# 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

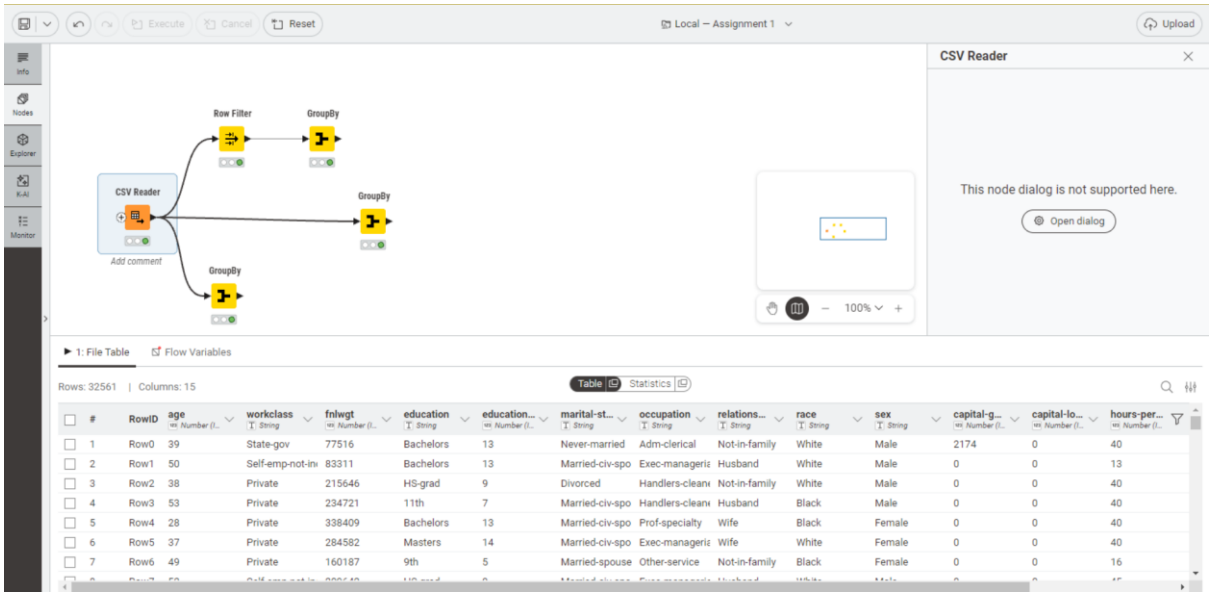
3) Calculate the averages of all numerical columns for each one of the 4 groups defined by sex and income values

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

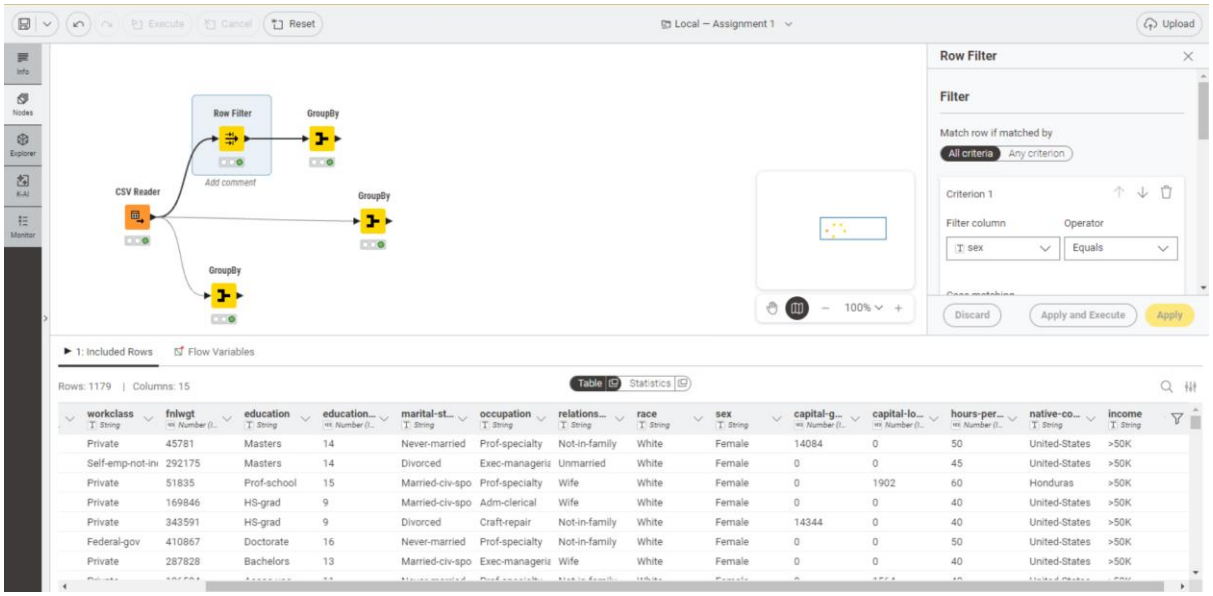
Notice that the last two aggregations should provide the same numbers!

**Step 1:** Read CSV File “adult.csv”

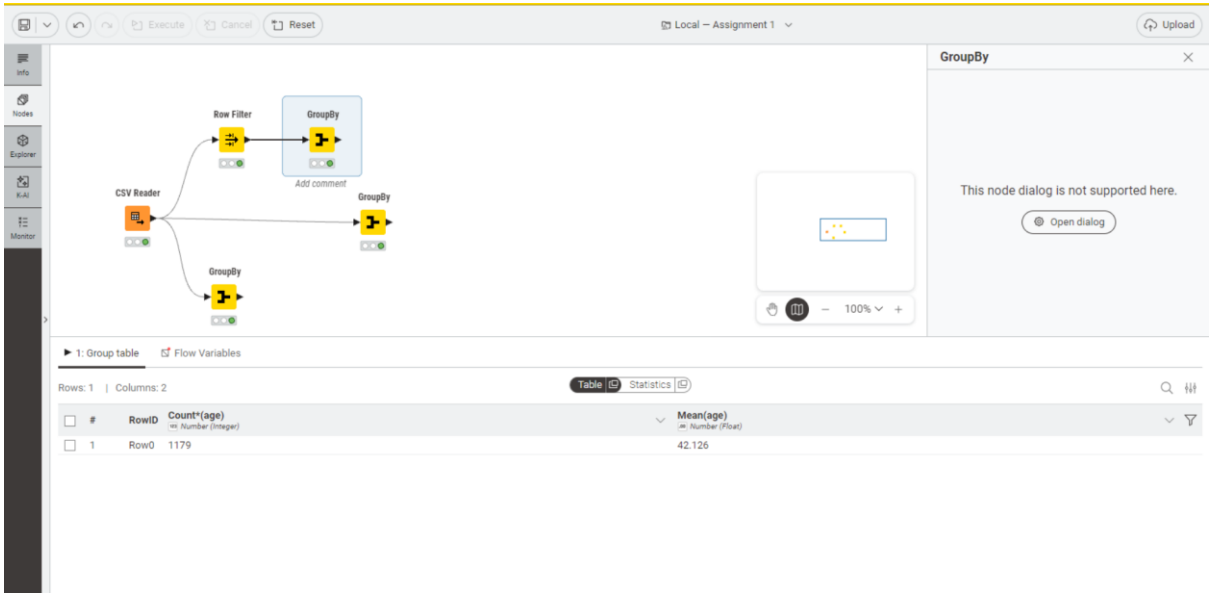



| # | RowID | age | workclass | fnlwgt | education | education... | marital-st... | occupation | relations... | race | sex | capital-g... | capital-lo... | hours-per... |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Row0 | 39 | State-gov | 77516 | Bachelors | 13 | Never-married | Adm-clerical | Not-in-family | White | Male | 2174 | 0 | 40 |
| 2 | Row1 | 50 | Self-emp-not-in | 83311 | Bachelors | 13 | Married-civ-spo | Exec-managerial | 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-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 |

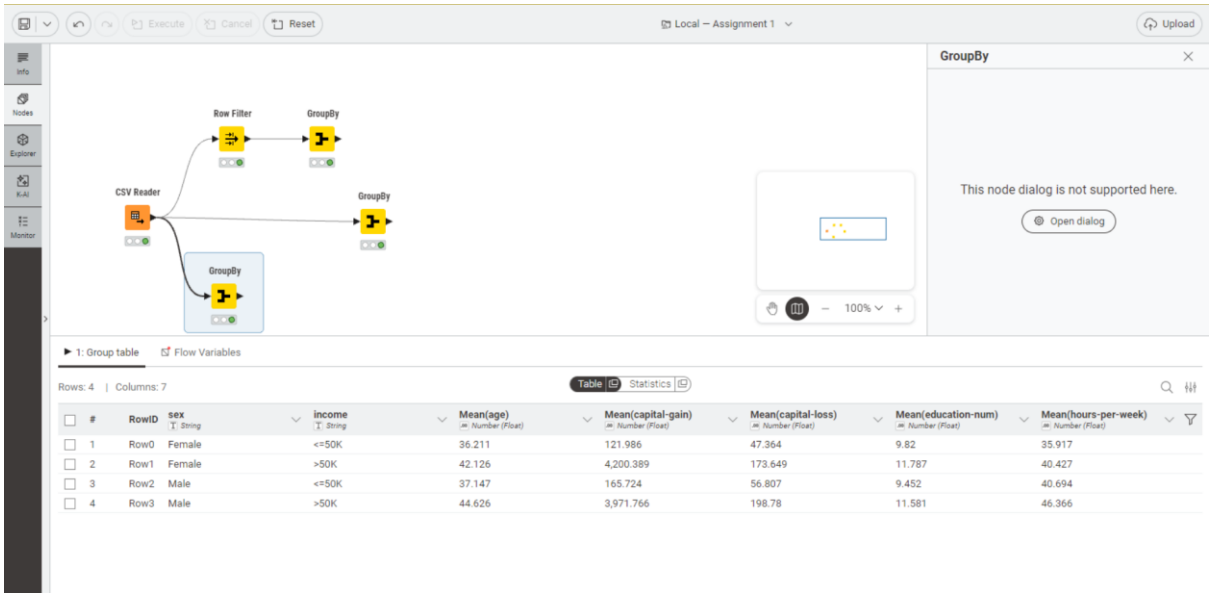
**Step 2:** Filter Row for Women with income >50K




**Step 3:** Use GroupBy node to calculate the count and average age of women with income >50K

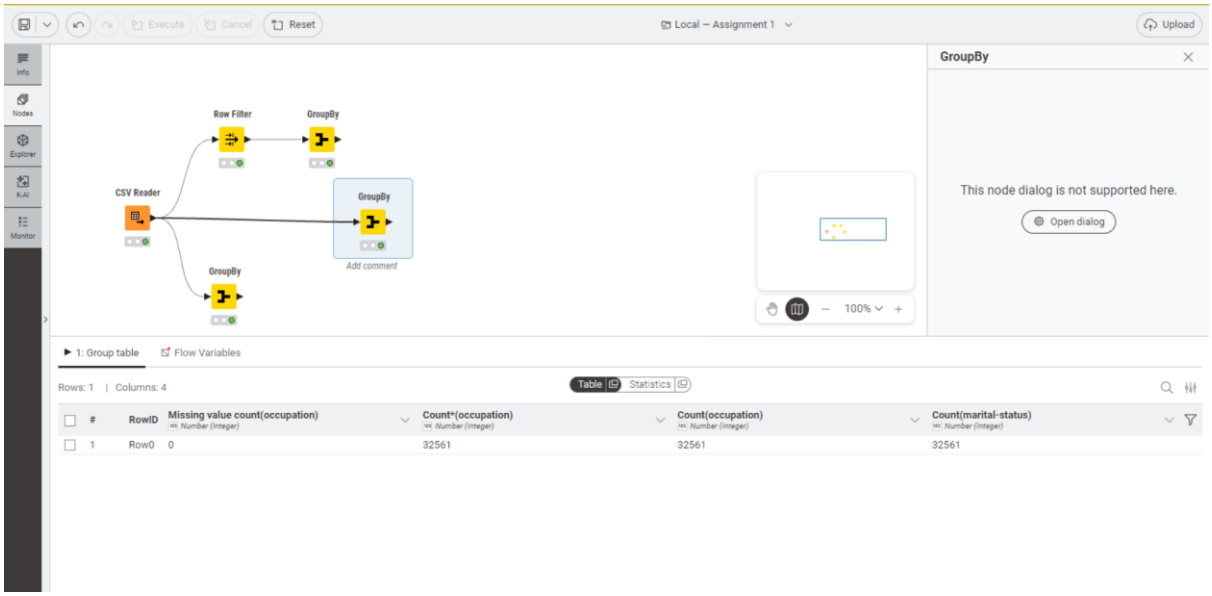


**Step 4:** Use GroupBy node to calculate the average of all numerical column for each of the 4- group defined by sex and income value




| # | 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




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