

# **Cleansing the customers file Using Data Preparation**

## **Group Members:**

- 1. Apurva Dhundur**
- 2. Sayli Bavdane**
- 3. Khushabu Gujar**
- 4. Kajal Ahire**
- 5. Vaibhav Nikam**
- 6. Aditya Katkamwar**
- 7. Dhananjay Sonawane**

## **Logging in to Talend Cloud Data Preparation**

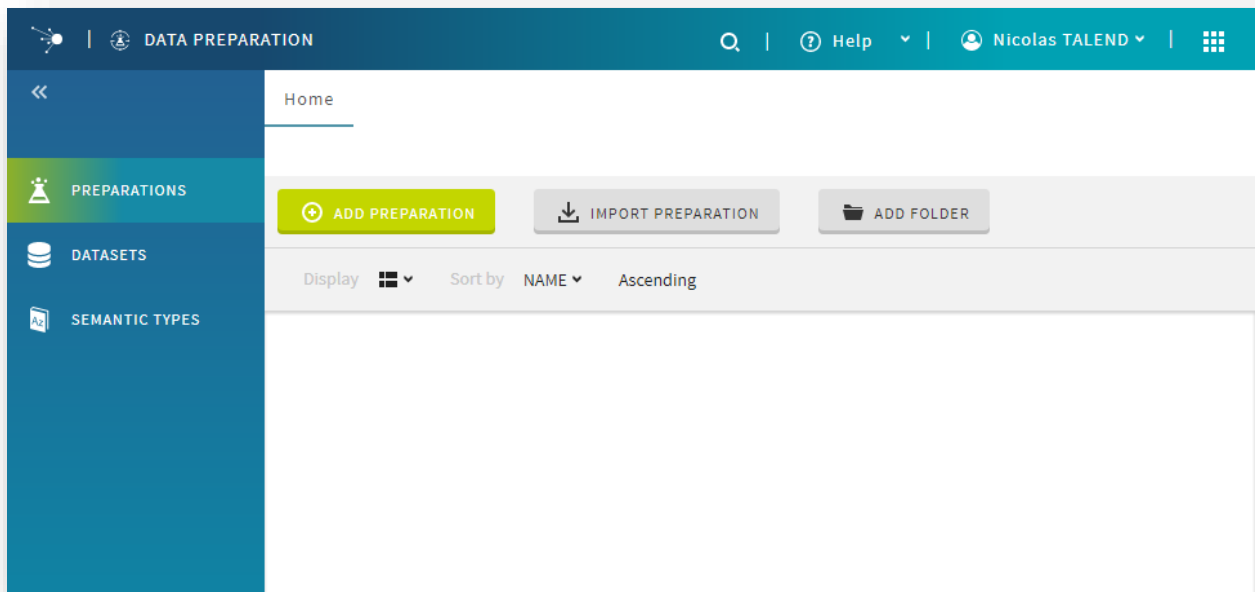
1. Select the website <https://www.talend.com/products/data-preparation/data-preparation-free-desktop>
2. Click on download for windows
3. After downloading the application

## **OPENING A DATASET FROM A LOCAL FILE**

You will now import the file containing the customer data and create your first preparation.

After logging in Talend Data Preparation, you are directed to the **Preparations** view.

This view shows all your preparations, in other words datasets on which you have started performing operations. It is empty for now, but this is where your work on the customer data will be saved. In this view, you can also add new preparations and organize them into folders.



To import the customer file containing the raw data, proceed as follows:

1. Click the **Add preparation** button.

ADD PREPARATION

Existing Datasets Find a Dataset

- Recent Datasets  
10 Last modified datasets
- Favorite Datasets
- Certified Datasets
- All Datasets
- Import File  
Import a local file

Preparation Name  
customers\_preparation

CANCEL CONFIRM

2. In the **Preparation Name** field, enter the name you want to give your preparation, customers\_preparation in this example.
3. Click **Import file**, and select the customers.xlsx file.
4. Click **Open**.

# Result

Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out	Salary_Out	Address
1	James	Butt	F	Under 18	K-12 Student	Single	8	6640 N Blue Gum St
2	Josephine	Darajly	M	56+	Self-Employed	Married	100,000-140,000	4 S Blue Ridge Blvd
3	ART	Venere	M	25-34	Scientist	Married	< 50,000	8 W Carrillos Ave #54
4	Lenna	Paprocki	M	45-49	Executive/Managerial	Divorced	150,000-190,000	639 Main St
5	Dorette	Follzer	M	25-34	Writer	Divorced	50,000-90,000	34 Center St
6	Sinona	Morasca	F	50-55	Homemaker	Married	100,000-140,000	3 McQuay Dr
7	Mitsue	Tallner	M	35-44	Academic/Educator	Divorced	100,000-140,000	7 Eads St
8	Leta	Willingard	M	25-34	Programmer	Divorced	100,000-140,000	7 W Jackson Blvd
9	Sage	Kluser	M	25-34	Technical/Engineer	Divorced	150,000-190,000	5 Boston Ave #88
10	Kristi	Hanner	F	35-44	Academic/Educator	Divorced	< 50,000	228 Runnymede Pl #280
11	Elaine	Waggoner	F	25-34	Academic/Educator	Divorced	150,000-190,000	2071 Jerrilda Ave
12	Abel	MacLean	M	25-34	Programmer	Divorced	< 50,000	37275 St. Rt 176 H
13	Kiley	Calderera	M	45-49	Academic/Educator	Divorced	150,000-190,000	25 E 75th St #63
14	Graciela	Ruta	M	35-44	Other	Divorced	< 50,000	90 Connecticut Ave N
15	Carmy	Albano	M	25-34	Executive/Managerial	Divorced	> 200,000	90 E Monmouth St
16	Hettie	Piquette	F	35-44	Other	Divorced	< 50,000	73 State Road 434 E
17	Hughan	Garrity	M	50-55	Academic/Educator	Divorced	> 200,000	60734 E Carrillo St
18	Sladys	Ron	F	18-24	Clerical/Admin	Divorced	50,000-90,000	322 New Horizon Blvd
19	Nali	Whitney	F	Under 18	K-12 Student	Single	8	1 State Route 27
20	Fletcher	Fiani	M	25-34	Sales/Marketing	Divorced	> 200,000	394 Manchester Blvd
21	Bette	Niska	M	18-24	Self-Employed	Divorced	100,000-190,000	6 S 33rd St
22	Veronika	Shroye	M	18-24	Scientist	Married	< 50,000	6 Greenleaf Ave
23	Willard	Kilmetz	M	35-44	Other	Divorced	< 50,000	618 N Wallace Ave
24	Maryann	Rosier	F	25-34	Executive/Managerial	Divorced	< 50,000	74 S Westgate St
25	Allison	SUDARSKI	M	18-24	College/Grad Student	Single	100,000-140,000	3273 State St
26	Allene	Turnbull	M	25-34	Executive/Managerial	Divorced	< 50,000	1 Central Ave
27	Daniel	Cady	M	25-34	Lawyer	Divorced	150,000-190,000	86 W 88th St #8873
28	Emelin	Choi	F	25-34	Academic/Educator	Divorced	150,000-190,000	2 Cedar Ave #84
29	Willow	Rucko	M	35-44	Executive/Managerial	Single	100,000-140,000	90991 Thornburn Ave

## Cleansing the customers file

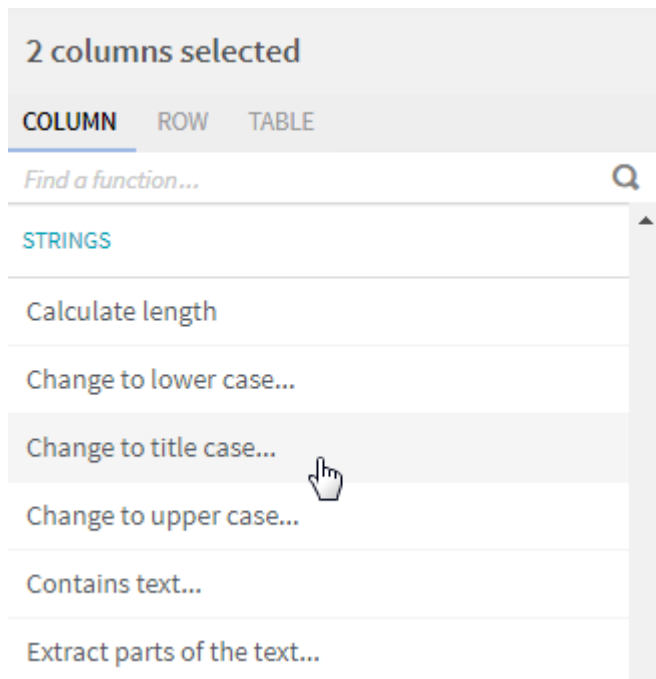
### Harmonizing the case

1. Click the header of the **First\_Name** column to select its content.
2. While pressing the Ctrl key, click the header of the **Last\_Name** column.

The two columns are now selected and you modify them both at the same time.

	First_Name	Last_Name	Gender
integer	first_name	text	
1	James	Butt	F
2	Josephine	Darakjy	M
3	ART	Venere	M
4	Lenna	Paprocki	M
5	Donette	Foller	M
6	Simona	Morasca	F
7	Mitsue	Tollner	M
8	Leota	dilliard	M
9	Sage	Wieser	M
10	kris	Marrier	F

3. In the **Functions panel** located in the upper right side of the screen, find **Change to title case** in the list of functions.



4. Click **Change to title case** to apply the function on the two columns.

All the names now begin with a Capital letter, with the rest in lower case.

integer	First_Name First Name	Last_Name text	Gender
1	James	Butt	F
2	Josephine	Darakjy	M
3	Art	Venere	M
4	Lenna	Paprocki	M
5	Donette	Foller	M
6	Simona	Morasca	F
7	Mitsue	Tollner	M
8	Leota	Dilliard	M
9	Sage	Wieser	M
10	Kris	Marrier	F

## Removing whitespaces

If some whitespaces have been mistakenly introduced in your data, you can apply the **Remove trailing and leading characters** function to clean them.

There is still some work to do in the **First\_Name** column, as well as the **Last\_Name** column. Indeed, you can see white boxes in front or behind some names.

9	Sage	Wie
10	Kris	Mar
11	Minna	Ami

To remove the whitespaces in the cells, proceed as follows:

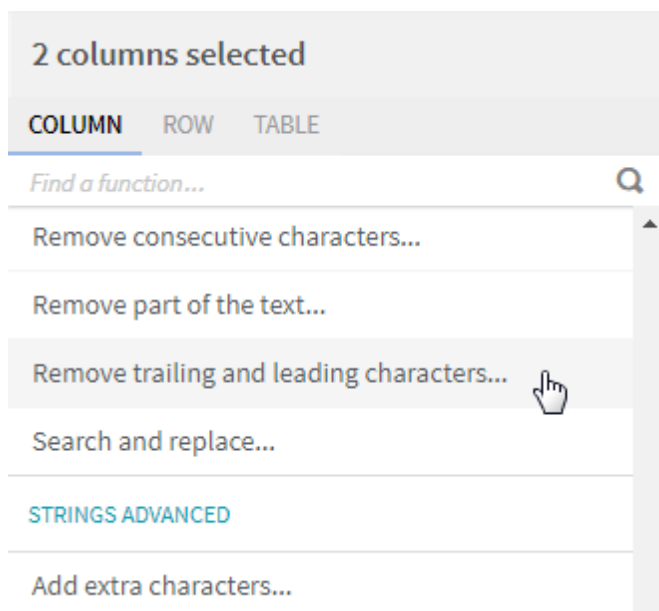
1. Click the header of the **First\_Name** column to select its content.

	First_Name	LAST
integer	First Name	
3	Art	Vene
4	Lenna	Pap
5	Donette	Foll
6	Simona	Mora
7	Mitsue	Toll
8	Leota	idil
9	Sage	Wies
10	Kris	Marr

2. While keeping the Ctrl button pressed, click the header of the **Last\_Name** column.

The two columns are now selected, and you can apply a function to both columns in one action.

3. In the list of functions, click **Remove trailing and leading characters** to open the options for the associated function.



In the **Padding character** drop-down list, select **whitespace** and click **Submit**.

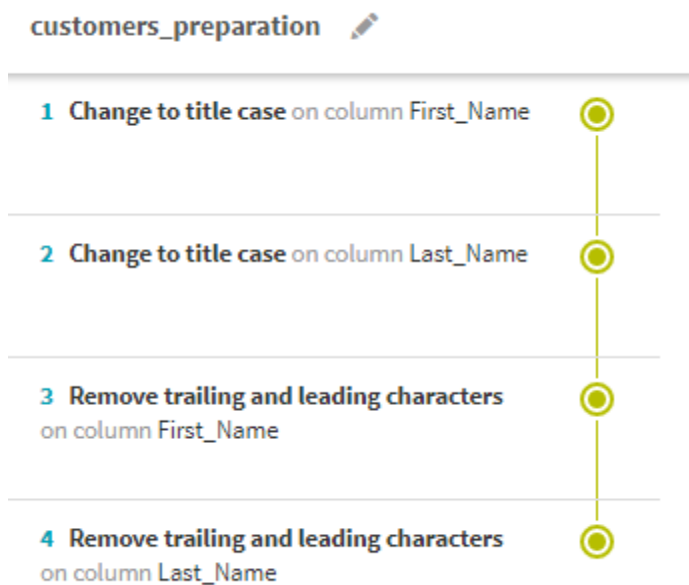
The white boxes have disappeared from the cells in both columns.

9	Sage	Wie
10	Kris	Mar
11	Minna	Ami

## Editing the recipe

The recipe in Talend Cloud Data Preparation, just like any cooking recipe, is the list of preparation steps applied to your data.

After completing four actions on your preparation, you might have noticed that every step was listed on the left side of the screen. This is the recipe of your preparation. Every function that has been applied on your data goes in the recipe.





For the sake of this example, you are going to manipulate the different items that make up your preparation.


To edit your preparation, proceed as follows:


1.To disable a specific recipe line, the third one for example, click the green round button to the right of it.




customers\_preparation 

1 Change to title case on column First\_Name 

2 Change to title case on column Last\_Name 

3 Remove trailing and leading characters on column First\_Name 

4 Remove trailing and leading characters on column Last\_Name 

Filters

Add a filter...

	Id	First_Name	Last_Name
	integer	First Name	text
1	1	James	Butt
2	2	Josephine	Darakjy
3	3	Art	Venere
4	4	Lenna	Paprocki
5	5	Donette	Foller
6	6	Simona	Morasca
7	7	Mitsue	Tollner
8	8	Leota	Dilliard
9	9	Sage	Wieser
10	10	Kris	Marrier
11	11	Minna	Amigon

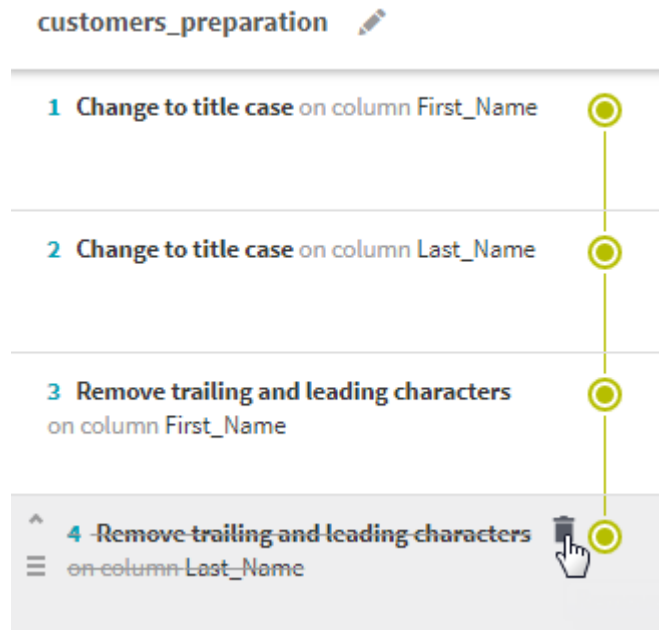
2. Because each preparation step is based on the previous one, disabling one recipe line also disables the following ones.

This operation allows you to look at the state of your data before you applied the function. In this case, you can see that the whitespaces in the **First\_Name** and **Last\_Name** columns can be found again. You can also hover your mouse over the green button for preview.

3. Click the green button next to the fourth recipe line to make the effects of the last two functions active again.

You can use this feature to disable the whole recipe and see your data in its original state. This can be useful if you want to make a before and after comparison of your data.

4. To delete a recipe line, the last one for example, hover over the line and click the trash can icon on the right.



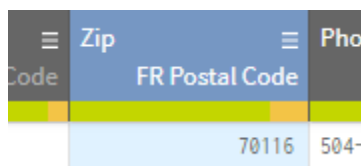
Unlike the disable button you used earlier, the trash can icon completely removes a line from the recipe.

5. Click the undo button on the top right part of the screen.

## Changing the semantic type

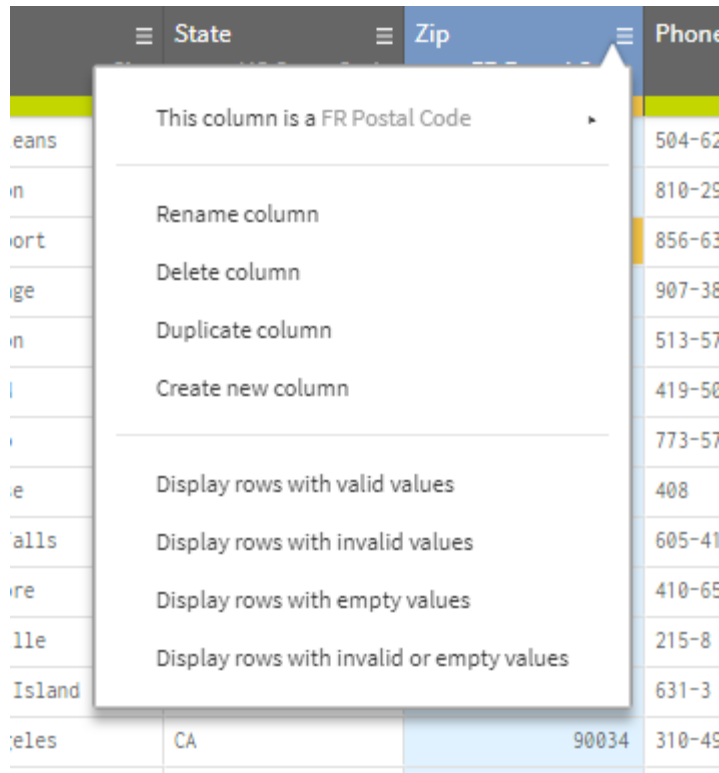
You can change the semantic type of your data to make sure that the data type of your column matches the actual values.

Talend Cloud Data Preparation automatically suggests a semantic type for your data. This type is specified under the header of each column. You can see in the **Zip** column that because they have the same number of digits, the US zip codes have been mistaken for French ones. You will now set the semantic type of the column to US postal code.



To modify the semantic type of a column, proceed as follows:

1. Click the options icon in the header of the **Zip** column.



The image shows a data table with columns: State, Zip, and Phone. The 'Zip' column header is highlighted in blue. A context menu is open over the 'Zip' header, displaying the following options:

- This column is a FR Postal Code
- Rename column
- Delete column
- Duplicate column
- Create new column
- Display rows with valid values
- Display rows with invalid values
- Display rows with empty values
- Display rows with invalid or empty values

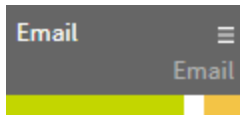
The table data is as follows:

	State	Zip	Phone
ians			504-62
in			810-29
ort			856-63
ge			907-38
in			513-57
			419-50
			773-57
e			408
alls			605-41
re			410-65
lle			215-8
Island			631-3
eles	CA	90034	310-45

2. In the drop-down menu, point your mouse over **this column is a fr\_postal\_code**.



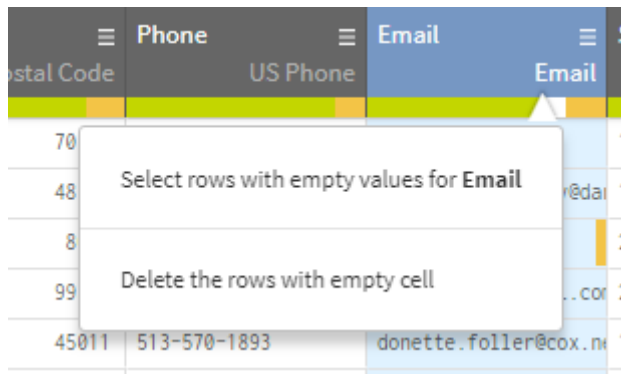
By looking at the quality bar under in the **Email** column header, you can see that there are empty cells and incorrect values among the data. You are going to remove them.



To use the quality bar to remove the lines containing those incorrect cells, proceed as follows:

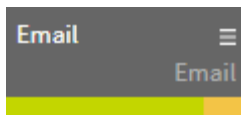
1. Click the white part of the quality bar, in the header of the **Email** column.

A drop-down menu opens.



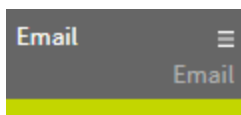
2. Click **Delete the rows with empty cells**.

The empty cells of the **Email** columns have been deleted and only the invalid values, represented by the orange bar, remain.



3. Repeat the last two steps, but this time, click the orange part of the quality bar, and select **Delete the rows with invalid cells**.

The **Email** column is now cleaned of all invalid data or empty cells.



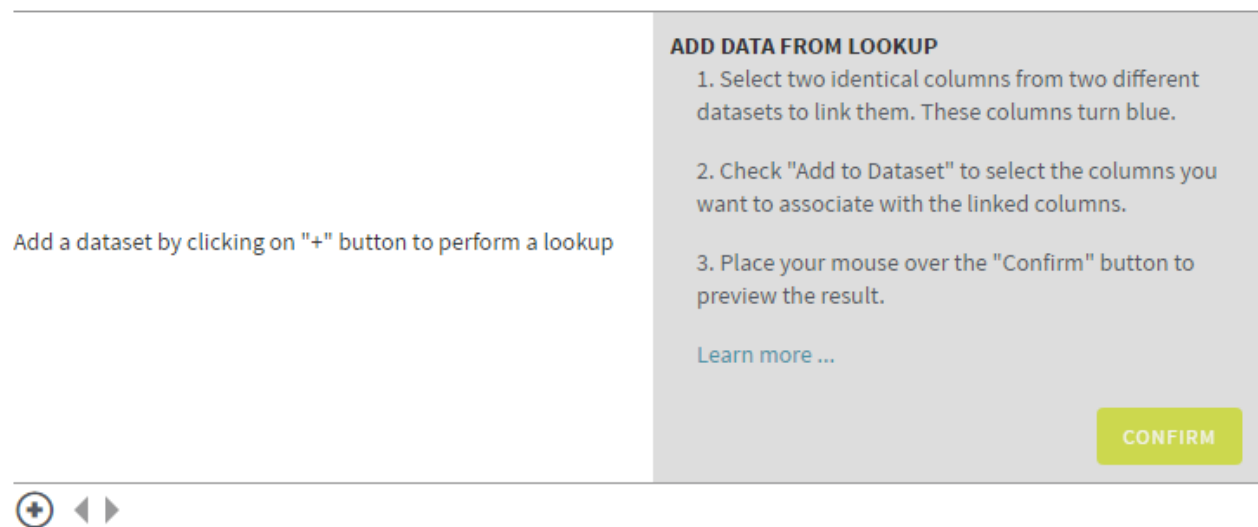
4. Use the quality bar to remove the invalid cells from the **Zip** and **Phone** columns.

## Blending data

1. Click the header of the **State** column to select its content.
2. Click the **Lookup** icon in the upper part of the screen.



The **Add data from lookup** panel opens at the bottom of the screen.



3. Click the + icon to select the dataset you want to add.

The list of previously imported datasets opens. In your case, only States is available.

4. Select the check box next to **States** and then click **Add**.

The States dataset opens in the bottom part of the screen. You can see that it is only made of two columns, including **State** that can also be found in your current preparation.

5. Select the **State** column in both your preparation and the dataset, so that they appear in blue.

Your preparation and the dataset can only be linked together if they have a column with information in common, the US State codes in this case

	MaritalStatus_Out	Salary_Out	Address	City	State	Zip	Phone	Email	SubDate	
	text	text	Address Line	City	US State Code	US Postal Code	US Phone	Email	date	
1	Single	0	6649 N Blue Gum St	New Orleans	LA	70116	504-621-8927	jbutt@gmail.com	17-Mar-2016	
2	Married	100,000-149,999	4 B Blue Ridge Blvd	Brighton	MI	48116	810-292-9388	josephine_darakjy@da	15-Mar-2013	
4	Divorced	150,000-199,999	639 Main St	Anchorage	AK	99501	907-385-4412	lpaprocki@hotmail.co	24-Nov-2013	
5		50,000-99,999	34 Center St	Hamilton	OH	45011	513-570-1893	donette_foller@cox.n	17-Apr-2012	
6	Married	100,000-149,999	3 McAuley Dr	Ashland	OH	44805	419-503-2484	simona@morasca.com	13-Apr-2016	
9	Divorced	150,000-199,999	5 Boston Ave #88	Sioux Falls	SD	57105	605-414-2147	sage_wieser@cox.net	20-Apr-2013	
10	Divorced	< 50,000	228 Runamuck Pl #280	Baltimore	MD	21224	410-655-8723	kris@gmail.com	31-Dec-2012	
13		150,000-199,999	25 E 75th St #69	Los Angeles	CA	90034	310-498-5651	kiley.caldarera@aol.	08-Mar-2014	
14	Divorced	< 50,000	98 Connecticut Ave N	Chagrin Falls	OH	44023	440-780-8425	gruta@cox.net	12-Jun-2013	
15		> 200,000	56 E Morehead St	Laredo	Texas	78045	956-537-6195	calbares@gmail.com	25-Jun-2011	
16		< 50,000	73 State Road 434 E	Phoenix	AZ	85013	602-277-4385	mattie@aol.com	01-Dec-2009	
17	Divorced	> 200,000	69734 E Carrillo St	Mc Minnville	TN	37110	931-313-9635	meaghan@hotmail.com	26-Dec-2015	
18		50,000-99,999	322 New Horizon Blvd	Milwaukee	WI	53207	414-661-9598	gladys.rimbrin.org	08-Sep-2011	
19	Single	0	1 State Route 27	Taylor	MI	48180	313-288-7937	yuki_whobrey@aol.com	17-Sep-2012	

State Code

US State Code

City

1	WA	West
2	MT	West
3	OR	West
4	ID	West
5	WY	West
6	CA	West
7	NV	West
8	UT	West

Region

City

☒ Add to Dataset

West

West

West

ADD DATA FROM LOOKUP

1. Select two identical columns from two different datasets to link them. These columns turn blue.  
 2. Check "Add to Dataset" to select the columns you want to associate with the linked columns.  
 3. Place your mouse over the "Confirm" button to preview the result.

Learn more...

CONFIRM

In the States dataset, select the check box **Add to Dataset** under the **Region** column header to add it to your current preparation

Region

City

☒ Add to Dataset

West

West

West

Point your mouse over the **Confirm** button to preview the changes

	MaritalStatus_Out	Salary_Out	Address	City	State	Region	Zip	Phone	Email
	text	text	Address Line	City	US State Code	text	US Postal Code	US Phone	Email
1	Single	0	6649 N Blue Gum St	New Orleans	LA	South East	70116	504-621-8927	jbutt@gmail.com
2	Married	100,000-149,999	4 B Blue Ridge Blvd	Brighton	MI	Mid West	48116	810-292-9388	josephine_darakjy@da
4	Divorced	150,000-199,999	639 Main St	Anchorage	AK	West	99501	907-385-4412	lpaprocki@hotmail.co
5		50,000-99,999	34 Center St	Hamilton	OH	Mid West	45011	513-570-1893	donette_foller@cox.n
6	Married	100,000-149,999	3 Mcauley Dr	Ashland	OH	Mid West	44805	419-503-2484	simona@morasca.com
9	Divorced	150,000-199,999	5 Boston Ave #88	Sioux Falls	SD	Mid West	57105	605-414-2147	sage_wieser@cox.net
10	Divorced	< 50,000	228 Runamuck Pl #280	Baltimore	MD	North East	21224	410-655-8723	kris@gmail.com
13		150,000-199,999	25 E 75th St #69	Los Angeles	CA	West	90034	310-498-5651	kiley.caldarera@aol.
14	Divorced	< 50,000	98 Connecticut Ave N	Chagrin Falls	OH	Mid West	44023	440-780-8425	gruta@cox.net
15		> 200,000	56 E Morehead St	Laredo	Texas		78045	956-537-6195	calbares@gmail.com
16		< 50,000	73 State Road 434 E	Phoenix	AZ	South West	85013	602-277-4385	mattie@aol.com
17	Divorced	> 200,000	69734 E Carrillo St	Mc Minnville	TN	South East	37110	931-313-9635	meaghan@hotmail.com
18		50,000-99,999	322 New Horizon Blvd	Milwaukee	WI	Mid West	53207	414-661-9598	gladys_rim@rim.org
19	Single	0	1 State Route 27	Taylor	MI	Mid West	48180	313-288-7937	yuki_whobrey@aol.com
20	Divorced	> 200,000	394 Manchester Blvd	Rockford	IL	Mid West	61109	815-828-2147	fletcher_flosi@yahoo
21		150,000-199,999	6 S 33rd St	Aston	PA	North East	19014	610-545-3615	bette_nicka@cox.net
22	Married	< 50,000	6 Greenleaf Ave	San Jose	CA	West	95111	408-540-1785	vinouye@aol.com

	State Code	Region
	US State Code	City
1	WA	West
2	MT	West
3	OR	West
4	ID	West
5	WY	West
6	CA	West
7	NV	West
8	UT	West

**ADD DATA FROM LOOKUP**

1. Select two identical columns from two different datasets to link them. These columns turn blue.
2. Check "Add to Dataset" to select the columns you want to associate with the linked columns.
3. Place your mouse over the "Confirm" button to preview the result.

[Learn more...](#)

**CONFIRM**

Click the **Confirm** button to apply the changes and add the **Region** column to your preparation

# Applying a value to all cells

Applying a certain value to many cells at once can save you a lot of time when correcting invalid cells.

The **State** column is the last column containing incorrect data. This column lists the States from which the customers have rented a movie, using a two-letter code. You can notice that among all the other US state codes, the occurrences of **Texas** stand out as errors

CA
OH
Texas
AZ



Rather than simply deleting the corresponding lines with the quality bar like you did before, you are going to correct one of the invalid cells, and apply the new value to all the cells with the same error. To replace the occurrences of **Texas** with the correct value, proceed as follows:

1. In the **State** column, double-click one of the occurrences of **Texas**.

You can now edit the content of the cell, and a menu with a check box opens.

City	State US State Code	Region	City	Zip
	OH	Mid West		
	OH	Mid West		
	SD	Mid West		
	MD	North East		
	CA	West		
	OH	Mid West		
	Texas			
	<input type="checkbox"/> Apply to all cells with this value			
	WI	Mid West		
	MI	Mid West		

1. Instead of **Texas**, type TX, which is the correct two-letter code.
2. Select the check box **Apply to all cells with this value**.

City	State US State Code	Region	City	Zip
	OH	Mid West		
	OH	Mid West		
	SD	Mid West		
	MD	North East		
	CA	West		
	OH	Mid West		
	TX			
	WI	Mid West		
	MI	Mid West		

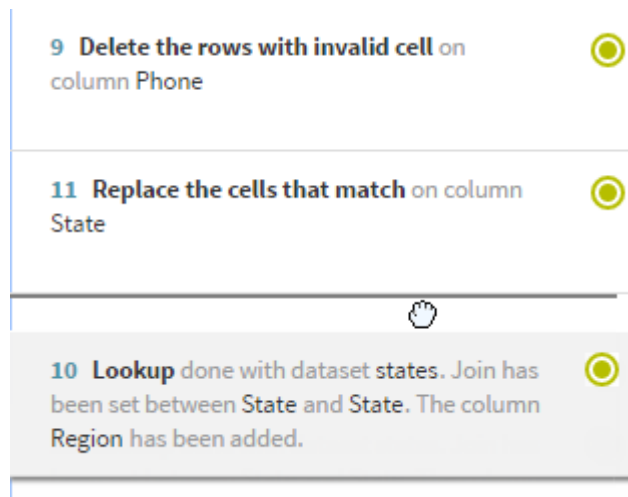
Press the **Enter** key

## Reordering preparation steps

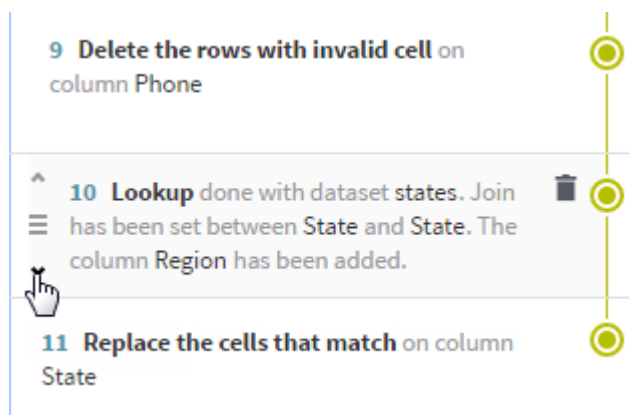
1.Point your mouse over the lookup step.

2.To move the lookup step from the second-last position to the last position, you can:

- Either drag the recipe step and drop it at the bottom of your recipe.



- the grey line shows where the recipe will be placed.
- Or click the up arrow on the left of your recipe step to move it down.

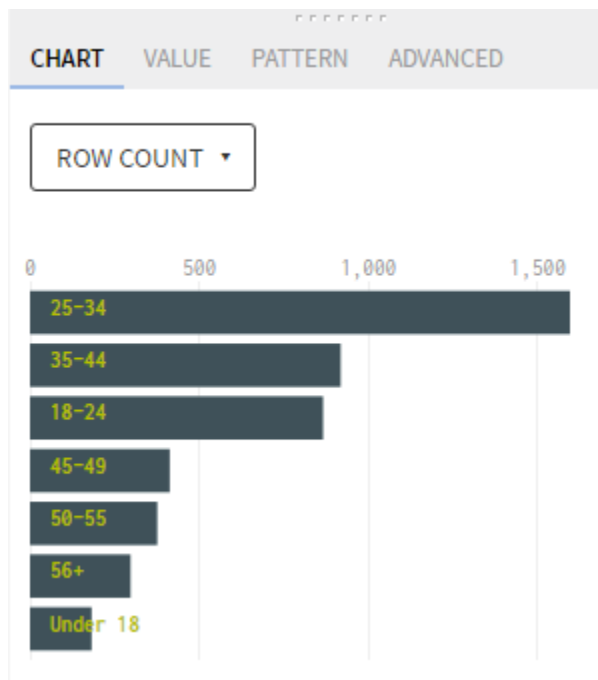


## Using charts to filter

1. Click the header of the **Age** column to select its content.

Age	Occ
Under 18	K-12
56+	Self
45-49	Exec
25-34	Writ

The graphical representation of the column's content is displayed in the form of an horizontal bar chart, on the bottom right side of the screen. Each bar represents the number of occurrences of an age group. Hovering over each bar displays information about the data.



In the chart, click the bar labeled **Under 18**

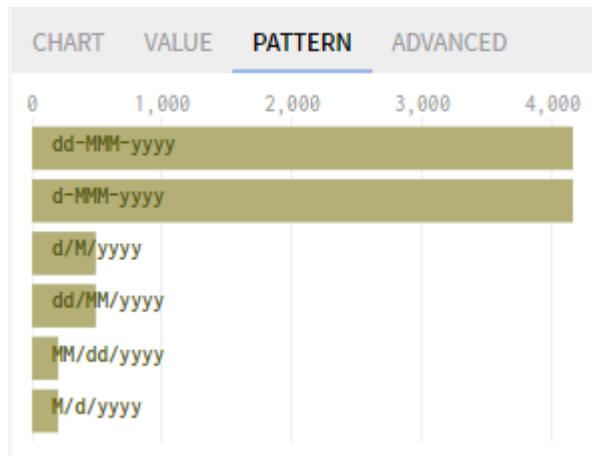
1. To clear the filter, simply click the **x** icon, on the right of the filter.

## Harmonizing the date format

1. Click the header of the **SubDate** column to select its content

mail	SubDate	date	NU
	17-Mar-2016		
	15-Mar-2013		
	24-Nov-2013		

2. In the statistics box on the bottom right, click **Pattern**.



3. To standardize the date format, click **Change Date Format...** in the functions list.

The screenshot shows a web interface for a date function. At the top is a header 'SubDate' with tabs for 'COLUMN', 'ROW', and 'TABLE'. Below the tabs is a search bar with the placeholder text 'Find a function...'. A dropdown menu is open, titled 'Change date format...'. Inside the menu, there is a checkbox labeled 'Create new column'. Below that, there are two dropdown menus: 'Current format:' with the selected option 'I don't know, best guess', and 'New format:' with the selected option 'ISO 8601 date'. At the bottom of the menu is a yellow 'SUBMIT' button and a link that says 'Learn more...'.

1. A menu opens, where you can specify the current date formats, and the desired one.
2. In the **Current format** drop-down list, leave **I don't know, best guess** selected.
3. In the **New format** drop-down list, select **custom**.
4. In the **Your format** field, type dd-MMM-yyyy.

SubDate

COLUMN ROW TABLE

Find a function...

Change date format...

☐ Create new column

Current format:  
I don't know, best guess

New format:  
Other

Your format:  
dd-MMM-yyyy

SUBMIT

[Learn more...](#)

The dd-MMM-yyyy format is the most suited since it is the one that already had the most occurrences.

## Finding and grouping similar content

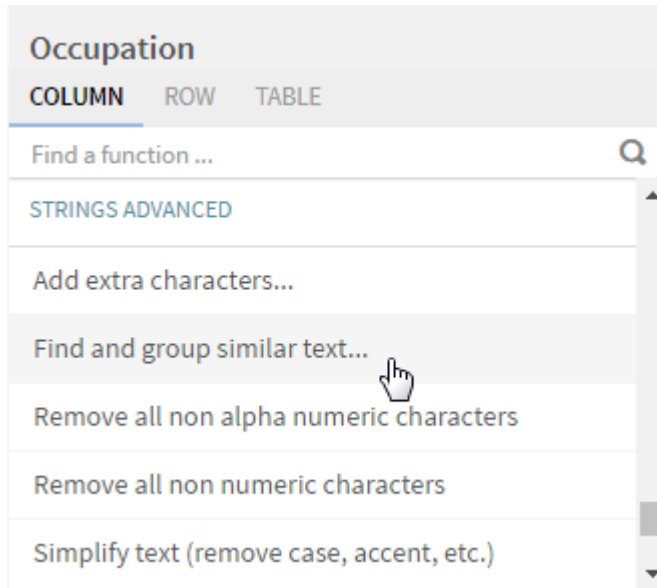
1. Click the header of the **Occupation** column to select its content.

≡ text	Occupation text	≡ text
	K-12 Student	Singl
	Self-Employed	Marri
	Executive/Managerial	Divor

You can confirm in the statistics box that there are occurrences of job titles that only slightly differ.

2. In the functions list, select **Find and Group Similar Text....**

The **Find and group similar text** menu opens.



All similar occupations are grouped together in the second column. In this case, **College/Grad Student** and **College Student**. The third column suggests an occupation title that could replace the values in the second column. You can choose another value from the drop-down list, or type a whole new one. Clear the check boxes in front of the values or groups of values you want to leave unchanged.

3. In the drop-down list of the third column, select **College Student**.



## FIND AND GROUP SIMILAR TEXT



Replace all similar values with the right one (i.e. cluster on fuzzy matching)

<input checked="" type="checkbox"/>	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> College Student	Replace value: <b>College/Grad Student</b> ▼
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> College/Grad Student	

SUBMIT

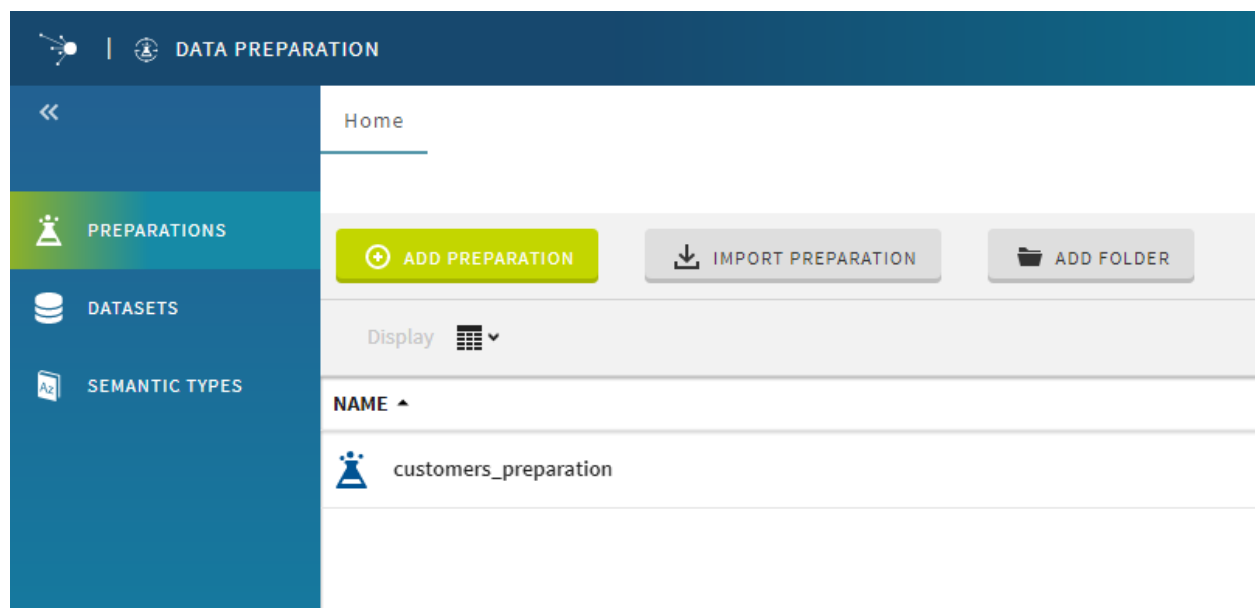
4. Click **Submit**.

## Sharing a preparation

1. Click the white **X** icon on the top right of the screen to exit your preparation.

Remember that your preparation is automatically saved after each step.

You are now in the **Preparations** view, where you can see customers\_preparation in the list.

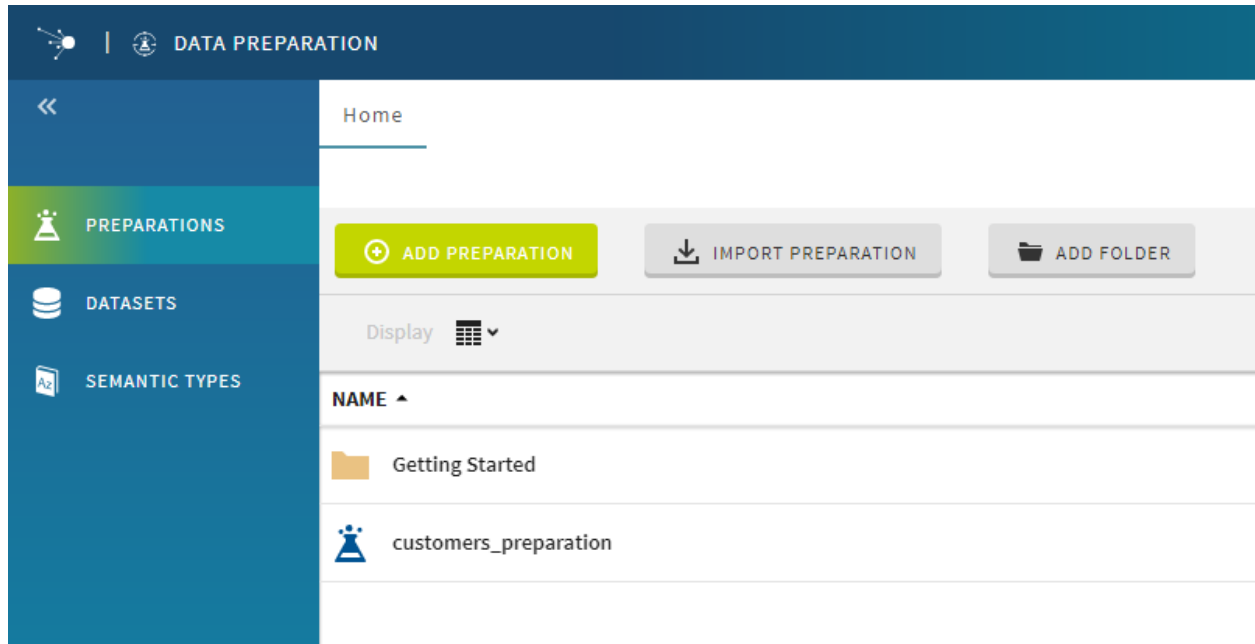


1. Click the **Add folder** icon.

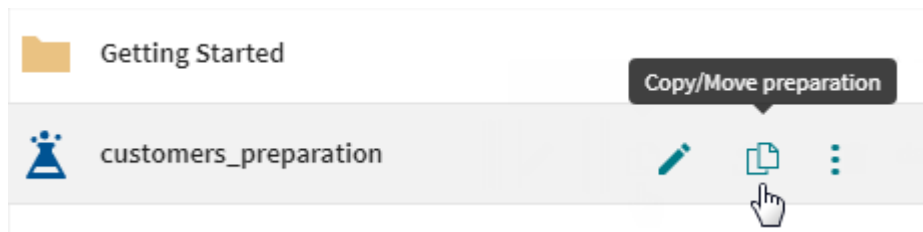
A window opens where you need to enter a name for your folder.

2. Type Getting Started in the empty field and click **OK**.

The Getting Started folder now appears in the list in the **Preparations** view.

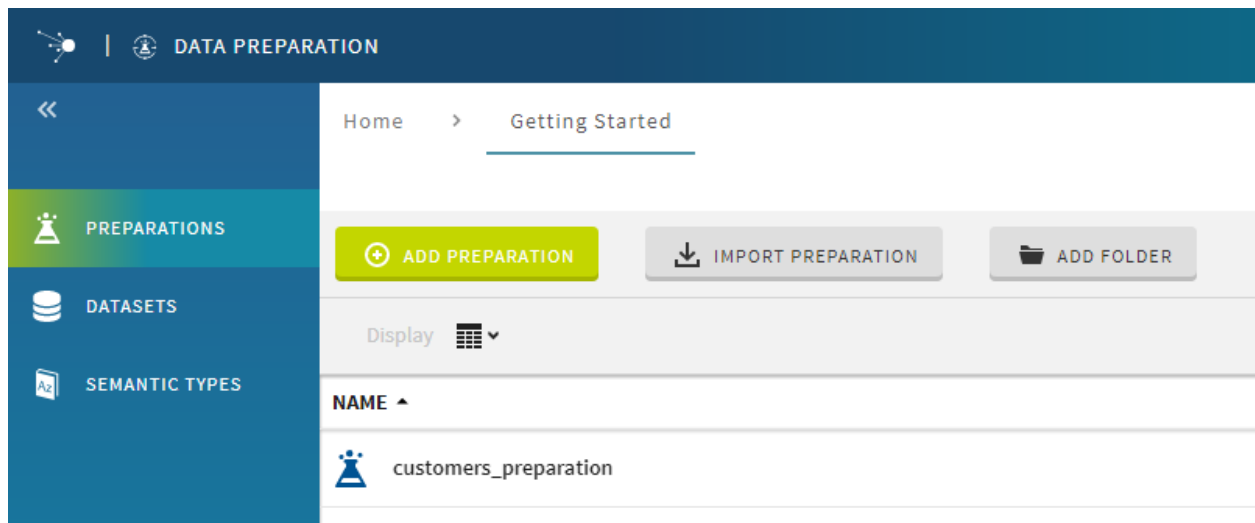


Point your mouse over customers\_preparation in order to display the available options and click the **Copy or Move** icon.



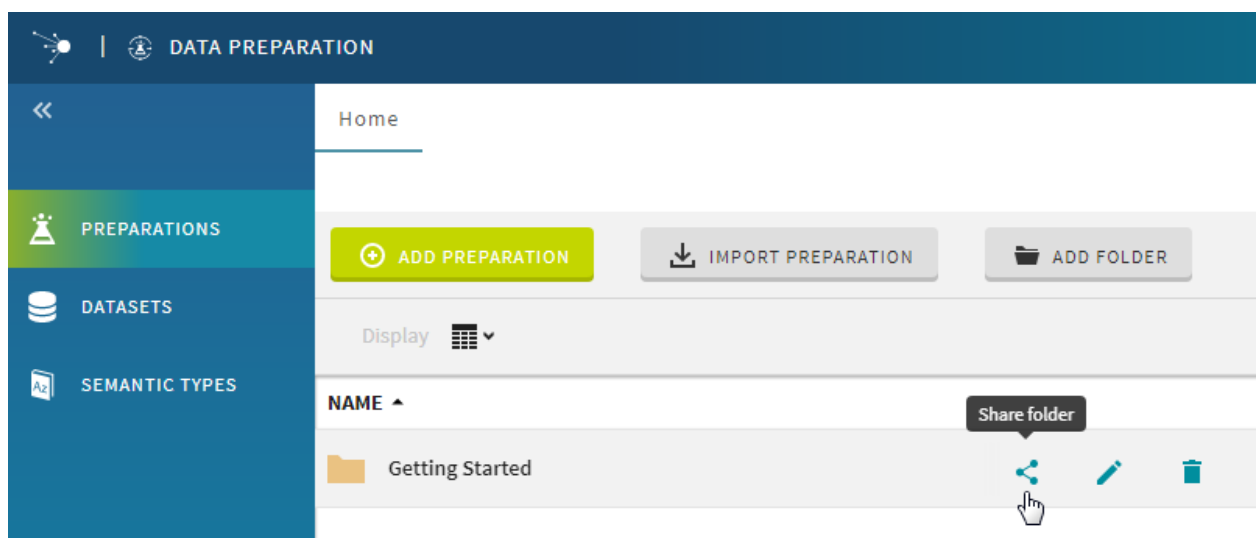
- 1.
2. The **Copy/Move item** window opens, where you can select the destination folder for your preparation.
3. Choose the Getting Started folder and click **Move**.

Your preparation is now located in the Getting Started folder, as you can see in the path above your preparation.



6. Click **Home** in the path to go back to the **Preparations** view.

7. Point your mouse over the Getting Started folder in order to display the available options and click the **Share folder** icon.



8. The **Share content for folder** window opens.

9. Browse the **All Users and Groups** list or use the **Find user/group** search bar to select a user or group that is part of your project.

10. Click a user or group and click **Add to List** to add them to the list of contributors.

## SHARE CONTENT FOR FOLDER: **GETTING STARTED**



### Current Collaborators

Find a collaborator 



Nicolas Talend  
Owner

← Add to List

### All Users and Groups

Find user/group 

QA  
back  
doc  
front

Abdelaziz Talend  
Boubacar Talend  
Charles Talend  
Christophe Talend

CANCEL

CONFIRM

1. Repeat the last step to add more contributors if necessary and click **Confirm**.