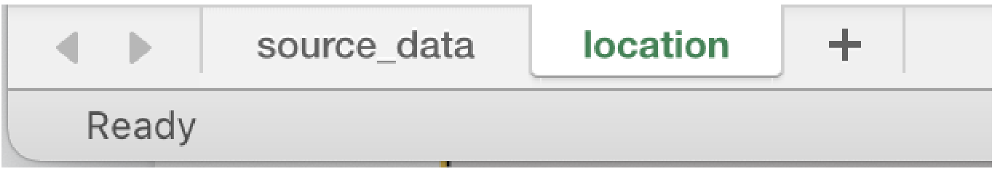
Step 0 – Prepare the data

0.1 – Rename the columns to match the relational model

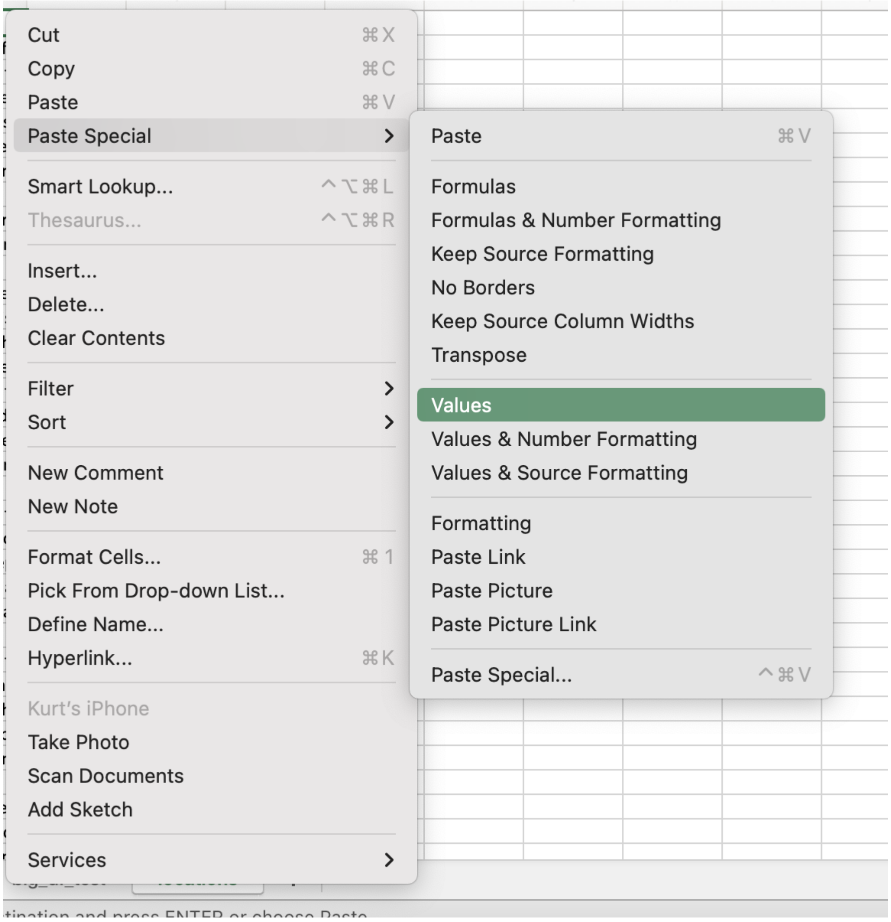
Step 1 – Create the "Location" table

1.1 – Create a "location" worksheet by renaming 'Sheet2' to 'location'



1.2 – In the "source" sheet, normalise the text in the scene_location column:

- 1.2.1 – Insert a new column into Column E (right next to the scene_location column). Also name it 'scene_location'
- 1.2.2 – Paste **FORMULA 1** into cell E2
- 1.2.3 – Apply the formula to the rest of the cells. To do this, go to the bottom right corner of cell E2 and double click or click on E2 and drag the formula all the way to the last row.
- 1.2.4 Once column E is fully populated with the normalized text, select the entire column and copy. With column E ('scene_location_norm') selected, right click and select 'Paste Special -> Values'. This will overwrite the column with the values. If you did this correctly, all the formulas should be removed and you should now only see the raw normalized text.



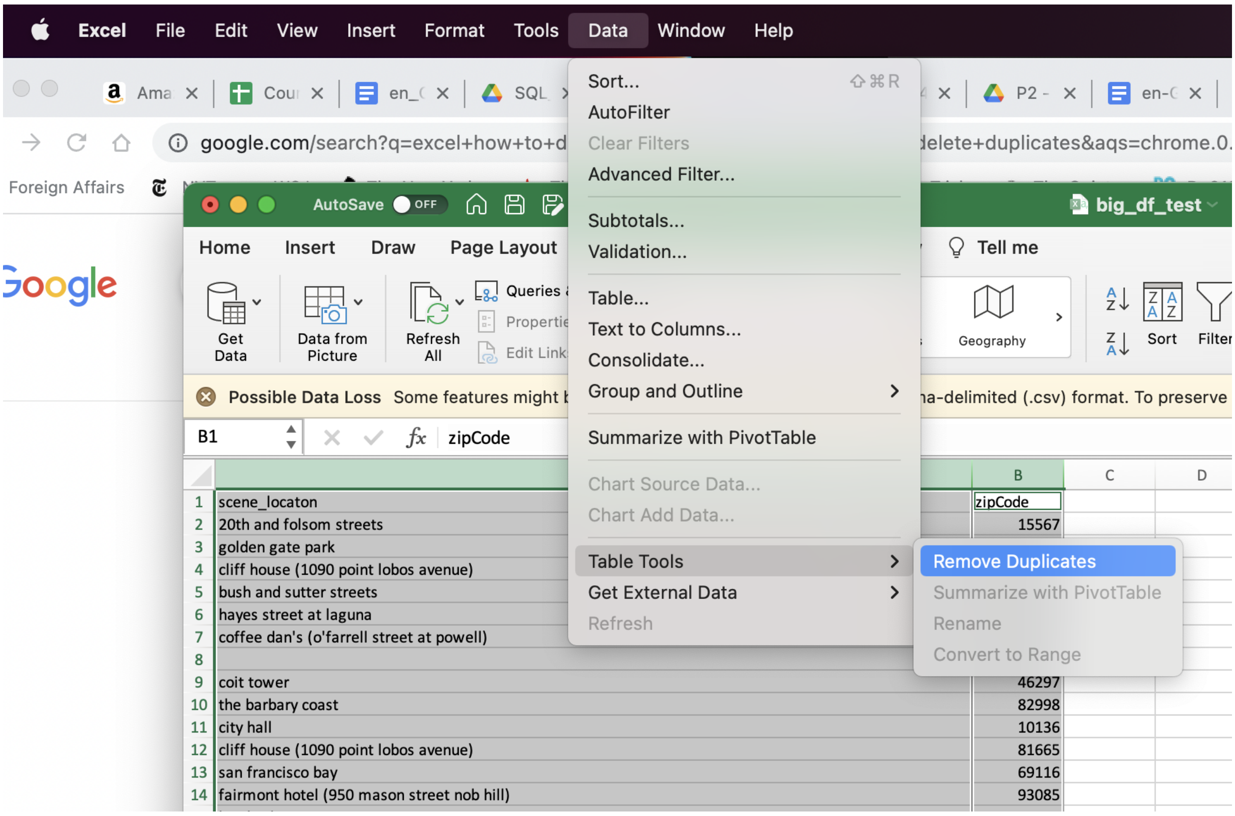
An example of “Paste Special -> Values”

1.3 – Now copy and paste the following columns from the “source” sheet into the “location” sheet in the same order:

- scene_location (column E)
- zip_code (column I)

1.4– **Delete duplicates*** (=identical rows) from the “location” table

- To do this, select / highlight the full range of both columns (shortcut = CNTRL + A), then go to (Data -> Table Tools -> Click Remove Duplicates)
-

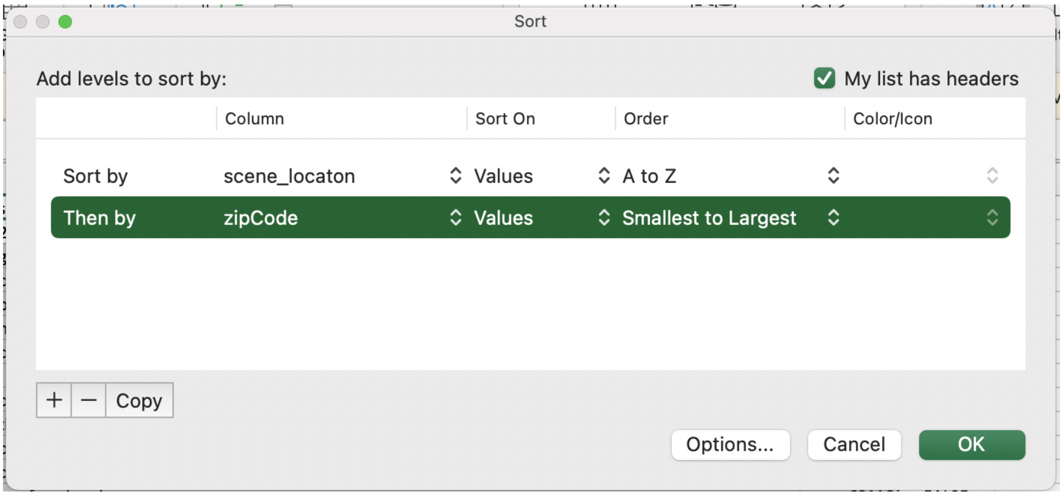


B1 *fx* zipCode

	A	B	C	D
1	scene_locaton	zipCode		
2	20th and folsom streets	15567		
3	golden gate park			
4	cliff house (1090 point lobos avenue)			
5	bush and sutter streets			
6	hayes street at laguna			
7	coffee dan's (o'farrell street at powell)			
8				
9	coit tower	46297		
10	the barbary coast	82998		
11	city hall	10136		
12	cliff house (1090 point lobos avenue)	81665		
13	san francisco bay	69116		
14	fairmont hotel (950 mason street nob hill)	93085		

1.5 – Check the uniqueness of the primary key (scene_location, zip_code)

- 1.5.1 - Go to Data -> Sort. Sort by scene_location by clicking ‘Column’ and choose scene_location. Add another sorting row using the ‘+’ button and choose to also sort by zip_code.



Add levels to sort by: ☑ My list has headers

	Column	Sort On	Order	Color/Icon
Sort by	scene_locaton	Values	A to Z	
Then by	zipCode	Values	Smallest to Largest	

- 1.5.2 – In cell C3, paste FORMULA 2
- 1.5.3 – Copy this formula down to the rest of column C (by dragging

the selection down the same way as you did in 1.2.3).
- 1.5.4 – Filter* the rows so that only TRUE values are shown in column C
- 1.5.5 – Delete rows where the value in column C is TRUE (if any).
- 1.5.5 – Remove the filter*, the rows with a value of FALSE will reappear
- 1.5.6 – Delete column C

1.6 – Create a location_id column:
- 1.6.1 – Type “location_id” in C1.
- 1.6.2 – Type 1 in C2 then 2 in C3
- 1.6.3 – Select C2 and C3, then drag this selection down to the bottom of the sheet.

1.7 – Add the foreign key to the “source” sheet:
- 1.7.1 – Create a new column (column L) after the end_date column
- 1.7.2 – Type “fk_location” in L1.
- 1.7.3 – In L2, paste FORMULA 3 .
- 1.74 – Copy the formula down the rest of column L in the same way as you did in 1.5.3
- 1.7.5 - Convert all the formula-based values in column L into actual values by selecting all of column L, copying, and then doing Paste Special -> Values back into column L. In column L all the base formulas should now be overwritten with the raw values.
- 1.7.6 – Delete the columns (D, E, and I) from “source_data”: scene_location, the normalized scene_location, and zip_code.

Step 2 – Create the “Production” table

2.1
- 2.1.1– Create a “production” worksheet

2.2
- 2.2.1 – In the “source_data” sheet, add an empty column in Column C (next to Column B) for displaying the normalized title text. Normalise the text in the title column (column B, starting on cell C2 and dragging all the way down) using **FORMULA 4,** as you did in 1.2. Also name this column ‘title’. Convert all the formula-based values in column C into actual values by selecting all of column C, copying, and then doing Paste Special -> Values back into column C. In column C all the base formulas should now be overwritten with the raw values.

- 2.2.2 – Add another empty column as Column F next to Column E (production_co). Normalise the text in the production_co column (column E) using **FORMULA 5**, starting on cell F2 and dragging all the way down. Also name this column 'production_co.' Convert all the formula-based values in column F into actual values by selecting all of column F, copying, and then doing Paste Special -> Values back into column F. In column F all the base formulas should now be overwritten with the raw values.

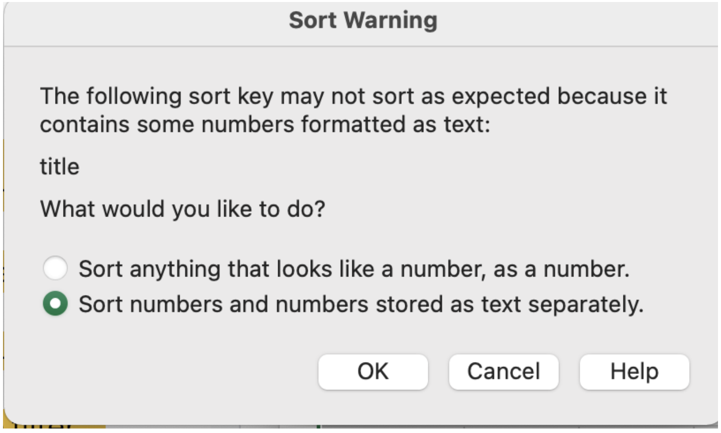
2.3 – Copy the following columns from the "source" sheet and Paste Special -> Values into "production" sheet (in this same order):

- filming_type (column H)
- release_year (column D)
- title (normalized; column C). Make sure it is labeled 'title' in the "production" sheet
- production_co (normalized; column F). Make sure it is labeled "production_co" in the "production" sheet

2.4 – Delete duplicates the same way you did in 1.4.

2.5 – Check the uniqueness of the primary key (title, prod_company)

- 2.5.1 – Sort using the following two columns: title, prod_company.



Sort Warning

The following sort key may not sort as expected because it contains some numbers formatted as text:

title

What would you like to do?

○ Sort anything that looks like a number, as a number.
● Sort numbers and numbers stored as text separately.

Note: you may get a sort warning. Click 'OK' on the default option that comes up.

- 2.5.2 – In cell E3 of "production", paste FORMULA 6
- 2.5.3 – Copy the formula down the rest of column E in the same way as you did in 1.6.3
- 2.5.4 – Look for rows in the file where column E is set to TRUE (using CTRL-F to search)
- 2.5.5 – Delete all rows where column E is true (if there are any). Then remove the filter.:
- 2.5.6 **[step missing from video]** – Delete column E

2.6 *[missing from video]* – Create a production_id column in column E in the same way as you did in 1.6

2.7 *[missing from video]* – Add the foreign key “fk_production” to the sheet “source_data” by creating a new column L after the “production_co” column and copying **FORMULA 7** in the same way as you did in 1.7. Column L should now be fully populated with integer values. Overwrite the formula-driven values in column L by doing Copy and Paste Special -> Values. Column L should have all raw values no formulas. Then, delete the columns from “source_data” that were copied into the “production” sheet (B, C, D, E, F, H)

Step 3 – Create the “Production Company” table

3.1 *[missing from video]* – Create a “production_co” worksheet

3.2 – (no normalisation of text here)

3.3 – In the “production_co” sheet, copy the production_co column from the “production” sheet and rename it to “name” following the relational model.

3.4 – Delete duplicates in the “production_co” table in the same way as you did in 1.4

3.5 – Sort values by name

3.6 *[missing]* – Create a “production_co_id” column in the same way as you did in 1.6

3.7 *[missing]* – Add the foreign key “fk_production_co” to the “production” sheet by creating a new column F after the “production_id” column and copying **FORMULA 8** in the same way as you did in 1.7.

3.8 Overwrite the formula-driven values in column F of “production” sheet by doing Copy and Paste Special -> Values over its own cells. Column F should now have all raw values and no formulas.

3.9 Then, delete the columns (only column D) from “production” that were copied into the “production_co” sheet

Step 4 – Create the “Duration” table

4.1 *[missing from video]* – Create a “duration” worksheet

4.2 – In the “source” sheet, create in Column G called ‘composite_key’. Copy ***Formula 9a*** *into G2 and drag the formula all the way down the column. You should see 4 columns (director, start_date, end_date, and fk_production) concatenated into a single composite string value.* Overwrite the formula-driven values in column G by doing Copy and Paste Special -> Values over its own cells. Column G should now have all raw values and no formulas.

4.3 [missing from video] – In the “duration” sheet, copy the following columns (in the same order they are presented on the “source_data” sheet):

- director

- start_date
- end_date
- fk_production

4.4 In the “duration” sheet, create a column E called “composite_key”. Copy ***Formula 9b*** *into E2 and drag the formula all the way down the column. You should see 4 columns (director, start_date, end_date, and fk_production) concatenated into a single composite string value.* Overwrite the formula-driven values in column F by doing Copy and Paste Special -> Values over its own cells. Column F should now have all raw values and no formulas.

4.5*[missing from video]*– Delete duplicates in the “duration” table in the same way as you did in 1.4

- 4.5.1 [missing] – Check the uniqueness constraint of the primary key by sorting by composite_key, then, and copying FORMULA 10 into column F in the same way as you did in 2.5.1 – 2.5.4. All values should be False. Delete Column F.

4.6. [missing from video] – On the “duration” sheet in column F, next to the composite_key, create a duration_id column in the same way as you did in 1.6

4.7. *[missing from video]* – Add the foreign key “fk_duration” to the sheet “source_data” by creating a new column H after the “composite_key” column and copying **FORMULA 11** in the same way as you did in 1.7. Then, copy the resulting values in this column and Paste Special -> Values to overwrite and fix the actual values to the cells. Delete the columns (B, C, D, F, and G) from “source_data”.

Step 5 – Create the “Director” table

5.1 *[missing from video]* – Create a “director” worksheet

5.2 – (no normalisation here)

5.3 *[missing from video]* – In the “director” sheet, copy the "director" column that exists in the "duration" sheet, then rename it to "name" following the relational model.

5.4 *[missing from video]*– Delete duplicates in the "director" table in the same way as you did in 1.4

5.5 – (no primary key uniqueness check here). Sort the name column alphabetically.

5.6 *[missing from video]* – Create a director_id column in the same way as you did in 1.6

5.7 *[missing from video]* – Add the foreign key "fk_director" to the "duration" sheet by creating a new column G after the "director" column and copying **FORMULA 12** in the same way as you did in 1.7. Then, copy the resulting values in this column (G) and Paste Special -> Values to overwrite and fix the actual values to the cells. Then, delete the column A and E from "duration".

Final step – Create the "Duration_Location_assoc" table

6.1 *[missing from video]* – Rename the "source_data" table to "duration_location_assoc"

6.2 [*missing from video*] – Delete the 'Scene ID' column

6.3 [*missing from video*] – Delete duplicates (there should be none)

At the end, **export***each worksheet to a csv file (one csv file per table)

* these operations differ depending on whether you're using Excel or LibreOffice Calc. The differences are detailed in the "Guidance notes"