



Data Mining – LAB 2

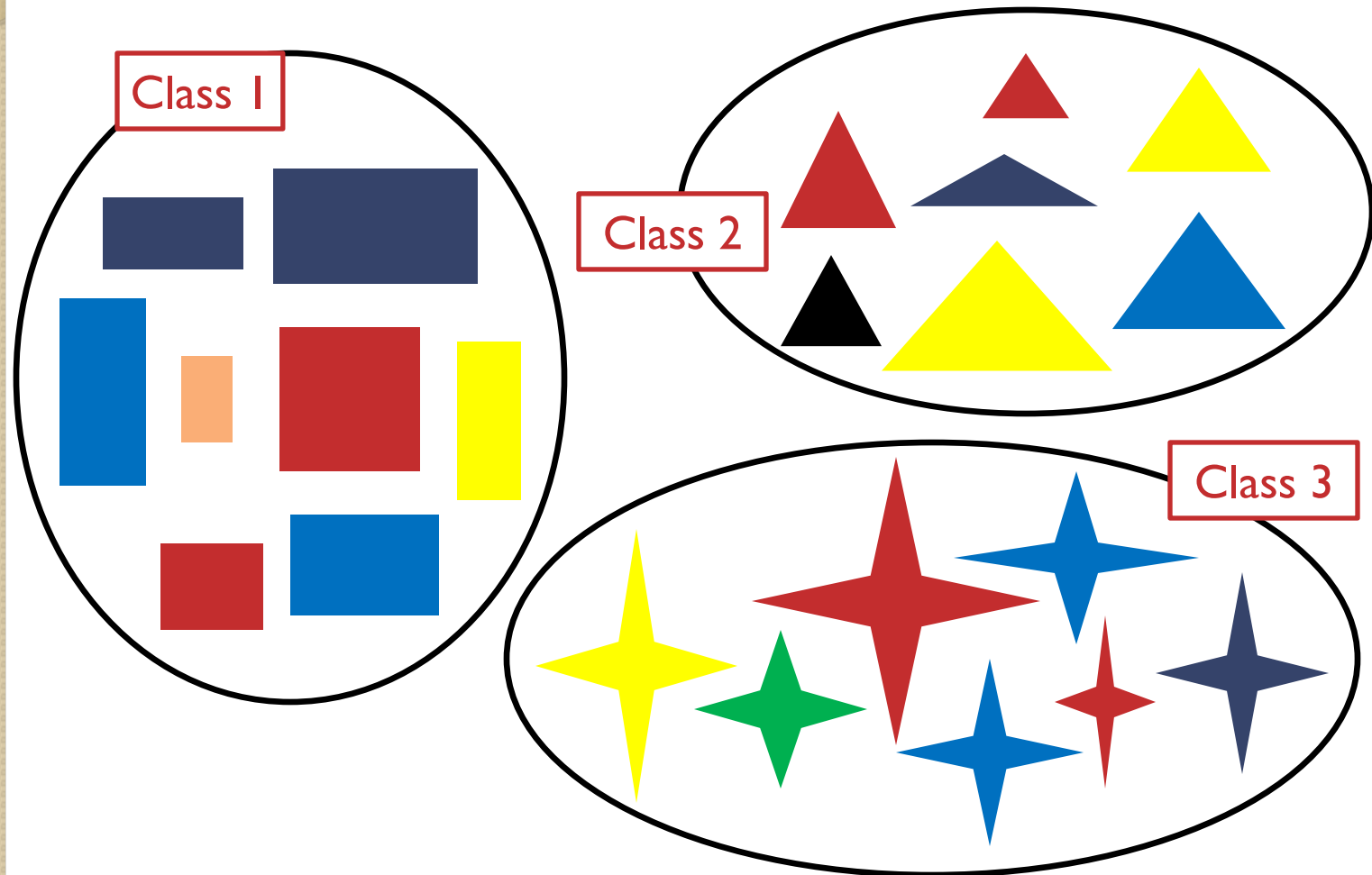
Classification – Decision Tree (C5.0)

Data File

- In Virtualbox image:
 - c:\data\data-play.csv
 - c:\data\data-Lab2.csv
- Download from:
 - www.comp.polyu.edu.hk/~csamak/data/data-play.csv
 - www.comp.polyu.edu.hk/~csamak/data/data-Lab2.mdb

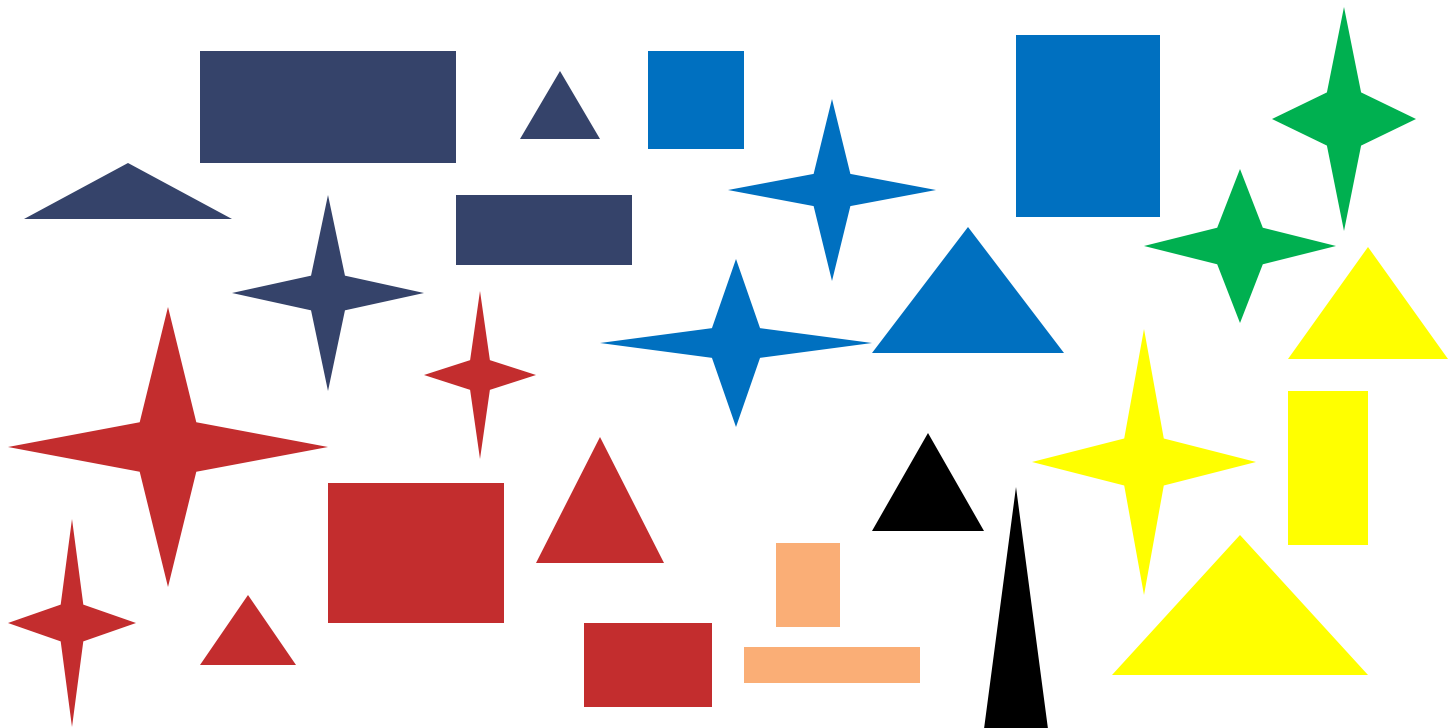
Classification

- With predefined class – e.g., in shape



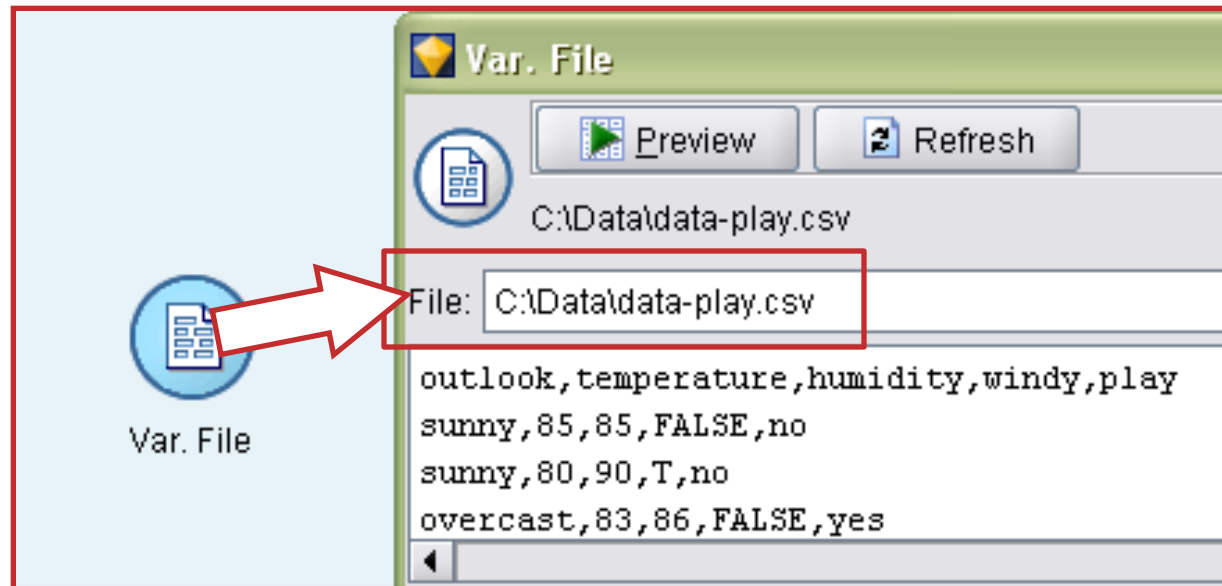
Clustering

- No class is defined in advance
- Shape, Color or Size



Import Data – data-play.csv

- Use “Var. File” to import the data file, data-play.csv



Set Target

1. Change tag to "Type"
2. Click button, "Clear All Values"
3. Click button, "Read Values"
4. Highlight "play" and change Direction to "Out"

data-play.csv

Preview Ref

C:\Data\data... csv

Read Values Clear Values Clear All Values

Field	Type	Values	Missing	Check	Direction
outlook	Set	overcast,ra...		None	In
temperature	Range	[64,85]		None	In
humidity	Range	[65,96]		None	In
windy	Set	"" ,F,FALSE,...		None	In
play	Flag	yes/no		None	In

☒ View current fields ☐ View unused field settings

File Data Filter Types Annotations

OK Cancel Apply

Direction dropdown options: In, Out, Both, None, Partition, Split

Data Understanding



Table (5 fields, 14 records)

	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	T	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	F	yes

Data Audit of [outlook temperature humidity windy play]

Field	Graph	Type	Min	Max	Mean	Std. Dev	Skewness	Unique	Valid
outlook		Set	--	--	--	--	--	3	14
temperature		Range	64	85	73.571	6.572	0.369	--	14
humidity		Range	65	96	81.917	11.156	-0.260	--	12
windy									
play									

Data Audit of [outlook temperature humidity windy play]

Complete fields (%): 0.6 Complete records (%): 157142

Field	Type	Outliers	Extremes	Action
outlook	Set	--	--	--
temperature	Range	0	0	None
humidity	Range	0	0	None
windy	Set	--	--	--
play	Flag	--	--	--

% Complete		
	% Complete	Valid Records
100	100	14
100	100	14
85.714	92.857	13
92.857	100	14
100		

Audit **Quality** Annotations

OK

Data Understanding

- Can you find errors in the dataset?

Data Cleaning

- Replace missing value (blanks) with specified value

The screenshot illustrates the process of cleaning data in Orange3. It features two main windows: 'data-play.csv' and 'humidity Values'.

data-play.csv window:

- Buttons: Preview, Refresh, Read Values, Clear Values, Clear All.
- Table with columns: Field, Type, Values, Missing, Check.
- Fields listed: outlook (Set), temperature (Range), humidity (Range), windy (Set), play (Flag).
- Buttons at the bottom: File, Data, Filter, Types, Apply, OK, Cancel.

humidity Values window:

- Buttons: OK, Cancel, Help.
- Fields: Type (Range), Storage (Integer), Model Field... (empty).
- Values section: Read from data, Pass, Specify values and la... (selected).
- Lower: 65, Upper: 96.
- Check values section: None, Nullify, Coerce (selected), Discard, Warn, Abort.
- Description: (empty).

Annotations:

- Red box around the 'Specify...' option in the 'data-play.csv' window's 'Types' dropdown.
- Red box around the 'Coerce' option in the 'humidity Values' window's 'Check values' dropdown.
- Red box around the 'OK' button in the 'humidity Values' window.

Data Cleaning

Table (5 fields, 14 records)

	outlook	temperature	humidity
1	sunny	85	85
2	sunny	80	90
3	overcast	83	86
4	rainy	70	96
5	rainy	68	\$null\$
6	rainy	65	70
7	overcast	64	65
8	sunny	72	95
9	sunny	69	70
10	rainy	75	\$null\$
11	sunny	75	70
12	overcast	72	90
13	overcast	81	75
14	rainy	71	91

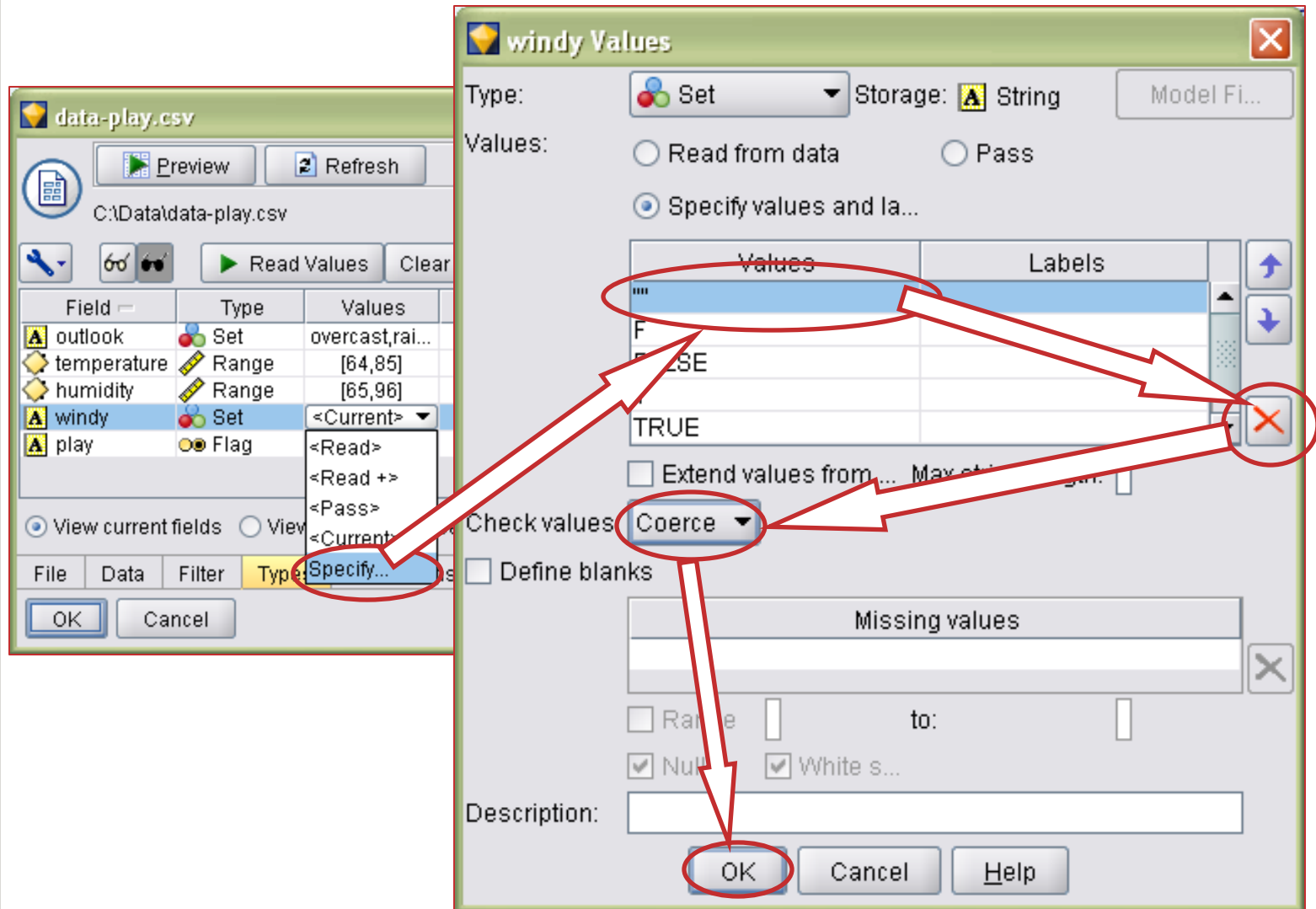
Table (5 fields, 14 records) #1

	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	T	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	F	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	F	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80		yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	T	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no

Table Annotations

OK

Data Cleaning



Data Cleaning

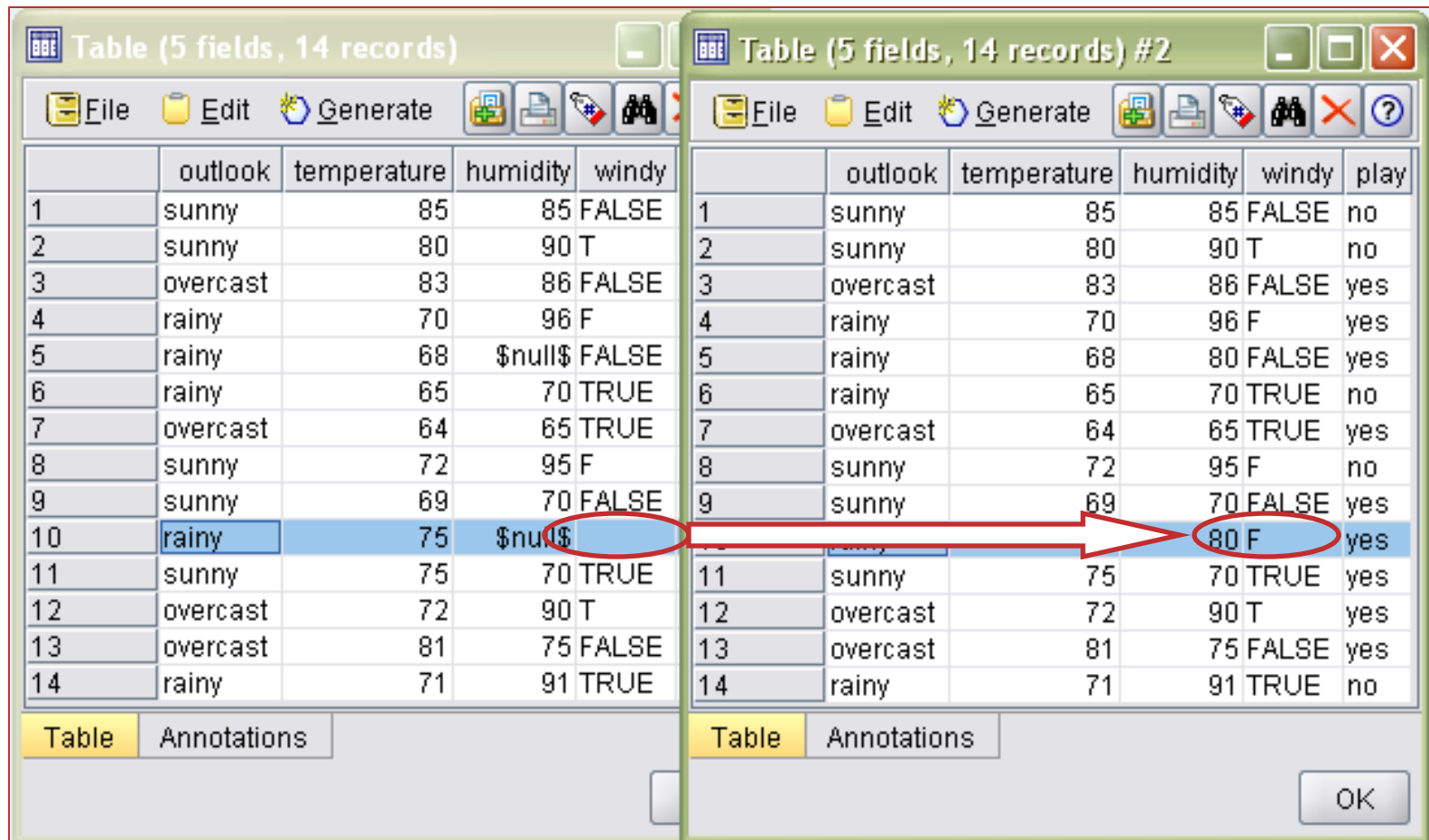


Table (5 fields, 14 records)

	outlook	temperature	humidity	windy
1	sunny	85	85	FALSE
2	sunny	80	90	T
3	overcast	83	86	FALSE
4	rainy	70	96	F
5	rainy	68	\$null\$	FALSE
6	rainy	65	70	TRUE
7	overcast	64	65	TRUE
8	sunny	72	95	F
9	sunny	69	70	FALSE
10	rainy	75	\$null\$	
11	sunny	75	70	TRUE
12	overcast	72	90	T
13	overcast	81	75	FALSE
14	rainy	71	91	TRUE

Table (5 fields, 14 records) #2

	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	T	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	F	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	F	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80 F		yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	T	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no

Table Annotations

Table Annotations

OK

Data Cleaning

The screenshot shows the 'Reclassify' dialog box in a data tool. The dialog is configured as follows:

- Mode:** ☒ Single ☐ Multiple
- Reclassify into:** ☐ New field ☒ Existing field
- Reclassify field:** windy
- New field name:** Reclassify1
- Reclassify values:**
 - ☒ Get ☒ Copy ☐ Clear new ☐ Auto...
- | Original value | New value |
|----------------|-----------|
| F | FALSE |
| FALSE | FALSE |
| T | TRUE |
| TRUE | TRUE |
- For unspecified values use:** ☒ Original ... ☐ Default ... undef
- Buttons:** OK, Cancel, Apply, Reset

Red arrows indicate the flow of configuration: from the 'Reclassify' tool icon to the dialog, from 'Existing field' to the 'windy' field, from 'Copy' to the value mapping table, and from 'Original ...' to the 'OK' button.

Data Cleaning

Table (5 fields, 14 records)

File Edit Generate

	outlook	temperature	humidity	windy
1	sunny	85	85	FALSE
2	sunny	80	90	T
3	overcast	83	86	FALSE
4	rainy	70	96	F
5	rainy	68	\$null\$	FALSE
6	rainy	65	70	TRUE
7	overcast	64	65	TRUE
8	sunny	72	95	F
9	sunny	69	70	FALSE
10	rainy	75	\$null\$	
11	sunny	75	70	TRUE
12	overcast	72	90	T
13	overcast	81	75	FALSE
14	rainy	71	91	TRUE

Table Annotations

Table (5 fields, 14 records) #3

File Edit Generate

	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes
12	overcast	72	90	TRUE	yes
13	overcast	81	75	FALSE	yes
14	rainy	71	91	TRUE	no

Table Annotations

OK

Data Transformation

Set “No. of bins” to “3” for
“temperature” and “humidity”

The screenshot illustrates the configuration of the Binning widget in Orange3. The workflow shows a 'data-play.csv' file being processed by a 'Reclassify' widget and then a 'Binning' widget. The Binning widget is configured with the following settings:

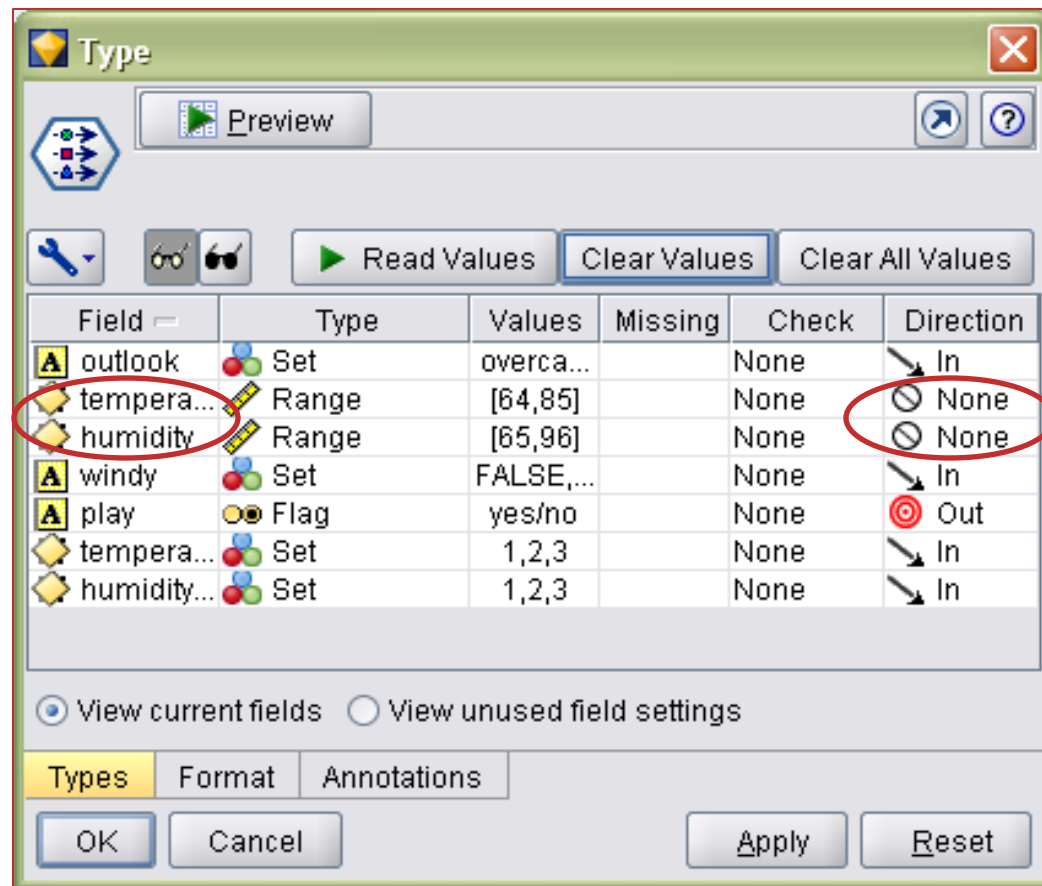
- Bin fields:** temperature, humidity
- Binning method:** Fixed-width
- Name extension:** _BIN
- Add as:** Suffix
- No. of bins:** 3
- Thresholds:** Always recompute

The output table, titled 'Table (5 fields, 14 records) #3', displays the results of the binning process. The table includes columns for the original features and their binned values.

	outlook	temperature	humidity	windy	play	play	temperature_BIN	humidity_BIN
1	sunny	85	85	FALSE	no	no	3	2
2	sunny	80	90	TRUE	no	no	3	3
3	overcast	83	86	FALSE	yes	yes	3	3
4	rainy	70	95	FALSE	yes	yes	1	3
5	rainy	68	80	FALSE	yes	yes	1	2
6	rainy	65	70	TRUE	no	no	1	1
7	overcast	64	65	TRUE	yes	yes	1	1
8	sunny	72	95	FALSE	no	no	2	3
9	sunny	69	70	FALSE	yes	yes	1	1
10	rainy	75	80	FALSE	yes	yes	2	2
11	sunny	75	70	TRUE	yes	yes	2	1
12	overcast	72	90	TRUE	yes	yes	2	3
13	overcast	81	75	FALSE	yes	yes	3	1
14	rainy	71	91	TRUE	no	no	2	3

Ready to Build the Model

- Use a “Type” node to refresh the memory and set the following



Decision Tree – C5.0

- Add the node, “C5.0” under “Modeling”
- Right click the node and “Execute”

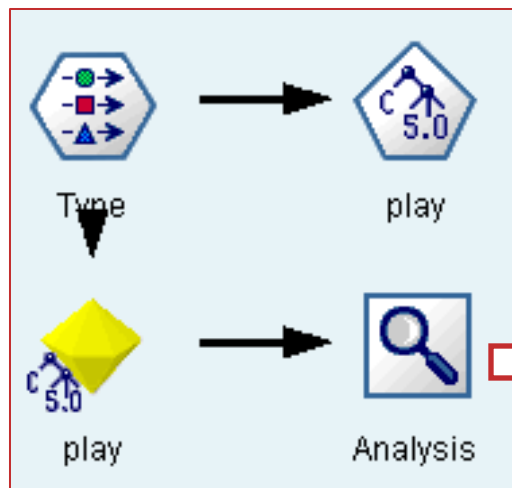
The screenshot displays the Orange3 software interface. The main window shows a decision tree model for the 'play' variable. The tree structure is as follows:

- Node 0 (Root):
 - no (35.714, 5) → Node 1
 - yes (64.286, 9) → Node 2
- Node 1 (overcast):
 - no (0.000, 0) → Node 3
 - yes (100.000, 4) → Node 4
- Node 2 (rainy):
 - no (40.000, 2) → Node 5
 - yes (60.000, 3) → Node 6
- Node 3 (FALSE):
 - no (0.000, 0) → Node 7
 - yes (100.000, 3) → Node 8
- Node 4 (TRUE):
 - no (100.000, 2) → Node 9
 - yes (0.000, 0) → Node 10
- Node 5 (humidity_BIN):
 - 1.000 → Node 11
 - 2.000; 3.000 → Node 12
- Node 6 (humidity_BIN):
 - 1.000 → Node 13
 - 2.000; 3.000 → Node 14
- Node 7 (humidity_BIN):
 - 1.000 → Node 15
 - 2.000; 3.000 → Node 16

The right side of the image shows the 'Modeling' palette with the 'C5.0' node highlighted by a red circle. A red arrow points from the 'C5.0' node in the palette to the 'C5.0' node in the main window, indicating the execution of the model.

Is the result good enough?

- Link the model, “play” with “Type”
- Add the “Analysis” node



The screenshot shows the 'Analysis of [play]' window. The window has a menu bar with 'File' and 'Edit', and buttons for 'Collapse All' and 'Expand All'. The main content area shows a tree view with the following structure:

- [-] Results for output field play
 - [-] Comparing \$C-play with play
 - Correct 14 100%
 - Wrong 0 0%
 - Total 14

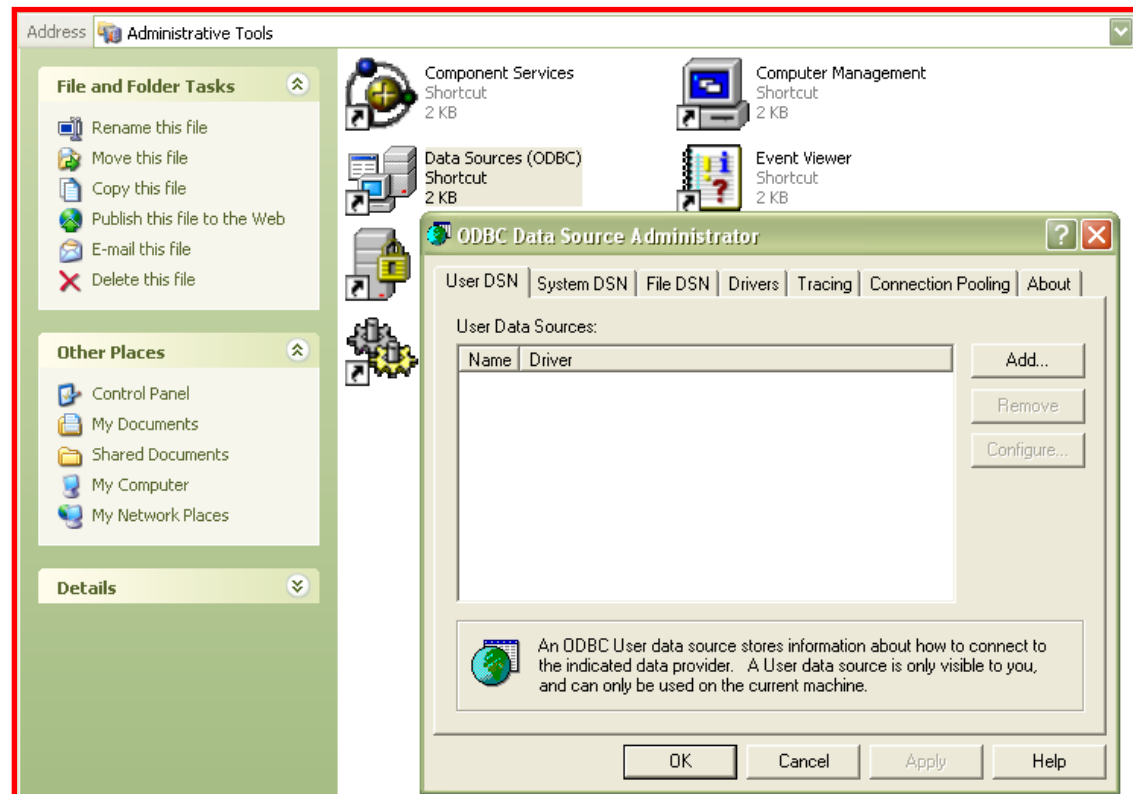
The table of results is circled in red:

Correct	14	100%
Wrong	0	0%
Total	14	

At the bottom of the window, there are tabs for 'Analysis' and 'Annotations', and an 'OK' button.

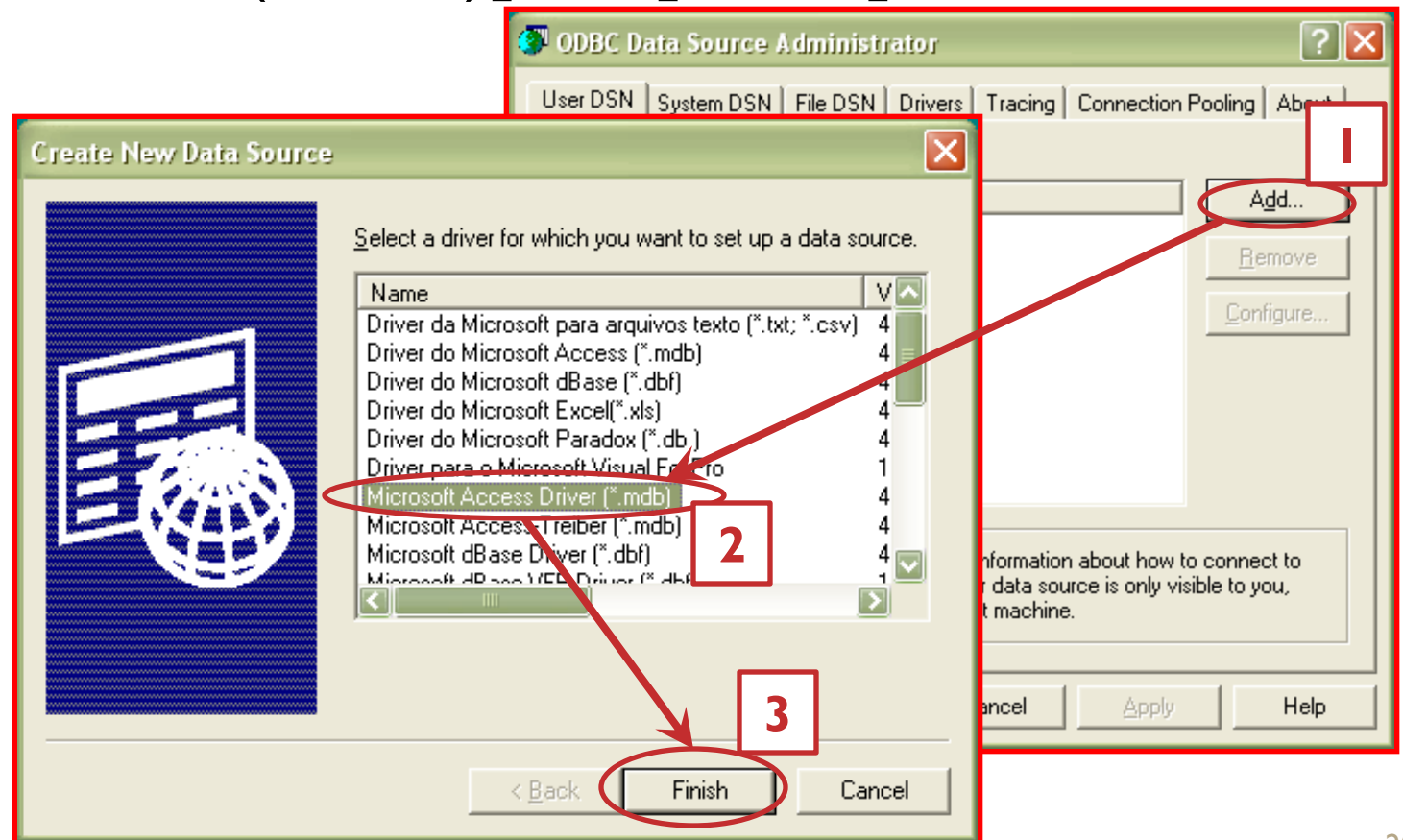
Another Example – data-Lab2.mdb

- To connect a database in PASW, a data source must be prepared first
- To create a data source, [Control Panel] → [Administrative Tools] → [Data Sources (ODBC)]



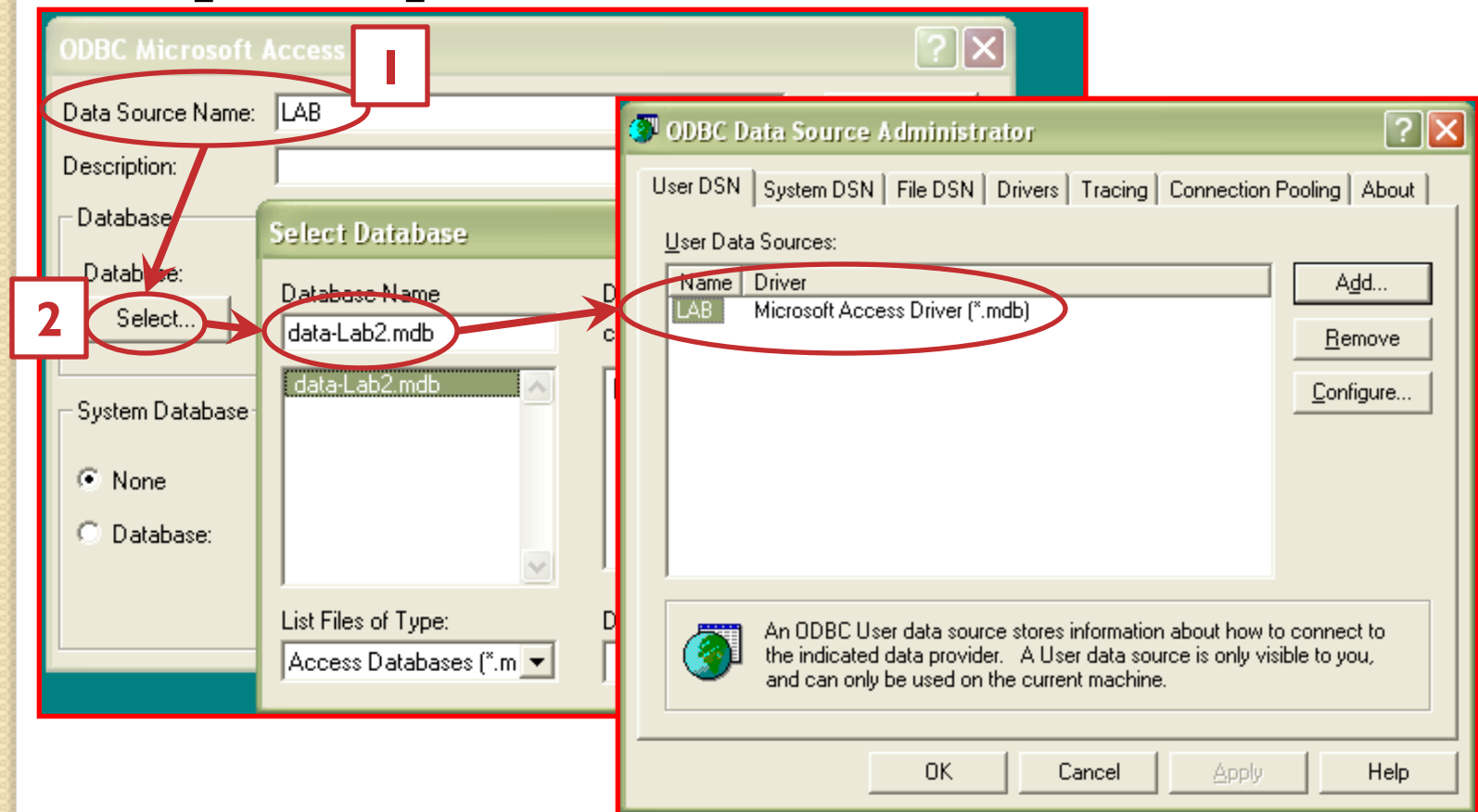
Create Data Source

- Click [Add] → Choose [Microsoft Access Driver (*.mdb)] → [Finish]



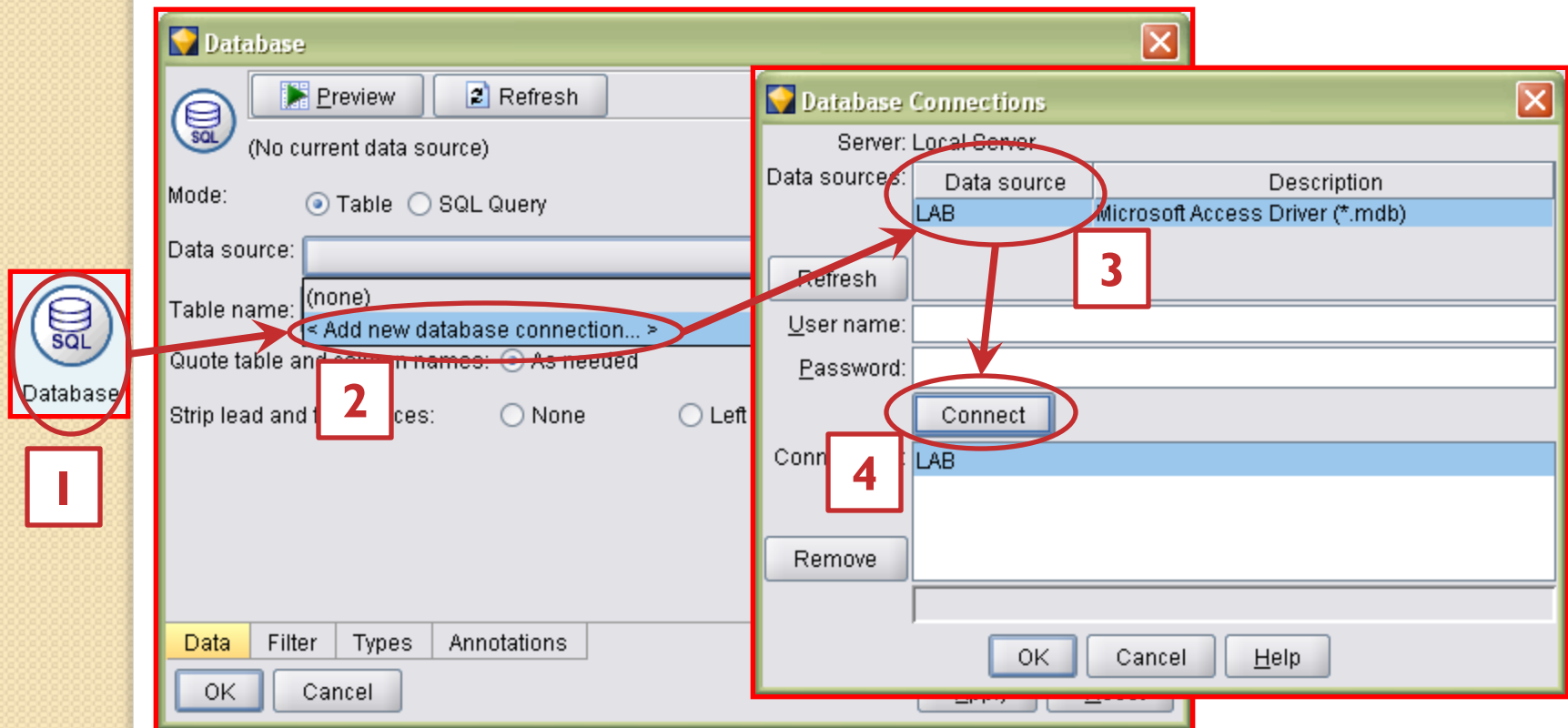
Create Data Source

- Name the [Data Source Name] → [Select] → Local the database file

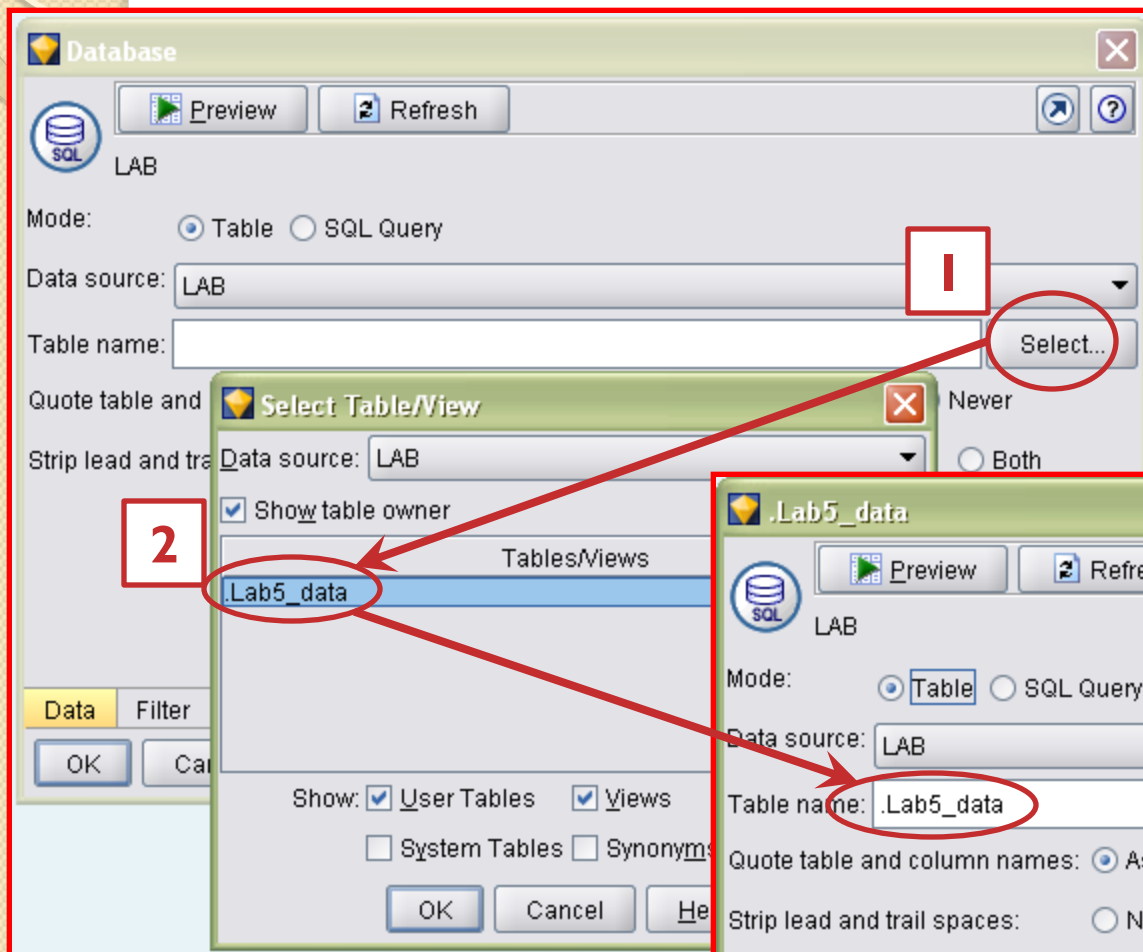


Import Data

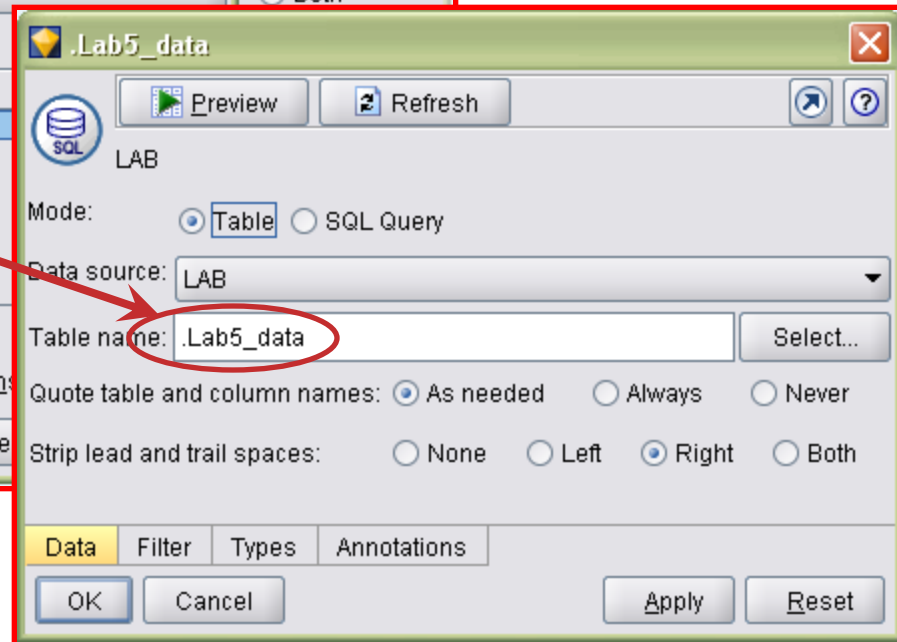
- Use [Database] icon
- Choose [<Add new database connection ...>]
- Select the [Data source] [LAB] → [Connect]



Import Data

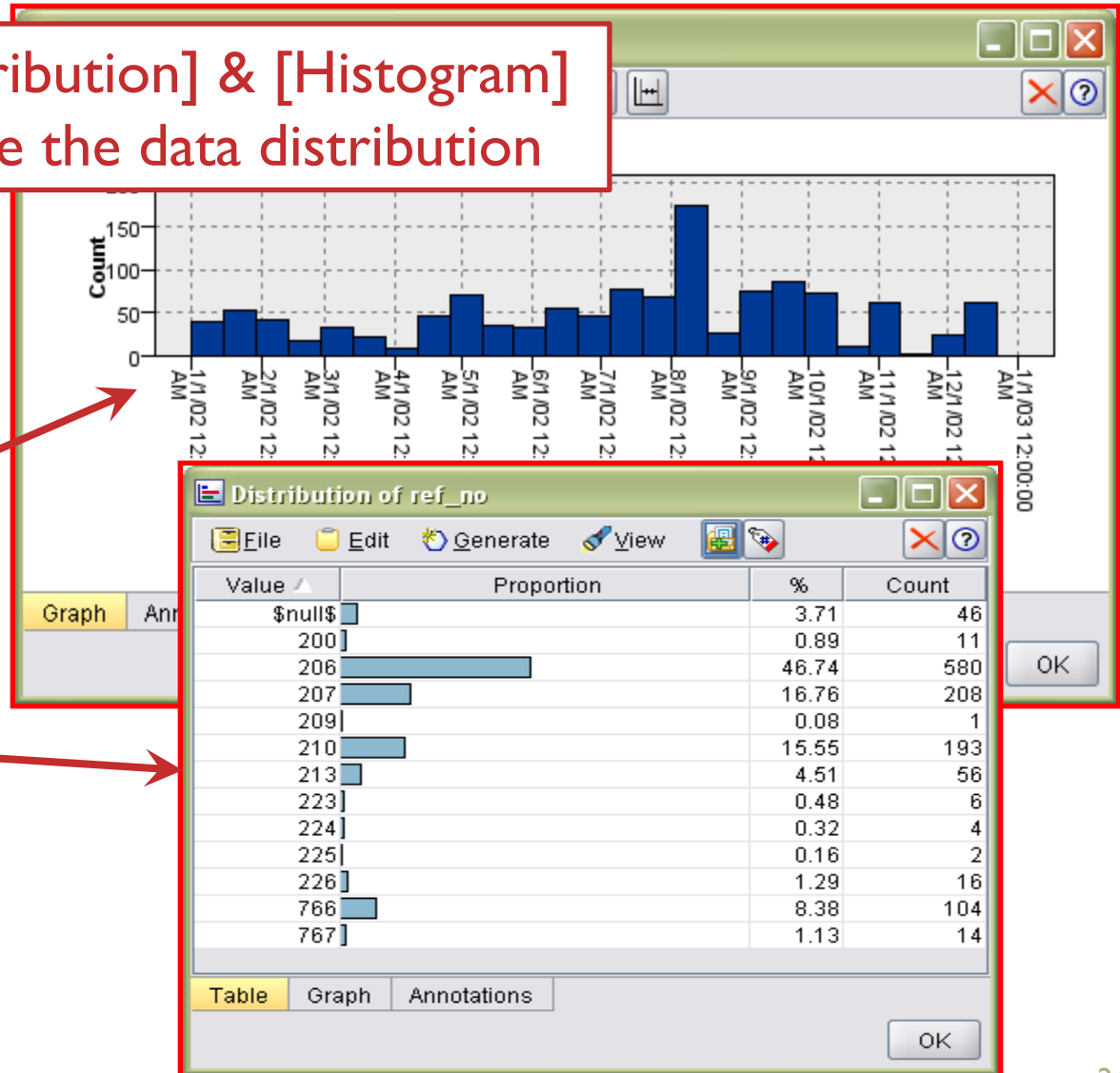
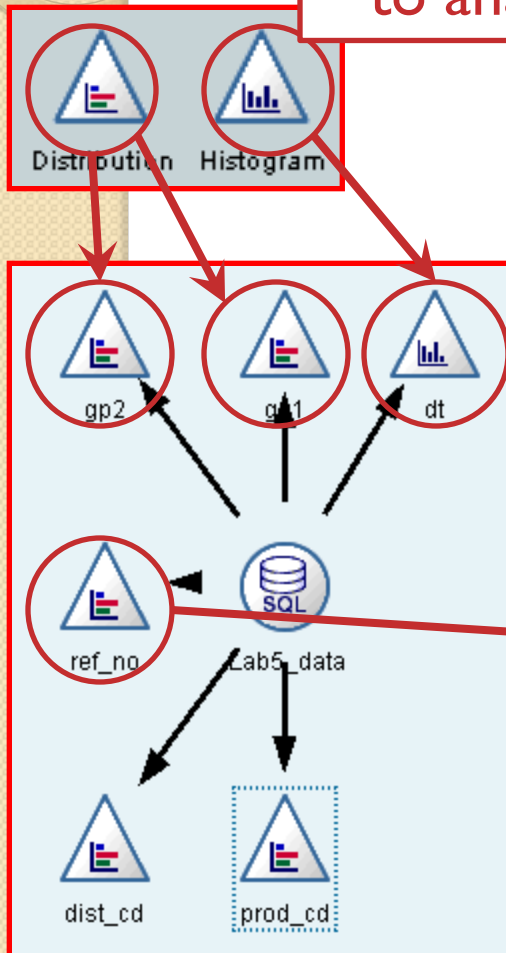


- [Select] → Choose the table
- Read values



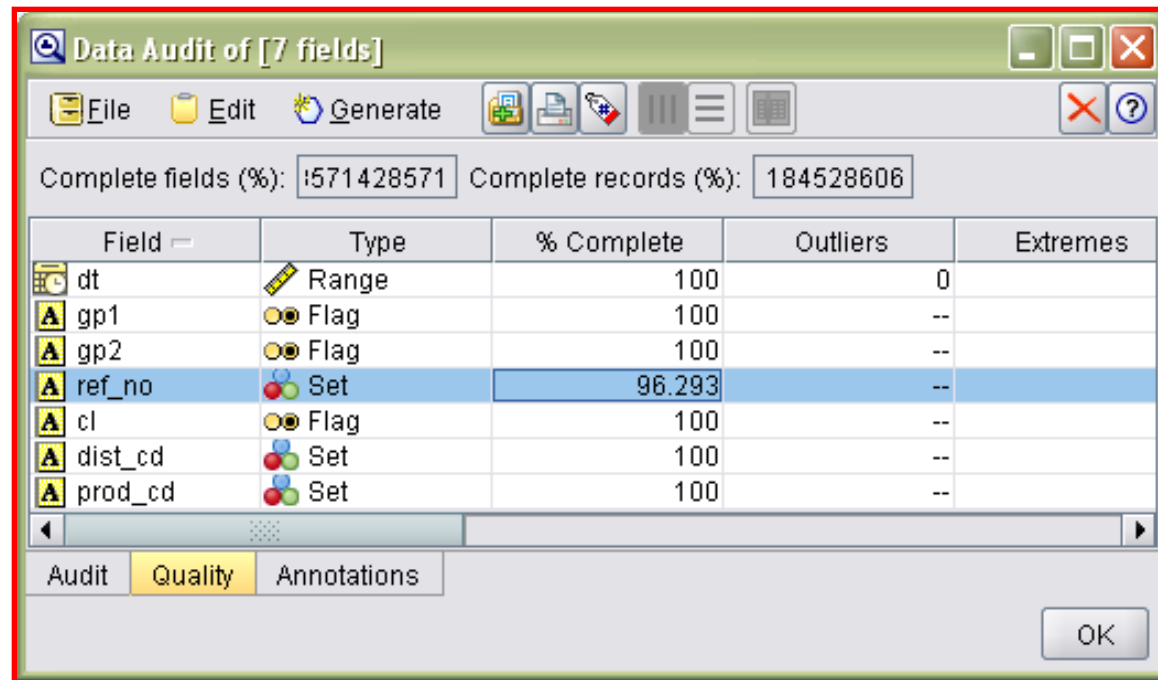
Data Understanding

Use [Distribution] & [Histogram] to analyze the data distribution



Data Understanding

- Use [Data Audit] to verify the [Quality] of data
- [ref_no] is not 100% complete
- Advise methods to fix the problem,
- i.e. [Data Cleaning]



Data Audit of [7 fields]

Complete fields (%): 1571428571 Complete records (%): 184528606

Field	Type	% Complete	Outliers	Extremes
dt	Range	100	0	
gp1	Flag	100	--	
gp2	Flag	100	--	
ref_no	Set	96.293	--	
cl	Flag	100	--	
dist_cd	Set	100	--	
prod_cd	Set	100	--	

Audit Quality Annotations

OK

Data Transformation

- Derive new attributes: [Day], [Weekday] & [Time]

The screenshot shows a data transformation workflow with three 'Derive' widgets. The 'Time' widget is highlighted with a dashed border. Below the widgets, a table displays the output data, with a red box highlighting the first three rows.

Day	Weekday	Time
16	Tuesday	16
16	Tuesday	16
26	Tuesday	18

Day: `datetime_day(dt)`
Weekday: `datetime_day_name(datetime_weekday(dt))`
Time: `datetime_hour(datetime_time(dt))`

Data Transformation

- Divide [Time] into intervals (binning)

The screenshot shows the 'TimeFrame' tool configuration window. The 'Derive as: Set' option is selected. The 'Field type' is set to 'Set'. The 'Default value' is 'default'. The 'Set field to' table is configured with the following conditions:

Set field to	If this condition is true
Night	Time < 6
Morning	Time >= 6 and Time < 12
Afternoon	Time >= 12 and Time < 18
Evening	Time >= 18 and Time < 24

The 'Preview' tab shows the resulting 'TimeFrame' field values for the input data. The data is as follows:

st_cd	prod_cd	Day	Weekday	Time	TimeFrame
E	P30	13	Tuesday	0	Night
N	P30	13	Tuesday	0	Night
E	P30	18	Sunday	19	Evening
N	P30	18	Sunday	19	Evening
N	P27	7	Monday	21	Evening
N	P30	23	Friday	8	Morning
E	P30	23	Friday	8	Morning
N	P30	12	Thursday	22	Evening
N	P30	12	Thursday	22	Evening
N	P27	2	Wednesday	1	Night
N	P27	15	Tuesday	12	Afternoon
N	P27	15	Tuesday	12	Afternoon
N	P39	18	Friday	18	Evening
N	P39	18	Friday	18	Evening
N	P30	19	Saturday	13	Afternoon
N	P30	19	Saturday	13	Afternoon
N	P30	21	Monday	17	Afternoon
N	P30	21	Monday	17	Afternoon
N	P28	23	Wednesday	20	Evening
N	P27	2	Wednesday	1	Night

Data Type Re-Define

- Use a [Type] node to change the data type of [Day] and [Time] from [Range] to [Set]

The image displays two side-by-side screenshots of the 'Type' dialog box, illustrating the process of re-defining data types for 'Day' and 'Time' fields.

Left Screenshot (Initial State):

Field	Type	Values	Missing
TID	Typeless		
dt	Range	[2002-01-0...	
gp1	Flag	N/N	
gp2	Flag	"3"/"2"	
ref_no	Set	"200","206"...	
cl	Flag	"109"/"109"	
dist_cd	Set	CHG,CLK,...	
prod_cd	Set	P12,P13,P...	
Day	Range	[1,31]	
Weekday	Set	Friday,Mon...	
Time	Range	[0,23]	
TimeFra...	Set	Afternoon,E...	

Right Screenshot (Re-defined State):

Field	Type	Values	Missing	Check	Direction
TID	Typeless			None	None
dt	Range	[2002-01-0...		None	In
gp1	Flag	N/N		None	In
gp2	Flag	"3"/"2"		None	In
ref_no	Set	"200","206"...		None	In
cl	Flag	"109"/"109"		None	In
dist_cd	Set	CHG,CLK,...		None	In
prod_cd	Set	P12,P13,P...		None	In
Day	Set	1,2,3,4,5,6,...		None	In
Weekday	Set	Friday,Mon...		None	In
Time	Set	0,1,2,3,4,5,...		None	In
TimeFrame	Set	Afternoon,...		None	In

The 'Type' dialog box includes a 'Preview' button, 'Read Values', 'Clear Values', and 'Clear All Values' buttons. At the bottom, there are tabs for 'Types', 'Format', and 'Annotations', and buttons for 'OK', 'Cancel', 'Apply', and 'Reset'.

Data Cleaning - Question

- How do we know whether or not there are **duplicate records**?
- In PASW, under [Record Ops], there is an icon named [Aggregate]. It helps to detect and remove duplicate records.

Generate the Model

- Use [C5.0]

The screenshot displays the C5.0 model generation interface. On the left, a sidebar shows the 'Type' dropdown set to 'C5.0' and the 'prod_cd' target selected. The 'Inputs' list includes 'ref_no', 'dist_cd', 'Day', 'Weekday', and 'TimeFrame'. The 'Partition' and 'Splits' sections are empty. The 'Use weight field' checkbox is unchecked. The 'Fields' tab is active, showing a list of rules and their associated modes. The 'Model' tab is also visible, showing a 'Variable Importance' chart. The chart has a horizontal axis from 0.0 to 1.0. The 'ref_no' variable is highlighted in blue, indicating its importance. The 'dist_cd' variable is also shown. The 'View' dropdown is set to 'Variable Importance'. The 'Apply' and 'Reset' buttons are at the bottom right.

prod_cd

Type: C5.0

Target: prod_cd

Inputs: ref_no, dist_cd, Day, Weekday, TimeFrame

Partition:

Splits:

☐ Use weight field

Fields | Model | Cost

OK | Execute

prod_cd

File | Generate | View | Preview

1 | 2 | All | % | i

```
ref_no in ["200"] [Mode: P12] ⇒ P12
ref_no in ["206" "207" "223" "766"] [Mode: P27] ⇒ P27
ref_no in ["209" "213" "224" "225" "226" "767"] [Mode: P30] ⇒ P30
ref_no in ["210"] [Mode: P30]
  dist_cd in ["CHG" "CSW" "CWE" "CWN" "KBY"] [Mode: P30] ⇒ P30
  dist_cd in ["CLK"] [Mode: P27] ⇒ P27
  dist_cd in ["HHM" "KCG"] [Mode: P30] ⇒ P30
```

Variable Importance

ref_no

dist_cd

0.0 0.2 0.4 0.6 0.8 1.0

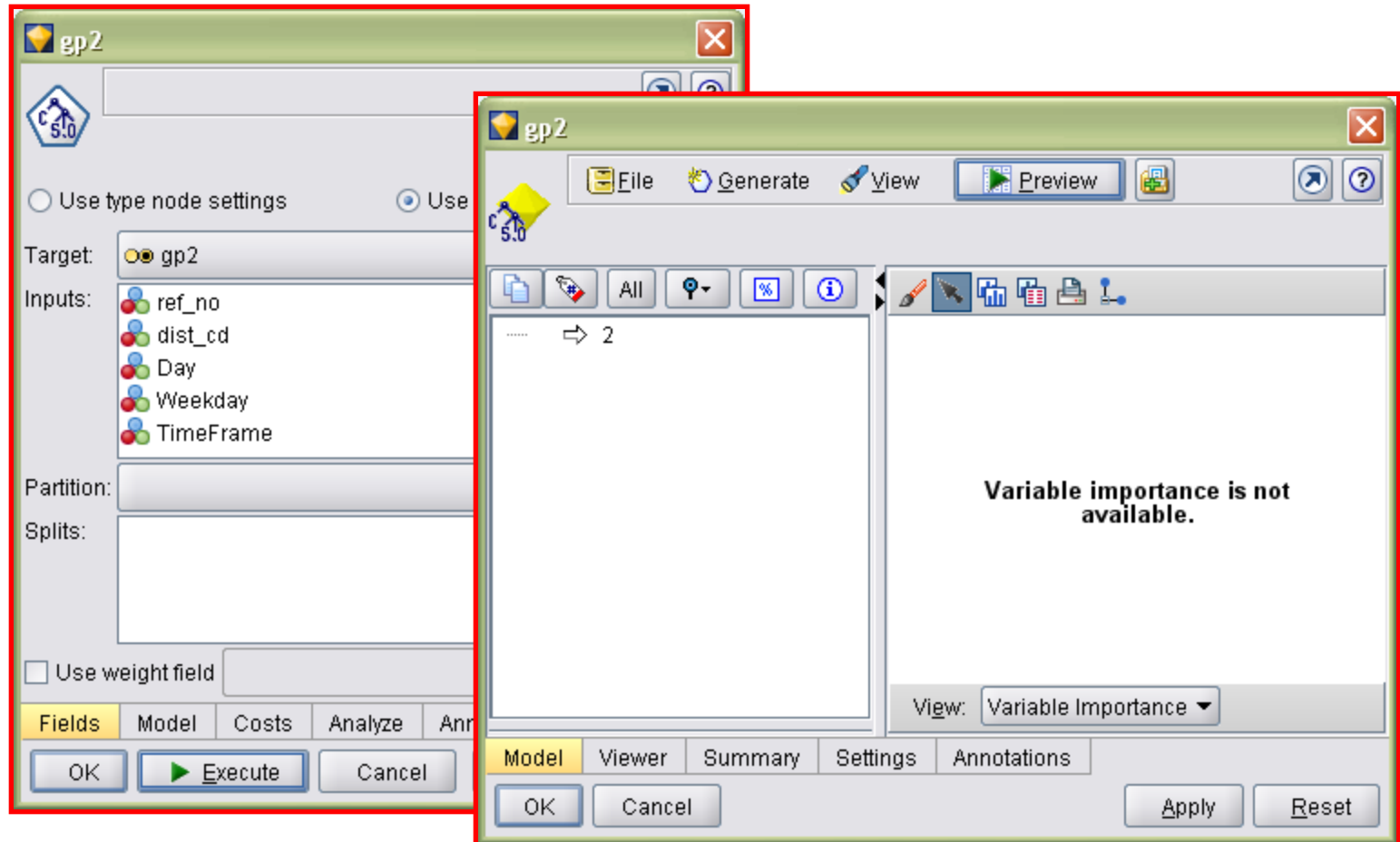
dist_cd ref_no

View: Variable Importance

Apply | Reset

Another Model with **No Rule**

- Use same inputs but target changes to [gp2]



Generate a **Balance Node**

- [Generate] → [Balance Node (boost)]

The screenshot illustrates the process of generating a balance node in Orange3. It shows three main components: a data table, a workflow canvas, and a model viewer.

1. Data Table (Distribution of gp2 #1): A table with two columns: 'Value' and an unlabeled column. The 'Value' column contains the numbers 2 and 3.

2. Workflow Canvas: A diagram showing the flow of data processing. It starts with a 'Type' node, followed by a '(generated)' node, and then a 'gp2' node. The 'gp2' node is labeled with a red '3'. A red box labeled '1' highlights the 'Generate' button in the top menu bar. A red box labeled '2' highlights the 'Balance Node (boost)' option in the 'Derive Node for' dropdown menu.

3. Model Viewer (gp2): A window showing the results of the model. It includes a 'Variable Importance' bar chart and a list of variables with their importance scores.

Variable Importance Chart:

Variable	Importance
Weekday	0.38
TimeFrame	0.35
Day	0.18
ref_no	0.08
dist_cd	0.05

Model Variables:

Variable	Mode	Score
TimeFrame	in ["Afternoon" "Evening" "Night"]	[Mode: 2] ⇒ 2
TimeFrame	in ["Morning"]	[Mode: 3]
Weekday	in ["Friday"]	[Mode: 3]
Day	in [1 2 9 16 19 21 22 23 25 26]	[Mode: 2] ⇒ 2
Day	in [3 4 5 6 7 10 11 12 13 14 15 17 18 20 24 27 28 29]	[Mode: 3]
Day	in [8]	[Mode: 3]
ref_no	in ["200" "209" "210" "213" "223" "224" "225" "226"]	[Mode: 3] ⇒ 3
ref_no	in ["206"]	[Mode: 3] ⇒ 3
ref_no	in ["207"]	[Mode: 2] ⇒ 2
Weekday	in ["Monday" "Sunday" "Thursday" "Tuesday" "Wednesday"]	[Mode: 3]
Weekday	in ["Saturday"]	[Mode: 3]
dist_cd	in ["CHG" "CWN" "KCG"]	[Mode: 2] ⇒ 2
dist_cd	in ["CLK"]	[Mode: 3]
Day	in [1 2 3 4 6 8 9 10 11 13 14 15 16 18 19 20 21 22]	[Mode: 3] ⇒ 3
Day	in [5]	[Mode: 3] ⇒ 3
Day	in [7 12 17 23 27]	[Mode: 2] ⇒ 2
dist_cd	in ["CSW" "CWE" "HHM" "KBY"]	[Mode: 3] ⇒ 3

Model can be formed after the [Balance] node has been added