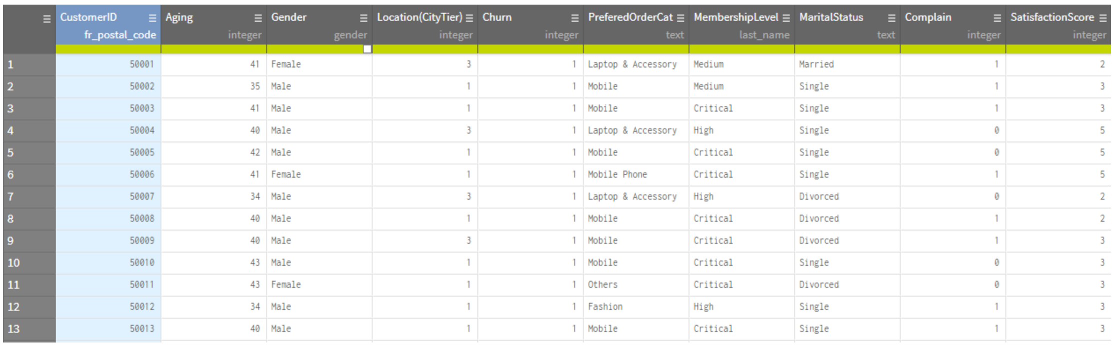
Talend Preparation

Dataset1: Customer

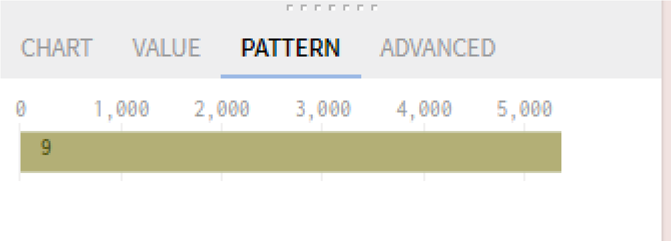
Before



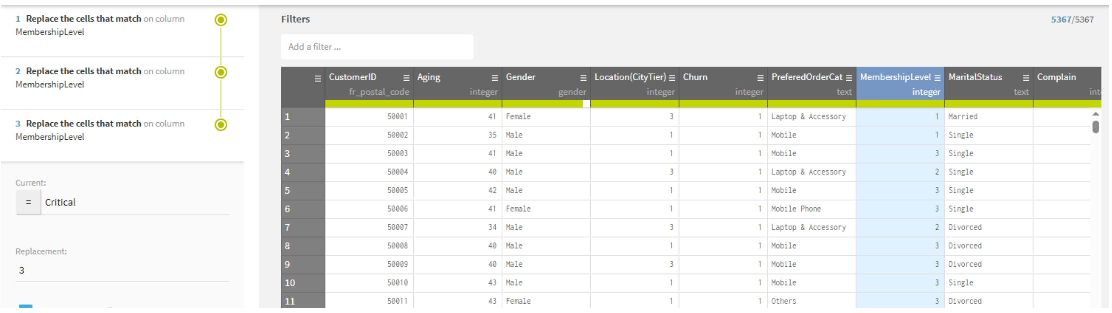
	CustomerID	Aging	Gender	Location(CityTier)	Churn	PreferedOrderCat	MembershipLevel	MaritalStatus	Complain	SatisfactionScore
	fr_postal_code	integer	gender	integer	integer	text	last_name	text	integer	integer
1	50001	41	Female	3	1	Laptop & Accessory	Medium	Married	1	2
2	50002	35	Male	1	1	Mobile	Medium	Single	1	3
3	50003	41	Male	1	1	Mobile	Critical	Single	1	3
4	50004	40	Male	3	1	Laptop & Accessory	High	Single	0	5
5	50005	42	Male	1	1	Mobile	Critical	Single	0	5
6	50006	41	Female	1	1	Mobile Phone	Critical	Single	1	5
7	50007	34	Male	3	1	Laptop & Accessory	High	Divorced	0	2
8	50008	40	Male	1	1	Mobile	Critical	Divorced	1	2
9	50009	40	Male	3	1	Mobile	Critical	Divorced	1	3
10	50010	43	Male	1	1	Mobile	Critical	Single	0	3
11	50011	43	Female	1	1	Others	Critical	Divorced	0	3
12	50012	34	Male	1	1	Fashion	High	Single	1	3
13	50013	40	Male	1	1	Mobile	Critical	Single	1	3

- **Operation1:**
  Number of missing values for each column:
  Gender:2

- **Operation 2:** Check "Parten" to make sure each column is formatted uniformly



- **Operation 3:** Modified the "MembershipLevel" column, adjusting the original levels "Medium, High, Critical" to numbers "1, 2, 3" to facilitate subsequent processing.



1 Replace the cells that match on column MembershipLevel

2 Replace the cells that match on column MembershipLevel

3 Replace the cells that match on column MembershipLevel

Current: = Critical

Replacement: 3

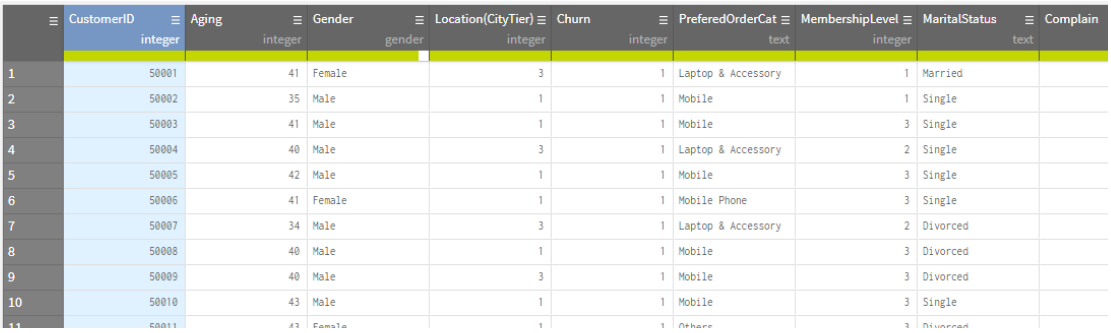
Filters

Add a filter ...

5367/5367

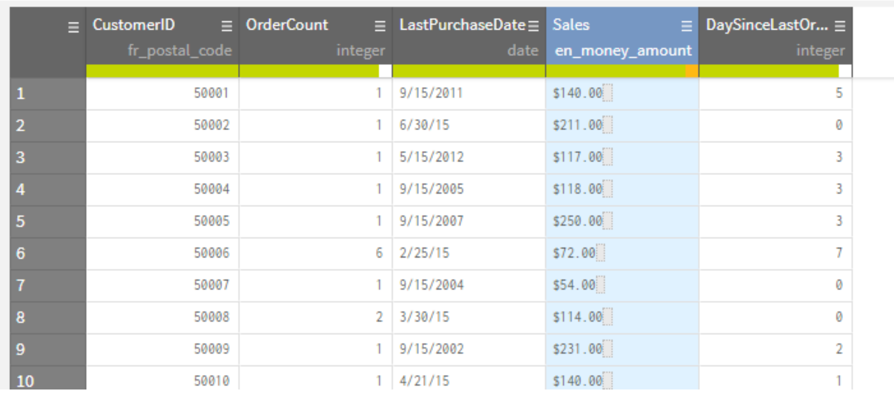
	CustomerID	Aging	Gender	Location(CityTier)	Churn	PreferedOrderCat	MembershipLevel	MaritalStatus	Complain
	fr_postal_code	integer	gender	integer	integer	text	integer	text	int
1	50001	41	Female	3	1	Laptop & Accessory	1	Married	
2	50002	35	Male	1	1	Mobile	1	Single	
3	50003	41	Male	1	1	Mobile	3	Single	
4	50004	40	Male	3	1	Laptop & Accessory	2	Single	
5	50005	42	Male	1	1	Mobile	3	Single	
6	50006	41	Female	1	1	Mobile Phone	3	Single	
7	50007	34	Male	3	1	Laptop & Accessory	2	Divorced	
8	50008	40	Male	1	1	Mobile	3	Divorced	
9	50009	40	Male	3	1	Mobile	3	Divorced	
10	50010	43	Male	1	1	Mobile	3	Single	
11	50011	43	Female	1	1	Others	3	Divorced	

- **Operation 4:** Modify the data type of "Customer ID"



	CustomerID	Aging	Gender	Location(CityTier)	Churn	PreferedOrderCat	MembershipLevel	MaritalStatus	Complain
	integer	integer	gender	integer	integer	text	integer	text	
1	50001	41	Female	3	1	Laptop & Accessory	1	Married	
2	50002	35	Male	1	1	Mobile	1	Single	
3	50003	41	Male	1	1	Mobile	3	Single	
4	50004	40	Male	3	1	Laptop & Accessory	2	Single	
5	50005	42	Male	1	1	Mobile	3	Single	
6	50006	41	Female	1	1	Mobile Phone	3	Single	
7	50007	34	Male	3	1	Laptop & Accessory	2	Divorced	
8	50008	40	Male	1	1	Mobile	3	Divorced	
9	50009	40	Male	3	1	Mobile	3	Divorced	
10	50010	43	Male	1	1	Mobile	3	Single	
11	50011	43	Female	1	1	Others	3	Divorced	

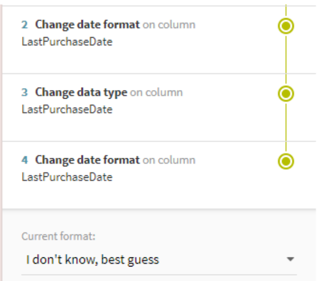
Dataset2: Order
Before



	CustomerID	OrderCount	LastPurchaseDate	Sales	DaySinceLastOr...
	fr_postal_code	integer	date	en_money_amount	integer
1	50001	1	9/15/2011	$140.00	5
2	50002	1	6/30/15	$211.00	0
3	50003	1	5/15/2012	$117.00	3
4	50004	1	9/15/2005	$118.00	3
5	50005	1	9/15/2007	$250.00	3
6	50006	6	2/25/15	$72.00	7
7	50007	1	9/15/2004	$54.00	0
8	50008	2	3/30/15	$114.00	0
9	50009	1	9/15/2002	$231.00	2
10	50010	1	4/21/15	$140.00	1

- **Operation1:**
  Number of missing values for each column:
  OrderCount:245
  DaySinceLastOrder:282

- **Operation2:** Unified date format

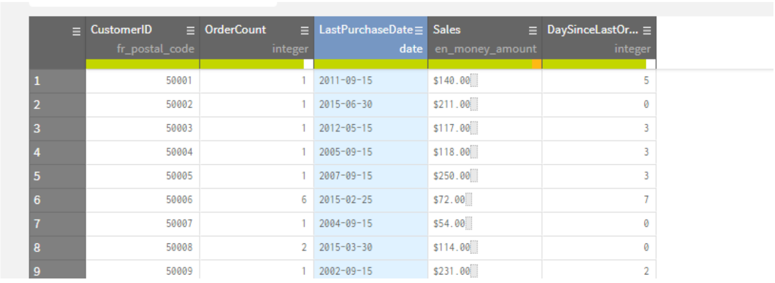


2 Change date format on column LastPurchaseDate

3 Change data type on column LastPurchaseDate

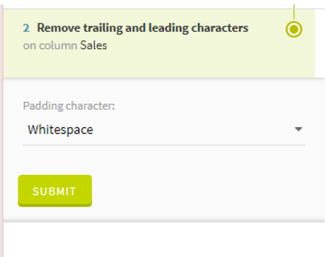
4 Change date format on column LastPurchaseDate

Current format: I don't know, best guess



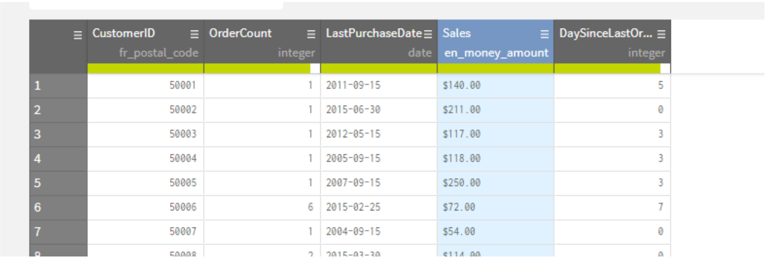
	CustomerID	OrderCount	LastPurchaseDate	Sales	DaySinceLastOr...
	fr_postal_code	integer	date	en_money_amount	integer
1	50001	1	2011-09-15	$140.00	5
2	50002	1	2015-06-30	$211.00	0
3	50003	1	2012-05-15	$117.00	3
4	50004	1	2005-09-15	$118.00	3
5	50005	1	2007-09-15	$250.00	3
6	50006	6	2015-02-25	$72.00	7
7	50007	1	2004-09-15	$54.00	0
8	50008	2	2015-03-30	$114.00	0
9	50009	1	2002-09-15	$231.00	2

- **Operation3:** Format the "Sales" column



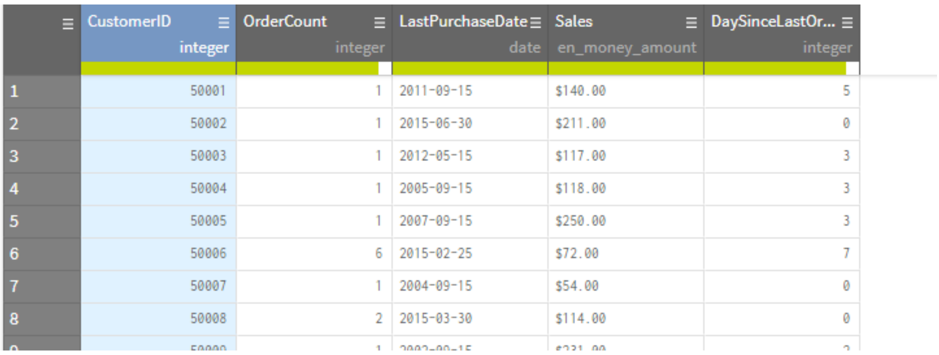
2 Remove trailing and leading characters on column Sales

Padding character: Whitespace



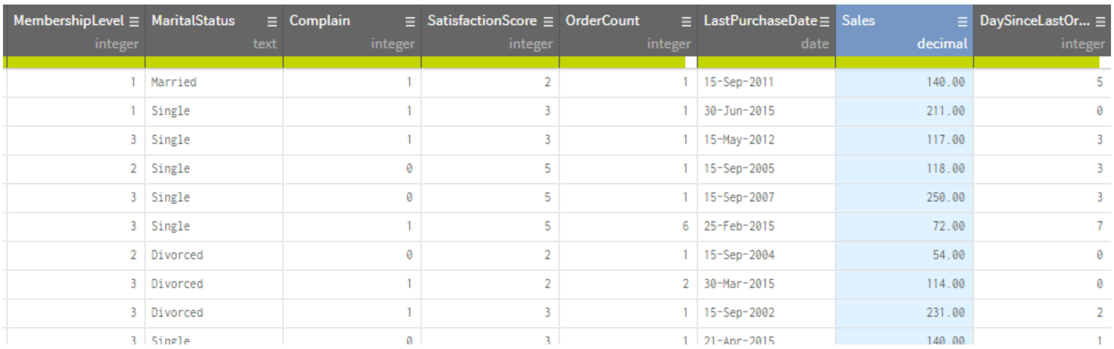
	CustomerID	OrderCount	LastPurchaseDate	Sales	DaySinceLastOr...
	fr_postal_code	integer	date	en_money_amount	integer
1	50001	1	2011-09-15	$140.00	5
2	50002	1	2015-06-30	$211.00	0
3	50003	1	2012-05-15	$117.00	3
4	50004	1	2005-09-15	$118.00	3
5	50005	1	2007-09-15	$250.00	3
6	50006	6	2015-02-25	$72.00	7
7	50007	1	2004-09-15	$54.00	0
8	50008	2	2015-03-30	$114.00	0

- **Operation4:** Modify the data type of "Customer ID"



	CustomerID	OrderCount	LastPurchaseDate	Sales	DaySinceLastOr...
	integer	integer	date	en_money_amount	integer
1	50001	1	2011-09-15	$140.00	5
2	50002	1	2015-06-30	$211.00	0
3	50003	1	2012-05-15	$117.00	3
4	50004	1	2005-09-15	$118.00	3
5	50005	1	2007-09-15	$250.00	3
6	50006	6	2015-02-25	$72.00	7
7	50007	1	2004-09-15	$54.00	0
8	50008	2	2015-03-30	$114.00	0
9	50009	1	2002-09-15	$231.00	2

- **Operation5:** Remove extra symbols for “sales”



MembershipLevel integer	MaritalStatus text	Complain integer	SatisfactionScore integer	OrderCount integer	LastPurchaseDate date	Sales decimal	DaySinceLastOr... integer
1	Married	1	2	1	15-Sep-2011	140.00	5
1	Single	1	3	1	30-Jun-2015	211.00	0
3	Single	1	3	1	15-May-2012	117.00	3
2	Single	0	5	1	15-Sep-2005	118.00	3
3	Single	0	5	1	15-Sep-2007	250.00	3
3	Single	1	5	6	25-Feb-2015	72.00	7
2	Divorced	0	2	1	15-Sep-2004	54.00	0
3	Divorced	1	2	2	30-Mar-2015	114.00	0
3	Divorced	1	3	1	15-Sep-2002	231.00	2
3	Single	0	3	1	21-Apr-2015	140.00	1