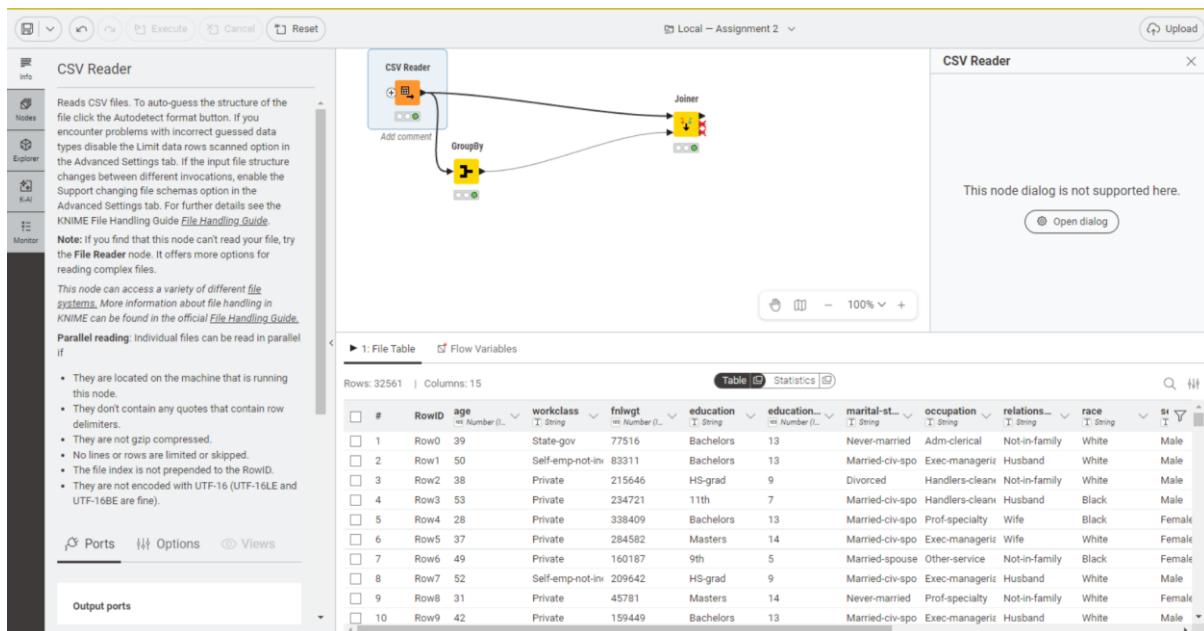


Assignment 02

- 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



Power BI and KNIME

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

The screenshot shows the KNIME interface. On the left, the 'Nodes' panel has 'GroupBy' selected. The main workspace shows a flow starting with a 'CSV Reader' node, followed by a 'GroupBy' node, and then a 'Joiner' node. A configuration dialog for the 'GroupBy' node is open on the right, displaying the following text:

```

Info 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 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.
  
```

Below the configuration dialog is a preview table titled '1: Group table' showing the aggregated data:

| # | RowID | sex | income | Mean(age) | Count(age) |
|---|-------|--------|--------|-----------|------------|
| 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

The screenshot shows the KNIME interface. On the left, the 'Nodes' panel has 'Joiner' selected. The main workspace shows a flow starting with a 'CSV Reader' node, followed by a 'GroupBy' node, and then a 'Joiner' node. A configuration dialog for the 'Joiner' node is open on the right, showing the 'Matching Criteria' section with 'All of the following' selected. The preview table titled '1: Join result' shows the joined data:

| # | sex | capital-g... | capital-i... | hours-per... | native-co... | income | sex (Right) | income (... | Mean(age) | Count(a... |
|----|------|--------------|--------------|--------------|---------------|--------|-------------|-------------|-----------|------------|
| te | Male | 2174 | 0 | 40 | United-States | <=50K | Female | <=50K | 36.211 | 9592 |
| te | Male | 0 | 0 | 13 | United-States | >50K | Female | >50K | 42.126 | 1179 |
| te | Male | 0 | 0 | 40 | United-States | <=50K | Male | <=50K | 37.147 | 15128 |
| sk | Male | 0 | 0 | 40 | United-States | >50K | Male | >50K | 44.626 | 6662 |