

Step-by-Step Guide to Testing OpenRefine with the CSV

Step 1: Install and Launch OpenRefine

1. Download OpenRefine:

- Visit the official OpenRefine download page.(<https://openrefine.org/docs>)
- Download the version that matches your operating system (Windows, macOS, or Linux).











2. Install OpenRefine:

- For Windows: After downloading, extract the .zip file. Open the folder and run refine.bat.
- For macOS: Drag the extracted OpenRefine application into the "Applications" folder and double-click to open.
- For Linux: Extract the .tar.gz file and open a terminal window, navigate to the extracted folder, and run ./refine.

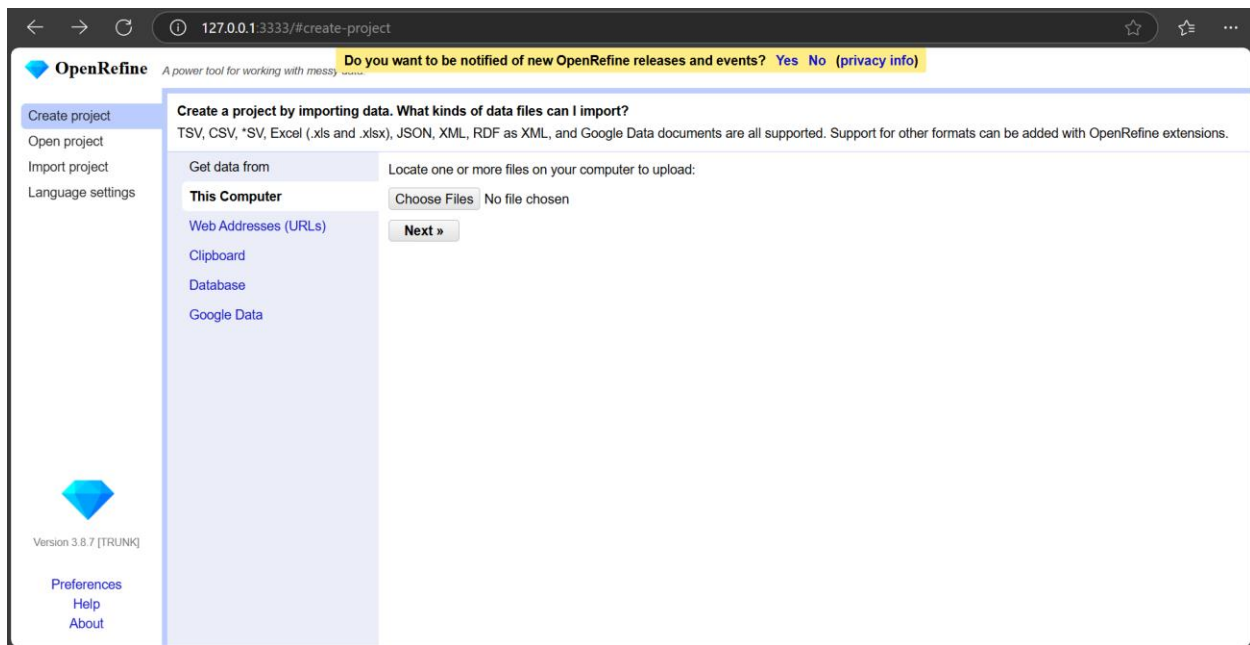
3. Start OpenRefine:

- Once OpenRefine is running, open your web browser and go to <http://127.0.0.1:3333/> to access the OpenRefine interface.

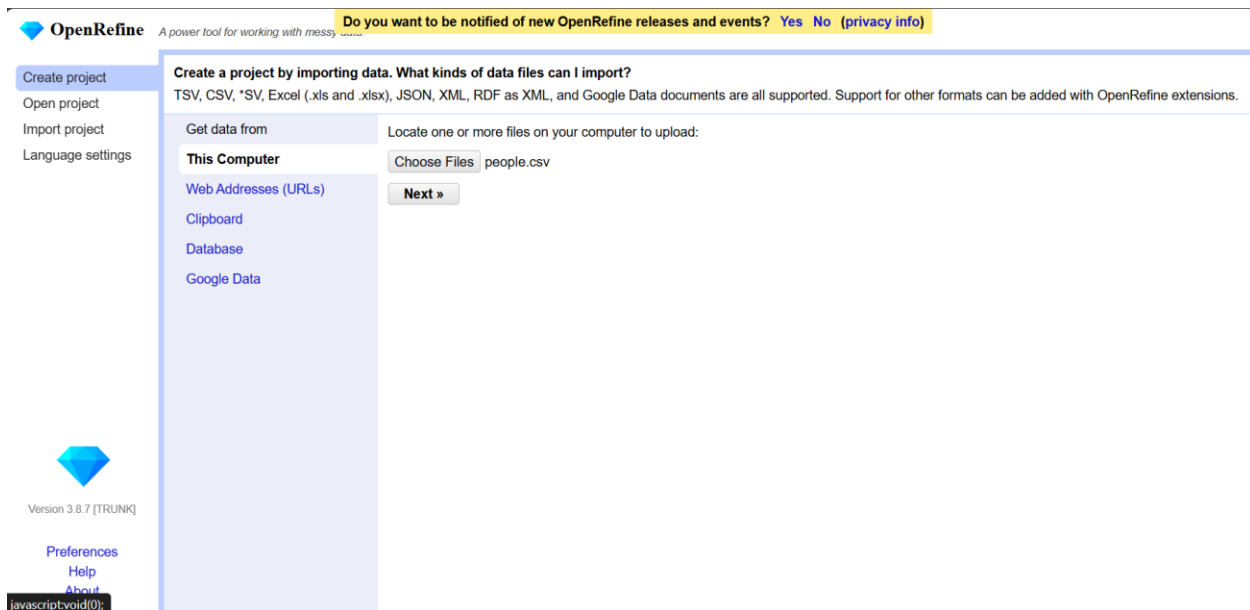
For Windows: After downloading, extract the .zip file. Open the folder and run refine.bat or openrefine.exe.

| Name | Date modified | Type | Size |
|--|---------------------|------------------------|--------|
| ▼ Today | | | |
|  LICENSE.txt | 12/21/2024 10:31 PM | Text Document | 2 KB |
|  licenses.xml | 12/21/2024 10:31 PM | Microsoft Edge HTM... | 13 KB |
|  openrefine.exe | 12/21/2024 10:31 PM | Application | 125 KB |
|  openrefine.l4j.ini | 12/21/2024 10:31 PM | Configuration settings | 1 KB |
|  README.md | 12/21/2024 10:31 PM | Markdown Source File | 4 KB |
|  refine.bat | 12/21/2024 10:31 PM | Windows Batch File | 10 KB |
|  refine.ini | 12/21/2024 10:31 PM | Configuration settings | 2 KB |
|  licenses | 12/21/2024 10:31 PM | File folder | |
|  server | 12/21/2024 10:31 PM | File folder | |
|  webapp | 12/21/2024 10:31 PM | File folder | |

Click "**Choose Files**", find the people.csv file you downloaded earlier, and select it.



Click "Next"



Click "Create Project" to load the file into OpenRefine.

OpenRefine A power tool for working with messy data

Do you want to be notified of new OpenRefine releases and events? [Yes](#) [No](#) ([privacy info](#))

Create project [start over](#) Configure parsing options Project name Tags [Create project »](#)

Open project
Import project
Language settings

| | Name | Age | City | Email |
|----|---------------|-----|---------------|-------------------------|
| 1. | john smith | 28 | New York | john.smith@email.com |
| 2. | jane doe | 32 | San Francisco | jane_doe@email.com |
| 3. | alice johnson | 25 | Los Angeles | alice.johnson@email.com |
| 4. | bob brown | 27 | New York | bob.brown@email.com |
| 5. | charlie davis | 30 | Chicago | charlie.davis@email.com |
| 6. | eve white | 29 | San Francisco | eve.white@email.com |
| 7. | eve white | 29 | San Francisco | eve_white@email.com |

Version 3.8.7 [TRUNK]

[Preferences](#)
[Help](#)
[About](#)

Parse data as

CSV / TSV / separator-based files

Line-based text files
Fixed-width field text files
PC-Axis text files
JSON files
MARC files
JSON-LD files
RDF/N3 files
RDF/N-Triples files
RDF/Turtle files

Character encoding [Update preview](#)

☐ Disable auto preview

Columns are separated by

☒ commas (CSV)
☐ tabs (TSV)
☐ custom ,

☒ Use character " to enclose cells containing column separators
☐ Trim leading & trailing whitespace from strings
Escape special characters with \

☐ Ignore first 0 line(s) at beginning of file
☒ Parse next 1 line(s) as column headers
☐ Column names (comma separated)

☐ Discard initial 0 row(s) of data
☐ Load at most 0 row(s) of data

☐ Attempt to parse cell text into numbers
☒ Store blank rows
☒ Store blank cells as nulls
☐ Store file source
☐ Store archive file

1.Transform the Data

Click the drop-down menu for the **Name** column.

OpenRefine people csv [Permalink](#) [Open...](#) [Export](#) [Help](#)

Facet / Filter [Undo / Redo 0 / 0](#) [7 rows](#) [Extensions Wikibase](#)

Using facets and filters

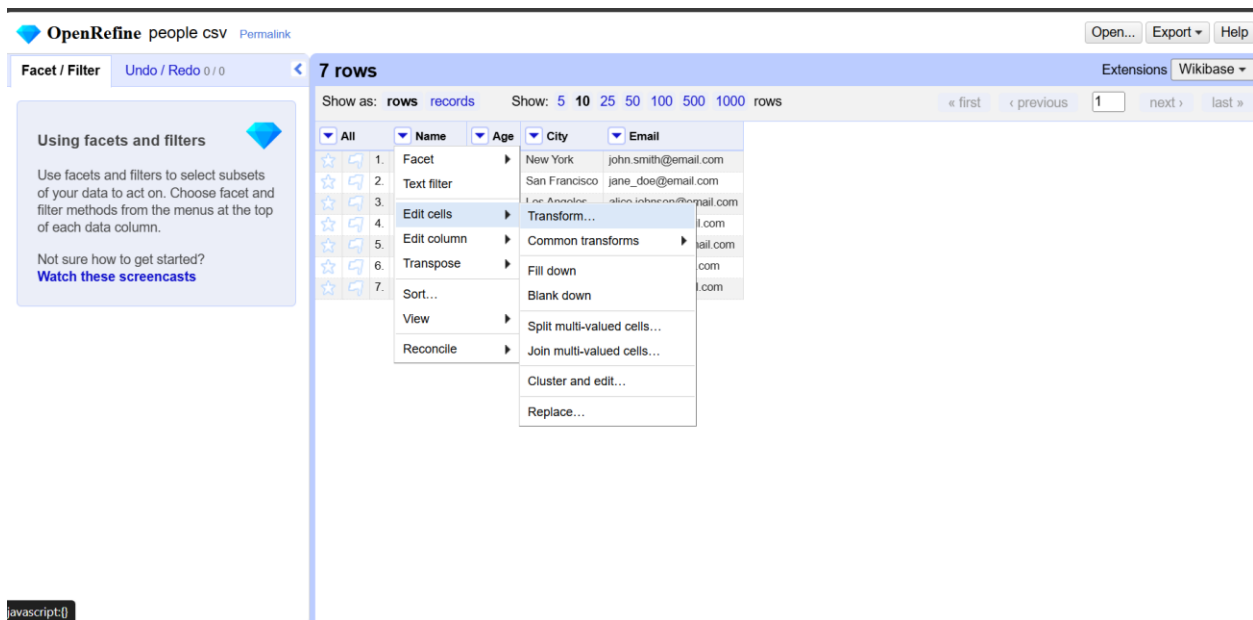
Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

Show as: [rows](#) [records](#) Show: [5](#) [10](#) [25](#) [50](#) [100](#) [500](#) [1000](#) rows [« first](#) [« previous](#) [1](#) [next »](#) [last »](#)

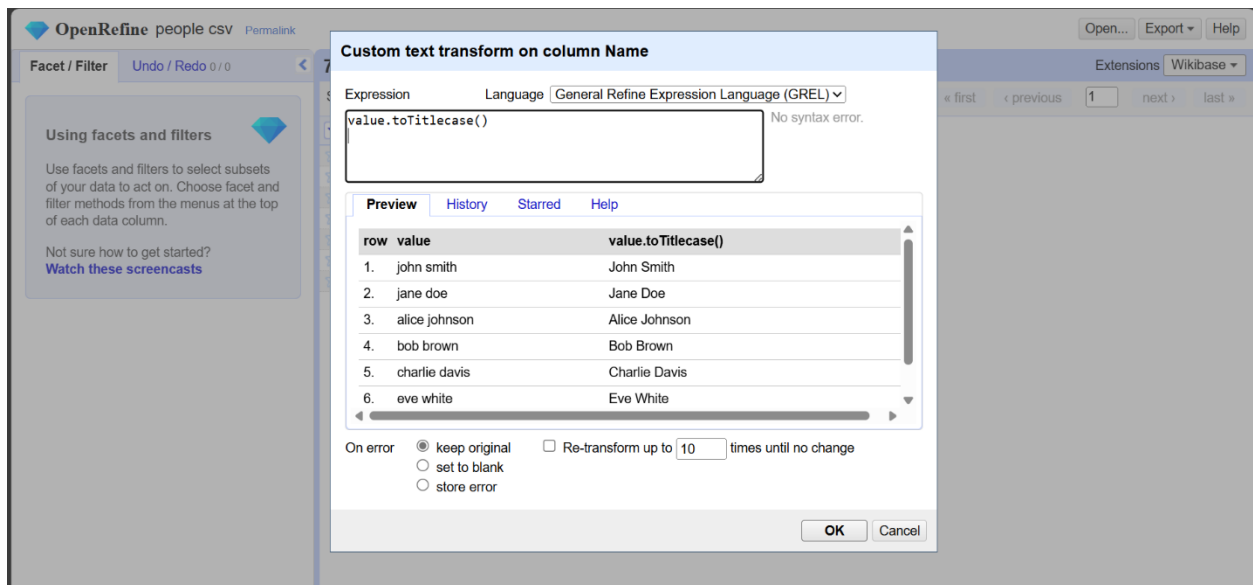
| | All | Name | Age | City | Email |
|----|---------------|------|---------------|-------------------------|-------|
| 1. | john smith | 28 | New York | john.smith@email.com | |
| 2. | jane doe | 32 | San Francisco | jane_doe@email.com | |
| 3. | alice johnson | 25 | Los Angeles | alice.johnson@email.com | |
| 4. | bob brown | 27 | New York | bob.brown@email.com | |
| 5. | charlie davis | 30 | Chicago | charlie.davis@email.com | |
| 6. | eve white | 29 | San Francisco | eve.white@email.com | |
| 7. | eve white | 29 | San Francisco | eve_white@email.com | |

Select **Edit cells > Transform**.



In the transformation dialog, enter this expression:

`"value.toTitlecase()"`



Result

OpenRefine people csv [Permalink](#)

Text transform on 7 cells in column Name:
grel:value.toTitlecase() [Undo](#)

Facet / Filter [Undo / Redo](#) 1 / 1 [7 rows](#)

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows « first

| | All | Name | Age | City | Email |
|----|-----|---------------|-----|---------------|-------------------------|
| 1. | | John Smith | 28 | New York | john.smith@email.com |
| 2. | | Jane Doe | 32 | San Francisco | jane_doe@email.com |
| 3. | | Alice Johnson | 25 | Los Angeles | alice.johnson@email.com |
| 4. | | Bob Brown | 27 | New York | bob.brown@email.com |
| 5. | | Charlie Davis | 30 | Chicago | charlie.davis@email.com |
| 6. | | Eve White | 29 | San Francisco | eve.white@email.com |
| 7. | | Eve White | 29 | San Francisco | eve_white@email.com |

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

2.Remove Duplicates in the Email Column:

To clean up duplicate email addresses (e.g., eve.white@email.com and eve_white@email.com), click the drop-down menu for the **Email** column.

Select **Edit cells > Cluster and edit....**

OpenRefine people csv [Permalink](#) [Open...](#) [Export](#) [Help](#)

Facet / Filter [Undo / Redo](#) 0 / 0 [7 rows](#) Extensions [Wikibase](#)

Show as: **rows** records Show: 5 10 25 50 100 500 1000 rows « first < previous 1 next > last »

| | All | Name | Age | City | Email |
|----|-----|---------------|-----|---------------|-------|
| 1. | | john smith | 28 | New York | |
| 2. | | jane doe | 32 | San Francisco | |
| 3. | | alice johnson | 25 | Los Angeles | |
| 4. | | bob brown | 27 | New York | |
| 5. | | charlie davis | 30 | Chicago | |
| 6. | | eve white | 29 | San Francisco | |
| 7. | | eve white | 29 | San Francisco | |

Facet

Text filter

Edit cells

Edit column

Transpose

Sort...

View

Reconcile

Transform...

Common transforms

Fill down

Blank down

Split multi-valued cells...

Join multi-valued cells...

Cluster and edit...

Replace...

OpenRefine will display groups of similar values. Review the clusters and merge any duplicates (like the two entries for "eve.white").

Cluster and edit column "Email"

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 Key collision Keying function Fingerprint ☒ Auto-update 1 cluster found

| Cluster size | Row count | Values in cluster | Merge? | New cell value |
|--------------|-----------|---|--------------------------|---------------------|
| 2 | 2 | <ul style="list-style-type: none">eve.white@email.comeve_white@email.com | <input type="checkbox"/> | eve.white@email.com |

Click "Browse this Cluster"

Cluster and edit column "Email"

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 Key collision Keying function Fingerprint ☒ Auto-update 1 cluster found

| Cluster size | Row count | Values in cluster | Merge? | New cell value |
|--------------|-----------|---|-------------------------------------|---------------------|
| 2 | 2 | <ul style="list-style-type: none">eve.white@email.comeve_white@email.com Browse this cluster | <input checked="" type="checkbox"/> | eve.white@email.com |

Click "edit"

The screenshot shows the OpenRefine web application interface. At the top, the title bar says "OpenRefine people.csv" with a "Permalink" button. Below the title bar, there are buttons for "Open...", "Export", and "Help". The main interface is divided into several sections:

- Facet / Filter:** On the left, there is a facet for "Email" with 7 choices. The choices are listed with their counts: "alice.johnson@email.com" (1), "bob.brown@email.com" (1), "charlie.davis@email.com" (1), "eve_white@email.com" (1), "eve.white@email.com" (1), "jane_doe@email.com" (1), and "john.smith@email.com" (1). The "eve.white@email.com" choice is selected and highlighted in blue. There are buttons for "Cluster", "exclude", and "edit" next to the selected choice.
- 2 matching rows (7 total):** In the center, there is a table showing the first two rows of the data. The table has columns for "All", "Name", "Age", "City", and "Email". The rows are:

| | | | | |
|---|-----------|----|---------------|---------------------|
| 6 | eve white | 29 | San Francisco | eve.white@email.com |
| 7 | eve white | 29 | San Francisco | eve_white@email.com |
- Extensions:** On the right, there is a dropdown menu for "Extensions" with "Wikibase" selected.

Edit and Click “Apply”

The screenshot shows the OpenRefine interface with the 'people.csv' dataset. The 'Email' facet is active on the left, showing a list of email addresses. A modal dialog is open over the 'Email' column, allowing the user to edit the cell 'ever.white@email.com'. The dialog has an 'Apply' button and a 'Cancel' button. The main table shows 2 matching rows (7 total) with columns: All, Name, Age, City, and Email.

| All | Name | Age | City | Email |
|-------------------------------------|-----------|-----|---------------|----------------------|
| <input checked="" type="checkbox"/> | eve white | 29 | San Francisco | eve.white@email.com |
| <input checked="" type="checkbox"/> | eve white | 29 | San Francisco | ever.white@email.com |

Click “Reset all”

The screenshot shows the OpenRefine interface with the 'people.csv' dataset. The 'Email' facet is active on the left, showing a list of email addresses. The 'Reset all' button is visible in the top left corner of the facet panel. The main table shows 7 rows with columns: All, Name, Age, City, and Email.

| All | Name | Age | City | Email |
|-------------------------------------|---------------|-----|---------------|-------------------------|
| <input checked="" type="checkbox"/> | john smith | 28 | New York | john.smith@email.com |
| <input checked="" type="checkbox"/> | jane doe | 32 | San Francisco | jane_doe@email.com |
| <input checked="" type="checkbox"/> | alice johnson | 25 | Los Angeles | alice.johnson@email.com |
| <input checked="" type="checkbox"/> | bob brown | 27 | New York | bob.brown@email.com |
| <input checked="" type="checkbox"/> | charlie davis | 30 | Chicago | charlie.davis@email.com |
| <input checked="" type="checkbox"/> | eve white | 29 | San Francisco | ever.white@email.com |
| <input checked="" type="checkbox"/> | eve white | 29 | San Francisco | eve.white@email.com |

3.Facet by City:

To group and analyze your data by city, click the drop-down menu for the **City** column.

Choose **Facet > Text facet**.

The screenshot shows the OpenRefine interface with the 'people.csv' dataset. The 'City' column menu is open, showing options like 'Facet', 'Text filter', 'Edit cells', 'Edit column', 'Transpose', 'Sort...', 'View', and 'Reconcile'. The 'Text facet' option is selected. The main table shows 7 rows with columns: All, Name, Age, City, and Email.

| All | Name | Age | City | Email |
|-------------------------------------|---------------|-----|-------------|-------------------------|
| <input checked="" type="checkbox"/> | john smith | 28 | Facet | Text facet |
| <input checked="" type="checkbox"/> | jane doe | 32 | Text filter | Numeric facet |
| <input checked="" type="checkbox"/> | alice johnson | 25 | Edit cells | Timeline facet |
| <input checked="" type="checkbox"/> | bob brown | 27 | Edit column | Scatterplot facet... |
| <input checked="" type="checkbox"/> | charlie davis | 30 | Transpose | Custom text facet... |
| <input checked="" type="checkbox"/> | eve white | 29 | Sort... | Custom numeric facet... |
| <input checked="" type="checkbox"/> | eve white | 29 | View | Customized facets |
| <input checked="" type="checkbox"/> | eve white | 29 | Reconcile | |

A facet will appear on the left side of the screen, showing all unique cities and the number of records for each. You can click on any city to filter the data for that city.

The screenshot shows the OpenRefine interface with a CSV file named 'people.csv'. On the left, a facet for the 'City' column is active, showing 4 choices: Chicago (1), Los Angeles (1), New York (2), and San Francisco (3). The main table displays 7 rows of data. The columns are All, Name, Age, City, and Email. The data is as follows:

| | All | Name | Age | City | Email |
|----|---------------|------|---------------|-------------------------|-------|
| 1. | john smith | 28 | New York | john.smith@email.com | |
| 2. | jane doe | 32 | San Francisco | jane_doe@email.com | |
| 3. | alice johnson | 25 | Los Angeles | alice.johnson@email.com | |
| 4. | bob brown | 27 | New York | bob.brown@email.com | |
| 5. | charlie davis | 30 | Chicago | charlie.davis@email.com | |
| 6. | eve white | 29 | San Francisco | ever.white@email.com | |
| 7. | eve white | 29 | San Francisco | eve_white@email.com | |

4.Split the Name Column:

If you want to split the **Name** column into **First Name** and **Last Name**, click the drop-down menu for the **Name** column.

Select **Edit column > Split into several columns....**

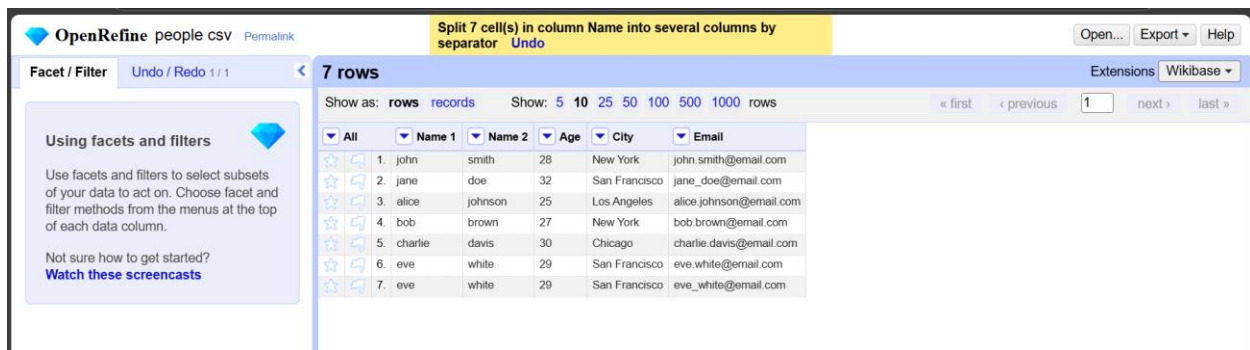
The screenshot shows the OpenRefine interface with the 'Edit column' menu open for the 'Name' column. The menu options are:

- Facet
- Text filter
- Edit cells
- Edit column
 - Split into several columns...
 - Join columns...
- Transpose
- Sort...
- View
 - Add column based on this column...
 - Add column by fetching URLs...
- Reconcile
 - Add columns from reconciled values...
 - Rename this column...
 - Remove this column
 - Move column to beginning
 - Move column to end
 - Move column left
 - Move column right

In the dialog that appears, choose **Space** as the separator.

The screenshot shows the 'Split column "Name" into several columns' dialog box. The 'How to split column' section has two options: 'by separator' (selected) and 'by field lengths'. The 'by separator' option has a text input field for the separator (set to ' ') and a checkbox for 'regular expression'. The 'Split into' field is set to 'columns at most (leave blank for no limit)'. The 'After Splitting' section has two checkboxes: 'Guess cell type' and 'Remove this column', both of which are checked. The 'List of integers separated by commas, e.g., 5, 7, 15' field is empty. The 'OK' and 'Cancel' buttons are at the bottom.

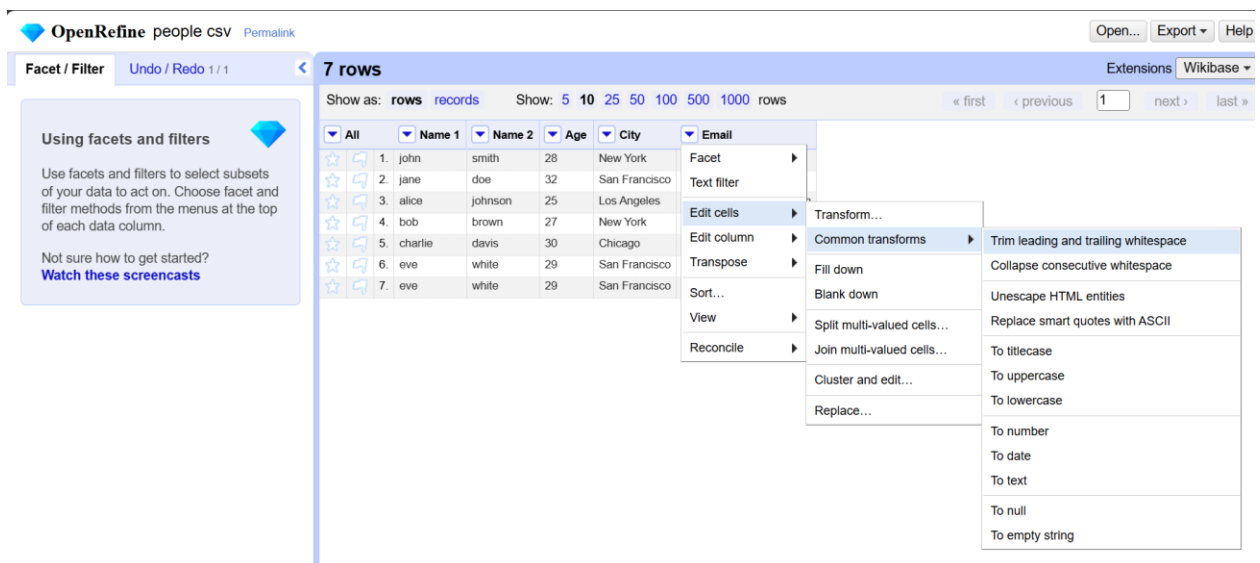
Click **OK** to split the column into two: "First Name" and "Last Name".



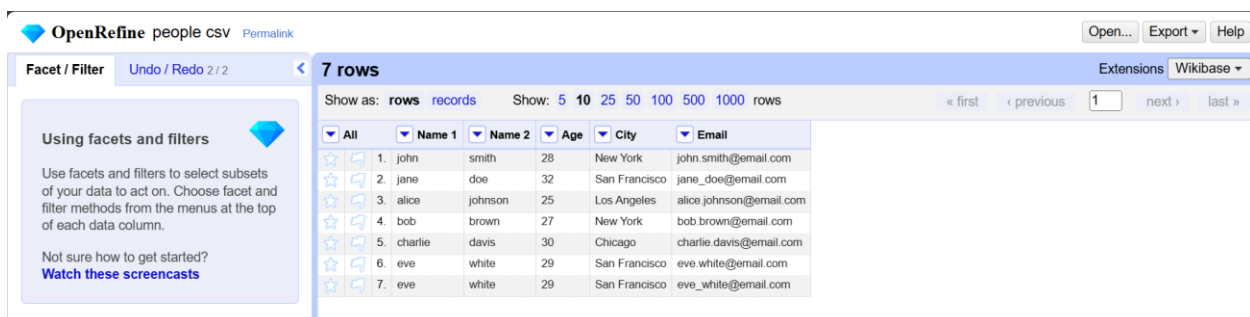
5.Trim Whitespace from Emails:

If there are any extra spaces in the email addresses, click the drop-down menu for the **Email** column.

Select **Edit cells > Trim leading and trailing spaces**.



This will clean any extra spaces in the email addresses.



5.Export the Cleaned Data

Once you've cleaned and transformed the data, you can export the cleaned dataset:

Export Data:

- At the top-right of the OpenRefine interface, click on the **Export** button.
- Select **Export to CSV** (or choose any other format you prefer).

The screenshot shows the OpenRefine web interface. At the top, there's a header with the OpenRefine logo, the file name 'people.csv', and a 'Permalink' button. Below the header, there's a 'Facet / Filter' section on the left with a 'Using facets and filters' sidebar. The main area displays a table with 7 rows of data. The columns are labeled 'All', 'Name 1', 'Name 2', 'Age', 'City', and 'Email'. The data rows are numbered 1 through 7. On the right side, there's a top bar with 'Open...', 'Export', and 'Help' buttons. The 'Export' button is clicked, and a dropdown menu is visible, listing various export formats: 'OpenRefine project archive to file', 'Tab-separated value', 'Comma-separated value', 'HTML table', 'Excel (.xls)', 'Excel 2007+ (.xlsx)', 'ODF spreadsheet', 'Custom tabular...', 'SQL...', 'Templating...', 'OpenRefine project archive to Google Drive...', 'Google Sheets...', 'Wikibase edits...', 'QuickStatements file', and 'Wikibase schema'.

OpenRefine people.csv Permalink

Facet / Filter Undo / Redo 2 / 2 7 rows

Show as: rows records Show: 5 10 25 50 100 500 1000 rows

Using facets and filters

Use facets and filters to select subsets of your data to act on. Choose facet and filter methods from the menus at the top of each data column.

Not sure how to get started?
[Watch these screencasts](#)

| | All | Name 1 | Name 2 | Age | City | Email |
|----|---------|---------|--------|---------------|-------------------------|-------|
| 1. | john | smith | 28 | New York | john.smith@email.com | |
| 2. | jane | doe | 32 | San Francisco | jane_doe@email.com | |
| 3. | alice | johnson | 25 | Los Angeles | alice.johnson@email.com | |
| 4. | bob | brown | 27 | New York | bob.brown@email.com | |
| 5. | charlie | davis | 30 | Chicago | charlie.davis@email.com | |
| 6. | eve | white | 29 | San Francisco | eve.white@email.com | |
| 7. | eve | white | 29 | San Francisco | eve_white@email.com | |

Open... Export Help

OpenRefine project archive to file

Tab-separated value

Comma-separated value

HTML table

Excel (.xls)

Excel 2007+ (.xlsx)

ODF spreadsheet

Custom tabular...

SQL...

Templating...

OpenRefine project archive to Google Drive...

Google Sheets...

Wikibase edits...

QuickStatements file

Wikibase schema