

A faint, light gray network diagram in the background, consisting of several interconnected circles of varying sizes. One circle in the center-left is filled with a solid dark blue color, while the others are hollow with gray outlines. Lines connect the circles, forming a web-like structure.

Oracle R Enterprise

Proširenja su u paketima

Krešimir Bokulić
Branko Radovanović

Multicom

- **Glavna područja ekspertize:**

- Data Mining
- Obračun i naplata (**Billing**)
- Upravljanje matičnim podacima (**MDM**)
- Skladišta podataka (**DWH**) i Poslovna Inteligencija (**BI**)
- **B2B**
- Upravljanje korisničkim procesima (**CRM**)



Reference

Telco

...

Finance

...

Public/Utility



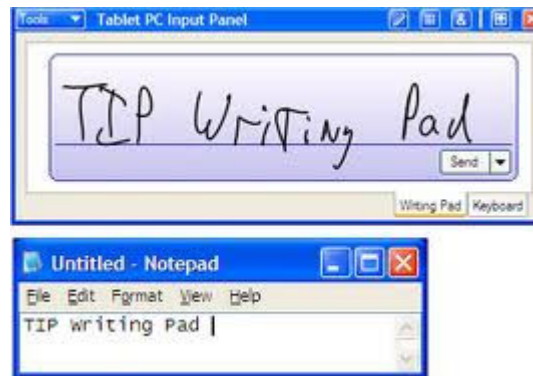
City of Zagreb



Advanced Analytics - svuda oko nas



Spam filters



Prepoznavanje rukopisa

Ads ⓘ

Google AdWords

www.google.hr/AdWords ▾

Vaš oglas na svjetskoj tražilici.

Oglašavajte na Googleu!

Online Oglašavanje

www.interartmedia.net/GoogleAdwords ▾

Unaprijedite svoje poslovanje!

Adwords™ certificirani partner

Expedia Hotels

www.expedia.ie/Hotels ▾

Book Online & Save up to 50%.

Great deals on Hotels

Hotels.com: Cheap Hotels

www.hotels.com/Cheap-Hotels ▾

Exclusive Deals, Central Locations!

Search & Book Cheap Hotels online.

Google ads



Loyalty cards



Credit risk

Frequently Bought Together



Price for all three: **\$77.97**

[Add all three to Cart](#) [Add all three to Wish List](#)

[Show availability and shipping details](#)

- This item:** Python Pocket Reference (Pocket Reference (O'Reilly)) by Mark Lutz Paperback **\$8.99**
- Learning Python, 5th Edition by Mark Lutz Paperback **\$38.99**
- Python Cookbook by David Beazley Paperback **\$29.99**

Customers Who Bought This Item Also Bought

			
C++ Pocket Reference by Kyle Loudon ★★★★☆ (21) Paperback \$9.45 ✓Prime	JavaScript Pocket Reference (Pocket ...) by David Flanagan ★★★★☆ (14) Paperback \$10.16 ✓Prime	Python Essential Reference (4th ...) by David M. Beazley ★★★★★ (71) Paperback \$27.21 ✓Prime	Perl Pocket Reference by Johan Vromans ★★★★★ (18) Paperback \$10.06 ✓Prime

Amazon recommendation engine

Što je R?

- R je Open Source jezik i okolina za statističke proračune i grafiku
- Stvoren 1994 kao alternativa SAS-u i SPSS-u
- Preko 2 milijuna R korisnika u svijetu
- Tisuće open source paketa na CRAN mreži
- CRAN – Comprehensive R Archive Network



CRAN
[Mirrors](#)
[What's new?](#)
[Task Views](#)
[Search](#)

About R
[R Homepage](#)
[The R Journal](#)

Software
[R Sources](#)
[R Binaries](#)
[Packages](#)
[Other](#)

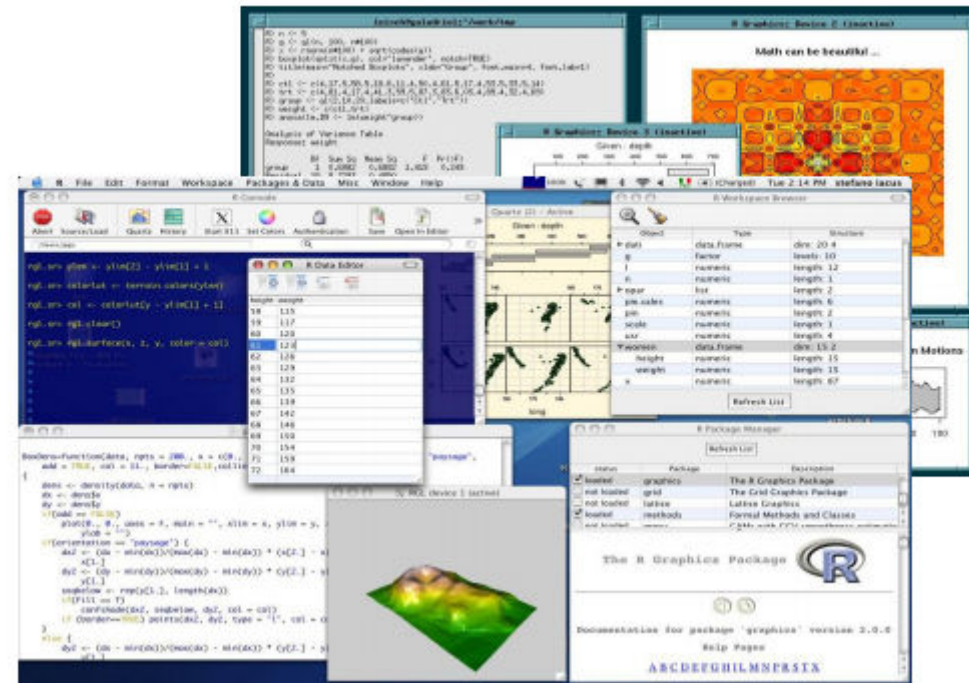
Documentation
[Manuals](#)
[FAQs](#)
[Contributed](#)

	CRAN Task Views
Bayesian	Bayesian Inference
ChemPhys	Chemometrics and Computational Physics
ClinicalTrials	Clinical Trial Design, Monitoring, and Analysis
Cluster	Cluster Analysis & Finite Mixture Models
DifferentialEquations	Differential Equations
Distributions	Probability Distributions
Econometrics	Computational Econometrics
Environmetrics	Analysis of Ecological and Environmental Data
ExperimentalDesign	Design of Experiments (DoE) & Analysis of Experimental Data
Finance	Empirical Finance
Genetics	Statistical Genetics
Graphics	Graphic Displays & Dynamic Graphics & Graphic Devices & Visualization
HighPerformanceComputing	High-Performance and Parallel Computing with R
MachineLearning	Machine Learning & Statistical Learning
MedicalImaging	Medical Image Analysis
MetaAnalysis	Meta-Analysis
Multivariate	Multivariate Statistics
NaturalLanguageProcessing	Natural Language Processing
NumericalMathematics	Numerical Mathematics
OfficialStatistics	Official Statistics & Survey Methodology
Optimization	Optimization and Mathematical Programming
Pharmacokinetics	Analysis of Pharmacokinetic Data
Phylogenetics	Phylogenetics, Especially Comparative Methods
Psychometrics	Psychometric Models and Methods

Zašto koristiti R?

R okolina je:

- Proširiva
- Omogućuje kvalitetnu grafiku
- Jednostavna za instalaciju
- Produktivna
- Fleksibilna
- Besplatna
- Otvorena



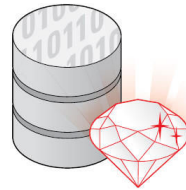
R je statistički jezik sličan Base SAS –u i SPSS-u

Oracle Advanced Analytics

Oracle Advanced Analytics:

- Oracle R Enterprise
 - Integrira open source programski jezik R unutar Oracle baze podataka
- Oracle Data Mining
 - SQL & PL/SQL fokusiran na in-database data mining –u

ORACLE®



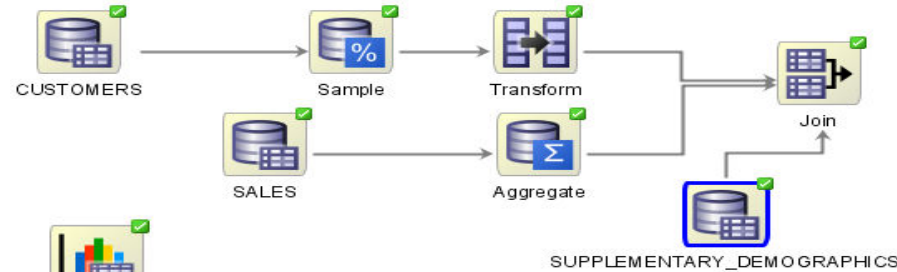
- Opcija Oracle 11gR2 enterprise baze
- Omogućuje naprednu analitiku unutar baze podataka

Oracle Data Miner

Tables and Views



Transformations

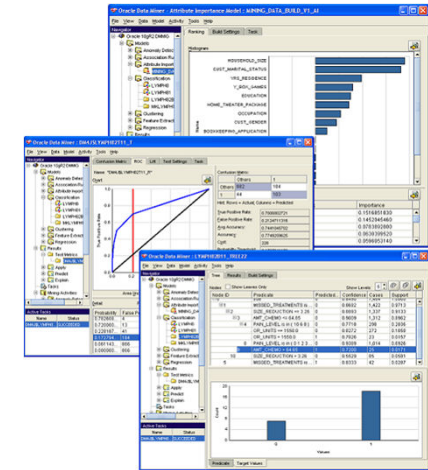
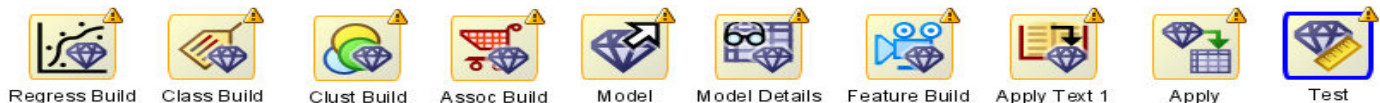


Explore Data



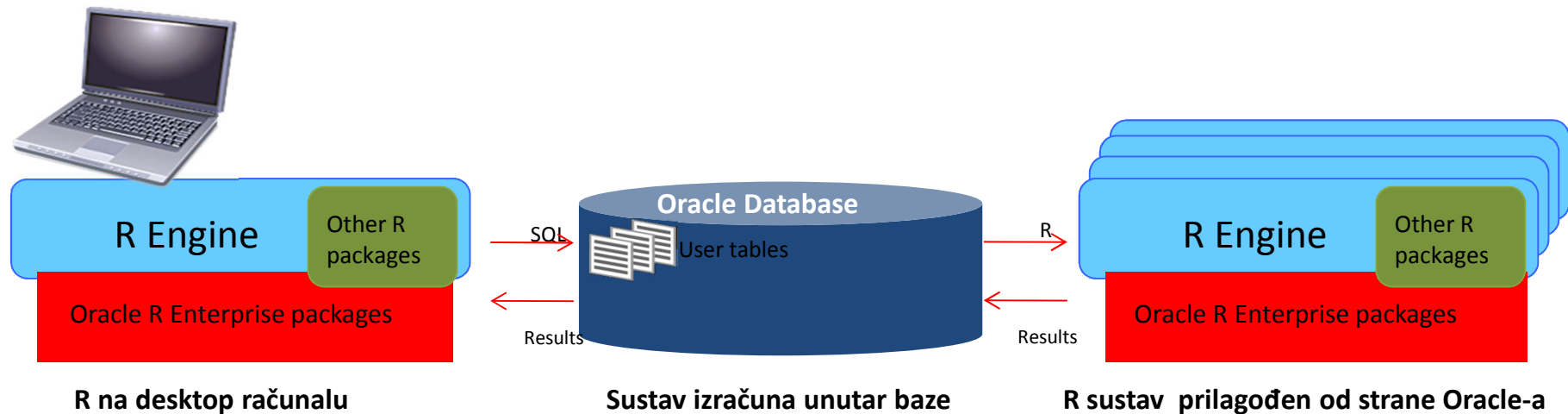
Explore Data

Modeling



- Obuhvaćeni svi aspekti izrade modela: dohvat, transformacija, procesiranje podataka, te izrada i evaluacija modela
- Interaktivna vizualizacija i izrada izvještaja, uz set predefiniраниh izvještaja i analitika
- Integriran u SQL developer, programiranje nepotrebno
- Tijek procesa gradnje modela očuvan unutar grafičkog workflow-a

Oracle R Enterprise



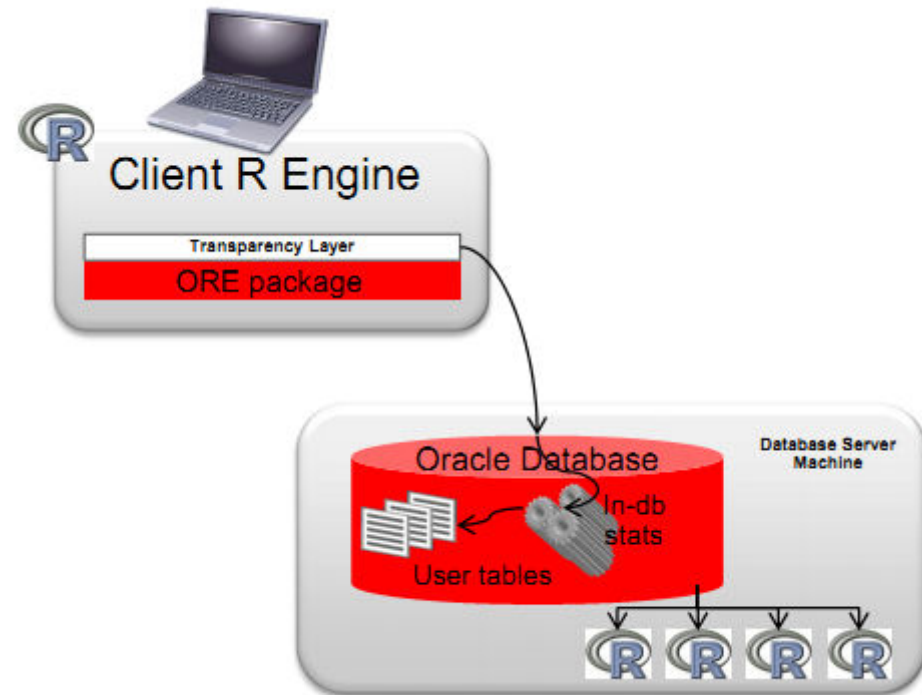
- R-SQL Transparency Framework presreće R funkcije i prosljeđuje ih serveru
- Transformacije podataka, statističke funkcije i napredne analitičke funkcije
- Grafički prikaz rezultata
- Tijek izvršavanja se nalazi unutar R skripte

- Omogućuje obradu velikih količina podataka
- Pristup tablicama, view-ovima, eksternim tablicama i DB linkovima
- Koristi SQL paralelizam
- Iskorištava postojeće statističke i data mining SQL funkcije i algoritme

- Skaliranje na više R sustava kako bi se iskoristio paralelizam
- Koristi map-reduce tip algoritama

Oracle R Enterprise poboljšanja

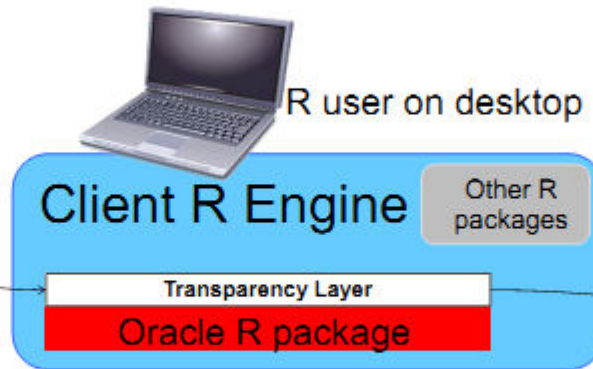
- Eliminira problem memorije na klijentskom računalu
- Omogućuje paralelizaciju i optimizaciju upita na bazi
- Paralelizira izvršavanja R skripte multiplicirajući R engine
- Omogućuje pozivanje R skripti kroz SQL i PL/SQL



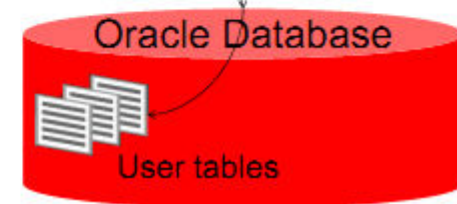
Transparency Layer

SELECT primjer: dohvat podataka iz baze u klijent R

```
class(ONTIME_S)  
dim(ONTIME_S)  
ontime <- ore.pull(ONTIME_S)  
class(ontime)  
dim(ontime)
```



```
select *  
from ONTIME_S
```



R skripta na serveru (Embedded R Script)

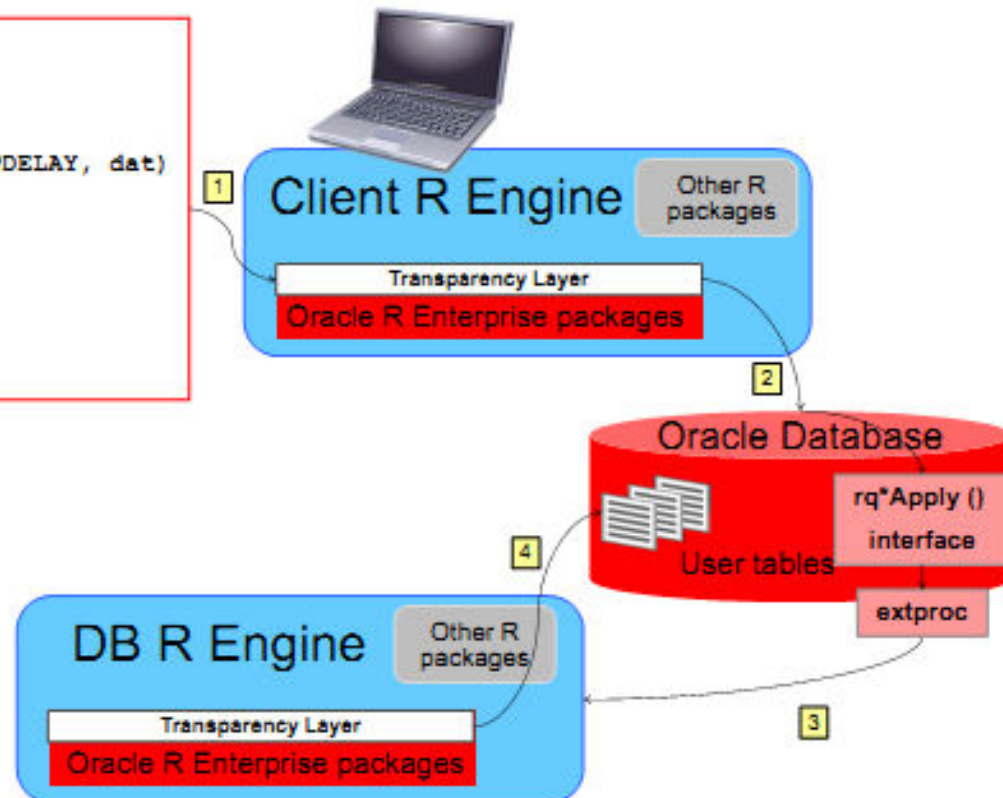
Izvršavanje kroz R okolinu (generiran LM):

```

mod <- ore.doEval(
  function(param) {
    dat <- ore.pull(ONTIME_S)
    mod <- lm(ARRDELAY ~ DISTANCE + DEPDELAY, dat)
    mod
  });
mod_local <- ore.pull(mod)
class(mod_local)
summary(mod_local)

```

Goal: Build a single regression model using Transparency Layer in DB R Engine
 Data explicitly loaded into R memory at DB R Engine using ore.pull()
 Result "mod" returned as a model object



R skripta na serveru (Embedded R Script)

R funkcije za embedded R skripte:

R Interface function	Purpose
<code>ore.doEval()</code>	Invoke stand-alone R script
<code>ore.tableApply()</code>	Invoke R script with full table as input
<code>ore.rowApply()</code>	Invoke R script on one row at a time, or multiple rows in chunks
<code>ore.groupApply()</code>	Invoke R script on data partitioned by grouping column
<code>ore.indexApply()</code>	Invoke R script N times
<code>ore.scriptCreate()</code>	Create an R script in the database
<code>ore.scriptDrop()</code>	Drop an R script in the database

Comprehensive R Archive Network

- Mreža R paketa koji proširuju osnovne funkcionalnosti R-a
- Paketi pisani u R, Javi, C i Fortran jeziku
- Preko 5800 dodatnih paketa



Popis popularnijih paketa:

- Plyr
- Reshape2
- Stringr
- Ggplot2
- googleVis
- klaR
- Glmnet
- Survival
- Xml
- Parallel
- Xts

Information Value

$$IV = \sum (DistributionGood_i - DistributionBad_i) \times \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

$$Weight\ of\ Evidence = \ln\left(\frac{DistributionGood_i}{DistributionBad_i}\right)$$

Age Group	Total Number of loans	Number of Bad loans	Number of Good Loans	% Bad loans	Name of Coarse Groups	Distribution of loans	Distribution Bad (DB)	Distribution Good (DG)	WOE	DG - DB	(DG - DB) * WOE
21-30	4821	206	4615	4.3%	G1	0.079	0.135	0.078	-0.553	-0.057	0.0318
30-36	10266	357	9909	3.5%	G2	0.169	0.235	0.167	-0.339	-0.067	0.0228
36-48	32926	776	32150	2.4%	G3	0.542	0.510	0.542	0.062	0.032	0.0020
48-60	12788	183	12605	1.4%	G4	0.210	0.120	0.213	0.570	0.092	0.0527
Total	60801	1522	59279							Information Value -->	0.1093

ORE i Poziv paketa 1

```
1 # 1. Način izračuna IV vrijednosti varijabli
2 # Povlačenje podataka iz baze
3
4 library(klar)
5 library(ORE)
6
7 #Oracle User connection
8 if(!ore.is.connected())
9   ore.connect
10  ore.sync()
11
12
13 fm(list = ls(all = TRUE))
14 dt <- ore.pull(INSUR_CUST_LTV_SAMPLE_BIN)
15 grouping_column <- 'BUY_INSURANCE'
16
17
18 dt[is.na(dt)] <- "N/A" # Zamjena null vrijednosti
19 for (i in which(sapply(dt, class) != "factor")) dt[[i]] <- as.factor(dt[[i]]) # Pretvaranje u faktore
20 gc <- dt[[grouping_column]]
21 dt[[grouping_column]] <- NULL
22 woemodel <- woe(dt, grouping=gc, appont=FALSE, zeroadj=0.5) # Izračun koristeći klar paket
23 iv <- woemodel$IV
24 as.data.frame(iv)
```

```
> iv <- woemodel$IV
> as.data.frame(iv)
      iv
AGE_BIN      0.08070162
BANK_FUNDS_BIN_Q 3.14069519
CAR_OWNERSHIP  0.06358286
MONEY_MONTHLY_OVERDRA_BIN 1.80768571
N_TRANS_ATM    1.82629870
SALARY_BIN     0.02009627
> |
```

ORE i Poziv paketa 2

```

1 # 2. Način izračuna IV vrijednosti varijabli
2 # Povlačenje podataka iz baze i koristimo funkciju koja koristi klar paket
3
4 library (ORE)
5
6 #Oracle User connection
7 if(!ore.is.connected())
8   ore.connect
9 ore.sync()
10
11
12 rm(list = ls(all = TRUE))
13
14
15 # Definiranje funkcije za izračun IV koristeći klar paket
16 iv <- function(x, grouping_column, ...) {
17   library(klar)
18   x[is.na(x)] <- "N/A"
19   for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
20   gc <- x[[grouping_column]]
21   x[[grouping_column]] <- NULL
22   woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
23   iv <- woemodel$IV
24   return(data.frame(name = names(iv), iv))
25 }
26
27 #-----
28 |
29 dt <- ore.pull(INSUR_CUST_LTV_SAMPLE_BIN)
30 grouping_column <- 'BUY_INSURANCE'
31
32 iv(dt,grouping_column)
33

```

```

> grouping_column <- 'BUY_INSURANCE'
>
> iv(dt,grouping_column)

```

	name	iv
1	AGE_BIN	0.08070162
2	BANK_FUNDS_BIN_Q	3.14069519
3	CAR_OWNERSHIP	0.06358286
4	MONEY_MONTHLY_OVERDRA_BIN	1.80768571
5	N_TRANS_ATM	1.82629870
6	SALARY_BIN	0.02009627
..	.	.

ORE i Poziv paketa 3

```

1 # 3. Način izračuna IV vrijednosti varijabli
2 # slanje f-je kroz parametar
3
4 library (ORE)
5
6 #Oracle User connection
7 if(!ore.is.connected())
8   ore.connect(user=
9 ore.sync()
10
11
12 rm(list = ls(all = TRUE))
13
14
15 # Definiranje funkcije za izračun IV koristeći klar paket
16 iv<- function(x, grouping_column, ...)
17 {
18   library(klar)
19   x[is.na(x)] <- "N/A"
20   for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
21   gc <- x[[grouping_column]]
22   x[[grouping_column]] <- NULL
23   woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
24   iv <- woemodel$IV
25   return(data.frame(name = names(iv), iv))
26 }
27
28
29 #-----
30
31 dt <- INSUR_CUST_LTV_SAMPLE_BIN
32 grouping_column <- 'BUY_INSURANCE'
33
34 ore.tableApply(dt,FUN =iv,grouping_column=grouping_column)
35
36
37 > grouping_column <- 'BUY_INSURANCE'
38 >
39 > ore.tableApply(dt,FUN =iv,grouping_column=grouping_column)
40
41      name      iv
42 AGE_BIN      AGE_BIN 0.08070162
43 BANK_FUNDS_BIN_Q  BANK_FUNDS_BIN_Q 3.14069519
44 CAR_OWNERSHIP      CAR_OWNERSHIP 0.06358286
45 MONEY_MONTHLY_OVERDRA_BIN MONEY_MONTHLY_OVERDRA_BIN 1.80768571
46 N_TRANS_ATM        N_TRANS_ATM 1.82629870
47 SALARY_BIN         SALARY_BIN 0.02009627

```


ORE i Poziv paketa 4

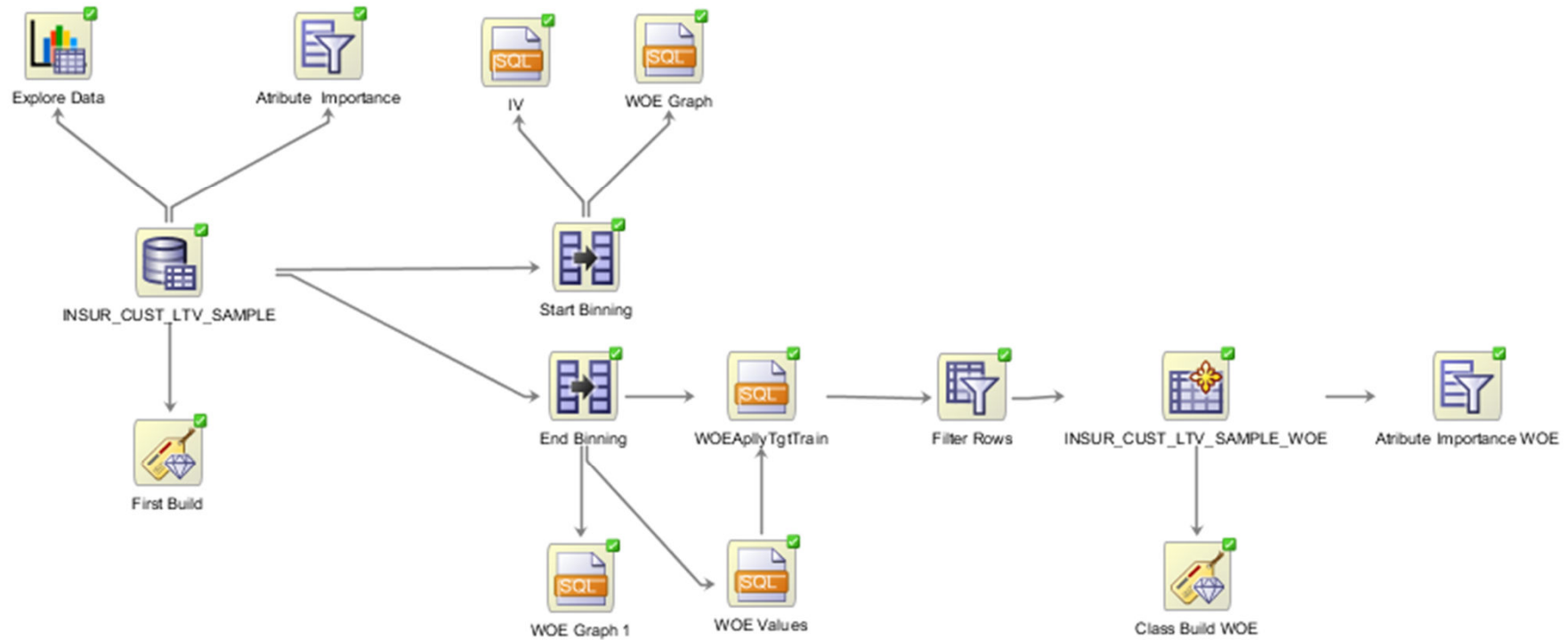
```

1 # 4. Način izračuna IV vrijednosti varijabli
2 # kreiranje f-je na bazi i njezin poziv
3
4 library (ORE)
5
6 #Oracle user connection
7 if(!ore.is.connected())
8   ore.connect(user=, password=)
9 ore.sync()
10
11 rm(list = ls(all = TRUE))
12
13 # Definiranje funkcije za izračun IV koristeći klar paket
14 ore.scriptDrop("woe.iv")
15 ore.scriptCreate("woe.iv",
16   function(x, grouping_column, ...)
17 {
18   library(klar)
19   x[is.na(x)] <- "N/A"
20   for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
21   gc <- x[[grouping_column]]
22   x[[grouping_column]] <- NULL
23   woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
24   iv <- woemodel$IV
25   return(data.frame(name = names(iv), iv))
26 }
27 )
28
29
30 #-----
31
32 dt <- INSUR_CUST_LTV_SAMPLE_BIN
33 grouping_column <- 'BUY_INSURANCE'
34
35 ore.tableApply(dt, FUN.NAME="woe.iv", grouping_column=grouping_column)
36

```

name	iv
AGE_BIN	0.08070162
BANK_FUNDS_BIN_Q	3.14069519
CAR_OWNERSHIP	0.06358286
MONEY_MONTHLY_OVERDRA_BIN	1.80768571
N_TRANS_ATM	1.82629870
SALARY_BIN	0.02009627

ODM WF i R paketi



ODM WF i R paketi

Source Column: AGE




Transform Type: Binning

Output Column: AGE_BIN Auto

Binning

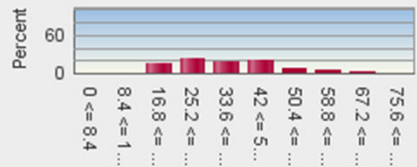
Binning Type: Custom

Bin Name	Lower Bound
1	No Lower Bound
2	29.0
3	41.0

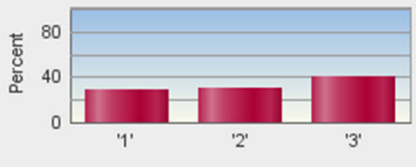
Reset   

Statistics

AGE


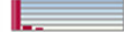
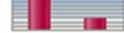






AGE_BIN

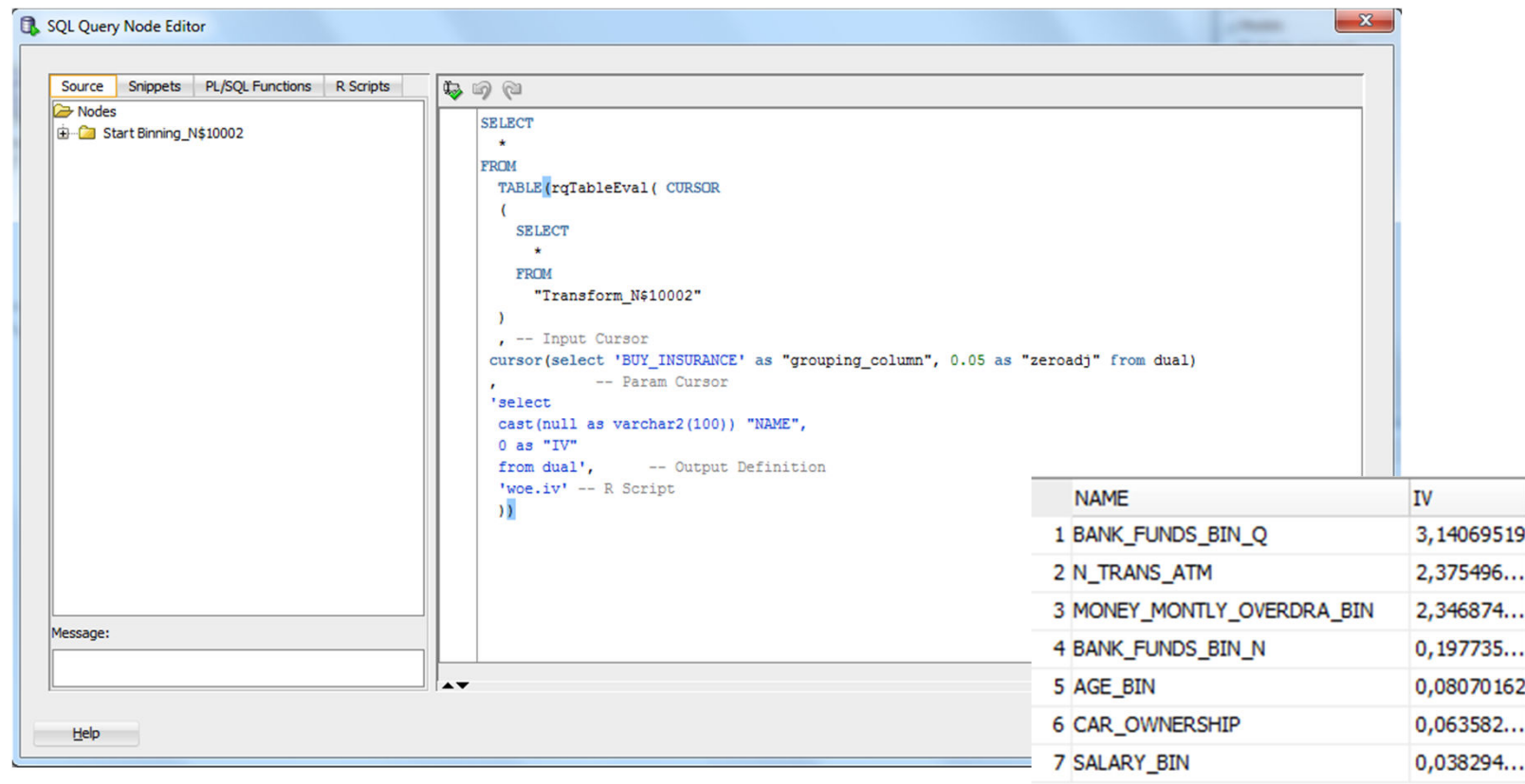


Update Statistics when Finished

Help OK Cancel

Name	Histogram	Data Type
AGE		NUMBER
BANK_FUNDS		NUMBER
BUY_INSURANCE		VARCHAR2
CAR_OWNERSHIP		NUMBER
MONEY_MONTHLY_OVERDRAWN		NUMBER
N_TRANS_ATM		NUMBER
SALARY		NUMBER

ODM WF i R paketi



The screenshot shows the 'SQL Query Node Editor' window. The main editor area contains the following SQL query:

```

SELECT
*
FROM
TABLE(rqTableEval( CURSOR
(
SELECT
*
FROM
"Transform_N$10002"
)
, -- Input Cursor
cursor(select 'BUY_INSURANCE' as "grouping_column", 0.05 as "zeroadj" from dual)
, -- Param Cursor
'select
cast(null as varchar2(100)) "NAME",
0 as "IV"
from dual', -- Output Definition
'woe.iv' -- R Script
))
  
```

Below the query editor, a table displays the results of the query:

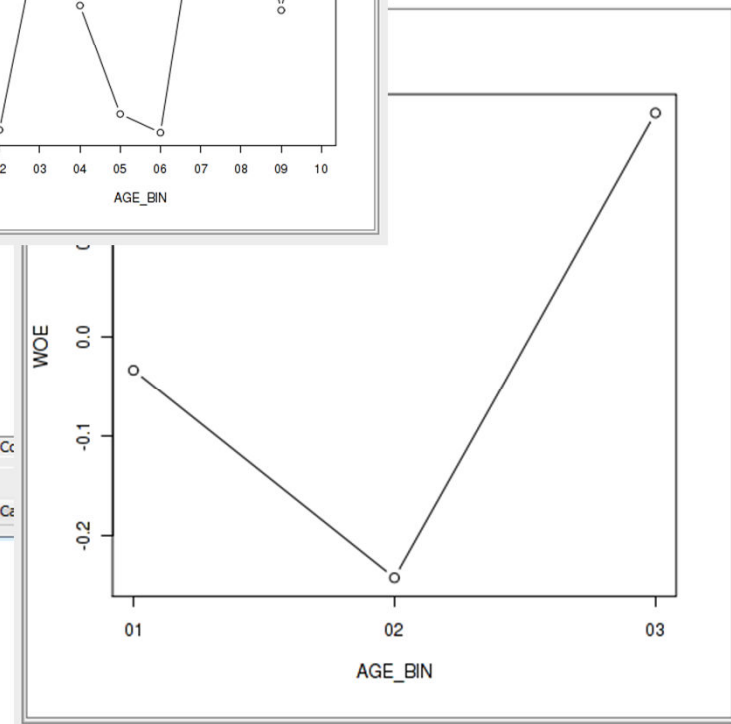
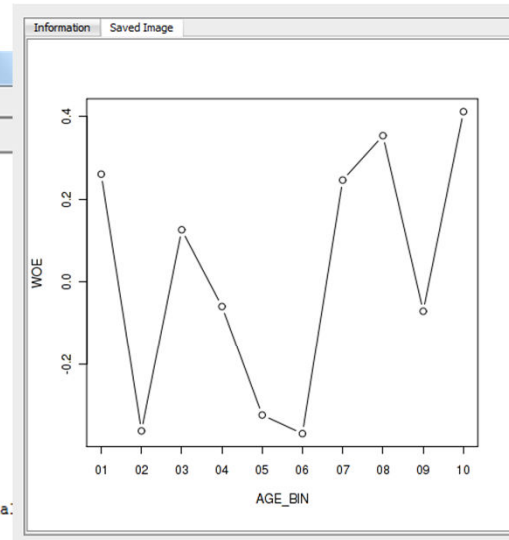
	NAME	IV
1	BANK_FUNDS_BIN_Q	3,14069519
2	N_TRANS_ATM	2,375496...
3	MONEY_MONTHLY_OVERDRA_BIN	2,346874...
4	BANK_FUNDS_BIN_N	0,197735...
5	AGE_BIN	0,08070162
6	CAR_OWNERSHIP	0,063582...
7	SALARY_BIN	0,038294...

ODM WF i R paketi

```

SELECT
*
FROM
TABLE(rqTableEval( CURSOR
(
SELECT
"End Binning_N$10012"."BUY_INSURANCE"
,"End Binning_N$10012"."AGE_BIN"
,"End Binning_N$10012"."BANK_FUNDS_BIN"
,"End Binning_N$10012"."MONEY_MONTHLY_OVERDRA_BIN"
,"End Binning_N$10012"."N_TRANS_ATM_BIN"
FROM
"End Binning_N$10012"
)
, -- Input Cursor
cursor(select 'BUY_INSURANCE' as "grouping_column", 0.05 as "zeroadj" from dual
, -- Param Cursor
'select
cast(null as varchar2(100)) as "VARIABLE",
cast(null as varchar2(100)) as "VAR_LEVEL",
0 as "WOE"
from dual', -- Output Definition
'woe.woe2' -- R Script
))

```



Zaključak

- Značajno proširenje funkcija R-a kroz CRAN pakete
- Omogućuje korištenje istih kroz R kod ali i kroz Oracle Data miner GUI

Popis popularnijih paketa:

- Plyr
- Reshape2
- Stringr
- Ggplot2
- googleVis
- klaR
- Glmnet
- Survival
- Xml
- Parallel
- Xts



...Hvala!



Q&A

kresimir.bokulic@multicom.hr
branko.radovanovic@multicom.hr