KONFERENCA
PORTOROŽ, 15. DO 17. MAJ 2017

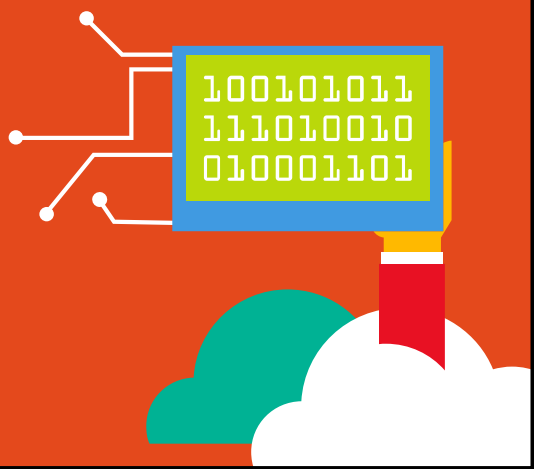
Microsoft

KONFERENCA

Microsoft

Introducing R

Dejan Sarka



TEHNOLOGIJA

Introduction

- Dejan Sarka
  - dsarka@solidq.com, dsarka@siol.net, @DejanSarka
  - MCT, Data Platform MVP
  - 30 years of data modeling, data mining and data quality
- 14 books, writing more
- 15+ courses

Microsoft

Reach us with #ntk17

KONFERENCA

Agenda

- Introducing R
- R in SQL Server Database Engine
- R in MS BI Suite
  - SSRS
  - SSIS
  - Power BI Desktop
  - Azure ML

Microsoft

Reach us with #ntk17

KONFERENCA

About R

- The R statistical programming language is a free open source package based on the S language developed by Bell Labs
- R written as a research project by Ross Ihaka and Robert Gentleman
  - Now developed by a group of statisticians called 'the R core team', with a home page at www.r-project.org
- R is available free of charge and is distributed under the terms of the Free Software Foundation's GNU General Public License
  - Available for Windows, Mac OS X, and Linux

About R

- R can run interactively
- R is a programming language for analyzing data
  - Many statistical functions are already built in
  - Excellent graphic functionality
  - Contributed packages expand the functionality to cutting edge research
- Some drawbacks as well
  - Generating code to complete tasks is required
  - Used to be the sole province of academic statisticians
  - It's open – different procedures for the same task
  - Free version not scalable

Getting R

- Install R from r-project.org
  - R packages extend the language - you need to be able to download zip files
  - Regular updates
- R defaults to an interactive mode
- R Console – command prompt or GUI?
  - A prompt ">" is presented to users
  - Each input expression is evaluated and a result returned

R Tools

- RStudio IDE is a powerful user interface for R
  - It's free and open source, and works on Windows, Mac, and Linux
  - Install it from Rstudio.com
  - Regular updates
- R Tools for Visual Studio
  - Open source plug-in
  - Syntax-aware editing, a command-line REPL, and interactive debugging

R Language Basics (1)

- R is a *functional* language
- You don't type commands but rather call *functions* to achieve results, even quit
  > q()
- Other common functions
  > help(<topic>) or ?<topic>
  > license()
  > contributors() and citation()
  > options() e.g. options(cmdhelp=TRUE) to get compiled help (default installation option)
  > source() code from file and sink() results to a file

R Language Basics (2)

- R is *case sensitive*
- Comments can be put almost anywhere, starting with a hash mark ('#')
- Commands are separated either by a semi-colon (';'), or by a newline
  - Commands can be grouped together into one compound expression by braces ('{' and '}')
- The entities that R creates are known as *objects*
  - The collection of current objects is the *workspace*
    > objects() to list the current objects
    > rm(<object>) to remove an object from the workspace

Storing Code and Objects

- At the end of each R session you are given the opportunity to save all the currently available objects
  - The objects are written to a file called .RData in the current directory, and the command lines used in the session are saved to a file called .Rhistory
- RStudio can work with script files
  - Called .R

Microsoft

Reach us with #ntk17

KONFERENCA

R Collections and Objects

- *Matrices* or more generally *arrays* are multi-dimensional generalizations of *vectors*
- *Factors* provide compact ways to handle categorical data – distinct values are *levels*
- *Lists* are a general form of vector in which the various elements or *components* need not be of the same type
- *Data frames* are matrix-like structures, in which the columns can be of different types
- *Functions* can be stored in the project's workspace - a simple way to extend R

Microsoft

Reach us with #ntk17

KONFERENCA

Working with Vectors and Matrices

- A vector **x** of order *p* (or dimension p) is a column of p numbers
  - x <- c(2,0,0,4) # Use the *combine* function to generate a vector
- An *m* x *n* matrix **X** is a rectangular array of scalar values
  - Use the *matrix* function to generate a matrix from a vector
  - Use the *array* function to generate a 2-dim array from a vector
- Matrix operations
  - Addition, subtraction
  - Multiplication – only for matrices where the number of rows of the first one is equal to the number of columns of the second one
  - Combine by rows or by columns

Using SQL Server Data in R

- Get a SQL Server ODBC drive
- Create a system DSN
- Install RODBC package for R
- Activate the package
- Create a connection object
- Read the data into a data frame

```
con <- odbcConnect("AWDW2014", uid="RUser",
   pwd="Pa$$w0rd");
TM <- as.data.frame(sqlQuery(con,
  "SELECT CustomerKey, MaritalStatus, Gender,
    Region, BikeBuyer
   FROM dbo.vTargetMail"), stringsAsFactors = TRUE);
```

Graphics

- A simple histogram plot

```
Education = factor(Education, order=TRUE,
                   levels=c("Partial High School",
                            "High School","Partial College",
                            "Bachelors", "Graduate Degree"));
plot(Education, main = 'Education',
     xlab='Education', ylab ='Number of Cases',
     col="purple");
```

- Grouped bars

```
barplot(nofcases,
        main='Number of cars owned and bike buyer gruped',
        xlab='BikeBuyer', ylab ='NumberCarsOwned',
        legend=rownames(nofcases),
        col=c("black", "blue", "red", "orange", "yellow"),
        beside=TRUE);
```


Statistics

- Summary of the dataset

```
summary(TM);
```

- Detailed functions

```
mean(Age);
median(Age);
sd(Age);
```

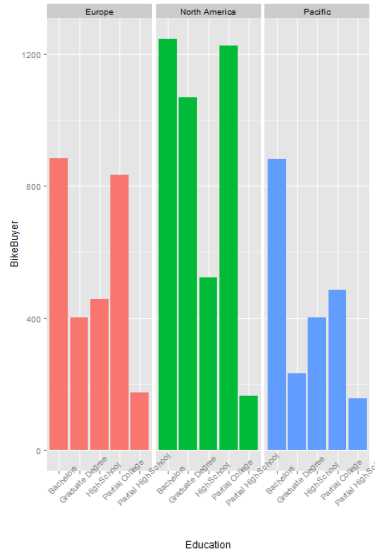
- Custom functions – skewness and kurtosis example

```
skewkurt <- function(p){
  avg <- mean(p)
  cnt <- length(p)
  stdev <- sd(p)
  skew <- sum((p-avg)^3/stdev^3)/cnt
  kurt <- sum((p-avg)^4/stdev^4)/cnt-3
  return(c(skewness=skew, kurtosis=kurt))};
skewkurt(Age);
```

The Most Popular Graphical Library - ggplot2

- ggplot2 implements a grammar of graphics (gg)
- Simple, consistent functions to build charts

```
ggplot(data=TM,
aes(x=Education,y=BikeBuyer,
fill=Region))
geom_bar(stat="identity")
facet_grid(.~Region)  theme(legend.position="none",
axis.text.x=
element_text(angle=45))
```



Data Mining in R

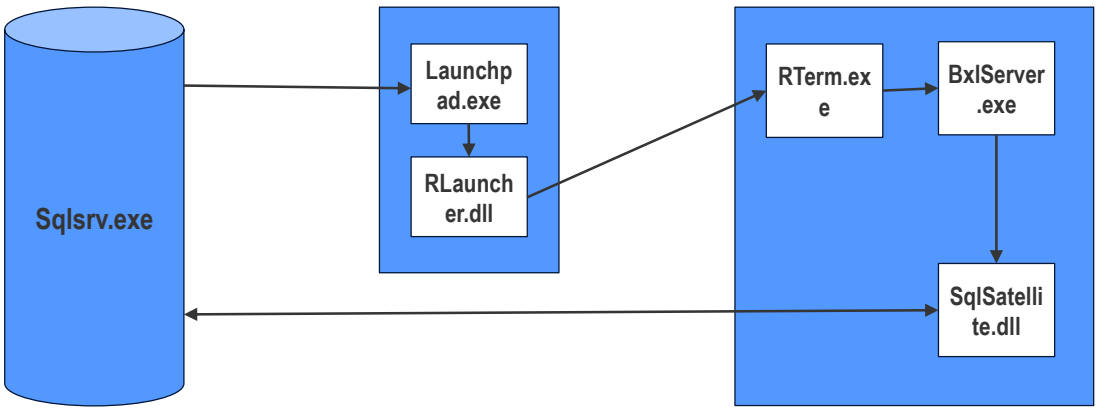
- Many, many algorithms in different packages
  - All popular algorithms
  - Can become confusing

```
# Package party (Decision Trees)
install.packages("party", dependencies = TRUE);
library("party");
# Train the model
TMDT <- ctree(BikeBuyer ~ NumberCarsOwned + Region,
              data = df_TM);
# Show the results
plot(TMDT, type = "simple");
```

Microsoft R

- Two flavors
  - **R Services (In-Database) –** this is the installation which integrates R in SQL Server
  - **Microsoft R Server** – this is a standalone R server with the same open and scalable packages that runs on multiple platforms
- Scalable functions
  - **RevoScaleR** – this is a set of parallelized scalable R functions for processing data, data overview and preliminary analysis, and machine learning models
  - **RevoPemaR –** this package allows you to write custom parallel external algorithms
  - **MicrosoftML –** more scalable algorithms coming soon

R in SQL Server



Execute R Script

```
sp_execute_external_script @language = N'language' ,
 @script = N'script',
 @input_data_1 = ] 'input_data_1‘
 [ , @input_data_1_name = ] N'input_data_1_name' ]
 [ , @output_data_1_name = 'output_data_1_name' ]
 [ WITH <execute_option> [ ,...n ] ] [;]

<execute_option>::=
{ { RESULT SETS UNDEFINED } |
{ RESULT SETS NONE } |
{ RESULT SETS ( <result_sets_definition> [,...n ] ) } }
```


RevoScaleR

- Library of fast, highly scalable R functions on multiple processors
- Don’t need to install the package or load the library if Revolution R Enterprise is installed with SQL Server
- Set of functions for:
  - Input / output
  - Data manipulations
  - Descriptive statistics and cross-tabulation
  - Statistical, data mining, and machine learning modeling
  - Graphing
  - Distributed computing
  - Hadoop convenience
  - Utility

Using R in SSRS

- No R data source
- Use SQL Server R integration
  - Execute T-SQL procedures that call external R script that creates a graph
  - Render the graph in appropriate MIME type, e.g. JPEG
  - Store graph in a SQL Server table to a VARBINARY(max) column
  - Read the binary value in a SSRS data set
  - Store the binary value in the SSRS Image control
    - Select database source and appropriate MIME type

Using R in SSIS

- No "Execute R Script" task or "R Script" data source
  - Execute Process Task in the Control Flow, call RScript in command prompt
  - Use the RDotNet library (CodePlex project), register all libraries needed, use R in the Script task in the Control Flow or Script component in the Data Flow
  - Use SQL Server 2016 R integration and execute T-SQL procedures that call external R script
    - Execute SQL task in the Control Flow
    - OLE DB source in the Data Flow

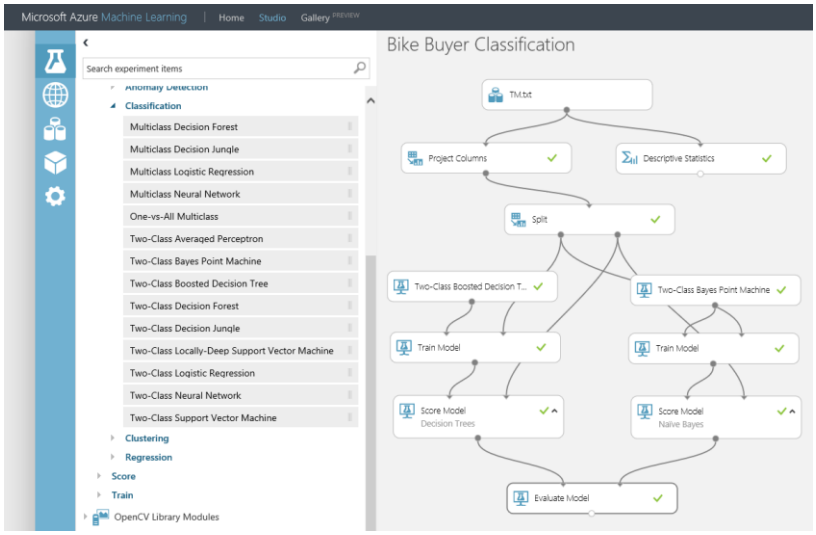
Using R in Power BI

- To run R scripts in Power BI Desktop, you need to install R on your local machine
  - Create the script in your local R development environment
- In Power BI Desktop, the *R Script data connector* is found in Get Data
  - To run your R Script, select Get Data > More…, then select Other > R Script
  - Only data frames are imported
- *R Script Visual control* enables R graphics on any Power BI data model
  - Creates a data frame from a projection of the data model data
  - Then uses R script o generate the plots
- *R Custom Visuals* at Office Store

Microsoft

KONFERENCA

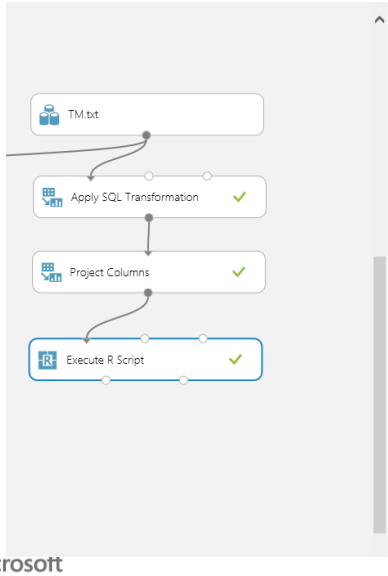
Azure ML Experiment



Microsoft

KONFERENCA

Extensibility



Execute R Script

R Script

```
# load("src/yourData.rdata");

# create a distance matrix from the data
ds <- dist(df_TM, method = "euclidean")

# Hierarchical clustering model
TMCL <- hclust(ds, method="ward.D2")

# Display the dendrogram
plot(TMCL)

# Cut tree into 6 clusters
groups <- cutree(TMCL, k=6)

# Draw dendogram with red borders around the 6 clusters
rect.hclust(TMCL, k=6, border="red")


# Select data.frame to be sent to the output Dataset port
maml.mapOutputPort("df_TM");
```

Random Seed

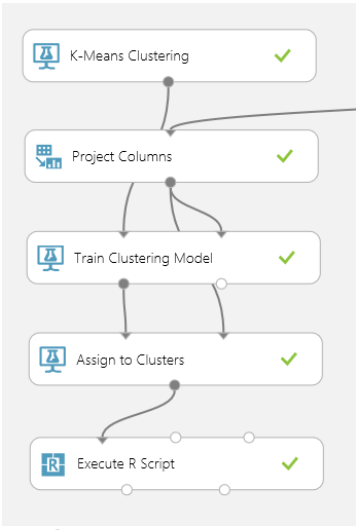
START TIME	3/21/2015 4:34:48 PM
END TIME	3/21/2015 4:34:48 PM
ELAPSED TIME	0:00:00.000
STATUS CODE	Finished
STATUS DETAILS	Task output was present in output cache

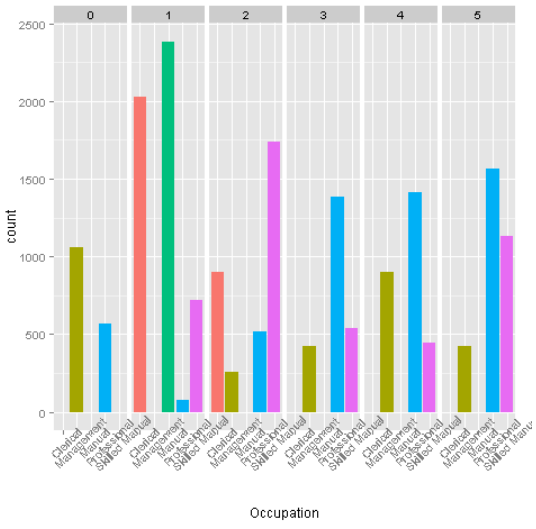
Microsoft

Reach us with #ntk17

KONFERENCA

Azure ML and R Visualizations



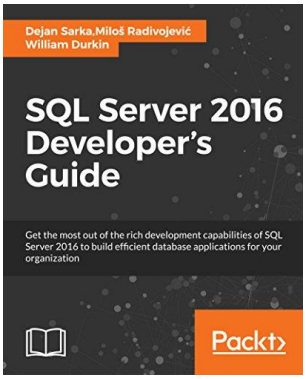


Microsoft

Reach us with #ntk17

KONFERENCA

Q & A



- Only for attendees: 40% discount on e-book!
  - Discount Code: **EBSQLS40**
  - Start date: May 1st 2017
    Expiry date: May 30th 2017
  - Create a login on the Packt site www.packtpub.com and add the book to the cart
  - Click "View Cart"
  - "Do you have a promo code?" field enter (code provided above)
  - Click the "Apply" button to apply the discount

• Thank you!

Microsoft

Reach us with #ntk17

KONFERENCA

Microsoft

KONFERENCA
PORTOROŽ, 15. DO 17. MAJ 2017