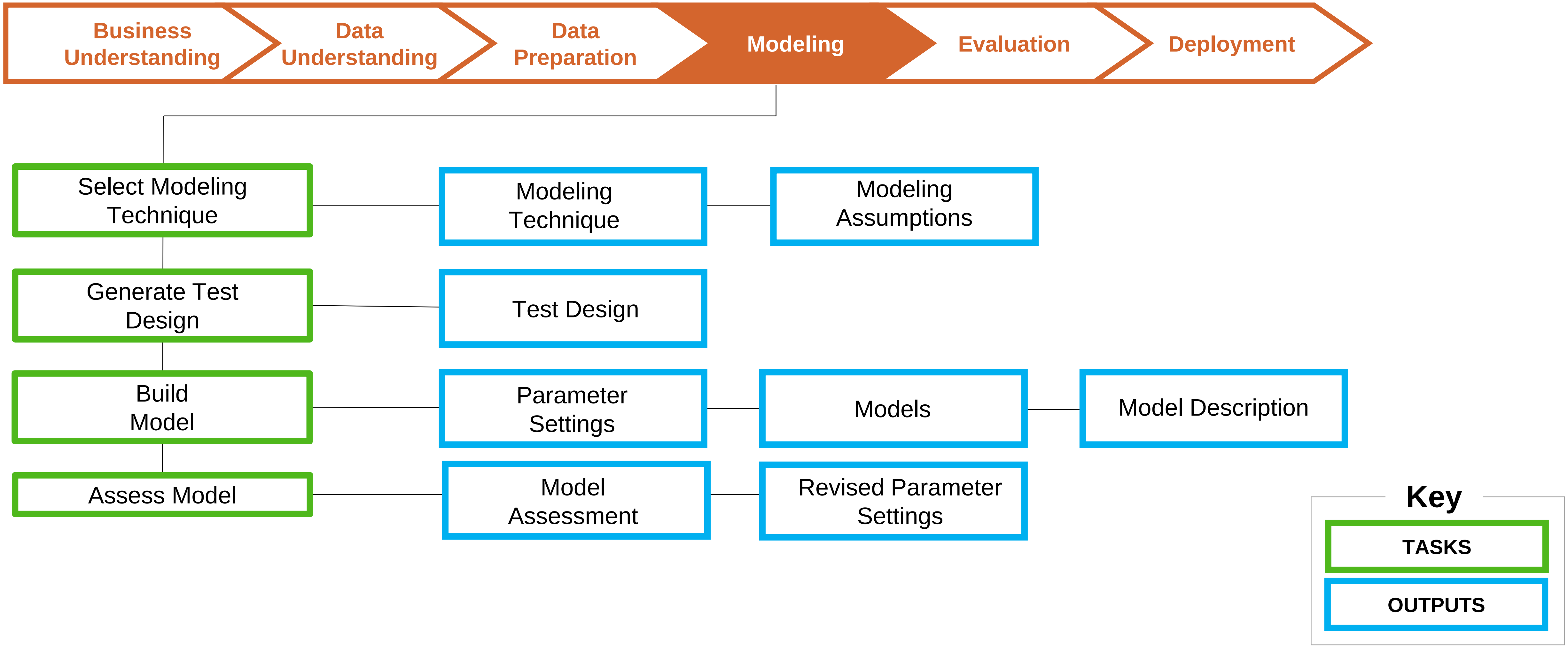
Week 3 Unit 1: **Modeling Phase – Overview**

open.sap.com

Modeling Phase – Overview

CRISP-DM – Phase 4: Modeling



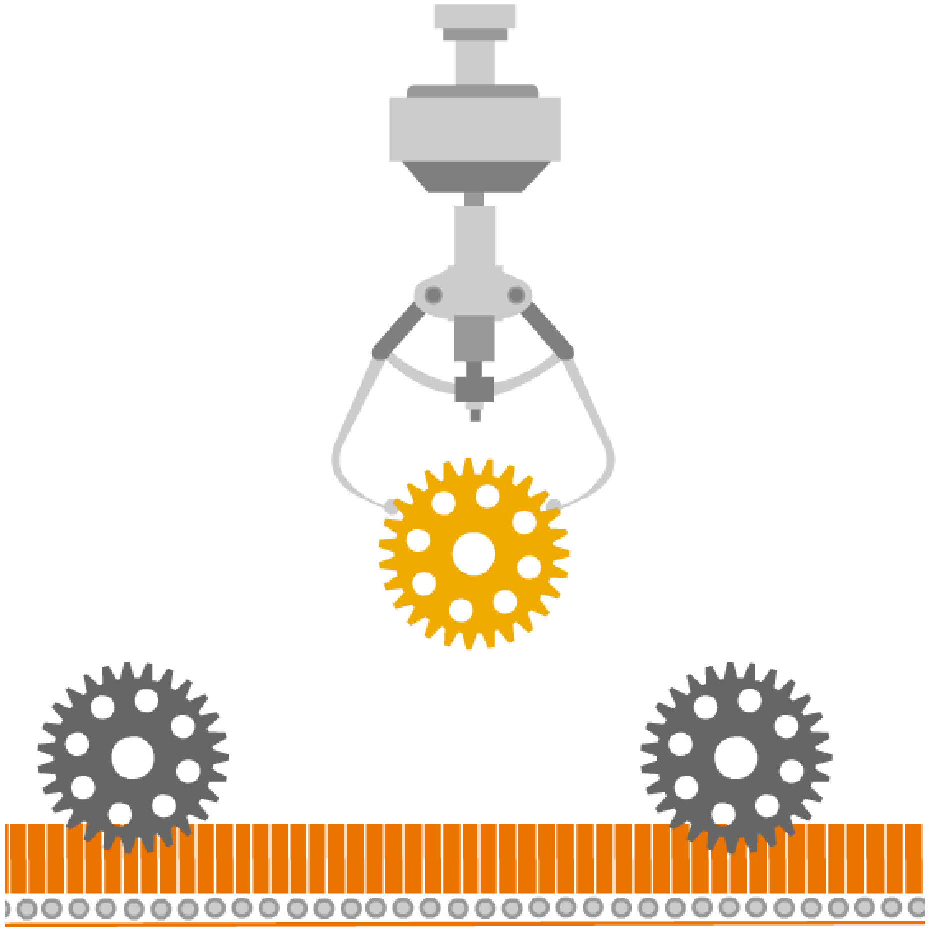
Modeling Phase – Overview

Phase 4.1: Select Modeling Technique

- **Task**
  – Select the actual modeling technique that is to be used.

- **Output - Modeling Technique**
  – Document the modeling technique that is to be used.

- **Output - Modeling Assumptions**
  – Record any such assumptions made.



Modeling Phase – Overview

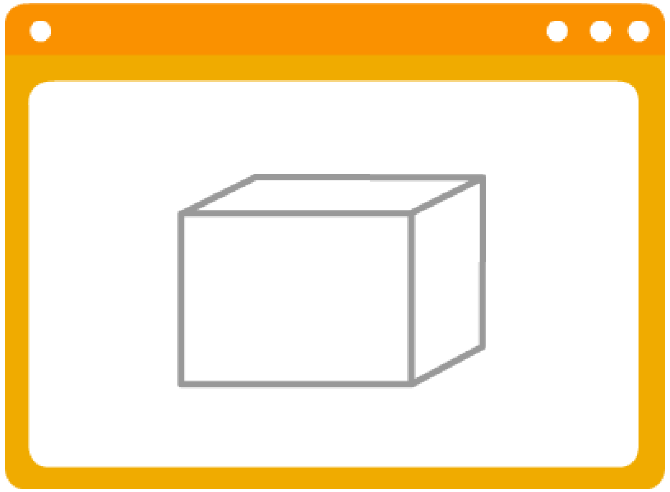
Phase 4.2: Generate Test Design

- **Task**
  - Before we actually build a model, we need to generate a procedure or mechanism to test the model's quality and validity.
- **Output - Test Design**
  - Describe the intended plan for training, testing, and evaluating the models.

Modeling Phase – Overview

Phase 4.3: Build Model

- **Task**
  - Run the modeling tool on the prepared dataset to create one or more models.
- **Output - Parameter Settings**
  - List the parameters and their chosen value, along with the rationale for the choice of parameter settings.
- **Output - Models**
  - These are the actual models produced by the modeling tool, not a report.
- **Output - Model Description**
  - Describe the resultant model.



Modeling Phase – Overview

Phase 4.4: Assess Model

- **Task**
  - Interpret the models according to domain knowledge, the data mining success criteria, and the desired test design.
- **Output - Model Assessment**
  - Summarize results, list qualities of generated models, and rank their quality in relation to each other.
- **Output - Revised Parameter Settings**
  - Revise parameter settings and tune them for the next run in the Build Model task.

Thank you

Contact information:

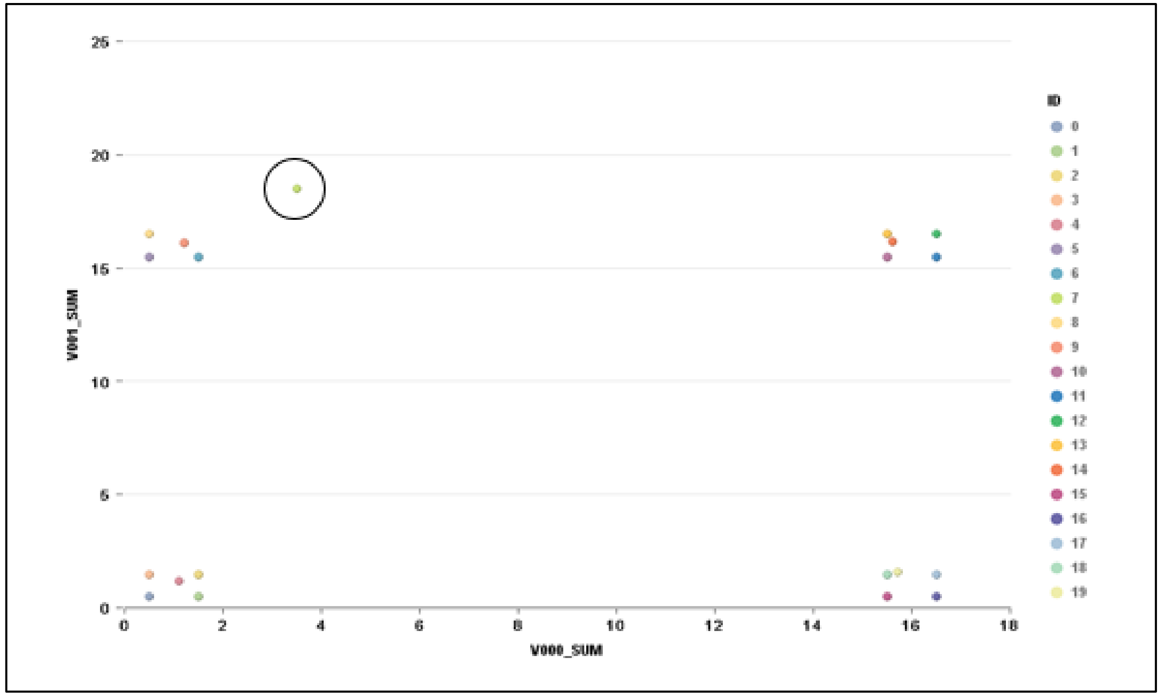
open@sap.com

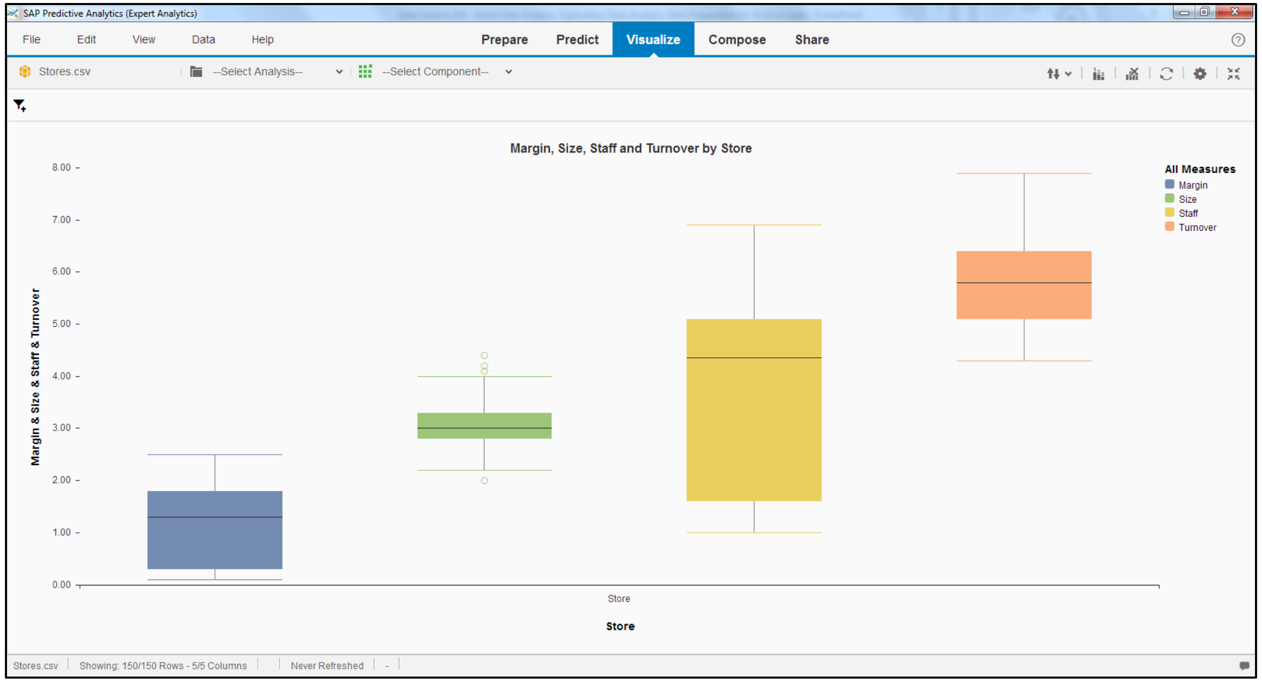
open.sap.com

Week 3 Unit 2: **Detecting Anomalies**

Detecting Anomalies

Outliers



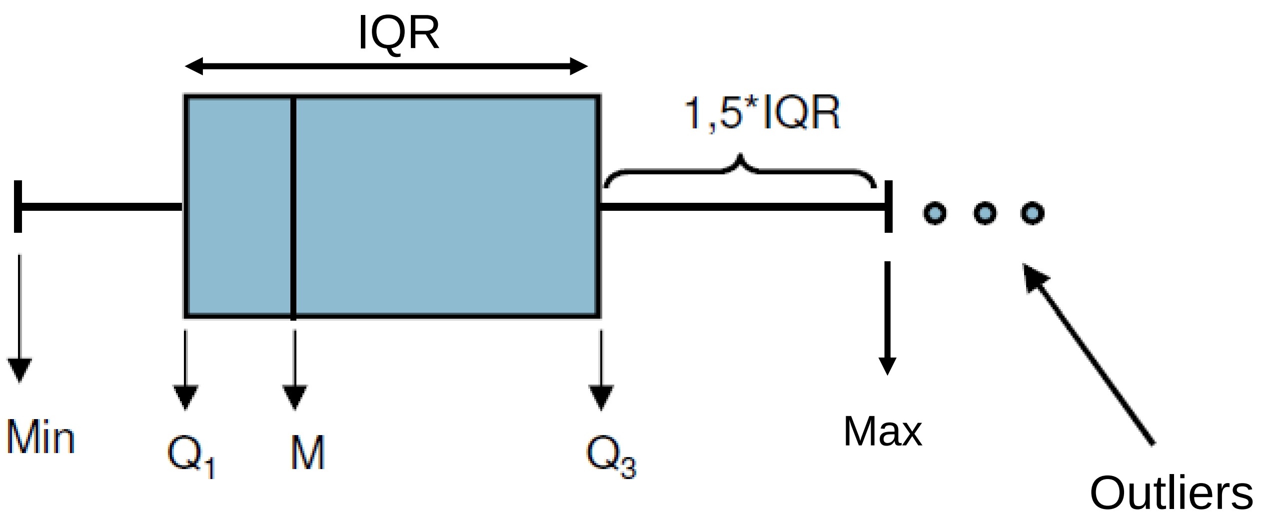


Outlier

Box Plot

Detecting Anomalies
Box plot

- A box plot is a visual representation of five numbers:
  - Median M (50%)
  - First Quartile Q1 (0-25% of data)
  - Third Quartile Q3 (75-100% of data)
  - Minimum
  - Maximum



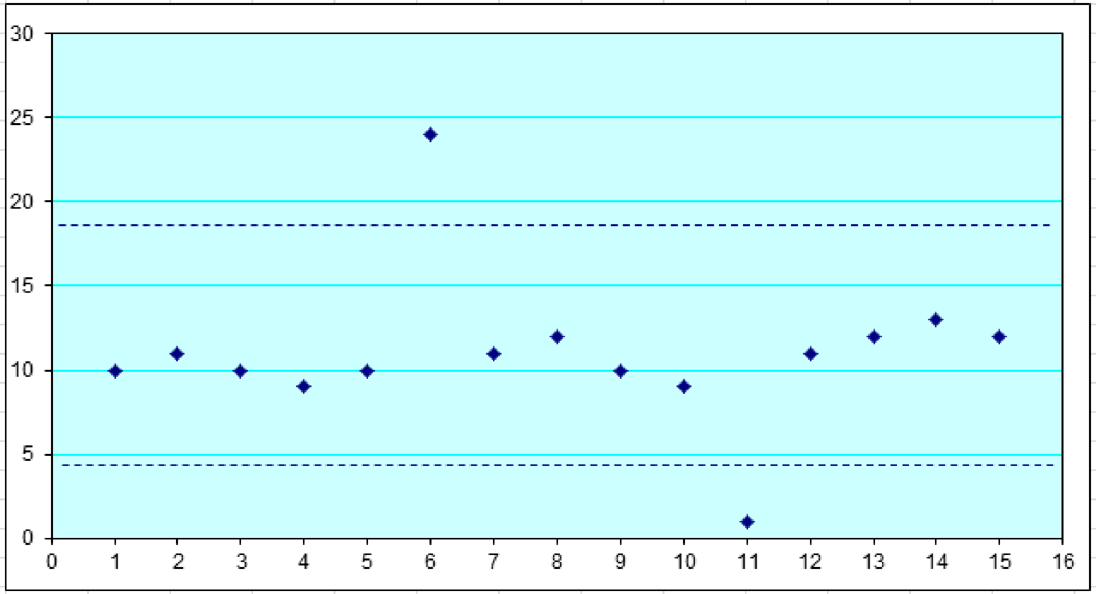
Detecting Anomalies

Outlier algorithms – Inter-quartile range (IQR) test: worked example

Outlier Test Example 1 - Inter Quartile Range Test

Value Name	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14	P15
Time series	10	11	10	9	10	24	11	12	10	9	1	11	12	13	12

Number of Values	15
Lower Quartile	10
Median	11
Upper Quartile	12
Inter Quartile Range	2
Inter Quartile Range Multiplier	3
Upper Limit	18
Lower Limit	4
Number of Values above Upper Limit	1
Number of Values below Lower Limit	1
Value1 above Upper Limit Value Name	P6
Value1 above Upper Limit Value Number	24
Distance from Median	13
Distance from Median %	118.18
Distance from Median Absolute	13
Distance from Median Absolute %	118.18
Value1 below Lower Limit Value Name	P11
Value1 below Lower Limit Value Number	1
Distance from Median Absolute	-10
Distance from Median %	-90.91
Distance from Median Absolute	10
Distance from Median %	90.91



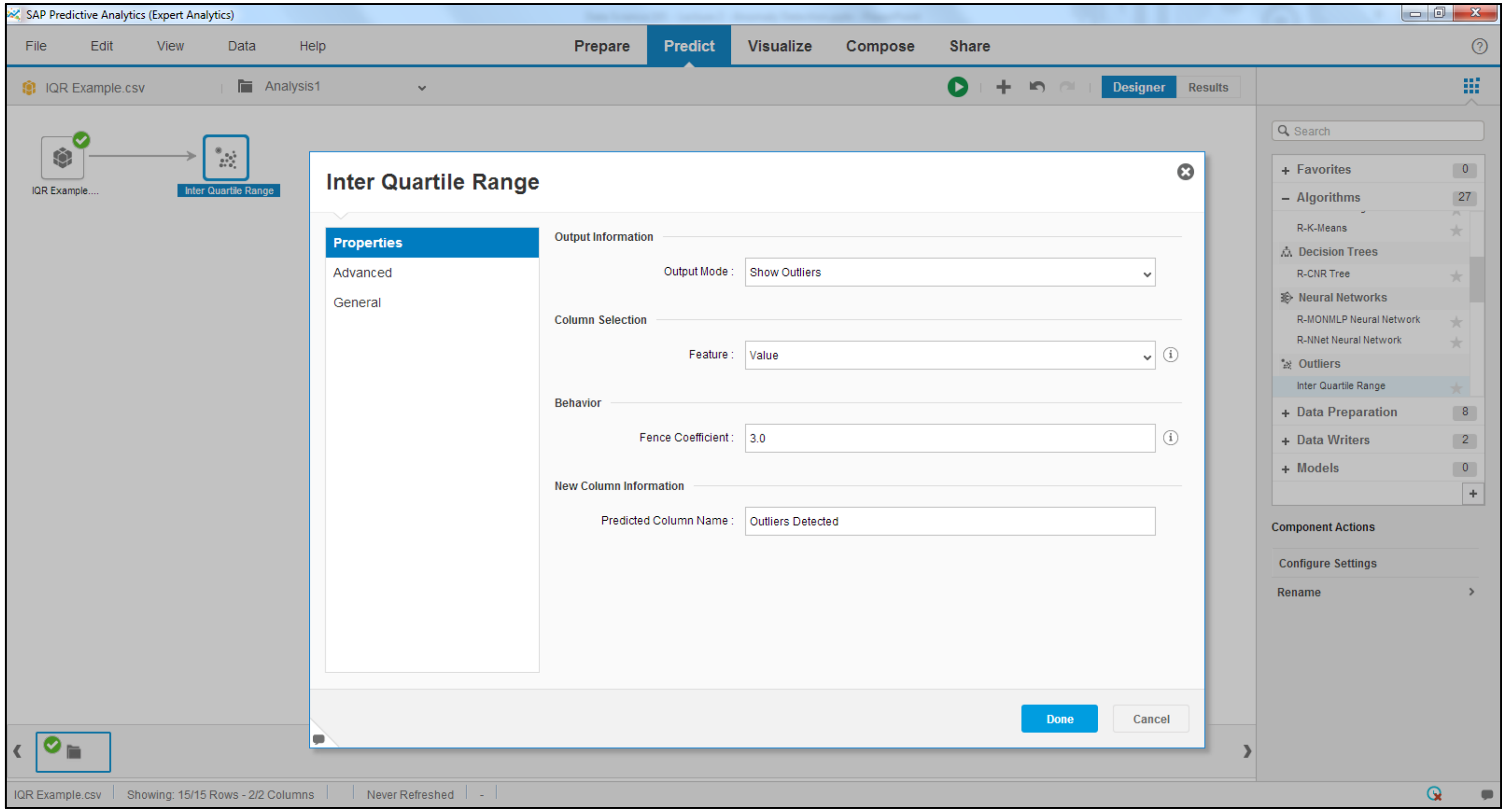
Detecting Anomalies

Outlier algorithms – Inter-quartile range (IQR) test: demonstration

IQR Example.csv

ABC Period	123 Value
P10	10
P11	11
P12	10
P13	9
P14	10
P15	24
P16	11
P17	12
P18	10
P19	9
P20	1
P21	11
P22	12
P23	13
P24	12

Input Data



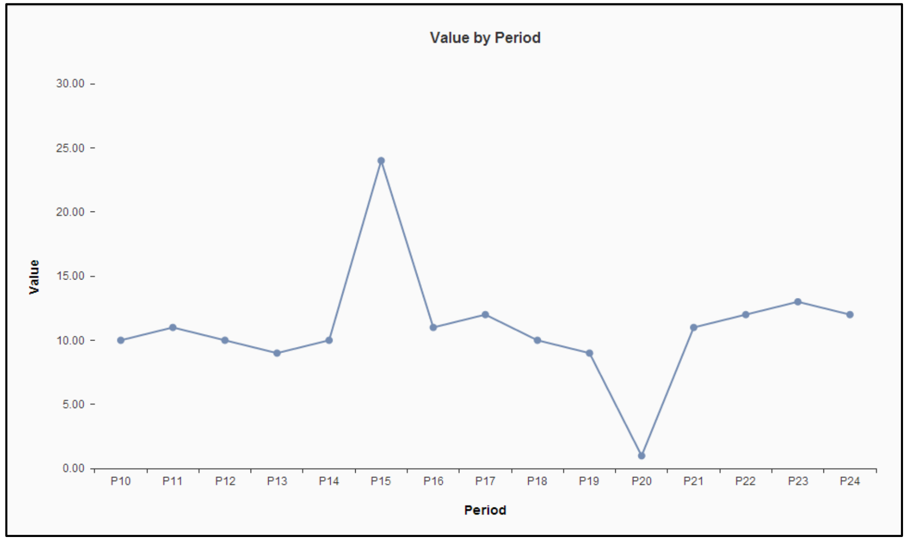
SAP Predictive Analytics

Detecting Anomalies

Outlier algorithms – Inter-quartile range (IQR) test: demonstration

ABC Period	123 Value	123 Outliers Detected
P10	10	0
P11	11	0
P12	10	0
P13	9	0
P14	10	0
P15	24	1
P16	11	0
P17	12	0
P18	10	0
P19	9	0
P20	1	1
P21	11	0
P22	12	0
P23	13	0
P24	12	0

Results with Outlier Flag



Graphical Plot of Data

IQR Example.csv | Analysis1

```
Information of the columns used in the algorithm
----------------------------------------------------
Independent Column
 Value : Integer

Inter Quartile Outlier Detection Summary
----------------------------------------------------
 First Quartile found at row: 4 with a value 10.0
 Third Quartile found at row: 11 with a value 12.0

 For a fence coef of 3.0
               Lower Fence value :4.0
               Upper Fence value :18.0

 Total Number of Outliers detected : 2


Data Set Summary
-----------------------------------

1. Mean                 :11.0
2. Standard Deviation   :4.535573676110727
3. Number values considered   :15
-----------------------------------
```

Statistical Summary for IQR Test

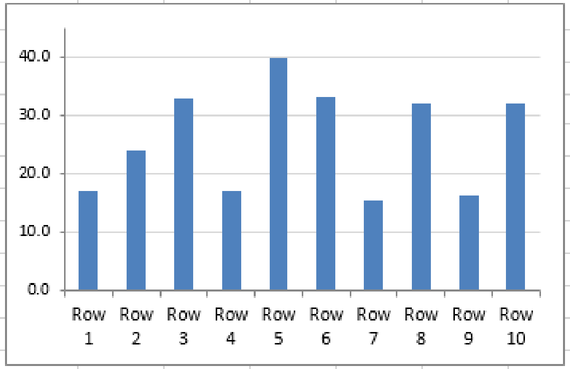
Detecting Anomalies

Nearest Neighbor outlier: worked example

K Nearest Neighbour Outlier Test

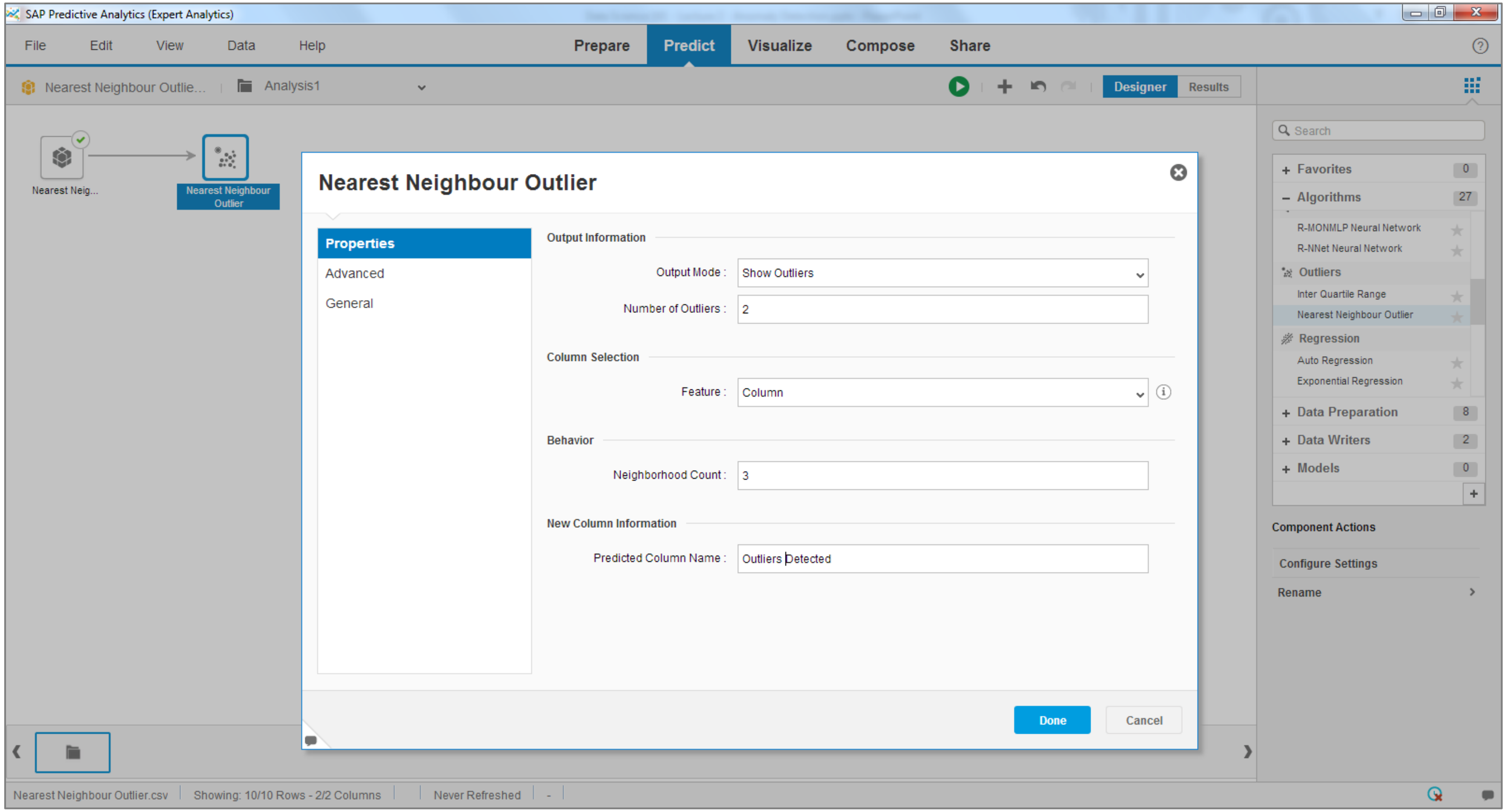
	Example
Number of neighbors to be used (K)	**K=3**
Number of outliers to be detected (N)	**N=2**

		Euclidean Distances														
Row	**Column**	Row 1	Row 2	Row 3	Row 4	Row 5	Row 6	Row 7	Row 8	Row 9	Row 10	Min	Next	Next	**Average**	
Row 1	16.9		7.1	16.0	0.2	22.9	16.1	1.4	15.0	0.7	15.2	0.2	0.7	1.4	0.77	
Row 2	24.0	7.1		8.9	6.9	15.8	9.0	8.5	7.9	7.8	8.1	6.9	7.1	7.8	7.27	**Largest 1**
Row 3	32.9	16.0	8.9		15.8	6.9	0.1	17.4	1.0	16.7	0.8	0.1	0.8	1.0	0.63	
Row 4	17.1	0.2	6.9	15.8		22.7	15.9	1.6	14.8	0.9	15.0	0.2	0.9	1.6	0.90	
Row 5	39.8	22.9	15.8	6.9	22.7		6.8	24.3	7.9	23.6	7.7	6.8	6.9	7.7	7.13	**Largest 2**
Row 6	33.0	16.1	9.0	0.1	17.4	6.8		17.5	1.1	16.8	0.9	0.1	0.9	1.1	0.70	
Row 7	15.5	1.4	8.5	17.4	1.6	24.3	17.5		16.4	0.7	16.6	0.7	1.4	1.6	1.23	
Row 8	31.9	15.0	7.9	1.0	14.8	7.9	1.1	16.4		15.7	0.2	0.2	1.0	1.1	0.77	
Row 9	16.2	0.7	7.8	16.7	0.9	23.6	16.8	0.7	15.7		15.9	0.7	0.7	0.9	0.77	
Row 10	32.1	15.2	8.1	0.8	15.0	7.7	0.9	16.6	0.2	15.9		0.2	0.8	0.9	0.63	



Detecting Anomalies

Nearest Neighbor outlier: demonstration



Detecting Anomalies

Nearest Neighbor outlier: demonstration

ABC Row	123 Column	123 Outliers ..
Row 1	16.90	0
Row 2	24.00	1
Row 3	32.90	0
Row 4	17.10	0
Row 5	39.80	1
Row 6	33.00	0
Row 7	15.50	0
Row 8	31.90	0
Row 9	16.20	0
Row 10	32.10	0

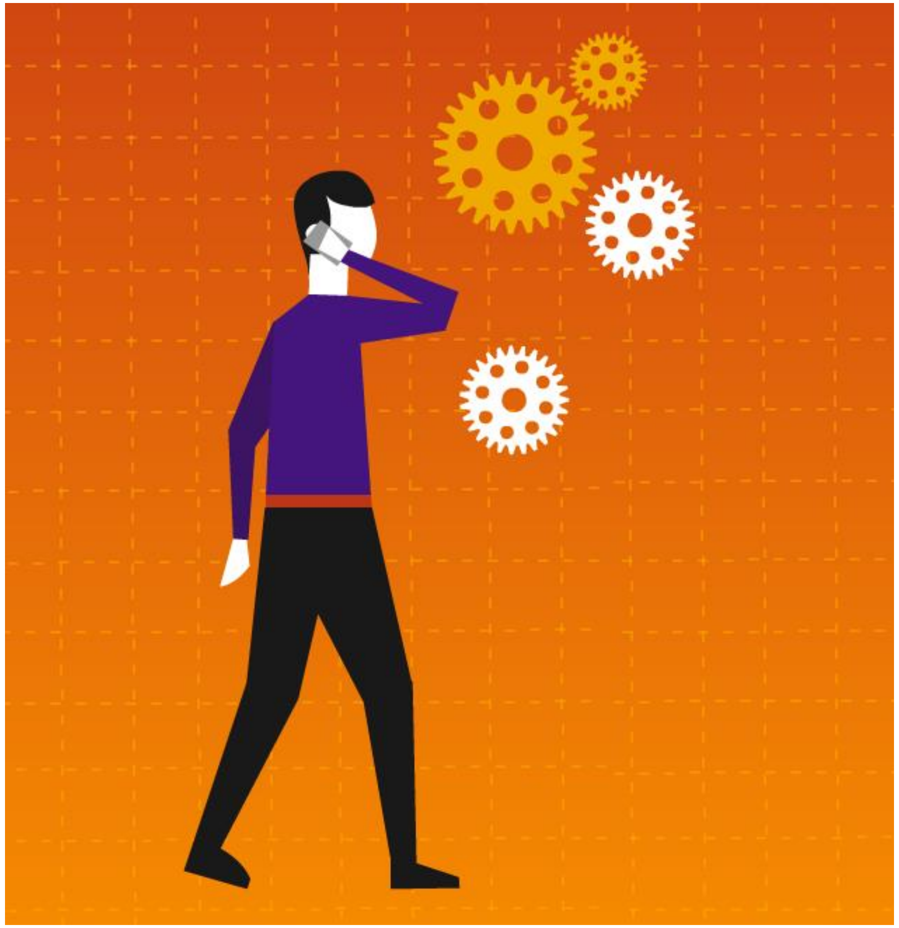
Results Table with Outlier Flag

Detecting Anomalies

Other methods for anomaly detection

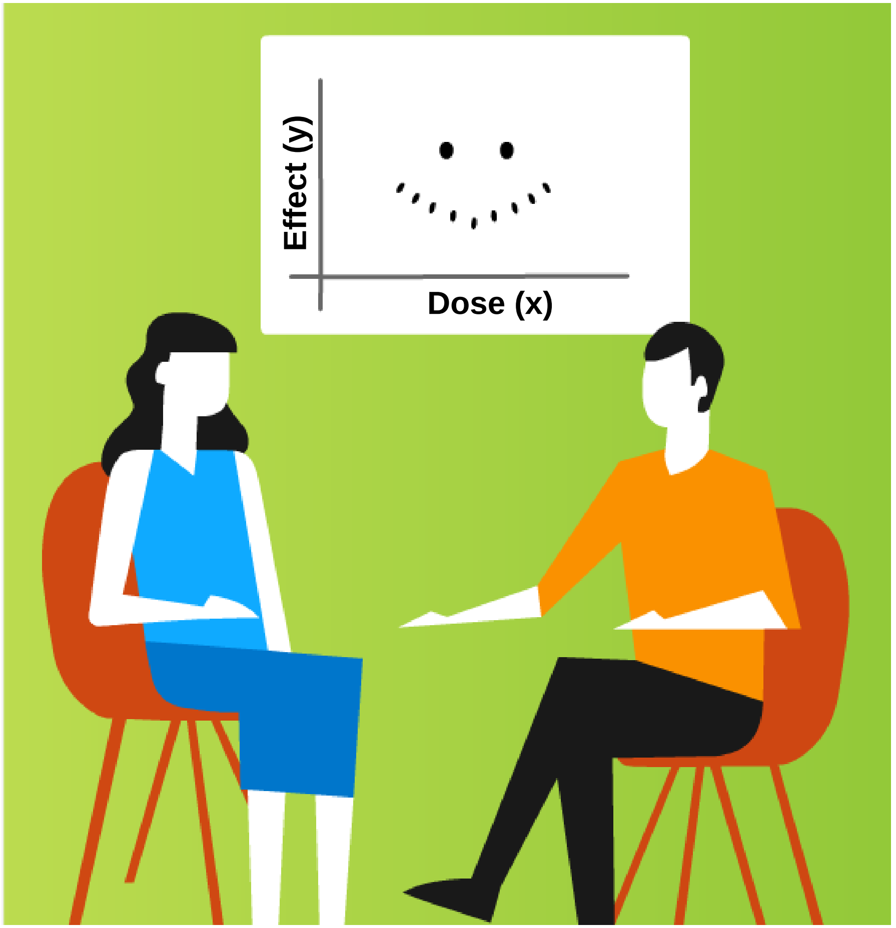
- There are a wide range of anomaly detection algorithms available apart from those previously described:
  - Cluster Modeling
  - Association Analysis – identifies rare occurrences
  - Principal Component Analysis
  - Distance-Based Failure Analysis
  - Link Analysis
  - …

- Anomalies can arise from what is ‘unusual’; but also what is ‘unexpected’
  - Build a model on observed data, score new data, examine the major variances of actual vs. predicted



Detecting Anomalies
Summary

- Outlier analysis is a key step in predictive analysis, as outliers can significantly affect a model.
- We can perform a visual analysis.
- We can use various algorithms.
- Outliers need to be investigated and not simply removed from the analysis.



Thank you

Contact information:

open@sap.com

Week 3 Unit 3: **Association Analysis**

openSAP
open.sap.com

Association Analysis

Introduction

Association Analysis

Demonstration

Record ID	Transaction ID	Product Name
1	160358	Product 1
2	160358	Product 2
3	160953	Product 2
4	160953	Product 3
5	160953	Product 7
6	162026	Product 1
7	162026	Product 3
8	220402	Product 4
9	236726	Product 5
10	271185	Product 6
11	271185	Product 2
12	271185	Product 4
13	323951	Product 5
14	323952	Product 2
15	377343	Product 1
16	584229	Product 2
17	584229	Product 3
18	584229	Product 4
19	608022	Product 1
20	681110	Product 4
21	681110	Product 5
22	681110	Product 6
23	710991	Product 5
24	710991	Product 6
25	710991	Product 7
26	716017	Product 6
27	740287	Product 2
28	805591	Product 1
29	905431	Product 5
30	905431	Product 6

Till-Roll Data

In our worked example, the most common ‘one-product’ occurrences are:

Product 1	5 times in 17 baskets =	5/17 =	0.294
Product 2	6 times in 17 baskets =	6/17 =	0.353
Product 3	3 times in 17 baskets =	3/17 =	0.176
Product 4	4 times in 17 baskets =	4/17 =	0.235
Product 5	5 times in 17 baskets =	5/17 =	0.294
Product 6	5 times in 17 baskets =	5/17 =	0.294
Product 7	2 times in 17 baskets =	2/17 =	0.118

Manual Analysis of Transactions

ABC Rules	123 Support
{} => {Product 7}	0.12
{} => {Product 1}	0.29
{} => {Product 3}	0.18
{} => {Product 4}	0.24
{} => {Product 5}	0.29
{} => {Product 6}	0.29
{} => {Product 2}	0.35

SAP Predictive Analytics Analysis

Association Analysis

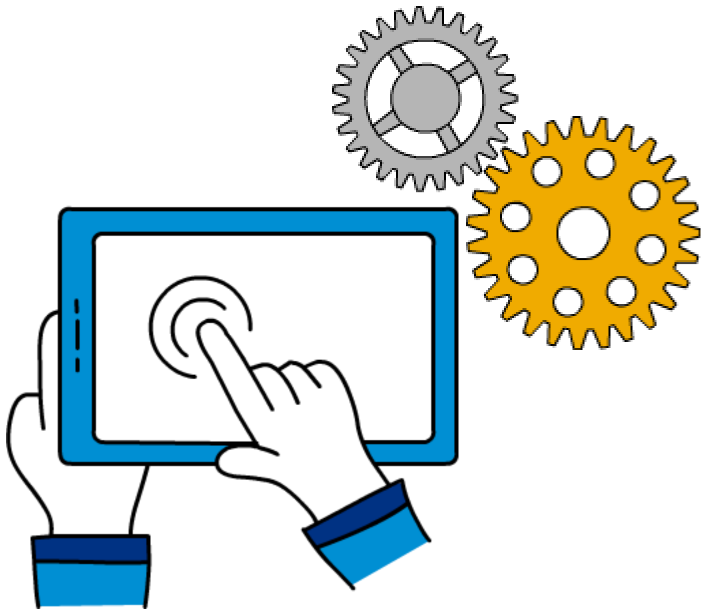
Rules

- Association analysis produces simple rules.
  For example:
  If {Product A}, {Product B} => {Product C}

- These are simple “If” “Then” type rules.

- This rule has two **antecedents** (product A and product B) and one **consequent** (product C).

- The **rule length** = 3, as there are 3 products.



Association Analysis

Support

Rules	No. of Baskets Supporting the Rule	Total Number of Baskets	Rule Support %
If 5 then 6	3	17	17.65%
If 2 then 3	2	17	11.76%
If 2 then 4	2	17	11.76%
If 4 then 6	2	17	11.76%
If 1 then 2	1	17	5.88%
If 1 then 3	1	17	5.88%
If 2 then 6	1	17	5.88%
If 2 then 7	1	17	5.88%
If 2 and 3 then 7	1	17	5.88%
If 6 and 2 then 4	1	17	5.88%
If 2 and 3 then 4	1	17	5.88%
If 4 and 5 then 6	1	17	5.88%
If 5 and 6 then 7	1	17	5.88%

Association Analysis
Confidence

Rule	No. of Baskets Supporting the Rule	Total Number of Baskets with Pa	Confidence
If Pa then Pb			
If Product 5 then Product 6	3	5	60%
If Product 2 then Product 3	2	6	33%
If Product 2 then Product 4	2	6	33%
If Product 4 then Product 6	2	4	50%
If Product 1 then Product 2	1	5	20%
If Product 1 then Product 3	1	5	20%
If Product 2 then Product 6	1	6	16%
If Product 2 then Product 7	1	6	16%

Rule	No. of Baskets Supporting the Rule	Total Number of Baskets with Pb	Confidence
If Pb then Pa			
If Product 6 then Product 5	3	5	60%
If Product 3 then Product 2	2	3	66%
If Product 4 then Product 2	2	4	50%
If Product 6 then Product 4	2	5	40%
If Product 2 then Product 1	1	6	16%
If Product 3 then Product 1	1	3	33%
If Product 6 then Product 2	1	5	20%
If Product 7 then Product 2	1	2	50%

Association Analysis

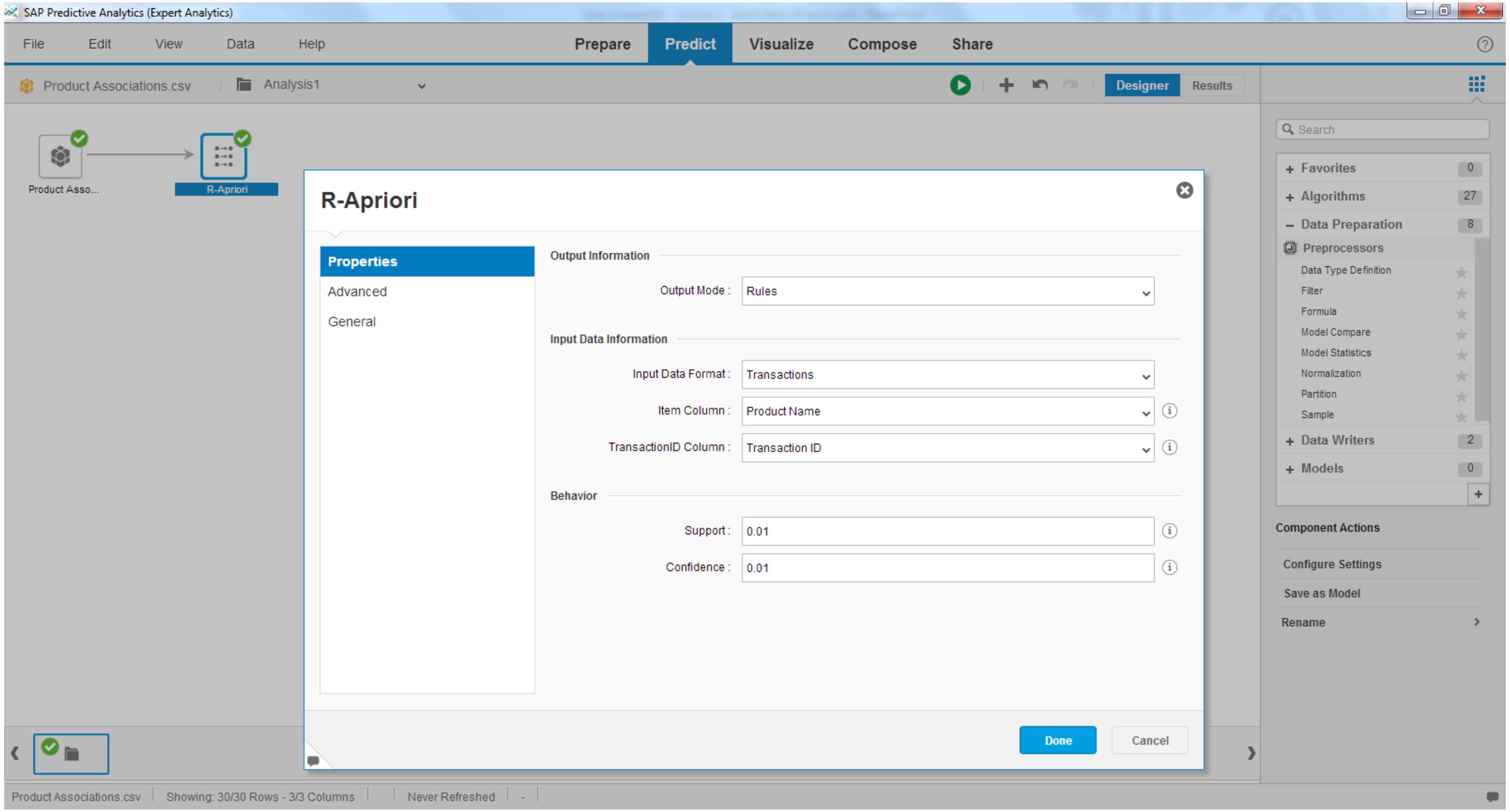
Lift

Rule	Confidence	Support Pb (Result)	Improvement / Rule Lift
If Product 5 then Product 6	60%	29.41% (5/17)	2.04
If Product 2 then Product 3	33%	17.65% (3/17)	1.87
If Product 2 then Product 4	33%	23.53% (4/17)	1.40
If Product 4 then Product 6	50%	29.41% (5/17)	1.70
If Product 1 then Product 2	20%	35.29% (6/17)	0.57
If Product 1 then Product 3	20%	17.65% (3/17)	1.13
If Product 2 then Product 6	16%	29.41% (5/17)	0.54
If Product 2 then Product 7	16%	11.76% (2/17)	1.36

Rule	Confidence	Support Pb (Result)	Improvement / Rule Lift
If Product 6 then Product 5	60%	29.41% (5/17)	2.04
If Product 3 then Product 2	66%	35.29% (6/17)	1.87

Association Analysis

SAP Predictive Analytics demonstration



Association Analysis

SAP Predictive Analytics demonstration results

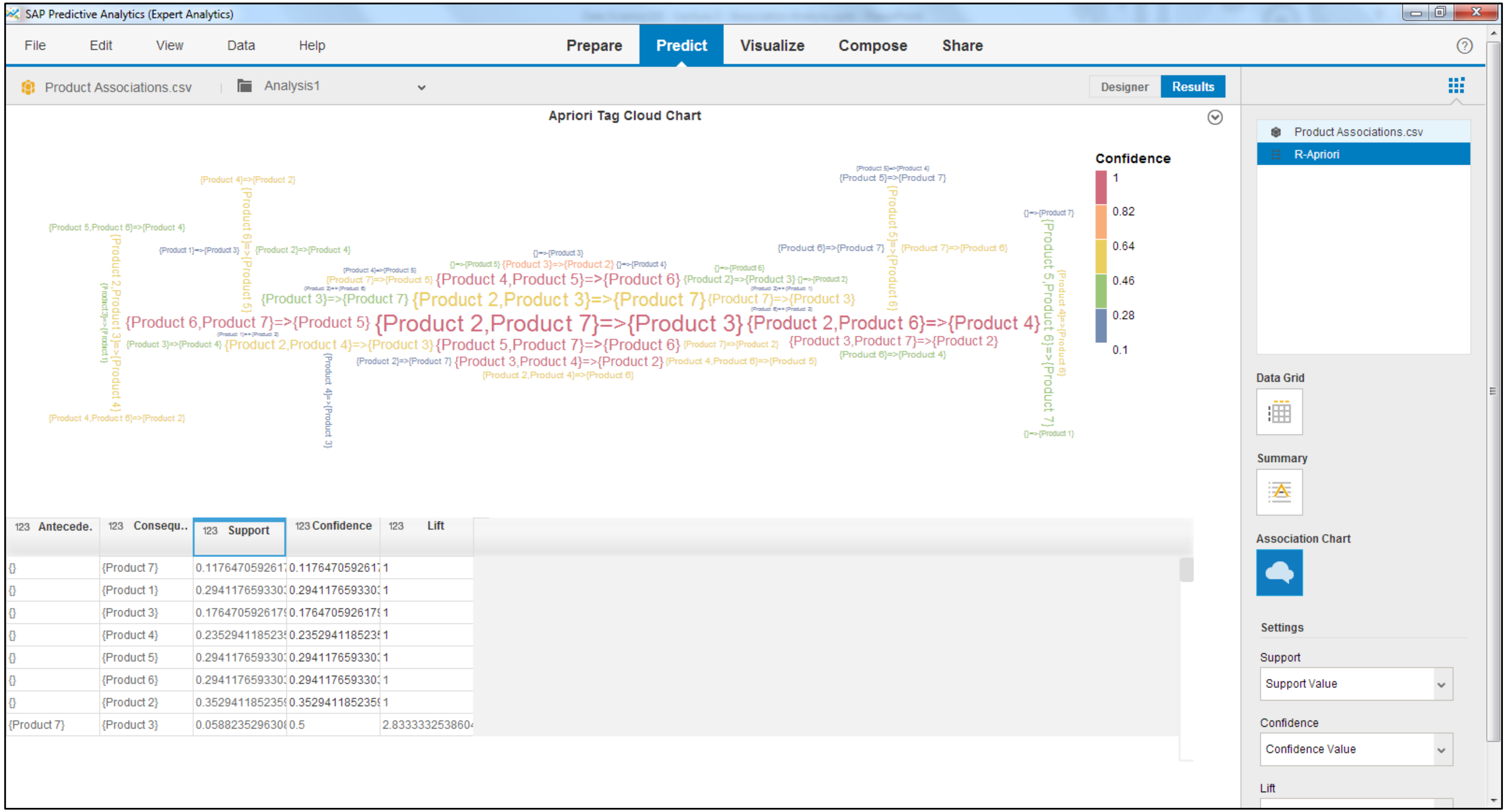
Product Associations.csv | Analysis1

ABC Rules	123 Support	123 Confidence	123 Lift
{} => {Product 7}	0.12	0.12	1.00
{} => {Product 1}	0.29	0.29	1.00
{} => {Product 3}	0.18	0.18	1.00
{} => {Product 4}	0.24	0.24	1.00
{} => {Product 5}	0.29	0.29	1.00
{} => {Product 6}	0.29	0.29	1.00
{} => {Product 2}	0.35	0.35	1.00
{Product 7} => {Product 3}	0.06	0.50	2.83
{Product 3} => {Product 7}	0.06	0.33	2.83
{Product 7} => {Product 5}	0.06	0.50	1.70
{Product 5} => {Product 7}	0.06	0.20	1.70
{Product 7} => {Product 6}	0.06	0.50	1.70
{Product 6} => {Product 7}	0.06	0.20	1.70
{Product 7} => {Product 2}	0.06	0.50	1.42
{Product 2} => {Product 7}	0.06	0.17	1.42
{Product 1} => {Product 3}	0.06	0.20	1.13
{Product 3} => {Product 1}	0.06	0.33	1.13
{Product 1} => {Product 2}	0.06	0.20	0.57
{Product 2} => {Product 1}	0.06	0.17	0.57
{Product 3} => {Product 4}	0.06	0.33	1.42
{Product 4} => {Product 3}	0.06	0.25	1.42
{Product 3} => {Product 2}	0.12	0.67	1.89
{Product 2} => {Product 3}	0.12	0.33	1.89

{Product 4} => {Product 5}	0.06	0.25	0.85
{Product 5} => {Product 4}	0.06	0.20	0.85
{Product 4} => {Product 6}	0.12	0.50	1.70
{Product 6} => {Product 4}	0.12	0.40	1.70
{Product 4} => {Product 2}	0.12	0.50	1.42
{Product 2} => {Product 4}	0.12	0.33	1.42
{Product 5} => {Product 6}	0.18	0.60	2.04
{Product 6} => {Product 5}	0.18	0.60	2.04
{Product 6} => {Product 2}	0.06	0.20	0.57
{Product 2} => {Product 6}	0.06	0.17	0.57
{Product 3,Product 7} => {Product 2}	0.06	1.00	2.83
{Product 2,Product 7} => {Product 3}	0.06	1.00	5.67
{Product 2,Product 3} => {Product 7}	0.06	0.50	4.25
{Product 5,Product 7} => {Product 6}	0.06	1.00	3.40
{Product 6,Product 7} => {Product 5}	0.06	1.00	3.40
{Product 5,Product 6} => {Product 7}	0.06	0.33	2.83
{Product 3,Product 4} => {Product 2}	0.06	1.00	2.83
{Product 2,Product 3} => {Product 4}	0.06	0.50	2.12
{Product 2,Product 4} => {Product 3}	0.06	0.50	2.83
{Product 4,Product 5} => {Product 6}	0.06	1.00	3.40
{Product 4,Product 6} => {Product 5}	0.06	0.50	1.70
{Product 5,Product 6} => {Product 4}	0.06	0.33	1.42
{Product 4,Product 6} => {Product 2}	0.06	0.50	1.42
{Product 2,Product 4} => {Product 6}	0.06	0.50	1.70
{Product 2,Product 6} => {Product 4}	0.06	1.00	4.25

Association Analysis

SAP Predictive Analytics demonstration output



Association Analysis
Summary

- **Strengths**
  - It produces clear and understandable results.
  - The calculations are straightforward and therefore easy to understand.
  - The results are actionable.
  - It is undirected data mining.

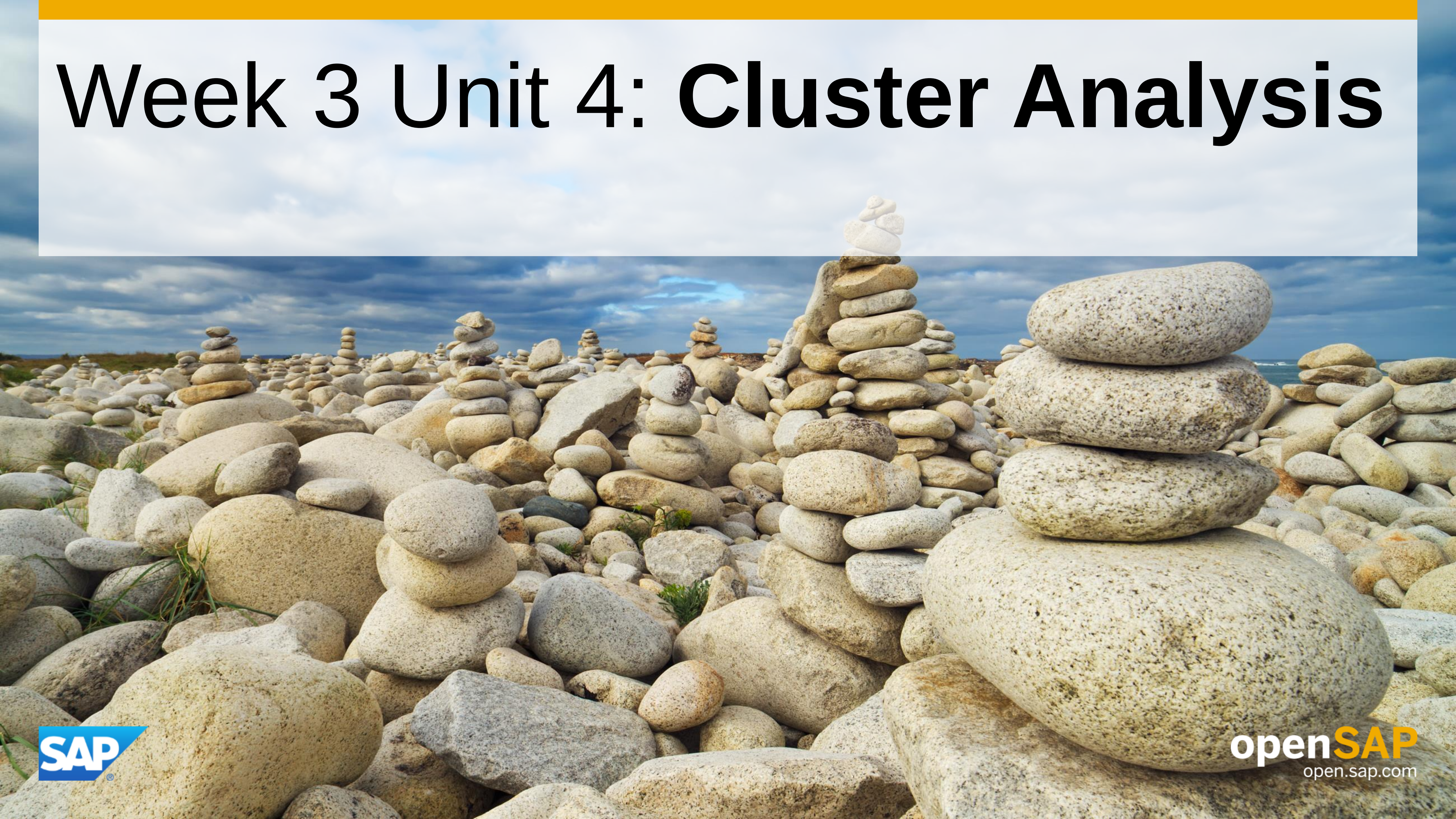
- **Weaknesses**
  - It requires exponentially more computations as the problem size grows.
  - Many of the results are often either trivial or inexplicable.
  - It discounts rare items.
  - It does not allow us to directly include any customer features (if they are available).

Thank you

Contact information:

open@sap.com

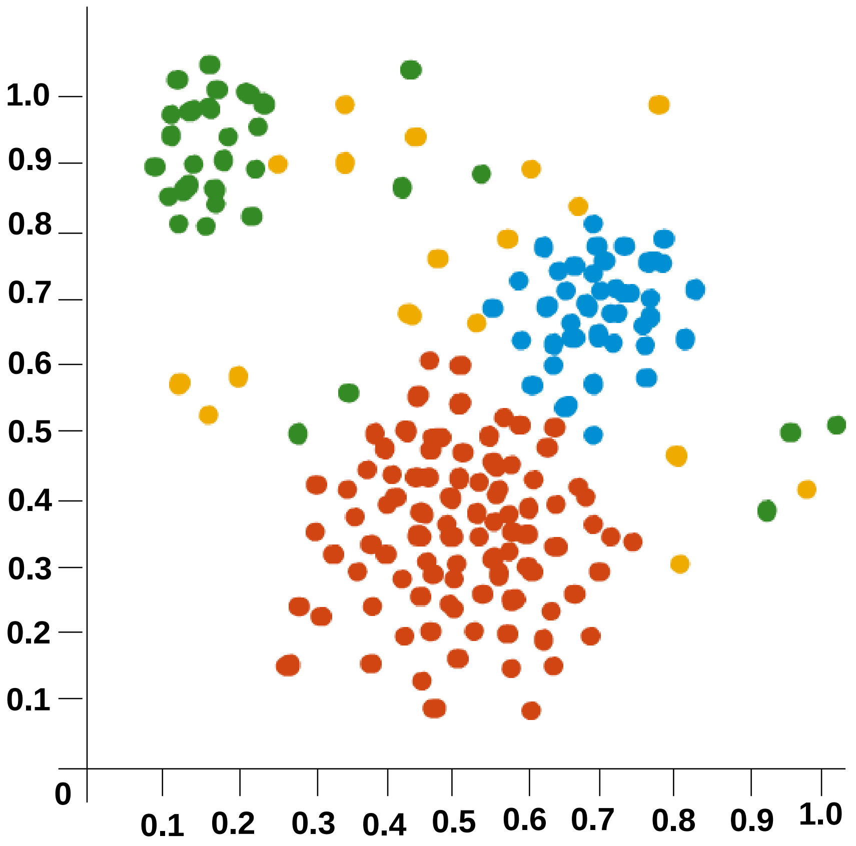
Week 3 Unit 4: **Cluster Analysis**



open.sap.com

Cluster Analysis
Introduction

Cluster analysis or **clustering** is the task of grouping a set of objects in such a way that objects in the same group (called a **cluster**) are more similar (homogeneous in some sense or another) to each other, but are very dissimilar to objects not belonging to that group (heterogeneous).



Cluster Analysis

ABC analysis

	ITEM	VALUE
1	Item1	15.4
2	Item2	50.4
3	Item3	55.4
4	Item4	40.9
5	Item5	30.4
6	Item6	25.6
7	Item7	18.4
8	Item8	10.5
9	Item9	46.5
10	Item10	10.4

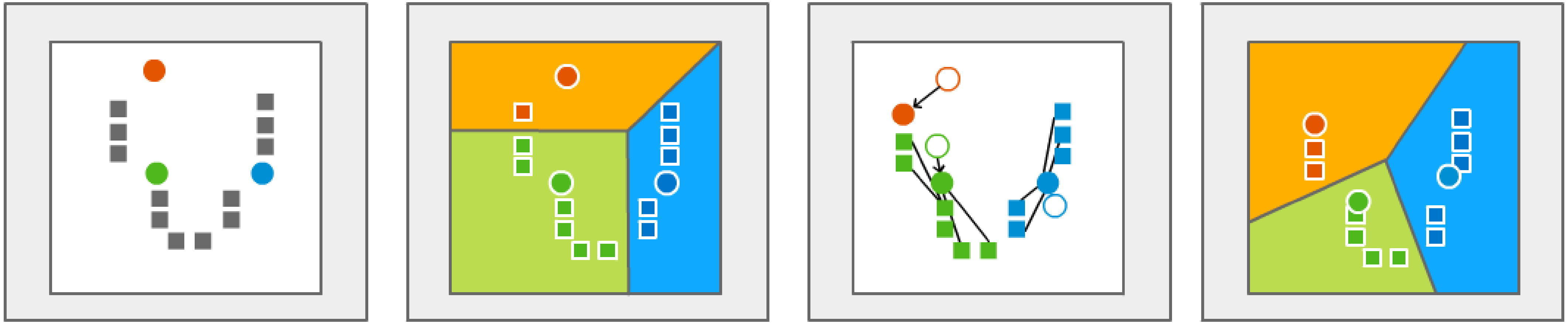
Input Data

	ABC	ITEM
1	A	Item3
2	A	Item2
3	B	Item9
4	B	Item4
5	B	Item5
6	C	Item6
7	C	Item7
8	C	Item1
9	C	Item8
10	C	Item10

ABC Analysis Output

Cluster Analysis

The K-means cluster analysis algorithm



1) ***k*** initial "means" (in this case k=3) are randomly generated within the data domain (shown in color).

2) ***k*** clusters are created by associating every observation with the nearest mean.

3) The centroid of each of the ***k*** clusters becomes the new mean.

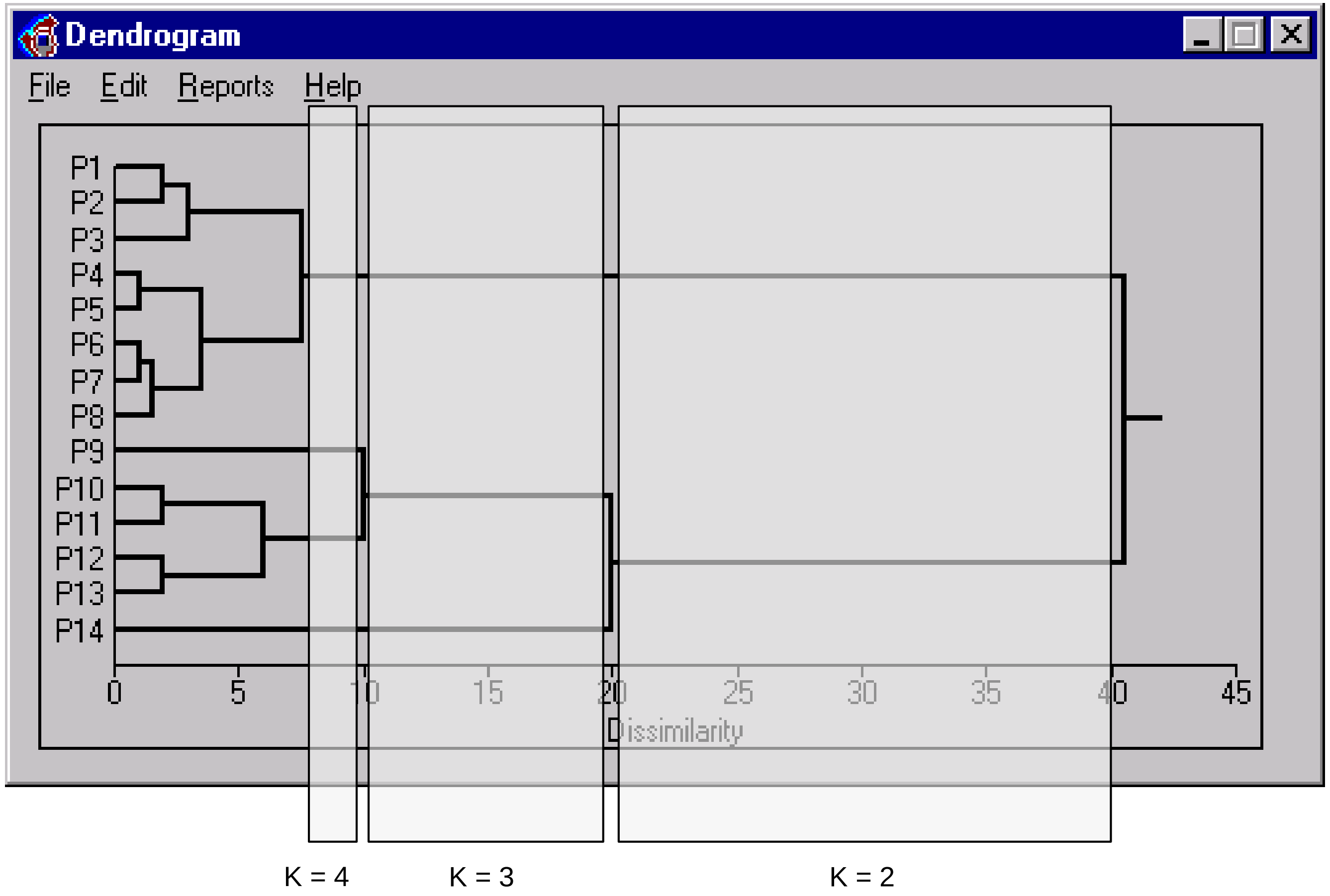
4) Steps 2 and 3 are repeated until convergence has been reached.

http://en.wikipedia.org/wiki/K-means_clustering

Cluster Analysis

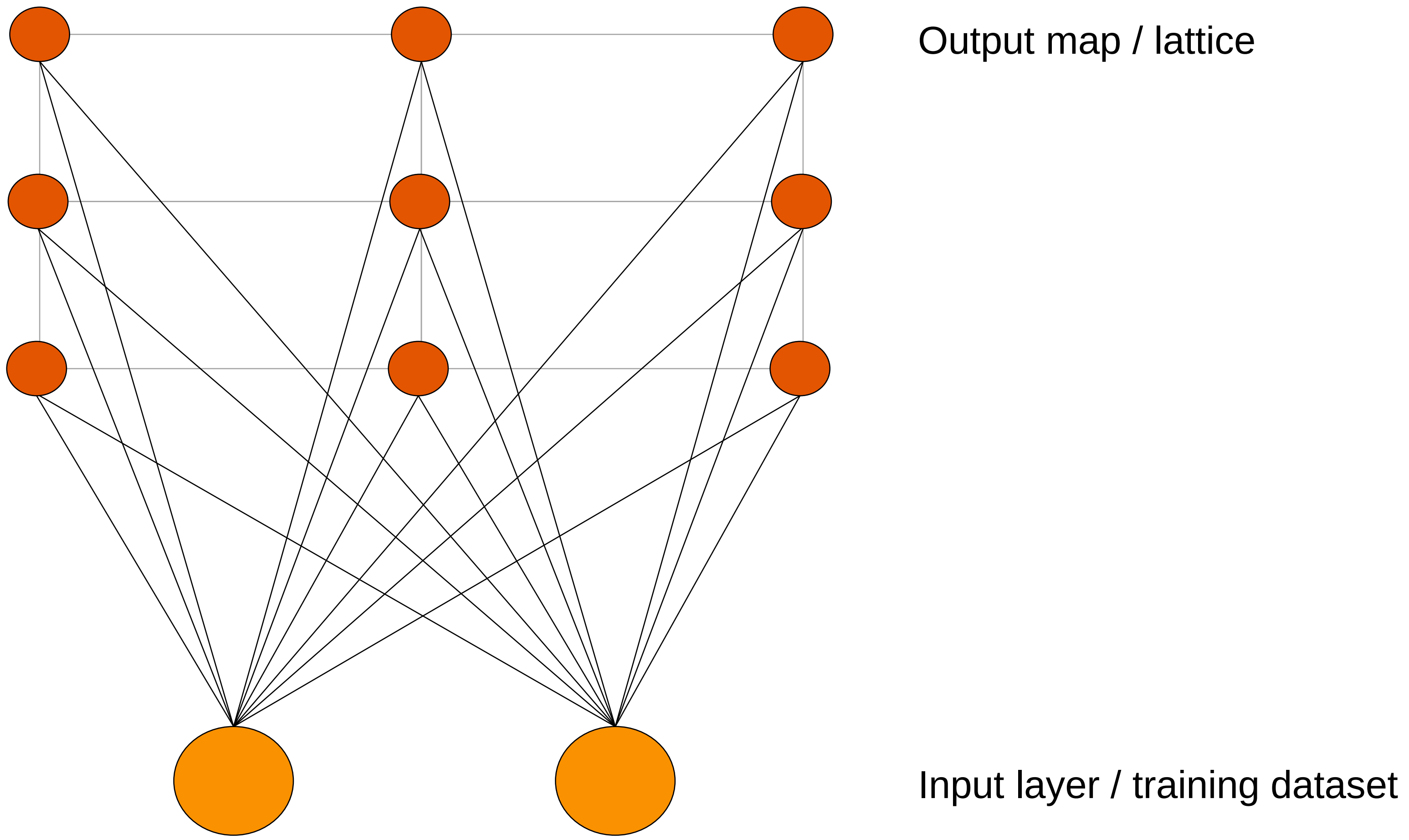
SAP HANA Predictive Analysis Library (PAL) hierarchical agglomeration

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13	P14
1	3	5	8	9	11	12	13	37	43	45	49	51	65



Cluster Analysis

SAP HANA PAL Kohonen self-organizing maps



Cluster Analysis
What makes a good segmentation?

The following key aspects of a segmentation model require consideration:

- Homogeneous – similarity of members within segments
- Heterogeneous – difference between segments
- Stable – segments should be stable over time so that appropriate business/marketing activity can be implemented
- Recognizable – segments must make sense to the business
- Meaningful/relevant – segments must be well defined and actionable
- Manageable – the number and complexity of segments (too few segments make the solution irrelevant, too many segments will be difficult to manage)

Cluster Analysis
Strengths and weaknesses

- **Strengths**
  - Automatic cluster detection is undirected
  - It is easy to understand and to apply
- **Weaknesses**
  - It can sometimes be difficult to interpret the results
  - The results may vary dependent on the choice of distance measure and variable weight
  - K-means is clearly driven by the choice of K
  - K-means can be sensitive to the initial choice of cluster centres
  - Outliers can become clusters

"Clustering is a great tool to use when you are faced with a large, complex data set with many variables and a lot of internal structure. At the start of a new data mining project, clustering is often the best first technique to turn to. It is rarely the only tool, however. Once automatic cluster detection has discovered regions of the data space that contain similar records, other data mining tools have a better chance of discovering rules and patterns within them." Berry & Linoff, Data Mining Techniques

Thank you

Contact information:

open@sap.com

Week 3 Unit 5: **Classification Analysis with Regression**

Classification Analysis with Regression

Introduction to regression

The formula for a simple regression line is represented as an equation:

$$Y = a + bx$$

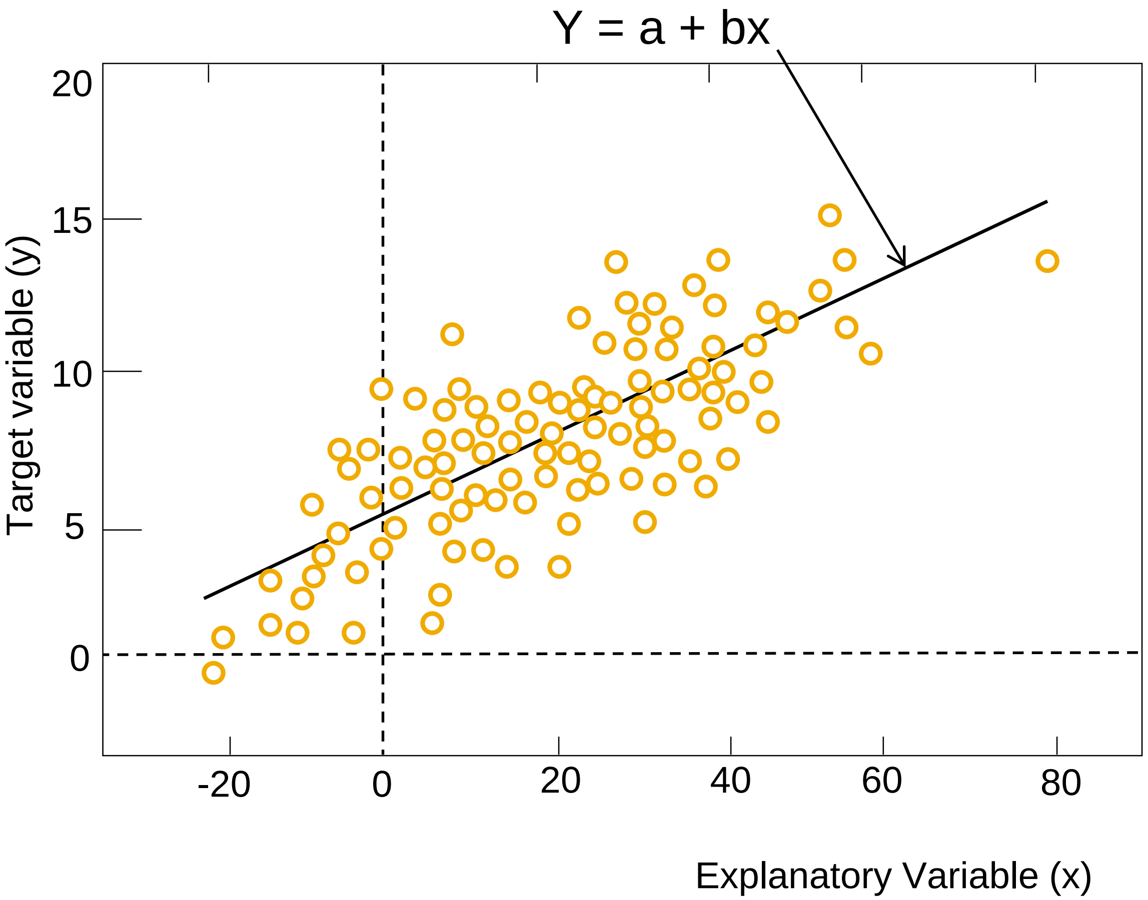
Where

Y is the target

a is the intercept (the level of *Y* where *x* is 0)

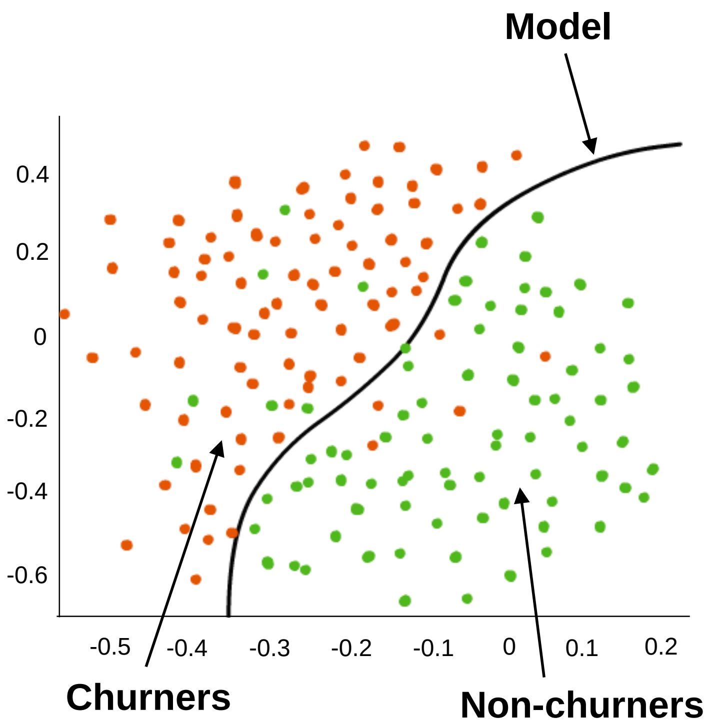
b is the slope of the line

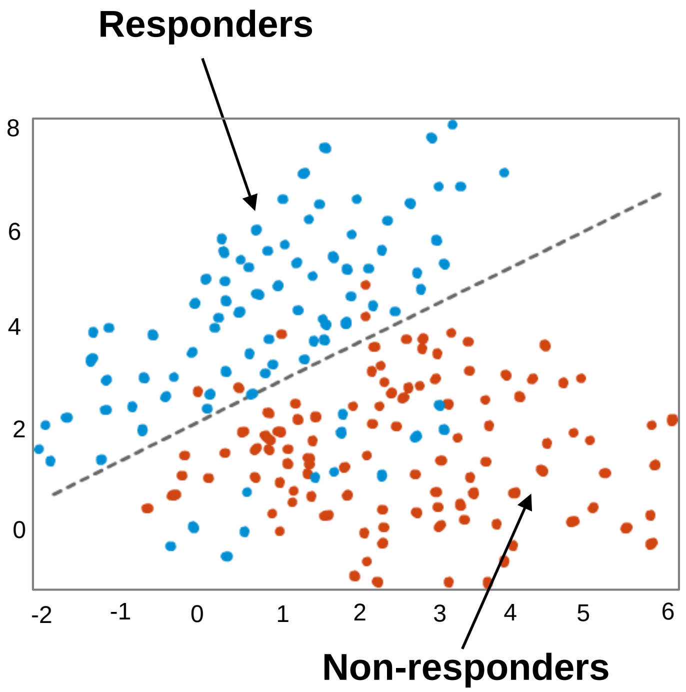
x is the explanatory variable

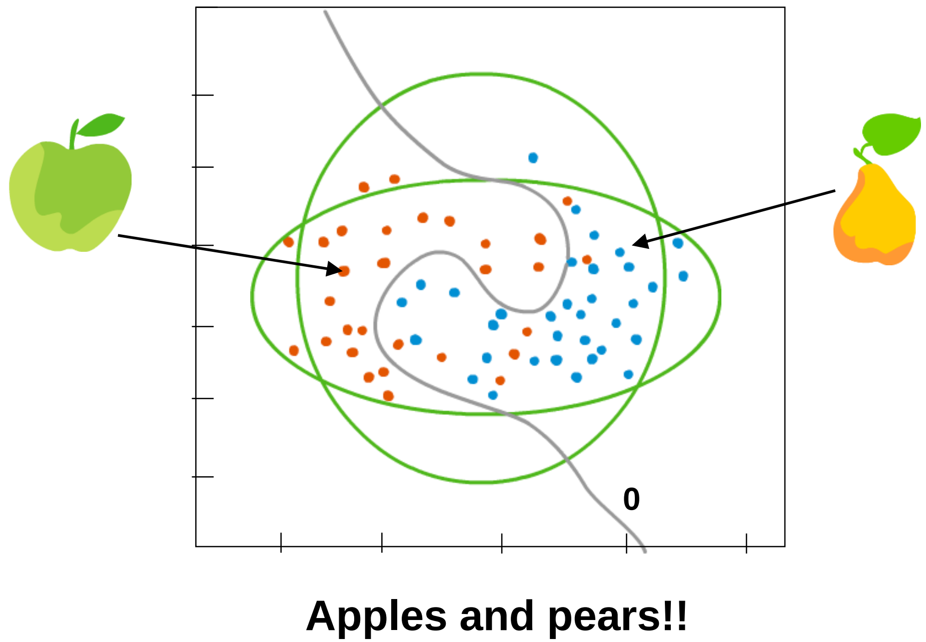


Classification Analysis with Regression

Introduction to classification





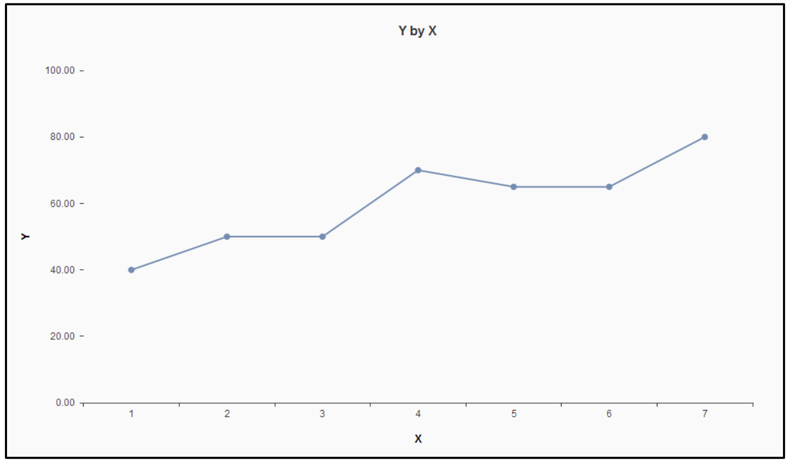


Classification Analysis with Regression

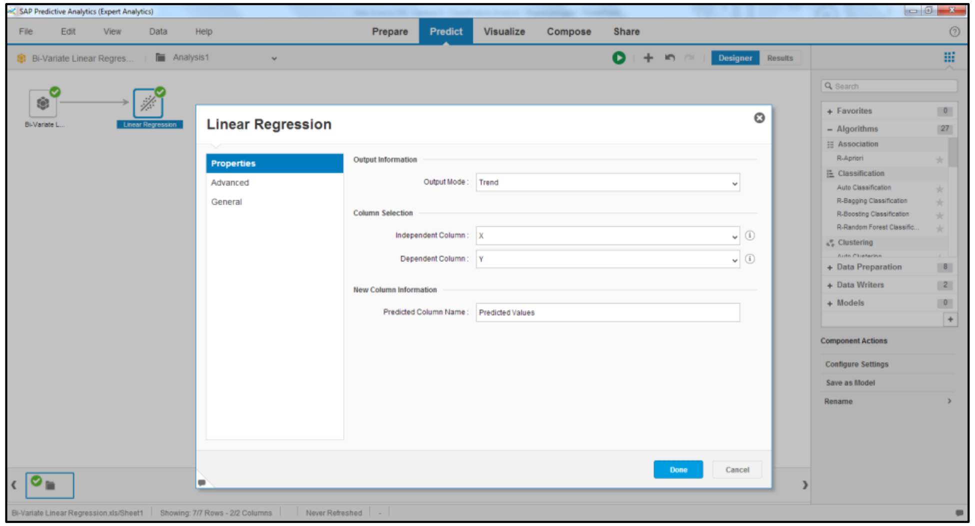
Linear regression analysis: demonstration

123 X	123 Y
1	40
2	50
3	50
4	70
5	65
6	65
7	80

Input Data



Visualization of Input Data



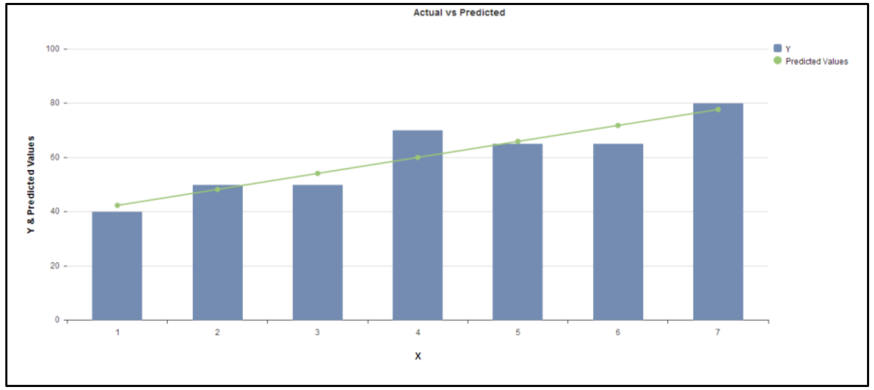
Linear Regression in SAP Predictive Analytics Expert

Classification Analysis with Regression

Linear regression analysis: demonstration

123 X	123 Y	123 Predicte..
1	40	42.32
2	50	48.21
3	50	54.11
4	70	60.00
5	65	65.89
6	65	71.79
7	80	77.68

Output Predicted Values



Visualization of Output Data

Bi-Variate Linear Regres... | Analysis1

```
Summary from SAP Algorithms
---------------------------------

Information of the columns used in the algorithm
----------------------------------------------------
Independent Column
 X : Integer
Dependent Column
 Y : Integer

The formula used is "Y = intercept + slope * X"
intercept 36.4286
slope 5.8929

The goodness of fit coefficient is 0.8146
The R-square factor is 0.8455
The f-value is 27.3618
The Standard Error of Estimate is 5.9612
The confidence levels for the slope are 8.7892 2.9965
The confidence levels for the intercept are 42.2213 30.6358
```

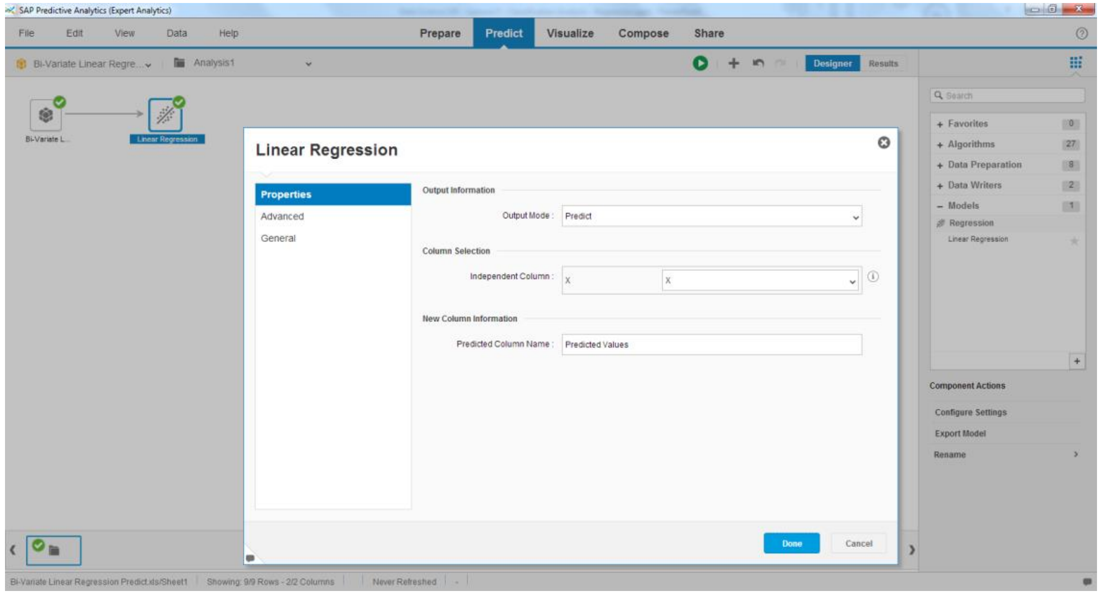
Linear Regression Model Summary Report

Classification Analysis with Regression

Applying the model – Model scoring: demonstration

123 X	123 Y
1	40
2	50
3	50
4	70
5	65
6	65
7	80
8	
9	

New Data



Apply the Regression Model in SAP Predictive Analytics

123 X	123 Y	123 Predicte..
1	40	42.32
2	50	48.21
3	50	54.11
4	70	60.00
5	65	65.89
6	65	71.79
7	80	77.68
8		83.57
9		89.46

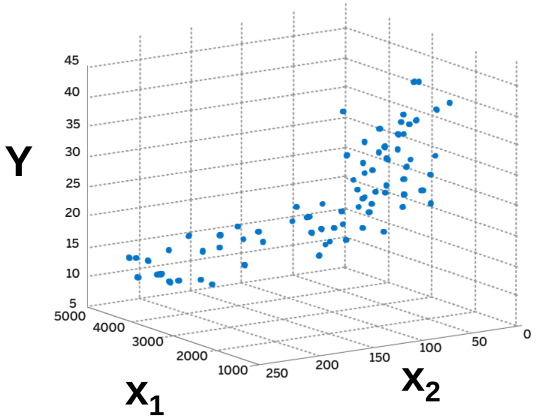
Create Predictions

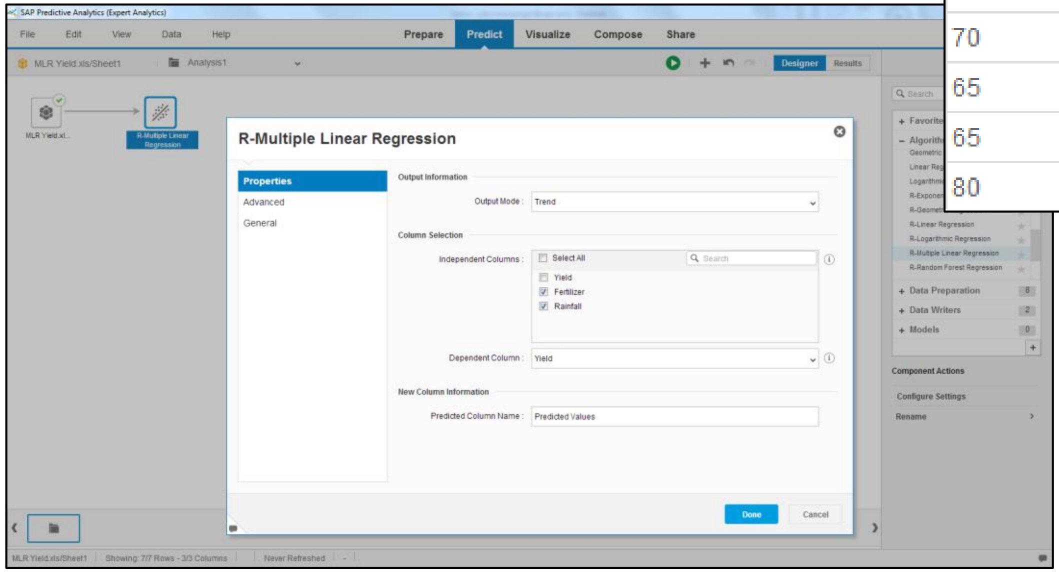
Classification Analysis with Regression

Multiple linear regression: demonstration

MLR Yield.xls/Sheet1 | Ana

123 Yield	123 Fertilizer	123 Rainfall
40	100	10
50	200	20
50	300	10
70	400	30
65	500	20
65	600	20
80	700	30





SAP Predictive Analytics

MLR Yield.xls/Sheet1 | Analysis1

123 Yield	123 Fertilizer	123 Rainfall	123 Predicte..
40	100	10	40.24
50	200	20	52.38
50	300	10	47.86
70	400	30	68.33
65	500	20	63.81
65	600	20	67.62
80	700	30	79.76

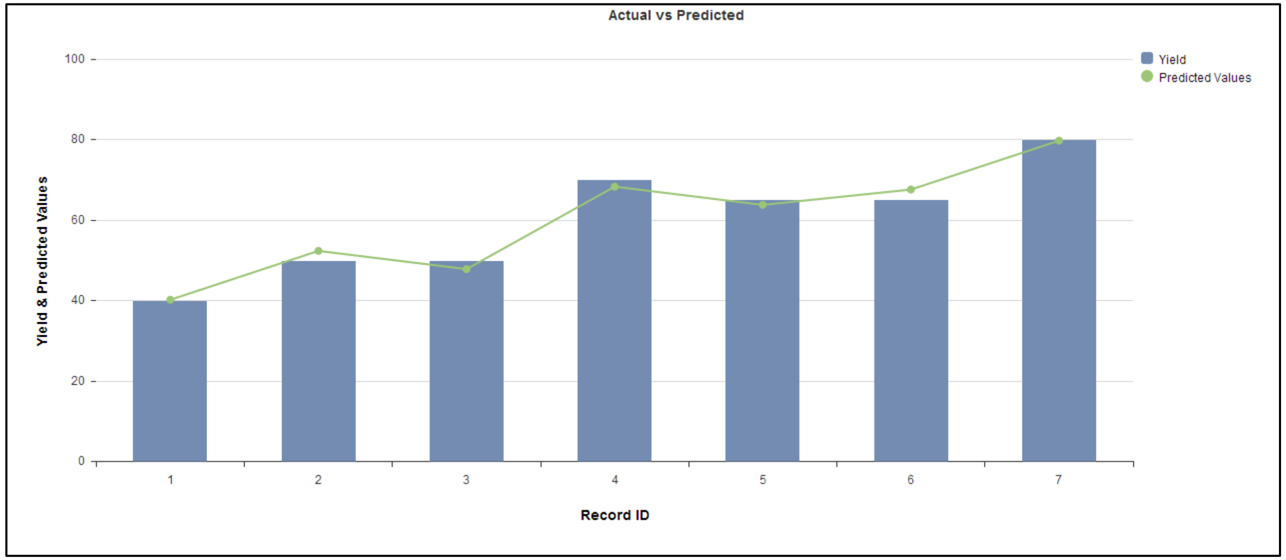
Output with Predicted Yield Values

Classification Analysis with Regression

Multiple linear regression: demonstration

MLR Yield.xls/Sheet1 | Analysis1

```
Summary of the model from R Scripts

Information of the columns used in the algorithm
----------------------------------------------------
Independent Columns
 Fertilizer : Integer
 Rainfall : Integer
Dependent Column
 Yield : Integer

Summary of the Model

Call:
lm(formula = Yield ~ Fertilizer + Rainfall, na.action = na.omit)

Residuals:
      1       2       3       4       5       6       7
-0.2381 -2.3810  2.1429  1.6667  1.1905 -2.6190  0.2381

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 28.095238   2.491482  11.277 0.000352 ***
Fertilizer   0.038095   0.005832   6.532 0.002838 **
Rainfall     0.833333   0.154303   5.401 0.005690 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.315 on 4 degrees of freedom
Multiple R-squared:  0.9814,    Adjusted R-squared:  0.972
F-statistic: 105.3 on 2 and 4 DF,  p-value: 0.0003472
```

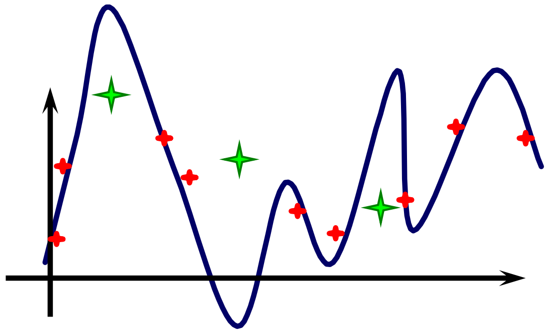


Plot of Actual and Predicted Values

Multiple Linear Regression Model Summary Report

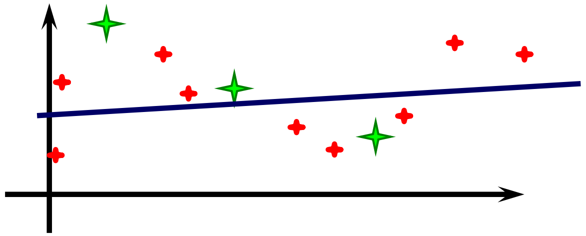
Classification Analysis with Regression

Over-fitting



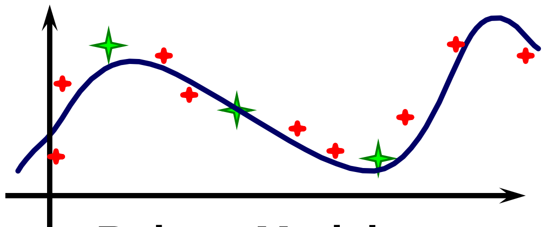
Over-Fit Model/Low Robustness

(No Training Error, High Test Error)



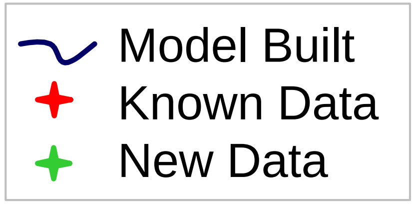
Under-Fit Model/High Robustness

(High Training Error = High Test Error)



Robust Model

(Low Training Error ≈ Low Test Error)



Classification Analysis with Regression

Summary

- **Strengths**
  - It is easy to understand and to apply
- **Weaknesses**
  - Significantly affected by outliers
  - Can suffer from over-fitting

Thank you

Contact information:

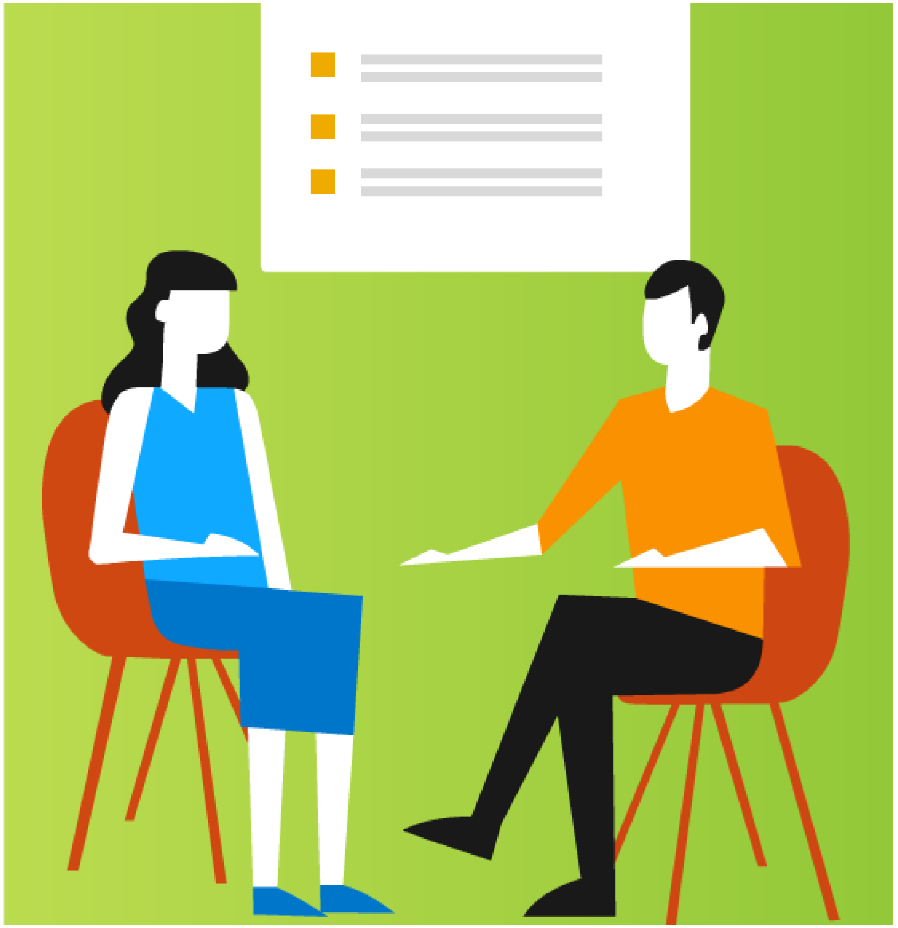
open@sap.com

Classification Analysis with Regression

Appendix

Additional Material

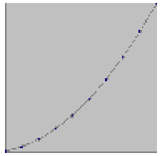
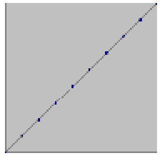
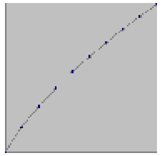
- Bi-Variate Regression Variations
- Polynomial Regression
- Logistic Regression



Classification Analysis with Regression

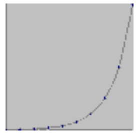
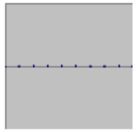
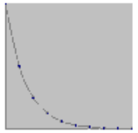
Bi-variate regression variations

Bi-Variate Geometric Regression $Y = a*X^b$

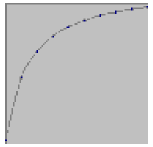
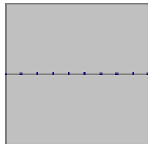
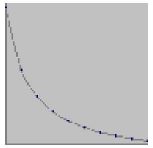
Geometric	Parameter Conditions	Curve
$Y = aX^b$	$b > 1$	
	$b = 1$	
	$0 < b < 1$	

Bi-Variate Exponential Regression $Y = a*b^X$

The Exponential curve, $y = ab^x$ is transformed to linear as $\log y = \log a + (\log b)\, x$

Exponential	Parameter Conditions	Curve
$y = ab^x$	$b > 1$	
	$b = 1$	
	$b < 1$	

Bi-Variate Natural Log Regression $Y = a + b * \log(X)$

Natural Log	Parameter Conditions	Curve
$y = a + b \log(x)$	$b > 0$	
	$b = 0$	
	$b < 0$	

There are a wide variety of other regressions that fit different types of curve to the data.

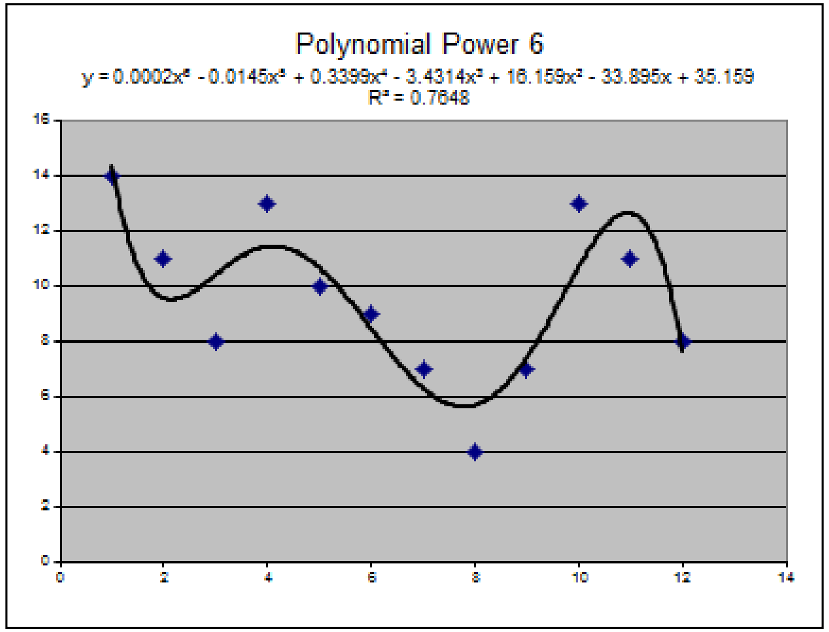
Classification Analysis with Regression

Polynomial regression

- The general form of the model is Y = β0 + β1 * X 1 + β2 * X 12 + β3 * X 13 + … + βn * X 1n where the parameters β0, β1,… βN are estimated using the principle of least squares.
- The value of N determines the degree of the polynomial.
- The SAP HANA Predictive Analysis Library (PAL) simply takes the variable X, raises it to the required degree of the polynomial, and then uses multiple linear regression to determine the model parameters β0, β1,… βN

- An example -

X	Y
1	14
2	11
3	8
4	13
5	10
6	9
7	7
8	4
9	7
10	13
11	11
12	8



Classification Analysis with Regression

Logistic regression

- In predictive analysis, we frequently come across applications where we want to predict a binary variable (0 or 1) or a categorical variable (yes or no). Such a dependent variable is also referred to as a dichotomous variable – something which is divided into two parts or classifications. The problem can be extended to predicting more than two integer values or categories. This is classification analysis.
- Examples are:
  - Churn analysis to predict the probability that a customer may leave/stay.
  - Success/failure of a medical treatment, dependent on dosage, patient's age, sex, weight, and severity of condition.
  - High/low cholesterol level, dependent on sex, age, whether a person smokes or not, etc.
  - Vote for/against political party, dependent on age, gender, education level, region, ethnicity, etc.
  - Yes/No or Agree/Disagree responses to questionnaire items in a survey.
  - There are a huge number of applications in a range of fields, including artificial neural networks, biology, biomathematics, demography, economics, chemistry, mathematical psychology, probability, sociology, political science, and statistics.
- Let's look at an example. We'll use a subset of the MTCARS dataset, which comprises fuel consumption and 10 aspects of automobile design and performance for 32 automobiles.

  http://stat.ethz.ch/R-manual/R-devel/library/datasets/html/mtcars.html

Classification Analysis with Regression

Logistic regression

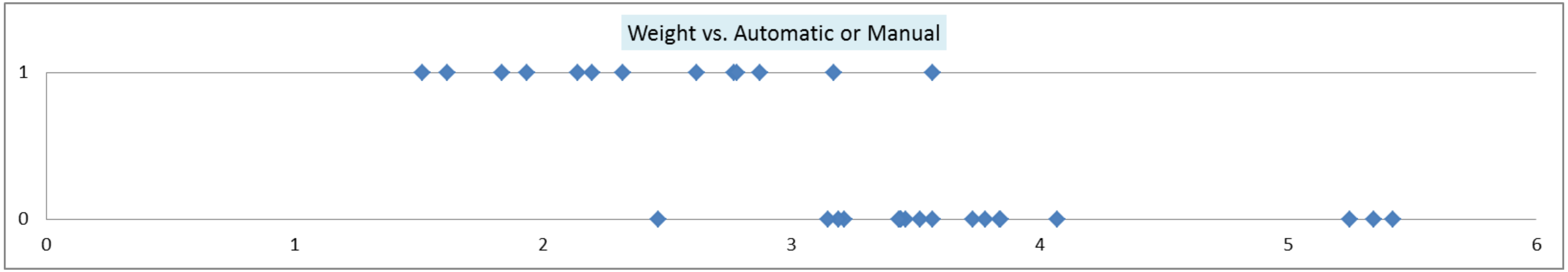
- MTCARS – the subset of the data we will use in our example is shown here. We wish to predict the type of vehicle (automatic or manual) depending on its horsepower and weight.
- The target is automatic/manual (a binary variable).
- The explanatory variables are horsepower and weight.

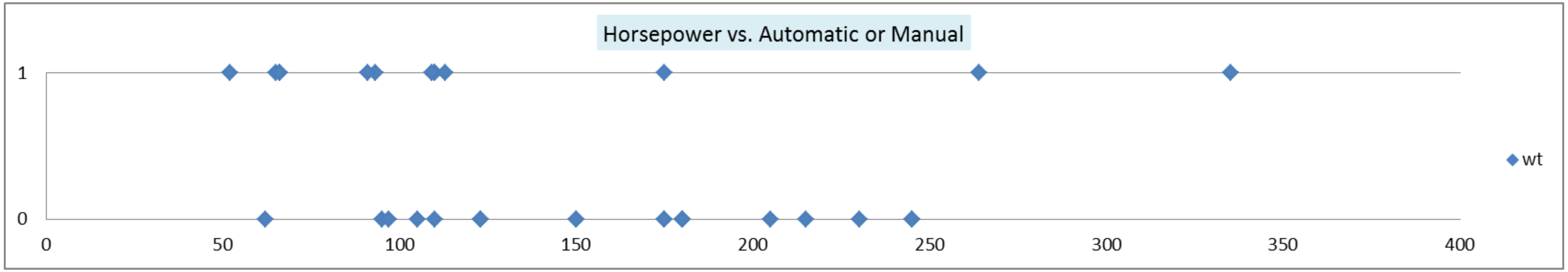
Vehicle	Horsepower	Weight	0 = Automatic, 1 = Manual
Mazda RX4	110	2.62	1
Mazda RX4 Wag	110	2.875	1
Datsun 710	93	2.32	1
Hornet 4 Drive	110	3.215	0
Hornet Sportabout	175	3.44	0
Valiant	105	3.46	0
Duster 360	245	3.57	0
Merc 240D	62	3.19	0
Merc 230	95	3.15	0
Merc 280	123	3.44	0
Merc 280C	123	3.44	0
Merc 450SE	180	4.07	0
Merc 450SL	180	3.73	0
Merc 450SLC	180	3.78	0
Cadillac Fleetwood	205	5.25	0
Lincoln Continental	215	5.424	0
Chrysler Imperial	230	5.345	0
Fiat 128	66	2.2	1
Honda Civic	52	1.615	1
Toyota Corolla	65	1.835	1
Toyota Corona	97	2.465	0
Dodge Challenger	150	3.52	0
AMC Javelin	150	3.435	0
Camaro Z28	245	3.84	0
Pontiac Firebird	175	3.845	0
Fiat X1-9	66	1.935	1
Porsche 914-2	91	2.14	1
Lotus Europa	113	1.513	1
Ford Pantera L	264	3.17	1
Ferrari Dino	175	2.77	1
Maserati Bora	335	3.57	1
Volvo 142E	109	2.78	1

Classification Analysis with Regression

Logistic regression

Here's a plot of the data – each explanatory variable vs. the target variable



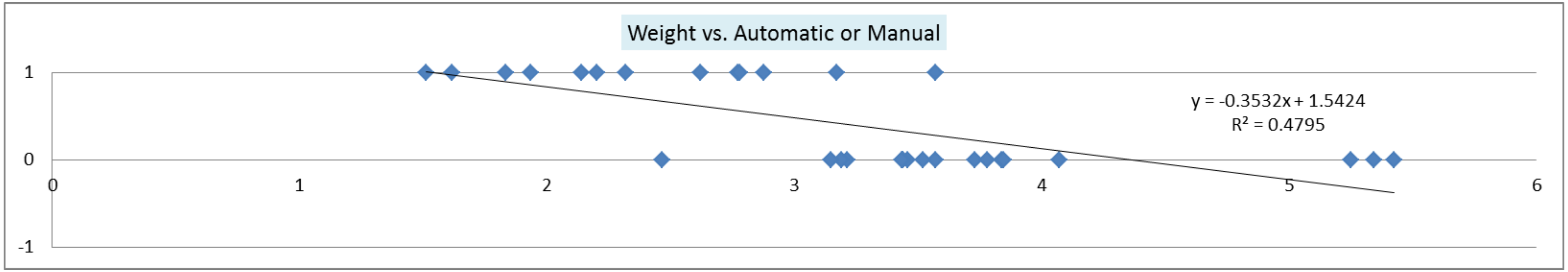


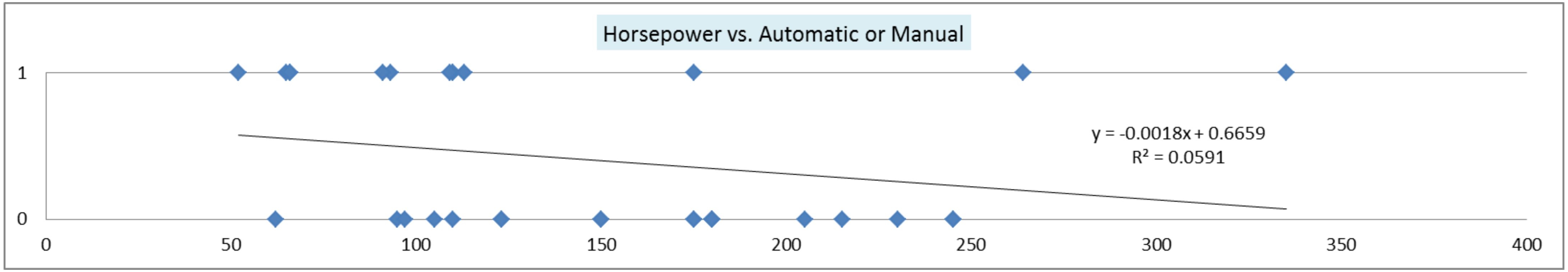
These plots emphasize the binary (0 or 1) nature of the target variable.

Classification Analysis with Regression

Logistic regression

A linear regression fit is a very poor fit:



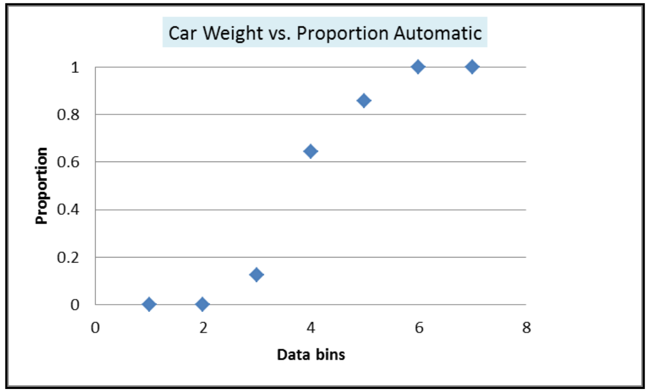


Classification Analysis with Regression

Logistic regression

- Linear regression is not suitable for other reasons. From a statistical perspective, the method assumes:
  - Linearity of the relationship between the dependent and independent variables
  - Normality of the error distribution of actual vs. fitted
  - Independence of the errors
  - A constant variance of the errors. Statisticians call this homoscedasticity – nice word ☺.
- All these assumptions do not apply when the dependent variable is dichotomous.
- Logistic regression is used for these types of binary dependent variable applications, whereby we estimate the probability that the outcome could be a 0 or 1.
- We predict the likelihood that Y is equal to 1 (rather than 0) given certain values of X. We think about predicting probabilities rather than the scores of the dependent variable.
- Here's a plot of the car weight vs. the proportion of 0 = automatic, 1 = manual – the likelihood / probability.

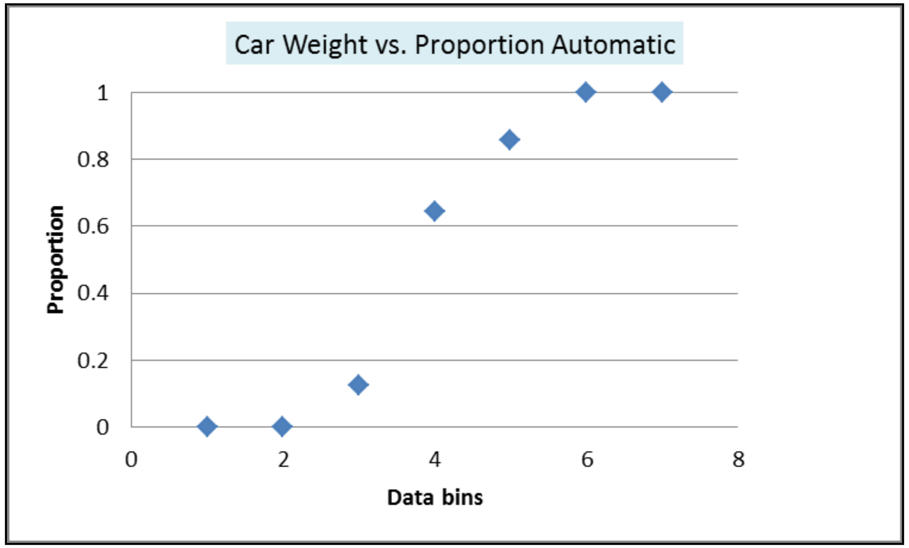
Data Bins	Automatic	Manual	Proportion
< 1.5	0	1	0
1.5- 2.5	1	7	0.125
2.6 - 3.5	9	5	0.643
3.6 - 4.5	6	1	0.857
4.6 - 5.5	3	0	1.000
> 5.5	1	0	1



Classification Analysis with Regression

Logistic regression

- This curve is very well known. It's a logistic curve or logistic function, and is a common sigmoid function. A generalized logistic curve can model the "S-shaped" behavior (abbreviated S-curve) of growth of a variable. The initial stage of growth is approximately exponential; then, as saturation begins, the growth slows, and at maturity, growth stops.



- A simple logistic function may be defined by the formula

  $p(t) = 1 / (1 + e^{-t})$

  where p(t) is the proportion or probability of the target group or class, usually coded as 1.

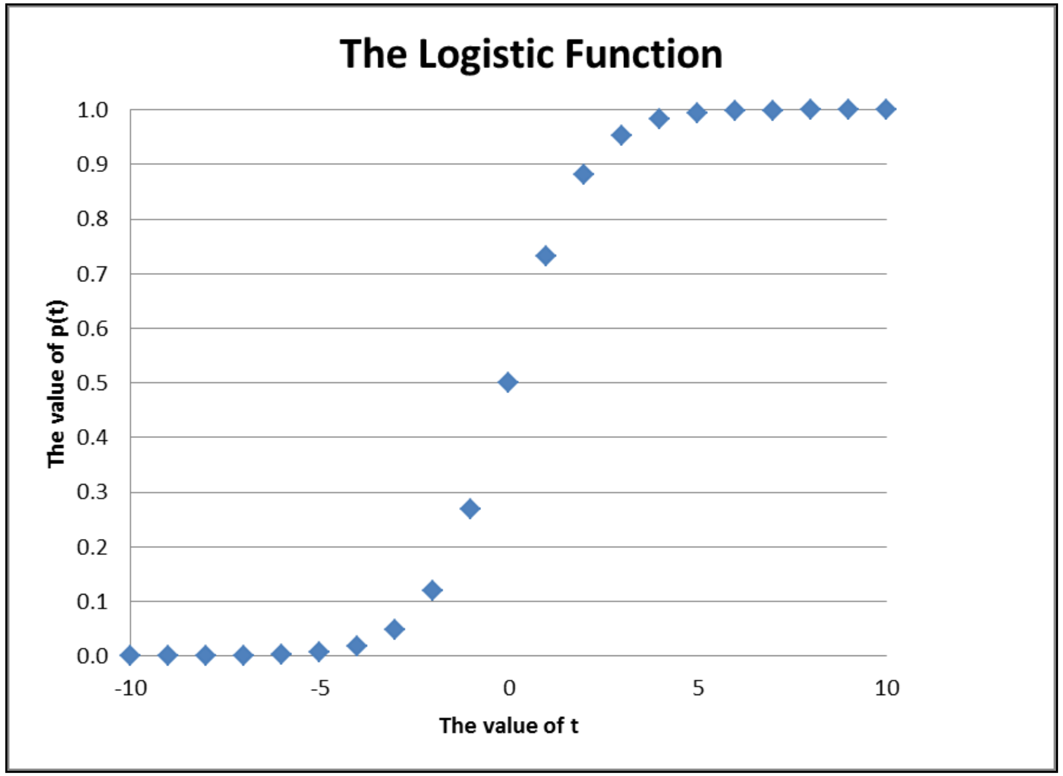
- The benefit of this curve is that the input values can range from $-\infty < t < \infty$ whilst the output ranges from 0 to 1, exactly the range for probability values.

Classification Analysis with Regression

Logistic regression

The curve $p(t) = 1 / (1 + e^{-t})$

The Logistic Function or Logisitc Curve			$p(t) = 1/(1+e^{-t})$
t	EXP(t)	EXP(-t)	= 1/(1+EXP(-t))
-10	0.000045	22026.465795	0.000045
-9	0.000123	8103.083928	0.000123
-8	0.000335	2980.957987	0.000335
-7	0.000912	1096.633158	0.000911
-6	0.002479	403.428793	0.002473
-5	0.006738	148.413159	0.006693
-4	0.018316	54.598150	0.017986
-3	0.049787	20.085537	0.047426
-2	0.135335	7.389056	0.119203
-1	0.367879	2.718282	0.268941
0	1.000000	1.000000	0.500000
1	2.718282	0.367879	0.731059
2	7.389056	0.135335	0.880797
3	20.085537	0.049787	0.952574
4	54.598150	0.018316	0.982014
5	148.413159	0.006738	0.993307
6	403.428793	0.002479	0.997527
7	1096.633158	0.000912	0.999089
8	2980.957987	0.000335	0.999665
9	8103.083928	0.000123	0.999877
10	22026.465795	0.000045	0.999955



Classification Analysis with Regression

Logistic regression

- The curve $p(t) = 1 / (1 + e^{-t})$ is described as the logistic curve
- The curve $p(t) = 1 / (1 + e^{-(\beta 0 + \beta 1 * X1)})$ is a logistic regression equation. If $\beta_0 = 0$ and $\beta_1 = 1$ then we have the first curve
- To be more general, the logistic regression equation for multiple explanatory variables is

$$Y = 1 / (1 + e^{-(\beta 0 + \beta 1 * X1 + \beta 2 * X2 + \dots \beta n * Xn)})$$

- Instead of finding the best fitting line by minimizing the squared residuals as in ordinary least squares, we use a different approach with logistic regression — Maximum Likelihood Estimation.
- This is a way of finding the smallest possible deviance between the observed and predicted values, similar to finding the best fitting line, using calculus. Maximum Likelihood Estimation uses different "iterations" in which it tries different solutions until it gets the smallest possible deviance or best fit. Once it has found the best solution, it provides a final value for the deviance, which is usually referred to as "negative two log likelihood“.

Classification Analysis with Regression

Logistic regression: worked example

- This example is based on a subset of the MTCARS data as shown previously.
- The model is the probability $Y = 1 / (1 + e^{-(\beta0 + \beta1 * X1 + \beta2 * X2)})$
- Specifically $\text{Prob}(Y=1)$ is $1 / (1 + e^{-(\beta0 + \beta1 * \text{Horsepower} + \beta2 * \text{Weight})})$

- The SAP HANA Predictive Analysis Library (PAL) procedure is LOGISTICREGRESSION
- The Input table structure is as follows –

Input Table

Table	Column	Column Data Type	Description	Constraint
Data	Columns	Integer or double	Variable X*n*	
	Type column	Integer	Variable TYPE	Only 0 and 1 are supported

- The data entered in SQLScript and shown in an SAP HANA table -

```
CREATE COLUMN TABLE DATA_TAB ("X1" DOUBLE,"X2"DOUBLE,"TYPE" INT);
INSERT INTO DATA_TAB VALUES (110,2.62,1);
INSERT INTO DATA_TAB VALUES (110,2.875,1);
INSERT INTO DATA_TAB VALUES (93,2.32,1);
INSERT INTO DATA_TAB VALUES (110,3.215,0);
INSERT INTO DATA_TAB VALUES (175,3.44,0);
INSERT INTO DATA_TAB VALUES (105,3.46,0);
INSERT INTO DATA_TAB VALUES (245,3.57,0);
INSERT INTO DATA_TAB VALUES (62,3.19,0);
…
```

SELECT * FROM DATA_TAB

	X1	X2	TYPE
1	110.0	2.62	1
2	110.0	2.875	1
3	93.0	2.32	1
4	110.0	3.215	0
5	175.0	3.44	0
6	105.0	3.46	0
7	245.0	3.57	0
8	62.0	3.19	0
9	95.0	3.15	0
10	123.0	3.44	0
11	123.0	3.44	0
12	180.0	4.07	0
13	180.0	3.73	0
14	180.0	3.78	0
15	205.0	5.25	0
16	215.0	5.424	0
17	230.0	5.345	0
18	66.0	2.2	1
19	52.0	1.615	1
20	65.0	1.835	1
21	97.0	2.465	0
22	150.0	3.52	0
23	150.0	3.435	0
24	245.0	3.84	0
25	175.0	3.845	0
26	66.0	1.935	1
27	91.0	2.14	1
28	113.0	1.513	1
29	264.0	3.17	1
30	175.0	2.77	1
31	335.0	3.57	1
32	109.0	2.78	1

Classification Analysis with Regression

Logistic regression: worked example

- The Result table

```
SELECT * FROM RESULTS_TAB
```

	ID	Ai
1	0	18.86629...
2	1	0.036255...
3	2	-8.083475...

- The model is Y= 1 / (1 + e – (18.86629 + 0.036255 * Horsepower – 8.084475 * Weight))

- To measure the “goodness of fit”, we can compare actual Y with fitted Y and build a classifier confusion matrix as in decision tree analysis. This is discussed later under Decision Trees.