

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 the KNIME workspace with a workflow containing a **CSV Reader** node and a **Joiner** node. The **CSV Reader** node is configured to read the **adult.csv** file. The **Joiner** node is connected to the **CSV Reader** node and a **GroupBy** node. The **GroupBy** node is configured to group by **sex** and **income**. The **Joiner** node is configured to join the **CSV Reader** node and the **GroupBy** node. The **Joiner** node is also connected to the **GroupBy** node. The **Joiner** node is also connected to the **GroupBy** node.

The **CSV Reader** node dialog is open, showing the file path and the autodetected schema. The schema includes columns for **RowID**, **age**, **workclass**, **fnlwgt**, **education**, **education-num**, **marital-status**, **occupation**, **relations**, **race**, and **sex**.

The **Joiner** node dialog is open, showing the join type and the join columns. The join type is set to **Inner Join** and the join columns are **sex** and **income**.

The **GroupBy** node dialog is open, showing the group by columns and the aggregation functions. The group by columns are **sex** and **income**, and the aggregation functions are **Average** and **Count**.

The **Table** view shows the resulting data table with 10 rows and 15 columns. The columns are **#**, **RowID**, **age**, **workclass**, **fnlwgt**, **education**, **education-num**, **marital-status**, **occupation**, **relations**, **race**, **sex**, **income**, **age_avg**, and **count**.

#	RowID	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relations	race	sex	income	age_avg	count
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-cleani	Not-in-family	White	Male			
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-cleani	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 details about the GroupBy node's functionality. The main workspace displays 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, titled '1: Group table', contains 4 rows and 6 columns: #, RowID, sex, income, Mean(age), and Count(age). The data is as follows:

#	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 details about the Joiner node's functionality. The main workspace displays 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, titled '1: Join result', contains 4 rows and 19 columns. The data is as follows:

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