

# 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 shows a KNIME workflow titled "Local – Assignment 2". The workflow consists of the following nodes and connections:

- CSV Reader**: This node is connected to a **Joiner** node.
- GroupBy**: This node is also connected to the **Joiner** node.
- Joiner**: This node has two inputs: one from the **GroupBy** node and one from another **CSV Reader** node.
- Table View**: This view displays the data from the joined tables, showing 32561 rows and 15 columns.

The **CSV Reader** node has a detailed description in its help panel:

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).

## 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 'GroupBy' node selected. The left panel displays the 'Info' tab of the 'GroupBy' node, which explains how it groups rows by unique values in specified columns and aggregates remaining columns. The main workspace shows a flow starting from a 'CSV Reader' node, followed by a 'GroupBy' node, and then a 'Joiner' node. The right panel shows the 'GroupBy' node dialog, which is noted as unsupported.

#	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 'Joiner' node selected. The left panel displays the 'Info' tab of the 'Joiner' node, explaining it combines two tables similar to a database join. The main workspace shows a flow starting from a 'CSV Reader' node, followed by a 'GroupBy' node, and then a 'Joiner' node. The right panel shows the 'Joiner' node dialog, which includes 'Matching Criteria' and 'Compare values in join columns by' options. The results table shows the joined data with additional columns for unmatched rows.

#	sex	capital-g...	capital-lo...	hours-per...	native-co...	income	sex (Right)	income (...)	Mean(age)	Count(age)
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