MASTERING KNIME: BUILD DATA PIPELINES AND DASHBOARDS VISUALLY

ARYAMAN SHARMA

1ST EDITION

Mastering KNIME: Build Data Pipelines and Dashboards Visually

1st Edition

Aryaman Sharma

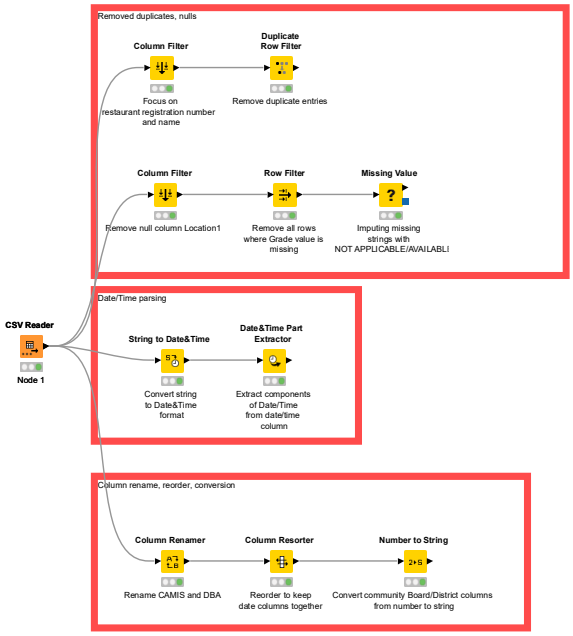
Mastering KNIME: Build Data Pipelines and Dashboards Visually

Disclaimer: This book is an independent guide and is in no way affiliated with or endorsed by KNIME AG. The information presented is based entirely on publicly available resources and the author's personal experience using the KNIME Analytics Platform. All trademarks, product names and company names or logos cited are the property of their respective owners.

First Edition: 2025
ISBN: 978-1-4452-7268-9

Contents

Community Board	NYC Community Board number where the restaurant is located.
Council District	City Council district number of the restaurant's location.
Census Tract	U.S. Census Bureau tract identifier.
BIN	NYC Department of Buildings Building Identification Number.
BBL	Borough, Block and Lot code for tax and property records.
NTA	Neighbourhood Tabulation Area code (used for NYC neighbourhood-level analysis).
Location Point1	Combined geographic location (Latitude, Longitude) in POINT format (used for mapping).

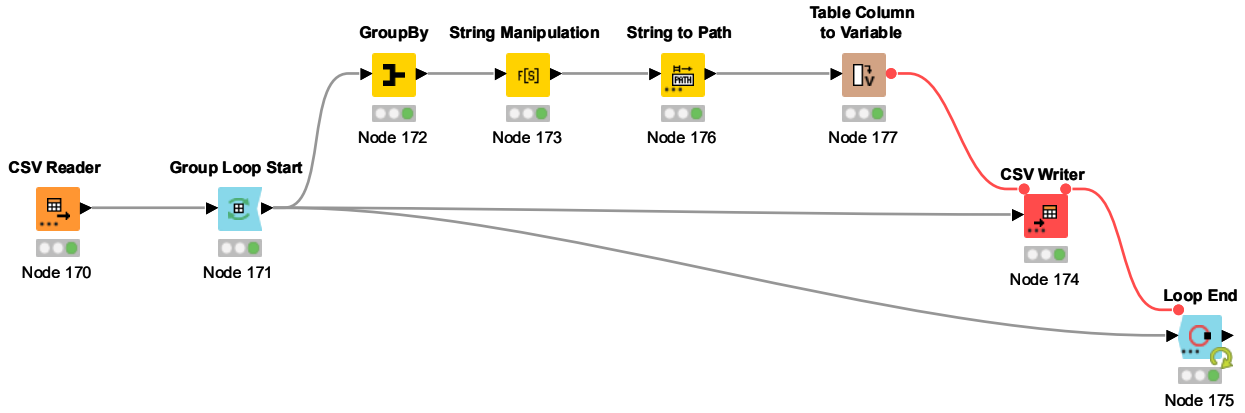


- Remove duplicates, nulls

The first step in any data preprocessing workflow is to remove duplicates and handle null values, as they can significantly impact the quality of downstream analyses. In KNIME, this process is carried out using nodes like the Duplicate Row Filter and Missing Value nodes. The Duplicate Row Filter node is configured to identify and remove duplicate records from the dataset based on selected columns, ensuring that repeated entries do not skew the results. Empty columns/rows or rows with missing values should also be filtered out. Furthermore, the Missing Value node is applied to address incomplete data. Depending on the analysis context, missing values can be either replaced with appropriate default values, such as the mean for numeric columns, or the affected rows can be removed entirely. These steps help trim the dataset to a clean and consistent state, making it ready for further processing.

Chapter 6: Automation and Reusability

- Loops and flow variables



In data processing workflows, especially when dealing with large and diverse customer datasets, it's often necessary to split records based on certain criteria such as geographic location. KNIME Analytics Platform provides a visual and modular approach to achieving this. In this example, we demonstrate a method to automatically separate rows based on the Country field and save each subset to its own CSV file.

The workflow begins with the CSV Reader node, which ingests the raw customer data file. This file contains various fields such as customer ID, name, contact details and most importantly, the country of origin. This node parses and loads the complete table into KNIME's memory, making it accessible for downstream operations.

Next, a Group Loop Start node is introduced. This node is configured to group data based on the Country column. Internally, this initiates a loop where each cycle processes only the rows corresponding to a specific country. In each iteration of the loop, only a single country's data is active, allowing for isolated processing.

We use a GroupBy node to first to get the country iteration the loop is currently in so that the same can be used while naming the dataset when extracting that country's dataset from original data. To ensure that file paths are valid and dynamic, a String Manipulation node is used. This node helps format the country name to avoid invalid characters in filenames. For example, spaces, slashes, or other special characters can be replaced with underscores or removed altogether. This ensures that file naming conventions remain clean and system compliant. The "C:/Users/aryam/OneDrive/Desktop/KNIME BOOK/COUNTRY SPLIT/" part must be renamed to the path one wants to save the final split datasets in.