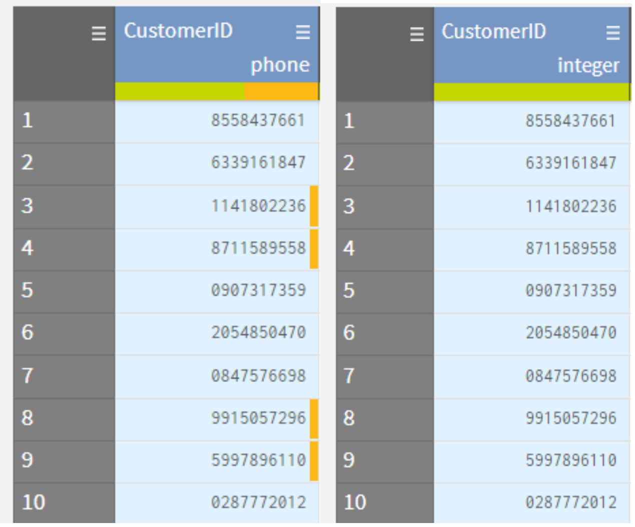
Part 3: Data Preparation Using Talend Data Preparation:

The data preprocessing is now moved to Talend Data Preparation (TDP) as TDP provide an easier yet robust way to process the data.

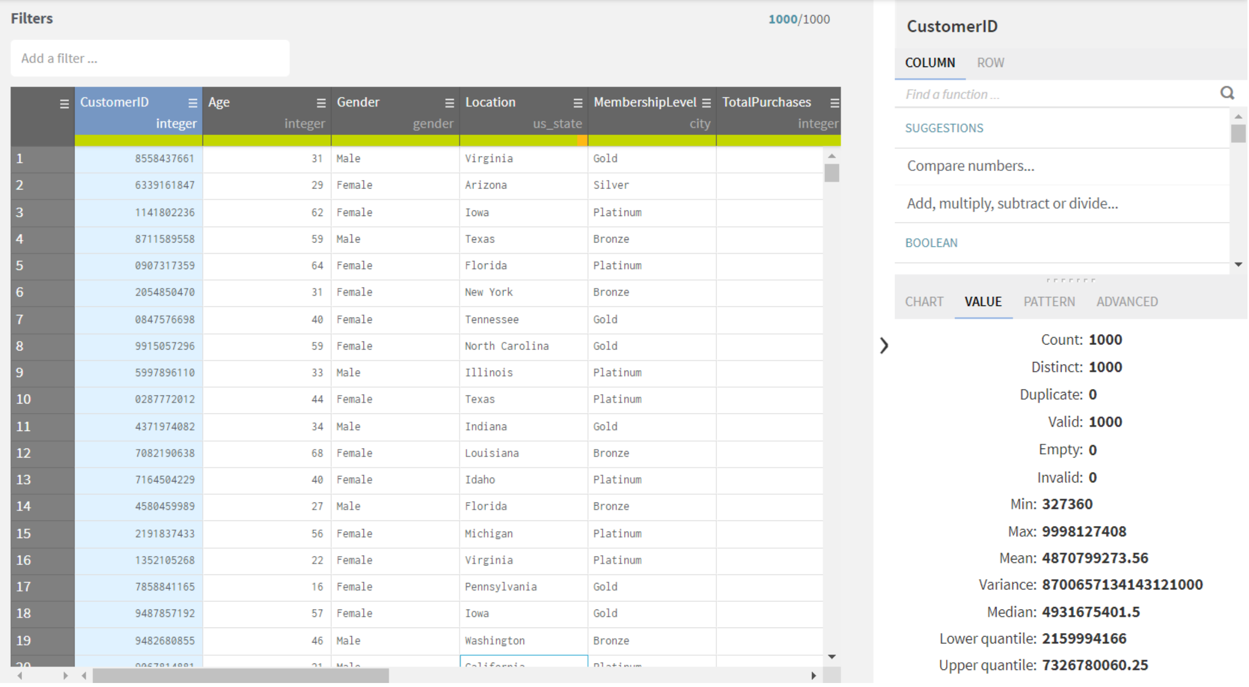
After importing the dataset to TDP, the data health bar for every column is inspected. The first step is to change the data type of CustomerID to Integer from Phone Number to eliminate the invalid data issues questioned by the tool. Note that TDP accepts all CustomerID values as integers, unlike TDI, indicating the CustomerID values from the original dataset might not have any issues with the data type.



	CustomerID phone
1	8558437661
2	6339161847
3	1141802236
4	8711589558
5	0907317359
6	2054850470
7	0847576698
8	9915057296
9	5997896110
10	0287772012

	CustomerID integer
1	8558437661
2	6339161847
3	1141802236
4	8711589558
5	0907317359
6	2054850470
7	0847576698
8	9915057296
9	5997896110
10	0287772012

Figure 8: Data health increased to perfect green after changing the data type of CustomerID.

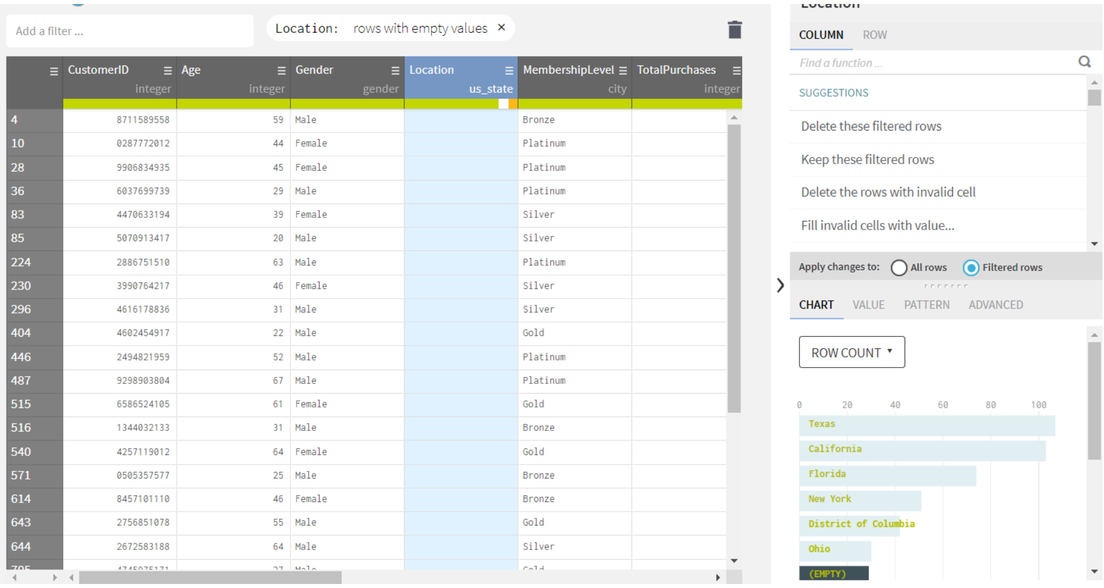


	CustomerID integer	Age integer	Gender gender	Location us_state	MembershipLevel city	TotalPurchases integer
1	8558437661	31	Male	Virginia	Gold	
2	6339161847	29	Female	Arizona	Silver	
3	1141802236	62	Female	Iowa	Platinum	
4	8711589558	59	Male	Texas	Bronze	
5	0907317359	64	Female	Florida	Platinum	
6	2054850470	31	Female	New York	Bronze	
7	0847576698	40	Female	Tennessee	Gold	
8	9915057296	59	Female	North Carolina	Gold	
9	5997896110	33	Male	Illinois	Platinum	
10	0287772012	44	Female	Texas	Platinum	
11	4371974082	34	Male	Indiana	Gold	
12	7082190638	68	Female	Louisiana	Bronze	
13	7164504229	40	Female	Idaho	Platinum	
14	4580459989	27	Male	Florida	Bronze	
15	2191837433	56	Female	Michigan	Platinum	
16	1352105268	22	Female	Virginia	Platinum	
17	7858841165	16	Female	Pennsylvania	Gold	
18	9487857192	57	Female	Iowa	Gold	
19	9482680855	46	Male	Washington	Bronze	

Count:	1000
Distinct:	1000
Duplicate:	0
Valid:	1000
Empty:	0
Invalid:	0
Min:	327360
Max:	9998127408
Mean:	4870799273.56
Variance:	8700657134143121000
Median:	4931675401.5
Lower quantile:	2159994166
Upper quantile:	7326780060.25

Figure 9: The right plane shows there is no duplicate CustomerID.

As shown in the output of Figure 9, as the duplicate check is performed, no duplicates are occurring at the CustomerID column which every row should have unique values. Hence, there is no action taken for the removal of duplicate data.



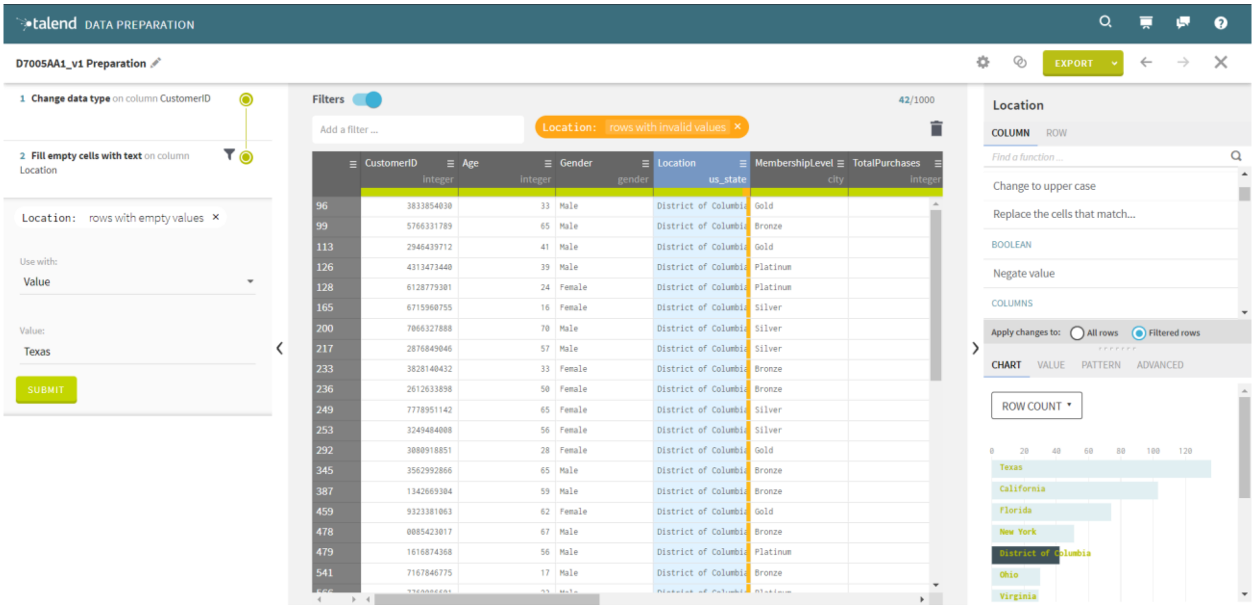
Location: rows with empty values

	CustomerID integer	Age integer	Gender gender	Location us_state	MembershipLevel city	TotalPurchases integer
4	8711589558	59	Male		Bronze	
10	0287772012	44	Female		Platinum	
28	9906834935	45	Female		Platinum	
36	6037699739	29	Male		Platinum	
83	4470633194	39	Female		Silver	
85	5070913417	20	Male		Silver	
224	2886751510	63	Male		Platinum	
230	3990764217	46	Female		Silver	
296	4616178836	31	Male		Silver	
404	4602454917	22	Male		Gold	
446	2494821959	52	Male		Platinum	
487	9298903804	67	Male		Platinum	
515	6586524105	61	Female		Gold	
516	1344032133	31	Male		Bronze	
540	4257119012	64	Female		Gold	
571	0505357577	25	Male		Bronze	
614	8457101110	46	Female		Bronze	
643	2756851078	55	Male		Gold	
644	2672583188	64	Male		Silver	

Apply changes to: ○ All rows ● Filtered rows

Figure 10: Handling missing values for Location.

Referring to Figure 10, there are missing values in location as designated during data generation. For this column, we select the state with the highest occurrence to perform data imputation for missing locations. The way to impute the data is shown in the left pane of Figure 8.



D7005AA1_v1 Preparation

1 **Change data type** on column CustomerID

2 **Fill empty cells with text** on column Location

Location: rows with empty values

Use with: Value

Value: Texas

Location: rows with invalid values

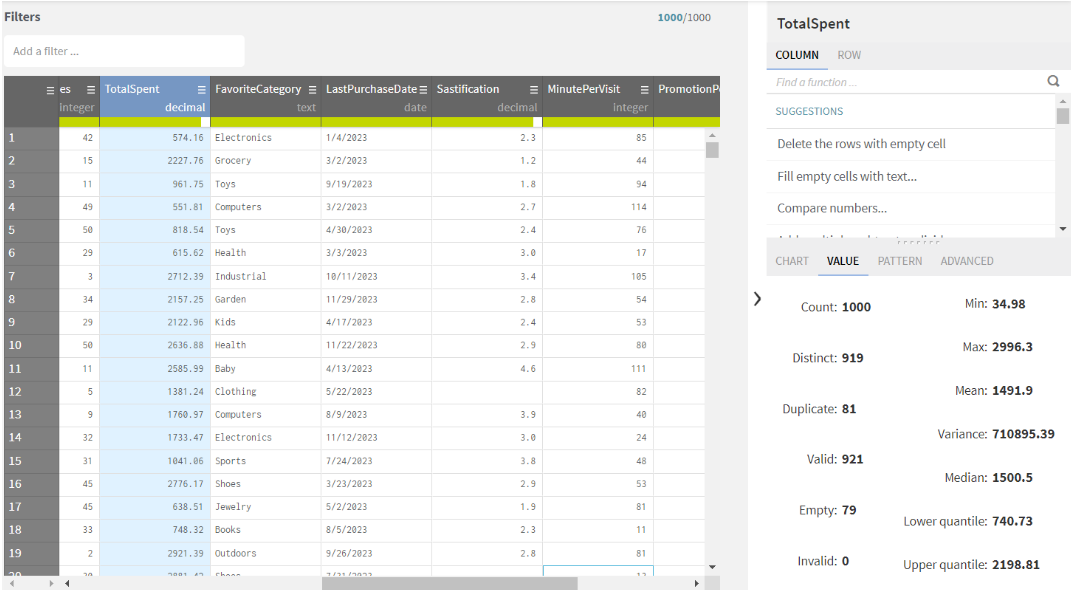
42/1000

	CustomerID integer	Age integer	Gender gender	Location us_state	MembershipLevel city	TotalPurchases integer
96	3833854030	33	Male	District of Columbi	Gold	
99	5766331789	65	Male	District of Columbi	Bronze	
113	2946439712	41	Male	District of Columbi	Gold	
126	4313473440	39	Male	District of Columbi	Platinum	
128	6128779301	24	Female	District of Columbi	Platinum	
165	6715960755	16	Female	District of Columbi	Silver	
200	7066327888	70	Male	District of Columbi	Silver	
217	2876849046	57	Male	District of Columbi	Silver	
233	3828140432	33	Female	District of Columbi	Bronze	
236	2612633898	50	Female	District of Columbi	Bronze	
249	7778951142	65	Female	District of Columbi	Silver	
253	3249484008	56	Female	District of Columbi	Silver	
292	3080918851	28	Female	District of Columbi	Gold	
345	3562992866	65	Male	District of Columbi	Bronze	
387	1342669304	59	Male	District of Columbi	Bronze	
459	9323381063	62	Female	District of Columbi	Gold	
478	0085423017	67	Male	District of Columbi	Bronze	
479	1616874368	56	Male	District of Columbi	Platinum	
541	7167846775	17	Male	District of Columbi	Bronze	

Apply changes to: ○ All rows ● Filtered rows

Figure 11: Location imputation and potential invalid location.

Besides missing values, there exist invalid values too. However, these values are the same, which is the District of Columbia, which is not a US state, but a federal district of the United States which not belong to any state. Hence, it is safe to ignore this flag as it is still an area of the United States.



Filters

Add a filter ...

1000/1000

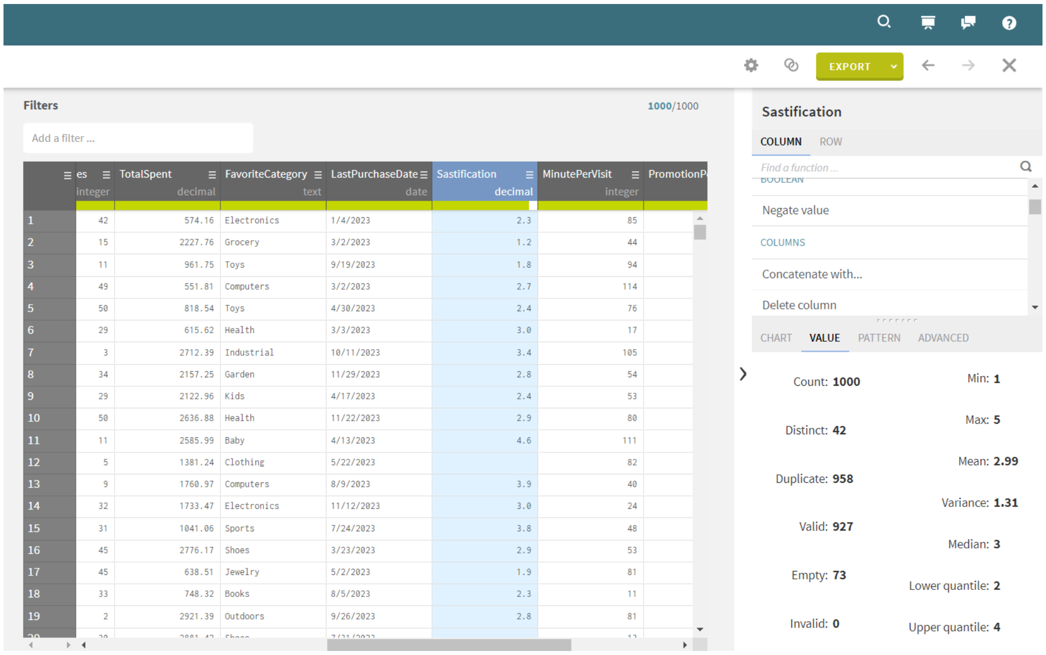
	es	TotalSpent	FavoriteCategory	LastPurchaseDate	Sastification	MinutePerVisit	PromotionP
	integer	decimal	text	date	decimal	integer	
1	42	574.16	Electronics	1/4/2023	2.3	85	
2	15	2227.76	Grocery	3/2/2023	1.2	44	
3	11	961.75	Toys	9/19/2023	1.8	94	
4	49	551.81	Computers	3/2/2023	2.7	114	
5	50	818.54	Toys	4/30/2023	2.4	76	
6	29	615.62	Health	3/3/2023	3.0	17	
7	3	2712.39	Industrial	10/11/2023	3.4	105	
8	34	2157.25	Garden	11/29/2023	2.8	54	
9	29	2122.96	Kids	4/17/2023	2.4	53	
10	50	2636.88	Health	11/22/2023	2.9	80	
11	11	2585.99	Baby	4/13/2023	4.6	111	
12	5	1381.24	Clothing	5/22/2023		82	
13	9	1760.97	Computers	8/9/2023	3.9	40	
14	32	1733.47	Electronics	11/12/2023	3.0	24	
15	31	1041.06	Sports	7/24/2023	3.8	48	
16	45	2776.17	Shoes	3/23/2023	2.9	53	
17	45	638.51	Jewelry	5/2/2023	1.9	81	
18	33	748.32	Books	8/5/2023	2.3	11	
19	2	2921.39	Outdoors	9/26/2023	2.8	81	

TotalSpent

Count: 1000
Distinct: 919
Duplicate: 81
Valid: 921
Empty: 79
Invalid: 0
Min: 34.98
Max: 2996.3
Mean: 1491.9
Variance: 710895.39
Median: 1500.5
Lower quantile: 740.73
Upper quantile: 2198.81

Figure 12: Deciding the imputation method of missing TotalSpent values.

For the TotalSpent, since the range of the value is large (34.98 to 2996.3), it is safer to select the median as the imputation of missing values to minimise the effect and bias caused by the potential outliers although the boxplot does not significantly skew to either side.



Filters

Add a filter ...

1000/1000

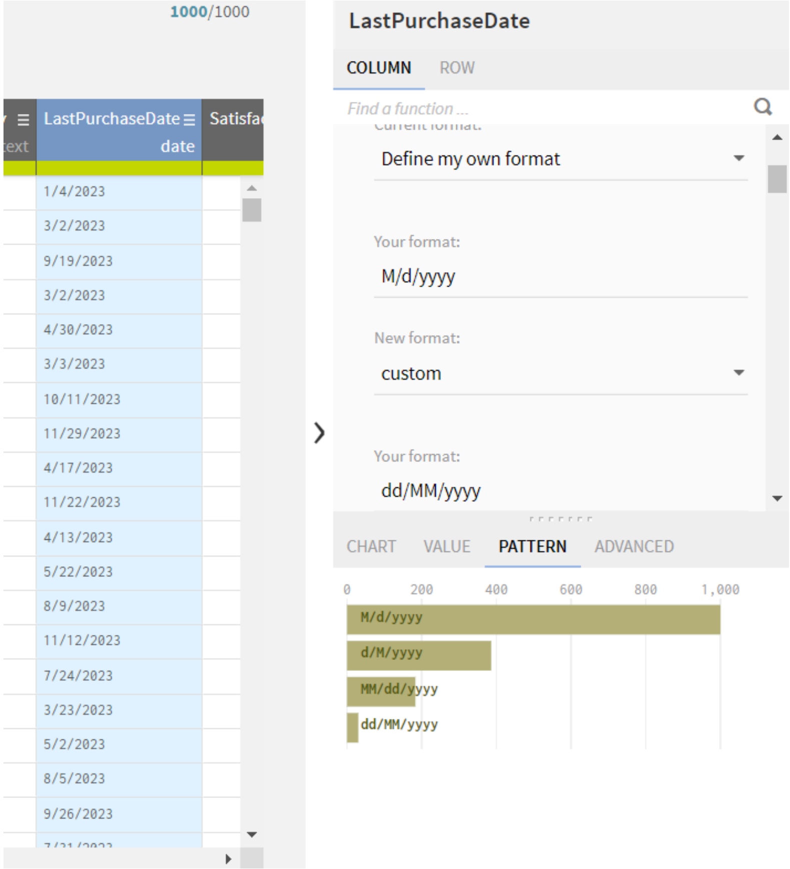
	es	TotalSpent	FavoriteCategory	LastPurchaseDate	Sastification	MinutePerVisit	PromotionP
	integer	decimal	text	date	decimal	integer	
1	42	574.16	Electronics	1/4/2023	2.3	85	
2	15	2227.76	Grocery	3/2/2023	1.2	44	
3	11	961.75	Toys	9/19/2023	1.8	94	
4	49	551.81	Computers	3/2/2023	2.7	114	
5	50	818.54	Toys	4/30/2023	2.4	76	
6	29	615.62	Health	3/3/2023	3.0	17	
7	3	2712.39	Industrial	10/11/2023	3.4	105	
8	34	2157.25	Garden	11/29/2023	2.8	54	
9	29	2122.96	Kids	4/17/2023	2.4	53	
10	50	2636.88	Health	11/22/2023	2.9	80	
11	11	2585.99	Baby	4/13/2023	4.6	111	
12	5	1381.24	Clothing	5/22/2023		82	
13	9	1760.97	Computers	8/9/2023	3.9	40	
14	32	1733.47	Electronics	11/12/2023	3.0	24	
15	31	1041.06	Sports	7/24/2023	3.8	48	
16	45	2776.17	Shoes	3/23/2023	2.9	53	
17	45	638.51	Jewelry	5/2/2023	1.9	81	
18	33	748.32	Books	8/5/2023	2.3	11	
19	2	2921.39	Outdoors	9/26/2023	2.8	81	

Sastification

Count: 1000
Distinct: 42
Duplicate: 958
Valid: 927
Empty: 73
Invalid: 0
Min: 1
Max: 5
Mean: 2.99
Variance: 1.31
Median: 3
Lower quantile: 2
Upper quantile: 4

Figure 13: Deciding the imputation method of missing Satisfaction values.

Regarding missing values of Satisfaction, the mean and median values are nearly identical which indicates that the distribution of values is possibly evenly distributed and both values are located at the middle of the range. Hence, the mean can be selected to be imputed. However, to ensure data formalisation which allows only one decimal place for input, the mean value of 2.99 is rounded off to 3.0 for imputation of missing values.



1000/1000

LastPurchaseDate date
1/4/2023
3/2/2023
9/19/2023
3/2/2023
4/30/2023
3/3/2023
10/11/2023
11/29/2023
4/17/2023
11/22/2023
4/13/2023
5/22/2023
8/9/2023
11/12/2023
7/24/2023
3/23/2023
5/2/2023
8/5/2023
9/26/2023

LastPurchaseDate

Current format: Define my own format

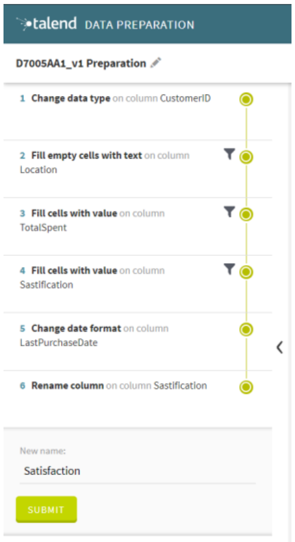
Your format: M/d/yyyy

New format: custom

Your format: dd/MM/yyyy

Figure 14: Transforming data format.

As this is a US-related dataset, the date format is recorded in American style. To change the date format to a format that is used more frequently in Malaysia, choose the “Change date format” at the right pane of the interface and apply the setting in Figure 14.



D7005AA1_v1 Preparation

1 Change data type on column CustomerID

2 Fill empty cells with text on column Location

3 Fill cells with value on column TotalSpent

4 Fill cells with value on column Sastification

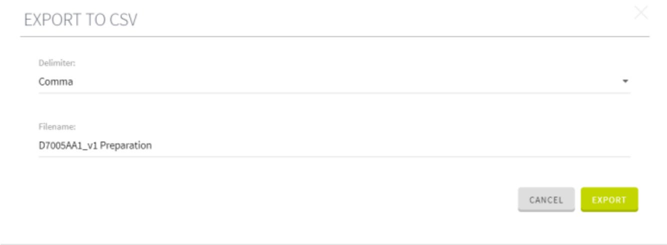
5 Change date format on column LastPurchaseDate

6 Rename column on column Sastification

New name: Satisfaction

Figure 15: Data processing workflow on Talend Data Preparation.

Finally, rename the unsatisfied column name (for example, wrong spelling) during the final check. As the data quality of all columns is satisfied, the processed dataset is exported into CSV format to the local environment as shown in Figure 16.


EXPORT TO CSV

Delimiter: Comma

Filename: D7005AA1_v1 Preparation

Figure 16: Exporting processed dataset.