Milestone 3 Processing of Data

WQD7005 – DATA MINING

ZULKANAIN BIN HASAN

WQD180031

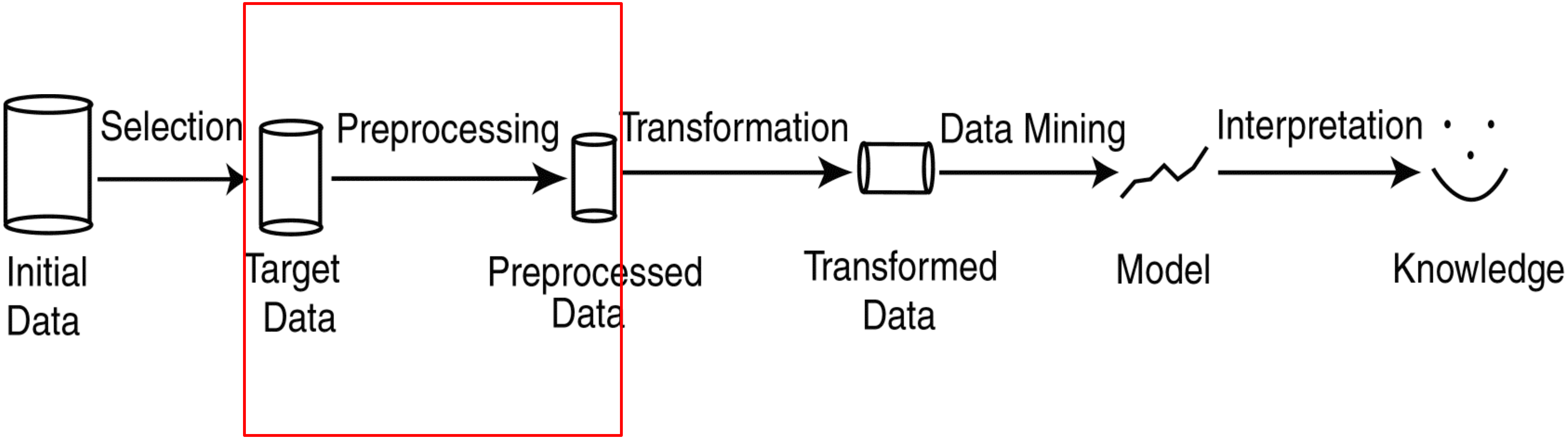
NUR ATHIRAH BT. RAZAK

WQD170072

Knowledge Discovery in Databases (KDD): process of finding useful information and patterns in data:

1. Selection (Pre-Mining 1): Obtain data from various sources
2. Preprocessing (Pre-Mining 2): Cleanse data

Milestone 3



Analysis Goal:
Findings a good stock counter for short term and long-term investment. Using historical data from KLSE Stock Market to predict future trend.

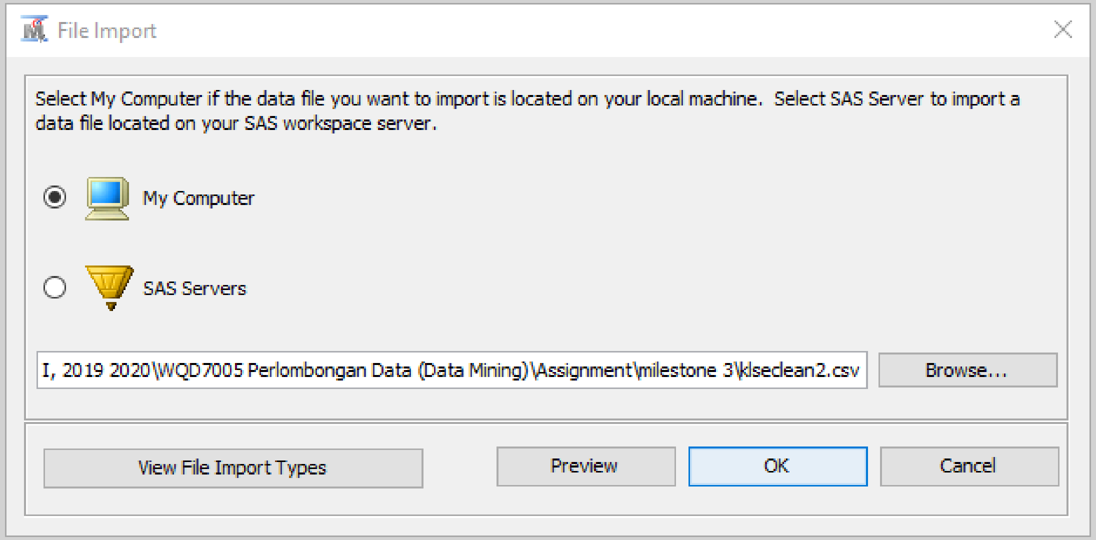
Analysis Data:

- Extract data from historical KLSE Stock Market
- Sample target based on adj-close data trend
- Actual sample target approximately 10%

Create SAS data file

Step 1: Import .csv file from local storage to project diagram by using File Import.

- Data have been pre-cleaning first before import to project diagram.
- Only target data have been used in this step.



Select My Computer if the data file you want to import is located on your local machine. Select SAS Server to import a data file located on your SAS workspace server.

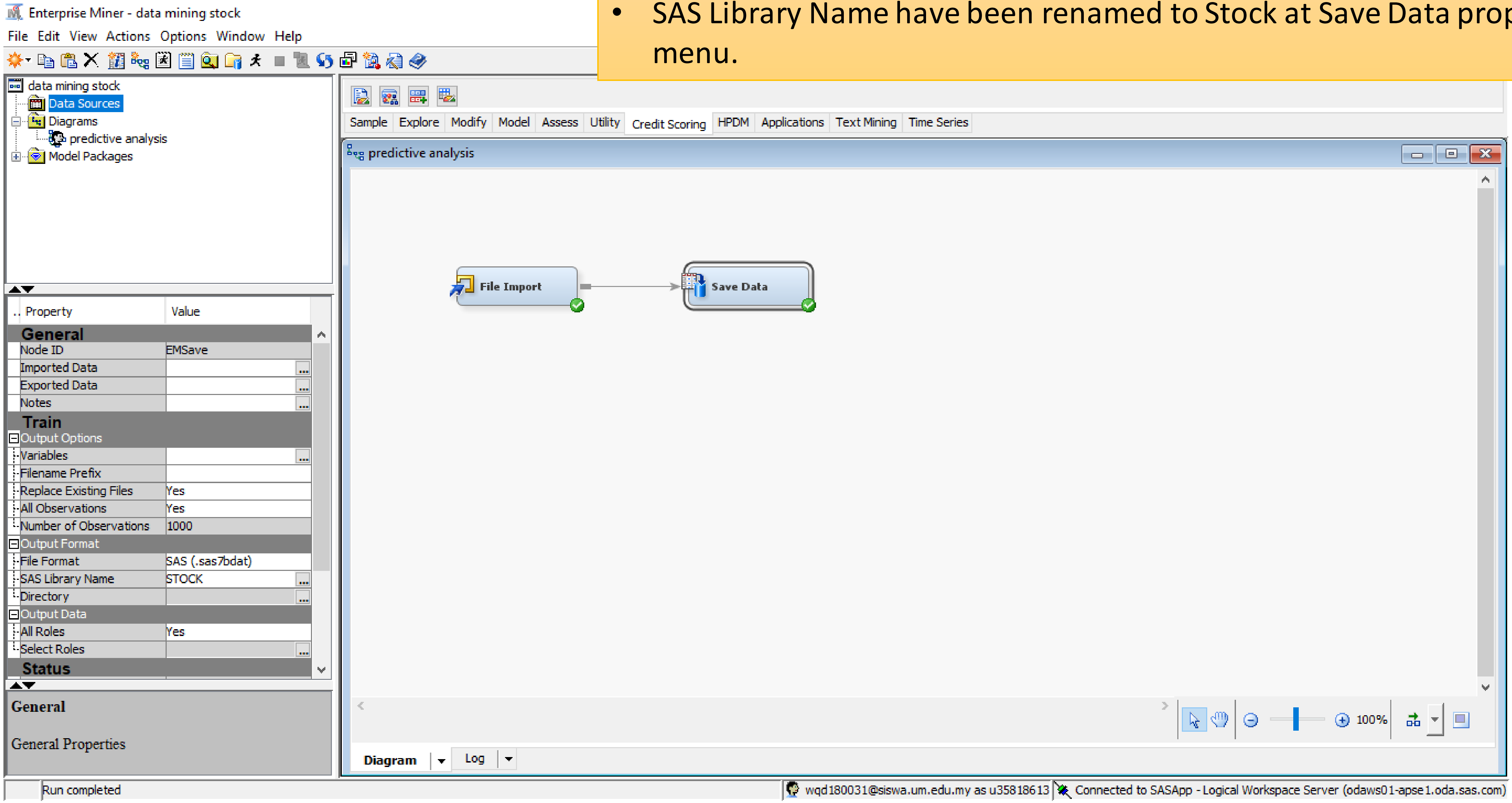
● My Computer

○ SAS Servers

I, 2019 2020\WQD7005 Perlombongan Data (Data Mining)\Assignment\milestone 3\klseclean2.csv

Step 2: Create linkages between File Import to Save Data in order to convert .csv file to SAS data file.

- SAS Library Name have been renamed to Stock at Save Data property menu.



Property	Value
General	
Node ID	EMSave
Imported Data	...
Exported Data	...
Notes	...
Train	
Output Options	
Variables	...
Filename Prefix	
Replace Existing Files	Yes
All Observations	Yes
Number of Observations	1000
Output Format	
File Format	SAS (.sas7bdat)
SAS Library Name	STOCK
Directory	...
Output Data	
All Roles	Yes
Select Roles	...
Status	

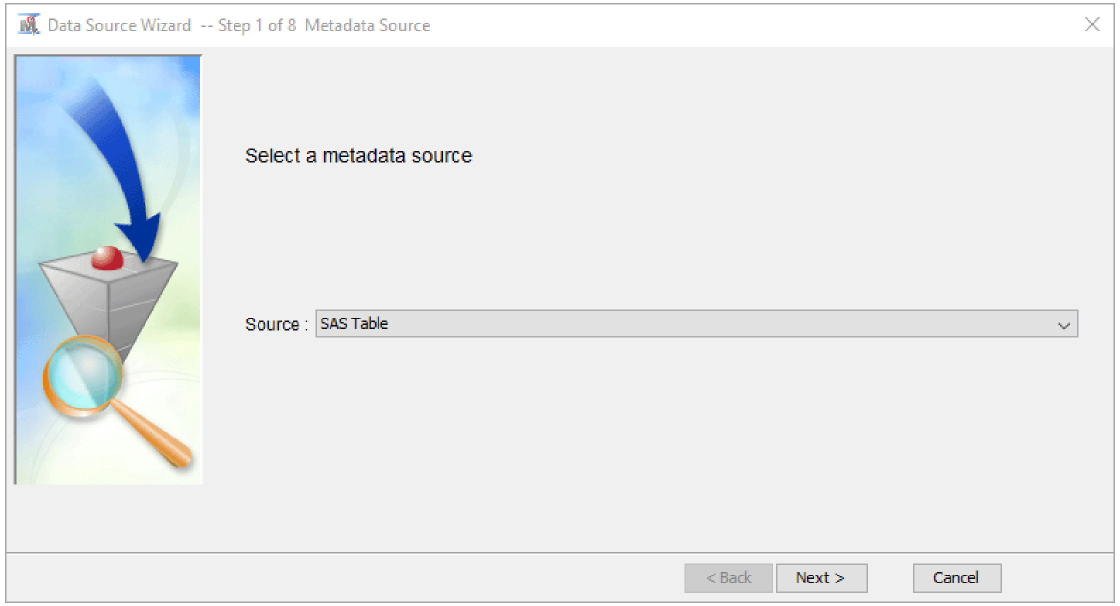
General

General Properties

Run completed

wqd180031@siswa.um.edu.my as u35818613

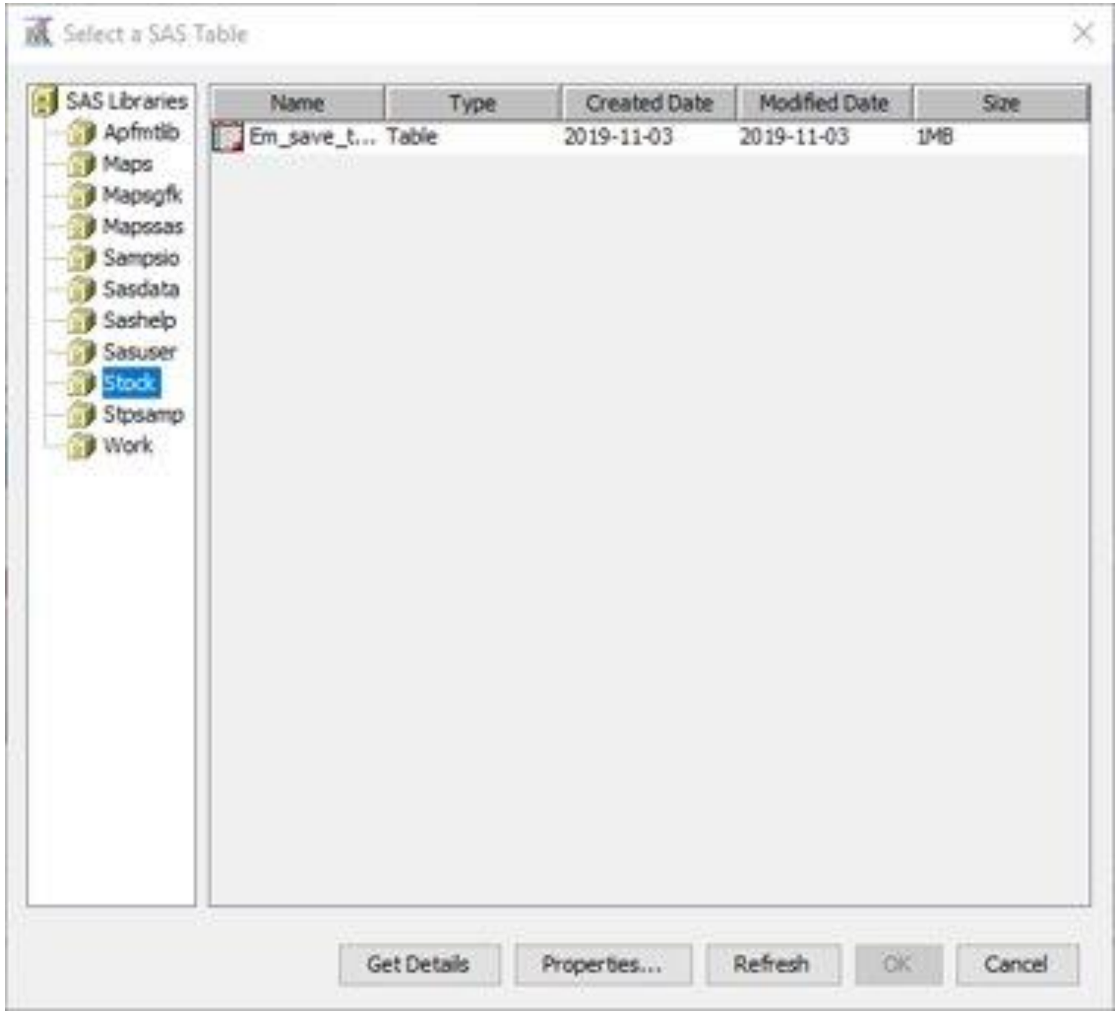
Connected to SASApp - Logical Workspace Server (odaws01-apse1.oda.sas.com)



Data Source Wizard -- Step 1 of 8 Metadata Source

Select a metadata source

Source : SAS Table



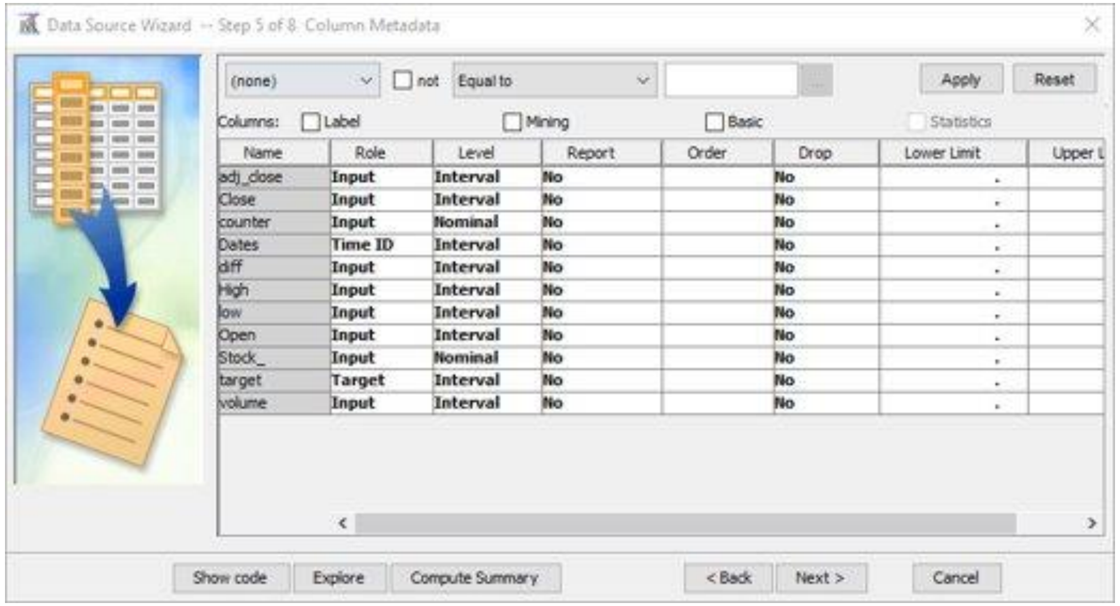
Select a SAS Table

Name	Type	Created Date	Modified Date	Size
Em_save_t...	Table	2019-11-03	2019-11-03	1MB

Step 3: Select a metadata source that have been create by previous step.

- For this case, data stored in SAS Library – Stock as per previous setup.

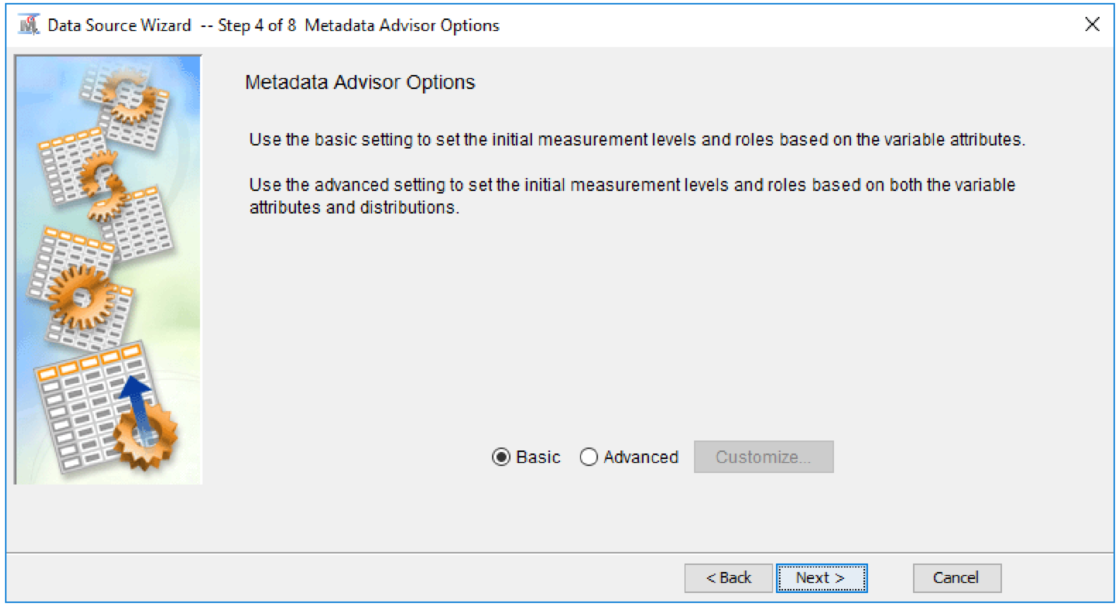
- Screening Column Metadata by inspecting all variables for their Role and Level.
- Target have been identified for this data and can be proceed with setting the Metadata Advisor Options (Basic or Advance).



Data Source Wizard -- Step 5 of 8 Column Metadata

Columns: □ Label □ Mining □ Basic □ Statistics

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper L
adj_close	Input	Interval	No		No	.	
Close	Input	Interval	No		No	.	
counter	Input	Nominal	No		No	.	
Dates	Time ID	Interval	No		No	.	
diff	Input	Interval	No		No	.	
High	Input	Interval	No		No	.	
low	Input	Interval	No		No	.	
Open	Input	Interval	No		No	.	
Stock_	Input	Nominal	No		No	.	
target	Target	Interval	No		No	.	
volume	Input	Interval	No		No	.	



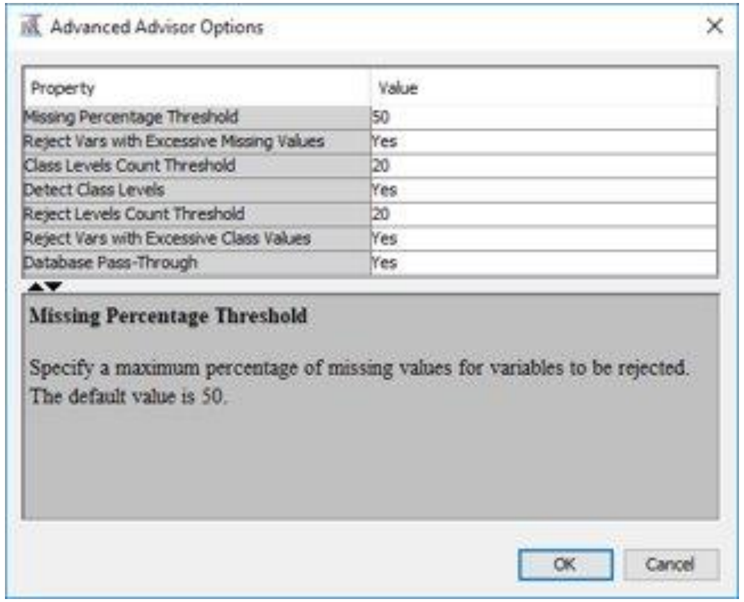
Data Source Wizard -- Step 4 of 8 Metadata Advisor Options

Metadata Advisor Options

Use the basic setting to set the initial measurement levels and roles based on the variable attributes.

Use the advanced setting to set the initial measurement levels and roles based on both the variable attributes and distributions.

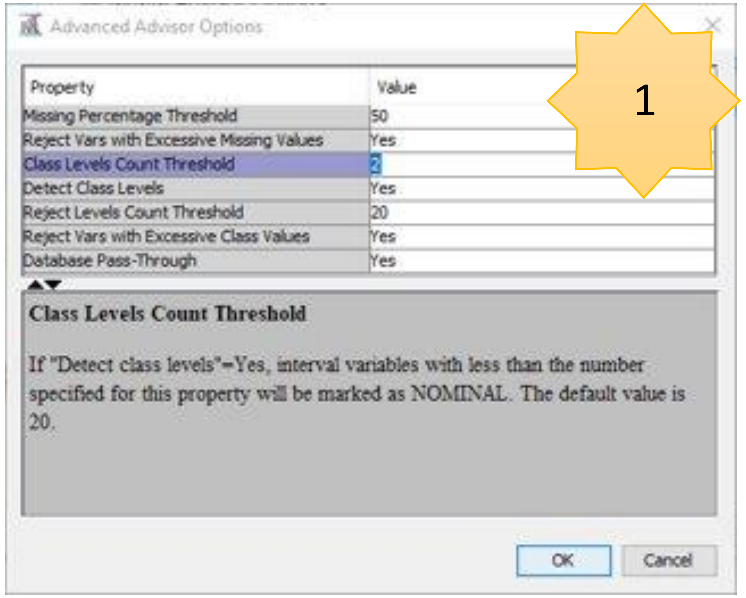
● Basic ○ Advanced



Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Values	Yes
Class Levels Count Threshold	20
Detect Class Levels	Yes
Reject Levels Count Threshold	20
Reject Vars with Excessive Class Values	Yes
Database Pass-Through	Yes

Missing Percentage Threshold

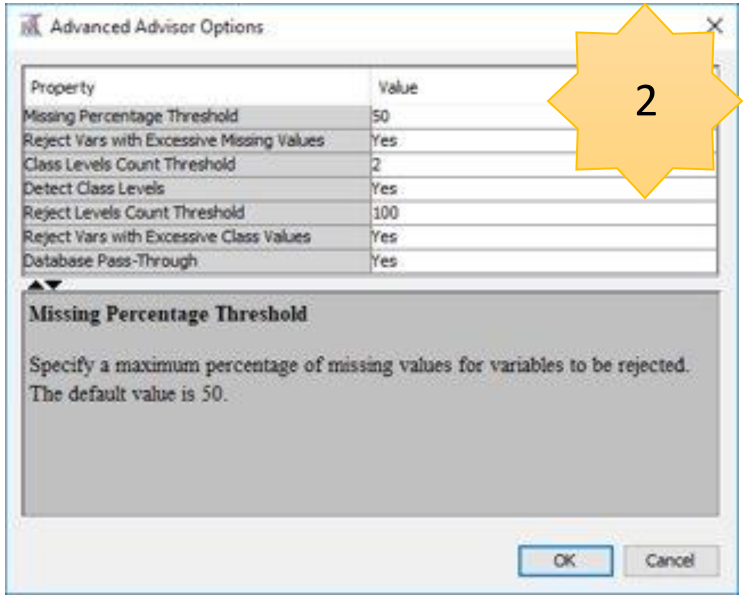
Specify a maximum percentage of missing values for variables to be rejected. The default value is 50.



Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Values	Yes
Class Levels Count Threshold	2
Detect Class Levels	Yes
Reject Levels Count Threshold	20
Reject Vars with Excessive Class Values	Yes
Database Pass-Through	Yes

Class Levels Count Threshold

If "Detect class levels"=Yes, interval variables with less than the number specified for this property will be marked as NOMINAL. The default value is 20.



Property	Value
Missing Percentage Threshold	50
Reject Vars with Excessive Missing Values	Yes
Class Levels Count Threshold	2
Detect Class Levels	Yes
Reject Levels Count Threshold	100
Reject Vars with Excessive Class Values	Yes
Database Pass-Through	Yes

Missing Percentage Threshold

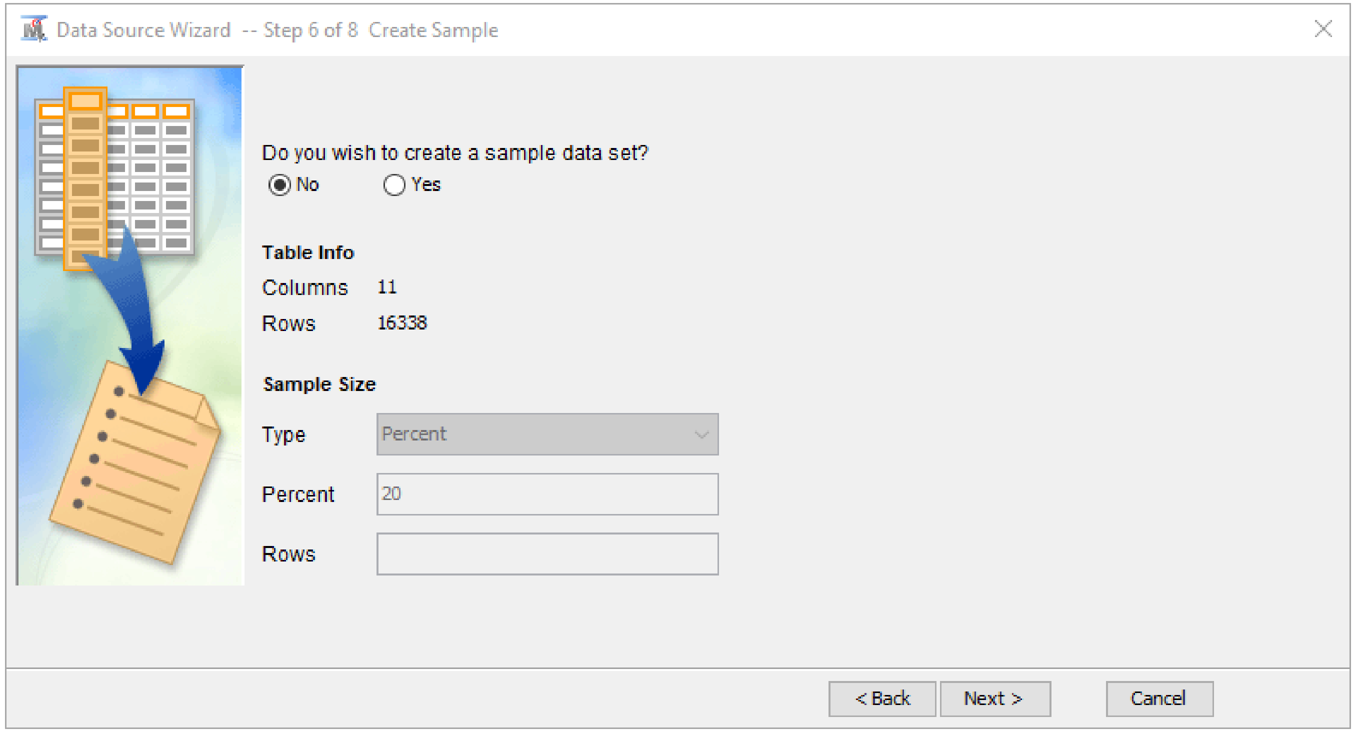
Specify a maximum percentage of missing values for variables to be rejected. The default value is 50.

In Advance Advisor Options, 2 property have been changed:

i. Class Levels Count Threshold from 2 to 20

ii. Reject Levels Count Threshold from 20 to 100

1



Data Source Wizard -- Step 6 of 8 Create Sample

Do you wish to create a sample data set?

● No ○ Yes

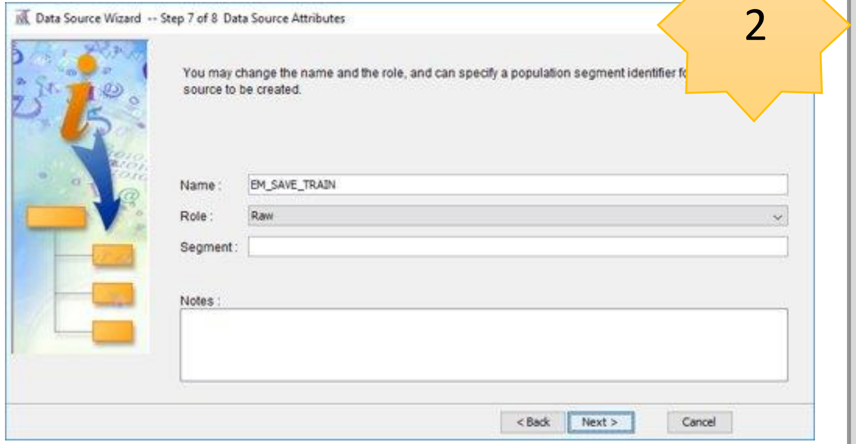
Table Info

Columns	11
Rows	16338

Sample Size

Type	Percent
Percent	20
Rows	

2

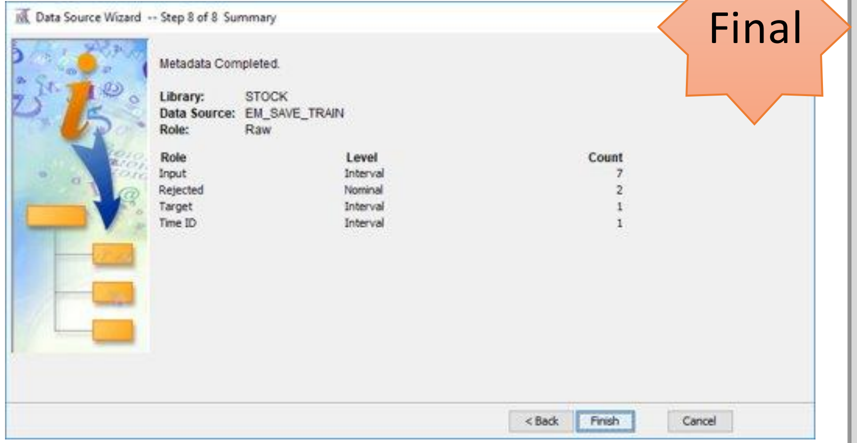


Data Source Wizard -- Step 7 of 8 Data Source Attributes

You may change the name and the role, and can specify a population segment identifier fo
source to be created.

Name :	EM_SAVE_TRAIN
Role :	Raw
Segment :	
Notes :	

Final



Data Source Wizard -- Step 8 of 8 Summary

Metadata Completed.

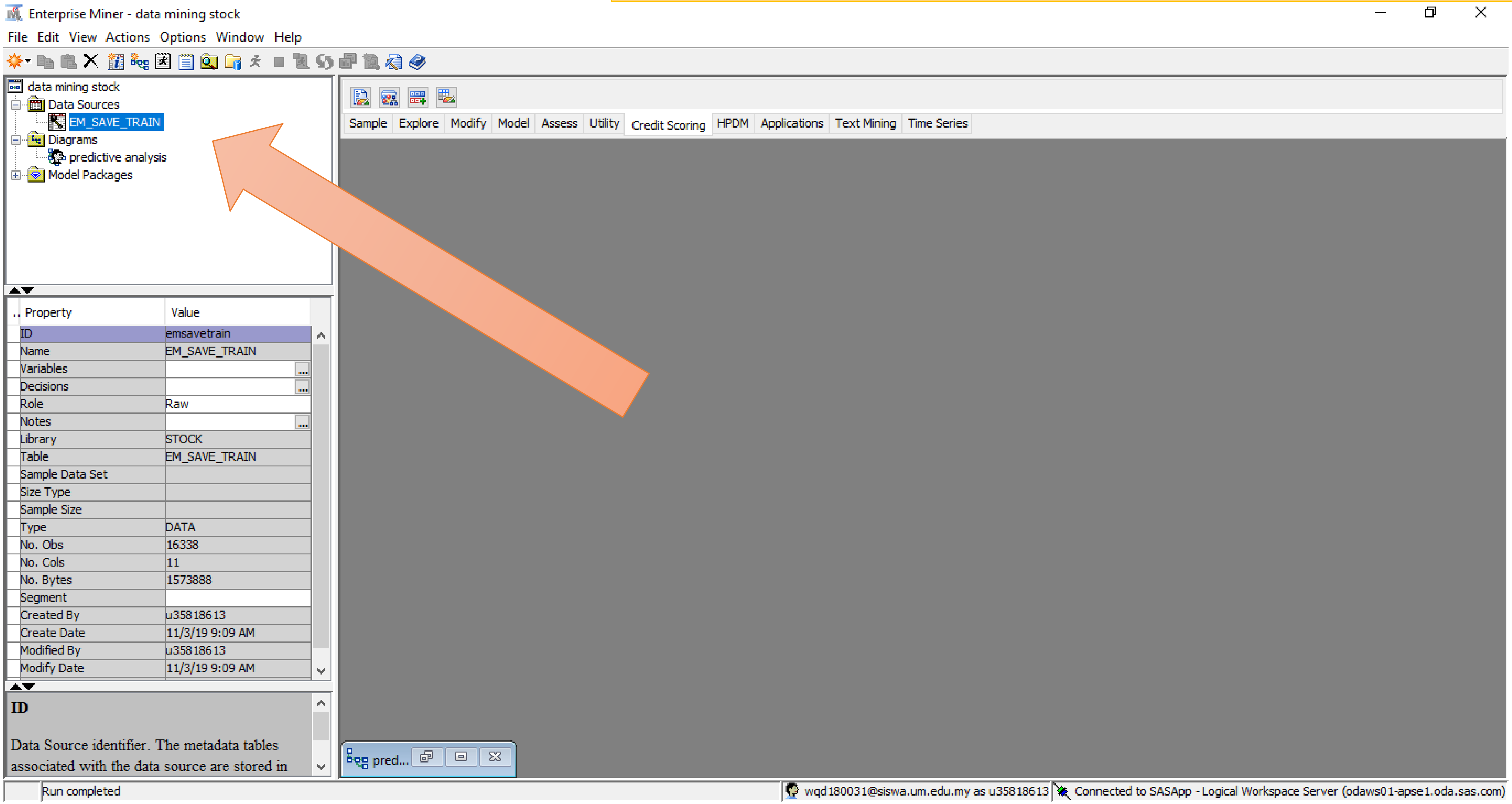
Library: STOCK
Data Source: EM_SAVE_TRAIN
Role: Raw

Role	Level	Count
Input	Interval	7
Rejected	Nominal	2
Target	Interval	1
Time ID	Interval	1

Final step for this part shows:

i. 7 role as Input
ii. 2 role as Rejected for Nominal Level
iii. 1 role as Target

Once success, SAS data file can be seen in Data Source section.



Property	Value
ID	emsavetrain
Name	EM_SAVE_TRAIN
Variables	...
Decisions	...
Role	Raw
Notes	...
Library	STOCK
Table	EM_SAVE_TRAIN
Sample Data Set	
Size Type	
Sample Size	
Type	DATA
No. Obs	16338
No. Cols	11
No. Bytes	1573888
Segment	
Created By	u35818613
Create Date	11/3/19 9:09 AM
Modified By	u35818613
Modify Date	11/3/19 9:09 AM

ID

Data Source identifier. The metadata tables associated with the data source are stored in

Run completed

wqd180031@siswa.um.edu.my as u35818613

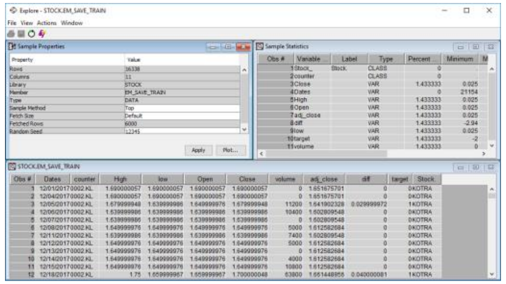
Connected to SASApp - Logical Workspace Server (odaws01-apse1.oda.sas.com)

Exploring source data

In this part, 2 property have been modified:

1. Sample Method to random
2. Fetch Size to max

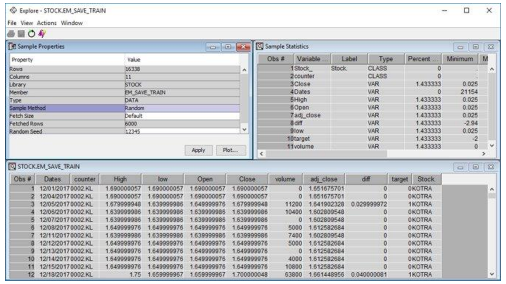
In this explore menu it shows total 11 columns and as variables available in this source data. No. of rows available is 16338.



Explore - STOCK.EM_SAVE_TRAIN

Sample Properties

Property	Value
Rows	16338
Columns	11
Library	STOCK
Member	EM_SAVE_TRAIN
Type	DATA
Sample Method	Top
Fetch Size	Default
Fetched Rows	6000
Random Seed	12345

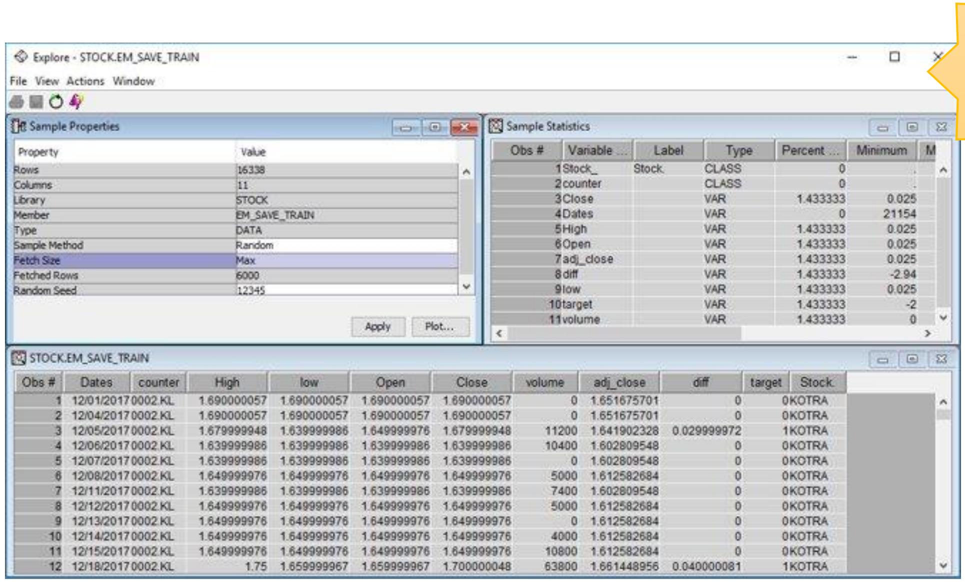


1

Explore - STOCK.EM_SAVE_TRAIN

Sample Properties

Property	Value
Rows	16338
Columns	11
Library	STOCK
Member	EM_SAVE_TRAIN
Type	DATA
Sample Method	Random
Fetch Size	Default
Fetched Rows	6000
Random Seed	12345



2

Explore - STOCK.EM_SAVE_TRAIN

Sample Properties

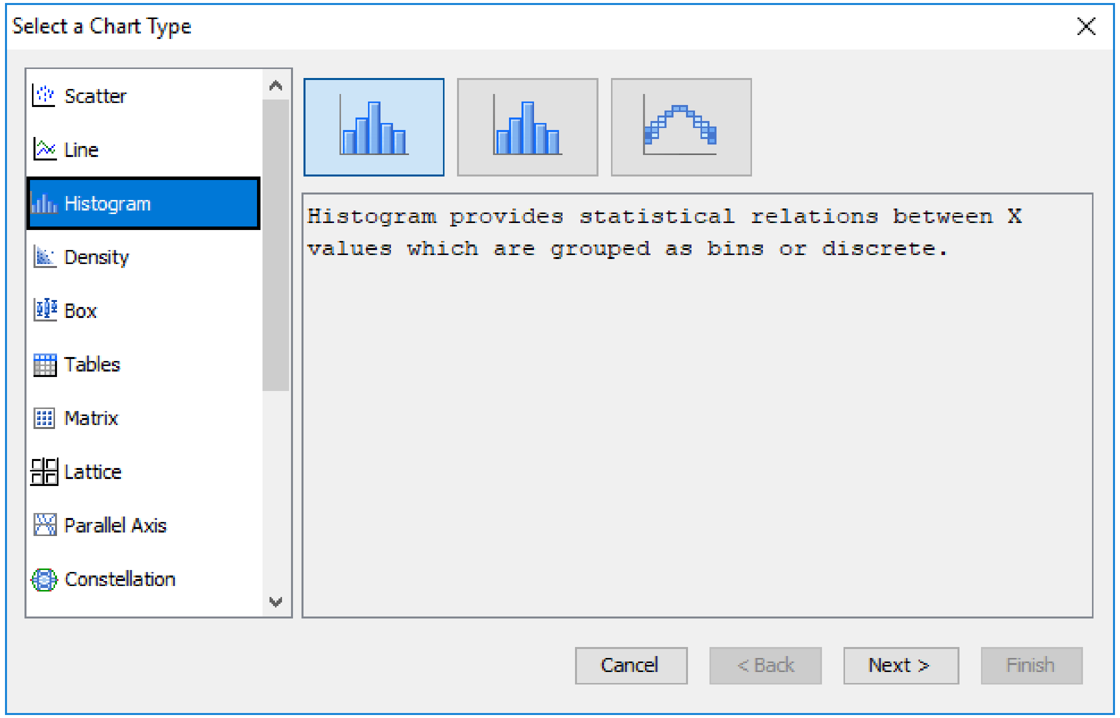
Property	Value
Rows	16338
Columns	11
Library	STOCK
Member	EM_SAVE_TRAIN
Type	DATA
Sample Method	Random
Fetch Size	Max
Fetched Rows	6000
Random Seed	12345

Sample Statistics

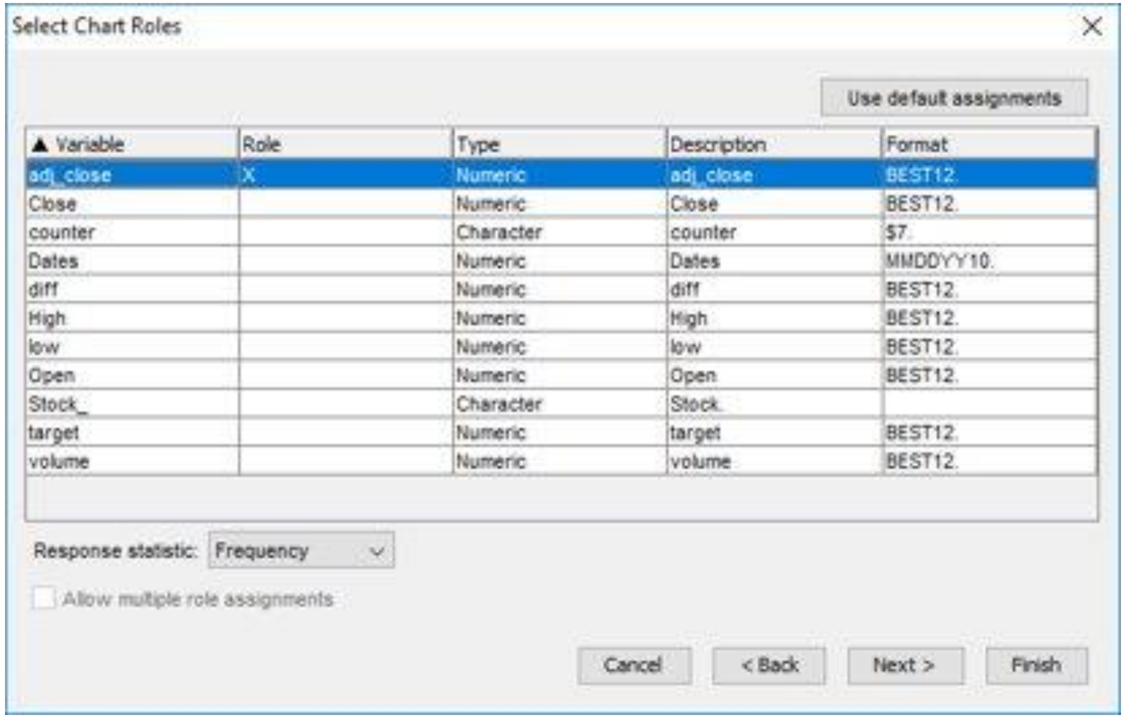
Obs #	Variable ...	Label	Type	Percent ...	Minimum	M
1	Stock_	Stock	CLASS	0	.	
2	counter		CLASS	0	.	
3	Close		VAR	1.433333	0.025	
4	Dates		VAR	0	21154	
5	High		VAR	1.433333	0.025	
6	Open		VAR	1.433333	0.025	
7	adj_close		VAR	1.433333	0.025	
8	diff		VAR	1.433333	-2.94	
9	low		VAR	1.433333	0.025	
10	target		VAR	1.433333	-2	
11	volume		VAR	1.433333	0	

STOCK.EM_SAVE_TRAIN

Obs #	Dates	counter	High	low	Open	Close	volume	adj_close	diff	target	Stock
1	12/01/2017	0002.KL	1.690000057	1.690000057	1.690000057	1.690000057	0	1.651675701	0	0	KOTRA
2	12/04/2017	0002.KL	1.690000057	1.690000057	1.690000057	1.690000057	0	1.651675701	0	0	KOTRA
3	12/05/2017	0002.KL	1.679999948	1.639999986	1.649999976	1.679999948	11200	1.641902328	0.029999972	1	KOTRA
4	12/06/2017	0002.KL	1.639999986	1.639999986	1.639999986	1.639999986	10400	1.602809548	0	0	KOTRA
5	12/07/2017	0002.KL	1.639999986	1.639999986	1.639999986	1.639999986	0	1.602809548	0	0	KOTRA
6	12/08/2017	0002.KL	1.649999976	1.649999976	1.649999976	1.649999976	5000	1.612582684	0	0	KOTRA
7	12/11/2017	0002.KL	1.639999986	1.639999986	1.639999986	1.639999986	7400	1.602809548	0	0	KOTRA
8	12/12/2017	0002.KL	1.649999976	1.649999976	1.649999976	1.649999976	5000	1.612582684	0	0	KOTRA
9	12/13/2017	0002.KL	1.649999976	1.649999976	1.649999976	1.649999976	0	1.612582684	0	0	KOTRA
10	12/14/2017	0002.KL	1.649999976	1.649999976	1.649999976	1.649999976	4000	1.612582684	0	0	KOTRA
11	12/15/2017	0002.KL	1.649999976	1.649999976	1.649999976	1.649999976	10800	1.612582684	0	0	KOTRA
12	12/18/2017	0002.KL	1.75	1.659999967	1.659999967	1.700000048	63800	1.661448956	0.040000081	1	KOTRA



```
Histogram provides statistical relations between X
values which are grouped as bins or discrete.
```



▲ Variable	Role	Type	Description	Format
adj_close	X	Numeric	adj_close	BEST12.
Close		Numeric	Close	BEST12.
counter		Character	counter	$7.
Dates		Numeric	Dates	MMDDYY10.
diff		Numeric	diff	BEST12.
High		Numeric	High	BEST12.
low		Numeric	low	BEST12.
Open		Numeric	Open	BEST12.
Stock_		Character	Stock	
target		Numeric	target	BEST12.
volume		Numeric	volume	BEST12.

Response statistic: Frequency

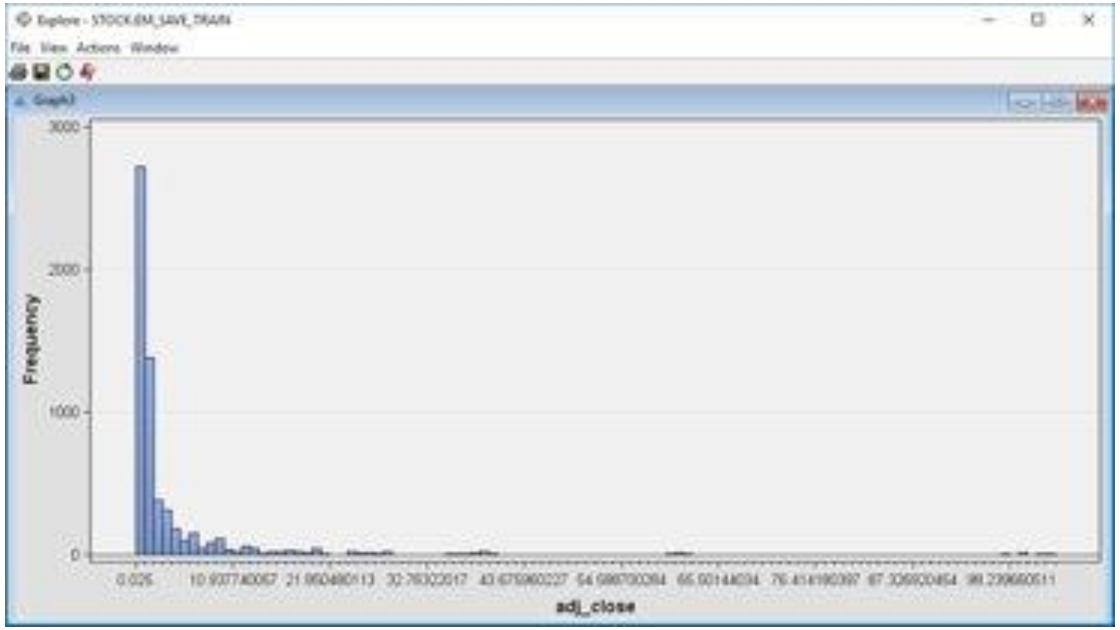
□ Allow multiple role assignments

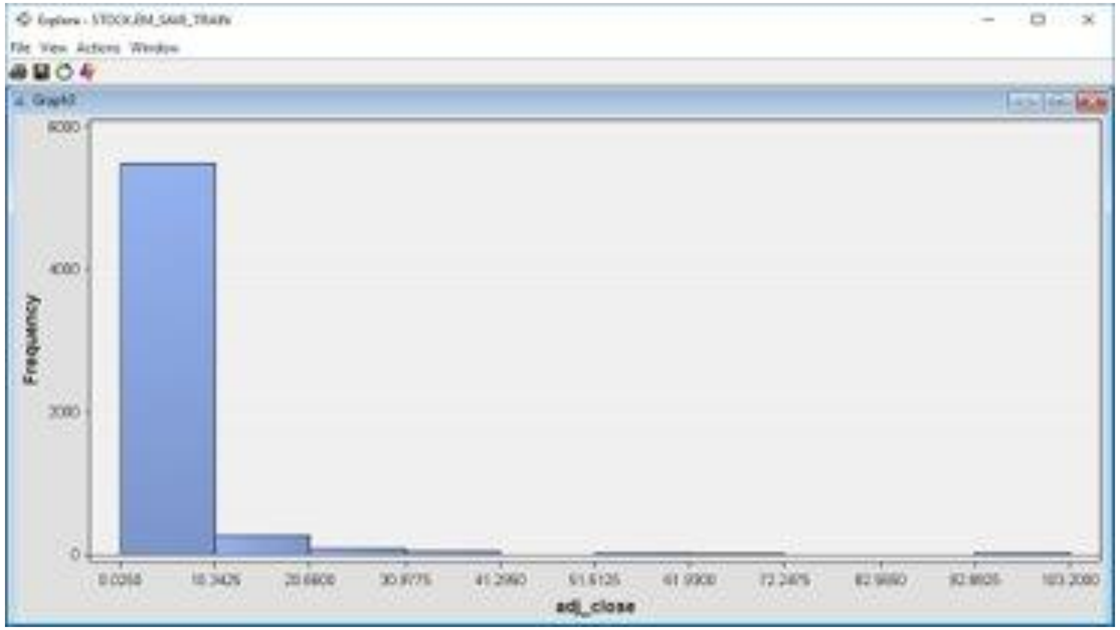
Creating a Histogram for a Single Variable – adj_close

- Set adj_close as Role X

Histogram shows:

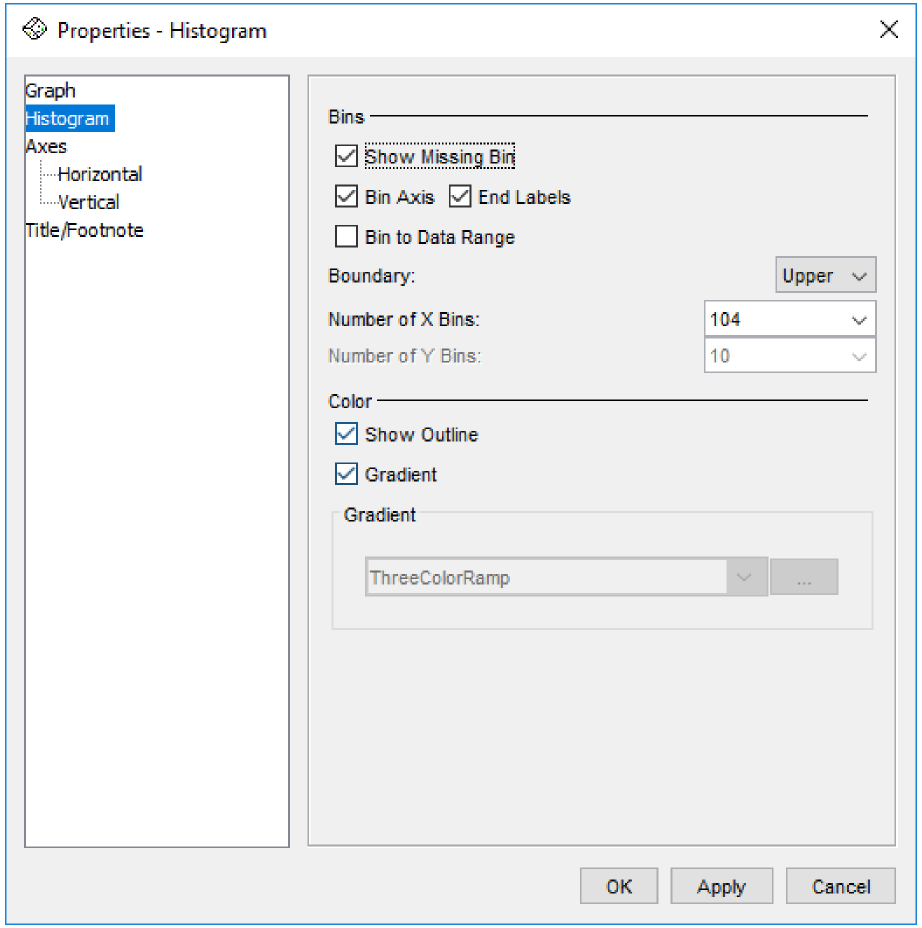
- The minimum value is 0.025
- The maximum value is 103.2





Left histogram is a default view while right histogram is an individual value after setting to histogram property

Adding a “Missing” Bin to a Histogram



Bins

☑ Show Missing Bin

☑ Bin Axis ☑ End Labels

☐ Bin to Data Range

Boundary: Upper

Number of X Bins: 104

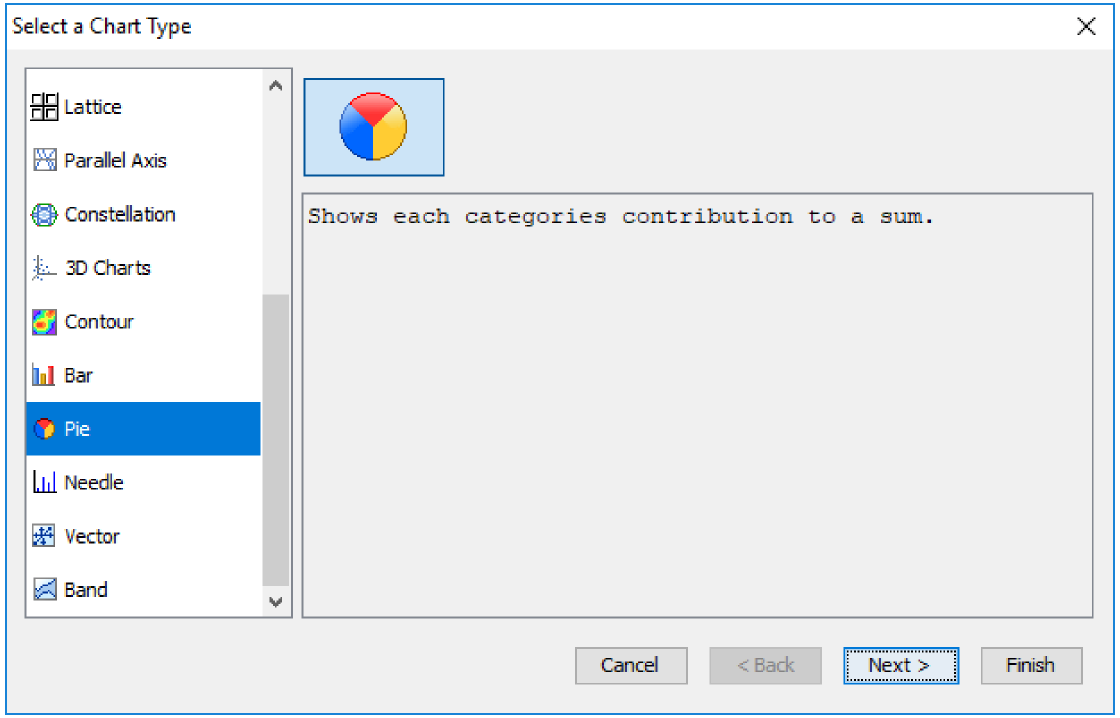
Number of Y Bins: 10

Color

☑ Show Outline

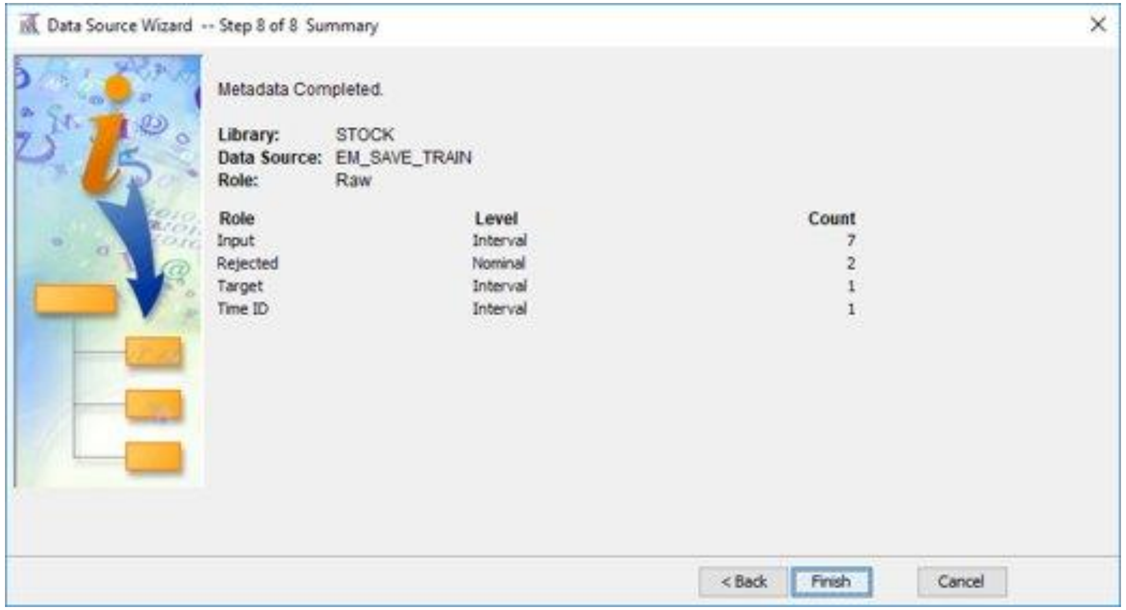
☑ Gradient

Gradient: ThreeColorRamp



Select a Chart Type

Shows each categories contribution to a sum.



Data Source Wizard -- Step 8 of 8 Summary

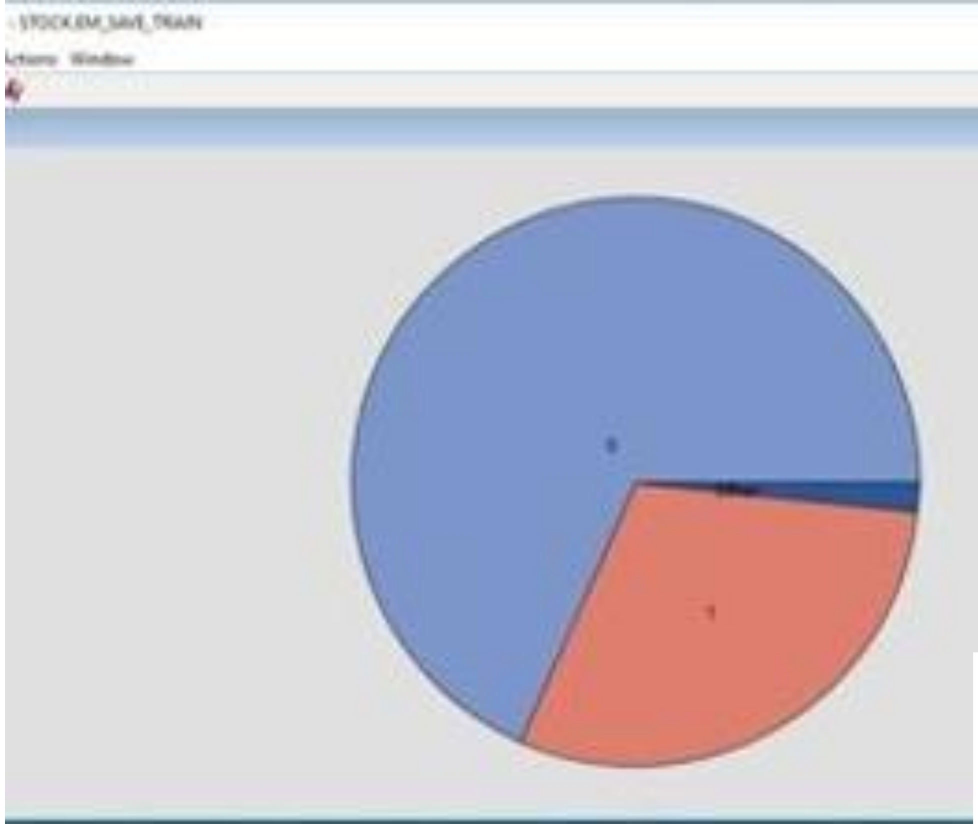
Metadata Completed.

Library: STOCK
Data Source: EM_SAVE_TRAIN
Role: Raw

Role	Level	Count
Input	Interval	7
Rejected	Nominal	2
Target	Interval	1
Time ID	Interval	1

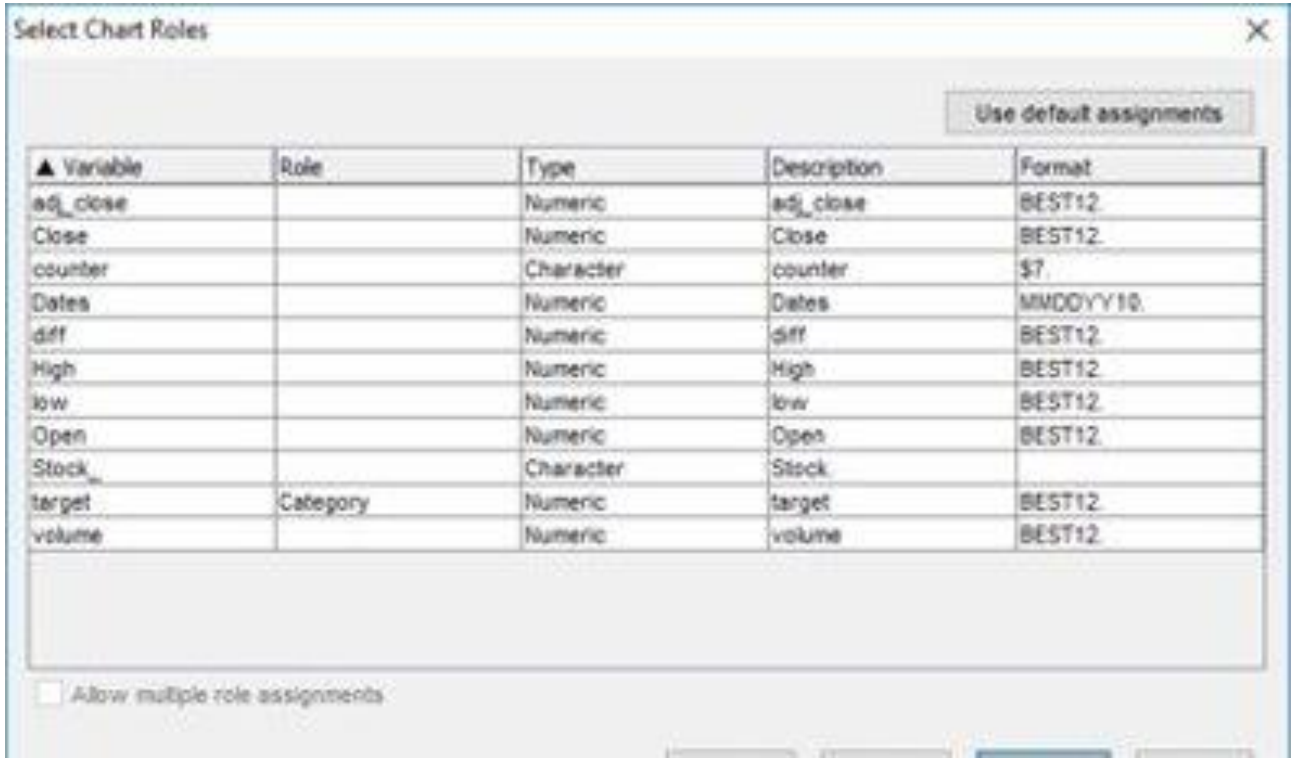
Adding Plots to the Explore Window

- Create Pie chart



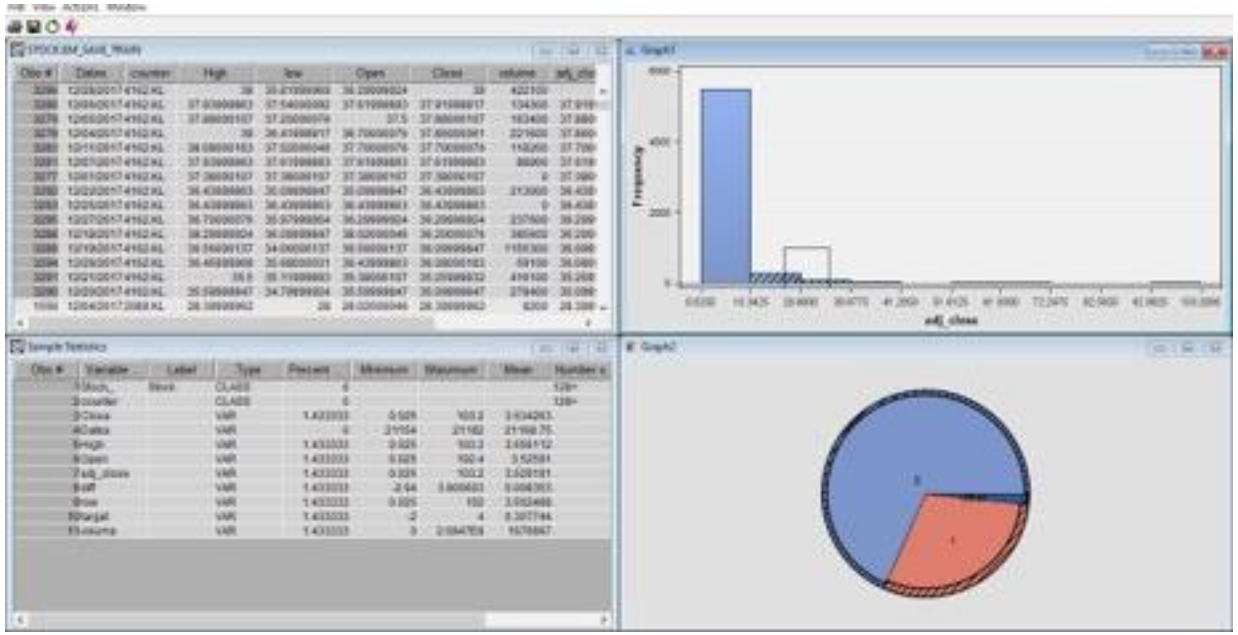
This step require to set Target Role as Category

As a results, Pie chart shows (60:40):
Target 0 = 4098 frequency
Target 1 = 1799 frequency



▲ Variable	Role	Type	Description	Format
adj_close		Numeric	adj_close	BEST12.
Close		Numeric	Close	BEST12.
counter		Character	counter	$7.
Dates		Numeric	Dates	MMDDYY10.
diff		Numeric	diff	BEST12.
High		Numeric	High	BEST12.
low		Numeric	low	BEST12.
Open		Numeric	Open	BEST12.
Stock		Character	Stock	
target	Category	Numeric	target	BEST12.
volume		Numeric	volume	BEST12.

□ Allow multiple role assignments



Sample Statistics

Obs #	Variable	Label	Type	Percent	Minimum	Maximum	Mean	Number o
	1 Stock_	Stock	CLASS	0				128+
	2 counter		CLASS	0				128+
	3 Close		VAR	1.433333	0.525	103.2	3.634263	
	4 Dates		VAR	0	21154	21182	21168.75	
	5 High		VAR	1.433333	0.525	103.2	3.658112	
	6 Open		VAR	1.433333	0.525	102.4	3.52581	
	7 adj_close		VAR	1.433333	0.525	103.2	3.628181	
	8 diff		VAR	1.433333	-2.94	3.800003	0.006353	
	9 low		VAR	1.433333	0.525	102	3.592488	
	10 range		VAR	1.433333	-2	4	0.397744	
	11 volume		VAR	1.433333	0	2.584759	1578867	

Exploring Variable Associations

- This shows all view about this data and how the variable associates. As an example, when point selected to histogram, pie chart and table shows same proportion to its related point.