Babu Banarasi Das University

Case Study

on

Comparative Analysis of Categorical and Continuous Fields with Visualization in SPSS Modeler Using Hospital Dataset

SUBMITTED TO:

Mr. Robin T. Falaq

SUBMITTED BY:

Name:

Kalpana G.

Ekta M.

Practical-: Deriving Insights from Hospital Data through Distinct, Aggregate, and Flag Transformation Nodes

Definition:

This practical demonstrates how to analyse and understand the relationships between different data fields in IBM SPSS Modeller. Using nodes such as Distinct, Aggregate, Type, SetToFlag, and visualisation tools, we can explore the dataset to identify unique values, summarise information, classify fields, and examine the relationship between continuous variables, such as Age and Billing Amount.

Outcome / Learning

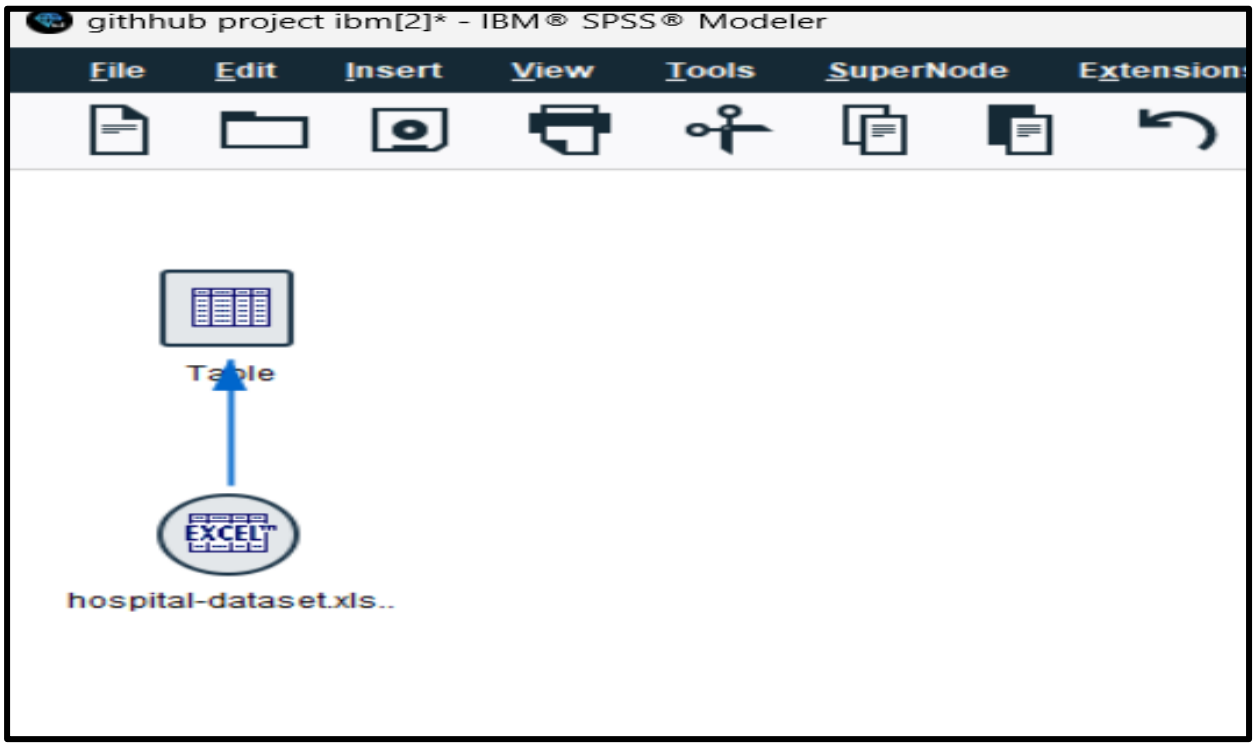
- Learned how to **import and view datasets** in SPSS Modeler.
- Understood the use of the **Distinct** node to remove duplicates and study unique entries.
- Used the **Aggregate** node to summarise data and derive meaningful insights.
- Reclassified and formatted variables using the **Type** node.
- Applied **SetToFlag** to convert selected values into binary flags.
- Visualised the **relationship between two continuous fields (Age vs Billing Amount)** using graphical analysis.

Working

First, the dataset was imported into IBM SPSS Modeler using the Excel source node and viewed through a Table node. The **Distinct** node was applied to identify unique records, and the **Aggregate** node was used to summarise numerical values. The **Type** node was then used to assign correct measurement levels to each field, followed by the **SetToFlag** node to convert selected conditions into binary values. Finally, the relationship between the two continuous fields ***(Age and Billing Amount)*** was visualised to observe patterns and trends in the data.

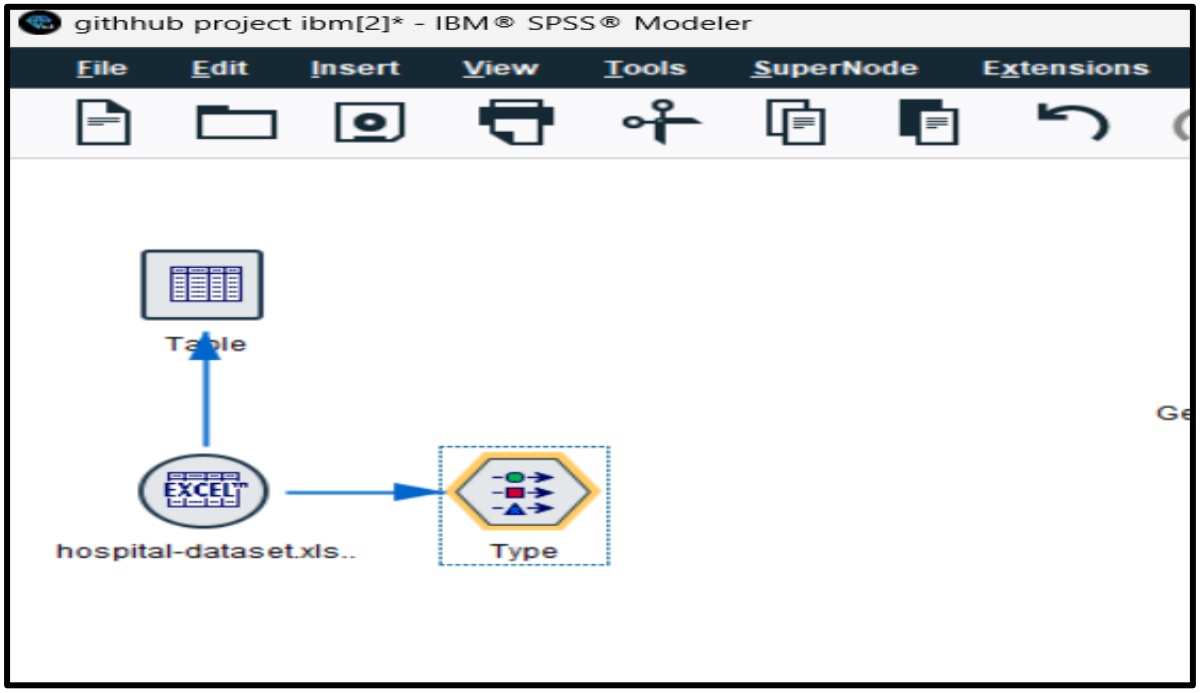
Step 1: Import Data & Preview Summary Statistics

First, we import the dataset containing patient details into SPSS Modeler and view it using a **Table node** to understand the available fields.



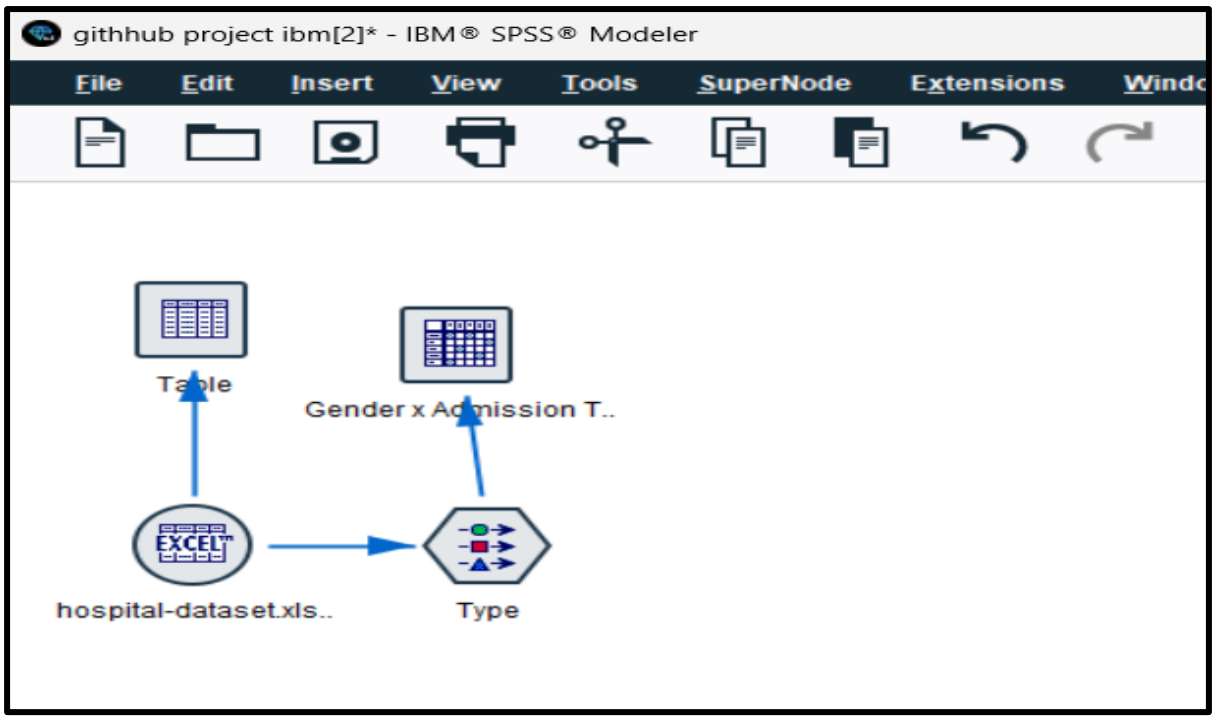
Step 2: Define Field Roles and Types

Add a **Type** node and ensure **Age** and **Billing Amount** are set as numeric/continuous. Mark identifier fields (ID, Account no.) as ID, and set categorical fields as nominal. Correct measurement levels if necessary.



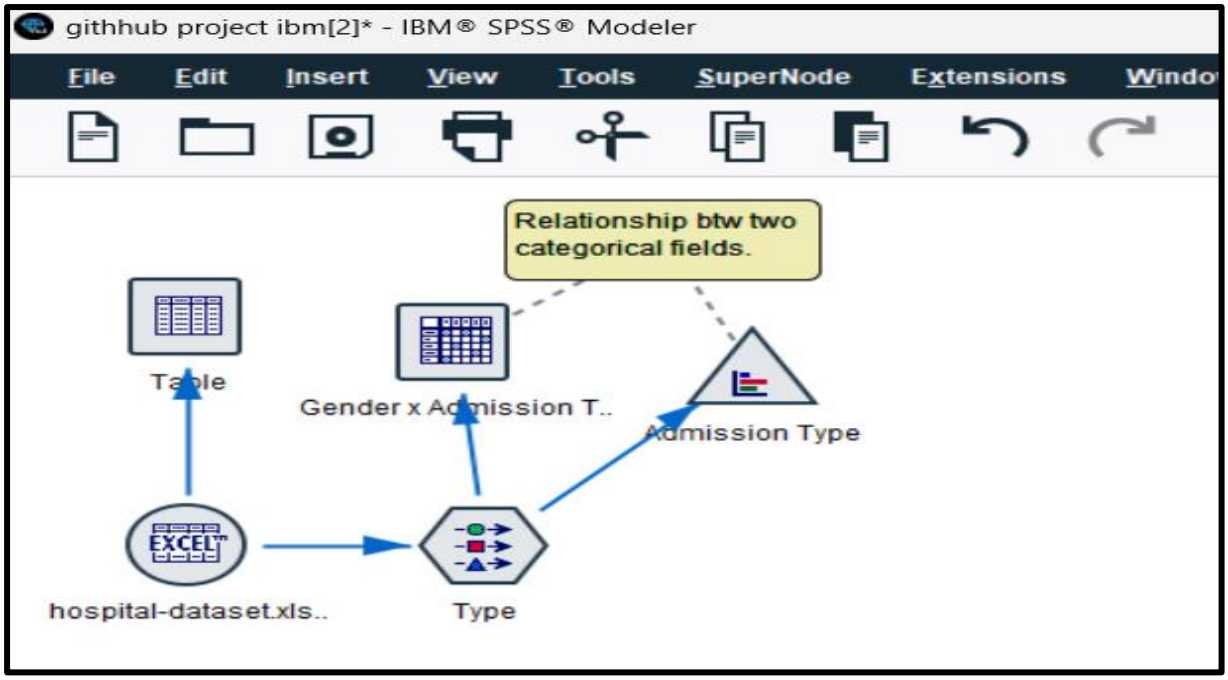
Step 3 — Create a cross-tabulation

Use a crosstab/table (labelled **Gender x Admission T.)** to tabulate **Gender** versus **Admission Type**. Configure counts and percentages as required.



Step 4: Generate Graphical Relationship Between Two Categorical Variables

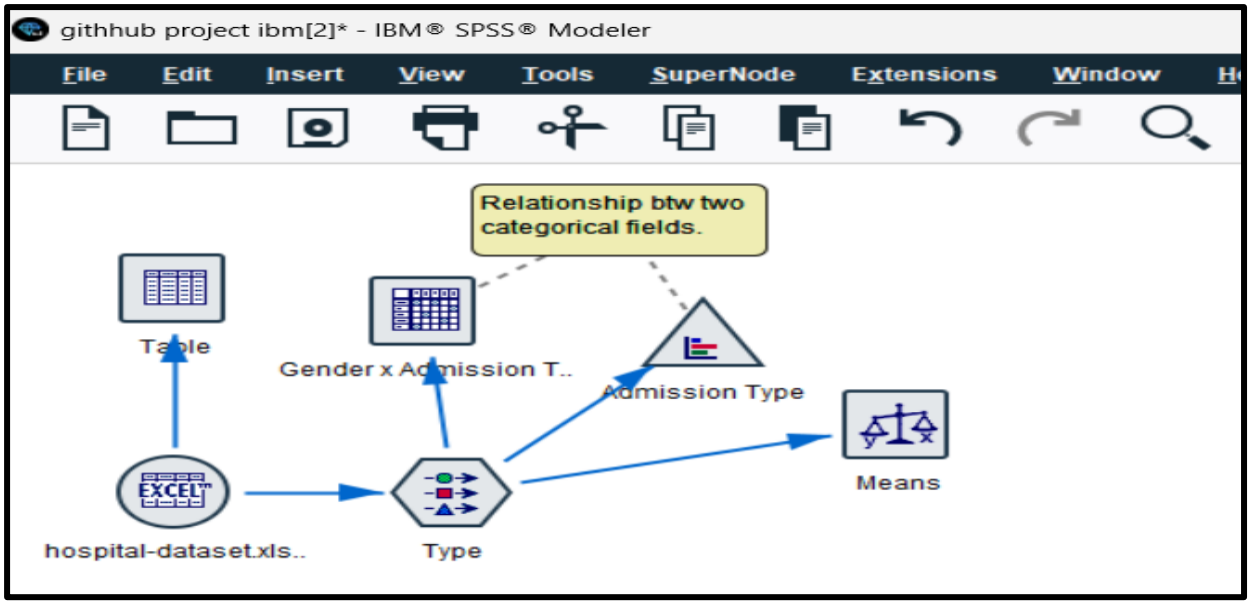
- Add the **Admission Type** chart node (visual graph node) and connect it to the Type node.
- Configure the chart to compare **Gender vs Admission Type**.
- Run the node to view a bar chart or stacked visualisation that shows how admission type varies by gender.



Step 5 — Compute group means for a categorical vs continuous relationship

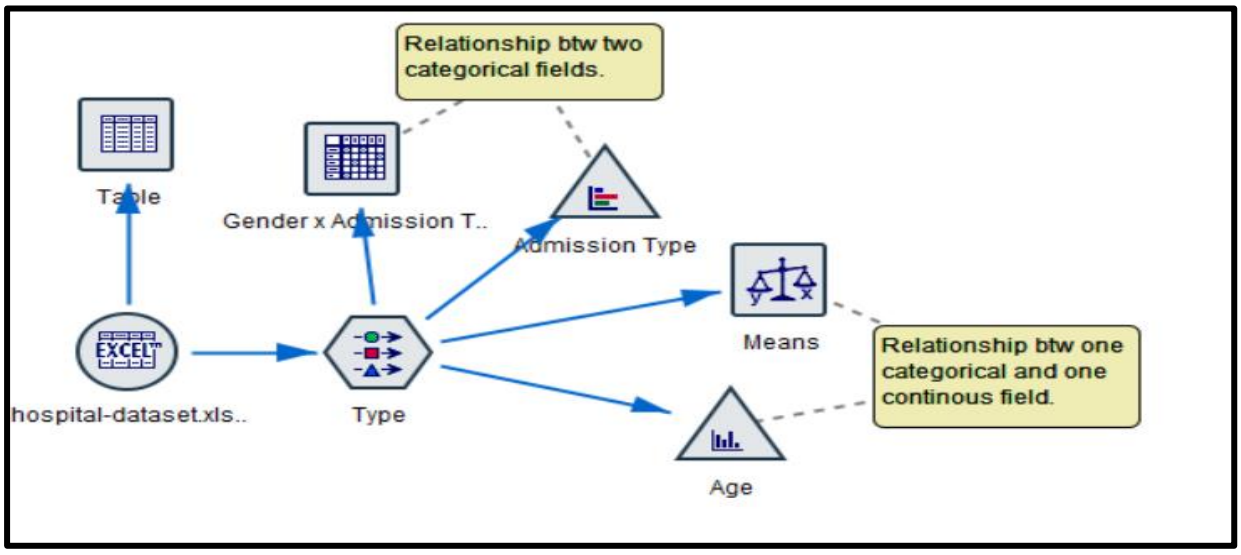
From the **Type node,** add a **Means node** to examine a categorical → continuous relationship (for example, **Admission Type → Age**). Configure the Means node:

- Grouping field = **Admission Type** (categorical).
- Summary field(s) = **Age** (continuous).
- Select summary measures to compute (mean, median, count, std. dev.).
    - Run the Means node to produce the table of average age per admission type and other descriptive statistics.



Step 6 — Interpret categorical→continuous relationship

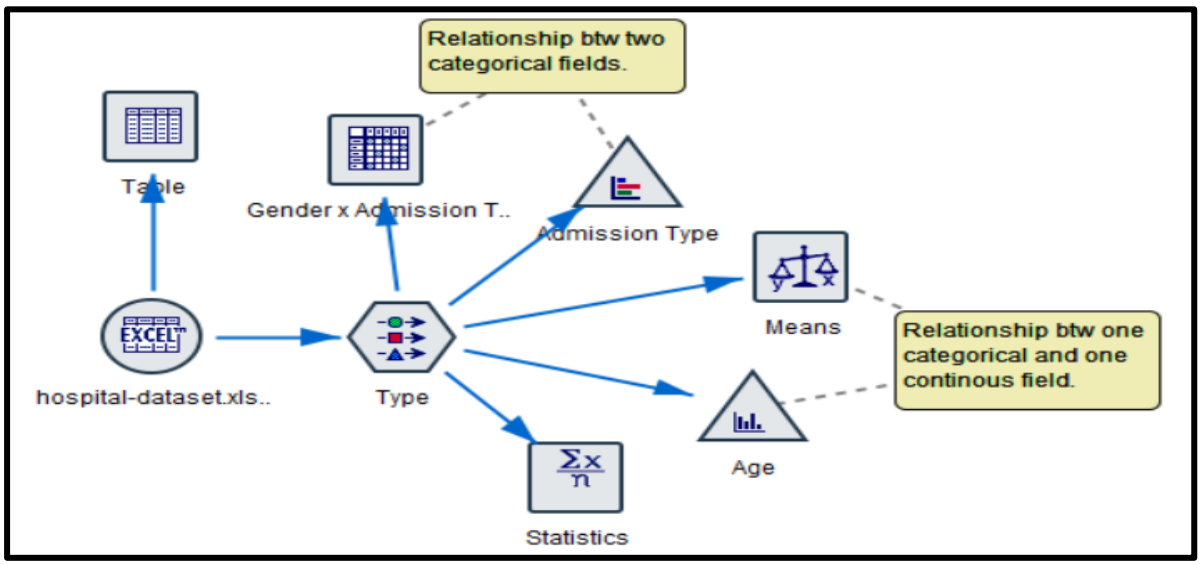
From the Means table and Age distribution, interpret how **Admission Type** relates to **Age** (e.g., which admission type has the highest mean age, and are the differences meaningful). Consider a statistical follow-up if needed.



Step 7: Calculate Summary Statistics

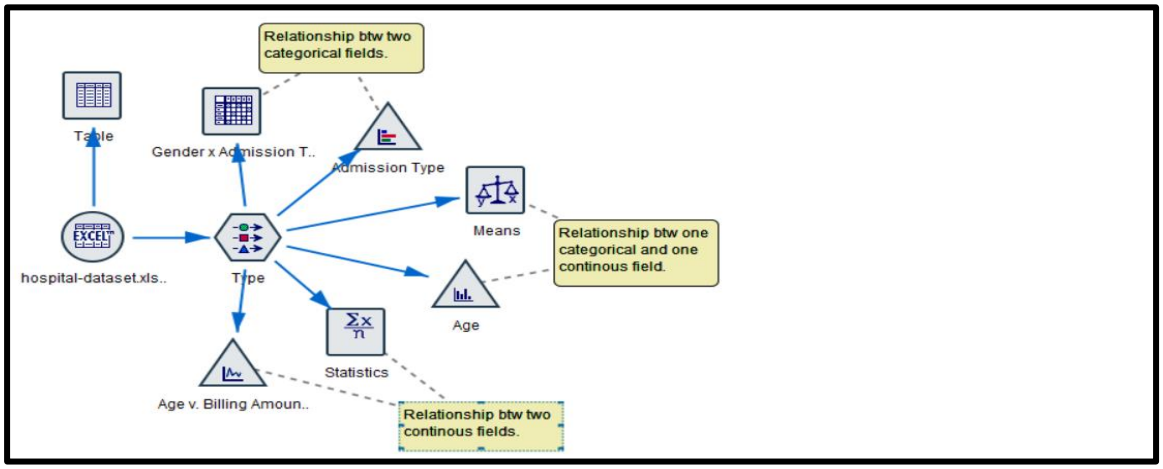
1. Drag a **Statistics** node and connect it to the **Type** node.

2. Open it and select **Descriptive Statistics** → choose fields like **Age**.
3. Run the node to view mean, min, max, and standard deviation.



Step 8 — Investigate the relationship between two continuous fields

Use a **Plot / Scatter** or **Statistics (Correlation)** node to examine **Age vs Billing Amount**. Configure X = Age, Y = Billing Amount; optionally overlay a trend line. Run the node and/or compute Pearson/Spearman correlation.

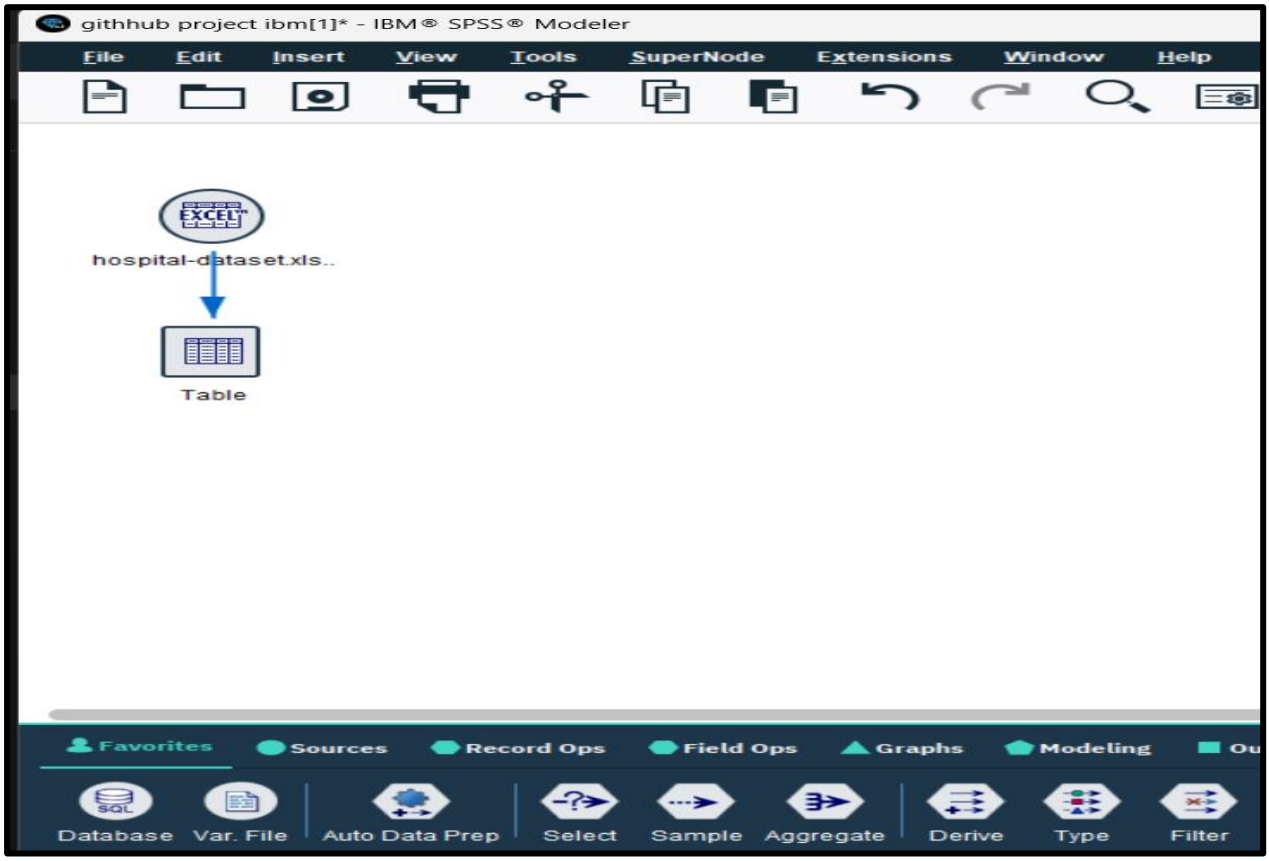


Step 9 — Add the Excel source node and attach a Table node to preview data

From the **Sources** (or Favorites) palette drag an **Excel** (or Var. File → Excel) node onto the canvas.

Why: this node will read the spreadsheet file into Modeler.

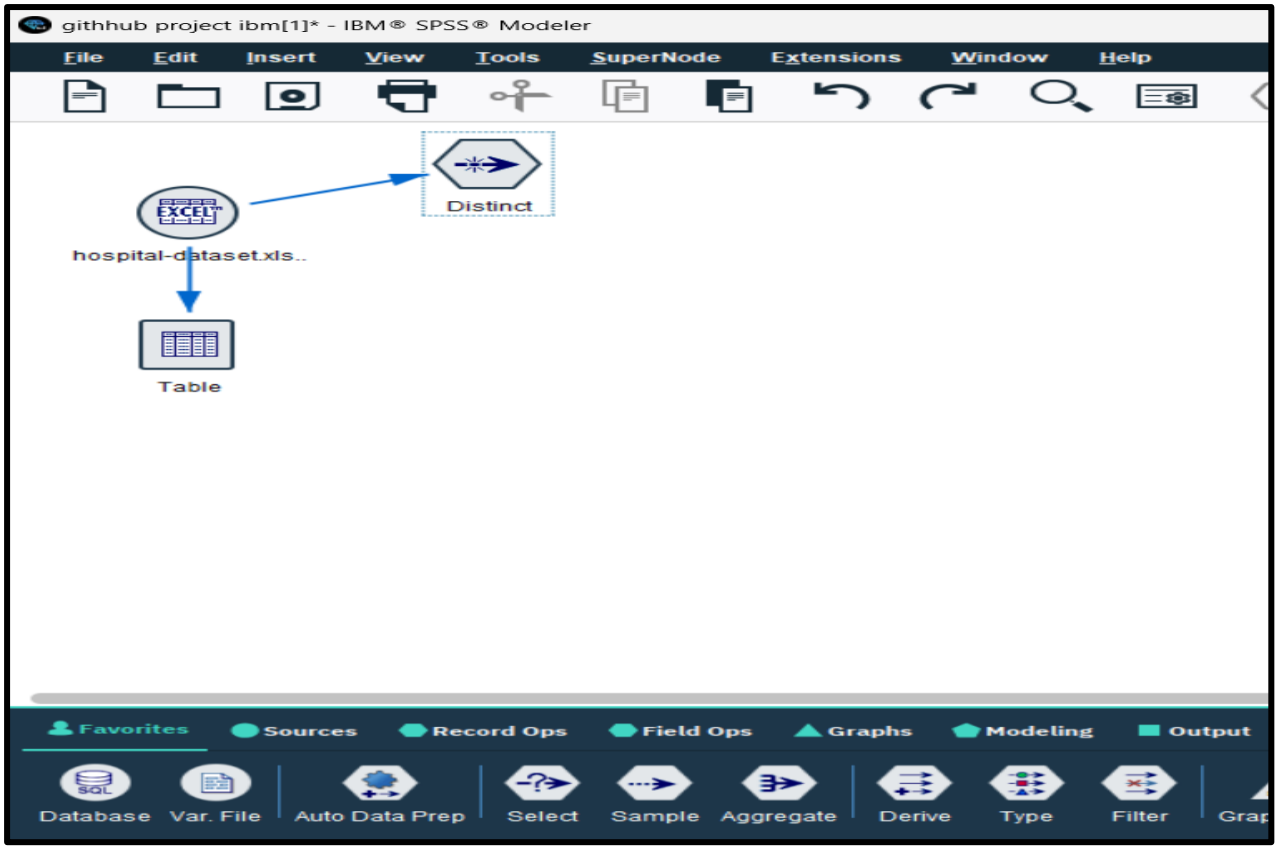
Drag a **Table** node from the Output palette and connect it to the Excel node.



Step 10 — Insert a Distinct Node to Remove Duplicate Records

1. From the **Record Ops** or **Favourites** palette, drag a **Distinct** node to the canvas.
2. Connect the **Excel node → Distinct node**.

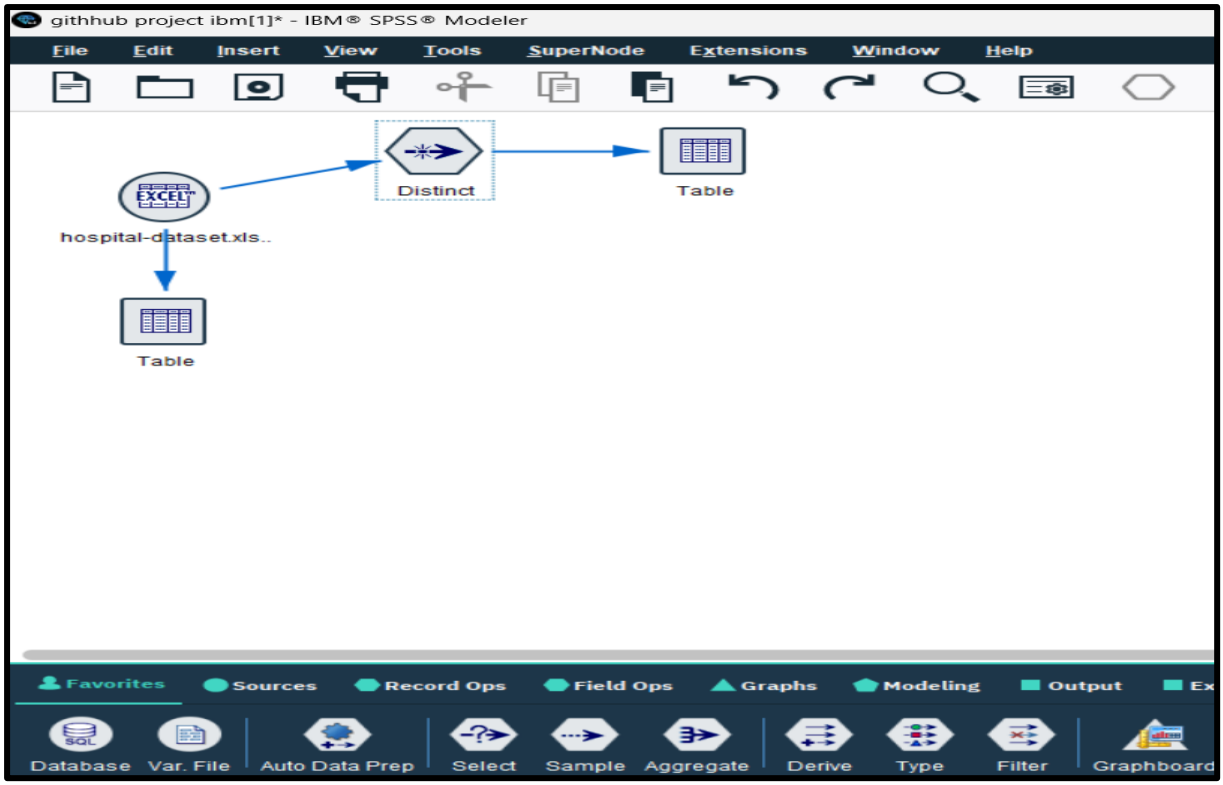
   This node is used to identify and remove duplicate records.



Step 11— View Data After Removing Duplicates

- Double-click the Distinct node.
- By default, it considers all fields when checking duplicates.
- If needed, select specific fields to check uniqueness (e.g., Patient_ID).
- Click **OK** to apply.

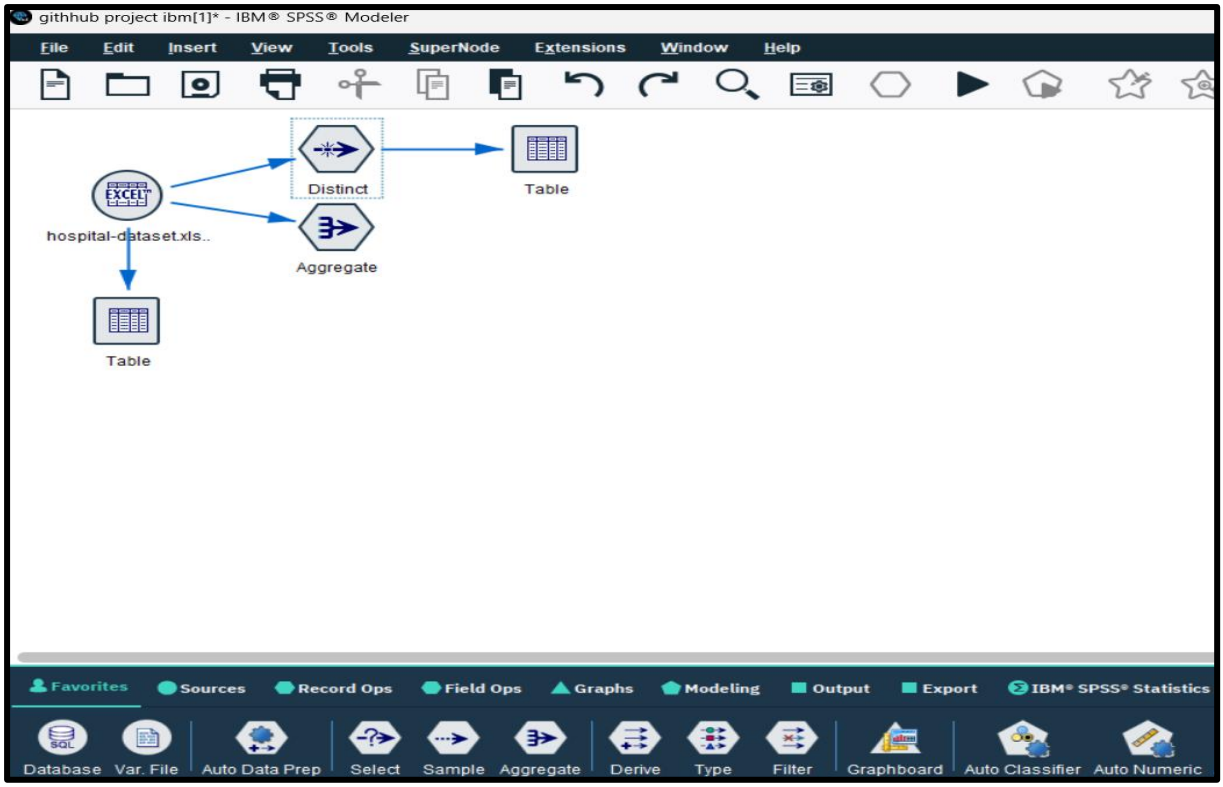
- Drag another **Table** node to the canvas.
- Connect the **Distinct** node → **Table** node.
- Right-click the **second Table** node → click **Run**.
- The output now displays the dataset **with duplicates removed**.



Step 12— Add Aggregate Node and Configure Aggregate Node

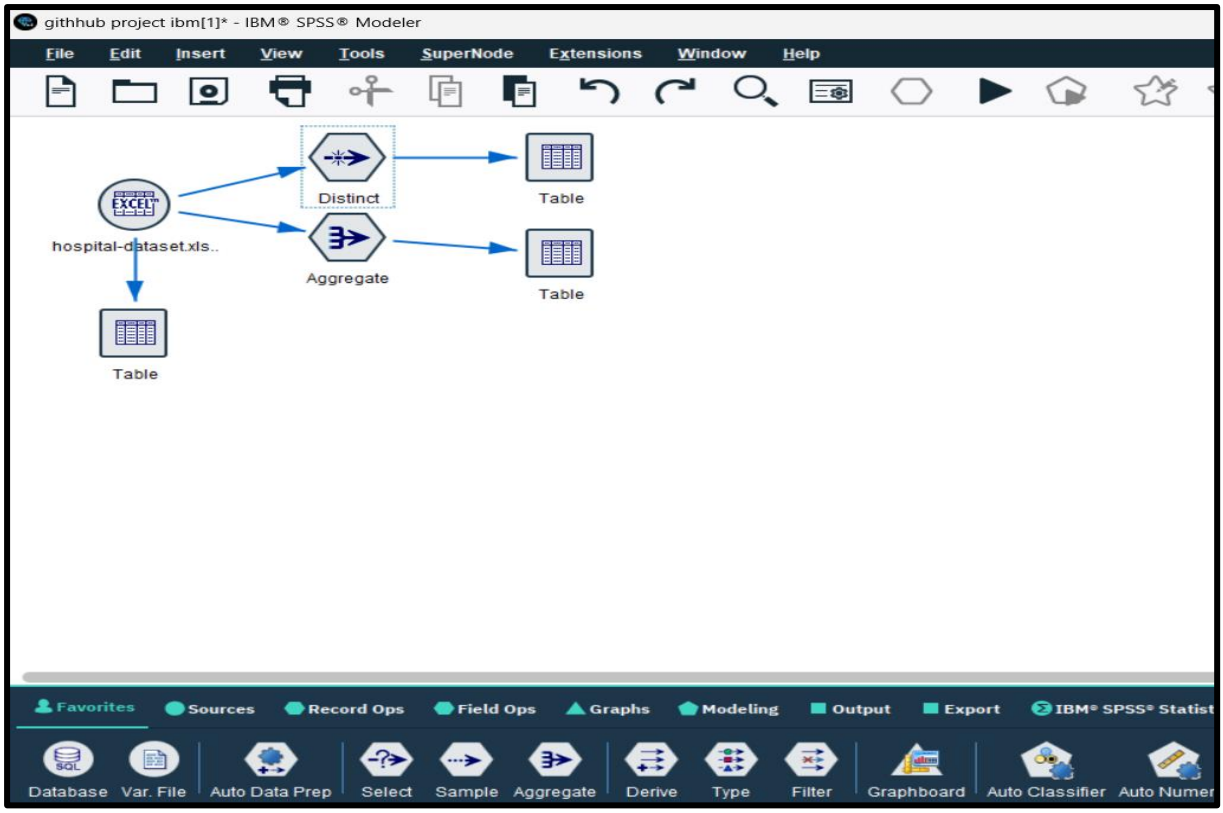
- From the Record Ops / Aggregate palette, drag the Aggregate node to the canvas.
- Connect the Excel node → Aggregate node.

- Double-click the **Aggregate** node.
- Select a **Grouping Field** (e.g., Gender, Admission_Type, Ward_No etc.).
- Grouping is used to categorize data.
- Select one or more **Summary Functions** such as:
- **Count** (to count number of patients)
- **Mean** (to calculate average age, treatment cost, etc.)
- Click **OK** to apply settings.



Step 13 — View aggregated results

- Drag a Table node and connect Aggregate → Table.
  Run this Table node to display the grouped summary table (e.g., patient counts per gender, average age per admission type).

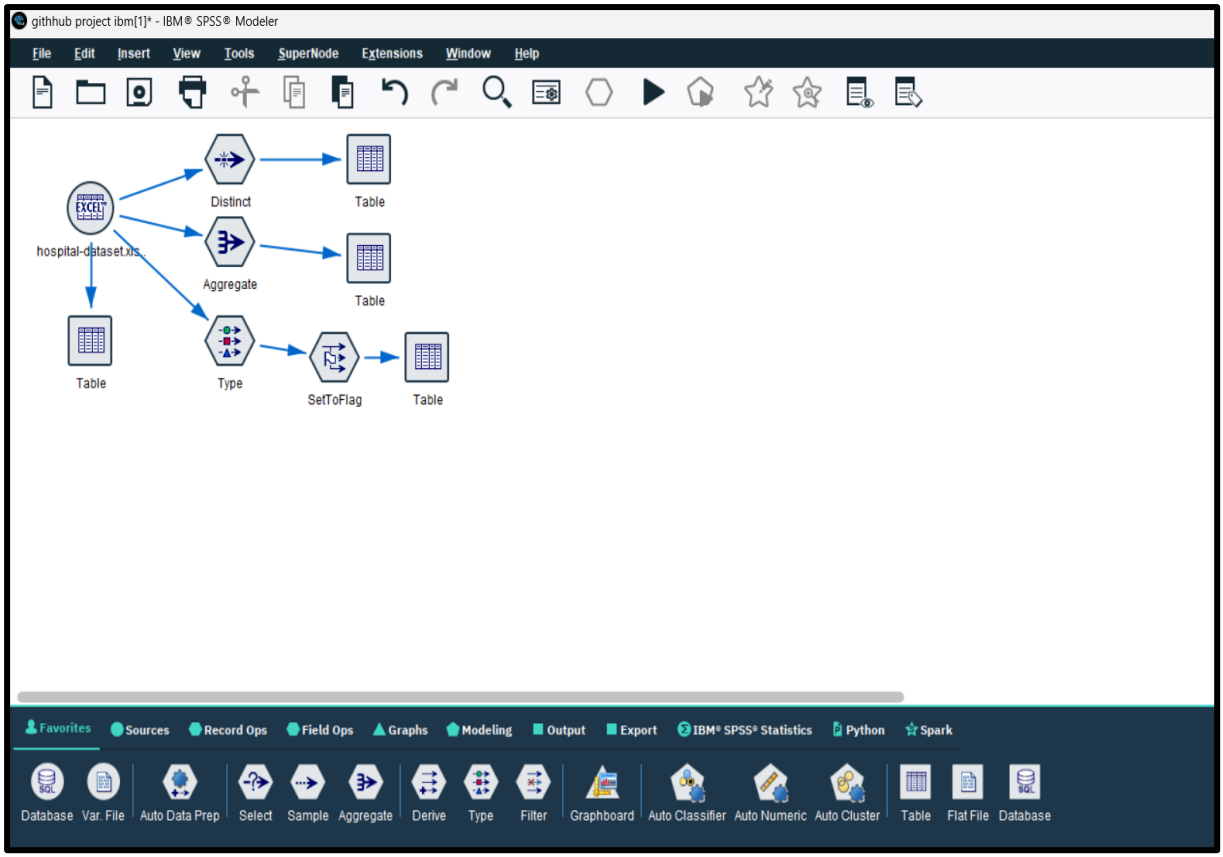


Step 14 — Apply a Type node (set roles & measurement)

- Drag a Type node onto the canvas (icon with colored shapes).
- Connect either the Excel node or an upstream node (Distinct or Aggregate) to the Type node input depending on which dataset you want to adjust. In your screenshot the Type node is fed by the Excel node.

- Double-click the Type node to configure each field:

- Role: set fields to ID, Input, Target, None as appropriate (e.g., PatientID → ID; Age, Charges → Input; Readmitted → Target).
- Measurement: set Nominal for categorical fields, Continuous for numeric fields, or Ordinal if ordered categories.
- Value labels: add or edit value labels for categorical fields (e.g., 1 = Male, 2 = Female), if needed.
- Missing value handling: note which fields have missing values (Type node doesn't impute, but flags types).
- Click OK.


githhub project ibm[1]* - IBM® SPSS® Modeler
File Edit Insert View Tools SuperNode Extensions Window Help
hospital-dataset.xls
Distinct
Table
Aggregate
Table
Table
Type
SetToFlag
Table
Favorites Sources Record Ops Field Ops Graphs Modeling Output Export IBM® SPSS® Statistics Python Spark
Database Var. File Auto Data Prep Select Sample Aggregate Derive Type Filter Graphboard Auto Classifier Auto Numeric Auto Cluster Table Flat File Database