

Babu Banarasi Das

University



Case Study: Data Preparation and Quality Improvement of Health Risk Datasets

SUBMITTED TO

Mr. Robin Tyagi Sir

SUBMITTED BY:

Name: Gunjan Singh Solanki

Roll No: 1230258186

Class: BCADS 33

Name: Farwa Fatima

Roll No: 1230258175

Class: BCADS 33

Definition:

In this practical, the objective is to merge, cleanse, and standardize multiple health-related datasets containing demographic and behavioural data. The task involves handling missing values, correcting data inconsistencies, and ensuring categorical uniformity across fields such as *Gender*, *Smoking*, *Alcohol Consumption*, *Marital Status*, and *Exercise Habits*. Additionally, derived calculations such as *BMI* (*Body Mass Index*) are computed for further analysis.

Required Tool:

IBM SPSS Modeler

Outcomes/Learning:

- Learned how to combine multiple datasets using the **Append Node** for unified analysis.
- Understood how to identify and remove duplicate records to maintain data accuracy.
- Gained experience in handling missing values using **Median Filler** and **Reclassify Nodes**.
- Learned how to standardize inconsistent categorical fields (Gender, Smoking, Alcohol, Married, and Exercise).
- Developed the ability to derive new calculated fields such as **Corrected BMI** using CLEM expressions.

Working:

SPSS Modeler follows the **CRISP-DM (Cross-Industry Standard Process for Data Mining)** framework, which includes six stages — *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, and *Deployment*.

In this practical, the focus is on the **Data Preparation** stage, where multiple datasets are cleaned, transformed, and standardized for consistent analysis.

The two datasets — *Health_Risk_Female_Data* and *Health_Risk_Male_Data* — are imported and combined using the **Append Node** to create a unified dataset. Data types are standardized using the **Type Node**, and redundant records are eliminated using the **Remove Duplicates Node**.

Next, **Reclassify Nodes** are used to resolve categorical inconsistencies in the *Gender*, *Smoking*, *Alcohol*, and *Marital Status* fields, ensuring uniform labelling.

Missing numeric values in *Age*, *Weight*, and *Height* are filled using **Median Filler Nodes** to ensure balanced data distribution. The **Exercise** field is then standardized by replacing blank or missing entries with the text “**Not Available**” using a **Reclassify Node**.

Finally, a **Derive Node** is used to compute a corrected *BMI* (*Body Mass Index*) value using a CLEM formula that accounts for height and weight.

Each step is validated using **Table Nodes**, ensuring data quality before proceeding to further modeling or analysis.

Steps:

Step 1: Importing and Viewing Datasets

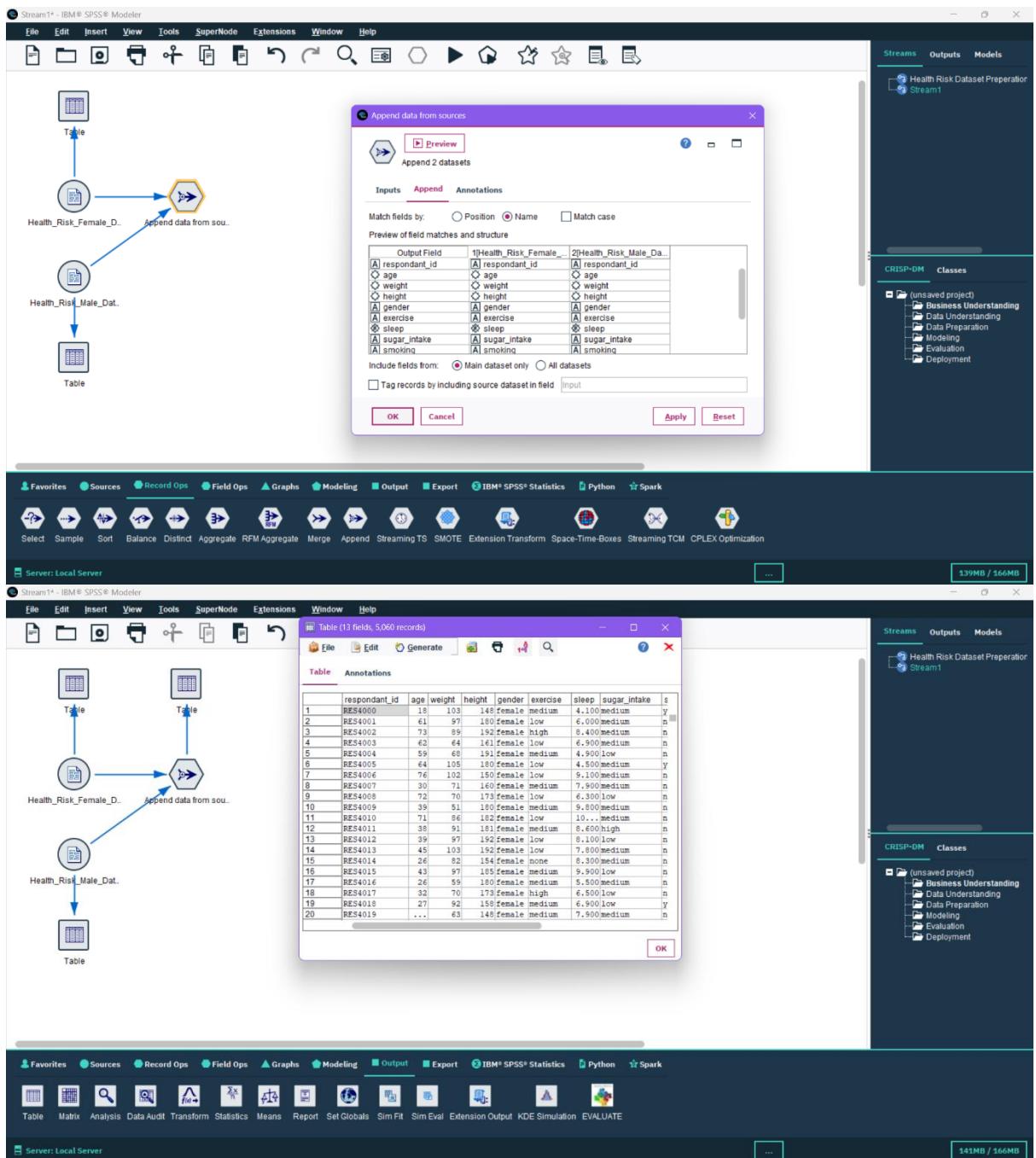
- Open IBM SPSS Modeler and create a new stream.
- From the **Sources** tab, add two **Var. File Nodes** to import *Health_Risk_Female_Data* and *Health_Risk_Male_Data*.
- Click **Apply → OK** to load both datasets.
- Connect **Table Nodes** to preview the records.

The screenshot illustrates the workflow for importing and viewing datasets in IBM SPSS Modeler:

- Top Left:** A Var. File Node for "Health_Risk_Female_Dat." is open, showing its configuration. It reads from "D:\BCIbm spss\project2\Health_Risk_Female_Dataset.csv". The "File" tab is selected, and the "OK" button is visible at the bottom.
- Top Right:** A Var. File Node for "Health_Risk_Male_Dat." is open, showing its configuration. It reads from "D:\BCIbm spss\project2\Health_Risk_Male_Dataset.csv". The "File" tab is selected, and the "OK" button is visible at the bottom.
- Middle:** The "Sources" tab is active, showing two nodes: "Health_Risk_Female_Dat." and "Health_Risk_Male_Dat.". Both nodes have a "Table" icon connected to them, indicating they are being previewed.
- Bottom:** A Stream window titled "am1" is displayed. It contains a "Table" node connected to the "Health_Risk_Female_Dat." and "Health_Risk_Male_Dat." nodes. The "Annotations" tab is selected, showing the first 20 records of the combined dataset. The columns are: respondent_id, age, weight, height, gender, exercise, sleep, sugar_intake, t. The data includes rows such as RES4000, RES4001, RES4002, etc., with various demographic and health values.
- Bottom Navigation:** The main menu bar includes "File", "Edit", "Insert", "View", "Tools", "SuperNode", "Extensions", "Window", and "Help". The toolbar below includes icons for Analytic Server, Database, Var. File, Fixed File, Statistics File, Data Collection, IBM Cognos Analytics, TM1 Import, TWC Import, SAS File, Excel, XML, User Input, Sim Gen, Extension Import, Data View, JSON, and Geospatial.

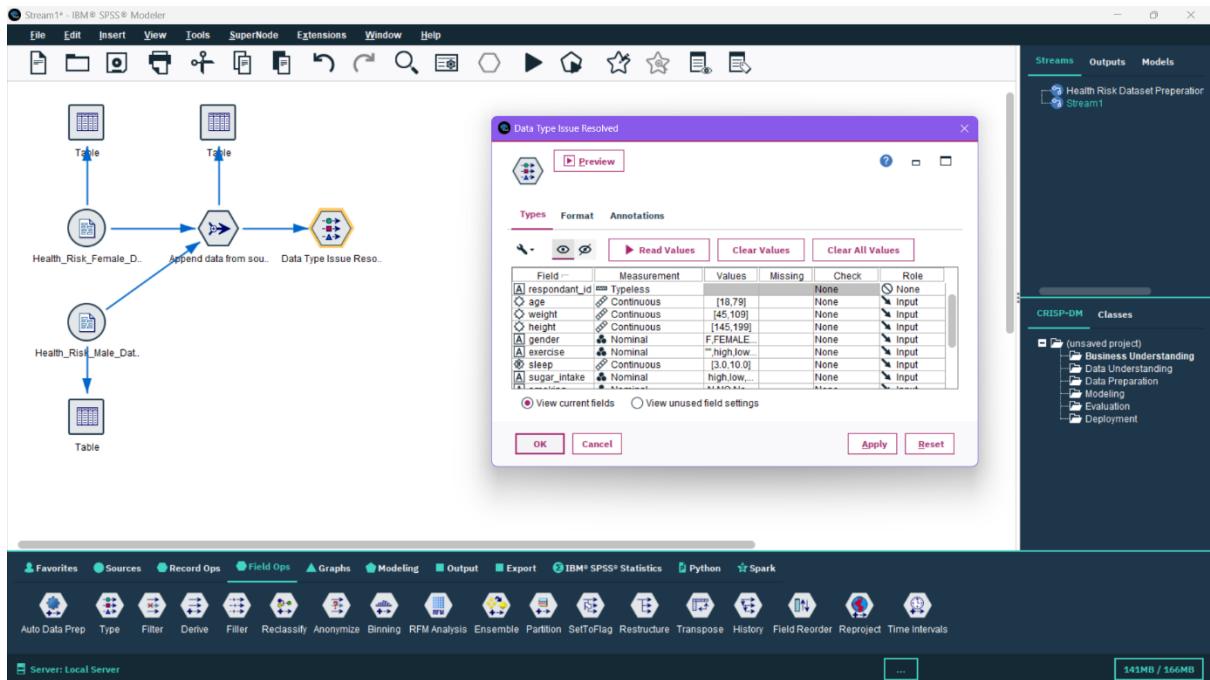
Step 2: Merging Datasets Using Append Node

- Go to the **Record Ops** tab and double-click **Append**.
- Connect both input datasets to the Append Node.
- Click **Apply → OK** to merge and create a unified dataset.
- Add a **Table Node** to verify the appended data.



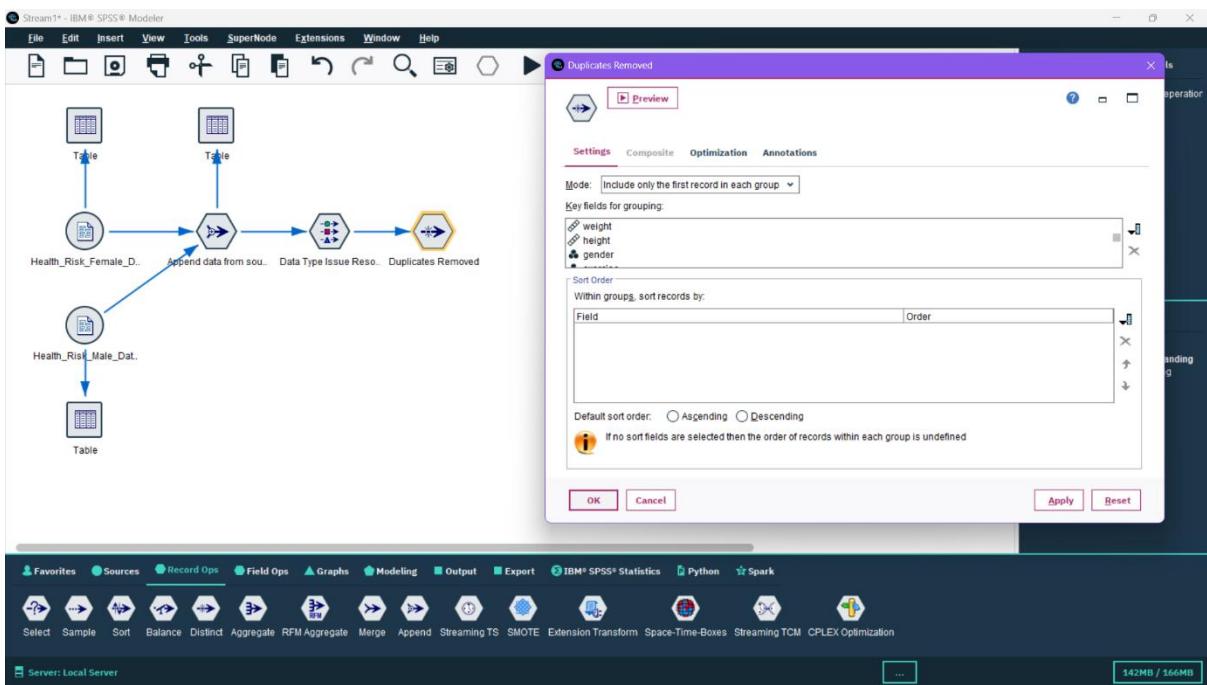
Step 3: Resolving Data Type Issues

- Connect a **Type Node** to the Append Node.
- Click **Read Values** and update incorrect data types or measurement levels.
- Apply and run the node to standardize field definitions.



Step 4: Removing Duplicate Records

- Add a **Remove Duplicates Node** to eliminate repeated entries.
- Select key identifiers such as *Patient ID* or *Name*.
- Run the node and verify the results through a **Table Node**.



StreamIt! - IBM SPSS Modeler

Streams Outputs Models

Health Risk Dataset Preparator Stream

Table (13 fields, 5,000 records)

Annotations

	respondent_id	age	weight	height	gender	exercise	sleep	sugar_intake	s
1	RES1000	56	67	195M	male	low	6.100	medium	Y
2	RES1001	69	76	174M	male	high	6.400	medium	n
3	RES1002	46	106	153M	female	high	6.400	low	Y
4	RES1003	32	54	196M	male	medium	8.500	medium	n
5	RES1004	60	98	195M	male	high	8.000	low	n
6	RES1005	25	96	160M	male	medium	3.800	medium	n
7	RES1006	78	64	168M	male	medium	9.900	high	n
8	RES1007	38	76	194M	male	low	6.600	medium	Y
9	RES1008	56	58	158M	male	medium	9.600	medium	n
10	RES1009	75	94	147M	none	8.100	medium	Y	
11	RES1010	61	72	178M	high	7.400	low	Y	
12	RES1011	40	95	174M	high	6.400	medium	n	
13	RES1012	28	61	164M	male	medium	10.100	medium	n
14	RES1013	28	61	160M	male	low	6.600	medium	n
15	RES1014	41	61	153M	male	medium	6.000	medium	n
16	RES1015	70	61	157M	male	low	6.100	medium	Y
17	RES1016	53	61	146M	male	low	7.000	medium	n
18	RES1017	57	61	151M	male	medium	7.000	low	n
19	RES1018	41	61	192M	high	8.500	medium	Y	
20	RES1019	20	61	165M	none	8.100	high	B	

OK

Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM SPSS Statistics Python Spark

Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals Sim Fit Sim Eval Extension Output KDE Simulation EVALUATE

Server: Local Server 143MB / 166MB

Step 5: Correcting Gender Inconsistencies

- Connect a **Reclassify Node** to the output stream.
- Map inconsistent entries (e.g., “M”, “Male”, “F”, “Female”, Etc.) into standardized categories (“male”, “female”).
- Apply changes and review results using a **Table Node**.

StreamIt! - IBM SPSS Modeler

gender inconsistency resolved

Settings Annotations

Mode: Single Multiple
Redeclassify into: New field Existing field

Redeclassify field: gender

New field name: Redeclassify0

Redeclassify values:

Original value	New value
F	female
FEMALE	female
Female	female
M	male
MALE	male

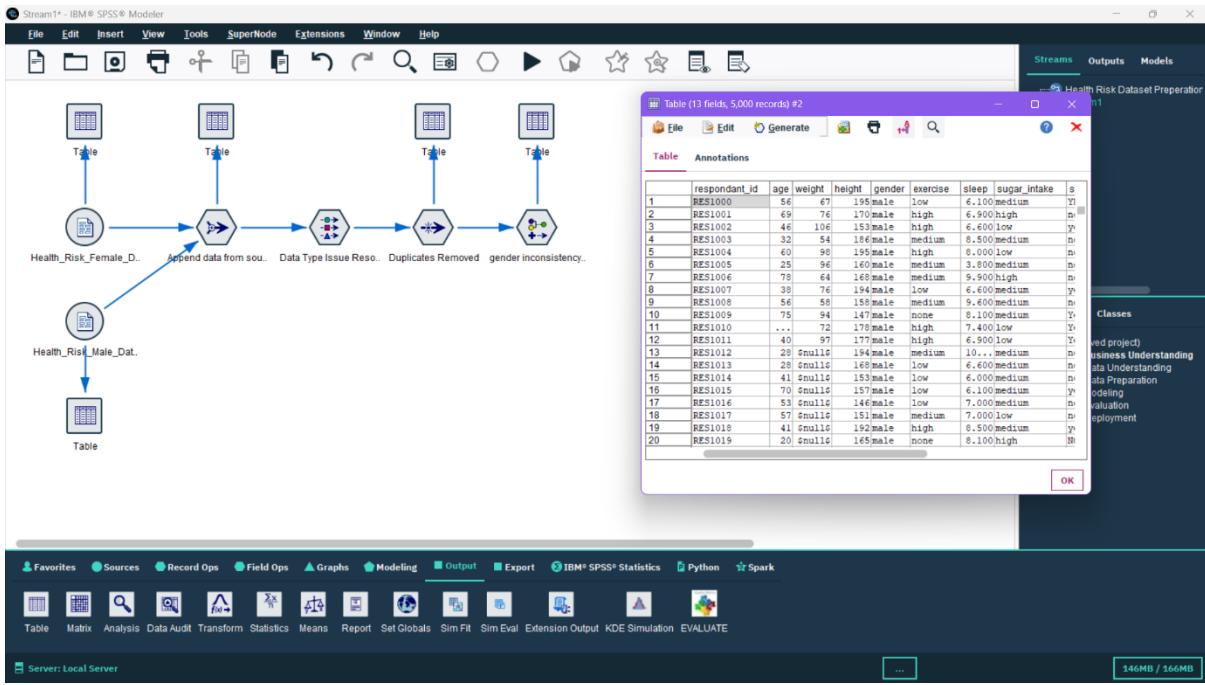
For unspecified values use: Original value Default value

OK Cancel Apply Reset

Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM SPSS Statistics Python Spark

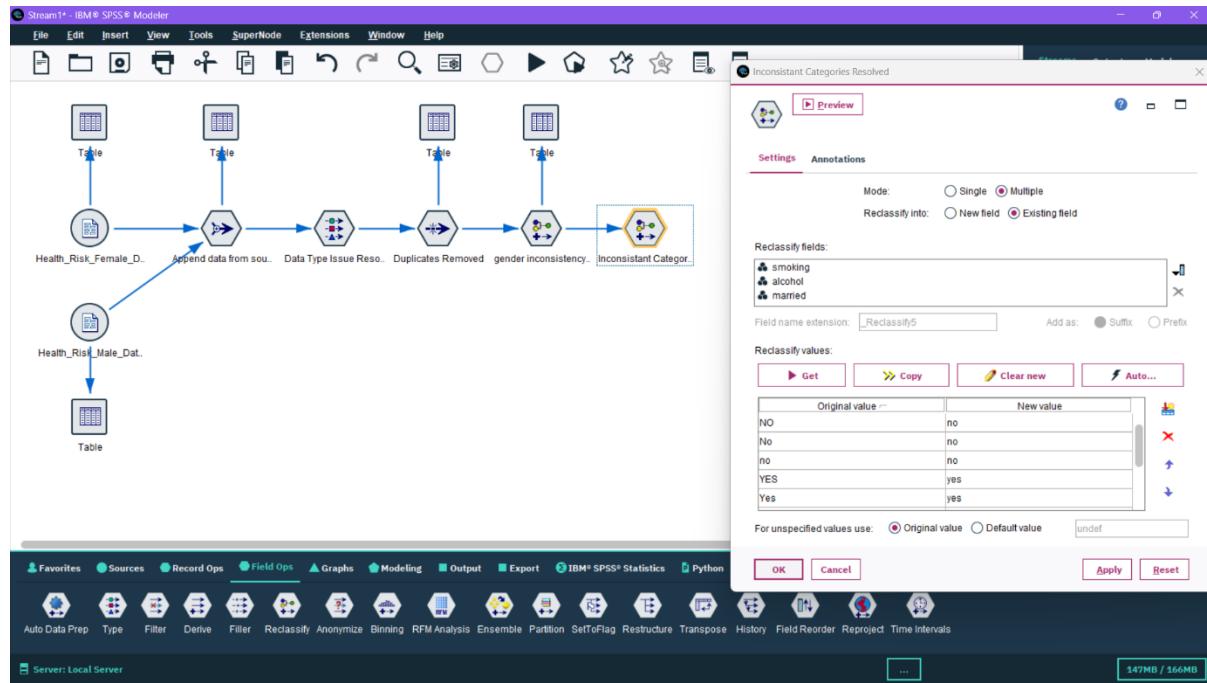
Auto Data Prep Type Filter Derive Filter Reclassify Anonymize Binning RFM Analysis Ensemble Partition SetToFlag Restructure Transpose History Field Reorder Reproj Time Intervals

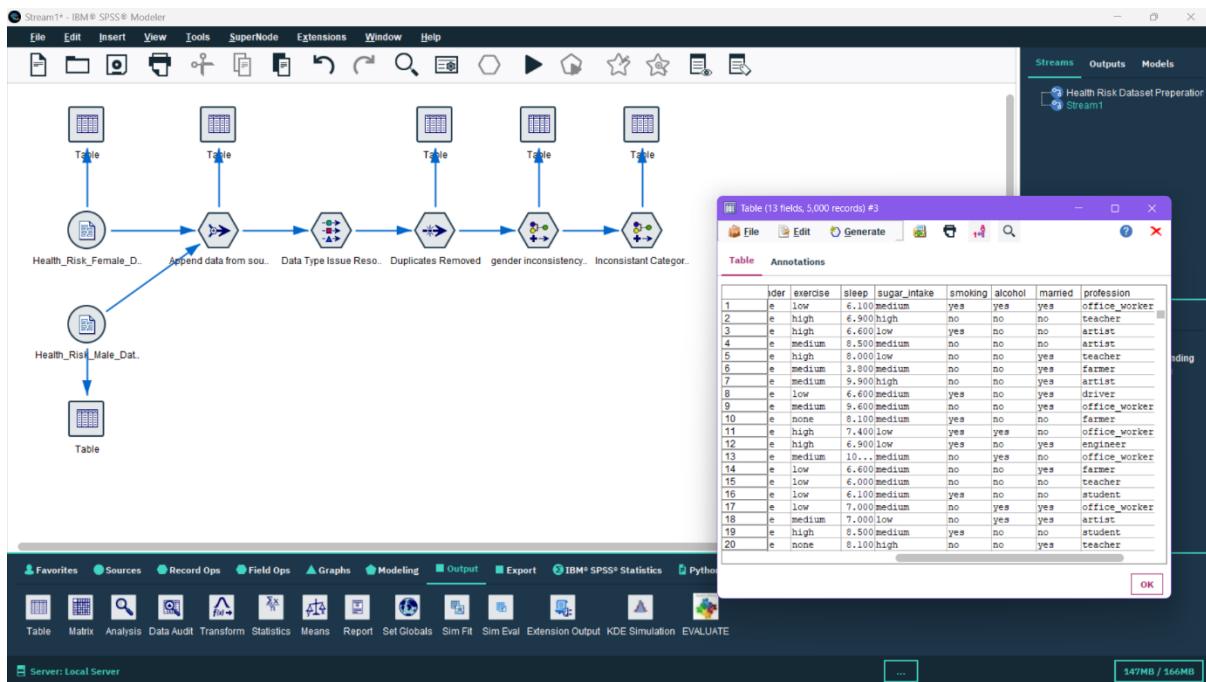
Server: Local Server 144MB / 166MB



Step 6: Standardizing Smoking, Alcohol, and Marital Status Fields

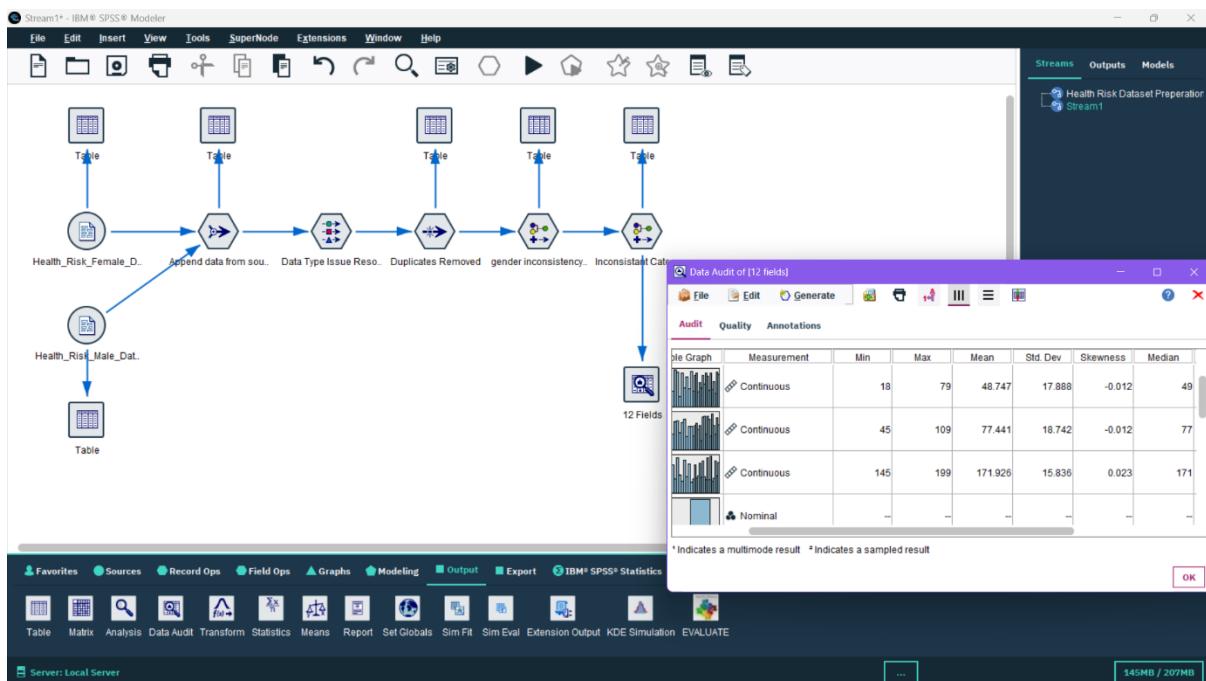
- Add another **Reclassify Node** to clean inconsistent categories for *Smoking*, *Alcohol*, and *Married* columns.
- Standardize values (e.g., “Y/N,” “Yes/No,” “YES/NO”) into a uniform format (“yes”/“no”).
- Apply, then verify with a **Table Node**.





Step 7: Filling Missing Numeric Values Using Median Filler

- Add a **Data Audit Node** to view important information about data fields.
- Add **Median Filler Nodes** for *Age*, *Weight*, and *Height* fields.
- Configure each to replace missing values with median values.
- Validate results through **Table Nodes**.



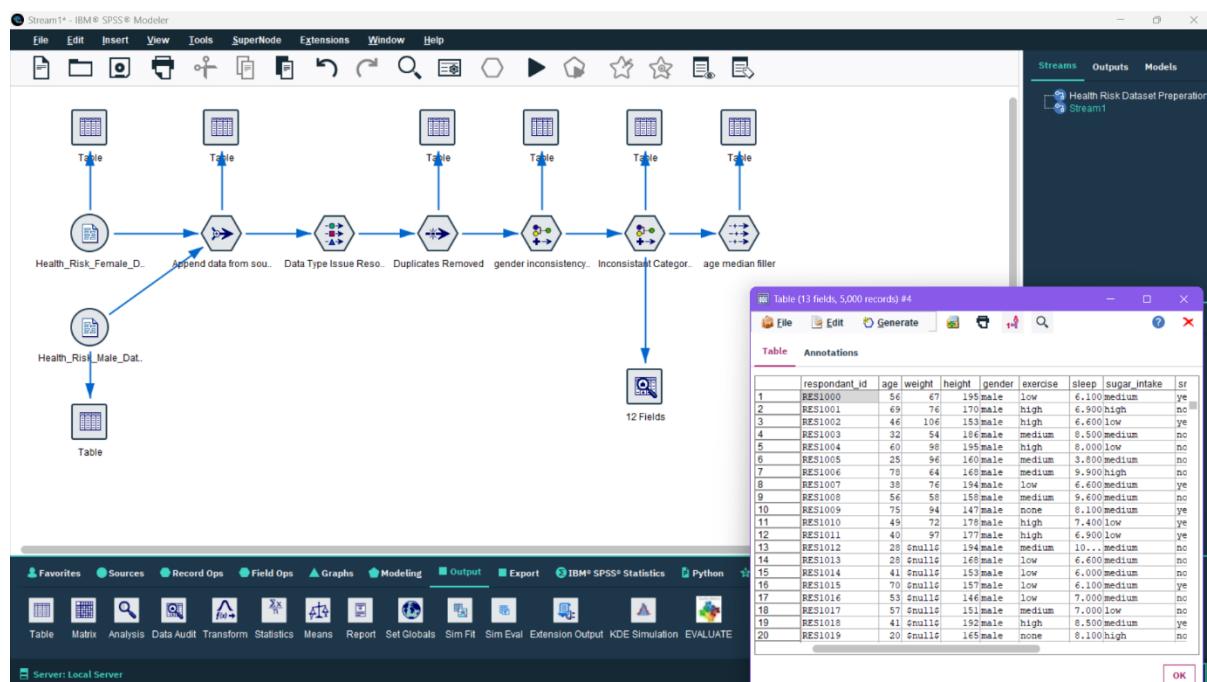
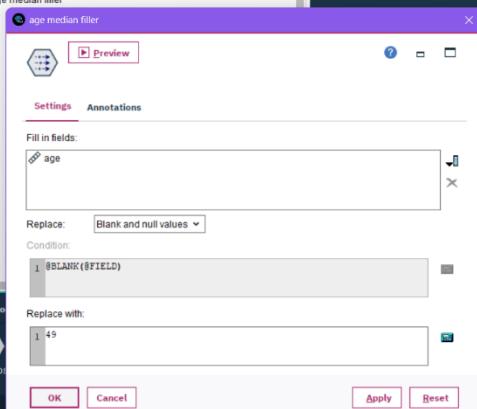
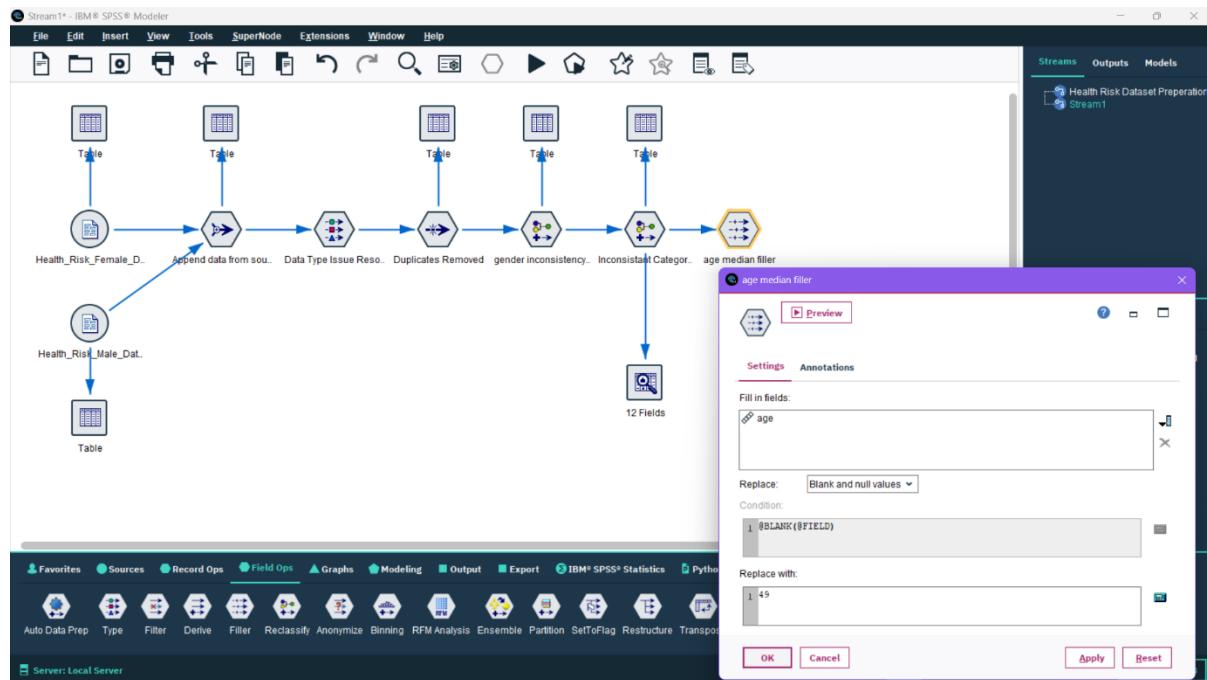
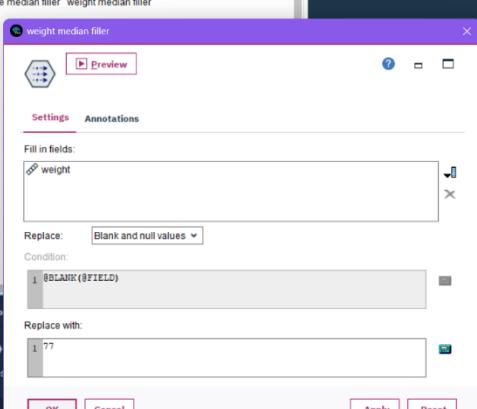
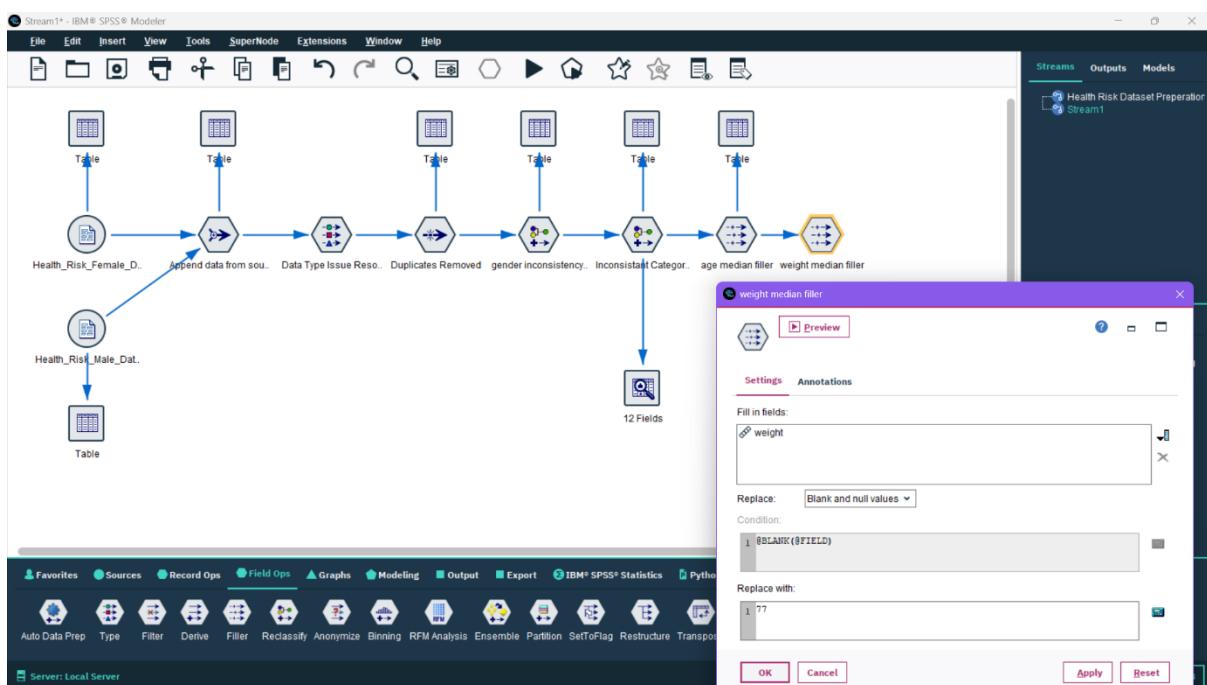


Table (13 fields, 5,000 records) #4

	respondent_id	age	weight	height	gender	exercise	sleep	sugar_intake	sr
1	RES1000	56	67	195	male	low	6.100	medium	ye
2	RES1001	69	76	170	male	high	6.900	high	no
3	RES1002	46	106	153	male	high	6.600	low	ye
4	RES1003	32	54	186	male	medium	8.500	medium	no
5	RES1004	60	98	195	male	high	8.000	low	no
6	RES1005	25	95	160	male	medium	3.800	medium	no
7	RES1006	79	74	165	male	medium	9.900	high	no
8	RES1007	30	76	194	male	low	6.400	medium	ye
9	RES1008	56	58	155	male	medium	9.600	medium	no
10	RES1009	75	94	147	male	none	8.100	medium	ye
11	RES1010	49	72	178	male	high	7.400	low	ye
12	RES1011	40	97	177	male	high	6.900	low	ye
13	RES1012	28	#null#	194	male	medium	10...	medium	no
14	RES1013	28	#null#	165	male	low	6.600	medium	no
15	RES1014	41	#null#	153	male	low	6.000	medium	no
16	RES1015	70	#null#	157	male	low	6.100	medium	ye
17	RES1016	53	#null#	144	male	low	7.000	medium	no
18	RES1017	57	#null#	151	male	medium	7.000	low	no
19	RES1018	41	#null#	192	male	high	8.500	medium	ye
20	RES1019	20	#null#	165	male	none	8.100	high	no

OK



Stream1 - IBM® SPSS® Modeler

Table (13 fields, 5,000 records) #5

respondent_id	age	weight	height	gender	exercise	sleep	sugar_intake	smoker
RESL000	56	67	195	male	low	6.100	medium	ye
RESL001	69	76	170	male	high	6.400	high	no
RESL002	46	106	153	male	high	6.400	low	ye
RESL003	32	54	186	male	medium	8.500	medium	no
RESL004	60	98	195	male	high	8.000	low	no
RESL005	25	96	160	male	medium	3.800	medium	no
RESL006	78	64	169	male	medium	9.900	high	no
RESL007	38	76	194	male	low	6.600	medium	ye
RESL008	56	58	159	male	medium	9.600	medium	no
RESL009	75	94	147	male	none	8.100	medium	ye
RESL010	49	72	179	male	high	7.400	low	ye
RESL011	40	97	177	male	high	6.800	low	no
RESL012	28	77	194	male	medium	10...	medium	no
RESL013	28	77	169	male	low	6.400	medium	no
RESL014	41	77	153	male	low	6.000	medium	no
RESL015	70	77	157	male	low	6.100	medium	no
RESL016	53	77	146	male	low	7.000	medium	no
RESL017	57	77	151	male	medium	7.000	low	no
RESL018	41	77	192	male	high	8.500	medium	ye
RESL019	20	77	165	male	none	8.100	high	no

Stream1 - IBM® SPSS® Modeler

height median filter

Settings Annotations

Fill in fields:

Replace: Blank and null values

Condition:

Replace with:

OK Cancel Apply Reset

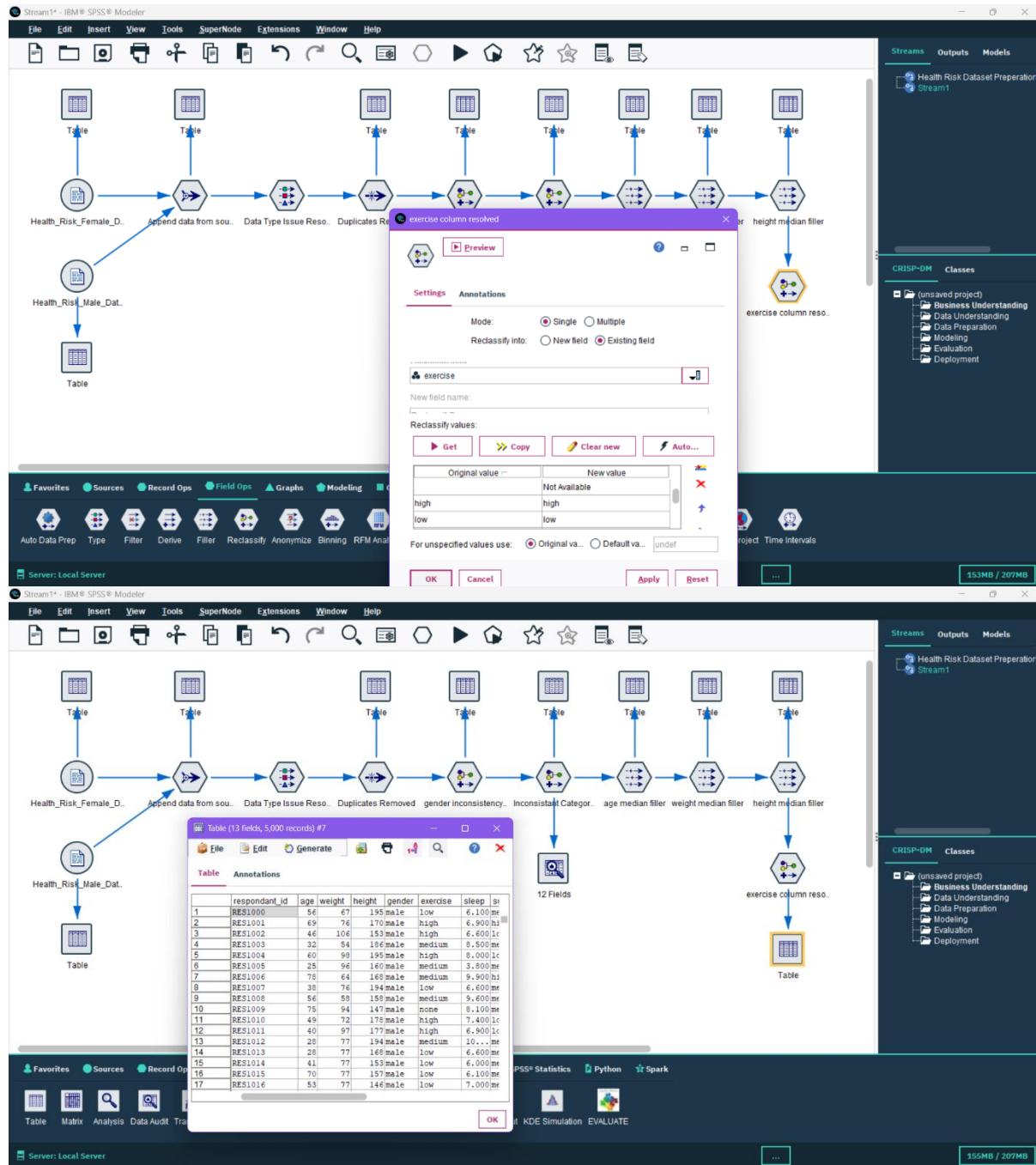
Stream1 - IBM® SPSS® Modeler

Table (13 fields, 5,000 records) #6

respondent_id	age	weight	height	gender	exercise	sleep	sugar_intake	smoker
ESL000	56	67	195	male	low	6.100	medium	ye
ESL001	69	76	170	male	high	6.400	high	no
ESL002	46	106	153	male	high	6.400	low	ye
ESL003	32	54	186	male	medium	8.500	medium	no
ESL004	60	98	195	male	high	8.000	low	no
ESL005	25	96	160	male	medium	3.800	medium	no
ESL006	78	64	169	male	medium	9.900	high	no
ESL007	38	76	194	male	low	6.600	medium	ye
ESL008	56	58	159	male	medium	9.600	medium	no
ESL009	75	94	147	male	none	8.100	medium	ye
ESL010	49	72	179	male	high	7.400	low	ye
ESL011	40	97	177	male	high	6.800	low	no
ESL012	28	77	194	male	medium	10...	medium	no
ESL013	28	77	169	male	low	6.400	medium	no
ESL014	41	77	153	male	low	6.000	medium	no
ESL015	70	77	157	male	low	6.100	medium	no
ESL016	53	77	146	male	low	7.000	medium	no
ESL017	57	77	151	male	medium	7.000	low	no
ESL018	41	77	192	male	high	8.500	medium	ye
ESL019	20	77	165	male	none	8.100	high	no

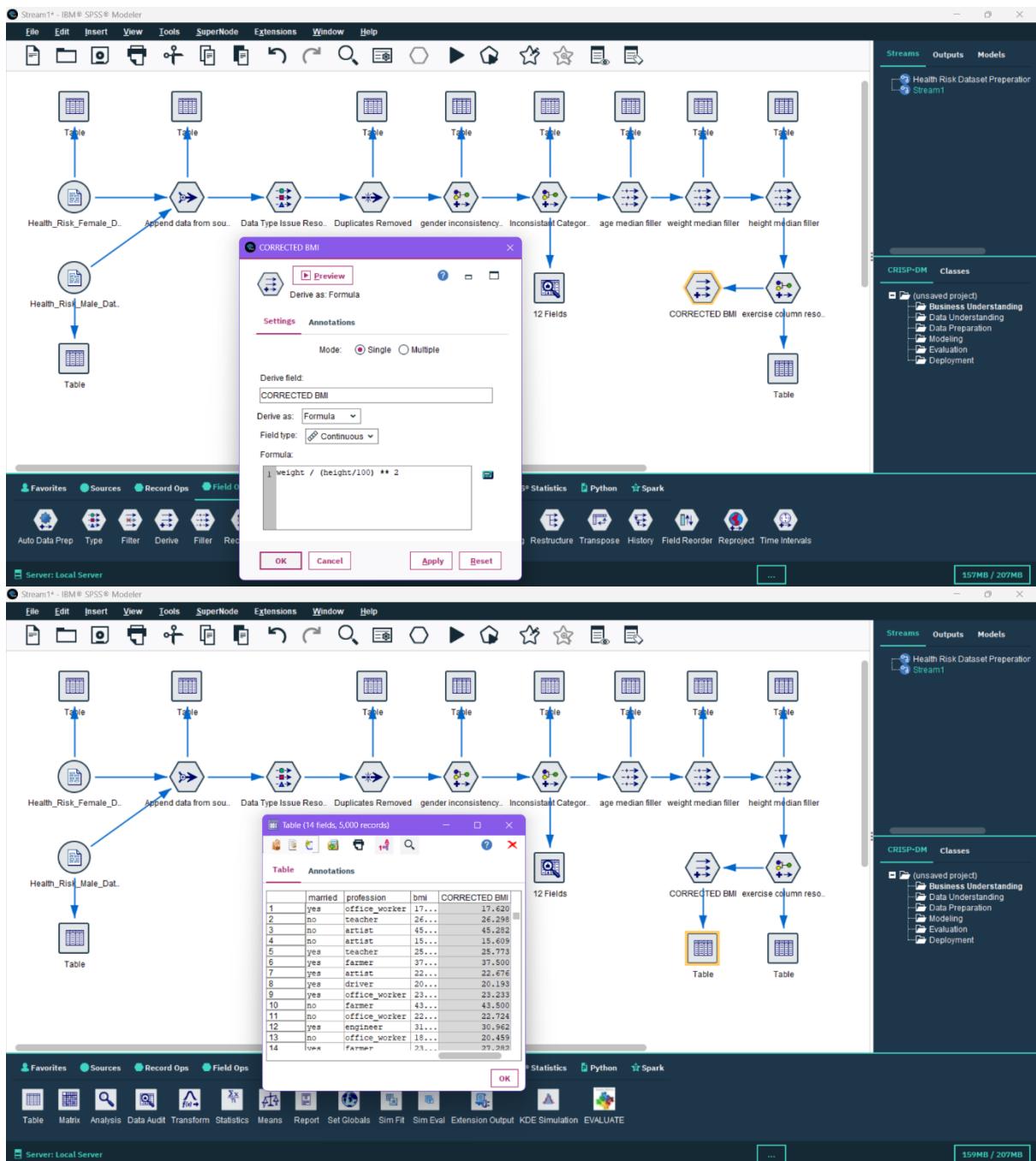
Step 8: Handling Missing Exercise Data

- Add a **Reclassify Node** for the *Exercise* field.
- Replace all blank or missing values with “Not Available.”
- Apply and validate results using a **Table Node**.



Step 9: Deriving Corrected BMI Field

- Add a **Derive Node** and name it *Corrected BMI*.
- Under Derive Type, select *Formula* and use a CLEM expression such as: $weight / (height/100) ^\star 2$
- Run the node and verify the output through a **Table Node**.



Step 10: Validating the Final Dataset

- Check all transformation nodes sequentially with connected **Table Nodes** to ensure completeness, accuracy, and consistency before exporting the final cleaned dataset.