Babu Banarasi Das University

School of Computer Applications

Case Study

on

Retail Sales Data Preparation and Discount Pattern Analysis

SUBMITTED TO: Mr. Robin Tyagi

SUBMITTED BY:

Name: Daya Yadav - 1230258149

Name: Deepak Kumar - 1230258151

Name: Devansh Kumar Singh - 1230258159

Definition

The project aims to prepare, clean, and analyze retail store sales data to identify the key patterns behind discount allocation and improve business decision-making. Using IBM SPSS Modeler 18.6, the dataset undergoes systematic preprocessing, cleaning, transformation, and exploratory analysis to ensure high-quality data for predictive or descriptive modelling.

Outcome / Learning

- Learned complete data preparation workflow in IBM SPSS Modeler.
- Understood difference between real and integer storage types.
- Practiced data cleaning, transformation, and field derivation.
- Built logical conditions to detect discount eligibility.
- Generated visual and tabular insights about customer behaviour.

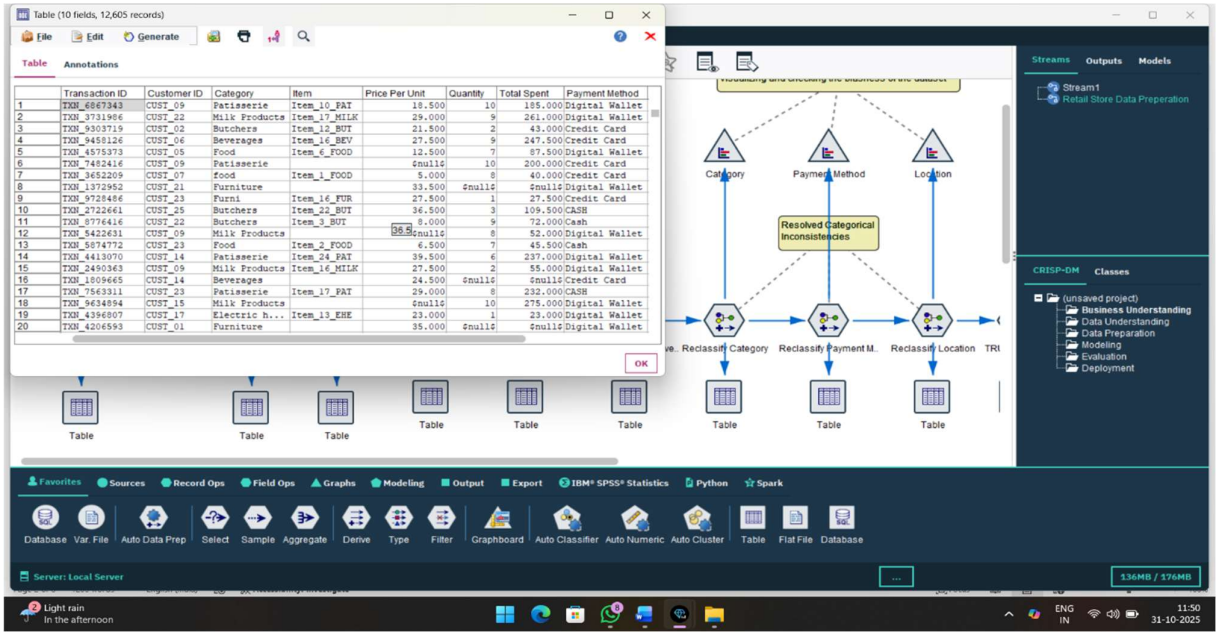
Required Tool: IBM SPSS Modeler

Working: Imported the dataset, corrected data types, cleaned duplicates and missing values, standardized categories, created new discount-related fields, and removed outliers. Verified each step using Table and Graph nodes in IBM SPSS Modeler.

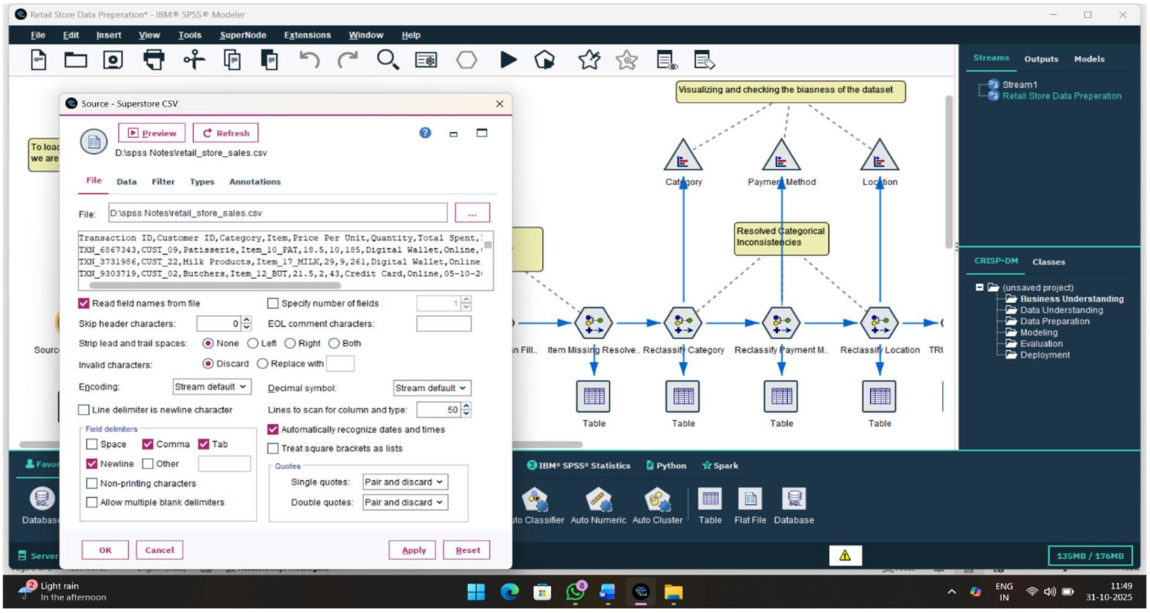
Step 1: Data Import

Purpose: To bring the dataset into SPSS Modeler and make it ready for analysis.

- **Role of Var. File Node:** Imports the retail store sales dataset and defines each field's metadata such as variable names, storage type, and field role (Input, Target, or None).



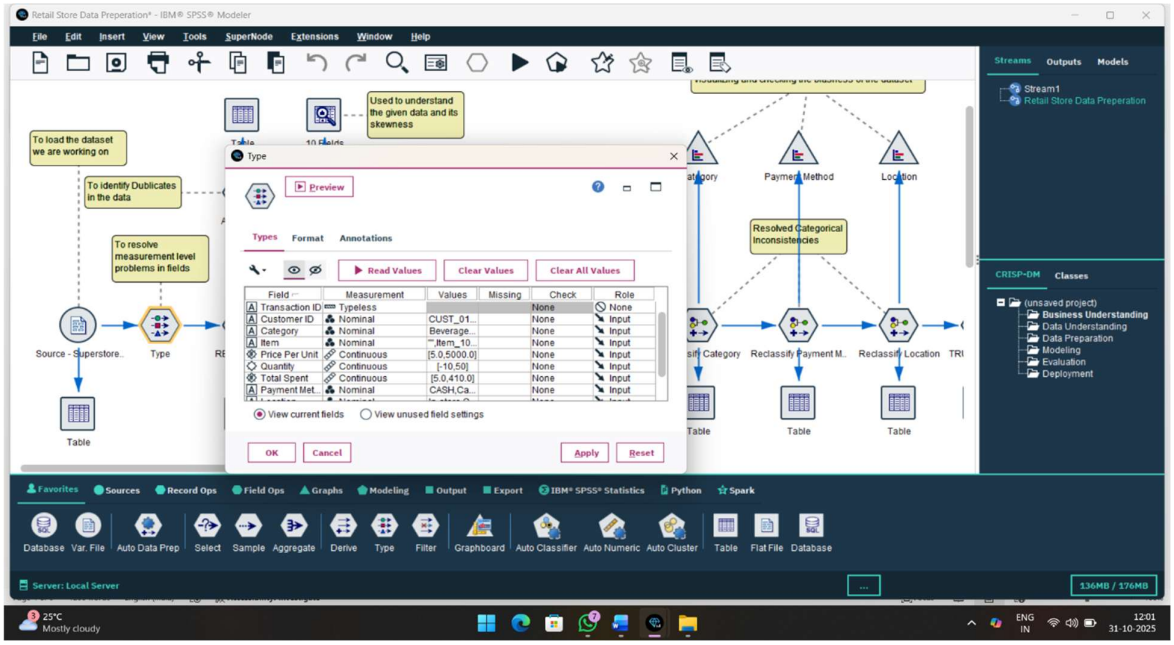
- **Role of Table Node:** Displays the imported data in a tabular format to visually verify that all fields and values are correctly loaded.



Step 2: Data Type Correction

Purpose: To assign accurate measurement levels and storage types to each variable for proper interpretation in the model flow.

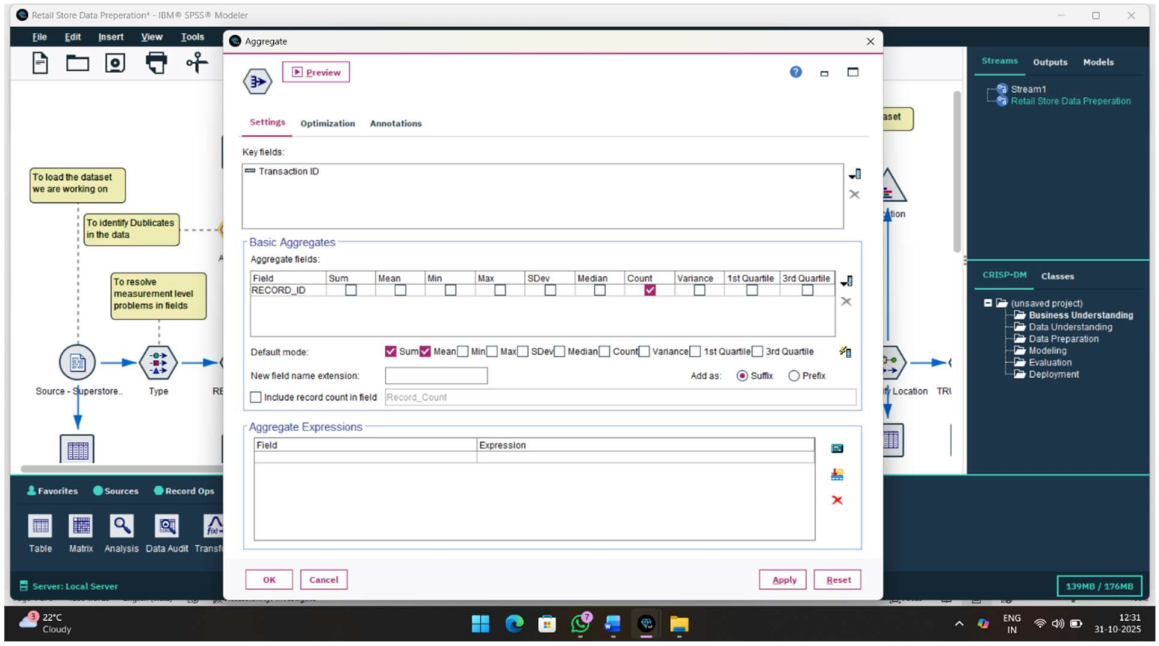
- **Role of Type Node:** Corrects and defines field measurement levels — *Nominal*, *Ordinal*, *Flag*, *Continuous*, *Categorical*, and *Typeless* — ensuring each variable is recognized correctly according to its data nature.

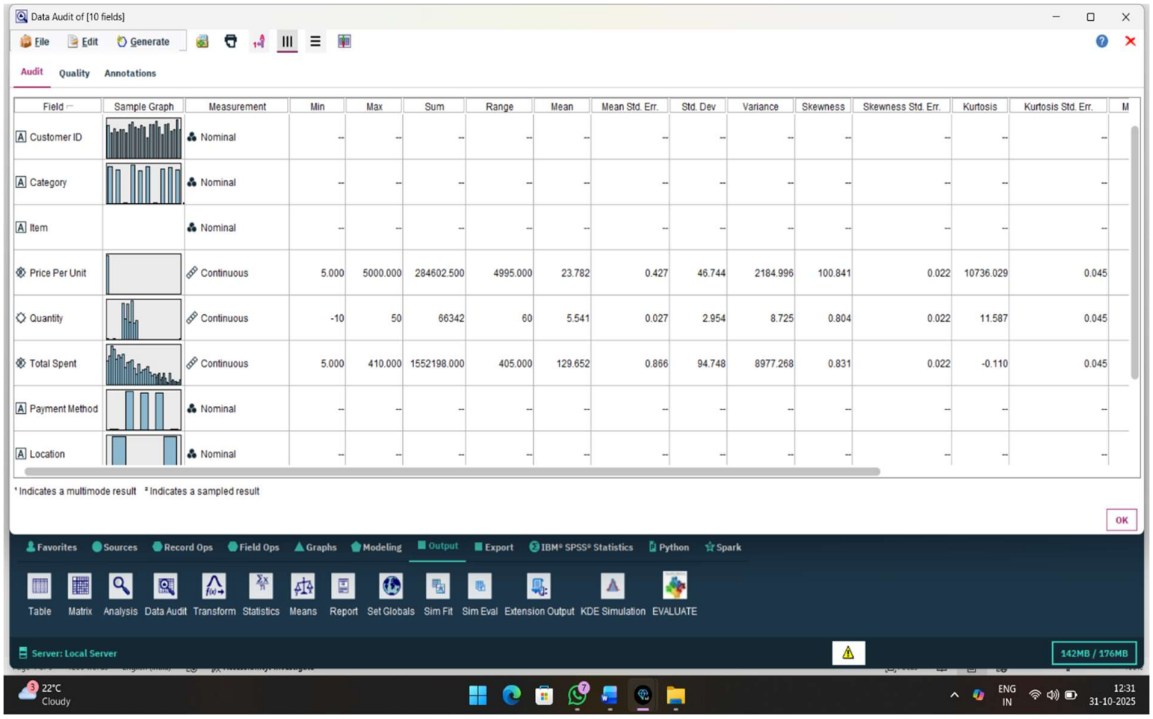


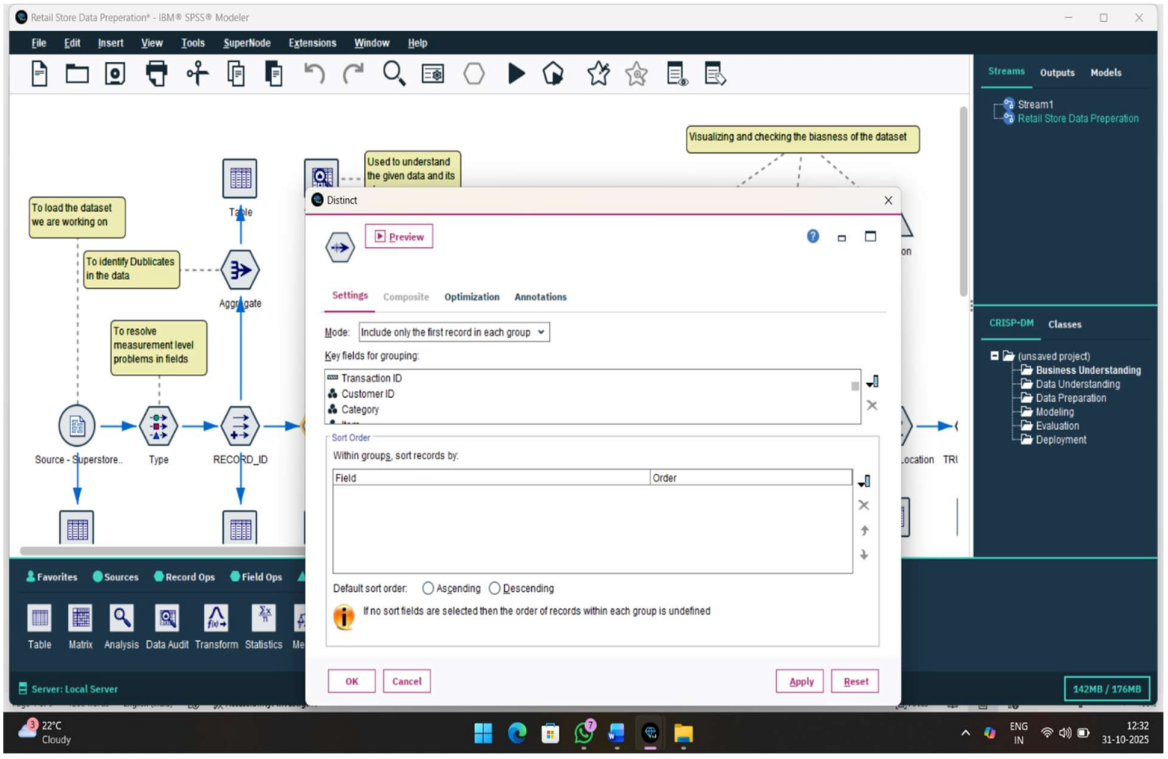
Step 3: Data Cleaning

Purpose: To identify and correct issues like duplicates, missing data, and inconsistent categorical values for a cleaner dataset.

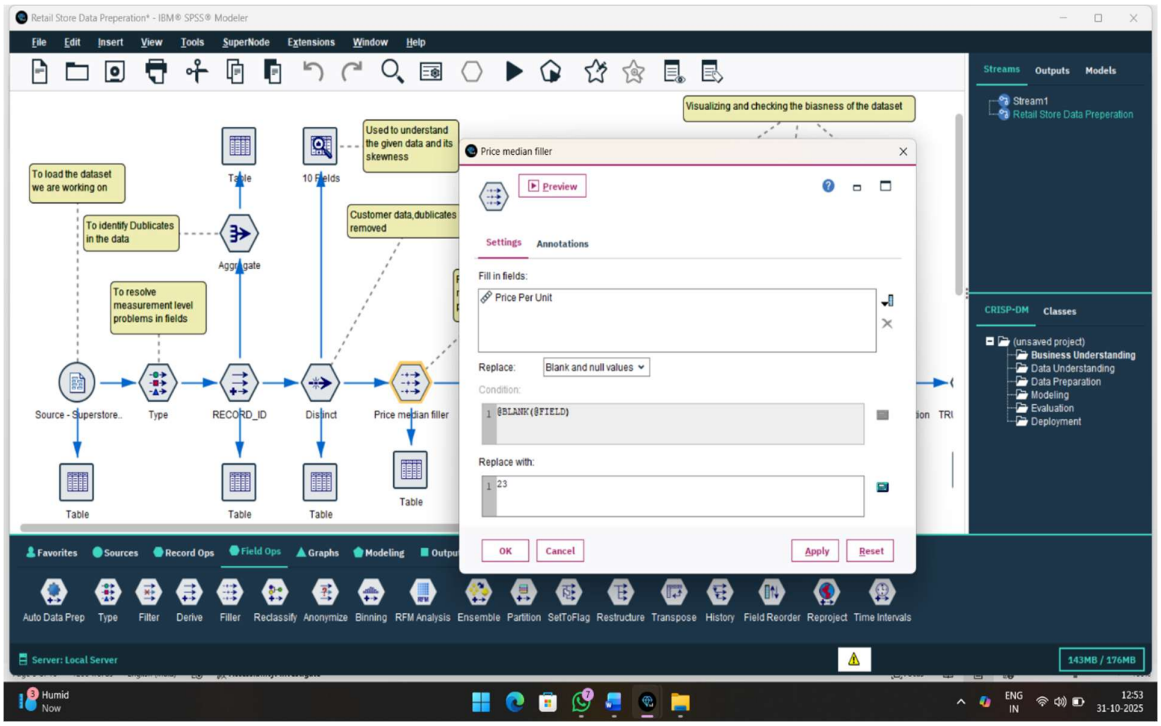
- **Role of Aggregate Node, Table Node, Data Audit Node, and Distinct Node** – *Duplicacy Removal:*
  Used to identify, summarize, and remove duplicate records while verifying results in tabular form.

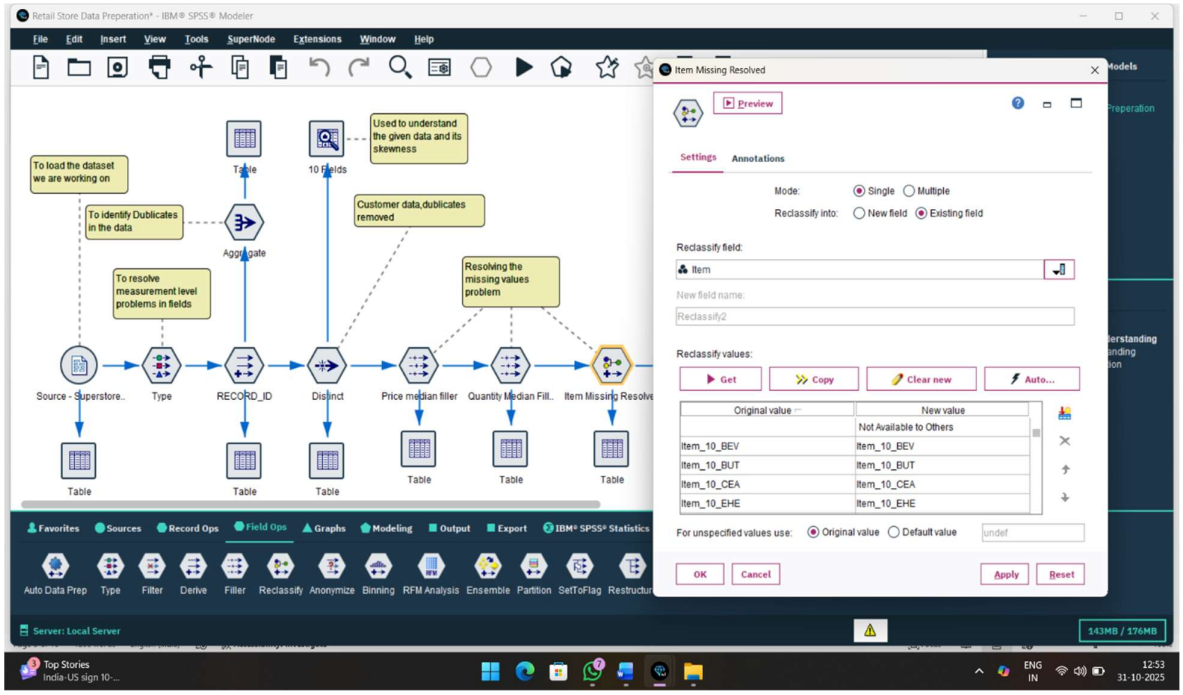




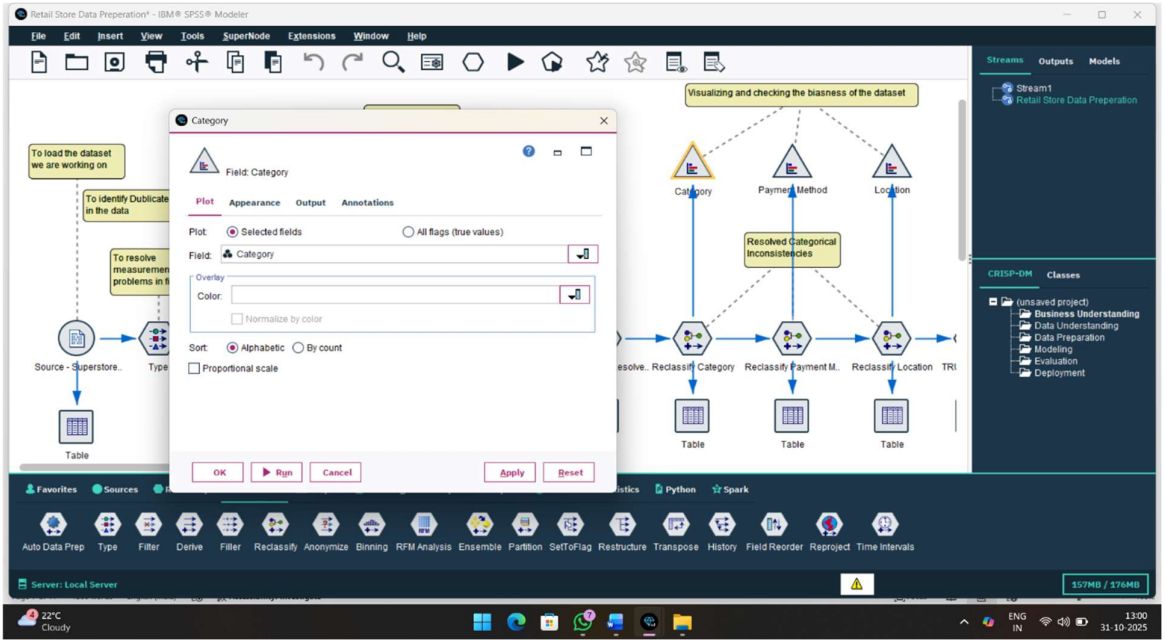


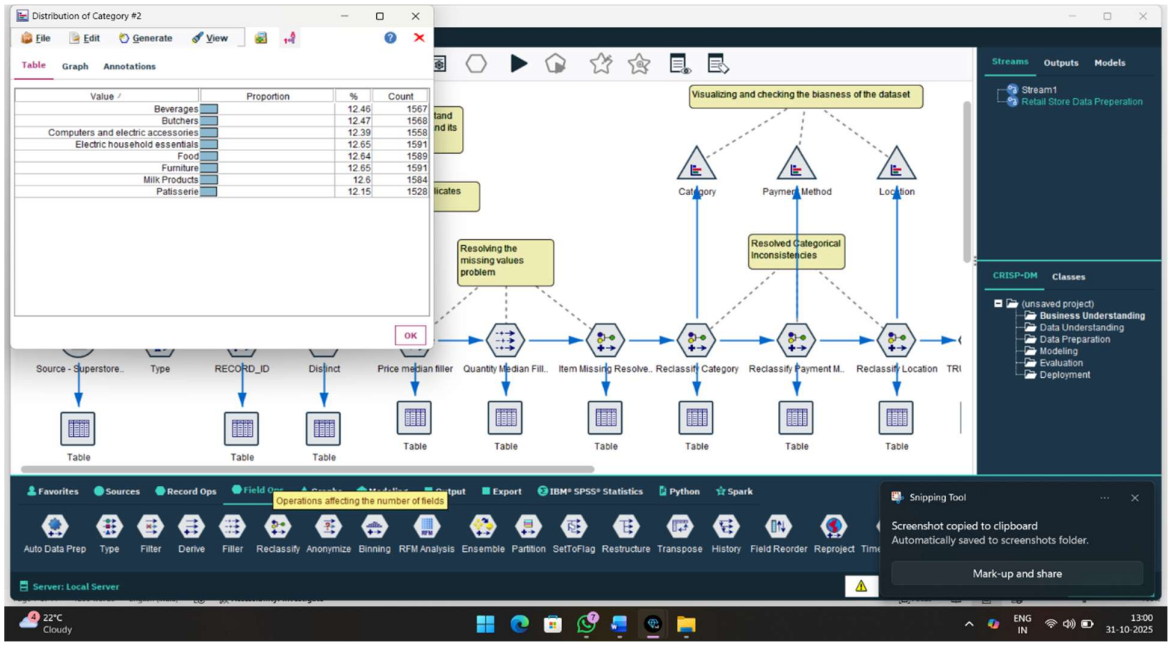
- **Role of Filler Nodes, Reclassify Nodes, and Table Nodes** – *Missing Value Resolution:* Used to fill or replace missing values appropriately and confirm the accuracy of updates.





- **Role of Reclassify Nodes, Graph Nodes, and Table Nodes** – *Categorical Inconsistency Resolution:*
  Used to standardize inconsistent category names and visually inspect distributions for uniformity.

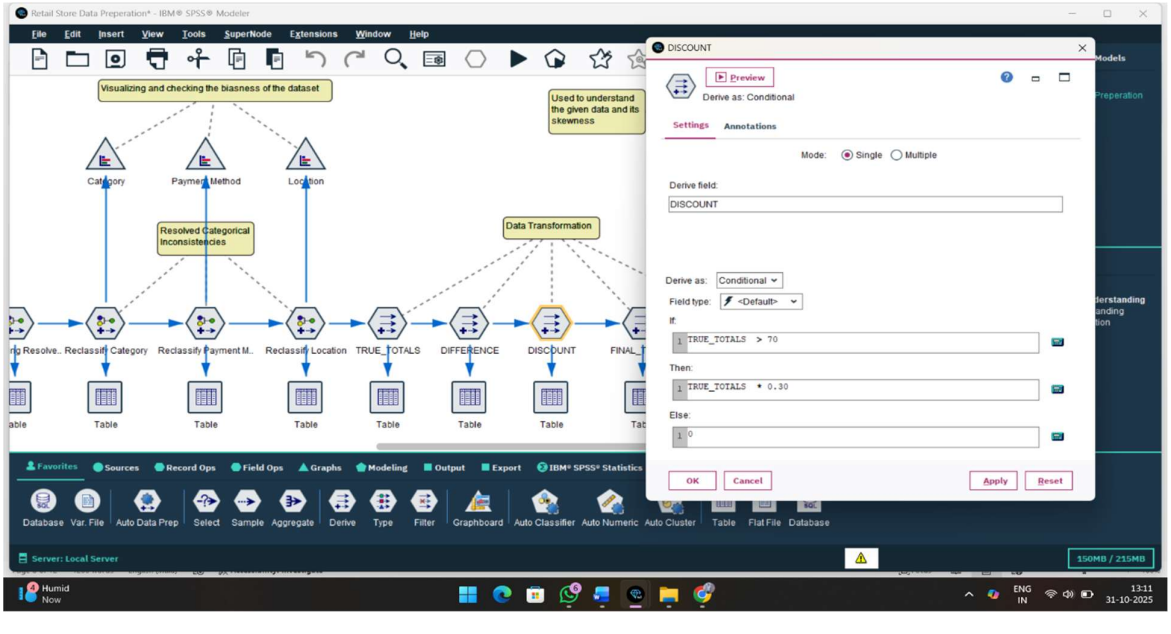




Step 4: Deriving New Variables

Purpose: To create new meaningful fields that help in understanding customer spending and discount behaviour.

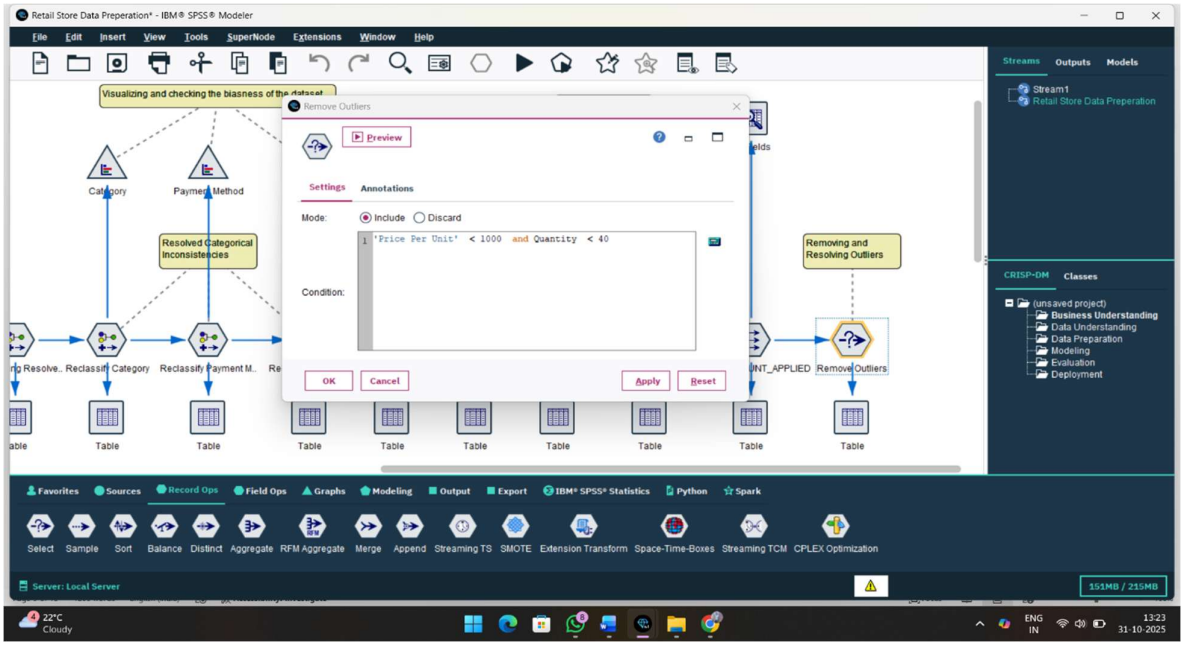
- **Role of Derive Nodes and Table Nodes:**
  Sequential Derive Nodes — **TRUE_TOTALS**, **DIFFERENCE**, **DISCOUNT**, **FINAL_TOTALS**, and **NEW_DISCOUNT_APPLIED** — were used to generate new calculated variables related to total spending, difference tracking, discount application, and final payment analysis.
  Table Nodes were placed after each Derive Node to review the output and ensure every transformation reflected correctly in the dataset.

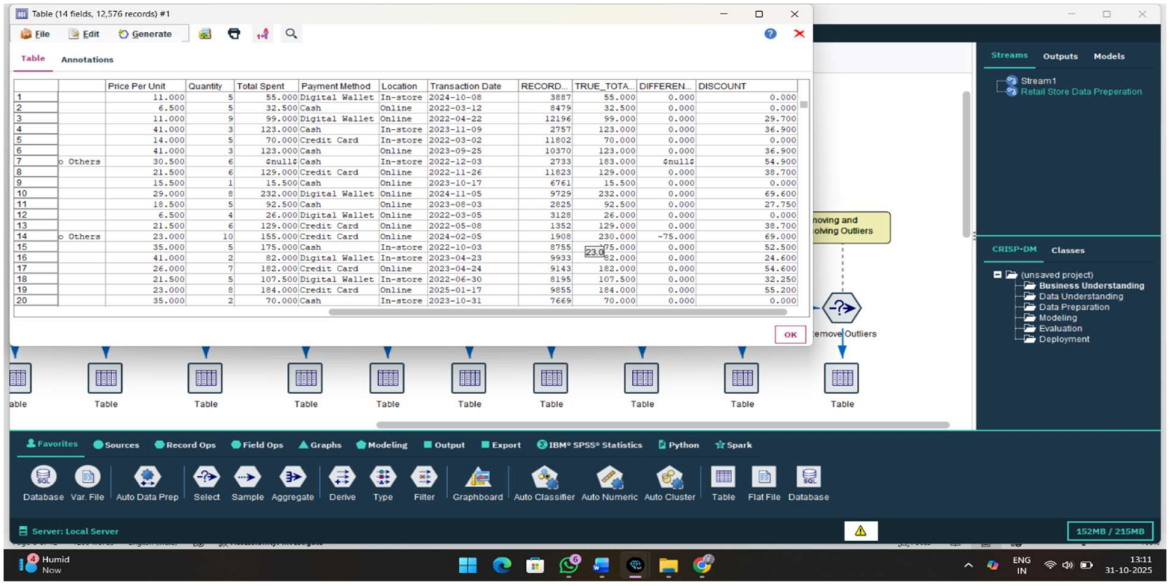


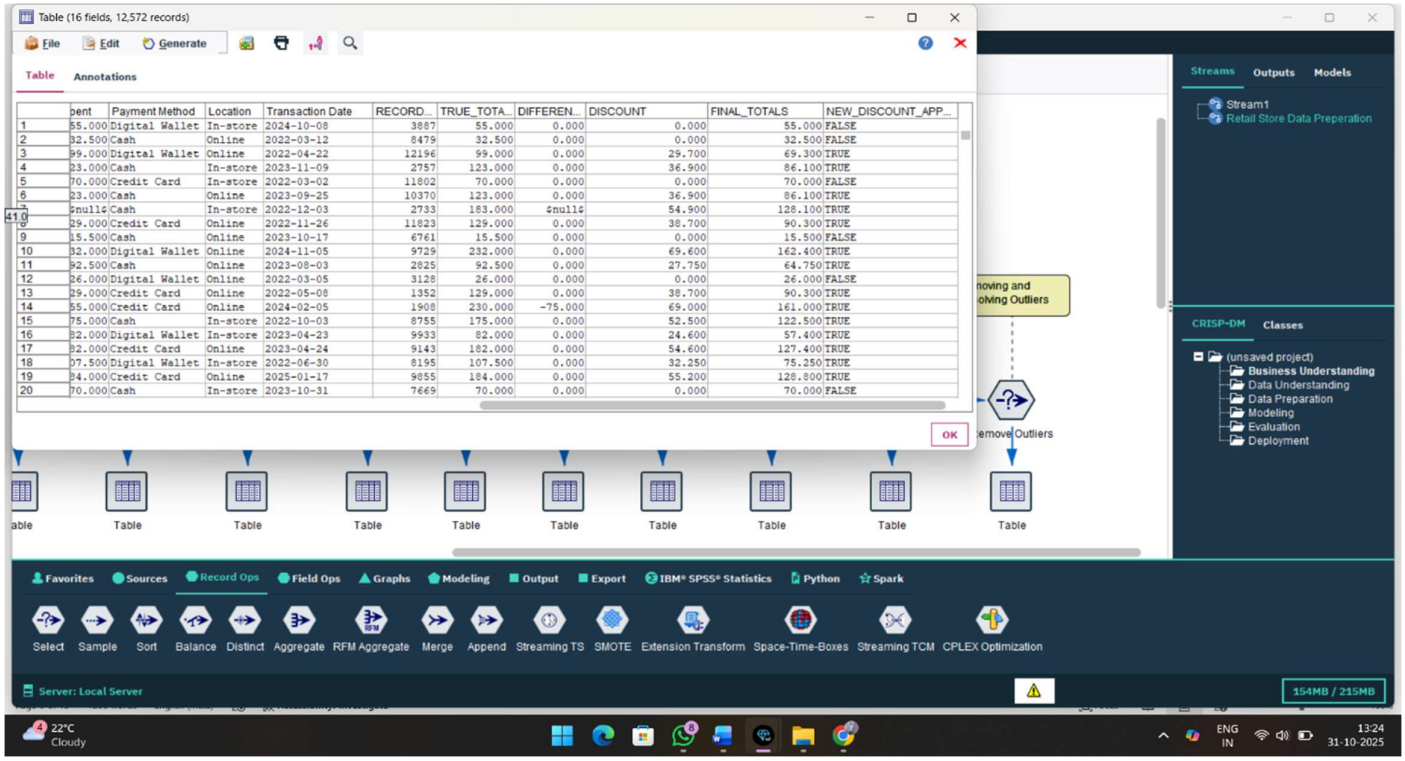
Step 5: Data Validation (Outlier Detection and Refinement)

Purpose: To validate the dataset by filtering out extreme or irrelevant data points to maintain result accuracy.

- **Role of Select Node and Table Node:**
  The Select Node was used to remove outliers or invalid records based on logical thresholds, and the Table Node displayed the refined data for final verification.







	pent	Payment Method	Location	Transaction Date	RECORD...	TRUE_TOTA...	DIFFEREN...	DISCOUNT	FINAL_TOTALS	NEW_DISCOUNT_APP...
1	55.000	Digital Wallet	In-store	2024-10-08	3887	55.000	0.000	0.000	55.000	FALSE
2	32.500	Cash	Online	2022-03-12	8479	32.500	0.000	0.000	32.500	FALSE
3	99.000	Digital Wallet	Online	2022-04-22	12196	99.000	0.000	29.700	69.300	TRUE
4	23.000	Cash	In-store	2023-11-09	2757	123.000	0.000	36.900	86.100	TRUE
5	70.000	Credit Card	In-store	2022-03-02	11802	70.000	0.000	0.000	70.000	FALSE
6	23.000	Cash	Online	2023-09-25	10370	123.000	0.000	36.900	86.100	TRUE
7	$null$	Cash	In-store	2022-12-03	2733	183.000	$null$	54.900	128.100	TRUE
8	29.000	Credit Card	Online	2022-11-26	11823	129.000	0.000	38.700	90.300	TRUE
9	15.500	Cash	Online	2023-10-17	6761	15.500	0.000	0.000	15.500	FALSE
10	32.000	Digital Wallet	Online	2024-11-05	9729	232.000	0.000	69.600	162.400	TRUE
11	92.500	Cash	Online	2023-08-03	2825	92.500	0.000	27.750	64.750	TRUE
12	26.000	Digital Wallet	Online	2022-03-05	3128	26.000	0.000	0.000	26.000	FALSE
13	29.000	Credit Card	Online	2022-05-08	1352	129.000	0.000	38.700	90.300	TRUE
14	55.000	Credit Card	Online	2024-02-05	1908	230.000	-75.000	69.000	161.000	TRUE
15	75.000	Cash	In-store	2022-10-03	8755	175.000	0.000	52.500	122.500	TRUE
16	82.000	Digital Wallet	In-store	2023-04-23	9933	82.000	0.000	24.600	57.400	TRUE
17	82.000	Credit Card	Online	2023-04-24	9143	182.000	0.000	54.600	127.400	TRUE
18	07.500	Digital Wallet	In-store	2022-06-30	8195	107.500	0.000	32.250	75.250	TRUE
19	84.000	Credit Card	Online	2025-01-17	9855	184.000	0.000	55.200	128.800	TRUE
20	70.000	Cash	In-store	2023-10-31	7669	70.000	0.000	0.000	70.000	FALSE