What is Data wrangling?

Data wrangling is a technique to

* Cleaning the data
* Organizing the data
* Transforming the data

We do apply data wrangling on original data. After all the pre-processing steps this data becomes useful for analysis.

Steps of data wrangling in Trifacta API

* Importing data
* Creating or adding a flow
* Building recipe
* Generating output
* Validation

Validation step checks all the transformations that has done to dataset under the recipe section.

**How to add a dataset to flow?**

Go to flow…then select the option create flow

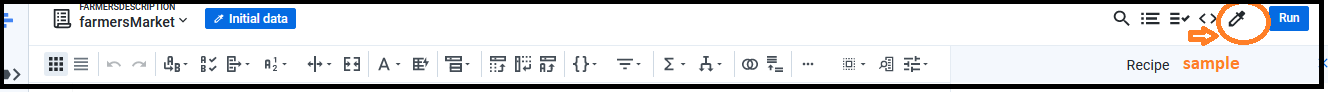
Click on connect to your data …then select the dataset which you want to transform.

Click on Edit recipe

The page will load your data.

For details, please check the documents.

**What is sample?**

****

The dataset which display in flow is not the full data. It is full only when the size of dataset is smaller than 10MB. If the dataset is larger than 10MB then we can visualize only a subset of dataset. We can select samples based on following options as

* Random
* Filter-based
* Anomaly-based
* Stratified
* Cluster-based

There are two types of views in trifecta.

1. Tabular view
2. Column view



We usually apply transform on grid view data.

**Dataset** : Farmers Markets in the United States

<https://www.kaggle.com/datasets/madeleineferguson/farmers-markets-in-the-united-states?resource=download>

1. Apply transform to updateTime

Click on column details(updateTime)

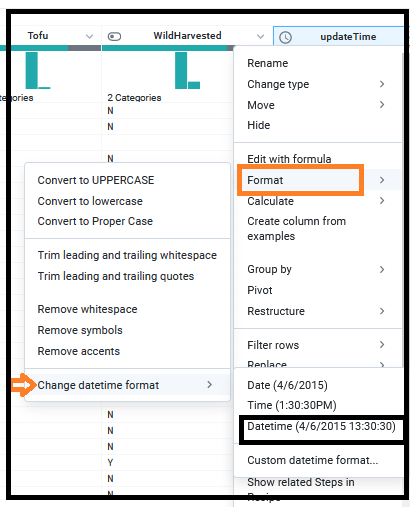


Column “updateTime” has mismatched values.

**Format dates**

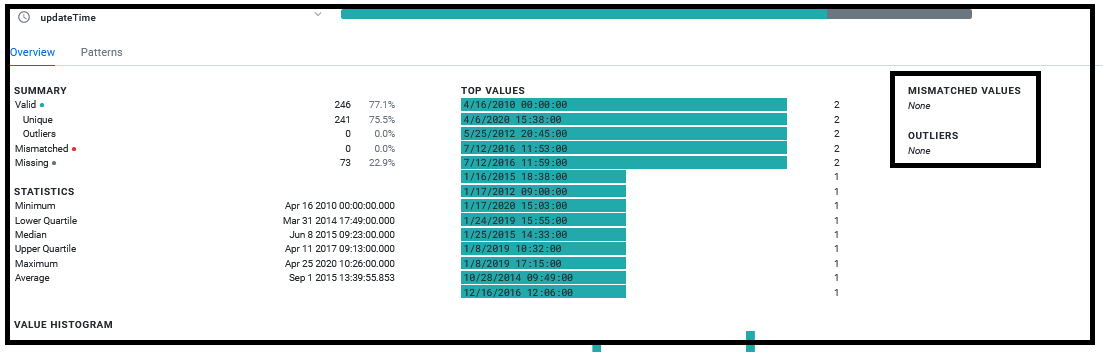
**How to remove mismatched values related to datetime format?**

Select the option format and then change datetime format



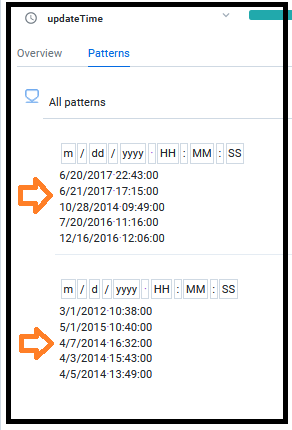
**Result**

Go to column details and check for mismatched values (does it exist ?).



Click on column details.

It will show: how many patterns that the column contains.



The result shows that the column “updateTime” contains 2 patterns.

Click on the first pattern and check the suggestions from RHS.

Select the option convert and click on change format of update time using 2 patterns and click Add.

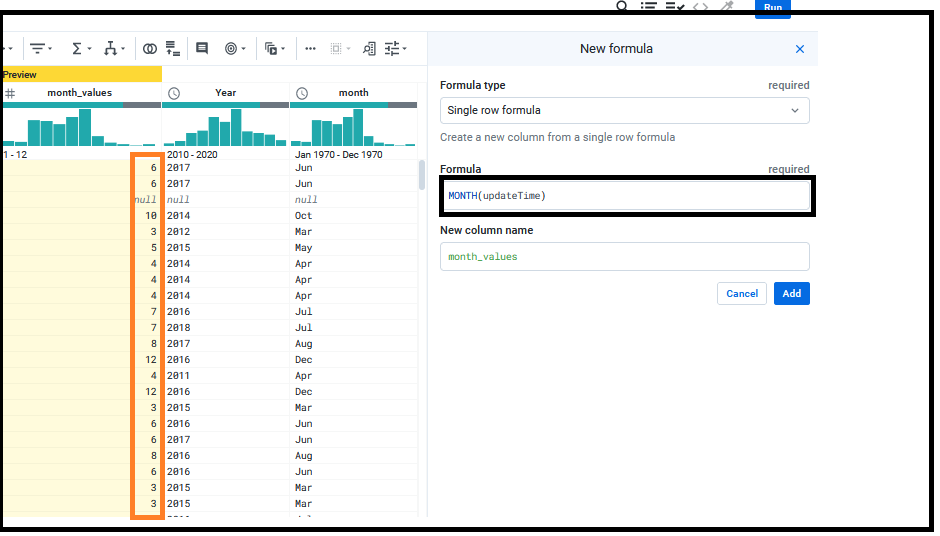
Check the preview option, what it is going to change…….

Then go to column details and check the details about the “updateTime” column.

Q. How to remove null values from updateTime?

Steps:

1. Go to New formula
2. DATEFORMAT(updateTime, 'MMM')
3. Create a new column name month and click add.

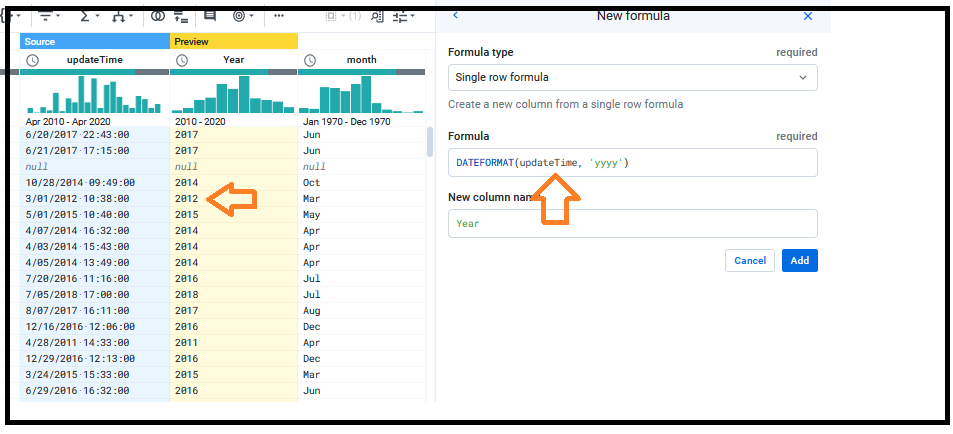


OR

You can use the option extract from column dropdown options.

* Datetime
  + It has options year, month, day of the month/year etc..
  + You can choose month or year.
  + Create a new column.

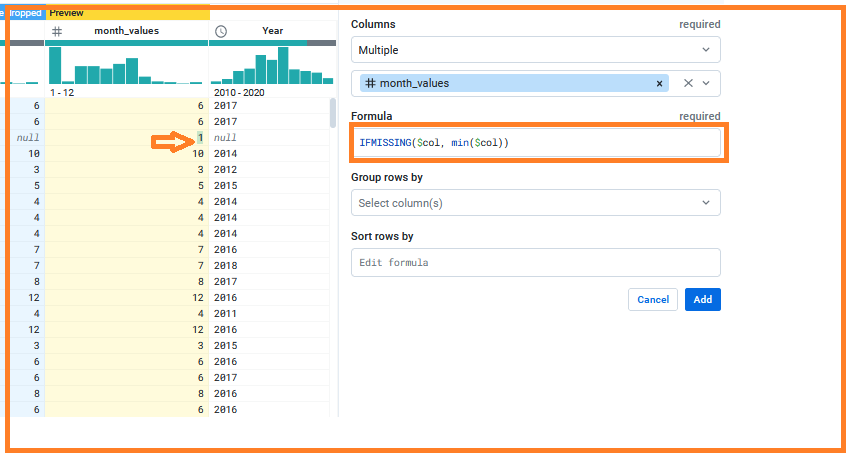
1. Create a new column name ‘year’ and click add.



1. Find min value for columns ‘year’ and ‘month’ columns.

Steps:

Go to the suggestion box, then select set option and click on Edit.

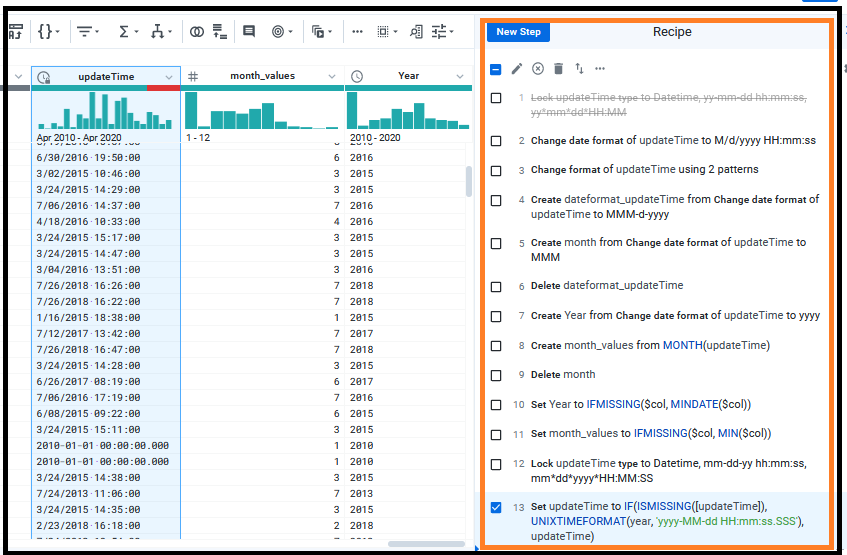


* + IFMISSING($col, MIN($col))
* Apply same method to ‘year’ column also.
  + IFMISSING($col, mindate($col))

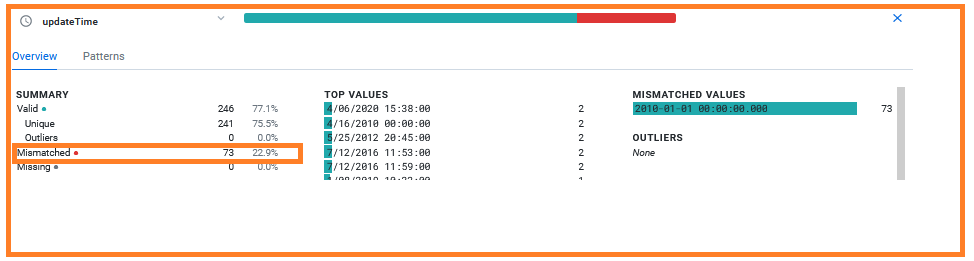
Next task is to replace the ‘null’ values from updateTime column by using month\_values and year columns.

**Steps**

* Go to suggestion box
  + Select the option **set** and write the below given formula.
* IF(ISMISSING(updateTime),UNIXTIMEFORMAT(year, "yyyy-MM-dd HH:mm:ss.SSS"),updateTime)

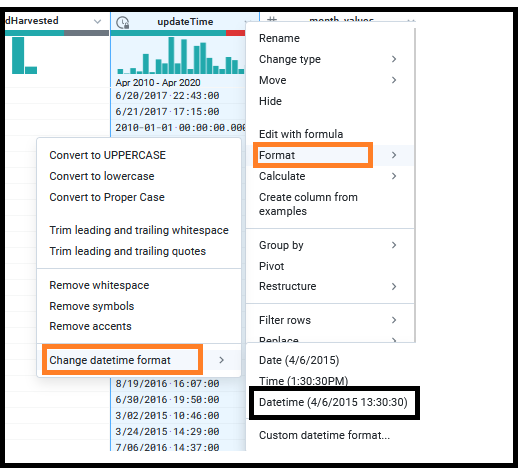


**How to remove Mismatched values?**

****

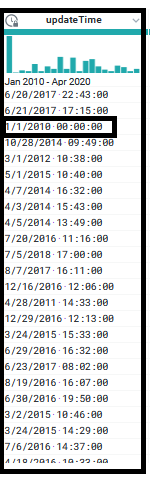
Click on ‘updateTime’ column and select the option format. We will get following options as

* Convert to uppercase
* Convert to lower case
* Remove whitespace
* Remove symbols
* Remove accents
* Change date time format
* Select the option Change date time format and select the option Datetime.



**Result**

You can check the column ‘updateTime’ has no mismatched values again and the value 2010-01-01 00:00:00:0000 is replaced with 1/1/2020.00:00:00.



End