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.**

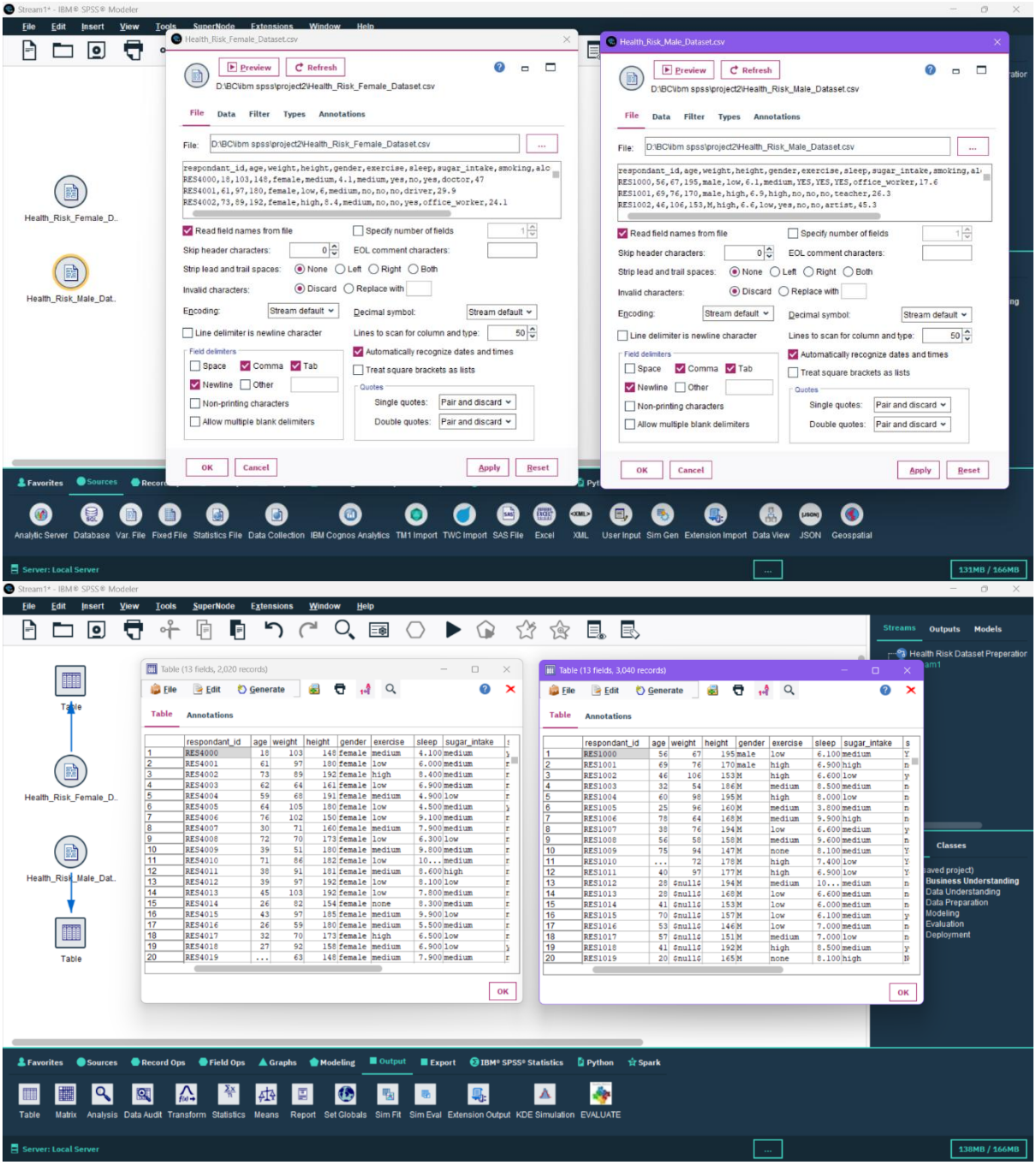
Next, **Reclassify Nodes** are used to resolve categorical inconsistencies in the *Gender, Smoking, Alcohol,* and *Marital Status* fields, ensuring uniform labelling.
Missing numeric values in *Age, Weight,* and *Height* are filled using **Median Filler Nodes** to ensure balanced data distribution. The **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:

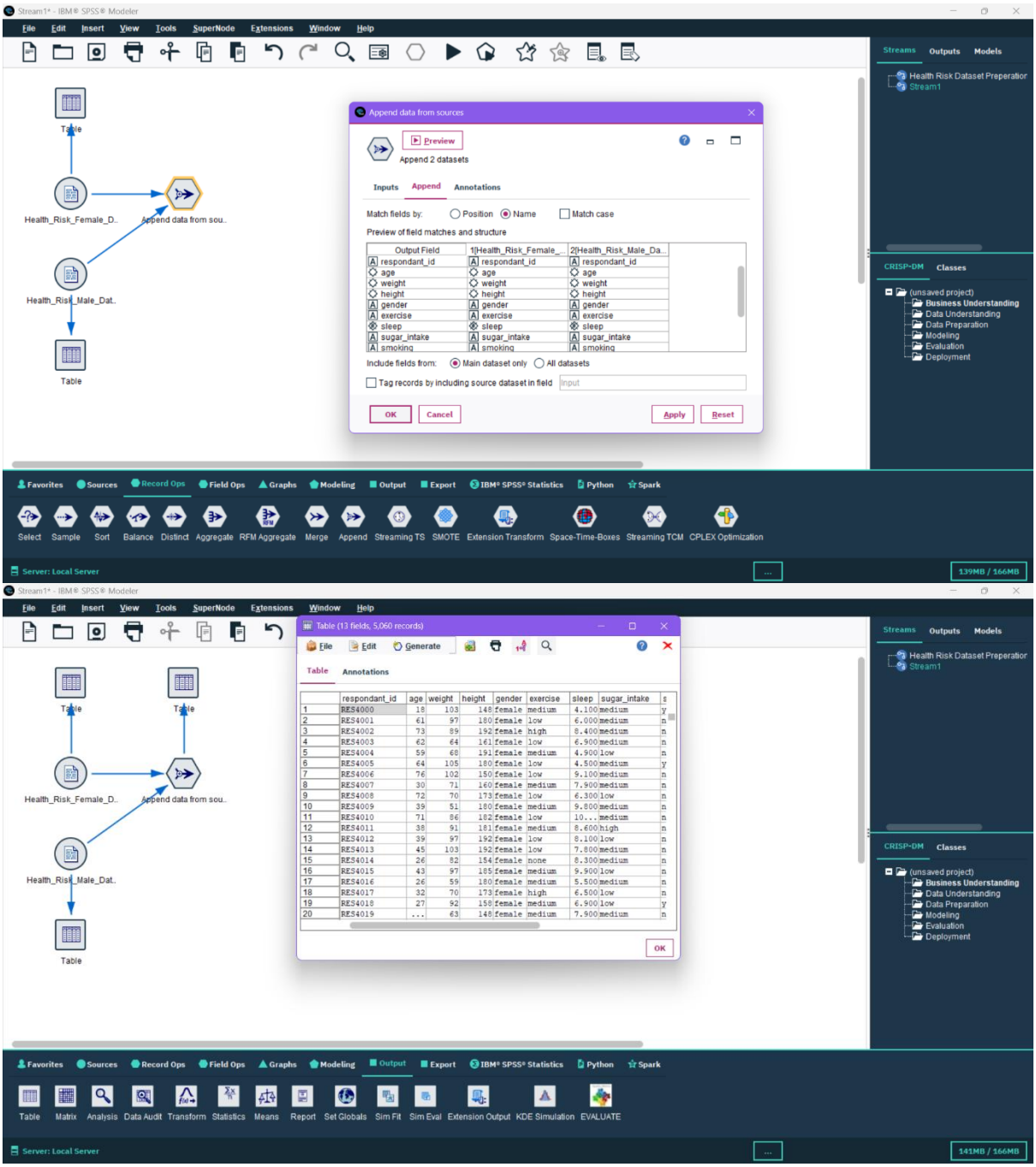
Step 1: Importing and Viewing Datasets

- 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.



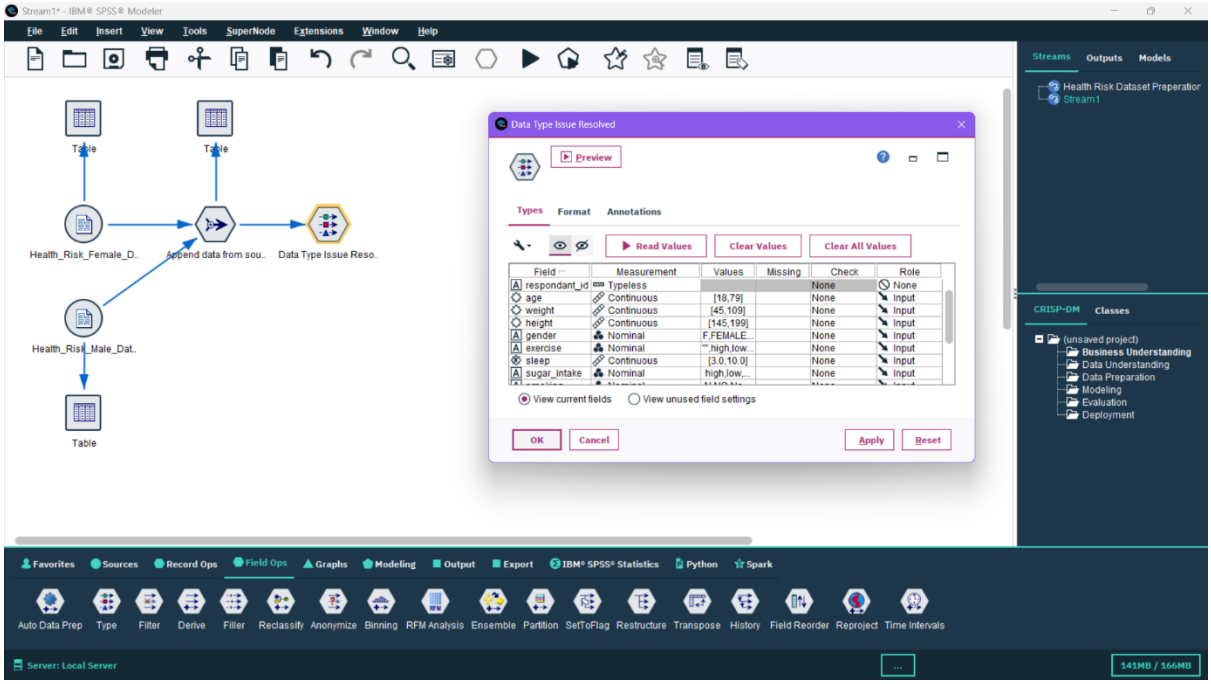
Step 2: Merging Datasets Using Append Node

- 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.



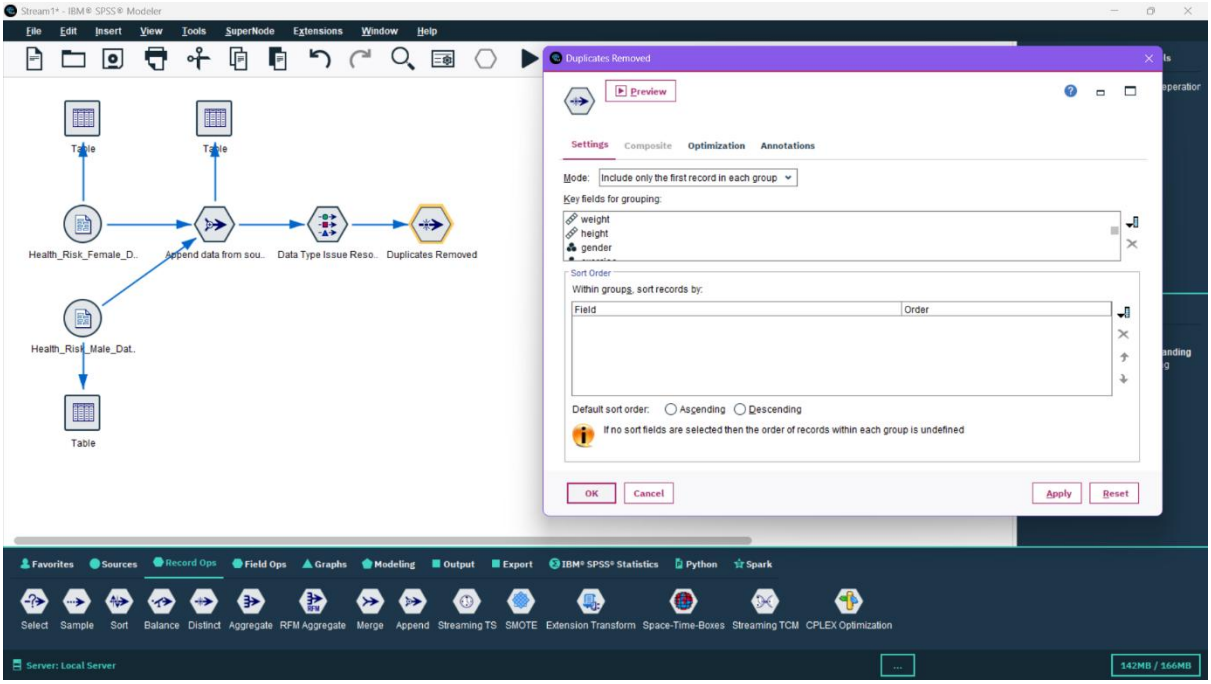
Step 3: Resolving Data Type Issues

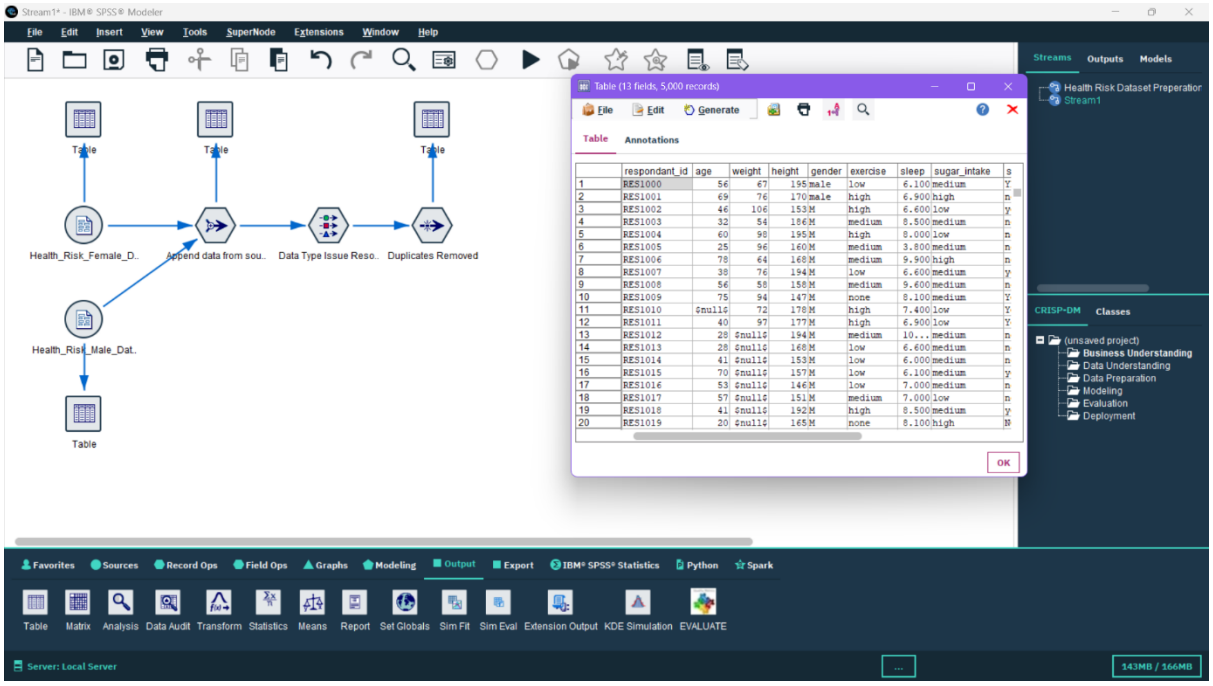
- Connect a **Type Node** to the Append Node.
- Click **Read Values** and update incorrect data types or measurement levels.
- Apply and run the node to standardize field definitions.



Step 4: Removing Duplicate Records

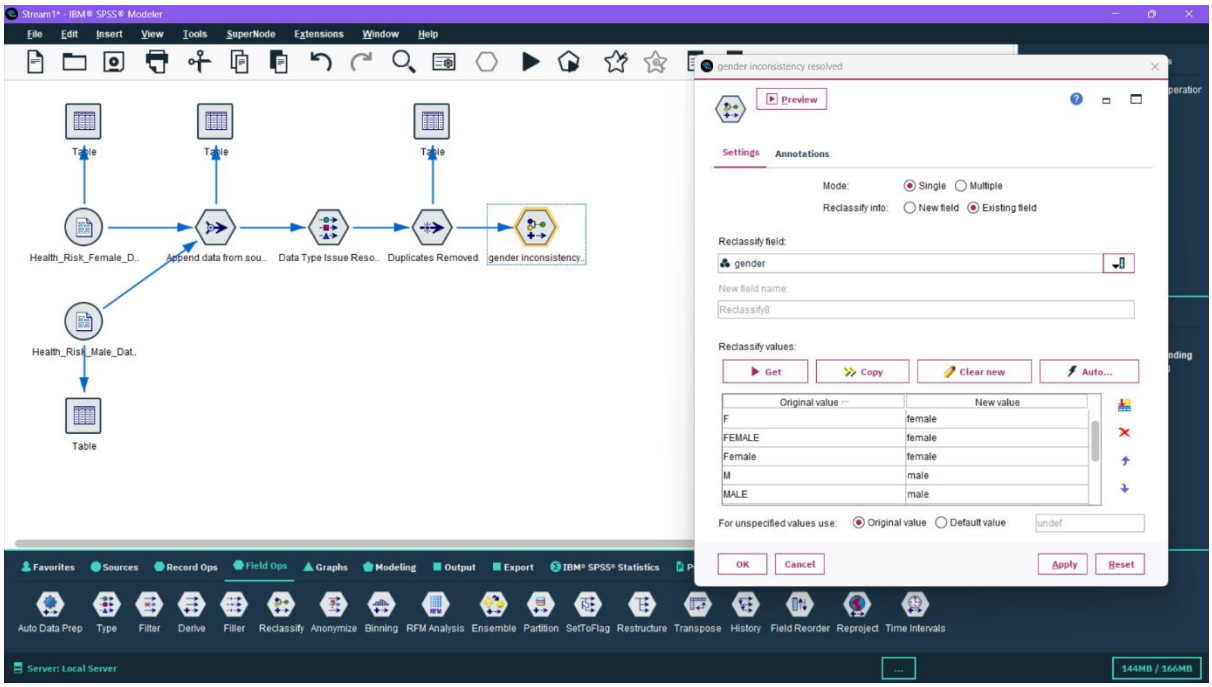
- Add a **Remove Duplicates Node** to eliminate repeated entries.
- Select key identifiers such as *Patient ID* or *Name.*
- Run the node and verify the results through a **Table Node.**

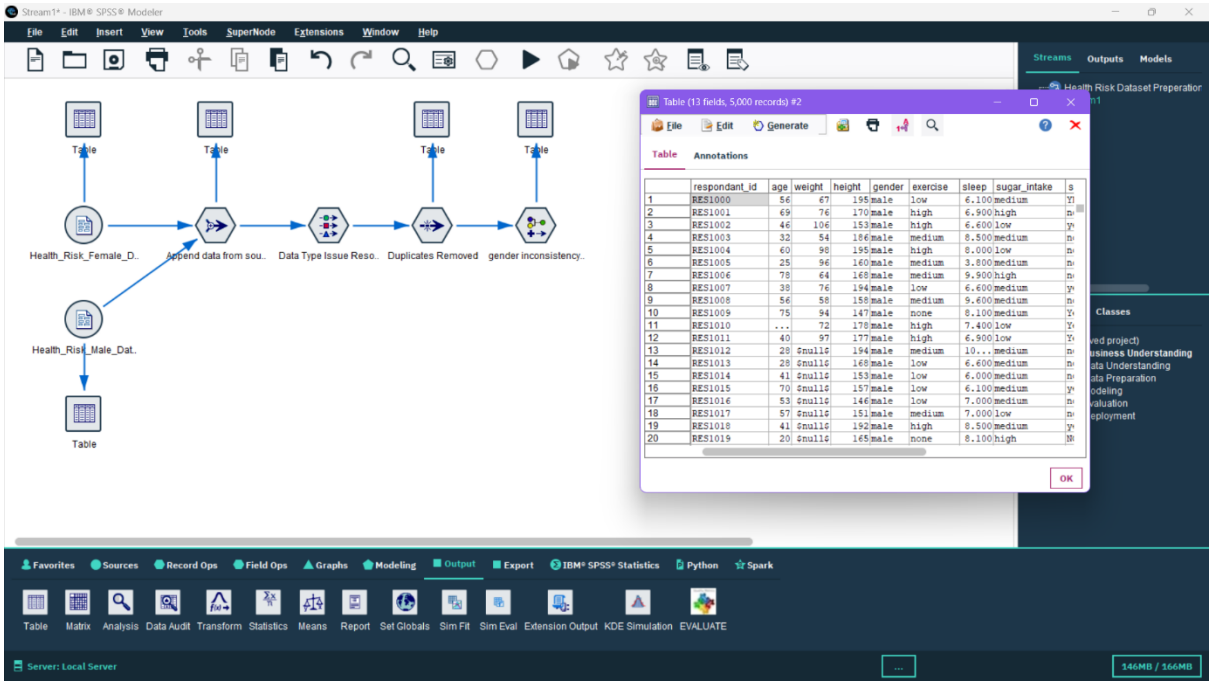




Step 5: Correcting Gender Inconsistencies

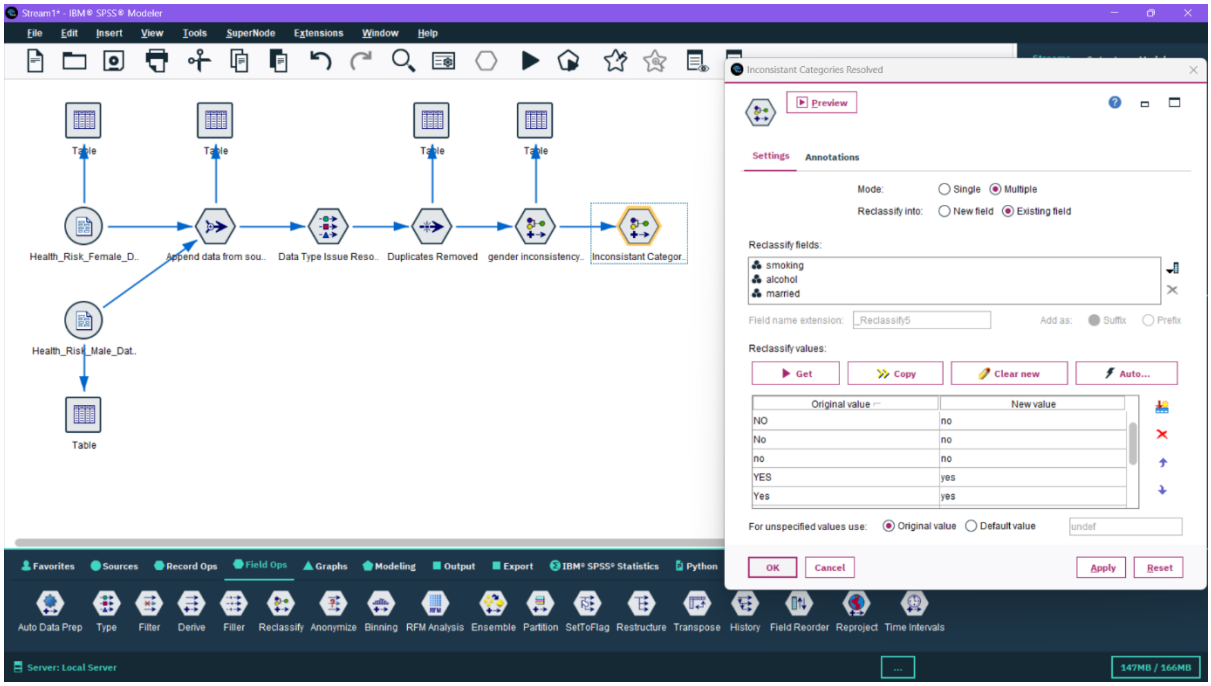
- Connect a **Reclassify Node** to the output stream.
- Map inconsistent entries (e.g., “M”, “Male”, “F”, “Female”, Etc.) into standardized categories (“male”,” female”).
- Apply changes and review results using a **Table Node.**

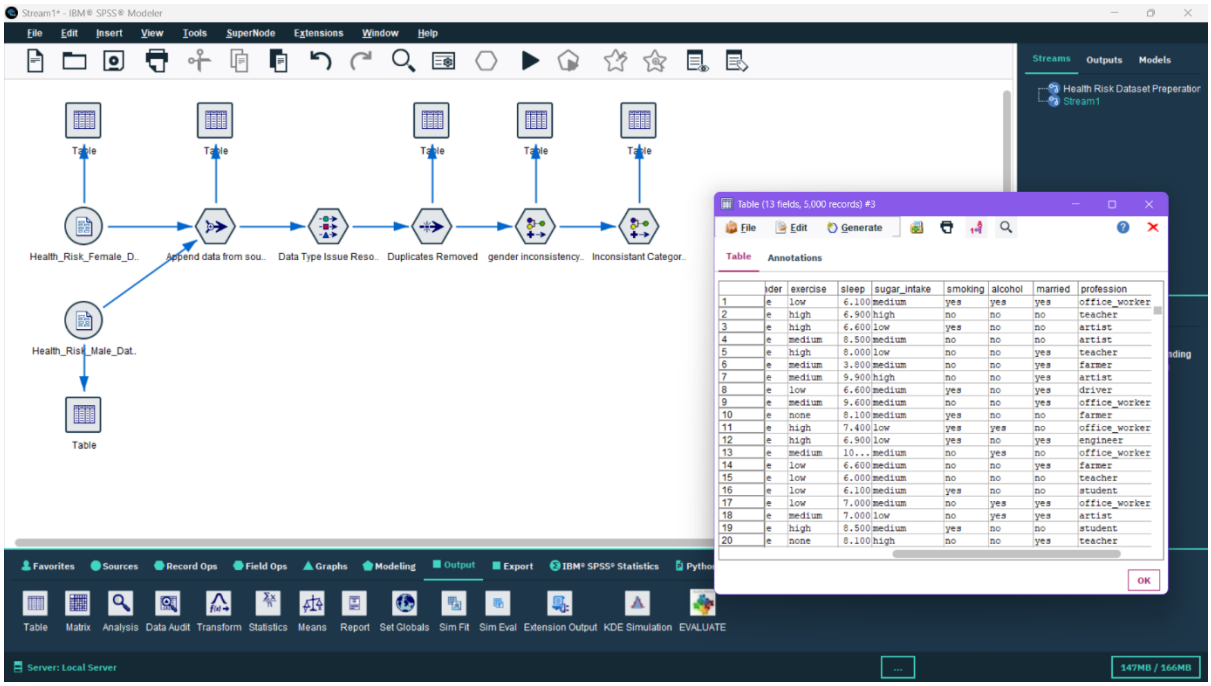




Step 6: Standardizing Smoking, Alcohol, and Marital Status Fields

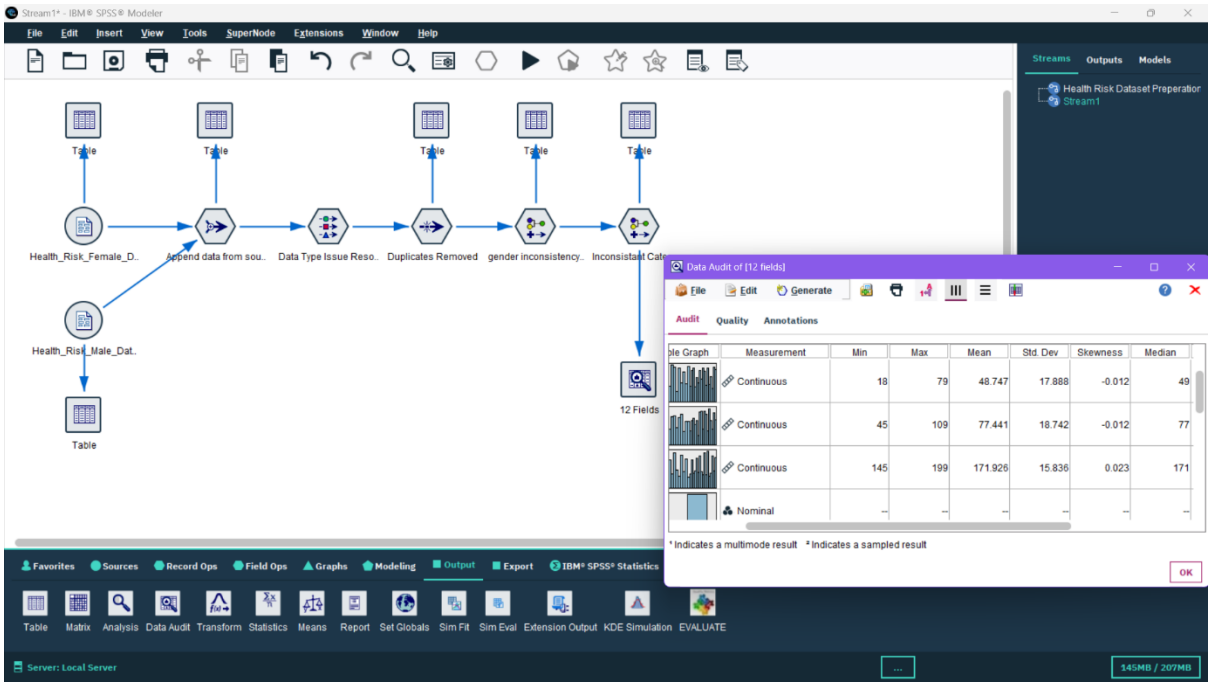
- Add another **Reclassify Node** to clean inconsistent categories for *Smoking, Alcohol,* and *Married* columns.
- Standardize values (e.g., “Y/N,” “Yes/No,” “YES/NO”) into a uniform format (“yes”/”no”).
- Apply, then verify with a **Table Node.**

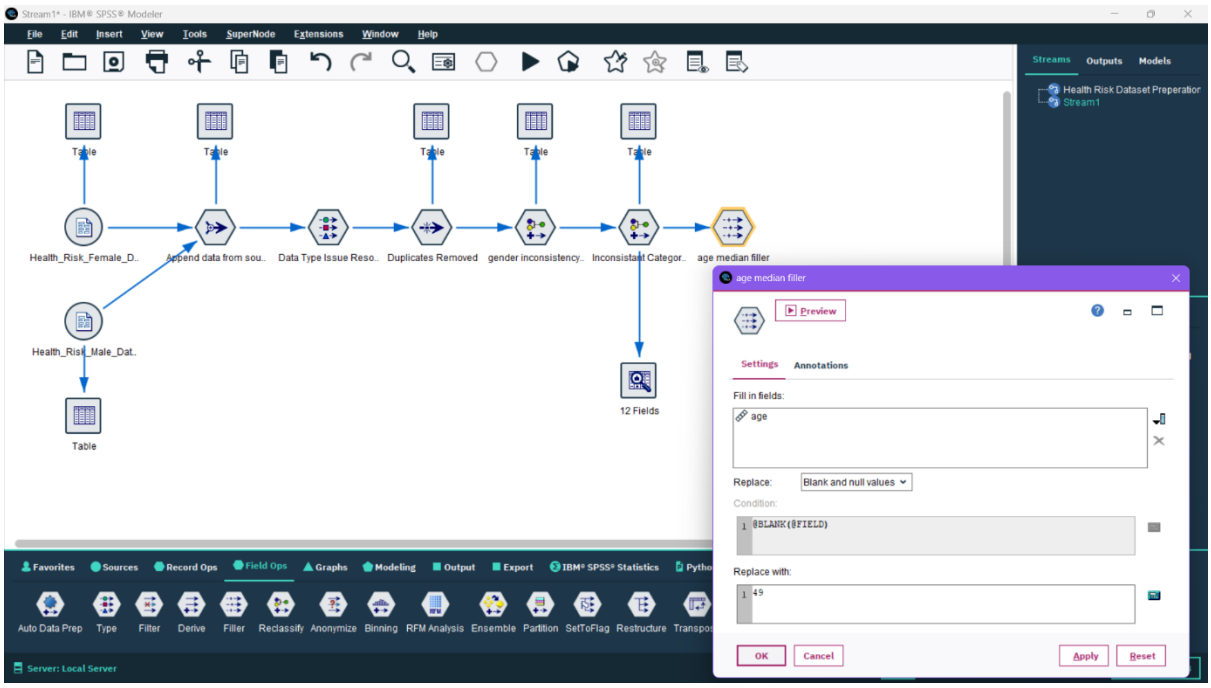


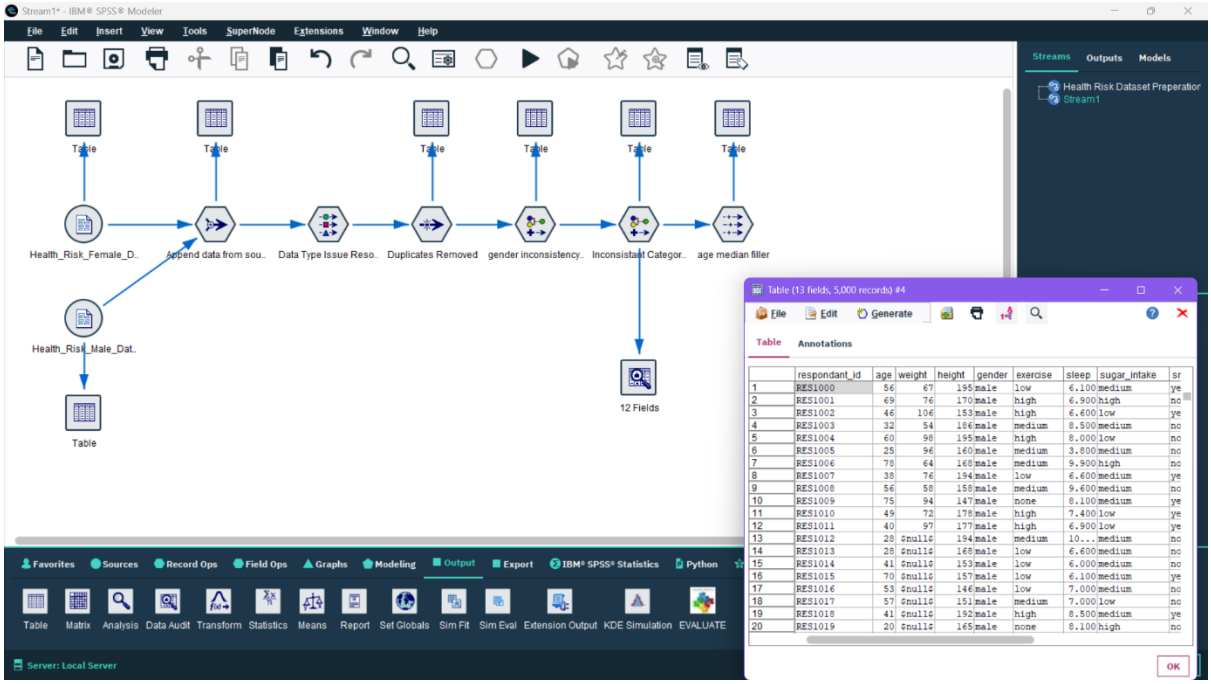


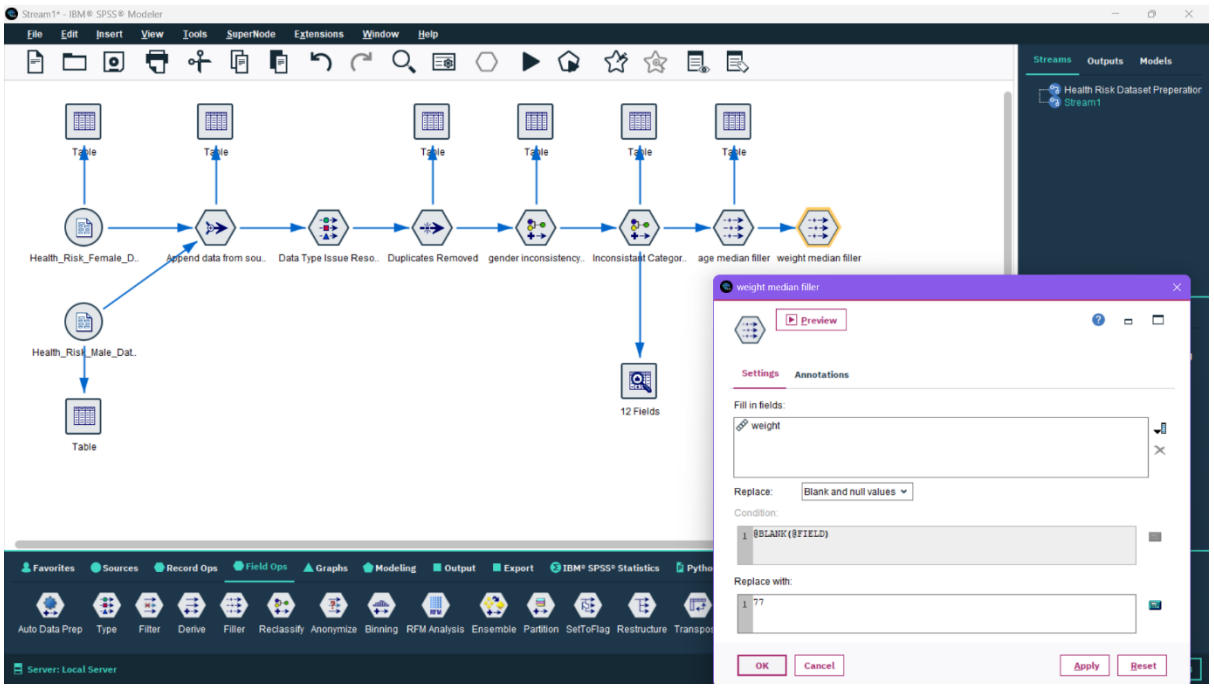
Step 7: Filling Missing Numeric Values Using Median Filler

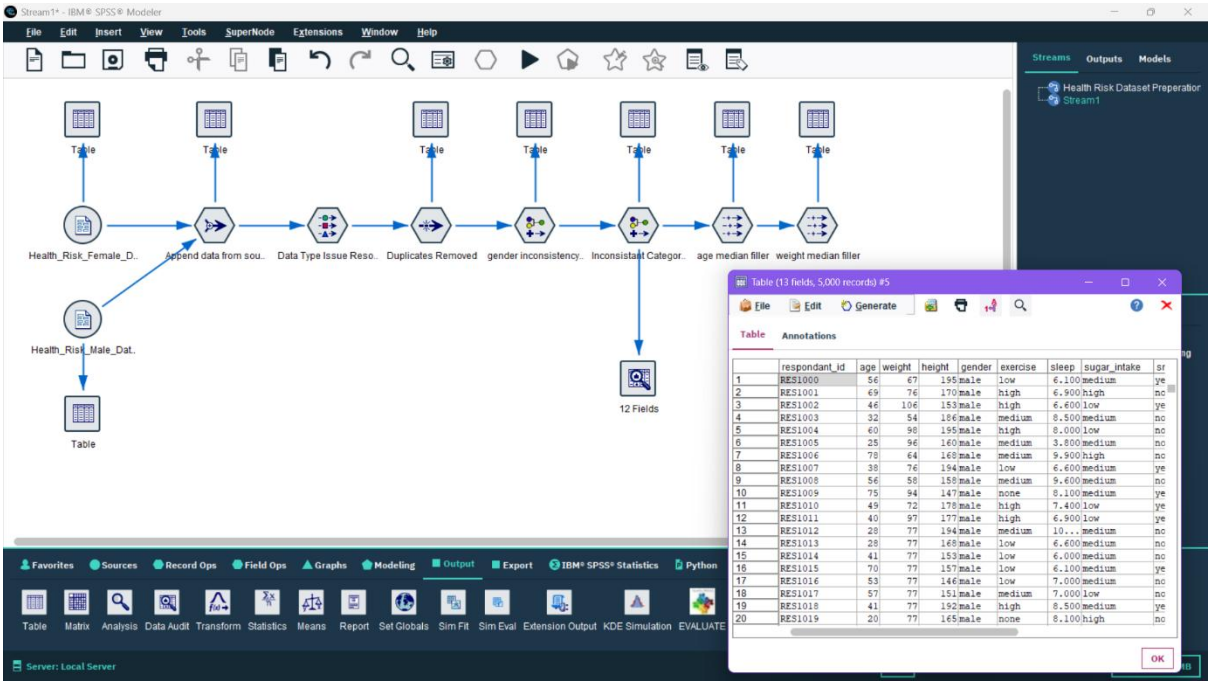
- Add a **Data Audit Node** to view important information about data fields.
- Add **Median Filler Nodes** for *Age, Weight,* and *Height* fields.
- Configure each to replace missing values with median values.
- Validate results through **Table Nodes.**

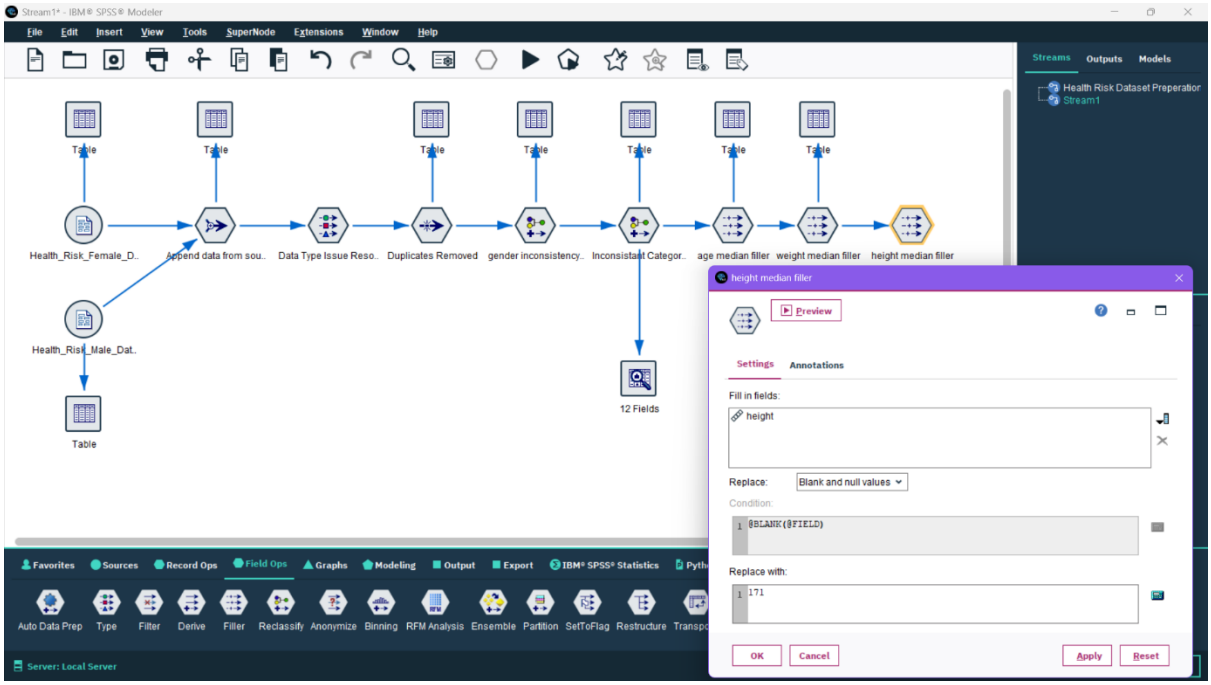


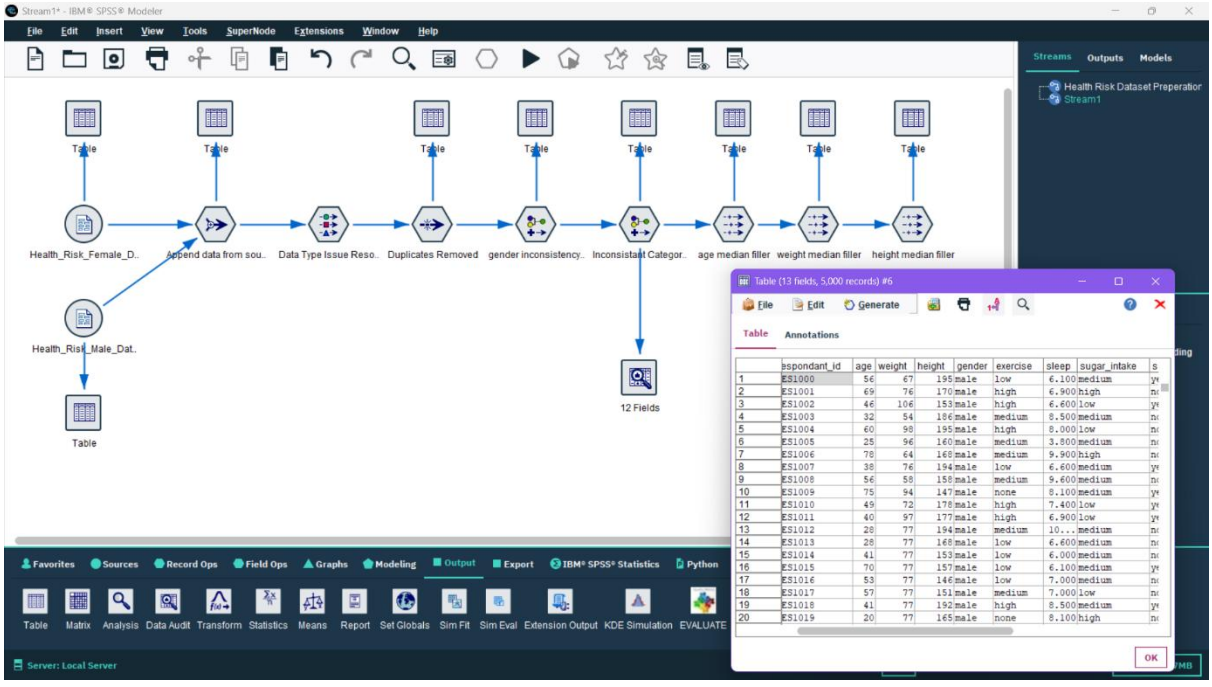

Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
Health_Risk_Male_Dat..
12 Fields
Table
age median filler
Preview
Settings Annotations
Fill in fields:
age
Replace:
Blank and null values
Condition:
@BLANK(@FIELD)
Replace with:
49
OK Cancel Apply Reset
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python
Auto Data Prep Type Filter Derive Filler Reclassify Anonymize Binning RFM Analysis Ensemble Partition SetToFlag Restructure Transpos
Server: Local Server


Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
Health_Risk_Male_Dat..
12 Fields
Table
Table (13 fields, 5,000 records) #4
File Edit Generate
Table Annotations
respondant_id age weight height gender exercise sleep sugar_intake sr
RES1000 56 67 195 male low 6.100 medium ye
RES1001 69 76 170 male high 6.900 high no
RES1002 46 106 153 male high 6.600 low ye
RES1003 32 54 186 male medium 8.500 medium no
RES1004 60 98 195 male high 8.000 low no
RES1005 25 96 160 male medium 3.800 medium no
RES1006 78 64 168 male medium 9.900 high no
RES1007 38 76 194 male low 6.600 medium ye
RES1008 56 58 158 male medium 9.600 medium no
RES1009 75 94 147 male none 8.100 medium ye
RES1010 49 72 178 male high 7.400 low ye
RES1011 40 97 177 male high 6.900 low ye
RES1012 28 $null$ 194 male medium 10...medium no
RES1013 28 $null$ 168 male low 6.600 medium no
RES1014 41 $null$ 153 male low 6.000 medium no
RES1015 70 $null$ 157 male low 6.100 medium ye
RES1016 53 $null$ 146 male low 7.000 medium no
RES1017 57 $null$ 151 male medium 7.000 low no
RES1018 41 $null$ 192 male high 8.500 medium ye
RES1019 20 $null$ 165 male none 8.100 high no
OK
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python
Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals Sim Fit Sim Eval Extension Output KDE Simulation EVALUATE
Server: Local Server


Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
weight median filler
Health_Risk_Male_Dat..
12 Fields
Table
weight median filler
Preview
Settings Annotations
Fill in fields:
weight
Replace:
Blank and null values
Condition:
@BLANK(@FIELD)
Replace with:
77
OK Cancel Apply Reset
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python
Auto Data Prep Type Filter Derive Filler Reclassify Anonymize Binning RFM Analysis Ensemble Partition SetToFlag Restructure Transpos
Server: Local Server

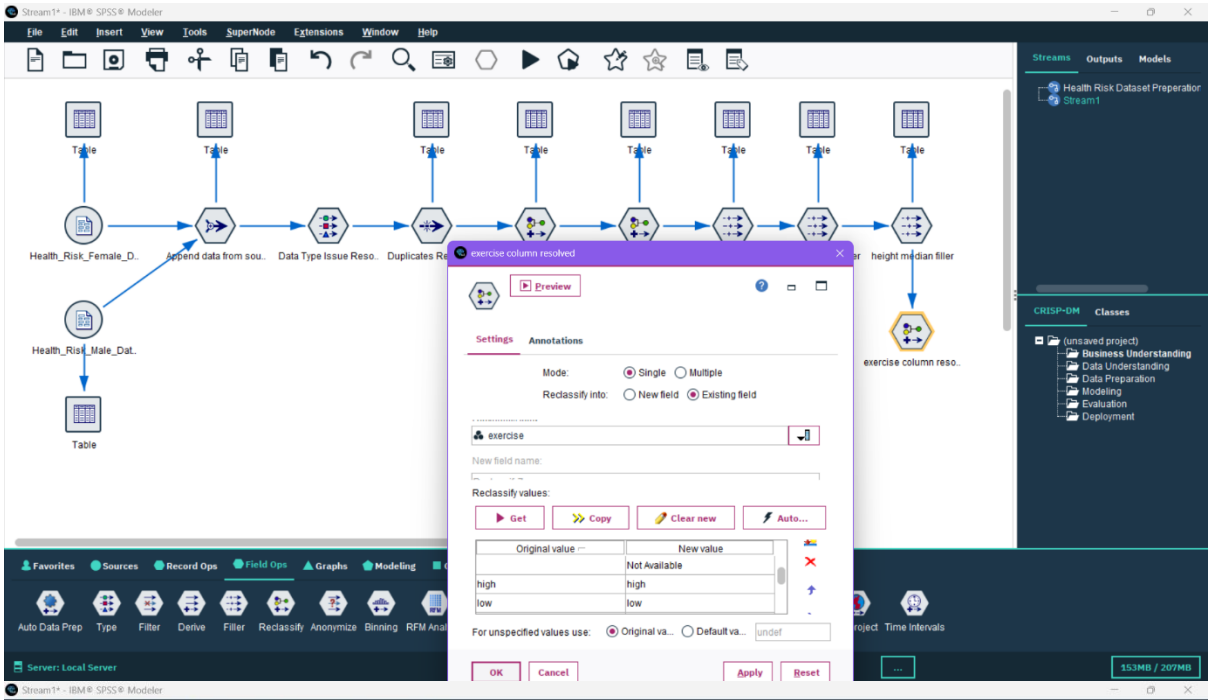

Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Table
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
weight median filler
Health_Risk_Male_Dat..
12 Fields
Table (13 fields, 5,000 records) #5
File Edit Generate
Table Annotations
respondant_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 96 160 male medium 3.800 medium no
7 RES1006 78 64 168 male medium 9.900 high no
8 RES1007 38 76 194 male low 6.600 medium ye
9 RES1008 56 58 158 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 77 194 male medium 10... medium no
14 RES1013 28 77 168 male low 6.600 medium no
15 RES1014 41 77 153 male low 6.000 medium no
16 RES1015 70 77 157 male low 6.100 medium ye
17 RES1016 53 77 146 male low 7.000 medium no
18 RES1017 57 77 151 male medium 7.000 low no
19 RES1018 41 77 192 male high 8.500 medium ye
20 RES1019 20 77 165 male none 8.100 high no
OK
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python
Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals Sim Fit Sim Eval Extension Output KDE Simulation EVALUATE
Server: Local Server

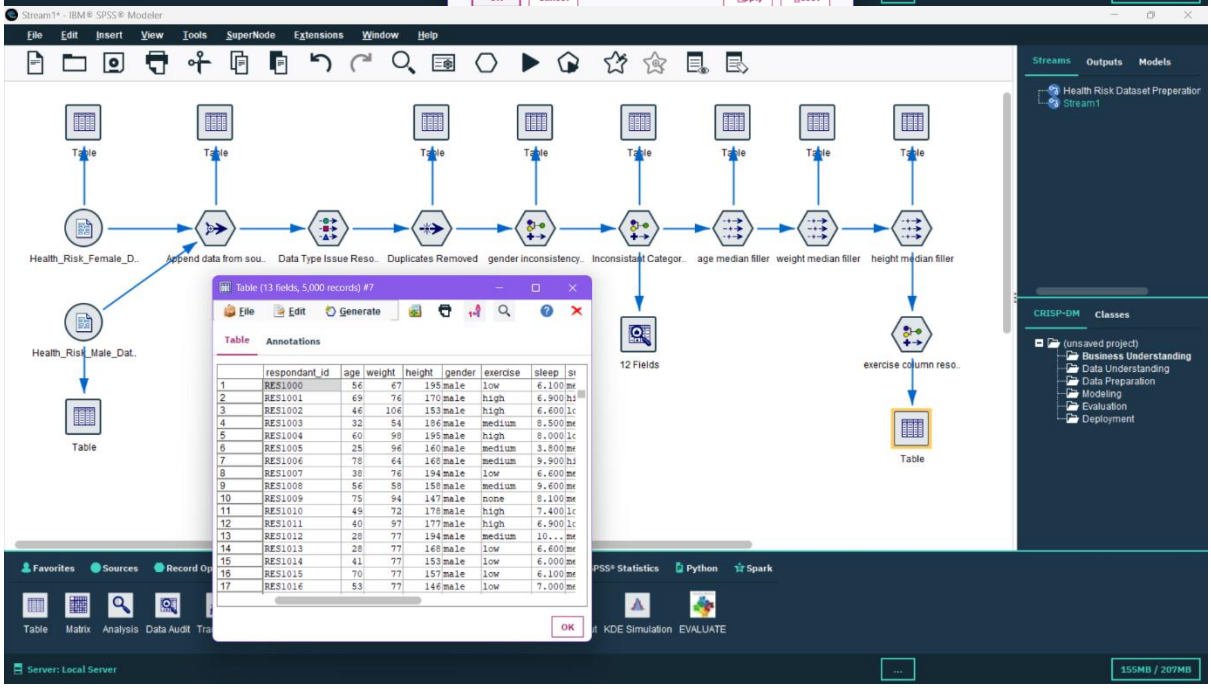

Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Table
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
weight median filler
height median filler
Health_Risk_Male_Dat..
12 Fields
height median filler
Preview
Settings Annotations
Fill in fields:
height
Replace:
Blank and null values
Condition:
1 @BLANK(@FIELD)
Replace with:
1 171
OK Cancel Apply Reset
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics
Auto Data Prep Type Filter Derive Filler Reclassify Anonymize Binning RFM Analysis Ensemble Partition SetToFlag Restructure
Server: Local Server


Stream1* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
Streams Outputs Models
Health Risk Dataset Preperation
Stream1
Table
Health_Risk_Female_D..
Append data from sou..
Data Type Issue Reso..
Duplicates Removed
gender inconsistency..
Inconsistant Categor..
age median filler
weight median filler
height median filler
Health_Risk_Male_Dat..
12 Fields
Table (13 fields, 5,000 records) #6
File Edit Generate
Table Annotations
OK
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python
Table Matrix Analysis Data Audit Transform Statistics Means Report Set Globals Sim Fit Sim Eval Extension Output KDE Simulation EVALUATE
Server: Local Server

Step 8: Handling Missing Exercise Data

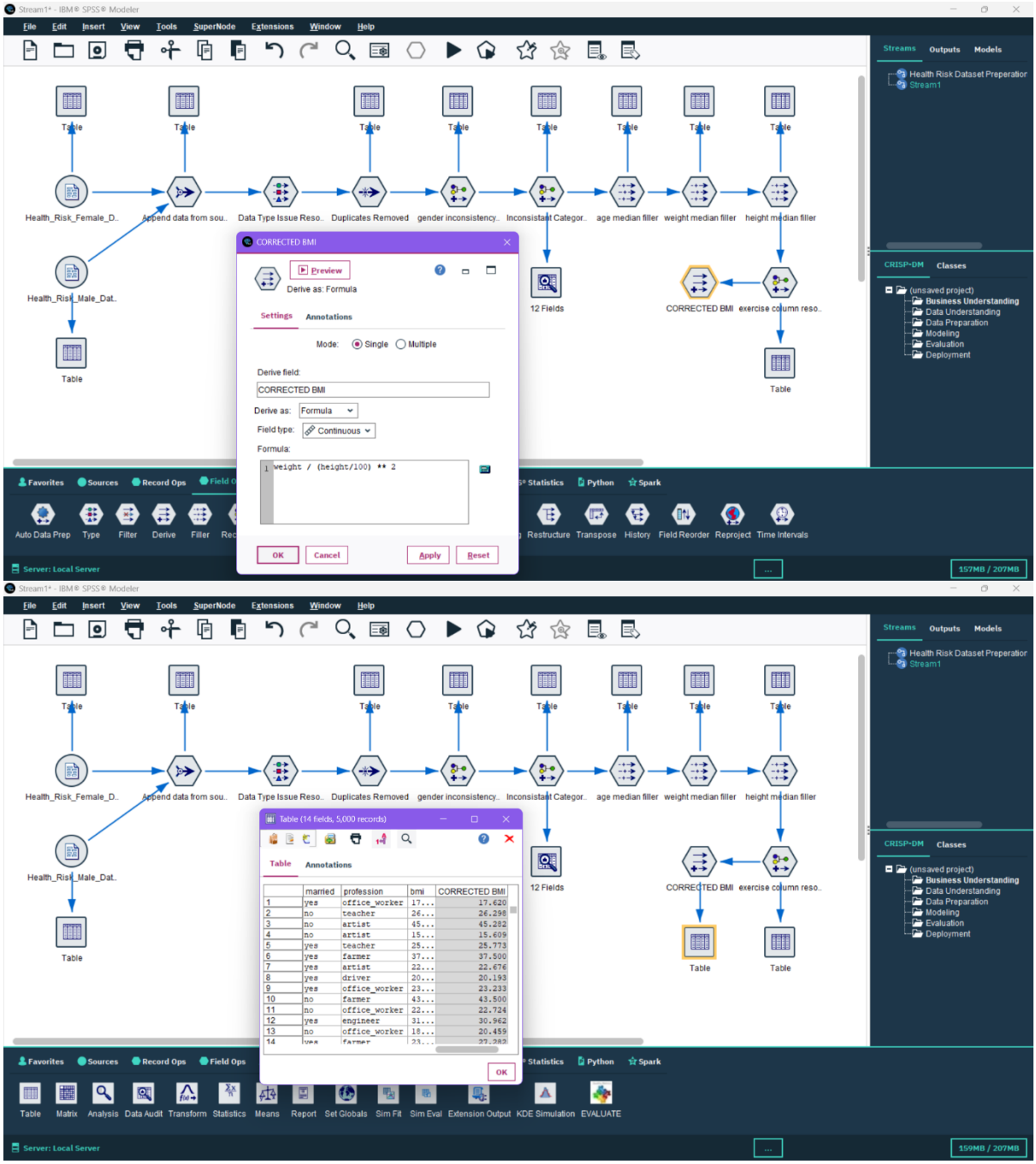
- Add a **Reclassify Node** for the *Exercise* field.
- Replace all blank or missing values with "Not Available."
- Apply and validate results using a **Table Node.**





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) ** 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.