Lecture 5
How to Conduct Data Preparation Phase?

Kim, Yang Sok

Dept. of MIS, Keimyung University

Table of Content

Introduction

Introduction

- **Data miners spend most of their time on Data Preparation. Most data used for data mining was originally collected and preserved for other purposes and needs some refinement before it is ready to use for modeling.**

- **The data preparation phase includes five tasks. These are**
  - Selecting data
  - Cleaning data
  - Constructing data
  - Integrating data
  - Formatting data

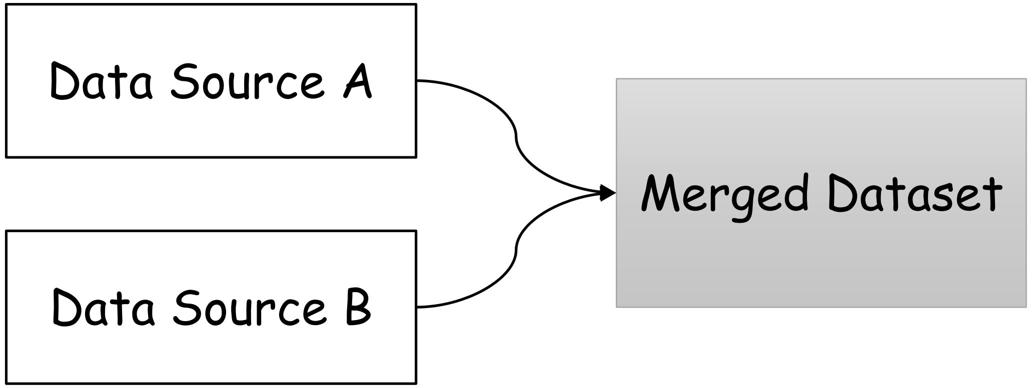
- **Rapidminer supports data preparation through Data Blending and Data Cleansing packages.**

Task & Deliverables

Tasks & Deliverables

Integrating Data

- **Your data may now be in several disparate datasets.**
- **You'll need to merge some or all of those disparate datasets together to get ready for the modeling phase.**
- **The deliverable for this task is the merged dataset.**



- Merging dataset refers to joining together two or more datasets that have different information about the same objects.
- At this stage, it may also be advisable to generate new records.
- It may also be recommended to generate aggregate values. Aggregation refers to operations where new values are computed by summarizing information from multiple records and/or tables.

Task & Deliverables

Selecting Data

- **Now you will decide which portion of the data that you have is actually going to be used for data mining.**

- **The criteria you might use to make this decision include the relevance of the data to your data mining goals, the quality of the data, and also technical constraints such as limits on data volume or data types.**

- **Note that data selection covers selection of attributes (columns) as well as selection of records (rows) in a table.**

- **The deliverable for this task is the rationale for inclusion and exclusion. In it, you'll explain what data will, and will not, be used for further data-mining work.**

Task & Deliverables

Cleaning Data

- **The data that you've chosen to use is unlikely to be perfectly clean (error-free). You need to raise the data quality to the level required by the selected analysis techniques.**

- **This may involve the selection of clean subsets of the data, the insertion of suitable defaults, or more ambitious techniques such as the estimation of missing data by modeling.**

- **The deliverable for this task is the data-cleaning report, which documents, in excruciating detail, every decision and action used to clean your data.**
  - This report should cover and refer to each data quality problem that was identified in the verify data quality task in the data-understanding phase of the process.
  - You report should also address the potential impact on results of the choices you have made during data cleaning.

Task & Deliverables

Constructing Data

- **This task includes constructive data preparation operations such as the production of derived attributes, complete new records, or transformed values for existing attributes.**

- **Deliverables for this task include two reports:**
  - Derived attributes: A report that describes what new fields (columns) you have constructed, how you did it, and why.
  - Generated records: A report that describes what new cases (rows) you have constructed, how you did it, and why.

Tasks & Deliverables

Formatting Data

- Data often comes to you in formats other than the ones that are most convenient for modeling. (Format changes are usually driven by the design of your tools.) So convert those formats now.

- Formatting transformations refers primarily to syntactic modifications made to the data that do not change its meaning, but might be required by the modeling tool.

- Some tools have requirements on the order of the attributes, such as the first field being a unique identifier for each record or the last field being the outcome field the model is to predict.

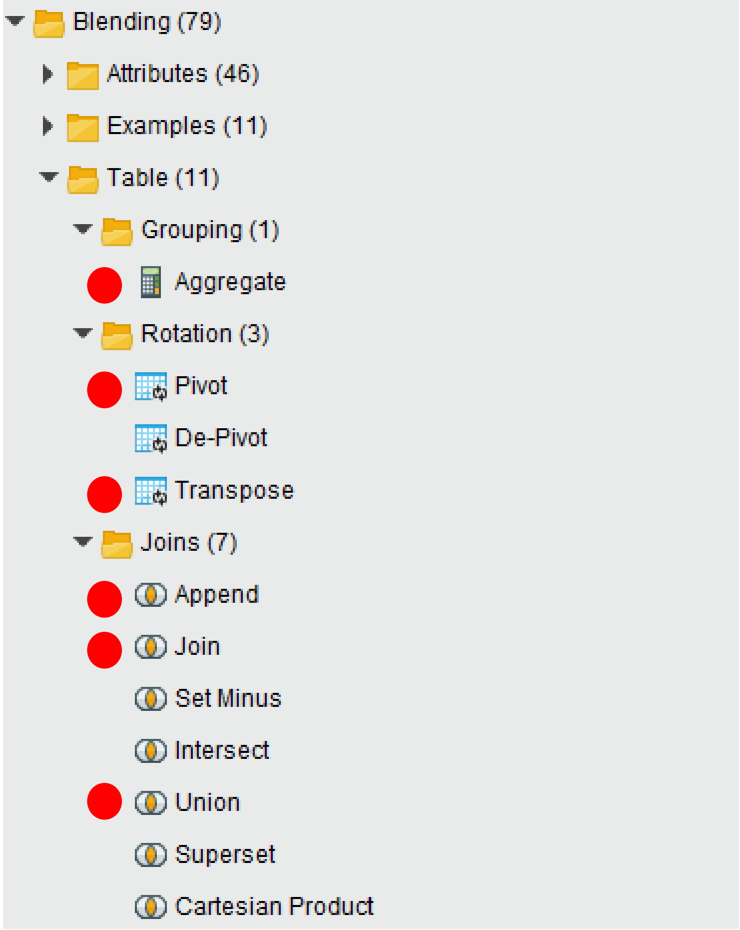
- The deliverable for this task is your reformatted data. (And a little report describing the changes you have made would be a smart thing to include.)

Exercise 1: Integrating Data with Rapidminer

Exercise 1: Integration Data with Rapidminer

Data Blending

- Data can be exists multiple sources, so it is necessary to integrate them to conduct data analysis.
- Rapidminer provides operators related to data integration in Blending > Table package.
- The package has sub-package that support data integration, such as grouping, rotation, and joins.



Exercise 1: Integration Data with Rapidminer

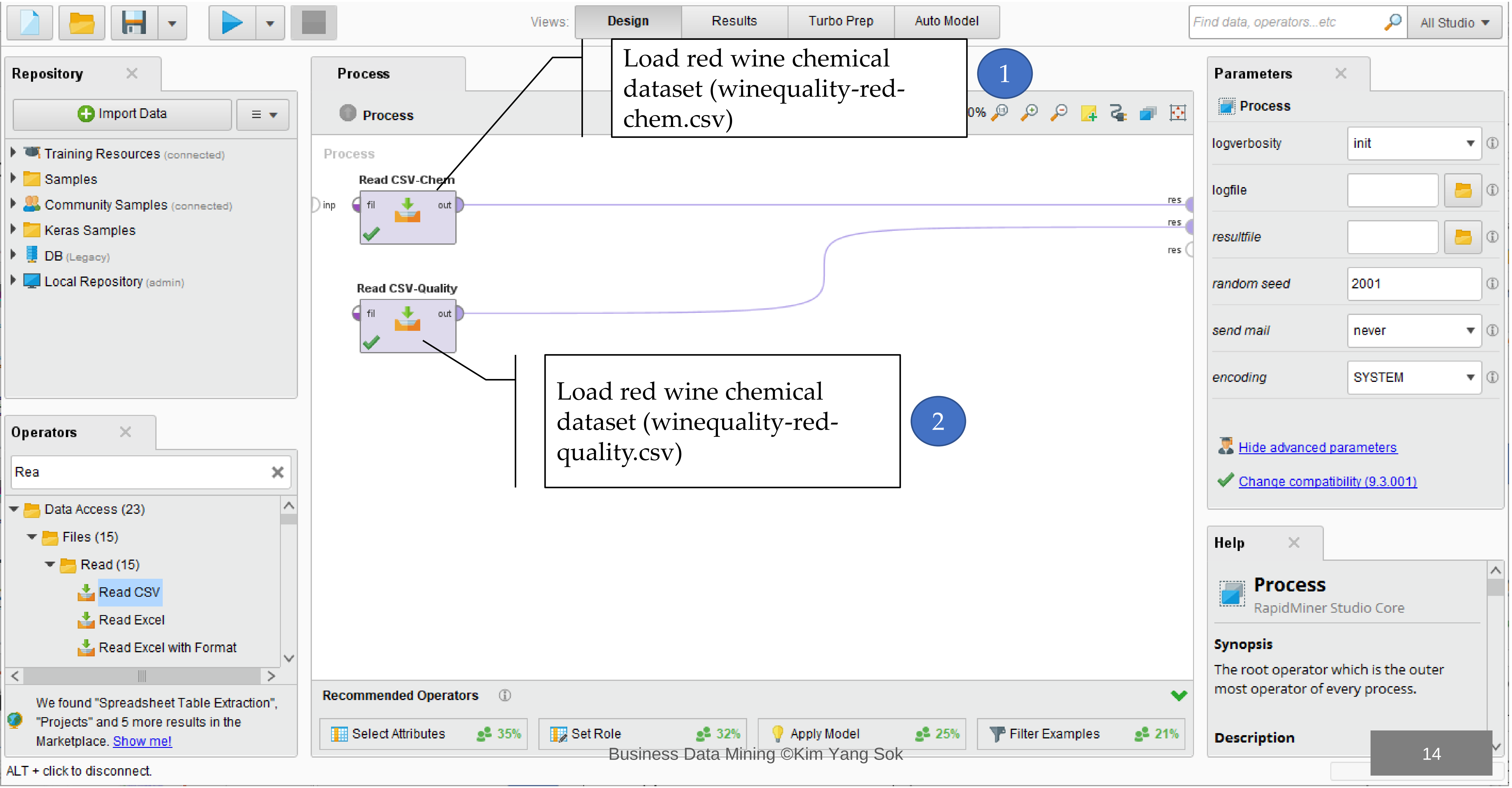
Task & Steps

- **Context**
  - Wine quality data and wine chemical data stored separate in two different tables and we wish to integrate them.

- **Tasks**
  - Load two datasets and integrate them into single dataset

- **Steps**
  1. Load red wine chemical properties data
  2. Load red wind quality data
  3. Combine two datasets into single data
  4. Do steps 1 ~ 3 for white wind data
  5. Combine red and white wine data

Load chemical data of red wine



Load chemical data of red wine

Filter (12 / 12 attributes): Search for Attributes

Name	Type	Missing	Min	Max	Average
fixed acidity	Real	0	4.600	15.900	8.320
volatile acidity	Real	21	0.120	1.580	0.527
citric acid	Real	0	0	1	0.271
residual sugar	Real	0	0.900	15.500	2.539
chlorides	Real	0	0.012	0.611	0.087
free sulfur dioxide	Integer	19	1	72	15.818
total sulfur dioxide	Integer	0	6	289	46.467
density	Real	0	0.990	1.004	0.997
pH	Real	0	2.740	4.010	3.311
sulphates	Real	0	0.330	2	0.658
alcohol	Real	0	8.400	14.900	10.423
id	Integer	0	1	1599	800

A statistical summary of chemical data of red wine

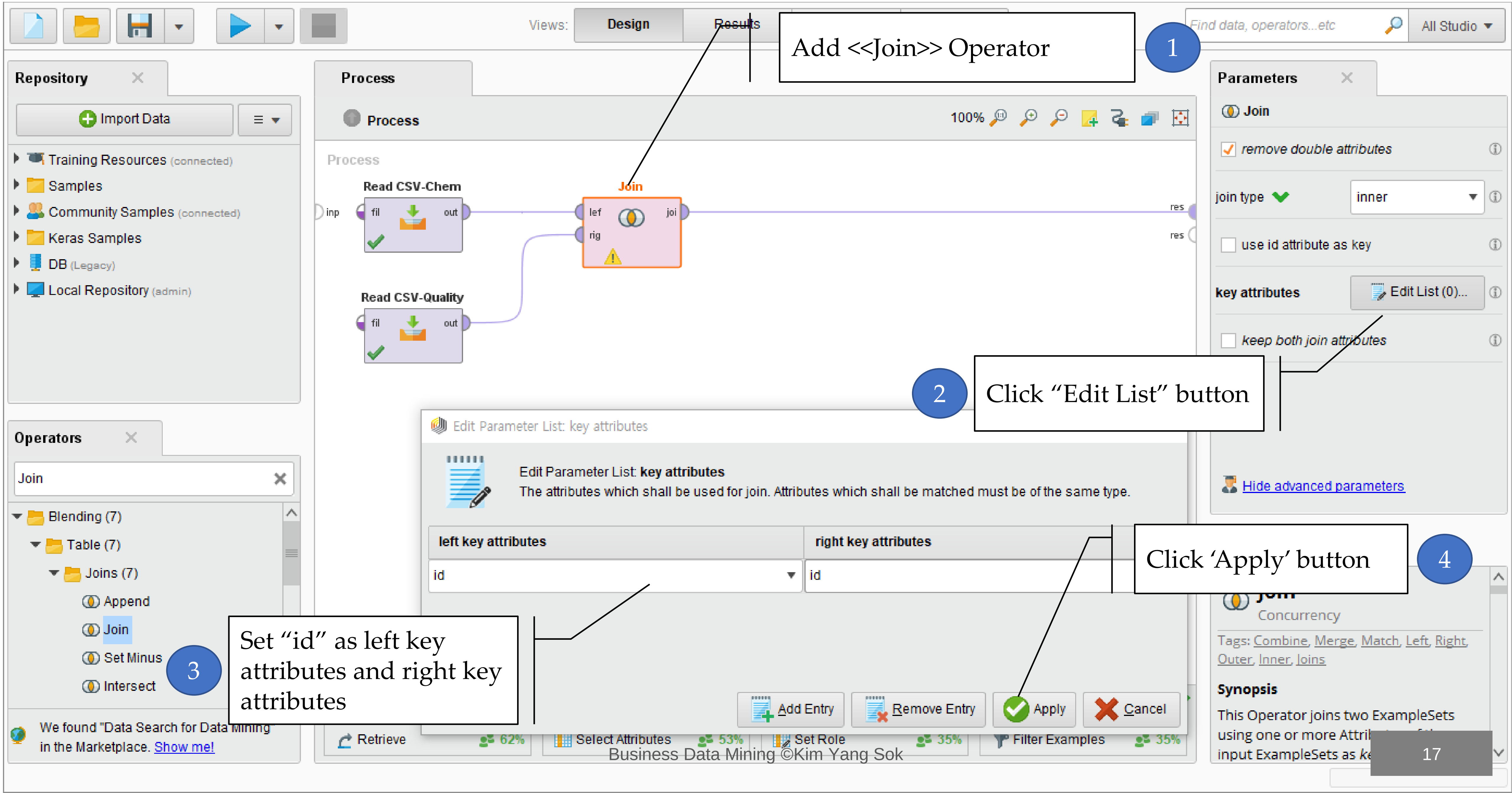
Load quality data of red wind

Filter (2 / 2 attributes): Search for Attributes

Name	Type	Missing	Statistics		
			Min	Max	Average
quality	Integer	0	3	8	5.636
id	Integer	0	1	1599	800

A statistical summary of quality data of red wine

Combine two datasets into single data



Combine two datasets into single data

Views: Design | Results | Turbo Prep | Auto Model

Find data, operators...etc | All Studio

Result History | ExampleSet (Join)

Data | Statistics | Visualizations | Annotations

Open in: Turbo Prep | Auto Model

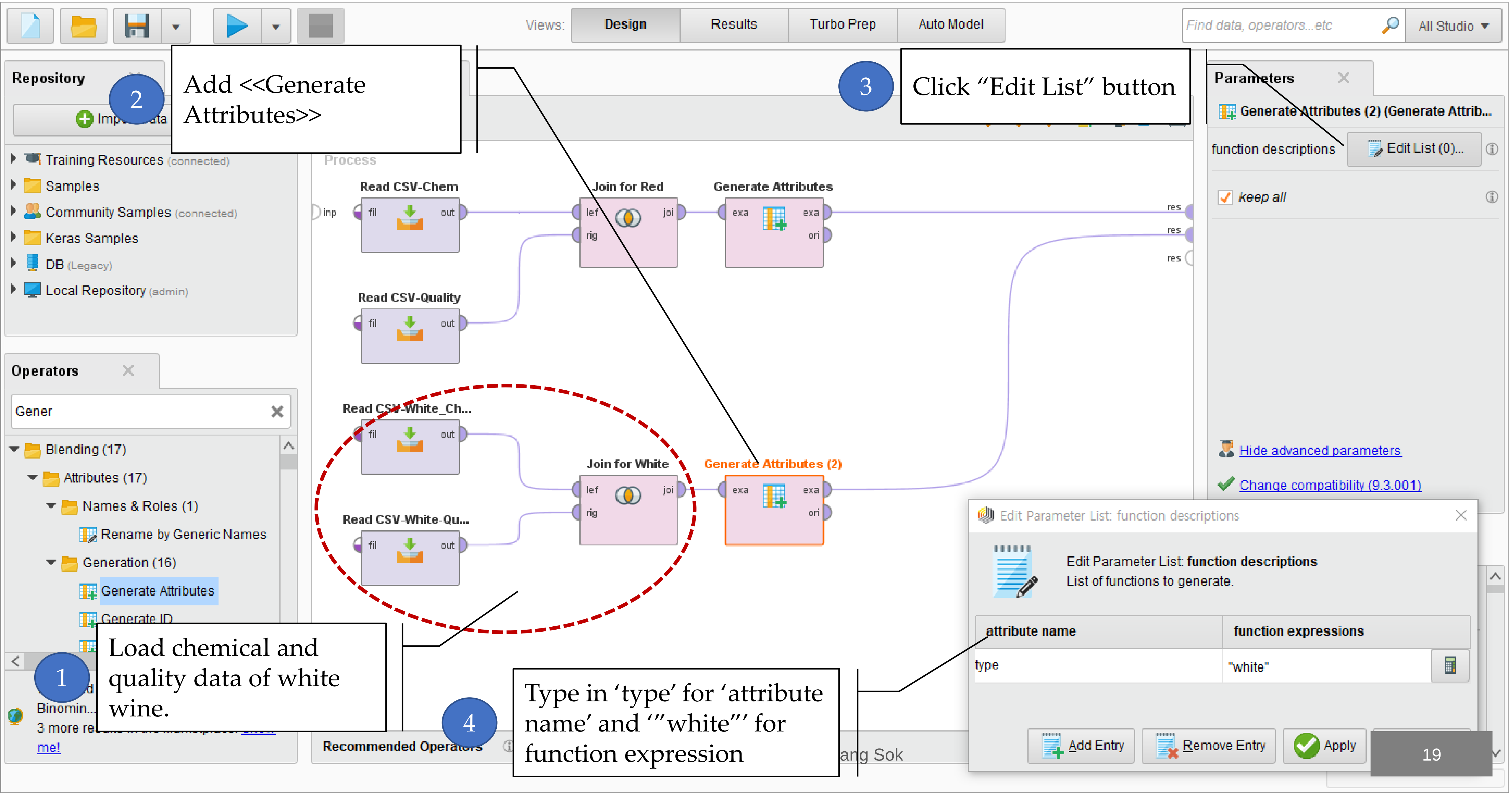
Filter (1,599 / 1,599 examples): all

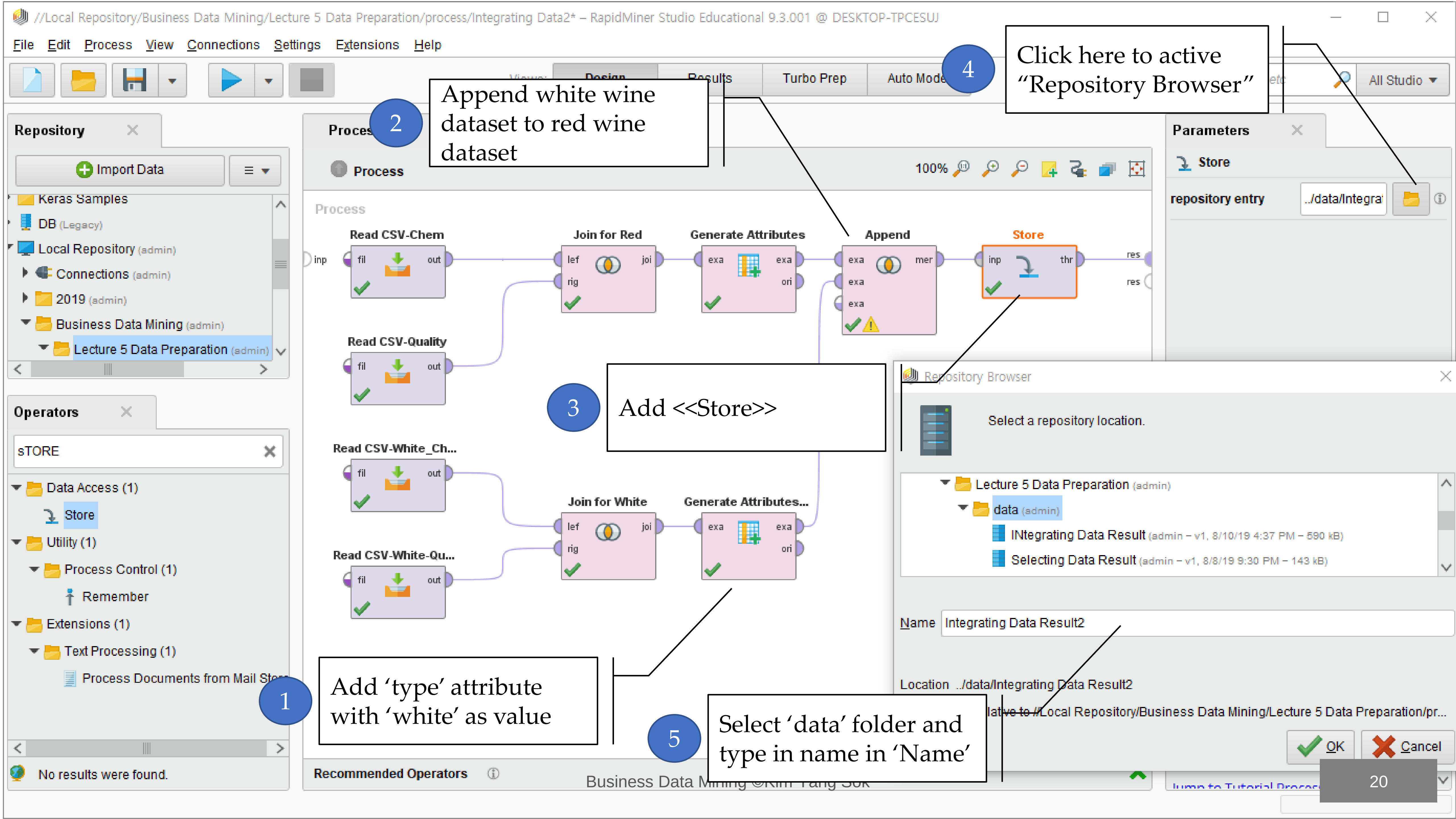
lity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	id	quality
	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	9.400	1	5
	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680	[illegible]	[illegible]	[illegible]
	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650	[illegible]	[illegible]	[illegible]
	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580	[illegible]	[illegible]	[illegible]
	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560	[illegible]	[illegible]	[illegible]
	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560	9.400	6	5
	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460	9.400	7	5
	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470	10	8	7
	0.580	0.020	2	0.073	9	18	0.997	3.360	0.570	9.500	9	7
	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	10	5
	0.580	0.080	1.800	0.097	15	65	0.996	3.280	0.540	9.200	11	5
	0.500	0.360	6.100	0.071	17	102	0.998	3.350	0.800	10.500	12	5
	0.615	0	1.600	0.089	16	59	0.994	3.580	0.520	9.900	13	5
	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560	9.100	14	5
	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880	9.200	15	5
	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930	9.200	16	5

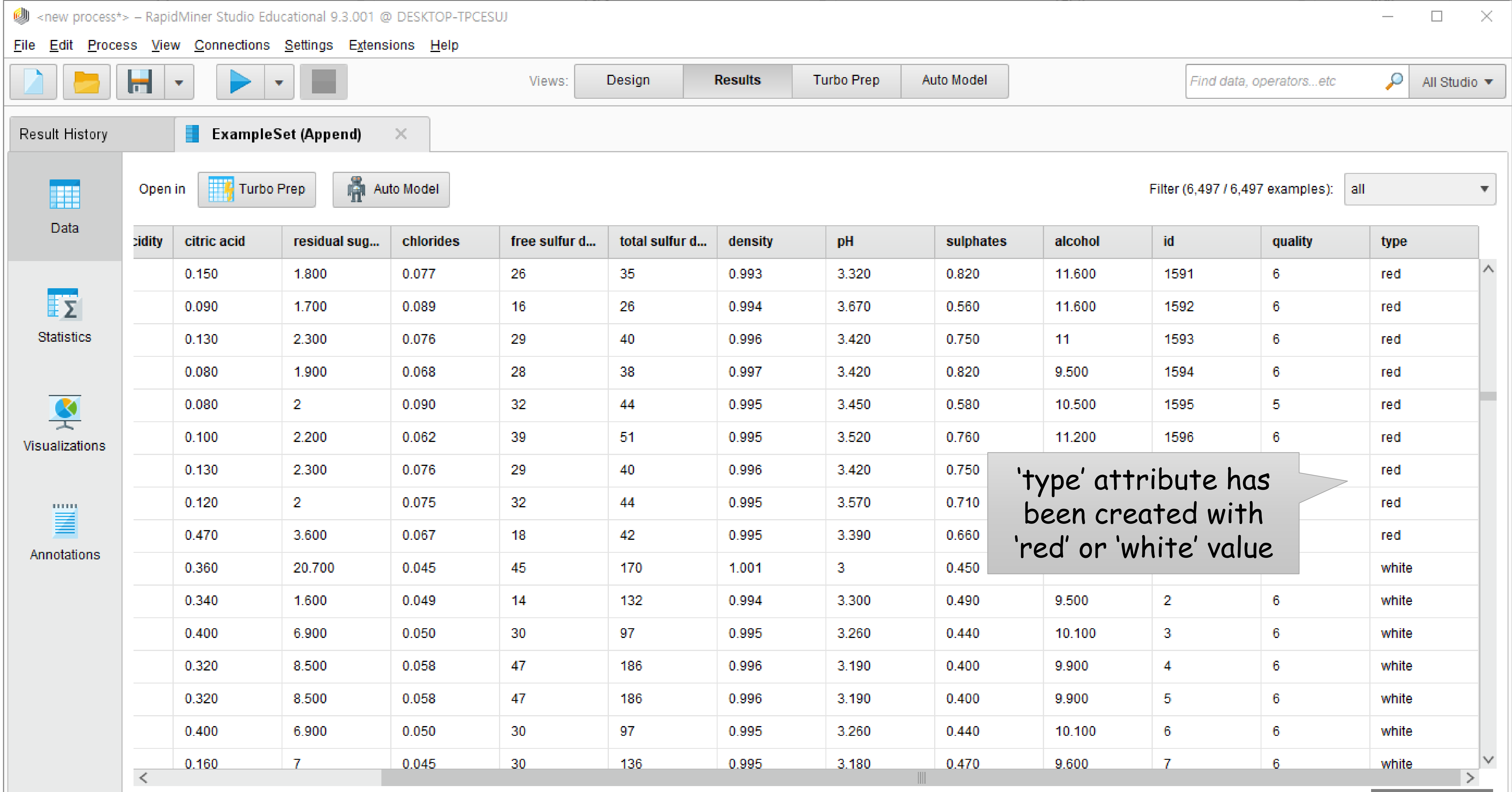
'quality' is combined with chemical attributes

ExampleSet (1,599 examples, 0 special attributes, 13 regular attributes)

Do steps 1 ~ 3 for white wind data




//Local Repository/Business Data Mining/Lecture 5 Data Preparation/process/Integrating Data2* – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ
File Edit Process View Connections Settings Extensions Help
Design Results Turbo Prep Auto Model
All Studio
Repository
Import Data
Keras Samples
DB (Legacy)
Local Repository (admin)
Connections (admin)
2019 (admin)
Business Data Mining (admin)
Lecture 5 Data Preparation (admin)
Operators
sTORE
Data Access (1)
Store
Utility (1)
Process Control (1)
Remember
Extensions (1)
Text Processing (1)
Process Documents from Mail Store
No results were found.
Process
100%
Read CSV-Chem
Join for Red
Generate Attributes
Append
Store
Read CSV-Quality
Read CSV-White_Ch...
Join for White
Generate Attributes...
Read CSV-White-Qu...
Recommended Operators
Parameters
Store
repository entry
../data/Integra
1
Add ‘type’ attribute with ‘white’ as value
2
Append white wine dataset to red wine dataset
3
Add <<Store>>
4
Click here to active “Repository Browser”
5
Select ‘data’ folder and type in name in ‘Name’
Repository Browser
Select a repository location.
Lecture 5 Data Preparation (admin)
data (admin)
INtegrating Data Result (admin – v1, 8/10/19 4:37 PM – 590 kB)
Selecting Data Result (admin – v1, 8/8/19 9:30 PM – 143 kB)
Name Integrating Data Result2
Location ../data/Integrating Data Result2
OK
Cancel



cidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulphates	alcohol	id	quality	type
	0.150	1.800	0.077	26	35	0.993	3.320	0.820	11.600	1591	6	red
	0.090	1.700	0.089	16	26	0.994	3.670	0.560	11.600	1592	6	red
	0.130	2.300	0.076	29	40	0.996	3.420	0.750	11	1593	6	red
	0.080	1.900	0.068	28	38	0.997	3.420	0.820	9.500	1594	6	red
	0.080	2	0.090	32	44	0.995	3.450	0.580	10.500	1595	5	red
	0.100	2.200	0.062	39	51	0.995	3.520	0.760	11.200	1596	6	red
	0.130	2.300	0.076	29	40	0.996	3.420	0.750				red
	0.120	2	0.075	32	44	0.995	3.570	0.710				red
	0.470	3.600	0.067	18	42	0.995	3.390	0.660				red
	0.360	20.700	0.045	45	170	1.001	3	0.450				white
	0.340	1.600	0.049	14	132	0.994	3.300	0.490	9.500	2	6	white
	0.400	6.900	0.050	30	97	0.995	3.260	0.440	10.100	3	6	white
	0.320	8.500	0.058	47	186	0.996	3.190	0.400	9.900	4	6	white
	0.320	8.500	0.058	47	186	0.996	3.190	0.400	9.900	5	6	white
	0.400	6.900	0.050	30	97	0.995	3.260	0.440	10.100	6	6	white
	0.160	7	0.045	30	136	0.995	3.180	0.470	9.600	7	6	white

ExampleSet (6,497 examples, 0 special attributes, 14 regular attributes)

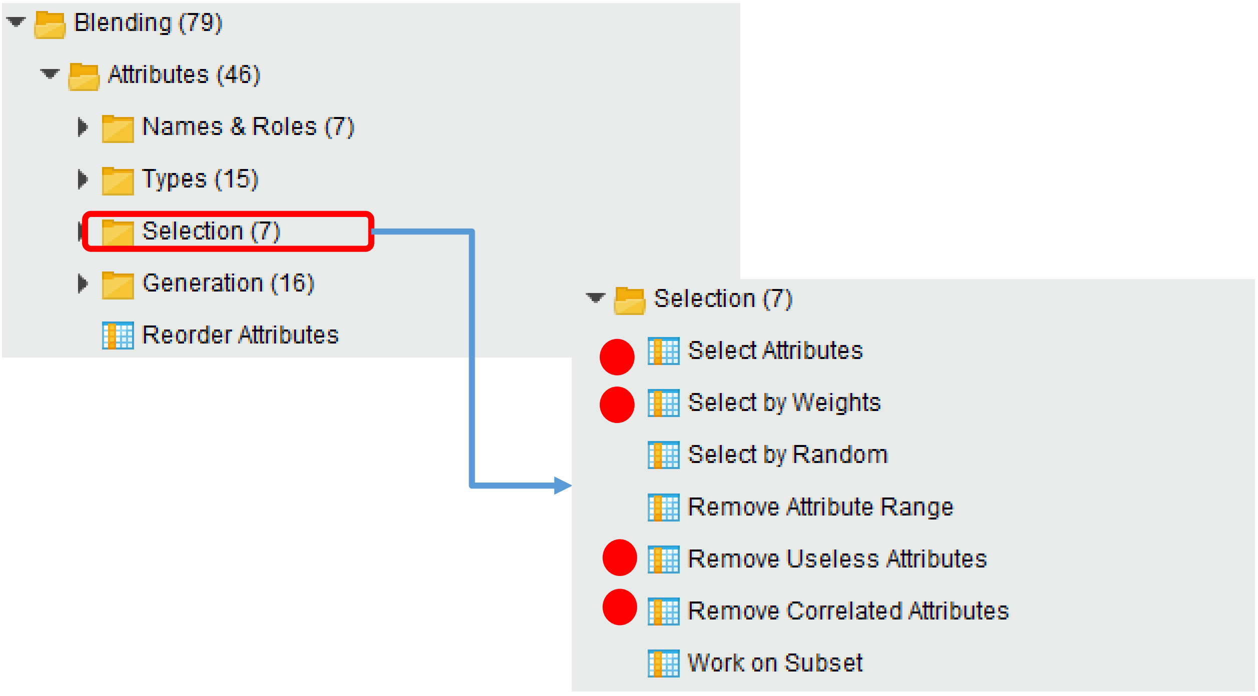
Exercise 2: Selecting Data with Rapidminer

Exercise 2: Selecting Data with Rapidminer

Data Selection

- Data selection covers selection of attributes (columns) as well as selection of examples (rows) in a dataset.

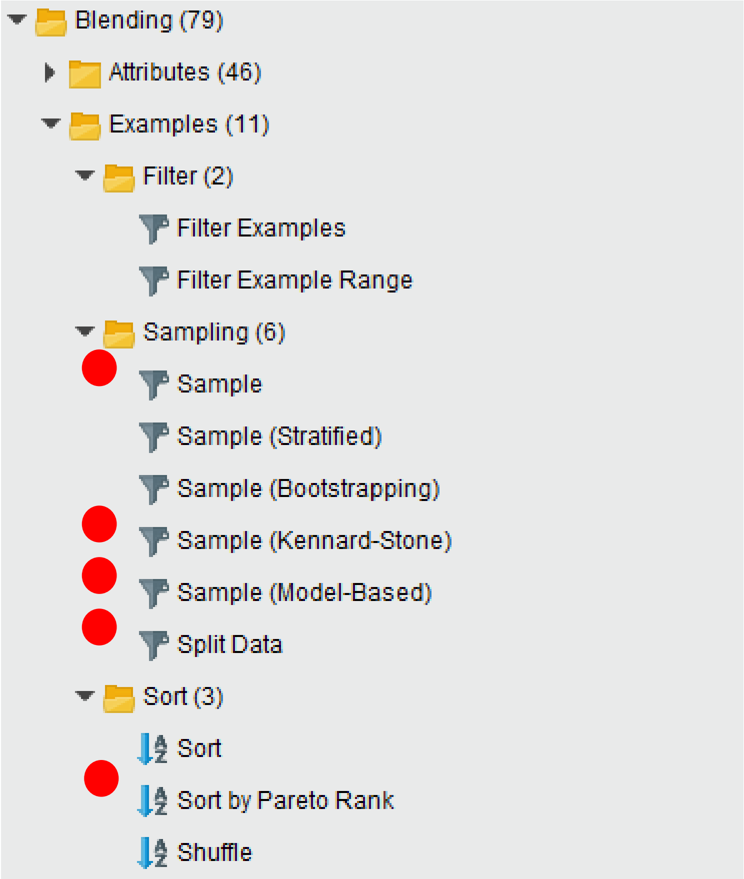
- You can select attributes manually(e.g., Select Attributes) or automatically (e.g., Select by Weight, Remove Useless Attributes, Remove Correlated Attributes).



Exercise 2: Selecting Data with Rapidminer

Data Selection

- **You can select examples using filtering (e.g., Filter Examples and sampling operators (e.g., Sample)**
- **Sampling is used to select subsets of examples for analysis. Sampling types are as follows:**
  - Linear Sampling: Select subset of examples sequentially(linearly)
  - Random Sampling: Select subset of examples randomly
  - Stratified Sampling: Select subset of examples randomly but class distribution unchanged.
- **Filtering is used to select subsets of examples using filtering condition/s.**



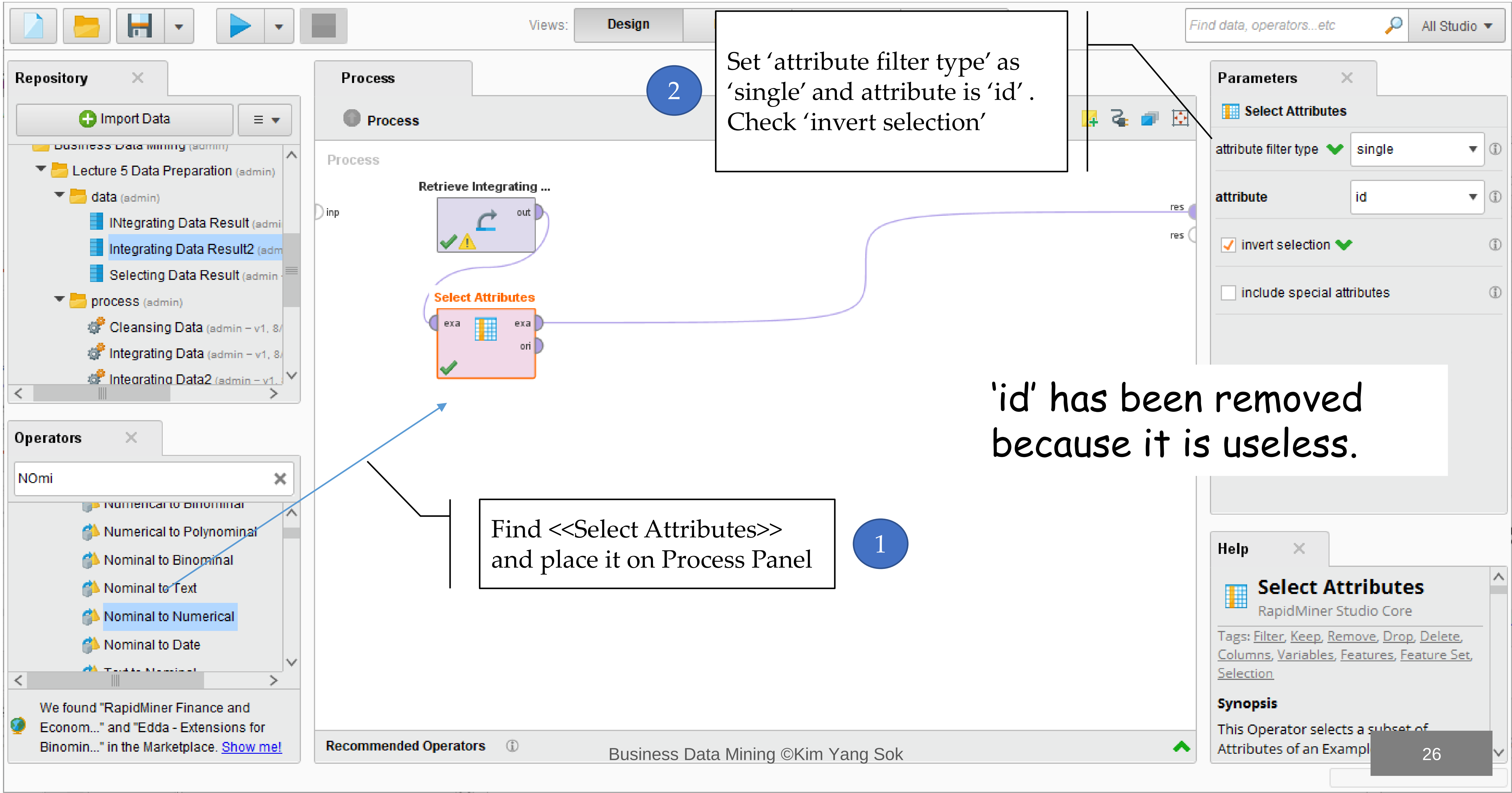
Exercise 2 : Selecting Data with Rapidminer

Task & Steps

- **Tasks**
  - Select useful subset of dataset from the merged dataset using attribute / example selection operators

- **Steps**
  1. Select attributes except 'id'
  2. Change 'type' into numeric using dummy coding
  3. Remove useless attributes
  4. Remove correlated attributes
  5. Filter examples with certain attribute's threshold
  6. Get subset of examples using sampling techniques
  7. Save "Selecting Data" Result

Select attributes except 'id'.



2

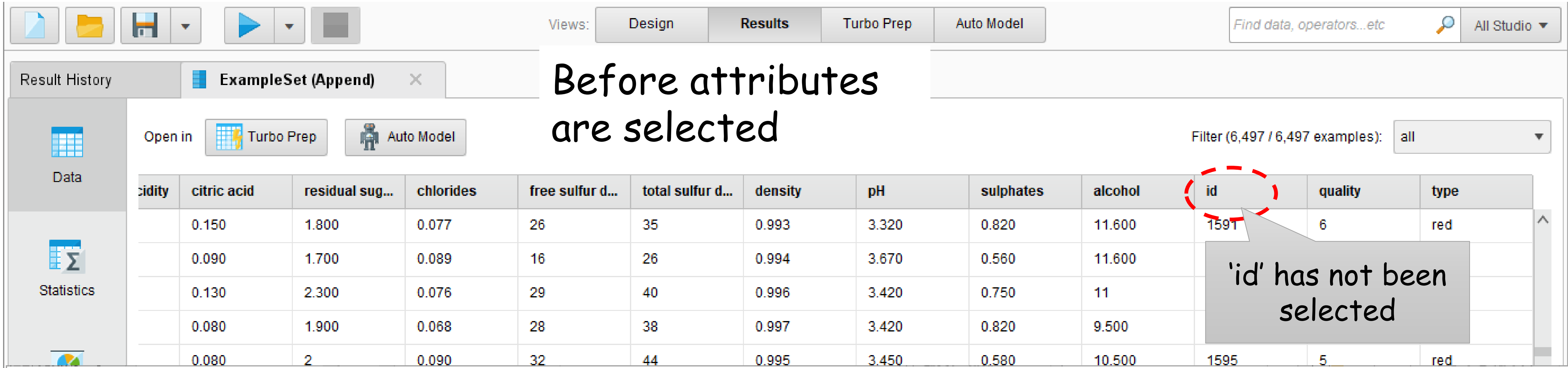
Set 'attribute filter type' as 'single' and attribute is 'id' . Check 'invert selection'

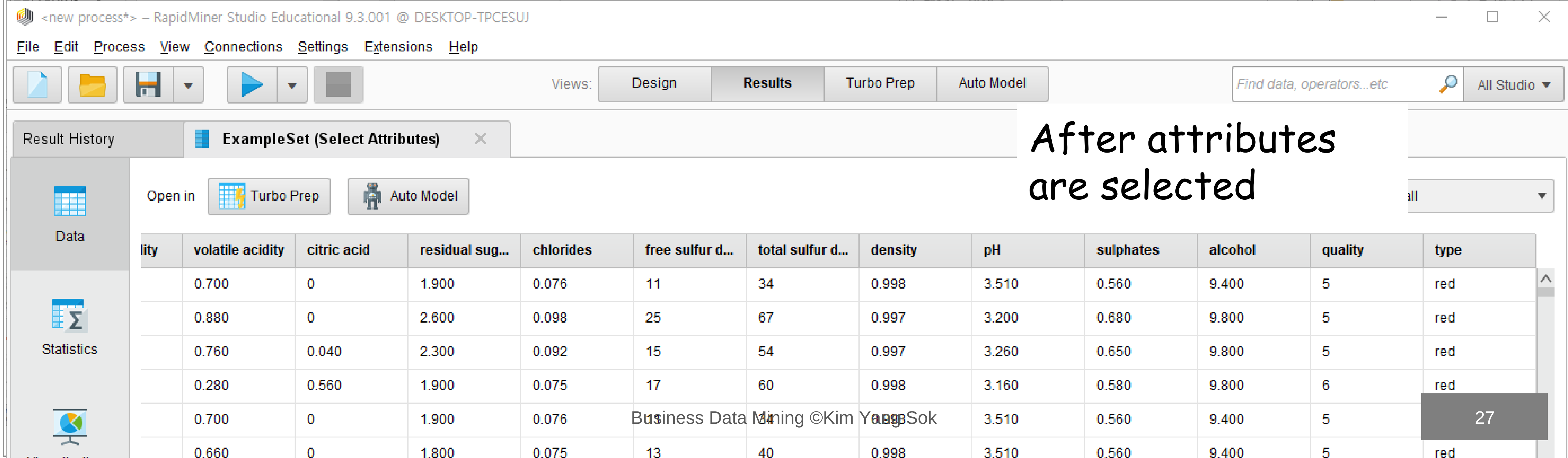
'id' has been removed because it is useless.

Find <<Select Attributes>> and place it on Process Panel

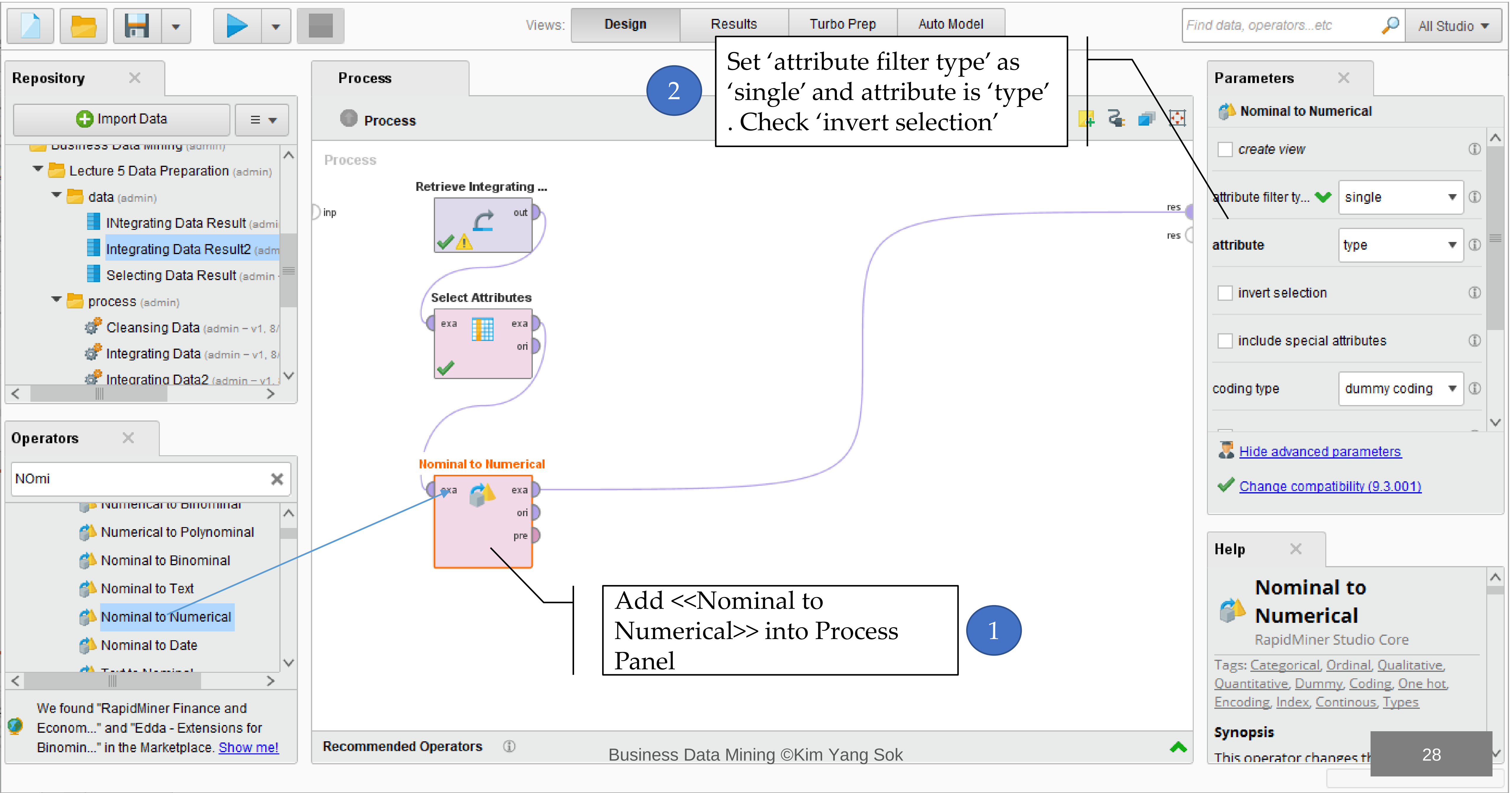
1

Select attributes except 'id'.

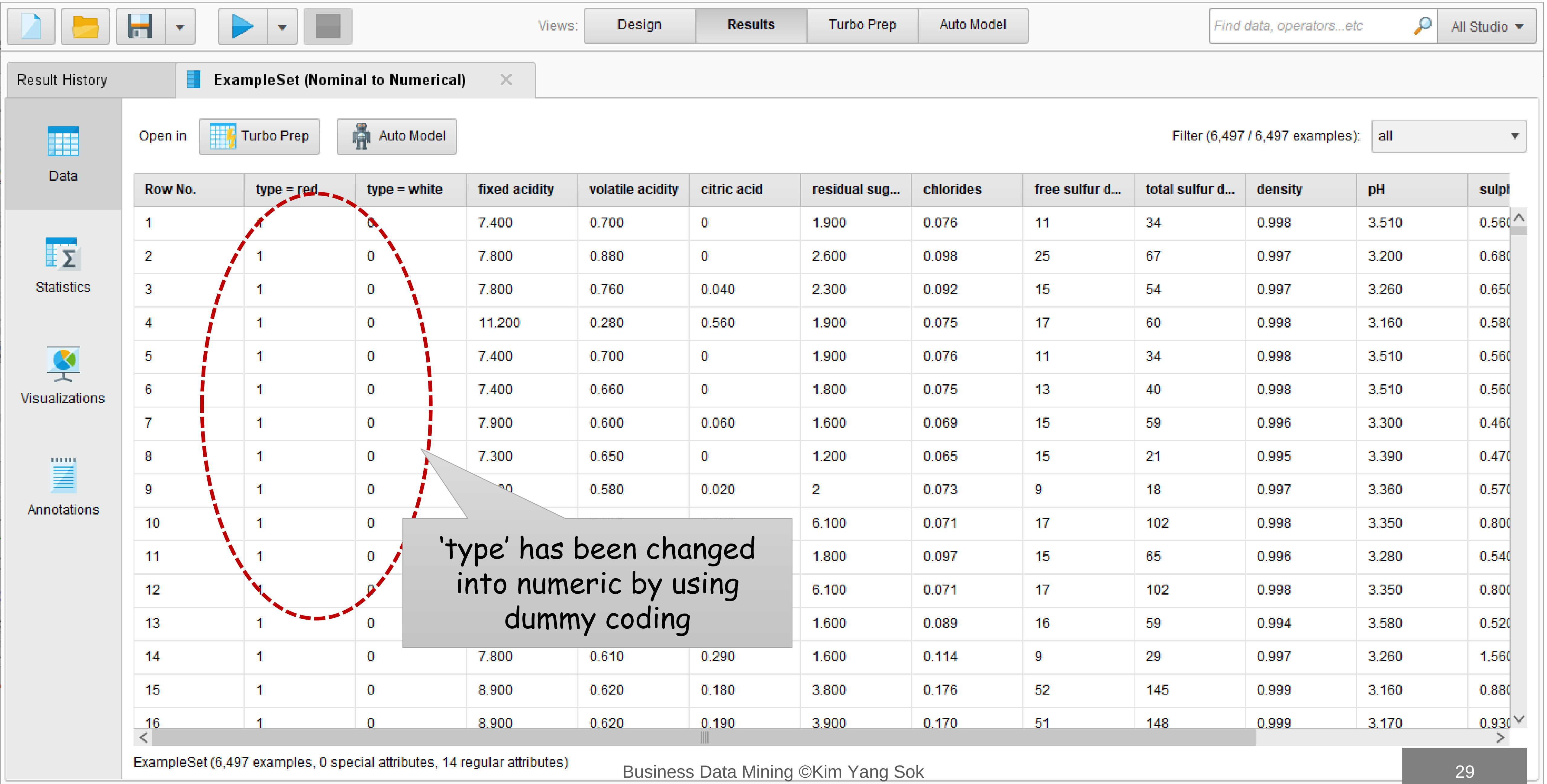




Change 'type' into numeric using dummy coding



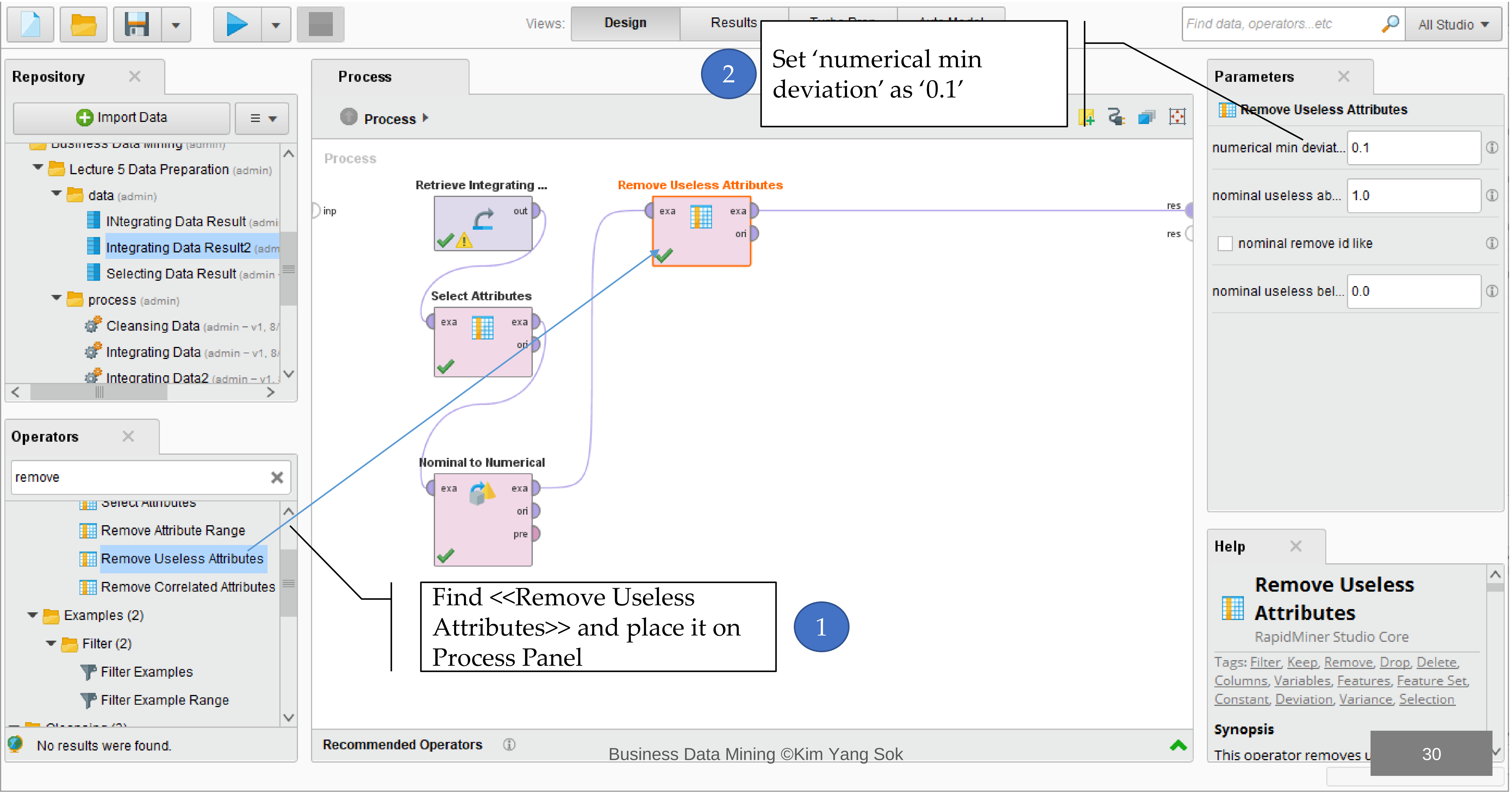
Change 'type' into numeric using dummy coding



Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulpl
1	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
2	1	0	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.680
3	1	0	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.650
4	1	0	11.200	0.280	0.560	1.900	0.075	17	60	0.998	3.160	0.580
5	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.560
6	1	0	7.400	0.660	0	1.800	0.075	13	40	0.998	3.510	0.560
7	1	0	7.900	0.600	0.060	1.600	0.069	15	59	0.996	3.300	0.460
8	1	0	7.300	0.650	0	1.200	0.065	15	21	0.995	3.390	0.470
9	1	0		0.580	0.020	2	0.073	9	18	0.997	3.360	0.570
10	1	0				6.100	0.071	17	102	0.998	3.350	0.800
11	1	0				1.800	0.097	15	65	0.996	3.280	0.540
12	1	0				6.100	0.071	17	102	0.998	3.350	0.800
13	1	0				1.600	0.089	16	59	0.994	3.580	0.520
14	1	0	7.800	0.610	0.290	1.600	0.114	9	29	0.997	3.260	1.560
15	1	0	8.900	0.620	0.180	3.800	0.176	52	145	0.999	3.160	0.880
16	1	0	8.900	0.620	0.190	3.900	0.170	51	148	0.999	3.170	0.930

'type' has been changed into numeric by using dummy coding

Remove useless attributes based on deviation



Remove useless attributes based on deviation

'chloride' is removed from dataset

'density' is removed from dataset

ExampleSet (Nominal to Numerical)

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	chlorides	free sulfur d...	total sulfur d...	density	pH	sulp
1	1	0	7.400	0.700	0	1.900	0.076	11	34	0.998	3.510	0.56
2	1	0	7.800	0.880	0	2.600	0.098	25	67	0.997	3.200	0.68
3	1	0	7.800	0.760	0.040	2.300	0.092	15	54	0.997	3.260	0.65
4	1	0	11.200	0.280	0.560	1.90	[illegible]	[illegible]	[illegible]	0.998	3.160	0.58
5	1	0	7.400	0.700	0	1.90	[illegible]	[illegible]	[illegible]	0.998	3.510	0.56
6	1	0	7.400	0.660	0	1.80	[illegible]	[illegible]	[illegible]	0.998	3.510	0.56

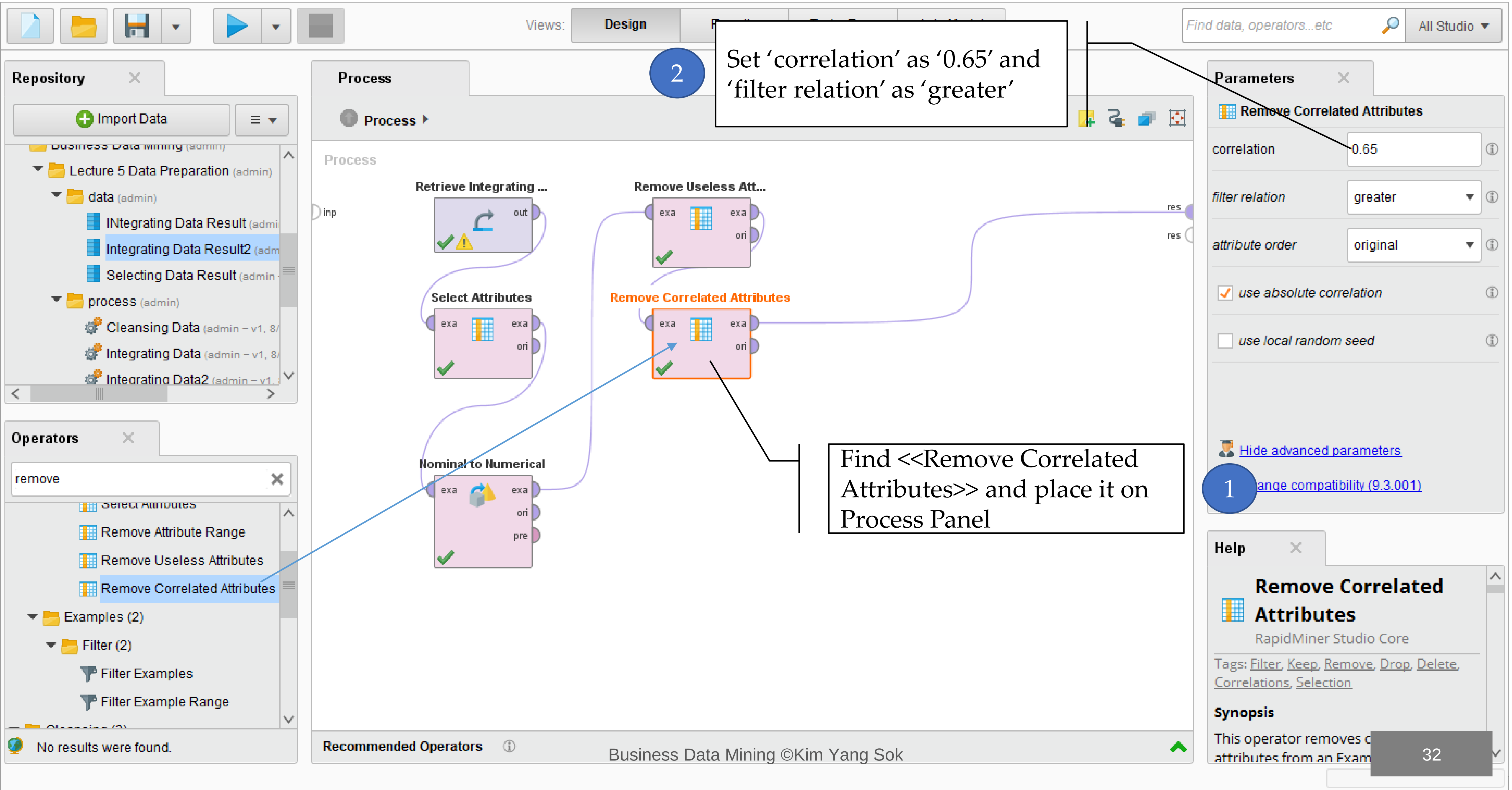
After removing useless attributes

ExampleSet (Remove Useless Attributes)

Filter (6,497 / 6,497 examples): all

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	total sulfur d...	pH	sulphates	alcohol	quali
1	1	0	7.400	0.700	0	1.900	11	34	3.510	0.560	9.400	5
2	1	0	7.800	0.880	0	2.600	25	67	3.200	0.680	9.800	5
3	1	0	7.800	0.760	0.040	2.300	15	54	3.260	0.650	9.800	[illegible]
4	1	0	11.200	0.280	0.560	1.900	17	60	3.160	0.580	9.800	6

Remove correlated attributes



Remove correlated attributes

Views: Design | Results | Turbo Prep | Auto Model

Result History | ExampleSet (Remove Useless Attributes)

Data | Statistics | Visualizations

Open in: Turbo Prep | Auto Model

Filter (6,497 / 6,497 examples): all

Row No.	type = red	type = white	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	total sulfur d...	pH	sulphates	alcohol	quali
1	1	0	7.400	0.700	0	1.900	11	34	3.510	0.560	9.400	5
2	1	0	7.800	0.880		2.600	25	67	3.200	0.680	9.800	5
3	1	0					15	54	3.260	0.650	9.800	5
4	1	0					17	60	3.160	0.580	9.800	6
5	1	0					11	34	3.510	0.560	9.400	5
6	1	0	7.400	0.660	0	1.800	13					5

'type=white' and 'total sulfur dioxide' is removed from dataset

After removing correlated attributes

<new process*> – RapidMiner Studio Educational 9.3.001 @ DESKTOP-TPCESUJ

File Edit Process View Connections Settings Extensions Help

Views: Design | Results | Turbo Prep | Auto Model

Result History | ExampleSet (Remove Useless Attributes) | ExampleSet (Remove Correlated Attributes)

Data | Statistics

Open in: Turbo Prep | Auto Model

Filter (6,497 / 6,497 examples): all

Row No.	type = red	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	pH	sulphates	alcohol	quality
1	1	7.400	0.700	0	1.900	11	3.510	0.560	9.400	5
2	1	7.800	0.880	0	2.600	25	3.200	0.680	9.800	5
3	1	7.800	0.760	0.040	2.300	15	3.260	0.650	9.800	5
4	1	11.200	0.280	0.560	1.900	17	3.160	0.580	9.800	6

Filter examples with certain attribute's threshold

Views: Design | Results | Turbo Prep | Auto Model

Result History | ExampleSet (Remove Correlated Attributes)

Data | Statistics | Visualizations | Annotations

Open in Turbo Prep | Auto Model

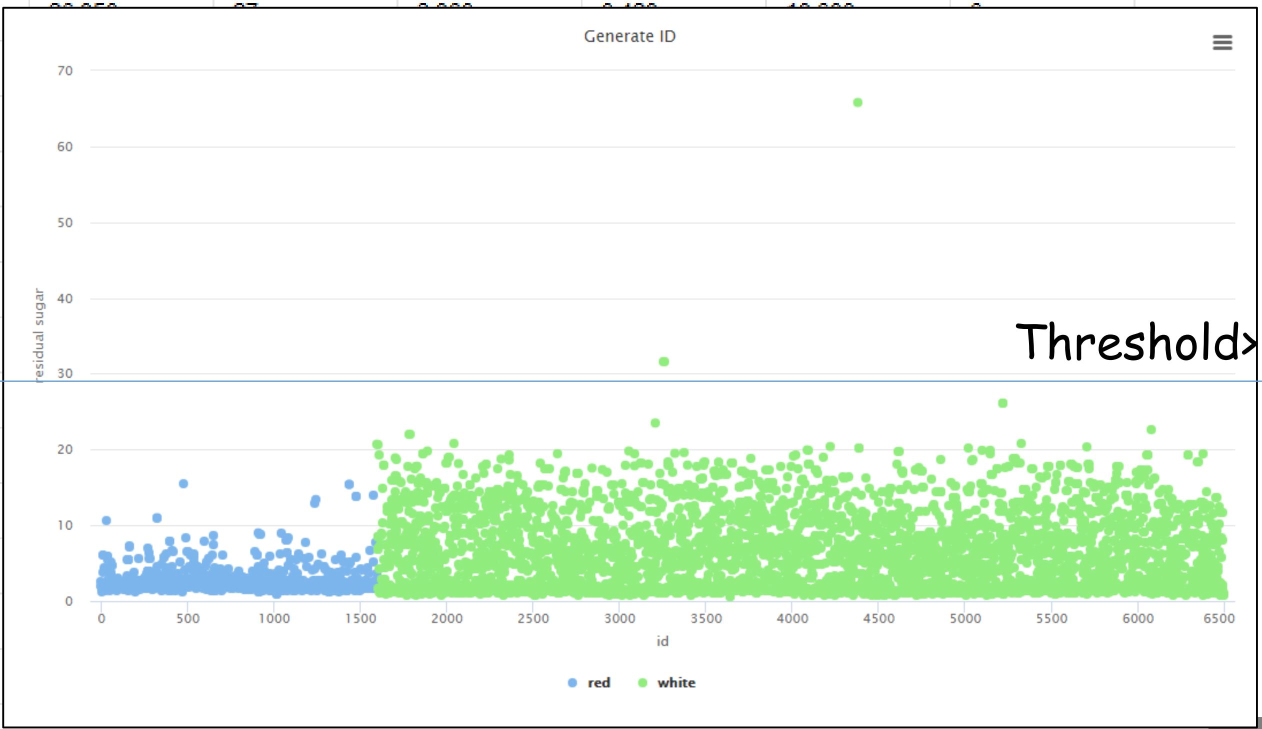
Sort examples with 'residual sugar' attribute in descending order

1

Filter (6,497 / 6,497 examples): all

These three examples have too high value → They might be outliers

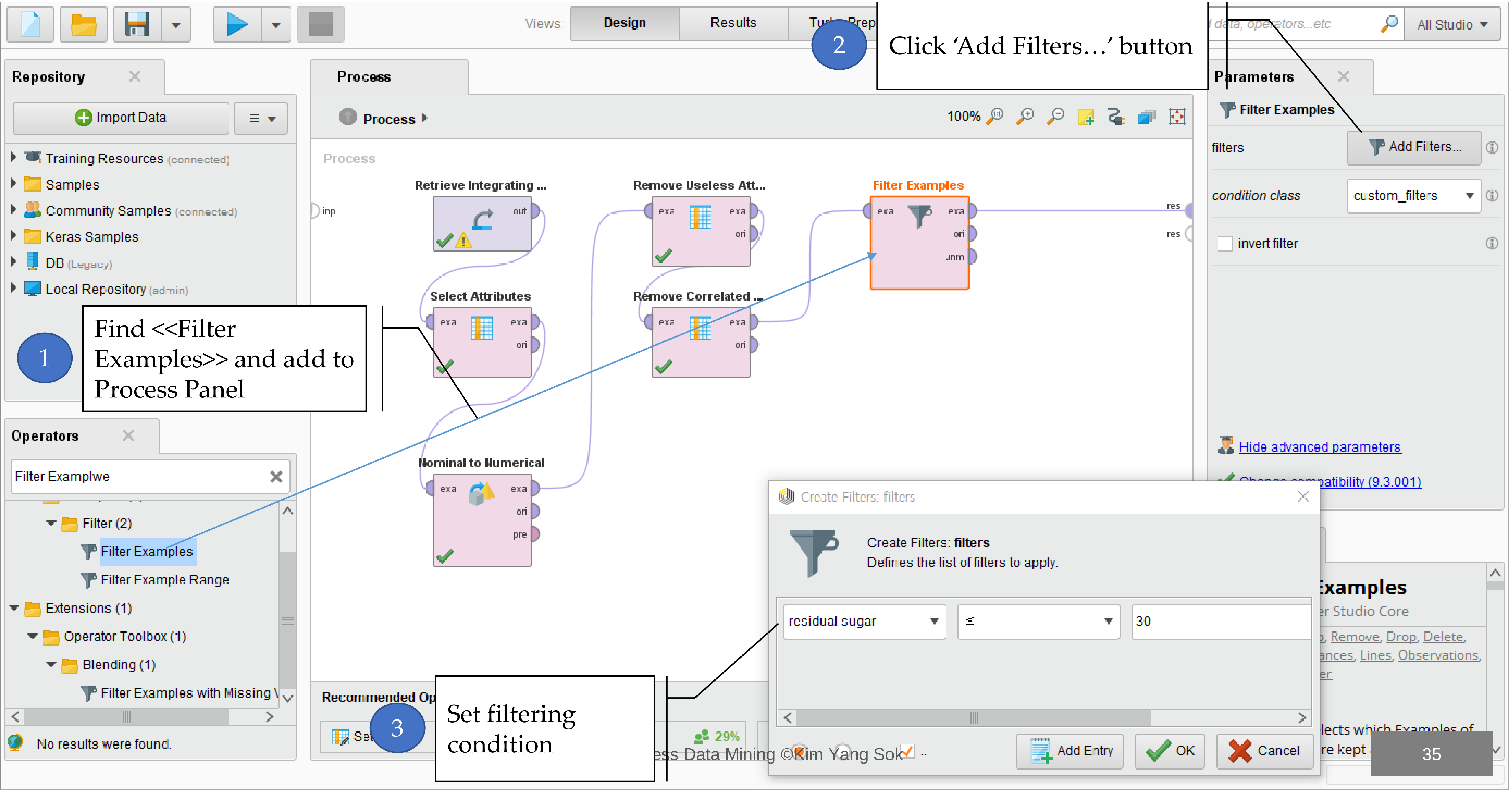
Row No.	type			itric acid	residual ... ↓	free sulfur d...	pH	sulphates	alcohol	quality
4381	0			.600	65.800	8	3.390	0.690	11.700	6
3253	0				31.600	35	3.150	0.380	8.800	6
3263	0			.280	31.600	35	3.150	0.380	8.800	6
5219	0	6.800	0.450	0.280						
5223	0	6.800	0.450	0.280						
3208	0	6.900	0.270	0.490						
6080	0	5.900	0.220	0.450						
1782	0	6.800	0.280	0.400						
1791	0	6.800	0.280	0.400						
2044	0	6.900	0.240	0.360						
5330	0	6.200	0.220	0.200						
1600	0	7	0.270	0.360						
1607	0	7	0.270	0.360						
4220	0	6.500	0.280	0.280						
5707	0	6.800	0.300	0.260						
4385	0	6.400	0.240	0.250						



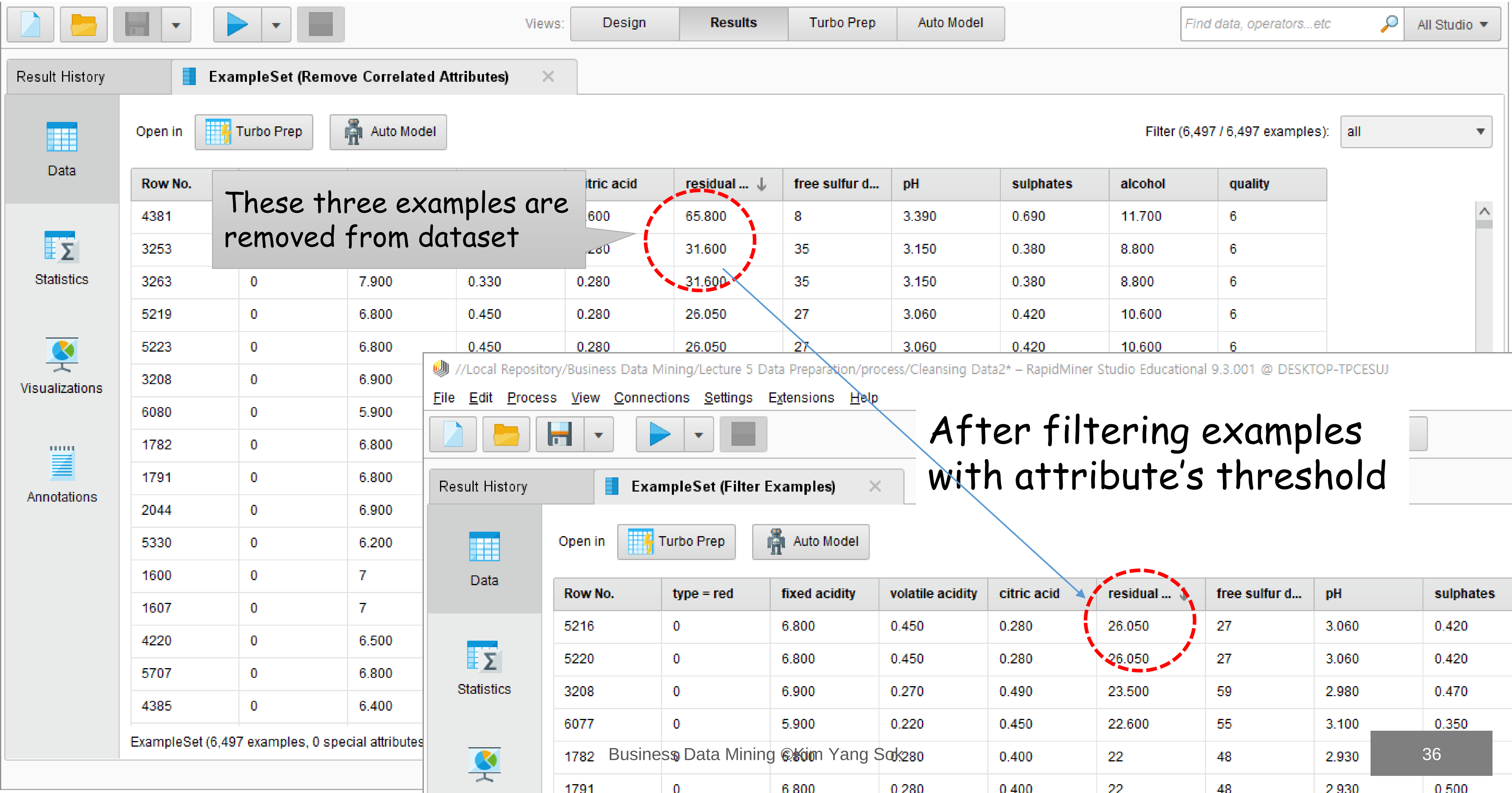
Threshold>30

ExampleSet (6,497 examples, 0 special attributes, 10 regular attributes)

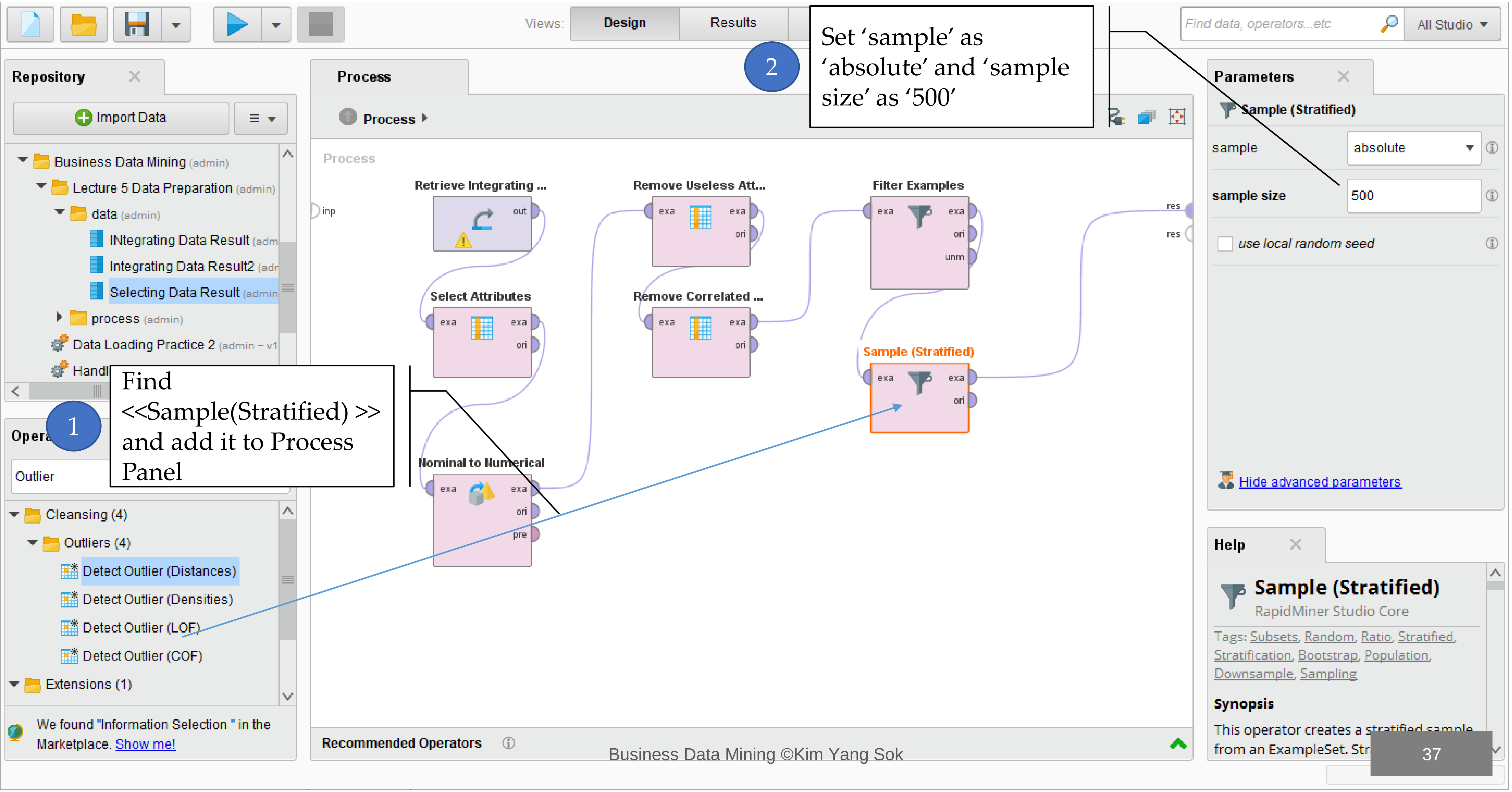
Filter examples with certain attribute's threshold



Filter examples with certain attribute's threshold



Get subset of examples using stratified sampling techniques



Get subset of examples using sampling techniques

Views: Design | Results | Turbo Prep | Auto Model

Find data, operators...etc | All Studio

Result History | ExampleSet (Sample (Stratified))

Data

Statistics

Visualizations

Annotations

Open in Turbo Prep | Auto Model

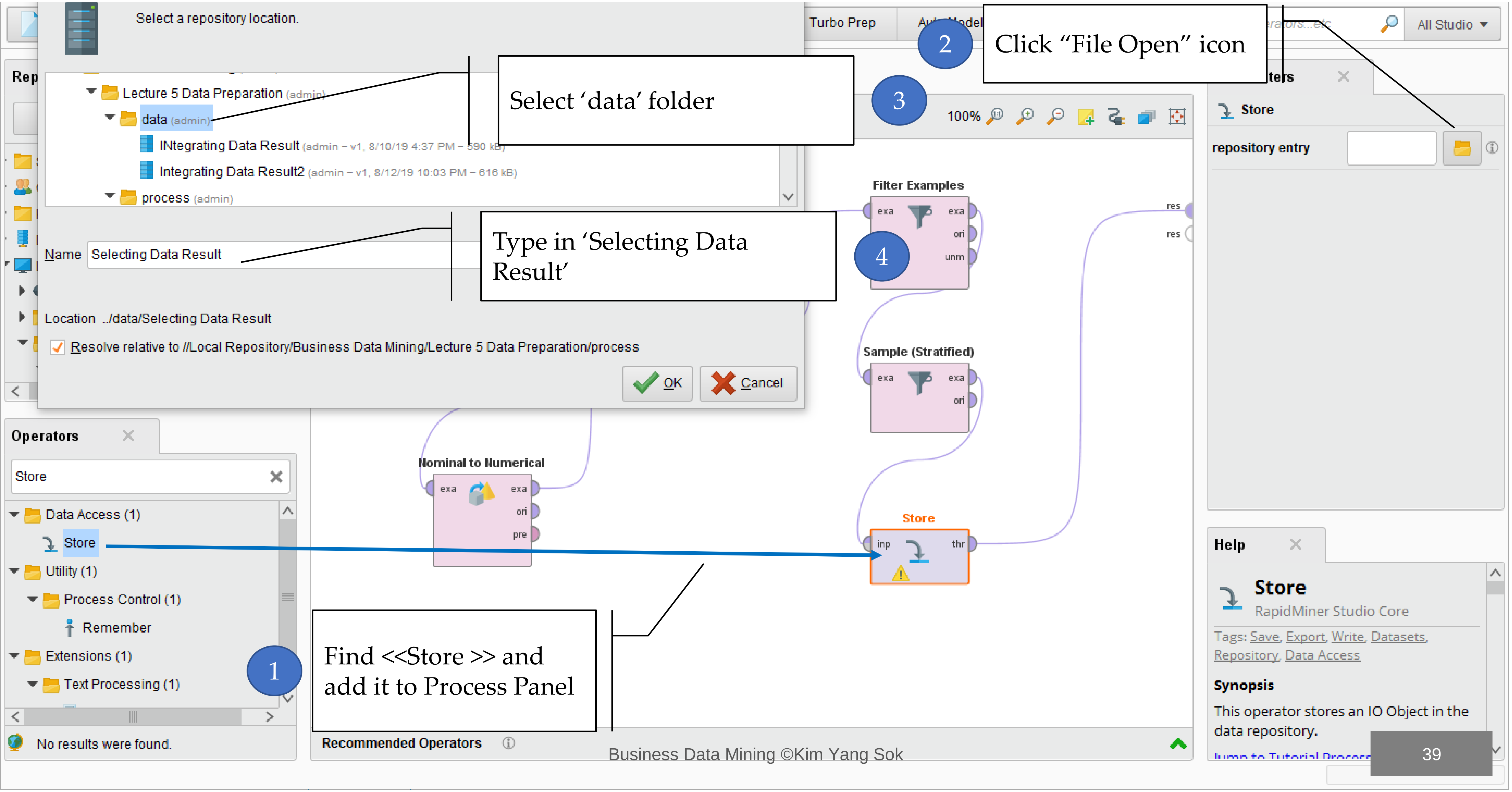
Filter (500 / 500 examples): all

Row No.	type = red	fixed acidity	volatile acidity	citric acid	residual sug...	free sulfur d...	pH	sulphates	alcohol	ty
1	1	6.700	0.580	0.080	1.800	15	3.280			
2	1	7.500	0.500	0.360	6.100	17	3.350			
3	1	6.900	0.685	0	2.500	22	3.460			
4	1	7.800	0.645	0	5.500	5	3.400	0.550	9.600	6
5	1	8.100	0.380	0.280	2.100	13	3.230	0.730	9.700	7
6	1	7.500	0.630	0.120	5.100	50	3.260	0.770	9.400	5
7	1	7.700	0.630	0.080	1.900	15	3.320	0.540	9.500	6
8	1	6.900	0.550	0.150	2.200	19	3.410	0.590	10.100	5
9	1	7	0.620	0.080	1.800	8	3.480	0.530	9	5
10	1	7	0.690	0.080	1.800	22	3.340	0.540	9.200	6
11	1	9	0.620	0.040	1.900	27	3.160	0.700	9.400	5
12	1	7.300	0.330	0.470	2.100	5	3.330	0.530	10.300	6
13	1	9.200	0.520	1	3.400	32	2.740	2	9.400	4
14	1	7.800	0.630	0.480	1.700	14	3.190	0.620	9.500	5
15	1	6.800	0.640	0.100	2.100	18	3.340	0.520	10.200	5
16	1	8.800	0.610	0.140	2.400	10	3.190	0.590	9.500	5

A total of 500 examples are selected.

ExampleSet (500 examples, 0 special attributes, 10 regular attributes)

Save Selecting Data Results

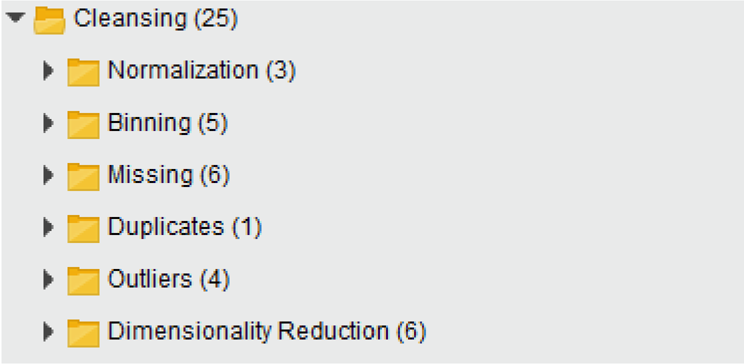


Exercise 3: Cleansing Data with Rapidminer

Exercise: Cleansing Data with Rapidminer

Data Cleansing

- Data cleansing aims to raise the data quality to the level required by the selected analysis techniques.
- Rapidminer supports data cleansing through Normalization, Binning, Missing, Duplicates, Outliers and Dimensionality Reduction packages.



Data Cleansing - Missing Values

- **Many real-world datasets may contain missing values for various reasons.**
- **One way to handle this problem is to get rid of the observations that have missing data. However, you will risk losing data points with valuable information.**
- **A better strategy would be to impute the missing values. In other words, we need to infer those missing values from the existing part of the data.**
- **Note that imputation does not necessarily give better results.**

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Missing Values

- **In order to handle missing values, it is necessary to understand types of missing values.**
- **Missing at Random (MAR):**
  - Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- **Missing Completely at Random (MCAR):**
  - The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- **Missing not at Random (MNAR):**
  - Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

Safe to remove the data with missing values

Removing observations with missing values can produce a bias in the model. So we have to be really careful before removing observations.

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Missing Values

- **Some algorithms can factor in the missing values and learn the best imputation values for the missing data based on the training loss reduction (ie. XGBoost).**

- **Some others have the option to just ignore them.**

- **However, other algorithms will panic and throw an error complaining about the missing values. In that case, you will need to handle the missing data and clean it before feeding it to the algorithm.**

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Missing Values

- **Imputation Using (Mean/Median) Values:**
  - This works by calculating the mean/median of the non-missing values in a column and then replacing the missing values within each column separately and independently from the others. It can only be used with numeric data.

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean() →

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

  - Pros:
    - Easy and fast.
    - Works well with small numerical datasets.
  - Cons:
    - Doesn’t factor the correlations between features. It only works on the column level.
    - Will give poor results on encoded categorical features (do NOT use it on categorical features).
    - Not very accurate.
    - Doesn’t account for the uncertainty in the imputations.

- **Imputation Using (Most Frequent) or (Zero/Constant) Values**
  - Most Frequent is another statistical strategy to impute missing values and YES!! It works with categorical features (strings or numerical representations) by replacing missing data with the most frequent values within each column.
  - Pros:
    - Works well with categorical features.
  - Cons:
    - It also doesn't factor the correlations between features.
    - It can introduce bias in the data.

  - Zero or Constant imputation — as the name suggests — it replaces the missing values with either zero or any constant value you specify

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

df.fillna(0)

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	0.0
1	9	0.0	9.0	0	7.0
2	19	17.0	0.0	9	0.0

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Missing Values

- **We create a predictive model to estimate values that will substitute the missing data.**

- **In this case, we divide our data set into two sets: One set with no missing values for the variable (training) and another one with missing values (test).**
  - Can use various prediction algorithm (e.g., k-NN, logistic regression, linear regression), depending on variables.

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Missing Values

- **In conclusion, there is no perfect way to compensate for the missing values in a dataset.**

- **Each strategy can perform better for certain datasets and missing data types but may perform much worse on other types of datasets.**

- **There are some set rules to decide which strategy to use for particular types of missing values, but beyond that, you should experiment and check which model works best for your dataset.**

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Normalization

- **Normalization aims to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the ranges of values or losing information.**
- **Zscore: Converts all values to a z-score.**
  - The values in the column are transformed using the following formula:

$$z = \frac{\mathrm{x} - mean(x)}{stdev(x)}$$

  - Mean and standard deviation are computed for each column separately. Population standard deviation is used.
- **MinMax: The min-max normalizer linearly rescales every feature to the [0,1] interval.**
  - Rescaling to the [0,1] interval is done by shifting the values of each feature so that the minimal value is 0, and then dividing by the new maximal value (which is the difference between the original maximal and minimal values).
  - The values in the column are transformed using the following formula:

$$z = \frac{\mathrm{x} - min(x)}{[\max(x) - min(x)]}$$

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Binning

- **Binning or discretization is the process of transforming numerical variables into categorical counterparts. e.g., an example is to bin values for Age into categories such as 20-39, 40-59, and 60-79.**
- **Unsupervised Binning methods transform numerical variables into categorical counterparts but do not use the target (class) information.**
  - Equal Width Binning
    - The algorithm divides the data into k intervals of equal size. The width of intervals is:

      w = (max-min)/k
    - And the interval boundaries are:

      min+w, min+2w, ... , min+(k-1)w
  - Equal Frequency Binning
    - The algorithm divides the data into k groups which each group contains approximately same number of values. For the both methods, the best way of determining k is by looking at the histogram and try different intervals or groups.

Data Cleansing - Binning

- **Supervised binning methods transform numerical variables into categorical counterparts and refer to the target (class) information when selecting discretization cut points.**

- **Entropy-based binning is an example of a supervised binning method. The entropy (or the information content) is calculated based on the class label.**
  - Intuitively, it finds the best split so that the bins are as pure as possible that is the majority of the values in a bin correspond to have the same class label.
  - Formally, it is characterized by finding the split with the maximal information gain.

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Outlier Detection

- **An outlier is an observation point that is distant from other observations.**

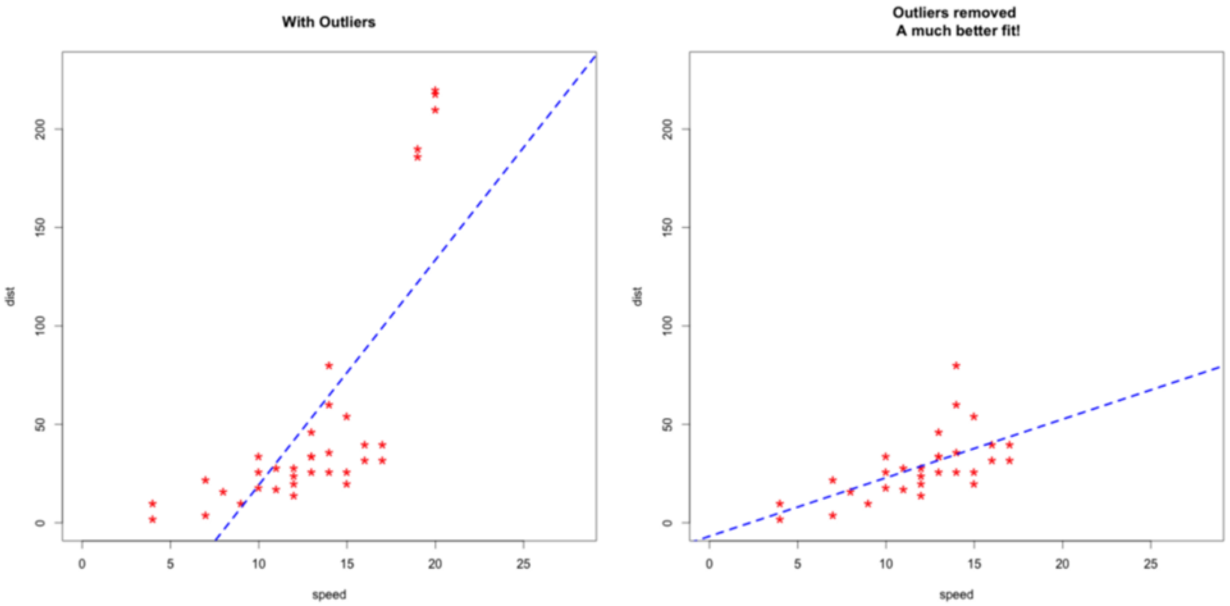
- **Outliers exist due to one of the four following reasons:**
  - Incorrect data entry can cause data to contain extreme cases.
  - A second reason for outliers can be failure to indicate codes for missing values in a dataset.
  - Another possibility is that the case did not come from the intended sample.
  - And finally, the distribution of the sample for specific variables may have a more extreme distribution than normal.

Exercise 3: Cleansing Data with Rapidminer

Data Cleansing - Outliers

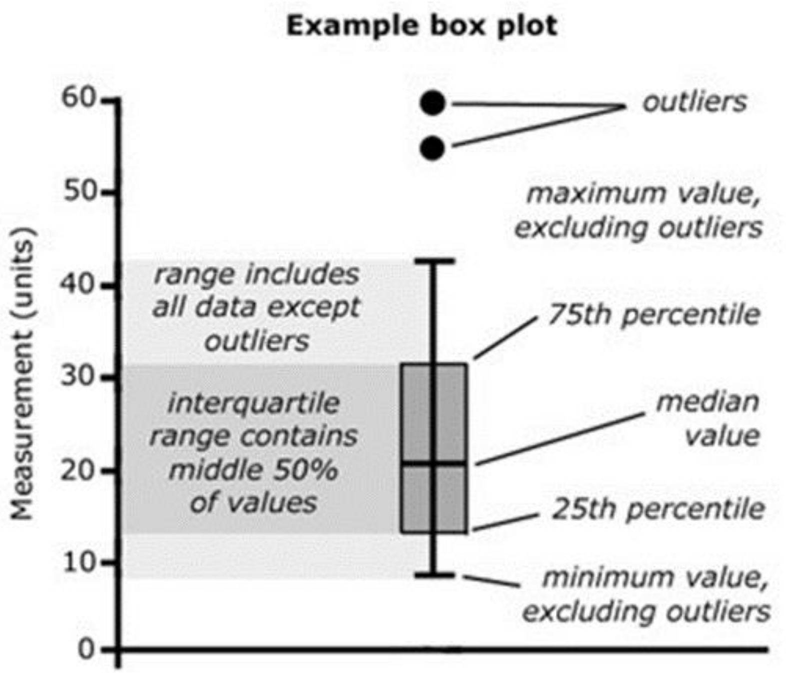
- **While outliers are attributed to a rare chance and may not necessarily be fully explainable, outliers in data can distort predictions and affect the accuracy, if you don't detect and handle them.**

- **The contentious decision to consider or discard an outlier needs to be taken at the time of building the model.**
  - Outliers can drastically bias/change the fit estimates and predictions. It is left to the best judgement of the analyst to decide whether treating outliers is necessary and how to go about it.



Data Cleansing - Outlier Detection

- **Univariate Approach:**
  - A univariate outlier is a data point that consists of an extreme value on one variable.
- **The Box Plot Rule**
  - For a given continuous variable, outliers are those observations that lie outside 1.5 * IQR, where IQR, the 'Inter Quartile Range' is the difference between 75th and 25th quartiles. This is also known as "The Box Plot Rule".
  - The box plot rule is the simplest statistical technique that has been applied to detect univariate outliers. Typically, in the Univariate Outlier Detection Approach look at the points outside the whiskers in a box plot.



Data Cleansing - Outlier Detection

- **Multivariate Approach:**
  - Declaring an observation as an outlier based on a just one (rather unimportant) feature could lead to unrealistic inferences. When you have to decide if an individual entity (represented by row or observation) is an extreme value or not, it better to collectively consider the features (X's) that matter.

- **A multivariate outlier is a combination of unusual scores on at least two variables.**

- **Several methods are used to identify outliers in multivariate datasets. Two of the widely used methods are:**
  - Distance-based Outlier Detection
  - Density-based Outlier Detection

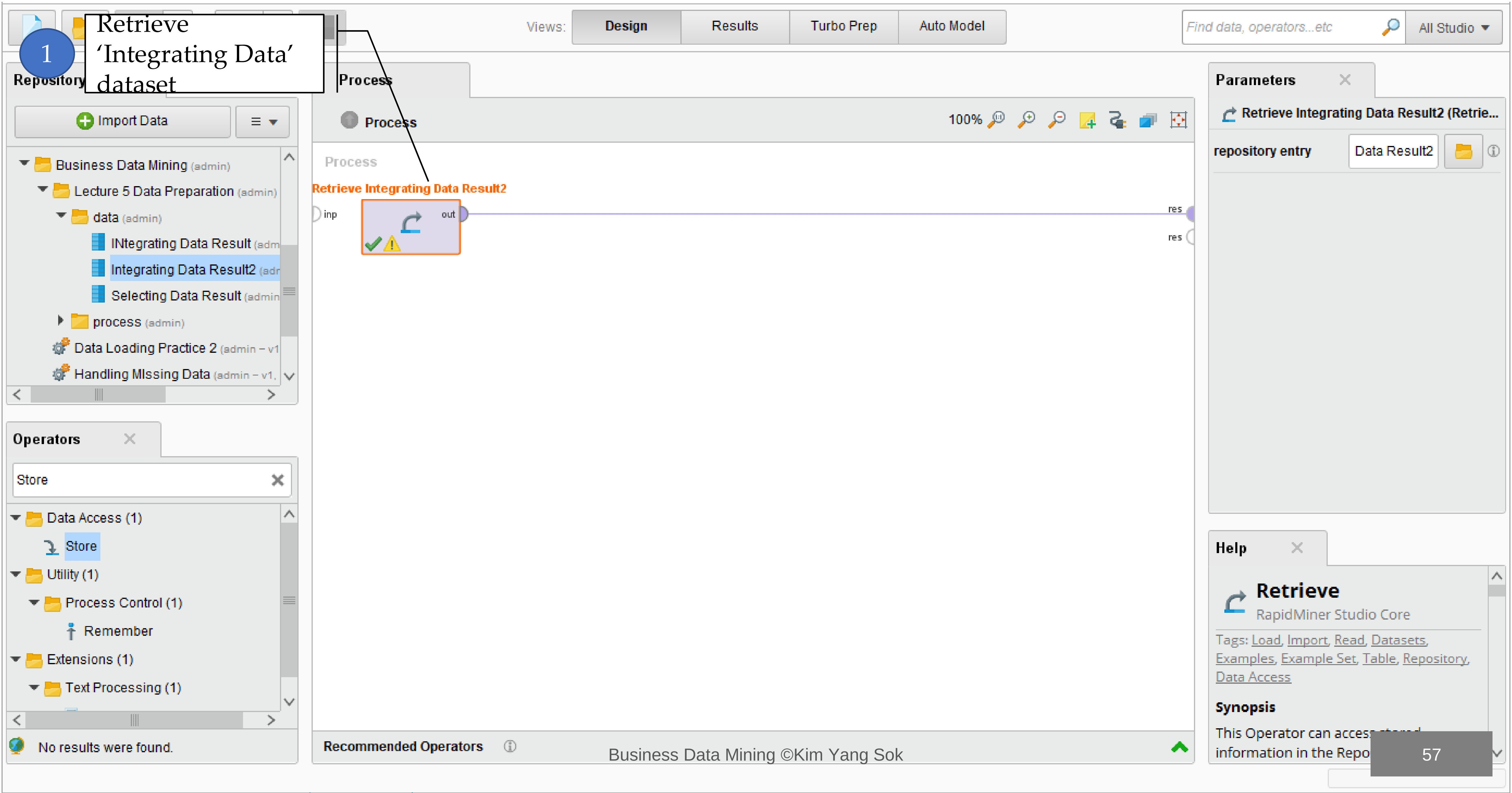
Exercise: Cleansing Data with Rapidminer

Task & Steps

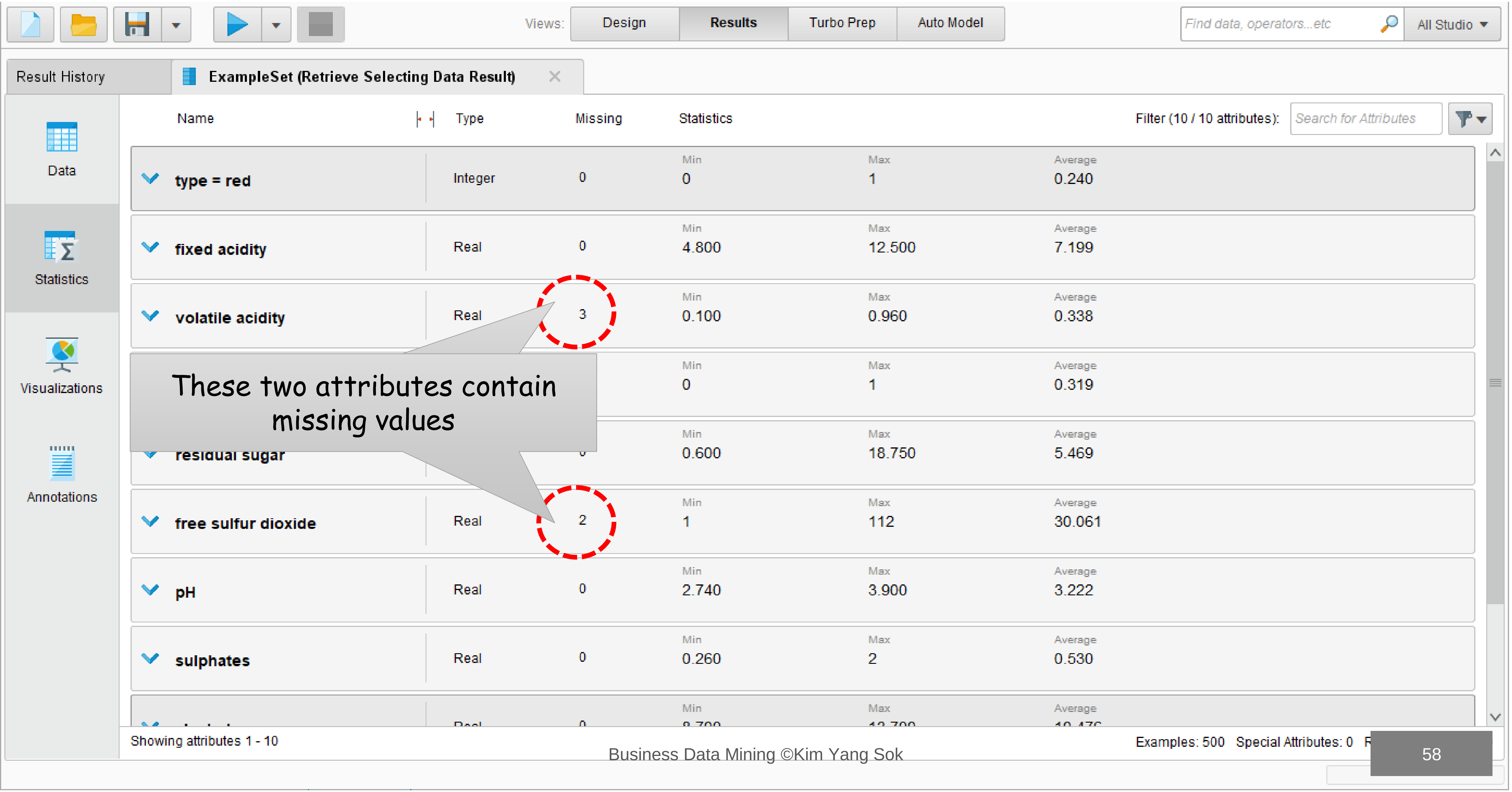
- **Task**
  - Improve data quality of dataset by applying data cleansing techniques

- **Steps**
  1. Import dataset from Local Repository (Integrating Data Result)
  2. Handling missing values
  3. Remove outliers using algorithm
  4. Change ‘alcohol’ attribute into categorical attribute by binning
  5. Normalize numerical attributes

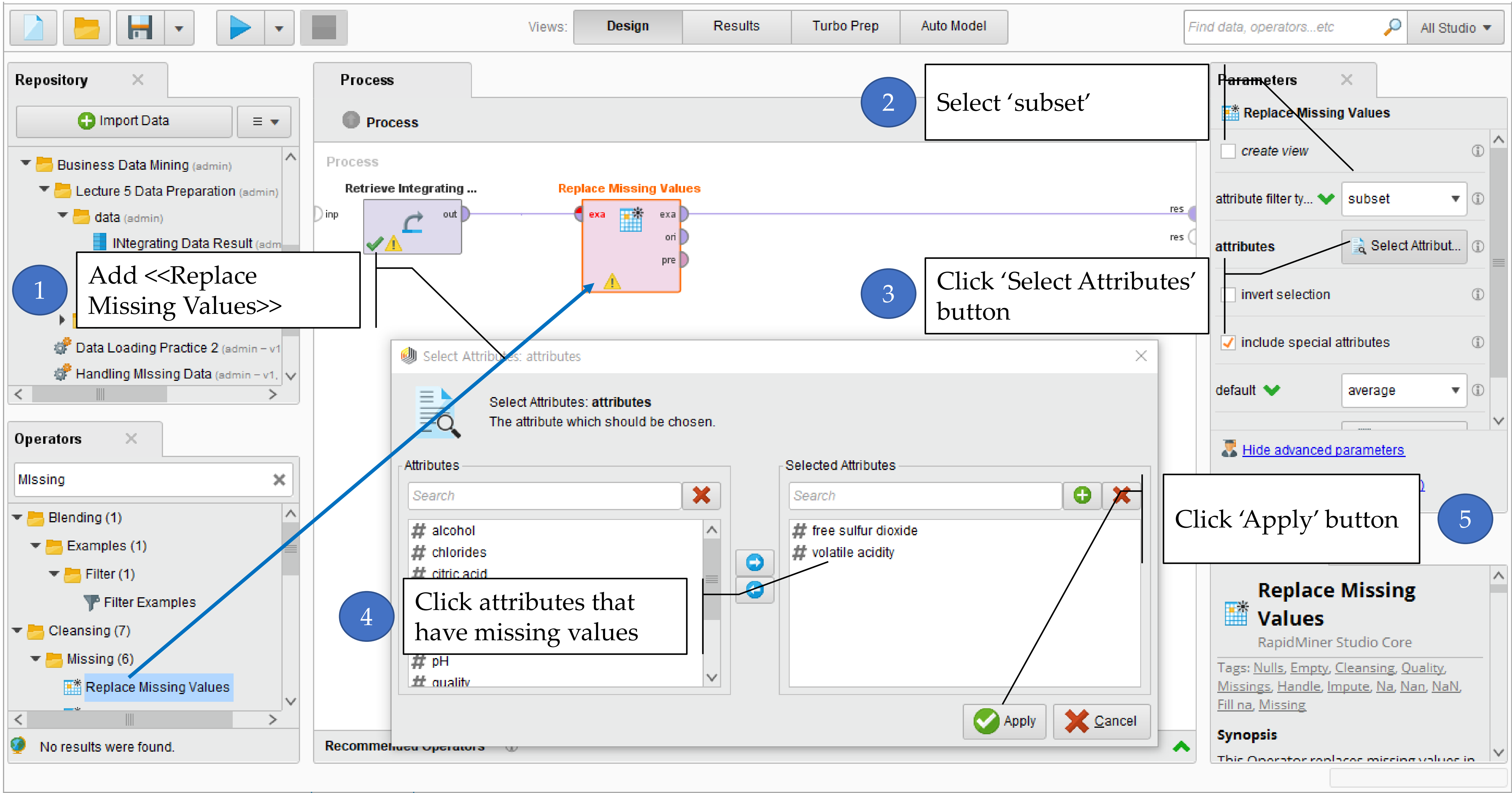
Import dataset from Local Repository (Integrating Data Result)



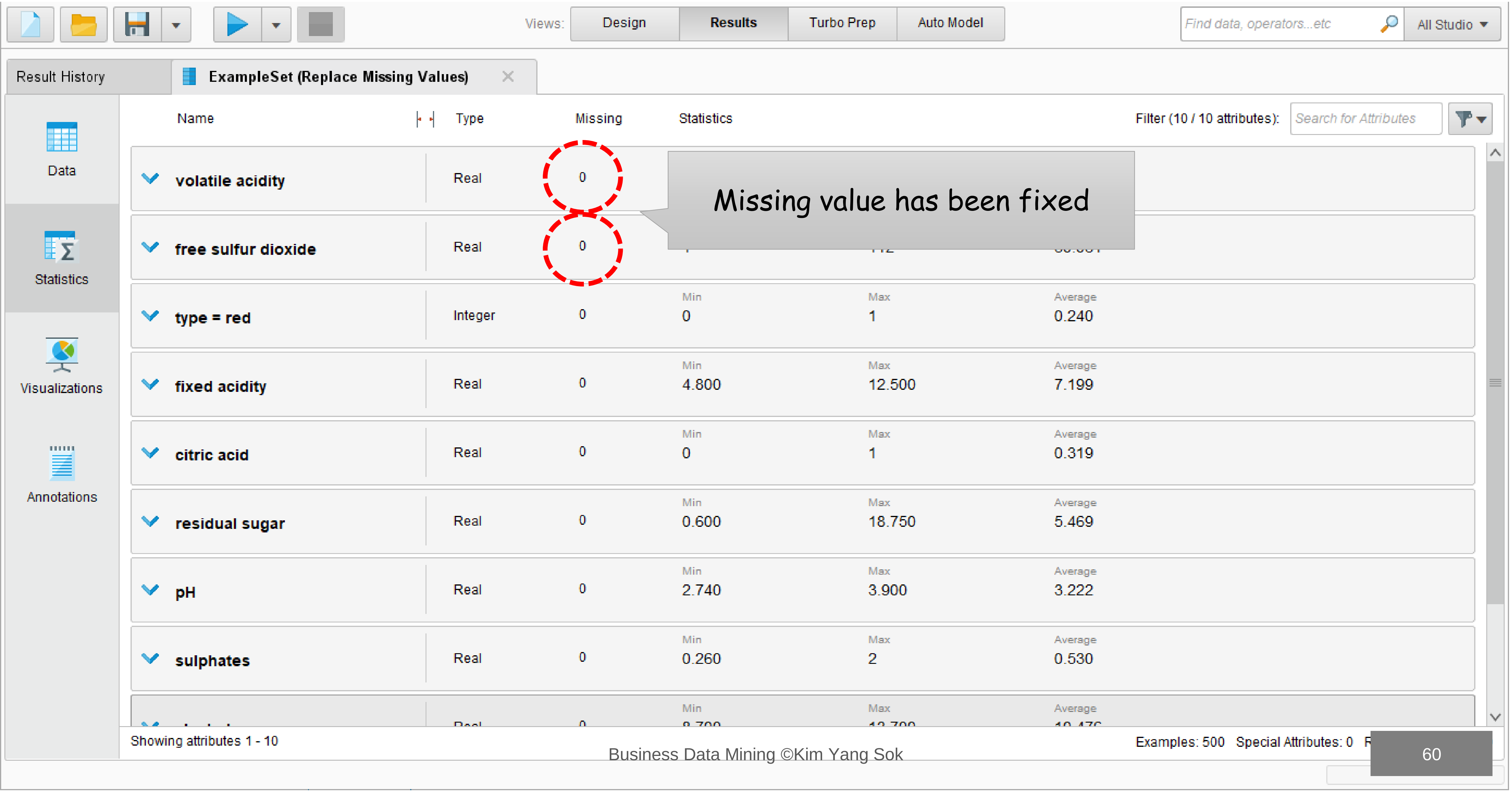
Handling missing values



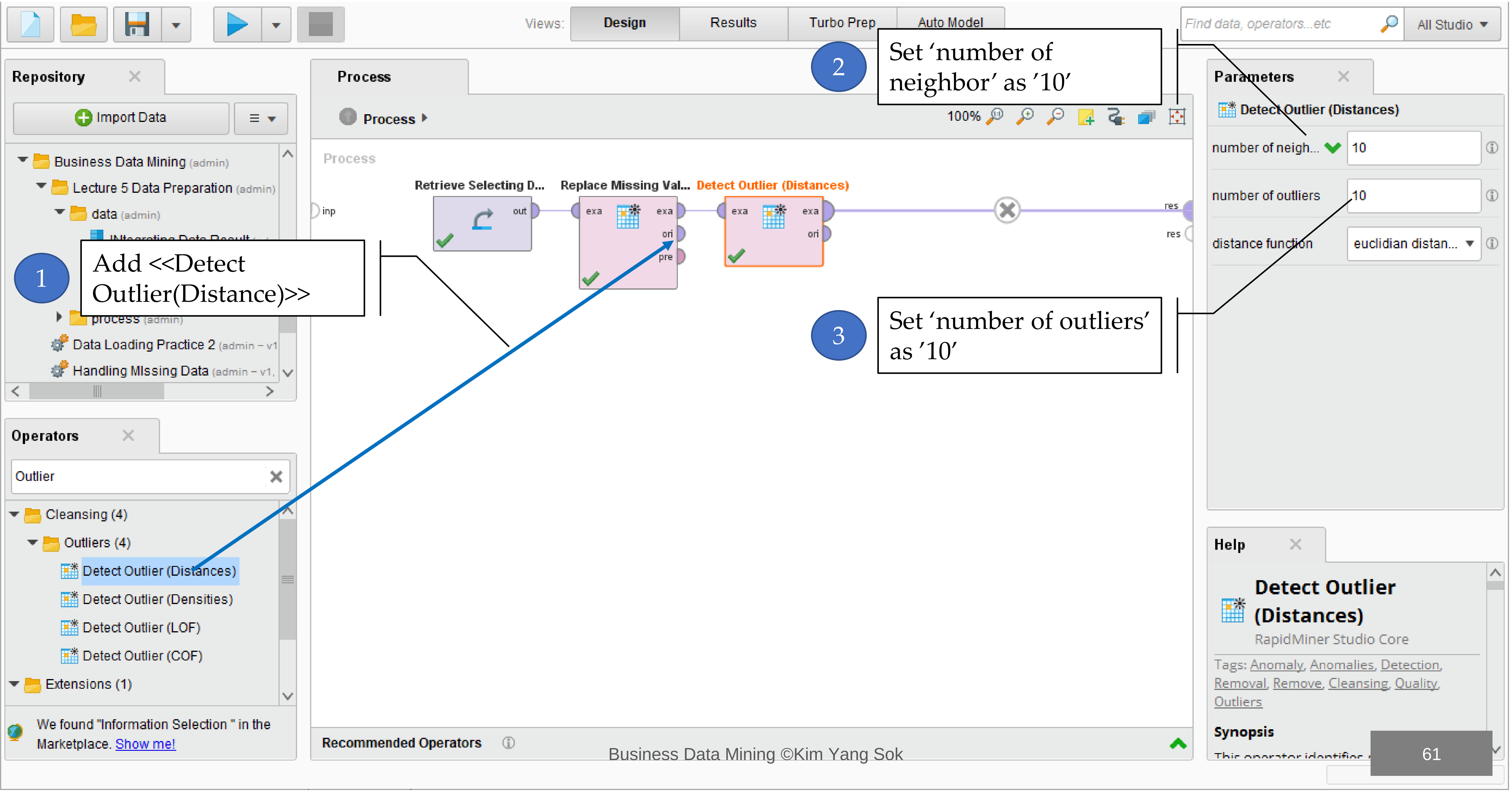
Handling missing values



Handling missing values



Remove outliers using algorithm

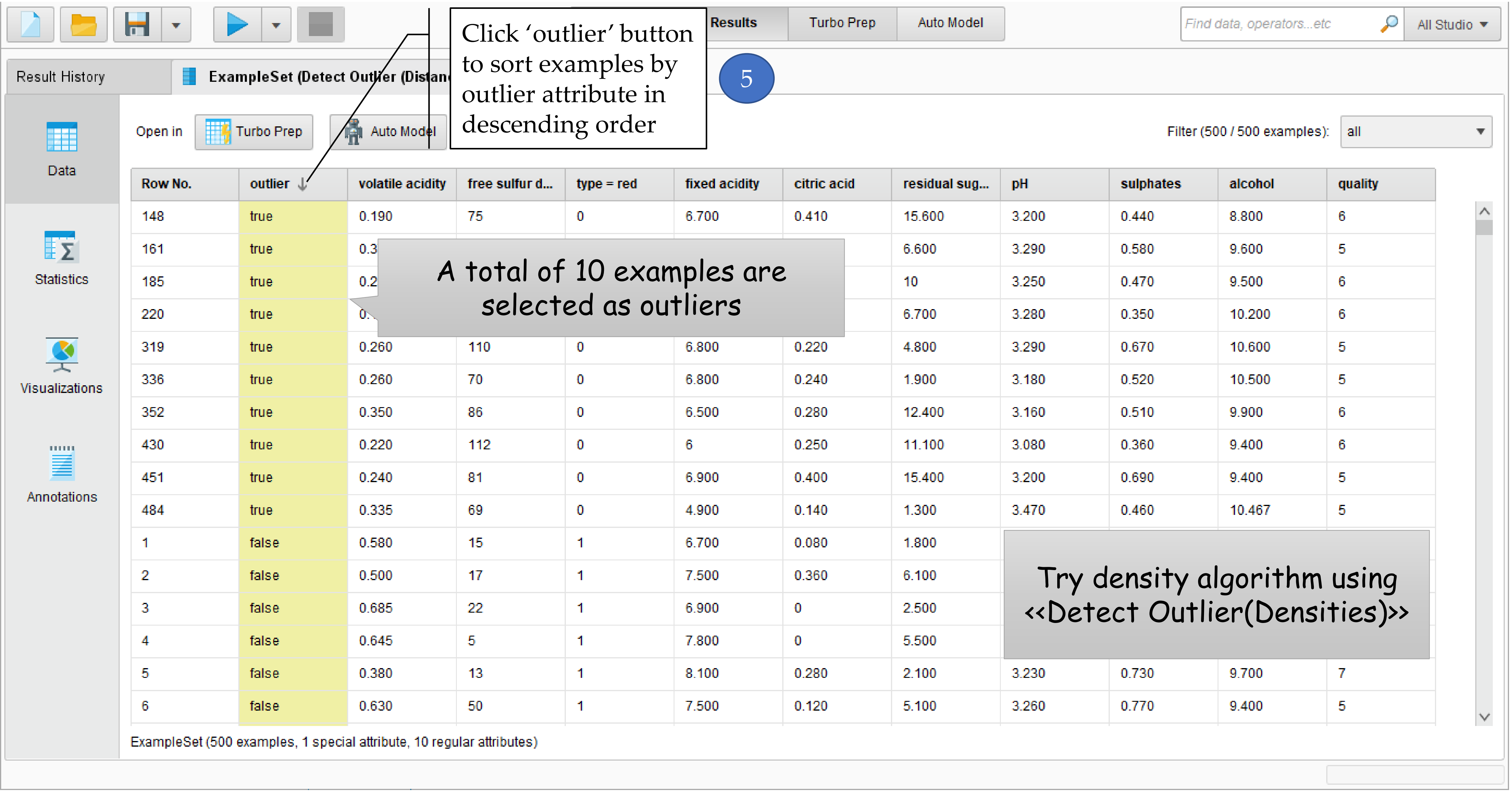


1 Add <<Detect Outlier(Distance)>>

2 Set 'number of neighbor' as '10'

3 Set 'number of outliers' as '10'

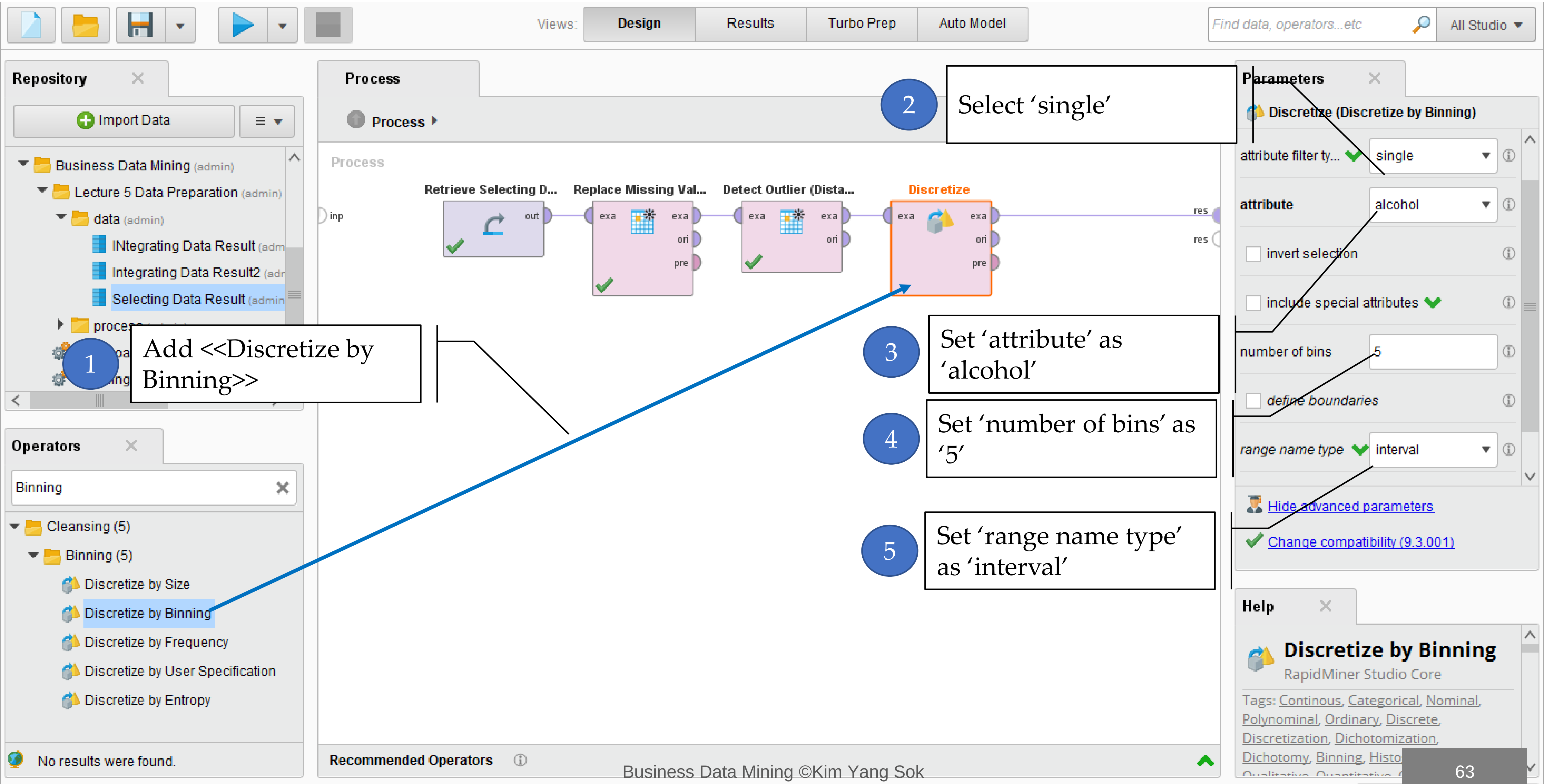
Remove outliers using algorithm



Row No.	outlier	volatile acidity	free sulfur d...	type = red	fixed acidity	citric acid	residual sug...	pH	sulphates	alcohol	quality
148	true	0.190	75	0	6.700	0.410	15.600	3.200	0.440	8.800	6
161	true	0.3	[illegible]	[illegible]	[illegible]	[illegible]	6.600	3.290	0.580	9.600	5
185	true	0.2	[illegible]	[illegible]	[illegible]	[illegible]	10	3.250	0.470	9.500	6
220	true	0.	[illegible]	[illegible]	[illegible]	[illegible]	6.700	3.280	0.350	10.200	6
319	true	0.260	110	0	6.800	0.220	4.800	3.290	0.670	10.600	5
336	true	0.260	70	0	6.800	0.240	1.900	3.180	0.520	10.500	5
352	true	0.350	86	0	6.500	0.280	12.400	3.160	0.510	9.900	6
430	true	0.220	112	0	6	0.250	11.100	3.080	0.360	9.400	6
451	true	0.240	81	0	6.900	0.400	15.400	3.200	0.690	9.400	5
484	true	0.335	69	0	4.900	0.140	1.300	3.470	0.460	10.467	5
1	false	0.580	15	1	6.700	0.080	1.800	[illegible]	[illegible]	[illegible]	[illegible]
2	false	0.500	17	1	7.500	0.360	6.100	[illegible]	[illegible]	[illegible]	[illegible]
3	false	0.685	22	1	6.900	0	2.500	[illegible]	[illegible]	[illegible]	[illegible]
4	false	0.645	5	1	7.800	0	5.500	[illegible]	[illegible]	[illegible]	[illegible]
5	false	0.380	13	1	8.100	0.280	2.100	3.230	0.730	9.700	7
6	false	0.630	50	1	7.500	0.120	5.100	3.260	0.770	9.400	5

ExampleSet (500 examples, 1 special attribute, 10 regular attributes)

Change 'alcohol' attribute into categorical attribute by binning



Change 'alcohol' attribute into categorical attribute by binning

Views: Design | Results | Turbo Prep | Auto Model

Result History | ExampleSet (Discretize)

Data | Statistics | Visualizations | Annotations

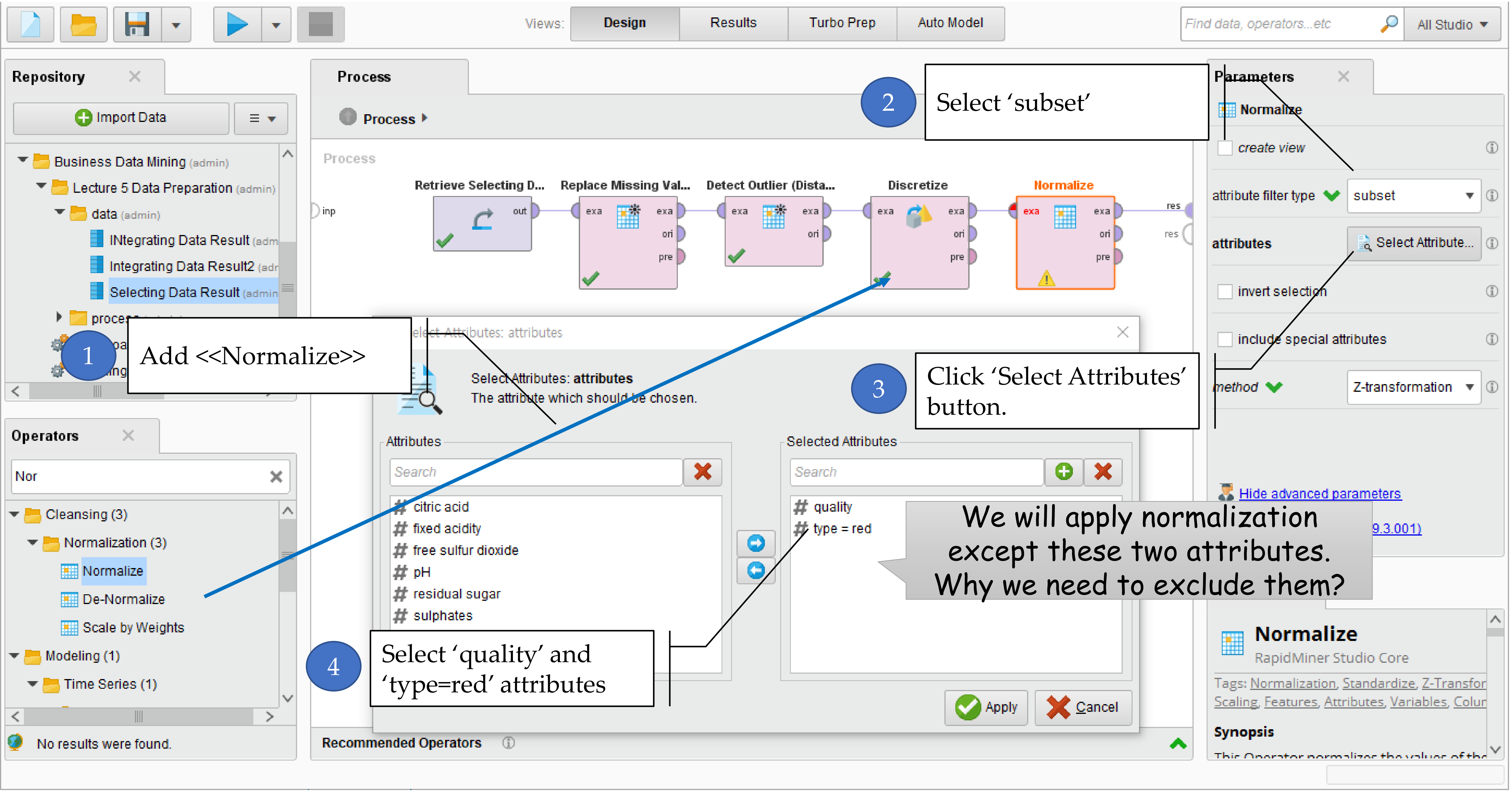
Open in Turbo Prep | Auto Model

Filter (500 / 500 examples): all

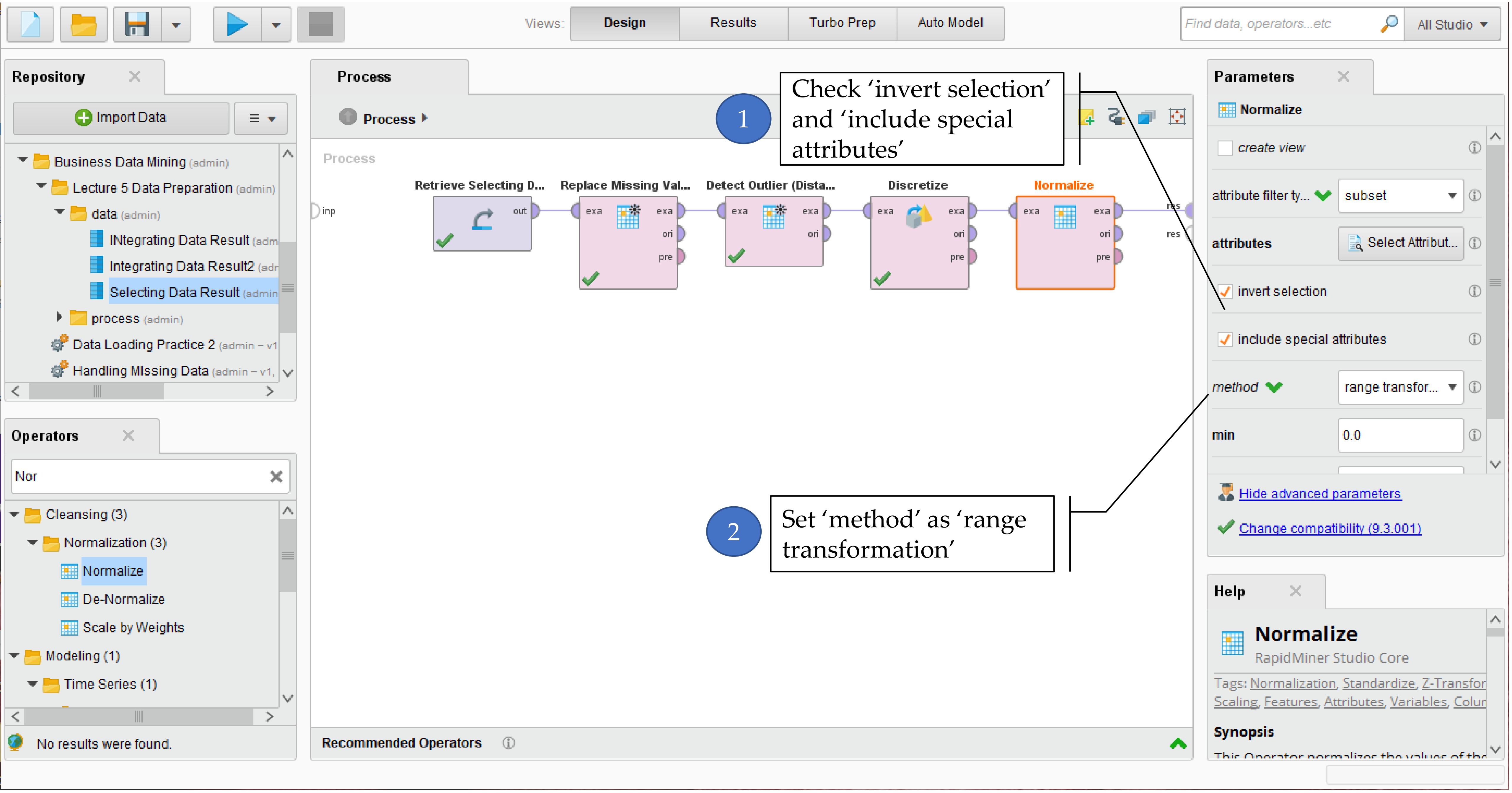
Row No.	outlier	alcohol	volatile acidity	free sulfur d...	type = red	fixed acidity	citric acid	residual sug...	pH	sulphates	quality
1	false	[-∞ - 9.7]	0.580	15	1	6.700	0.080	1.800	3.280	0.540	5
2	false	[9.7 - 10.7]	0.500	17	1	7.500	0.360	6.100	3.350	0.800	5
3	false	[9.7 - 10.7]	0.685	22	1	6.900	0	2.500	3.460	0.570	6
4	false	[-∞ - 9.7]	0.645	5	1	7.800	0	5.500	3.400	0.550	6
5	false	[-∞ - 9.7]	0.380	13	1	8.100	0.280	2.100	3.230	0.730	7
6	false	[-∞ - 9.7]	0.630	50	1	7.500	0.120	5.100	3.260	0.770	5
7	false	[-∞ - 9.7]	0.630	15	1	7.700	0.080	1.900	3.320	0.540	6
8	false	[9.7 - 10.7]	0.550	19	1	6.900	0.150	2.200	3.410	0.590	5
9	false	[-∞ - 9.7]	0.620	8	1	7	0.080	1.800	3.480	0.530	5
10	false	[-∞ - 9.7]	0.690	22	1	7	0.080	1.800	3.340	0.540	6
11	false	[-∞ - 9.7]	0.620	27	1	9	0.040	1.900	3.160	0.700	5
12	false	[9.7 - 10.7]	0.330	5	1	7.300	0.470	2.100	3.330	0.530	6
13	false	[-∞ - 9.7]	0.520	32	1	9.200	1	3.400	2.740	2	4
14	false	[-∞ - 9.7]	0.630	14	1	7.800	0.480	1.700	3.190	0.620	5
15	false	[9.7 - 10.7]	0.640	18	1	6.800	0.100	2.100	3.340	0.520	5
16	false	[-∞ - 9.7]	0.610	10	1	8.800	0.140	2.400	3.190	0.590	5

ExampleSet (500 examples, 1 special attribute, 10 regular attributes)

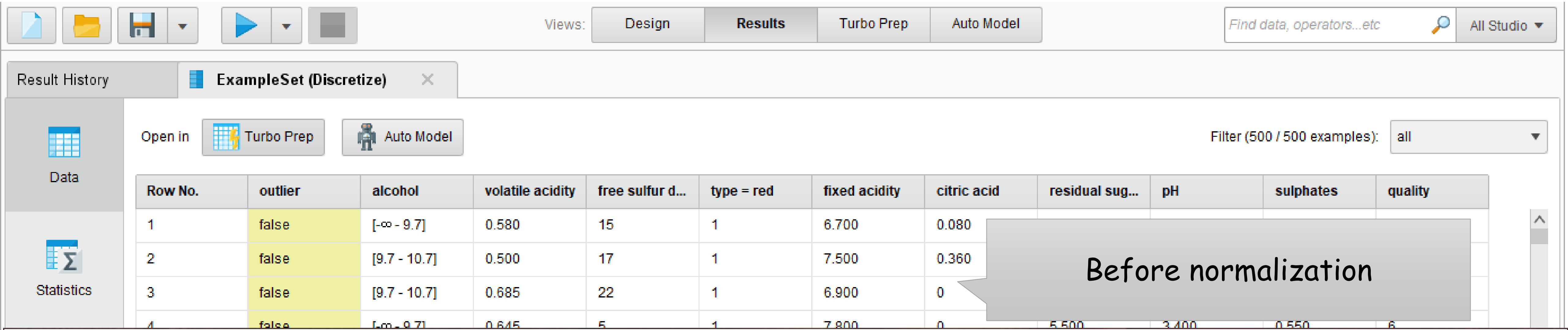
Normalize numerical attributes

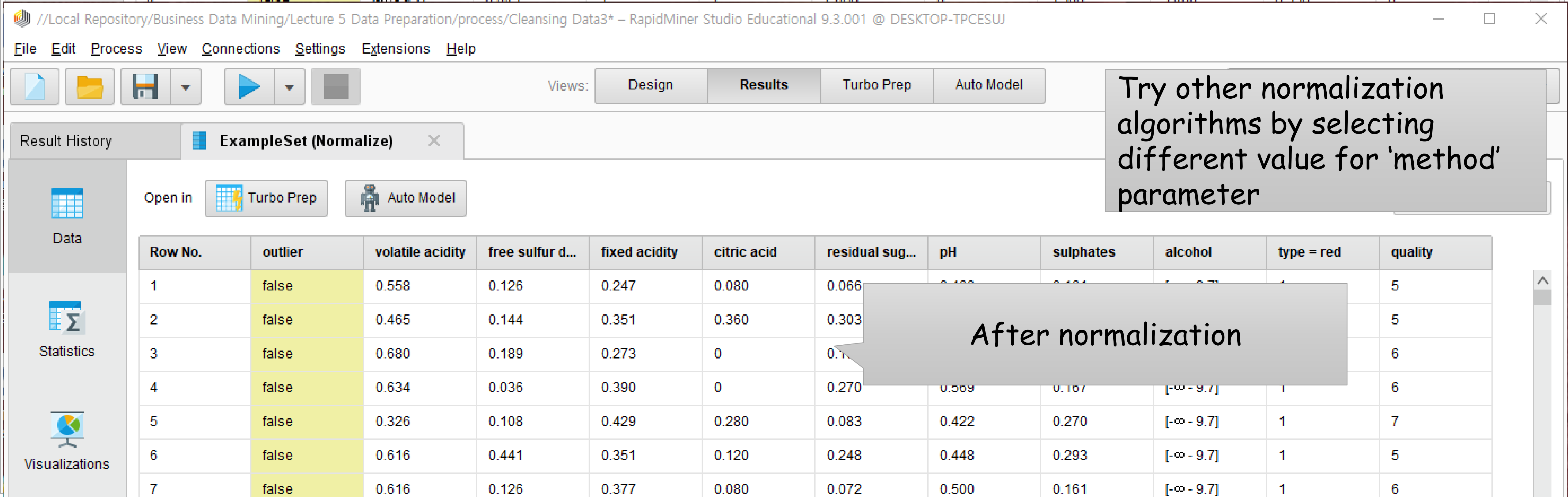


Normalize numerical attributes



Normalize numerical attributes





Conclusion

Conclusion

- Various techniques can be applied to data preparation phase such as data integration, data selection, and data cleansing.

- Correct data preparation is essential in data mining in order to get better performance and stable analysis results. If you have data problems in modeling phase, you can return this phase to resolve the problems.

- Now you have prepared dataset, you can go to next phase of CRISP-DM, the Modeling phase.

QUESTIONS?