

Business Intelligence Technologies: Practical 4: Correlation and Regression

Weighting: 4%

The US Police Department is concerned with the levels of violent crime which seems to be on the increase and would like some help in identifying the factors that are associated with this and how the information they have gathered on 100 cities (see dataset below) might be used to predict levels of violent crime in other areas.

Follow a CRISP methodology and use the Crime dataset (saved in Excel) to conduct correlation and linear regression analysis, documenting each step in the method and evaluating your results.

Assessment Criteria:

Correct and documented processes following the CRISP methodology, showing sound understanding of principles of correlation and regression analysis: 2 marks
Documentation and evaluation of results: 2 marks

Attribute information:

Note that you will need to rename the attributes (from X1 to X7) to something more meaningful prior to starting your analysis.

The data (X1, X2, X3, X4, X5, X6, X7) are for each city:

X1 = total overall reported crime rate per 1 million residents

X2 = reported violent crime rate per 100,000 residents

X3 = annual police funding in \$/resident

X4 = % of people 25 years+ with 4 yrs. of high school

X5 = % of 16 to 19 year-olds not in high school and not high school graduates.

X6 = % of 18 to 24 year-olds in college

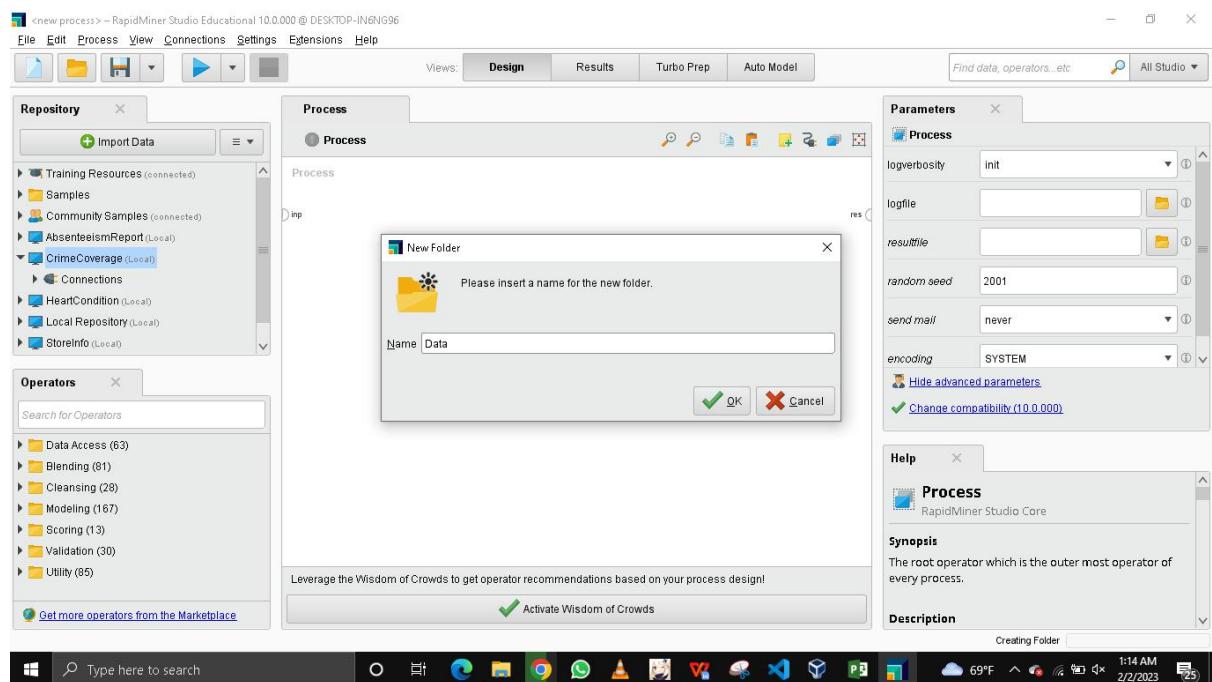
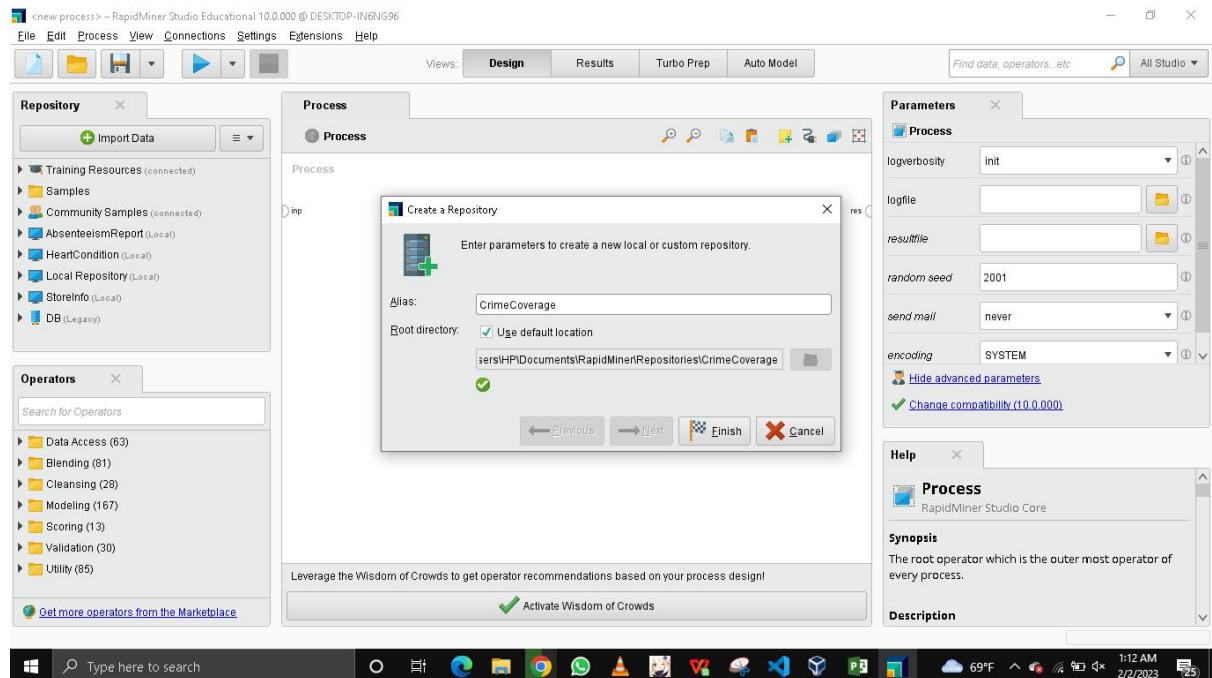
X7 = % of people 25 years+ with at least 4 years of college (25+ w/4yrs College)

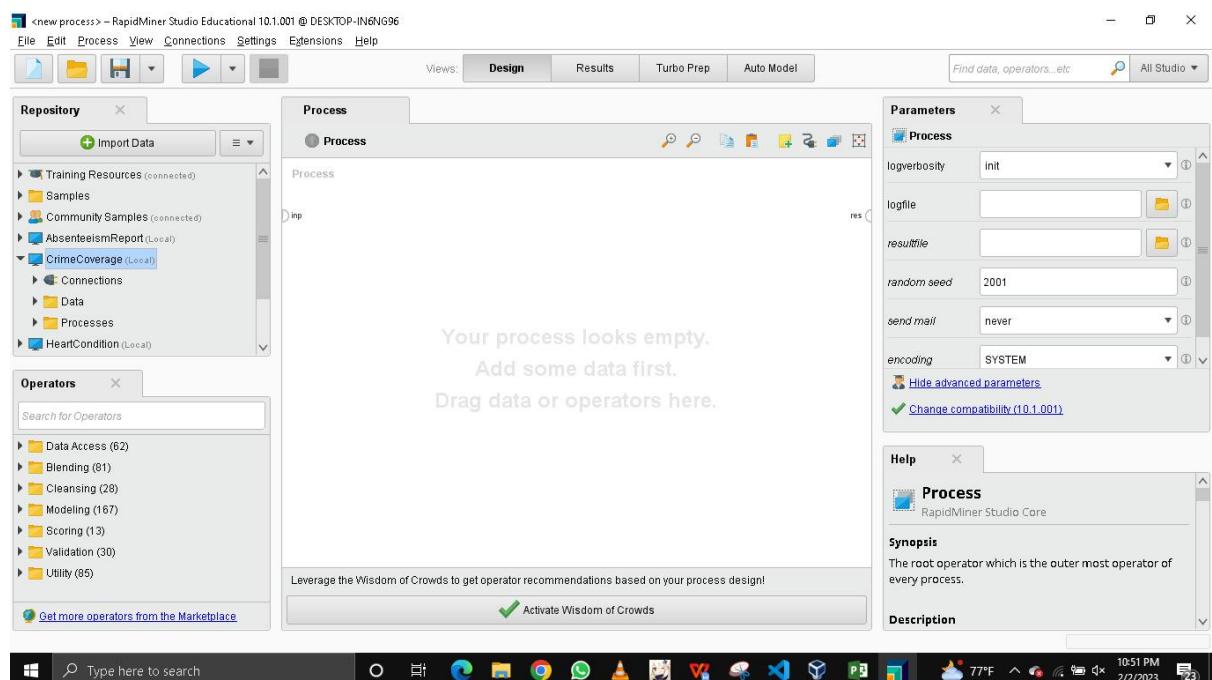
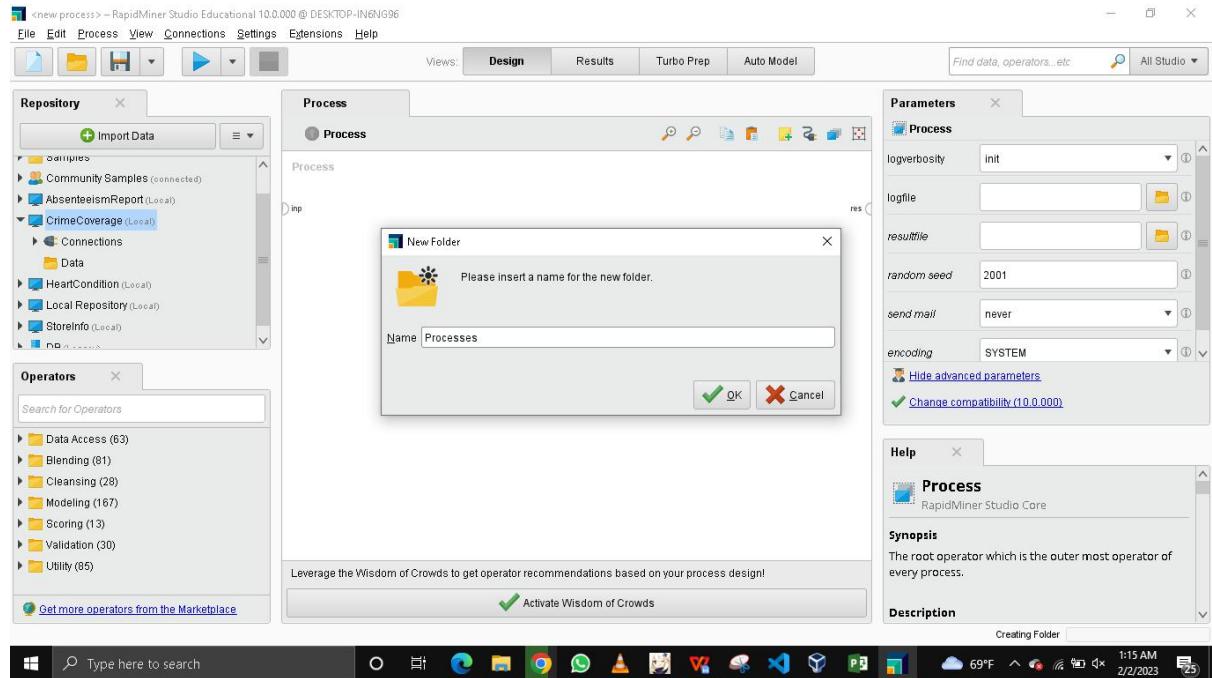
You should ensure that you:

- 1) Produce a correlation matrix, and provide analysis of the results.
- 2) Produce a linear regression model using a suitable attribute as your dependent variable, test your model (and for one example in the dataset, manually calculate the predicted value using the linear regression formula $y=mx+b$), and provide analysis of the results and any advice you might give to the US Police Department based on your findings.

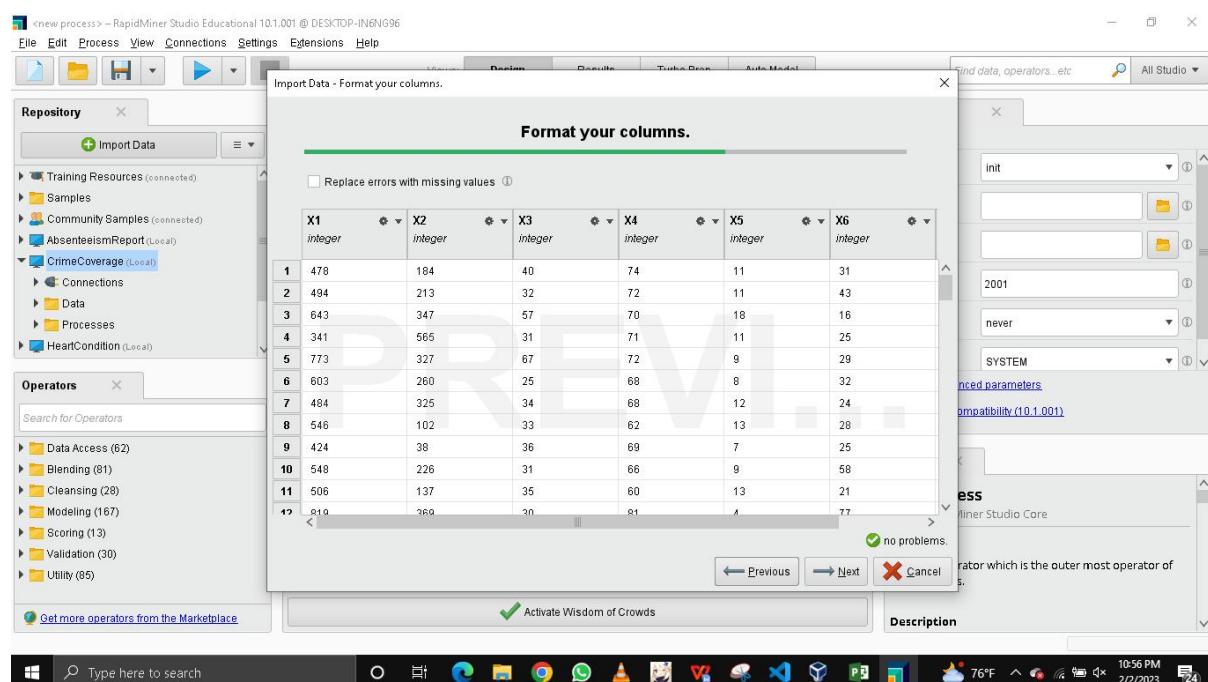
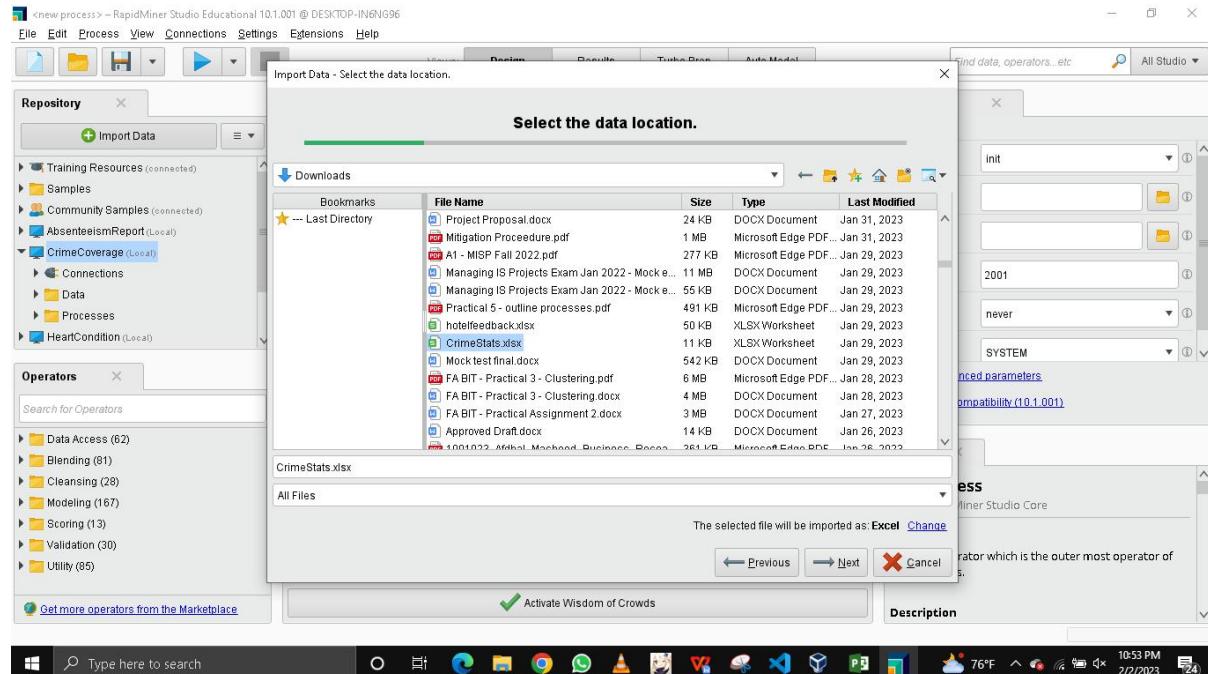
Practical Documentation

1. Created a Repository labeled “CrimeCoverage” and created two subfolders named “Data” and “Processes” respectively





2. Imported the CrimeStats table into the Data folder of the CrimeCoverage Repository as shown below with all except the table's column headings unchanged



<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Tutorial Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	X2 integer	X3 integer	X4 integer	X5 integer	X6 integer	X7 integer
1	184	40	74	11	31	20
2	213	32	72	11	43	18
3	347	57	70	18	16	16
4	565	31	71	11	25	19
5	327	67	72	9	29	24
6	260	25	68	8	32	15
7	325	34	68	12	24	14
8	102	33	62	13	28	11
9	38	36	69	7	25	12
10	226	31	66	9	58	15
11	137	35	60	13	21	9
12	260	30	81	4	77	26

no problems.

← Previous Next → Cancel

Activate Wisdom of Crowds

Description

Repository Operators

Import Data

Training Resources Samples Community Samples (connected) AbsenteeismReport (Local) CrimeCoverage (Local) Connections Data Processes HeartCondition (Local)

Operators

Search for Operators

Data Access (62) Blending (81) Cleansing (28) Modeling (167) Scoring (13) Validation (30) Utility (85)

Get more operators from the Marketplace

Windows Type here to search 10:57 PM 76°F 2/2/2023

<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Tutorial Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Total Crim... Name: Total Crime Rate (per 1M Residents) Type: integer	X2 integer	X3 integer	X4 integer	X5 integer	X6 integer
1				74	11	31
2	494	213	32	72	11	43
3	643	347	57	70	18	16
4	341	565	31	71	11	25
5	773	327	67	72	9	29
6	603	260	25	68	8	32
7	484	325	34	68	12	24
8	546	102	33	62	13	28
9	424	38	36	69	7	25
10	548	226	31	66	9	58
11	506	137	35	60	13	21
12	810	260	30	81	4	77

no problems.

← Previous Next → Cancel

Activate Wisdom of Crowds

Description

Repository Operators

Import Data

Training Resources Samples Community Samples (connected) AbsenteeismReport (Local) CrimeCoverage (Local) Connections Data Processes HeartCondition (Local)

Operators

Search for Operators

Data Access (62) Blending (81) Cleansing (28) Modeling (167) Scoring (13) Validation (30) Utility (85)

Get more operators from the Marketplace

Windows Type here to search 11:20 PM 76°F 2/2/2023

<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Tutorial Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Total Crim...	Violent Cri...	X3	X4	X5	X6
	integer	integer			integer	integer
1	478				11	31
2	494	213	32	72	11	43
3	643	347	57	70	18	16
4	341	565	31	71	11	25
5	773	327	67	72	9	29
6	603	260	25	68	8	32
7	484	325	34	68	12	24
8	546	102	33	62	13	28
9	424	38	36	69	7	25
10	548	226	31	66	9	58
11	506	137	35	60	13	21
12	810	360	30	81	4	77

no problems.

← Previous Next →

Activate Wisdom of Crowds

Description

init
2001
never
SYSTEM

ESS
RapidMiner Studio Core

operator which is the outer most operator of

Type here to search

11:21 PM 76°F 2/2/2023

<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Tutorial Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Total Crim...	Violent Cri...	X3	X4	X5	X6
	integer	integer			integer	integer
1	478	184				31
2	494	213	32	72	11	43
3	643	347	57	70	18	16
4	341	565	31	71	11	25
5	773	327	67	72	9	29
6	603	260	25	68	8	32
7	484	325	34	68	12	24
8	546	102	33	62	13	28
9	424	38	36	69	7	25
10	548	226	31	66	9	58
11	506	137	35	60	13	21
12	810	360	30	81	4	77

no problems.

← Previous Next →

Activate Wisdom of Crowds

Description

init
2001
never
SYSTEM

ESS
RapidMiner Studio Core

operator which is the outer most operator of

Type here to search

11:23 PM 76°F 2/2/2023

<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Train Data Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Total Crime	Violent Crime	Annual Police	% HS Graduates	X5	X6	X7
1	478	184	40	72	11	43	31
2	494	213	32	70	18	16	25
3	643	347	57	71	11	29	29
4	341	565	31	62	13	28	28
5	773	327	67	72	9	32	21
6	603	260	25	68	8	24	25
7	484	325	34	68	12	20	20
8	546	102	33	62	13	21	21
9	424	38	36	69	7	25	25
10	548	226	31	66	9	58	58
11	506	137	35	60	13	21	21
12	510	360	20	81	4	77	36

no problems.

Activate Wisdom of Crowds

Description

init
2001
never
SYSTEM

ESS
RapidMiner Studio Core

operator which is the outer most operator of

Type here to search

11:25 PM 2/2/2023

<new process> – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Design Results Train Data Auto Model

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Violent Crime	Annual Police	% HS Graduates	% Not in HS	X6	X7
1	184	40	74	11	43	18
2	213	32	72	18	16	16
3	347	57	70	11	25	19
4	565	31	71	11	29	24
5	327	67	72	9	32	15
6	260	25	68	8	24	14
7	325	34	68	12	28	11
8	102	33	62	13	77	36
9	38	36	69	7	25	12
10	226	31	66	9	58	15
11	137	35	60	13	21	9
12	360	20	81	4	77	36

no problems.

Activate Wisdom of Crowds

Description

init
2001
never
SYSTEM

ESS
RapidMiner Studio Core

operator which is the outer most operator of

Type here to search

11:27 PM 2/2/2023

The screenshot shows the RapidMiner Studio interface with the title bar "RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96". The menu bar includes File, Edit, Process, View, Connections, Settings, Extensions, Help. On the left, there are two collapsed panes: "Repository" and "Operators". The main workspace has tabs for "Dataset", "Results", "Tutorial Cases", and "Auto Models". A context menu is open over a dataset table, with the "Format your columns..." option selected. This brings up a modal dialog titled "Format your columns." with the sub-instruction "Import Data - Format your columns.". It contains a checkbox "Replace errors with missing values" and a table view of data. The table has columns: Violent Crime (integer), Annual Pol. Exp. (integer), % HS Grad. (integer), % Not in HS (integer), % College-Going (18-24) (integer), and X7 (integer). The last row is highlighted with a yellow border, and a tooltip shows "Name: % College-Going (18-24)" and "Type: integer". At the bottom of the dialog are buttons for "no problems.", "Previous", "Next", "Cancel", and "Activate Wisdom of Crowds". To the right of the dialog, a vertical sidebar shows search results for "init", "2001", "never", and "SYSTEM".

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Violent Crime integer	Annual Pol. Exp. integer	% HS Grad. integer	% Not in HS integer	% College-Going (18-24) integer	X7
1	184	40	74	11	43	18
2	213	32	72	11	43	18
3	347	57	70	18	16	16
4	565	31	71	11	25	19
5	327	67	72	9	29	24
6	260	25	68	8	32	15
7	325	34	68	12	24	14
8	102	33	62	13	28	11
9	38	36	69	7	25	12
10	226	31	66	9	58	15
11	137	35	60	13	21	9
12	360	30	81	4	77	36

no problems.

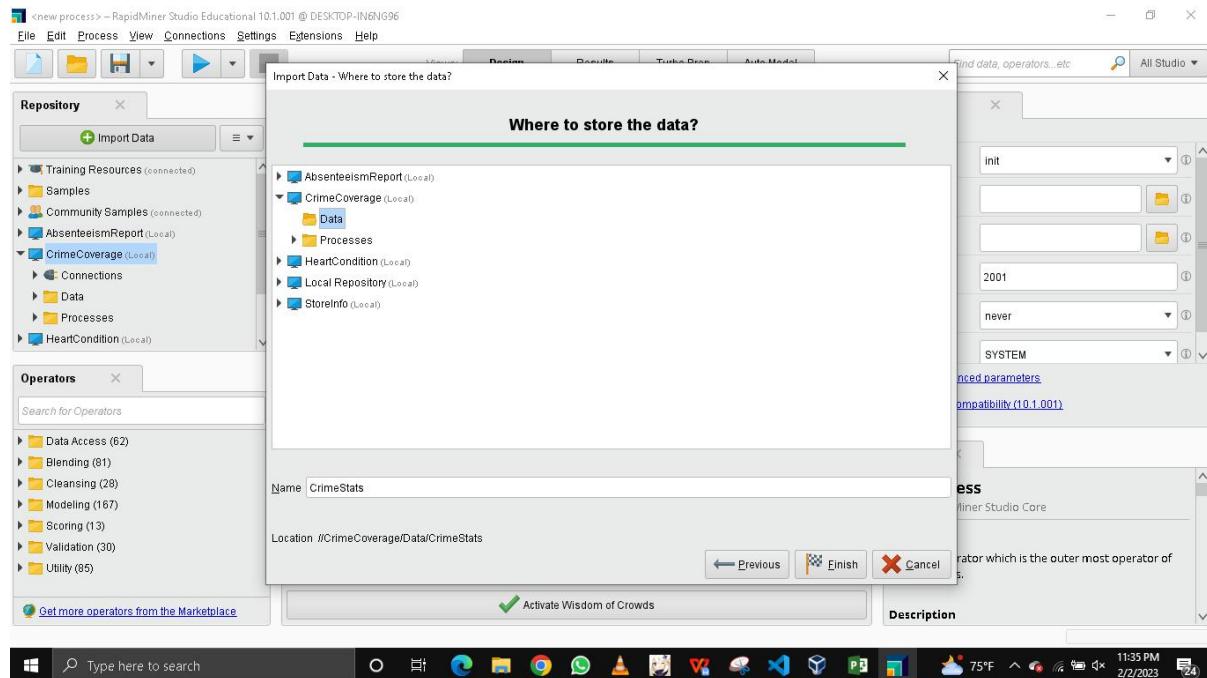
Previous Next Cancel

Activate Wisdom of Crowds

Description

The screenshot shows the RapidMiner Studio interface with the following details:

- File Bar:** File, Edit, Process, View, Connections, Settings, Extensions, Help.
- Toolbars:** Repository, Operators, Data, Results, Models.
- Repository Panel:** Shows connected resources like Training Resources, Samples, Community Samples, AbsenteeismReport, CrimeCoverage (Local), and HeartCondition (Local).
- Operators Panel:** Shows categories like Data Access (62), Blending (81), Cleansing (28), Modeling (167), Scoring (13), Validation (30), and Utility (85).
- Central Area:** A dialog titled "Import Data - Format your columns." with the sub-section "Format your columns." It contains a table with 12 columns and 13 rows of data. The first row is a header. The last row has a red asterisk indicating an error. A checkbox "Replace errors with missing values" is present. To the right of the table, there are two dropdowns: "Name: % College Graduates (25+)" and "Type: integer".
- Process Builder Panel:** On the right, it shows the process flow with operators like "init", "2001", "never", and "SYSTEM". It also displays compatibility information (10.1.001) and a note about the outermost operator.
- Bottom Bar:** Shows system icons (Windows, search, taskbar) and the status bar with "11:32 PM" and "75°F".



Below are the results of the completed import with the ranges of the values within the CrimeStats table shown for further detail

Row No.	Total Crime...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS /...	% College-G...	% College Gr...
1	478	184	40	74	11	31	20
2	494	213	32	72	11	43	18
3	643	347	57	70	18	16	16
4	341	565	31	71	11	25	19
5	773	327	67	72	9	29	24
6	603	260	25	68	8	32	15
7	484	325	34	68	12	24	14
8	546	102	33	62	13	28	11
9	424	38	36	69	7	25	12
10	548	226	31	66	9	58	15
11	506	137	35	60	13	21	9
12	819	369	30	81	4	77	36
13	541	109	44	66	9	37	12
14	491	809	32	67	11	37	16
15	514	29	30	65	12	35	11

ExampleSet(100 examples, 0 special attributes, 7 regular attributes)

Results View:

Row No.	Total Crime ...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS ...	% College-G...	% College Gr...
1	478	184	40	74	11	31	20
2	494	213	32	72	11	43	18
3	643	347	57	70	18	16	16
4	341	565	31	71	11	25	19
5	773	327	67	72	9	29	24
6	603	260	25	68	8	32	15
7	484	325	34	68	12	24	14
8	546	102	33	62	13	28	11
9	424	38	36	69	7	25	12
10	548	226	31	66	9	58	15
11	506	137	35	60	13	21	9
12	819	369	30	81	4	77	36
13	541	109	44	66	9	37	12
14	491	809	32	67	11	37	16
15	514	29	30	65	12	35	11

ExampleSet (100 examples, 0 special attributes, 7 regular attributes)

Repository View:

- Training Resources (connected)
- Samples
- Community Samples (connected)
- AbsenteeismReport (Local)
- CrimeCoverage (Local)
 - Connections
 - Data
 - CrimeStats (2/2/2023, 11:35 PM - 8 kB)
- Processes
- HeartCondition (Local)
- Local Repository (Local)

Press 'F3' for focus.

3. Retrieved the CrimeStats table into the Process panel and checked for any missing values; saving the process as “01-import CrimeStats table” in the Processes subfolder in the CrimeCoverage Repository

Process View:

Process

Retrieve CrimeStats

Operators View:

Search for Operators

- Data Access (62)
- Blending (61)
- Cleansing (28)
- Modeling (167)
- Scoring (13)
- Validation (30)
- Utility (65)

Parameters View:

repository entry: meCoverage/Data/CrimeStats

Help View:

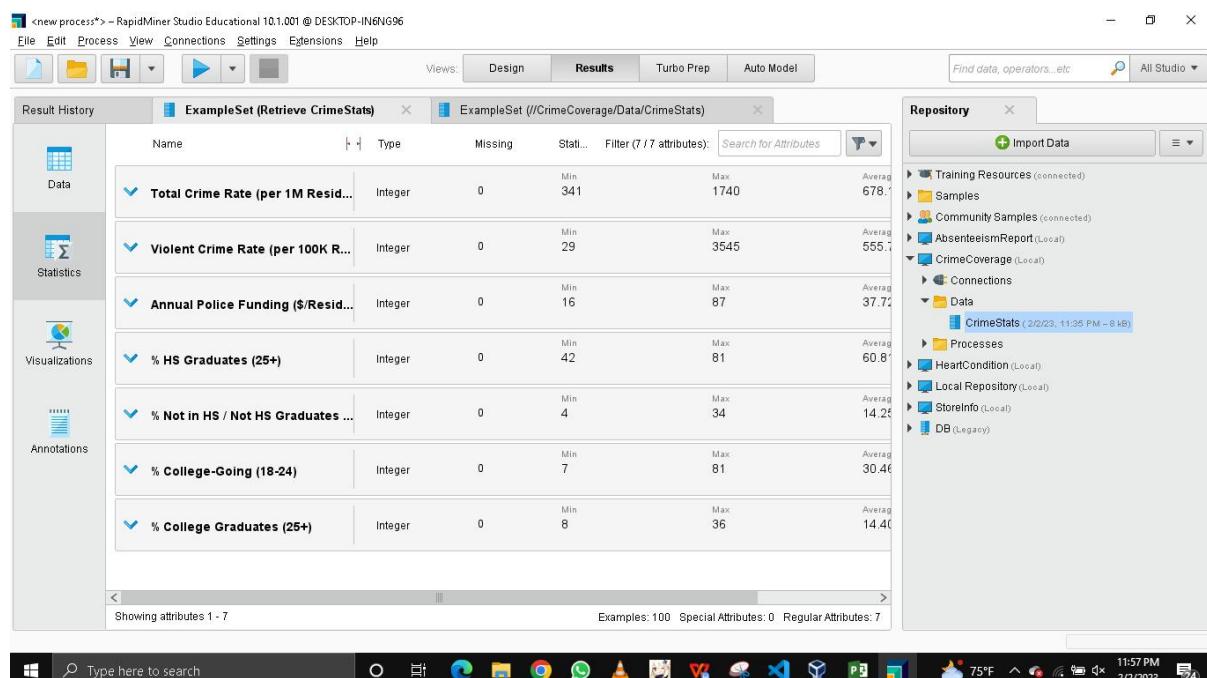
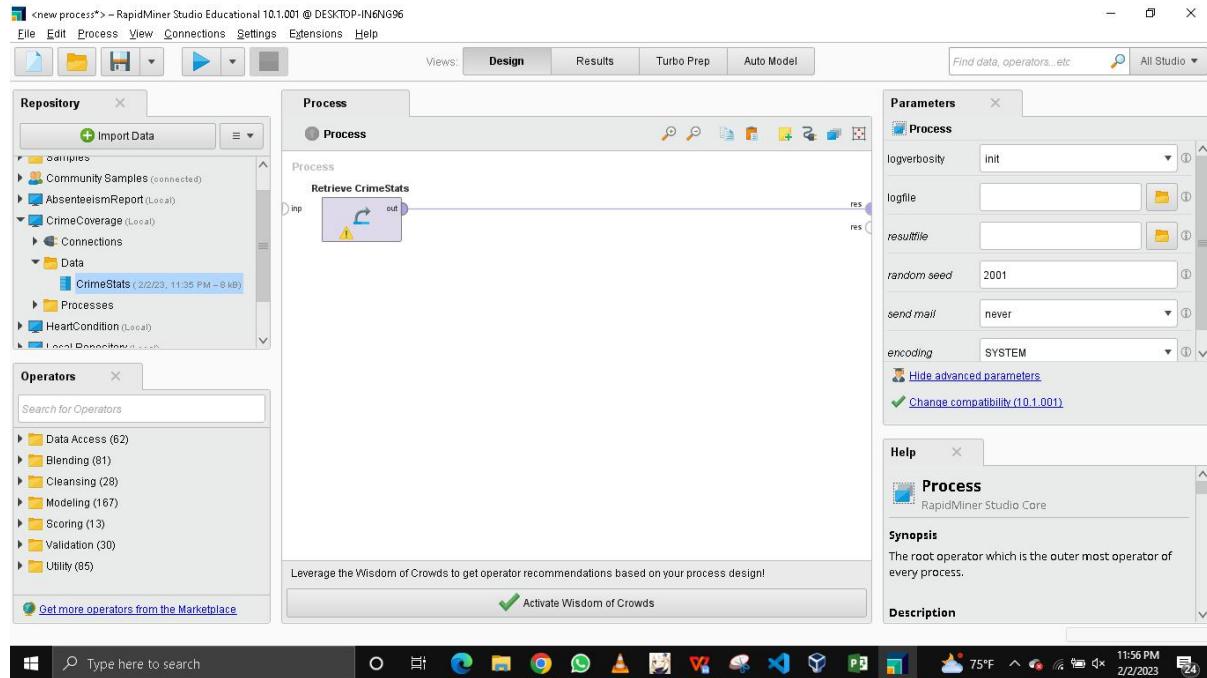
Retrieve

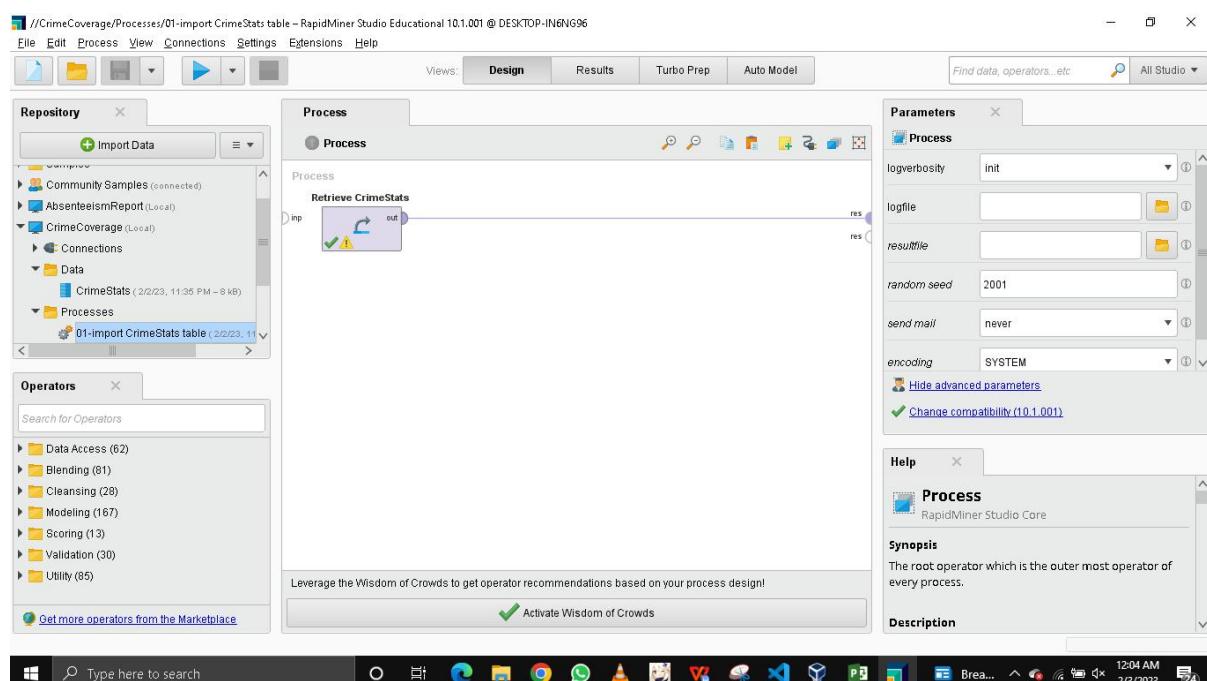
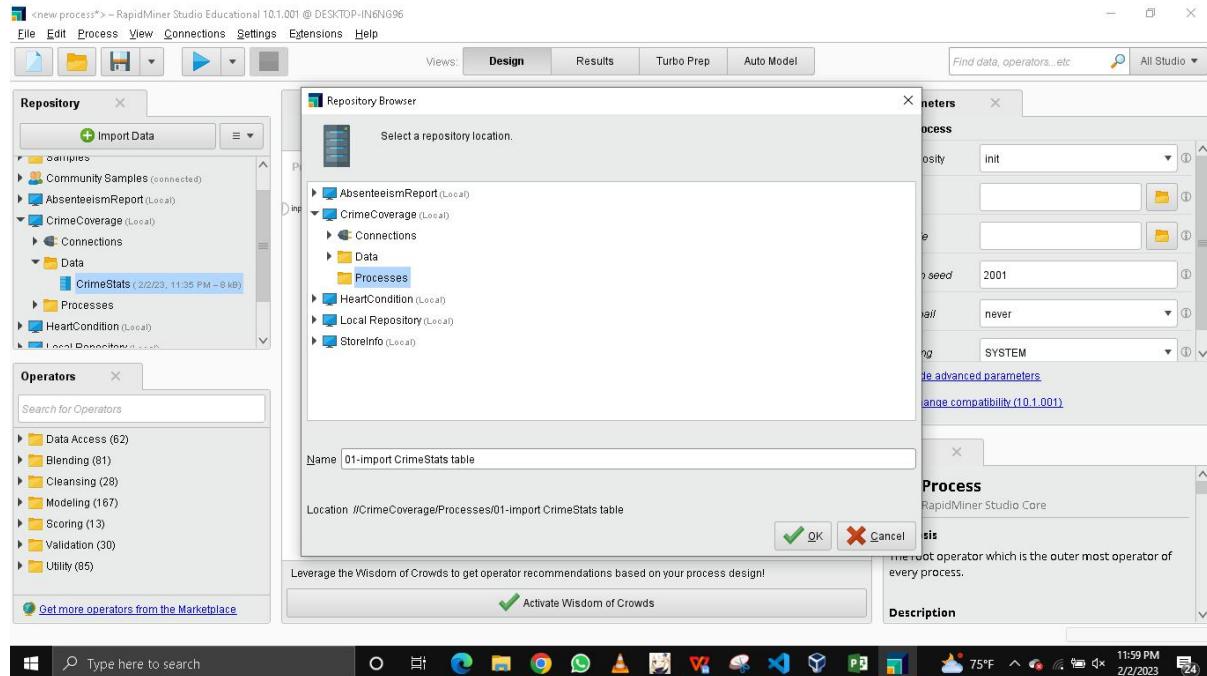
RapidMiner Studio Core

Tags: Load, Import, Read, Datasets, Examples, Example Set, Table, Repository, Data Access

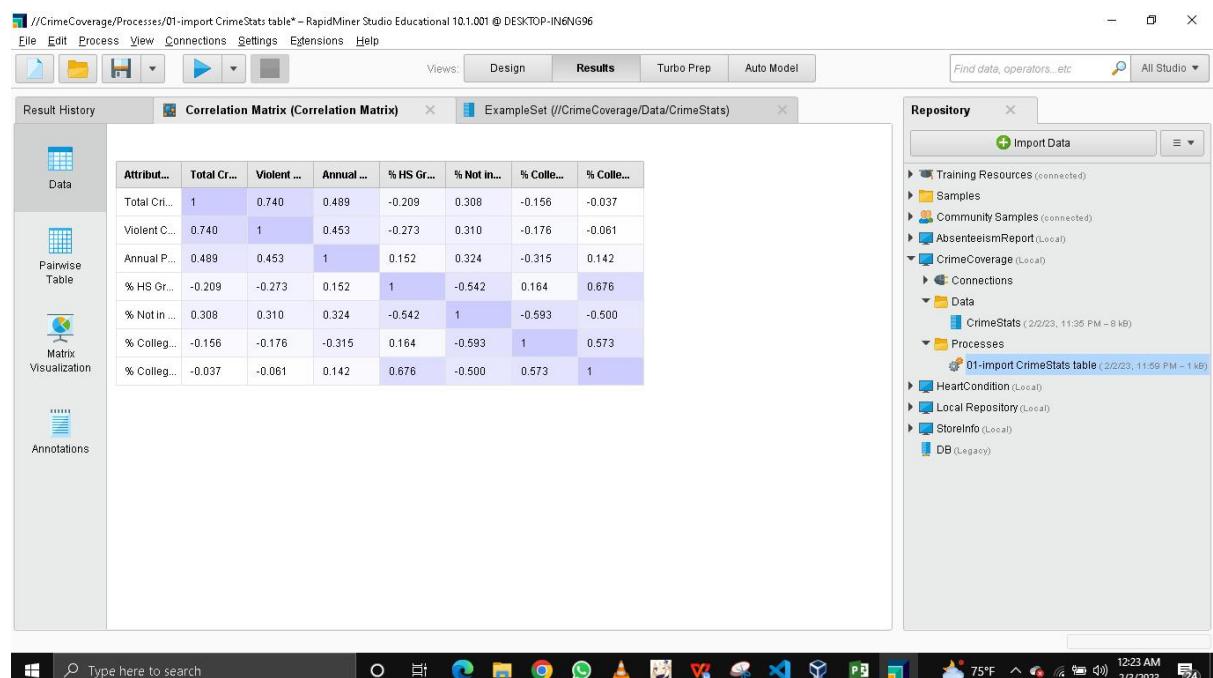
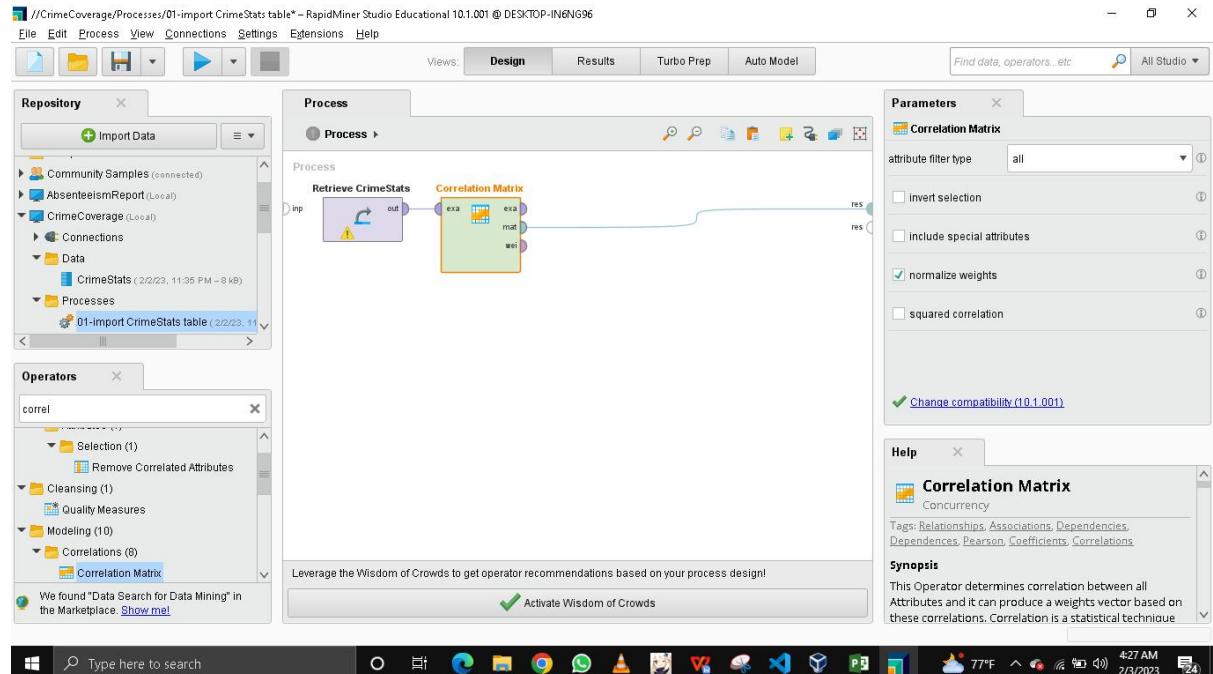
Synopsis

This Operator can access stored information in the Repository and load them into the Process.



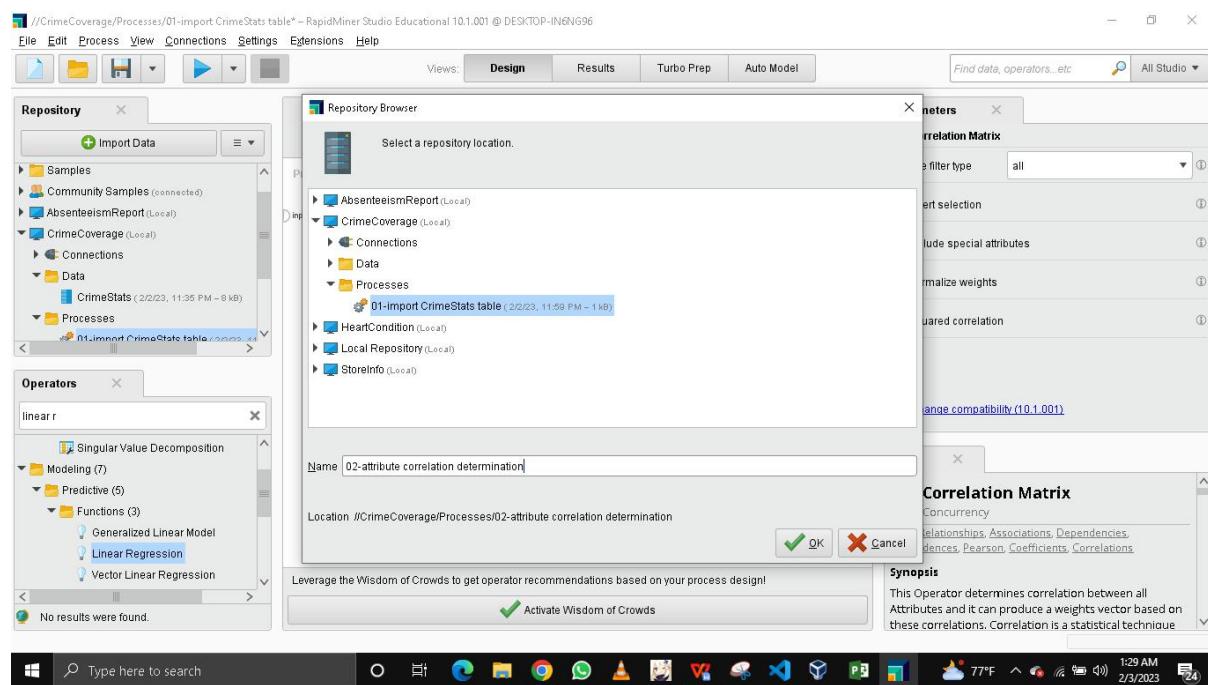


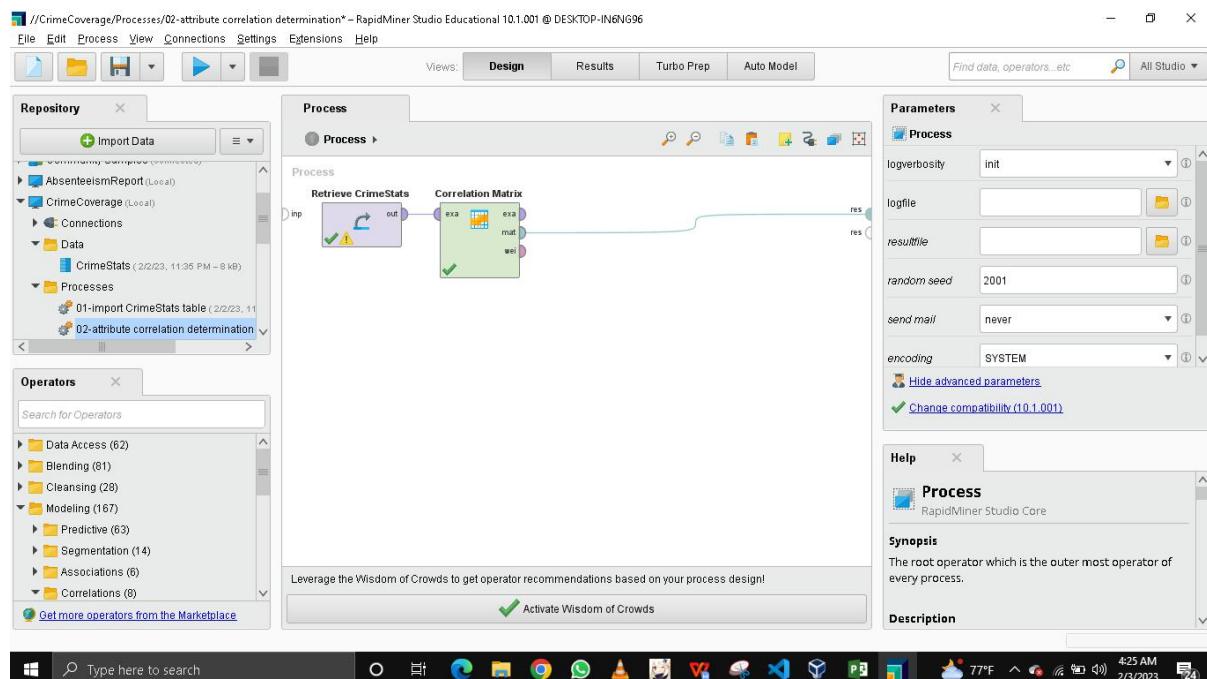
4. Used the “Correlation Matrix” operator to analyze the correlation between each pair of attributes and saved the process as “02-attribute correlation determination”



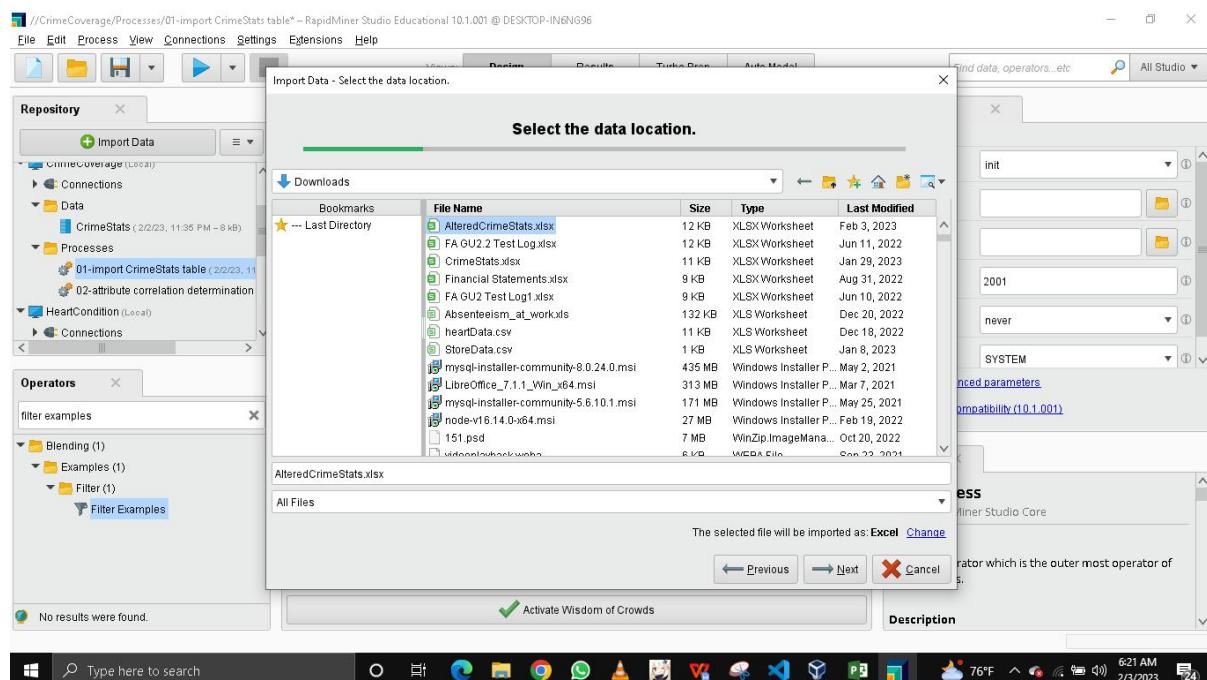
As indicated by the above Correlation Matrix;

- The Total Crime Rate (per 1M Residents) and Violent Crime Rate (per 100K Residents) share a coefficient of 0.74 indicating a strong positive correlation between the two attributes
- The % HS Graduates (25+) and % College Graduates (25+) share a coefficient of 0.676 indicating a strong positive correlation between the two attributes
- The % Not in HS / Not HS Graduates and % College-Going (18-24) share a coefficient of -0.593 indicating a strong negative correlation between the two attributes





5. Imported a new “AlteredCrimeStats” table into the Data folder of the CrimeCoverage Repository with the table’s column headings changed to match those of the “CrimeStats” table and saved the process as 03-import AlteredCrimeStats table



//CrimeCoverage/Processes/01-import CrimeStats table - RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	X1 integer	X2 integer	X3 integer	X4 integer	X5 integer	X6 integer
1	478	184	40	74	11	31
2	494	?	32	72	11	43
3	643	347	57	70	18	16
4	341	565	31	71	11	25
5	773	?	67	72	9	29
6	603	260	25	68	8	32
7	484	325	34	68	12	24
8	546	?	33	62	13	28
9	424	38	36	69	7	25
10	548	226	31	66	9	58
11	506	?	35	60	13	21
12	810	360	20	81	4	77

✓ no problems.

← Previous Next → ✖ Cancel

Activate Wisdom of Crowds

Description

No results were found.

Windows Type here to search 6:24 AM 2/3/2023

//CrimeCoverage/Processes/01-import CrimeStats table - RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Import Data - Format your columns.

Format your columns.

Replace errors with missing values ⓘ

	Total Crim... integer	Violent Cri... integer	Annual Pol... integer	% HS Grad... integer	% Not in HS... integer	% College-... integer
1	478	184	40	74	11	31
2	494	?	32	72	11	43
3	643	347	57	70	18	16
4	341	565	31	71	11	25
5	773	?	67	72	9	29
6	603	260	25	68	8	32
7	484	325	34	68	12	24
8	546	?	33	62	13	28
9	424	38	36	69	7	25
10	548	226	31	66	9	58
11	506	?	35	60	13	21
12	810	360	20	81	4	77

✓ no problems.

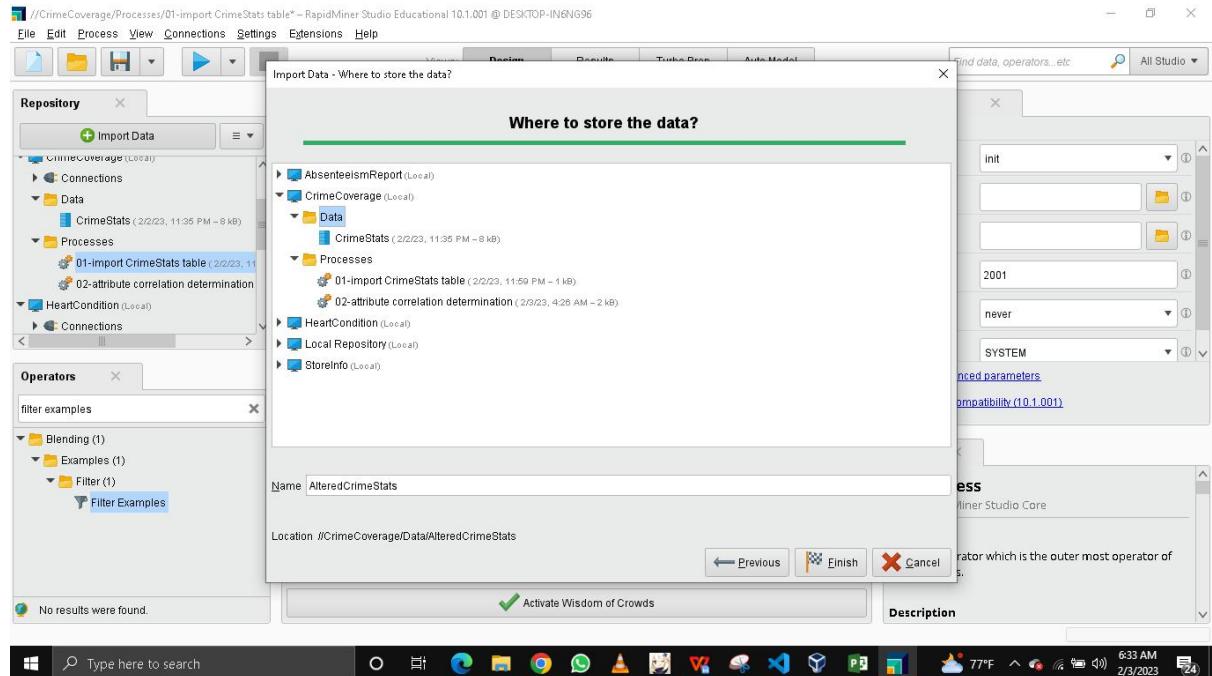
← Previous Next → ✖ Cancel

Activate Wisdom of Crowds

Description

No results were found.

Windows Type here to search 6:32 AM 2/3/2023



Row No.	Total Crime ...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS /...	% College-G...	College Gra...
1	478	184	40	74	11	31	20
2	494	?	32	72	11	43	18
3	643	347	57	70	18	16	16
4	341	565	31	71	11	25	19
5	773	?	67	72	9	29	24
6	603	260	25	68	8	32	15
7	494	325	34	68	12	24	14
8	546	?	33	62	13	28	11
9	424	38	36	69	7	25	12
10	548	226	31	66	9	58	15
11	506	?	35	60	13	21	9
12	819	369	30	81	4	77	36
13	541	109	44	66	9	37	12
14	491	?	32	67	11	37	16
15	514	29	30	65	12	35	11

ExampleSet (100 examples, 0 special attributes, 7 regular attributes)

//CrimeCoverage/Processes/03-import AlteredCrimeStats table* – RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operators... etc All Studio

Result History ExampleSet //CrimeCoverage/Data/AlteredCrimeStats Correlation Matrix (Correlation Matrix)

Data Statistics Visualizations Annotations

Open in Turbo Prep Auto Model Filter (100 / 100 examples): all

Row No.	Total Crime ...	Violent Crime...	Annual Polic...	% HS Gradu...	% Not in HS ...	% College-G...	College Gra...
1	478	184	40	74	11	31	20
2	494	?	32	72	11	43	18
3	643	347	57	70	18	16	16
4	341	565	31	71	11	25	19
5	773	?	67	72	9	29	24
6	603	260	25	68	8	32	15
7	484	325	34	68	12	24	14
8	546	?	33	62	13	28	11
9	424	38	36	69	7	25	12
10	548	226	31	66	9	58	15
11	506	?	35	60	13	21	9
12	819	369	30	81	4	77	36
13	541	109	44	66	9	37	12
14	491	?	32	67	11	37	16
15	514	29	30	65	12	35	11

ExampleSet (100 examples, 0 special attributes, 7 regular attributes)

Repository Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- AbsenteeismReport (Local)
- CrimeCoverage (Local)
 - Connections
 - AlteredCrimeStats

AlteredCrimeStats

Data table
Number of examples = 100
7 attributes:

Role	Name	Type	Range	Missing
Total Crime...	# integer	=[341 - 17...]	= 0	
Violent Crime...	# integer	= [29 - 3549]	= 20	
Annual Polic...	# integer	= [16 - 87]	= 0	
% HS Gradu...	# integer	= [42 - 81]	= 0	
% Not in HS ...	# integer	= [4 - 34]	= 0	
% College-G...	# integer	= [7 - 81]	= 0	

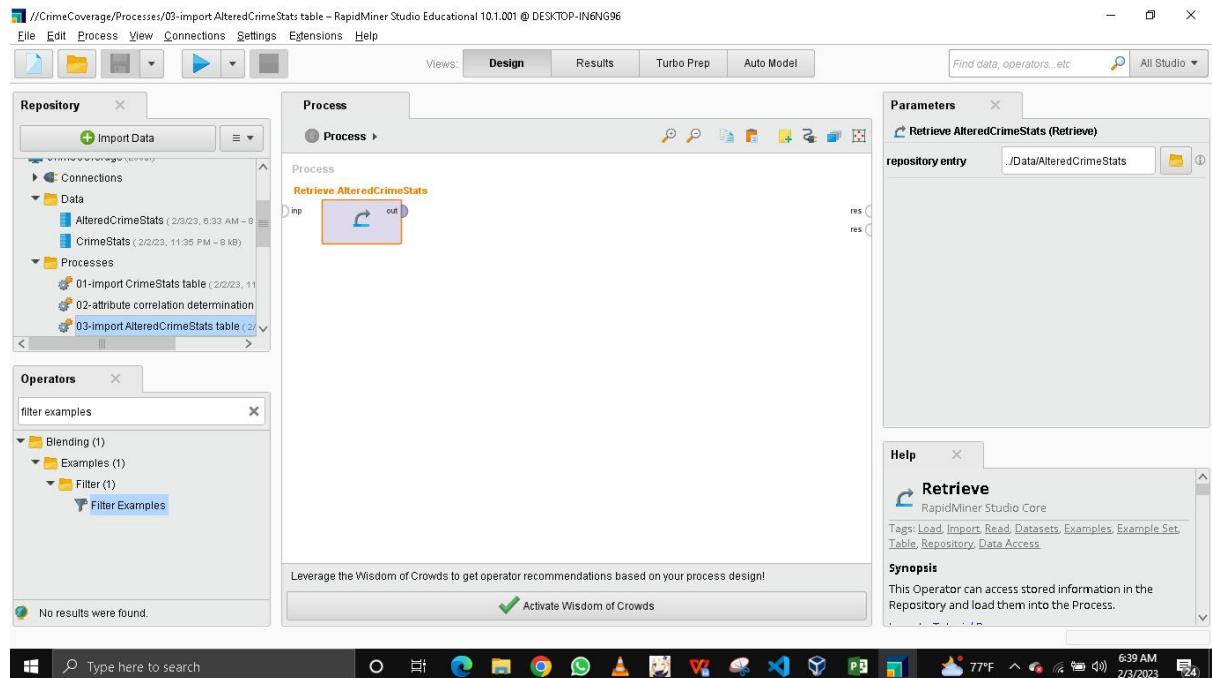
04-performance evaluation (12/18/22, 10:07 PM -)

Press F3* for focus.

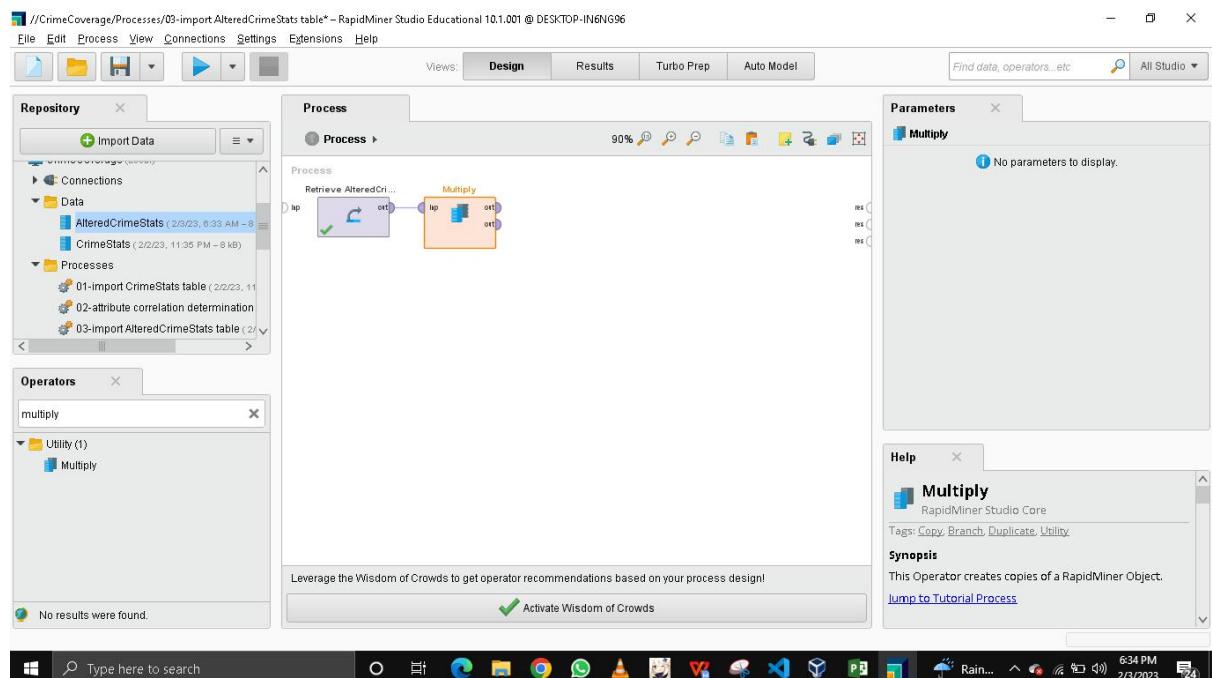
6:49 AM 77°F 2/3/2023

The screenshot shows the RapidMiner Studio interface with the following details:

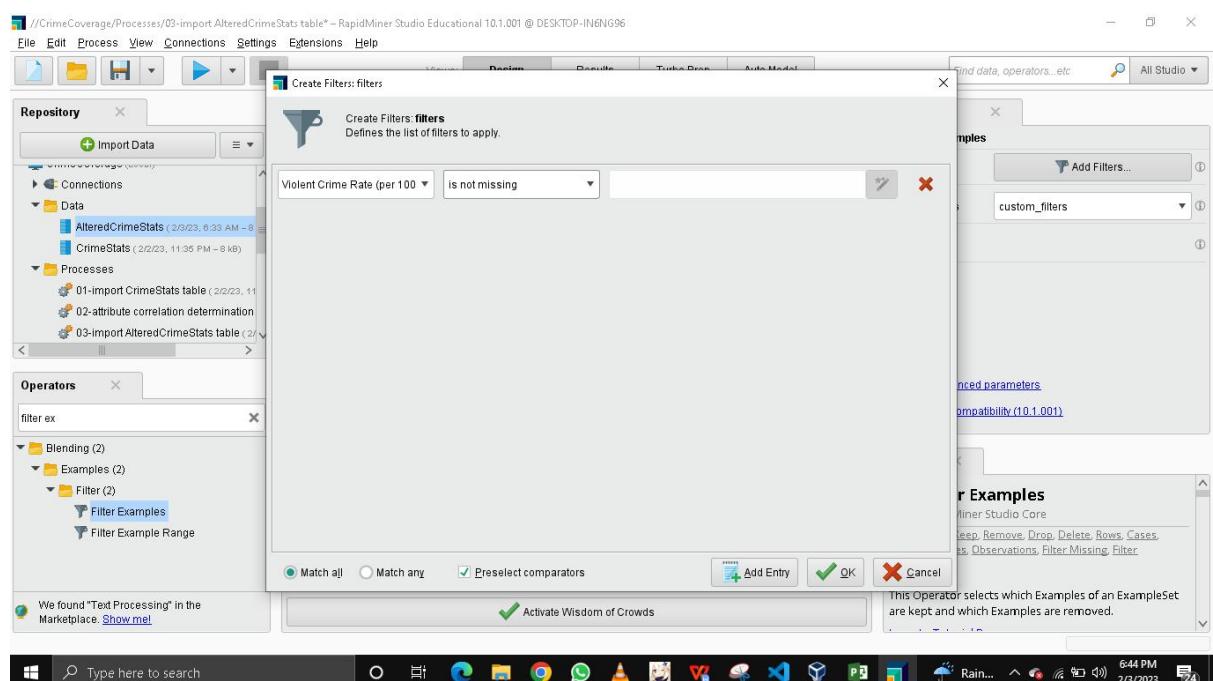
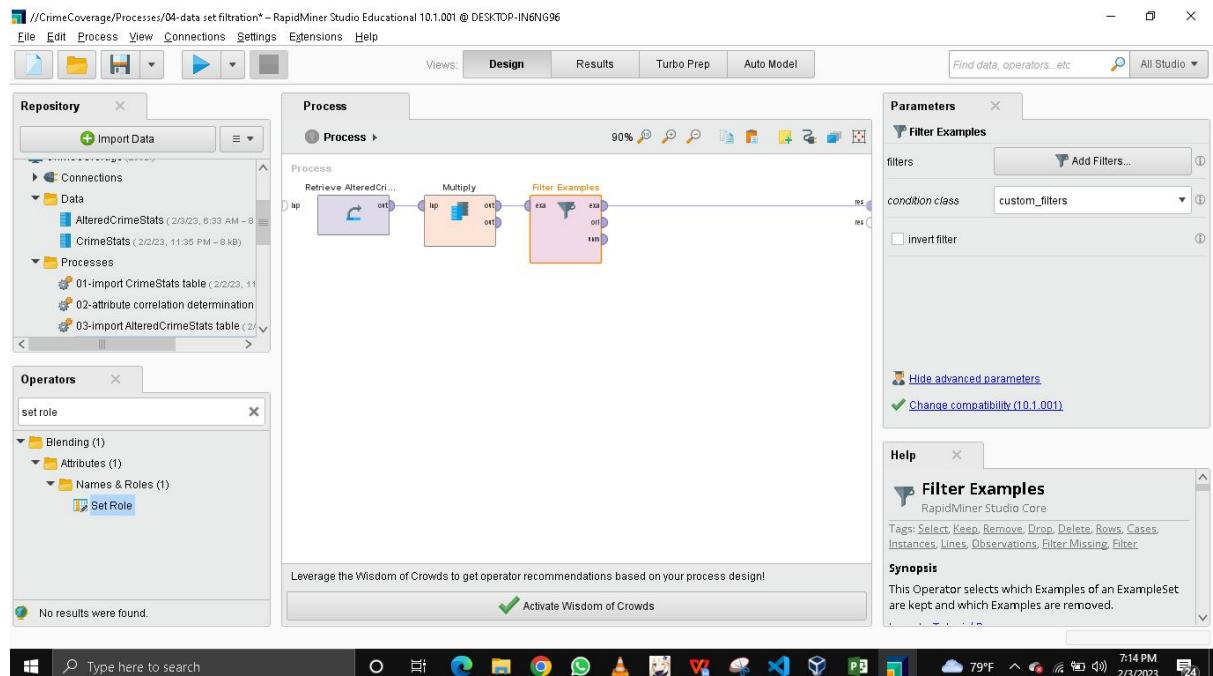
- Title Bar:** //CrimeCoverage/Processes/01-import CrimeStats table - RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96
- Menu Bar:** File, Edit, Process, View, Connections, Settings, Extensions, Help
- Toolbars:** Views, Design, Results, Turbo Prep, Auto Model
- Search Bar:** Find data, operators...etc, All Studio
- Repository Browser:** Select a repository location. The tree view shows:
 - Data
 - AlteredCrimeStats (2/2/23, 6:33 AM - 8 kB)
 - CrimeStats (2/2/23, 11:35 PM - 8 kB)
 - Processes
 - 01-import CrimeStats table (2/2/23, 11:59 PM - 1 kB)
 - 02-attribute correlation determination
 - HeartCondition (Local)
 - Connections
- Operators:** filter examples
 - Blending (1)
 - Examples (1)
 - Filter (1)
 - Filter ExamplesNo results were found.
- Create New Process Dialog:** Name: 03-import AlteredCrimeStats table, Location: //CrimeCoverage/Processes/03-import AlteredCrimeStats table, OK and Cancel buttons, Activate Wisdom of Crowds checkbox.
- Retrieve Dialog:** Retrieve AlteredCrimeStats (Retrieve), Repository entry: ./Data/AlteredCrimeStats, OK and Cancel buttons.
- Retrieval Summary:** This Operator can access stored information in the Repository and load them into the Process.
- System Status:** 77°F, 2/13/2023, 6:38 AM

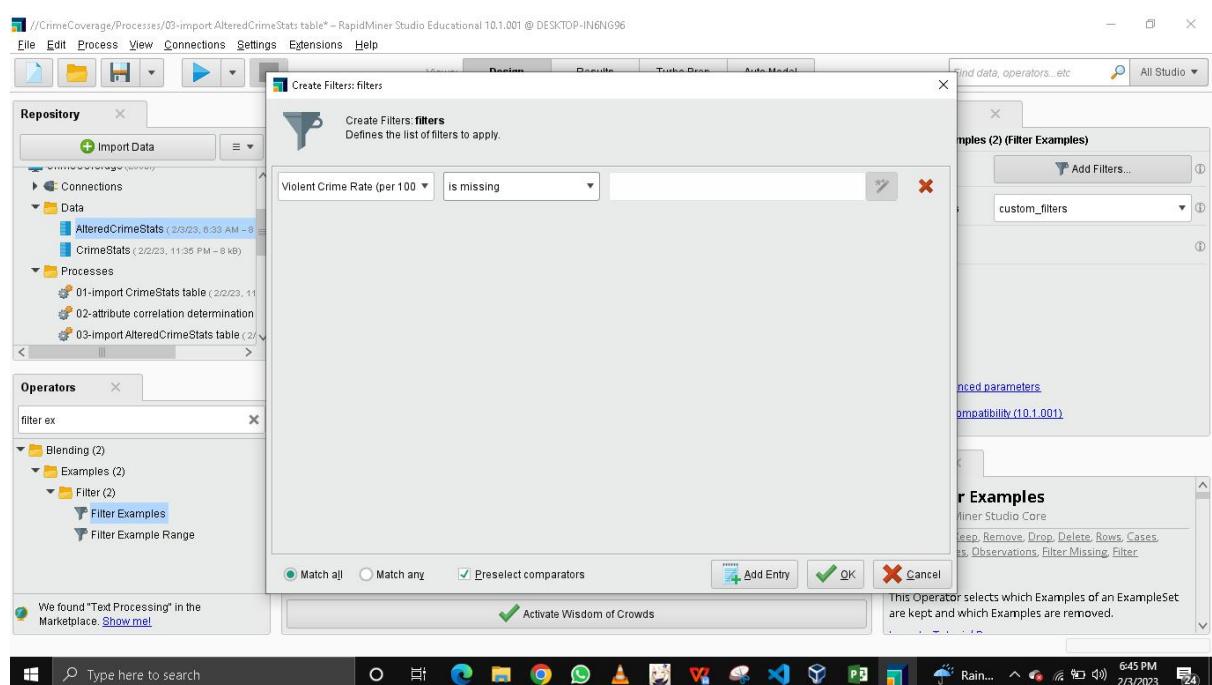
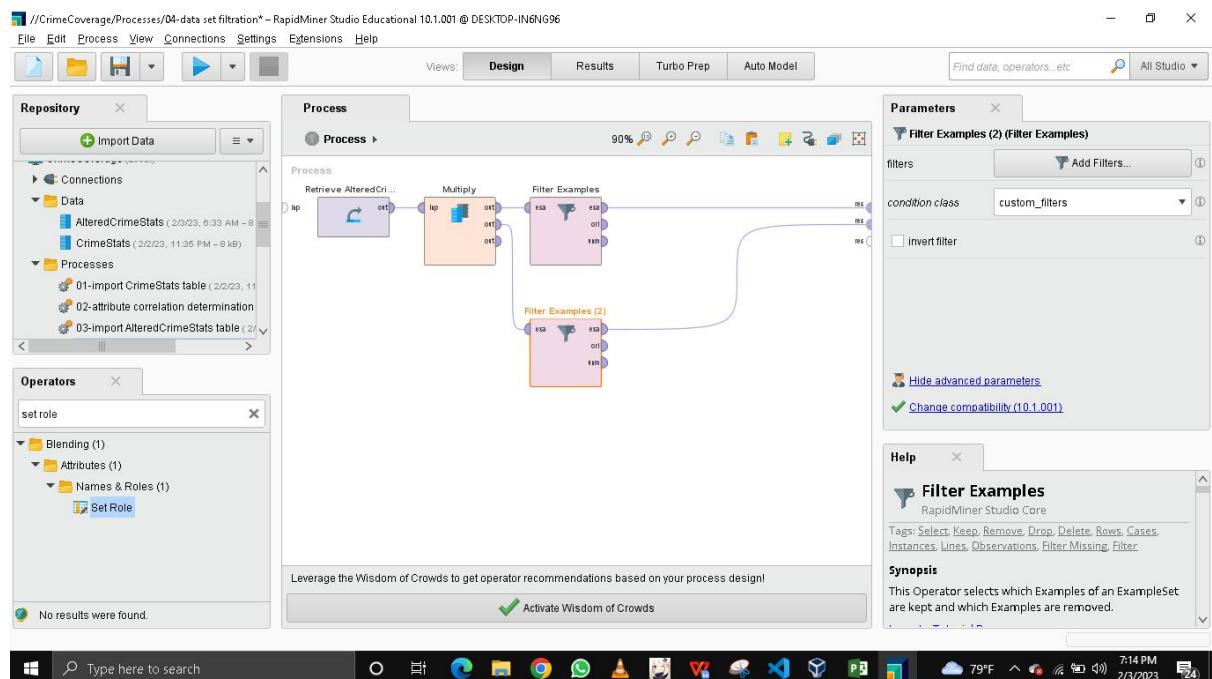


6. Used the “Multiply” operator to duplicate the AlteredCrimeStats table data in order to prepare test data and model data through a filtering process



7. Used “Filter Examples” operators on each copy of the AlteredCrimeStats tables to only show fields under the “Violent Crime Rate (per 100k Residents)” column containing defined values and non defined values to prepare the training and test data sets respectively; saving the process as “04-data set filtration”





Shown below are the results of the filtration process for the training and test data sets respectively

Result History ExampleSet (Filter Examples) ExampleSet (Filter Examples (2))

Views: Design Results Turbo Prep Auto Model

Find data, operators... etc. All Studio

Data

Open in: Turbo Prep Auto Model

Filter (80 / 80 examples): all

Row No.	Total Crime ...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS /...	% College-G...	College Gra...
1	478	184	40	74	11	31	20
2	643	347	57	70	18	16	16
3	341	565	31	71	11	25	19
4	603	260	25	68	8	32	15
5	484	325	34	68	12	24	14
6	424	38	36	69	7	25	12
7	548	226	31	66	9	58	15
8	819	369	30	81	4	77	36
9	541	109	44	66	9	37	12
10	514	29	30	65	12	35	11
11	371	245	16	64	10	42	14
12	437	148	36	62	7	81	27
13	570	387	30	59	15	31	16
14	432	98	23	56	15	50	15
15	357	218	35	54	14	27	13

ExampleSet(80 examples, 0 special attributes, 7 regular attributes)

Repository

Import Data

- Training Resources (connected)
 - Samples
 - Community Samples (connected)
 - AbsenteeismReport (Local)
 - CrimeCoverage (Local)
 - Connections
 - Data
 - AlteredCrimeStats (2/2/23, 6:33 AM - 8 kB)
 - CrimeStats (2/2/23, 11:35 PM - 8 kB)
 - Processes
 - 01-import CrimeStats table (2/2/23, 11:59 PM - 1 kB)
 - 02-attribute correlation determination (2/2/23, 4:26 AM)
 - 03-import AlteredCrimeStats table (2/2/23, 6:38 AM - 1 kB)
 - 04-data set filtration (2/3/23, 6:59 PM - 2 kB)
- HeartCondition (Local)
 - Connections
 - Data
 - Processes
 - 01-import heartData table (12/18/22, 9:11 PM - 905 bytes)
 - 02-build decision tree (12/18/22, 9:42 PM - 3 kB)
 - 03-apply the model (12/18/22, 9:19 AM - 3 kB)

Result History ExampleSet (Filter Examples) ExampleSet (Filter Examples (2))

Views: Design Results Turbo Prep Auto Model

Find data, operators... etc. All Studio

Data

Open in: Turbo Prep Auto Model

Filter (20 / 20 examples): all

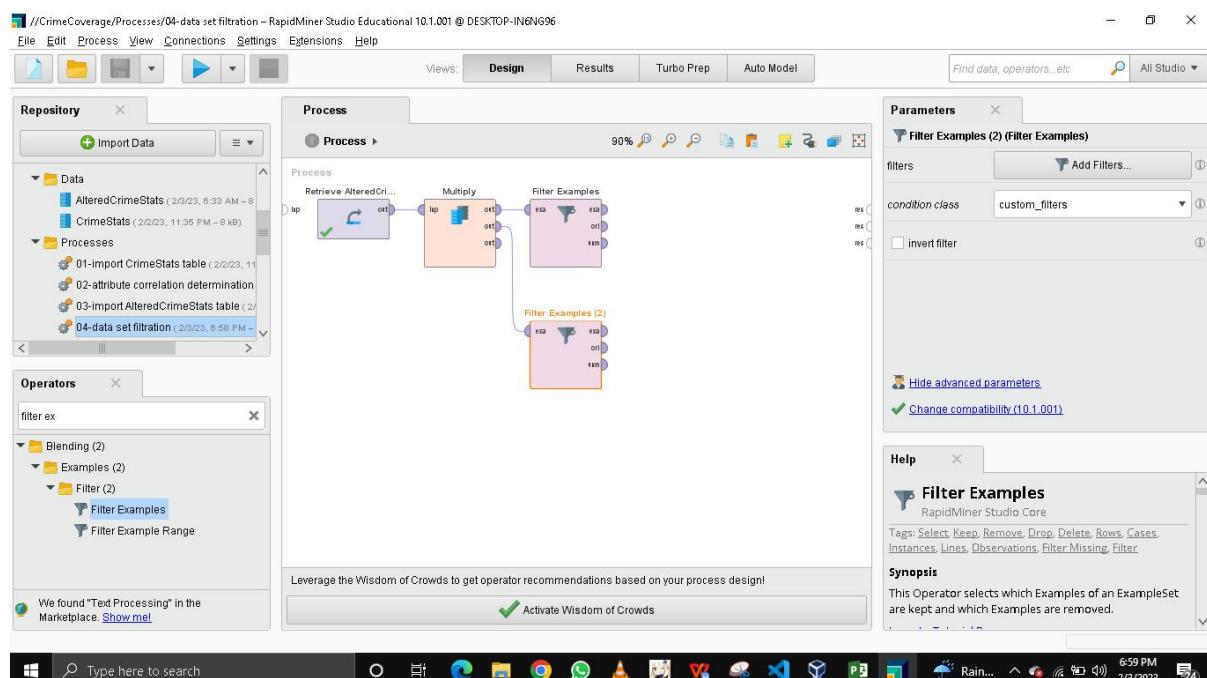
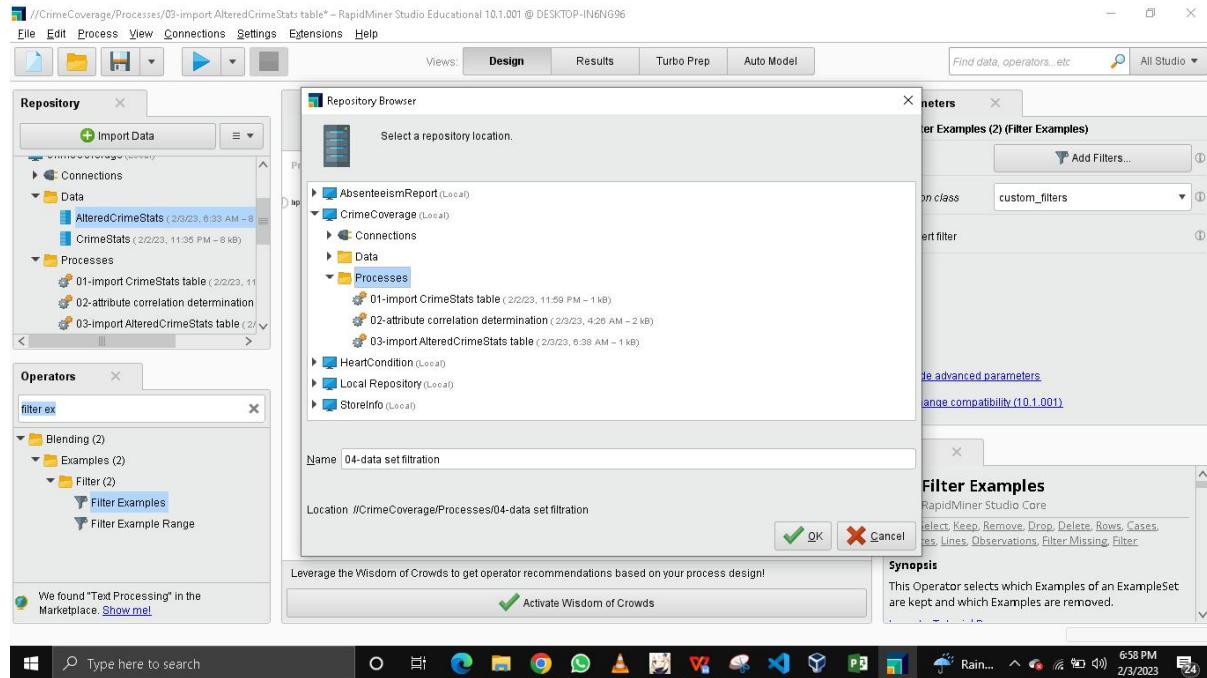
Row No.	Total Crime ...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS /...	% College-G...	College Gra...
1	494	?	32	72	11	43	18
2	773	?	67	72	9	29	24
3	546	?	33	62	13	28	11
4	506	?	35	60	13	21	9
5	491	?	32	67	11	37	16
6	457	?	29	64	12	21	10
7	619	?	33	46	22	24	8
8	439	?	31	61	19	14	12
9	805	?	46	57	14	25	12
10	1419	?	43	59	14	21	13
11	863	?	20	69	23	7	12
12	940	?	66	67	26	18	16
13	343	?	31	72	11	25	19
14	819	?	30	81	4	77	36
15	437	?	36	62	7	81	27

ExampleSet(20 examples, 0 special attributes, 7 regular attributes)

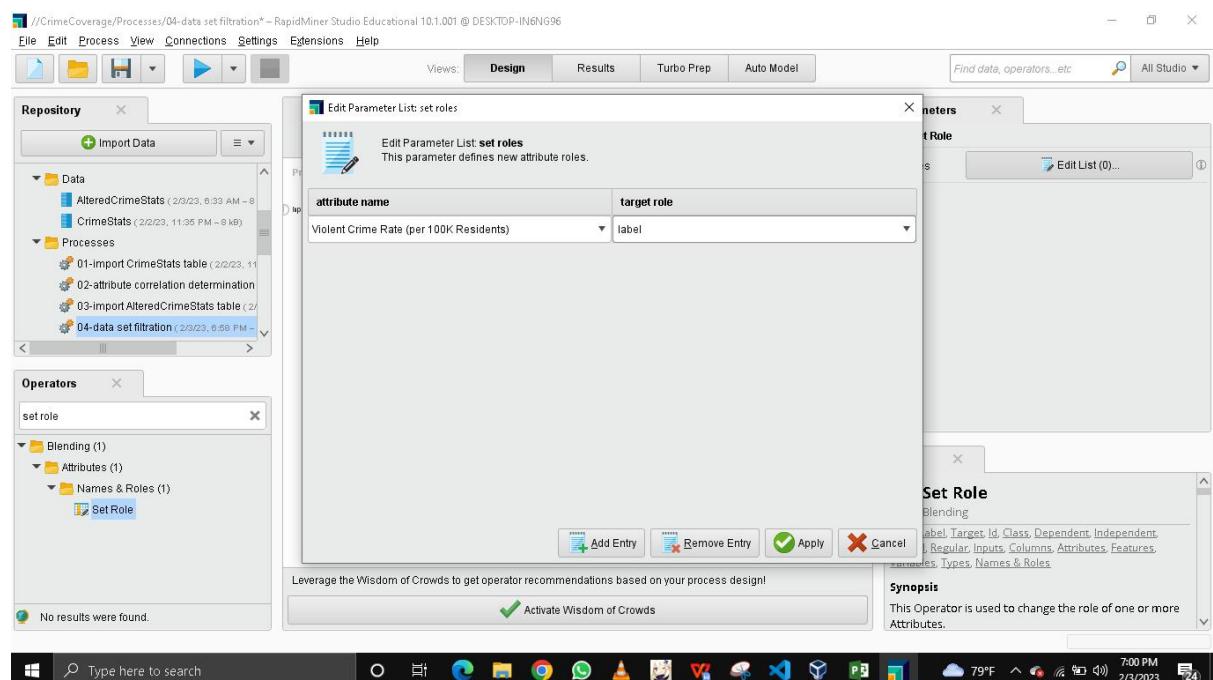
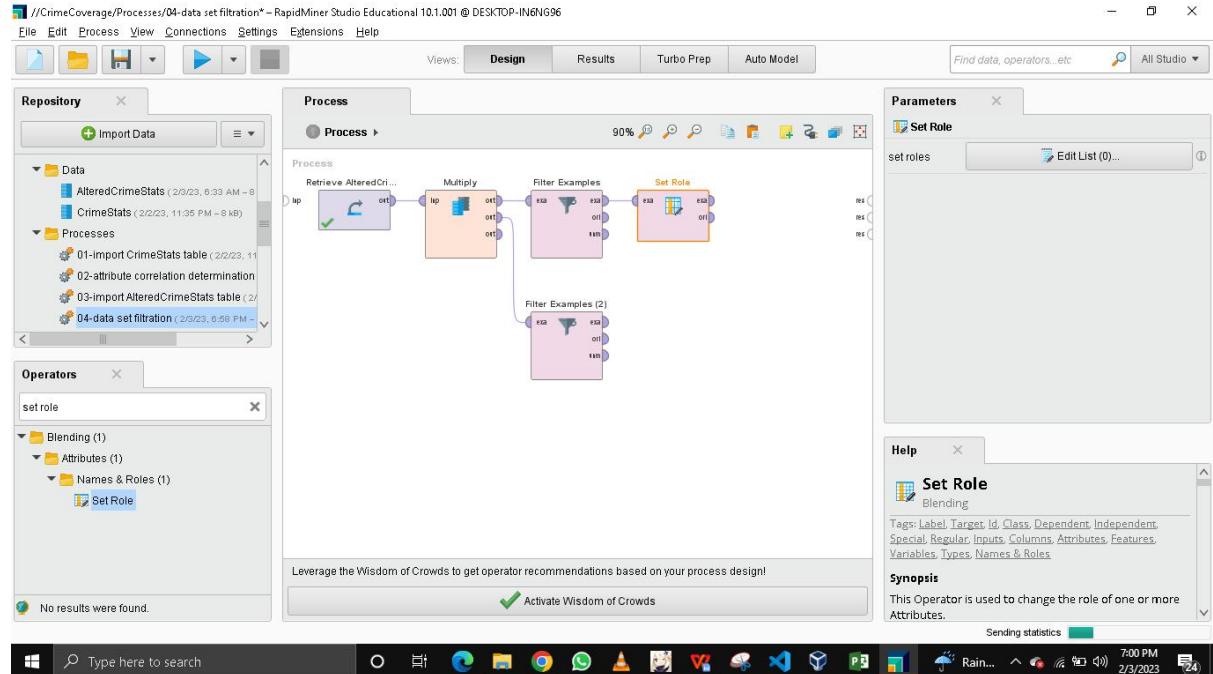
Repository

Import Data

- Training Resources (connected)
 - Samples
 - Community Samples (connected)
 - AbsenteeismReport (Local)
 - CrimeCoverage (Local)
 - Connections
 - Data
 - AlteredCrimeStats (2/2/23, 6:33 AM - 8 kB)
 - CrimeStats (2/2/23, 11:35 PM - 8 kB)
 - Processes
 - 01-import CrimeStats table (2/2/23, 11:59 PM - 1 kB)
 - 02-attribute correlation determination (2/2/23, 4:26 AM)
 - 03-import AlteredCrimeStats table (2/2/23, 6:38 AM - 1 kB)
 - 04-data set filtration (2/3/23, 6:59 PM - 2 kB)
- HeartCondition (Local)
 - Connections
 - Data
 - Processes
 - 01-import heartData table (12/18/22, 9:11 PM - 905 bytes)
 - 02-build decision tree (12/18/22, 9:42 PM - 3 kB)
 - 03-apply the model (12/18/22, 9:19 AM - 3 kB)



8. Used the “Set Role” operator on the filtered training data set and classified the “Violent Crime Rate (per 100k Residents)” column within the AlteredCrimeStats table as the label



//CrimeCoverage/Processes/04-data set filtration* - RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operators...etc All Studio

ExampleSet (/CrimeCoverage/Data/AlteredCrimeStats)

ExampleSet (Set Role) **ExampleSet (Filter Examples (2))**

Data Statistics Visualizations Annotations

Open in Turbo Prep Auto Model Filter (80 / 80 examples): all

Row No.	Violent Crime Rate (per 100K Residents)	Total Crime Rate (per 1M Residents)	Annual Police Funding (\$/Resident)	% HS Graduates (25+)	% Not in HS / Not HS Graduates	% College-Going (18-24)	College Graduates (25+)
1	184	478	40	74	11	31	20
2	347	643	57	70	18	16	16
3	565	341	31	71	11	25	19
4	260	603	25	68	8	32	15
5	325	484	34	68	12	24	14
6	38	424	36	69	7	25	12
7	226	548	31	66	9	58	15
8	369	819	30	81	4	77	36
9	109	541	44	66	9	37	12
10	29	514	30	65	12	35	11
11	245	371	16	64	10	42	14
12	148	437	36	62	7	81	27
13	387	570	30	59	15	31	16
14	98	432	23	56	15	50	15

ExampleSet (80 examples, 1 special attribute, 6 regular attributes)

Repository Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- AbsenteeismReport (Local)
- CrimeCoverage (Local)
 - Connections
 - Data
 - AlteredCrimeStats (2/2/23, 6:33 AM - 8 kB)
 - CrimeStats (2/2/23, 11:35 PM - 8 kB)
 - Processes
 - 01-import CrimeStats table (2/2/23, 11:59 PM - 1 kB)
 - 02-attribute correlation determination (2/2/23, 4:2 kB)
 - 03-import AlteredCrimeStats table (2/2/23, 6:38 AM - 9 kB)
 - 04-data set filtration (2/2/23, 6:58 PM - 2 kB)
- HeartCondition (Local)
 - Connections
 - Data
 - Processes
 - 01-import heartData table (12/18/22, 9:11 PM - 99 kB)
 - 02-build decision tree (12/18/22, 9:42 PM - 3 kB)
 - 03-apply the model (12/18/22, 9:19 AM - 3 kB)

Windows Type here to search 7:03 PM 2/3/2023

//CrimeCoverage/Processes/04-data set filtration* - RapidMiner Studio Educational 10.1.001 @ DESKTOP-IN6NG96

File Edit Process View Connections Settings Extensions Help

Views: Design Results Turbo Prep Auto Model Find data, operators...etc All Studio

ExampleSet (/CrimeCoverage/Data/AlteredCrimeStats)

ExampleSet (Set Role)

Name	Type	Missing	Statist...	Filter (7 / 7 attributes):	Search for Attributes
Label					
Violent Crime Rate (per 100K Residents)	Integer	0	Min 29	Max 3545	Ave 57
Total Crime Rate (per 1M Residents)	Integer	0	Min 341	Max 1740	Ave 68
Annual Police Funding (\$/Resident)	Integer	0	Min 16	Max 87	Ave 38
% HS Graduates (25+)	Integer	0	Min 42	Max 81	Ave 60
% Not in HS / Not HS Graduates	Integer	0	Min 4	Max 34	Ave 14
% College-Going (18-24)	Integer	0	Min 13	Max 81	Ave 30
College Graduates (25+)	Integer	0	Min 8	Max 36	Ave 14

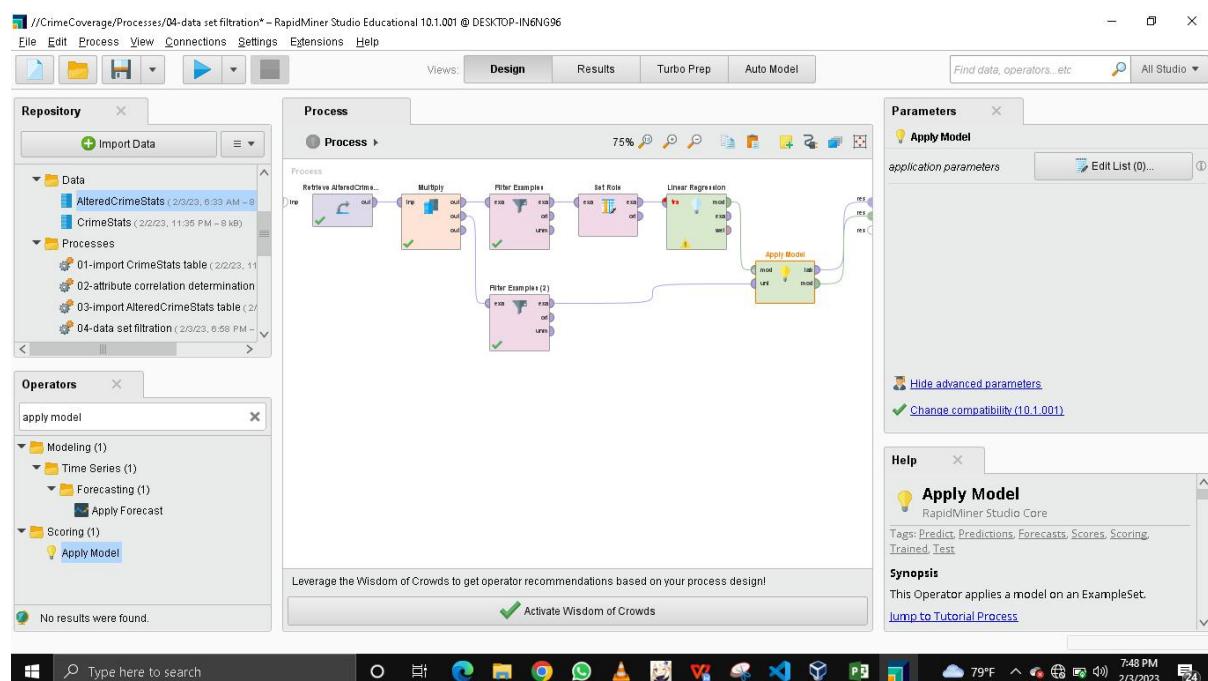
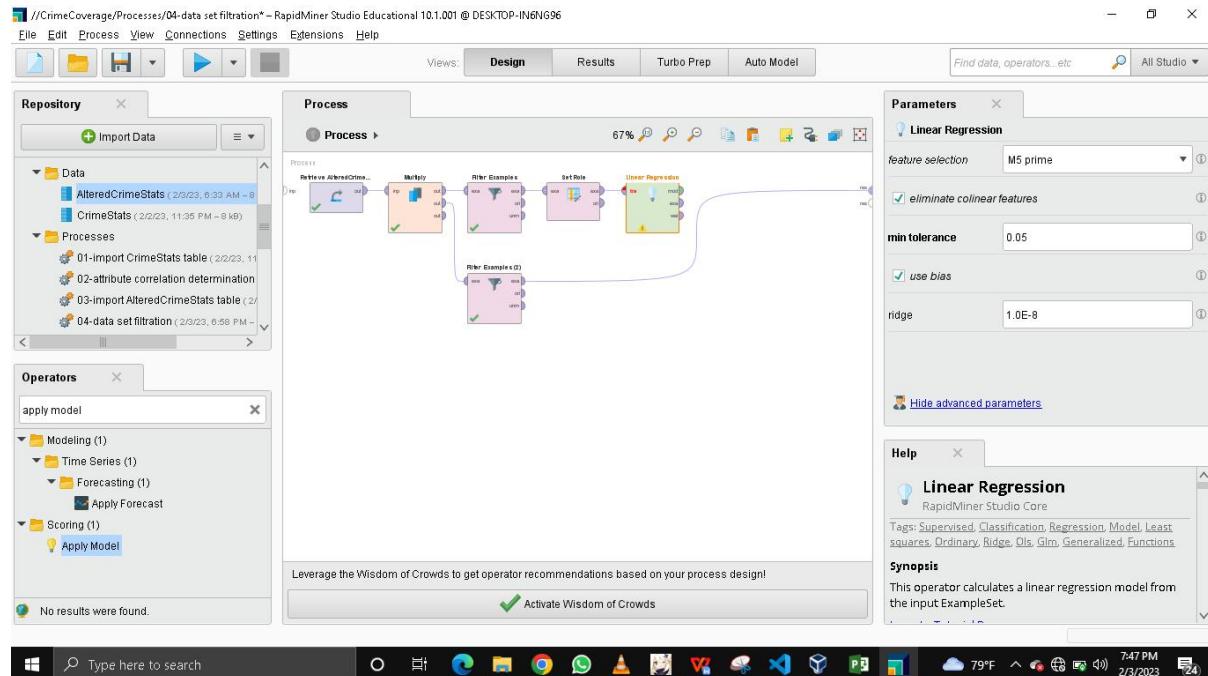
Showing attributes 1 - 7 Examples: 80 Special Attributes: 1 Regular Attributes: 6

Repository Import Data

- Training Resources (connected)
- Samples
- Community Samples (connected)
- AbsenteeismReport (Local)
- CrimeCoverage (Local)
 - Connections
 - Data
 - AlteredCrimeStats (2/2/23, 6:33 AM - 8 kB)
 - CrimeStats (2/2/23, 11:35 PM - 8 kB)
 - Processes
 - 01-import CrimeStats table (2/2/23, 11:59 PM - 1 kB)
 - 02-attribute correlation determination (2/2/23, 4:2 kB)
 - 03-import AlteredCrimeStats table (2/2/23, 6:38 AM - 9 kB)
 - 04-data set filtration (2/2/23, 6:58 PM - 2 kB)
- HeartCondition (Local)
 - Connections
 - Data
 - Processes
 - 01-import heartData table (12/18/22, 9:11 PM - 99 kB)
 - 02-build decision tree (12/18/22, 9:42 PM - 3 kB)
 - 03-apply the model (12/18/22, 9:19 AM - 3 kB)

Windows Type here to search 7:04 PM 2/3/2023

9. Used the “Linear Regression” and “Apply Model” operators alongside each other to conduct linear regression analysis on the “Violent Crime Rate (per 100k Residents)” column for both datasets and saved the process as “05-linear regression analysis”



Shown below are the Violent Crime Rate (per 100k Residents) predicted values and the coefficient indexes of each attribute

Result History ExampleSet (Apply Model) LinearRegression (Linear Regression)

Row No. ↑	predictionV...	Total Crime ...	Violent Crim...	Annual Polic...	% HS Gradu...	% Not in HS /...	% College-G...	Coll
1	74.451	494	?	32	72	11	43	18
2	1077.265	773	?	67	72	9	29	24
3	260.070	546	?	33	62	13	28	11
4	270.708	506	?	35	60	13	21	9
5	179.574	491	?	32	67	11	37	16
6	104.131	457	?	29	64	12	21	10
7	492.178	619	?	33	46	22	24	8
8	129.384	439	?	31	61	19	14	12
9	883.532	805	?	46	57	14	25	12
10	1624.844	1419	?	43	59	14	21	13
11	332.299	863	?	20	69	23	7	12
12	882.755	940	?	66	67	26	18	16
13	32.656	343	?	31	72	11	25	19
14	669.672	819	?	30	81	4	77	36

ExampleSet(20 examples, 1 special attribute, 7 regular attributes)

Result History ExampleSet (Apply Model) LinearRegression (Linear Regression)

Attribute	Coefficient	Std. Error	Std. Coefficient	Tolerance	t-Stat	p-Value	Code
Total Crime R...	1.272	0.185	0.615	0.623	6.875	0.000	****
Annual Police ...	9.439	4.648	0.211	0.745	2.031	0.046	**
% HS Graduat...	-24.528	8.276	-0.397	0.958	-2.964	0.004	***
% Not in HS /...	-21.805	13.406	-0.193	0.777	-1.627	0.108	
% College-Gol...	-7.278	5.100	-0.154	0.974	-1.427	0.158	
College Gradua...	28.685	16.034	0.215	0.980	1.789	0.078	*
(Intercept)	946.633	601.284	?	?	1.574	0.120	

Linear Regression Accuracy Verification

In order to verify the accuracy of the predictions made on the test data using the training data; the usage of the $y = xm + b$ formula is necessary. Each attribute of the formula is defined as follows;

- b - Intercept Value
- m - Attribute Coefficient Value
- x - Table Attribute Value
- y - True Value

Shown below is an example using the first two row of test data as values for the equation:

$$(494 \times 1.272) + (32 \times 9.439) + (72 \times (-24.528)) + (11 \times (-21.805)) + (43 \times (-7.278)) + (18 \times 28.685) + 946.633 \\ = 74.554$$

Given that the data in the screenshots above are predictions of the dependent variable and not true values it is expected that the values differ. The difference in these values is called the residual error which in this case is a value of 0.103 between the two values.

$$74.554 - 74.451 = 0.103$$

For better accuracy and reliability on these results the second row of test data is used as well:

$$(773 \times 1.272) + (67 \times 9.439) + (72 \times (-24.528)) + (9 \times (-21.805)) + (29 \times (-7.278)) + (24 \times 28.685) + 946.663 \\ = 1077.449$$

With the residual error of this set of data being a value of 0.184 between the two values.

$$1077.449 - 1,077.265 = 0.184$$

With the results above an assumption declaring that the residual error lies within a range of 0.1 and 0.2 can be made however is not absolute.

In terms of the prediction accuracy of the linear regression method using the above results a calculation can be made as follows:

$$((1,077.265/1077.449) * 100 + (74.451/74.554) * 100) / 2 = 99.922383865\%$$

It is important to note that this data however accurate is simply theoretical and not a realistic representative of the actual Violent Crime Rate (per 100k Residents) in the United States

