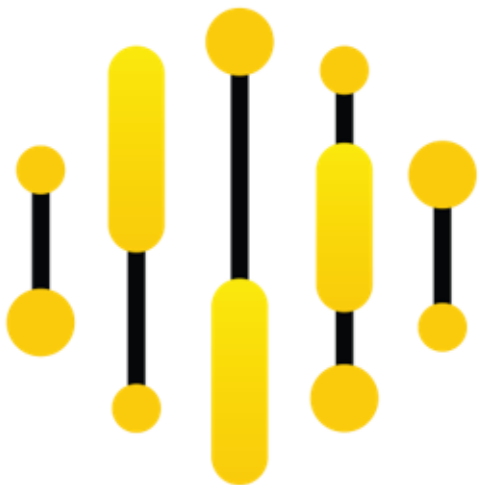


# Training Course: Intermediate Data Engineering Data Preprocessing with RapidMiner

Titirat Boonchuaychu  
Data Engineer



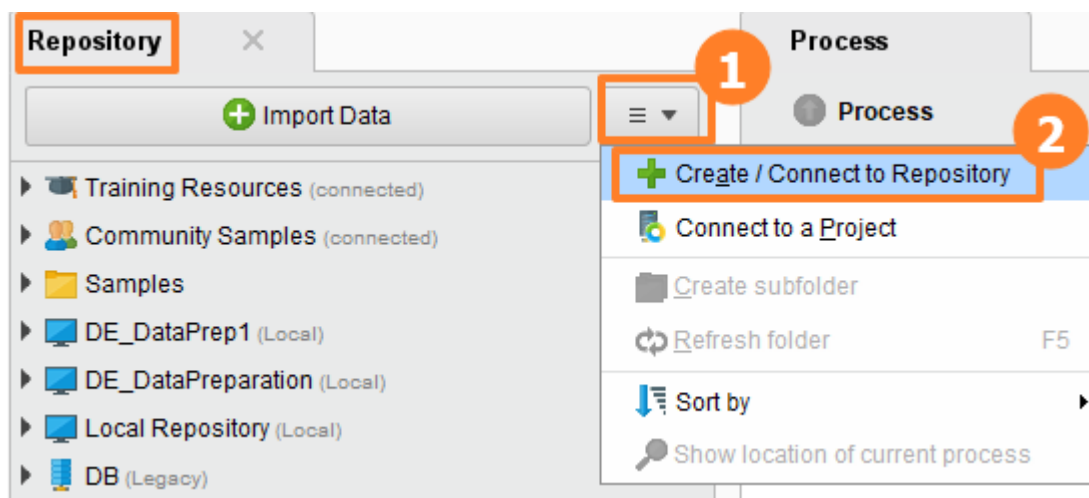
## Workshop: Create/ Connect to Repository

---

Training Course: Intermediate Data Engineering

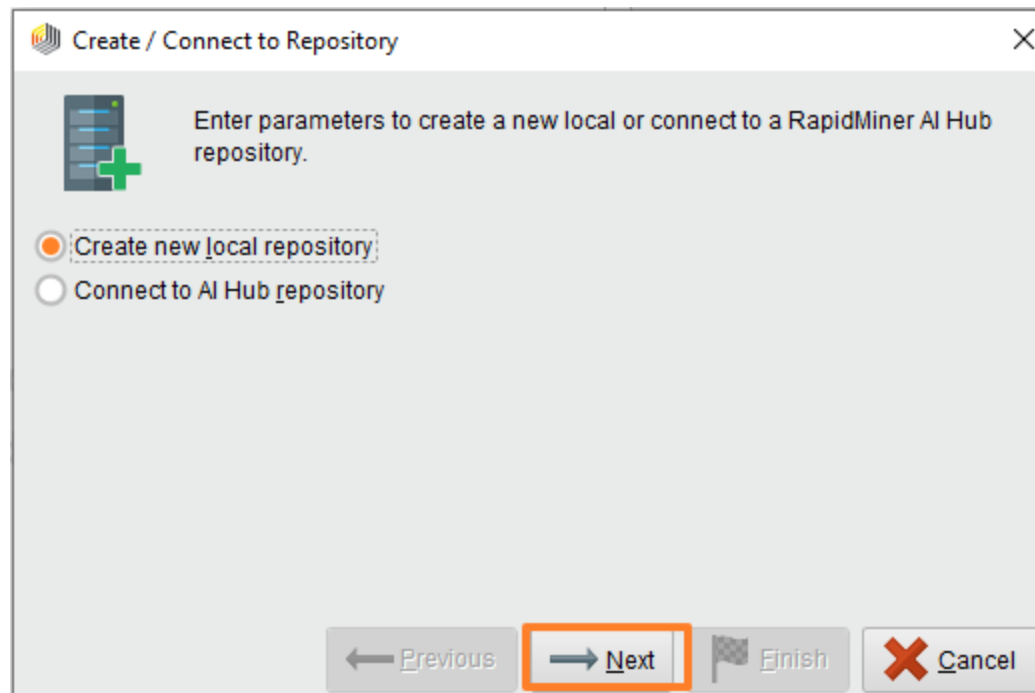
# Workshop: Create/ Connect to Repository

- สร้างโฟลเดอร์ไว้ที่ไดรฟ์ D: ชื่อว่า **DE\_DataPreprocessing**
- เปิดโปรแกรม RapidMiner Studio
- ไปที่แถบ **Repository** และคลิกที่ปุ่ม  จากนั้นเลือก **Create / Connect to Repository**




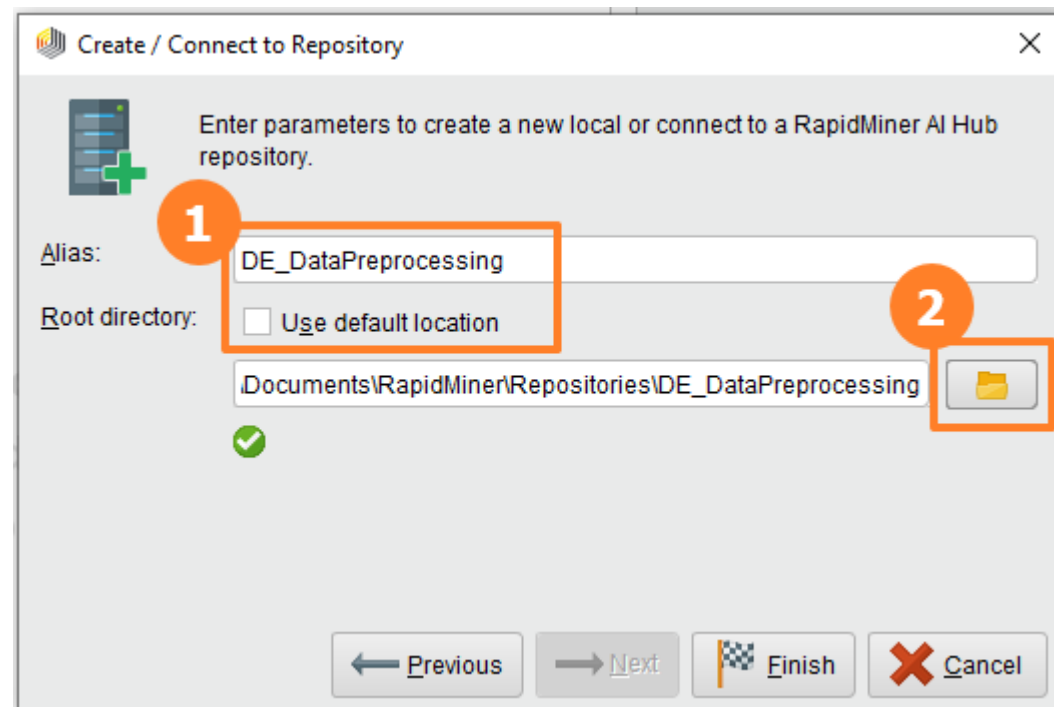
# Workshop: Create/ Connect to Repository

- เลือก Create new local repository และกดปุ่ม Next



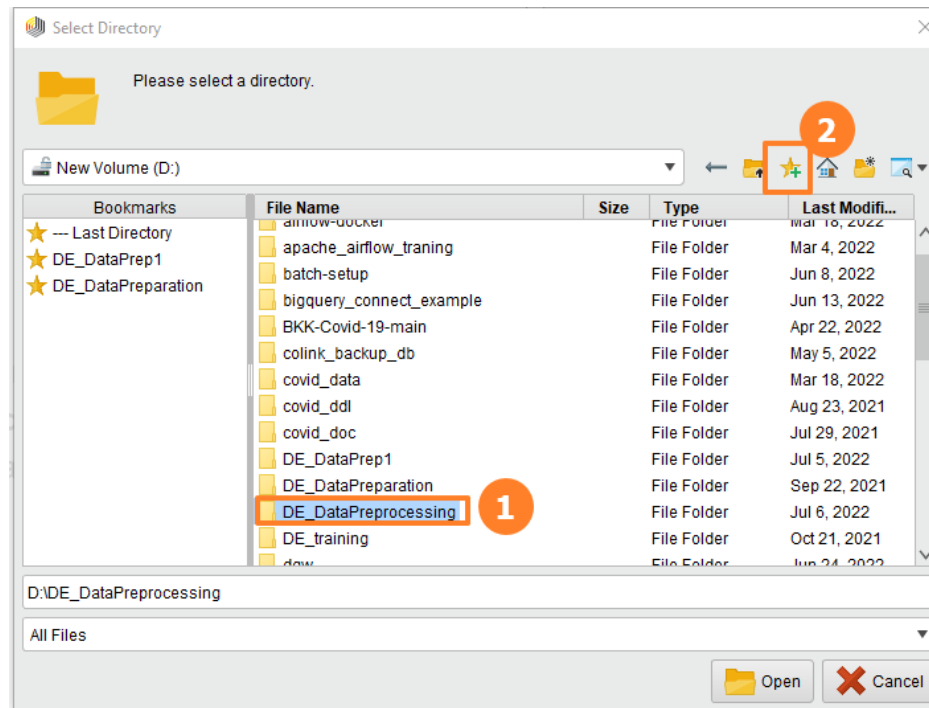
# Workshop: Create/ Connect to Repository

- ตั้งชื่อ Repository (Alias) เป็น **DE\_DataPreprocessing** (ชื่อเดียวกับชื่อไฟล์เดอร์)
- คลิก **Unchecked** ที่ **Use default location** เพื่อระบุโฟลเดอร์ที่จะใช้จัดเก็บเอง
- กดปุ่ม  แล้วเลือกโฟลเดอร์ **DE\_DataPreprocessing** ที่สร้างไว้ในไดรฟ์ **D:** (จากขั้นตอนแรก)



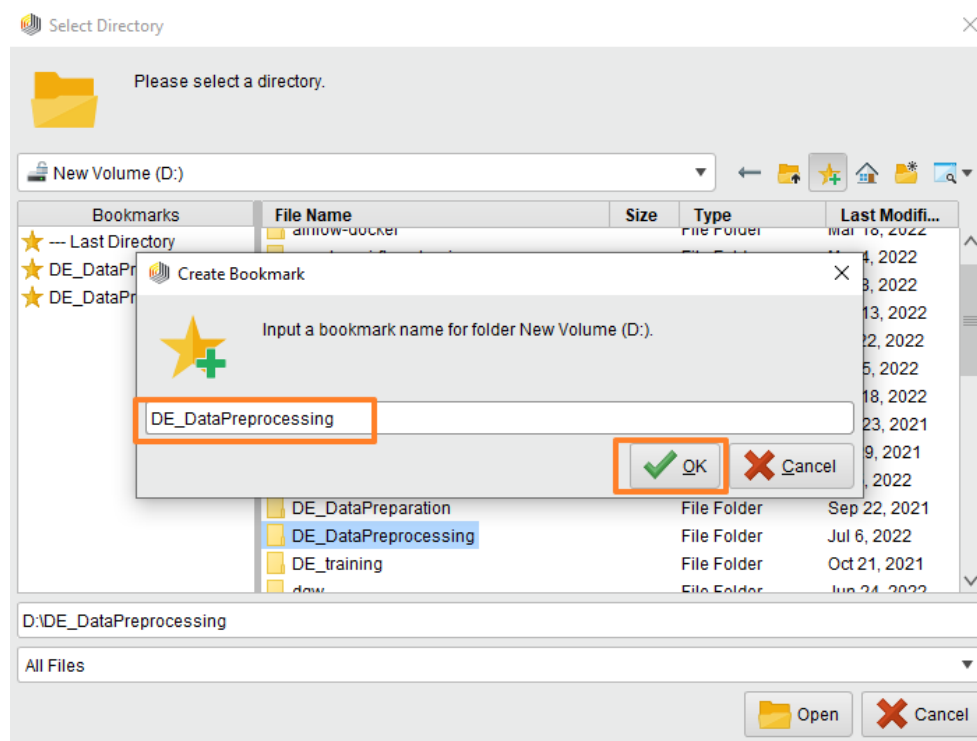
# Workshop: Create/ Connect to Repository

- เลือกโฟลเดอร์ **DE\_DataPreprocessing** ที่สร้างไว้ในไดรฟ์ D: (จากขั้นตอนแรก)
- กดปุ่ม **Add Bookmarks** (รูปดาว มีเครื่องหมายบวก) เพื่อให้สะดวกในการใช้งานในครั้งต่อไป



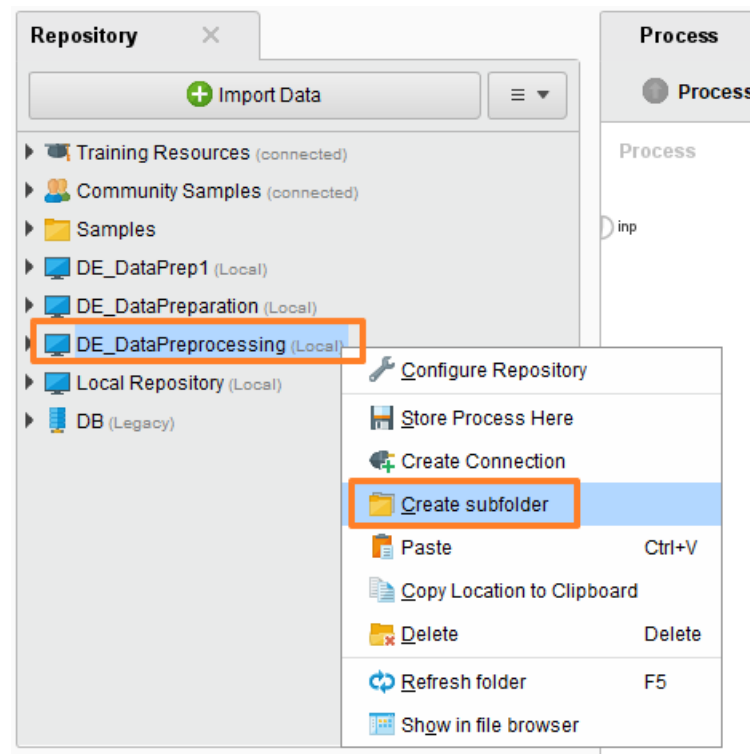
# Workshop: Create/ Connect to Repository

- กรอกชื่อ Bookmark เป็น **DE\_DataPreprocessing** และกดปุ่ม OK



# Workshop: Create/ Connect to Repository

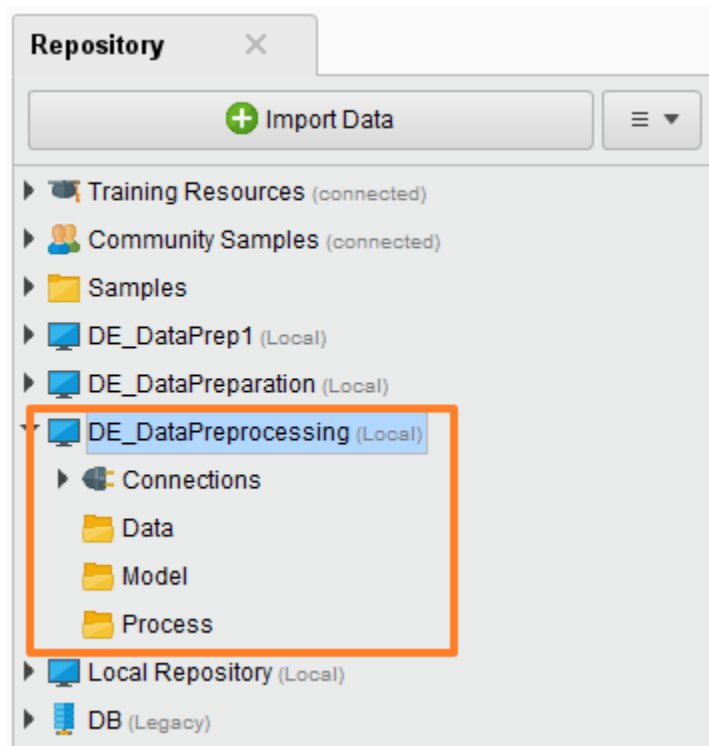
- เมื่อกลับไปดูที่แถบ Repository จะปรากฏ Repository: **DE\_DataPreprocessing** ที่สร้างขึ้น
- คลิกขวาที่ Repository: DE\_DataPreprocessing แล้วเลือก **Create subfolder** เพื่อสร้างโฟลเดอร์ย่อย สำหรับแยกเก็บไฟล์ประเภทต่างๆ

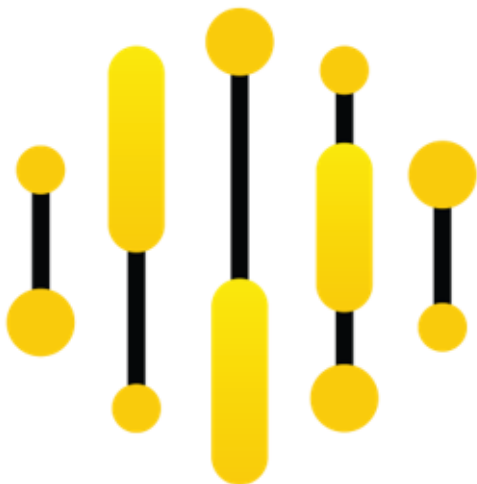




# Workshop: Create/ Connect to Repository

- สร้างโฟลเดอร์ย่อยไว้ 3 โฟลเดอร์ ตั้งชื่อว่า **Data**, **Model** และ **Process** ตามลำดับ





## Workshop: Import/ Access Data

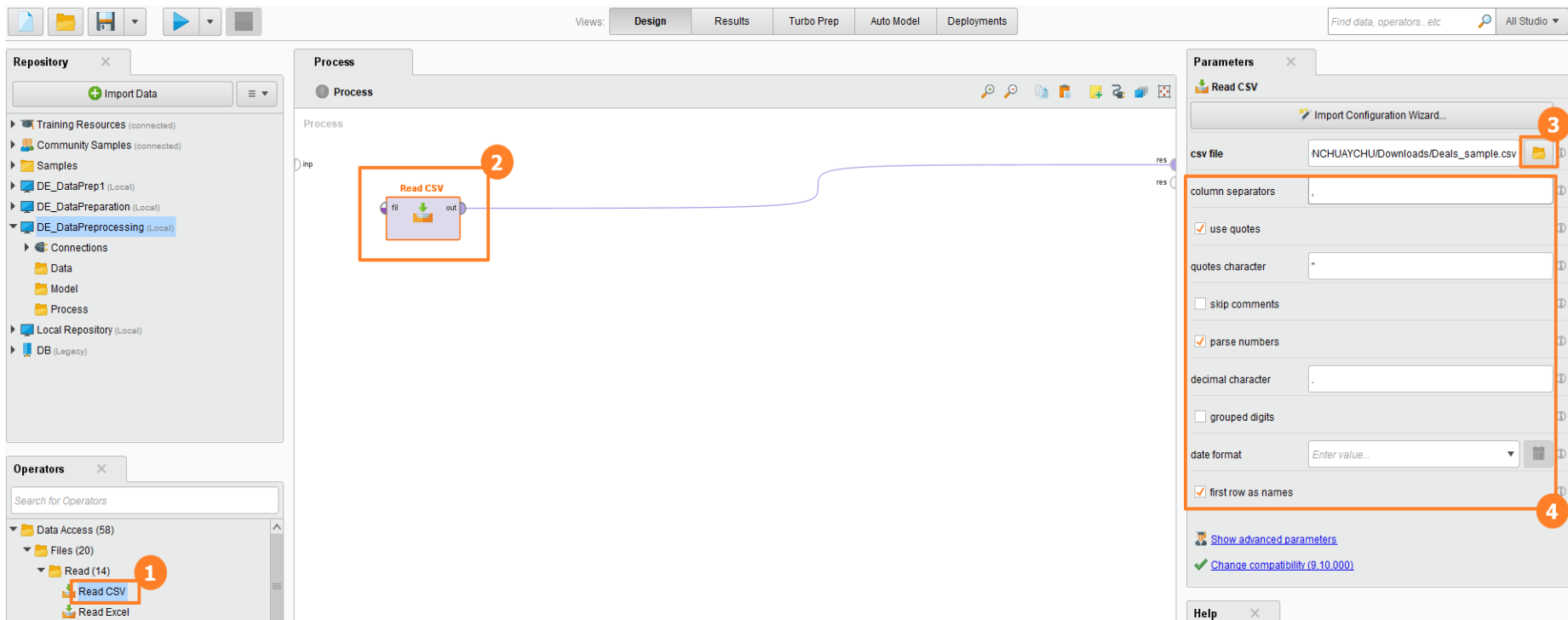
---

Training Course: Intermediate Data Engineering

# Workshop: Import/ Access Data

## Import Data from CSV

- แถบ **Operators** เลือก **Data Access > Files > Read > Read CSV** และลากมาวางในพื้นที่ **Process**
- กดปุ่ม  เพื่อเลือก **CSV file** ที่ต้องการนำเข้าข้อมูล
- กำหนดค่า **Parameters** ดังรูป



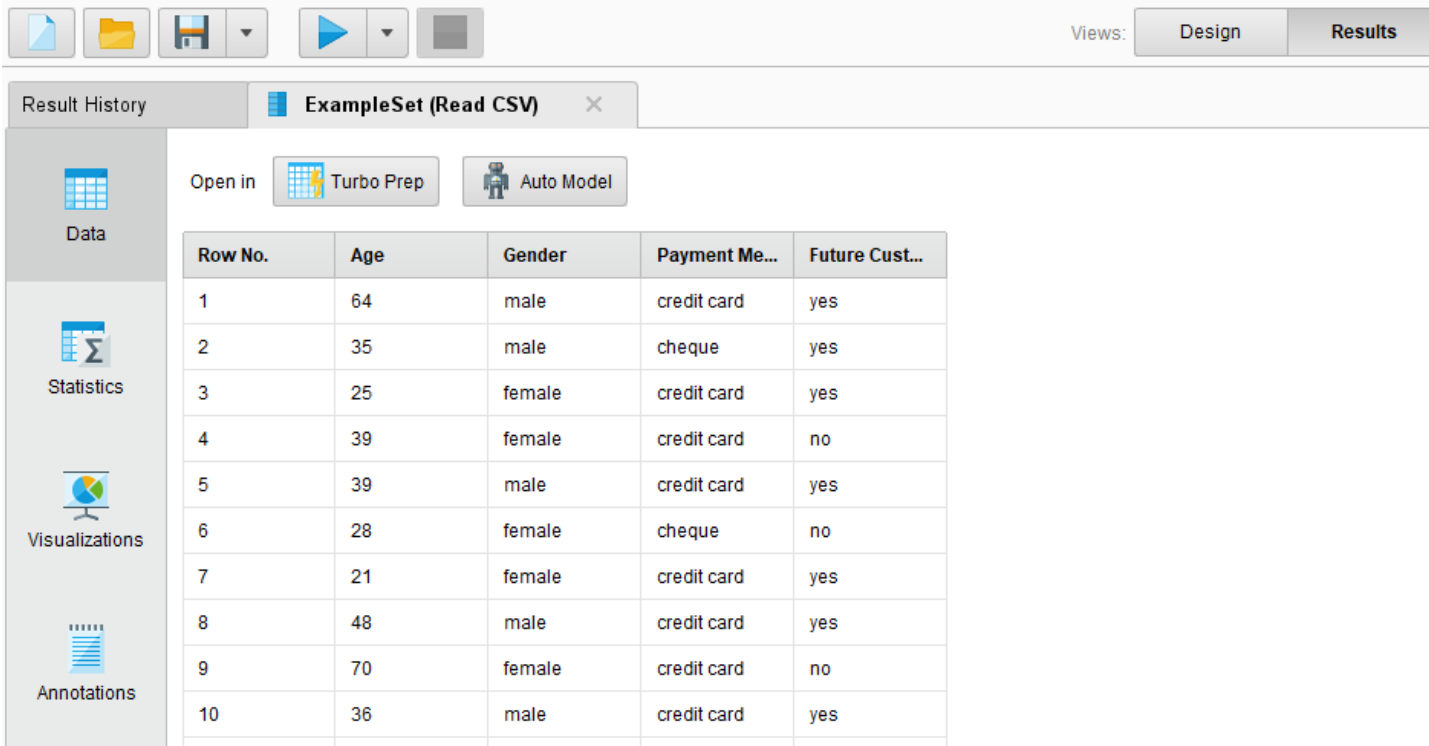
The screenshot displays the Data Engineering Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources. The 'Read CSV' operator is highlighted in the 'Data Access > Files > Read' path, marked with a red circle and the number 1.
- Process:** The central workspace where the 'Read CSV' operator is placed and connected to the data source, marked with a red circle and the number 2.
- Parameters:** A panel on the right showing the configuration for the 'Read CSV' operator. The 'csv file' field is set to 'NCHUAYCHU/Downloads/Deals\_sample.csv', marked with a red circle and the number 3. The 'column separators' field is set to a comma. The 'use quotes' checkbox is checked. The 'quotes character' field is set to a double quote. The 'skip comments' checkbox is unchecked. The 'parse numbers' checkbox is checked. The 'decimal character' field is set to a period. The 'grouped digits' checkbox is unchecked. The 'date format' field is set to 'Enter value...'. The 'first row as names' checkbox is checked, marked with a red circle and the number 4.

# Workshop: Import/ Access Data

## Import Data from CSV

- กดปุ่ม **Run**  แล้วดูผลลัพธ์ในแถบ **Results** เลือกเมนู **Data** เพื่อดูรายการข้อมูล



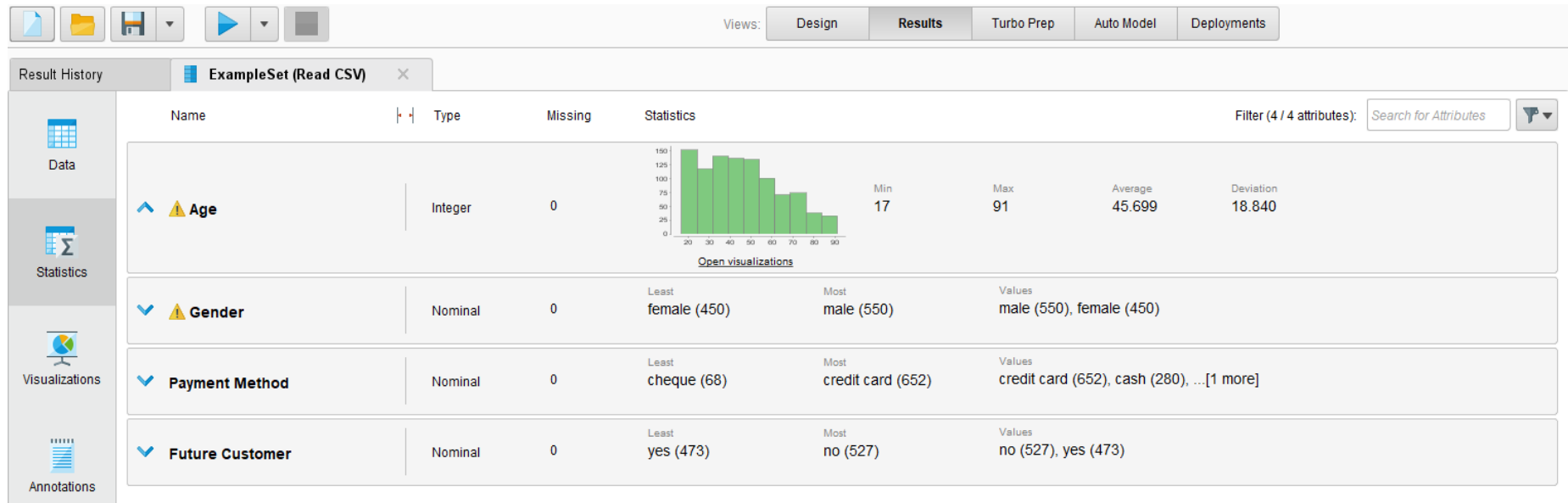
The screenshot shows the software interface with the **Results** view selected. The **ExampleSet (Read CSV)** tab is active. The **Data** menu is selected in the left sidebar. The table displays 10 rows of data with the following columns: Row No., Age, Gender, Payment Me..., and Future Cust....

Row No.	Age	Gender	Payment Me...	Future Cust...
1	64	male	credit card	yes
2	35	male	cheque	yes
3	25	female	credit card	yes
4	39	female	credit card	no
5	39	male	credit card	yes
6	28	female	cheque	no
7	21	female	credit card	yes
8	48	male	credit card	yes
9	70	female	credit card	no
10	36	male	credit card	yes

# Workshop: Import/ Access Data

## Import Data from CSV

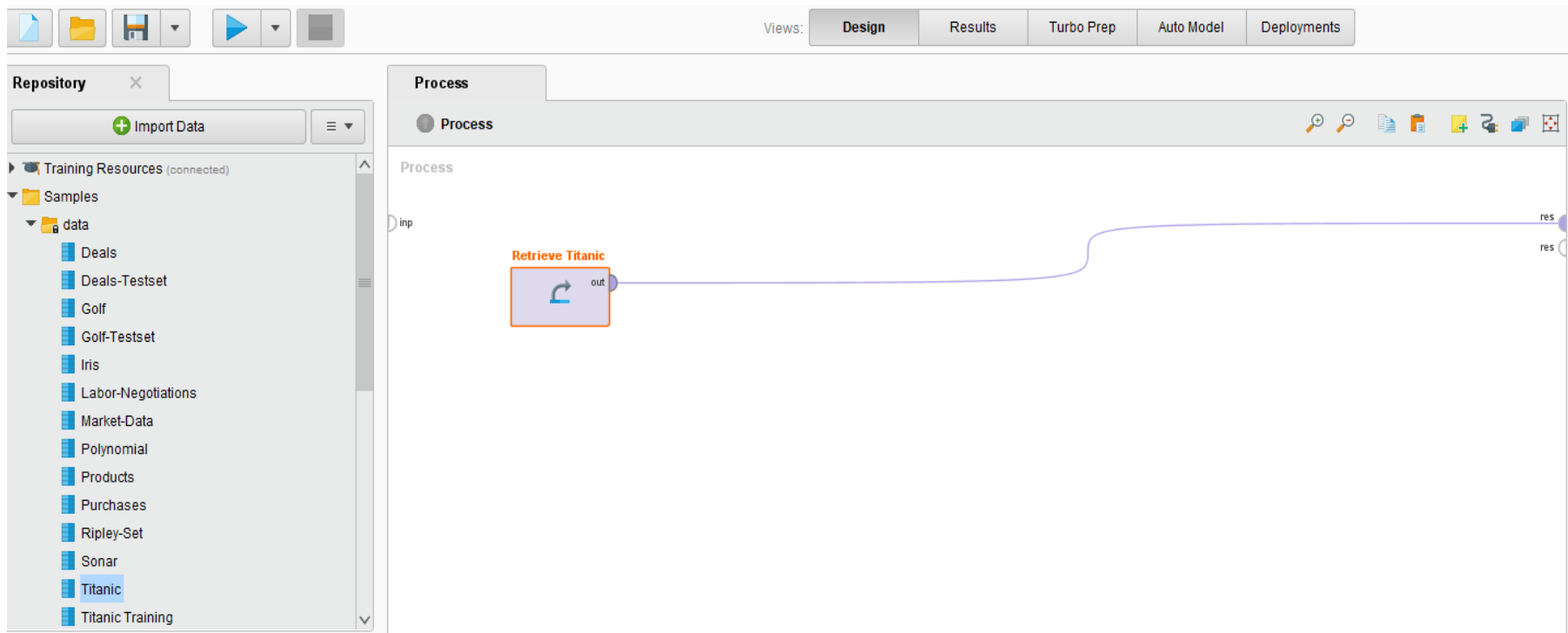
- เลือกเมนู **Statistics** เพื่อดูสถิติของข้อมูล



# Workshop: Import/ Access Data

## Import Dataset in RapidMiner

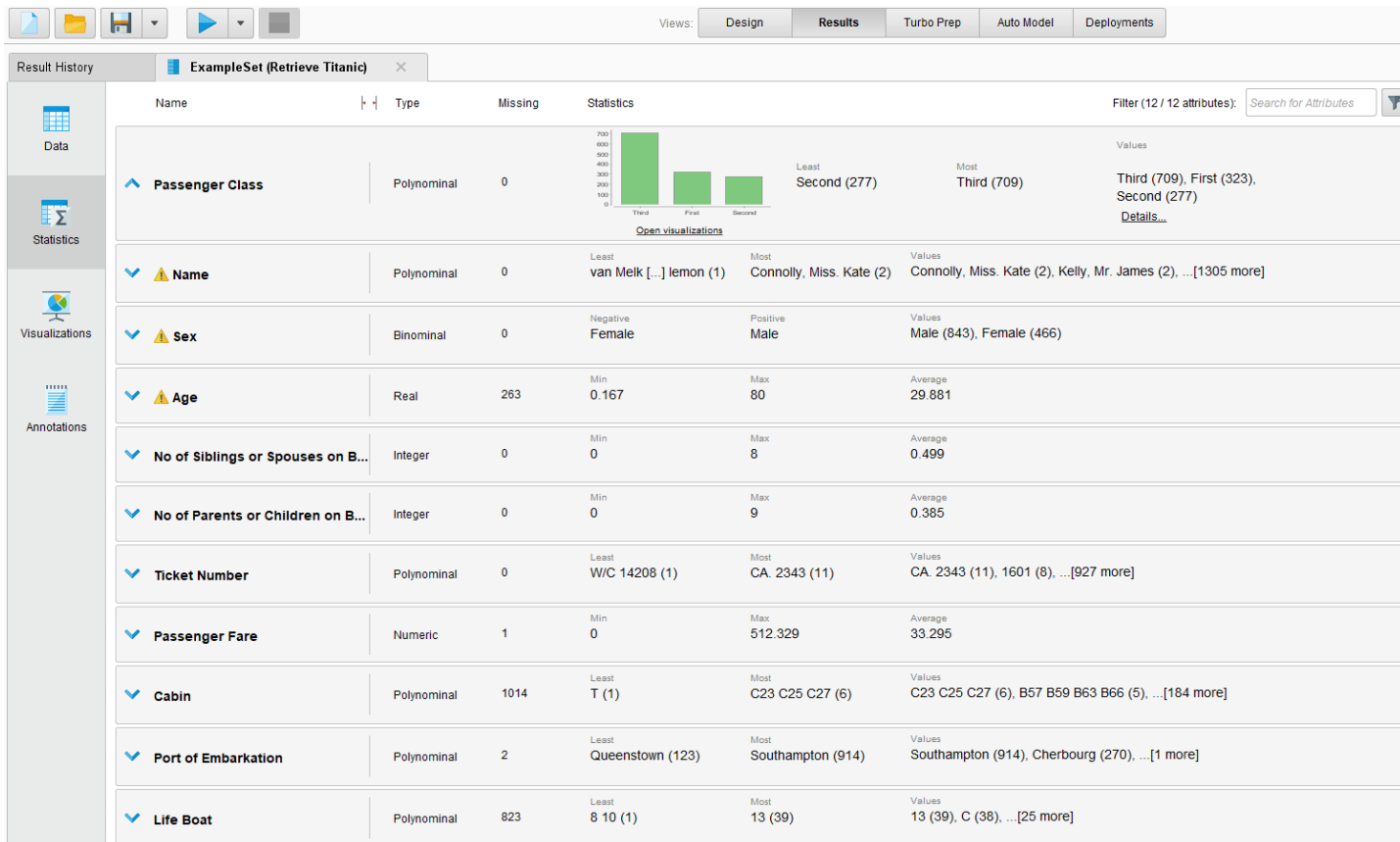
- เลือกข้อมูลจากไฟล์เดอร์ Samples > data > **Titanic** แล้วลากมาวางในพื้นที่ Process

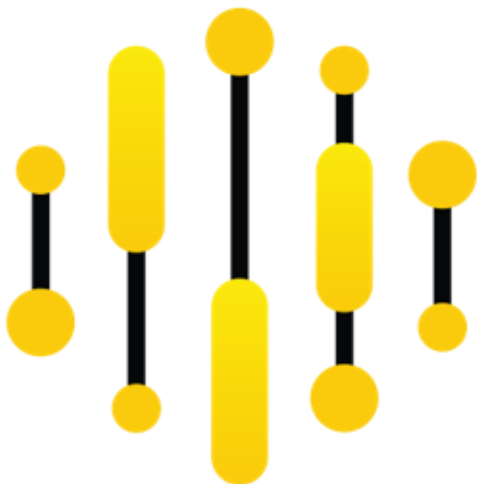


# Workshop: Import/ Access Data

## Import Dataset in RapidMiner

- กดปุ่ม Run  แล้วดูผลลัพธ์ในแถบ Results เลือกเมนู **Statistics** เพื่อดูสถิติของข้อมูล





## Workshop: Data cleaning

---

Training Course: Intermediate Data Engineering



# Workshop: Data cleaning

## Handle Missing data - Remove Missing Values

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Declare Missing Value** แล้วลากมาวางในพื้นที่ **Process** และกำหนดค่า **Parameters** ดังรูป

The screenshot displays the GBDi software interface with the following components:

- Repository:** A list of data sources on the left, including 'data', 'Deals', 'Deals-Testset', 'Golf', 'Golf-Testset', 'Iris', 'Labor-Negotiations', 'Market-Data', 'Polynomial', 'Products', 'Purchases', 'Ripley-Set', 'Sonar', 'Titanic', 'Titanic Training', 'Titanic Unlabeled', and 'Transactions'.
- Operators:** A search bar at the top contains the text 'declare'. Below it, a list of operators is shown, with 'Declare Missing Value' highlighted under the 'Missing (1)' category. A red circle with the number '1' is placed next to this operator.
- Process:** The central workspace shows a workflow. It starts with a 'Retrieve Titanic' operator, followed by a 'Declare Missing Value' operator (highlighted with a red box). A red circle with the number '2' is placed next to the 'Declare Missing Value' operator. The workflow is connected by a line from the 'out' port of 'Retrieve Titanic' to the 'in' port of 'Declare Missing Value'.
- Parameters:** A panel on the right shows the configuration for the 'Declare Missing Value' operator. The 'attribute filter type' is set to 'all'. There are checkboxes for 'invert selection' and 'include special attributes', both of which are unchecked. The 'mode' is set to 'nominal'. The 'nominal value' is set to '?'. A red circle with the number '2' is placed next to the 'Parameters' panel.

# Workshop: Data cleaning

## Handle Missing data - Remove Missing Values

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Filter Examples** แล้วลากมาวางในพื้นที่ Process และกดปุ่ม Show advanced parameters

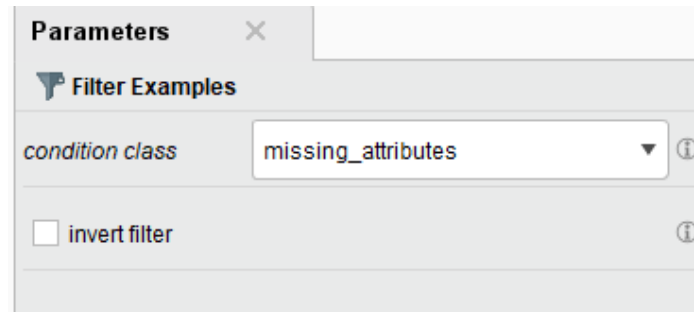
The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A list of datasets on the left, with 'Titanic' selected.
- Operators:** A panel on the bottom left showing a search for 'filter'. Under the 'Filter' category, 'Filter Examples' is highlighted with a red box and a circled '1'.
- Process:** The central workspace showing a workflow: 'Retrieve Titanic' → 'Declare Missing Values' → 'Filter Examples'. The 'Filter Examples' operator is highlighted with a red box.
- Parameters:** A panel on the right for the 'Filter Examples' operator. It shows an 'invert filter' checkbox. A red box highlights the 'Show advanced parameters' button, which is circled with a '2'.
- Help:** A panel at the bottom right showing the 'Filter Examples' operator's description and tags.

# Workshop: Data cleaning


## Handle Missing data - Remove Missing Values

- ที่แถบ Parameters กำหนด condition class เป็น missing\_attributes ดังรูป



# Workshop: Data cleaning

## Handle Missing data - Remove Missing Values

- เมื่อกดปุ่ม Run  และดูข้อมูลจะปรากฏข้อมูลที่มี Missing Values ดังรูป

Views: Design Results Turbo Prep Auto Model Deployments

Result History ExampleSet (Filter Examples) x

Open in Turbo Prep Auto Model

Filter (1,129 / 1,129 examples): all

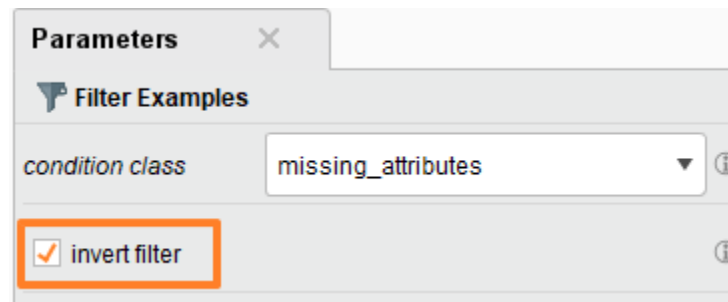
Row No.	Passenger ...	Name	Sex	Age	No of Sibling...	No of Parent...	Ticket Numb...	Passenger F...	Cabin	Port of Emb...	Life Boat	Survived
1	First	Allison, Miss. ...	Female	2	1	2	113781	151.550	C22 C26	Southampton	?	No
2	First	Allison, Mr. H...	Male	30	1	2	113781	151.550	C22 C26	Southampton	?	No
3	First	Allison, Mrs. ...	Female	25	1	2	113781	151.550	C22 C26	Southampton	?	No
4	First	Andrews, Mr. ...	Male	39	0	0	112050	0	A36	Southampton	?	No
5	First	Artagaveytia, ...	Male	71	0	0	PC 17609	49.504	?	Cherbourg	?	No
6	First	Astor, Col. Jo...	Male	47	1	0	PC 17757	227.525	C62 C64	Cherbourg	?	No
7	First	Barber, Miss. ...	Female	26	0	0	19877	78.850	?	Southampton	6	Yes
8	First	Baumann, Mr...	Male	?	0	0	PC 17318	25.925	?	Southampton	?	No
9	First	Baxter, Mr. Qu...	Male	24	0	1	PC 17558	247.521	B58 B60	Cherbourg	?	No
10	First	Bidois, Miss. ...	Female	42	0	0	PC 17757	227.525	?	Cherbourg	4	Yes
11	First	Birnbaum, Mr...	Male	25	0	0	13905	26	?	Cherbourg	?	No
12	First	Blackwell, Mr...	Male	45	0	0	113784	35.500	T	Southampton	?	No
13	First	Borebank, Mr...	Male	42	0	0	110489	26.550	D22	Southampton	?	No
14	First	Bowen, Miss...	Female	45	0	0	PC 17608	262.375	?	Cherbourg	4	Yes
15	First	Bradley, Mr. G...	Male	?	0	0	111427	26.550	?	Southampton	9	Yes
16	First	Brady, Mr. Jo...	Male	41	0	0	113054	30.500	A21	Southampton	?	No
17	First	Brandeis, Mr. ...	Male	48	0	0	PC 17591	50.496	B10	Cherbourg	?	No
18	First	Brewe, Dr. Art...	Male	?	0	0	112379	39.600	?	Cherbourg	?	No
19	First	Butt, Major, Ar...	Male	45	0	0	113050	26.550	B38	Southampton	?	No
20	First	Cairns, Mr. Al...	Male	?	0	0	113798	31	?	Southampton	?	No
21	First	Candee, Mrs....	Female	53	0	0	PC 17606	27.446	?	Cherbourg	6	Yes
22	First	Carlsson, Mr...	Male	33	0	0	695	5	B51 B53 B55	Southampton	?	No
23	First	Carrau, Mr. Fr...	Male	28	0	0	113059	47.100	?	Southampton	?	No
24	First	Carrau, Mr. J...	Male	17	0	0	113059	47.100	?	Southampton	?	No
25	First	Case, Mr. Ho...	Male	49	0	0	19924	26	?	Southampton	?	No

ExampleSet (1,129 examples, 0 special attributes, 12 regular attributes)

# Workshop: Data cleaning

## Handle Missing data - Remove Missing Values

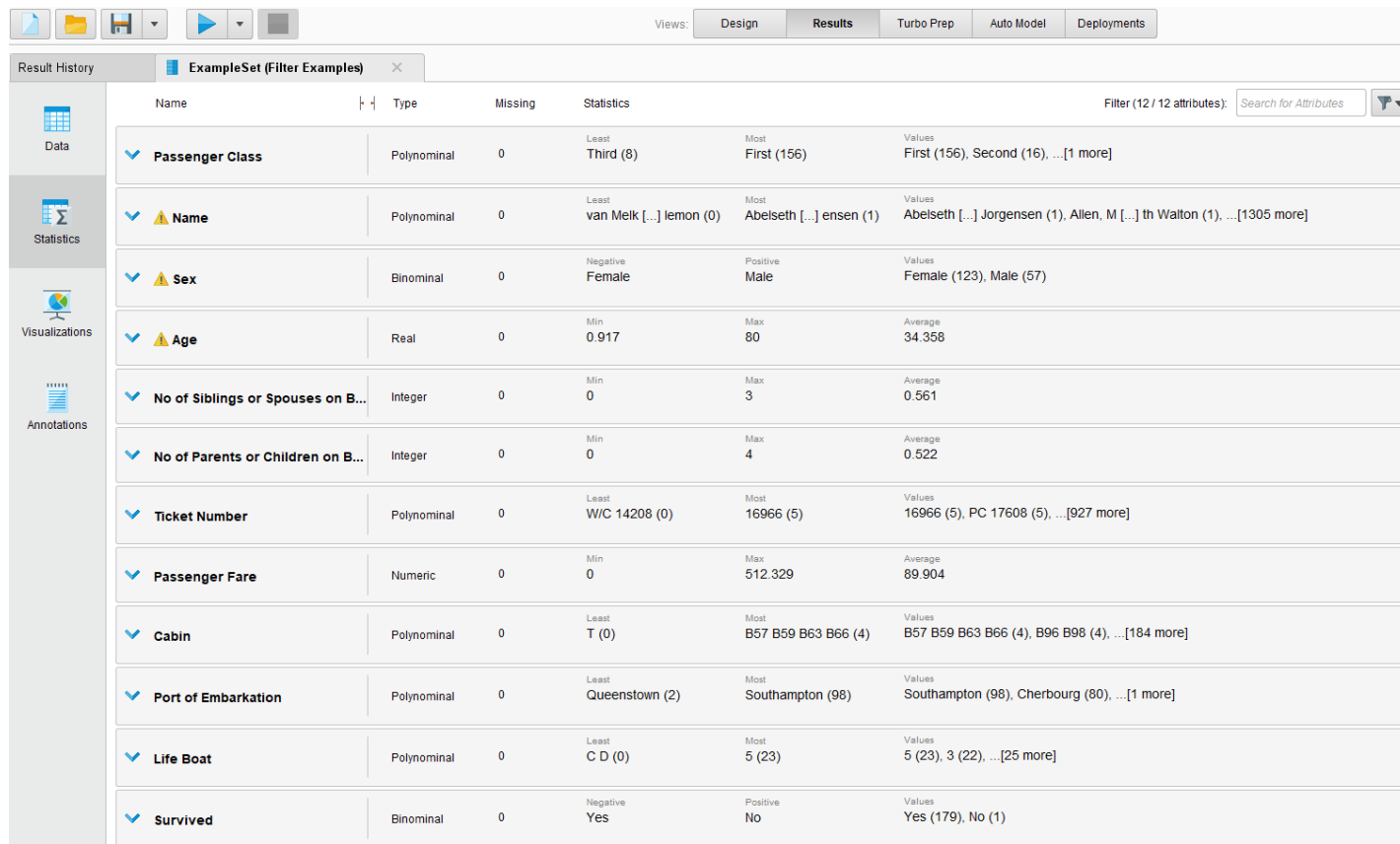
- ที่แถบ Parameters คลิก invert filter



# Workshop: Data cleaning

## Handle Missing data - Remove Missing Values

- เมื่อกดปุ่ม Run แล้วดูผลลัพธ์ใน **Statistics** จะปรากฏข้อมูลที่ไม่ใช่ Missing Values ดังรูป



Views: Design Results Turbo Prep Auto Model Deployments

Result History ExampleSet (Filter Examples) x

Filter (12 / 12 attributes): Search for Attributes

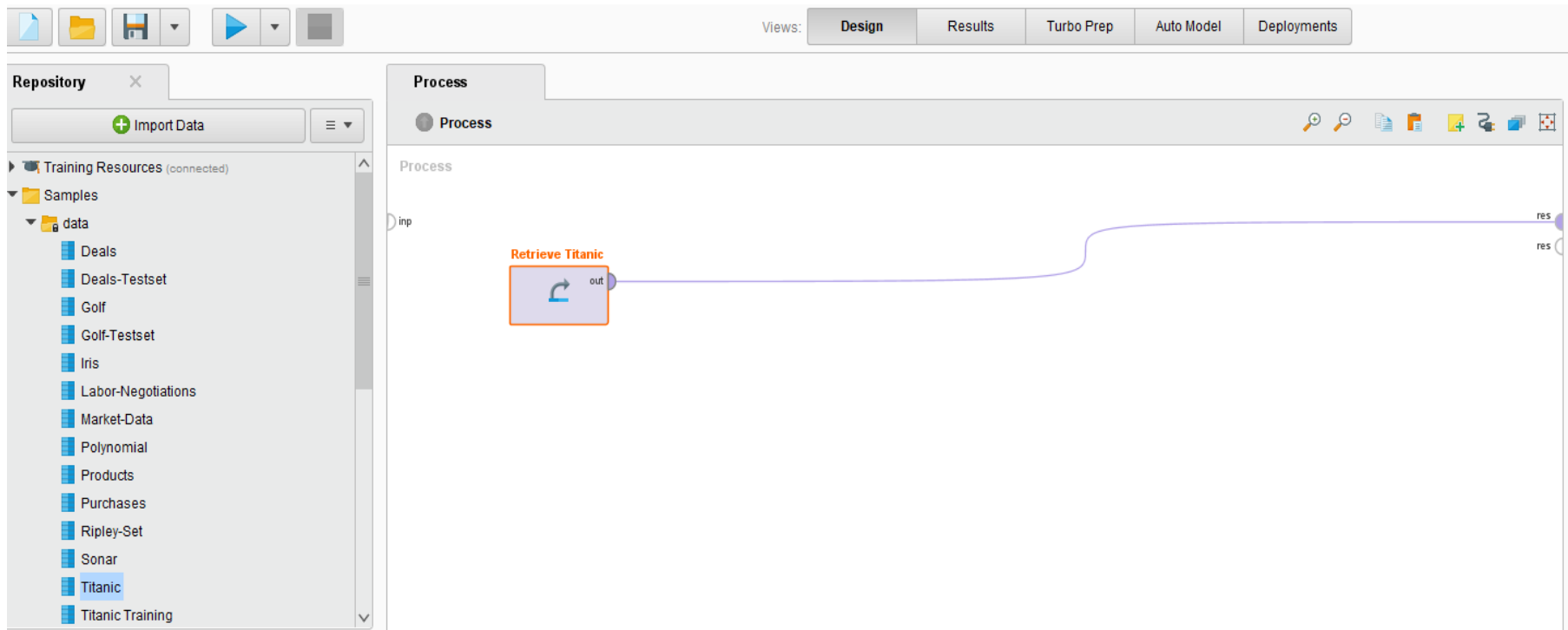
Name	Type	Missing	Statistics		
Passenger Class	Polynomial	0	Least Third (8)	Most First (156)	Values First (156), Second (16), ...[1 more]
Name	Polynomial	0	Least van Melk [...] lemon (0)	Most Abelseth [...] ensen (1)	Values Abelseth [...] Jorgensen (1), Allen, M [...] th Walton (1), ...[1305 more]
Sex	Binominal	0	Negative Female	Positive Male	Values Female (123), Male (57)
Age	Real	0	Min 0.917	Max 80	Average 34.358
No of Siblings or Spouses on B...	Integer	0	Min 0	Max 3	Average 0.561
No of Parents or Children on B...	Integer	0	Min 0	Max 4	Average 0.522
Ticket Number	Polynomial	0	Least W/C 14208 (0)	Most 16966 (5)	Values 16966 (5), PC 17608 (5), ...[927 more]
Passenger Fare	Numeric	0	Min 0	Max 512.329	Average 89.904
Cabin	Polynomial	0	Least T (0)	Most B57 B59 B63 B66 (4)	Values B57 B59 B63 B66 (4), B96 B98 (4), ...[184 more]
Port of Embarkation	Polynomial	0	Least Queenstown (2)	Most Southampton (98)	Values Southampton (98), Cherbourg (80), ...[1 more]
Life Boat	Polynomial	0	Least C D (0)	Most 5 (23)	Values 5 (23), 3 (22), ...[25 more]
Survived	Binominal	0	Negative Yes	Positive No	Values Yes (179), No (1)

# Workshop: Data cleaning

## Handle Missing data - Replace Missing Values

- เลือกข้อมูลจากโฟลเดอร์ Samples > data > **Titanic** แล้วลากมาวางในพื้นที่

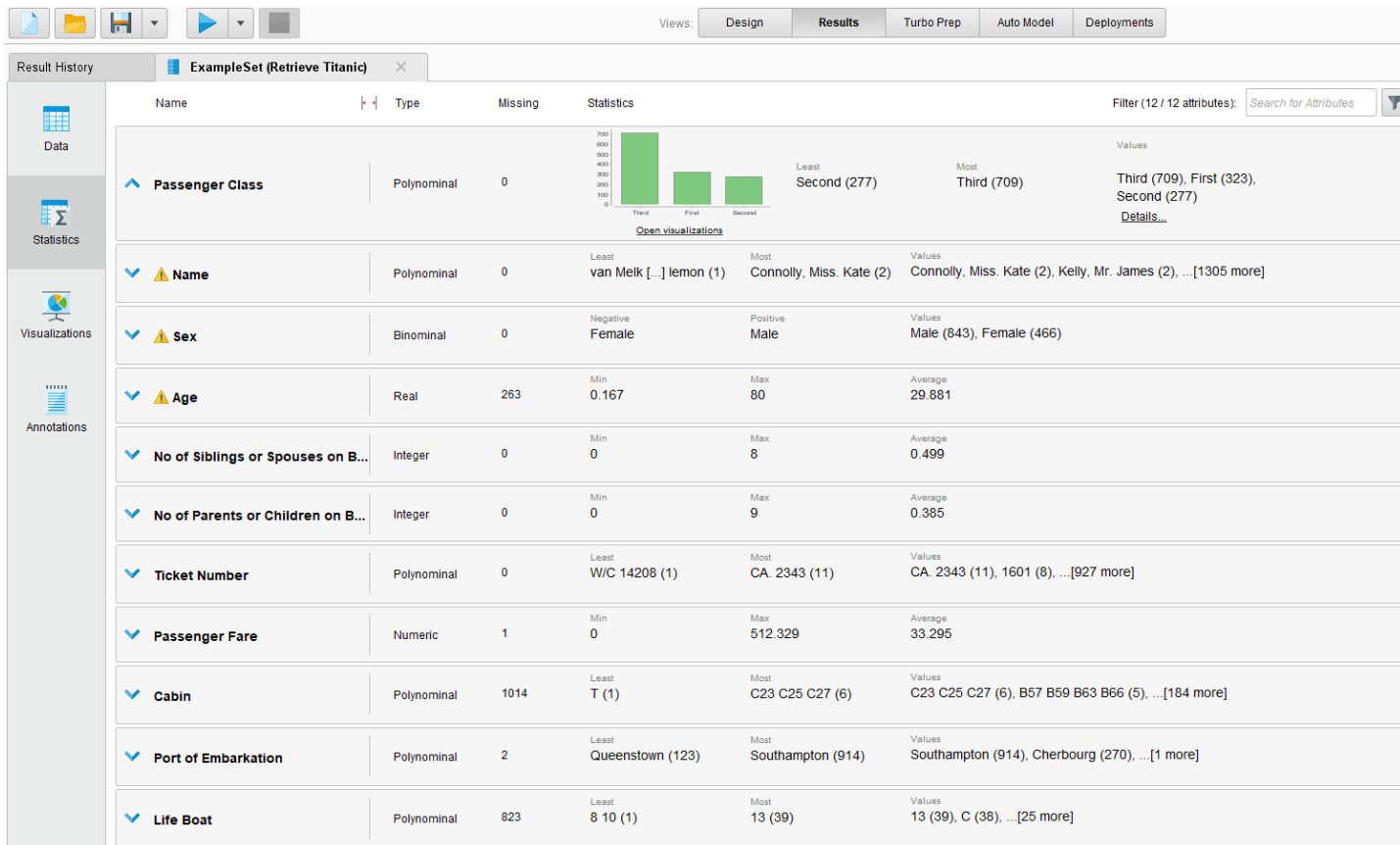
### Process



# Workshop: Data cleaning

## Handle Missing data - Replace Missing Values

- กดปุ่ม Run แล้วดูผลลัพธ์ในแถบ Results เลือกเมนู **Statistics** เพื่อดูสถิติของข้อมูล





# Workshop: Data cleaning

## Handle Missing data - Replace Missing Values

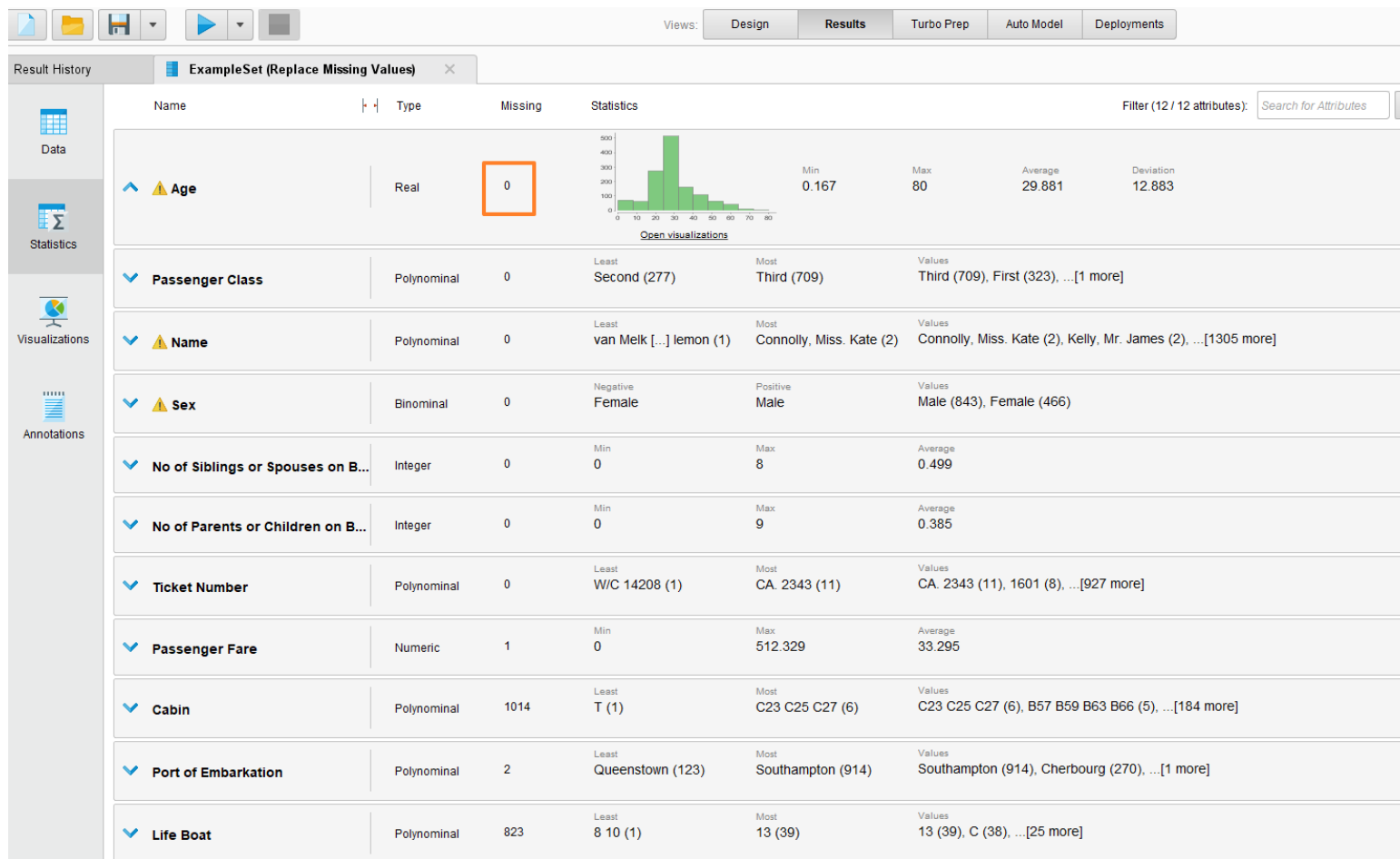
- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Replace Missing Values** แล้วกำหนดค่า Parameters ดังรูป

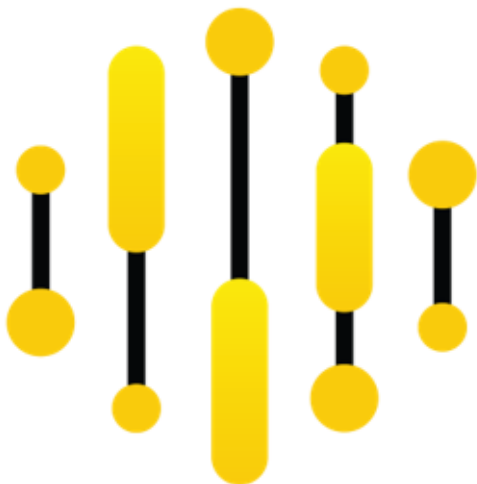
The screenshot displays the RapidMiner Studio interface. On the left, the 'Repository' pane shows a list of training resources, with 'Titanic' selected. Below it, the 'Operators' pane shows a search bar and a list of operators, with 'Replace Missing Values' highlighted under the 'Missing' category (marked with a red circle 1). The central 'Process' pane shows a workflow diagram with a 'Retrieve Titanic' operator connected to a 'Replace Missing Values' operator. The 'Parameters' pane on the right shows the configuration for the 'Replace Missing Values' operator (marked with a red circle 2). The parameters are: 'attribute filter type' set to 'single', 'attribute' set to 'Age', 'invert selection' and 'include special attributes' are unchecked, and 'default' is set to 'average'. At the bottom, a 'Help' pane shows the 'Replace Missing Values' operator documentation.

# Workshop: Data cleaning

## Handle Missing data - Replace Missing Values

- กดปุ่ม Run แล้วดูค่าสถิติข้อมูล





## Workshop: Store data

---

Training Course: Intermediate Data Engineering

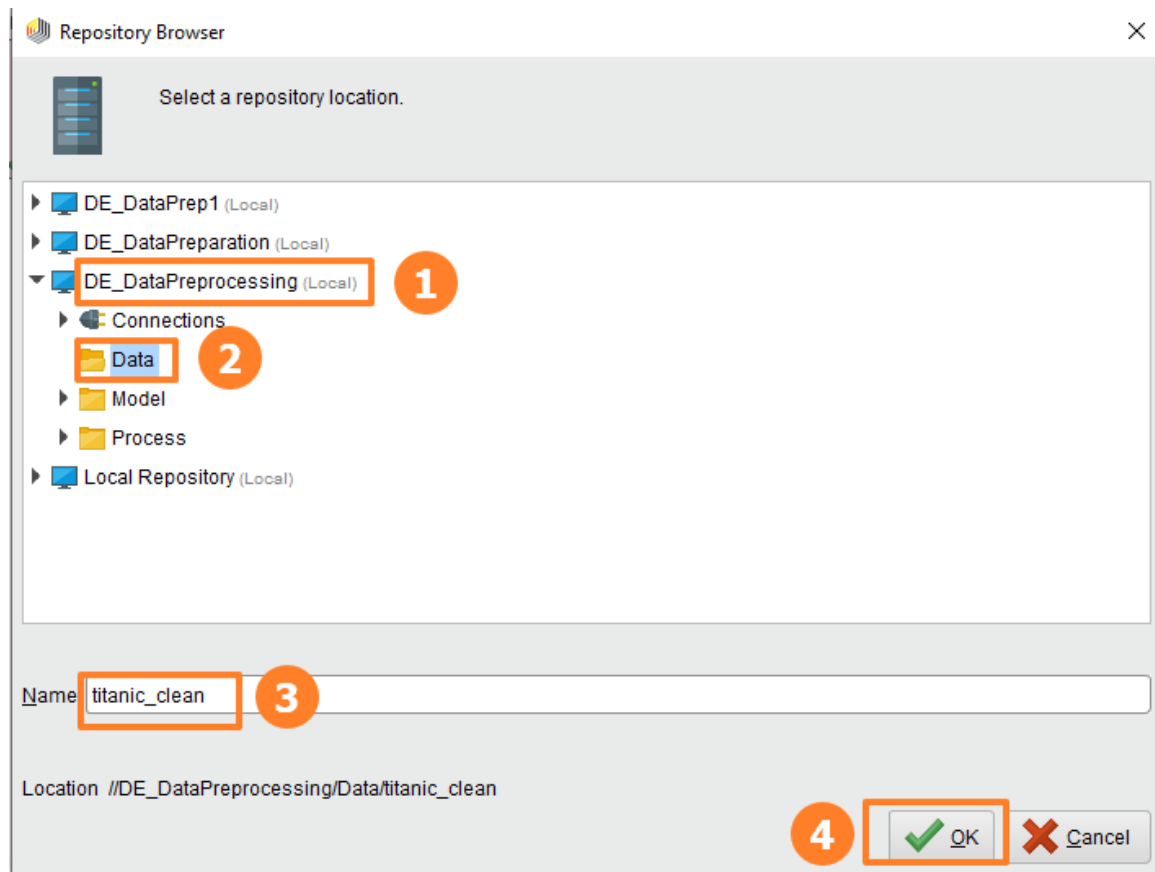
# Workshop: Store data

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Store** แล้วกดปุ่มรูปไฟล์เดอร์ในแถบ Parameters เพื่อกำหนดค่าการบันทึกข้อมูล

The screenshot displays the RapidMiner Studio interface during a data storage workflow. The main canvas shows a process flow: 'Retrieve Titanic' (green checkmark) connects to 'Replace Missing Values' (green checkmark), which then connects to the 'Store' operator (orange box with a warning icon). The 'Store' operator has two output ports labeled 'res'. On the left, the 'Repository' pane shows a tree structure with 'Samples' > 'data' > 'Titanic' selected. Below it, the 'Operators' pane shows 'Data Access (58)' with 'Store' highlighted and circled in red, labeled with a red '1'. On the right, the 'Parameters' pane for the 'Store' operator shows 'repository entry' with a file folder icon circled in red and labeled with a red '2'. At the bottom right, a 'Help' pane for the 'Store' operator is visible, showing 'RapidMiner Studio Core' and tags: 'Save, Export, Write, Datasets, Repository, Data Access'.

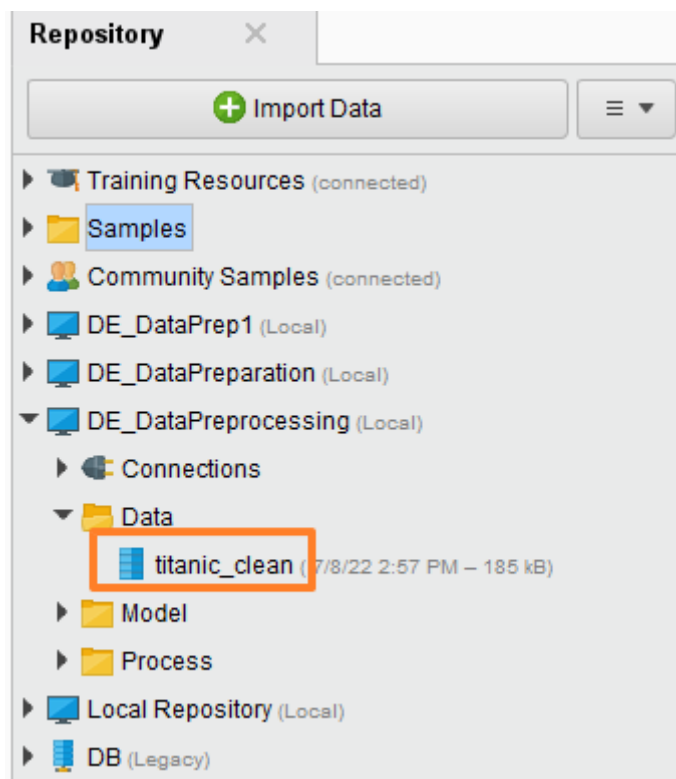
# Workshop: Store data

- คลิกเลือก Repository: **DE\_DataPreprocessing**
- คลิกเลือกโฟลเดอร์ **Data**
- ตั้งชื่อไฟล์เป็น **titanic\_clean** และกดปุ่ม **OK**



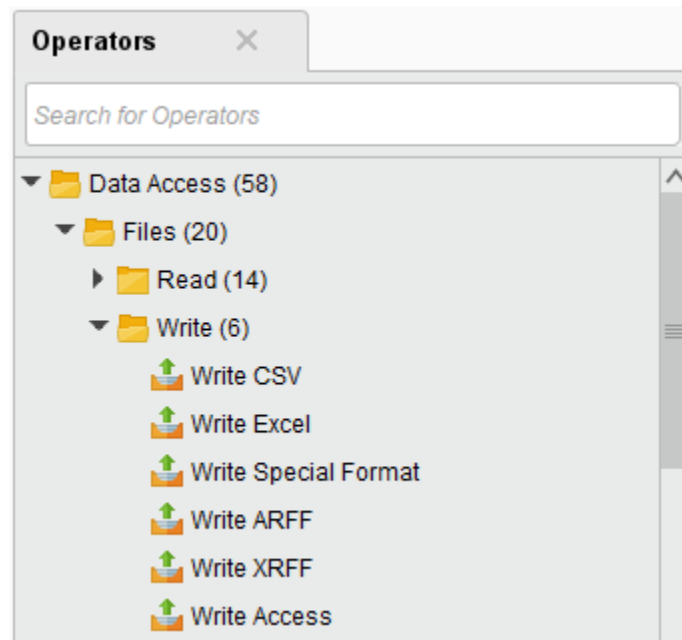
# Workshop: Store data

- กดปุ่ม Run  และดูที่แถบ Repository จะปรากฏไฟล์ชื่อ **titanic\_clean** ในโฟลเดอร์ Data ดังรูป



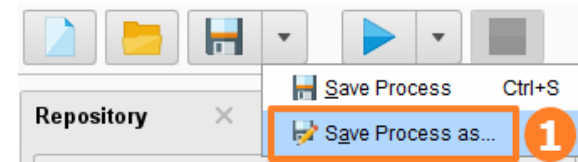
# Workshop: Store data

- หากต้องการบันทึกเป็นไฟล์ประเภทอื่น ๆ ให้ใช้ **Operators** ในกลุ่มของ **Write** แล้วเลือกตามประเภทของไฟล์ที่ต้องการบันทึก เช่น **Write CSV** ใช้บันทึกไฟล์เป็นรูปแบบของ **CSV**

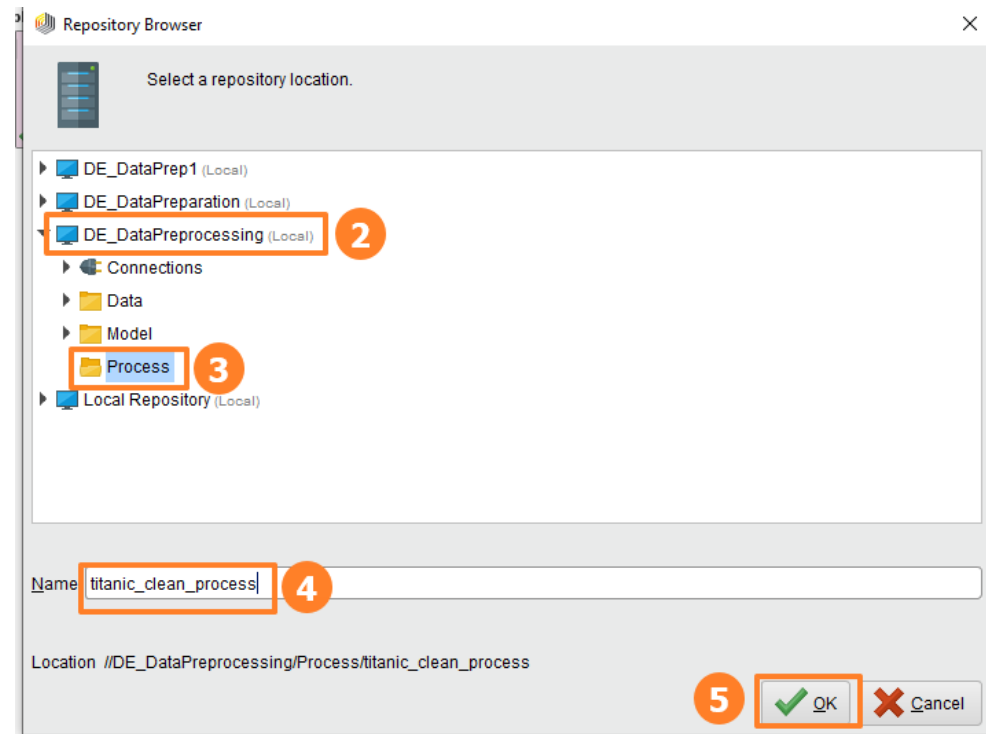


# Workshop: Store data

- คลิกเลือก Save Process as...



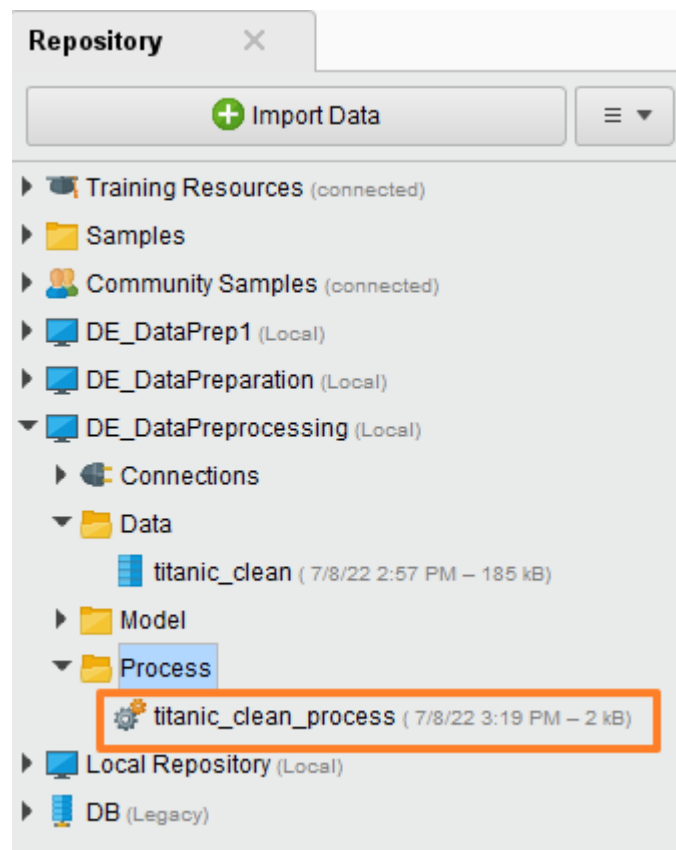
- คลิกเลือก Repository: DE\_DataPreprocessing
- คลิกเลือกโฟลเดอร์ Process
- ตั้งชื่อไฟล์เป็น titanic\_clean\_process และกดปุ่ม OK

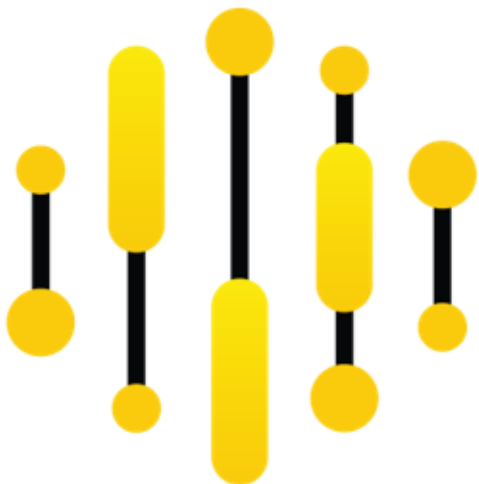




# Workshop: Store data

- ที่แถบ **Repository** จะปรากฏไฟล์ดังรูป





## Workshop: Data cleaning (Cont.)

---

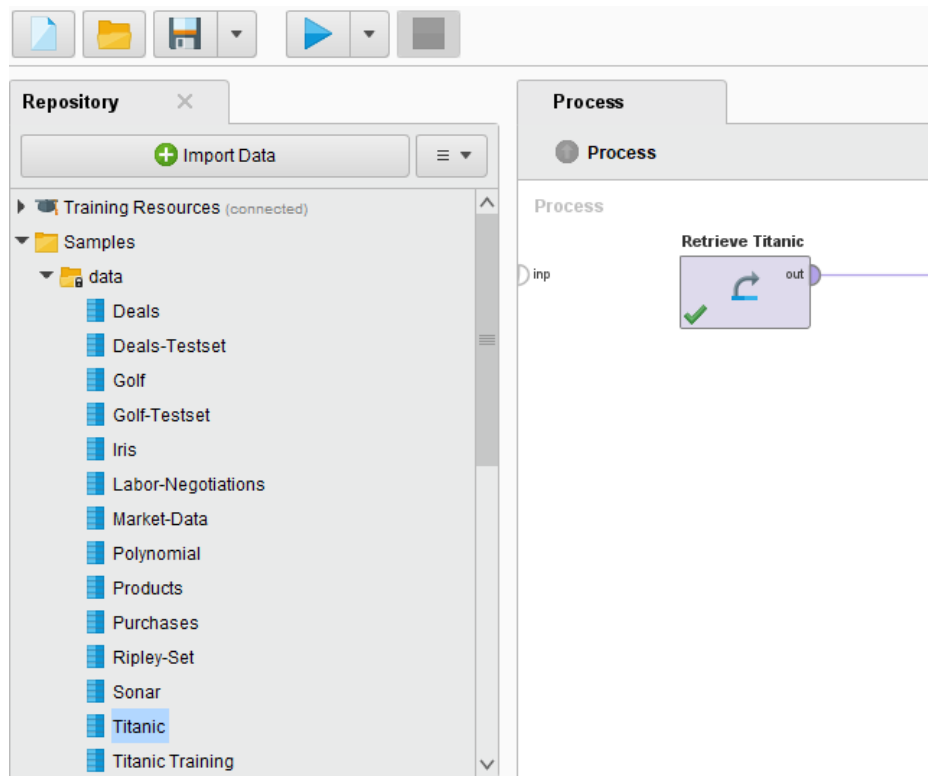
Training Course: Intermediate Data Engineering

# Workshop: Data cleaning (Cont.)

## Filtering

- เลือกข้อมูลจากโฟลเดอร์ Samples > data > Titanic แล้วลากมาวางในพื้นที่

## Process



# Workshop: Data cleaning (Cont.)

## Filtering

- ที่แถบ Operators เลือก Operators ชื่อ **Filter Examples** แล้วลากมาวางในพื้นที่ Process จากนั้นกดปุ่ม **Add filters...**

The screenshot displays the RapidMiner Studio interface during a data cleaning workshop. The 'Repository' pane on the left lists various datasets, with 'Titanic' selected. The 'Operators' pane at the bottom left shows a list of operators, with 'Filter Examples' highlighted by a red box and a red circle with the number 1. The 'Process' area in the center shows a workflow with 'Retrieve Titanic' and 'Filter Examples' operators. The 'Parameters' pane on the right shows the configuration for the 'Filter Examples' operator, with the 'Add Filters...' button highlighted by a red box and a red circle with the number 2.

# Workshop: Data cleaning (Cont.)

## Filtering

- กำหนดค่า **Parameters** ดังรูปและกดปุ่ม **OK**

Create Filters: filters

Create Filters: **filters**  
Defines the list of filters to apply.

Age z 1

☒ Match all ☐ Match any ☒ Preselect comparators

Add Entry OK Cancel

# Workshop: Data cleaning (Cont.)

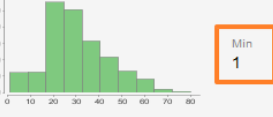
## Filtering

- กดปุ่ม **Run** และดูค่าสถิติของข้อมูล

Views: Design Results Turbo Prep Auto Model Deployments

Result History: ExampleSet (Filter Examples) ExampleSet (//Samples/data/Titanic)

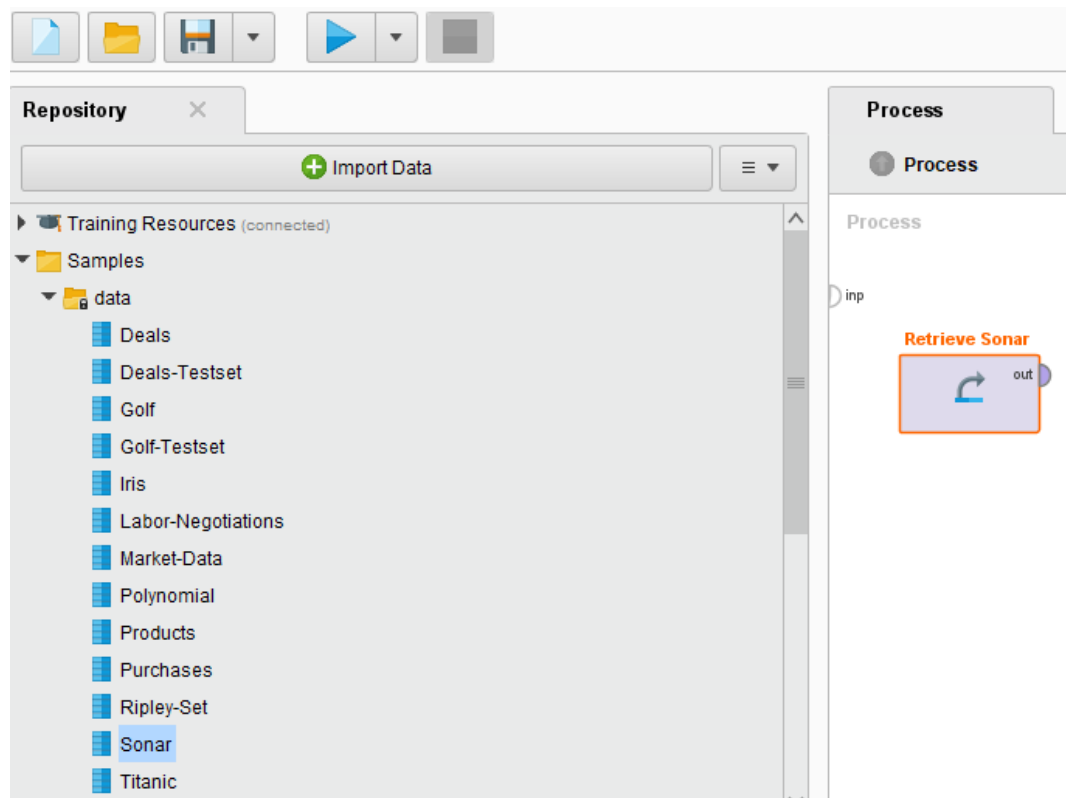
Filter (12 / 12 attributes): Search for Attributes

Name	Type	Missing	Statistics		Values
Passenger Class	Polynomial	0	Least Second (257)	Most Third (494)	Third (494), First (283), ...[1 more]
Name	Polynomial	0	Least van Melk [...] lemon (0)	Most Connolly, Miss. Kate (2)	Connolly, Miss. Kate (2), Kelly, Mr. James (2), ...[1305 more]
Sex	Binomial	0	Negative Female	Positive Male	Male (650), Female (384)
Age	Real	0			Max: 80 Average: 30.220 Deviation: 14.147
No of Siblings or Spouses on B...	Integer	0	Min 0	Max 8	Average 0.499
No of Parents or Children on B...	Integer	0	Min 0	Max 6	Average 0.409
Ticket Number	Polynomial	0	Least W./C. 6609 (0)	Most CA 2144 (8)	Values CA 2144 (8), 3101295 (7), ...[927 more]
Passenger Fare	Numeric	1	Min 0	Max 512.329	Average 36.777
Cabin	Polynomial	763	Least F38 (0)	Most C23 C25 C27 (6)	Values C23 C25 C27 (6), B57 B59 B63 B66 (5), ...[184 more]
Port of Embarkation	Polynomial	2	Least Queenstown (50)	Most Southampton (773)	Values Southampton (773), Cherbourg (209), ...[1 more]

# Workshop: Data cleaning (Cont.)

## Removing outliers

- เลือกข้อมูลจากโฟลเดอร์ Samples > data > **Sonar** แล้วลากมาวางในพื้นที่  
Process



# Workshop: Data cleaning (Cont.)

## Removing outliers

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Select Attributes** แล้วไปที่แถบ Parameters และกำหนด attributes filter type เป็น subset แล้วคลิก Select Attribute... ดังรูป

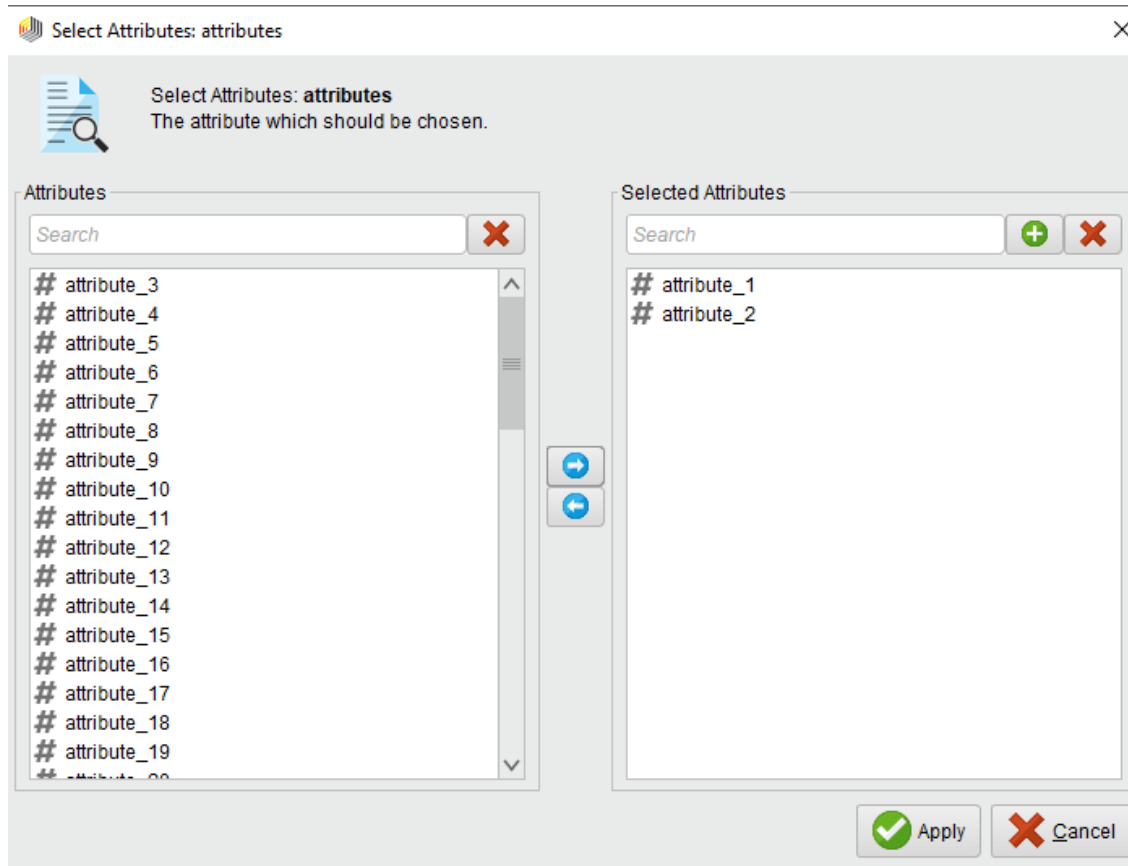
The screenshot displays the RapidMiner Studio interface with three main panels: Repository, Process, and Parameters. In the Repository panel, the 'data' folder is expanded, showing various datasets like 'Deals', 'Golf', 'Iris', etc. In the Process panel, a workflow is shown with 'Retrieve Sonar' and 'Select Attributes' operators. The 'Select Attributes' operator is highlighted with a red box and a red circle labeled '1'. In the Parameters panel, the 'attribute filter type' is set to 'subset', and the 'attributes' field is highlighted with a red box and a red circle labeled '2'. The 'Select Attributes...' button is also visible. A 'Help' panel at the bottom right shows the 'Select Attributes' operator documentation.



# Workshop: Data cleaning (Cont.)

## Removing outliers

- คลิกเลือก **attribute\_1** และ **attribute\_2** ดังรูป และกดปุ่ม **Apply**



# Workshop: Data cleaning (Cont.)

## Removing outliers

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Detect Outlier (Distances)** ลากเส้นเชื่อมให้สมบูรณ์

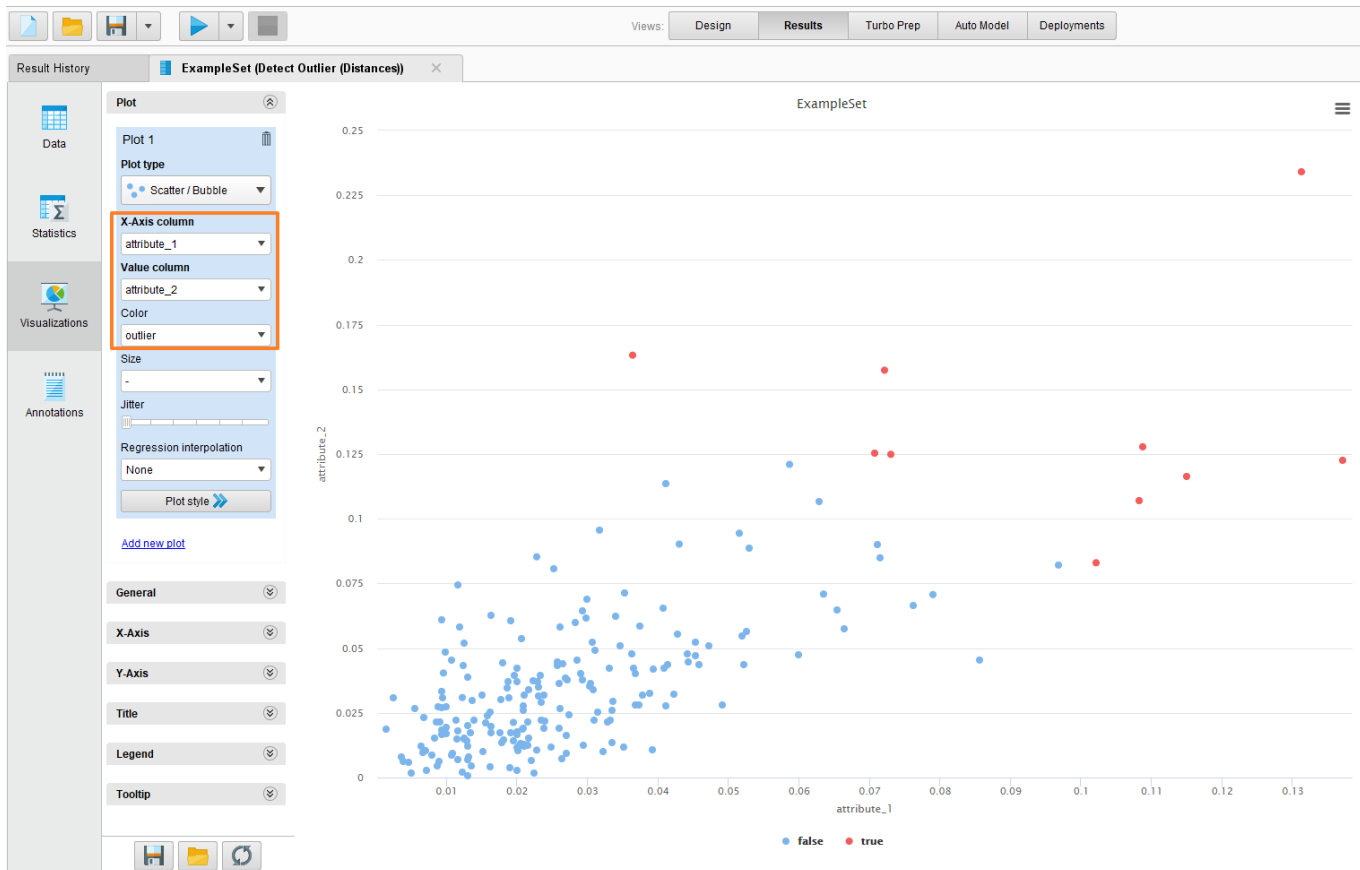
The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing training resources. The 'data' folder is expanded, listing datasets like Deals, Golf, Iris, and Sonar. The 'Sonar' dataset is selected.
- Operators:** A panel on the bottom left with a search bar containing 'outlier'. Under the 'Outliers (4)' category, the 'Detect Outlier (Distances)' operator is highlighted with a red box.
- Process:** The central workspace shows a workflow diagram. It starts with a 'Retrieve Sonar' operator, followed by a 'Select Attributes' operator, and then the 'Detect Outlier (Distances)' operator. The output of the 'Detect Outlier (Distances)' operator is connected to a 'res' output port.
- Parameters:** A panel on the right shows the configuration for the 'Detect Outlier (Distances)' operator. The parameters are: 'number of neighbors' set to 10, 'number of outliers' set to 10, and 'distance function' set to 'euclidian distance'.
- Help:** A panel at the bottom right provides documentation for the 'Detect Outlier (Distances)' operator, including tags like 'Anomaly', 'Anomalies', 'Detection', 'Removal', 'Remove', 'Cleansing', 'Quality', and 'Outliers'.

# Workshop: Data cleaning (Cont.)

## Removing outliers

- กดปุ่ม Run แล้วดูผลลัพธ์ในแถบ Results เมนู Visualizations



# Workshop: Data cleaning (Cont.)

## Removing outliers

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Filter Examples** แล้วไปที่แถบ Parameters กดปุ่ม **Add Filters...**

The screenshot displays the RapidMiner Studio interface during a data cleaning workflow. The main canvas shows a process flow starting with 'Retrieve Sonar', followed by 'Select Attributes', 'Detect Outlier (Distance)', and finally 'Filter Examples'. The 'Filter Examples' operator is highlighted with an orange box and a red circle labeled '2'. Below the main canvas, the 'Operators' panel is visible, showing a search for 'filter' and a list of operators. 'Filter Examples' is selected and highlighted with a red circle labeled '1'. On the right side, the 'Parameters' panel for 'Filter Examples' is open, showing the 'Add Filters...' button highlighted with an orange box. The 'Invert filter' checkbox is unchecked. At the bottom right, a 'Help' panel for 'Filter Examples' is visible, providing additional information about the operator.

# Workshop: Data cleaning (Cont.)

## Removing outliers

- กำหนดเงื่อนไขในการกรองข้อมูล ดังรูป แล้วกดปุ่ม **OK**

Create Filters: filters

Create Filters: filters  
Defines the list of filters to apply.

outlier equals false

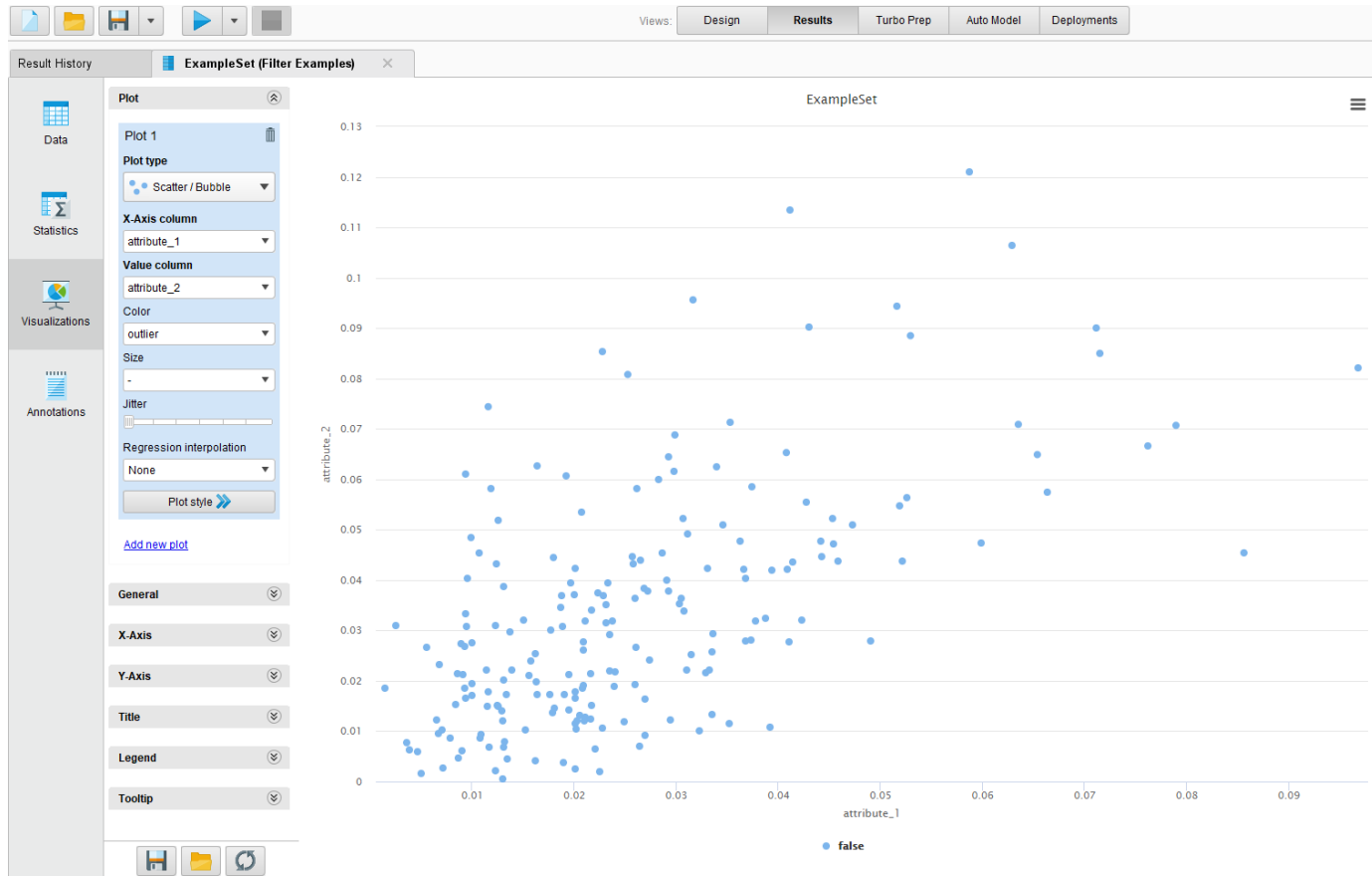
☒ Match all ☐ Match any ☒ Preselect comparators

Add Entry OK Cancel

# Workshop: Data cleaning (Cont.)

## Removing outliers

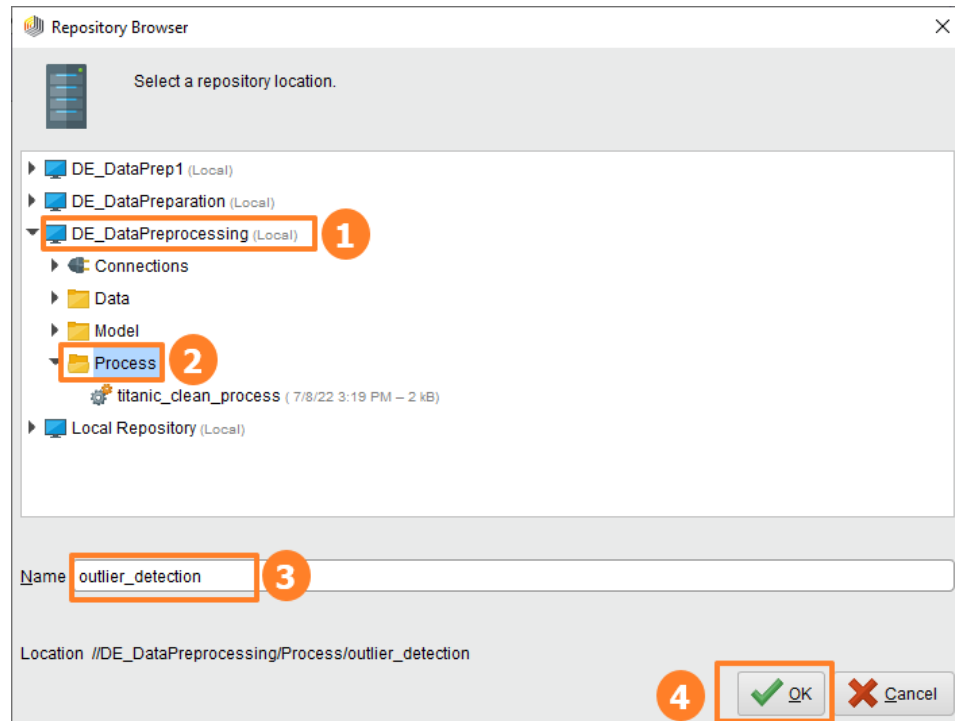
- กดปุ่ม Run แล้วดูผลลัพธ์ในหน้าต่าง Results



# Workshop: Data cleaning (Cont.)

## Removing outliers

- คลิกเลือก **Save Process as...**
- คลิกเลือก Repository: **DE\_DataPreprocessing**
- คลิกเลือกโฟลเดอร์ **Process**
- ตั้งชื่อไฟล์เป็น **outlier\_detection**



# Workshop: Data cleaning (Cont.)

## Mixed data values

- ที่แถบ **Operators** เลือก **Operator** ที่ชื่อว่า **Read CSV** และลากมาวางใน **พื้นที่ Process**
- กดปุ่มโฟลเดอร์เพื่อเลือก **CSV file** ที่ต้องการนำเข้าข้อมูล
- กำหนดค่า **Parameters** ดังรูป


The screenshot displays the Data Science Studio interface with three main panels:

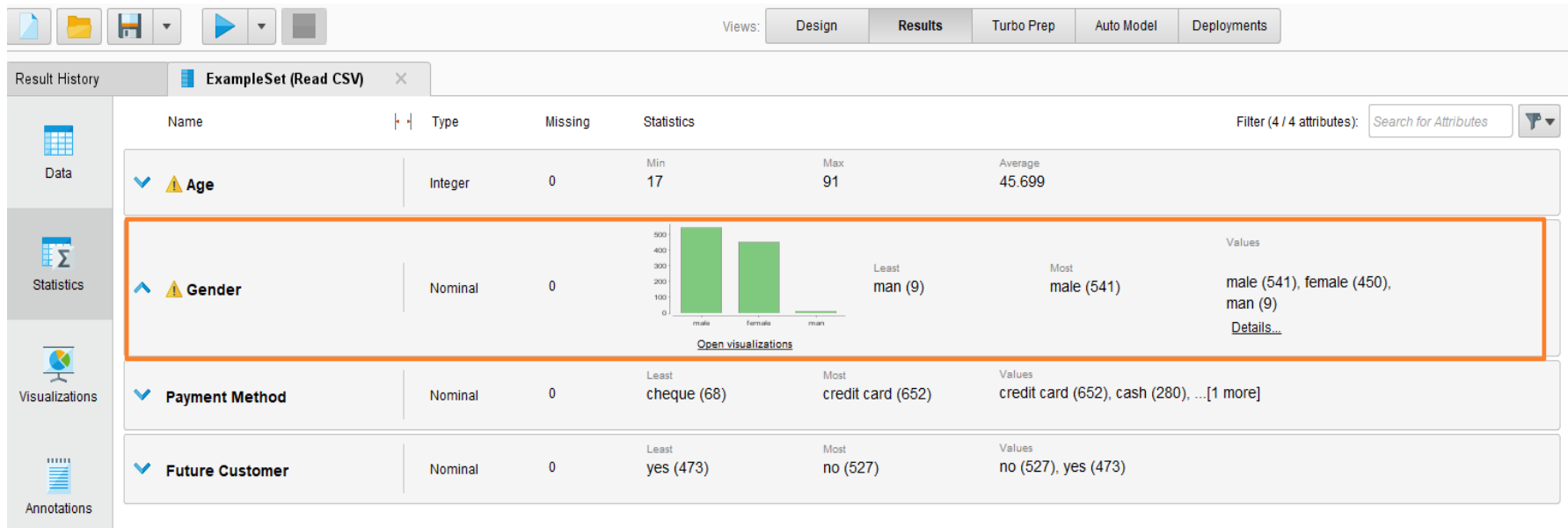
- Repository Panel (Left):** Shows a tree view of the project structure. Under the 'Process' folder, the 'Read CSV' operator is highlighted with an orange box and a red circle labeled '1'.
- Process Panel (Center):** Shows a canvas with a 'Read CSV' operator icon. A line connects it to the 'Parameters' panel on the right.
- Parameters Panel (Right):** Shows the configuration for the 'Read CSV' operator. The 'csv file' field is set to 'C:/Users/TITIRATBOONCH' and is highlighted with an orange box and a red circle labeled '2'. The 'column separators' field is set to '.' and is highlighted with an orange box and a red circle labeled '3'. Other parameters like 'use quotes', 'skip comments', 'parse numbers', 'decimal character', 'grouped digits', 'date format', and 'first row as names' are also visible.



# Workshop: Data cleaning (Cont.)

## Mixed data values

- กดปุ่ม Run  และดูค่าสถิติ (man or male)



# Workshop: Data cleaning (Cont.)

## Mixed data values

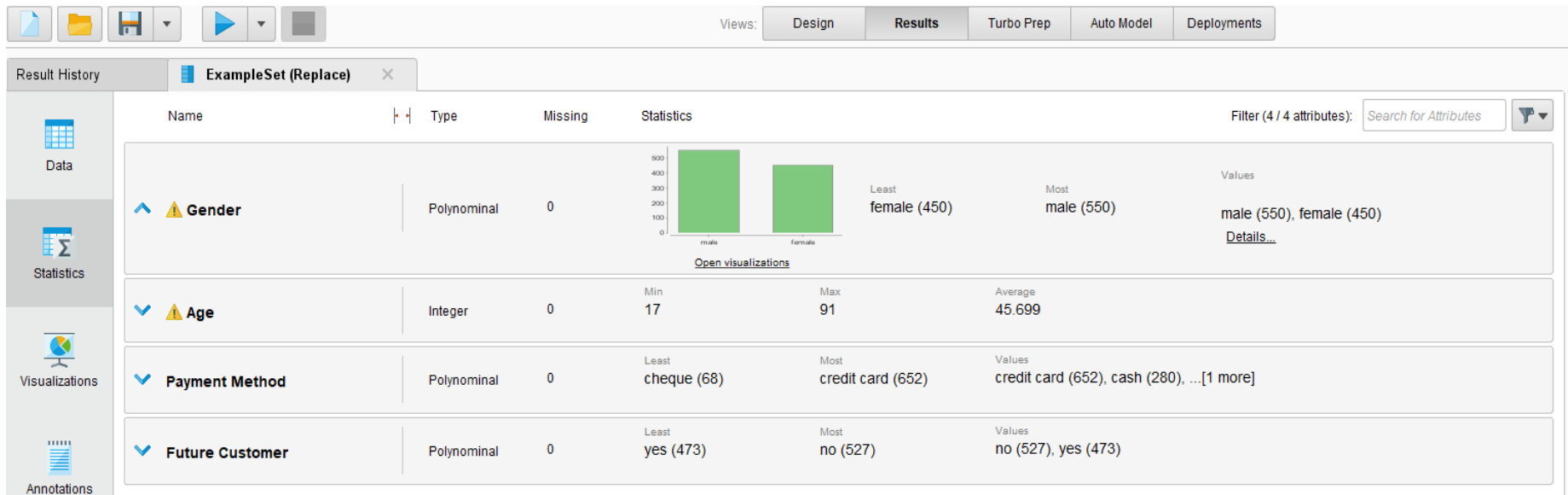
- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Replace** และลากมาวางใน พื้นที่ Process
- กำหนดค่า Parameters ดังรูป

The screenshot displays the Data Engineering Studio interface. On the left, the 'Repository' pane shows a tree view of data assets, with the 'Process' folder expanded. The 'Operators' pane at the bottom left shows a search for 'replace', with the 'Replace' operator highlighted and labeled with a red circle '1'. The central 'Process' pane shows a workflow diagram with a 'Read CSV' operator connected to a 'Replace' operator. On the right, the 'Parameters' pane for the 'Replace' operator is open, labeled with a red circle '2'. It shows the following configuration: 'attribute filter type' set to 'single', 'attribute' set to 'Gender', 'invert selection' and 'include special attributes' are unchecked, 'replace what' set to 'man', and 'replace by' set to 'male'. At the bottom of the parameters pane, there are links for 'Show advanced parameters' and 'Change compatibility (9.10.000)', and a 'Help' button at the very bottom.

# Workshop: Data cleaning (Cont.)

## Mixed data values

- กดปุ่ม Run  และดูค่าสถิติ



# Workshop: Data cleaning (Cont.)

## Mixed data values

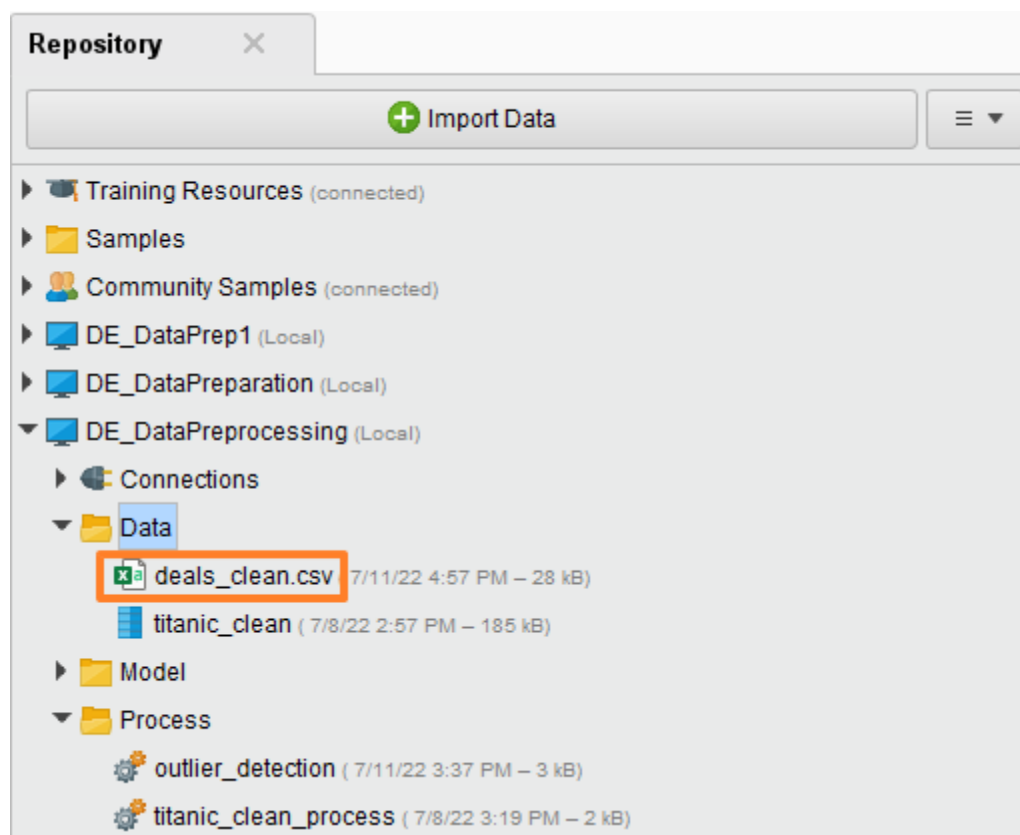
- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Write CSV** และลากมาวางในพื้นที่ Process
- กดที่ปุ่มโฟลเดอร์ เพื่อกำหนดค่าการบันทึกข้อมูล โดยให้บันทึกไฟล์ไว้ที่ไดรฟ์ D: > **DE\_DataPreprocessing > Data** และตั้งชื่อไฟล์ว่า **deals\_clean**
- เลือก column separator เป็น comma (,)

The screenshot displays the Alteryx software interface. On the left, the 'Repository' pane shows a tree structure with 'Data' > 'deals\_clean' selected. Below it, the 'Operators' pane shows the 'Write CSV' operator selected under 'Files' > 'Write'. The central 'Process' pane shows a workflow with three operators: 'Read CSV', 'Replace', and 'Write CSV'. The 'Write CSV' operator is highlighted with a red box and a red circle with the number 1. On the right, the 'Parameters' pane for the 'Write CSV' operator is shown. The 'csv file' parameter is set to 'issingData/deals\_clean.csv' (highlighted with a red box and a red circle with the number 2). The 'column separator' parameter is set to ',' (highlighted with a red box and a red circle with the number 3). Other parameters like 'write attribute names', 'quote nominal values', 'date format', and 'append to file' are also visible.

# Workshop: Data cleaning (Cont.)

## Mixed data values

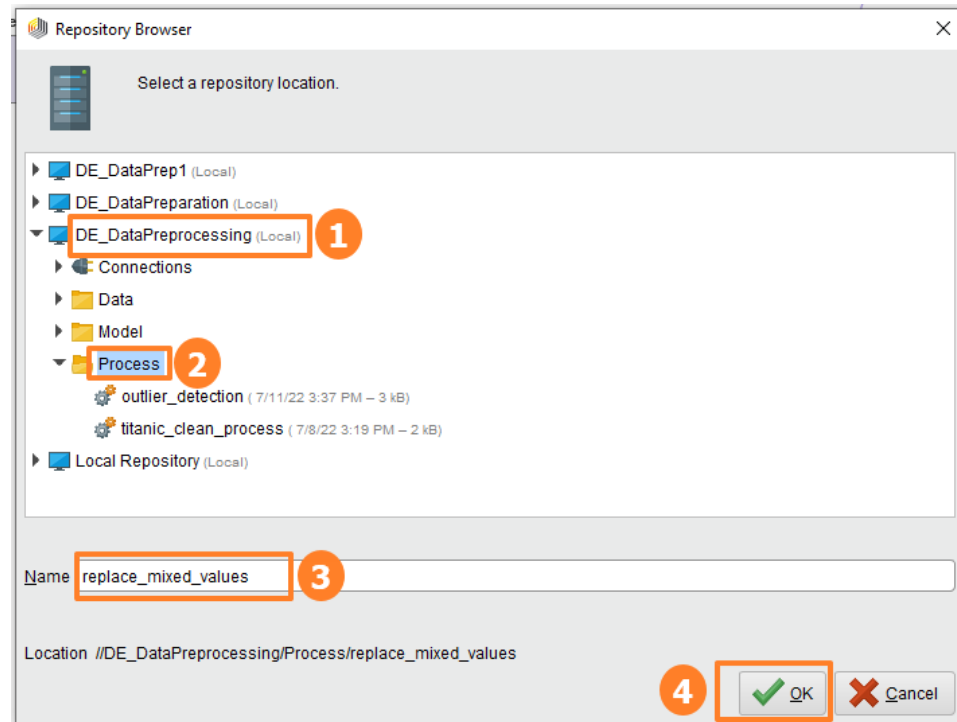
- กดปุ่ม Run  และดูที่ Repository



# Workshop: Data cleaning (Cont.)

## Mixed data values

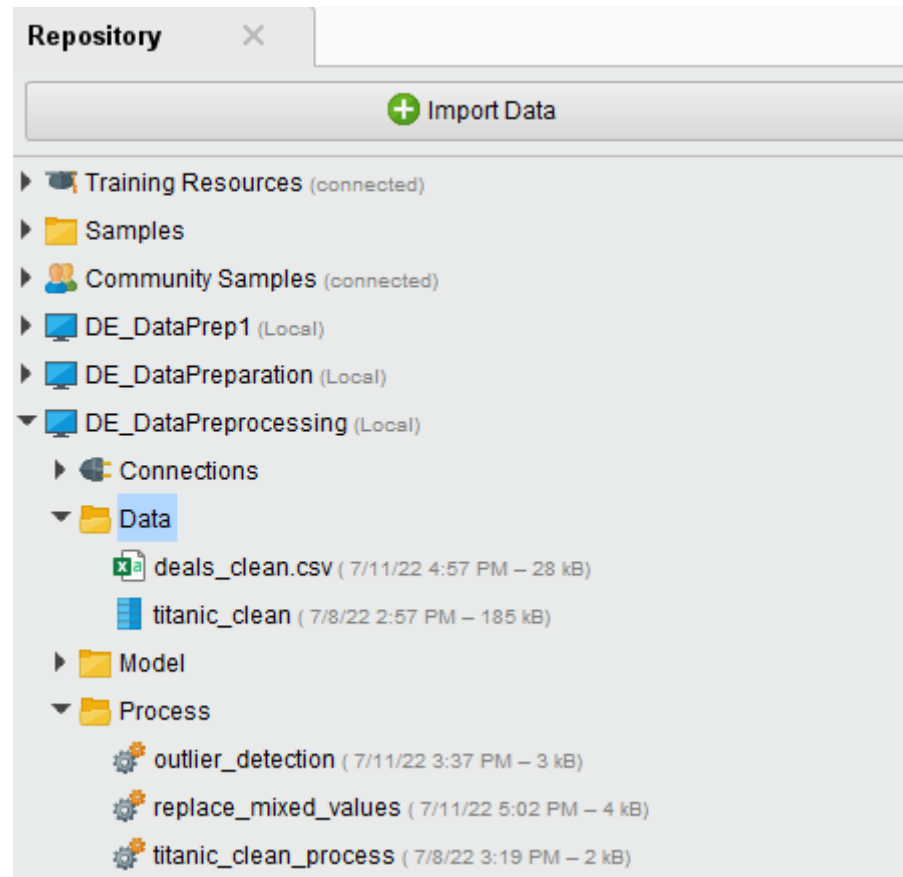
- คลิกเลือก **Save Process as...**
- คลิกเลือก Repository: **DE\_DataPreprocessing**
- คลิกเลือกโฟลเดอร์ **Process**
- ตั้งชื่อไฟล์เป็น **replace\_mixed\_values**

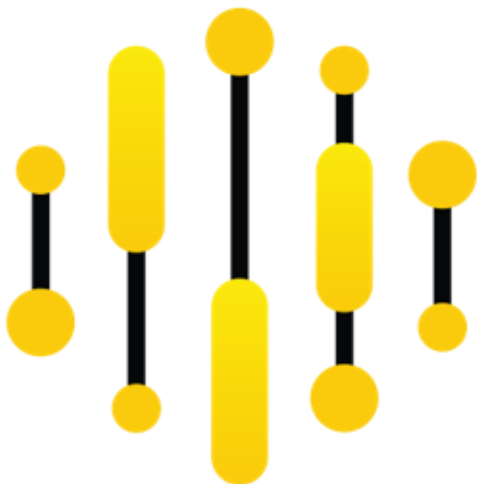


# Workshop: Data cleaning (Cont.)

## Mixed data values

- คู่มือแบบ Repository จะได้ผลลัพธ์ดังรูป





# Workshop: Data integration

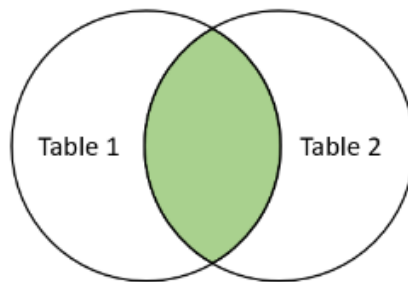
---

Training Course: Intermediate Data Engineering

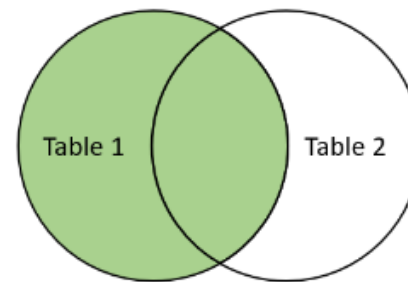


# Workshop: Data integration

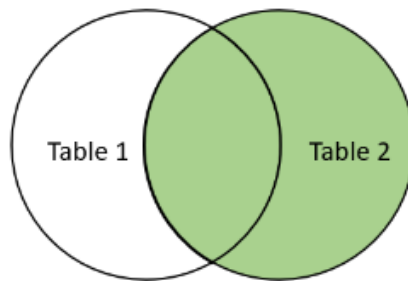
- JOIN



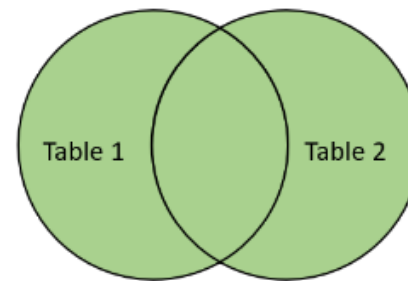
**INNER JOIN**



**LEFT JOIN**



**RIGHT JOIN**

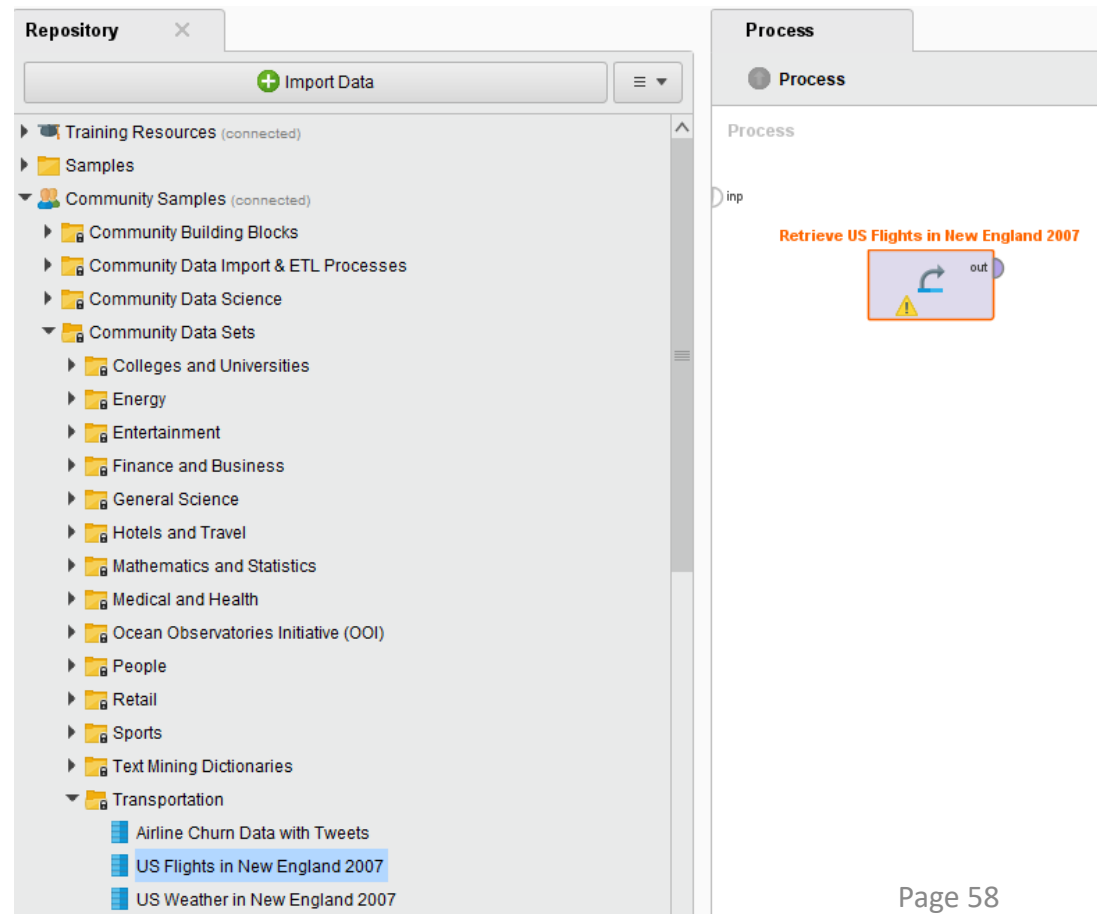


**FULL JOIN**

Source: [JOIN](#)

# Workshop: Data integration

- เลือกข้อมูลจากโฟลเดอร์ Community Samples > Community Data Sets > Transportation > **US Flights in New England 2007** แล้วลากมาวางในพื้นที่ Process



# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Filter Examples** แล้วกดปุ่ม **Add Filters...** ในแถบ Parameters เพื่อกำหนดค่าการกรองข้อมูล

The screenshot displays the RapidMiner Studio interface during a data integration workshop. The main workspace shows a process flow with two operators: 'Retrieve US Flights L...' and 'Filter Examples'. The 'Filter Examples' operator is highlighted with a red box and a red circle labeled '2'. The 'Parameters' panel on the right shows the 'Filter Examples' operator's configuration, with the 'Add Filters...' button highlighted by a red box and a red circle labeled '2'. The 'Operators' panel on the left shows the 'Filter Examples' operator selected, with a red box and a red circle labeled '1' around it. The 'Help' panel on the bottom right provides information about the 'Filter Examples' operator, including its tags and synopsis.

# Workshop: Data integration

- กำหนดเงื่อนไขการกรองข้อมูล ดังรูป

Create Filters: filters

Create Filters: filters  
Defines the list of filters to apply.

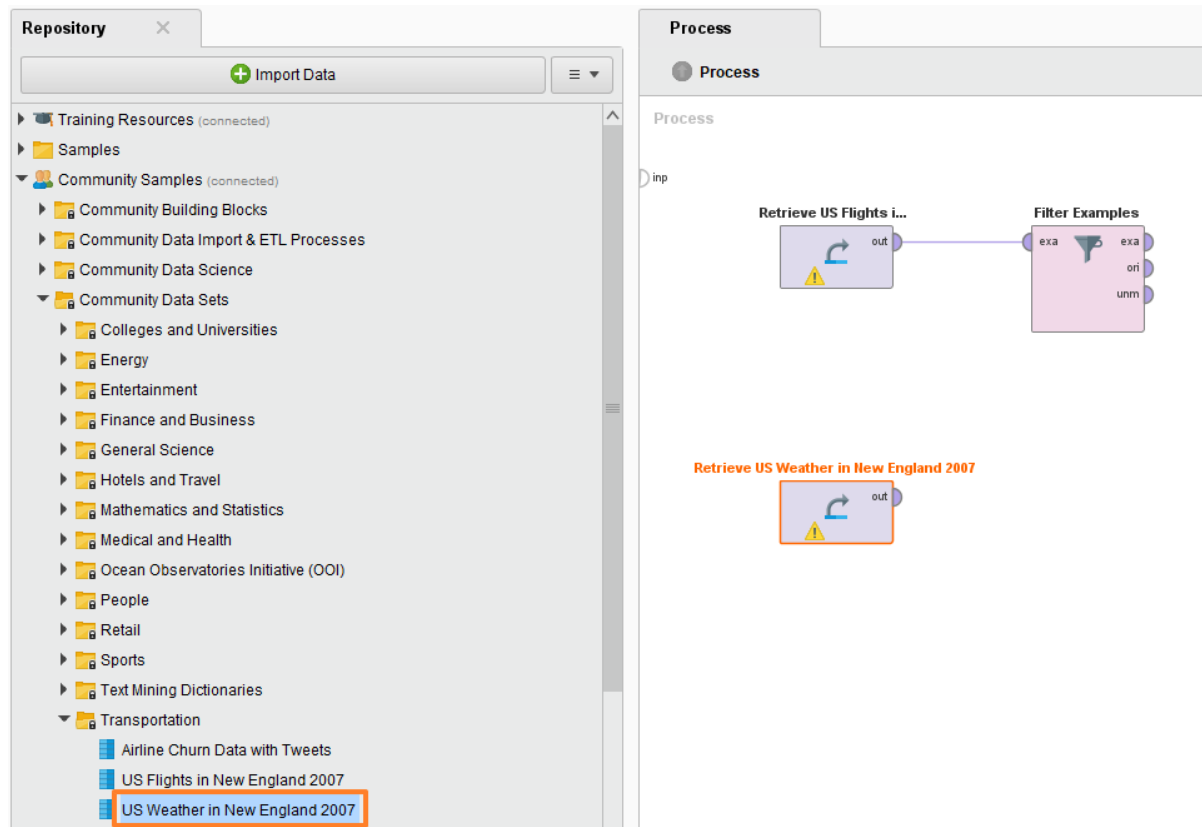
Origin equals BOS

☒ Match all ☐ Match any ☒ Preselect comparators

Add Entry OK Cancel

# Workshop: Data integration

- เลือกข้อมูลจากโฟลเดอร์ Community Samples > Community Data Sets > Transportation > **US Weather in New England 2007** แล้วลากมาวางในพื้นที่ Process



# Workshop: Data integration

- ที่แถบ **Operators** เลือก **Operator** ที่ชื่อว่า **Generate Copy** แล้ว  
กำหนดค่า **Parameters** ดังรูป เพื่อสร้างคอลัมน์ใหม่ชื่อว่า **W\_DAY** โดยนำค่า  
มาจากคอลัมน์ **PST**

The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources. The **Operators** tab is active, and the **Generate Copy** operator is highlighted with a red box and a red circle labeled '1'.
- Process:** The central workspace showing a process flow. It includes two 'Retrieve' operators (one for flights, one for weather) connected to a 'Filter Examples' operator and a 'Generate Copy' operator. The 'Generate Copy' operator is highlighted with a red box and a red circle labeled '2'.
- Parameters:** A panel on the right showing the configuration for the 'Generate Copy' operator. The 'attribute name' is set to 'PST' and the 'new name' is set to 'W\_DAY'. A red box highlights these two fields, and a red circle labeled '2' is next to the panel title.
- Help:** A panel at the bottom right showing the 'Generate Copy' operator's help text.

# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Generate Copy** แล้ว  
กำหนดค่า Parameters ดังรูป เพื่อสร้างคอลัมน์ใหม่ชื่อว่า **W\_MONTH** โดย  
นำค่ามาจากคอลัมน์ **PST**

The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources like 'Training Resources', 'Samples', and 'Community Samples'. The 'Import Data' button is visible at the top.
- Process:** The central workspace showing a workflow. It includes operators like 'Retrieve US Flights I...', 'Filter Examples', 'Retrieve US Weather...', 'Generate Copy', and 'Generate Copy (2)'. The 'Generate Copy (2)' operator is highlighted with an orange box and a red circle with the number '2'.
- Parameters:** A panel on the right showing the configuration for the 'Generate Copy (2) (Generate Copy)' operator. It has two fields: 'attribute name' set to 'PST' and 'new name' set to 'W\_MONTH'. Both fields are highlighted with an orange box and a red circle with the number '1'.
- Operators:** A panel on the bottom left showing a list of operators. The 'Generate Copy' operator is highlighted with an orange box and a red circle with the number '1'.

# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Date to Numerical** แล้ว  
กำหนดค่า Parameters ดังรูป เพื่อกำหนดค่าวันที่ (day) ให้กับคอลัมน์  
**W\_DAY**

The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources like Training Resources, Samples, and Community Samples.
- Process:** The central workspace showing a workflow. It includes operators like 'Retrieve US Flights L...', 'Filter Examples', 'Retrieve US Weather...', 'Generate Copy', 'Generate Copy (2)', and 'Date to Numerical'.
- Operators:** A panel on the bottom left with a search bar containing 'date to'. It lists several operators, with 'Date to Numerical' highlighted and marked with a red circle and the number 1.
- Parameters:** A panel on the right titled 'Date to Numerical' (marked with a red circle and the number 2). It contains the following settings:
  - attribute name:** W\_DAY
  - time unit:** day
  - day relative to:** month
- Help:** A panel at the bottom right showing the 'Date to Numerical' operator's description and version information.



# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Date to Numerical** แล้ว  
กำหนดค่า Parameters ดังรูป เพื่อกำหนดค่าเดือน (month) ให้กับคอลัมน์  
**W\_MONTH**

The screenshot displays the RapidMiner Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources like Training Resources, Samples, and Community Samples.
- Process:** The central workspace showing a workflow. It includes operators like 'Retrieve US Flights i...', 'Filter Examples', 'Retrieve US Weather...', 'Generate Copy', 'Generate Copy (2)', 'Date to Numerical', and 'Date to Numerical (2)'. The 'Date to Numerical (2)' operator is highlighted with an orange box and a red circle with the number 2.
- Operators:** A panel on the bottom left showing a search for 'date to'. Under the 'Types (4)' section, the 'Date to Numerical' operator is selected and highlighted with an orange box and a red circle with the number 1.
- Parameters:** A panel on the right showing the configuration for the 'Date to Numerical (2) (Date to Numerical)' operator. The parameters are:
  - attribute name: W\_MONTH
  - time unit: month
  - month relative to: yearThe parameter table is highlighted with an orange box.
- Help:** A panel at the bottom right showing the 'Date to Numerical' operator's help text, including 'RapidMiner Studio Core' and 'Type: Date, Time, DateTime, Continuous, Numerical'.

# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Join** ลากเส้นเชื่อมต่อให้สมบูรณ์ แล้วไปที่แถบ Parameters เลือก join type เป็น inner แล้วคลิก Edit List (0)... ดังรูป

The screenshot displays the Data Engineering Studio interface with the following components:

- Repository:** A sidebar on the left showing a tree structure of data sources under 'Community Data Sets'.
- Process:** The central workspace showing a workflow. It includes a 'Retrieve US Flights L...' operator, a 'Filter Examples' operator, and a 'Join' operator. Below these, there is a sequence of operators: 'Retrieve US Weather...', 'Generate Copy', 'Generate Copy (2)', 'Date to Numerical', and 'Date to Numerical (2)'. The 'Join' operator is highlighted with a red box and a red circle with the number '2'.
- Parameters:** A panel on the right showing the configuration for the 'Join' operator. The 'join type' is set to 'inner'. The 'key attributes' section has an 'Edit List (0)...' button, which is highlighted with a red box and a red circle with the number '1'.
- Operators:** A panel at the bottom left showing a list of operators. The 'Join' operator is selected and highlighted with a red box and a red circle with the number '1'.

# Workshop: Data integration

- กำหนดค่า **key attributes** ที่ใช้เชื่อมโยงข้อมูล ดังรูป

Edit Parameter List: key attributes

Edit Parameter List: **key attributes**  
The attributes which shall be used for join. Attributes which shall be matched must be of the same type.

left key attributes	right key attributes
Origin	AirportCode
Month	W_MONTH
DayofMonth	W_DAY

Add Entry Remove Entry Apply Cancel

# Workshop: Data integration

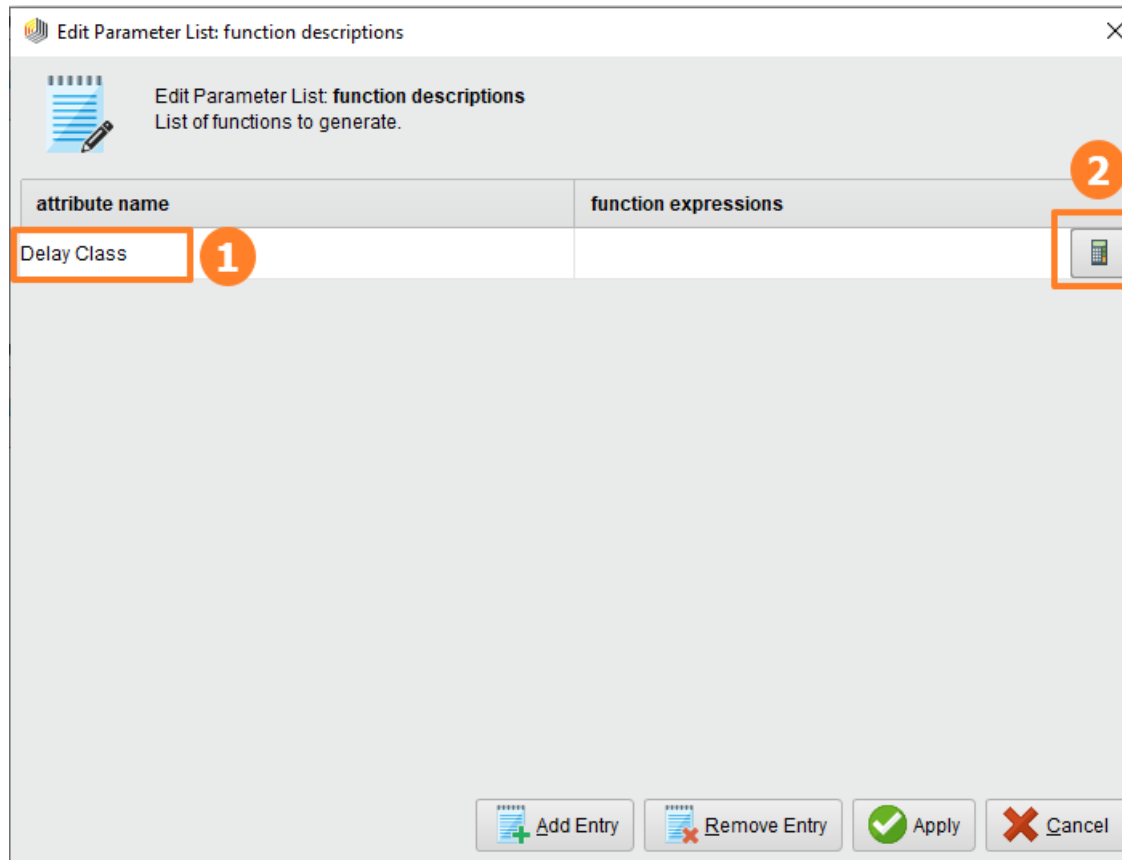
- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Generate Attributes** แล้ว  
ไปที่แถบ Parameters คลิก Edit List (0)... ดังรูป

The screenshot displays the Data Engineering Studio interface with the following components:

- Repository:** A tree view on the left showing various data sources like Training Resources, Samples, and Community Samples. The 'Community Data Sets' folder is expanded.
- Operators:** A panel at the bottom left showing a search for 'generate'. Under the 'Generation' category, the 'Generate Attributes' operator is highlighted with a red box and a red circle labeled '1'.
- Process:** The central workspace showing a workflow. It includes operators like 'Retrieve US Flights I...', 'Filter Examples', 'Join', 'Retrieve US Weather...', 'Generate Copy', 'Generate Copy (2)', 'Date to Numerical', and 'Date to Numerical (2)'. The 'Generate Attributes' operator is highlighted with a red box and a red circle labeled '2'.
- Parameters:** A panel on the right for the 'Generate Attributes' operator. It shows 'function descriptions' and a button 'Edit List (0)...' which is highlighted with a red box and a red circle labeled '2'. Below this are links for 'Show advanced parameters' and 'Change compatibility (9.10.000)'.

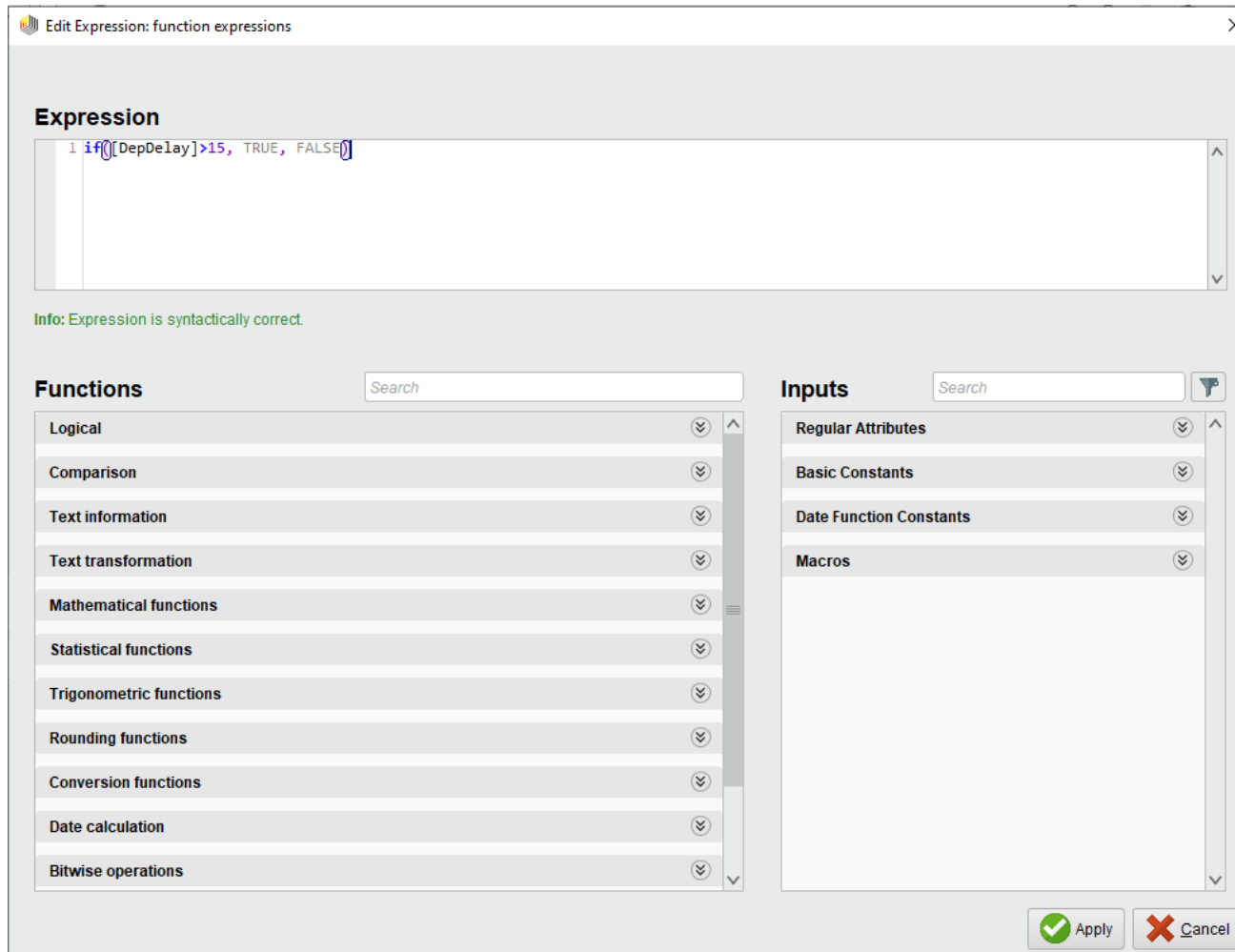
# Workshop: Data integration

- กำหนดค่า **attribute name** เป็น **Delay Class** แล้วกดปุ่มรูปเครื่องคิดเลข เพื่อใส่เงื่อนไข



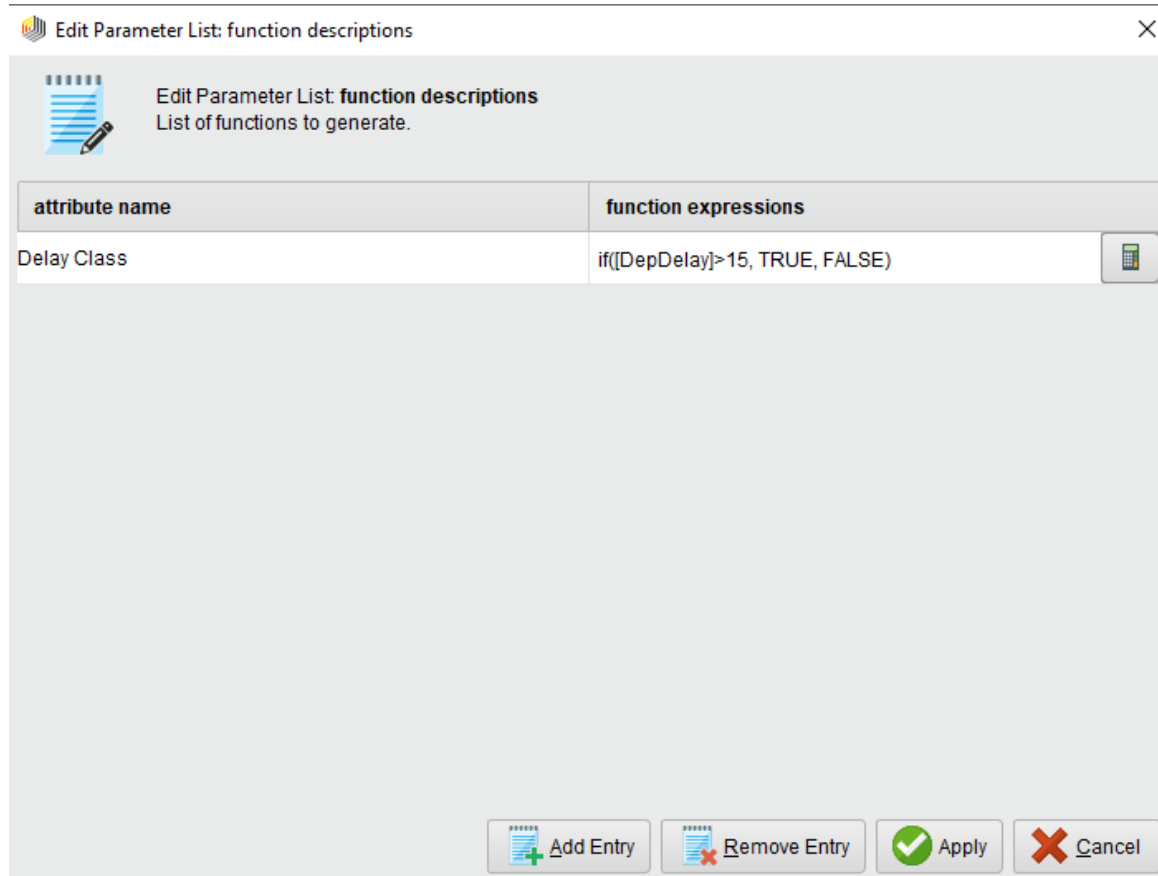
# Workshop: Data integration

- ใส่สูตร `if([DepDelay]>15, TRUE, FALSE)`



# Workshop: Data integration

- เมื่อกลับมายังหน้าต่าง **function descriptions** จะได้เงื่อนไขดังรูป



Edit Parameter List: function descriptions

Edit Parameter List: **function descriptions**  
List of functions to generate.

attribute name	function expressions
Delay Class	if([DepDelay]>15, TRUE, FALSE)

Buttons: Add Entry, Remove Entry, Apply, Cancel

# Workshop: Data integration

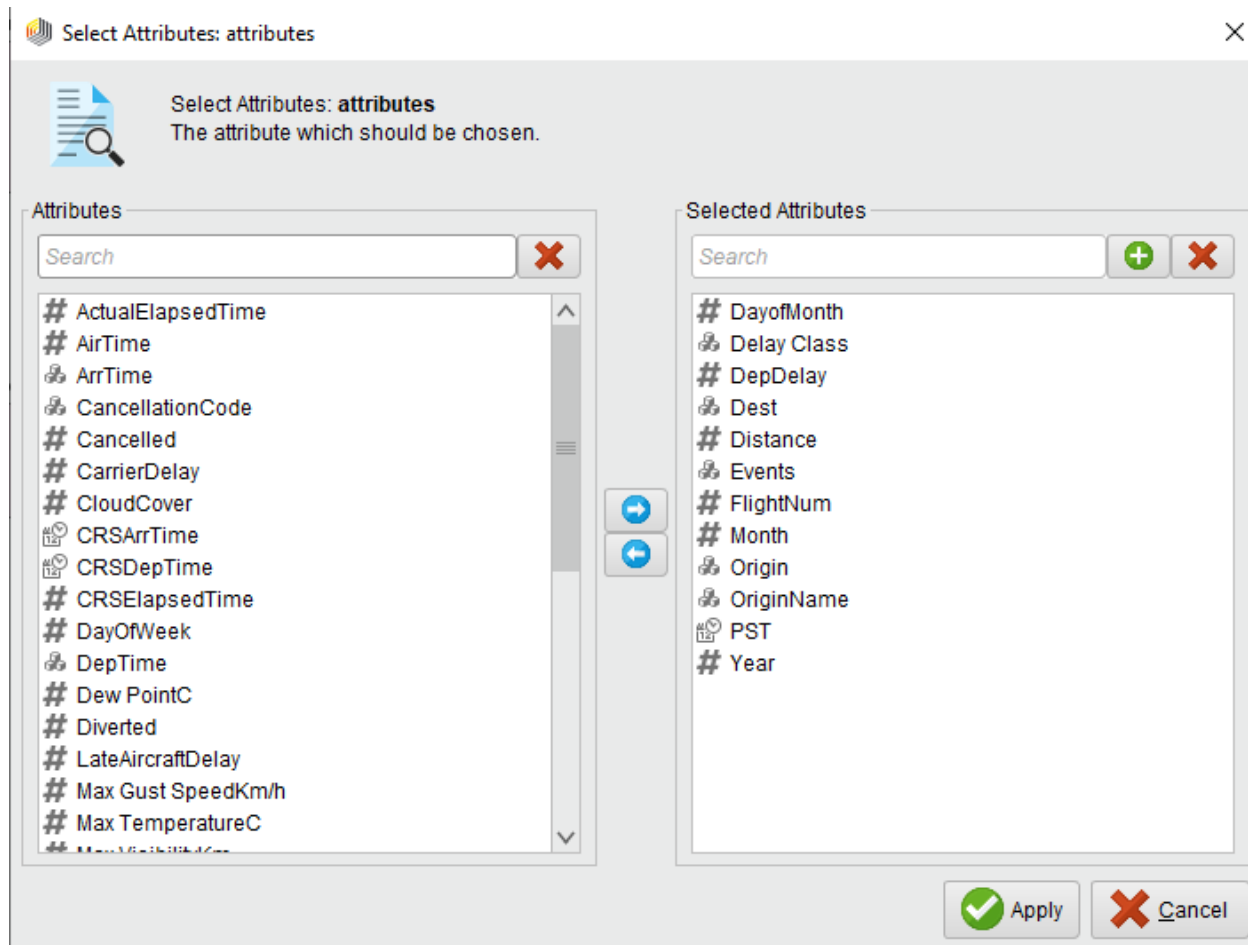
- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Select Attributes** แล้วไปที่แถบ Parameters กำหนด attributes filter type เป็น subset แล้วคลิก Select Attribute... ดังรูป

The screenshot displays the Orange3 data mining software interface. The main workspace shows a workflow with several operators: 'Filter Examples', 'Join', 'Generate Attributes', and 'Select Attributes'. The 'Select Attributes' operator is highlighted with a red box and a red circle with the number 2. The 'Parameters' panel on the right shows the configuration for the 'Select Attributes' operator, with 'attribute filter type' set to 'subset' and a red box around the 'Select Attributes...' button. The 'Repository' panel on the left shows a list of data sources, with 'Community Data Science' selected. The 'Operators' panel at the bottom left shows a list of operators, with 'Select Attributes' selected and highlighted with a red box and a red circle with the number 1.



# Workshop: Data integration

- เลือกคอลัมน์ที่ต้องการ ดังรูป



# Workshop: Data integration

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Filter Examples** แล้วกดปุ่ม **Add Filters...** ในแถบ Parameters เพื่อกำหนดค่าการกรองข้อมูล

The screenshot displays the RapidMiner Studio interface with the following components:


- Repository:** A tree view on the left showing various data sources like 'Training Resources', 'Samples', and 'Community Data Sets'.
- Operators:** A panel on the bottom left with a search bar containing 'filter'. Under the 'Filter (2)' category, the 'Filter Examples' operator is highlighted with a red box and a red circle labeled '1'.
- Process:** The central workspace showing a workflow diagram. The workflow includes:
  - 'Generate Copy (2)'
  - 'Date to Numerical' (twice)
  - 'Join'
  - 'Generate Attributes'
  - 'Select Attributes'
  - 'Filter Examples (2)' (highlighted with a red box)
- Parameters:** A panel on the right for the 'Filter Examples (2) (Filter Examples)' operator. It features a red box around the 'Add Filters...' button, which is also labeled with a red circle '2'. Below this button is an 'invert filter' checkbox.
- Help:** A panel at the bottom right showing the 'Filter Examples' operator's documentation, including tags and a synopsis.

# Workshop: Data integration

- กำหนดเงื่อนไขการกรองข้อมูล ดังรูป

The screenshot shows a dialog box titled "Create Filters: filters" with a close button (X) in the top right corner. Inside the dialog, 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 rule entry: "Delay Class" (selected from a dropdown), "equals" (selected from a dropdown), and "true" (entered in a text field). To the right of the text field are two icons: a star and a red X. At the bottom of the dialog, there are three radio buttons: "Match all" (selected), "Match any", and "Preselect comparators" (checked). To the right of these are three buttons: "Add Entry" (with a plus icon), "OK" (with a green checkmark), and "Cancel" (with a red X).

# Workshop: Data integration

- กดปุ่ม Run  และดูผลลัพธ์ในแถบ Results

Views: Design Results Turbo Prep Auto Model Deployments

Result History ExampleSet (Filter Examples (2))

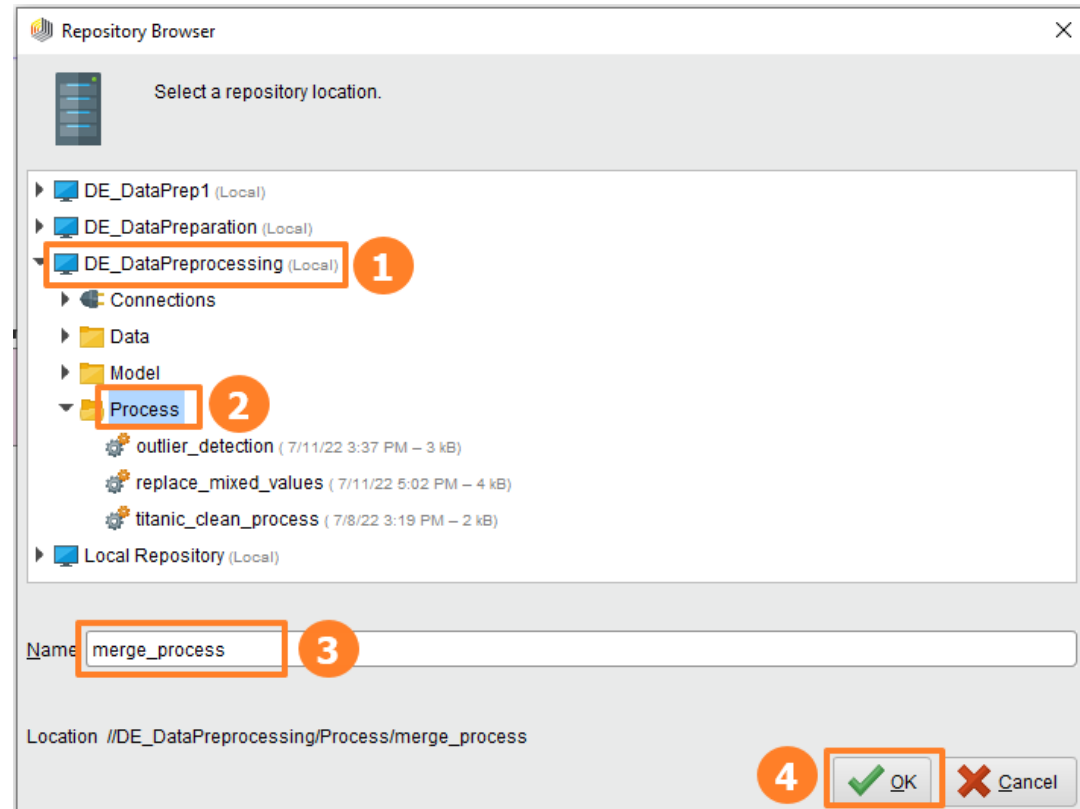
Open in Turbo Prep Auto Model Filter (713 / 713 examples): all

Row No.	Year	Month	DayofMonth	FlightNum	DepDelay	Origin	OriginName	Dest	Distance	Events	PST	Delay Class
1	2007	1	26	2781	16	BOS	Boston	CLE	563	Normal	Jan 26, 2007	true
2	2007	1	22	3148	18	BOS	Boston	CLE	563	Snow	Jan 22, 2007	true
3	2007	1	26	2128	23	BOS	Boston	CLE	563	Normal	Jan 26, 2007	true
4	2007	1	6	1299	60	BOS	Boston	EWB	200	Rain	Jan 6, 2007	true
5	2007	1	19	3148	34	BOS	Boston	CLE	563	Rain-Snow	Jan 19, 2007	true
6	2007	1	16	2128	41	BOS	Boston	CLE	563	Rain	Jan 16, 2007	true
7	2007	1	22	2781	17	BOS	Boston	CLE	563	Snow	Jan 22, 2007	true
8	2007	1	31	2128	33	BOS	Boston	CLE	563	Normal	Jan 31, 2007	true
9	2007	1	20	1213	42	BOS	Boston	EWB	200	Normal	Jan 20, 2007	true
10	2007	1	1	7394	38	BOS	Boston	IAD	413	Rain	Jan 1, 2007	true
11	2007	1	3	7271	20	BOS	Boston	IAD	413	Normal	Jan 3, 2007	true
12	2007	1	4	7340	26	BOS	Boston	IAD	413	Normal	Jan 4, 2007	true
13	2007	1	5	7340	275	BOS	Boston	IAD	413	Rain	Jan 5, 2007	true
14	2007	1	5	7394	77	BOS	Boston	IAD	413	Rain	Jan 5, 2007	true
15	2007	1	6	7160	134	BOS	Boston	IAD	413	Rain	Jan 6, 2007	true
16	2007	1	7	7160	46	BOS	Boston	IAD	413	Normal	Jan 7, 2007	true
17	2007	1	8	7340	35	BOS	Boston	IAD	413	Rain	Jan 8, 2007	true
18	2007	1	8	7394	56	BOS	Boston	IAD	413	Rain	Jan 8, 2007	true
19	2007	1	10	7271	28	BOS	Boston	IAD	413	Normal	Jan 10, 2007	true
20	2007	1	11	7340	18	BOS	Boston	IAD	413	Normal	Jan 11, 2007	true
21	2007	1	11	7394	27	BOS	Boston	IAD	413	Normal	Jan 11, 2007	true
22	2007	1	13	7258	145	BOS	Boston	IAD	413	Rain	Jan 13, 2007	true
23	2007	1	14	7208	26	BOS	Boston	IAD	413	Rain	Jan 14, 2007	true
24	2007	1	14	7394	38	BOS	Boston	IAD	413	Rain	Jan 14, 2007	true
25	2007	1	15	7208	22	BOS	Boston	IAD	413	Rain	Jan 15, 2007	true

ExampleSet (713 examples, 0 special attributes, 12 regular attributes)

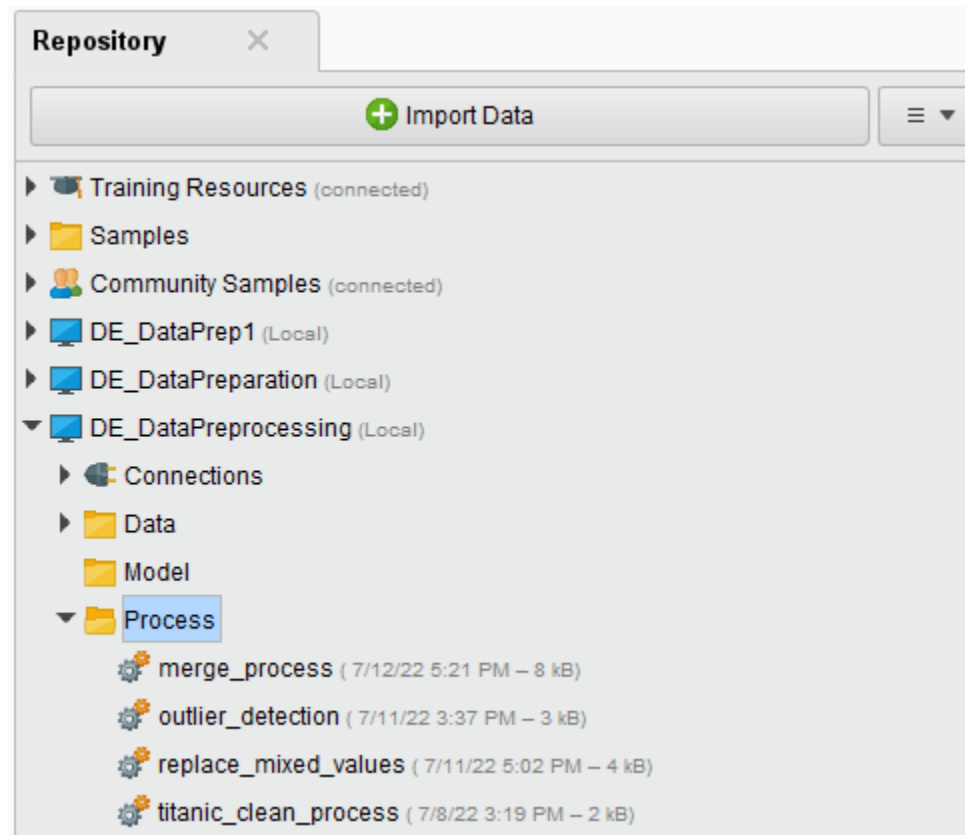
# Workshop: Data integration

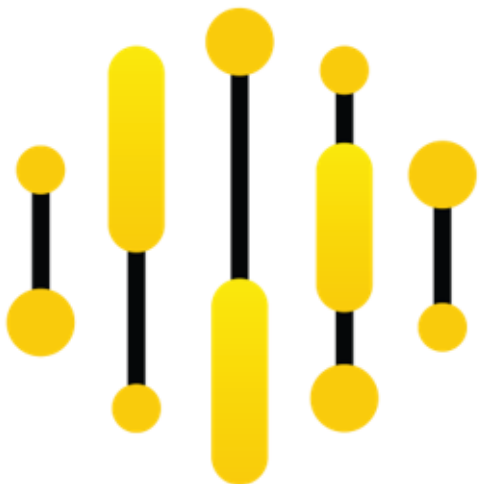
- คลิกเลือก **Save Process as...**
- คลิกเลือก Repository: **DE\_DataPreprocessing**
- คลิกเลือกโฟลเดอร์ **Process**
- ตั้งชื่อไฟล์เป็น **merge\_process**



# Workshop: Data integration

- กลับมาดูที่ Repository





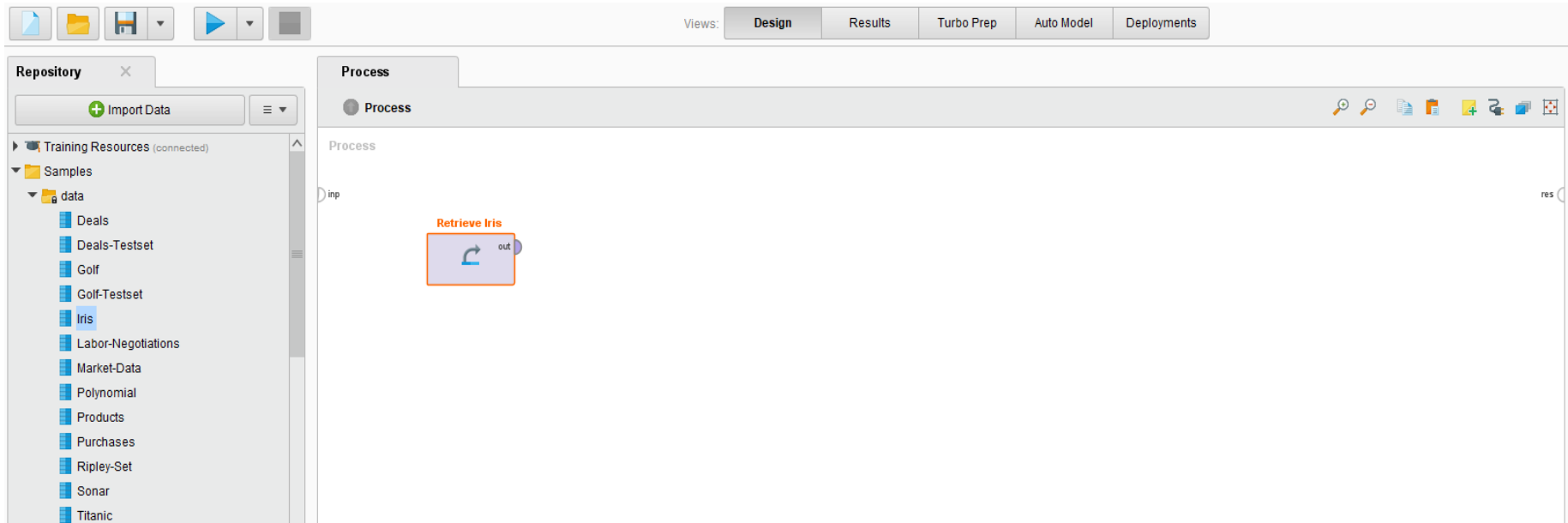
# Workshop: Data transformation

---

Training Course: Intermediate Data Engineering

# Workshop: Data transformation

- เลือกข้อมูลจากไฟล์เดอร์ Samples > data > Iris แล้วลากมาวางในพื้นที่ Process





# Workshop: Data transformation

- ที่แถบ Operators เลือก Operator ที่ชื่อว่า **Normalize** และลากมาวางใน พื้นที่ Process และกำหนดค่า Parameters ดังรูป

The screenshot displays the Data Engineering Studio interface with the following components:

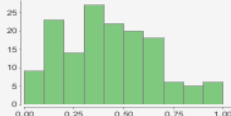
- Repository:** A tree view on the left showing 'Training Resources (connected)' with a 'Samples' folder containing a 'data' folder. The 'data' folder lists various datasets, including 'Iris'.
- Process:** The central workspace showing a workflow. A 'Retrieve Iris' operator is connected to a 'Normalize' operator. The 'Normalize' operator is highlighted with a red box and a red circle with the number '1'.
- Operators:** A panel on the bottom left showing a search for 'normalize'. Under the 'Cleansing (2)' category, the 'Normalize' operator is highlighted with a red box and a red circle with the number '1'.
- Parameters:** A panel on the right showing the configuration for the 'Normalize' operator. The 'attribute filter type' is set to 'all'. The 'method' is set to 'range transform...'. The 'min' value is '0.0' and the 'max' value is '1.0'. The 'invert selection' and 'include special attributes' checkboxes are unchecked. The panel is highlighted with a red box and a red circle with the number '2'.

# Workshop: Data transformation

- กดปุ่ม **Run**  แล้วดูผลลัพธ์ในแถบ **Results** เลือกเมนู **Statistics** เพื่อดูสถิติของข้อมูล

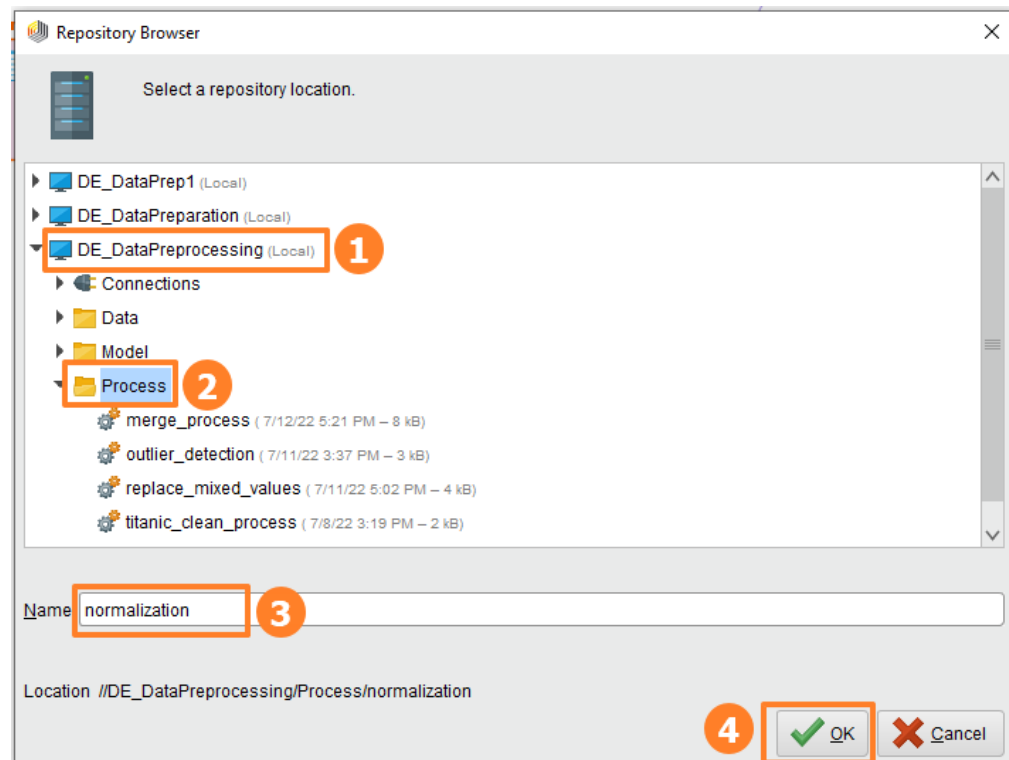
Views: Design Results Turbo Prep Auto Model Deployments

Result History: ExampleSet (Normalize)

Name	Type	Missing	Statistics		Filter (6 / 6 attributes): <input type="text" value="Search for Attributes"/>
id	Nominal	0	Least id_99 (1)	Most id_1 (1)	Values id_1 (1), id_10 (1), ...[148 more]
label	Nominal	0	Least Iris-virginica (50)	Most Iris-setosa (50)	Values Iris-setosa (50), Iris-versicolor (50), ...[1 more]
a1	Real	0	 Min 0 Max 1 Average 0.429 Deviation 0.230		<a href="#">Open visualizations</a>
a2	Real	0	Min 0	Max 1	Average 0.439
a3	Real	0	Min 0	Max 1	Average 0.468
a4	Real	0	Min 0	Max 1	Average 0.458

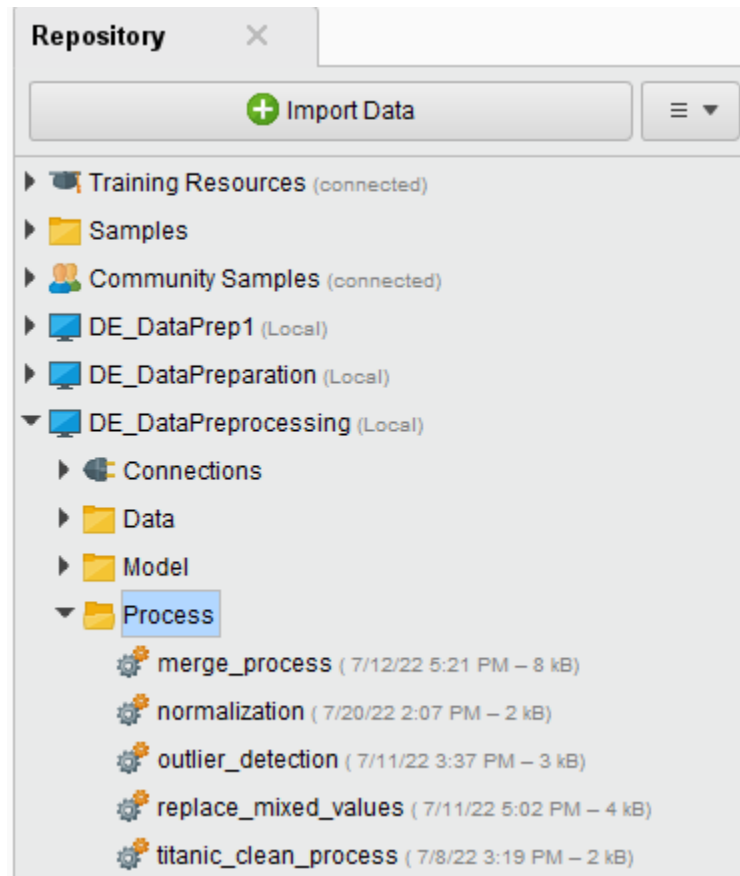
# Workshop: Data transformation

- คลิกเลือก **Save Process as...**
- คลิกเลือก Repository: **DE\_DataPreprocessing**
- คลิกเลือกโฟลเดอร์ **Process**
- ตั้งชื่อไฟล์เป็น **normalization** และกดปุ่ม **OK**



# Workshop: Data transformation

- คู่มือแบบ Repository จะได้ผลลัพธ์ดังรูป



# End Workshop

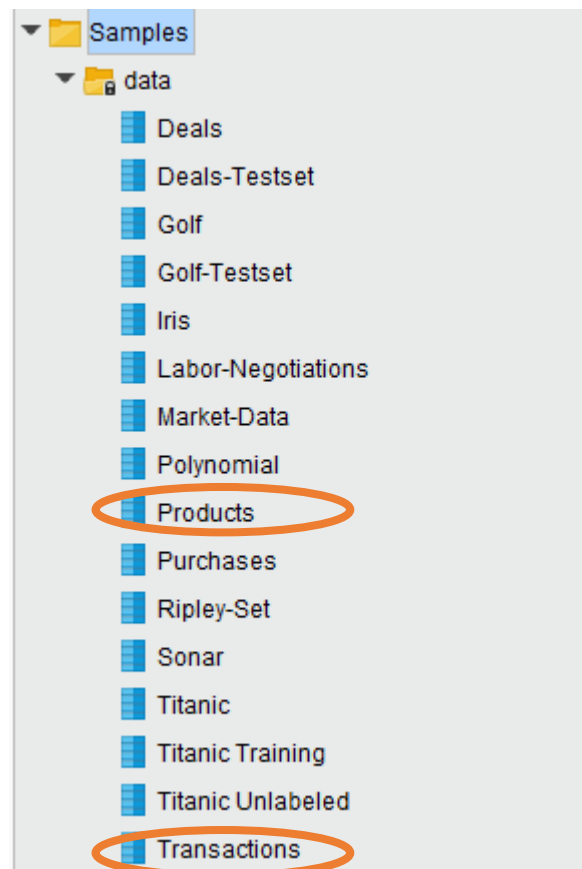


# Workshop: เพิ่มเติม

- ข้อมูล Deals\_sample\_2.csv

Customer	Customer Type	Payment Type	Purchases	Sales	Refunds	Country	Continent
10000	Person	Cash	120000	150000	240	Canada	America
10001	Company	Cash		651750	1043	Japan	Asia
10002	Company	Credit Card	451000	563750	902	Mexico	America
10003	Company	Transfer	565000	706250	1130	Spain	Europe
10004	Personal	Transfers	512300		1024	Argentina	America
10005	Person	Transfer	415500	519375	0	Canada	America
10006	Company	Credit Card	696300	870375	1392	EEUU	America
10007	Person	Cash	741000	926250	1482	Chile	America
10008	Company	Cash	541000	676250	1082	EEUU	America
10009	Company	Cashs	83000	103750	166	EEUU	America
10010	Company	Cash	454100	567925	910	EEUU	America
10011	Person	Transfers	520033	650041	1041	EEUU	America
10012	Person	Credit Card	452000	565000	904	Canada	America
10013	Person	Transfer	352000	440000	704	Germany	Europe
10014	Company	Transfer	241010	301262	480	EEUU	America
10015	Company	Credit Card	560122	700152	1120	Mexico	America
10016	Person	Credit Card	362200	452750	0	Canada	America
10017	Person	Cash	452230	565287	903	Japan	Asia
10018	Company	Cash	521000	651250	1042	Spain	Europe

- ต้องการทราบชื่อสินค้าที่มียอดขายรวมตั้งแต่ 60 ขึ้นขึ้นไป จากข้อมูล Transactions และ Products



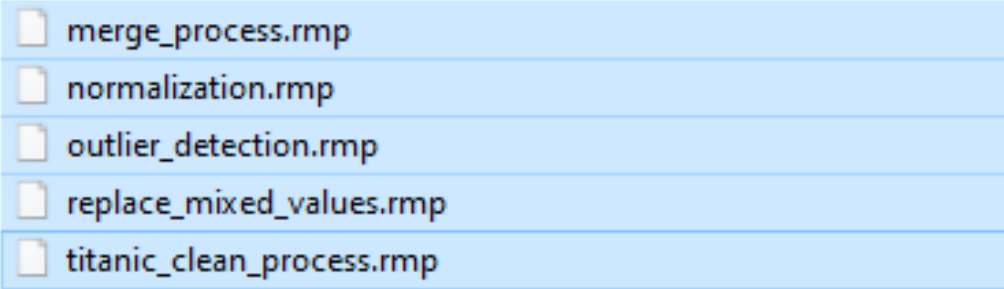
# Assignment






ส่งงานไฟล์ (ที่ไดรฟ์ D: > โฟลเดอร์ DE\_DataPreprocessing > Process)

- merge\_process.rmp
- normalization.rmp
- outlier\_detection.rmp
- replace\_mixed\_values.rmp
- titanic\_clean\_process.rmp

New Volume (D:) > DE\_DataPreprocessing > Process

Name



 merge_process.rmp
 normalization.rmp
 outlier_detection.rmp
 replace_mixed_values.rmp
 titanic_clean_process.rmp