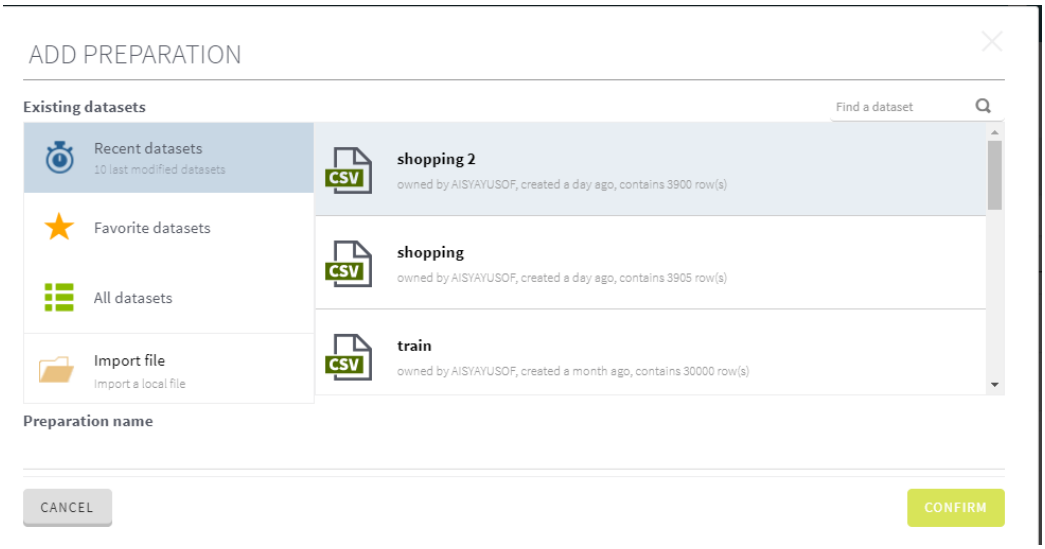
WQD7005 AA1

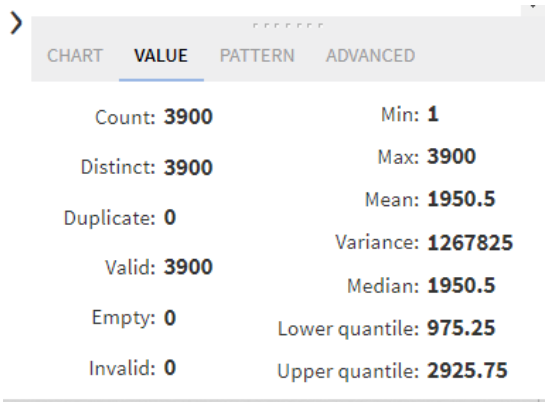
Nur Aisyah Yusof (22072845)

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

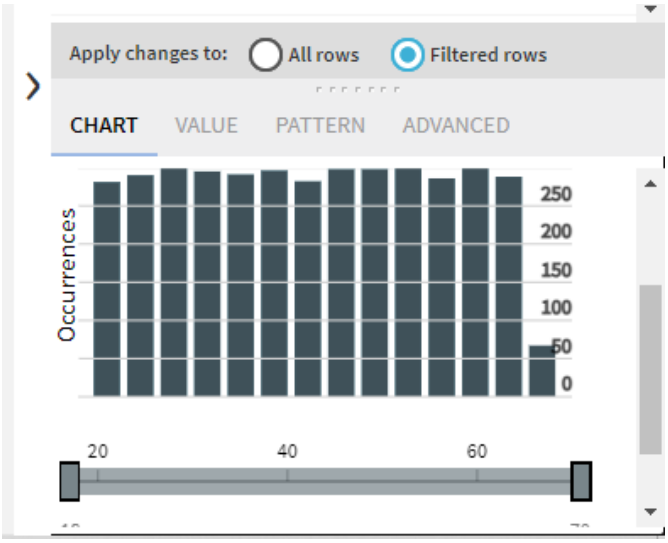
1) Import the dataset into TDP.



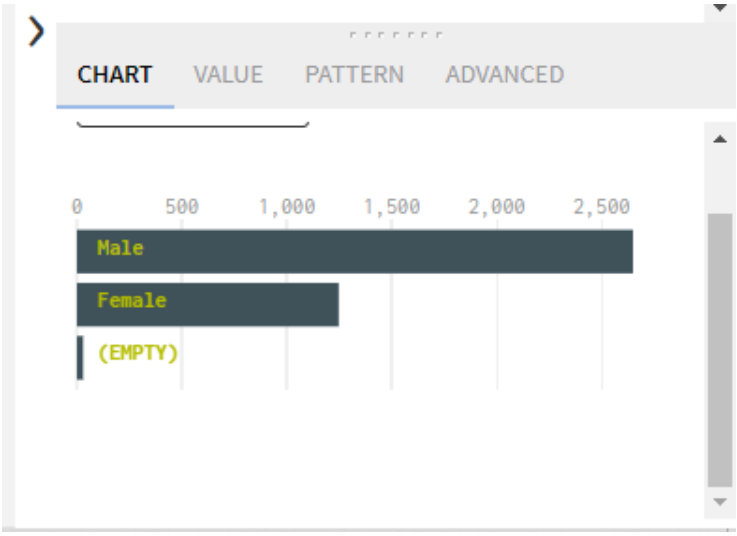
2) After importing the dataset into TDP, the dataset variables were checked whether have null value, invalid value, basic statistic, and the visualization chart etc. Here are a few examples of the variable's details.



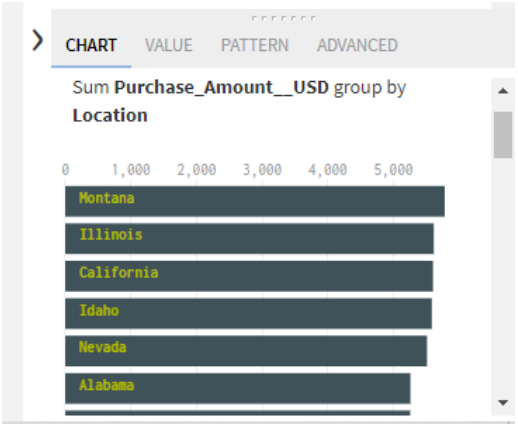
Customer ID Column show no duplicate and null value.



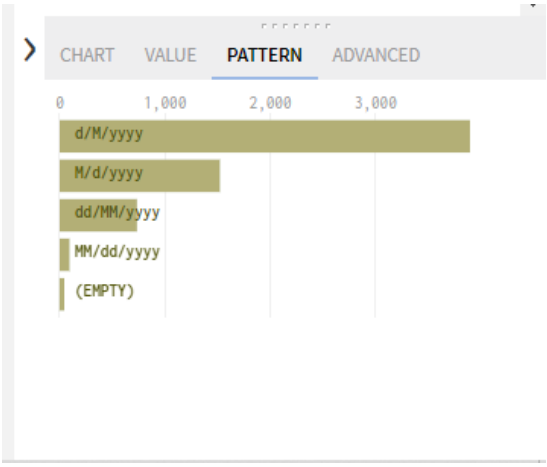
Age Column show the range age between 18 to 70.



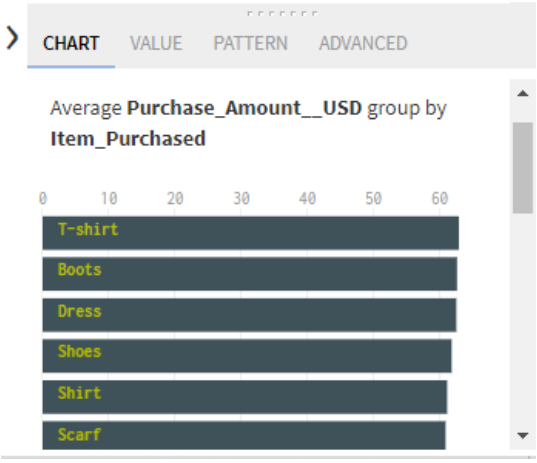
Gender Column show number of male are greater than female.



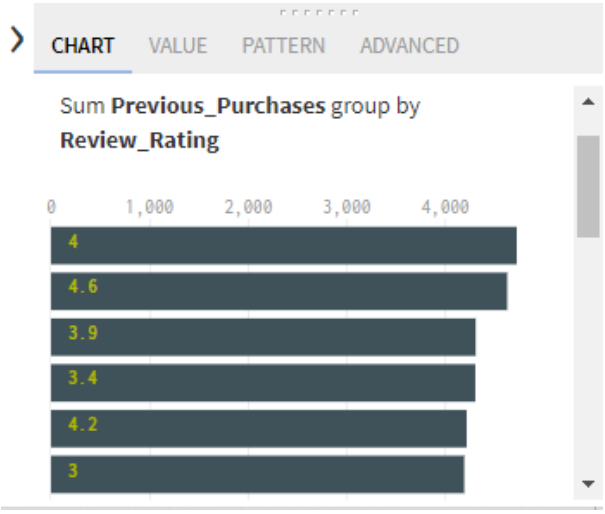
Montana has the highest purchase amount from the other location.



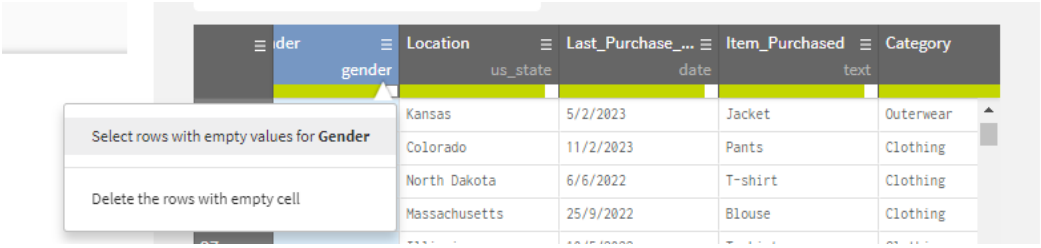
The last purchase date column show there is different in date format.



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

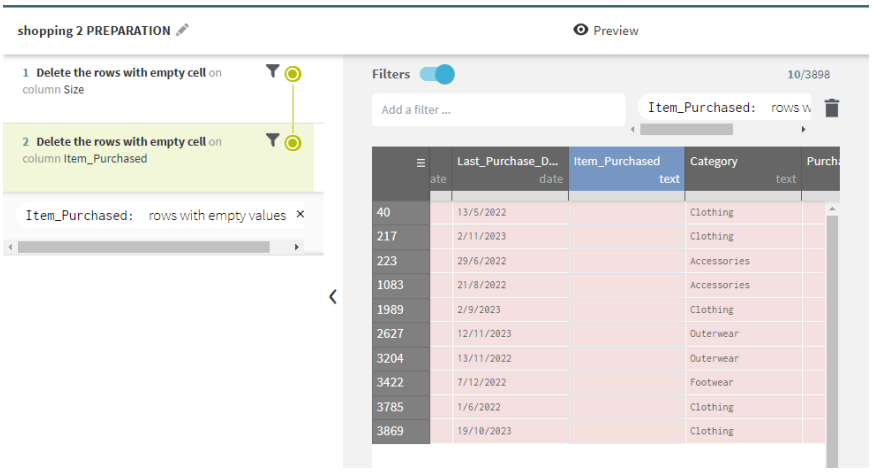


Most customers give review rating 4 from their previous purchase.

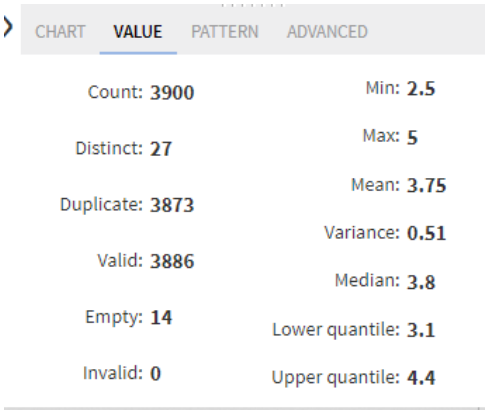


Null value can also be detected at the green bar. The white bar shows there is null value in the column.

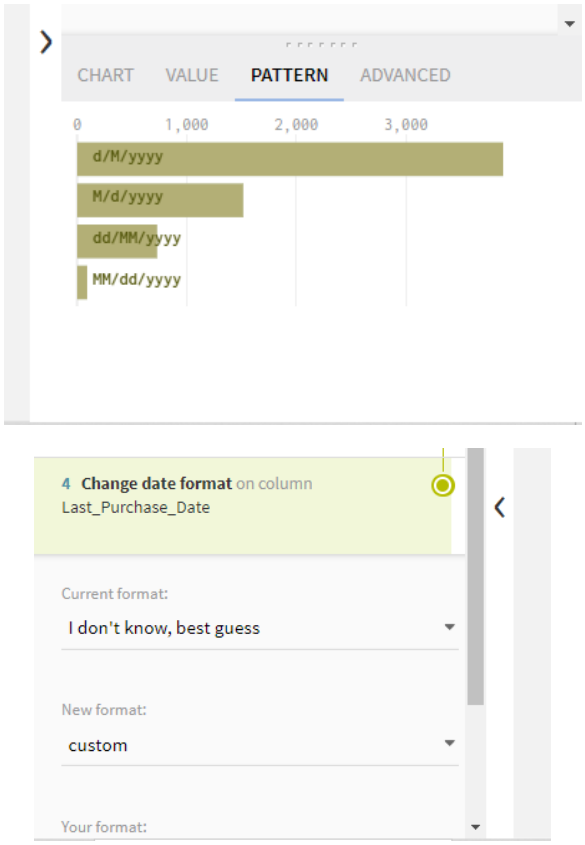
3) From the previous steps, I have decided to remove the rows that contain null value as the missing value were classified as MCAR and it might affect the quality of data later.



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



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.