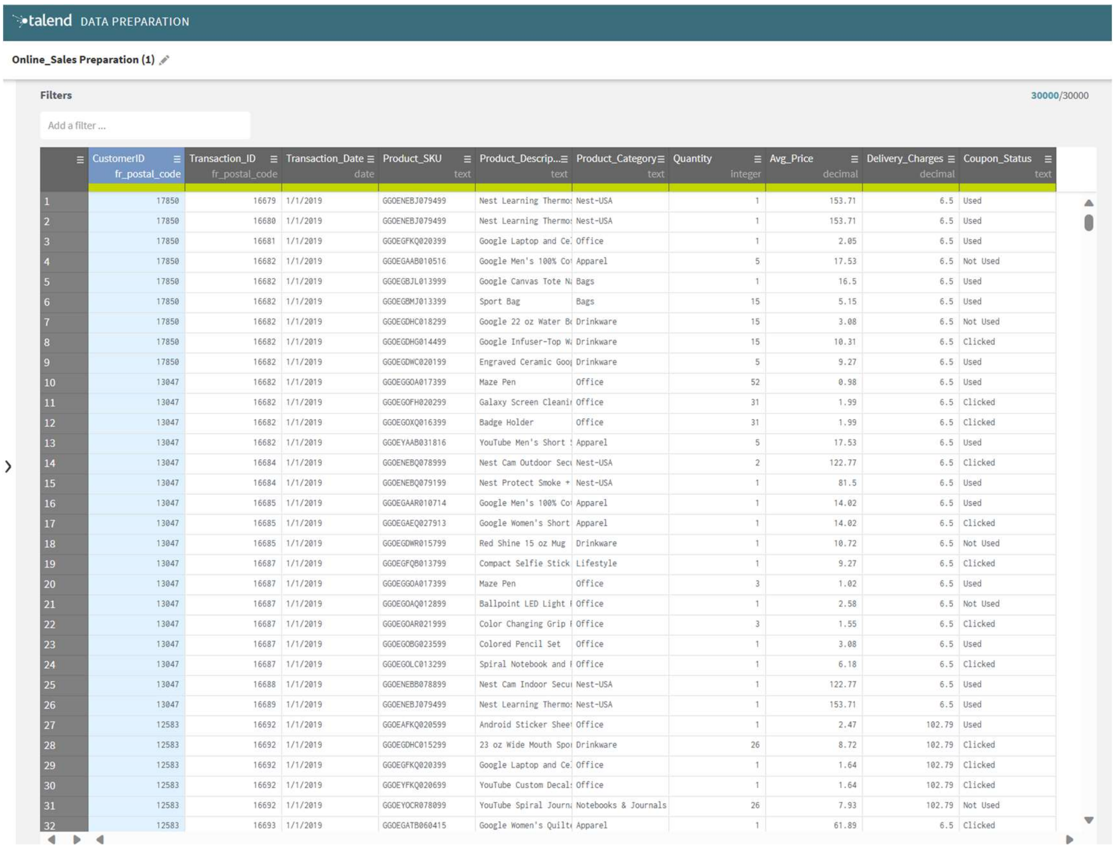
Talend Data Preparation

The original Online_Sales.csv dataset has 52924 rows while Talend Data Preparation can only access 30000 rows at most, so I split the dataset into two files, one with 30000 rows and the other with 22924 rows, as shown below:



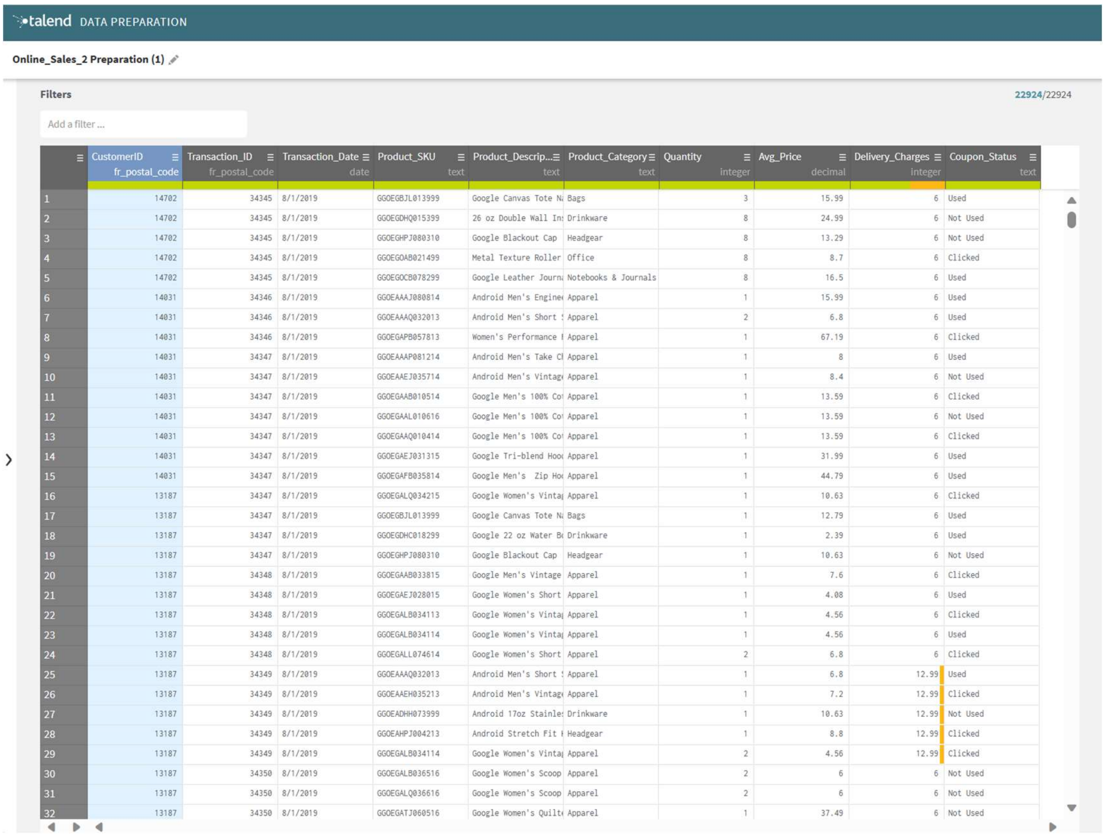
Online_Sales Preparation (1)

Filters

Add a filter ...

30000/30000

	CustomerID	Transaction_ID	Transaction_Date	Product_SKU	Product_Descrip...	Product_Category	Quantity	Avg_Price	Delivery_Charges	Coupon_Status
	fr_postal_code	fr_postal_code	date	text	text	text	integer	decimal	decimal	text
1	17850	16679	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
2	17850	16680	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
3	17850	16681	1/1/2019	GGOEGFKQ020399	Google Laptop and Ce	Office	1	2.05	6.5	Used
4	17850	16682	1/1/2019	GGOEGAAB010516	Google Men's 100% Co	Apparel	5	17.53	6.5	Not Used
5	17850	16682	1/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	1	16.5	6.5	Used
6	17850	16682	1/1/2019	GGOEGBMJ013399	Sport Bag	Bags	15	5.15	6.5	Used
7	17850	16682	1/1/2019	GGOEGDHC018299	Google 22 oz Water B	Drinkware	15	3.08	6.5	Not Used
8	17850	16682	1/1/2019	GGOEGDHG014499	Google Infuser-Top W	Drinkware	15	10.31	6.5	Clicked
9	17850	16682	1/1/2019	GGOEGDWC020199	Engraved Ceramic Goo	Drinkware	5	9.27	6.5	Used
10	13047	16682	1/1/2019	GGOEGGOA017399	Maze Pen	Office	52	0.98	6.5	Used
11	13047	16682	1/1/2019	GGOEGOFH020299	Galaxy Screen Cleani	Office	31	1.99	6.5	Clicked
12	13047	16682	1/1/2019	GGOEGOXQ016399	Badge Holder	Office	31	1.99	6.5	Clicked
13	13047	16682	1/1/2019	GGOEYAAB031816	YouTube Men's Short	Apparel	5	17.53	6.5	Used
14	13047	16684	1/1/2019	GGOENEBQ078999	Nest Cam Outdoor Sec	Nest-USA	2	122.77	6.5	Clicked
15	13047	16684	1/1/2019	GGOENEBQ079199	Nest Protect Smoke +	Nest-USA	1	81.5	6.5	Used
16	13047	16685	1/1/2019	GGOEGAAR010714	Google Men's 100% Co	Apparel	1	14.02	6.5	Used
17	13047	16685	1/1/2019	GGOEGAEQ027913	Google Women's Short	Apparel	1	14.02	6.5	Clicked
18	13047	16685	1/1/2019	GGOEGDWR015799	Red Shine 15 oz Mug	Drinkware	1	10.72	6.5	Not Used
19	13047	16687	1/1/2019	GGOEGFQB013799	Compact Selfie Stick	Lifestyle	1	9.27	6.5	Clicked
20	13047	16687	1/1/2019	GGOEGGOA017399	Maze Pen	Office	3	1.02	6.5	Used
21	13047	16687	1/1/2019	GGOEGOAQ012899	Ballpoint LED Light	Office	1	2.58	6.5	Not Used
22	13047	16687	1/1/2019	GGOEGOAR021999	Color Changing Grip	Office	3	1.55	6.5	Clicked
23	13047	16687	1/1/2019	GGOEGOBG023599	Colored Pencil Set	Office	1	3.08	6.5	Used
24	13047	16687	1/1/2019	GGOEGOLC013299	Spiral Notebook and	Office	1	6.18	6.5	Clicked
25	13047	16688	1/1/2019	GGOENEBB078899	Nest Cam Indoor Secu	Nest-USA	1	122.77	6.5	Used
26	13047	16689	1/1/2019	GGOENEBJ079499	Nest Learning Thermo	Nest-USA	1	153.71	6.5	Used
27	12583	16692	1/1/2019	GGOEAFKQ020599	Android Sticker Shee	Office	1	2.47	102.79	Used
28	12583	16692	1/1/2019	GGOEGDHC015299	23 oz Wide Mouth Spo	Drinkware	26	8.72	102.79	Clicked
29	12583	16692	1/1/2019	GGOEGFKQ020399	Google Laptop and Ce	Office	1	1.64	102.79	Clicked
30	12583	16692	1/1/2019	GGOEYFKQ020699	YouTube Custom Decal	Office	1	1.64	102.79	Clicked
31	12583	16692	1/1/2019	GGOEYOCR078099	YouTube Spiral Journ	Notebooks & Journals	26	7.93	102.79	Not Used
32	12583	16693	1/1/2019	GGOEGATB060415	Google Women's Quilt	Apparel	1	61.89	6.5	Clicked



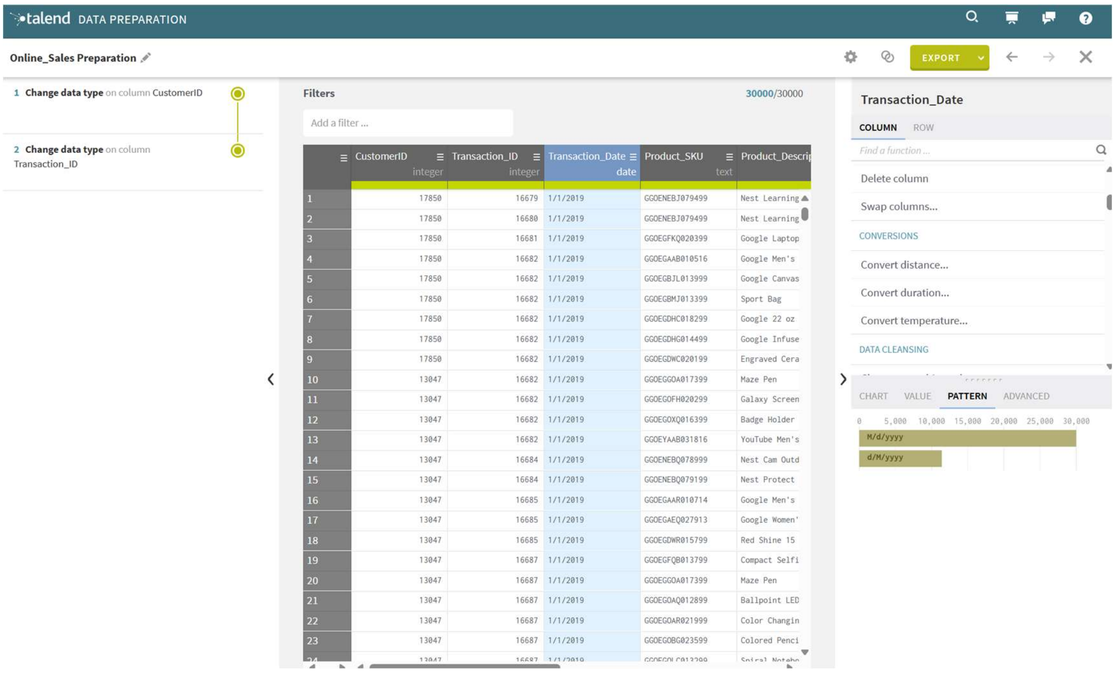
Online_Sales_2 Preparation (1)

22924/22924

	CustomerID fr_postal_code	Transaction_ID fr_postal_code	Transaction_Date date	Product_SKU text	Product_Descrip... text	Product_Category text	Quantity integer	Avg_Price decimal	Delivery_Charges integer	Coupon_Status text
1	14702	34345	8/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	3	15.99	6	Used
2	14702	34345	8/1/2019	GGOEGDHQ015399	26 oz Double Wall In	Drinkware	8	24.99	6	Not Used
3	14702	34345	8/1/2019	GGOEGHPJ080310	Google Blackout Cap	Headgear	8	13.29	6	Not Used
4	14702	34345	8/1/2019	GGOEGOAB021499	Metal Texture Roller	Office	8	8.7	6	Clicked
5	14702	34345	8/1/2019	GGOEGOCB078299	Google Leather Journ	Notebooks & Journals	8	16.5	6	Used
6	14031	34346	8/1/2019	GGOEAAAJ080814	Android Men's Engine	Apparel	1	15.99	6	Used
7	14031	34346	8/1/2019	GGOEAAAQ032013	Android Men's Short	Apparel	2	6.8	6	Used
8	14031	34346	8/1/2019	GGOEGAPB057813	Women's Performance	Apparel	1	67.19	6	Clicked
9	14031	34347	8/1/2019	GGOEAAAP081214	Android Men's Take C	Apparel	1	8	6	Used
10	14031	34347	8/1/2019	GGOEAAEJ035714	Android Men's Vintag	Apparel	1	8.4	6	Not Used
11	14031	34347	8/1/2019	GGOEGAAB010514	Google Men's 100% Co	Apparel	1	13.59	6	Clicked
12	14031	34347	8/1/2019	GGOEGAAL010616	Google Men's 100% Co	Apparel	1	13.59	6	Not Used
13	14031	34347	8/1/2019	GGOEGAAQ010414	Google Men's 100% Co	Apparel	1	13.59	6	Clicked
14	14031	34347	8/1/2019	GGOEGAEJ031315	Google Tri-blend Hoo	Apparel	1	31.99	6	Used
15	14031	34347	8/1/2019	GGOEGAFB035814	Google Men's Zip Ho	Apparel	1	44.79	6	Used
16	13187	34347	8/1/2019	GGOEGALQ034215	Google Women's Vinta	Apparel	1	10.63	6	Clicked
17	13187	34347	8/1/2019	GGOEGBJL013999	Google Canvas Tote N	Bags	1	12.79	6	Used
18	13187	34347	8/1/2019	GGOEGDHC018299	Google 22 oz Water B	Drinkware	1	2.39	6	Used
19	13187	34347	8/1/2019	GGOEGHPJ080310	Google Blackout Cap	Headgear	1	10.63	6	Not Used
20	13187	34348	8/1/2019	GGOEGAAB033815	Google Men's Vintage	Apparel	1	7.6	6	Clicked
21	13187	34348	8/1/2019	GGOEGAEJ028015	Google Women's Short	Apparel	1	4.08	6	Used
22	13187	34348	8/1/2019	GGOEGALB034113	Google Women's Vinta	Apparel	1	4.56	6	Clicked
23	13187	34348	8/1/2019	GGOEGALB034114	Google Women's Vinta	Apparel	1	4.56	6	Used
24	13187	34348	8/1/2019	GGOEGALL074614	Google Women's Short	Apparel	2	6.8	6	Clicked
25	13187	34349	8/1/2019	GGOEAAAQ032013	Android Men's Short	Apparel	1	6.8	12.99	Used
26	13187	34349	8/1/2019	GGOEAAEH035213	Android Men's Vintag	Apparel	1	7.2	12.99	Clicked
27	13187	34349	8/1/2019	GGOEADHH073999	Android 17oz Stainle	Drinkware	1	10.63	12.99	Not Used
28	13187	34349	8/1/2019	GGOEAHPJ004213	Android Stretch Fit	Headgear	1	8.8	12.99	Clicked
29	13187	34349	8/1/2019	GGOEGALB034114	Google Women's Vinta	Apparel	2	4.56	12.99	Clicked
30	13187	34350	8/1/2019	GGOEGALB036516	Google Women's Scoop	Apparel	2	6	6	Not Used
31	13187	34350	8/1/2019	GGOEGALQ036616	Google Women's Scoop	Apparel	2	6	6	Not Used
32	13187	34350	8/1/2019	GGOEGATJ060516	Google Women's Quilt	Apparel	1	37.49	6	Not Used

To make the data type more accurate and convenient for subsequent processing, I change the data type of CustomerID, Transaction_ID into integer, and Delivery_Charges into decimal.

In addition, the format of the Transaction_Date is not uniform, some shows 'M/d/yyyy', the others show 'd/M/yyyy', so I change the date format into 'yyyy-MM-dd' for subsequent processing:



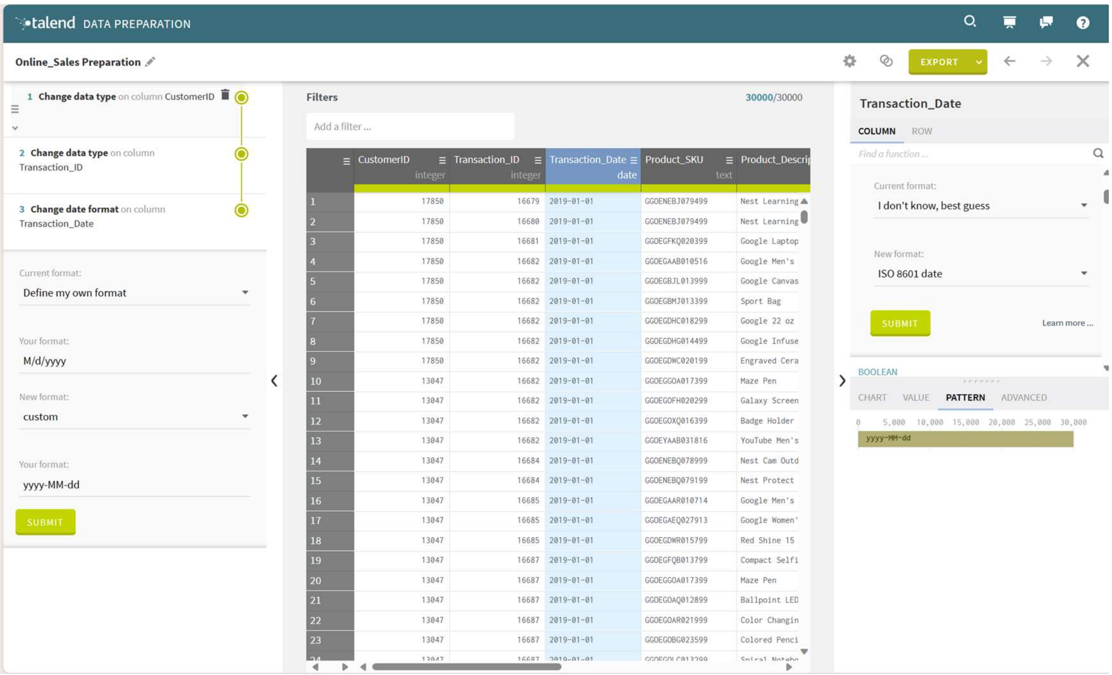
Online_Sales Preparation

1 Change data type on column CustomerID

2 Change data type on column Transaction_ID

30000/30000

	CustomerID integer	Transaction_ID integer	Transaction_Date date	Product_SKU text	Product_Descrip
1	17850	16679	1/1/2019	GGOENEBJ079499	Nest Learning
2	17850	16680	1/1/2019	GGOENEBJ079499	Nest Learning
3	17850	16681	1/1/2019	GGOEGFKQ020399	Google Laptop
4	17850	16682	1/1/2019	GGOEGAAB010516	Google Men's
5	17850	16682	1/1/2019	GGOEGBJL013999	Google Canvas
6	17850	16682	1/1/2019	GGOEGBMJ013399	Sport Bag
7	17850	16682	1/1/2019	GGOEGDHC018299	Google 22 oz
8	17850	16682	1/1/2019	GGOEGDHG014499	Google Infuse
9	17850	16682	1/1/2019	GGOEGDWC020199	Engraved Cera
10	13047	16682	1/1/2019	GGOEGGOA017399	Maze Pen
11	13047	16682	1/1/2019	GGOEGOFH020299	Galaxy Screen
12	13047	16682	1/1/2019	GGOEGOXQ016399	Badge Holder
13	13047	16682	1/1/2019	GGOEYAAB031816	YouTube Men's
14	13047	16684	1/1/2019	GGOENEBQ078999	Nest Cam Outd
15	13047	16684	1/1/2019	GGOENEBQ079199	Nest Protect
16	13047	16685	1/1/2019	GGOEGAAR010714	Google Men's
17	13047	16685	1/1/2019	GGOEGAEQ027913	Google Women'
18	13047	16685	1/1/2019	GGOEGDWR015799	Red Shine 15
19	13047	16687	1/1/2019	GGOEGFQB013799	Compact Selfi
20	13047	16687	1/1/2019	GGOEGGOA017399	Maze Pen
21	13047	16687	1/1/2019	GGOEGOAQ012899	Ballpoint LED
22	13047	16687	1/1/2019	GGOEGOAR021999	Color Changin
23	13047	16687	1/1/2019	GGOEGOBG023599	Colored Penci
24	13047	16687	1/1/2019	GGOEGOLC013299	Spiral Notebo



talend DATA PREPARATION

Online_Sales Preparation

1 Change data type on column CustomerID

2 Change data type on column Transaction_ID

3 Change date format on column Transaction_Date

Current format: Define my own format

Your format: M/d/yyyy

New format: custom

Your format: yyyy-MM-dd

SUBMIT

Filters

Add a filter ...

30000/30000

	CustomerID integer	Transaction_ID integer	Transaction_Date date	Product_SKU text	Product_Descrip
1	17850	16679	2019-01-01	GGOENEBJ079499	Nest Learning
2	17850	16680	2019-01-01	GGOENEBJ079499	Nest Learning
3	17850	16681	2019-01-01	GGOEGFKQ020399	Google Laptop
4	17850	16682	2019-01-01	GGOEGAAB010516	Google Men's
5	17850	16682	2019-01-01	GGOEGBJL013999	Google Canvas
6	17850	16682	2019-01-01	GGOEGBMJ013399	Sport Bag
7	17850	16682	2019-01-01	GGOEGDHC018299	Google 22 oz
8	17850	16682	2019-01-01	GGOEGDHG014499	Google Infuse
9	17850	16682	2019-01-01	GGOEGDWC020199	Engraved Cera
10	13047	16682	2019-01-01	GGOEGGOA017399	Maze Pen
11	13047	16682	2019-01-01	GGOEGOFH020299	Galaxy Screen
12	13047	16682	2019-01-01	GGOEGOXQ016399	Badge Holder
13	13047	16682	2019-01-01	GGOEYAAB031816	YouTube Men's
14	13047	16684	2019-01-01	GGOENEBQ078999	Nest Cam Outd
15	13047	16684	2019-01-01	GGOENEBQ079199	Nest Protect
16	13047	16685	2019-01-01	GGOEGAAR010714	Google Men's
17	13047	16685	2019-01-01	GGOEGAEQ027913	Google Women'
18	13047	16685	2019-01-01	GGOEGDWR015799	Red Shine 15
19	13047	16687	2019-01-01	GGOEGFQB013799	Compact Selfi
20	13047	16687	2019-01-01	GGOEGGOA017399	Maze Pen
21	13047	16687	2019-01-01	GGOEGOAQ012899	Ballpoint LED
22	13047	16687	2019-01-01	GGOEGOAR021999	Color Changin
23	13047	16687	2019-01-01	GGOEGOBG023599	Colored Penci

Transaction_Date

COLUMN ROW

Find a function ...

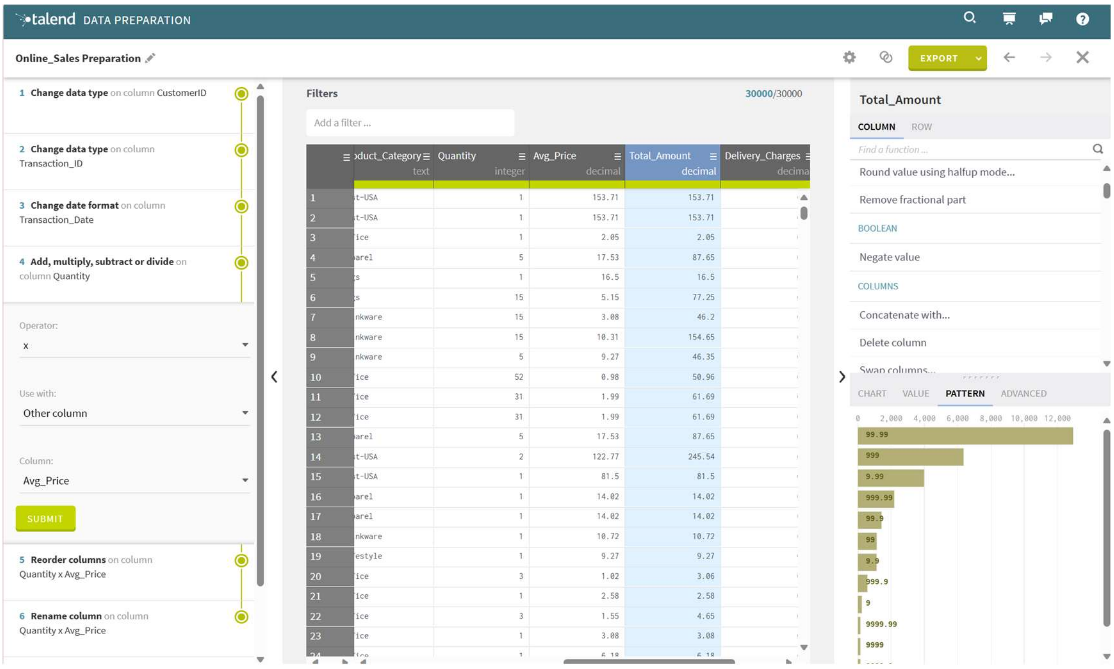
Current format: I don't know, best guess

New format: ISO 8601 date

SUBMIT Learn more ...

BOOLEAN

For subsequent analysis, I calculate the ‘Total_Amount’ spent on each transaction using ‘Quantity’ mutyplies ‘Avg_Price’, and rename the new column as ‘Total_Amount’.



talend DATA PREPARATION

Online_Sales Preparation

1 Change data type on column CustomerID

2 Change data type on column Transaction_ID

3 Change date format on column Transaction_Date

4 Add, multiply, subtract or divide on column Quantity

Operator: x

Use with: Other column

Column: Avg_Price

SUBMIT

5 Reorder columns on column Quantity x Avg_Price

6 Rename column on column Quantity x Avg_Price

Filters

Add a filter ...

30000/30000

	oduct_Category text	Quantity integer	Avg_Price decimal	Total_Amount decimal	Delivery_Charges decima
1	t-USA	1	153.71	153.71	
2	t-USA	1	153.71	153.71	
3	ice	1	2.05	2.05	
4	arel	5	17.53	87.65	
5	s	1	16.5	16.5	
6	s	15	5.15	77.25	
7	nkware	15	3.08	46.2	
8	nkware	15	10.31	154.65	
9	nkware	5	9.27	46.35	
10	ice	52	0.98	50.96	
11	ice	31	1.99	61.69	
12	ice	31	1.99	61.69	
13	arel	5	17.53	87.65	
14	t-USA	2	122.77	245.54	
15	t-USA	1	81.5	81.5	
16	arel	1	14.02	14.02	
17	arel	1	14.02	14.02	
18	nkware	1	10.72	10.72	
19	estyle	1	9.27	9.27	
20	ice	3	1.02	3.06	
21	ice	1	2.58	2.58	
22	ice	3	1.55	4.65	
23	ice	1	3.08	3.08	

Total_Amount

COLUMN ROW

Find a function ...

Round value using halfup mode...

Remove fractional part

BOOLEAN

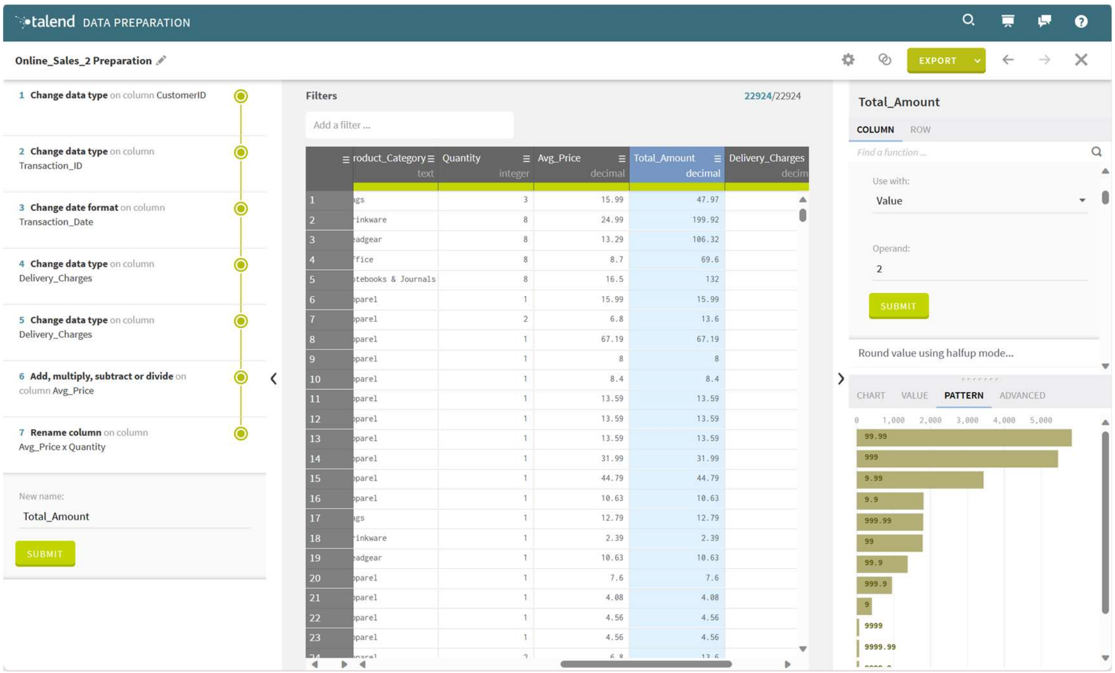
Negate value

COLUMNS

Concatenate with...

Delete column

Swap columns...



Online_Sales_2 Preparation

1 **Change data type** on column CustomerID

2 **Change data type** on column Transaction_ID

3 **Change date format** on column Transaction_Date

4 **Change data type** on column Delivery_Charges

5 **Change data type** on column Delivery_Charges

6 **Add, multiply, subtract or divide** on column Avg_Price

7 **Rename column** on column Avg_Price x Quantity

New name: Total_Amount

Filters

Add a filter ...

22924/22924

	roduct_Category	Quantity	Avg_Price	Total_Amount	Delivery_Charges
	text	integer	decimal	decimal	decim
1	gs	3	15.99	47.97	
2	rinkware	8	24.99	199.92	
3	eadgear	8	13.29	106.32	
4	ffice	8	8.7	69.6	
5	otebooks & Journals	8	16.5	132	
6	pparel	1	15.99	15.99	
7	pparel	2	6.8	13.6	
8	pparel	1	67.19	67.19	
9	pparel	1	8	8	
10	pparel	1	8.4	8.4	
11	pparel	1	13.59	13.59	
12	pparel	1	13.59	13.59	
13	pparel	1	13.59	13.59	
14	pparel	1	31.99	31.99	
15	pparel	1	44.79	44.79	
16	pparel	1	10.63	10.63	
17	gs	1	12.79	12.79	
18	rinkware	1	2.39	2.39	
19	eadgear	1	10.63	10.63	
20	pparel	1	7.6	7.6	
21	pparel	1	4.08	4.08	
22	pparel	1	4.56	4.56	
23	pparel	1	4.56	4.56	
24	pparel	2	6.8	13.6	

Total_Amount

COLUMN ROW

Find a function ...

Use with: Value

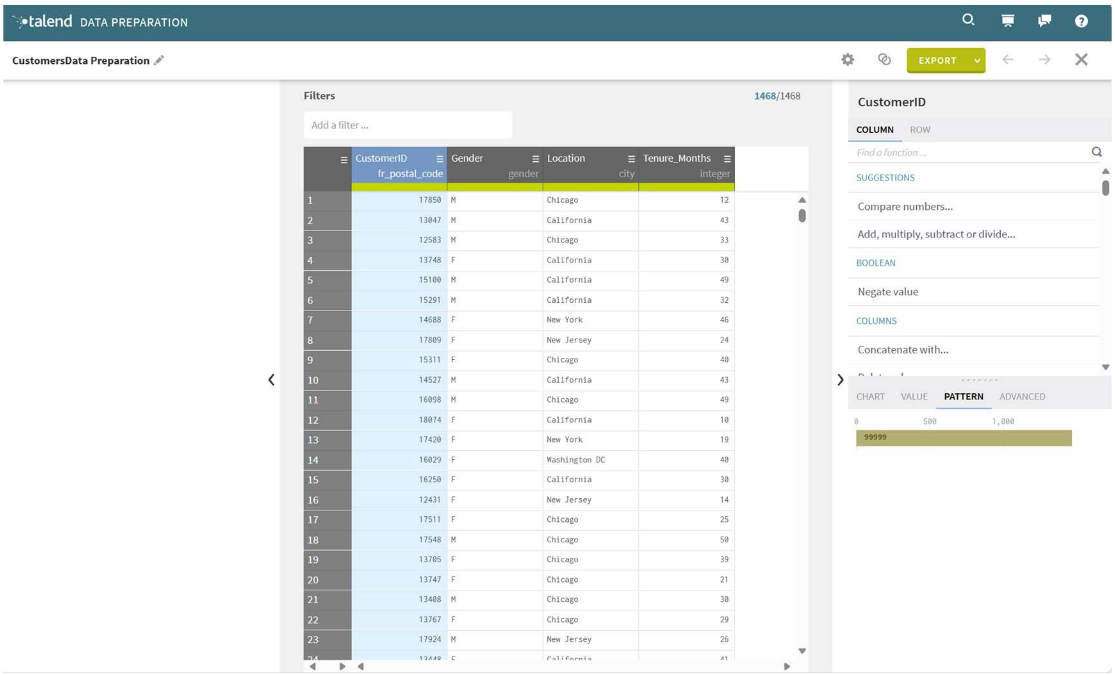
Operand: 2

Round value using halfup mode...

CHART VALUE PATTERN ADVANCED

After preprocessing, the two splitted sets of Online_Sales.csv file are cleaned and then combined as Online_Sales_preprocessed.xlsx.

The original CustomerData.xlsx file is as follows, there are no missing value or invalid value at all, and it need no more preprocessing:



CustomersData Preparation

Filters

Add a filter ...

1468/1468

	CustomerID	Gender	Location	Tenure_Months
	fr_postal_code	gender	city	integer
1	17850	M	Chicago	12
2	13047	M	California	43
3	12583	M	Chicago	33
4	13748	F	California	30
5	15100	M	California	49
6	15291	M	California	32
7	14688	F	New York	46
8	17809	F	New Jersey	24
9	15311	F	Chicago	40
10	14527	M	California	43
11	16098	M	Chicago	49
12	18074	F	California	10
13	17420	F	New York	19
14	16029	F	Washington DC	40
15	16250	F	California	30
16	12431	F	New Jersey	14
17	17511	F	Chicago	25
18	17548	M	Chicago	50
19	13705	F	Chicago	39
20	13747	F	Chicago	21
21	13408	M	Chicago	30
22	13767	F	Chicago	29
23	17924	M	New Jersey	26
24	13448	F	California	41

CustomerID

COLUMN ROW

Find a function ...

SUGGESTIONS

Compare numbers...

Add, multiply, subtract or divide...

BOOLEAN

Negate value

COLUMNS

Concatenate with...

CHART VALUE PATTERN ADVANCED