



# Talend Data Preparation Free Desktop Installing Guide

พิชิตชัย พิมพ์โคตร

	Page
1. Requirements for Talend Data Preparation Free Desktop .....	3
2. Installing Talend Data Preparation Free Desktop .....	4
• Download Installer Package	
• Set-up Talend Data Preparation on Windows	
• Set-up Talend Data Preparation on Mac	
3. Configuring the language of the interface.....	7
4. How to add a Preparation?.....	8
5. Manipulate data.....	9

# Requirements for Talend Data Preparation Free Desktop

## Hardware requirements

Processor	64-bit processor is required
Allocated memory	1GB minimum
Disk space	500MB minimum + datasets = 5GB recommended

## Software requirements

Operating system	<ul style="list-style-type: none"><li>• Windows 7 64-bit or more recent</li><li>• Mac OS X 10.7 "Lion" or more recent</li></ul>
------------------	---

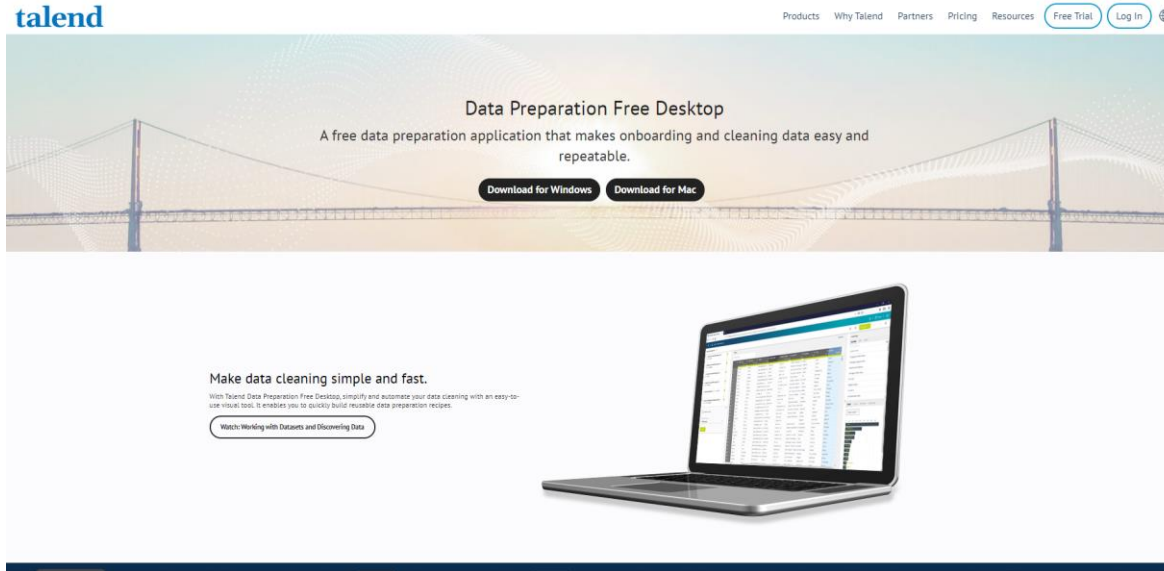
## Compatible Web browsers

Mozilla Firefox / Firefox ESR	Latest version
Microsoft Internet Explorer	11
Microsoft Edge	Latest version
Apple Safari	10
Google Chrome	Latest version

Here is the software and hardware information required and recommended to get started with Talend Data Preparation.

### Java:

There are no specific Java requirements for most of Windows and Apple computers. However, if you want to install the Apache version of Talend Data Preparation, you must have Oracle Java 8 64-bit installed on your computer. The default Windows 32-bit version is not supported, only the 64-bit version is.



**Download Talend Data  
Preparation here  
>> Click <<**

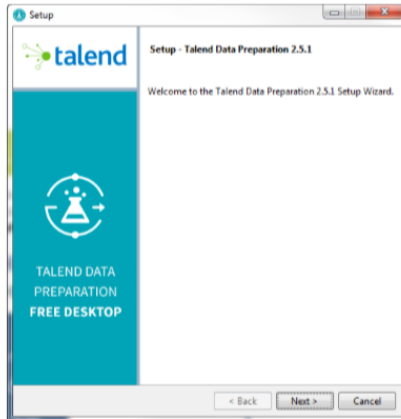
**Select your operating system,  
and the download starts  
automatically.**

Link download : [https://www.talend.com/products/data-preparation/data-preparation-free-desktop/?qt-product\\_tos\\_download\\_new=5](https://www.talend.com/products/data-preparation/data-preparation-free-desktop/?qt-product_tos_download_new=5)

# Set-up Talend Data Preparation on Windows

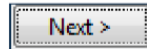
1

Locate the file you have just downloaded and double click **Talend-DataPreparation-Free-Desktop-2.5.exe**



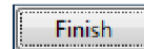
2

Click **Next** through the setup and use the default settings.



3

Click **Finish** once the Install is complete.



4

Use the Desktop icon or the shortcut on the Start menu to begin using the Talend Data Preparation tool.



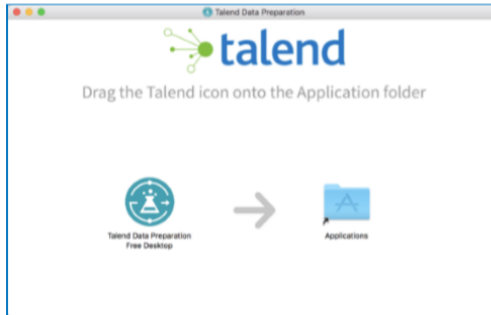
# Set-up Talend Data Preparation on Mac

1

Double-click the **Talend-DataPreparation-Free-Desktop-2.5.dmg** file to open the package.

2

Drag and drop into the Applications folder.



3

Talend Data Preparation will now be in your list of **Applications**. Locate the icon and double click to open the application.

4

To disable **App Nap** and ensure optimal performance, follow this quick procedure:

1. Open the Terminal from the `/Applications/Utilities` folder.

2. Enter the following command:

```
defaults write  
org.talend.dataprep  
NSAppSleepDisabled  
-bool YES
```



1

Open the  
<TDP\_Installation\_Path>  
/dataprep/config/applic  
ation.properties  
configuration file.

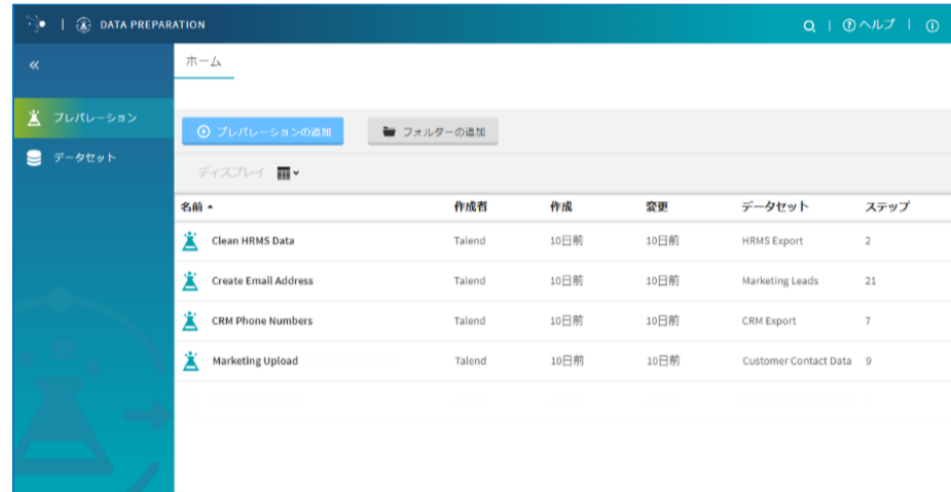
2

For the dataprep.locale  
parameter, enter one of the three  
supported values:

- **en-US** for English
- **fr-FR** for French
- **ja-JP** for Japanese

3

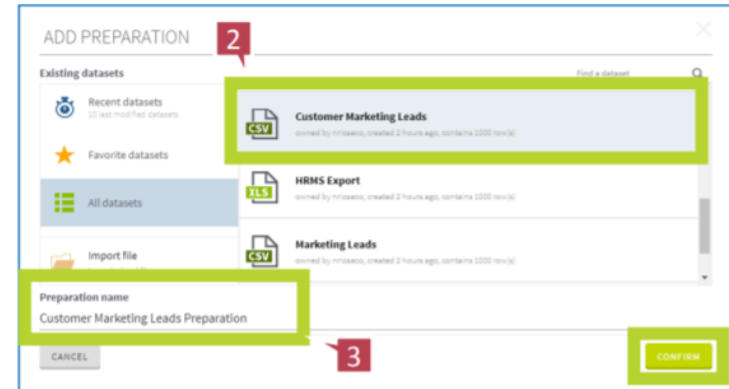
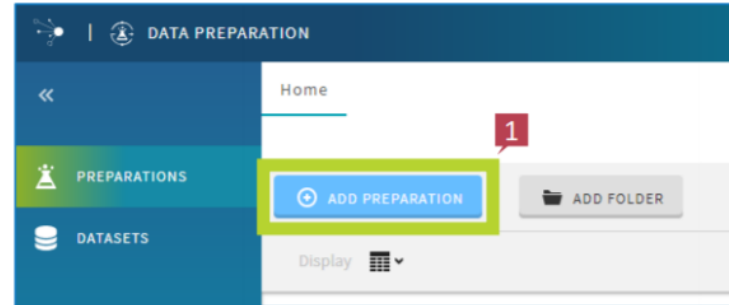
Restart Talend Data Preparation  
Free Desktop.



# How to add a preparation?

To get started on the example:

1. Click the **Add Preparation** button from the **Preparations** view.
2. The **Add Preparation** dialog opens. Click the **Customers Marketing Leads** dataset from the **All Datasets** list.
3. Choose a **name** for your preparation.
4. Click **Confirm** to open the preparation and start cleansing the data from the dataset.





## Simple cleansing examples

First, let's fix the **Name** column:

1. Click the header of the **Name** column.
2. While pressing the **Ctrl** key, click the header of the **last\_name** column. Both columns are now selected. You can also use **Shift + click** to select multiple columns.
3. In the top right corner is a box with all the **Functions** available. You can search for functions or use one of the suggested functions to improve your data.
4. You may need to scroll, down depending on your screen resolution, to find the **Change to upper case** function. Hover over the function to preview its effect on the data. Click the function to apply the changes to the two selected columns.

Here we are cleaning up the customer name fields to do some basic standardization, You can see that there are mixed case names, leading and training Spaces and the last name has been defined as an incorrect type.

The screenshot shows a data preparation interface with a table of customer data. The 'Name' and 'last\_name' columns are selected, indicated by a green box and a red arrow labeled '1'. A second red arrow labeled '2' points to the 'last\_name' column header. A third red arrow labeled '3' points to the 'Functions' panel on the right. A fourth red arrow labeled '4' points to the 'Change to upper case...' function in the list. The table data shows various names, some with mixed case and leading/trailing spaces.

NAME	last_name
771143 Jason	Alexander
770396 Lillian	Simpson
524952 WALTER	Ruiz
744980 Joshua	Hunt
484856 Mildred	Flores
958818 Victor	Gonzalez
595842 Joshua	Simmons
149872 Beverly	Wright

## Simple cleansing examples

To execute basic formatting and cleansing:

Name column, continued.

1. While looking at the data, you will see grey boxes in front or behind some names, for example "Joshua".
2. To remove those grey boxes, search for and select the **Remove trailing and leading characters** function.
3. Leave the **Create new column** checkbox clear. In the **Padding character** drop-down list, select **Whitespace** and click **Submit**.

Several functions, including this one, allow you to output the result of the transformation in a new column by selecting the **Create new column** checkbox. If you do not select it, the function will be applied in the current column.

The screenshot displays the DATA PREPARATION interface. The main window shows a table with columns: ID, Name, Job title, Location, Company, City, State, Date, Campaign ID, and Lead score. A red arrow points from the 'Name' column header to a zoomed-in view of the data. The zoomed-in view shows a table with columns: ID, Name, and Job title. The 'Name' column contains values like 'JOSHUA', 'MILDR', 'VICTOR', 'JOSHUA', 'BEVERLY', 'FRED', and 'JOSEPH'. A green box highlights the 'JOSHUA' entry, which has a grey box behind it. A red arrow points from this entry to the 'Remove trailing and leading characters...' function configuration panel. The panel has a 'Name' dropdown set to 'COLUMN', a 'Remove trailing and leading characters...' checkbox checked, a 'Create new column' checkbox unchecked, a 'Padding character' dropdown set to 'Whitespace', and a 'SUBMIT' button. A red arrow points from the 'SUBMIT' button back to the data table.

## Recipes

1. Every time you apply a function, it is added to the recipe panel on the left.
2. To delete a recipe item, point your mouse over the line item and click the trash can icon.
3. To rename a preparation, click the pencil icon and enter a new name.
4. The recipe panel can be hidden by clicking the arrow.
5. To export the result of your preparation, click **Export** and select a file type.

Because you created this preparation using the **Add Preparation** button, you do not need to save anything. Every new preparation step is automatically saved.

The screenshot shows the 'DATA PREPARATION' window for 'Customer Marketing Leads Preparation'. The left panel lists steps: 'Change to upper case on column Name', 'Change to upper case on column last\_name', and 'Remove trailing and leading characters'. The right panel shows a table of 15 rows with columns: id, Name (First Name, Last Name), email, job\_title, company, city, and state. Red callouts are placed over the interface: 1 points to the arrow to hide the recipe panel, 2 points to the trash can icon, 3 points to the pencil icon, 4 points to the 'Remove step' button, and 5 points to the 'Export' button.

Multiple preparations can be saved and created for a single dataset.

Remember that the original data from your dataset remains unchanged.

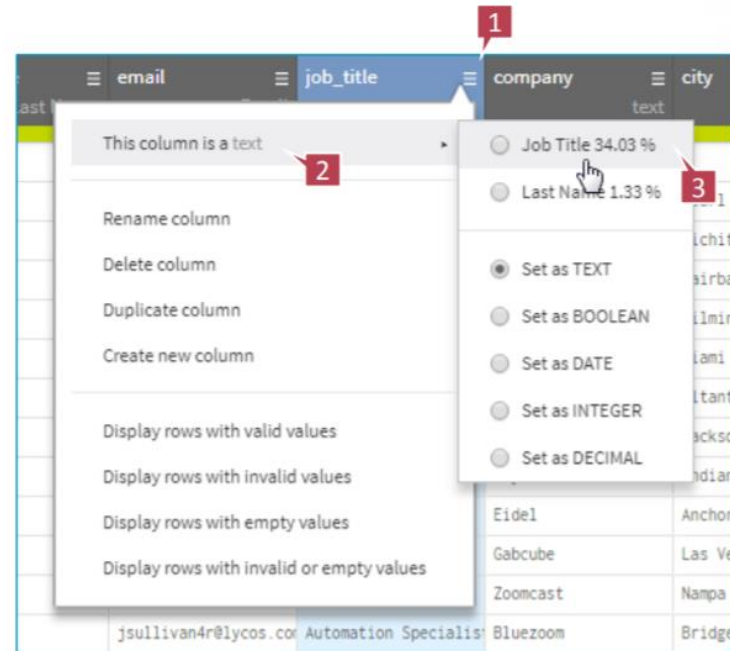
## Semantic type

Talend Data Preparation automatically suggests the proper semantic type for each column of your datasets. It will help you to further discover the data. But you can change those suggestion at any time, based on your own experience.

The suggested semantic type for the **job\_title** column is **Text**. Let's change it to a more meaningful one, **Job Title** in this case.

1. Click the **menu icon** on the column header and select a new semantic type.
2. Point your mouse over **This column is a text**.
3. Select **Job Title** as new semantic type.

The Enterprise Edition of Talend Data Preparation allows you to create custom semantic types, as well as editing or removing the default ones.



## Data quality bar

Under each column is a data quality bar that displays the amount of fields that have correct data, empty fields, or incorrect data. Each of these 3 are represented by a color.

- **Green** – Data matches the cell format
- **White** – Empty cells
- **Orange** – Data in the cell does not match the cell format

id	Name	last_name	email	job_title
integer	First Name	Last Name	Email	text

Let's take a closer look at the quality bar for the **email** column. Exact numbers and percentages can be found by pointing your mouse over each color.

- **Green** – 979 cells have data in the correct format
- **White** – 20 empty cells
- **Orange** – 1 cells have entries in an incorrect format

Click any color to select, delete, or clear the cells with data in an invalid format. Click the orange section and click **Select rows with invalid values** for the **email** column to display the entries with an incorrect format.

The first screenshot shows the email column header with a tooltip indicating 979 valid values (98%). The second screenshot shows the same header with a tooltip indicating 20 empty values (2%). The third screenshot shows the same header with a tooltip indicating 1 invalid values (0%). The fourth screenshot shows a context menu with options: 'Select rows with invalid values for email', 'Clear the cells with invalid values', and 'Delete the rows with invalid cell'.

Don't forget to clear the filter to return to the full list.

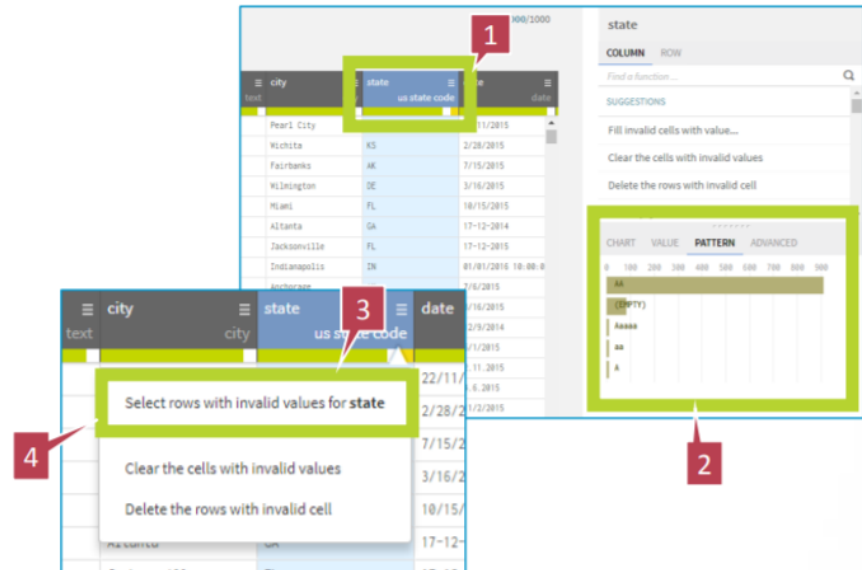
The screenshot shows the Filters section with a filter applied to the email column: 'email : rows with invalid values'. The filter is active, and the count 1/1000 is displayed.

## Basic text manipulation

How to filter invalid rows:

1. Click the header of the **state** column.
2. In the bottom right is a **Pattern** table. Point your mouse over the rows to see counts. The top row indicates that 911 records contain a 2 letter state code. **You can click a bar to isolate those records (to remove the filter, click the x in the filter, or click the bin icon in the filter bar).**
3. On the data quality bar, click the **orange** section.
4. Click **Select rows with invalid values for state**.
5. 7 rows that contain invalid information will be displayed.

Here we are cleaning and changing the values in a field with invalid values. You will see how you can use the charts to help filter the data as well as change values directly in the grid.





## Basic text manipulation

How to filter invalid rows:

1. To edit the text value in a field, **double-click one of the cells** that contain **Texas**. Change **Texas** to **TX**. **DO NOT** hit Enter yet!
2. Under the cell that you are editing is a check box with the label **Apply to all cell with this value**. Check that box. **NOW** hit Enter! You have changed all cells with the value **Texas** to **TX**.
3. That should leave you with 2 rows with incorrect data. Check out the different functions and **you pick the one you want** to use to fix the invalid state codes!
4. Once all actions and functions are applied, your **data quality bar** under the **state** column should now only contain **green and white**.
5. Click the **x** in the **state: rows with invalid values** filter to return to the full list.

Filters: state: rows with invalid values x

	last_name	email	job_title	company	city	state	date	camp
	Last Name	Email	Job Title		test	Airport	US State Code	date
213	KENNEDY	ekennedy@youtu.be	Executive Secretary	Filipodia	Cedar			HOOKEY
754	MATTHEWS	amathewsg@oup.io	Staff Scientist	Eazzy	Dallas	TX	11/13/2015	HOOKEY
756	LOPEZ	jlopez@geocities.jp	Clinical Specialist	Flipstorm	Austin			HOOKEY
757	CRAWFORD	ecrawford@nasa.gov	Administrative Assis	Ozu	Dallas			BIKE_3
765	SHAW	jshaw@mduic.edu	Occupational Therapi	Rooxo	Piano			TRAIL
961	WEBB	rwebb@theguardian.	Administrative Assis	Thoughtlink	Dallas	Texas	10/28/2015	HOOKEY
985	WALKER					E	8/11/2015	TRAIL

state	date
us state code	
HI	22/11/
KS	2/28/2

## Recipes

Each function that has been applied has been added to our recipe. Looking at the last steps in the recipe, it is easy to identify that we changed all fields that had **Texas** listed as a state to **TX**.

3 Remove trailing and leading characters on column Name

4 Change semantic domain on column job\_title

5 Search and replace on column state

state : rows with invalid values x

☐ Create new column

Search for:  
= Texas

Replace with:  
TX

☐ Overwrite entire cell

SUBMIT

6 Delete these filtered rows on column city

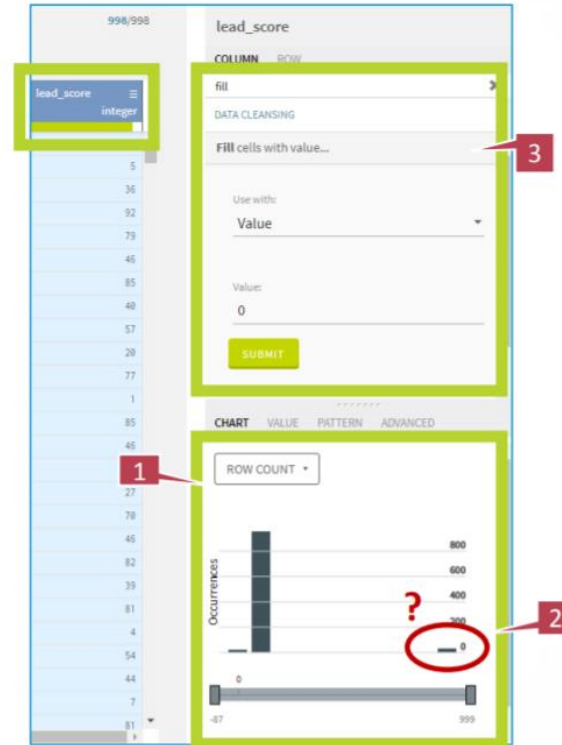


## Basic numeric manipulation

Next, let's look at the **lead\_score** column.

1. Select the **lead\_score** column. You will see that it contains basic integer values. But look at the **histogram graph** at the bottom right. The data is being skewed by some large value.
2. Click the **blue bar** on the far right of the graph. It should return 31 records with the value of 999. It looks like the default is set to 999. This time, to change the data, you will use a function called **Fill cell with value**.
3. Type **Fill** into the search field in the upper right. Click the function **Fill cell with value**. Set the value to **0** and click **Submit**.

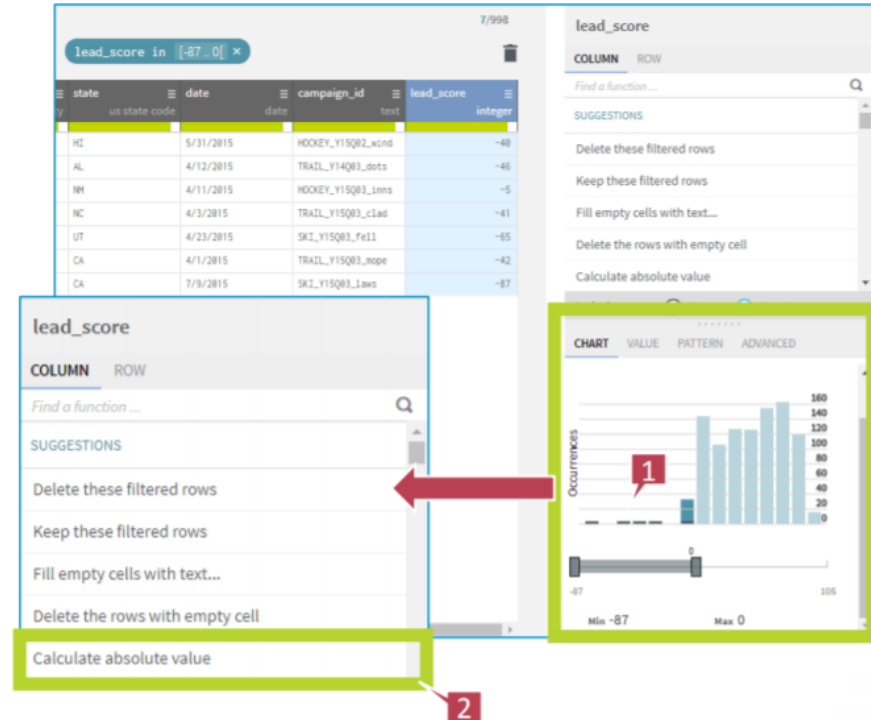
Here, we are cleaning and changing outliers in a numeric field. You will see how you can use charts to help filter the data as well as change values directly in the data grid.



## Basic numeric manipulation

**lead\_score** column, continued.

1. If you take a close look at the graph for the **lead\_score** column, you will notice that there are negative lead scores.
2. Since we cannot have negative lead scores, let's remove those values. Under the suggested functions, click **Calculate absolute value**. This will keep the rows with their respective lead score numbers while dropping the negative sign.



## Date cleansing and formatting

Next, look at the **date** column.

1. Click the header of the **date** column, then change the view on the right to **Pattern**. This gives you a better view of the different date formats and masking used. Some dates are formatted in the European standards and others are formatted in the US standard. Some contain - and other /.
2. To standardize the dates, click **Change date format** under the suggested functions. Select a pre-existing format or type one in. Click **Submit** when ready.

How many times do we see a spreadsheet with all kinds of crazy data formats and standards? We all know that Excel can reformat a date field, but when the dates are a mix of European standards and US standard with different masking, Excel starts to break DOWN!

The screenshot shows a data table with columns: state, date, campaign\_id, lead\_score. The 'date' column contains various date formats. The sidebar on the right shows the 'Pattern' view selected. The 'Change date format...' option is highlighted. A dialog box for 'Change date format...' is open, showing the current format and a list of suggested formats.

state	date	campaign_id	lead_score
KI	22/11/2015	HOOKEY_Y15Q81_cant	5
KS	2/28/2015	RUL_Y14Q82_dead	36
AK	7/15/2015	TRADL_Y14Q84_purr	92
DE	3/16/2015	HOOKEY_Y14Q82_nose	79
FL	18/15/2015	HOOKEY_Y15Q84_chun	46
GA	17-12-2014	TRADL_Y15Q83_hold	85
FL	17-12-2015	TRADL_Y14Q83_noon	40
IN	01/01/2015: 18:00:00	TRADL_Y15Q84_rossy	57
AK	7/6/2015	BDKE_Y14Q82_hurt	29
NV	5/16/2015	HOOKEY_Y15Q82_boos	77
ID	12/9/2014		
CT	6/1/2015		
KI	2.11		
OR	4.8.		
NV	11/2/2015		
GA	12/25/2014		
KS	8/31/2015		
CT	6/4/2015		
CT	12/2/2014		

Change date format...

☐ Create new column

Current format:  
I don't know, best guess

New format:  
Other

Your format:  
MM.dd.yyyy

SUBMIT

## Date cleansing and formatting

Modifying recipes is simple.

1. From the **recipe** on the left, highlight the last action.
2. In the drop-down for the **Change date format** operation, select **Other** (design your custom pattern). Enter **dd-MMMM-yyyy** (Date formatting is case sensitive so pay attention to the case).
3. Once you click **Submit**, the change will take effect. You can delete a step from the recipe list of actions on the left or click the green dot to inactivate that action.
4. You can also **reorder the steps of your recipe through drag & drop**. You will save time if you realize that a column you applied a function on, still does not fully contain the expected data.

The screenshot shows a data manipulation interface. On the left, a 'Recipe' list contains two actions: '8 Calculate absolute value on column lead\_score' and '9 Change date format on column date'. The 'Change date format' action is selected, and its configuration panel is open. It shows 'Current format: I don't know, best guess' and 'New format: Other'. A green box highlights the 'Your format: MM.dd.yyyy' field and the 'SUBMIT' button. A red number '2' points to the 'SUBMIT' button. A second configuration panel is shown below it, with 'Your format: dd-MMMM-yyyy' and a 'SUBMIT' button, highlighted by a green box. A red number '3' points to this 'SUBMIT' button. On the right, a 'Filters' panel is visible above a data table. The table has columns: email, job\_title, company, and text. It contains 12 rows of data. A red number '1' points to the 'Change date format' action in the recipe list.

	email	job_title	company	text
2	jalexander44@gmail.com	Chemical Engineer	Abata	
3	lsimpsonf7@gmail.com	Desktop Support Tech	Canimbo	
4	wruizlz@gmail.com	Geological Engineer	Yakitri	
5	jhuntek@last.fm	Financial Advisor	Oyope	
6	mflores06@earthlink.net	Nurse	Edgebiab	
7	vgonzalez8c@npr.org	Sales Associate	Ntag	
8	jsimmonsa5@newyorker.com	Occupational Therapist	Oba	
9	bwright38@arizona.edu	Biostatistician	Skynoodle	
10	frodriqueznc@fotki.com	Director of Sales	Eidel	
11	jpetersonnm@sohu.com	Research Nurse	Gabcube	
12	dmartint@java.com	Speech Pathologist	Zoomcast	

## Data masking

You can easily mask sensitive data.

1. Click the **email** column to select its content.
2. In the function list, search for **Mask data (Obfuscation)**.
3. Click it to apply the function on the email entries.
4. All the characters before @ are replaced by XXX, while the rest is left unchanged. This is the effect of the data masking function on entries whose semantic type is email. But the effects of the data masking will be different depending on a column's semantic type.

When manipulating sensitive data such as names, addresses, credit card or social security numbers, you might want to mask this data. To protect the original data, you will use the data masking function to generate functional substitutes.

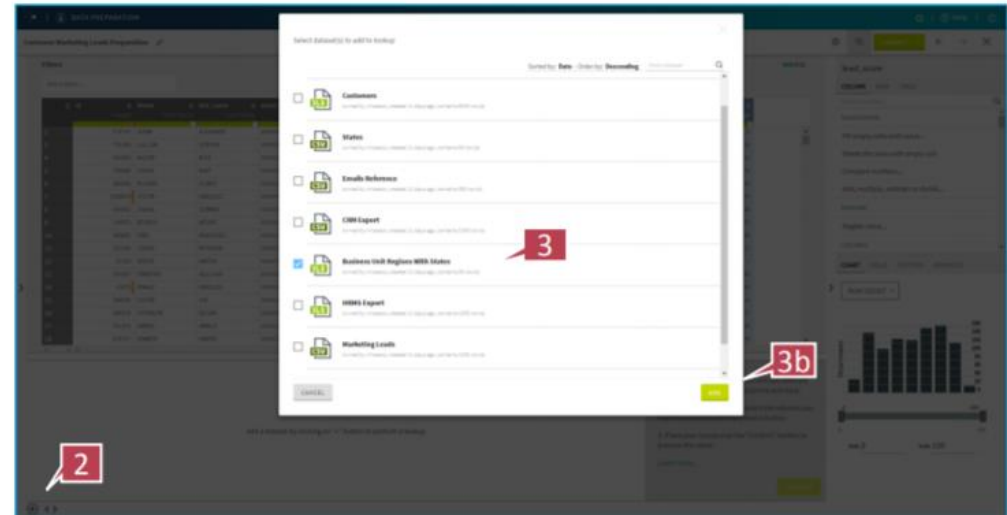
The screenshot illustrates the process of applying a data masking function in the Data Preparation tool. The main window shows a table with columns including 'email'. A red callout '1' points to the 'email' column header. A second red callout '2' points to the 'Functions' pane on the right, where 'Mask data (Obfuscation)' is selected. A third red callout '3' points to the 'Mask data (obfuscation)' option in the 'DATA MASKING' category. A fourth red callout '4' points to the 'email' column in the preview table, which now displays masked email addresses (e.g., 'XXXXXXXXXXXX@gmail.com').

email	ROW
XXXXXXXXXXXX	
XXXXXXXXXXXX@gmail.com	
XXXXXXXXXXXX@gmail.com	
XXXXXXXX@gmail.com	
XXXXXXXX@last.fm	
XXXXXXXXXXXX@earthlink.net	
XXXXXXXXXXXX@npr.org	

## Data blending

Data blending is about connecting data from different sources. It allows you to take data from another preloaded dataset and add them into the dataset you are currently working on.

1. Click the **Lookup** icon.
2. All the datasets that you have loaded plus some preloaded are available to choose from by clicking the **+** icon.
3. Click the checkbox in front of **Business Unit Regions With States** and click **Add**.



## Data blending

Data blending, continued.

1. Click the **column you would like to blend**, the **state** column in your current dataset.
2. At the bottom, add the region information by clicking **Add to dataset** under the **Region** column header.
3. **Point your mouse over the Confirm button** to preview the changes, that are displayed in green. To accept the changes, click **Confirm**.

The screenshot displays the Alteryx Data Preparation interface for a dataset titled "Customer Marketing Leads Preparation". The main table lists customer data with columns: ID, Name, First Name, Last Name, Email, Title, Job Title, Company, City, Region, State, and Campaign ID. A secondary table at the bottom shows the "Region" column with a list of regions. A callout box labeled "2" points to the "Add to Dataset" button in the Region column header. A callout box labeled "3" points to the "Confirm" button in the bottom right corner. A callout box labeled "1" points to the "State" column in the main table. A callout box labeled "3b" points to the "Confirm" button in the top right corner. A "Preview" window on the right shows the resulting data with the "State" column highlighted in green. A "Confirm" button is also visible in the bottom right corner of the main interface.



# Manipulate data

## Group and standardize

Group and standardize allows you to find cells that have similar content and group them together by changing the text to match.

1. Click the **job\_title** column header.
2. The chart on the bottom right displays the large amount of slightly different job titles. To reduce the number of job titles, let's group similar job titles together.
3. In the search field, search for **group**.
4. Click the **Find and group similar text** function.

The screenshot displays the 'Group and standardize' tool interface. On the left, a table lists job titles, companies, cities, and states. The 'job\_title' column header is highlighted with a red box labeled '1'. On the right, a search bar contains the text 'group', with a red box labeled '3' next to it. Below the search bar, the 'Find and group similar text...' function is highlighted with a red box labeled '4'. On the bottom right, a bar chart shows the frequency of job titles, with 'Therapist' being the most frequent. A red box labeled '2' is placed near the chart. The interface also includes a 'ROW COUNT' dropdown and a 'CHART' tab.

job_title	company	city	state
Chemical Engineer	Abata	Pearl City	HI
Desktop Support Tech	Cambo	Wichita	KS
Geological Engineer	Vakitra	Fairbanks	AK
Financial Advisor	Oyap	Wilmington	DE
Nurse	Edgemlab	Miami	FL
Sales Associate	Wtag	Atlanta	GA
Occupational Therapist	Oba	Jacksonville	FL
Biostatistician	Skyhoodle	Indianapolis	IN
Director of Sales	Eloel	Anchorage	AK
Research Nurse	Gaboule	Las Vegas	NV
Speech Pathologist	Zooncast	Nampa	ID
Automation Specialist	Bluesoon	Bridgeport	CT
Automation Specialist	Shuffletag	Racine	WI
Librarian	Skallith	Bond	OR
Actuary	Rhyloo	Manhattan	NY
Senior Editor	Tazzy	Columbus	GA
Structural Engineer	Dynava	Overland Park	KS
Help Desk Operator	Gabtone	Orange	CT
Senior Sales Associate	Npath	Chester	CT
VP Marketing	Guzz	New Haven	CT
Research Associate	Tavu	Prosspect	CT
Tax Accountant	Devbug	New Haven	CT
Professor	Blogpad	East Lyme	CT
Financial Analyst	Chatterpoint	New Haven	CT
Systems Administrator	Fivedridge	Greenville	DE
Junior Executive	Kuimber	Wilmington	DE
Librarian	Fendfish	Pike Creek	DE
Nurse	Youfeed	Greenville	DE



## Find and group similar text

Group and standardize, continued.

1. All similar job titles are grouped together in the second column
2. The third column suggests a job title that could **replace** the data in the second column. You can **use the drop-down list to choose a different job title or type in an appropriate job title.**
3. If you do not want to change a specific job title, leave the check box in front of the job title **clear**.
4. If you do not want to change a group of job titles, leave the check box in front of the first column **clear**.
5. Click **Submit** when finished.

The screenshot shows a web interface titled "FIND AND GROUP SIMILAR TEXT" with a close button in the top right. Below the title is a subtitle: "Replace all similar values with the right one (i.e. cluster on fuzzy matching)". The interface is divided into three main columns:

- Column 1:** A header row with a checkbox and the text "These values have been found". Below this, there are three rows of checkboxes. The first row has a checkbox with a red callout "3" pointing to it. The second row has a checkbox. The third row has a checkbox with a red callout "4" pointing to it.
- Column 2:** A header row with the text "This value will be kept". Below this, there are three rows of job titles. The first row contains "Health Coach" and "Health Coach!". The second row contains "Administrative Assistant" and "Administrative Officer". The third row contains "Account Executive", "Account Representative", "Account Representative", "Accountant", and "Accounting Assistant".
- Column 3:** A header row with the text "Replace values:". Below this, there are three rows of drop-down menus. The first row shows "Health Coach". The second row shows "Administrative Assistant" with a red callout "2" pointing to the drop-down arrow. The third row shows "Accountant".

At the bottom of the interface is a yellow "SUBMIT" button with a red callout "5" pointing to it.