Data Mining – LAB 2

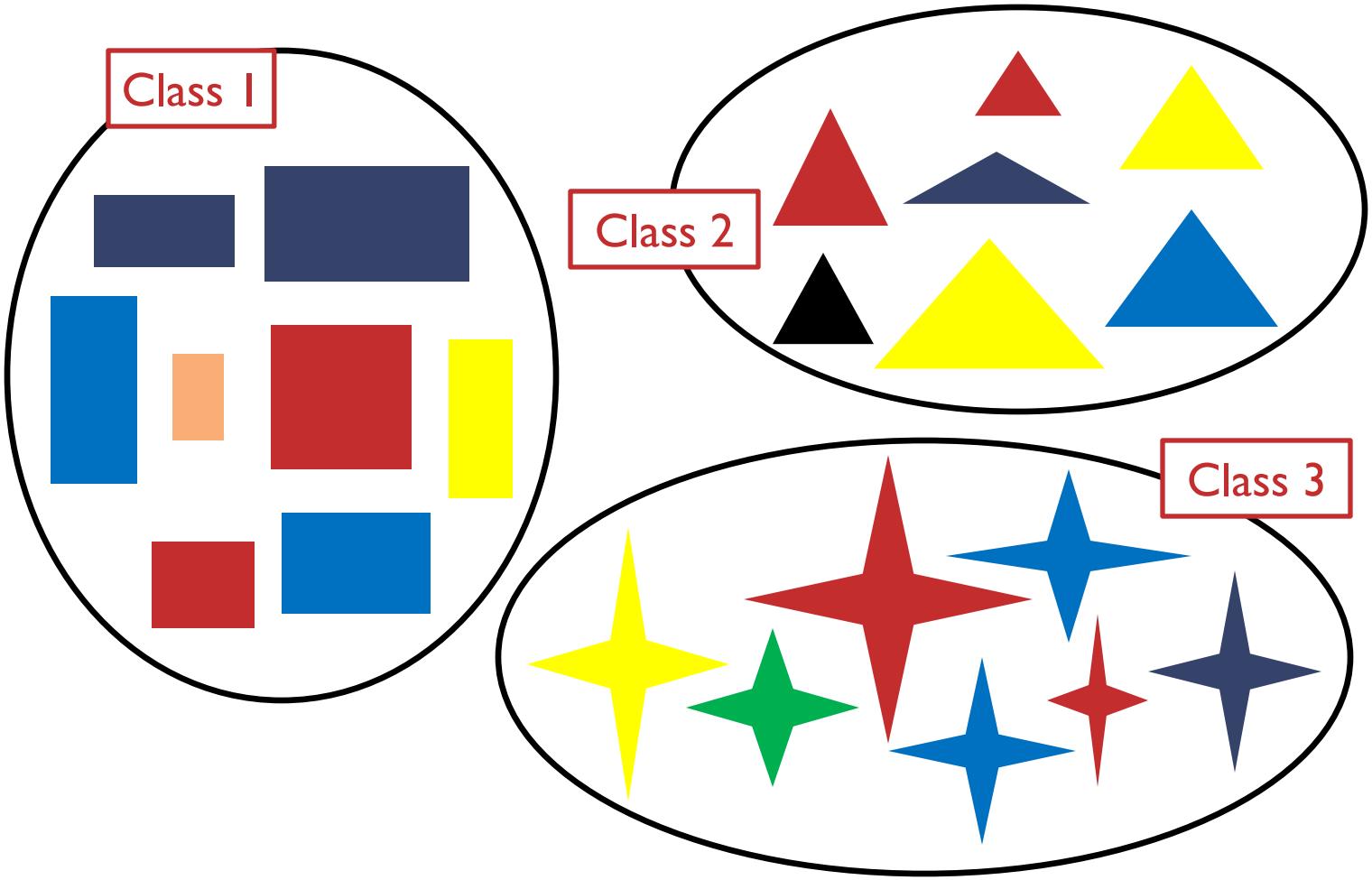
Classification – Decision Tree (C5.0)

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

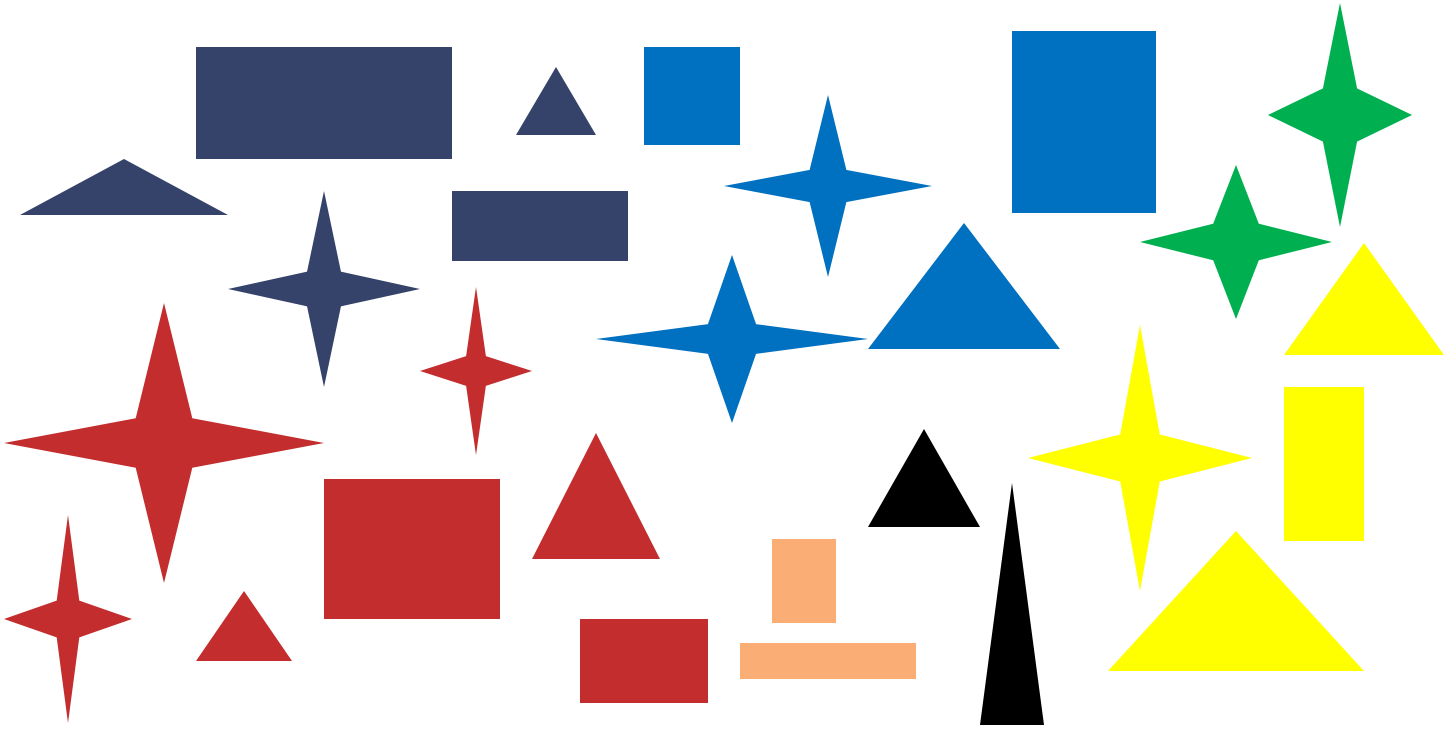
Classification

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



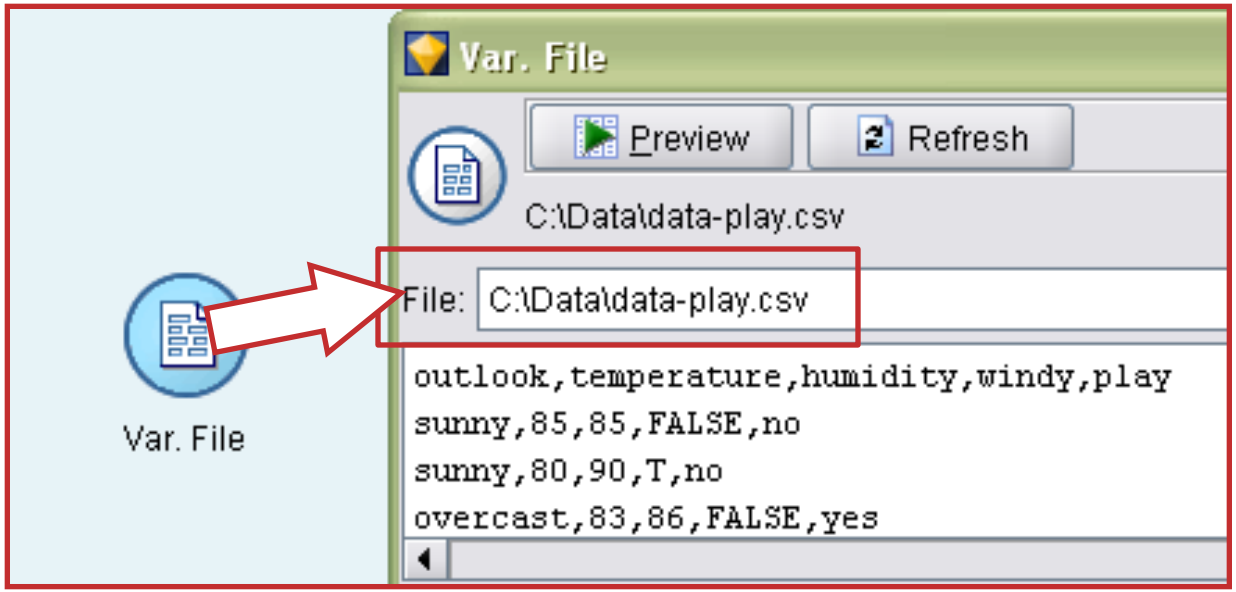
Clustering

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



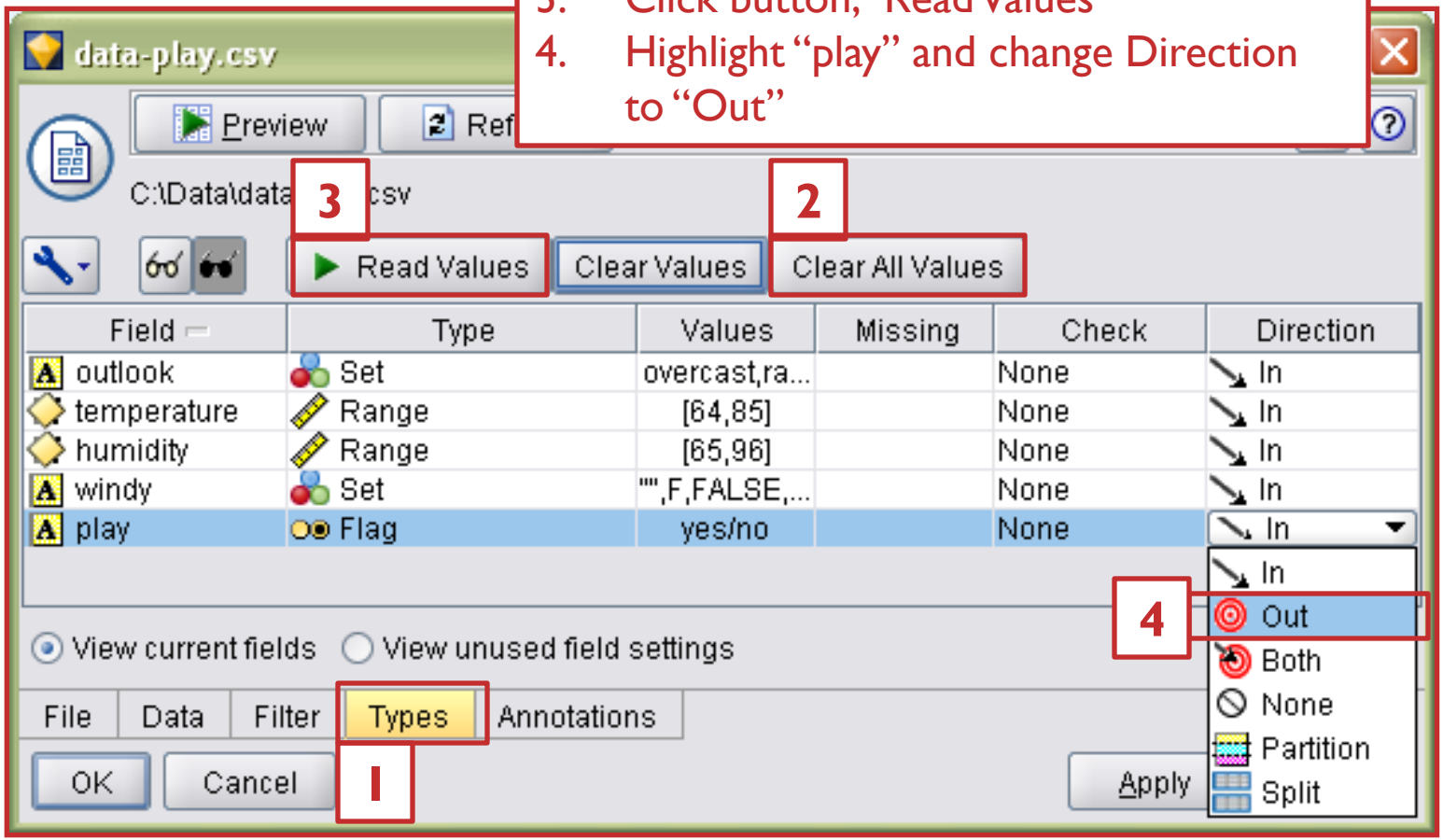
Import Data – data-play.csv

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

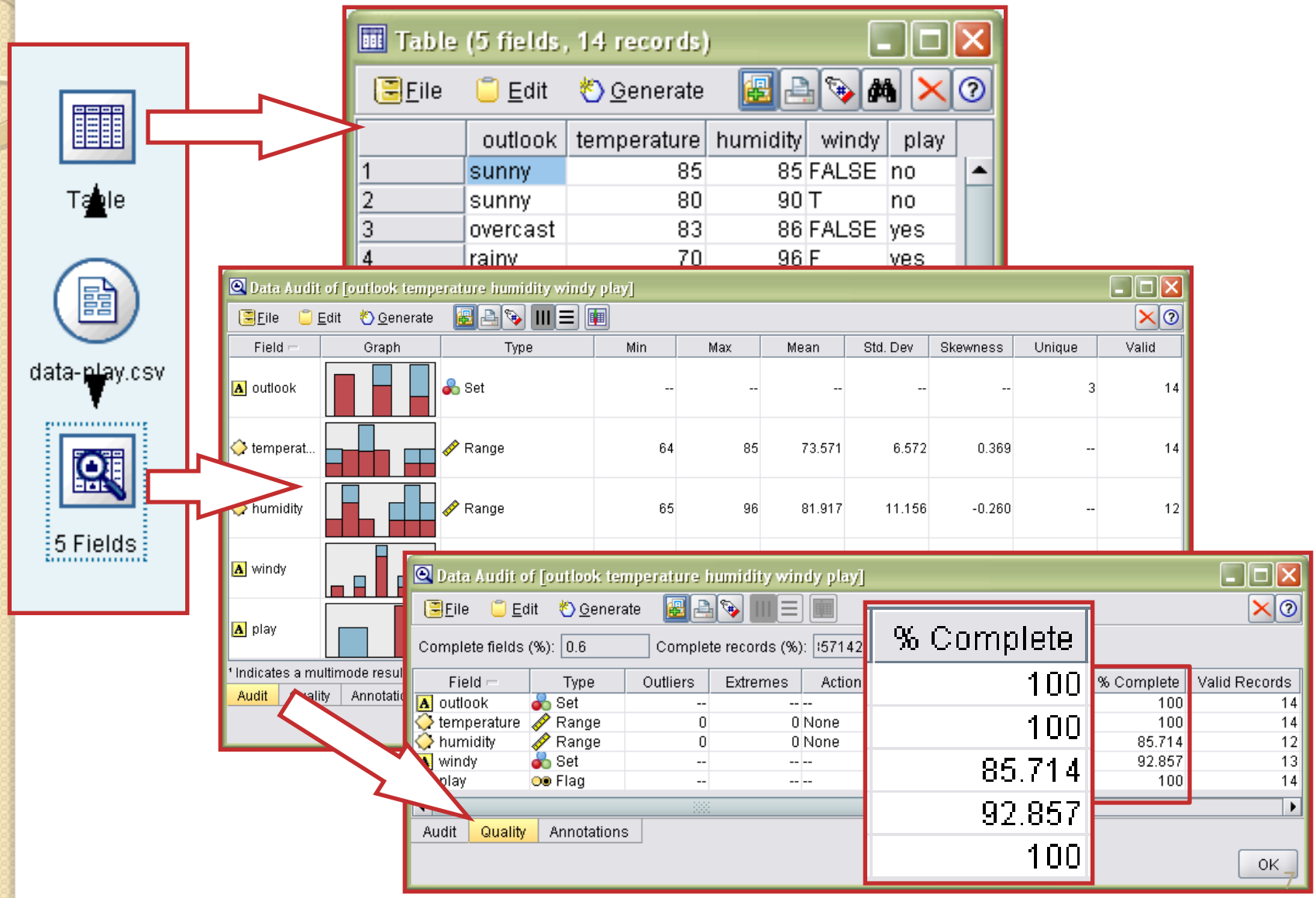


Set Target

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 Understanding

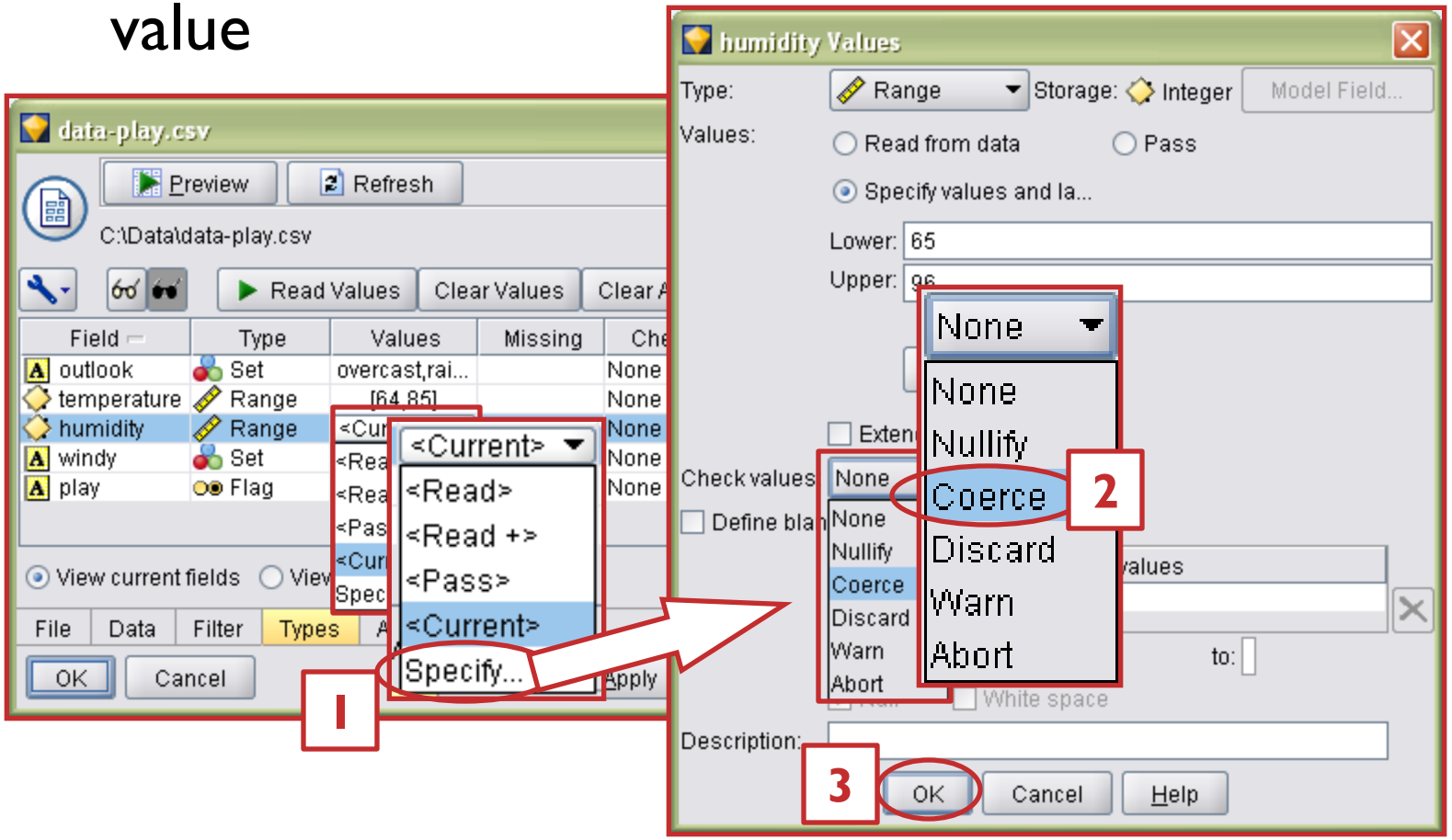


Data Understanding

- Can you find errors in the dataset?

Data Cleaning

- Replace missing value (blanks) with specified value



Data Cleaning

Table (5 fields, 14 records)

File Edit Generate

	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 Annotations

Table (5 fields, 14 records) #1

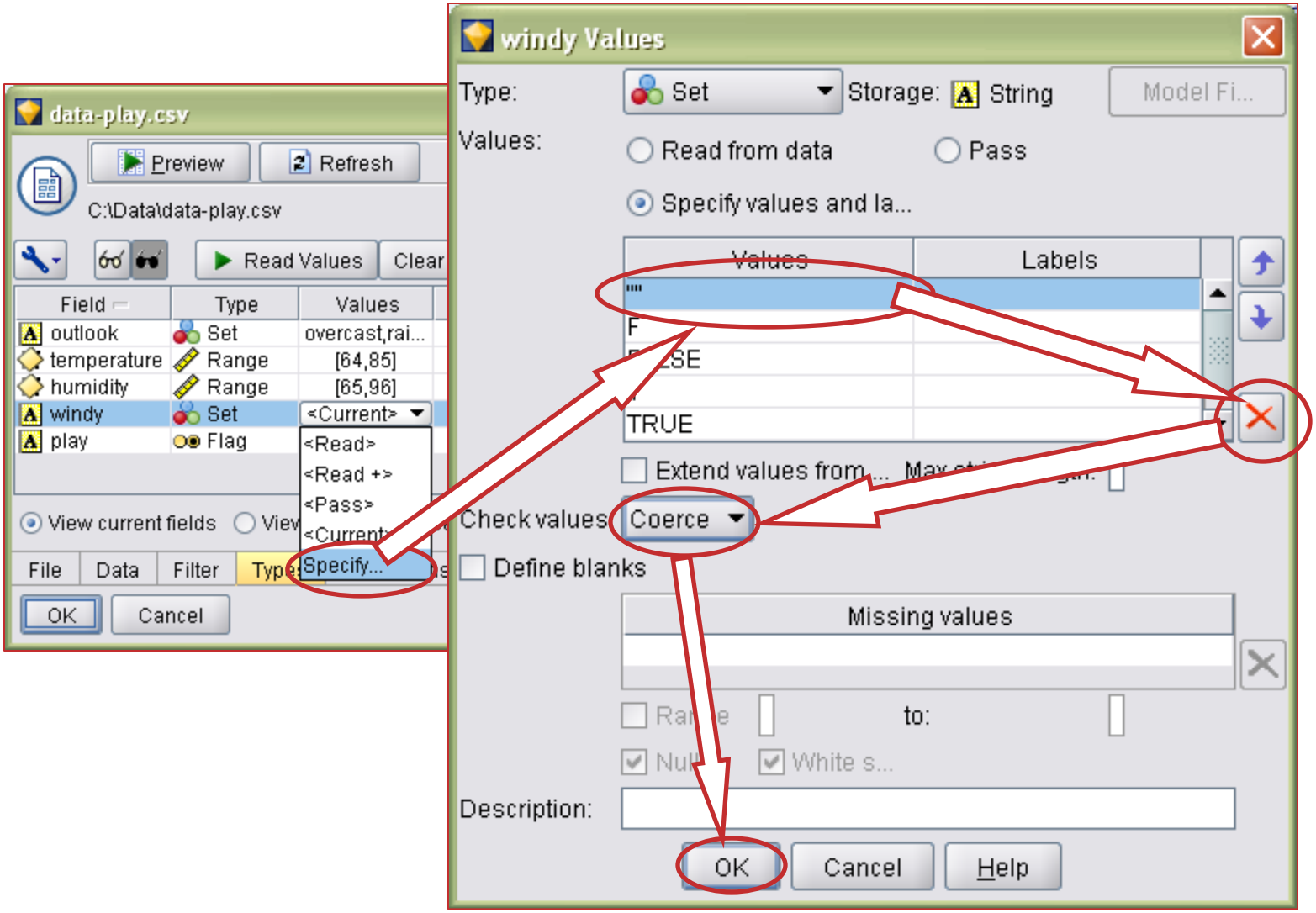
File Edit Generate

	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
[illegible]	[illegible]	[illegible]	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
[illegible]	[illegible]	[illegible]	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)

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) #2

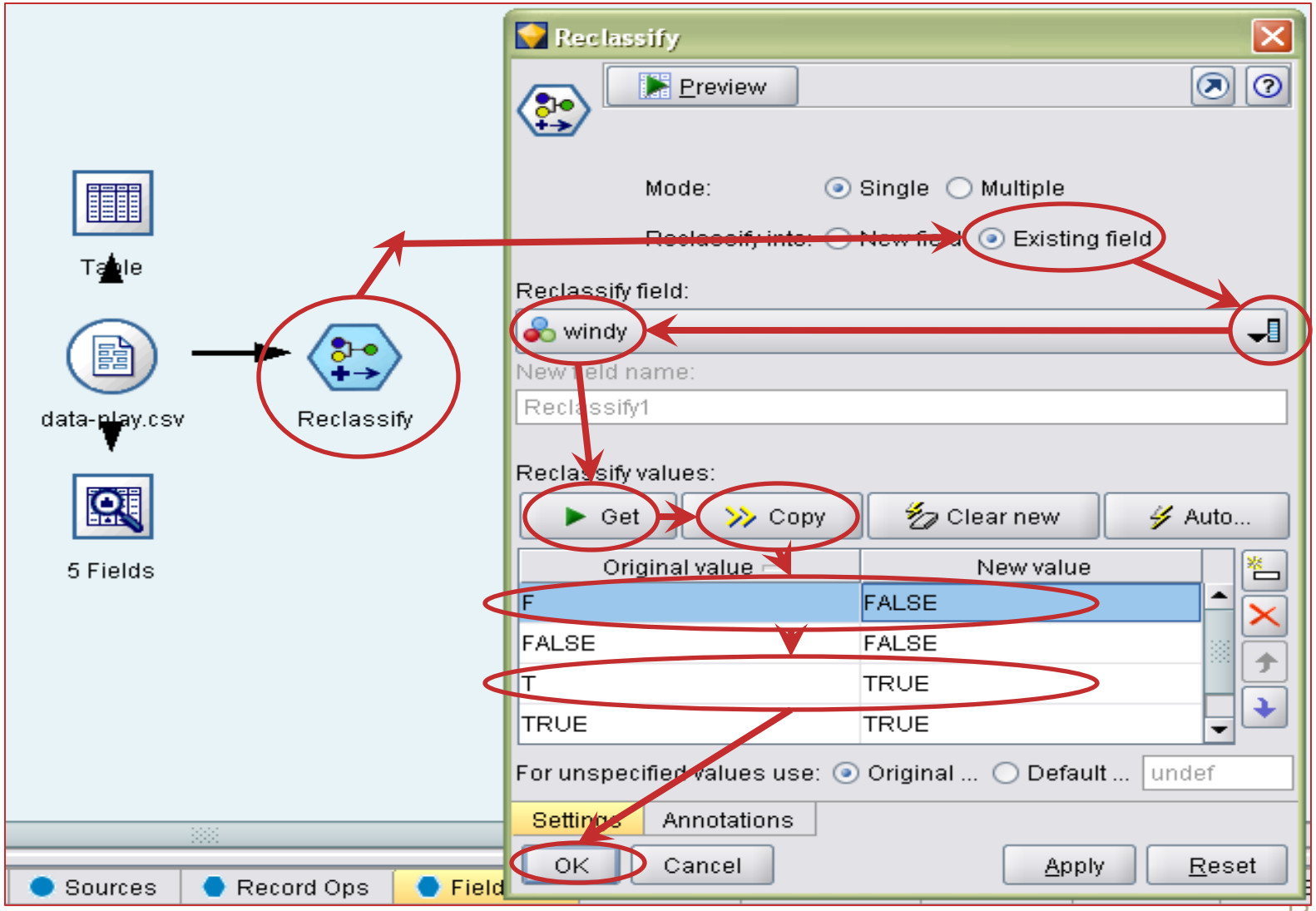
File Edit Generate

	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

OK

Data Cleaning



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

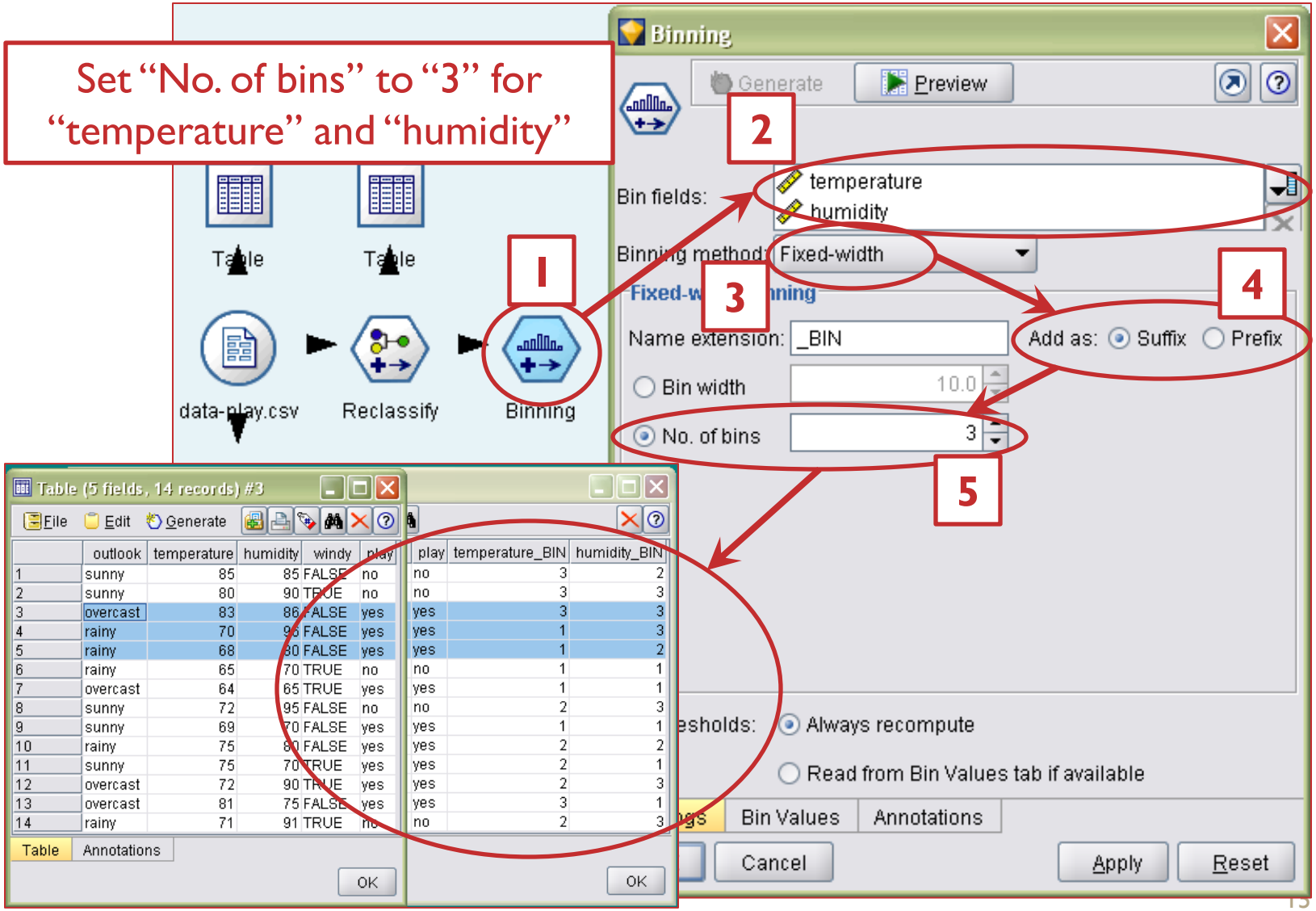
	outlook	temperature	humidity	windy	play
1	sunny	85	85	FALSE	no
2	sunny	80	90	TRUE	no
3	overcast	83	86	FALSE	yes
4	rainy	70	96	FALSE	yes
5	rainy	68	80	FALSE	yes
6	rainy	65	70	TRUE	no
7	overcast	64	65	TRUE	yes
8	sunny	72	95	FALSE	no
9	sunny	69	70	FALSE	yes
10	rainy	75	80	FALSE	yes
11	sunny	75	70	TRUE	yes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

Table Annotations

OK

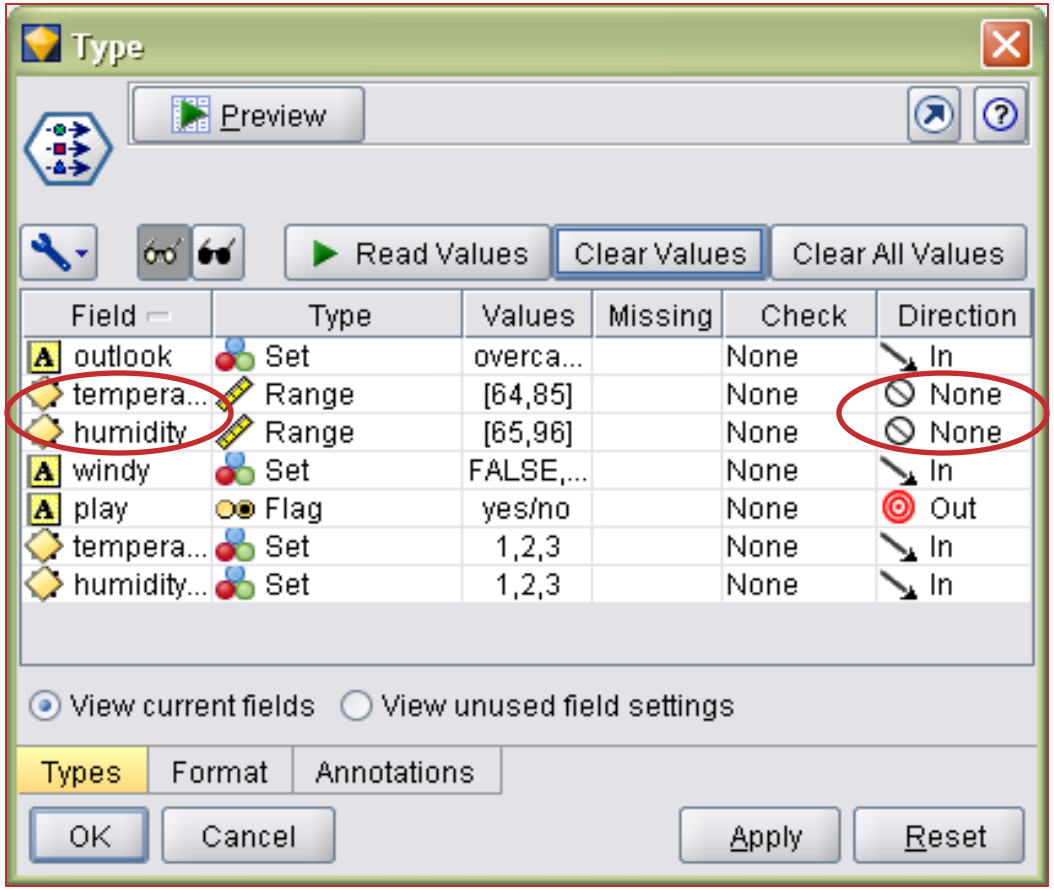
Data Transformation

Set "No. of bins" to "3" for "temperature" and "humidity"



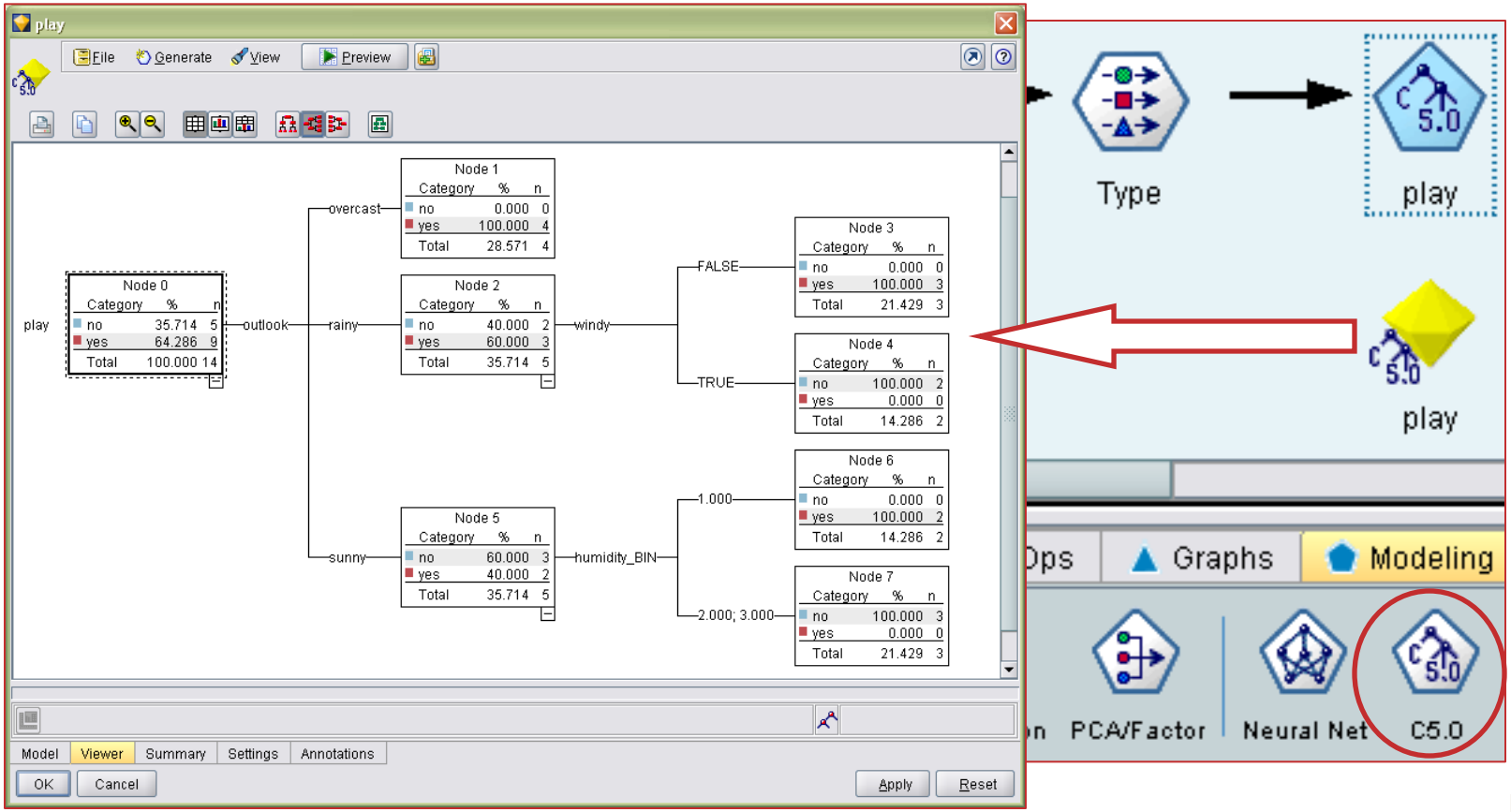
Ready to Build the Model

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



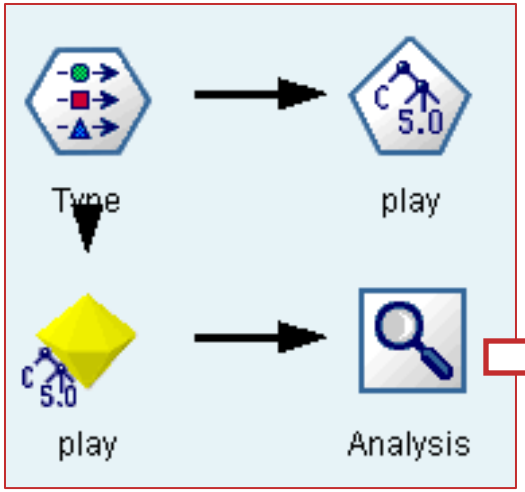
Decision Tree – C5.0

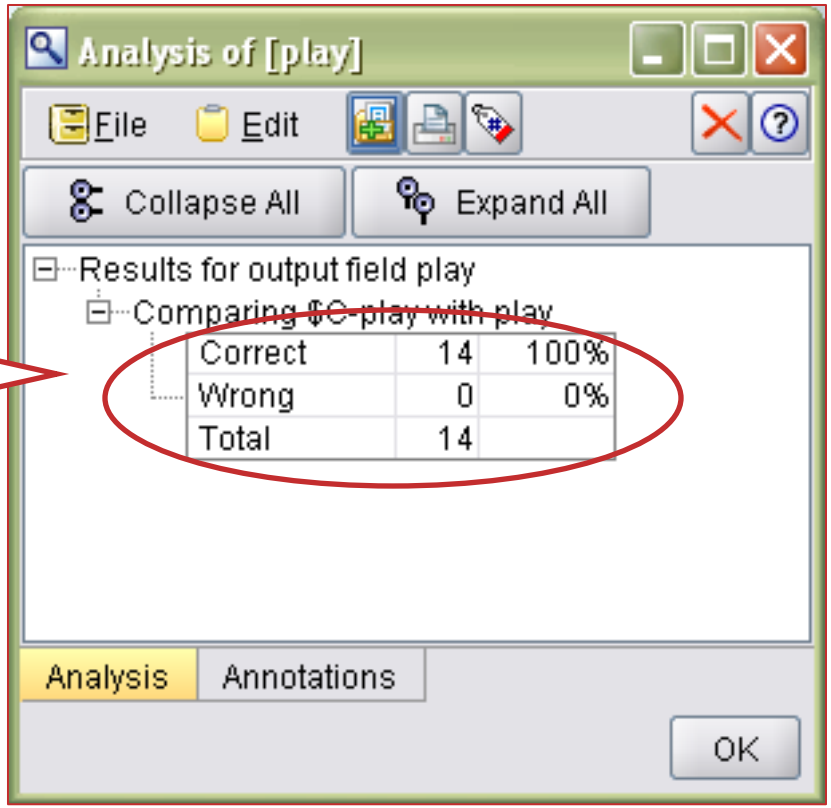
- Add the node, "C5.0" under "Modeling"
- Right click the node and "Execute"



Is the result good enough?

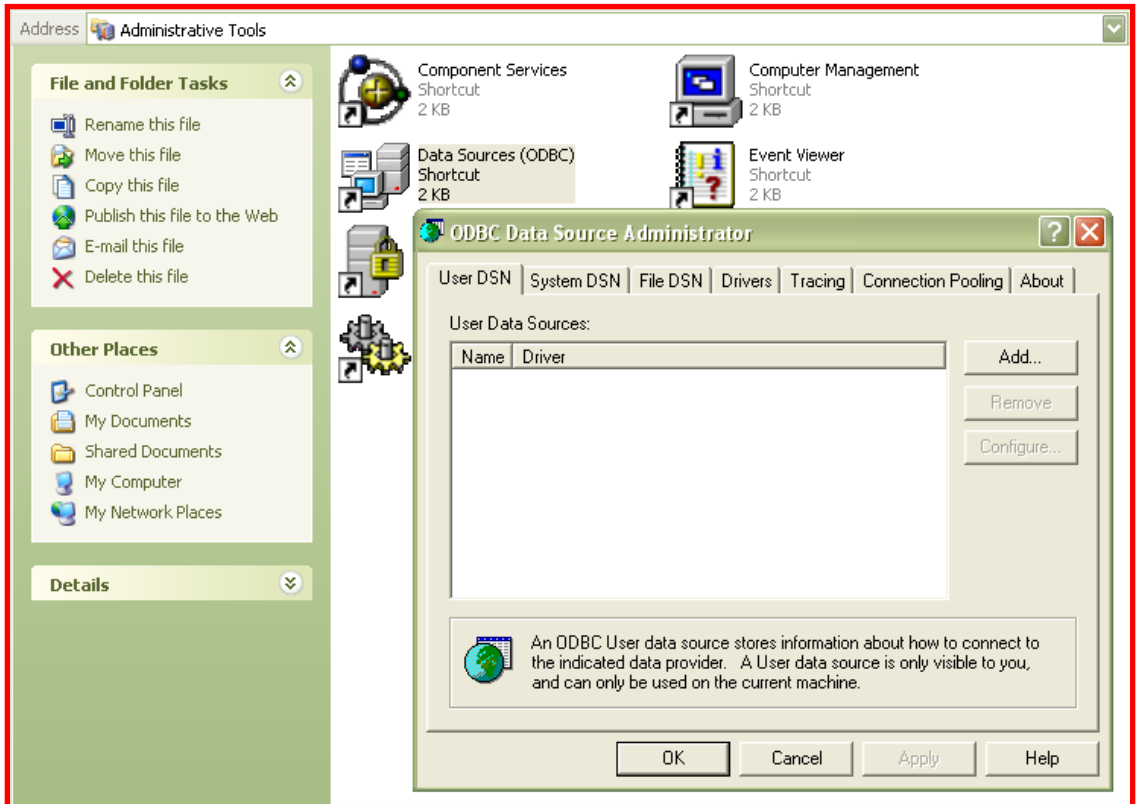
- Link the model, "play" with "Type"
- Add the "Analysis" node





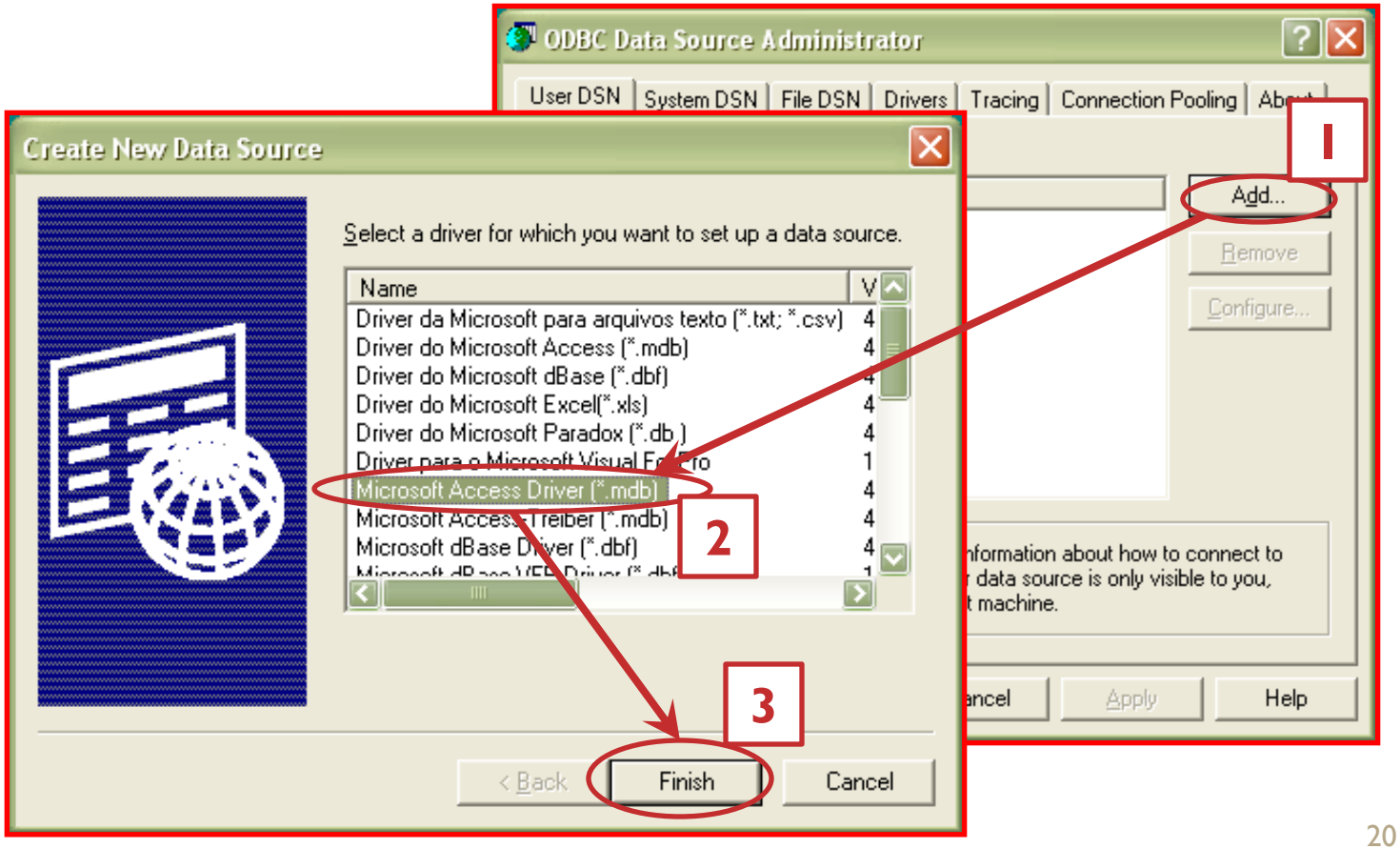
Another Example – data-Lab2.mdb

- 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)]



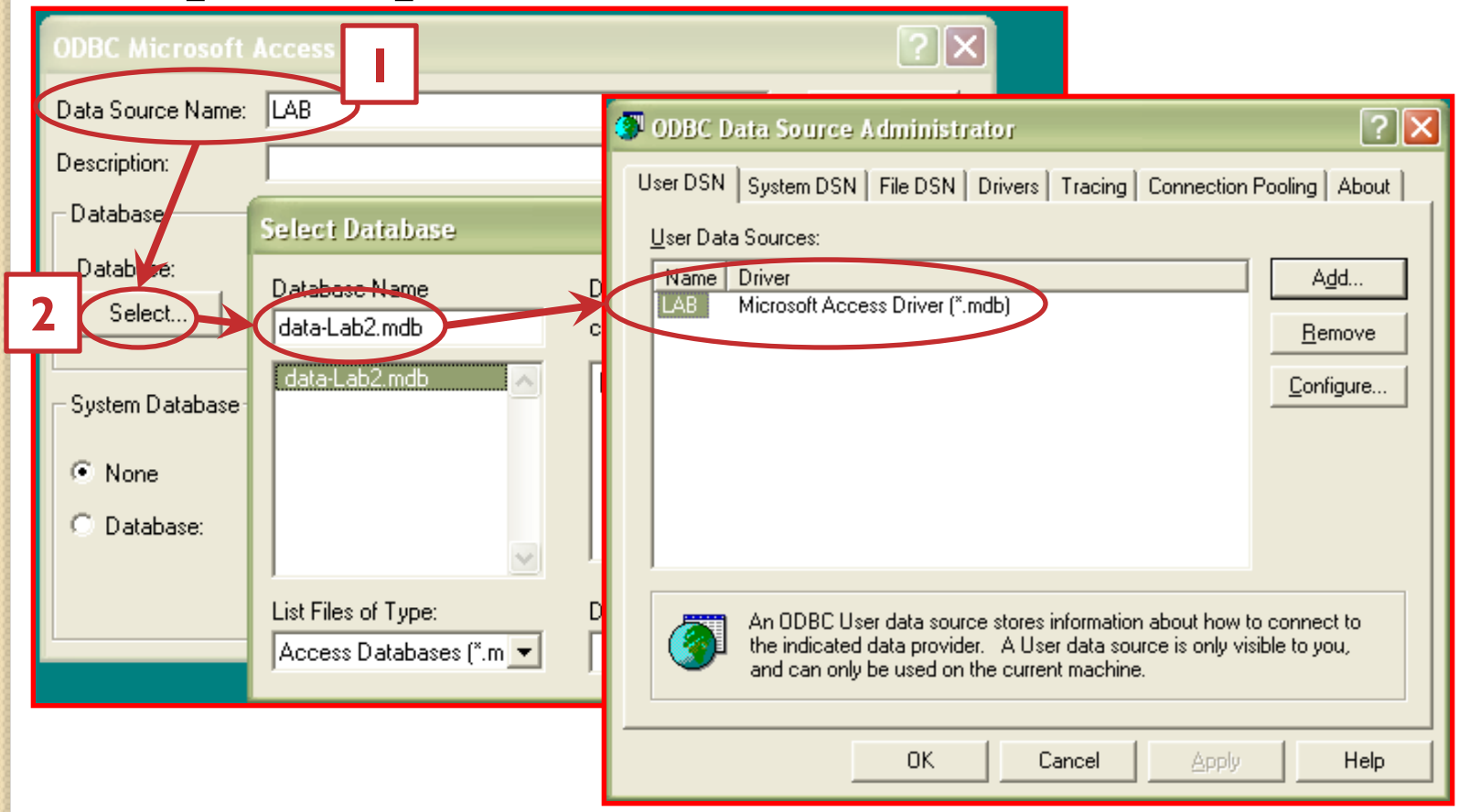
Create Data Source

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



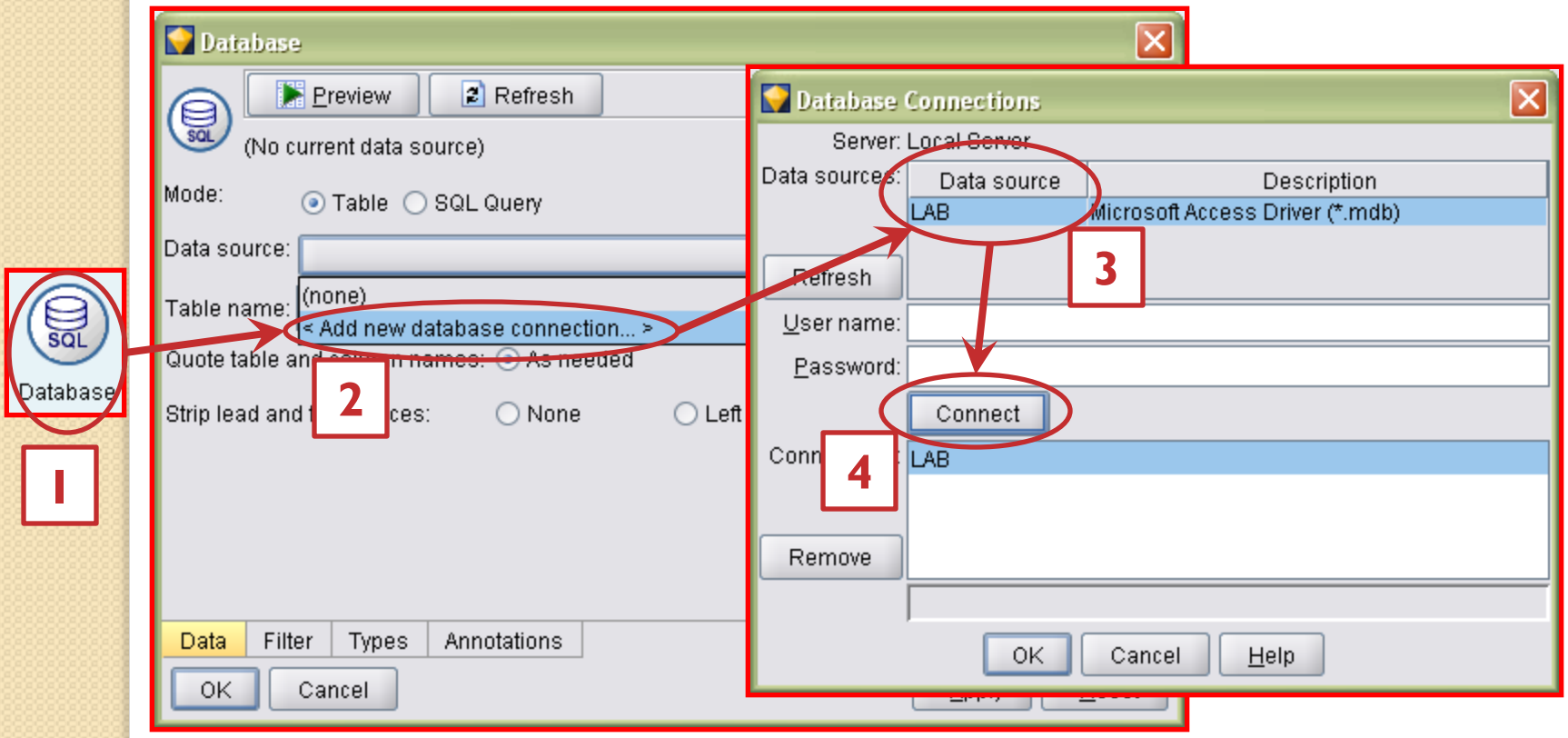
Create Data Source

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

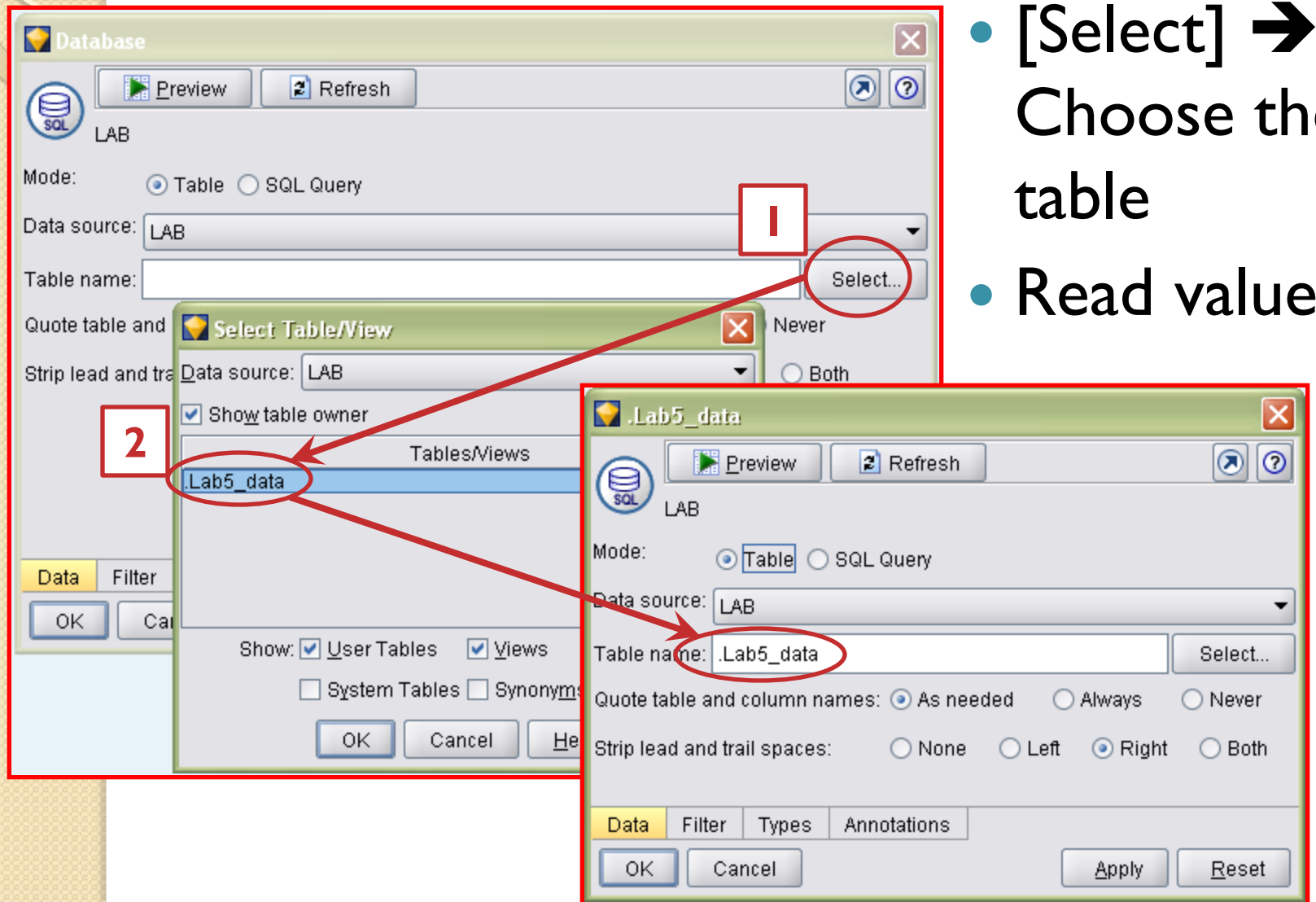


Import Data

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



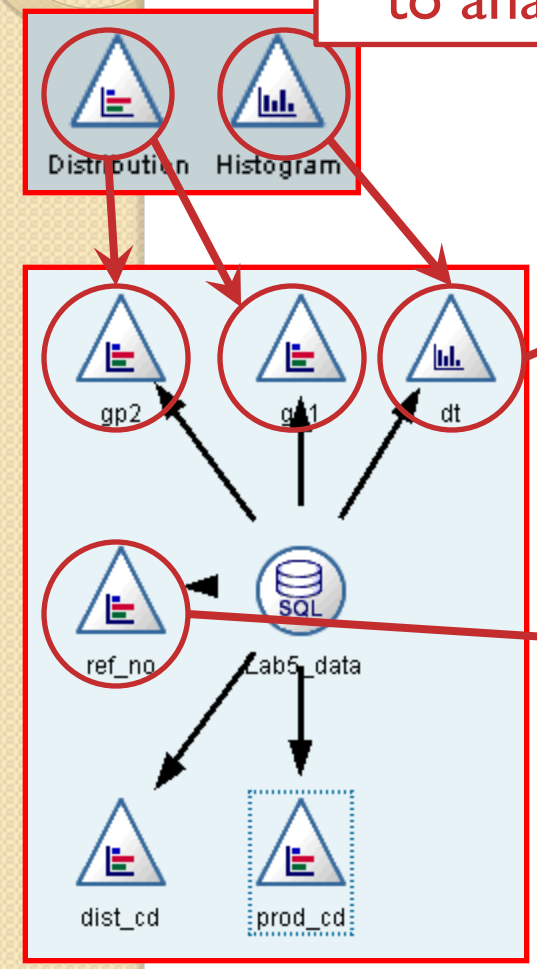
Import Data

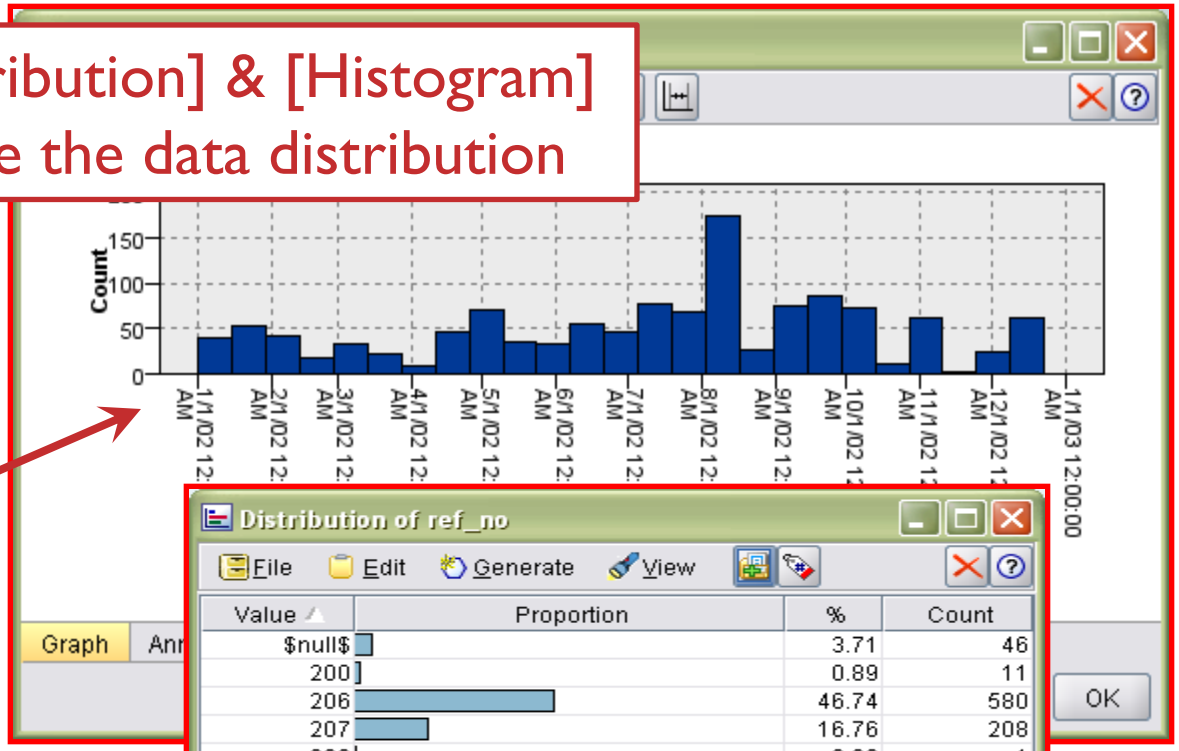


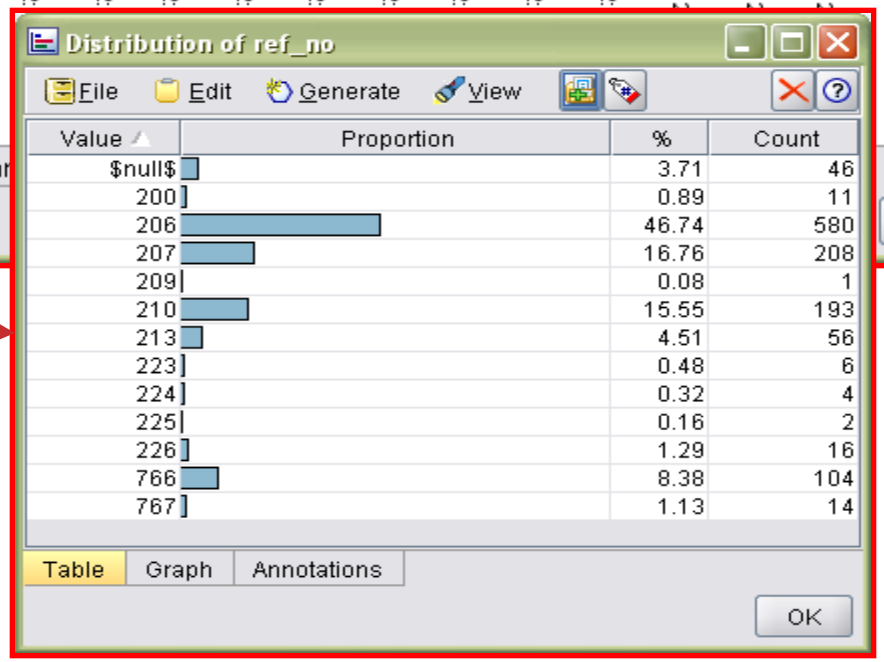
- [Select] ➔ Choose the table
- Read values

Data Understanding

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



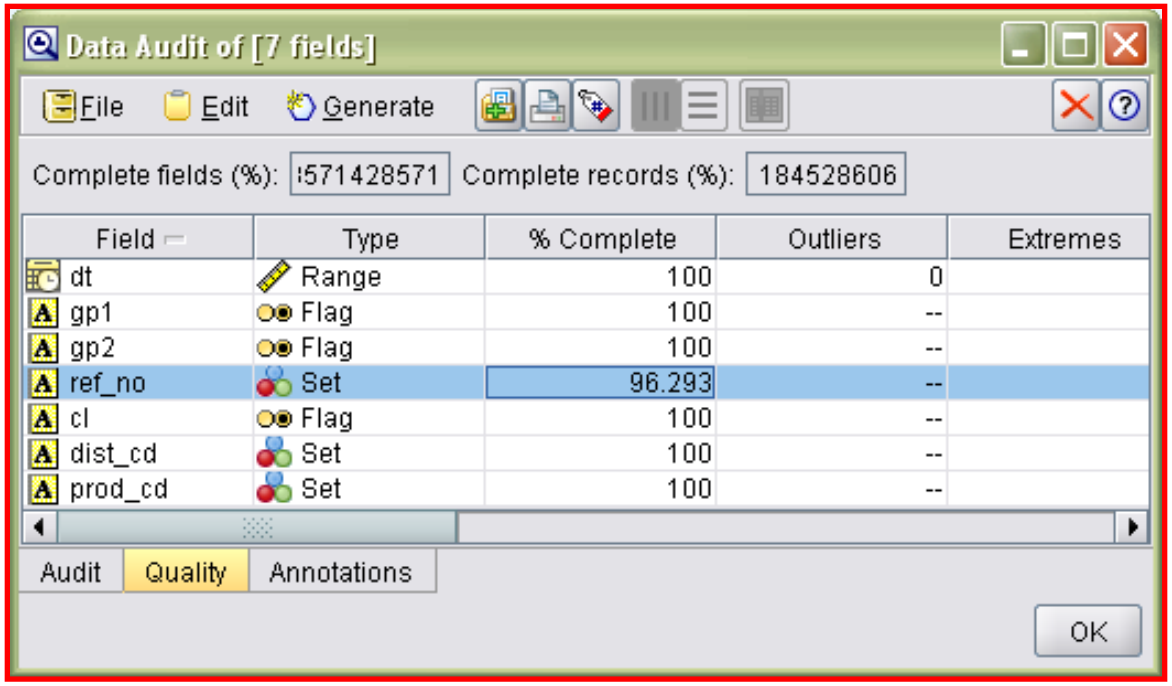




Value	Proportion	%	Count
$null$		3.71	46
200		0.89	11
206		46.74	580
207		16.76	208
209		0.08	1
210		15.55	193
213		4.51	56
223		0.48	6
224		0.32	4
225		0.16	2
226		1.29	16
766		8.38	104
767		1.13	14

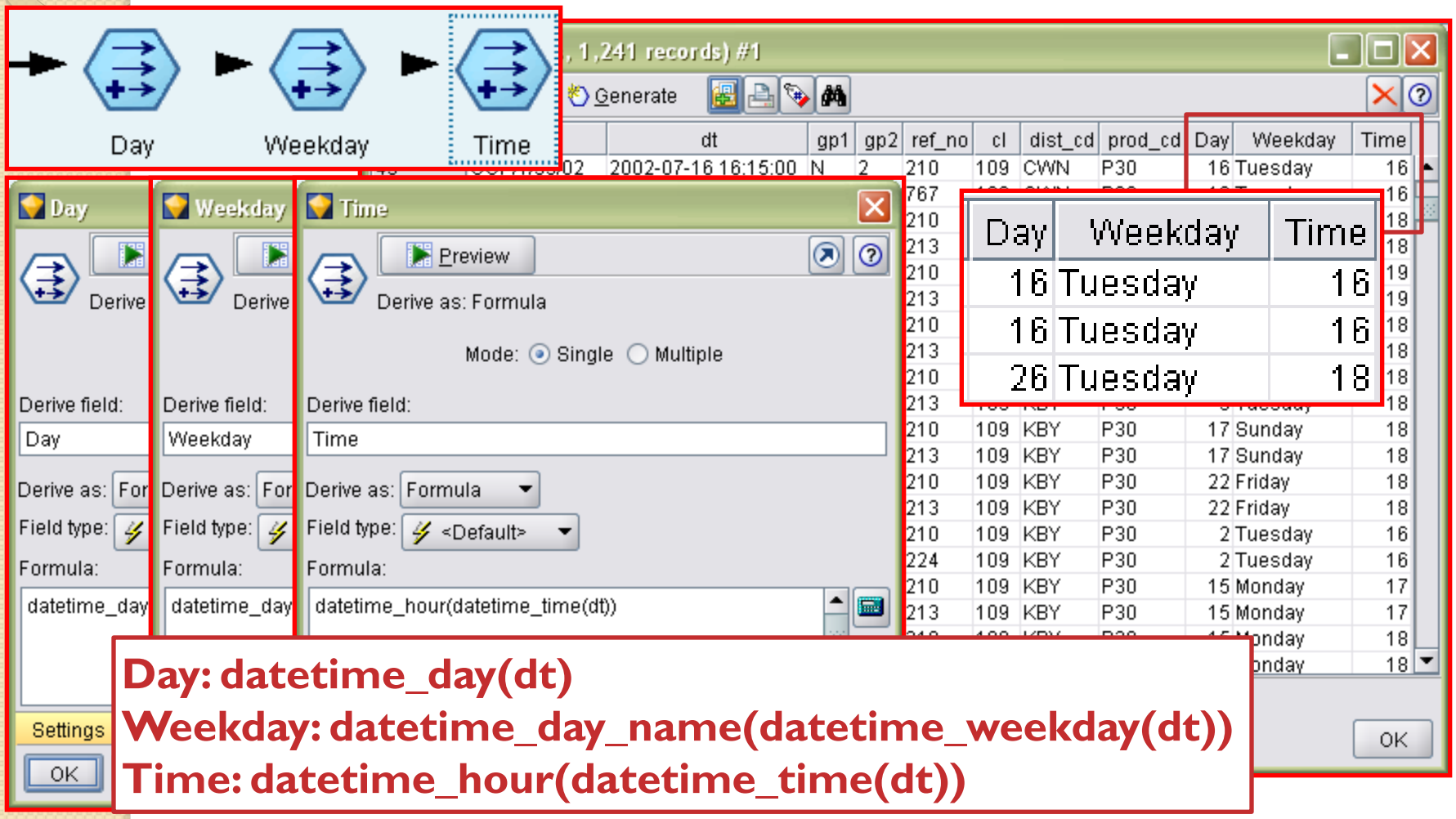
Data Understanding

- 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 Transformation

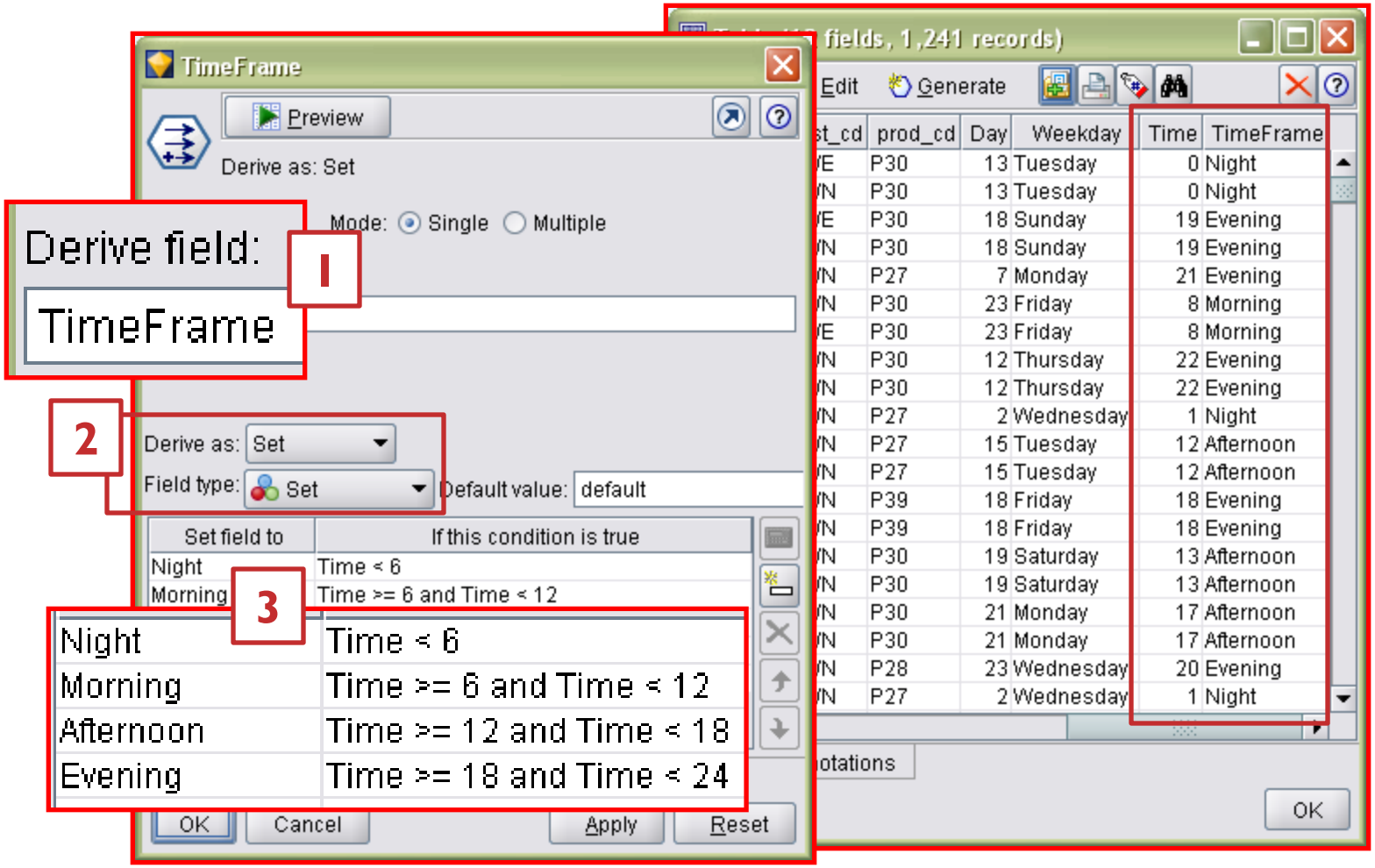
- Derive new attributes: [Day], [Weekday] & [Time]



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

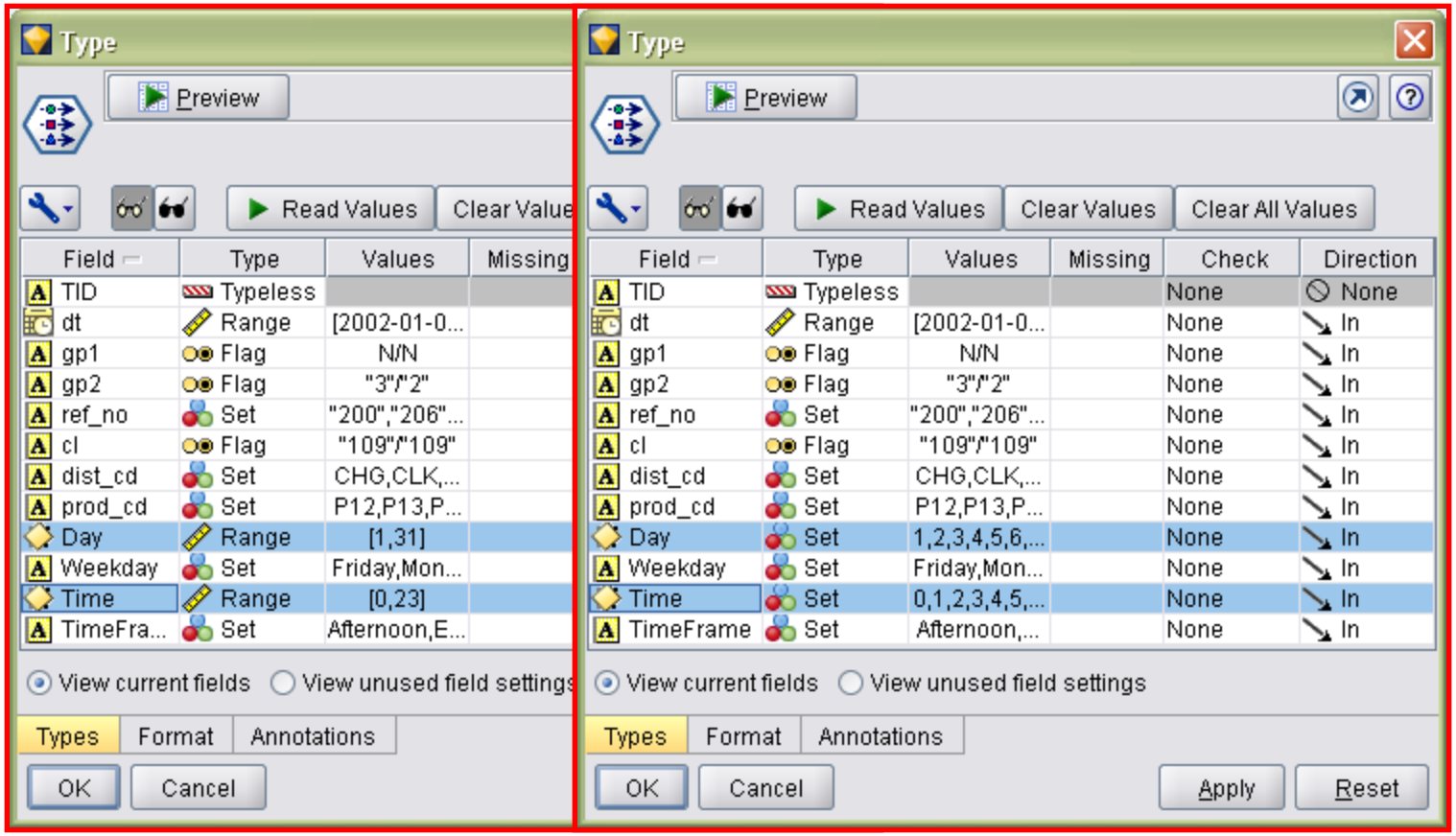
Data Transformation

- Divide [Time] into intervals (binning)



Data Type Re-Define

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



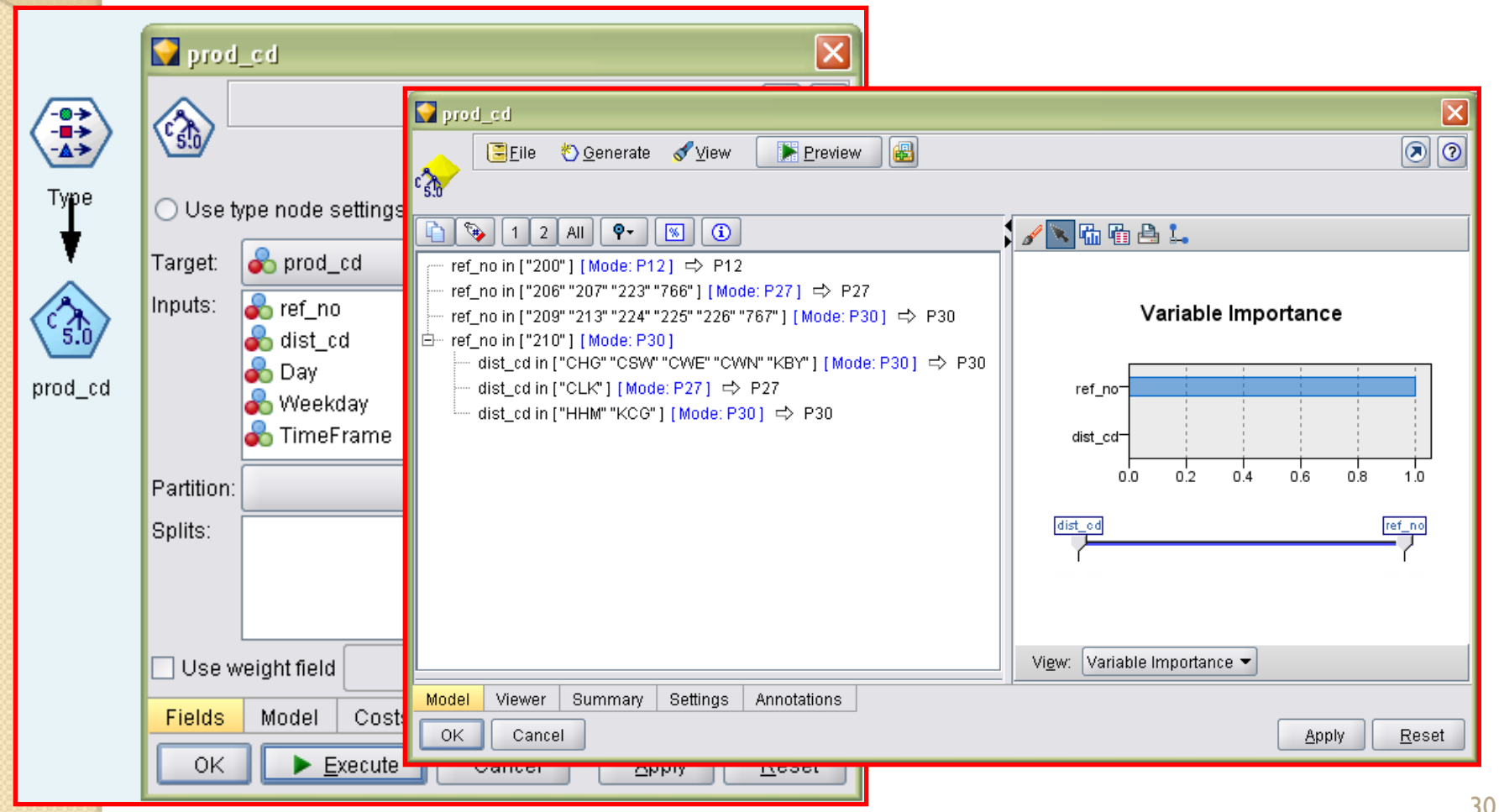
Data Cleaning - Question

- 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.

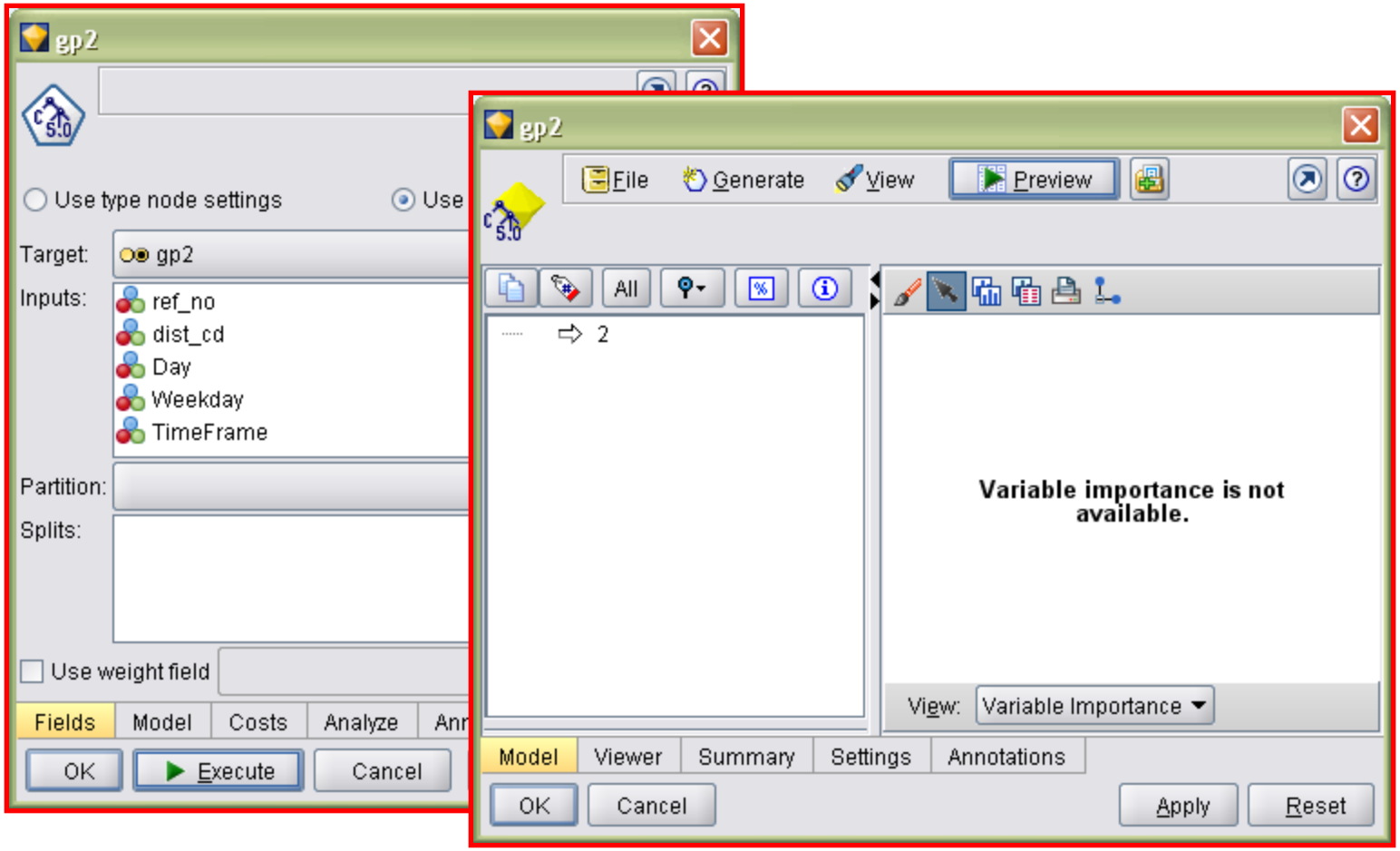
Generate the Model

- Use [C5.0]



Another Model with **No Rule**

- Use same inputs but target changes to [gp2]



Generate a **Balance** Node

- [Generate] ➔ [Balance Node (boost)]