Deduplicate Transform

> **NOTE:** Transforms are a part of the underlying language, which is not directly accessible to users. This content is maintained for reference purposes only. For more information on the user-accessible equivalent to transforms, see *Transformation Reference*.

Removes exact duplicate rows from your dataset. Duplicate rows are identified by exact, case-sensitive matches between values.

For example, two strings with different capitalization do not match.

Basic Usage

```
deduplicate
```

Output: Rows that are exact duplicates of previous rows are removed from the dataset.

Syntax and Parameters

There are no parameters for this transform.

Examples

> **Tip:** For additional examples, see *Common Tasks*.

Matches and non-matches for Deduplicate Transform

Source:

For example, your dataset looks like the following, which contains three sets of very similar records. The second row of each set is different in one column than the previous one.

Name	Date	Score
Joe Jones	1/2/03	88
joe jones	1/2/03	88
Jane Jackson	2/3/04	77
Jane Jackson	February 3, 2004	77
Jill Johns	3/4/05	66
Jill Johns	3/4/05	66.00

Transformation:

Transformation Name	Remove duplicate rows

If you remove duplicate rows on this dataset, no rows are previewed. This preview indicates that no rows will be removed as duplicates. You might need to clean up the data before you can remove any duplicate rows.

Your first step should be get your capitalization consistent. Try the following:

Transformation Name	`Edit column with formula`
Parameter: Columns	`Name`
Parameter: Formula	`proper(Name)`

All entries in the `Name` column now appear as proper names. Next, you can clean up the score column by normalizing numeric values to the same format. Try the following:

Transformation Name	`Edit column with formula`
Parameter: Columns	`Score`
Parameter: Formula	`numformat(Score, '##.00')`

The above transformation normalizes the numeric formats to include two-digits after the decimal point always, which forces all numbers to be the same format. You can use the ## format string here, too.

Use the following to fix the Date column:

Transformation Name	`Replace text or pattern`
Parameter: Column	`Date`
Parameter: Find	`'February 3, 2004'`
Parameter: Replace with	`'2/3/04'`

Now, you can deduplicate your dataset:

Transformation Name	`Remove duplicate rows`

Results:

Name	Date	Score
Joe Jones	1/2/03	88.00
Jane Jackson	2/3/04	77.00
Jill Johns	3/4/05	66.00