Cleaning Data in Spreadsheets

**Introduction**

Now that you have some experience cleaning dirty data, get your hands even "dirtier" with complex cleaning. Before starting any analysis project, it's important to confirm that the data is clean. If not, then it must be cleaned.

Consider this real-world situation:

A San Francisco marketing agency wants to contact local boba tea shops about the potential of creating a coupon book for them. The agency plans to visit the top-rated shops within a 10-mile radius of the center of their target area. To help with planning, the agency asks a data analyst to review external Yelp data related to ratings and locations of boba tea shops**. The analyst makes a spreadsheet from an online source. Unfortunately, the data is not in the greatest shape.**

Your job is to figure out what is making this dataset dirty and clean it up. Here is the dataset: [San Francisco Boba Tea Shop Location Info](https://drive.google.com/u/0/uc?id=1RLsNIOS1rpJycE6J8d_Ruxw7uSryDFom&export=download). Click on the link to the dataset file to download it to your computer. The data includes the following information:

|  |  |
| --- | --- |
| **Field** | **Description** |
| id | a unique identifier for each boba shop |
| name | name of boba shop |
| rating | Yelp rating (0 to 5 stars) |
| address | street address |
| city | city |
| lat-long | latitude and longitude |

The original data from Yelp API was found [here](https://www.kaggle.com/vnxiclaire/bobabayarea).

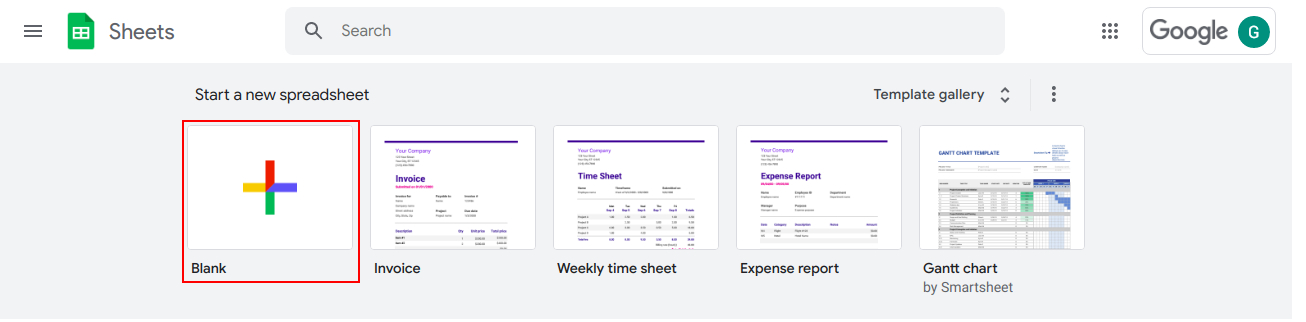
**What you'll do**

In this lab, you'll explore several spreadsheet functions that will help you better understand the data that you work with. You are free to use Google Sheets, Microsoft Excel, Libra Office 3, or another spreadsheet application of your choice. Keep in mind that different applications may have slightly different steps.

By the end of this lab, you'll be able to:

* Identify if and why a dataset is dirty
* Remove duplicate data
* Use the COUNTIF and TRIM functions to clean data
* Use the SPLIT and CONCATENATE functions to clean data
* Use LEFT, RIGHT, MID, and/or LEN to retrieve parts of a cell within a dataset

**You will have 60 minutes to complete this lab.**



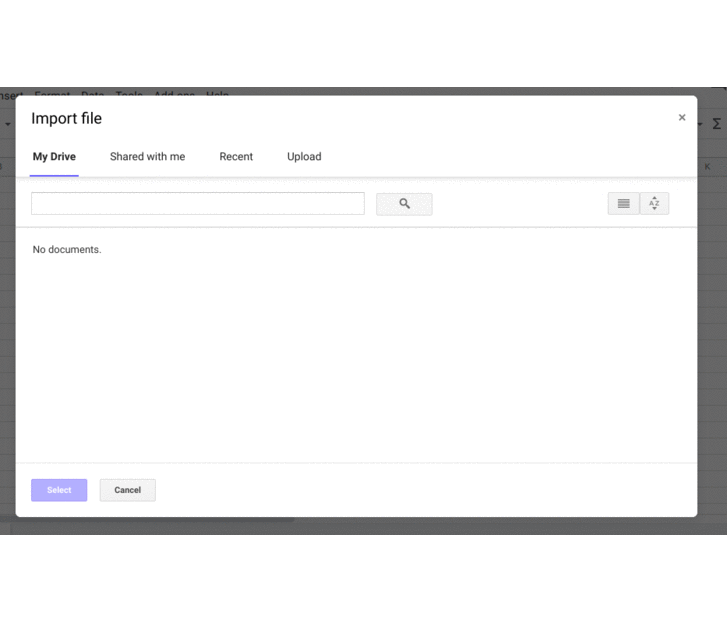
**Identify the dirty elements in the data**

First, import the data into a spreadsheet. Go to the top menu bar, and click **File → Import**. Select the **Upload** tab. You will be prompted to upload the file.

Click on the option **Select a file from your device**.

Select the file that you downloaded earlier to be uploaded ([San Francisco Boba Tea Shop Location Info](https://drive.google.com/u/0/uc?id=1RLsNIOS1rpJycE6J8d_Ruxw7uSryDFom&export=download)).

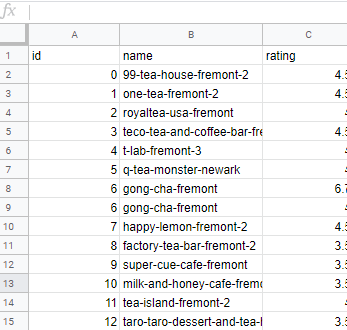
In the next popup window, continue with the default selections and click **Import data** button.



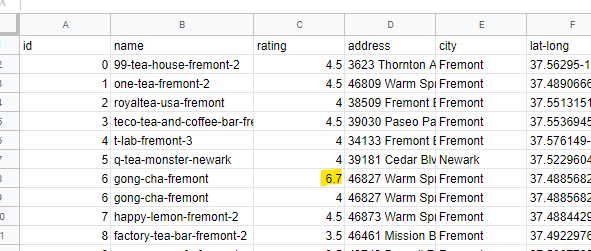
If you didn't rename the file, it will be called *San Francisco Boba Tea Shop Location Info.csv*.

During the first part of this process, you will not be using any functions; you will just identify what type of cleaning should be done. You will work with five Columns: name (column B), rating (column C), and location (Columns D-F). The goal is to create a list of the best boba shops in decreasing order of rating.

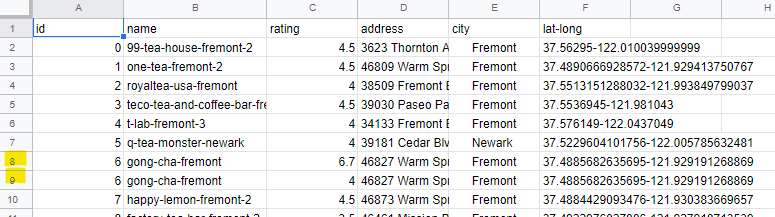
Now, consider what problems you see within the data. First, many of the names in column B are confusing because they contain hyphens. Also, the names of establishments should be capitalized. **As a data analyst, your job is to present data that is readable, accurate, and visually appealing.**

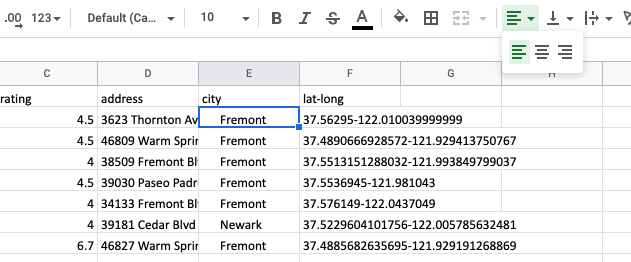


Second, ratings should fall within 0 and 5. However, at least one rating falls outside of that range (row 8):

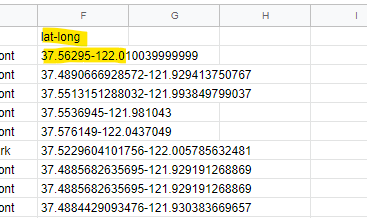


Third, there is at least one duplicate line, as seen in rows 8 and 9:



Also, there seems to be a spacing issue with the city column. Notice that the column is not formatted as **Center align**. Therefore, there are about five extra spaces that have been added to the left of the city name. The data is not technically wrong in this format, but it is not as visually appealing as it can be. 

Finally, in order for someone to be able to use this data, the latitude and longitude should be in separate columns.

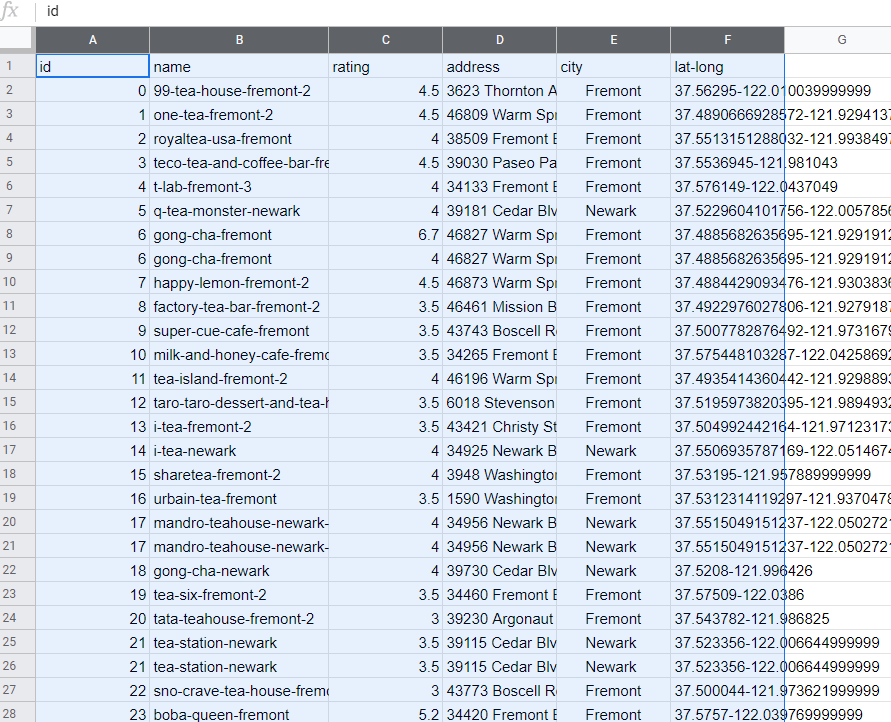


Your goal is to fix these errors to create a clean dataset for the marketing firm.

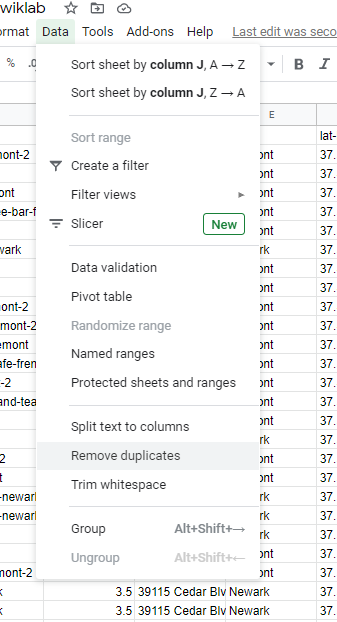
**Removing duplicates**

The first step is to eliminate any duplicates. To do this, first remove duplicate rows. For best practice, this should be done even if duplicates are not easily seen, as in this file.

Highlight Columns A through F.

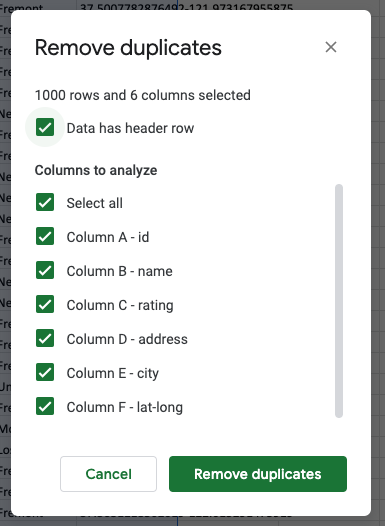


Next, in the top menu bar, select **Data → Remove duplicates.**

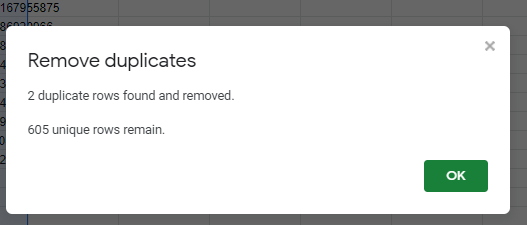


Then, select **Data has header row** and make sure all other columns are selected under **Columns to analyze.**

Once everything has been selected, click **Remove duplicates.**

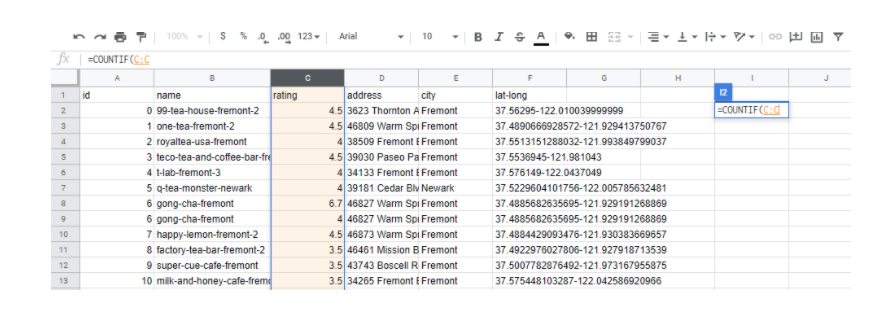


If done correctly, two duplicate rows will be deleted, and 605 remain.



**Use the COUNTIF and TRIM functions to clean data**

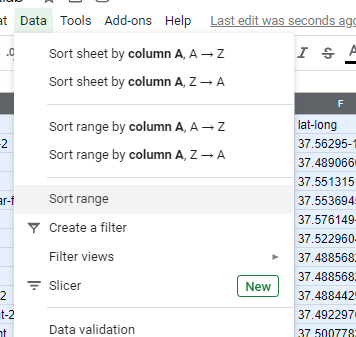
Next, clean up any data that does not make sense. Yelp ratings should be less than 5 and greater than 0, so now you will determine how many entries are inaccurate. Use the COUNTIF function to perform this task. In cell I2, type =COUNTIF(C:C,">5"). The first entry C:C refers to the range where you are counting the data. In this case, you are referring to Column C, which contains the Yelp ratings. Instead of typing C:C, you can also select the entire column.



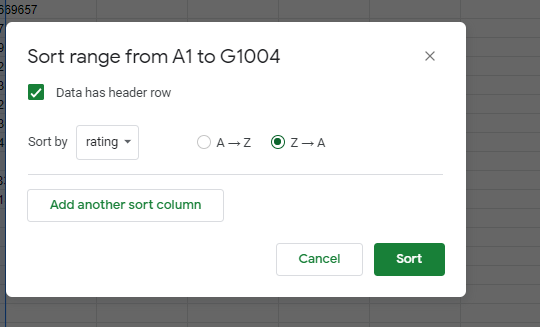
The second entry, >5, tells the function to count numbers greater than 5. Press ENTER or RETURN. You will see that the function has returned a value of 9. Therefore, you have 9 rows that have a value greater than 5.

As a data analyst, it's your job to decide what to do with these incorrect values. One effective approach would be to look them up on Yelp to find their actual rating. But for the purposes of this lab, just replace them with the number 5.

An easy way to replace the incorrect values is to sort the dataset numerically from largest to smallest rating. Highlight Columns A through F, and then from the top menu bar, select **Data → Sort range**.



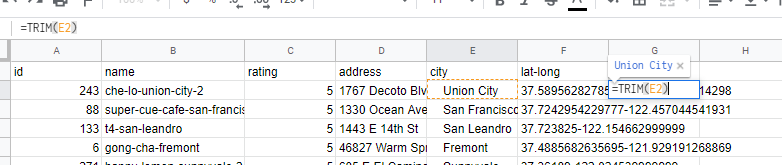
Then, select the **Data has header row** and sort by rating **Z →A.**



After clicking Sort, you should see the 9 rows that have incorrect data. Replace those rating values with 5. After replacing all the 9 ratings with 5, you should see that the new value in cell I2 is 0. This confirms that we no longer have values above 5. Since we don’t need this information anymore, you can now delete the formula from cell I2.

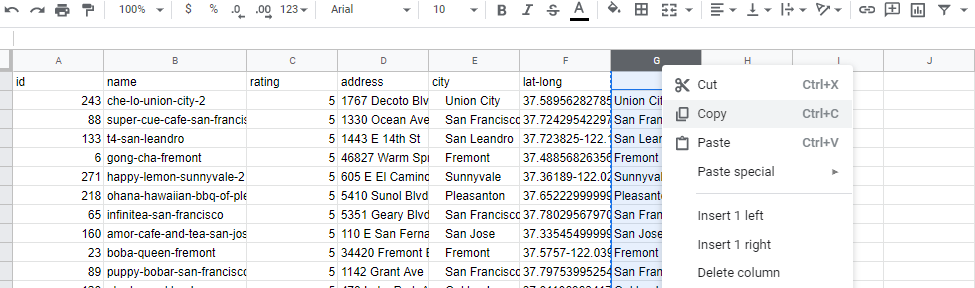
Next, clean up the city names. There is some extra spacing in front of each city name in Column E. Use the TRIM to remove the additional space in those entries.

In cell G2, type =TRIM(E2):

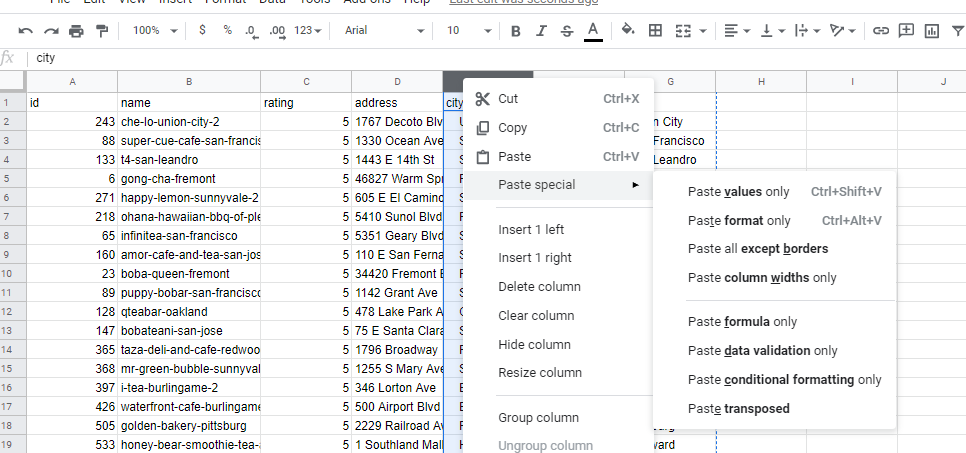


Pressing enter will correct the spacing. Double-click on the blue square in the bottom right corner of the cell to copy the formula down the entire column. Next, you will replace Column E with these entries.

Select the G column, right-click, and select **Copy.**



Then select Column E, right-click, and select **Paste special → Paste values only.**



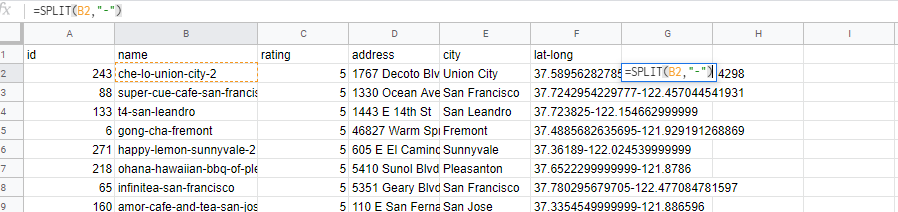
Rename the column as city, then select column G, right-click, and select Delete column. Your first rows should look like this:



**Use the SPLIT and CONCATENATE functions to clean data**

Next, clean up the names of the boba shops by removing the hyphens and capitalizing all of the words. Use the SPLIT function to split each word in the cell into its column and remove the hyphens.

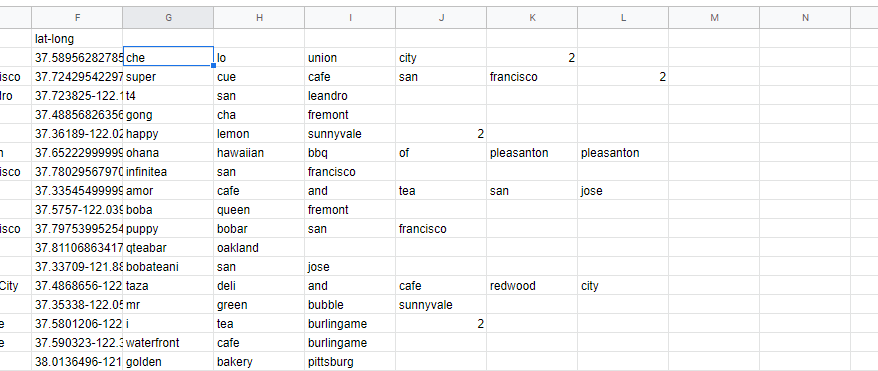
In cell G2 type =SPLIT(B2,"-"). The first entry refers to the cell being split, B2. The second entry refers to the fact that you are dividing the cell based on the hyphen.



After pressing ENTER or RETURN, the result shows each fragment of the cell surrounding a hyphen in a different cell:



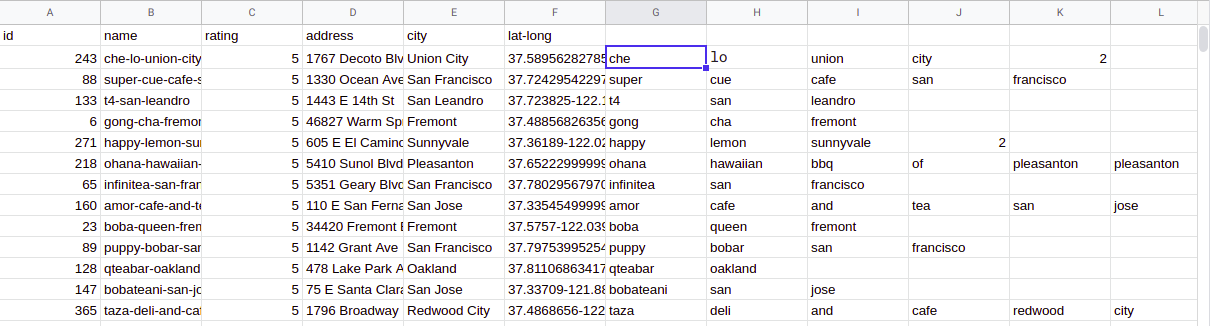
Double-clicking on the blue square in the bottom right corner will correct every entry in Column B based on the hyphens.

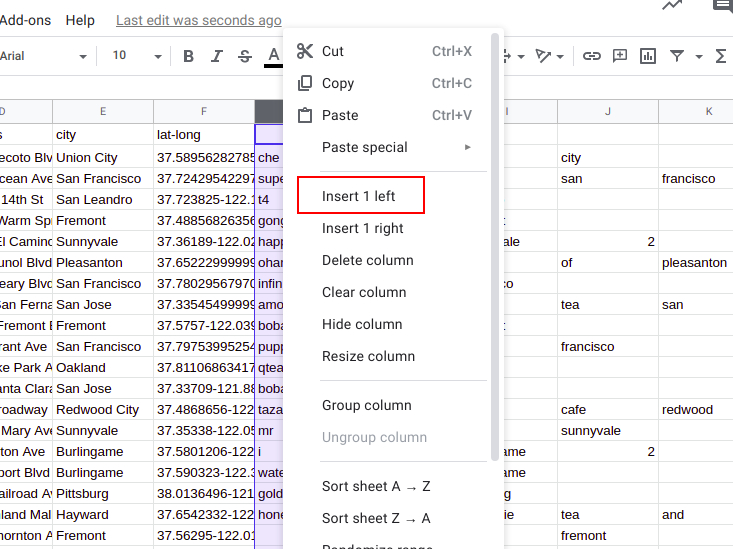


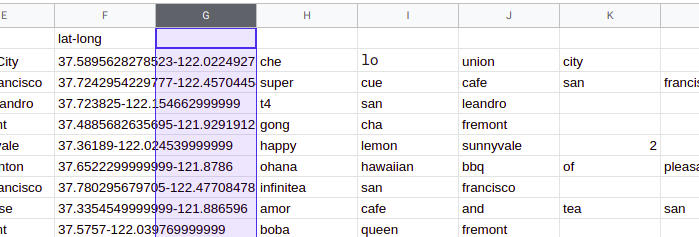
Next, capitalize each part of the name, and join the cells back together. Use the PROPER function to capitalize words and the CONCATENATE function to rejoin the names.

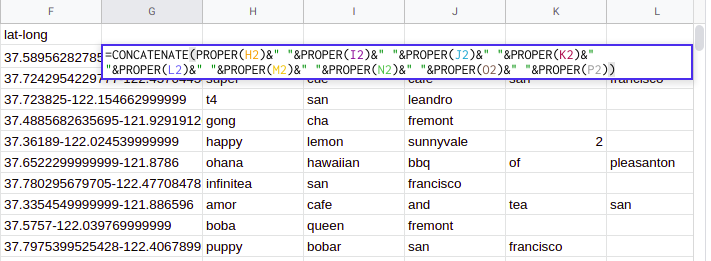
The function =PROPER(G2) returns the value of cell G2 with the first letter capitalized and =CONCATENATE(G2&H2) joins the cells in G2 and H2.

Now, insert a new column to the left of G column. The new column added is now named as G.

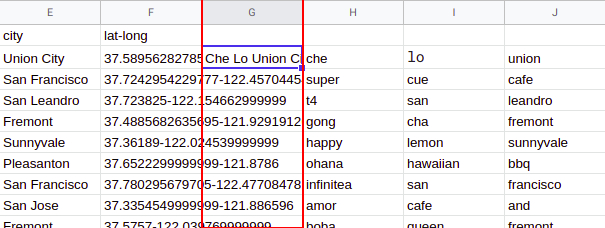


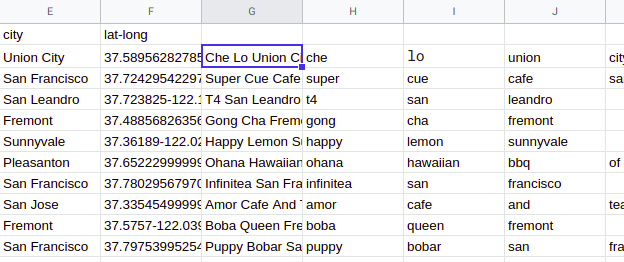




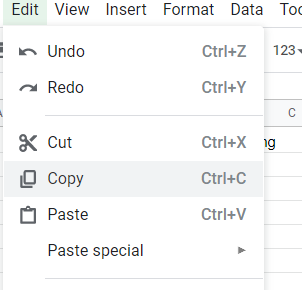
The longest entries from the SPLIT function are 9 cells long and have entries in Column O. Look for the longest chain because, in order to create a formula to concatenate, it's essential to include all portions of a boba shop's name. Type the function in cell G2: =CONCATENATE(PROPER(H2)&" "&PROPER(I2)&" "&PROPER(J2)&" "&PROPER(K2)&" "&PROPER(L2)&" "&PROPER(M2)&" "&PROPER(N2)&" "&PROPER(O2)&" "&PROPER(P2)). 

Then, press ENTER or RETURN to create a clear and proper business name.

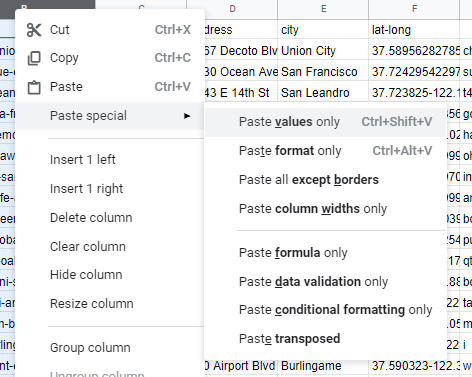




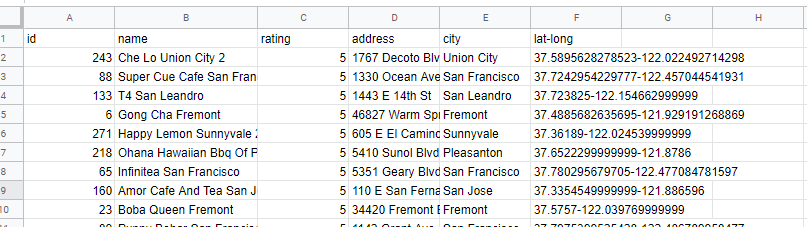
Copy the function down to row 605 by clicking the blue square and dragging it down to row 605. Then, copy and paste the entire column back into Column B. To do this, highlight column G, then, from the top menu bar select **Edit → Copy.**



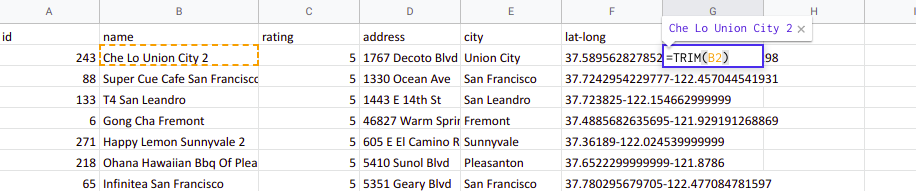
After highlighting Column B, from the top menu bar, select **Edit → Paste values only.**



Next, rename the column as name, delete Columns G and H. The first few rows should look like this:



Finally, trim Column B. In most of the cells, there is now added space after the name of the boba shop. Therefore, in cell G2 type =TRIM(B2).



Pressing enter will correct the spacing.

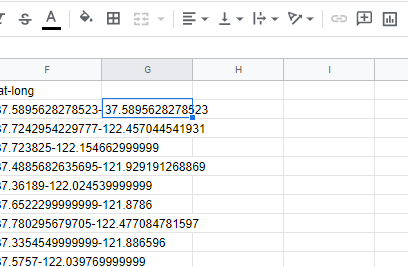
Now, copy that function down to row 605. Then, copy and paste the entire column back into Column B. To do this, highlight Column G, then, from the top menu bar, select **Edit → Copy.** Then select Column B, right-click, and select **Paste special → Paste values only.** Then, rename this column as name. The appearance of the spreadsheet will not change after this step, but the names in cell B will be shorter. This makes it look cleaner and clearer.

**Use LEFT, RIGHT, MID and/or LEN to retrieve parts of a cell within a dataset**

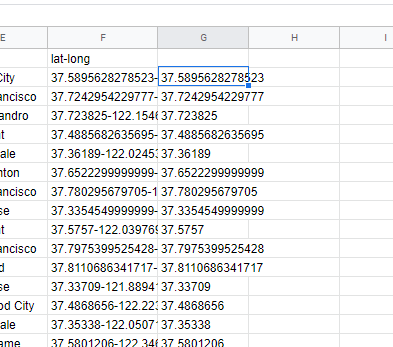
To complete the cleanup, split the latitude and longitude (Column F) into two cells, and make them shorter. To do this, use the LEFT, RIGHT, and LEN functions.

To enter only the latitude into a cell, use the LEFT function to insert all values to the left of the hyphen in cell G2. In cell G2, type =LEFT(F2,FIND("-",F2)-1). The FIND function determines the position of the hyphen in the string, and the LEFT function returns all the values to the left of it. Inserting the "-1" ensures that the hyphen is not returned because the second input in the function is equal to the number of characters desired to be extracted.

After pressing ENTER or RETURN, the following should be in cell G2:

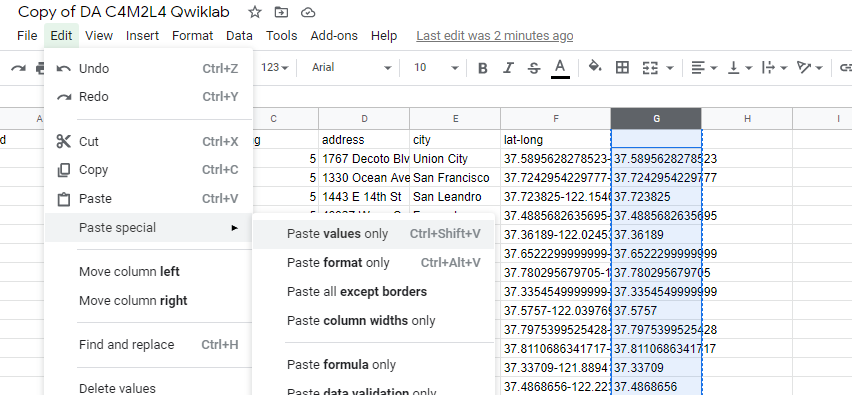


Right-click on the bottom right corner of cell G2 to copy the formula down and populate the entire row in the same manner.

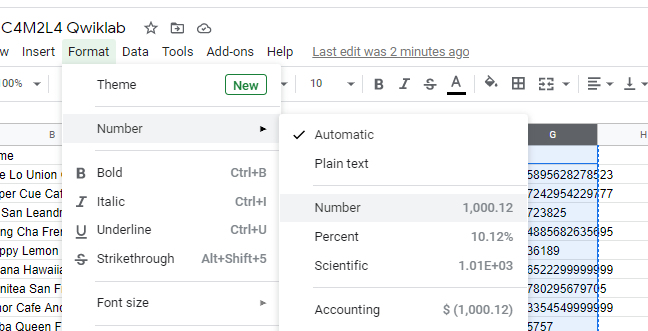


As a data analyst, your job is to make sure data is easily readable. Therefore, you should be consistent with the number of decimal places shown in each cell.

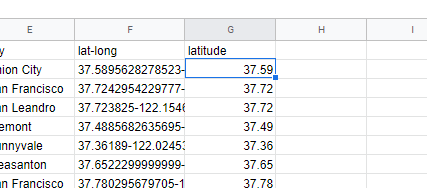
Highlight all of Column G. From the top menu bar, select **Edit → Copy.** Then, keeping Column G highlighted, select **Edit → Paste special → Paste values only.**



Keeping the entire column highlighted, format the data as numbers. From the top menu bar select **Format → Number → Number.**

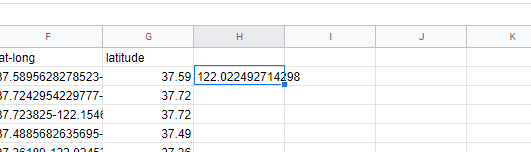


The result is a latitude value showing two decimal places. Place a new header in cell G2 as latitude.



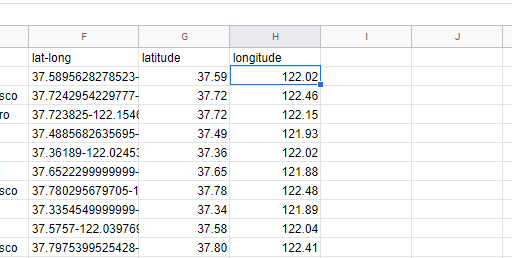
Next, complete a similar process with longitude, but use the RIGHT function. Using the RIGHT function in cell H2 will return all values to the right of the hyphen. In cell H2, type =RIGHT(F2,LEN(F2)-FIND("-",F2)). The LEN function measures the length of the cell, and the FIND function subtracts the length of the values to the left of the cell. This makes the number of characters removed equal to those to the right of the hyphen. This entire function returns all values to the right of the hyphen.

After pressing ENTER or RETURN, the following should be in cell H2:

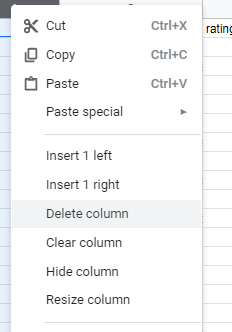


Now, complete the same steps for longitude. Right-click the bottom right corner of the cell to copy the formula down the entire column. Then, copy and paste as values. Finally, change the format of the entire cell to numbers, and rename this column as longitude.

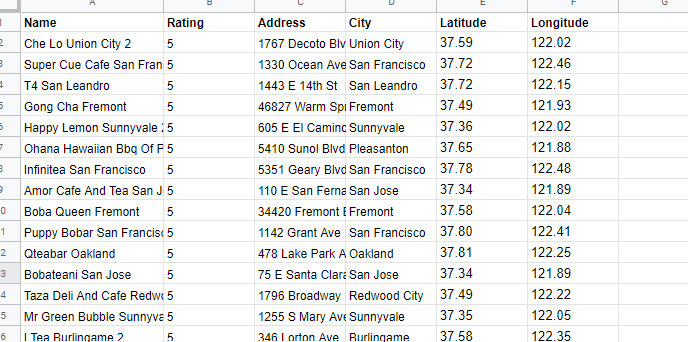
The result looks like this:



Because the data is now split up into two columns, Column F is not needed. Delete the entire column to clean up the spreadsheet. Check to see if the data is in a format that you would be comfortable giving to the marketing company. Perhaps you want to make sure all columns have the same left alignment and there is capitalization and boldface in your titles. You could also remove any columns with data that is not needed. For instance, the id field in Column A is not necessary for this analysis project. Delete the column by right clicking on any value in the column and then select **Delete column**.



Here is the cleaned dataset:



This dataset is in great shape to pass on to the marketing company!