

SE634: Artificial Intelligence in Industrial Engineering

Data preprocessing and introduction to RapidMiner Studio

Introduction to RapidMiner Studio



<https://rapidminer.com/>

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

Gartner®

- ❖ 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.




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




Market Basket Analysis
Find products frequently



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

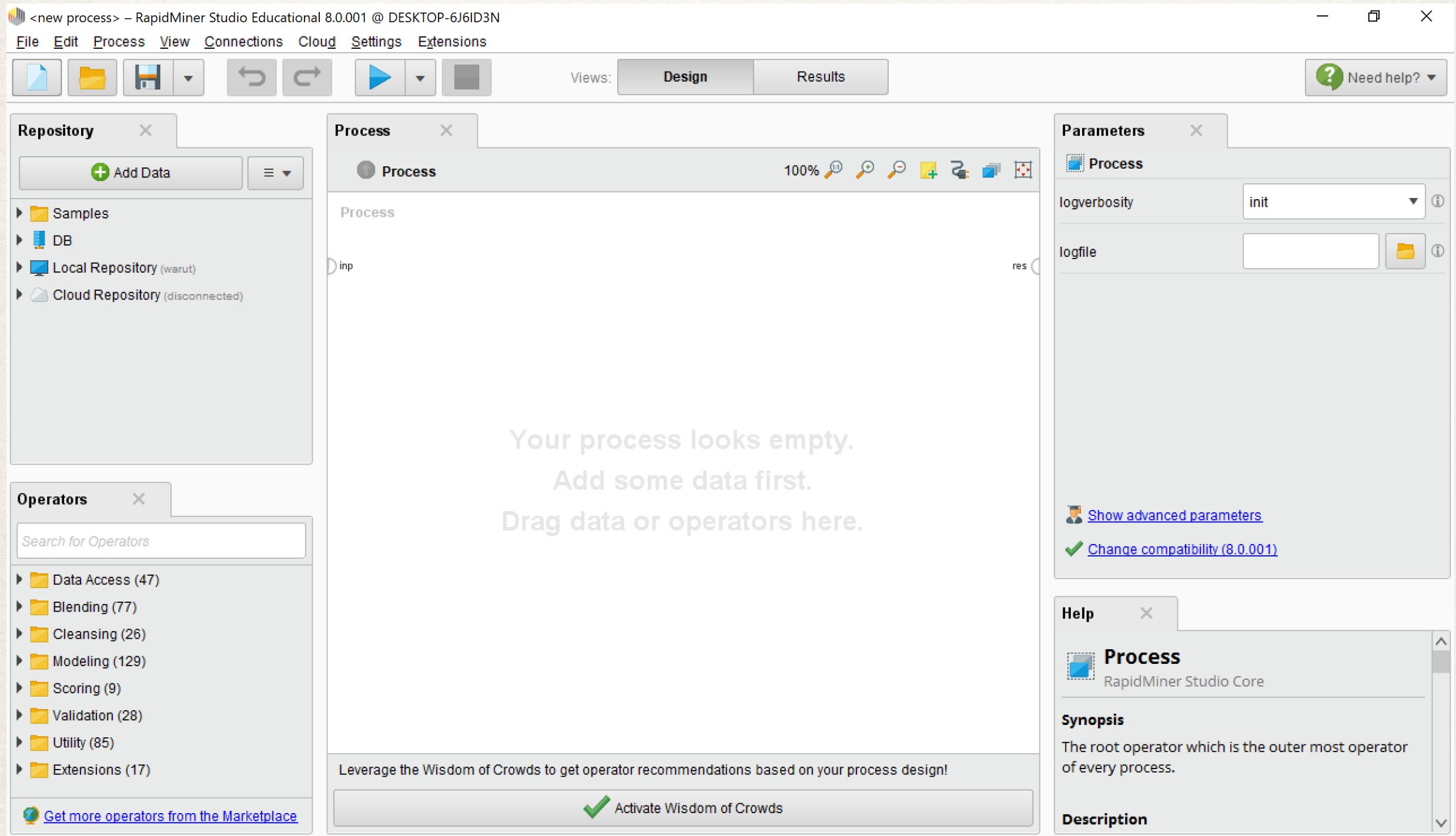


Credit Risk Modeling
Model credit default risk by training an optimized Support Vector Machine (SVM) model.

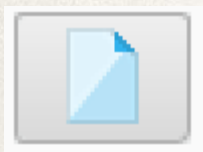
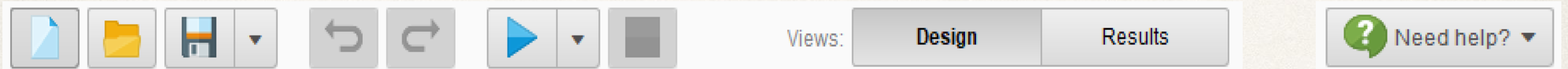


Predictive Maintenance
Model equipment failures to

Interface



Menu



Create new process



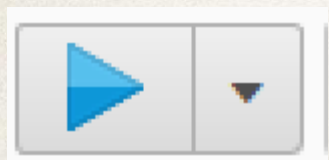
Open existing process



Save



Undo and redo

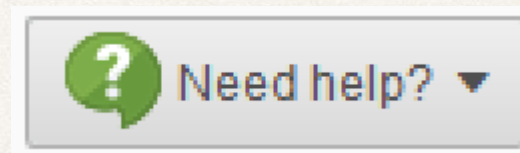


Run



Go to design view

Go to result view



Support and help resources

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 |

ID

Attribute

Label

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

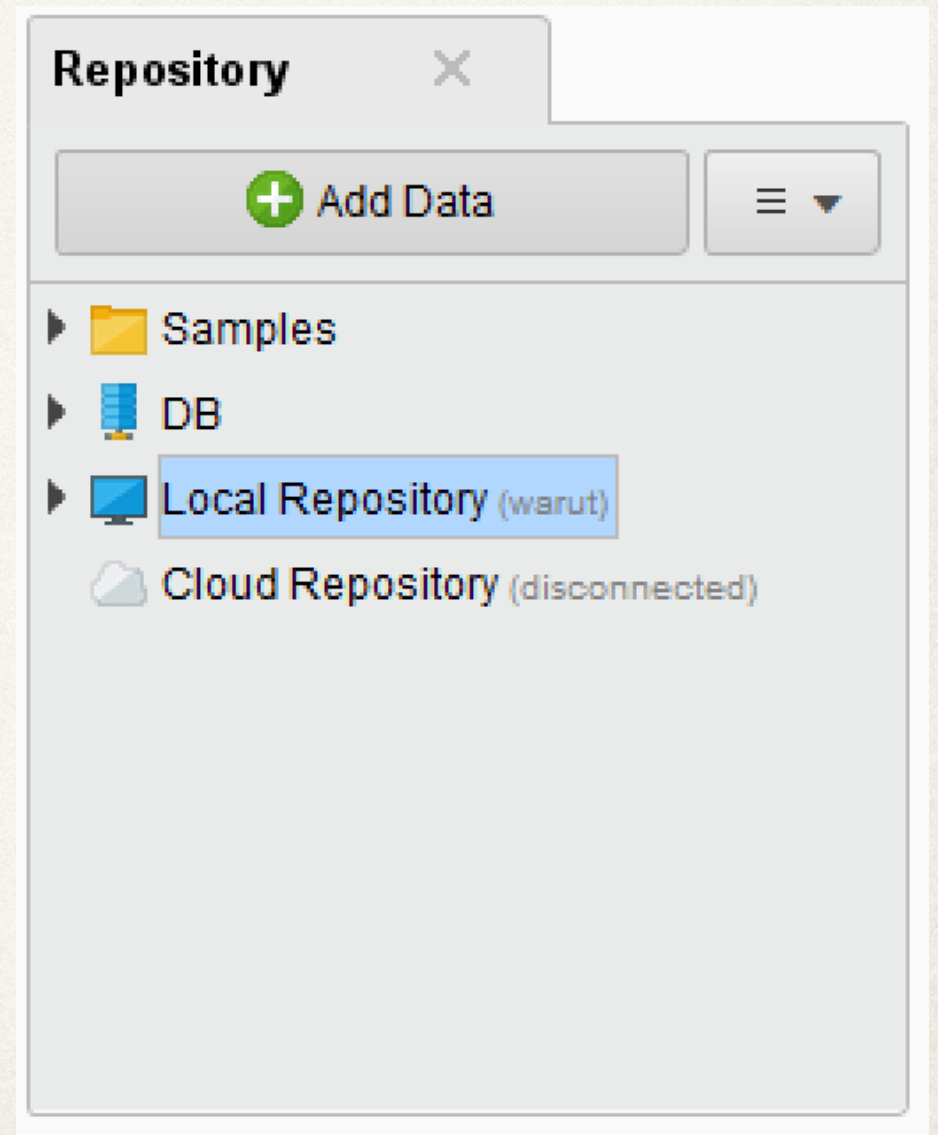
Binomial

Numeric

Binomial

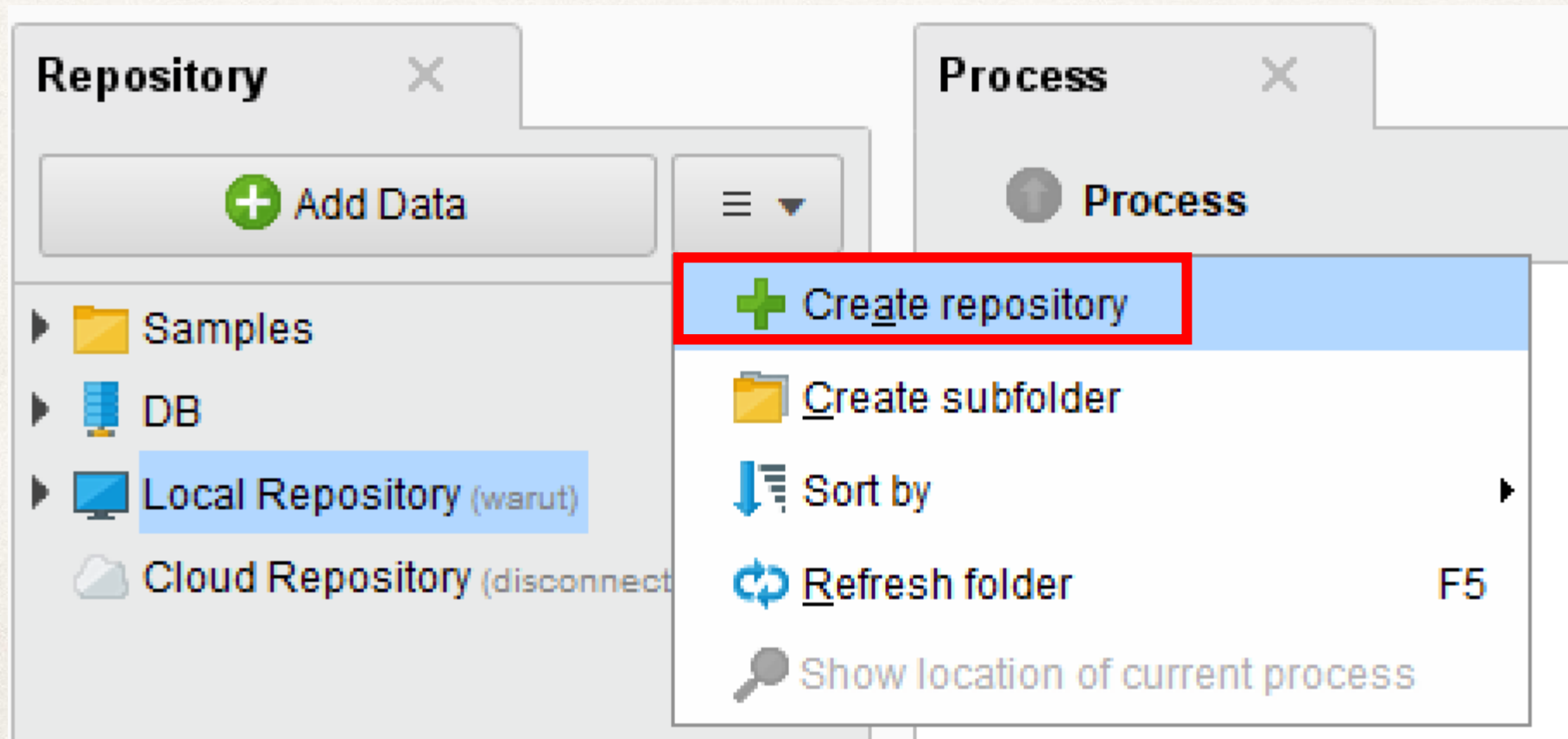
Data Management

- ❖ 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



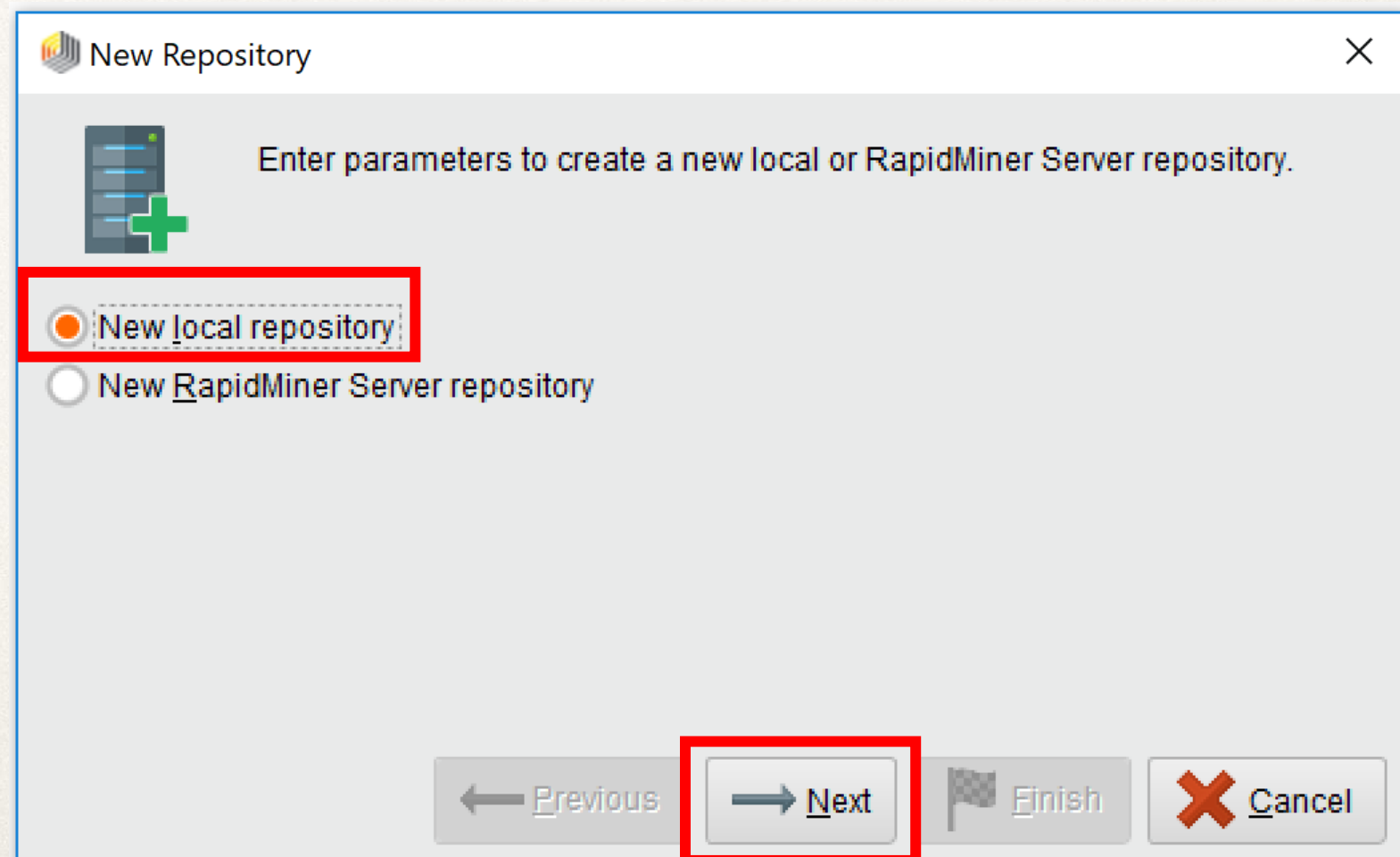
Data Management

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



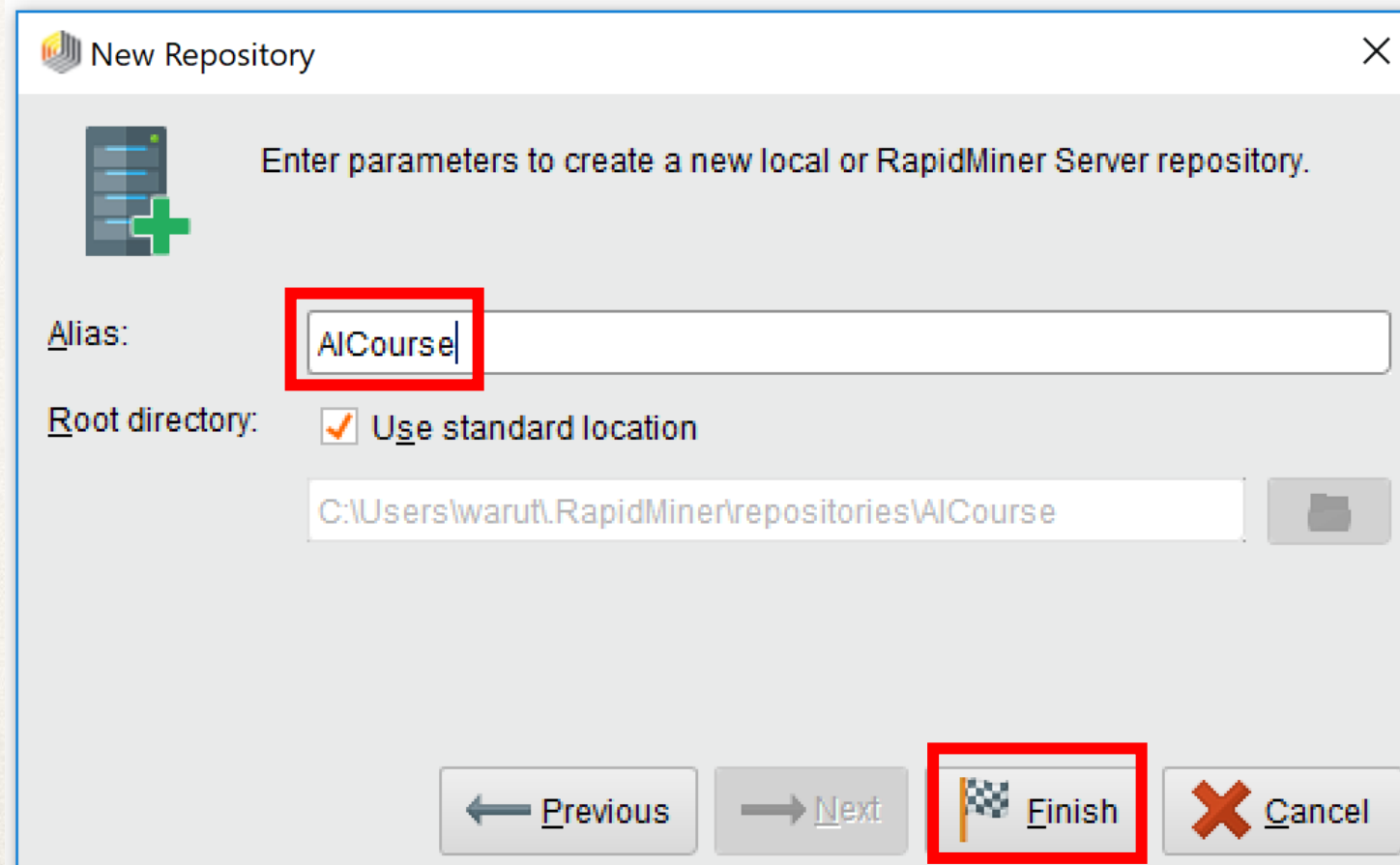
Data Management

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



Data Management

- ❖ Create new repository (Cont.)
 - ❖ Change Alias to “AICourse”
 - ❖ Press “Finish”



New Repository

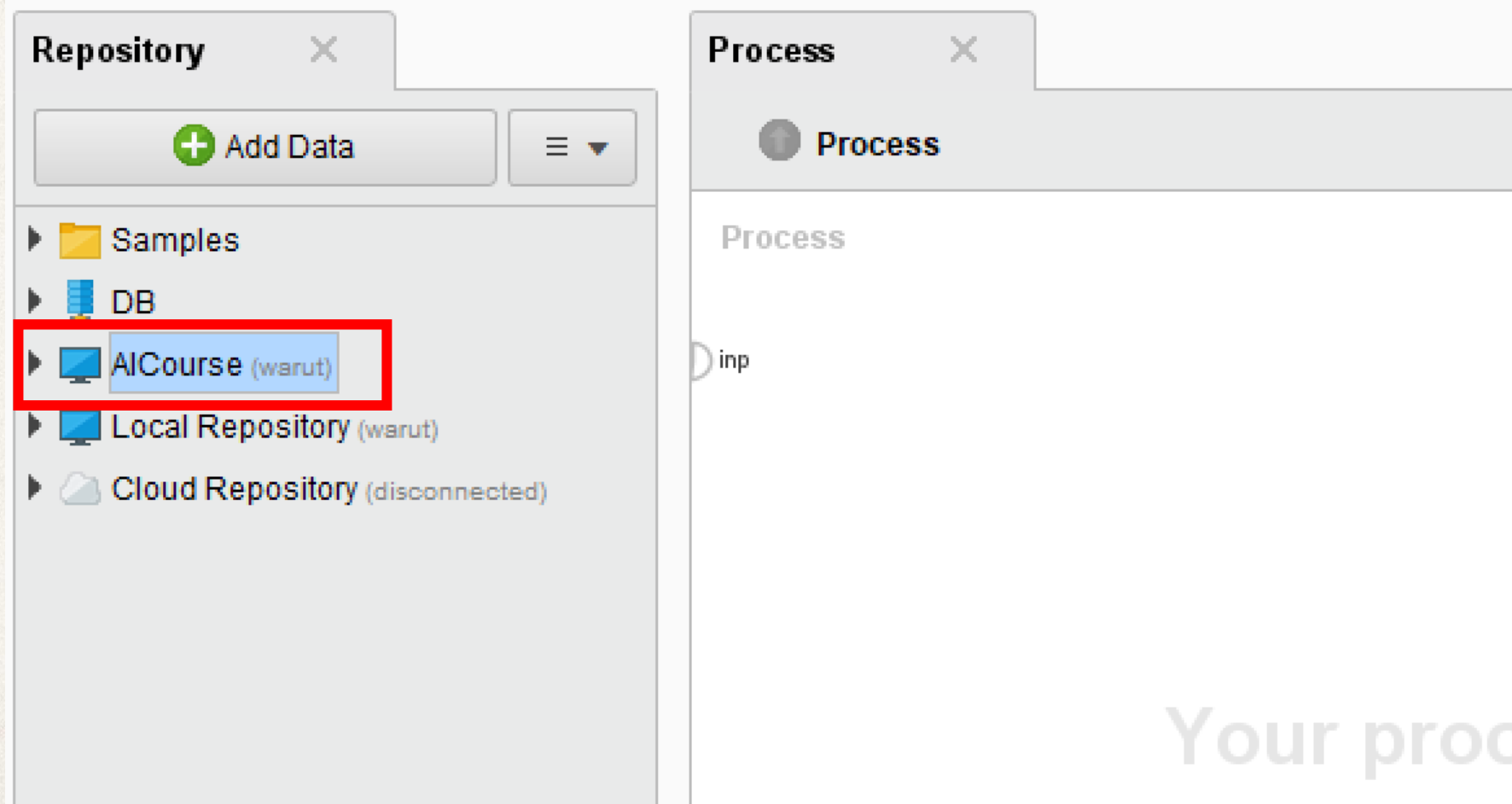
Enter parameters to create a new local or RapidMiner Server repository.

Alias:

Root directory: ☒ Use standard location

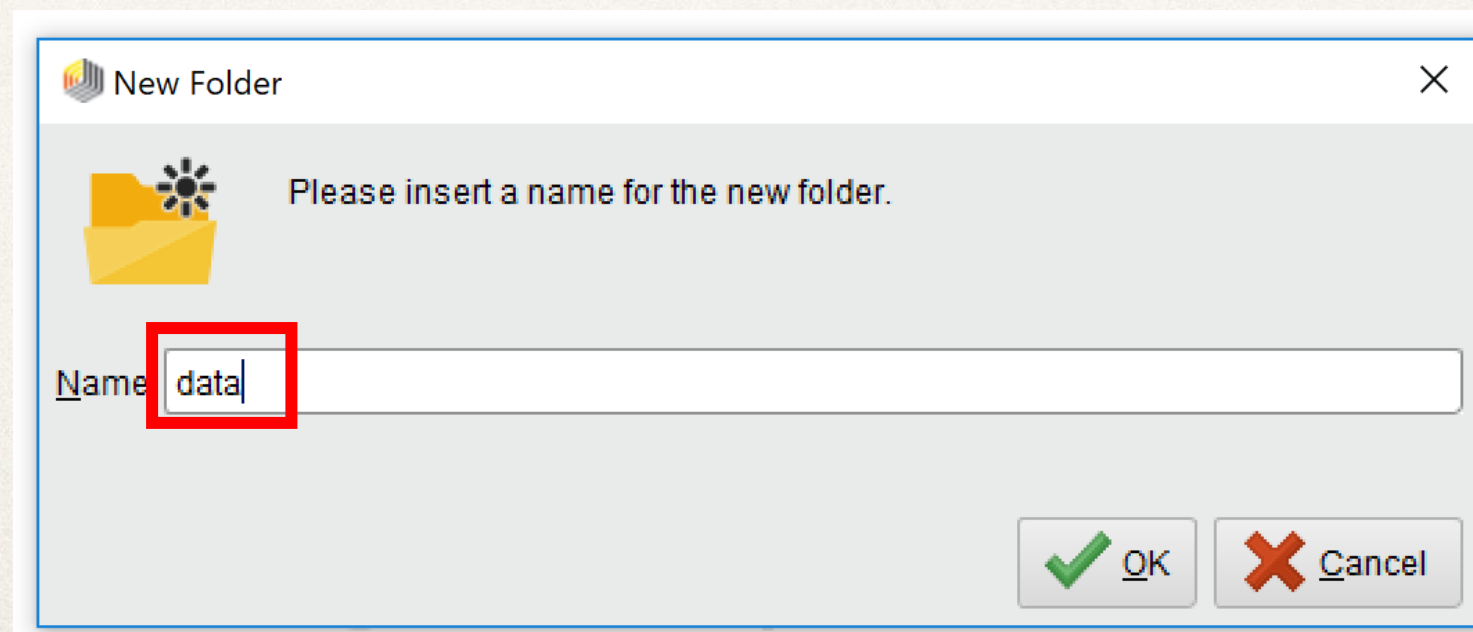
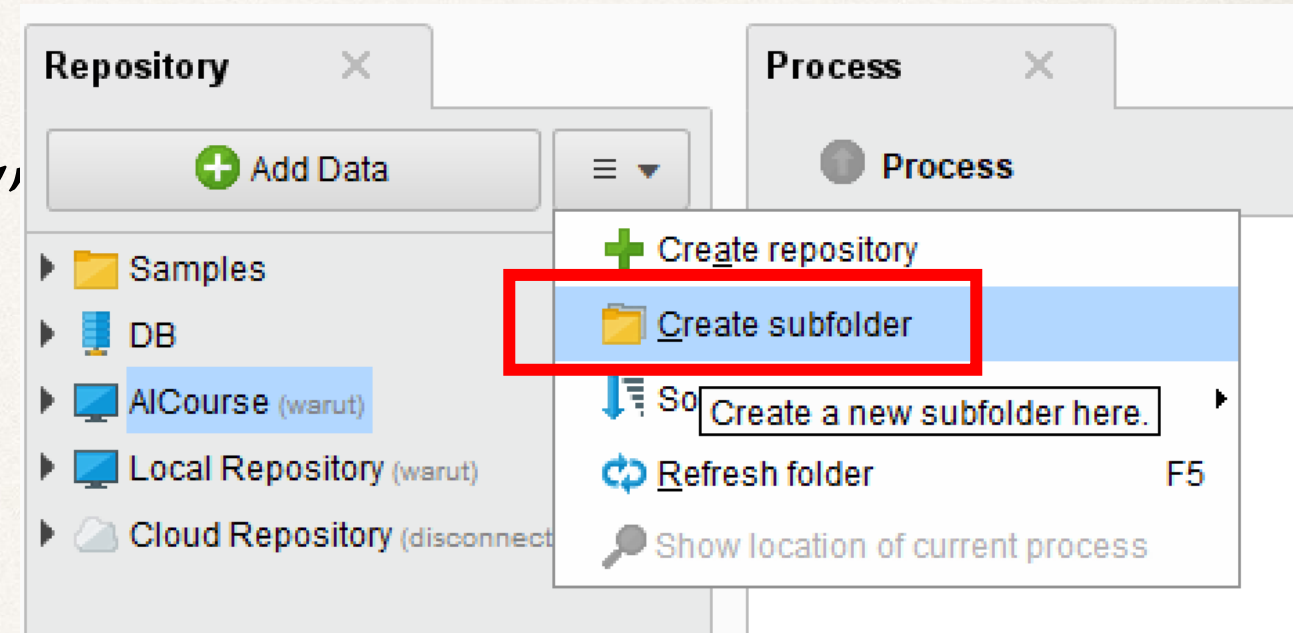
Data Management

- ❖ “AICourse” will be appeared in repository.



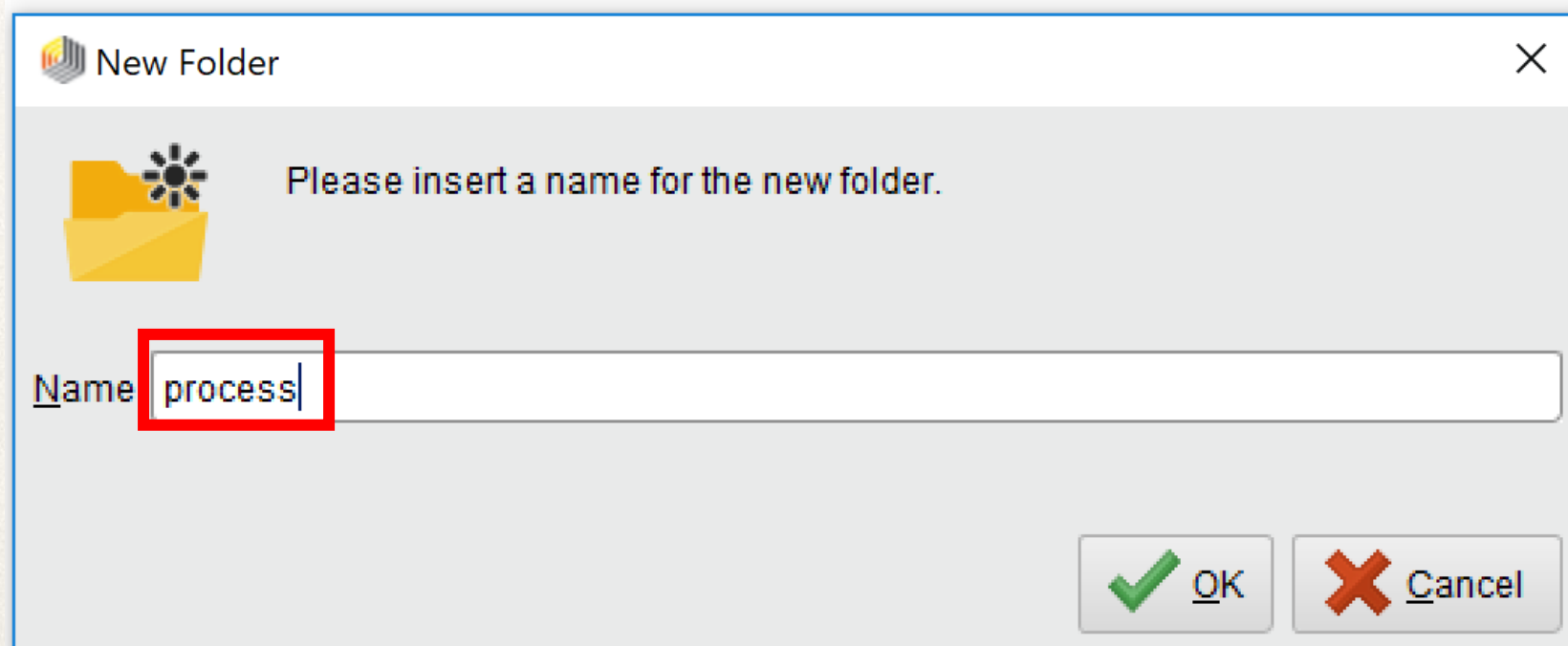
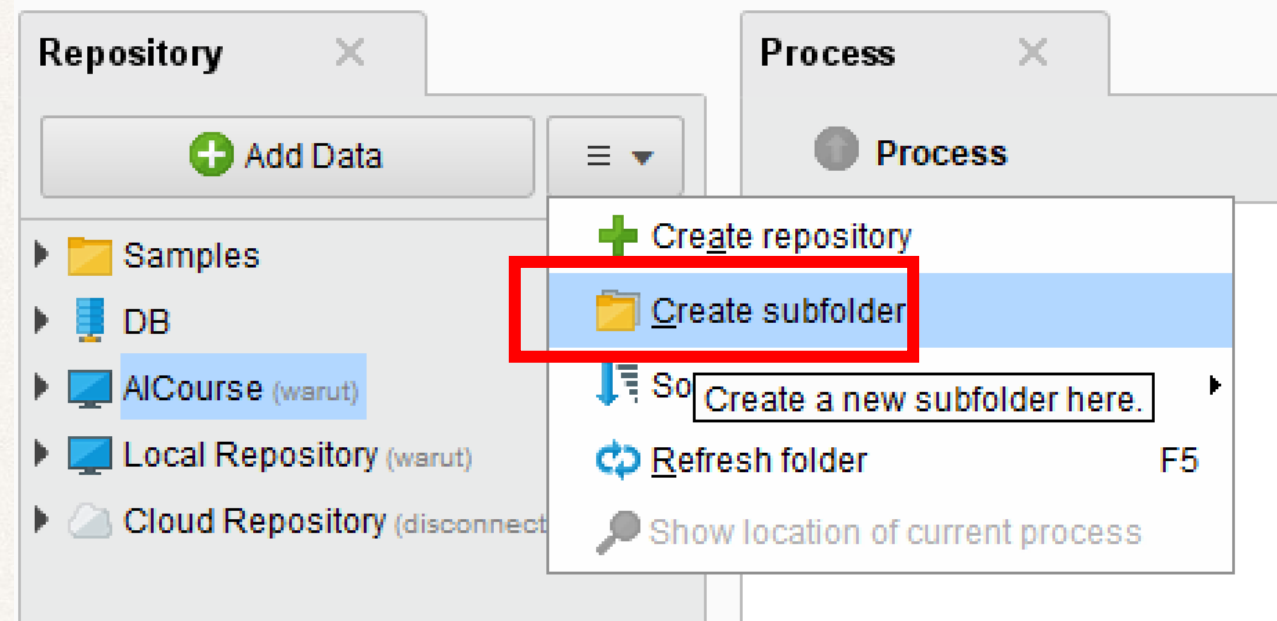
Data Management

- ❖ Create folder in “AICourse”
- ❖ “data” for data storage



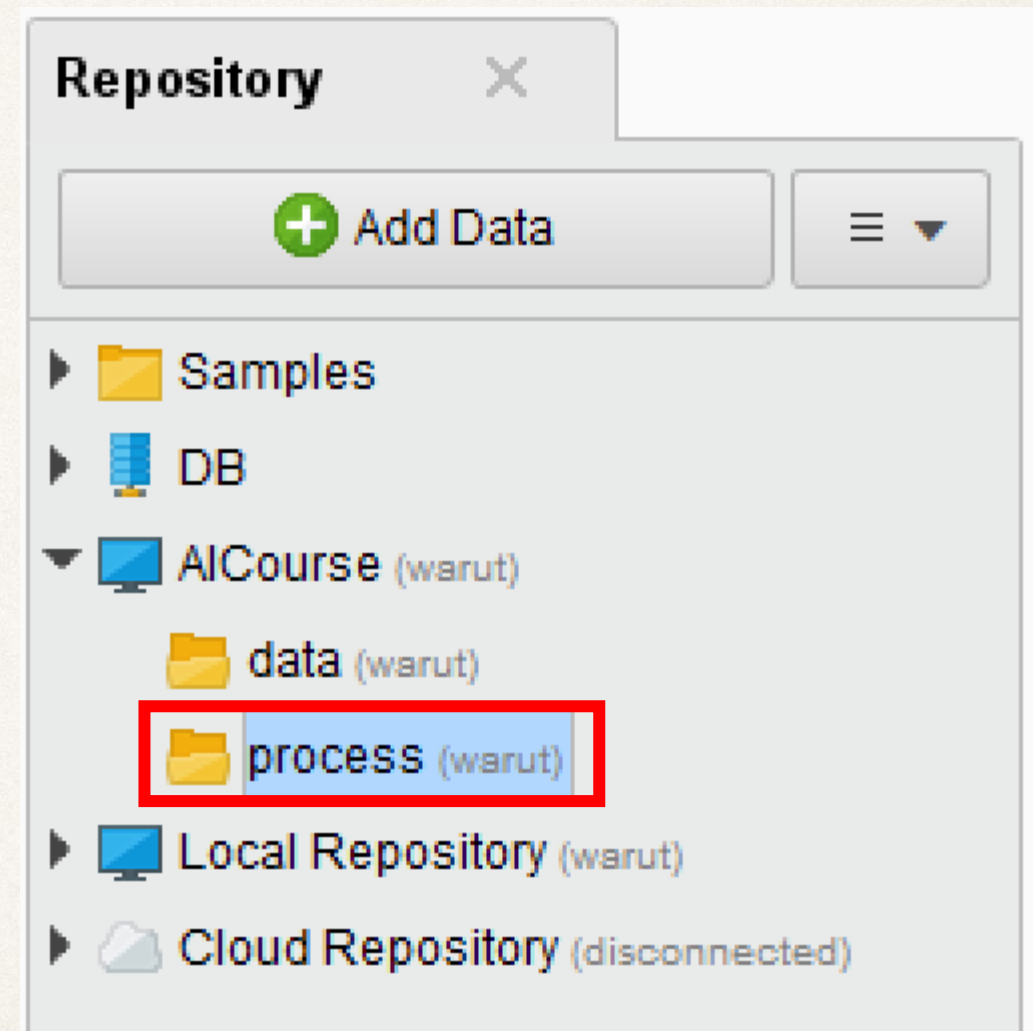
Data Management

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



Data Management


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

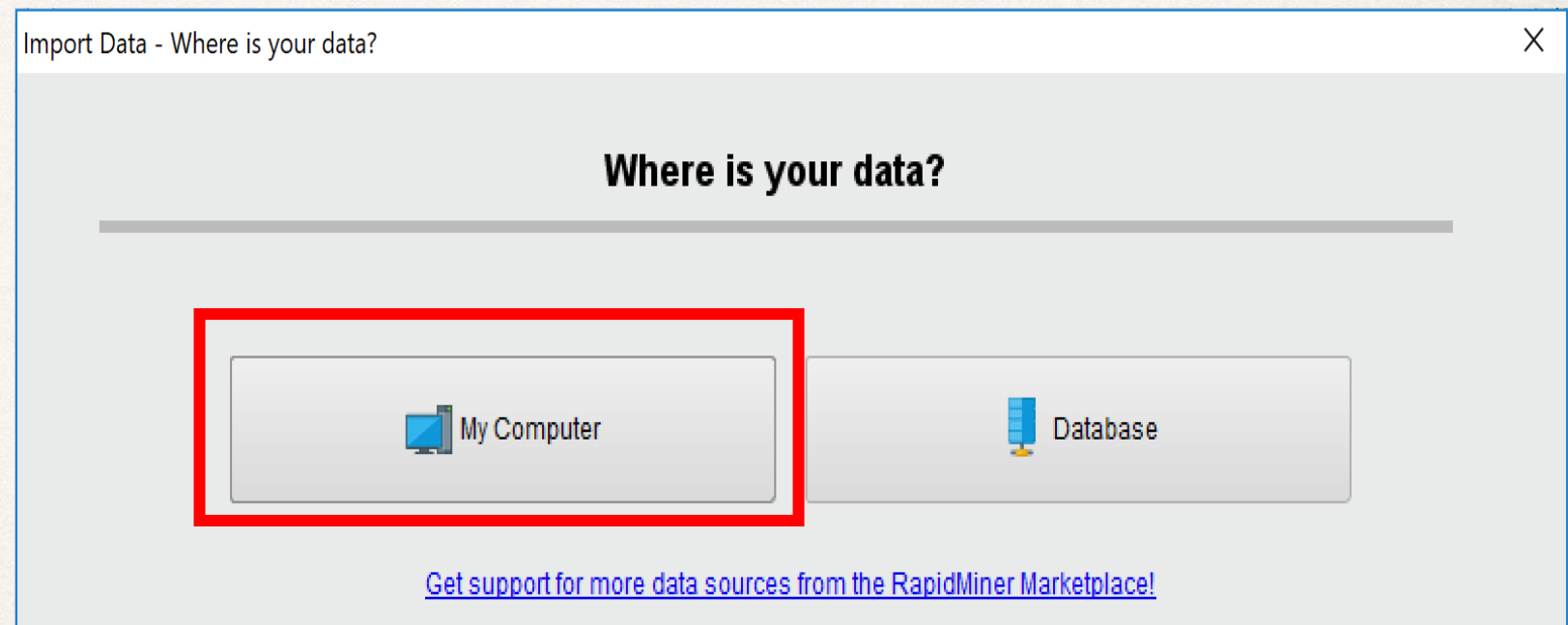
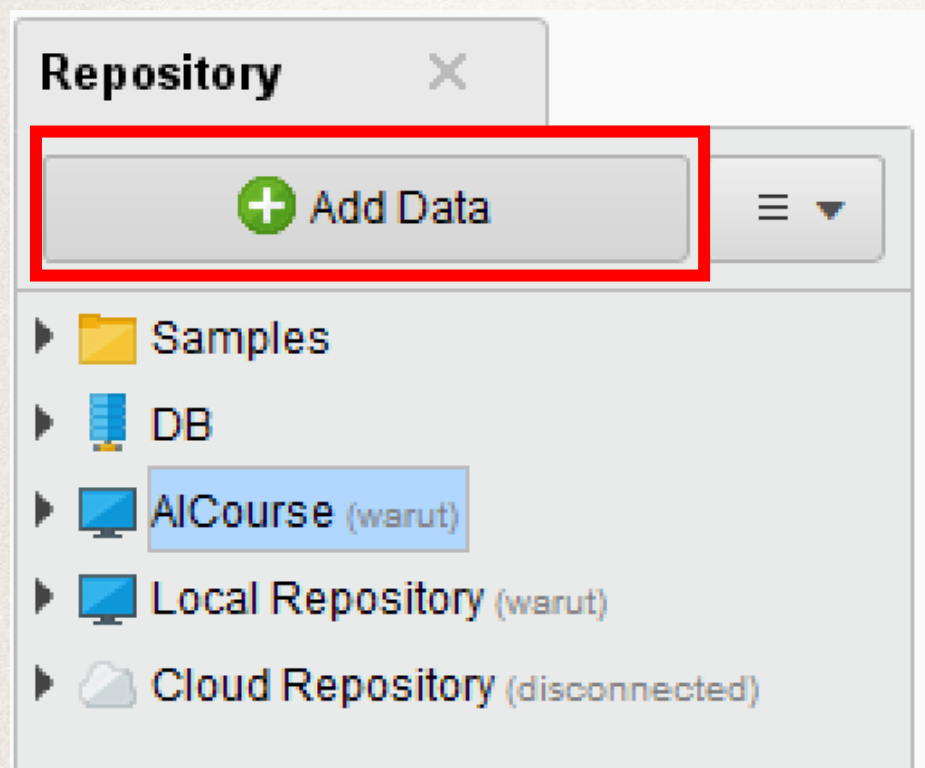


Data Pre-Processing with RapidMiner

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

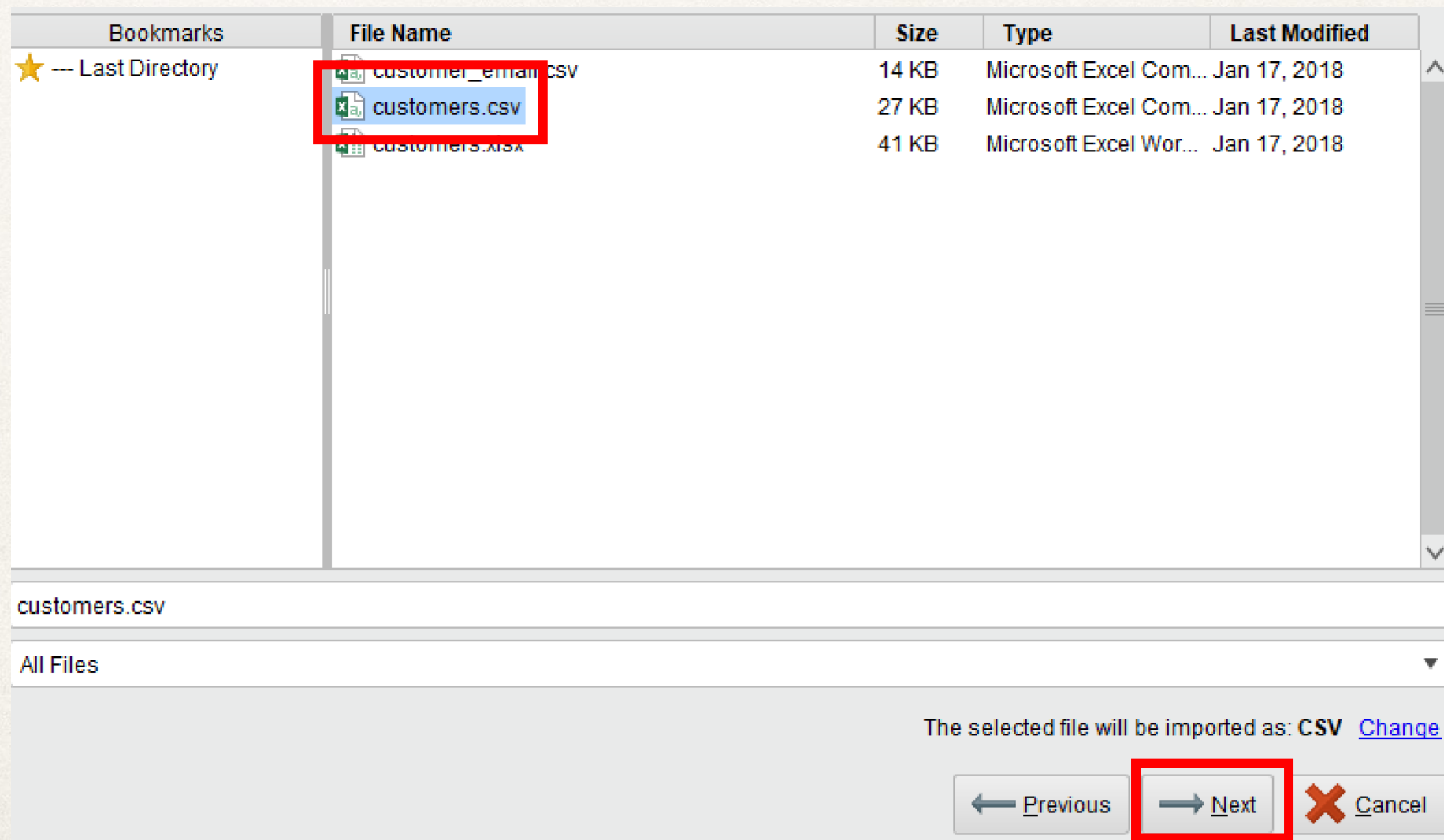
Import Data into Repository

- ❖ Click 
- ❖ Choose “My Computer”



Import Data into Repository

- ❖ Choose “customer.csv”



Import Data into Repository

- ❖ For Column Separator, choose Comma “,”

Import Data - Specify your data format

Specify your data format

☒ Header Row
Start Row
Column Separator **Comma “,”**
File Encoding
Escape Character
Decimal Character
☒ Use Quotes
☐ Trim Lines
☒ Skip Comments

| | customer... | age | gender | region | income | married | children | car | response |
|----|-------------|-----|--------|-------------|---------|---------|----------|-----|----------|
| 1 | ID12101 | 48 | FEMALE | | 17546 | NO | 1 | NO | NO |
| 2 | ID12102 | 40 | MALE | TOWN | 30085.1 | YES | 3 | YES | NO |
| 3 | ID12103 | 51 | FEMALE | INNER_CI... | 16575.4 | YES | 0 | YES | YES |
| 4 | ID12104 | 23 | FEMALE | TOWN | 20375.4 | YES | 3 | NO | NO |
| 5 | ID12105 | 57 | FEMALE | RURAL | 50576.3 | YES | 0 | NO | YES |
| 6 | ID12106 | 57 | WOMAN | TOWN | 37869.6 | YES | 2 | NO | YES |
| 7 | ID12107 | 22 | MALE | RURAL | 8877.07 | NO | 0 | NO | NO |
| 8 | ID12108 | 58 | MALE | TOWN | 24946.6 | YES | 0 | YES | YES |
| 9 | ID12109 | 37 | FEMALE | SUBURB... | 25304.3 | YES | 2 | YES | NO |
| 10 | ID12110 | 54 | MAN | TOWN | 24212.1 | YES | 2 | YES | YES |
| 11 | ID12111 | 5 | FEMALE | TOWN | 50000.0 | YES | 0 | NO | YES |

no problems.

Previous **Next** Cancel

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 | city |
|----|-------------|-----|------------|------------|-----------|------------|------|
| | polynomial | | polynomial | polynomial | real | polynomial | city |
| 1 | ID12101 | | FEMALE | ? | 17546.000 | NO | |
| 2 | ID12102 | | ? | ? | 30085.100 | YES | |
| 3 | ID12103 | | FEMALE | INNER_CITY | 16575.400 | | |
| 4 | ID12104 | 23 | FEMALE | TOWN | 20375.400 | | |
| 5 | ID12105 | 57 | FEMALE | RURAL | 50576.300 | | |
| 6 | ID12106 | 57 | WOMAN | TOWN | 37869.600 | | |
| 7 | ID12107 | 22 | MALE | RURAL | 8877.070 | | |
| 8 | ID12108 | 58 | MALE | TOWN | 24946.600 | | |
| 9 | ID12109 | 37 | FEMALE | SUBURBAN | 25304.300 | | |
| 10 | ID12110 | 54 | MAN | TOWN | 24212.100 | | |
| 11 | ID12111 | 6 | FEMALE | TOWN | 59803.900 | | |
| 12 | ID12112 | 52 | FEMALE | ? | 26658.800 | NO | |

Change role dialog:

Please enter the new role:

id

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 <i>polynomial</i> | income <i>real</i> | married <i>polynomial</i> | children <i>integer</i> | car <i>polynomial</i> | response <i>polynomial</i> |
|----|-----------------------------|-----------------------|------------------------------|----------------------------|--------------------------|-------------------------------|
| 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 | YES | |
| 11 | TOWN | 59803.900 | YES | 0 | NO | |
| 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 - Where to store the data? ×

Where to store the data?

▼ AICourse (warut)

- data (warut)
- ▶ process (warut)

▶ Local Repository (warut)

▶ Cloud Repository (disconnected)

Name **customers**

Location //AICourse/data/customers



← Previous **Finish** ✕ Cancel

Import Data into Repository

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

Result History

ExampleSet (//AICourse/data/customers) x

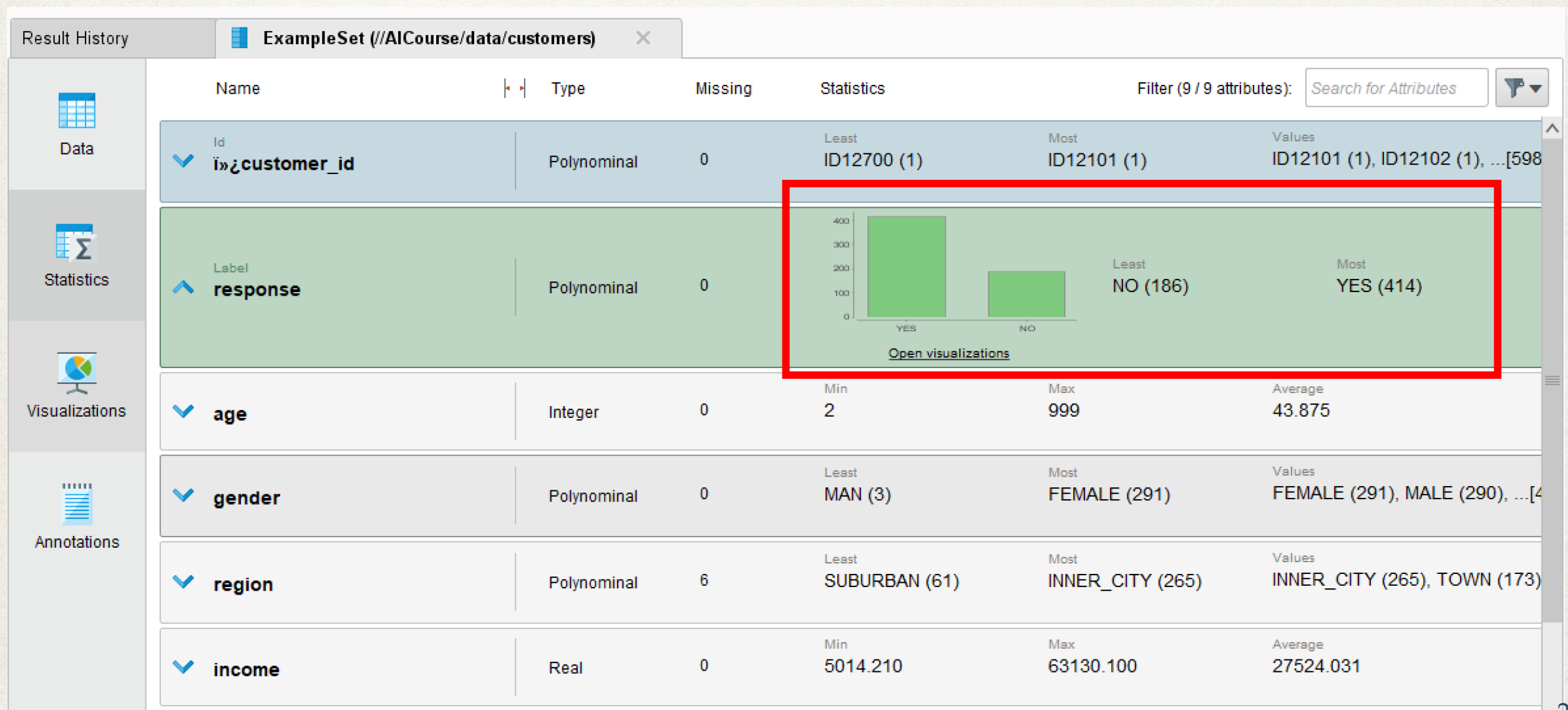
Open in  Turbo Prep  Auto Model

Filter (600 / 600 examples): all

| 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 |

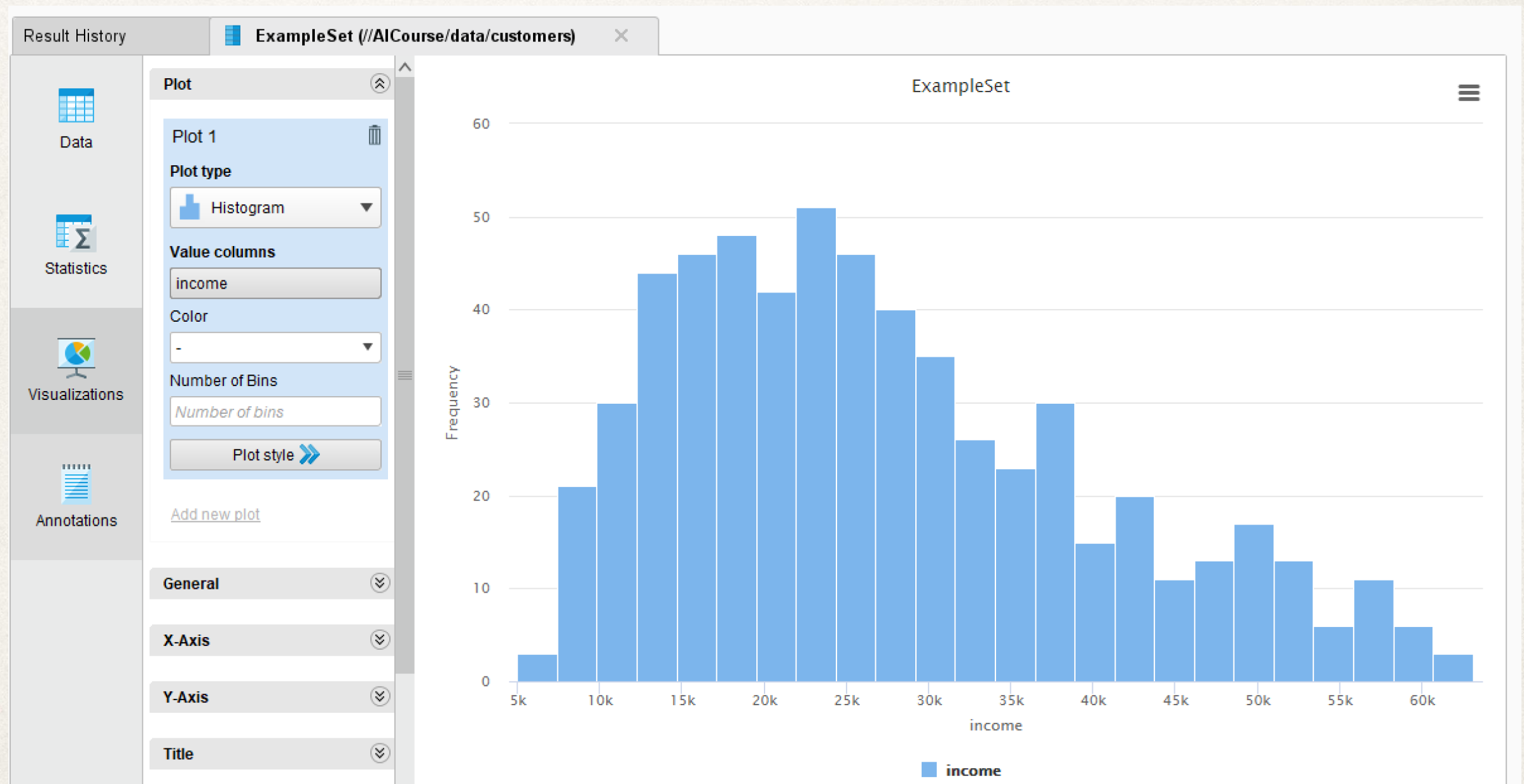
Import Data into Repository

❖ Statistics



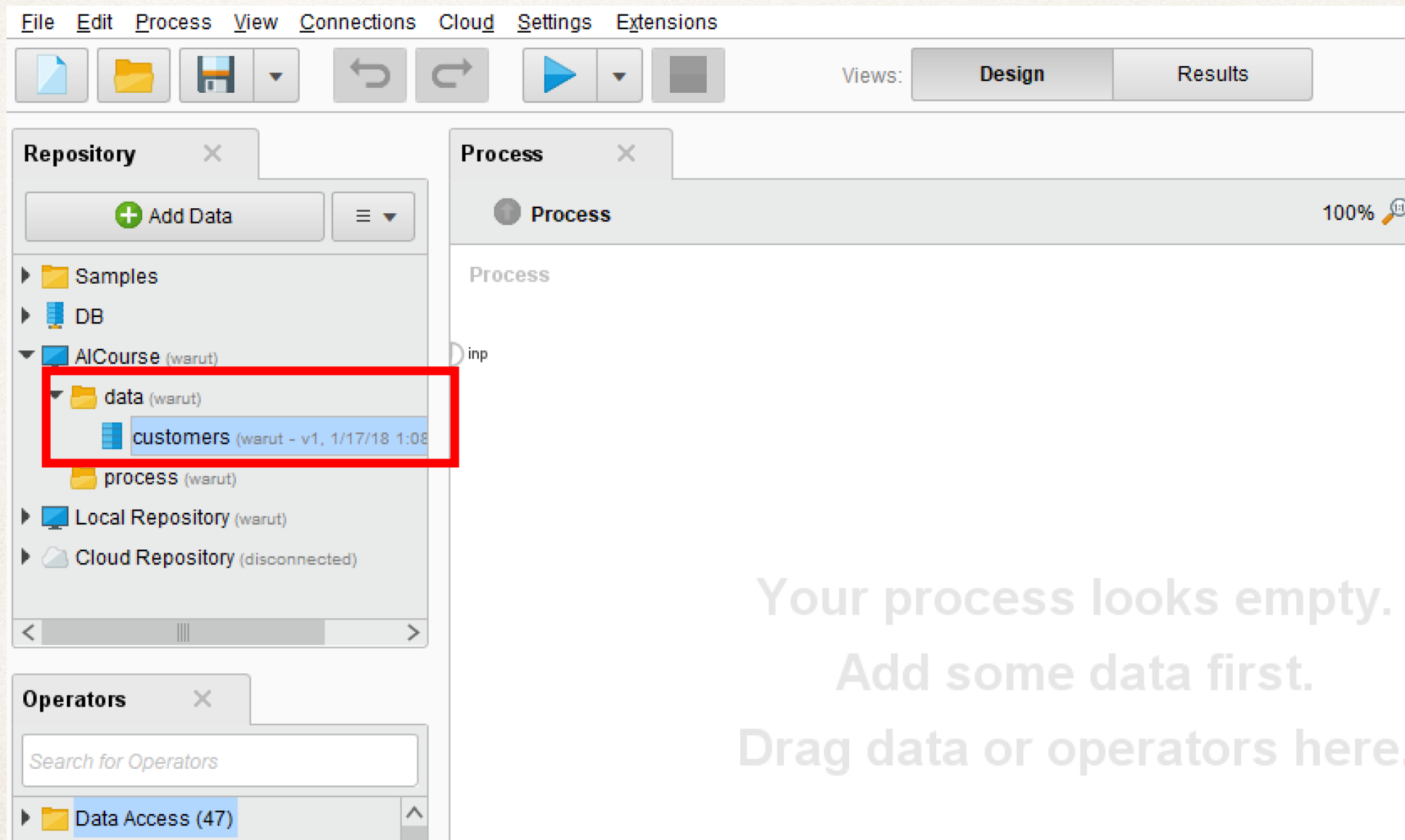
Import Data into Repository

✿ 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 (//AICourse/data/customers) X



Data


Statistics


Visualizations

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 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 |
|------------------------------|--|--|-----------------------|
| | | | all |
| | | | no_missing_attributes |
| | | | missing_attributes |
| | | | no_missing_labels |
| | | | missing_labels |
| | | | YES |
| | | | YES |
| | | | NO |
| | | | |
| | | | |

Data Exploration

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

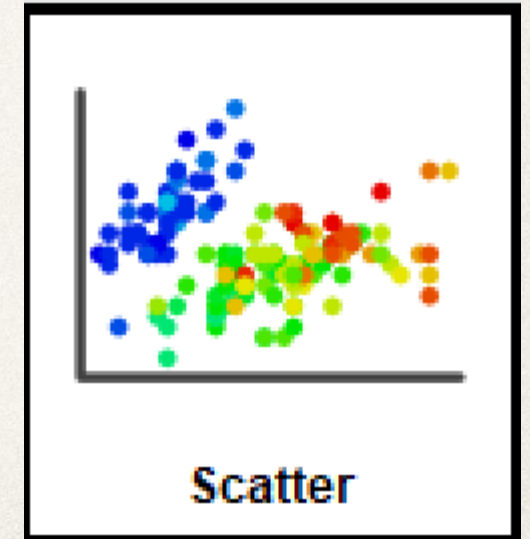
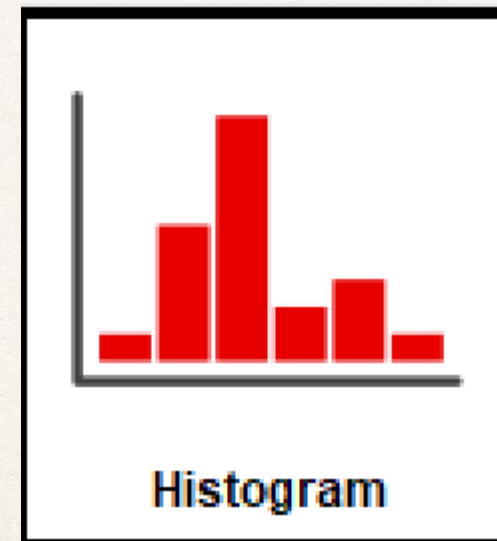
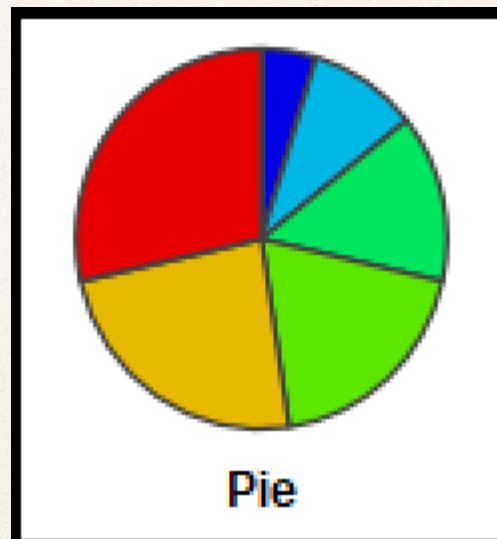
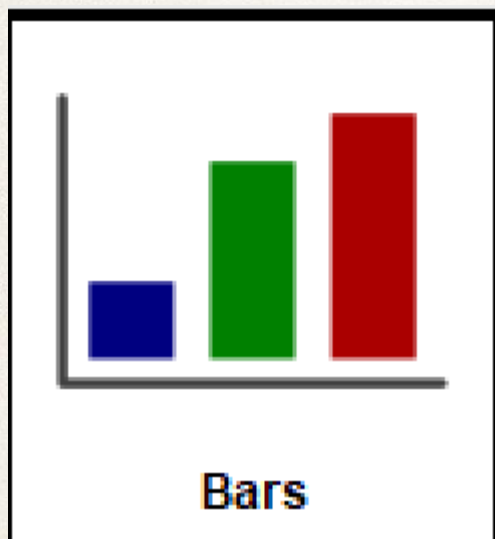
The screenshot shows a software interface with a menu bar (File, Edit, Process, View, Connections, Cloud, Settings, Extensions) and a toolbar with icons for file operations, undo, redo, and execution. Below the toolbar, there are tabs for 'Result History' and 'ExampleSet (Read CSV)'. On the left, a sidebar contains icons for 'Data' and 'Statistics'. The main area displays a table with the following columns: Name, Type, Missing, Statistics, and Filter (9 / 9 attributes). The table contains three rows of data:

| Name | Type | Missing | Statistics | Filter (9 / 9 attributes) |
|--|------------|---------|----------------------|---------------------------|
| <input checked="" type="checkbox"/> customer_id | Polynomial | 0 | Least ID12700 (1) | Most ID12101 (1) |
| <input checked="" type="checkbox"/> response | Polynomial | 0 | Least NO (186) | Most YES (414) |
| <input checked="" type="checkbox"/> age | Integer | 0 | Min 2 | Max 999 |

The 'Statistics' column for 'customer_id' also shows 'Values: ID12101 (1), ID12102'. The 'Statistics' column for 'response' also shows 'Values: YES (414), NO (186)'. The 'Statistics' column for 'age' also shows 'Average: 43.788'.

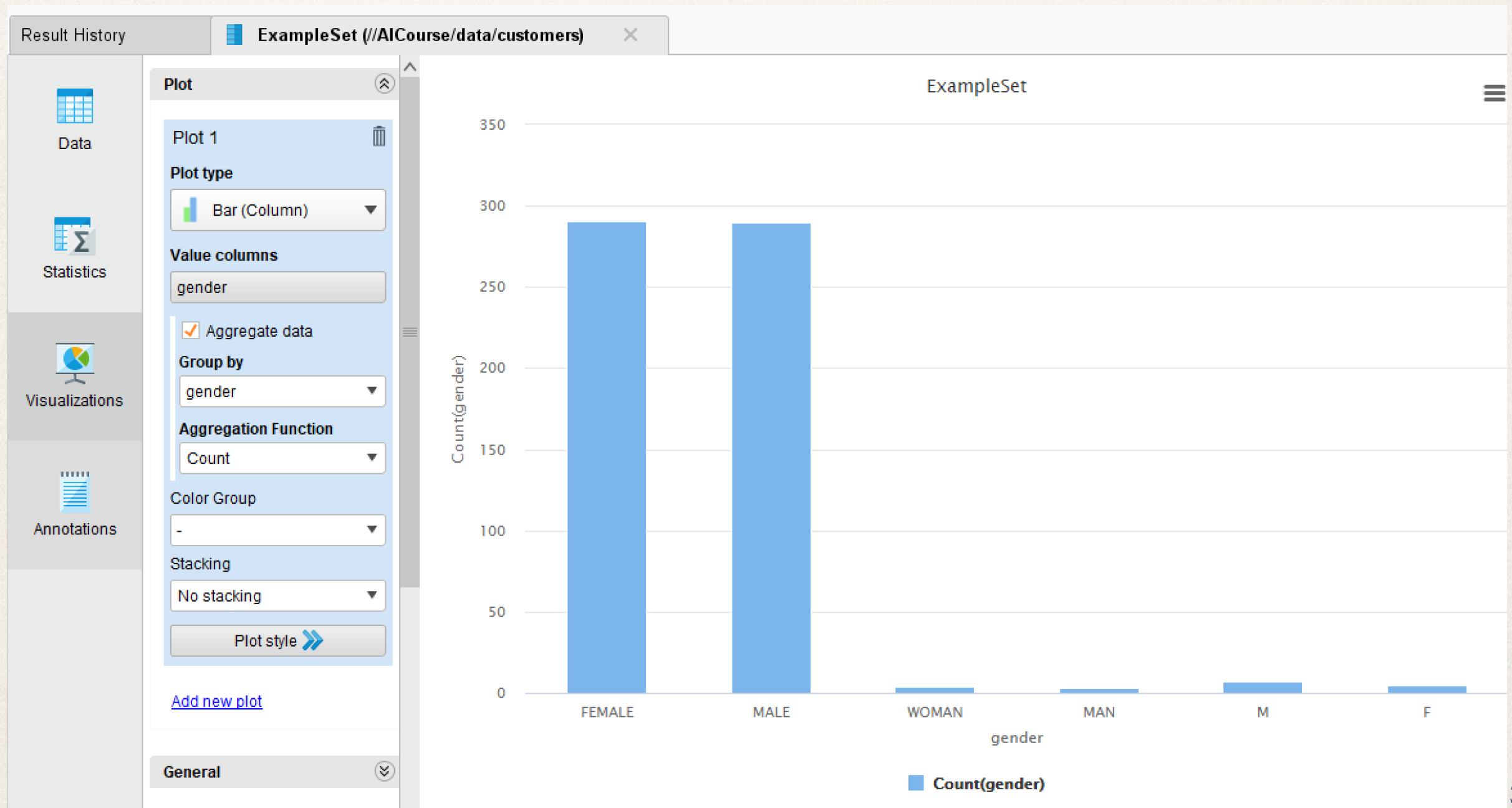
Data Visualization

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



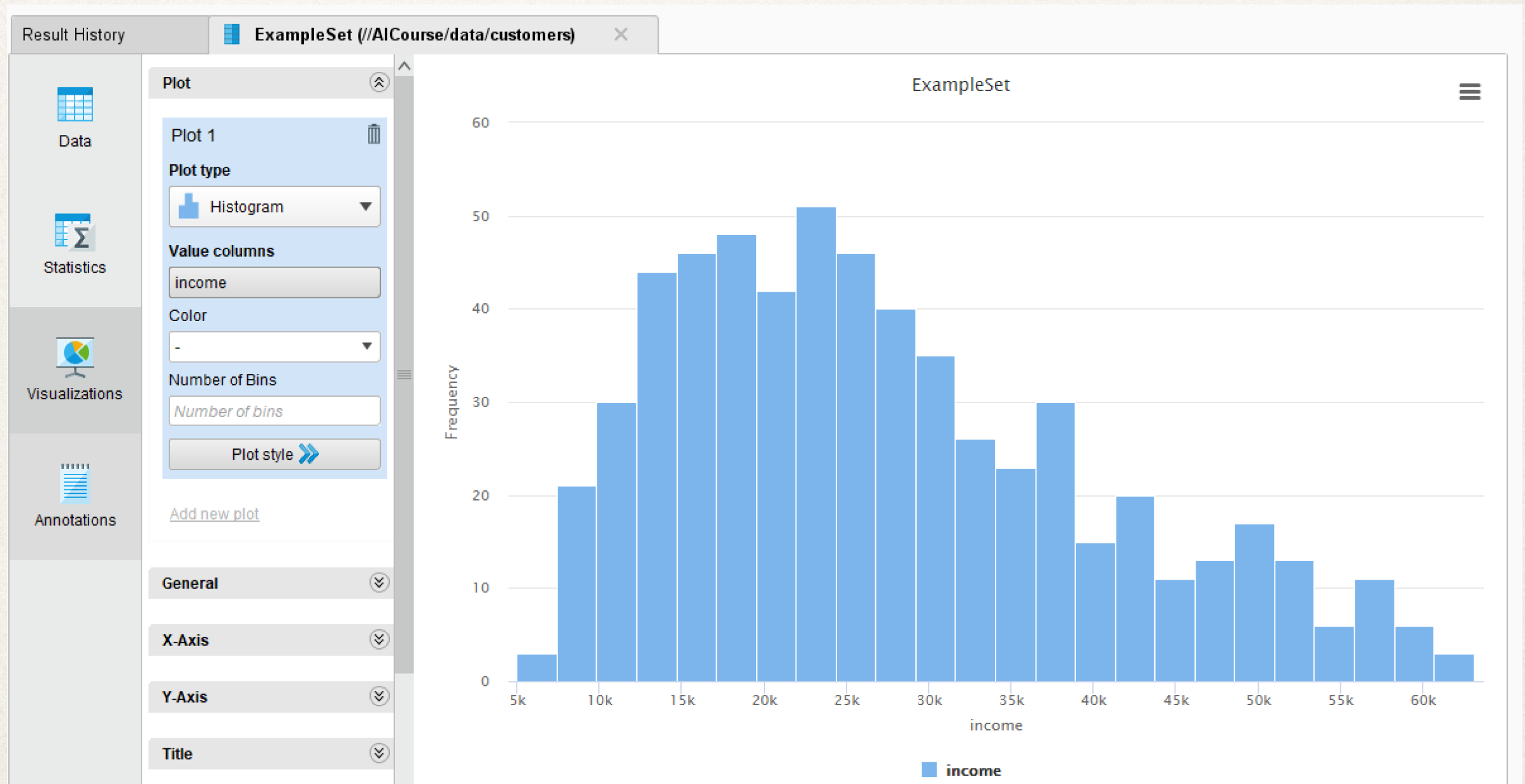
Data Visualization

❖ Bar chart of gender attribute



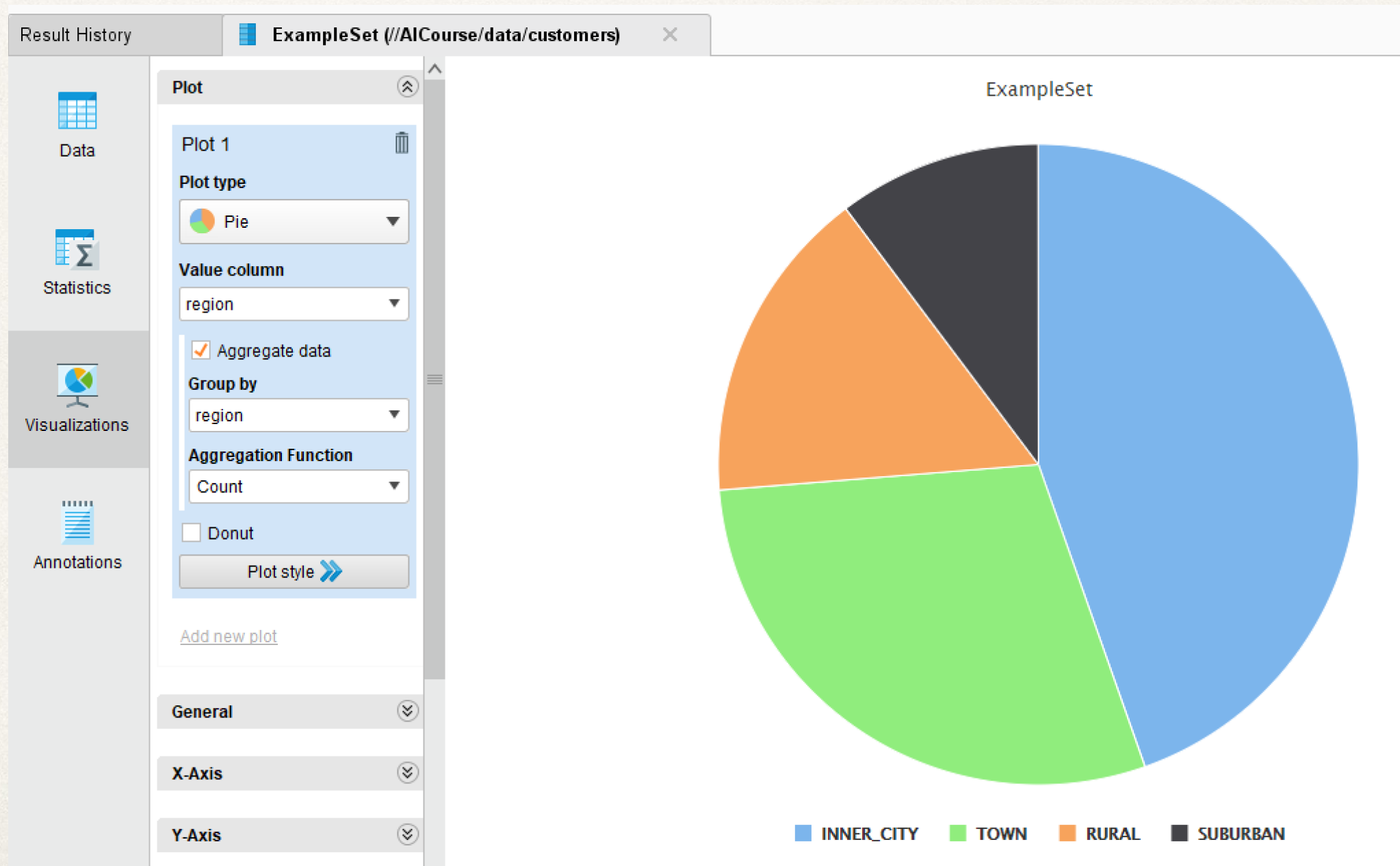
Data Visualization

❖ Histogram of **income** attribute



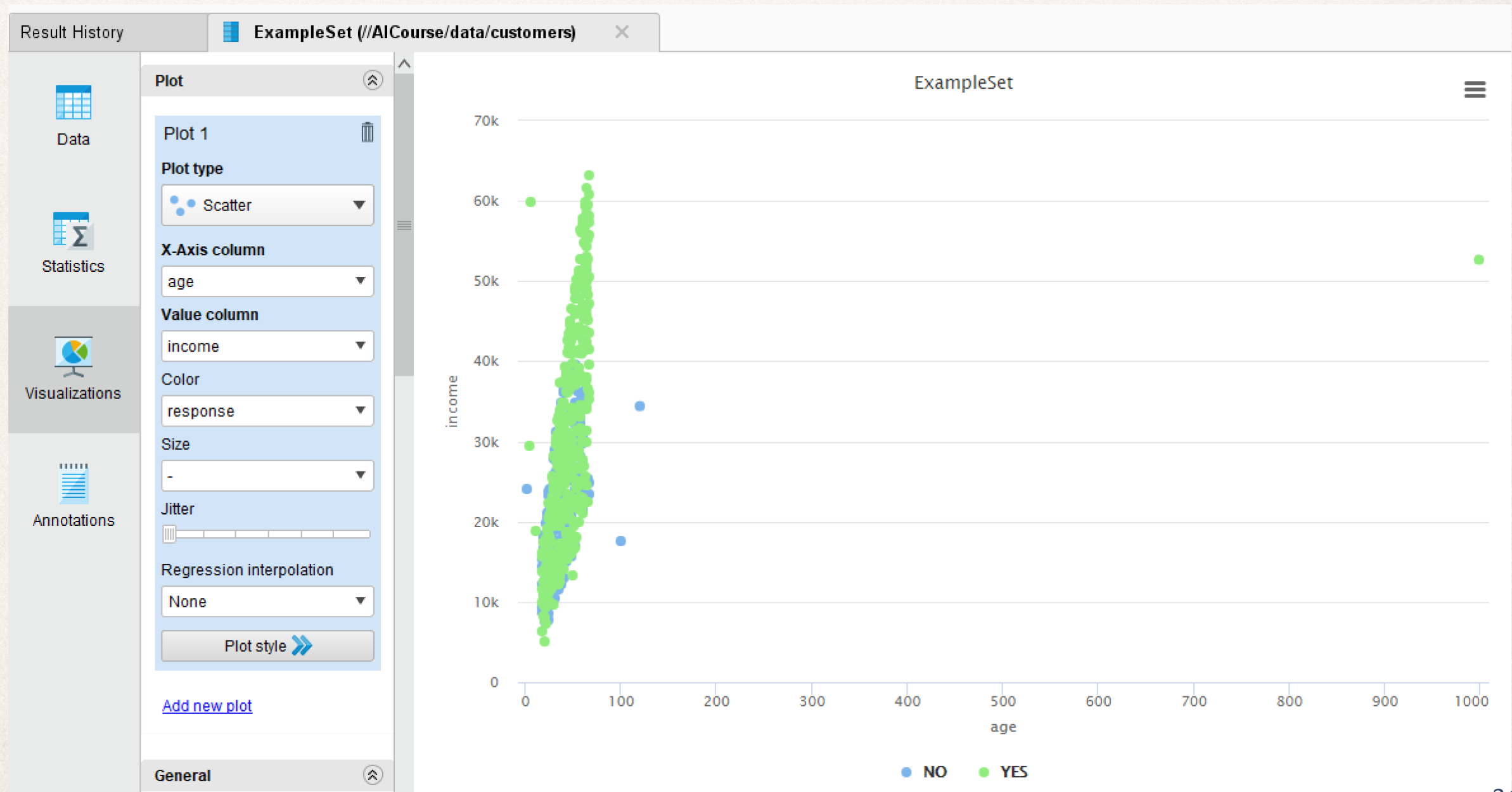
Data Visualization

❖ Pie chart of **region** attribute



Data Visualization

- ❖ Scatter plot of **age** and **income** attributes



Data Visualization

- ✿ Different types of chart for different proposes

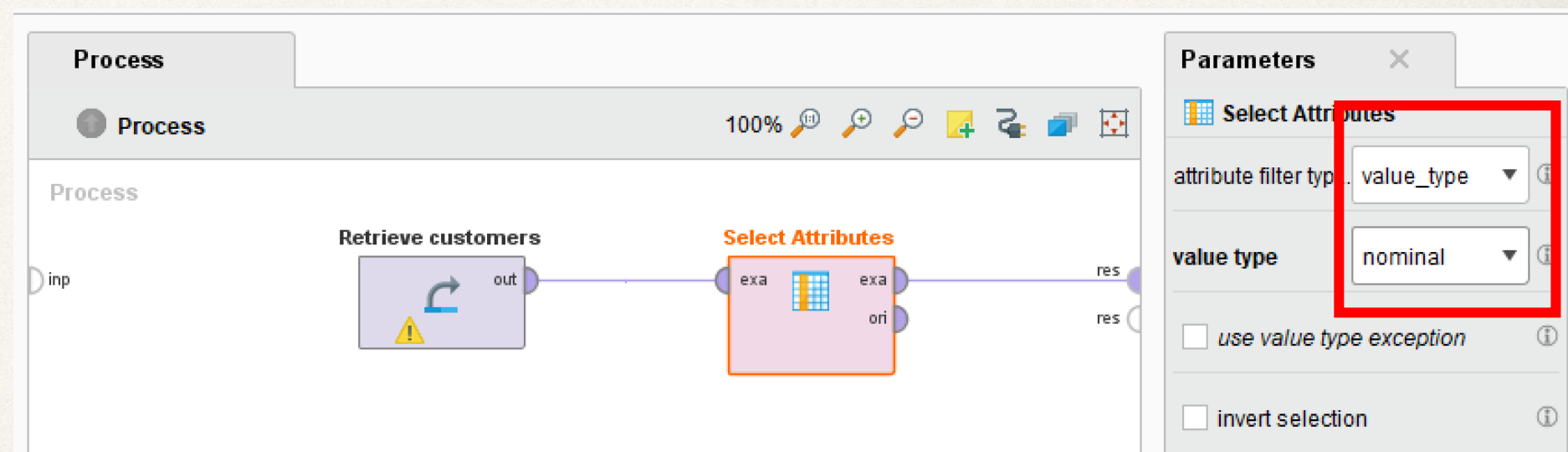
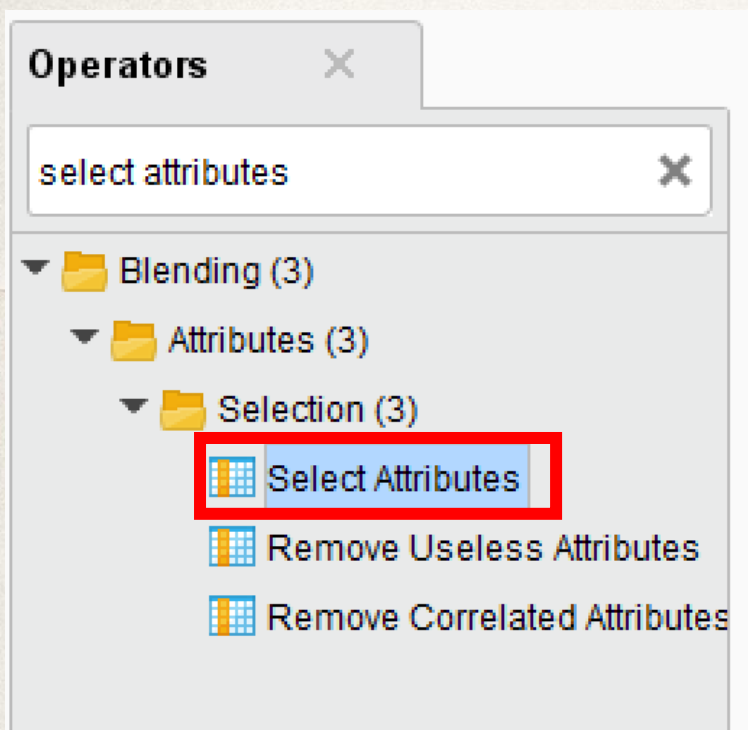


Data Preparation

- ❖ 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 × **ExampleSet (Select Attributes)** ×


Data


Statistics


Charts

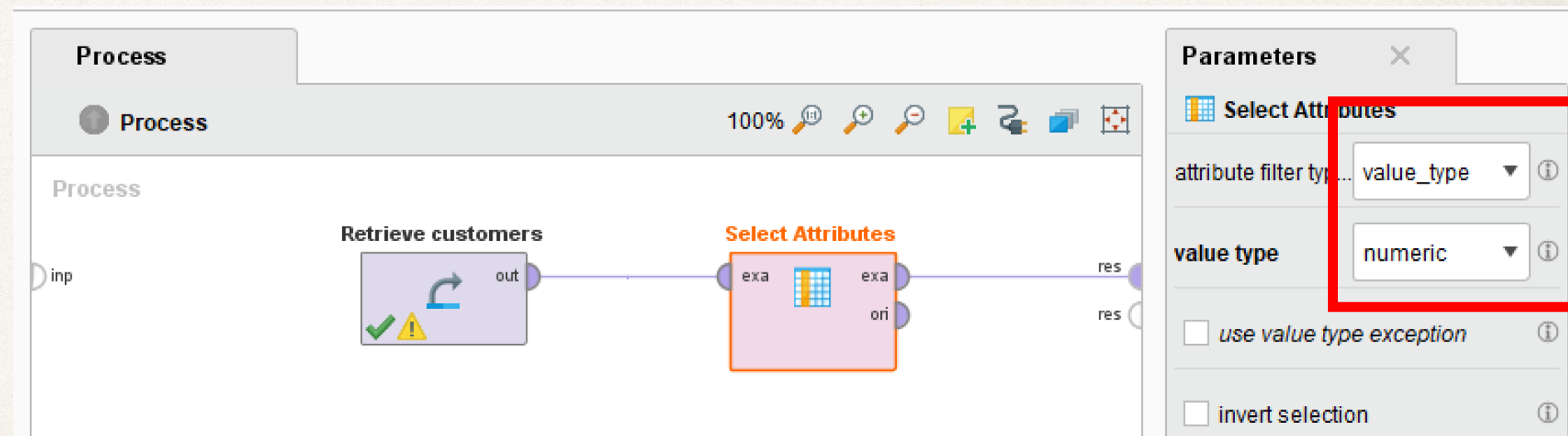
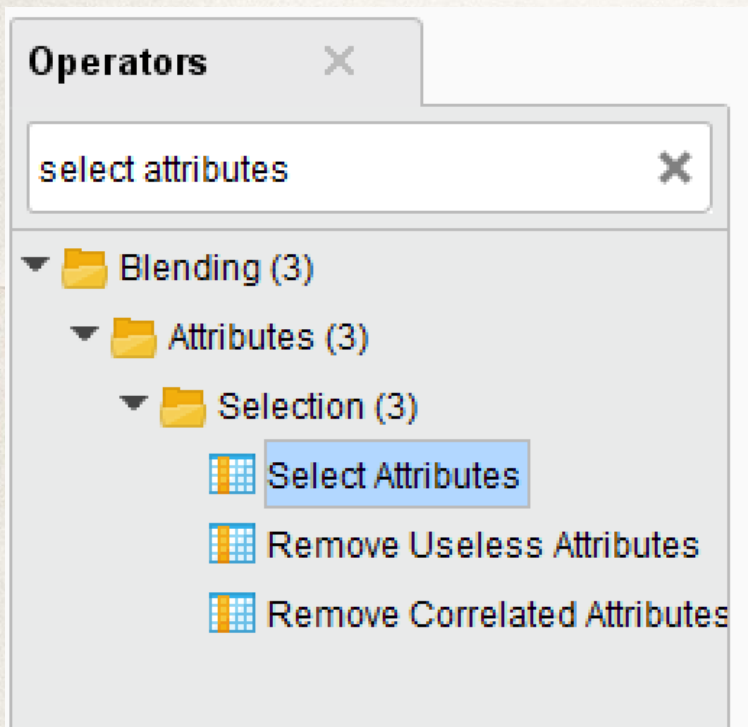


ExampleSet (600 examples, 2 special attributes, 4 regular attributes) Filter (600 / 600 examples): all ▼

| 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 |

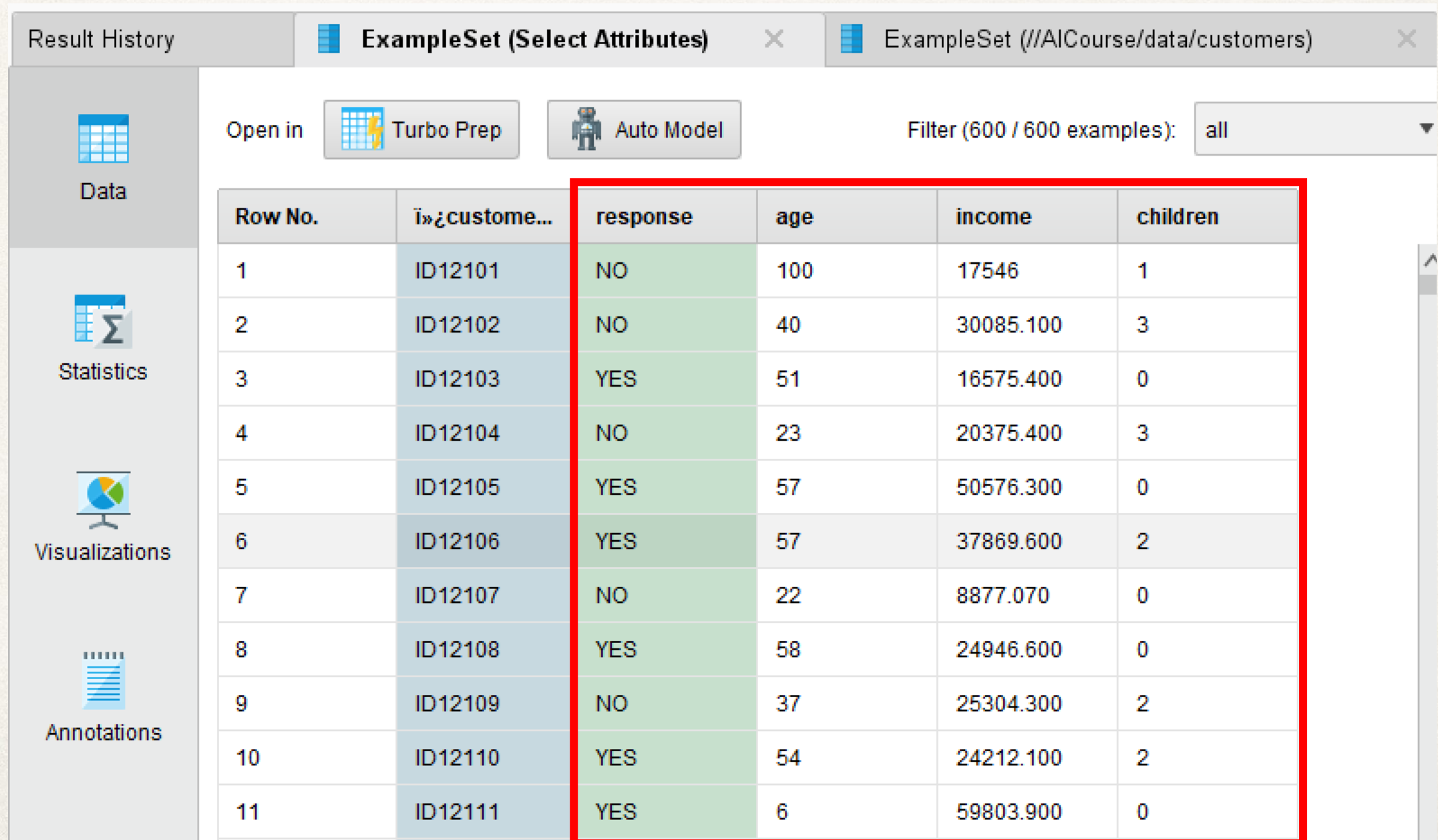
Select Attributes by Type

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



Select Attributes by Type

- ❖ The result shows only numeric attributes.

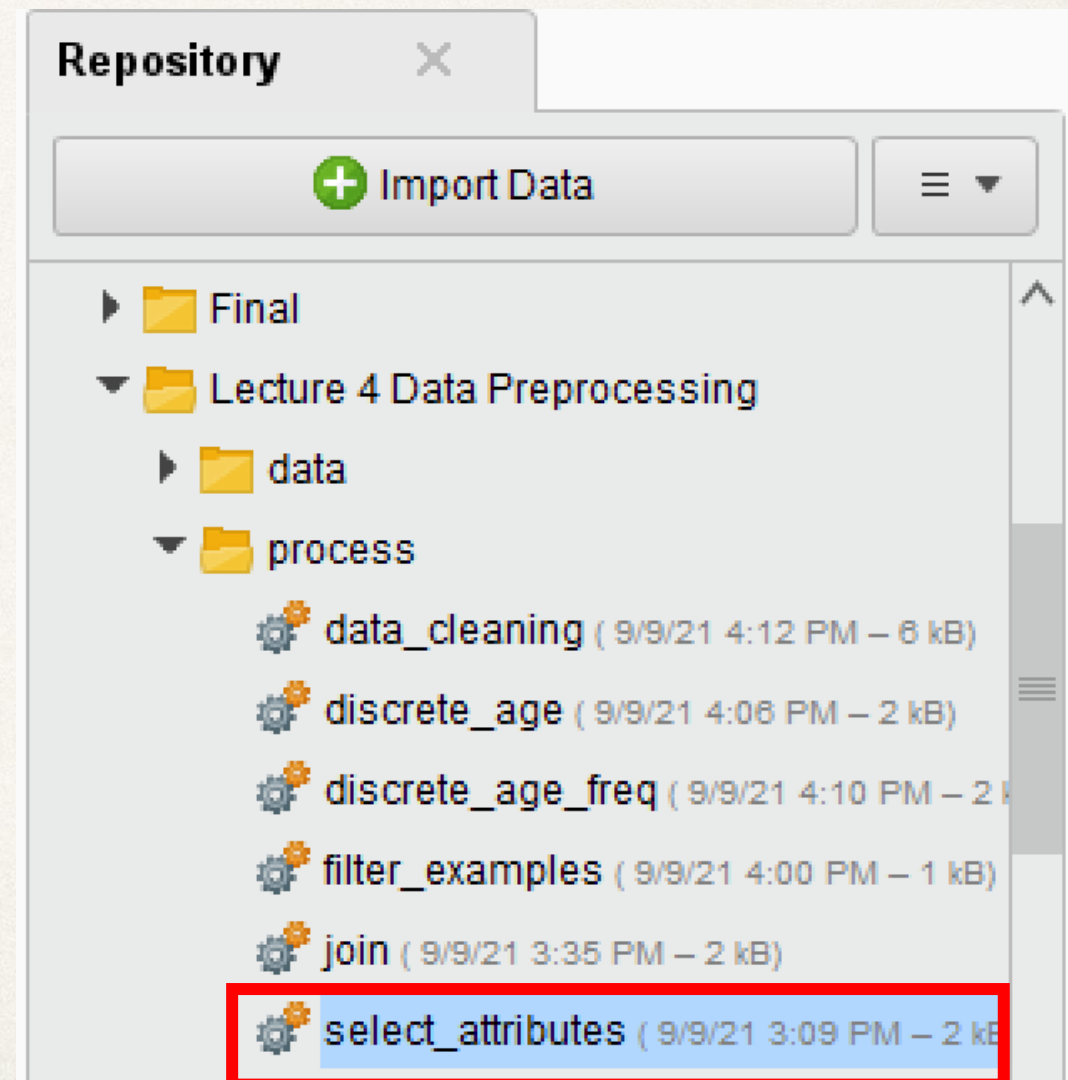
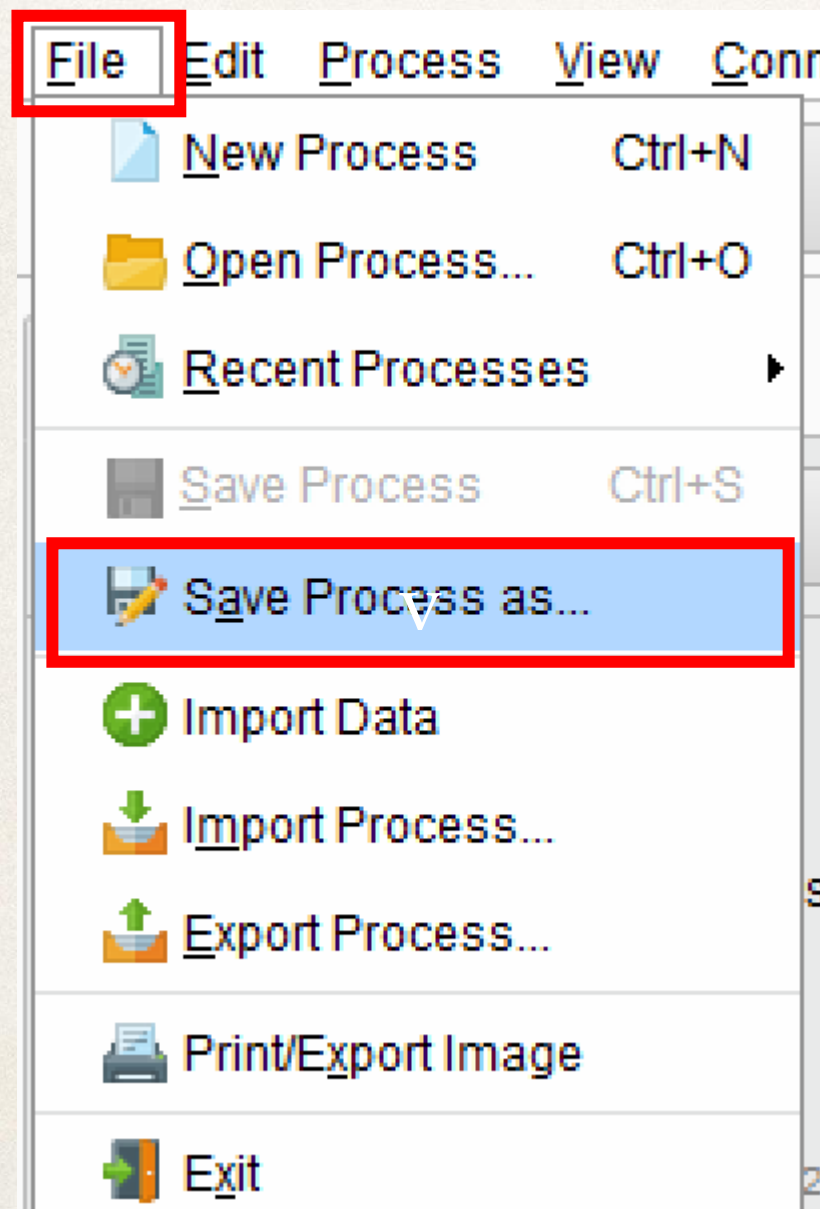


The screenshot shows a data analysis tool interface. At the top, there are tabs for 'Result History', 'ExampleSet (Select Attributes)', and 'ExampleSet (//AICourse/data/customers)'. Below the tabs, there are buttons for 'Open in Turbo Prep' and 'Auto Model', and a filter dropdown set to 'all'. On the left side, there is a sidebar with icons for 'Data', 'Statistics', 'Visualizations', and 'Annotations'. The main area displays a table with 11 rows of customer data. A red box highlights the columns 'response', 'age', 'income', and 'children'.

| Row No. | ID customer... | 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 |

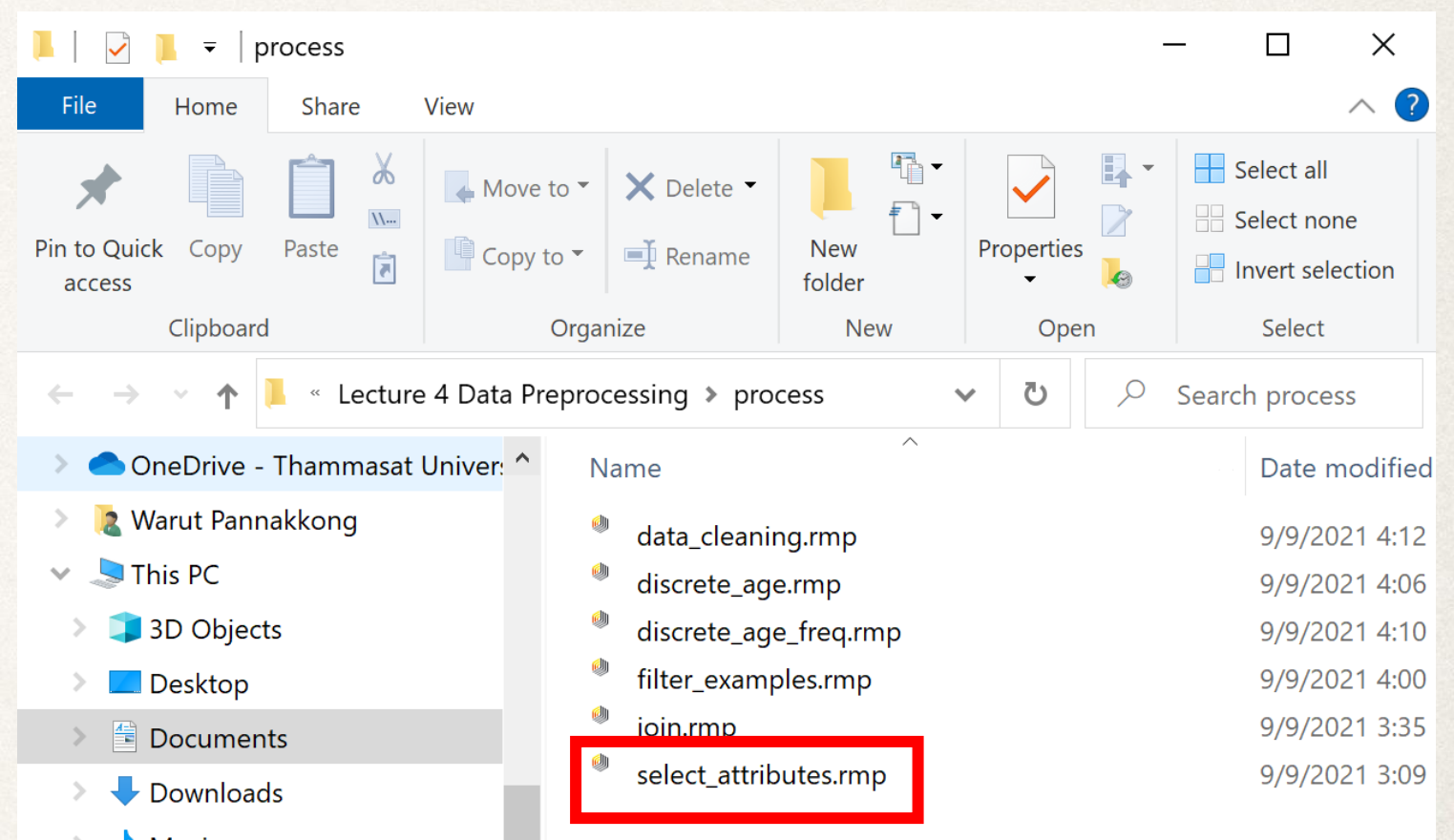
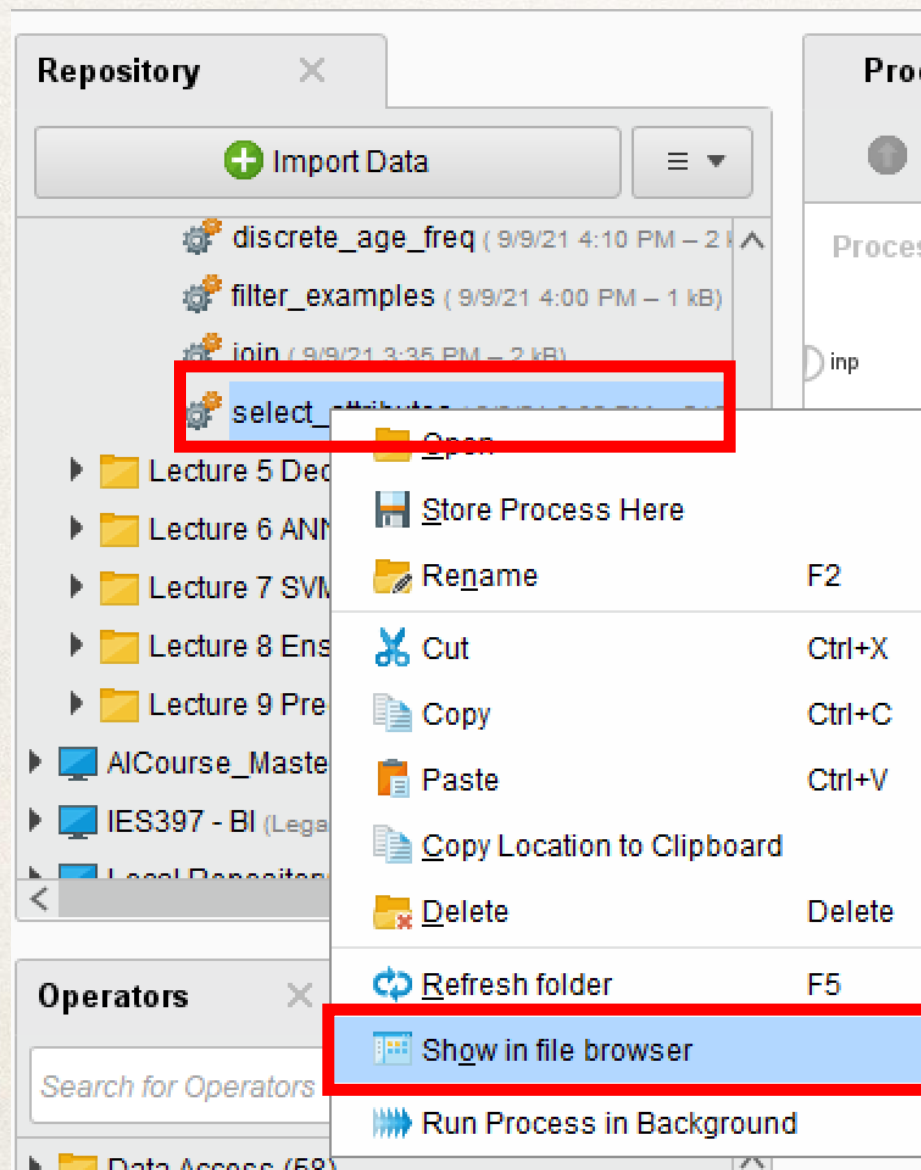
How to Save a Process into Repository

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



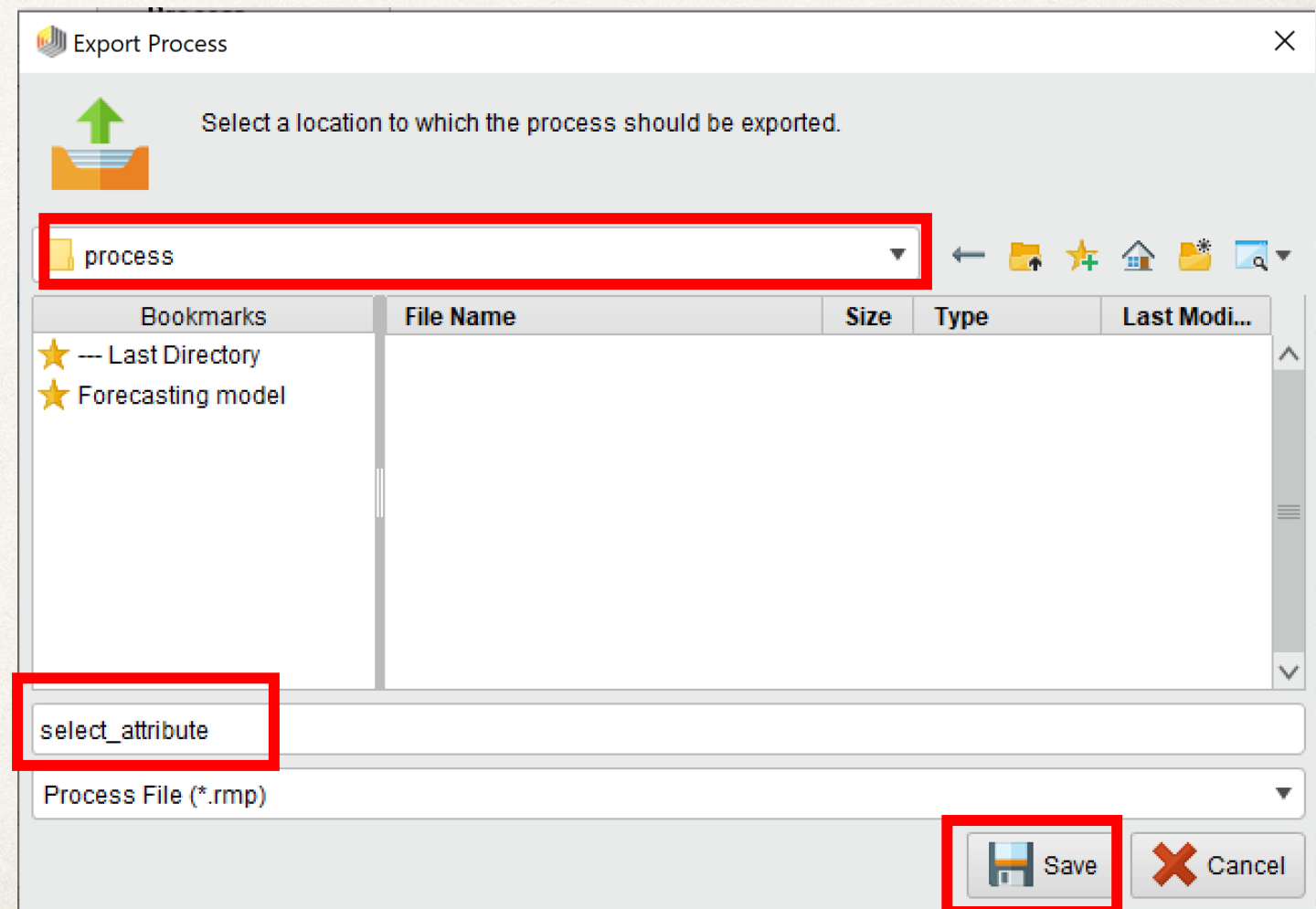
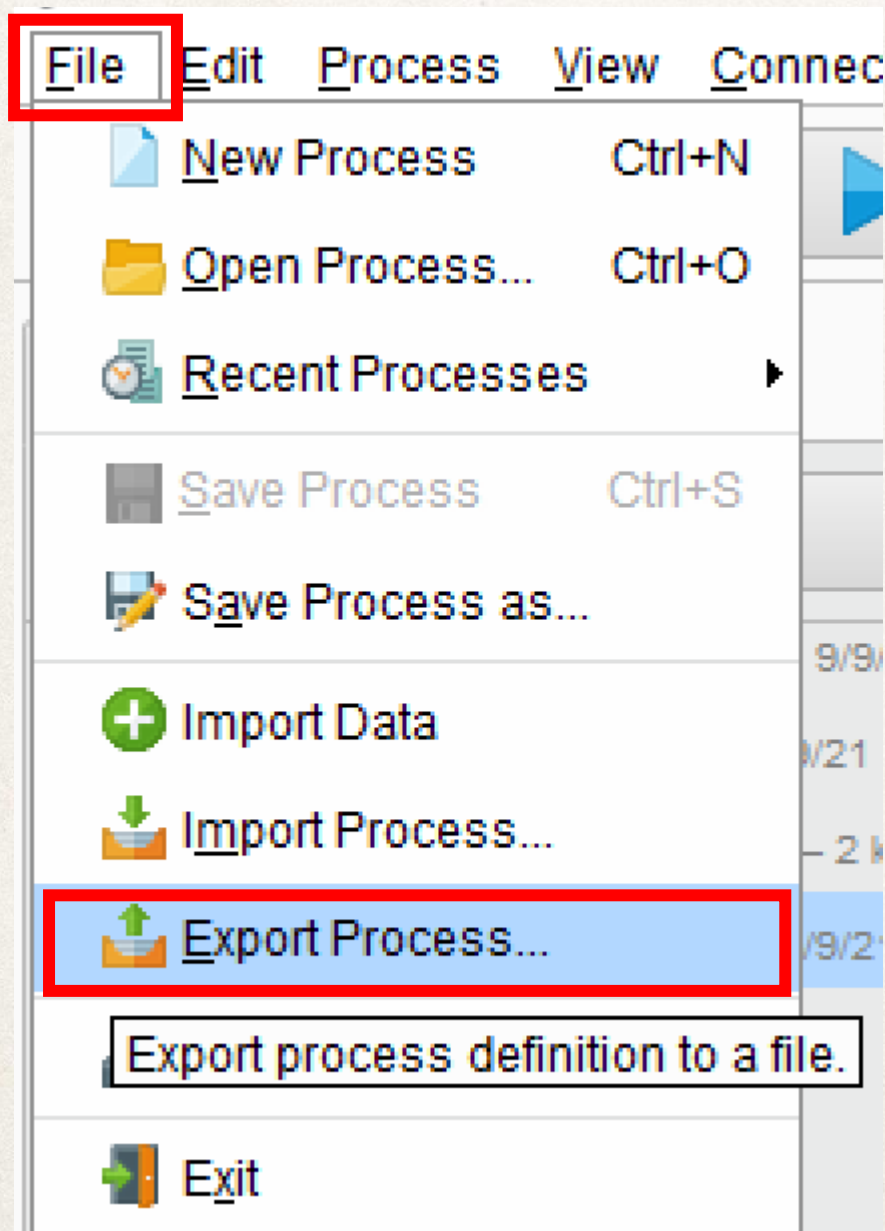
How to Show the Saved Process File (.rmp) in Repository

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



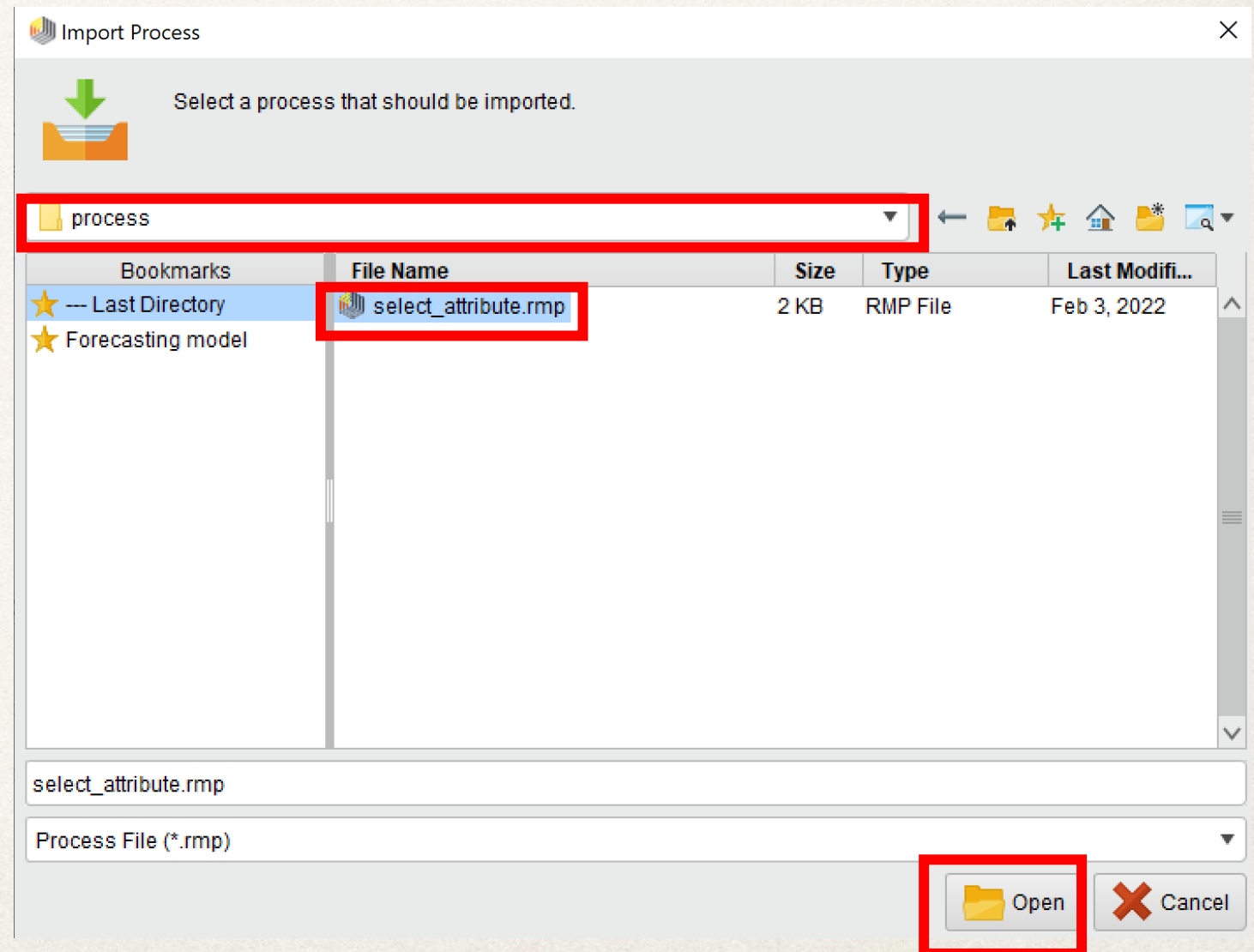
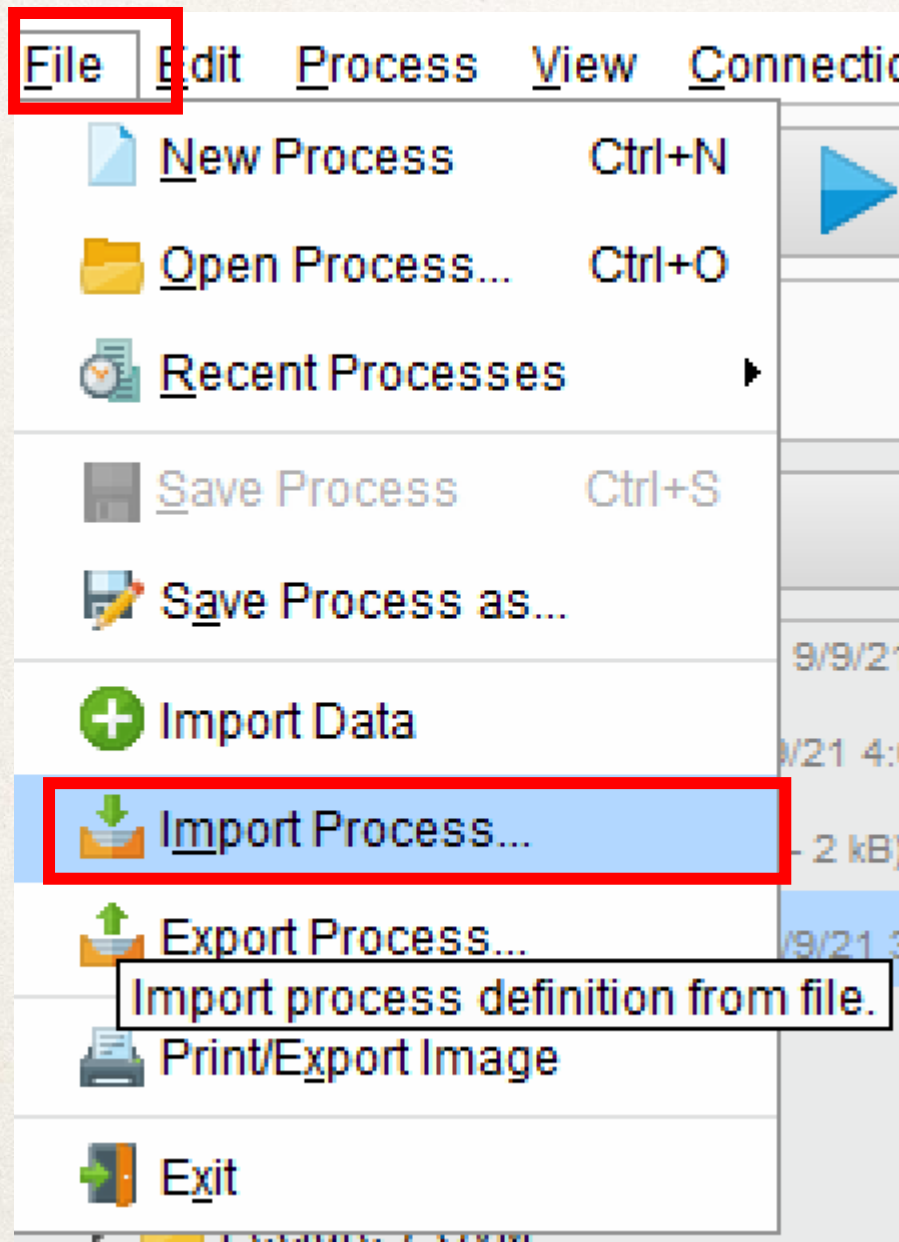
How to Export a Process File (.rmp)

❖ File > Export Process



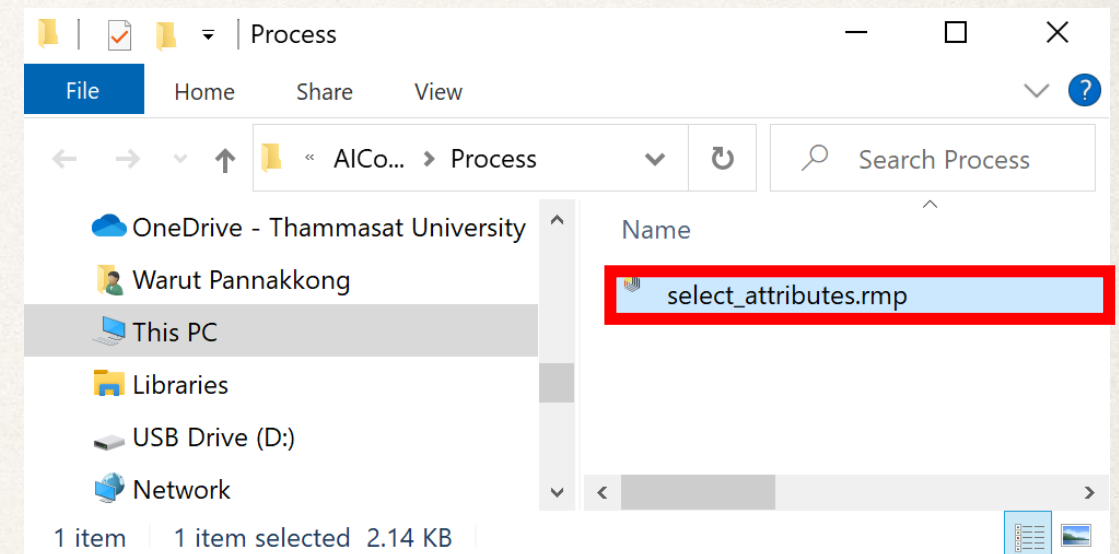
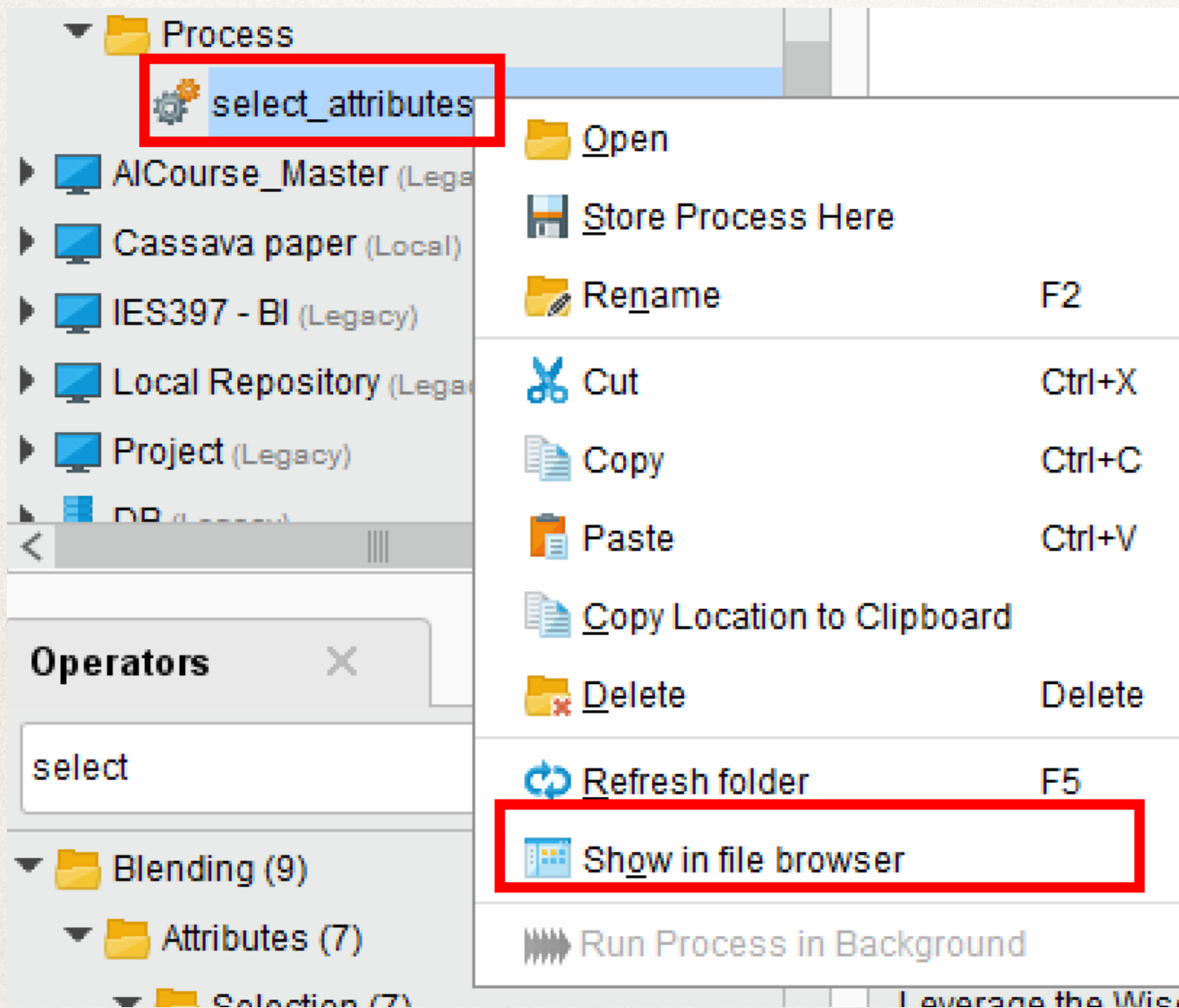
How to Import a Process File (.rmp)

❖ 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**

The screenshot displays the Alteryx interface with three main panels: Operators, Process, and Parameters.

- Operators Panel:** A search bar contains "select attributes". The results are categorized under "Blending (3)", "Attributes (3)", and "Selection (3)". The "Select Attributes" operator is highlighted.
- Process Panel:** A workflow diagram showing a "Retrieve customers" operator connected to a "Select Attributes" operator. The "Select Attributes" operator has a yellow warning icon.
- Parameters Panel:** The "Select Attributes" operator's configuration is shown. The "attribute filter type" is set to "subset". The "attributes" field is set to "Select Attribute...". The "invert selection" and "include special attributes" checkboxes are unchecked.

Select Attributes by Name

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

Select Attributes: attributes

Select Attributes: **attributes**
The attribute which should be chosen.

Attributes

Search

- car
- # children
- gender
- married
- region
- response
- i»¿customer_id

Selected Attributes

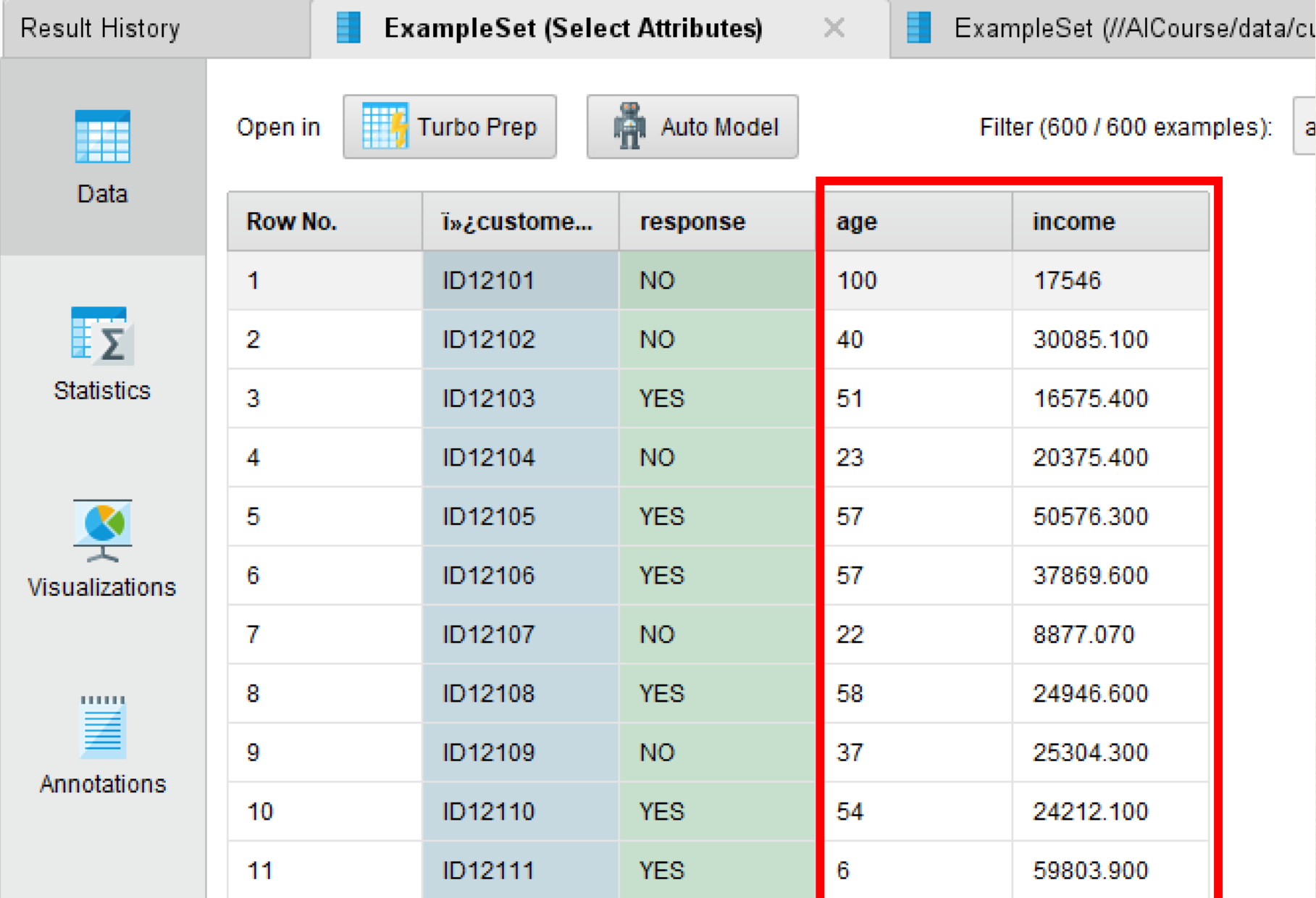
Search

- # age
- # income

Apply Cancel

Select Attributes by Name

- ❖ The result shows only the selected attributes.



The screenshot shows a software interface with a sidebar on the left containing icons for Data, Statistics, Visualizations, and Annotations. The main area displays a table titled 'ExampleSet (Select Attributes)'. Above the table are buttons for 'Open in Turbo Prep' and 'Auto Model', and a filter status 'Filter (600 / 600 examples):'. The table has five columns: 'Row No.', 'customer...', 'response', 'age', and 'income'. The last two columns, 'age' and 'income', are highlighted with a red border. The table contains 11 rows of data.

| Row No. | customer... | 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 |

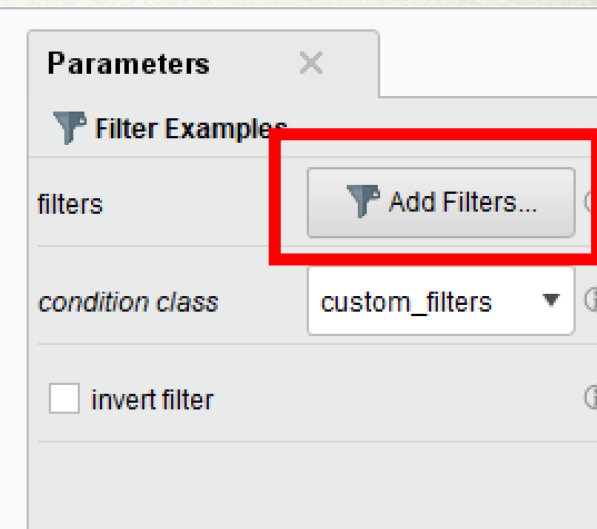
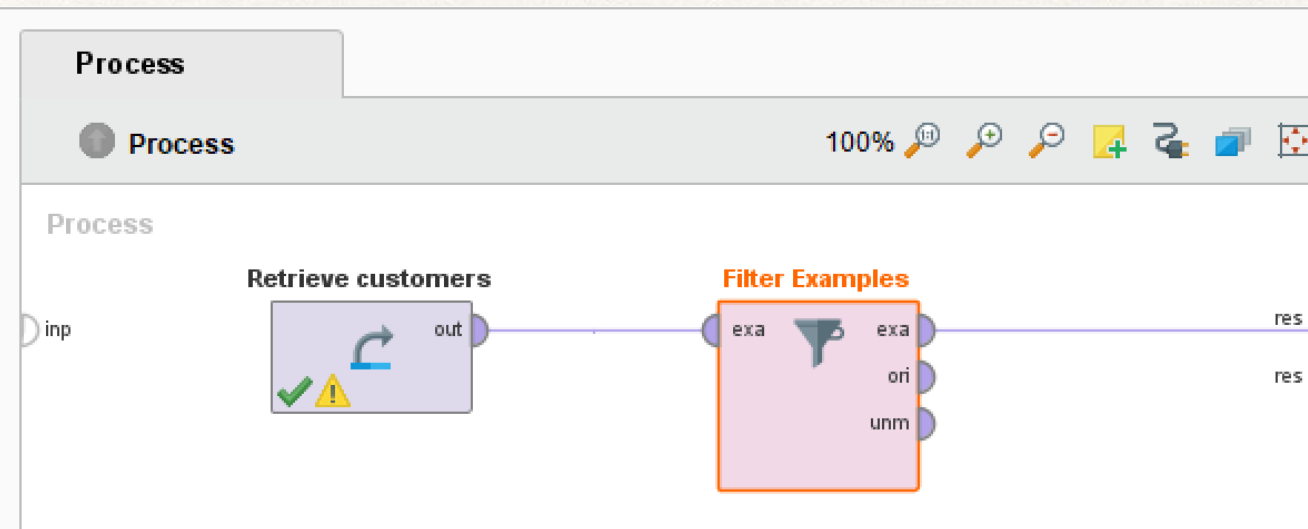
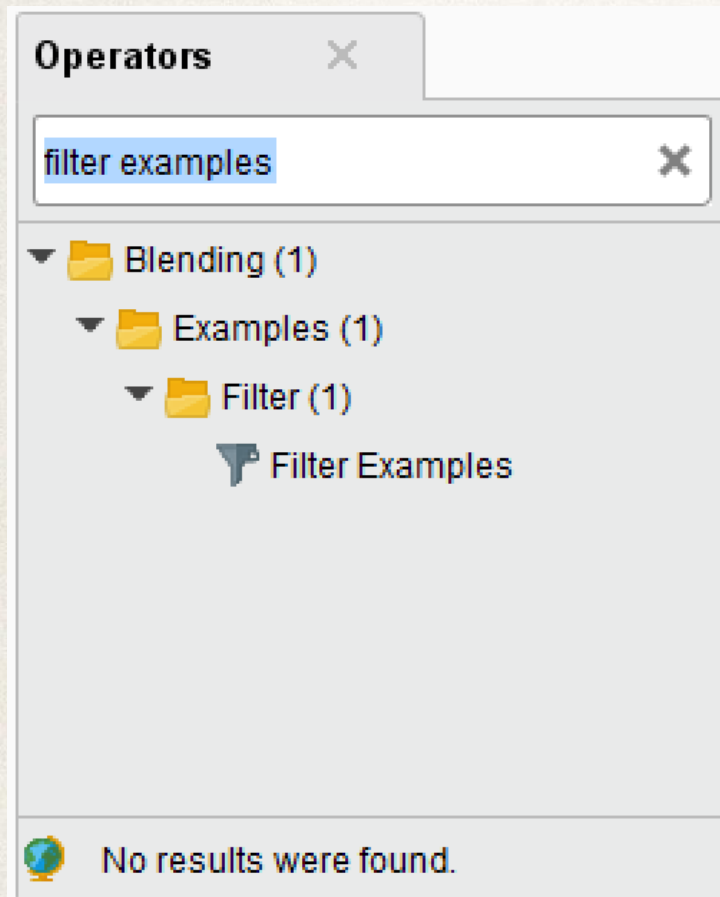
Data Preparation

- ❖ 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

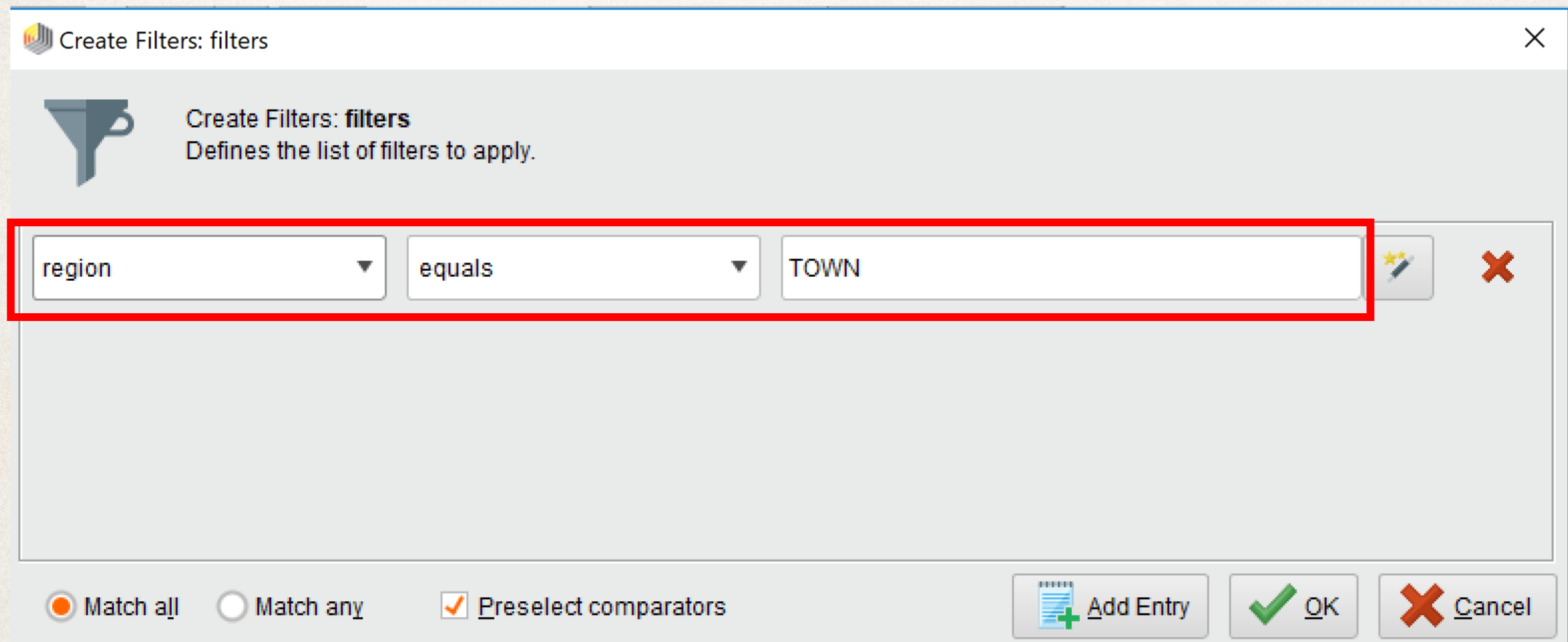
❖ Choose Filter Examples operator

❖ Press



Filter Examples by Conditions

- ❖ Set filter condition: Region equals town.



The screenshot shows a dialog box titled "Create Filters: filters" with a close button (X) in the top right corner. Below the title bar, there is a funnel icon and the text "Create Filters: filters" and "Defines the list of filters to apply." Below this, there is a filter entry row highlighted with a red border. The row contains three fields: a dropdown menu with "region" selected, a dropdown menu with "equals" selected, and a text input field containing "TOWN". To the right of the text input field are two icons: a yellow star and a red X. At the bottom of the dialog box, there are three radio buttons: "Match all" (selected), "Match any", and "Preselect comparators" (checked). To the right of these radio buttons are three buttons: "Add Entry" (with a plus icon), "OK" (with a green checkmark icon), and "Cancel" (with a red X icon).

Filter Examples by Conditions

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

Result History

ExampleSet (Filter Examples)

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`

Data import wizard - Step 4 of 4

This wizard guides you to import your data.
Step 4: RapidMiner Studio uses strongly typed attributes. In this step, you can define the data types of your attributes. Furthermore, RapidMiner Studio assigns roles to the attributes, defining what they can be used for by the individual operators. These roles can be also defined here. Finally, you can rename attributes or deselect them entirely.

Date format:

☒ Preview uses only first 100 rows.

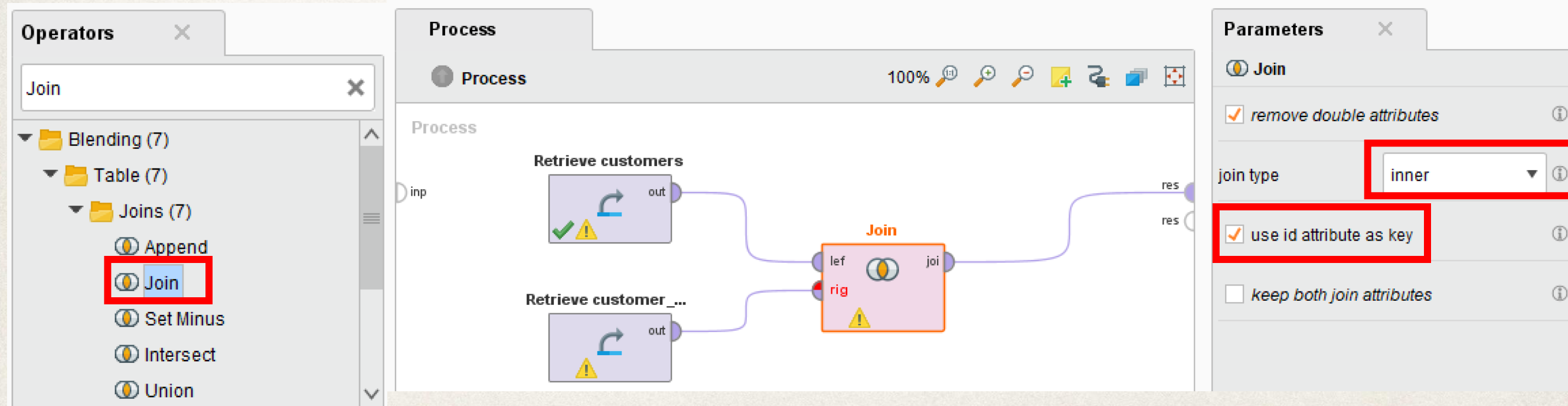
| <input checked="" type="checkbox"/> | <input checked="" type="checkbox"/> | | |
|-------------------------------------|-------------------------------------|-------------|------------|
| | | customer_id | email |
| | | polyno... | polyno... |
| | | id | attribute |
| | | ID12101 | cugx@g... |
| | | ID12102 | kxbc@g... |
| | | ID12103 | nxms@g... |
| | | ID12104 | gezk@g... |
| | | ID12105 | tipx@gm... |

☒ 0 errors. ☒ Ignore errors ☐ Show only errors

| Row, Column | Error | Original value | Message |
|-------------|-------|----------------|---------|
|-------------|-------|----------------|---------|

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)

Data

Statistics

Visualizations

Annotations

Open in

Turbo Prep

Auto Model

Filter (600 / 600 examples): all

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

Data Preparation

- ❖ Preprocessing

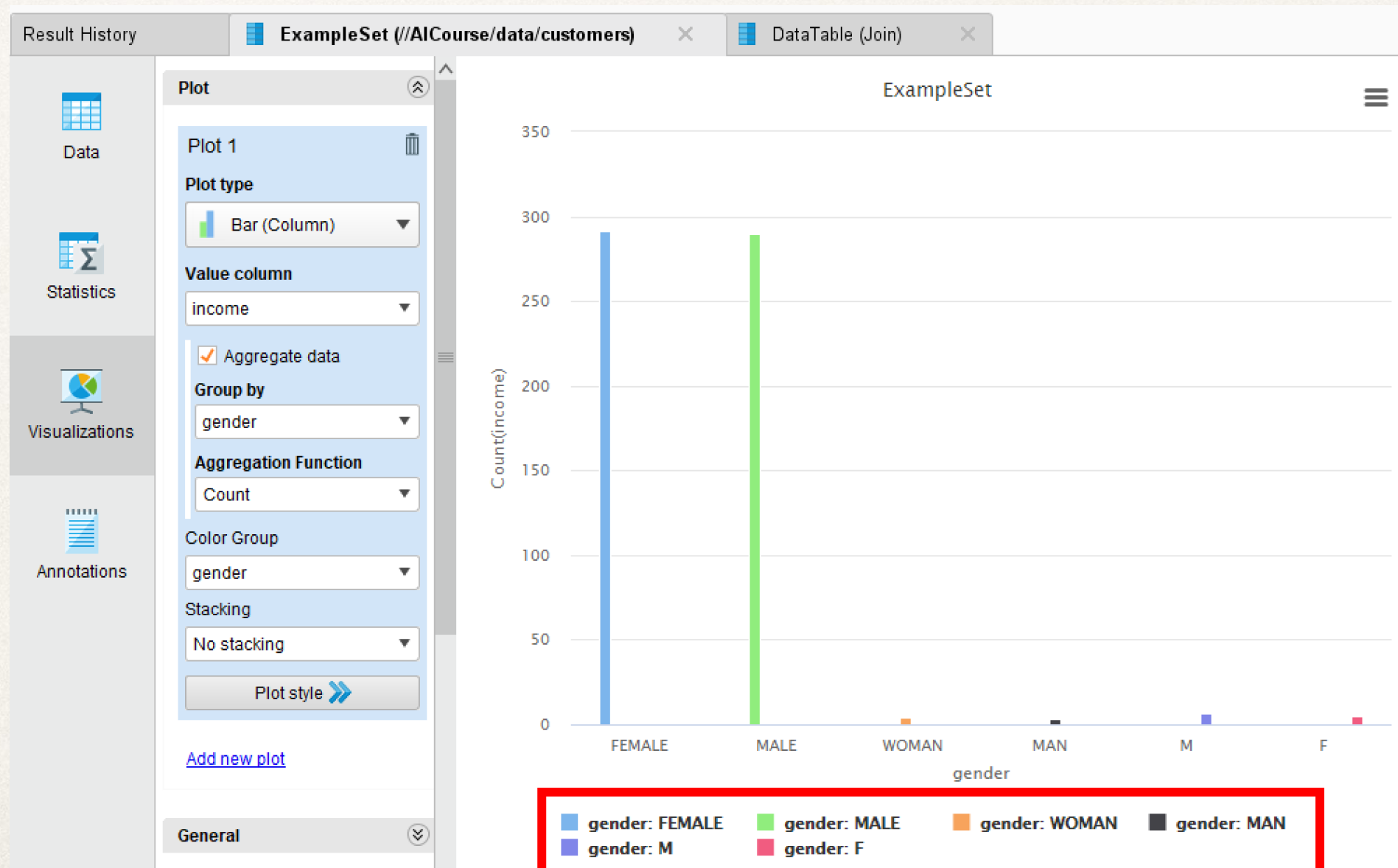
- ❖ Select attributes
 - ❖ Select by type of attributes
 - ❖ Select by specific attributes
- ❖ Filter examples by conditions
- ❖ Join data from multiple

sources

- ❖ **Deal with incomplete data**
 - ❖ **Inconsistent data**
 - ❖ Missing data
- ❖ Data transformation
 - ❖ Discretization (numeric to nominal)
 - ❖ User defined
 - ❖ Equal frequency

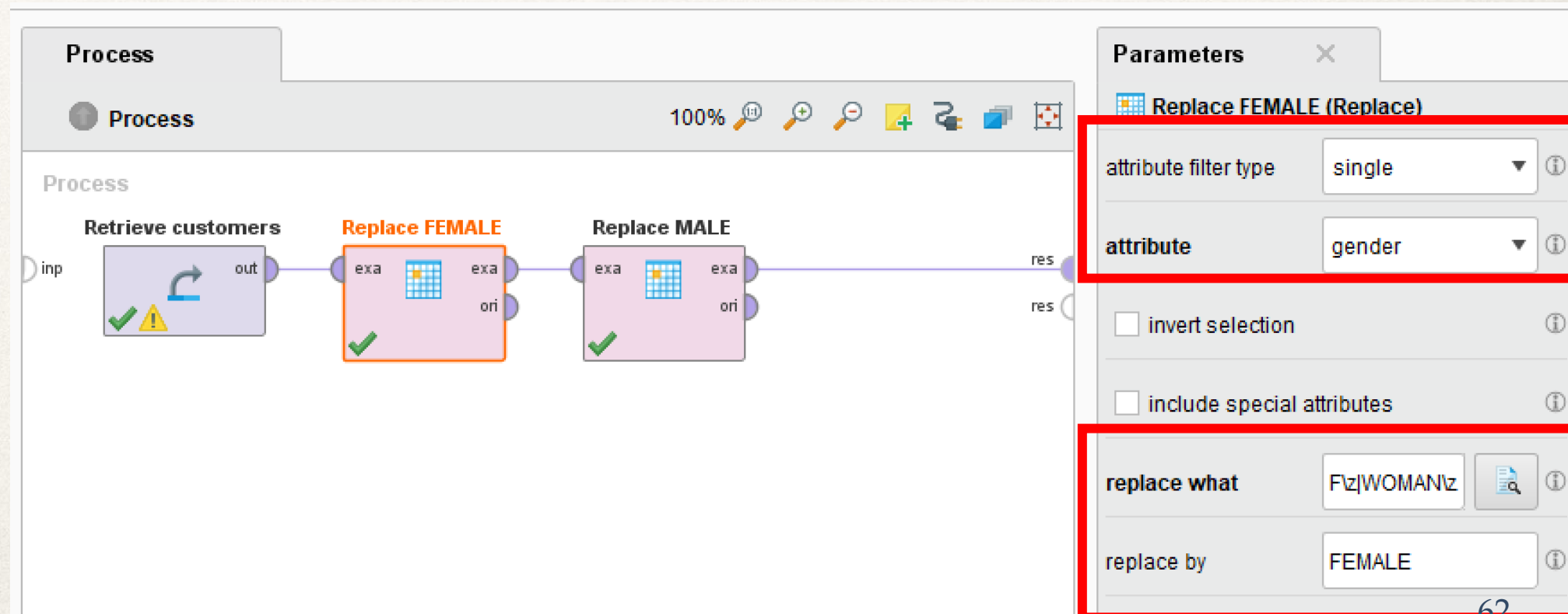
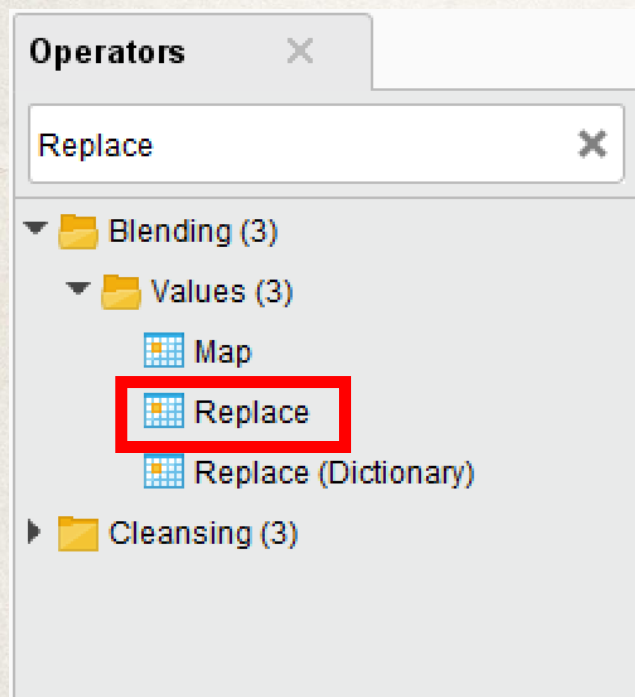
Replace Inconsistent Data

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



Replace Inconsistent Data

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



Replace Inconsistent Data

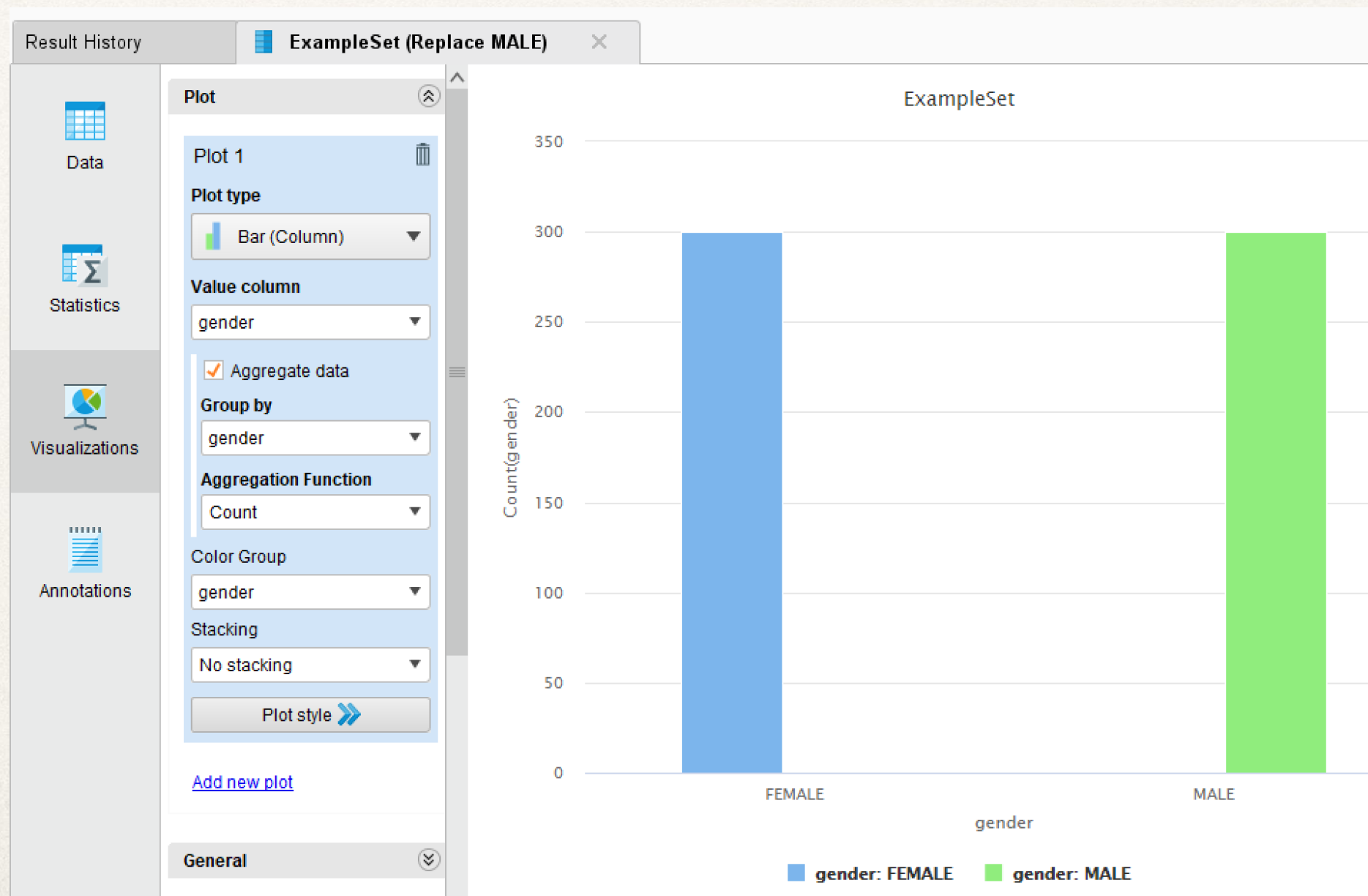
- ❖ Choose **Replace** operator
- ❖ replace what = **M\z|MAN\z**
- ❖ replace by = **MALE**

The screenshot displays the Alteryx Designer interface with three main panels: Operators, Process, and Parameters.

- Operators Panel:** Located on the left, it shows a search bar with "Replace" entered. Under the "Values (3)" folder, the "Replace" operator is highlighted with a red box.
- Process Panel:** The central workspace shows a workflow. It starts with a "Retrieve customers" operator, followed by a "Replace FEMALE" operator, and then a "Replace MALE" operator. The "Replace MALE" operator is highlighted with a red box. The workflow ends with a "res" output.
- Parameters Panel:** Located on the right, it shows the configuration for the "Replace MALE (Replace)" operator. The "attribute filter type" is set to "single" and the "attribute" is set to "gender". The "replace what" field contains the regex "M\z|MAN\z" and the "replace by" field contains "MALE". Both the parameter configuration area and the "replace what" field are highlighted with red boxes.

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 × **ExampleSet (//AICourse/data/customers)** ×


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

- ❖ Choose **Replace Missing Values** operator
- ❖ attribute filter type = **single** attribute = **region**
- ❖ default = **value** replenishment = **N/A**

The screenshot displays the Alteryx workflow editor with the 'Replace Missing Values' operator configured. The 'Operators' pane on the left shows the 'Replace Missing Values' operator selected under the 'Missing (3)' category. The 'Process' pane in the center shows a workflow starting with 'Retrieve customers', followed by 'Replace FEMALE', 'Replace MALE', 'Replace Missing Values Region' (highlighted in orange), and 'Replace Missing Values'. The 'Parameters' pane on the right shows the configuration for the 'Replace Missing Values Region (Repla...' operator:

- attribute filter type**: single
- attribute**: region
- default**: value
- replenishment value**: N/A

Replace Missing Data

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

The screenshot displays the Alteryx workflow editor with the 'Replace Missing Values' operator configured. The 'Operators' pane on the left shows the 'Replace Missing Values' operator selected under the 'Missing' category. The 'Process' pane in the center shows a workflow starting with 'Retrieve customers', followed by 'Replace FEMALE', 'Replace MALE', 'Replace Missing V', and finally 'Replace Missing Values Children'. The 'Parameters' pane on the right shows the configuration for the 'Replace Missing Values Children' operator, with the following settings highlighted by red boxes:

- attribute filter type**: single
- attribute**: children
- default**: zero

The 'columns' section at the bottom of the parameters pane shows an 'Edit List (0)...' button.

Replace Missing Data

- ❖ No missing data for **region** and **children** attributes

Result History

ExampleSet (Replace Missing Values Children)

ExampleSet (//AICourse/data/customers)

Data

Statistics

Visualizations

Annotations

Name

Type

Missing

Statisti...

Filter (9 / 9 attributes):

Search for Attributes

Id

customer_id

Polynomial

0

Least ID12700 (1)

Most ID12101 (1)

Value ID12

Label

response

Polynomial

0

Least NO (186)

Most YES (414)

Value YES

children

Integer

0

Min 0

Max 3

Avera 1.00

region

Polynomial

0

Least N/A (6)

Most INNER_CITY (265)

Value INNE

gender

Polynomial

0

Least MALE (300)

Most FEMALE (300)

Value FEM

age

Integer

0

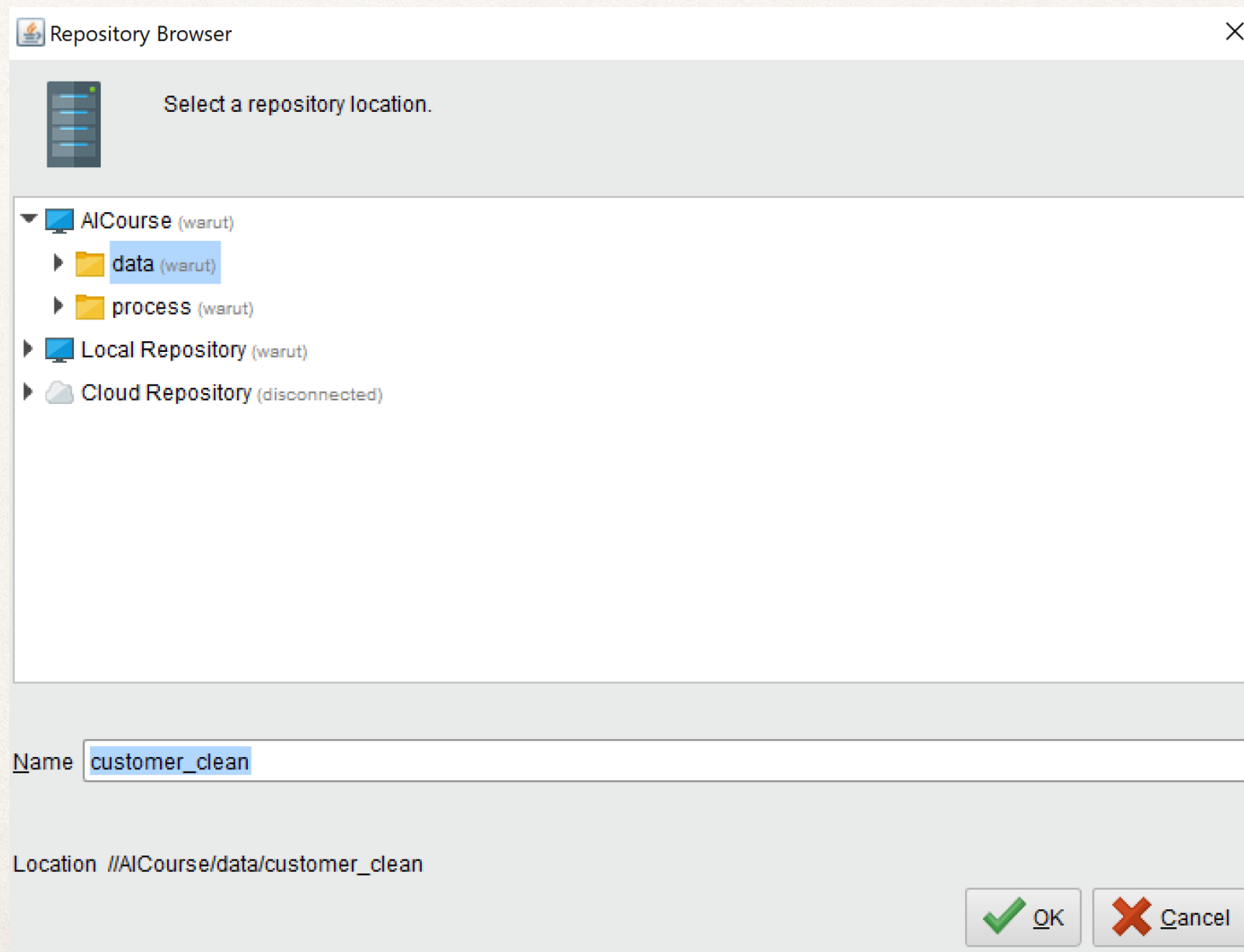
Min 2

Max 999

Avera 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 |

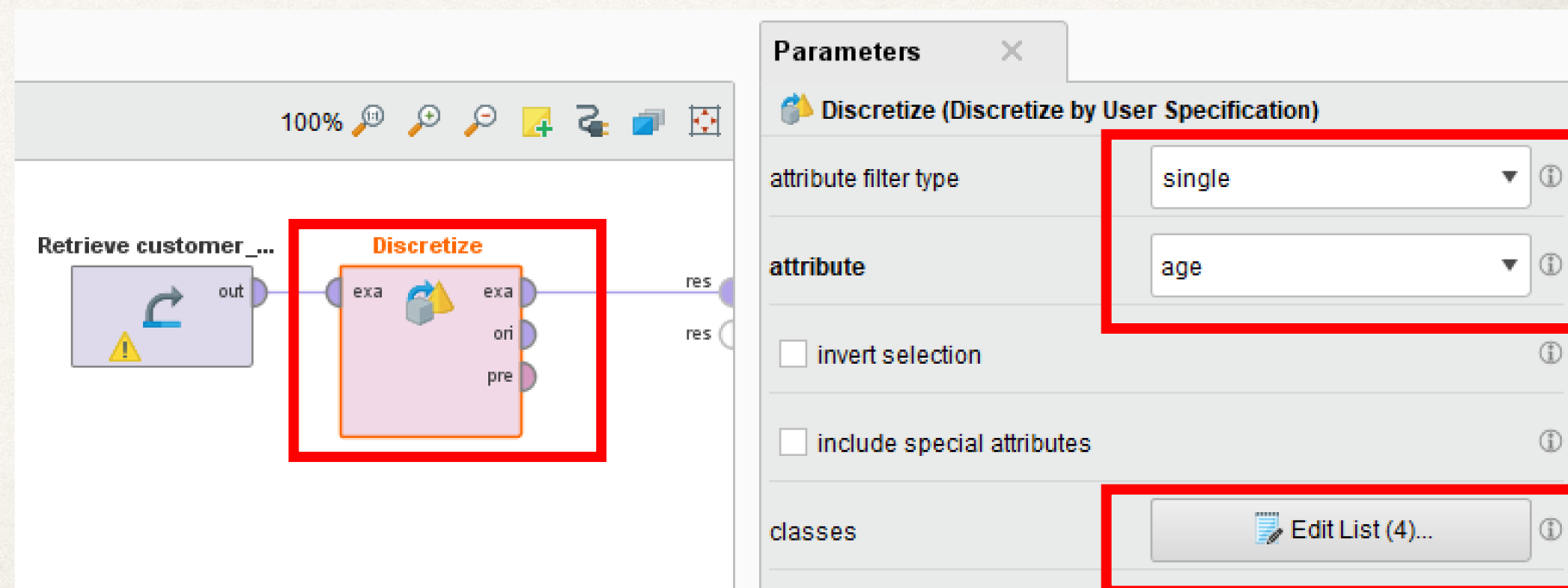
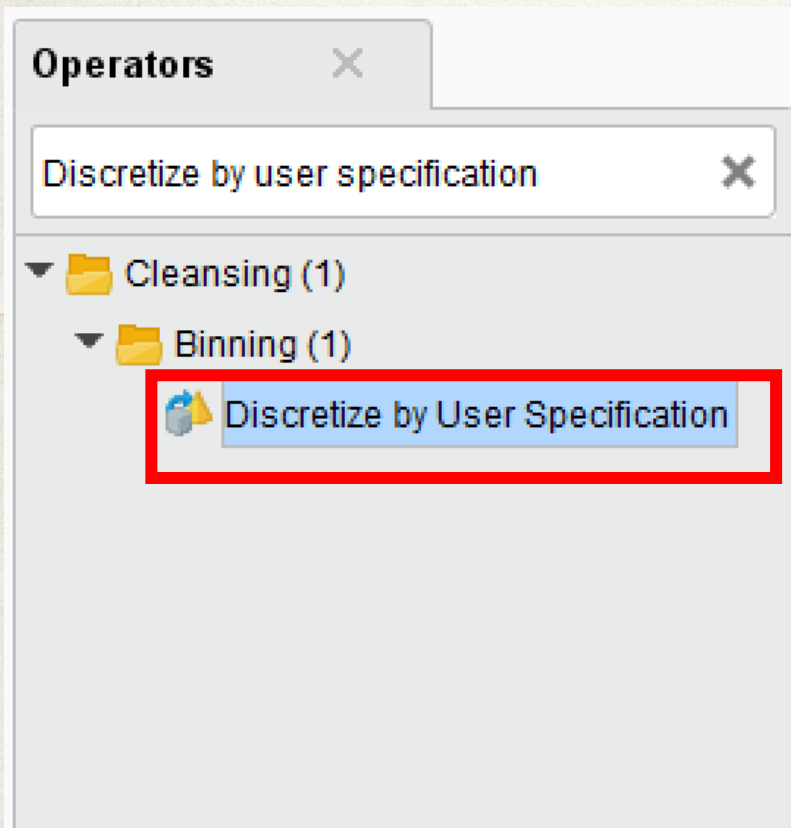
Conditions:

- $18 \leq$ 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





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

ExampleSet (Discretize)

Data

Statistics

Charts

Advanced Charts

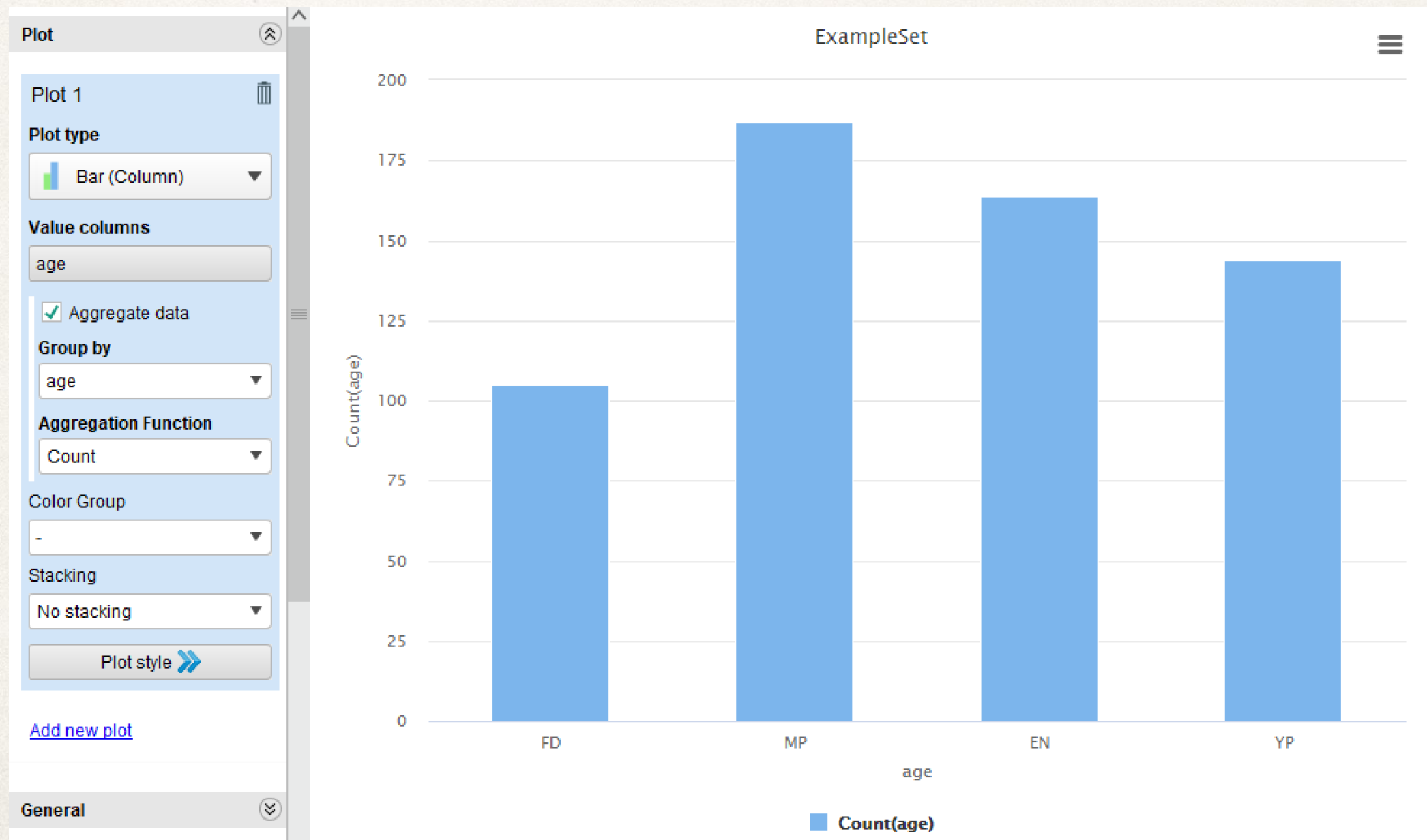
ExampleSet (600 examples, 2 special attributes, 7 regular attributes)

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

The screenshot displays a data mining software interface with three main components:

- Operators Panel (Left):** A tree view showing the 'Discretize by Frequency' operator under the 'Binning (1)' category.
- Canvas (Center):** A workflow diagram showing a 'Retrieve customer_...' operator connected to a 'Discretize' operator. The 'Discretize' operator has inputs 'exa', 'ori', and 'pre', and an output 'res'.
- Parameters Panel (Right):** A configuration window for the 'Discretize (Discretize by Frequency)' operator. The parameters are:
 - attribute filter type:** single
 - attribute:** age
 - invert selection:** ☐
 - include special attributes:** ☐
 - number of bins:** 3

Discretization: Equal Frequency

❖ The result after transformation

Result History

ExampleSet (Discretize)

Data

Statistics

Visualizations

Annotations

Open in

Turbo Prep

Auto Model

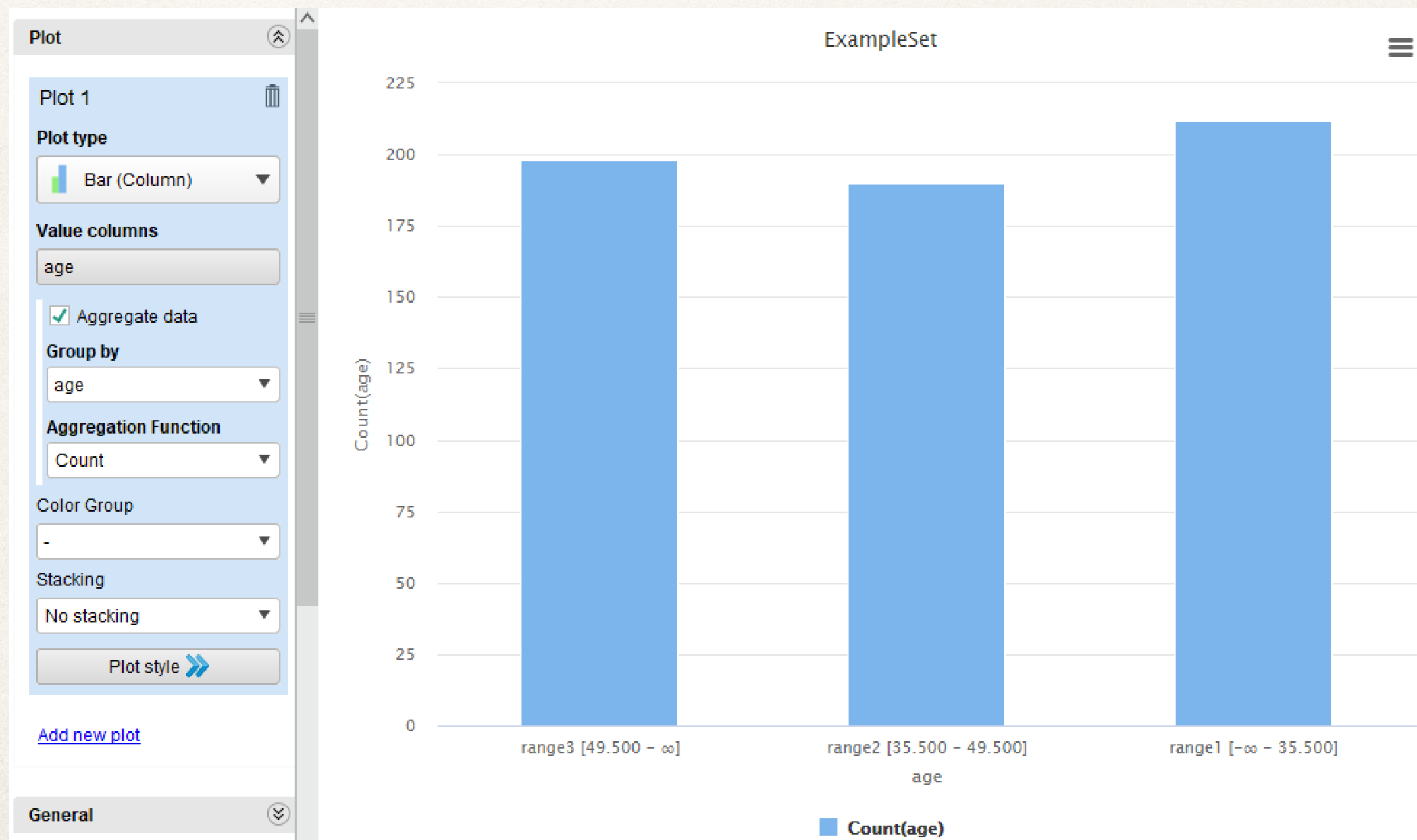
Filter (600 / 600 examples):

all

| Row No. | ID customer... | 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>