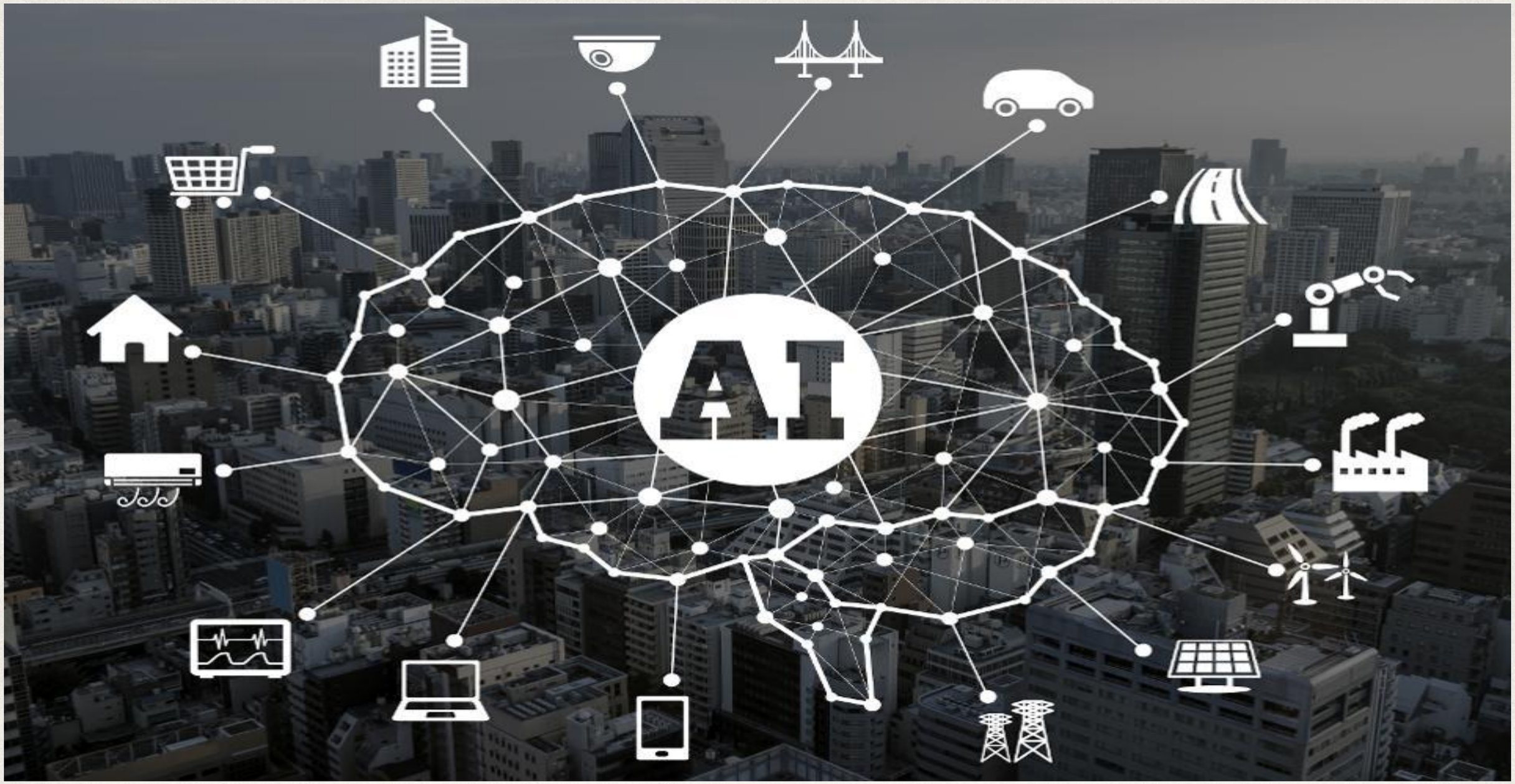


SE634: Artificial Intelligence in Industrial Engineering

Data preprocessing and introduction to RapidMiner Studio

SE634 Assoc. Prof. Dr. Warut Pannakkong Sirindhorn International Institute of Technology, Thammasat University

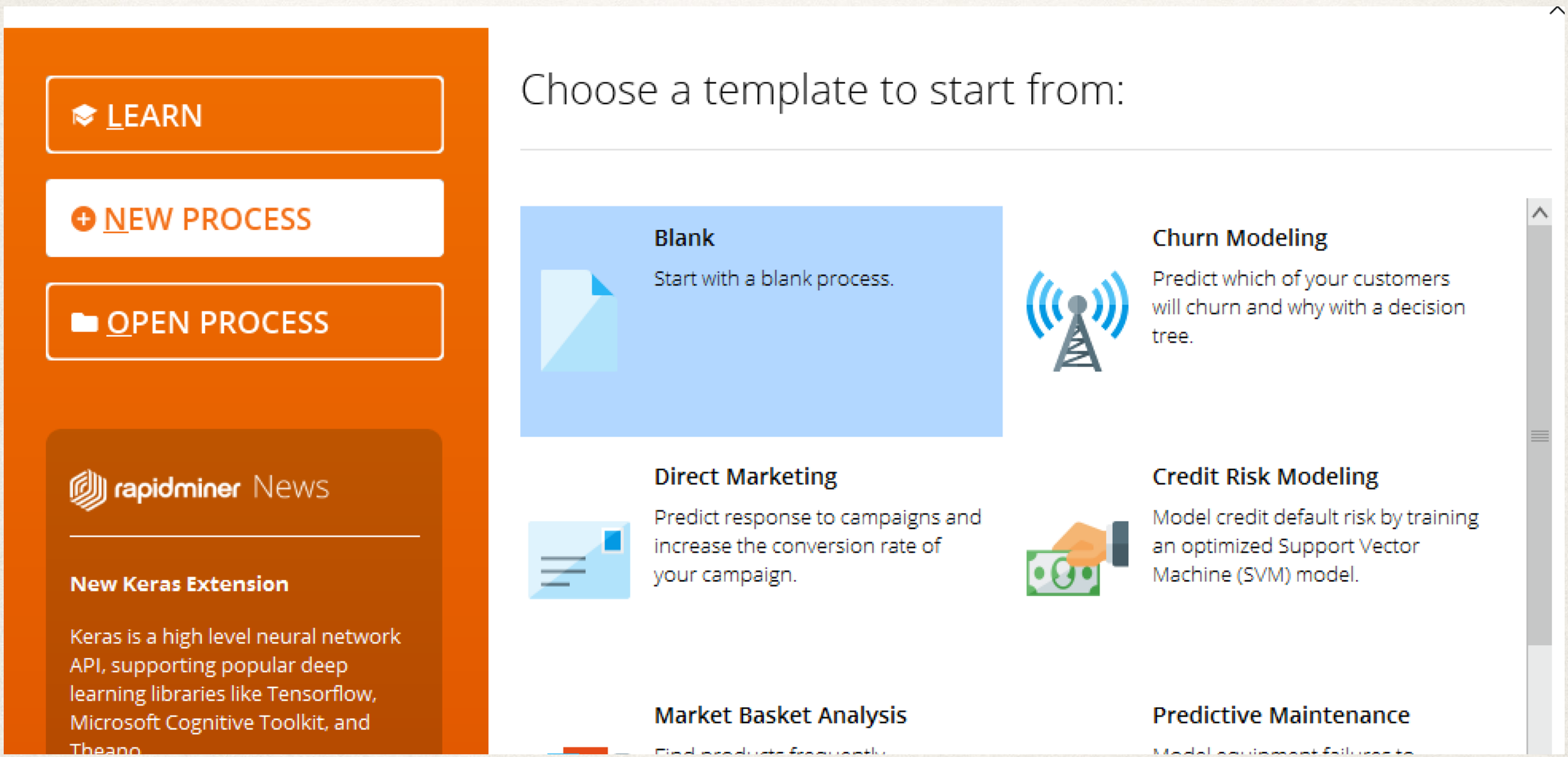
Introduction to RapidMiner Studio

https://rapidminer.com/

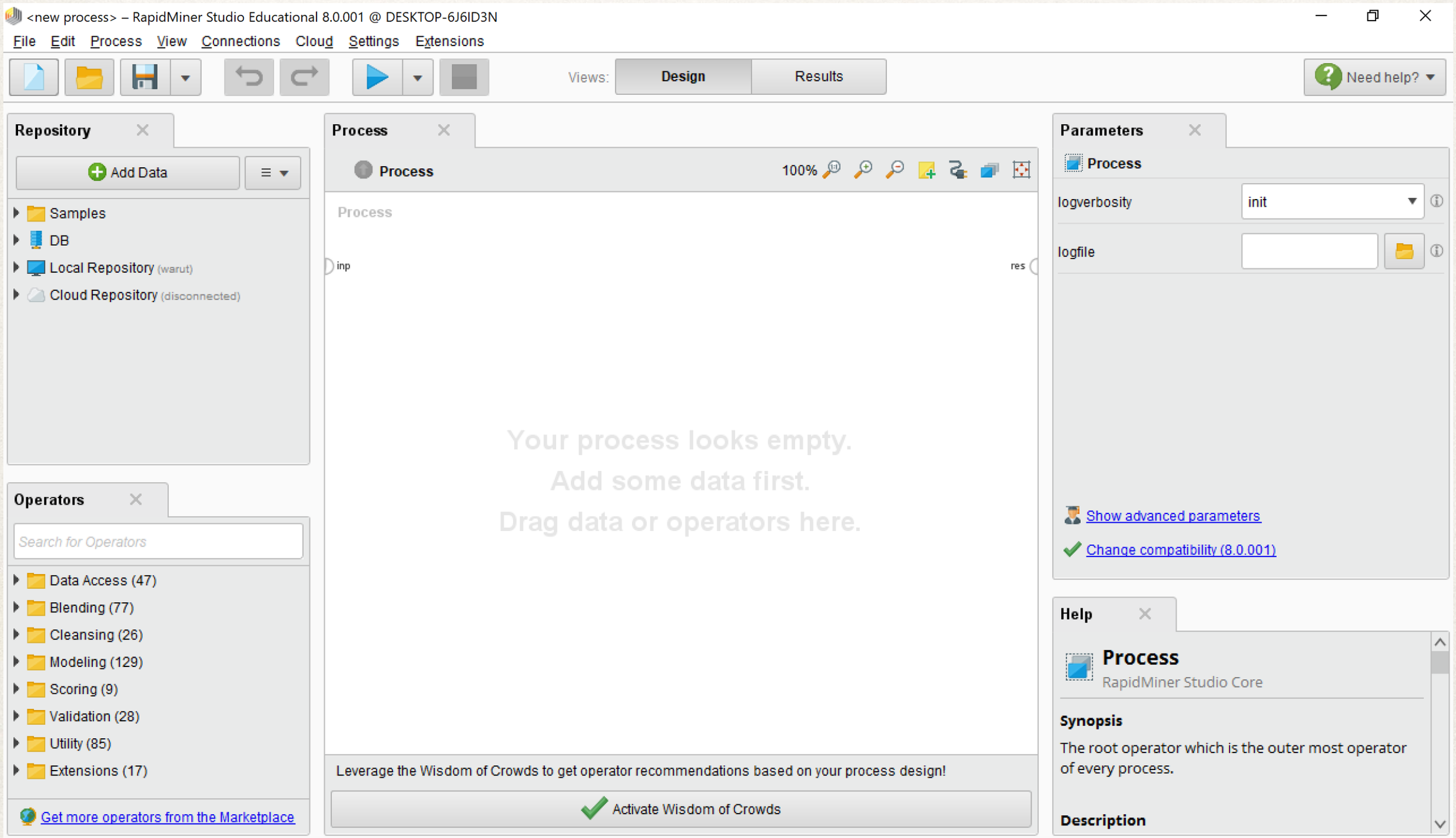
https://my.rapidminer.com/nexus/account/index.html#signup

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

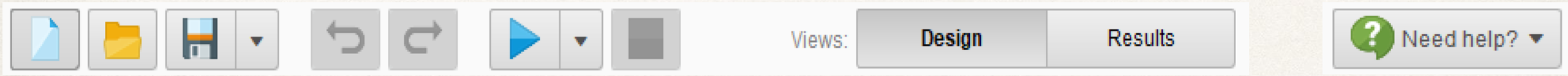
Home Screen


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
Predictive Maintenance

Interface


<new process> – RapidMiner Studio Educational 8.0.001 @ DESKTOP-6J6ID3N
File Edit Process View Connections Cloud Settings Extensions
Views: Design Results
Need help?
Repository
Add Data
Samples
DB
Local Repository (warut)
Cloud Repository (disconnected)
Operators
Search for Operators
Data Access (47)
Blending (77)
Cleansing (26)
Modeling (129)
Scoring (9)
Validation (28)
Utility (85)
Extensions (17)
Get more operators from the Marketplace
Process
Process
100%
Process
inp
res
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
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

Menu



Create new process

Open existing process

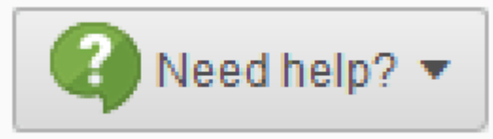
Save

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

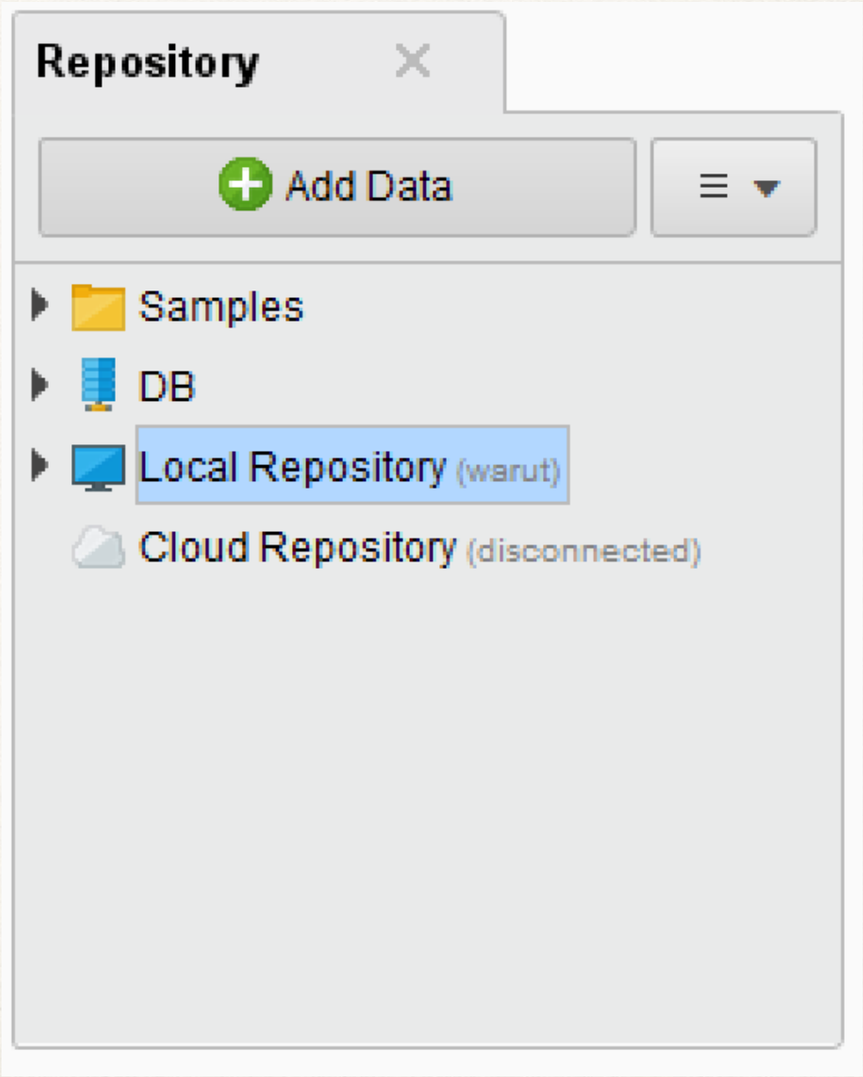
Value Type

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

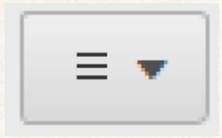
customer_id	age	gender	region	income	married	children	car	response
ID12101	48	FEMALE		17546	NO	1	NO	NO
ID12102	40	MALE	TOWN	30085.1	YES	3	YES	NO
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES
ID12104	23	FEMALE	TOWN	20375.4	YES	3	NO	NO
Polynomial	Numeric	Binomial		Numeric	Binomial	Numeric	Binomial	

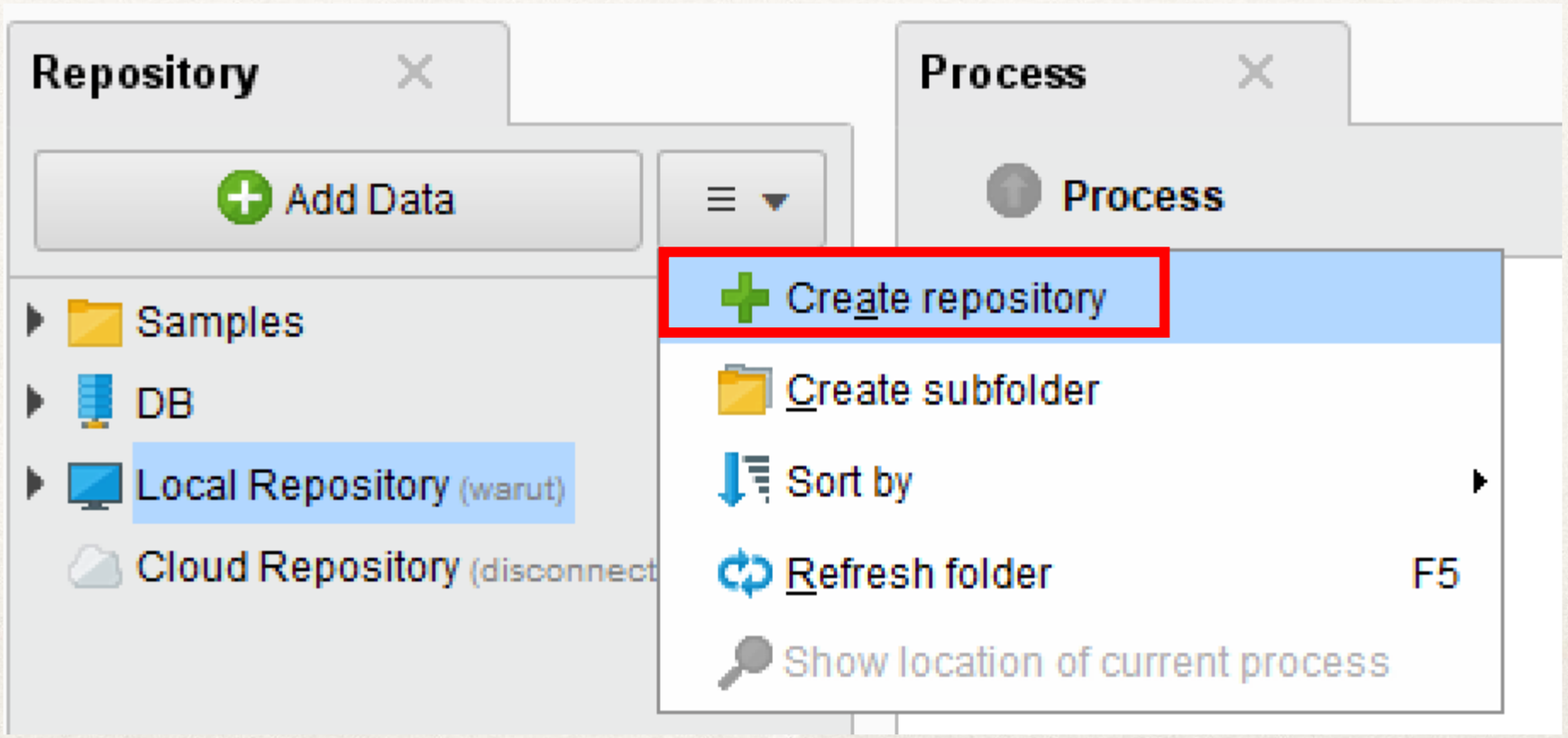
Data Management

- 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



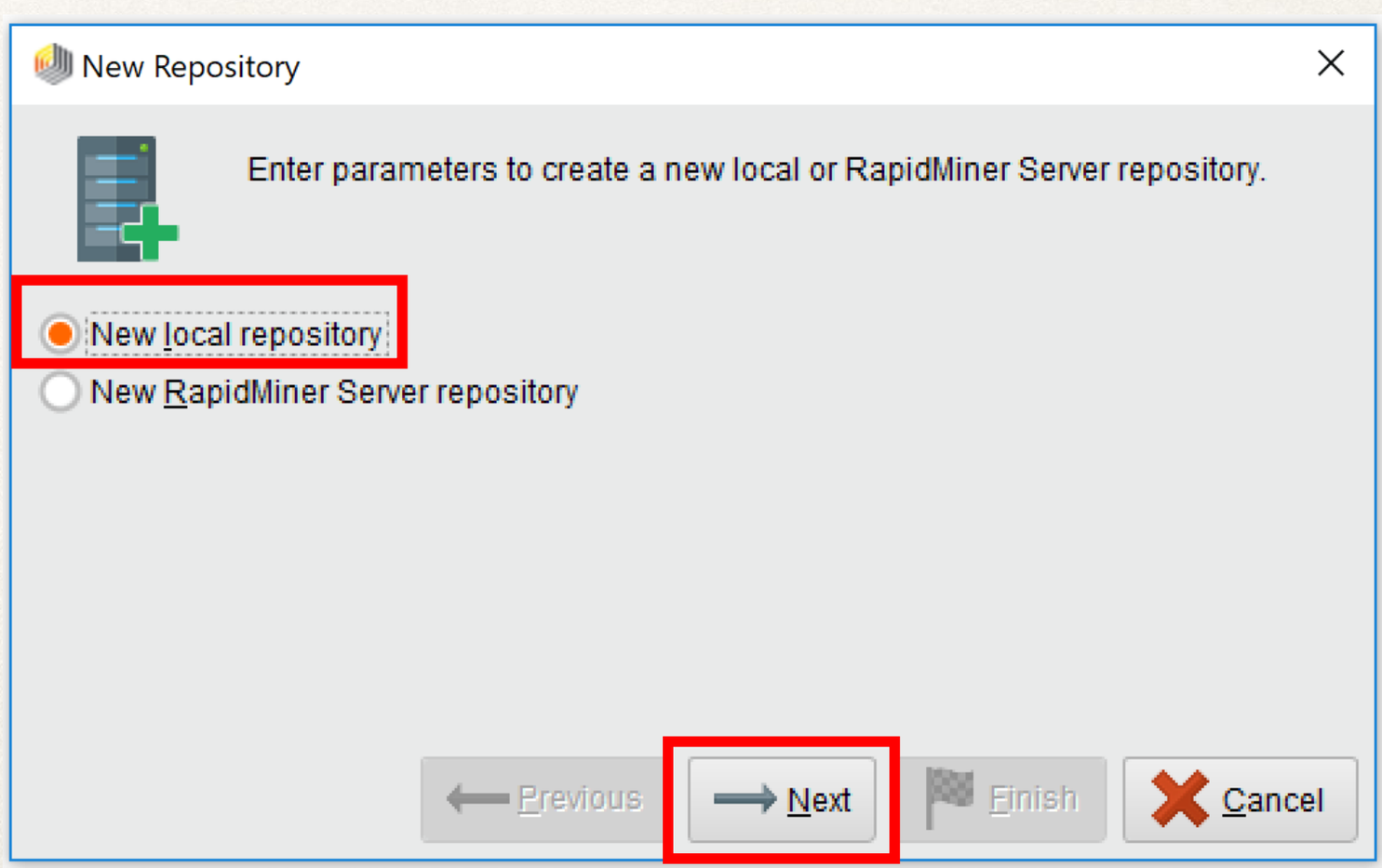
Data Management

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



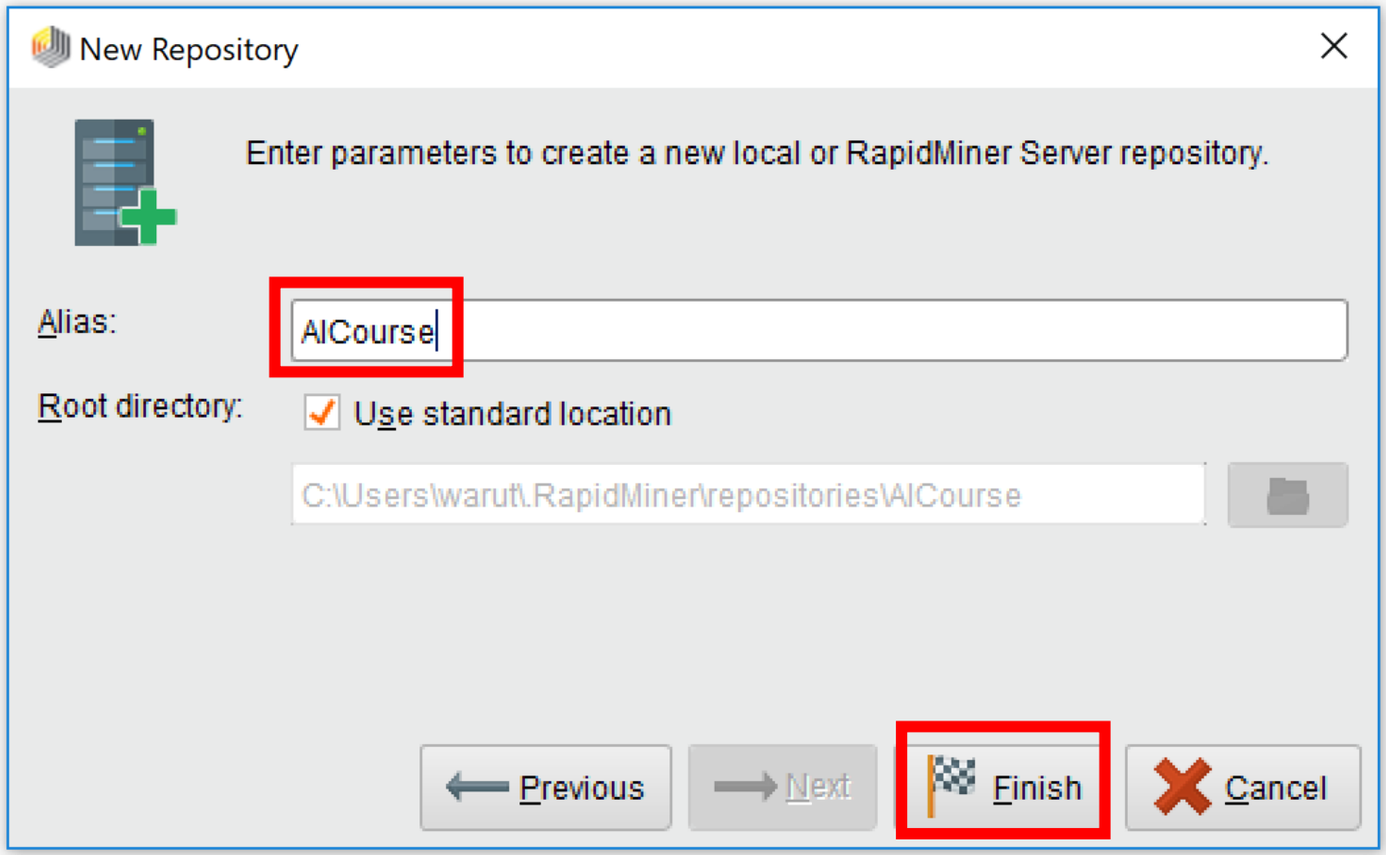
Data Management

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



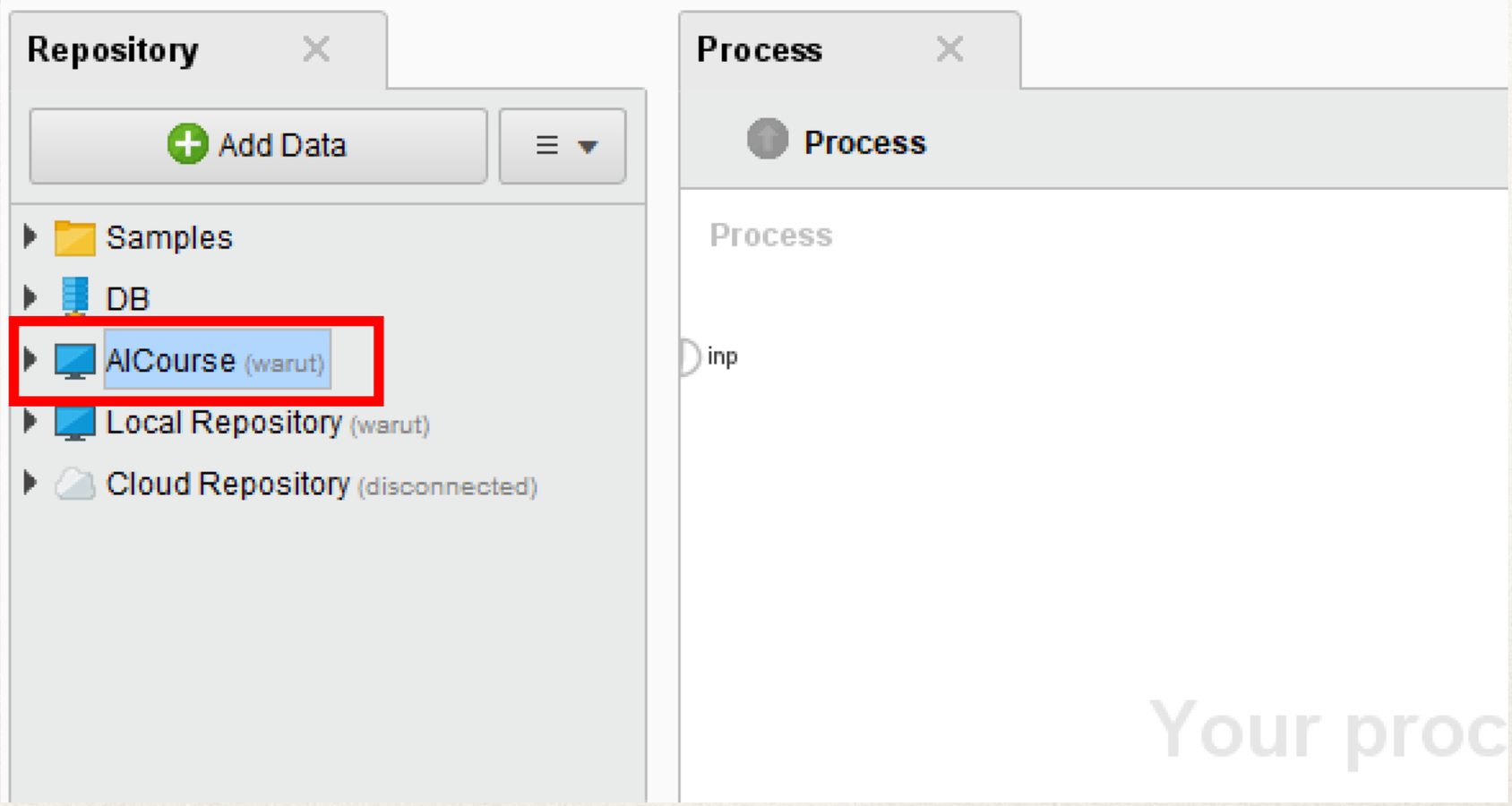
Data Management

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



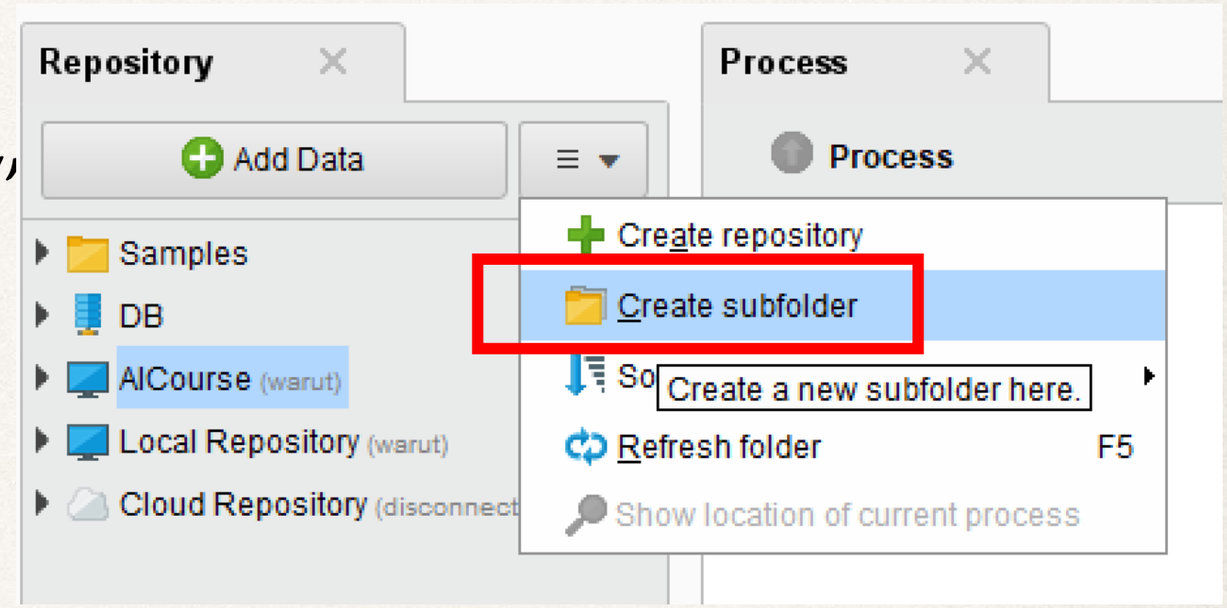
Data Management

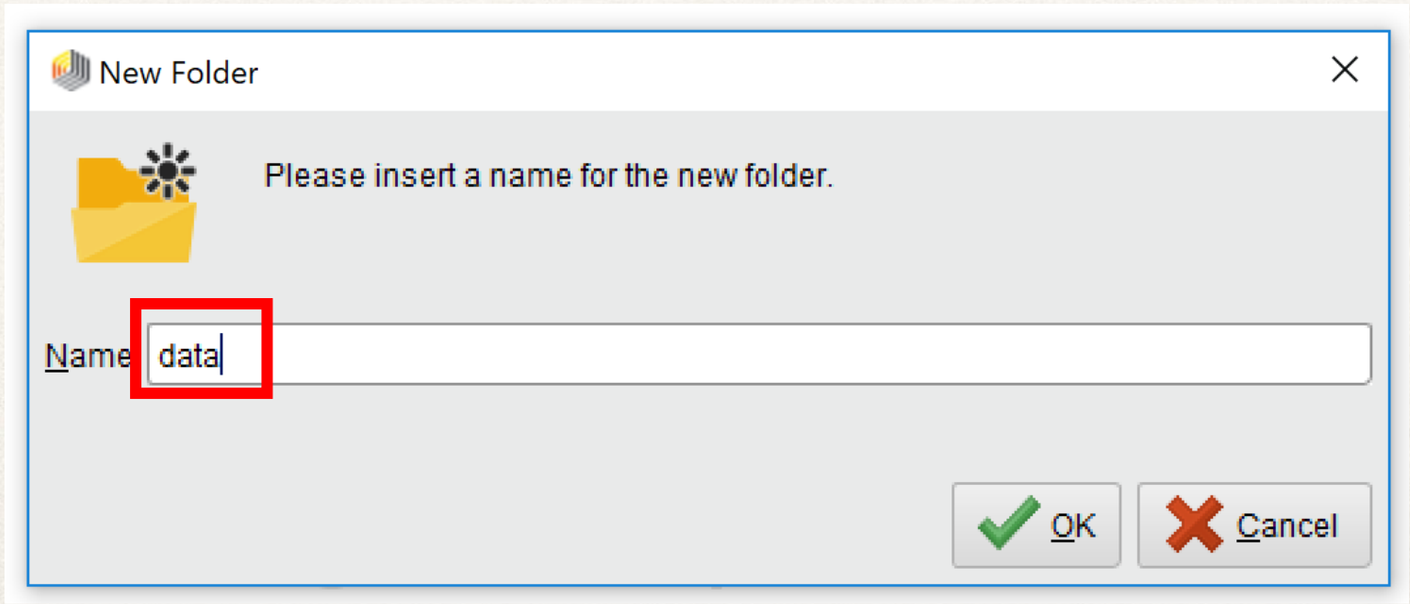
- “AICourse” will be appeared in repository.



Data Management

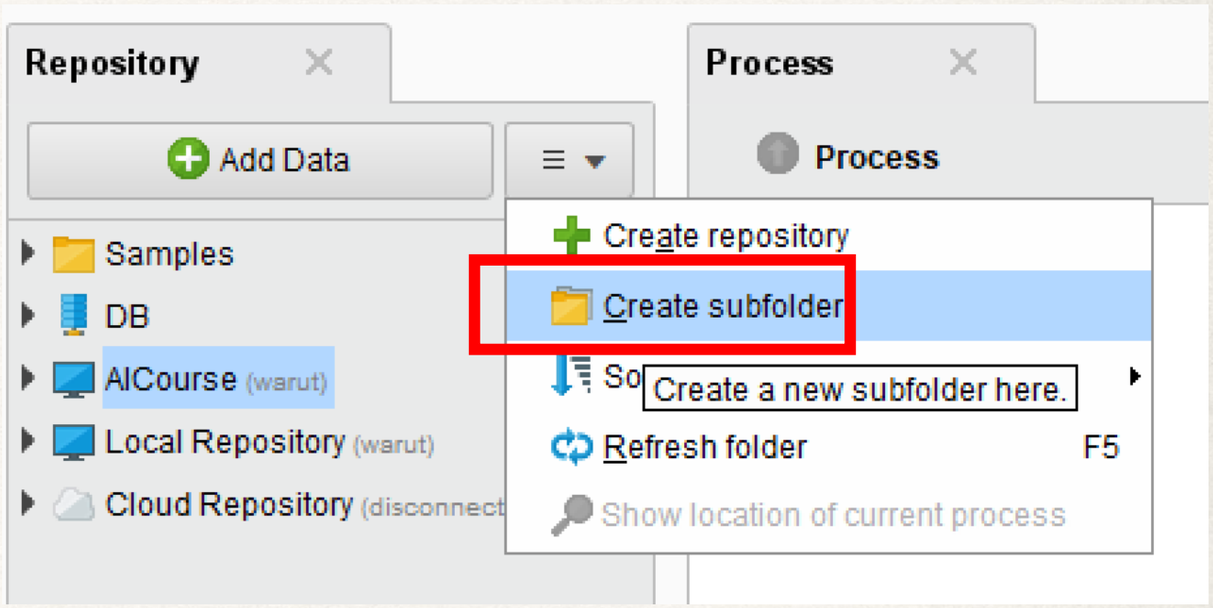
- Create folder in “AICourse”
  - “data” for data storage

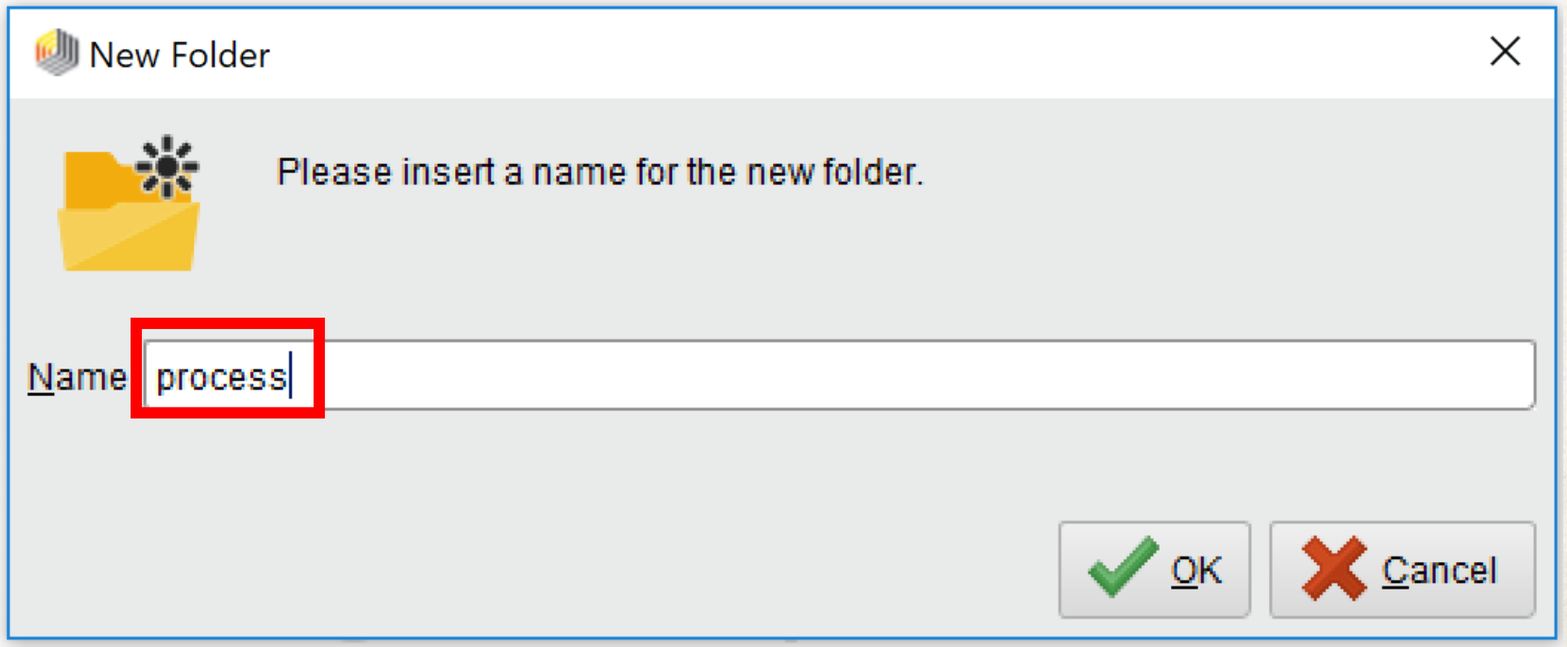




Data Management

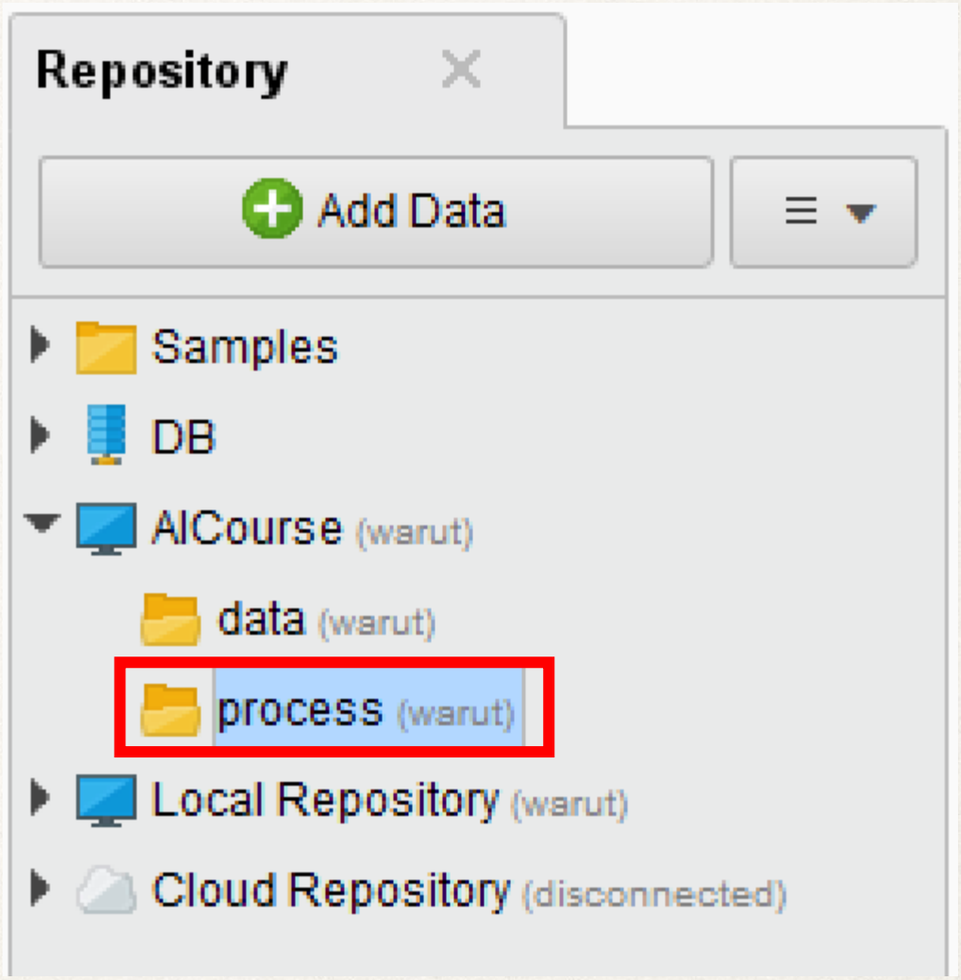
- Create folder in "AICourse"
  - "data" for data storage
  - "process" for processes storage





Data Management

- Create folder in "AICourse"
  - "data" for data storage
  - "process" for processes storage



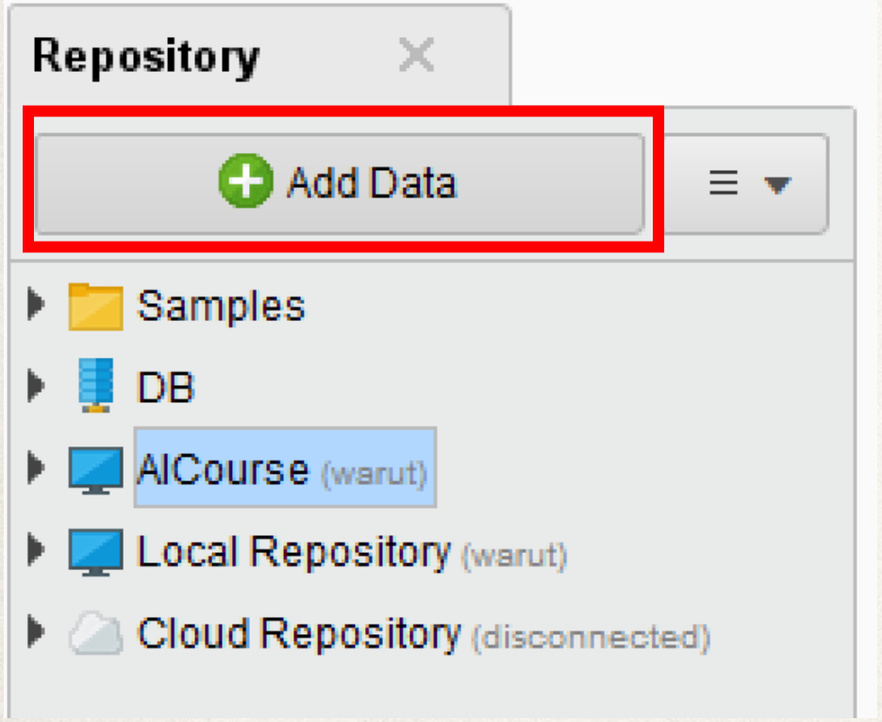
Data Pre-Processing with RapidMiner

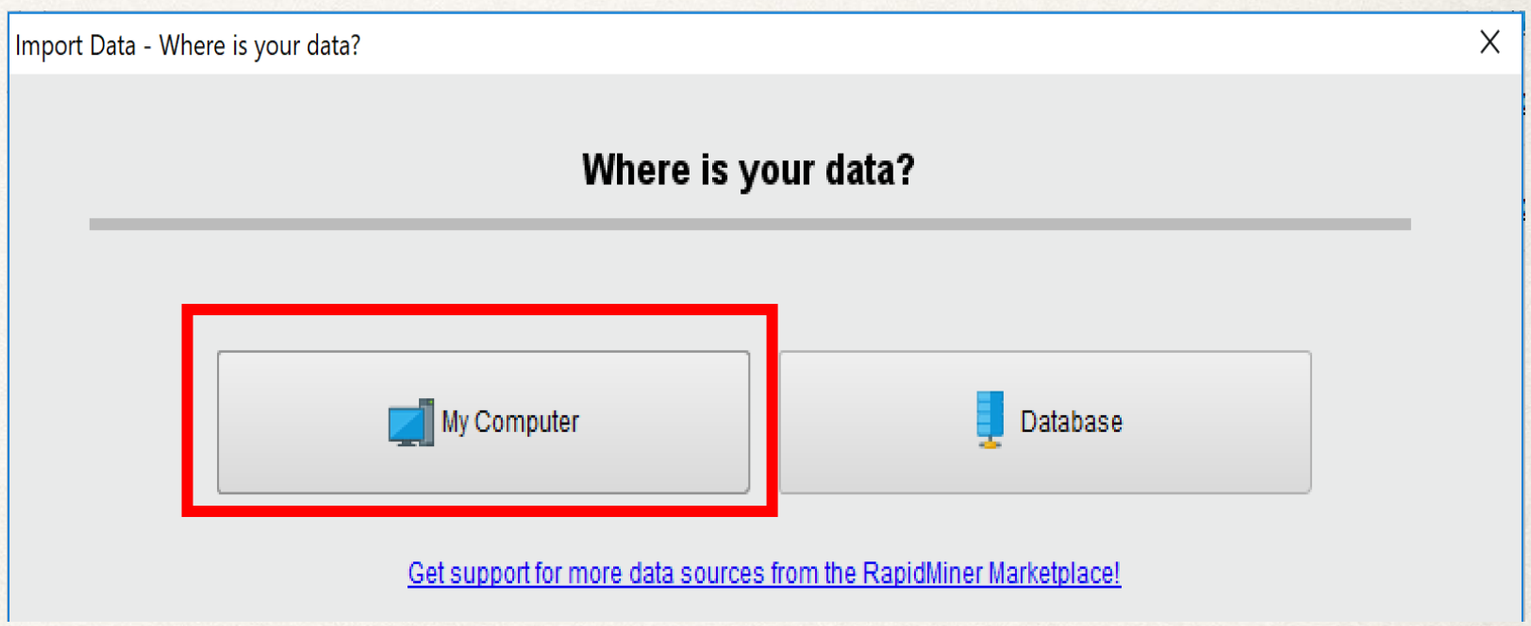
- Data management
- **Import data**
- Data exploration
- Data preparation

Import Data into Repository

- Click 

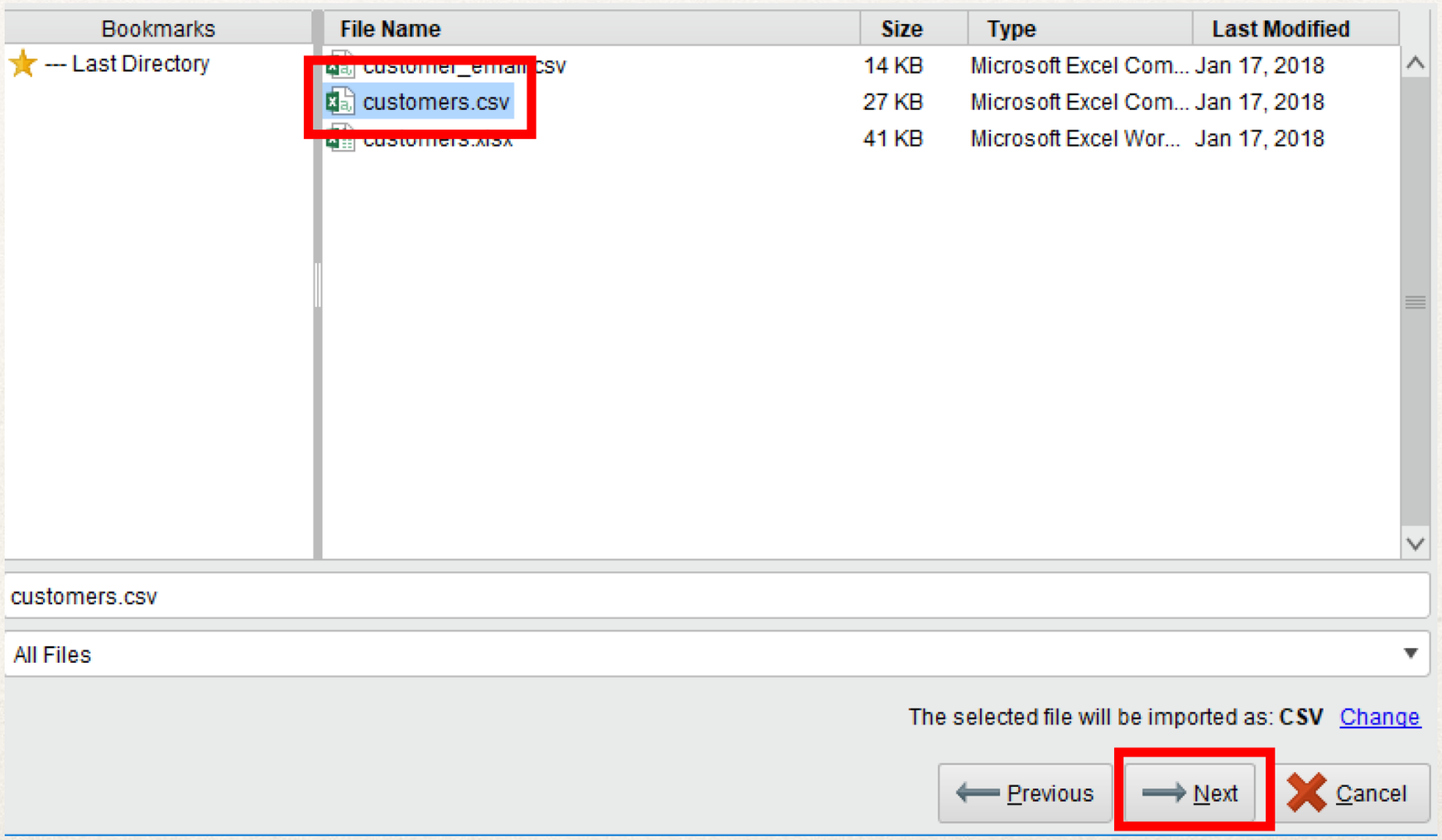
- Choose “My Computer”





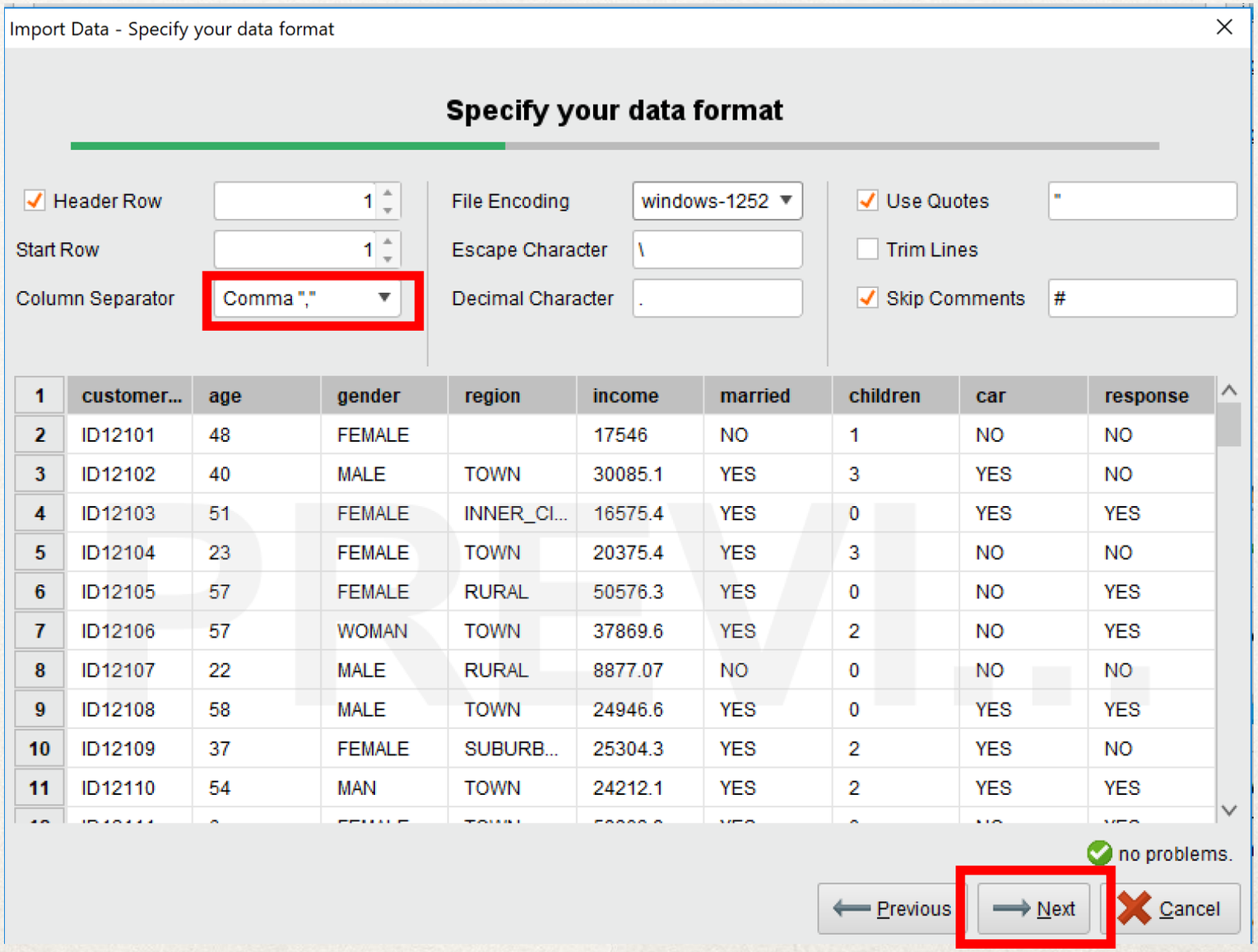
Import Data into Repository

- Choose “customer.csv”



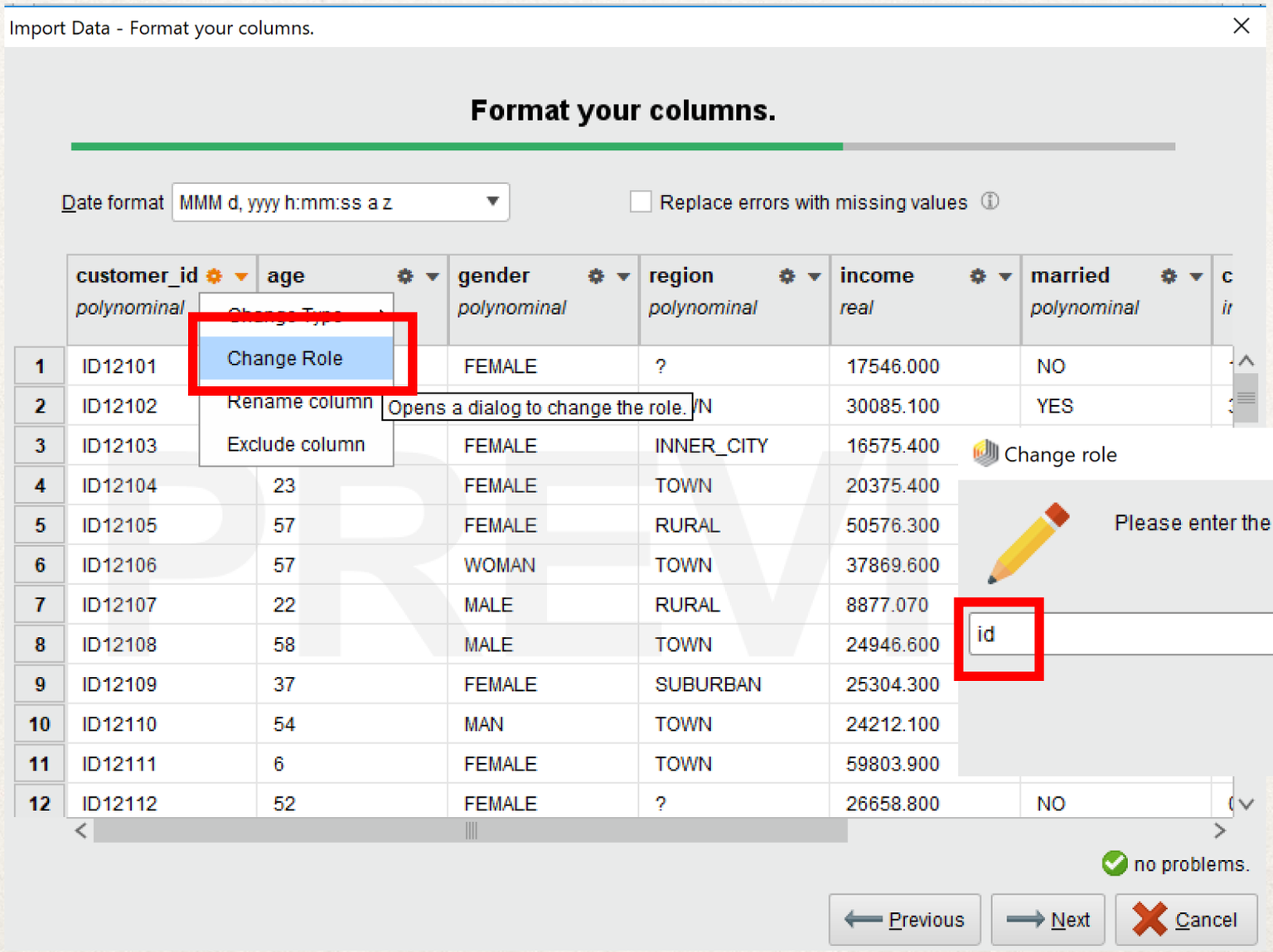
Import Data into Repository

- For Column Separator, choose Comma “,”



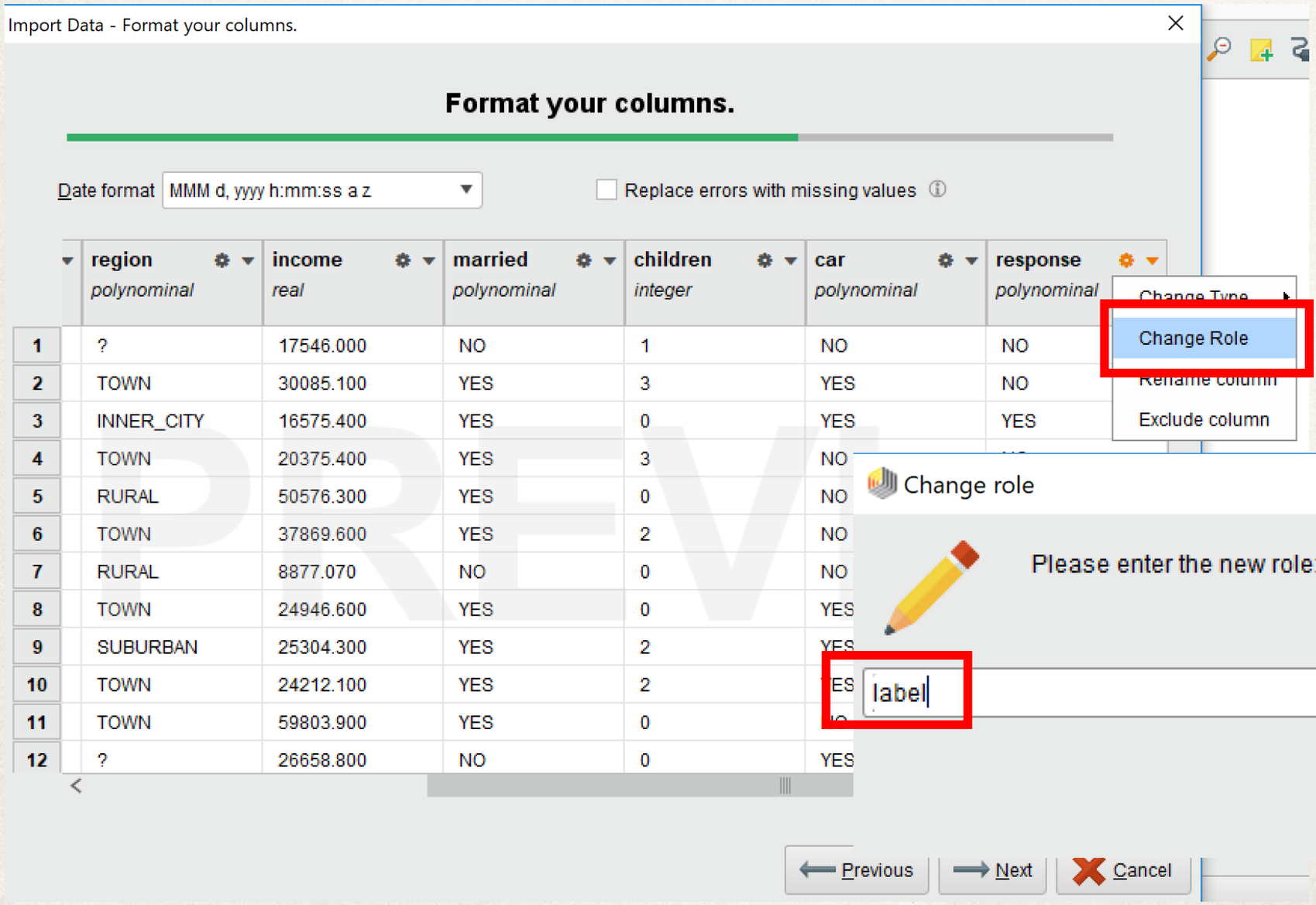
Import Data into Repository

- Change role of customer_id to id



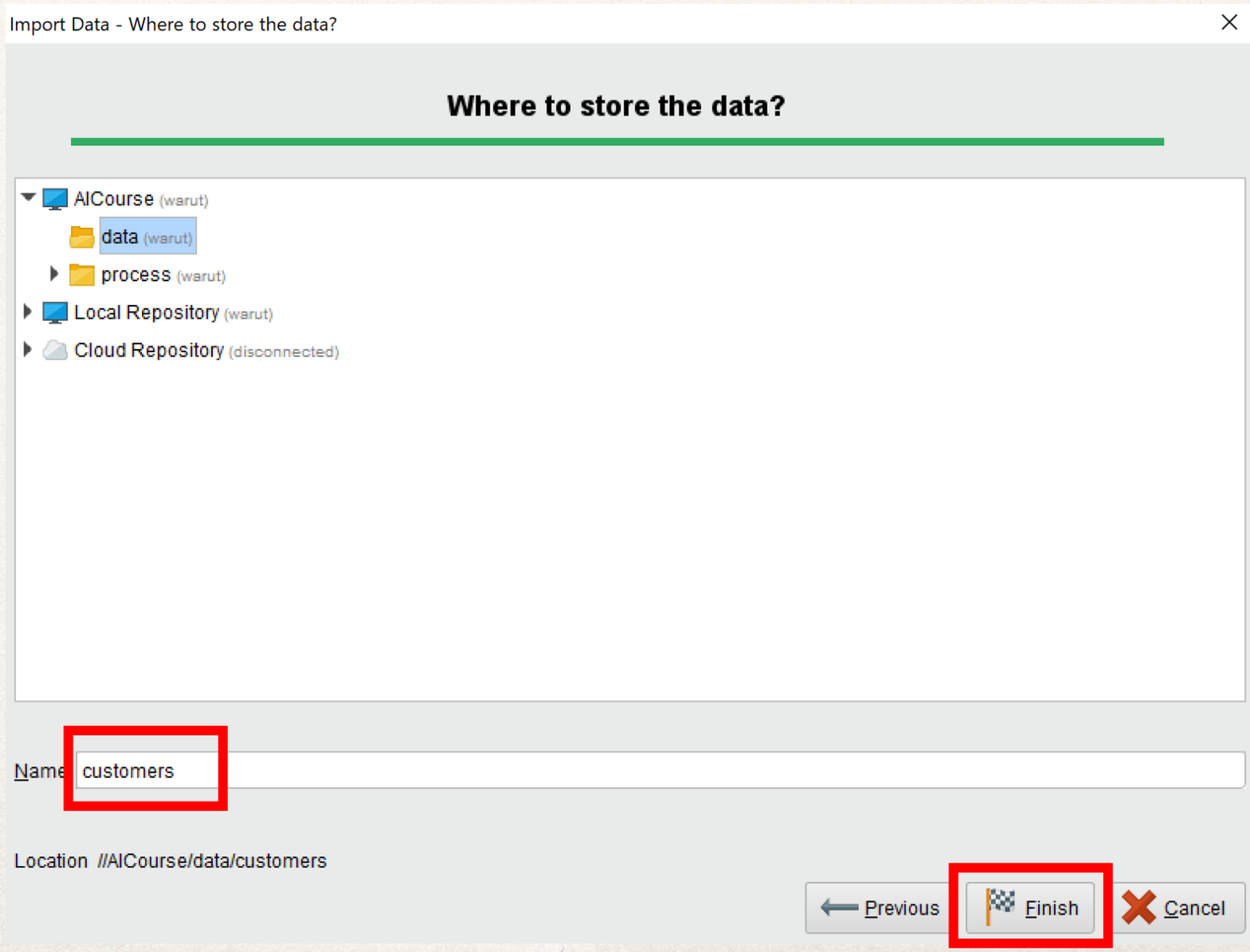
Import Data into Repository

- Change role of response to label



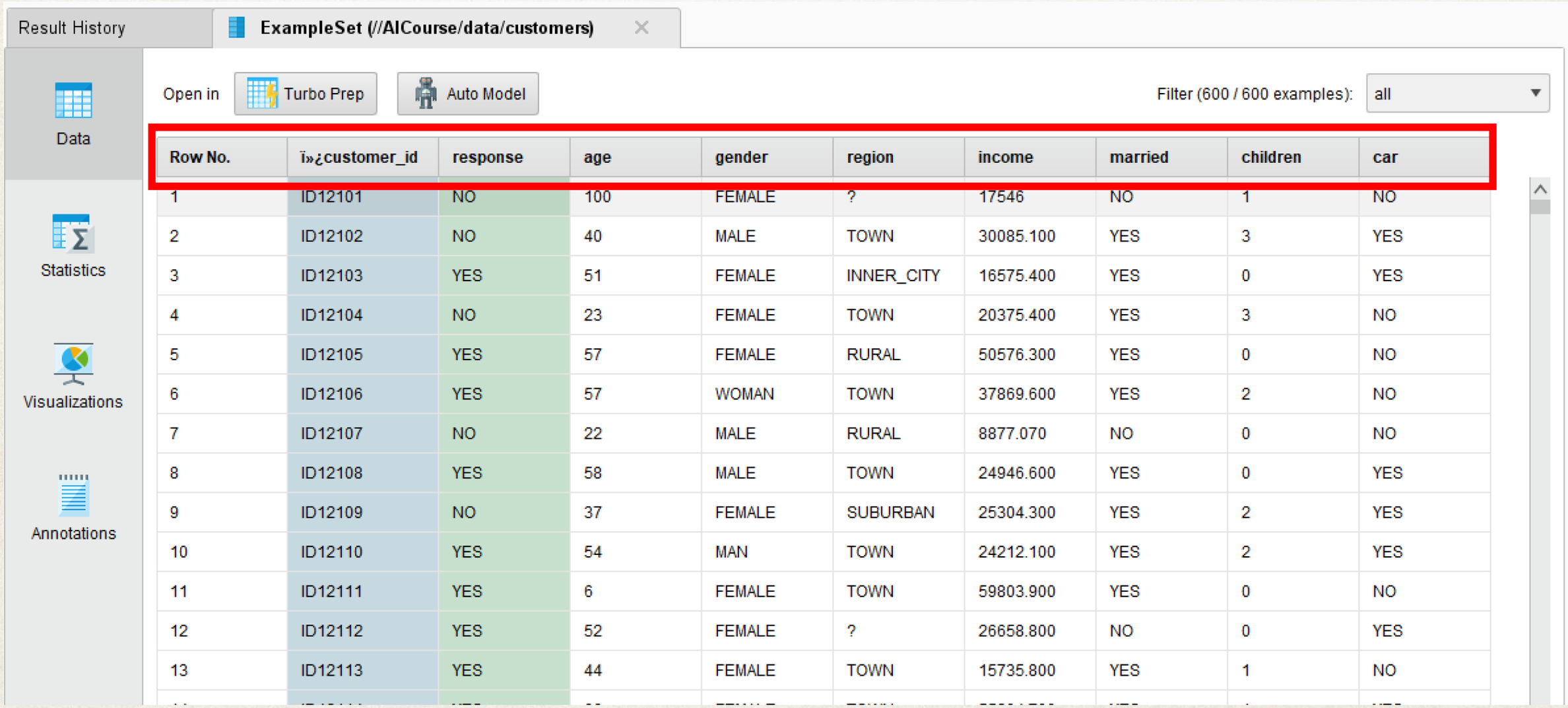
Import Data into Repository

- Name = customers
- Save into folder data under AICourse



Import Data into Repository

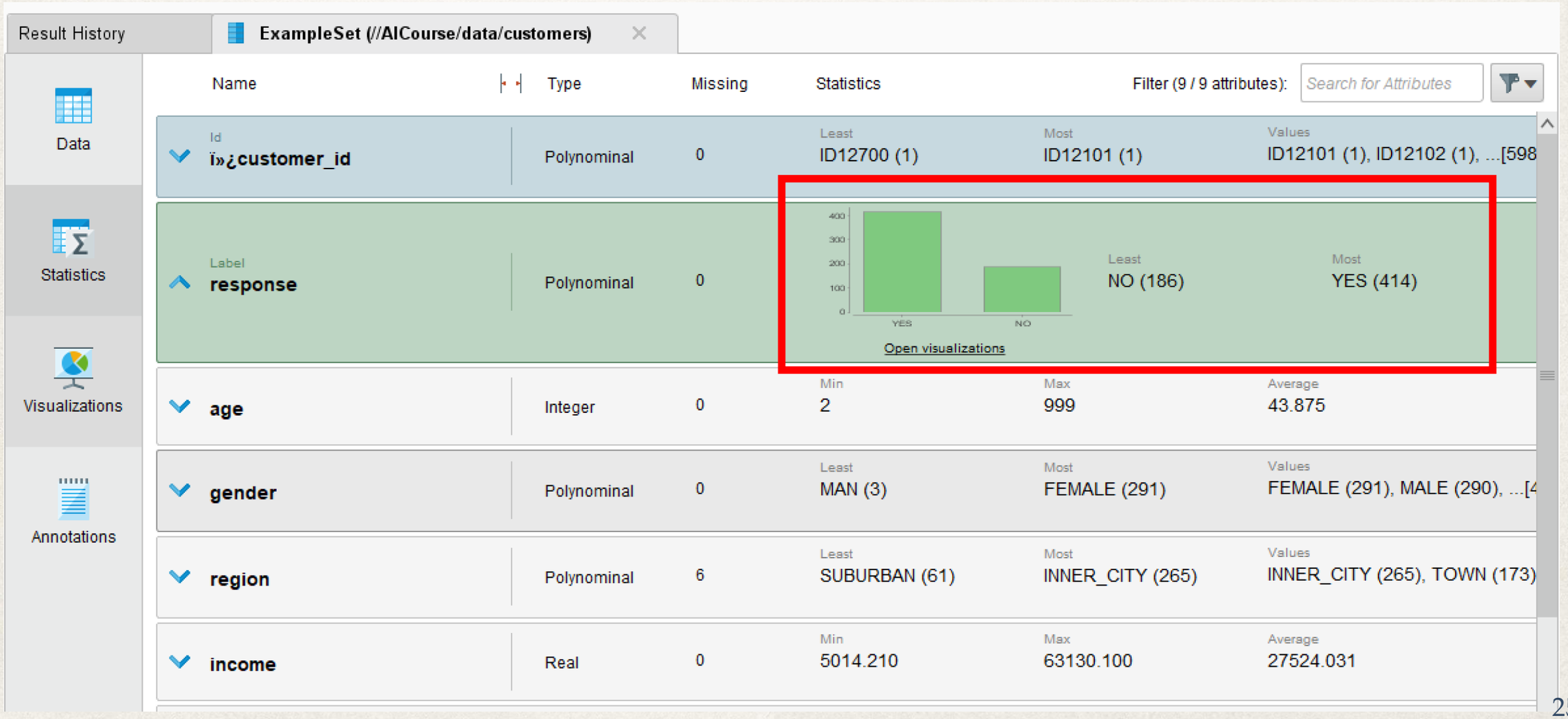
- Click at header of each column to sort the data



Row No.	ï»¿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

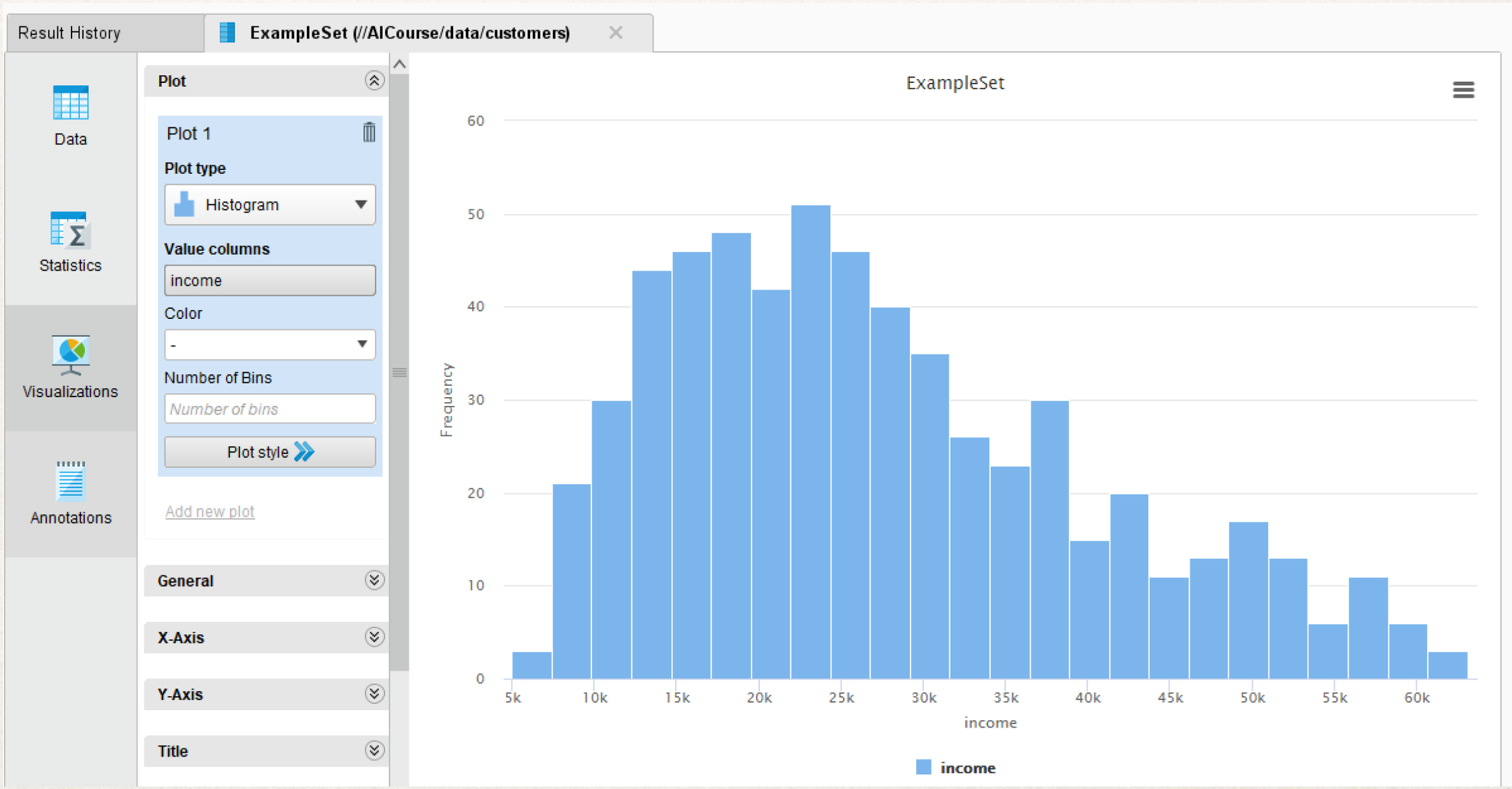
Import Data into Repository

- Statistics



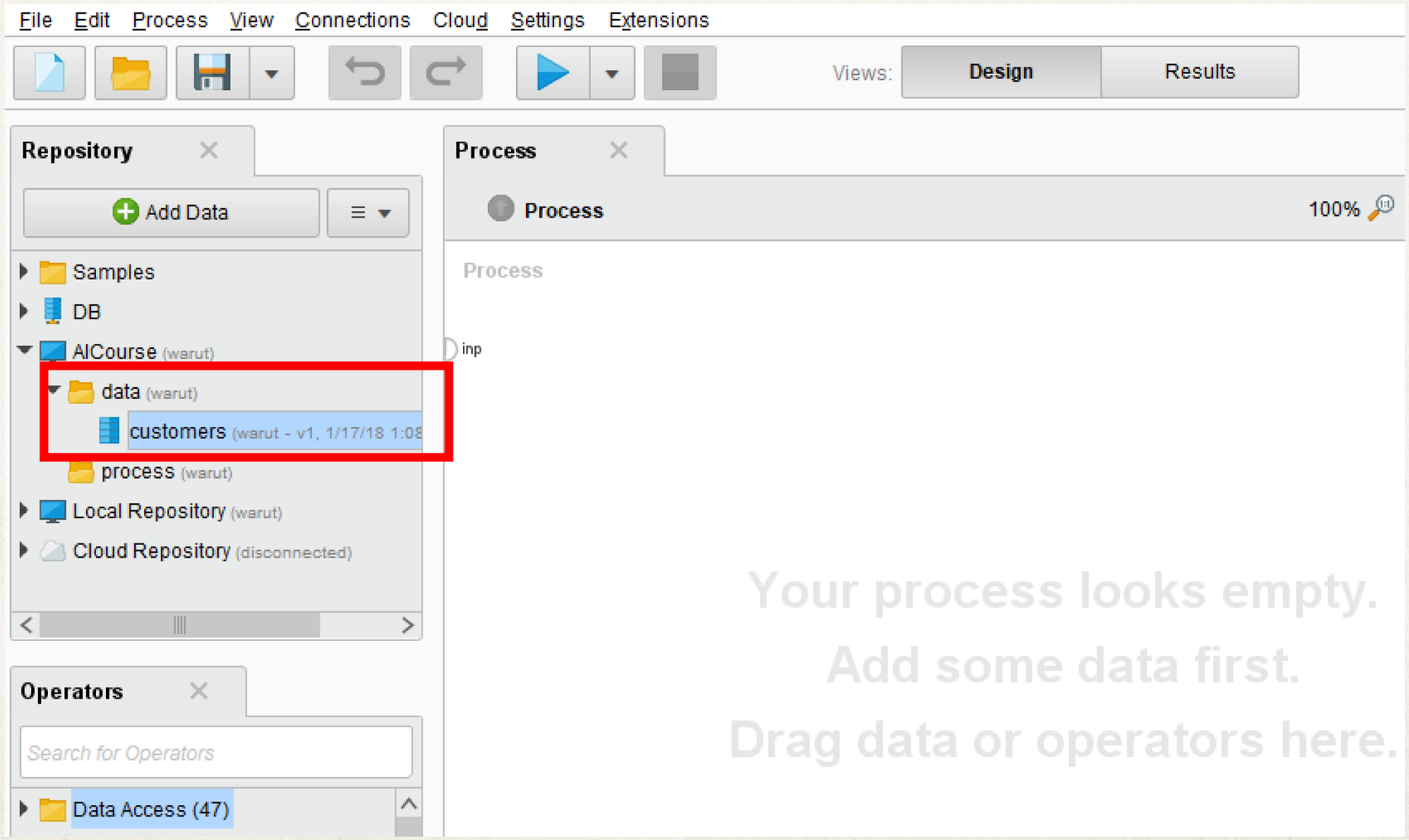
Import Data into Repository

- 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)

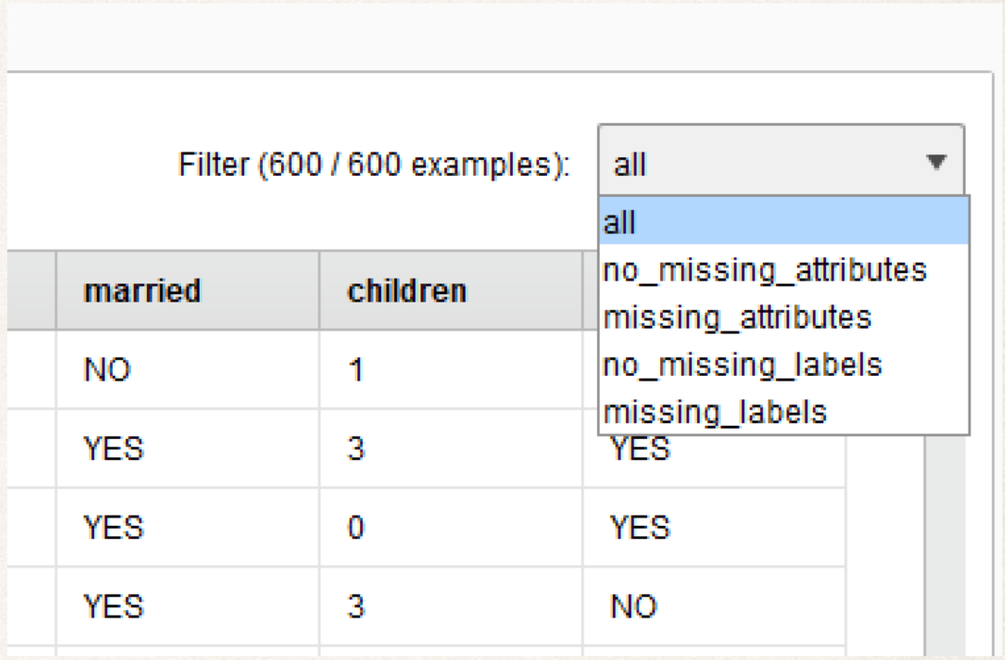
Data
Statistics
Visualizations

Open in Turbo Prep | Auto Model | Filter (60

Row No.	ï»¿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

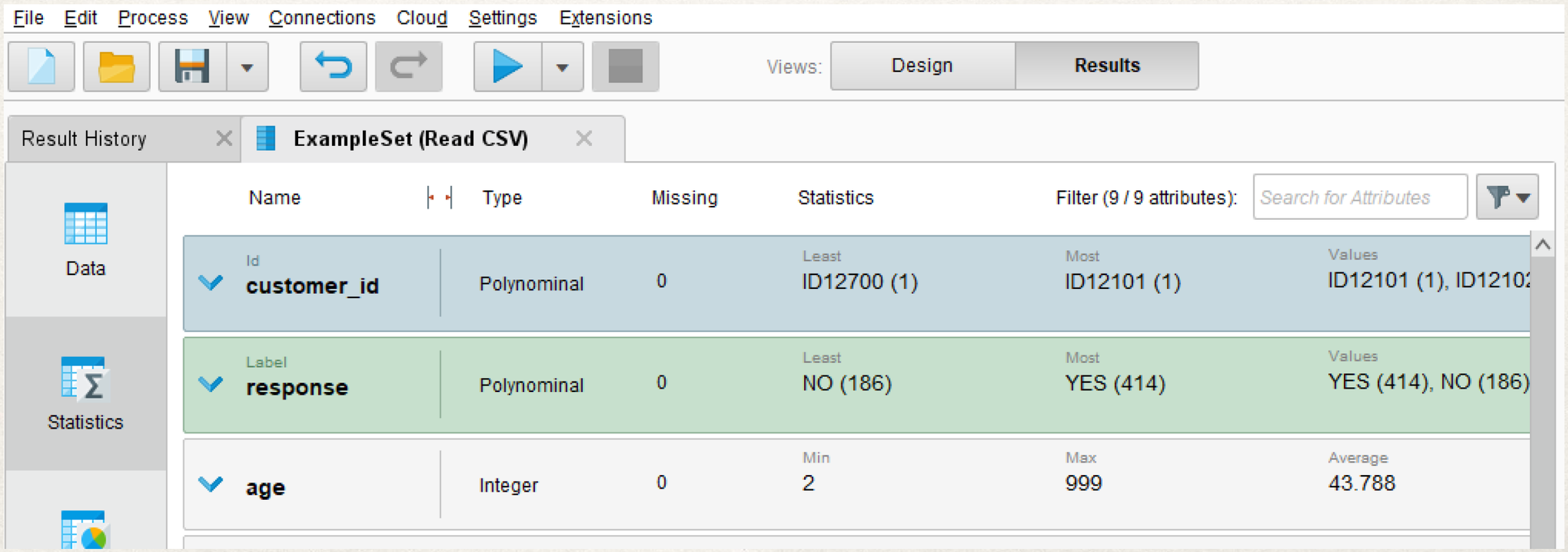
Data Exploration

- 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



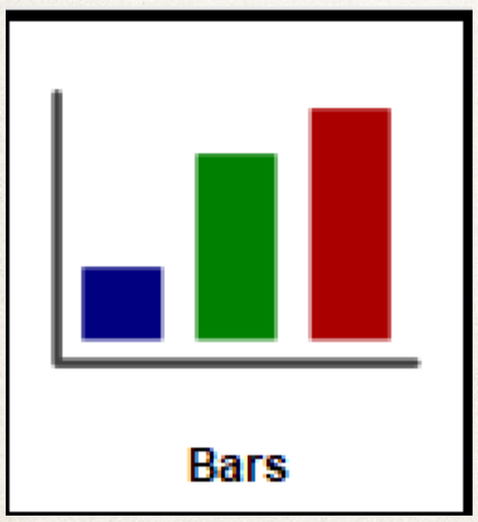
Data Exploration

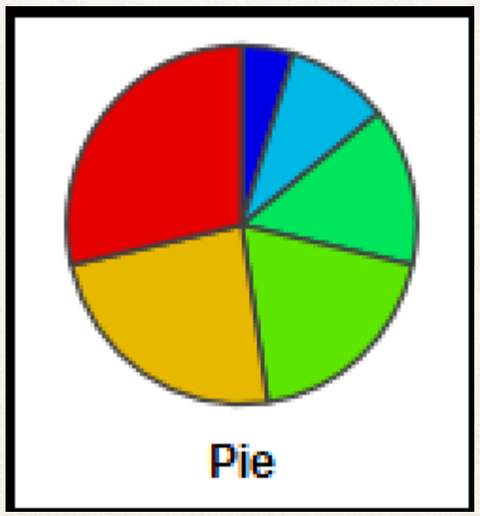
- Statistics
  - Name
  - Type
  - Miss
  - Min, Max, Average, Deviation, Least, Most, Values

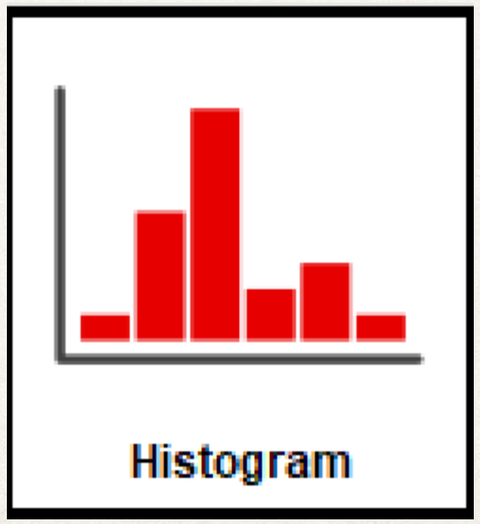


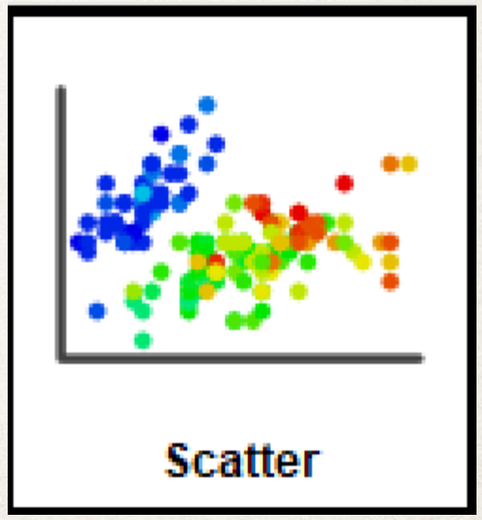
Data Visualization

- Charts
  - Bars
  - Histogram
  - Pie
  - Scatter



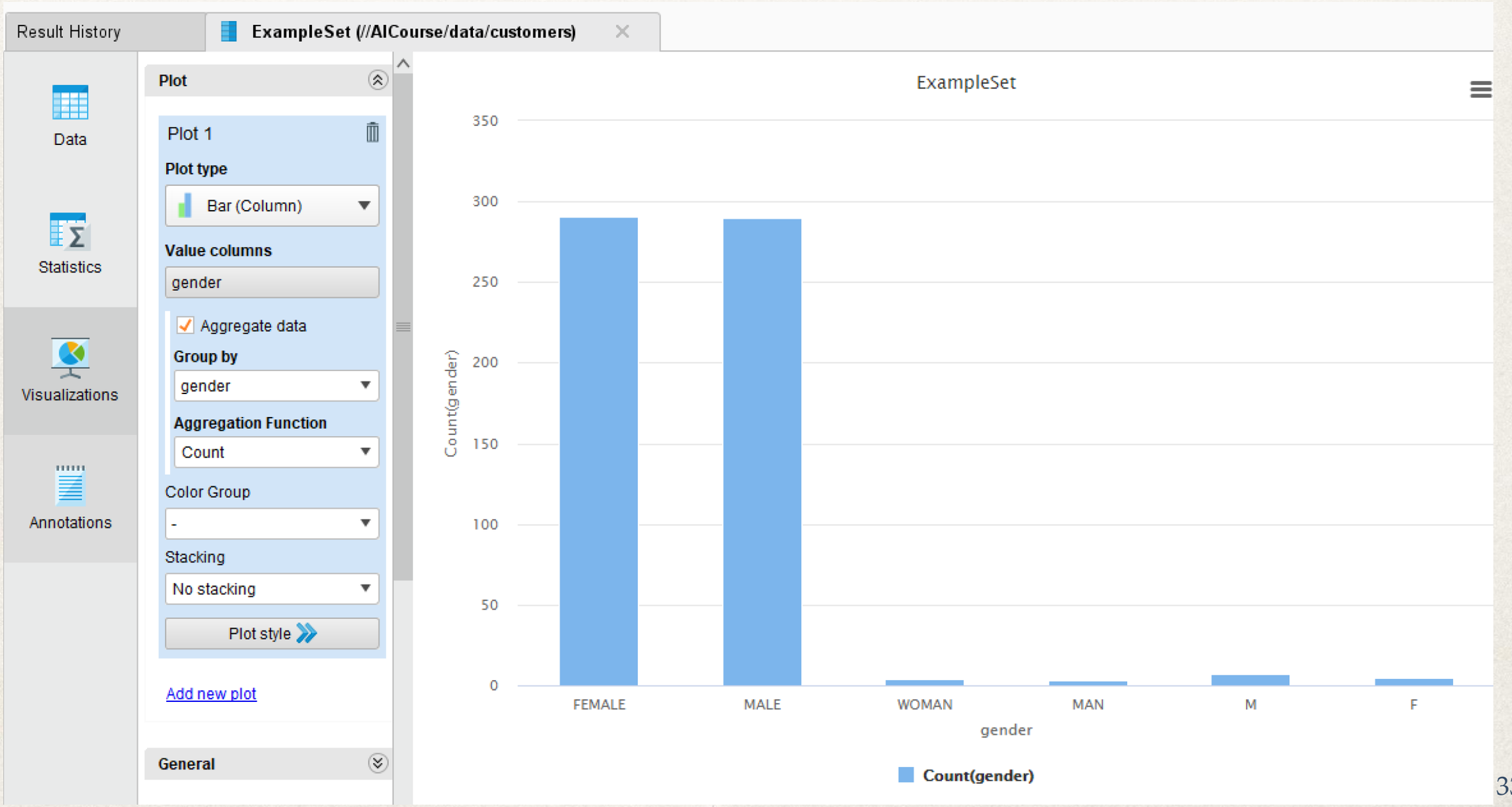






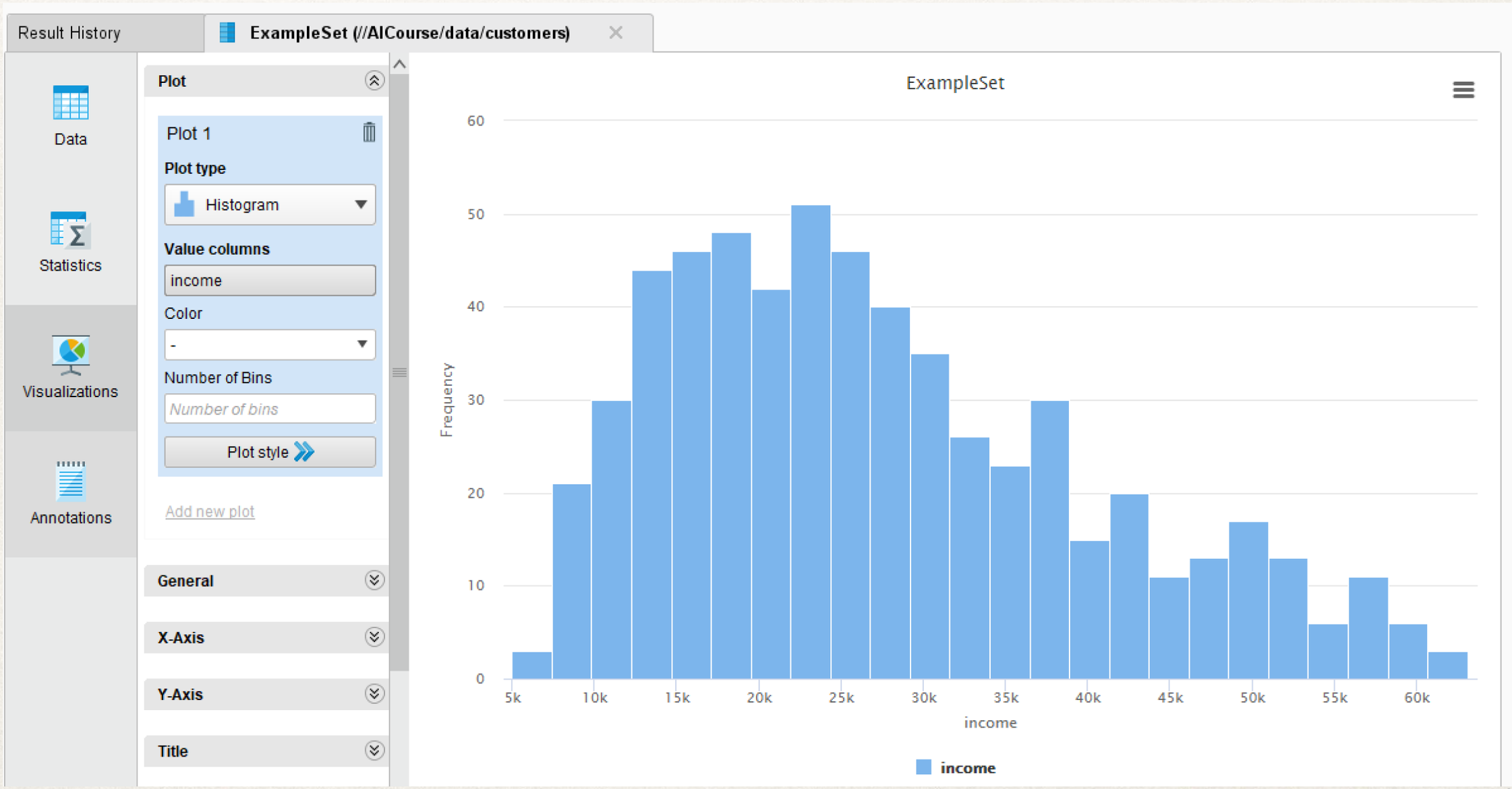
Data Visualization

- Bar chart of gender attribute



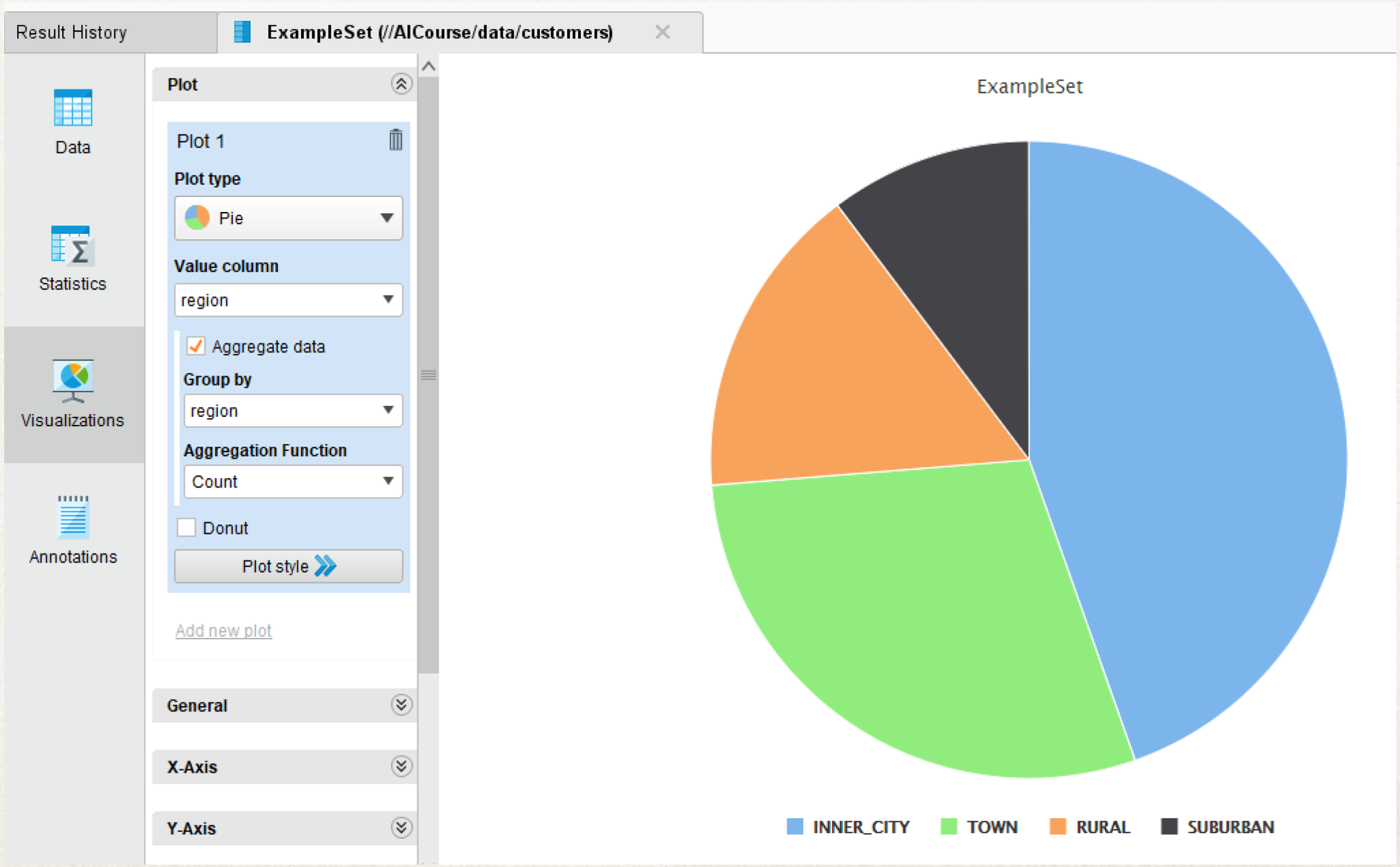
Data Visualization

- Histogram of income attribute



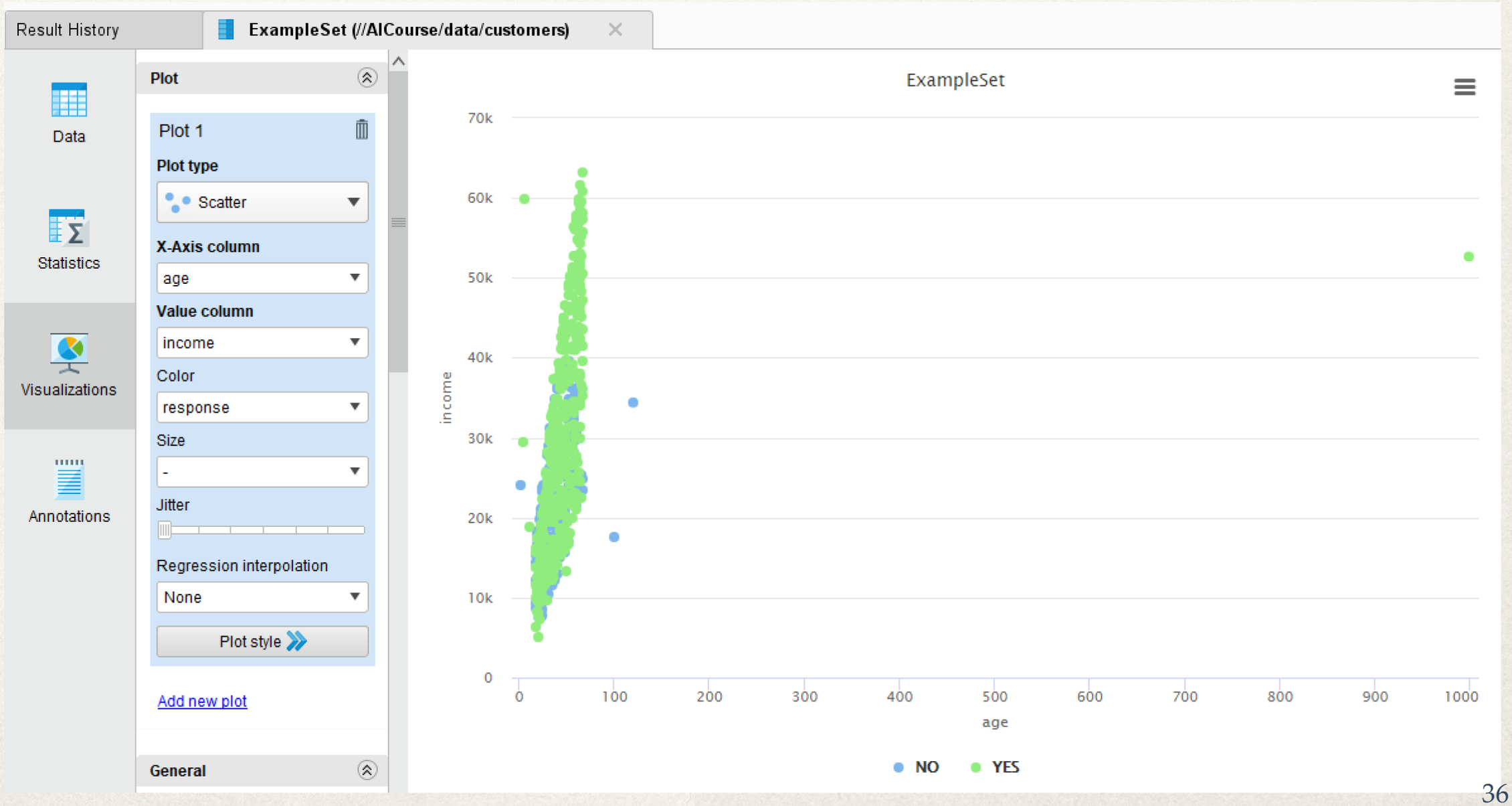
Data Visualization

- Pie chart of region attribute



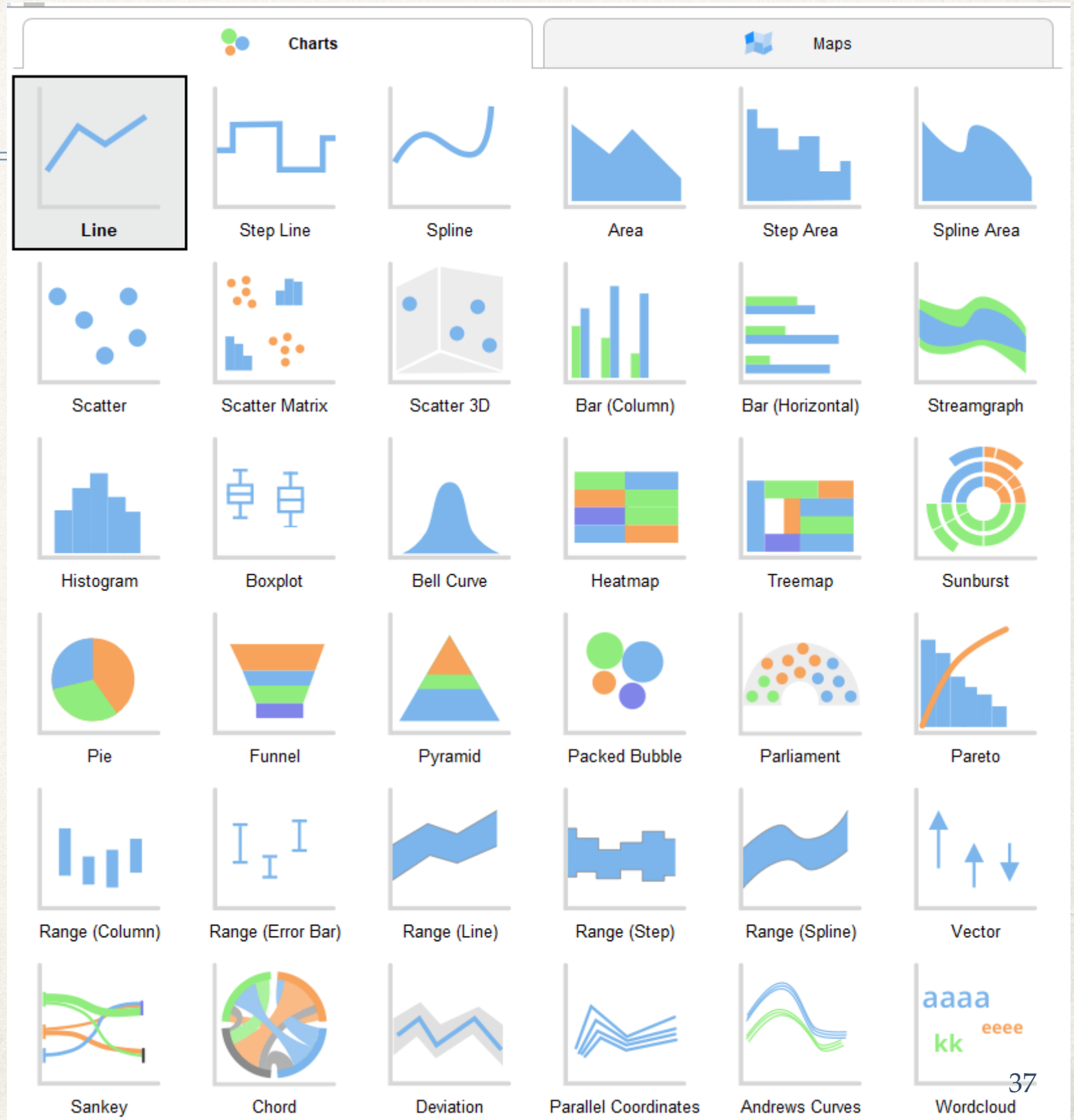
Data Visualization

- Scatter plot of age and income attributes



Data Visualization

- Different types of chart for different proposes

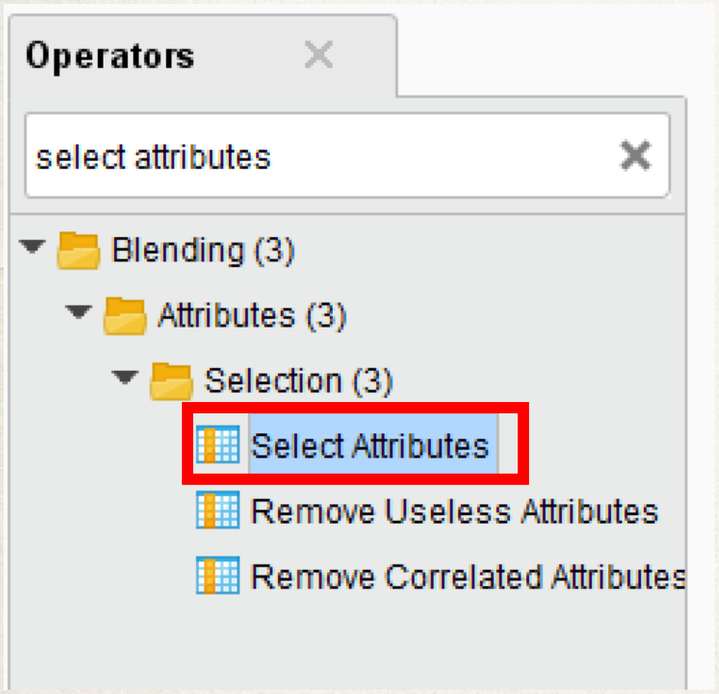


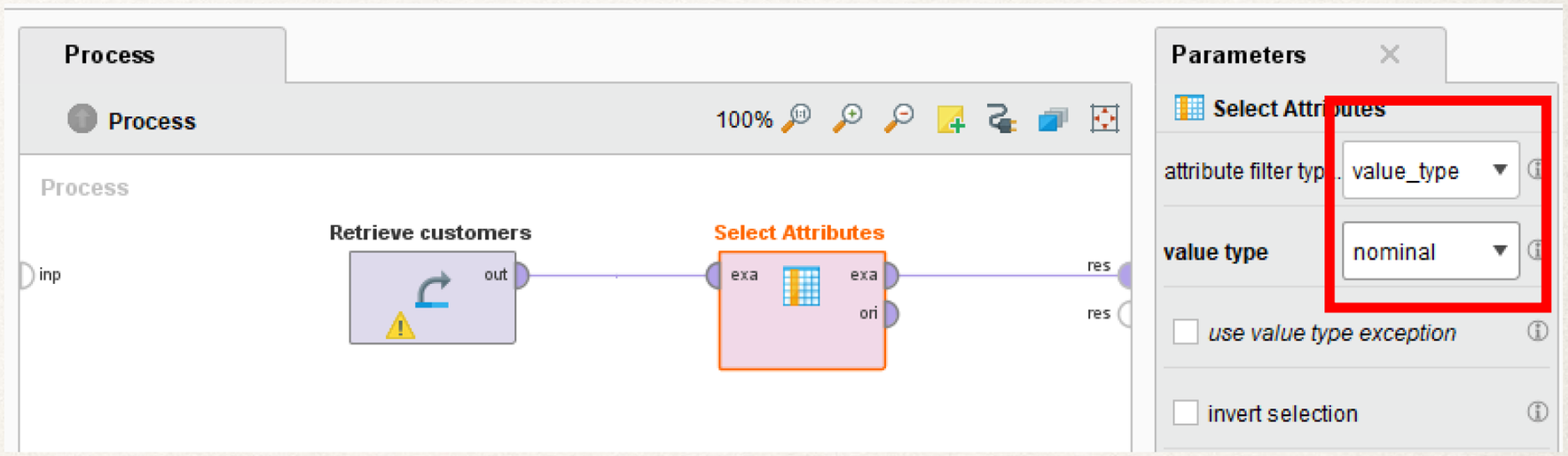
Data Preparation

- 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)

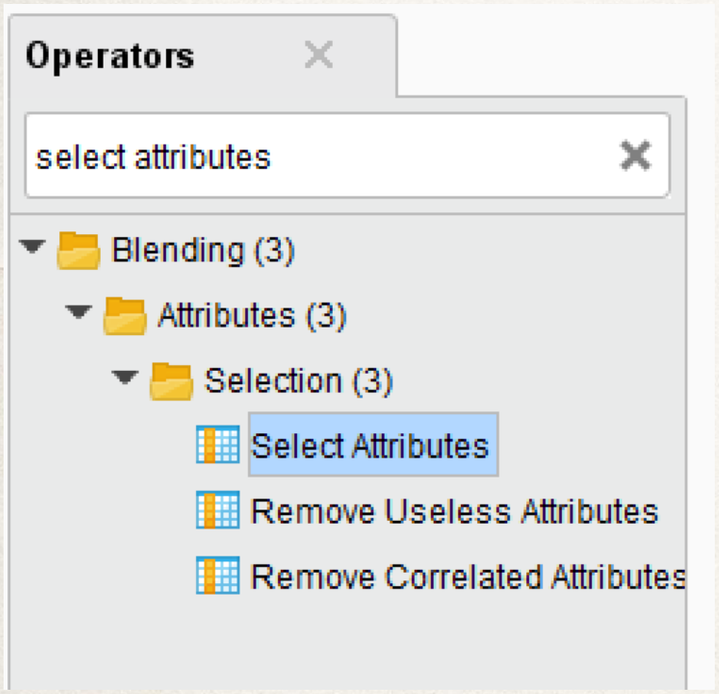
Data | Statistics | Charts

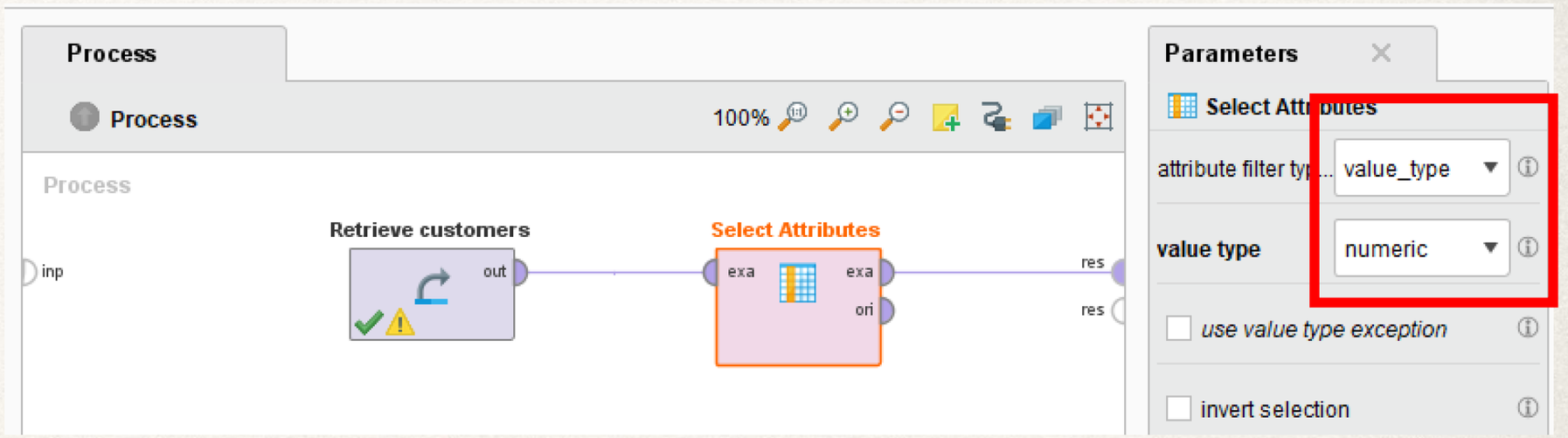
ExampleSet (600 examples, 2 special attributes, 4 regular attributes) Filter (600 / 600 examples): all

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

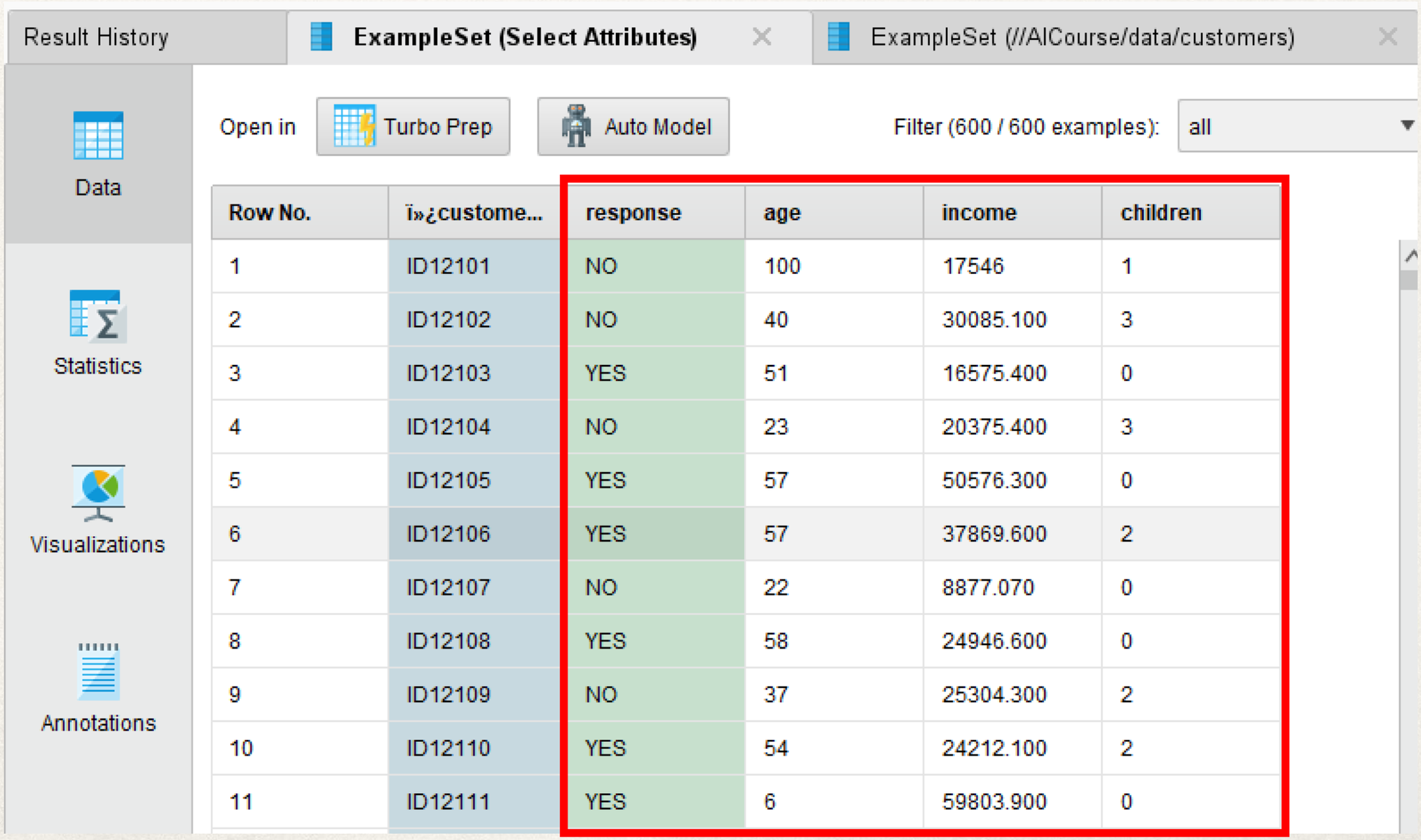
- Choose Select Attributes operator
- attribute filter type = value_type
- value type = numeric





Select Attributes by Type

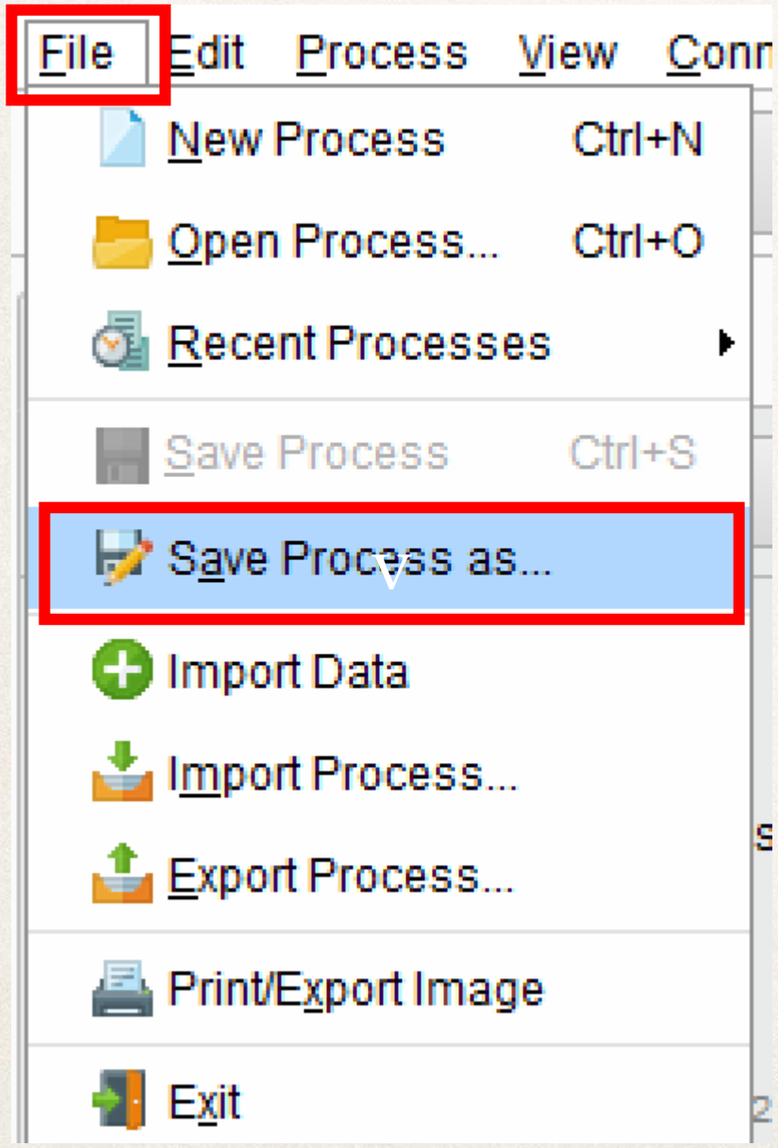
- The result shows only numeric attributes.

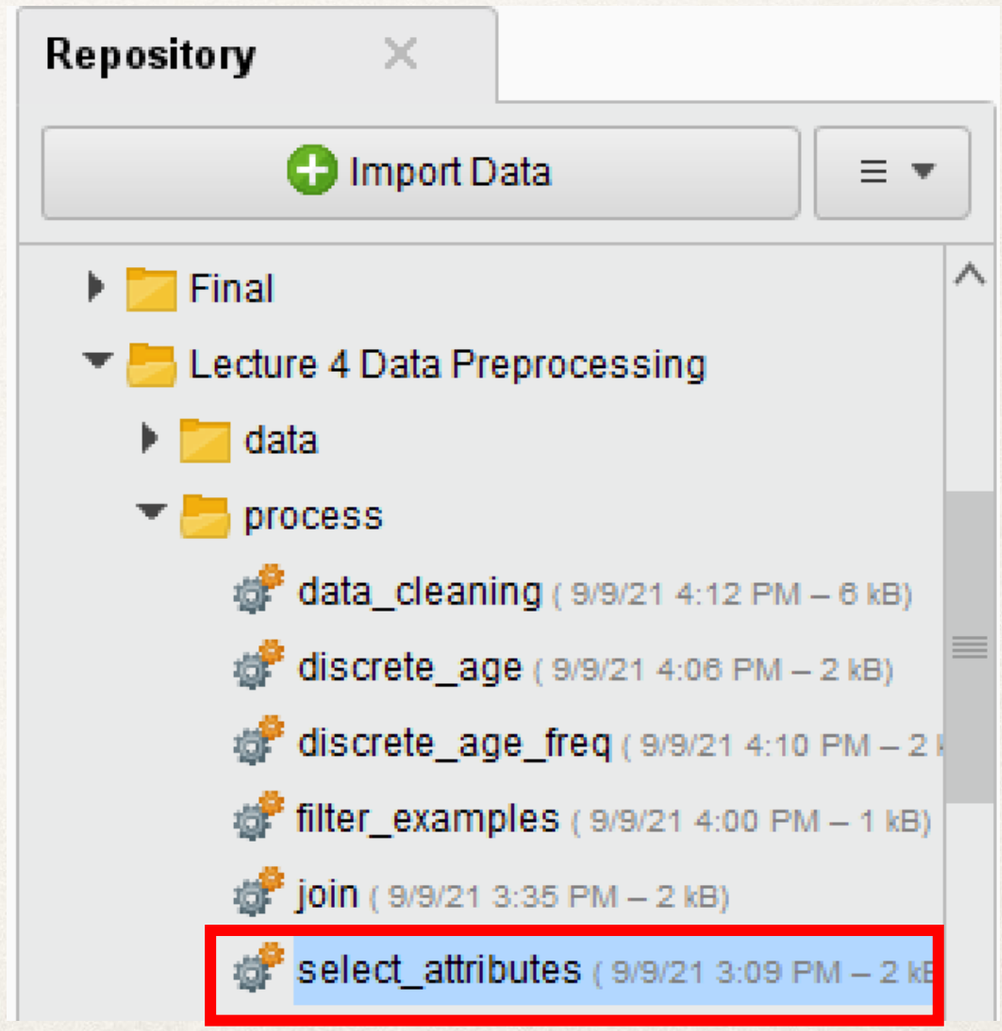


Row No.	ï»¿custome...	response	age	income	children
1	ID12101	NO	100	17546	1
2	ID12102	NO	40	30085.100	3
3	ID12103	YES	51	16575.400	0
4	ID12104	NO	23	20375.400	3
5	ID12105	YES	57	50576.300	0
6	ID12106	YES	57	37869.600	2
7	ID12107	NO	22	8877.070	0
8	ID12108	YES	58	24946.600	0
9	ID12109	NO	37	25304.300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

How to Save a Process into Repository

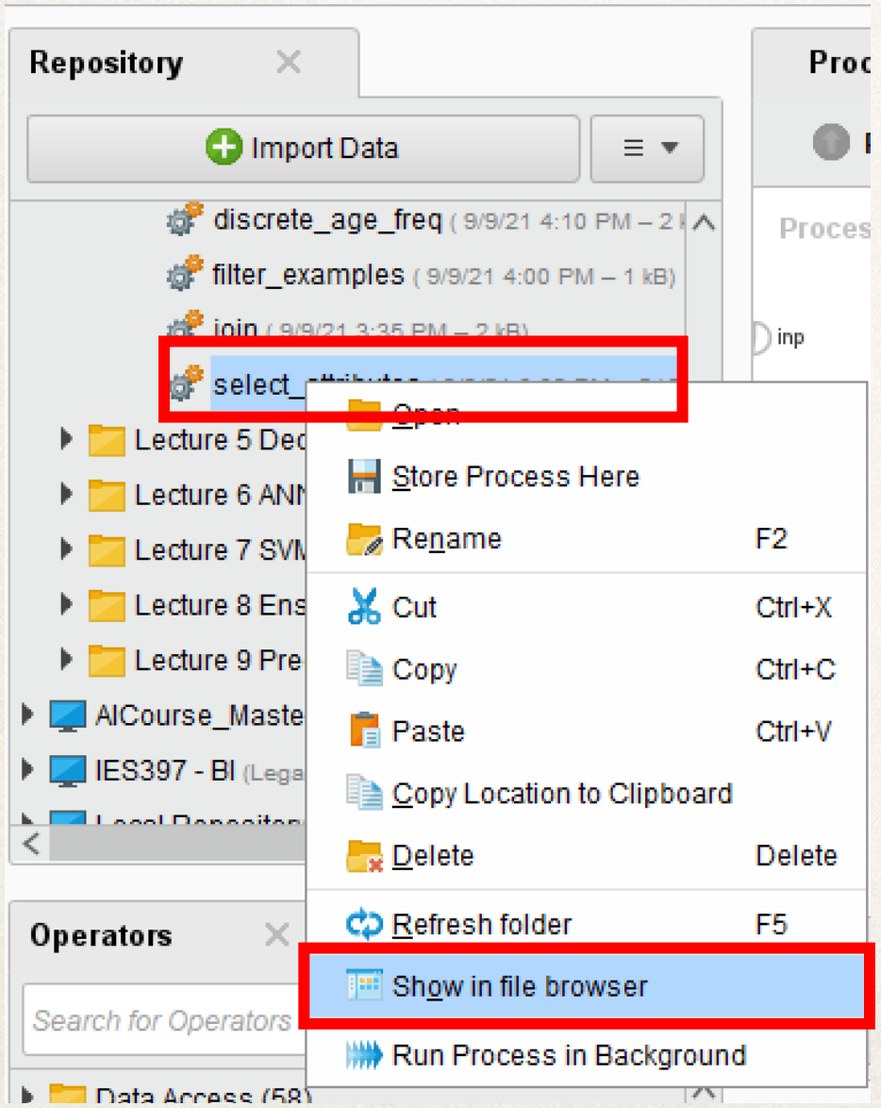
- File > Save Process as > .rmp file is saved in repository folder

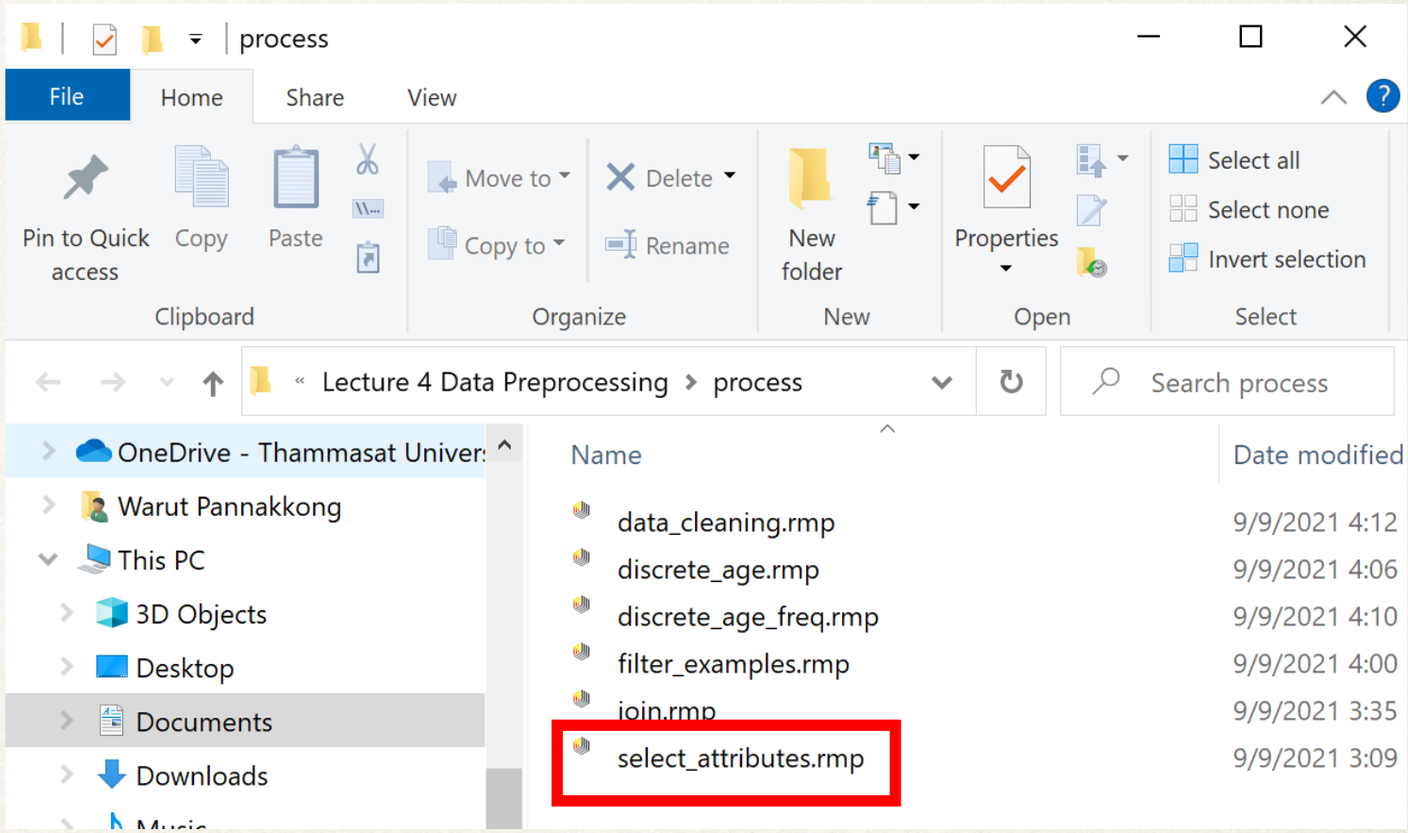




How to Show the Saved Process File (.rmp) in Repository

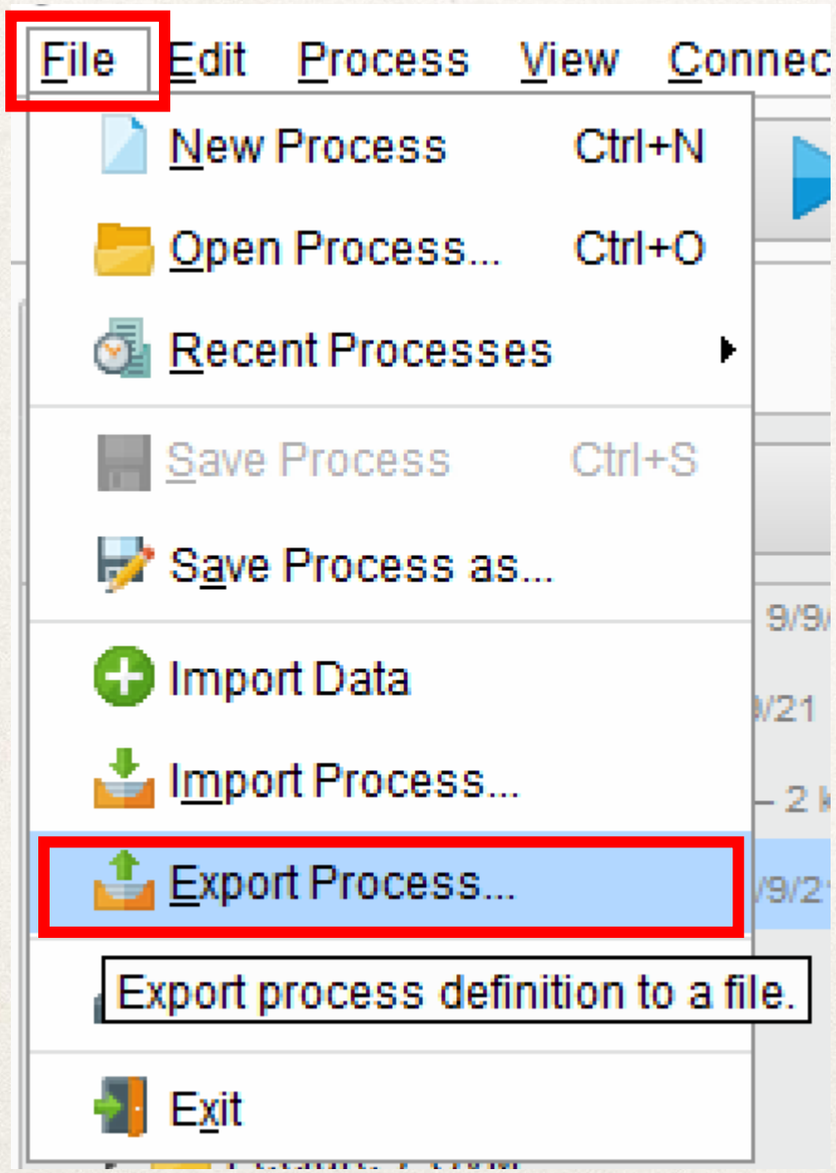
- Right click on the process > Select “Show in file browser”

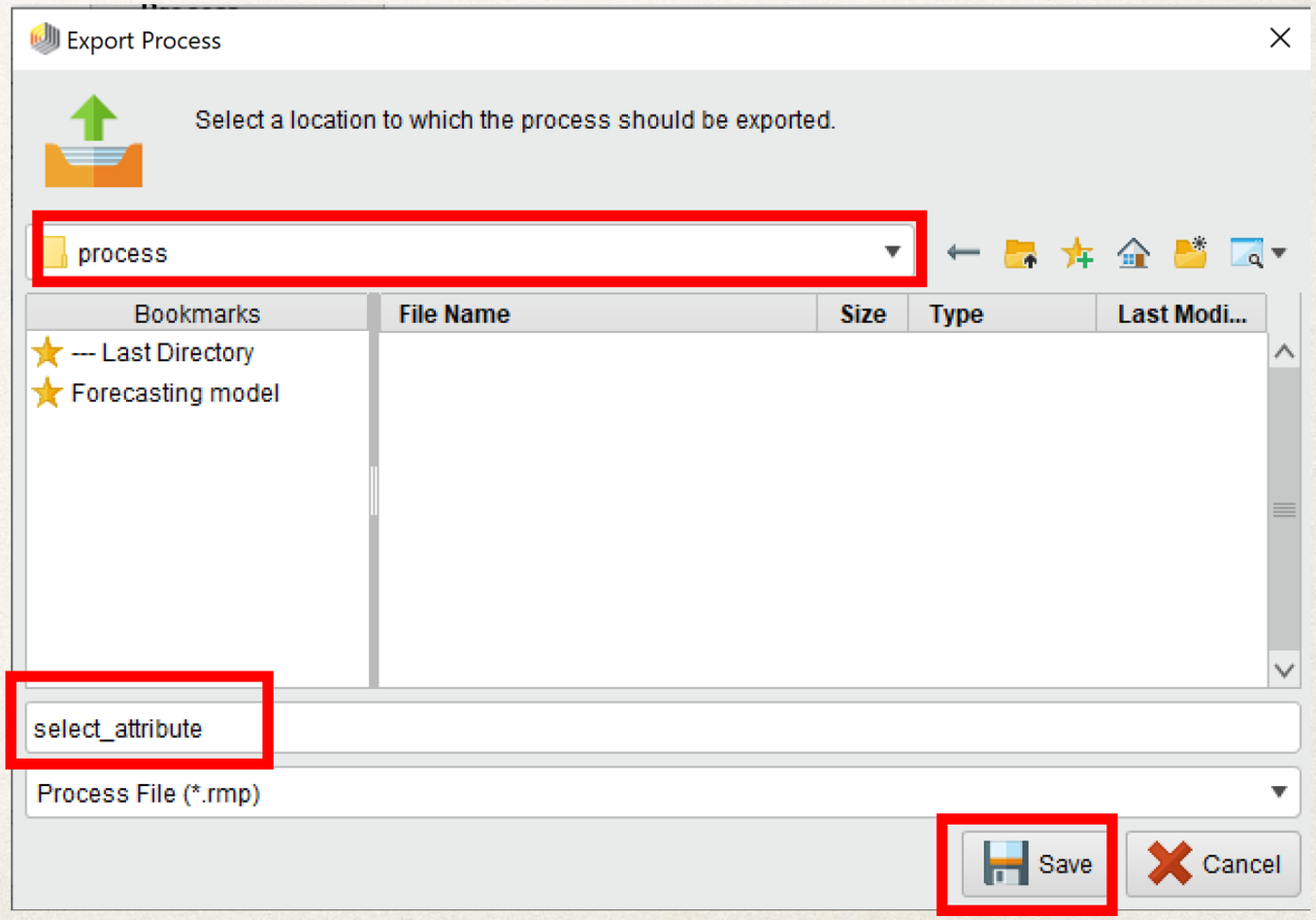




How to Export a Process File (.rmp)

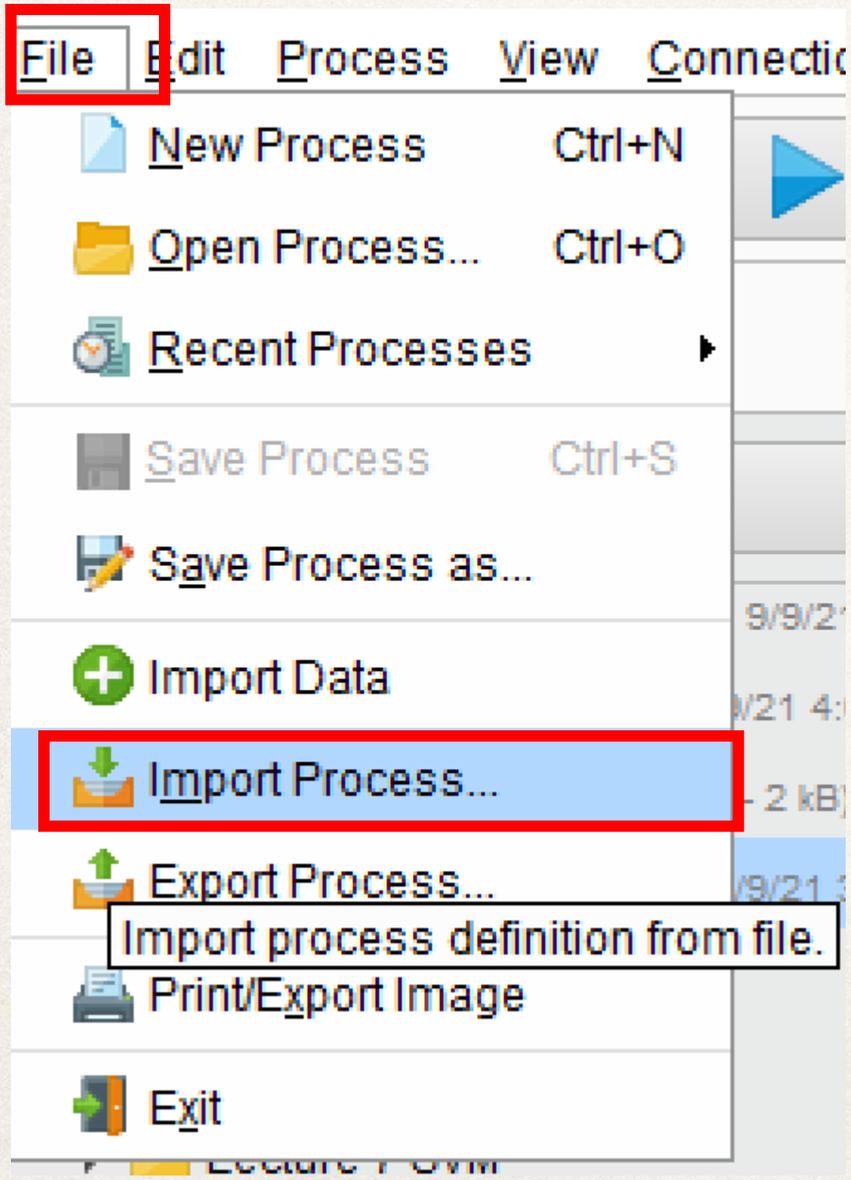
- File > Export Process

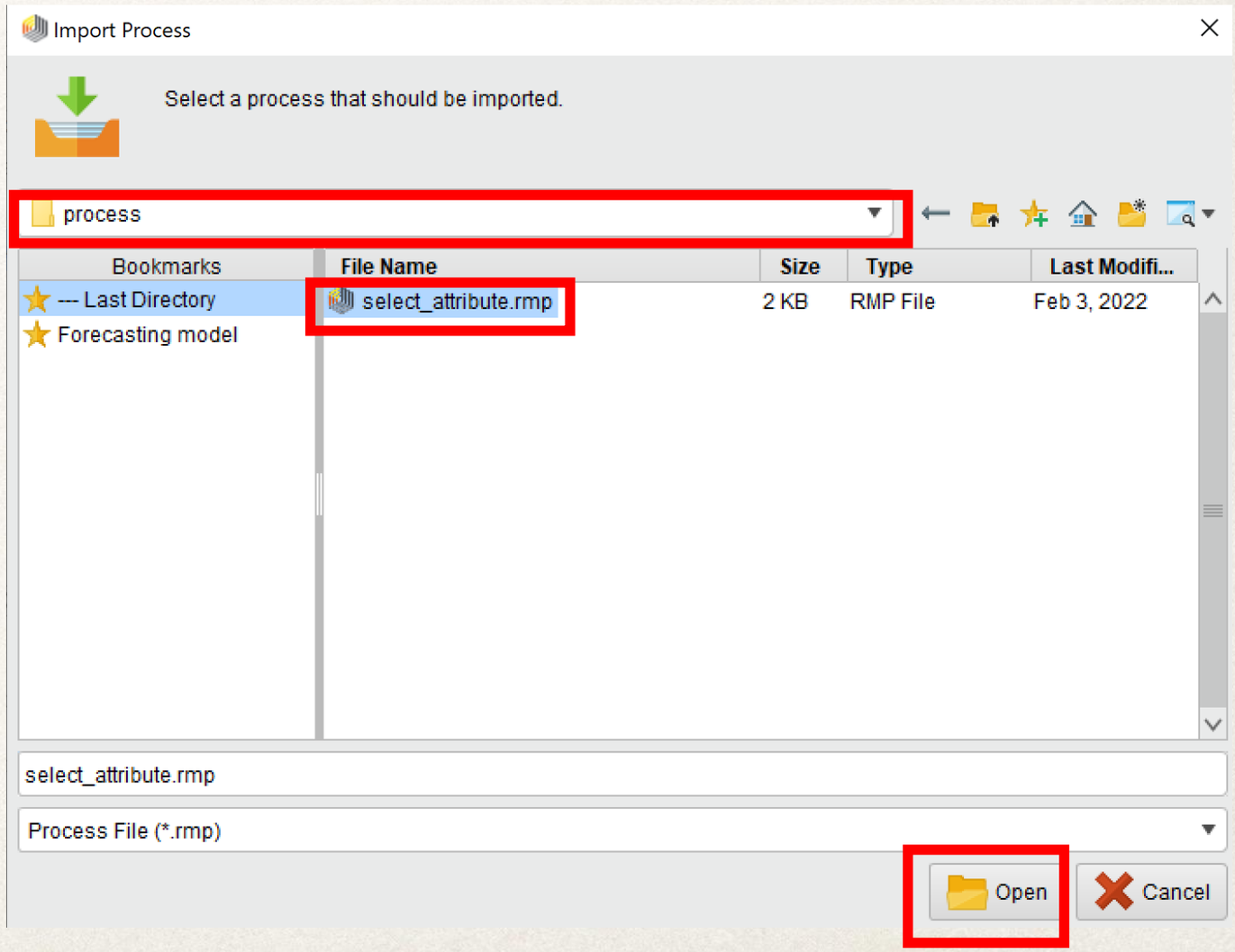




How to Import a Process File (.rmp)

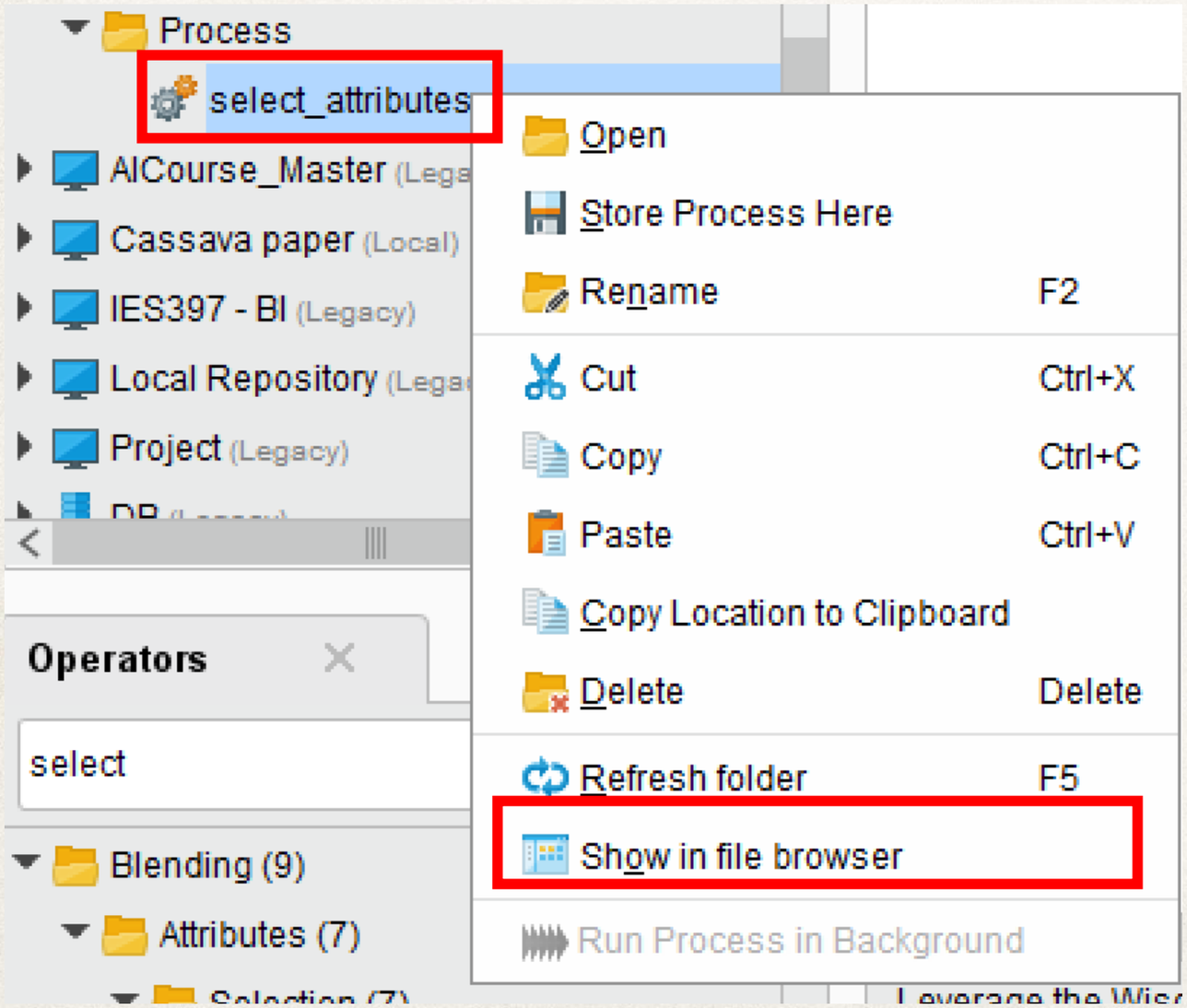
- File > Export Process

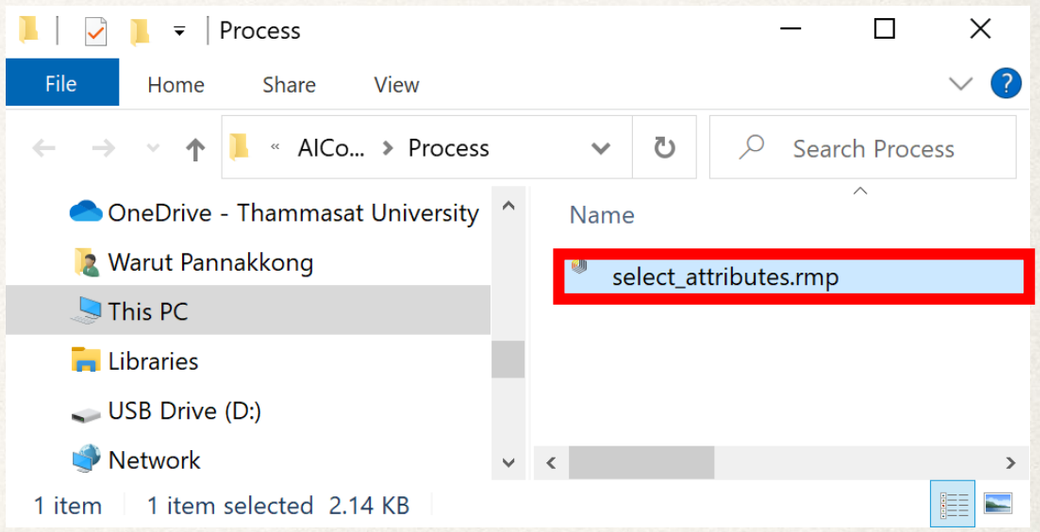




How to Import a Process File (.rmp)

- Browse the file on the PC



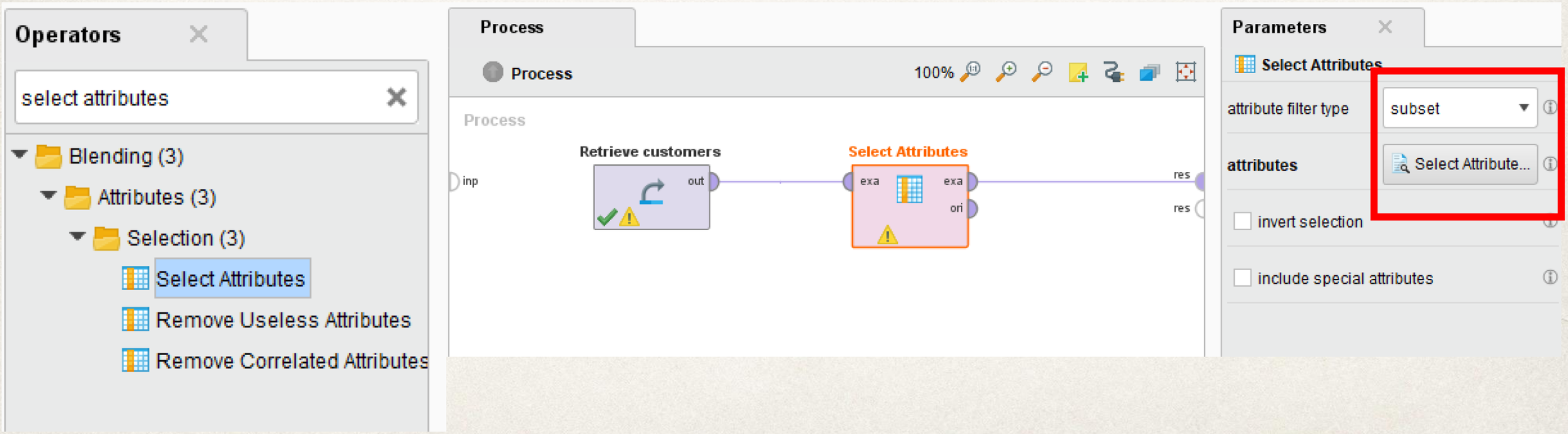


Data Preparation

- **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

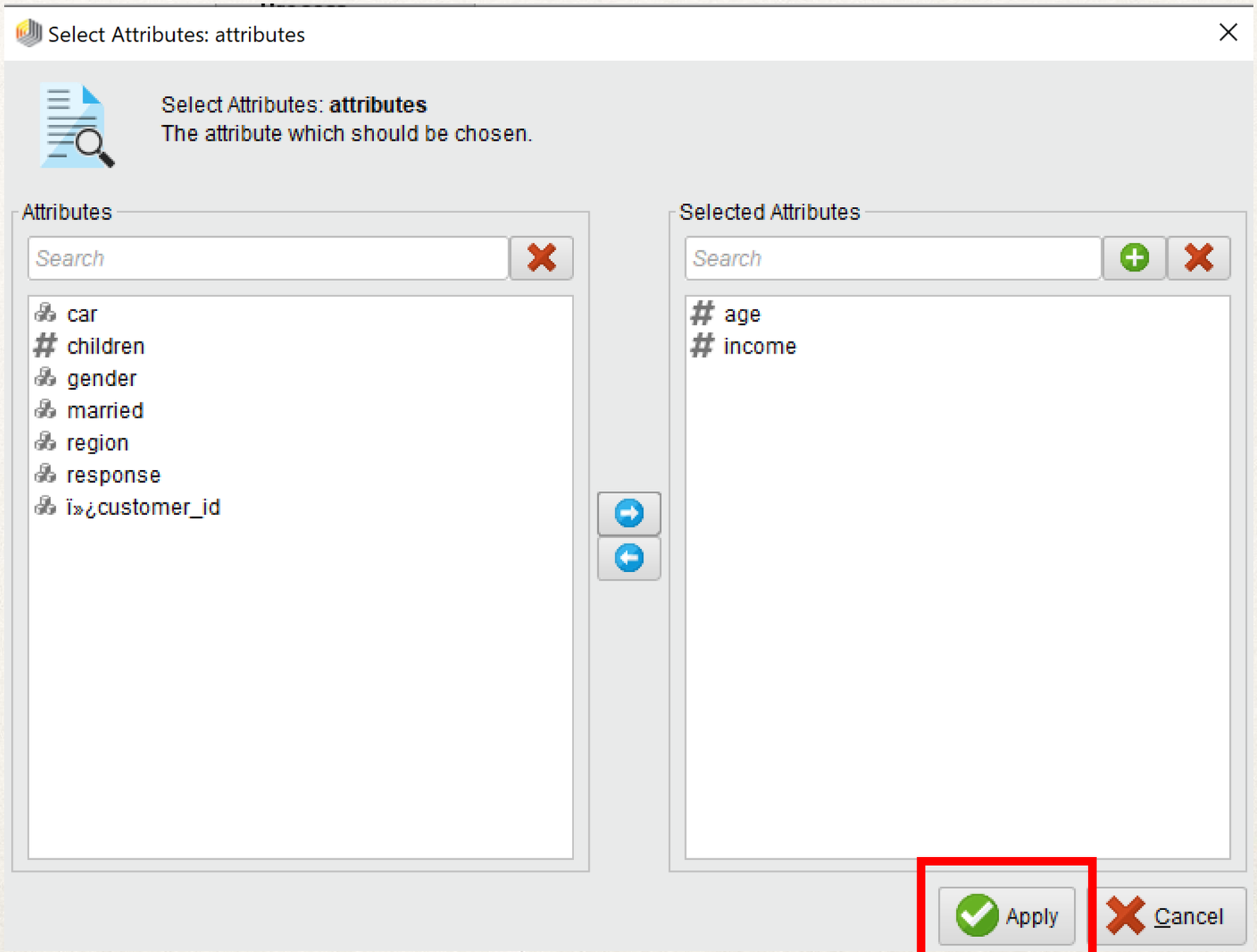
Select Attributes by Name

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



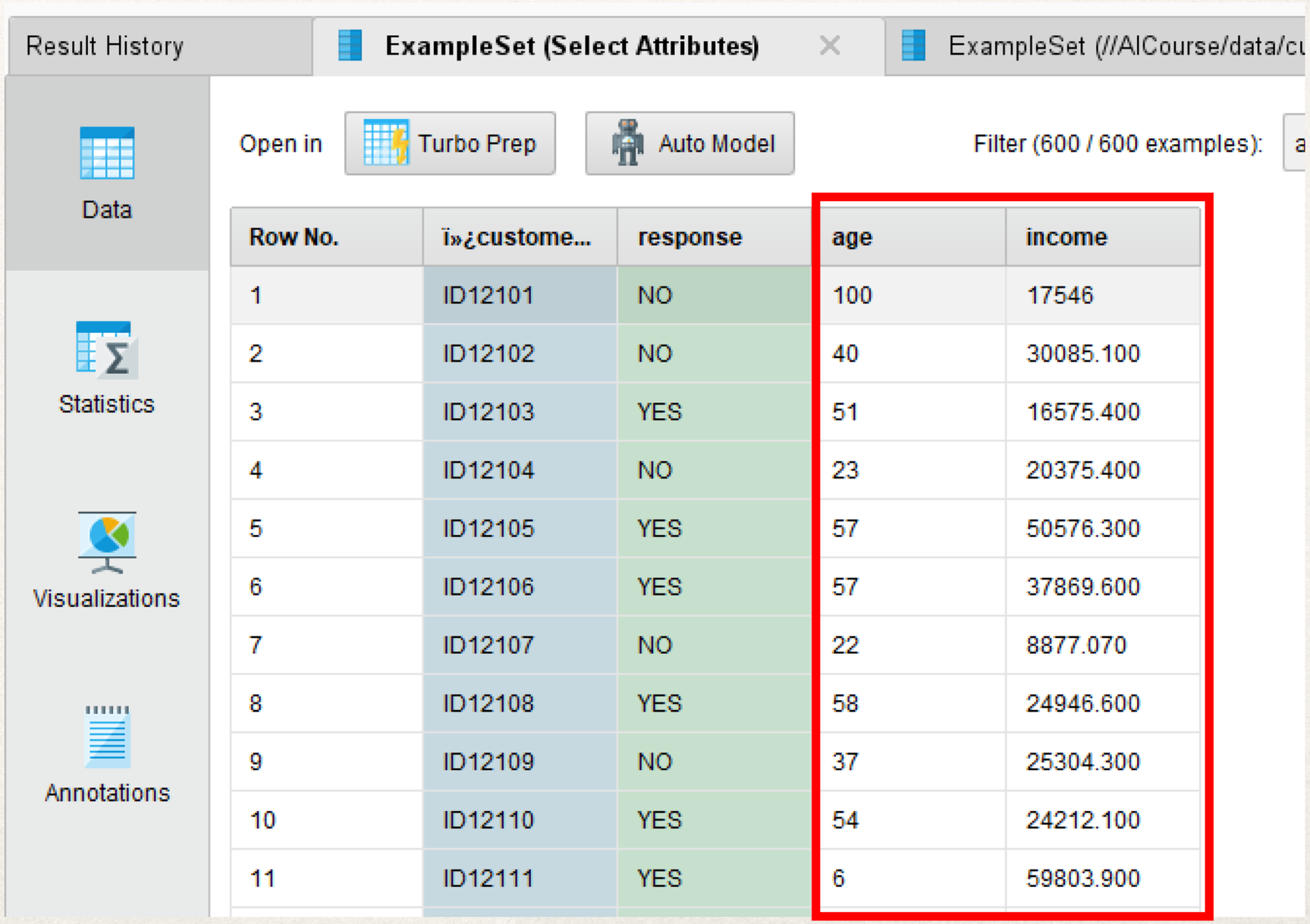
Select Attributes by Name

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



Select Attributes by Name

- The result shows only the selected attributes.

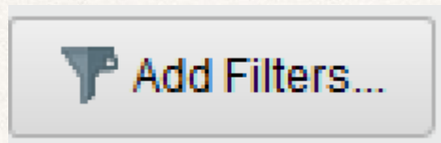


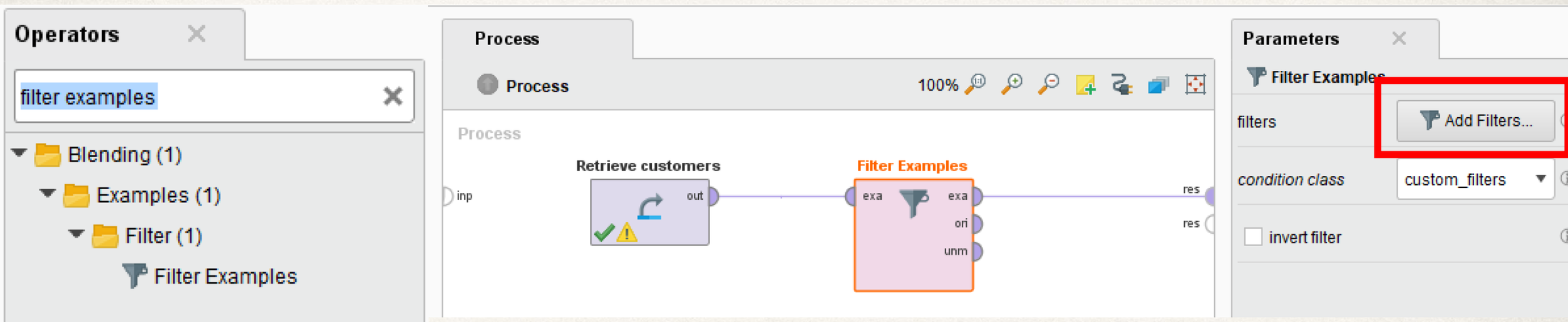
Row No.	ï»¿custome...	response	age	income
1	ID12101	NO	100	17546
2	ID12102	NO	40	30085.100
3	ID12103	YES	51	16575.400
4	ID12104	NO	23	20375.400
5	ID12105	YES	57	50576.300
6	ID12106	YES	57	37869.600
7	ID12107	NO	22	8877.070
8	ID12108	YES	58	24946.600
9	ID12109	NO	37	25304.300
10	ID12110	YES	54	24212.100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

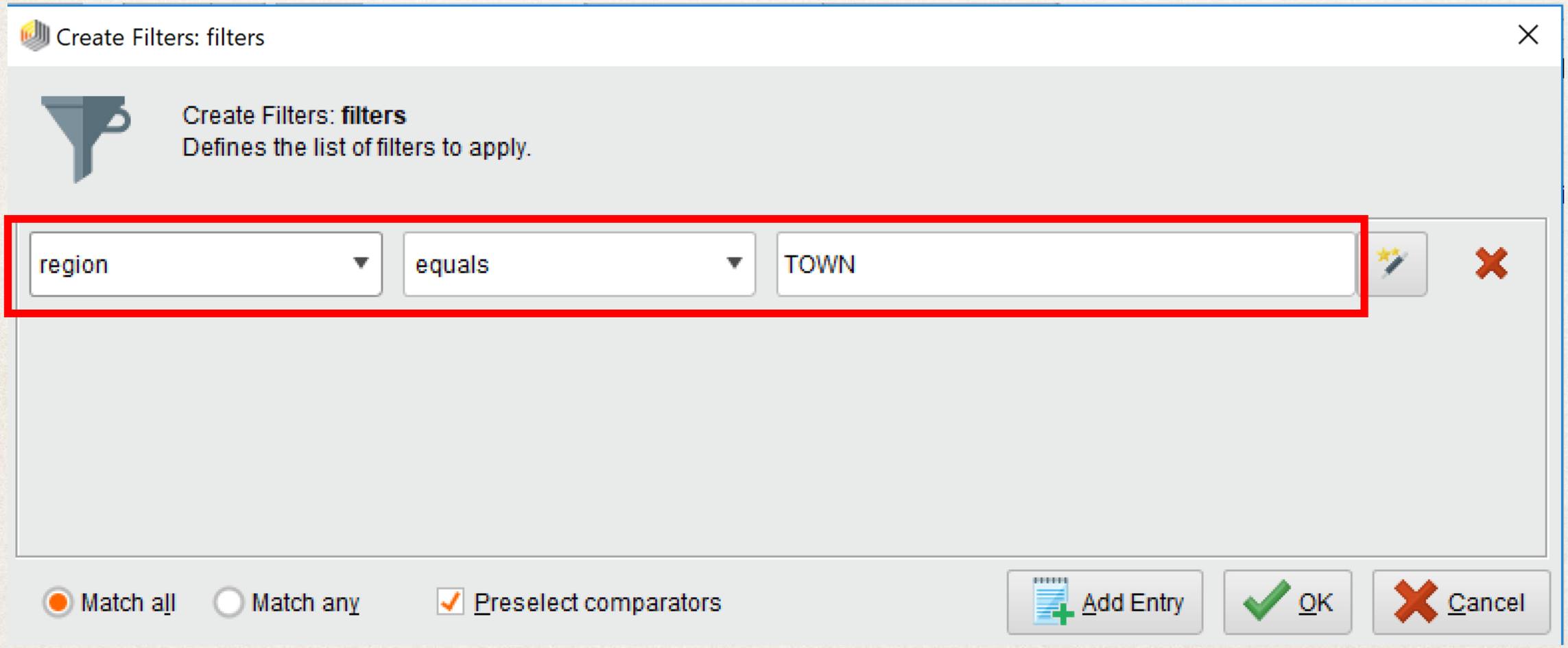
Filter Examples by Conditions

- 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 | ExampleSet (Filter Examples)

Data | Statistics | Charts | Advanced

ExampleSet (173 examples, 2 special attributes, 7 regular attributes)

Filter (173 / 173 examples): all

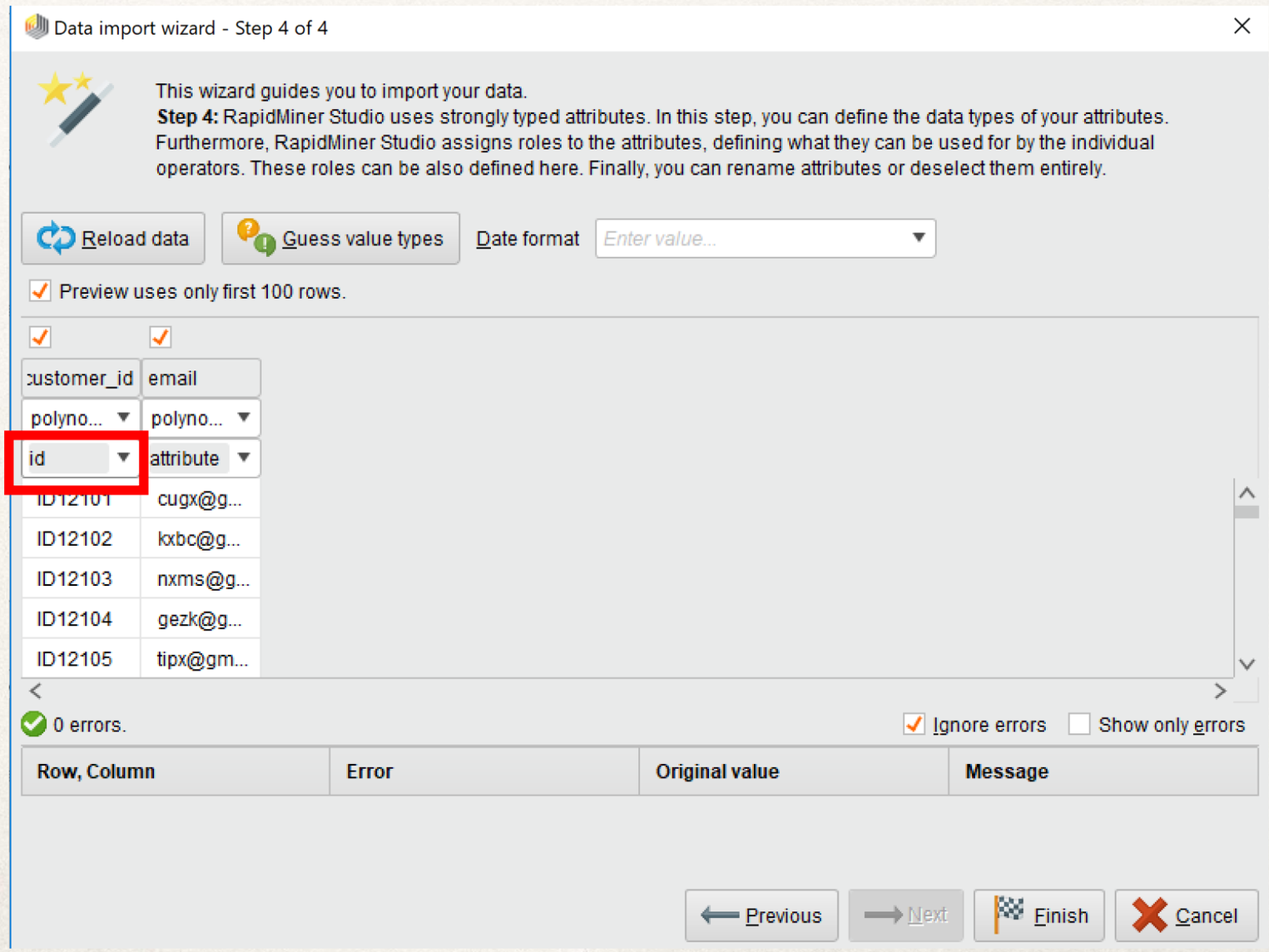
Row No.	customer_id	response	age	gender	region	income	married	children
1	ID12102	NO	40	MALE	TOWN	30085.100	YES	3
2	ID12104	NO	23	FEMALE	TOWN	20375.400	YES	3
3	ID12106	YES	57	WOMAN	TOWN	37869.600	YES	2
4	ID12108	YES	58	MALE	TOWN	24946.600	YES	0
5	ID12110	YES	54	MAN	TOWN	24212.100	YES	2
6	ID12111	YES	6	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

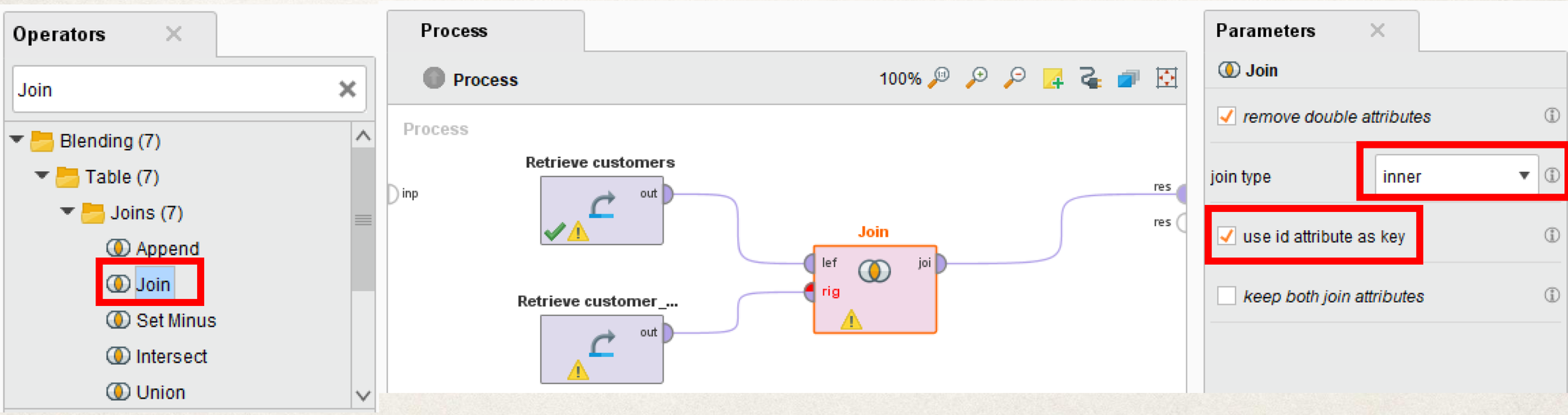
Join Data from Multiple Sources

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



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)

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

Data

Statistics

Visualizations

Annotations

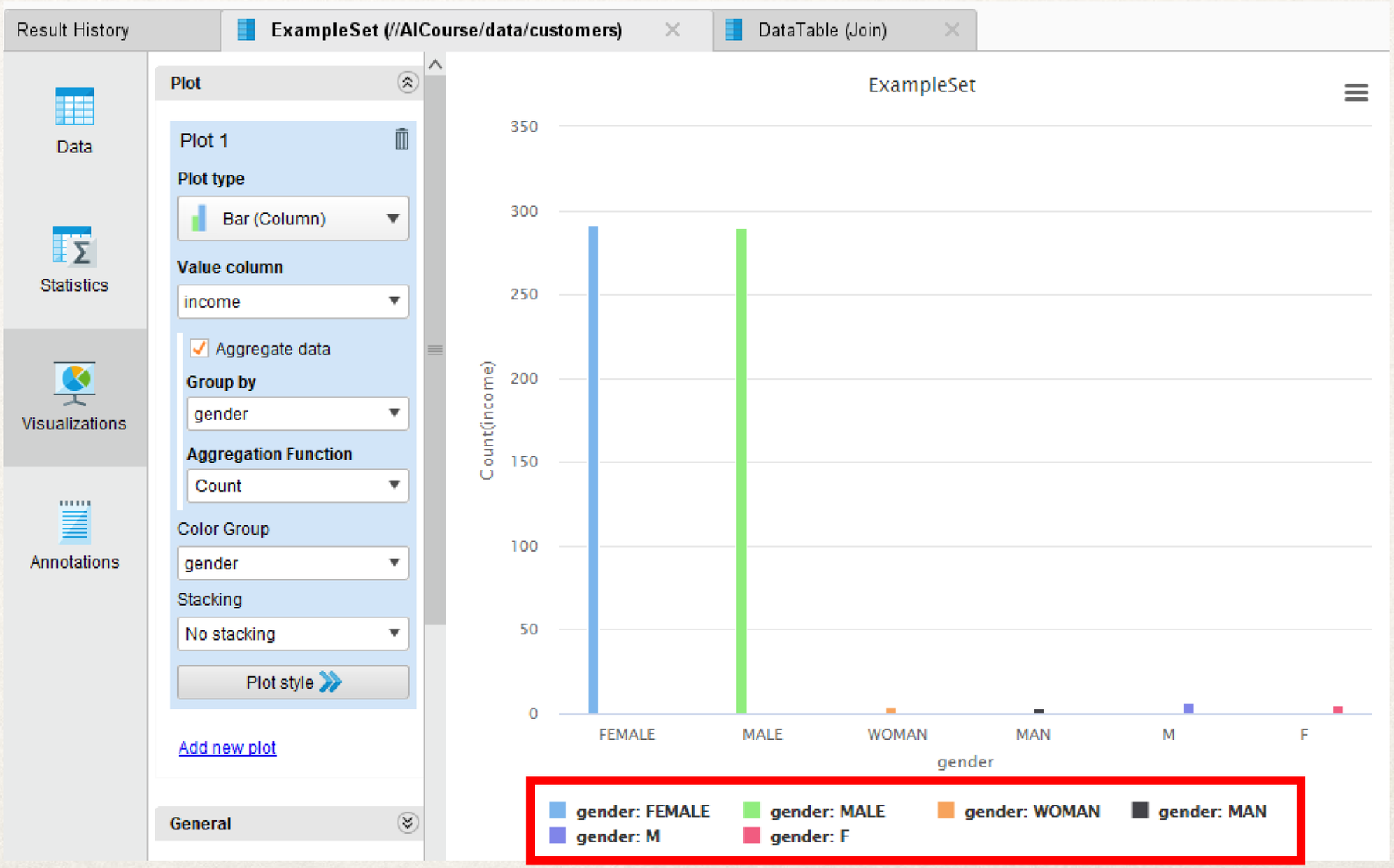
Row No.	ï»¿custome...	response	email	age	gender	region	income	married
1	ID12101	NO	cugx@gmail....	100	FEMALE	?	17546	NO
2	ID12102	NO	kxbc@gmail....	40	MALE	TOWN	30085.100	YES
3	ID12103	YES	nxms@gmail...	51	FEMALE	INNER_CITY	16575.400	YES
4	ID12104	NO	gezk@gmail....	23	FEMALE	TOWN	20375.400	YES
5	ID12105	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

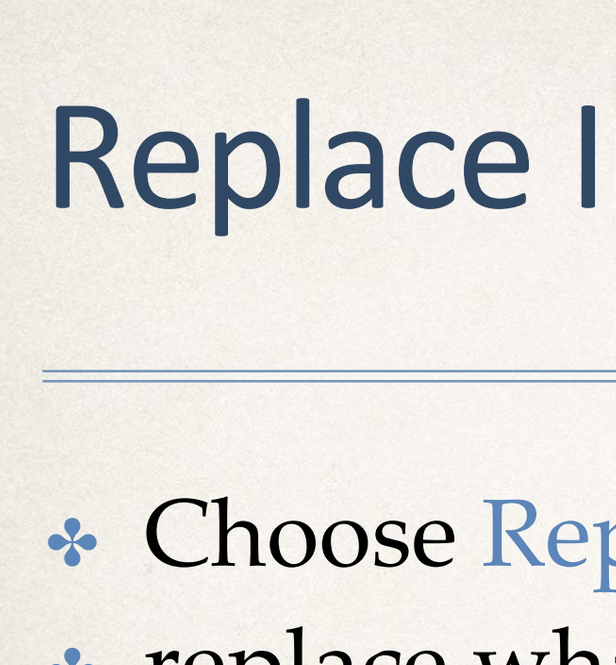
Replace Inconsistent Data

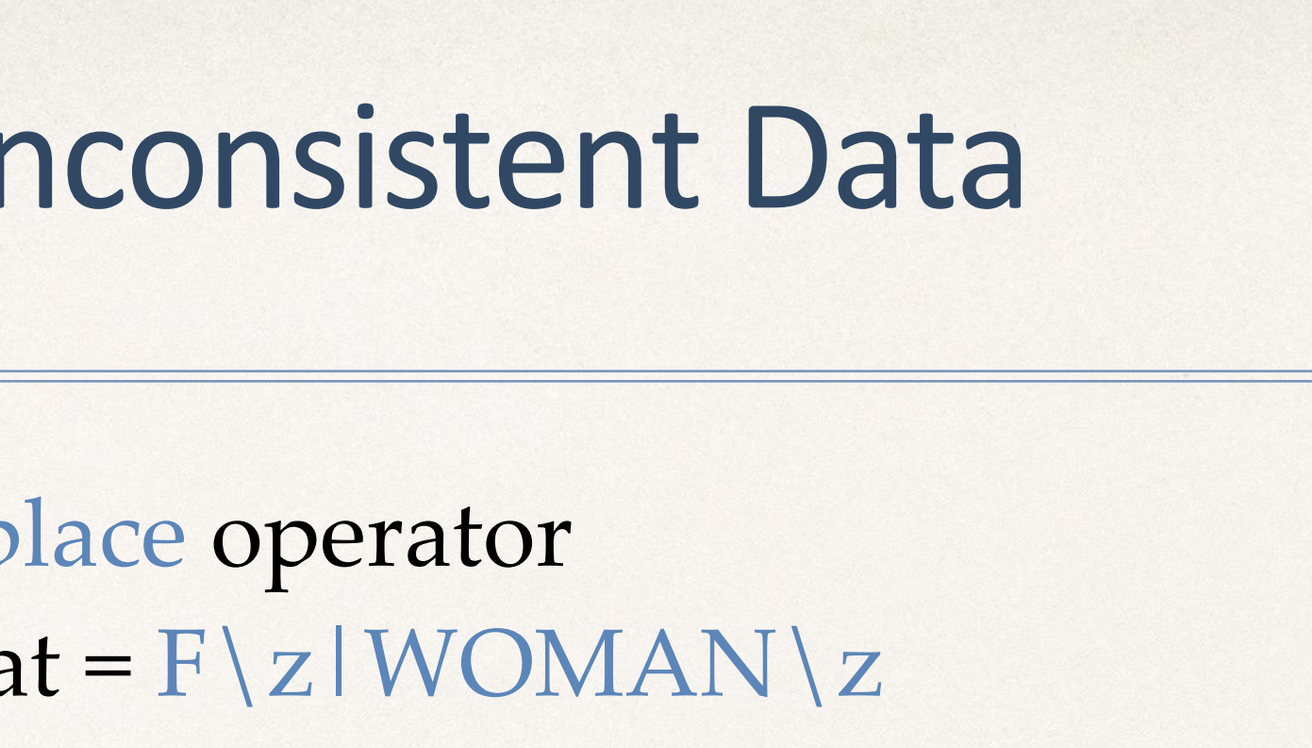
- FEMALE, WOMAN, and F
- MALE, MAN, and M

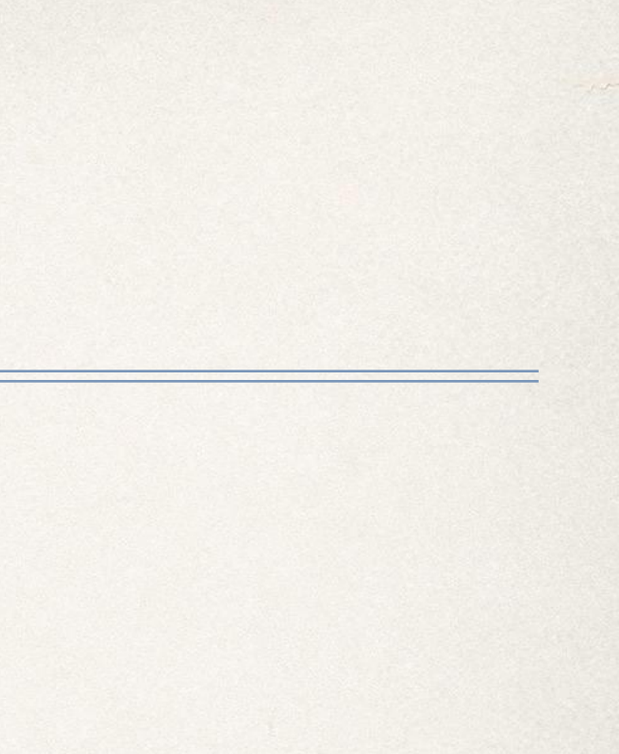


Replace Inconsistent Data

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

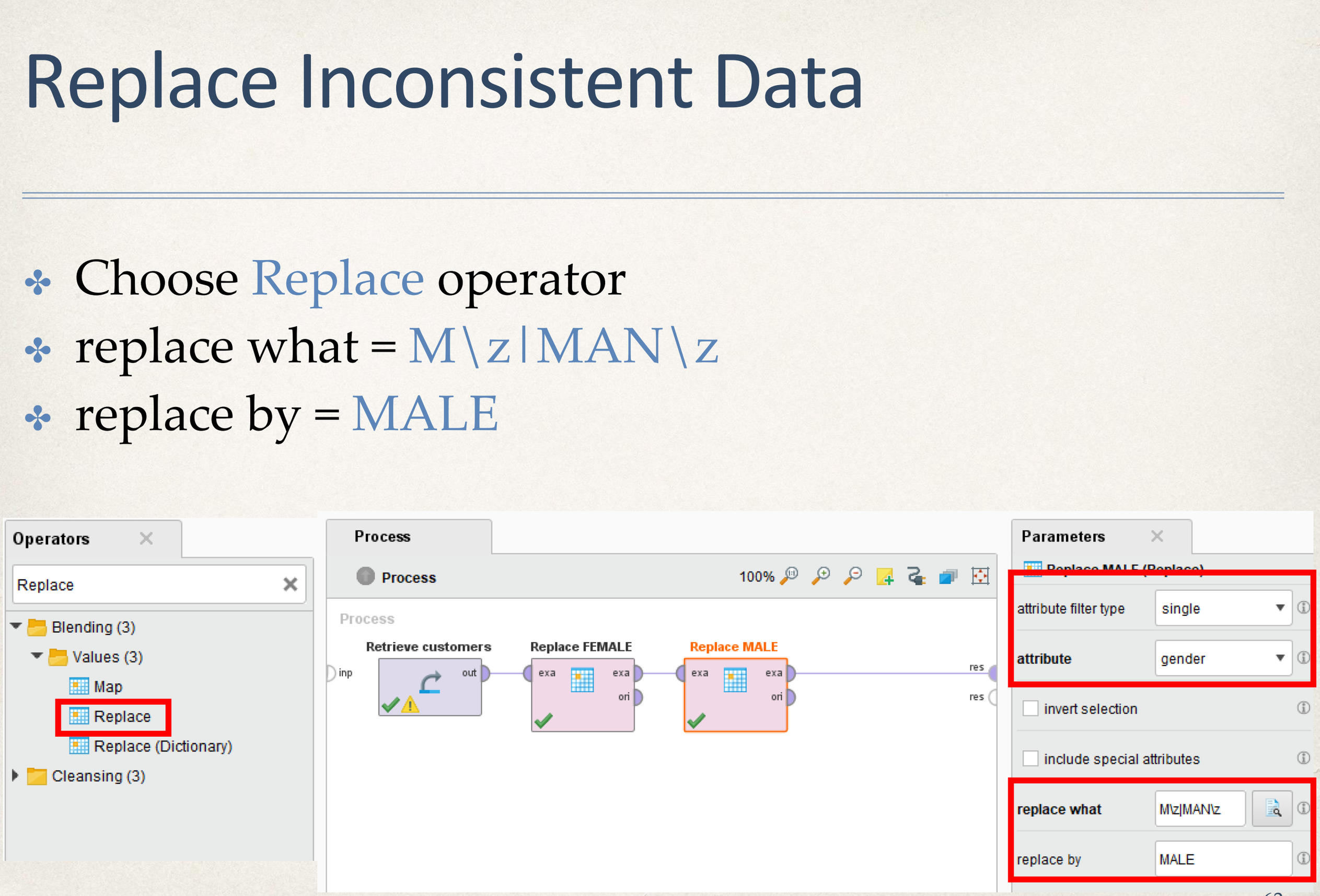






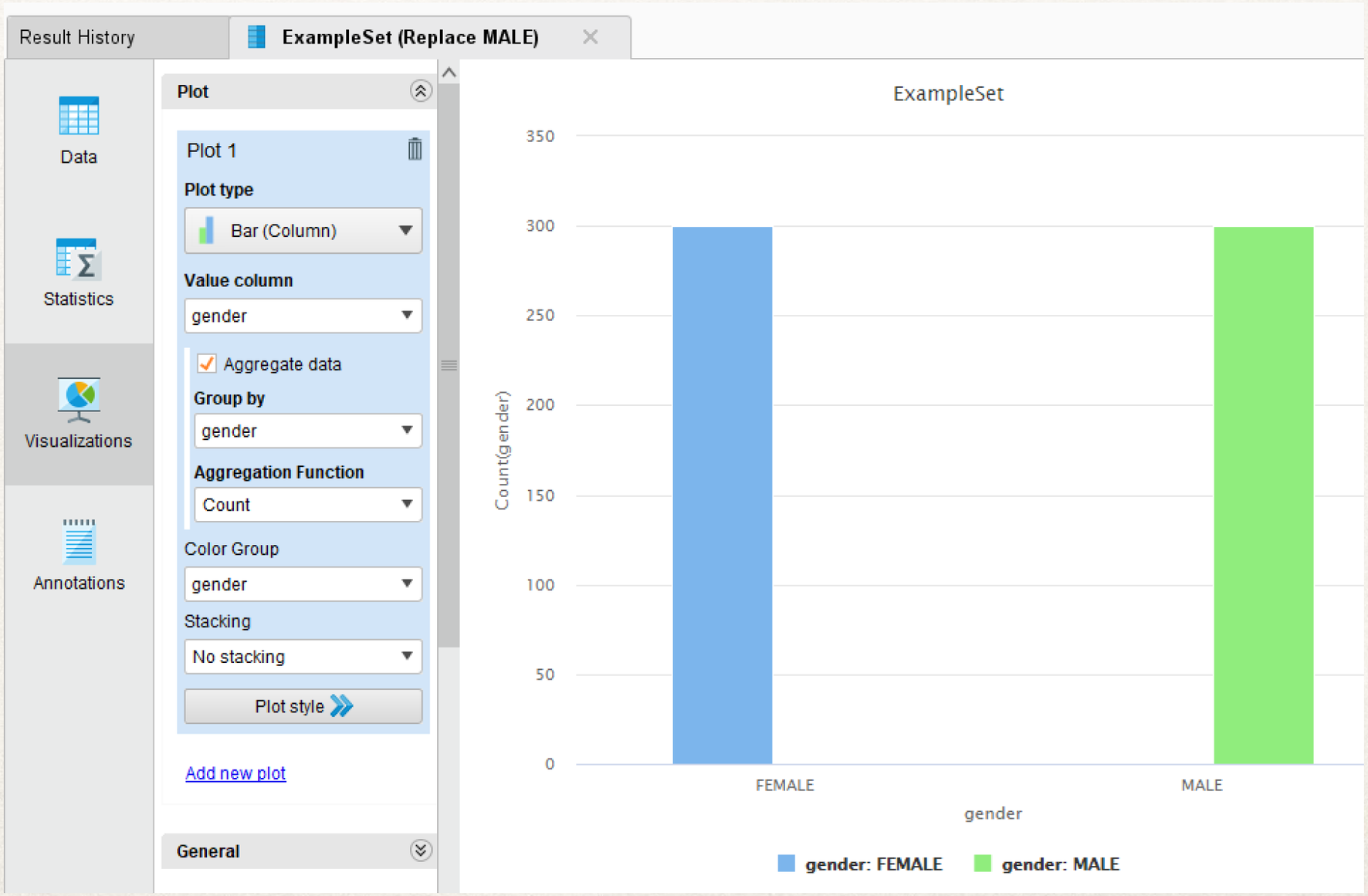
Replace Inconsistent Data

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



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

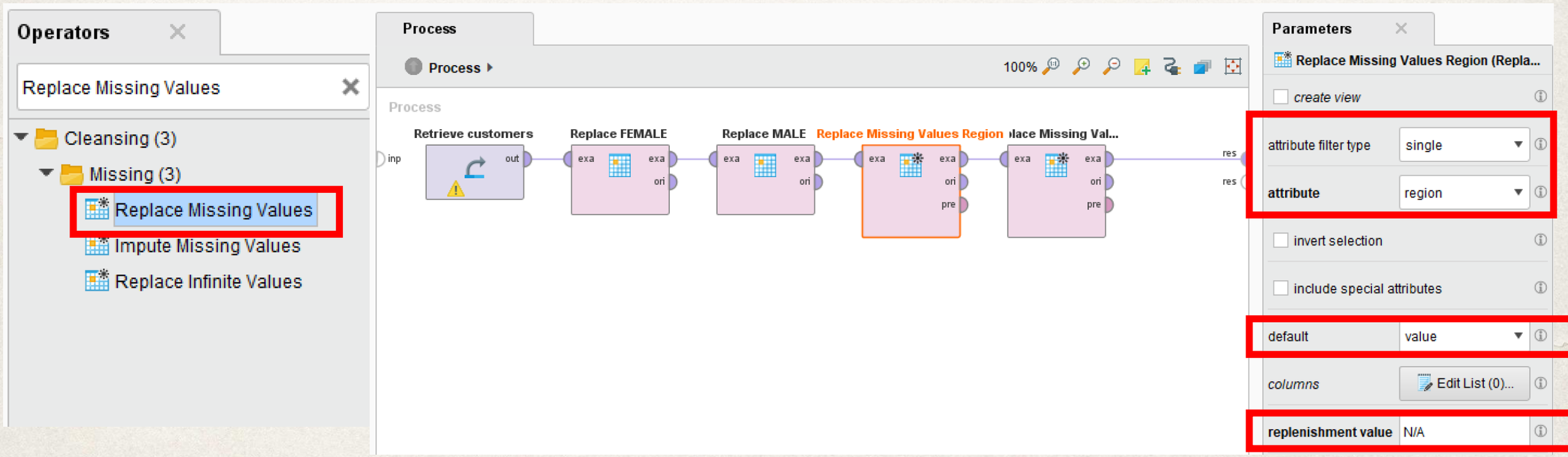
Row No.	custom...	response	age	gender	region	income	married	children	car
1	ID12101	NO	48	FEMALE	?	17546	NO	1	NO
2	ID12112	YES	52	FEMALE	?	26658.800	NO	0	YES
3	ID12123	YES	54	MALE	?	38446.600	YES	0	NO
4	ID12133	YES	45	MALE	?	23443.200	YES	1	YES
5	ID12138	NO	36	FEMALE	RURAL	13381	NO	?	YES
6	ID12144	NO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

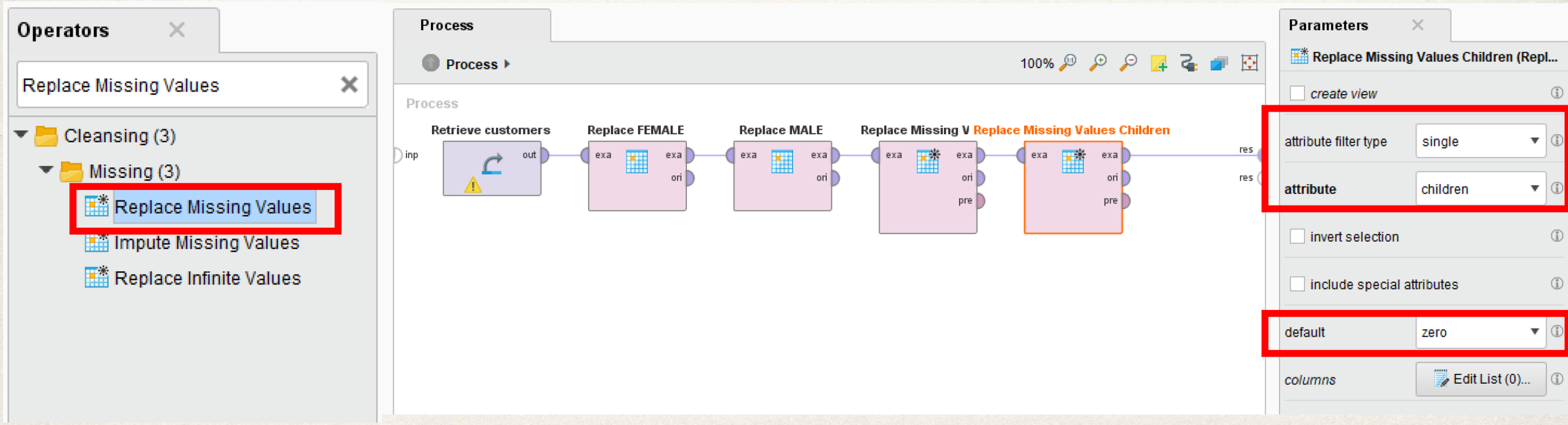
Replace Missing Data

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



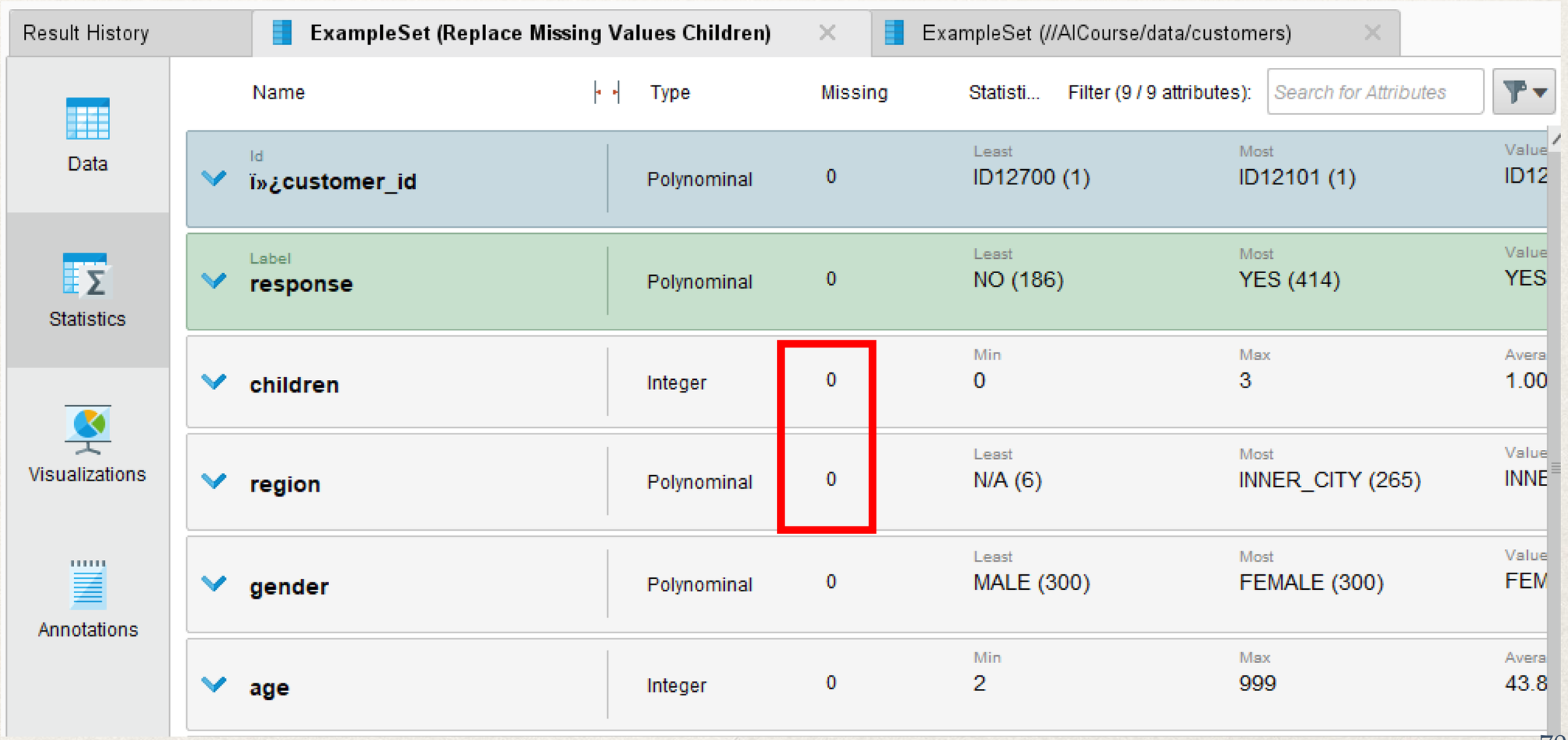
Replace Missing Data

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



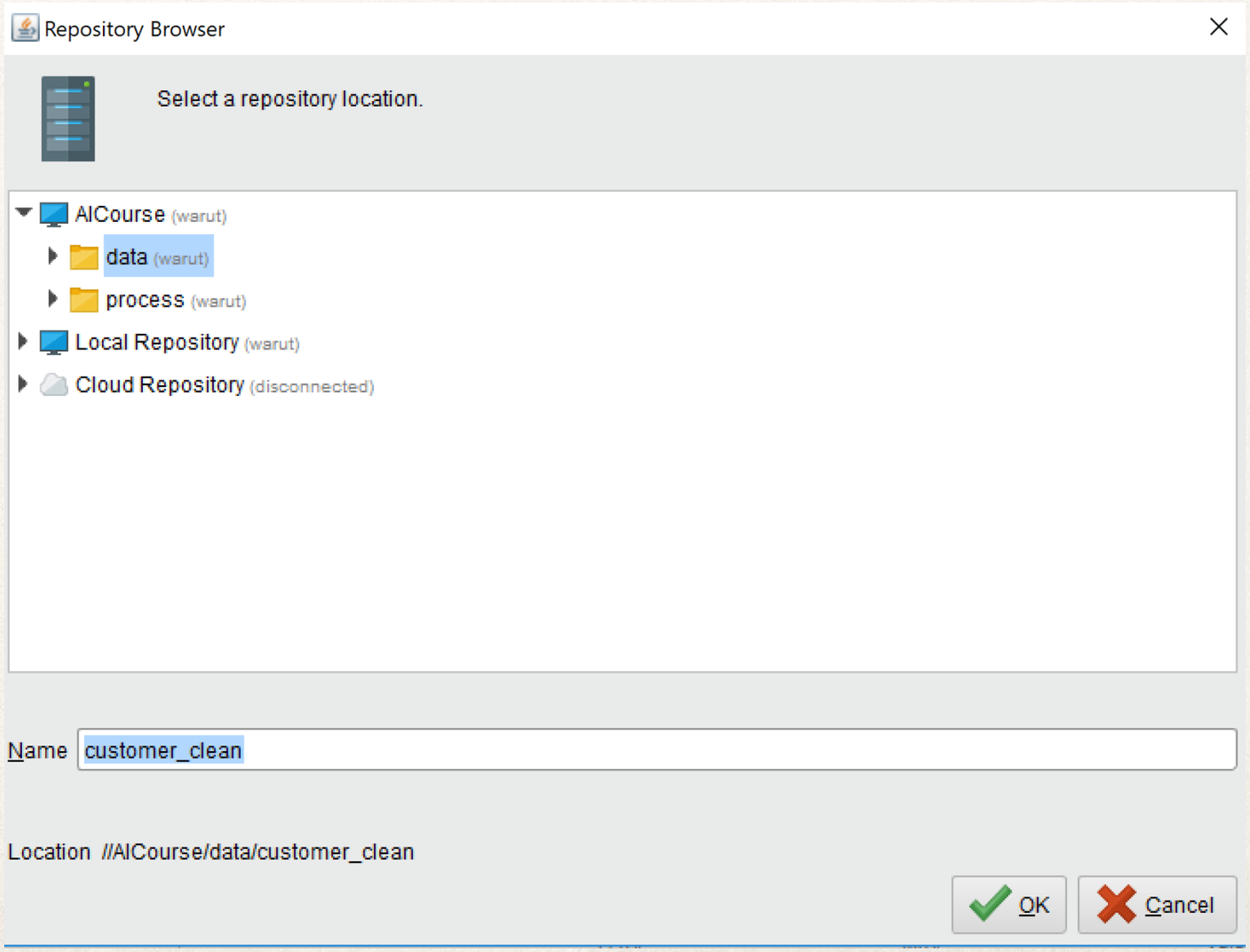
Replace Missing Data

- No missing data for region and children attributes



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 used defined conditions.

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

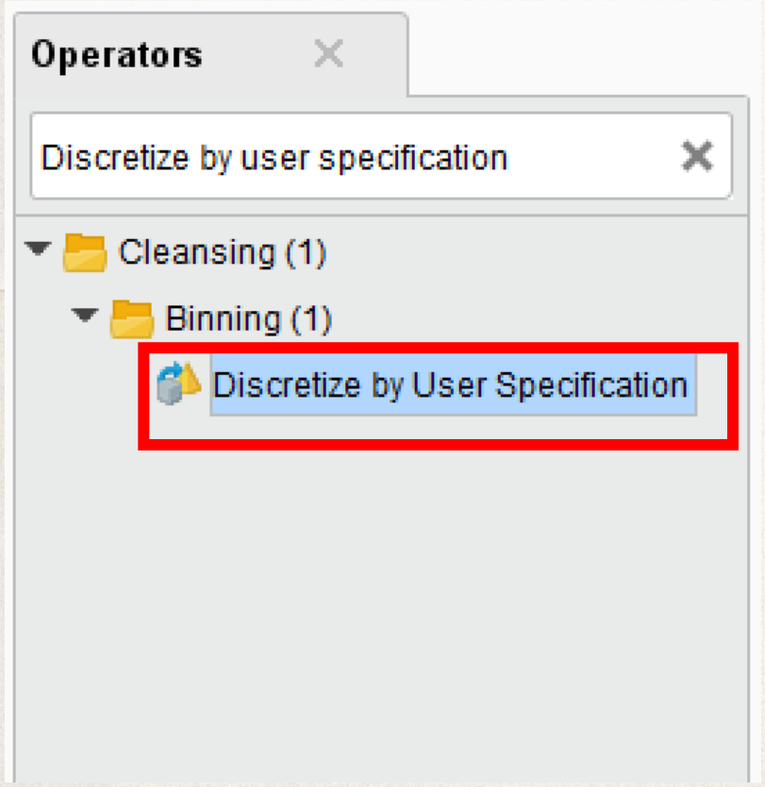
Conditions:

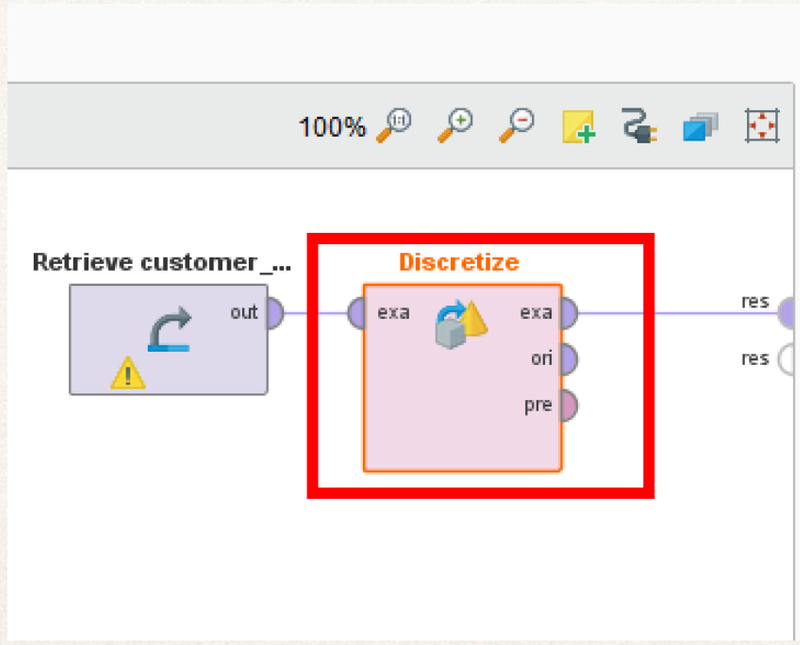
- 18 ≤ Financially Dependent (FD)
- > 18 and ≤ 30 Young Professional (YP)
- > 30 and ≤ 45 Married Professional (MP)
- > 45 and ≤ 60 Empty Nesters (EN)
- > 60 and ≤ 100 Retired ®

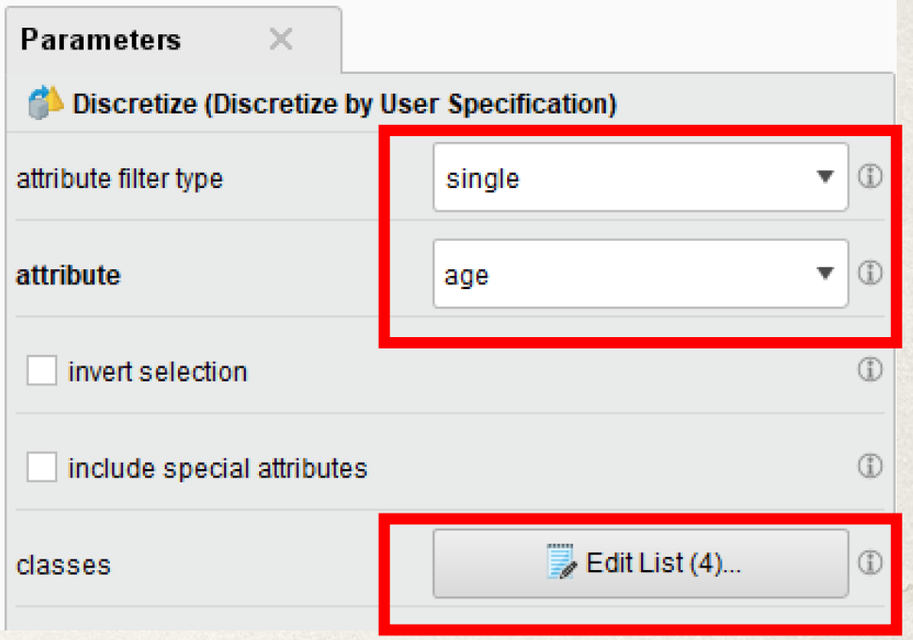
No.	Age
1	FD
2	YP
3	YP
4	YP
5	YP
6	MP
7	MP
8	MP
9	EN

Discretization: User Defined

- Choose Discretize by User Specification operator

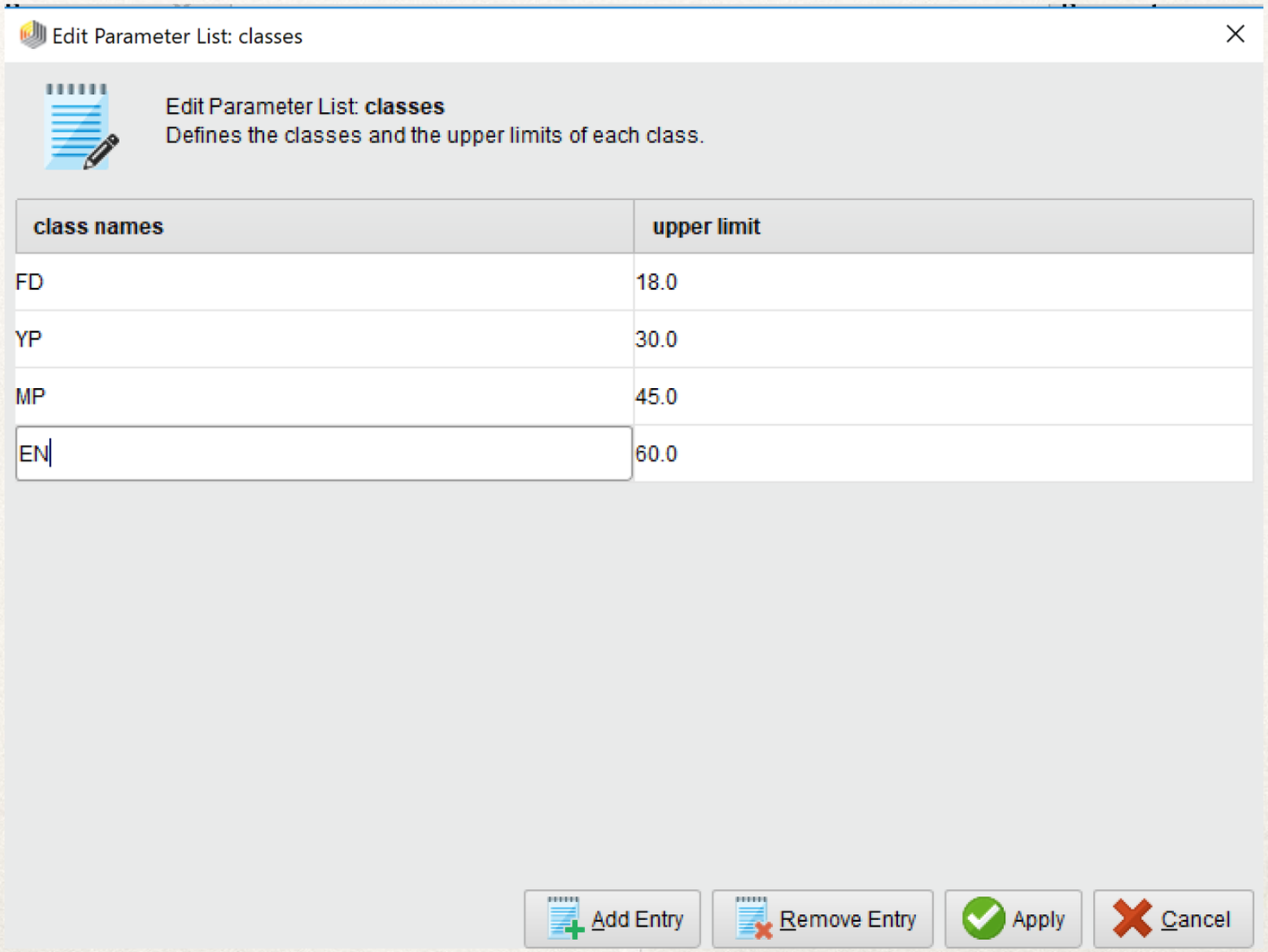






Discretization: User Defined

- Set the conditions based on user defined conditions



Discretization: User Defined

- The result after transformation

Result History | ExampleSet (Discretize)

Data | Statistics | Charts | Advanced Charts

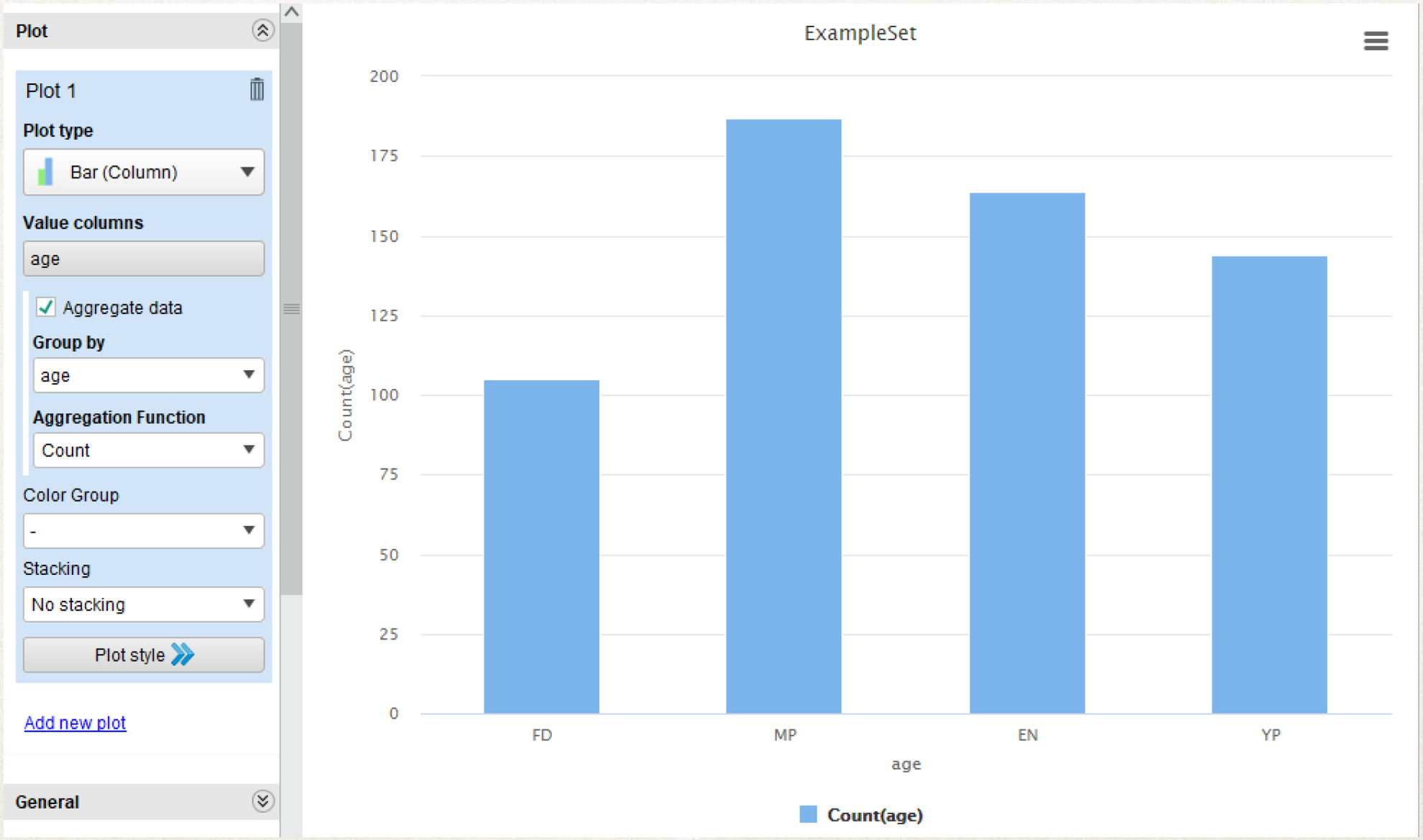
ExampleSet (600 examples, 2 special attributes, 7 regular attributes)

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

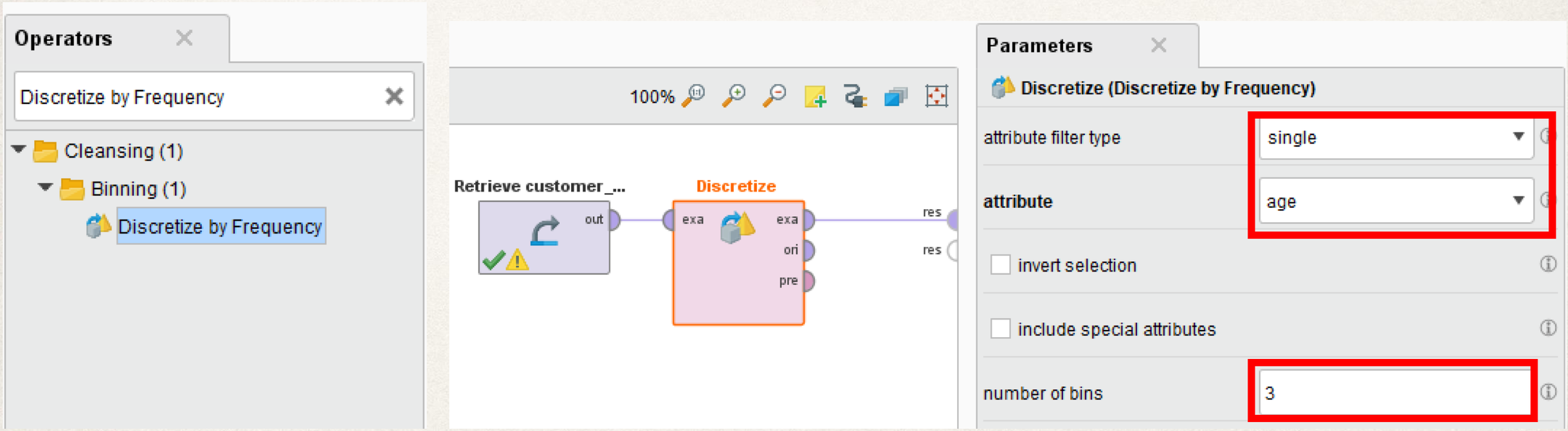


Data Preparation

- 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



Discretization: Equal Frequency

- The result after transformation

Result History | ExampleSet (Discretize)

Data | Statistics | Visualizations | Annotations

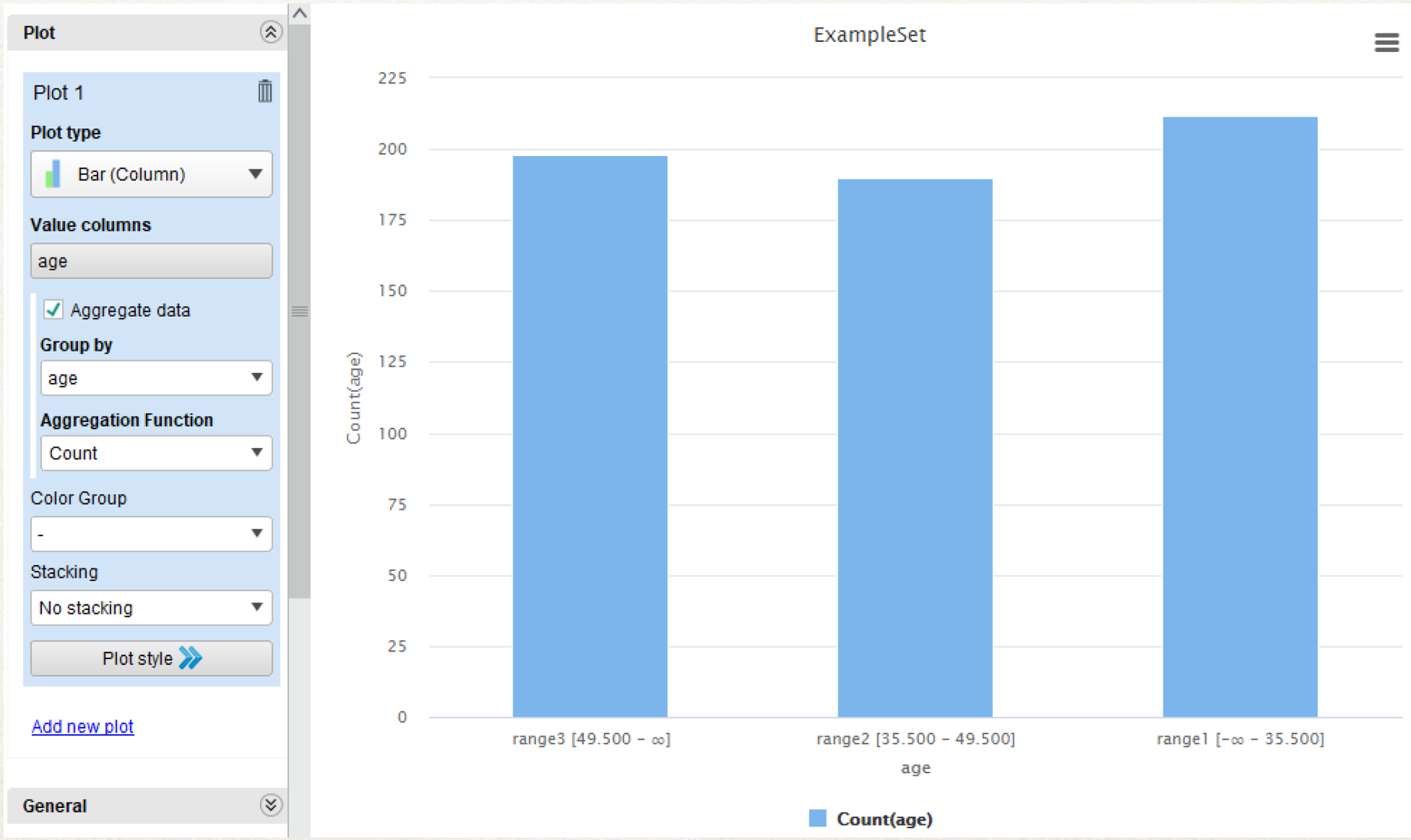
Open in Turbo Prep | Auto Model

Filter (600 / 600 examples): all

Row No.	ï»¿custome...	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