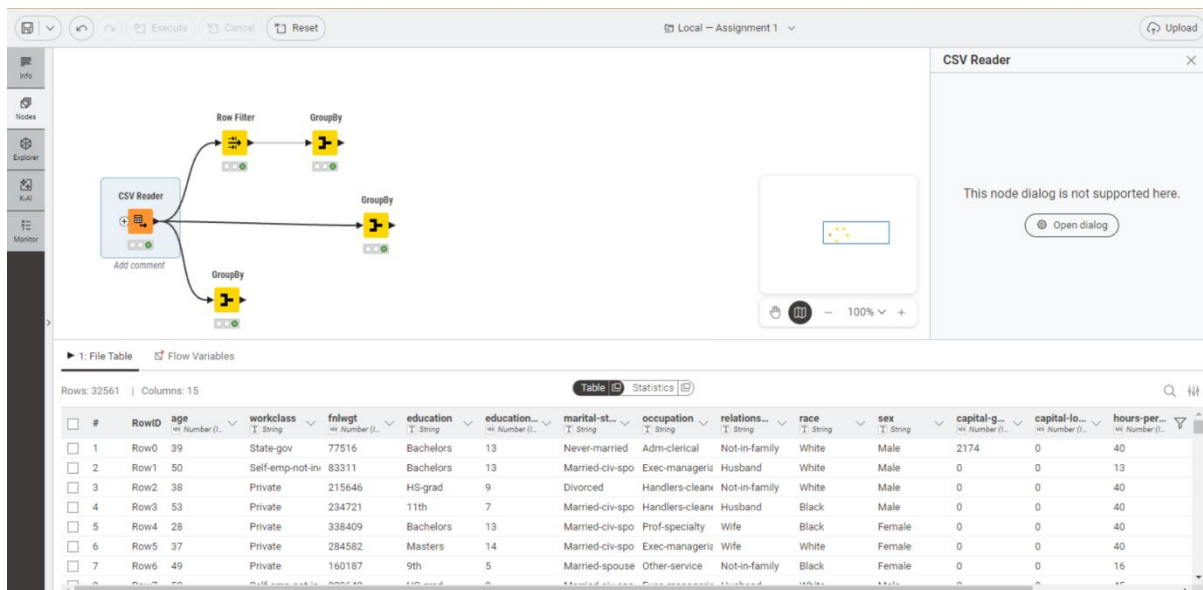


Knime - Assignment 1

- 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 shows the KNIME software interface. The main workspace displays a workflow with the following nodes: CSV Reader, Row Filter, and two GroupBy nodes. The CSV Reader node is connected to the Row Filter node, which is then connected to the first GroupBy node. The CSV Reader node is also connected to the second GroupBy node. The Row Filter node is configured to filter for 'sex = female' and 'income > 50000'. The first GroupBy node is configured to group by 'sex' and 'income'. The second GroupBy node is configured to group by 'sex' and 'income'. The CSV Reader node dialog is open, showing the file path and the 'Add comment' checkbox.

Below the workflow, the 'File Table' view shows the first 7 rows of the data:

#	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-spo	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-spo	Handlers-cleaner	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 the Power BI Desktop interface with a data flow from a CSV Reader to a Row Filter node and then to a GroupBy node. The Row Filter node is configured with the following criteria:

- Match row if matched by: All criteria
- Criterion 1: Filter column: sex, Operator: Equals, Value: F
- Criterion 2: Filter column: income, Operator: Greater than, Value: 50K

The resulting table shows 1179 rows and 15 columns. The columns are: workclass, fnlwgt, education, education-num, marital-st, occupation, relations, race, sex, capital-g, capital-lo, hours-per, native-co, and income.

workclass	fnlwgt	education	education-num	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 Power BI Desktop interface with a data flow from a CSV Reader to a Row Filter node and then to a GroupBy node. The GroupBy node is configured with the following criteria:

- Group by: sex
- Aggregates: Count*(age), Mean*(age)

The resulting table shows 1 row and 2 columns. The columns are: # and RowID. The data is as follows:

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

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 Power BI Desktop interface with a data flow from a CSV Reader to a Row Filter and then to a GroupBy node. The GroupBy node is configured to calculate the mean of numerical columns grouped by sex and income. The resulting table shows 4 rows with columns for sex, income, and means for age, capital-gain, capital-loss, education-num, and hours-per-week.

#	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

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 Power BI Desktop interface with a data flow from a CSV Reader to a Row Filter and then to a GroupBy node. The GroupBy node is configured to calculate the count of missing and non-missing values for occupation and marital-status. The resulting table shows 1 row with columns for missing value count(occupation), Count*(occupation), Count(occupation), and Count(marital-status).

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