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

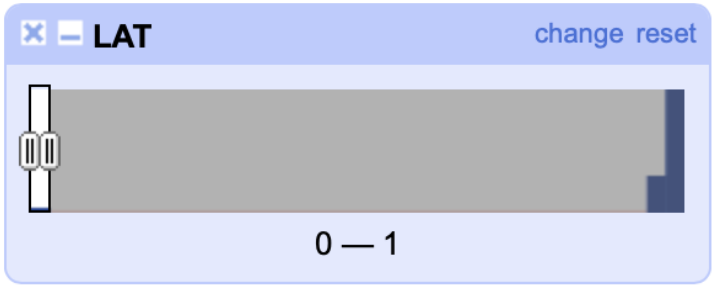
- **Redundant Columns:**
    - Duplicate information across multiple columns (e.g., having both "Full Name" and separate "First Name" and "Last Name" columns). In OpenRefine, the redundant column(s) can be deleted By going to All > Edit columns > Re-order / remove …
- **Inconsistent Relationships:**
    - 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

LOCATION	Cross Street	LAT	LON
19300 BUSINESS CENTER DR		0	0

After

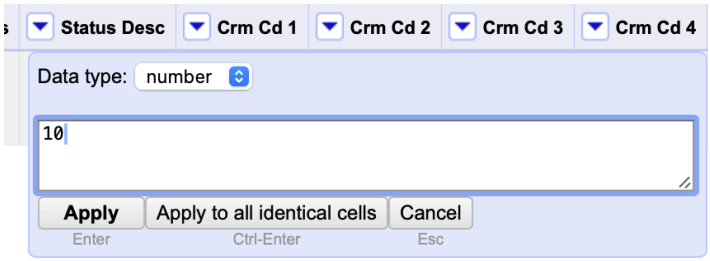
LOCATION	Cross Street	LAT	LON
19300 BUSINESS CENTER DR		34.230892	-118.553713

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.

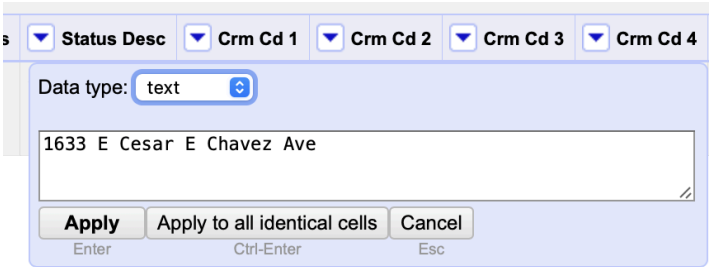
   Solution: "FREEWAY" was removed from the "Cross Street" field by clicking "Edit," deleting the text, and applying the change. Additionally, the data type for "LOCATION" was updated from number to text, and the complete address was entered.

Before



LOCATION	Cross Street	LAT	LON
10	FREEWAY	34.0517	-118.2178

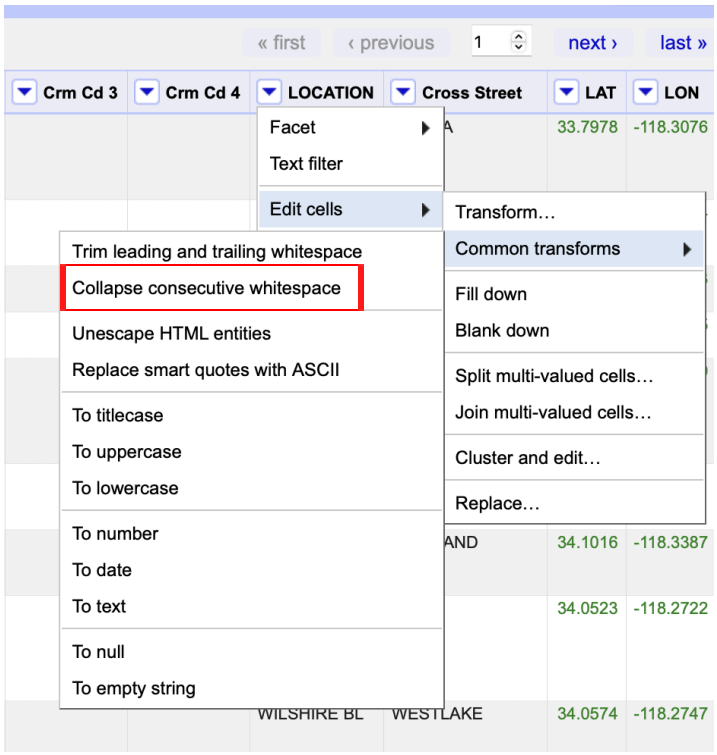
After



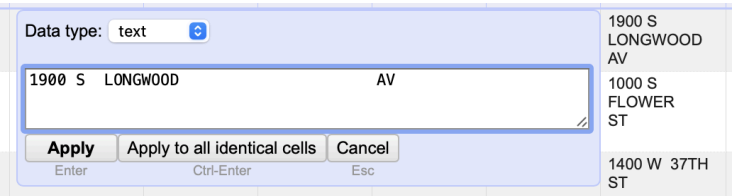
LOCATION	Cross Street	LAT	LON
1633 E Cesar E Chavez Ave		34.0517	-118.2178

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.

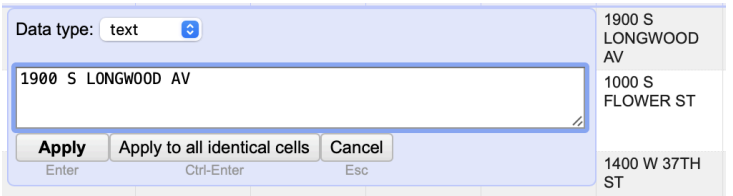
   Solution: Apply "Collapse consecutive whitespace" to standardize the formatting.



Before

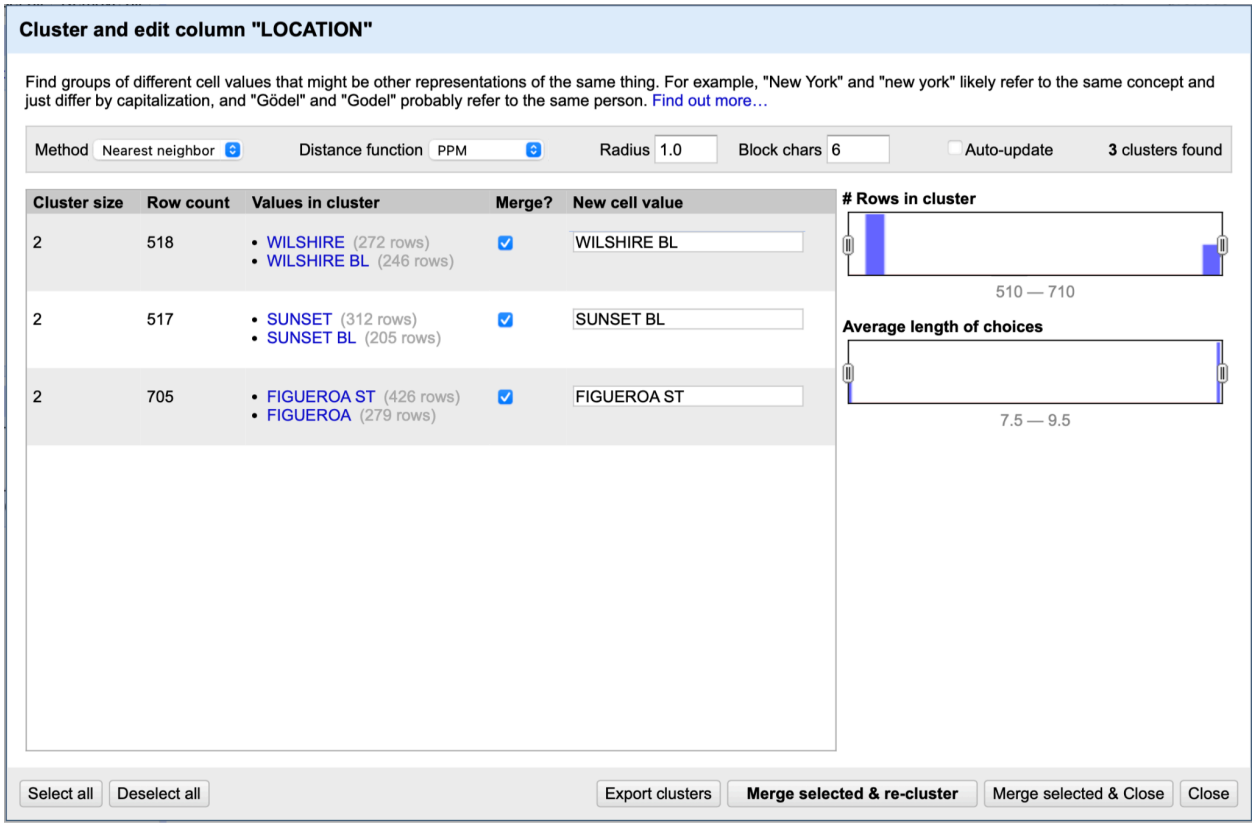


After



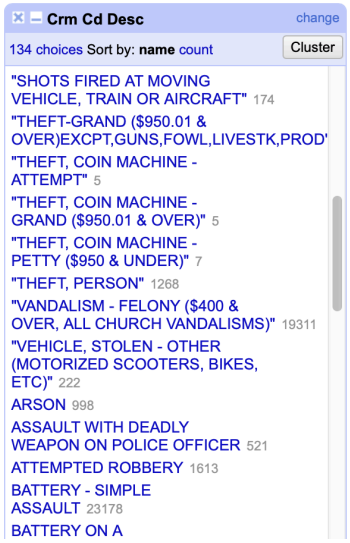
4. In the "LOCATION" field, several entries can be grouped together. The accuracy of the suggested merges was verified using Google Maps, confirming that the values refer to the same location.

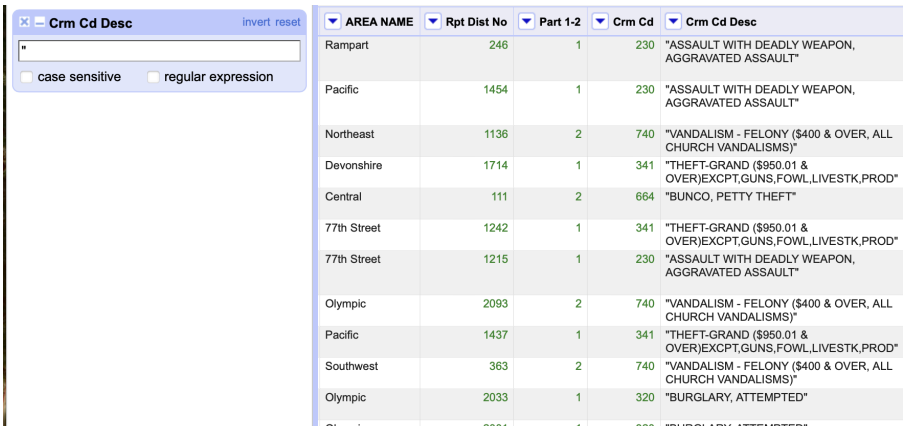
   Solution: Merge "WILSHIRE" with "WILSHIRE BL," "SUNSET" with "SUNSET BL," and "FIGUEROA" with "FIGUEROA ST."

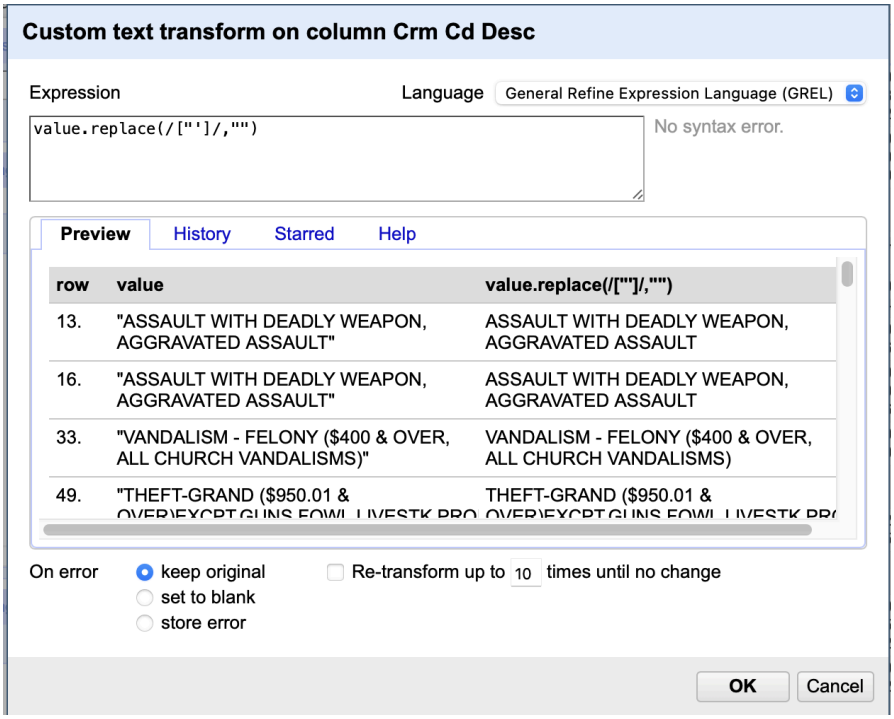


5. Less than half of the entries in the “Crm Cd Desc” column are enclosed in quotation marks. The inconsistency of having some values quoted while others are not, despite identical entries, can create a messy appearance that decreases both consistency and professionalism. This issue is also present in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

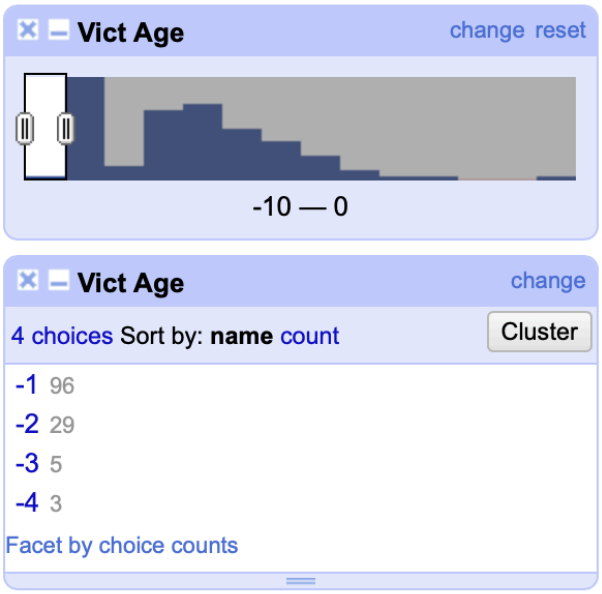
   Solution: Apply a text filter to identify entries that have quotation marks, followed by a custom text transformation on the column to ensure uniformity.



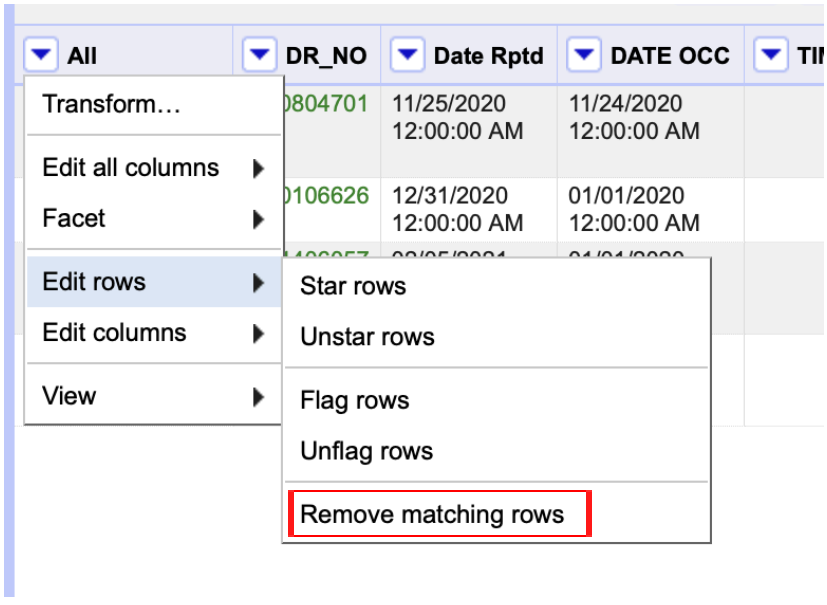




6. Certain ages are inaccurately recorded as negative numbers (e.g., -2, -1). These are logical errors that compromise data integrity and accuracy.

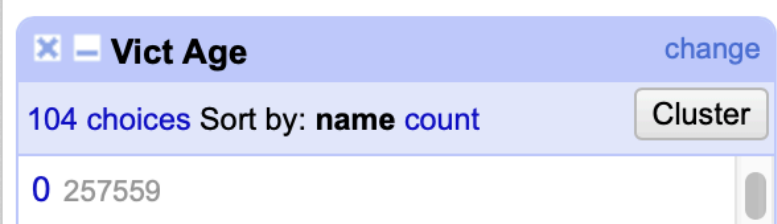


Solution: Remove the rows containing negative victim ages by adjusting the slider to select negative values in a numeric facet, then choose All > Edit Rows > Remove Matching Rows.



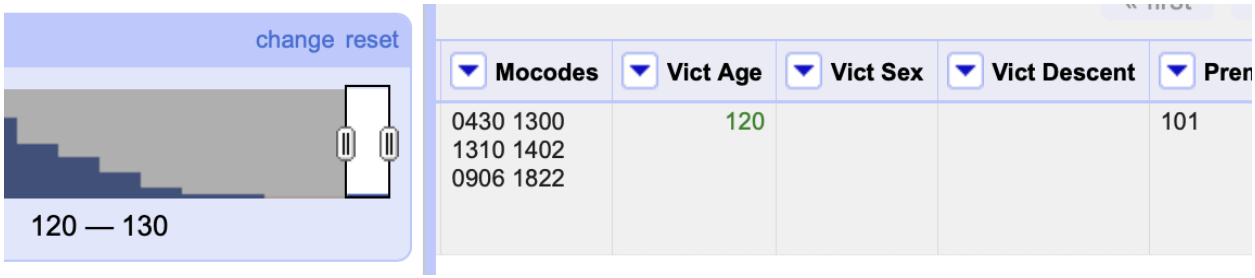
7. Using the text faucet tool, entries with missing age values (represented by "0") can be identified and selected. Since the age information cannot be retrieved, the entries can be removed.

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



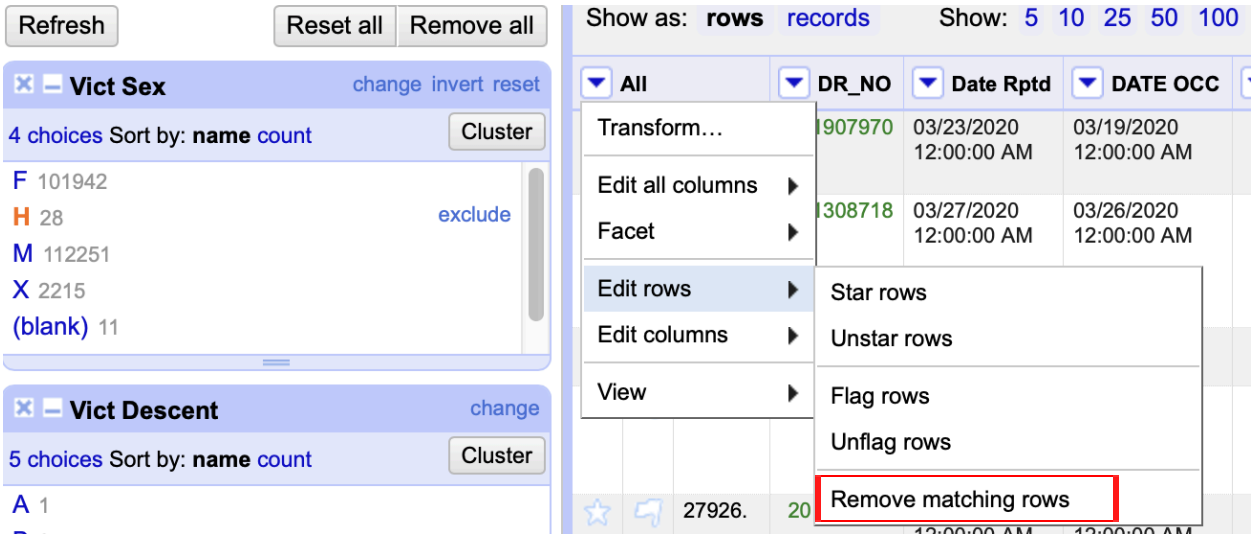
8. The 120-130 range in the numeric faucet contains only a single entry, which may warrant further investigation.

   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.

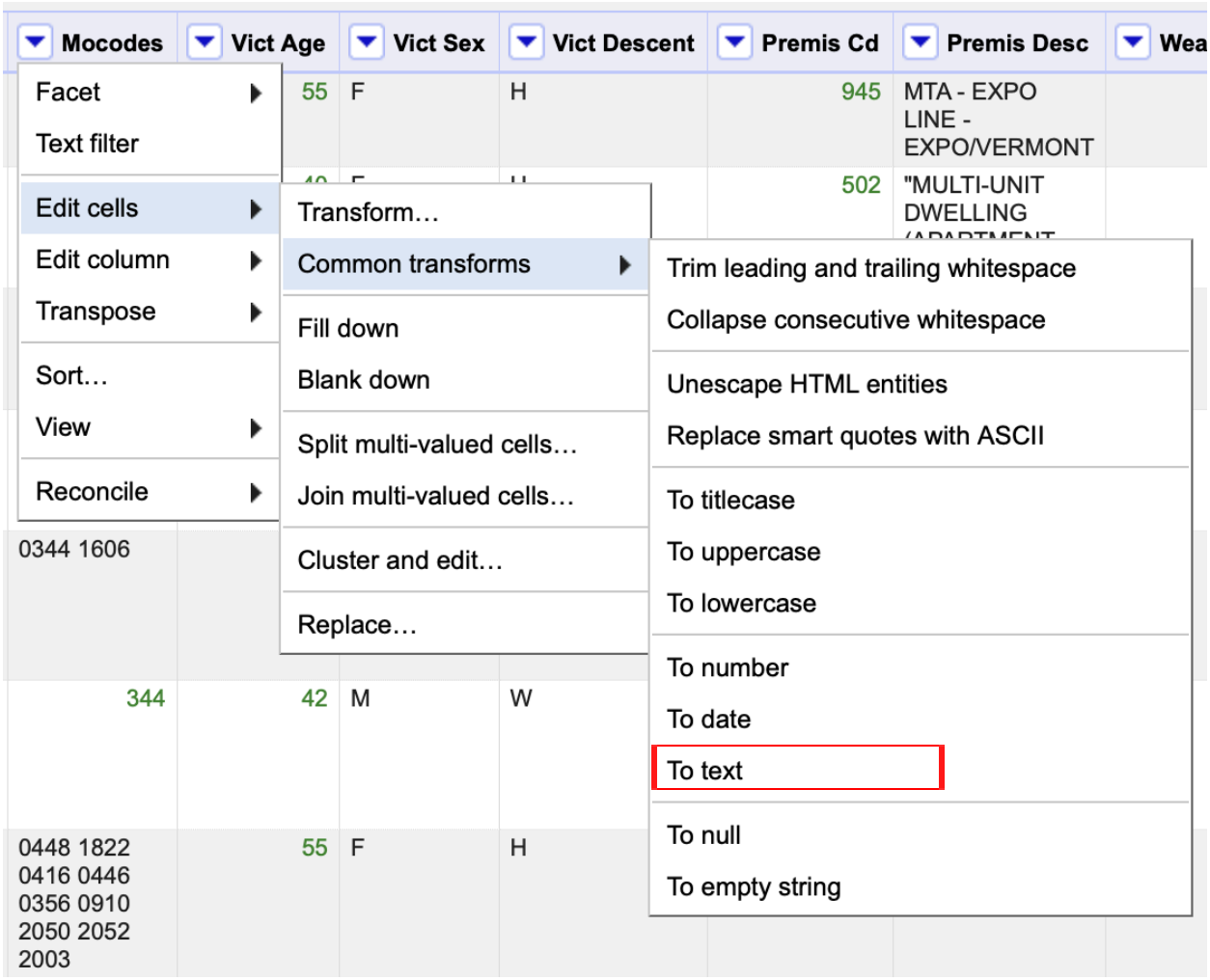


10. Vict Sex, Vict Decent, and Vict Sex fields contain 'X' or are left blank.

   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.

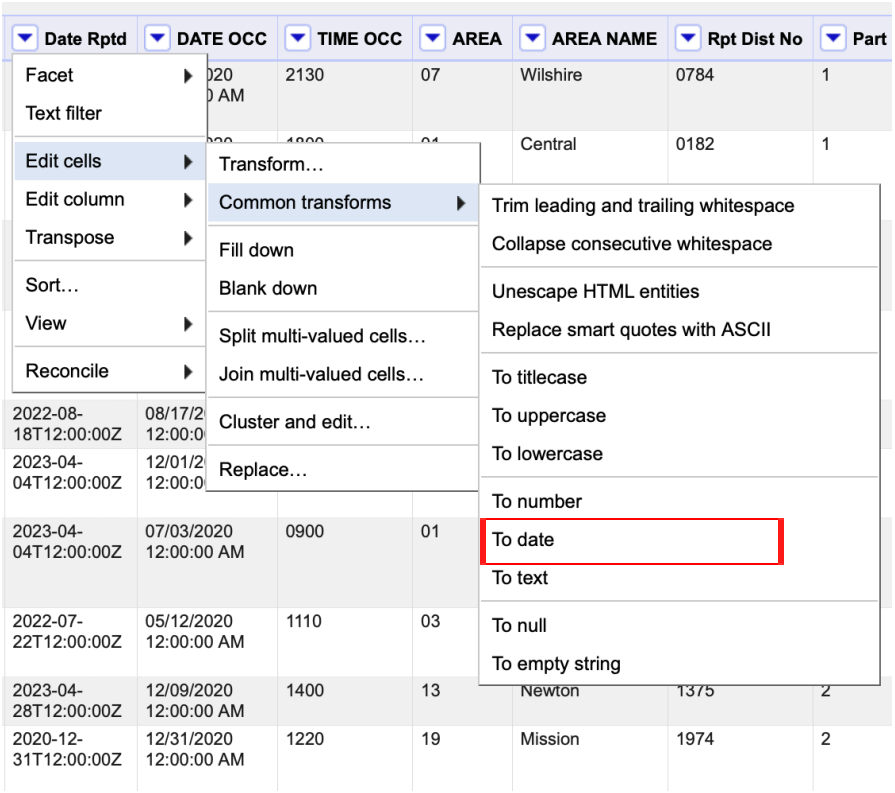
11. Most "Mocodes" are stored as text, while a few are formatted as numbers, leading to data inconsistencies. The numeric format strips leading zeros from "Mocodes," which can result in misidentification or loss of critical information.

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

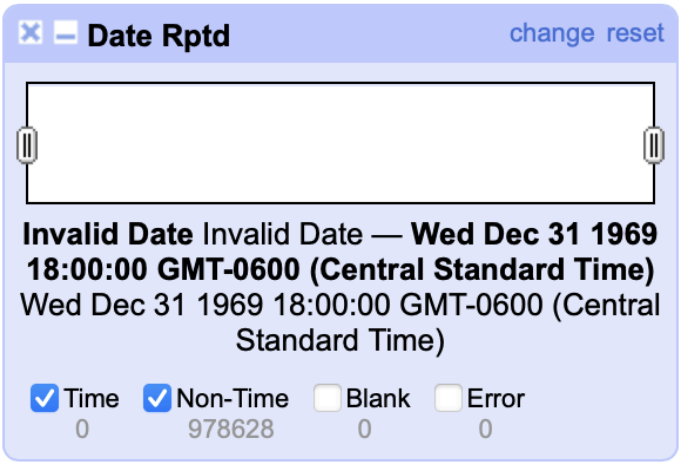


12. The dates in the "Date Rptd" and "DATE OCC" columns are not properly formatted because they are stored as text. This improper data type leads to issues with the "Timeline faucet," as the dates are being treated incorrectly.

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



Before



After