



talend

# Talend Data Preparation Getting Started Guide

8.0

Last updated: 2022-06-27

# Contents

<b>Copyright.....</b>	<b>3</b>
<b>Talend Data Preparation Getting Started Guide.....</b>	<b>4</b>
<b>Reading client data from an Excel file.....</b>	<b>5</b>
<b>Logging in to Talend Data Preparation.....</b>	<b>6</b>
<b>Opening a dataset from a local file.....</b>	<b>8</b>
<b>Cleansing the customers file.....</b>	<b>10</b>
Harmonizing the case.....	10
Removing whitespaces.....	11
Editing the recipe.....	12
Changing the semantic type.....	14
Working with the quality bar.....	16
Blending data.....	16
Applying a value to all cells.....	19
Reordering a preparation step.....	21
Using charts to filter.....	22
Using charts to calculate absolute value.....	23
setting a unique date format.....	26
Finding and grouping similar content.....	29
<b>Sharing the finalized preparation.....</b>	<b>31</b>
<b>Exporting the result of your preparation.....</b>	<b>34</b>
<b>What's next?.....</b>	<b>36</b>

# Copyright

Adapted for 7.3.1. Supersedes previous releases.

Copyright © 2021 Talend. All rights reserved.

The content of this document is correct at the time of publication.

However, more recent updates may be available in the online version that can be found on [Talend Help Center](#).

## Notices

All brands, product names, company names, trademarks and service marks are the properties of their respective owners.

## End User License Agreement

The software described in this documentation is provided under **Talend**'s End User Software and Subscription Agreement ("Agreement") for commercial products. By using the software, you are considered to have fully understood and unconditionally accepted all the terms and conditions of the Agreement.

To read the Agreement now, visit [http://www.talend.com/legal-terms/us-eula?utm\\_medium=help&utm\\_source=help\\_content](http://www.talend.com/legal-terms/us-eula?utm_medium=help&utm_source=help_content).

# Talend Data Preparation Getting Started Guide

Talend Data Preparation is a self-service application that enables information workers to cut hours out of their work day by simplifying and expediting the laborious and time-consuming process of preparing data for analysis or other data-driven tasks.

Talend Data Preparation runs on top of the Talend Integration Platform and delivers enterprise-class capabilities together with connectivity to virtually any data source. It fosters collaboration between business people who know the data best and central organizations, like IT or Risk Management, that define the rules and policies for data accessibility and governance.

It includes:

- Integration and cataloging.
- Data discovery and profiling.
- Cleansing, standardizing and shaping.
- Enriching and connecting datasets.
- Operationalizing Talend Data Preparation.

# Reading client data from an Excel file

Imagine your company is a provider of movie rental and streaming video services.

In the examples described in this chapter you have to cleanse a file containing data about your customers, such as their names, age, contact details, subscription dates or numbers of rentals. You want to make this data clearer and easier to work with.

The examples provided in this chapter assume that:

- You have installed Talend Data Preparation. For more information, see the Talend Installation Guide.
- You are registered in Talend Administration Center as a Talend Data Preparation user.

This scenario describes a variety of actions that you can perform on your client file to remove or correct inconsistencies and improve the quality of your data.

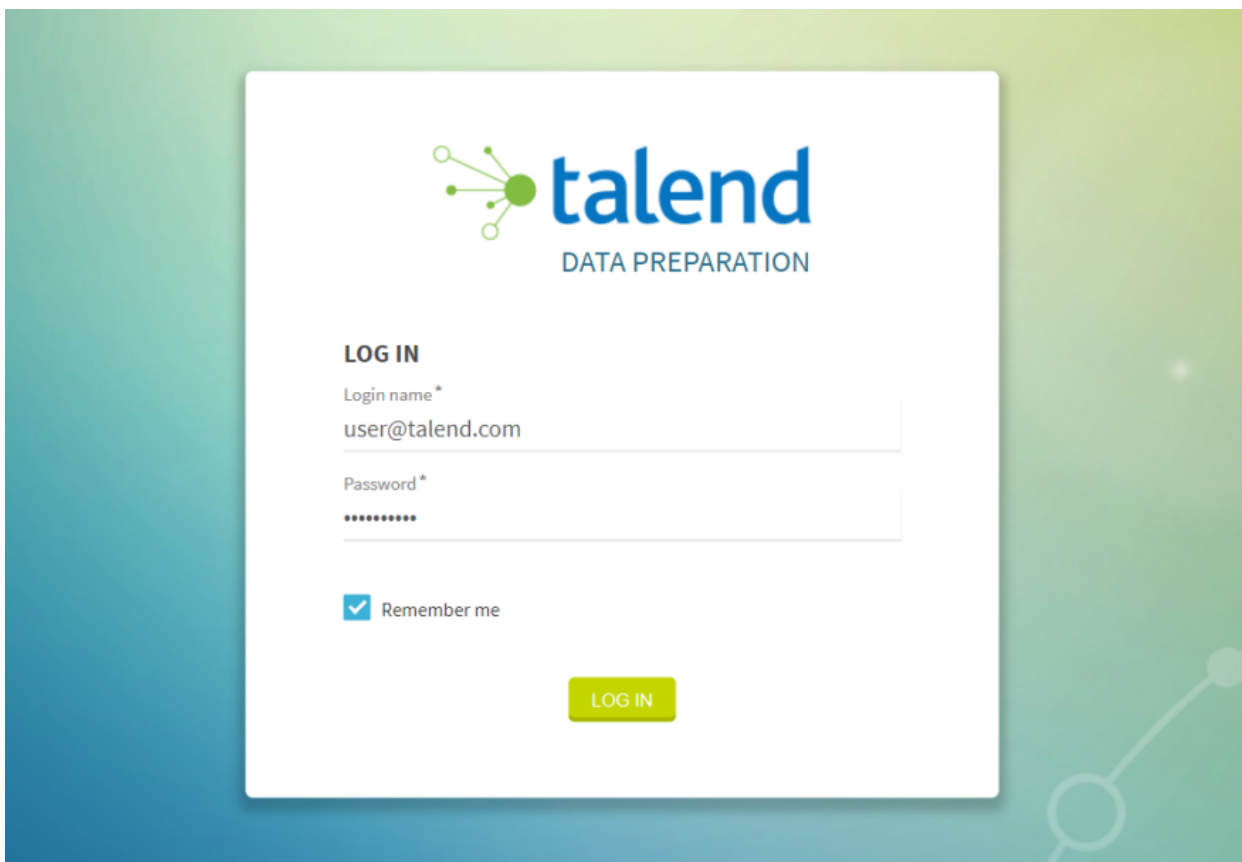
# Logging in to Talend Data Preparation

The first step on the way to using Talend Data Preparation is to access and log in to the web application.

To log in Talend Data Preparation and start using it, proceed as follows:

## Procedure

1. Contact your admin to get the URL to Talend Data Preparation. By default, the URL to Talend Data Preparation is `http://<host>:9999`, where <host> is the host of the machine on which Talend Data Preparation is installed.
2. Follow this link to access the Talend Data Preparation homepage to log in.
3. Type your user email address and user password in the corresponding fields.

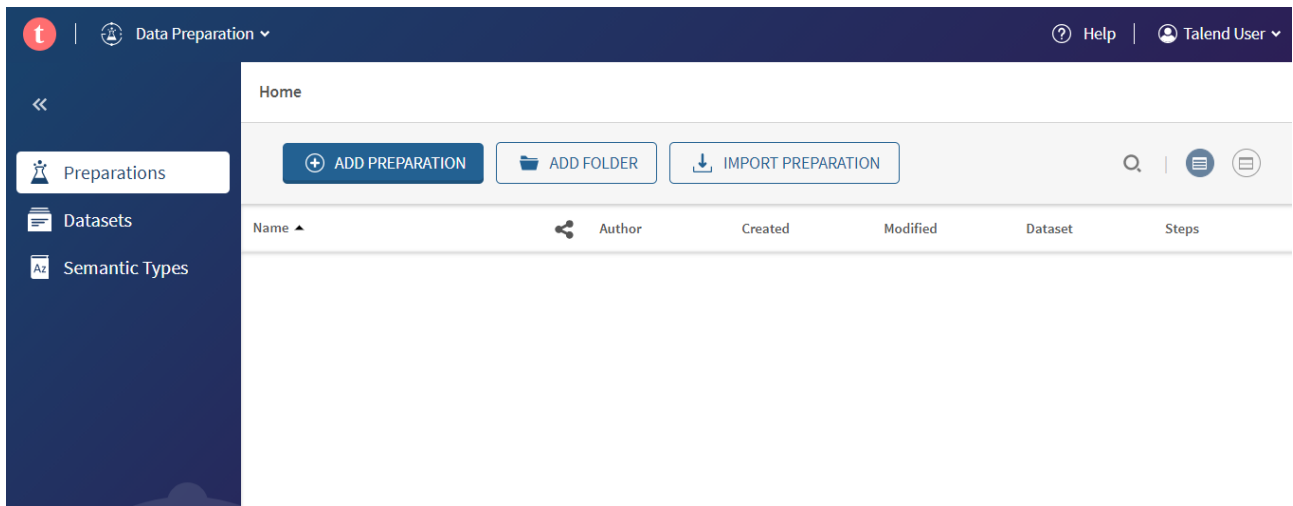


If you entered a wrong value for one of the fields, an error message is displayed.

4. Click **Log in**

## Results

You have reached the Talend Data Preparation home page

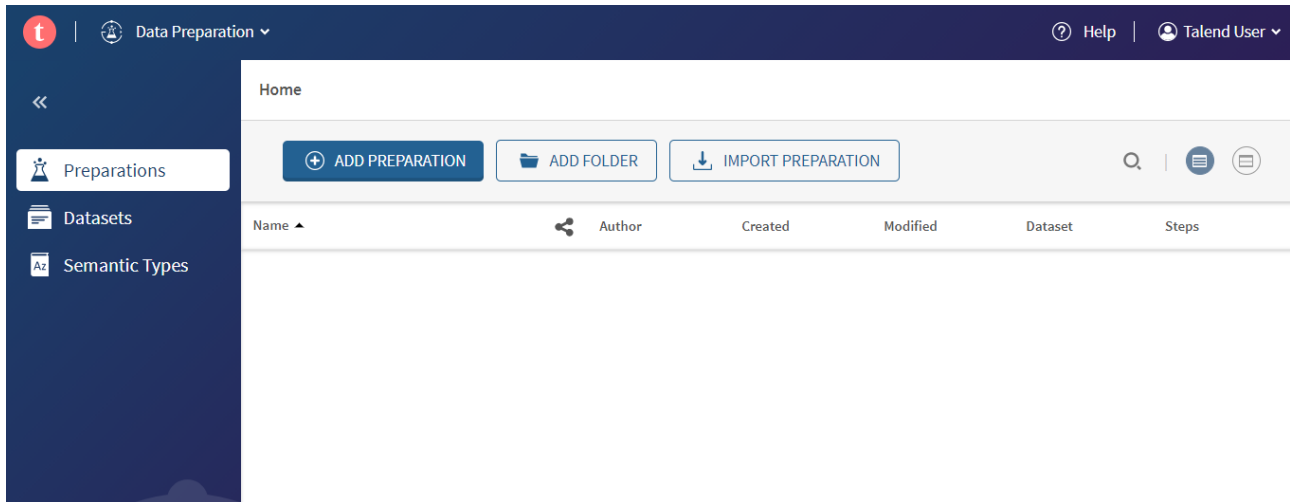


# Opening a dataset from a local file

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:

## Before you begin

Retrieve the `customers.xlsx` file from the **Downloads** tab of the documentation page.

## Procedure

1. From the left panel menu, select **Datasets** to open the list of datasets.
2. To import the `customers.xlsx` dataset that you have previously downloaded, click the **Add dataset** button and browse your files to open it.
3. Click the **Go back** arrow on the top left to return to the dataset list.
4. From the left panel menu, select **Preparations**.
5. Click the **Add preparation** button.



**Add a preparation**
✕

Preparation name\*

customers\_preparation

Dataset\*

Select an existing dataset

**customers**

owned by Talend User, 1 minute ago

CANCEL
SUBMIT

6. In the **Preparation Name** field, enter the name you want to give your preparation, `customers_preparation` in this example.
7. Click **Import file**, and select the `customers.xlsx` file.
8. Click **Open**.

### Results

Your dataset opens in the form of a preparation with an empty recipe. Your data has not been modified yet, but has been saved as a preparation on which you can start applying preparation steps.

Because you imported the `customers.xlsx` dataset, and created the corresponding preparation using the **Add preparation** button, every change made to `customers_preparation` will be automatically saved. As for the raw dataset you imported, it can be viewed in the **Datasets** view, and the data remains unchanged.

DATA PREPARATION
Help | Message | Settings

**customers\_preparation**

Dataset: customers

EXPORT

Filters

Add a filter...

	Id	First_Name	Last_Name	Gender	Age	Occupation	MaritalStatus_Out
	Integer	First Name	text	Gender	text	text	text
1	1	James	Butt	F	Under 18	K-12 Student	Single
2	2	Josephine	Darakjy	M	56+	Self-Employed	Married
3	3	ART	Venere	M	25-34	Scientist	Married
4	4	Lenna	Paprocki	M	45-49	Executive/Manager...	Divorced
5	5	Donette	Foller	M	25-34	Writer	
6	6	Simona	Horasca	F	50-55	Homemaker	Married
7	7	Mitsue	Tollner	M	35-44	Academic/Educator	Divorced
8	8	Leota	dilliard	M	25-34	Programmer	
9	9	Sage	Wieser	M	25-34	Technical/Engineer	Divorced
10	10	kris	Marrier	F	35-44	Academic/Educator	Divorced
11	11	minna	Anigon	F	25-34	Academic/Educator	Divorced
12	12	Abel	Maclead	M	25-34	Programmer	Divorced
13	13	Kiley	Caldarera	M	45-49	Academic/Educator	
14	14	Graciela	Ruta	M	35-44	Other	Divorced
15	15	Camy	Albares	M	25-34	Executive/Manager...	
16	16	Mattie	Poquette	F	35-44	Other	
17	17	Meaghan	Garufi	M	50-55	Academic/Educator	Divorced

Id

COLUMN ROW TABLE

Find a function...

SUGGESTIONS

Add, multiply, subtract or divide...

Compare numbers...

BOOLEAN

Negate value...

COLUMNS

Concatenate with...

Delete column

CHART VALUE PATTERN ADVANCED

Row count \*

# Cleansing the customers file

Now that the working environment is set up, you can start working on your data to cleanse it and improve it.

The examples provided in this chapter assume that:

- You have retrieved the `customers.xlsx` file from the **Downloads** tab of the documentation page.
- You have opened the `customers_preparation` preparation previously created.

This scenario describes various cleansing and formatting operations that you can perform on the customers data.

## Harmonizing the case

The first action will be to fix the column listing the first names of the customers.

You can notice that in the `customers.xlsx` file, there are some inconsistencies among the names in the **First\_Name** and **Last\_Name** columns. Some begin with a capital letter, some are entirely in lower case, and others are entirely in upper case.

To harmonize the style of all the cells of a column, proceed as follows:

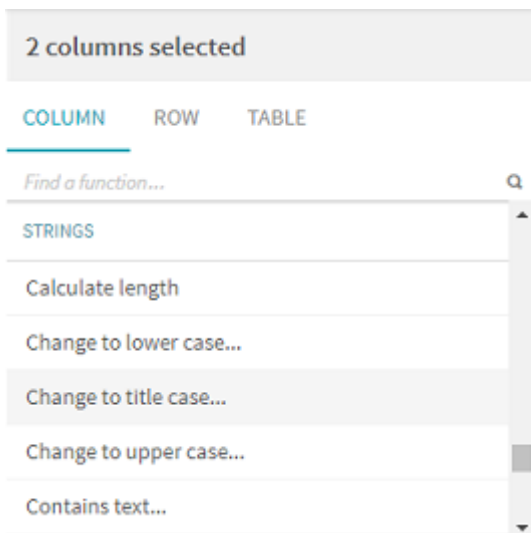
### Procedure

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 First Name	Last_Name text	Ger Gen
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.



This box lists all the functions available to apply on your data. Search for the function you want, or choose among the suggested ones. Hover over a function in the list to preview the effect it will have on your data. Click the function to apply the changes.

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.

	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 Whitespaces** 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.

10	Kris	Marrier
----	------	---------

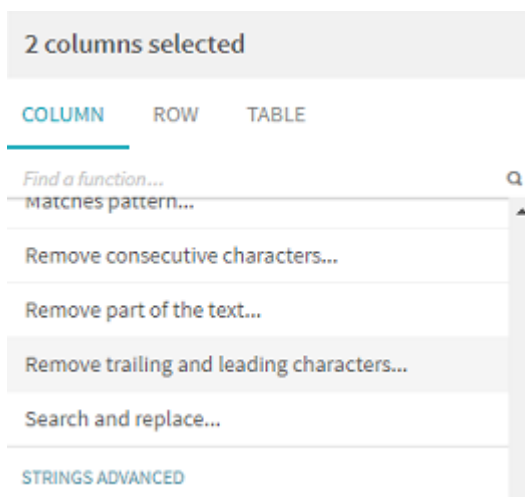
To remove the whitespaces in the cells, proceed as follows:

## Procedure

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

	First_Name First Name	Last_Name Last Name
3	Art	Ve
4	Lenna	Pa
5	Donette	Fo
6	Simona	Mo
7	Mitsue	To
8	Leota	Di
9	Sage	Wi
10	Kris	Ma

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.



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

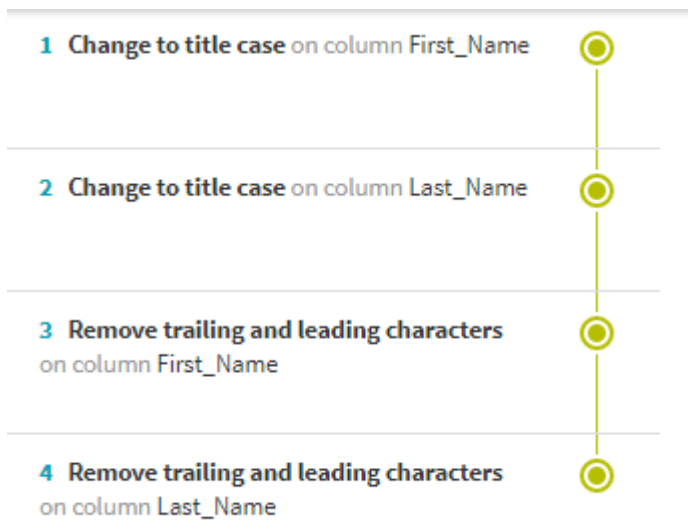
The white boxes have disappeared from the cells in both columns.

9	Sage	Wies
10	Kris	Marr
11	Minna	Amie

## Editing the recipe

The recipe in Talend 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:

#### Procedure

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

	Id integer	First_Name First Name	Last_Name text
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

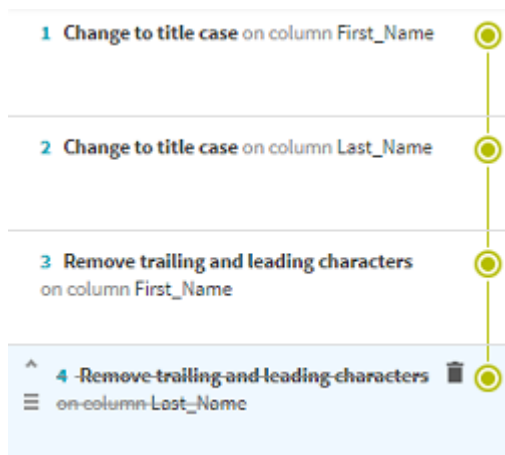
Because each preparation step is based on the previous one, disabling one recipe line also disables the following ones.

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.

2. 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.

3. 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.

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



### Results

Your preparation is now back at the state it was at the beginning of this procedure. Remember that every modification can be undone. You have now learned how editing the recipe can impact the preparation.

## 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.

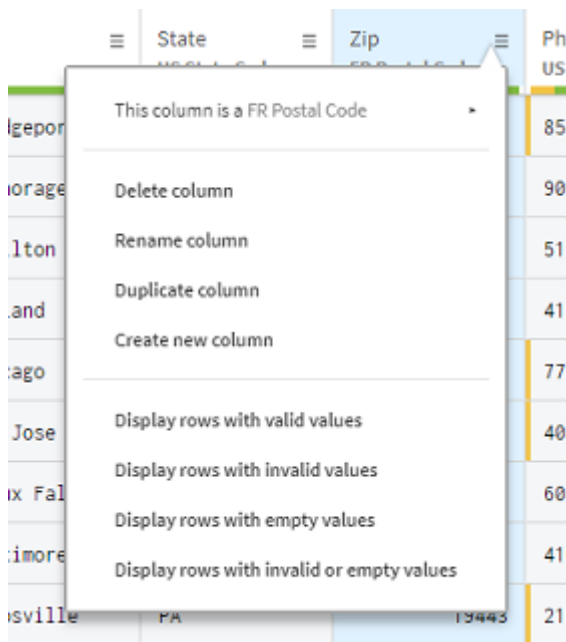
Talend Data Preparation automatically suggests a semantic type for your data. This type is specified under the header of each column. You can see in the **Zip** column that because they have the same number of digits, the US zip codes have been mistaken for French ones. You will now set the semantic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

≡	Zip	≡	Phc
	FR Postal Code		US I
	8014		856

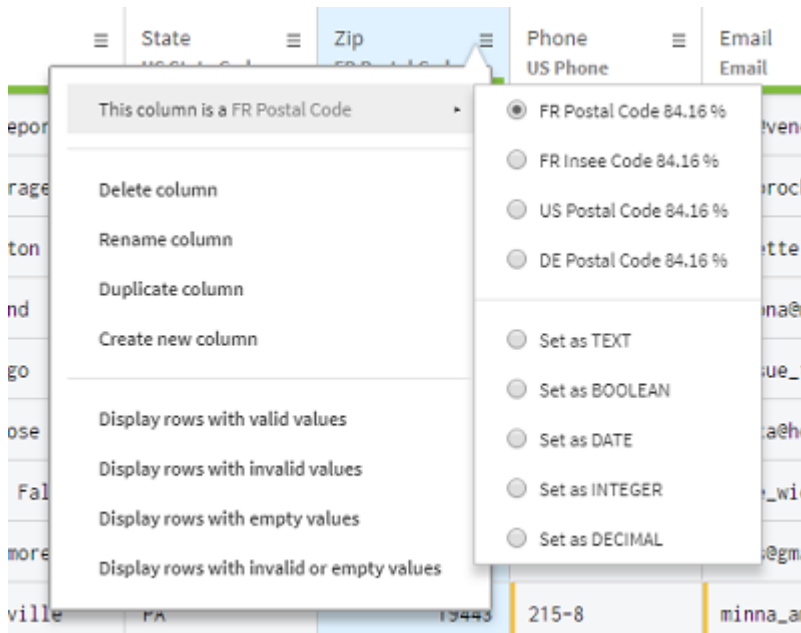
To modify the semantic type of a column, proceed as follows:

### Procedure

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



2. In the drop-down menu, point your mouse over **this column is a FR Postal Code**.

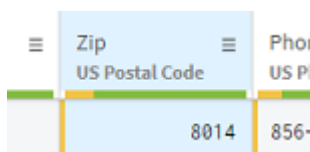


A list of suggested semantic types opens.

3. Click **US Postal Code**

## Results

The semantic type of the **Zip** column has now been set to **US Postal Code**.



More generally, if the semantic type proposed by Talend Data Preparation for one column is not the desired one, you can change it at any time, based on your own experience.

## Working with the quality bar

The quickest way to identify incorrect data is to look at the quality bar.

Under each column is a quality bar that displays the amount of fields that have correct data, incorrect data or empty fields. Each category is represented by a color:

- Green for data that matches the cell format
- White for empty cells
- Orange for data that does not match the cell format

Click any color to select, delete or clear the cells with data in an invalid format. Hovering over the colors allows you to display the exact number of lines for each category, as well as the percentage it represents in a column.

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.

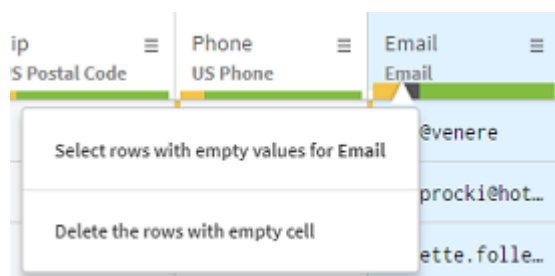


To use the quality bar to remove the lines containing those incorrect cells, proceed as follows:

### Procedure

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

A drop-down menu opens.



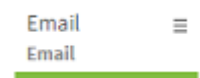
2. Click **Delete the rows with empty cell**.

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 cell**.

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.

### Results

The only remaining column with invalid data is now **State**, but you are going to treat it in a different way.

## Blending data

The Lookup feature allows you to take data from an existing dataset and add it to your preparation.

This example assumes that:

- You have retrieved the `states.csv` file from the **Downloads** tab of the documentation page.



- You have added `states.csv` to your list of datasets in Talend Data Preparation. For more information about how to import a dataset, see [Opening a dataset from a local file](#) on page 8.

In this example, you want to add more geographical information on your customers, thanks to a reference file that you possess: the `States` dataset. This dataset contains the list of the US State codes, and their corresponding region. You will dynamically use the data from this dataset to complement your preparation. This will allow you to add information about each customer's subscription region, based on their State code.

To blend the data from another dataset in your preparation, proceed as follows:

#### Procedure

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



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

Add a dataset by clicking on "+" button to perform a lookup

**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 **+** icon to select the dataset you want to add.  
The list of previously imported datasets opens. In your case, only `States` is available.
- 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.
- 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.

	Salary_Out text	Address Address Line	City text	State US State Code	Zip US Postal Code	Phone US Phone	Email Email	SubDate date
4	150,000-199,9...	639 Main St	Anchorage	AK	99501	907-385-4412	lpaprocki@hot...	24-Nov
5	50,000-99,999	34 Center St	Hamilton	OH	45011	513-570-1893	donette.folle...	17-Apr
6	100,000-149,9...	3 McAuley Dr	Ashland	OH	44805	419-503-2484	simona@morasc...	13-Apr
9	150,000-199,9...	5 Boston Ave ...	Sioux Falls	SD	57105	605-414-2147	sage_wieser@c...	20-Apr
10	< 50,000	228 Runamuck ...	Baltimore	MD	21224	410-655-8723	kris@gmail.com	31-Dec
13	150,000-199,9...	25 E 75th St ...	Los Angeles	CA	90034	310-498-5651	kiley.caldare...	08-Mar
14	< 50,000	98 Connecticu...	Chagrin Falls	OH	44023	440-780-8425	gruta@cox.net	12-Jun
15	> 200,000	56 E Morehead...	Laredo	Texas	78045	956-537-6195	calbares@mai...	25-Jun
16	< 50,000	73 State Road...	Phoenix	AZ	85013	602-277-4385	mattie@aol.com	01-Dec

	State US State <input type="checkbox"/> Add to Dataset	State Code US State Code	Region text <input type="checkbox"/> Add to Dataset
1	Washington	WA	West
2	Montana	MT	West
3	Oregon	OR	West
4	Idaho	ID	West
5	Wyoming	WY	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

states

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

Region  
text  
☒ Add to Dataset

West

West

West

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

	Salary_Out text	Address Address Line	City Airport	State US State Code	Region text	Zip US Postal Code	Phone US Phone	Email
4	150,000-199,9...	639 Main St	Anchorage	AK	West	99501	907-385-4412	lp
5	50,000-99,999	34 Center St	Hamilton	OH	Mid West	45011	513-570-1893	dc
6	100,000-149,9...	3 Mcauley Dr	Ashland	OH	Mid West	44805	419-503-2484	si
9	150,000-199,9...	5 Boston Ave ...	Sioux Falls	SD	Mid West	57105	605-414-2147	sa
10	< 50,000	228 Runamuck ...	Baltimore	MD	North East	21224	410-655-8723	kr
13	150,000-199,9...	25 E 75th St ...	Los Angeles	CA	West	90034	310-498-5651	ki
14	< 50,000	98 Connecticu...	Chagrin Falls	OH	Mid West	44023	440-780-8425	gr
15	> 200,000	56 E Morehead...	Laredo	Texas		78045	956-537-6195	ca
16	< 50,000	73 State Road...	Phoenix	AZ	South West	85013	602-277-4385	ma

	State US State <input type="checkbox"/> Add to Dataset	State Code US State Code	Region text <input checked="" type="checkbox"/> Add to Dataset
1	Washington	WA	West
2	Montana	MT	West
3	Oregon	OR	West
4	Idaho	ID	West
5	Wyoming	WY	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**

states

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

### Results

Your data now includes a new information about the subscription region of your customer, that you extracted from a reference file.

## 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:

### Procedure

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.

	State US State Code	Region text
	MD	North East
s	CA	West
lls	OH	Mid West
	Texas	
	<input type="checkbox"/> Apply to all cells with this value	
le	TN	South East
	WI	Mid West

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

	State US State Code	Region text
	MD	North East
s	CA	West
lls	OH	Mid West
	TX	
	<input checked="" type="checkbox"/> Apply to all cells with this value	
le	TN	South East
	WI	Mid West

4. Press the **Enter** key.

### Results

All the occurrences of **Texas** have been replaced by the correct **TX** State code and the quality bar now indicates that all the data in the **State** column is correct.

Note that when the **State** column is selected, the data is visualized in the form of an interactive map of the United-States in the **Data profiling** panel.

CHART   VALUE   PATTERN   ADVANCED

Row count ▾



### State distribution



## Reordering a preparation step

In Talend Data Preparation, each preparation step you apply on your data is based on the previous one. As a consequence, if you already applied many preparation steps to your data, but forgot one small change at the beginning, you would not achieve the expected result.

In a preparation with many steps, you have the possibility to rearrange your preparation steps so that the changes take effect in the right order.

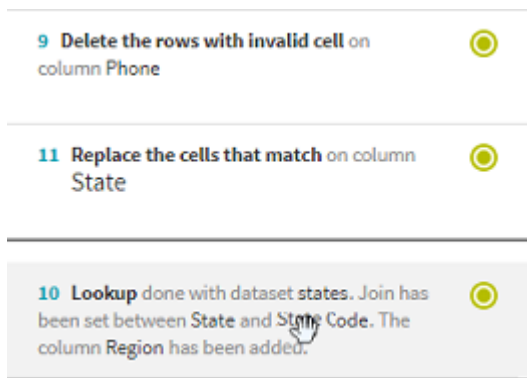
In the earlier steps, you have performed a lookup on the **State** column just before making a correction to this column, where the Texas entry was in the wrong format. Because that last change on the State column was performed after the lookup, some information is now missing from the **Region** column that was added with the lookup.

State US State Code	Region text
MD	North East
CA	West
OH	Mid West
TX	
AZ	South West
TN	South East

You are going to take the lookup step and place it as the last step of the preparation to make sure it includes all the States.

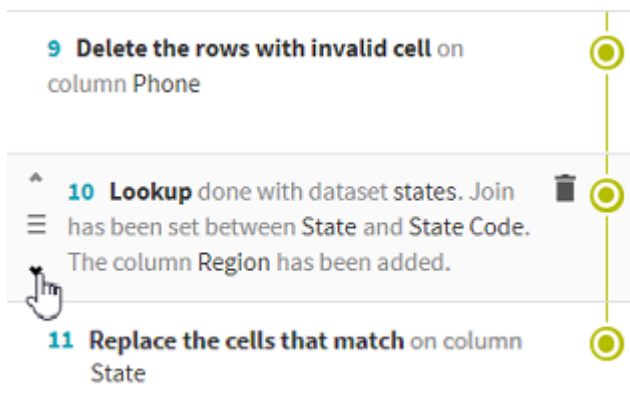
#### Procedure

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:
  - Drag the recipe step and drop it at the bottom of your recipe.



the grey line shows where the recipe will be placed.

- Click the up arrow on the left of your recipe step to move it down.



## Results

Your preparation is automatically updated with the correct sequence of actions, and the **Region** column now includes Texas.

CA	West	
OH	Mid West	
TX	South West	
AZ	South West	
TN	South East	

## Using charts to filter

An easy way to apply a filter on your data is to use the charts showing the graphical representation of your data.

Filters can be used to isolate values, thus making it easier to apply functions on a specific category of data. Using the quality bar to select empty or invalid cells is one solution, but in this example, you are going to learn how to filter data via the charts located on the bottom right of the interface. In Talend Data Preparation, the statistics for each column are available in the form of a chart.

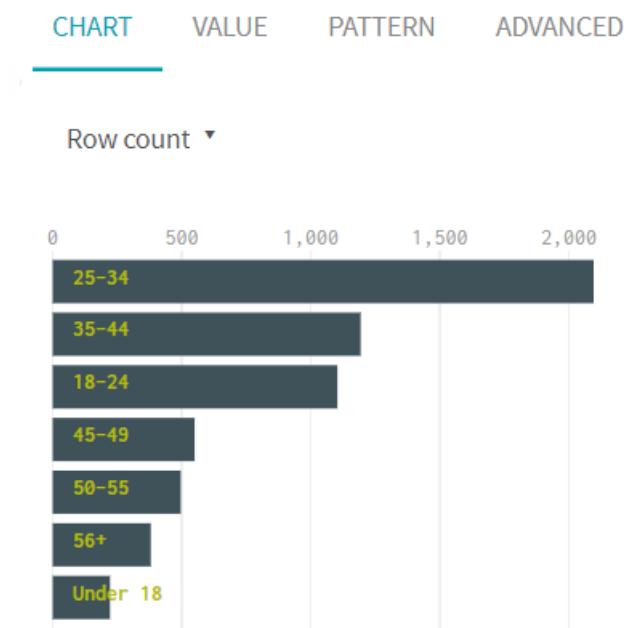
Imagine you want to have a better understanding of the age distribution of your customers, and identify those that are under the age of 18. To filter this specific group of customers using a chart, proceed as follows:

### Procedure

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

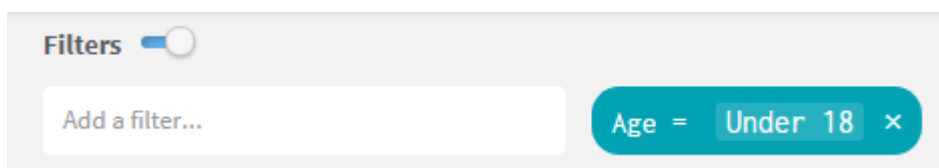
Age	Count
Under 18	1,000
56+	500
45-49	400
25-34	2,000

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**.

Your preparation now only displays the lines corresponding to clients under the age of 18. On the upper part of the screen, you can see that the filter `Age=Under 18` is currently active. You now have the possibility to apply functions to this group of customers only.



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

### Results

The preparation now displays the full list of customers again.

## Using charts to calculate absolute value

Calculating the absolute value of a number is one of the various mathematical functions available to use on your data.

If you take a close look at the **Number\_of\_rentals** column, you will notice that some of the numbers have a negative value.

	356
	778
	-4
	100
	78

These cells are not marked as incorrect in the quality bar because they still fit the semantic type automatically set as `integer`. Nevertheless, this is unusable data. As a consequence, you are going to apply a function to remove the negative sign for all these numbers.

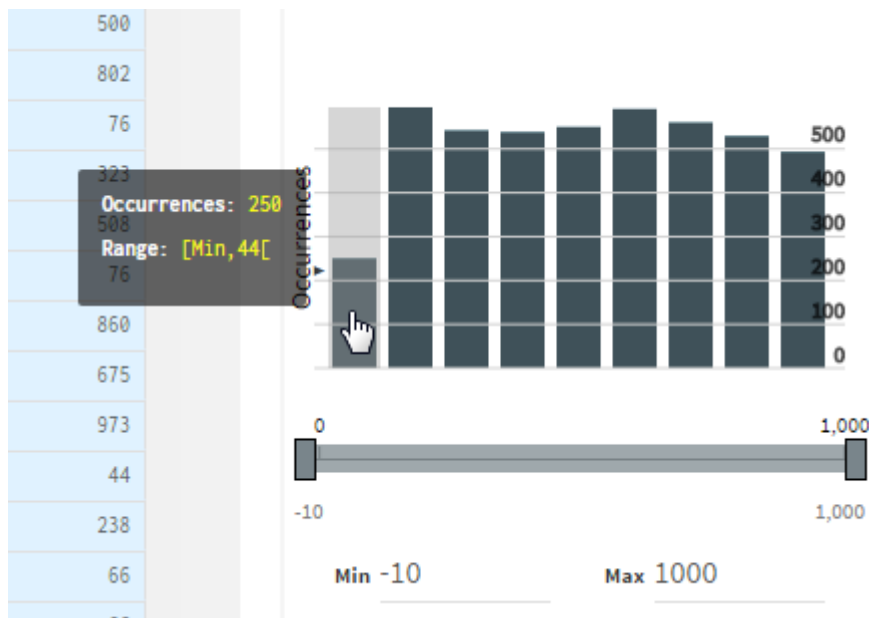
To calculate the absolute value of your data, proceed as follows:

#### Procedure

1. Click the header of the **Number\_of\_rentals** column to select its content.

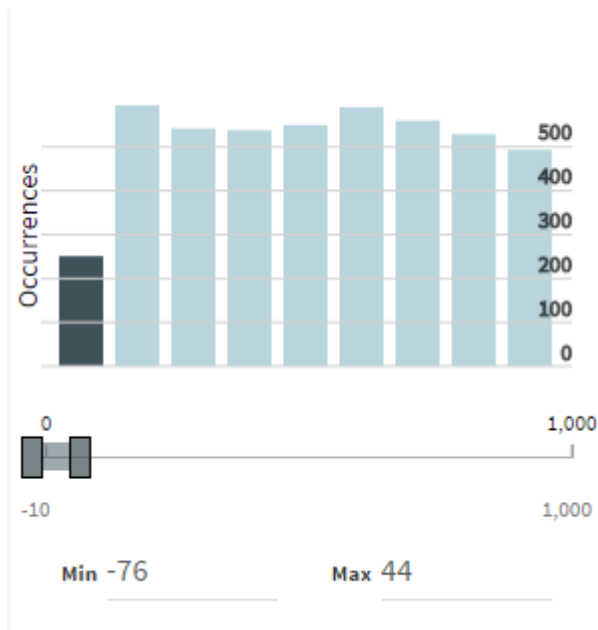
	Number_of_rentals
	integer
2016	541
2013	994
2013	586
2012	82

In the statistics box, you can clearly see that some values range between **-10** and **0**.



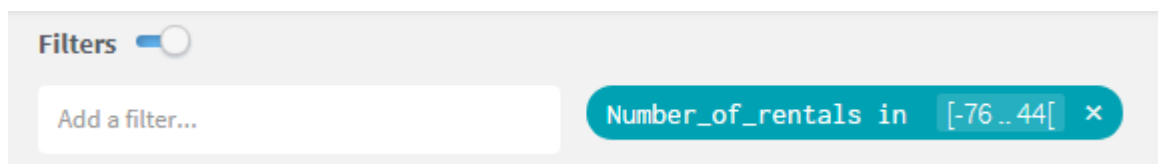
2. In the vertical bar chart at the bottom right of the screen, click the first bar from the left.



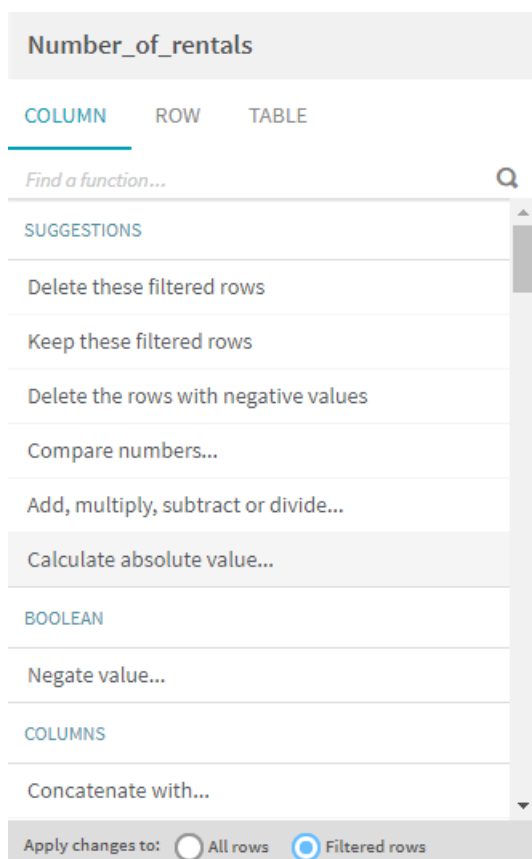


This bar represents all the occurrences of the values that are equal or below **0**.

A filter has now been applied on your data. Your preparation now only displays the lines with a value equal or below zero for the number of rentals. You can now apply a function only on those cells.



3. Under the functions list, in front of **Apply changes to:**, select the **Filtered rows** radio button.
4. In the functions list, click **Calculate Absolute Value**.



All the negative values have been converted.

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

## Results

Your preparation now displays all your data again. If you take another look at the statistics box for the **Number\_of\_rentals** column, you can see that the minimum value is now **0** instead of **-10**. You have thus improved the quality and usability of your data.

## setting a unique date format

Talend Data Preparation supports many different date formats, which you can harmonize to improve your data.

You can see in the **SubDate** column that even if your data respects the semantic type set as `date`, they do not follow only one date format. As a consequence, European and American standards, `-` and `/` are coexisting.

03-Feb-2016
24-Feb-2010
12/04/2007
26-Dec-2011
03/07/2007
10-Jul-2011

You are going to harmonize the **SubDate** column and set only one date format for all your data. To do so:

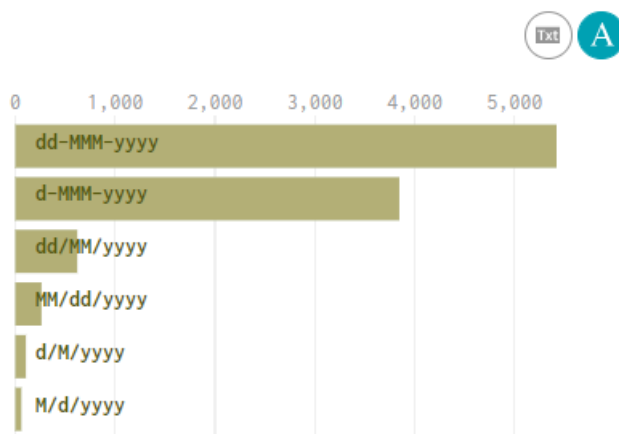
## Procedure

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

	SubDate date	Nur inte
.c...	17-Mar-2016	
ar...	15-Mar-2013	
ot...	24-Nov-2013	

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

CHART    VALUE    **PATTERN**    ADVANCED



This tab gives you a better view of the different date formats currently used. Some dates follow the European standard, while other follow the American format. In any case, you can see that the `dd-MMM-yyyy` format is the most commonly used.

- To standardize the date format, click **Change date format...** in the functions list.

A menu opens, where you can specify the current date formats, and the desired one.

- In the **Current format** drop-down list, leave **I don't know, best guess** selected.
- In the **New format** drop-down list, select **Other**.
- 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-MM-yyyy

SUBMIT

[Learn more...](#)

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

## Results

The **SubDate** column now follows only one date format, which make it easier to read. You can also notice that the recipe highlights your last action and it is even possible to modify the date format again, directly from the recipe.

12 Calculate absolute value on column Number\_of\_rentals

13 Change date format on column SubDate

☐ Create new column

Current format:  
I don't know, best guess

New format:  
Other

Your format:  
dd-MM-yyyy

SUBMIT

## Finding and grouping similar content

Finding and grouping similar text can be used to harmonize content with only small variations.

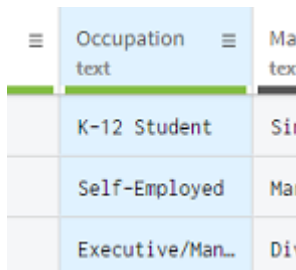
**Note:** The `Find and group similar text` function does not support Asian characters.

In the `customers.xlsx` file, there is information about the occupation of your clients. Some of the values are closely similar to each other, for example **College/Grad Student** and **College Student**. A way to improve the readability, and thus the quality of your data, would be to regroup some of these values together.

To find and group similar content, proceed as follows:

### Procedure

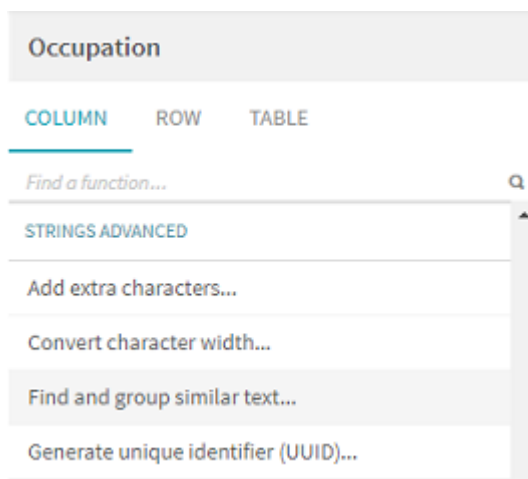
1. Click the header of the **Occupation** column to select its content.



Occupation	Ma tex
K-12 Student	Sir
Self-Employed	Ma
Executive/Man...	Div

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.

**Find and group similar text** ✕

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

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

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**.
4. Click **Submit**.

#### Results

All the occurrences of **College/Grad Student** and **College Student** have been regrouped under **College Student**, the new harmonized value.

# Sharing the finalized preparation

To make your preparation accessible to the other members of a project, you can use the share feature.

This example assumes that you and other users are registered Talend Data Preparation users.

Users or groups of users from your organization will be able to open and edit your preparations according to the rights they were given. Sharing a preparation is only possible if your preparation is located in a folder.

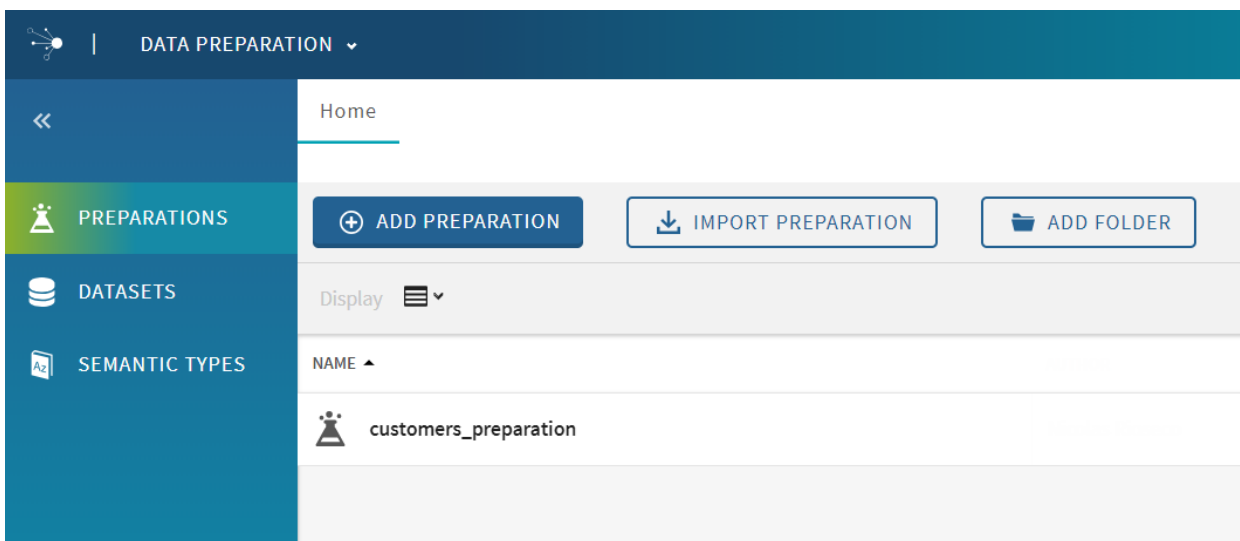
To create and share a preparation folder, proceed as follows:

## Procedure

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.

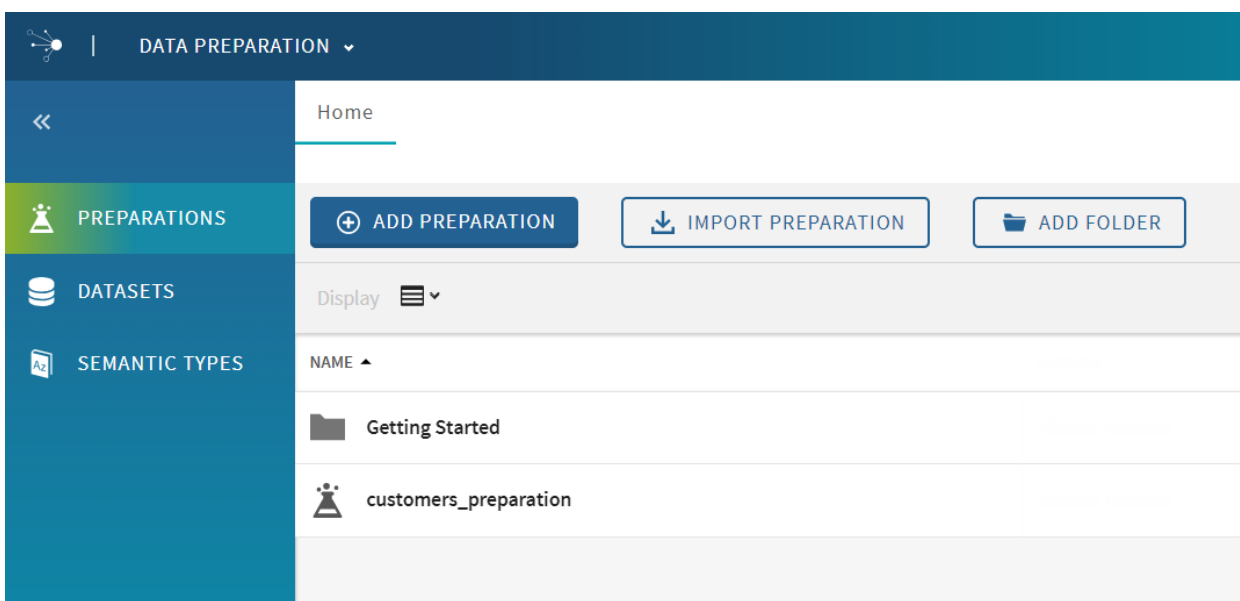


2. Click the **Add folder** icon.

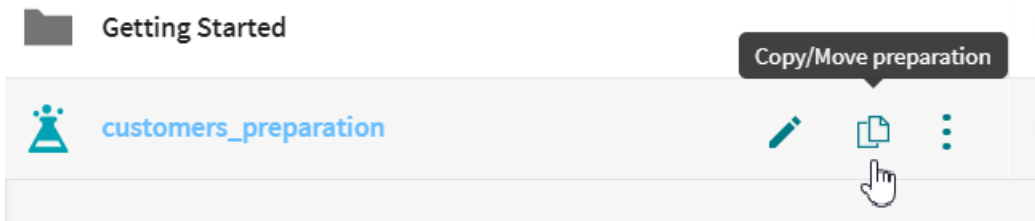
A window opens where you need to enter a name for your folder.

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

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



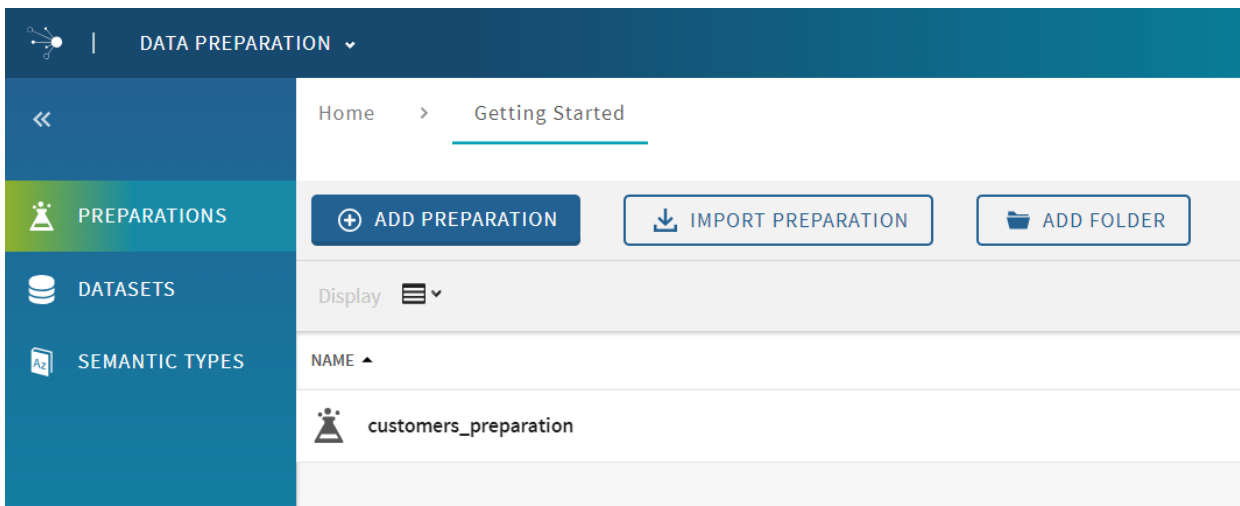
- Point your mouse over `customers_preparation` in order to display the available options and click the **Copy or Move** icon.



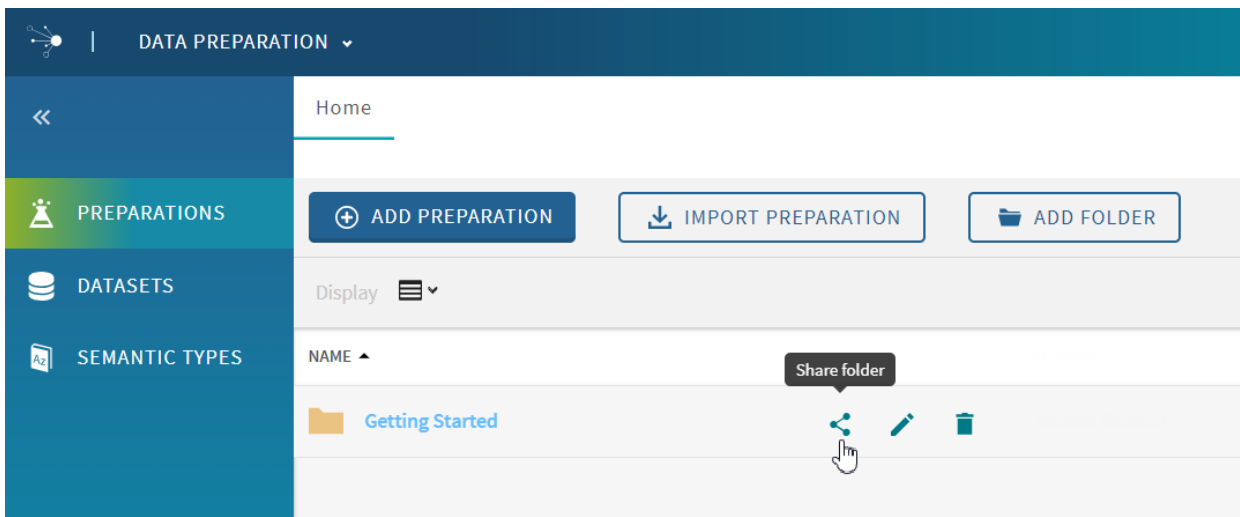
The **Copy/Move item** window opens, where you can select the destination folder for your preparation.

- 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.



- Click **Home** in the path to go back to the **Preparations** view.
- Point your mouse over the `Getting Started` folder in order to display the available options and click the **Share folder** icon.



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

- 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.
- 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

Q



Nicolas Talend  
Owner

← ADD TO LIST

CANCEL

All users and groups

Find user/group

Q

Marketing  
Sales

Anne Talend  
Arnaud Talend  
**Eric Talend**  
Stéphane Talend

CONFIRM

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

### Results

You can see from the change in the `Getting Started` folder icon, that it was successfully shared with the selected users.



If other users share a preparation folder with you, the folder icon will be blue.

# Exporting the result of your preparation

Once your preparation is complete, you may want to export the data you have cleansed.

To export your preparation as a local file, proceed as follows:

## Procedure

1. Click the **Export** button in the upper right side of the screen.

**Export**

☐ Current sample ☒ **All data**

---

☐ Local CSV file

☐ Local XLSX file

☐ Local TABLEAU file

☐ Amazon S3

Here you can choose between the CSV, XLSX or Tableau format to export your preparation. In this example, you are going to export it as an XLSX file.

2. Click the **Local XLSX file** radio button from the list.
3. In the **Filename** field, enter a name for your file, `customers_preparation` in this example.

**Export to XLSX**

☐ Current sample    ☒ **All data**

---

☐ Local CSV file

☒ **Local XLSX file**

Filename:  
customers\_preparation

☐ Local TABLEAU file

☐ Amazon S3

4. Click **Confirm**.

#### Results

The data you cleansed using your preparation has been exported to a local file.

# What's next?

Now that you have completed this basic data cleansing scenario, you can learn about Talend Data Preparation more advanced features:

- Using your preparations directly in the flow of your Talend Jobs in Talend Studio. See "Operationalizing a recipe in a Talend Job".
- Importing larger files, and working on a sample before applying your preparation to the whole data. See "Working on large datasets".
- Connecting to various sources to import your data in self-service.
  - "Adding a dataset from a database"
  - "Adding a dataset from HDFS"
  - "Adding a dataset from Amazon S3"
  - "Adding a dataset from Salesforce"

Look for the aforementioned articles on the documentation portal.