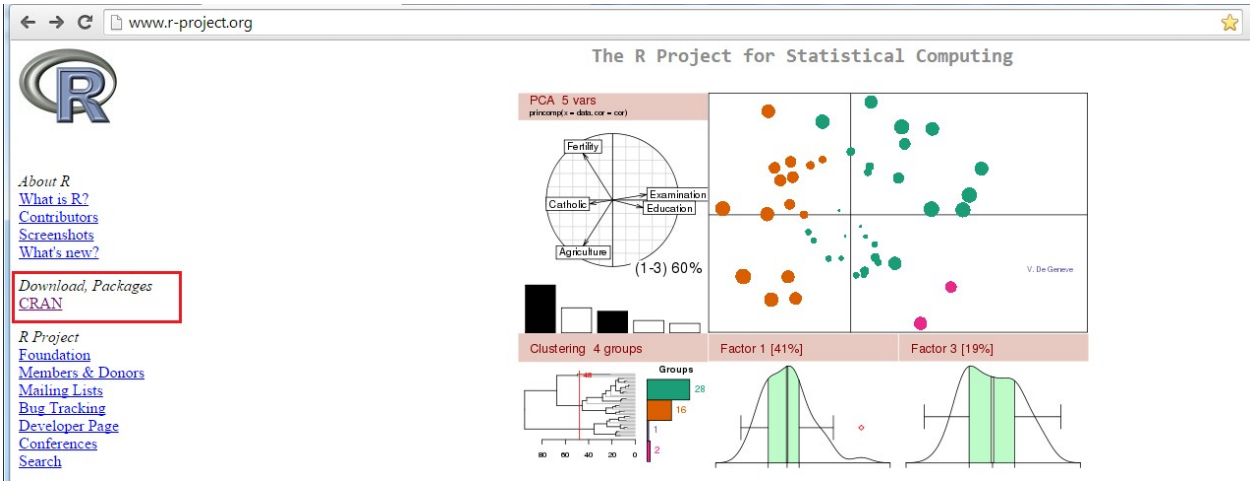
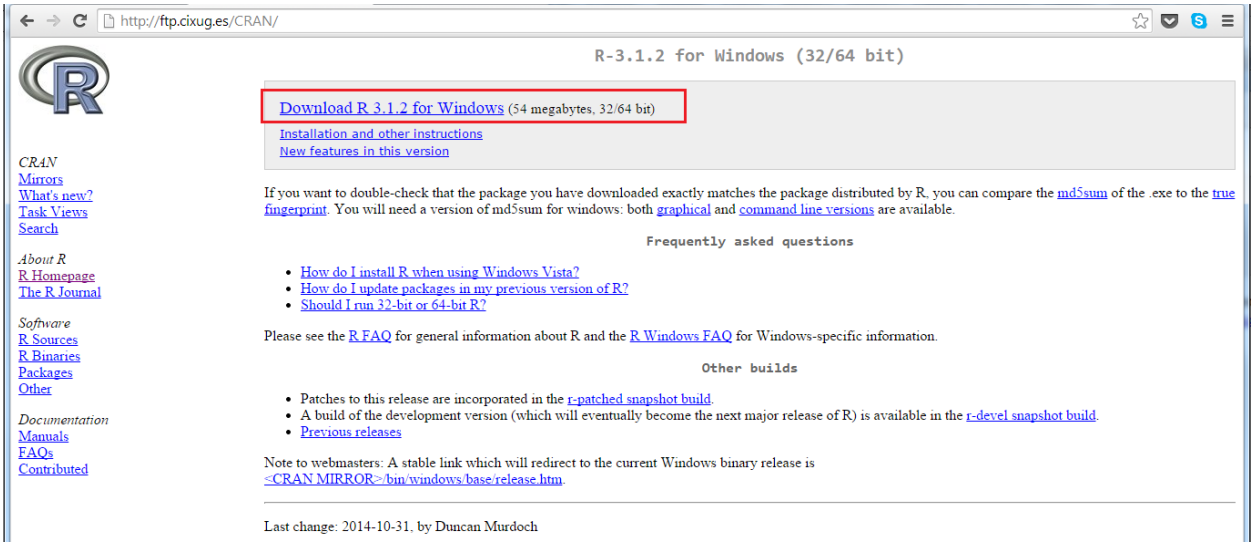
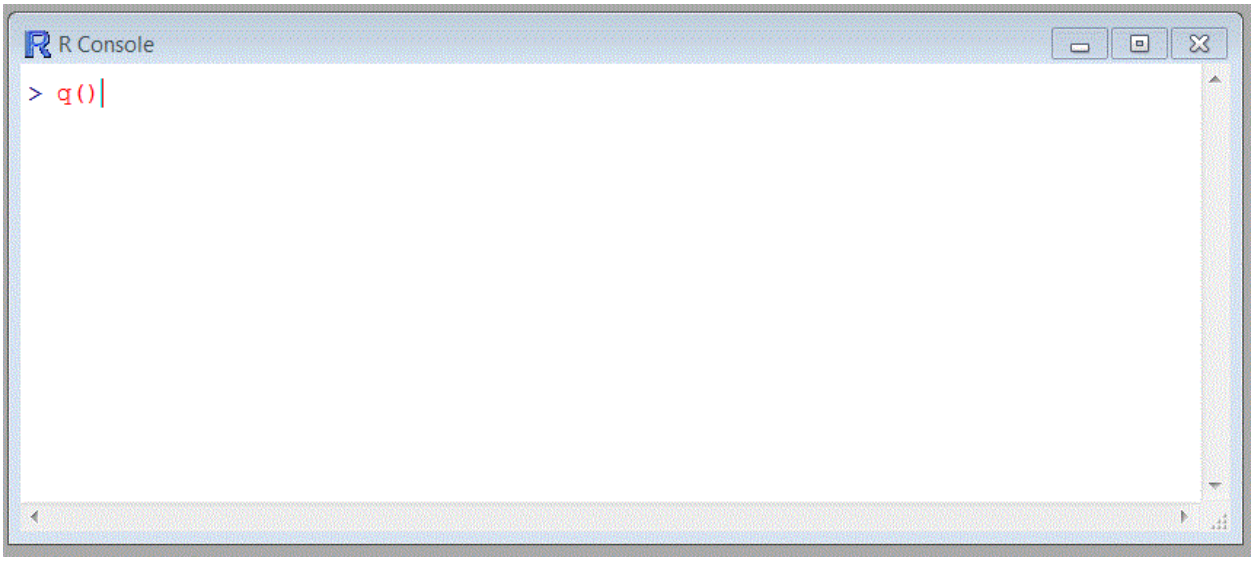
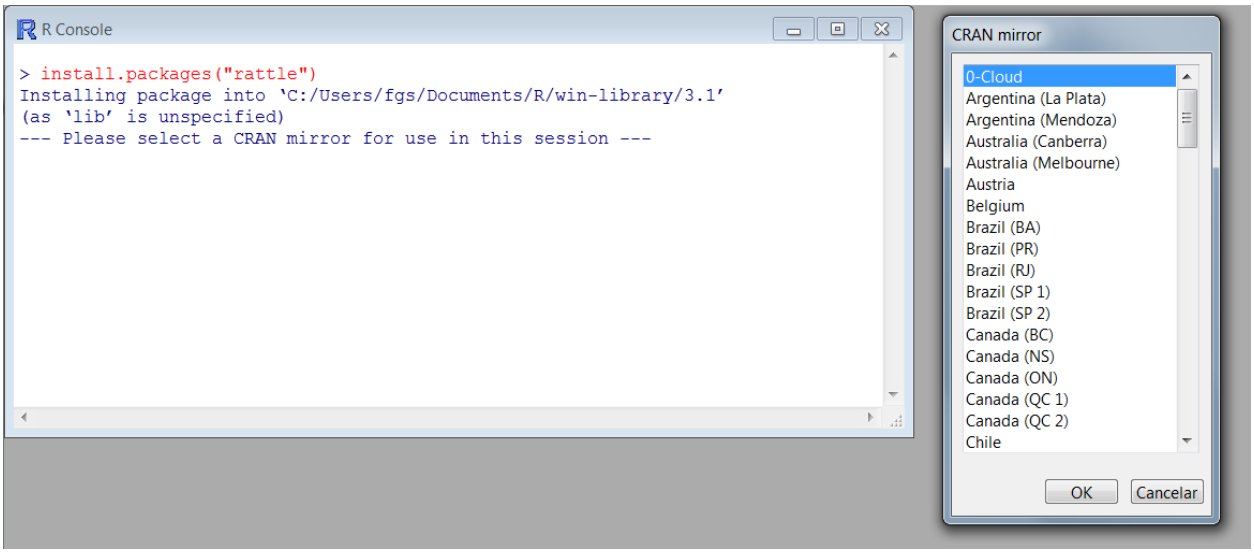
Predictive Analysis Using Rattle and Qlik Sense

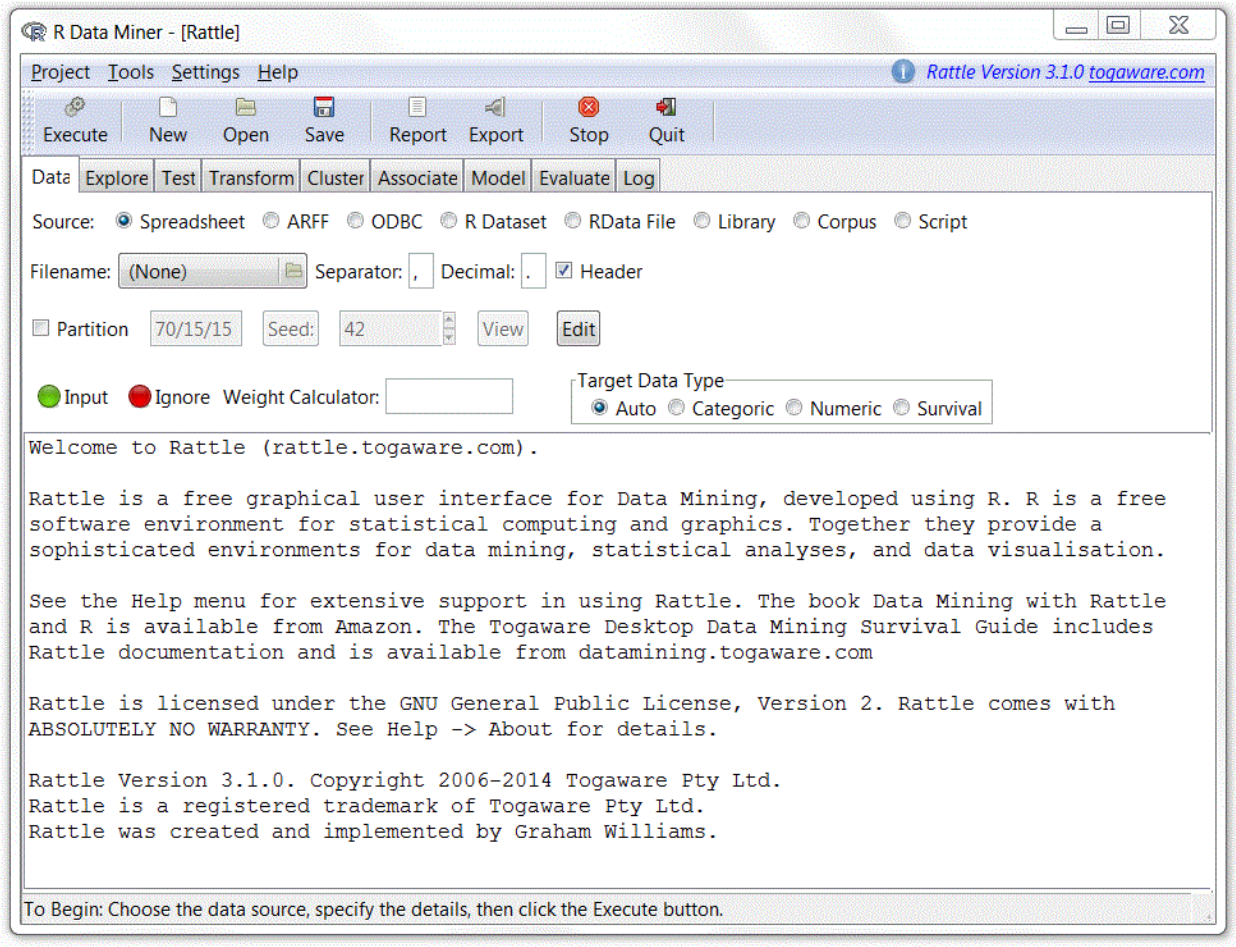
Chapter 1: Getting Ready with Predictive Analytics



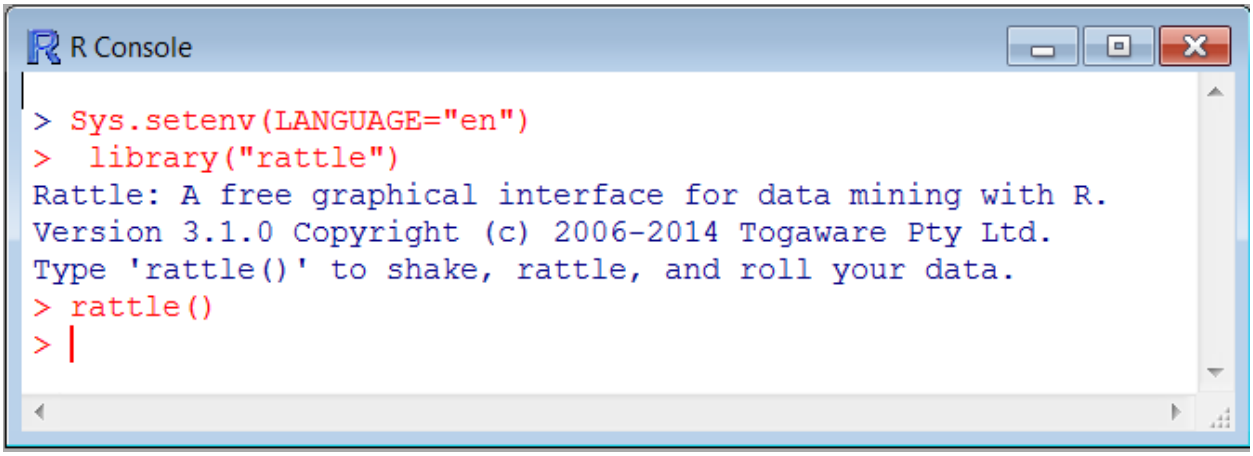

http://ftp.cixug.es/CRAN/
R-3.1.2 for Windows (32/64 bit)
Download R 3.1.2 for Windows (54 megabytes, 32/64 bit)
Installation and other instructions
New features in this version
If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.
Frequently asked questions
How do I install R when using Windows Vista?
How do I update packages in my previous version of R?
Should I run 32-bit or 64-bit R?
Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.
Other builds
Patches to this release are incorporated in the r-patched snapshot build.
A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
Previous releases
Note to webmasters: A stable link which will redirect to the current Windows binary release is <CRAN MIRROR>/bin/windows/base/release.htm.
Last change: 2014-10-31, by Duncan Murdoch
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

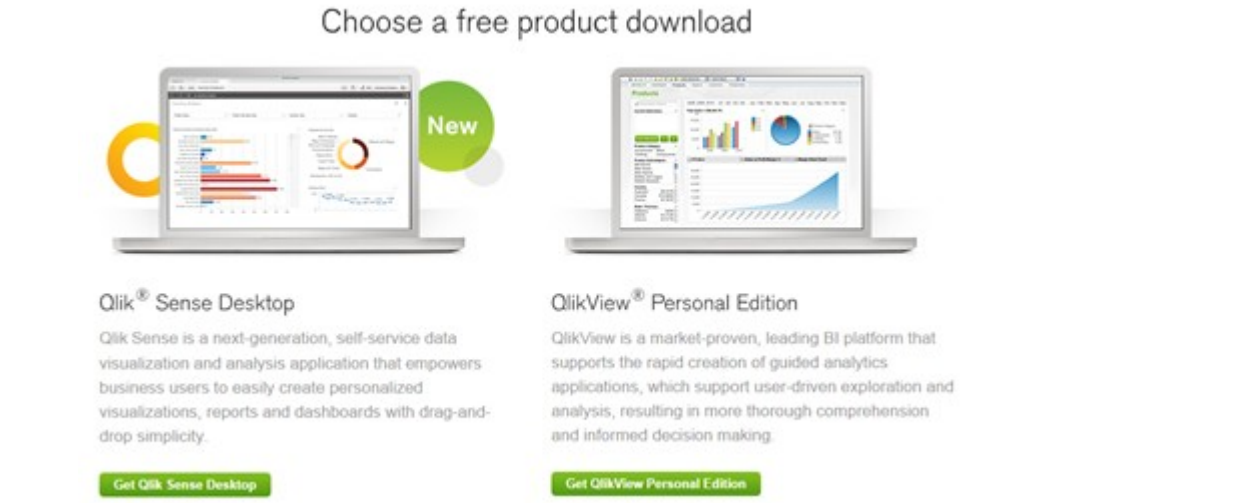

R Console
> q()

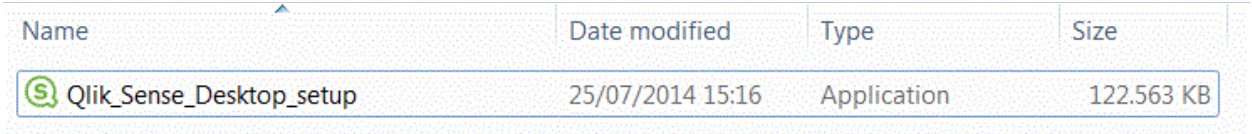

R Console
> install.packages("rattle")
Installing package into 'C:/Users/fgs/Documents/R/win-library/3.1'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
CRAN mirror
0-Cloud
Argentina (La Plata)
Argentina (Mendoza)
Australia (Canberra)
Australia (Melbourne)
Austria
Belgium
Brazil (BA)
Brazil (PR)
Brazil (RJ)
Brazil (SP 1)
Brazil (SP 2)
Canada (BC)
Canada (NS)
Canada (ON)
Canada (QC 1)
Canada (QC 2)
Chile
OK
Cancelar

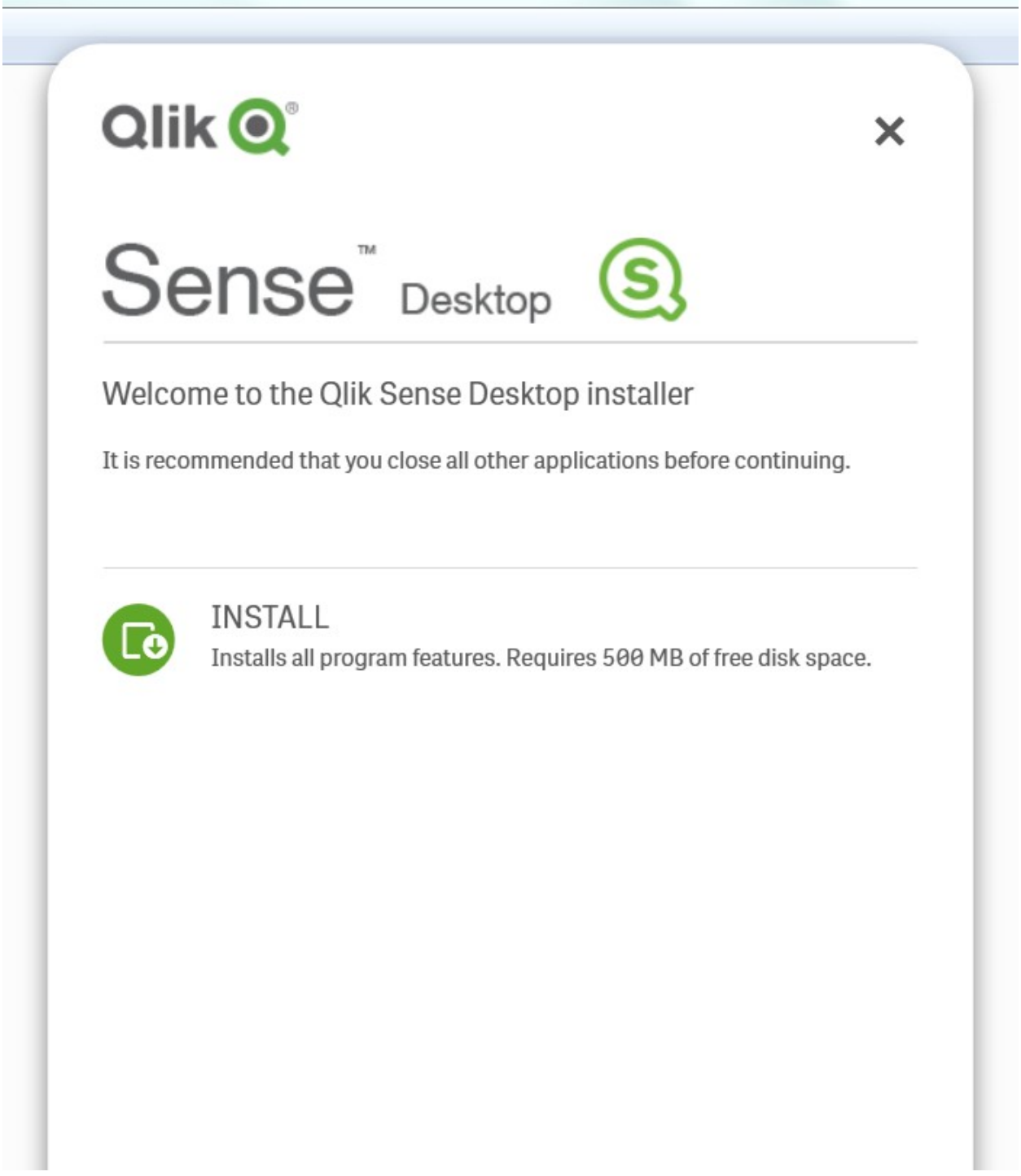

R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 3.1.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename: (None) Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
Welcome to Rattle (rattle.togaware.com).
Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.
See the Help menu for extensive support in using Rattle. The book Data Mining with Rattle and R is available from Amazon. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togaware.com
Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.
Rattle Version 3.1.0. Copyright 2006-2014 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
To Begin: Choose the data source, specify the details, then click the Execute button.

Execute
New
Open
Save
Report
Export
Stop
Quit

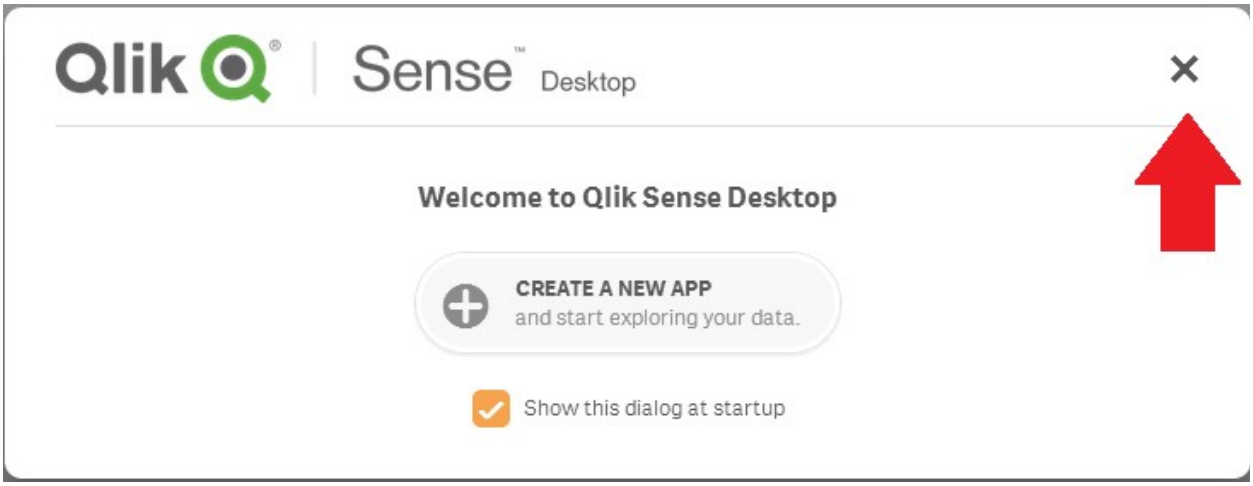

R Console
> Sys.setenv(LANGUAGE="en")
> library("rattle")
Rattle: A free graphical interface for data mining with R.
Version 3.1.0 Copyright (c) 2006-2014 Togaware Pty Ltd.
Type 'rattle()' to shake, rattle, and roll your data.
> rattle()
>


Choose a free product download
New
Qlik® Sense Desktop
Qlik Sense is a next-generation, self-service data visualization and analysis application that empowers business users to easily create personalized visualizations, reports and dashboards with drag-and-drop simplicity.
Get Qlik Sense Desktop
QlikView® Personal Edition
QlikView is a market-proven, leading BI platform that supports the rapid creation of guided analytics applications, which support user-driven exploration and analysis, resulting in more thorough comprehension and informed decision making.
Get QlikView Personal Edition

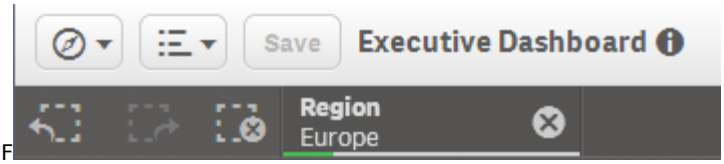

Name
Date modified
Type
Size
Qlik_Sense_Desktop_setup
25/07/2014 15:16
Application
122.563 KB


Qlik
Sense™ Desktop
Welcome to the Qlik Sense Desktop installer
It is recommended that you close all other applications before continuing.
INSTALL
Installs all program features. Requires 500 MB of free disk space.

Qlik Sense
Desktop


Qlik
Sense
Desktop
Welcome to Qlik Sense Desktop
CREATE A NEW APP
and start exploring your data.
Show this dialog at startup

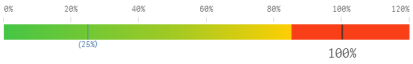
F


Save
Executive Dashboard
Region
Europe

KPI Dashboard

Expenses vs. Target = 100%

Expense Target = 25%



Revenue vs Last Year Revenue = 114%

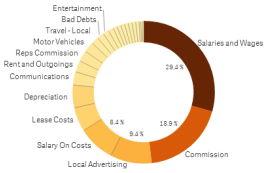
Target Revenue Increase = 110%

Account Receivable Overdue = 25%

Account Receivable Target = 22%

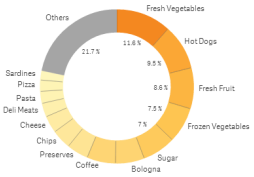
Expenses by Accounts *

Total Expenses = $36,215,584



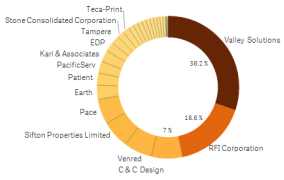
Revenue by Products

Total Revenue = $67,119,290



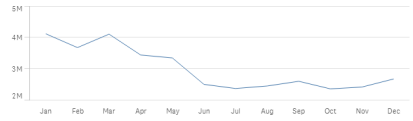
Amount Overdue by Customers *

Total Overdue = $208,879



** The data set contains negative or zero values that cannot be shown in this chart.*

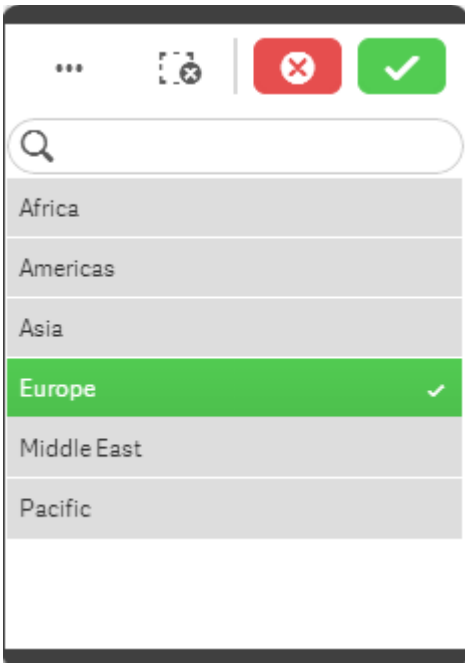
Expense Trend

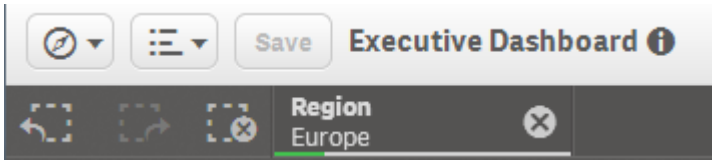


Revenue Trend

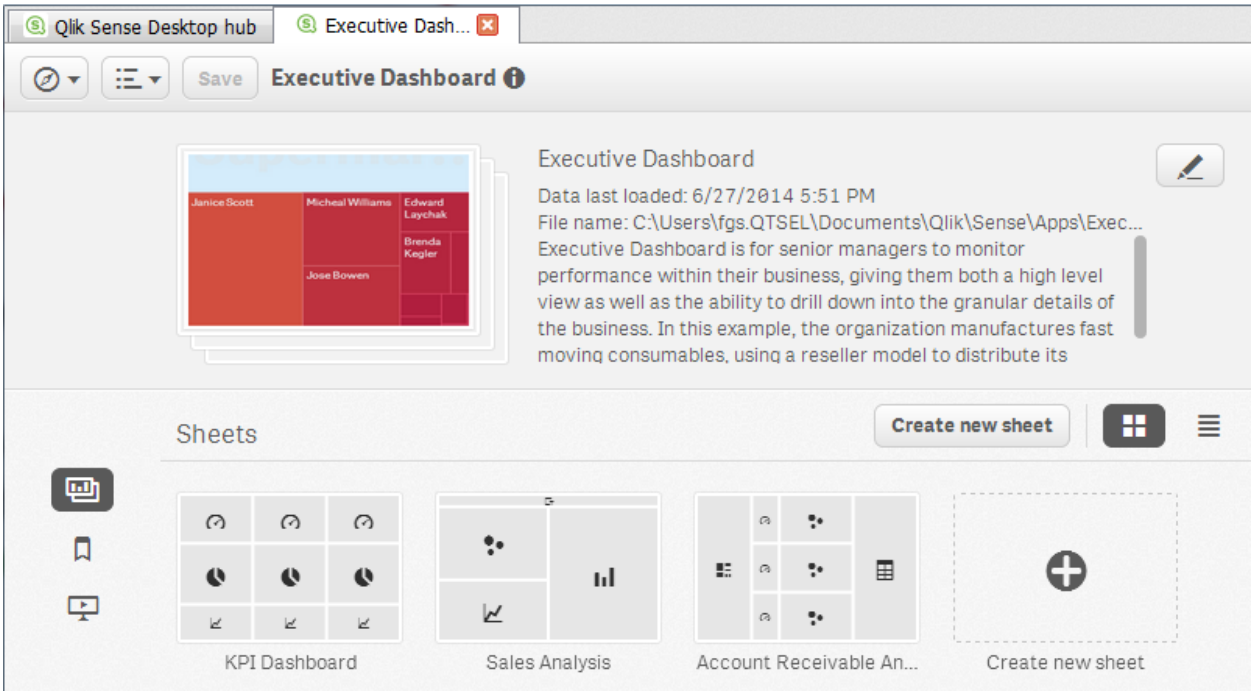
** The data set contains negative or zero values that cannot be shown in this chart.*

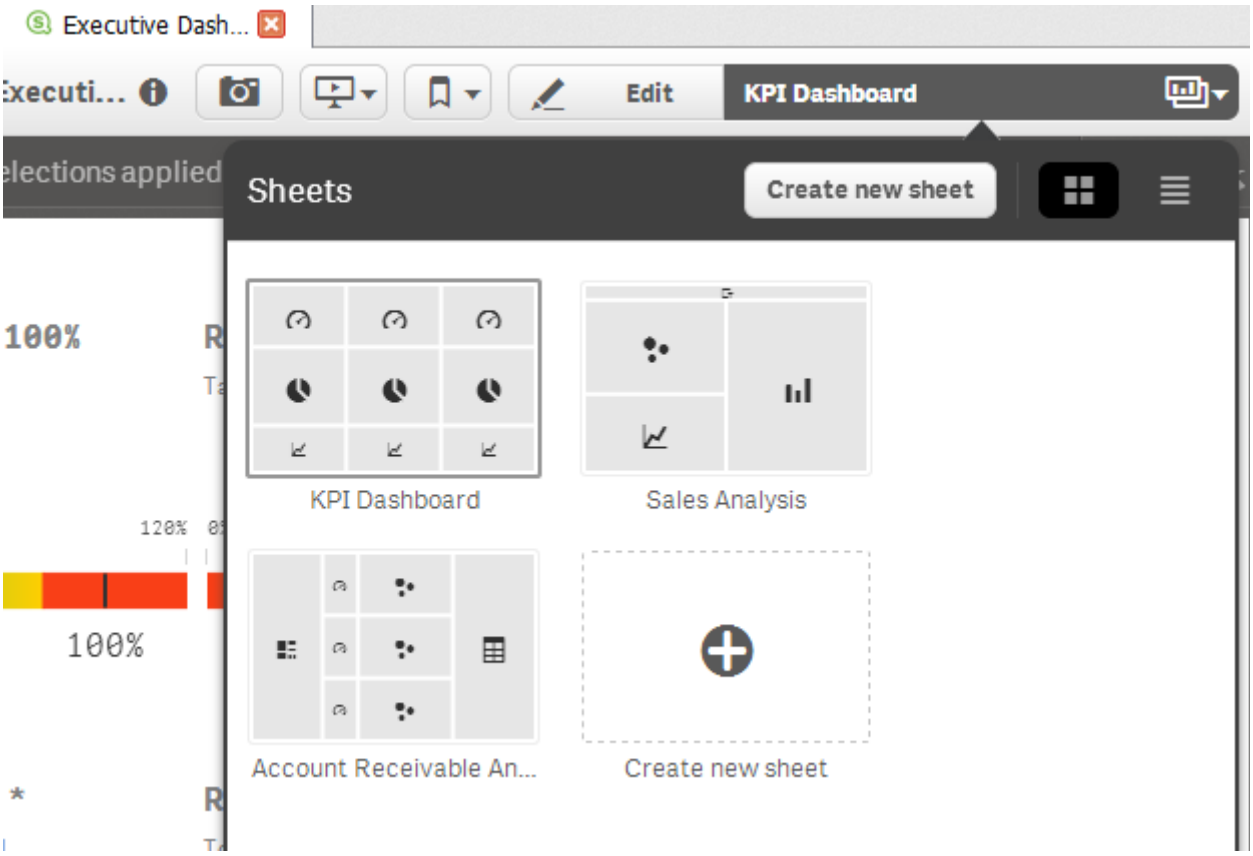
AR Trend

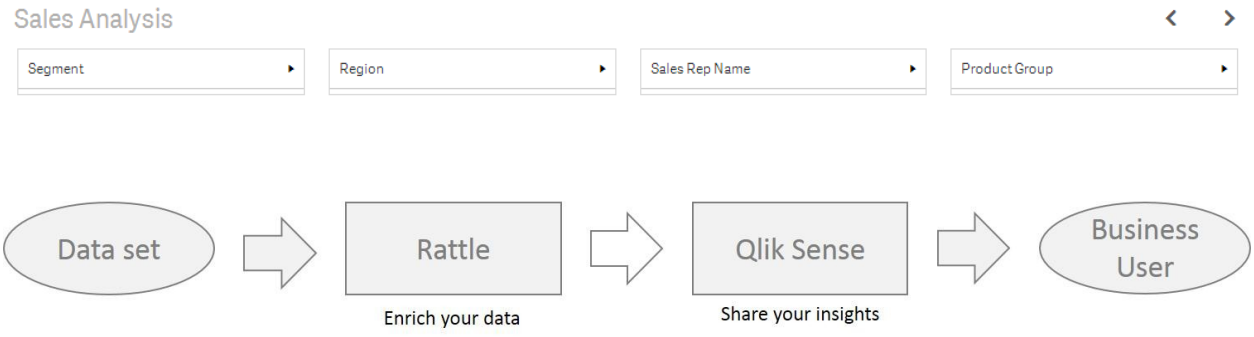

Africa
Americas
Asia
Europe
Middle East
Pacific


Save
Executive Dashboard
Region
Europe

R x64 3.1.2
R i386 3.1.2


Qlik Sense Desktop hub
Executive Dash...
Save
Executive Dashboard
Janice Scott
Micheal Williams
Edward Laychak
Brenda Kegler
Jose Bowen
Executive Dashboard
Data last loaded: 6/27/2014 5:51 PM
File name: C:\Users\fgs.QTSEL\Documents\Qlik\Sense\Apps\Exec...
Executive Dashboard is for senior managers to monitor performance within their business, giving them both a high level view as well as the ability to drill down into the granular details of the business. In this example, the organization manufactures fast moving consumables, using a reseller model to distribute its
Sheets
Create new sheet
KPI Dashboard
Sales Analysis
Account Receivable An...
Create new sheet

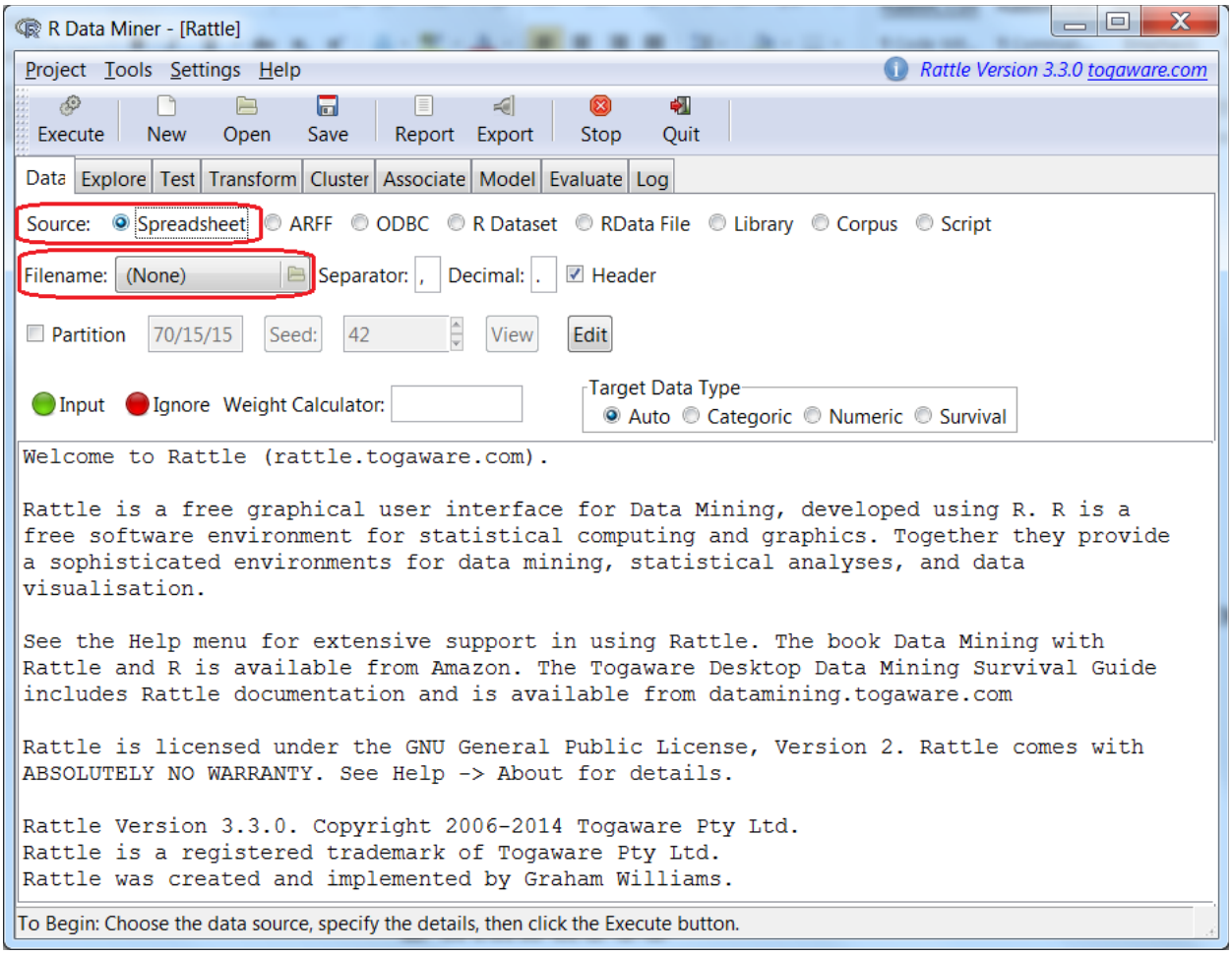

Executive Dash...
Edit
KPI Dashboard
Sheets
Create new sheet
100%
120%
100%
KPI Dashboard
Sales Analysis
Account Receivable An...
Create new sheet

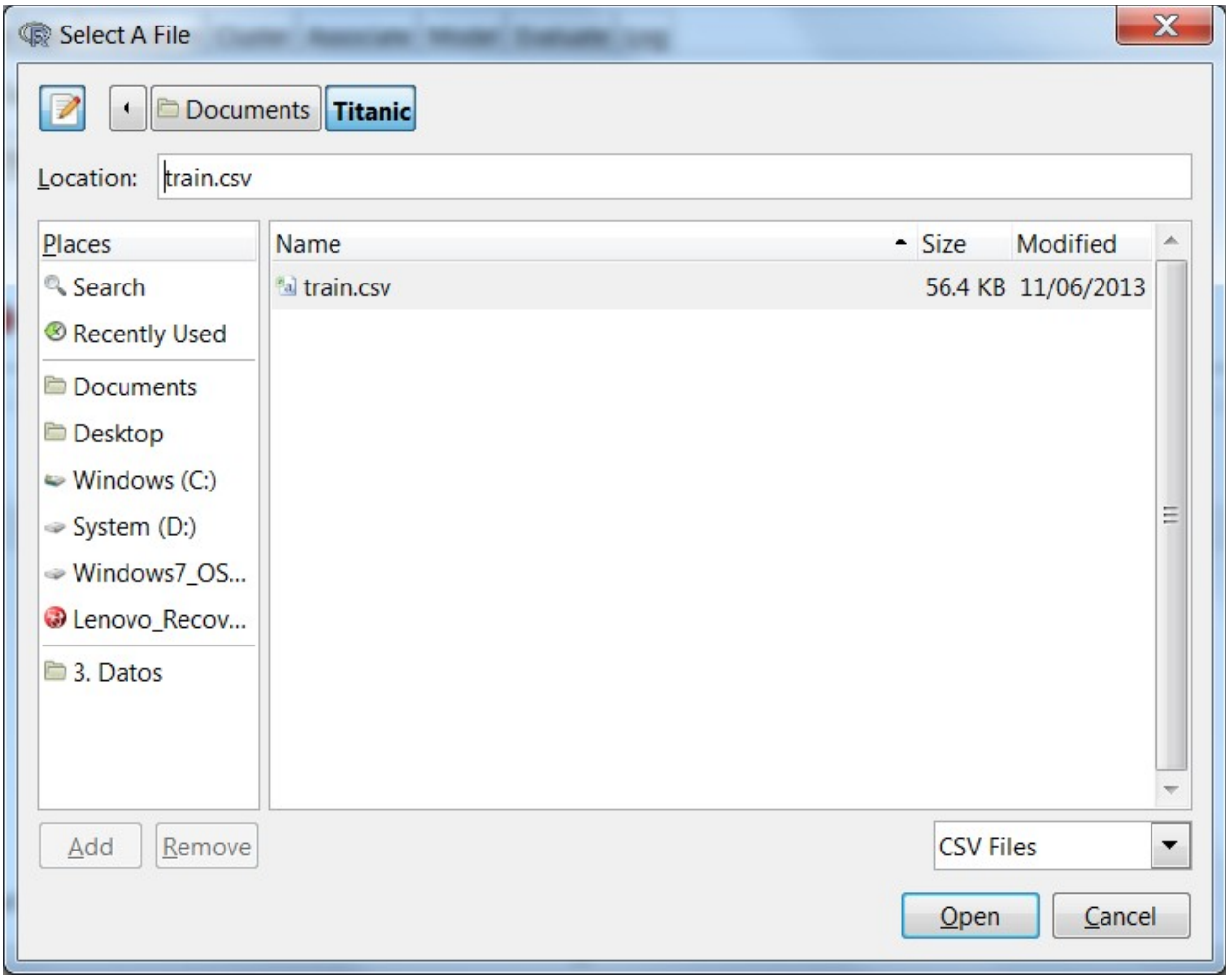

Sales Analysis
Segment
Region
Sales Rep Name
Product Group

Data set
Rattle
Enrich your data
Qlik Sense
Share your insights
Business User

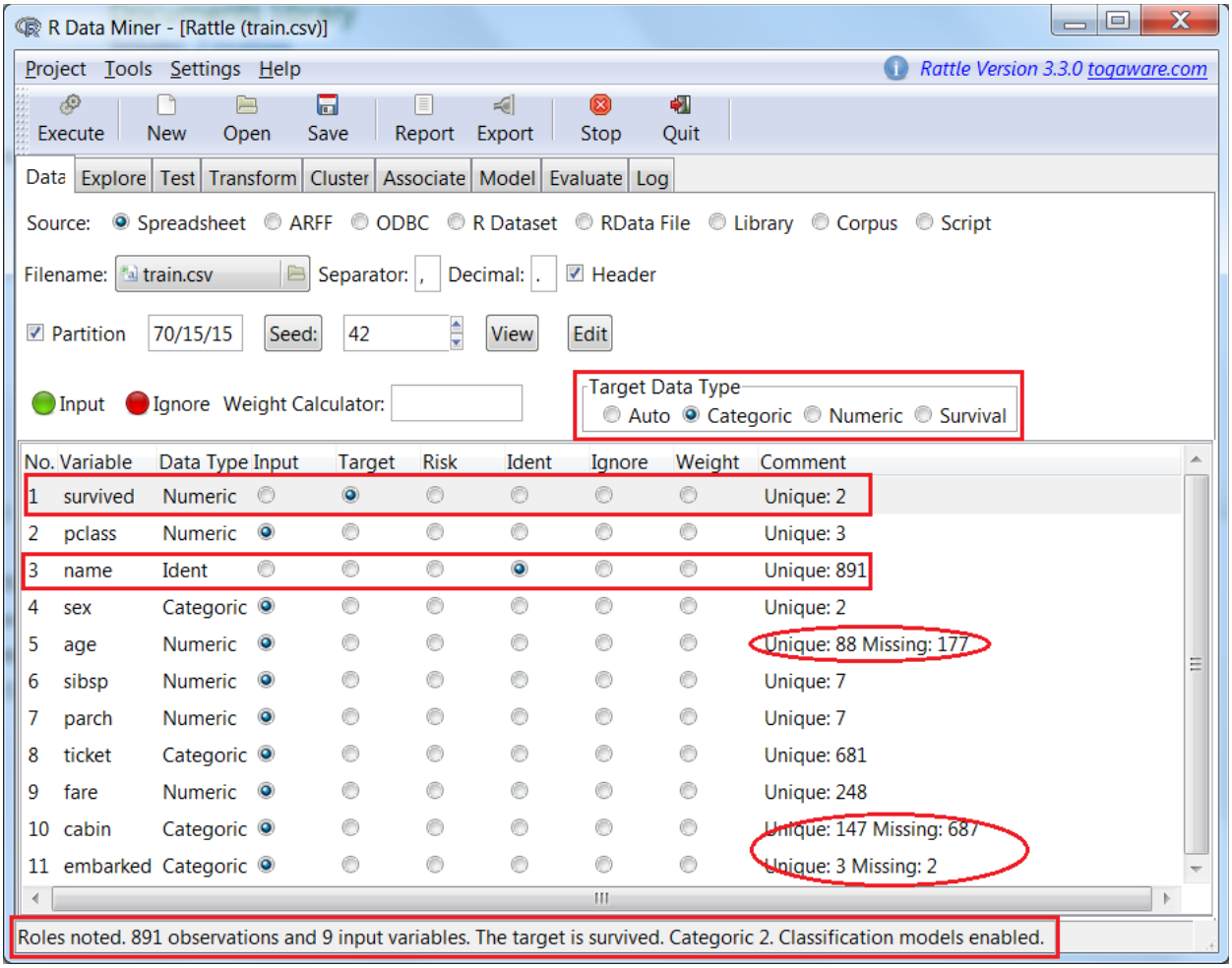
Chapter 2: Preparing Your Data

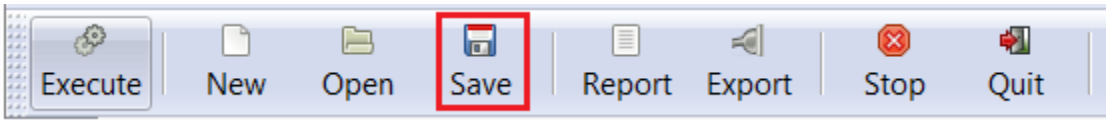
PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2.	7.925		S
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S
6	0	3	Moran, Mr. James	male		0	0	330877	8.4583		Q
7	0	1	McCarthy, Mr. Timothy J	male	54	0	0	17463	51.8625	E46	S
8	0	3	Palsson, Master. Gosta Leonard	male	2	3	1	349909	21.075		S
9	1	3	Johnson, Mrs. Oscar W (Elisabeth Vilhelmina Berg)	female	27	0	2	347742	11.1333		S
10	1	2	Nasser, Mrs. Nicholas (Adele Achem)	female	14	1	0	237736	30.0708		C
11	1	3	Sandstrom, Miss. Marguerite Rut	female	4	1	1	PP 9549	16.7	G6	S


R Data Miner - [Rattle]
Project Tools Settings Help
Rattle Version 3.3.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename: (None) Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
Welcome to Rattle (rattle.togaware.com).
Rattle is a free graphical user interface for Data Mining, developed using R. R is a free software environment for statistical computing and graphics. Together they provide a sophisticated environments for data mining, statistical analyses, and data visualisation.
See the Help menu for extensive support in using Rattle. The book Data Mining with Rattle and R is available from Amazon. The Togaware Desktop Data Mining Survival Guide includes Rattle documentation and is available from datamining.togaware.com
Rattle is licensed under the GNU General Public License, Version 2. Rattle comes with ABSOLUTELY NO WARRANTY. See Help -> About for details.
Rattle Version 3.3.0. Copyright 2006-2014 Togaware Pty Ltd.
Rattle is a registered trademark of Togaware Pty Ltd.
Rattle was created and implemented by Graham Williams.
To Begin: Choose the data source, specify the details, then click the Execute button.

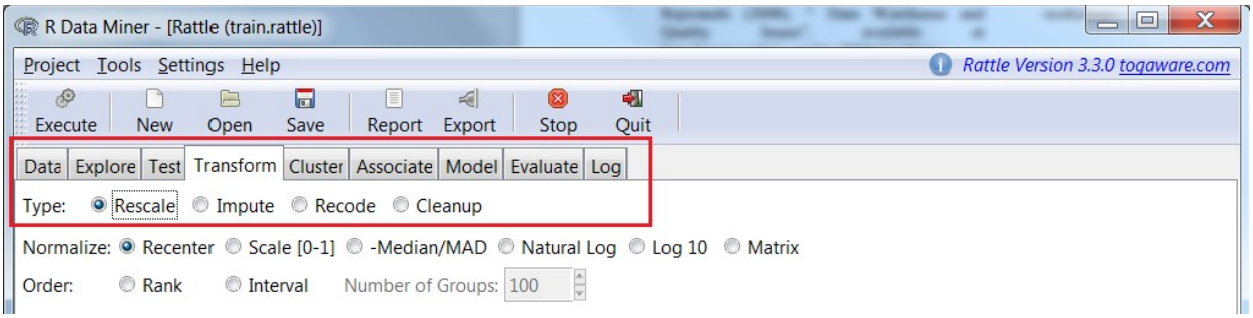

Select A File
Documents
Titanic
Location:
train.csv
Places
Search
Recently Used
Documents
Desktop
Windows (C:)
System (D:)
Windows7_OS...
Lenovo_Recov...
3. Datos
Name
Size
Modified
train.csv
56.4 KB
11/06/2013
Add
Remove
CSV Files
Open
Cancel

Execute
New
Open
Save
Report
Export
Stop
Quit


R Data Miner - [Rattle (train.csv)]
Project Tools Settings Help
Rattle Version 3.3.0 togaware.com
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename: train.csv Separator: , Decimal: . Header
Partition 70/15/15 Seed: 42 View Edit
Input Ignore Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
No. Variable Data Type Input Target Risk Ident Ignore Weight Comment
1 survived Numeric Unique: 2
2 pclass Numeric Unique: 3
3 name Ident Unique: 891
4 sex Categoric Unique: 2
5 age Numeric Unique: 88 Missing: 177
6 sibsp Numeric Unique: 7
7 parch Numeric Unique: 7
8 ticket Categoric Unique: 681
9 fare Numeric Unique: 248
10 cabin Categoric Unique: 147 Missing: 687
11 embarked Categoric Unique: 3 Missing: 2
Roles noted. 891 observations and 9 input variables. The target is survived. Categoric 2. Classification models enabled.


Execute
New
Open
Save
Report
Export
Stop
Quit

Partition
70/15/15
Seed:
42
View
Edit

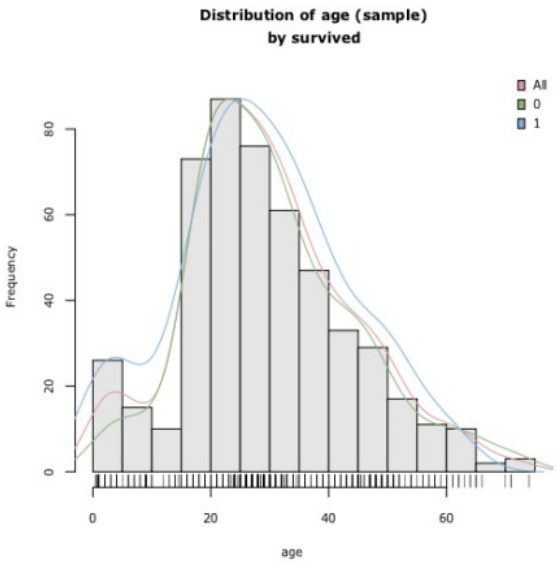


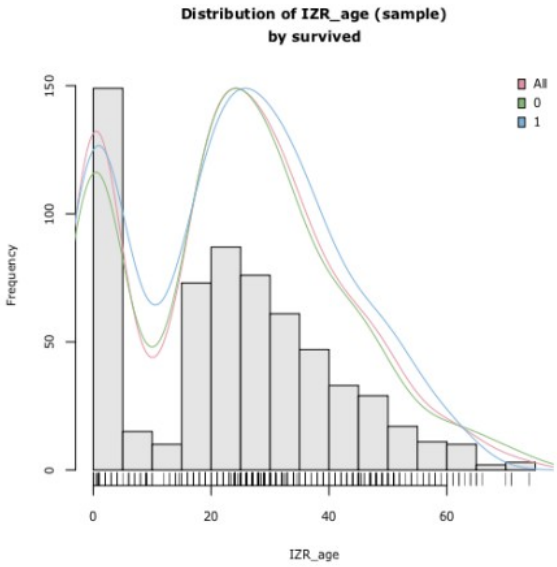
Type: Rescale Impute Recode Cleanup

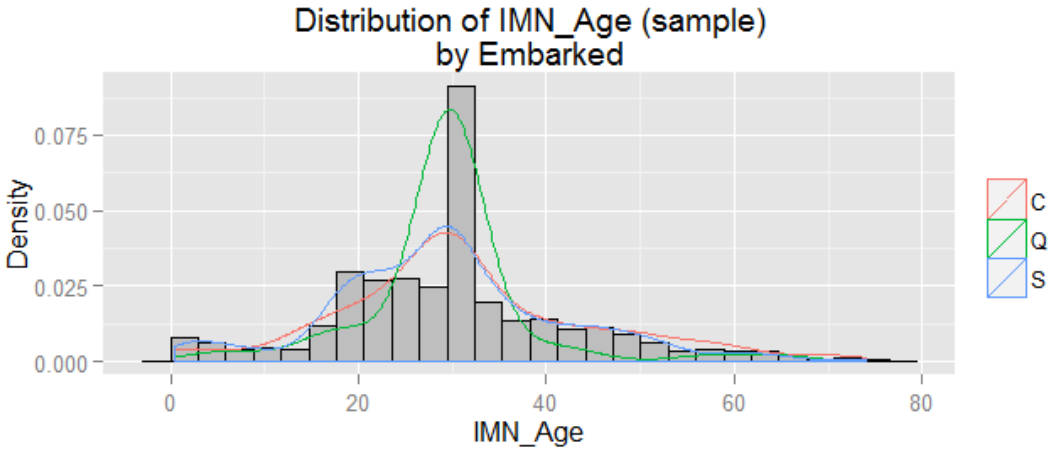
Select the required imputation method and the variables to apply this to, then click Execute:

Zero/Missing Mean Median Mode Constant:

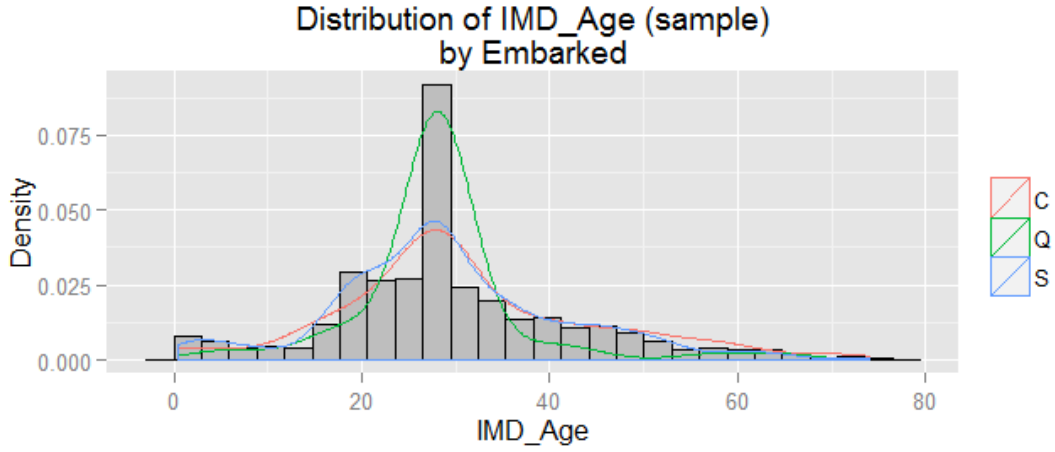
Interval Number of Groups: 10


Distribution of age (sample)
by survived
All
0
1
Frequency
0
20
40
60
80
0
20
40
60
age

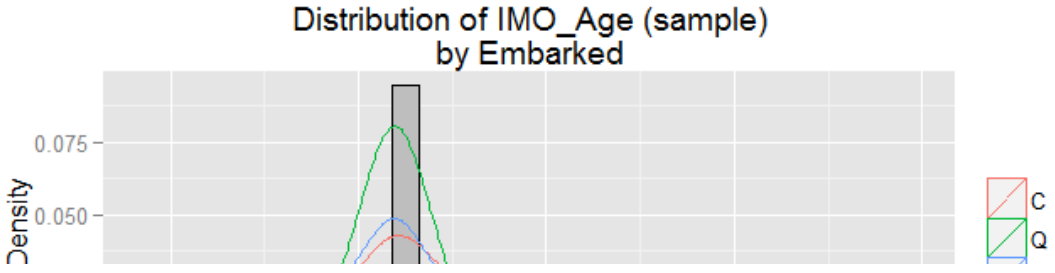

Distribution of IZR_age (sample)
by survived
All
0
1
Frequency
0
50
100
150
0
20
40
60
IZR_age



Rattle 2015-Mar-26 23:38:52



Rattle 2015-Mar-26 23:40:20 fgs

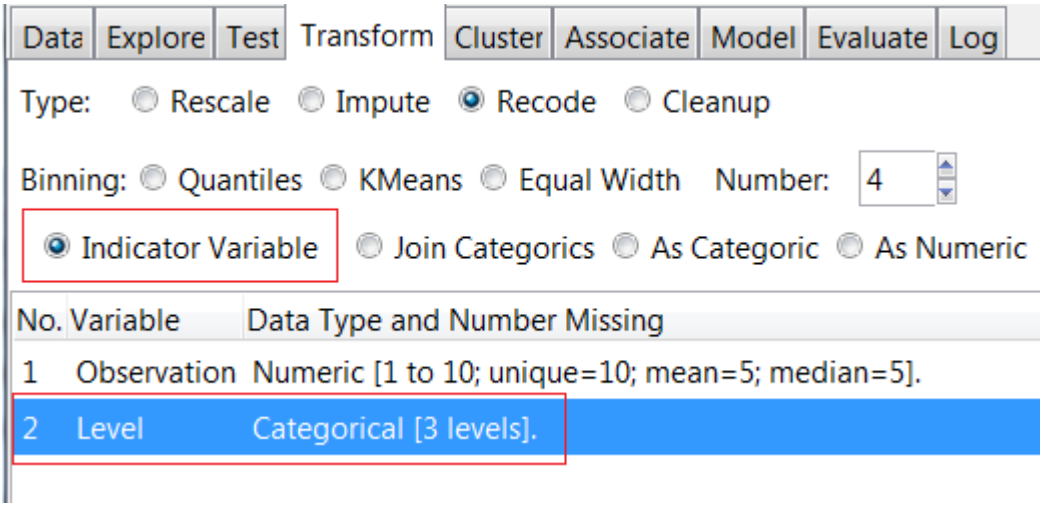


Type: Rescale Impute Recode Cleanup

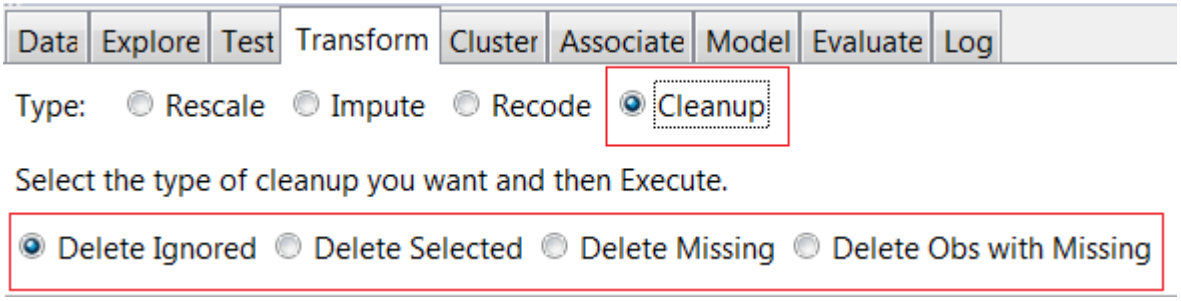
Binning: Quantiles KMeans Equal Width Number: 4

Indicator Variable Join Categorics As Categoric As Numeric

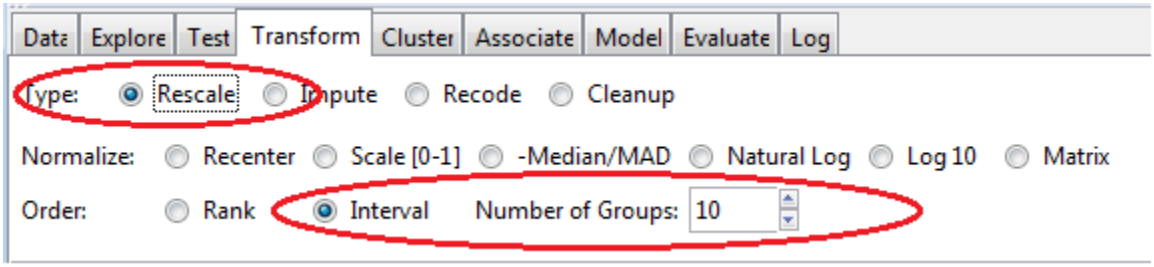
Observation	Level	Beginner indicator	Medium indicator	Advanced Indicator
1	Beginner	1	0	0
2	Medium	0	1	0
3	Medium	0	1	0
4	Beginner	1	0	0
5	Advanced	0	0	1
6	Advanced	0	0	1
7	Medium	0	1	0
8	Beginner	1	0	0
9	Beginner	1	0	0
10	Advanced	0	0	1



Observation	Level	Sex	TJN_Level_Sex
1	Begginer	Male	Begginer_Male
2	Medium	Male	Medium_Male
3	Medium	Male	Medium_Male
4	Begginer	Male	Begginer_Male
5	Advanced	Female	Advanced_Female
6	Advanced	Female	Advanced_Female
7	Medium	Female	Medium_Female
8	Begginer	Female	Begginer_Female
9	Begginer	Female	Begginer_Female
10	Advanced	Male	Advanced_Male



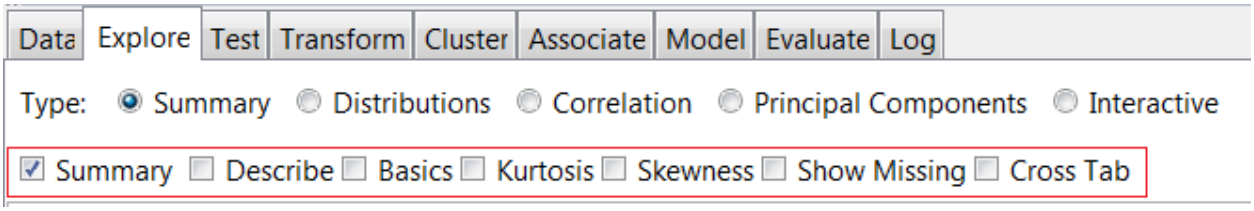
No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	survived	Numeric	●						Unique: 2
2	pclass	Numeric	●						Unique: 3
3	name	Categoric				●			Unique: 891
4	sex	Categoric	●						Unique: 2
5	age	Numeric	●						Unique: 88 Missing: 177
6	sibsp	Numeric	●						Unique: 7
7	parch	Numeric	●						Unique: 7
8	ticket	Categoric	●						Unique: 681
9	fare	Numeric					●		Unique: 248
10	cabin	Categoric	●						Unique: 147 Missing: 687
11	embarked	Categoric		●					Unique: 3 Missing: 2
12	RRK_fare	Numeric	●						Unique: 248



No.	Variable	Data Type and Number Missing
1	PassengerId	Numeric [1 to 891; unique=891; mean=446; median=446].
2	Survived	Numeric [0 to 1; unique=2; mean=0; median=0].
3	Pclass	Numeric [1 to 3; unique=3; mean=2; median=3].
4	Name	Categorical [891 levels].
5	Sex	Categorical [2 levels].
6	Age	Numeric [0.42 to 80.00; unique=88; mean=29.70; median=28.00; miss=177; ignored].
7	SibSp	Numeric [0 to 8; unique=7; mean=0; median=0].
8	Parch	Numeric [0 to 6; unique=7; mean=0; median=0].
9	Ticket	Categorical [681 levels].
10	Fare	Numeric [0.00 to 512.33; unique=248; mean=32.20; median=14.45].
11	Cabin	Categorical [147 levels; miss=687].
12	Embarked	Categorical [3 levels; miss=2].
13	RIN_Age_10	Numeric [0.00 to 9.00; unique=10; mean=3.29; median=3.00; miss=177].
14	BE10_Age	Categorical [10 levels; miss=177].

No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	PassengerId	Numeric	○	○	○	◉	○	○	Unique: 891
2	Survived	Numeric	◉	○	○	○	○	○	Unique: 2
3	Pclass	Numeric	◉	○	○	○	○	○	Unique: 3
4	Name	Categoric	○	○	○	◉	○	○	Unique: 891
5	Sex	Categoric	◉	○	○	○	○	○	Unique: 2
6	Age	Numeric	○	○	○	○	◉	○	Unique: 88 Missing: 177
7	SibSp	Numeric	◉	○	○	○	○	○	Unique: 7
8	Parch	Numeric	◉	○	○	○	○	○	Unique: 7
9	Ticket	Categoric	◉	○	○	○	○	○	Unique: 681
10	Fare	Numeric	◉	○	○	○	○	○	Unique: 248
11	Cabin	Categoric	◉	○	○	○	○	○	Unique: 147 Missing: 687
12	Embarked	Categoric	○	◉	○	○	○	○	Unique: 3 Missing: 2
13	R01_Age	Numeric	◉	○	○	○	○	○	Unique: 88 Missing: 177

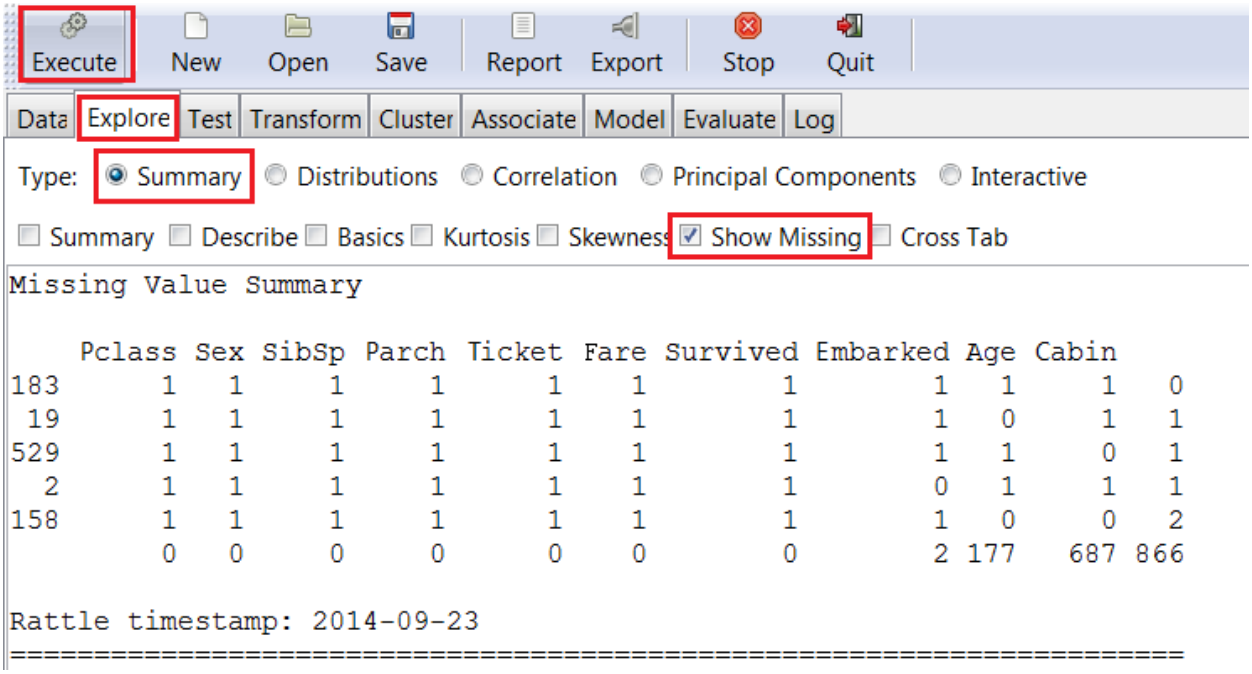
Chapter 3: Exploring and Understanding Your Data

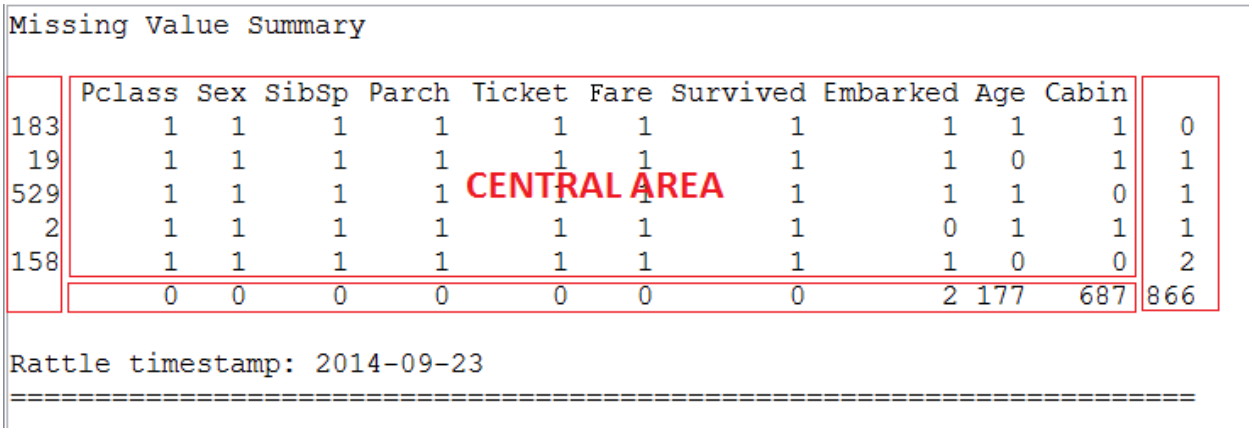


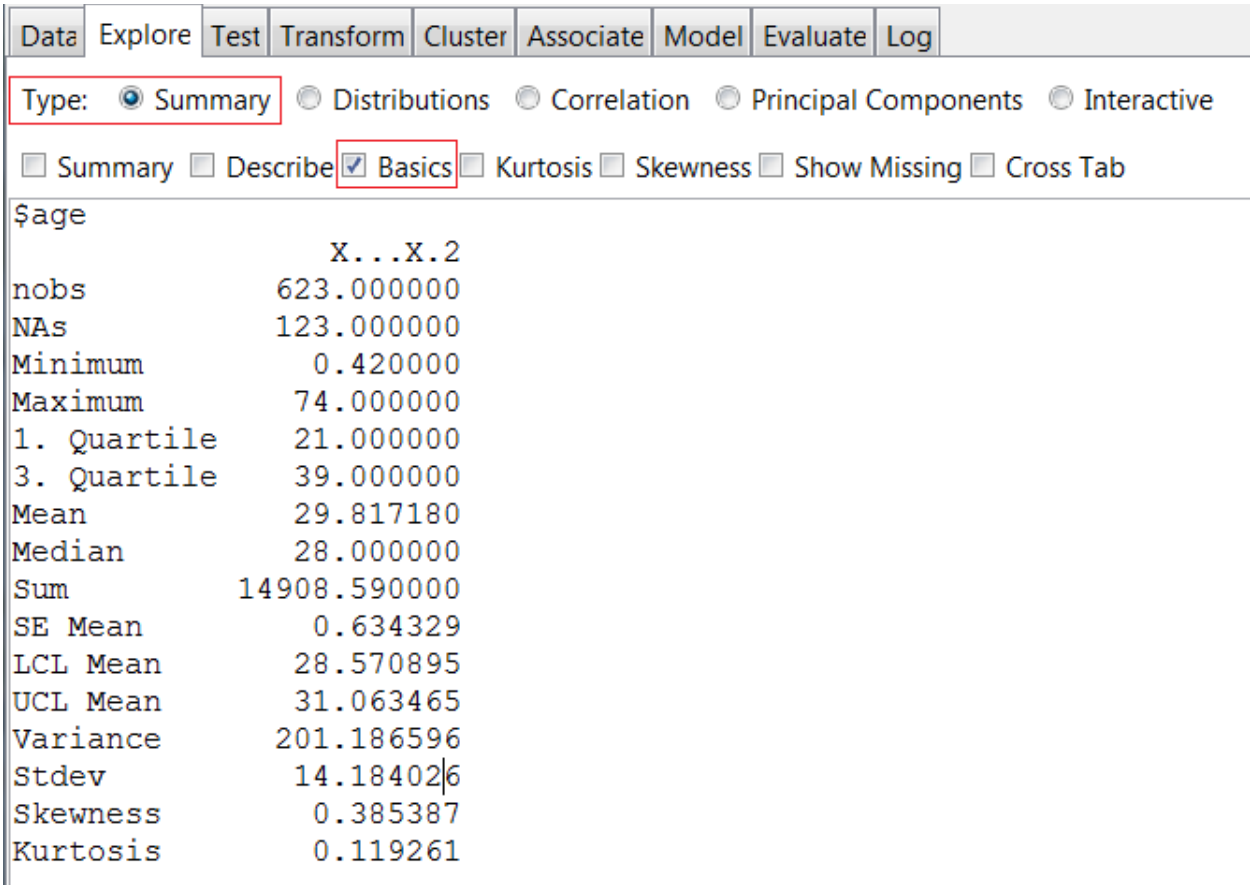

```
      Age
 Min.   : 0.42
 1st Qu.:21.00
 Median :28.00
 Mean   :29.82
 3rd Qu.:39.00
 Max.   :74.00
 NA's   :123
```

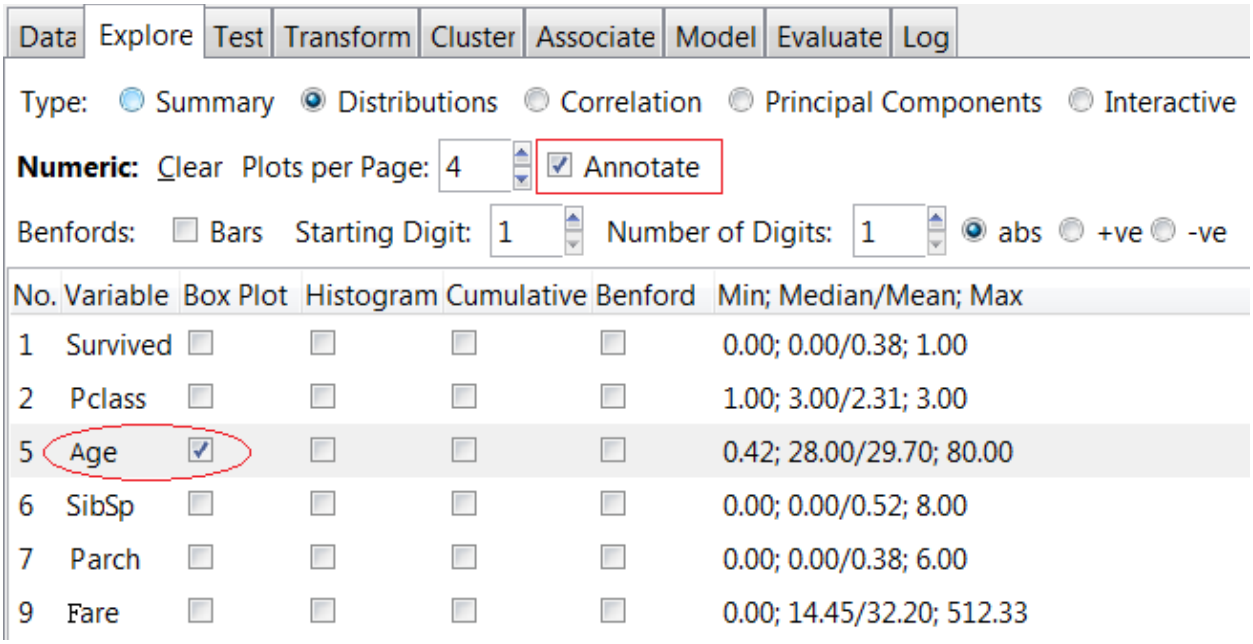
```
Embarked
C   :117
Q   : 56
S   :449
NA's:  1
```

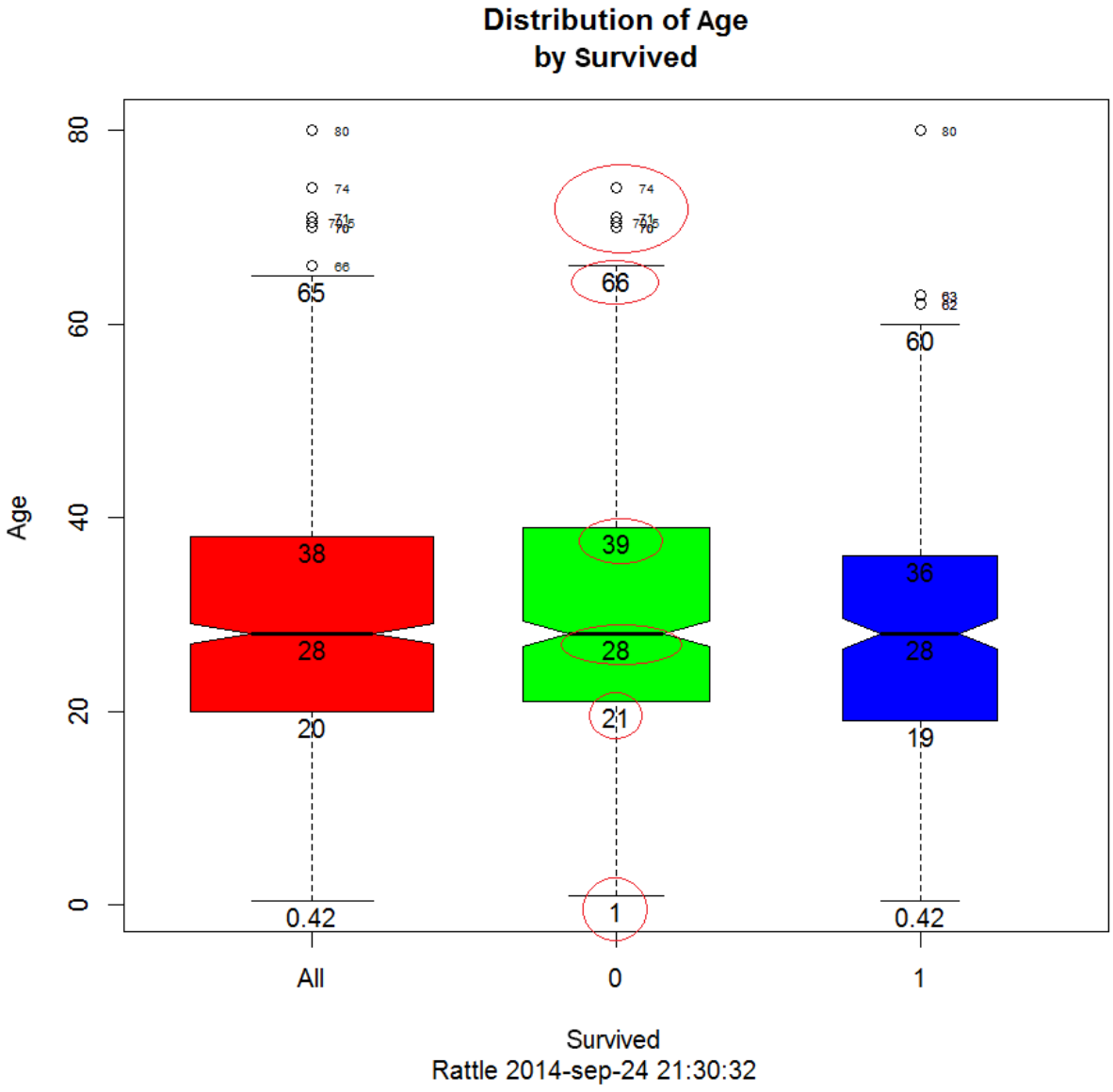
$$Variance = \frac{\sum(X - Median)^2}{N}$$

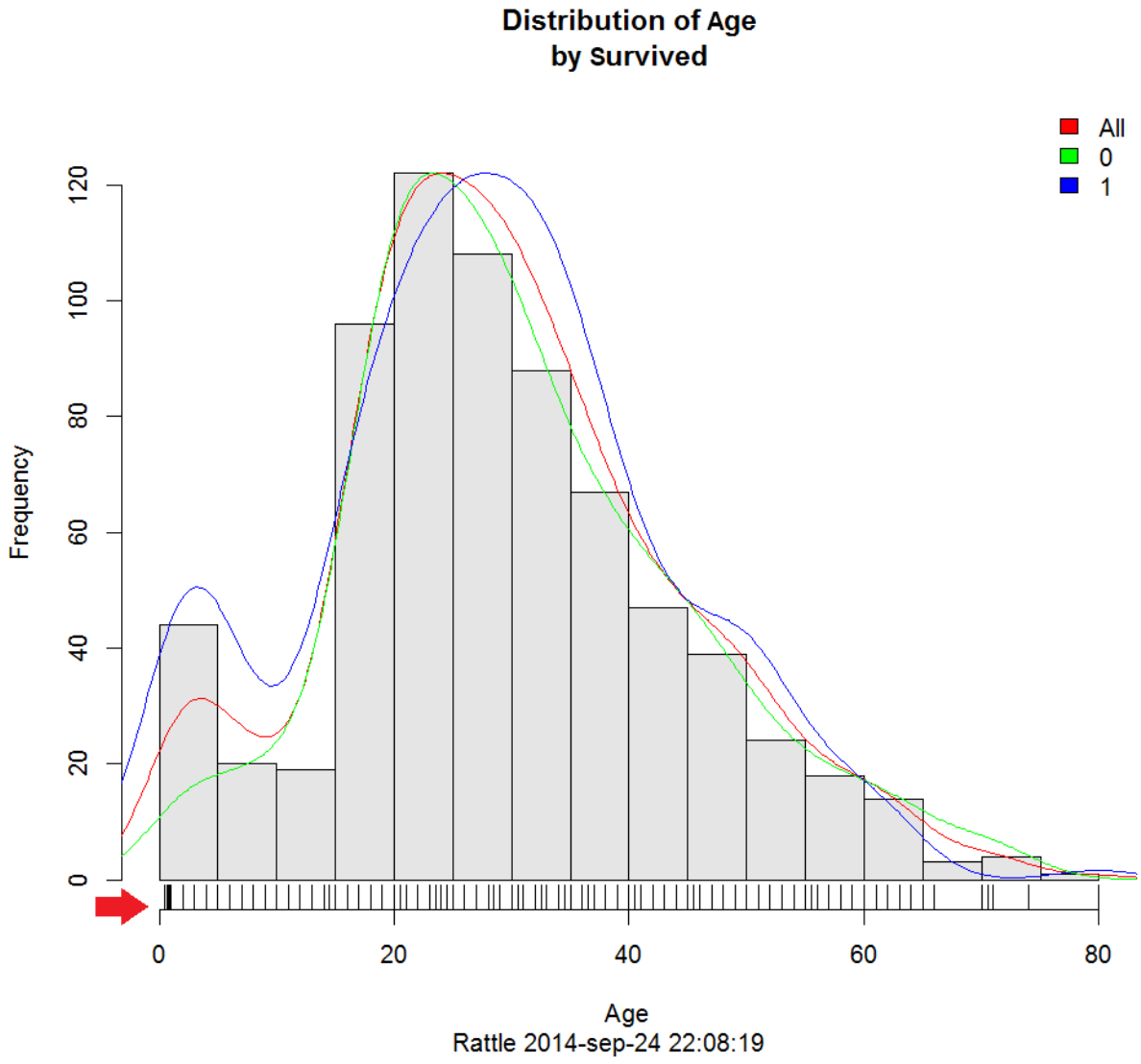

Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type:
Summary
Distributions
Correlation
Principal Components
Interactive
Summary
Describe
Basics
Kurtosis
Skewness
Show Missing
Cross Tab
Missing Value Summary
Pclass Sex SibSp Parch Ticket Fare Survived Embarked Age Cabin
183 1 1 1 1 1 1 1 1 1 1 0
19 1 1 1 1 1 1 1 1 0 1 1
529 1 1 1 1 1 1 1 1 1 0 1
2 1 1 1 1 1 1 1 0 1 1 1
158 1 1 1 1 1 1 1 1 0 0 2
0 0 0 0 0 0 0 2 177 687 866
Rattle timestamp: 2014-09-23

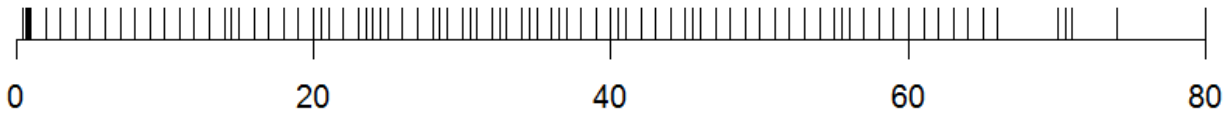

Missing Value Summary
Pclass Sex SibSp Parch Ticket Fare Survived Embarked Age Cabin
183 1 1 1 1 1 1 1 1 1 1 0
19 1 1 1 1 1 1 1 1 0 1 1
529 1 1 1 1 1 1 1 1 1 0 1
2 1 1 1 1 1 1 1 0 1 1 1
158 1 1 1 1 1 1 1 1 0 0 2
0 0 0 0 0 0 0 2 177 687 866
CENTRAL AREA
Rattle timestamp: 2014-09-23

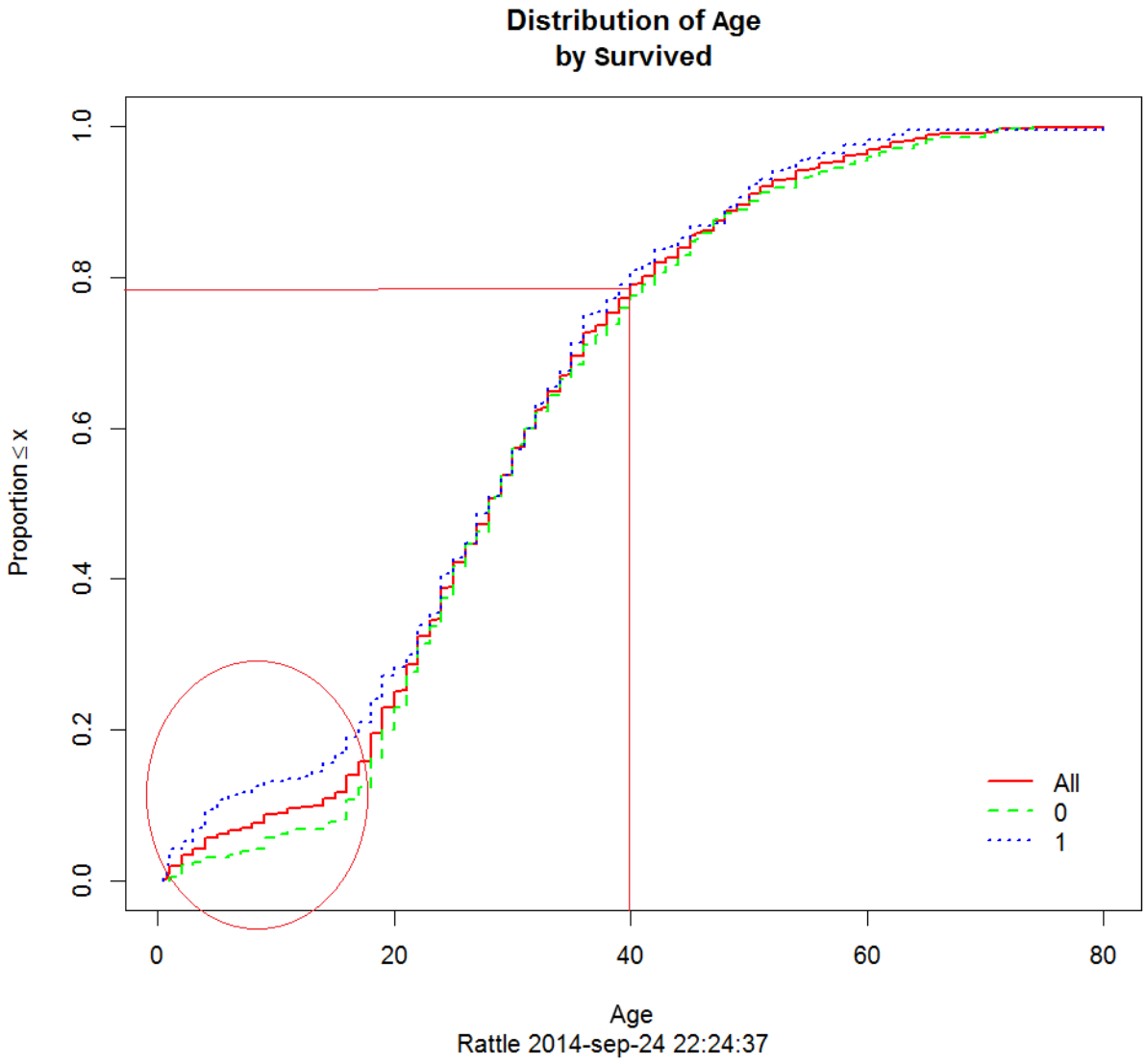

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type:
Summary
Distributions
Correlation
Principal Components
Interactive
Summary
Describe
Basics
Kurtosis
Skewness
Show Missing
Cross Tab
$age
X...X.2
nobs 623.000000
NAs 123.000000
Minimum 0.420000
Maximum 74.000000
1. Quartile 21.000000
3. Quartile 39.000000
Mean 29.817180
Median 28.000000
Sum 14908.590000
SE Mean 0.634329
LCL Mean 28.570895
UCL Mean 31.063465
Variance 201.186596
Stdev 14.184026
Skewness 0.385387
Kurtosis 0.119261

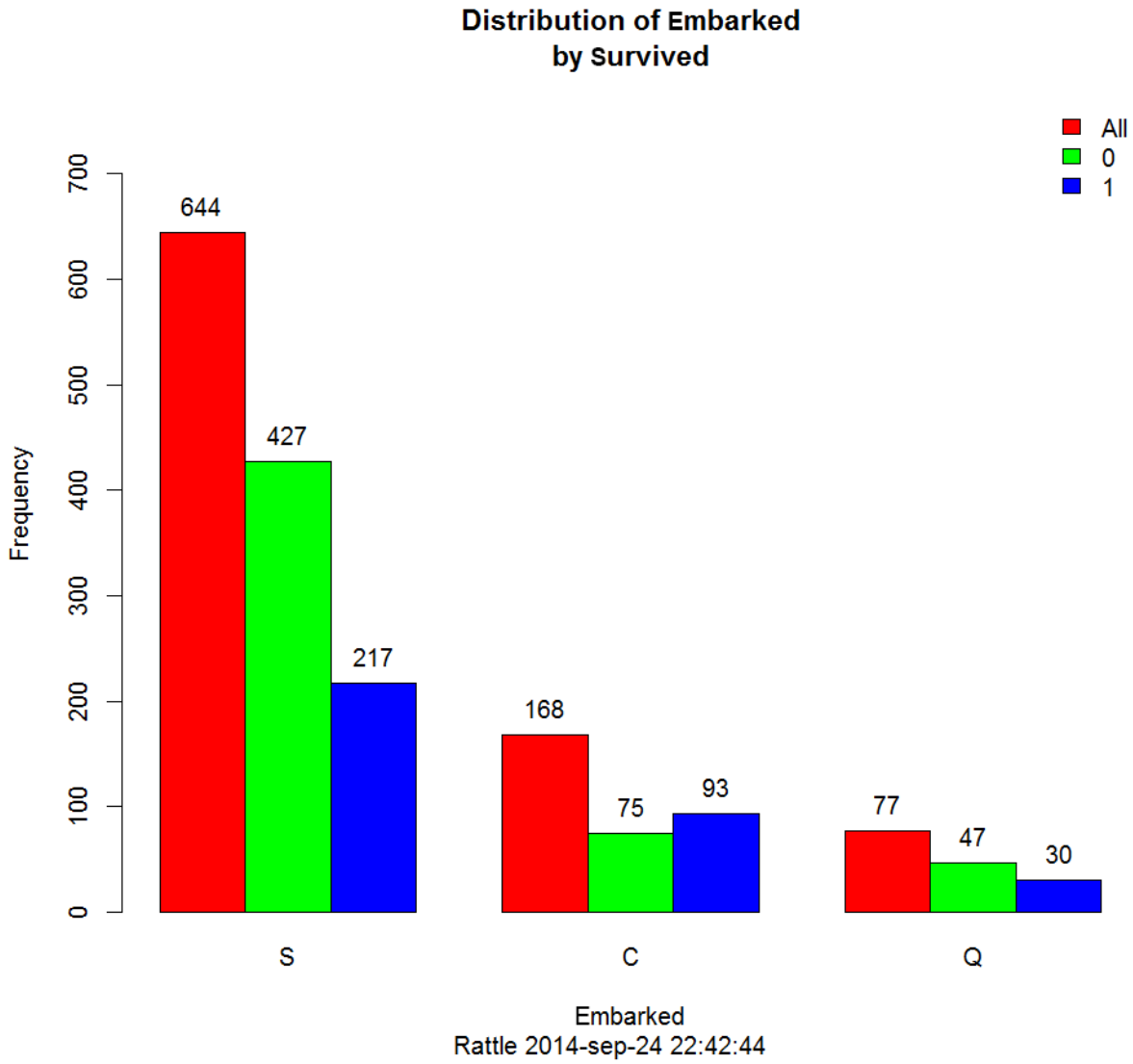

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Summary Distributions Correlation Principal Components Interactive
Numeric: Clear Plots per Page: 4 Annotate
Benfords: Bars Starting Digit: 1 Number of Digits: 1 abs +ve -ve
No. Variable Box Plot Histogram Cumulative Benford Min; Median/Mean; Max
1 Survived 0.00; 0.00/0.38; 1.00
2 Pclass 1.00; 3.00/2.31; 3.00
5 Age 0.42; 28.00/29.70; 80.00
6 SibSp 0.00; 0.00/0.52; 8.00
7 Parch 0.00; 0.00/0.38; 6.00
9 Fare 0.00; 14.45/32.20; 512.33

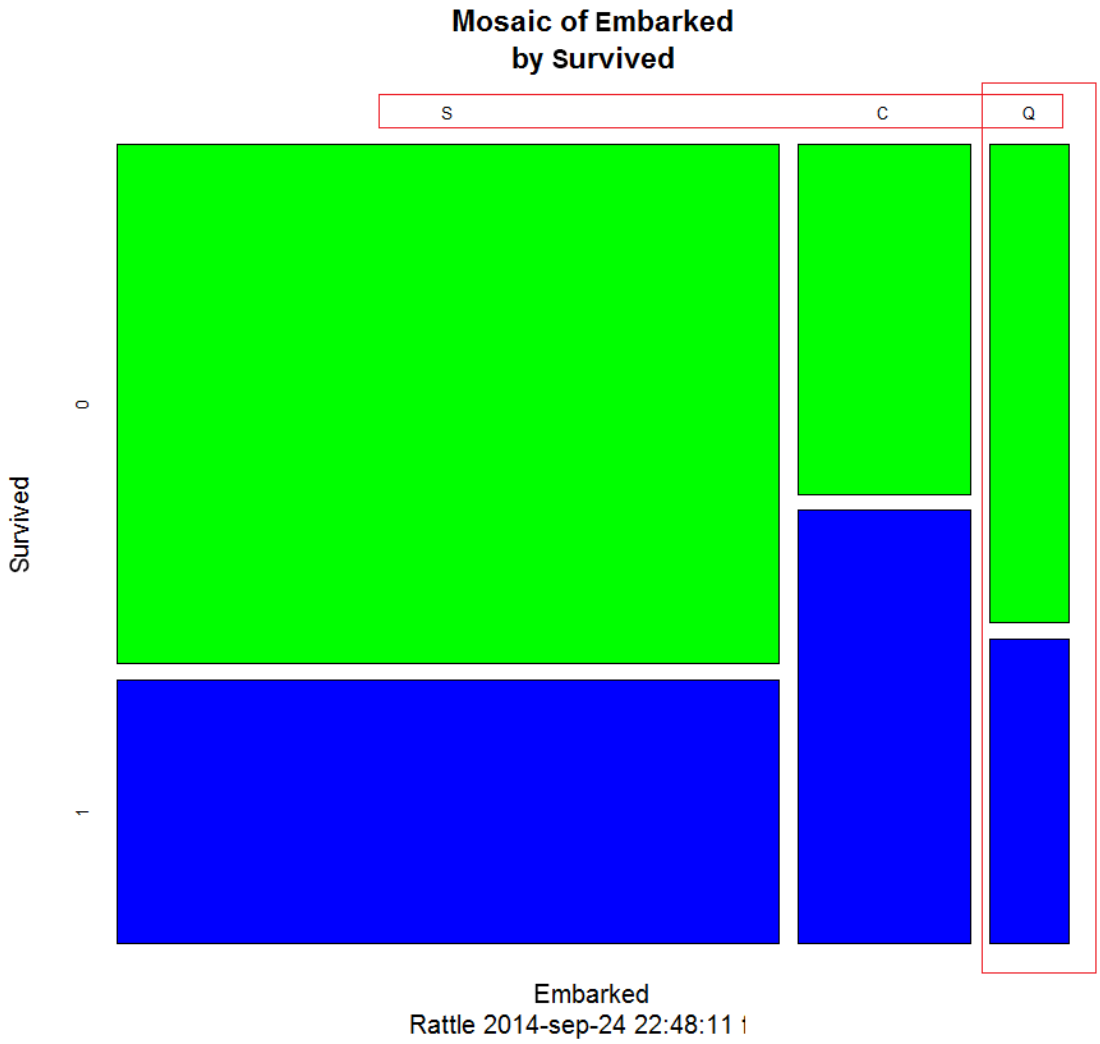

Distribution of Age
by Survived
Age
0
20
40
60
80
80
74
71
70
66
65
38
28
20
0.42
74
71
70
66
39
28
21
1
80
63
62
60
36
28
19
0.42
All
0
1
Survived
Rattle 2014-sep-24 21:30:32

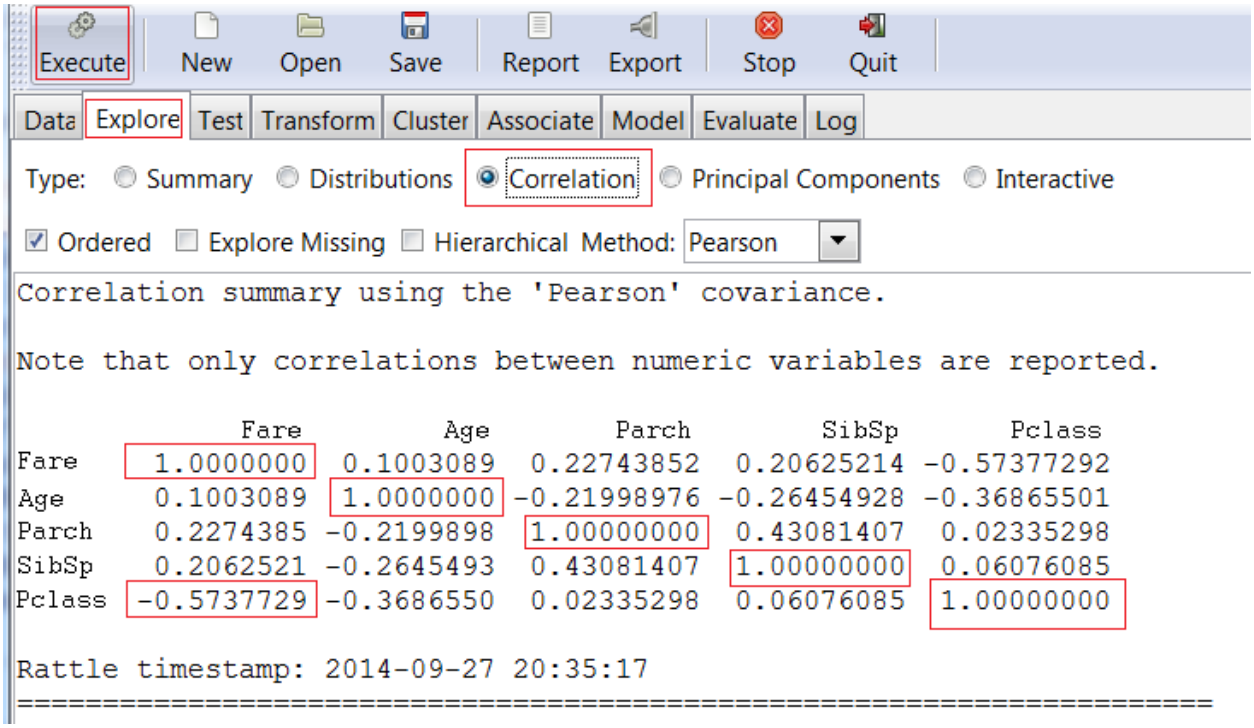

Distribution of Age
by Survived
All
0
1
Frequency
0
20
40
60
80
100
120
0
20
40
60
80
Age
Rattle 2014-sep-24 22:08:19

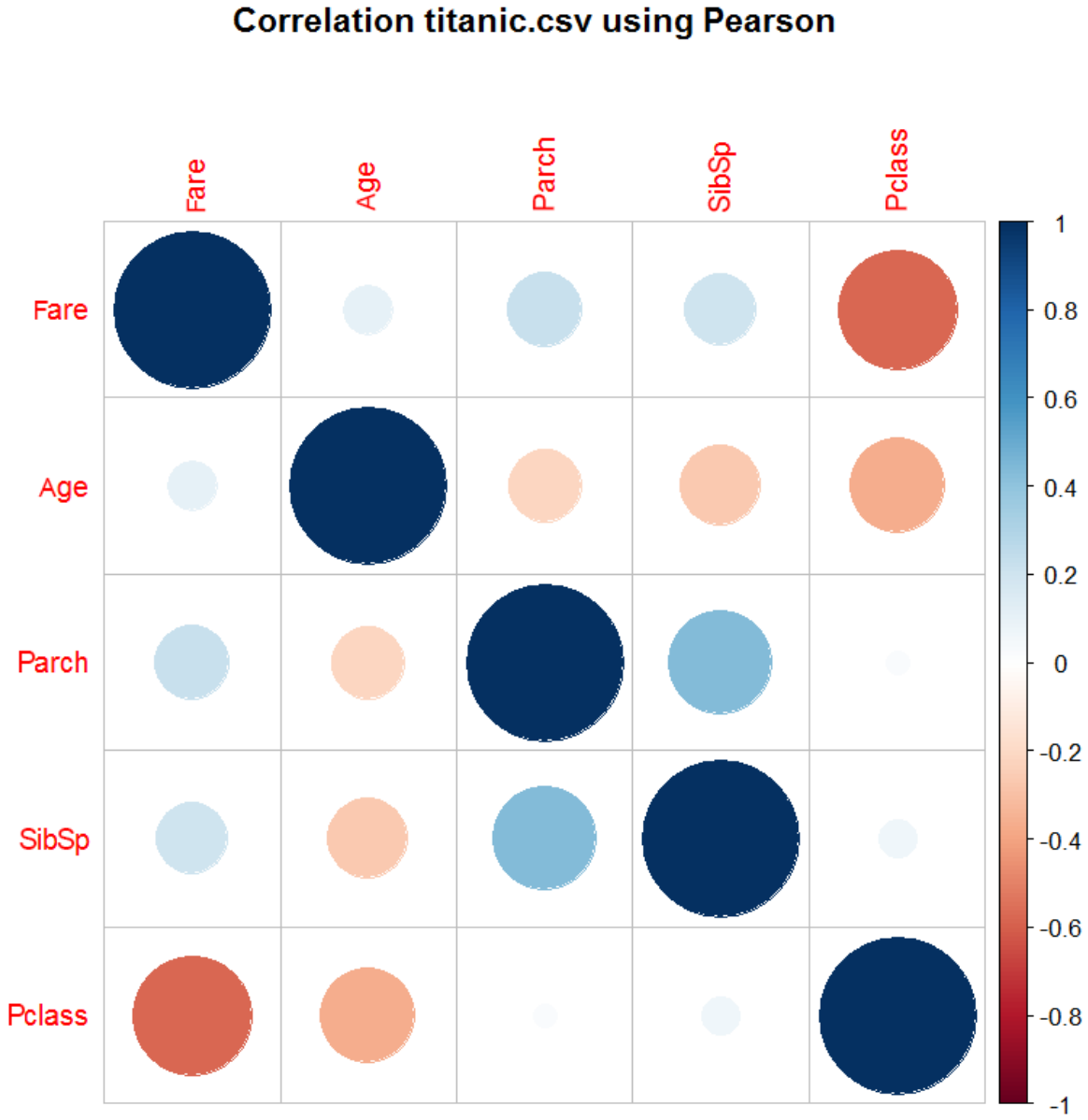

0
20
40
60
80

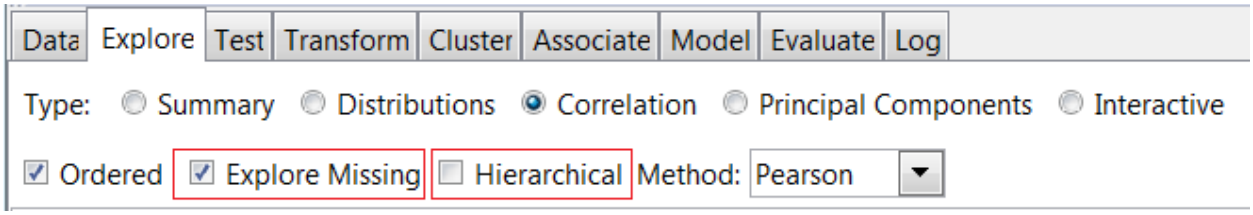

Distribution of Age
by Survived
Proportion ≤ x
0.0
0.2
0.4
0.6
0.8
1.0
0
20
40
60
80
Age
All
0
1
Rattle 2014-sep-24 22:24:37

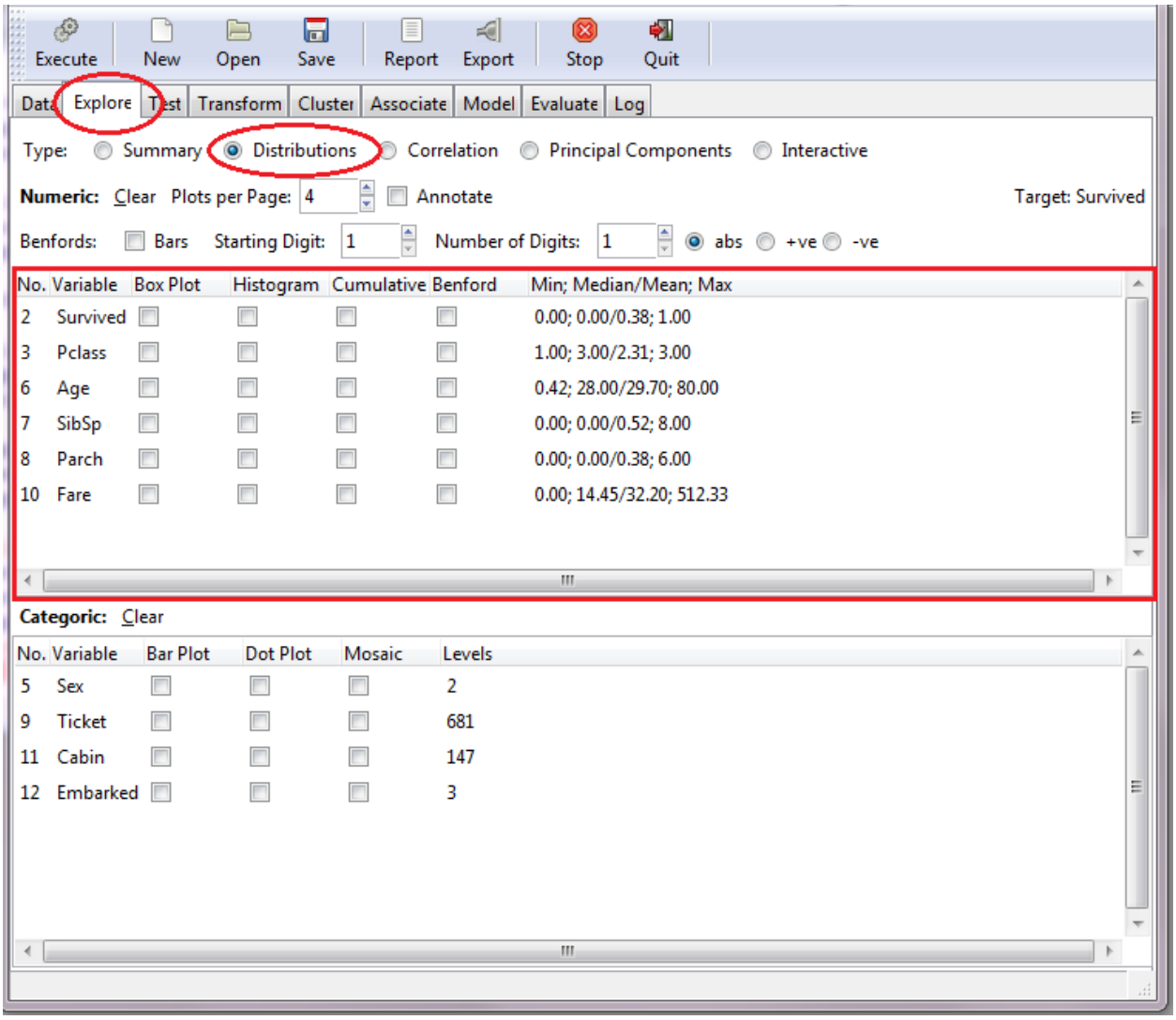

Distribution of Embarked
by Survived
All
0
1
Frequency
0
100
200
300
400
500
600
700
644
427
217
168
75
93
77
47
30
S
C
Q
Embarked
Rattle 2014-sep-24 22:42:44

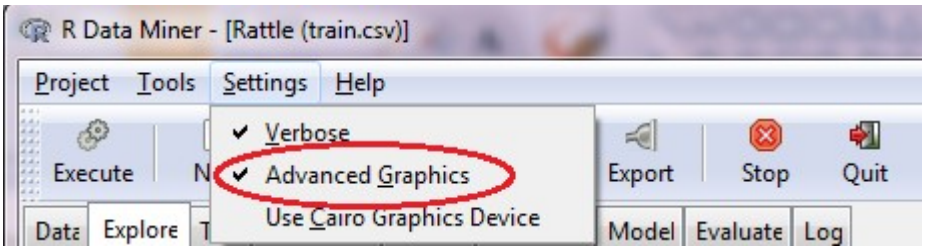

Mosaic of Embarked
by Survived
S
C
Q
Survived
0
1
Embarked
Rattle 2014-sep-24 22:48:11

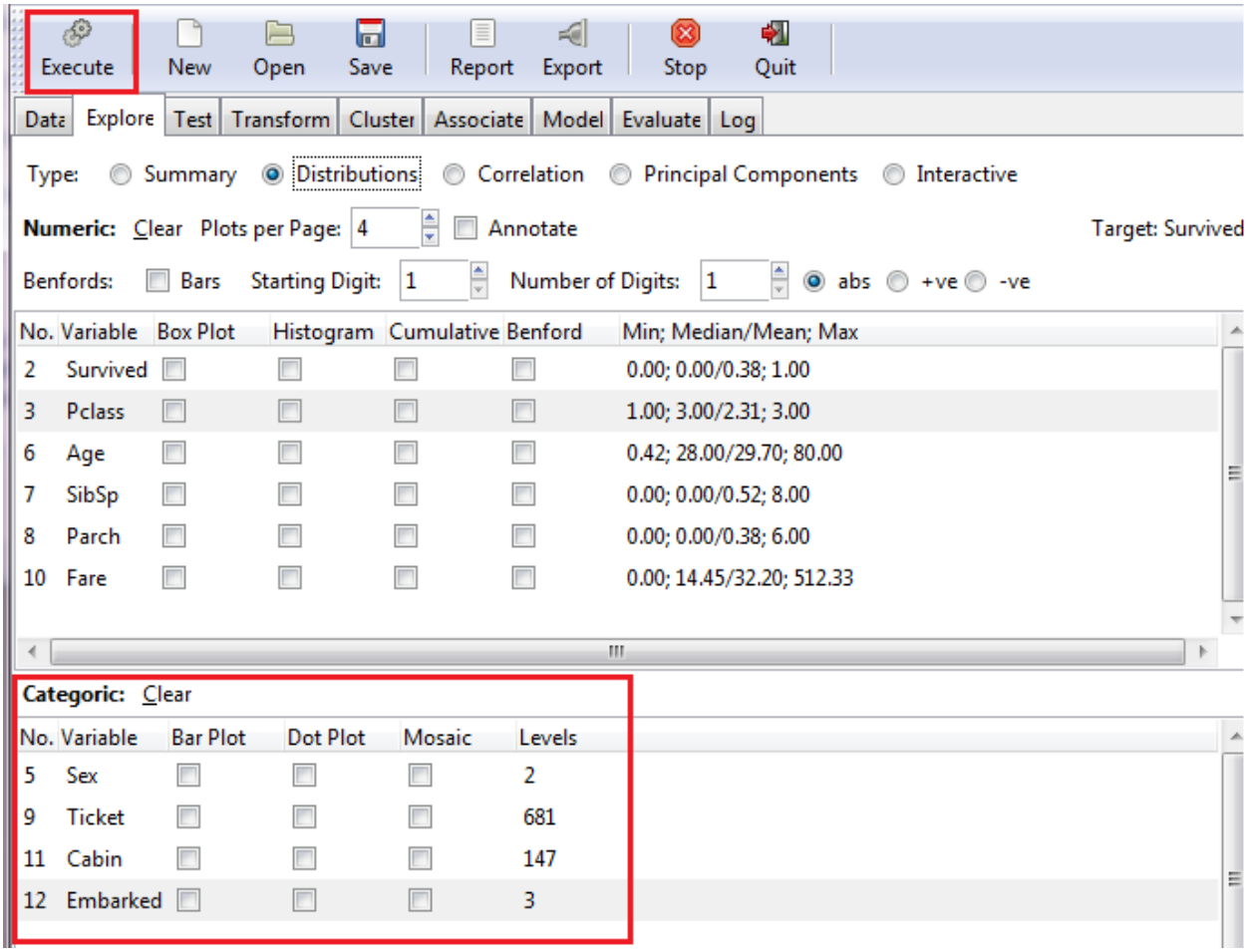

Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type:
Summary
Distributions
Correlation
Principal Components
Interactive
Ordered
Explore Missing
Hierarchical
Method:
Pearson
Correlation summary using the 'Pearson' covariance.
Note that only correlations between numeric variables are reported.
Fare Age Parch SibSp Pclass
Fare 1.0000000 0.1003089 0.22743852 0.20625214 -0.57377292
Age 0.1003089 1.0000000 -0.21998976 -0.26454928 -0.36865501
Parch 0.2274385 -0.2199898 1.00000000 0.43081407 0.02335298
SibSp 0.2062521 -0.2645493 0.43081407 1.00000000 0.06076085
Pclass -0.5737729 -0.3686550 0.02335298 0.06076085 1.00000000
Rattle timestamp: 2014-09-27 20:35:17


Correlation titanic.csv using Pearson
Fare
Age
Parch
SibSp
Pclass
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

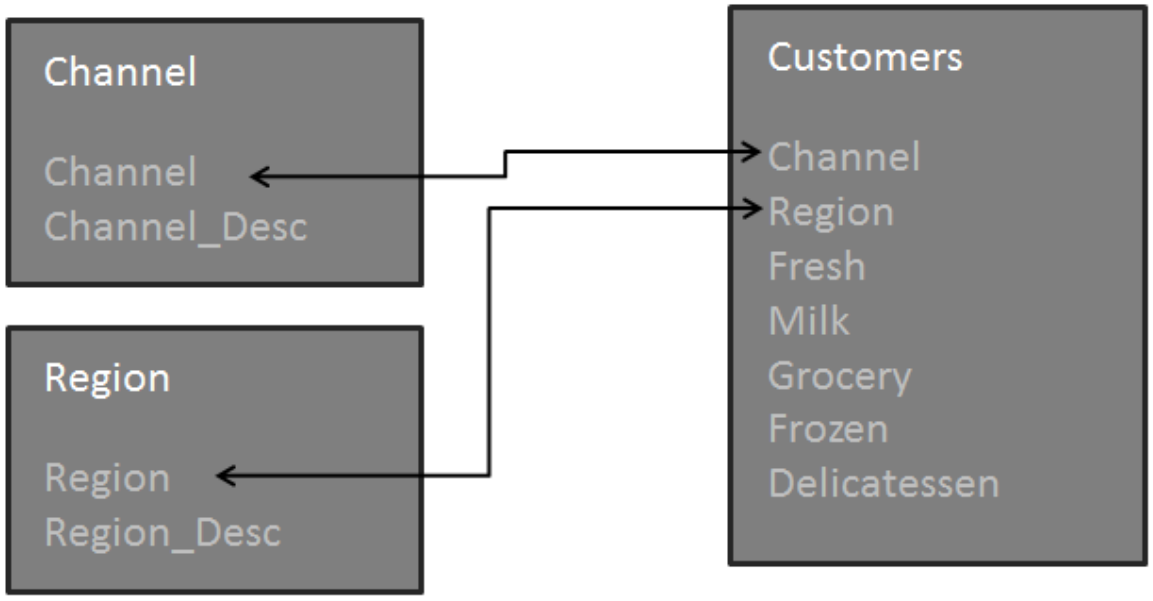

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type:
Summary
Distributions
Correlation
Principal Components
Interactive
Ordered
Explore Missing
Hierarchical
Method:
Pearson


Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Summary Distributions Correlation Principal Components Interactive
Numeric: Clear Plots per Page: 4 Annotate Target: Survived
Benfords: Bars Starting Digit: 1 Number of Digits: 1 abs +ve -ve
No. Variable Box Plot Histogram Cumulative Benford Min; Median/Mean; Max
2 Survived 0.00; 0.00/0.38; 1.00
3 Pclass 1.00; 3.00/2.31; 3.00
6 Age 0.42; 28.00/29.70; 80.00
7 SibSp 0.00; 0.00/0.52; 8.00
8 Parch 0.00; 0.00/0.38; 6.00
10 Fare 0.00; 14.45/32.20; 512.33
Categoric: Clear
No. Variable Bar Plot Dot Plot Mosaic Levels
5 Sex 2
9 Ticket 681
11 Cabin 147
12 Embarked 3


R Data Miner - [Rattle (train.csv)]
Project
Tools
Settings
Help
Verbose
Advanced Graphics
Use Cairo Graphics Device
Execute
Export
Stop
Quit
Data
Explore
Model
Evaluate
Log


Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Summary Distributions Correlation Principal Components Interactive
Numeric: Clear Plots per Page: 4 Annotate Target: Survived
Benfords: Bars Starting Digit: 1 Number of Digits: 1 abs +ve -ve
No. Variable Box Plot Histogram Cumulative Benford Min; Median/Mean; Max
2 Survived 0.00; 0.00/0.38; 1.00
3 Pclass 1.00; 3.00/2.31; 3.00
6 Age 0.42; 28.00/29.70; 80.00
7 SibSp 0.00; 0.00/0.52; 8.00
8 Parch 0.00; 0.00/0.38; 6.00
10 Fare 0.00; 14.45/32.20; 512.33
Categoric: Clear
No. Variable Bar Plot Dot Plot Mosaic Levels
5 Sex 2
9 Ticket 681
11 Cabin 147
12 Embarked 3

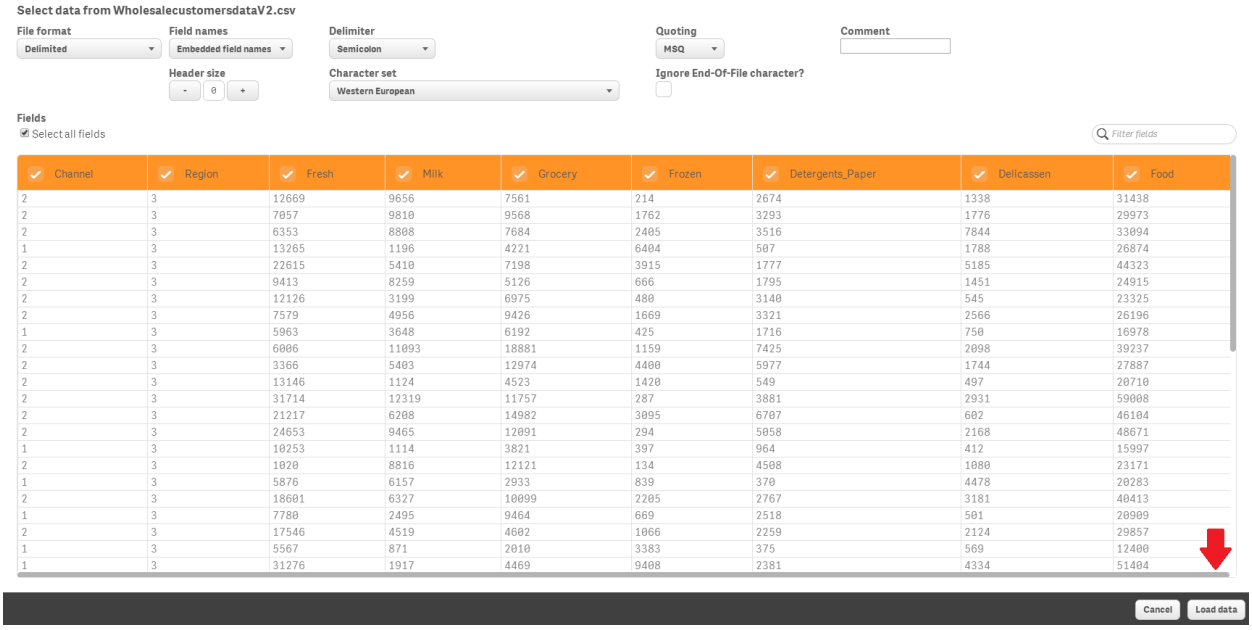
Chapter 4: Creating Your First Qlik Sense Application

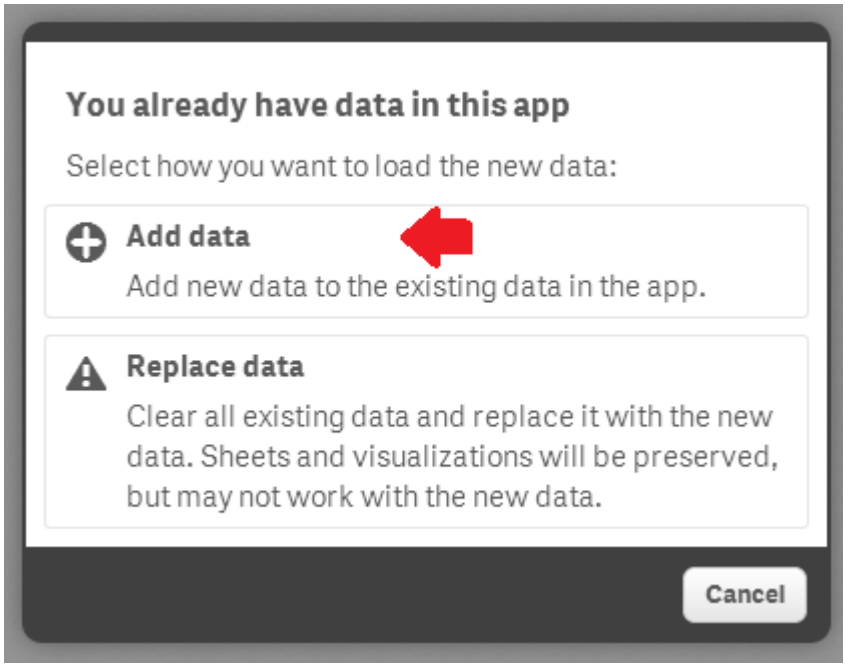


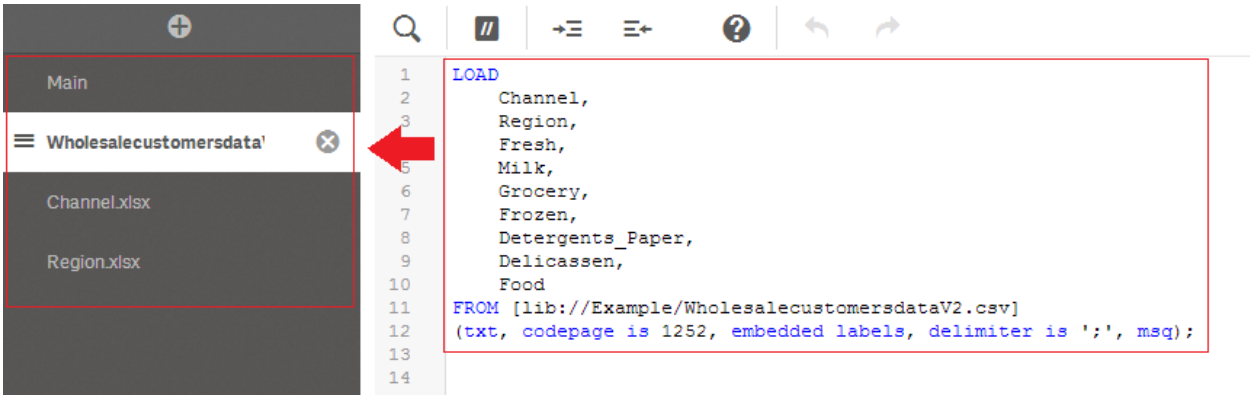
Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
2	3	12669	9656	7561	214	2674	1338
2	3	7057	9810	9568	1762	3293	1776
2	3	6353	8808	7684	2405	3516	7844
1	3	13265	1196	4221	6404	507	1788
2	3	22615	5410	7198	3915	1777	5185
2	3	9413	8259	5126	666	1795	1451

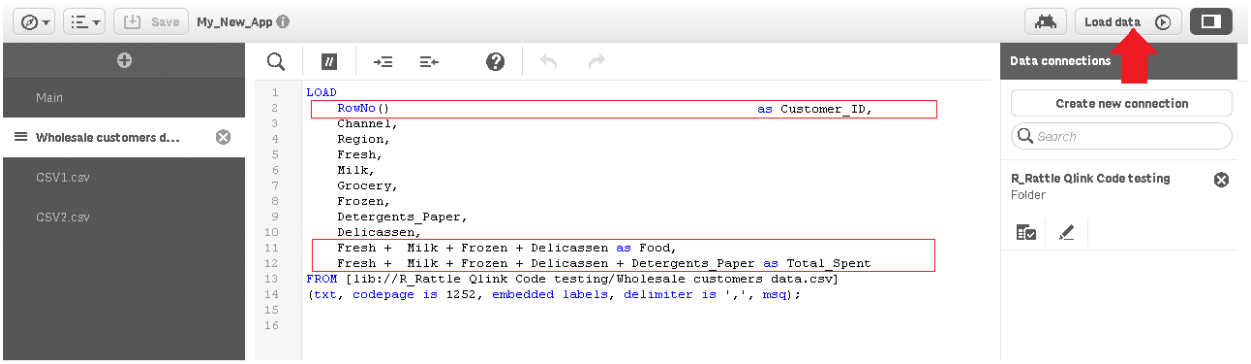
Channel	Channel_Desc
1	Horeca
2	Retail

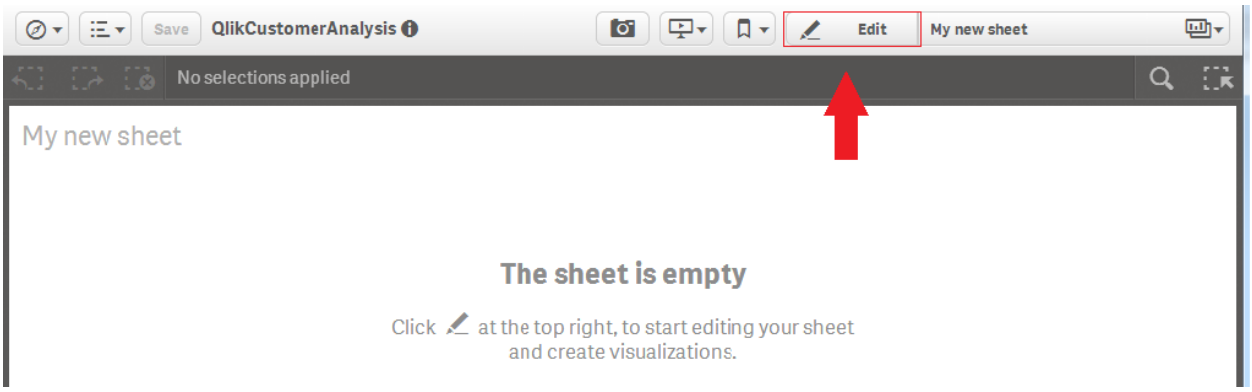
Region	Region_Desc
1	Lisbon
2	Porto
3	Other

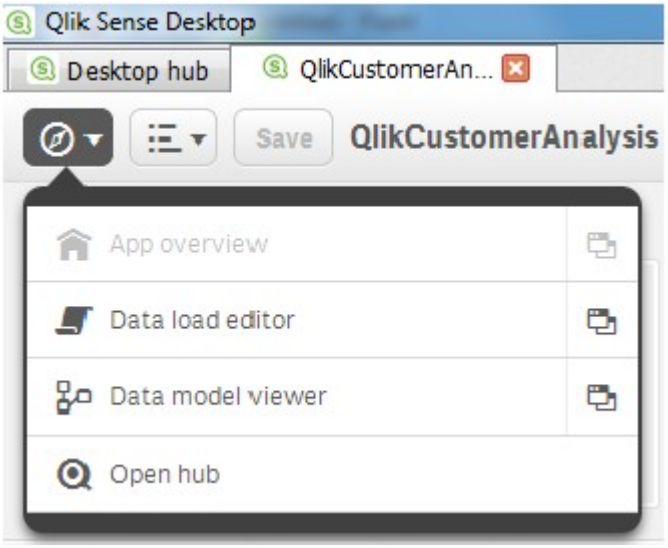



You already have data in this app
Select how you want to load the new data:
Add data
Add new data to the existing data in the app.
Replace data
Clear all existing data and replace it with the new data. Sheets and visualizations will be preserved, but may not work with the new data.
Cancel

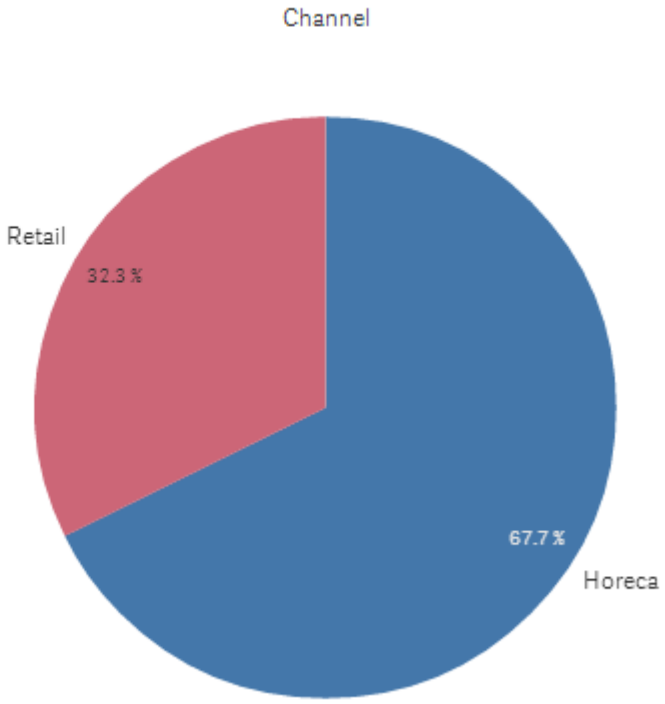

Main
Wholesalecustomersdata'
Channel.xlsx
Region.xlsx
LOAD
Channel,
Region,
Fresh,
Milk,
Grocery,
Frozen,
Detergents_Paper,
Delicassen,
Food
FROM [lib://Example/WholesalecustomersdataV2.csv]
(txt, codepage is 1252, embedded labels, delimiter is ';', msq);

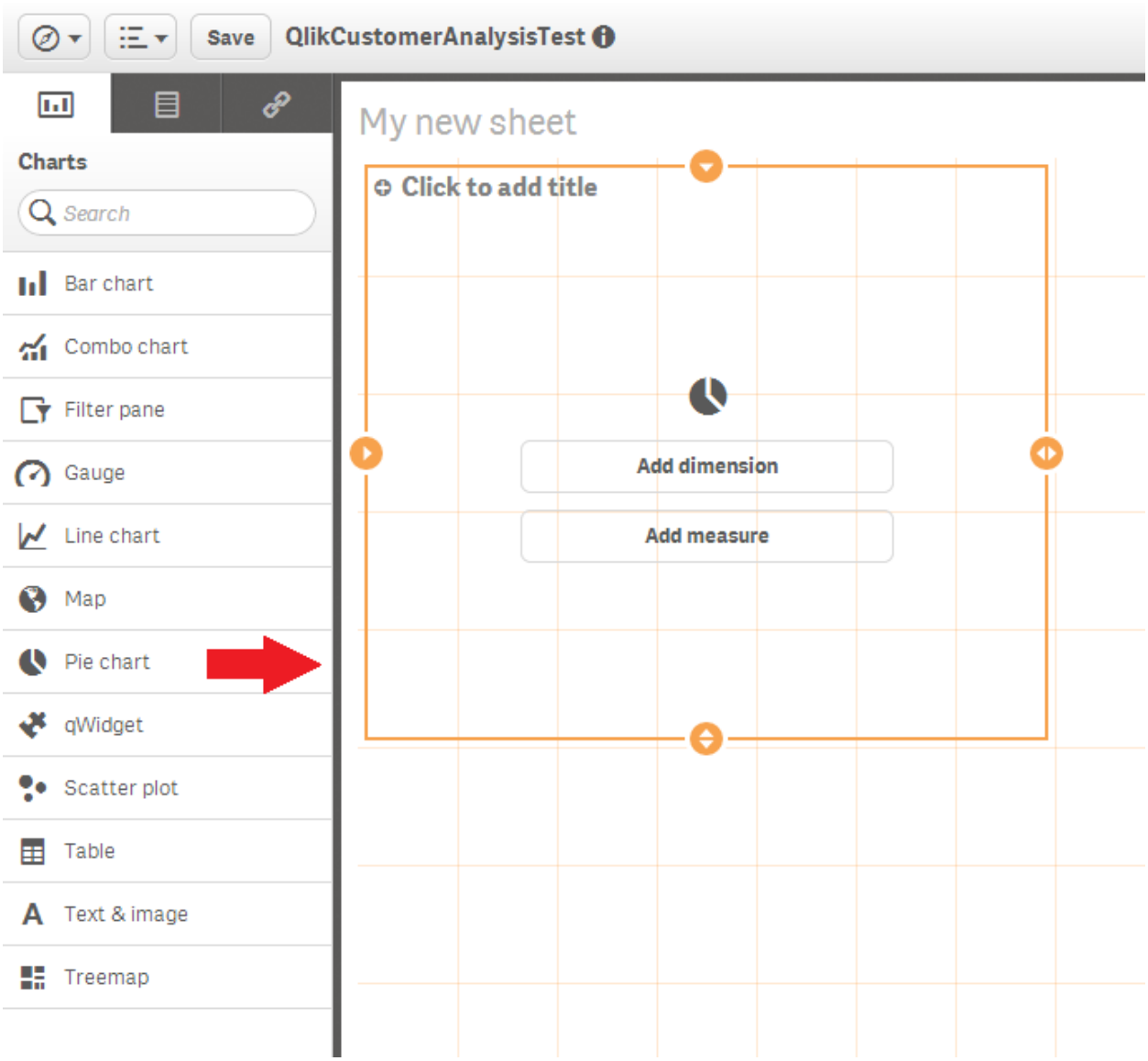

Save
My_New_App
Load data
Main
Wholesale customers d...
CSV1.csv
CSV2.csv
Data connections
Create new connection
Search
R_Rattle Qlink Code testing
Folder
LOAD
RowNo() as Customer_ID,
Channel,
Region,
Fresh,
Milk,
Grocery,
Frozen,
Detergents_Paper,
Delicassen,
Fresh + Milk + Frozen + Delicassen as Food,
Fresh + Milk + Frozen + Delicassen + Detergents_Paper as Total_Spent
FROM [lib://R_Rattle Qlink Code testing/Wholesale customers data.csv]
(txt, codepage is 1252, embedded labels, delimiter is ',', msq);

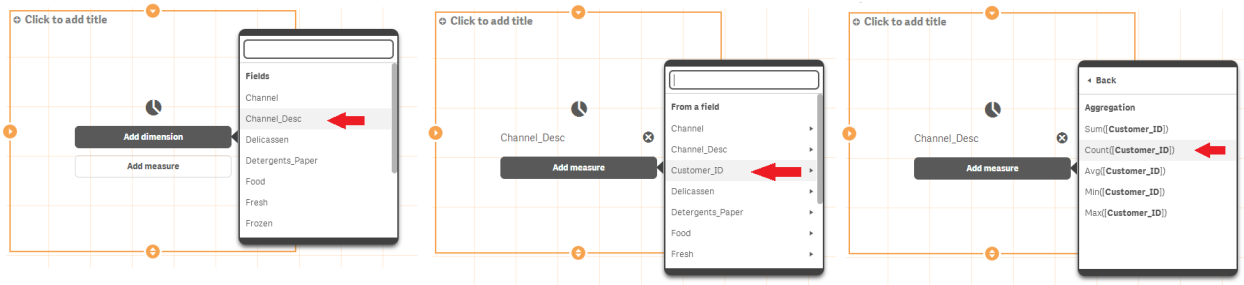

Save
QlikCustomerAnalysis
Edit
My new sheet
No selections applied
My new sheet
The sheet is empty
Click at the top right, to start editing your sheet
and create visualizations.


Qlik Sense Desktop
Desktop hub
QlikCustomerAn...
Save
QlikCustomerAnalysis
App overview
Data load editor
Data model viewer
Open hub

Channel Distribution



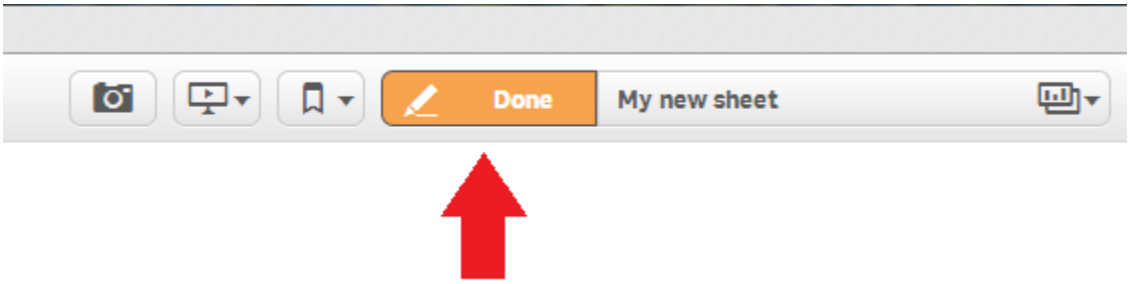

Save
QlikCustomerAnalysisTest
Charts
Search
Bar chart
Combo chart
Filter pane
Gauge
Line chart
Map
Pie chart
qWidget
Scatter plot
Table
Text & image
Treemap
My new sheet
Click to add title
Add dimension
Add measure



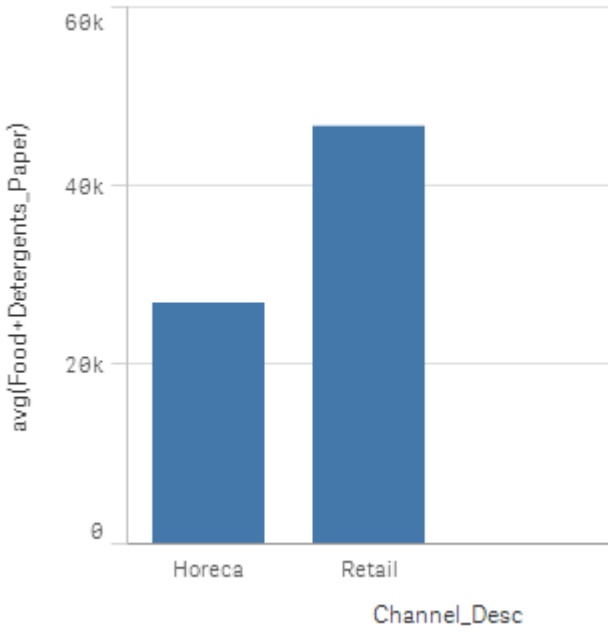
Select a dimension

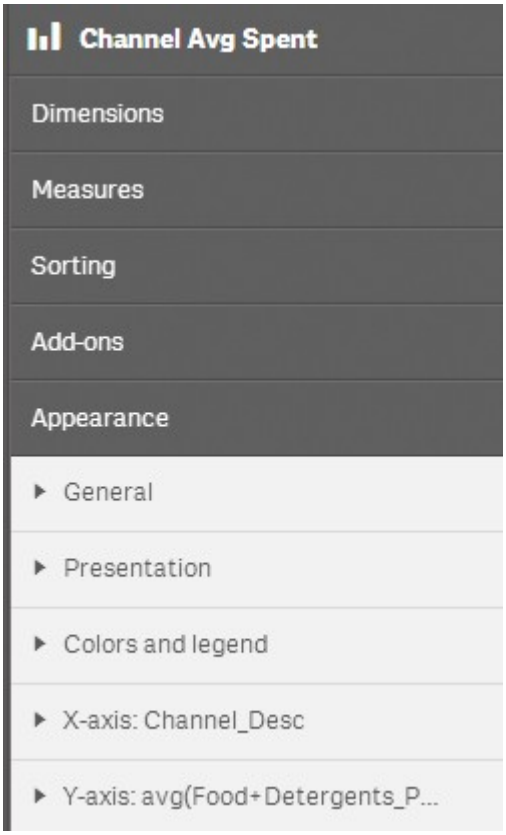
Select a measure

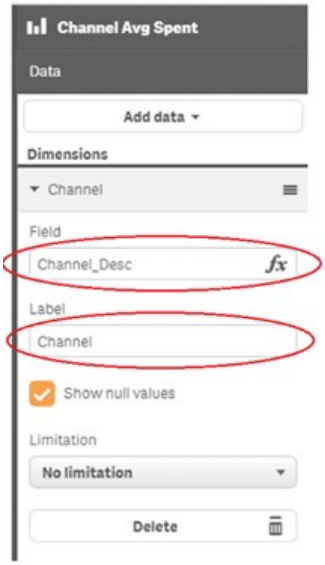
Select an operation



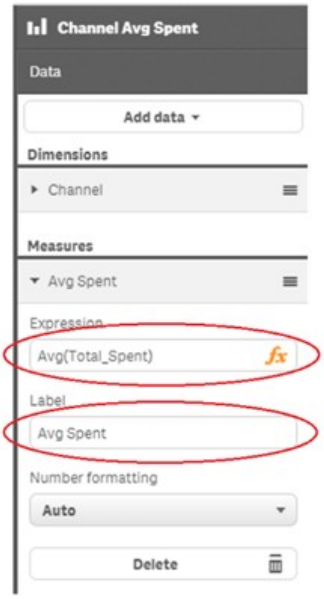
Channel Avg Spent



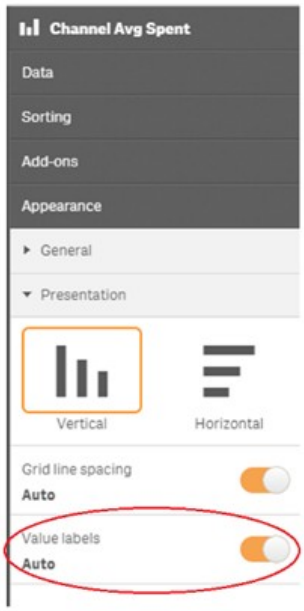

Channel Avg Spent
Dimensions
Measures
Sorting
Add-ons
Appearance
General
Presentation
Colors and legend
X-axis: Channel_Desc
Y-axis: avg(Food+Detergents_P...



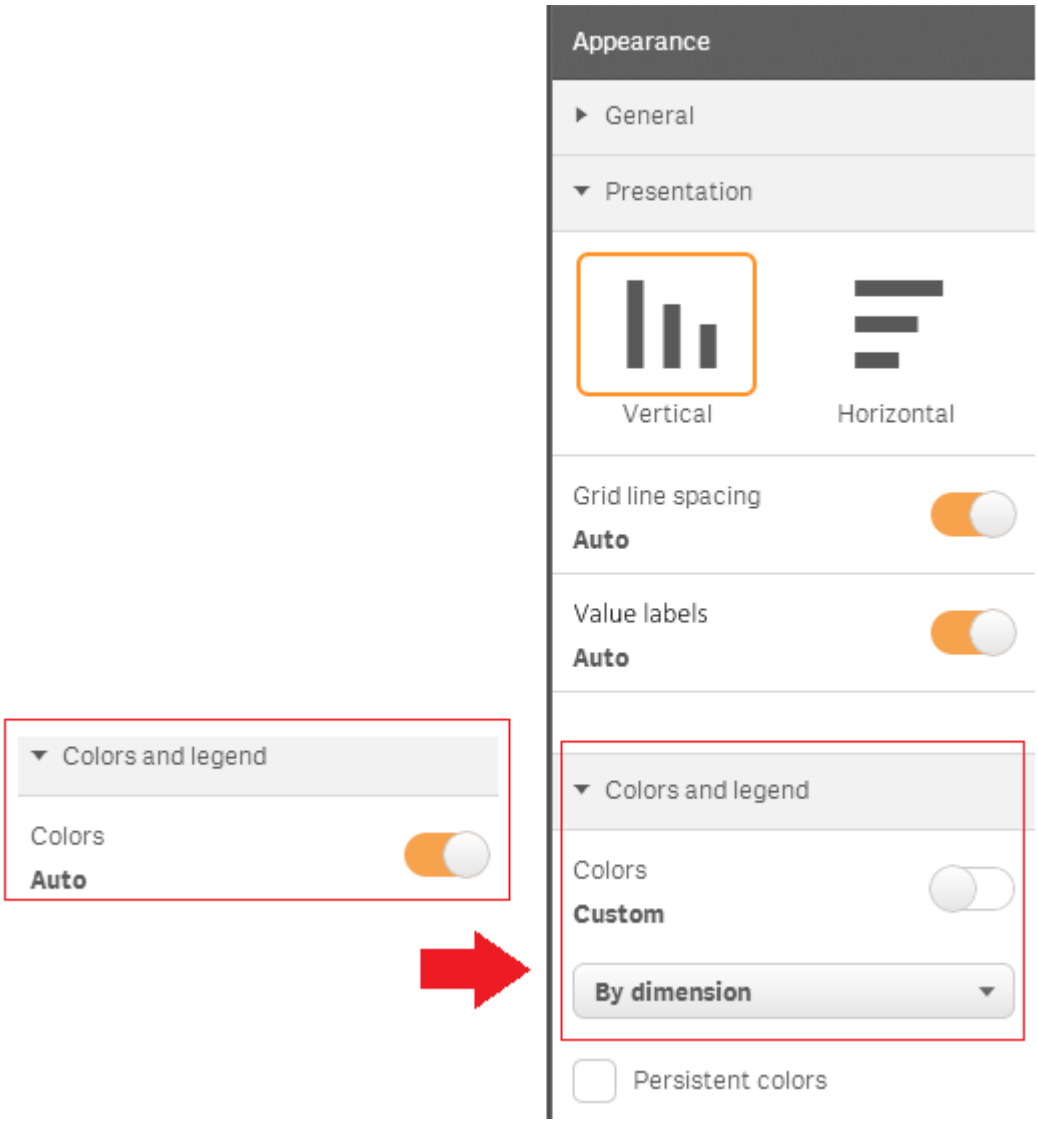
1 – Use Channel_Desc as dimension and label it 'Channel'.



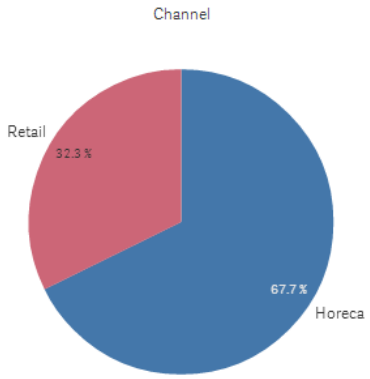
2 – Create a measure and label it 'Avg Spent'.



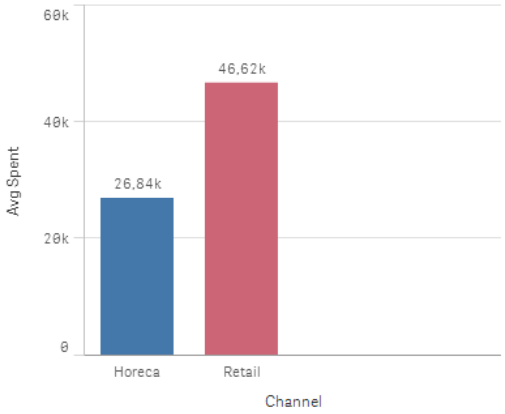
3 – Turn on 'Value labels'.

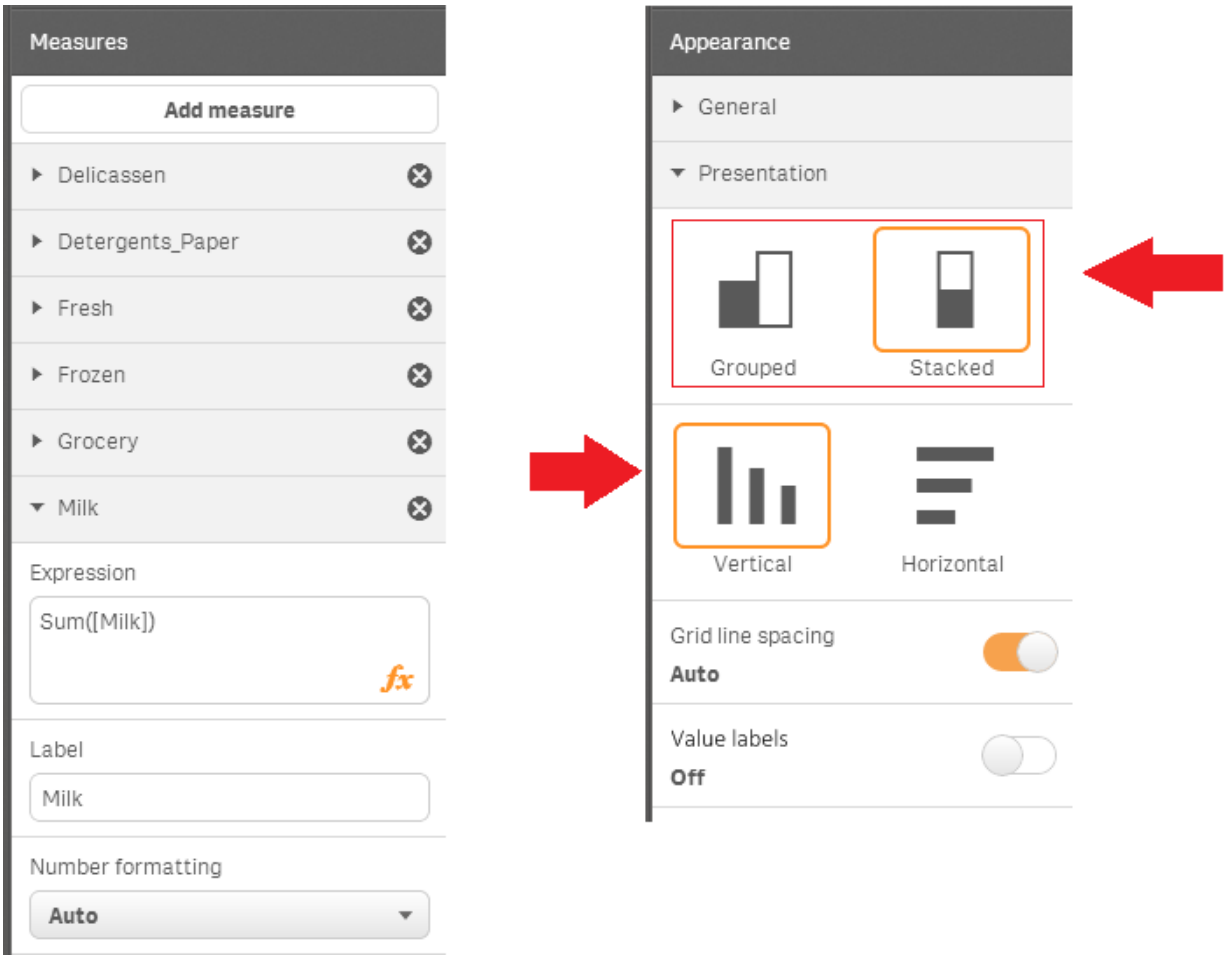

Appearance
General
Presentation
Vertical
Horizontal
Grid line spacing
Auto
Value labels
Auto
Colors and legend
Colors
Auto
Colors and legend
Colors
Custom
By dimension
Persistent colors

% Customers by Channel

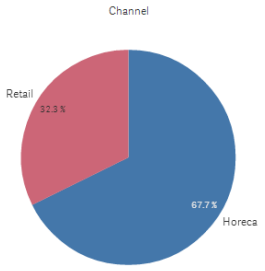


Avg Spent by Channel

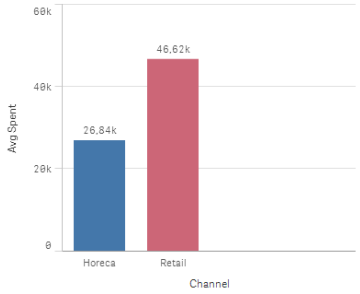



Measures
Add measure
Delicassen
Detergents_Paper
Fresh
Frozen
Grocery
Milk
Expression
Sum([Milk])
Label
Milk
Number formatting
Auto
Appearance
General
Presentation
Grouped
Stacked
Vertical
Horizontal
Grid line spacing
Auto
Value labels
Off

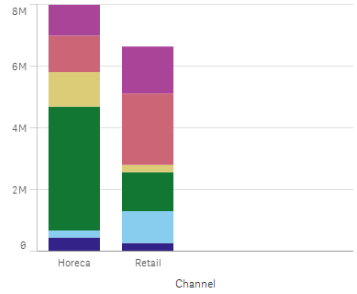
% Customers by Channel

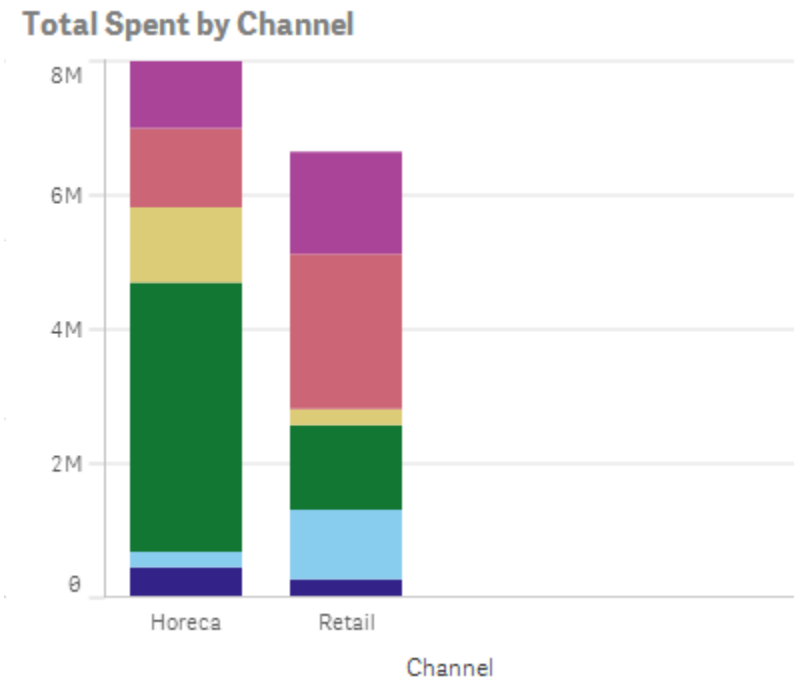


Avg Spent by Channel

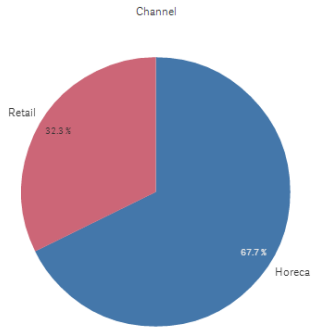


Total Spent by Channel

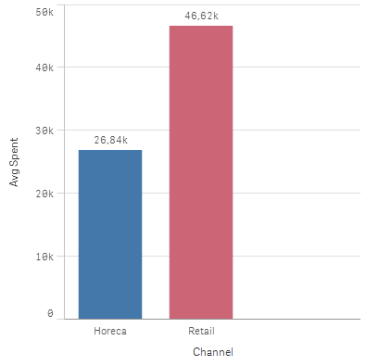




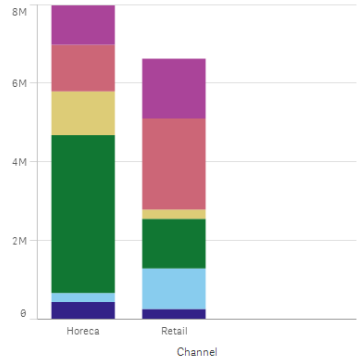
% Customers by Channel



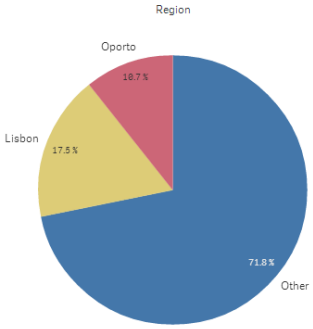
Avg Spent by Channel



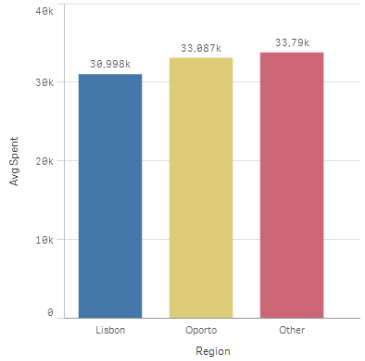
Total Spent by Channel



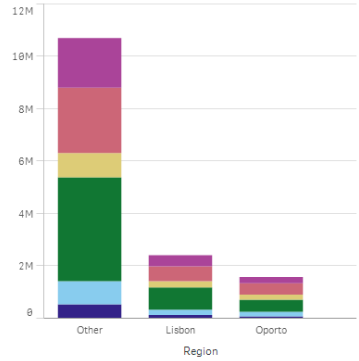
% Customers by Region

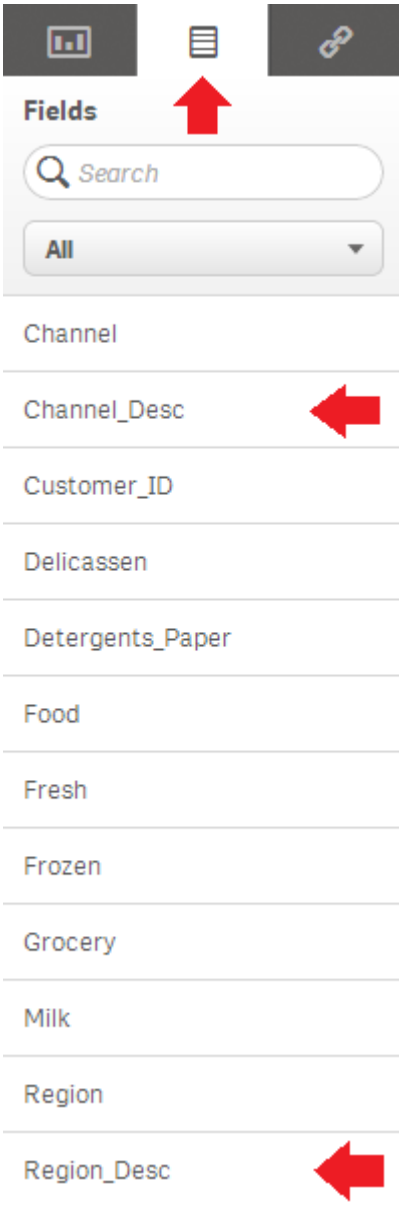


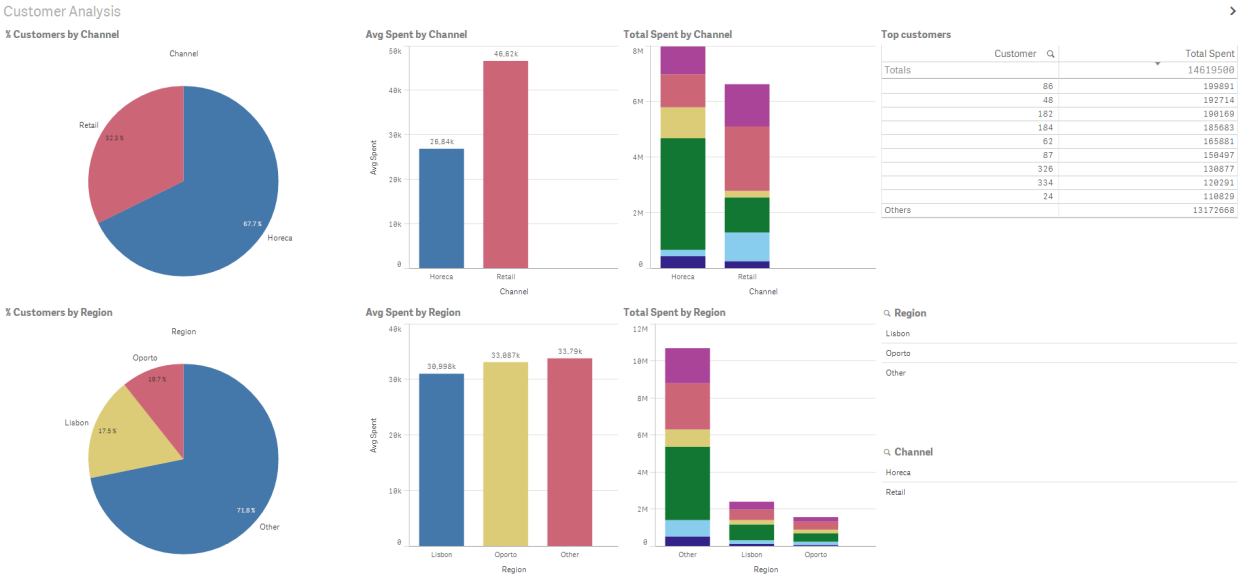
Avg Spent by Region

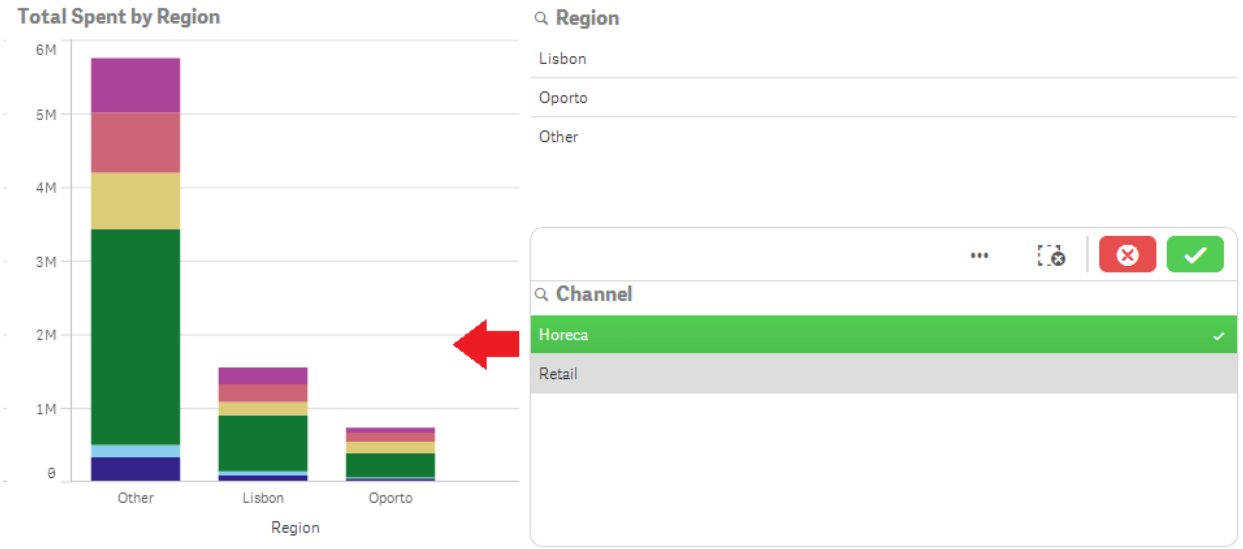


Total Spent by Region




Fields
Search
All
Channel
Channel_Desc
Customer_ID
Delicassen
Detergents_Paper
Food
Fresh
Frozen
Grocery
Milk
Region
Region_Desc

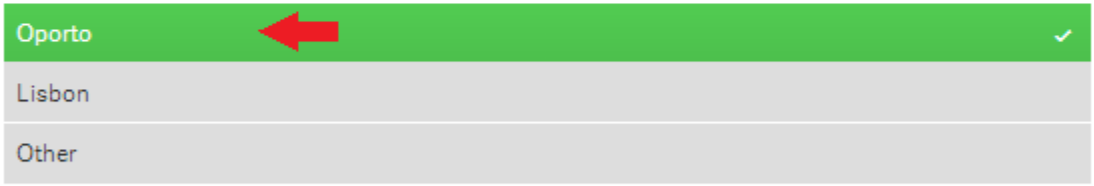

Customer Analysis
% Customers by Channel
Channel
Retail
32.3 %
67.7 %
Horeca
Avg Spent by Channel
50k
40k
30k
20k
10k
0
Avg Spent
26.84k
46.62k
Horeca
Retail
Channel
Total Spent by Channel
8M
6M
4M
2M
0
Horeca
Retail
Channel
Top customers
Customer
Total Spent
Totals
14619500
86 199891
48 192714
182 190169
184 185683
62 165881
87 150497
326 130877
334 120291
24 110829
Others 13172668
% Customers by Region
Region
Oporto
10.7 %
Lisbon
17.5 %
71.8 %
Other
Avg Spent by Region
40k
30k
20k
10k
0
Avg Spent
30,998k
33,887k
33,79k
Lisbon
Oporto
Other
Region
Total Spent by Region
12M
10M
8M
6M
4M
2M
0
Other
Lisbon
Oporto
Region
Region
Lisbon
Oporto
Other
Channel
Horeca
Retail


Total Spent by Region
6M
5M
4M
3M
2M
1M
0
Other
Lisbon
Oporto
Region
Region
Lisbon
Oporto
Other
Channel
Horeca
Retail

Top customers

Customer	Total Spent
Totals	719150
326	130877
312	43784
325	35629
295	33410
329	32124
333	31972
324	29064
297	28687
323	25311
Others	328292

Region

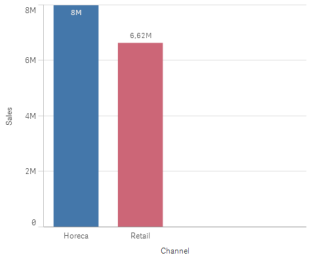


Channel

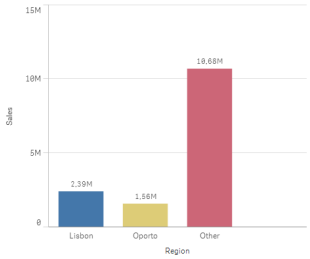
Channel
Channel
Channel_Desc
Region
Region
Region_Desc
Wholesale customers data
Channel
Region
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen

360° Analysis

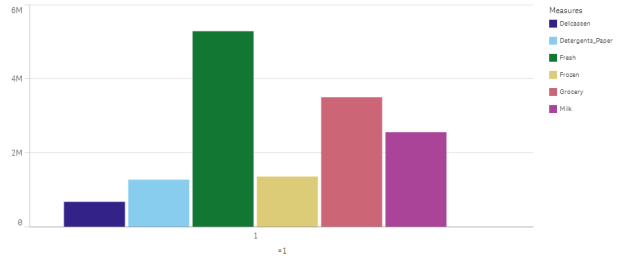
Total Sales by Channel



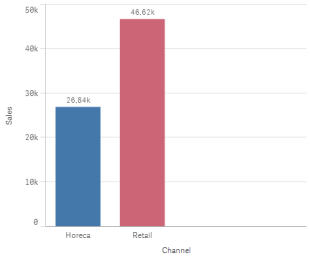
Total Sales by Region



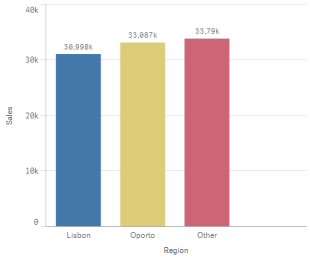
Total Spent by Product



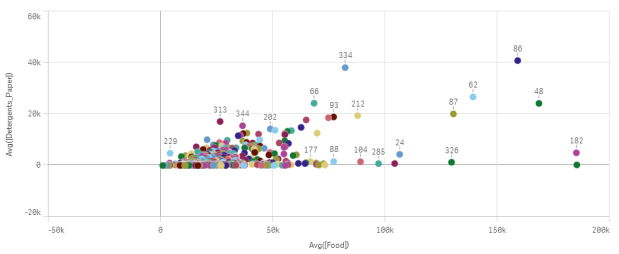
Avg Sales by Channel

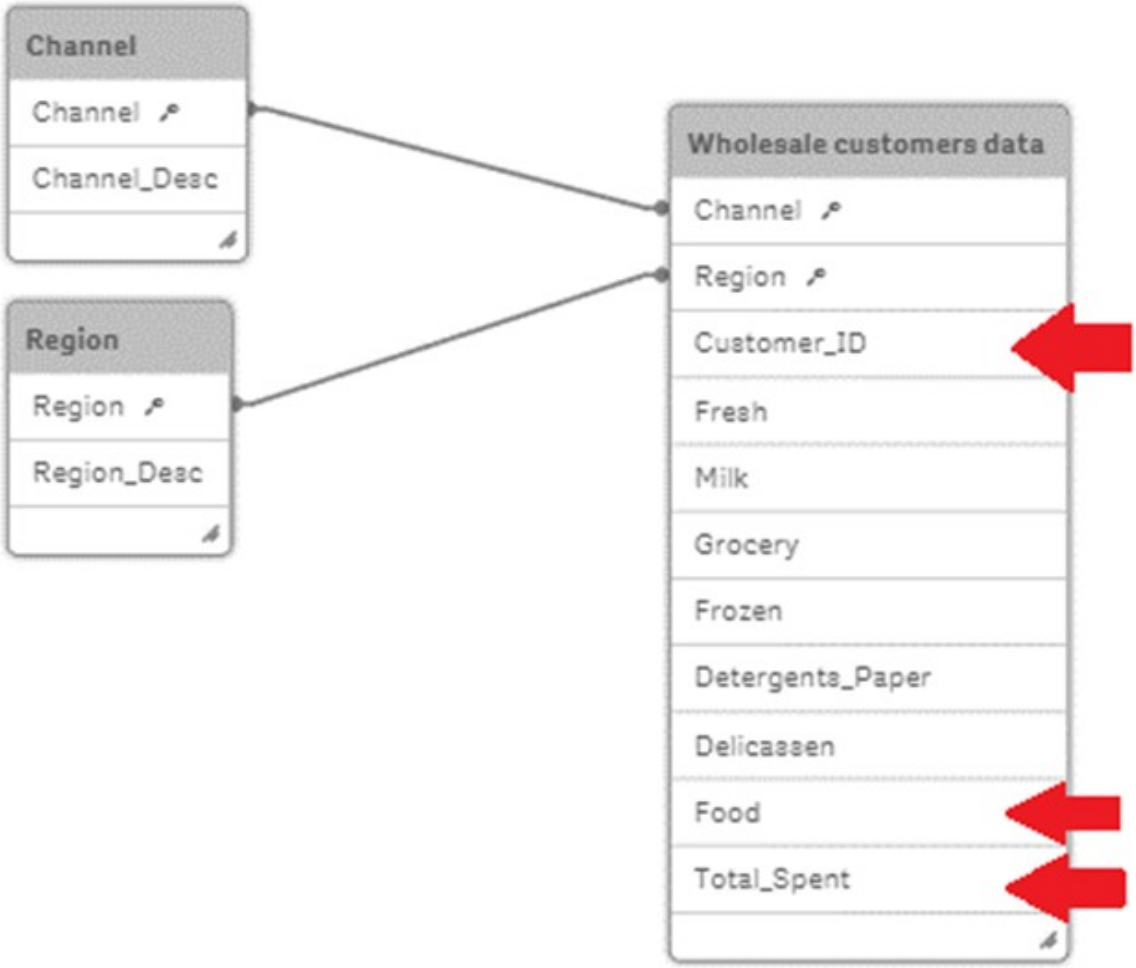


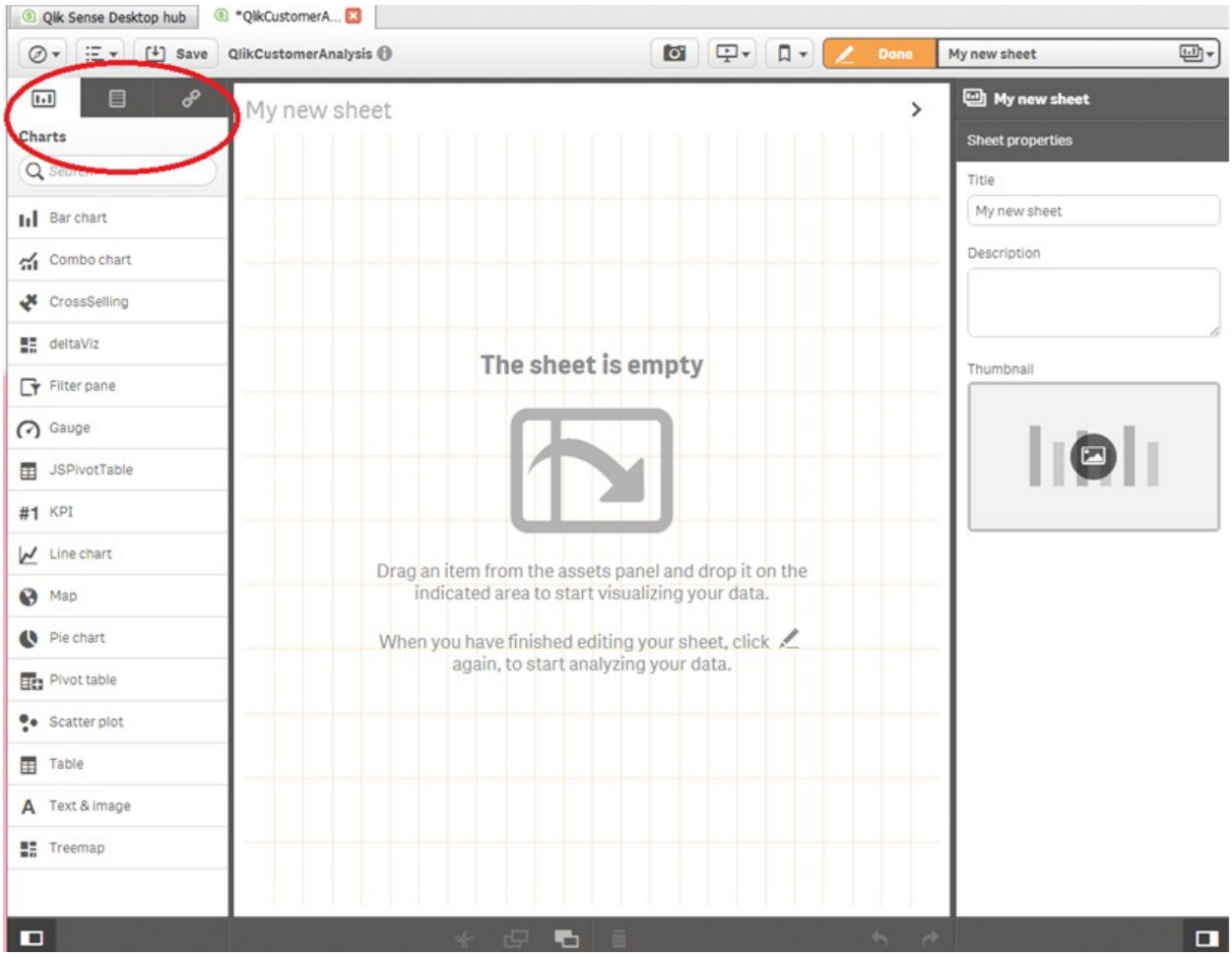
Avg Sales by Region



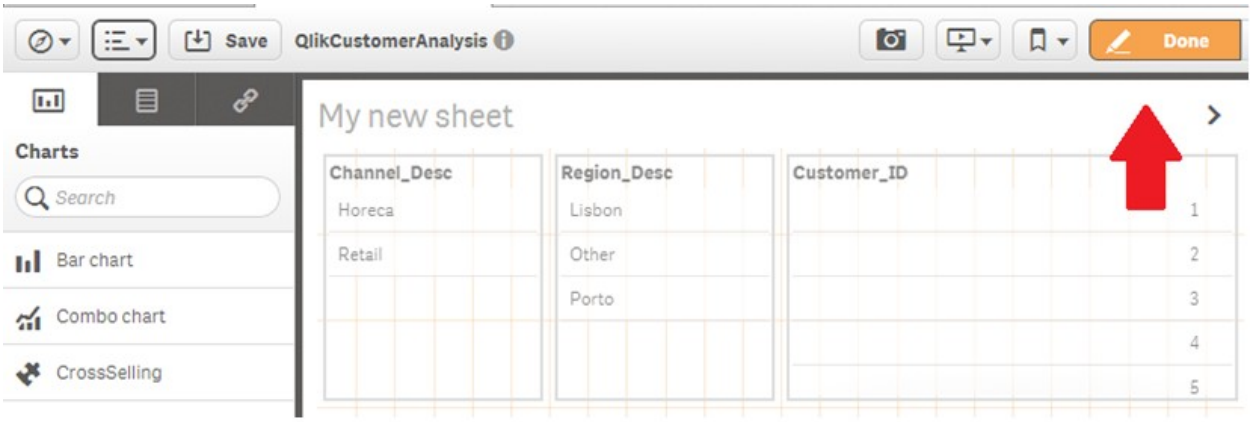
Customer's Behaviour

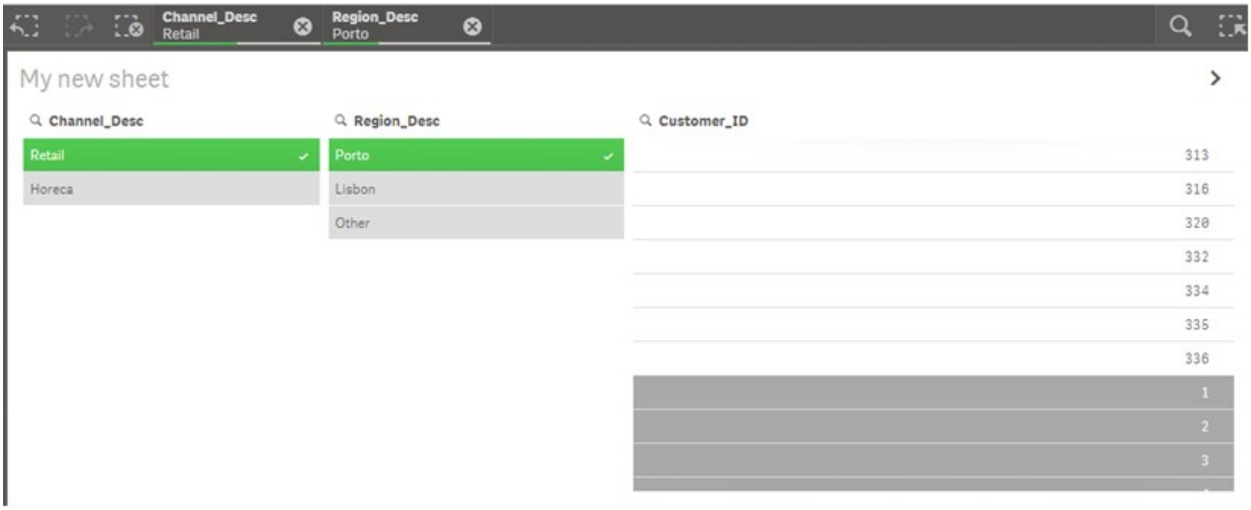



Channel
Channel
Channel_Desc
Region
Region
Region_Desc
Wholesale customers data
Channel
Region
Customer_ID
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
Food
Total_Spent

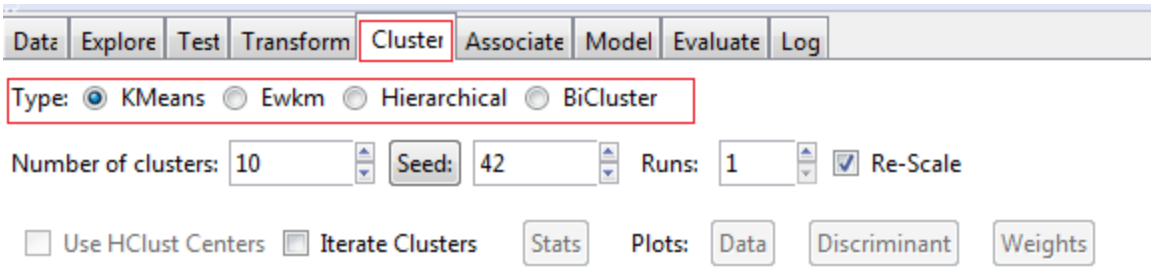

Qlik Sense Desktop hub
*QlikCustomerA...
Save
QlikCustomerAnalysis
Done
My new sheet
Charts
Bar chart
Combo chart
CrossSelling
deltaViz
Filter pane
Gauge
JSPivotTable
KPI
Line chart
Map
Pie chart
Pivot table
Scatter plot
Table
Text & image
Treemap
My new sheet
The sheet is empty
Drag an item from the assets panel and drop it on the indicated area to start visualizing your data.
When you have finished editing your sheet, click again, to start analyzing your data.
My new sheet
Sheet properties
Title
My new sheet
Description
Thumbnail

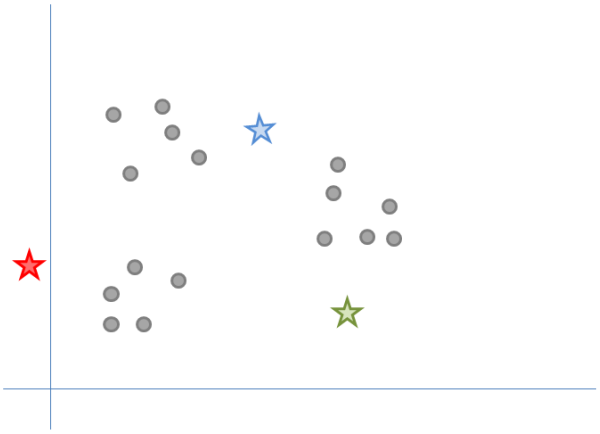
Horeca
Retail


Save
QlikCustomerAnalysis
Done
Charts
Search
Bar chart
Combo chart
CrossSelling
My new sheet
Channel_Desc
Horeca
Retail
Region_Desc
Lisbon
Other
Porto
Customer_ID
1
2
3
4
5

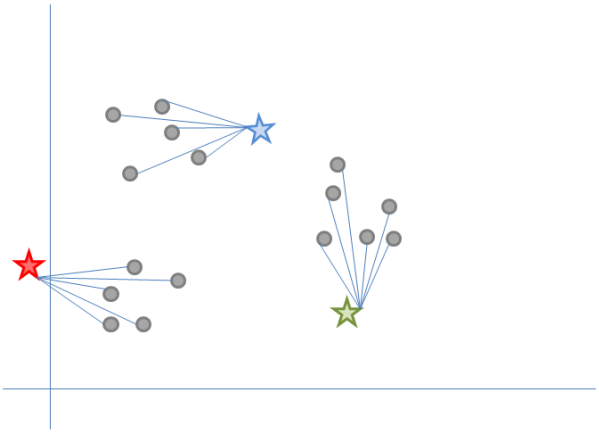

Channel_Desc
Retail
Region_Desc
Porto
My new sheet
Channel_Desc
Retail
Horeca
Region_Desc
Porto
Lisbon
Other
Customer_ID
313
316
320
332
334
335
336
1
2
3

Chapter 5: Clustering and Other Unsupervised Learning

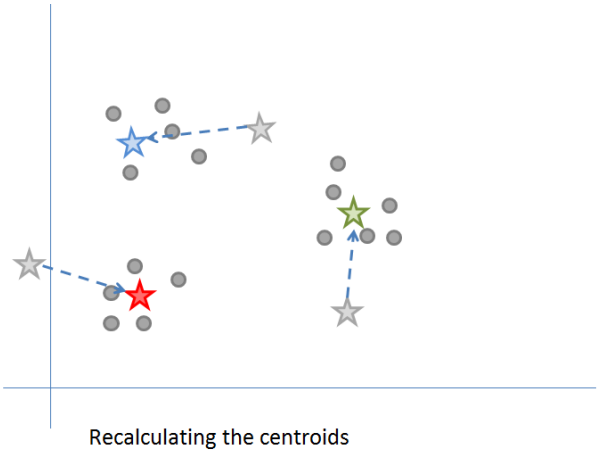




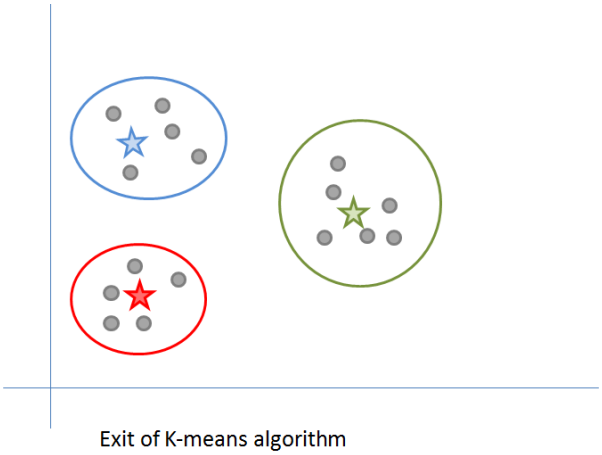
K-means Initialization



Associating each observation to a cluster



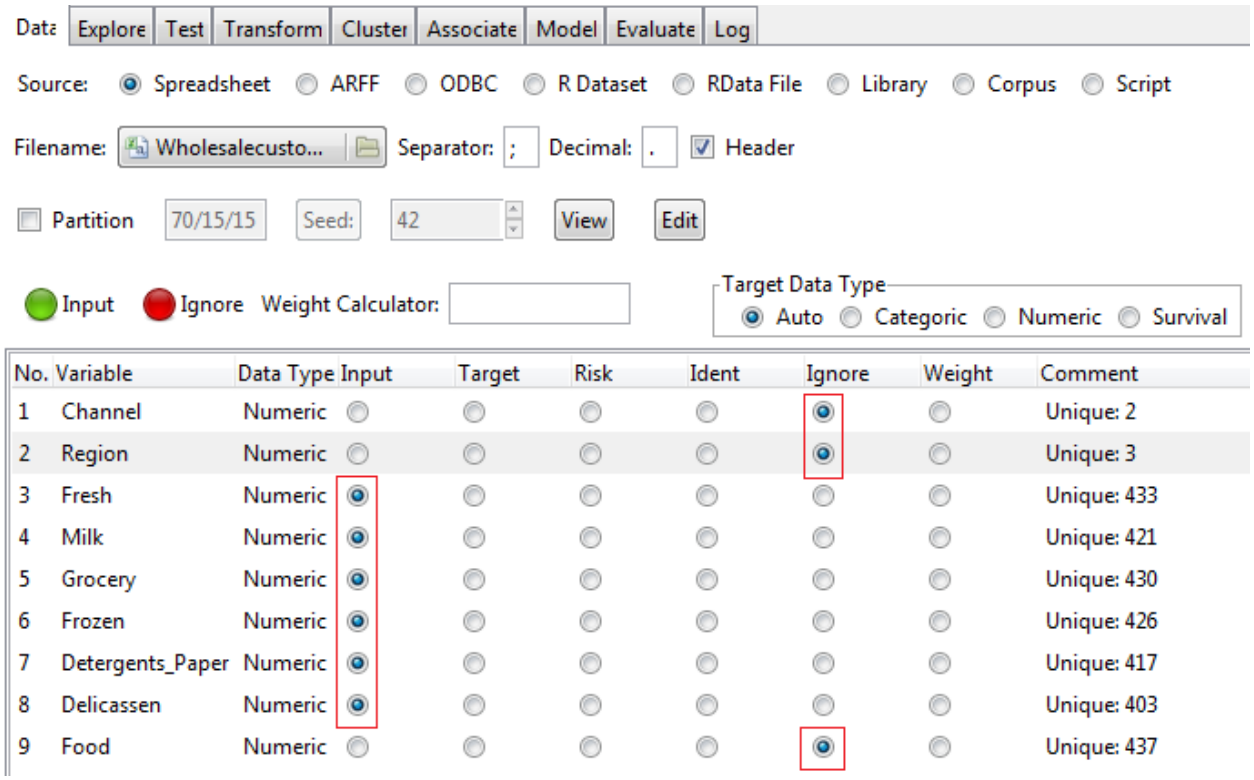
Recalculating the centroids

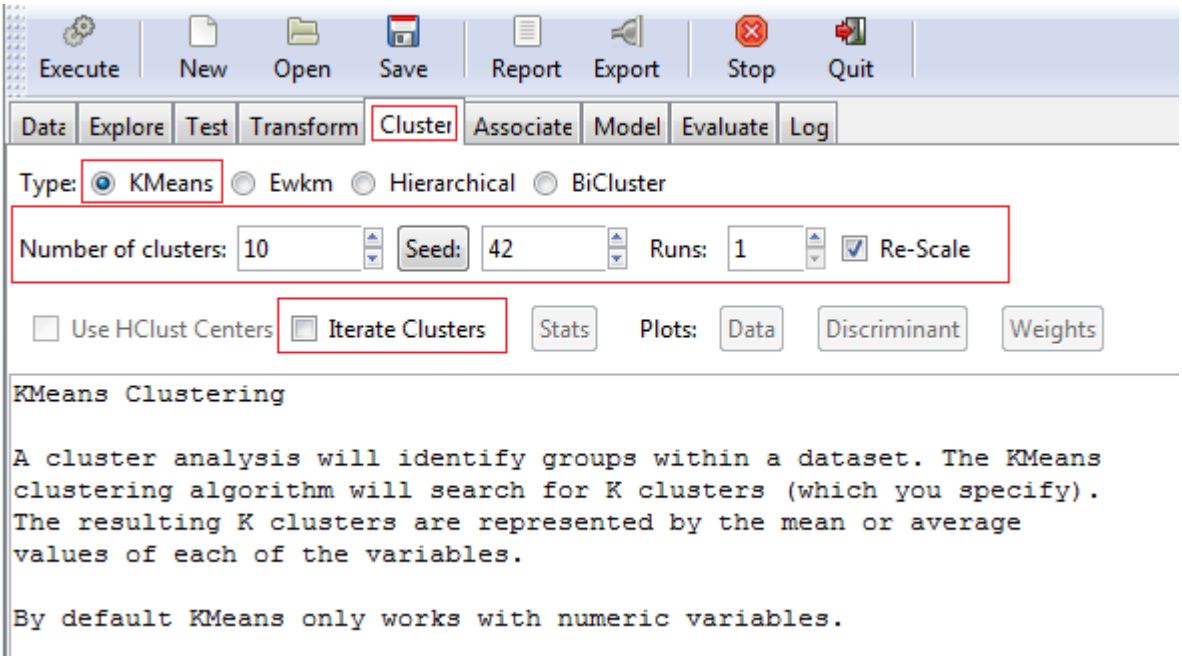


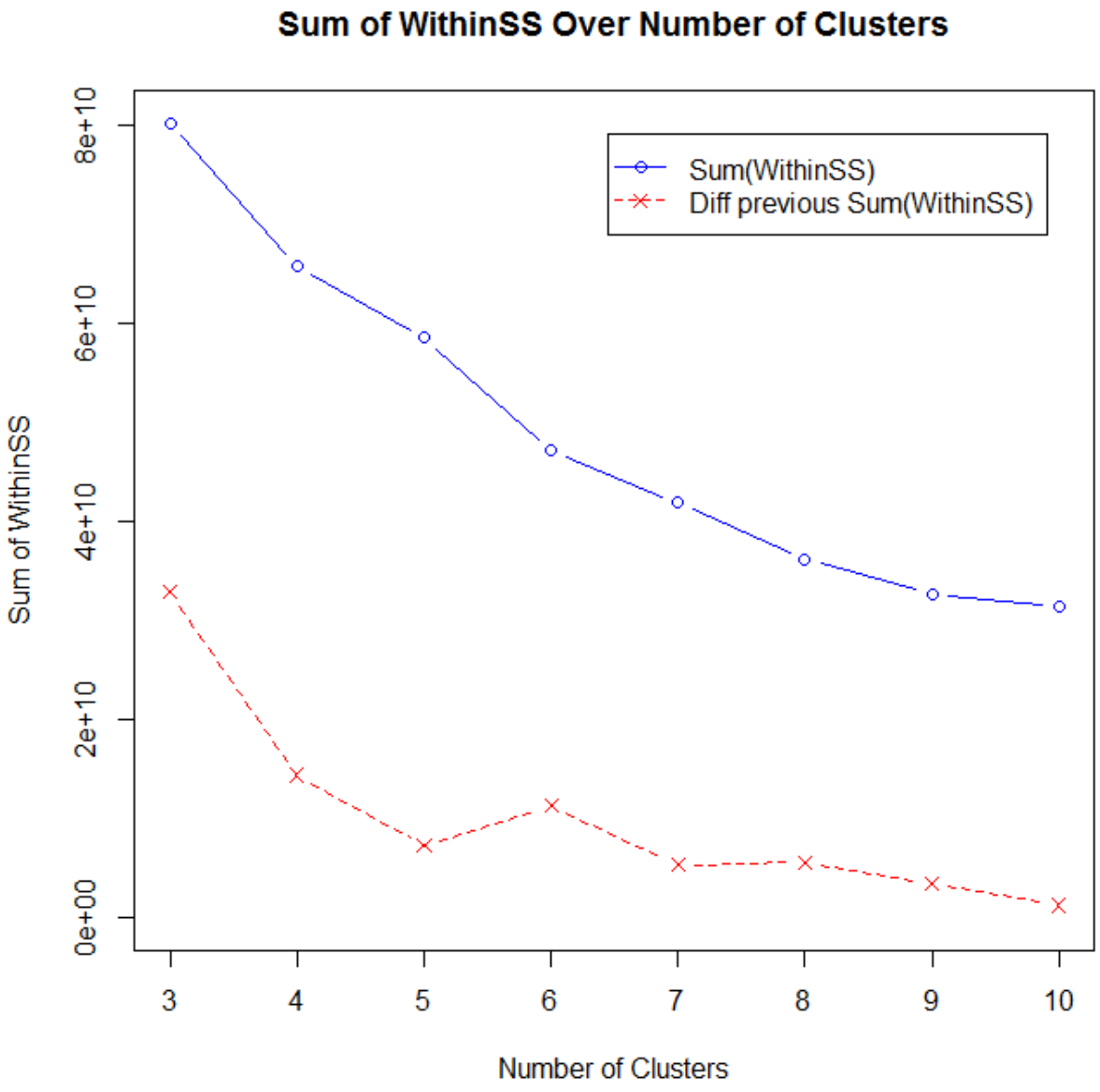
Exit of K-means algorithm

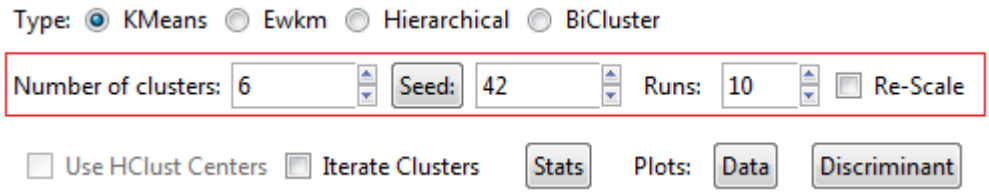
Cluster centers:

	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen
1	46518.000	3438.682	4785.091	5249.955	801.0455	2147.227
2	20996.317	3826.433	5098.087	4157.010	1119.9423	1679.712
3	20031.286	38084.000	56126.143	2564.571	27644.5714	2548.143
4	60571.667	30120.333	17314.667	38049.333	2153.0000	20700.667
5	5996.482	3368.597	4206.765	2418.283	1282.9469	1001.305
6	5076.654	12288.526	18814.526	1605.000	8254.3974	1830.513

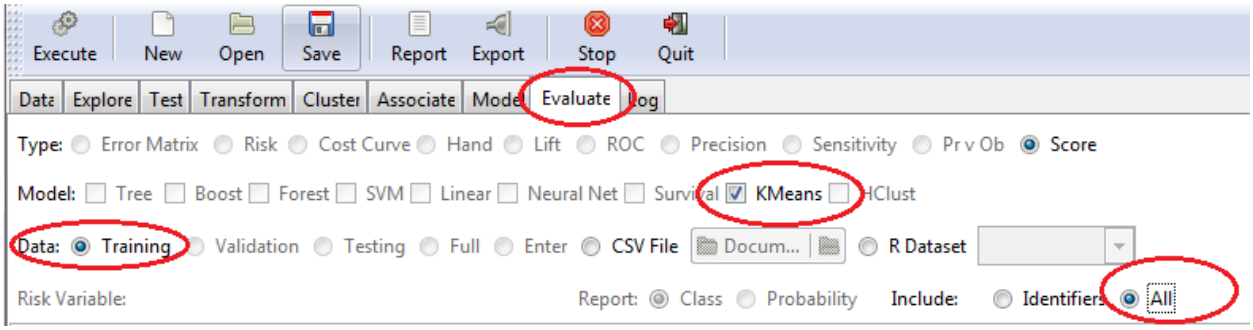

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Source: Spreadsheet ARFF ODBC R Dataset RData File Library Corpus Script
Filename: Wholesalecusto...
Separator: ;
Decimal: .
Header
Partition
70/15/15
Seed:
42
View
Edit
Input
Ignore
Weight Calculator:
Target Data Type
Auto Categoric Numeric Survival
No. Variable Data Type Input Target Risk Ident Ignore Weight Comment
1 Channel Numeric Unique: 2
2 Region Numeric Unique: 3
3 Fresh Numeric Unique: 433
4 Milk Numeric Unique: 421
5 Grocery Numeric Unique: 430
6 Frozen Numeric Unique: 426
7 Detergents_Paper Numeric Unique: 417
8 Delicassen Numeric Unique: 403
9 Food Numeric Unique: 437


Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type:
KMeans
Ewkm
Hierarchical
BiCluster
Number of clusters:
10
Seed:
42
Runs:
1
Re-Scale
Use HClust Centers
Iterate Clusters
Stats
Plots:
Data
Discriminant
Weights
KMeans Clustering
A cluster analysis will identify groups within a dataset. The KMeans clustering algorithm will search for K clusters (which you specify). The resulting K clusters are represented by the mean or average values of each of the variables.
By default KMeans only works with numeric variables.

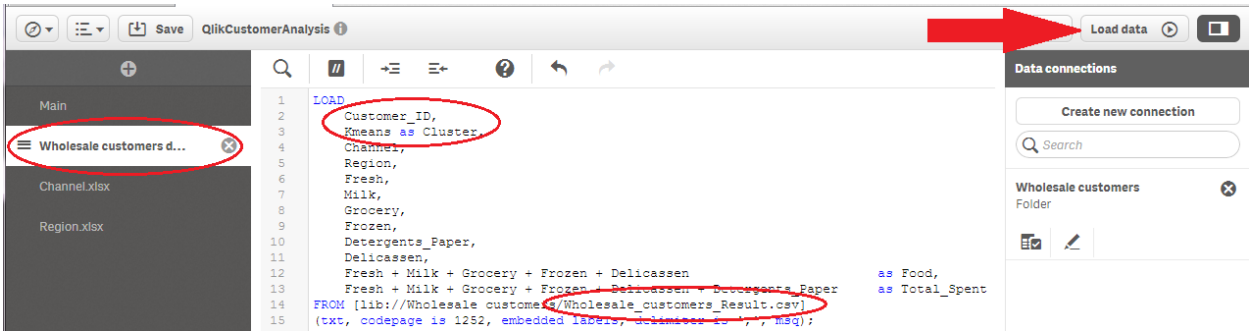

Sum of WithinSS Over Number of Clusters
Sum(WithinSS)
Diff previous Sum(WithinSS)
Sum of WithinSS
0e+00
2e+10
4e+10
6e+10
8e+10
3
4
5
6
7
8
9
10
Number of Clusters

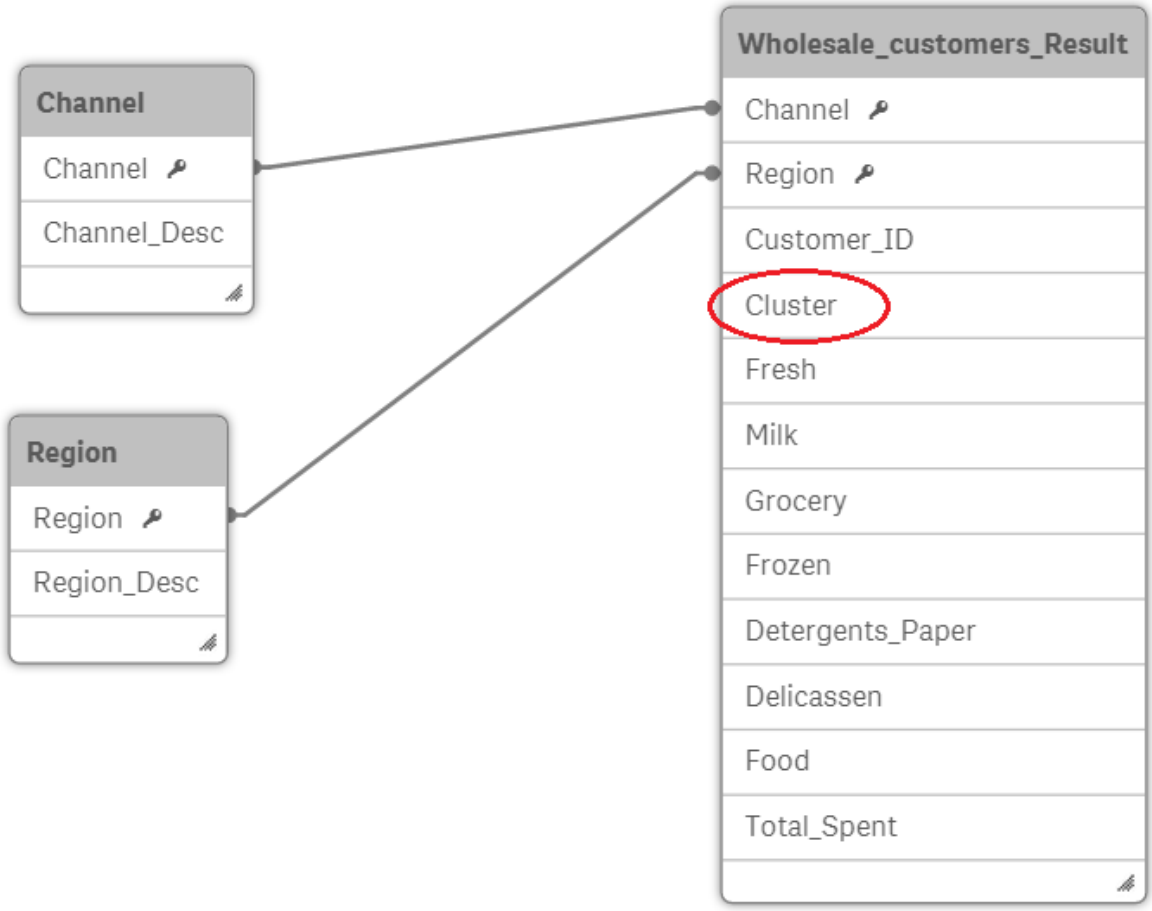


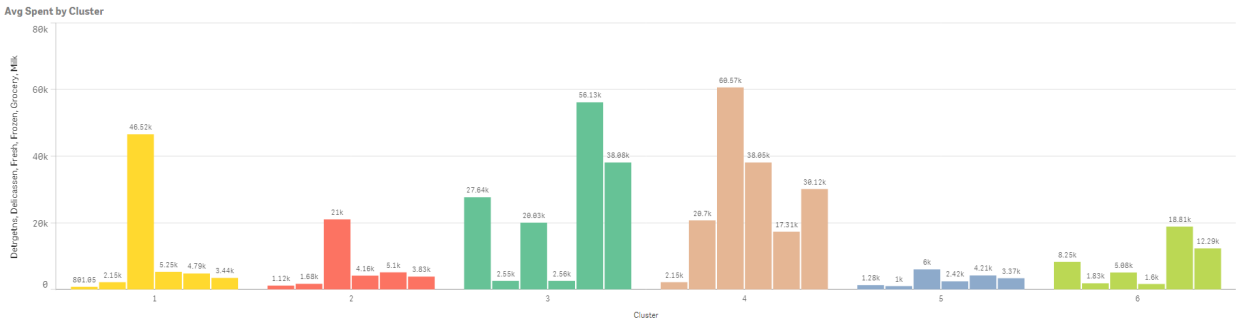

```
Cluster centers:

      Fresh      Milk   Grocery    Frozen Detergents_Paper Delicassen
1 46518.000  3438.682  4785.091  5249.955         801.0455   2147.227
2 20996.317  3826.433  5098.087  4157.010        1119.9423   1679.712
3 20031.286 38084.000 56126.143  2564.571       27644.5714   2548.143
4 60571.667 30120.333 17314.667 38049.333        2153.0000  20700.667
5  5996.482  3368.597  4206.765  2418.283        1282.9469   1001.305
6  5076.654 12288.526 18814.526  1605.000        8254.3974   1830.513
```

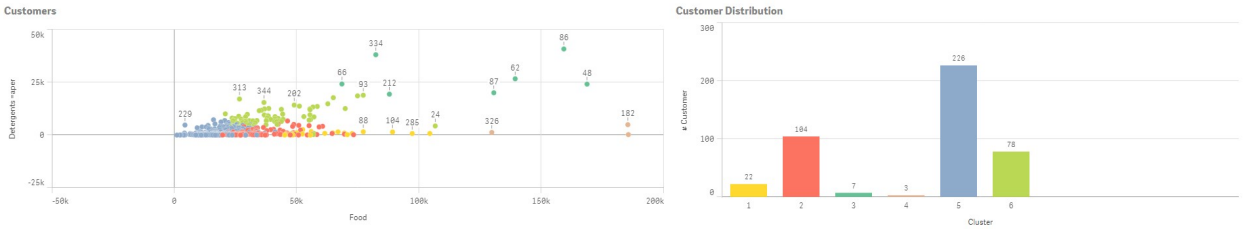


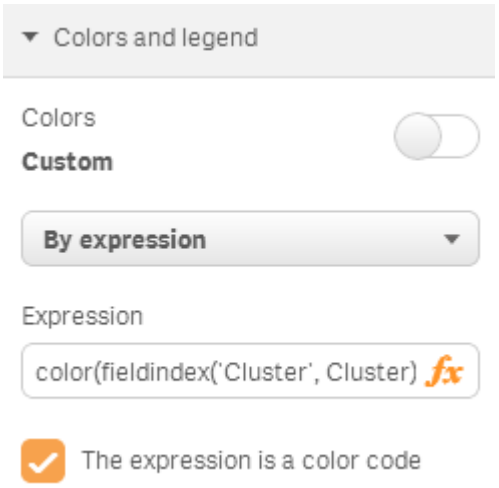
Customer_ID	Channel	Region	Fresh	Milk	Grocery	Frozen	Detergents_Paper	Delicassen	kmeans
1	2	3	12669	9656	7561	214	2674	1338	5
2	2	3	7057	9810	9568	1762	3293	1776	5
3	2	3	6353	8808	7684	2405	3516	7844	5
4	1	3	13265	1196	4221	6404	507	1788	2
5	2	3	22615	5410	7198	3915	1777	5185	2
6	2	3	9413	8259	5126	666	1795	1451	5
7	2	3	12126	3199	6975	480	3140	545	5
8	2	3	7579	4956	9426	1669	3321	2566	5
9	1	3	5963	3648	6192	425	1716	750	5
10	2	3	6006	11093	18881	1159	7425	2098	6

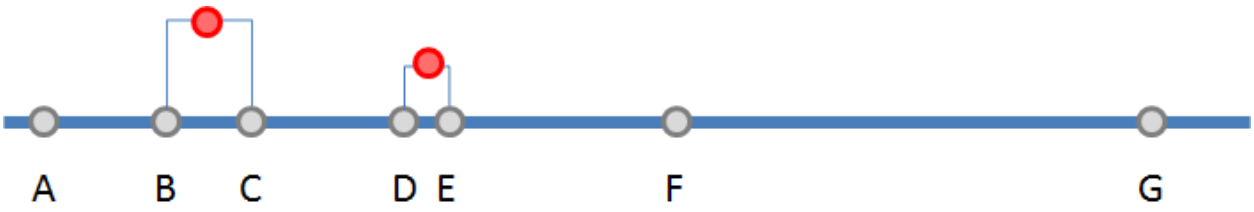



Channel
Channel
Channel_Desc
Region
Region
Region_Desc
Wholesale_customers_Result
Channel
Region
Customer_ID
Cluster
Fresh
Milk
Grocery
Frozen
Detergents_Paper
Delicassen
Food
Total_Spent

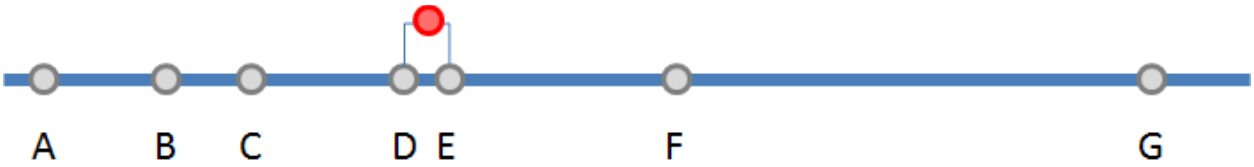

Avg Spent by Cluster
Detrgetns, Delicassen, Fresh, Frozen, Grocery, Milk
80k
60k
40k
20k
0
Cluster
1
2
3
4
5
6

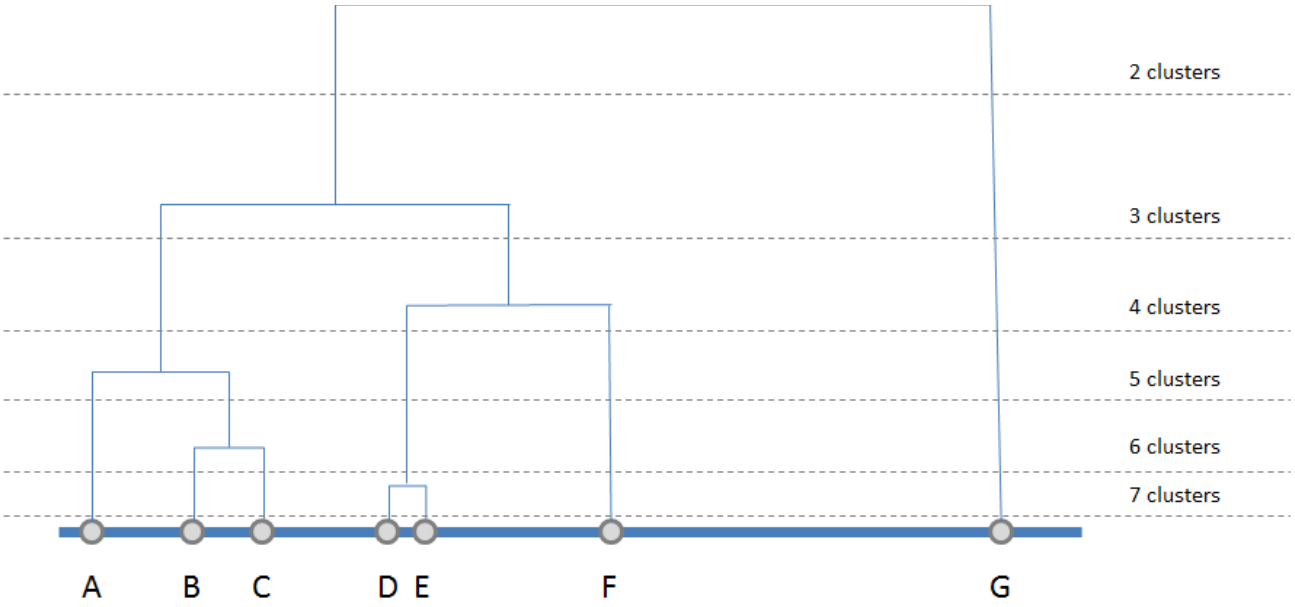

Customers
Food
Customer Distribution
Cluster

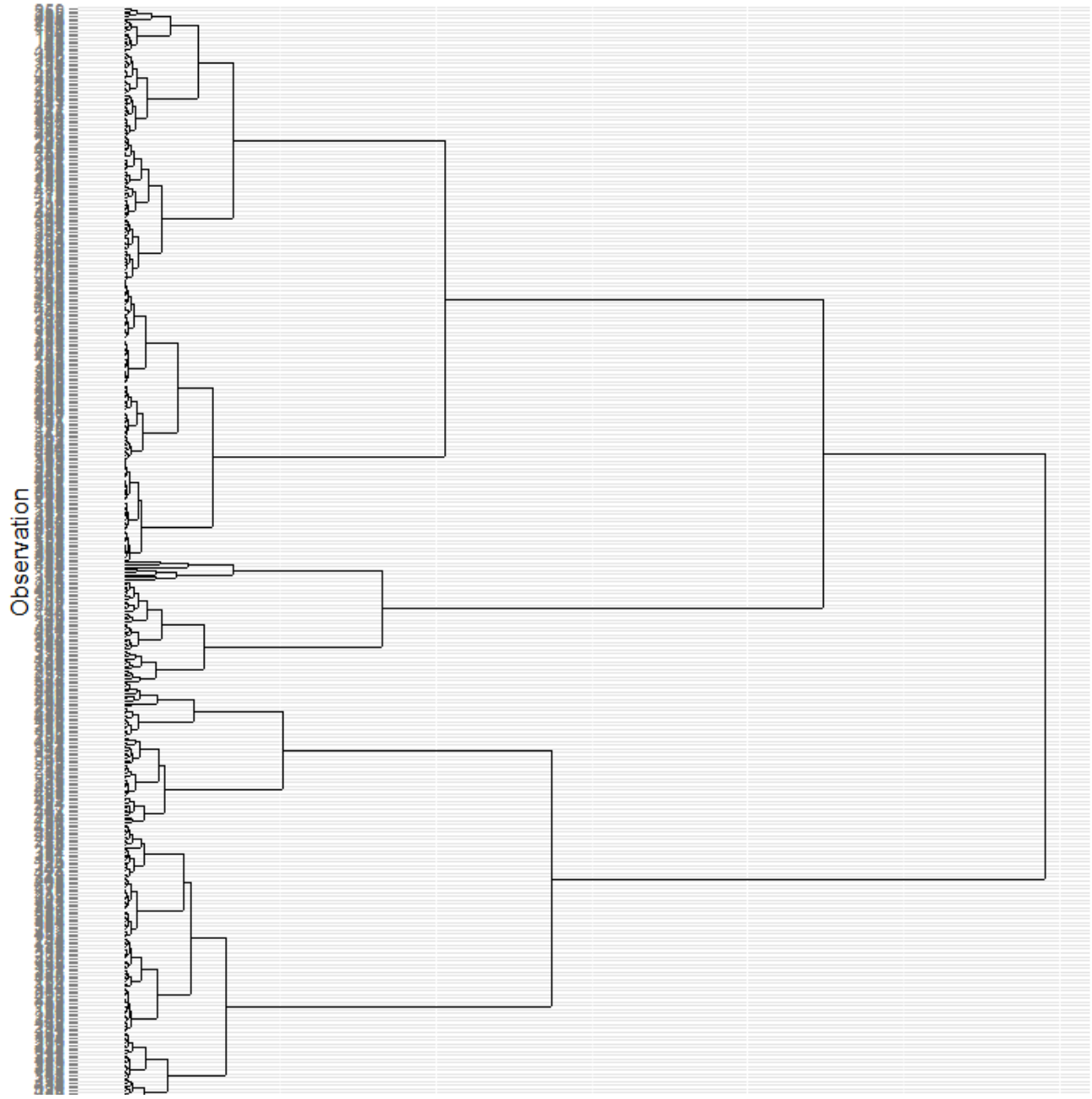

Colors and legend
Colors
Custom
By expression
Expression
color(fieldindex('Cluster', Cluster)
The expression is a color code


A
B
C
D
E
F
G

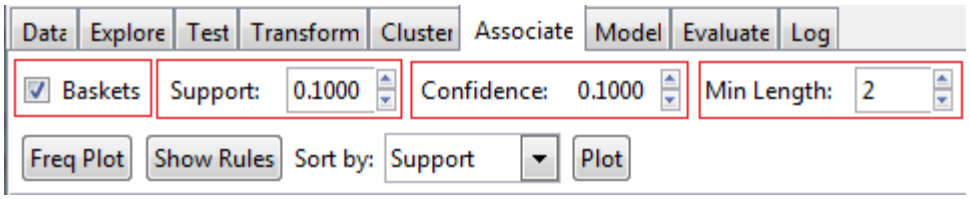
A
B
C
D
E
F
G

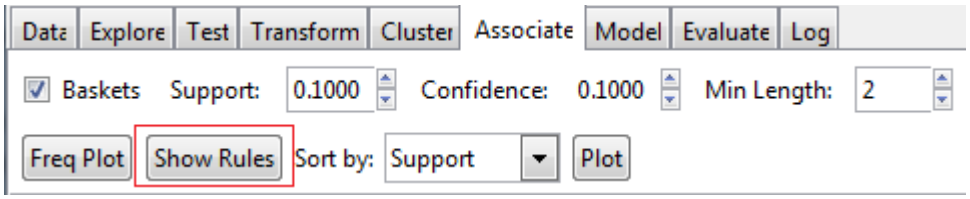

A
B
C
D
E
F
G

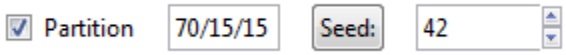

2 clusters
3 clusters
4 clusters
5 clusters
6 clusters
7 clusters
A
B
C
D
E
F
G

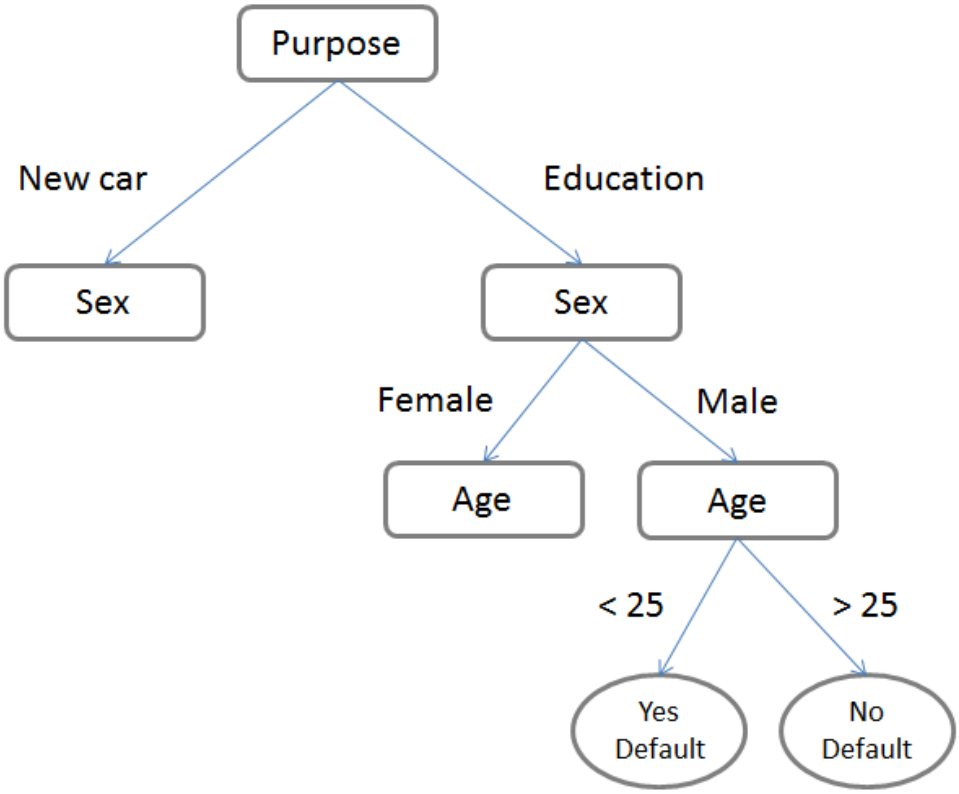

Cluster Dendrogram Wholesalecustomersdata.csv
Observation

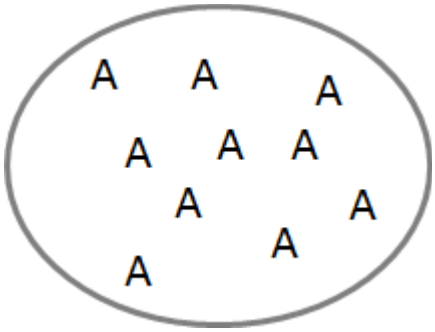
Transaction 1	Burger	Chicken	Potatoes		
Transaction 2	Burger	Onions			
Transaction 3	Onions	Boots			
Transaction 4	Burger	Onions	Chicken		
Transaction 5	Burger	Onions	Chicken	Clothes	Potatoes
Transaction 6	Chicken	Clothes	Potatoes		
Transaction 7	Chicken	Potatoes	Clothes		

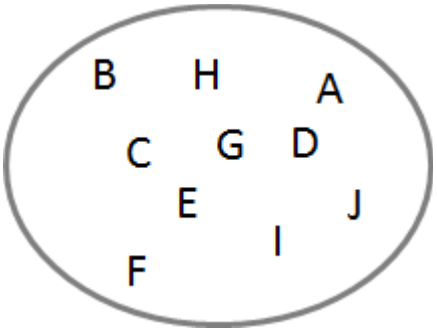


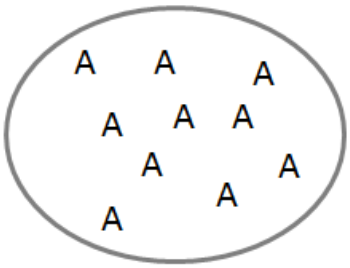



Partition
70/15/15
Seed:
42

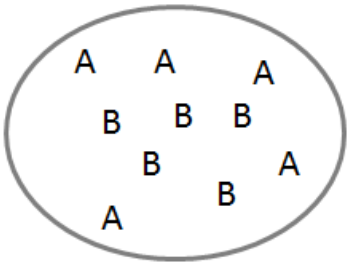

Purpose
New car
Education
Sex
Sex
Female
Male
Age
Age
< 25
> 25
Yes
Default
No
Default


A A A
A A A
A A
A
A

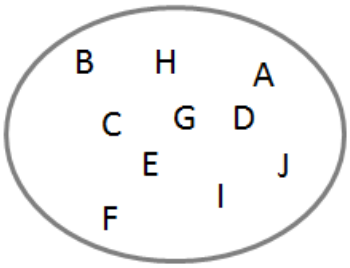

B H A
C G D
E J
I
F



$$E = -P_A * log_2 P_A = -\left(\frac{10}{10}\right) * log_2 \frac{10}{10} = 0$$



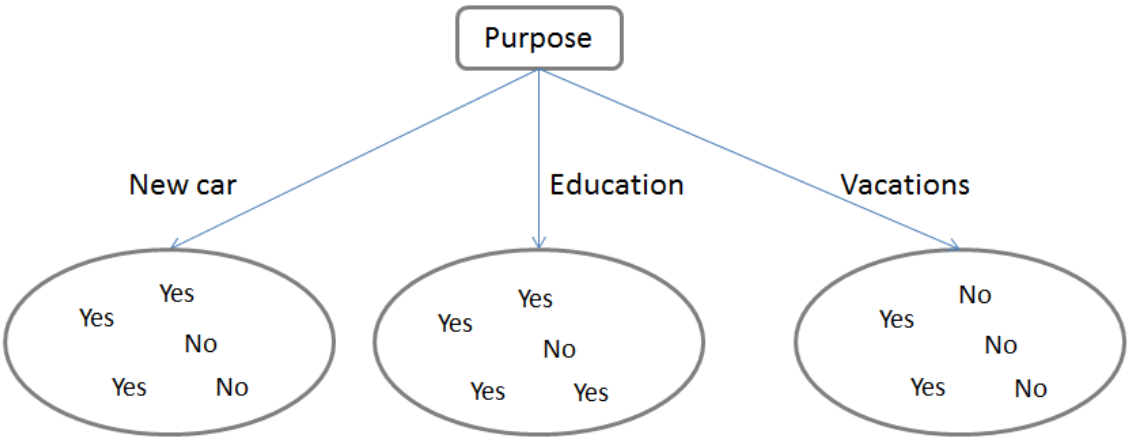
$$E = -(P_A * log_2 P_A + P_B * log_2 P_B) =$$
$$-\left(\left(\frac{5}{10}\right) * log_2 \frac{5}{10} + \left(\frac{5}{10}\right) * log_2 \frac{5}{10}\right) = 1$$



$$E = -(P_A * log_2 P_A + \ldots + P_J * log_2 P_J) = 3.32$$

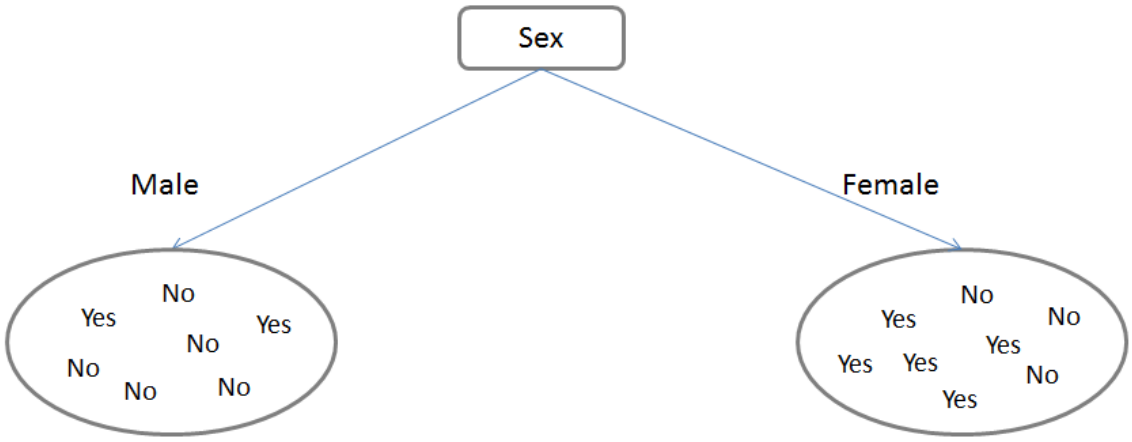
Purpose	Sex	Age	Default?
New Car	Male	< 25	Yes
New Car	Male	25 - 65	No
New Car	Female	< 25	Yes
New Car	Female	25 - 65	Yes
New Car	Male	25 - 65	No
Education	Male	< 25	Yes
Education	Male	< 25	No
Education	Female	< 25	Yes
Education	Female	< 25	Yes
Education	Female	< 25	Yes
Vacations	Female	< 25	Yes
Vacations	Female	> 65	No
Vacations	Female	< 25	Yes
Vacations	Male	25 - 65	No
Vacations	Male	> 65	No

$$E = -\left(P_{yes} * log_2 P_{Yes} + P_{No} * log_2 P_{No}\right) = -\left(\frac{9}{15} * log_2 \frac{9}{15} + \frac{6}{15} * log_2 \frac{6}{15}\right) = 0.97$$



$E_{new\ car} = 0.97$ $E_{education} = 0.72$ $E_{vacations} = 0.97$

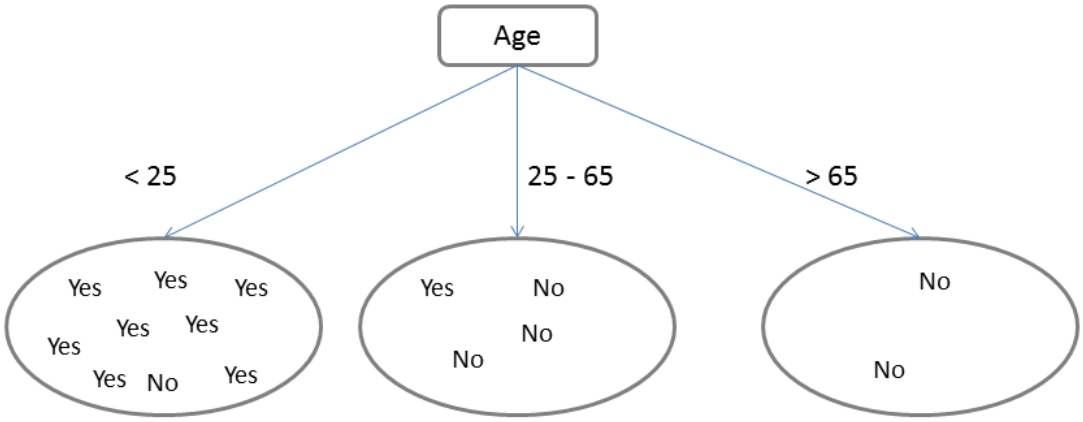
$$E_{purpose} = \frac{5}{15} * 0.97 + \frac{5}{15} * 0.72 + \frac{5}{15} * 0.97 = 0.89$$



$E_{Male} = 0.86$

$E_{Female} = 0.95$

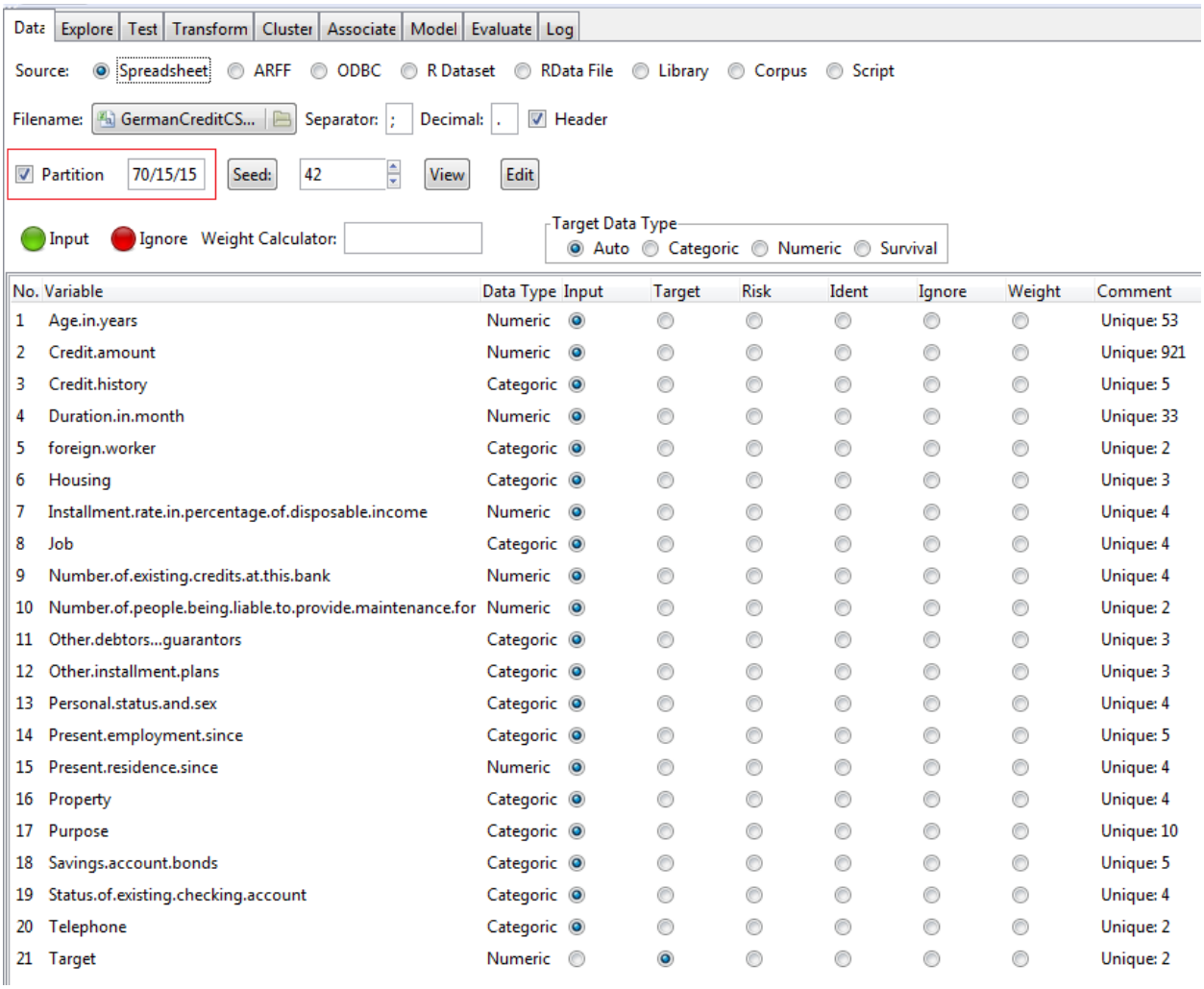
$$E_{Sex} = \frac{7}{15} * 0.86 + \frac{8}{15} * 095 = 0.91$$



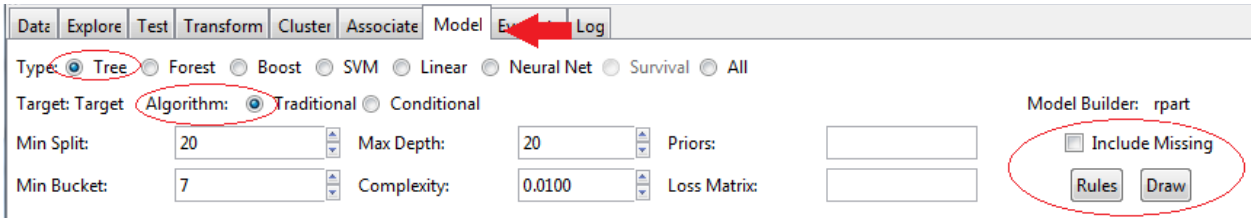
$E_{<25} = 0.50$ $E_{25-65} = 0.81$ $E_{>65} = 0$

$$E_{purpose} = \frac{9}{15} * 0.97 + \frac{4}{15} * 0.72 + 0 = 0.52$$

```
A11 6 A34 A43 1169 A65 A75 4 A93 A101 4 A121 67 A143 A152 2 A173 1 A192 A201 1
A12 48 A32 A43 5951 A61 A73 2 A92 A101 2 A121 22 A143 A152 1 A173 1 A191 A201 2
A14 12 A34 A46 2096 A61 A74 2 A93 A101 3 A121 49 A143 A152 1 A172 2 A191 A201 1
A11 42 A32 A42 7882 A61 A74 2 A93 A103 4 A122 45 A143 A153 1 A173 2 A191 A201 1
A11 24 A33 A40 4870 A61 A73 3 A93 A101 4 A124 53 A143 A153 2 A173 2 A191 A201 2
A14 36 A32 A46 9055 A65 A73 2 A93 A101 4 A124 35 A143 A153 1 A172 2 A192 A201 1
A14 24 A32 A42 2835 A63 A75 3 A93 A101 4 A122 53 A143 A152 1 A173 1 A191 A201 1
A12 36 A32 A41 6948 A61 A73 2 A93 A101 2 A123 35 A143 A151 1 A174 1 A192 A201 1
A14 12 A32 A43 3059 A64 A74 2 A91 A101 4 A121 61 A143 A152 1 A172 1 A191 A201 1
A12 30 A34 A40 5234 A61 A71 4 A94 A101 2 A123 28 A143 A152 2 A174 1 A191 A201 2
```



No.	Variable	Data Type	Input	Target	Risk	Ident	Ignore	Weight	Comment
1	Age.in.years	Numeric	●	○	○	○	○	○	Unique: 53
2	Credit.amount	Numeric	●	○	○	○	○	○	Unique: 921
3	Credit.history	Categoric	●	○	○	○	○	○	Unique: 5
4	Duration.in.month	Numeric	●	○	○	○	○	○	Unique: 33
5	foreign.worker	Categoric	●	○	○	○	○	○	Unique: 2
6	Housing	Categoric	●	○	○	○	○	○	Unique: 3
7	Installment.rate.in.percentage.of.disposable.income	Numeric	●	○	○	○	○	○	Unique: 4
8	Job	Categoric	●	○	○	○	○	○	Unique: 4
9	Number.of.existing.credits.at.this.bank	Numeric	●	○	○	○	○	○	Unique: 4
10	Number.of.people.being.liable.to.provide.maintenance.for	Numeric	●	○	○	○	○	○	Unique: 2
11	Other.debtors...guarantors	Categoric	●	○	○	○	○	○	Unique: 3
12	Other.installment.plans	Categoric	●	○	○	○	○	○	Unique: 3
13	Personal.status.and.sex	Categoric	●	○	○	○	○	○	Unique: 4
14	Present.employment.since	Categoric	●	○	○	○	○	○	Unique: 5
15	Present.residence.since	Numeric	●	○	○	○	○	○	Unique: 4
16	Property	Categoric	●	○	○	○	○	○	Unique: 4
17	Purpose	Categoric	●	○	○	○	○	○	Unique: 10
18	Savings.account.bonds	Categoric	●	○	○	○	○	○	Unique: 5
19	Status.of.existing.checking.account	Categoric	●	○	○	○	○	○	Unique: 4
20	Telephone	Categoric	●	○	○	○	○	○	Unique: 2
21	Target	Numeric	○	●	○	○	○	○	Unique: 2

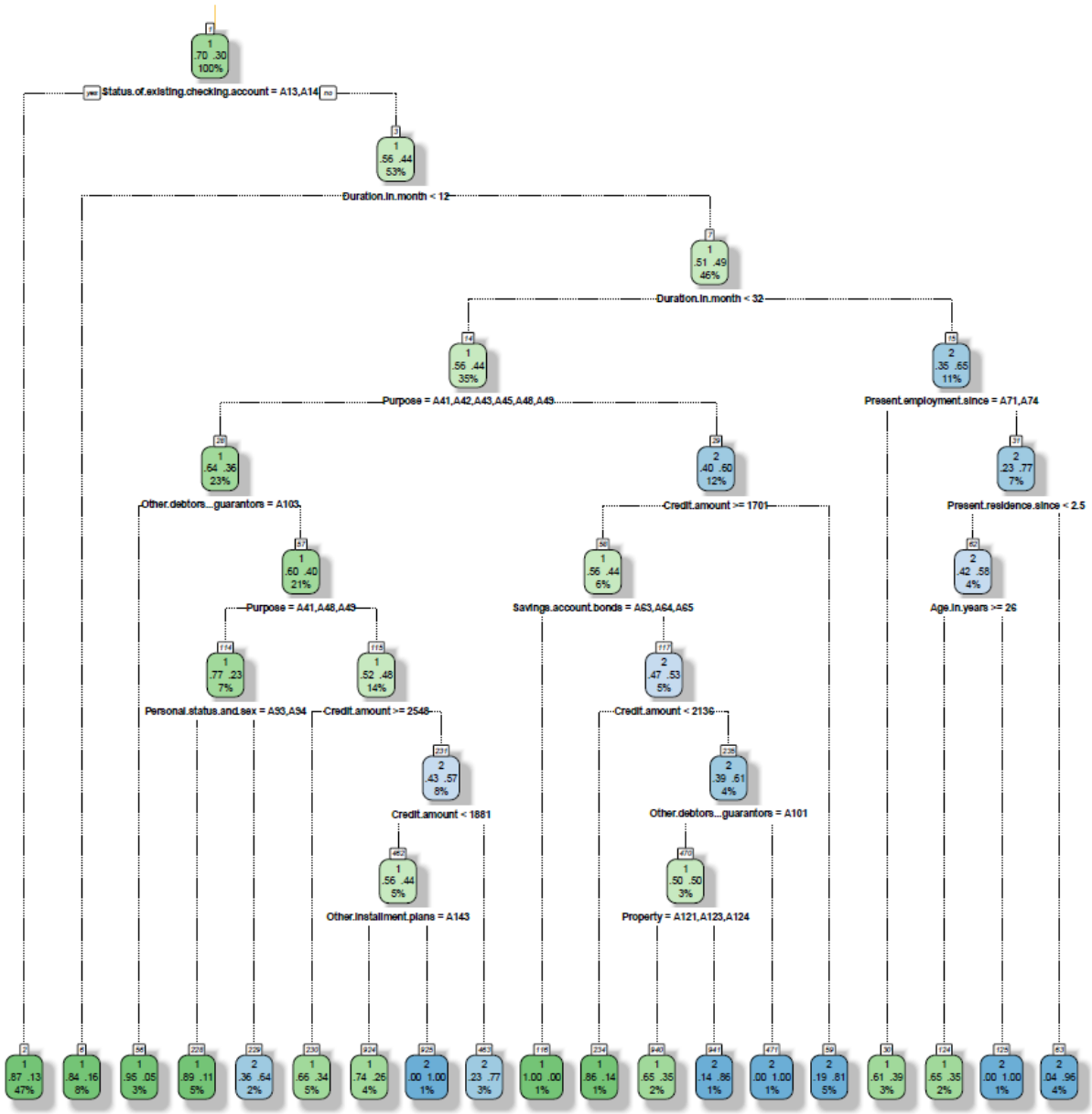


```
Summary of the Decision Tree model for Classification (built using 'rpart'):

n= 700

node), split, n, loss, yval, (yprob)
      * denotes terminal node

  1) root 700 209 1 (0.70142857 0.29857143)
    2) Status.of.existing.checking.account=A13,A14 326  44 1 (0.86503067 0.13496933) *
    3) Status.of.existing.checking.account=A11,A12 374 165 1 (0.55882353 0.44117647)
```

1) root 700 209 1 (0.70142857 0.29857143)

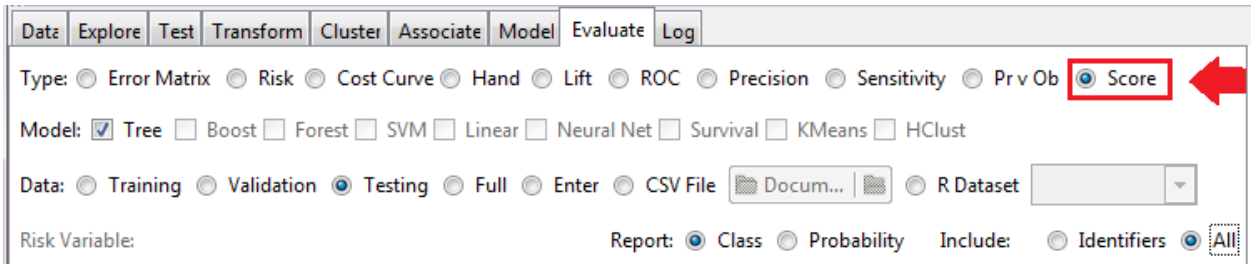
2) Status.of.existing.checking.account=A13,A14 326 44 1 (0.86503067 0.13496933) *

3) Status.of.existing.checking.account=A11,A12 374 165 1 (0.55882353 0.44117647)

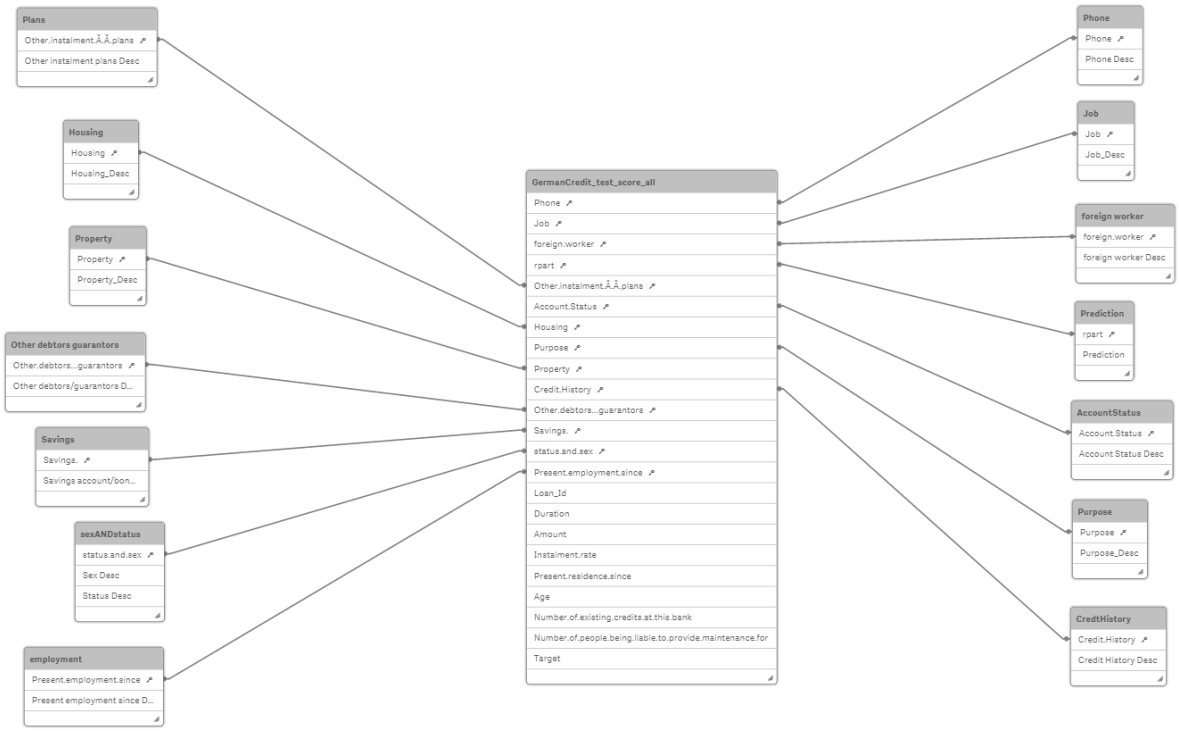

1
.70 .30
100%
yes
Status.of.existing.checking.account = A13,A14
no
1
.56 .44
53%
Duration.in.month < 12
1
.51 .49
46%
Duration.in.month < 32
1
.56 .44
35%
Purpose = A41,A42,A43,A45,A48,A49
2
.35 .65
11%
Present.employment.since = A71,A74
1
.64 .36
23%
Other.debtors...guarantors = A103
2
.40 .60
12%
Credit.amount >= 1701
2
.23 .77
7%
Present.residence.since < 2.5
1
.60 .40
21%
Purpose = A41,A48,A49
1
.56 .44
6%
Savings.account.bonds = A63,A64,A65
2
.42 .58
4%
Age.in.years >= 26
1
.77 .23
7%
1
.52 .48
14%
2
.47 .53
5%
Personal.status.and.sex = A93,A94
Credit.amount >= 2548
Credit.amount < 2136
2
.43 .57
8%
2
.39 .61
4%
Credit.amount < 1881
Other.debtors...guarantors = A101
1
.56 .44
5%
1
.50 .50
3%
Other.installment.plans = A143
Property = A121,A123,A124
1
.87 .13
47%
1
.84 .16
8%
1
.95 .05
3%
1
.89 .11
5%
2
.36 .64
2%
1
.66 .34
5%
1
.74 .26
4%
2
.00 1.00
1%
2
.23 .77
3%
1
1.00 .00
1%
1
.86 .14
1%
1
.65 .35
2%
2
.14 .86
1%
2
.00 1.00
1%
2
.19 .81
5%
1
.61 .39
3%
1
.65 .35
2%
2
.00 1.00
1%
2
.04 .96
4%

```
================================================================================
Tree as rules:

 Rule number: 125 [Target=2 cover=9 (1%) prob=1.00]
   Status.of.existing.checking.account=A11,A12
   Duration.in.month>=11.5
   Duration.in.month>=31.5
   Present.employment.since=A72,A73,A75
   Present.residence.since< 2.5
   Age.in.years< 26.5
```

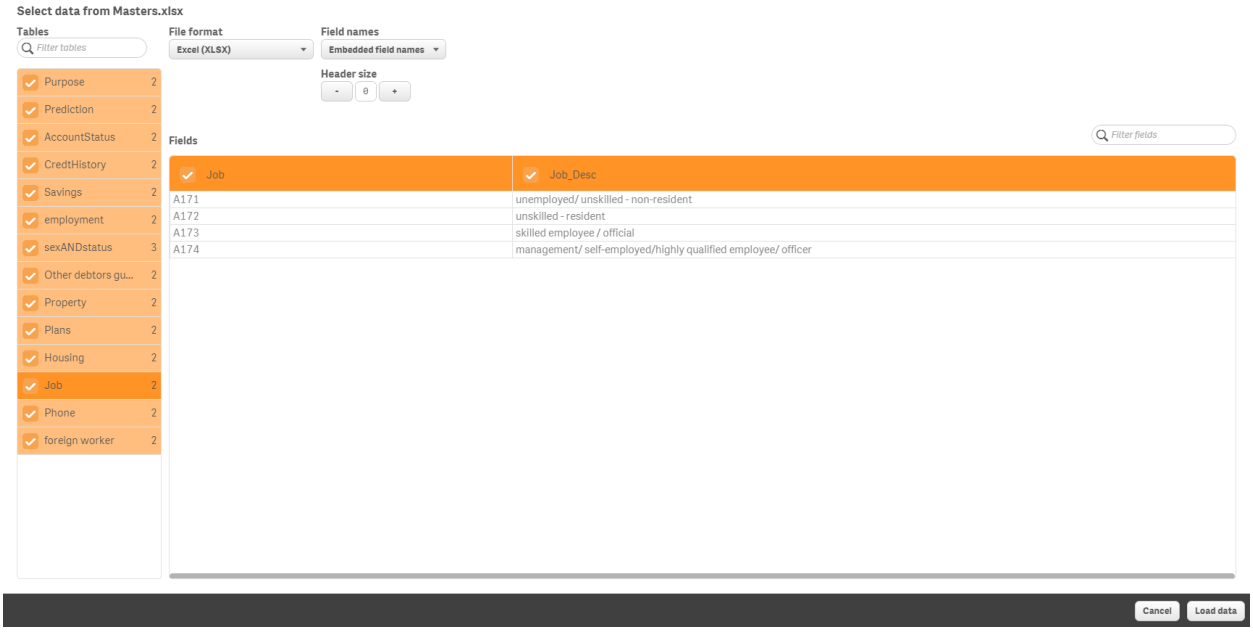
```
// Rule number: 125 [Target=2 cover=9 (1%) prob=1.00]
if (("Status of existing checking account" = 'A11' OR "Status of existing checking account" = 'A12') AND
     ("Duration in month">=11.5 AND "Duration in month">=31.5) AND
     ("Present residence since" = 'A72' OR "Present residence since" = 'A73' OR "Present residence since" = 'A75') AND
     ( "Present residence since" < 2.5) AND
     ("Age in years" < 26.5), 2, 0) AS Prediction,
```



	Account.S	Duration	Credit.His	Purpose	Amount	Savings.	Present.e	Instalmen	status.anc	Other.det	Present.re	Property	Age	Other.inst	Housing	Number.c	Job	Number.c	Phone	foreign.w	Target	rpart
2	A14	12	A32	A43	3059	A64	A74	2	A91	A101	4	A121	61	A143	A152	1	A172	1	A191	A201	1	1
3	A11	30	A30	A49	8072	A65	A72	2	A93	A101	3	A123	25	A141	A152	3	A173	1	A191	A201	1	1
4	A12	24	A32	A41	12579	A61	A75	4	A92	A101	2	A124	44	A143	A153	1	A174	1	A192	A201	2	2
5	A11	6	A32	A42	1374	A61	A73	1	A93	A101	2	A121	36	A141	A152	1	A172	1	A192	A201	1	1
6	A12	18	A32	A49	1913	A64	A72	3	A94	A101	3	A121	36	A141	A152	1	A173	1	A192	A201	1	1
7	A14	12	A34	A49	1264	A65	A75	4	A93	A101	4	A124	57	A143	A151	1	A172	1	A191	A201	1	1
8	A14	36	A32	A43	2299	A63	A75	4	A93	A101	4	A123	39	A143	A152	1	A173	1	A191	A201	1	1



Purpose	Purpose_Description
A40	car (new)
A41	car (used)
A42	furniture/equipment
A43	radio/television
A44	domestic appliances
A45	repairs
A46	education
A47	vacation
A48	retraining
A49	business
A410	others

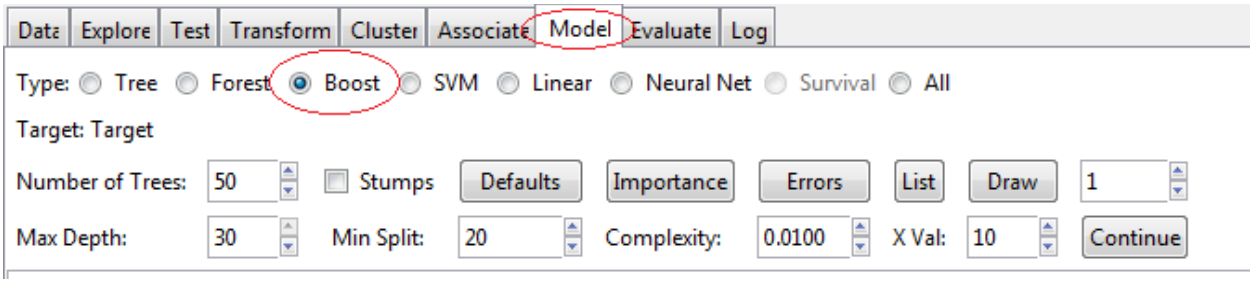

Select data from Masters.xlsx
Tables
Filter tables
Purpose 2
Prediction 2
AccountStatus 2
CredtHistory 2
Savings 2
employment 2
sexANDstatus 3
Other debtors gu... 2
Property 2
Plans 2
Housing 2
Job 2
Phone 2
foreign worker 2
File format
Excel (XLSX)
Field names
Embedded field names
Header size
-
0
+
Fields
Filter fields
Job
Job_Desc
A171
unemployed/ unskilled - non-resident
A172
unskilled - resident
A173
skilled employee / official
A174
management/ self-employed/highly qualified employee/ officer
Cancel
Load data

New Loan Applications
Prediction
Bad
Good
Account Status Desc
Age
Amount
Credit History Desc
Duration
foreign worker Desc
Housing_Desc
Job_Desc
Number.of.people.being.liable.to.provide.ma...
Number.of.existing.credits.at.this.bank
Number.of.people.being.liable.to.provide.ma...
Other instalment plans Desc
Phone Desc
Present.residence.since
Property_Desc
Number of Applications
Prediction
Bad
22%
78%
Good
Credit Amount
Prediction
Bad
36.5%
63.5%
Good
Purpose
Applications
40
20
0
business
car (new)
car (used)
domestic appliance
education
furniture/equipment
others
radio/television
repairs
retraining
vacation
Purpose_Desc , Prediction
Credit Distribution
Duration
80
60
40
20
0
0
2k
4k
6k
8k
10k
12k
14k
16k
Credit Amount

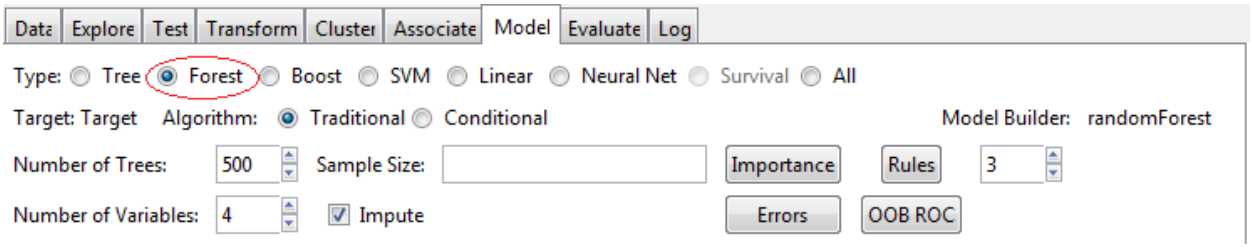
$$Tree_x(new\ observation) = Classification_x$$

$$Tree\left(new\ observation\right) = \frac{\sum_{i=1}^{m} Tree_i\left(new\ observation\right)}{m}$$

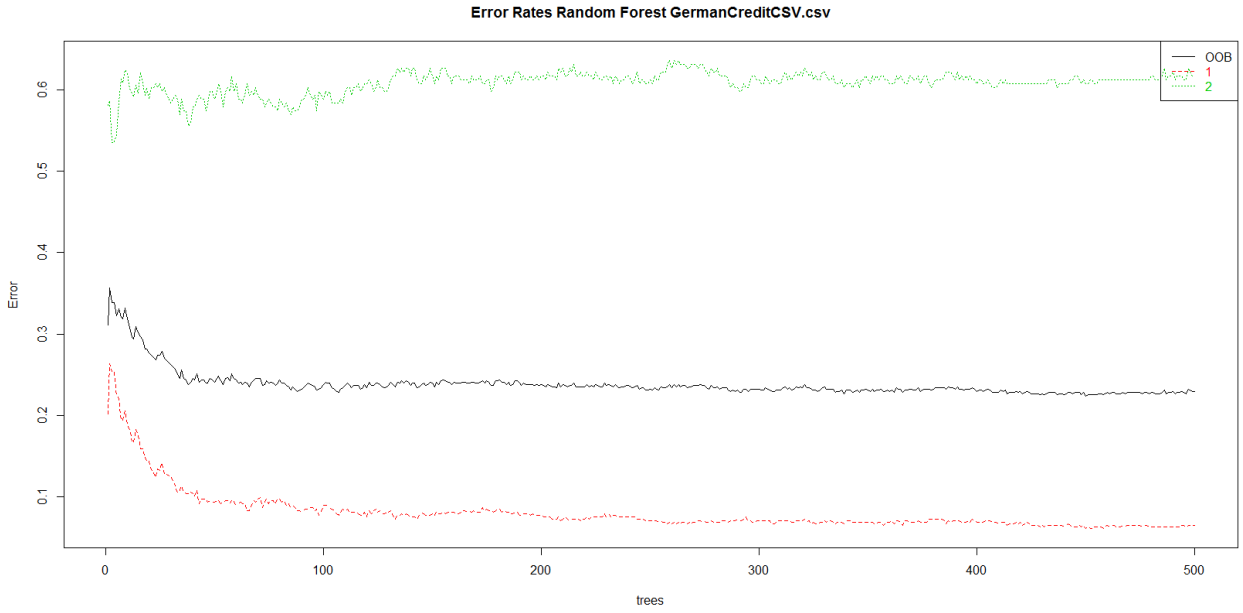
$$Tree(new\ observation) = \sum_{i=1}^{m} Weight_i * Tree_i(new\ observation)$$

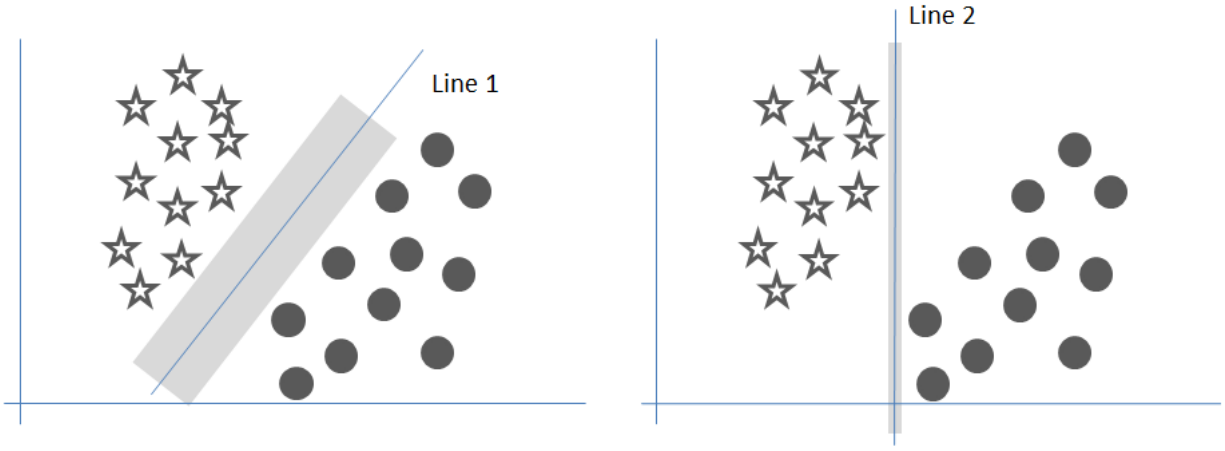


```
Summary of the Ada Boost model:

Call:
ada(Target ~ ., data = crs$dataset[crs$train, c(crs$input, crs$target)],
    control = rpart.control(maxdepth = 30, cp = 0.01, minsplit = 20,
        xval = 10), iter = 50)

Loss: exponential Method: discrete   Iteration: 50

Final Confusion Matrix for Data:
          Final Prediction
True value   1   2
         1 481  10
         2  54 155

Train Error: 0.091

Out-Of-Bag Error:  0.123  iteration= 46
```

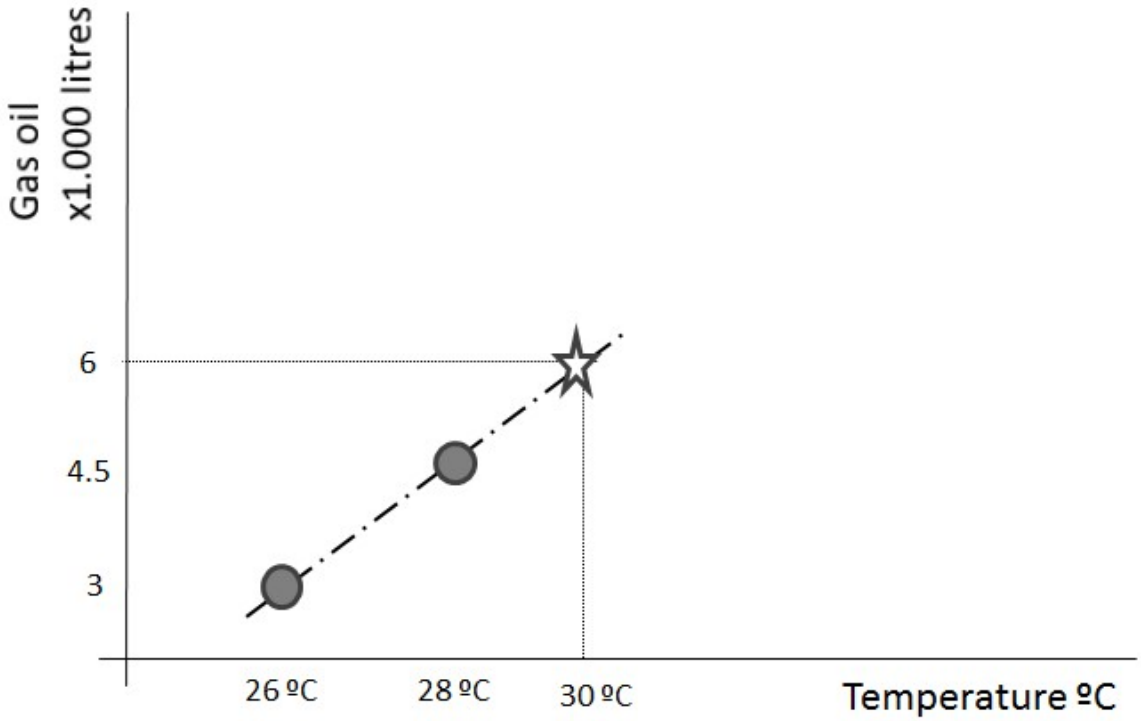


```
Summary of the Random Forest Model
==================================

Number of observations used to build the model: 700
Missing value imputation is active.

Call:
 randomForest(formula = as.factor(Target) ~ .,
              data = crs$dataset[crs$sample, c(crs$input, crs$target)],
              ntree = 500, mtry = 4, importance = TRUE, replace = FALSE, na.action = na.roughfix)

               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 23%
Confusion matrix:
    1  2 class.error
1 459 32  0.06517312
2 129 80  0.61722488
```



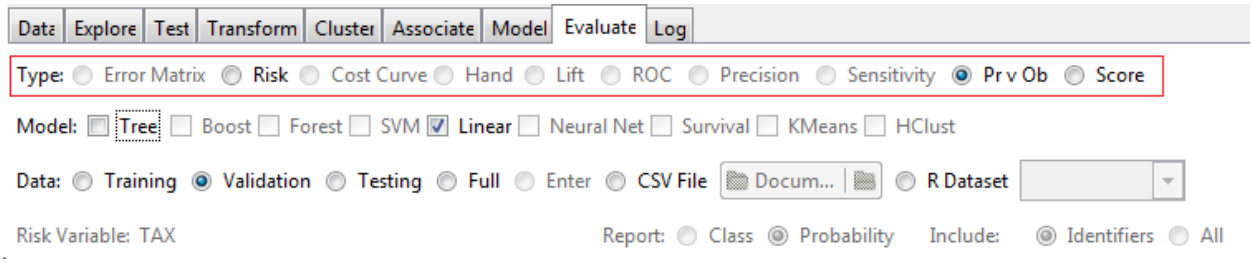


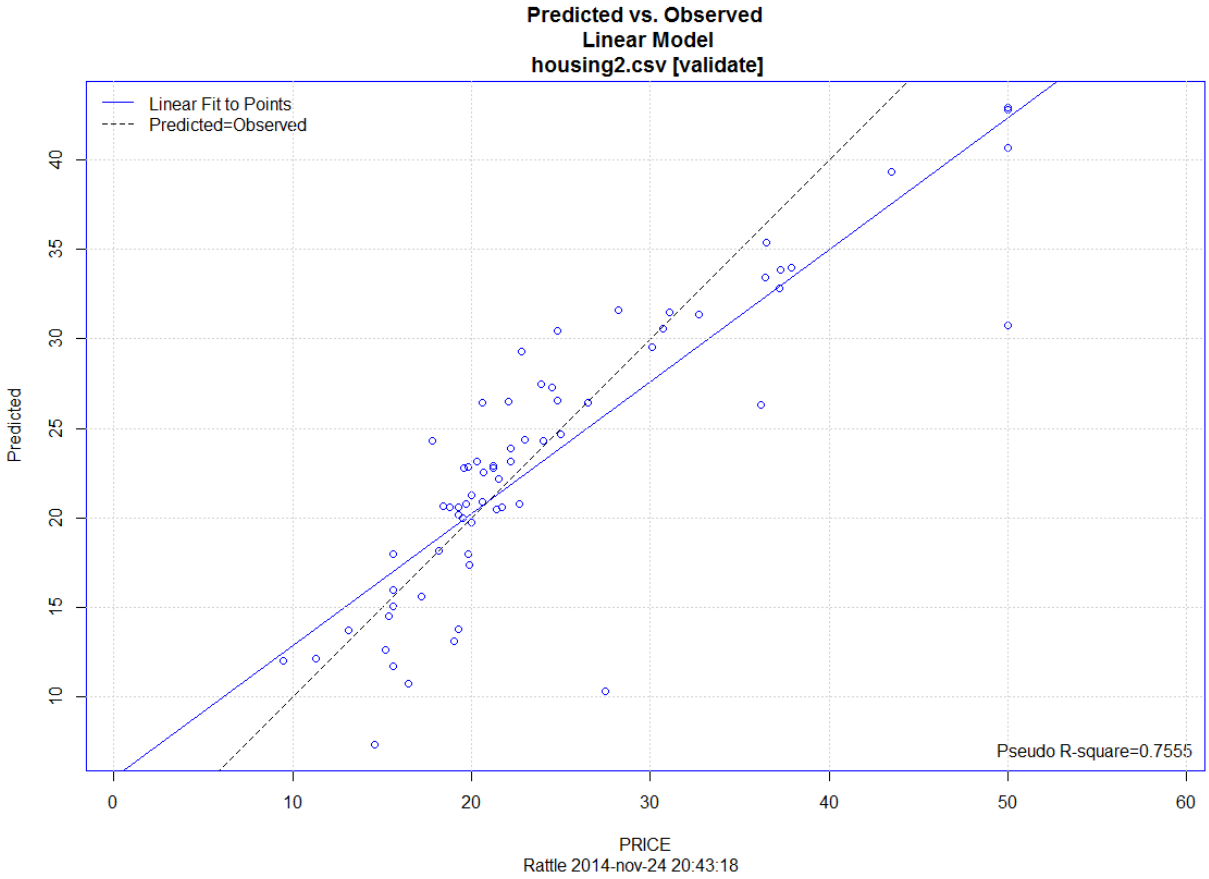
$$Gas\ oil\ Sold = a + b * Temperature$$

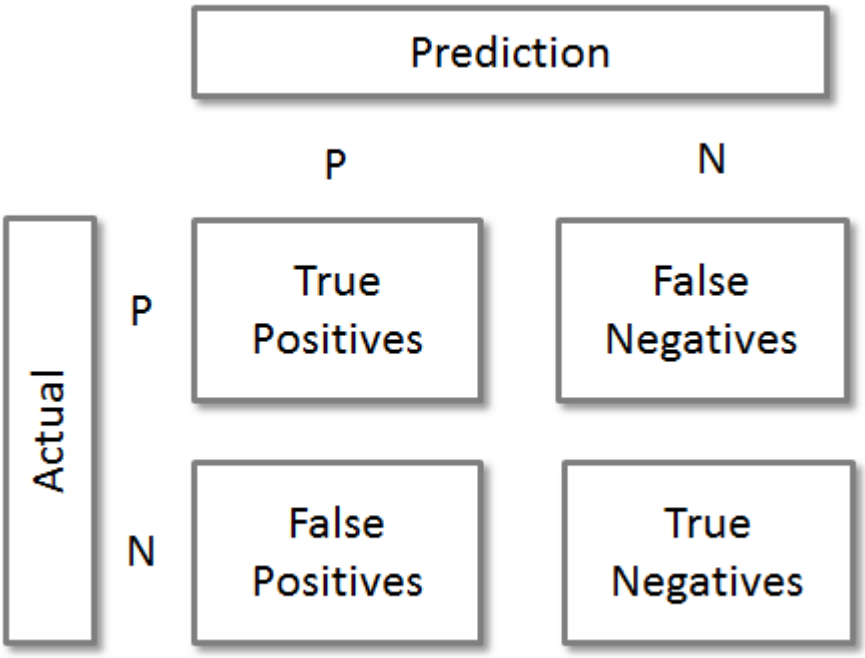

Gas oil
x1.000 litres
6
4.5
3
26 ºC
28 ºC
30 ºC
Temperature ºC

Chapter 7: Model Evaluation

```
Root node error: 209/700 = 0.29857

n= 700

        CP nsplit rel error  xerror     xstd
1 0.036683      0   1.00000 1.00000 0.057932
2 0.023923      4   0.80861 0.99522 0.057852
3 0.016746      6   0.76077 0.93780 0.056839
4 0.012759      8   0.72727 0.96172 0.057273
5 0.011962     14   0.61244 0.96651 0.057358
6 0.010000     18   0.56459 0.96651 0.057358
```

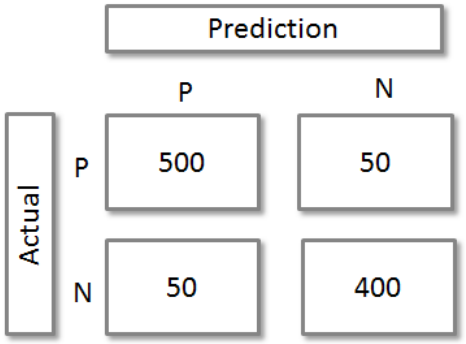



Predicted vs. Observed
Linear Model
housing2.csv [validate]
Linear Fit to Points
Predicted=Observed
Predicted
PRICE
Pseudo R-square=0.7555
Rattle 2014-nov-24 20:43:18



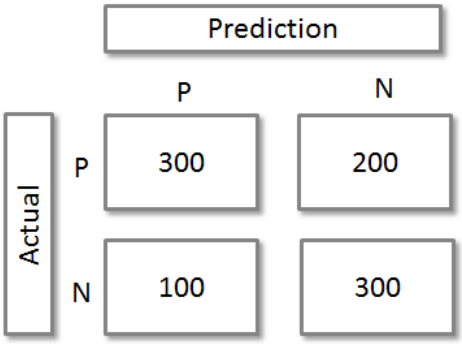
$$Accuracy = \frac{Correctly\ classified\ observations}{Total\ Predicted\ observations}$$

Classifier A



$$Accuracy = \frac{900}{1.000} = 0.9$$

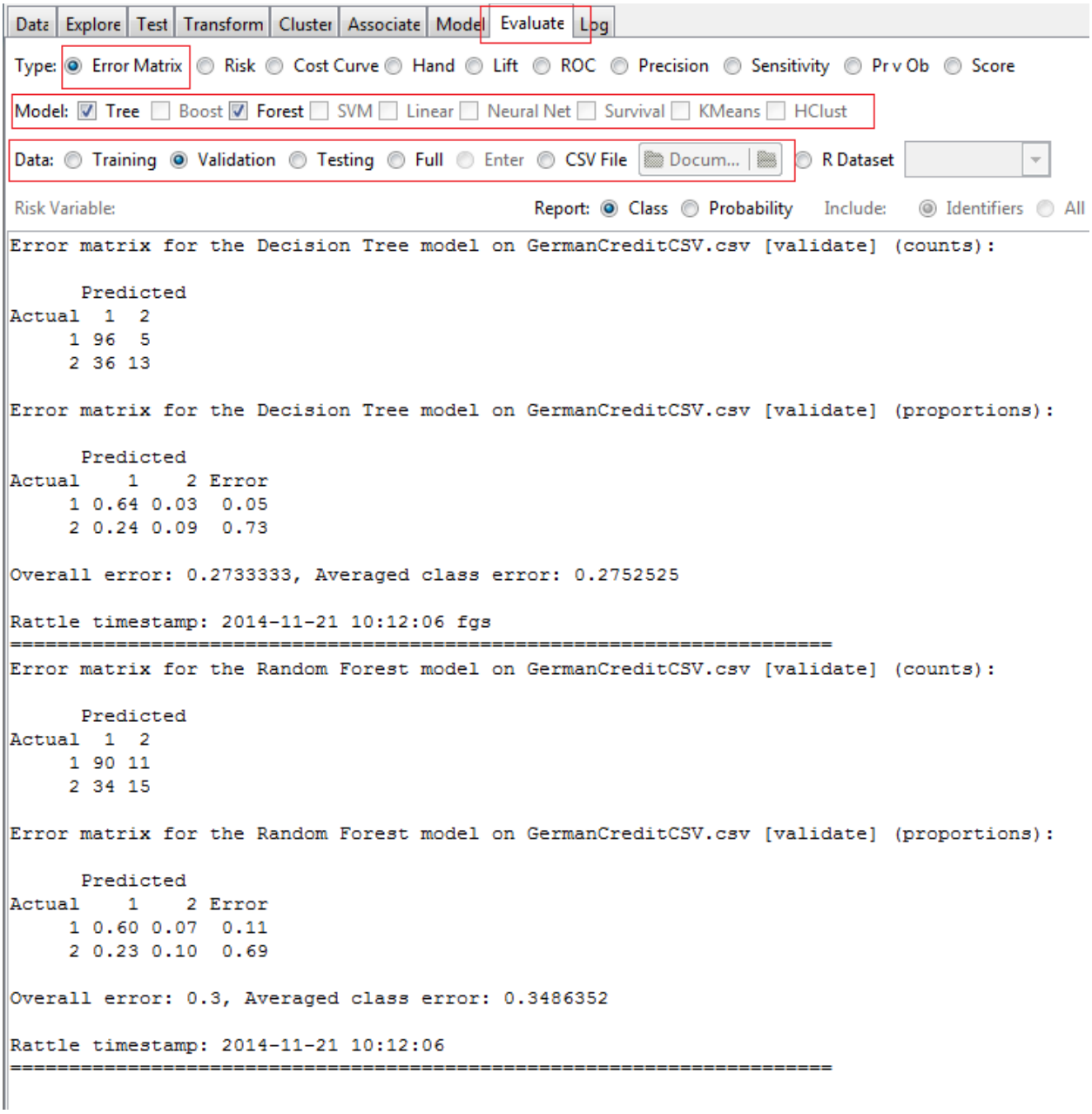
Classifier B

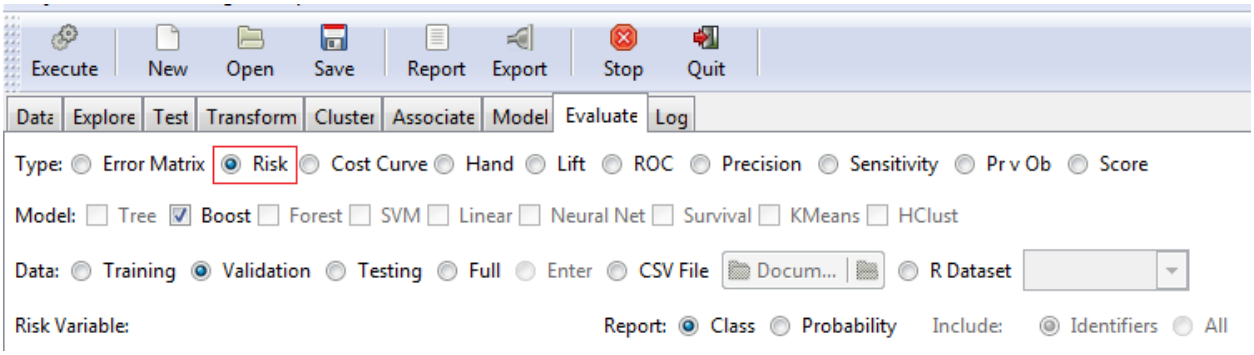


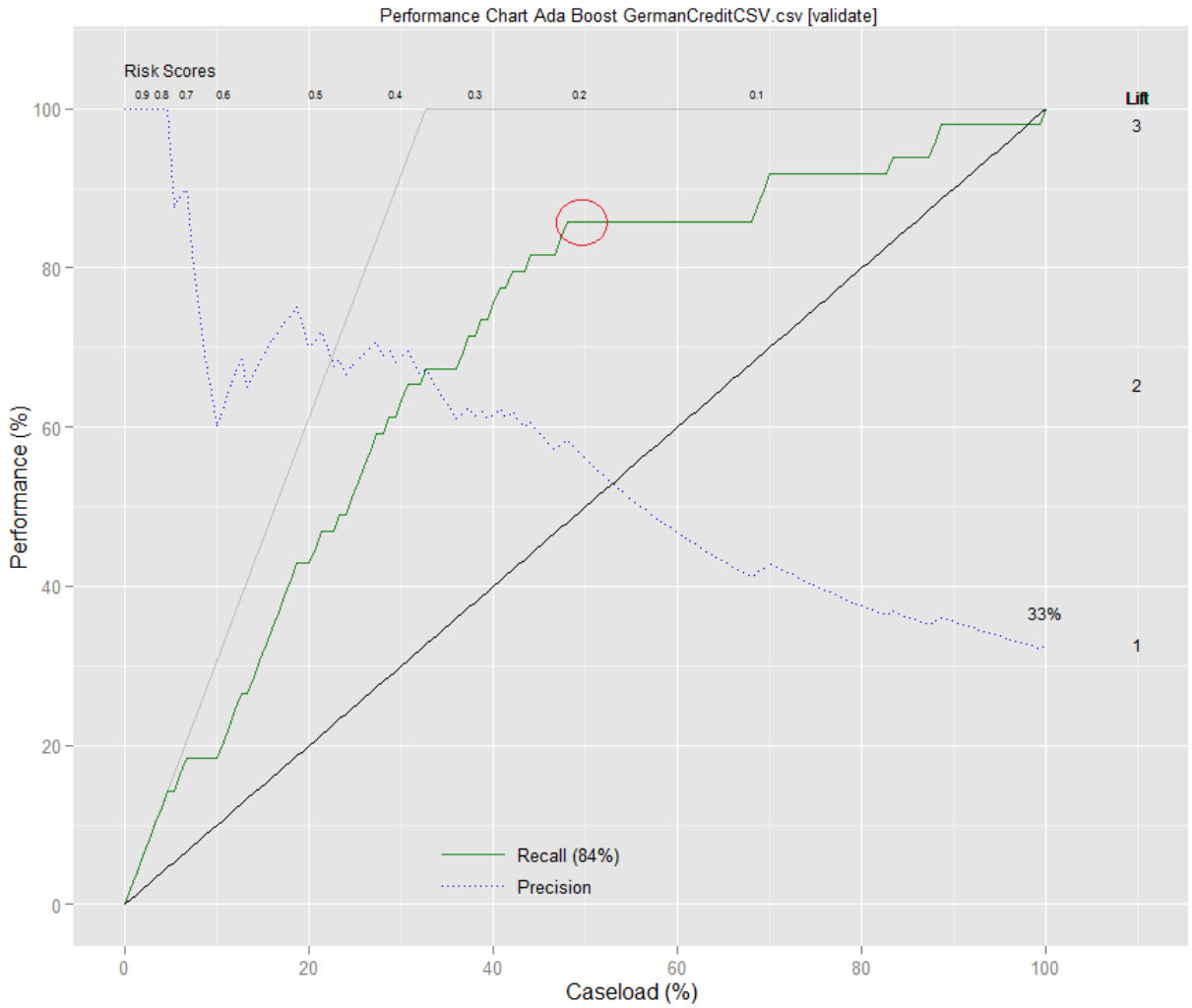
$$Accuracy = \frac{600}{1.000} = 0.6$$

$$Sensitivity = \frac{True\ Positive}{Positive} = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$Sensitivity = \frac{True\ Negative}{Negative} = \frac{True\ Negative}{True\ Negative + False\ Negative}$$

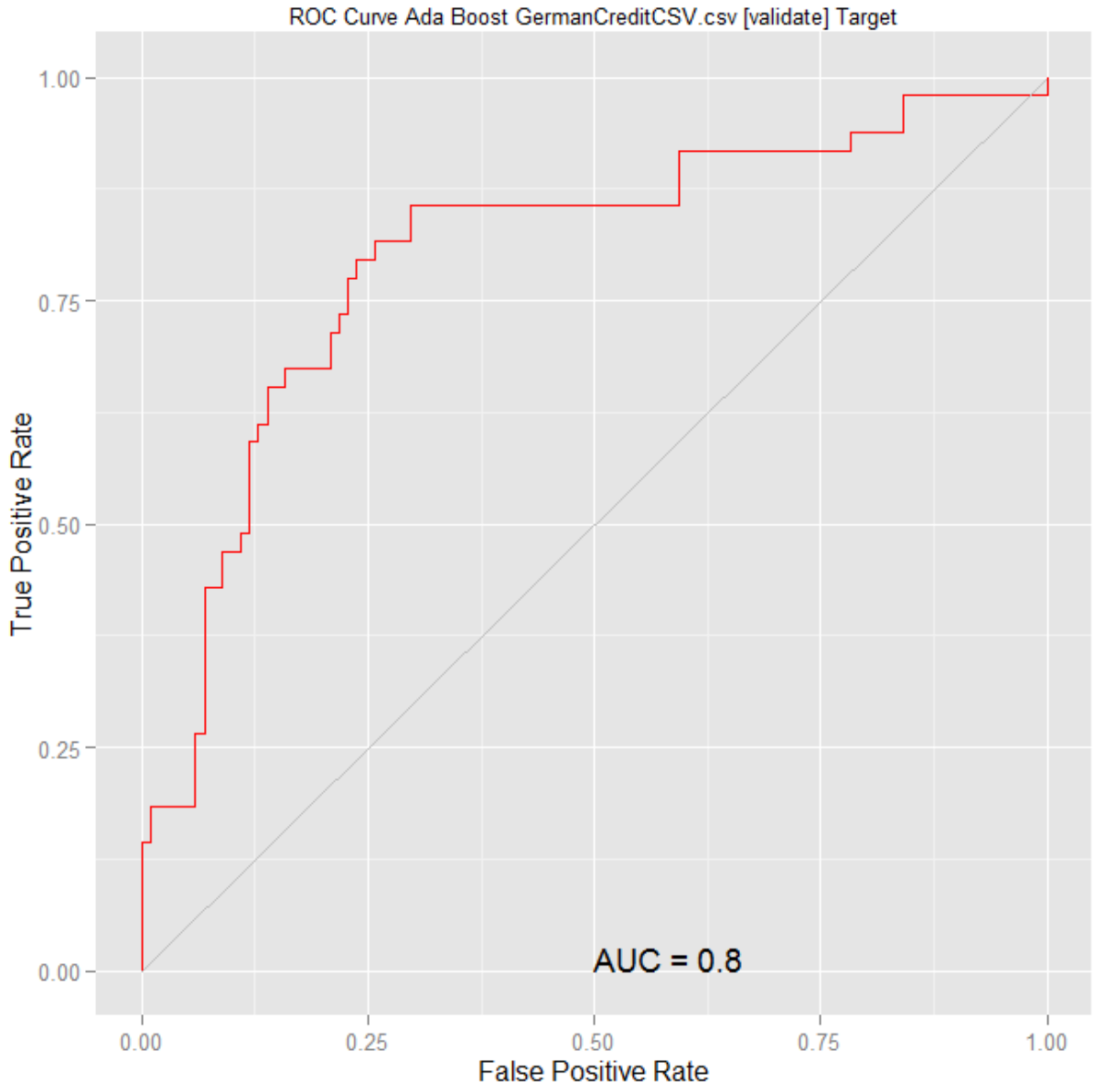

Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score
Model: Tree Boost Forest SVM Linear Neural Net Survival KMeans HClust
Data: Training Validation Testing Full Enter CSV File Docum... R Dataset
Risk Variable:
Report: Class Probability
Include: Identifiers All
Error matrix for the Decision Tree model on GermanCreditCSV.csv [validate] (counts):
Predicted
Actual 1 2
1 96 5
2 36 13
Error matrix for the Decision Tree model on GermanCreditCSV.csv [validate] (proportions):
Predicted
Actual 1 2 Error
1 0.64 0.03 0.05
2 0.24 0.09 0.73
Overall error: 0.2733333, Averaged class error: 0.2752525
Rattle timestamp: 2014-11-21 10:12:06 fgs
Error matrix for the Random Forest model on GermanCreditCSV.csv [validate] (counts):
Predicted
Actual 1 2
1 90 11
2 34 15
Error matrix for the Random Forest model on GermanCreditCSV.csv [validate] (proportions):
Predicted
Actual 1 2 Error
1 0.60 0.07 0.11
2 0.23 0.10 0.69
Overall error: 0.3, Averaged class error: 0.3486352
Rattle timestamp: 2014-11-21 10:12:06


Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score
Model: Tree Boost Forest SVM Linear Neural Net Survival KMeans HClust
Data: Training Validation Testing Full Enter CSV File Docum... R Dataset
Risk Variable:
Report: Class Probability
Include: Identifiers All

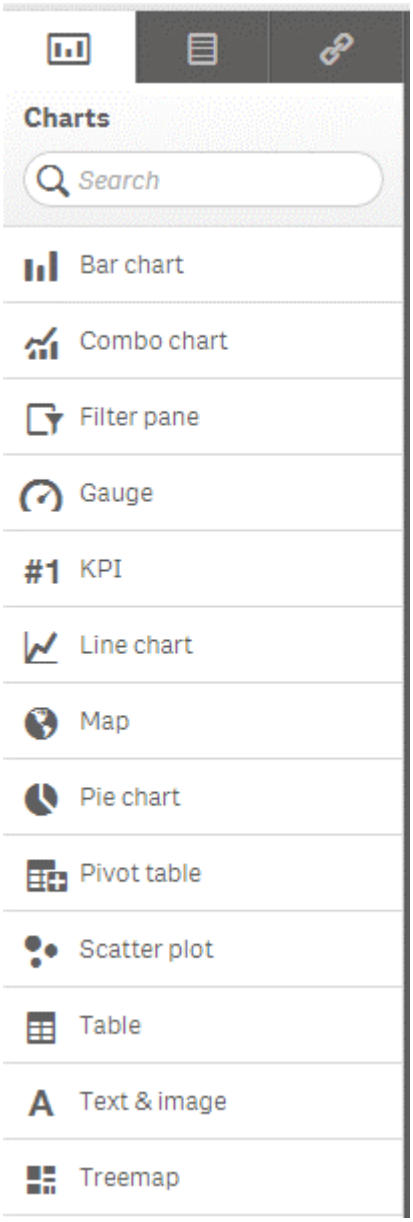


$$Precision = \frac{True\ Positive}{True\ Positive - False\ Positive}$$

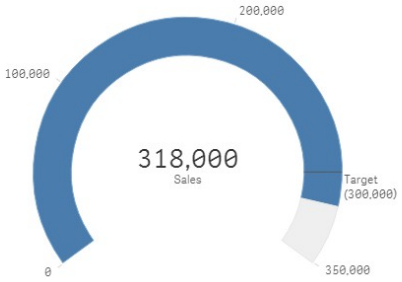
$$Recall = Sensitivity = \frac{True\ Positive}{Positive} = \frac{True\ Positive}{True\ Positive - False\ Negative}$$



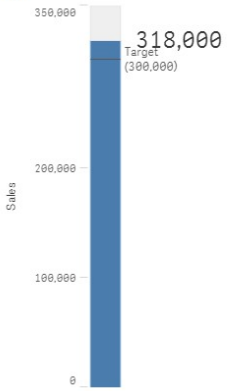
Chapter 8: Visualizations, Data Applications, Dashboards, and Data Storytelling



Radial Gauge



Bar Gauge



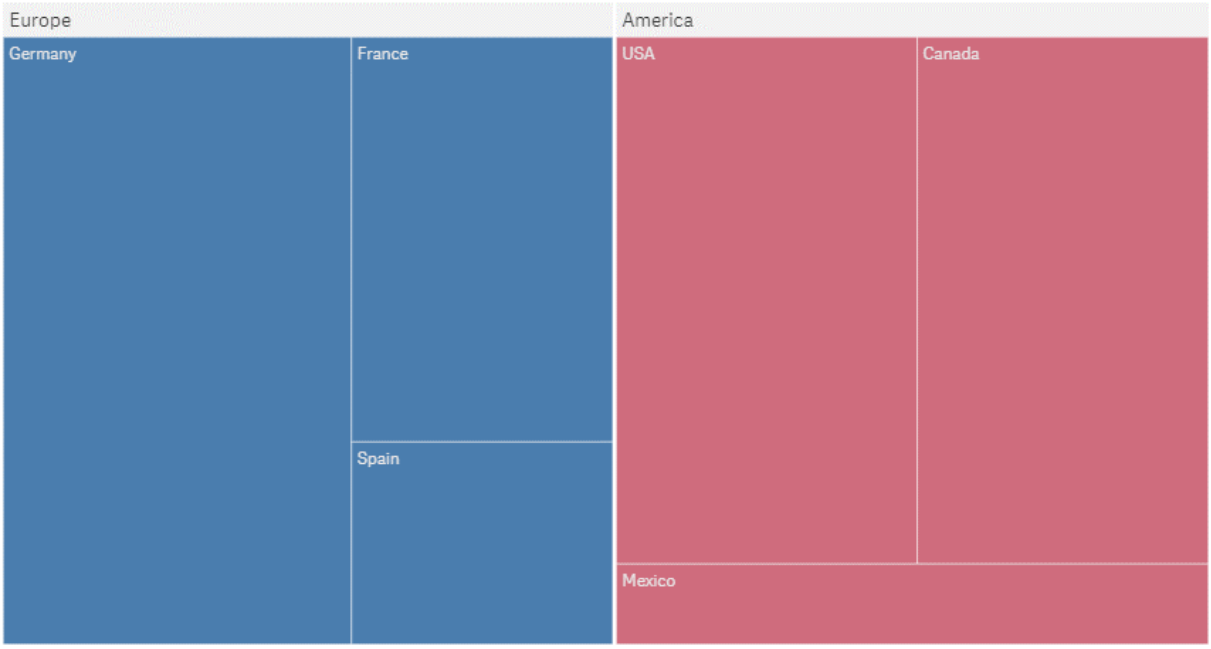
KPI

Target 300k

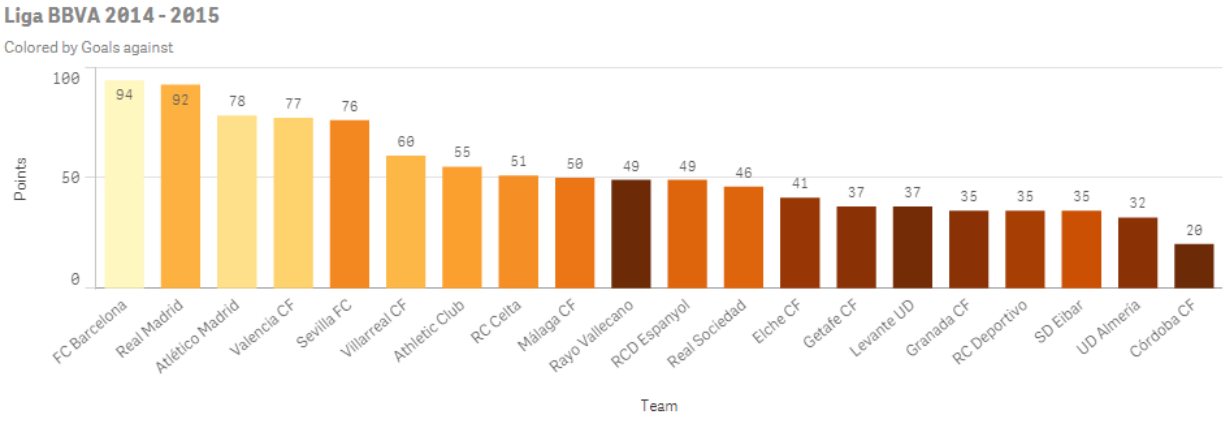
Sales

318k

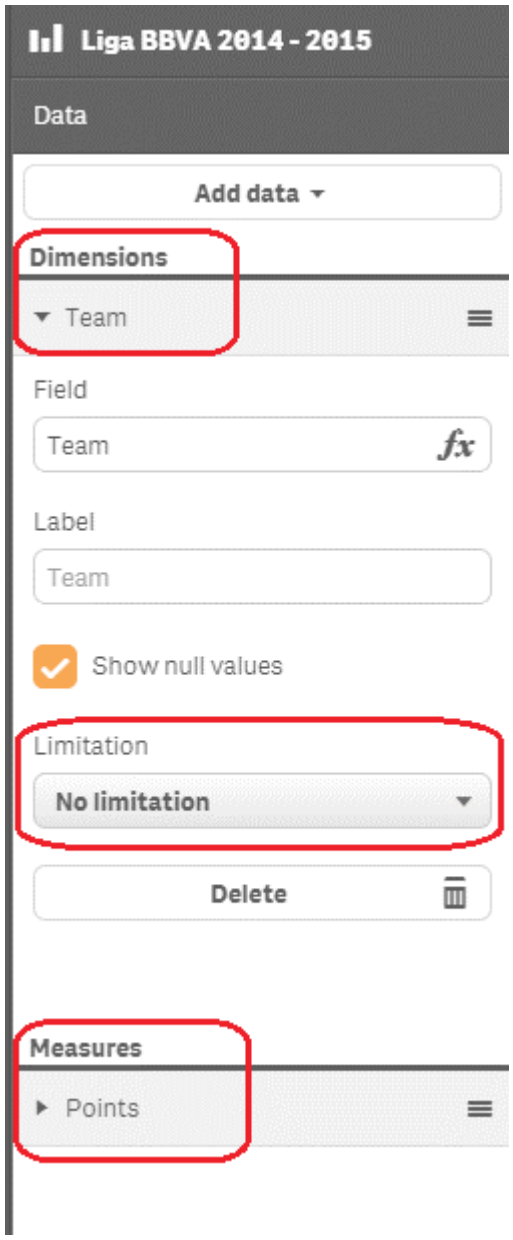
Sales by Geography



Rank	Team	Points	Played	Won	Drawn	Lost	Golas For	Goals Against
1	FC Barcelona	94	38	30	4	4	110	21
2	Real Madrid	92	38	30	2	6	118	38
3	Atlético Madrid	78	38	23	9	6	67	29
4	Valencia CF	77	38	22	11	5	70	32
5	Sevilla FC	76	38	23	7	8	71	45
6	Villarreal CF	60	38	16	12	10	48	37
7	Athletic Club	55	38	15	10	13	42	41
8	RC Celta	51	38	13	12	13	47	44
9	Málaga CF	50	38	14	8	16	42	48
10	RCD Espanyol	49	38	13	10	15	47	51
11	Rayo Vallecano	49	38	15	4	19	46	68
12	Real Sociedad	46	38	11	13	14	44	51
13	Elche CF	41	38	11	8	19	35	62
14	Levante UD	37	38	9	10	19	34	67
15	Getafe CF	37	38	10	7	21	33	64
16	RC Deportivo	35	38	7	14	17	35	60
17	Granada CF	35	38	7	14	17	29	64
18	SD Eibar	35	38	9	8	21	34	55
19	UD Almería	32	38	8	8	22	35	64
20	Córdoba CF	20	38	3	11	24	22	68



Final Classification


Liga BBVA 2014 - 2015
Data
Add data
Dimensions
Team
Field
Team
Label
Team
Show null values
Limitation
No limitation
Delete
Measures
Points

8	RC Celta	51	38	13	12	13	47	44
9		50	38	14	8	16	42	48
10	RCD Espanyol	49	38	13	10	15	47	51

Limitation

No limitation

Limitation

Fixed number

Top | **Bottom**

10 fx

Calculated on measure: Points

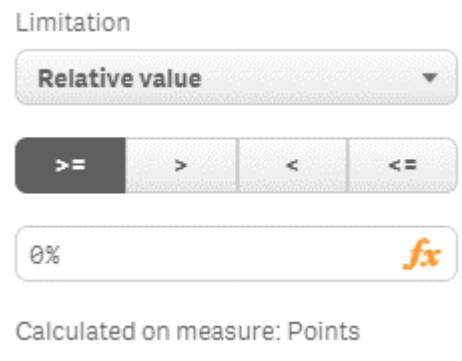
Limitation

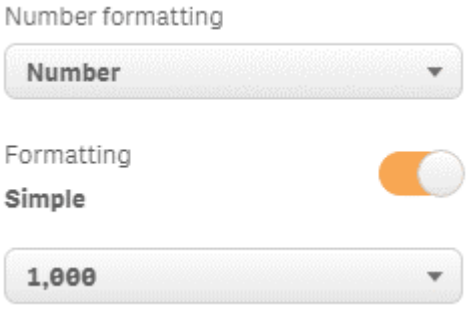
Exact value

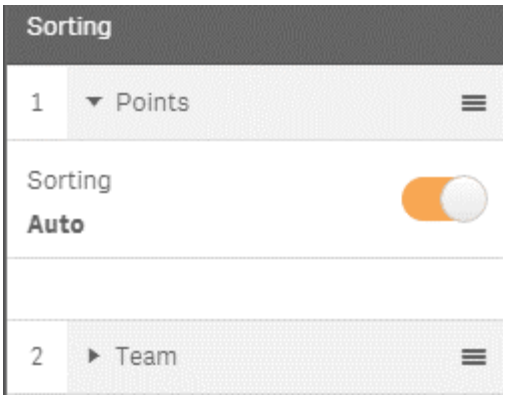
>= | **>** | **<** | **<=**

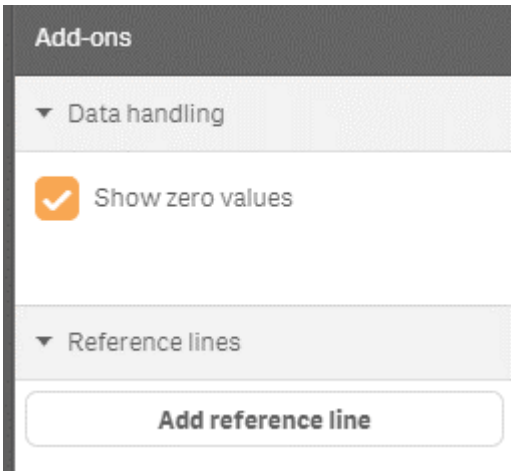
0 fx

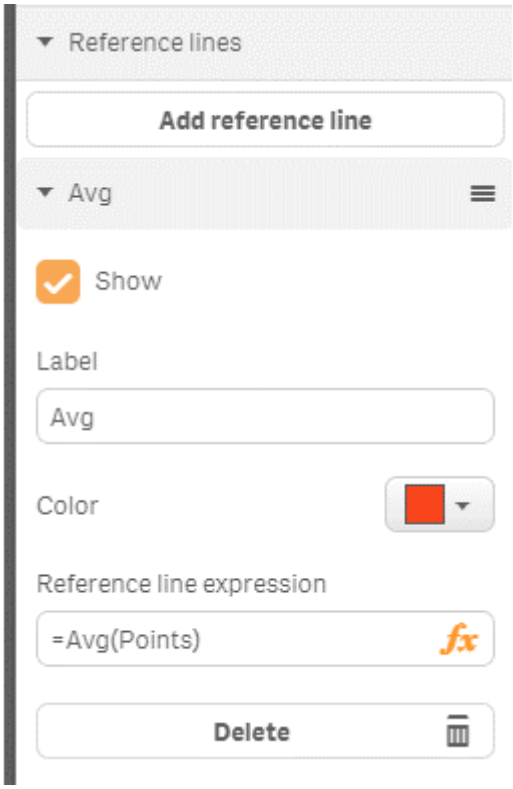
Calculated on measure: Points

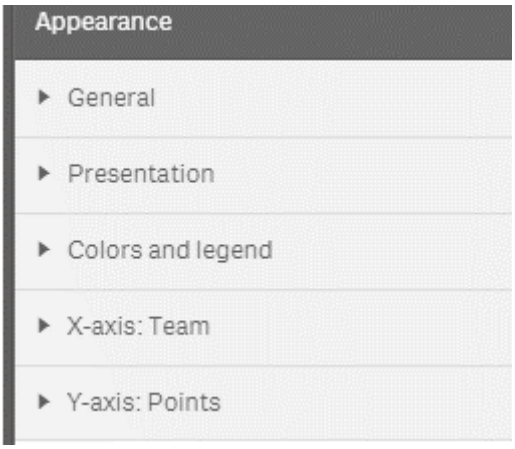

Limitation
Relative value
>=
>
<
<=
0%
Calculated on measure: Points

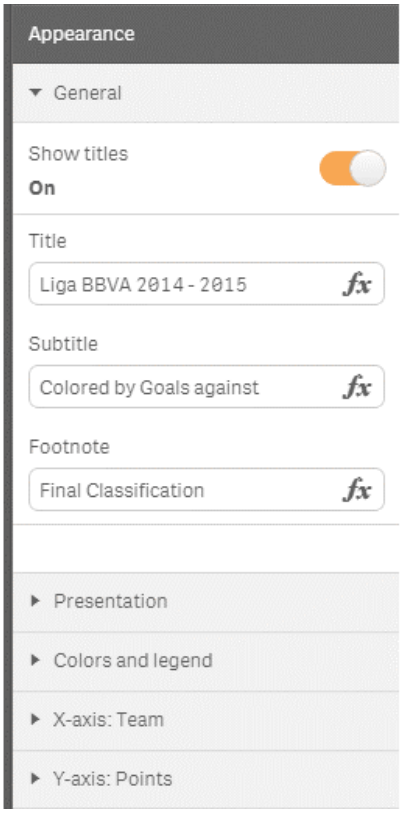

Number formatting
Number
Formatting
Simple
1,000

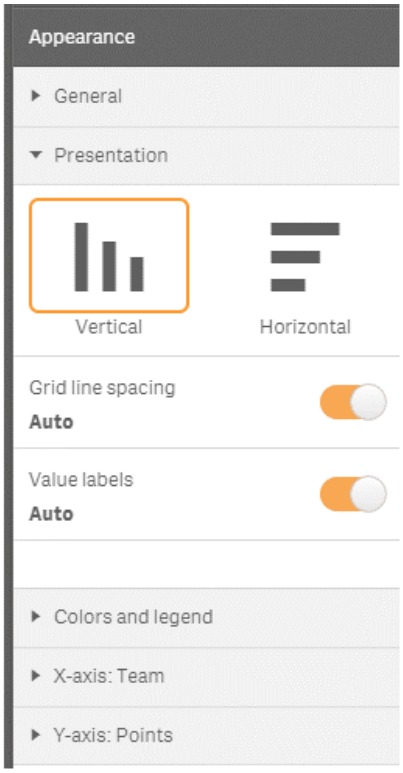

Sorting
1
Points
Sorting
Auto
2
Team

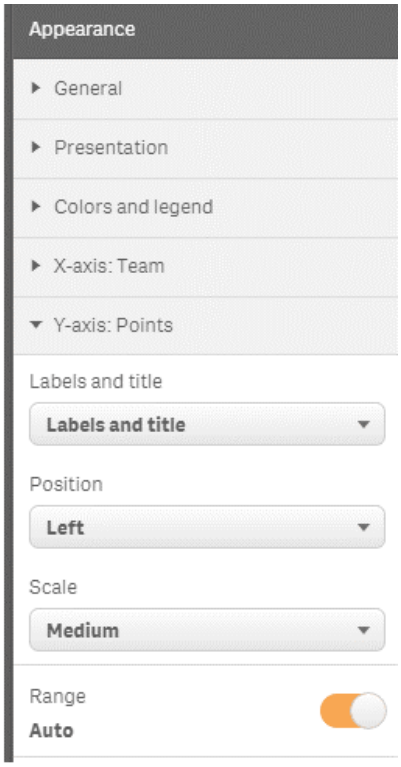

Add-ons
Data handling
Show zero values
Reference lines
Add reference line

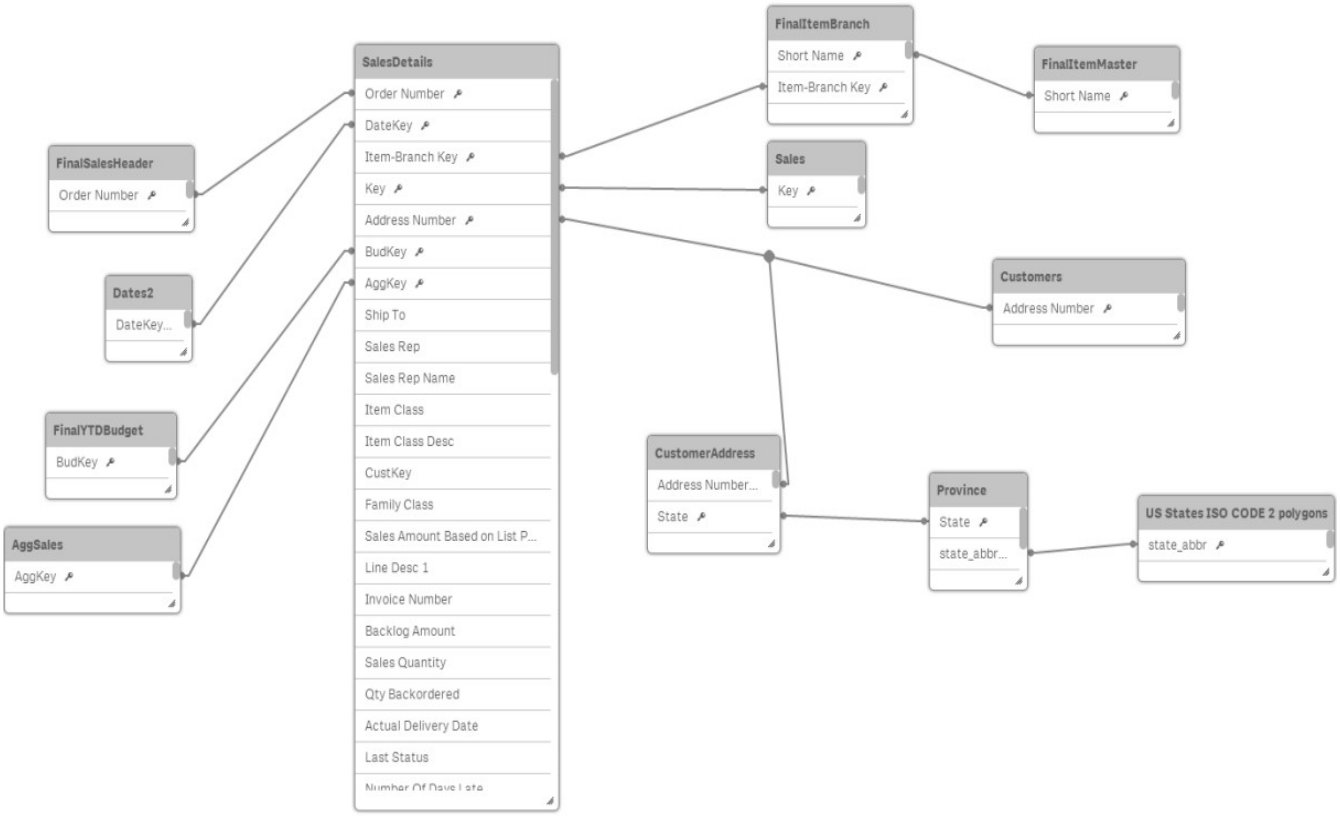

Reference lines
Add reference line
Avg
Show
Label
Avg
Color
Reference line expression
=Avg(Points)
Delete

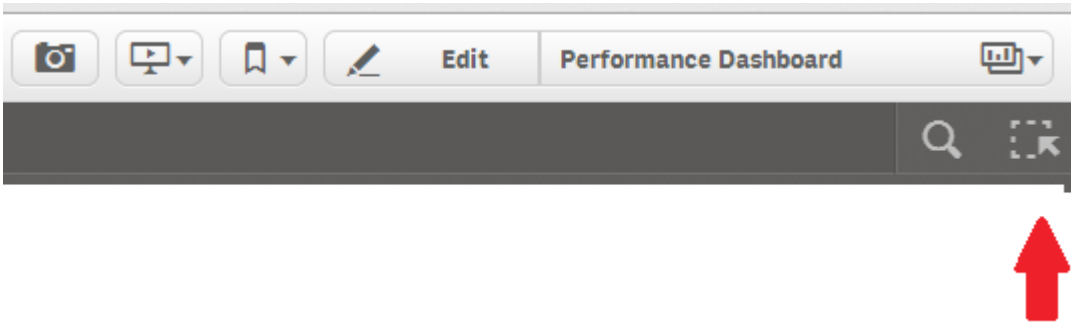

Appearance
General
Presentation
Colors and legend
X-axis: Team
Y-axis: Points

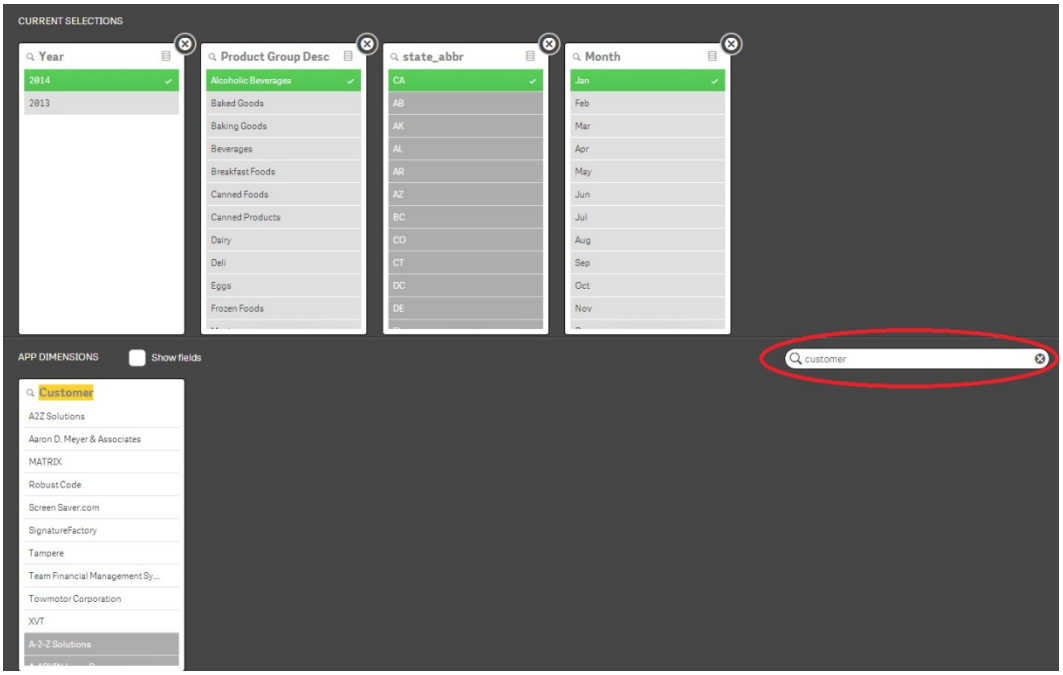

Appearance
General
Show titles
On
Title
Liga BBVA 2014 - 2015
Subtitle
Colored by Goals against
Footnote
Final Classification
Presentation
Colors and legend
X-axis: Team
Y-axis: Points


Appearance
General
Presentation
Vertical
Horizontal
Grid line spacing
Auto
Value labels
Auto
Colors and legend
X-axis: Team
Y-axis: Points


Appearance
General
Presentation
Colors and legend
X-axis: Team
Y-axis: Points
Labels and title
Labels and title
Position
Left
Scale
Medium
Range
Auto


SalesDetails
Order Number
DateKey
Item-Branch Key
Key
Address Number
BudKey
AggKey
Ship To
Sales Rep
Sales Rep Name
Item Class
Item Class Desc
CustKey
Family Class
Sales Amount Based on List P...
Line Desc 1
Invoice Number
Backlog Amount
Sales Quantity
Qty Backordered
Actual Delivery Date
Last Status
FinalItemBranch
Short Name
Item-Branch Key
FinalItemMaster
Short Name
FinalSalesHeader
Order Number
Sales
Key
Dates2
DateKey...
Customers
Address Number
FinalYTDBudget
BudKey
CustomerAddress
Address Number...
State
Province
State
state_abbr...
US States ISO CODE 2 polygons
state_abbr
AggSales
AggKey


Edit
Performance Dashboard


CURRENT SELECTIONS
Year
2014
2013
Product Group Desc
Alcoholic Beverages
Baked Goods
Baking Goods
Beverages
Breakfast Foods
Canned Foods
Canned Products
Dairy
Deli
Eggs
Frozen Foods
state_abbr
CA
AB
AK
AL
AR
AZ
BC
CO
CT
DC
DE
Month
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
APP DIMENSIONS
Show fields
customer
Customer
A2Z Solutions
Aaron D. Meyer & Associates
MATRIX
Robust Code
Screen Saver.com
SignatureFactory
Tampere
Team Financial Management Sy...
Towmotor Corporation
XVT
A-2-Z Solutions

Performance Dashboard

Division

Region

State

Sales Rep

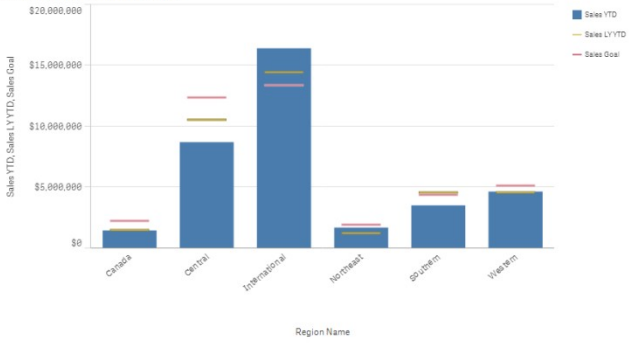
- Judy Thurman
- Stewart Wind
- Lee Chin
- TAGnology
- Brad Taylor
- John Greg
- Cheryle Sincock
- Brenda Gibson
- David Laychak
- Martha Richard
- Amalia Craig
- Max Blagburn
- Karl Anderson
- David Howard
- Angelen Carter
- Amanda Honda
- Amelia Fields
- Donna Brown
- Deborah Halmon
- Peggie Hurt
- Scott Powell

Sales Variation
-1.71%
YTD: $36,173,882
LYTD: $36,803,513

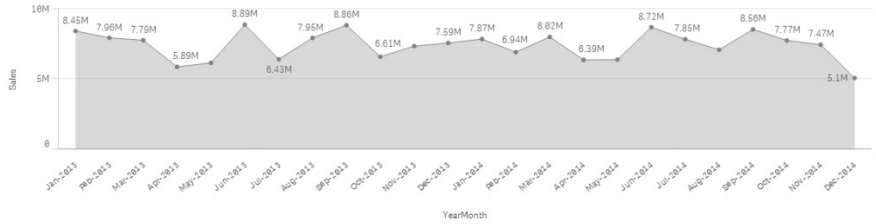
Margin Variation
2.13%
YTD: $15,849,041
LYTD: $15,519,111

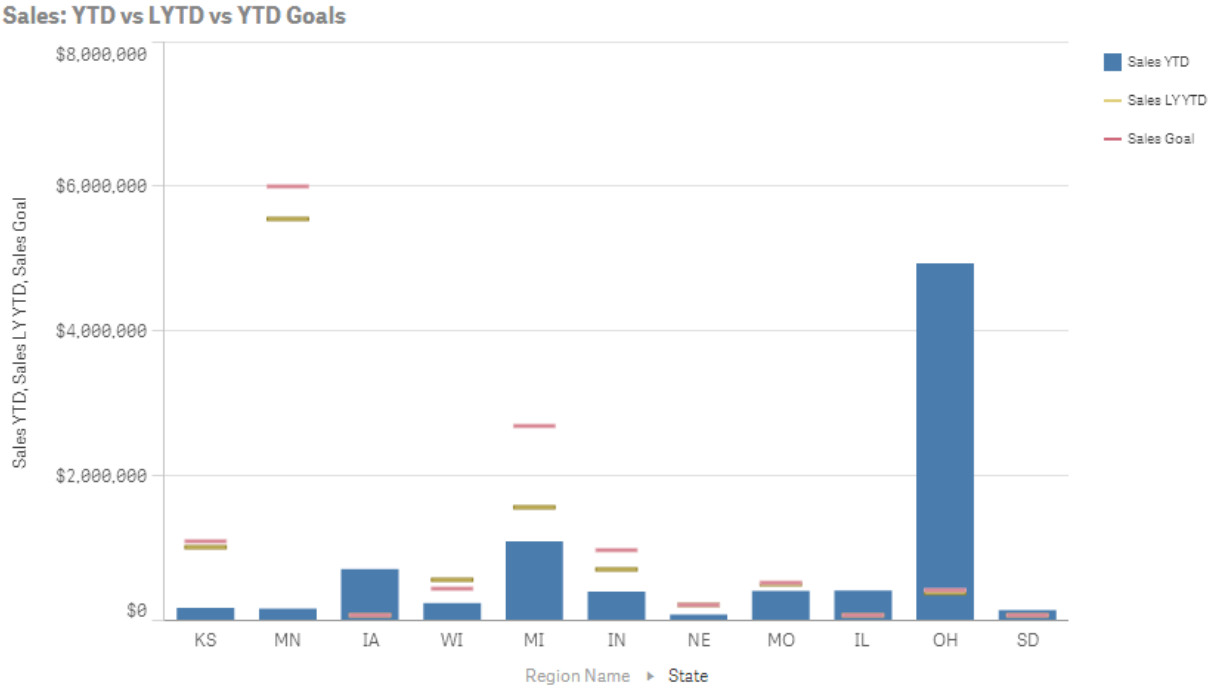
Sales vs Budget
-8.29%
Sales YTD: $36,173,882
Sales Goal YTD: $39,444,059

Sales: YTD vs LYTD vs YTD Goals

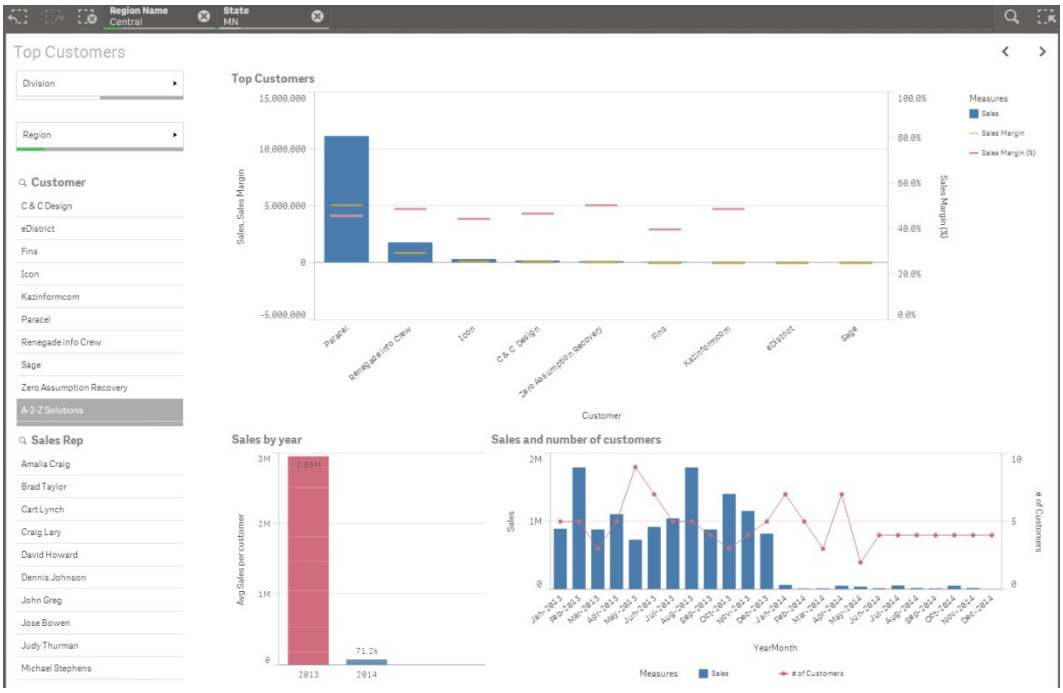


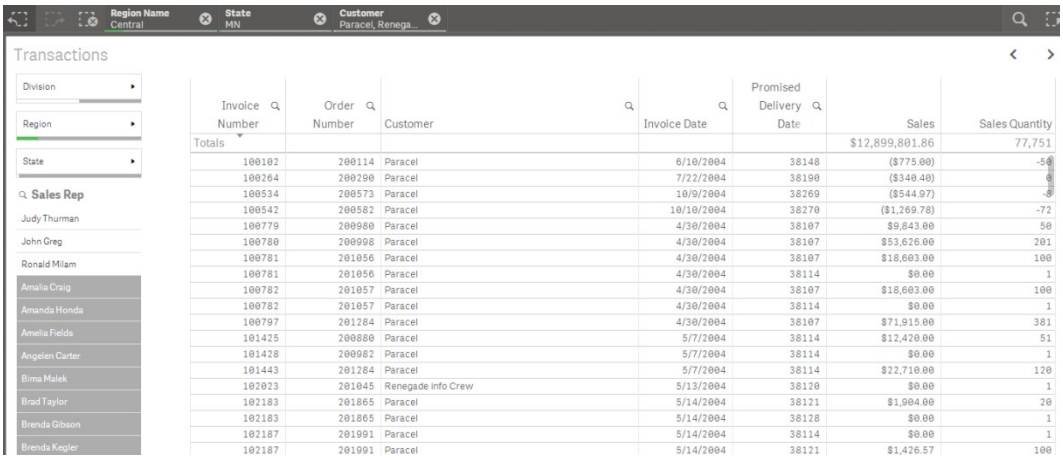
Sales Evolution



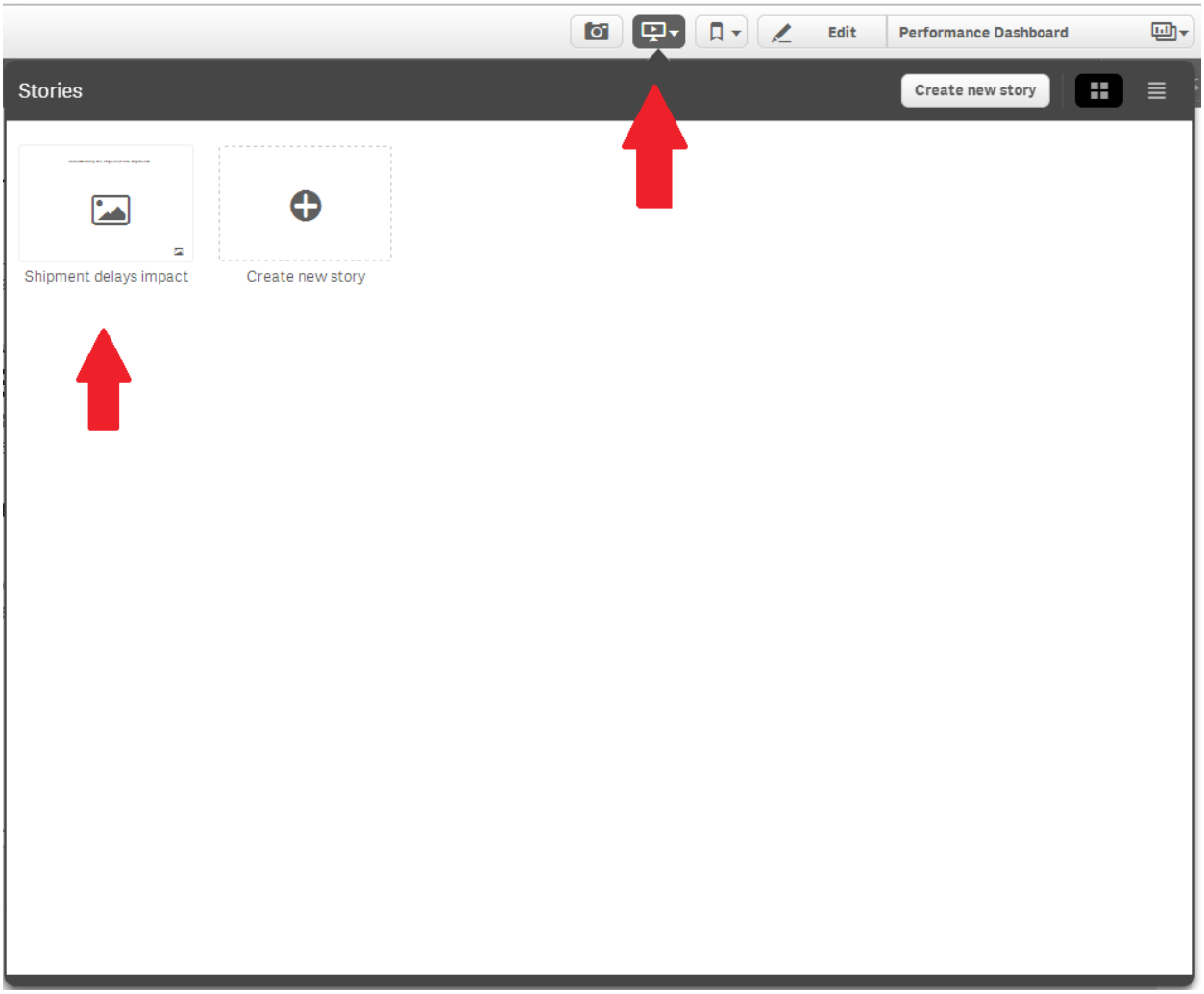

Sales: YTD vs LYTD vs YTD Goals
$8,000,000
$6,000,000
$4,000,000
$2,000,000
$0
Sales YTD, Sales LYYTD, Sales Goal
KS
MN
IA
WI
MI
IN
NE
MO
IL
OH
SD
Region Name ▸ State
Sales YTD
Sales LYYTD
Sales Goal

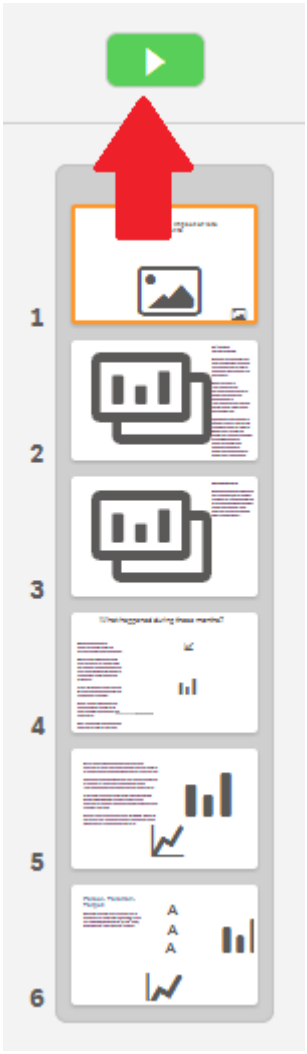
Region Name
Central
State
MN

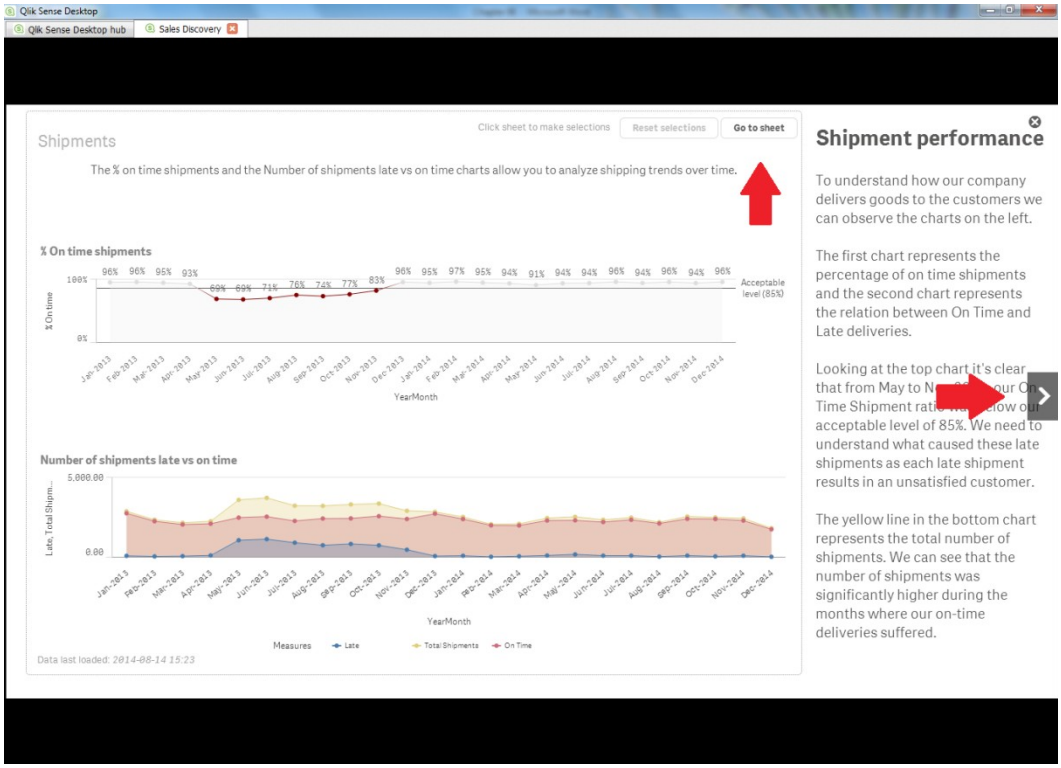


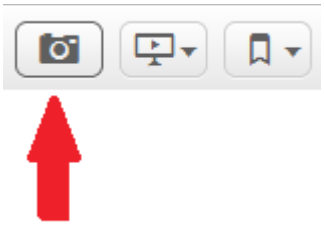


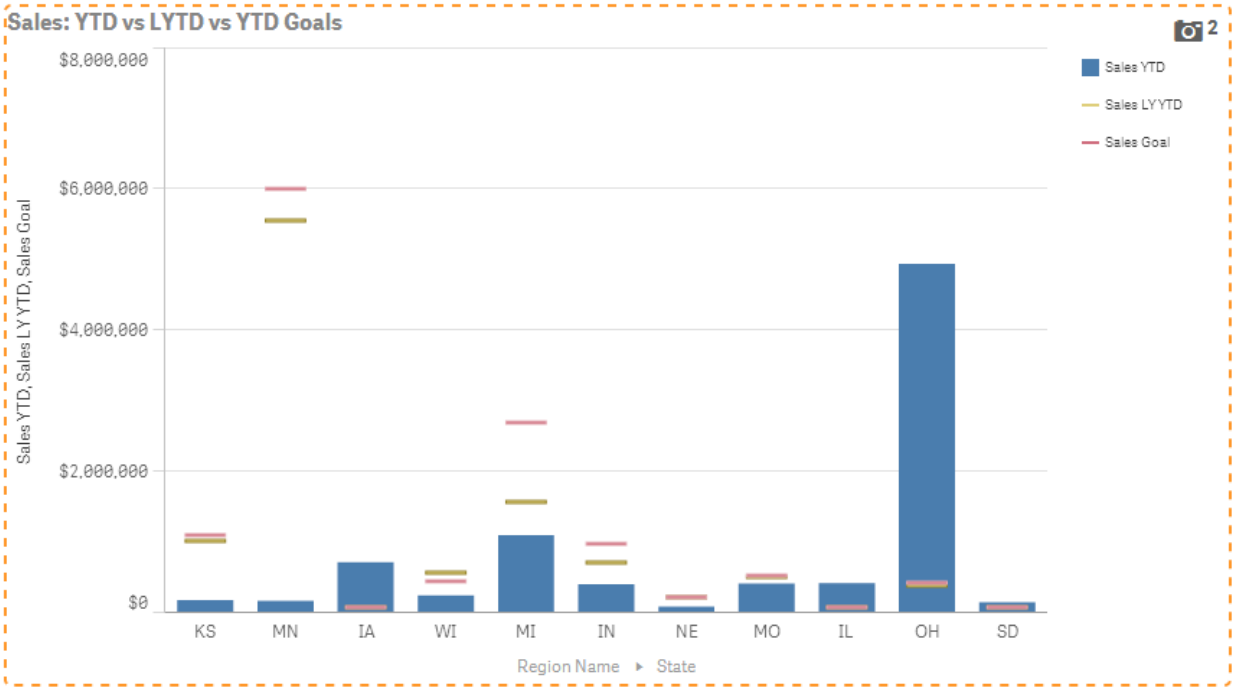
Invoice Number	Order Number	Customer	Invoice Date	Promised Delivery Date	Sales	Sales Quantity
Totals					$12,899,801.86	77,751
100102	200114	Paracel	6/10/2004	38148	($775.00)	-50
100264	200290	Paracel	7/22/2004	38190	($340.40)	-8
100534	200573	Paracel	10/9/2004	38269	($544.97)	-9
100542	200582	Paracel	10/10/2004	38270	($1,269.78)	-72
100779	200980	Paracel	4/30/2004	38107	$9,843.00	50
100780	200998	Paracel	4/30/2004	38107	$53,626.00	201
100781	201056	Paracel	4/30/2004	38107	$18,603.00	100
100781	201056	Paracel	4/30/2004	38114	$0.00	1
100782	201057	Paracel	4/30/2004	38107	$18,603.00	100
100782	201057	Paracel	4/30/2004	38114	$0.00	1
100797	201284	Paracel	4/30/2004	38107	$71,915.00	381
101425	200880	Paracel	5/7/2004	38114	$12,420.00	51
101428	200982	Paracel	5/7/2004	38114	$0.00	1
101443	201284	Paracel	5/7/2004	38114	$22,710.00	120
102023	201045	Renegade info Crew	5/13/2004	38120	$0.00	1
102183	201865	Paracel	5/14/2004	38121	$1,904.00	20
102183	201865	Paracel	5/14/2004	38128	$0.00	1
102187	201991	Paracel	5/14/2004	38114	$0.00	1
102187	201991	Paracel	5/14/2004	38121	$1,426.57	100

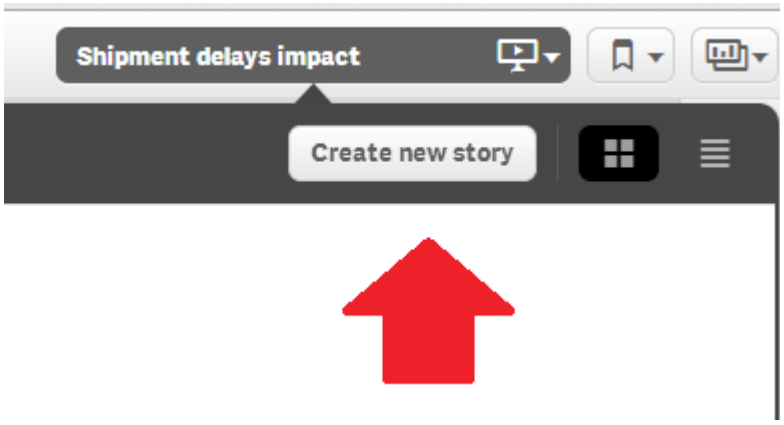

Edit
Performance Dashboard
Stories
Create new story
Shipment delays impact
Create new story


1
2
3
4
5
6

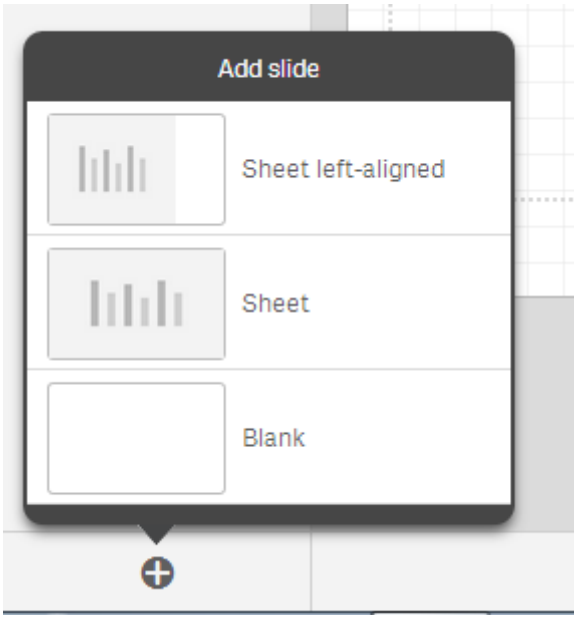

Qlik Sense Desktop
Qlik Sense Desktop hub
Sales Discovery
Shipments
Click sheet to make selections
Reset selections
Go to sheet
The % on time shipments and the Number of shipments late vs on time charts allow you to analyze shipping trends over time.
% On time shipments
Acceptable level (85%)
YearMonth
Number of shipments late vs on time
Measures
Late
Total Shipments
On Time
Data last loaded: 2014-08-14 15:23
Shipment performance
To understand how our company delivers goods to the customers we can observe the charts on the left.
The first chart represents the percentage of on time shipments and the second chart represents the relation between On Time and Late deliveries.
Looking at the top chart it's clear that from May to N
our On Time Shipment rati
low our acceptable level of 85%. We need to understand what caused these late shipments as each late shipment results in an unsatisfied customer.
The yellow line in the bottom chart represents the total number of shipments. We can see that the number of shipments was significantly higher during the months where our on-time deliveries suffered.

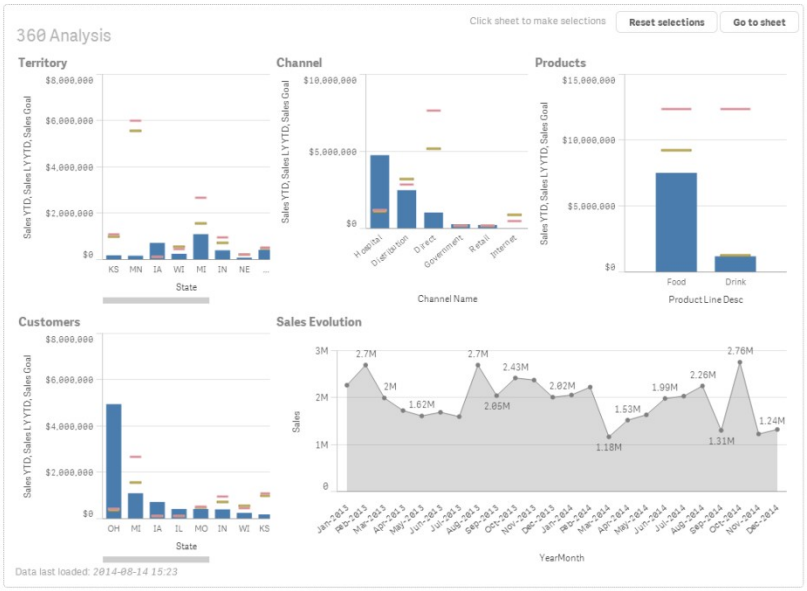



Sales: YTD vs LYTD vs YTD Goals
2
$8,000,000
$6,000,000
$4,000,000
$2,000,000
$0
Sales YTD, Sales LY YTD, Sales Goal
KS
MN
IA
WI
MI
IN
NE
MO
IL
OH
SD
Region Name ▸ State
Sales YTD
Sales LY YTD
Sales Goal


Shipment delays impact
Create new story

Camera tool.
Text
Shapes
Effects
Images

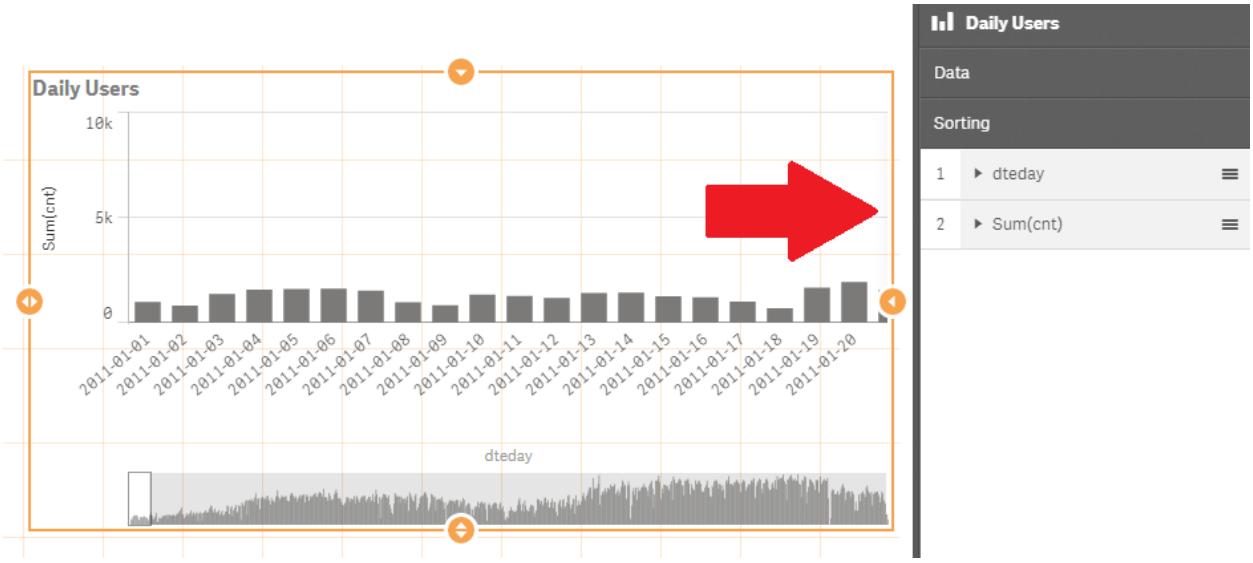

Add slide
Sheet left-aligned
Sheet
Blank



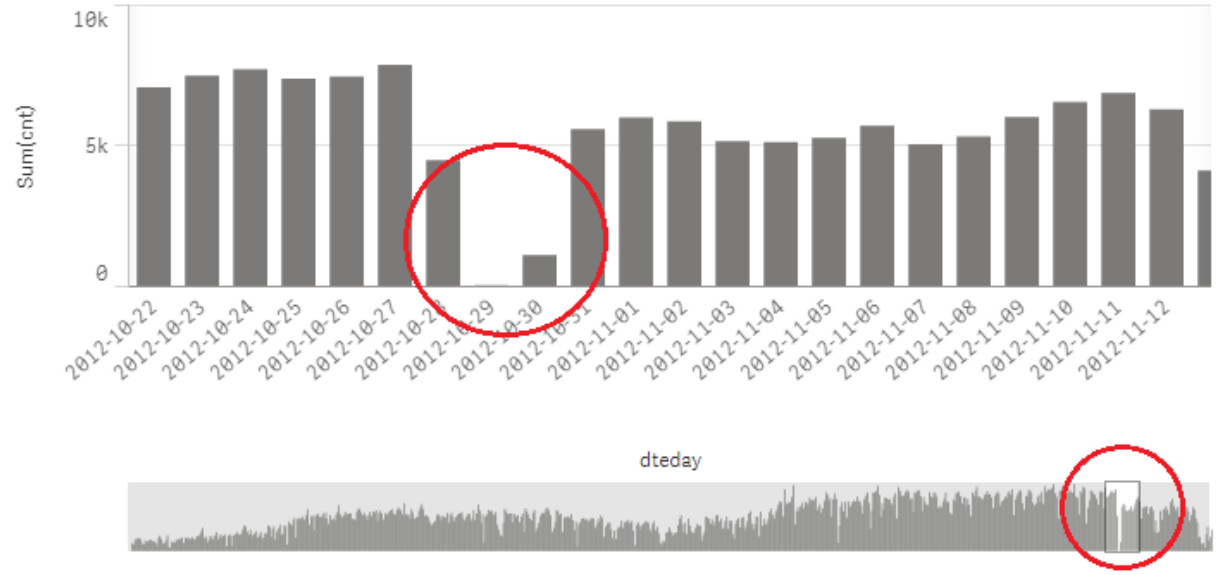
Use this slide to answer the questions of your public.

You can make selections in the charts.

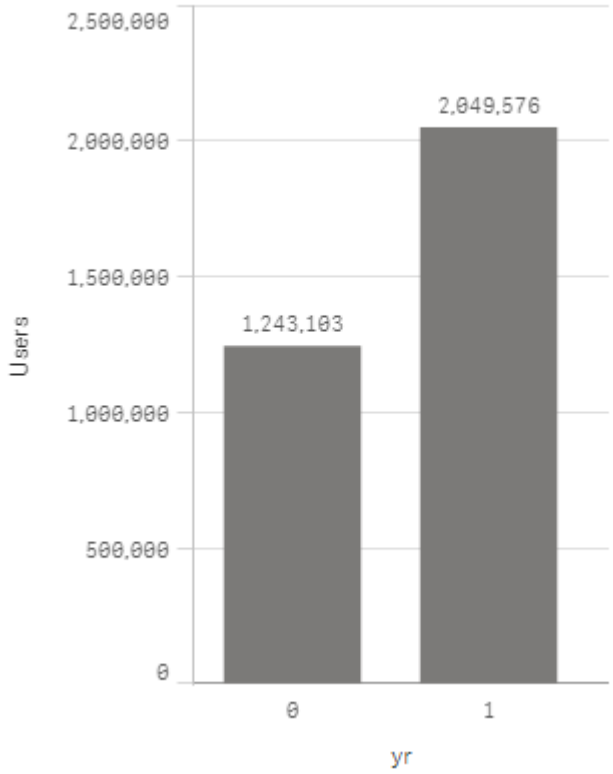
Chapter 9: Developing a Complete Application



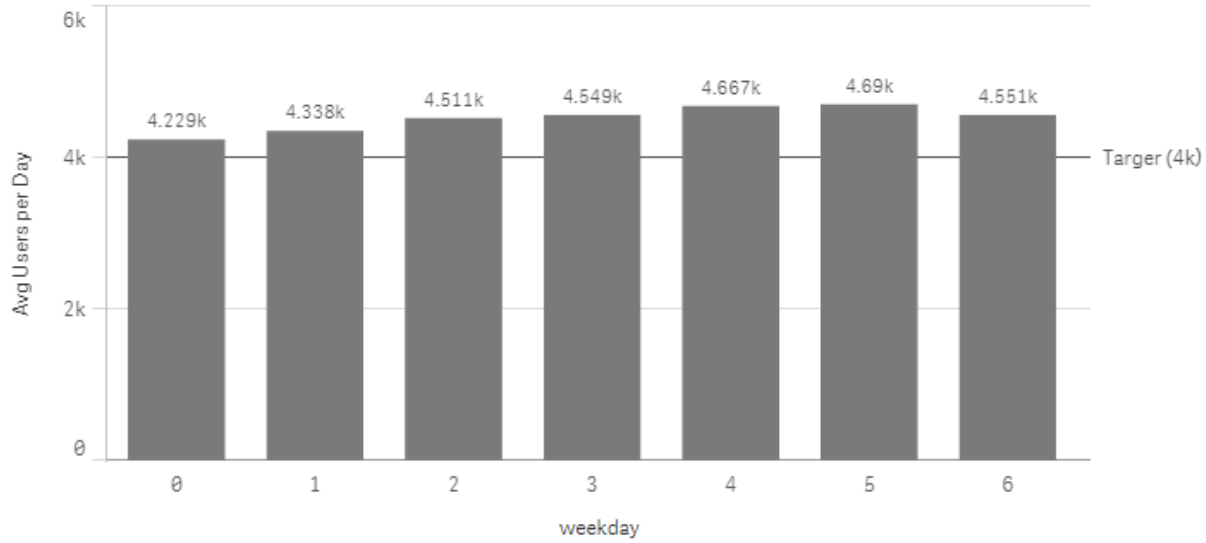
Daily Users


10k
5k
0
Sum(cnt)
2012-10-22
2012-10-23
2012-10-24
2012-10-25
2012-10-26
2012-10-27
2012-10-28
2012-10-29
2012-10-30
2012-10-31
2012-11-01
2012-11-02
2012-11-03
2012-11-04
2012-11-05
2012-11-06
2012-11-07
2012-11-08
2012-11-09
2012-11-10
2012-11-11
2012-11-12
dteday

Year

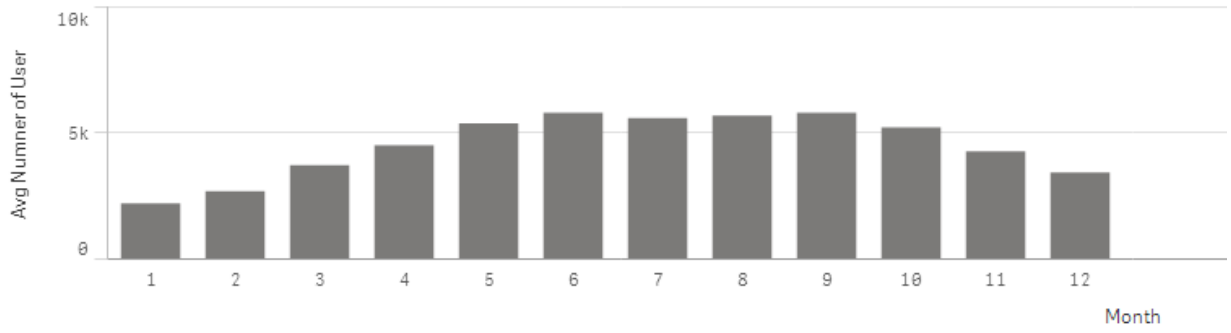

2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
Users
2,049,576
1,243,103
0
1
yr

Day of the Week

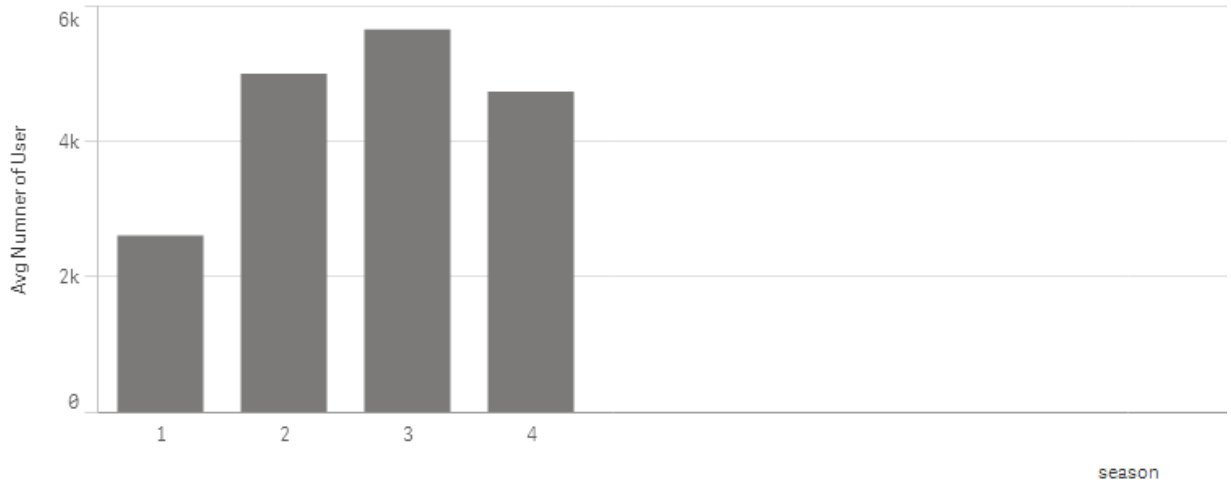


Months and Seasons

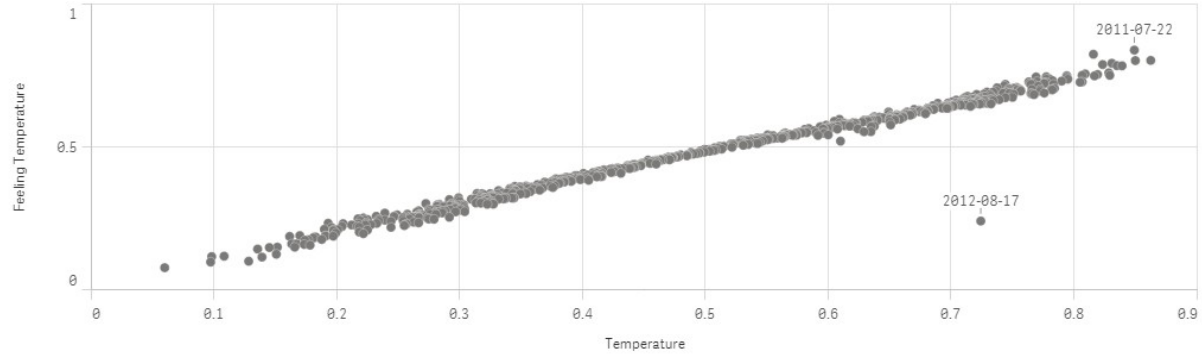
Months



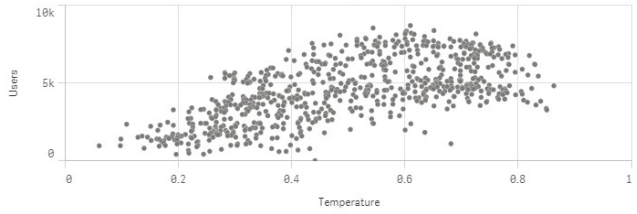
Seasond



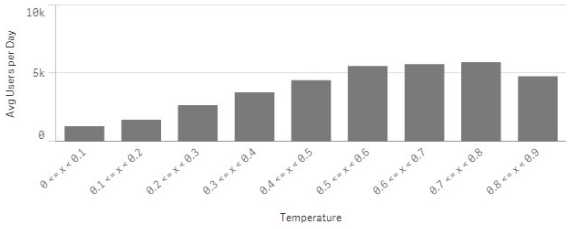
Temperature Vs Feeling Temerature

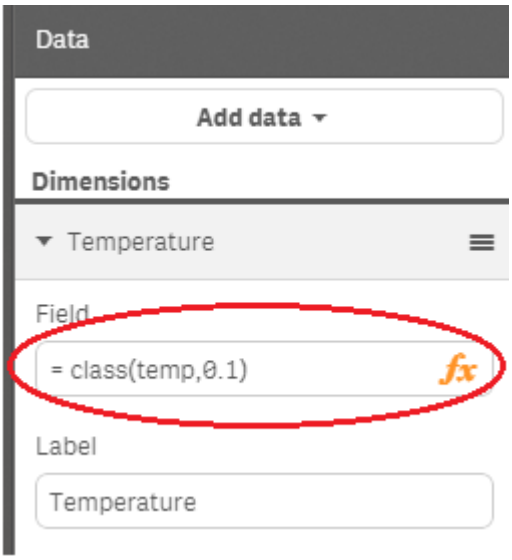


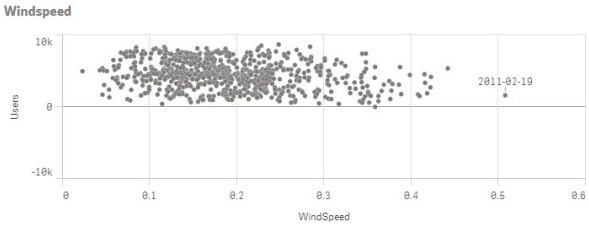
Temperature

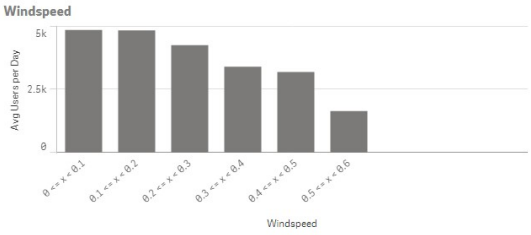


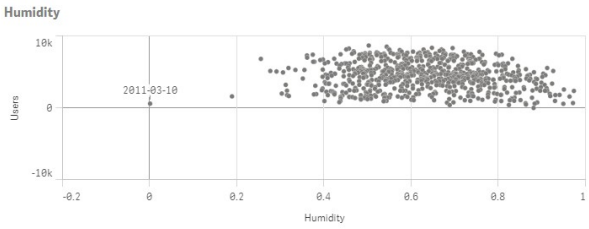
Temperature

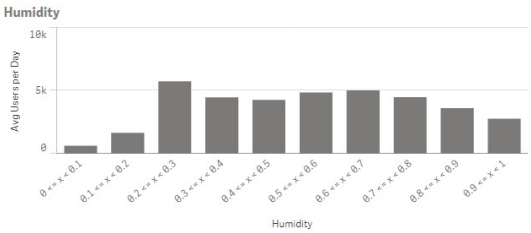



Data
Add data
Dimensions
Temperature
Field
= class(temp,0.1)
Label
Temperature


Windspeed
10k
0
-10k
Users
2011-02-19
0
0.1
0.2
0.3
0.4
0.5
0.6
WindSpeed


Windspeed
5k
2.5k
0
Avg Users per Day
0 <= x < 0.1
0.1 <= x < 0.2
0.2 <= x < 0.3
0.3 <= x < 0.4
0.4 <= x < 0.5
0.5 <= x < 0.6
Windspeed


Humidity
10k
0
-10k
Users
2011-03-10
-0.2
0
0.2
0.4
0.6
0.8
1
Humidity


Humidity
10k
5k
0
Avg Users per Day
0 <= x < 0.1
0.1 <= x < 0.2
0.2 <= x < 0.3
0.3 <= x < 0.4
0.4 <= x < 0.5
0.5 <= x < 0.6
0.6 <= x < 0.7
0.7 <= x < 0.8
0.8 <= x < 0.9
0.9 <= x < 1
Humidity

Dashboard

The target is 4.000 users per day.

Year
0
1

mnth
1
2
3
4
5
6
7
8

weekday
0
1
2
3
4
5
6

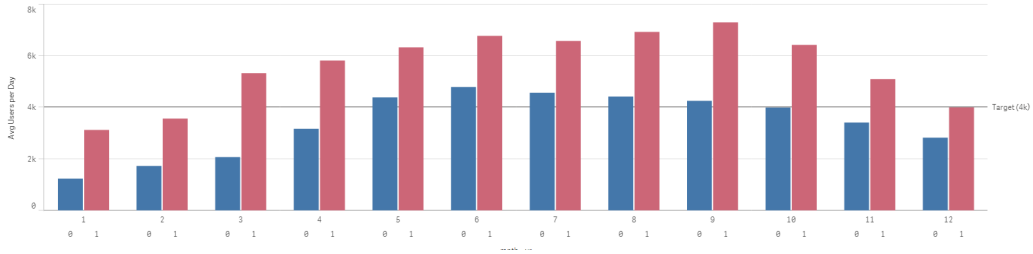
Users
3.29M

Avg Users per Day
4.5k

Avg Registered Users per Day
3.66k

Avg Casual Users per Day
848.2

Monthly Activity



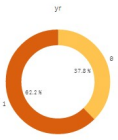
Analysis

Target 4.000

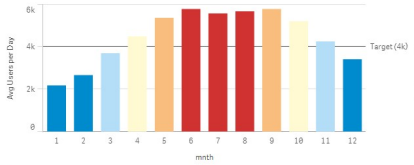
Avg Users per Day
4.5k

Users
3.29M

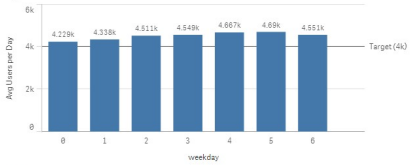
Year



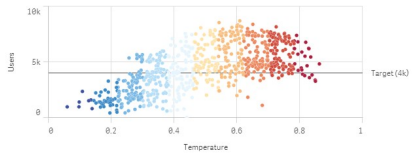
Month



Day of the Week

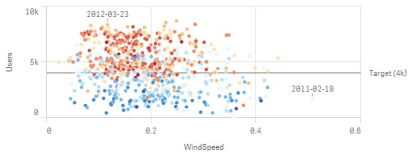


Temperature

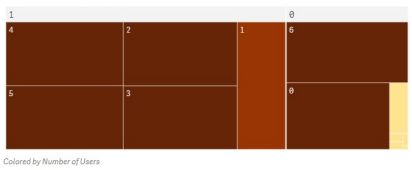


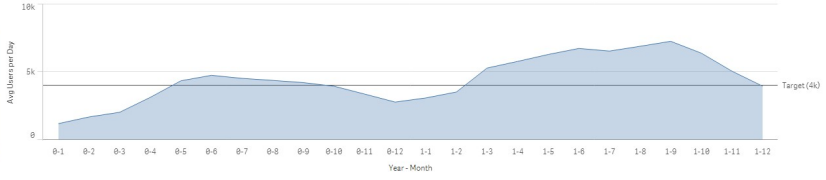
Humidity

Windspeed

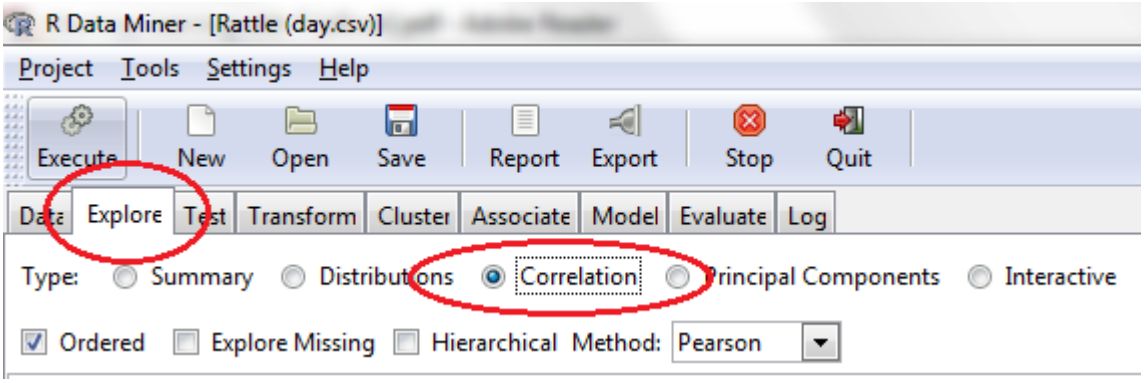


Working day and Weekday

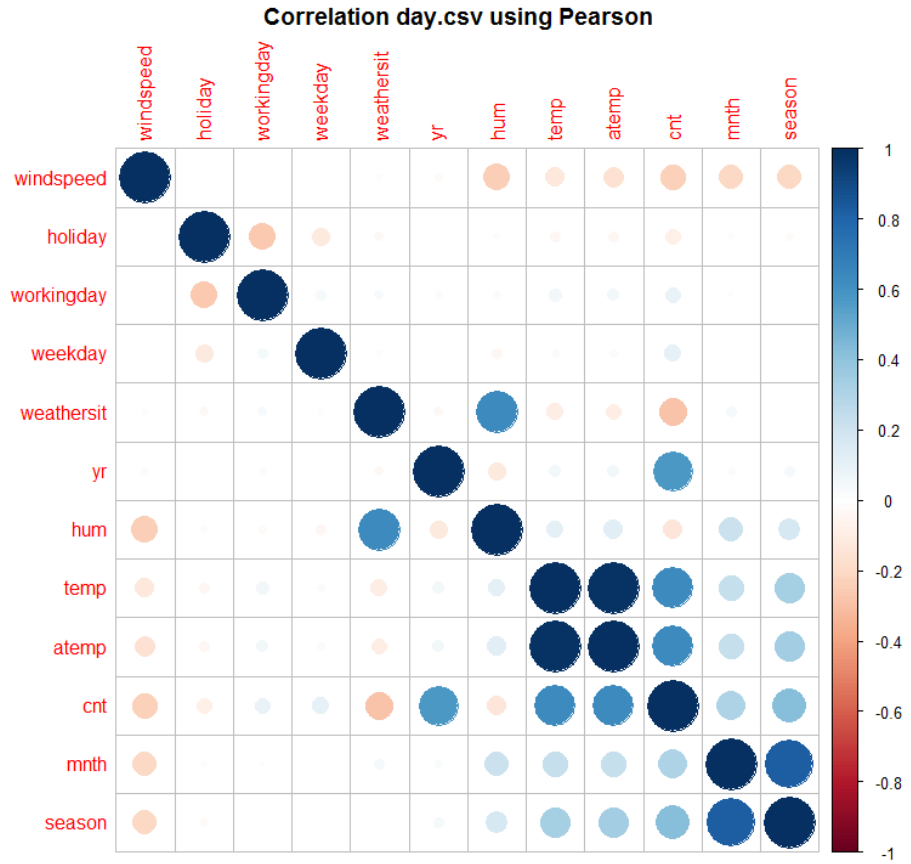


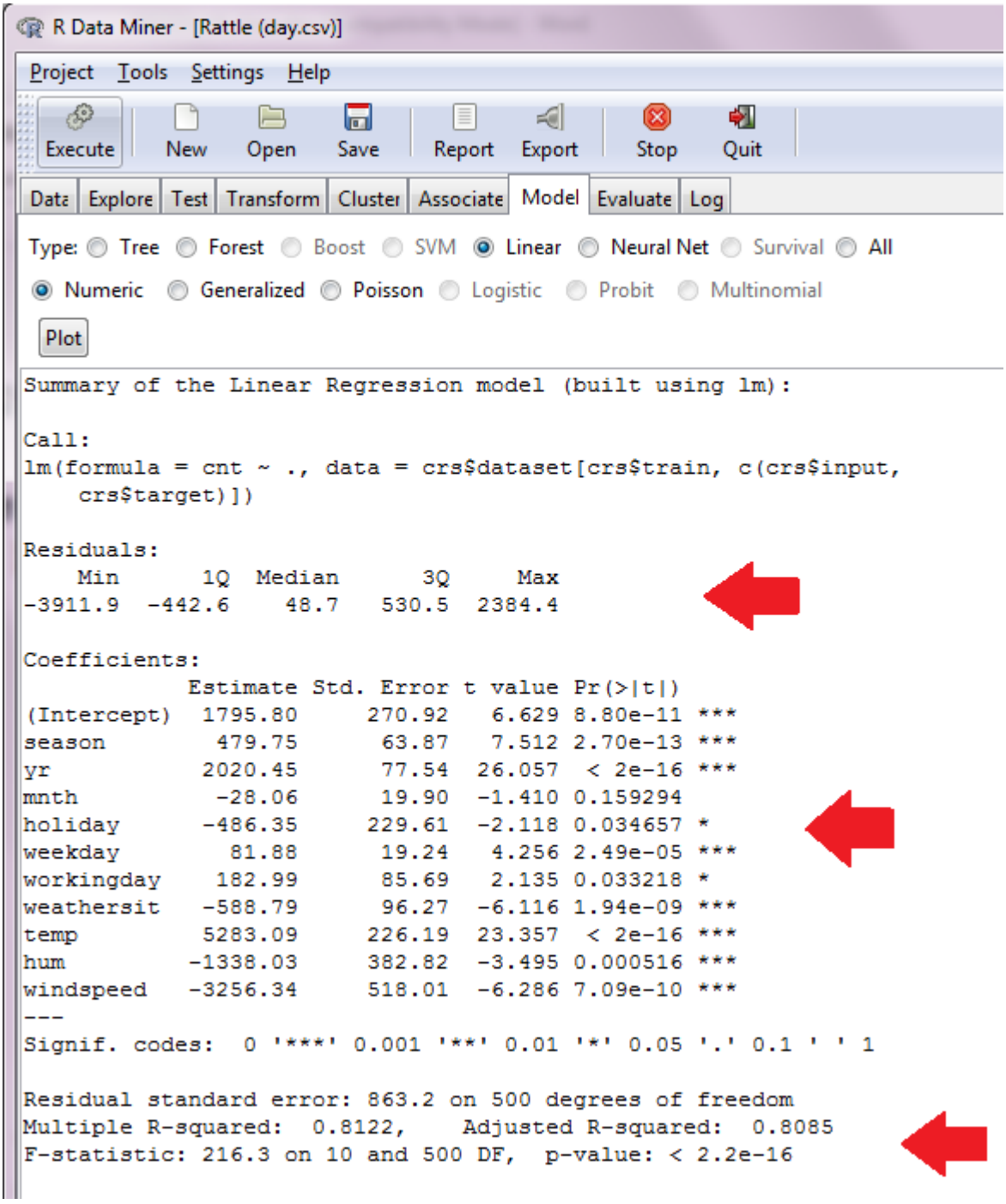


Analysis
Target 4.000
Avg Users per Day
2.5k
Users
695.3k
Year
Month
Day of the Week
Temperature
Humidity
Windspeed
Working day and Weekday
Colored by Number of Users
Year - Month

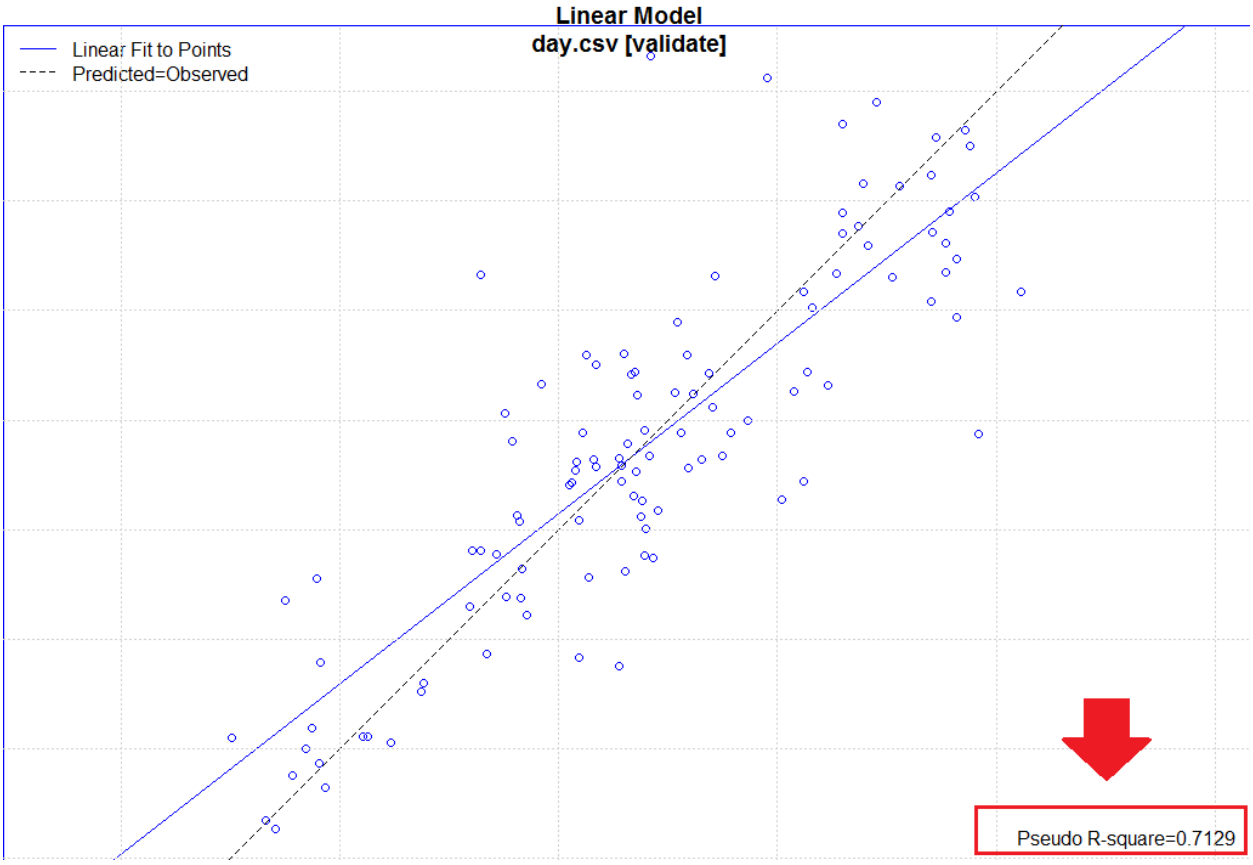

R Data Miner - [Rattle (day.csv)]
Project Tools Settings Help
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Summary Distributions Correlation Principal Components Interactive
Ordered Explore Missing Hierarchical Method: Pearson

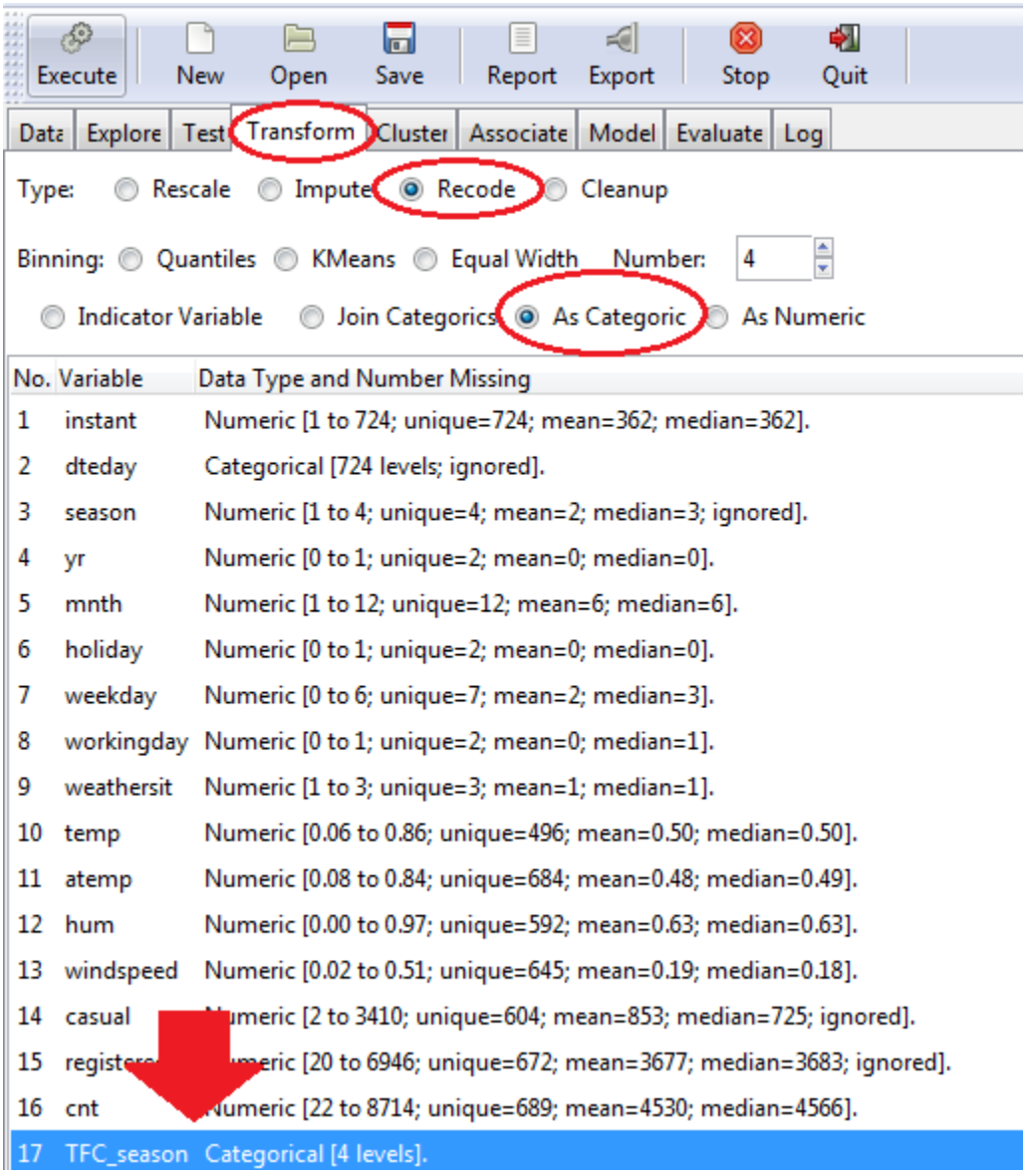
```
Correlation summary using the 'Pearson' covariance.

Note that only correlations between numeric variables are reported.

               windspeed      holiday   workingday       weekday   weathersit
windspeed    1.000000000 -0.002138282 -0.001560627  0.0072097243  0.019696893
holiday     -0.002138282  1.000000000 -0.265080924 -0.1167294529 -0.035618710
workingday  -0.001560627 -0.265080924  1.000000000  0.0409517959  0.036544724
weekday      0.007209724 -0.116729453  0.040951796  1.0000000000  0.018991914
weathersit   0.019696893 -0.035618710  0.036544724  0.0189919137  1.000000000
yr          -0.026605114  0.001759249  0.022853155 -0.0007176284 -0.031865947
hum         -0.249789799 -0.019085193 -0.022536858 -0.0421432549  0.635618502
temp        -0.129407940 -0.042122706  0.059877976  0.0277297718 -0.098747332
atemp       -0.157154074 -0.047527991  0.057232999  0.0236027677 -0.094983995
cnt         -0.232535932 -0.083692129  0.095895392  0.1040646910 -0.288818791
mnth        -0.205734775  0.019274288 -0.010169220  0.0073114706  0.041825772
season      -0.206270773 -0.024476785 -0.001246263  0.0030517333  0.007360641
                      yr          hum        temp       atemp         cnt
windspeed   -0.0266051144 -0.24978980 -0.12940794 -0.15715407 -0.23253593
holiday      0.0017592489 -0.01908519 -0.04212271 -0.04752799 -0.08369213
workingday   0.0228531545 -0.02253686  0.05987798  0.05723300  0.09589539
weekday     -0.0007176284 -0.04214325  0.02772977  0.02360277  0.10406469
weathersit  -0.0318659473  0.63561850 -0.09874733 -0.09498399 -0.28881879
yr           1.0000000000 -0.11515648  0.05666152  0.05191096  0.57115393
hum         -0.1151564848  1.00000000  0.11056228  0.12845472 -0.13944589
temp         0.0566615209  0.11056228  1.00000000  0.98977296  0.63082829
atemp        0.0519109640  0.12845472  0.98977296  1.00000000  0.63077395
cnt          0.5711539293 -0.13944589  0.63082829  0.63077395  1.00000000
mnth         0.0292674143  0.21963639  0.23059473  0.23885273  0.30162668
season       0.0410655502  0.17984302  0.33149921  0.34015894  0.42383148
                     mnth      season
windspeed   -0.205734775 -0.206270773
holiday      0.019274288 -0.024476785
workingday  -0.010169220 -0.001246263
weekday      0.007311471  0.003051733
weathersit   0.041825772  0.007360641
yr           0.029267414  0.041065550
hum          0.219636389  0.179843016
temp         0.230594733  0.331499211
atemp        0.238852732  0.340158943
cnt          0.301626677  0.423831483
mnth         1.000000000  0.829079430
season       0.829079430  1.000000000

Rattle timestamp: 2015-06-25 02:53:38
```

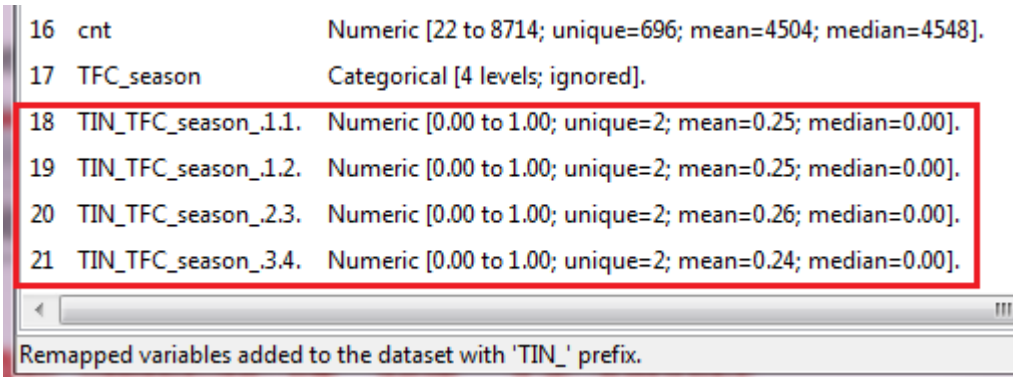

Correlation day.csv using Pearson
windspeed
holiday
workingday
weekday
weathersit
yr
hum
temp
atemp
cnt
mnth
season
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1


R Data Miner - [Rattle (day.csv)]
Project Tools Settings Help
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Tree Forest Boost SVM Linear Neural Net Survival All
Numeric Generalized Poisson Logistic Probit Multinomial
Plot
Summary of the Linear Regression model (built using lm):
Call:
lm(formula = cnt ~ ., data = crs$dataset[crs$train, c(crs$input,
crs$target)])
Residuals:
Min 1Q Median 3Q Max
-3911.9 -442.6 48.7 530.5 2384.4
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1795.80 270.92 6.629 8.80e-11 ***
season 479.75 63.87 7.512 2.70e-13 ***
yr 2020.45 77.54 26.057 < 2e-16 ***
mnth -28.06 19.90 -1.410 0.159294
holiday -486.35 229.61 -2.118 0.034657 *
weekday 81.88 19.24 4.256 2.49e-05 ***
workingday 182.99 85.69 2.135 0.033218 *
weathersit -588.79 96.27 -6.116 1.94e-09 ***
temp 5283.09 226.19 23.357 < 2e-16 ***
hum -1338.03 382.82 -3.495 0.000516 ***
windspeed -3256.34 518.01 -6.286 7.09e-10 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 863.2 on 500 degrees of freedom
Multiple R-squared: 0.8122, Adjusted R-squared: 0.8085
F-statistic: 216.3 on 10 and 500 DF, p-value: < 2.2e-16


Linear Model
day.csv [validate]
Linear Fit to Points
Predicted=Observed
Pseudo R-square=0.7129

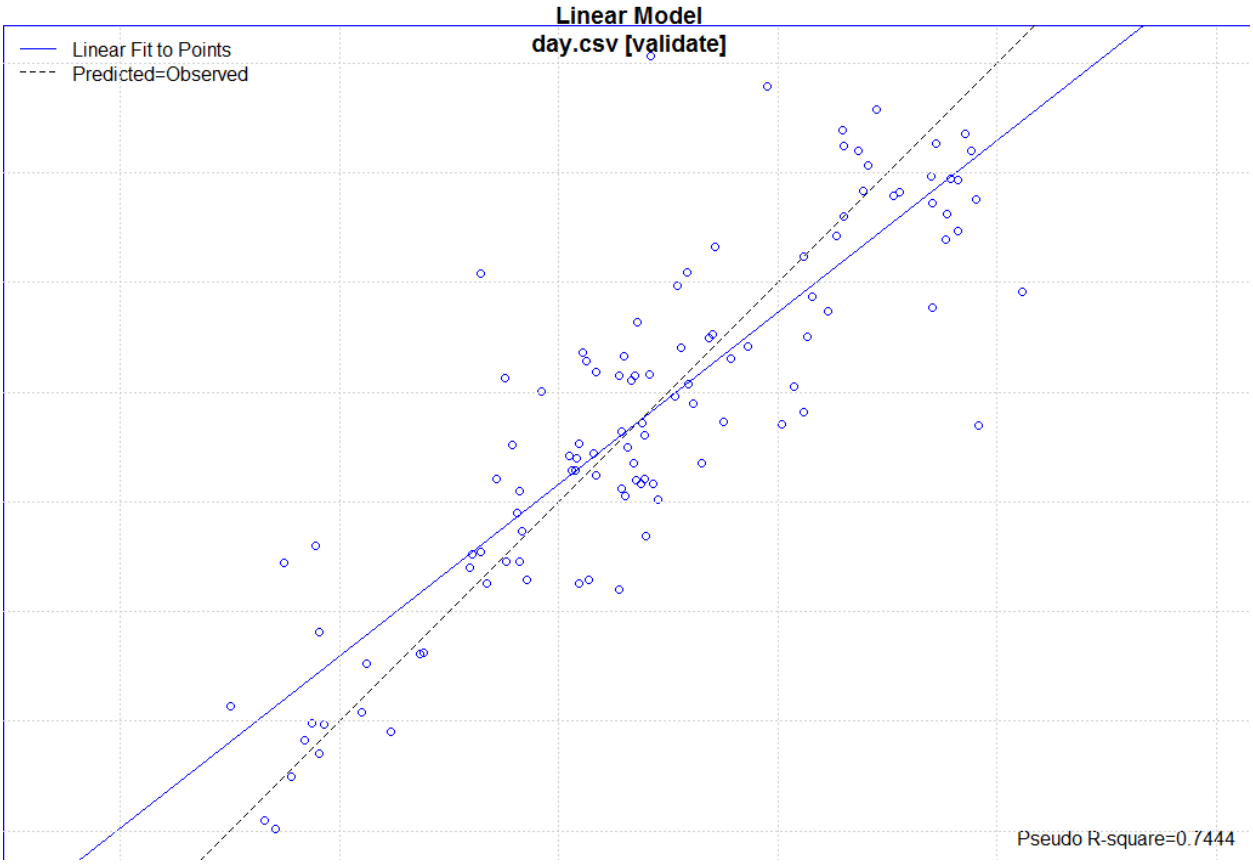

Execute
New
Open
Save
Report
Export
Stop
Quit
Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Rescale Impute Recode Cleanup
Binning: Quantiles KMeans Equal Width Number: 4
Indicator Variable Join Categorics As Categoric As Numeric
No. Variable Data Type and Number Missing
1 instant Numeric [1 to 724; unique=724; mean=362; median=362].
2 dteday Categorical [724 levels; ignored].
3 season Numeric [1 to 4; unique=4; mean=2; median=3; ignored].
4 yr Numeric [0 to 1; unique=2; mean=0; median=0].
5 mnth Numeric [1 to 12; unique=12; mean=6; median=6].
6 holiday Numeric [0 to 1; unique=2; mean=0; median=0].
7 weekday Numeric [0 to 6; unique=7; mean=2; median=3].
8 workingday Numeric [0 to 1; unique=2; mean=0; median=1].
9 weathersit Numeric [1 to 3; unique=3; mean=1; median=1].
10 temp Numeric [0.06 to 0.86; unique=496; mean=0.50; median=0.50].
11 atemp Numeric [0.08 to 0.84; unique=684; mean=0.48; median=0.49].
12 hum Numeric [0.00 to 0.97; unique=592; mean=0.63; median=0.63].
13 windspeed Numeric [0.02 to 0.51; unique=645; mean=0.19; median=0.18].
14 casual
15
16 cnt
17 TFC_season Categorical [4 levels].

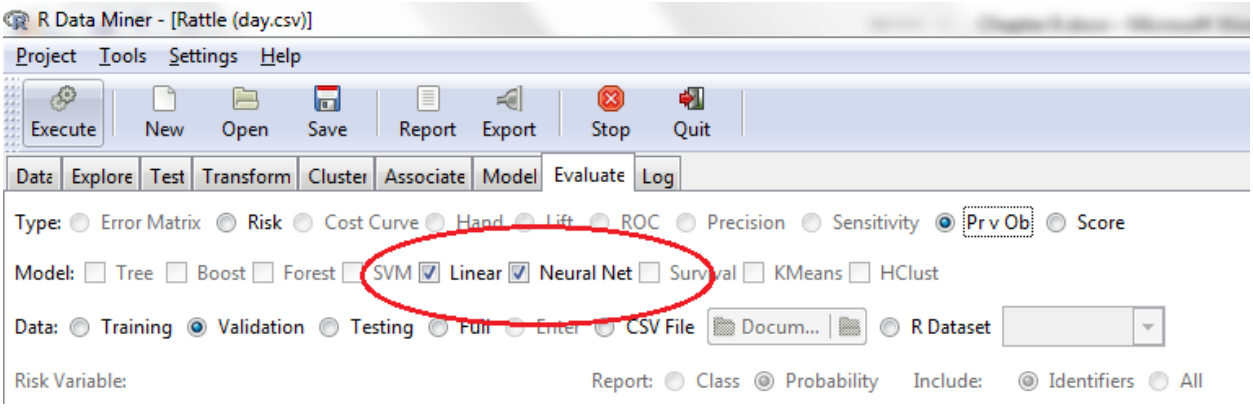


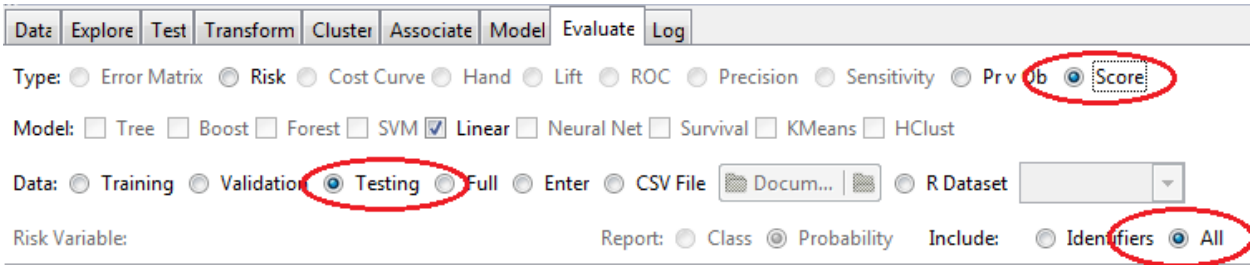
16	cnt	Numeric [22 to 8714; unique=696; mean=4504; median=4548].
17	TFC_season	Categorical [4 levels; ignored].
18	TIN_TFC_season_.1.1.	Numeric [0.00 to 1.00; unique=2; mean=0.25; median=0.00].
19	TIN_TFC_season_.1.2.	Numeric [0.00 to 1.00; unique=2; mean=0.25; median=0.00].
20	TIN_TFC_season_.2.3.	Numeric [0.00 to 1.00; unique=2; mean=0.26; median=0.00].
21	TIN_TFC_season_.3.4.	Numeric [0.00 to 1.00; unique=2; mean=0.24; median=0.00].

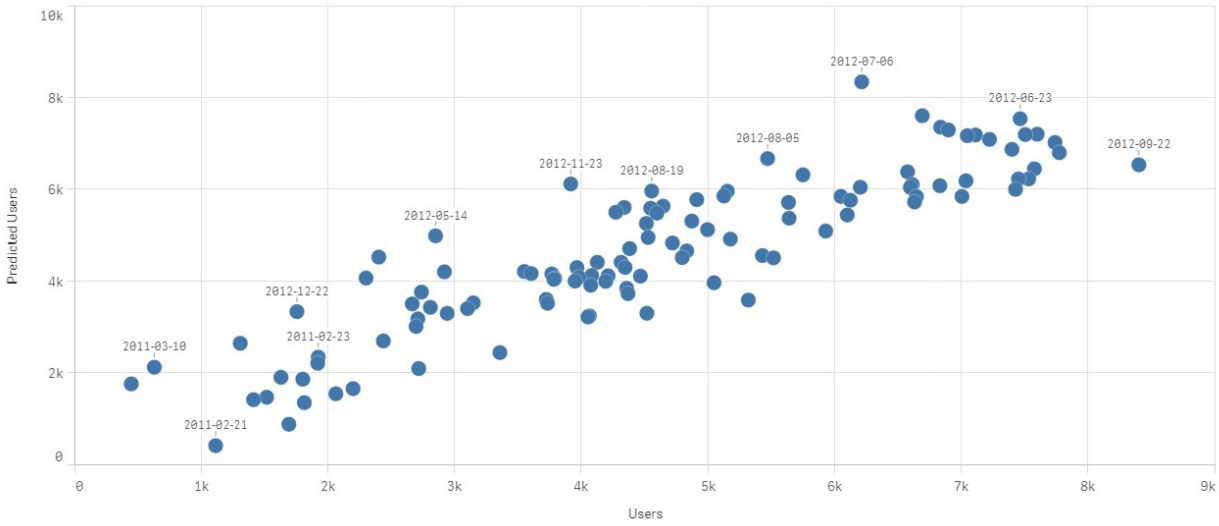
Remapped variables added to the dataset with 'TIN_' prefix.

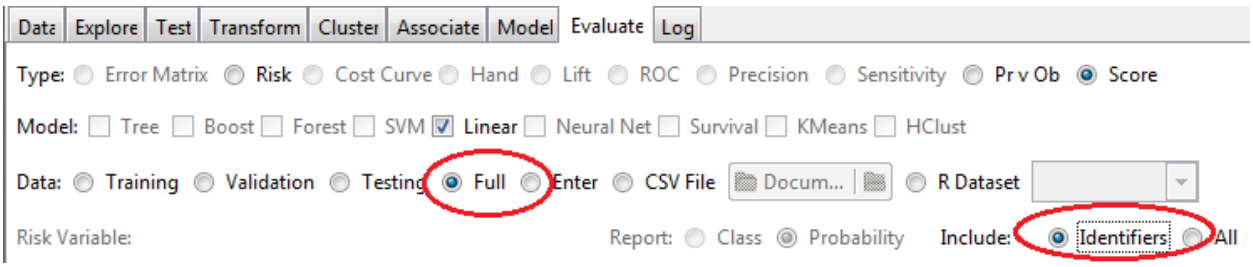
17	TFC_season	Categoric	○	○	○	○	◉	○	Unique: 4
18	TIN_TFC_season_.1.1.	Numeric	◉	○	○	○	○	○	Unique: 2
19	TIN_TFC_season_.1.2.	Numeric	◉	○	○	○	○	○	Unique: 2
20	TIN_TFC_season_.2.3.	Numeric	◉	○	○	○	○	○	Unique: 2
21	TIN_TFC_season_.3.4.	Numeric	◉	○	○	○	○	○	Unique: 2

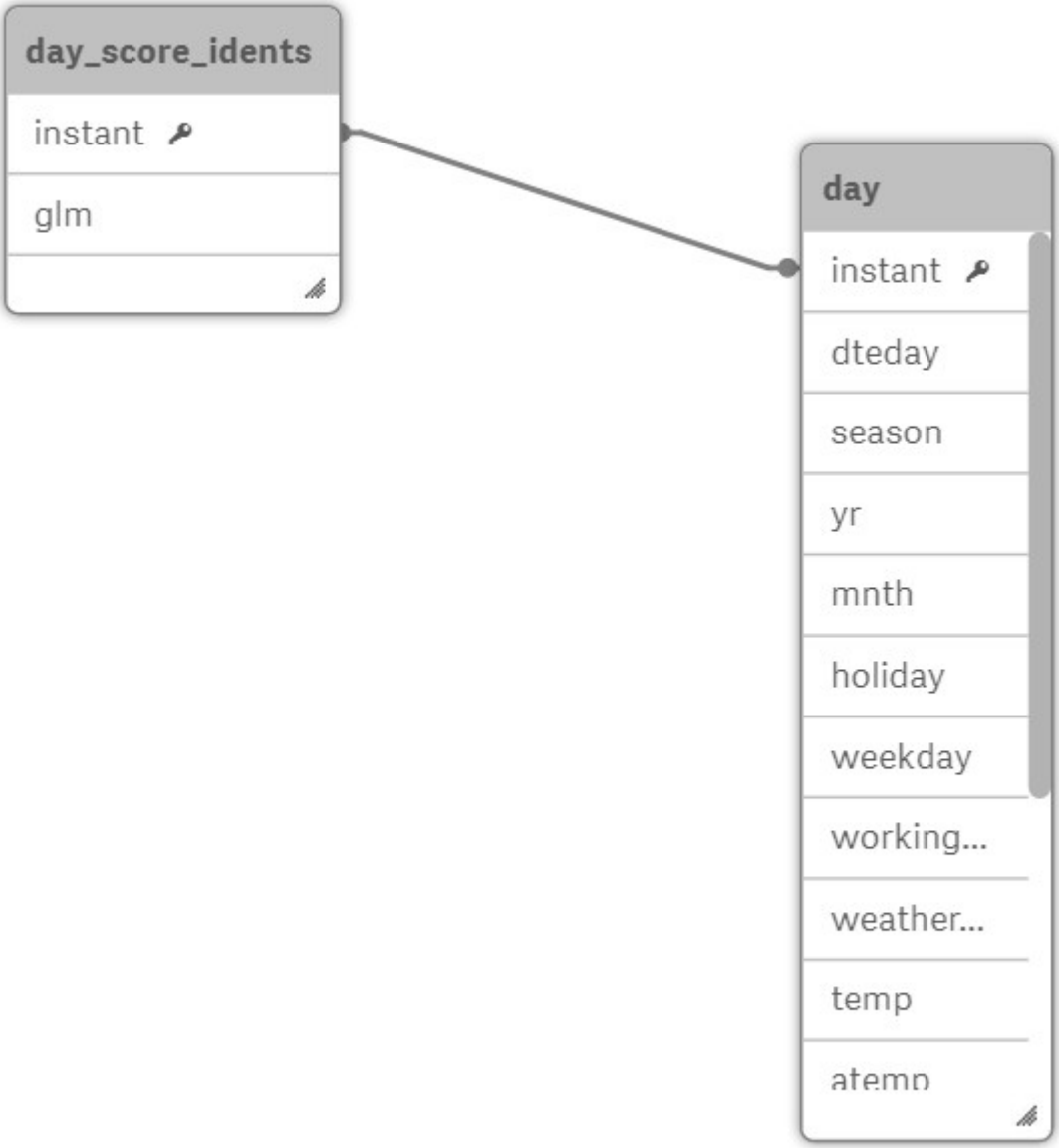

Linear Model
day.csv [validate]
Linear Fit to Points
Predicted=Observed
Pseudo R-square=0.7444


R Data Miner - [Rattle (day.csv)]
Project Tools Settings Help
Execute New Open Save Report Export Stop Quit
Data Explore Test Transform Cluster Associate Model Evaluate Log
Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score
Model: Tree Boost Forest SVM Linear Neural Net Survival KMeans HClust
Data: Training Validation Testing Full Enter CSV File Docum... R Dataset
Risk Variable:
Report: Class Probability Include: Identifiers All

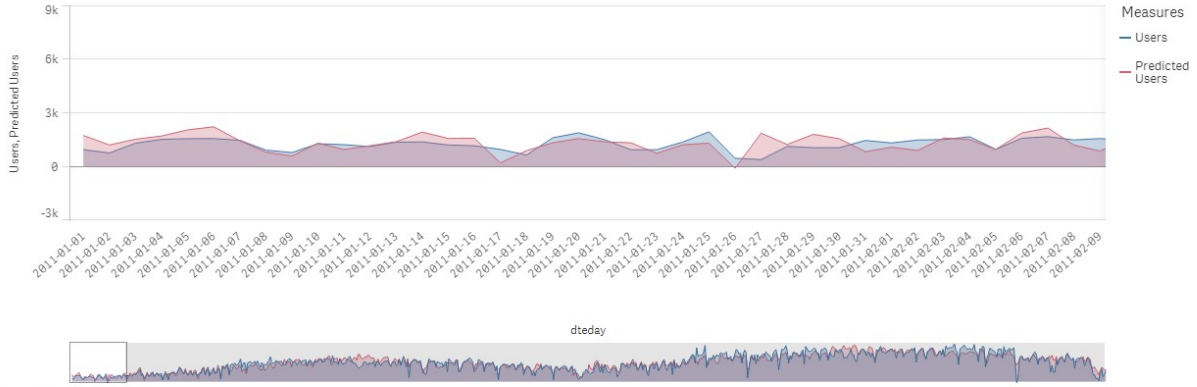

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score
Model: Tree Boost Forest SVM Linear Neural Net Survival KMeans HClust
Data: Training Validation Testing Full Enter CSV File Docum... R Dataset
Risk Variable:
Report: Class Probability
Include: Identifiers All


10k
8k
6k
4k
2k
0
Predicted Users
0
1k
2k
3k
4k
5k
6k
7k
8k
9k
Users
2012-07-06
2012-06-23
2012-08-05
2012-09-22
2012-11-23
2012-08-19
2012-05-14
2012-12-22
2011-02-23
2011-03-10
2011-02-21

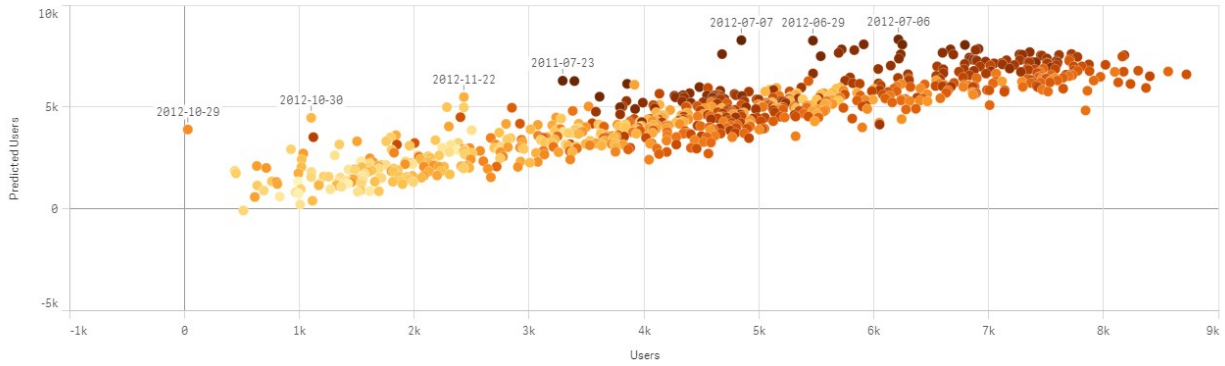

Data
Explore
Test
Transform
Cluster
Associate
Model
Evaluate
Log
Type: Error Matrix Risk Cost Curve Hand Lift ROC Precision Sensitivity Pr v Ob Score
Model: Tree Boost Forest SVM Linear Neural Net Survival KMeans HClust
Data: Training Validation Testing Full Enter CSV File Docum... R Dataset
Risk Variable:
Report: Class Probability
Include: Identifiers All


day_score_idents
instant
glm
day
instant
dteday
season
yr
mnth
holiday
weekday
working...
weather...
temp
atemp

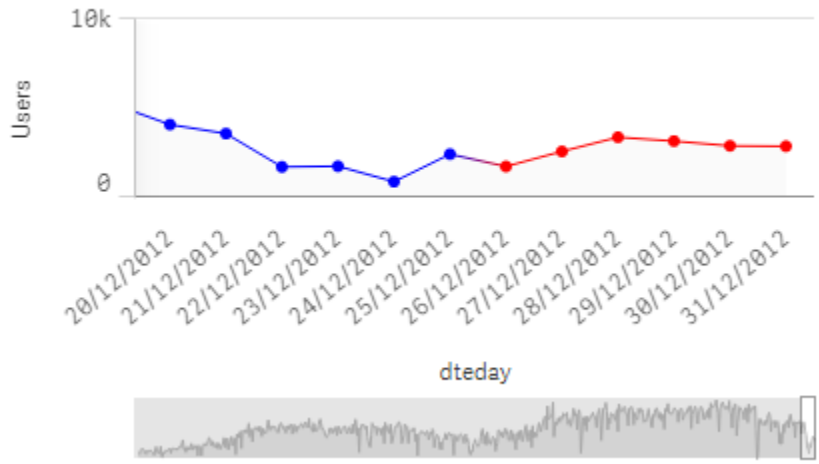
Evolution Predicted Vs Actual



Users Vs Predicted Users



Forecast


10k
Users
0
20/12/2012
21/12/2012
22/12/2012
23/12/2012
24/12/2012
25/12/2012
26/12/2012
27/12/2012
28/12/2012
29/12/2012
30/12/2012
31/12/2012
dteday