

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 three main nodes: a "CSV Reader" node on the left, a "Joiner" node on the right, and a "GroupBy" node in the center. Arrows indicate data flow from the CSV Reader to the GroupBy node, and from the GroupBy node to the Joiner node. The "GroupBy" node has a "Count" output port connected to the Joiner. The "Joiner" node has an "Inner Join" output port connected back to the GroupBy node. The "Joiner" node also has an "Outer Join" output port, which is connected to a second "CSV Reader" node on the right. This second "CSV Reader" node is shown with a tooltip: "This node dialog is not supported here." Below the nodes is a "File Table" view displaying the first 10 rows of the adult.csv dataset. The columns listed are RowID, #, RowID, age, workclass, fnlwgt, education, education, marital-st..., occupation, relations..., race, and sex.

#	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-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaner	Not-in-family	White	Male
4	Row3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaner	Husband	Black	Male
5	Row4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female
6	Row5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	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-inc	209642	HS-grad	9	Married-civ-spouse	Exec-managerial	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-spouse	Exec-managerial	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 following components:

- GroupBy Node Info:** Describes the node's function: "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." It also details how to define columns to aggregate.
- GroupBy Node Execution:** A flow diagram showing a CSV Reader node connected to a GroupBy node, which then connects to a Joiner node. The GroupBy node has a "Add comment" field.
- GroupBy Node Results:** A table titled "1: Group table" showing the aggregated data. The columns are RowID, #, RowID, sex, income, Mean(age), and Count(age). The data shows four groups: Female <=50K (Row0), Female >50K (Row1), Male <=50K (Row2), and Male >50K (Row3).

#	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 following components:

- Joiner Node Info:** Describes the node's function: "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." It includes external resources and port definitions.
- Joiner Node Execution:** A flow diagram showing a CSV Reader node connected to a GroupBy node, which then connects to a Joiner node. The Joiner node has a "Matching Criteria" dialog open, set to "All of the following".
- Joiner Node Results:** A table titled "1: Join result" showing the joined data. The columns are the same as the GroupBy table, plus additional columns from the original CSV: capital-gdp, capital-income, hours-per-week, native-country, income, sex (Right), income (>=50K), Mean(age), and Count(age). The data shows the same four groups as the previous step, now joined with their respective capital and hours data.

#	RowID	sex	capital-gdp	capital-income	hours-per-week	native-country	income	sex (Right)	income (>=50K)	Mean(age)	Count(age)
1	Row0	Female	2174	0	40	United-States	<=50K	Female	<=50K	36.211	9592
2	Row1	Female	0	0	13	United-States	<=50K	Female	>50K	42.126	1179
3	Row2	Male	0	0	40	United-States	<=50K	Male	<=50K	37.147	15128
4	Row3	Male	0	0	40	United-States	<=50K	Male	>50K	44.626	6662