Data Mining – LAB 3

Clustering – K-Means

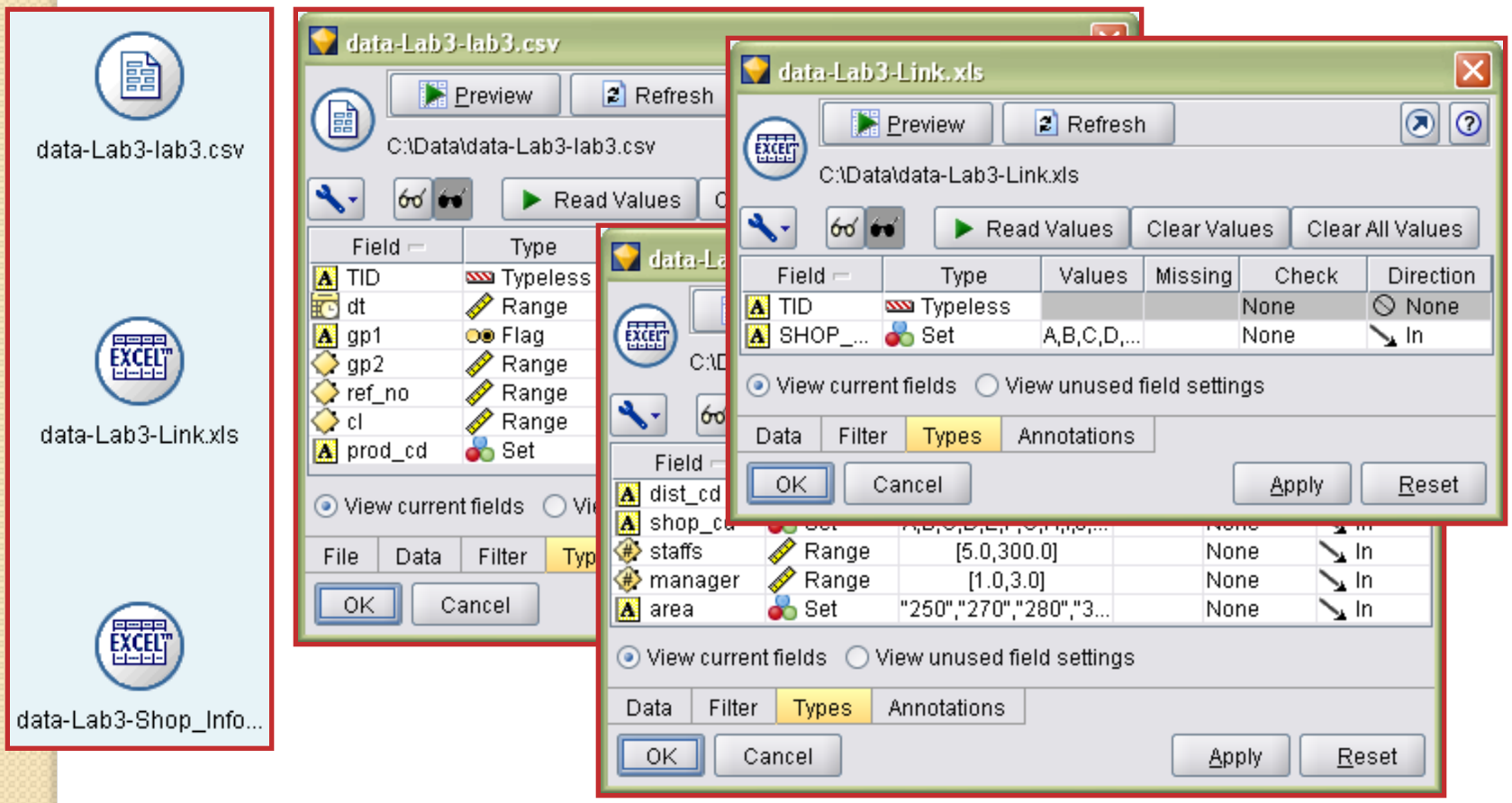
Data File

- In Virtualbox image:
  - c:\data\data-Lab3-lab3.csv
  - c:\data\data-Lab3-Link.xls
  - c:\data\data-Lab3-Shop_Info.xls
- Download from:
  - *www.comp.polyu.edu.hk/~csamak/data/Lab3.zip*
  - *www.comp.polyu.edu.hk/~csamak/data/data-Lab3.mdb*

For *data-Lab3.mdb*, it does not work probably in the **VM**. So, do the lab with the "Lab3.zip" if you use the virtual machine.

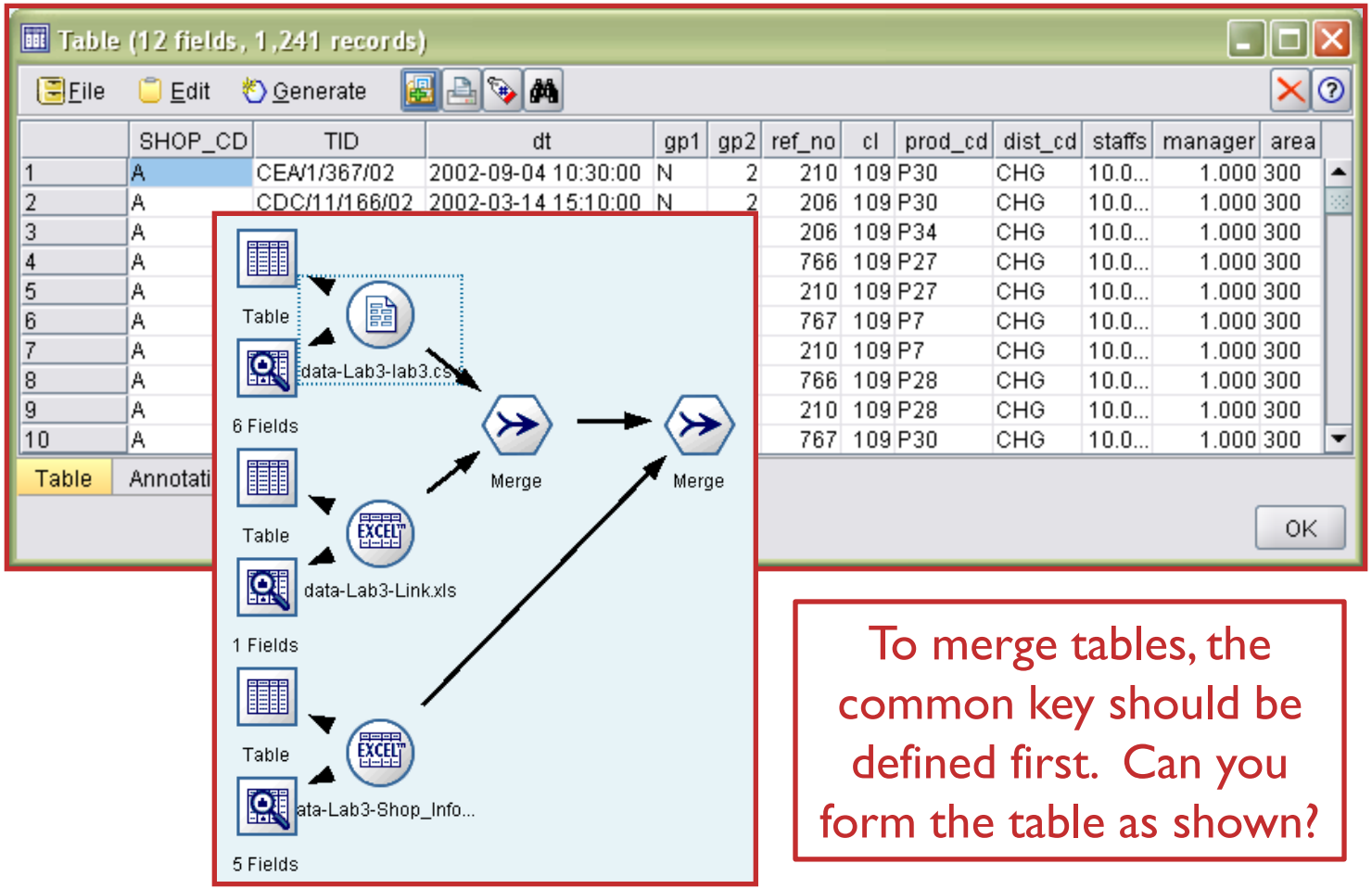
Import Data from MS Excel File

- Use 3 "Var. File" nodes to import the three data files



Import Data

- Merge the three tables into ONE



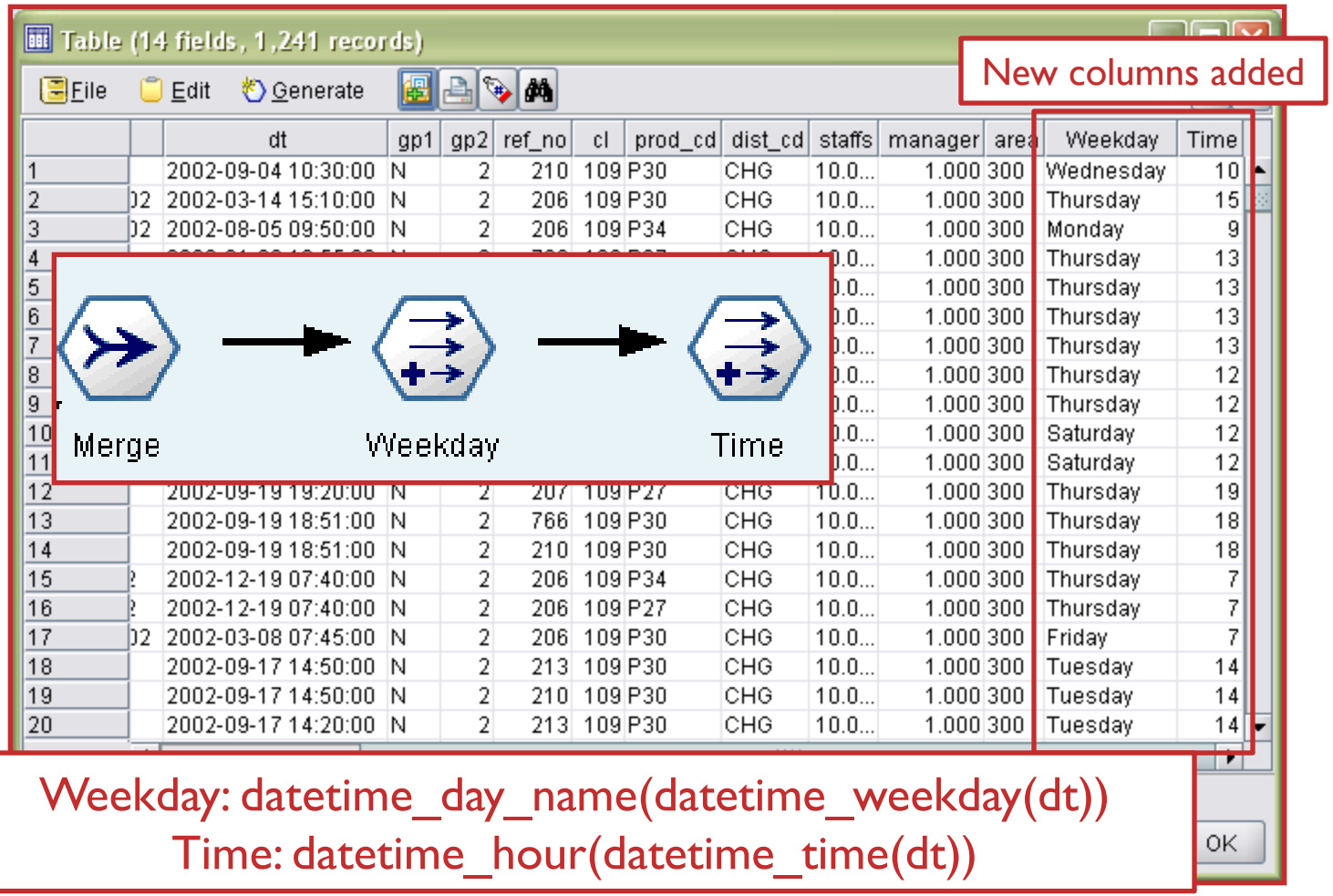
To merge tables, the common key should be defined first. Can you form the table as shown?

Data Understanding/Cleaning

- What are the errors in the data set?

- Suggest methods to correct them.

Data Transformation

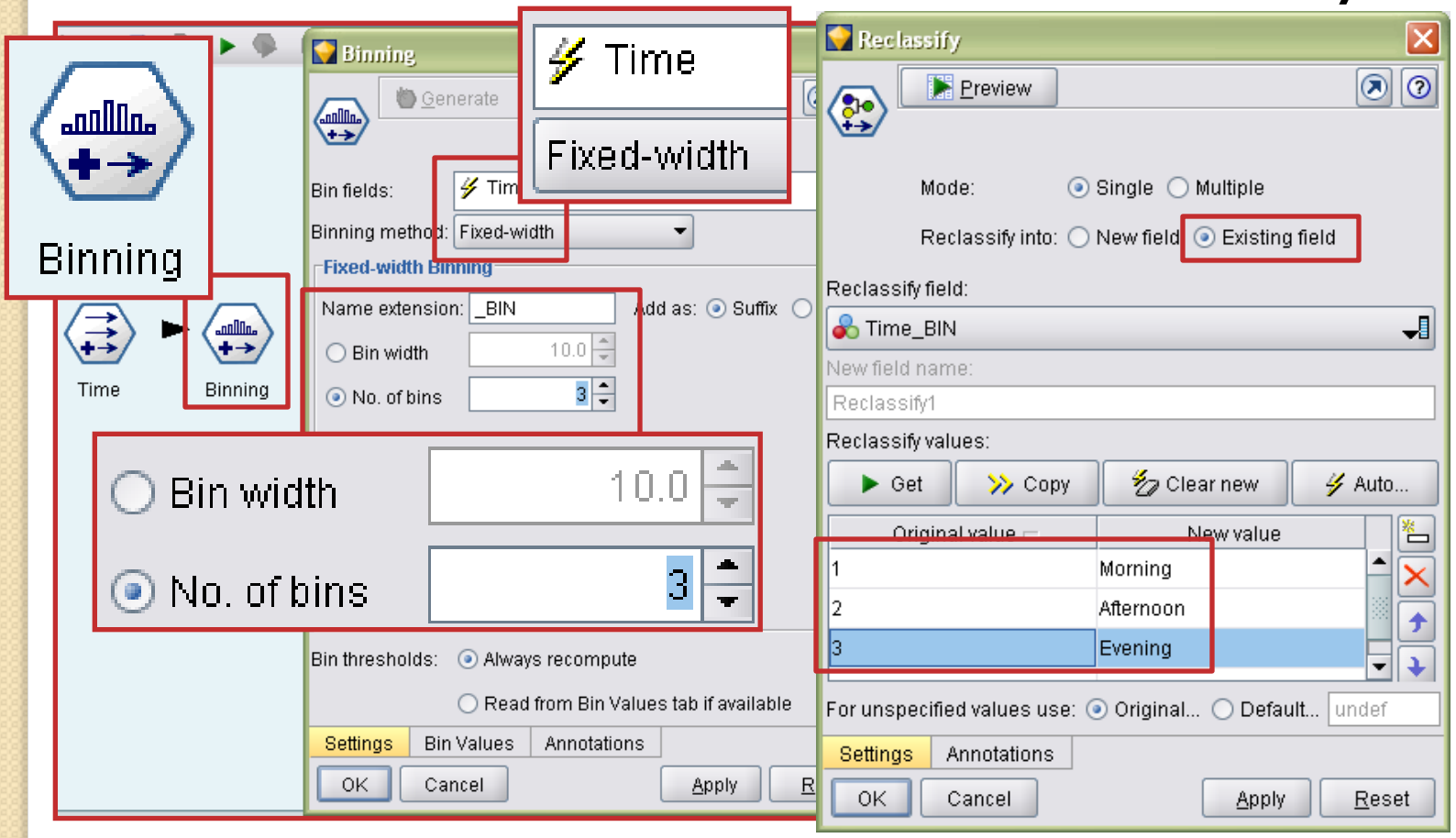


		dt	gp1	gp2	ref_no	cl	prod_cd	dist_cd	staffs	manager	area	Weekday	Time
1		2002-09-04 10:30:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Wednesday	10
2	02	2002-03-14 15:10:00	N	2	206	109	P30	CHG	10.0...	1.000	300	Thursday	15
3	02	2002-08-05 09:50:00	N	2	206	109	P34	CHG	10.0...	1.000	300	Monday	9
4									0.0...	1.000	300	Thursday	13
5									0.0...	1.000	300	Thursday	13
6									0.0...	1.000	300	Thursday	13
7									0.0...	1.000	300	Thursday	13
8									0.0...	1.000	300	Thursday	12
9									0.0...	1.000	300	Thursday	12
10									0.0...	1.000	300	Saturday	12
11									0.0...	1.000	300	Saturday	12
12		2002-09-19 19:20:00	N	2	207	109	P27	CHG	10.0...	1.000	300	Thursday	19
13		2002-09-19 18:51:00	N	2	766	109	P30	CHG	10.0...	1.000	300	Thursday	18
14		2002-09-19 18:51:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Thursday	18
15	?	2002-12-19 07:40:00	N	2	206	109	P34	CHG	10.0...	1.000	300	Thursday	7
16	?	2002-12-19 07:40:00	N	2	206	109	P27	CHG	10.0...	1.000	300	Thursday	7
17	02	2002-03-08 07:45:00	N	2	206	109	P30	CHG	10.0...	1.000	300	Friday	7
18		2002-09-17 14:50:00	N	2	213	109	P30	CHG	10.0...	1.000	300	Tuesday	14
19		2002-09-17 14:50:00	N	2	210	109	P30	CHG	10.0...	1.000	300	Tuesday	14
20		2002-09-17 14:20:00	N	2	213	109	P30	CHG	10.0...	1.000	300	Tuesday	14

Weekday: datetime_day_name(datetime_weekday(dt))
Time: datetime_hour(datetime_time(dt))

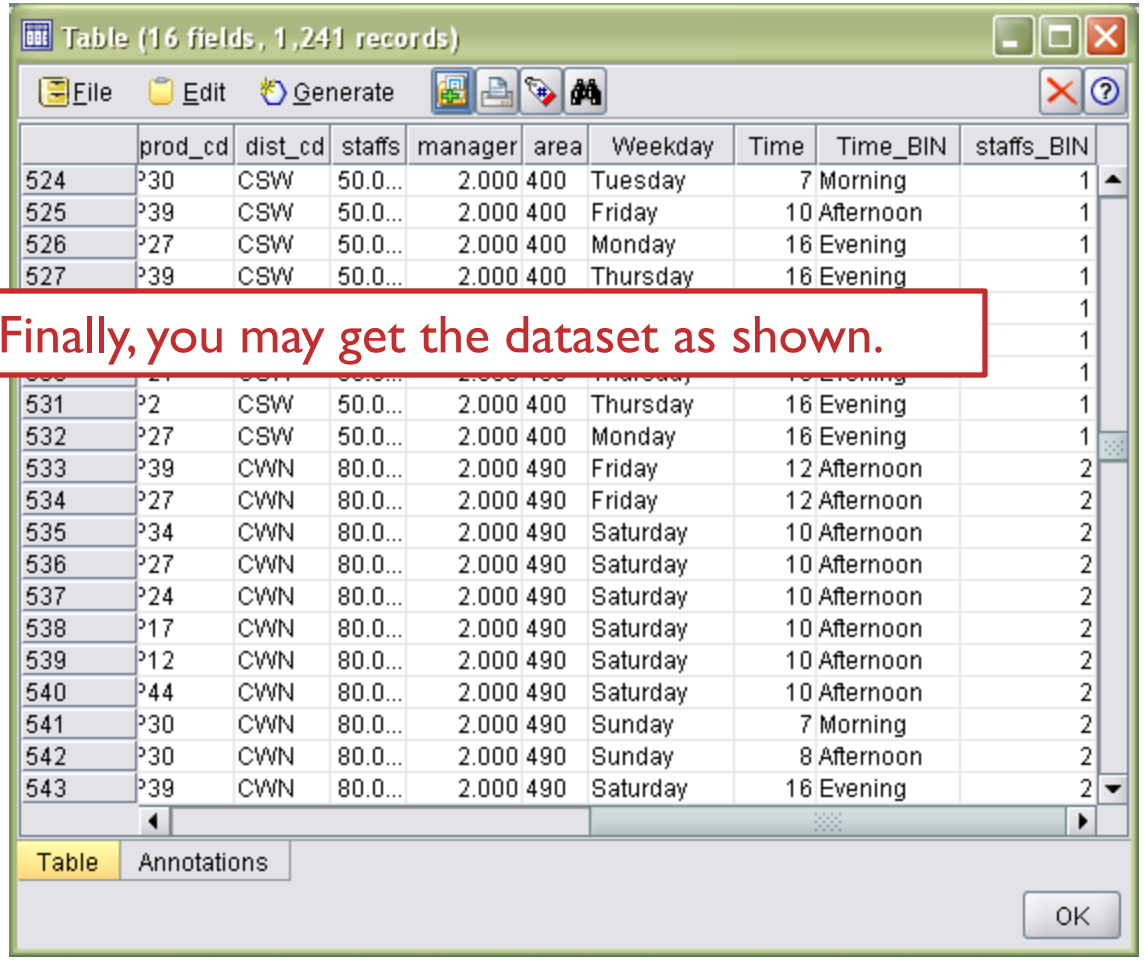
Data Transformation – Discretization

- Divide "Time" into 3 intervals (binning) and rename the values with node, Reclassify



Data Transformation – Discretization

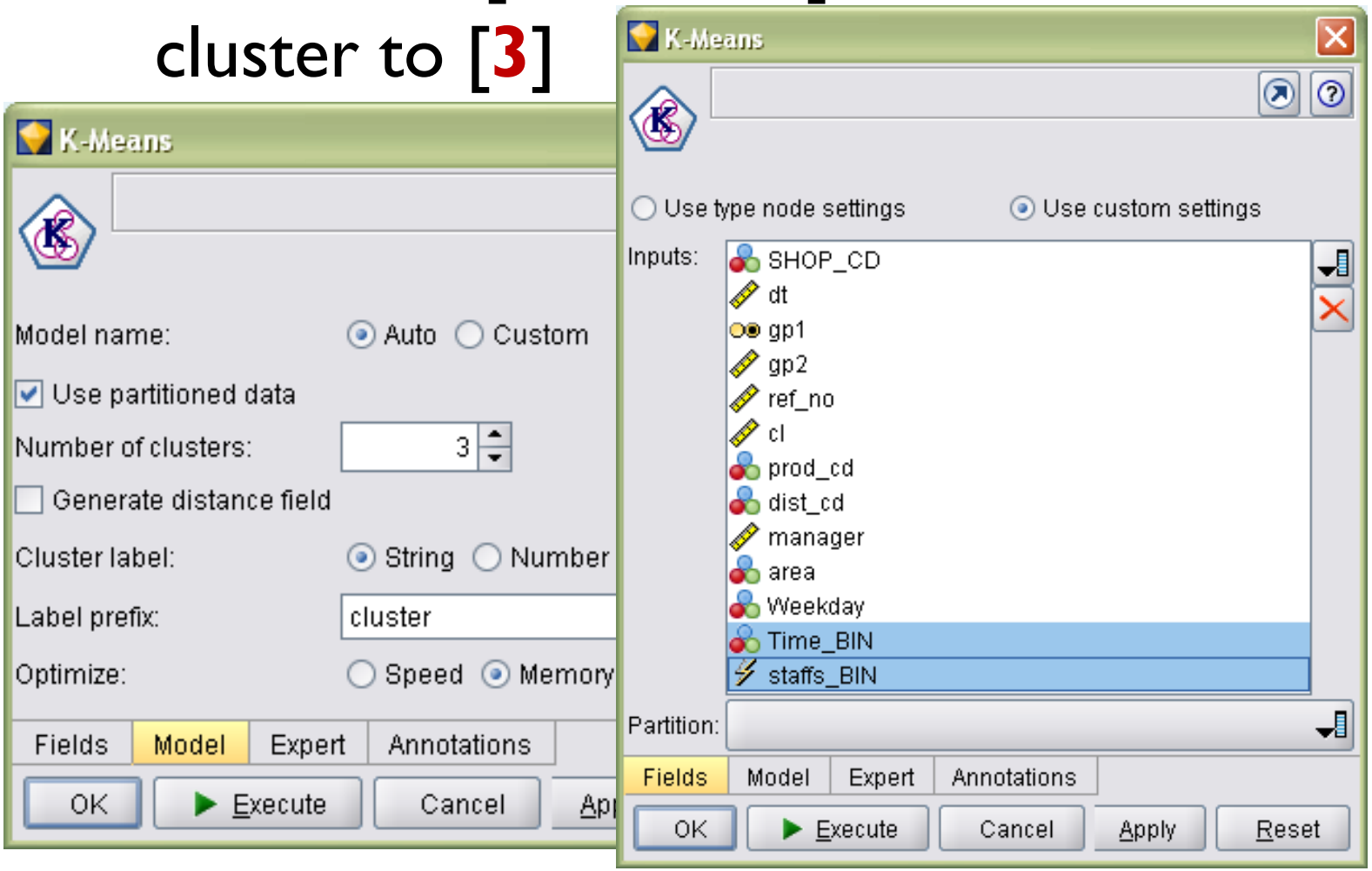
- Divide "staffs" into 5 intervals (Fixed-width)



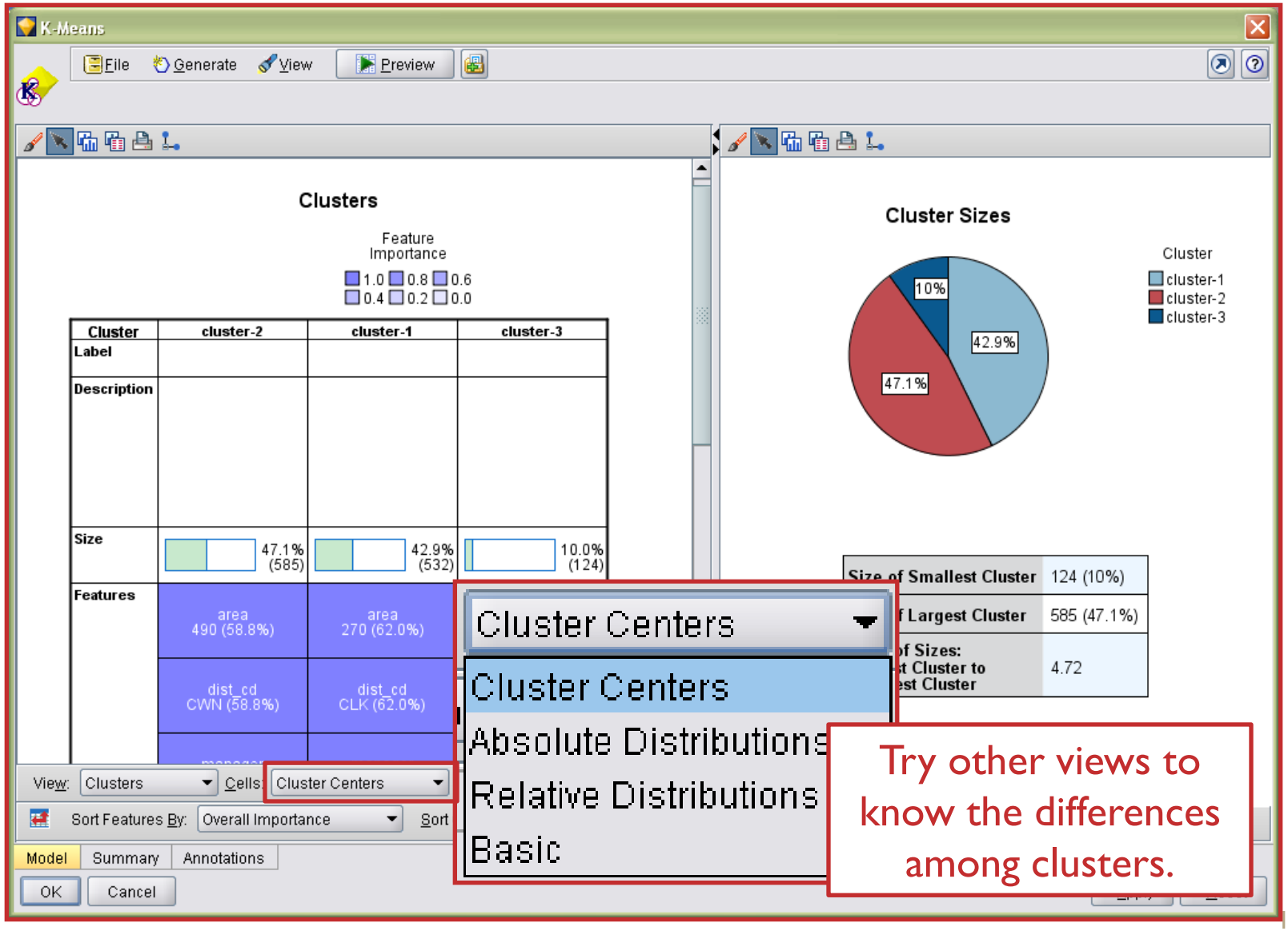
	prod_cd	dist_cd	staffs	manager	area	Weekday	Time	Time_BIN	staffs_BIN
524	P30	CSW	50.0...	2.000	400	Tuesday	7	Morning	1
525	P39	CSW	50.0...	2.000	400	Friday	10	Afternoon	1
526	P27	CSW	50.0...	2.000	400	Monday	16	Evening	1
527	P39	CSW	50.0...	2.000	400	Thursday	16	Evening	1
									1
									1
									1
531	P2	CSW	50.0...	2.000	400	Thursday	16	Evening	1
532	P27	CSW	50.0...	2.000	400	Monday	16	Evening	1
533	P39	CWN	80.0...	2.000	490	Friday	12	Afternoon	2
534	P27	CWN	80.0...	2.000	490	Friday	12	Afternoon	2
535	P34	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
536	P27	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
537	P24	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
538	P17	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
539	P12	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
540	P44	CWN	80.0...	2.000	490	Saturday	10	Afternoon	2
541	P30	CWN	80.0...	2.000	490	Sunday	7	Morning	2
542	P30	CWN	80.0...	2.000	490	Sunday	8	Afternoon	2
543	P39	CWN	80.0...	2.000	490	Saturday	16	Evening	2

Clustering – Divide "Customers into 3 Clusters

- Add node, [**K-Means**] and set number of cluster to [**3**]



Result


K-Means
File
Generate
View
Preview
Clusters
Feature Importance
1.0 0.8 0.6
0.4 0.2 0.0
Cluster
cluster-2
cluster-1
cluster-3
Label
Description
Size
47.1% (585)
42.9% (532)
10.0% (124)
Features
area 490 (58.8%)
area 270 (62.0%)
dist_cd CWN (58.8%)
dist_cd CLK (62.0%)
Cluster Sizes
Cluster
cluster-1
cluster-2
cluster-3
10%
42.9%
47.1%
Size of Smallest Cluster
124 (10%)
Largest Cluster
585 (47.1%)
Sizes:
Cluster to
Cluster
4.72
Cluster Centers
Cluster Centers
Absolute Distributions
Relative Distributions
Basic
View: Clusters
Cells: Cluster Centers
Sort Features By: Overall Importance
Sort
Model
Summary
Annotations
OK
Cancel
Try other views to know the differences among clusters.