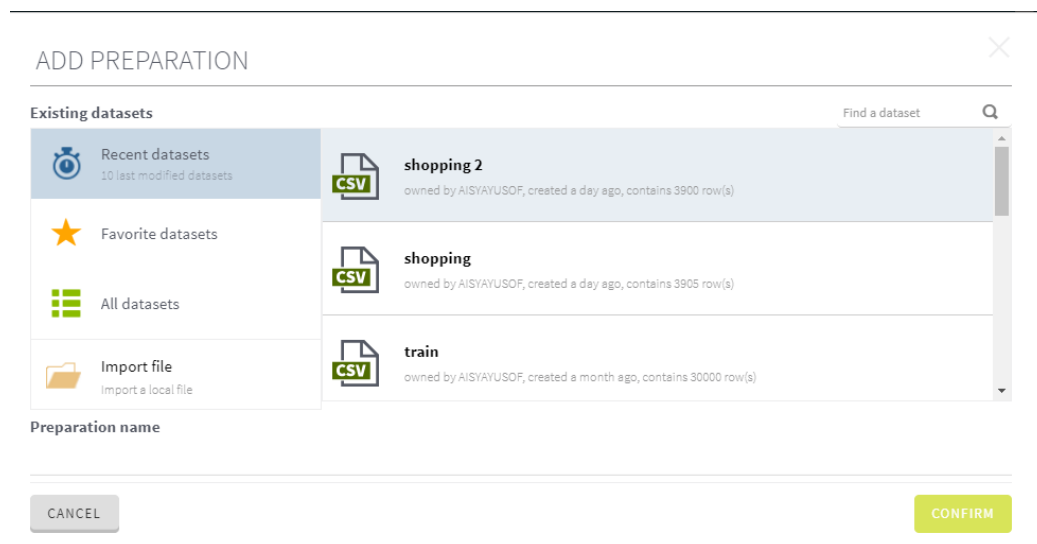
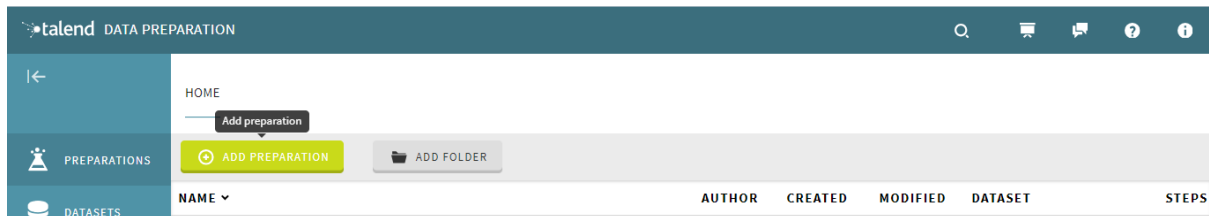


WQD7005 AA1

Nur Aisyah Yusof (22072845)

To check the data quality and data visualization, Talend Data Preparation (TDP) is used :-

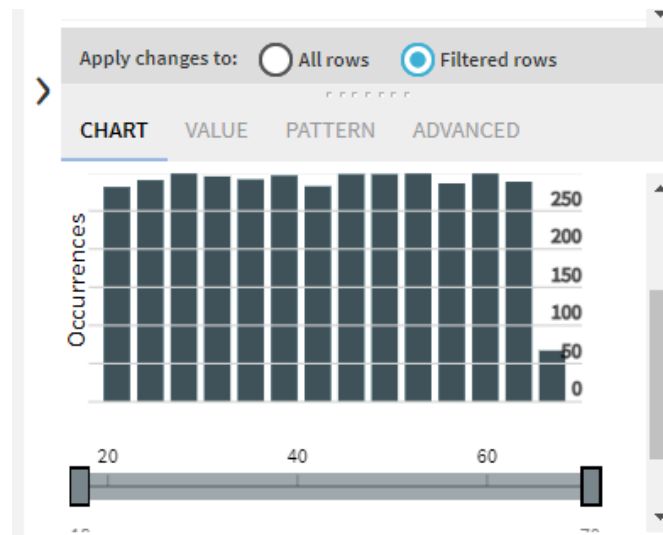
- 1) Import the dataset into TDP.



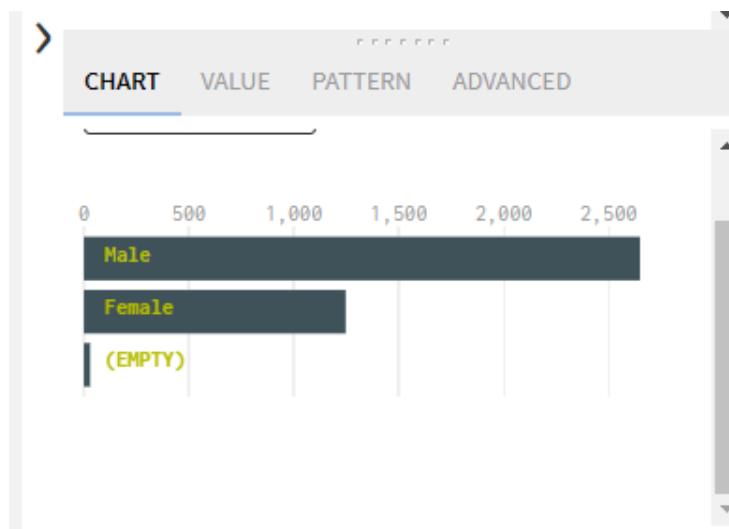
- 2) After importing the dataset into TDP, the dataset variables were checked whether have null value, invalid value, basic statistic, and the visualization chart etc. Here are a few examples of the variable's details.

CHART		VALUE		PATTERN		ADVANCED	
Count: 3900		Min: 1					
Distinct: 3900		Max: 3900					
Duplicate: 0		Mean: 1950.5					
Valid: 3900		Variance: 1267825					
Empty: 0		Median: 1950.5					
Invalid: 0		Lower quantile: 975.25					
		Upper quantile: 2925.75					

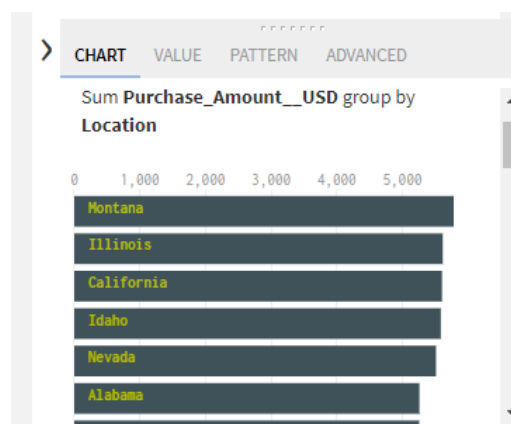
Customer ID Column show no duplicate and null value.



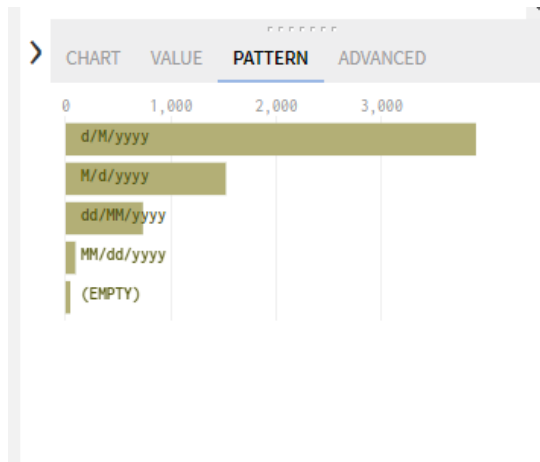
Age Column show the range age between 18 to 70.



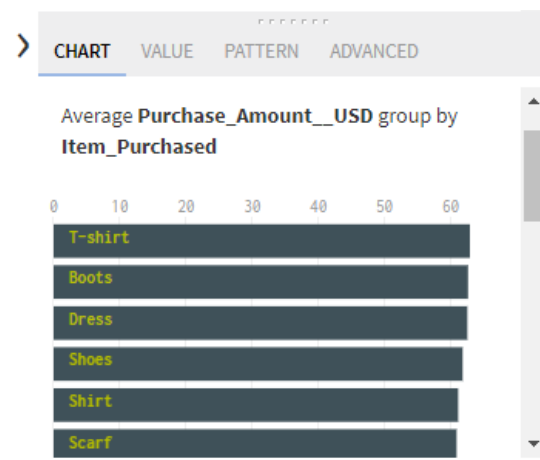
Gender Column show number of male are greater than female.



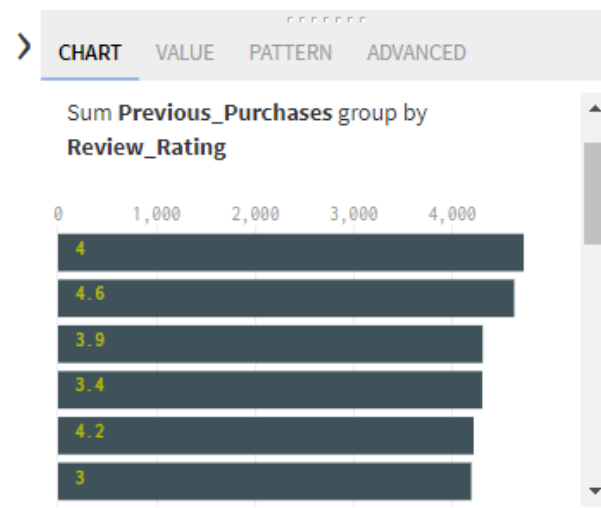
Montana has the highest purchase amount from the other location.



The last purchase date column show there is different in date format.



T-shirt is the average purchase amount in item purchase.



Most customers give review rating 4 from their previous purchase.

gender	Location us_state	Last_Purchase_... date	Item_Purchased text	Category
	Kansas	5/2/2023	Jacket	Outerwear
	Colorado	11/2/2023	Pants	Clothing
	North Dakota	6/6/2022	T-shirt	Clothing
	Massachusetts	25/9/2022	Blouse	Clothing

Null value can also be detected at the green bar. The white bar shows there is null value in the column.

- 3) From the previous steps, I have decided to remove the rows that contain null value as the missing value were classified as MCAR and it might affect the quality of data later.

shopping 2 PREPARATION

1 Delete the rows with empty cell on column Size

2 Delete the rows with empty cell on column Item_Purchased

Filters: 10/3898

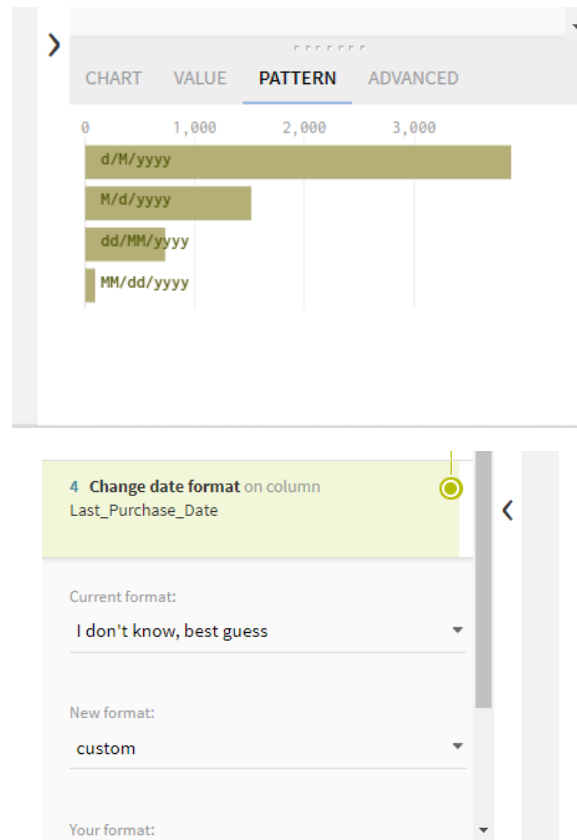
Add a filter ... Item_Purchased: rows with empty values

	Last_Purchase_D...	Item_Purchased	Category	Purch
40	13/5/2022		Clothing	
217	2/11/2023		Clothing	
223	29/6/2022		Accessories	
1083	21/8/2022		Accessories	
1989	2/9/2023		Clothing	
2627	12/11/2023		Outerwear	
3204	13/11/2022		Outerwear	
3422	7/12/2022		Footwear	
3785	1/6/2022		Clothing	
3869	19/10/2023		Clothing	

- 4) However, the review rating column was expected to be dependent on subscription status. Thus, the column will later be modified in SAS EM.

	VALUE	PATTERN	ADVANCED
Count:	3900		Min: 2.5
Distinct:	27		Max: 5
Duplicate:	3873		Mean: 3.75
Valid:	3886		Variance: 0.51
Empty:	14		Median: 3.8
Invalid:	0		Lower quantile: 3.1
			Upper quantile: 4.4

- 5) As for the last purchase date column, there are different format pattern shown, will make it hard for the analysis later. Thus, the format pattern was changed into one uniform pattern which is dd-MM-yyyy.



- 6) Lastly, the clean data was exported into csv files for the next step.

