

Assignment 2

- 1) Read the adult.csv file available in the **data** folder on the KNIME Hub. The data are provided by the **UCI Machine Learning Repository**.
- 2) Calculate the average age and count for each one of the 4 groups defined by sex and income values
- 3) Join the two aggregated values to the original table

Step 1: Read the adult.csv file

The screenshot displays the KNIME workspace with a workflow consisting of a **CSV Reader** node, a **GroupBy** node, and a **Joiner** node. The **CSV Reader** node is selected, and its configuration dialog is open on the right. The dialog shows the file path and the autodetect format button. Below the dialog, a preview of the data table is visible, showing the first 10 rows of the adult.csv file. The table has 15 columns: #, RowID, age, workclass, fnlwgt, education, education..., marital-st..., occupation, relations..., race, and sex. The data is as follows:

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex
1	Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male
2	Row1	50	Self-emp-not-in	83311	Bachelors	13	Married-civ-spo	Exec-manageriz	Husband	White	Male
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleanz	Not-in-family	White	Male
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-cleanz	Husband	Black	Male
5	Row4	28	Private	338409	Bachelors	13	Married-civ-spo	Prof-specialty	Wife	Black	Female
6	Row5	37	Private	284582	Masters	14	Married-civ-spo	Exec-manageriz	Wife	White	Female
7	Row6	49	Private	160187	9th	5	Married-spouse	Other-service	Not-in-family	Black	Female
8	Row7	52	Self-emp-not-in	209642	HS-grad	9	Married-civ-spo	Exec-manageriz	Husband	White	Male
9	Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female
10	Row9	42	Private	159449	Bachelors	13	Married-civ-spo	Exec-manageriz	Husband	White	Male

Step 2: Calculate the average age and count for each one of the 4 groups defined by sex and income values

The screenshot shows the KNIME interface with the GroupBy node selected. The left sidebar contains the 'Info' tab, which provides documentation for the GroupBy node. The main workspace shows a workflow diagram with a CSV Reader node connected to a GroupBy node, which is then connected to a Joiner node. The GroupBy node's configuration dialog is open, showing the 'Manual Aggregation' tab. The output table is displayed below the dialog.

GroupBy Node Info:

Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns.

The columns to aggregate can be either defined by selecting the columns directly, by name based on a search pattern or based on the data type. Input columns are handled in this order and only considered once e.g. columns that are added directly on the 'Manual Aggregation' tab are ignored even if their name matches a search pattern on the 'Pattern Based Aggregation' tab or their type matches a defined type on the 'Type Based Aggregation' tab. The same holds for columns that are added based on a search pattern. They are ignored even if they match a criterion that has been defined in the 'Type Based Aggregation' tab.

The 'Manual Aggregation' tab allows you to change the aggregation method of more than one column. In order to do so select each column to change, open the context menu with a right mouse click and select the aggregation method to use.

In the 'Pattern Based Aggregation' tab you can assign aggregation methods to columns based on a search pattern. The pattern can be either a string with wildcards or a regular expression. Columns where the name matches the pattern but where the data type is not compatible with the selected aggregation method are ignored. Only columns that have not been selected as group column or that have not been selected as aggregation column on the 'Manual Aggregation' tab are considered.

GroupBy Node Configuration:

Group columns: sex, income

Aggregation methods:

- Mean(age): Mean (Float)
- Count*(age): Count (Integer)

Output Table:

#	RowID	sex	income	Mean(age)	Count*(age)
1	Row0	Female	<=50K	36.211	9592
2	Row1	Female	>50K	42.126	1179
3	Row2	Male	<=50K	37.147	15128
4	Row3	Male	>50K	44.626	6662

Step 3: Join the two aggregated values to the original value

The screenshot shows the KNIME interface with the Joiner node selected. The left sidebar contains the 'Info' tab, which provides documentation for the Joiner node. The main workspace shows a workflow diagram with a CSV Reader node connected to a GroupBy node, which is then connected to a Joiner node. The Joiner node's configuration dialog is open, showing the 'Matching Criteria' tab. The output table is displayed below the dialog.

Joiner Node Info:

This node combines two tables similar to a join in a database. It combines each row from the top input port with each row from the bottom input port that has identical values in selected columns. Rows that remain unmatched can also be output.

Joiner Node Configuration:

Matching Criteria:

- Match: All of the following
- Compare values in join columns by: Value and type

Output Table:

sex	capital-g...	capital-lo...	hours-per...	native-co...	income	sex (Right)	income (...)	Mean(age)	Count*(a...
Male	2174	0	40	United-States	<=50K	Female	<=50K	36.211	9592
Male	0	0	13	United-States	<=50K	Female	>50K	42.126	1179
Male	0	0	40	United-States	<=50K	Male	<=50K	37.147	15128
Male	0	0	40	United-States	<=50K	Male	>50K	44.626	6662