1/1 point



1.

By now, you've been introduced to some useful techniques for cleaning spreadsheet data, such as sorting and filtering. In this activity, you'll continue to

to help clean data.

Imagine you are a data analyst working for a marketing agency based in San Francisco. The marketing agency wants to contact local boba tea shops to inquire about a potential collaboration for a new marketing campaign. The agency plans to visit the top-rated shops within a 10-mile radius of the center of their target

area. To assist with planning, the agency asks your team to review external data related to ratings and locations of boba tea shops in San Francisco. One of your teammates has created a spreadsheet from an online source. However, the data is not in the greatest shape. Your assignment is to identify the dirty elements in the dataset and clean them up. By the time you complete this activity, you will be able to identify dirty elements in a dataset, remove duplicate data, and use the COUNTIF and SPLIT functions

develop your data-cleaning skills by using spreadsheet functions.

What you will need The dataset includes the following column headers:

Column Header Description id a unique identifier for each boba shop

name of boba shop name rating Yelp rating (0 to 5 stars) street address address city city latitude and longitude lat-long

To get started, access the spreadsheet that contains the data. Click the link and make a copy of the spreadsheet ...

Or, if you don't have a Google account, you may download the dataset directly from the attachment below:

Identify the dirty elements in your data

San Francisco Boba Tea Shop Location Info

the name (B), address (D), and lat-long (F) columns.

describes the data your spreadsheet contains.

identify the dirty elements in your data.

This is not a comprehensive list of every potential problem, but is a great starting point for data cleaning. First, there is at least one duplicate line (rows 20 and 21) in your dataset.

- 17 mandro-teahouse-newark-3 4 34956 Newark Blvd Newark 17 mandro-teahouse-newark-3 4 34956 Newark Blvd Newark
- Second, all Yelp ratings should fall between 0 and 5. However, at least one rating (in cell C8) falls outside of that range.
- 0 99-tea-house-fremont-2 4.5 3623 Thornton Ave 37.56295-122.010039999999 Fremont 1 one-tea-fremont-2 4.5 46809 Warm Springs Blvd Fremont

37.56295-122.010039999999 37.4890666928572-121.929413750767 37.5513151288032-121.993849799037 37.5536945-121.981043 37.576149-122.0437049 37.5229604101756-122.005785632481
37.5513151288032-121.993849799037 37.5536945-121.981043 37.576149-122.0437049 37.5229604101756-122.005785632481
37.5536945-121.981043 37.576149-122.0437049 37.5229604101756-122.005785632481
37.576149-122.0437049 37.5229604101756-122.005785632481
37.5229604101756-122.005785632481
07 40050000050505 104 000101000000
37.4885682635695-121.929191268869
37.4885682635695-121.929191268869
37.4884429093476-121.930383669657

Your goal is to fix these errors and help create a clean dataset for analysis. You can address each issue in turn. Remove duplicates

To start, select columns A through F.

Remove duplicates

Data has header row

Column A - id

Column B - name

1000 rows and 6 columns selected

2. Then, in the menu bar, choose Data, then Data Cleanup, and select Remove duplicates.

section, make sure the relevant columns (id, name) are selected.

Clean your data

Columns to analyze Select all

Column D - address Column E - city Column F - lat-long Remove duplicates Cancel Once everything has been selected, click Remove duplicates.

- 3. Select columns A through F. Then, from the menu bar, choose Data, then Sort range, and select Advanced range sorting options.

6 gong-cha-fremont 271 happy-lemon-sunnyvale-2

65 infinitea-san-francisco 160 amor-cafe-and-tea-san-jose

Sort range from A1 to F1000

Data has header row

Sort by rating -

- 23 boba-queen-fremont
- 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 you don't need this information anymore.
- lat-long 37.5895628278523-122.022492714298 122.0224927 37.58956283 37.7242954229777-122.457044541931

Ctrl+X **%** Cut □ Copy Ctrl+C Paste Ctrl+V

long 37.58956283 122.0224927 37.72429542 122.4570445 37.723825 122.154663 37.48856826 121.9291913 122.02454 37.36189 37.65223 121.8786 37.78029568 122.4770848 37.335455 121.886596

Select Column G, right-click, and choose Paste special and Paste values only. 14. Then, select column H, right-click, and choose Delete column.

The TRIM function

Correct The SPLIT function divides text around a specified character or string, and puts each fragment of text into a separate cell in the row. Spreadsheet functions are useful tools for data cleaning, and knowing how to use functions effectively is a key part of every data analyst's skill set.

Cleaning Data with Spreadsheet Functions: Spreadsheet functions can significantly improve data cleaning efficiency and effectiveness:

2. In the text box below, write 2-3 sentences (40-60 words) in response to each of the following questions:

How can spreadsheet functions help you clean data more efficiently and effectively?

By leveraging the power of readability, accuracy, and visual appeal, you can transform your data presentations into insightful tools that drive action and impact. And by mastering spreadsheet functions, you can clean data more efficiently and effectively, paving the way for accurate and trustworthy analysis

Remember: * Choosing the right functions depends on your specific data and cleaning goals. * Combining functions can create powerful workflows for complex data cleansing tasks. * Continuously learning new functions and exploring advanced techniques will further enhance your data cleaning skills.

* **Credibility and trust:** Accurate data builds trust and strengthens your conclusions. Misleading or incorrect information can undermine your analysis

* **Engagement and action:** Visually appealing presentations capture attention, sparking interest and driving engagement with your findings. This can lead

* **Improving data accuracy:** Functions like TRIM and LEFT/RIGHT remove unwanted spaces or extract specific characters, ensuring data consistency and

* **Identifying anomalies and outliers:** Functions like AVERAGE and STANDARD DEVIATION help identify data points that deviate from the norm, prompting

* **Sorting and filtering:** Functions like SORT and FILTER organize and focus your data to quickly identify specific patterns or trends, streamlining your

rating address city lat-long 2 royaltea-usa-fremont 4 38509 Fremont Blvd Fremont 5 3 teco-tea-and-coffee-bar-fremont 37.5536945-121.981043 4.5 39030 Paseo Padre Pkwy Fremont 6 4 t-lab-fremont-3 Fremont 37.576149-122.0437049 4 34133 Fremont Blvd 5 q-tea-monster-newark 4 39181 Cedar Blvd Newark 6 gong-cha-fremont 6.7 46827 Warm Springs Blvd Fremont 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. lat-long Now you know what issues to focus your attention on during the cleaning process.

Column C - rating

5. If done correctly, 3 duplicate rows will be found and removed and 604 rows will remain. Correct the ratings data =COUNTIF(C:C, ">5") address lat-long city 4.5 3623 Thornton Ave Fremont 4.5 46809 Warm Springs Blvd Fremont 4 38509 Fremont Blvd

Add another sort column

218 ohana-hawaiian-bbq-of-pleasanton-pleasanton

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

Next, add a column header. In cell H1, type: long. Columns F and G should look like this:

37.48856826

37.78029568 37.335455

37.36189 37.65223

37.5757

-121.9291913 -122.02454

-121.8786 -122.4770848

-121.886596

-122.03977

Confirmation and reflection

The COUNTIF function

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

The SPLIT function The CONCATENATE function

Why Readable, Accurate, and Visually Appealing Data Matters: For a data analyst, presenting information in a clear, accurate, and aesthetically pleasing way is paramount for several reasons: * **Clarity and understanding:** Readable data ensures your audience (stakeholders, colleagues, etc.) can readily understand the key points and insights

you've extracted.

and credibility.

reducing errors.

* **Automating repetitive tasks:** Functions like COUNTIF, VLOOKUP, and CONCATENATE automate data cleansing tasks like finding duplicates, merging

further investigation and potential data correction.

information, and formatting entries, saving you time and effort.

to actionable decisions based on your insights.

analysis process.

that inspires informed decisions.

✓ Correct Congratulations on completing this hands-on activity! In this activity, you used spreadsheet functions to clean "dirty" data in a spreadsheet. Cleaning data is an important part of the data analysis process. If data analysis is based on bad or "dirty" data, it may be biased, erroneous, and uninformed. Knowing how to effectively use spreadsheet functions to work with data is an essential skill for every data analyst. In upcoming activities, you will continue to learn more about spreadsheet functions and how they can help you analyze your data.

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 122.03977 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, type =G2*-1. The asterisk is the operator for multiplication. Press Enter. Still in cell H2, double-click on the fill handle to fill in the rest of the values. 12. Now, replace the longitude data in column G with the new data in column H. Select column H, right-click, and choose Copy. G long 37.58956283 -122.0224927 37.72429542 -122.4570445 37.723825 -122.154663

Which of the following functions divides text around a specified character or string and puts each fragment of text into a separate cell in the row?

Paste format only Ctrl+Alt+V Insert 1 left 37.723825-122.154 37.4885682635695 Paste all except borders Insert 1 right 37.36189-122.0245 Paste column widths only Delete column 37.6522299999999 37.780295679705-37.335454999999 Paste formula only 37.5757-122.03976 Hide column Paste data validation only 37.7975399525428 Resize column 37.8110686341717 Paste conditional formatting only boundary of the column heading. 37.5757

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, type 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 Clean up the latitude and longitude data Next, clean up the latitude and longitude data by placing each value in a separate column. You 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, type =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. F 37.5895628278523-122.022492714298 =SPLIT(F2, "-") 37.7242954229777-122.457044541931 Press Enter. The result shows each fragment of text in a different cell. Н Select cell G2 again. In cell G2, double-click on the fill handle to split all the remaining lat-long entries. Now add column headers to the two new columns (G and H). In cell G1, type lat. In cell H1, type long. % Cut Ctrl+X 1+ - B - CO I II F - Z -Copy Ctrl+C lat-long Paste special 37.5895628278523-122.022492714298 37.58956283 37.7242954229777-122.457044541931 Insert 2 left 37 723825-122 154662999999 37.723825 37.4885682635695-121.929191268869 Insert 2 right 37.36189-122.024539999999 37.36189 37.6522299999999-121.8786 37.65223 Delete columns G - H 37.780295679705-122.477084781597 37.78029568 Clear columns G - H 37.3354549999999-121.886596 37.335455 37.5757-122.039769999999 37.5757 Hide columns G - H 37.79753995 37.7975399525428-122.406789958477 Resize columns G - H 37.8110686341717-122.24723573774 37.81106863

37.4890666928572-121.929413750767 37.5513151288032-121.993849799037 37.5229604101756-122.005785632481 37.4885682635695-121.929191268869

As a data analyst, your 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 1. Rename your spreadsheet. Click Untitled Spreadsheet and enter a new name. You can use the name sf_boba_tea_shop_data or a similar name that

2. If you want to get a better view of your data, you can make the columns wider by dragging the right boundary of the column heading. This may apply to 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. 37.5515049151237-122.050272187505 37.5515049151237-122.050272187505

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.

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

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. The COUNTIF function quickly counts how many items in a range of cells meet a given criterion. 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, 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. You'll notice that the function returns a value of 9. This tells you that your dataset contains 9 entries that have a rating greater than 5. 37.56295-122.010039999999 37.4890666928572-121.929413750767 37.5513151288032-121.993849799037 As a data analyst, it's your 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.

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). name rating 243 che-lo-union-city-2 9.2 88 super-cue-cafe-san-francisco-2 8.9 133 t4-san-leandro 7.4

> 6.2 5.7

5.6

5.4

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.

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.

Paste special 37.5895628278523 37.7242954229777

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

* **Efficiency and communication:** Clearly presented data simplifies communication, eliminating confusion and facilitating quicker decision-making based on your analysis.

As a data analyst, why is it important for the data you present to be readable, accurate, and visually appealing?

1/1 point