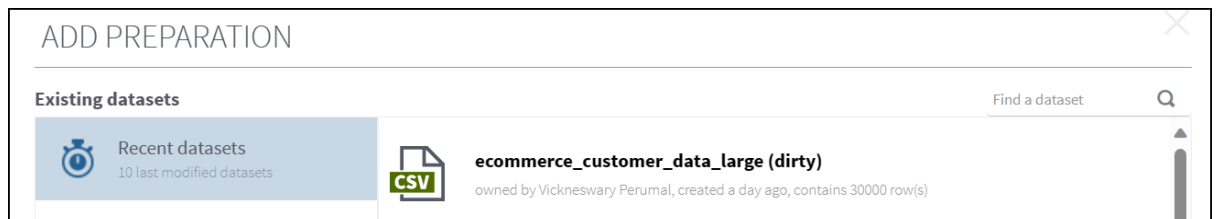


Talend Data Preparation

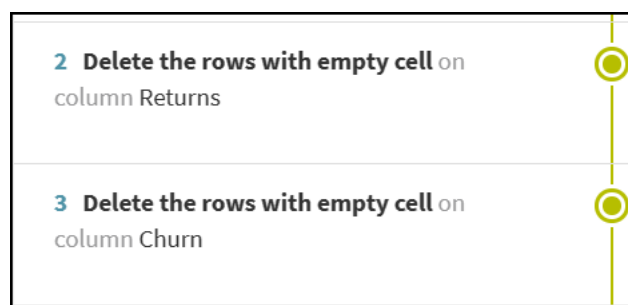
1. The dataset "ecommerce_customer_data_large (dirty)" was successfully loaded into Talend Data Preparation for cleaning and preparation.



2. Data Cleaning:

Addressing Missing Values:

- Rows with missing values in the "Returns" and "Churn" columns were deleted due to their low count (less than 10), minimizing impact on overall analysis.



3. Correcting Data Inconsistencies:

- The "Find and group similar text" function was used to:
Address typo errors in the "Product_Category" column.
Standardize data formats in the "Gender" and "Payment Method" columns.

talend DATA PREP

ecommerce_customer_d

1 Delete row #6

2 Delete the rows with en column Returns

3 Delete the rows with en column Churn

4 Find and group similar t Clothing

5 Delete the rows with im column Total Purchase Am

FIND AND GROUP SIMILAR TEXT

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

<input checked="" type="checkbox"/>	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Clothings <input checked="" type="checkbox"/> Clothing	Replace value: Clothing
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Homee <input checked="" type="checkbox"/> Home	Replace value: Home

SUBMIT

FIND AND GROUP SIMILAR TEXT

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

<input checked="" type="checkbox"/>	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> CASH <input checked="" type="checkbox"/> Cash	Replace value: Cash
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Paypal <input checked="" type="checkbox"/> PayPal	Replace value: PayPal

SUBMIT

FIND AND GROUP SIMILAR TEXT

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

<input checked="" type="checkbox"/>	These values have been found	This value will be kept
<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/> Female <input checked="" type="checkbox"/> female	Replace value: Female

SUBMIT

4. Removing Invalid Data:

- Rows with invalid records in the "Total Purchase Amount" column were deleted due to their limited quantity (less than 5).

5 Delete the rows with invalid cell on column Total Purchase Amount

5. The "Fill cells with value" function rectified inconsistencies in the "Gender" column.

7 Fill cells with value on column Gender

Gender = F x

Use with:

Value

Value:

Female

SUBMIT

ecommerce_customer_data_large (dirty) PREPARATION

Gender = M x

Use with:

Value

Value:

Male

SUBMIT

6. Standardizing Date Formats:

- The "Change date format" function ensured consistency in date representations throughout the dataset.

13 Change date format on column Purchase Date

Current format:

I don't know, best guess

New format:

custom

Your format:

dd-MM-yyyy

SUBMIT

7. Overview of the data preprocessing. With the completion of data preprocessing, the dataset is now prepared for further analysis and modeling.

