Lab Assignment 1

Were Vincent Ouma

Lab Assignment 1

a)Data Quality Check:

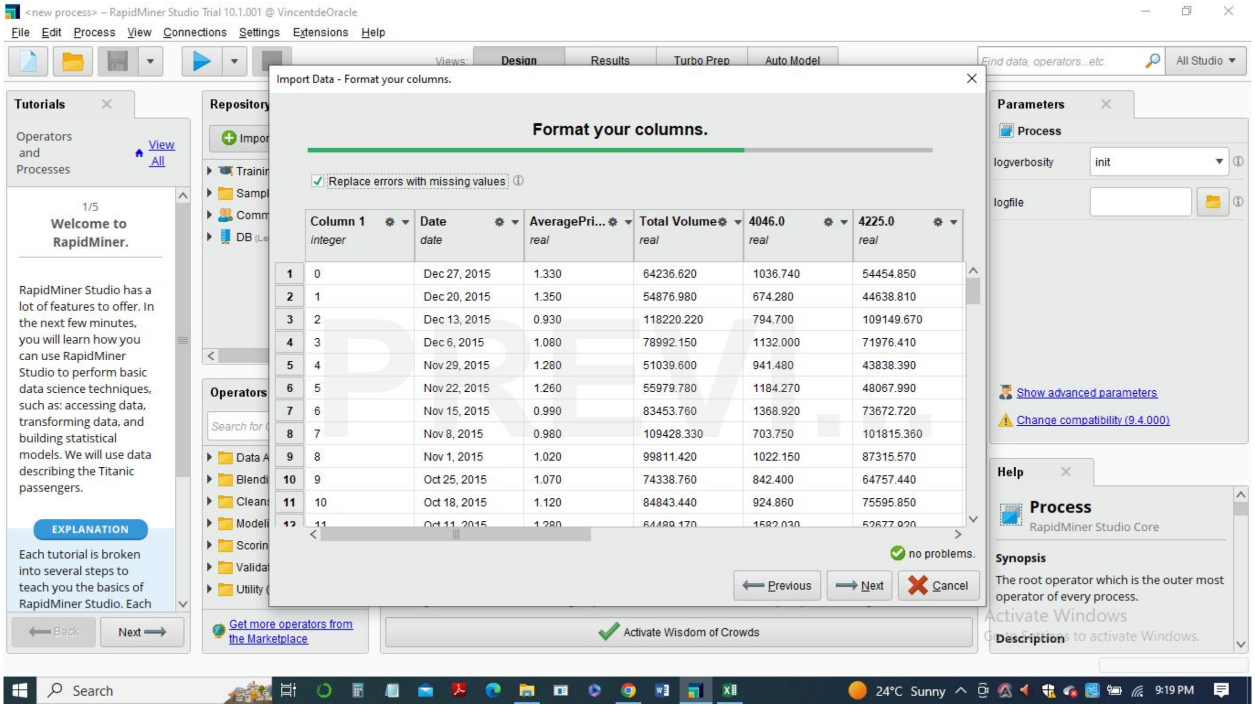
a.1) In RapidMiner Studio, the data quality is checked by using the “Data Quality” operator. This operator detect any bias in the dataset by calculating descriptive statistics of the dataset and visualizing the distribution of the data. It can also detect outliers and missing values. The “Data Quality” operator also provides information about the number of rows and columns in the dataset.

a.2) To handle the data quality issues such as duplications and missing values, the “Data Cleansing” operator is used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to remove all duplicate values or to replace missing values with the median of the column feature in that specific “Region” variable.

a.3) The number of rows and columns in the dataset is 18249 and 13, respectively.

b)Data Cleaning and Preprocessing:

b.1) In RapidMiner Studio, the first column in the dataset is removed by using the “Select Attributes” operator. This operator allows the user to select or deselect the attributes that should be included in the dataset. The ‘year’ variable can be treated as nominal by using the “Nominal to Numeric” operator. This operator allows the user to convert the nominal values to numeric values.



<new process> – RapidMiner Studio Trial 10.1.001 @ VincentdeOracle
File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model
Find data, operators...etc
All Studio
Tutorials
Repository
Import
Training
Samples
Community
DB
Operators
Search for Operators
Data Access
Blending
Cleansing
Modeling
Scoring
Validation
Utility
Get more operators from the Marketplace
Activate Wisdom of Crowds
Parameters
Process
logverbosity init
logfile
Show advanced parameters
Change compatibility (9.4.000)
Help
Search
24°C Sunny
9:19 PM

Operators and Processes View All

1/5

Welcome to RapidMiner.

RapidMiner Studio has a lot of features to offer. In the next few minutes, you will learn how you can use RapidMiner Studio to perform basic data science techniques, such as: accessing data, transforming data, and building statistical models. We will use data describing the Titanic passengers.

EXPLANATION

Each tutorial is broken into several steps to teach you the basics of RapidMiner Studio. Each

Back Next

Import Data - Format your columns.

Format your columns.

☑ Replace errors with missing values

	Column 1 integer	Date date	AveragePri... real	Total Volume real	4046.0 real	4225.0 real
1	0	Dec 27, 2015	1.330	64236.620	1036.740	54454.850
2	1	Dec 20, 2015	1.350	54876.980	674.280	44638.810
3	2	Dec 13, 2015	0.930	118220.220	794.700	109149.670
4	3	Dec 6, 2015	1.080	78992.150	1132.000	71976.410
5	4	Nov 29, 2015	1.280	51039.600	941.480	43838.390
6	5	Nov 22, 2015	1.260	55979.780	1184.270	48067.990
7	6	Nov 15, 2015	0.990	83453.760	1368.920	73672.720
8	7	Nov 8, 2015	0.980	109428.330	703.750	101815.360
9	8	Nov 1, 2015	1.020	99811.420	1022.150	87315.570
10	9	Oct 25, 2015	1.070	74338.760	842.400	64757.440
11	10	Oct 18, 2015	1.120	84843.440	924.860	75595.850
12	11	Oct 11, 2015	1.280	64489.170	1582.030	52677.920

✓ no problems.

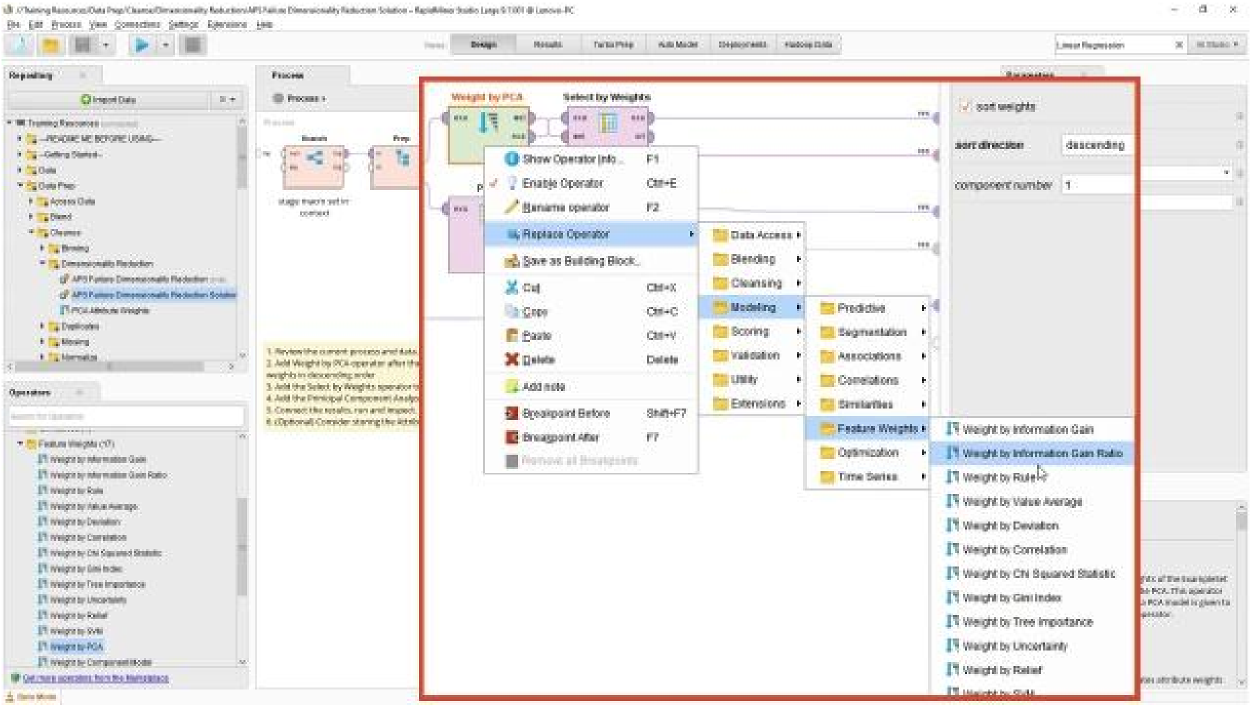
Previous Next Cancel

Process
RapidMiner Studio Core

Synopsis
The root operator which is the outer most operator of every process.

Description

b.2) To check for duplicate values and remove them, the "Data Cleansing" operator is used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to remove all duplicate values.

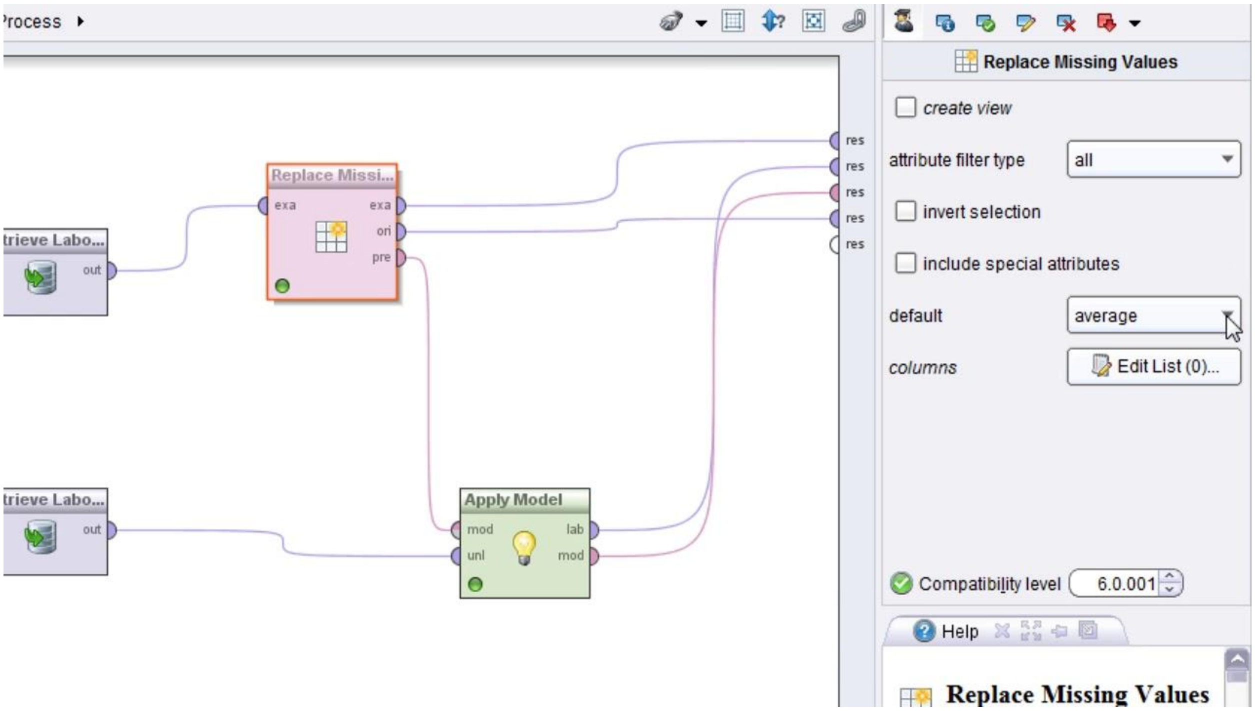


☑ sort weights

sort direction: descending

component number: 1

b.3) To check for missing values, the “Data Cleansing” operator is also used. This operator allows the user to select the columns to be cleaned and specify the rules for cleaning the data. For example, the user can specify the rule to replace missing values with the median of the column feature in that specific “Region” variable. If most column values in a data record are missing, the data record can be removed. The correlation between the variables is found by using the “Correlation Matrix” operator. This operator calculates the correlation coefficient between all the variables in the dataset and visualizes the correlation matrix. The result of the correlation matrix is shown in the screenshot below.



Replace Missing Values

□ create view

attribute filter type: all

□ invert selection

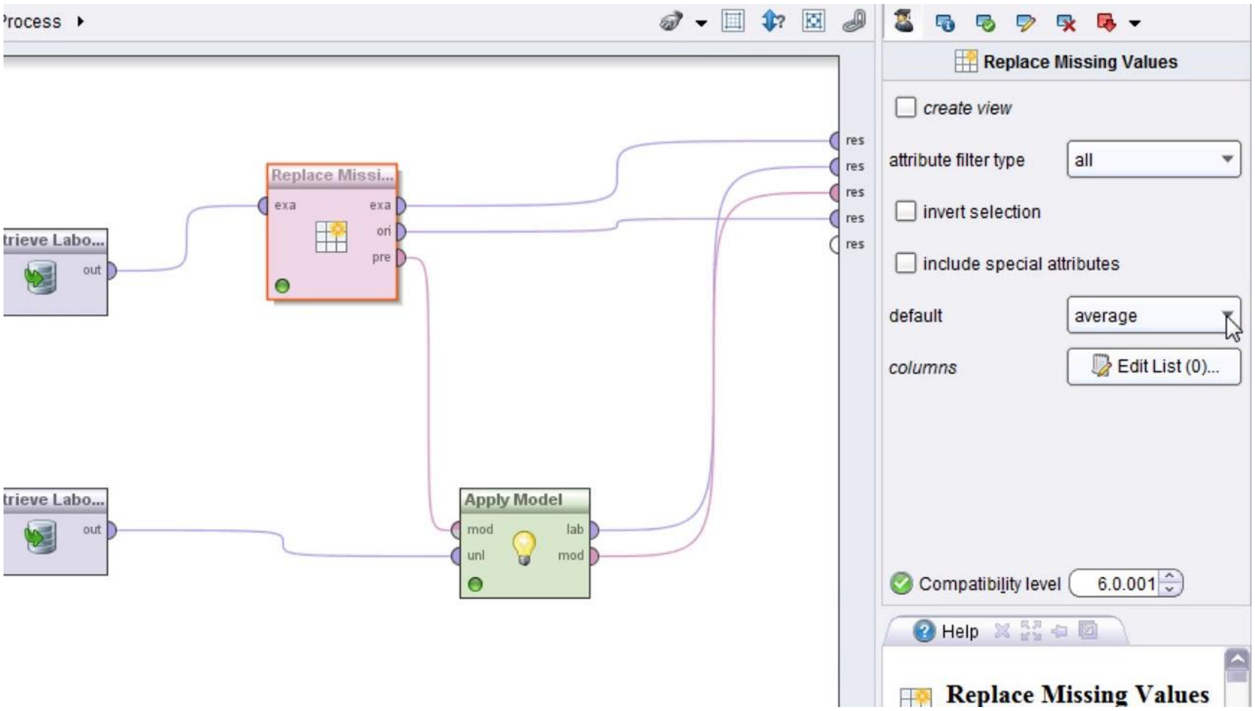
□ include special attributes

default: average

columns: Edit List (0)...

Compatibility level 6.0.001

Replace Missing Values


Replace Missing Values

□ create view

attribute filter type: all

□ invert selection

□ include special attributes

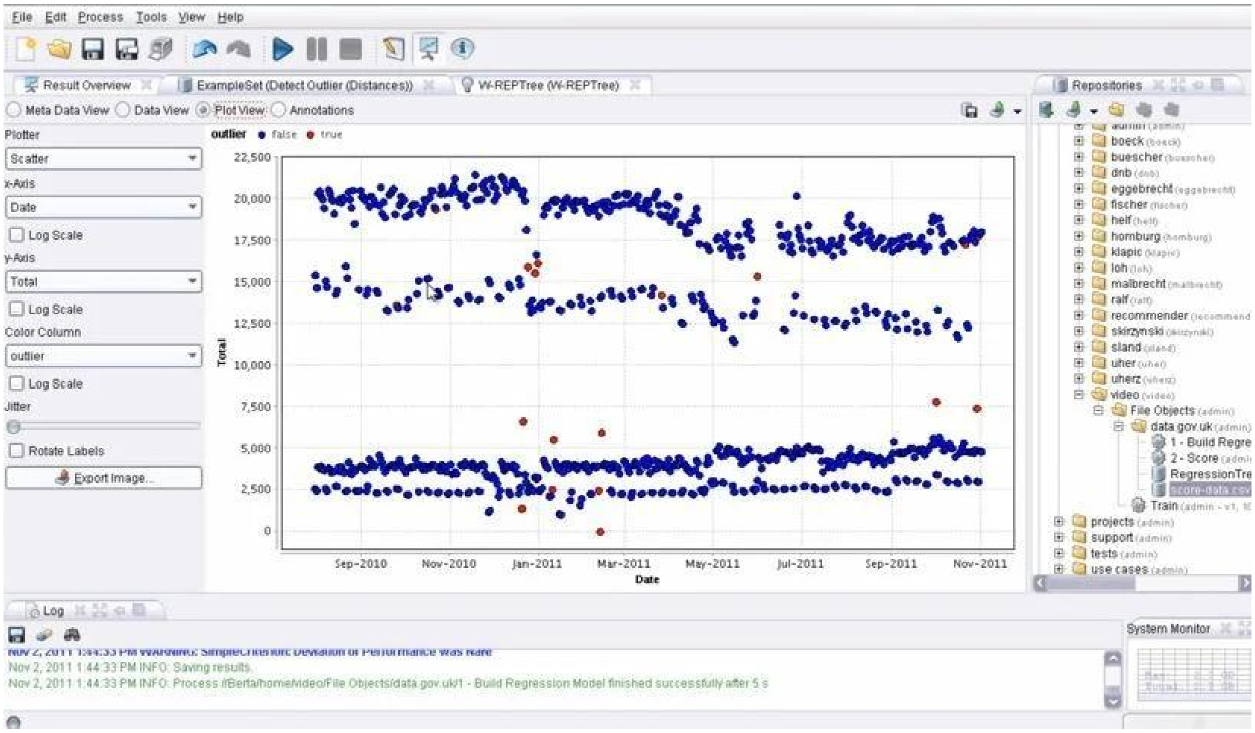
default: average

columns: Edit List (0)...

Compatibility level 6.0.001

Replace Missing Values

b.4) The correlation between the variables can affect the model accuracy. If two variables are highly correlated, they might provide redundant information, which can lead to overfitting of the model. On the other hand, if two variables are not correlated, they can provide complementary information which can help improve the accuracy of the model.



Plotter
Scatter
x-Axis
Date
☐ Log Scale
y-Axis
Total
☐ Log Scale
Color Column
outlier
☐ Log Scale
☐ Rotate Labels

```
Nov 2, 2011 1:44:33 PM WARNING: SimpleCriterion: Deviation of Performance was NaN
Nov 2, 2011 1:44:33 PM INFO: Saving results.
Nov 2, 2011 1:44:33 PM INFO: Process //Berta/home/video/File Objects/data.gov.uk/1 - Build Regression Model finished successfully after 5 s
```