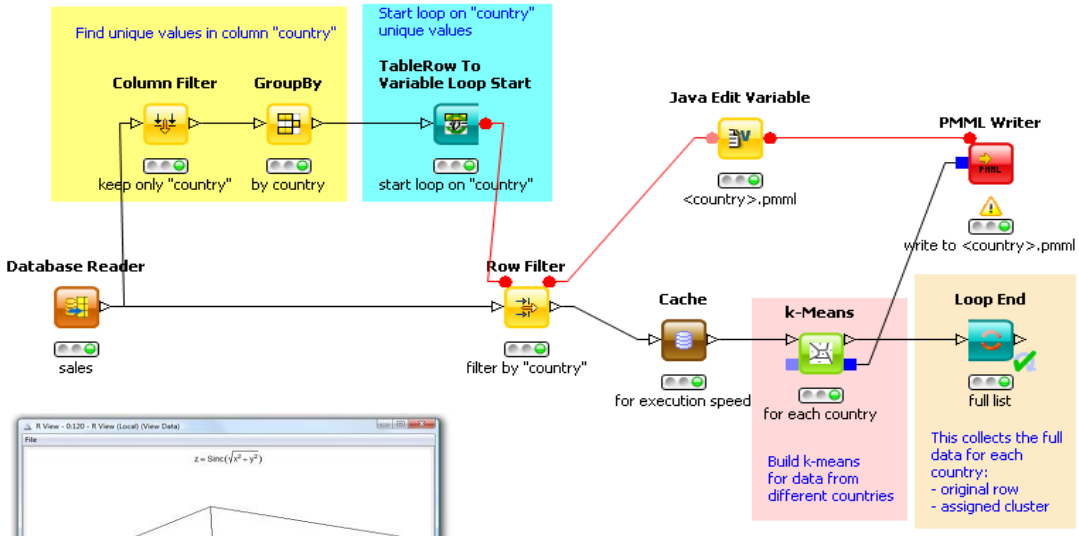
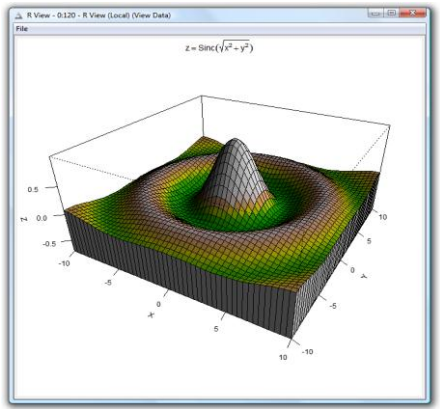
Rosaria Silipo, Michael P. Mazanetz

The KNIME Cookbook

Recipes for the Advanced User

KNIME press





Exercise 3

product	Sum(quantity)	Sum(amount)	
prod_1	74	2590	
prod_2	130	5200	
prod_3	106	8480	
prod_4	1	3	

This book has been updated for KNIME 2.5.

For information regarding permissions and sales, write to:

KNIME Press
Technoparkstr. 1
8005 Zurich
Switzerland

knimepress@knime.com

ISBN: 978-3-9523926-0-7

Table of Contents

Acknowledgements

We would like to thank a number of people for their help and encouragement in writing this book.

In particular, we would like to thank Thomas Gabriel for the advice on how to discover the many possibilities of integrating R into KNIME, Bernd Wiswedel for answering our endless questions about calling external web services from inside a workflow, and Iris Adae for explaining the most advanced features of some of the Time Series nodes.

Special thanks go to Peter Ohl for reviewing the book contents and making sure that they comply with KNIME intended usage and to Heather Fyson for reviewing the book's English written style.

Finally, we would like to thank the whole KNIME Team, and especially Dominik Morent, for their support in publishing and advertising this book.

Chapter 1. Introduction

1.1. Purpose and Structure of this Book

KNIME is a powerful tool for data analysis and data visualization. It provides a complete environment for data analysis which is fairly simple and intuitive to use. This, coupled with the fact that KNIME is open source, has led many thousands of professionals to use KNIME. In addition, other software vendors develop KNIME extensions in order to integrate their tools into KNIME. KNIME nodes are now available that reach beyond customer relationship management and business intelligence, extending into the field of finance, the life sciences, biotechnology, pharmaceutical and chemical industries. Thus, the archetypal KNIME user is no longer necessarily a data mining expert, although his/her goal is still the same: to understand data and to extract useful information.

This book was written with the intention of building upon the reader's first experience with KNIME. It expands on the topics that were covered in the first KNIME user guide ("The KNIME Beginner´s Luck" [1]) and introduces more advanced functionality. In the first KNIME user guide [1], we described the basic principles of KNIME and showed how to use it. We demonstrated how to build a basic workflow to model, visualize, and manipulate data, and how to build reports. Here, we complete these descriptions by introducing the reader to more advanced concepts. A summary of the chapters provides you with a short overview of the actual concepts we discuss.

Chapter 2 describes the nodes needed to read and write data from and into a database. Reading and writing into a database are the basic operations necessary for any, even very simple data warehousing strategies.

Chapter 3 introduces the DateTime object and the nodes to turn a String column into a DateTime column, to format it, to extract a time difference and so on. The DateTime object provides the basis for working with time series.

A very important concept for the KNIME workflow is the concept of "workflow variables". Workflow variables enable external parameters to be introduced into a workflow. Chapter 4 describes what a workflow variable is, how to create it, and how to edit it inside the workflow if needed.

KNIME is a new software tool, very easy to use and already empowered with a lot of useful functionalities. One of the strongest points of KNIME, however, is the openness of the platform to other data analysis software. R, for example, is one of the richest and oldest open source data analysis software available and is equipped with a wealth of graphical libraries and pre-programmed statistical functions. Where KNIME might come short then, the many pre-implemented functions of the R libraries can be conveniently taken advantage of. A few KNIME nodes, indeed, allow for the integration of R scripts into the KNIME workflow execution. So, if you are or have been an R expert for long time and have a number of R scripts your

fingertips and ready to use, you can easily recycle them in your new KNIME workflows. Chapter 5 illustrates how to include and run R scripts in KNIME workflows. We describe how to install the R package, how to connect KNIME and R, and which nodes and R code can be combined together to produce plots, data models, or just simple data manipulations.

As of KNIME 2.4, it is also possible to call external web services and to collect the results inside a KNIME workflow. Chapter 6 describes the node that can connect to and run external web services.

Most data operations in KNIME are executed on a data matrix. This means that an operation is executed on all data rows, one after the other. This is a big advantage in terms of speed and programming compactness. However, from time to time, a workflow also needs to run its data through a real loop. Chapter 7 introduces a few nodes that implement loops: from a simple "for" or "while" cycle to a more complex loop on a list of values.

Chapter 8 illustrates the use of logical switches to change the workflow path upon compliance with some predefined condition.

Chapter 9 is an extension of chapter 6 in "The KNIME Beginner´s Luck" [1]: it describes a number of advanced features of the KNIME reporting tool. First of all, it explains how to introduce parameters into a report and how workflow variables and report parameters are connected. Later on in the chapter, a few more functions are discussed which can be used to create a more dynamic report.

Chapter 10 concludes this book and gives a few suggestions on how to manage memory in KNIME and how to run workflows in batch mode. These tips are not necessary in order to run workflows, but they can come in handy if a workflow has to be run multiple times or if the size of the workflow is particularly large.

In this introductory chapter, we list the data and the example workflows that have been built for this book and note the additional software required to run some of these example workflows.

1.2. Data and Workflows

In the course of this book we put together a few workflows to show how KNIME works. In each chapter we build one or more workflows and we expect the reader to build a few more in the exercises. The data and workflows used and implemented in this book are available in the "Download Zone". You should receive a link to the "Download Zone" together with this book. In the "Download Zone" you will find a folder for each chapter, containing the chapter's example workflows, and a subfolder called "Exercises", containing the solutions to the exercises in the corresponding chapter. You will also find a folder called "Data", which contains the data used to run the workflows.

Structure of the "Download Zone"

Chapter 2

- Database_Operations.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip
  - Exercise3.zip

Chapter 3

- DateTime_Manipulation.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip

Chapter 4

- Workflow_Vars.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip
  - Exercise3.zip

Chapter 5

- R Snippet Example.zip
- Plotting Example 1.zip
- Plotting Example 2.zip
- Plotting Example 3.zip
- R Model Example.zip
- R PMML Node.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip
  - Exercise3.zip

Chapter 6

- WebServiceNodes.zip
- Exercises
  - Exercise1.zip

Chapter 10

- Batch Mode.zip
- Batch Mode Databases.zip
- Simple Batch Example.zip
- Exercises
  - Create Database.zip
  - Exercise1.zip

Chapter 7

- Chunk Loop.zip /Chunk Loop 2 Ports.zip
- Counting Loop 1.zip /Counting Loop 2.zip
- Loop on List of Columns.zip
- Loop on List of Values.zip
- Loop with final Condition.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip
  - Exercise3.zip
  - Exercise4.zip

Chapter 8

- Automated IF Switch.zip
- CASE Switch.zip
- Empty table Replacer.zip
- Java IF & Tables.zip
- Manual IF Switch.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip

Chapter 9

- New Projects.zip
- Traffic lights.zip
- Exercises
  - Exercise1.zip
  - Exercise2.zip
  - Exercise3.zip

Data

- sales.csv, wrong_sales_file.txt
- sales/sales_*.csv
- cars-85.csv
- dates.txt
- slump_test.csv
- Projects.txt
- Totals Projects.csv
- images/*.png
- ZIP_CODES.zip
- ChEBI_original.csv

This book is not meant as an exhaustive reference for KNIME, although many useful workflows and aspects of KNIME are demonstrated through worked examples. This text is intended to give you the confidence to use the advanced functions in KNIME to manage and mine your own data.

The data files used for the exercises and the example workflows were either generated by the authors or downloaded from the UCI Machine Learning Repository [2], a public data repository (http://archive.ics.uci.edu/ml/datasets). For the data sets belonging to the UCI Repository, the full link is provided below.

Data sets from the UCI Machine Learning Repository [2]:

- Automobile: http://archive.ics.uci.edu/ml/datasets/Automobile
- Slump_test: http://archive.ics.uci.edu/ml/datasets/Concrete+Slump+Test

1.3. Additional Software used in this Book

Two additional applications are required to run some of the examples used throughout this book: a database (we used PostGreSQL) and the R Project software. These tools extend the capabilities of KNIME and are freely available for download.

1.3.1. Database

In Chapter 2 of this book, we illustrate the KNIME approach to database operations. As the chapter progresses, we build an entire workflow to show the usage of database nodes and their potential in a practical example.

For the example to work, though, you have to install one of the many available database software tools. We chose to install the PostGreSQL database software. This choice was not motivated by KNIME considerations, since the KNIME database nodes work equally well for all other database tools, such as MySQL, MS SQL, Oracle, just to cite three among the most commonly used databases. This choice was largely motivated by the fact that PostGreSQL is open source and can be installed by the reader at no additional cost.

More information about PostGreSQL can be found at the database web site http://www.postgresql.org/ . The database software can be downloaded from http://www.postgresql.org/download/. From the binary packages, download the one-click installer for your machine and follow the installation instructions.

After installation of the database tool, start the administration console. If you have installed PostGreSQL, the administration console is called “pgAdminIII” (for Windows users, “All Programs” -> “PostgreSQL x.y” -> “pgAdminIII”).

From the administration console create a:

- User with username “knime_user” and password “knime_user”
- Database named “Book2” accessible by user “knime_user”
- Table named “sales” by importing the “sales.csv” file available in the Download Zone of the book

Note that the “date” field in the newly created “sales” table should have type Date.

If you do not know your way around a database, you can always use the “Database Writer” node to import the “sales.csv” file into the database. We have encountered the “Database Writer” node already in the “KNIME Beginner’s Luck” book [1]. However, for the readers who are new to this node, it is described in chapter 2.

You also need the JDBC database driver for your database to communicate with the KNIME database nodes. If you are using PostGreSQL, its database driver is available for download from http://jdbc.postgresql.org/download.html.

1.3.2. R-Project

R is a free software environment for data manipulation, statistical computing, and graphical visualization.

In Chapter 5 we explore the R KNIME extensions. These extensions make it possible to open R views, build models within R and run snippets of R code. The KNIME R plug-in contains all the nodes to run R from inside KNIME, but requires the R-project (R binaries) to be already installed locally on the machine or the R Server component “Rserve” and be accessible by KNIME.

For Windows, the R binaries can be installed via the KNIME update site. However, if you want a custom installation of R, in a pre-defined location, you need to install R directly from the R-project web site. On Linux and Mac you always need to install R beforehand. That is, in order to run the KNIME nodes for R, you need to download and install the R-package locally OR download, install, and run the R server component “Rserve”.

Install R

A link to the latest versions of R and Rserve, which are compatible with KNIME, can be found on the KNIME Extensions web site http://www.knime.org/downloads/extensions).

The R project software for Windows can also be installed directly in KNIME from the KNIME Update site. From KNIME, go to the top menu:

- Select “Help” –> “Install New Software”

- In the “Available Software” window, select the KNIME update site

Alternatively

- Select “File” -> “Install KNIME Extensions …”

Next:

- Open the “KNIME & Extensions” group
- Select “KNIME R Integration (Windows Binaries)”
- Click “Next” and follow the installation instructions

Installing R directly from the KNIME update site is the simplest way to obtain R. This method is sufficient for most applications.

Alternatively, you can download the R and/or the RServe packages directly from the “Comprehensive R Archive Network (CRAN)” homepage at http://cran.r-project.org.

Note. The “KNIME R Statistics Integration (Windows Binaries)” package is only applicable on Windows. If you need R to run on another operating system, you need to download the appropriate software package from the CRAN homepage.

Note. There are two KNIME Extensions for R: “KNIME R Statistics Integration” installs the KNIME plugin with the R nodes; “KNIME R Statistics Integration (Windows Binaries)” installs R on your machine.

The R software download (“KNIME R Statistics Integration (Windows Binaries)”) from the KNIME Extensions site already includes around 50 R packages that are ready for use in KNIME: a variety of plotting and statistical R functions and some R data sets. If you download R from the CRAN, you also need to install the following packages separately: “QSARdata”, “corrgram”, “lattice”, “car”, “alr3”, “ggplot2”, “circular”, “reshape”, and “drc”.

Install the Packages Required for KNIME R Templates

If you want to download a package from CRAN, the general procedure involves navigating to the “Packages” page by clicking “Packages” in the “Software” section. This page contains details on how to install new packages and further information and documentation on the available R packages.

However, downloading and installing a package is best done using the install.packages function in the R editor.

Open your R application and execute the following command (you might be asked to select a mirror for downloading the packages):

```
install.packages(c('QSARdata'))
install.packages(c('corrgram', 'lattice', 'car', 'alr3', 'ggplot2', 'circular', 'reshape', 'drc'))
```

Once the command has been run, close the R editor.

Note. Remember: You do not need to install these packages if you have installed the R-Project from the KNIME Extensions update site ("KNIME R Statistics Integration (Windows Binaries)").

Setup an R-Server (if needed)

Download and install the latest version of Rserve from RForge, http://www.rforge.net/Rserve.

Follow the instructions in the documentation for your operating system. Once Rserve has been downloaded and the package installed, the R-Server can be started (see an example in http://www.rforge.net/Rserve/example.html).

1.3.3. External Software Summary

The table below is a summary of the installations and configurations of external software tools used in this book.

Software type	Name	Download from:
Database	PostGreSQL (or other database software) with: ○ user: "knime_user", "knime_user" ○ Database: "Book2" ○ Table: "sales" imported from file "sales.csv" from the Download Zone	PostGreSQL: http://www.postgresql.org/download/ MySQL: http://dev.mysql.com/downloads/
R	R (local) / Rserve (server)	R-project: http://cran.r-project.org/ RForge: http://www.rforge.net/Rserve/ KNIME Extensions: http://www.knime.org/downloads/extensions