

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 software interface. On the left, the 'CSV Reader' node is selected, showing its configuration options. The main workspace shows a flow diagram with a 'CSV Reader' node connected to a 'Joiner' node. The 'Joiner' node is also connected to a 'GroupBy' node. The 'GroupBy' node is configured with 'age' as the aggregation function and 'sex' and 'income' as the grouping variables. The 'Joiner' node is configured to join the 'CSV Reader' node and the 'GroupBy' node. The 'Joiner' node is also connected to a 'GroupBy' node. The 'Joiner' node is configured with 'age' as the aggregation function and 'sex' and 'income' as the grouping variables. The 'Joiner' node is also connected to a 'GroupBy' node. The 'Joiner' node is configured with 'age' as the aggregation function and 'sex' and 'income' as the grouping variables.

CSV Reader

Reads CSV files. To auto-guess the structure of the file click the Autodetect format button. If you encounter problems with incorrect guessed data types disable the Limit data rows scanned option in the Advanced Settings tab. If the input file structure changes between different invocations, enable the Support changing file schemas option in the Advanced Settings tab. For further details see the KNIME File Handling Guide [File Handling Guide](#).

Note: If you find that this node can't read your file, try the **File Reader** node. It offers more options for reading complex files.

This node can access a variety of different [file systems](#). More information about file handling in KNIME can be found in the official [File Handling Guide](#).

Parallel reading: Individual files can be read in parallel if

- They are located on the machine that is running this node.
- They don't contain any quotes that contain row delimiters.
- They are not gzip compressed.
- No lines or rows are limited or skipped.
- The file index is not prepended to the RowID.
- They are not encoded with UTF-16 (UTF-16LE and UTF-16BE are fine).

Ports Options Views

Output ports

1: File Table

Flow Variables

#	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-manageri	Husband	White	Male
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-clean	Not-in-family	White	Male
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-clean	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-manageri	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-manageri	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-manageri	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 a workflow containing a CSV Reader, a GroupBy node, and a Joiner node. The GroupBy node is configured to aggregate data by sex and income. The output table is displayed below the node.

#	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 a workflow containing a CSV Reader, a GroupBy node, and a Joiner node. The Joiner node is configured to join the two aggregated tables back to the original data. The output table is displayed below the node.

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