

Power BI and KNIME 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”

The screenshot shows a KNIME workflow interface. On the left, there's a vertical toolbar with icons for Info, Notes, Explorer, K-Means, and Monitor. The main workspace contains a flow diagram with several nodes: a 'CSV Reader' node at the bottom left, followed by a 'Row Filter' node, a 'GroupBy' node, another 'GroupBy' node, and a final 'GroupBy' node at the bottom right. Arrows connect these nodes sequentially. To the right of the flow diagram is a 'CSV Reader' dialog box with the message "This node dialog is not supported here." and a "Open dialog" button. Below the flow diagram is a 'File Table' view showing the first few rows of the 'adult.csv' dataset. The table has 15 columns: RowID, #, age, workclass, fnlwgt, education, education-num, marital-status, occupation, relations, race, sex, capital-gain, capital-loss, hours-per-week, and capital-loss-number. The data includes various categories like 'Private', 'HS-grad', 'Bachelors', etc., and numerical values like 2174, 0, 40, etc.

Power BI and KNIME Assignment 1

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

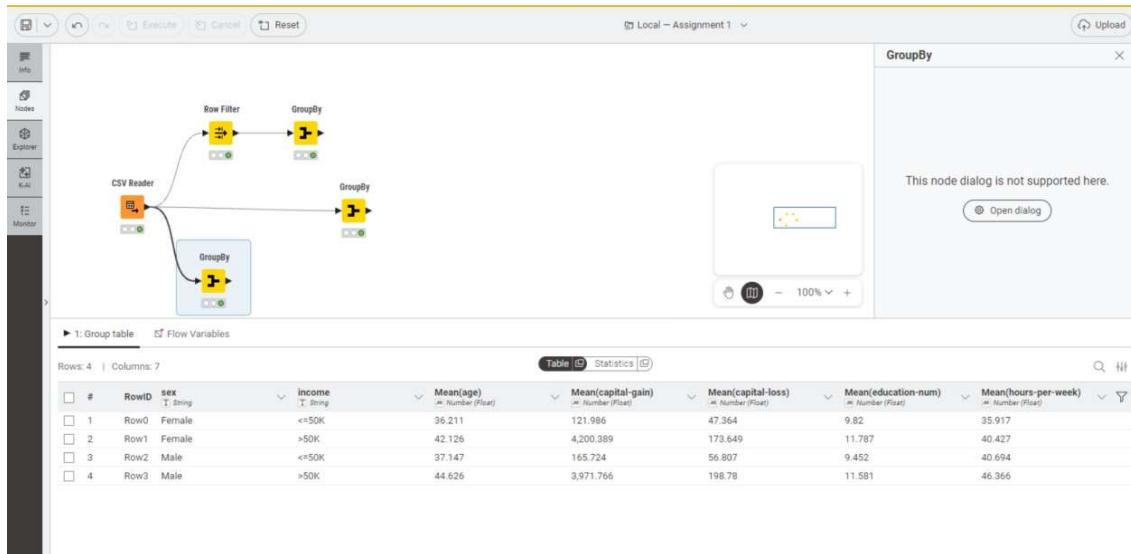
The screenshot shows a KNIME workflow titled "Local - Assignment 1". The workflow starts with a "CSV Reader" node, followed by a "Row Filter" node with the condition "sex = Female". This is followed by two "GroupBy" nodes. The first "GroupBy" node has no output variables selected. The second "GroupBy" node has the output variable "Count(*)" selected. A "Table" view shows the results of the first "GroupBy" node, which includes columns for "workclass", "fnlwgt", "education", "education-num", "marital-status", "occupation", "relations", "race", "sex", "capital-gain", "capital-loss", "hours-per-week", "native-country", and "income". The second "GroupBy" node's output is not visible in the screenshot.

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

The screenshot shows a KNIME workflow titled "Local - Assignment 1". The workflow starts with a "CSV Reader" node, followed by a "Row Filter" node with the condition "sex = Female". This is followed by two "GroupBy" nodes. The first "GroupBy" node has no output variables selected. The second "GroupBy" node has the output variables "Count(*)" and "Mean(age)" selected. A "Table" view shows the results of the second "GroupBy" node, which includes columns for "#", "RowID", "Count*(age)", and "Mean(age)". The value for "Count*(age)" is 1179 and the value for "Mean(age)" is 42.126. A tooltip on the second "GroupBy" node states: "This node dialog is not supported here." An "Open dialog" button is also present.

Power BI and KNIME Assignment 1

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



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 marital-status

