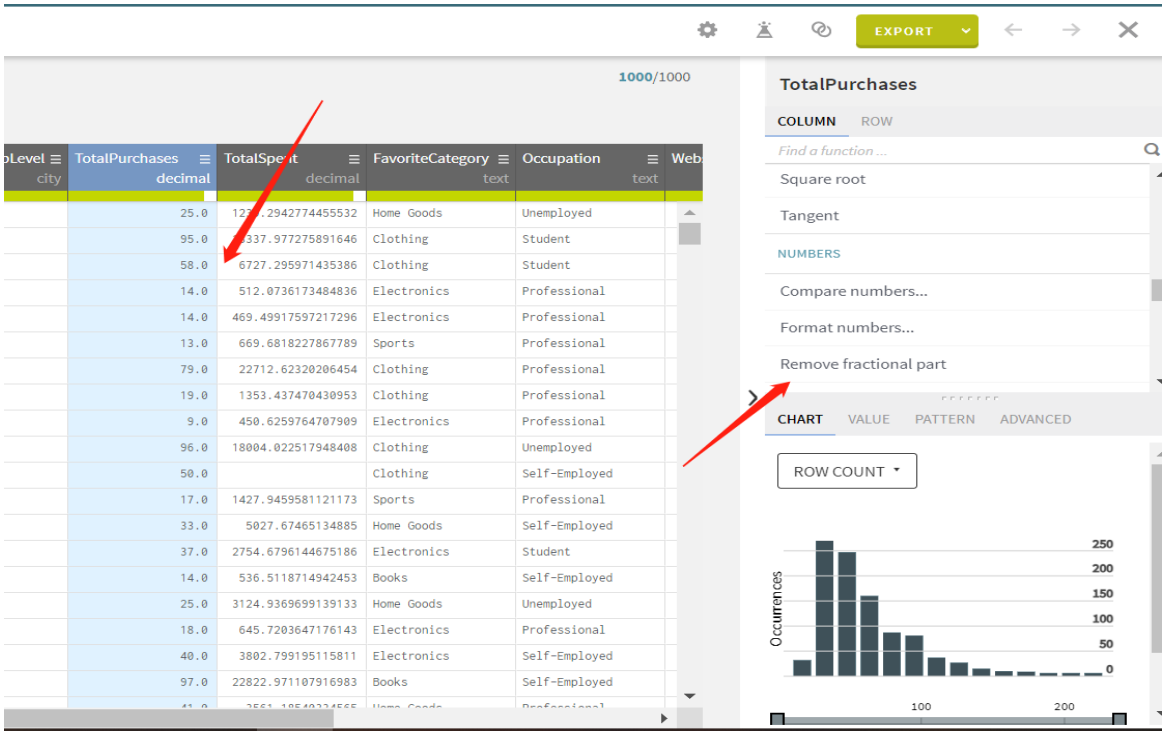
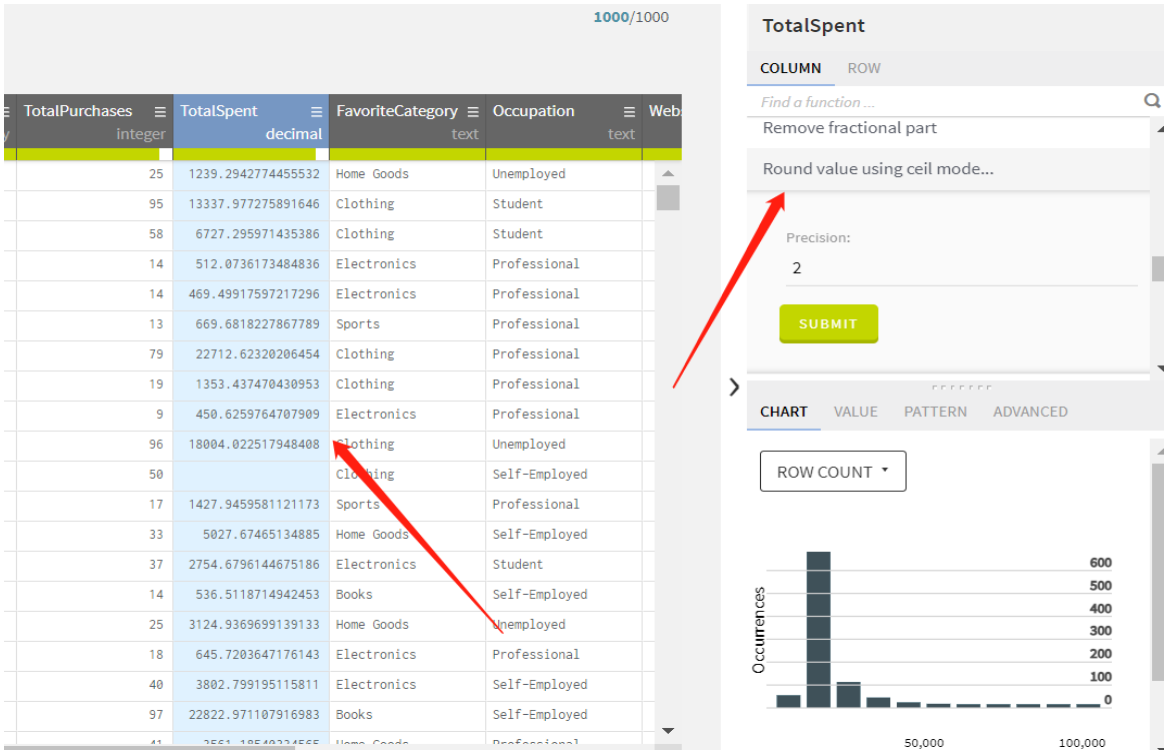
Data Preparation (DP)

After merging the distinct datasets, we need to conduct initial exploration and processing of the data.

- 2.1 It is not reasonable for TotalPurchases to have decimal points, so I removed the decimal points from TotalPurchases.

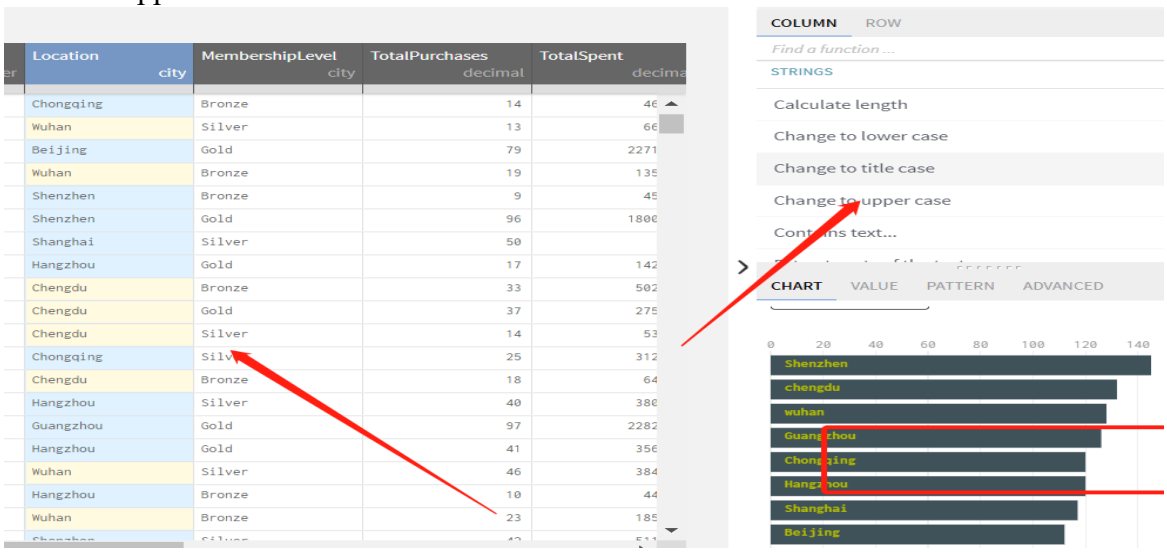


- 2.2 Many decimal places in TotalSpent are not reasonable, so I retained two decimal places for TotalSpent.

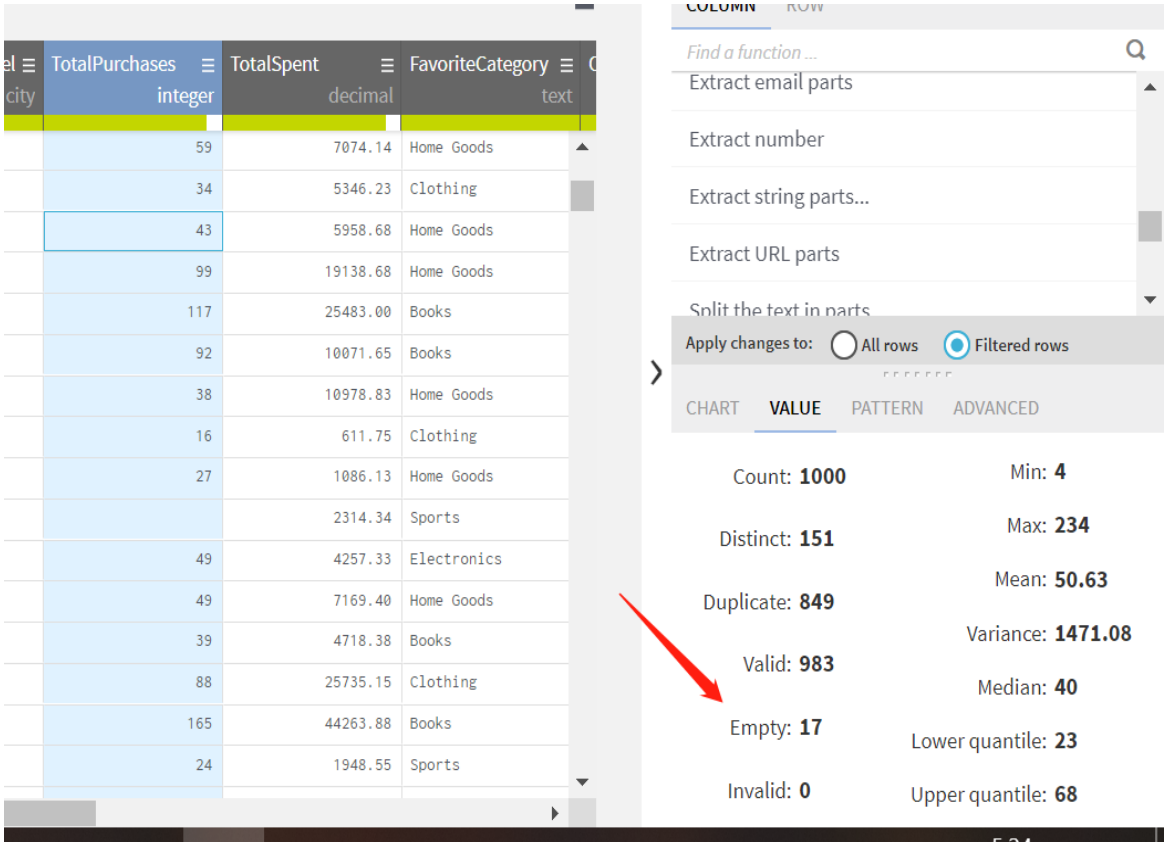


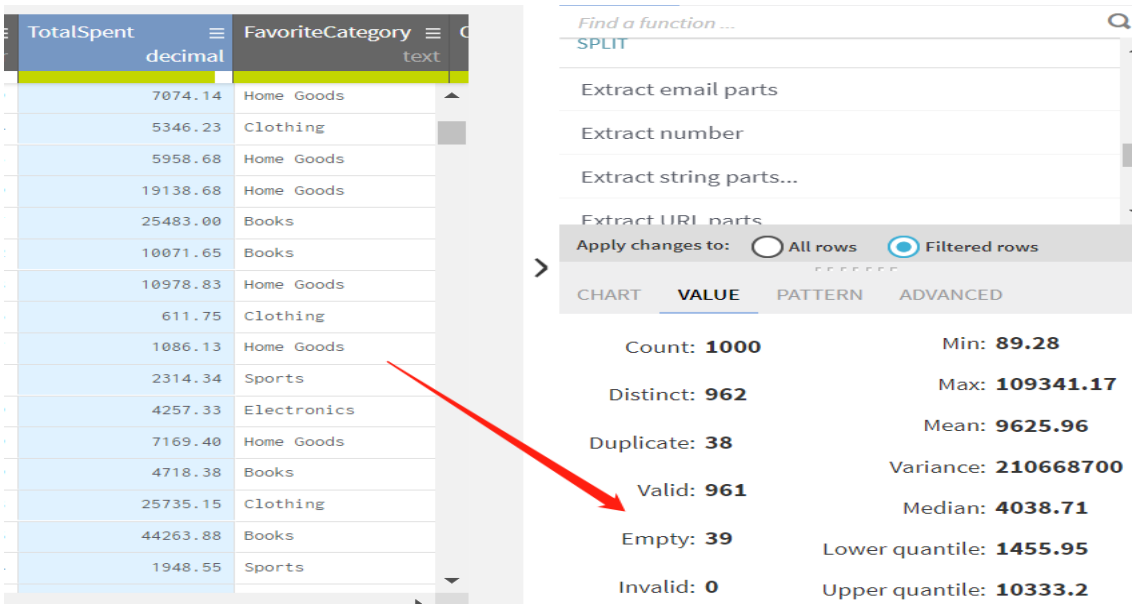
- 2.3 The capitalization of city names in the location data is inconsistent, with

some having uppercase initial letters and others having lowercase initial letters. This is not reasonable, so I have standardized all city names to have uppercase initial letters.

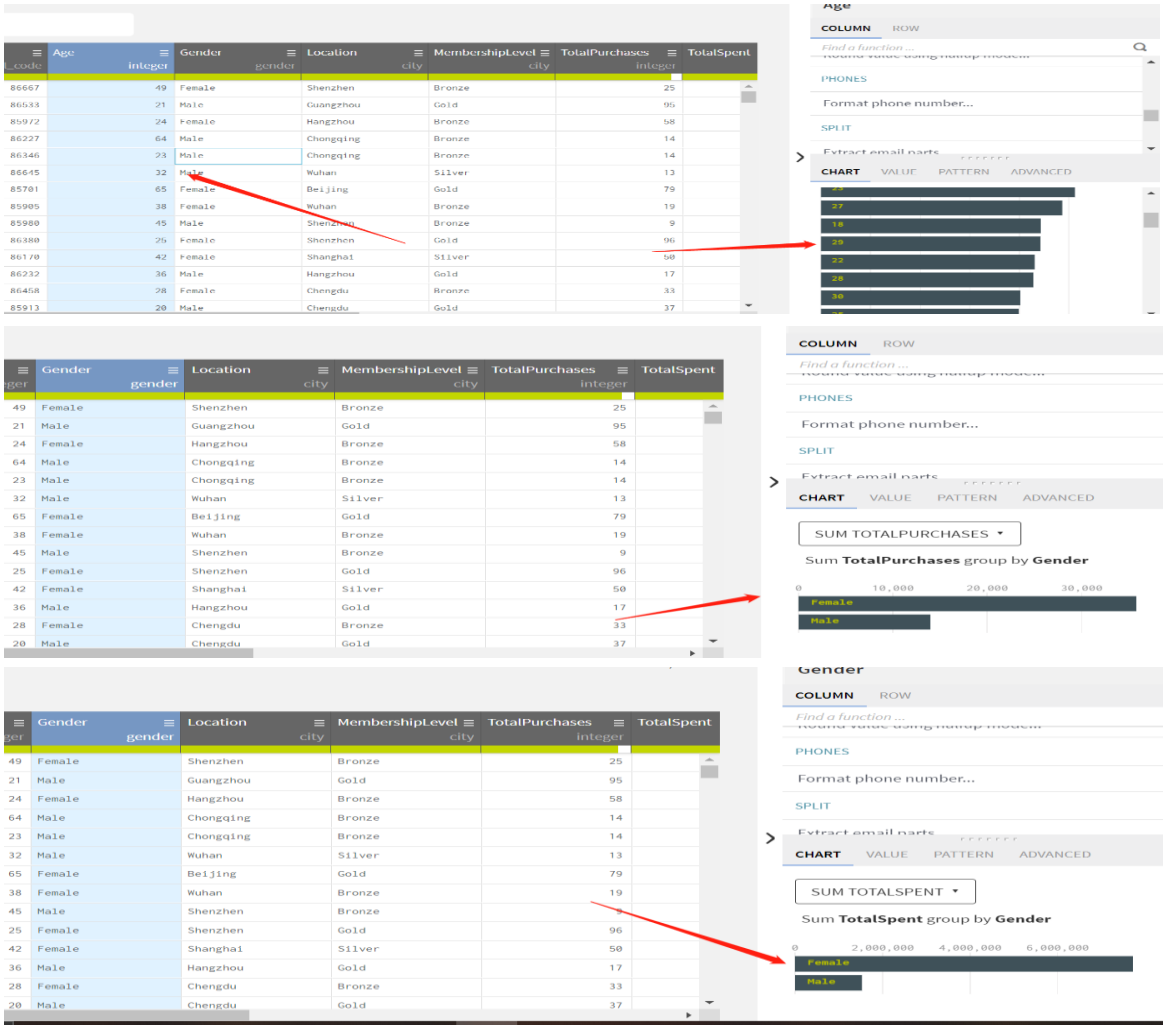


- 2.4 I observed that both TotalPurchases and TotalSpent have missing values, with 17 and 39 missing values respectively. I won't handle them at this point; instead, I will examine the overall structure of the data inside SAS EM before proceeding with any imputation or treatment.

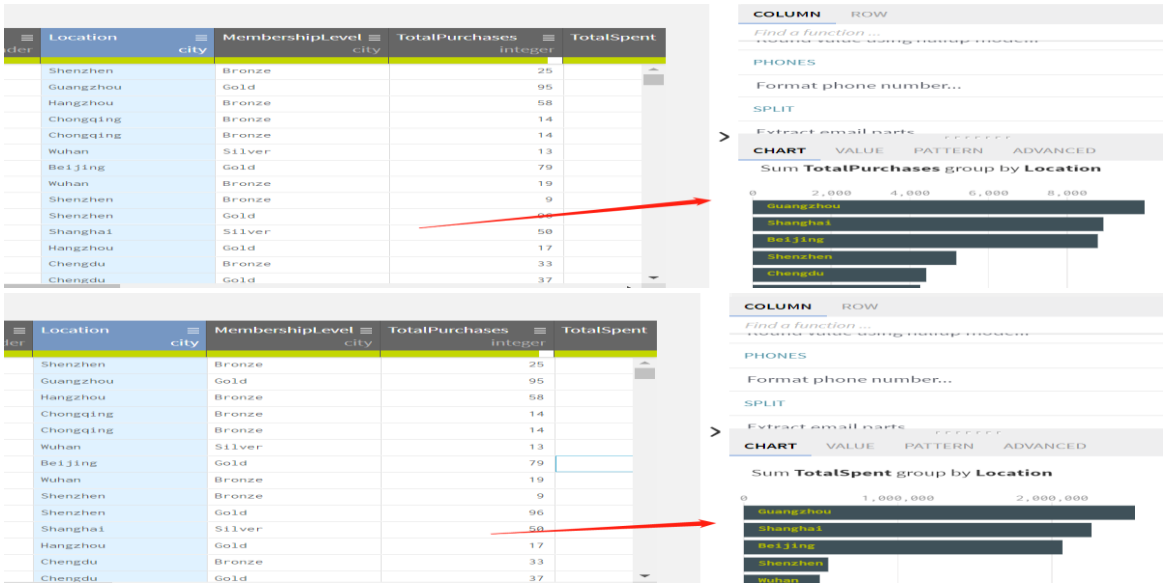




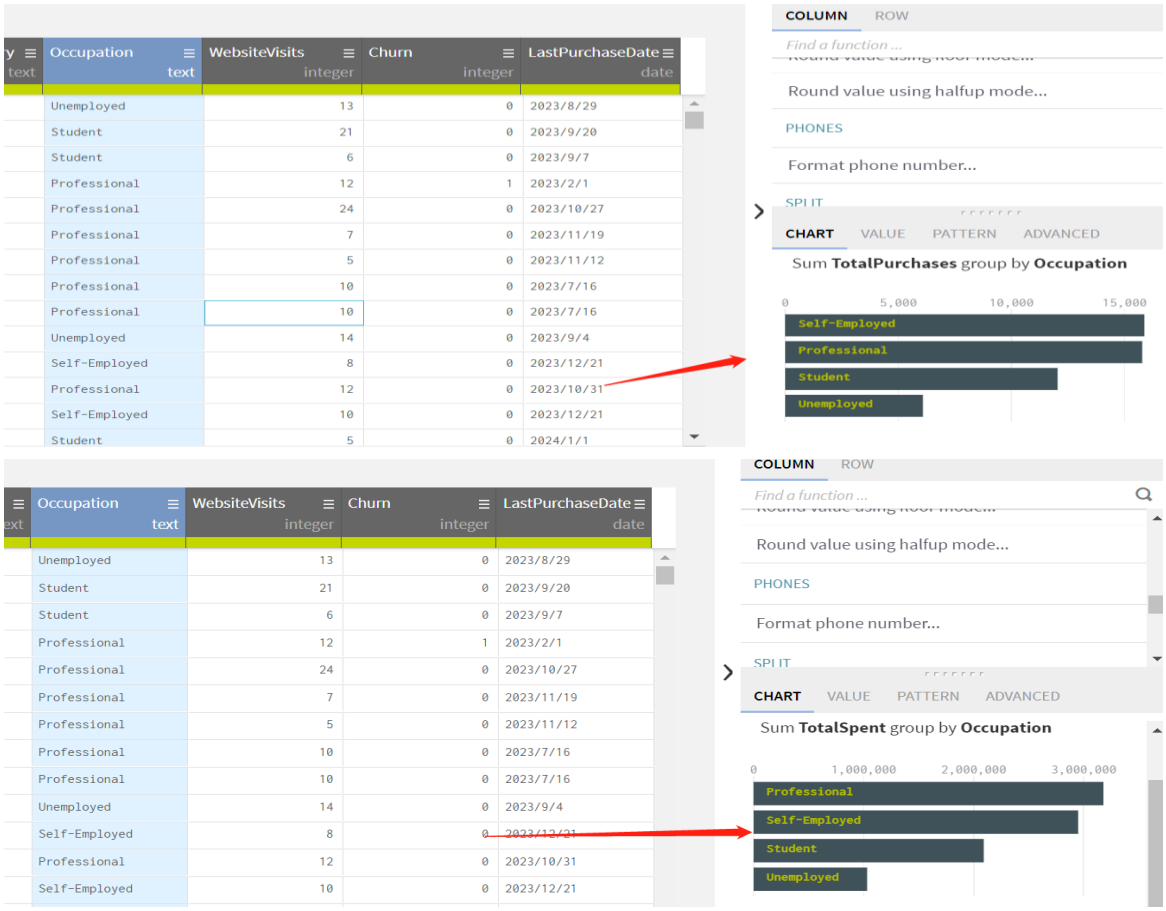
- 2.5 Observing this dataset, it appears that the majority of individuals are young, and the spending capacity of females is significantly higher than that of males. This aligns well with real-world scenarios.



- 2.6 It can be observed that major cities like Beijing, Guangzhou, and Shanghai exhibit significantly higher purchase frequencies and total expenditures compared to other cities.



- 2.7 It is noticeable that individuals with employment or student status tend to have significantly higher purchase frequencies and total expenditures compared to those who are unemployed.



- 2.8 After completing the conversion of the data format and cleaning the data, as well as the initial exploration of data correlations, we exported the data and continued to SAS EM to process the analysis