Week 4 Practical Lab: Simple ETL Process using KNIME

Useful resources:

- **KNIME Analytics Platform** download link:
  https://www.knime.com/downloads/download-knime
  (It is recommender to download the installer version 561 MB for Windows)
- **Example Data set, Sales Data**
  https://sistc.learnbook.com.au/mod/resource/view.php?id=3632&redirect=1

Objective:

- Implement ETL process using KNIME on a sample data set

Instructions:

1. Open the KNIME software installed on your computer.
2. Create a new workflow from File-> New -> **New KNIME Workflow** -> Next -> *Enter a project name* -> Finish.
3. Select a data set of your choice from the given link (https://archive.ics.uci.edu/ml/index.php) and open the file using the appropriate node reader. (In this example, the following data set is chosen: https://sistc.learnbook.com.au/mod/resource/view.php?id=3632&redirect=1)
4. Execute the File Reader node and observe data through ***File Table*** option

File Table - 0:3 - File Reader (sales records)

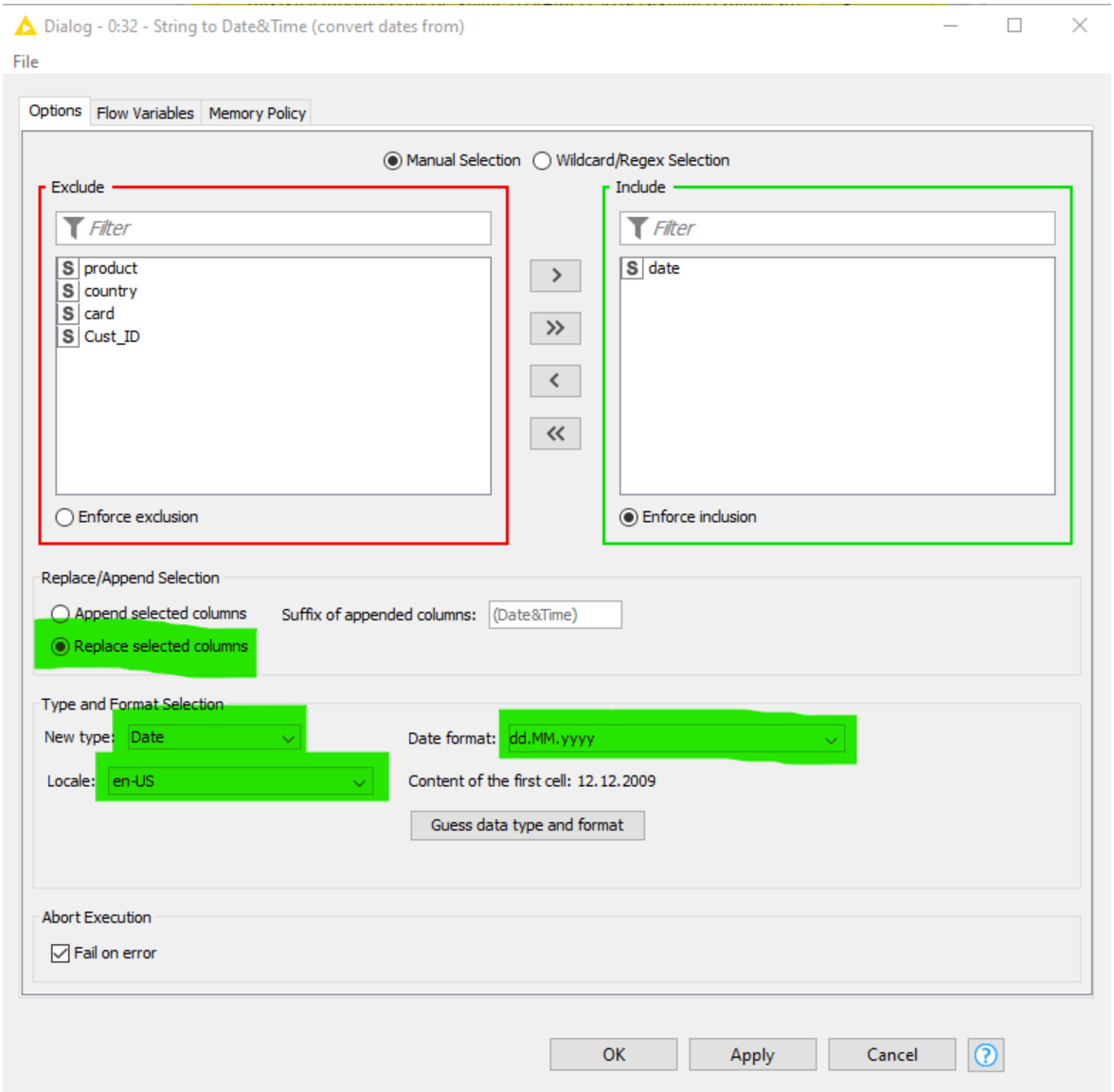
File Edit Hilite Navigation View

Table "sales_2008-2011.csv" - Rows: 47 Spec - Columns: 7 Properties Flow Variables

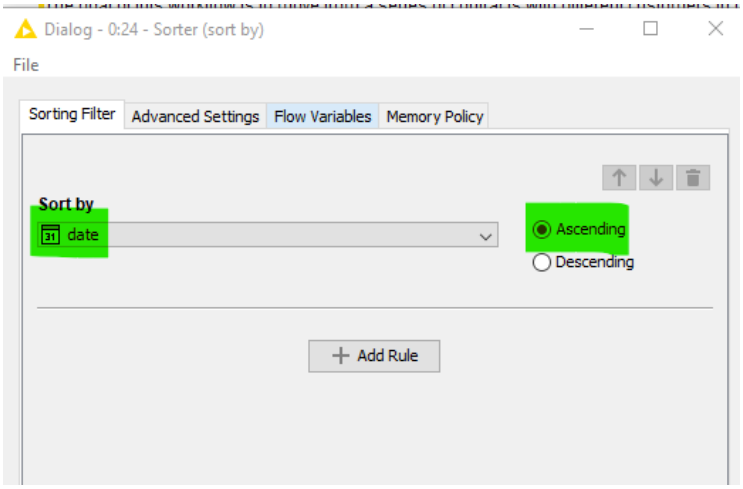
Row ID	product	country	date	quantity	amount	card	Cust_ID
Row0	prod_1	China	12.12.2009	1	35	Y	Cust_2
Row1	prod_2	Germany	01.02.2011	1	40	Y	Cust_1
Row2	prod_3	USA	17.03.2010	1	80	Y	Cust_3
Row3	prod_1	China	28.06.2010	10	350	Y	Cust_5
Row4	prod_2	Germany	31.03.2010	5	200	Y	Cust_4
Row5	prod_3	USA	20.08.2009	20	1600	?	Cust_3
Row6	prod_1	USA	11.10.2010	2	70	Y	Cust_6
Row7	prod_2	Germany	22.11.2009	15	600	N	Cust_1
Row8	prod_3	Germany	13.01.2010	1	80	?	Cust_4
Row9	prod_1	USA	04.07.2009	2	70	Y	Cust_3
Row10	prod_2	USA	20.01.2010	2	80	?	Cust_6
Row11	prod_3	Germany	14.09.2010	2	160	?	Cust_1
Row12	prod_1	Brazil	17.07.2010	5	175	?	Cust_7
Row13	prod_2	USA	07.07.2010	12	480	?	Cust_3
Row14	prod_4	unknown	12.12.2008	1	3	?	Cust_8
Row15	prod_3	China	02.01.2011	8	640	?	Cust_2
Row16	prod_1	Germany	20.03.2011	11	385	N	Cust_4
Row17	prod_2	USA	22.02.2010	6	240	?	Cust_6
Row18	prod_3	China	10.05.2009	2	160	?	Cust_2
Row19	prod_1	Germany	06.03.2011	10	350	?	Cust_4
Row20	prod_2	Germany	22.06.2010	6	240	?	Cust_1
Row21	prod_3	Brazil	08.06.2009	15	1200	?	Cust_7
Row22	prod_1	Germany	02.12.2009	1	35	Y	Cust_1
Row23	prod_2	Germany	11.02.2011	1	40	?	Cust_4
Row24	prod_3	China	12.03.2010	1	80	?	Cust_5
Row25	prod_1	China	28.08.2010	10	350	N	Cust_2
Row26	prod_2	Germany	31.08.2010	5	200	?	Cust_1

a. From the file table, it can be seen that, the date column is in **String** format. Therefore, it needs to converted to **date** format.

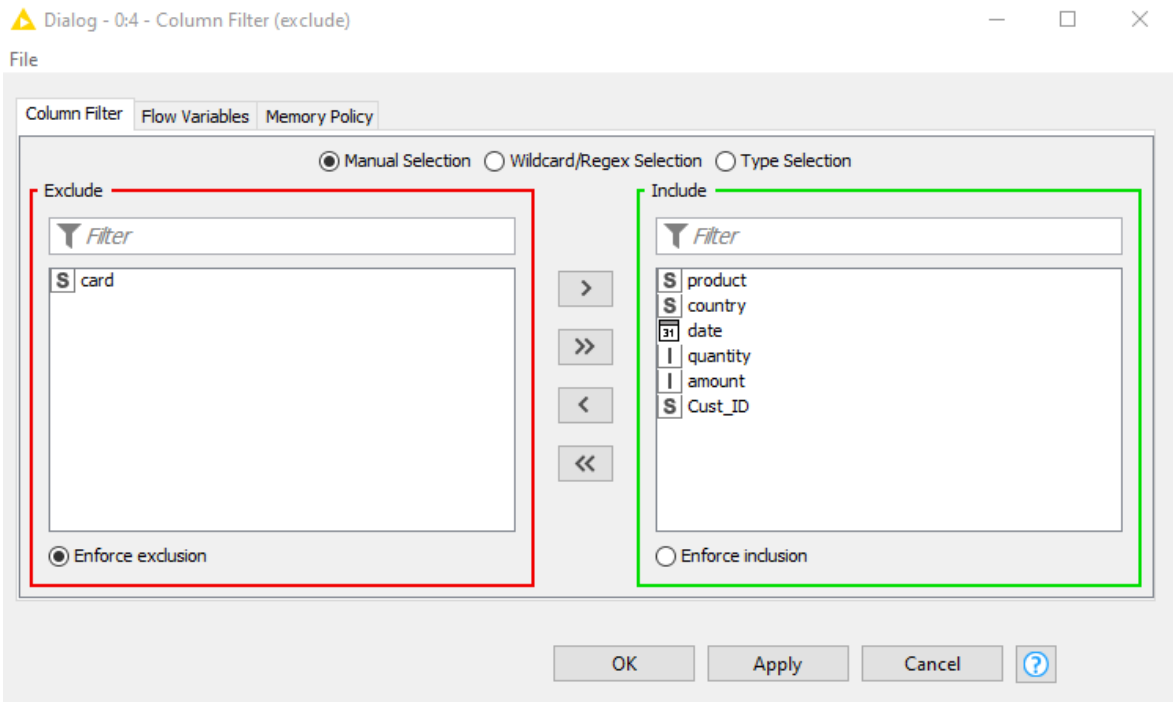
5. To convert the column from String to Date, take the ***String to Date&Time*** node and configured as follows. In other cases, you can convert string to number, number to string etc.



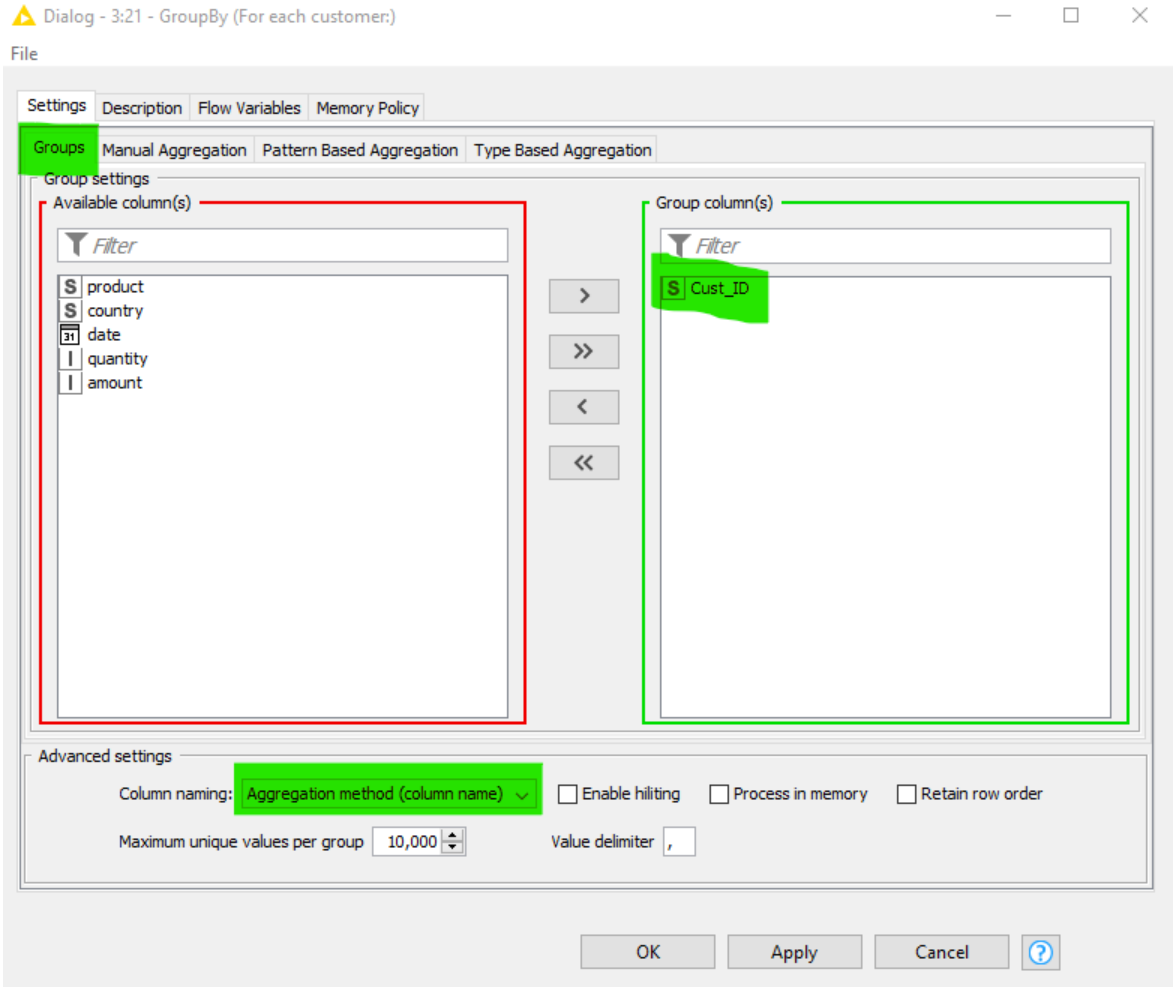
6. To sort the data set based on date, take the **Sorter** node and configure as follows.

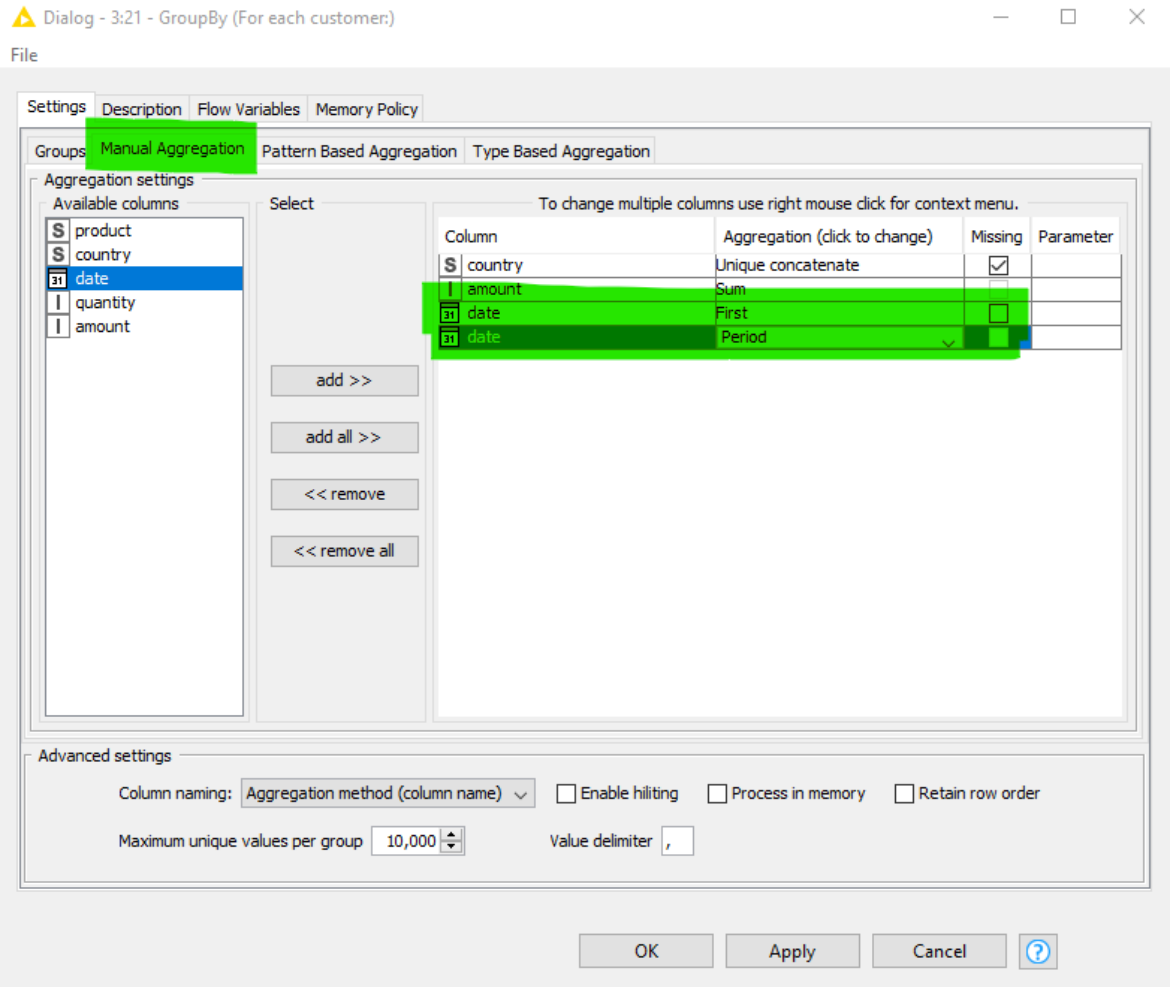


7. To exclude a column, connect the **Column Filter** node and configure as follows.

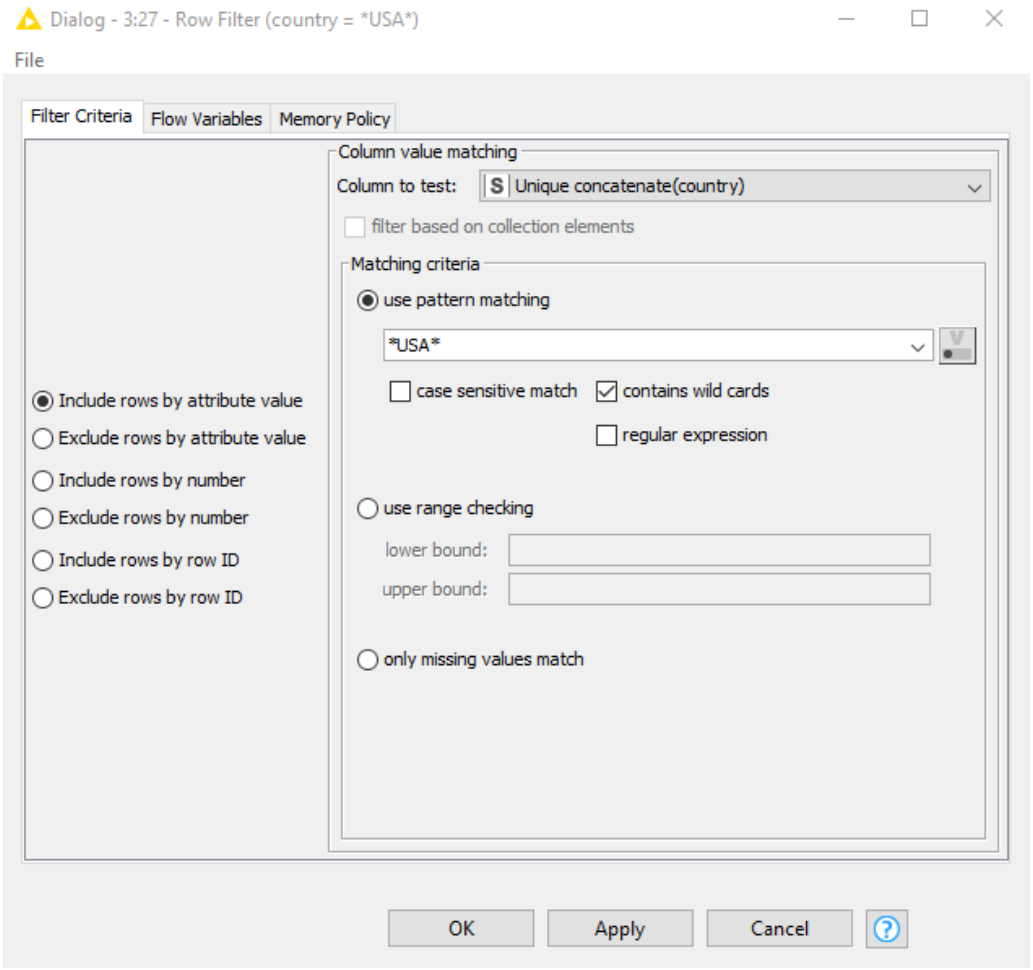


8. To move from a series of contracts to a one-row summary for each customer, select the GroupBy node and configure accordingly.

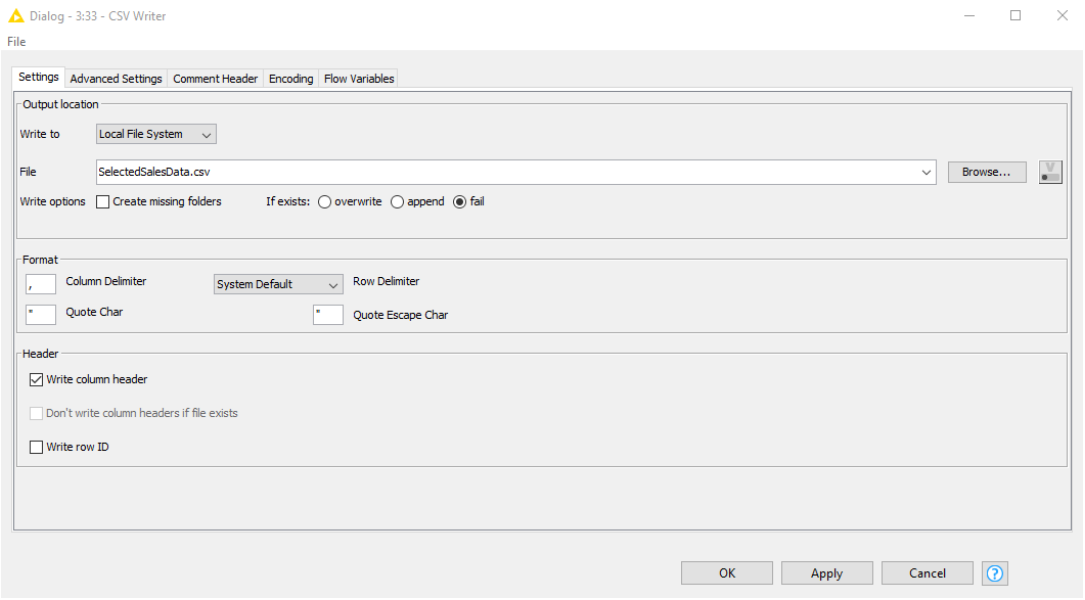

Dialog - 3:21 - GroupBy (For each customer:)
File
Settings
Description
Flow Variables
Memory Policy
Groups
Manual Aggregation
Pattern Based Aggregation
Type Based Aggregation
Group settings
Available column(s)
Filter
product
country
date
quantity
amount
Group column(s)
Filter
Cust_ID
Advanced settings
Column naming:
Aggregation method (column name)
Enable hiliting
Process in memory
Retain row order
Maximum unique values per group
10,000
Value delimiter
,
OK
Apply
Cancel



9. To select customers from a specific country, use the **Row Filter** node and configure as follows:



10. As the last step, write the data as a csv file using the CSV Writer node and execute.



11. The final workflow can be look like this