K.R. MANGALAM UNIVERSITY

THE COMPLETE WORLD OF EDUCATION

Data Analysis with Power BI & KNIME (ETSEDA115)

MCA (AI & ML)- Sem 1

Assignment 1

Roll No:-

2501940072

Submitted by:-

KULSUM BANOO

Submitted to:-

MR. MOHAMMAD AIJAZ

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

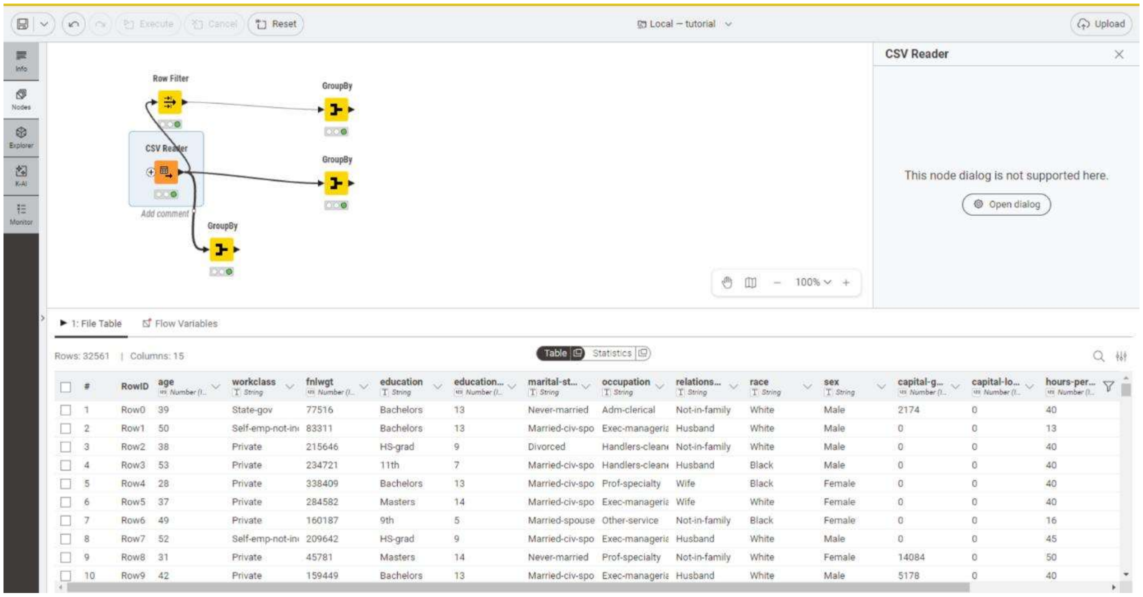
3) Calculate the averages of all numerical columns for each one of the 4 groups defined by sex and income values

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!

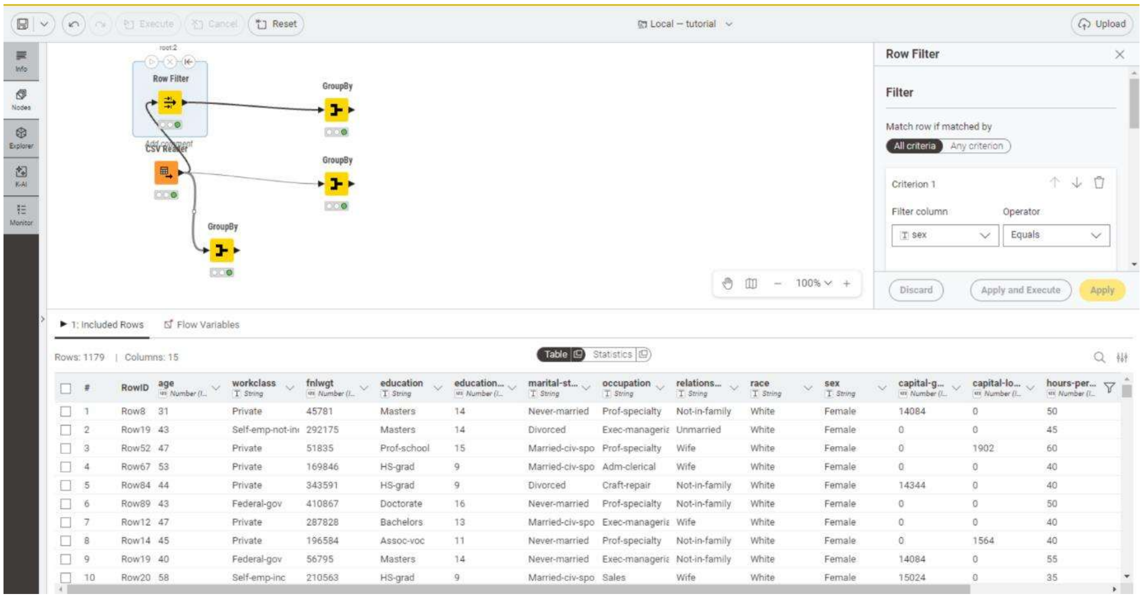
1) Read the adult.csv file



Rows: 32561 | Columns: 15

#	RowID	age	workclass	fnlwgt	education	education...	marital-st...	occupation	relations...	race	sex	capital-g...	capital-lo...	hours-per...
		Number (I...	String	Number (I...	String	Number (I...	String	String	String	String	String	Number (I...	Number (I...	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
8	Row7	52	Self-emp-not-in	209642	HS-grad	9	Married-civ-spo	Exec-manageria	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-manageria	Husband	White	Male	5178	0	40

2) A) Filter Female and Income >50k using Row Filter



Row Filter

Filter

Match row if matched by: All criteria | Any criterion

Criterion 1

Filter column: sex

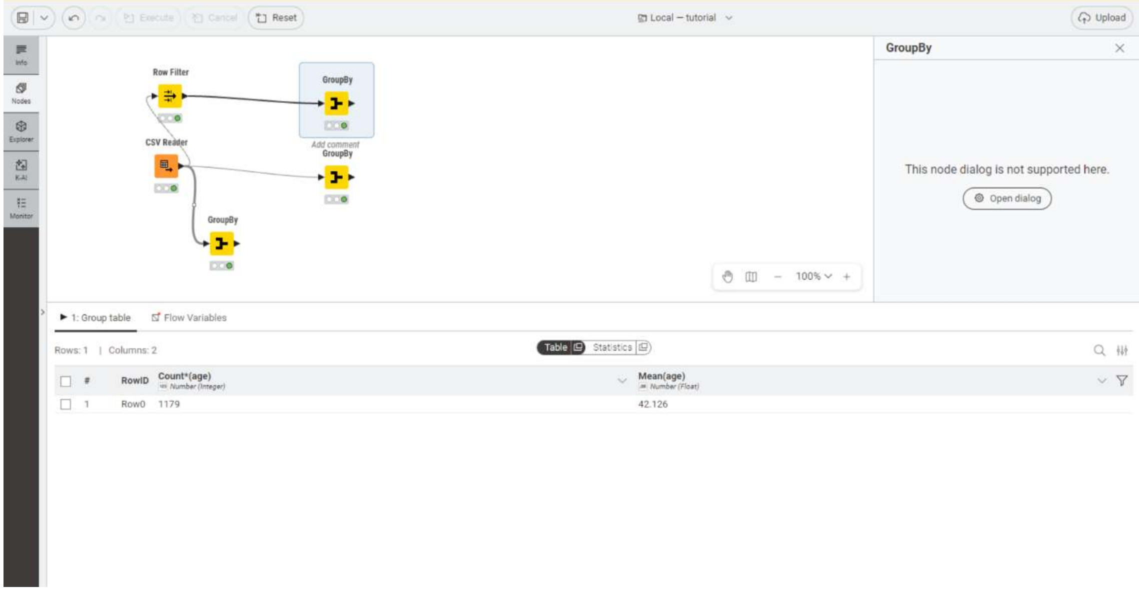
Operator: Equals

1: Included Rows | Flow Variables

Rows: 1179 | Columns: 15

#	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	Row8	31	Private	45781	Masters	14	Never-married	Prof-specialty	Not-in-family	White	Female	14084	0	50
2	Row19	43	Self-emp-not-in	292175	Masters	14	Divorced	Exec-manageria	Unmarried	White	Female	0	0	45
3	Row52	47	Private	51835	Prof-school	15	Married-civ-spo	Prof-specialty	Wife	White	Female	0	1902	60
4	Row67	53	Private	169846	HS-grad	9	Married-civ-spo	Adm-clerical	Wife	White	Female	0	0	40
5	Row84	44	Private	343591	HS-grad	9	Divorced	Craft-repair	Not-in-family	White	Female	14344	0	40
6	Row89	43	Federal-gov	410867	Doctorate	16	Never-married	Prof-specialty	Not-in-family	White	Female	0	0	50
7	Row12	47	Private	287828	Bachelors	13	Married-civ-spo	Exec-manageria	Wife	White	Female	0	0	40
8	Row14	45	Private	196584	Assoc-voc	11	Never-married	Prof-specialty	Not-in-family	White	Female	0	1564	40
9	Row19	40	Federal-gov	56795	Masters	14	Never-married	Exec-manageria	Not-in-family	White	Female	14084	0	55
10	Row20	58	Self-emp-inc	210563	HS-grad	9	Married-civ-spo	Sales	Wife	White	Female	15024	0	35

2) B) Calculate the Count and Average age of women with income >50k



GroupBy

This node dialog is not supported here.

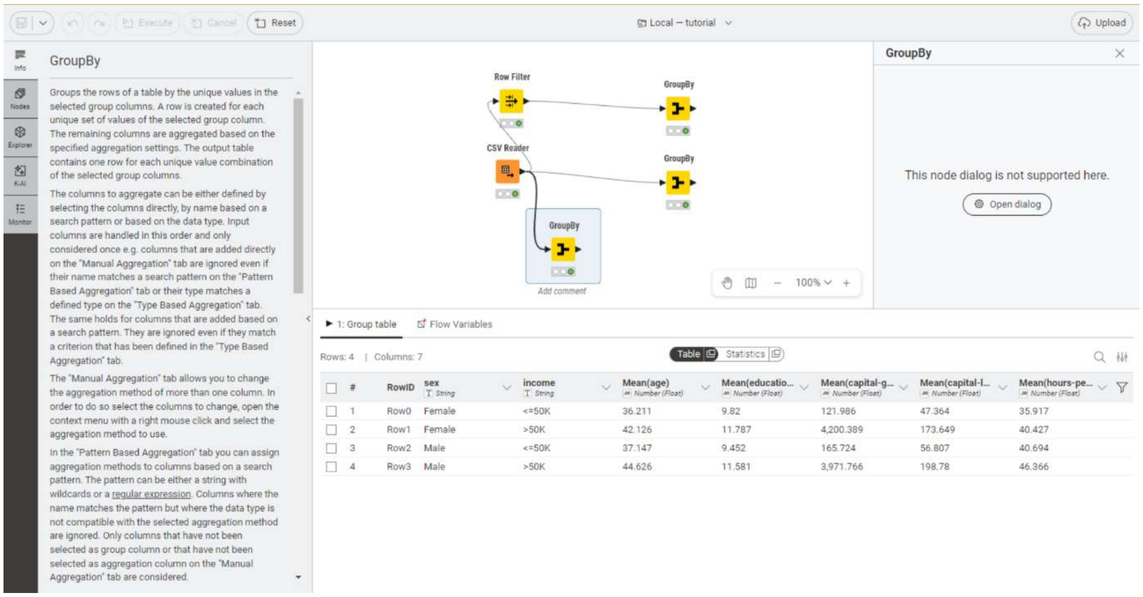
Open dialog

1: Group table | Flow Variables

Rows: 1 | Columns: 2

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

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



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.

GroupBy

1: Group table | Flow Variables

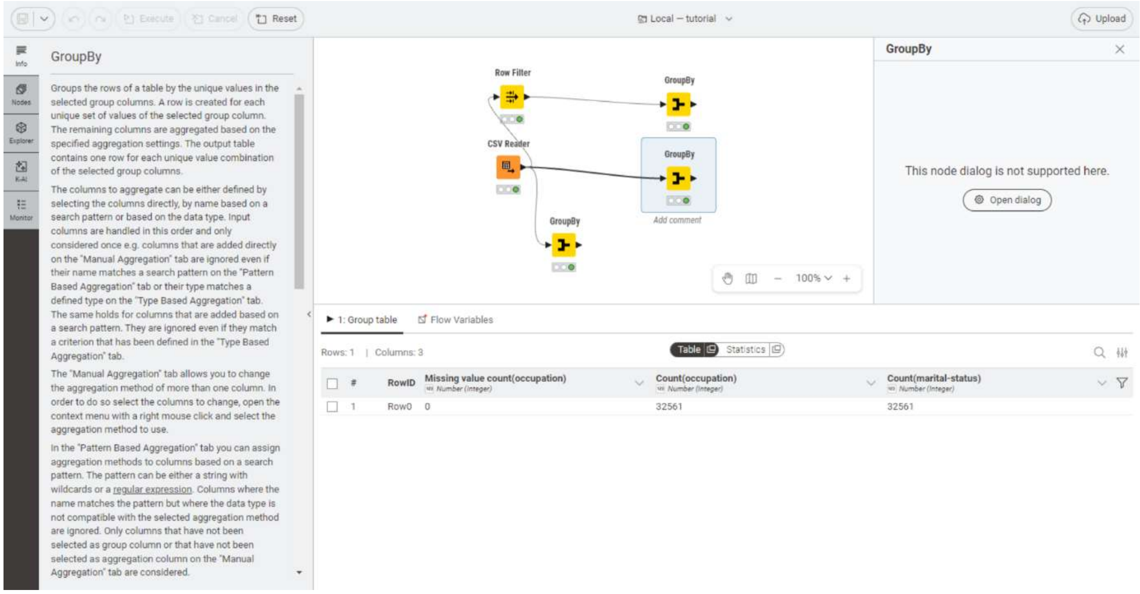
Rows: 4 | Columns: 7

Table | Statistics

□	#	RowID	sex String	income String	Mean(age) Number (Float)	Mean(educatio... Number (Float)	Mean(capital-g... Number (Float)	Mean(capital-l... Number (Float)	Mean(hours-pe... Number (Float)
□	1	Row0	Female	<=50K	36.211	9.82	121.986	47.364	35.917
□	2	Row1	Female	>50K	42.126	11.787	4,200.389	173.649	40.427
□	3	Row2	Male	<=50K	37.147	9.452	165.724	56.807	40.694
□	4	Row3	Male	>50K	44.626	11.581	3,971.766	198.78	46.366

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



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.

GroupBy

1: Group table | Flow Variables

Rows: 1 | Columns: 3

Table | Statistics

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