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

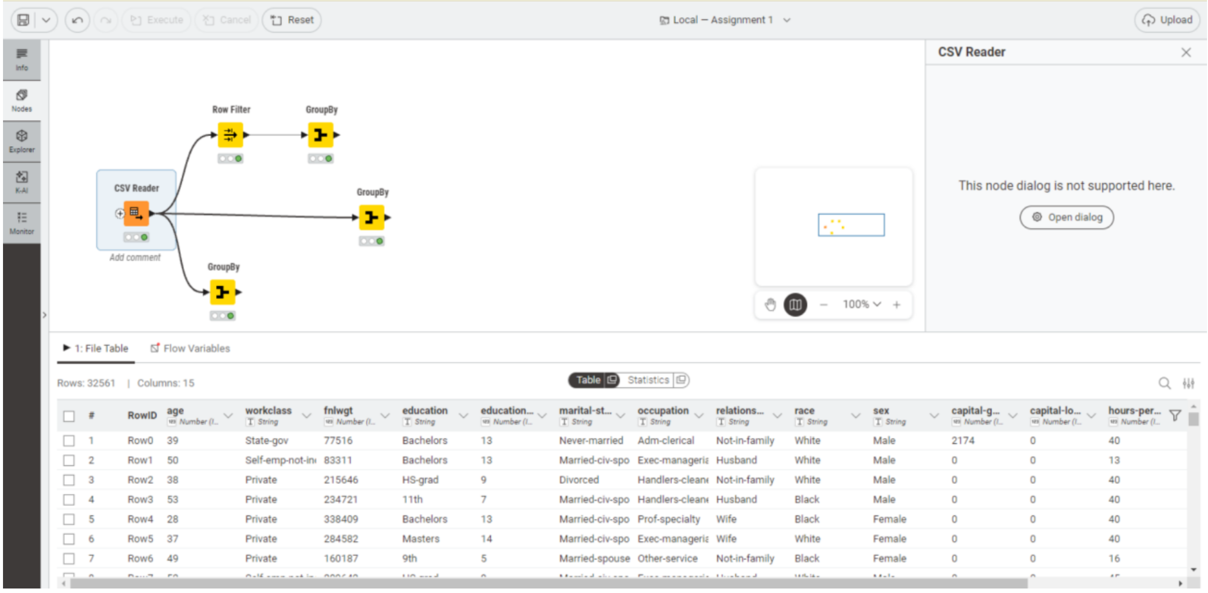
3) Calculate the averages of all numerical columns for each one of the 4 groups defined by sex and income values

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"

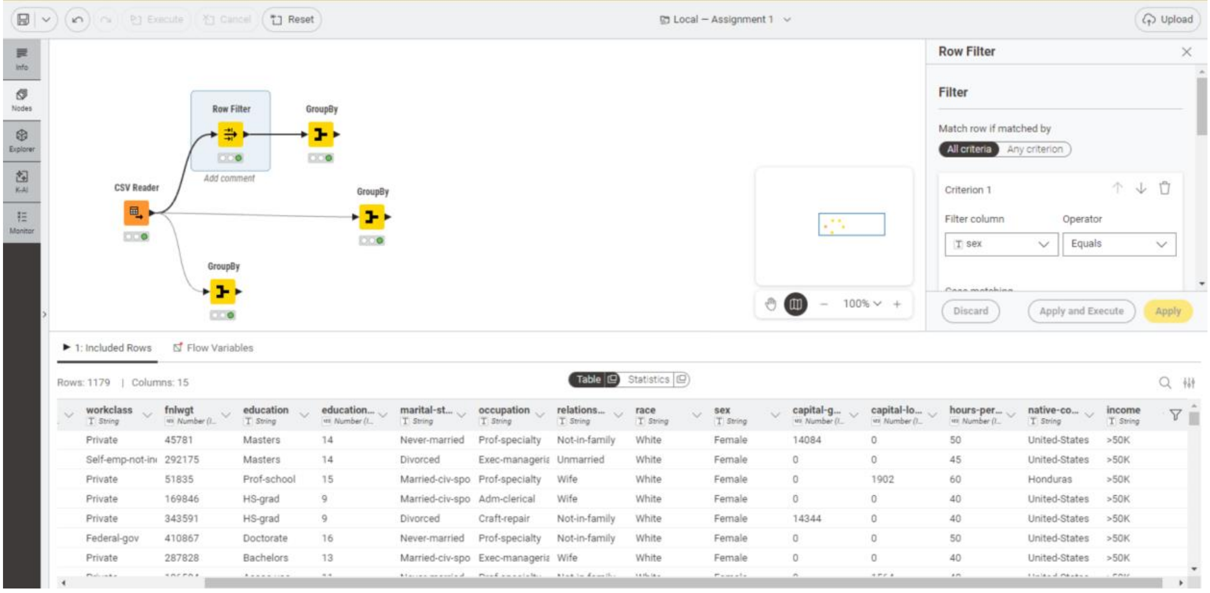


This node dialog is not supported here.

Rows: 32561 | Columns: 15

#	RowID	age Number (I...	workclass String	fnlwgt Number (I...	education String	education... Number (I...	marital-st... String	occupation String	relations... String	race String	sex String	capital-g... Number (I...	capital-lo... Number (I...	hours-per... Number (I...
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-in	83311	Bachelors	13	Married-civ-spo	Exec-manageria	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-manageria	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



Row Filter

Filter

Match row if matched by: All criteria | Any criterion

Criterion 1

Filter column: sex

Operator: Equals

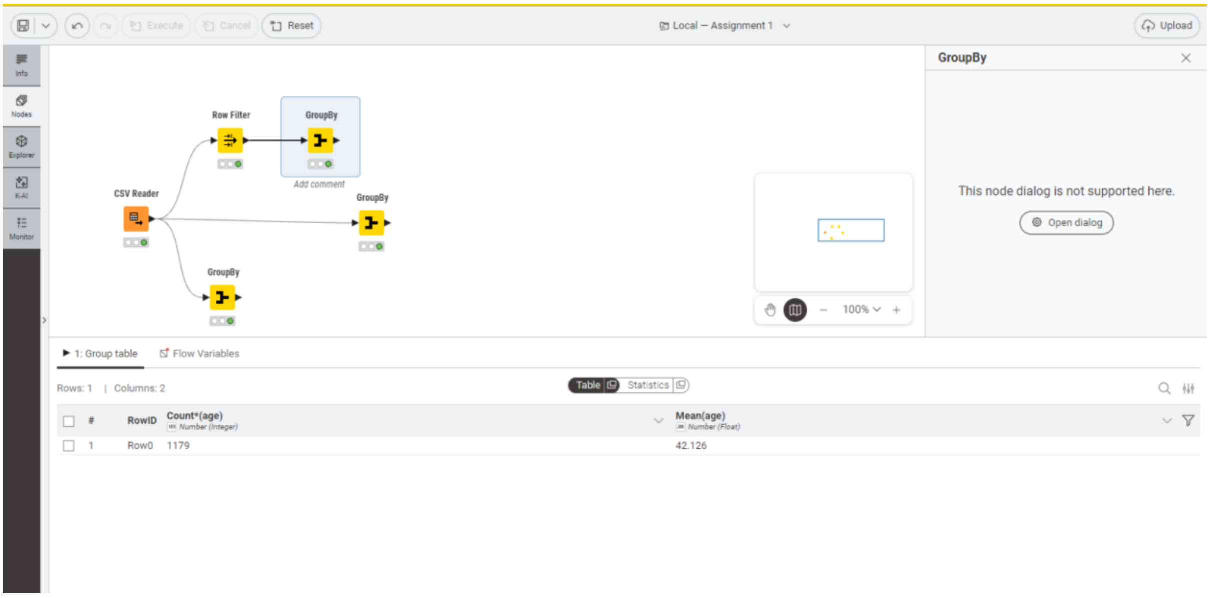
1: Included Rows | Flow Variables

Rows: 1179 | Columns: 15

Table | Statistics

workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-lo...	hours-per...	native-co...	income
String	Number (I...	String	Number (I...	String	String	String	String	String	Number (I...	Number (I...	Number (I...	String	String
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-manageri	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-manageri	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



This node dialog is not supported here.

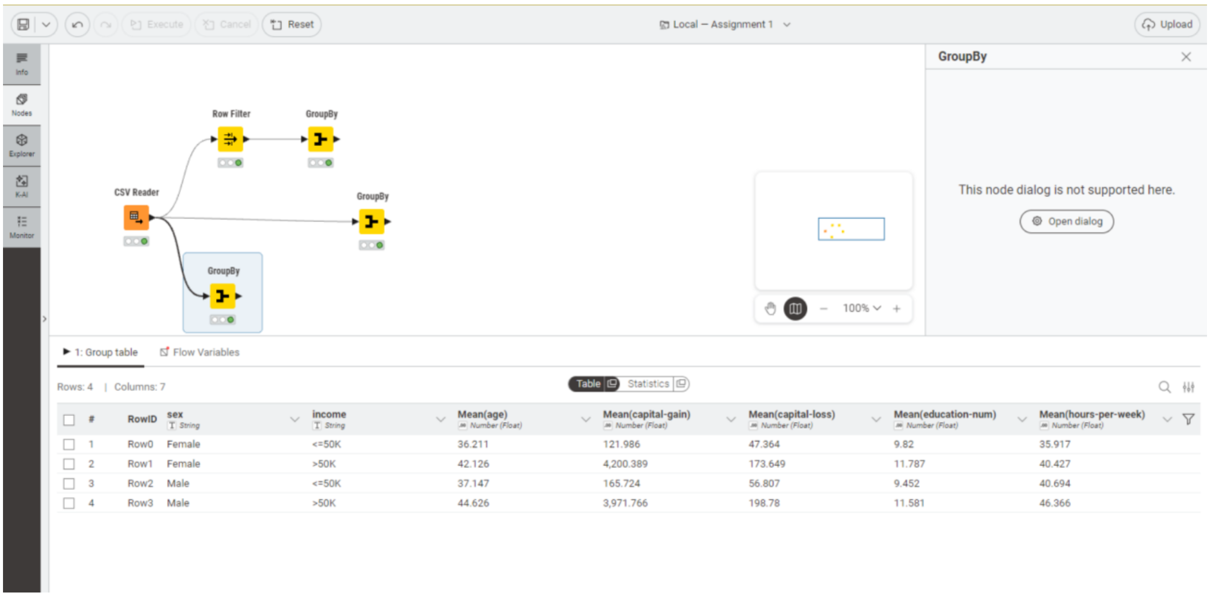
1: Group table | Flow Variables

Rows: 1 | Columns: 2

Table | Statistics

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

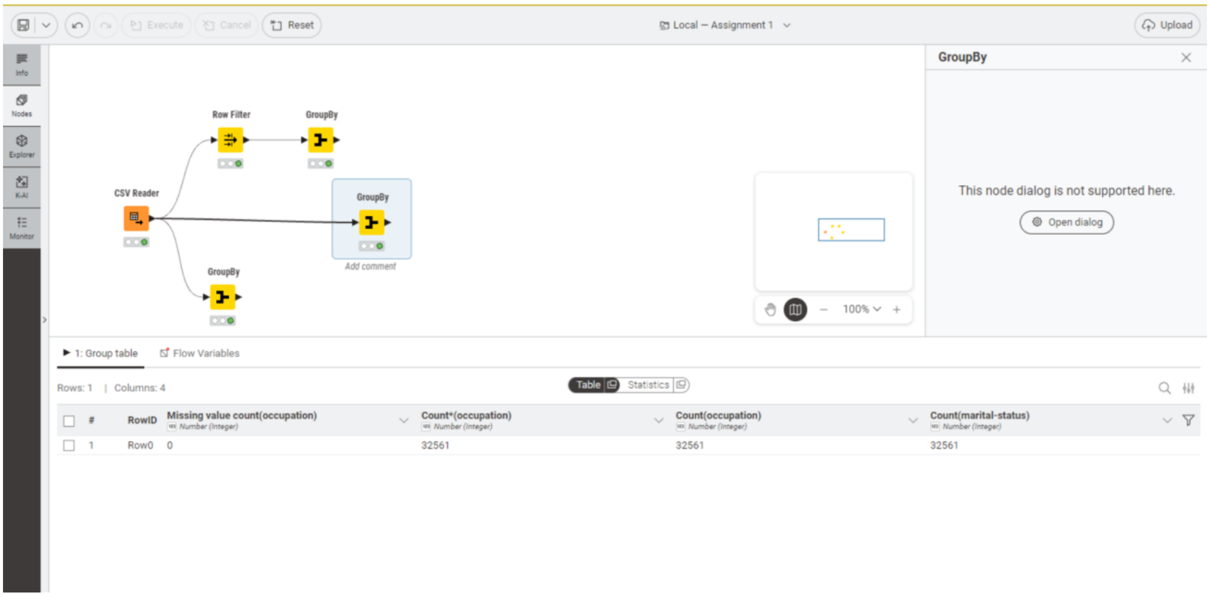
Step 4: Use GroupBy node to calculate the average of all numerical column for each of the 4- group defined by sex and income value



Rows: 4 | Columns: 7

#	RowID	sex String	income String	Mean(age) Number (Float)	Mean(capital-gain) Number (Float)	Mean(capital-loss) Number (Float)	Mean(education-num) Number (Float)	Mean(hours-per-week) Number (Float)
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 martial-status



Rows: 1 | Columns: 4

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