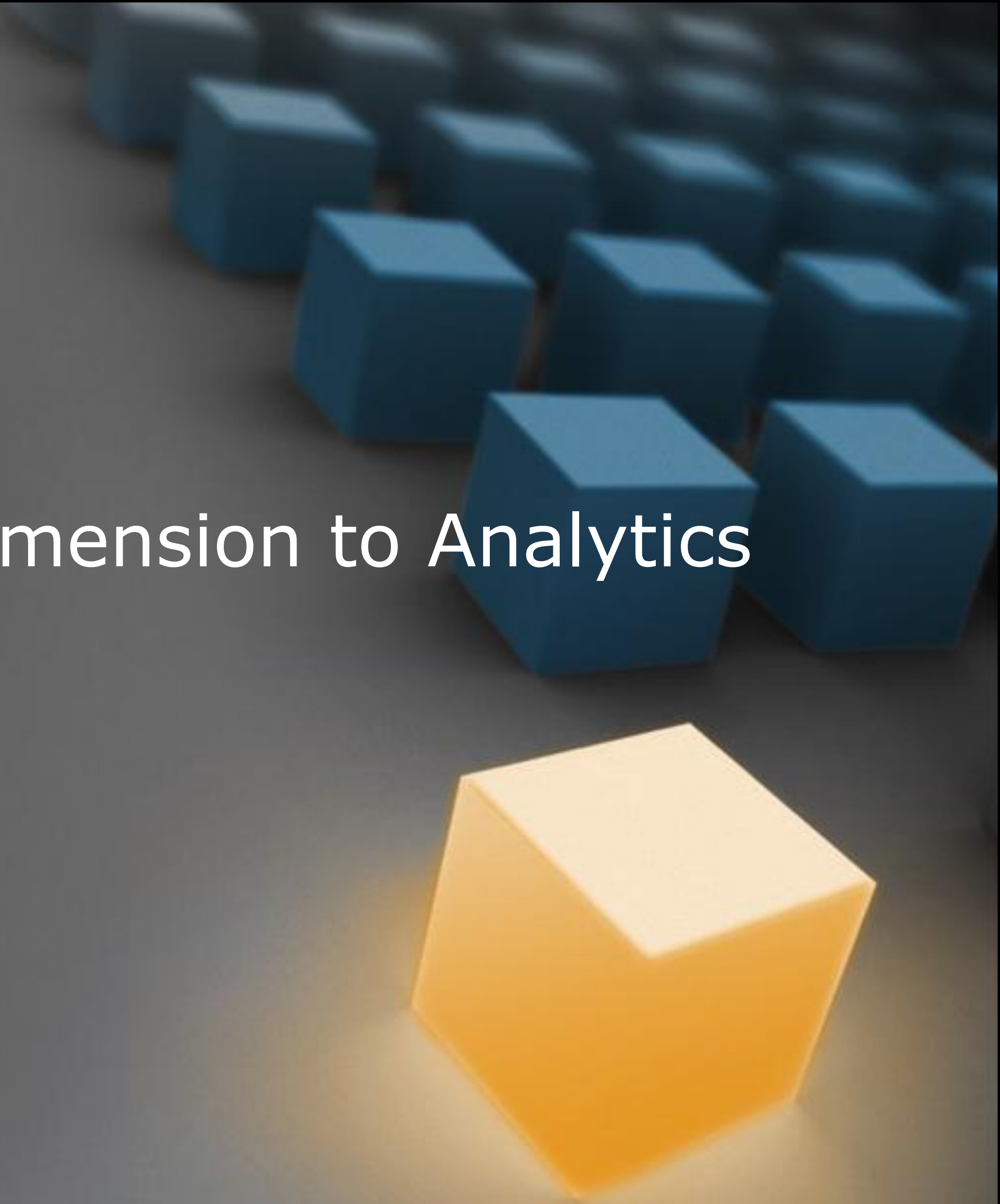


R – Adding a new dimension to Analytics

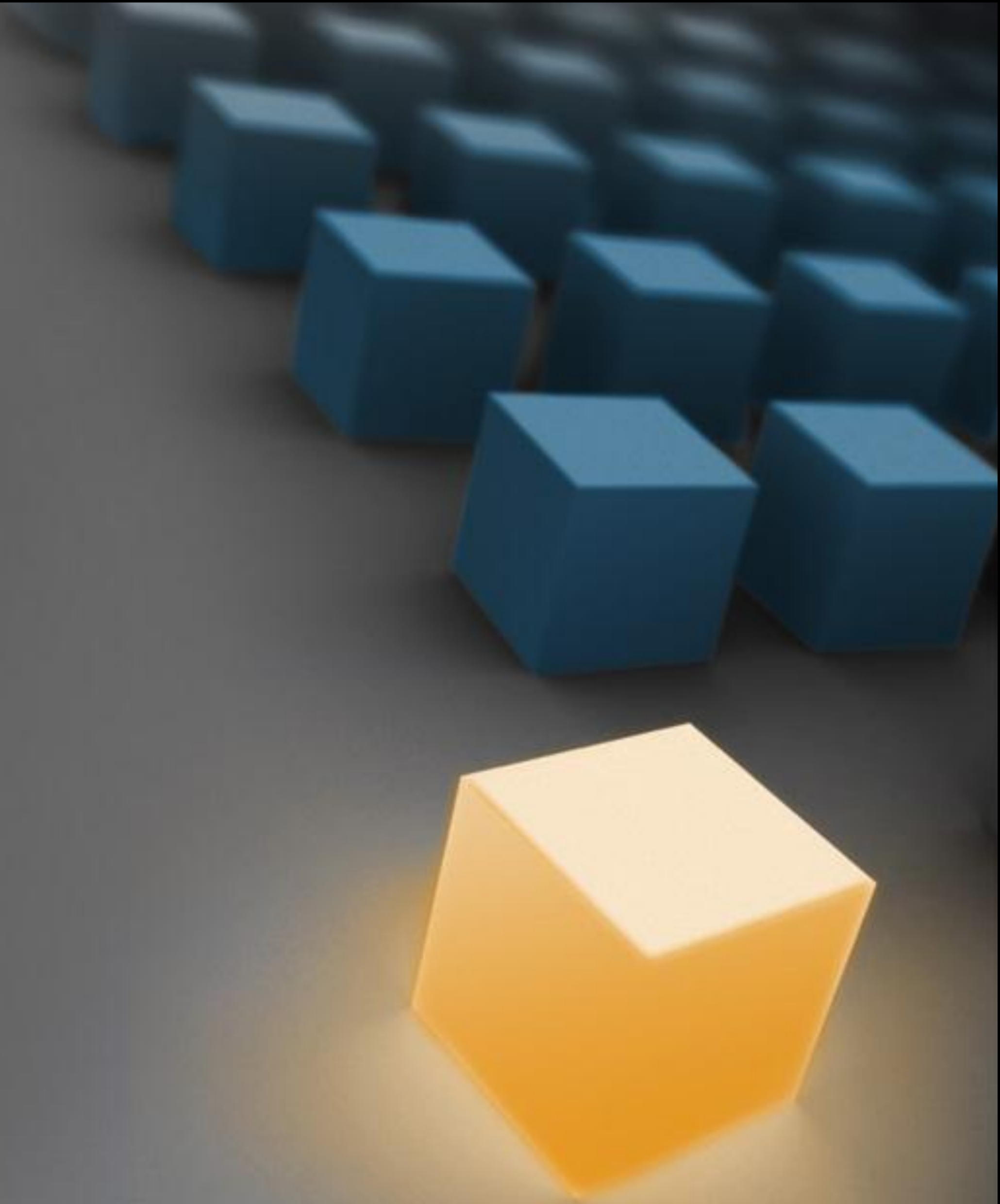
Christiane Raffeiner
Advanced Analytics Consultant
Teradata Center of Expertise

TERADATA®
Raising Intelligence

Agenda

- Introduction
- R IDEs and R GUIs
- Teradata R
- R use in DWH environment
  - Preparation of data in DWH
  - Modeling in R
  - Applying of Models in DWH
- Advanced Analytics Use Case with R

Introduction

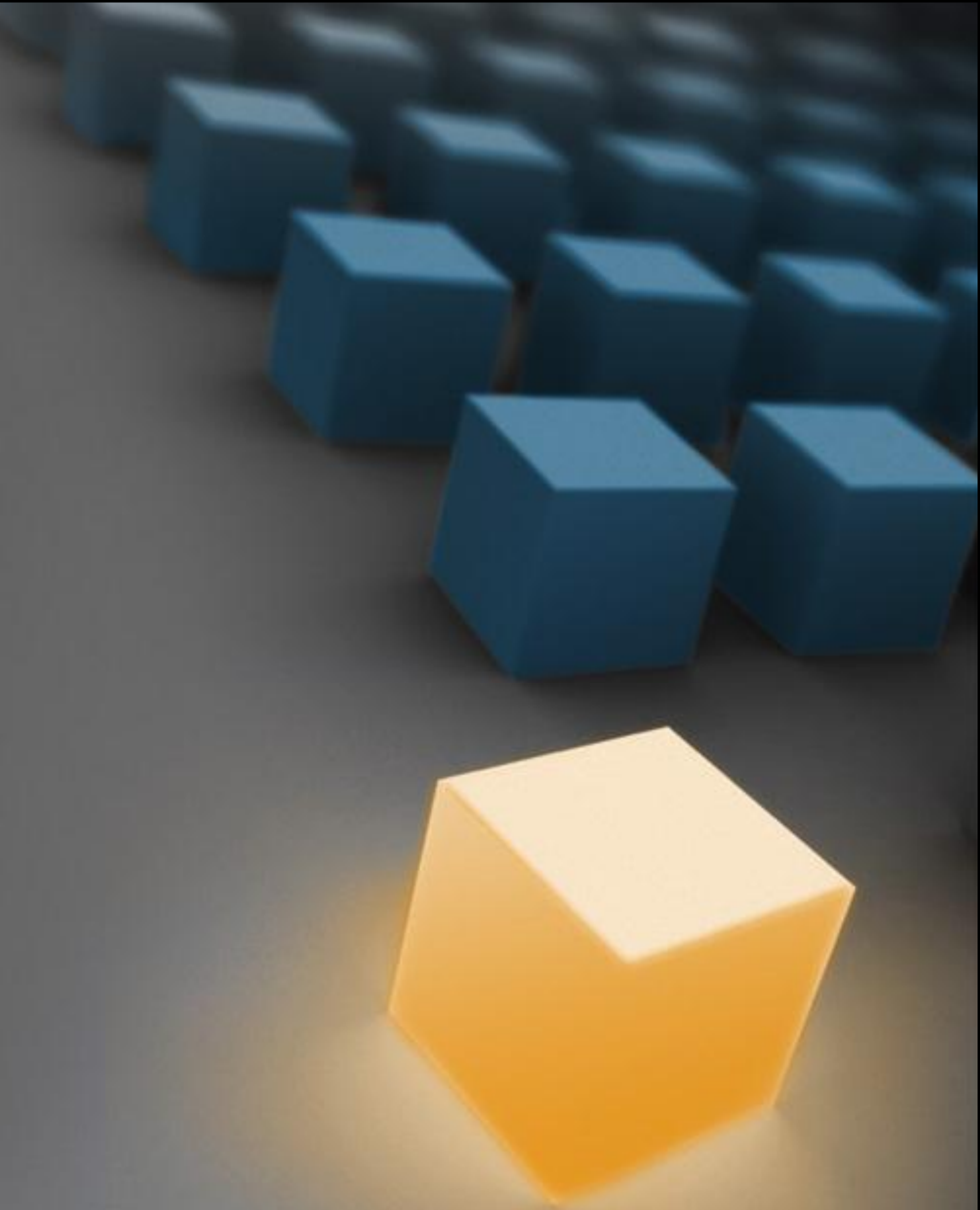


Raising Intelligence

What is R?

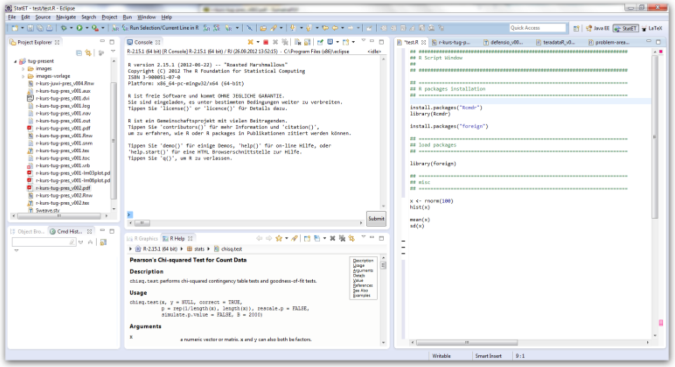
- R is a system for statistical computation and graphics
  - > Open Source Statistics Package
  - > Core functions plus 1000s of different packages
  - > Consist a programming language plus a run-time environment – R console
- R can be used
  - > interactively via expressions from the command line
  - > through support of related GUIs or Editors
  - > by writing your own functions
- Use
  - > Growing number of data analyst inside corporations & academia
  - > Ideal starting point in Pilots & Proof of Concepts
  - > R as add on to existing commercial products to extend functionality

R IDEs and R GUIs



Integrated Development Environments
(Free and open source)

- TinnR
  - > Probably one of the first IDEs for R
  - > Relatively simple (yet efficient) code editor
  - > Only available for Windows
- RStudio
  - > Comfortable to install
  - > Available for Linux, Mac, and Windows
  - > Integrated help, graphics, object browser, etc.
- Eclipse (with StatET plugin)
  - > Very powerful IDE
  - > Highly configurable, lots of functions and shortcuts
  - > Eclipse can be used for various purposes, including Teradata Access



R-2.15.1 (64 bit) [R Console] R-2.15.1 (64 bit) / RJ (26.09.2012 13:52:15) - C:\Program Files (x86)\eclipse <idle>

```
R version 2.15.1 (2012-06-22) -- "Roasted Marshmallows"
Copyright (C) 2012 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R ist freie Software und kommt OHNE JEGLICHE GARANTIE.
Sie sind eingeladen, es unter bestimmten Bedingungen weiter zu verbreiten.
Tippen Sie 'license()' or 'licence()' für Details dazu.

R ist ein Gemeinschaftsprojekt mit vielen Beitragenden.
Tippen Sie 'contributors()' für mehr Information und 'citation()',
um zu erfahren, wie R oder R packages in Publikationen zitiert werden können.

Tippen Sie 'demo()' für einige Demos, 'help()' für on-line Hilfe, oder
'help.start()' für eine HTML Browserschnittstelle zur Hilfe.
Tippen Sie 'q()', um R zu verlassen.
```

```
## ##########################################################################
## R Script Window
##
## ##########################################################################

## ==========================================================================
## R packages installation
## ==========================================================================

install.packages("Rcmdr")
library(Rcmdr)

install.packages("foreign")

## ==========================================================================
## load packages
## ==========================================================================

library(foreign)

## ==========================================================================
## misc
## ==========================================================================

x <- rnorm(100)
hist(x)

mean(x)
sd(x)
```

R-2.15.1 (64 bit) ▸ stats ▸ chisq.test

Pearson's Chi-squared Test for Count Data

Description

chisq.test performs chi-squared contingency table tests and goodness-of-fit tests.

Usage

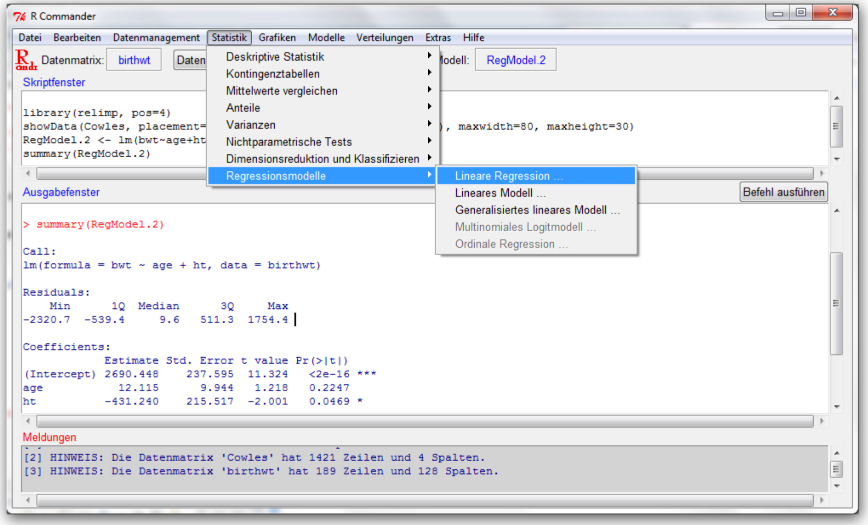
```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Arguments

x a numeric vector or matrix. x and y can also both be factors.

R GUIs
(Free and open source)

- RCOMMANDER
  - > Offers point-and-click graphics surface
  - > Contains basic analyses
  - > Good for starters



Datenmatrix: birthwt

Modell: RegModel.2

Skriptfenster

```
library(relimp, pos=4)
showData(Cowles, placement=                    ), maxwidth=80, maxheight=30)
RegModel.2 <- lm(bwt~age+ht
summary(RegModel.2)
```

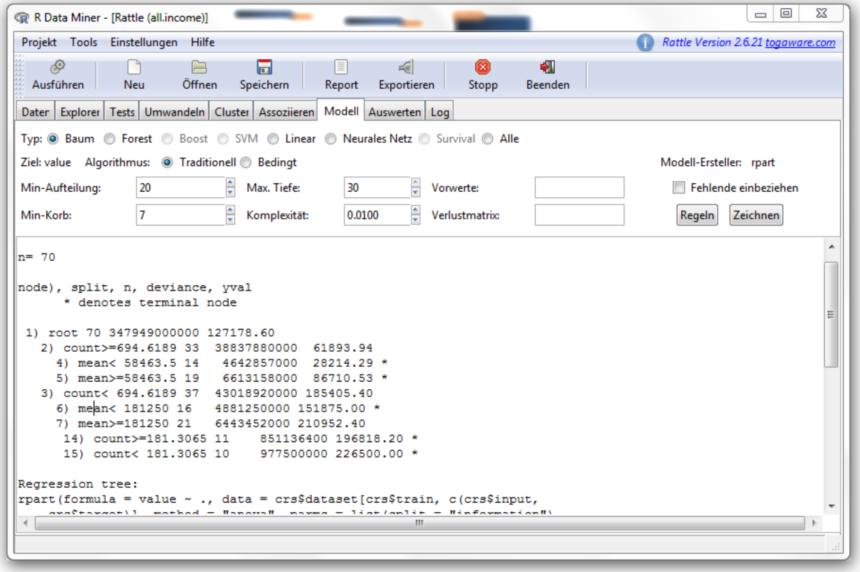
Ausgabefenster

```
> summary(RegModel.2)

Call:
lm(formula = bwt ~ age + ht, data = birthwt)

Residuals:
    Min      1Q  Median      3Q     Max
-2320.7  -539.4     9.6   511.3  1754.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2690.448    237.595  11.324   <2e-16 ***
age           12.115      9.944   1.218   0.2247
ht          -431.240    215.517  -2.001   0.0469 *
```

Meldungen

```
[2] HINWEIS: Die Datenmatrix 'Cowles' hat 1421 Zeilen und 4 Spalten.
[3] HINWEIS: Die Datenmatrix 'birthwt' hat 189 Zeilen und 128 Spalten.
```

- RATTLE
  - > Graphics User Interface
  - > For data mining
  - > Can access Teradata by ODBC



Typ: ● Baum ○ Forest ○ Boost ○ SVM ○ Linear ○ Neurales Netz ○ Survival ○ Alle

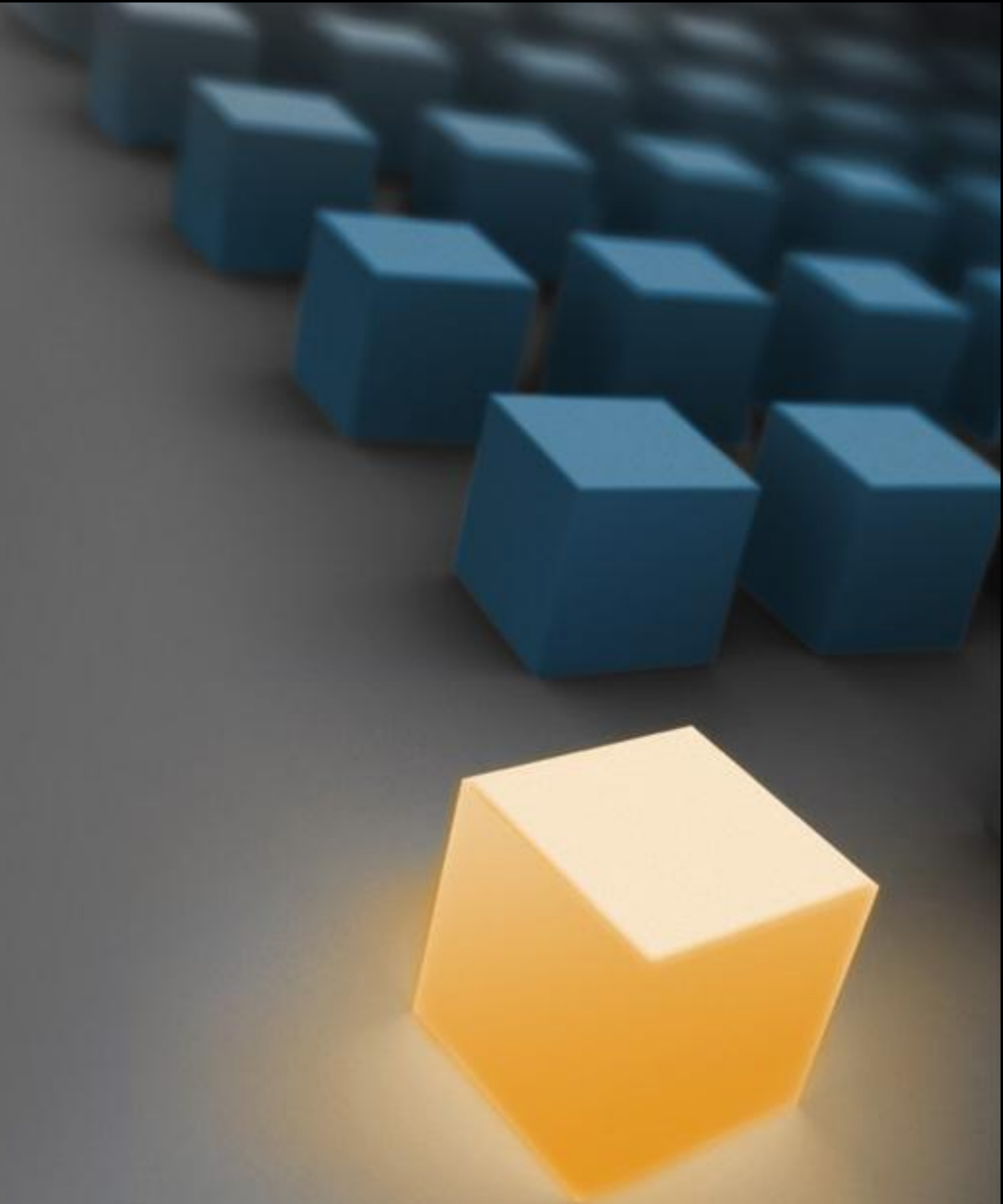
Ziel: value Algorithmus: ● Traditionell ○ Bedingt Modell-Ersteller: rpart

Min-Aufteilung: 20 Max. Tiefe: 30 Vorwerte: ☐ Fehlende einbeziehen

Min-Korb: 7 Komplexität: 0.0100 Verlustmatrix:

```
n= 70

node), split, n, deviance, yval
      * denotes terminal node

 1) root 70 347949000000 127178.60
   2) count>=694.6189 33  38837880000  61893.94
     4) mean< 58463.5 14   4642857000  28214.29 *
     5) mean>=58463.5 19   6613158000  86710.53 *
   3) count< 694.6189 37  43018920000 185405.40
     6) mean< 181250 16   4881250000 151875.00 *
     7) mean>=181250 21   6443452000 210952.40
      14) count>=181.3065 11    851136400 196818.20 *
      15) count< 181.3065 10    977500000 226500.00 *

Regression tree:
rpart(formula = value ~ ., data = crs$dataset[crs$train, c(crs$input,
    crs$target)], method = "anova", parms = list(split = "information"))
```

Teradata R



TERADATA®

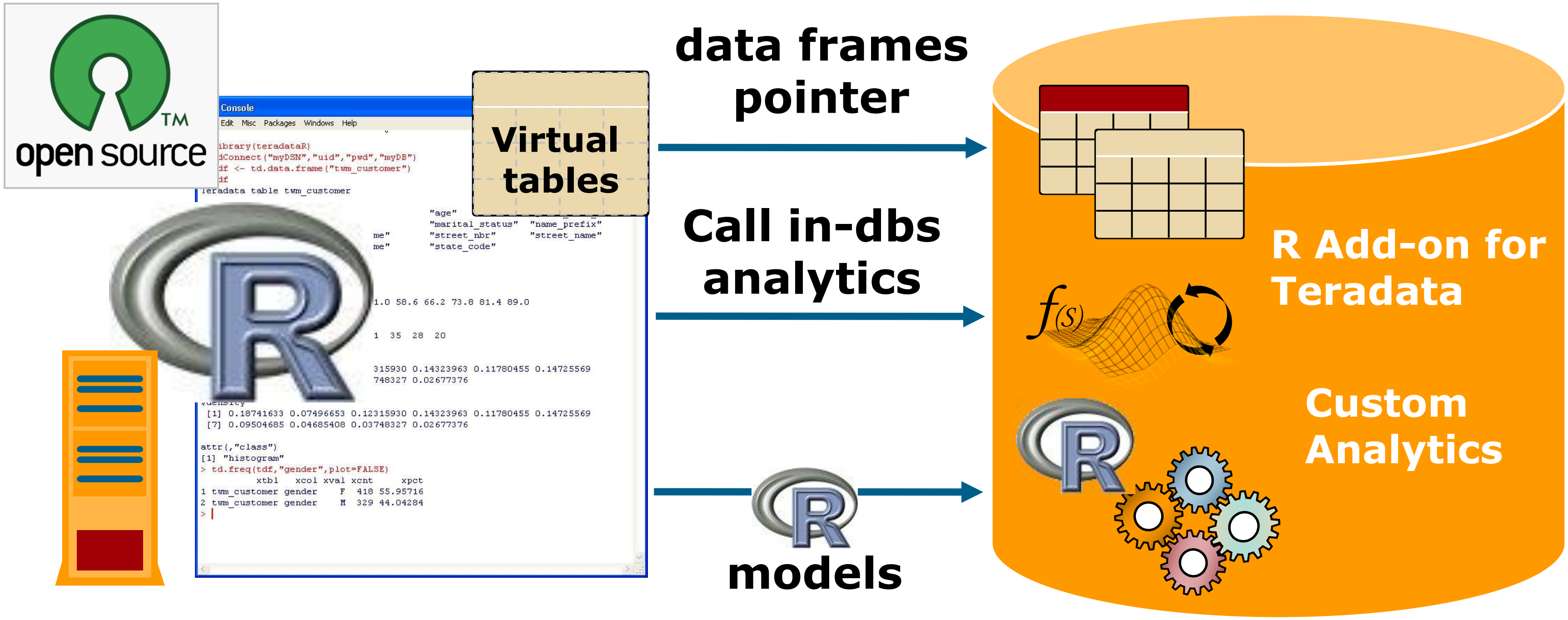
Raising Intelligence

Advanced Analytics

- Teradata in-database advanced analytics includes the following:
  - Partner optimizations with SAS, IBM SPSS Modeler, KXEN
  - Teradata Warehouse Miner
  - Emerging technology: R

- Benefits
  - Eliminate data movement to accelerate the process
  - Lifts all “big data” limitations with Teradata’s scalability
  - Leverages the parallel processing of the database

R Add-on for Teradata



```
ibrary(teradataR)
dConnect("myDSN","uid","pwd","myDB")
df <- td.data.frame("twm_customer")
df
Teradata table twm_customer

                                    "age"
                                    "marital_status"  "name_prefix"
                        me"         "street_nbr"      "street_name"
                        me"         "state_code"



                        1.0 58.6 66.2 73.8 81.4 89.0


                        1  35  28  20


                        315930 0.14323963 0.11780455 0.14725569
                        748327 0.02677376

$density
 [1] 0.18741633 0.07496653 0.12315930 0.14323963 0.11780455 0.14725569
 [7] 0.09504685 0.04685408 0.03748327 0.02677376

attr(,"class")
[1] "histogram"
> td.freq(tdf,"gender",plot=FALSE)
          xtbl   xcol xval xcnt     xpct
1 twm_customer gender    F  418 55.95716
2 twm_customer gender    M  329 44.04284
> 
```

R Package *teradataR*

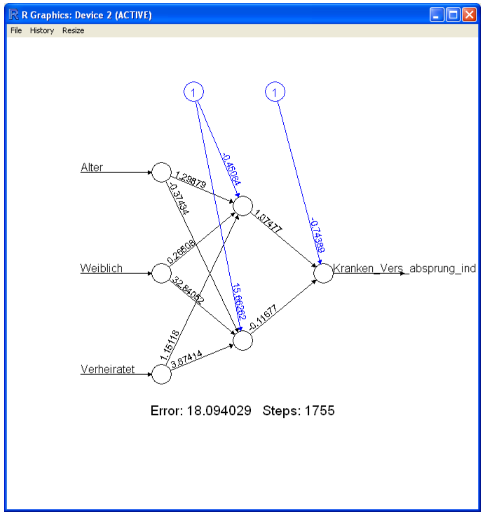
- Simplifies connection to Teradata
- Establishes a data frame pointer (virtual table) to Teradata tables
- Provides over 40 in-database analytical functions
- Custom Analytics (udfs) created are callable by R
- R models can run in-database via PMML

R Use in DWH Environment

TERADATA

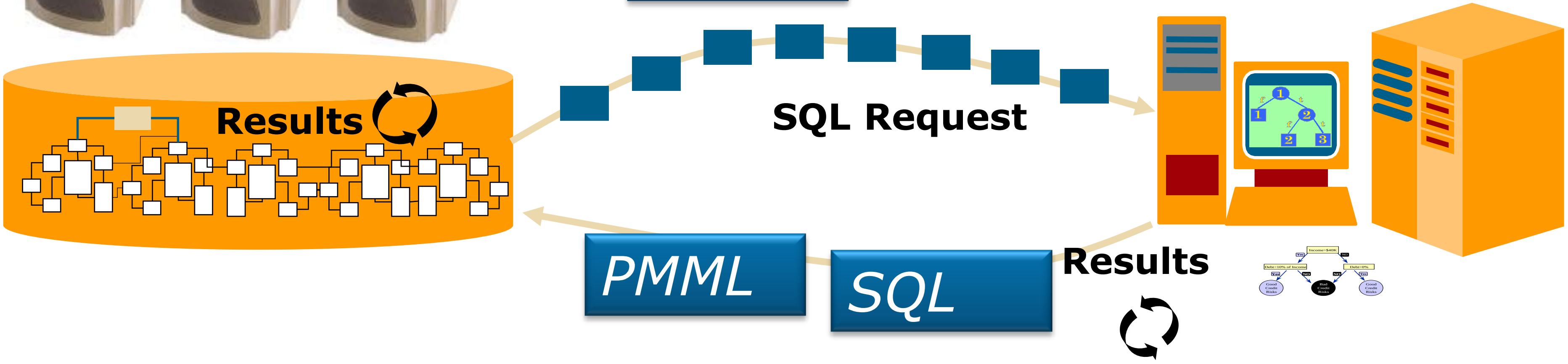
Raising Intelligence

R and Teradata Architecture



Error: 18.094029 Steps: 1755

ODBC



TERADATA Raising Intelligence

SQL – RODBC package

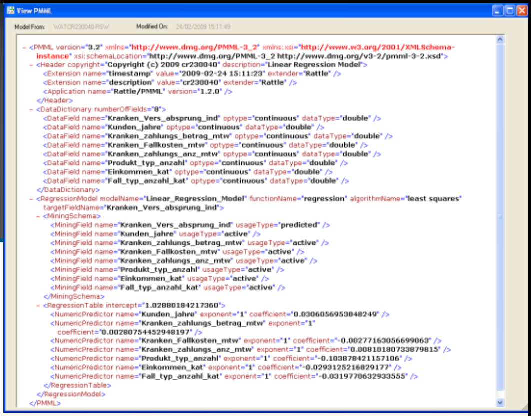
- RODBC package implements ODBC database connectivity
- RODBC allows direct access to tables stored in the DWH

- Functions provided
  - Internal odbc* commands
  - Sql* functions to read, save, copy and manipulate data

- SQL* arguments can be used within R and are passed through to the DWH
  - sqlSave
  - sqlCopy
  - sqlQuery
  - sqlFetch

R Package *RODBC*

PMML – an open data format



Model From: WATCR230040.RSW Modified On: 24/02/2009 15:11:49

```
- <PMML version="3.2" xmlns="http://www.dmg.org/PMML-3_2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
  instance" xsi:schemaLocation="http://www.dmg.org/PMML-3_2 http://www.dmg.org/v3-2/pmml-3-2.xsd">
 - <Header copyright="Copyright (c) 2009 cr230040" description="Linear Regression Model">
    <Extension name="timestamp" value="2009-02-24 15:11:23" extender="Rattle" />
    <Extension name="description" value="cr230040" extender="Rattle" />
    <Application name="Rattle/PMML" version="1.2.0" />
   </Header>
 - <DataDictionary numberOfFields="8">
    <DataField name="Kranken_Vers_absprung_ind" optype="continuous" dataType="double" />
    <DataField name="Kunden_jahre" optype="continuous" dataType="double" />
    <DataField name="Kranken_zahlungs_betrag_mtw" optype="continuous" dataType="double" />
    <DataField name="Kranken_Fallkosten_mtw" optype="continuous" dataType="double" />
    <DataField name="Kranken_zahlungs_anz_mtw" optype="continuous" dataType="double" />
    <DataField name="Produkt_typ_anzahl" optype="continuous" dataType="double" />
    <DataField name="Einkommen_kat" optype="continuous" dataType="double" />
    <DataField name="Fall_typ_anzahl_kat" optype="continuous" dataType="double" />
   </DataDictionary>
 - <RegressionModel modelName="Linear_Regression_Model" functionName="regression" algorithmName="least squares"
    targetFieldName="Kranken_Vers_absprung_ind">
  - <MiningSchema>
     <MiningField name="Kranken_Vers_absprung_ind" usageType="predicted" />
     <MiningField name="Kunden_jahre" usageType="active" />
     <MiningField name="Kranken_zahlungs_betrag_mtw" usageType="active" />
     <MiningField name="Kranken_Fallkosten_mtw" usageType="active" />
     <MiningField name="Kranken_zahlungs_anz_mtw" usageType="active" />
     <MiningField name="Produkt_typ_anzahl" usageType="active" />
     <MiningField name="Einkommen_kat" usageType="active" />
     <MiningField name="Fall_typ_anzahl_kat" usageType="active" />
    </MiningSchema>
  - <RegressionTable intercept="1.02880184217360">
     <NumericPredictor name="Kunden_jahre" exponent="1" coefficient="0.0306056953848249" />
     <NumericPredictor name="Kranken_zahlungs_betrag_mtw" exponent="1"
       coefficient="0.00280754452948197" />
     <NumericPredictor name="Kranken_Fallkosten_mtw" exponent="1" coefficient="-0.00277163056699063" />
     <NumericPredictor name="Kranken_zahlungs_anz_mtw" exponent="1" coefficient="0.00810180733879815" />
     <NumericPredictor name="Produkt_typ_anzahl" exponent="1" coefficient="-0.103878421157106" />
     <NumericPredictor name="Einkommen_kat" exponent="1" coefficient="-0.0293125216829177" />
     <NumericPredictor name="Fall_typ_anzahl_kat" exponent="1" coefficient="-0.0319770632933555" />
    </RegressionTable>
   </RegressionModel>
  </PMML>
```

- Predictive Model Markup Language (PMML)
  - is an XML based standard to describe statistical and data mining models
  - since 1998 developed by the Data Mining Group (www.dmg.org)
  - describes the data inputs to data mining models, some algorithm specific transformations used to prepare data for data mining, and the parameters which define the models themselves
- Application-independent method of defining models
  - Users can develop models within one vendor's application, and use the model in any other application supporting PMML
  - Requires a PMML consumer to read PMML output from a PMML producer and create an executable form of the model for scoring purposes

R Package *pmml*

TERADATA
Raising Intelligence

Advanced Analytics Use Case

Raising Intelligence

RETAIL Price Optimization

TERADATA

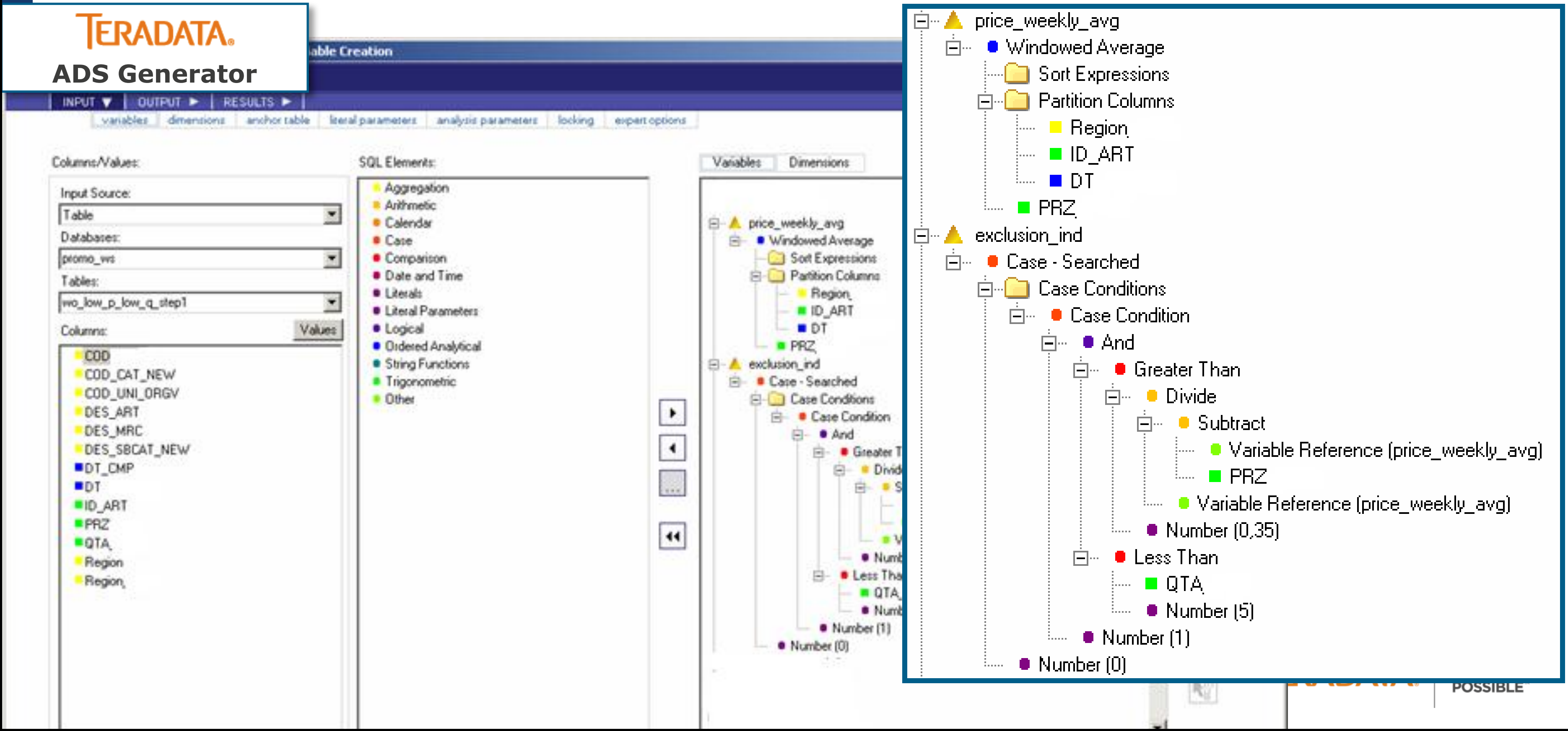
Raising Intelligence

Pilot Objectives

- Main Goals:
  - Understand how to define promotional sales prices
  - Simulation of price changes and their impact on sales quantity using a predefined set of stores and chosen articles of predefined categories
  - Estimation of the products price elasticity and cross price elasticity between products of the chosen category:
    - Price Elasticity, Cross Selling & Cannibalization

Step 1: Data Preparation

- Create Analytical Data Set for Modeling in R
  - > Combination of very close price points
- Data Filtering
  - > Articles/stores with minimum number of observations and minimum price spread
  - > Elimination of Outliers using descriptive statistics



TERADATA

ADS Generator

...able Creation
INPUT ▼ | OUTPUT ► | RESULTS ►
variables | dimensions | anchor table | literal parameters | analysis parameters | locking | expert options
Values
Variables | Dimensions

Columns/Values:

Input Source: Table

Databases: promo_ws

Tables: wo_low_p_low_q_step1

Columns:

- COD
- COD_CAT_NEW
- COD_UNI_ORGV
- DES_ART
- DES_MRC
- DES_SBCAT_NEW
- DT_CMP
- DT
- ID_ART
- PRZ
- QTA
- Region
- Region

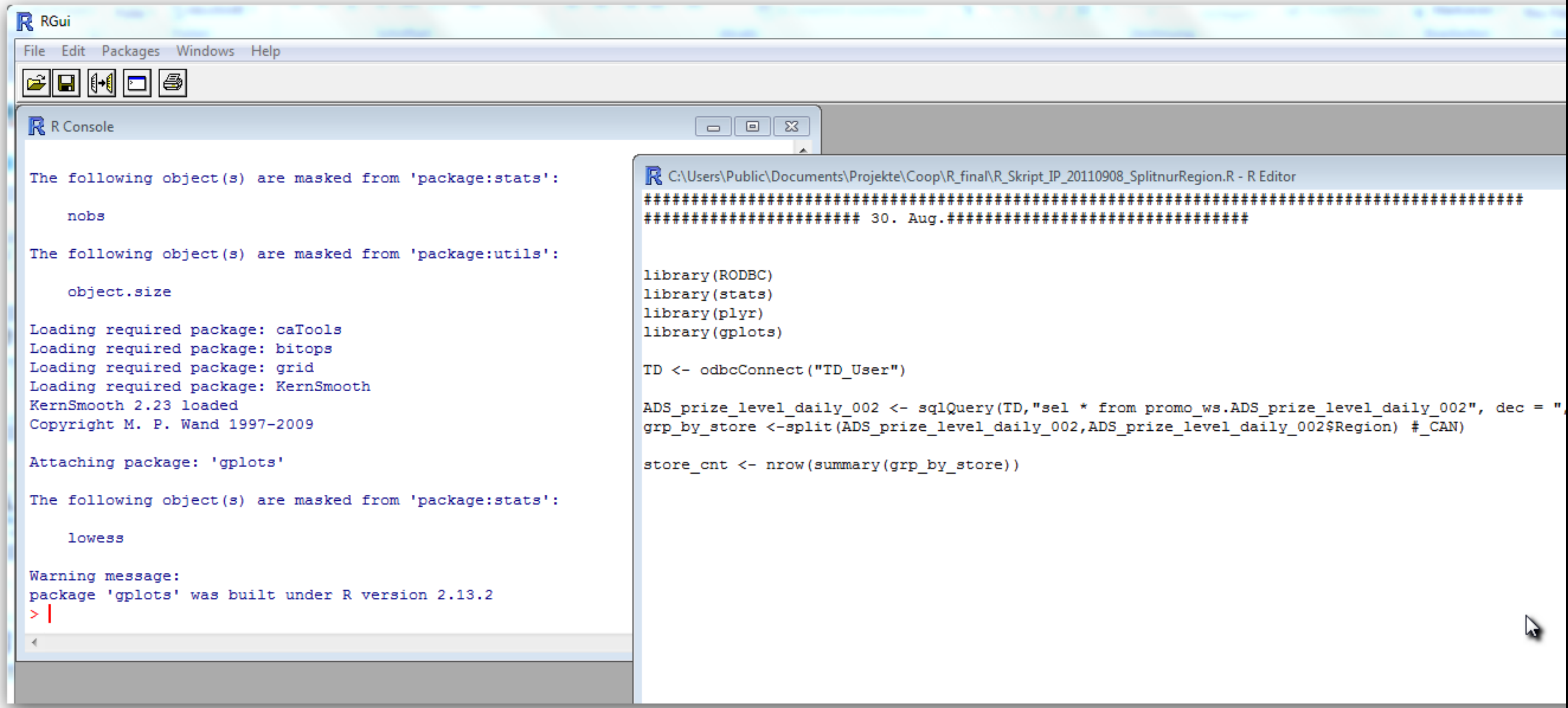
SQL Elements:

- Aggregation
- Arithmetic
- Calendar
- Case
- Comparison
- Date and Time
- Literals
- Literal Parameters
- Logical
- Ordered Analytical
- String Functions
- Trigonometric
- Other

```
price_weekly_avg
    Windowed Average
        Sort Expressions
        Partition Columns
            Region
            ID_ART
            DT
        PRZ
exclusion_ind
    Case - Searched
        Case Conditions
            Case Condition
                And
                    Greater Than
                        Divide
                            Subtract
                                Variable Reference (price_weekly_avg)
                                PRZ
                            Variable Reference (price_weekly_avg)
                        Number (0,35)
                    Less Than
                        QTA
                        Number (5)
                Number (1)
        Number (0)
```

Step 2: ODBC connection to Teradata database

- R libraries used:
  - „RODBC“ for ODBC connection to TD
  - „stats“ for regression
  - „gplot“ for plotting results



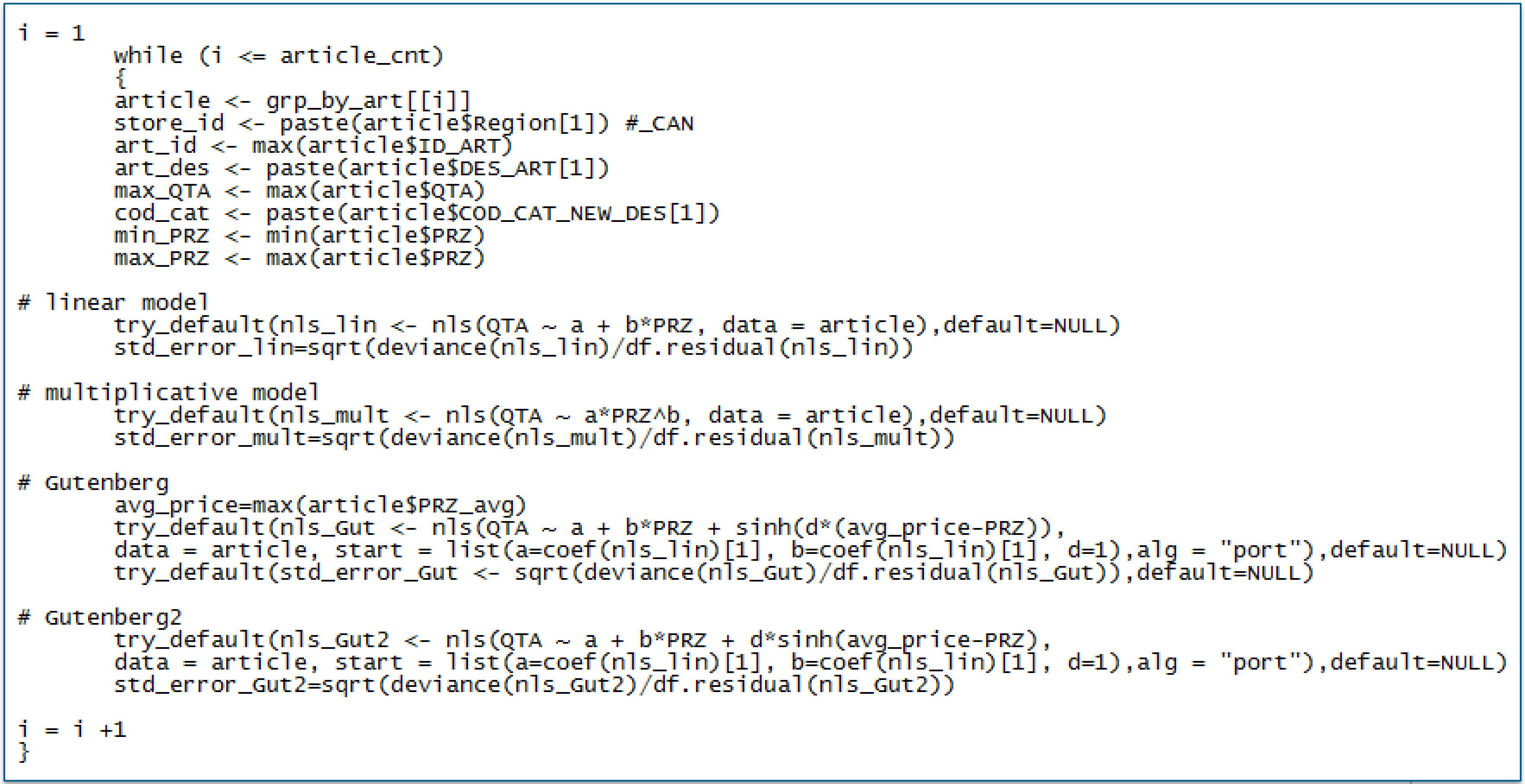
RGui
File Edit Packages Windows Help
R Console
C:\Users\Public\Documents\Projekte\Coop\R_final\R_Skript_IP_20110908_SplitnurRegion.R - R Editor

```
The following object(s) are masked from 'package:stats':

    nobs

The following object(s) are masked from 'package:utils':

    object.size

Loading required package: caTools
Loading required package: bitops
Loading required package: grid
Loading required package: KernSmooth
KernSmooth 2.23 loaded
Copyright M. P. Wand 1997-2009

Attaching package: 'gplots'

The following object(s) are masked from 'package:stats':

    lowess

Warning message:
package 'gplots' was built under R version 2.13.2
> 
```

```
###############################################################################################
####################### 30. Aug.#################################


library(RODBC)
library(stats)
library(plyr)
library(gplots)

TD <- odbcConnect("TD_User")

ADS_prize_level_daily_002 <- sqlQuery(TD,"sel * from promo_ws.ADS_prize_level_daily_002", dec = ",
grp_by_store <-split(ADS_prize_level_daily_002,ADS_prize_level_daily_002$Region) #_CAN)

store_cnt <- nrow(summary(grp_by_store))
```

Step 3: Price Elasticity Models with R

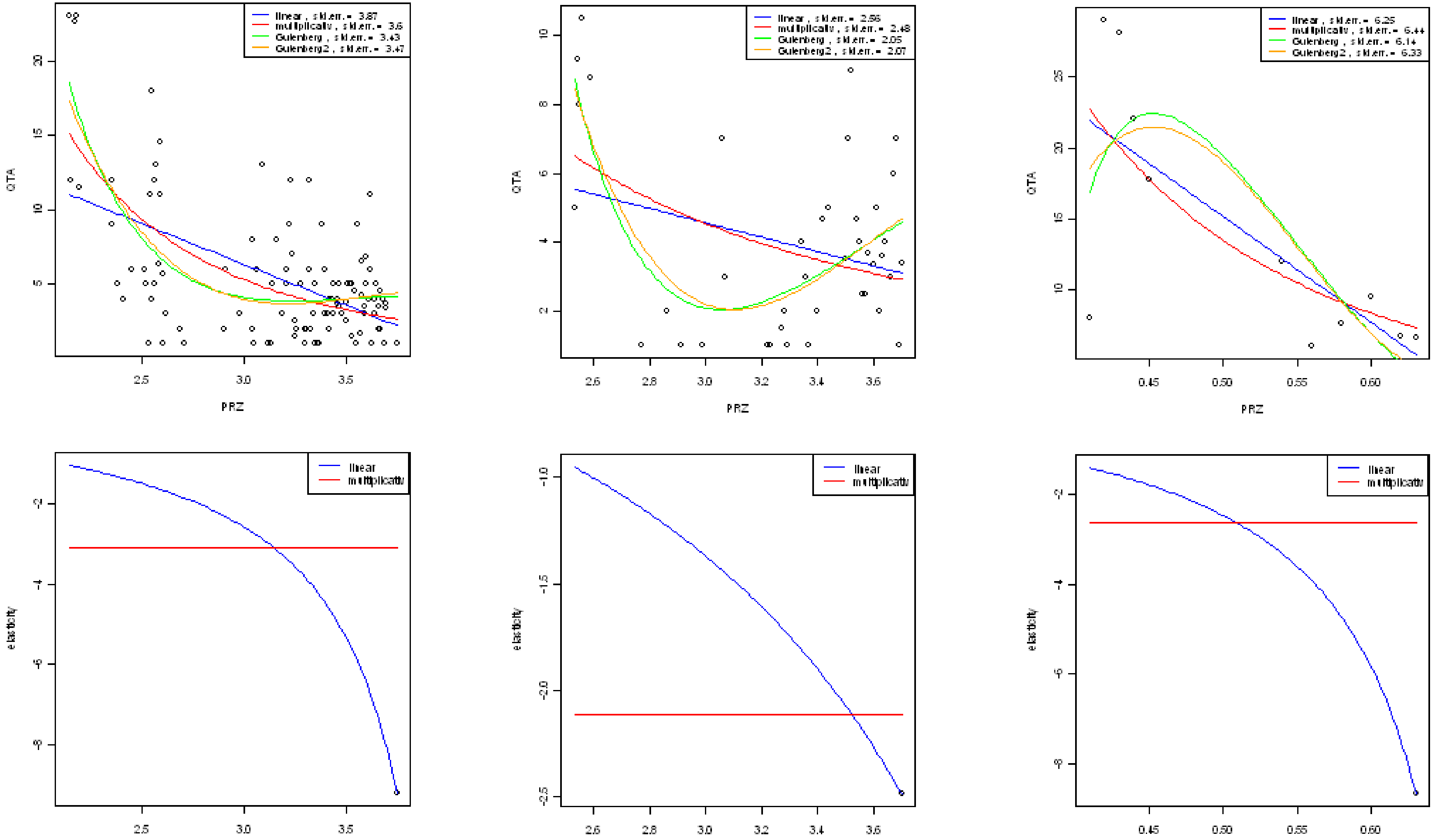
- Four different models created per article/region:
  - Linear model
  - Multiplicative model
  - Two Gutenberg-models



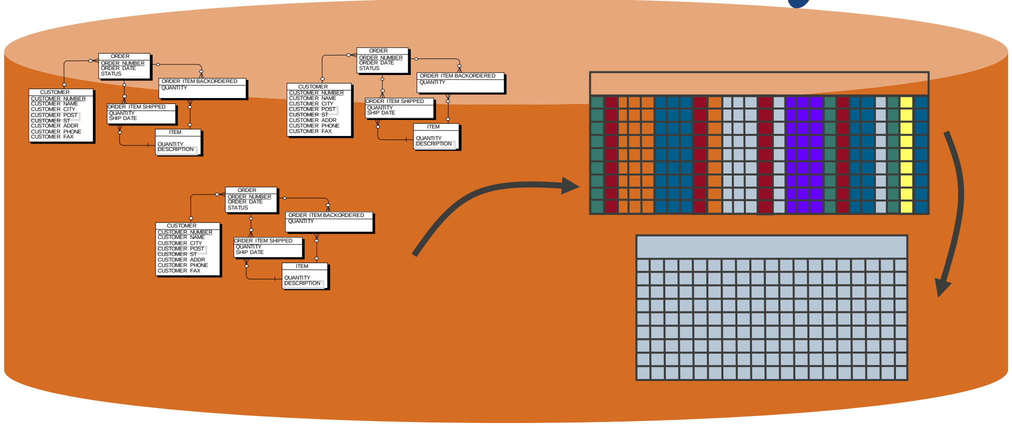
```
i = 1
       while (i <= article_cnt)
       {
       article <- grp_by_art[[i]]
       store_id <- paste(article$Region[1]) #_CAN
       art_id <- max(article$ID_ART)
       art_des <- paste(article$DES_ART[1])
       max_QTA <- max(article$QTA)
       cod_cat <- paste(article$COD_CAT_NEW_DES[1])
       min_PRZ <- min(article$PRZ)
       max_PRZ <- max(article$PRZ)

# linear model
       try_default(nls_lin <- nls(QTA ~ a + b*PRZ, data = article),default=NULL)
       std_error_lin=sqrt(deviance(nls_lin)/df.residual(nls_lin))

# multiplicative model
       try_default(nls_mult <- nls(QTA ~ a*PRZ^b, data = article),default=NULL)
       std_error_mult=sqrt(deviance(nls_mult)/df.residual(nls_mult))

# Gutenberg
       avg_price=max(article$PRZ_avg)
       try_default(nls_Gut <- nls(QTA ~ a + b*PRZ + sinh(d*(avg_price-PRZ)),
       data = article, start = list(a=coef(nls_lin)[1], b=coef(nls_lin)[1], d=1),alg = "port"),default=NULL)
       try_default(std_error_Gut <- sqrt(deviance(nls_Gut)/df.residual(nls_Gut)),default=NULL)

# Gutenberg2
       try_default(nls_Gut2 <- nls(QTA ~ a + b*PRZ + d*sinh(avg_price-PRZ),
       data = article, start = list(a=coef(nls_lin)[1], b=coef(nls_lin)[1], d=1),alg = "port"),default=NULL)
       std_error_Gut2=sqrt(deviance(nls_Gut2)/df.residual(nls_Gut2))

i = i +1
}
```

Step 4: Model Evaluation - R plots

- Developed R-script to build price elasticity models automatically for all articles and plot results
- Results are automatically plotted and saved for evaluation



Overall Analytical Process
Project Steps

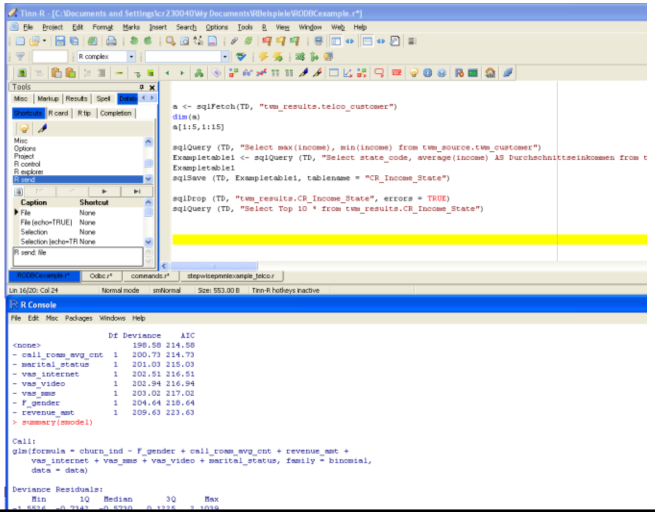


1. Derive, Aggregate, Transform to create Analytical Data Set (ADS)

2. Access ADS from R via RODBC

3. Create Model in R

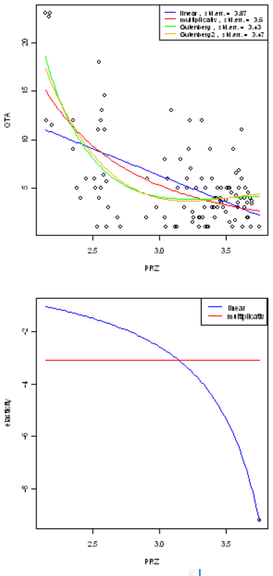
4. Evaluate Model



Tinn-R - [C:\Documents and Settings\cr230040\My Documents\RBeispiele\RODBCexample.r*]
File Project Edit Format Marks Insert Search Options Tools R View Window Web Help
R complex
Tools
Misc Markup Results Spell Datab
Shortcuts R card R tip Completion
Misc
Options
Project
R control
R explorer
R send
Caption Shortcut
File None
File (echo=TRUE) None
Selection None
Selection (echo=TR None
R send: file
RODBCexample.r* Odbc.r* commands.r* stepwisepmmlexample_telco.r
Lin 16/20: Col 24 Normal mode smNormal Size: 553.00 B Tinn-R hotkeys inactive
R Console
File Edit Misc Packages Windows Help

```
a <- sqlFetch(TD, "twm_results.telco_customer")
dim(a)
a[1:5,1:15]

sqlQuery (TD, "Select max(income), min(income) from twm_source.twm_customer")
Exampletable1 <- sqlQuery (TD, "Select state_code, average(income) AS Durchschnittseinkommen from t
Exampletable1
sqlSave (TD, Exampletable1, tablename = "CR_Income_State")

sqlDrop (TD, "twm_results.CR_Income_State", errors = TRUE)
sqlQuery (TD, "Select Top 10 * from twm_results.CR_Income_State")
```

```
                    Df Deviance    AIC
<none>                   198.58 214.58
- call_roam_avg_cnt  1   200.73 214.73
- marital_status     1   201.03 215.03
- vas_internet       1   202.51 216.51
- vas_video          1   202.94 216.94
- vas_mms            1   203.02 217.02
- F_gender           1   204.64 218.64
- revenue_amt        1   209.63 223.63
> summary(smodel)

Call:
glm(formula = churn_ind ~ F_gender + call_roam_avg_cnt + revenue_amt +
    vas_internet + vas_mms + vas_video + marital_status, family = binomial,
    data = data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5526  -0.7342  -0.5730   0.1225   2.1039
```

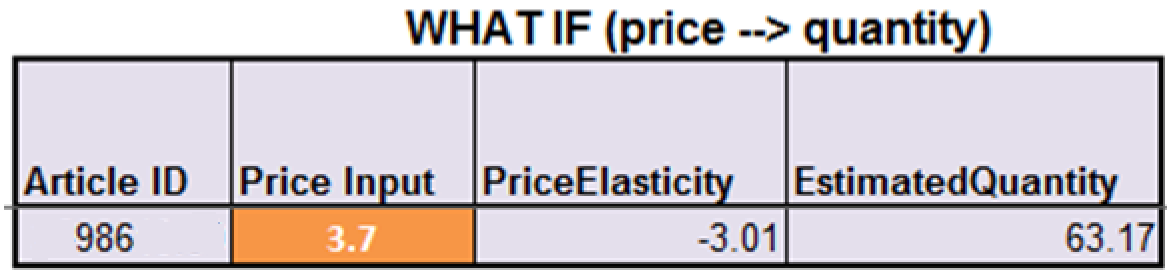


5. Matrix with regression results and coefficients is written back to TD database

6. Results can now be accessed by Front End via ODBC connection

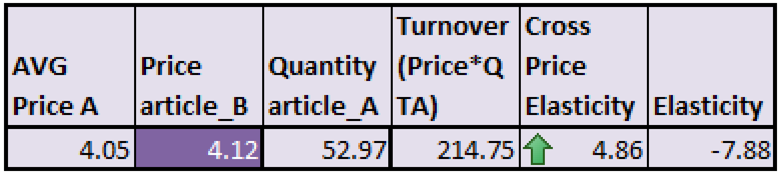
TERADATA. THE BEST DECISION POSSIBLE™

Using Results – What if Analysis

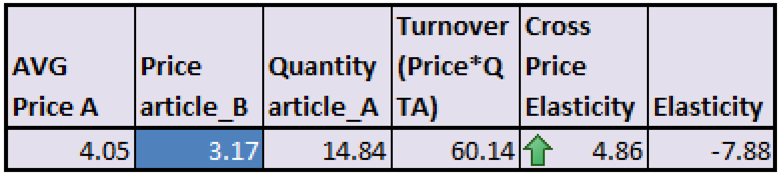


WHAT IF (price --> quantity)

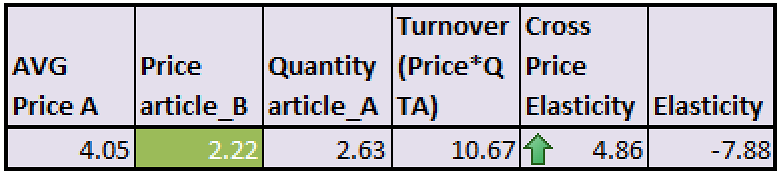
Article ID	Price Input	PriceElasticity	EstimatedQuantity
986	3.7	-3.01	63.17



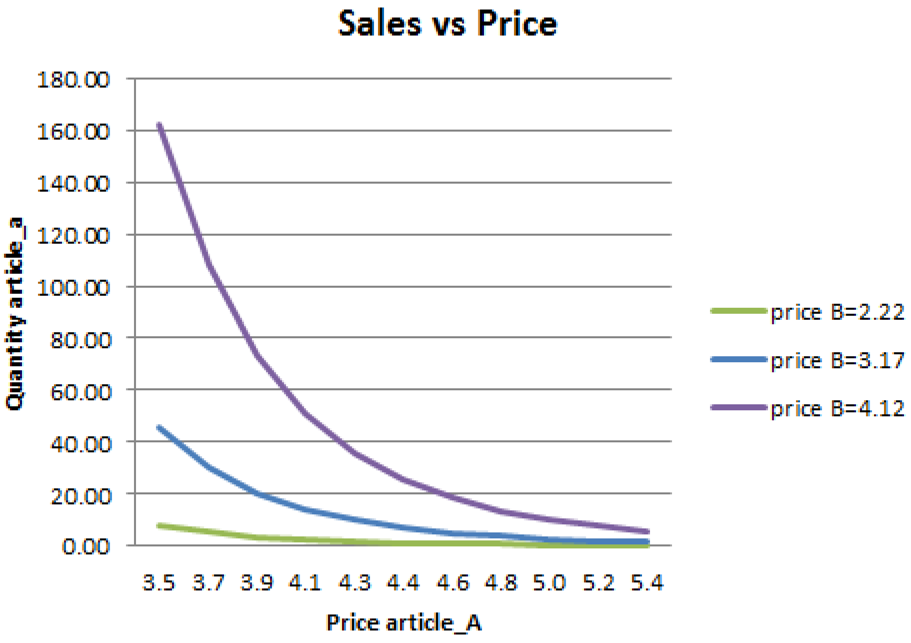
			Turnover	Cross	
AVG	Price	Quantity	(Price*Q	Price	
Price A	article_B	article_A	TA)	Elasticity	Elasticity
4.05	4.12	52.97	214.75	⇧ 4.86	-7.88



			Turnover	Cross	
AVG	Price	Quantity	(Price*Q	Price	
Price A	article_B	article_A	TA)	Elasticity	Elasticity
4.05	3.17	14.84	60.14	⇧ 4.86	-7.88



			Turnover	Cross	
AVG	Price	Quantity	(Price*Q	Price	
Price A	article_B	article_A	TA)	Elasticity	Elasticity
4.05	2.22	2.63	10.67	⇧ 4.86	-7.88



TERADATA. THE BEST DECISION POSSIBLE™

Summary

- Advantages
  - Open source - free
  - Choice of thousands of analytical functions
  - Flexibility and extendability
  - Possibility to go beyond commercial products
  - Use R as add-on or single data mining tool
  - Easy fit into existing working processes
  - R can interact with other programs easily
  - Expert Interaction and Support through Forums
  - Fast implementation of new algorithms due to active community

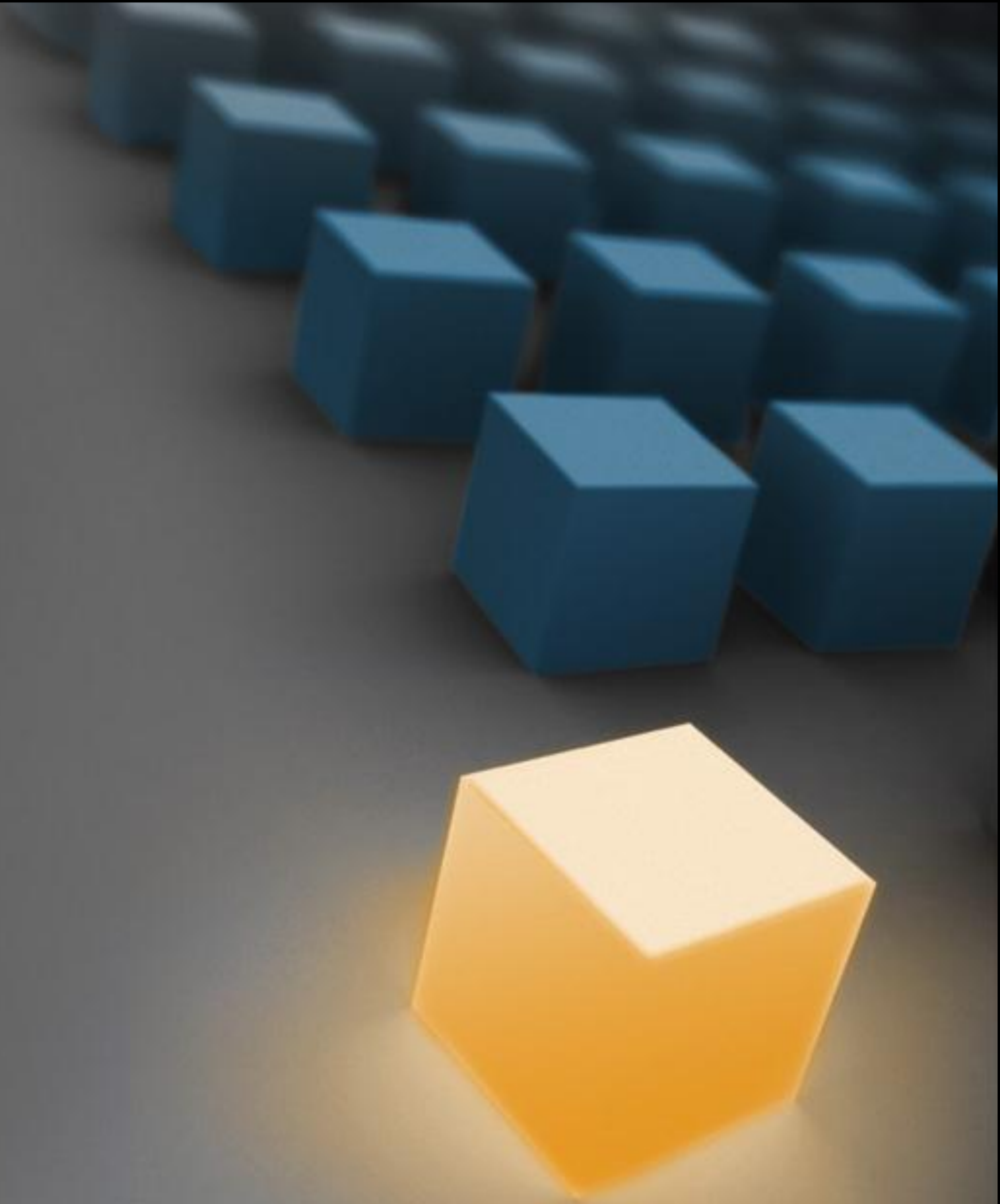
- Challenges
  - Unsupported tools need strong solution partnerships
  - Code based work maintenace requires well defined processes
  - Buy in traditional GUI users and grow R skills
  - Sound statistical knowledge important for exploiting full functionality
  - Limitations in-memory processing
  - Rather slow in computation time for some calculations

TERADATA

Christiane Raffeiner
Analytical Consultant
Vienna Advanced Analytics
Center of Expertise

Teradata GmbH
christiane.raffeiner@teradata.com

Teradata.at



Backup

TERADATA®

Raising Intelligence

Advanced Analytics Best Practice

Data Exploration

Explore all the data directly in the database with Teradata Profiler

Data Preparation

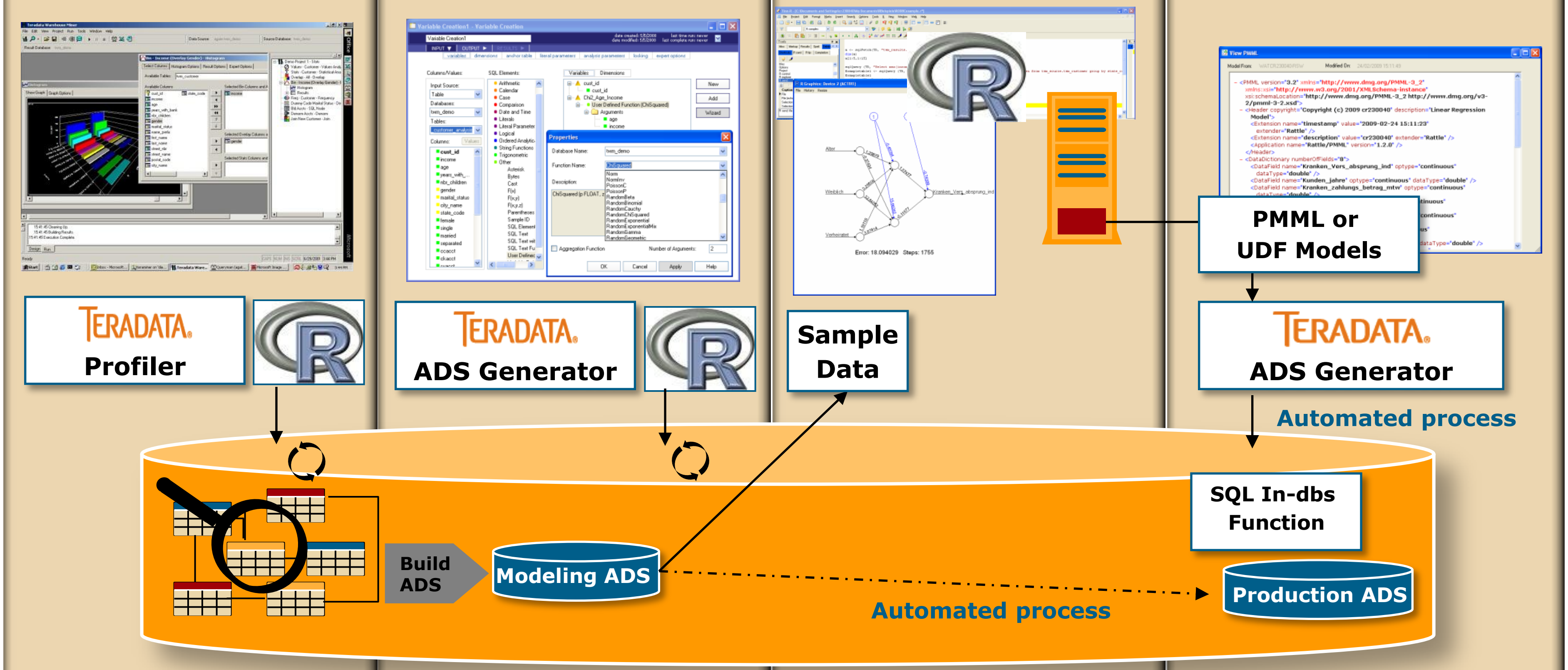
Transform and aggregate data in the database with Teradata ADS Generator

Model Development

Sample your ADS data and build your model on an R client

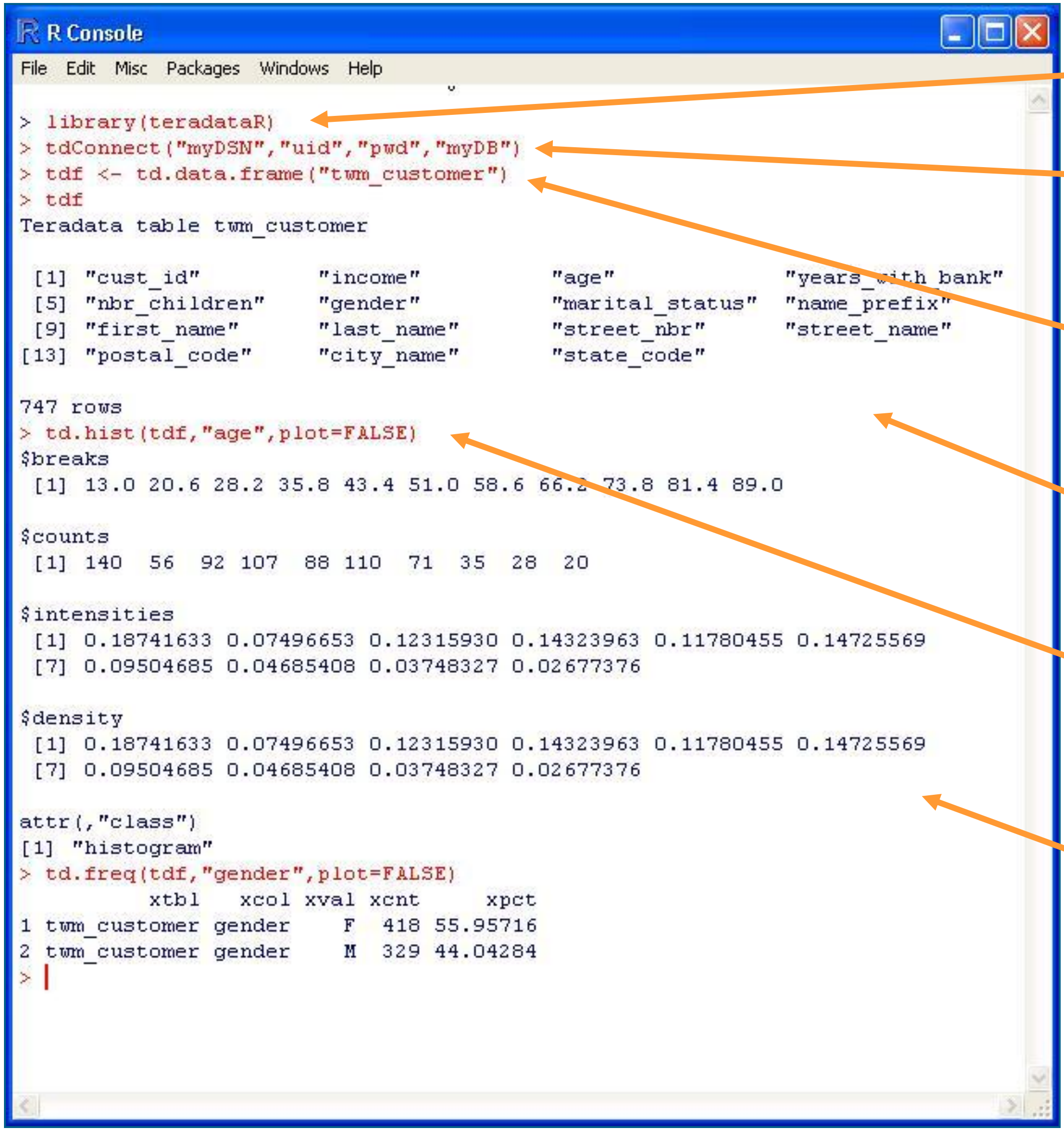
Model Deployment

Converts your R PMML model to SQL automatically generates the production ADS



Teradata Raising Intelligence

Sample R Code



```
> library(teradataR)
> tdConnect("myDSN","uid","pwd","myDB")
> tdf <- td.data.frame("twm_customer")
> tdf
Teradata table twm_customer

 [1] "cust_id"         "income"          "age"             "years_with_bank"
 [5] "nbr_children"    "gender"          "marital_status"  "name_prefix"
 [9] "first_name"      "last_name"       "street_nbr"      "street_name"
[13] "postal_code"     "city_name"       "state_code"

747 rows
> td.hist(tdf,"age",plot=FALSE)
$breaks
 [1] 13.0 20.6 28.2 35.8 43.4 51.0 58.6 66.2 73.8 81.4 89.0

$counts
 [1] 140  56  92 107  88 110  71  35  28  20

$intensities
 [1] 0.18741633 0.07496653 0.12315930 0.14323963 0.11780455 0.14725569
 [7] 0.09504685 0.04685408 0.03748327 0.02677376

$density
 [1] 0.18741633 0.07496653 0.12315930 0.14323963 0.11780455 0.14725569
 [7] 0.09504685 0.04685408 0.03748327 0.02677376

attr(,"class")
[1] "histogram"
> td.freq(tdf,"gender",plot=FALSE)
          xtbl    xcol xval xcnt     xpct
1 twm_customer gender    F  418 55.95716
2 twm_customer gender    M  329 44.04284
> 
```

1) Teradata library

2) Connect to Teradata

3) Establish a pointer to a Teradata table

4) View Teradata table variables

5) Call in-database analytics

6) View results