yes's in each of the columns that are used to create it. First, insert the new column to the right of *romantic (column W)*. Then name the first cell: *positivity\_score*. To total all the yes's, use the **COUNTIF** function. In cell X2, type the following formula:

=COUNTIF(P2:W2,"=yes")

Then hit ENTER. Just as you did before, go back and select that cell, then double click the fill handle to copy this formula down the column.



You should now have a new column named *positivity\_score,* that is the total of all the yes's in each of the columns that were used to create it.