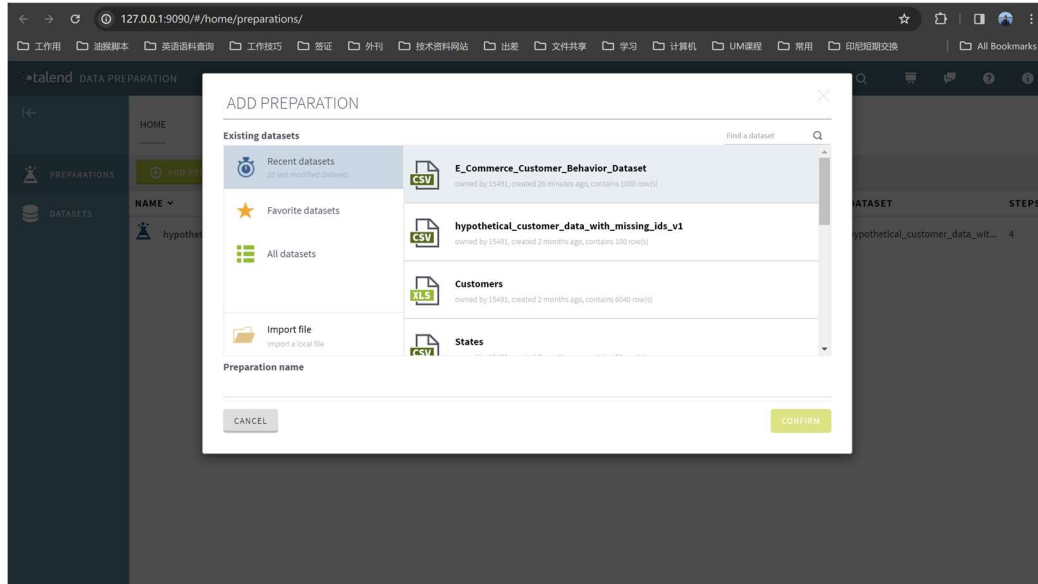


Talend Data Preparation: Data Pre-Processing

Talend Data Preparation:



The screenshot shows the main interface of the Talend Data Preparation tool. The left sidebar has 'PREPARATIONS' and 'DATASETS' sections. The 'DATASETS' section is active, showing a list of datasets. The table below represents the data shown in the interface:

NAME	AUTHOR	CREATED	MODIFIED	DATASET	STEPS
E_Commerce_Customer_Behavior_Dataset Preparation	15491	a few seconds ago	a few seconds ago	E_Commerce_Customer_Beh...	0
hypothetical_customer_data_with_missing_ids_v1 Preparation	15491	2 months ago	2 months ago	hypothetical_customer_data...	4

Find patterns and visualization for all the columns/variables:

We can observe the white gap for each variables represents the missing values.

talend DATA PREPARATION

1E_Commerce_Customer_Behavior_Dataset Preparation

1 Delete the rows with empty cell on column Gender

Filters

Add a filter ...

	CustomerID	Churn	Tenure	PreferredLoginID	Age	Gender
	integer	integer	integer	text	decimal	gender
1	50001	0	1	2 Tablet	50.0	Male
2	50002	1	1	Computer	50.0	Other
3	50003	0	1	Mobile	54.0	Male
4	50004	0	1	Computer	52.0	Male
5	50005	0	5	Tablet	67.0	Male
6	50006	0	4	Computer	63.0	Other
7	50007	0	1	Tablet	27.0	Other
8	50008	0	5	Mobile	56.0	Female
9	50009	0	1	Computer	28.0	Female
10	50010	0	5	Computer	53.0	Other
11	50011	0	5	Mobile	41.0	Other
12	50012	0	2	Mobile	27.0	Other
13	50013	0	4	Tablet	39.0	Male
14	50014	0	4	Computer	37.0	Female
15	50015	0	5	Tablet	29.0	Male
16	50016	0	2	Mobile	61.0	Female
17	50017	0	5	Mobile	45.0	Other
18	50018	0	5	Mobile	40.0	Male

CustomerID

COLUMN ROW

Find a function ...

SUGGESTIONS

Compare numbers...


Add, multiply, subtract or divide...

BOOLEAN

Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT *



Filters

Add a filter ...

	Age	Gender	Location	MembershipLevel	TotalPurchases	TotalSpent
	decimal	gender	text	city	decimal	decimal
1	58.0	Male	Suburban	Platinum	49.0	18237.707351897
2	55.0	Other	Rural	Platinum	20.0	633.75784591921
3	54.0	Male	Suburban	Bronze	87.0	6106.8631806448
4	52.0	Male	Rural	Platinum	50.0	4759.407207638
5	67.0	Male	Rural	Silver	16.0	7425.915140907
6	63.0	Other	Urban	Gold	59.0	26165.387854226
7	27.0	Other	Rural	Silver	15.0	2508.606097404
8	56.0	Female	Rural	Bronze	12.0	4461.209360291
9	28.0	Female	Urban	Bronze	9.0	4465.59358993
10	53.0	Other	Rural	Bronze	40.0	18588.272306745
11	41.0	Other	Rural	Platinum	24.0	4312.245401668
12	27.0	Other	Suburban	Silver	87.0	27849.129297995
13	39.0	Male	Suburban	Gold	19.0	8552.431767280
14	37.0	Female	Rural	Bronze	17.0	5453.107616315
15	29.0	Male	Suburban	Bronze	96.0	42098.59480281
16	32.0	Female	Suburban	Silver	27.0	5169.539003545
17	45.0	Other	Suburban	Bronze	54.0	13863.129821054
18	40.0	Male	Urban	Silver	100.0	22410.543001986

Age

COLUMN ROW

Find a function ...

SUGGESTIONS

Delete the rows with empty cell

Fill empty cells with text...

Compare numbers...

Add, multiply, subtract or divide...

Round value using halfup mode...

CHART VALUE PATTERN ADVANCED

Count: **995** Min: **18**

Distinct: **54** Max: **70**

Duplicate: **941** Mean: **44.37**

Valid: **990** Variance: **228.36**

Empty: **5** Median: **44**

Invalid: **0** Lower quantile: **32**

Upper quantile: **57**

Filters

Add a filter ...

	Location	MembershipLevel	TotalPurchases	TotalSpent	FavoriteCategory	LastPurchase
	gender	city	decimal	decimal	text	text
1	Suburban	Platinum	49.0	18237.707351897774	Home Goods	2023-02-
2	Rural	Platinum	20.0	633.7578459192816	Electronics	2023-07-
3	Suburban	Bronze	87.0	6106.8631806448275	Home Goods	2023-10-
4	Rural	Platinum	50.0	4759.40720763638	Home Goods	2023-04-
5	Rural	Silver	16.0	7425.915140907109	Home Goods	2023-11-
6	Urban	Gold	59.0	26165.38785422678	Home Goods	2023-05-
7	Rural	Silver	15.0	2508.606097404642	Home Goods	2023-02-
8	Rural	Bronze	12.0	4461.209360291154	Clothing	2023-08-
9	Urban	Bronze	9.0	4465.59358993141	Electronics	2023-09-
10	Rural	Bronze	40.0	18588.272306745566	Home Goods	2023-04-
11	Rural	Platinum	24.0	4312.245401668005	Home Goods	2023-04-
12	Suburban	Silver	87.0	27849.129297995158	Clothing	2023-12-
13	Suburban	Gold	19.0	8552.431767280506	Home Goods	2023-10-
14	Rural	Bronze	17.0	5453.107616315044	Electronics	2023-02-
15	Suburban	Bronze	96.0	42098.59480281735	Electronics	2023-12-
16	Suburban	Silver	27.0	5169.539003545292	Home Goods	2023-04-
17	Suburban	Bronze	54.0	13863.129821054747	Home Goods	2023-05-
18	Urban	Silver	100.0	22410.543001986385	Home Goods	2023-09-

MembershipLevel

COLUMN ROW

Find a function ...

SUGGESTIONS

Delete the rows with empty cell

Fill empty cells with text...

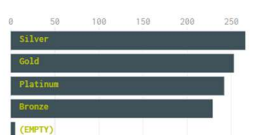
Change to upper case

Replace the cells that match...

Change to lower case

CHART VALUE PATTERN ADVANCED

ROW COUNT *



talend DATA PREPARATION

1E_Commerce_Customer_Behavior_Dataset Preparation

1 Delete the rows with empty cell on column Gender

Filters

Add a filter ...

996/995

	CustomerID	Churn	tenure	PreferredLogin...	Age	Gender
	text	boolean	integer	text	integer	text
1	50001	0	1	2 Tablet	58.0	Male
2	50002	1	1	Computer	55.0	Other
3	50003	0	1	Mobile	54.0	Male
4	50004	0	1	Computer	52.0	Male
5	50005	0	5	Tablet	67.0	Male
6	50006	0	4	Computer	63.0	Other
7	50007	0	1	Tablet	27.0	Other
8	50008	0	5	Mobile	56.0	Female
9	50009	0	1	Computer	28.0	Female
10	50010	0	5	Computer	53.0	Other
11	50011	0	5	Mobile	41.0	Other
12	50012	0	2	Mobile	27.0	Other
13	50013	0	4	Tablet	39.0	Male
14	50014	0	4	Computer	37.0	Female
15	50015	0	5	Tablet	29.0	Male
16	50016	0	2	Mobile	33.0	Female
17	50017	0	5	Mobile	45.0	Other
18	50018	0	5	Mobile	40.0	Male

Churn

COLUMN ROW

Find a function ...

SUGGESTIONS

Compare numbers...

Add, multiply, subtract or divide...

BOOLEAN

Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT

Occurrences

Filters

Add a filter ...

995/995

	Location	MembershipLevel	TotalPurchases	TotalSpent	FavoriteCategory	LastPurchase
	text	text	integer	decimal	text	text
1	Suburban	Platinum	49.0	10237.707351897774	Home Goods	2023-02-
2	Rural	Platinum	20.0	633.7578459192816	Electronics	2023-07-
3	Suburban	Bronze	87.0	6106.8631806448275	Home Goods	2023-10-
4	Rural	Platinum	50.0	4759.40720763638	Home Goods	2023-04-
5	Rural	Silver	16.0	7425.915140907109	Home Goods	2023-11-
6	Urban	Gold	59.0	26165.38785422678	Home Goods	2023-05-
7	Rural	Silver	15.0	2508.606097404642	Home Goods	2023-02-
8	Rural	Bronze	12.0	4461.209360291154	Clothing	2023-08-
9	Urban	Bronze	9.0	4465.59358993141	Electronics	2023-09-
10	Rural	Bronze	40.0	18588.272306745566	Home Goods	2023-04-
11	Rural	Platinum	24.0	4312.245401668005	Home Goods	2023-04-
12	Suburban	Silver	87.0	27849.129297995158	Clothing	2023-12-
13	Suburban	Gold	19.0	8552.431767280506	Home Goods	2023-10-
14	Rural	Bronze	17.0	5453.107616315044	Electronics	2023-02-
15	Suburban	Bronze	96.0	42098.59480281735	Electronics	2023-12-
16	Suburban	Silver	27.0	5169.539803545292	Home Goods	2023-04-
17	Suburban	Bronze	54.0	13863.129821054747	Home Goods	2023-05-
18	Urban	Silver	100.0	22410.543001986385	Home Goods	2023-09-

Location

COLUMN ROW

Find a function ...

SUGGESTIONS

Change to upper case

Replace the cells that match...

Change to lower case

BOOLEAN

Negate value

CHART VALUE PATTERN ADVANCED

ROW COUNT

Filters

Add a filter ...

995/995

	TotalPurchases	TotalSpent	FavoriteCategory	LastPurchaseDate	Occupation	FrequencyOfWe...
	decimal	decimal	text	date	text	integer
1	49.0	10237.707351897774	Home Goods	2023-02-12	Student	7
2	20.0	633.7578459192816	Electronics	2023-07-11	Professional	14
3	87.0	6106.8631806448275	Home Goods	2023-10-25	Self-Employed	8
4	50.0	4759.40720763638	Home Goods	2023-04-19	Student	26
5	16.0	7425.915140907109	Home Goods	2023-11-09	Professional	25
6	59.0	26165.38785422678	Home Goods	2023-05-08	Student	12
7	15.0	2508.606097404642	Home Goods	2023-02-27	Self-Employed	3
8	12.0	4461.209360291154	Clothing	2023-08-13	Self-Employed	12
9	9.0	4465.59358993141	Electronics	2023-09-08	Self-Employed	3
10	40.0	18588.272306745566	Home Goods	2023-04-06	Retired	25
11	24.0	4312.245401668005	Home Goods	2023-04-23	Retired	27
12	87.0	27849.129297995158	Clothing	2023-12-17	Retired	30
13	19.0	8552.431767280506	Home Goods	2023-10-12	Professional	26
14	17.0	5453.107616315044	Electronics	2023-02-13	Self-Employed	20
15	96.0	42098.59480281735	Electronics	2023-12-30	Student	13
16	27.0	5169.539803545292	Home Goods	2023-04-25	Self-Employed	18
17	54.0	13863.129821054747	Home Goods	2023-05-05	Retired	17
18	100.0	22410.543001986385	Home Goods	2023-09-01	Retired	14

Occupation

COLUMN ROW

Find a function ...

SUGGESTIONS

Change to upper case

Replace the cells that match...

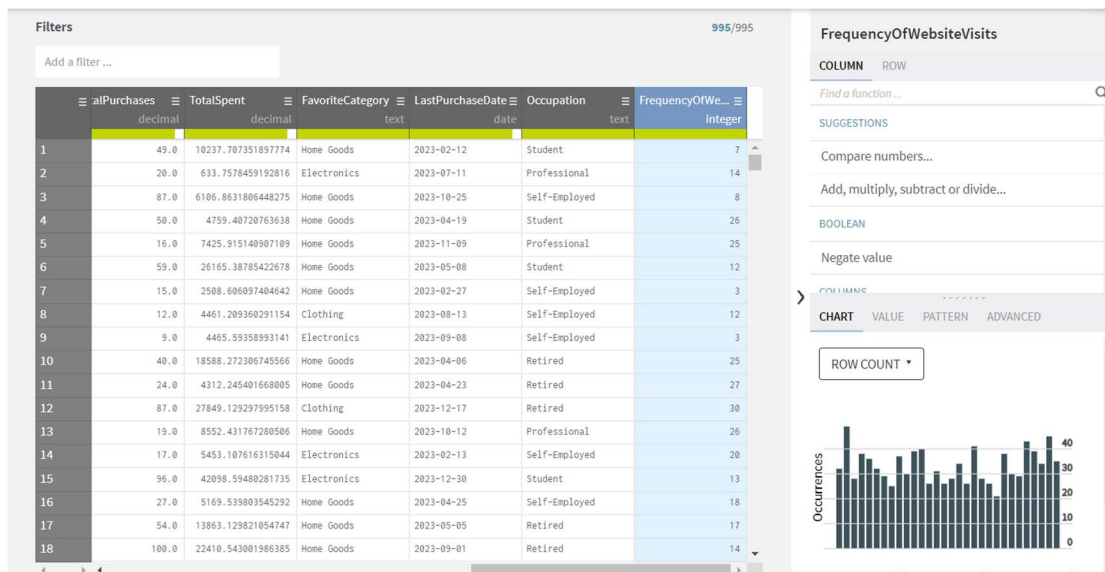
Change to title case

Change to lower case

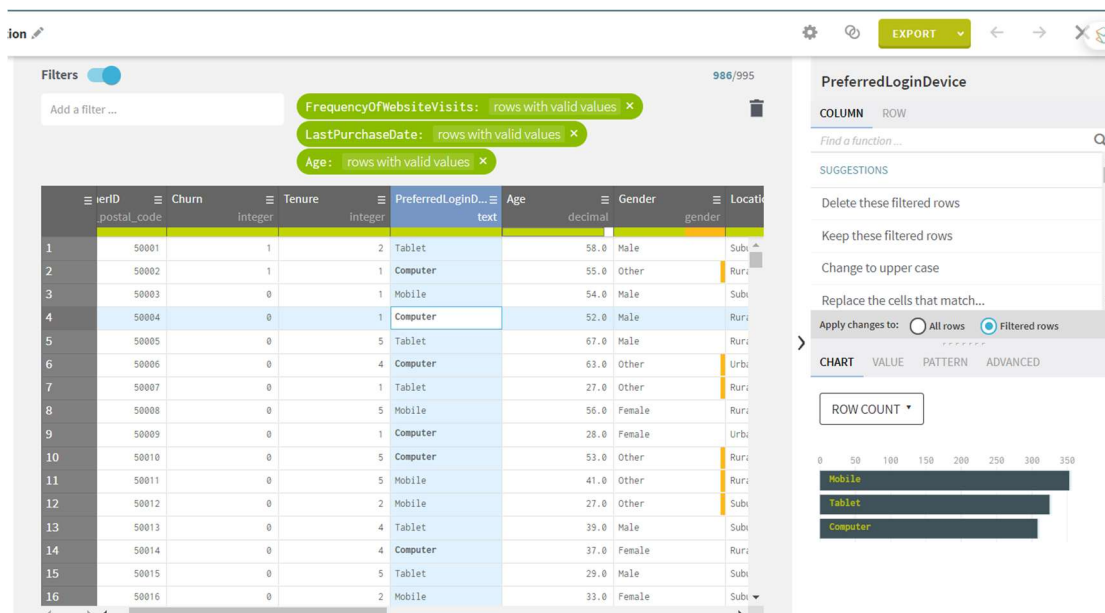
BOOLEAN

CHART VALUE PATTERN ADVANCED

ROW COUNT

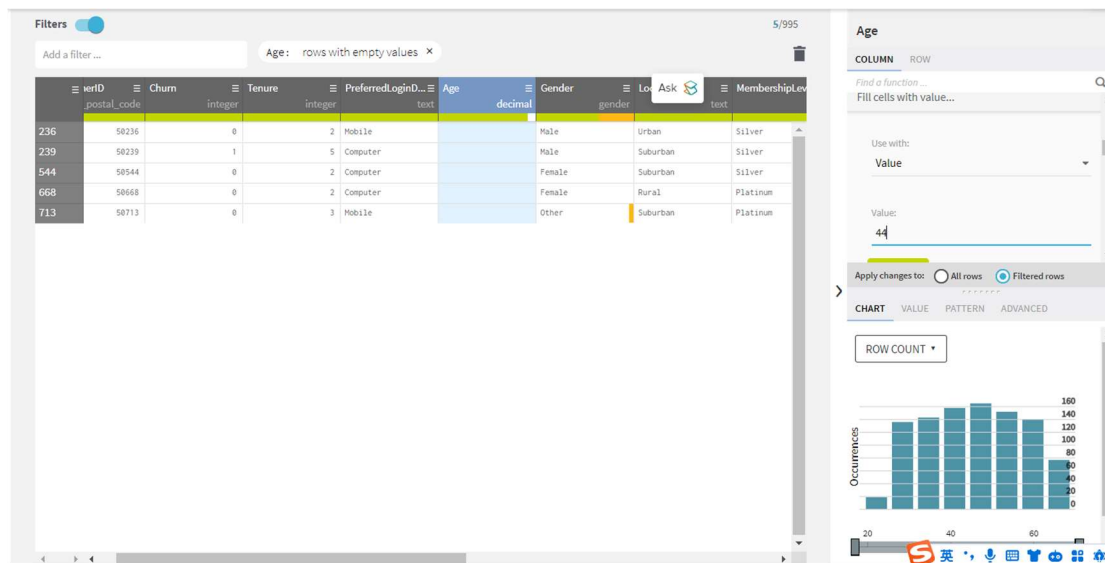


We can select the rows with valid values:

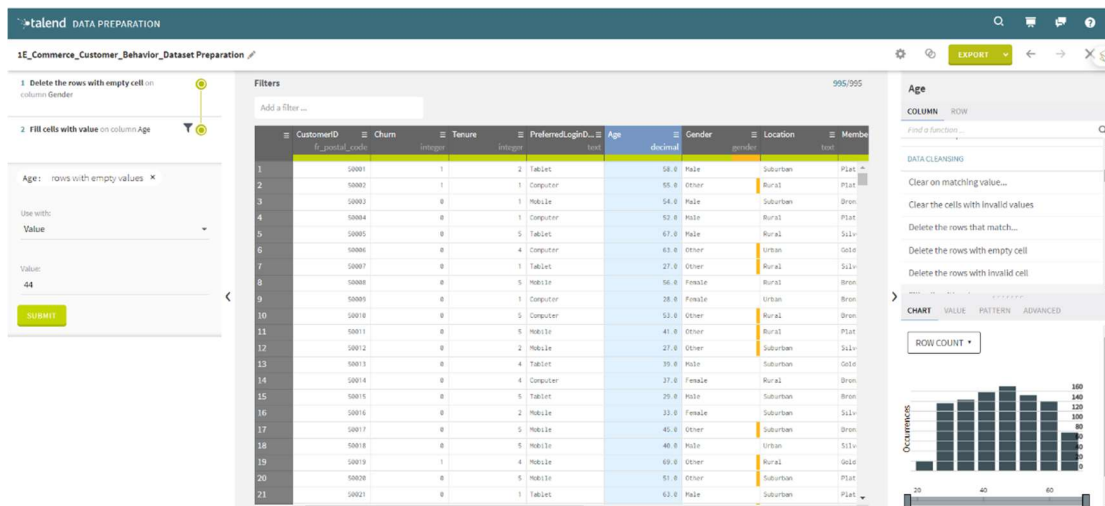


Sample of filling the missing values, including age (numeric), MembershipLevel (string), and other variables (delete unimportant rows):

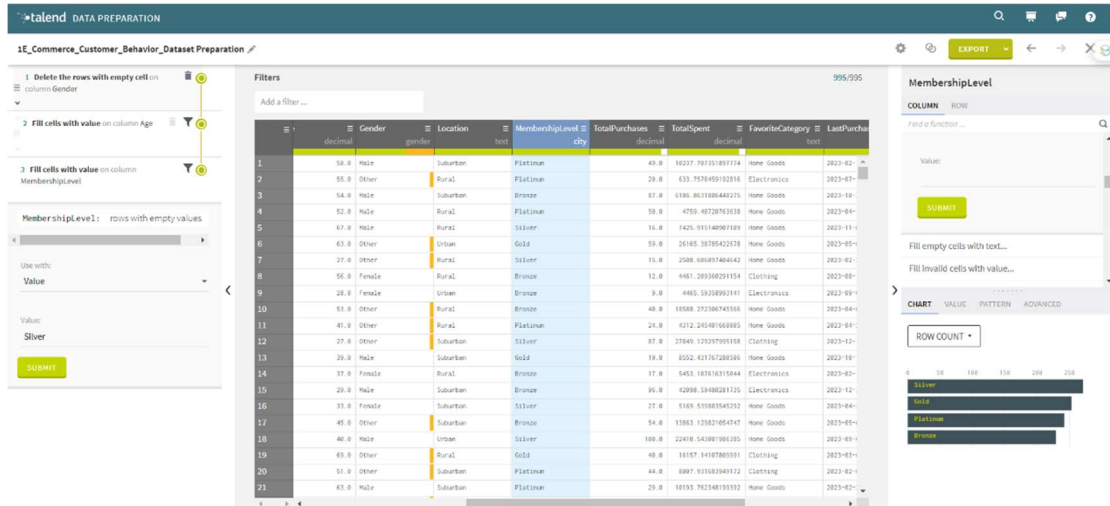
First select the rows with empty cells.



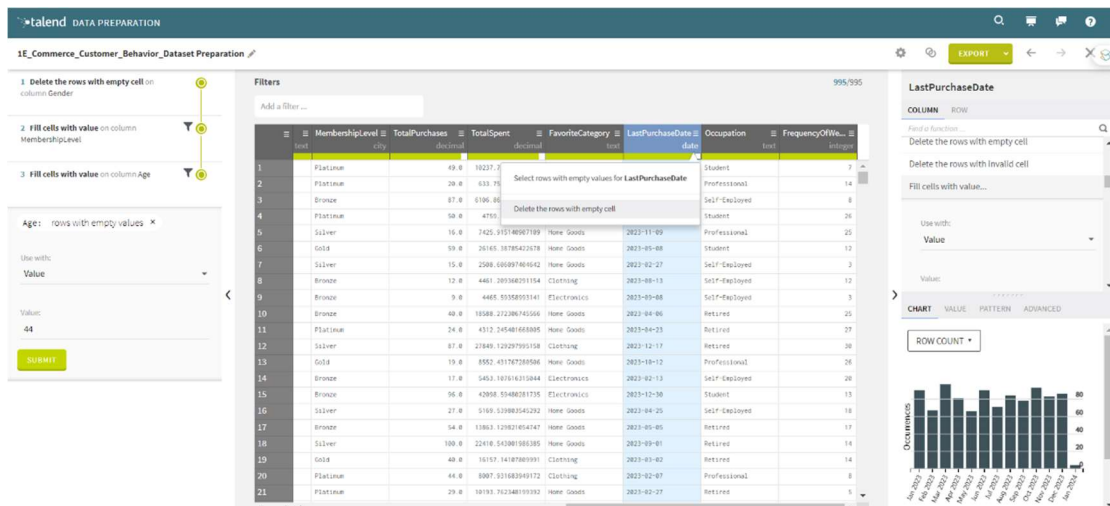
And then fill the missing values with median/mean number.



For the variables with missing value type of string/text, use the same method, impute the most frequent value based on the distribution.



Other variables, delete unimportant rows with empty values.



And then we finish the data pre-processing part.

