

SE634: Artificial Intelligence in Industrial Engineering

Data preprocessing and introduction to RapidMiner Studio

Introduction to RapidMiner Studio



RAPIDMINER

Gartner®

<https://rapidminer.com/>

<https://my.rapidminer.com/nexus/account/index.html#signup>

- ✿ RapidMiner is a Visionary in the 2021 Gartner Magic Quadrant for Data Science and Machine Learning Platforms

Home Screen

 **LEARN**

 **NEW PROCESS**

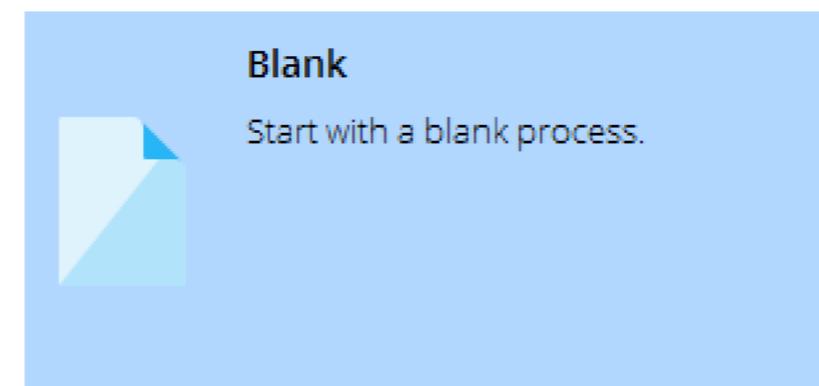
 **OPEN PROCESS**

 **rapidminer News**

New Keras Extension

Keras is a high level neural network API, supporting popular deep learning libraries like Tensorflow, Microsoft Cognitive Toolkit, and Theano.

Choose a template to start from:



Blank

Start with a blank process.



Churn Modeling

Predict which of your customers will churn and why with a decision tree.



Direct Marketing

Predict response to campaigns and increase the conversion rate of your campaign.



Credit Risk Modeling

Model credit default risk by training an optimized Support Vector Machine (SVM) model.

Market Basket Analysis

Find associations between items.

Predictive Maintenance

Model equipment failure to

Interface

The screenshot shows the RapidMiner Studio Educational 8.0.001 interface. The window title is "<new process> – RapidMiner Studio Educational 8.0.001 @ DESKTOP-6J6ID3N". The menu bar includes File, Edit, Process, View, Connections, Cloud, Settings, and Extensions. The toolbar contains icons for Add Data, Repository, Operators, and Process. The Views dropdown is set to "Design". The main area displays a process canvas with the message "Your process looks empty. Add some data first. Drag data or operators here." A bottom banner encourages using the Wisdom of Crowds to get operator recommendations. The right side features a Parameters panel with "logverbosity" set to "init" and "logfile" as a folder path, and a Help panel for the Process operator. The Operators panel lists categories like Data Access, Blending, Cleansing, Modeling, Scoring, Validation, Utility, and Extensions. The Repository panel shows connections to Samples, DB, Local Repository (warut), and Cloud Repository (disconnected). The status bar at the bottom right shows page number 4.

File Edit Process View Connections Cloud Settings Extensions

Add Data Repository Operators Process

Views: Design Results

Need help?

Repository

- Samples
- DB
- Local Repository (warut)
- Cloud Repository (disconnected)

Process

Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Activate Wisdom of Crowds

Parameters

Process

- logverbosity: init
- logfile: [Folder icon]

Show advanced parameters

Change compatibility (8.0.001)

Help

Process

RapidMiner Studio Core

Synopsis

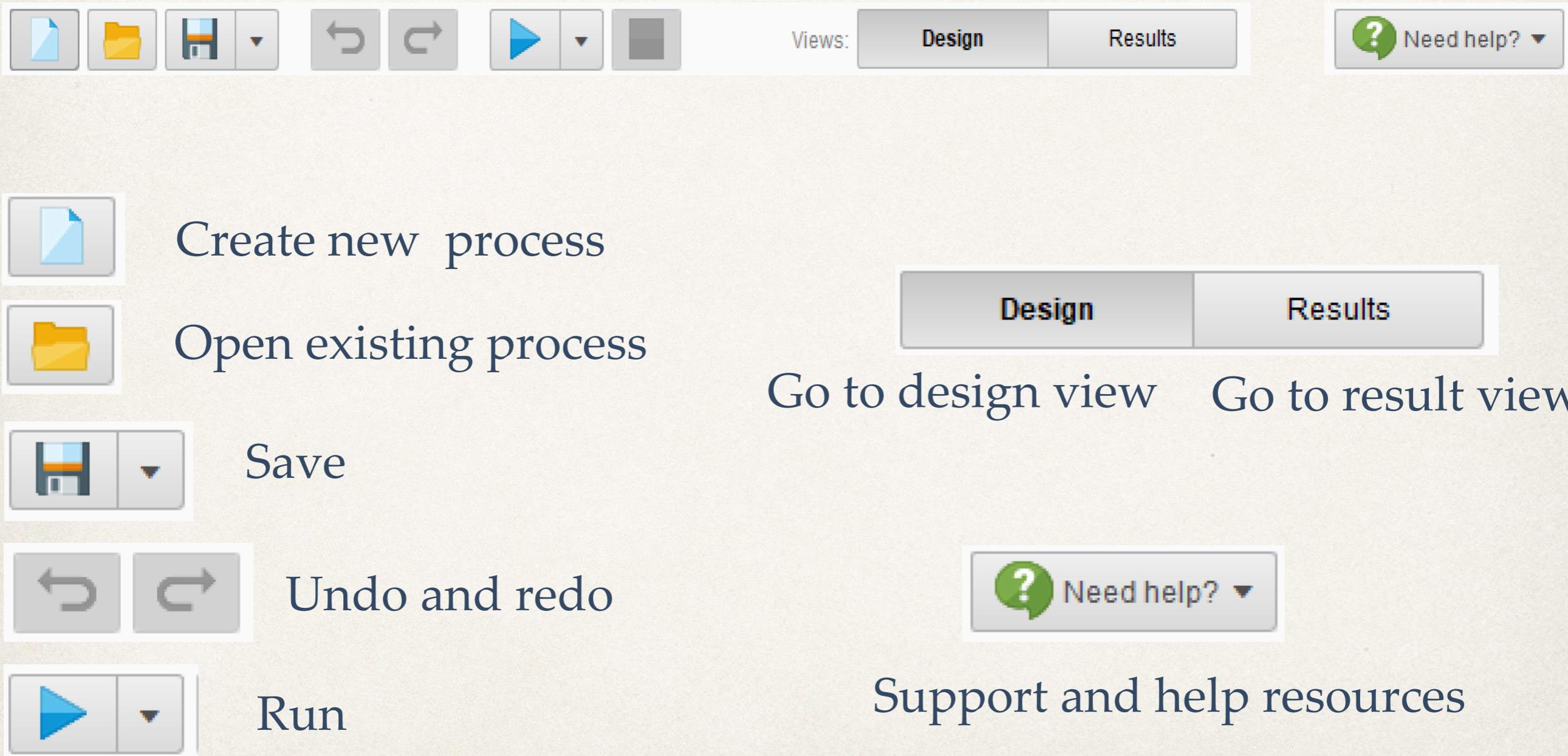
The root operator which is the outer most operator of every process.

Description

Get more operators from the Marketplace

4

Menu



Data Pre-Processing with RapidMiner

- ❖ Data management
- ❖ Import data
- ❖ Data exploration
- ❖ Data preparation

Data

- * Data is presented in table format
 - * Row = Example
 - * Column = Attribute that has three typical roles
 - * ID = Identification or primary key
 - * Attribute = Feature or independent variable
 - * Label = Class or dependent variable

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE		17546	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO



Value Type

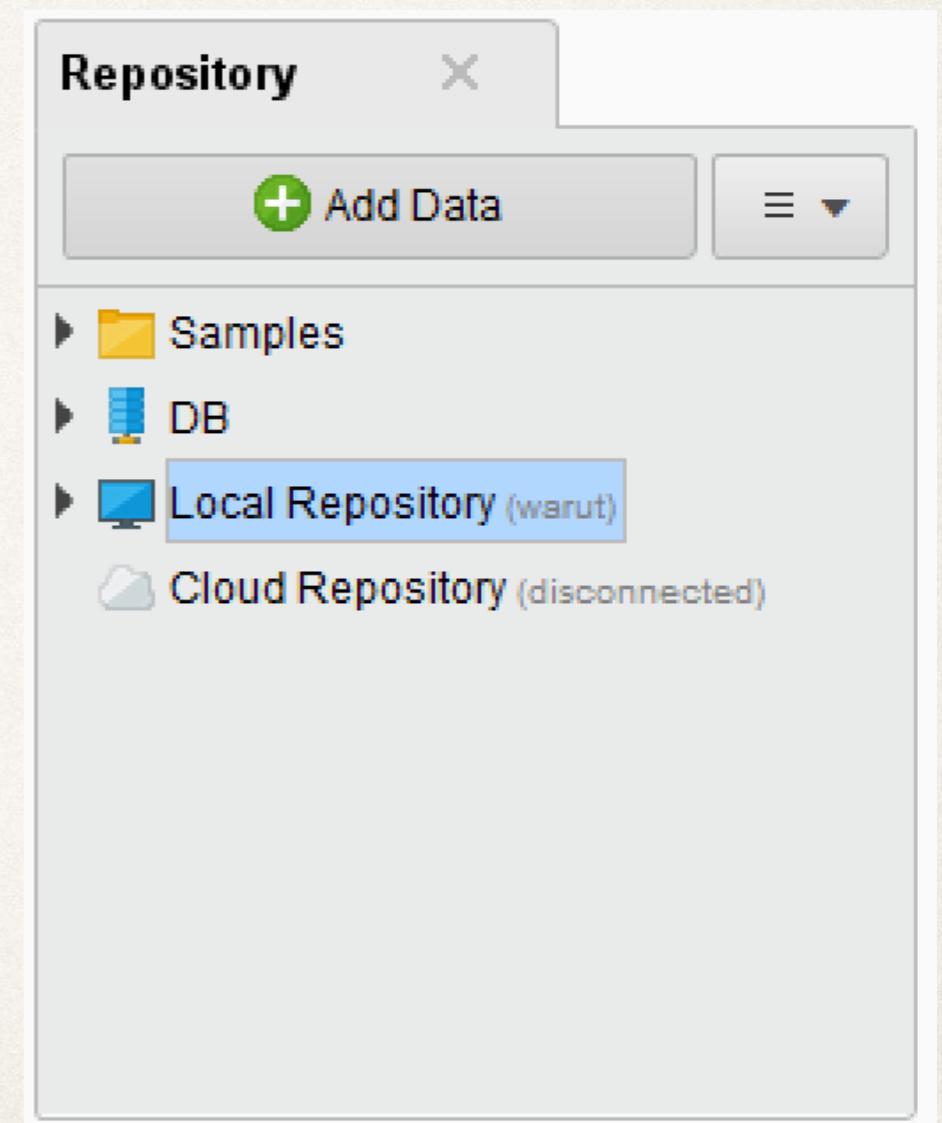
- ✿ Polynomial = Categorical data (>two categories)
- ✿ Binomial = Categorical data (two categories)
- ✿ Numeric or integer
- ✿ Text

customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE		17546	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO

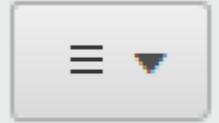
Polynomial Numeric Binomial Numeric Binomial Numeric Binomial

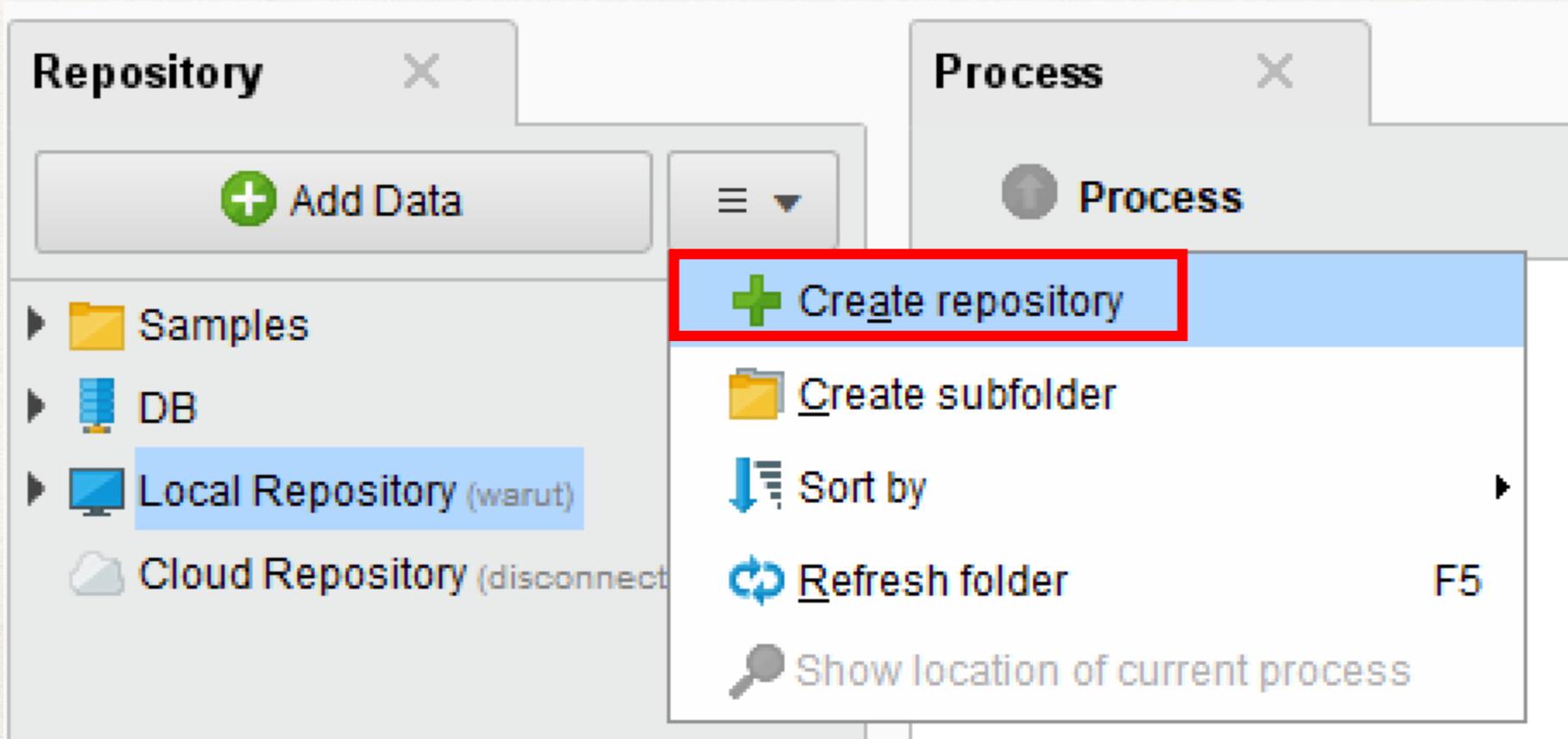
Data Management

- ❖ Repository
 - ❖ Storage area for data and processes
 - ❖ No need to load the file for each experimental run
- ❖ Components in repository
 - ❖ Add new repository
 - ❖ Process sample
 - ❖ Data and processes in repository



Data Management

- ✿ Create new repository
- ✿ Click 
- ✿ Choose “Create repository”



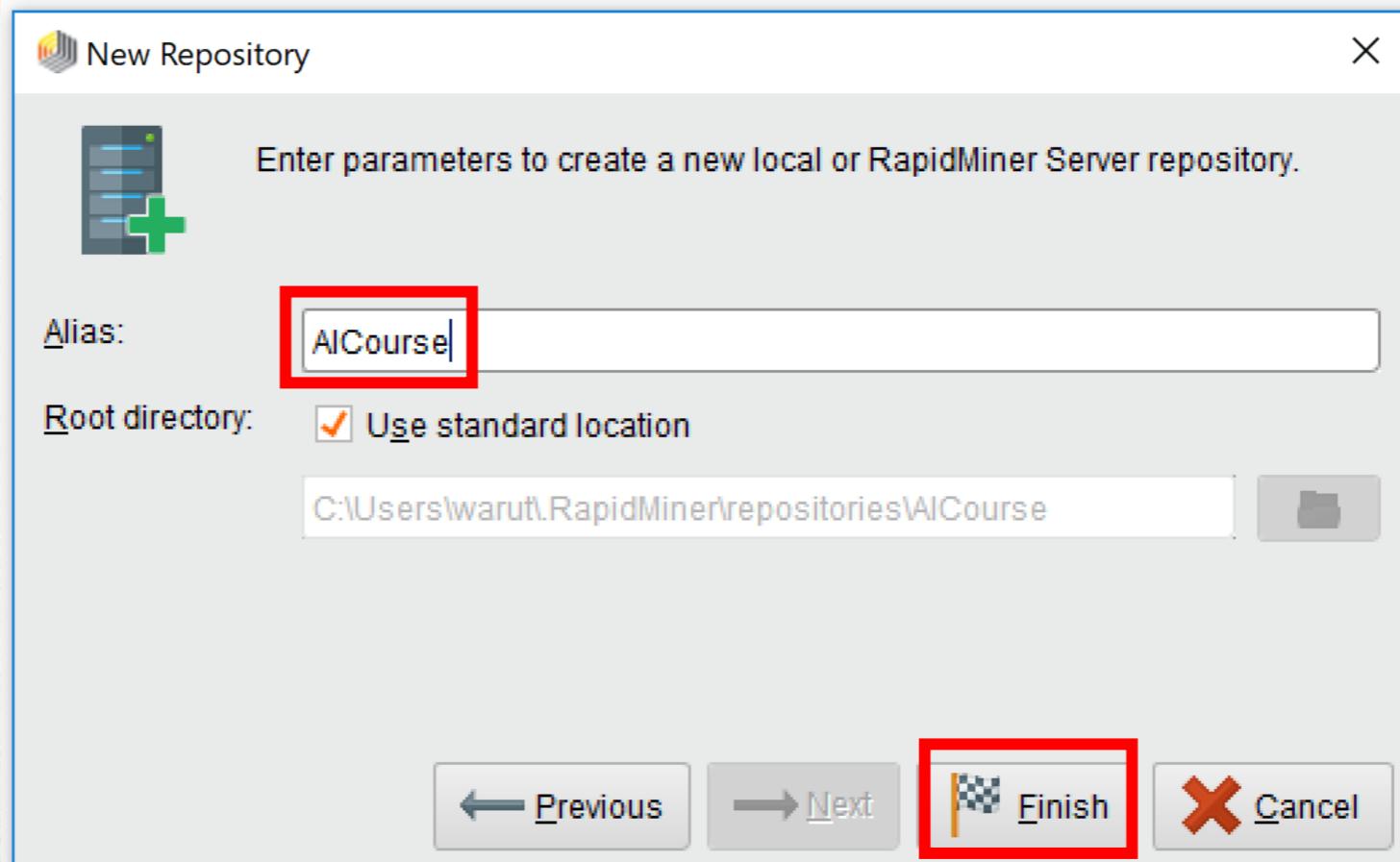
Data Management

- ✿ Create new repository (Cont.)
 - ✿ Choose “New local repository”
 - ✿ Click “Next”



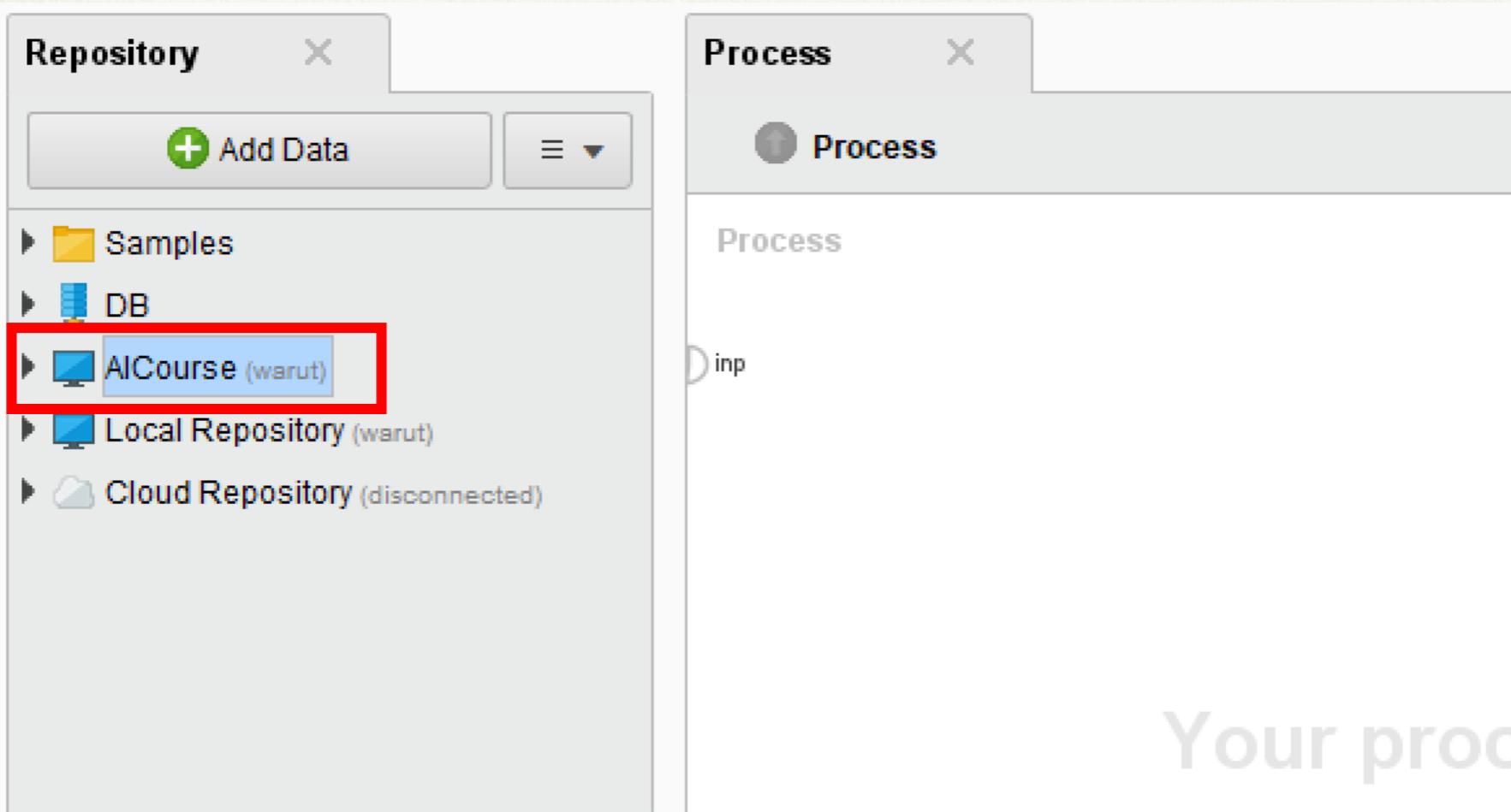
Data Management

- * Create new repository (Cont.)
 - * Change Alias to “AI Course”
 - * Press “Finish”



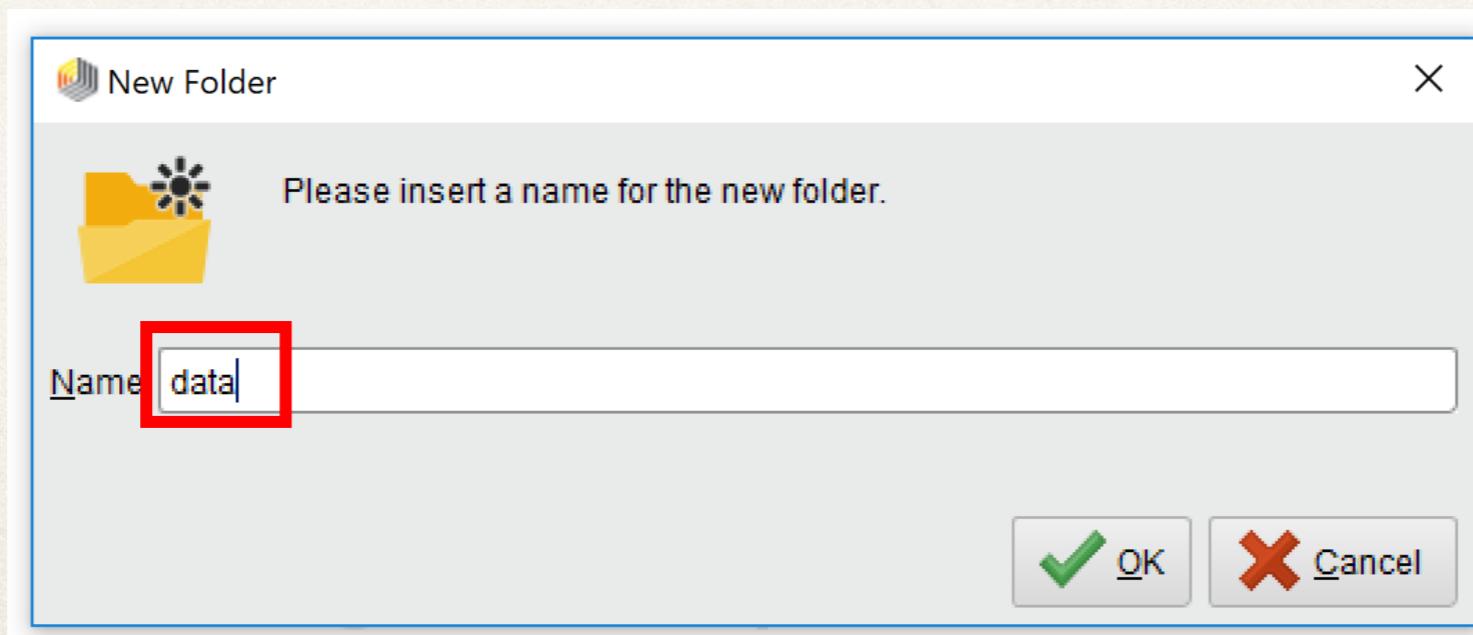
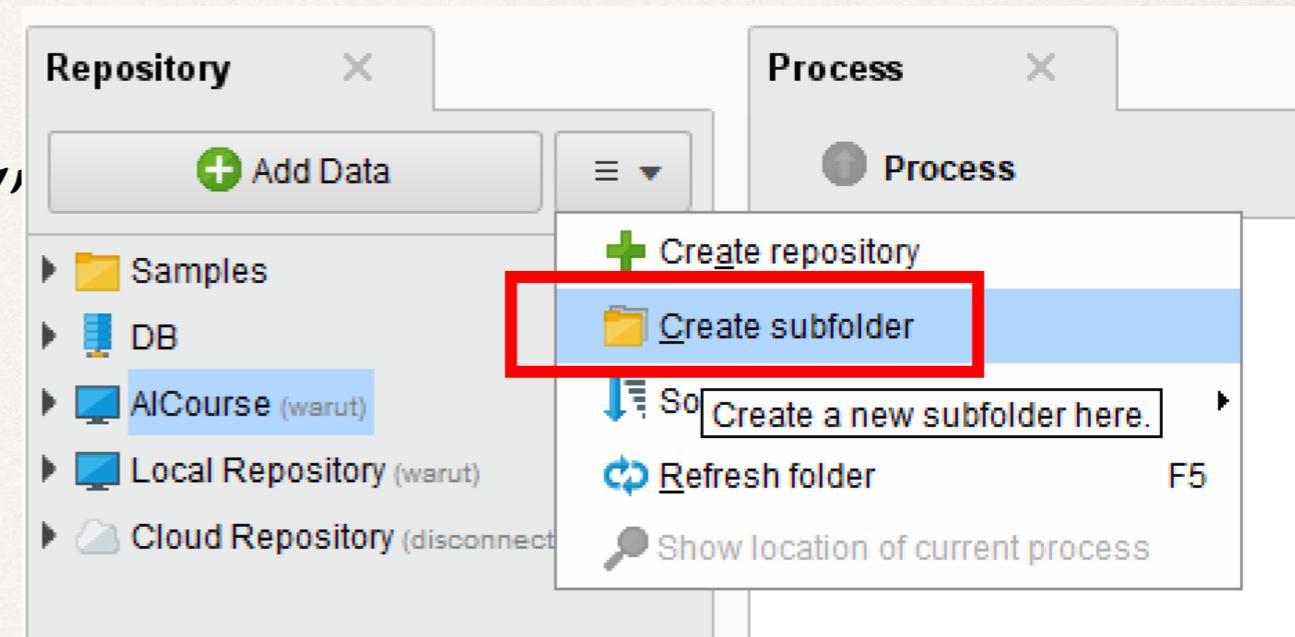
Data Management

- ✿ “AICourse” will be appeared in repository.



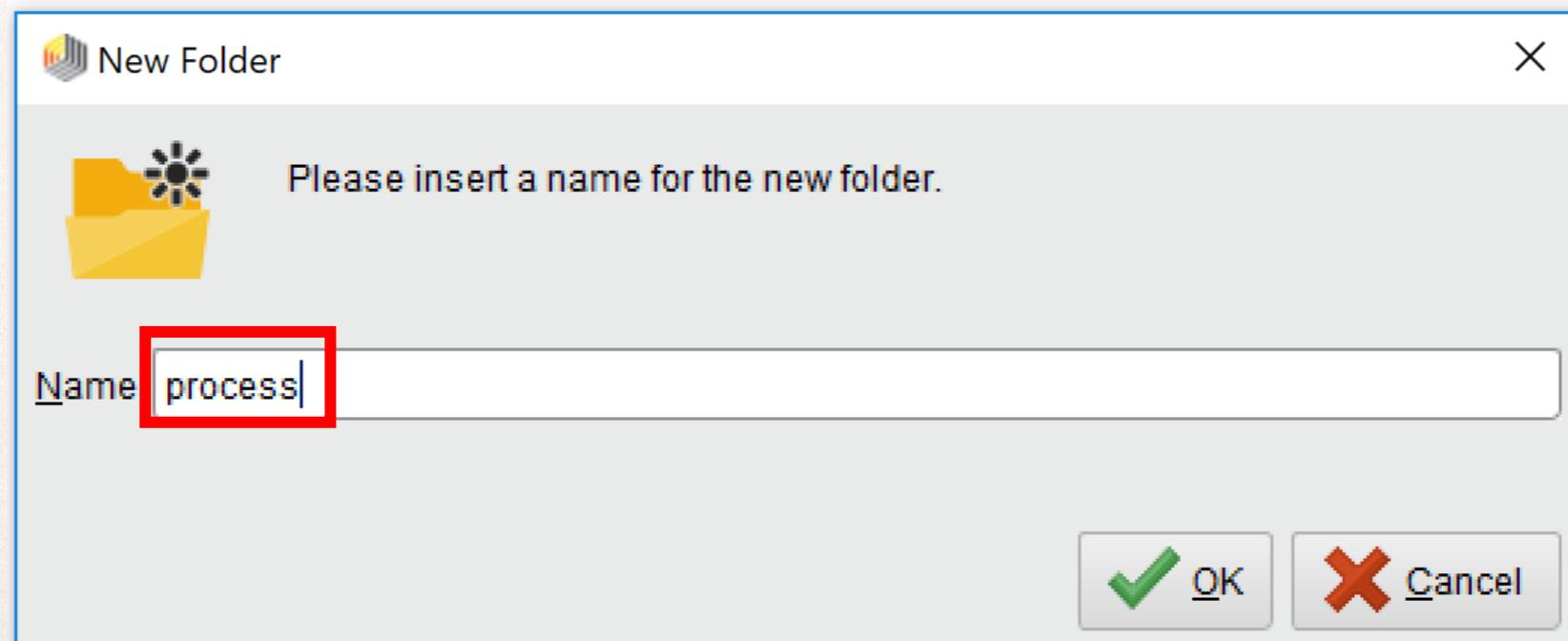
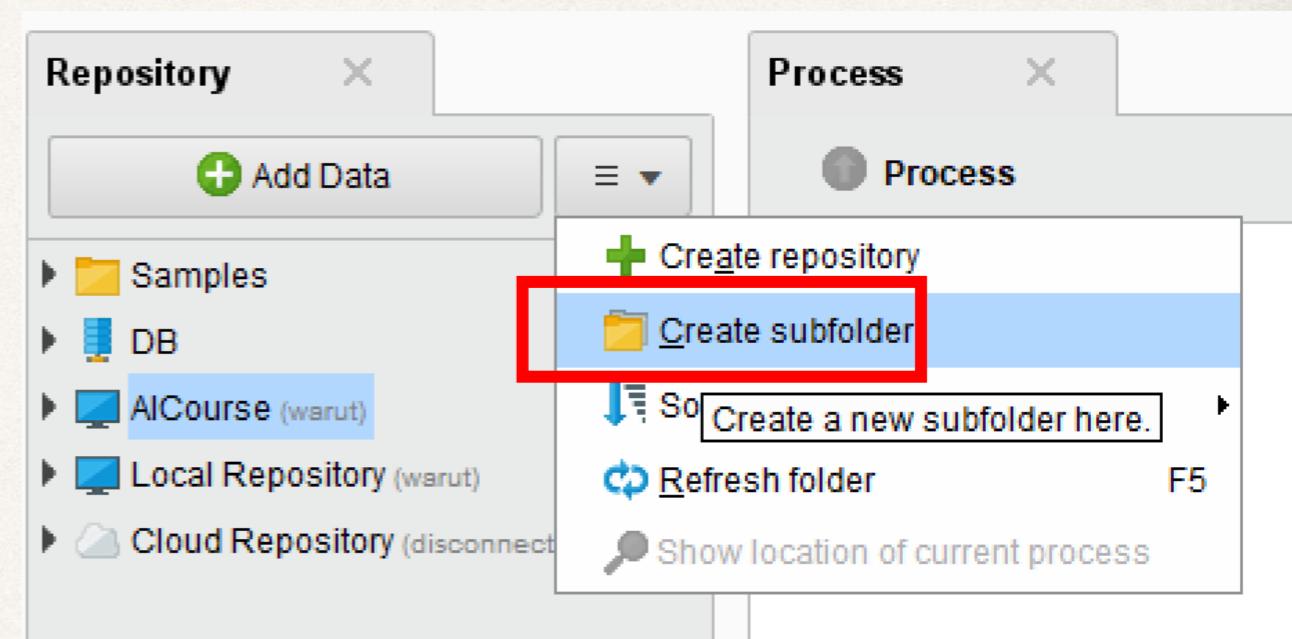
Data Management

- ✿ Create folder in “AI Course”
 - ✿ “data” for data storage



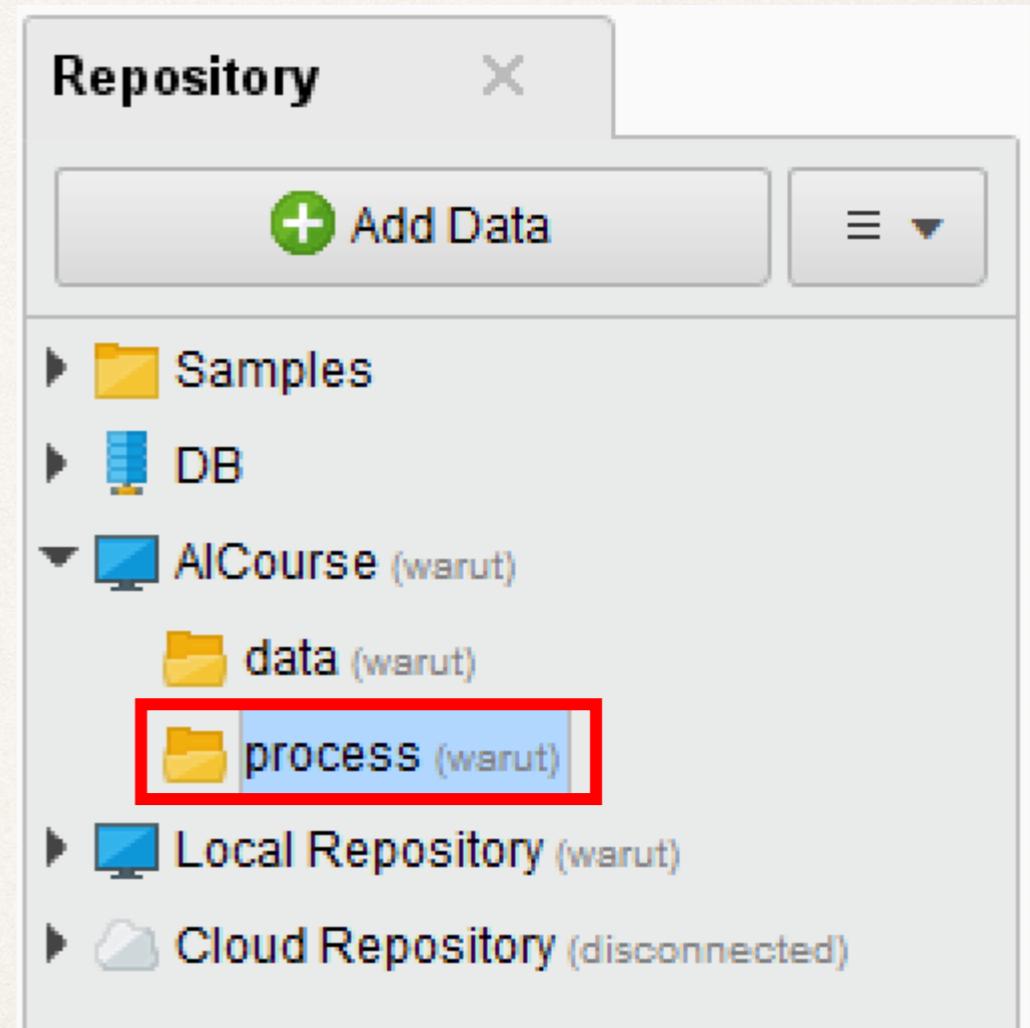
Data Management

- ✿ Create folder in “AI Course”
 - ✿ “data” for data storage
 - ✿ “process” for processes storage



Data Management

- ✿ Create folder in “AICourse”
 - ✿ “data” for data storage
 - ✿ “process” for processes storage

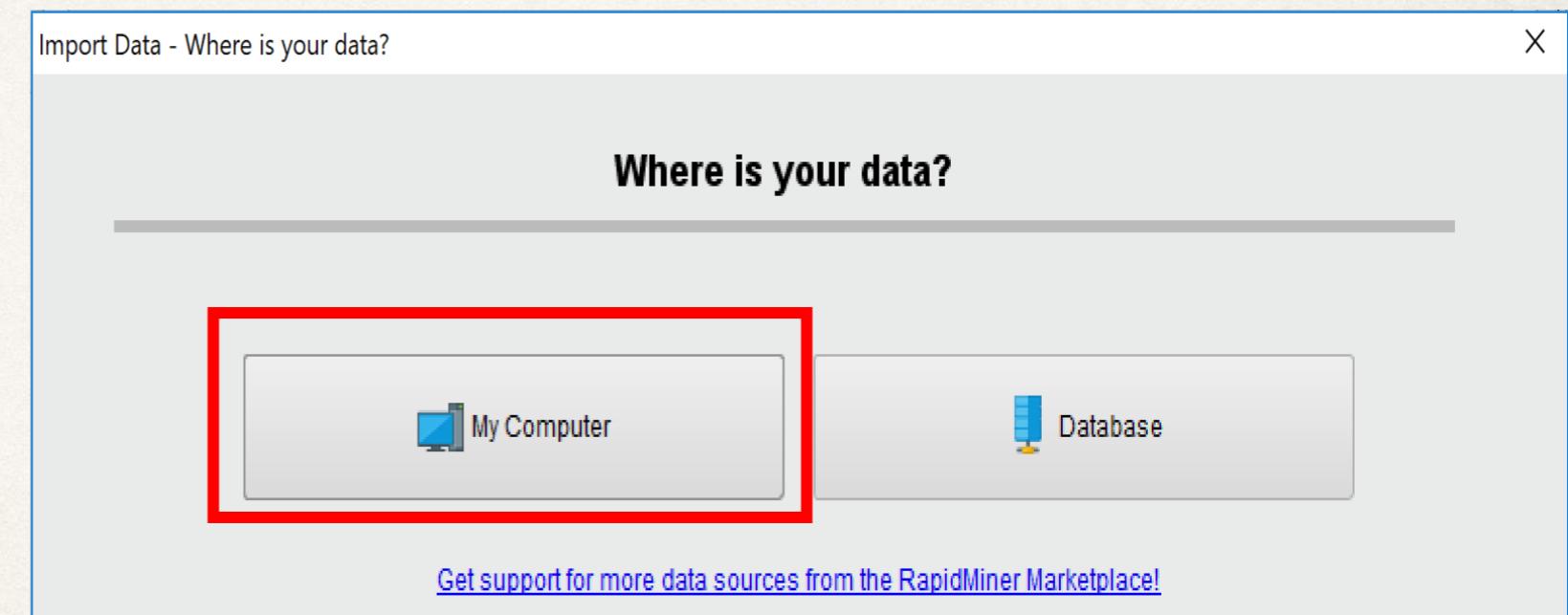
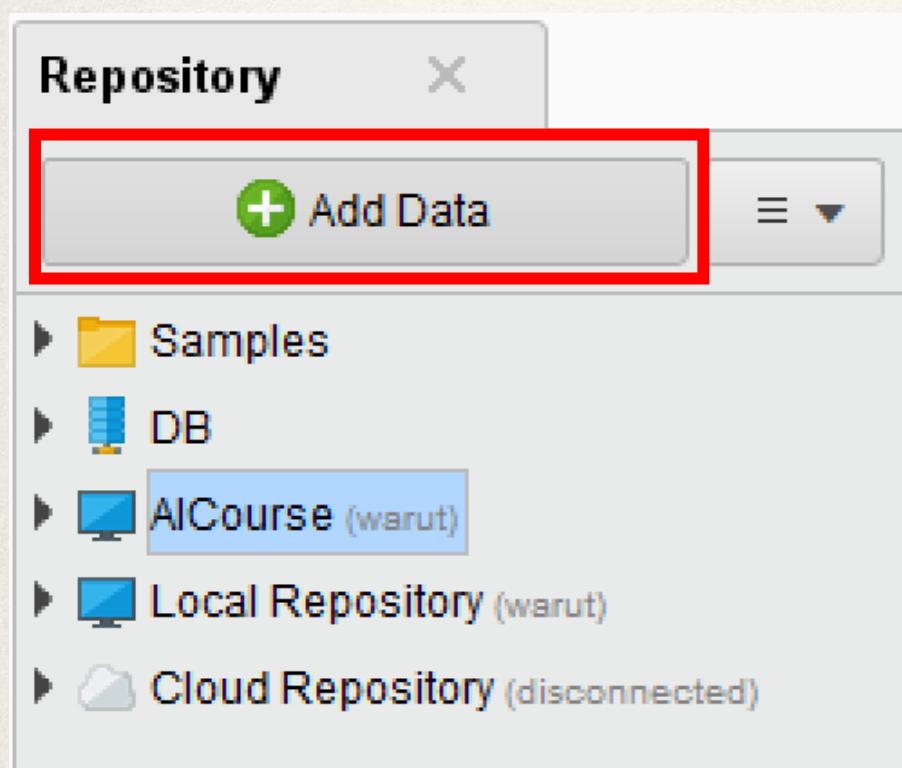


Data Pre-Processing with RapidMiner

- ✿ Data management
- ✿ **Import data**
- ✿ Data exploration
- ✿ Data preparation

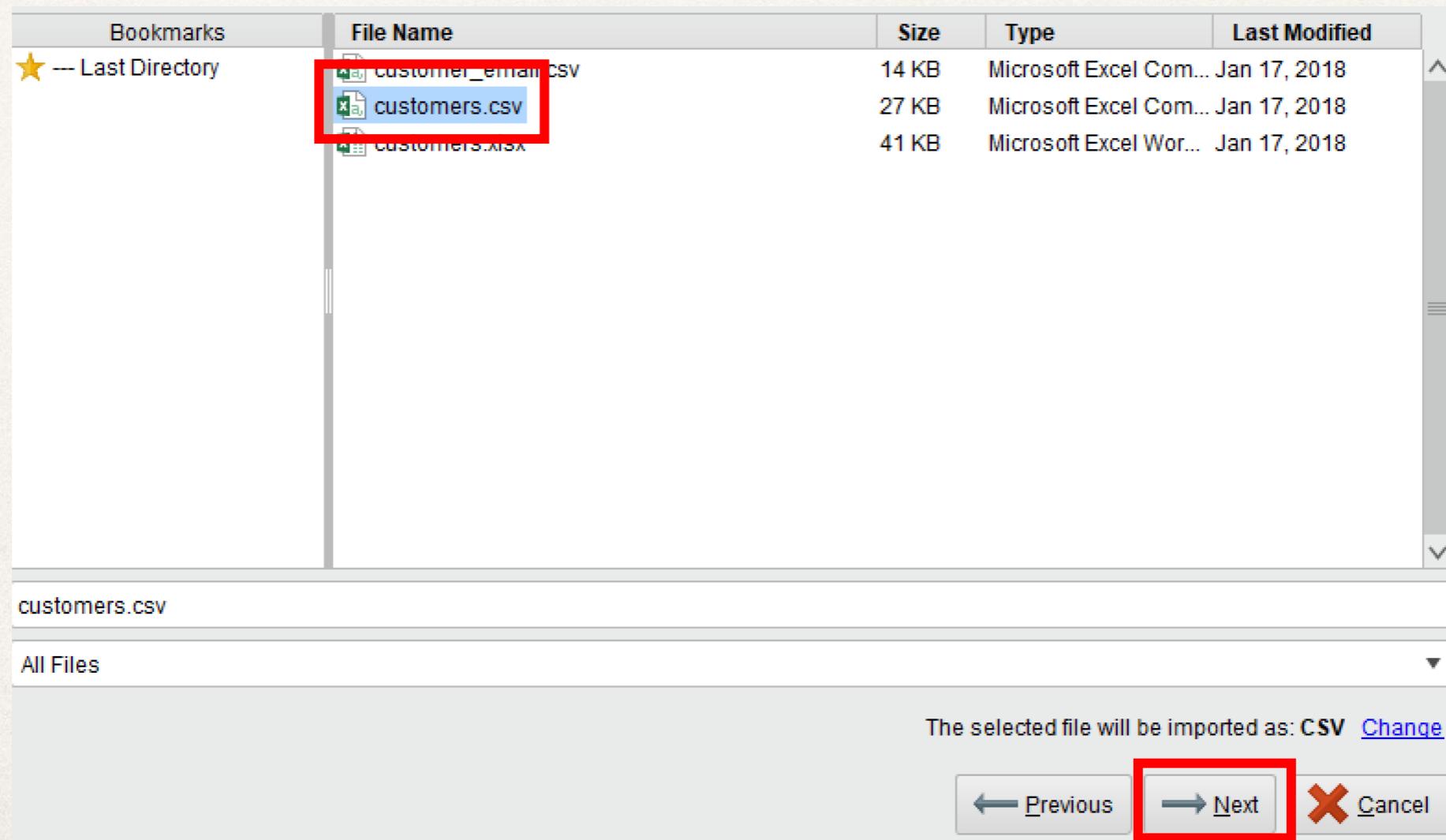
Import Data into Repository

- * Click 
- * Choose “My Computer”



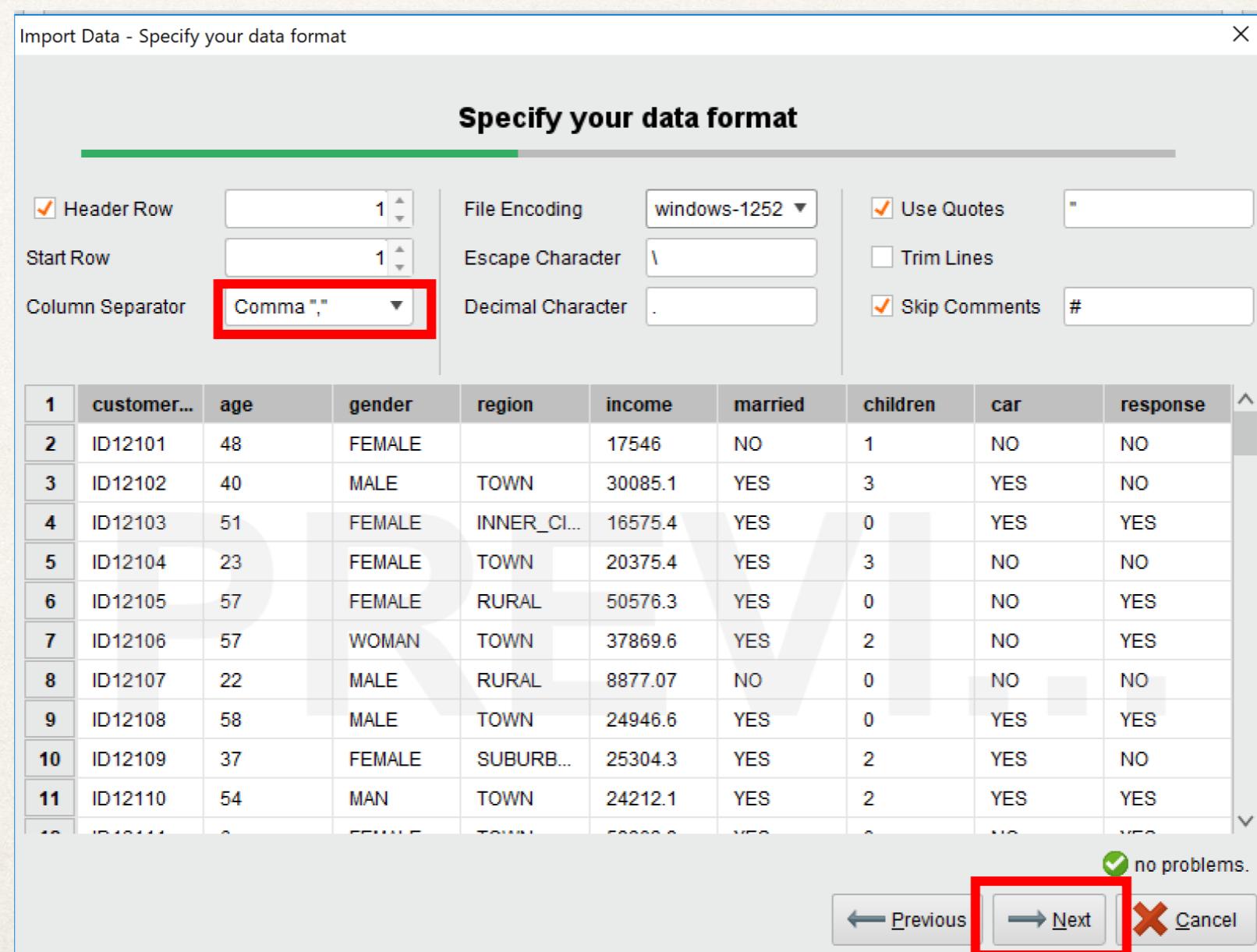
Import Data into Repository

- Choose “customer.csv”



Import Data into Repository

- For Column Separator, choose Comma “,”



Import Data into Repository

- ❖ Change role of `customer_id` to `id`

Import Data - Format your columns.

Format your columns.

Date format: MMM d, yyyy h:mm:ss a z ▾ Replace errors with missing values ⓘ

	customer_id	age	gender	region	income	married	c
1	ID12101	23	FEMALE	?	17546.000	NO	ir
2	ID12102	57	FEMALE	RURAL	30085.100	YES	ir
3	ID12103	57	WOMAN	TOWN	16575.400	NO	ir
4	ID12104	22	MALE	RURAL	20375.400	NO	ir
5	ID12105	58	MALE	TOWN	37869.600	NO	ir
6	ID12106	37	FEMALE	SUBURBAN	8877.070	NO	ir
7	ID12107	37	FEMALE	TOWN	24946.600	NO	ir
8	ID12108	54	MAN	TOWN	25304.300	NO	ir
9	ID12109	6	FEMALE	TOWN	24212.100	NO	ir
10	ID12110	52	FEMALE	?	59803.900	NO	ir
11	ID12111	52	FEMALE	?	26658.800	NO	ir
12	ID12112	52	FEMALE	?	NO	NO	ir

Change Role Change Type Rename column Exclude column Opens a dialog to change the role. ⓘ

Change role: Please enter the new role:

OK Cancel

no problems.

Previous Next Cancel

Import Data into Repository

- ❖ Change role of response to label

Import Data - Format your columns.

Format your columns.

Date format: MMM d, yyyy h:mm:ss a z ▾ Replace errors with missing values ⓘ

	region	income	married	children	car	response
1	?	17546.000	NO	1	NO	NO
2	TOWN	30085.100	YES	3	YES	NO
3	INNER_CITY	16575.400	YES	0	YES	YES
4	TOWN	20375.400	YES	3	NO	...
5	RURAL	50576.300	YES	0	NO	...
6	TOWN	37869.600	YES	2	NO	...
7	RURAL	8877.070	NO	0	NO	...
8	TOWN	24946.600	YES	0	YES	...
9	SUBURBAN	25304.300	YES	2	YES	...
10	TOWN	24212.100	YES	2	ES	label
11	TOWN	59803.900	YES	0	ES	...
12	?	26658.800	NO	0	YES	...

Change Type ▾

- Change Role
- Rename column
- Exclude column

Change role

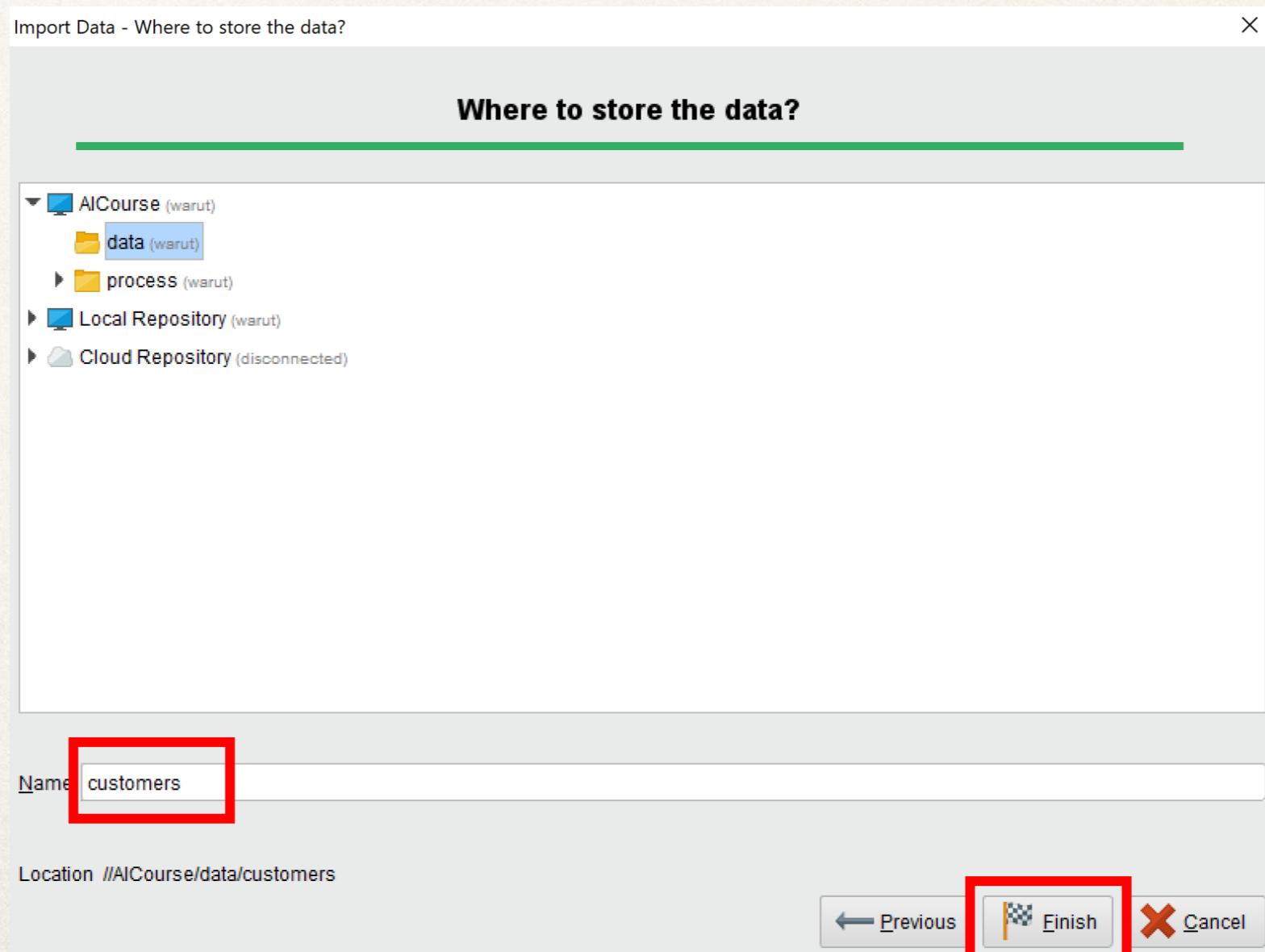
Please enter the new role:

label

OK Cancel Previous Next Cancel

Import Data into Repository

- ✿ Name = **customers**
- ✿ Save into folder **data** under **AICourse**



Import Data into Repository

- ✿ Click at header of each column to sort the data

Result History ExampleSet (/AI.Course/data/customers) X

Open in Turbo Prep Auto Model Filter (600 / 600 examples): all ▾

Data Statistics Visualizations Annotations

Row No.	customer_id	response	age	gender	region	income	married	children	car
1	ID12101	NO	100	FEMALE	?	17546	NO	1	NO
2	ID12102	NO	40	MALE	TOWN	30085.100	YES	3	YES
3	ID12103	YES	51	FEMALE	INNER_CITY	16575.400	YES	0	YES
4	ID12104	NO	23	FEMALE	TOWN	20375.400	YES	3	NO
5	ID12105	YES	57	FEMALE	RURAL	50576.300	YES	0	NO
6	ID12106	YES	57	WOMAN	TOWN	37869.600	YES	2	NO
7	ID12107	NO	22	MALE	RURAL	8877.070	NO	0	NO
8	ID12108	YES	58	MALE	TOWN	24946.600	YES	0	YES
9	ID12109	NO	37	FEMALE	SUBURBAN	25304.300	YES	2	YES
10	ID12110	YES	54	MAN	TOWN	24212.100	YES	2	YES
11	ID12111	YES	6	FEMALE	TOWN	59803.900	YES	0	NO
12	ID12112	YES	52	FEMALE	?	26658.800	NO	0	YES
13	ID12113	YES	44	FEMALE	TOWN	15735.800	YES	1	NO
...

Import Data into Repository

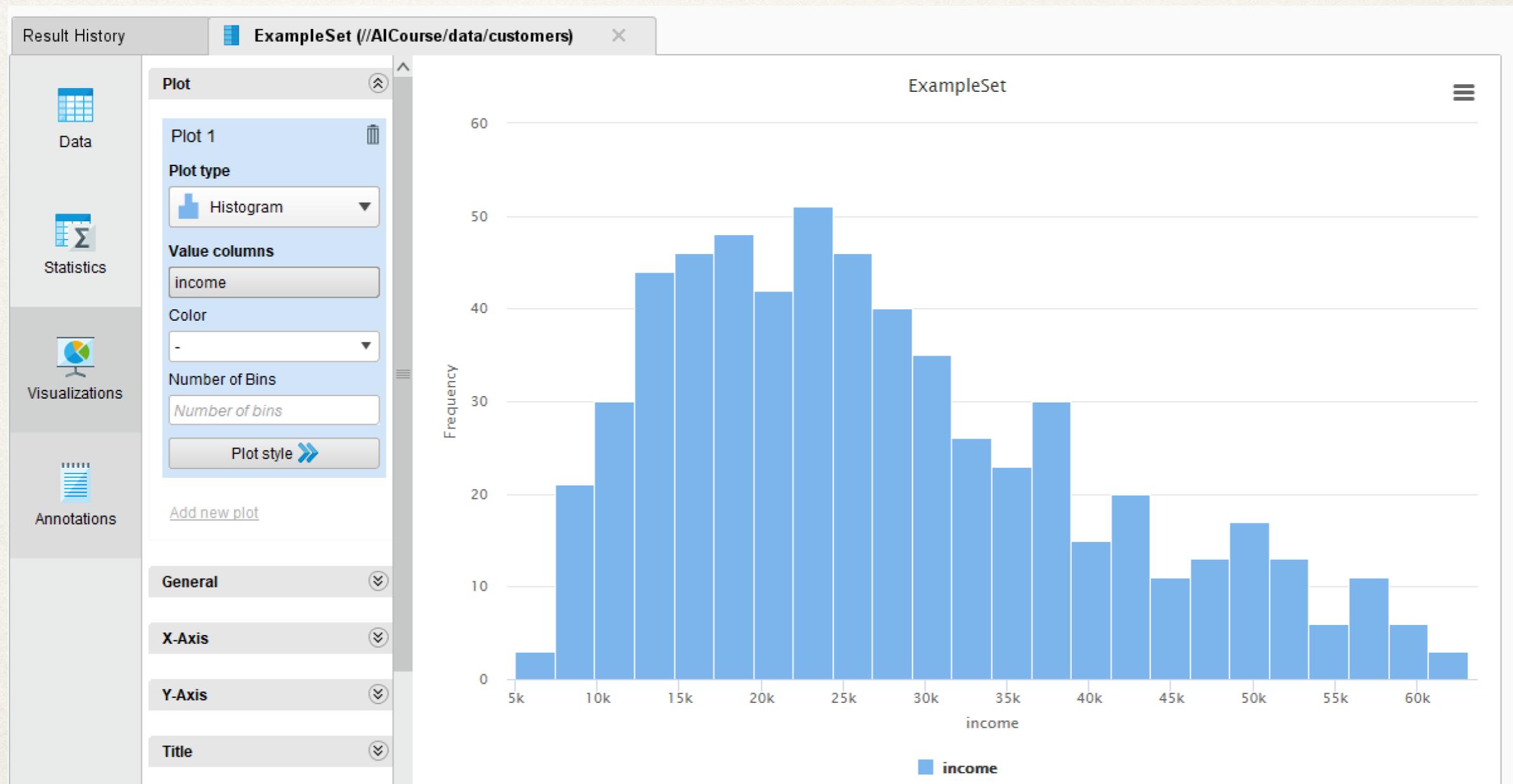
❖ Statistics

Result History ExampleSet //AI.Course/data/customers

Name	Type	Missing	Statistics	Filter (9 / 9 attributes):	Search for Attributes	Filter
customer_id	Polynomial	0	Least ID12700 (1) Most ID12101 (1) Values ID12101 (1), ID12102 (1), ...[598]			
response	Polynomial	0	Least NO (186) Most YES (414)		Open visualizations	
age	Integer	0	Min 2 Max 999 Average 43.875			
gender	Polynomial	0	Least MAN (3) Most FEMALE (291) Values FEMALE (291), MALE (290), ...[4]			
region	Polynomial	6	Least SUBURBAN (61) Most INNER_CITY (265) Values INNER_CITY (265), TOWN (173)			
income	Real	0	Min 5014.210 Max 63130.100 Average 27524.031			

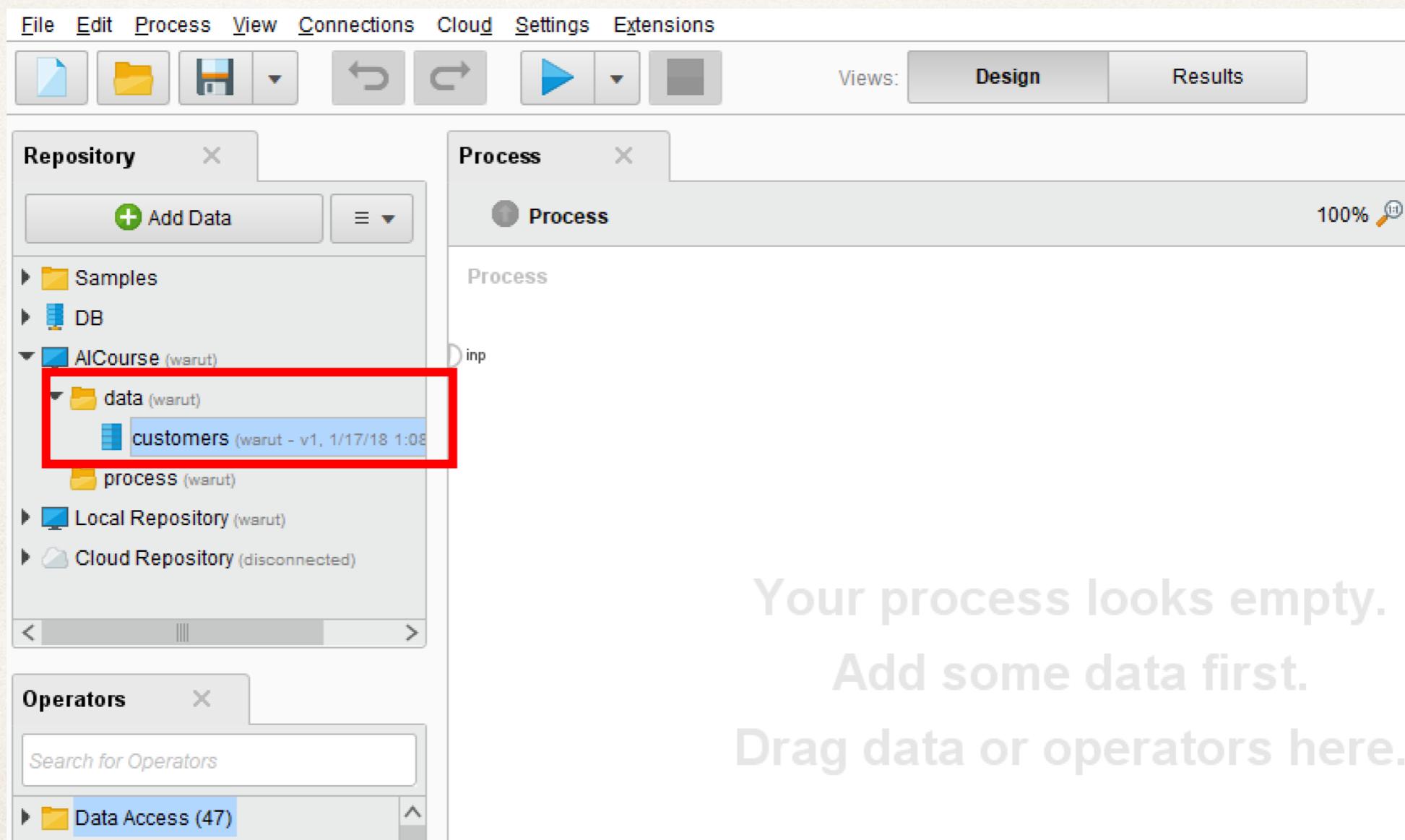
Import Data into Repository

❖ Visualizations



Import Data into Repository

- ✿ The data is stored in repository.



Data Pre-Processing with RapidMiner

- ✿ Data management
- ✿ Import data
- ✿ **Data exploration**
- ✿ Data preparation

Data Exploration

- ❖ Data
- ❖ Statistics
- ❖ Visualizations

Result History ExampleSet (//AI.Course/data/customers) X

Open in [Turbo Prep](#) [Auto Model](#) Filter (60)

Row No.	customer_id	response	age	gender	region	income	married
1	ID12101	NO	100	FEMALE	?	17546	NO
2	ID12102	NO	40	MALE	TOWN	30085.100	YES
3	ID12103	YES	51	FEMALE	INNER_CITY	16575.400	YES
4	ID12104	NO	23	FEMALE	TOWN	20375.400	YES
5	ID12105	YES	57	FEMALE	RURAL	50576.300	YES
6	ID12106	YES	57	WOMAN	TOWN	37869.600	YES
...

Data **Statistics** **Visualizations**

Data Exploration

- ✿ Data
 - ✿ ExampleSet: All data in the file
 - ✿ Filter
 - ✿ All
 - ✿ No_missing_attributes
 - ✿ Missing_attributes
 - ✿ no_missing_labels
 - ✿ Missing_labels
 - ✿ Sorting by multiple attributes
 - ✿ Ctrl (hold) + Click attribute name

Filter (600 / 600 examples): all

The screenshot shows a data exploration interface. At the top, there is a dropdown menu labeled "Filter (600 / 600 examples):" with the option "all" selected. Below the dropdown is a table with two columns: "married" and "children". The data in the table is as follows:

	married	children
NO	1	
YES	3	
YES	0	YES
YES	3	NO

A vertical scroll bar is visible on the right side of the table.

Data Exploration

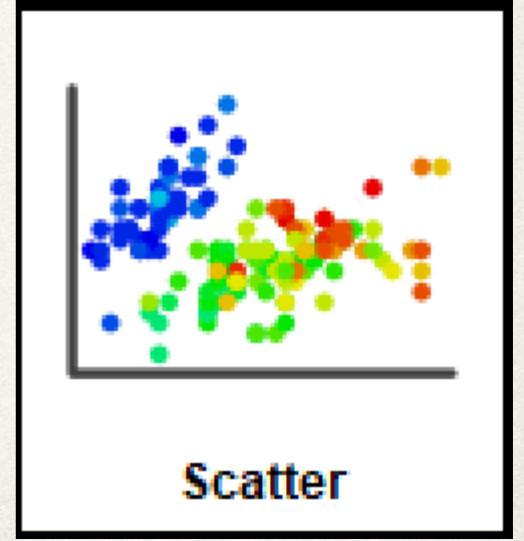
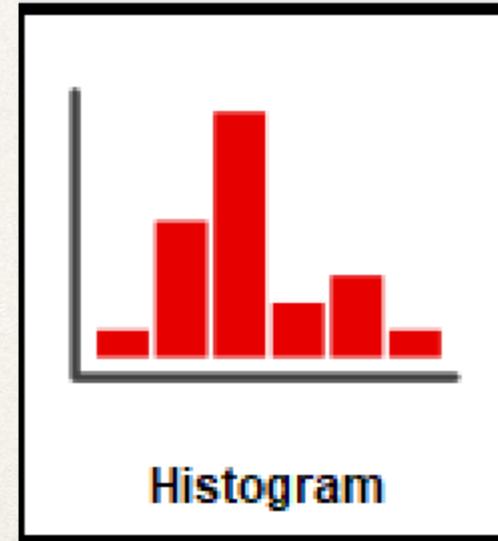
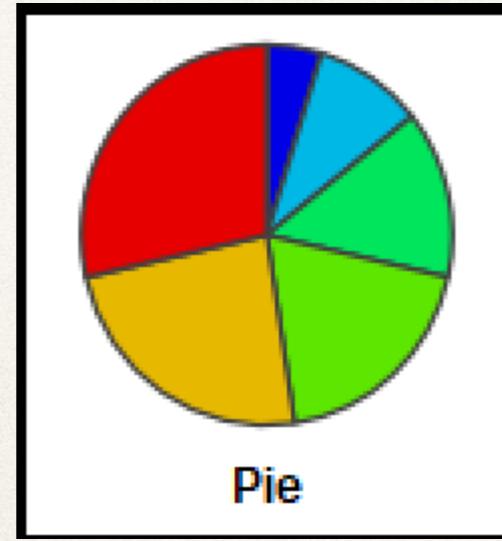
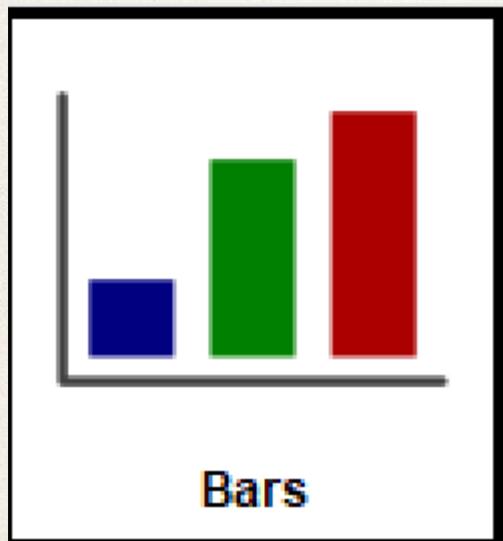
- ❖ Statistics
 - ❖ Name
 - ❖ Type
 - ❖ Miss
 - ❖ Min, Max, Average, Deviation, Least, Most, Values

The screenshot shows the KNIME interface with the 'ExampleSet (Read CSV)' view open. The top menu bar includes File, Edit, Process, View, Connections, Cloud, Settings, and Extensions. Below the menu is a toolbar with icons for file operations and navigation. The main window displays a table of attributes with their statistics. On the left, there are tabs for Data, Statistics, and another tab that is partially visible. The 'Data' tab is currently selected.

Name	Type	Missing	Statistics	Filter (9 / 9 attributes):	Search for Attributes
customer_id	Polynomial	0	Least ID12700 (1) Most ID12101 (1)	ID12101 (1), ID12102 (1)	
response	Polynomial	0	Least NO (186) Most YES (414)	YES (414), NO (186)	
age	Integer	0	Min 2 Max 999	Average 43.788	

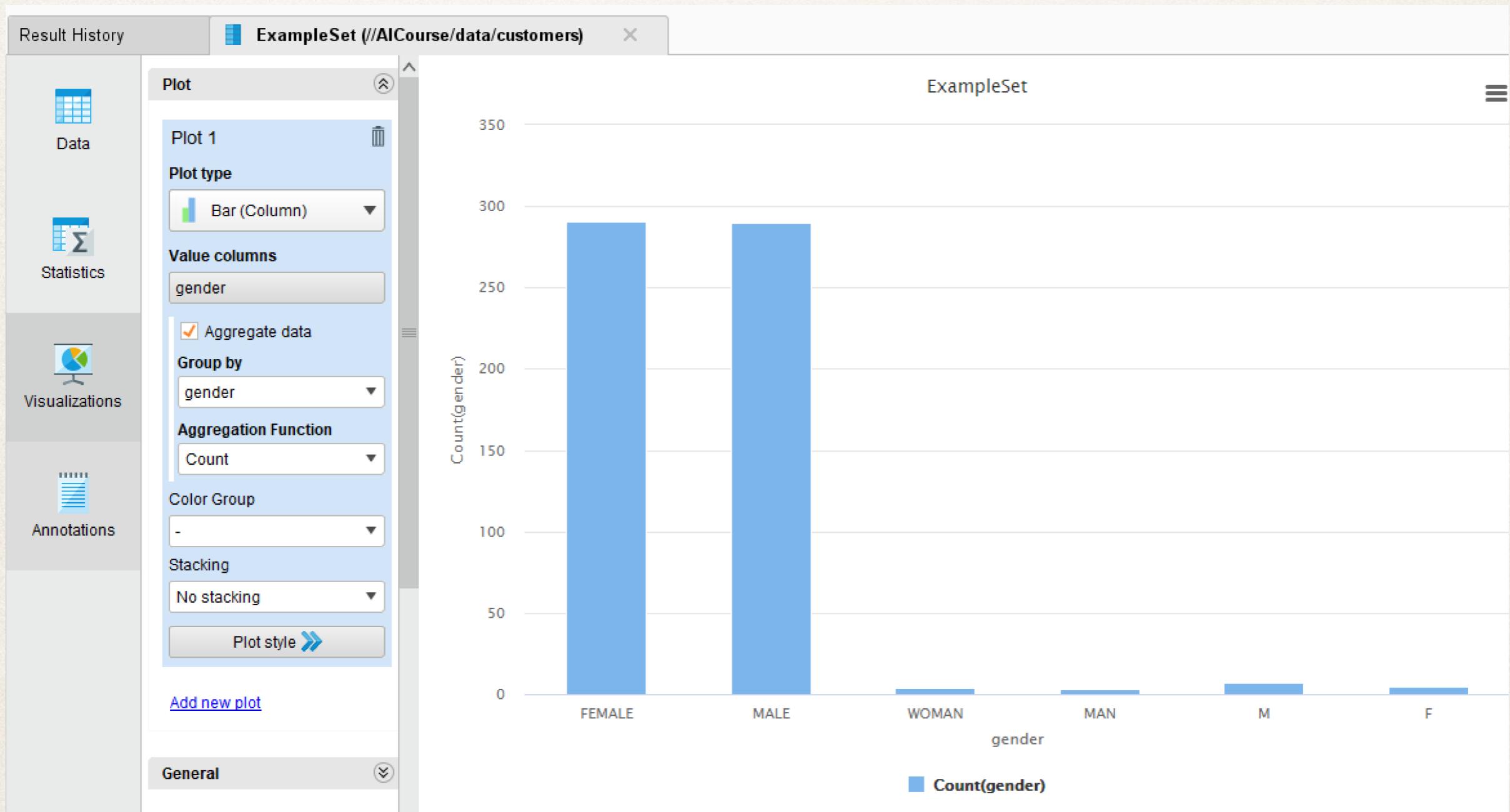
Data Visualization

- ❖ Charts
 - ❖ Bars
 - ❖ Histogram
 - ❖ Pie
 - ❖ Scatter



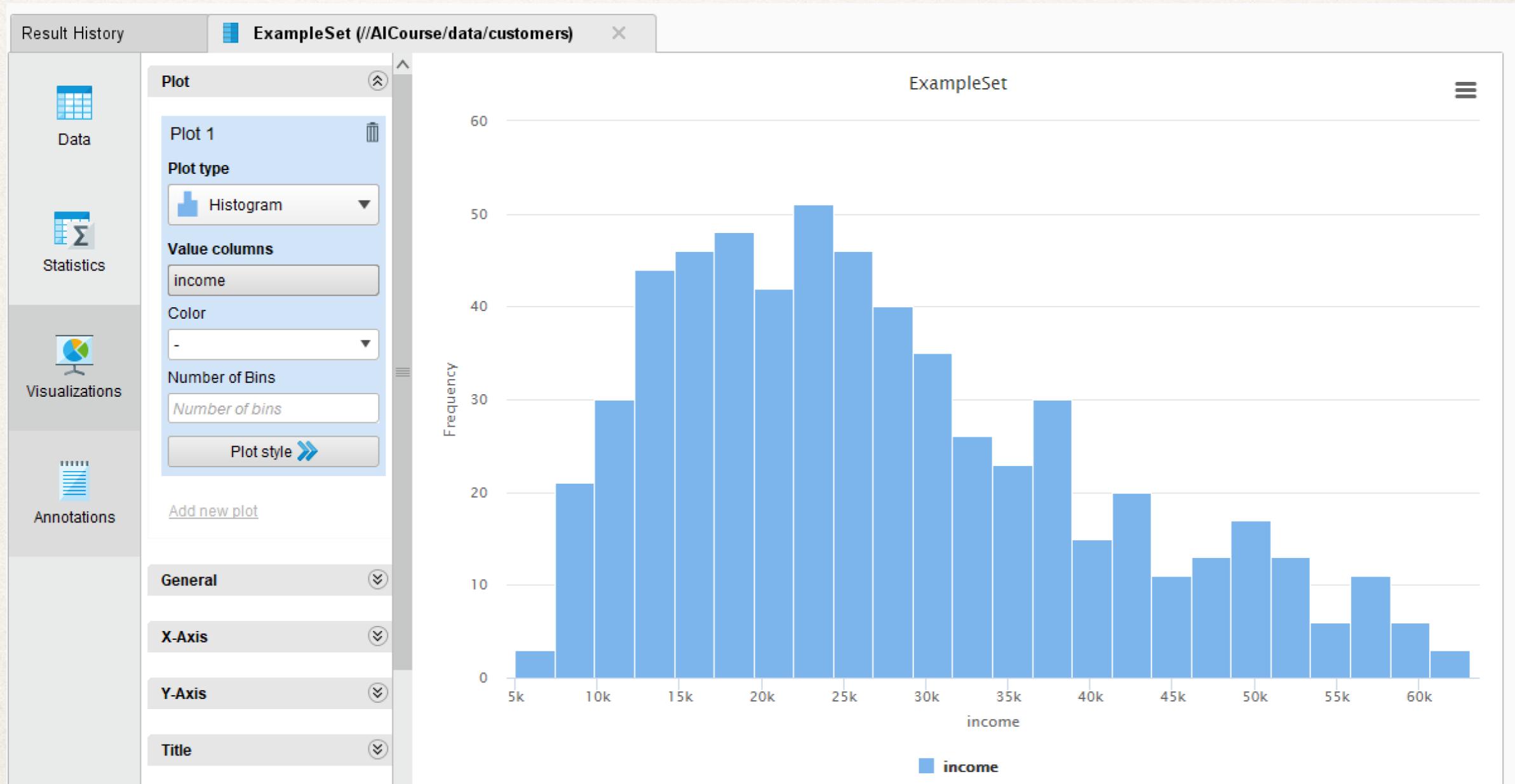
Data Visualization

✿ Bar chart of gender attribute



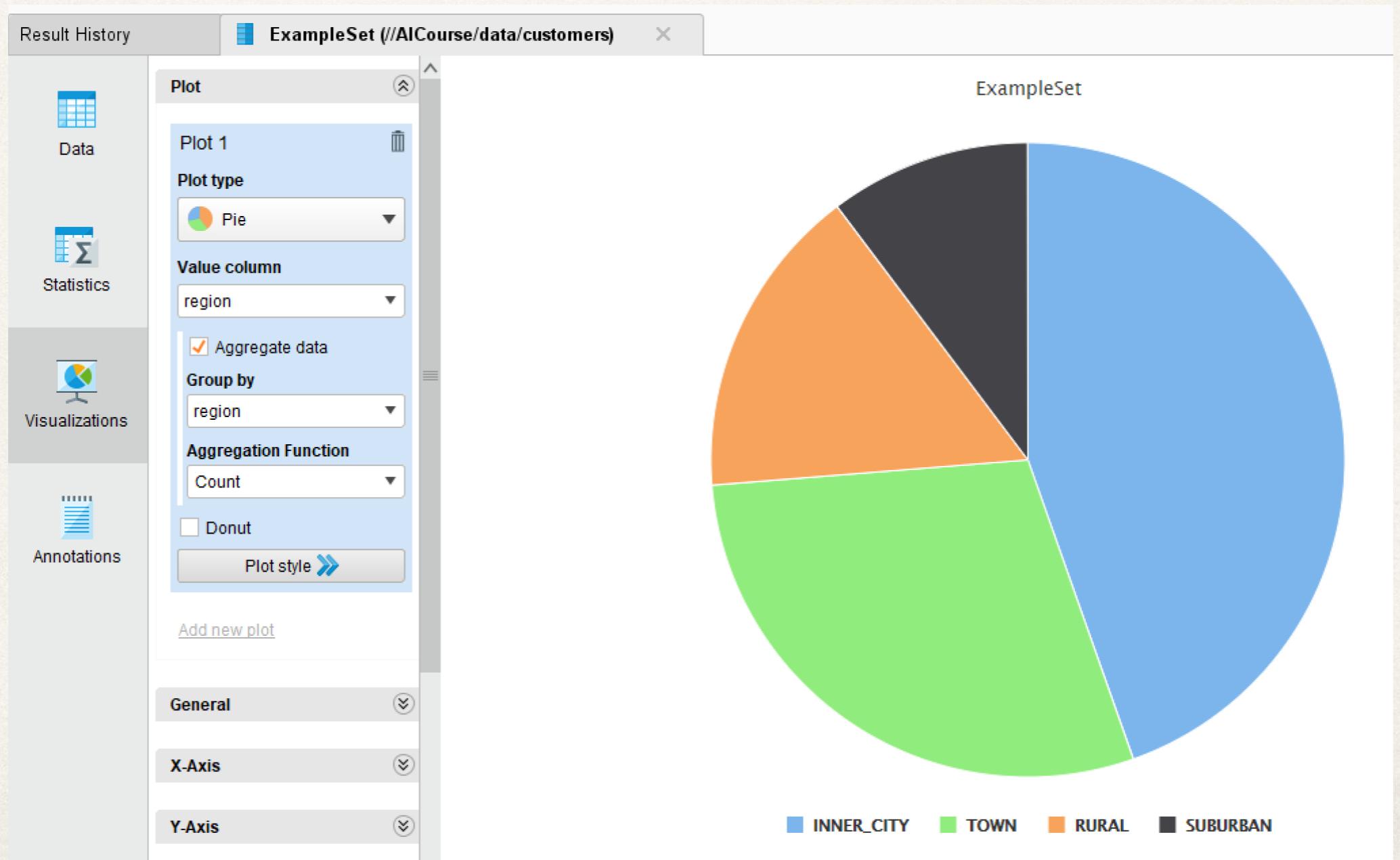
Data Visualization

✿ Histogram of income attribute



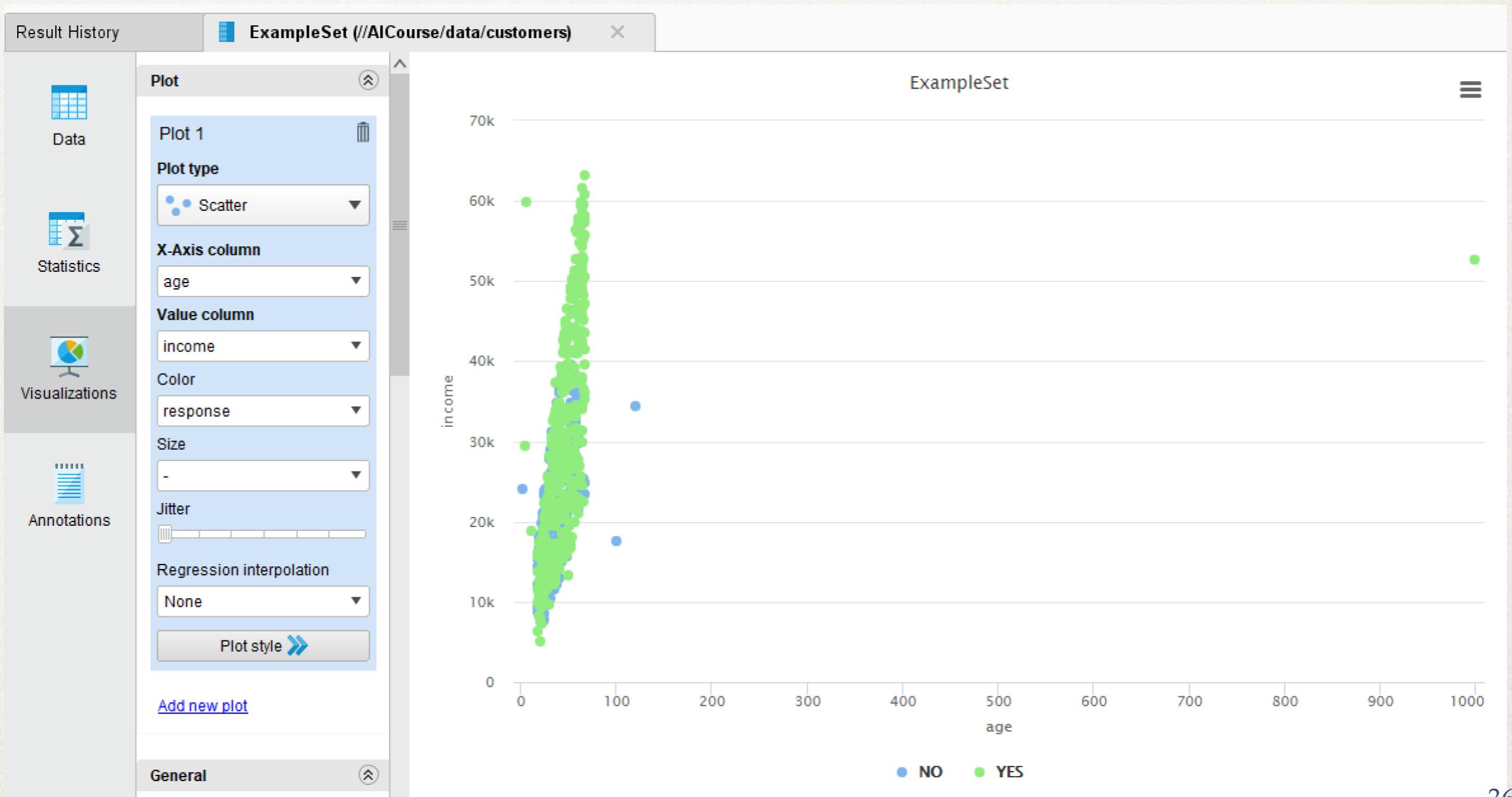
Data Visualization

- ✿ Pie chart of **region** attribute



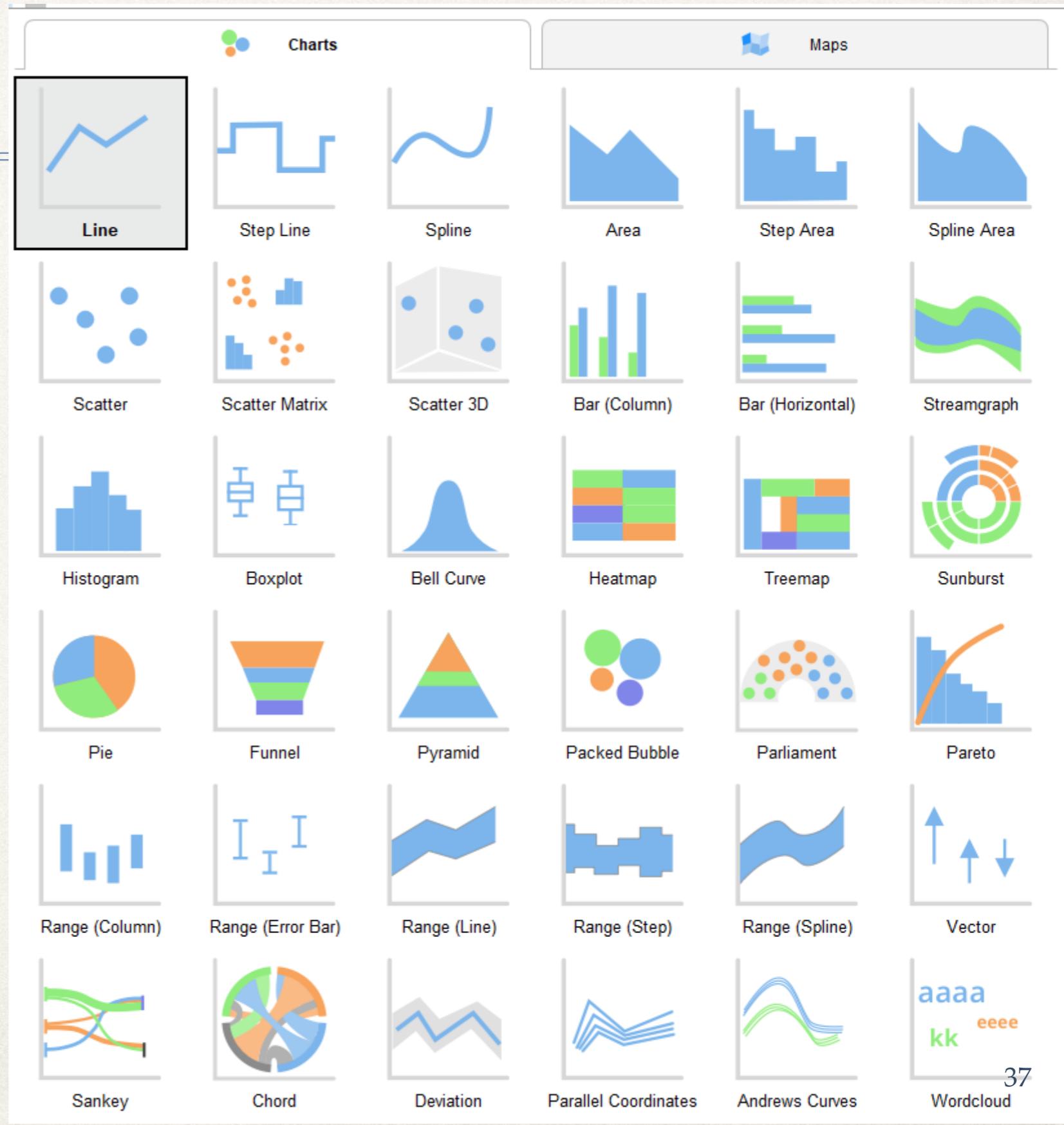
Data Visualization

- ❖ Scatter plot of age and income attributes



Data Visualization

- ✿ Different types of chart for different proposes

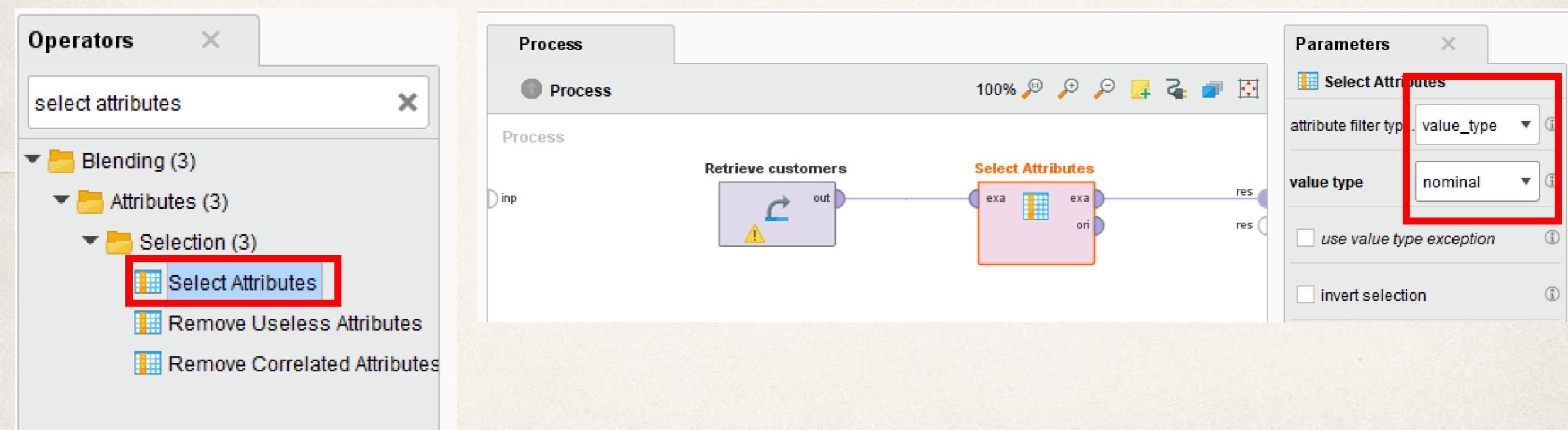


Data Preparation

- ✿ Preprocessing
 - ✿ Select attributes
 - ✿ Select by type of attributes
 - ✿ Select by specific attributes
 - ✿ Filter examples by conditions
 - ✿ Join data from multiple sources
- ✿ Deal with incomplete data
 - ✿ Inconsistent data
 - ✿ Missing data
 - ✿ Outliers
- ✿ Data transformation
 - ✿ Discretization (numeric to nominal)
 - ✿ User defined
 - ✿ Equal frequency

Select Attributes by Type

- ✿ Choose **Select Attributes** operator
- ✿ attribute filter type = **value_type**
- ✿ value type = **nominal**



Select Attributes by Type

- ❖ The result shows only nominal attributes.

Result History X ExampleSet (Select Attributes) X

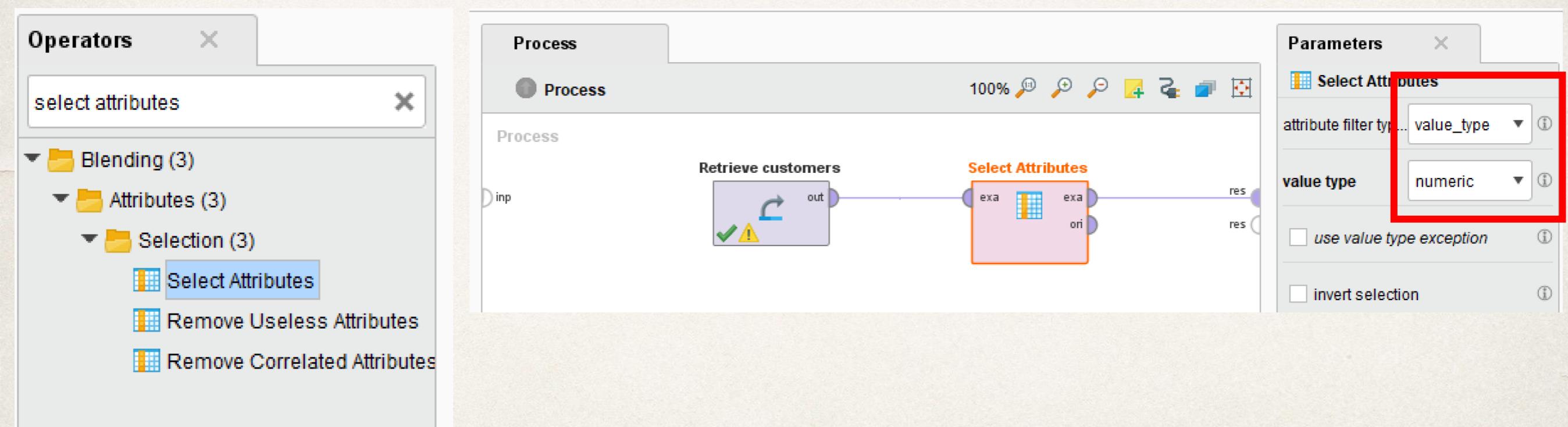
ExampleSet (600 examples, 2 special attributes, 4 regular attributes) Filter (600 / 600 examples): all

Data Statistics Charts

Row No.	customer_id	response	gender	region	married	car
1	ID12101	NO	FEMALE	?	NO	NO
2	ID12102	NO	MALE	TOWN	YES	YES
3	ID12103	YES	FEMALE	INNER_CITY	YES	YES
4	ID12104	NO	FEMALE	TOWN	YES	NO
5	ID12105	YES	FEMALE	RURAL	YES	NO
6	ID12106	YES	WOMAN	TOWN	YES	NO
7	ID12107	NO	MALE	RURAL	NO	NO
8	ID12108	YES	MALE	TOWN	YES	YES
9	ID12109	NO	FEMALE	SUBURBAN	YES	YES

Select Attributes by Type

- ✿ Choose **Select Attributes** operator
- ✿ attribute filter type = **value_type**
- ✿ value type = **numeric**



Select Attributes by Type

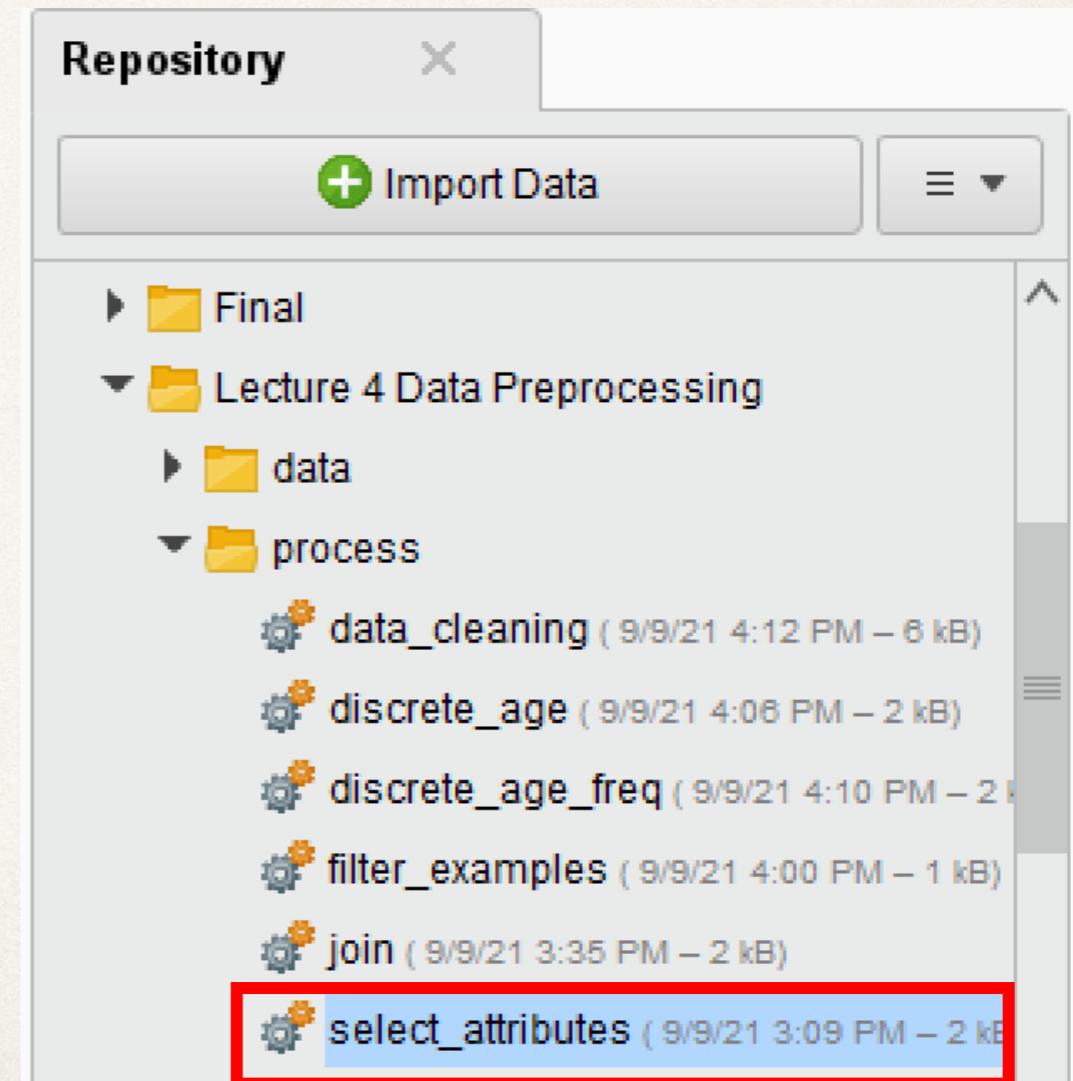
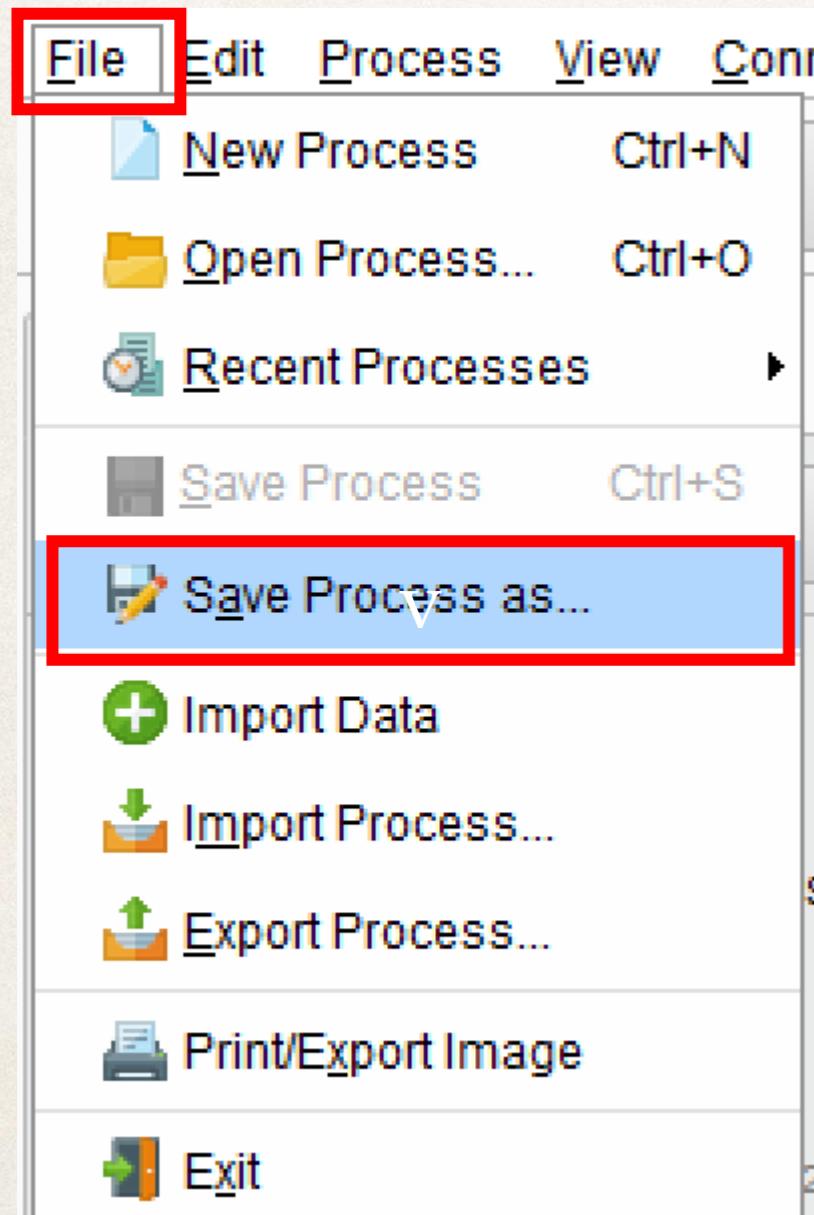
- ❖ The result shows only numeric attributes.

The screenshot shows the KNIME interface with the 'ExampleSet (Select Attributes)' tab selected. On the left, there's a sidebar with icons for Data, Statistics, Visualizations, and Annotations. The main area displays a table with 11 rows and 5 columns. The columns are labeled 'Row No.', 'ID', 'response', 'age', 'income', and 'children'. The 'response' column contains categorical values ('NO', 'YES'), while 'age', 'income', and 'children' are numeric. A red box highlights the last three columns. The table data is as follows:

Row No.	ID	response	age	income	children
1	ID12101	NO	100	17546	1
2	ID12102	NO	40	30085.100	3
3	ID12103	YES	51	16575.400	0
4	ID12104	NO	23	20375.400	3
5	ID12105	YES	57	50576.300	0
6	ID12106	YES	57	37869.600	2
7	ID12107	NO	22	8877.070	0
8	ID12108	YES	58	24946.600	0
9	ID12109	NO	37	25304.300	2
10	ID12110	YES	54	24212.100	2
11	ID12111	YES	6	59803.900	0

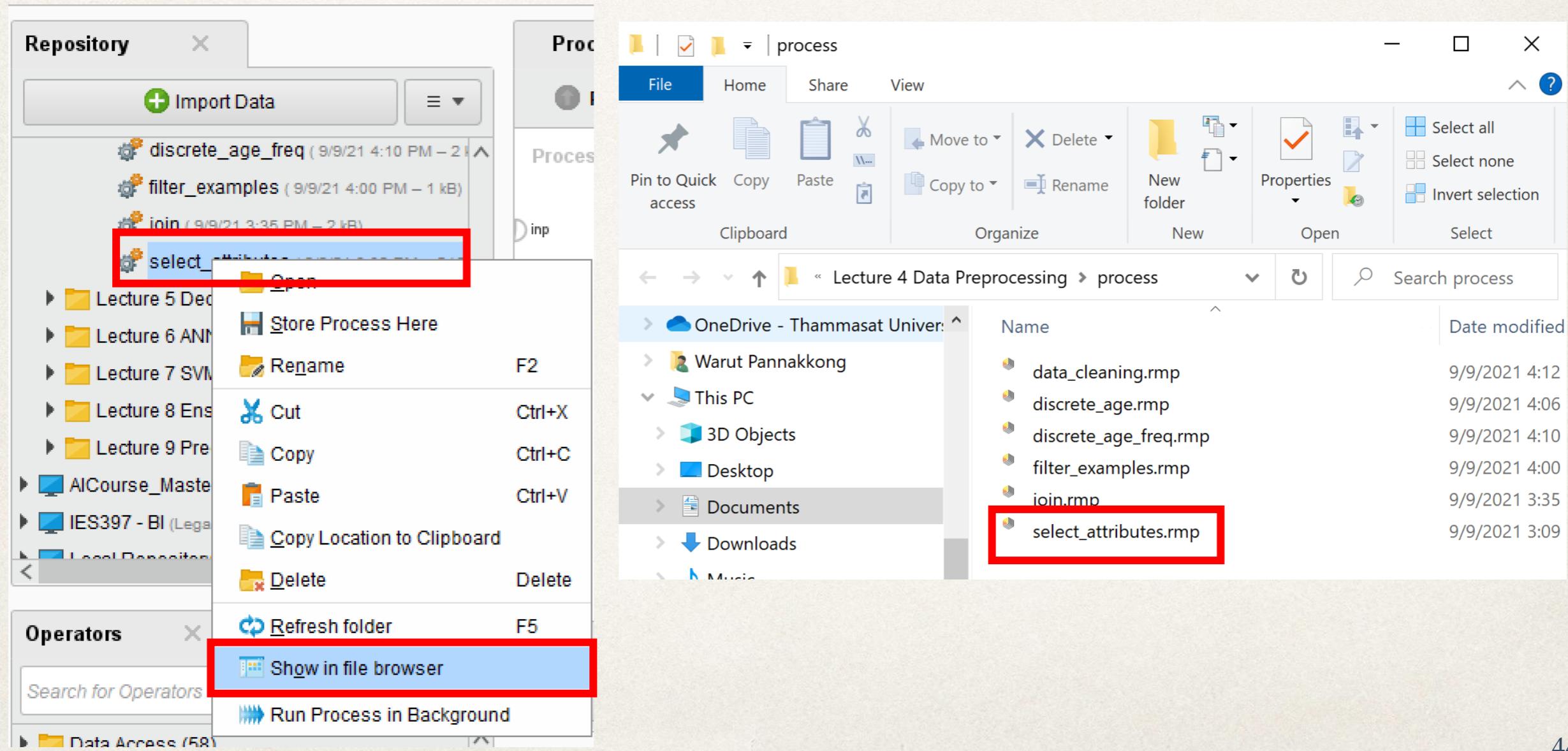
How to Save a Process into Repository

- File > Save Process as > .rmp file is saved in repository folder



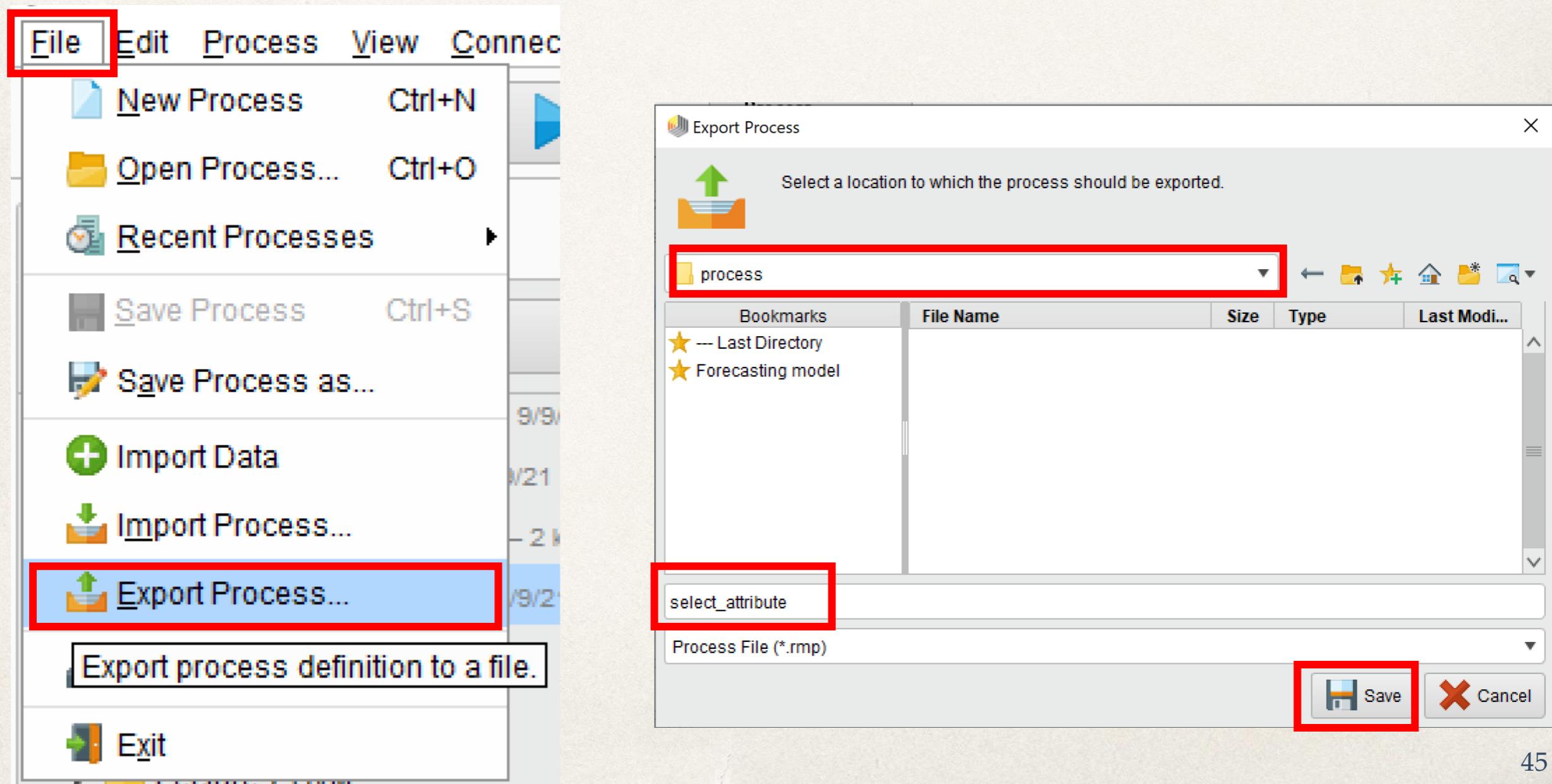
How to Show the Saved Process File (.rmp) in Repository

- Right click on the process > Select “Show in file browser”



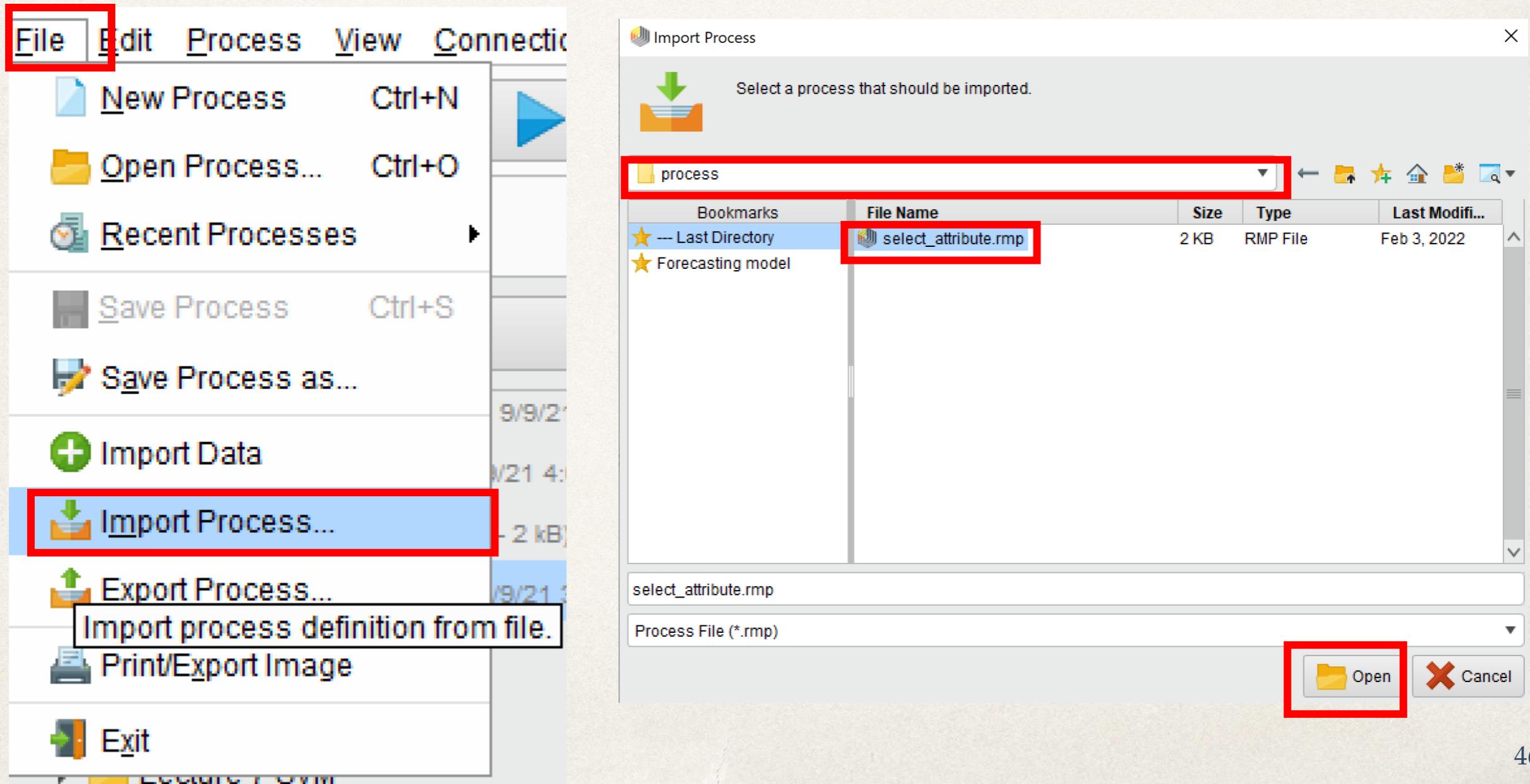
How to Export a Process File (.rmp)

- ❖ File > Export Process



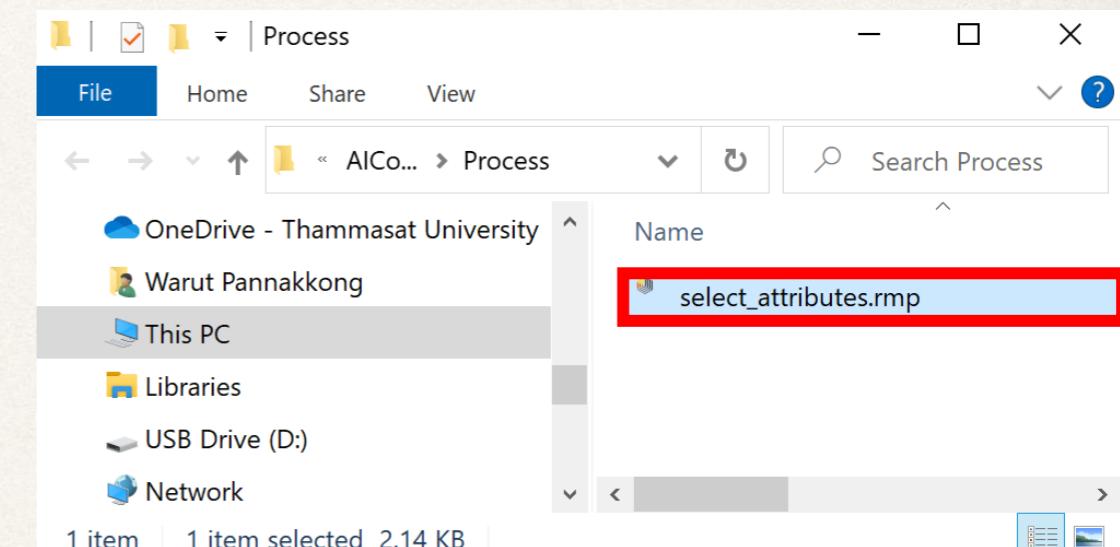
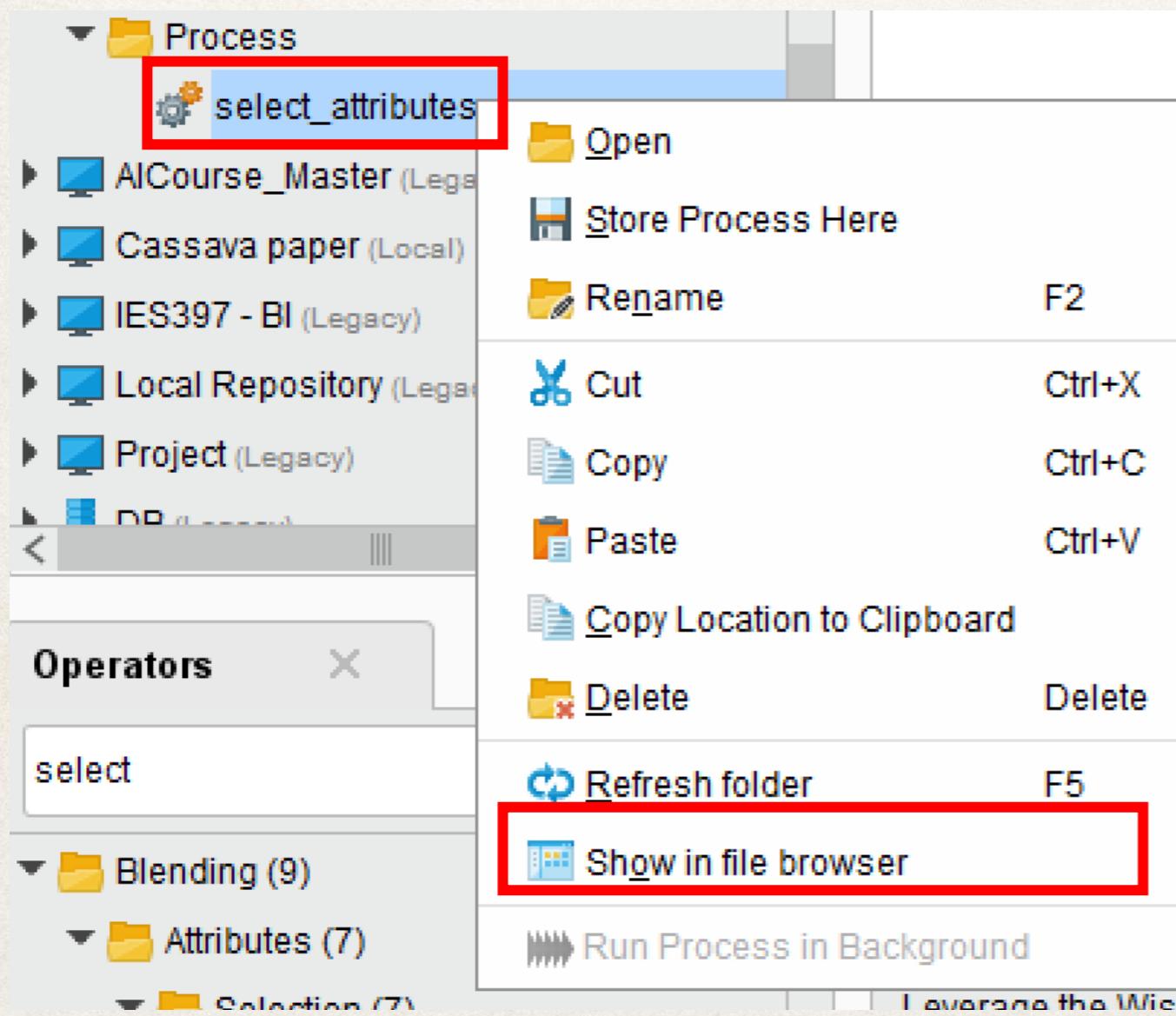
How to Import a Process File (.rmp)

- ❖ File > Export Process



How to Import a Process File (.rmp)

- ✿ Browse the file on the PC

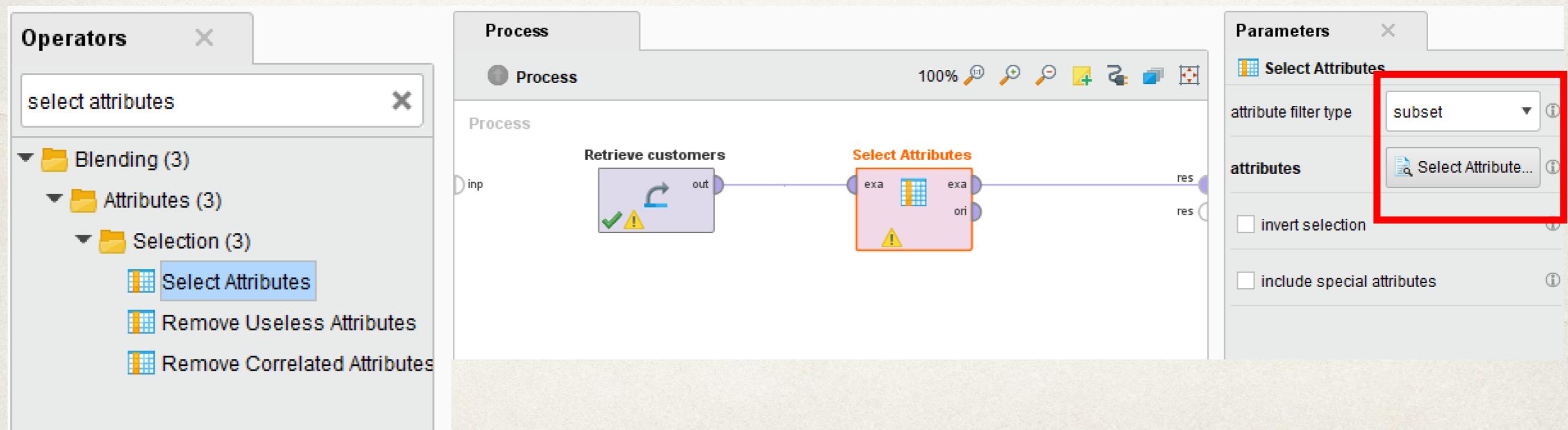


Data Preparation

- ✿ **Preprocessing**
 - ✿ **Select attributes**
 - ✿ Select by type of attributes
 - ✿ **Select by specific attributes**
 - ✿ Filter examples by conditions
 - ✿ Join data from multiple sources
- ✿ Deal with incomplete data
 - ✿ Inconsistent data
 - ✿ Missing data
- ✿ Data transformation
 - ✿ Discretization (numeric to nominal)
 - ✿ User defined
 - ✿ Equal frequency

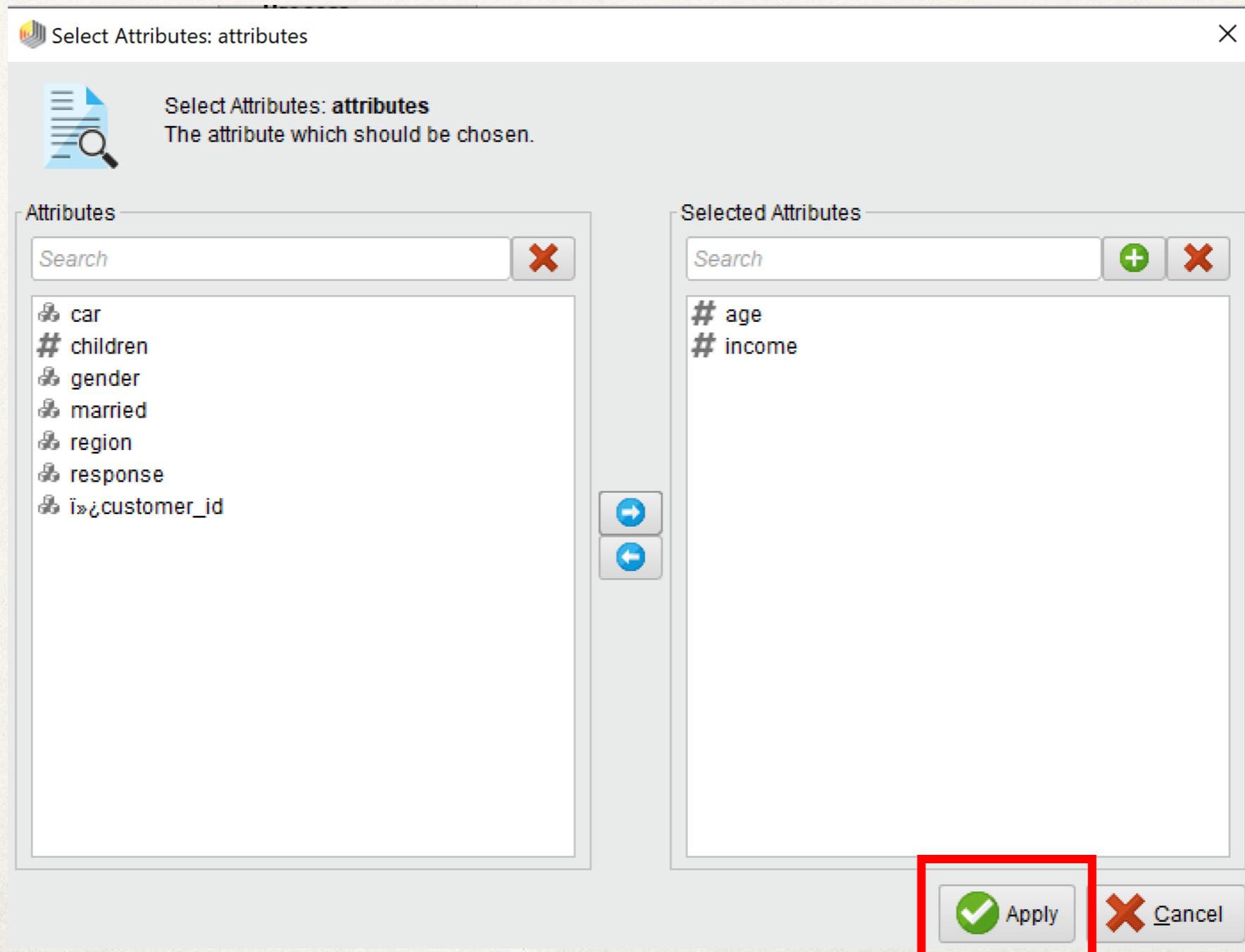
Select Attributes by Name

- ✿ Choose Select Attributes operator
- ✿ attribute filter type = subset
- ✿ value type = Select Attributes



Select Attributes by Name

- Select the name(s) of attribute that will be shown in the result.



Select Attributes by Name

- The result shows only the selected attributes.

The screenshot shows the KNIME interface with the following details:

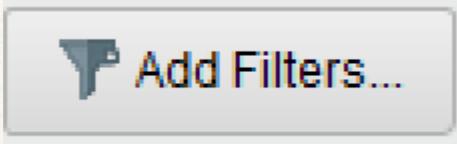
- Result History:** Shows "ExampleSet (Select Attributes)"
- Data View:** Shows a table with columns: Row No., ID, response, age, and income. The columns "age" and "income" are highlighted with a red border.
- Open In:** Buttons for "Turbo Prep" and "Auto Model".
- Filter:** Shows "Filter (600 / 600 examples):" followed by a dropdown menu.
- Left Sidebar:** Categories: Data, Statistics, Visualizations, Annotations.

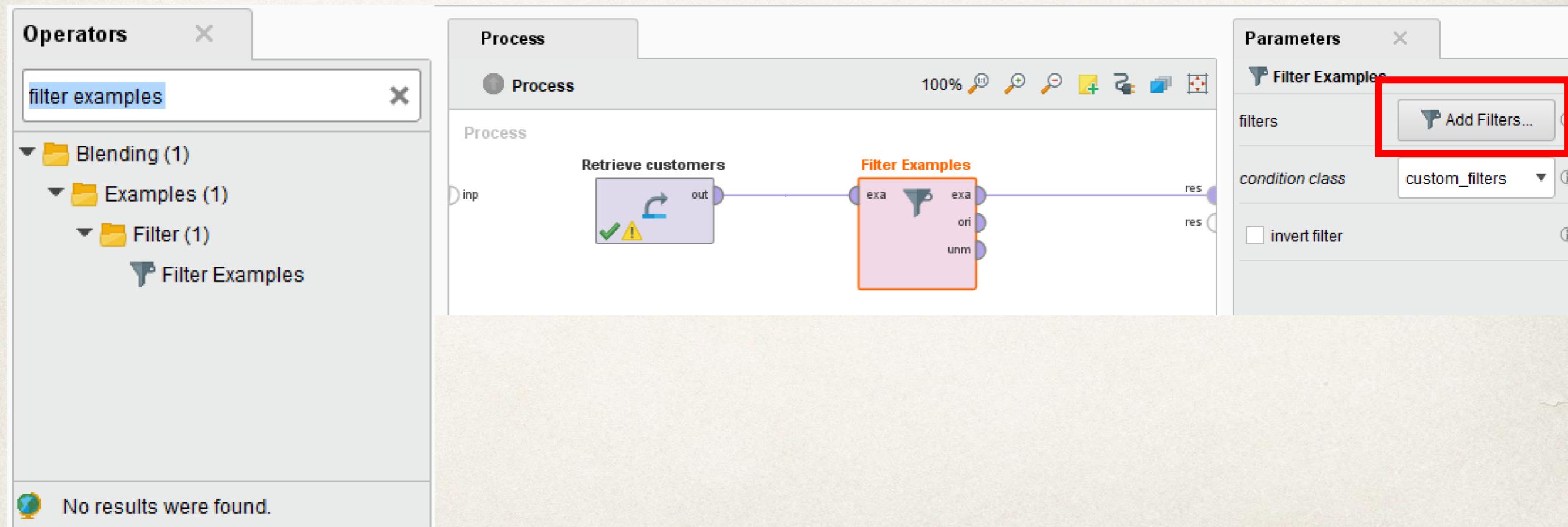
Row No.	ID	response	age	income
1	ID12101	NO	100	17546
2	ID12102	NO	40	30085.100
3	ID12103	YES	51	16575.400
4	ID12104	NO	23	20375.400
5	ID12105	YES	57	50576.300
6	ID12106	YES	57	37869.600
7	ID12107	NO	22	8877.070
8	ID12108	YES	58	24946.600
9	ID12109	NO	37	25304.300
10	ID12110	YES	54	24212.100
11	ID12111	YES	6	59803.900

Data Preparation

- ❖ Preprocessing
 - ❖ Select attributes
 - ❖ Select by type of attributes
 - ❖ Select by specific attributes
 - ❖ **Filter examples by conditions**
 - ❖ Join data from multiple sources
 - ❖ Deal with incomplete data
 - ❖ Inconsistent data
 - ❖ Missing data
 - ❖ Data transformation
 - ❖ Discretization (numeric to nominal)
 - ❖ User defined
 - ❖ Equal frequency

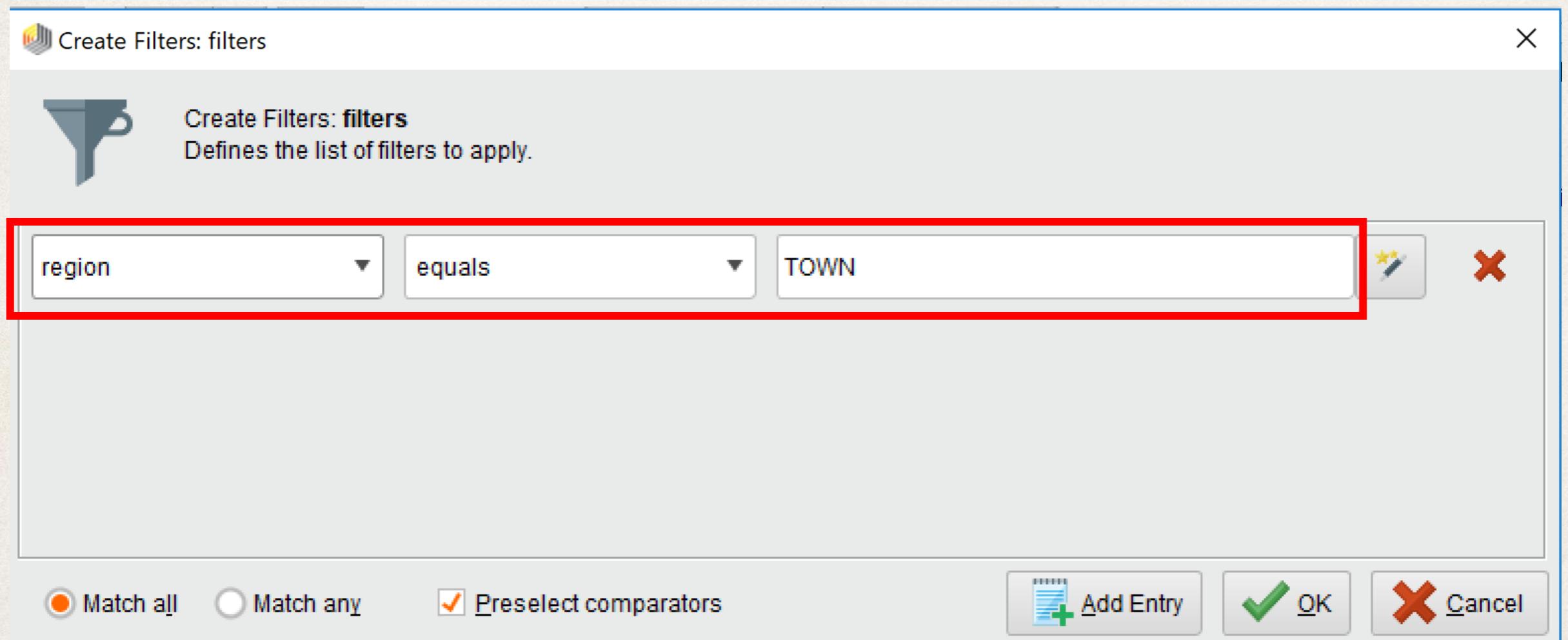
Filter Examples by Conditions

- ✿ Choose Filter Examples operator
- ✿ Press 



Filter Examples by Conditions

- ❖ Set filter condition: Region equals town.



Filter Examples by Conditions

- The result shows only the examples satisfying the filter condition.

Result History X ExampleSet (Filter Examples) X

Data Statistics Charts Advanced

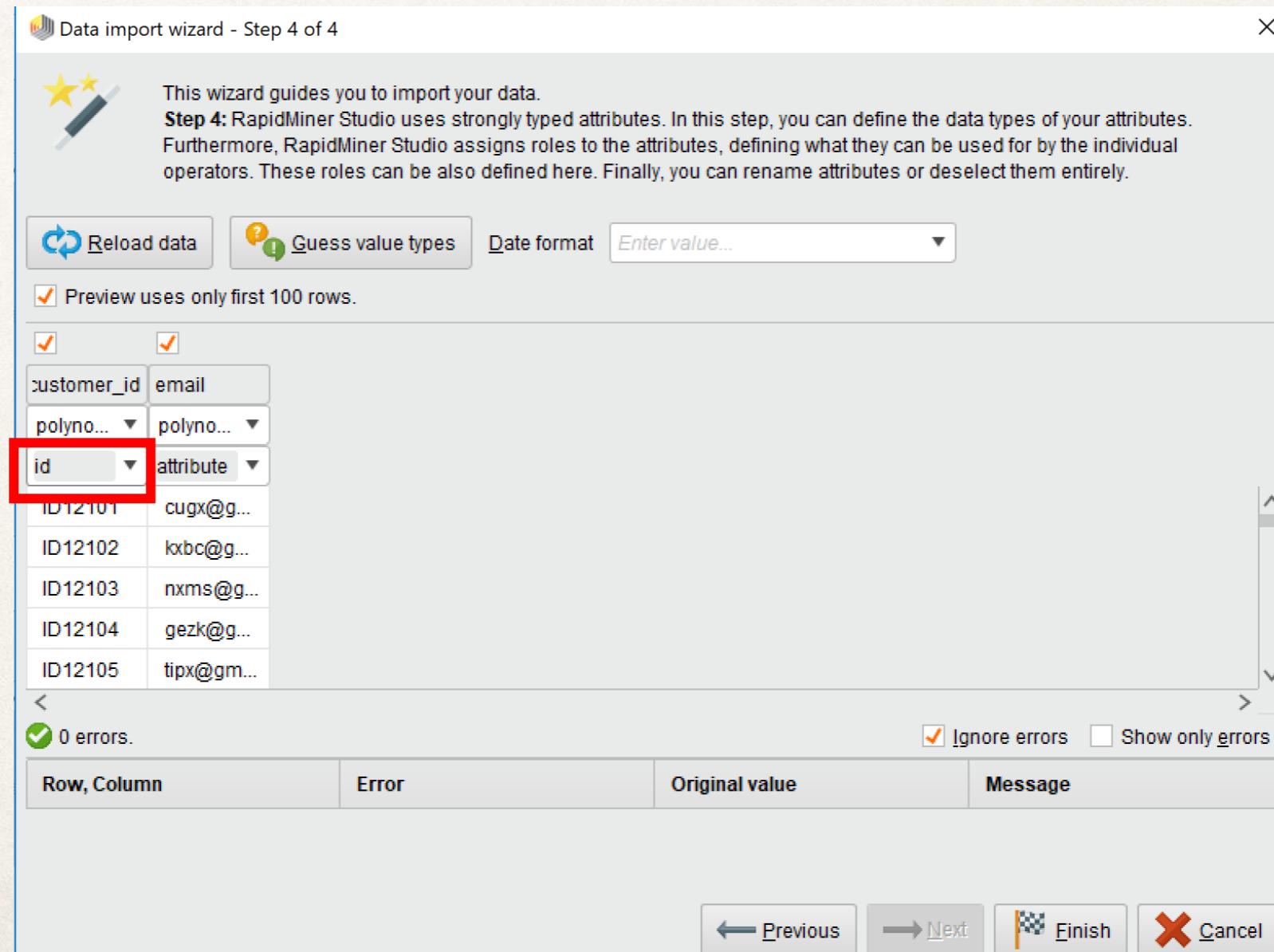
ExampleSet (173 examples, 2 special attributes, 7 regular attributes)									Filter (173 / 173 examples):	all
Row No.	customer_id	response	age	gender	region	income	married	children		
1	ID12102	NO	40	MALE	TOWN	30085.100	YES	3	^	
2	ID12104	NO	23	FEMALE	TOWN	20375.400	YES	3		
3	ID12106	YES	57	WOMAN	TOWN	37869.600	YES	2		
4	ID12108	YES	58	MALE	TOWN	24946.600	YES	0		
5	ID12110	YES	54	MAN	TOWN	24212.100	YES	2		
6	ID12111	YES	6	FEMALE	TOWN	59803.900	YES	0		
7	ID12113	YES	44	FEMALE	TOWN	15735.800	YES	1		
8	ID12114	YES	66	FEMALE	TOWN	55204.700	YES	1		
9	ID12117	NO	37	FEMALE	TOWN	17729.800	YES	2		
10	ID12120	YES	31	M	TOWN	22522.800	YES	0		

Data Preparation

- ✿ Preprocessing
 - ✿ Select attributes
 - ✿ Select by type of attributes
 - ✿ Select by specific attributes
 - ✿ Filter examples by conditions
 - ✿ **Join data from multiple sources**
- ✿ Deal with incomplete data
 - ✿ Inconsistent data
 - ✿ Missing data
- ✿ Data transformation
 - ✿ Discretization (numeric to nominal)
 - ✿ User defined
 - ✿ Equal frequency

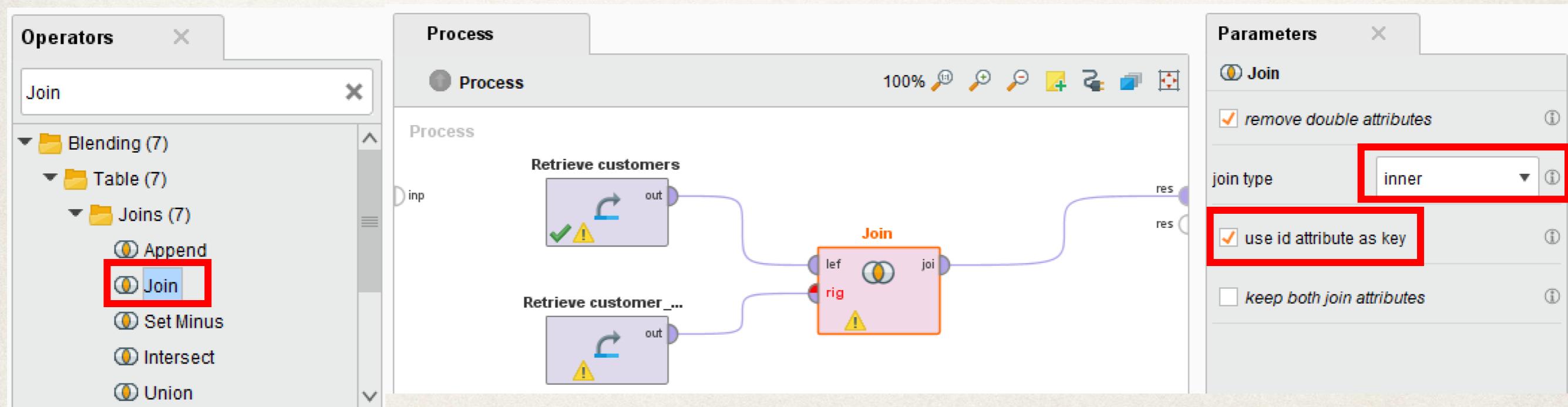
Join Data from Multiple Sources

- ❖ Load `customer_email.csv` into repository
- ❖ Set role of `customer_id` = `id`



Join Data from Multiple Sources

- * Choose Join operator
- * Connect the ports
- * join type = inner
- * Select use id attribute as key



Join Data from Multiple Sources

- ✿ The email accounts have been included into the

Result History **DataTable (Join)** X

Open in [Turbo Prep](#) [Auto Model](#) Filter (600 / 600 examples): all

Data

Row No.	Customer ID	response	email	age	gender	region	income	married
1	ID12101	NO	cugx@gmail.com	100	FEMALE	?	17546	NO
2	ID12102	NO	kxbc@gmail.com	40	MALE	TOWN	30085.100	YES
3	ID12103	YES	nxms@gmail.com	51	FEMALE	INNER_CITY	16575.400	YES
4	ID12104	NO	gezk@gmail.com	23	FEMALE	TOWN	20375.400	YES
5	ID12105	YES	tipx@gmail.com	57	FEMALE	RURAL	50576.300	YES
6	ID12106	YES	bosx@gmail.com	57	WOMAN	TOWN	37869.600	YES
7	ID12107	NO	qvqa@gmail.com	22	MALE	RURAL	8877.070	NO
8	ID12108	YES	twhd@gmail.com	58	MALE	TOWN	24946.600	YES
9	ID12109	NO	gtwi@gmail.com	37	FEMALE	SUBURBAN	25304.300	YES
10	ID12110	YES	fuvq@gmail.com	54	MAN	TOWN	24212.100	YES
11	ID12111	YES	iffy@gmail.com	6	FEMALE	TOWN	59803.900	YES

Statistics

Visualizations

Annotations

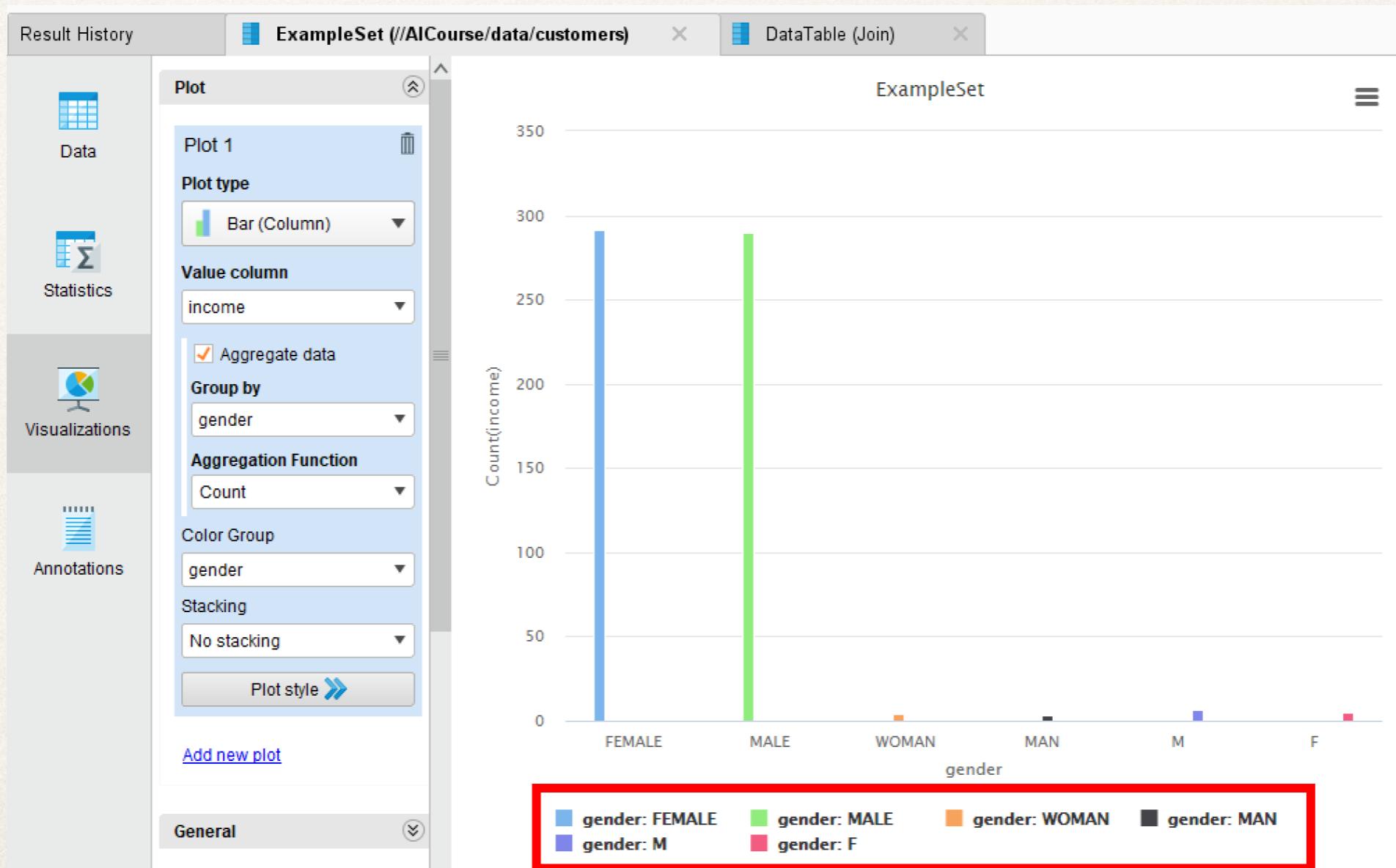
59

Data Preparation

- ❖ Preprocessing
 - ❖ Select attributes
 - ❖ Select by type of attributes
 - ❖ Select by specific attributes
 - ❖ Filter examples by conditions
 - ❖ Join data from multiple sources
- ❖ **Deal with incomplete data**
 - ❖ **Inconsistent data**
 - ❖ Missing data
- ❖ Data transformation
 - ❖ Discretization (numeric to nominal)
 - ❖ User defined
 - ❖ Equal frequency

Replace Inconsistent Data

- ✿ FEMALE, WOMAN, and F
- ✿ MALE, MAN, and M



Replace Inconsistent Data

- ✿ Choose Replace operator
- ✿ replace what = F\z|WOMAN\z
- ✿ replace by = FEMALE

The screenshot shows a data processing interface with three main panels: Operators, Process, and Parameters.

Operators Panel: Shows a tree view of operators. The "Replace" operator under "Values" is highlighted with a red box.

Process Panel: Displays a process flow starting with a "Retrieve customers" step, followed by two "Replace" steps: "Replace FEMALE" and "Replace MALE". The "Replace FEMALE" step is highlighted with a red box.

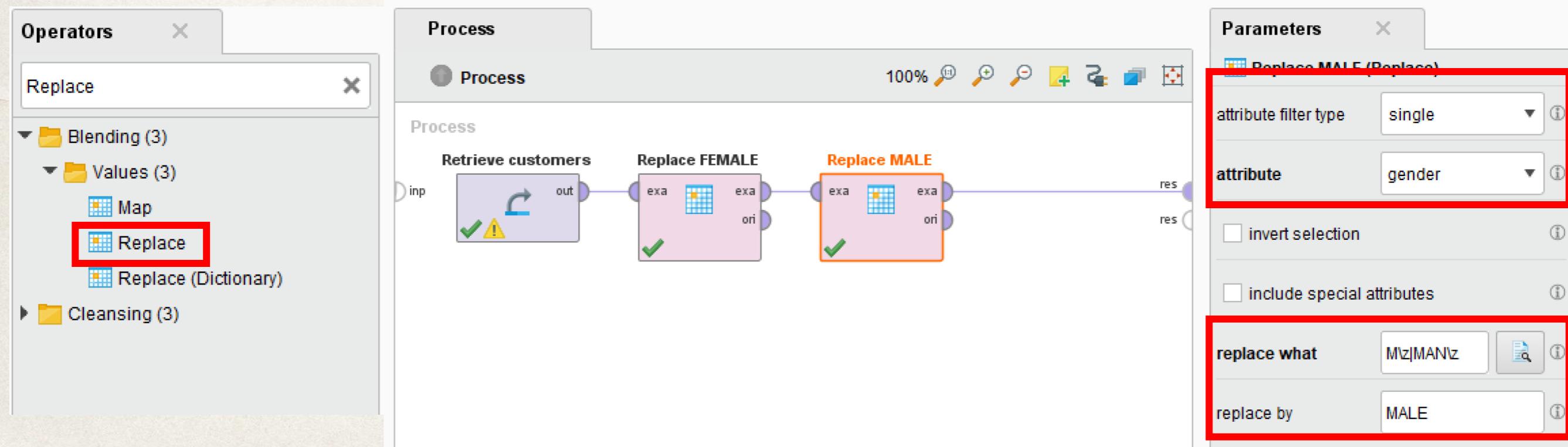
Parameters Panel: Shows the configuration for the selected "Replace FEMALE" operator. The parameters are:

- attribute filter type: single
- attribute: gender
- invert selection: unchecked
- include special attributes: unchecked
- replace what: F\z|WOMAN\z
- replace by: FEMALE

A red box highlights the "attribute" and "replace what" fields.

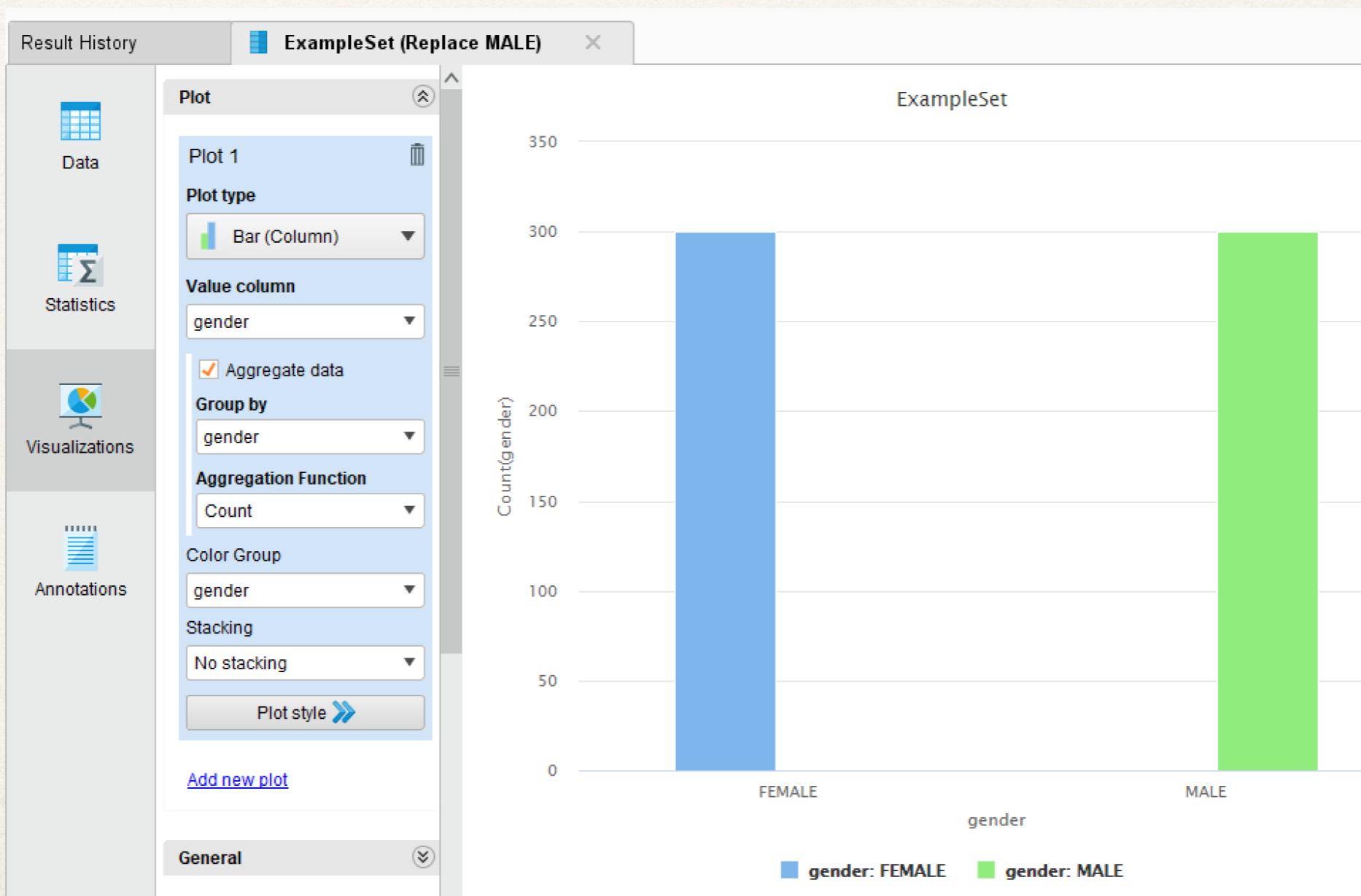
Replace Inconsistent Data

- ✿ Choose Replace operator
- ✿ replace what = M\z|MAN\z
- ✿ replace by = MALE



Replace Inconsistent Data

- ❖ The result after replacing the inconsistent data



Data Preparation

- ✿ Preprocessing
 - ✿ Select attributes
 - ✿ Select by type of attributes
 - ✿ Select by specific attributes
 - ✿ Filter examples by conditions
 - ✿ Join data from multiple sources
- ✿ **Deal with incomplete data**
 - ✿ Inconsistent data
 - ✿ **Missing data**
- ✿ Data transformation
 - ✿ Discretization (numeric to nominal)
 - ✿ User defined
 - ✿ Equal frequency

Replace Missing Data

- ✿ Data is not available.
- ✿ There is an error from filling data.

Result History X ExampleSet (//AICourse/data/customers) X

Data Statistics Charts

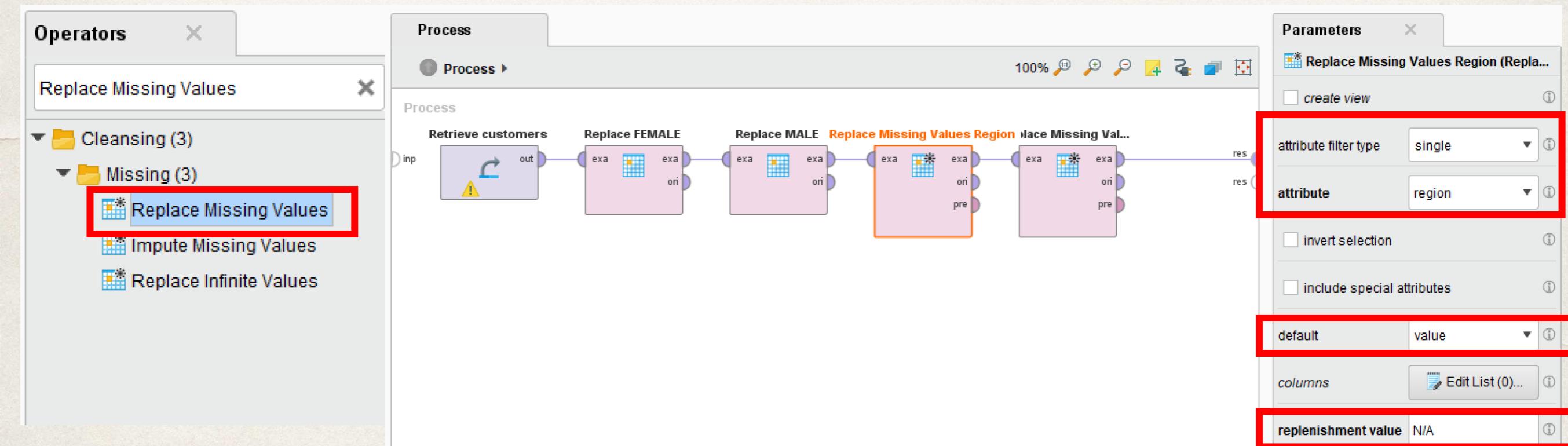
ExampleSet (600 examples, 2 special attributes, 7 regular attributes) Filter (11 / 600 examples): missing_attributes ▾

Row No.	custom...	response	age	gender	region	income	married	children	car
1	ID12101	NO	48	FEMALE	?	17546	NO	1	NO
2	ID12112	YES	52	FEMALE	?	26658.800	NO	0	YES
3	ID12123	YES	54	MALE	?	38446.600	YES	0	NO
4	ID12133	YES	45	MALE	?	23443.200	YES	1	YES
5	ID12138	NO	36	FEMALE	RURAL	13381	NO	?	YES
6	ID12144	NO	32	F	TOWN	27571.500	YES	?	YES
7	ID12150	YES	47	FEMALE	?	17867.300	YES	2	YES

Replace Missing Data

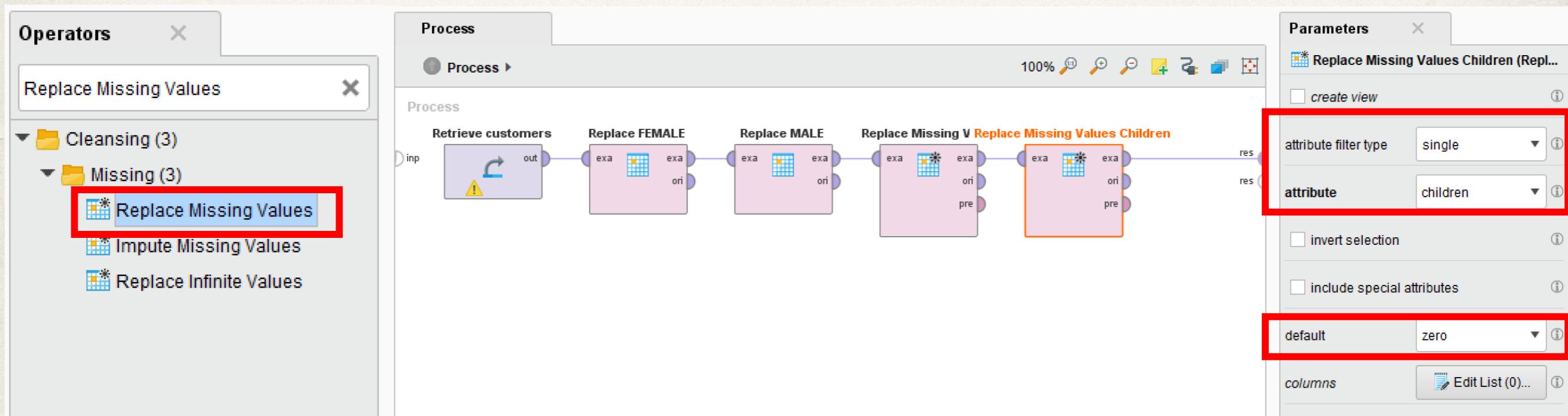
- ✿ How to replace?
- ✿ N/A
- ✿ Average
- ✿ Mode
- ✿ 0 (zero)

Replace Missing Data



Replace Missing Data

- ✿ Choose Replace Missing Values operator
 - ✿ attribute filter type = single attribute = children
 - ✿ default = zero



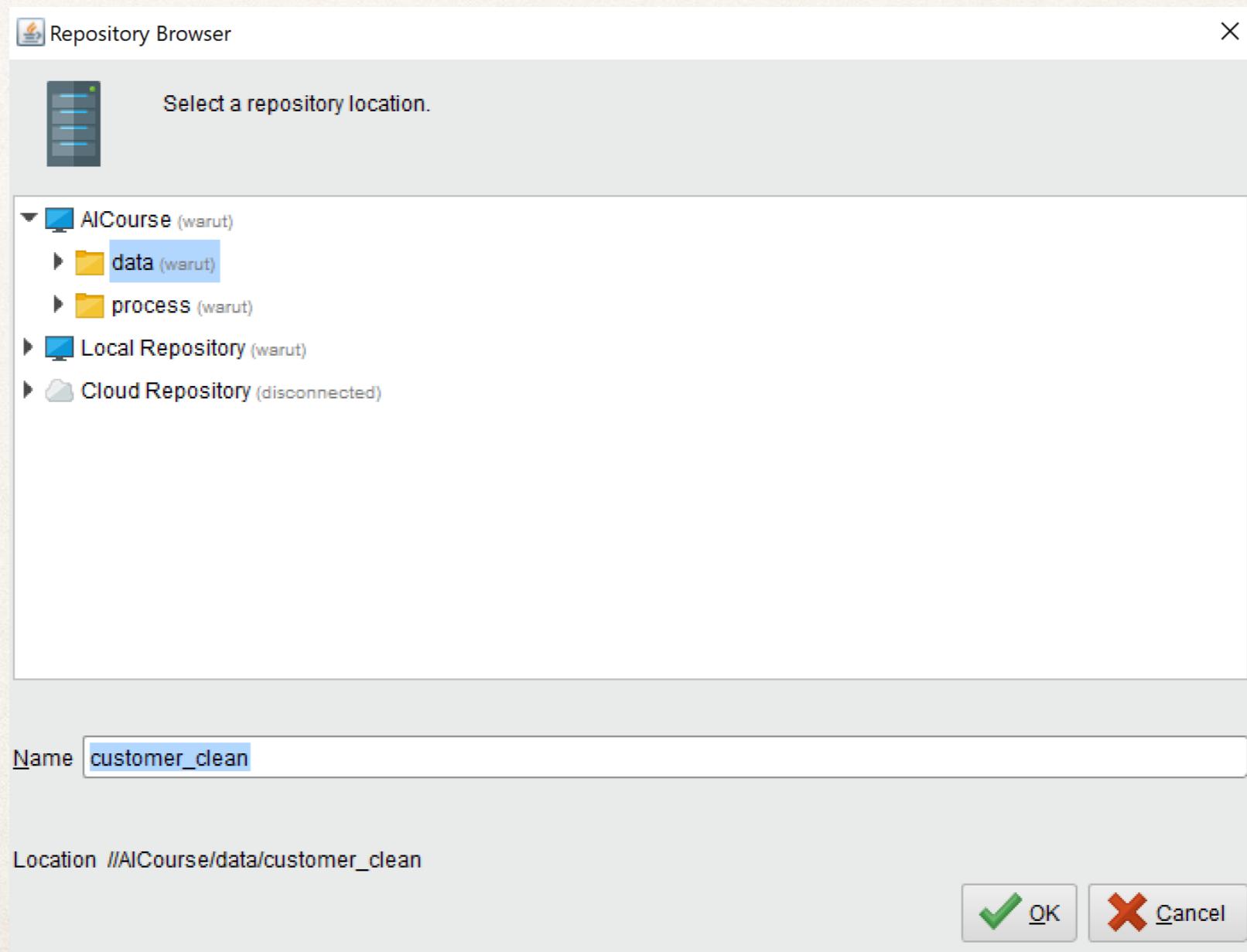
Replace Missing Data

- ✿ No missing data for `region` and `children` attributes

Result History		ExampleSet (Replace Missing Values Children)		ExampleSet (//AI.Course/data/customers)	
	Name	Type	Missing	Statisti...	Filter (9 / 9 attributes): Search for Attributes
Data	Id ↳ <code>customer_id</code>	Polynomial	0	Least ID12700 (1)	Most ID12101 (1) Value ID12
Statistics	Label ↳ <code>response</code>	Polynomial	0	Least NO (186)	Most YES (414) Value YES
Visualizations	children	Integer	0	Min 0	Max 3 Average 1.00 Value
Annotations	region	Polynomial	0	Least N/A (6)	Most INNER_CITY (265) Value INNER_CITY
	gender	Polynomial	0	Least MALE (300)	Most FEMALE (300) Value FEM
	age	Integer	0	Min 2	Max 999 Average 43.8

Replace Missing Data

- ❖ Save the data into the repository as `customer_clean`



Data Preparation

- ✿ Preprocessing
 - ✿ Select attributes
 - ✿ Select by type of attributes
 - ✿ Select by specific attributes
 - ✿ Filter examples by conditions
 - ✿ Join data from multiple sources
- ✿ Deal with incomplete data
 - ✿ Inconsistent data
 - ✿ Missing data
- ✿ **Data transformation**
 - ✿ **Discretization (numeric to nominal)**
 - ✿ **User defined**
 - ✿ Equal frequency

Discretization: User Defined

- ❖ Transform numerical data to nominal data based on user defined conditions.

No.	Age
1	15
2	20
3	20
4	20
5	25
6	40
7	45
8	45
9	50

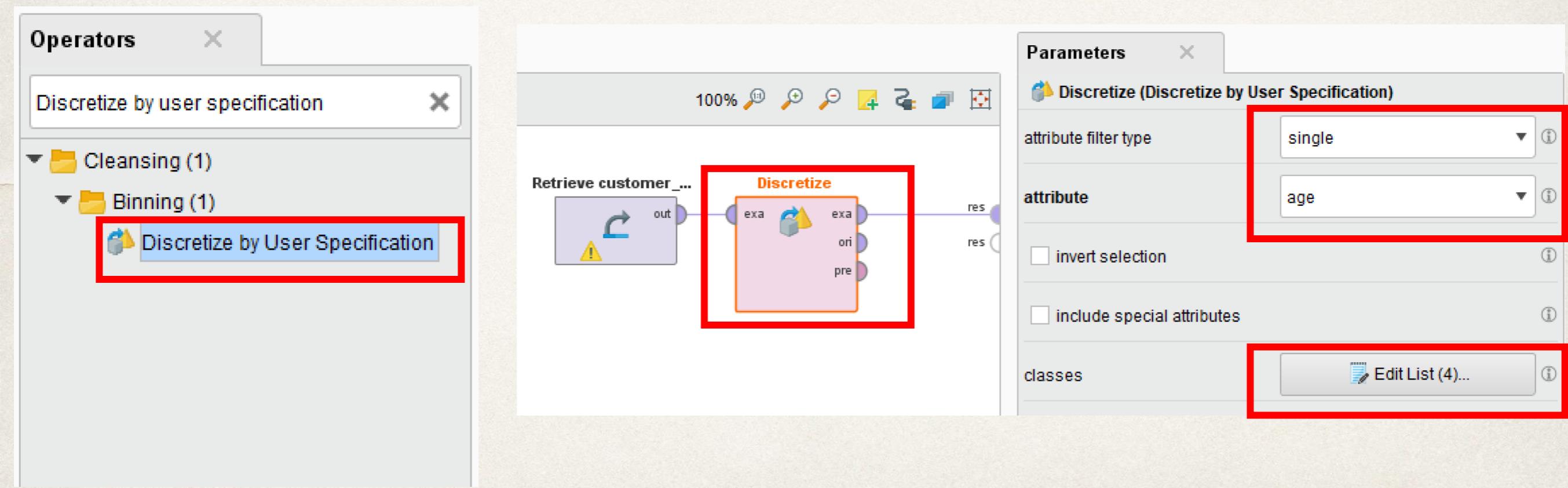
Conditions:

- $18 \leq$ Financially Dependent (FD)
- > 18 and ≤ 30 Young Professional (YP)
- > 30 and ≤ 45 Married Professional (MP)
- > 45 and ≤ 60 Empty Nesters (EN)
- > 60 and ≤ 100 Retired ®

No.	Age
1	FD
2	YP
3	YP
4	YP
5	YP
6	MP
7	MP
8	MP
9	EN

Discretization: User Defined

- ❖ Choose Discretize by User Specification operator



Discretization: User Defined

- ❖ Set the conditions based on user defined conditions

Edit Parameter List: classes X

 Edit Parameter List: classes
Defines the classes and the upper limits of each class.

class names	upper limit
FD	18.0
YP	30.0
MP	45.0
EN	60.0

 Add Entry  Remove Entry  Apply  Cancel

Discretization: User Defined

- ❖ The result after transformation

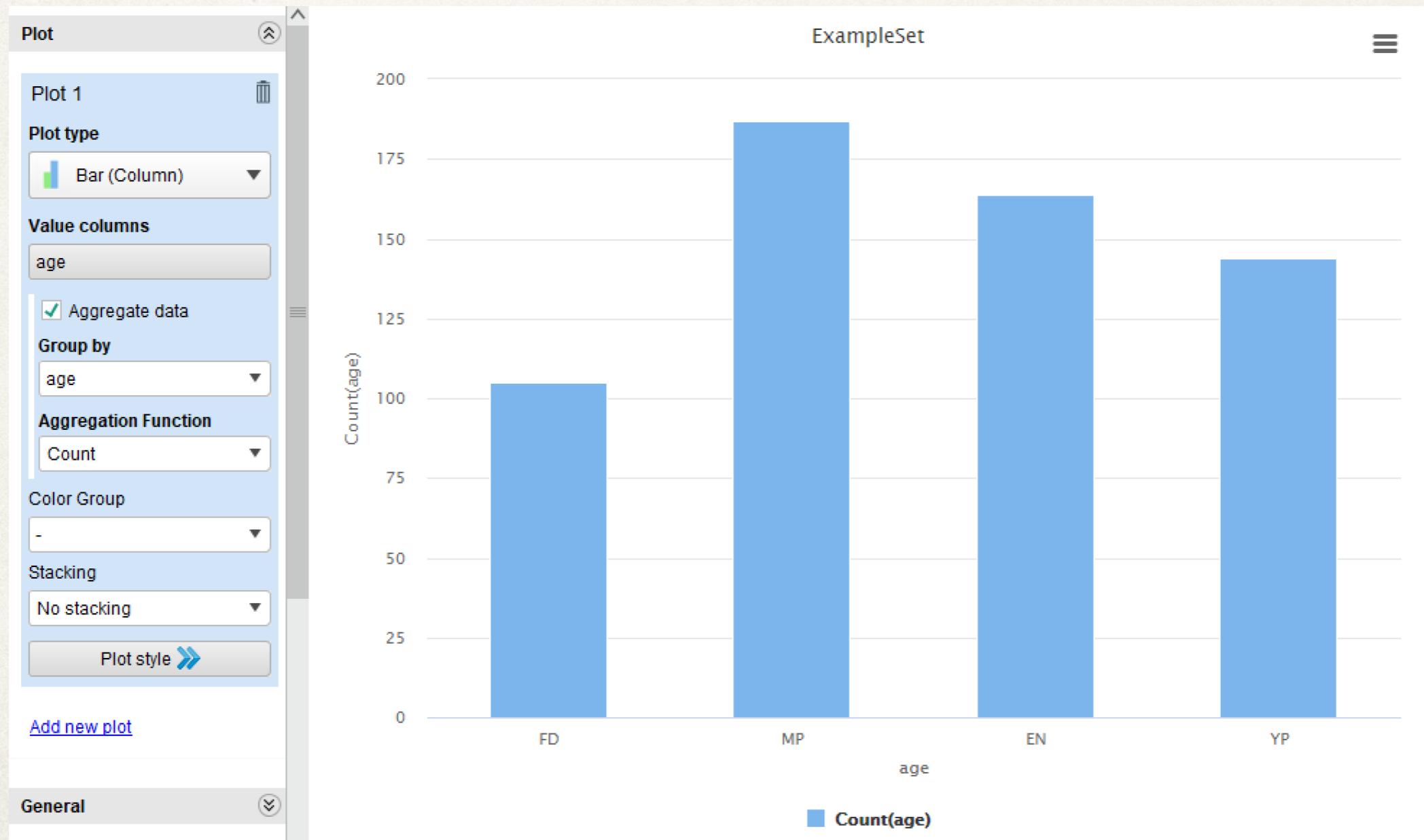
Result History X ExampleSet (Discretize) X

ExampleSet (600 examples, 2 special attributes, 7 regular attributes) Filter (600 / 600 examples): all ▾

Row No.	customer_id	response	age	children	region	gender	income	married
1	ID12101	NO	EN	1	N/A	FEMALE	17546	NO
2	ID12102	NO	MP	3	TOWN	MALE	30085.100	YES
3	ID12103	YES	EN	0	INNER_CITY	FEMALE	16575.400	YES
4	ID12104	NO	YP	3	TOWN	FEMALE	20375.400	YES
5	ID12105	YES	EN	0	RURAL	FEMALE	50576.300	YES
6	ID12106	YES	EN	2	TOWN	FEMALE	37869.600	YES
7	ID12107	NO	YP	0	RURAL	MALE	8877.070	NO
8	ID12108	YES	EN	0	TOWN	MALE	24946.600	YES
9	ID12109	NO	MP	2	SUBURBAN	FEMALE	25304.300	YES
10	ID12110	YES	EN	2	TOWN	MALE	24212.100	YES

Discretization: User Defined

- ❖ The result after transformation

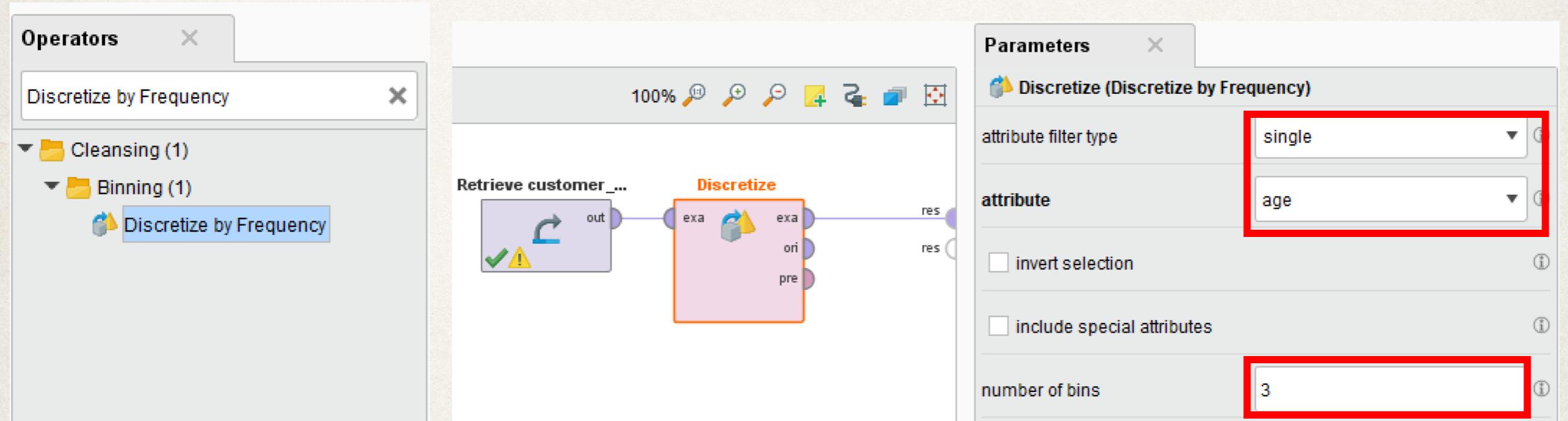


Data Preparation

- ✿ Preprocessing
 - ✿ Select attributes
 - ✿ Select by type of attributes
 - ✿ Select by specific attributes
 - ✿ Filter examples by conditions
 - ✿ Join data from multiple sources
- ✿ Deal with incomplete data
 - ✿ Inconsistent data
 - ✿ Missing data
- ✿ **Data transformation**
 - ✿ **Discretization (numeric to nominal)**
 - ✿ User defined
 - ✿ **Equal frequency**

Discretization: Equal Frequency

- ❖ Transform numerical data to nominal data by discretizing the numerical data into a user-specified number of bins
- ❖ Choose Discretize by Frequency operator



Discretization: Equal Frequency

- ❖ The result after transformation

Result History ExampleSet (Discretize) X

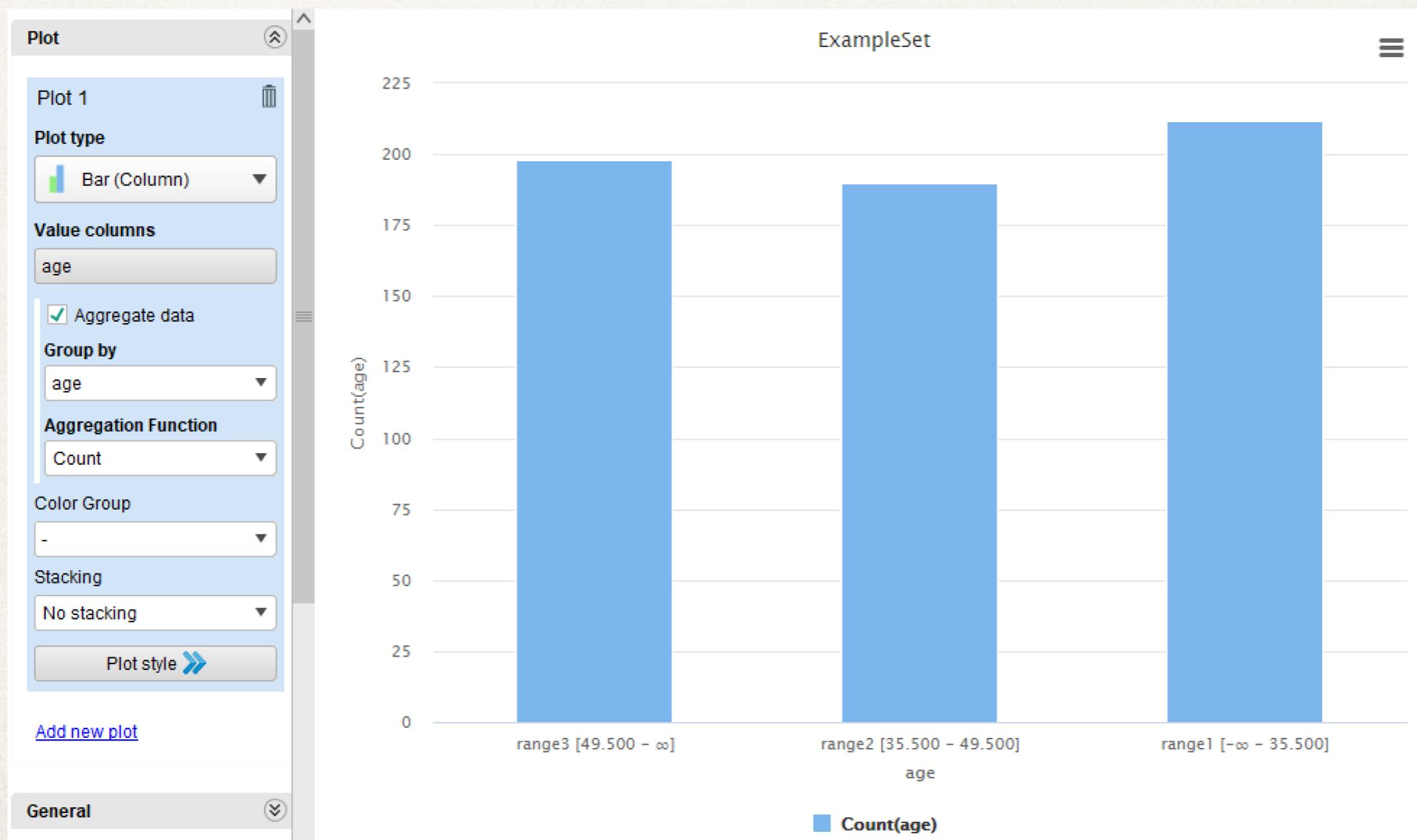
Open in [Turbo Prep](#) [Auto Model](#) Filter (600 / 600 examples): all

Data Statistics Visualizations Annotations

Row No.	ID	response	age	children	region	gender	income	married
1	ID12101	NO	range3 [49.500 - ∞]	1	N/A	FEMALE	17546	NO
2	ID12102	NO	range2 [35.500 - 49.500]	3	TOWN	MALE	30085.100	YES
3	ID12103	YES	range3 [49.500 - ∞]	0	INNER_CITY	FEMALE	16575.400	YES
4	ID12104	NO	range1 [-∞ - 35.500]	3	TOWN	FEMALE	20375.400	YES
5	ID12105	YES	range3 [49.500 - ∞]	0	RURAL	FEMALE	50576.300	YES
6	ID12106	YES	range3 [49.500 - ∞]	2	TOWN	FEMALE	37869.600	YES
7	ID12107	NO	range1 [-∞ - 35.500]	0	RURAL	MALE	8877.070	NO
8	ID12108	YES	range3 [49.500 - ∞]	0	TOWN	MALE	24946.600	YES
9	ID12109	NO	range2 [35.500 - 49.500]	2	SUBURBAN	FEMALE	25304.300	YES
10	ID12110	YES	range3 [49.500 - ∞]	2	TOWN	MALE	24212.100	YES

Discretization: Equal Frequency

- ❖ The result after transformation



Reference

Ekasit Pacharawongsakda (2017). *Practical Data Mining with RapidMiner Studio 8.*

<http://www.dataminingtrend.com>

<http://facebook/datacube.th>