



Data Preparation

No more missing values in the *Discount* attribute.

Result History ExampleSet (Replace Missing Values) ExampleSet (//Lo

Name	Type	Missing	Statis
Discount	Real	0	Min 0.0:
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earlie Jan
Store Number	Integer	0	Min 100
Product_ID	Integer	0	Min 156



UNIVERSITY OF SANTO TOMAS

Dealing with Miscoded entries

using

Altair® AI Studio





Data Preparation

Go back to **Design** view.

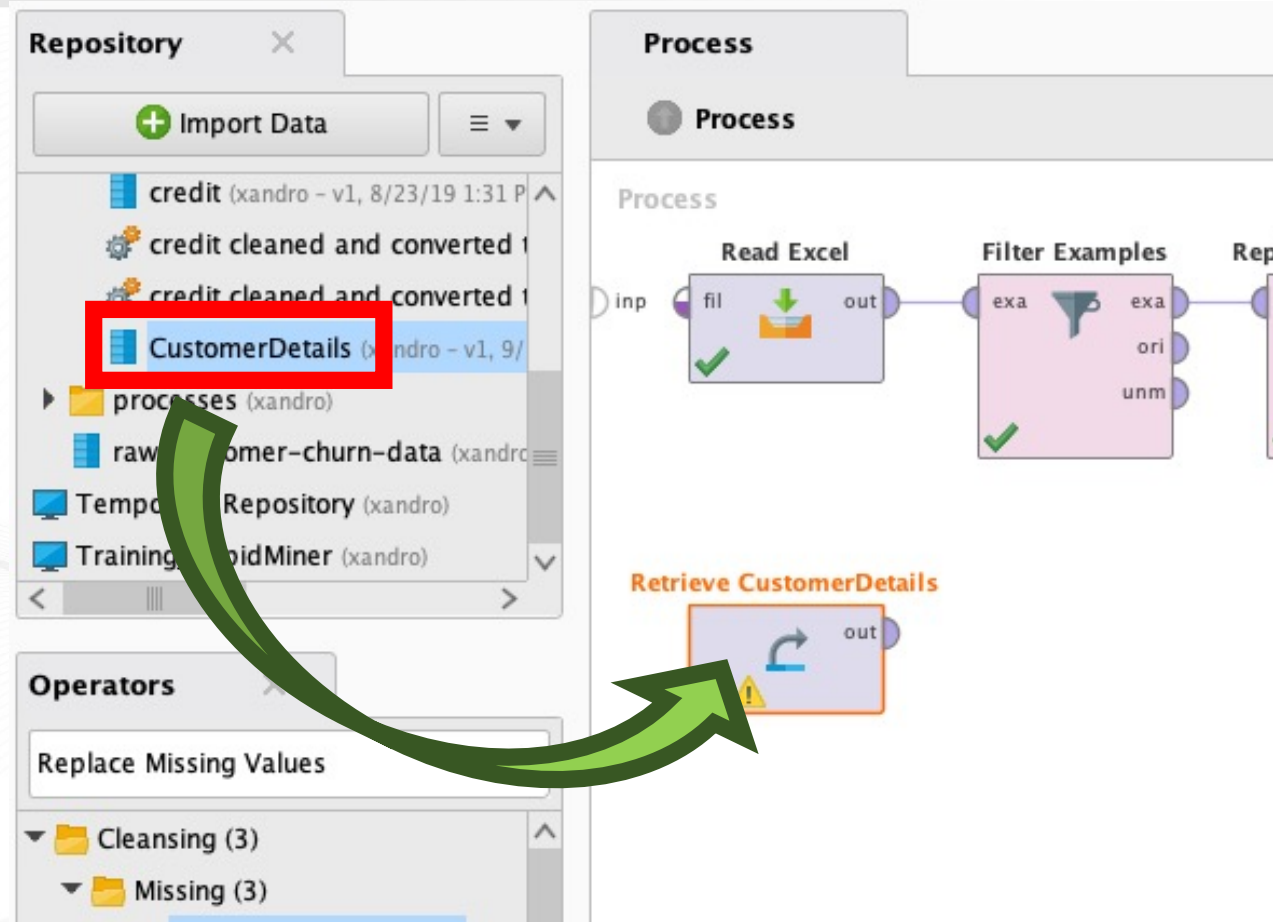
Views: **Design** Results

Result History ExampleSet (Replace Missing Values) ExampleSet (//Local

Name	Type	Missing	Statistic
Discount	Real	0	Min 0.010
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earliest Jan 2,

Data Preparation

- Instead of the Order Details data, we will use the **Customer Details** data.
- Drag and drop the **Customer Details** in the canvas.



The screenshot displays a data preparation tool interface with three main panels:

- Repository:** A list of data sources. The item **CustomerDetails** is highlighted with a red box. Other items include **credit**, **credit cleaned and converted**, **processes**, **raw customer-churn-data**, **Temporary Repository**, and **Training Data Miner**.
- Operators:** A list of data processing operations. The **Retrieve CustomerDetails** operator is highlighted with a green arrow pointing from the **CustomerDetails** data source in the Repository panel.
- Process:** A canvas showing a workflow. The workflow includes a **Read Excel** operator (with a green checkmark), followed by a **Filter Examples** operator (with a green checkmark). The **Retrieve CustomerDetails** operator is shown below the main workflow, with a green arrow pointing to it from the Repository panel.



Data Preparation

- The **Customer Details** data can be viewed in the Results view.

Views: Design **Results** Turbo Prep

Result History ExampleSet (Replace Missing Values) ExampleSet (//Local Repository/)

	Name	Type	Missing	Statistics	Filter (7 / 7)
Data	✓ Responder	Polynomial	100	Least Yes (329)	
Statistics	✓ First Name	Polynomial	0	Least ZENIA (1)	
	✓ Last Name	Polynomial	0	Least ZUELSDORF (1)	
Visualizations	^ Sex	Polynomial	1		
Annotations	✓ Z9_Latitude	Real	0	Min 4.565	
	✓ Z9_Longitude	Real	0	Min 111.026	

Open visualizations



Data Preparation

- Notice in the statistics tab, the **Gender** attribute has miscoded entries.

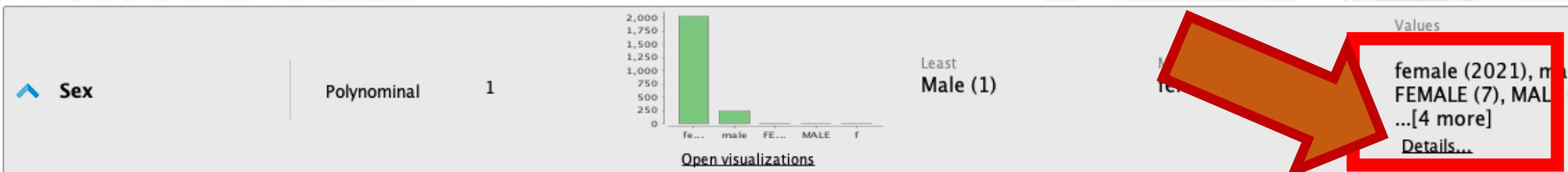
The screenshot shows a software interface with a toolbar at the top containing icons for file operations and execution. Below the toolbar, there are tabs for 'Design', 'Results', and 'Turbo Prep'. The main area displays a table of attributes with columns for Name, Type, Missing, and Statistics. The 'Sex' attribute is highlighted with a red box, and a bar chart is visible next to it.

Name	Type	Missing	Statistics
✓ Responder	Polynomial	100	Least Yes (329)
✓ First Name	Polynomial	0	Least ZENIA (1)
✓ Last Name	Polynomial	0	Least ZUELSDORF (1)
^ Sex	Polynomial	1	Least
✓ Z9_Latitude	Real	0	Min 4.565
✓ Z9_Longitude	Real	0	Min 111.026

Data Preparation


- Notice in the statistics tab, that the **Gender** attribute has miscoded entries.

Click Details...




Data Preparation

- Notice in the statistics tab, that the **Gender** attribute has miscoded entries.



Index	Nominal value	Absolute count	Fraction
1	female	2021	0.892
2	male	230	0.102
3	FEMALE	7	0.003
4	MALE	2	0.001
5	f	2	0.001
6	m	2	0.001
7	female	1	0.000
8	Male	1	0.000

1

 Close



Data Preparation

Go back to **Design** view.

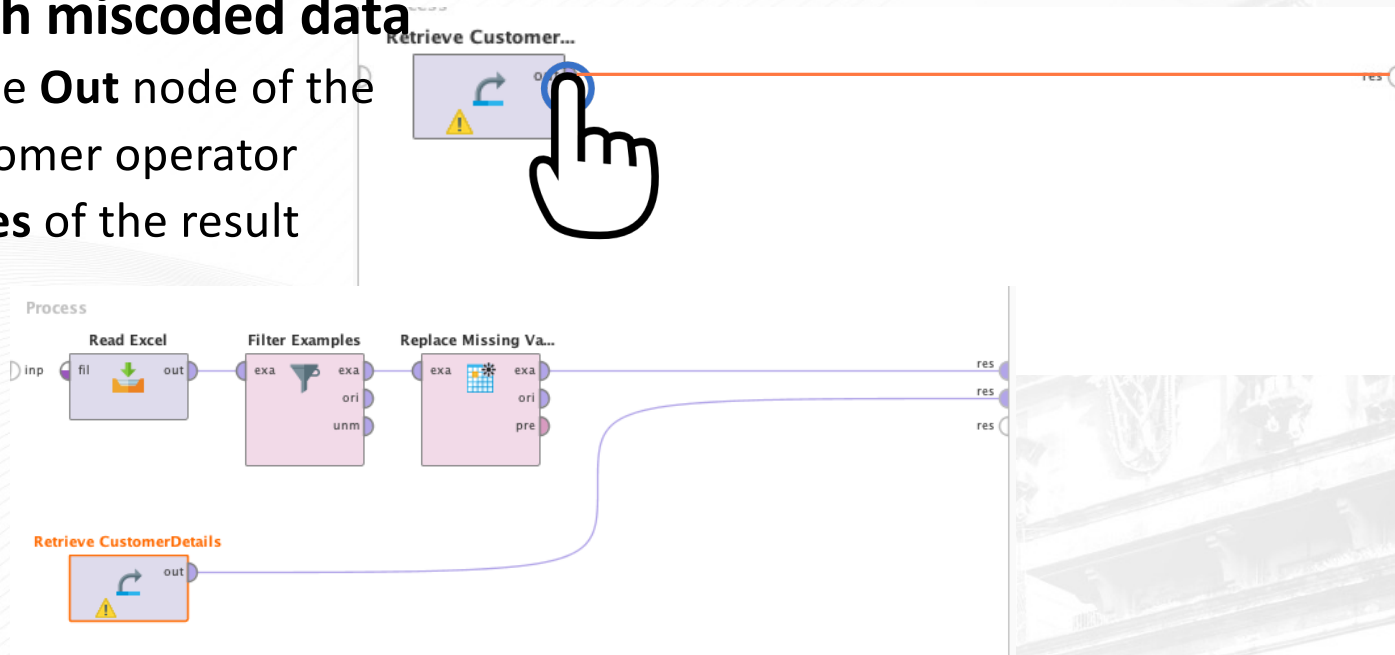
The screenshot shows a software interface for data preparation. At the top, there is a toolbar with icons for file operations and execution. Below the toolbar, a 'View' menu is open, with the 'Design' option highlighted by a red rectangle. The main area displays a table of variables with columns for Name, Type, Missing, and Statistics. The table includes variables like Customer ID, Responder, First Name, Last Name, and Sex. A sidebar on the left contains icons for Data, Statistics, and Visualizations.

Name	Type	Missing	Statistics
Customer ID	Integer	0	Min 2
Responder	Polynomial	100	Least Yes (329)
First Name	Polynomial	0	Least ZENIA (1)
Last Name	Polynomial	0	Least ZUELSDORF (1)
Sex	Polynomial	1	

Data Preparation

3. Dealing with miscoded data

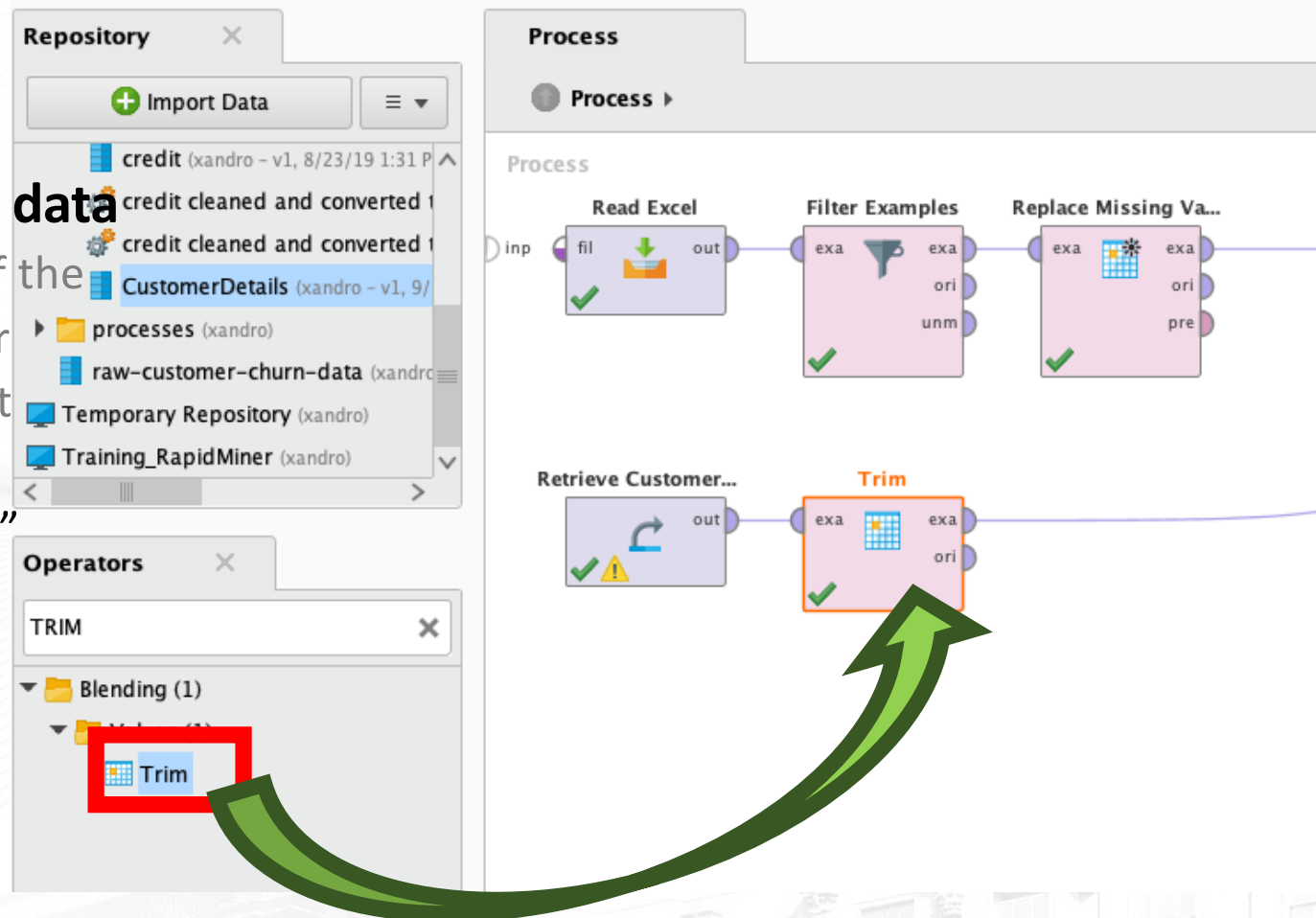
- Connect the **Out** node of the Retrieve Customer operator and **second res** of the result knob.



Data Preparation

3. Dealing with miscoded data

- Connect the **Out** node of the Retrieve Customer operator and **second res** of the result knob.
- To remove “white spaces” in the encoding, use the **TRIM** operator.

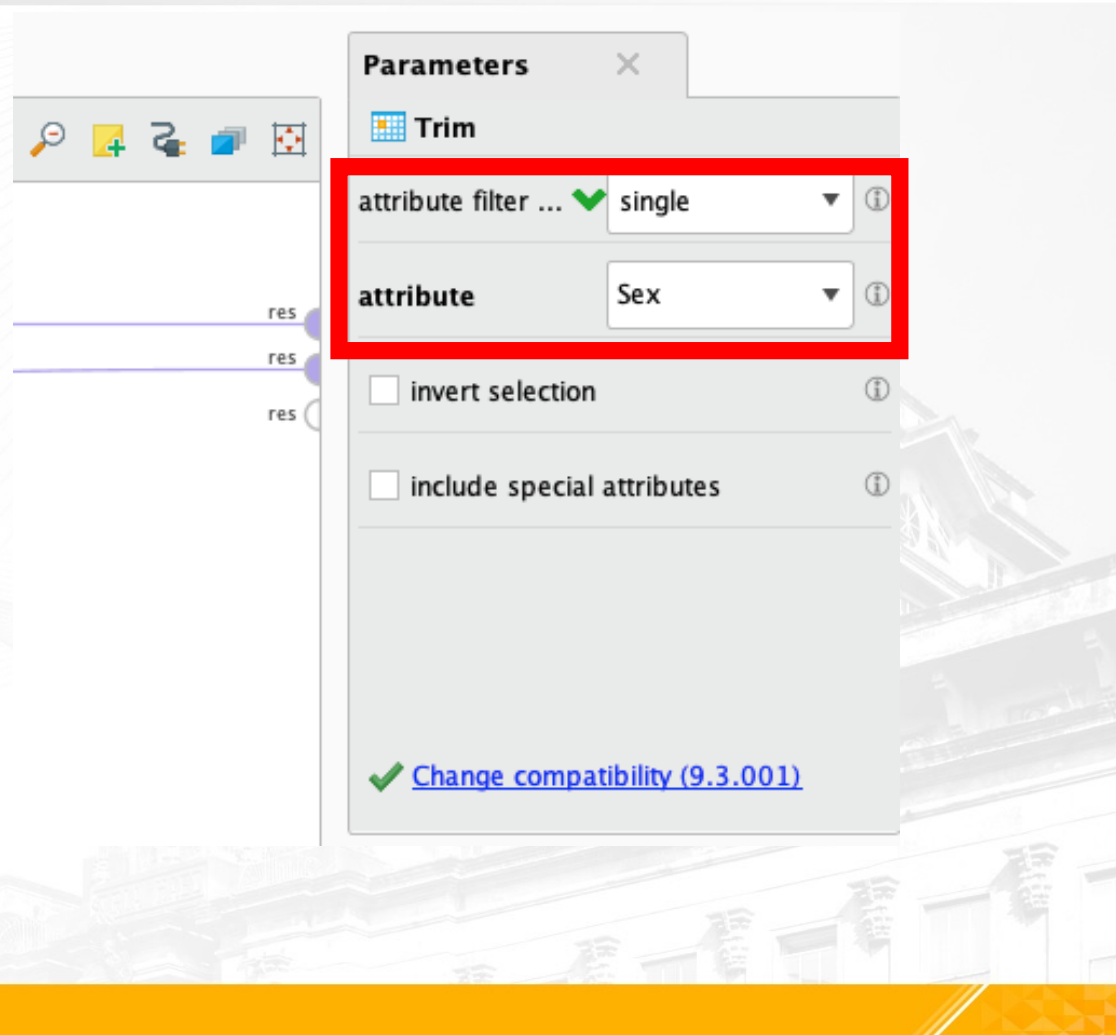


Data Preparation

3. Dealing with miscoded data

- Select single if trimming shall be applied to a single attribute.

Then click the **Play** button.





Data Preparation

You may see the *trimming* result by viewing the statistics.

The screenshot shows a software interface with a table of data. The table has columns for Name, Type, Missing, and Statistics. The 'ExampleSet (Trim)' tab is highlighted with a red box. The 'Statistics' icon is also highlighted with a red box. A red arrow points to a 'Details...' link in the statistics section. The statistics section shows a bar chart and a list of values: female (2022), male (23), FEMALE (7), MALE (2), ...[3 more].

Name	Type	Missing	Statistics
Polynomial	1		Least Male (1) Most female (2022)

Click Details...

Data Preparation

You may see the *trimming* result by viewing the statistics.

Before

Index	Nominal value	Absolute count	Fraction
1	female	2021	0.892
2	male	230	0.102
3	FEMALE	7	0.003
4	MALE	2	0.001
5	f	2	0.001
6	m	2	0.001
7	female	1	0.000
8	Male	1	0.000

Close



After

Index	Nominal value	Absolute count	Fraction
1	female	2022	0.892
2	male	230	0.102
3	FEMALE	7	0.003
4	MALE	2	0.001
5	f	2	0.001
6	m	2	0.001
7	Male	1	0.000

Close



Data Preparation

Go back to **Design** view.

Views: **Design** Results

Result History

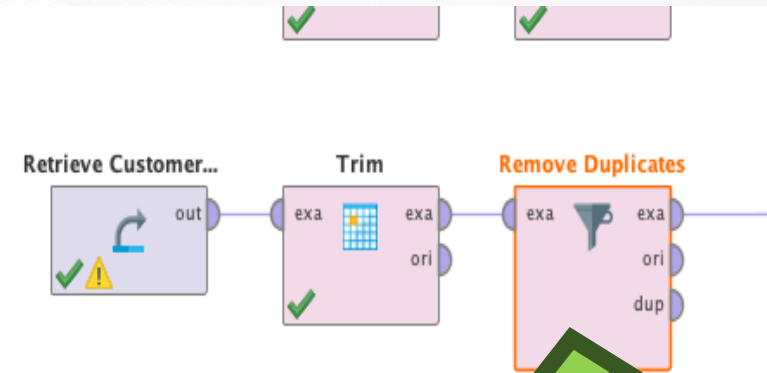
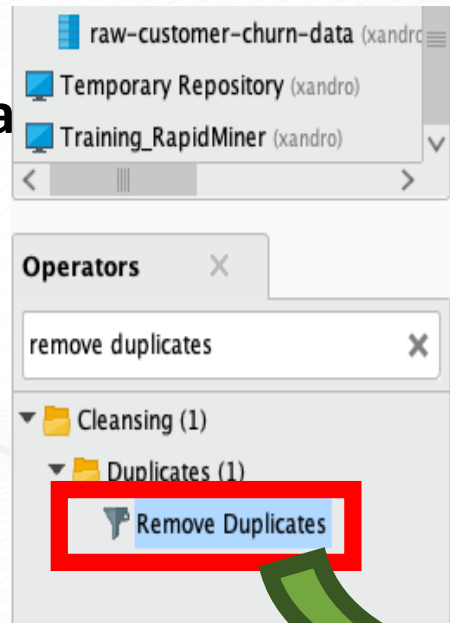
- ExampleSet (Replace Missing Values) x
- ExampleSet (//Local

Name	Type	Missing	Statistic
Discount	Real	0	Min 0.010
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earliest Jan 2,

Data Preparation

3. Dealing with miscoded data

- To remove “duplicates” in the encoding, use the **Remove Duplicates** operator.

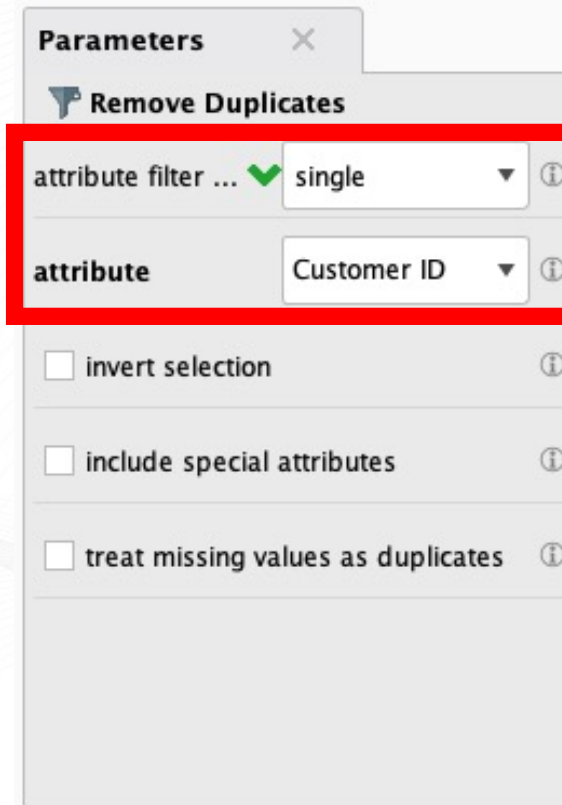


Data Preparation

3. Dealing with miscoded data

- Select single if trimming shall be applied to a single attribute.

Then click the **Play** button.



The screenshot shows a 'Parameters' dialog box with a close button (X) in the top right corner. The title of the dialog is 'Remove Duplicates'. Below the title, there are three main sections. The first section is 'attribute filter ...' with a green checkmark icon and a dropdown menu set to 'single'. This section is highlighted with a red border. The second section is 'attribute' with a dropdown menu set to 'Customer ID'. The third section contains three checkboxes: 'invert selection', 'include special attributes', and 'treat missing values as duplicates', each with an information icon (i) to its right.

This will retain only one entry if duplicate Customer IDs have been found.



Data Preparation

Still, 2267 cases are retained, indicating that there are no duplicates in Customer IDs.

Filter (2,267 / 2,267 examples):

Last Name	Z9_Latitude	Z9_Longitu...
SMITH	14.558	121.079
CARRERA	14.508	121.154
WACK	14.727	121.020
GAULDEN	14.607	120.974
WRIGHT	14.639	121.054
GONZALEZ	14.543	120.934
MAYBERRY	14.514	121.100



Data Preparation

Go back to **Design** view.

Views: **Design** Results

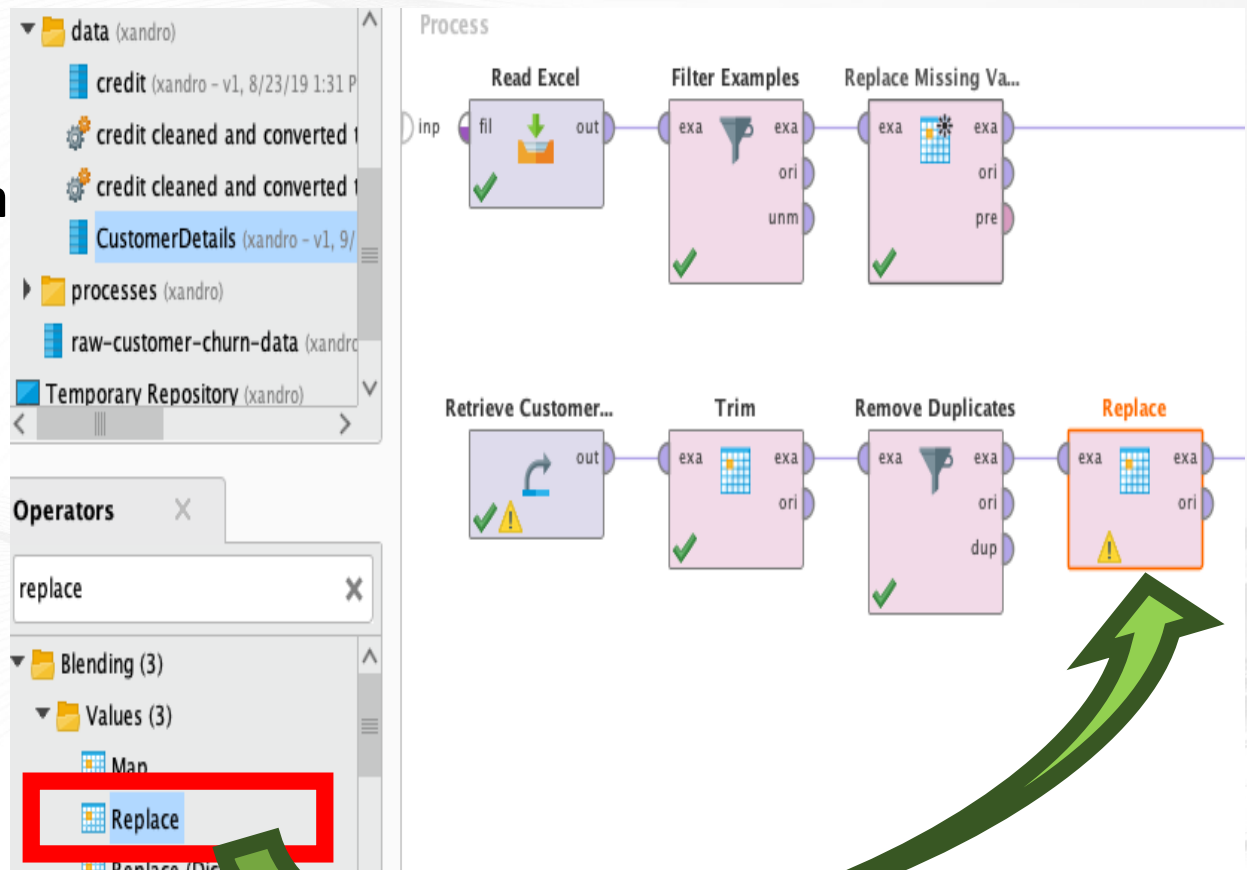
Result History ExampleSet (Replace Missing Values) ExampleSet (//Local

Name	Type	Missing	Statistic
Discount	Real	0	Min 0.010
Customer_ID	Integer	0	Min 10
Order_ID	Integer	0	Min 193
Order_Date	Date	0	Earliest Jan 2,

Data Preparation

3. Dealing with miscoded data

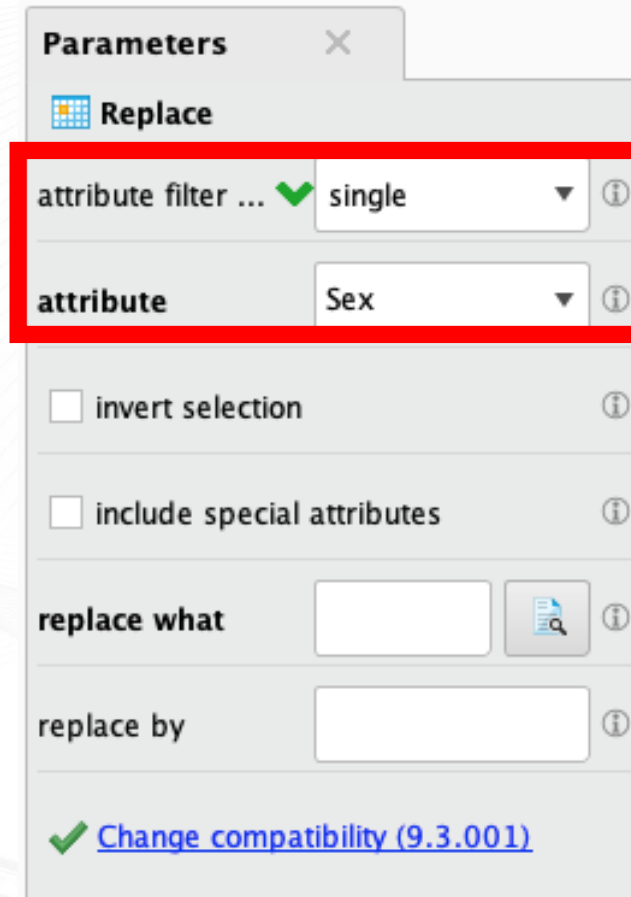
- To recode miscoded values, use the **REPLACE** operator.



Data Preparation

3. Dealing with miscoded data

Select single if replacing of values shall be applied to a single attribute.



Parameters

Replace

attribute filter ... single

attribute Sex

invert selection

include special attributes


replace what

replace by


[Change compatibility \(9.3.001\)](#)

Data Preparation


Temporarily,
female → girl



Index	Nominal value	Absolute count	Fraction
1	female	2022	0.892
2	male	230	0.102
3	FEMALE	7	0.003
4	MALE	2	0.001
5	f	2	0.001
6	m	2	0.001
7	Male	1	0.000



Parameters ×


 **Replace**

attribute filter ... single ⓘ

attribute Sex ⓘ

invert selection ⓘ

include special attributes ⓘ

replace what female  ⓘ

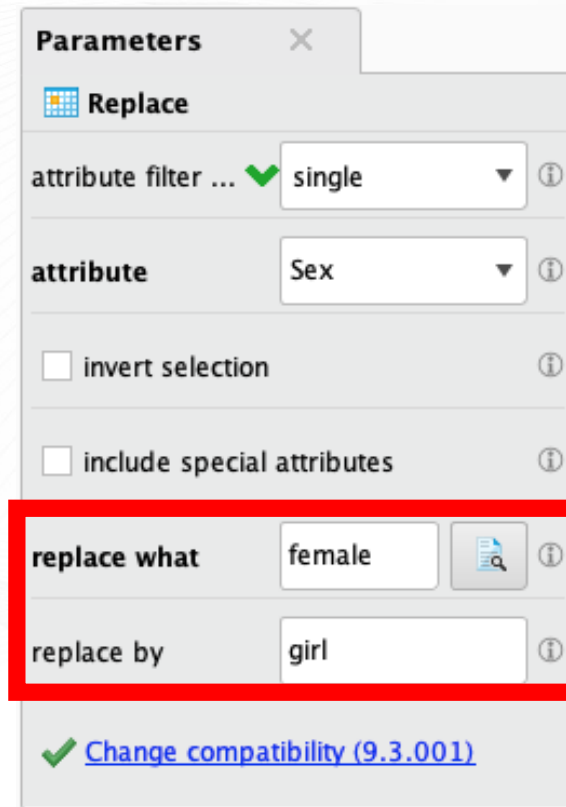
replace by girl ⓘ

[Change compatibility \(9.3.001\)](#)


Data Preparation


3. Dealing with miscoded data


Select single if replacing of values shall be applied to a single attribute.





Parameters [X]



 **Replace**


attribute filter ... single 

attribute Sex 

invert selection 

include special attributes 

replace what female  

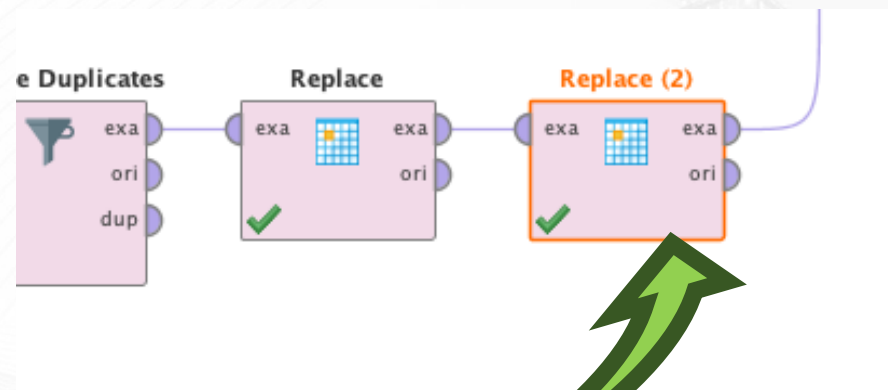
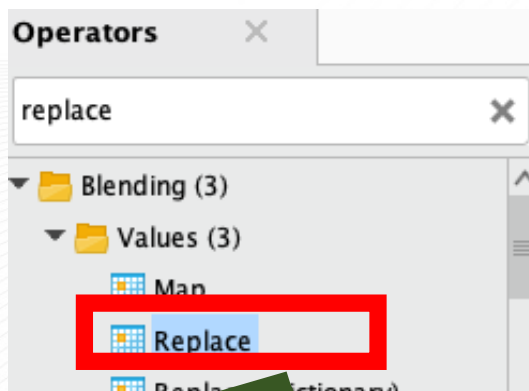
replace by girl 

[Change compatibility \(9.3.001\)](#)

Data Preparation

3. Dealing with miscoded data

- Add another **REPLACE** operator,









Data Preparation


Replace FEMALE with girl.



Replace (2) (Replace)


attribute filter type  single 

attribute Sex 

invert selection 

include special attributes 

replace what FEMALE  

replace by girl 



Data Preparation

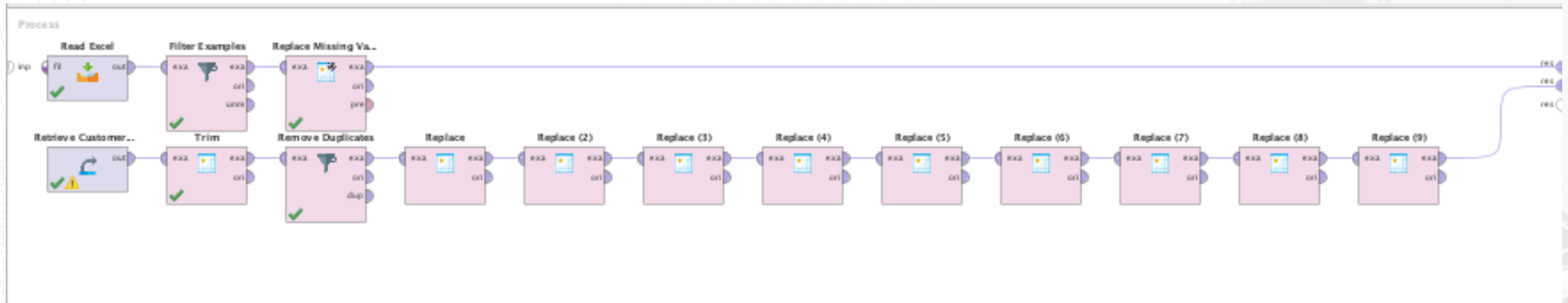
3. Dealing with miscoded data

- Add another **REPLACE** operator replacing male with boy;
- Add another **REPLACE** operator replacing m with boy;
- Add another **REPLACE** operator replacing f with girl;
- Add another **REPLACE** operator replacing MALE with boy;
- Add another **REPLACE** operator replacing Male with boy;

To replace back girl and boy to female and male, respectively,

- Add another **REPLACE** operator replacing girl with female;
- Add another **REPLACE** operator replacing boy with male.

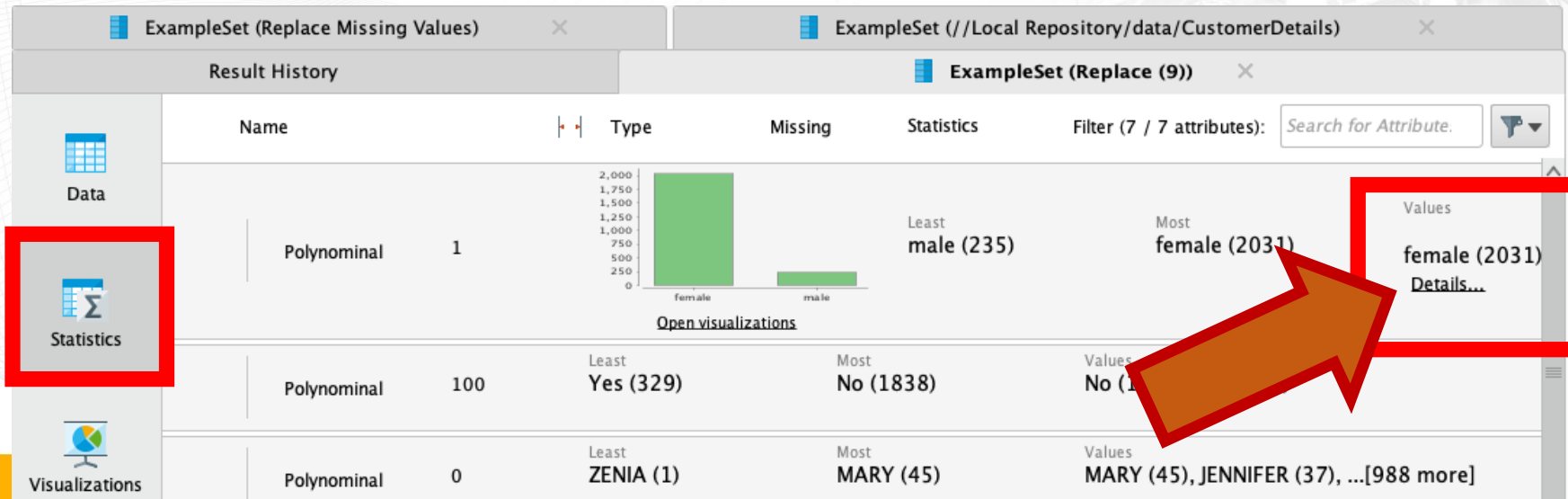
Data Preparation



Data Preparation

3. Dealing with miscoded data

- Click the **Play** button to verify the process.
- Once done, click Statistics, and on the Sex attribute, click Details...



The screenshot shows a software interface with two tabs: "ExampleSet (Replace Missing Values)" and "ExampleSet (//Local Repository/data/CustomerDetails)". The main window displays a "Result History" table with columns for Name, Type, Missing, Statistics, and Filter (7 / 7 attributes). A bar chart is visible for the "Sex" attribute, showing counts for "female" and "male". A red box highlights the "Statistics" icon in the left sidebar, and another red box highlights the "Details..." link for the "female (2031)" value in the "Values" column. A large red arrow points from the "Details..." link towards the right side of the screen.

Name	Type	Missing	Statistics	Filter (7 / 7 attributes):
Polynomial	1		Least male (235) Most female (2031)	Search for Attribute.
Polynomial	100	Least Yes (329) Most No (1838)	Values No (1)	
Polynomial	0	Least ZENIA (1) Most MARY (45)	Values MARY (45), JENNIFER (37), ...[988 more]	

Data Preparation

3. Dealing with miscoded data

- Click the **Play** button to verify the process.
- Once done, click Statistics, and on the Sex attribute, click Details...



Index	Nominal value	Absolute count	Fraction
1	female	2031	0.896
2	male	235	0.104



Data Preparation

You **may** impute missing values using **REPLACE MISSING VALUES** operator in other attributes.

Name	Type	Missing	Statistics	Filter (7 / 7 attributes): <input type="text" value="Search for Attribute"/>
Sex	Polynomial	1	Least male (242)	Most female (2024)
Responder	Polynomial	100	Least Yes (329)	Most No (1838)



UNIVERSITY OF SANTO TOMAS

Selecting and Setting Roles of Attributes

using

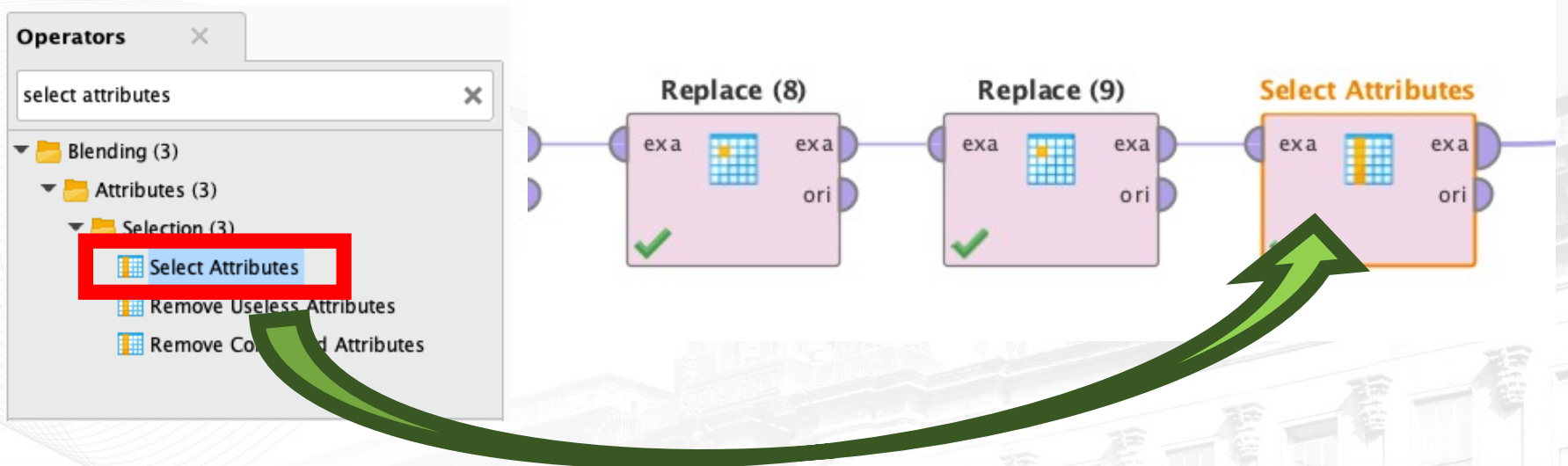


Altair[®] AI Studio

Data Preparation

4. Selecting the Attributes for Analysis

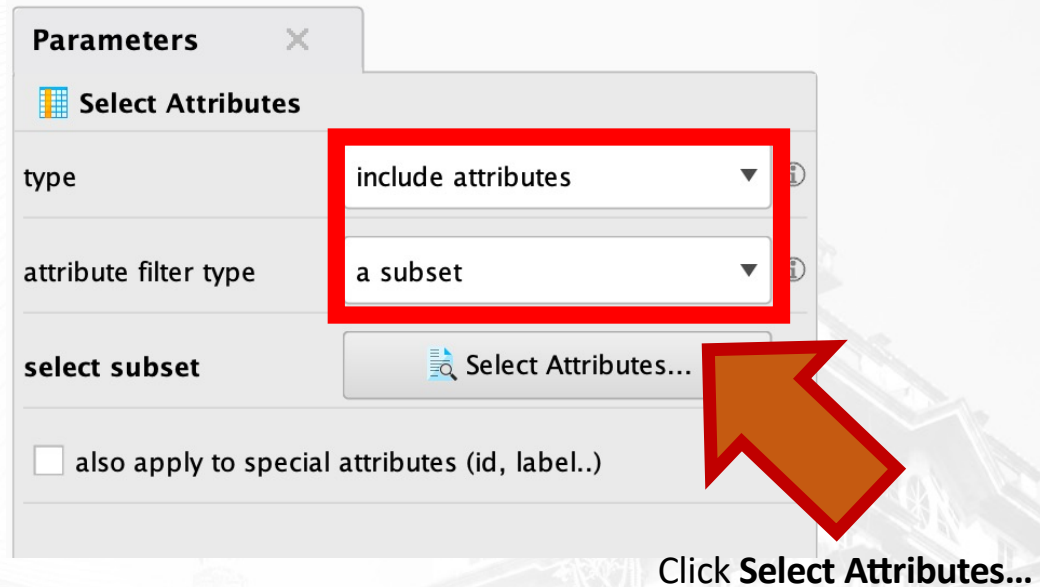
- Use the **Select Attributes** operator to select the attributes needed in the analysis.



Data Preparation

4. Selecting the Attributes for Analysis

- You can select all the attributes, single, and or a subset.



The screenshot shows a 'Parameters' dialog box with a close button (X) in the top right corner. Below the title bar is a tab labeled 'Select Attributes' with a grid icon. The dialog contains the following fields and controls:

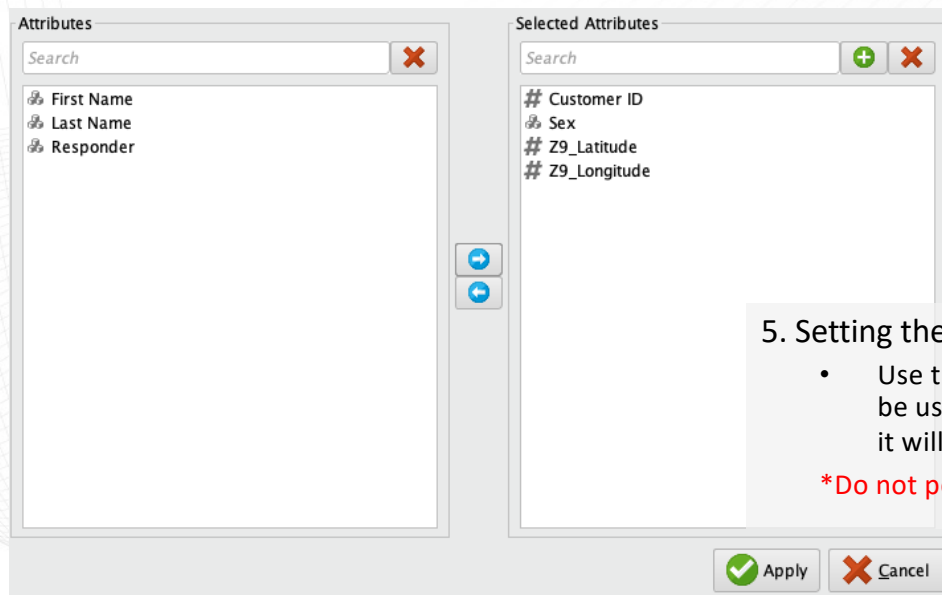
- type**: A dropdown menu with 'include attributes' selected. This field is highlighted with a red rectangular border.
- attribute filter type**: A dropdown menu with 'a subset' selected.
- select subset**: A button with a magnifying glass icon and the text 'Select Attributes...'. A large red arrow points to this button.
- also apply to special attributes (id, label..)**

Below the dialog box, the text 'Click Select Attributes...' is displayed.

Data Preparation

4. Selecting the Attributes for Analysis

- Select the Attributes that will be used for analysis.

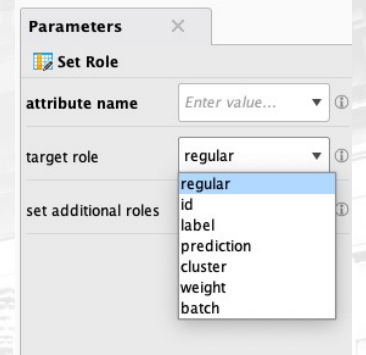


This will remove the names and Responder attribute in the final data.

5. Setting the role that an attribute to perform.

- Use the **Set Role** operator to tag the attribute that will be used as the label (Target Variable) or any other role it will act in the analysis.

**Do not perform this yet on this topic.*





UNIVERSITY OF SANTO TOMAS

Combining Data Sets

using

Altair® AI Studio

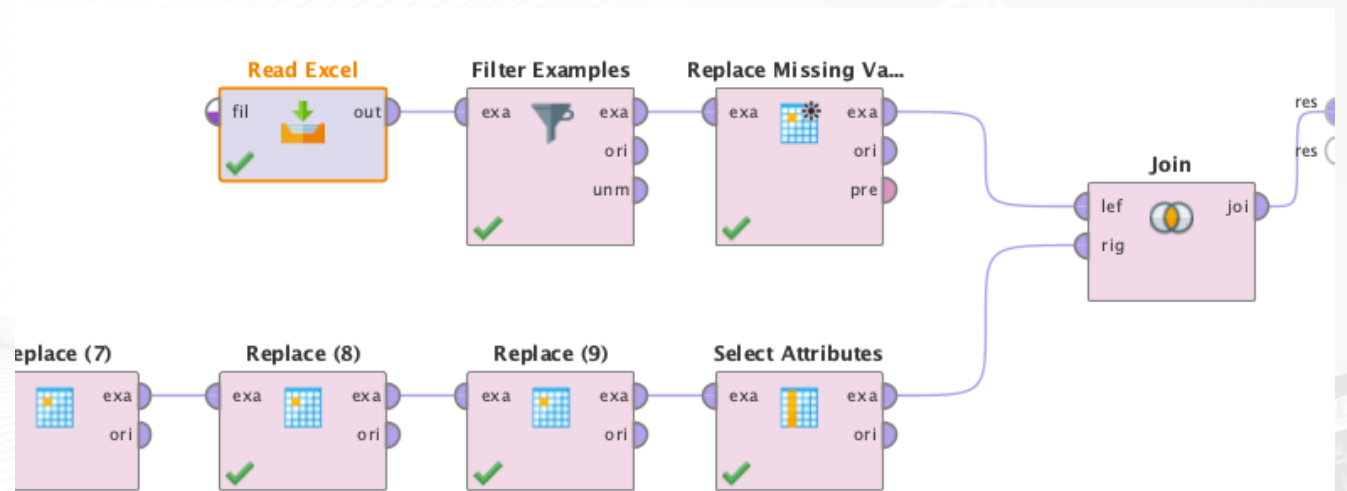


Data Preparation

6. Joining Two Data Sets

If two data sets are needed to be merged in order to make an analysis, use the **Join** operator.

- Connect the first data set or its result in the left node of the **Join** operator and the other data set at the right node.

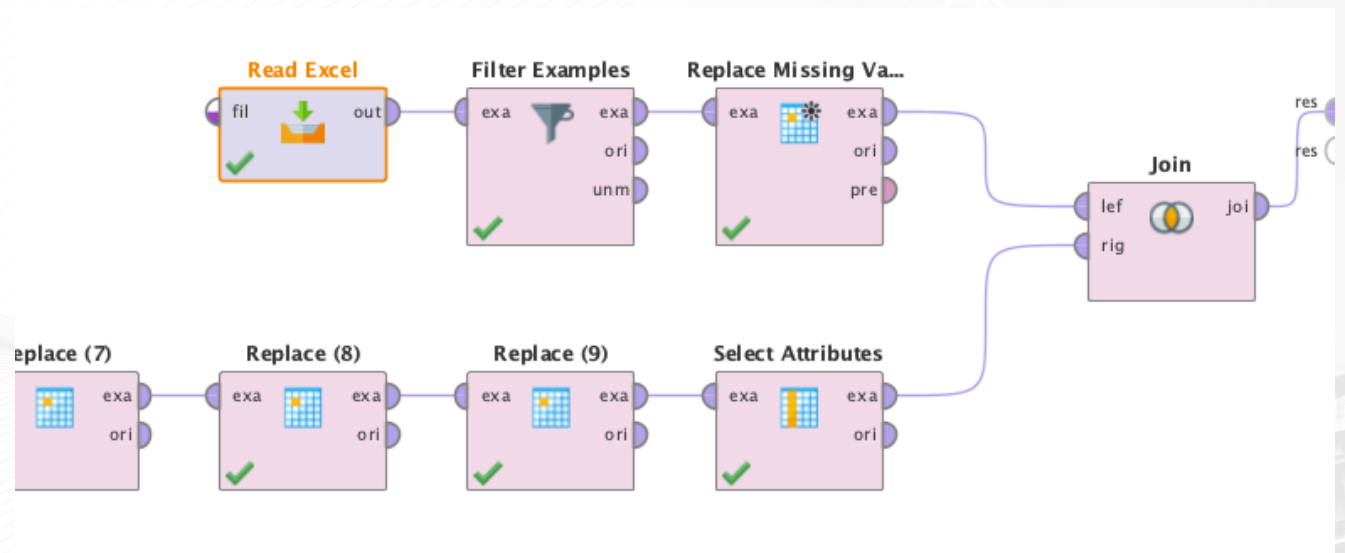


Data Preparation

6. Joining Two Data Sets

If two data sets are needed to be merged in order to make an analysis, use the **Join** operator.

- Connect the first data set or its result in the left node of the **Join** operator and the other data set at the right node.

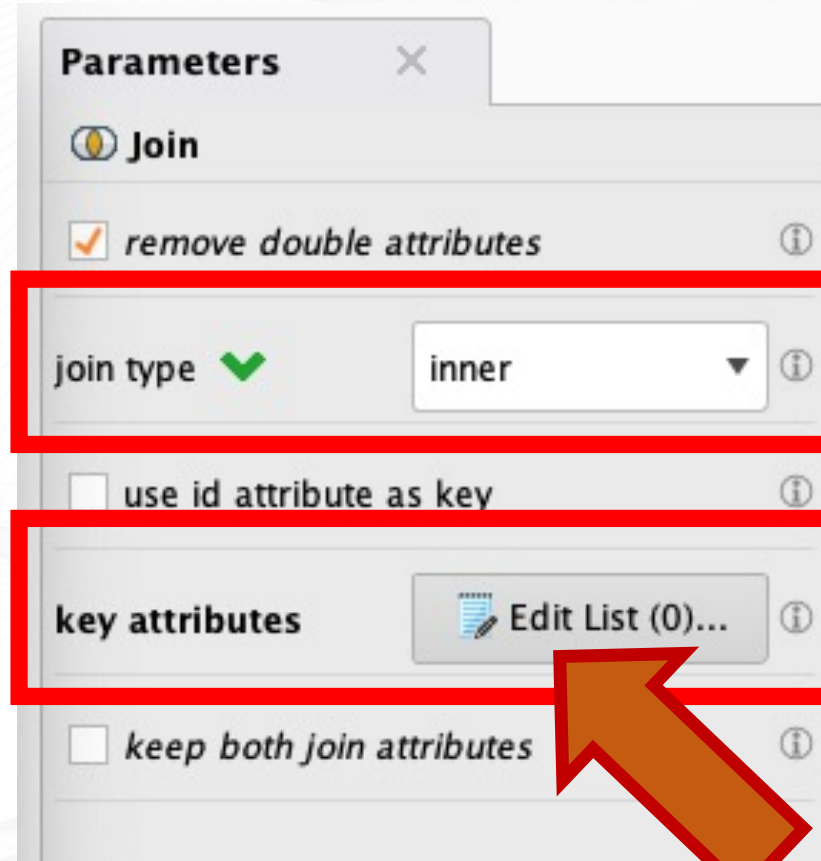


Data Preparation

6. Joining Two Data Sets

In the parameter tab, use **Inner** as join type.

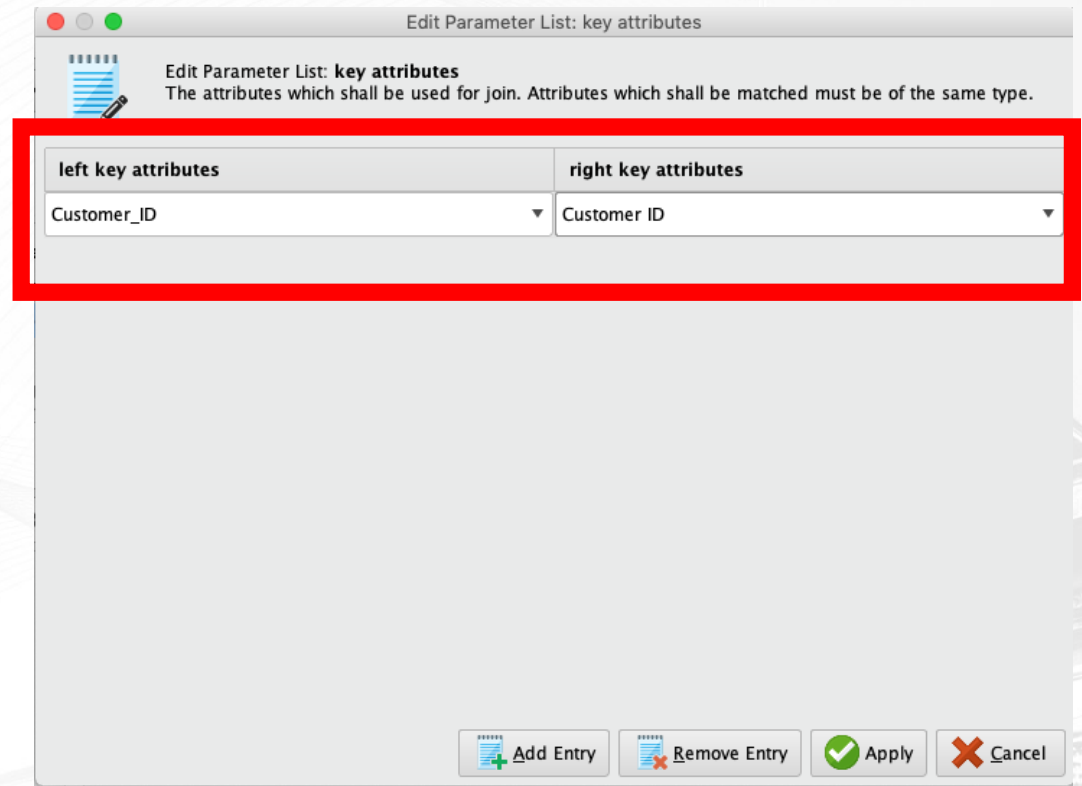
Click **Edit List**.



Data Preparation

6. Joining Two Data Sets

Select the attribute on the first data (left) and the second data (right) that will be used in matching the two data sets.





Data Preparation

6. Joining Two Data Sets

Select the attribute on the first data (left) and the second data (right) that will be used in matching the two data sets.

Click **Apply**, then click the **Play** button.

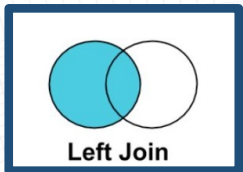
Open in Turbo Prep Auto Model Filter (2,262 / 2,262 examples): all

Row No.	Discount	Customer_L...	Order_ID	Order_Date	Store Num...	Product_ID	Unit_Price	Quantity
1	0.090	101	88209	Apr 26, 2014	100	16474	145.450	13
2	0.090	102	16710	Apr 26, 2014	103	16474	145.450	53
3	0.100	129	86698	Nov 30, 2015	101	16124	111.030	43
4	0.090	180	90784	Nov 4, 2015	100	16238	291.730	21
5	0.070	180	90785	Dec 14, 2015	100	16566	140.980	21
6	0.080	1025	89013	Nov 9, 2014	108	16182	284.980	20
7	0.080	1027	89016	Dec 29, 2015	103	15956	449.990	20
8	0.090	1030	89622	Apr 22, 2014	100	15653	175.990	11
9	0.060	1043	87851	Sep 8, 2015	100	16083	270.970	12
10	0.060	1056	90215	Jan 15, 2013	106	16244	113.980	12
11	0.090	1060	58628	Mar 25, 2012	105	15872	138.750	23
12	0.060	1074	86424	Jan 27, 2014	107	16474	145.450	11
13	0.080	1106	45824	Dec 8, 2012	103	16253	140.810	81
14	0.090	1118	86773	Nov 15, 2013	102	16361	209.840	26

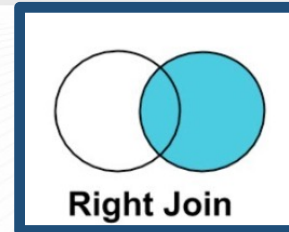
Joining two data sets

Data file 1		
ID	Sex	Amount Deposited
1	M	3000
2	M	4000
4	M	20000
5	F	8000
7	F	10000
8	F	40000

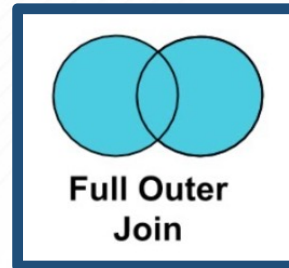
Data file 2		
ID	Age	Total Balance
1	25	15000
2	29	20000
3	30	35000
4	38	210000
6	51	80000
8	45	120000



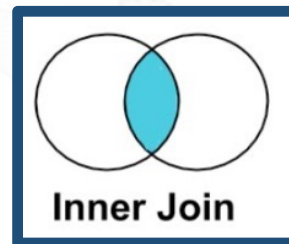
ID	Sex	Amount Deposited	Age	Total Balance
1	M	3000	25	15000
2	M	4000	29	20000
4	M	20000	38	210000
5	F	8000	-	-
7	F	10000	-	-
8	F	40000	45	120000



ID	Sex	Amount Deposited	Age	Total Balance
1	M	3000	25	15000
2	M	4000	29	20000
3	-	-	30	35000
4	M	20000	38	210000
6	-	-	51	80000
8	F	40000	45	120000



ID	Sex	Amount Deposited	Age	Total Balance
1	M	3000	25	15000
2	M	4000	29	20000
3	-	-	30	35000
4	M	20000	38	210000
5	F	8000	-	-
6	F	40000	51	80000
7	F	10000	-	-
8	F	40000	45	120000



ID	Sex	Amount Deposited	Age	Total Balance
1	M	3000	25	15000
2	M	4000	29	20000
4	M	20000	38	210000
8	F	40000	45	120000



Data Preparation

7. You may save your work using File >> **Save Process**.

8. You may also creating a new data set from the cleaned/pre-process data.

- You may add the **Store** operator to create a RapidMiner data set from the process;
- and use the **Write ***** operator to store the data in a format you want.



UNIVERSITY OF SANTO TOMAS



Data Preparation using Altair[®] AI Studio

