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**)

Deloitte. 2010 Technology Fast 50 First place in 2009. category "Rising Stars"

ISO 9001:2008 certified

multicom

Reference

Telco ...

T-Mobile

T-Com

Optima

H1

vip Vipnet

vip Vipmobile

tušmobil

Metronet

Македонски Телекомуникации

tel

GTW

Finance ...

HPB Hrvatska poštanska banka

Fina

MBU

ERSTE

otpbanka

Zagrebačka banka UniCredit Group

Public/Utility ....

Ministarstvo gospodarstva

Ministry of Foreign Affairs and European Integration

City of zagreb

HEP Operator prijenosnog sustava d.o.o.

SIEMENS

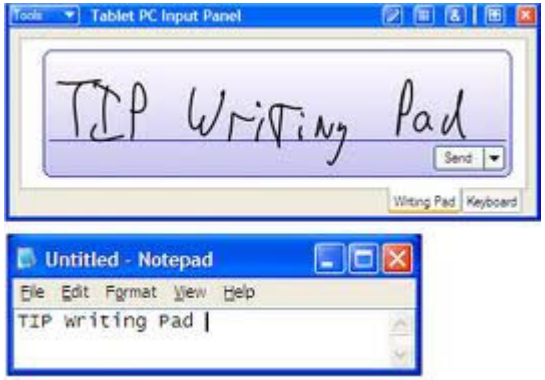
HOPS

KONČAR Končar - Inženjering za energetiku i transport d.d.

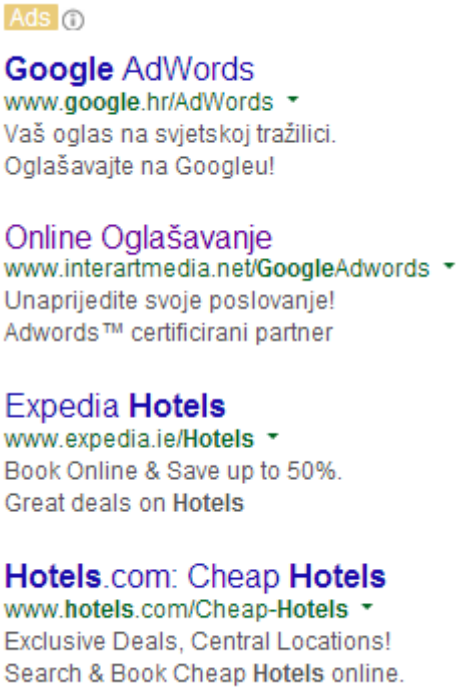
multicom

Advanced Analytics - svuda oko nas

Spam filters



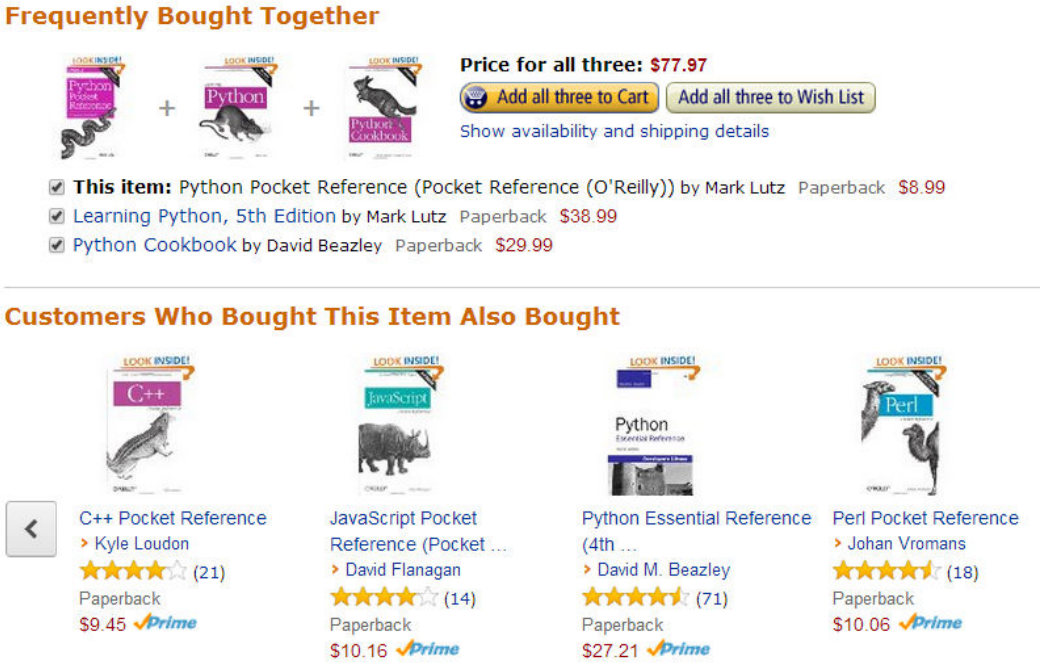
Prepoznavanje rukopisa



Google ads

Loyality cards

Credit risk



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

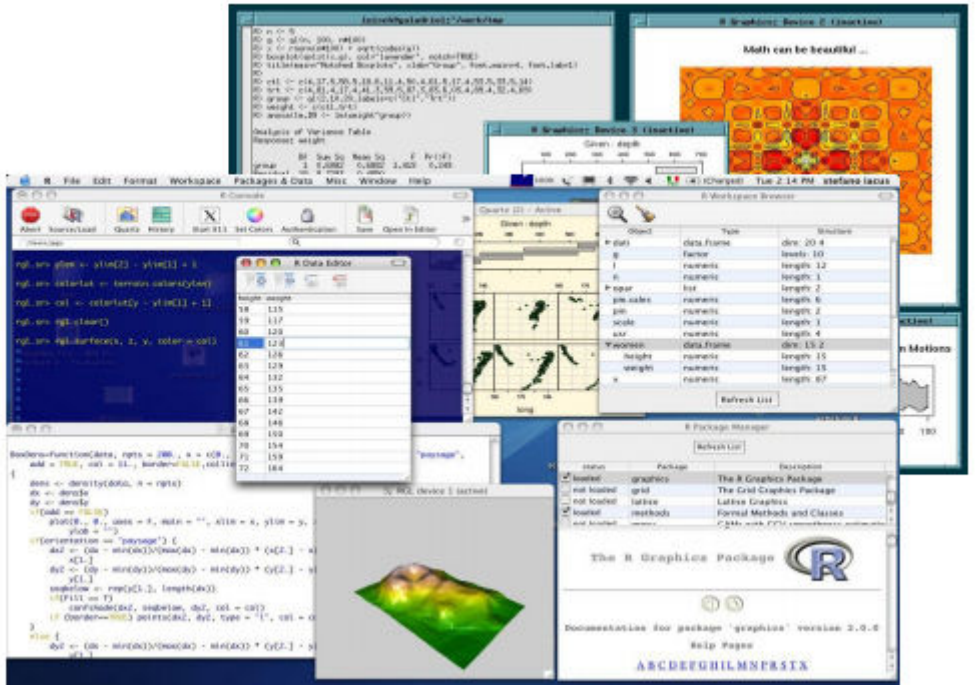
CRAN Task Views

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

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

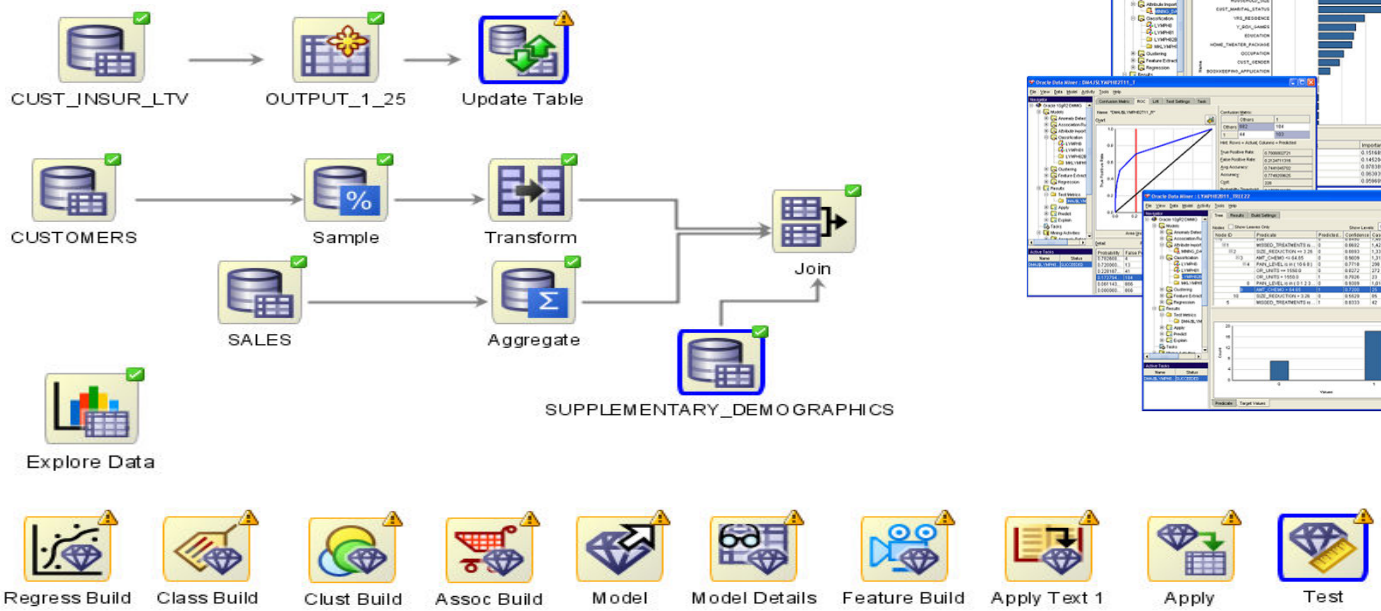
Oracle Data Miner

Tables and Views

Transformations

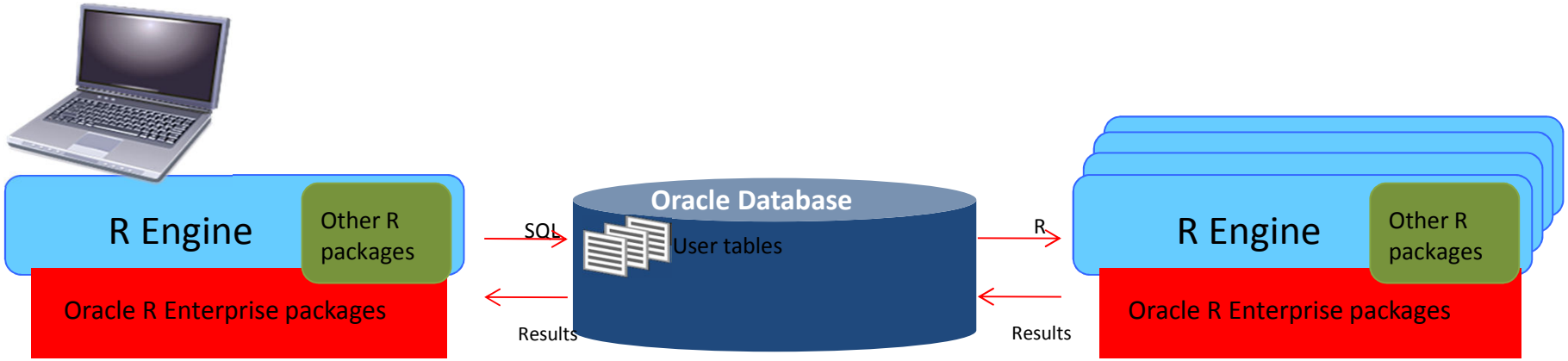
Explore Data

Modeling



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

Oracle R Enterprise



R na desktop računalu

- 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

Sustav izračuna unutar baze

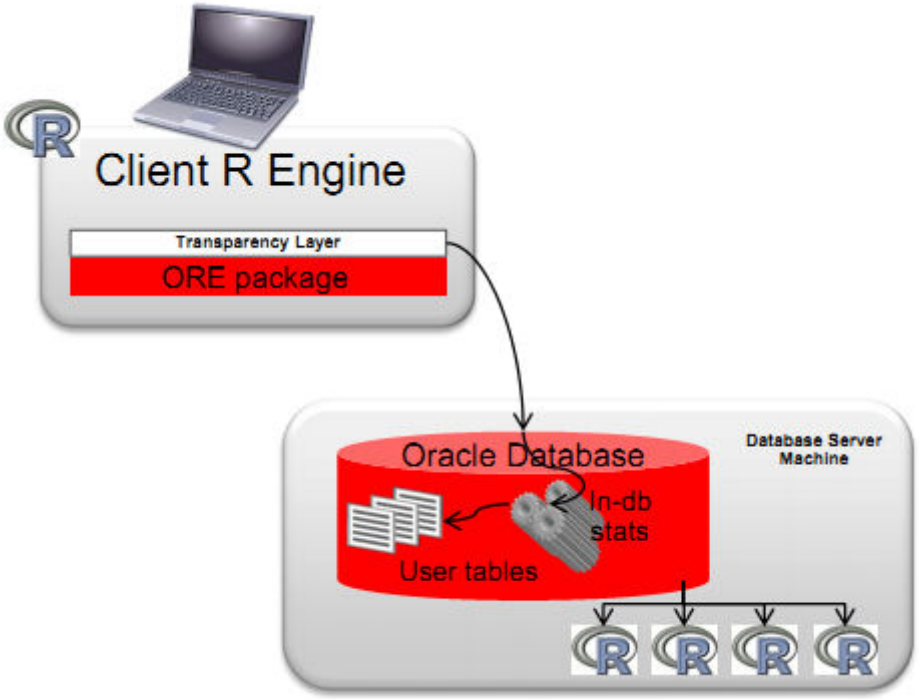
- Omogućuje obradu velikih količina podataka
- Pristup tablicama , view-evima, eksternim tablica i DB linkovima
- Koristi SQL paralelizam
- Iskorištava postojeće statističke i data mining SQL funkcije i algoritme

R sustav prilagođen od strane Oracle-a

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

Oracle R Enterprise poboljšanja

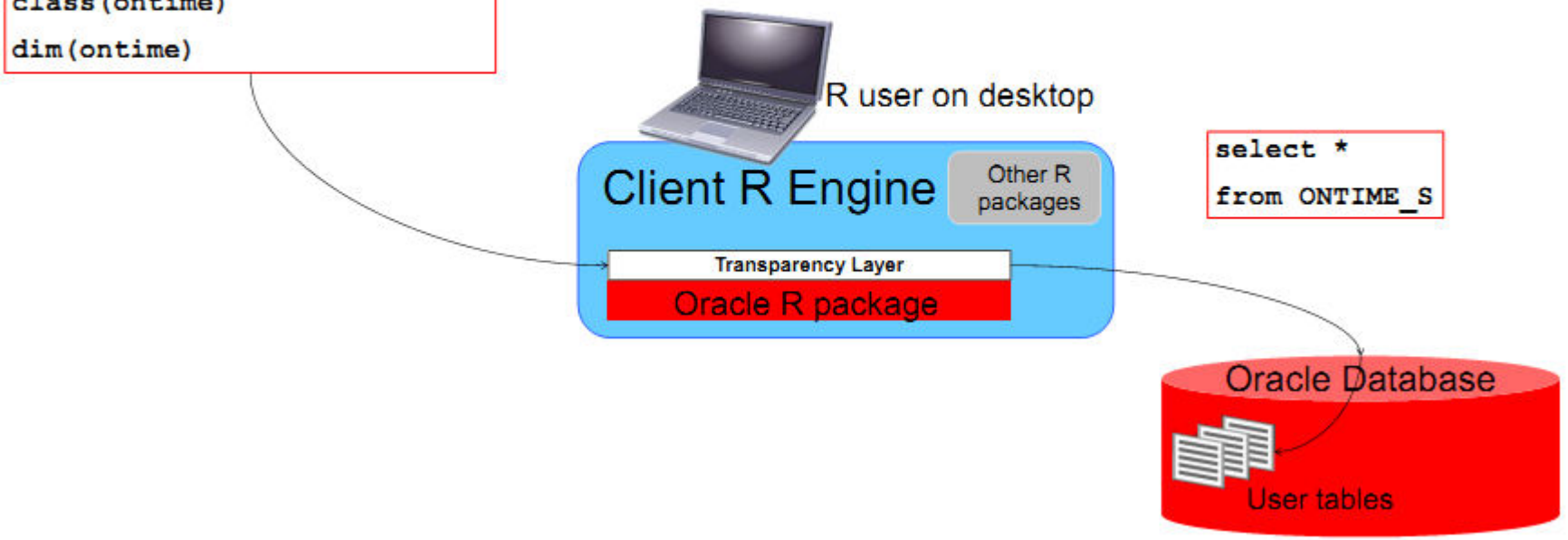
- Eliminira problem memorije na klijentskom računalu
- Omogućuje paralelizaciju i optimizaciju upita na bazi
- Paralelizira izvršavanja R skripte multiplicirajući R engine
- Omogućuje pozivanje R skripti kroz SQL i PL/SQL



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)
```



R skripta na serveru (Embedded R Script)

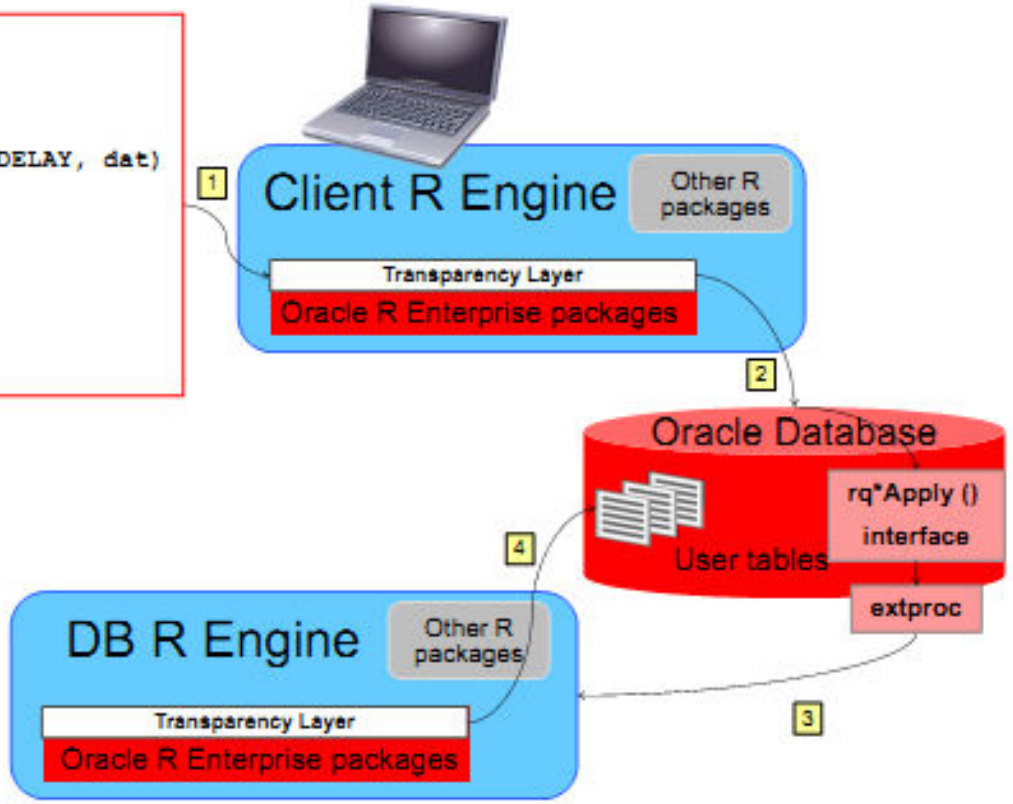
Izvršavanje kroz R okolinu (generiran LM):

```
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 srkipte:

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

Popis popularnijih paketa:

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

Information Value

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

$$WeightofEvidence = ln(\frac{DistributionGood_i}{DistributionBad_i})$$

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 rm(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_MONTLY_OVERDRA_BIN 1.80768571
N_TRANS_ATM             1.82629870
SALARY_BIN              0.02009627
>
```

ORE i Poziv paketa 2

```
# 2. Način izračuna IV vrijednosti varijabli
# Povlačenje podataka iz baze i koristimo funkciju koja koristi klaR paket

library (ORE)

#Oracle User connection
if(!ore.is.connected())
  ore.connect
ore.sync()


rm(list = ls(all = TRUE))


# Definiranje funkcije za izračun IV koristeći klaR paket
iv <- function(x, grouping_column, ...) {
  library(klaR)
  x[is.na(x)] <- "N/A"
  for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
  gc <- x[[grouping_column]]
  x[[grouping_column]] <- NULL
  woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
  iv <- woemodel$IV
  return(data.frame(name = names(iv), iv))
}

#-----------------------------------------------------


dt <- ore.pull(INSUR_CUST_LTV_SAMPLE_BIN)
grouping_column <- 'BUY_INSURANCE'

iv(dt,grouping_column)
```

```
> 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_MONTLY_OVERDRA_BIN 1.80768571
5                 N_TRANS_ATM 1.82629870
6                  SALARY_BIN 0.02009627
```

ORE i Poziv paketa 3

```
# 3. Način izračuna IV vrijednosti varijabli
# Slanje f-je kroz parametar

library (ORE)

#Oracle User connection
if(!ore.is.connected())
  ore.connect(user=
ore.sync()


rm(list = ls(all = TRUE))


# Definiranje funkcije za izračun IV koristeći klaR paket
iv<-   function(x, grouping_column, ...)
  {
    library(klaR)
    x[is.na(x)] <- "N/A"
    for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
    gc <- x[[grouping_column]]
    x[[grouping_column]] <- NULL
    woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
    iv <- woemodel$IV
    return(data.frame(name = names(iv), iv))
  }


#-------------------------------------------------


dt <- INSUR_CUST_LTV_SAMPLE_BIN
grouping_column <- 'BUY_INSURANCE'

ore.tableApply(dt,FUN =iv,grouping_column=grouping_column)
```

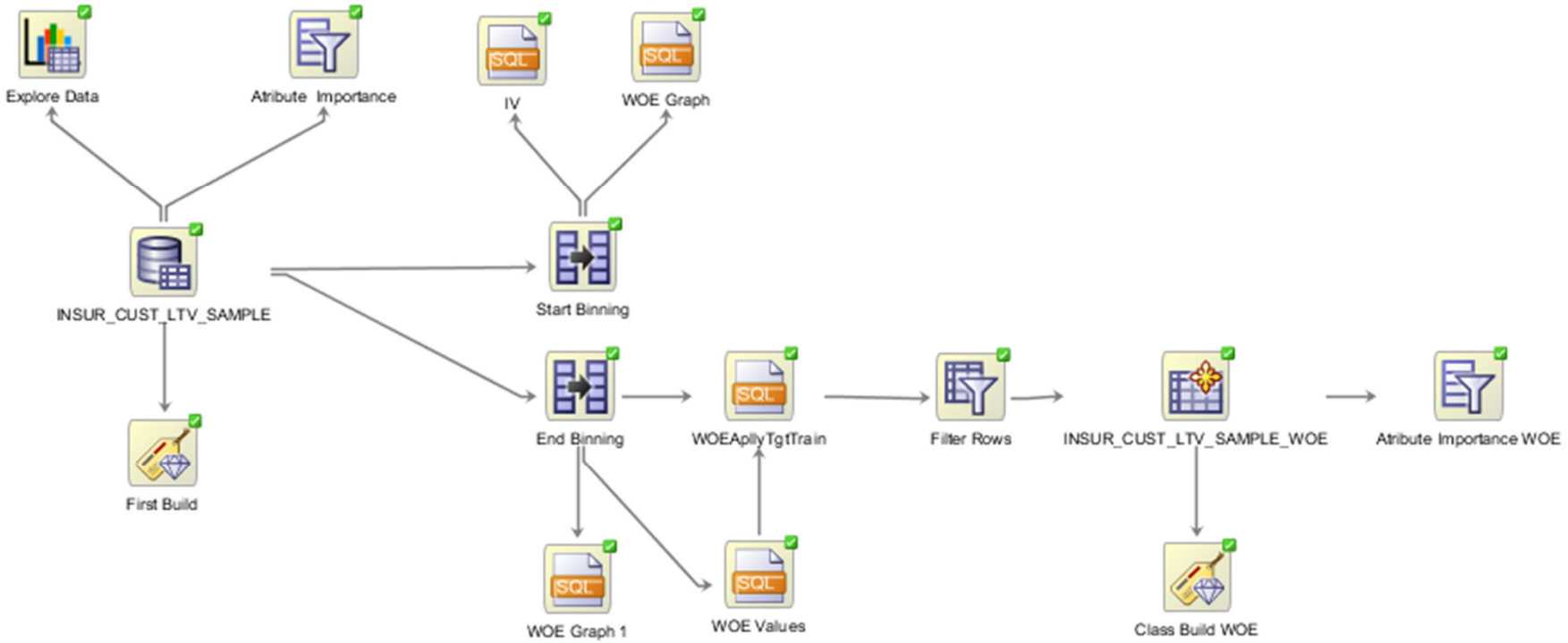
```
> grouping_column <- 'BUY_INSURANCE'
>
> ore.tableApply(dt,FUN =iv,grouping_column=grouping_column)
                                                 name          iv
AGE_BIN                                       AGE_BIN 0.08070162
BANK_FUNDS_BIN_Q                     BANK_FUNDS_BIN_Q 3.14069519
CAR_OWNERSHIP                           CAR_OWNERSHIP 0.06358286
MONEY_MONTLY_OVERDRA_BIN MONEY_MONTLY_OVERDRA_BIN 1.80768571
N_TRANS_ATM                                 N_TRANS_ATM 1.82629870
SALARY_BIN                                   SALARY_BIN 0.02009627
```

ORE i Poziv paketa 4

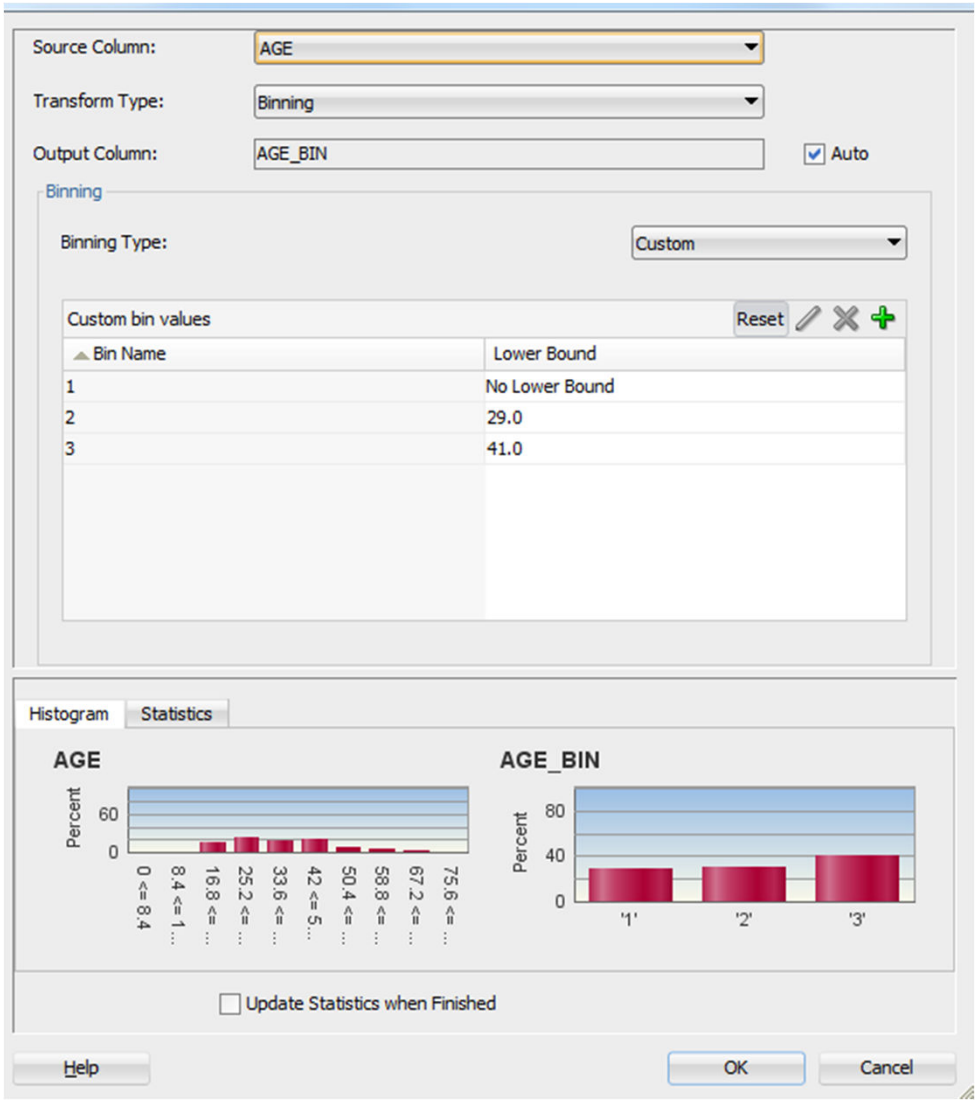
```
# 4. Način izračuna IV vrijednosti varijabli
# Kreiranje f-je na bazi i njezin poziv

library (ORE)

#Oracle User connection
if(!ore.is.connected())
  ore.connect(user=
ore.sync()


rm(list = ls(all = TRUE))

# Definiranje funkcije za izračun IV koristeći klaR paket
ore.scriptDrop("woe.iv")
ore.scriptCreate("woe.iv",
 function(x, grouping_column, ...)
{
  library(klaR)
  x[is.na(x)] <- "N/A"
  for (i in which(sapply(x, class) != "factor")) x[[i]] <- as.factor(x[[i]])
  gc <- x[[grouping_column]]
  x[[grouping_column]] <- NULL
  woemodel <- woe(x, grouping=gc, appont=FALSE, zeroadj=0.5)
  iv <- woemodel$IV
  return(data.frame(name = names(iv), iv))
}
)

#-----------------------------------------------------

dt <- INSUR_CUST_LTV_SAMPLE_BIN
grouping_column <- 'BUY_INSURANCE'

ore.tableApply(dt,FUN.NAME="woe.iv",grouping_column=grouping_column)
```

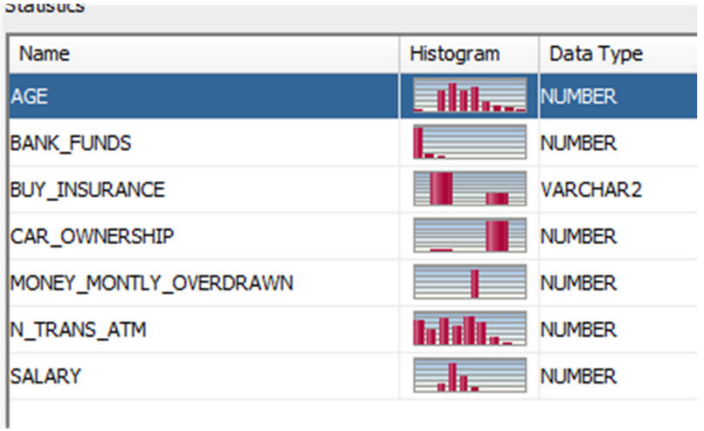
```
> dt <- INSUR_CUST_LTV_SAMPLE_BIN
> grouping_column <- 'BUY_INSURANCE'
>
> ore.tableApply(dt,FUN.NAME="iv",grouping_column=grouping_column)
                                                    name          iv
AGE_BIN                                          AGE_BIN 0.08070162
BANK_FUNDS_BIN_Q                        BANK_FUNDS_BIN_Q 3.14069519
CAR_OWNERSHIP                              CAR_OWNERSHIP 0.06358286
MONEY_MONTLY_OVERDRA_BIN MONEY_MONTLY_OVERDRA_BIN 1.80768571
N_TRANS_ATM                                  N_TRANS_ATM 1.82629870
SALARY_BIN                                    SALARY_BIN 0.02009627
>
```

ODM WF i R paketi

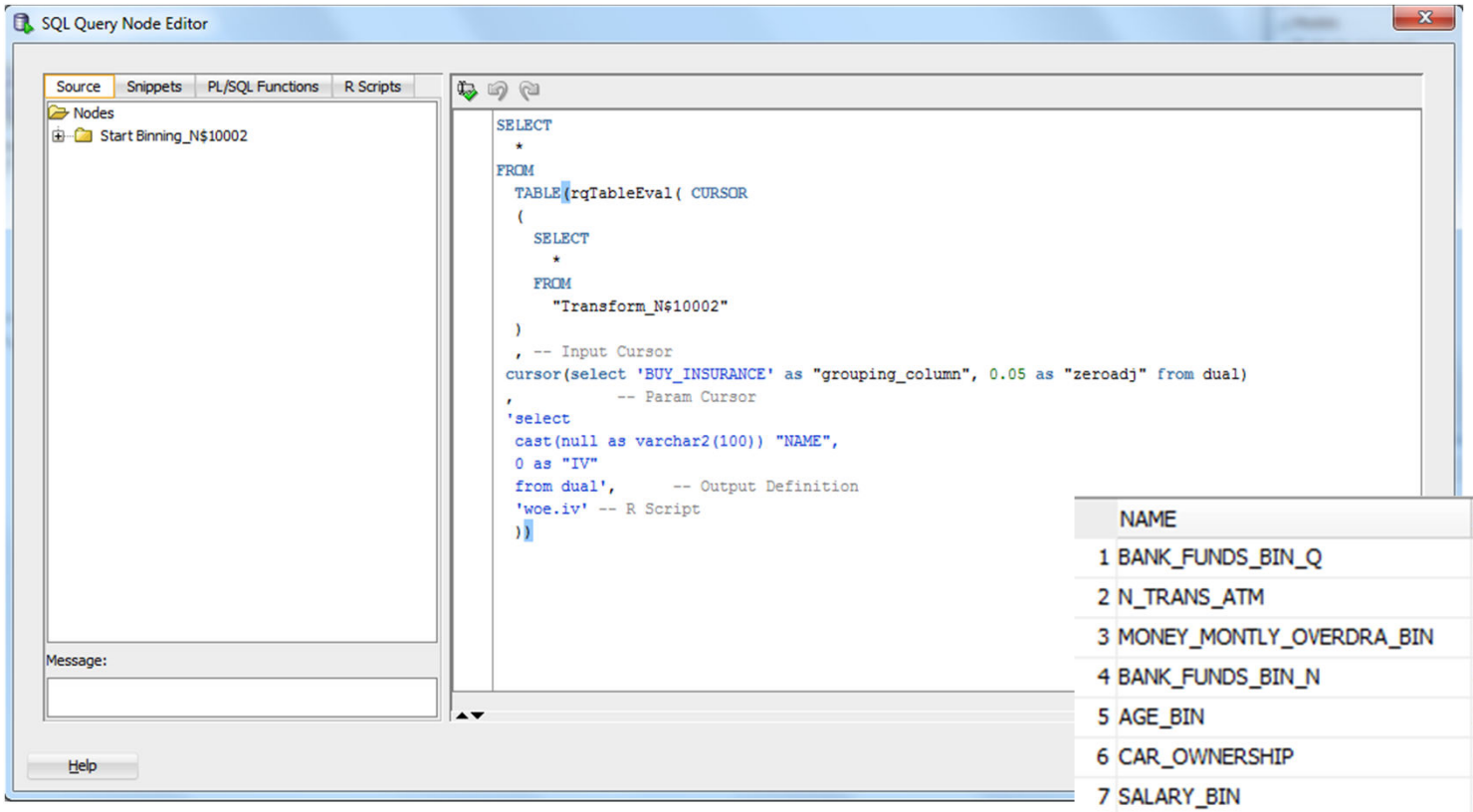


ODM WF i R paketi



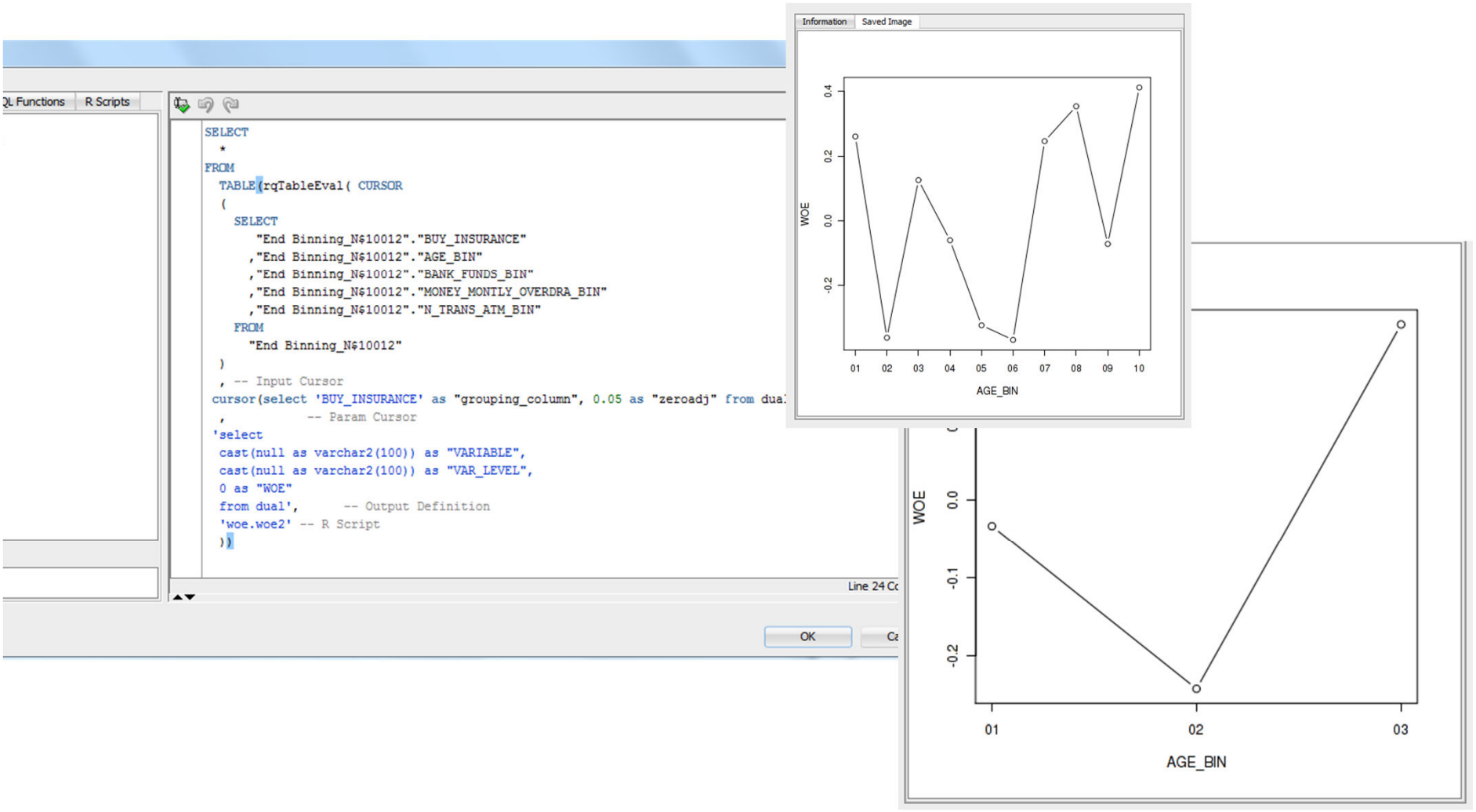


ODM WF i R paketi



	NAME	IV
1	BANK_FUNDS_BIN_Q	3,14069519
2	N_TRANS_ATM	2,375496...
3	MONEY_MONTLY_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



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
- Surival
- Xml
- Parallel
- Xts

...Hvala!

Q&A

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