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

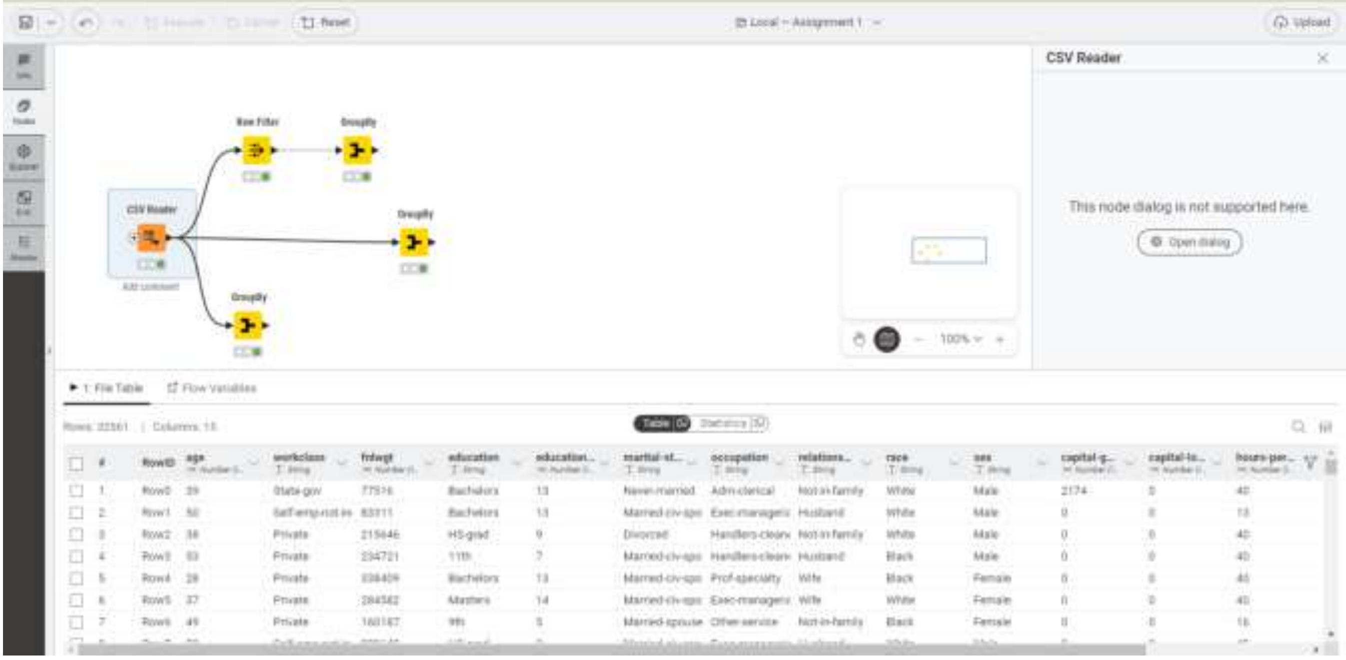
3) Calculate the averages of all numerical columns for each one of the 4 groups defined by sex and income values

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!

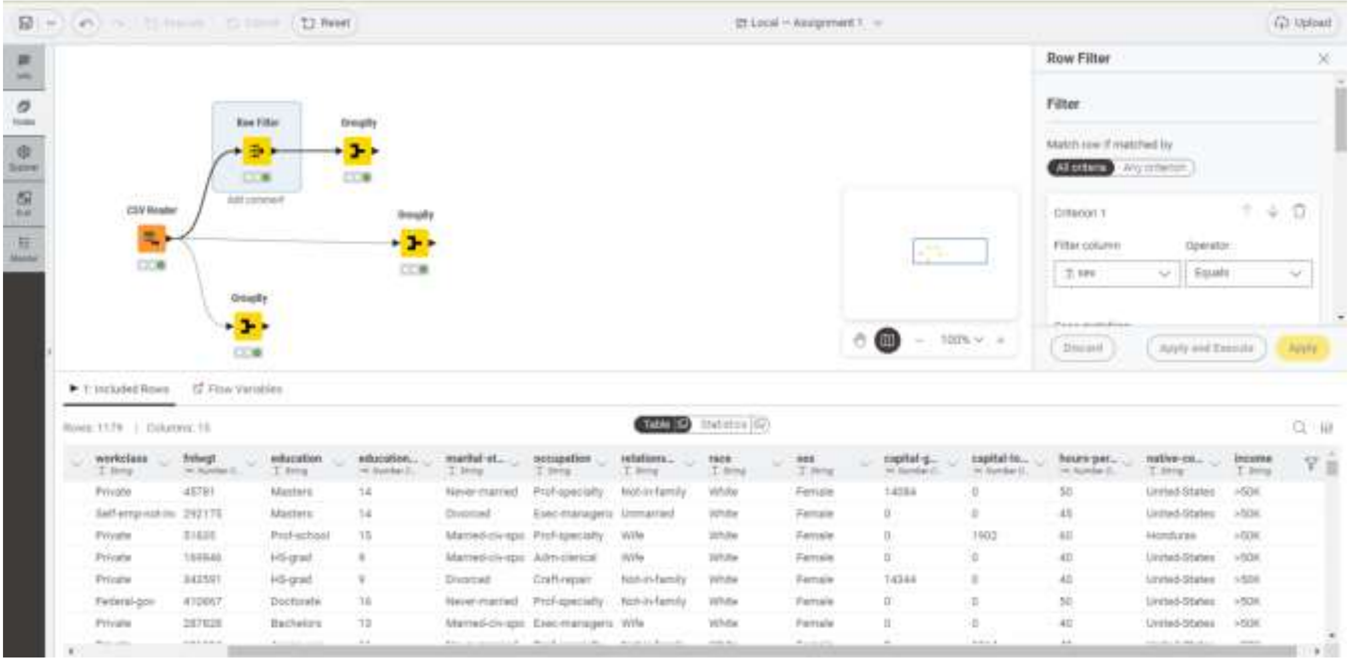
Step 1: Read CSV File "adult.csv"



Rows: 32561 | Columns: 15

#	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-in	83311	Bachelors	13	Married-civ-spo	Exec-manageri	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-manageri	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

Discard | Apply and Execute | Apply

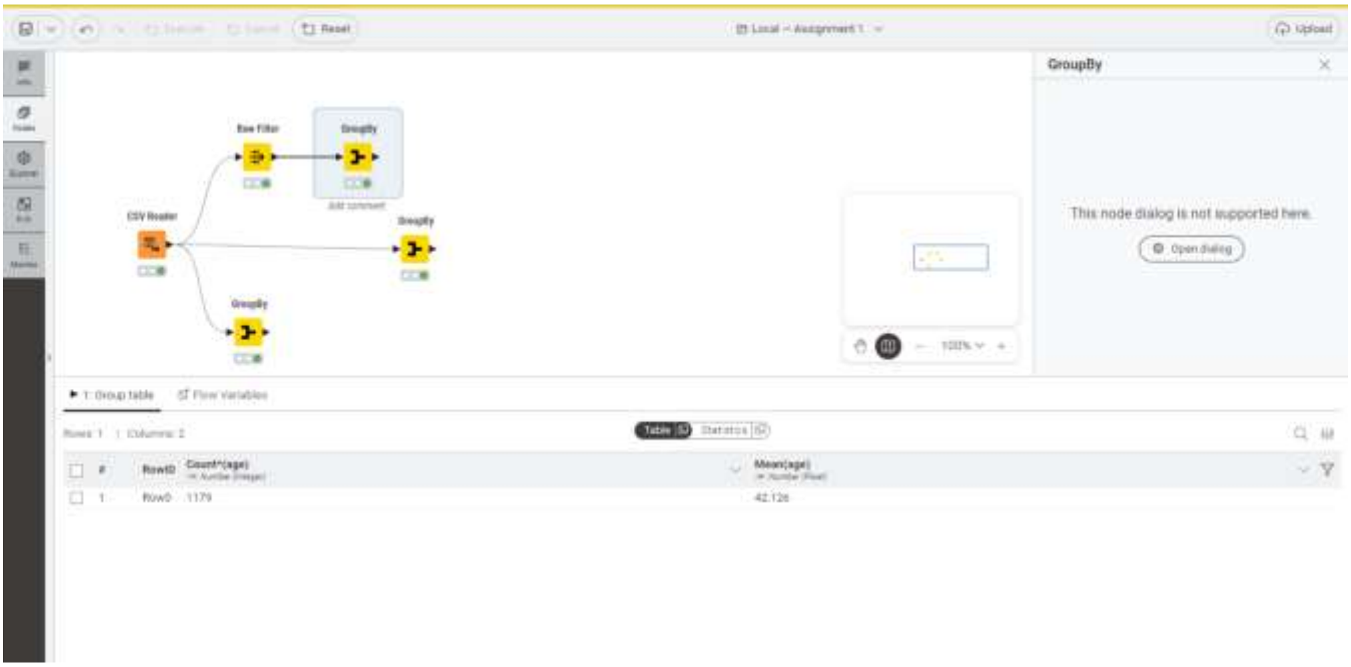
1: Included Rows | Flow Variables

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
Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50	United-States	>50K
Self-emp-not-inc	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



GroupBy

This node dialog is not supported here.

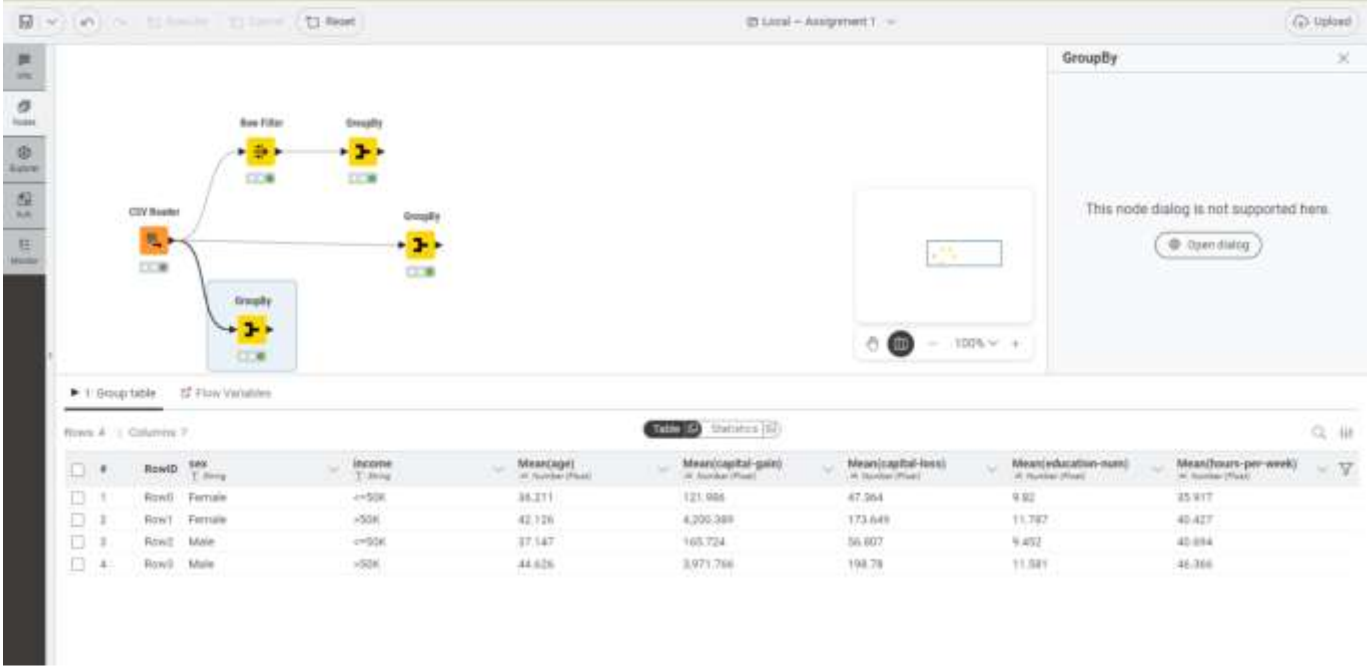
1: Group table | Flow Variables

Rows: 1 | Columns: 2

Table | Statistics

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

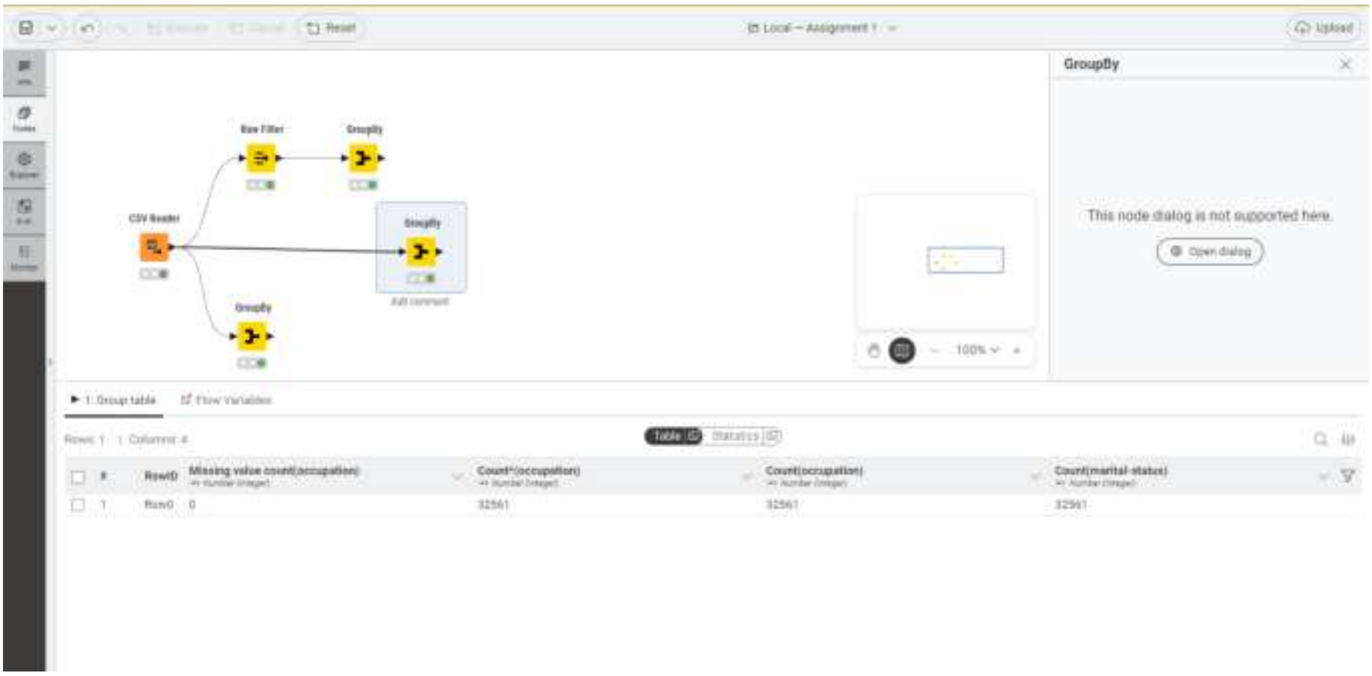
Step 4: Use GroupBy node to calculate the average of all numerical column for each of the 4-group defined by sex and income value



Rows: 4 | Columns: 7

#	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 martial-status



Rows: 1 | Columns: 4

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

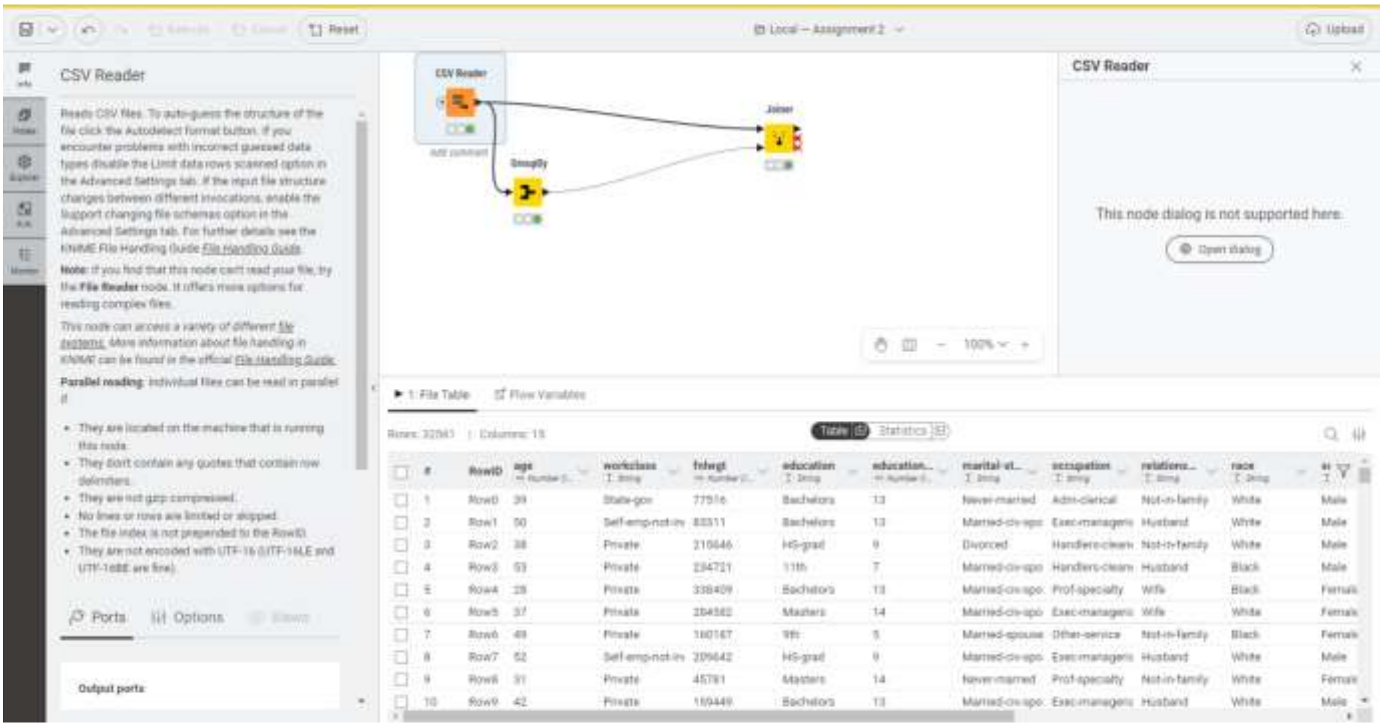
Assignment 2

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



CSV Reader

Reads CSV files. To auto-guess the structure of the file click the Autodetect format button. If you encounter problems with incorrect guessed data types disable the Limit data rows scanned option in the Advanced Settings tab. If the input file structure changes between different invocations, enable the Support changing file schemas option in the Advanced Settings tab. For further details see the KNIME File Handling Guide File Handling Guide.

Note: If you find that this node can't read your file, try the **File Reader** node. It offers more options for reading complex files.

This node can access a variety of different file systems. More information about file handling in KNIME can be found in the official File Handling Guide.

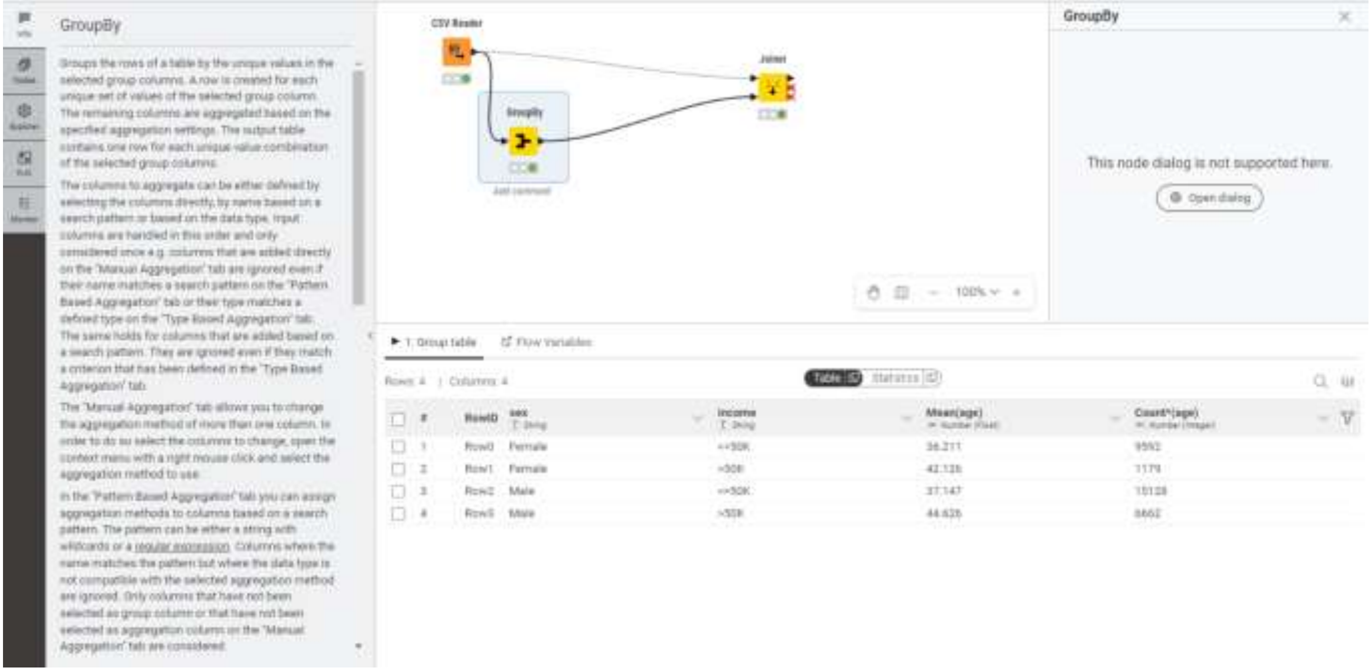
Parallel reading: Individual files can be read in parallel if

- They are located on the machine that is running this node.
- They don't contain any quotes that contain row delimiters.
- They are not gzip compressed.
- No lines or rows are limited or skipped.
- The file index is not prepended to the RowID.
- They are not encoded with UTF-16 (UTF-16LE and UTF-16BE are fine).

Rows: 32561 | Columns: 15

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	se
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-managen	Husband	White	Male
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-clean	Not-in-family	White	Male
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-clean	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-managen	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-managen	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-managen	Husband	White	Male

Step 2: Calculate the average age and count for each one of the 4 groups defined by sex and income values



GroupBy

Groups the rows of a table by the unique values in the selected group columns. A row is created for each unique set of values of the selected group column. The remaining columns are aggregated based on the specified aggregation settings. The output table contains one row for each unique value combination of the selected group columns.

The columns to aggregate can be either defined by selecting the columns directly, by name based on a search pattern or based on the data type. Input columns are handled in this order and only considered once e.g. columns that are added directly on the "Manual Aggregation" tab are ignored even if their name matches a search pattern on the "Pattern Based Aggregation" tab or their type matches a defined type on the "Type Based Aggregation" tab. The same holds for columns that are added based on a search pattern. They are ignored even if they match a criterion that has been defined in the "Type Based Aggregation" tab.

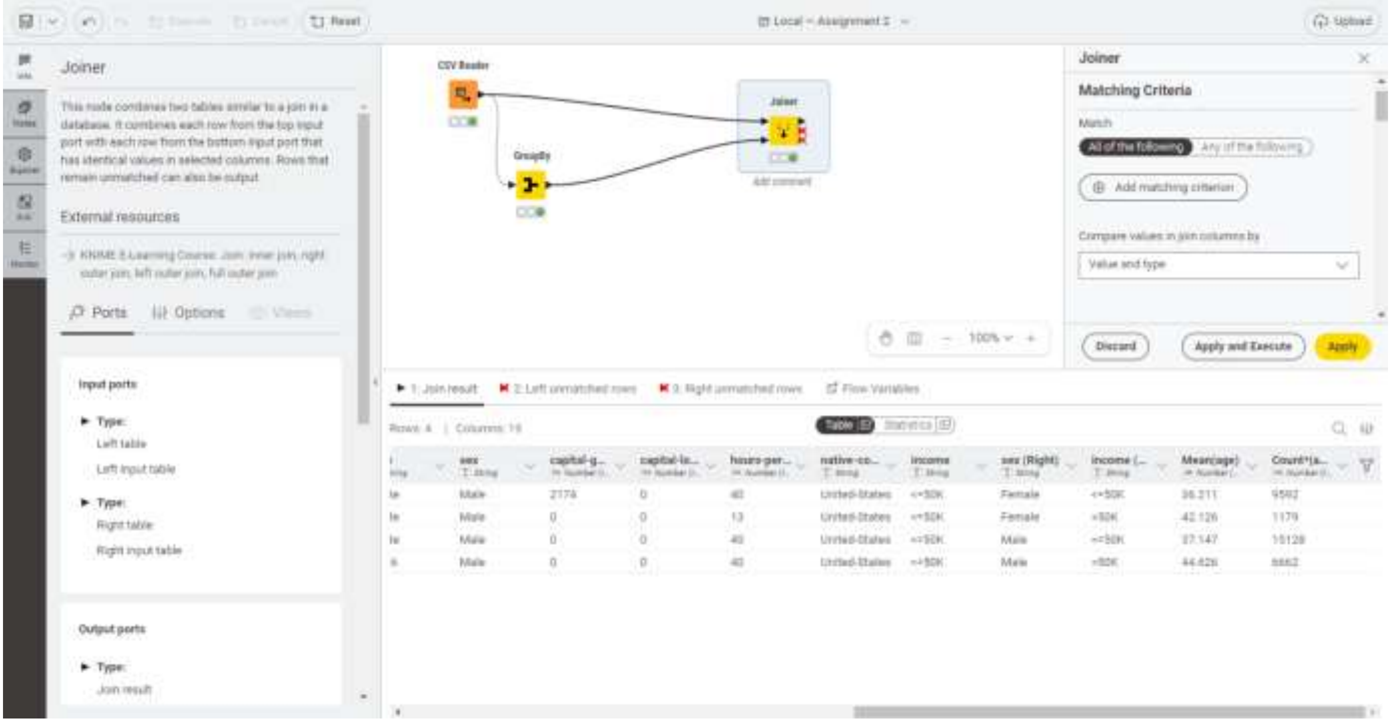
The "Manual Aggregation" tab allows you to change the aggregation method of more than one column. In order to do so select the columns to change, open the context menu with a right mouse click and select the aggregation method to use.

In the "Pattern Based Aggregation" tab you can assign aggregation methods to columns based on a search pattern. The pattern can be either a string with wildcards or a regular expression. Columns where the name matches the pattern but where the data type is not compatible with the selected aggregation method are ignored. Only columns that have not been selected as group column or that have not been selected as aggregation column on the "Manual Aggregation" tab are considered.

Rows: 4 | Columns: 4

#	RowID	sex String	income String	Mean(age) Number (Float)	Count*(age) Number (Integer)
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



Joiner

This node combines two tables similar to a join in a database. It combines each row from the top input port with each row from the bottom input port that has identical values in selected columns. Rows that remain unmatched can also be output.

External resources

KNIME E-Learning Course: Join: inner join, right outer join, left outer join, full outer join

Input ports

Type:
Left table
Left input table

Type:
Right table
Right input table

Output ports

Type:
Join result

Joiner

Matching Criteria

Match
All of the following | Any of the following

Add matching criterion

Compare values in join columns by
Value and type

Rows: 4 | Columns: 16

i ring	sex String	capital-g... Number (I...	capital-lo... Number (I...	hours-per... Number (I...	native-co... String	income String	sex (Right) String	income (... String	Mean(age) Number (...	Count*(a... Number (I...
le	Male	2174	0	40	United-States	<=50K	Female	<=50K	36.211	9592
le	Male	0	0	13	United-States	<=50K	Female	>50K	42.126	1179
le	Male	0	0	40	United-States	<=50K	Male	<=50K	37.147	15128
le	Male	0	0	40	United-States	<=50K	Male	>50K	44.626	6662

Assignment 3

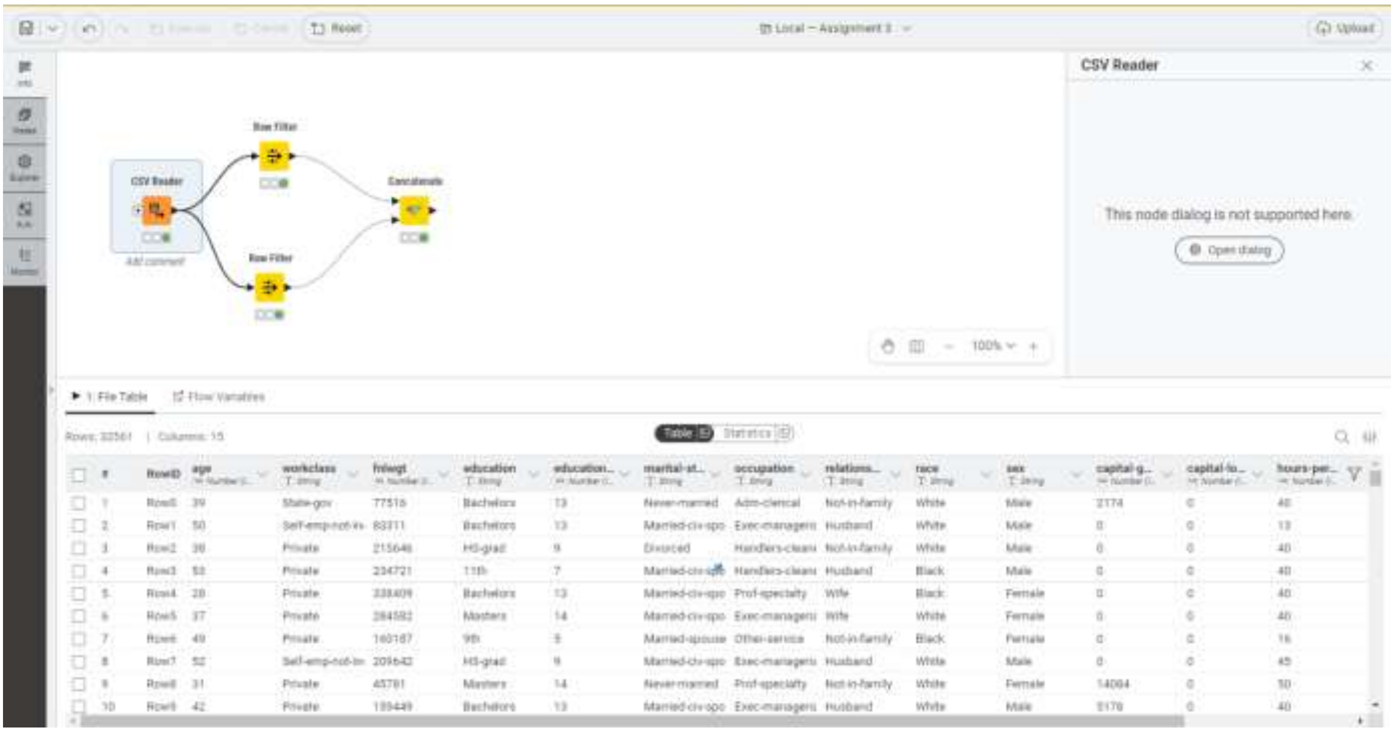
1) Read the adult.csv file available in the data folder on the KNIME Hub. The data are provided by the UCI Machine Learning Repository.

2) Extract people with age between 20 and 40 (both included) and working in a workclass starting with "S"

3) Extract people with age between 40 and 60 (both included) and working in a workclass starting with "P"

4) Concatenate both subsets into a single data table

Step 1: Read the adult.csv file

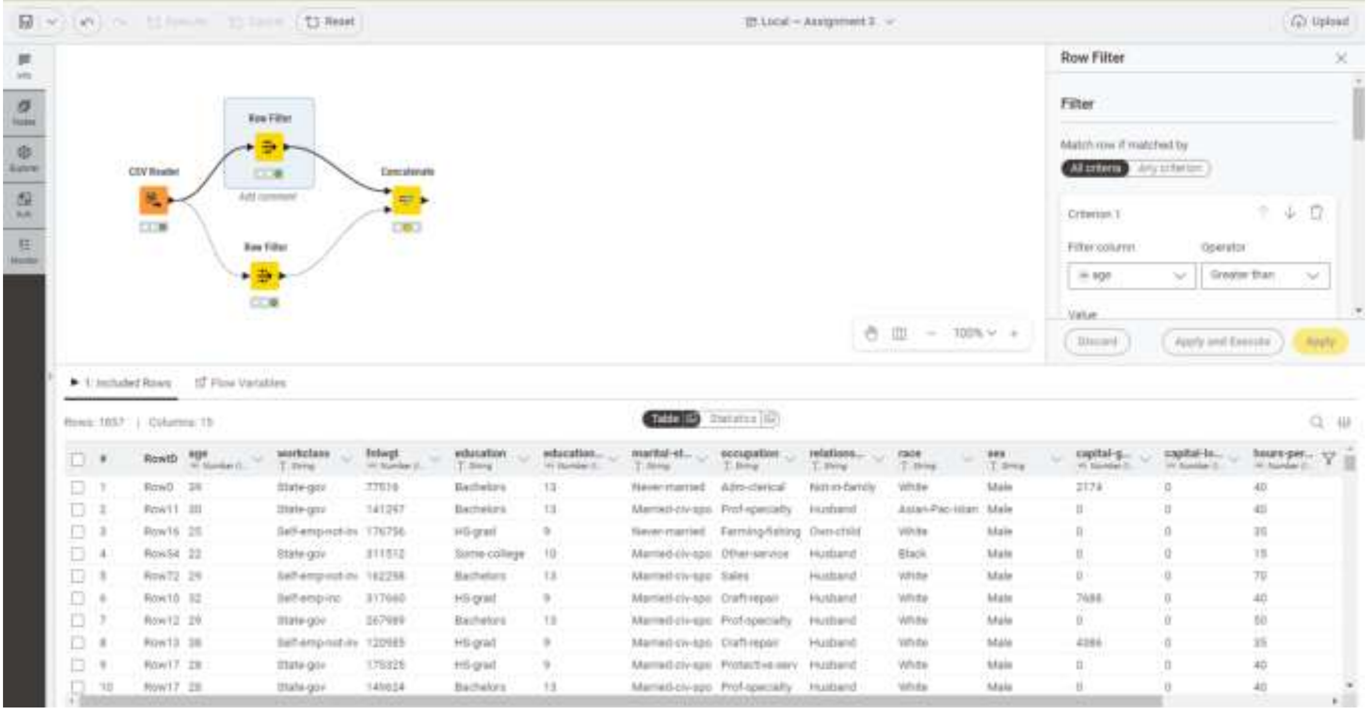


This node dialog is not supported here.

Rows: 32561 | Columns: 15

#	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-in	83311	Bachelors	13	Married-civ-spo	Exec-managerı	Husband	White	Male	0	0	13
3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaı	Not-in-family	White	Male	0	0	40
4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-cleaı	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-managerı	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
8	Row7	52	Self-emp-not-in	209642	HS-grad	9	Married-civ-spo	Exec-managerı	Husband	White	Male	0	0	45
9	Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50
10	Row9	42	Private	159449	Bachelors	13	Married-civ-spo	Exec-managerı	Husband	White	Male	5178	0	40

Step 2: Extract people with age between 20 and 40 (both included) and working in a work class starting with “S”



Row Filter

Filter

Match row if matched by

All criteria | Any criterion

Criterion 1

Filter column: age

Operator: Greater than

Value

Discard | Apply and Execute | Apply

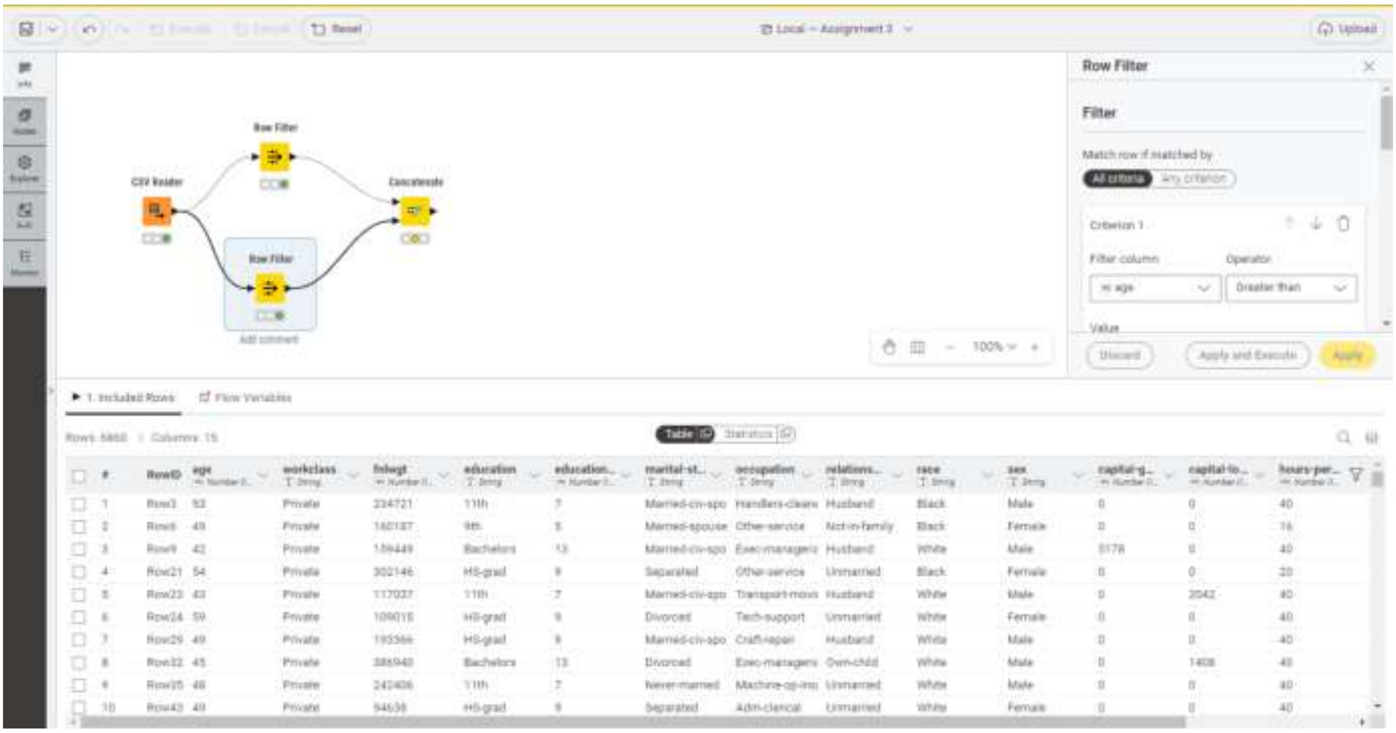
1: Included Rows | Flow Variables

Rows: 1657 | Columns: 15

Table | Statistics

#	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	Row11	30	State-gov	141297	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	Asian-Pac-Islan	Male	0	0	40
3	Row16	25	Self-emp-not-inc	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35
4	Row54	22	State-gov	311512	Some-college	10	Married-civ-spo	Other-service	Husband	Black	Male	0	0	15
5	Row72	29	Self-emp-not-inc	162298	Bachelors	13	Married-civ-spo	Sales	Husband	White	Male	0	0	70
6	Row10	32	Self-emp-inc	317660	HS-grad	9	Married-civ-spo	Craft-repair	Husband	White	Male	7688	0	40
7	Row12	29	State-gov	267989	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	White	Male	0	0	50
8	Row13	38	Self-emp-not-inc	120985	HS-grad	9	Married-civ-spo	Craft-repair	Husband	White	Male	4386	0	35
9	Row17	28	State-gov	175325	HS-grad	9	Married-civ-spo	Protective-serv	Husband	White	Male	0	0	40
10	Row17	28	State-gov	149624	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	White	Male	0	0	40

Step 3: Extract People with age between 40 and 60 (both included) and working in a work class starting with “P”



Row Filter

Filter

Match row if matched by

All criteria | Any criterion

Criterion 1

Filter column: age

Operator: Greater than

Value

Discard | Apply and Execute | Apply

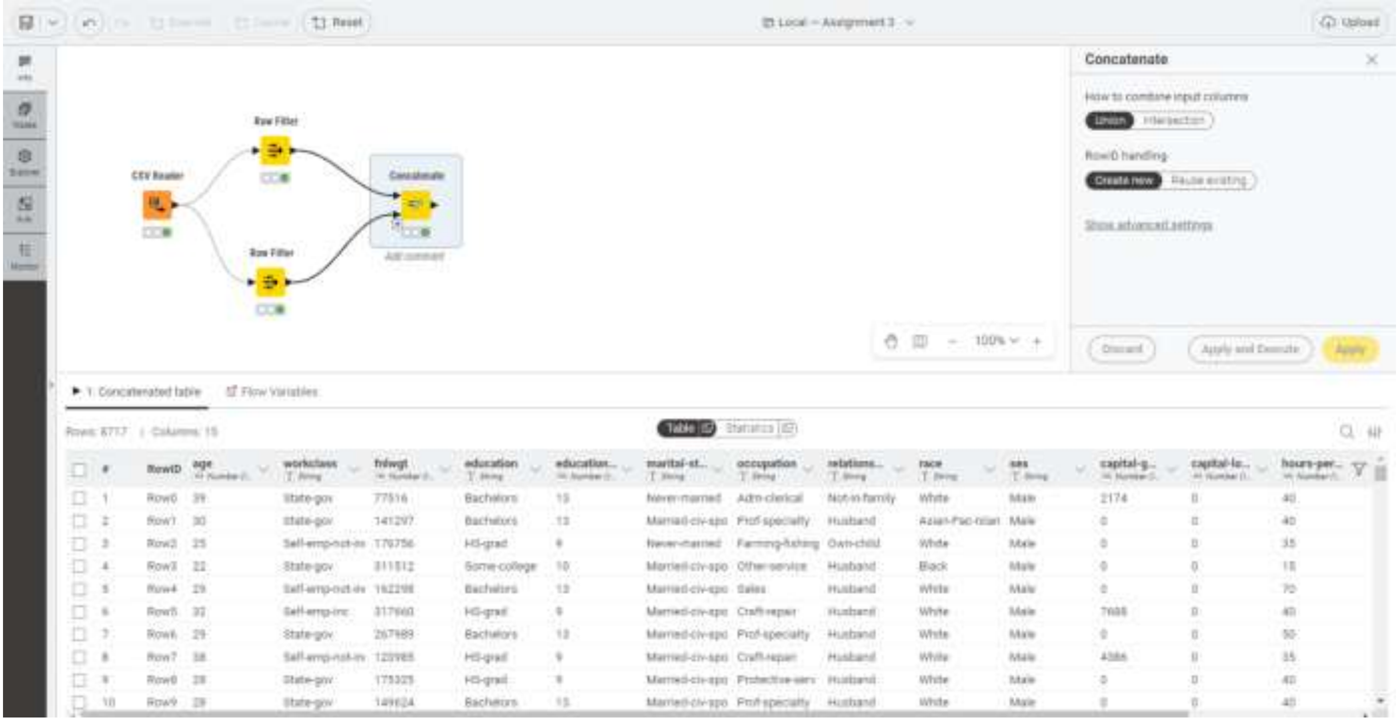
1: Included Rows | Flow Variables

Rows: 5868 | Columns: 15

Table | Statistics

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-lo...	hours-per...
1	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-cleane	Husband	Black	Male	0	0	40
2	Row6	49	Private	160187	9th	5	Married-spouse	Other-service	Not-in-family	Black	Female	0	0	16
3	Row9	42	Private	159449	Bachelors	13	Married-civ-spo	Exec-manageri	Husband	White	Male	5178	0	40
4	Row21	54	Private	302146	HS-grad	9	Separated	Other-service	Unmarried	Black	Female	0	0	20
5	Row23	43	Private	117037	11th	7	Married-civ-spo	Transport-movi	Husband	White	Male	0	2042	40
6	Row24	59	Private	109015	HS-grad	9	Divorced	Tech-support	Unmarried	White	Female	0	0	40
7	Row29	49	Private	193366	HS-grad	9	Married-civ-spo	Craft-repair	Husband	White	Male	0	0	40
8	Row32	45	Private	386940	Bachelors	13	Divorced	Exec-manageri	Own-child	White	Male	0	1408	40
9	Row35	48	Private	242406	11th	7	Never-married	Machine-op-insi	Unmarried	White	Male	0	0	40
10	Row43	49	Private	94638	HS-grad	9	Separated	Adm-clerical	Unmarried	White	Female	0	0	40

Step 4: Concatenate both subsets into a single data



Concatenate

How to combine input columns

[✓ Union] Intersection

RowID handling

[✓ Create new] Reuse existing

Show advanced settings

Rows: 8717 | Columns: 15

#	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	30	State-gov	141297	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	Asian-Pac-Islan	Male	0	0	40
3	Row2	25	Self-emp-not-in	176756	HS-grad	9	Never-married	Farming-fishing	Own-child	White	Male	0	0	35
4	Row3	22	State-gov	311512	Some-college	10	Married-civ-spo	Other-service	Husband	Black	Male	0	0	15
5	Row4	29	Self-emp-not-in	162298	Bachelors	13	Married-civ-spo	Sales	Husband	White	Male	0	0	70
6	Row5	32	Self-emp-inc	317660	HS-grad	9	Married-civ-spo	Craft-repair	Husband	White	Male	7688	0	40
7	Row6	29	State-gov	267989	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	White	Male	0	0	50
8	Row7	38	Self-emp-not-in	120985	HS-grad	9	Married-civ-spo	Craft-repair	Husband	White	Male	4386	0	35
9	Row8	28	State-gov	175325	HS-grad	9	Married-civ-spo	Protective-serv	Husband	White	Male	0	0	40
10	Row9	28	State-gov	149624	Bachelors	13	Married-civ-spo	Prof-specialty	Husband	White	Male	0	0	40