

Assignment 01

- 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 count and average age of women with income >50K
- 3) Calculate the averages of all numerical columns for each one of the 4 groups defined by sex and income values
- 4) Calculate

- the number of missing values in the occupation column
- the number of non-missing rows in the occupation column
- the number of rows in the occupation column
- the number of rows in the marital-status column

Notice that the last two aggregations should provide the same numbers!

Step 1: Read CSV File “adult.csv”

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-lo...	hours-per...
1	Row0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40
2	Row1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaner	Not-in-family	White	Male	0	0	40
4	Row3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaner	Husband	Black	Male	0	0	40
5	Row4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40
6	Row5	37	Private	284582	Masters	14	Married-civ-spouse	Exec-managerial	Wife	White	Female	0	0	40
7	Row6	49	Private	160187	9th	5	Married-spouse-absent	Other-service	Not-in-family	Black	Female	0	0	16
8	Row7	50	Govt-work	451321	HS-grad	0	Married-spouse-absent	Adm-clerical	Not-in-family	White	Male	0	0	40
9	Row8	40	Private	434374	HS-grad	0	Married-spouse-absent	Handlers-cleaner	Not-in-family	White	Male	0	0	40
10	Row9	39	Private	398906	HS-grad	0	Married-spouse-absent	Prof-specialty	Not-in-family	White	Male	0	0	40

Power BI and KNIME

Step 2: Filter Row for Women with income >50K

The screenshot shows a KNIME workflow titled "Local - Assignment 1". The process starts with a "CSV Reader" node, followed by a "GroupBy" node. Then, the output goes through a "Row Filter" node, another "GroupBy" node, and finally a third "GroupBy" node. A "Table" view at the bottom displays the filtered data, which includes columns such as workclass, fnlwgt, education, occupation, sex, and income. The data shows several rows where the income is greater than 50K.

workclass	fnlwgt	education	occupation	sex	income
Private	45781	Masters	Prof-specialty	Female	>50K
Self-emp-not-inc	291275	Masters	Divorced	Female	>50K
Private	51835	Prof-school	Exec-managerial	Female	>50K
Private	169846	HS-grad	Married-civ-spouse	Female	>50K
Private	343591	HS-grad	Adm-clerical	Female	>50K
Federal-gov	410867	Doctorate	Craft-repair	Female	>50K
Private	287828	Bachelors	Never-married	Female	>50K

Step 3: Use GroupBy node to calculate the count and average age of women with income >50K

The screenshot shows a continuation of the KNIME workflow. The process starts with a "CSV Reader" node, followed by a "GroupBy" node. Then, the output goes through a "Row Filter" node and another "GroupBy" node. A "Group table" view at the bottom displays the results of the grouping, showing the count and mean age for women with an income greater than 50K.

#	RowID	Count*(age)	Mean(age)
1	Row0	1179	42.126

Power BI and KNIME

Step 4: Use GroupBy node to calculate the average of all numerical column for each of the 4- group defined by sex and income value

The screenshot shows a KNIME workflow interface. On the left, there's a tree view for 'Nodes' (CSV Reader, Row Filter, GroupBy), 'Explorer', and 'Monitor'. The main workspace contains a 'Group table' view with the following data:

#	RowID	sex	income	Mean(age)	Mean(capital-gain)	Mean(capital-loss)	Mean(education-num)	Mean(hours-per-week)
1	Row0	Female	<=50K	36.211	121.986	47.364	9.82	35.917
2	Row1	Female	>50K	42.126	4,200.389	173.649	11.787	40.427
3	Row2	Male	<=50K	37.147	165.724	56.807	9.452	40.694
4	Row3	Male	>50K	44.626	3,971.766	198.78	11.581	46.366

A 'GroupBy' node is highlighted in blue, indicating it's selected. A tooltip says 'This node dialog is not supported here.' and has a 'Open dialog' button.

Step 5: Use GroupBy node to calculate Missing value count for occupation, non-missing value count for occupation, no of rows in occupation column, no of rows in marital-status

The screenshot shows a KNIME workflow interface similar to the previous one. The 'Group table' view now displays the following data:

#	RowID	Missing value count(occupation)	Count*(occupation)	Count(occupation)	Count(marital-status)
1	Row0	0	32561	32561	32561

A 'GroupBy' node is highlighted in blue. A tooltip says 'This node dialog is not supported here.' and has a 'Open dialog' button.