

- Notice that the last two aggregations should provide the same numbers!

1) Read the adult.csv file

Local - tutorial
Upload

```

graph LR
    CSVReader[CSV Reader] --> RF[Row Filter]
    CSVReader --> AC[Add Comment]
    RF --> GB1[GroupBy]
    AC --> GB2[GroupBy]
        
```

CSV Reader

This node dialog is not supported here.

[Open dialog](#)

► 1: File Table
🔗 Flow Variables

#	RowId	age	workclass	fnlwgt	education	education_...	marital_st...	occupation	relations...	race	sex	capital_g...	capital-la...	hours-per...
		T: String w/ Number [...]		w/ Number [...]	T: String	w/ Number [...]	T: String	T: String	T: String	T: String	T: String	w/ Number [...]	w/ Number [...]	w/ Number [...]
❑ 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-manageriz	Husband	White	Male	0	0	13
❑ 3	Row2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleanz	Not-in-family	White	Male	0	0	40
❑ 4	Row3	53	Private	234721	11th	7	Married-civ-spo	Handlers-cleanz	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-manageriz	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-manageriz	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-manageriz	Husband	White	Male	5178	0	40

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

2 Power BI Assignment 1

Row Filter

Filter

Match row if matched by

All criteria Any criterion

Criterion 1

Filter column Operator

sex Equals

Discard Apply and Execute Apply

► 1: Included Rows Flow Variables

Rows: 1179 | Columns: 15

#	RowID	age	workclass	fnlwgt	education	education_	marital-st	occupation	relations	race	sex	capital-g	capital-lo	hours-per
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-manageris	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-manageris	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-manageris	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)	Mean*(age)
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

The screenshot shows the KNIME GroupBy node configuration. The left pane displays the 'Info' tab with documentation. The main workspace shows a flow diagram with a 'Row Filter' node connected to a 'CSV Reader' node, which then feeds into a 'GroupBy' node. The right pane shows the 'GroupBy' node dialog, which is not supported in this version.

1. Group table | Flow Variables

Rows: 4 | Columns: 7

#	RowID	sex	income	Mean(age)	Mean(educatio...	Mean(capital-g...	Mean(capital-L...	Mean(hours-pe...
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

The screenshot shows the KNIME GroupBy node configuration. The left pane displays the 'Info' tab with documentation. The main workspace shows a flow diagram with a 'Row Filter' node connected to a 'CSV Reader' node, which then feeds into a 'GroupBy' node. The right pane shows the 'GroupBy' node dialog, which is not supported in this version.

1. Group table | Flow Variables

Rows: 1 | Columns: 3

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