MSc Data Science and Advanced Computing
CSMDM21 - Data Analytics and Mining

Practical 2: Data Transformation in KNIME

Exercise 1: Data Transformation and Visualization

Workflow 1: Category to Number

1. Create a KNIME workflow to read **iris.csv** (download from Blackboard).
2. Transform the names of flower species into numbers (Iris-setosa: 0; Iris-versicolor: 1; Iris-virginica: 2). This can be done in two ways:
    a. Using the *Cell Replacer* node. Read **iris_file_class_names.csv** (download from Blackboard): it has two columns, iris species and a corresponding numerical ID. Replace the species label with a numerical value in the data table.
    b. Using the *Category to Number* node. Select this node and read its description to understand how to use it.
3. Find a node to convert back the numerical class values (0, 1, 2) into the corresponding species labels (Iris-setosa, Iris-versicolor, Iris-virginica).

Workflow 2: Principal Component Analysis (PCA) for Visual Exploration

1. Create a KNIME workflow to read the iris dataset.
2. Use *Color Manager* to associate a color to each class (iris species).
3. Apply *PCA* to the four numerical attributes to generate the top two principal components (PCA0 and PCA1): use these two PCA attributes to generate a scatter plot.
4. Select some data points in the scatter plot and apply *HiLite selected*.
5. Use an *Interactive Table* to view the data table: apply *Filter – Show Hilited Only*.

Exercise 2: Data Aggregation

Workflow 1: Data Aggregation

1. Create a KNIME workflow to read **customer_file_1.csv** (download from Blackboard).
2. Try generating the following tables by using the node *GroupBy* or the node *Pivoting*.

Table 1: Products count

Products	CustomerKey (Count*)
CO Investment	961
Fund Manager+	2036
Gold Investment	2002
P+B Investment	3161
Private Investment	3391

Table 2: Products count (sorted in ascending/descending order)

Products	CustomerKey (Count*)
CO Investment	961
Gold Investment	2002
Fund Manager+	2036
P+B Investment	3161
Private Investment	3391

3. Pivot the customer data (one with duplicate removed) across Gender column to find the average age. It should produce three tables:
   a. Table 3: Average age of male and female per product.

Products	F+Mean(Age)	M+Mean(Age)
CO Investment	48.071	47.287
Fund Manager+	48.71	49.603
Gold Investment	47.949	47.826
P+B Investment	48.065	47.454
Private Investment	48.632	48.712

   b. Table 4: Average age per product irrespective of male and female.

Products	Mean(Age)
CO Investment	47.667
Fund Manager+	49.149
Gold Investment	47.887
P+B Investment	47.764
Private Investment	48.673

   c. Table 5: Average age of male and female across all product.

Products	F+Mean(Age)	M+Mean(Age)
Across all product	48.326	48.251

Plot the three tables using *Bar Chart*.

Solutions:
Sample solutions to these exercises are available on Blackboard (Bb) in two forms: images of the workflows and the actual KNIME workflows (as a single zip archive). You should first try to build your own workflows for each exercise. During the practical session, you may use the images to see the proposed solutions and to reconstruct them. During or at the end of the practical session, the archives with the actual KNIME workflows will be made available for you to import and test them. These are KNIME archives (file extension ". knwf ") that can be imported in KNIME. Note that after importing the KNIME archives, you may need to change the file locations of the source data file and the output file destination folder.

More Practice with the KNIME Example Server

The KNIME examples server is useful to learn about various nodes and to investigate typical tasks: **EXAMPLES (knime@hub.knime.com)** is available in the KNIME explorer panel.

Login into the KNIME examples server (no username/password is required), open and test the following workflows.

- 02_ETL_Data_Manipulation | 00_Basic_Examples:
  - 00_Visual_Analysis_of_Sales_Data
    Filter rows and columns of sales data and generate some visualisation. Used nodes include *row/column filter, Colour Manager, Stacked Area Chart, Pie/Donut Chart*
  - 01_Example_for_Standard_Preprocessing
    Perform row and column filtering, concatenation, and binning. Used nodes include *splitter, row/column filter, reference row/column filter, concatenate*.
  - 02_ETL_Basics
    A more complex example for ETL basics operations on sales data. This workflow performs an aggregation task: it process a series of contracts with different customers in different countries to generate a one-row summary description for each one of the customers.
- 04_Analytics | 02_Statistics | 01_Simple_Example_with_Statistics
  This workflow splits the columns of the input file into numeric and nominal ones, performs some filtering and then joins the data back again. A data view (box plot) and statistics are generated. Used nodes include *splitter, row filter, joiner, box plot*.