UNIVERSITÀ DEGLI STUDI DI TERAMO

Big Data Analytics

Maneggiare i dati in KNIME

Prof.ssa Romina Eramo
Università degli Studi di Teramo
Dipartimento di Scienze della Comunicazione
reramo@unite.it

Obiettivo della lezione

» Imparare a gestire le **tabelle di dati** in KNIME
  - **Caricare** e **salvare** dati
  - **Pulire**, **aggregare**, **filtrare**, **combinare** tabelle
  - **Esportare** i risultati

» La lezione termina con un **tutorial completo**:
  - analisi delle **vendite di un negozio online di articoli da regalo**.

Caricare e salvare dati

» Nodi fondamentali della sezione IO (Input/Output)

Operazione	Nodo principale	Descrizione
Lettura file	File Reader / Excel Reader	Carica dati da CSV, TXT, o Excel
Scrittura file	CSV Writer / Excel Writer	Esporta i risultati in formato testuale o Excel

» Suggerimento pratico:

- Trascina direttamente il file nel **Workflow Editor** → KNIME sceglierà automaticamente il nodo di lettura corretto.
- I formati più usati: **CSV** (Comma-Separated Values) e **XLS (Excel)**.

File Reader

Il nodo File Reader

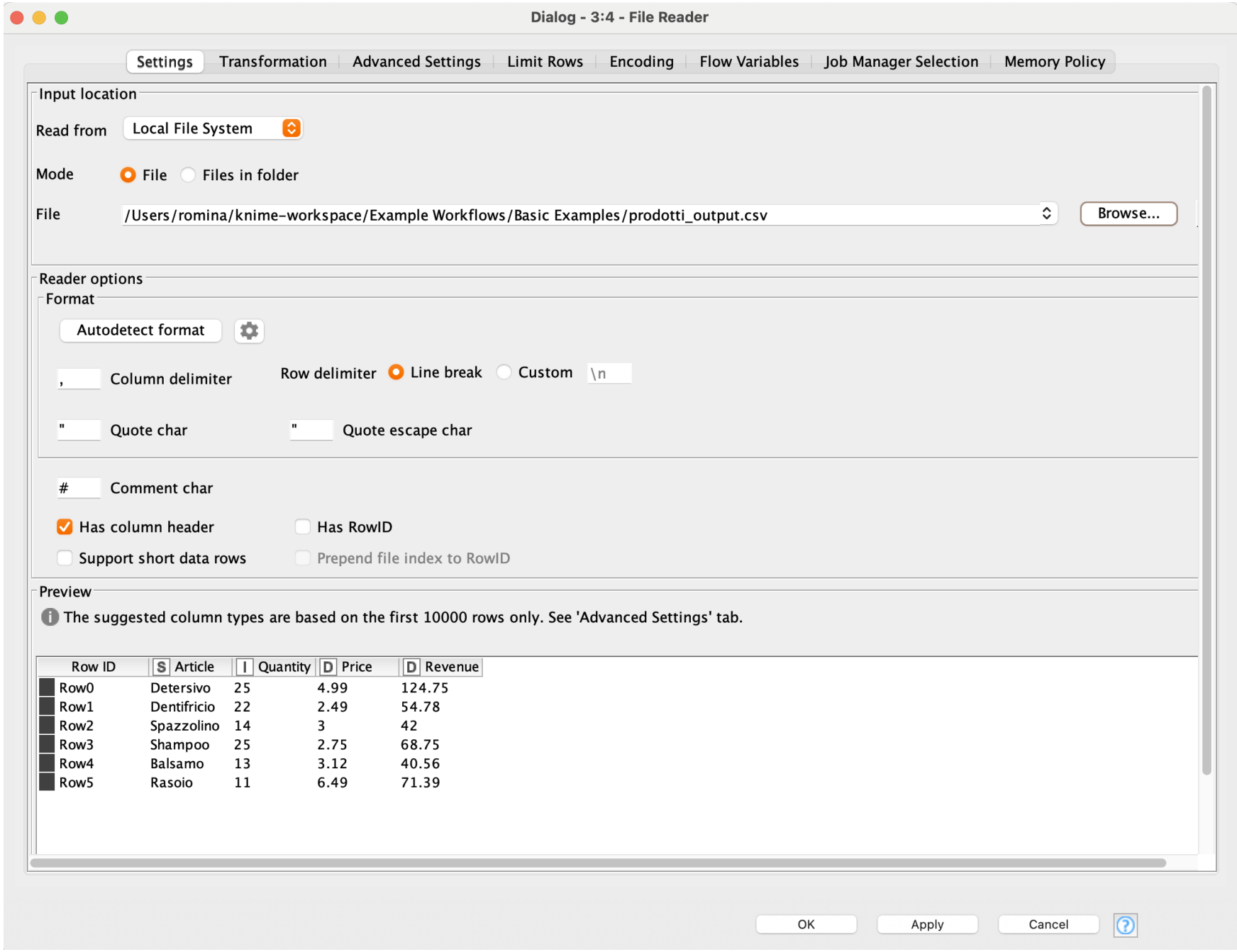
» **Funzione:** leggere file di testo (CSV, TXT, ecc.)

» È il **nodo più versatile e diffuso** per la lettura dei dati

» **Configurazione base** ***(finestra Dialog)***

- Indica **percorso o URL** del file.
- KNIME tenta il **riconoscimento automatico del formato** (separatore, presenza intestazioni, encoding, ecc.).
- Visualizza un'**anteprima della tabella letta**.

» Usa il pulsante **Quick Scan** per analizzare solo le prime 50 righe (utile con file molto grandi).


Dialog - 3:4 - File Reader
Settings
Transformation
Advanced Settings
Limit Rows
Encoding
Flow Variables
Job Manager Selection
Memory Policy
Input location
Read from
Local File System
Mode
File
Files in folder
File
/Users/romina/knime-workspace/Example Workflows/Basic Examples/prodotti_output.csv
Browse...
Reader options
Format
Autodetect format
, Column delimiter
Row delimiter
Line break
Custom
\n
" Quote char
" Quote escape char
Comment char
Has column header
Has RowID
Support short data rows
Prepend file index to RowID
Preview
The suggested column types are based on the first 10000 rows only. See 'Advanced Settings' tab.
Row ID | S Article | I Quantity | D Price | D Revenue
Row0 | Detersivo | 25 | 4.99 | 124.75
Row1 | Dentifricio | 22 | 2.49 | 54.78
Row2 | Spazzolino | 14 | 3 | 42
Row3 | Shampoo | 25 | 2.75 | 68.75
Row4 | Balsamo | 13 | 3.12 | 40.56
Row5 | Rasoio | 11 | 6.49 | 71.39
OK
Apply
Cancel

Funzionalità avanzate del File Reader

» **Dalla finestra di configurazione Dialog** è possibile gestire:

- **Short Lines** → cosa fare con righe incomplete
- **Limit Rows** → quante righe saltare o leggere
- **Ignore Spaces** → rimuovere spazi finali
- **Rename columns** o **cambiare data type** direttamente dall'anteprima
- ….

 *Il File Reader è l'alleato ideale per leggere file testuali complessi o non standard.*

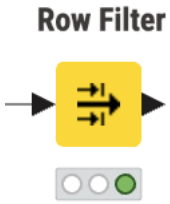
Ripulire i dati

» **Perché è necessario ripulire i dati?**
- Ogni processo di acquisizione dati può generare errori.
- Formati non rispettati, campi mancanti, valori errati o corrotti.
- Anche file trasmessi possono deteriorarsi.
- Tecniche di machine learning sono robuste, ma **errori in ingresso = errori in uscita**.

» **Concetto chiave**
- **GIGO – Garbage In, Garbage Out**
- Se entrano dati sporchi, usciranno risultati inaffidabili.

» **Obiettivo del preprocessing**
- Eliminare errori e incoerenze.
- Filtrare valori scorretti o inutili.
- Gestire stringhe e anomalie nelle tabelle in ingresso.

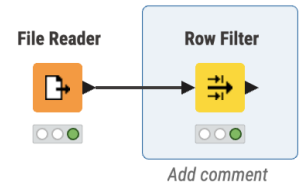
Il nodo Row Filter



» **Cos'è il Row Filter?**
- Un nodo che permette di **selezionare o eliminare righe** di una tabella.
- Rimuove le righe che **non soddisfano un criterio di validità**.

» **Modalità di selezione**
- **By attribute value** → in base ai valori di una colonna.
- **By number** → in base al numero di riga.
- **By row ID** → in base all'identificativo di riga.

» **Due funzioni principali**
- **Output matching rows** → mantiene solo le righe che rispondono al criterio.
- **Output non-matching rows** → elimina le righe che rispondono al criterio.



Row Filter

Filter

Criterion 1

Filter column: Revenue

Operator: Greater than

Value: 50

Add filter criterion

Output

Column domains: Retain | Compute

Filter behavior: Output matching rows | Output non-matching rows

Discard | Apply and Execute | Apply

100%

▶ 1: Included Rows | Flow Variables

Rows: 4 | Columns: 4

Table | Statistics

#	RowID	Article String	Quantity Number (Integer)	Price Number (Float)	Revenue Number (Float)
1	Row0	Detersivo	25	4.99	124.75
2	Row1	Dentifricio	22	2.49	54.78
3	Row3	Shampoo	25	2.75	68.75
4	Row5	Rasoio	11	6.49	71.39

Row Filter: criteri

1. **Pattern matching**
   - Usa * (qualsiasi numero di caratteri) e ? (un carattere).
   - Esempio: mantenere righe in cui una colonna inizia per “a”.
   - Possibile uso di **espressioni regolari**.
2. **Range checking**
   - Imposta un valore minimo e/o massimo.
   - Esempio: mantenere solo righe con valore ≥ 0.
3. **Only missing value match**
   - Identifica righe con valori mancanti.
   - Elimina o mantiene righe con il campo vuoto.
4. **Include/Exclude rows by number**
   - Seleziona righe in base alla loro posizione nella tabella.
   - Esempi:
   - mantenere solo le prime 10 righe;
   - escludere le ultime 100 righe.

Le espressioni regolari, o *regex* , sono molto utili se si ha a che fare con stringhe di testo. Per saperne di più potete consultare il manuale *Espressioni regolari* di Marco Berio (Apogeo, 2007), o cercare online per ulteriori tutorial. Per esercitarvi e testare le vostre espressioni regolari potete usare una webapp come `https://regexr.com/`.

Tabella 3.7 Metacaratteri più comuni usati nelle espressioni regolari.

Metacarattere	Utilizzo	Esempio
`.`	Indica un qualsiasi singolo carattere.	`.iao` identifica stringhe che sono formate da un qualsiasi carattere seguito da `iao`, come `ciao` e `miao`.
`*`	Indica zero o più occorrenze del carattere precedente.	`a*bc` identifica stringhe che iniziano con 0 o più ripetizioni di `a` seguite da `bc`, come `abc`, `aabc`, `bc`. Di conseguenza, `.*` indica qualsiasi stringa mentre `.*ic.*` identifica tutte le stringhe che includono `ic` in una qualsiasi posizione come `bici`, `icaro`, `tic`.
`+`	Indica una o più occorrenze del carattere precedente.	`a+bc` identifica stringhe che iniziano con una `a` o ripetizioni di `a`, seguite da `bc`, come `abc` e `aabc` (non `bc`). Di conseguenza, `.+` indica ogni stringa non vuota: `.*ic.+` identifica tutte le stringhe che includono `ic` ma hanno almeno un carattere dopo `ic`, come `bici, icaro` (non `tic` o `bic`).
`[ ]`	Indica un singolo carattere tra quelli inclusi tra parentesi. Con `-` si identificano gli intervalli.	`[abc]iao` (o `[a-c]iao`) identifica le stringhe che iniziano con `a`, `b` o `c`, seguiti da `iao` come `ciao` (ma non `miao`).
`[^ ]`	Indica ogni singolo carattere diverso da quelli inclusi tra parentesi.	`[^abc]iao` identifica le stringhe che iniziano per qualsiasi carattere, tranne `a`, `b` o `c` e che proseguono con `iao` come `miao` (ma non `ciao`).
`^`	Indica l'inizio della stringa.	`^[cC].*` indentifica le stringhe che iniziano con la lettera `c` minuscola o maiuscola.
`$`	Indica la fine della stringa.	`.*[AEIOUaeiou]$` indentifica le stringhe che finiscono con una vocale.

Missing Value

» **Perché gestire i valori mancanti?**
- Celle vuote possono alterare analisi e risultati.
- Esempio: risposte mancanti in un sondaggio.
- Serve una strategia univoca per trattare i *missing*.

» **Funzione del nodo**
- Definisce cosa fare con celle vuote per ogni singola colonna.
- Impostazioni nel pannello *Column Settings*.

Missing Value: Opzioni disponibili

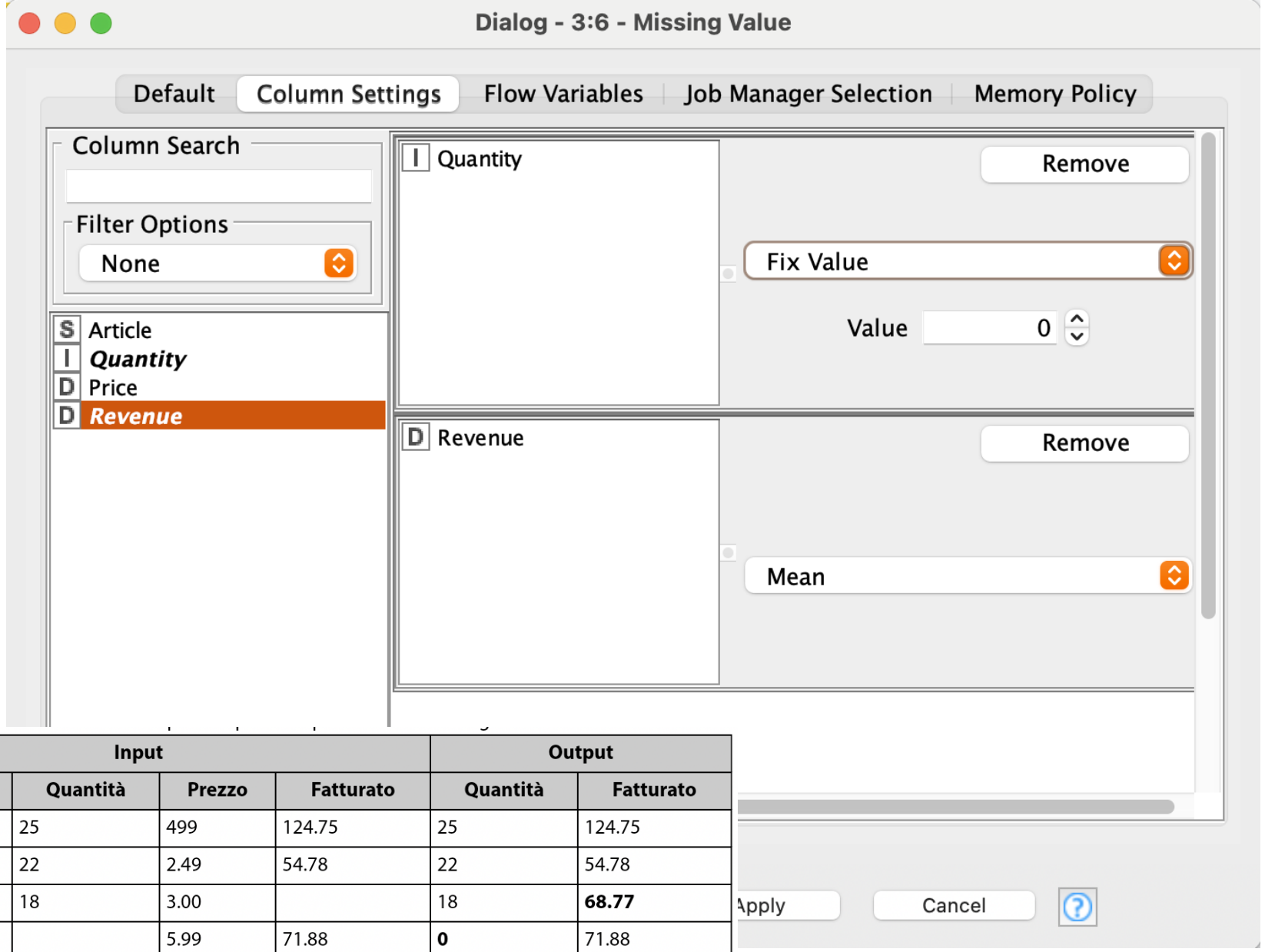
» **Strategie di sostituzione**
- **Fix Value** → sostituzione con valore fisso.
- **Mean/Median** → media o mediana.
- **Maximum/Minimum** → valore massimo/minimo.
- **Most Frequent Value** → valore più ricorrente.
- **Previous/Next Value** → valore della riga precedente/successiva.

» **Strategie avanzate**
- **Linear Interpolation** → interpolazione tra righe vicine.
- **Remove Row** → eliminare la riga contenente valori mancanti.

Missing Value: Impostazioni globali

» **Pannello Default**
  - Permette di definire un'unica regola per:
  - colonne testuali,
  - colonne numeriche intere,
  - colonne con valori decimali.

» **Esempio (vedi slide successiva)**
  - Valore mancante in *Fatturato* → riempito con la media (68.77).
  - Valore mancante in *Quantità* → sostituito con valore fisso (0).



Esempio:

Input				Output	
Prodotto	**Quantità**	**Prezzo**	**Fatturato**	**Quantità**	**Fatturato**
Detersivo	25	499	124.75	25	124.75
Dentifricio	22	2.49	54.78	22	54.78
Spazzolino	18	3.00		18	**68.77**
Collutorio		5.99	71.88	**0**	71.88
Shampoo	25	2.75	68.75	25	68.75
Balsamo	13	3.12	40.56	13	40.56

String Manipulation

» Molte tabelle contengono colonne con **testo**, non solo numeri.

» Le stringhe richiedono **funzioni specifiche** per la loro manipolazione.

» Il nodo *String Manipulation* offre molti strumenti per pulire e trasformare il testo.

Selezione delle funzioni per stringhe

Funzione	Descrizione	Esempio
upperCase (x)	Riporta tutte le lettere della stringa in maiuscolo. Esiste anche la funzione opposta lowerCase().	`upperCase("tessa") =TESSA"`
strip(x)	Elimina gli spazi all'inizio e alla fine della stringa.	`strip(" ciao ")="ciao"`
length(x)	Conta tutti i caratteri, spazi inclusi, della stringa.	`length("KNIME")=5`
compare (x,y)	Compara le stringhe *x* e *y*. Se uguali, restituisce 0, altrimenti 1 o -1 in base al loro ordinamento relativo.	`compare("abc","abc")=0`
replace (x,y,z)	Rimpiazza tutte le occorrenze della sottostringa *y* individuate nella stringa *x* e le sostituisce con *z*.	`replace("ciao come","c","t")="tiao tome")`
join(x,y)	Unisce le stringhe, concatenandole.	`join("Ci","ao")="Ciao"`
toDouble (x)	Converte la stringa in valore numerico decimale. Esistono anche toInt(), toLong() ecc.	`toDouble("32")=32.0`

Come usare String Manipulation

» **Interfaccia (*Dialog*)**
  - Al centro: elenco delle funzioni disponibili.
  - A destra: descrizione ed esempi della funzione selezionata.
  - A sinistra: colonne della tabella da manipolare.

» **Aggiungere una funzione**
  - Doppio clic sulla funzione nell'elenco centrale.

» **Applicare la funzione a una colonna**
  - Doppio clic sulla colonna nell'elenco a sinistra.

» **Modificare la funzione**
  - Editare la formula nella casella di testo al centro.

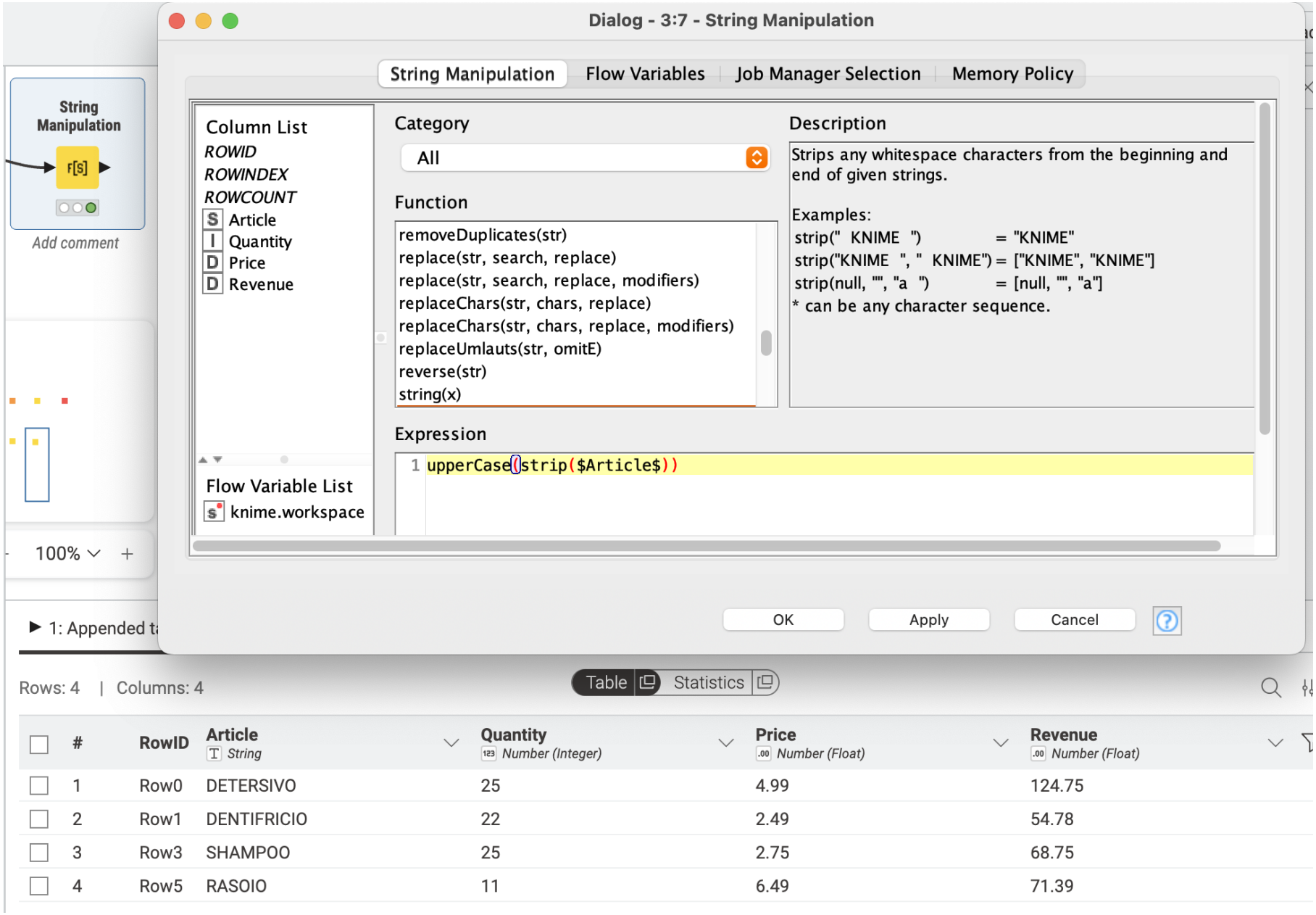
Esempio pratico

» **Obiettivo**
  - Uniformare il testo nella colonna *Prodotto*.

» **Passaggi:**
  - Eliminare spazi superflui → strip()
  - Convertire in maiuscolo → upperCase()
  - Applicare alla colonna *Prodotto*

» **Risultato**
  - upperCase(strip($Prodotto$))

» **Output**
  - Inserire come nuova colonna (*Append Column*)
  - Oppure sostituire quella esistente (*Replace Column*)



#	RowID	Article String	Quantity Number (Integer)	Price Number (Float)	Revenue Number (Float)
1	Row0	DETERSIVO	25	4.99	124.75
2	Row1	DENTIFRICIO	22	2.49	54.78
3	Row3	SHAMPOO	25	2.75	68.75
4	Row5	RASOIO	11	6.49	71.39

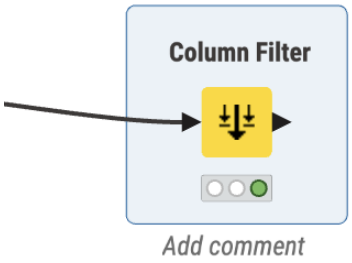
Column Filter

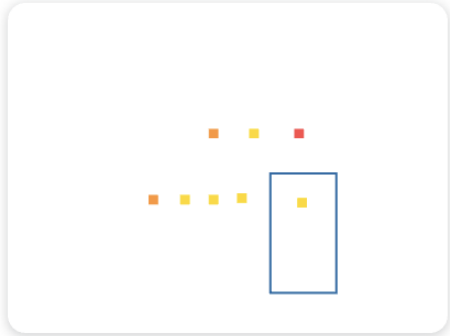
» **Perché usarlo?**

- Troppe colonne rallentano il workflow
- Rendono difficile la lettura e l'analisi
- Eliminare le colonne inutili migliora chiarezza ed efficienza

» **Come funziona?**

- Doppio clic sulle colonne da escludere.
- Le colonne escluse finiscono nella lista dedicata.
- Piccolo gesto → **grande miglioramento nella leggibilità del workflow**.





100%

Column Filter

Column filter

Manual | Wildcard | Regex | Type

Search Aa

Excludes	Includes
Price	Article
	Quantity
	Revenue

Discard | Apply and Execute | Apply

▶ 1: Filtered table | Flow Variables

Rows: 4 | Columns: 3

Table | Statistics

#	RowID	Article *String*	Quantity *Number (Integer)*	Revenue *Number (Float)*
1	Row0	DETERSIVO	25	124.75
2	Row1	DENTIFRICIO	22	54.78
3	Row3	SHAMPOO	25	68.75
4	Row5	RASOIO	11	71.39

Combinare le tabelle

» **Perché combinare tabelle?**
  - I dati sono spesso distribuiti tra più tabelle.
  - Ogni tabella contiene informazioni parziali.
  - Per analizzare categorie, calcolare valori aggregati o incrociare informazioni → serve unire le tabelle.
  - La tecnica usata è la **join (congiunzione relazionale)**.

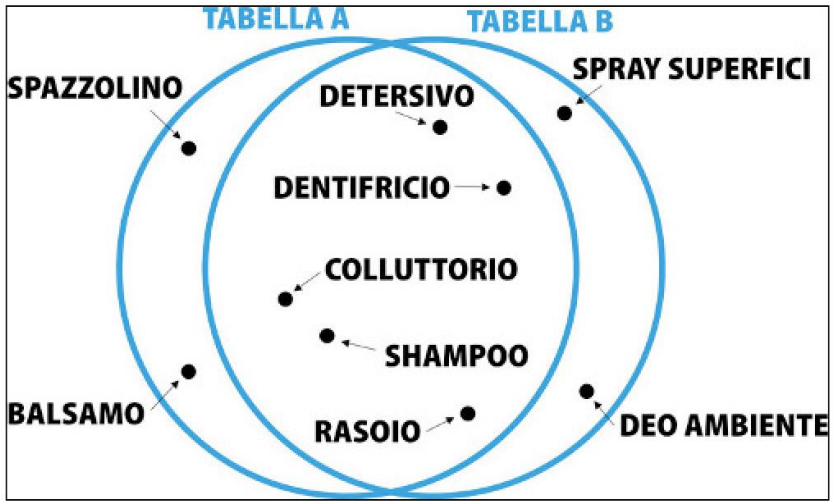
» **Obiettivo**
  - Creare una **terza tabella** che rappresenti la combinazione delle due tabelle di origine.

TABELLA A			TABELLA B	
Prodotto	**Quantità**	**Fatturato**	**Prodotto**	**Categoria**
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	SPRAY SUPERFICI	Cura casa
SPAZZOLINO	18	68.773	DEO AMBIENTE	Cura casa
COLLUTORIO	0	71.88	DENTIFRICIO	Igiene orale
SHAMPOO	25	68.75	COLLUTORIO	Igiene orale
BALSAMO	13	40.56	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona

Quanto fatturato è generato per ogni categoria merceologica?

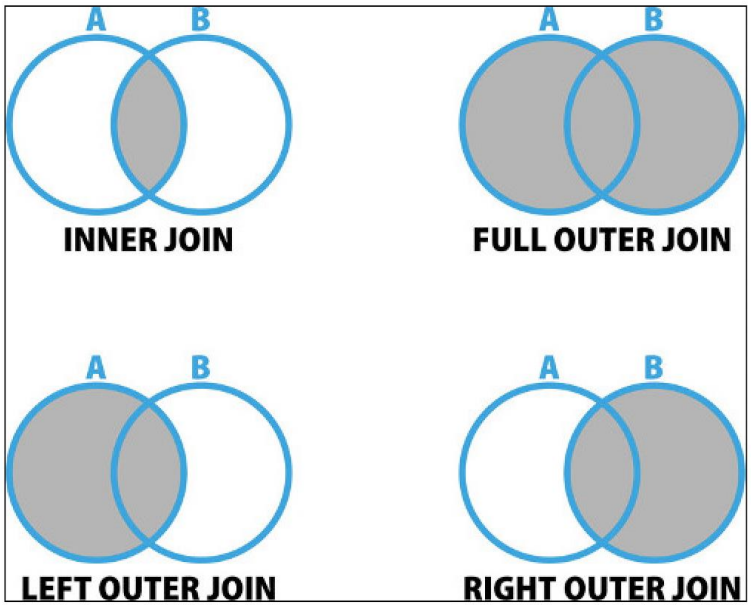
TABELLA A			TABELLA B	
Prodotto	**Quantità**	**Fatturato**	**Prodotto**	**Categoria**
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	SPRAY SUPERFICI	Cura casa
SPAZZOLINO	18	68.773	DEO AMBIENTE	Cura casa
COLLUTORIO	0	71.88	DENTIFRICIO	Igiene orale
SHAMPOO	25	68.75	COLLUTORIO	Igiene orale
BALSAMO	13	40.56	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona

- Le due tabelle condividono una chiave comune (la colonna Prodotto) che identifica in maniera univoca gli elementi
- Una parte dei prodotti sono presenti sia in tabella A sia in tabella B, altri sono solo presenti in una delle due tabelle



Tipi di Join

- **Inner Join**
  - Mantiene solo le righe presenti in entrambe le tabelle.
- **Full Outer Join**
  - Mantiene tutte le righe di entrambe le tabelle, con valori NULL dove mancano corrispondenze.
- **Left Outer Join**
  - Mantiene tutte le righe della tabella sinistra; la tabella destra contribuisce quando possibile.
- **Right Outer Join**
  - Mantiene tutte le righe della tabella destra; la tabella sinistra contribuisce quando possibile.



Inner Join

- **Risultato:** solo elementi presenti sia in A sia in B.
- Gli elementi non comuni vengono esclusi.

Full Outer Join

- **Risultato:** tutti gli elementi di entrambe le tabelle.
- Dove mancano valori → NULL.

Left / Right Join

- Left Join: priorità alla tabella A.
- Right Join: priorità alla tabella B.

INNER JOIN

Prodotto.A	Quantità	Fatturato	Prodotto.B	Categoria
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	DENTIFRICIO	Igiene orale
COLLUTORIO	0	71.88	COLLUTORIO	Igiene orale
SHAMPOO	25	68.75	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona

FULL OUTER JOIN

Prodotto.A	Quantità	Fatturato	Prodotto.B	Categoria
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	DENTIFRICIO	Igiene orale
COLLUTORIO	0	71.88	COLLUTORIO	Igiene orale
SHAMPOO	25	68.75	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona
SPAZZOLINO	18	68.73	?	?
BALSAMO	13	40.56	?	?
?	?	?	SPRAY	Cura casa
?	?	?	DEO AMBIENTE	Cura casa

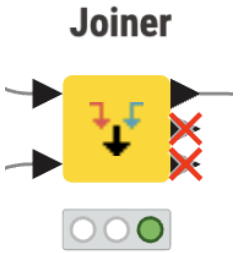
LEFT OUTER JOIN

Prodotto.A	Quantità	Fatturato	Prodotto.B	Categoria
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	DENTIFRICIO	Igiene orale
COLLUTORIO	0	71.88	COLLUTORIO	Igiene orale
SHAMPOO	25	68.75	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona
SPAZZOLINO	18	68.73	?	?
BALSAMO	13	40.56	?	?

RIGHT OUTER JOIN

Prodotto.A	Quantità	Fatturato	Prodotto.B	Categoria
DETERSIVO	25	124.75	DETERSIVO	Cura tessuti
DENTIFRICIO	22	54.78	DENTIFRICIO	Igiene orale
COLLUTORIO	0	71.88	COLLUTORIO	Igiene orale
SHAMPOO	25	68.75	SHAMPOO	Cura persona
RASOIO	8	51.92	RASOIO	Cura persona
?	?	?	SPRAY	Cura casa
?	?	?	DEO AMBIENTE	Cura casa

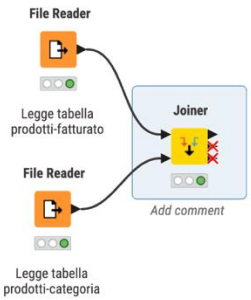
Il nodo Joiner in KNIME

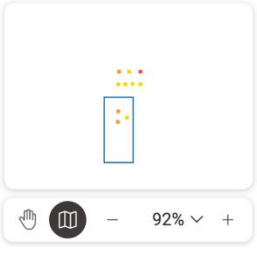


» **Funzioni principali**
  - Combina due tabelle in ingresso.
  - Permette di scegliere:
    - » il **tipo di join**,
    - » le **colonne chiave** della combinazione.

» **"Matching Criteria" e "Include in Output"**
  - Definizione delle **Joining Columns** (colonne su cui fare match).
  - Selezione tipo di join (Inner, Left, Right, Full).

» **"Output Columns"**
  - Scelta delle colonne da mantenere o eliminare.
  - Gestione delle colonne duplicate.





Joiner

Matching Criteria

Match

All of the following | Any of the following

Criterion 1

Top input ('left' table)

RowID

Bottom input ('right' table)

RowID

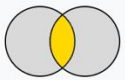
Add matching criterion

Compare values in join columns by

Value and type

Include in Output

- [x] Matching rows
- [ ] Left unmatched rows
- [ ] Right unmatched rows



Discard | Apply and Execute | Apply

Output Columns

Top input ('left' table)

Manual | Wildcard | Regex | Type

Search Aa

Excludes	Includes
No columns in this list.	Article Quantity Price Revenue *Any unknown column*

Bottom input ('right' table)

Manual | Wildcard | Regex | Type

Search Aa

Excludes	Includes
Article	Category *Any unknown column*

- [ ] Merge join columns

If there are duplicate column names

▶ 1: Join result | 2: Left unmatched rows | 3: Right unmatched rows | Flow Variables

Rows: 6 | Columns: 5

Table | Statistics

#	RowID	Article *String*	Quantity *Number (Integer)*	Price *Number (Float)*	Revenue *Number (Float)*	Category *String*
1	Row0_l	Detersivo	25	4.99	124.75	Cura tessuti
2	Row1_l	Dentifricio	22	2.49	54.78	Cura casa
3	Row2_l	Spazzolino	14	3	42	Cura casa
4	Row3_l	Shampoo	25	2.75	68.75	Igiene orale
5	Row4_l	Balsamo	13	3.12	40.56	Igiene orale
6	Row5_l	Rasoio	11	6.49	71.39	Cura persona

Aggregare e disaggregare tabelle

» Spesso serve ottenere:
  - tabelle riassuntive (aggregare)
  - tabelle dettagliate (disaggregare)

» Strumenti:
  - GroupBy
  - Pivoting

GroupBy

» A cosa serve?

- Raggruppare righe che condividono un valore comune.
- Calcolare aggregati (somma, media, conteggi, min/max, ecc.).
- Restituire **una riga per gruppo**.

» Come funziona?

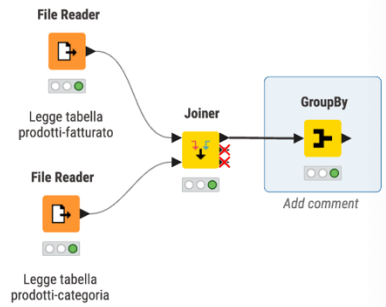
- Selezionare una o più colonne come “chiave” del gruppo.
- Applicare funzioni di aggregazione alle altre colonne.

GroupBy: Aggregazioni

Funzione	Descrizione
Sum	Il più semplice dei metodi di aggregazione, la semplice somma. Esistono anche *sum of logs* e *sum of squares* nel caso in cui vogliate sommare i logaritmi o i quadrati dei valori.
Mean/ Median	Media aritmetica e mediana dei valori nel gruppo. La mediana è il valore centrale della sequenza ordinata dei numeri da aggregare. È da preferire alla media nel caso in cui la popolazione dei numeri da aggregare possa contenere outlier, ovvero valori anomali.
Minimum/ Maximum	Il valore minimo o massimo nel gruppo. Nel caso di date, restituisce la data più indietro nel tempo (*min*) o quella più in avanti (*max*).
First/Last	Il primo (o ultimo) valore incontrato seguendo l'ordine originale della tabella sorgente.
Count	Il conteggio degli elementi nel gruppo. È disponibile anche *Unique count*, che non conta più volte i valori duplicati e ci restituisce il numero dei valori non ripetuti, unici.
Concatenate	Nel caso di stringhe, quest'aggregazione restituisce un'unica stringa che rappresenta la concatenazione di tutte le stringhe nel gruppo. Esiste anche *Unique concatenate*, che include le stringhe ripetute solo una volta.

» Pannello "Manual Aggregation"

- Seleziona colonne e funzioni una per una.
- Possibilità di applicare più aggregazioni alla stessa colonna.



Dialog - 3:11 - GroupBy

Settings | Description | Flow Variables | Job Manager Selection | Memory Policy

Groups | Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Group settings

Available column(s)

Filter

S Article
I Quantity
D Price
D Revenue

Group column(s)

Filter

S Category

Advanced settings

Column naming: Aggregation method (column name)

Enable hiliting | Process in memory | Retain row order

Value delimiter ,

OK | Apply | Cancel

Dialog - 3:11 - GroupBy

Settings | Description | Flow Variables | Job Manager Selection | Memory Policy

Groups | Manual Aggregation | Pattern Based Aggregation | Type Based Aggregation

Aggregation settings

Available columns

S Article
I Quantity
D Price
D Revenue

Select

add >>
add all >>
<< remove
<< remove all

To change multiple columns use right mouse click for context menu.

Column	Aggregation (click to change)	Missi...	Parameter
D Revenue	Sum		
I Quantity	Mean		

Advanced settings

Column naming: Keep original name(s)

Enable hiliting | Process in memory | Retain row order

Maximum unique values per group 10,000

Value delimiter ,

OK | Apply | Cancel

Output:

Rows: 4 | Columns: 3

Table | Statistics

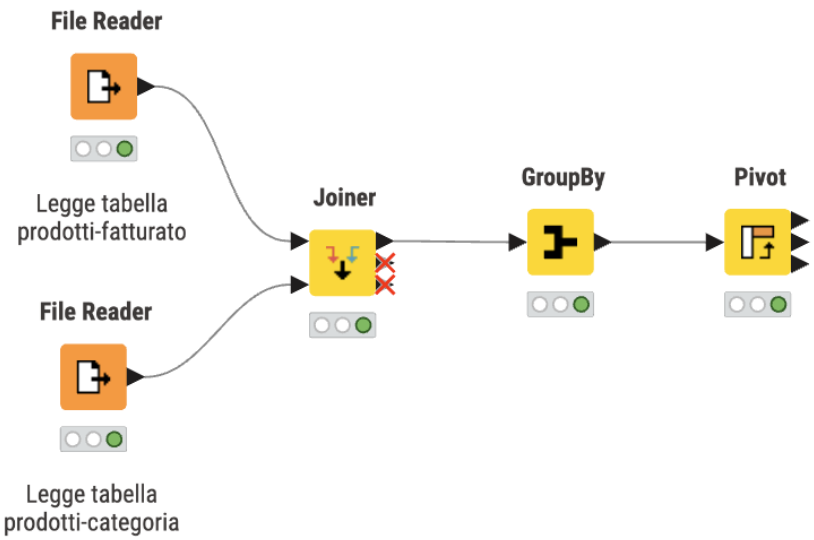
	#	RowID	Category String	Revenue Number (Float)	Quantity Number (Float)
	1	Row0	Cura casa	96.78	18
	2	Row1	Cura persona	71.39	11
	3	Row2	Cura tessuti	124.75	25
	4	Row3	Igiene orale	109.31	19

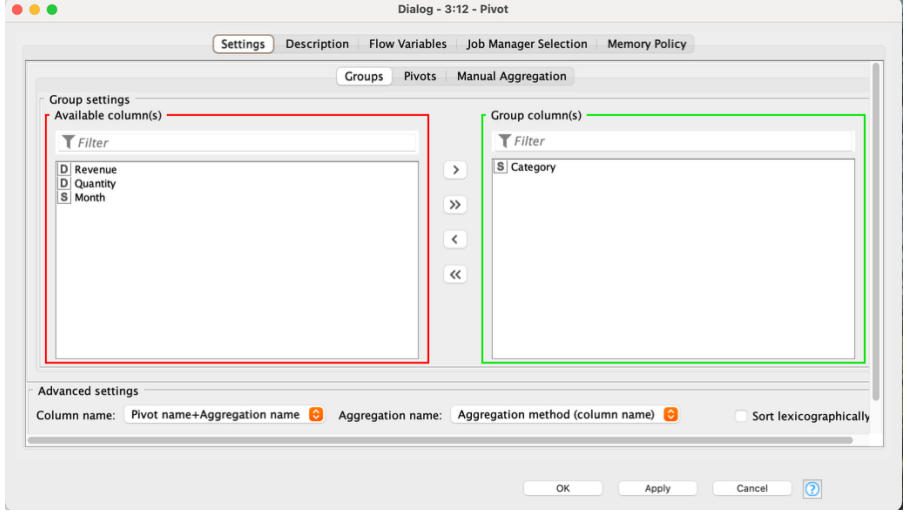
Pivoting

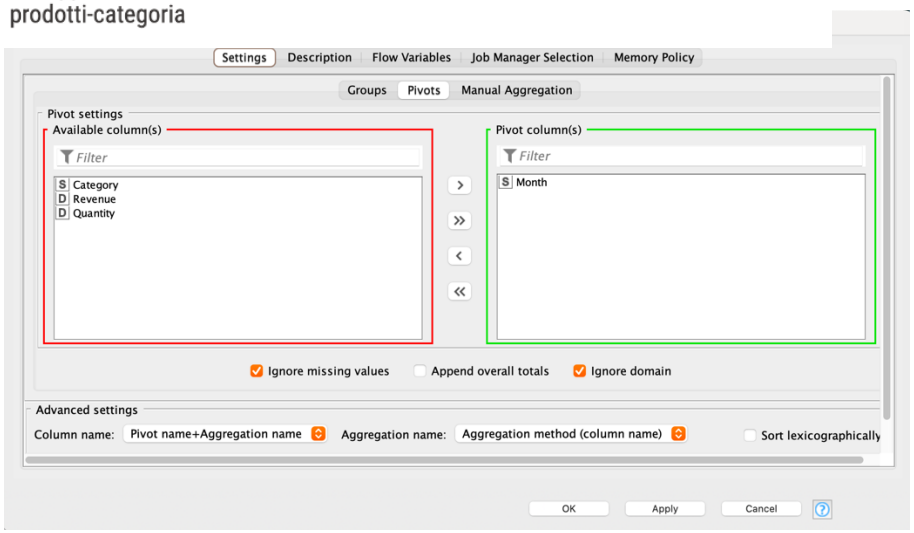
» **Quando usare una tabella pivot?**
  - Quando serve rappresentare gruppi **sia come righe che come colonne**.
  - Simile alle pivot di Excel.

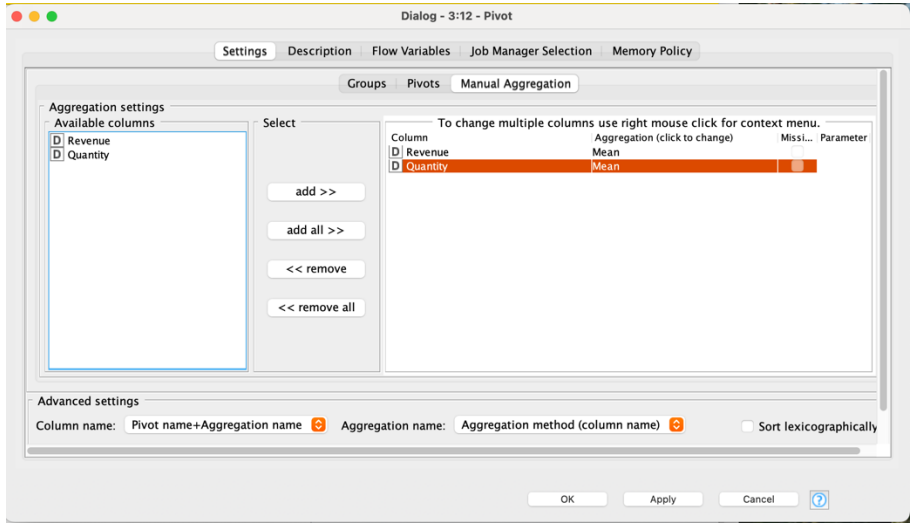
» **Come funziona in KNIME:**
  - Scelta delle colonne di:
    - » **Group** (linee della pivot)
    - » **Pivot** (colonne generate)

» **Aggregation** (valore da calcolare)
  - Output:
    - » Tabella pivotata
    - » Totali per gruppo
    - » Totali per pivot









Categoria	Aprile		Maggio	
	Quantità	Fatturato	Quantità	Fatturato
	Aprile+ Quantità	Aprile+ Fatturato	Maggio+ Quantità	Maggio+ Fatturato
Cura persona	13	23.49	20	36.85
Cura tessuti	11	54.89	14	69.86
Igiene orale	18	36.41	16	26.92

Calcolare formule matematiche

» **Perché servono le formule?**

- I valori delle tabelle spesso devono essere **modificati o trasformati**.
- Occorre calcolare rapporti, somme, differenze, percentuali, medie, ecc.
- Non si tratta ancora di machine learning, ma di **operazioni numeriche di base**.

» **Come farlo in KNIME?**

- Con nodi che **non richiedono codice** (Math Formula).
- Con nodi che **permettono l'uso di linguaggi di programmazione** (Java Snippet).

Math Formula: a cosa serve

» **Funzioni principali**
  - Implementa operazioni numeriche su colonne della tabella.
  - Le colonne diventano **variabili**.
  - Funzioni disponibili nella finestra:
  - round(), average(), ceil(), floor(), sin(), cos(), ecc.

» **Interfaccia**
  - Centro: elenco funzioni con descrizione.
  - Sinistra: variabili/colonne disponibili.
  - Basso: scelta se creare nuova colonna o sostituire quella esistente.

Math Formula: esempio pratico

» **Obiettivo**

- Calcolare **Prezzo = Fatturato / Quantità**

» **Procedura**

- Doppio clic sulla colonna **Fatturato**.
- Digitare **/**.
- Doppio clic sulla colonna **Quantità**.

→ Formula generata: **$Fatturato$ / $Quantità$**

Output

- Append Column → crea nuova colonna “Prezzo”
- Replace Column → sostituisce una colonna esistente

Java Snippet: quando serve

» **Quando usarlo?**

- Quando Math Formula e String Manipulation **non bastano**.
- Per logiche più complesse, cicli, condizioni, manipolazioni avanzate.
- Per introdurre **codice Java** direttamente nel workflow KNIME.

» **Punti di forza**

- Accesso a costrutti logici completi.
- Utilizzo di librerie Java.
- Massima flessibilità nella manipolazione dei dati.

Java Snippet: esempio pratico

» **Calcolo del prezzo**

```
out_Prezzo = c_Fatturato / c_Quantità;
```

» **Dettagli importanti**

- Serve il **punto e virgola** a fine istruzione.
- È necessario dichiarare il campo di output tramite *Add output field*.
- Il risultato è identico a Math Formula, ma il Java Snippet permette:
  - ✓ manipolazioni più complesse
  - ✓ logiche condizionali
  - ✓ operazioni iterative

Variabili speciali nel Java Snippet

» **Variabili predefinite disponibili**

- **ROWID** → ID univoco della riga
- **ROWINDEX** → indice numerico della riga (parte da 0)
- **ROWCOUNT** → numero totale delle righe della tabella

» **Quando servono?**

- Per operazioni “riga per riga” più avanzate
- Per creare indici, numerazioni, controlli condizionali

Math Formula o Java Snippet?

» **Scegli Math Formula se:**
- devi fare operazioni semplici;
- vuoi evitare codice;
- devi lavorare rapidamente su colonne numeriche.

» **Scegli Java Snippet se:**
- ti servono funzioni complesse o personalizzate;
- devi usare condizioni, cicli, librerie;
- vuoi espandere le capacità analitiche del workflow.

Tutorial: Analisi vendite e-commerce

» **Caso**
- Un azienda di e-commerce vuole dotarsi di un sistema di reporting

» **Dataset**
- Vendite di un negozio online di articoli da regalo.
- 3 file CSV: transazioni, anagrafica prodotti, anagrafica clienti.

» **Le domande a cui rispondere**
1. Quali sono i 10 articoli con più fatturato?
2. Quanto ha prodotto la categoria “orologi”?
3. Verso quali nazioni si è venduto di più?
4. In quale momento della giornata si vendono più articoli?

I dati a disposizione (CSV)

Transactions	InvoiceNo	Identificativo univoco della fattura
	StockCode	Codice articolo venduto
	Quantity	Quantità venduta per riga di transazione
	InvoiceDate	Data e ora della transazione
	UnitPrice	Prezzo per unità del prodotto
	CustomerID	Identificativo cliente
	Country	Nazione del cliente (presa dalla transazione)
ProductMaster	StockCode	Codice articolo (chiave univoca)
	Description	Descrizione del prodotto
CustomerMaster	CustomerID	Identificativo cliente (chiave univoca)
	Country	Nazione del cliente (fonte anagrafica)

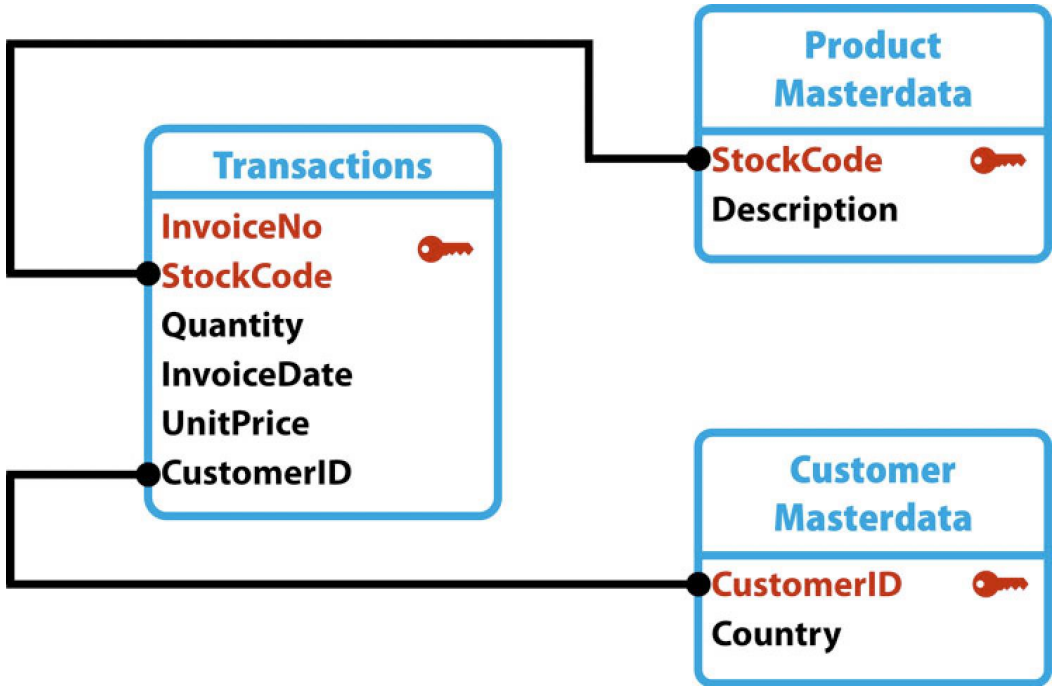
I dati a disposizione (CSV)

» I dataset possono essere scaricati qui:
  – https://hub.knime.com/adm/spaces/Public/Workflows/Data%20Analytics%20Made%20Easy/Chapter%203~xINn600UL4GbSsyw/

Chiavi di JOIN

Transactions	InvoiceNo	Identificativo univoco della fattura
	StockCode	Codice articolo venduto
	Quantity	Quantità venduta per riga di transazione
	InvoiceDate	Data e ora della transazione
	UnitPrice	Prezzo per unità del prodotto
	CustomerID	Identificativo cliente
	Country	Nazione del cliente (presa dalla transazione)
ProductMaster	StockCode	Codice articolo (chiave univoca)
	Description	Descrizione del prodotto
CustomerMaster	CustomerID	Identificativo cliente (chiave univoca)
	Country	Nazione del cliente (fonte anagrafica)

Diagramma entità-relazione semplificato



Osserviamo che:

» Una domanda ci chiede di valutare il fatturato in termini di destinazione della vendita, ovvero della nazione di residenza del cliente.
  - Fatturato e nazione sono in tabelle diverse (**transazioni** e **master data cliente**) che vanno, quindi, combinate;

» un'altra domanda ci richiede di valutare il fatturato di una specifica categoria (gli orologi).
  - Solo la descrizione dell'articolo può permetterci di individuare gli orologi: di conseguenza, è necessario combinare il **master data prodotto** con le **transazioni**.

Svolgimento

- Step 1: Import dei dati
  - Caricamento dei CSV
    - Trascinare i file nell'editor o usare e configurare **File Reader**
    - Controllare intestazioni, separatori, formati
    - Verificare ID di riga e tipi delle colonne
- Step 2: Unire le tabelle
  - Join: Transazioni + Master prodotti
    - Inner Join, per evitare I valori NULL
    - Join su **StockCode**
  - Join Transazioni + Master Clienti
    - Inner Join, per evitare I valori NULL
    - Join su **CustomerID**
      - È necessario prima convertire Marte Clienti(CustomerID) in String

Svolgimento

- Step 3: Pulizia dei dati
  - Gestione valori anomali
    - Quantity < 0 → rimuovere righe
    - UnitPrice = 0 → verificare / filtrare
  - Altri controlli utili
    - Eliminare StockCode speciali (POSTAGE, DISCOUNT…)
      - Row Filter → *Exclude rows by attribute value*
      - Pattern matching o Regex
- Step 4: Identificare gli orologi
  - Come riconoscere la categoria
    - Filtrare per sottocategoria “clock”
    - Aggregare usando Groupby, lasciando i gruppi vuoti

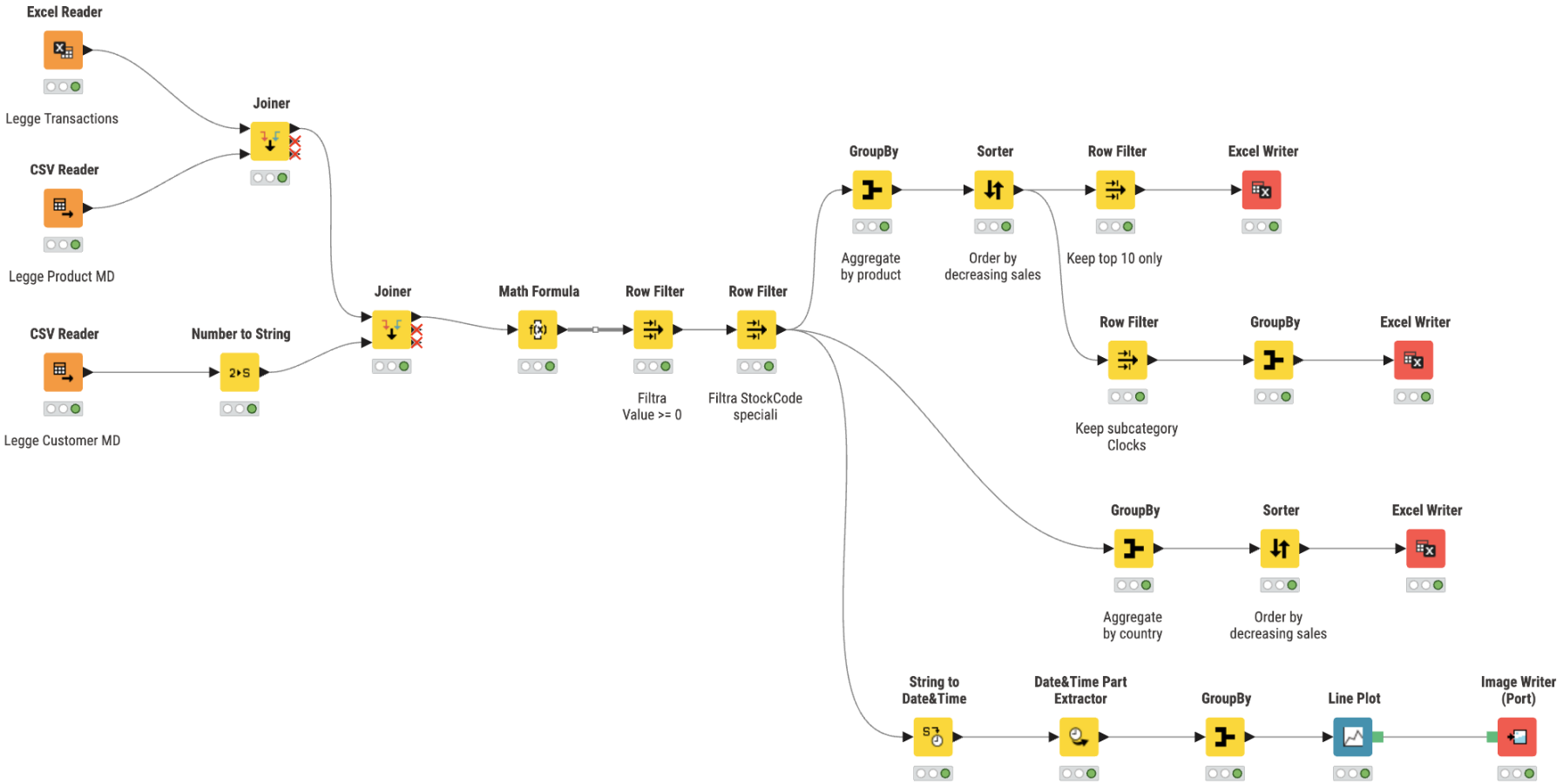
Svolgimento

- Step 5: fatturato per nazione
  - Raggruppare per nazione, aggregando per quantità, prezzo e fatturato
  - Aggiungere il nodo Sorter per per ordinare
- Step 6: Creare variabili temporali
  - Trasformare il campo data/ora
    - Nodo **String to Date&Time**
    - Nodo **Extract Date&Time Fields**
    - estrazione di: ora, giorno, mese, trimestre…
  - Utilità
    - Analisi temporale delle vendite
    - Identificazione fasce orarie strategiche
- Step 7: Aggregazioni
  - Con GroupBy raggruppare in base all'orario e aggregare in base alla somma dei fatturati

Svolgimento

- Step 8: Visualizzazioni
  - E’ possible scrivere i risultati in file Excel (risposte 1,2,3)
  - Grafici disponibili (per la risposta 4)
    - Line Plot → vendite per ora del giorno
    - Bar Chart → prodotti più venduti
    - Pie/Bar → distribuzione per nazione
    - Uso del nodo Image Writer
      - Permette di esportare il grafico come immagine

Workflow finale



Risultati ottenuti

» **Conclusioni dell'analisi**
- Identificati i 10 articoli con più fatturato
- Calcolato il fatturato totale della categoria orologi
- Classifica nazioni con vendite maggiori
- Grafico fatturato per ora del giorno → identificazione ore "di picco"

» **Benefici per il committente**
- Automazione di processi manuali (Excel)
- Migliore qualità dei dati
- Analisi ripetibile e scalabile
- Supporto alle decisioni di marketing e pricing
- Possibilità di estendere il sistema a nuovi dati