

# **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

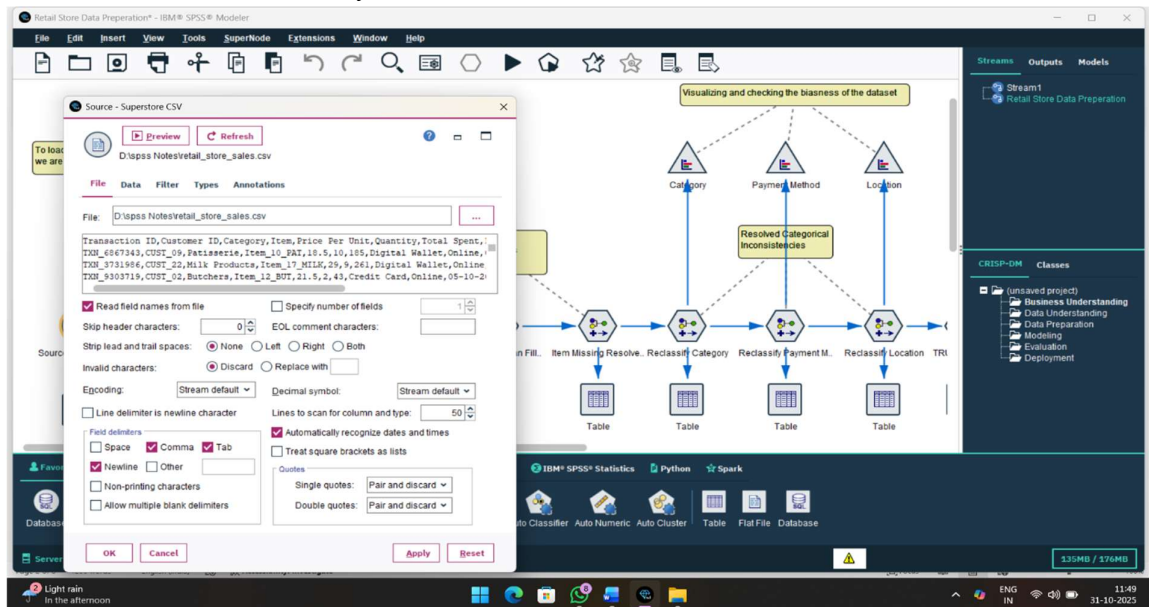
**Purpose:** To bring the dataset into SPSS Modeler and make it ready for analysis.

- **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).

The screenshot displays the IBM SPSS Modeler 18.6 interface. On the left, a 'Table' node shows a preview of the retail store sales dataset with 10 fields and 12,605 records. The table includes columns for Transaction ID, Customer ID, Category, Item, Price Per Unit, Quantity, Total Spent, and Payment Method. The main workspace shows a workflow diagram with nodes for 'Category', 'Payment Method', and 'Location', each connected to a 'Resolved Categorical Inconsistencies' node. The right sidebar shows the 'Streams' and 'Outputs' panels, with 'Stream1: Retail Store Data Preparation' selected. The bottom toolbar contains various nodes for data preparation, including 'Database', 'Var. File', 'Auto Data Prep', 'Select', 'Sample', 'Aggregate', 'Derive', 'Type', 'Filter', 'Graphboard', 'Auto Classifier', 'Auto Numeric', 'Auto Cluster', 'Table', 'Flat File', and 'Database'. The status bar at the bottom indicates the server is 'Local Server' and the project size is '136MB / 176MB'.

Transaction ID	Customer ID	Category	Item	Price Per Unit	Quantity	Total Spent	Payment Method
TXM_6867343	CUST_09	Patisserie	Item_10_PAT	18.500	10	185.000	Digital Wallet
TXM_3731986	CUST_22	Milk Products	Item_17_MILK	29.000	9	261.000	Digital Wallet
TXM_3003719	CUST_02	Butchers	Item_12_BUT	21.500	2	43.000	Credit Card
TXM_9458126	CUST_06	Beverages	Item_16_BEV	27.500	9	247.500	Credit Card
TXM_4575373	CUST_05	Food	Item_6_FOOD	12.500	7	87.500	Digital Wallet
TXM_7482416	CUST_09	Patisserie	Item_1_FOOD	200.000	10	200.000	Credit Card
TXM_3652209	CUST_07	Food	Item_1_FOOD	9.000	8	40.000	Credit Card
TXM_1372952	CUST_21	Furniture	Item_1_FOOD	33.500	Small	Small	Digital Wallet
TXM_9728486	CUST_23	Furni	Item_16_FUR	27.500	1	27.500	Credit Card
TXM_2722461	CUST_25	Butchers	Item_21_BUT	36.500	3	109.500	CASH
TXM_8776416	CUST_22	Butchers	Item_3_BUT	8.000	9	72.000	Cash
TXM_5422631	CUST_09	Milk Products	Item_10_PAT	66.5	Small	Small	Digital Wallet
TXM_5874772	CUST_23	Food	Item_2_FOOD	6.500	7	45.500	Cash
TXM_4413070	CUST_14	Patisserie	Item_24_PAT	39.500	6	237.000	Digital Wallet
TXM_2450363	CUST_09	Milk Products	Item_16_MILK	27.500	2	55.000	Digital Wallet
TXM_1809665	CUST_14	Beverages	Item_17_PAT	24.500	Small	Small	Credit Card
TXM_7563911	CUST_23	Patisserie	Item_17_PAT	29.000	8	232.000	CASH
TXM_9634894	CUST_15	Milk Products	Item_17_PAT	Small	10	275.000	Digital Wallet
TXM_4396807	CUST_17	Electric h...	Item_13_EHE	23.000	1	23.000	Digital Wallet
TXM_4206593	CUST_01	Furniture	Item_13_EHE	35.000	Small	Small	Digital Wallet

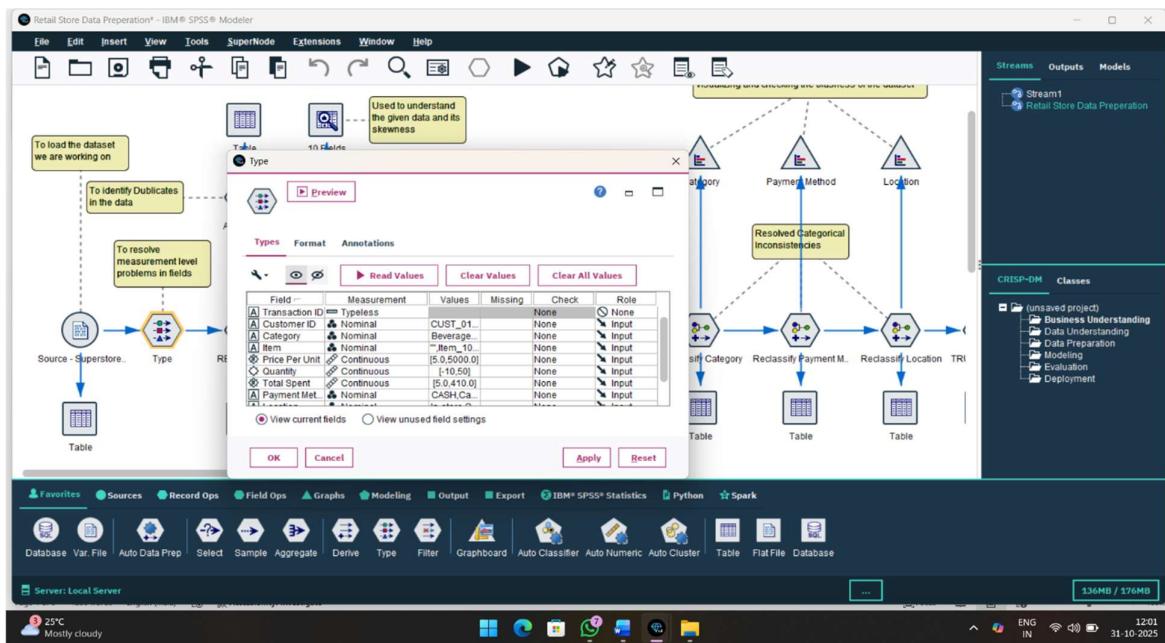
- **Role of Table Node:** Displays the imported data in a tabular format to visually verify that all fields and values are correctly loaded.



## Step 2: Data Type Correction

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

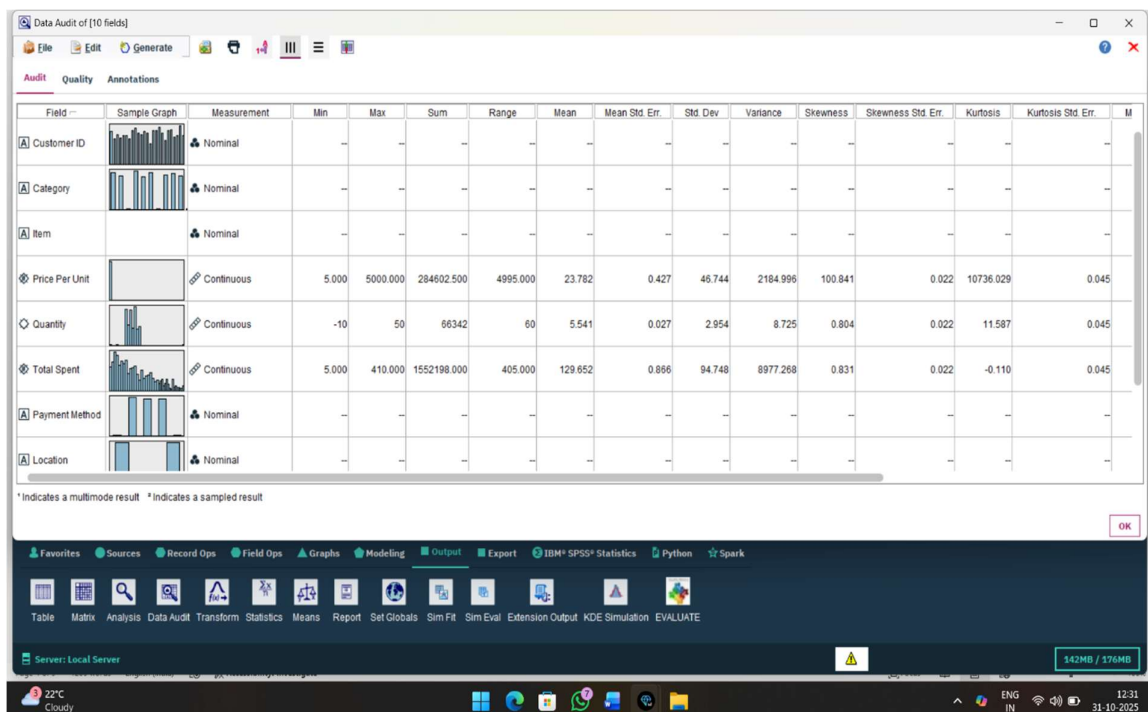
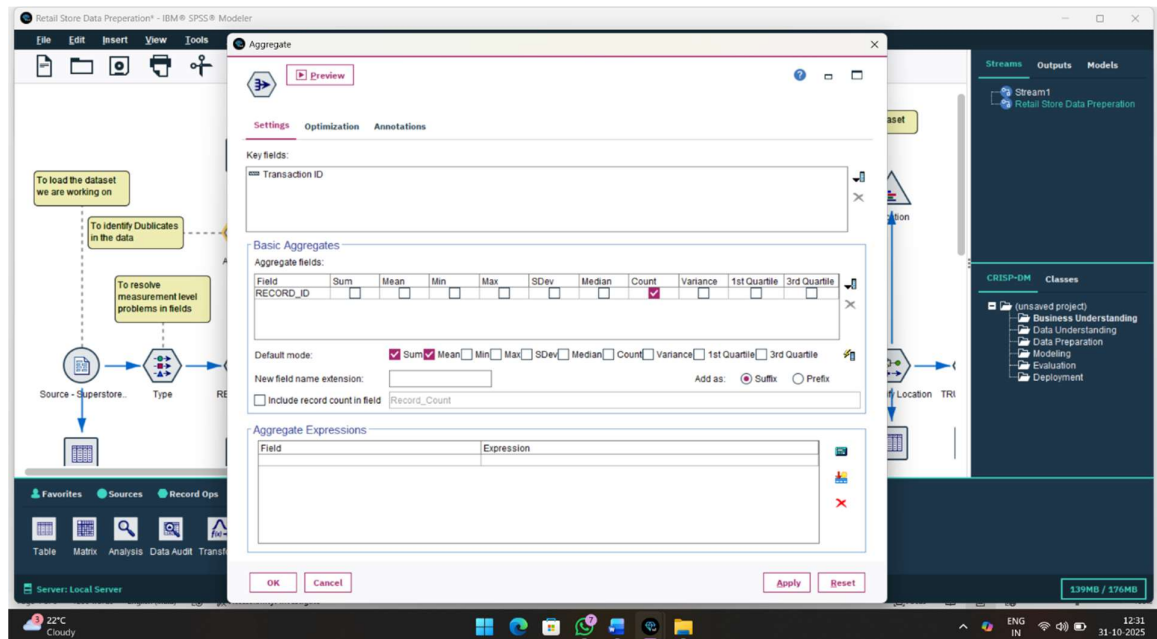
- **Role of Type Node:** Corrects and defines field measurement levels — *Nominal*, *Ordinal*, *Flag*, *Continuous*, *Categorical*, and *Typeless* — ensuring each variable is recognized correctly according to its data nature.

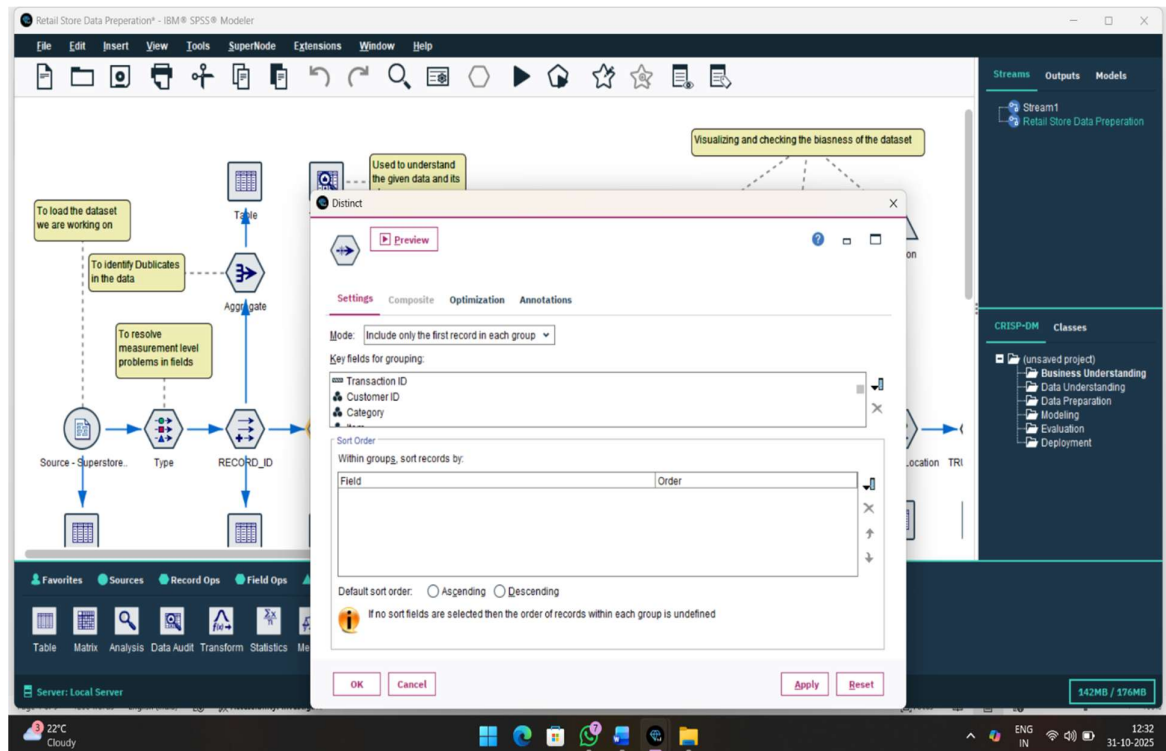


### Step 3: Data Cleaning

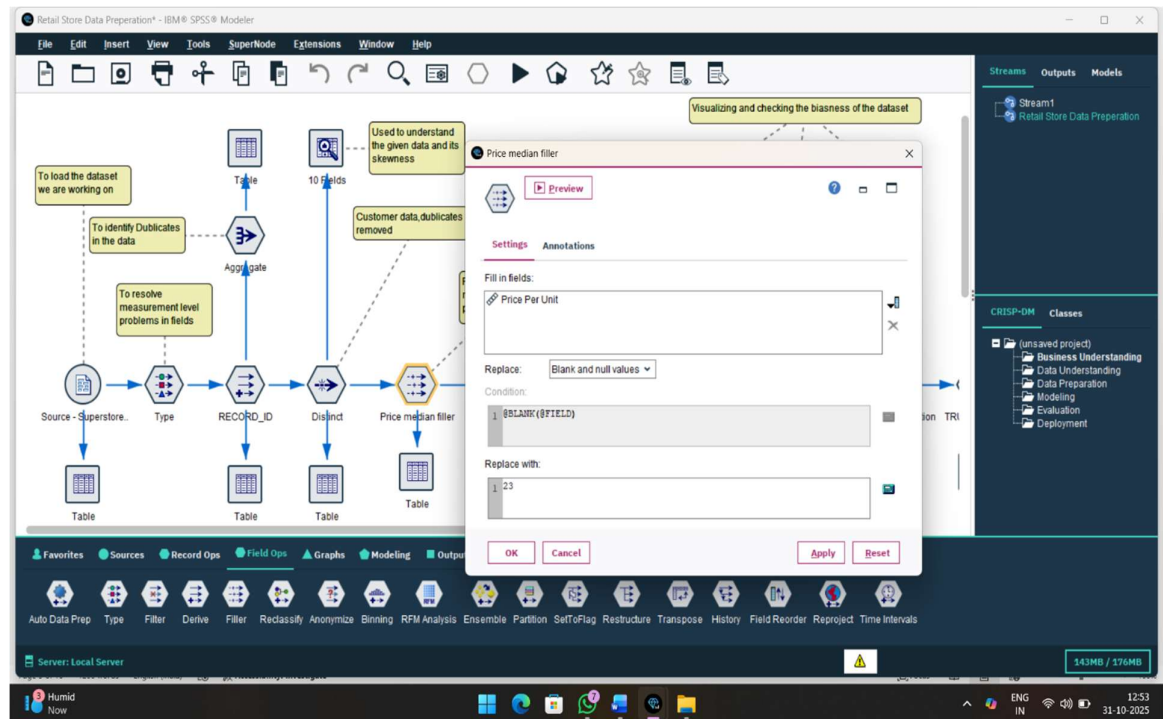
**Purpose:** To identify and correct issues like duplicates, missing data, and inconsistent categorical values for a cleaner dataset.

- 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.

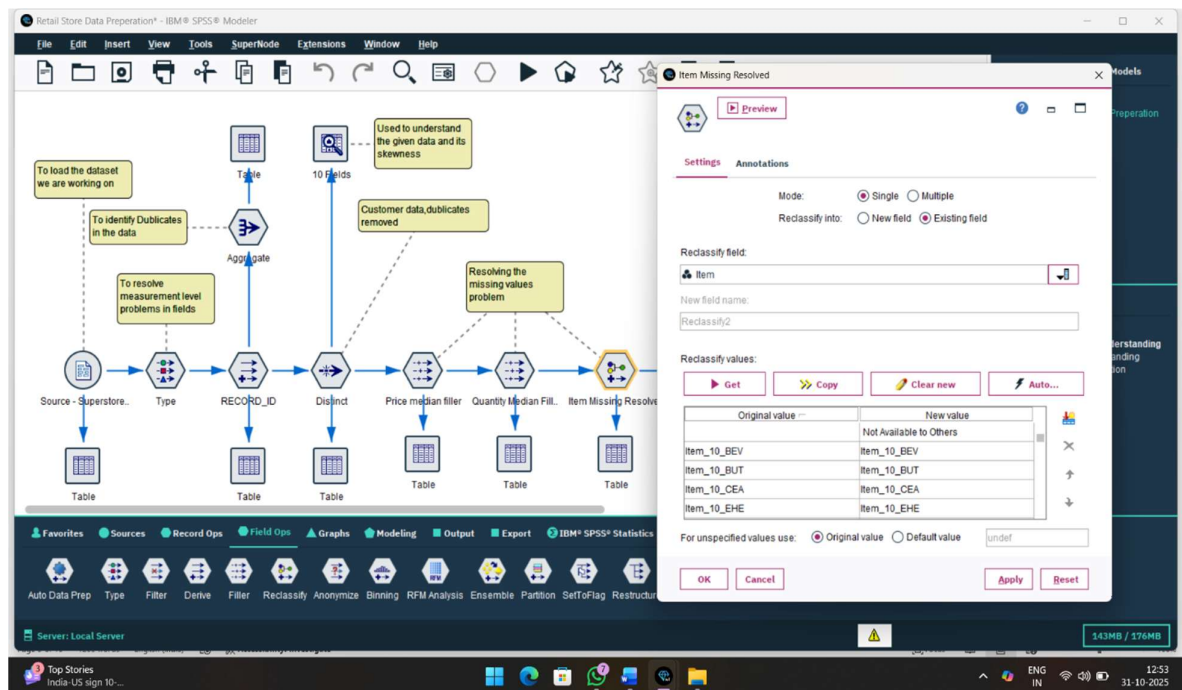




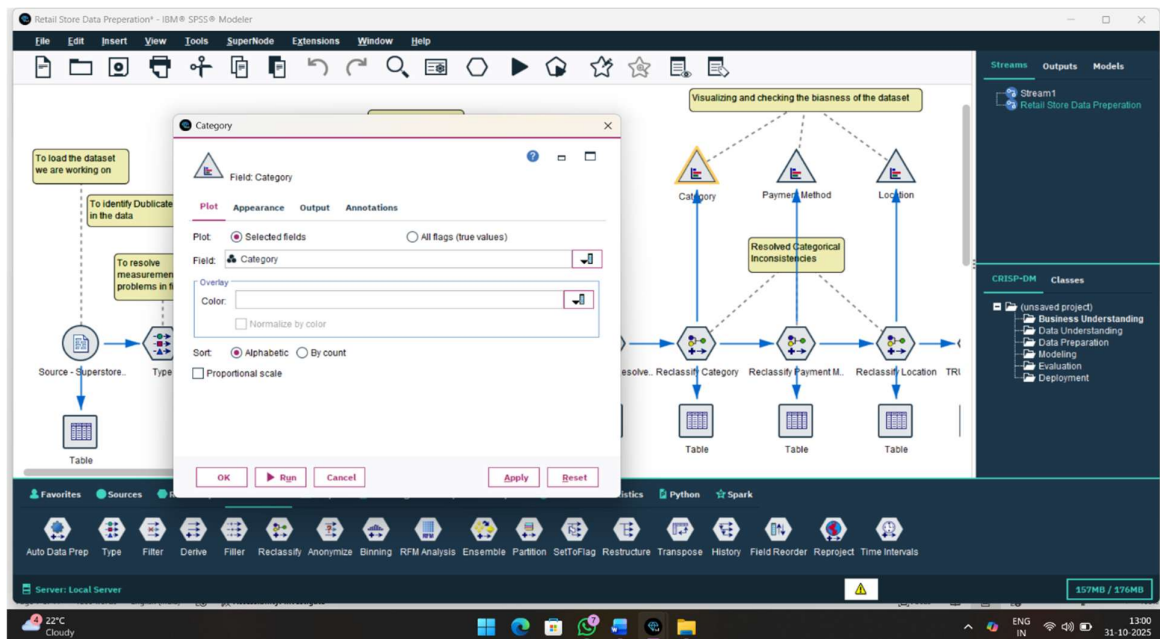
- **Role of Filler Nodes, Reclassify Nodes, and Table Nodes – *Missing Value Resolution:***  
Used to fill or replace missing values appropriately and confirm the accuracy of updates.

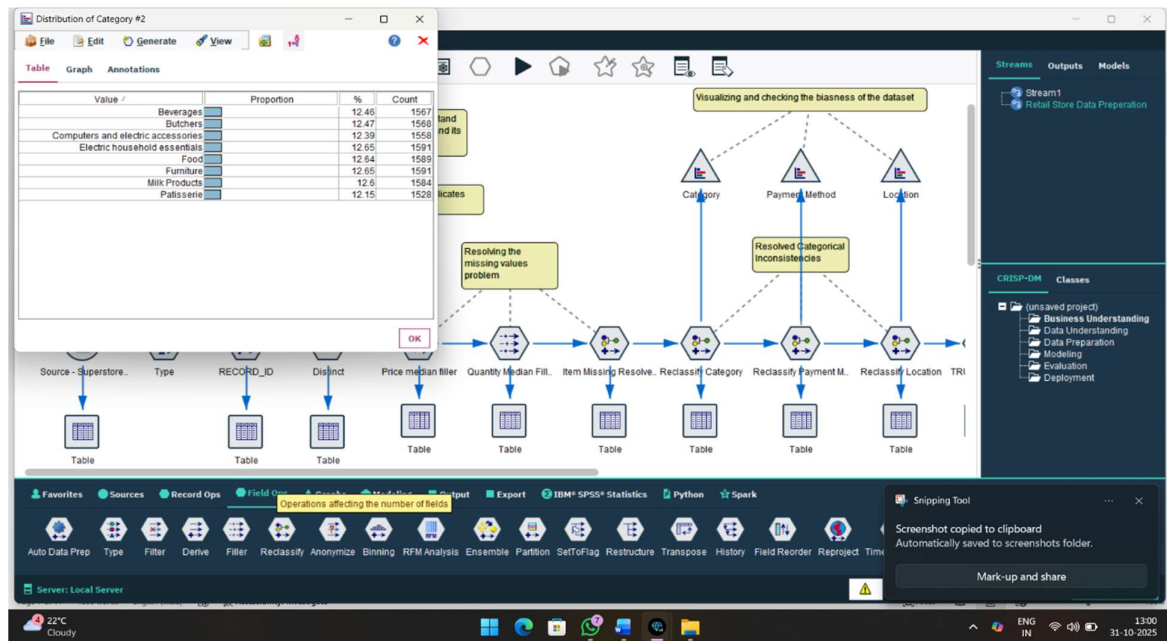






- Role of Reclassify Nodes, Graph Nodes, and Table Nodes – Categorical Inconsistency Resolution:**  
 Used to standardize inconsistent category names and visually inspect distributions for uniformity.

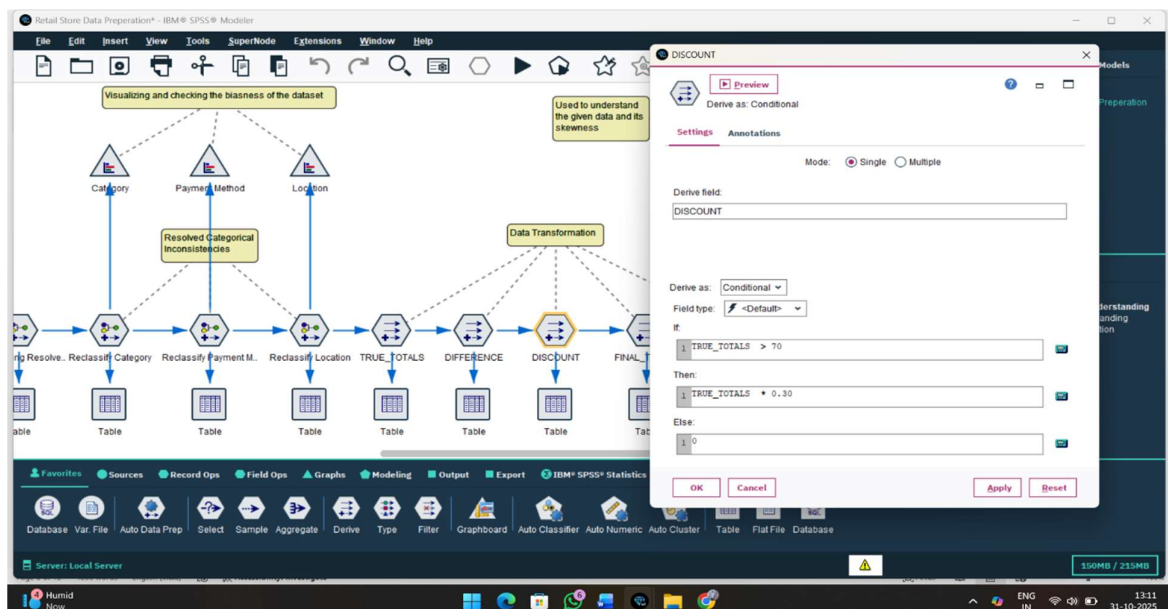




## Step 4: Deriving New Variables

**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.

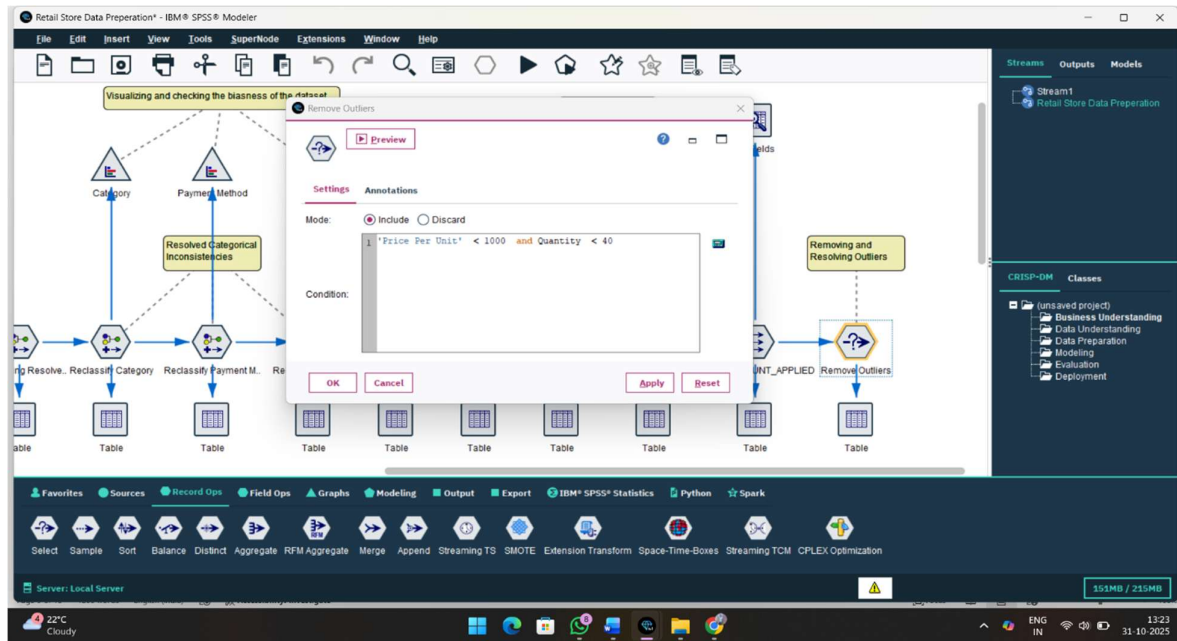


## Step 5: Data Validation (Outlier Detection and Refinement)

**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.



The screenshot shows a data table with 14 fields and 12,576 records. The table is displayed in a grid format with columns for 'Price Per Unit', 'Quantity', 'Total Spent', 'Payment Method', 'Location', 'Transaction Date', 'RECORD', 'TRUE TOTAL', 'DIFFEREN', and 'DISCOUNT'. The data is sorted by 'RECORD' in ascending order. The table is titled 'Table (14 fields, 12,576 records) #1'. The right sidebar shows the 'Streams' panel with 'Stream1: Retail Store Data Preparation' and the 'CRISP-DM' classes. The bottom toolbar includes various data processing tools like 'Database Var. File', 'Auto Data Prep', 'Select', 'Sample', 'Aggregate', 'Derive', 'Type', 'Filter', 'Graphboard', 'Auto Classifier', 'Auto Numeric', 'Auto Cluster', 'Table', 'Flat File', and 'Database'.

	Price Per Unit	Quantity	Total Spent	Payment Method	Location	Transaction Date	RECORD	TRUE TOTAL	DIFFEREN	DISCOUNT
1	11,000	5	55,000	Digital Wallet	In-store	2024-10-09	3887	55,000	0.000	0.000
2	6,500	5	32,500	Cash	Online	2022-03-12	8479	32,500	0.000	0.000
3	11,000	9	99,000	Digital Wallet	Online	2022-04-22	12194	99,000	0.000	29,700
4	41,000	3	123,000	Cash	In-store	2023-11-09	2197	123,000	0.000	36,900
5	14,000	5	70,000	Credit Card	In-store	2022-03-02	11802	70,000	0.000	0.000
6	41,000	3	123,000	Cash	Online	2023-09-25	10370	123,000	0.000	36,900
7	30,800	6	184,800	Cash	In-store	2022-12-03	2738	183,000	Small	44,900
8	21,800	6	129,000	Credit Card	Online	2022-11-24	11823	129,000	0.000	38,700
9	15,800	1	15,800	Cash	Online	2023-10-17	4761	15,500	0.000	0.000
10	29,000	8	232,000	Digital Wallet	Online	2024-11-05	9728	232,000	0.000	49,400
11	18,500	5	92,500	Cash	Online	2023-08-03	2825	92,500	0.000	27,750
12	6,500	4	26,000	Digital Wallet	Online	2022-03-05	3128	26,000	0.000	0.000
13	21,800	6	129,000	Credit Card	Online	2022-05-08	1352	129,000	0.000	38,700
14	23,000	10	155,000	Credit Card	Online	2024-02-05	1508	230,000	-75,000	69,000
15	35,000	5	175,000	Cash	In-store	2022-10-03	8755	175,000	0.000	52,500
16	41,000	2	82,000	Digital Wallet	In-store	2023-04-23	9933	82,000	0.000	24,600
17	26,000	7	182,000	Credit Card	Online	2023-04-24	9143	182,000	0.000	54,600
18	21,800	5	107,500	Digital Wallet	In-store	2022-06-30	8195	107,500	0.000	32,250
19	23,000	8	184,000	Credit Card	Online	2025-01-17	9855	184,000	0.000	55,200
20	35,000	2	70,000	Cash	In-store	2023-10-31	7669	70,000	0.000	0.000



Table (16 fields, 12,572 records)

FileEditGenerate

Annotations

	Amount	Payment Method	Location	Transaction Date	RECORD_ID	TRUE_TOTAL	DIFFERENCE	DISCOUNT	FINAL_TOTALS	NEW_DISCOUNT_APPLICATION
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	9479	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	23.000	Cash	In-store	2022-12-03	2733	183.000	0.000	54.900	128.100	TRUE
8	29.000	Credit Card	Online	2022-11-26	11823	129.000	0.000	36.700	90.300	TRUE
9	15.500	Cash	Online	2023-10-17	4761	15.500	0.000	0.000	15.500	FALSE
10	32.000	Digital Wallet	Online	2024-11-05	9729	232.000	0.000	69.400	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	36.700	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	32.000	Digital Wallet	In-store	2023-04-23	9933	82.000	0.000	24.600	57.400	TRUE
17	32.000	Credit Card	Online	2023-04-24	9143	182.000	0.000	54.600	127.400	TRUE
18	17.500	Digital Wallet	In-store	2022-06-30	8195	107.500	0.000	32.250	75.250	TRUE
19	24.000	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

moving and solving Outliers

remove Outliers

OK

Table

Table

Table

Table

Table

Table

Table

Table

Table

Table

FavoritesSourcesRecord OpsField OpsGraphsModelingOutputExportIBM® SPSS® StatisticsPythonSpark

SelectSampleSortBalanceDistinctAggregateRFM AggregateMergeAppendStreaming TSSMOTEExtension TransformSpace-Time-BoxesStreaming TCMCPLEX Optimization

Server: Local Server

154MB / 215MB

StreamsOutputsModels

Stream1

Retail Store Data Preparation

CRISP-DMClasses

(unsaved project)

Business Understanding

Data Understanding

Data Preparation

Modeling

Evaluation

Deployment

22°C Cloudy

ENG

IN

13:24

31-10-2025