Data Cleaning and Visualization of Boba Tea Shops in San Francisco

231030063 Tenzing Pema Thungon

Introduction

This project aims to clean and prepare a dataset containing information about boba tea shops in the San Francisco area. The dataset includes various columns such as shop ID, name, rating, address, city, latitude, and longitude. The goal is to identify and address dirty elements within the dataset, including duplicates and inconsistencies, to ensure accurate data for further analysis and visualization. Once the data is cleaned, a visualization will be created using Tableau to assist in the decision-making process for a potential marketing campaign collaboration.

Objectives

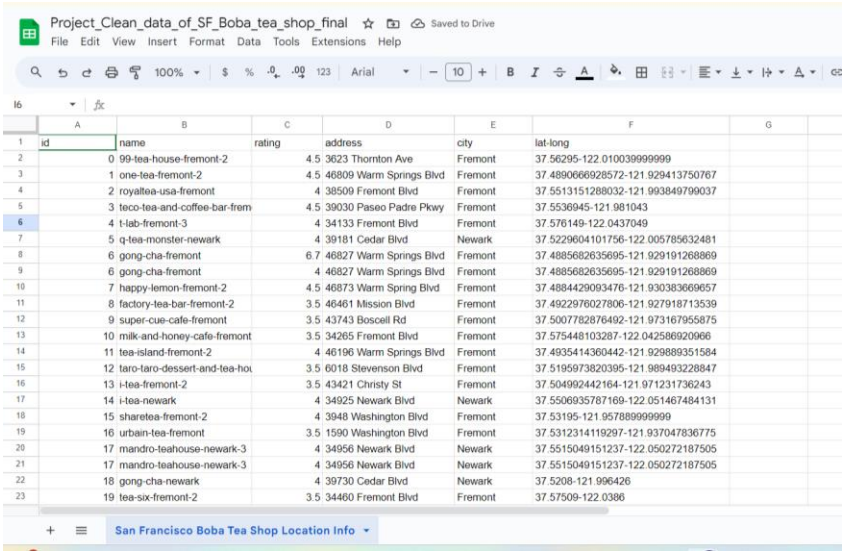
- Identify and correct dirty data elements, including duplicates and inconsistencies.
- Standardize the format of addresses, city names, and ratings.
- Remove duplicate entries to ensure data accuracy.
- Prepare the cleaned data for effective visualization using Tableau.
- Enable informed decision-making for the marketing campaign targeting boba tea shops in San Francisco.

Methodology

Using Google Spreadsheets for Accurate Data Cleaning and Tableau for Visualization.

Exercise 1: Use conditional formatting to highlight blank cells

Step 1: Access the Google spreadsheet

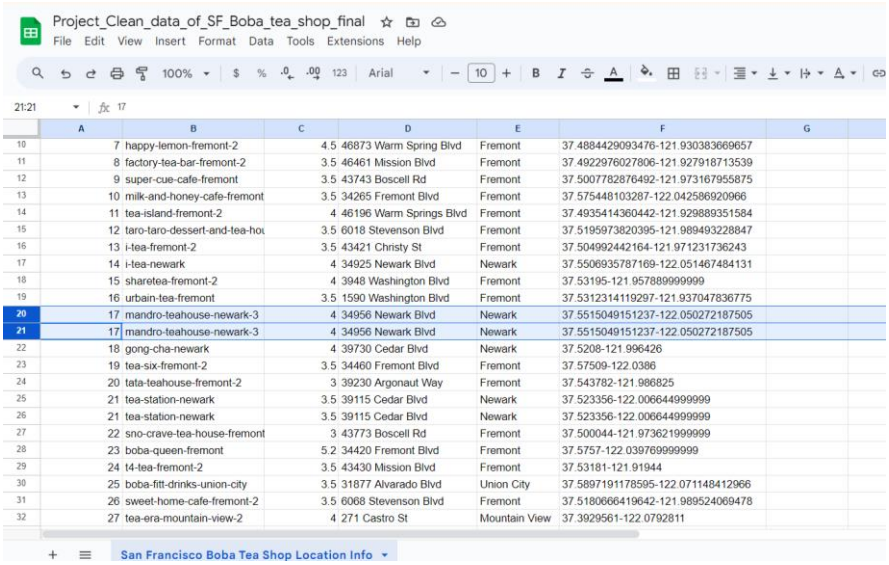


	A	B	C	D	E	F	G
1	id	name	rating	address	city	lat-long	
2	0	99-tea-house-fremont-2	4.5	3623 Thornton Ave	Fremont	37.56295-122.010039999999	
3	1	one-tea-fremont-2	4.5	46809 Warm Springs Blvd	Fremont	37.4890666928572-121.929413750767	
4	2	royaltea-usa-fremont	4	38509 Fremont Blvd	Fremont	37.5513151288032-121.993849799037	
5	3	teco-tea-and-coffee-bar-frem	4.5	39030 Paseo Padre Pkwy	Fremont	37.5536945-121.981043	
6	4	t-lab-fremont-3	4	34133 Fremont Blvd	Fremont	37.576149-122.0437049	
7	5	q-tea-monster-newark	4	39181 Cedar Blvd	Newark	37.5229604101756-122.005785632481	
8	6	gong-cha-fremont	6.7	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869	
9	6	gong-cha-fremont	4	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869	
10	7	happy-lemon-fremont-2	4.5	46873 Warm Spring Blvd	Fremont	37.4884429093476-121.930383669657	
11	8	factory-tea-bar-fremont-2	3.5	46461 Mission Blvd	Fremont	37.4922976027806-121.927918713539	
12	9	super-cue-cafe-fremont	3.5	43743 Boscell Rd	Fremont	37.5007782876492-121.973167955875	
13	10	milk-and-honey-cafe-fremont	3.5	34265 Fremont Blvd	Fremont	37.575448103287-122.042586920966	
14	11	tea-island-fremont-2	4	46196 Warm Springs Blvd	Fremont	37.4935414360442-121.929889351584	
15	12	taro-taro-dessert-and-tea-hou	3.5	6018 Stevenson Blvd	Fremont	37.5195973820395-121.989493228847	
16	13	i-tea-fremont-2	3.5	43421 Christy St	Fremont	37.504992442164-121.971231736243	
17	14	i-tea-newark	4	34925 Newark Blvd	Newark	37.5506935787169-122.051467484131	
18	15	sharetea-fremont-2	4	3948 Washington Blvd	Fremont	37.53195-121.957889999999	
19	16	urbain-tea-fremont	3.5	1590 Washington Blvd	Fremont	37.5312314119297-121.937047836775	
20	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505	
21	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505	
22	18	gong-cha-newark	4	39730 Cedar Blvd	Newark	37.5208-121.996426	
23	19	tea-six-fremont-2	3.5	34460 Fremont Blvd	Fremont	37.57509-122.0386	

Step 2: Identify the dirty elements in our data

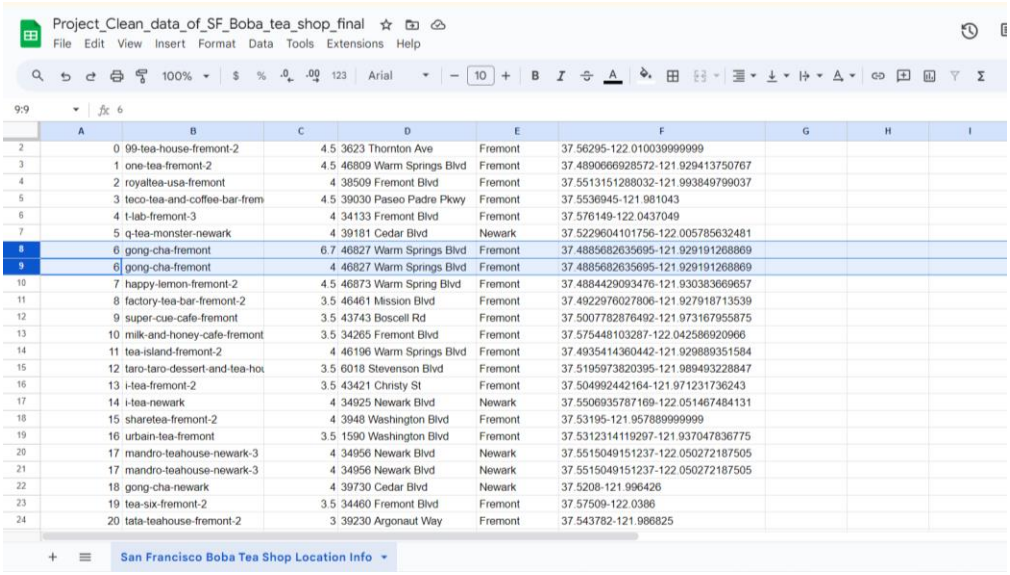
Our job is to present data that is readable, accurate, and visually appealing. Cleaning your data helps you achieve this goal. The first step is to identify the dirty elements in your data.

1. Rename our spreadsheet. Click Untitled Spreadsheet and enter a new name. You can use the name sf_boba_tea_shop_data or a similar name that describes the data our spreadsheet contains.
2. If we want to get a better view of our data, you can make the columns wider by dragging the right boundary of the column heading. This may apply to the name (B), address (D), and lat-long (F) columns.
3. Now, review your data and consider any problems you may need to address. The following are examples of errors that you can quickly identify and fix. This is not a comprehensive list of every potential problem but is a great starting point for data cleaning.



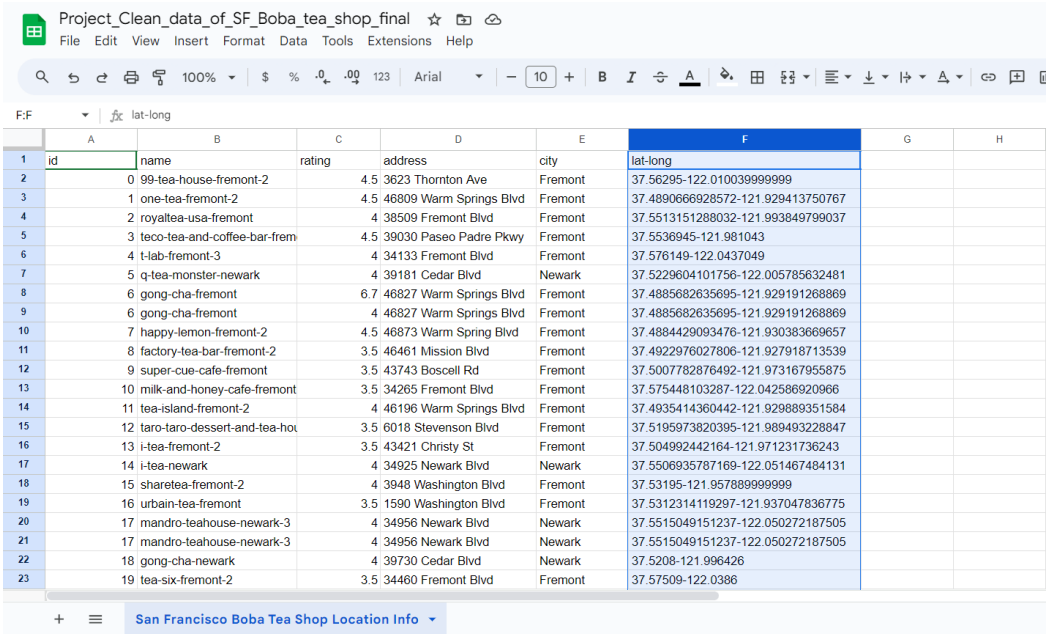
	A	B	C	D	E	F	G
10	7	happy-lemon-fremont-2	4.5	46873 Warm Spring Blvd	Fremont	37.4884429093476-121.930383669657	
11	8	factory-tea-bar-fremont-2	3.5	46461 Mission Blvd	Fremont	37.4922976027806-121.927918713539	
12	9	super-cue-cafe-fremont	3.5	43743 Boscell Rd	Fremont	37.5007782876492-121.973167955875	
13	10	milk-and-honey-cafe-fremont	3.5	34265 Fremont Blvd	Fremont	37.57544810328 7-122.042586920966	
14	11	tea-island-fremont-2	4	46196 Warm Springs Blvd	Fremont	37.4935414360442-121.929889351584	
15	12	taro-taro-dessert-and-tea-hou	3.5	6018 Stevenson Blvd	Fremont	37.5195973820395-121.989493228847	
16	13	i-tea-fremont-2	3.5	43421 Christy St	Fremont	37.504992442164-121.971231736243	
17	14	i-tea-newark	4	34925 Newark Blvd	Newark	37.5506935787169-122.051467484131	
18	15	sharetea-fremont-2	4	3948 Washington Blvd	Fremont	37.53195-121.957889999999	
19	16	urbain-tea-fremont	3.5	1590 Washington Blvd	Fremont	37.5312314119297-121.937047836775	
20	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505	
21	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505	
22	18	gong-cha-newark	4	39730 Cedar Blvd	Newark	37.5208-121.996426	
23	19	tea-six-fremont-2	3.5	34460 Fremont Blvd	Fremont	37.57509-122.0386	
24	20	tata-teahouse-fremont-2	3	39230 Argonaut Way	Fremont	37.543782-121.986825	
25	21	tea-station-newark	3.5	39115 Cedar Blvd	Newark	37.523356-122.006644999999	
26	21	tea-station-newark	3.5	39115 Cedar Blvd	Newark	37.523356-122.006644999999	
27	22	sno-crave-tea-house-fremont	3	43773 Boscell Rd	Fremont	37.500044-121.973621999999	
28	23	boba-queen-fremont	5.2	34420 Fremont Blvd	Fremont	37.5757-122.039769999999	
29	24	t4-tea-fremont-2	3.5	43430 Mission Blvd	Fremont	37.53181-121.91944	
30	25	boba-fitt-drinks-union-city	3.5	31877 Alvarado Blvd	Union City	37.5897191178595-122.071148412966	
31	26	sweet-home-cafe-fremont-2	3.5	6068 Stevenson Blvd	Fremont	37.5180666419642-121.989524069478	
32	27	tea-era-mountain-view-2	4	271 Castro St	Mountain View	37.3929561-122.0792811	

- First, we see there is at least one duplicate line (rows 20 and 21) in our dataset.



	A	B	C	D	E	F	G	H	I
2	0	99-tea-house-fremont-2	4.5	3623 Thornton Ave	Fremont	37.56295-122.010039999999			
3	1	one-tea-fremont-2	4.5	46809 Warm Springs Blvd	Fremont	37.4890666928572-121.929413750767			
4	2	royaltea-usa-fremont	4	38509 Fremont Blvd	Fremont	37.5513151288032-121.993849799037			
5	3	teco-tea-and-coffee-bar-frem	4.5	39030 Paseo Padre Pkwy	Fremont	37.5536945-121.981043			
6	4	t-lab-fremont-3	4	34133 Fremont Blvd	Fremont	37.576149-122.0437049			
7	5	q-tea-monster-newark	4	39181 Cedar Blvd	Newark	37.5229604101756-122.005785632481			
8	6	gong-cha-fremont	6.7	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869			
9	6	gong-cha-fremont	4	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869			
10	7	happy-lemon-fremont-2	4.5	46873 Warm Spring Blvd	Fremont	37.4884429093476-121.930383669657			
11	8	factory-tea-bar-fremont-2	3.5	46461 Mission Blvd	Fremont	37.4922976027806-121.927918713539			
12	9	super-cue-cafe-fremont	3.5	43743 Boscell Rd	Fremont	37.5007782876492-121.973167955875			
13	10	milk-and-honey-cafe-fremont	3.5	34265 Fremont Blvd	Fremont	37.575448103287-122.042586920966			
14	11	tea-island-fremont-2	4	46196 Warm Springs Blvd	Fremont	37.4935414360442-121.929889351584			
15	12	taro-taro-dessert-and-tea-hou	3.5	6018 Stevenson Blvd	Fremont	37.5195973820395-121.989493228847			
16	13	i-tea-fremont-2	3.5	43421 Christy St	Fremont	37.504992442164-121.971231736243			
17	14	i-tea-newark	4	34925 Newark Blvd	Newark	37.5506935787169-122.051467484131			
18	15	sharetea-fremont-2	4	3948 Washington Blvd	Fremont	37.53195-121.957889999999			
19	16	urbain-tea-fremont	3.5	1590 Washington Blvd	Fremont	37.5312314119297-121.937047836775			
20	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505			
21	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505			
22	18	gong-cha-newark	4	39730 Cedar Blvd	Newark	37.5208-121.996426			
23	19	tea-six-fremont-2	3.5	34460 Fremont Blvd	Fremont	37.57509-122.0386			
24	20	tata-teahouse-fremont-2	3	39230 Argonaut Way	Fremont	37.543782-121.986825			

- Second, all Yelp ratings should fall between 0 and 5. However, at least one rating (in cell C8) falls outside of that range.



	A	B	C	D	E	F	G	H
1	id	name	rating	address	city	lat-long		
2	0	99-tea-house-fremont-2	4.5	3623 Thornton Ave	Fremont	37.56295-122.010039999999		
3	1	one-tea-fremont-2	4.5	46809 Warm Springs Blvd	Fremont	37.4890666928572-121.929413750767		
4	2	royaltea-usa-fremont	4	38509 Fremont Blvd	Fremont	37.5513151288032-121.993849799037		
5	3	teco-tea-and-coffee-bar-frem	4.5	39030 Paseo Padre Pkwy	Fremont	37.5536945-121.981043		
6	4	t-lab-fremont-3	4	34133 Fremont Blvd	Fremont	37.576149-122.0437049		
7	5	q-tea-monster-newark	4	39181 Cedar Blvd	Newark	37.5229604101756-122.005785632481		
8	6	gong-cha-fremont	6.7	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869		
9	6	gong-cha-fremont	4	46827 Warm Springs Blvd	Fremont	37.4885682635695-121.929191268869		
10	7	happy-lemon-fremont-2	4.5	46873 Warm Spring Blvd	Fremont	37.4884429093476-121.930383669657		
11	8	factory-tea-bar-fremont-2	3.5	46461 Mission Blvd	Fremont	37.4922976027806-121.927918713539		
12	9	super-cue-cafe-fremont	3.5	43743 Boscell Rd	Fremont	37.5007782876492-121.973167955875		
13	10	milk-and-honey-cafe-fremont	3.5	34265 Fremont Blvd	Fremont	37.575448103287-122.042586920966		
14	11	tea-island-fremont-2	4	46196 Warm Springs Blvd	Fremont	37.4935414360442-121.929889351584		
15	12	taro-taro-dessert-and-tea-hou	3.5	6018 Stevenson Blvd	Fremont	37.5195973820395-121.989493228847		
16	13	i-tea-fremont-2	3.5	43421 Christy St	Fremont	37.504992442164-121.971231736243		
17	14	i-tea-newark	4	34925 Newark Blvd	Newark	37.5506935787169-122.051467484131		
18	15	sharetea-fremont-2	4	3948 Washington Blvd	Fremont	37.53195-121.957889999999		
19	16	urbain-tea-fremont	3.5	1590 Washington Blvd	Fremont	37.5312314119297-121.937047836775		
20	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505		
21	17	mandro-teahouse-newark-3	4	34956 Newark Blvd	Newark	37.5515049151237-122.050272187505		
22	18	gong-cha-newark	4	39730 Cedar Blvd	Newark	37.5208-121.996426		
23	19	tea-six-fremont-2	3.5	34460 Fremont Blvd	Fremont	37.57509-122.0386		

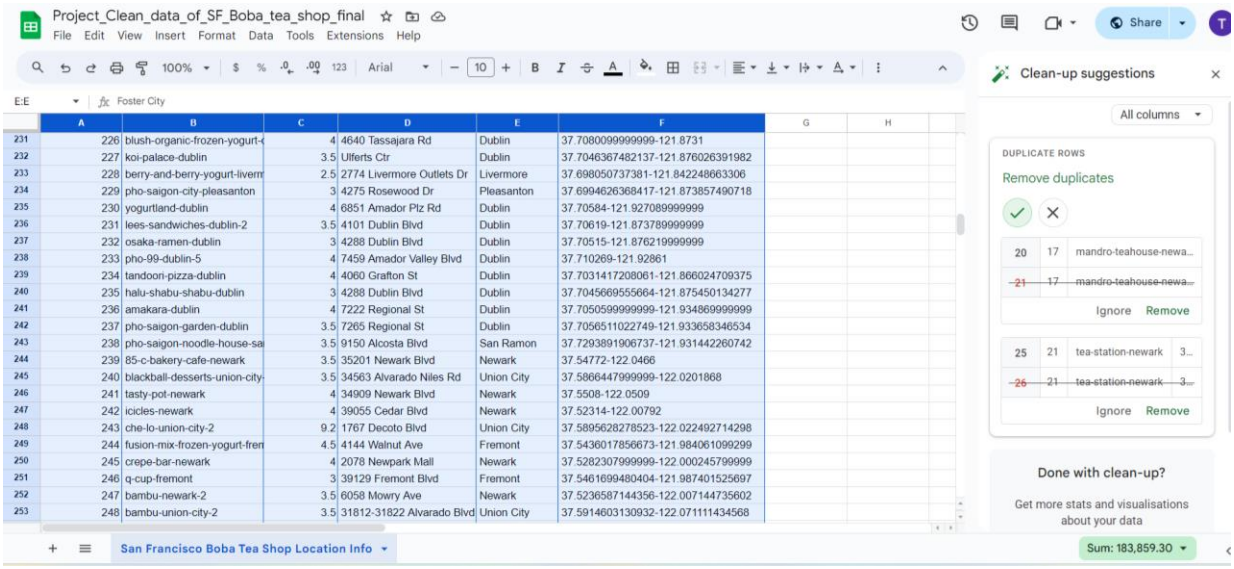
- Finally, the data for latitude and longitude is contained in a single column (F). In order for someone to be able to use this data for analysis, the two values should be in separate columns.

Step 3: Remove duplicates

The first step is to eliminate any duplicate entries from your dataset. As a best practice, duplicates should be removed even if they are not readily apparent.

1. To start, select columns A through F.

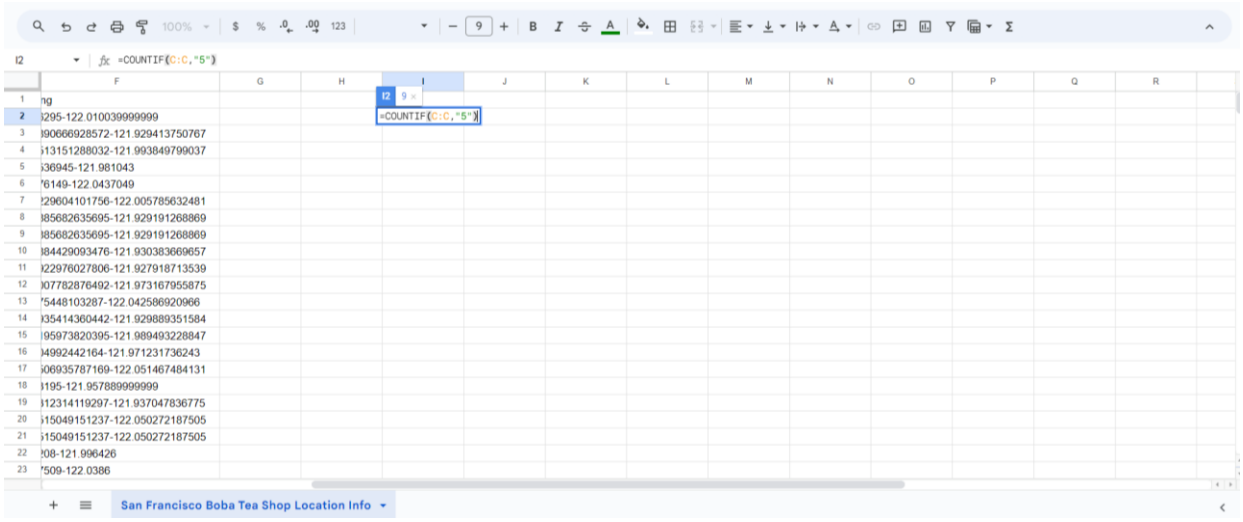
2. Then, in the menu bar, choose Data, then Data Cleanup, and select Remove duplicates.
3. In the pop-up window, click Data has header row. You want to remove duplicate boba shop id's and boba shop names. In the Columns to analyze section, make sure the relevant columns (id, name) are selected.

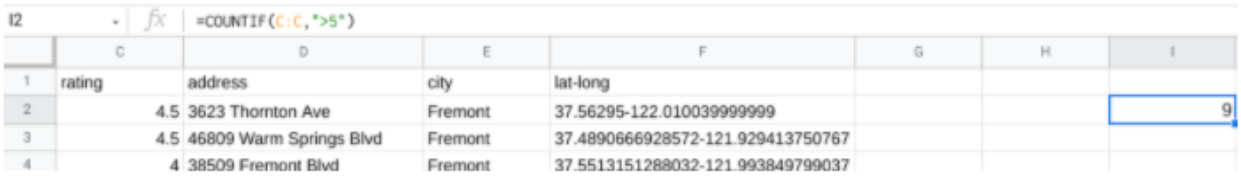


Step 4: Correct the ratings data

Next, clean up any data that does not make sense. Yelp ratings should be less than 5 and greater than 0. Now, you will determine how many entries are inaccurate and correct them. You can use the COUNTIF function to perform this task.

1. The COUNTIF function quickly counts how many items in a range of cells meet a given criterion. In cell I2, enter =COUNTIF(C:C,">5"). The first entry (C:C) refers to the range where you are counting the data. In this case, the range is the entire rating column (C), which contains the Yelp ratings. The second entry refers to the criterion (>5), and tells the function to count all the values greater than 5.
2. Press Enter.





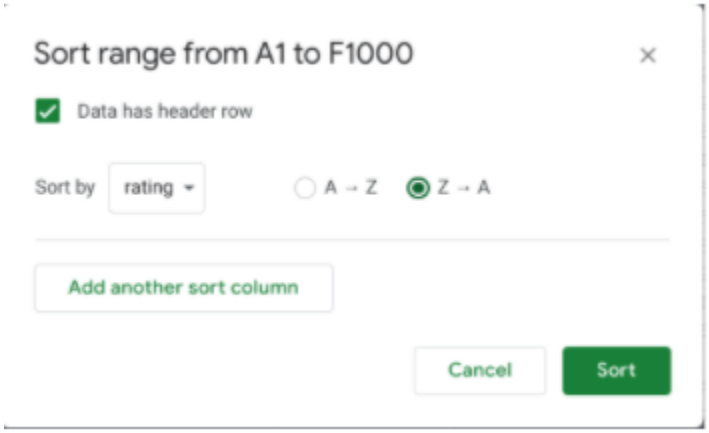
We notice that the function returns a value of 9. This tells us that our dataset contains 9 entries that have a rating greater than 5.

It's our job to decide what to do with incorrect values or to ask the dataset owner for advice if you're unsure. In this case, one effective approach would be to search on Yelp for the actual ratings. For this activity, you can just replace the incorrect ratings with the number 5. An efficient way to replace the ratings is to sort the data numerically from largest to smallest rating.

3. Select columns A through F.

4. Then, from the menu bar, choose Data, then Sort range, and select Advanced range sorting options.

5. In the pop-up window, check the box next to Data has header row. Sort by rating from Z →A. This way, the highest ratings will be listed first.

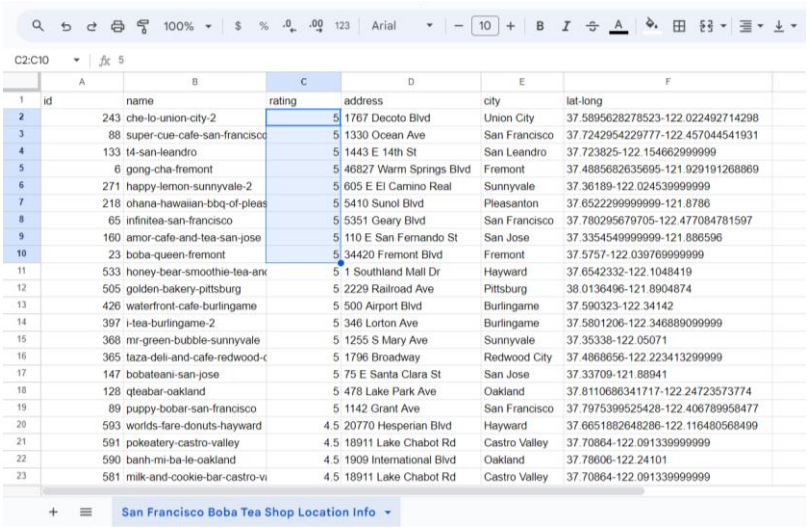


6. Click Sort. Check out your spreadsheet. At the start of the rating column, you should now find the 9 rows that have incorrect values (rating > 5).

	A	B	C
1	id	name	rating
2	243	che-lo-union-city-2	9.2
3	88	super-cue-cafe-san-francisco-2	8.9
4	133	t4-san-leandro	7.4
5	6	gong-cha-fremont	6.7
6	271	happy-lemon-sunnyvale-2	6.2
7	218	ohana-hawaiian-bbq-of-pleasanton-pleasanton	5.7
8	65	infinitea-san-francisco	5.6
9	160	amor-cafe-and-tea-san-jose	5.4
10	23	boba-queen-fremont	5.2

7. Next, select the range of cells C2:C10. Press delete to delete the values that are greater than 5.

8. Replace all the values with the number 5. In cell C2, enter 5. Then, drag the fill handle down to cell C10 to fill the remaining cells with 5.



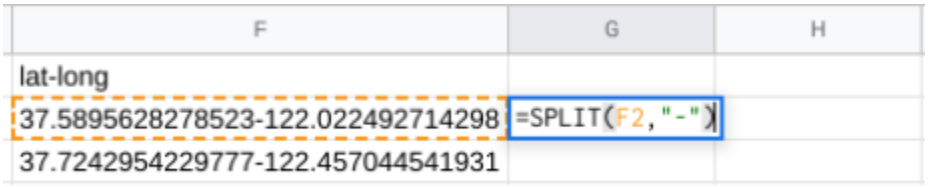
9. After replacing the incorrect ratings with the number 5, you may notice that the new value in cell I2 is 0. The output of the COUNTIF function now reflects the changes in your dataset. This confirms that the rating column no longer contains any values greater than 5.

10. Finally, delete the formula from cell I2 since we don't need this information anymore.

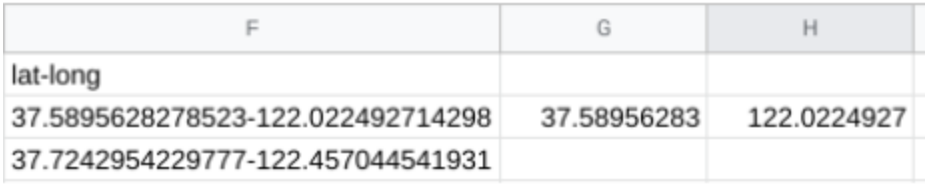
Step 5: Clean up the latitude and longitude data

Next, clean up the latitude and longitude data by placing each value in a separate column. We can use the SPLIT function to accomplish this task.

1. The SPLIT function divides text around a specified character or string and puts each fragment of text into a separate cell in the row. The SPLIT function will split the single lat-long column into two separate columns, one for latitude and the other for longitude. In cell G2, enter =SPLIT(F2,"-"). The first entry (F2) refers to the cell where the text is located. The second entry ("-") refers to the fact that you are dividing the text based on the minus sign.



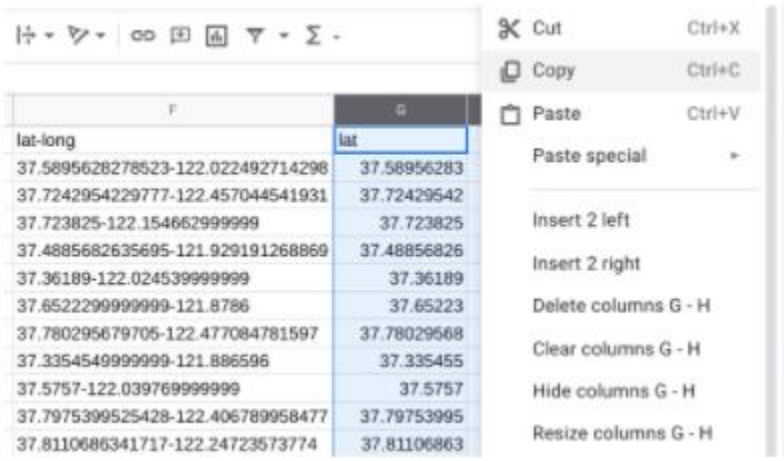
2. Press **Enter**. The result shows each fragment of text in a different cell.



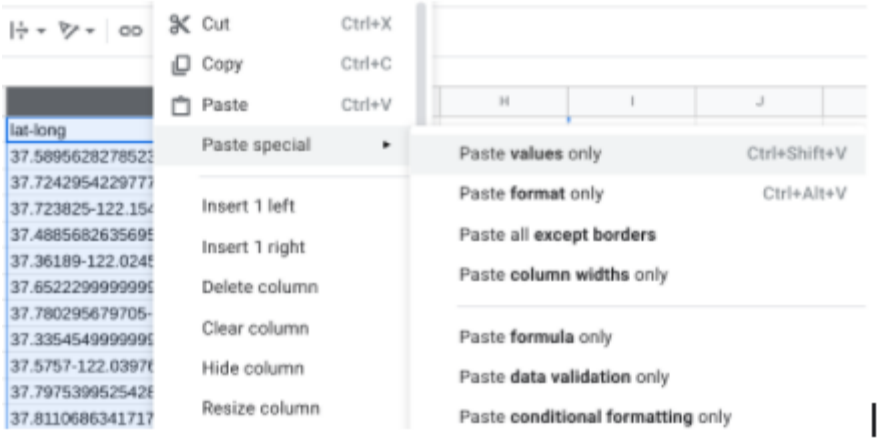
3. Select cell G2 again. In cell G2, double-click on the fill handle to split all the remaining **lat-long** entries.

4. Now add column headers to the two new columns (G and H). In cell G1, enter **lat**. In cell H1, enter **long**.

5. Next, replace the original **lat-long** data in column F with the new split entries in columns G and H. Select columns G and H, right-click, and choose **Copy.**



6. Then, select Column F, right-click, and choose **Paste special** and **Paste values only**.



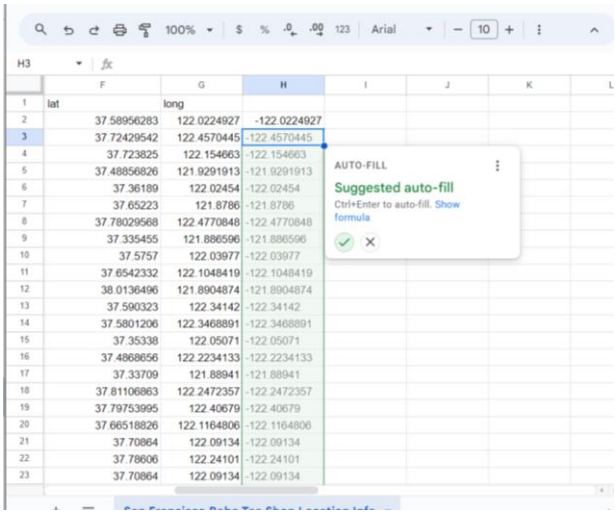
7. Now the new **lat** column is column F, and the new **long** column is column G. Adjust the width of the **lat** column (F) to fit the data by dragging the right boundary of the column heading.

F	G
lat	long
37.58956283	122.0224927
37.72429542	122.4570445
37.723825	122.154663
37.48856826	121.9291913
37.36189	122.02454
37.65223	121.8786
37.78029568	122.4770848
37.335455	121.886596
37.5757	122.03977

8. Next, select column H, right-click, and choose **Delete column.**

9. Finally, the longitude values should be negative so that they are accurate coordinates for mapping. To make the values in the **long** column negative, multiply them by **-1.** In cell H2, enter **=G2*-1**. The asterisk is the operator for multiplication. Press **Enter.**

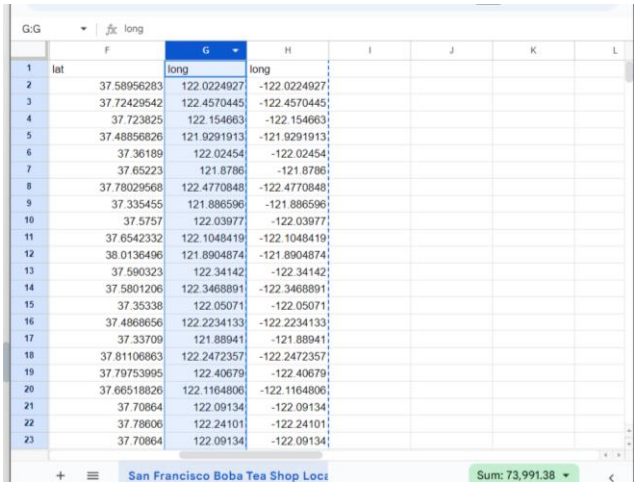
10. Still in cell H2, double-click on the fill handle to fill in the rest of the values.



11. Next, add a column header. In cell H1, enter: **long**.

12. Now, replace the longitude data in column G with the new data in column H. Select column H, right-click, and choose **Copy.**

13. Select Column G, right-click, and choose **Paste special** and **Paste values only**.

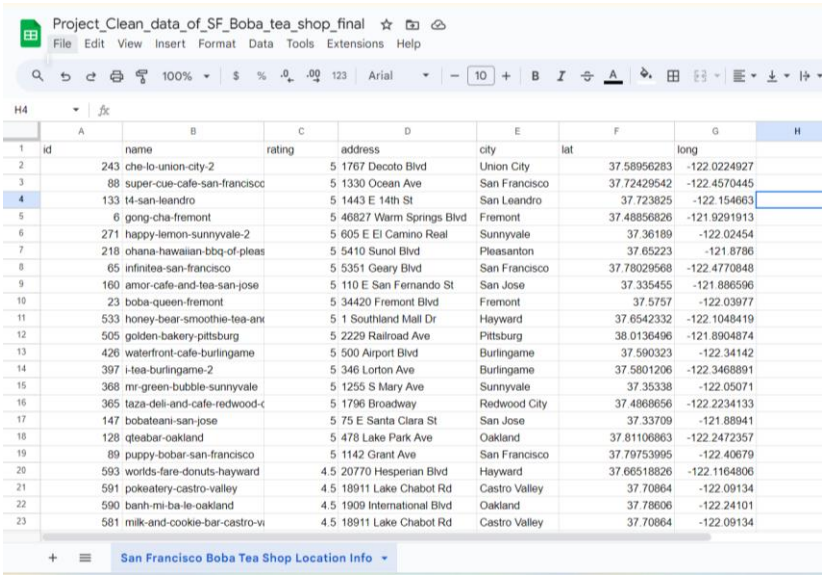


14. Then, select column H, right-click, and choose **Delete column.**

Columns F and G look like this:

F	G
lat	long
37.58956283	-122.0224927
37.72429542	-122.4570445
37.723825	-122.154663
37.48856826	-121.9291913
37.36189	-122.02454
37.65223	-121.8786
37.78029568	-122.4770848
37.335455	-121.886596
37.5757	-122.03977

Now our data is cleaner, clearer, and easier to use.



id	name	rating	address	city	lat	long
243	che-lo-union-city-2	5	1767 Decoto Blvd	Union City	37.58956283	-122.0224927
88	super-cue-cafe-san-francisco	5	1330 Ocean Ave	San Francisco	37.72429542	-122.4570445
133	t4-san-leandro	5	1443 E 14th St	San Leandro	37.723825	-122.154663
6	gong-cha-fremont	5	46827 Warm Springs Blvd	Fremont	37.48856826	-121.9291913
271	happy-lemon-sunnyvale-2	5	605 E El Camino Real	Sunnyvale	37.36189	-122.02454
218	ohana-hawaiian-bbq-of-pleas	5	5410 Sunol Blvd	Pleasanton	37.65223	-121.8786
65	infinitea-san-francisco	5	5351 Geary Blvd	San Francisco	37.78029568	-122.4770848
160	amor-cafe-and-tea-san-jose	5	110 E San Fernando St	San Jose	37.335455	-121.886596
23	boba-queen-fremont	5	34420 Fremont Blvd	Fremont	37.5757	-122.03977
533	honey-bear-smoothie-tea-an	5	1 Southland Mall Dr	Hayward	37.6542332	-122.1048419
505	golden-bakery-pittsburg	5	2229 Railroad Ave	Pittsburg	38.0136496	-121.8904874
426	waterfront-cafe-burlingame	5	500 Airport Blvd	Burlingame	37.590323	-122.34142
397	i-tea-burlingame-2	5	346 Lorton Ave	Burlingame	37.5801206	-122.3468891
368	mr-green-bubble-sunnyvale	5	1255 S Mary Ave	Sunnyvale	37.35338	-122.05071
365	taza-deli-and-cafe-redwood-c	5	1796 Broadway	Redwood City	37.4868656	-122.2234133
147	bobateani-san-jose	5	75 E Santa Clara St	San Jose	37.33709	-121.88941
128	qteabar-oakland	5	478 Lake Park Ave	Oakland	37.81106863	-122.2472357
89	puppy-bobar-san-francisco	5	1142 Grant Ave	San Francisco	37.79753995	-122.40679
593	worlds-fare-donuts-hayward	4.5	20770 Hesperian Blvd	Hayward	37.66518826	-122.1164806
591	pokeatery-castro-valley	4.5	18911 Lake Chabot Rd	Castro Valley	37.70864	-122.09134
590	banh-mi-ba-le-oakland	4.5	1909 International Blvd	Oakland	37.78606	-122.24101
581	milk-and-cookie-bar-castro-v	4.5	18911 Lake Chabot Rd	Castro Valley	37.70864	-122.09134

Result:

1. Data Deduplication:
    - Successfully identified and removed duplicate entries to ensure each boba tea shop had a unique representation, enhancing data integrity.
2. Data Validation and Correction:
    - Detected and corrected 9 entries with Yelp ratings exceeding the valid range (greater than 5), ensuring data accuracy and reliability.
3. Geospatial Data Enhancement:

- Separated latitude and longitude into distinct columns using the SPLIT function, optimizing the dataset for precise geospatial analysis and mapping.

4. Data Standardization:
    - Standardized addresses, city names, and ratings, ensuring uniformity across the dataset, which is crucial for consistent data processing and visualization.
5. Tableau Visual Analytics:
    - Developed comprehensive visualizations in Tableau:
        - Geospatial Distribution Map: Mapped all boba tea shops across multiple Bay Area cities, providing a clear visual representation of shop locations.
        - 5-Star Rating Analysis: Visualized the distribution of top-rated (5-star) boba tea shops, identifying key geographic clusters of high-quality shops.
        - Distance and Rating Correlation: Analyzed the relationship between shop ratings and their distance from a central point, offering insights into the spatial distribution of top-rated shops.
6. Data-Driven Decision Support:
    - The cleaned and visualized dataset was prepared to support data-driven decision-making, enabling targeted marketing strategies focused on the most promising boba tea shops in strategic Bay Area locations.

Conclusion:

The project successfully executed a comprehensive data cleaning and preparation process, validating and correcting critical data elements, and enhancing geospatial information for accurate analysis. The subsequent visual analytics in Tableau provided valuable insights into the distribution, ratings, and geographic spread of boba tea shops across the San Francisco Bay Area. These insights are instrumental in informing strategic decisions for a targeted marketing campaign, demonstrating the importance of meticulous data preparation and advanced visualization techniques in driving effective, data-driven strategies.