User Guide

Data Preparation- 4.4

Contents

1. About this Guide

1.1. Document History

Product Version	Date (Release date)	Description
Data Preparation 4.0	December 31st, 2018	First Release of the document
Data Preparation 4.2	March 25th, 2019	Updated document
Data Preparation 4.3	April 24th, 2019	Updated document
Data Preparation 4.4	June 7th, 2019	Updated document

1.2. Overview

This guide covers:

- Explanation and usage of all the Data Preparation options
- Explanation and usage of the Transforms
- Integration with Data Pipeline

1.3. Target Audience

This guide is aimed at users who wish to use BDB Data Preparation option to prepare and transform their business data.

2. Introduction

2.1. Introducing the Data Preparation

The Data Preparation module can turn any Business data into a cost-effective and custom-made experience. The Data Analysts can instantly detect anomalous records (rows with invalid or empty values) and purge the unwanted data sets in a few clicks using Machine-Learning based smart techniques and sampling. The users can identify errors and apply changes to data set from any source and export the analysis ready data in minutes. Automated detection of groups and categories in your data can be viewed through a frequency table. The user can filter the group in a single click and transform data matching the filter conditions and get intelligent Data Transformation suggestions based on data type and quality.

2.2. Supported Web Browsers

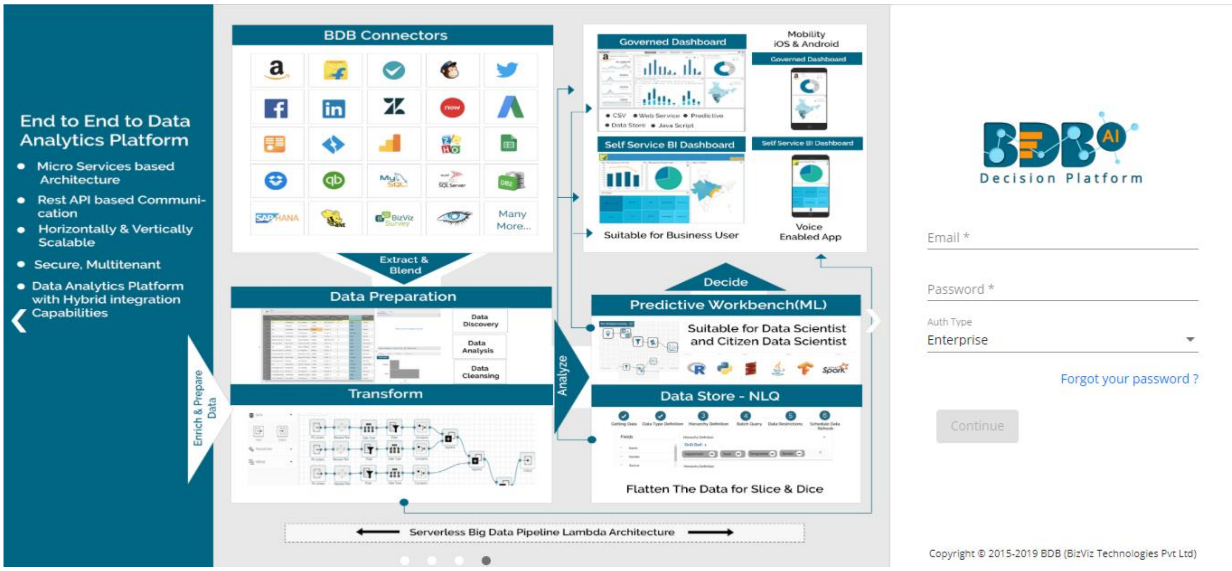
The BDB Platform is a web browser-based application. The users can run the BDB Platform and its various plugins on the below given versions of the browsers:

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

3. Getting Started with BDB Data Preparation

This section covers initial steps to access the BDB Dashboard Designer plugin using the BDB Platform.

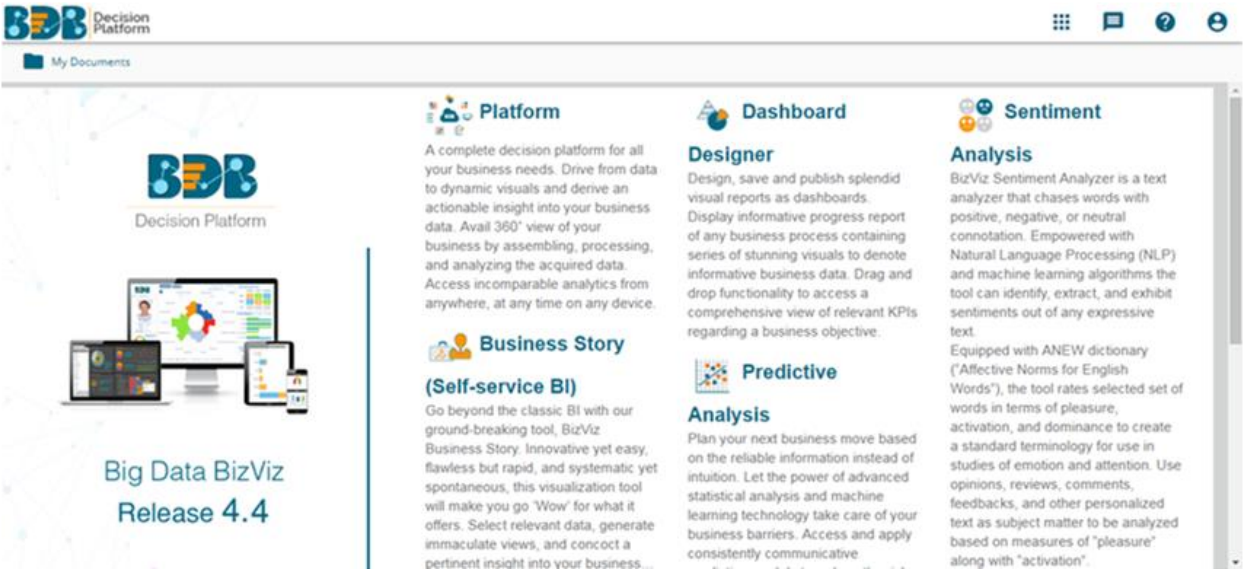
i) Open the BDB Enterprise Platform Link: https://app.bdb.ai
ii) Enter your credentials to log in to the platform.
iii) Click the ‘**Continue**’ option.



End to End to Data Analytics Platform

- Micro Services based Architecture
- Rest API based Communication
- Horizontally & Vertically Scalable
- Secure, Multitenant
- Data Analytics Platform with Hybrid integration Capabilities

iv) BDB Platform homepage opens (The below page appears only for the first time when the user login. Once the user creates some document, he gets directed to the homepage by default).



Platform

A complete decision platform for all your business needs. Drive from data to dynamic visuals and derive an actionable insight into your business data. Avail 360° view of your business by assembling, processing, and analyzing the acquired data. Access incomparable analytics from anywhere, at any time on any device.

Business Story (Self-service BI)

Go beyond the classic BI with our ground-breaking tool, BizViz Business Story. Innovative yet easy, flawless but rapid, and systematic yet spontaneous, this visualization tool will make you go 'Wow' for what it offers. Select relevant data, generate immaculate views, and concoct a pertinent insight into your business...

Dashboard Designer

Design, save and publish splendid visual reports as dashboards. Display informative progress report of any business process containing series of stunning visuals to denote informative business data. Drag and drop functionality to access a comprehensive view of relevant KPIs regarding a business objective.

Predictive Analysis

Plan your next business move based on the reliable information instead of intuition. Let the power of advanced statistical analysis and machine learning technology take care of your business barriers. Access and apply consistently communicative

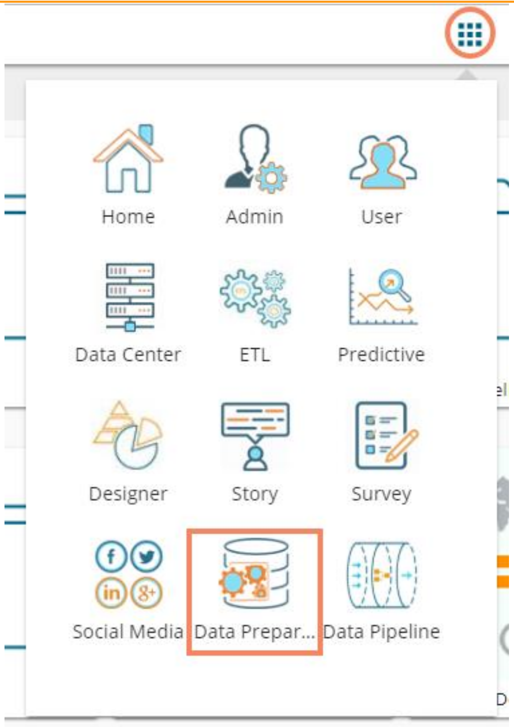
Sentiment Analysis

BizViz Sentiment Analyzer is a text analyzer that chases words with positive, negative, or neutral connotation. Empowered with Natural Language Processing (NLP) and machine learning algorithms the tool can identify, extract, and exhibit sentiments out of any expressive text.

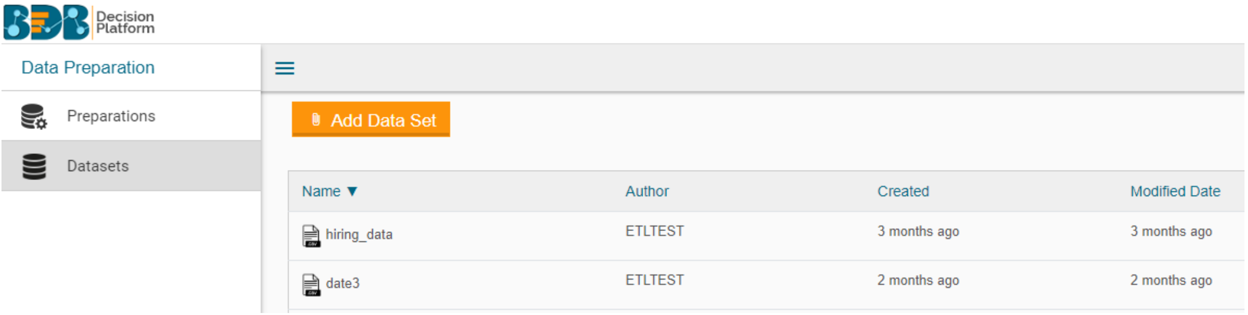
Equipped with ANEW dictionary ("Affective Norms for English Words"), the tool rates selected set of words in terms of pleasure, activation, and dominance to create a standard terminology for use in studies of emotion and attention. Use opinions, reviews, comments, feedbacks, and other personalized text as subject matter to be analyzed based on measures of "pleasure" along with "activation".

Note: The above screen opens only for those newly created users who have not yet created any document/folder using the BDB Platform.

v) Click on the ‘**App**’ menu button.
vi) Select the ‘**Data Preparation**’ plugin from the app menu.



vii) The Data Preparation landing page opens displaying the Datasets tab (by default)



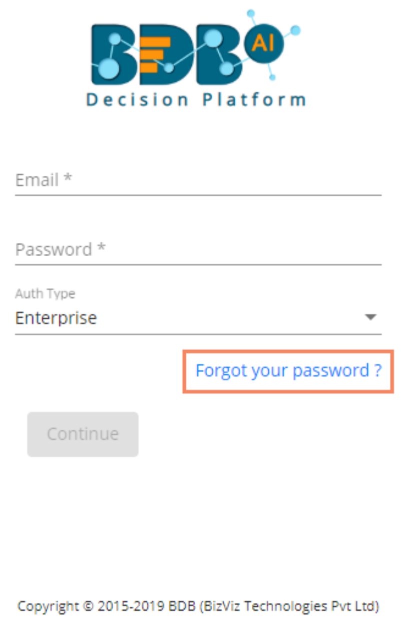
Name ▼	Author	Created	Modified Date
hiring_data	ETLTEST	3 months ago	3 months ago
date3	ETLTEST	2 months ago	2 months ago

3.1. Forgot Password Option

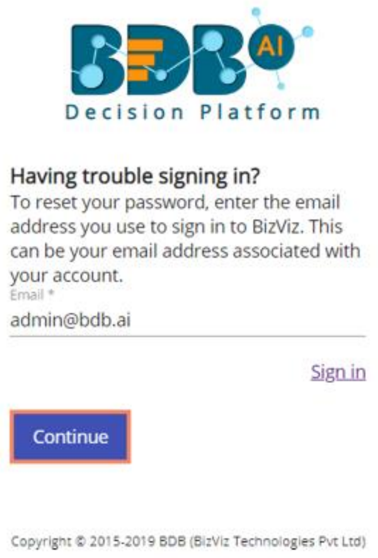
Users are provided with a choice to change the password on the Login page of the platform.

i) Navigate to the login page of the BDB Platform.

ii) Click the '**Forgot your password?**' option.



iii) Users get redirected to a new window.
iv) Provide the email id that is registered with BDB to send the reset password link.
v) Click the '**Continue**' option.

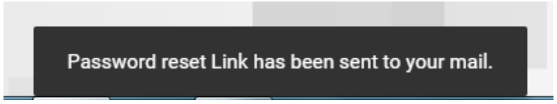


Having trouble signing in?

To reset your password, enter the email address you use to sign in to BizViz. This can be your email address associated with your account.

Email *
admin@bdb.ai

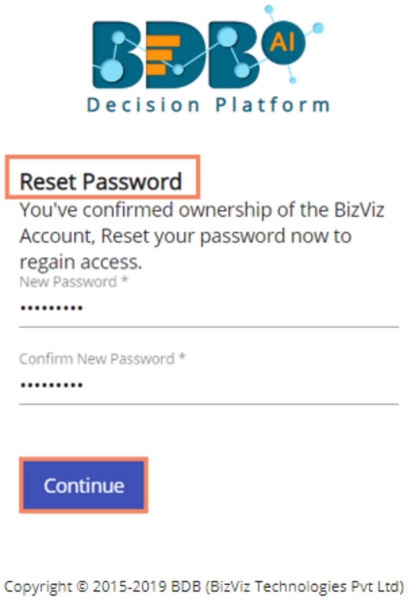
vi) Users may be redirected to select a space in case of multiple areas under one server link; they need to choose a space and click the '**Continue**' option once again. Otherwise, a message will pop-up to notify that the password reset link has been sent to the registered email.



Password reset Link has been sent to your mail.

vii) Click the link from your registered email.
viii) Users get redirected to the '**Reset Password**' page to set a new password.

ix) Set a new password.
x) Confirm the newly set password.
xi) Click the ‘**Continue’** option.



Decision Platform

Reset Password

You've confirmed ownership of the BizViz Account, Reset your password now to regain access.

New Password * ·········

Confirm New Password * ·········

Continue

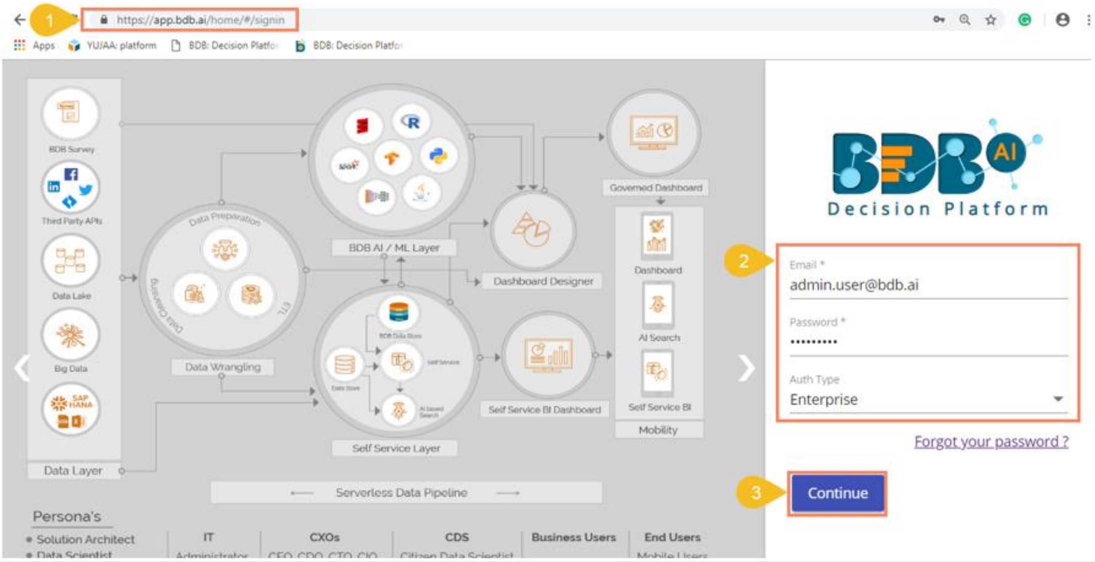
Copyright © 2015-2019 BDB (BizViz Technologies Pvt Ltd)

xii) The new password gets updated for the selected BDB account, and the user gets redirected back to the ‘**Log In**’ page of the BDB Platform.

3.2. Force Login

The ‘**Force Login**’ functionality has been introduced to control the number of active sessions up to three. The users can access only 3 sessions at a time when they try to access 4th session a warning message displays to inform that the user has consumed the permitted sessions and a click on the ‘**Force Login**’ would kill all those active sessions.

i) Navigate to the BDB Platform Login page.
ii) Enter the valid credentials to log in.
iii) Click the ‘**Continue**’ option.



Email * admin.user@bdb.ai

Password * ·········

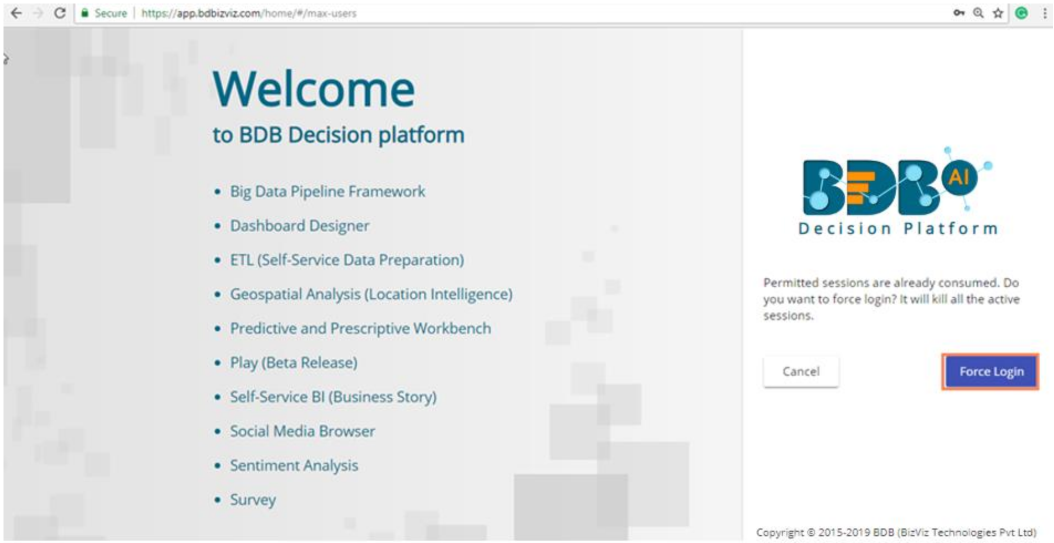
Auth Type Enterprise

Forgot your password ?

Continue

iv) The user will get the following message if the user already consumes the permitted active sessions (3 sessions at a time).

v) Click the '**Force Login**' option.



https://app.bdbizviz.com/home/#/max-users

Welcome
to BDB Decision platform

- Big Data Pipeline Framework
- Dashboard Designer
- ETL (Self-Service Data Preparation)
- Geospatial Analysis (Location Intelligence)
- Predictive and Prescriptive Workbench
- Play (Beta Release)
- Self-Service BI (Business Story)
- Social Media Browser
- Sentiment Analysis
- Survey

Permitted sessions are already consumed. Do you want to force login? It will kill all the active sessions.

Copyright © 2015-2019 BDB (BizViz Technologies Pvt Ltd)

vi) A warning message appears that the currently active sessions get killed for the user and the user has redirected to the log in a page of the BDB Platform.

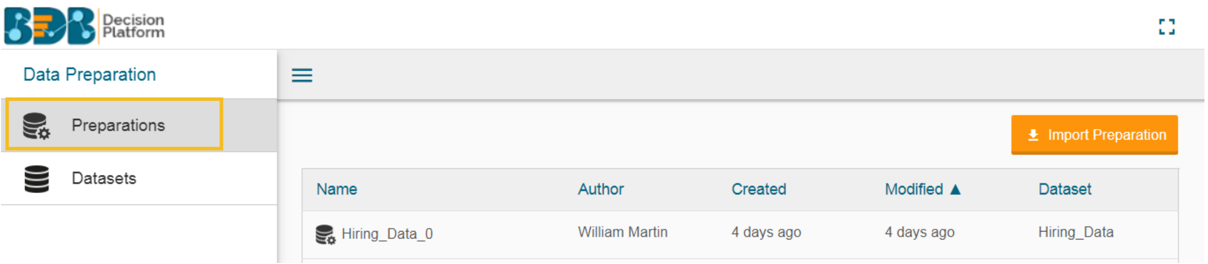
Note: The user can successfully login to the BDB Platform after selecting the '**Force Login'** option to log in the platform.

4. Data Preparation Landing Page

The landing page of the data preparation has two menus: 1) Preparations and 2) Datasets. The user can start the data preparation process by uploading a dataset, and the newly created preparation gets saved under the '**Preparations**' tab.

4.1. Preparations

The '**Preparations**' tab lists all the available preparations displaying Name, Author, when it was created, when it was last modified and using which data set it was created.



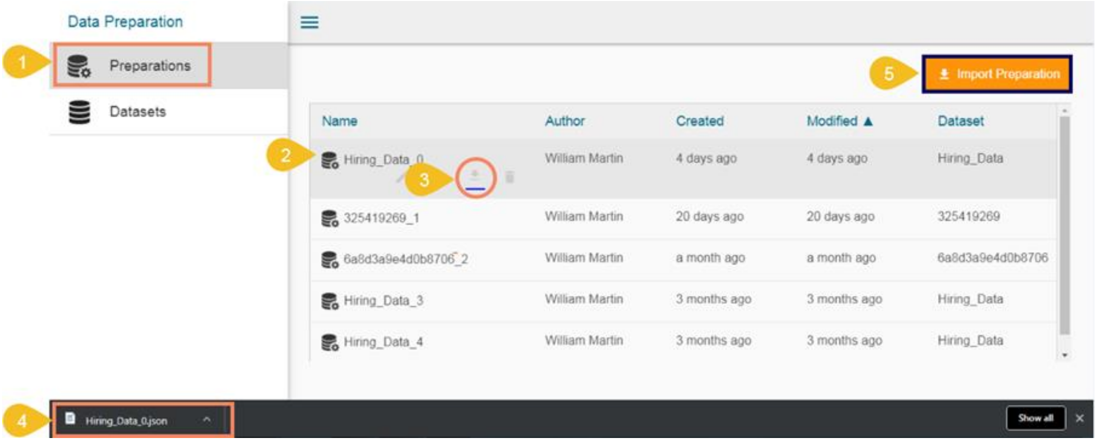
Name	Author	Created	Modified ▲	Dataset
Hiring_Data_0	William Martin	4 days ago	4 days ago	Hiring_Data

The users can continue adding more steps to the existing preparations. The user can import an existing preparation using the '**Import Preparation**' option.

4.1.1. Importing a Preparation

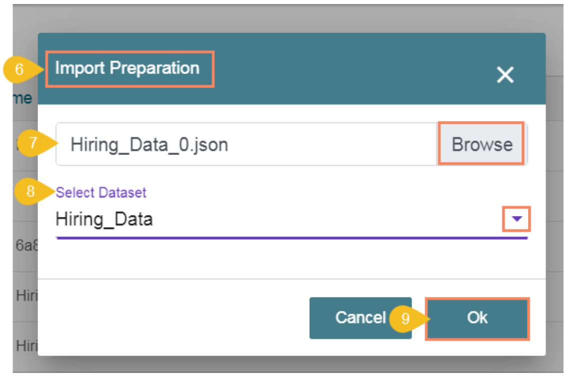
This feature can be used to apply a set of cleansing steps on a dataset with similar metadata.

i) Navigate to the Preparations list.
ii) Select a preparation from the list.
iii) Click the ‘**download**’ icon for the preparation.
iv) The selected Preparation gets downloaded in a json file.
v) Click the ‘**Import Preparation**’ option after the json file of the preparation gets downloaded.



Name	Author	Created	Modified ▲	Dataset
Hiring_Data_0	William Martin	4 days ago	4 days ago	Hiring_Data
325419269_1	William Martin	20 days ago	20 days ago	325419269
6a8d3a9e4d0b8706_2	William Martin	a month ago	a month ago	6a8d3a9e4d0b8706
Hiring_Data_3	William Martin	3 months ago	3 months ago	Hiring_Data
Hiring_Data_4	William Martin	3 months ago	3 months ago	Hiring_Data

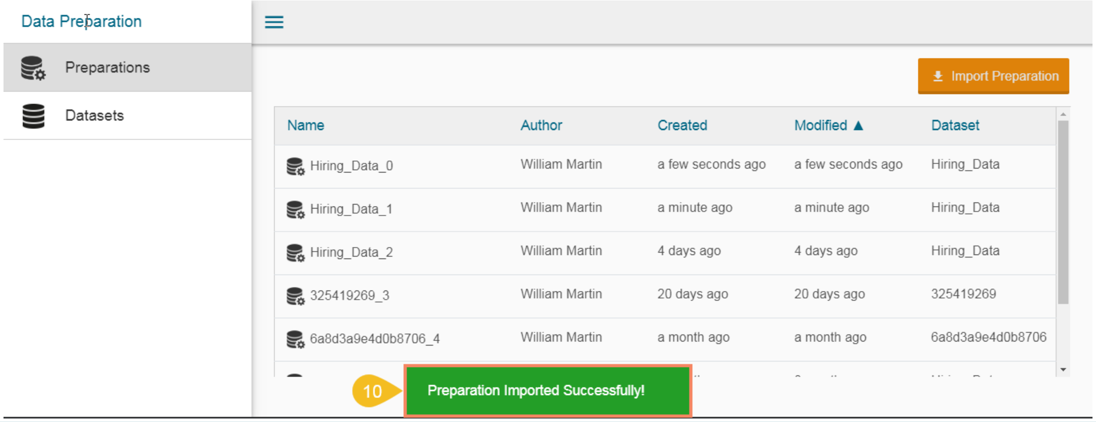
vi) The ‘**Import Preparation**’ window opens.
vii) Browse the downloaded json file.
viii) Select a dataset of similar metadata.
ix) Click the ‘**Ok**’ option.



Hiring_Data_0.json

Select Dataset
Hiring_Data

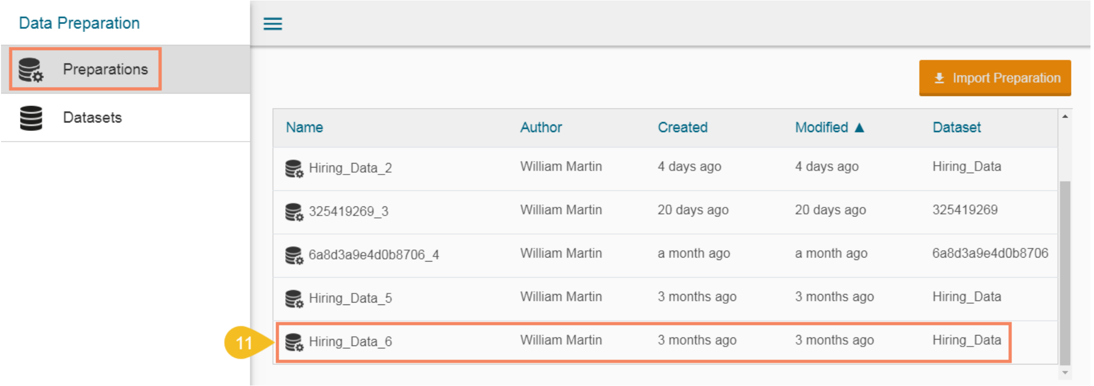
x) A success message appears.



Name	Author	Created	Modified ▲	Dataset
Hiring_Data_0	William Martin	a few seconds ago	a few seconds ago	Hiring_Data
Hiring_Data_1	William Martin	a minute ago	a minute ago	Hiring_Data
Hiring_Data_2	William Martin	4 days ago	4 days ago	Hiring_Data
325419269_3	William Martin	20 days ago	20 days ago	325419269
6a8d3a9e4d0b8706_4	William Martin	a month ago	a month ago	6a8d3a9e4d0b8706

Preparation Imported Successfully!

xi) The Preparation gets imported and applied to the selected dataset.



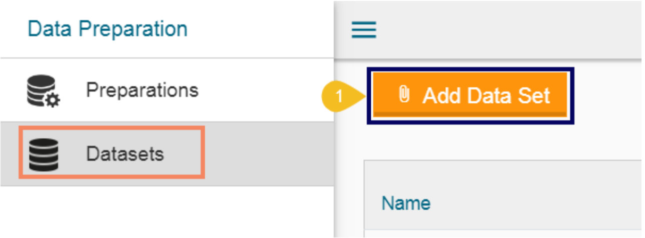
Name	Author	Created	Modified ▲	Dataset
Hiring_Data_2	William Martin	4 days ago	4 days ago	Hiring_Data
325419269_3	William Martin	20 days ago	20 days ago	325419269
6a8d3a9e4d0b8706_4	William Martin	a month ago	a month ago	6a8d3a9e4d0b8706
Hiring_Data_5	William Martin	3 months ago	3 months ago	Hiring_Data
Hiring_Data_6	William Martin	3 months ago	3 months ago	Hiring_Data

4.2. Datasets

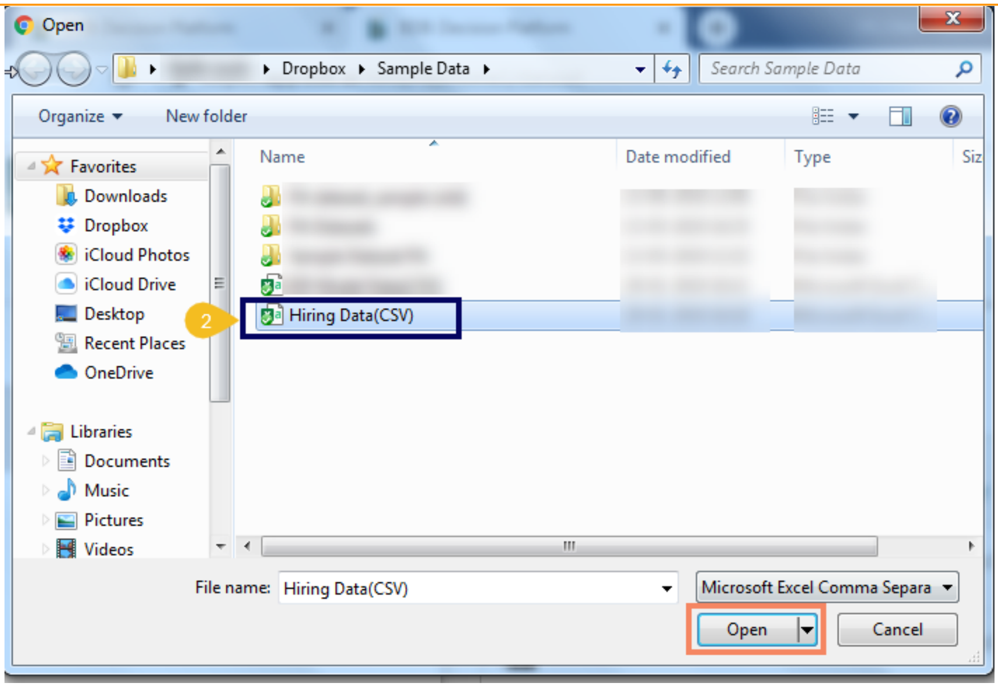
The '**Datasets**' section lists the data/inputs added to the system. The user can create a new preparation by selecting any of the listed datasets. The Datasets window also provides an option to add new datasets.

4.2.1. Adding a new Dataset

i) Click the '**Add Data Set**' icon.



ii) A new window opens redirecting the user to select a CSV Data set.



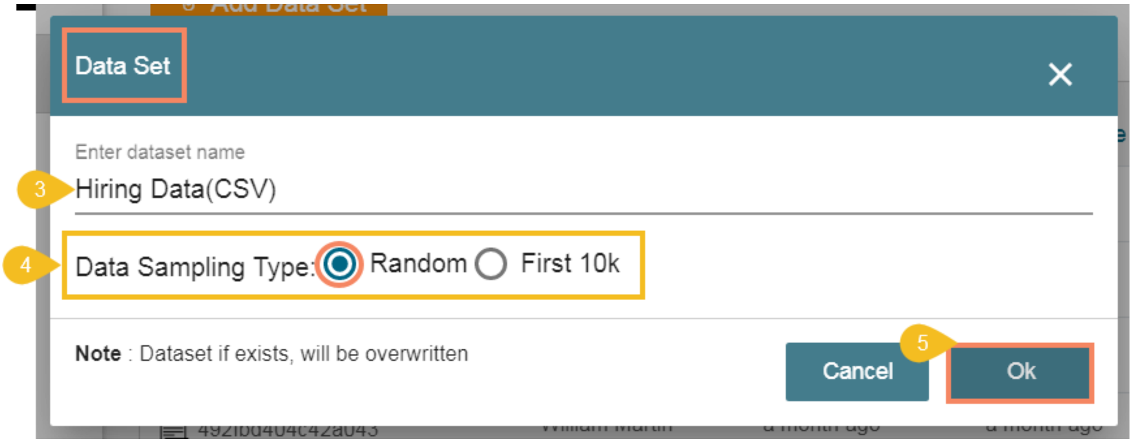
File name: Hiring Data(CSV)

Microsoft Excel Comma Separa

iii) After selecting a CSV dataset, it displays a '**Data Set**' window with the selected CSV file.

iv) The user can select a **Data Sampling Type** by marking the radio button.

v) Click the '**Ok**' option.



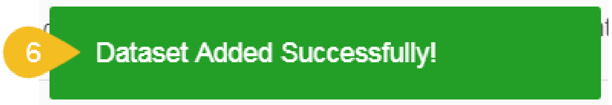
Enter dataset name

Hiring Data(CSV)

Data Sampling Type: ● Random ○ First 10k

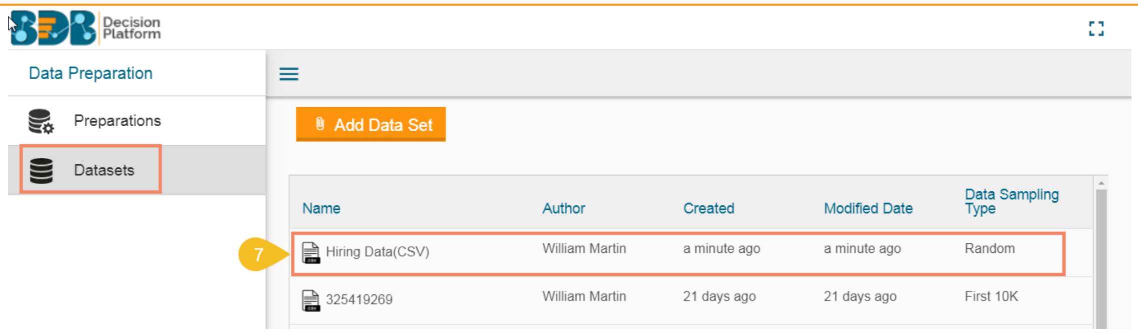
Note : Dataset if exists, will be overwritten

vi) A success message appears.



Dataset Added Successfully!

vii) The selected CSV File gets added to the Datasets page.



Name	Author	Created	Modified Date	Data Sampling Type
Hiring Data(CSV)	William Martin	a minute ago	a minute ago	Random
325419269	William Martin	21 days ago	21 days ago	First 10K

Note: The standalone version of the Data Preparation supports only CSV input of max 10k records. To work on other data sources and colossal volume, please use the ETL integrated version of data cleansing.

5. Data Grid

The data grid in the BDB Data Preparation is used for visualizing the data. The data displayed in the grid is a sample from the actual data set or complete data based on the data volume. The grid always shows the first 10 K rows in the dataset.

The displayed data in the grid changes based on the number of transforms performed on it.

5.1. Data Grid Header

The grid has a header which displays the column name from the dataset. The context menu in the header has an option to rename the column and delete the column. It also presents the data type of the column. It is analyzed based on the max match to any data type in the first 10K records.

Consider that a 10000 rows sample has 9000 integers and 1000 string values, the selected data type is Integer. Moreover, the 1000 rows get detected as invalid rows.

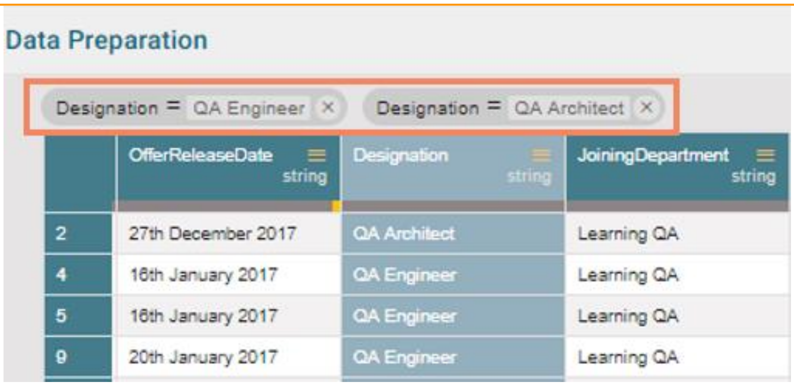
5.2. Data Types

The BDB Data Preparation supports the following data types:

1. Integer
2. Double
3. String
4. Date
5. Timestamp

5.3. Panel to List the Selected Filters.

When a filter is selected, it gets added to the filter panel on top of the grid. The added filter has an option to remove it by clicking the '**Close**' (X) mark.

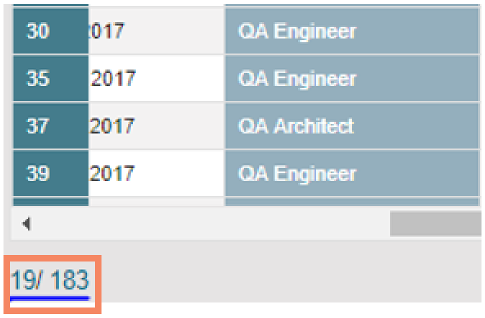


Data Preparation

Designation = QA Engineer × Designation = QA Architect ×

	OfferReleaseDate string	Designation string	JoiningDepartment string
2	27th December 2017	QA Architect	Learning QA
4	16th January 2017	QA Engineer	Learning QA
5	16th January 2017	QA Engineer	Learning QA
9	20th January 2017	QA Engineer	Learning QA

The left bottom of the grid displays the number of rows meeting the filter condition out of the total.



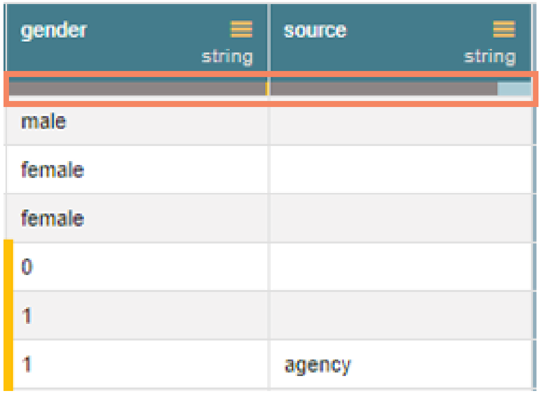
30	017	QA Engineer
35	2017	QA Engineer
37	2017	QA Architect
39	2017	QA Engineer

19/ 183

5.4. Data Quality Bar in the Grid

A Data Quality Bar appears in the header of the grid. The Data Quality is indicated through color coding, as explained below:

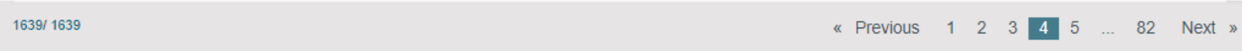
- Brown-Valid Data
- Orange– Invalid data
- Light blue -Blank data



gender string	source string
male	
female	
female	
0	
1	
1	agency

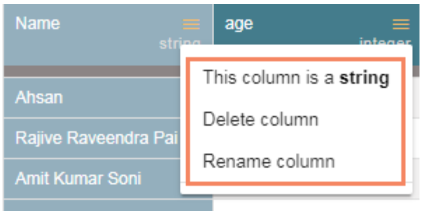
5.5. Pagination

Pagination is implemented for the grid data. The tool displays 20 records on each page. The maximum rows displayed for sampling is always 10k.



1639/ 1639 « Previous 1 2 3 4 5 ... 82 Next »

Note: The users can get information about the Column Type, option to Delete the column and option to Rename the column by clicking the ‘**Column Menu’** ≡ icon provided next to the column names in the data grid.

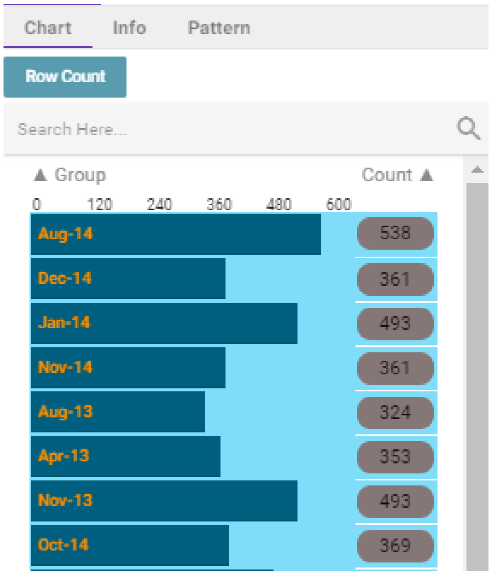


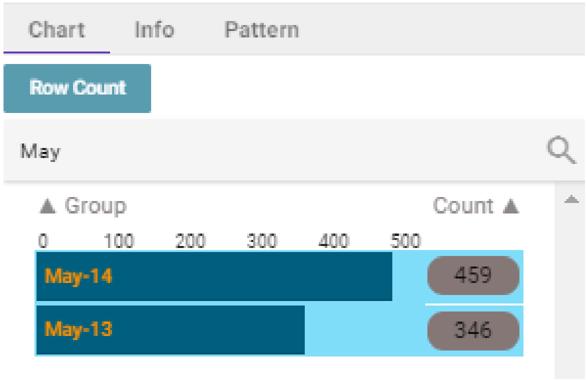
6. Summary Pane

The summary pane gives an overview of the data like different patterns of data, distinct values, and occurrences.

6.1. Charts

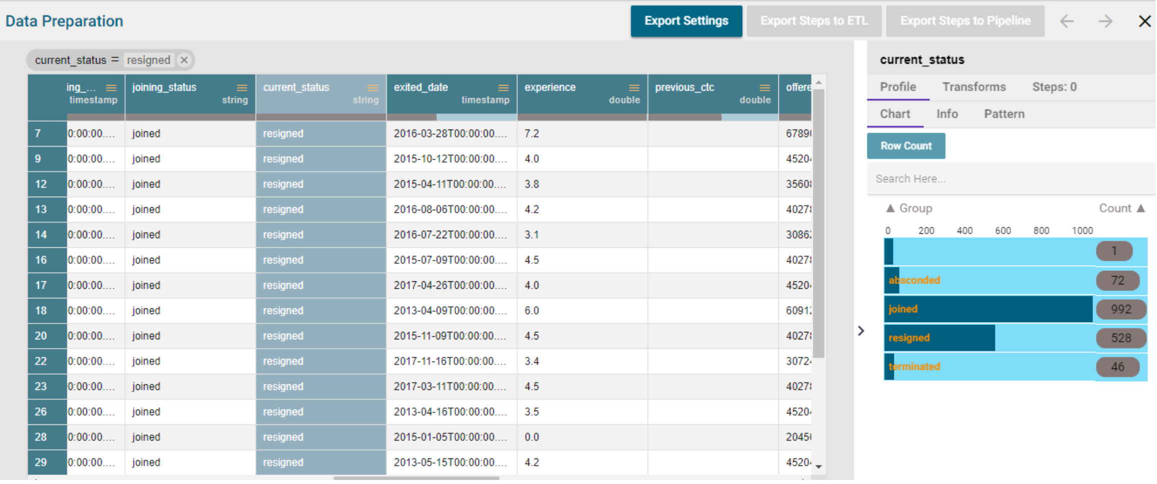
The in-built charts (Column and Bar charts) display the occurrence of each value. The Bar appears to display string value. The Column chart projects numeric value columns and dates.





The graph is interactive. When the user clicks on any bar, it adds a filter in the filter pane and filters the data displayed in the grid. Later the transform can be performed on the filtered data.

The chart can be sorted based on the group or the count of occurrence of a group.



Data Preparation

current_status = resigned

	ing_... timestamp	joining_status string	current_status string	exited_date timestamp	experience double	previous_ctc double	offere
7	0:00:00....	joined	resigned	2016-03-28T00:00:00....	7.2		6789
9	0:00:00....	joined	resigned	2015-10-12T00:00:00....	4.0		4520
12	0:00:00....	joined	resigned	2015-04-11T00:00:00....	3.8		3560
13	0:00:00....	joined	resigned	2016-08-06T00:00:00....	4.2		4027
14	0:00:00....	joined	resigned	2016-07-22T00:00:00....	3.1		3086
16	0:00:00....	joined	resigned	2015-07-09T00:00:00....	4.5		4027
17	0:00:00....	joined	resigned	2017-04-26T00:00:00....	4.0		4520
18	0:00:00....	joined	resigned	2013-04-09T00:00:00....	6.0		6091
20	0:00:00....	joined	resigned	2015-11-09T00:00:00....	4.5		4027
22	0:00:00....	joined	resigned	2017-11-16T00:00:00....	3.4		3072
23	0:00:00....	joined	resigned	2017-03-11T00:00:00....	4.5		4027
26	0:00:00....	joined	resigned	2013-04-16T00:00:00....	3.5		4520
28	0:00:00....	joined	resigned	2015-01-05T00:00:00....	0.0		2045
29	0:00:00....	joined	resigned	2013-05-15T00:00:00....	4.2		4520

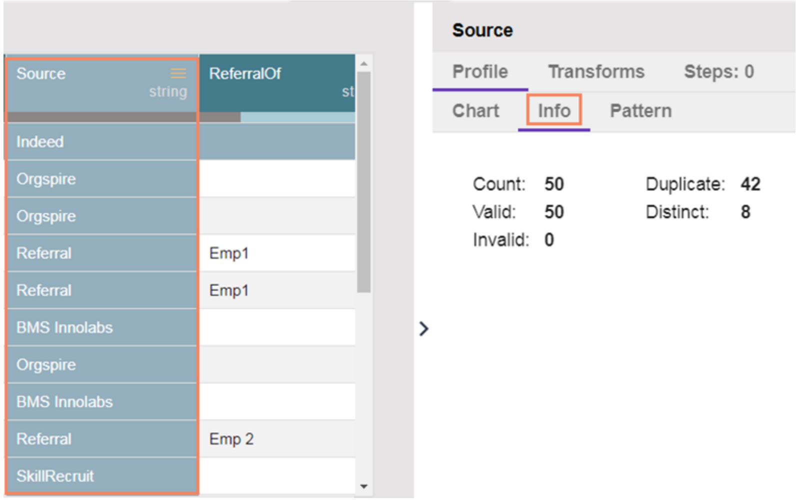
current_status

▲ Group	Count ▲
	1
absconded	72
joined	992
resigned	528
terminated	46

6.2. Info: Value/Statistics

The information tab displays value or statistics of the data. The following aspects are displayed about the chosen data when the column is of string type:

- Count of Rows
- Count of Duplicates
- Count of Valid Data
- Distinct Values
- Count of Invalid Data



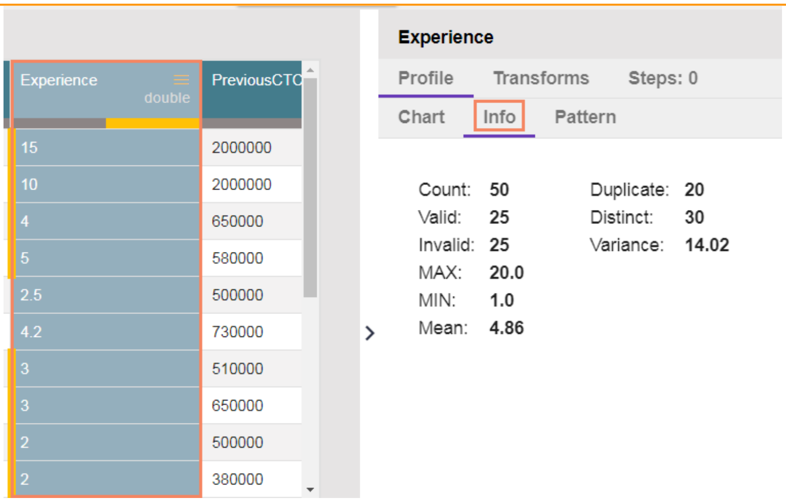
Source string	ReferralOf st
Indeed	
Orgspire	
Orgspire	
Referral	Emp1
Referral	Emp1
BMS Innolabs	
Orgspire	
BMS Innolabs	
Referral	Emp 2
SkillRecruit	

Source

Count: 50 Duplicate: 42
Valid: 50 Distinct: 8
Invalid: 0

When the selected column is of numeric type, the additional displayed information under the 'Info' tab is based on aggregation functions as mentioned below:

- Minimum
- Maximum
- Mean
- Variance

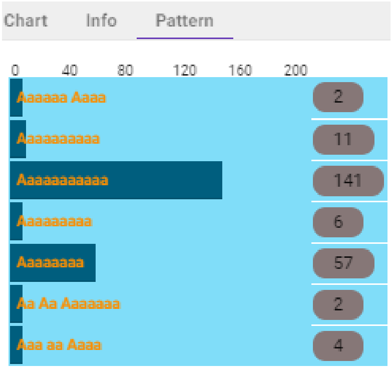


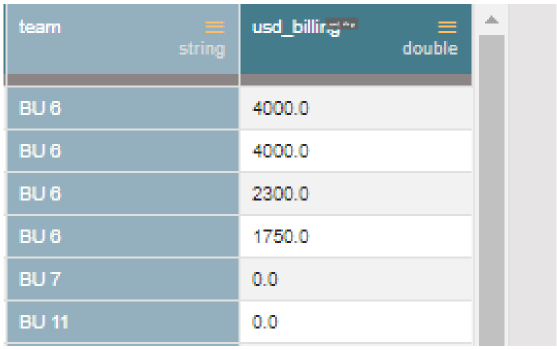
Experience double	PreviousCTC
15	2000000
10	2000000
4	650000
5	580000
2.5	500000
4.2	730000
3	510000
3	650000
2	500000
2	380000

Count:	50	Duplicate:	20
Valid:	25	Distinct:	30
Invalid:	25	Variance:	14.02
MAX:	20.0		
MIN:	1.0		
Mean:	4.86		

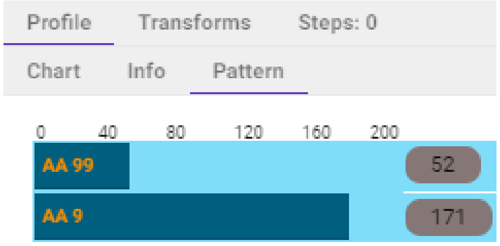
6.3. Pattern

This section focuses on how data pattern and occurrences of each pattern in the dataset sample get plotted in a chart.

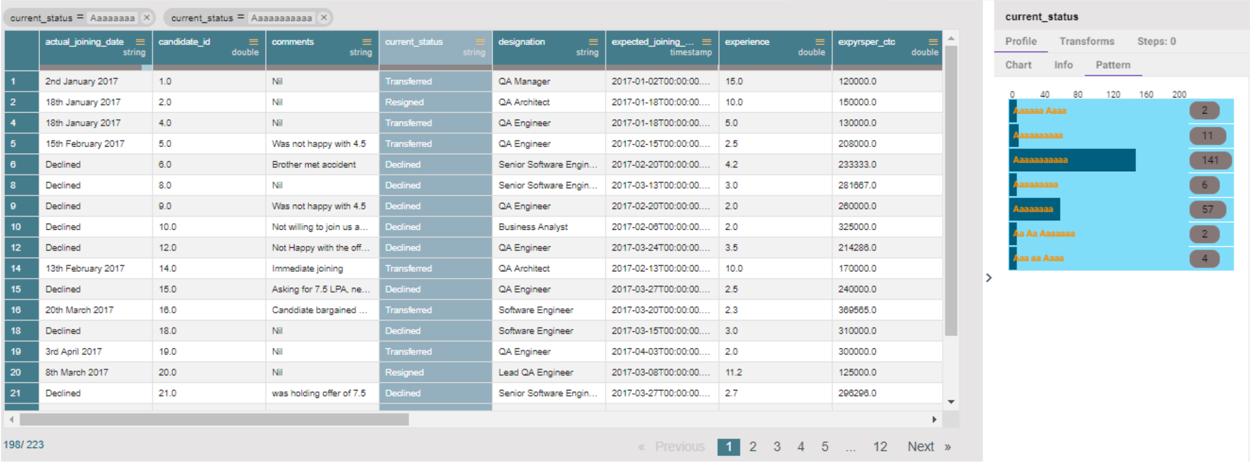




team string	usd_billing double
BU 6	4000.0
BU 6	4000.0
BU 6	2300.0
BU 6	1750.0
BU 7	0.0
BU 11	0.0



Note: The value displayed is not the actual value, and it's just a pattern of the value.



	actual_joining_date string	candidate_id double	comments string	current_status string	designation string	expected_joining_... timestamp	experience double	expyrsper_ctc double
1	2nd January 2017	1.0	Nil	Transferred	QA Manager	2017-01-02T00:00:00....	15.0	120000.0
2	18th January 2017	2.0	Nil	Resigned	QA Architect	2017-01-18T00:00:00....	10.0	150000.0
4	18th January 2017	4.0	Nil	Transferred	QA Engineer	2017-01-18T00:00:00....	5.0	130000.0
5	15th February 2017	5.0	Was not happy with 4.5	Transferred	QA Engineer	2017-02-15T00:00:00....	2.5	208000.0
6	Declined	6.0	Brother met accident	Declined	Senior Software Engin...	2017-02-20T00:00:00....	4.2	233333.0
8	Declined	8.0	Nil	Declined	Senior Software Engin...	2017-03-13T00:00:00....	3.0	281667.0
9	Declined	9.0	Was not happy with 4.5	Declined	QA Engineer	2017-02-20T00:00:00....	2.0	260000.0
10	Declined	10.0	Not willing to join us a...	Declined	Business Analyst	2017-02-06T00:00:00....	2.0	325000.0
12	Declined	12.0	Not Happy with the off...	Declined	QA Engineer	2017-03-24T00:00:00....	3.5	214286.0
14	13th February 2017	14.0	Immediate joining	Transferred	QA Architect	2017-02-13T00:00:00....	10.0	170000.0
15	Declined	15.0	Asking for 7.5 LPA, ne...	Declined	QA Engineer	2017-03-27T00:00:00....	2.5	240000.0
16	20th March 2017	16.0	Canddiate bargained ...	Transferred	Software Engineer	2017-03-20T00:00:00....	2.3	369565.0
18	Declined	18.0	Nil	Declined	Software Engineer	2017-03-15T00:00:00....	3.0	310000.0
19	3rd April 2017	19.0	Nil	Transferred	QA Engineer	2017-04-03T00:00:00....	2.0	300000.0
20	8th March 2017	20.0	Nil	Resigned	Lead QA Engineer	2017-03-08T00:00:00....	11.2	125000.0
21	Declined	21.0	was holding offer of 7.5	Declined	Senior Software Engin...	2017-03-27T00:00:00....	2.7	296296.0

198/ 223

Aaaaaa Aaaa	2
Aaaaaaaaaa	11
Aaaaaaaaaaa	141
Aaaaaaaaa	6
Aaaaaaaa	57
Aa Aa Aaaaaaa	2
Aaa aa Aaaa	4

6.4. Transforms

Data Preparation module provides a list of transforms that can be performed on the data to clean/prepare the data for insightful visualization.

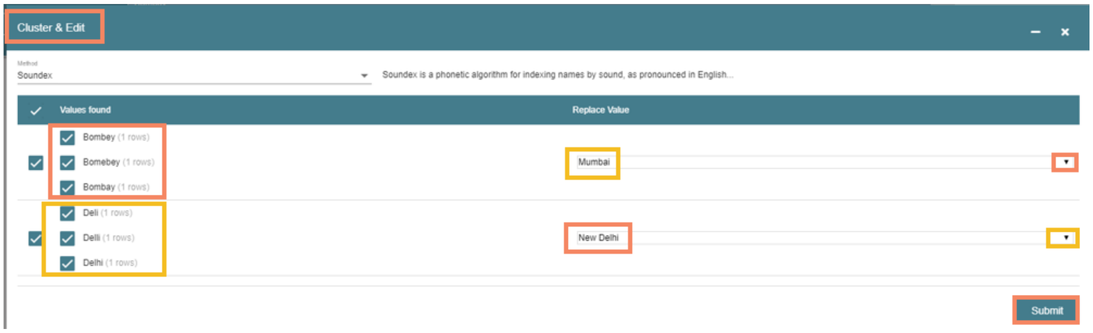
This section explains the details of the transforms.

6.4.1. Advanced

6.4.1.1. Cluster & Edit

The '**Cluster & Edit**' transform when applied groups the words with similar phonetic (Speech sound/Pronunciation) into a cluster. The user can apply this transform to replace function on that bucket to replace all those words at once. We can also exclude some value when replacing it with the new value. It works on the Soundex algorithm to cluster the data.

When the Cluster & Edit transform gets applied as follows:

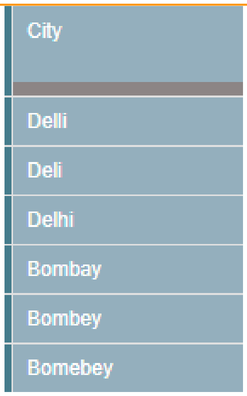


Method
Soundex

Soundex is a phonetic algorithm for indexing names by sound, as pronounced in English...

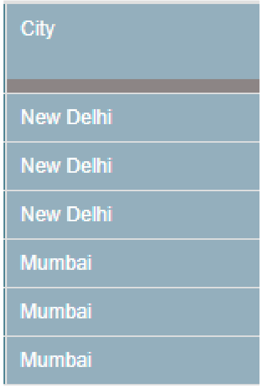
✓	Values found	Replace Value
☑	☑ Bombey (1 rows) ☑ Bomebey (1 rows) ☑ Bombay (1 rows)	Mumbai ▾
☑	☑ Deli (1 rows) ☑ Delli (1 rows) ☑ Delhi (1 rows)	New Delhi ▾

The existing column '**City**' with the following Phonetic variations:



City
Delli
Deli
Delhi
Bombay
Bombey
Bomebey

it gets converted into



City
New Delhi
New Delhi
New Delhi
Mumbai
Mumbai
Mumbai

6.4.1.2. Expression Editor

The Expression Editor transform has a collection of different function to manipulate the data like absolute, to date, from Unix time.

i) Select the ‘**Expression Editor**’ transform option using the ‘**Transforms**’ tab

ii) The Expression Editor window opens with the following information:

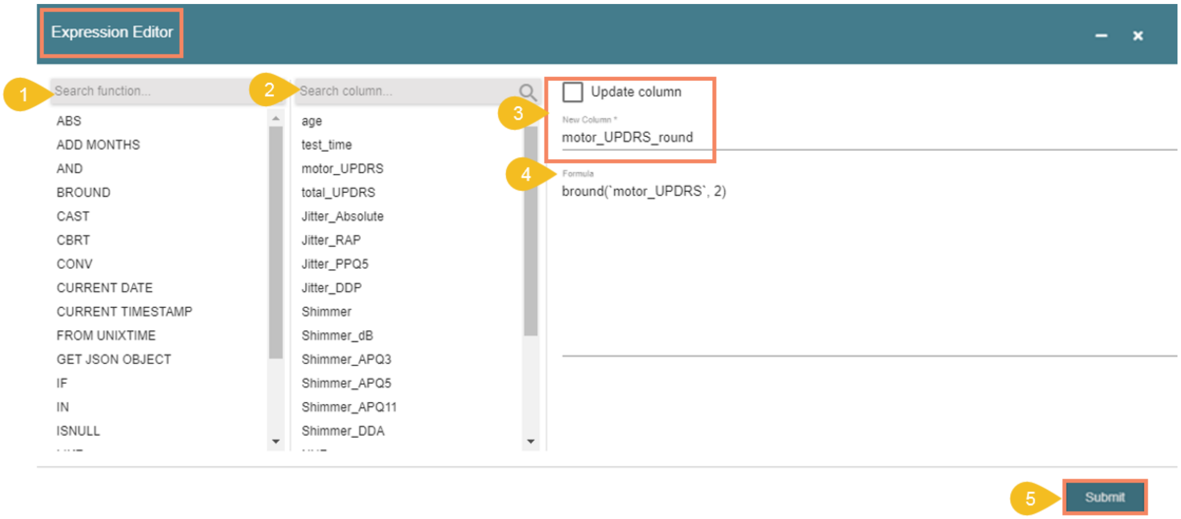
a. Search Function- Use double click to search/select a function from the displayed list. The selected function appears under the ‘**Formula**’ space.

b. Search Column- Use double click to search/select a column from the displayed list. The selected column appears under the ‘**Formula**’ space.

c. The user can select an existing column by enabling the ‘**Update column**’ option or create a New Column by entering the column name for the new column.

d. The selected function and column appear under the ‘**Formula**’ space.

e. Click the ‘**Submit**’ option.



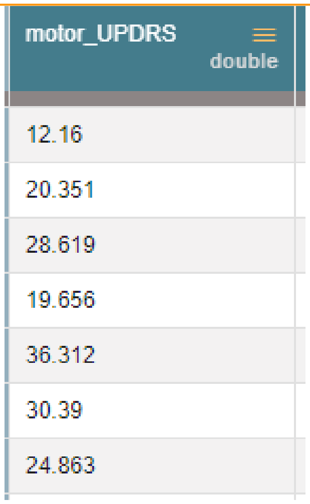
☐ Update column

New Column *: motor_UPDRS_round

Formula: bround(`motor_UPDRS`, 2)

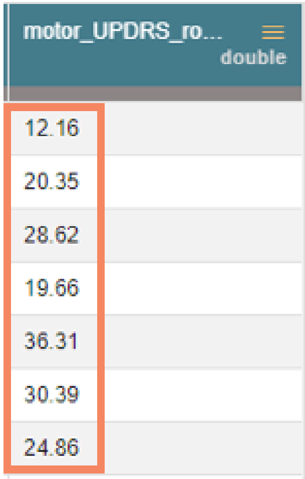
The new column gets added to the data grid with the updated data based on the applied formula.

The original data



motor_UPDRS double
12.16
20.351
28.619
19.656
36.312
30.39
24.863

gets converted into

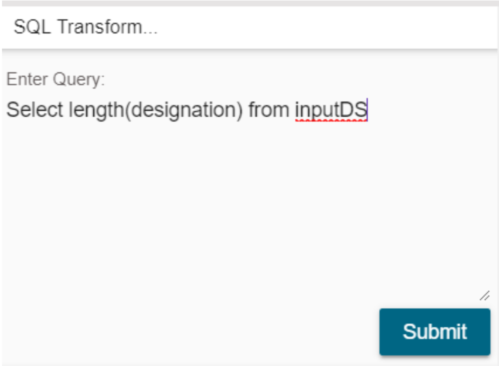


motor_UPDRS_ro... double
12.16
20.35
28.62
19.66
36.31
30.39
24.86

Note: In case of selecting an existing column, the data gets updated as per the applied formula in the column.

6.4.1.3. SQL Transform

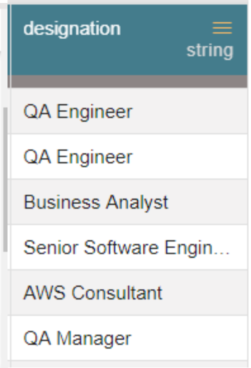
This transform allows the user to write SQL Query against the table as we can write in any SQL editor. This transform requires the table name to be mentioned as '**InputDS**' in the query.



SQL Transform...

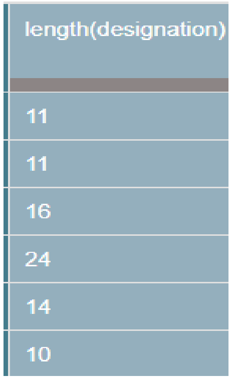
Enter Query:

```
Select length(designation) from inputDS
```



designation string
QA Engineer
QA Engineer
Business Analyst
Senior Software Engin...
AWS Consultant
QA Manager

gives



length(designation)
11
11
16
24
14
10

6.4.2. Columns

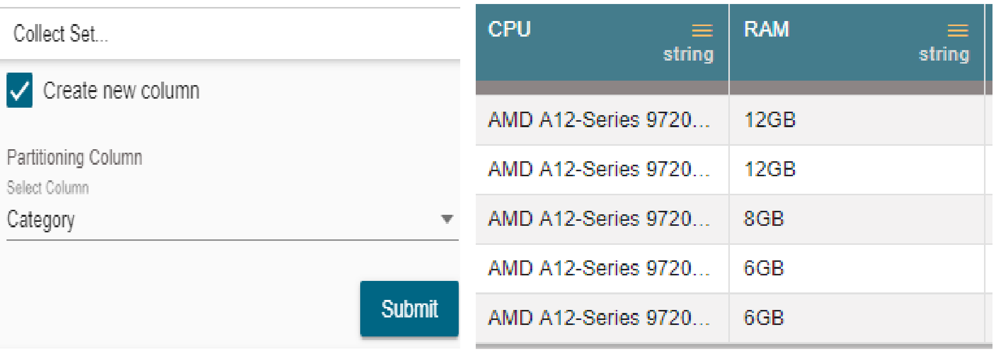
6.4.2.1. Cast to Types

It is a table-based operation. The profiling of a column is done based on the data type present in the majority. Let's say in column A; we have four integer value and one string value, then the data type of column gets profiled as the integer despite one string value present in it. The 'Cast to Types' transform removes the value with the invalid data type. In this case, it converts data with a string data type to the null value.

**Note: *Cast to types is a lossy transformation. There is a possibility of some data loss.*

6.4.2.2. Collect Set

The 'Collect Set' transform generates the list of all the unique values of the column based on the selected column. It performs group concatenation.

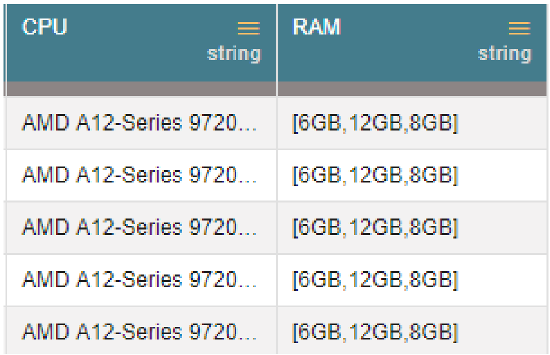


Collect Set...

☑ Create new column

Partitioning Column
Select Column
Category

CPU string	RAM string
AMD A12-Series 9720...	12GB
AMD A12-Series 9720...	12GB
AMD A12-Series 9720...	8GB
AMD A12-Series 9720...	6GB
AMD A12-Series 9720...	6GB



CPU string	RAM string
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]
AMD A12-Series 9720...	[6GB,12GB,8GB]

generates the list of all unique value

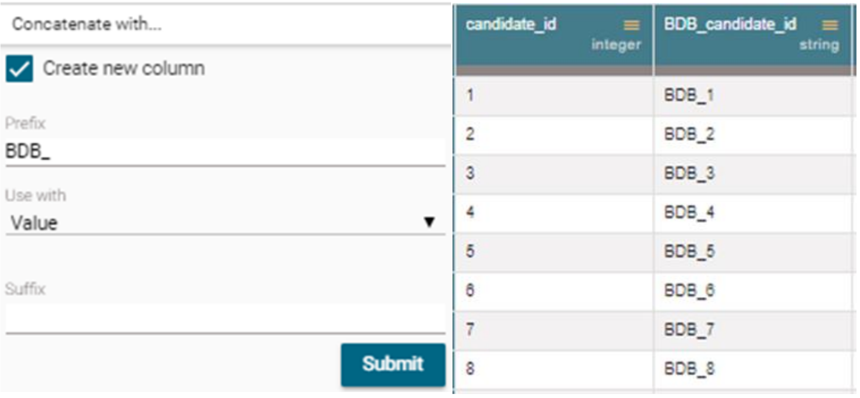
6.4.2.3. Concatenate with

The users can concatenate a column value with some other column or with some prefix/suffix.
To perform the transform, select the column to which data must be concatenated and select the 'concatenate with' transform. The available options are:

a. **Prefix:** Specify the value to be prefixed to the selected column value

b. **Use with:**
   i. Select the '**Value**' to add a Prefix/Suffix
   ii. Select '**Other column**' to concatenate two columns

c. **Suffix:** Specify the value to be suffixed to the selected column value returns when performed on the 'candidate_id' column.



Concatenate with...

☑ Create new column

Prefix
BDB_

Use with
Value

Suffix

candidate_id integer	BDB_candidate_id string
1	BDB_1
2	BDB_2
3	BDB_3
4	BDB_4
5	BDB_5
6	BDB_6
7	BDB_7
8	BDB_8

The users must select '**Use with Other column**' option to concatenate a value with another column and select the '**Use with Value**' option to add prefix/suffix.

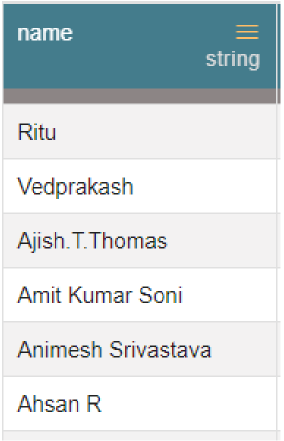
6.4.2.4. Delete Column

It deletes any selected column.

To perform the transform, select the column and click on the ‘**Delete Column**’ transform.

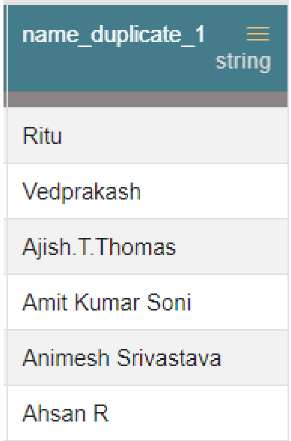
6.4.2.5. Duplicate Columns

It will create a duplicate of the selected column.



name string
Ritu
Vedprakash
Ajish.T.Thomas
Amit Kumar Soni
Animesh Srivastava
Ahsan R

gives



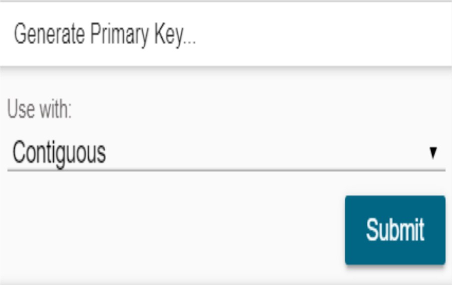
name_duplicate_1 string
Ritu
Vedprakash
Ajish.T.Thomas
Amit Kumar Soni
Animesh Srivastava
Ahsan R

6.4.2.6. Generate Primary Key

It generates the primary key for the table. It is a table-based operation.

Use with: The user gets two options to generate the primary key. Contiguous will generate the auto incremented value starting from 1.

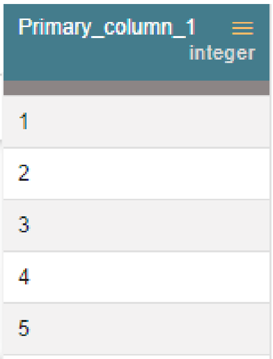
The ‘Non_contiguous’ option gets generated the unique and random integer value.



Generate Primary Key...

Use with:

Contiguous



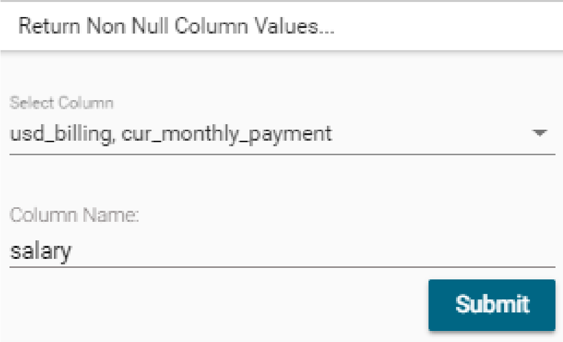
Primary_column_1 integer
1
2
3
4
5

6.4.2.7. Return Non-Null Column Values

The transform returns the first non-null value from the list of columns specified to a new column. To perform the transform, select the columns which must be checked for null and specify a column name for the result.

a. **Select Column:** Select the columns to be checked for null

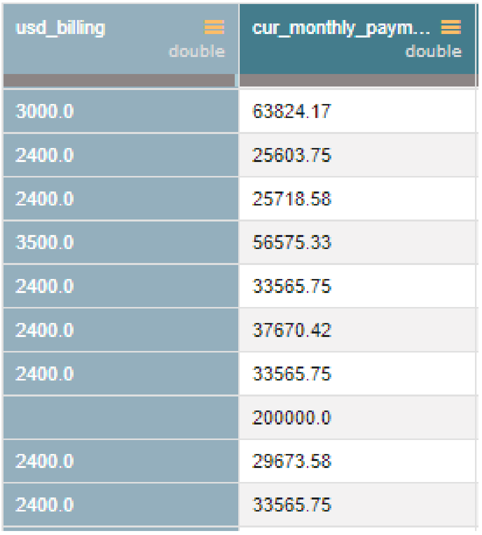
b. **Column name:** The name for the new result column returns



Return Non Null Column Values...

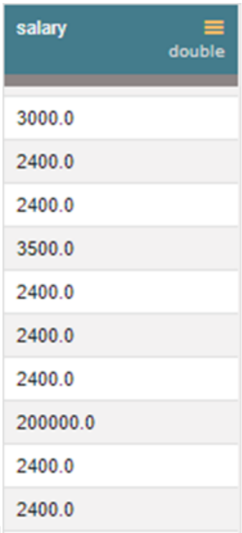
Select Column
usd_billing, cur_monthly_payment

Column Name:
salary



usd_billing double	cur_monthly_paym... double
3000.0	63824.17
2400.0	25603.75
2400.0	25718.58
3500.0	56575.33
2400.0	33565.75
2400.0	37670.42
2400.0	33565.75
	200000.0
2400.0	29673.58
2400.0	33565.75

returns the new result column



salary double
3000.0
2400.0
2400.0
3500.0
2400.0
2400.0
2400.0
200000.0
2400.0
2400.0

6.4.3. Conversions

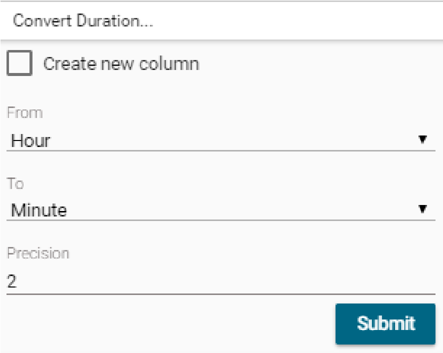
6.4.3.1. Convert Duration

The transform converts any duration (day, hour, minute, seconds, milliseconds) to any specified duration.

To perform the transform, select the column which has the duration to be converted and specify the duration type.

a. **From:** The type of source interval
b. **To:** The type of destination interval
c. **Precision:** The decimal points to be retained

Below is the snapshot of how the transform converts data:

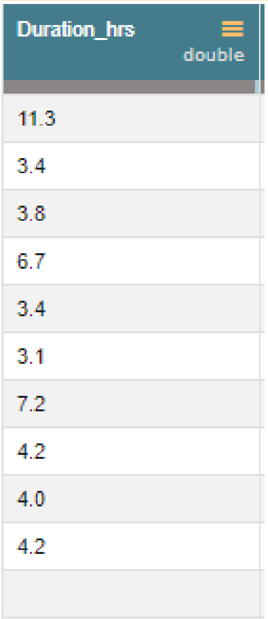

Convert Duration...

☐ Create new column

From: Hour ▾

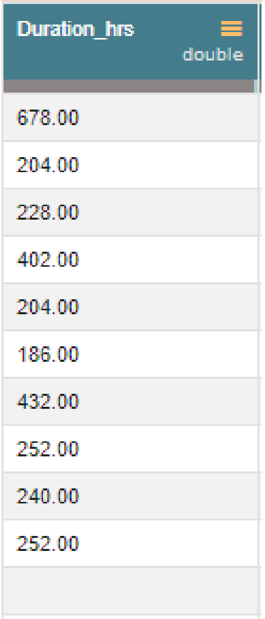
To: Minute ▾

Precision: 2



Duration_hrs double
11.3
3.4
3.8
6.7
3.4
3.1
7.2
4.2
4.0
4.2

converts to



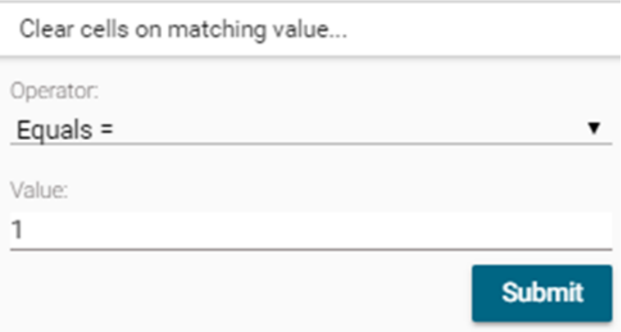
Duration_hrs double
678.00
204.00
228.00
402.00
204.00
186.00
432.00
252.00
240.00
252.00

6.4.4. Data Cleansing

6.4.4.1. Clear Cells on Matching Value

Clear the cell value on matching the condition specified. Operators include contains, equals, starts with, end with and regex match. Transform applies on the same column.

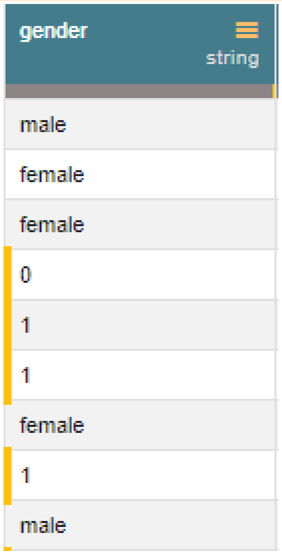
- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column


Clear cells on matching value...

Operator: Equals = ▾

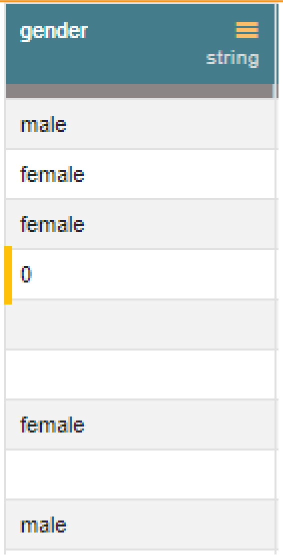
Value: 1

The value selected in the form clears the cell with 1 in the selected column.



gender ≡ string
male
female
female
0
1
1
female
1
male

turns



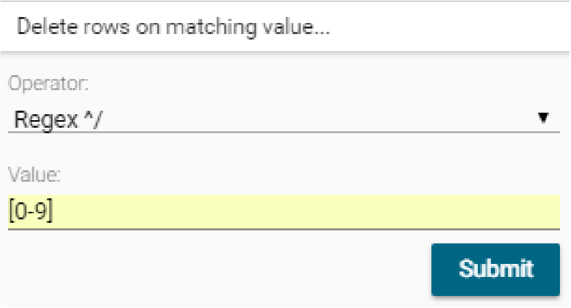
gender ≡ string
male
female
female
0
female
male

when above transformation is applied

6.4.4.2. Delete Rows on Matching Value

Delete the rows on matching the condition specified for that column. Operators include contains, equals, starts with, ends with and regex match.

- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column

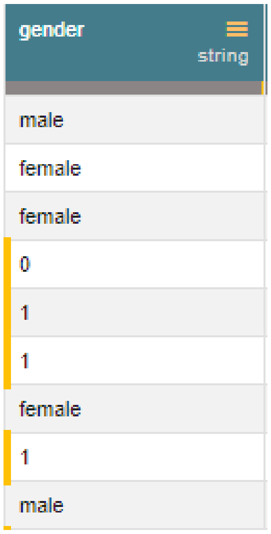


Delete rows on matching value...

Operator: Regex ^/

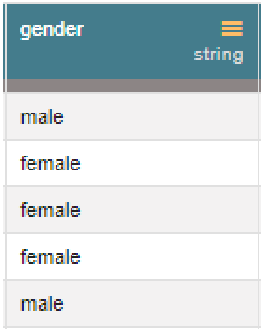
Value: [0-9]

The value selected in the form deletes the row with any numbers from 0-9 in the selected column.



gender ≡ string
male
female
female
0
1
1
female
1
male

turns to

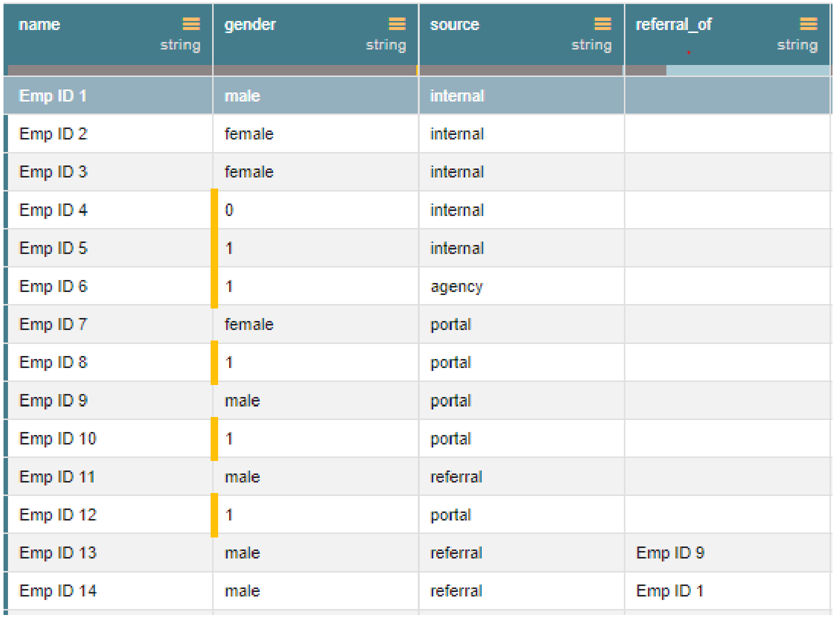


gender ≡ string
male
female
female
female
male

when the above transform is applied.

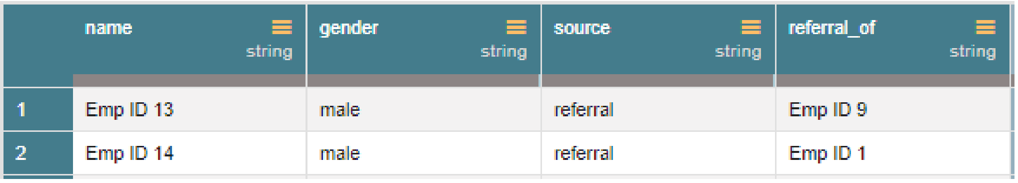
6.4.4.3. Delete Rows with Empty Cell

a. The transform deletes any row which has a blank value in the selected column. The transform does not have a form.



name	gender	source	referral_of
string	string	string	string
Emp ID 1	male	internal	
Emp ID 2	female	internal	
Emp ID 3	female	internal	
Emp ID 4	0	internal	
Emp ID 5	1	internal	
Emp ID 6	1	agency	
Emp ID 7	female	portal	
Emp ID 8	1	portal	
Emp ID 9	male	portal	
Emp ID 10	1	portal	
Emp ID 11	male	referral	
Emp ID 12	1	portal	
Emp ID 13	male	referral	Emp ID 9
Emp ID 14	male	referral	Emp ID 1

b. When we perform the transform on column “referral_of” it deletes all the rows which have an empty value in that column returning the data as below:

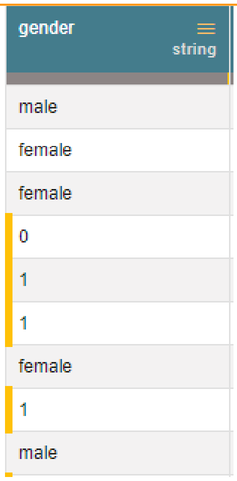


	name	gender	source	referral_of
	string	string	string	string
1	Emp ID 13	male	referral	Emp ID 9
2	Emp ID 14	male	referral	Emp ID 1

6.4.4.4. Delete Rows with Invalid Cell

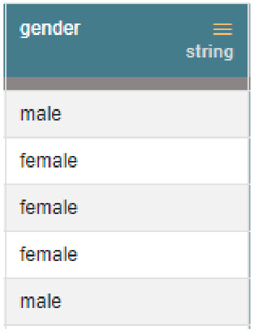
a. The transform deletes any row which has an invalid value in the selected column. The transform does not have form.

b. When we do the transform on the ‘gender’ column, it deletes all rows marked invalid as displayed below:



gender string
male
female
female
0
1
1
female
1
male

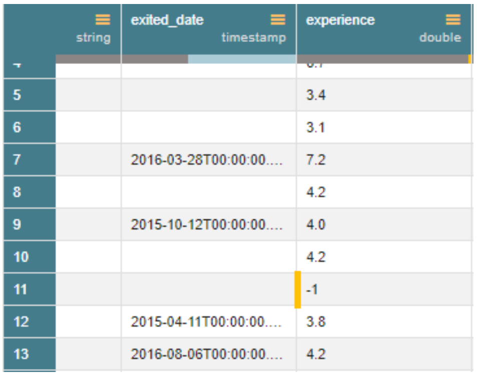
returns



gender string
male
female
female
female
male

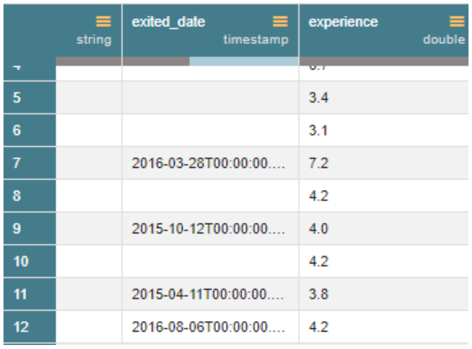
6.4.4.5. Delete Rows with Negative Values

1. It deletes the rows which have a negative value in the selected column. This transform does not have a form.
2. When this transform is applied to experience column, it deletes all rows with negative, as displayed below:



	string	exited_date timestamp	experience double
5			3.4
6			3.1
7		2016-03-28T00:00:00....	7.2
8			4.2
9		2015-10-12T00:00:00....	4.0
10			4.2
11			-1
12		2015-04-11T00:00:00....	3.8
13		2016-08-06T00:00:00....	4.2

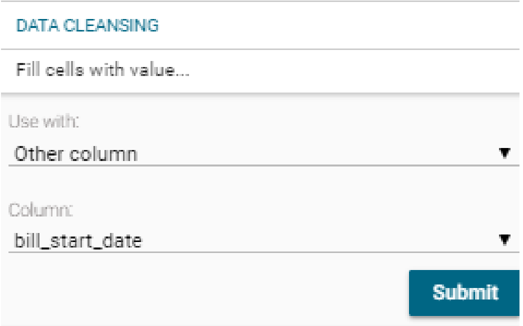
3. It returns the transformed column as displayed below:



	string	exited_date timestamp	experience double
5			3.4
6			3.1
7		2016-03-28T00:00:00....	7.2
8			4.2
9		2015-10-12T00:00:00....	4.0
10			4.2
11		2015-04-11T00:00:00....	3.8
12		2016-08-06T00:00:00....	4.2

6.4.4.6. Fill Cells with Value

It fills the selected column with a value or a value from another column.



DATA CLEANSING

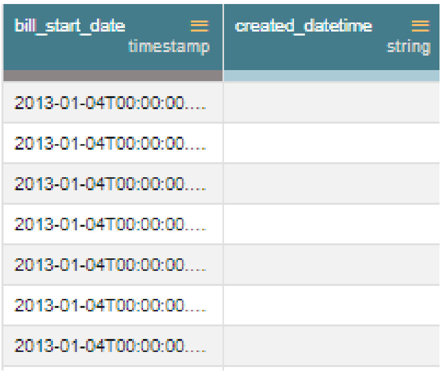
Fill cells with value...

Use with: Other column

Column: bill_start_date

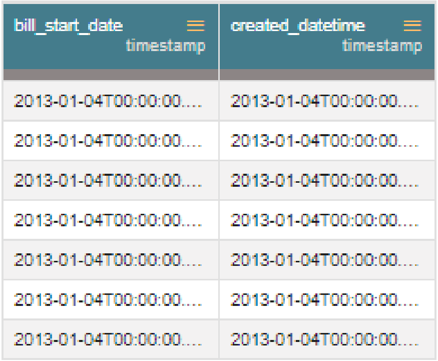
- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be filled, or the column with which the value must be replaced

When the above transform is applied to the below data on the column 'created_datetime,' it copies the value from the 'bill_start_date' column to the 'created_datetime' column.



bill_start_date timestamp	created_datetime string
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	
2013-01-04T00:00:00....	

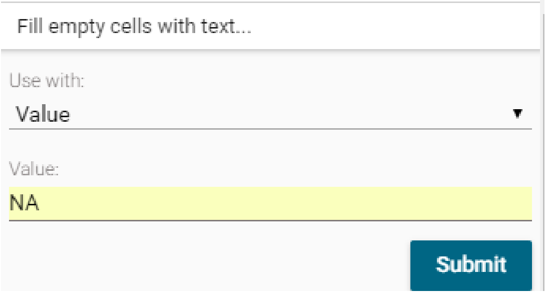
converts into



bill_start_date timestamp	created_datetime timestamp
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....
2013-01-04T00:00:00....	2013-01-04T00:00:00....

6.4.4.7. Fill Empty Cells with Text

It helps to fill the empty cells of a selected column with a value or a value from another column if the destination column is empty.



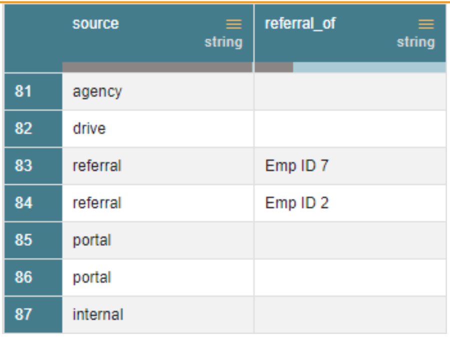
Fill empty cells with text...

Use with: Value

Value: NA

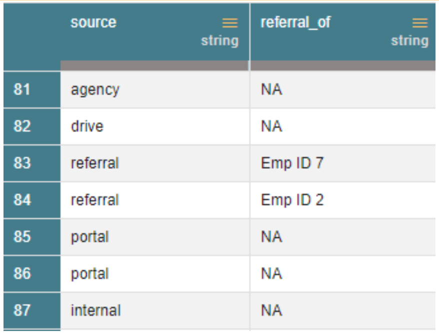
- **Use with:** Specify whether to fill with a value or another column value.
- **Column/ Value:** The value with which the column must be filled, or the column with which the value must be replaced.

When the transform is applied to the below data on column 'referral_of,' it fills the value 'NA' for all the empty cells of that column.



	source string	referral_of string
81	agency	
82	drive	
83	referral	Emp ID 7
84	referral	Emp ID 2
85	portal	
86	portal	
87	internal	

converts to



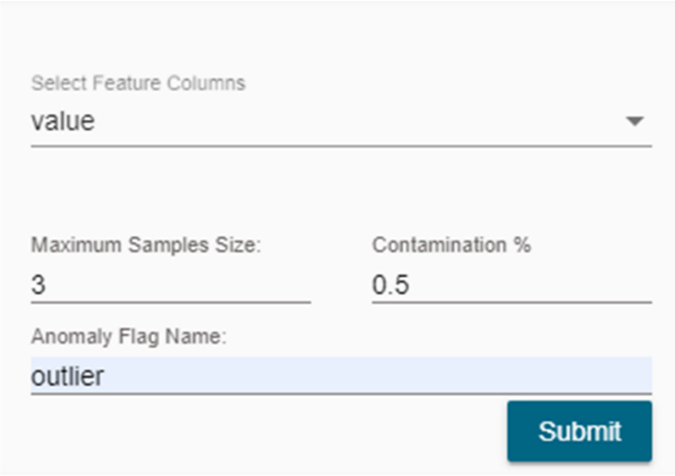
	source string	referral_of string
81	agency	NA
82	drive	NA
83	referral	Emp ID 7
84	referral	Emp ID 2
85	portal	NA
86	portal	NA
87	internal	NA

6.4.4.8. Find Anomaly

Anomaly detection is used to identify any anomaly present in the data. i.e., Outlier. Instead of looking for usual points in the data, it looks for any anomaly. It uses the **Isolation Forest** algorithm.

The '**Find Anomaly**' transform takes four parameters:

1. **Feature column:** We can select one or more column where we want to find the anomaly.
2. **Max Sample size:** Isolation forest takes the training data of a given sample size to find out the normal value in the dataset. The sample size can vary from 1 to 250 (both inclusive).
3. **Contamination (%):** It is the percentage of observations we believe to be outliers. It varies from 0 to 1 (both inclusive).
4. **Anomaly Flag Name:** The result is either 0 or 1. 0 means the data is standard data, and 1 means data is an outlier. This information gets stored in the new column given in the anomaly flag name.
5. Click the '**Submit**' option to detect anomaly from the selected data.



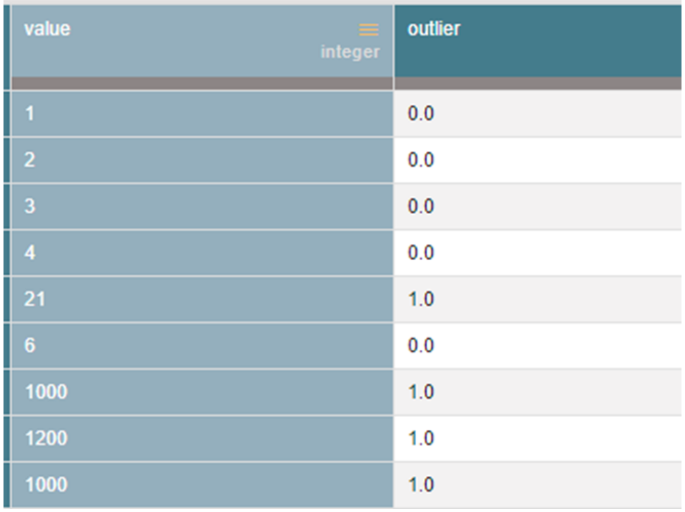
Select Feature Columns: value

Maximum Samples Size: 3

Contamination %: 0.5

Anomaly Flag Name: outlier

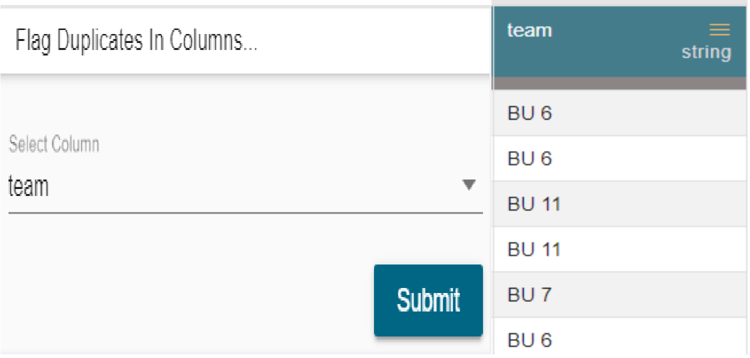
The anomaly gets store in the new column under the anomaly flag name (In this case, it is displayed under the '**outlier**' column).



value integer	outlier
1	0.0
2	0.0
3	0.0
4	0.0
21	1.0
6	0.0
1000	1.0
1200	1.0
1000	1.0

6.4.4.9. Flag Duplicates in Columns

This transform adds a new Boolean column based on duplicate values in the column. For original value it gives false, and for the duplicate value, it provides true value.

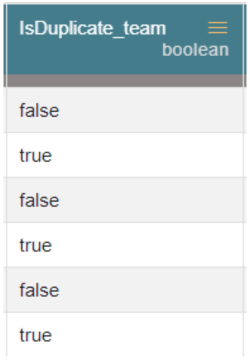


Flag Duplicates In Columns...

Select Column: team

team string
BU 6
BU 6
BU 11
BU 11
BU 7
BU 6

returns



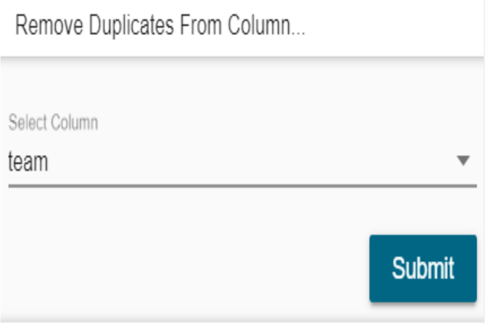
IsDuplicate_team boolean
false
true
false
true
false
true

6.4.4.10. Flag Duplicates in Tables

This transform adds a new Boolean column based on duplicate rows in the table. For original value it gives false, and for the duplicate value, it provides true value.

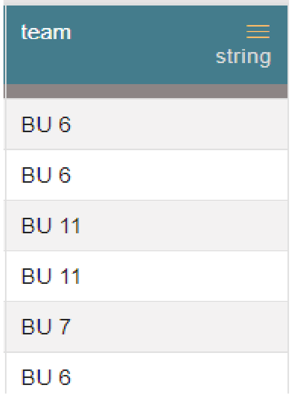
6.4.4.11. Remove Duplicates from Column

It removes duplicate values from the selected columns. This transform can be performed on a single as well as on multiple columns.



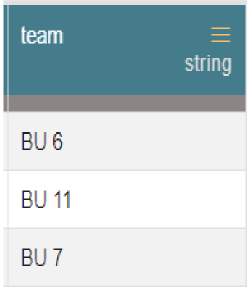
Remove Duplicates From Column...

Select Column
team



team string
BU 6
BU 6
BU 11
BU 11
BU 7
BU 6

converts to



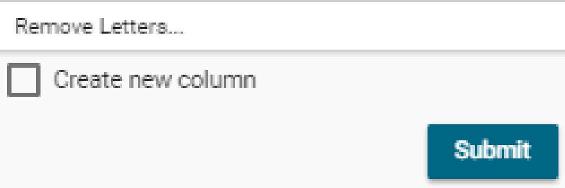
team string
BU 6
BU 11
BU 7

6.4.4.12. Remove Duplicates from Table

It Removes all duplicate rows from the table.

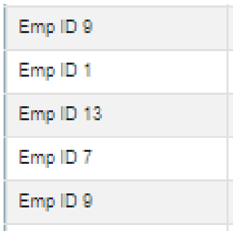
6.4.4.13. Remove Letters

It removes any letter present in the selected column. The users can either add a new column with the transformed value or overwrite the same column.

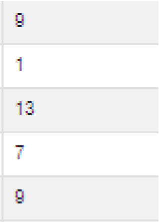


Remove Letters...

□ Create new column



Emp ID 9
Emp ID 1
Emp ID 13
Emp ID 7
Emp ID 9

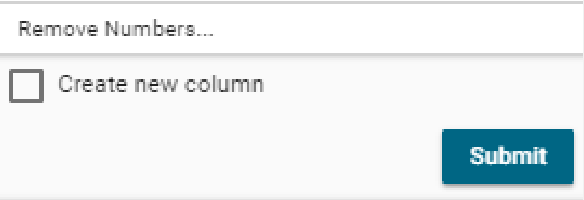


9
1
13
7
9

The selected column converts into after transformation.

6.4.4.14. Remove Numbers

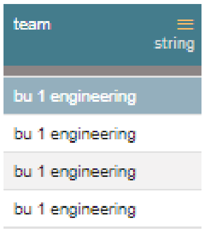
It removes any number present in the selected column. We can either add a new column with the transformed value or overwrite the same column.



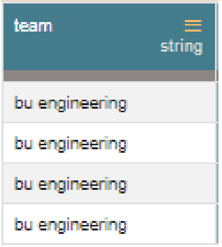
Remove Numbers...

□ Create new column

When the ‘Remove Numbers’ transform gets performed on a selected column,

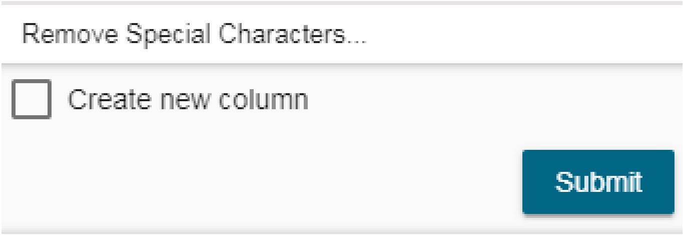

team
string
bu 1 engineering
bu 1 engineering
bu 1 engineering
bu 1 engineering

it removes numbers from the selected column

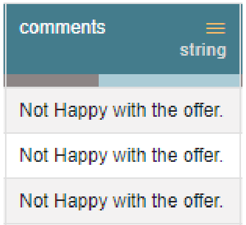

team
string
bu engineering
bu engineering
bu engineering
bu engineering

6.4.4.15. Remove Special Characters

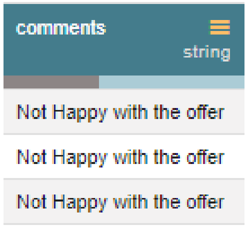
It removes any special character present in the selected column. Only letters, numbers, and spaces are retained. We can either add a new column with the transformed value or overwrite the same column.



When the transform '**Remove Special Characters**' gets performed on the selected column,


comments
string
Not Happy with the offer.
Not Happy with the offer.
Not Happy with the offer.

the punctuations get removed from the column

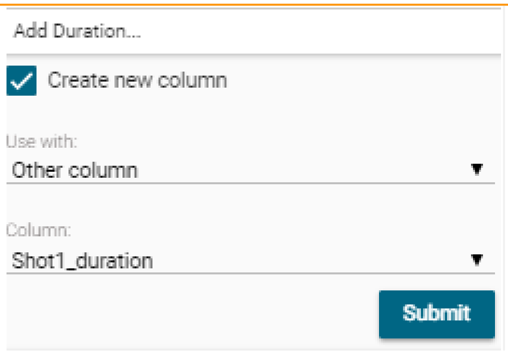

comments
string
Not Happy with the offer
Not Happy with the offer
Not Happy with the offer

6.4.5. Dates

6.4.5.1. Add Duration

The transform adds two-time values. It can either add the selected column with a time value or time from another column. The transform supports adding time into '**hh:mm:ss.mmm**' and '**hh:mm:ss**' formats.

- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be added, or the column with which the selected column value must be added.



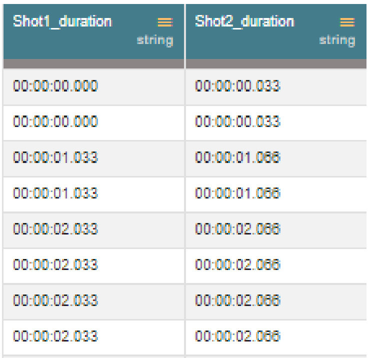
Add Duration...

☑ Create new column

Use with:
Other column ▼

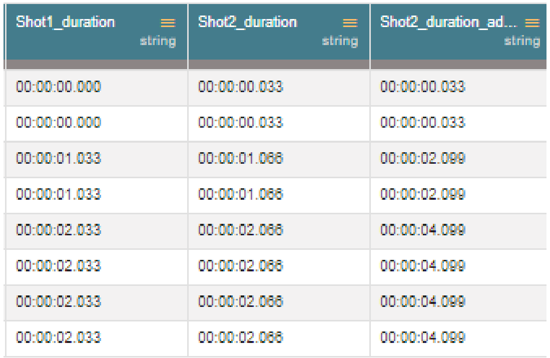
Column:
Shot1_duration ▼

The transform when performed on the data selecting ‘Shot1_duration’, it adds Shot1_duration and Shot2_duration and gives a new column with the result.



Shot1_duration string	Shot2_duration string
00:00:00.000	00:00:00.033
00:00:00.000	00:00:00.033
00:00:01.033	00:00:01.066
00:00:01.033	00:00:01.066
00:00:02.033	00:00:02.066
00:00:02.033	00:00:02.066
00:00:02.033	00:00:02.066
00:00:02.033	00:00:02.066

converts to



Shot1_duration string	Shot2_duration string	Shot2_duration_ad... string
00:00:00.000	00:00:00.033	00:00:00.033
00:00:00.000	00:00:00.033	00:00:00.033
00:00:01.033	00:00:01.066	00:00:02.099
00:00:01.033	00:00:01.066	00:00:02.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099
00:00:02.033	00:00:02.066	00:00:04.099

6.4.5.2. Add Interval to Date

It adds the time duration specified to the selected datetime column.

- **Input Format:** It is used to specify the format of the selected date column format. It can have values ‘Year first’, ‘Month first’, and ‘Day first.’
- **Value Type:** It specifies the type of duration which acts as the operand for the addition. The value type can be years, months, days, weeks, hours, minutes or milliseconds
- **Value:** The value or the operand that must be added with the selected column

Note: The transform supports datetime column of ‘**yyyy-mm-dd’** into the ‘**hh:mm:ss’** format.

6.4.5.3. Extract Time

Extract the time units from a selected column with a time value. The time units that get extracted include hours, minutes, seconds, milliseconds, and time to milliseconds.

- **Hours:** Extracts hours from a time
- **Minutes:** Extracts minutes from a time
- **Seconds**: Extracts seconds from a time
- **MilliSeconds:** Extracts milliseconds from a time
- **Time to MilliSeconds:** Converts the time given to milliseconds

Note : The transform supports time format like- hh:mm:ss:mmm, hh:mm:ss, hh:mm

6.4.5.4. Extract Date

It extracts the date part from a selected column with a date value.

The date parts that can be extracted include day, month, year, the day of the week, the day of the year and a week of the year.

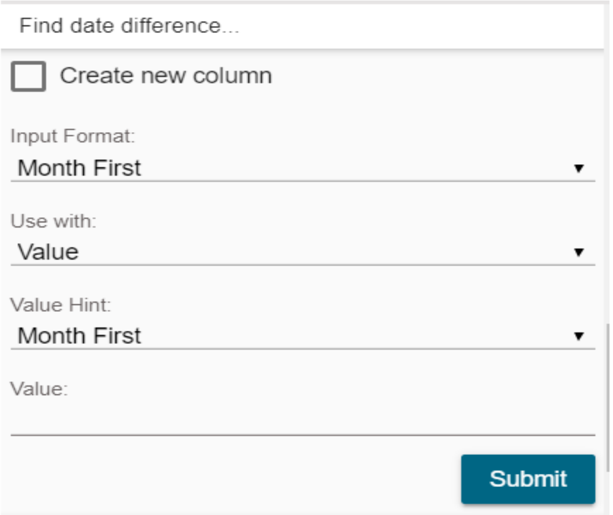
- **Day:** It extracts day from a date
- **Month:** It extracts the month from a date/datetime. We can specify the pattern in which the month value has to be returned. Month pattern can be 0-12, Jan - Dec or January - December
- **Year**: It extracts the year from a date. We can specify the pattern in which the year has to be returned. Year pattern can be in the 'yy' or 'yyyy' format.
- **Day of Week:** It returns the day of the week for the selected date. Day of week pattern can also be specified. The pattern can be 1-7, Sun-Sat or Sunday-Saturday
- **Day of Year:** It returns a number between 1 and 365, which indicates the sequential day number starting with day one on January 1st.
- **Week of Year:** It replaces a number between 1 and 53, which indicates the sequential week number beginning with 1 for the week January 1st falls.

Note: The transform supports Date and DateTime format (date hh:mm:ss)

6.4.5.5. Find Date Difference

The transform finds the difference between two date values. It can either subtract the selected column with a date value or date from another column. The transformed value can replace the existing column value or can be added as a new column.

- **Input Format**: Specifies the format of the given date column
- **Use with**: Specify whether to fill with a value or another column value
- **Value Hint**: Specifies format of value from which we want to find the difference
- **Value**: Pass the date value from where you want to find the date difference



Find date difference...

□ Create new column

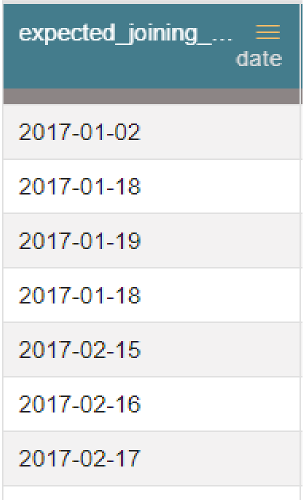
Input Format: Month First

Use with: Value

Value Hint: Month First

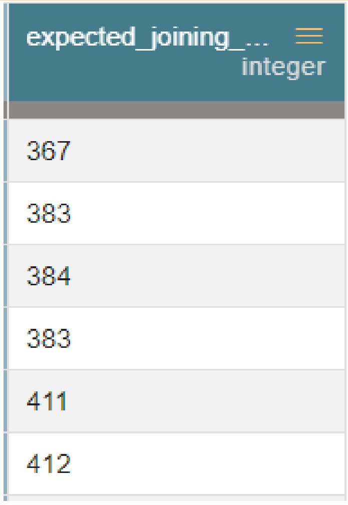
Value:

This transform gives the number of days by finding out the difference between the given date and value/date column which we have used.
Here value used is: 2016-01-01



expected_joining_... date
2017-01-02
2017-01-18
2017-01-19
2017-01-18
2017-02-15
2017-02-16
2017-02-17

converts to

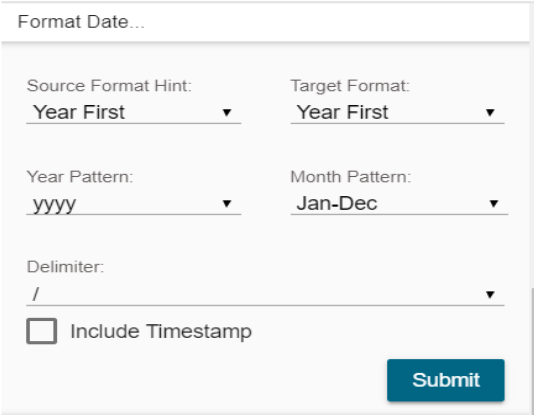


expected_joining_... integer
367
383
384
383
411
412

6.4.5.6. Format Date

The users can change the format of a date column by using this transform.

- **Source Format Hint:** Specifies the current format of the date column.
- **Target Format:** Specifies what we want first(Year, Month, Day) in our output format of the date column
- **Year Pattern:** Specifies format of the year (yyyy or yy) in the output date column.
- **Month Pattern:** It specifies the format of the month (number, Jan-Dec, January-December) in the output date column.
- **Delimiter:** Specifies Delimiter(like- slash, a hyphen, comma, full stop, space) for the output date column.
- **Include Timestamp:** It adds a timestamp to the current date format if enabled with a tick mark.



Format Date...

Source Format Hint: Year First

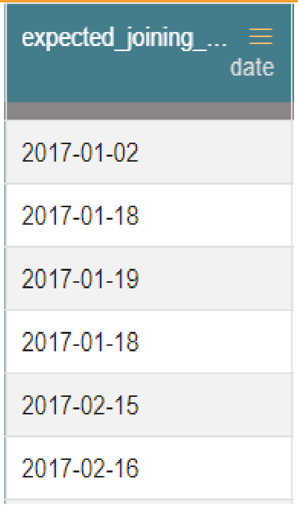
Target Format: Year First

Year Pattern: yyyy

Month Pattern: Jan-Dec

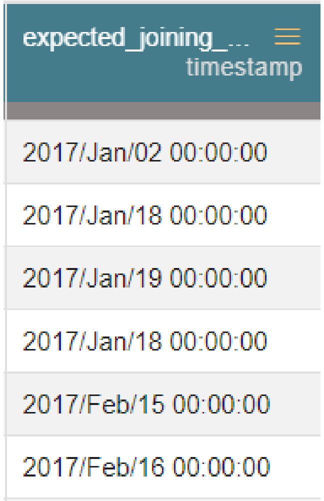
Delimiter: /

☐ Include Timestamp



expected_joining_... date
2017-01-02
2017-01-18
2017-01-19
2017-01-18
2017-02-15
2017-02-16

converts to

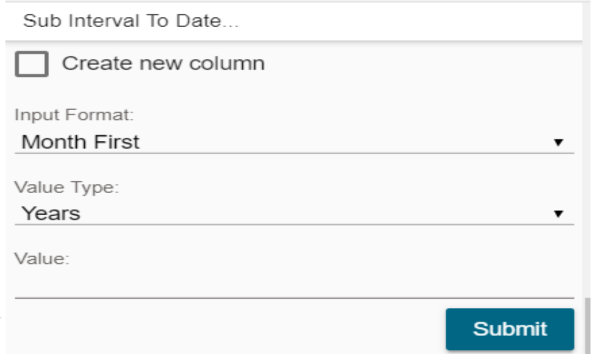


expected_joining_... timestamp
2017/Jan/02 00:00:00
2017/Jan/18 00:00:00
2017/Jan/19 00:00:00
2017/Jan/18 00:00:00
2017/Feb/15 00:00:00
2017/Feb/16 00:00:00

6.4.5.7. Sub Interval to Date

The 'Sub Interval to Date' transform subtracts specified value(interval) from the given date column. The transformed value can replace the existing column value or can be added as a new column.

- **Input Format**- Format of date column(given) should be specified here.
- **Value Type**-specifies what we want to subtract like years, months, days, weeks, etc.
- **Value**- specifies how many years(value type) we want to subtract.



Sub Interval To Date...

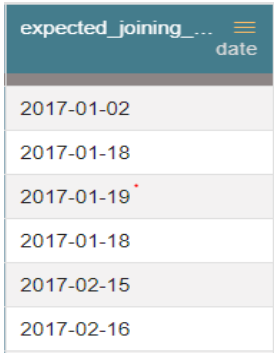
□ Create new column

Input Format: Month First

Value Type: Years

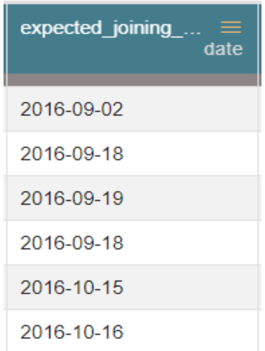
Value:

This transform when performed subtracts four months from the date column and gives this new column having the date which is four months back from the given date.



expected_joining_... date
2017-01-02
2017-01-18
2017-01-19
2017-01-18
2017-02-15
2017-02-16

converts to

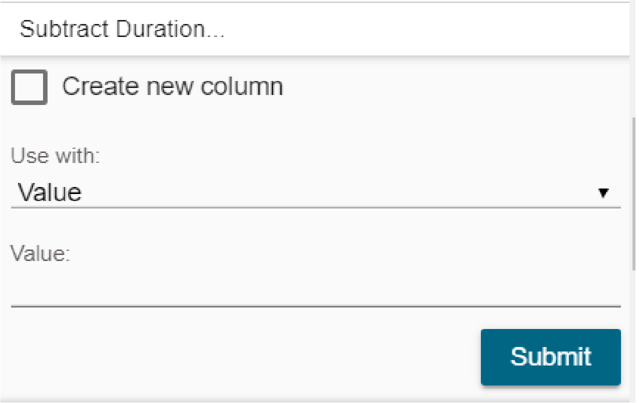


expected_joining_... date
2016-09-02
2016-09-18
2016-09-19
2016-09-18
2016-10-15
2016-10-16

6.4.5.8. Subtract Duration

The ‘Subtract Duration’ transform deducts the time values in two ways. It can either subtract the selected column with a time value or time from another column. The transform supports subtracting time into ‘**hh:mm:ss.mmm’ ,‘hh:mm:ss’** and **’hh:mm’** formats. The transformed value can replace the existing column value or can be added as a new column.

- **Use with:** Specify whether to fill with a value or another column value
- **Column/ Value:** The value with which the column must be subtracted, or the column with which the selected column value must be subtracted.



Subtract Duration...

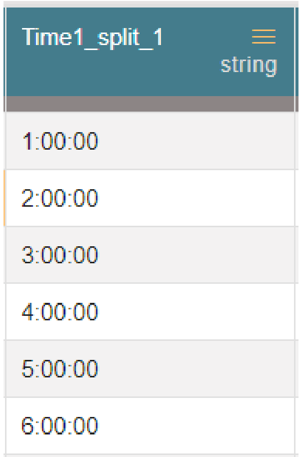
□ Create new column

Use with:

Value ▾

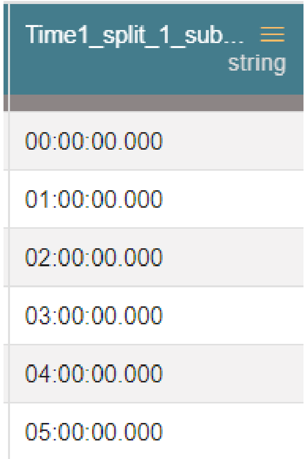
Value:

This transform when performed on Time1_split1 for subtracting 01:00:00 from this column provides a new column having values after deducting 01:00:00.



Time1_split_1 string
1:00:00
2:00:00
3:00:00
4:00:00
5:00:00
6:00:00

converts to



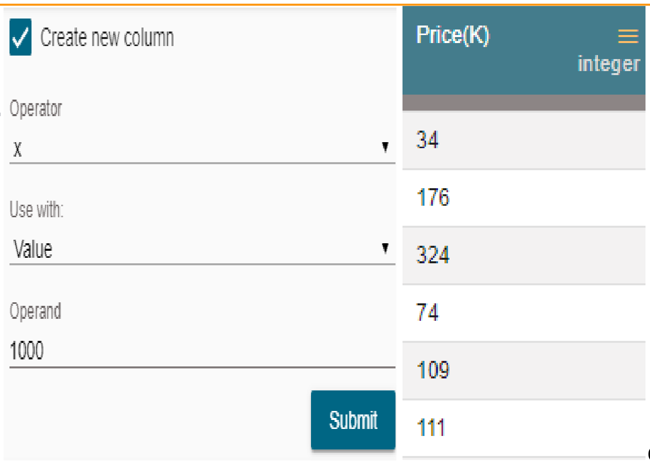
Time1_split_1_sub... string
00:00:00.000
01:00:00.000
02:00:00.000
03:00:00.000
04:00:00.000
05:00:00.000

6.4.6. Integer

6.4.6.1. Add, Multiply, Subtract or Divide

It performs the arithmetic operation on the selected numerical column.

- **Operator:** There is four arithmetic operation to choose from +, -, / and *.
- **Use with:** The operation can be performed between column-column and column-value.
- **Operand/Column:** The arithmetic operation needs two operands. The first operand is one on which the operation is being performed. The second operation can be either be a value or other numerical column based on the choice of use with an option.



☑ Create new column

Operator
x

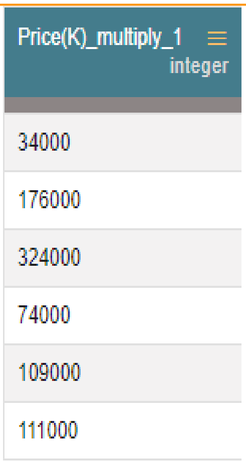
Use with:
Value

Operand
1000

Submit

Price(K) integer
34
176
324
74
109
111

converts to

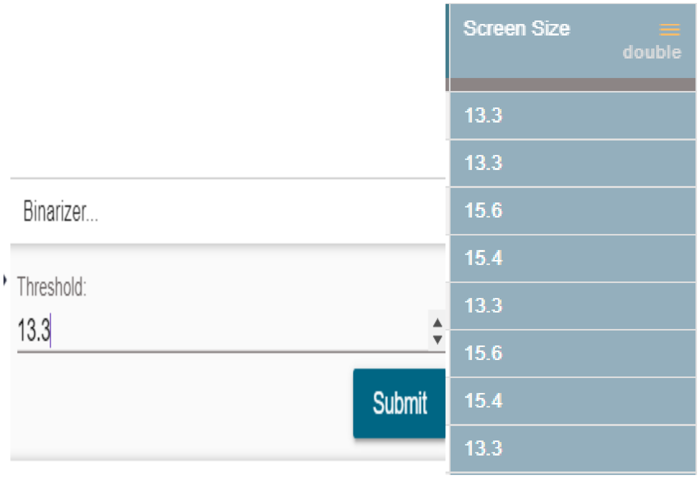


Price(K)_multiply_1 integer
34000
176000
324000
74000
109000
111000

6.4.7. ML

6.4.7.1. Binarizer

It converts the value of a numerical column to zero when the value in the column is less than or equals to the threshold value and one if the value in the column is greater than threshold value.



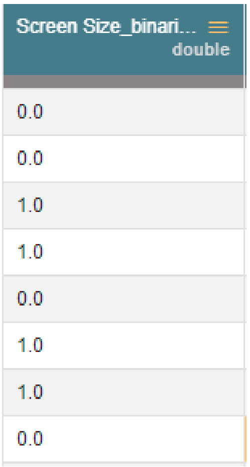
Binarizer...

Threshold:
13.3

Submit

Screen Size double
13.3
13.3
15.6
15.4
13.3
15.6
15.4
13.3

converts to



Screen Size_binari... double
0.0
0.0
1.0
1.0
0.0
1.0
1.0
0.0

6.4.8. Numbers

6.4.8.1. Max

It gives the maximum value from the selected columns row-wise. The selected column should be numerical and more than one.

6.4.8.2. Mean

It gives the average value of the selected columns row-wise. The selected column should be numerical and more than one.

6.4.8.3. Min

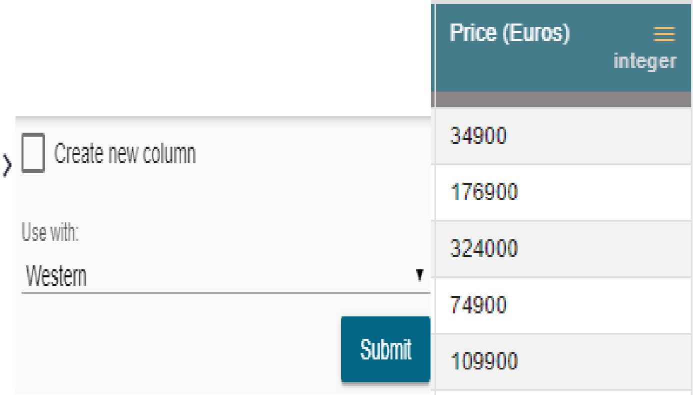
It gives the minimum value from the selected columns row-wise. The selected column should be numerical and more than one.

6.4.8.4. Negate

It complements the sing of a numeric value. If the value is positive, then a negative value comes and vice-versa.

6.4.8.5. Number Name

It converts the value of the selected column into words. The column must be of integer type.
Use with: It gives the users an option to convert word into either western format or Indian format.

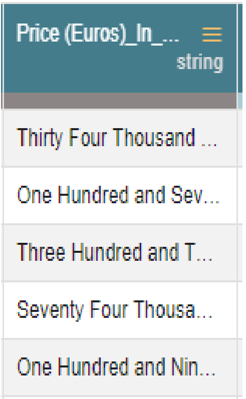


☐ Create new column

Use with: Western

Price (Euros) integer
34900
176900
324000
74900
109900

converts to



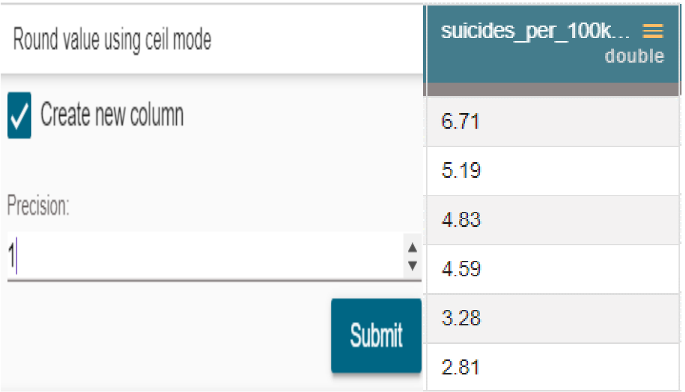
Price (Euros)_In_... string
Thirty Four Thousand ...
One Hundred and Sev...
Three Hundred and T...
Seventy Four Thousa...
One Hundred and Nin...

6.4.8.6. Remove Fractional Part

It removes the fractional part from the numerical column. The float column is converted into the integer data type.

6.4.8.7. Round Value using Ceil Mode

It replaces the number with a greater integer value if the number is between two integer value. The transformed value can replace the existing column value or can be added as a new column.



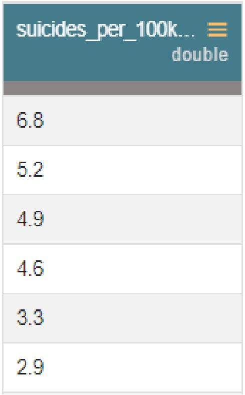
Round value using ceil mode

☑ Create new column

Precision: 1

suicides_per_100k... double
6.71
5.19
4.83
4.59
3.28
2.81

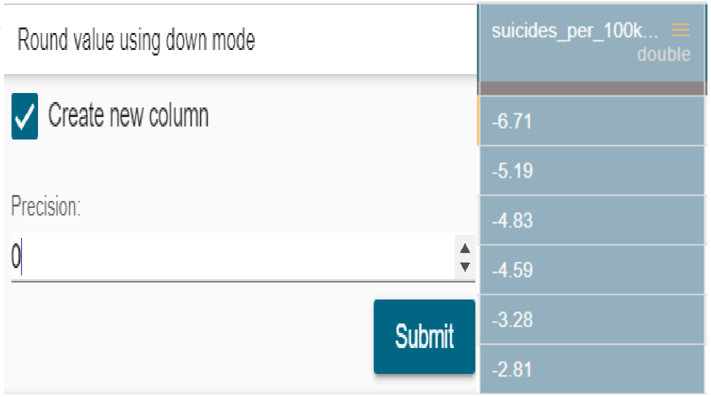
converts to



suicides_per_100k... double
6.8
5.2
4.9
4.6
3.3
2.9

6.4.8.8. Round Value using Down Mode

It rounds the number down to a specified digit or gives the specified number of decimals without any change in value. The transformed value can replace the existing column value or can be added as a new column.

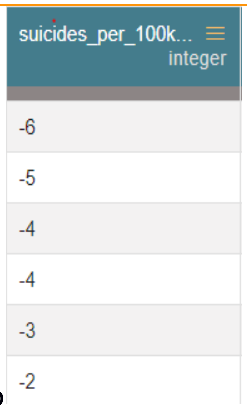

Round value using down mode

☑ Create new column

Precision:
0

suicides_per_100k... double
-6.71
-5.19
-4.83
-4.59
-3.28
-2.81

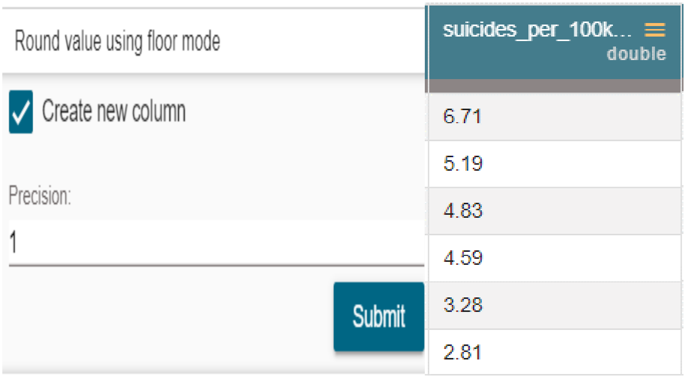
converts to



suicides_per_100k... integer
-6
-5
-4
-4
-3
-2

6.4.8.9. Round Value using Floor Mode

It replaces a number with the lesser integer value, if the number is between two integer value, or it rounds the number down to the nearest multiple of Specified significance. It does not consider weather next digit is 5 or less than or greater than 5. The transformed value can replace the existing column value or can be added as a new column.

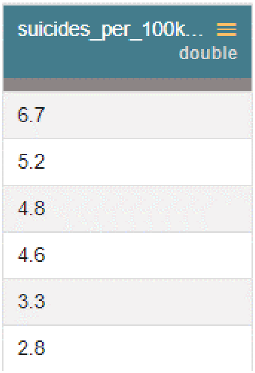

Round value using floor mode

☑ Create new column

Precision:
1

suicides_per_100k... double
6.71
5.19
4.83
4.59
3.28
2.81

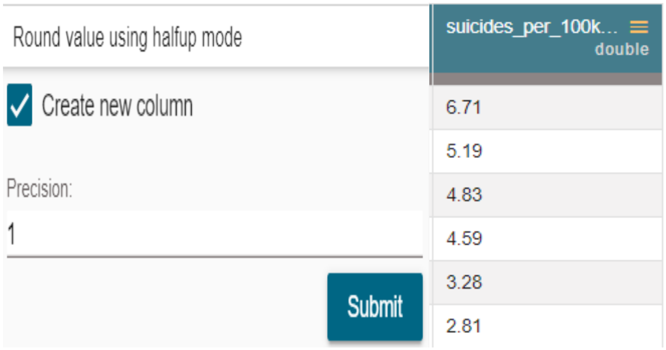
converts to



suicides_per_100k... double
6.7
5.2
4.8
4.6
3.3
2.8

6.4.8.10. Round Value using Half-up mode

It replaces a number with next integer value if its next digit is 5 or greater than 5. The transformed value can replace the existing column value or can be added as a new column.

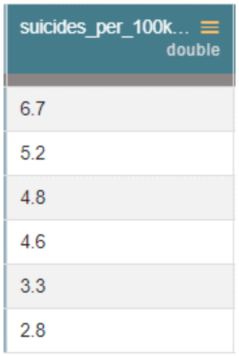

Round value using halfup mode

☑ Create new column

Precision:
1

suicides_per_100k... double
6.71
5.19
4.83
4.59
3.28
2.81

converts to



suicides_per_100k... double
6.7
5.2
4.8
4.6
3.3
2.8

6.4.9. String

6.4.9.1. Change to lower case

It converts the selected column value to the small case. The transformed value can replace the existing column value or can be added as a new column.

6.4.9.2. Change to Title Case

It converts the selected column value to title case. The transformed value can replace the existing column value or can be added as a new column.

6.4.9.3. Change to Upper Case

It converts the selected column value to capital letters. The transformed value can replace the existing column value or can be added as a new column.

6.4.9.4. Extract Substring at Position

It extracts the substring from the selected column based on the starting position and the length of the extract. The transformed value can replace the existing column value or can be added as a new column.

- **Position:** This value is required and is the start position. It can be both a positive or negative number. If it is a positive number, this function extracts from the beginning of the string. If it is a negative number, this function extracts from the end of the string.
- **Length:** This value is optional. It specifies the number of characters to extract. If omitted, the whole string is returned starting from the given position.

6.4.9.5. Extract Substring before Delimiter

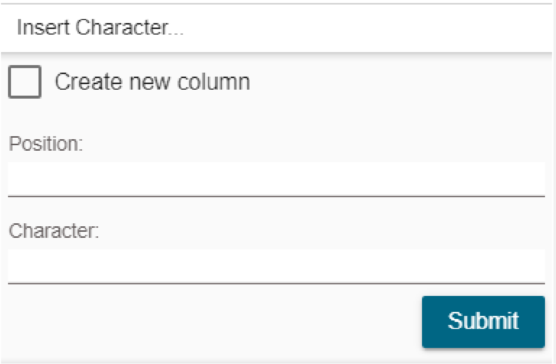
It extracts the substring from the selected column, before the 'n$^{th'}$ occurrence of the delimiter specified where 'n' is the count. The transformed value can replace the existing column value or can be added as a new column.

- **Delimiter:** The delimiter on whose occurrence the extract should happen.
- **Count:** This value is mandatory and specifies the count of occurrence of the delimiter before which the extract should happen.

6.4.9.6. Insert Character

It inserts the character entered after specified position. The transformed value can replace the existing column value or can be added as a new column.

- **Position:** The position in the cell value, after which the character must be inserted. We can even pass comma separated values. E.g., 2,4,6 insert the specified character after position 2, 4 & 6 of the cell values
- **Character**: The character that should be inserted after the specified positions



6.4.9.7. Remove Consecutive Characters

The transform removes the repeated whitespace or character and modifies the selected column /adds the result to a new column. It removes only the repetition.

- **Separator**: it has values whitespace /other. If whitespace, the transform searches for multiple white spaces and return a single-spaced value.
- **Custom repeated Character:** When a repeated character is '**Other**,' this provides an option to give the character whose consecutive occurrence must be searched.

6.4.9.8. Remove Part of Text

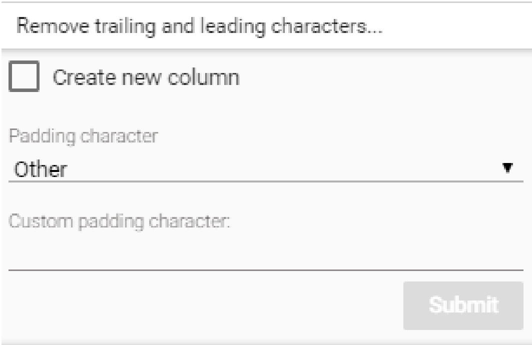
It matches and removes the matching part or entire value based on the condition. The transformed value can replace the existing column value or can be added as a new column.

- **Operator:** Select the operator required for matching from the list
- **Value:** The value or pattern to be searched for in the selected column

6.4.9.9. Remove Trailing and Leading Characters

It removes trailing and leading characters from the column. The transformed value can replace the existing column value or can be added as a new column.

- **Padding character:** Specify whether to remove whitespace or another character using the drop-down menu.
- **Custom padding character -** If '**other**' is selected as a padding character, specify which is the character to be removed.


Remove trailing and leading characters...
□ Create new column

Padding character
Other ▼

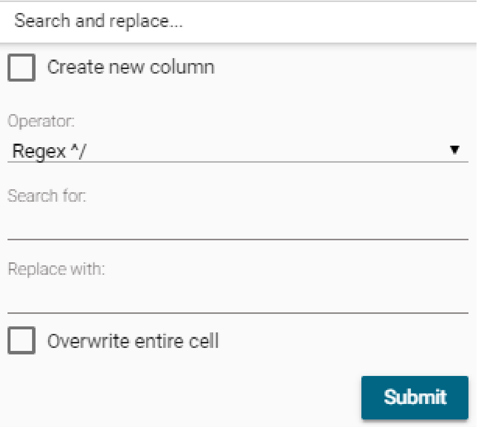
Custom padding character:
Submit

6.4.9.10. Search and Replace

It searches and replaces the matching part or entire value based on the option selected.

The transformed value can replace the existing column value or can be added as a new column.

Operator- Select the operator required for matching from the list. Operators include contains, equals, starts with, end with and regex match.
Value: It is the value or pattern to be searched for in the selected column.



Search and replace...

□ Create new column

Operator: Regex ^/

Search for:

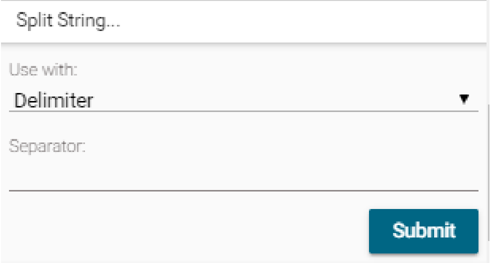
Replace with:

□ Overwrite entire cell

6.4.9.11. Split String

It splits the string based on condition. It displays new columns based on the number of delimiter and on position.

- **Use With:** Specify whether to split with a delimiter or at position
- **Delimiter:** The delimiter on whose occurrence the split should happen
- **Position:** After which position split should happen if use with is 'position.'

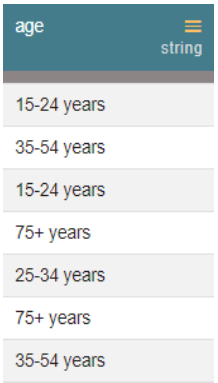


Split String...

Use with: Delimiter

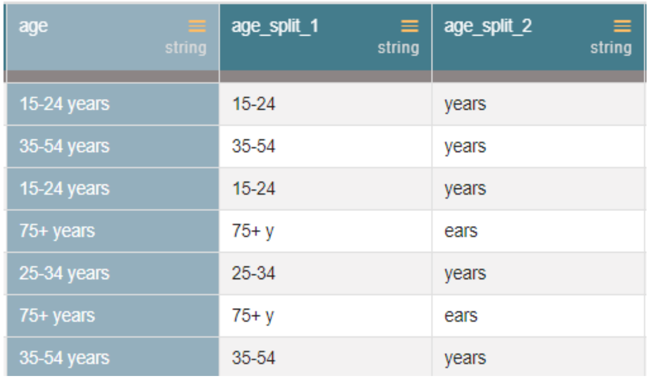
Separator:

Here splitting of the column is done based on position (after the 5th character)



age string
15-24 years
35-54 years
15-24 years
75+ years
25-34 years
75+ years
35-54 years

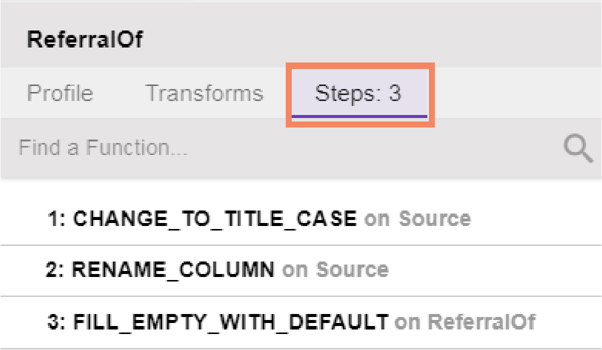
converts to



age string	age_split_1 string	age_split_2 string
15-24 years	15-24	years
35-54 years	35-54	years
15-24 years	15-24	years
75+ years	75+ y	ears
25-34 years	25-34	years
75+ years	75+ y	ears
35-54 years	35-54	years

6.5. Steps

This tab lists all the transforms that were performed on the data. It also gives a count of steps performed.

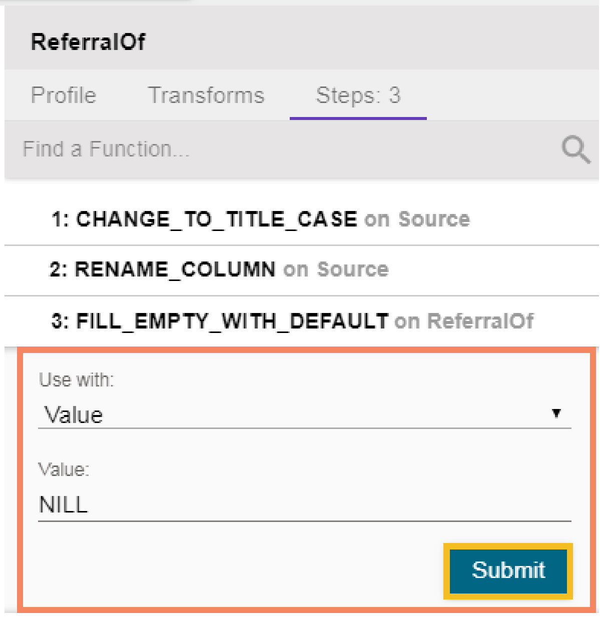


1: CHANGE_TO_TITLE_CASE on Source

2: RENAME_COLUMN on Source

3: FILL_EMPTY_WITH_DEFAULT on ReferralOf

The user can open any performed transform and edit it using the '**Steps**' tab.



1: CHANGE_TO_TITLE_CASE on Source

2: RENAME_COLUMN on Source

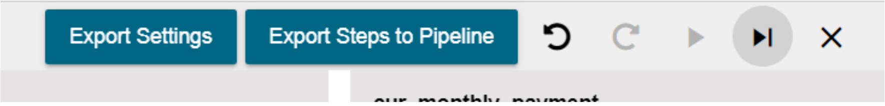
3: FILL_EMPTY_WITH_DEFAULT on ReferralOf

Use with: Value

Value: NILL

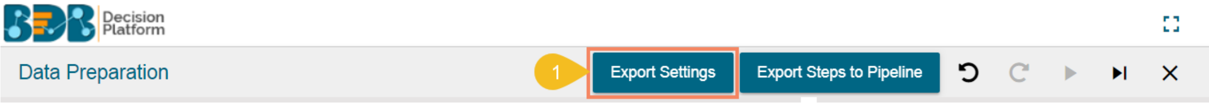
7. Navigation Pane

The navigation pane provides options to export the preparation steps in Elastic settings, move the steps out of the BDB Data Preparation. The navigation panel also has icons to perform Undo, Re-do, Replay Dirty, and Replay All options.

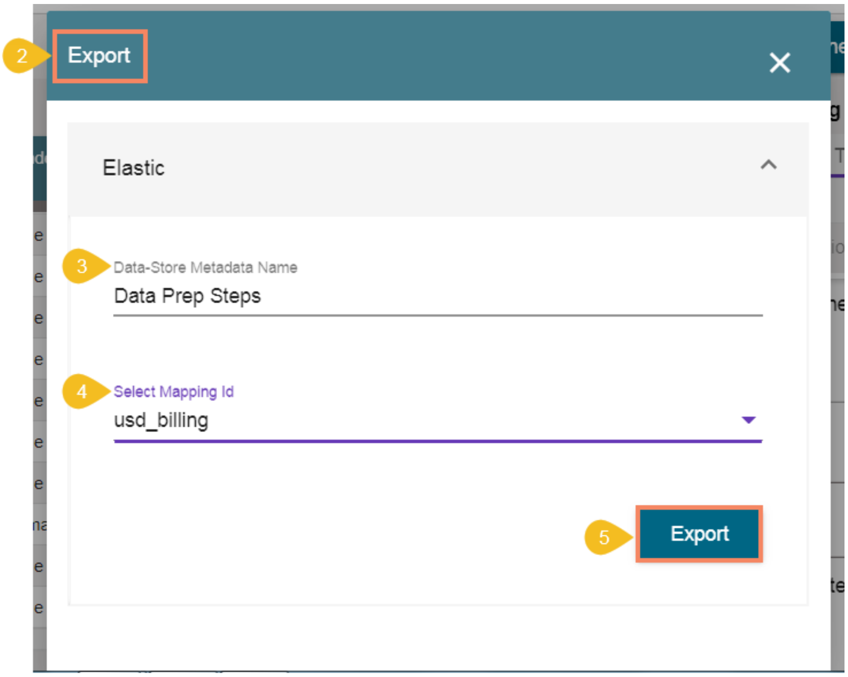


a. **Export Settings:** The '**Export Settings**' option redirects the user to specify the elastic settings into which the cleansed data must be moved.

   ○ Click the **'Export Settings'** option using the Navigation Pane**.**



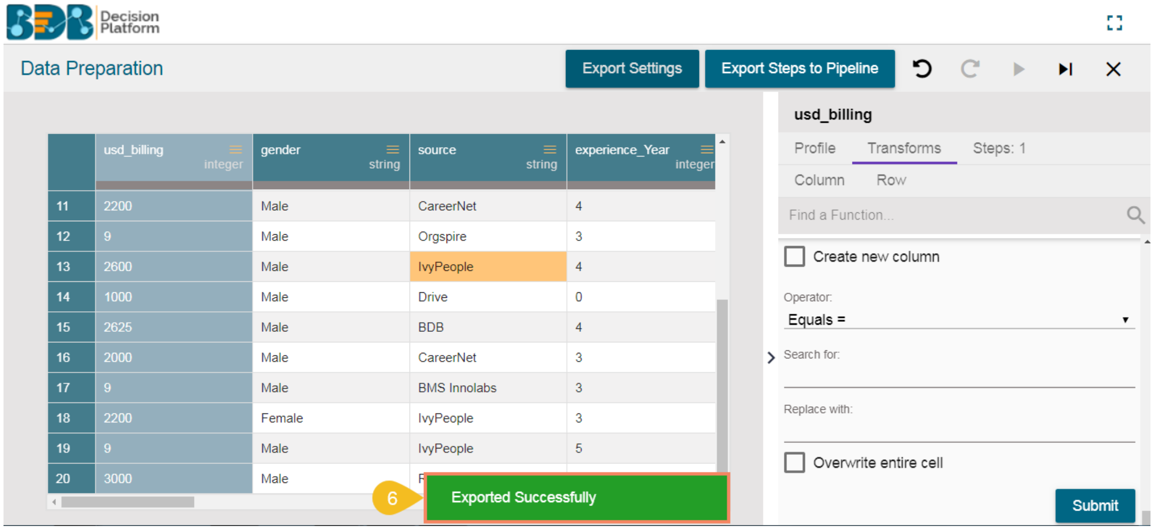
- The Export window opens.
- Provide the following details:
  - Data-Store Metadata Name: Provide a name for the data store metadata.
  - Select Mapping Id: Select a matching column from the drop-down menu.
  - Click the ‘**Export**’ option.



Data-Store Metadata Name: Data Prep Steps

Select Mapping Id: usd_billing

- A Success message appears to confirm.



usd_billing

	usd_billing integer	gender string	source string	experience_Year integer
11	2200	Male	CareerNet	4
12	9	Male	Orgspire	3
13	2600	Male	IvyPeople	4
14	1000	Male	Drive	0
15	2625	Male	BDB	4
16	2000	Male	CareerNet	3
17	9	Male	BMS Innolabs	3
18	2200	Female	IvyPeople	3
19	9	Male	IvyPeople	5
20	3000	Male	F	

☐ Create new column

Operator: Equals =

Search for:

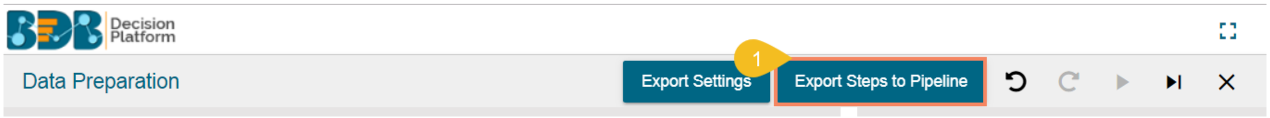
Replace with:

☐ Overwrite entire cell

Exported Successfully

- The settings get exported to the selected Elastic Settings.

b. **Export Steps to Pipeline**: This option provides an option to specify the name in which the steps/transforms created as part of cleansing must be exposed to the pipeline module of the platform.



c. **Undo** ↺ : Undo a list of last few transforms. This button gets enabled only if we have applied some transform on the data.

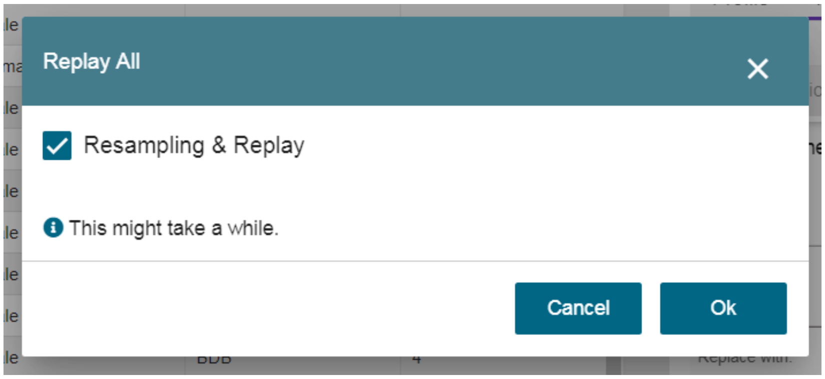
d. **Redo** ↻ : Redo a list of last few transforms, that was undone. If we have not undone any transform, then the '**redo**' icon gets disabled.

e. **Replay dirty** ▶ **:** The '**Replay Dirty**' option when applied on the data from a specific step it replays all the transforms which are listed after the selected transform in the list of steps.

- The '**Replay Dirty**' option gets enabled only when the user edits some transform step using the '**Steps**' tab.
- To indicate what all transform steps will be affected, the listed steps get colored in red.
- After the '**Replay Dirty**' function gets applied, all the steps that were colored in red become black and all the transforms get applied to the dataset.

f. **Replay All** ▶| **:** The Replay All option allows the user to resample the data and replay the steps on the new data sample. It is useful when there is a change in the underlying dataset. It updates the data in the grid applying all the steps (In case of edit or steps added after edit).

- Click the '**Replay All**' icon from the navigation pane.
- The '**Replay All**' window appears.
- Select '**Resampling & Replay'** option using the checkbox (if required).
- Click the '**Ok**' option.



ℹ This might take a while.

g. **Close the Preparation**: The user can exit from the preparation window and reach the landing page of data preparation.

Note: The standalone version of data preparation provides an option to export the prepared data to elastic so that that visualization modules can consume it.

8. Signing Out

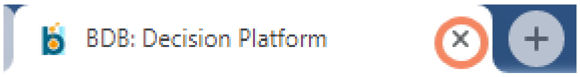
The users can Sign-out from the Data Preparation tab at any given stage, but preferable is that the users should complete all the preparation tasks they wish to perform and save it before closing the tab or singing out from the Platform.

The Signing Out process for the Data Preparation has two steps:

1. Closing the BDB Data Preparation

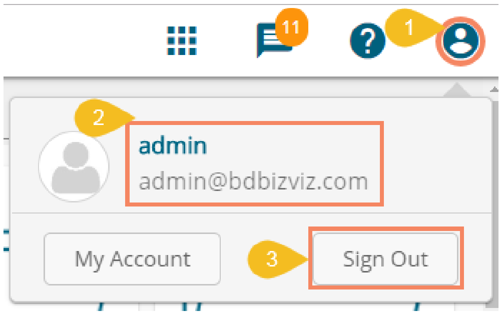
Once you have completed the Data Preparation tasks, save your work and close the Data Preparation tab.

Click the **'Close'** button (the 'X' on the right edge) from the Data Preparation tab.



2. Sign Out from the BDB Platform

i) Click the '**User**' icon on the Platform homepage.
ii) A menu appears with the logged in user details (User's name and email id).
iii) Click '**Sign Out**.'



iv) The user successfully signs off from the **BDB Platform**.

Note: Clicking on the '**Sign Out**' option redirects the user back to the login page of the BDB platform.