

Basic building blocks of KNIME for data analytics



End to End Data Science



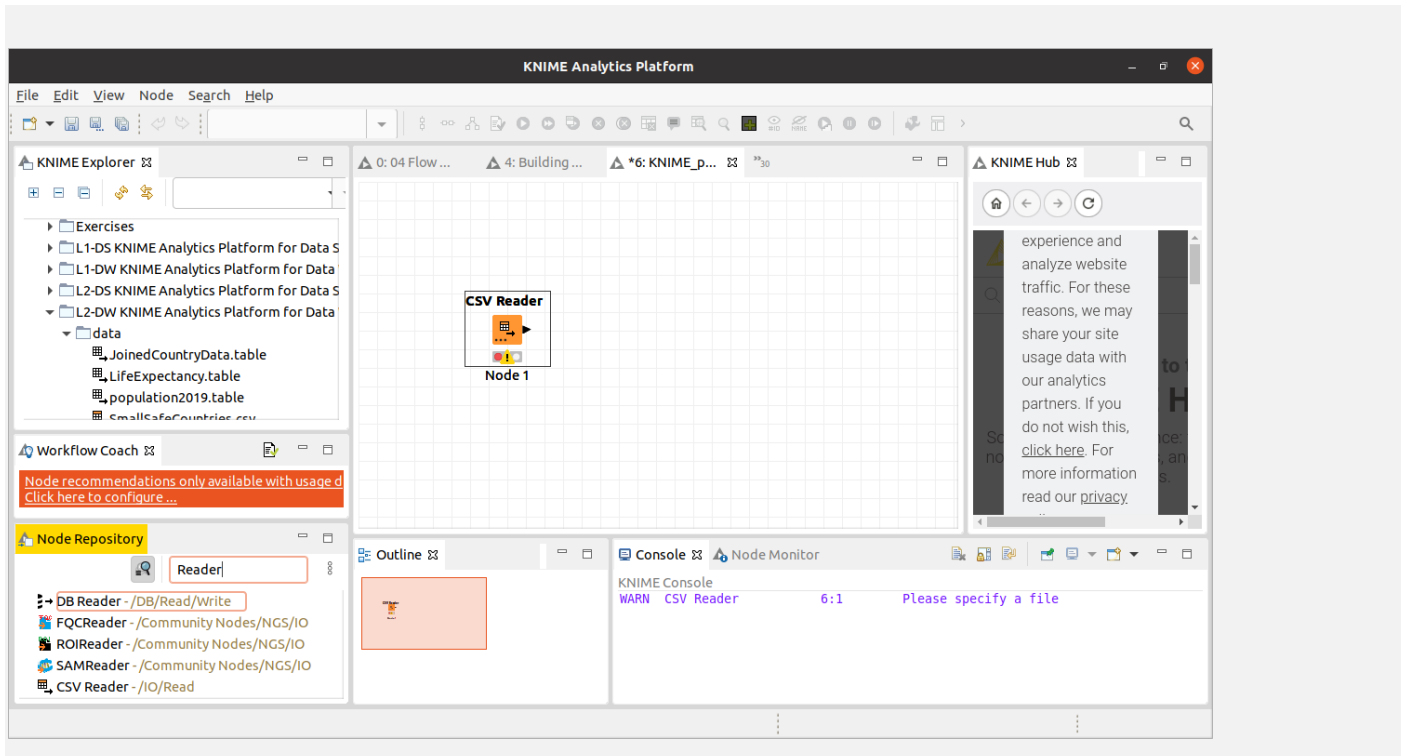
Introduction :

In this article we will talk about KNIME open-source software and how it can assist data scientists and data science enthusiasts to solve complex problems with little or no coding knowledge at all. In this article, we will get you started with KNIME as Data Analytics Platform. If you are new to KNIME, you can read the Introduction blog [here](#). To download KNIME, you can find .

I. Data Reading:

Usually, the first thing we should do when analyzing data is reading data. In 'Node Repository', we can see all kinds of Reader nodes such as CSV

Reader node, EXCEL Reader node, Table Reader node, and so on. All we need to do is simply drag and drop the node we want into 'Workflow Editor'.

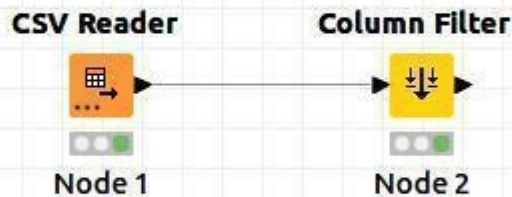


Right click the node, we can change the node's configuration; for example, we can select the path of data where we will get data from and then execute the node. If the node is executed successfully, the node will turn into Green and now we can look at the loaded data from the executed node to make sure that it has been imported properly.

II. Data Pre-processing:

1. *Filtering:*

Most of the time, we do not need all the information from our dataset. 'Row Filter' nodes and 'Column Filter' nodes help us select rows and columns that we want to use. This operation can be achieved by setting the configuration of the node in order to extract specific rows and columns we intend to use.



Dialog - 10:2 - Column Filter

File

Column Filter | Flow Variables | Job Manager Selection | Memory Policy

☒ Manual Selection ☐ Wildcard/Regex Selection ☐ Type Selection

Exclude

Filter

- ☐ Passengerid
- ☐ Survived
- ☐ Pclass
- ☐ SibSp
- ☐ Parch
- ☒ Ticket
- ☒ Fare
- ☒ Cabin
- ☒ Embarked

☒ Enforce exclusion

Include

Filter

- ☒ Name
- ☒ Sex
- ☒ Age

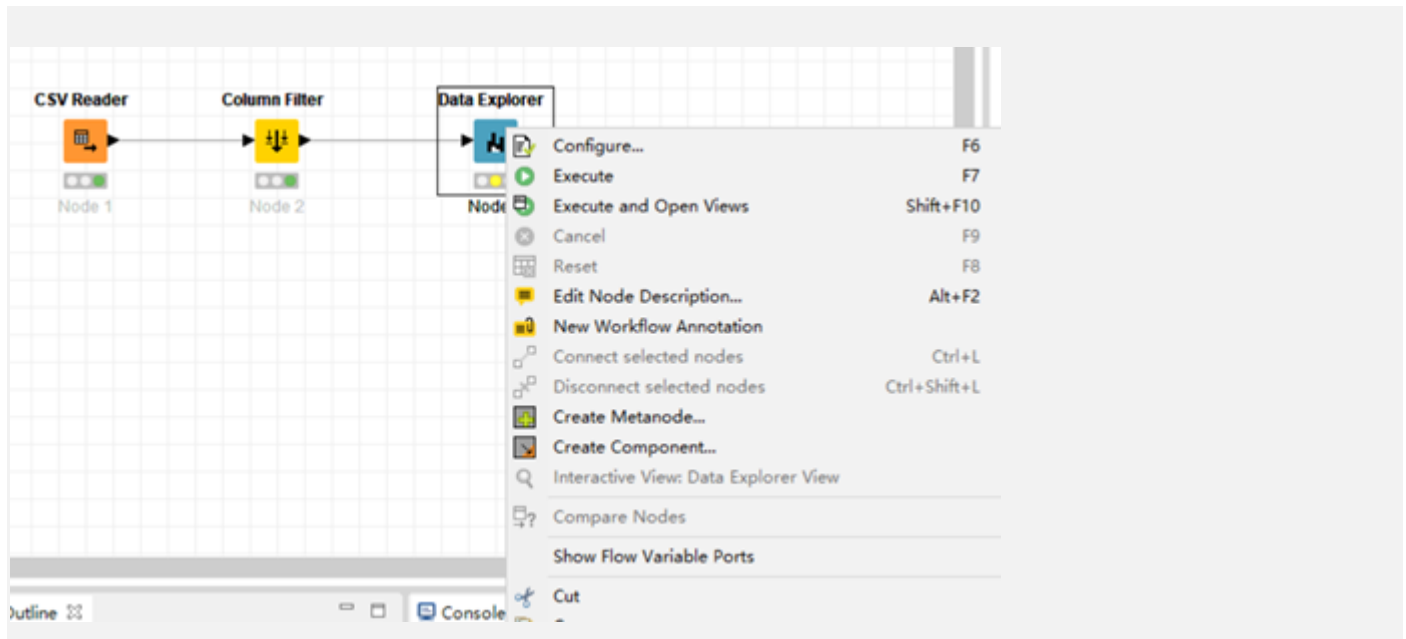
☐ Enforce inclusion

OK Apply Cancel ?

2. *Obtaining Description:*

After selecting the columns, we may want to see the description of the data; for example, we may want to get the feel of the data by looking at basic statistics of data. (i.e. minimum value, maximum value, mean value,

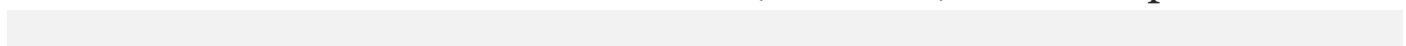
the standard deviation of our numeric data, and so on). All we need to do is to find the 'Data Explorer' node from the 'Node Repository' and drag it into 'Workflow Editor'. Later, we connect the current node ('node 2') with the recently imported node ('node 3') by connecting the 'Black Arrow' from tail to head between two nodes together. After that, we can now execute our new node by right-clicking on the 'node 3' and choosing 'Execute and Open views' option for executing our latest operation.




Now we can see the description of our data.

Column	Exclude Column	Minimum	Maximum	Mean	Standard Deviation	Variance	Skewness	Kurtosis	Overall Sum
Age	<input type="checkbox"/>	0.420	80	29.699	14.526	211.019	0.389	0.178	21205.170

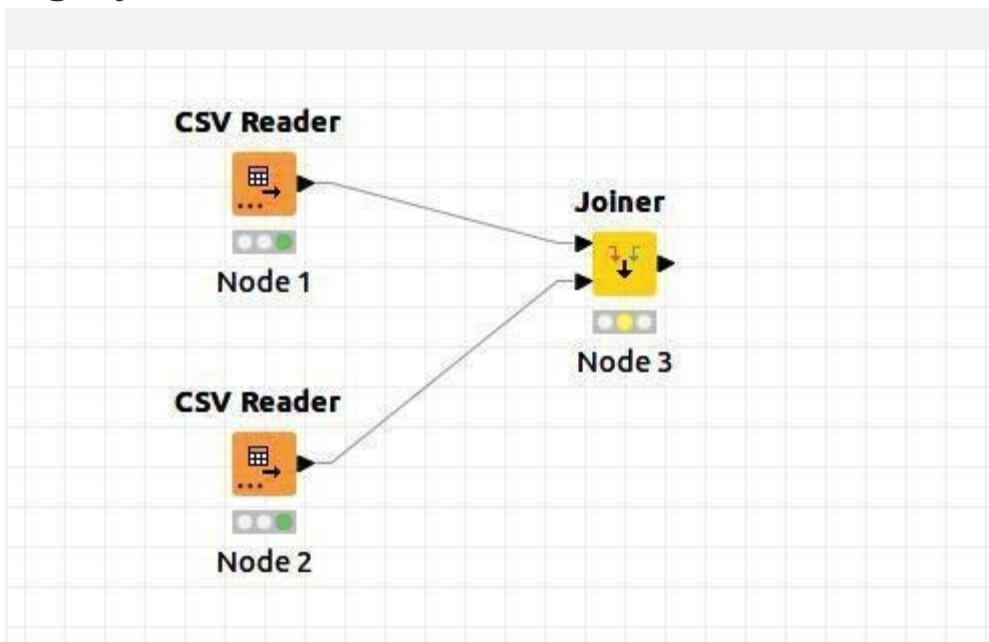
Additionally, KNIME gives us even more information. We can even see the distribution of data in each column (Bar chart) in this step.



Column	Exclude Column	No. missings	Unique values	All nominal values	Frequency Bar Chart
Name	<input type="checkbox"/>	0	891	Young, Miss. Marie Grice, Van Impe, Mr. Jean Baptiste, Danborn, Mr. Ernst Gilbert, McEvoy, Mr. Michael, Andrews, Mr. Thomas Jr, [...], Pears, Mr. Thomas Clinton, Emanuel, Miss. Virginia Ethel, Sage, Mr. George John Jr, Pavlovic, Mr. Stefo, Leitch, Miss. Jessie Wills	
Sex	<input type="checkbox"/>	0	2	male, female	

3. Combining data from multiple sources:

Sometimes, we may need to combine different datasets from various sources into one single dataset to get all necessary information we want to use. By using 'Joiner' node, we can join two datasets into one single dataset in any different joining mode such as Inner join, Left join, or Right join.





Dialog - 10:3 - Joiner

File

Joiner Settings | Column Selection | Flow Variables | Job Manager Selection | Memory Policy

Join Mode

Join mode: Inner Join

Joining Columns

☒ Match all of the following ☐ Match any of the following

Top Input ('left' table)	Bottom Input ('right' table)
S Name	S Name

Performance Tuning

Maximum number of open files: 200

☐ Enable hiliting

Row IDs

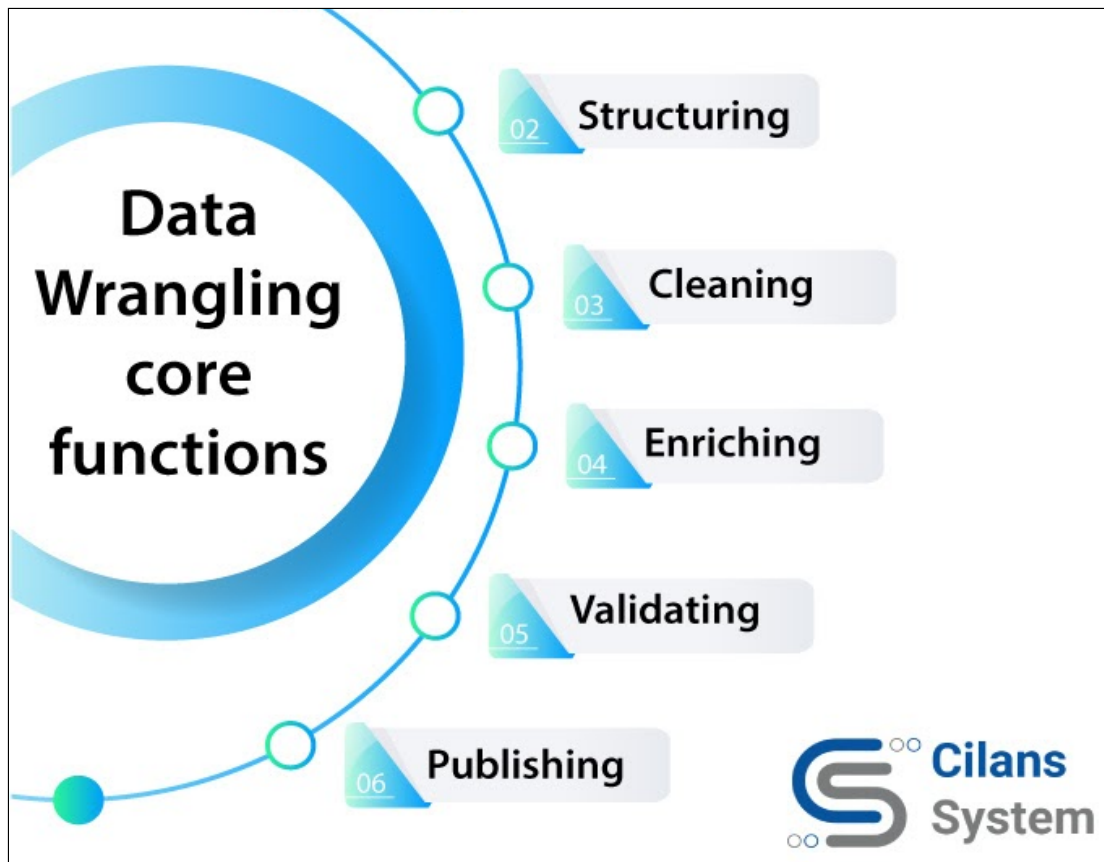
Row ID separator in joined table:

OK Apply Cancel ?

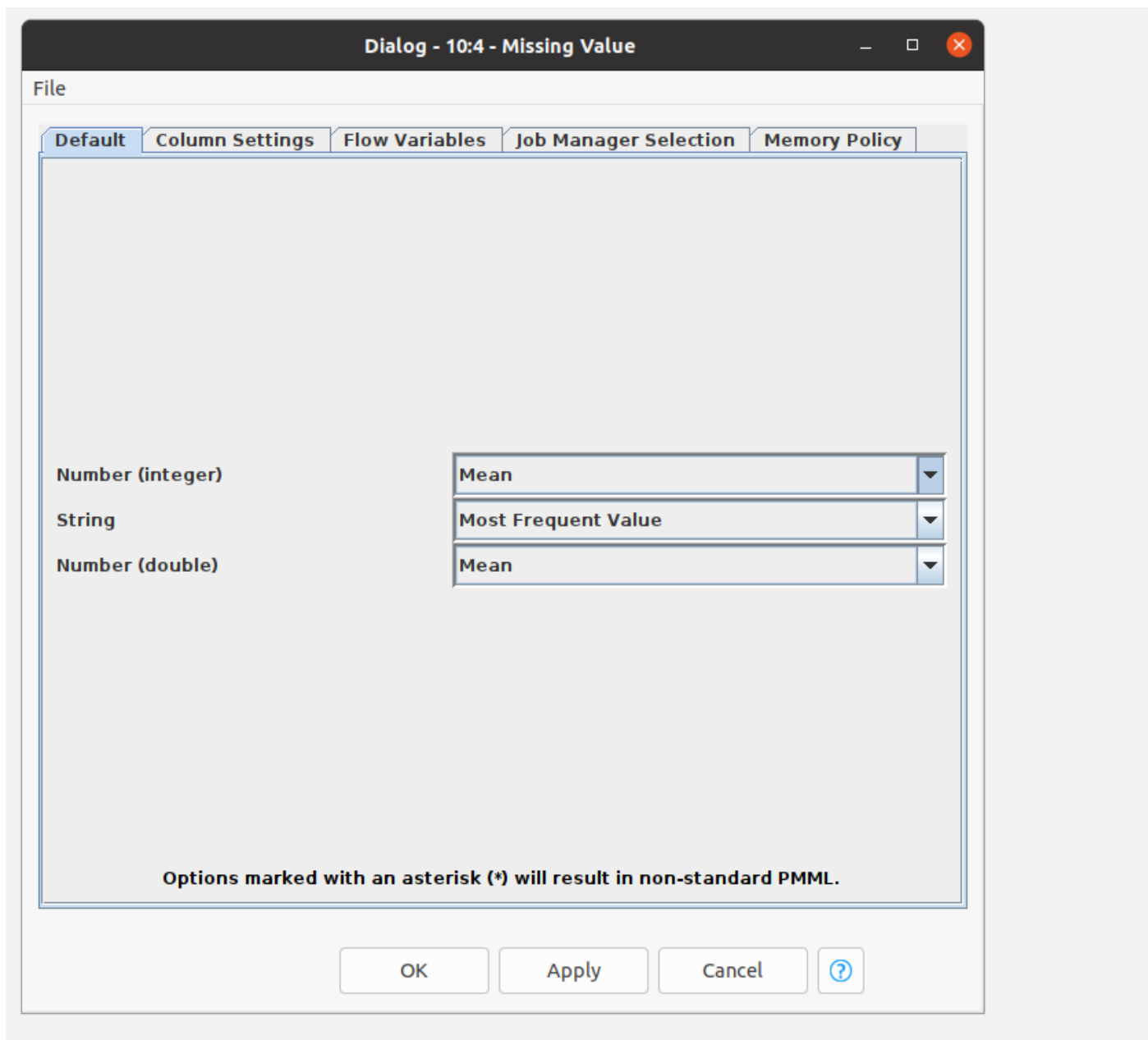
playes

4. *Removing the missing values:*

Data wrangling is an important part of Data analysis. Data cleaning plays a vital role as you can see in the diagram below



‘Missing Value’ node helps handle missing values found in cells of the input table. For example, we can replace missing values of numeric type with the mean value of that column. Similarly, the missing value of string type can also be replaced with the most frequent value occurring in that specific column.



5. *Sorting / Order :*

‘Sorter’ node helps sort the rows according to user-defined criteria. In the dialog box, we can select the columns according to which of our data should be sorted. Also, we can select how our data should be sorted in ascending or descending order.



Dialog - 10:5 - Sorter

File

Job Manager SelectionMemory Policy

Sorting FilterAdvanced SettingsFlow Variables

Sort by

D Age

↑

↓

🗑

☒ Ascending☐ Descending

Next by

S Sex

↑

↓

🗑

☒ Ascending☐ Descending

Next by

S Name

↑

↓

🗑

☒ Ascending☐ Descending

+ Add Rule

OK

Apply

Cancel

?

+91 7043122287
+91 794007058

+1 408 689 1891
+1 603 921 3957

Ahmedabad
New York

www.cilans.net
info@cilans.net



S Name	S Sex	D Age
Thomas, Master. Assad Alexander	male	0.42
Hamalainen, Master. Viljo	male	0.67
Baclini, Miss. Eugenie	female	0.75
Baclini, Miss. Helene Barbara	female	0.75
Caldwell, Master. Alden Gates	male	0.83
Richards, Master. George Sibley	male	0.83
Allison, Master. Hudson Trevor	male	0.92
Johnson, Miss. Eleanor Ileen	female	1
Nakid, Miss. Maria ("Mary")	female	1
Becker, Master. Richard F	male	1
Dean, Master. Bertram Vere	male	1
Goodwin, Master. Sidney Leonard	male	1
Mallet, Master. Andre	male	1
Panula, Master. Eino Viljami	male	1
Allison, Miss. Helen Loraine	female	2
Andersson, Miss. Ellis Anna Maria	female	2
Hirvonen, Miss. Hildur E	female	2
Quick, Miss. Phyllis May	female	2
Skoog, Miss. Margit Elizabeth	female	2
Strom, Miss. Telma Matilda	female	2
Navratil, Master. Edmond Roger	male	2
Palsson, Master. Gosta Leonard	male	2
Panula, Master. Urho Abraham	male	2

III. Model Selection and Data Analysis:

In KNIME, there are many analytic methods. In this example, we apply a Machine Learning algorithm called Random Forest to perform our analysis. We can just drag the 'Random Forest Learner' node from 'Node Repository' and drop it into our 'Workflow Editor'. Furthermore, we can set the configuration of our model node such as the number of Trees. We can now execute and train our model. After that if we want to make a prediction, we just drag the 'Random Forest Predictor' node from 'Node Repository' into the 'Workflow Editor' and execute. We can now see the prediction results.



Dialog - 10:7 - Random Forest Learner

File

Options | Flow Variables | Job Manager Selection | Memory Policy

Target Column: [S] Survived

Attribute Selection

☐ Use fingerprint attribute [no valid fingerprint input]

☒ Use column attributes

☒ Manual Selection ☐ Wildcard/Regex Selection

Exclude

Filter

[S] Name
[S] Ticket
[S] Cabin
[S] Embarked

☒ Enforce exclusion

Include

Filter

[I] PassengerId
[I] Pclass
[S] Sex
[D] Age
[I] SibSp
[I] Parch
[D] Fare

☐ Enforce inclusion

Misc Options

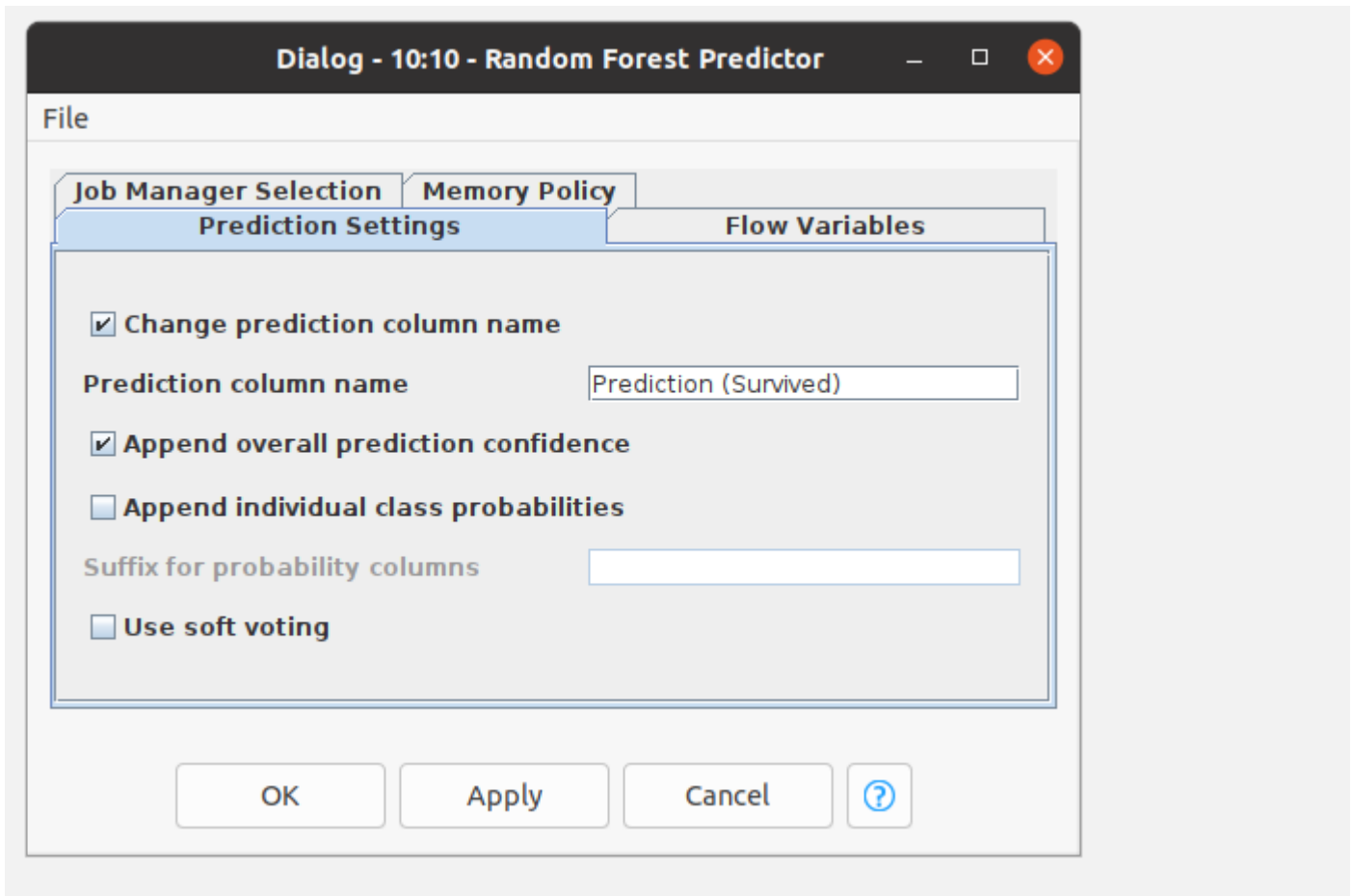
☐ Enable Hilighting (#patterns to store) 2,000

☐ Save target distribution in tree nodes (memory expensive - only important for tree view and PMML export)

Tree Options

Split Criterion: Information Gain Ratio

OK Apply Cancel ?



IV. Data Visualization:

In KNIME, there are many different kinds of plot nodes. For example, we can combine the 'Color Manager' node and 'Scatter Plot' node to customize colors and draw a scatter plot to show the distribution of age. We can select colors and choose which column will be on the x-axis and which column will be on the y-axis in the configuration dialog box.



Dialog - 10:11 - Color Manager

File

Color Settings

Flow Variables

Job Manager Selection

Memory Policy

Select one Column

S Sex

☒ Nominal

female

male

☐ Range

Preview

Palettes

Swatches

HSV

HSL

RGB

CMYK

Alpha

☒ Set 1

☐ Set 2

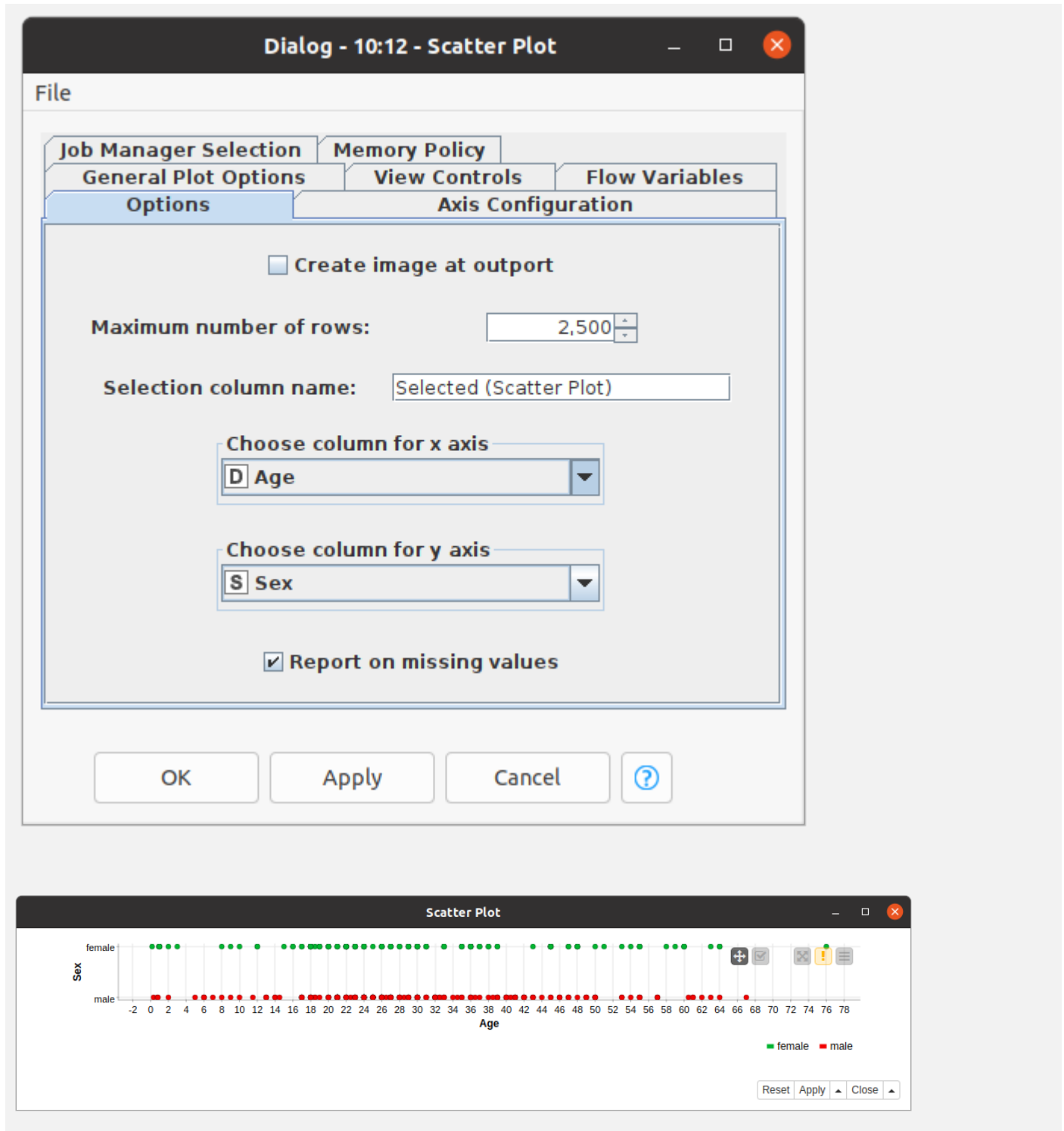
☐ Set 3 (colorblind safe)

☐ Custom

OK

Apply

Cancel



v. Conclusion

KNIME is a powerful platform which is easy to learn and use. When talking about the life cycle of Data Science, we are talking about data collection, data cleaning, data integration, analysis/modeling and



visualization. KNIME users can easily complete all of these steps in this single platform. Furthermore, as we have seen, anyone without coding background can also work on Data analysis problems. . It makes data analysis available for everyone, especially for the person who needs to analyze data only occasionally. We believe that the innovation of KNIME is beneficial to the overall Data Science community as it helps facility and introduce a powerful Analytics platform to newcomers and non-programmers.

Just try this out, and ping us if you have any queries:

Contact us on [Linkedin](#) or info@cilans.net

We will be posting more articles on Knime and Data Science in future. Check out [here](#), for future blogs.

Contact: [Team Cilans](#)