INTRODUCTION

Analyzing large and complex datasets using SQL, without the need for a database administrator or the need to set up and maintain complex infrastructures is one of the key aspects of business intelligence. Big Query is a fully managed, cloud native data warehouse that enables SQL queries using the processing power of Google's infrastructure, making it a powerful tool for data analysis and business intelligence. KNIME (Konstanz Information Miner) is an open-source data analytics platform that allows users to create data pipelines, analyze data and create reports and visualizations. It also features a wide range of built-in data processing, data visualization, and machine learning nodes, making it a powerful tool for data exploration, data preparation, and modeling. Data Bears will leverage KNIME, integrate Knime to extract data from various data sets and analyze it according to business logic. Eventually, we will visualize it, to tell data stories.

KNIME Analytics Platform is open-source software for creating data science applications and services. Intuitive, open, and continuously integrating new developments, KNIME makes understanding data and designing data science workflows and reusable components.

Accessible to everyone. With KNIME Analytics Platform, you can create visual workflows with an intuitive, drag and drop style graphical interface, without the need for coding. In this quick start guide we'll take you through the KNIME Workbench and show you how you can build your first workflow. Most of your questions will probably arise as soon as you start with a real project. In this situation, you'll find a lot of answers in the KNIME Workbench.

As a company, KNIME is dedicated to creating an accessible and open-source tool, and to that end KNIME Analytics Platform is free for anybody to download and use. KNIME is a full-featured and incredibly powerful tool for building data prep workflows of any level of complexity.

Whether you need to connect to Excel files, a Snowflake database, process audio files, analyze images, or build an interactive dashboard, KNIME can do it. Every day that we use KNIME, whether for a client project or personal education, we learn something new and exciting about the tool. Using KNIME is incredibly easy.

USECASE

KNIME is a strong and adaptable platform for data analytics that can be utilized for a variety of use cases across different sectors. The following are a some of the most typical KNIME usage cases:

Data cleaning, preprocessing, and transformation: KNIME can be used to clean, transform, and convert raw data into an easily analyzed format. It has a variety of tools for cleansing, organizing, and transforming data.

KNIME is a tool for predictive analytics, which may be used to find patterns and trends in data, create predictive models, and forecast future outcomes.

Image processing, which includes activities like segmentation, object detection, and feature extraction, is a task that KNIME can perform. It comes with a number of built-in nodes for processing images.

Text Analytics: Text analytics is the process of examining and drawing conclusions from unstructured data, such as text, and it can be done with KNIME. For text processing, it comes with a number of pre-built nodes, such as sentiment analysis, text categorization, and entity recognition.

Social Media Analytics: KNIME can be used to examine social media data and draw conclusions regarding user sentiment, behavior, and preferences.

KNIME can be used to analyze biological data, such as gene expression, protein structure, and genomics data, in the field of bioinformatics.

KNIME can be used for business intelligence, which entails deriving knowledge from data to assist in decision-making. It comes with several pre-built nodes for reporting, dashboarding, and data visualization.

Fraud Detection: KNIME is a tool that can be used to spot and stop fraudulent activities in financial transactions.

In real estate there are many more use cases where knime is generally used and few of them are

Predictive Analytics: KNIME can be used for predictive analytics in real estate to predict property values and rental rates based on historical data and market trends.

Portfolio Management: KNIME can be used for real estate portfolio management by analyzing property data to identify profitable properties and optimize portfolio performance.

Location Analysis: KNIME can be used for location analysis in real estate to identify ideal locations for new developments or investments

Risk Management: KNIME can be used for risk management in real estate to identify potential risks and minimize their impact.

Property Valuation: KNIME can be used for property valuation by analyzing property data to determine accurate property values.

SOFTWARE DETAILS

KNIME is an open-source platform for process automation, machine learning, and data analysis. It is highly modular in design, allowing users to quickly add and remove functionality as needed. These are some essential KNIME software specifics:

Platforms: KNIME is compatible with a wide number of operating systems thanks to its availability on Windows, Mac OS X, and Linux.

Structured, unstructured, and multimedia data are just a few of the several forms of data that KNIME can manage.

KNIME is capable of reading data from a wide range of sources, including local files, databases, online services, and cloud services.

Analytics: For the processing, modeling, visualization, and reporting of data, KNIME offers a variety of analytical tools. Additionally, it enables interaction with well-known analytics programs like R and python.

KNIME comes with a machine learning library that includes a variety of classification, regression, clustering, and association methods. Moreover, it allows for the integration of third-party machine learning frameworks like TensorFlow and Keras.

Workflow: KNIME has a drag-and-drop workflow interface that makes it simple for users to build elaborate data processing and analysis pipelines.

Collaboration: Using version control, history tracking, and the sharing of data and processes with others, KNIME enables users to work together on workflows.

Extensibility: KNIME's basic platform may be expanded with a variety of extensions and plugins to add features or interface with other applications.

KNIME, in general, is a strong and adaptable software platform for data analysis, machine learning, and process automation. It has a variety of features and capabilities that make it a popular choice for many data scientists and analysts.

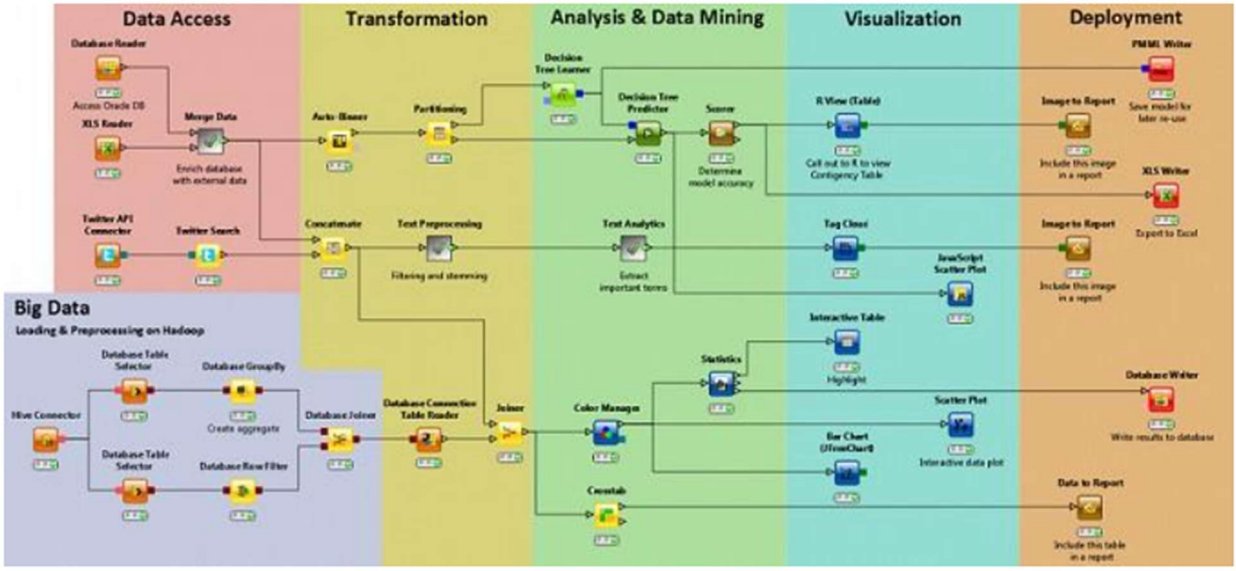
KNIME Extensions: The KNIME Analytics Platform is enhanced by a sizable and expanding collection of extensions, which are effectively plug-ins. The Picture Processing, Text Processing, and Cheminformatics extensions are a few of the more well-liked add-ons.

KNIME Cloud Analytics Platform: KNIME provides a cloud-based version of their software that enables customers to do machine learning and data analytics in the cloud. This platform offers a scalable, secure, and adaptable environment for data analytics and is completely integrated with the KNIME Analytics Platform.

Integrations made possible by KNIME partner companies: KNIME has collaborated with a number of tech companies to integrate their software with KNIME. Tableau, Alteryx, and Amazon Web Services are a few instances of partner integrations (AWS).

KNIME Community: KNIME has a sizable and vibrant user, developer, and partner community that actively participates in the software's development. To ensure that users are utilizing the KNIME Analytics Platform to its full potential, the community helps, instruction, and resources.

These are just a few KNIME's software specifics. KNIME is a robust and adaptable data analytics and machine learning tool that offers a variety of capabilities and integrations to fit the demands of different users and organizations.



Data science and machine learning software platforms include KNIME. Users can drag and drop nodes to build workflows that process, analyze, and visualize data using the graphical interface it offers. The software was created in Java and is compatible with Linux, Mac, and Windows.

These are some essential KNIME software specifics:

Data Integration: KNIME may be integrated with a wide range of data sources, including databases, cloud services, and online services. It also supports many data types.

Data preprocessing, transformation, and analysis are all possible using a variety of analytics and machine learning tools from KNIME. It supports several modeling methods, including association rule mining, classification, clustering, and regression.

Visualization: KNIME offers customers several visualization options that let them explore and comprehend their data. Users can alter the look of their visualizations and it offers interactive visualizations.

Collaboration: KNIME enables users to exchange data and workflows with others. It enables real-time user collaboration on workflows and version control.

SWOT Analysis of KNIME

Strengths:

User-friendly design: KNIME's drag-and-drop user interface is simple to use and intuitive, making it usable by people of different technical backgrounds.

Open-source: KNIME is free software that may be used by anybody. This makes it possible for community contributions and cooperation.

Scalability: KNIME is capable of handling high data volumes and may be installed on a range of hardware setups, including cloud-based services.

Versatility: KNIME offers many different features and functionalities, such as data integration, analytics and machine learning, and visualization.

Integration: R, Python, and Hadoop are just a few of the third-party tools and services that KNIME can be integrated with.

Weaknesses:

A high learning curve may be necessary for users who are unfamiliar with data science and machine learning ideas, even though KNIME's interface is user-friendly.

Support resources may be limited for users who need more help, even though KNIME offers documentation and a community forum.

KNIME can be resource-intensive, requiring strong hardware and computer capabilities to handle massive volumes of data and intricate workflows.

Opportunities:

Expanding market: There is a growing need for data science and machine learning tools, which presents KNIME with chances to grow both its user base and its functionalities.

Collaborations: To increase its possibilities for integration and support, KNIME has the chance to collaborate with other software and service providers.

Development that will never stop: KNIME will be able to innovate and improve its software in order to satisfy the changing demands of its users and the market.

Threats:

Competition: KNIME has competition from a variety of data science and machine learning tools available on the market.

Technological developments: As technology develops, it's possible that KNIME will become obsolete or unable to keep up with new capabilities and trends.

Data leaks and security breaches are possible with any software platform, which could damage KNIME's reputation and influence its user base.

Software Details

As an actively developed open-source software platform, KNIME constantly receives new software innovations and makes improvements to already existing capabilities. Here are a few KNIME software updates from recently:

Data connectivity, data visualization, machine learning, and analytics workflows are among the new capabilities and enhancements in KNIME Analytics Platform 4.4. Together with new data cleaning and processing nodes, it also features new tool integrations from third parties.

KNIME Server 4.13: This version of the server software provides a number of enhancements, such as new security features, performance upgrades, and improved support for large data situations.

KNIME Hub: Users can share workflows, nodes, and other resources with the community on the KNIME Hub's online platform. The Hub also offers a storefront where customers can get KNIME extensions from other developers.

Integration with other tools: R, Python, Hadoop, and cloud-based services like Amazon Web Services and Google Cloud Platform are just a few of the tools and services that KNIME is continuing to extend its connection with.

Contributions from the community: KNIME is an open-source software platform that benefits from contributions from its user base. New nodes and workflows, as well as bug fixes and enhancements to current functionality, are recent community contributions.

KNIME Analytics Platform 4.4: KNIME published version 4.4 of its analytics platform in November 2021. New machine learning algorithms, support for TensorFlow 2.0, enhanced performance, and additional workflow tools are all included in this edition.

KNIME Extension for Deep Learning 4.4: Version 4.4 of KNIME's extension for deep learning was published along with version 4.4 of the analytics platform. New deep learning nodes, support for TensorFlow 2.0, and enhanced usability are all part of this upgrade.

KNIME Server 4.12: KNIME published version 4.12 of its server software in October 2021. This release contains new capabilities for workflow management and collaboration in addition to speed, scalability, and security enhancements.

KNIME Hub: KNIME Hub is an online library of components, nodes, and processes that KNIME users can share and reuse. Improved search and discovery capabilities, GitHub connection, and improved security features are some of the most recent software updates for KNIME Hub.

KNIME Cloud Analytics Platform: KNIME delivers a cloud-based analytics platform with scalable compute and storage resources for managing massive data science and machine learning projects. Support for Kubernetes and additional tools for managing and monitoring processes are recent software additions for the KNIME Cloud Analytics Platform.

Overall, KNIME is a dynamic and ever-evolving software platform with a strong developer community and a dedication to continuously enhancing its offerings to better serve its consumers.

Future Developments

Being a software provider, KNIME continuously seeks to enhance and develop its platform in order to satisfy the shifting demands of both its users and the market. These are a few prospective advancements for KNIME in the future:

Better user experience: KNIME may keep concentrating on making its platform even more approachable and accessible for users with a range of technical skill levels.

Increasing automation and AI skills: KNIME may strive to create new features and capabilities to support these trends, such as enhanced computer vision and natural language processing (NLP) abilities, as the demand for automation and AI continues to rise.

Improved security features: KNIME may consider enhancing its security measures to make sure that user data is always safeguarded as cybersecurity threats continue to grow.

More real-time data processing support: As more businesses look to employ real-time data streams for machine learning and other applications, KNIME may provide new features to assist real-time data processing and analysis.

Increased support for use cases specific to certain industries: KNIME may continue to provide new features and capabilities to support use cases specific to certain industries, like healthcare analytics, supply chain optimization, and financial forecasting.

Integration with AI: KNIME is probably going to keep spending money on enhancing its AI capabilities. This can involve the development of fresh machine learning algorithms, integration of NLP, and improved deep learning capabilities.

Enhanced automation: KNIME is probably going to keep putting a lot of effort on automation, which will let users create more complicated workflows with less manual labor. Improved batch processing, automatic feature engineering, and more sophisticated data pretreatment capabilities could be examples of this.

Collaboration will likely be improved when KNIME makes investments in this direction, enabling teams to collaborate more successfully on data science and machine learning projects. These could include better project management tools, greater version control, and more sophisticated workflow sharing.

Improved cloud integration: KNIME is anticipated to invest in enhancing its integration with cloud-based applications as cloud computing becomes more popular. This might entail better Kubernetes support, more scalability, and more effective cloud-based processing.

More integrations with third-party products: KNIME is probably going to keep improving its ability to integrate with third-party applications, enabling users to operate smoothly across many platforms. This may entail deeper integrations with well-known machine learning frameworks and programming languages like Python and R.

More sophisticated data visualization: KNIME is anticipated to make investments in enhancing its data visualization capabilities, enabling users to communicate the outcomes of their research more effectively.

More chart formats, enhanced interactivity, and more complex customizability possibilities might be part of this.

Improved Artificial Intelligence (AI) Capabilities: KNIME is anticipated to concentrate on expanding its AI capabilities as AI becomes more common in the commercial sphere. This might entail adding fresh deep learning and artificial intelligence algorithms as well as extending its support for well-liked TensorFlow and PyTorch AI frameworks.

Features for Better Team Collaboration: KNIME may concentrate on creating new features for better workflow sharing and version control.

Support for cloud-based services will probably continue to grow, giving customers easier access to and deployment of workflows on cloud infrastructure.

Better Workflow Visualization: KNIME might concentrate on creating new visualization tools to assist users in better comprehending and interpreting their workflows, making it simpler to examine data and derive insights.

Increasing Emphasis on Data Privacy and Security: KNIME is anticipated to keep enhancing its security features and adopting new technologies to safeguard sensitive data as data privacy and security become more important to businesses and organizations.

Improved Support for Explainable AI: KNIME may concentrate on creating features and tools that make it simpler for people to comprehend and use AI models, contributing to the growth in explainability and transparency.

Overall, KNIME is expected to adapt and include new technologies and features as the area of data science and machine learning continues to develop to stay ahead of the curve and satisfy the needs of its users.

More automation: KNIME may keep developing its automation features, making it easier for users to automate routine operations and workflows.

Improved scalability: KNIME may concentrate on enhancing its scalability to handle larger datasets and more intricate procedures as data volumes continue to rise.

Improved machine learning: KNIME has already achieved tremendous advancements in the field of machine learning, but it may continue to create new features and algorithms to raise the bar even higher.

Integration with more outside tools: KNIME may keep putting more of an emphasis on integrating with other programs and services, making it simpler for users to add KNIME to their current processes and toolchains.

A more advanced and interactive visualization may be offered to users by KNIME as it continues to develop its visualization capabilities, allowing them to explore and comprehend their data.

Improved big data and cloud computing support: KNIME may concentrate on delivering stronger support for these technologies as they continue to gain popularity, making it simpler for users to handle and analyze enormous datasets in the cloud.

CONCLUSION

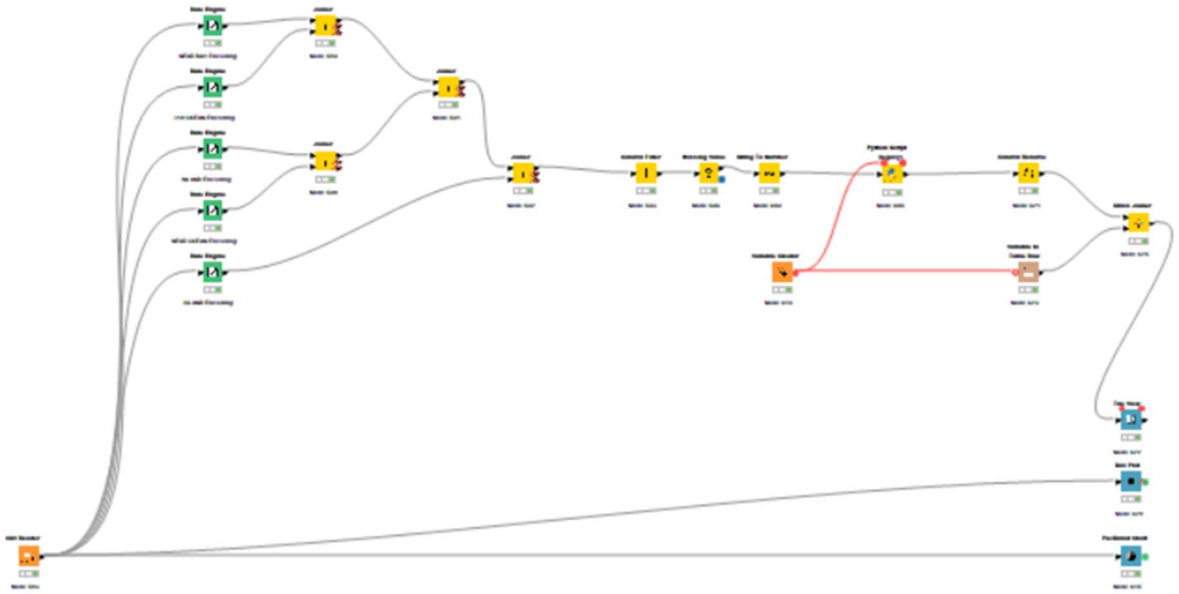
In conclusion, KNIME is a robust and adaptable platform for data analytics and machine learning that is widely utilized in a variety of sectors, including manufacturing, banking, and the health sciences. KNIME is a well-liked option for both novice and professional users due to its user-friendly drag-and-drop interface that makes it simple for users to design and deploy data workflows.

Because KNIME is extensible and open-source, users may easily link it with other programs and services to create a flexible and individualized environment for data analytics and machine learning projects. Furthermore, KNIME's active community and frequent updates make sure that the program is current with the most recent advancements in data science and machine learning.

Despite its advantages, KNIME has significant drawbacks, including a challenging learning curve for more complicated features and the requirement for technical knowledge to set up and operate more intricate workflows. KNIME is an effective tool for data scientists and analysts, however, who require an adaptable, scalable, and customizable platform for their data analytics and machine learning projects.

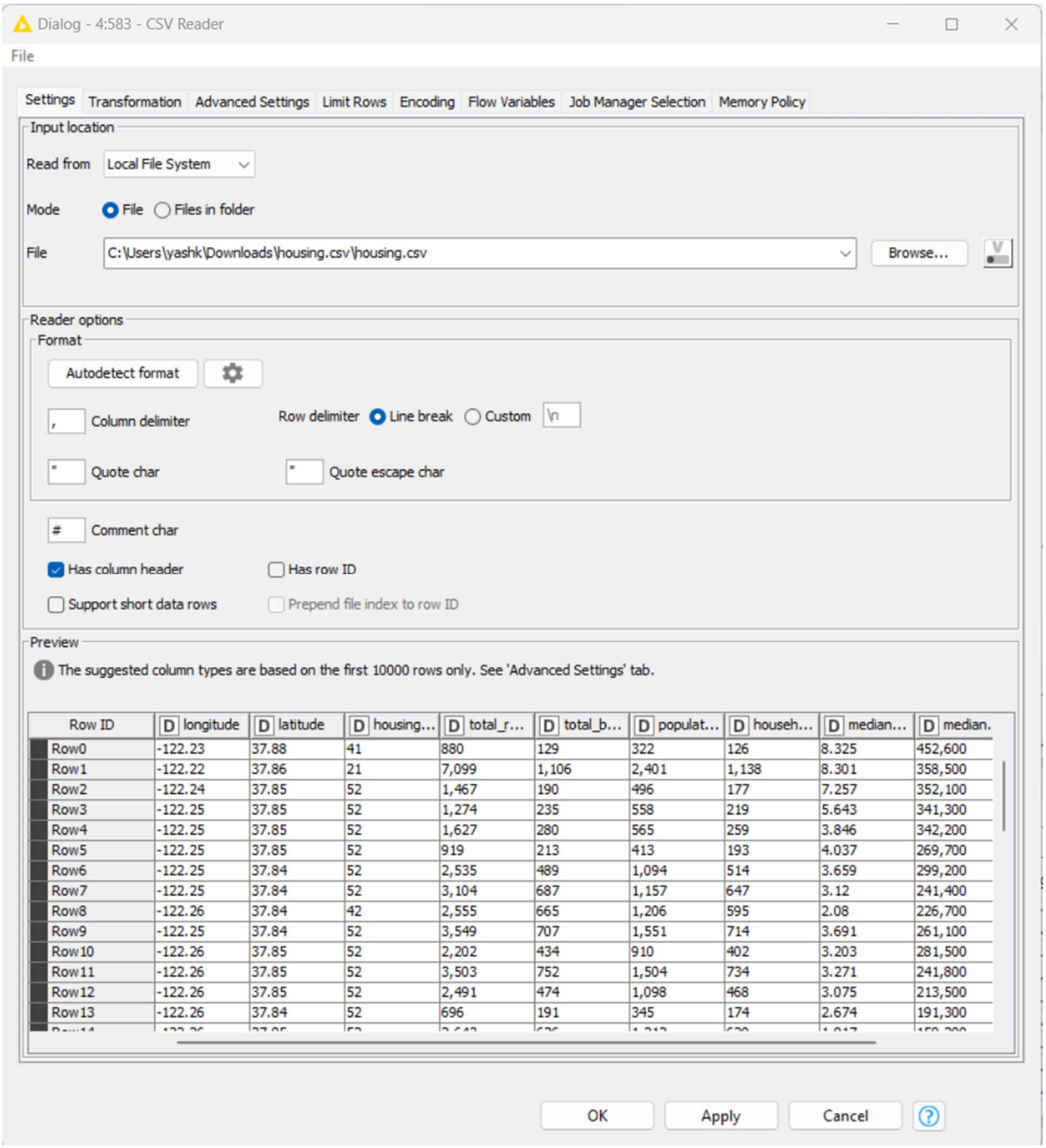
As a result, KNIME is a strong and adaptable platform for data science and machine learning that provides a variety of tools and features to let users explore and examine their data. It offers a user-friendly graphical interface for creating processes, a sizable library of pre-built nodes, and connections with tools and services from third parties. KNIME also provides robust machine learning capabilities, allowing users to apply a range of algorithms and techniques to their data. The platform is always changing, with frequent upgrades and the addition of new features to enhance its use and efficiency. With its simplicity of use, potent features, and emphasis on automation and integration.

APPENDIX – Tool Tutorial



Data Importing:

From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "CSV File Reader" node onto the workflow canvas. Then double-click on the node to open its configuration window and make configurations as shown below:

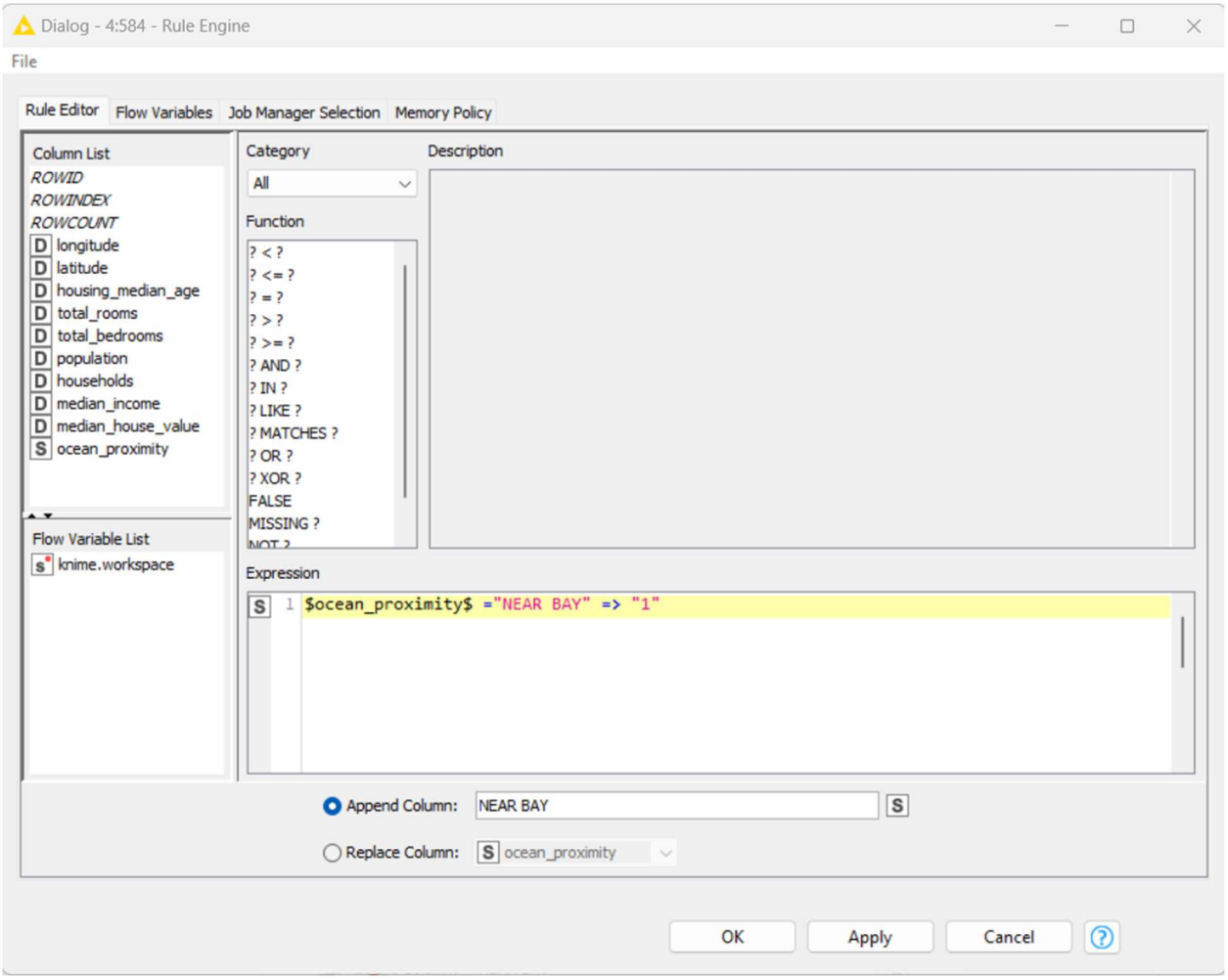


Row ID	longitude	latitude	housing...	total_r...	total_b...	populat...	househ...	median...	median.
Row0	-122.23	37.88	41	880	129	322	126	8.325	452,600
Row1	-122.22	37.86	21	7,099	1,106	2,401	1,138	8.301	358,500
Row2	-122.24	37.85	52	1,467	190	496	177	7.257	352,100
Row3	-122.25	37.85	52	1,274	235	558	219	5.643	341,300
Row4	-122.25	37.85	52	1,627	280	565	259	3.846	342,200
Row5	-122.25	37.85	52	919	213	413	193	4.037	269,700
Row6	-122.25	37.84	52	2,535	489	1,094	514	3.659	299,200
Row7	-122.25	37.84	52	3,104	687	1,157	647	3.12	241,400
Row8	-122.26	37.84	42	2,555	665	1,206	595	2.08	226,700
Row9	-122.25	37.84	52	3,549	707	1,551	714	3.691	261,100
Row10	-122.26	37.85	52	2,202	434	910	402	3.203	281,500
Row11	-122.26	37.85	52	3,503	752	1,504	734	3.271	241,800
Row12	-122.26	37.85	52	2,491	474	1,098	468	3.075	213,500
Row13	-122.26	37.84	52	696	191	345	174	2.674	191,300

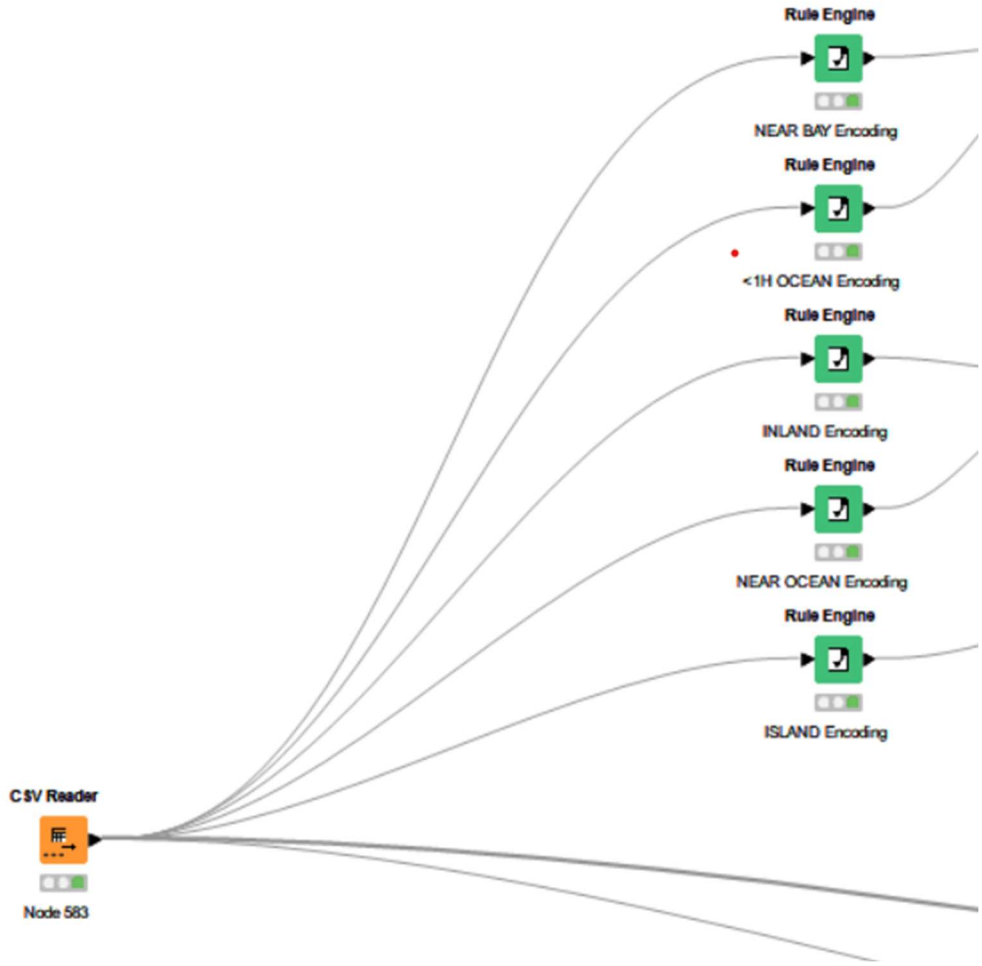
Data Encoding:

String input parameters that are used to predict the house cost need to be encoded in binary for the model to fit the regressor. Hence, splitting the “Ocean proximity” column using the rule engine. From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "Rule Engine" node onto the workflow canvas and double-click on the "Rule Engine" node to open its configuration window:

Using OneHotEncoding encode the Ocean Proximity column:



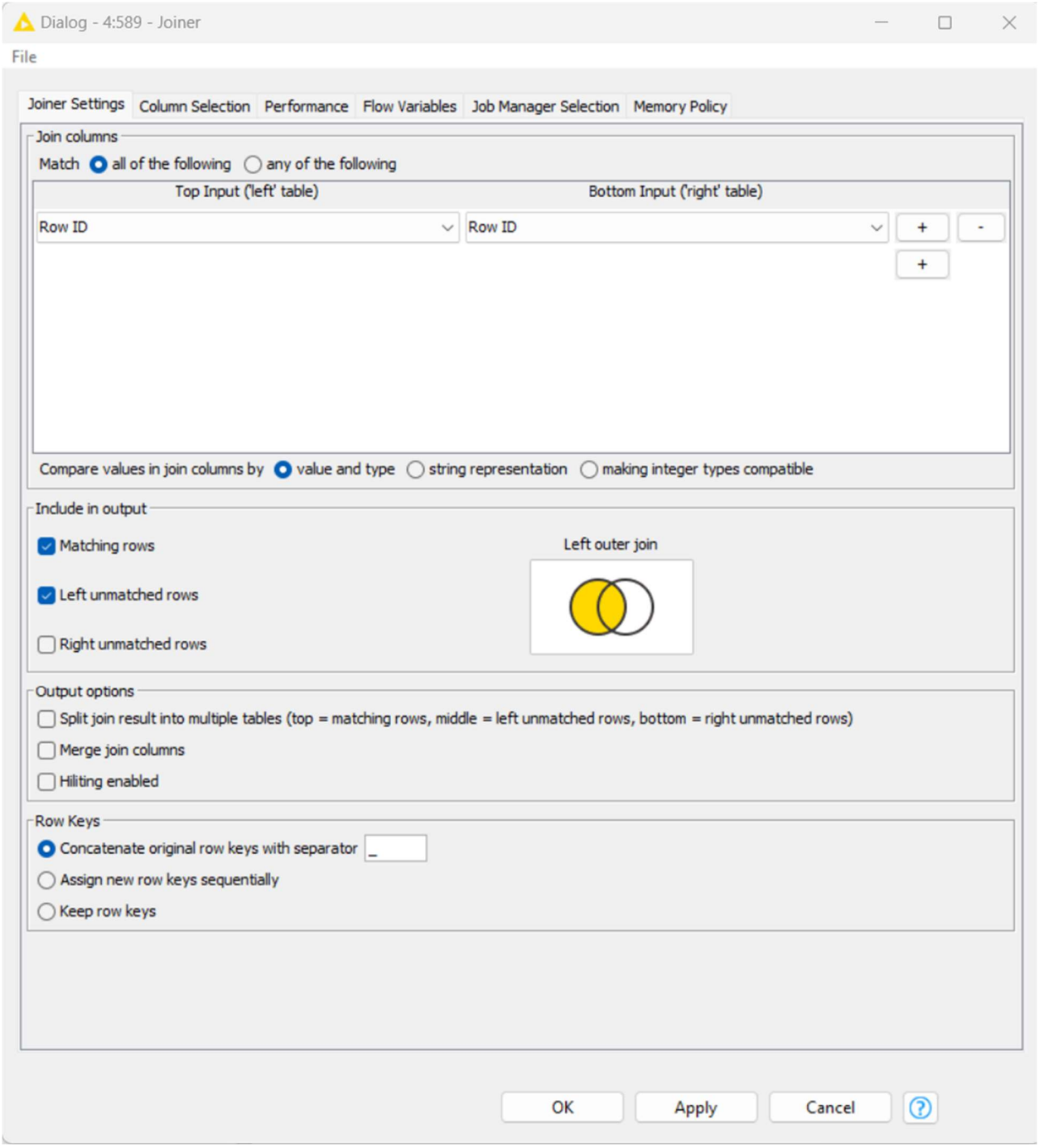
Repeat the process for all distinct values in Ocean Proximity.

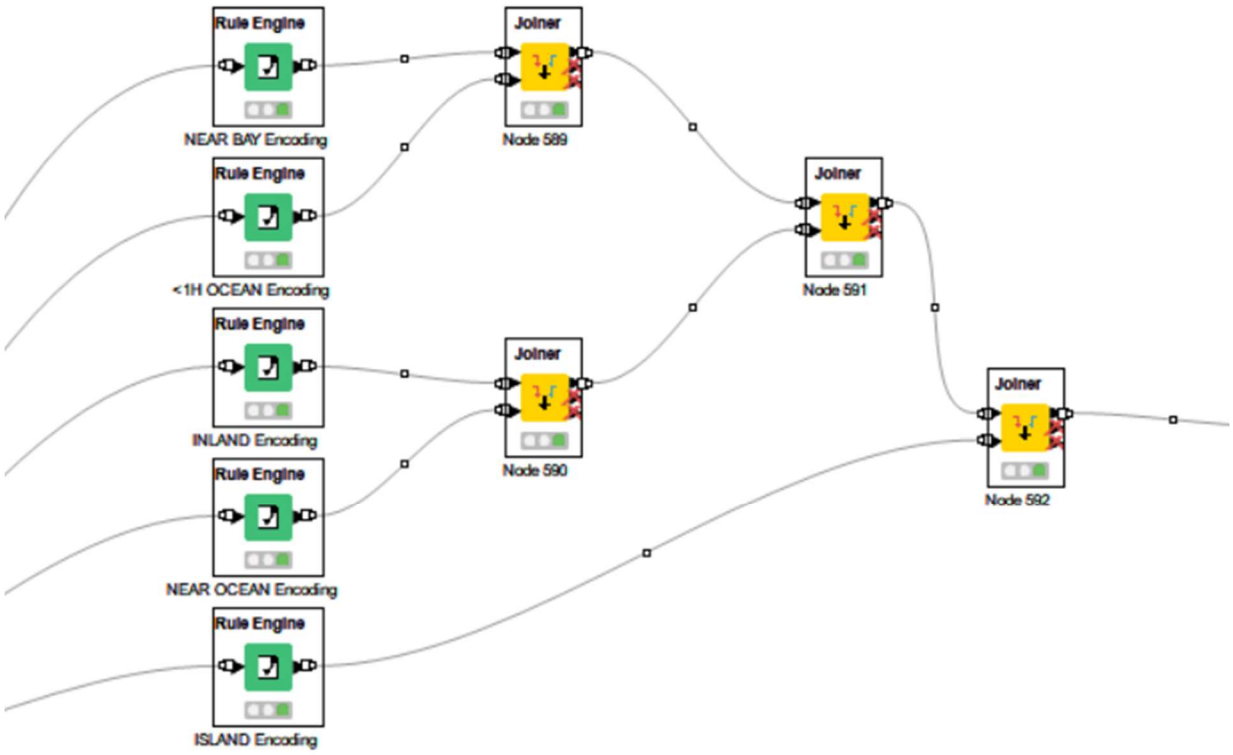


Column Joining:

The Five newly created columns need to be joined together to function as a single table.

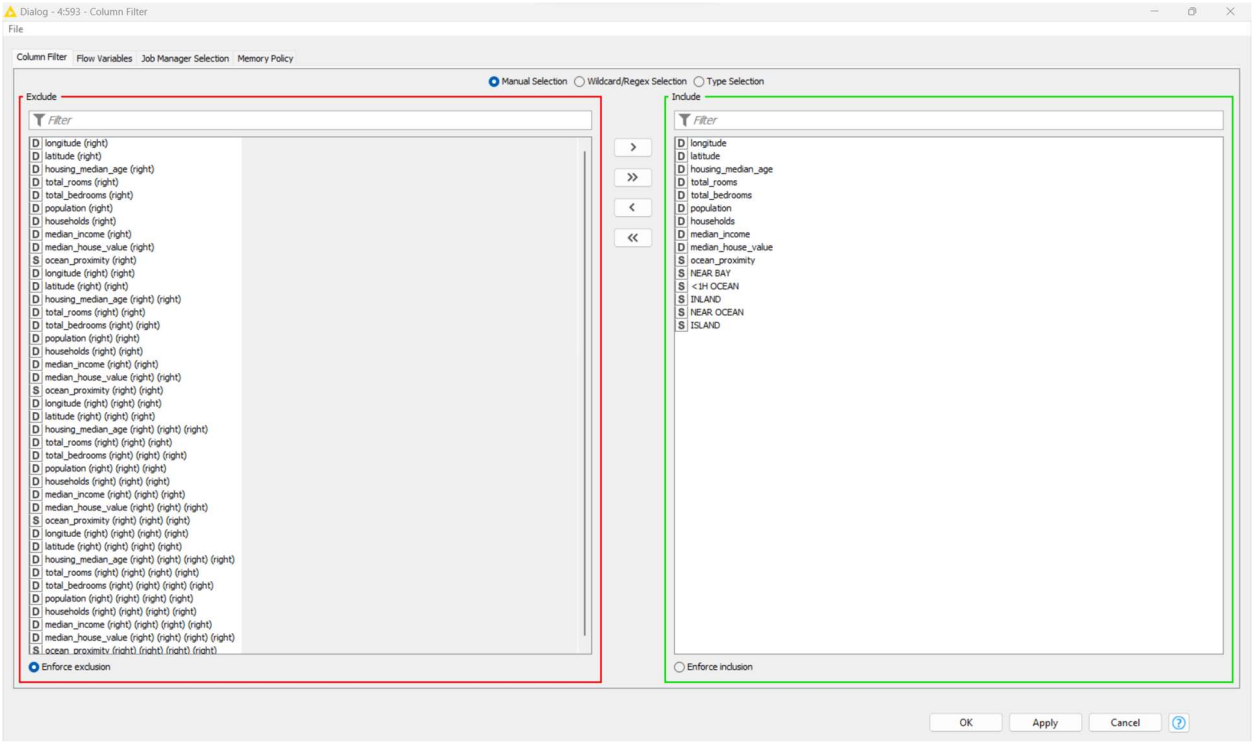
From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "Joiner" node onto the workflow canvas. Double-click on the "Joiner" node to open its configuration window. In the "Join Method" tab, select the left outer join. And select the default “Row ID” as the Primary Key (Surrogate Primary key) to join the tables. Repeat this process until all tables a joined.




Rule Engine
NEAR BAY Encoding
Joiner
Node 589
Rule Engine
<1H OCEAN Encoding
Joiner
Node 591
Rule Engine
INLAND Encoding
Joiner
Node 590
Joiner
Node 592
Rule Engine
NEAR OCEAN Encoding
Rule Engine
ISLAND Encoding

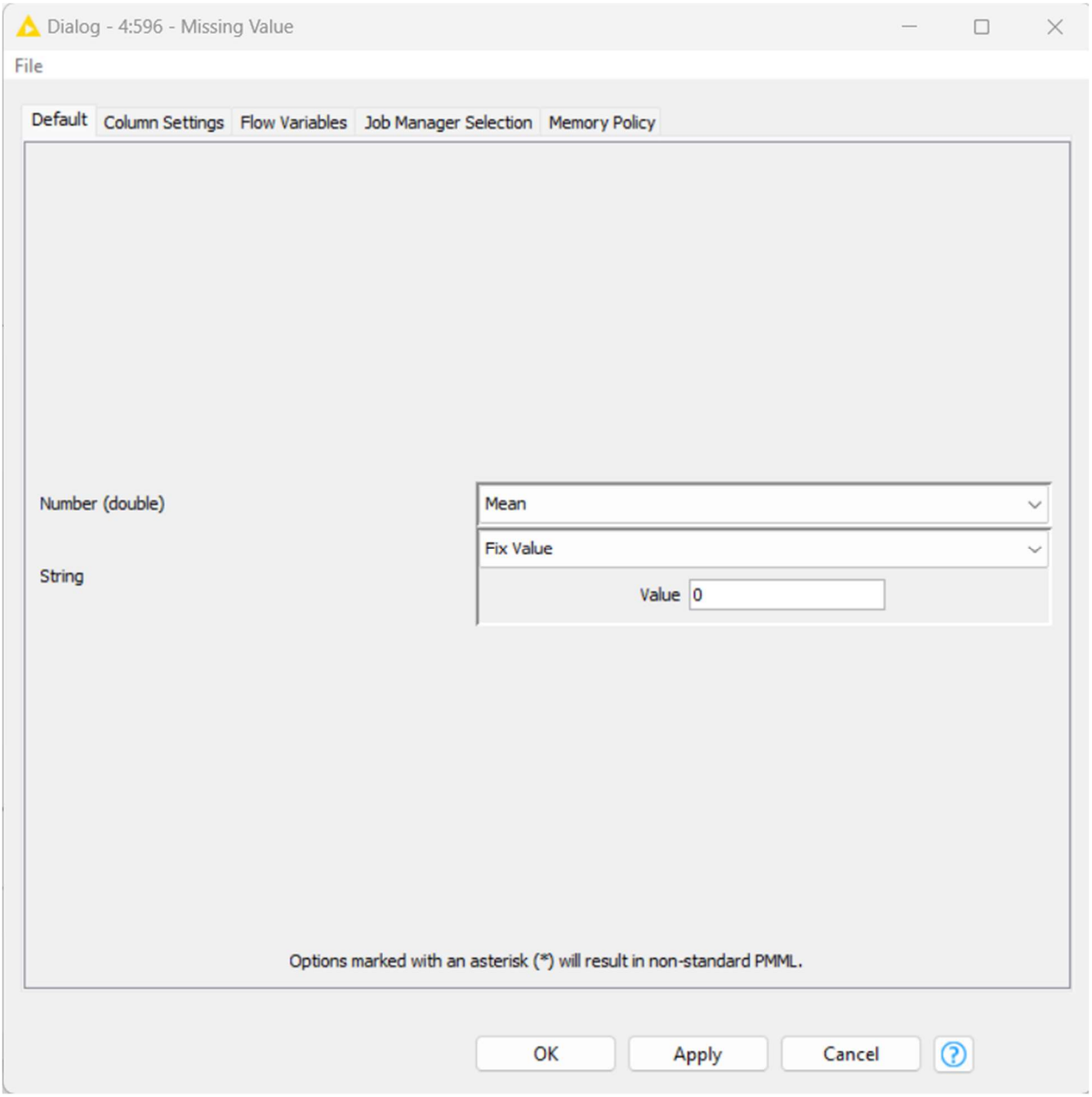
Column Filter:

Remove the duplicate columns created due to multiple joins in the above step, using Column Filter Node. From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "Column Filter" node onto the workflow canvas. Double-click on the "Column Filter" node to open its configuration window. In the "Include/Exclude" tab, select the required columns as shown below.



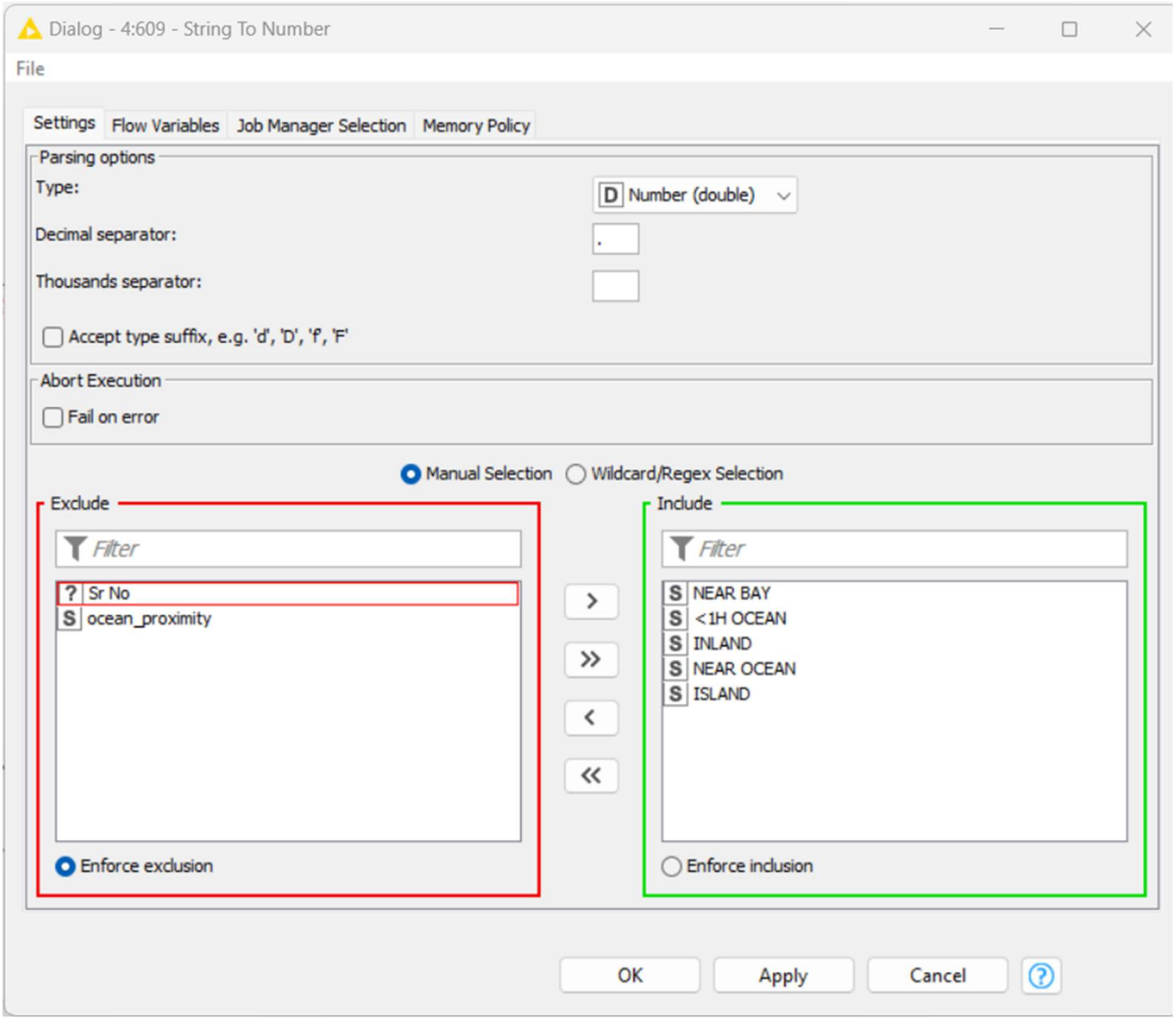
Missing Values:

Convert the null values created due to the join to “0” and all the missing num values to Avg of the column. From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "Missing Value" node onto the workflow canvas. Double-click on the "Missing Value" node to open its configuration window. In the "Missing Value Handling" tab, select the two methods to handle missing and null values as shown below. X`



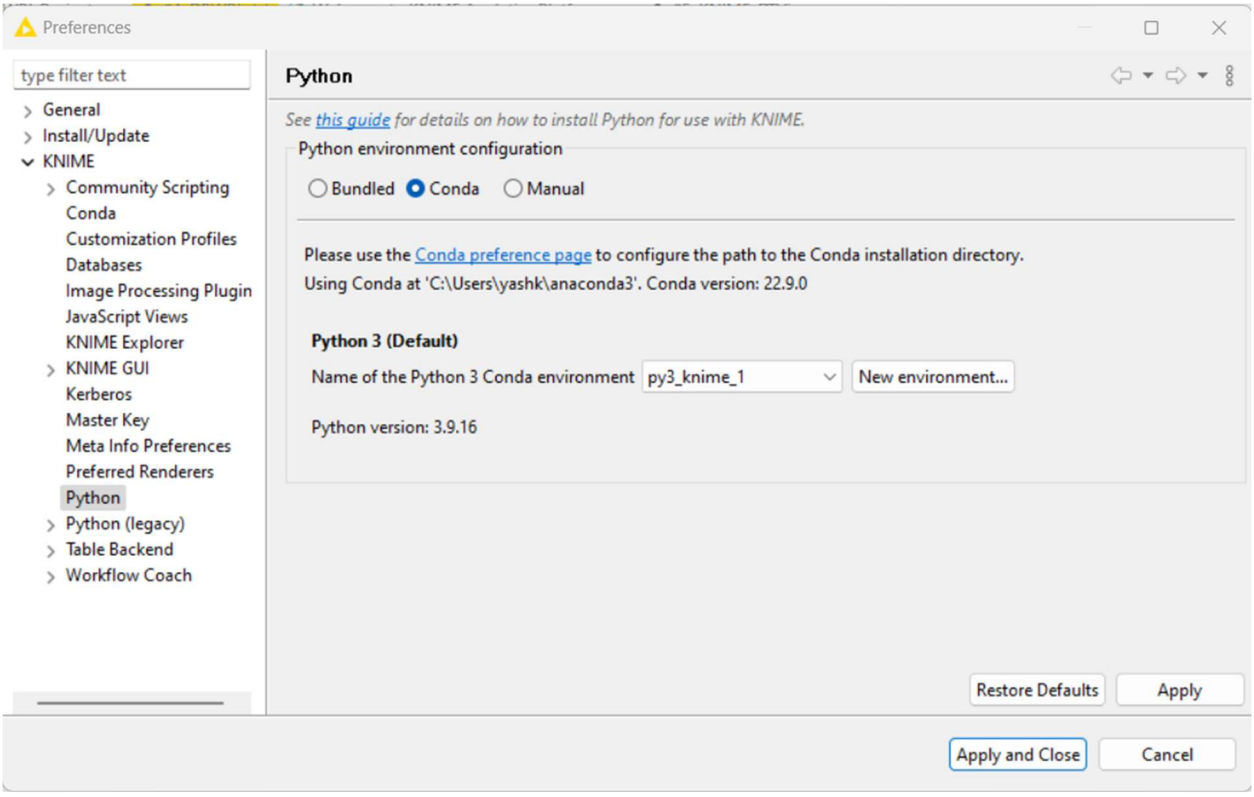
Convert string to Num:

String values created during HotEncoding need to be converted to num to fit into the regression model. From the "Node Repository" panel on the left-hand side of the screen, drag and drop the "String to Number" node onto the workflow canvas. Double-click on the "String to Number" node to open its configuration window. In the "Settings" tab, select the column or columns that you want to convert to numerical values as shown below.



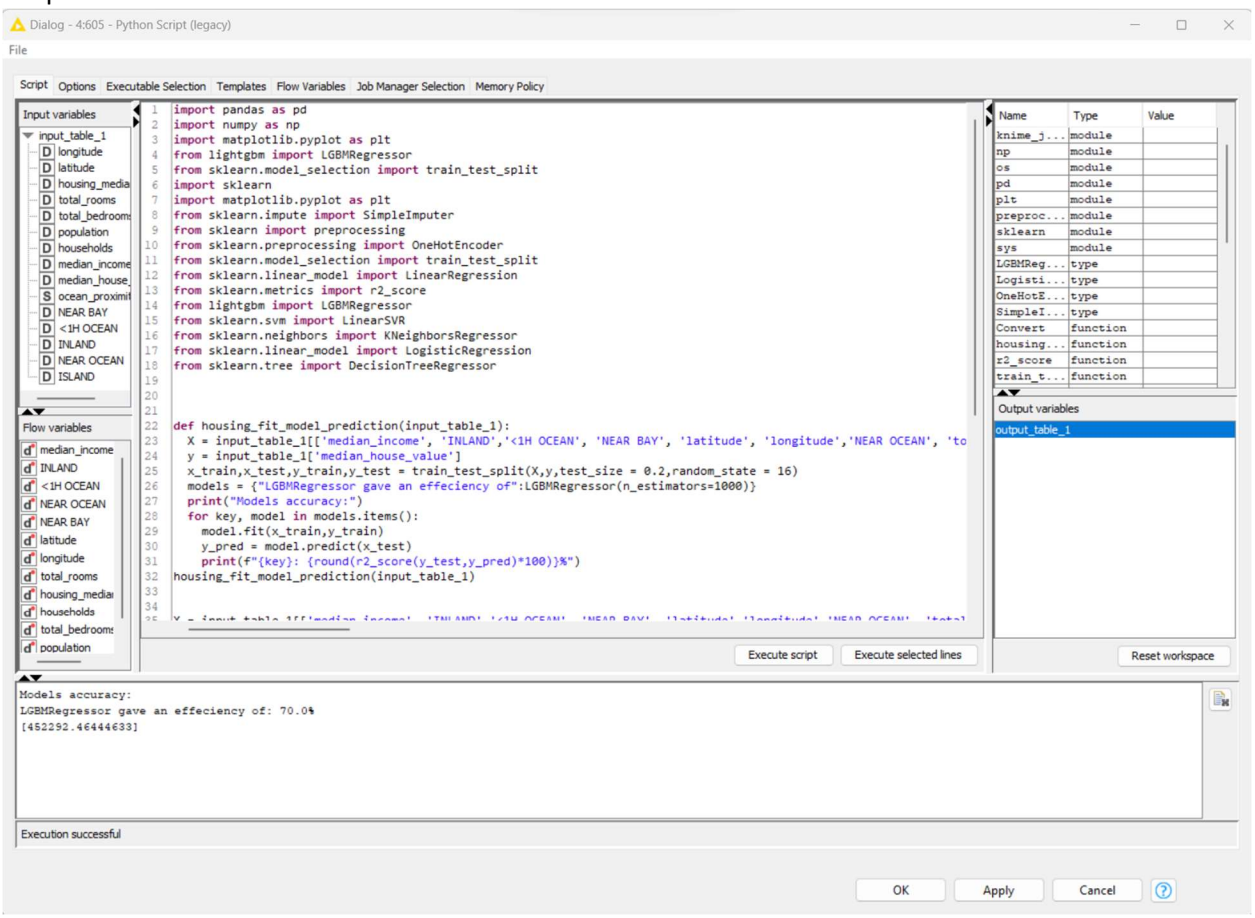
Python Script (legacy):

Note: To use the Python Script node in KNIME, we first need to have Python and the lightgbm package installed on our computer. Once this is done, set or create a new python environment through preferences to integrate KINME and Python to work together, as shown below.

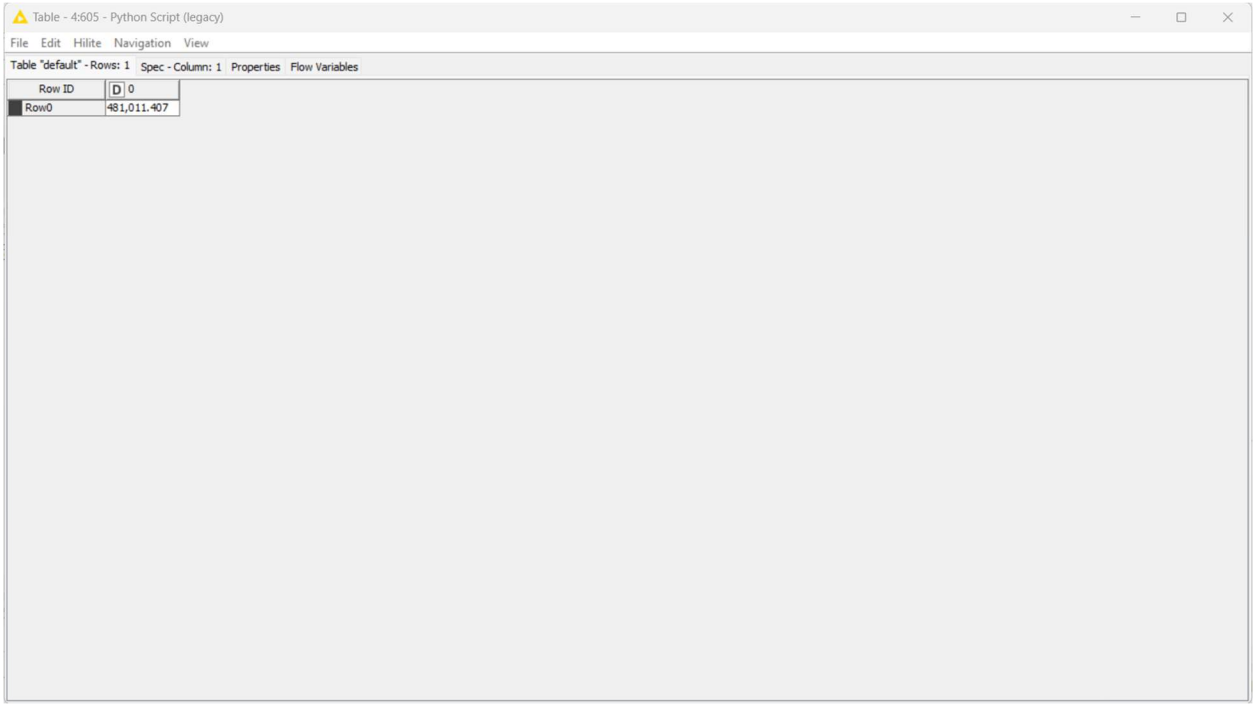


Next, drag and drop a Python Script node from the KNIME node repository onto the workflow canvas. Double-click the node to open the configuration dialogue, and then select the Input and Output ports as shown below. In the "Python Script" tab of the configuration dialogue, write the required Python code in the text editor. Once you have written your Python script, click "Execute" to run the node and generate the output table.

Script:

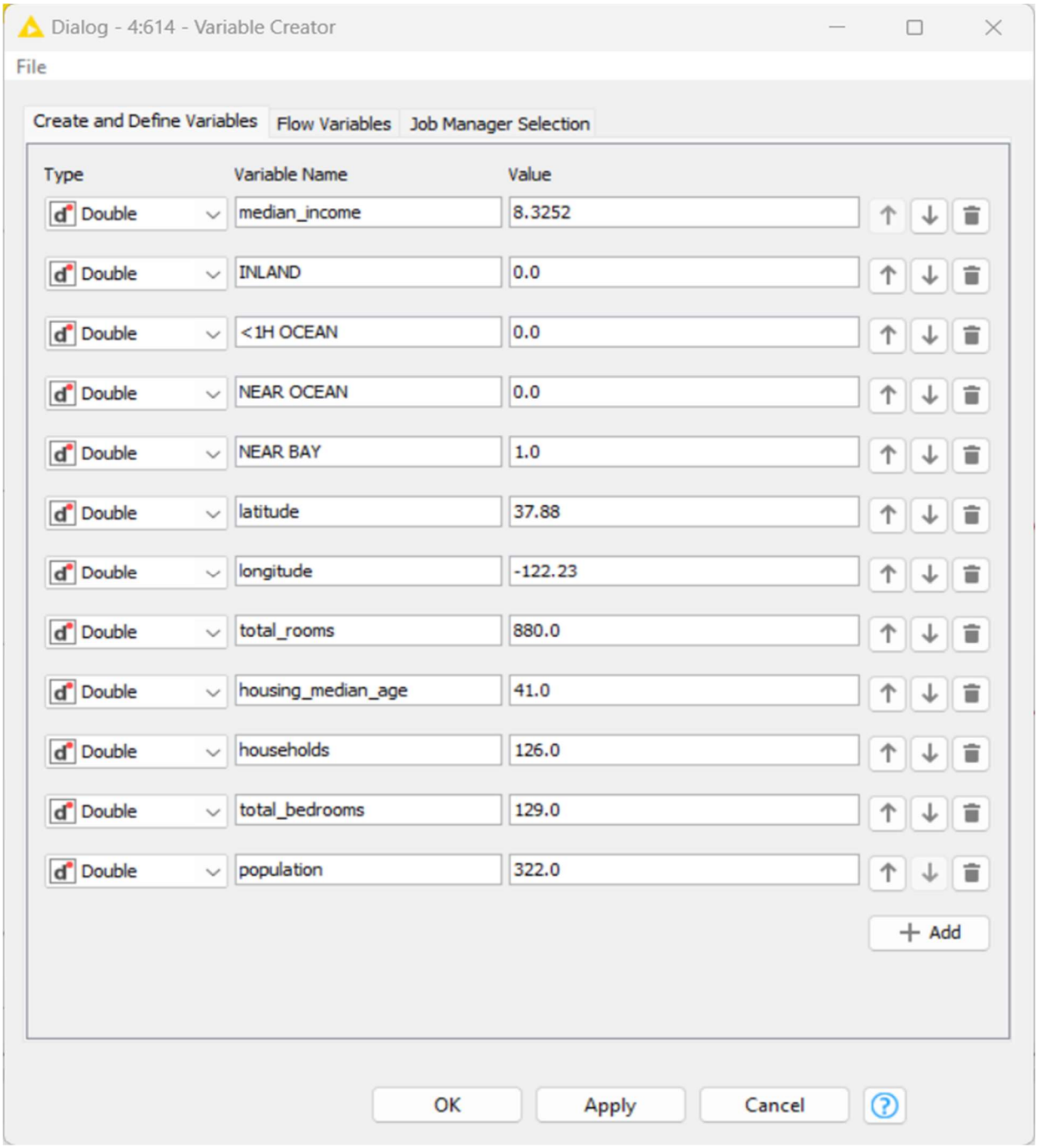


Predicted output:

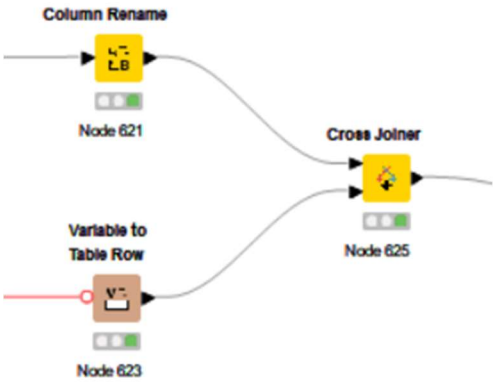


Variable Input:

In the KNIME Analytics Platform, the Variable Input node is used to define and manage flow variables within a workflow. Using this node define the values of the x-parameter for which the median house price is to be predicted. To do this, drag and drop a Variable Input node from the KNIME node repository onto the workflow canvas. Double-click the node to open the configuration dialogue, and then select the "Flow Variables" tab. In the "Flow Variables" tab, define the input variables, their data types, and values as shown below:



Generate output table:

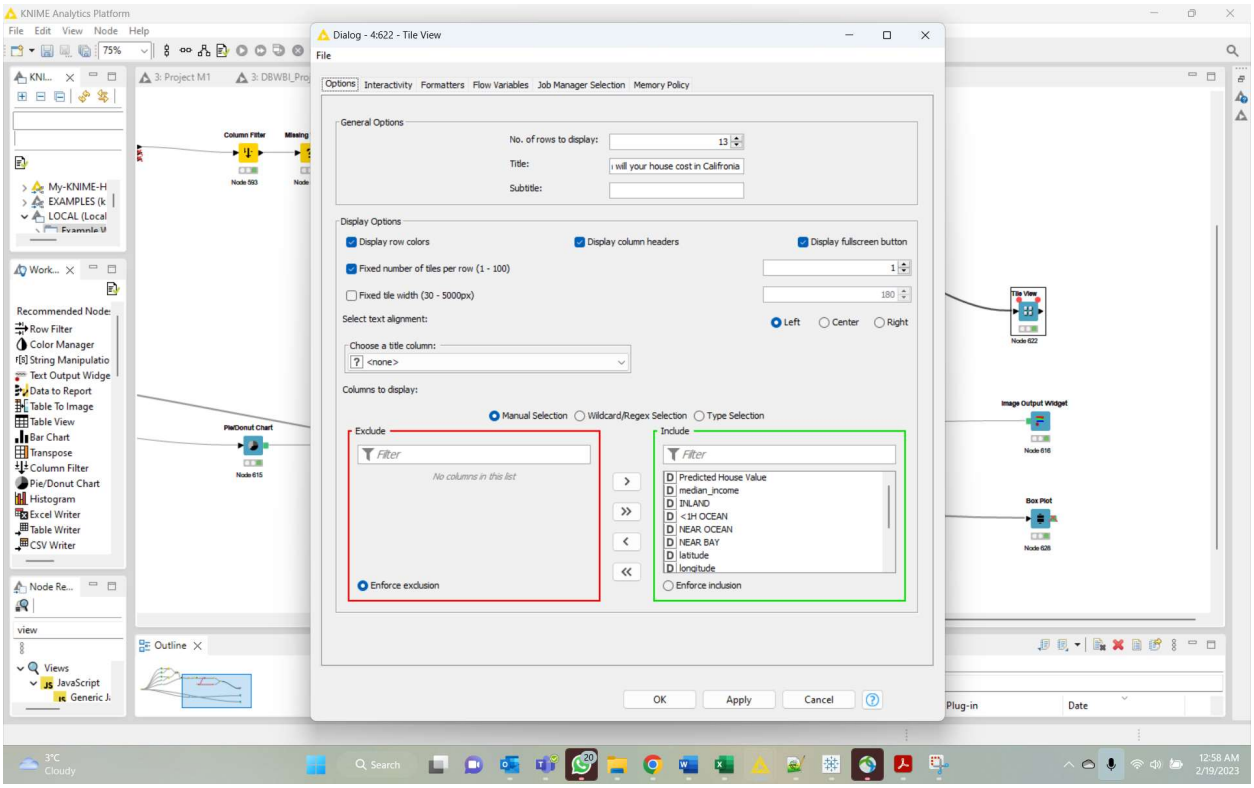


Using the combination of Column Rename, Variable to the table and Cross Joiner

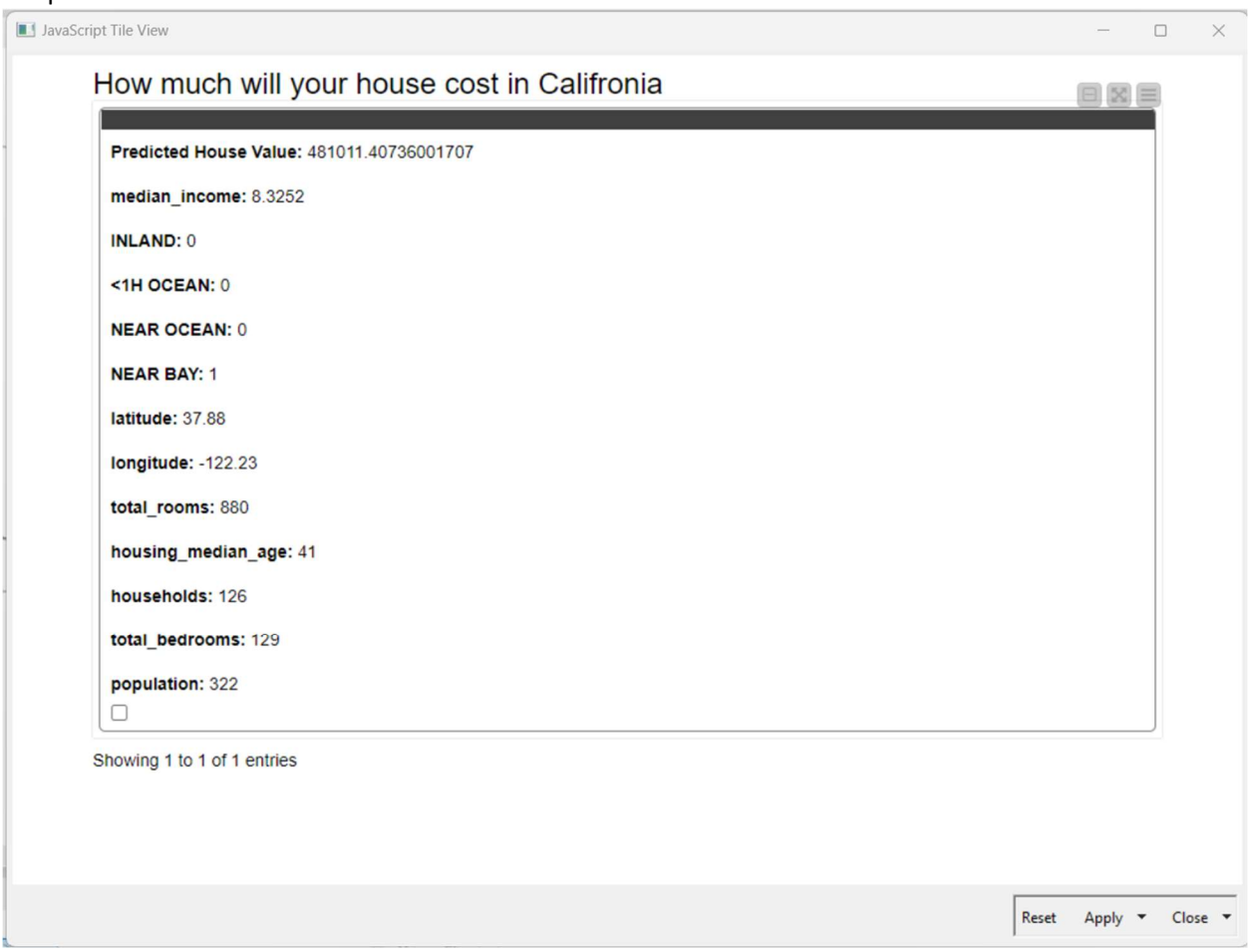
Row ID	Predicted House Value	median...	INLAND	<1H O...	NEAR ...	NEAR BAY	latitude	longitude	total_r...	housing...	househ...	total_b...	populat...
Row0_Row_0	481,011.407	8.325	0	0	0	1	37.88	-122.23	880	41	126	129	322

Data Visualization:

Tile View node is used to visualize and explore data in a grid-like format. Drag and drop a Tile View node from the KNIME node repository onto the workflow canvas. Connect the output of a data processing node to the input of the Tile View node. Double-click the Tile View node to open the configuration dialogue, and then select the columns you want to display in the tile view. Configure the node as shown below:

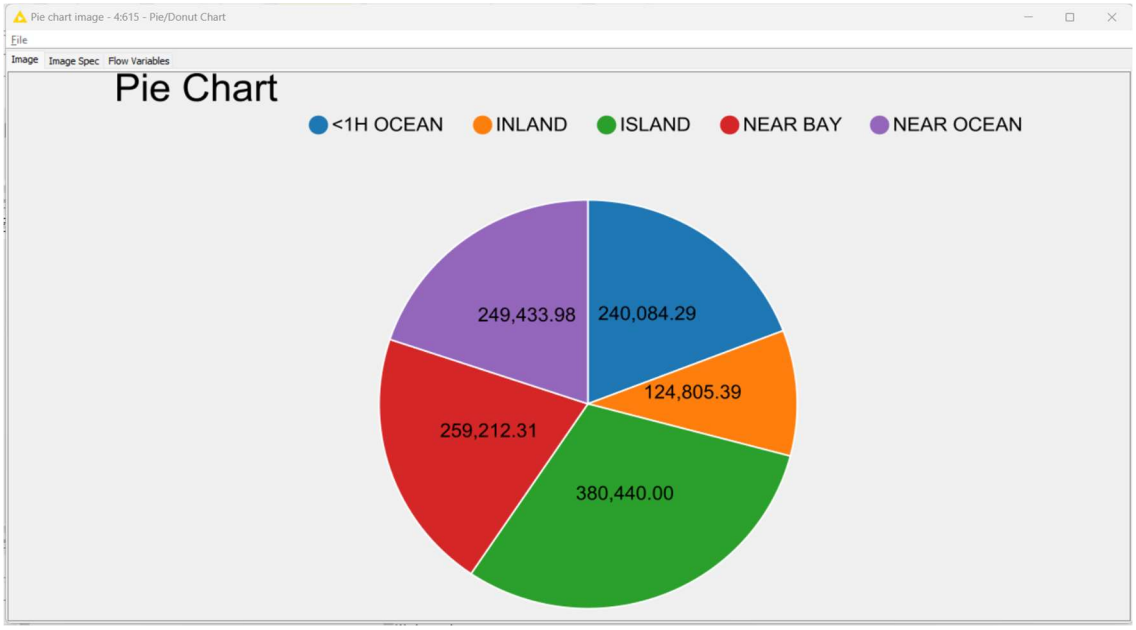


Output:


JavaScript Tile View
How much will your house cost in Califronia
Predicted House Value: 481011.40736001707
median_income: 8.3252
INLAND: 0
<1H OCEAN: 0
NEAR OCEAN: 0
NEAR BAY: 1
latitude: 37.88
longitude: -122.23
total_rooms: 880
housing_median_age: 41
households: 126
total_bedrooms: 129
population: 322
Showing 1 to 1 of 1 entries
Reset
Apply
Close

EDA:

Pie/Donut Chart



Box plot summarizing complete data set.