

# Cleaning Data Assignment

## Potential Problems in a Dataset

### 1. Inconsistency

- **Formatting:**
  - Different formats for the same data, such as varying date formats like mm/dd/yyyy vs dd/mm/yyyy, can lead to inconsistencies. To address this issue in OpenRefine, the Edit Cells > Transform feature can be utilized. A GREL expression can be written to standardize the date format.
- **Terminology:**
  - Variations in terms used for the same concept (e.g., "Emily Johnson" vs. "E. Johnson") can create inconsistencies. To address this in OpenRefine, the Text Facet and Clustering feature can be utilized. This functionality allows for the identification and merging of different terminologies into a single standardized term.
- **Casing:**
  - Mixed capitalization for the same value (e.g., "Product" vs. "product"). To address this issue in OpenRefine, the Edit cells > Common Transformations can change all values in a column to a consistent case (e.g., To lowercase or To titlecase).
- **Data Type:**
  - Mixing data types within a column (e.g., some age values as text while others are numeric). To address this issue in OpenRefine, the Edit cells > Common Transformations can change all values in a column to a consistent value (e.g., To number or To date).
- **Character Encoding:**
  - Issues with character representation (e.g., "MarÃa" instead of "María") affect text readability. Character encodings can be reset to address such errors. For instance, Excel allows users to select a character encoding from a list when importing a data file (Data > Get External Data > From Text).
- **Spacing:**
  - Inconsistent spacing within entries (e.g., extra or irregular spaces in entries). One way inconsistency can be revealed is by clicking "edit" in OpenRefine and observing excess spaces. One way to address this issue is to go to Edit Cells > Common Transformations > Collapse consecutive whitespace to standardize the spacing.

## 2. Missing or Incorrect Data

- **Missing Values:**
  - Data points that are not captured can create gaps in the dataset, which may appear as (blank) in text facets in OpenRefine. Missing values might also be entered as "Not Provided" or similar terms, potentially misleading users into thinking the information is complete. These gaps can be found with a text faucet and can be addressed by either locating the missing data or deleting the entry.
- **Incorrect Values:**
  - Typos or errors in data entry (e.g., "Apple" vs. "Appel"). One method for identifying and correcting these issues in OpenRefine, similar to variations in terminology for the same concept, is by using a Text Facet and the Clustering feature.
- **Data Type Mismatch:**
  - Storing all data within a column in an incorrect format (e.g., age column as text instead of a number) can lead to analysis issues. To resolve this in OpenRefine, navigate to Edit Cells > Common Transformations > To Number.
- **Misfielded Values:**
  - Entering data in the wrong field (e.g., "Australia" in a city field) can be resolved in OpenRefine by using facets. Start by creating a text facet for the city field to filter and identify entries containing 'Australia.' Clicking on the 'Australia' value in the facet displays the affected records. To correct the entries, select the rows with the incorrect data and navigate to Edit Cells > Transform. A formula such as `value.replace()` can be applied.
- **Logical Errors:**
  - Values exceeding logical limits (e.g., ages recorded as negative numbers). In OpenRefine, these can be found through a numeric faucet, and these entries can be either updated or removed as needed.
- **Inconsistent Units:**
  - Measurements recorded in different units (e.g., weight in both kilograms and pounds). To address this issue in OpenRefine, transformations can convert all measurements to a consistent unit.

### 3. Structural Issues

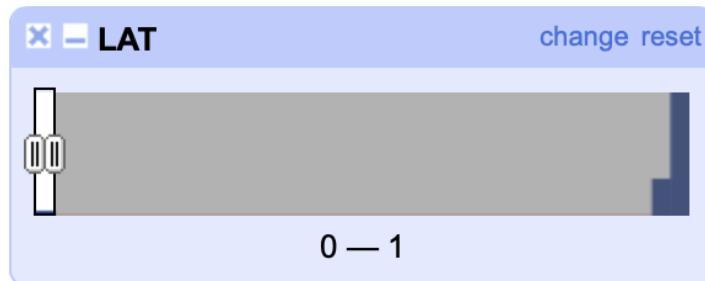
- **Redundant Columns:**
  - Duplicate information across multiple columns (e.g., having both "Full Name" and separate "First Name" and "Last Name" columns). In OpenRefine, the redundant column(s) can be deleted By going to All > Edit columns > Re-order / remove ...
- **Inconsistent Relationships:**
  - Data not adhering to expected relationships (e.g., total amounts not equaling quantity multiplied by unit price), indicating calculations errors. To address this issue in OpenRefine, click on the dropdown for the column (e.g., "Total Amount"), select "Edit column" > "Add column based on this column," then use a GREL expression to calculate the expected total. Next, filter the dataset to show rows where the existing total amount does not match the calculated total.
- **Complex Fields:**
  - Fields containing multiple values (e.g., "red|blue|green") may need to be split for meaningful analysis. To address this issue in OpenRefine, select the column with complex values, then choose Edit column > Split into several columns...

## Specific Problems in the Dataset, “Crime Data from 2020 to Present.”

URL: <https://catalog.data.gov/dataset/crime-data-from-2020-to-present>

**Description:** This dataset reflects incidents of crime in the City of Los Angeles dating back to 2020.

1. Most of the data consists of latitude and longitude coordinates that correspond to the “LOCATION” column of the reported crimes. However, a few entries are missing these coordinates, indicated by a value of "0." The following image below shows how the numeric facet was used to identify the missing values in “LAT” specifically.



**Solution:** The provided “LOCATION” details make it possible to find the missing latitude and longitude coordinates using Google Maps. For instance, the address “19300 BUSINESS CENTER DR” corresponds to a latitude of “34.230892” and a longitude of “-118.553713.”

To correct the data, the cell containing "0" was selected, followed by clicking "Edit" and entering the correct latitude. Afterward, "Apply" was clicked to save the changes. The same process was repeated for the longitude. The image below shows an example of an updated entry based on the previous photo.

Before				After			
LOCATION	CROSS STREET	LAT	LON	LOCATION	CROSS STREET	LAT	LON
19300 BUSINESS CENTER DR		0	0	19300 BUSINESS CENTER DR		34.230892	-118.553713

2. A numeric entry with the value "10" was found in the "LOCATION" field, which should contain text. Additionally, its "Cross Street" was listed as "FREEWAY," which is inconsistent. Entries in the "Premis Desc" column labeled "FREEWAY" typically have "LOCATION" and "Cross Street" fields that align with the provided latitude and longitude. Upon verifying the coordinates on Google Maps, the correct location was identified as 1633 E Cesar E Chavez Ave.

Although it might seem incorrect to list an avenue address for an incident involving a freeway, this follows the pattern observed in other entries. According to the dataset's landing page, "Address fields are only provided to the nearest hundred block to maintain privacy." In this case, 1633 E Cesar E Chavez Ave is situated near two potential freeways: the San Bernardino and Golden State.

**Solution:** "FREEWAY" was removed from the "Cross Street" field by clicking "Edit," deleting the text, and applying the change. Additionally, the data type for "LOCATION" was updated from number to text, and the complete address was entered.

Before

Status Desc | Crm Cd 1 | Crm Cd 2 | Crm Cd 3 | Crm Cd 4

Data type: number

10

Apply | Apply to all identical cells | Cancel  
Enter | Ctrl-Enter | Esc

LOCATION	CROSS STREET	LAT	LON
10	FREEWAY	34.0517	-118.2178

After

Status Desc | Crm Cd 1 | Crm Cd 2 | Crm Cd 3 | Crm Cd 4

Data type: text

1633 E Cesar E Chavez Ave

Apply | Apply to all identical cells | Cancel  
Enter | Ctrl-Enter | Esc

LOCATION	CROSS STREET	LAT	LON
1633 E Cesar E Chavez Ave		34.0517	-118.2178

3. The “LOCATION” entries have inconsistencies in spacing, and one way of confirming this is by clicking “edit,” on random entries and seeing the excess spaces.

**Solution:** Apply "Collapse consecutive whitespace" to standardize the formatting.

Crm Cd 3	Crm Cd 4	LOCATION	Cross Street	LAT	LON
		Facet	A	33.7978	-118.3076
		Text filter			
		Edit cells	Transform...		
		Trim leading and trailing whitespace	Common transforms		
		Collapse consecutive whitespace			
		Unescape HTML entities	Fill down		
		Replace smart quotes with ASCII	Blank down		
		To titlecase	Split multi-valued cells...		
		To uppercase	Join multi-valued cells...		
		To lowercase	Cluster and edit...		
		To number	Replace...		
		To date	AND	34.1016	-118.3387
		To text		34.0523	-118.2722
		To null			
		To empty string			
			WILSHIRE BL	WESTLAKE	34.0574 -118.2747

Before

After

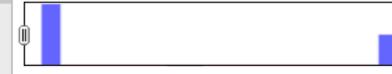
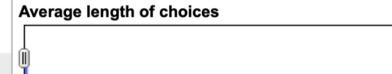
Data type: text	1900 S LONGWOOD AV	1000 S FLOWER ST	1400 W 37TH ST
1900 S LONGWOOD AV	1000 S FLOWER ST	1400 W 37TH ST	1900 S LONGWOOD AV  1000 S FLOWER ST  1400 W 37TH ST

4. In the "LOCATION" field, several entries can be grouped together. The accuracy of the suggested merges was verified using Google Maps, confirming that the values refer to the same location.

**Solution:** Merge "WILSHIRE" with "WILSHIRE BL," "SUNSET" with "SUNSET BL," and "FIGUEROA" with "FIGUEROA ST."

**Cluster and edit column "LOCATION"**

Find groups of different cell values that might be other representations of the same thing. For example, "New York" and "new york" likely refer to the same concept and just differ by capitalization, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method	Nearest neighbor	Distance function	PPM	Radius	1.0	Block chars	6	Auto-update	3 clusters found
Cluster size	Row count	Values in cluster	Merge?	New cell value	# Rows in cluster				
2	518	• WILSHIRE (272 rows) • WILSHIRE BL (246 rows)	<input checked="" type="checkbox"/>	WILSHIRE BL	 510 — 710				
2	517	• SUNSET (312 rows) • SUNSET BL (205 rows)	<input checked="" type="checkbox"/>	SUNSET BL	 7.5 — 9.5				
2	705	• FIGUEROA ST (426 rows) • FIGUEROA (279 rows)	<input checked="" type="checkbox"/>	FIGUEROA ST					

Select all Deselect all      Export clusters      Merge selected & re-cluster      Merge selected & Close      Close

5. Less than half of the entries in the “Crm Cd Desc” column are enclosed in quotation marks. The inconsistency of having some values quoted while others are not, despite identical entries, can create a messy appearance that decreases both consistency and professionalism. This issue is also present in several other columns, including “Weapon Desc,” and the same solution would apply to them.

**Solution:** Apply a text filter to identify entries that have quotation marks, followed by a custom text transformation on the column to ensure uniformity.

AREA NAME	Rpt Dist No	Part 1-2	Crm Cd	Crm Cd Desc
Rampart	246	1	230	"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"
Pacific	1454	1	230	"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"
Northeast	1136	2	740	"VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VANDALISMS)"
Devonshire	1714	1	341	"THEFT-GRAND (\$950.01 & OVER)EXCEPT,GUNS,FOWL,LIVESTK,PROD"
Central	111	2	664	"BUNCO, PETTY THEFT"
77th Street	1242	1	341	"THEFT-GRAND (\$950.01 & OVER)EXCEPT,GUNS,FOWL,LIVESTK,PROD"
77th Street	1215	1	230	"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"
Olympic	2093	2	740	"VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VANDALISMS)"
Pacific	1437	1	341	"THEFT-GRAND (\$950.01 & OVER)EXCEPT,GUNS,FOWL,LIVESTK,PROD"
Southwest	363	2	740	"VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VANDALISMS)"
Olympic	2033	1	320	"BURGLARY, ATTEMPTED"

### Custom text transform on column Crm Cd Desc

Expression: value.replace(/["']/,"")  
Language: General Refine Expression Language (GREL)

No syntax error.

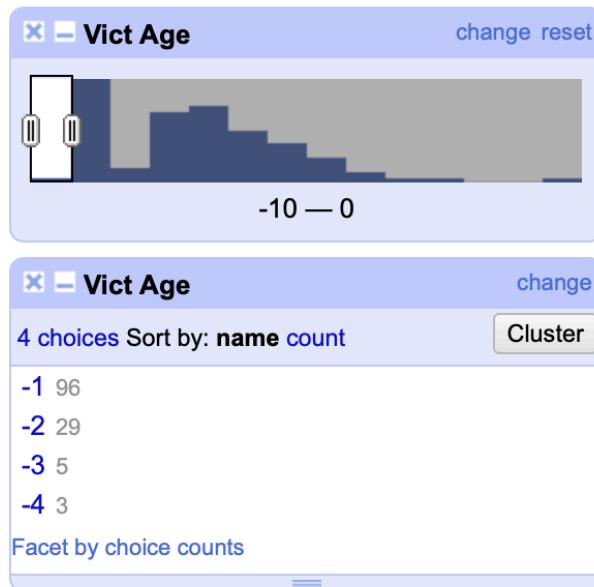
Preview	History	Starred	Help
row	value	value.replace(/["']/,"")	
13.	"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	
16.	"ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT"	ASSAULT WITH DEADLY WEAPON, AGGRAVATED ASSAULT	
33.	"VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VANDALISMS)"	VANDALISM - FELONY (\$400 & OVER, ALL CHURCH VANDALISMS)	
49.	"THEFT-GRAND (\$950.01 & OVER)EXCEPT,GUNS,FOWL,LIVESTK,PROD"	THEFT-GRAND (\$950.01 & OVER)EXCEPT,GUNS,FOWL,LIVESTK,PROD	

On error:  keep original  set to blank  store error

Re-transform up to 10 times until no change

OK Cancel

6. Certain ages are inaccurately recorded as negative numbers (e.g., -2, -1). These are logical errors that compromise data integrity and accuracy.



**Solution:** Remove the rows containing negative victim ages by adjusting the slider to select negative values in a numeric facet, then choose All > Edit Rows > Remove Matching Rows.

The screenshot shows a data editing interface with a sidebar menu. The 'Edit rows' option is selected, and the 'Remove matching rows' option is highlighted with a red box.

- All
- Transform...
- Edit all columns ►
- Facet ►
- Edit rows ►**
- Edit columns ►
- View ►

- 0804701 11/25/2020 11/24/2020
- 0106626 12/31/2020 01/01/2020
- 4000077 00/00/2001 01/01/2000

- Star rows
- Unstar rows
- Flag rows
- Unflag rows
- Remove matching rows**

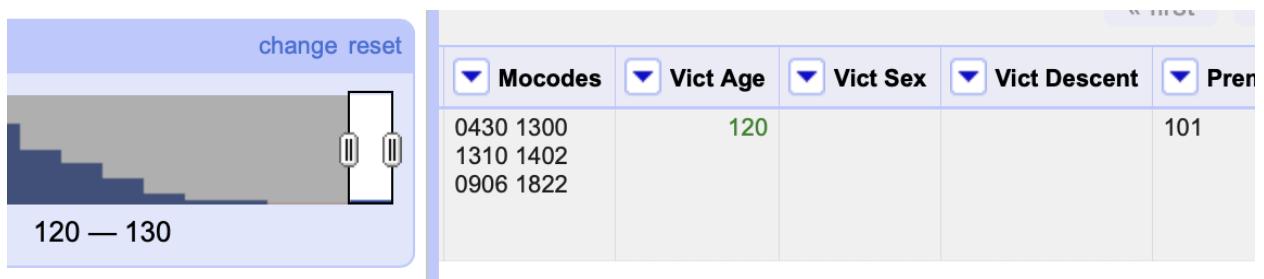
7. Using the text faucet tool, entries with missing age values (represented by "0") can be identified and selected. Since the age information cannot be retrieved, the entries can be removed.

**Solution:** Use the text faucet tool to select all entries with a value of "0," then navigate to: All > Edit Rows > Remove Matching Rows.



8. The 120-130 range in the numeric faucet contains only a single entry, which may warrant further investigation.

**Solution:** The extensive blank data across other columns in this entry suggests it may be an error. To address this, navigate to All > Edit Rows > Remove Matching Rows.



9. "H" is not a valid entry for "Vict Sex," as indicated on the dataset's landing page.

**Solution:** Use the text faucet tool to select all entries in the "Vict Sex" column with a value of "H," then navigate to: All > Edit Rows > Remove Matching Rows.

The screenshot shows the Text Faucet interface with the 'Vict Sex' column selected. The column has 4 choices: F 101942, H 28, M 112251, and (blank) 11. The choice 'H' is highlighted with a red box. To the right, a table view shows rows with columns DR\_NO, Date Rptd, and DATE OCC. A context menu is open over the table, with the 'Remove matching rows' option highlighted in a red box.

10. Vict Sex, Vict Decent, and Vict Sex fields contain 'X' or are left blank.

The screenshot shows the Text Faucet interface with the 'Vict Sex' column selected. The column has 3 choices: F 101942, M 112251, and X 2215. The choice 'X' is highlighted with a red box. To the right, a table view shows rows with columns DR\_NO, Date Rptd, and DATE OCC. A context menu is open over the table, with the 'Remove matching rows' option highlighted in a red box.

**Solution:** Use the text faucet tool to select all entries in the selected column with a value of "blank," and/or "X" then navigate to: All > Edit Rows > Remove Matching Rows.

11. Most "Mocodes" are stored as text, while a few are formatted as numbers, leading to data inconsistencies. The numeric format strips leading zeros from "Mocodes," which can result in misidentification or loss of critical information.

**Solution:** To ensure consistency, convert all entries to text by navigating to Edit cells > Common transforms > To text.

Mocodes	Vict Age	Vict Sex	Vict Descent	Premis Cd	Premis Desc	Wea		
Facet	55	F	H	945	MTA - EXPO LINE - EXPO/VERMONT			
Text filter	40	F	H	502	"MULTI-UNIT DWELLING APTMENT"			
Edit cells	Transform...			Trim leading and trailing whitespace				
Edit column	Common transforms			Collapse consecutive whitespace				
Transpose	Fill down			Unescape HTML entities				
Sort...	Blank down			Replace smart quotes with ASCII				
View	Split multi-valued cells...			To titlecase				
Reconcile	Join multi-valued cells...			To uppercase				
0344 1606	344	Cluster and edit...			To lowercase			
		Replace...			To number			
		42	M	W	To date			
		55	F	H	To text			
					To null			
					To empty string			
0448 1822								
0416 0446								
0356 0910								
2050 2052								
2003								

12. The dates in the "Date Rptd" and "DATE OCC" columns are not properly formatted because they are stored as text. This improper data type leads to issues with the "Timeline faucet," as the dates are being treated incorrectly.

**Solution:** Navigate to Edit cells > Common transforms > To date

Date Rptd	DATE OCC	TIME OCC	AREA	AREA NAME	Rpt Dist No	Part
Facet	020 0 AM	2130	07	Wilshire	0784	1
Text filter	000 1000 2000	01		Central	0182	1
Edit cells	Transform...					
Edit column	Common transforms					
Transpose	Fill down					
Sort...	Blank down					
View	Unescape HTML entities					
Reconcile	Replace multi-valued cells...					
2022-08-18T12:00:00Z	08/17/2022 12:00:00					
2023-04-04T12:00:00Z	12/01/2023 12:00:00					
2023-04-04T12:00:00Z	07/03/2020 12:00:00 AM	0900	01			
2022-07-22T12:00:00Z	05/12/2020 12:00:00 AM	1110	03			
2023-04-28T12:00:00Z	12/09/2020 12:00:00 AM	1400	13	Newton	1375	2
2020-12-31T12:00:00Z	12/31/2020 12:00:00 AM	1220	19	Mission	1974	2

Before

**Date Rptd** change reset

**Invalid Date Invalid Date — Wed Dec 31 1969 18:00:00 GMT-0600 (Central Standard Time)**  
**Wed Dec 31 1969 18:00:00 GMT-0600 (Central Standard Time)**

Time    Non-Time    Blank    Error

0	978628	0	0
---	--------	---	---

After

**Date Rptd** change reset

**Wed Jan 01 2020 06:00:00 GMT-0600 (Central Standard Time) Wed Jan 01 2020 06:00:00 GMT-0600 (Central Standard Time) — Mon Sep 16 2024 07:00:00 GMT-0500 (Central Daylight Time) Mon Sep 16 2024 07:00:00 GMT-0500 (Central Daylight Time)**