

Power BI and KNIME Assignment 1

- 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”

The screenshot displays the KNIME software interface. On the left, a sidebar contains icons for 'Info', 'Nodes', 'Execute', 'Cancel', and 'Reset'. The main workspace shows a flowchart with a 'CSV Reader' node connected to three 'GroupBy' nodes. A 'Row Filter' node is also present. The 'CSV Reader' node's dialog box is open on the right, showing a message: 'This node dialog is not supported here.' Below the flowchart, a table preview is visible, showing the first 15 rows of the 'adult.csv' file. The table has 15 columns: #, RowID, age, workclass, fnlwgt, education, education..., marital-st..., occupation, relations..., race, sex, capital-g..., capital-to..., and hours-per-... The data rows show various attributes for 15 different individuals.

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-to...	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-spo	Exec-managerial	Husband	White	Male	0	0	13
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-clean	Not-in-family	White	Male	0	0	40
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-clean	Husband	Black	Male	0	0	40
5	Row4	28	Private	338409	Bachelors	13	Married-civ-spo	Prof-specialty	Wife	Black	Female	0	0	40
6	Row5	37	Private	284582	Masters	14	Married-civ-spo	Exec-managerial	Wife	White	Female	0	0	40
7	Row6	49	Private	160187	9th	5	Married-spouse	Other-service	Not-in-family	Black	Female	0	0	16

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

The screenshot shows a KNIME workflow with the following nodes: CSV Reader, Row Filter, and two GroupBy nodes. The Row Filter node is configured with the filter: sex = Female AND income > 50K. The first GroupBy node is configured with 'sex' as the grouping variable. The second GroupBy node is configured with 'income' as the grouping variable. The data table below shows the filtered rows.

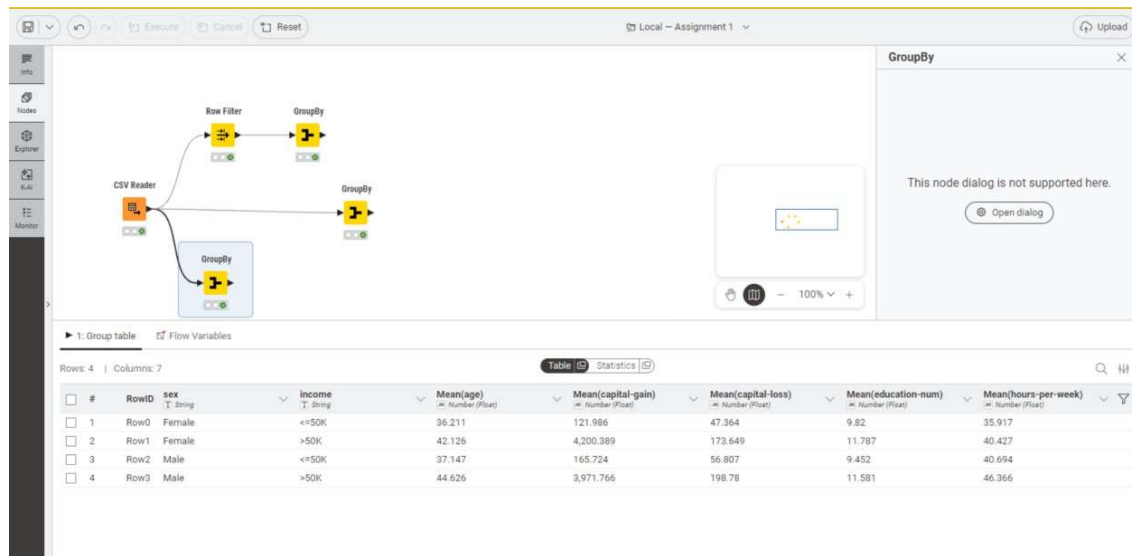
workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-lo...	hours-per...	native-co...	income
Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
Self-emp-not-in	292175	Masters	14	Divorced	Exec-managerial	Unmarried	White	Female	0	0	45	United-States	>50K
Private	51835	Prof-school	15	Married-civ-spo	Prof-specialty	Wife	White	Female	0	1902	60	Honduras	>50K
Private	169846	HS-grad	9	Married-civ-spo	Adm-clerical	Wife	White	Female	0	0	40	United-States	>50K
Private	343591	HS-grad	9	Divorced	Craft-repair	Not-in-family	White	Female	14344	0	40	United-States	>50K
Federal-gov	410867	Doctorate	16	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	50	United-States	>50K
Private	287828	Bachelors	13	Married-civ-spo	Exec-managerial	Wife	White	Female	0	0	40	United-States	>50K

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

The screenshot shows the same KNIME workflow as Step 2, but with the second GroupBy node configured to calculate the count and average age. The data table below shows the results of the GroupBy operation.

RowID	Count(age)	Mean(age)
1	1179	42.125

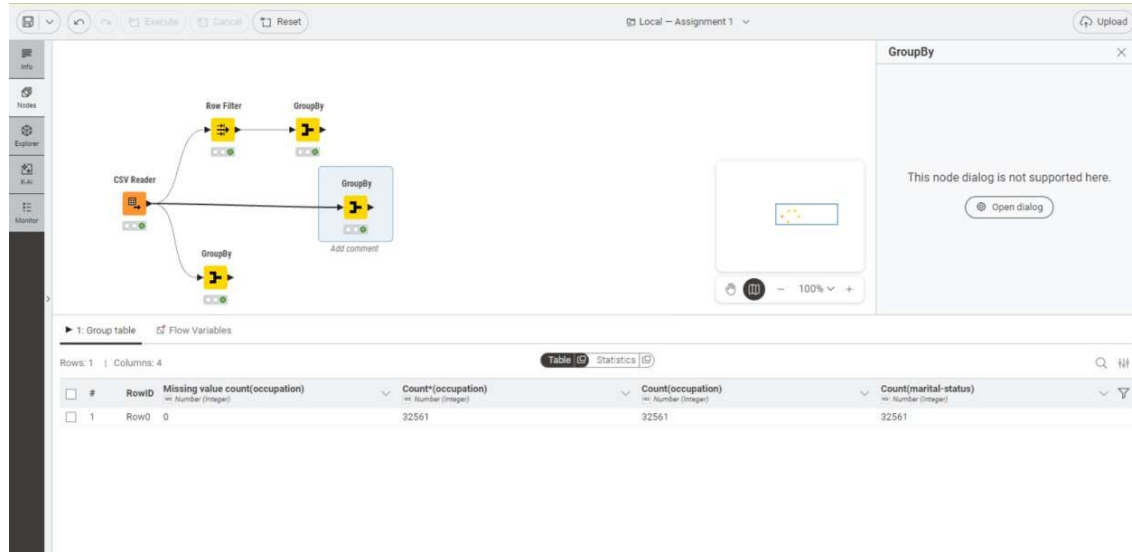
Step 4: Use GroupBy node to calculate the average of all numerical column for each of the 4 group defined by sex and income value



The KNIME workflow for Step 4 starts with a 'CSV Reader' node connected to a 'Row Filter' node. The 'Row Filter' node is connected to a 'GroupBy' node. The 'GroupBy' node is also connected to a 'Statistics' node. The 'Statistics' node is connected to a 'Table' node. The 'Table' node displays 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	35.211	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

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 KNIME workflow for Step 5 starts with a 'CSV Reader' node connected to a 'Row Filter' node. The 'Row Filter' node is connected to a 'GroupBy' node. The 'GroupBy' node is also connected to a 'Statistics' node. The 'Statistics' node is connected to a 'Table' node. The 'Table' node displays the following data:

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