Business Intelligence per i Big Data

Esercitazione di laboratorio N. 1

L'obiettivo dell'esercitazione è:

- **utilizzare il software Rapid Miner per effettuare il preprocessing di dati strutturati (relativi ad una campagna promozionale) e di dati non strutturati (ad esempio, dati testuali relativi ad un argomento specifico) per analisi successive.**

Dati strutturati

Il dataset denominato UsersSmall (UsersSmall.xls) è disponibile sul sito del corso (https://dbdmg.polito.it/dbdmg_web/index.php/2022/03/01/business-intelligence-per-big-data/). Esso raccoglie dati anagrafici e lavorativi relativi a circa 300 persone contattate da un'azienda per proporre loro l'iscrizione ad un loro servizio. Per tali utenti è noto se, dopo essere stati contattati, si sono iscritti al servizio proposto oppure no (valore del campo Response).

La lista completa degli attributi del dataset a disposizione (UsersSmall.xls) è riportata di seguito.

(1) Age
(2) Workclass
(3) Education
(4) Marital status
(5) Occupation
(6) Relationship
(7) Race
(8) Sex
(9) Native country
(10) Response.

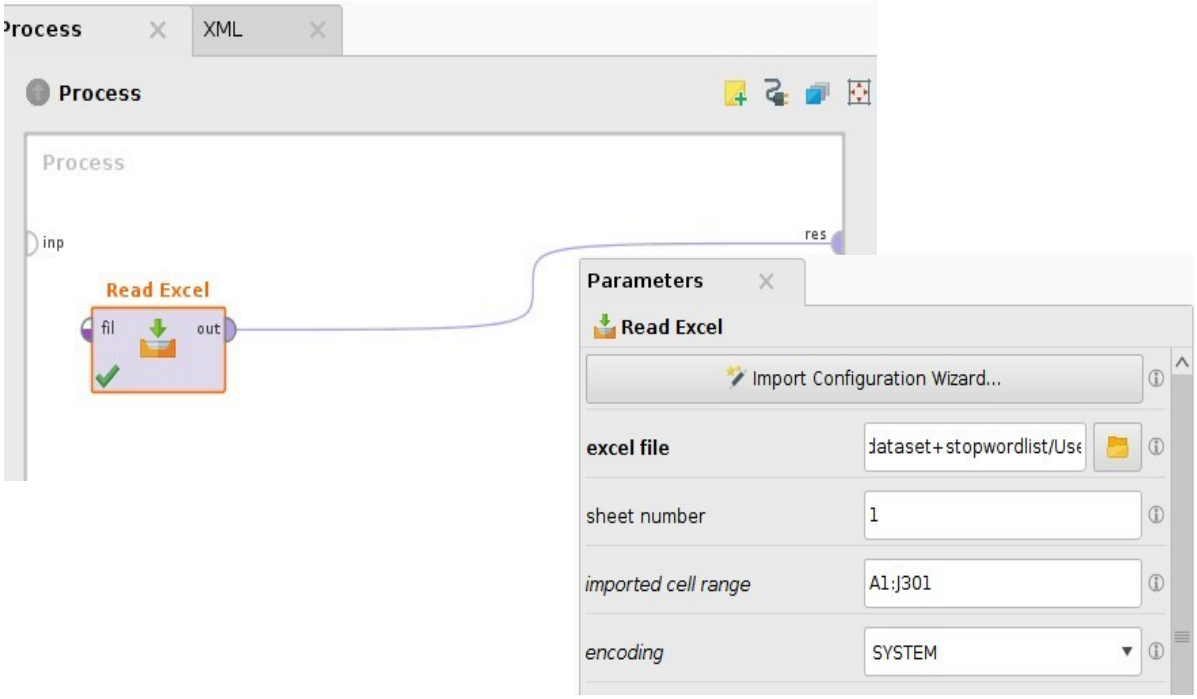
Dati testuali

Il dataset denominato Wikipedia (Wikipedia.zip) è disponibile sul sito del corso. Esso contiene una collezione di 12 articoli di Wikipedia, appartenenti a 3 differenti categorie. In particolare, i documenti appartengono ai seguenti argomenti: matematica, cibo, sport.

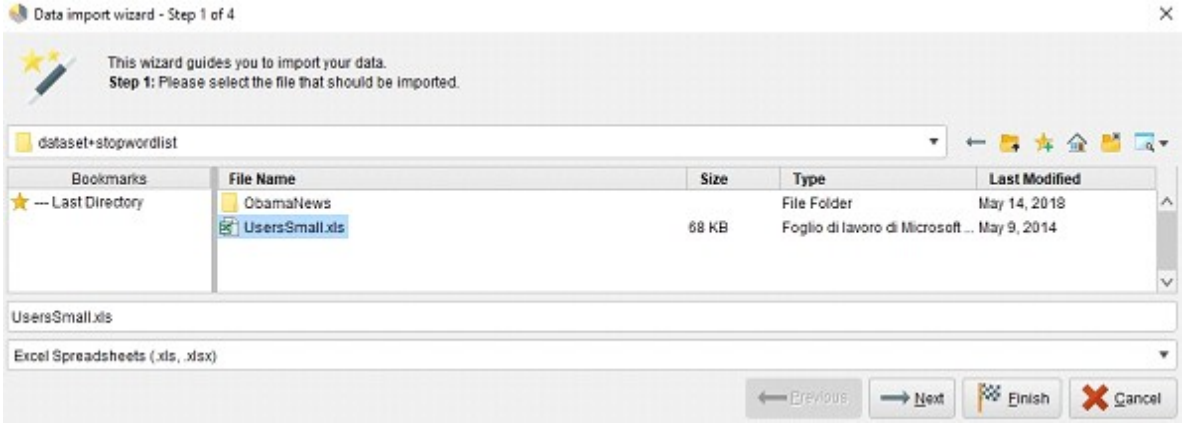
Preparazione dei dati strutturati

Obiettivo 1 – Import dei dati

- Nel pannello **Operators** cercare l'operatore **Read Excel** e trascinarlo nello spazio di lavoro.
- Importare il dataset ***UsersSmall.xls*** utilizzando la procedura guidata ***Import Configuration Wizard***.

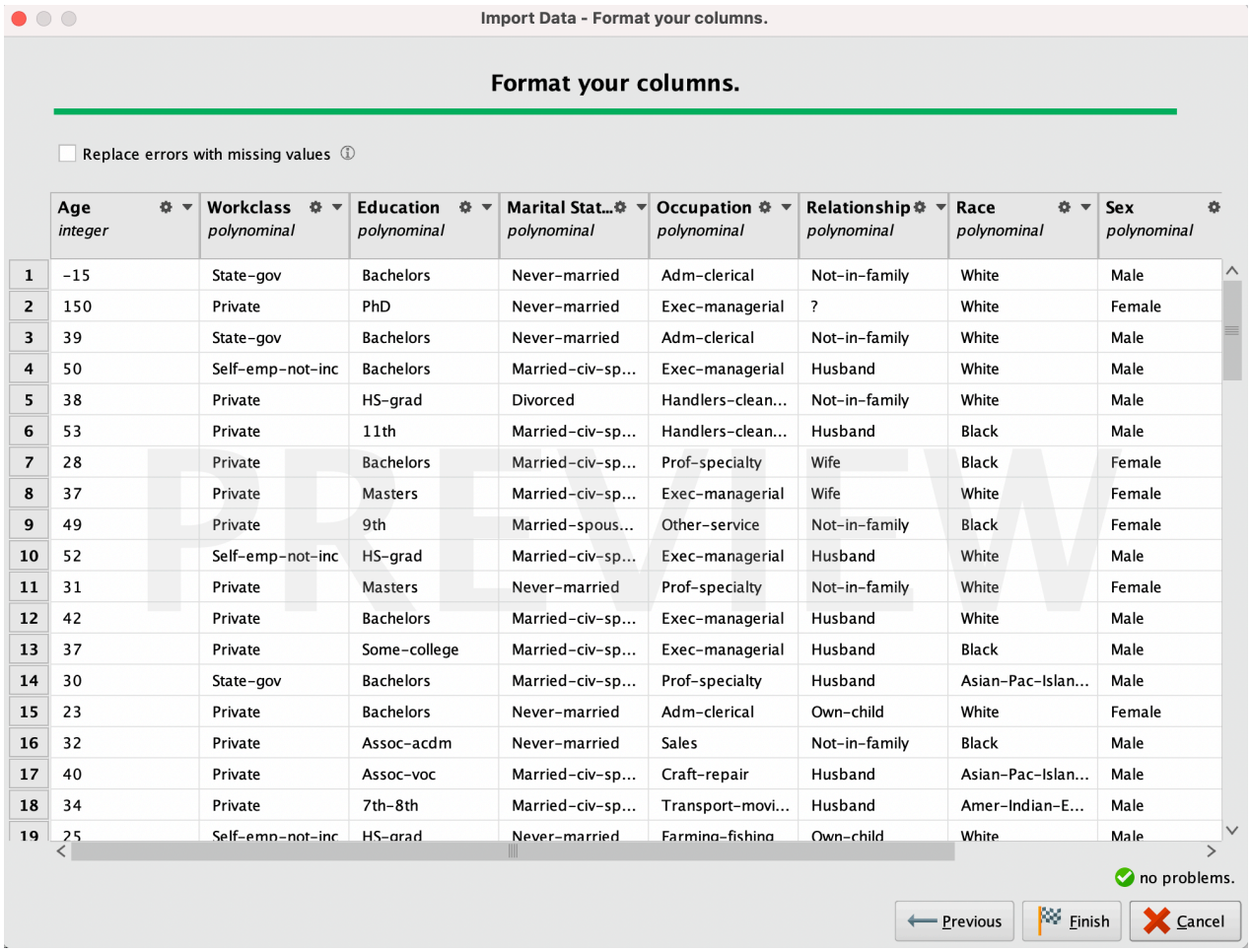


- Selezionare il file all'interno della cartella in cui si sono scompattati i file e selezionare *UsersSmall.xls*.



- Cliccare *Next*.
- Controllare che l'intera matrice di dati sia selezionata, in caso contrario selezionare **tutte le colonne**.
- Procedere con *Next*

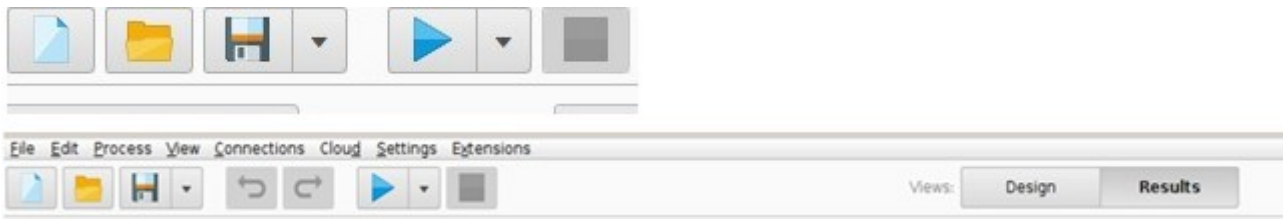
- Procedere con *Finish*. Si chiuderà il processo guidato.



	Age *integer*	Workclass *polynominal*	Education *polynominal*	Marital Stat... *polynominal*	Occupation *polynominal*	Relationship *polynominal*	Race *polynominal*	Sex *polynominal*
1	-15	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male
2	150	Private	PhD	Never-married	Exec-managerial	?	White	Female
3	39	State-gov	Bachelors	Never-married	Adm-clerical	Not-in-family	White	Male
4	50	Self-emp-not-inc	Bachelors	Married-civ-sp...	Exec-managerial	Husband	White	Male
5	38	Private	HS-grad	Divorced	Handlers-clean...	Not-in-family	White	Male
6	53	Private	11th	Married-civ-sp...	Handlers-clean...	Husband	Black	Male
7	28	Private	Bachelors	Married-civ-sp...	Prof-specialty	Wife	Black	Female
8	37	Private	Masters	Married-civ-sp...	Exec-managerial	Wife	White	Female
9	49	Private	9th	Married-spous...	Other-service	Not-in-family	Black	Female
10	52	Self-emp-not-inc	HS-grad	Married-civ-sp...	Exec-managerial	Husband	White	Male
11	31	Private	Masters	Never-married	Prof-specialty	Not-in-family	White	Female
12	42	Private	Bachelors	Married-civ-sp...	Exec-managerial	Husband	White	Male
13	37	Private	Some-college	Married-civ-sp...	Exec-managerial	Husband	Black	Male
14	30	State-gov	Bachelors	Married-civ-sp...	Prof-specialty	Husband	Asian-Pac-Islan...	Male
15	23	Private	Bachelors	Never-married	Adm-clerical	Own-child	White	Female
16	32	Private	Assoc-acdm	Never-married	Sales	Not-in-family	Black	Male
17	40	Private	Assoc-voc	Married-civ-sp...	Craft-repair	Husband	Asian-Pac-Islan...	Male
18	34	Private	7th-8th	Married-civ-sp...	Transport-movi...	Husband	Amer-Indian-E...	Male
19	25	Self-emp-not-inc	HS-grad	Never-married	Farming-fishing	Own-child	White	Male

- Collegare l'uscita dell'operatore Read Excel con res. Usare il tasto destro del mouse.

- Per lanciare un processo in RapidMiner usare il triangolo in alto nella barra dei processi.
- Per tornare nel processo principale, cliccare su **design**.



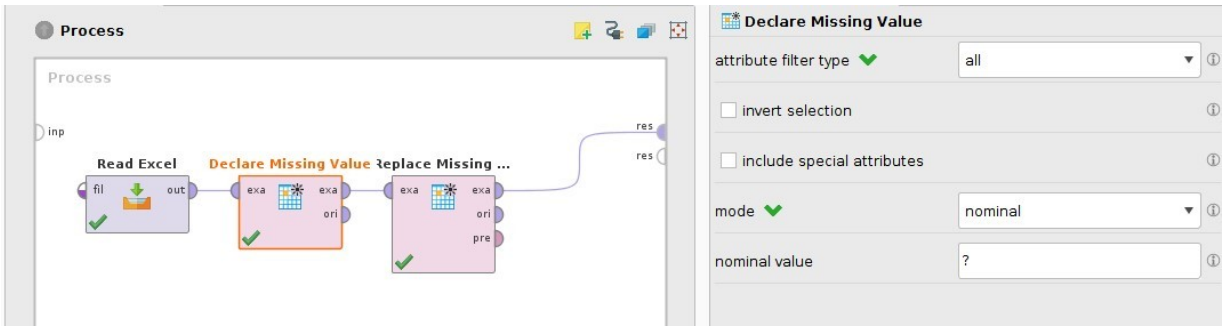
- Analizzare la semantica degli attributi e il loro ruolo a seconda degli obiettivi dell'analisi svolta.

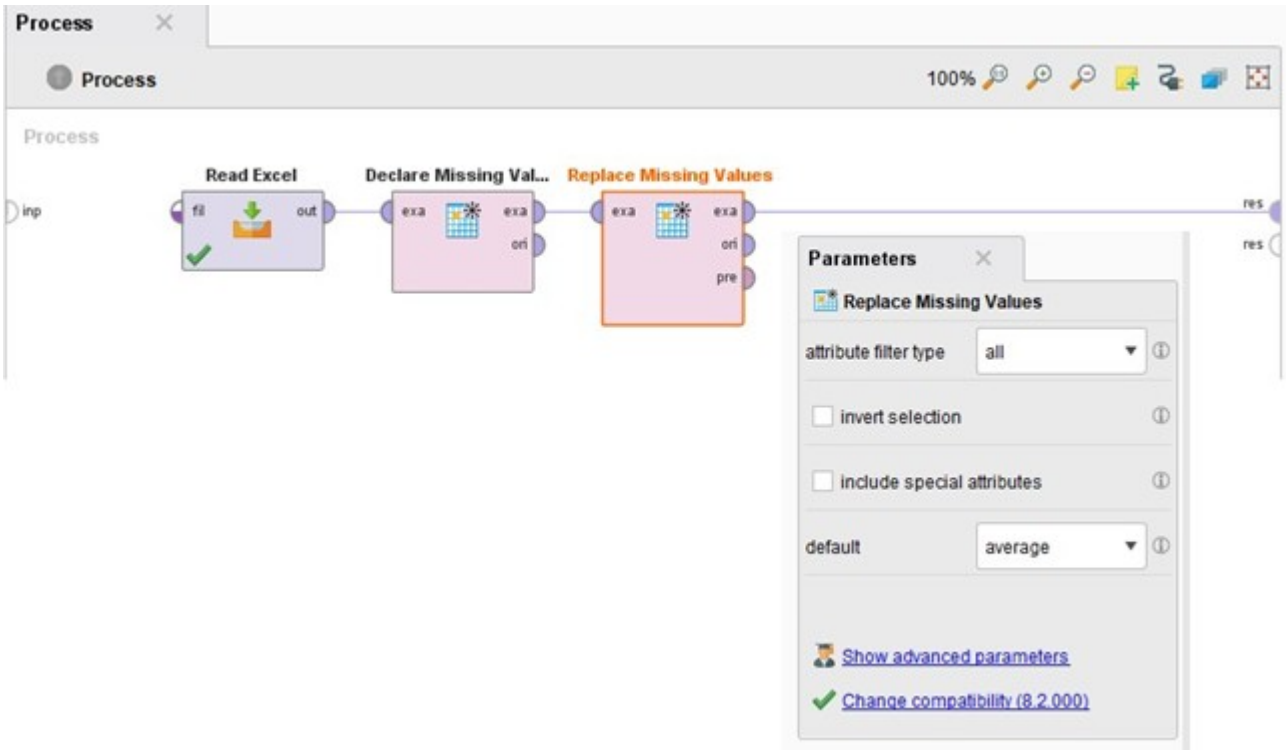
Obiettivo 2 – Gestione dei dati mancanti

Verificare la presenza di eventuali dati mancanti e gestirli con opportuni passi di trasformazione (operatori *Declare Missing values* e *Replace Missing Values*).

• Dichiarare per tutti gli attributi il '**?**' come valore NULL attraverso l'operatore **Declare Missing Value**.

• Sostituire i valori nulli dichiarati al punto precedente con il valore più frequente usando l'operatore **Replace Missing Values**.





Obiettivo 3 – Outlier detection

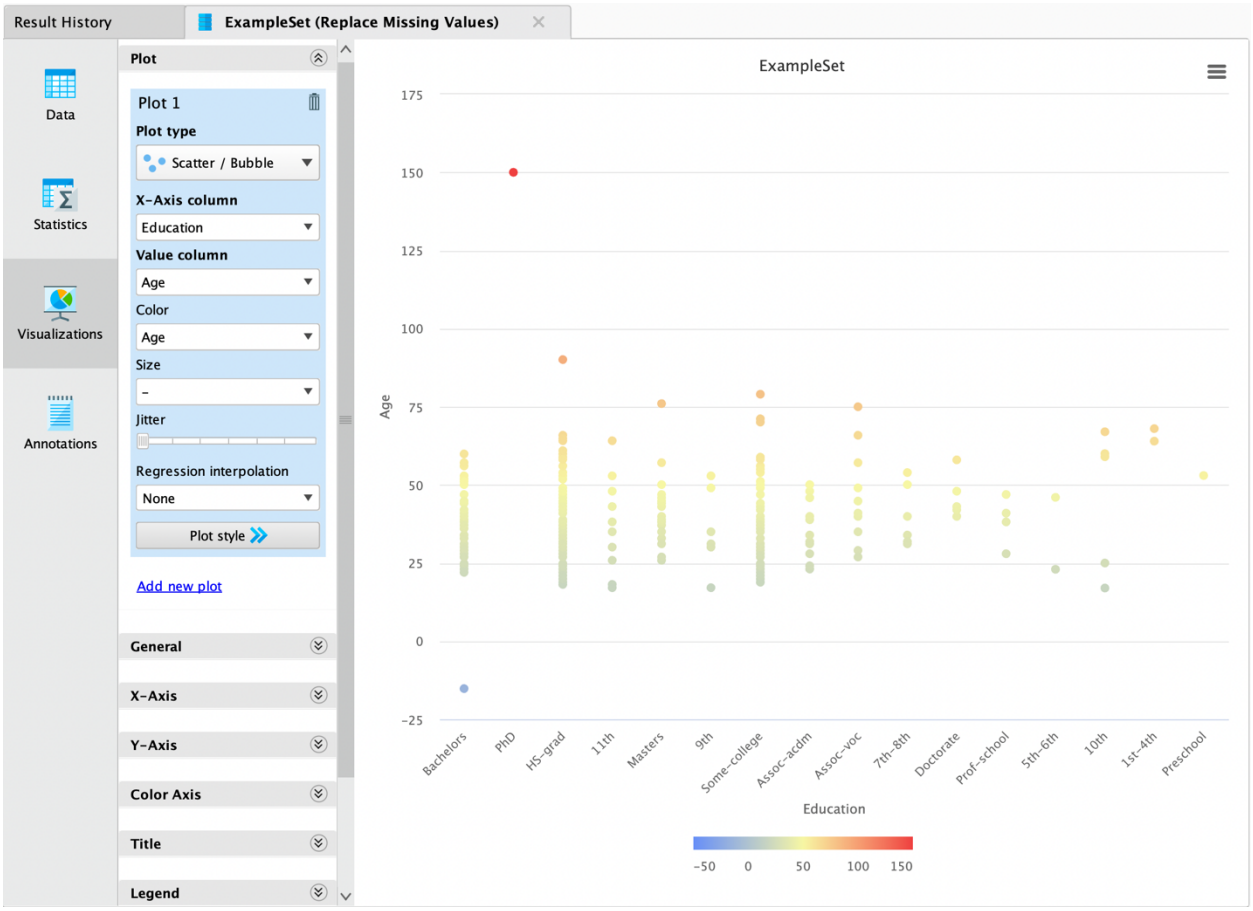
Verificare la presenza di outlier all'interno del dataset, utilizzando una strategia univariata.

- Nel tab R*isultati*, visualizzare le statistiche calcolate per il dataset da analizzare.
- Nel tab R*isultati*, plottare il grafico *Histogram* (10 bins) per l'attributo *Age*, selezionandolo tra i diversi Charts disponibili.

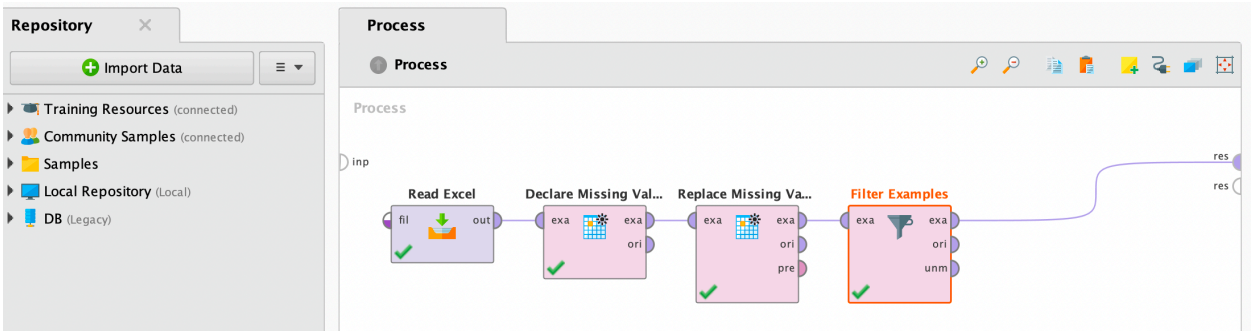
Sono presenti possibili outlier per il dataset in questione? Quali?

Verificare adesso in che modo gli outlier identificati per l'attributo *Age* sono distribuiti rispetto agli altri attributi.

- Nel tab Risultati, plottare il grafico Scatter/Bubble con l'attributo Age fisso sull'asse Y, e gli altri attributi a turno sull'asse X. Notate qualche situazione di interesse? Come sono distribuiti gli outlier in relazione all'etichetta del dataset (attributo *Response*)



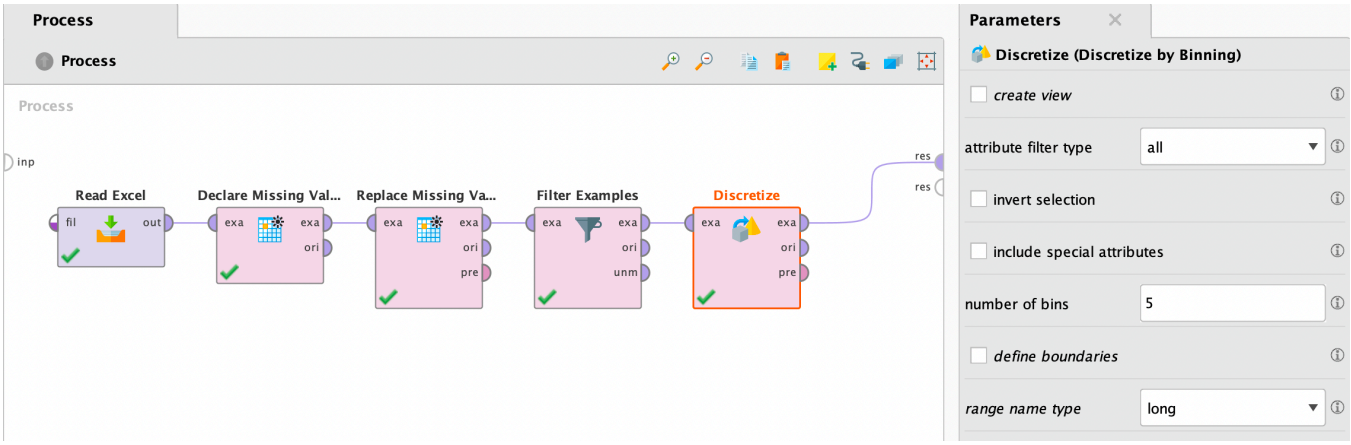
Rimuovere (per tutte le analisi successive) gli outlier identificati, utilizzando l'operatore *Filter Examples* e specificando gli opportuni filtri sull'attributo *Age.*
Provare adesso a plottare nuovamente il grafico Histogram relativo all'attributo age.



Obiettivo 4 – Discretizzazione

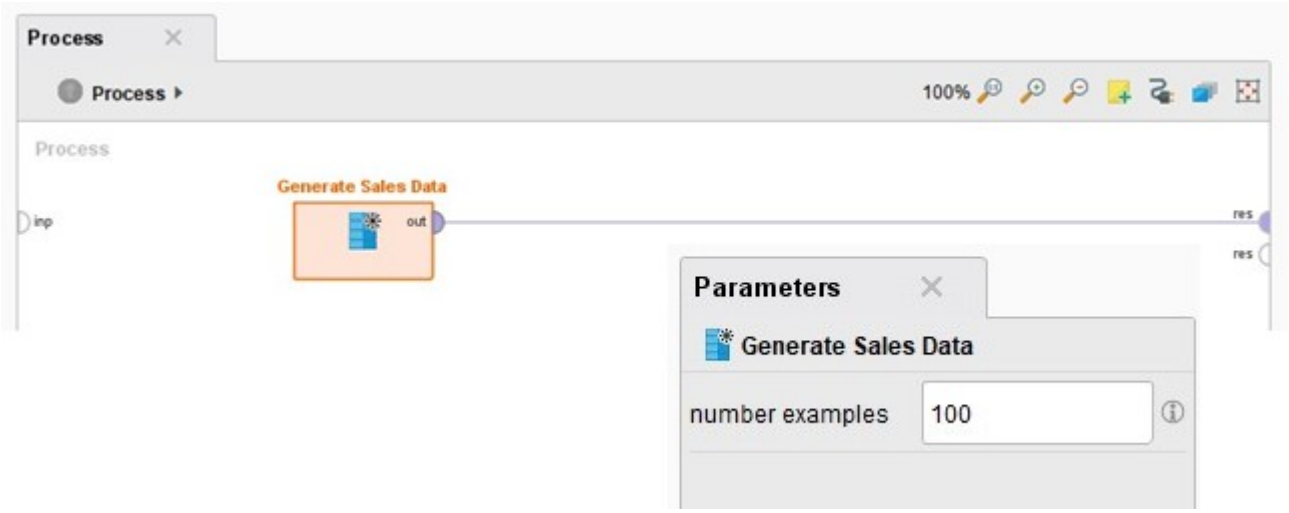
Verificare la presenza di attributi continui nei dati di origine.

- Discutere l'eventuale necessità di applicare un processo preliminare di discretizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati.
- Applicare diverse tecniche di discretizzazione (operatori *Discretize by binning*, *Discretize by frequency*, *Discretize by size*) e confrontare i risultati. Quale o quali attributi sono discretizzati? Qual è la differenza tra queste tre tecniche di discretizzazione?
- Come cambia il risultato utilizzando la tecnica *Discretize by binning* o *Discretize by frequency*, utilizzando in entrambi i casi il parametro *number of bins* uguale a 5?

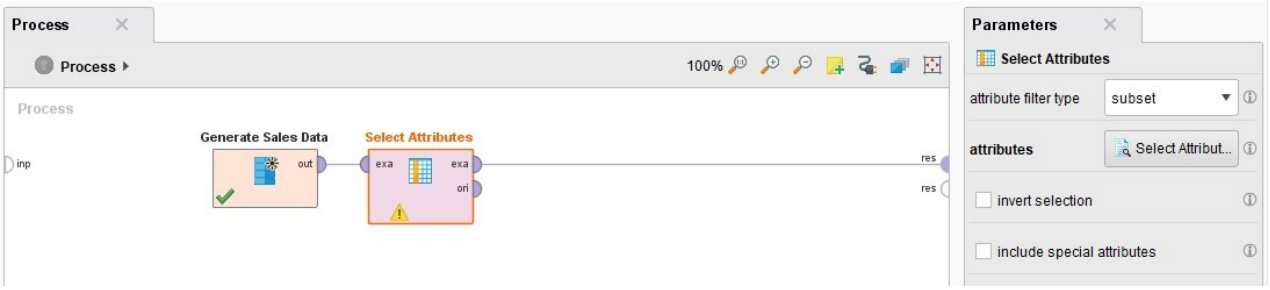


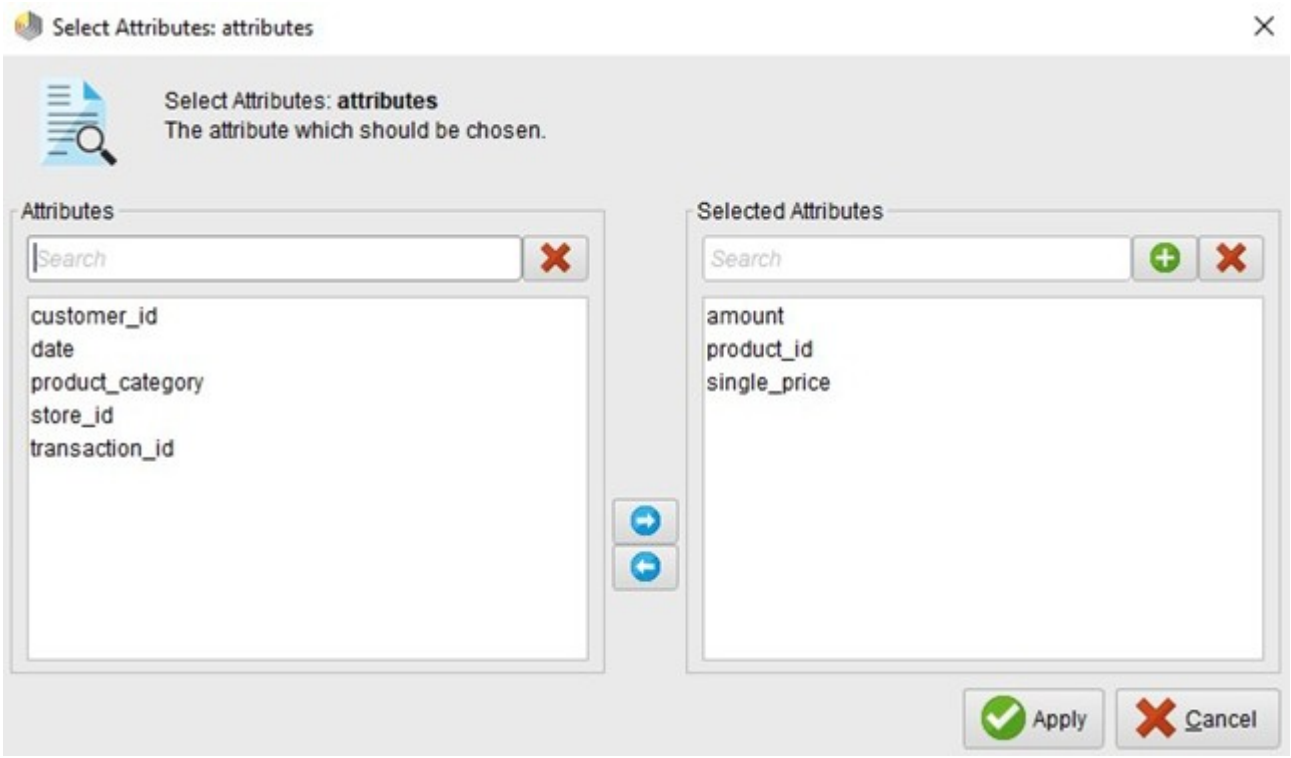
<u>Bonus: dati strutturati con attributi continui</u>

- Utilizzare l'operatore **Generate Sales Data**. Nella barra dei parametri, settare il parametro number examples a 100.

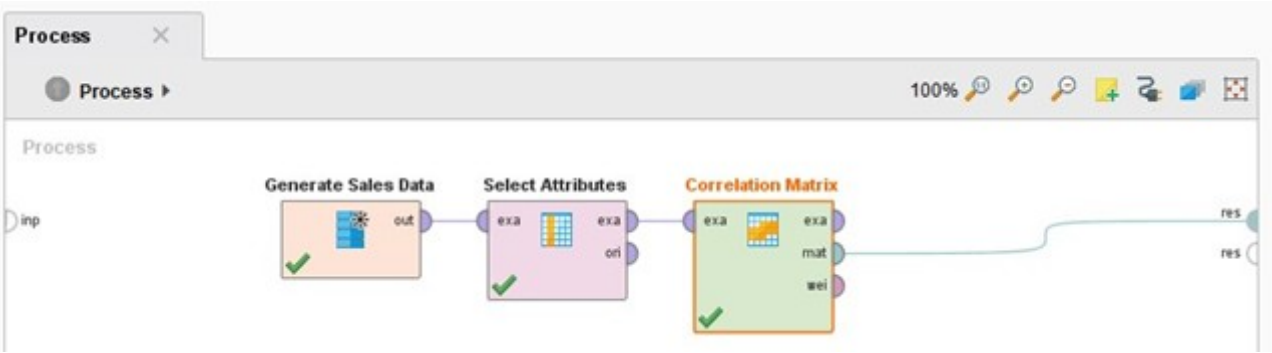


- Selezionare solo gli attributi numerici. Utilizzare l'operatore Select Attributes.



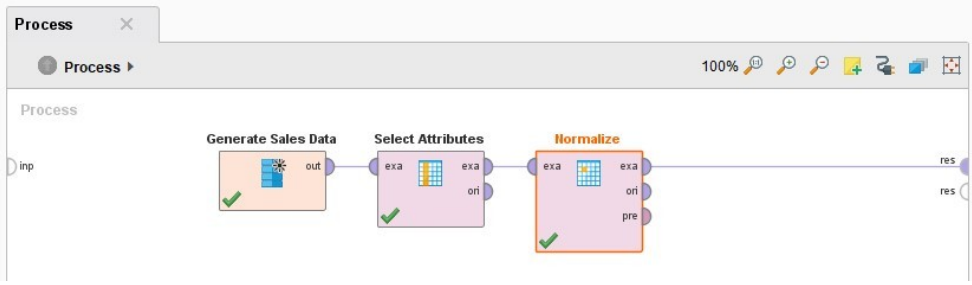


- Analizzare la correlazione tra coppie di attributi (operatore *Correlation Matrix*). Inserire l'operatore "Correlation Matrix" in coda al processo e visualizzare la rispettiva matrice collegando il plug-in del blocco denominato "mat" al plug-in "Result" sulla destra della finestra del processo principale. Il processo così generato sarà analogo al seguente:



- Esiste qualche correlazione elevata?
- Eliminare l'operatore correlation matrix (selezionarlo con il mouse e premere canc).

- Discutere l'eventuale necessità di applicare un processo preliminare di normalizzazione in funzione degli obiettivi dell'analisi e degli algoritmi di data mining utilizzati (operatore *Normalize*).

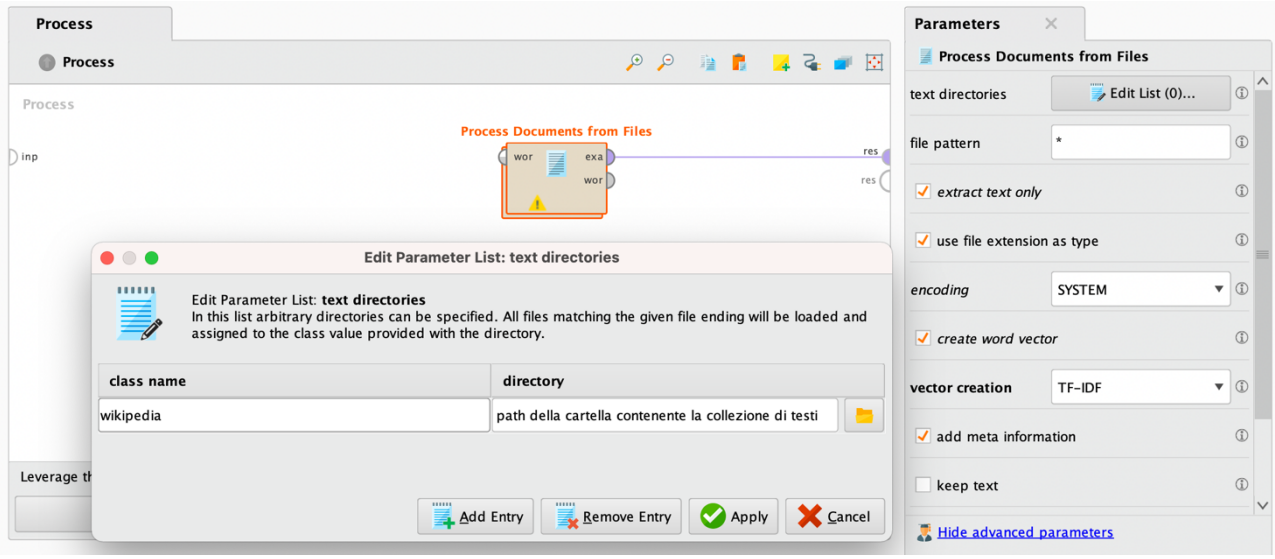


- Quali differenze ci sono tra le varie tecniche di normalizzazione disponibili?
- Quando è utile normalizzare i dati?

Preparazione dei dati testuali

Obiettivo 1 – Import dei dati

- Importare il dataset Wikipedia in Rapid Miner (operatore *Process Documents From Files*).



- Se volete avere l'informazione del testo all'interno dei risultati, spuntate la voce **Keep Text** nel pannello dei parametri dell'operatore **Process Documents from Files**.

Obiettivo 2 – Generazione dei token, stopwords e stemming

Il **TF-IDF** (*Term Frequency–Inverse Document Frequency*) è una funzione nota nel text mining utilizzata per misurare l'importanza di un termine rispetto ad una collezione di documenti. Il TF-IDF aumenta **proporzionalmente** al numero di volte che il termine è contenuto nel documento, ma cresce in maniera **inversamente proporzionale** con la frequenza del termine all'interno della collezione. In questo modo si possono penalizzare le parole molto frequenti che non danno rilevanza alla collezione e dare più importanza ai termini che in generale sono poco frequenti ma più rilevanti per l'analisi.

$$\mathrm{tfidf}_{i,j} = \mathrm{tf}_{i,j} \times \log\left(\frac{\mathrm{N}}{\mathrm{df}_i}\right)$$

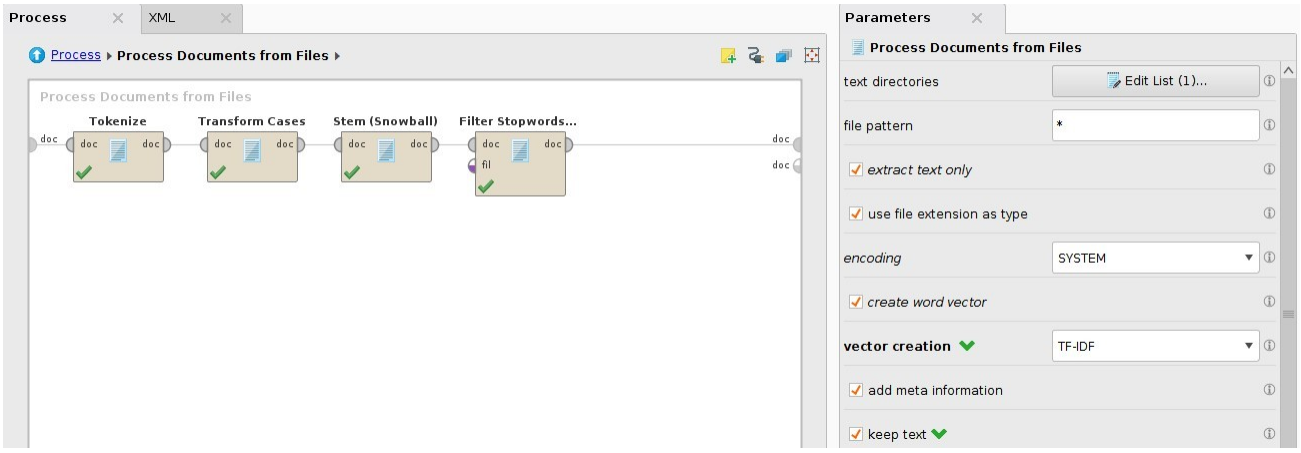
$\mathrm{tf}_{i,j}$ = total number of occurences of i in j
df_i = total number of documents (speeches) containing i
N = total number of documents (speeches)

L'operatore *Process Document from Files* ammette un sottoprocesso per poter pulire il dataset e trasformarlo in una tabella chiamata matrice documenti*termini. La tabella avrà una riga per ogni documento della collezione presente nella cartella letta e una colonna per ogni termine presente all'interno della collezione.

- Prima di creare la matrice, guardare l'output. Nel sottoprocesso (per entrare nel sottoprocesso, fare doppio click con il tasto sinistro sull'operatore *Process Document from Files*) collegare le uscite come in figura.

- Per tornare al processo principale, cliccare la freccia blu di fianco a *Process*.
- Collegare l'uscita exa con res ed eseguire il processo.

- Applicare i passi di pre-processing sul dataset testuale. Doppio click sull'operatore **Process Documents from Files**. Verrà aperto un sottoprocesso. Utilizzare i seguenti blocchi:
  - Il blocchetto **Tokenize**: splitta ogni documento della collezione Wikipedia in un vettore diparole. L'ordine delle parole non sarà più rispettato. Secondo te ha importanza ai fini dell'analisi? (Settare il parametro non letters).
  - Il blocchetto **Transform Cases**: Trasforma il testo in maiuscolo o minuscolo.
  - Il blocchetto **Stem (Snowball)**: Riduce le parole alla propria radice. La radice è quell'elemento linguistico irriducibile (non ulteriormente suddivisibile) che esprime il significato principale della parola. (Utilizzare la lingua italiana).
  - Il blocchetto **Filter Stopwords (Dictionary)**: Permette di eliminare le parole definite Stopword, parole che non hanno un particolare significato se isolate dal testo e quindi vengono ignorate dai programmi. Sono parole poco significative perché possono essere usate spesso all' interno delle frasi. Ad esempio articoli, congiunzioni e preposizioni non caratterizzano il significato di un testo, possono essere eliminate a monte di una analisi text mining. Carica il file **stopwordsEnglish.txt** presente sul sito del corso.



- Utilizzare la codifica UTF-8 per il file delle stopword.