Detailed Sales Data Analysis Workflow Summary

This document provides a detailed summary of the data analysis workflow represented in the provided process flow diagram. The workflow is designed for analyzing sales data, preparing it for modeling, building predictive models using the CHAID algorithm, and generating analytical outputs in table format.

1. Data Input
The process begins with the **sales_data.csv** file, which contains the raw sales dataset. This dataset is loaded into the workflow and serves as the foundation for all subsequent operations. Two connections are made from this dataset: **Fields:** Used to examine the metadata of the dataset, including variable names, data types, and possible value ranges. **Table:** Used to visualize and inspect the raw data before any transformation is applied.

2. Data Preparation
The preparation stage includes several nodes designed to clean and organize the dataset: **SORTED DATA:** The dataset is sorted to ensure that the data order is consistent. Sorting helps improve accuracy in subsequent steps that rely on sequential data. **Type:** This node validates and assigns the correct data types (e.g., numeric, categorical, string) to each variable. **Filler:** Missing or incomplete values are handled in this step. The filler node can apply imputation techniques or remove rows/columns with missing data. **SALES_CATEGORY:** A new field or derived variable is created based on sales values to categorize data (e.g., Low, Medium, High sales groups).

3. Partition
The **Partition** node splits the dataset into multiple subsets—typically training and testing samples. This partition ensures that model evaluation is unbiased and that predictive accuracy can be validated on unseen data.

4. Modeling (CHAID)
The workflow applies the **CHAID (Chi-squared Automatic Interaction Detector)** modeling technique twice: The first **CHAID** node generates an initial decision tree model predicting the **SALES_CATEGORY** variable. The second **CHAID** node refines or compares results, possibly adjusting splitting criteria or pruning settings to improve model performance. The CHAID model helps identify key variables influencing sales category and provides interpretable, rule-based outcomes.

5. Analysis
After modeling, the **Analysis** node is connected to the CHAID output to examine model performance. It enables visualization of decision tree results, node distributions, and statistical metrics such as p-values and chi-squared statistics.

6. Post-Processing
After analysis, post-processing steps are applied: **Filter:** Selects relevant records or model results based on defined criteria (e.g., removing outliers or focusing on specific sales categories). **Aggregate:** Summarizes data by grouping variables to compute totals, averages, or counts for better interpretation.

7. Output Tables
Each major step (raw data, sorted data, filtered data, aggregated data) is connected to a **Table** node. These tables represent intermediate and final outputs of the analysis pipeline. They can be exported for reporting, further statistical analysis, or visualization in external tools.

8. Summary

This workflow represents a complete data mining process designed for sales analysis. It includes: Data loading and exploration Data cleaning, transformation, and categorization Model building using CHAID for classification Model evaluation and performance analysis Result filtering, aggregation, and export The use of multiple CHAID models and structured partitioning ensures model reliability and interpretability for decision-making.