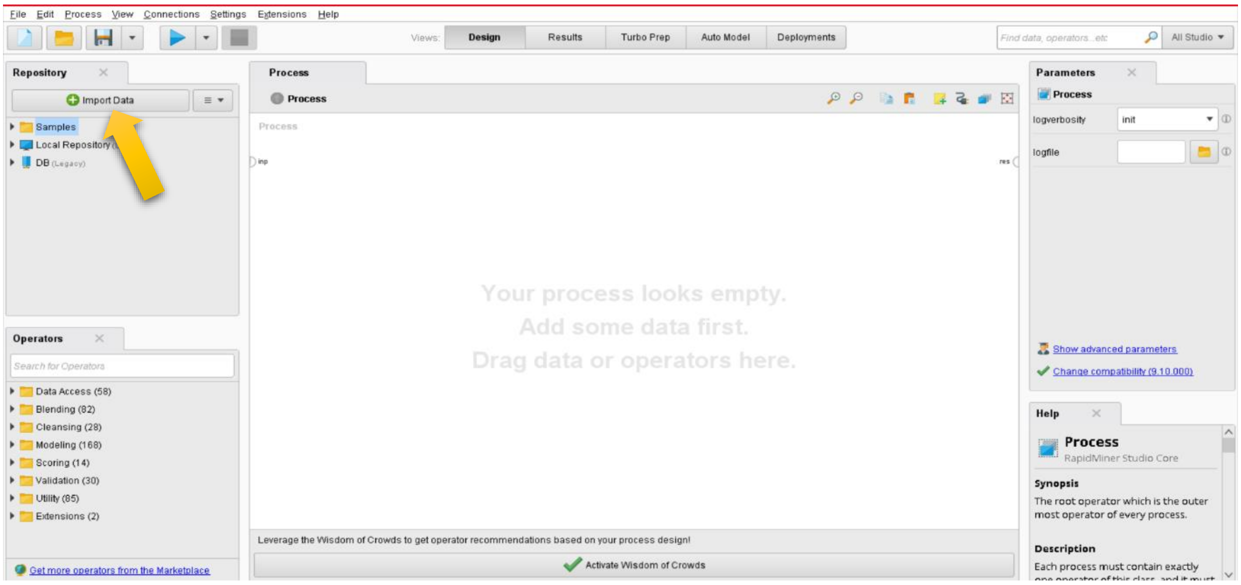
Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **01** Page No. **01**	Total Pages. **17**

Q1. Perform the steps to import data from your local computer.

Answer: For importing data, first we have a data set in our system.

Step 1: Open rapid miner and click on **Import Data.**



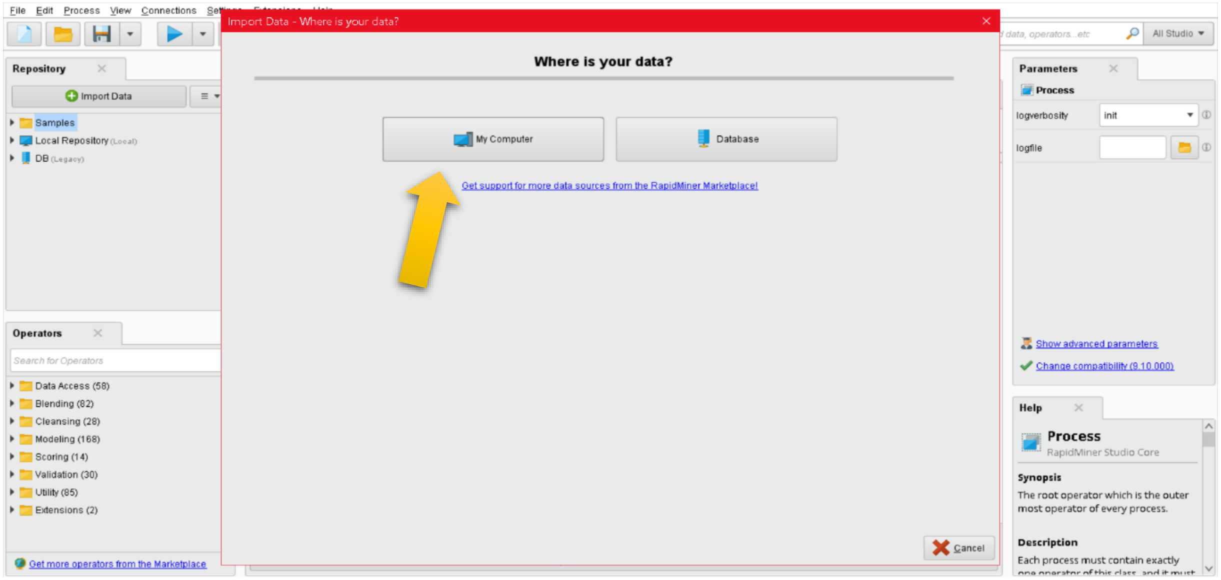
Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

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

Each process must contain exactly

Step 2: Now click on **My computer**.

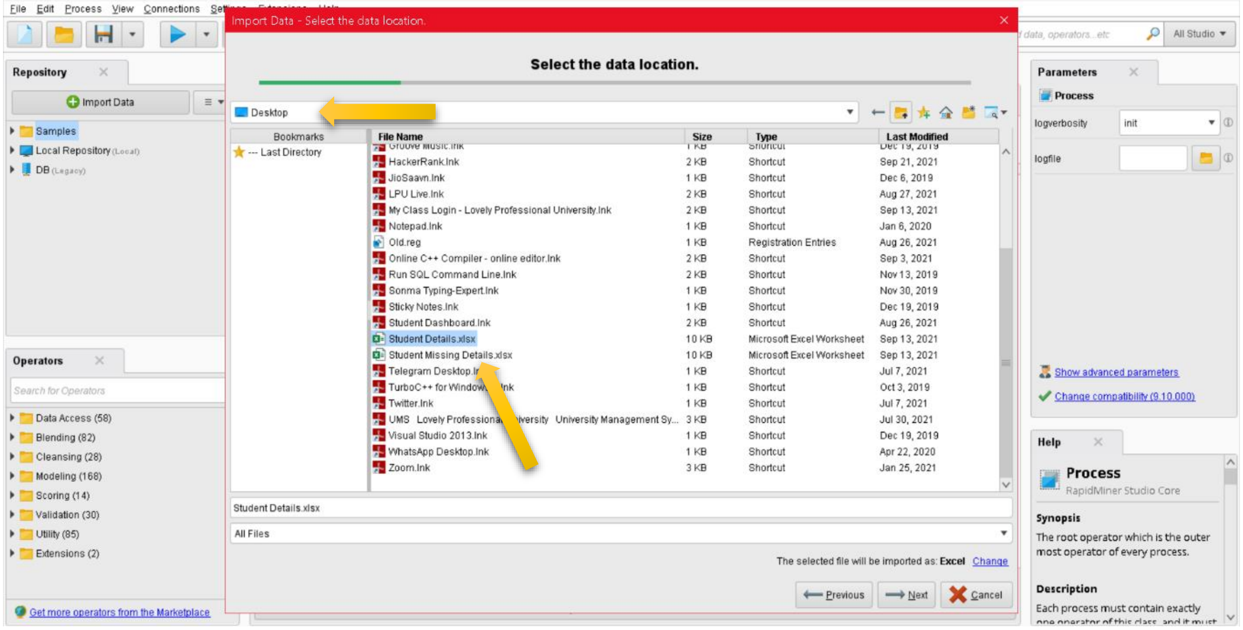


The root operator which is the outer most operator of every process.

Each process must contain exactly

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **01**	Page No. **02** Total Pages. **17**

Step 3: Select your **Data Location**.



Import Data - Select the data location.

Select the data location.

Desktop

Bookmarks: --- Last Directory

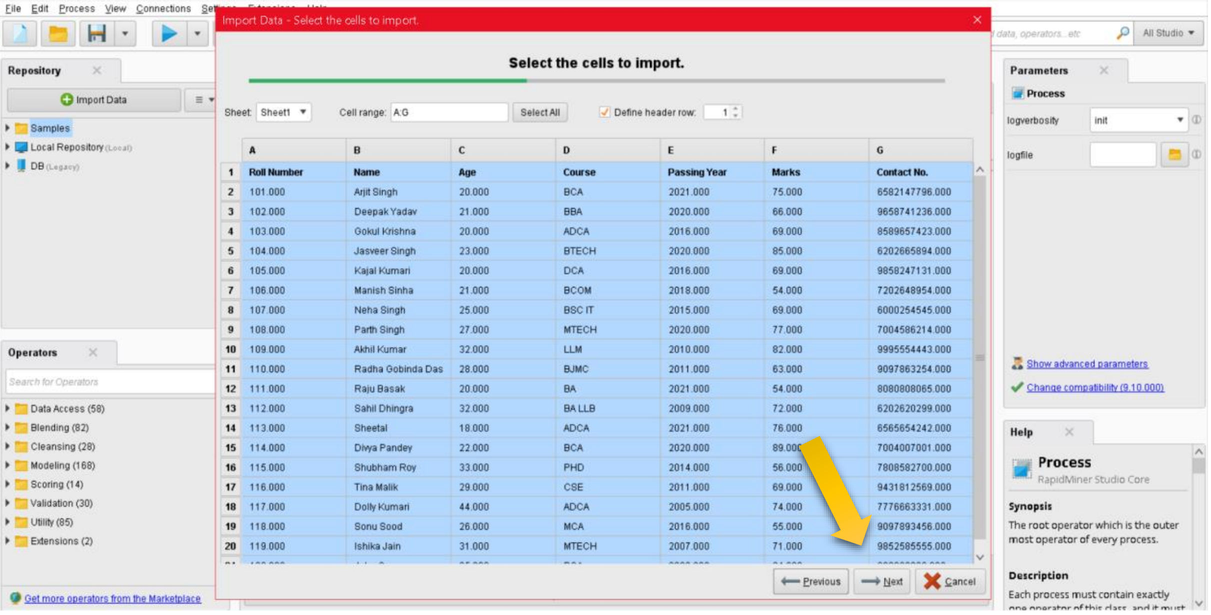
File Name	Size	Type	Last Modified
Groove Music.lnk	1 KB	Shortcut	Dec 19, 2019
HackerRank.lnk	2 KB	Shortcut	Sep 21, 2021
JioSaavn.lnk	1 KB	Shortcut	Dec 6, 2019
LPU Live.lnk	2 KB	Shortcut	Aug 27, 2021
My Class Login - Lovely Professional University.lnk	2 KB	Shortcut	Sep 13, 2021
Notepad.lnk	1 KB	Shortcut	Jan 6, 2020
Old.reg	1 KB	Registration Entries	Aug 26, 2021
Online C++ Compiler - online editor.lnk	2 KB	Shortcut	Sep 3, 2021
Run SQL Command Line.lnk	2 KB	Shortcut	Nov 13, 2019
Sonma Typing-Expert.lnk	1 KB	Shortcut	Nov 30, 2019
Sticky Notes.lnk	1 KB	Shortcut	Dec 19, 2019
Student Dashboard.lnk	2 KB	Shortcut	Aug 26, 2021
Student Details.xlsx	10 KB	Microsoft Excel Worksheet	Sep 13, 2021
Student Missing Details.xlsx	10 KB	Microsoft Excel Worksheet	Sep 13, 2021
Telegram Desktop.lnk	1 KB	Shortcut	Jul 7, 2021
TurboC++ for Windows.lnk	1 KB	Shortcut	Oct 3, 2019
Twitter.lnk	1 KB	Shortcut	Jul 7, 2021
UMS Lovely Professional University University Management Sy...	3 KB	Shortcut	Jul 30, 2021
Visual Studio 2013.lnk	1 KB	Shortcut	Dec 19, 2019
WhatsApp Desktop.lnk	1 KB	Shortcut	Apr 22, 2020
Zoom.lnk	3 KB	Shortcut	Jan 25, 2021

Student Details.xlsx

All Files

The selected file will be imported as: **Excel** Change

Step 4: Now your data set is showing, select the cells that you want to import and then click on **Next**.



Import Data - Select the cells to import.

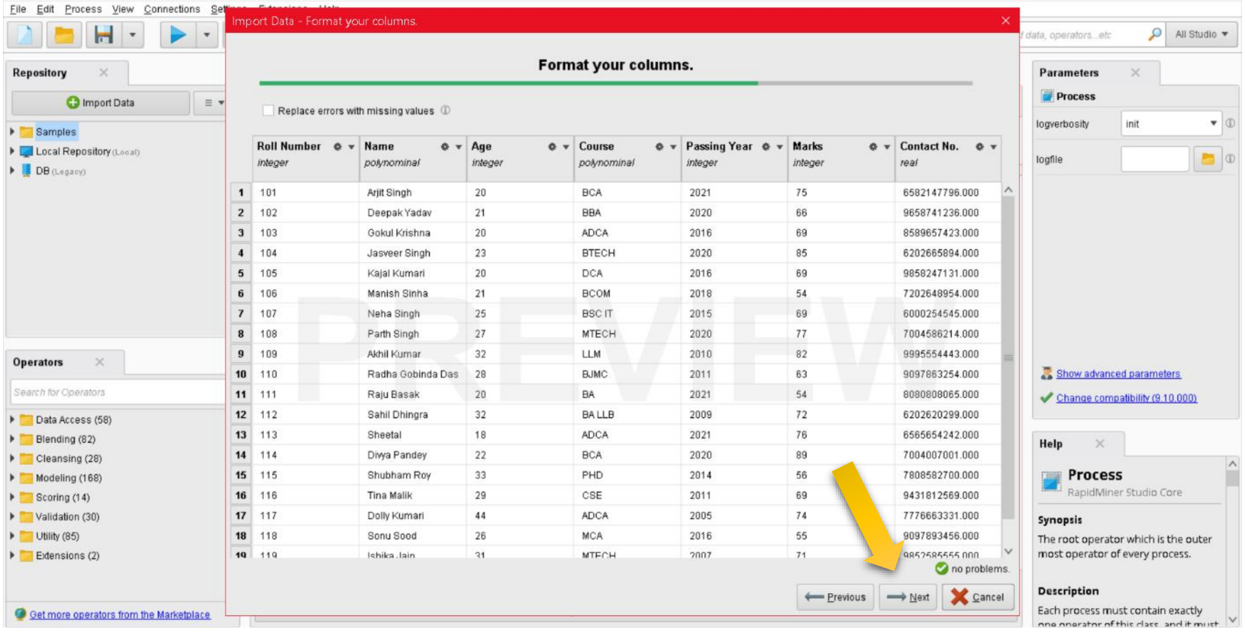
Select the cells to import.

Sheet: Sheet1 Cell range: A:G Select All ☑ Define header row: 1

	A	B	C	D	E	F	G
1	Roll Number	Name	Age	Course	Passing Year	Marks	Contact No.
2	101.000	Arjit Singh	20.000	BCA	2021.000	75.000	6582147796.000
3	102.000	Deepak Yadav	21.000	BBA	2020.000	66.000	9658741236.000
4	103.000	Gokul Krishna	20.000	ADCA	2016.000	69.000	8589657423.000
5	104.000	Jasveer Singh	23.000	BTECH	2020.000	85.000	6202665894.000
6	105.000	Kajal Kumari	20.000	DCA	2016.000	69.000	9858247131.000
7	106.000	Manish Sinha	21.000	BCOM	2018.000	54.000	7202648954.000
8	107.000	Neha Singh	25.000	BSC IT	2015.000	69.000	6000254545.000
9	108.000	Parth Singh	27.000	MTECH	2020.000	77.000	7004586214.000
10	109.000	Akhil Kumar	32.000	LLM	2010.000	82.000	9995554443.000
11	110.000	Radha Gobinda Das	28.000	BJMC	2011.000	63.000	9097863254.000
12	111.000	Raju Basak	20.000	BA	2021.000	54.000	8080808065.000
13	112.000	Sahil Dhingra	32.000	BA LLB	2009.000	72.000	6202620299.000
14	113.000	Sheetal	18.000	ADCA	2021.000	76.000	6565654242.000
15	114.000	Divya Pandey	22.000	BCA	2020.000	89.000	7004007001.000
16	115.000	Shubham Roy	33.000	PHD	2014.000	56.000	7808582700.000
17	116.000	Tina Malik	29.000	CSE	2011.000	69.000	9431812569.000
18	117.000	Dolly Kumari	44.000	ADCA	2005.000	74.000	7776663331.000
19	118.000	Sonu Sood	26.000	MCA	2016.000	55.000	9097893456.000
20	119.000	Ishika Jain	31.000	MTECH	2007.000	71.000	9852585555.000

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **01** Page No. **03**	Total Pages. **17**

Step 5: Your selected data set has been loaded, now click on **Next**.



Import Data - Format your columns.

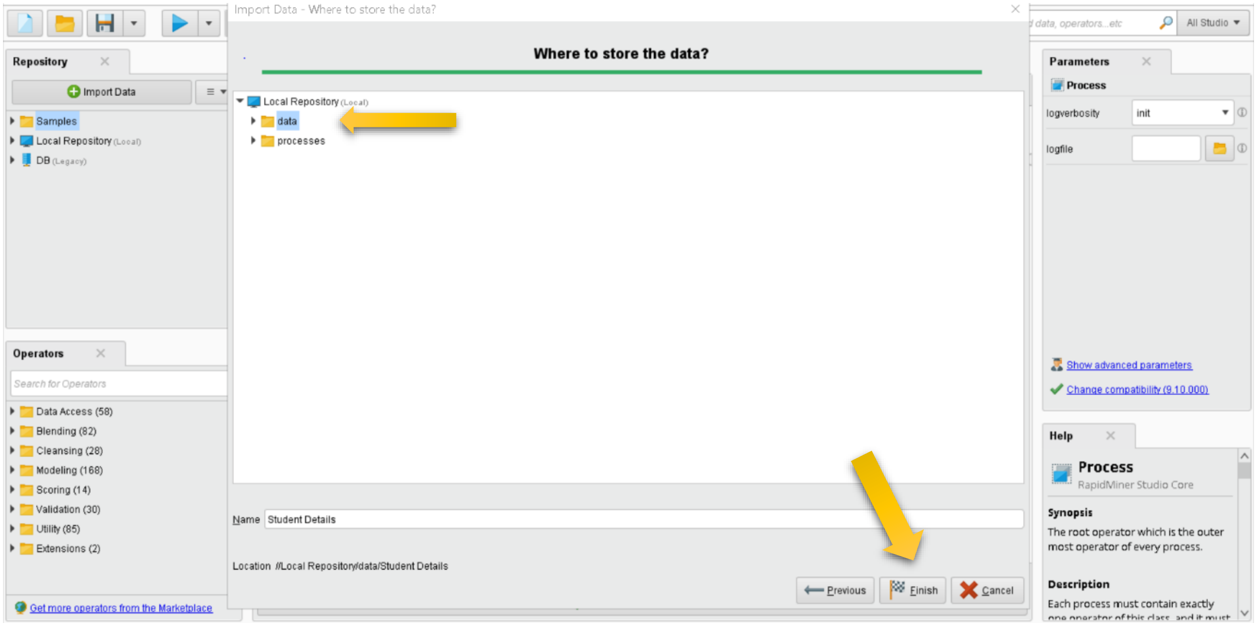
Format your columns.

□ Replace errors with missing values

	Roll Number *integer*	Name *polynominal*	Age *integer*	Course *polynominal*	Passing Year *integer*	Marks *integer*	Contact No. *real*
1	101	Arjit Singh	20	BCA	2021	75	6582147796.000
2	102	Deepak Yadav	21	BBA	2020	66	9658741236.000
3	103	Gokul Krishna	20	ADCA	2016	69	8589657423.000
4	104	Jasveer Singh	23	BTECH	2020	85	6202665894.000
5	105	Kajal Kumari	20	DCA	2016	69	9858247131.000
6	106	Manish Sinha	21	BCOM	2018	54	7202648954.000
7	107	Neha Singh	25	BSC IT	2015	69	6000254545.000
8	108	Parth Singh	27	MTECH	2020	77	7004586214.000
9	109	Akhil Kumar	32	LLM	2010	82	9995554443.000
10	110	Radha Gobinda Das	28	BJMC	2011	63	9097863254.000
11	111	Raju Basak	20	BA	2021	54	8080808065.000
12	112	Sahil Dhingra	32	BA LLB	2009	72	6202620299.000
13	113	Sheetal	18	ADCA	2021	76	6565654242.000
14	114	Divya Pandey	22	BCA	2020	89	7004007001.000
15	115	Shubham Roy	33	PHD	2014	56	7808582700.000
16	116	Tina Malik	29	CSE	2011	69	9431812569.000
17	117	Dolly Kumari	44	ADCA	2005	74	7776663331.000
18	118	Sonu Sood	26	MCA	2016	55	9097893456.000
19	119	Ishika Jain	31	MTECH	2007	71	9852585555.000

✓ no problems.

Step 6: Now select where you want to store your data set, you have two options **data/processes,** here I selected **data** and then click on **Finish**.



Import Data - Where to store the data?

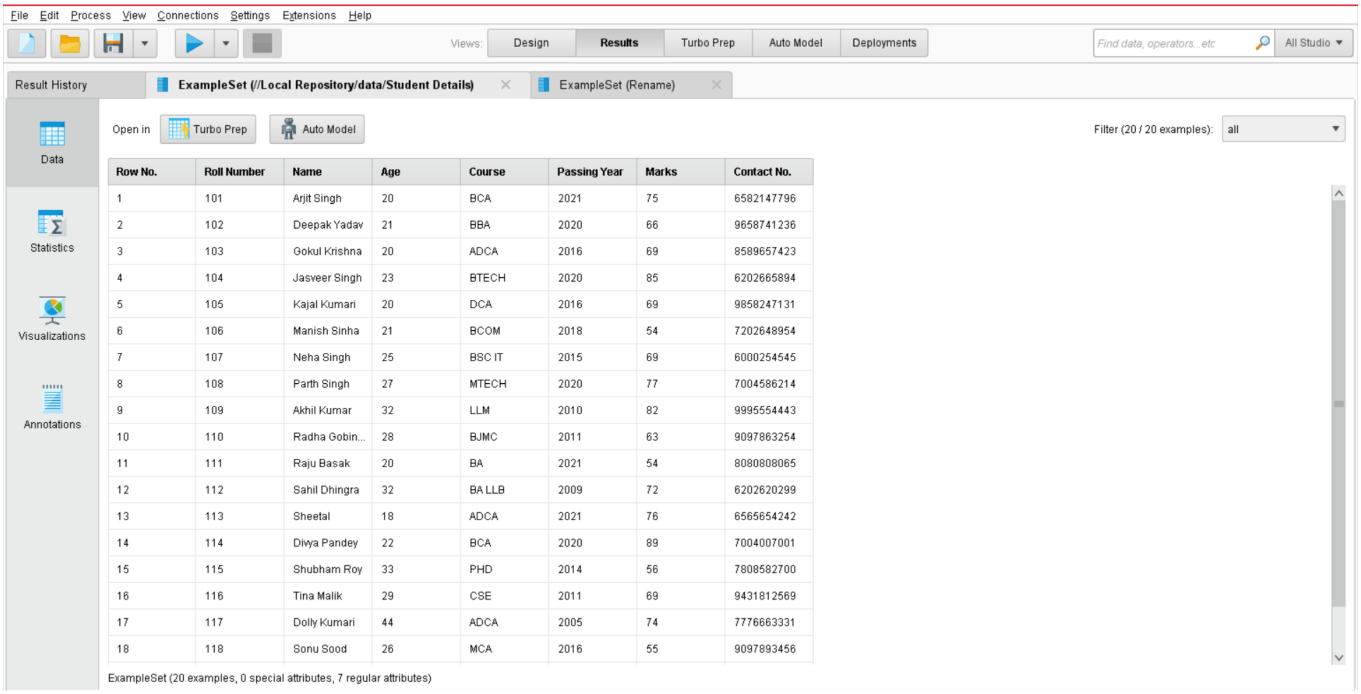
Where to store the data?

Name Student Details

Location //Local Repository/data/Student Details

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**	
Course Instructor: **Dr. Geeta Sharma**		
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**	
Question No. **01**	Page No. **04**	Total Pages. **17**

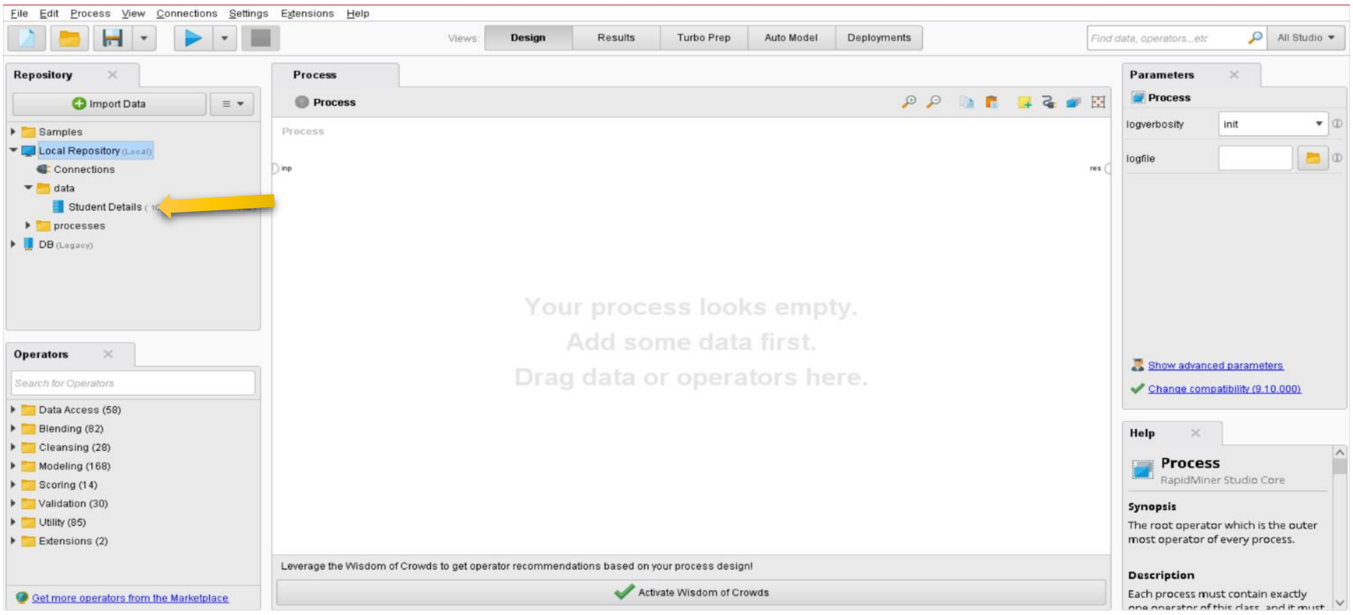
Step 7: Now you can see your data set has been **Imported Successfully.**



Row No.	Roll Number	Name	Age	Course	Passing Year	Marks	Contact No.
1	101	Arjit Singh	20	BCA	2021	75	6582147796
2	102	Deepak Yadav	21	BBA	2020	66	9658741236
3	103	Gokul Krishna	20	ADCA	2016	69	8589657423
4	104	Jasveer Singh	23	BTECH	2020	85	6202665894
5	105	Kajal Kumari	20	DCA	2016	69	9858247131
6	106	Manish Sinha	21	BCOM	2018	54	7202648954
7	107	Neha Singh	25	BSC IT	2015	69	6000254545
8	108	Parth Singh	27	MTECH	2020	77	7004586214
9	109	Akhil Kumar	32	LLM	2010	82	9995554443
10	110	Radha Gobin...	28	BJMC	2011	63	9097863254
11	111	Raju Basak	20	BA	2021	54	8080808065
12	112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
13	113	Sheetal	18	ADCA	2021	76	6565654242
14	114	Divya Pandey	22	BCA	2020	89	7004007001
15	115	Shubham Roy	33	PHD	2014	56	7808582700
16	116	Tina Malik	29	CSE	2011	69	9431812569
17	117	Dolly Kumari	44	ADCA	2005	74	7776663331
18	118	Sonu Sood	26	MCA	2016	55	9097893456

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

You can also see that your data set has been stored in the data section of the local repository as well.



logverbosity: init

logfile:

Your process looks empty.
Add some data first.
Drag data or operators here.

Leverage the Wisdom of Crowds to get operator recommendations based on your process design!

Process
RapidMiner Studio Core

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

Description
Each process must contain exactly one operator of this class, and it must

Q2. Perform following Transformation operations on your data you imported in Q1:

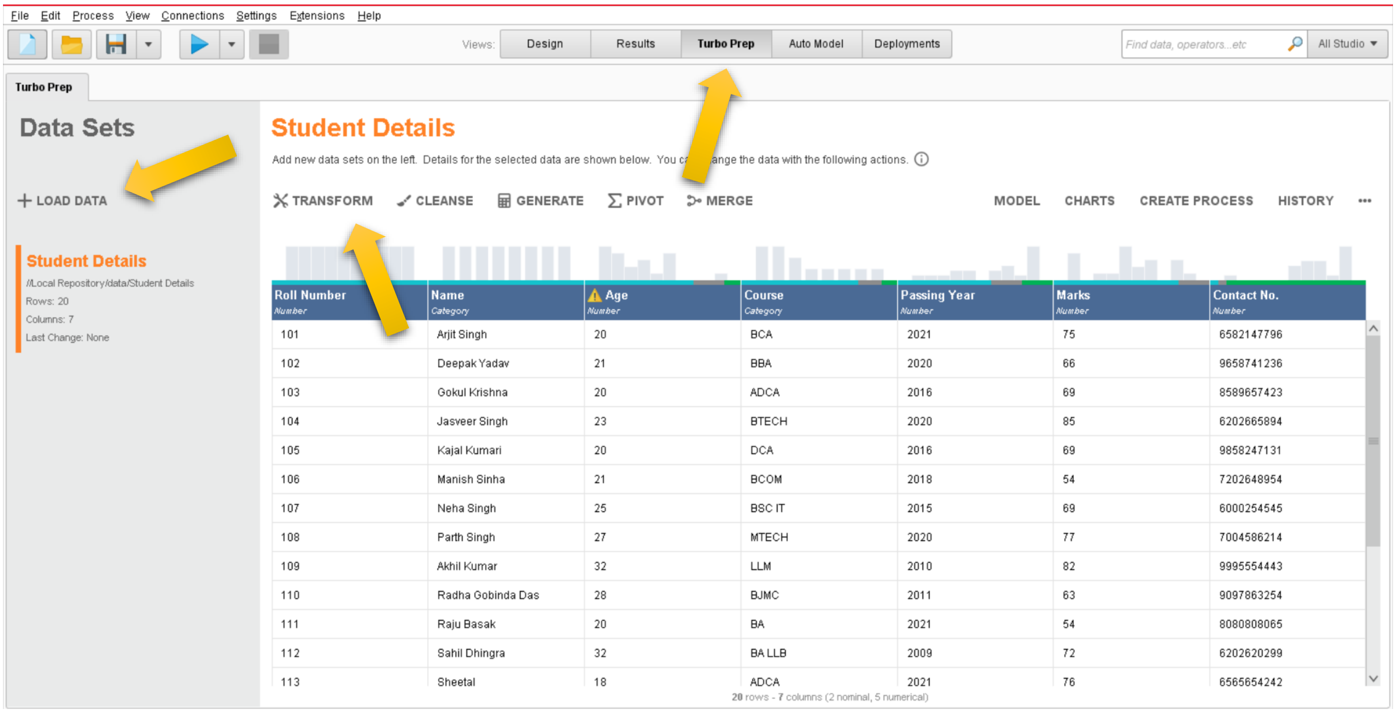
a) **Sorting**

b) **Filter numerical data**

c) **Filter String data**

d) **Remove attribute**

Answer:_ For performing all the following transformation in the Imported data set first we have to click on **Turbo Prep** and then click on **Load Data** and select the imported data set. (Imported data sets are in the local repository.) Now click on **TRANSFORM**.



File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Turbo Prep
Data Sets
+ LOAD DATA
Student Details
//Local Repository/data/Student Details
Rows: 20
Columns: 7
Last Change: None
TRANSFORM
CLEANSE
GENERATE
PIVOT
MERGE
MODEL
CHARTS
CREATE PROCESS
HISTORY

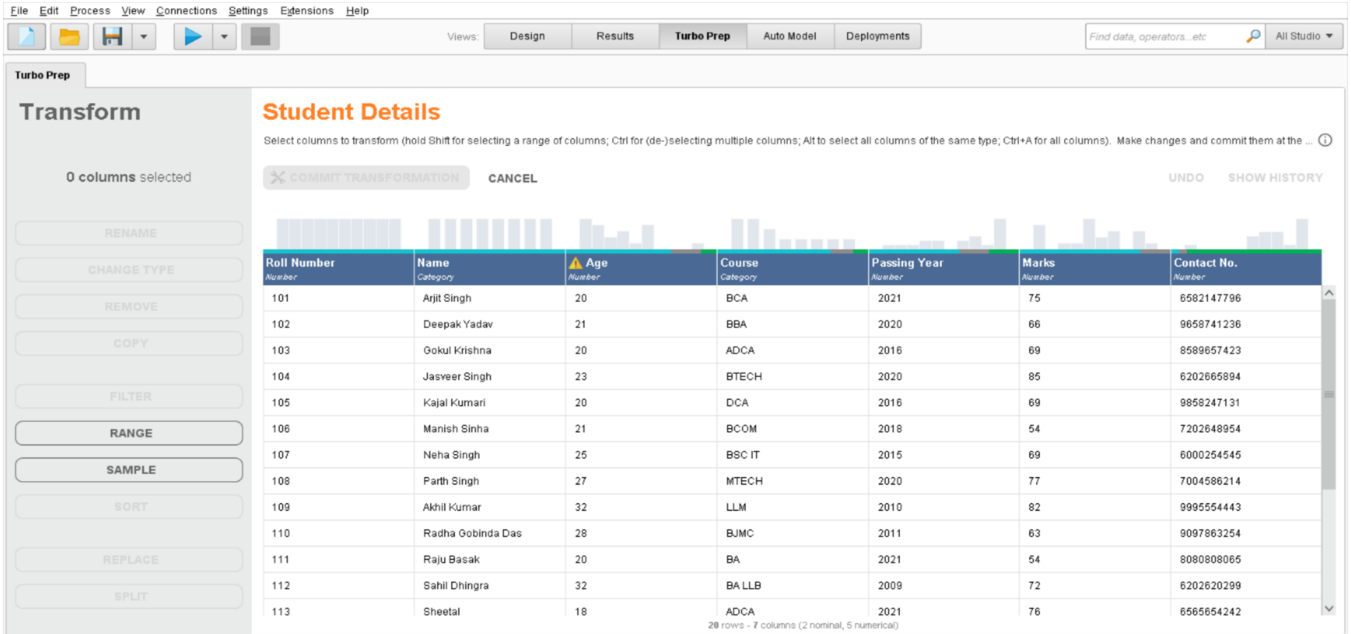
Student Details

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions.

Roll Number Number	Name Category	⚠ Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

This interface will come up after clicking on **TRANSFORM**.



Student Details

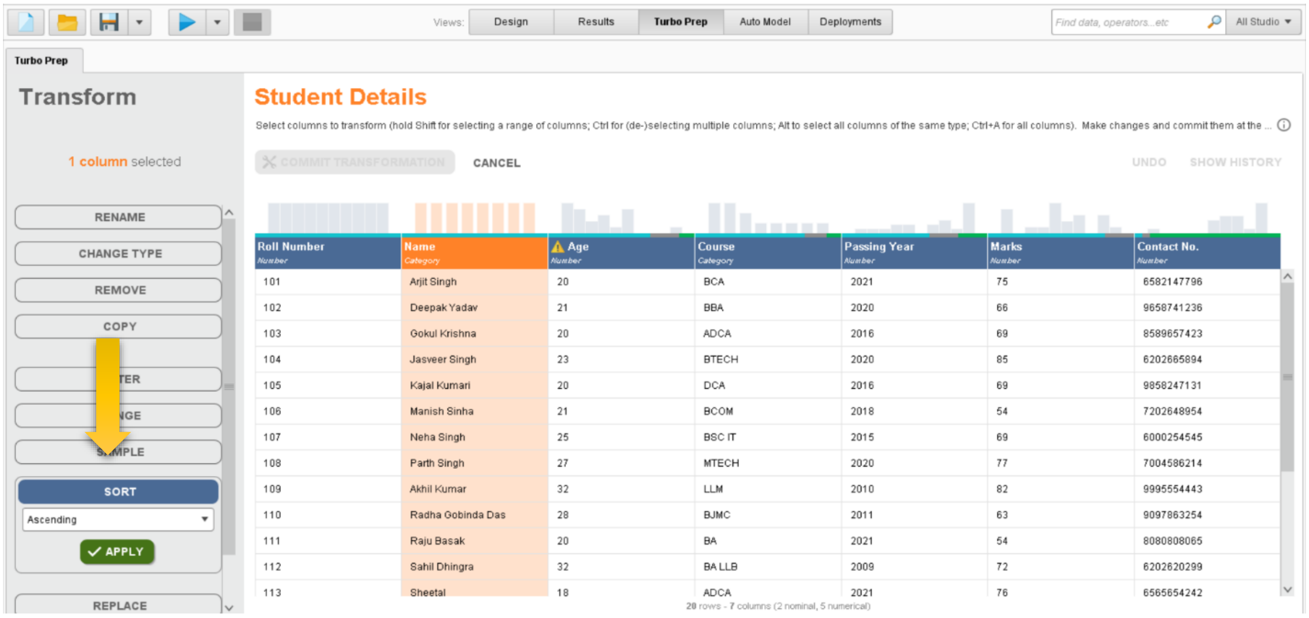
Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

a) Sorting

Step 1: First you have to select the attribute where you want to use the sorting then click on **Sort**. Now we have two options sorted in **Ascending/Descending.** Here I selected **Ascending.**



Ascending

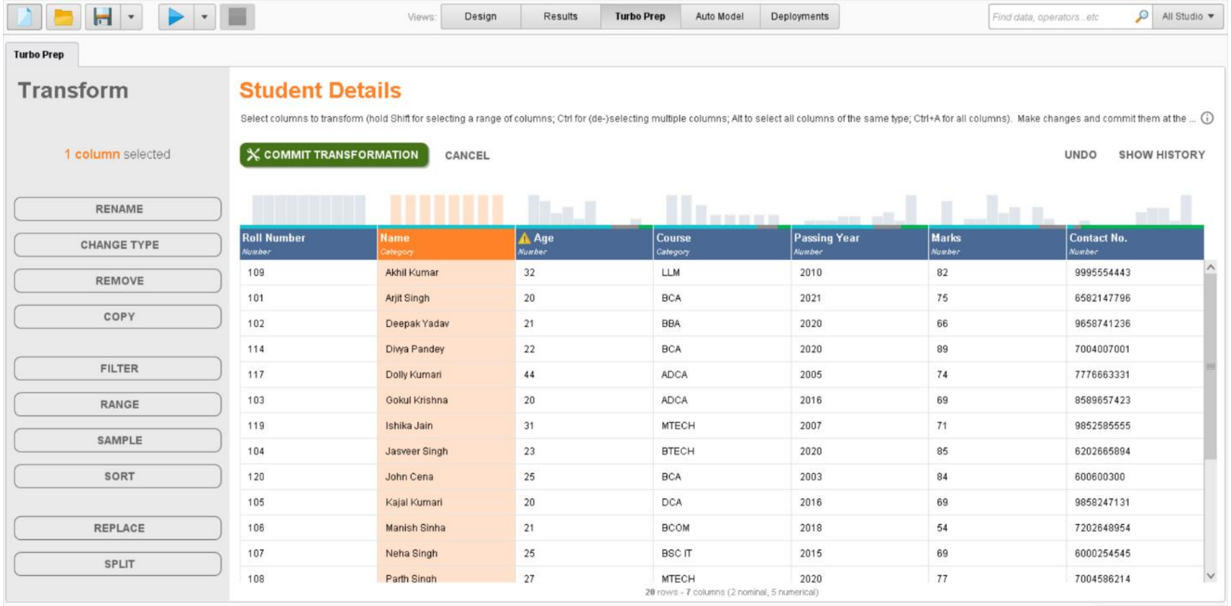
Student Details

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

Step 2: Now you can see the name attribute has been sorted in ascending order.



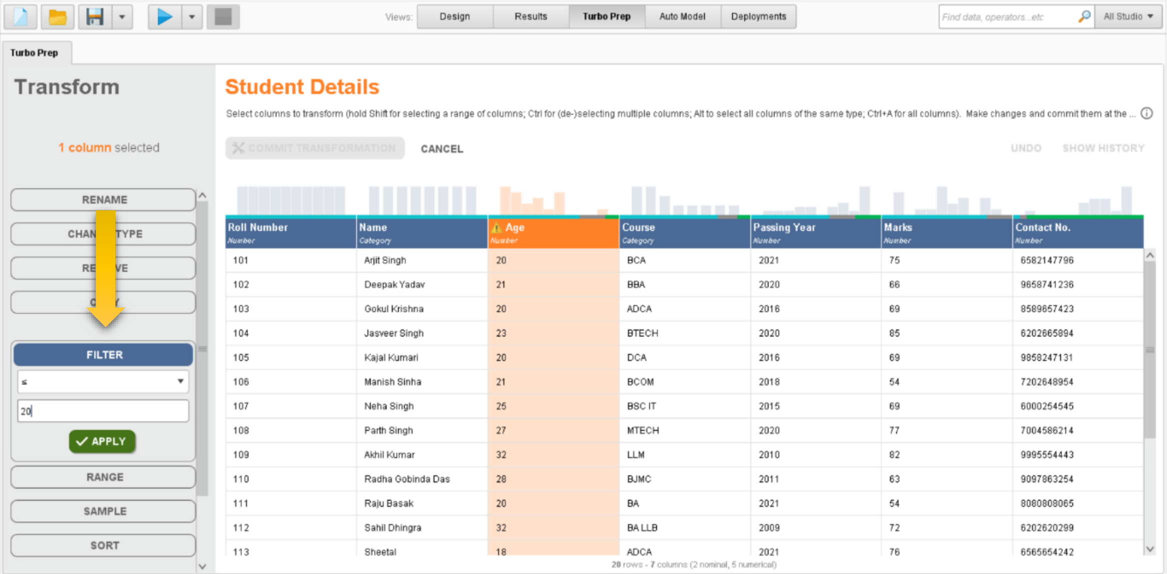
Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
109	Akhil Kumar	32	LLM	2010	82	9995554443
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
114	Divya Pandey	22	BCA	2020	89	7004007001
117	Dolly Kumari	44	ADCA	2005	74	7776663331
103	Gokul Krishna	20	ADCA	2016	69	8589657423
119	Ishika Jain	31	MTECH	2007	71	9852585555
104	Jasveer Singh	23	BTECH	2020	85	6202665894
120	John Cena	25	BCA	2003	84	600600300
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214

20 rows - 7 columns (2 nominal, 5 numerical)

b) Filter Numeric Data

Step 1: First we have to select numeric data attribute, here I selected **Age**. Now click on **Filter** and select filter operator like **=, >, <, ≠, ≤, ≥, is missing, is not missing.** Here I selected **≤20.**

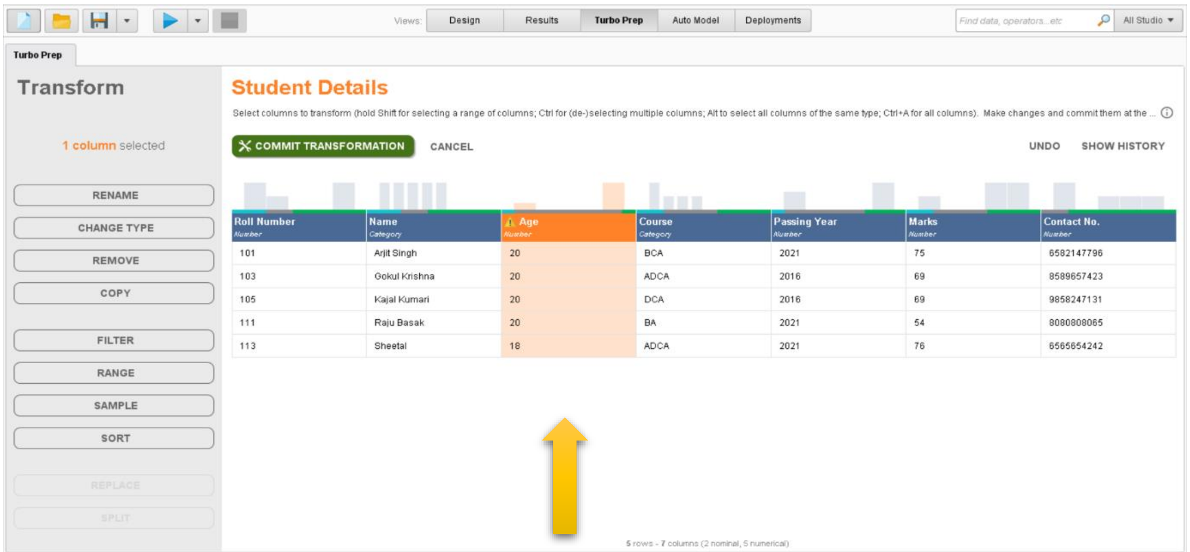


Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

Step 2: Here we can see our Age attribute is **Filtered** as per the given operation. **(≤20)**



Student Details

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

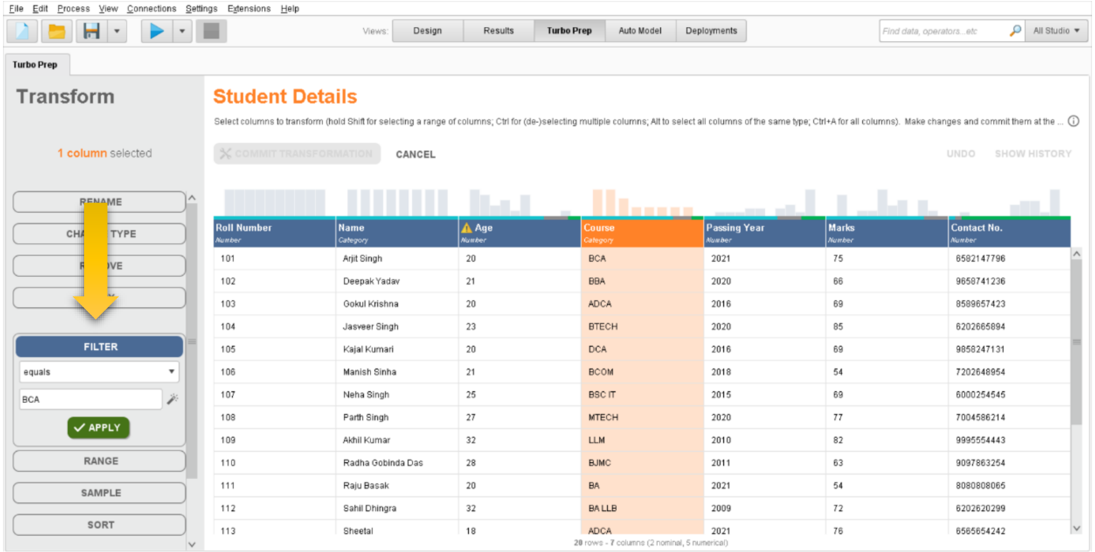
COMMIT TRANSFORMATION CANCEL UNDO SHOW HISTORY

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
103	Gokul Krishna	20	ADCA	2016	69	8589657423
105	Kajal Kumari	20	DCA	2016	69	9858247131
111	Raju Basak	20	BA	2021	54	8080808065
113	Sheetal	18	ADCA	2021	76	6565654242

5 rows - 7 columns (2 nominal, 5 numerical)

c) Filter String Data

Step 1: First we have to select string data attribute, here I selected **Course**. Now click on **Filter** and select filter operator like **equals, does not equal, is in, is not in, contains, does not contain, start with etc.** Here I selected **start with (BCA).**


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Turbo Prep
Transform
1 column selected
RENAME
CHANGE TYPE
REMOVE
FILTER
APPLY
RANGE
SAMPLE
SORT

equals

BCA

Student Details

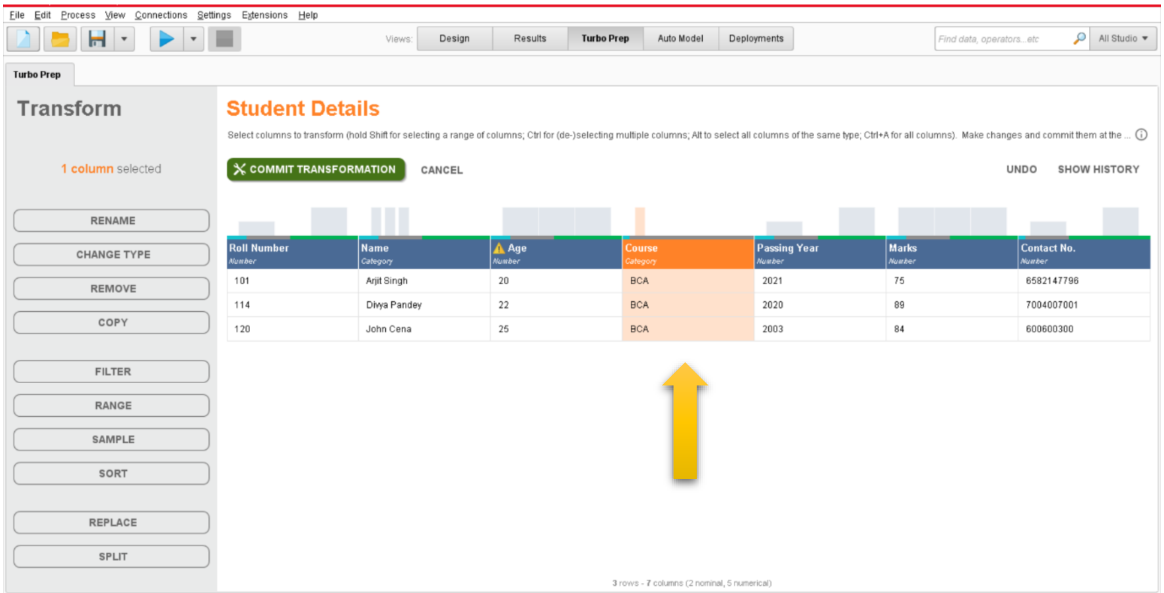
Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

COMMIT TRANSFORMATION CANCEL UNDO SHOW HISTORY

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

Step 2: Here we can see our Course attribute is **Filtered** as per the given operation. (**start with BCA)**



Student Details

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

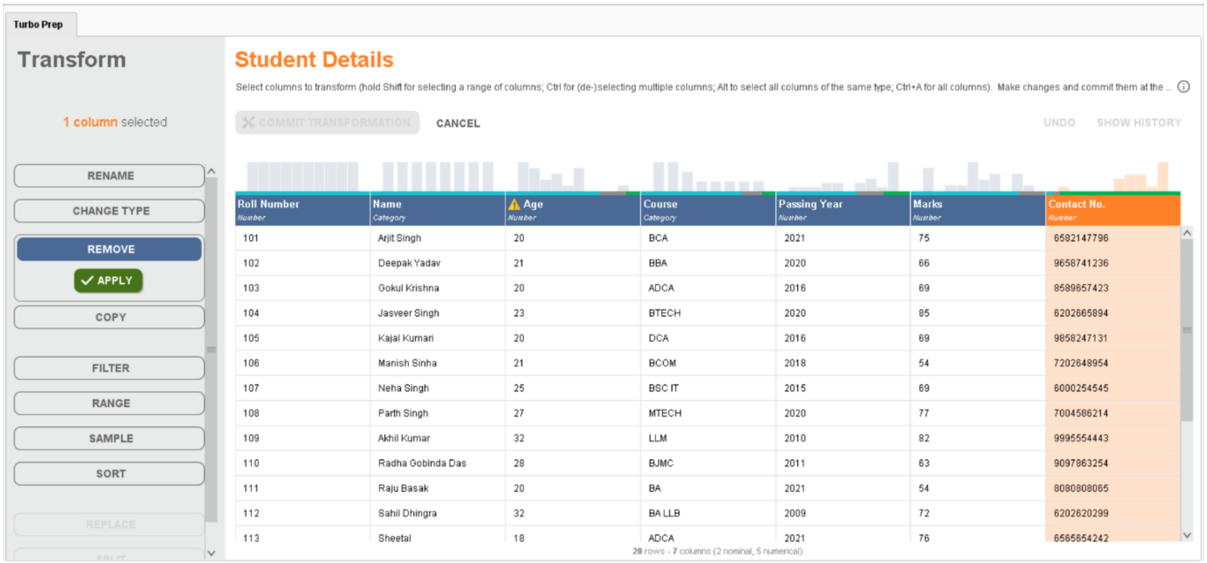
Roll Number Number	Name Category	⚠ Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
114	Divya Pandey	22	BCA	2020	89	7004007001
120	John Cena	25	BCA	2003	84	600600300

3 rows - 7 columns (2 nominal, 5 numerical)

d) Remove Attribute

Step 1: First we have to select the attribute which we want to remove from the data set. Here I selected **Contact No.** attribute.

Now click on **Remove** and then click on **Apply.**



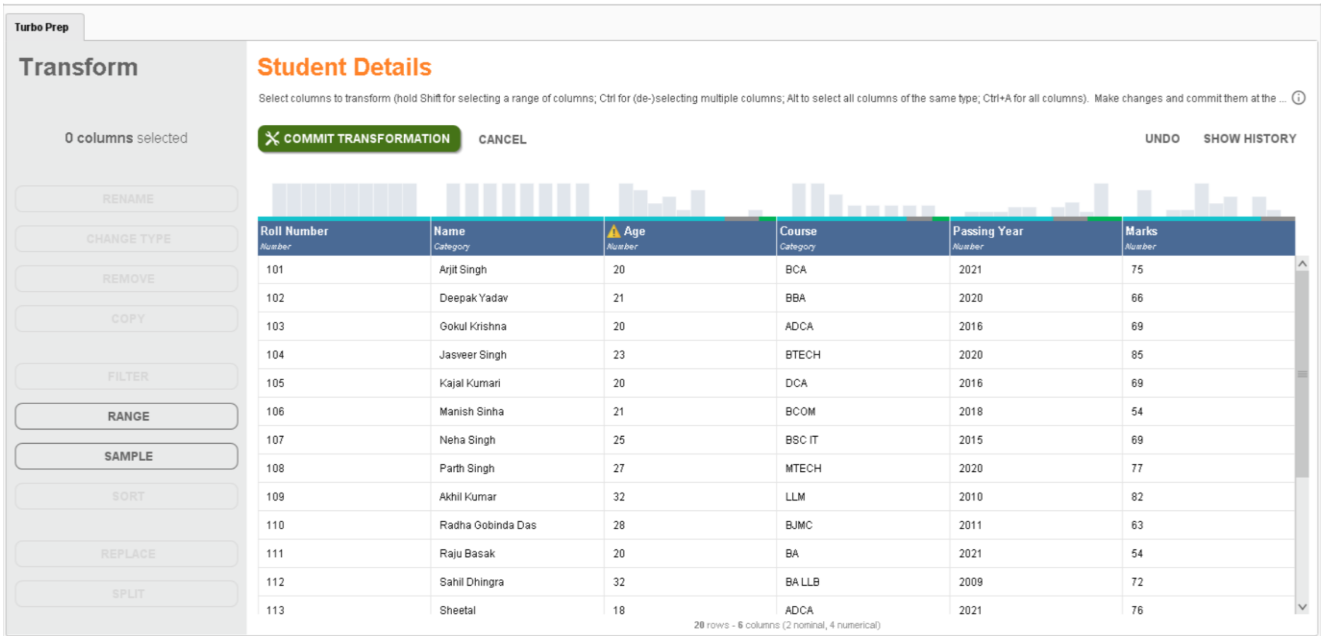
Student Details

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number Number	Name Category	⚠ Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

Step 2: Here we can see Contact No. attribute has been **Removed**.



Student Details

Select columns to transform (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the ...

Roll Number *Number*	Name *Category*	⚠ Age *Number*	Course *Category*	Passing Year *Number*	Marks *Number*
101	Arjit Singh	20	BCA	2021	75
102	Deepak Yadav	21	BBA	2020	66
103	Gokul Krishna	20	ADCA	2016	69
104	Jasveer Singh	23	BTECH	2020	85
105	Kajal Kumari	20	DCA	2016	69
106	Manish Sinha	21	BCOM	2018	54
107	Neha Singh	25	BSC IT	2015	69
108	Parth Singh	27	MTECH	2020	77
109	Akhil Kumar	32	LLM	2010	82
110	Radha Gobinda Das	28	BJMC	2011	63
111	Raju Basak	20	BA	2021	54
112	Sahil Dhingra	32	BA LLB	2009	72
113	Sheetal	18	ADCA	2021	76

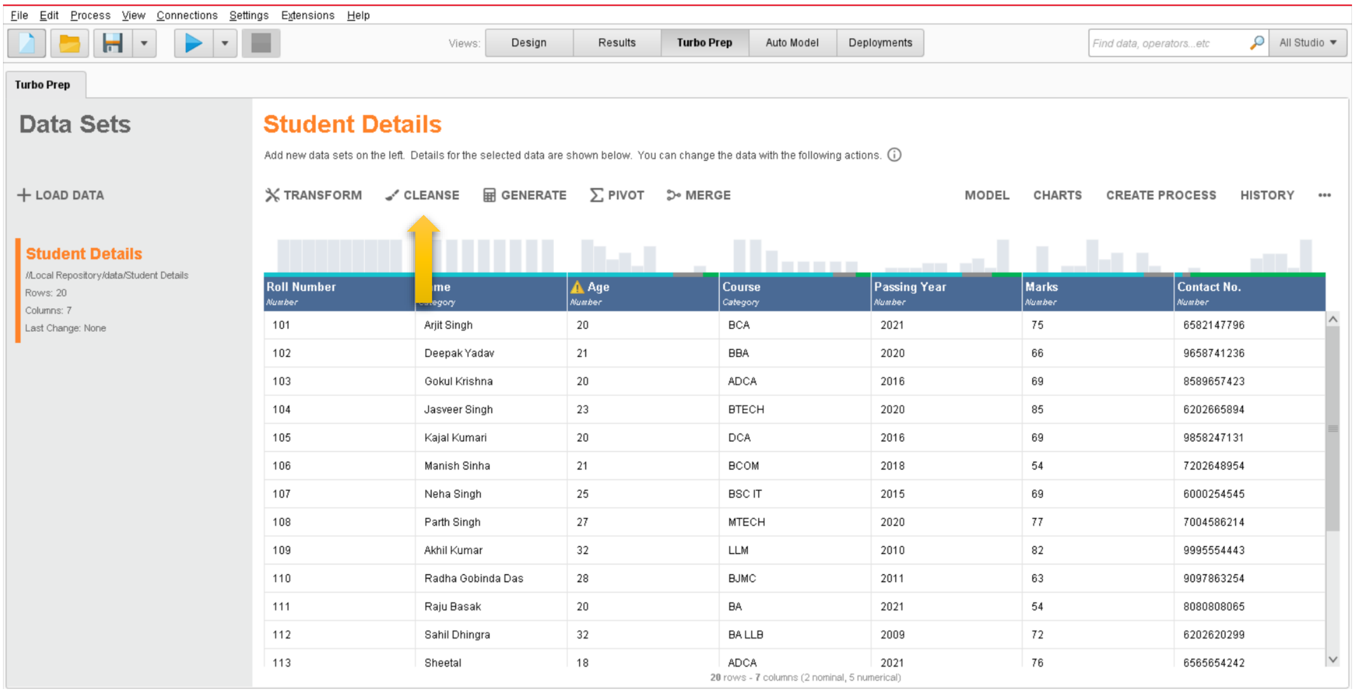
20 rows - 6 columns (2 nominal, 4 numerical)

Q3. Perform following Cleanse operations on your data.

a) Auto Cleanse
b) Normalisation
c) Discretization

Answer: For performing all the following transformation in the Imported data set first we have to click on **Turbo Prep** and then click on **Load Data** and select the imported data set. (Imported data sets are in the local repository.) Now click on **CLEANSE**.

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**	
Course Instructor: **Dr. Geeta Sharma**		
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**	
Question No. **03**	Page No. **11**	Total Pages. **17**


File Edit Process View Connections Settings Extensions Help
Views: Design Results Turbo Prep Auto Model Deployments
Find data, operators...etc
All Studio
Turbo Prep
TRANSFORM CLEANSE GENERATE PIVOT MERGE
MODEL CHARTS CREATE PROCESS HISTORY

Data Sets

+ LOAD DATA

Student Details
//Local Repository/data/Student Details
Rows: 20
Columns: 7
Last Change: None

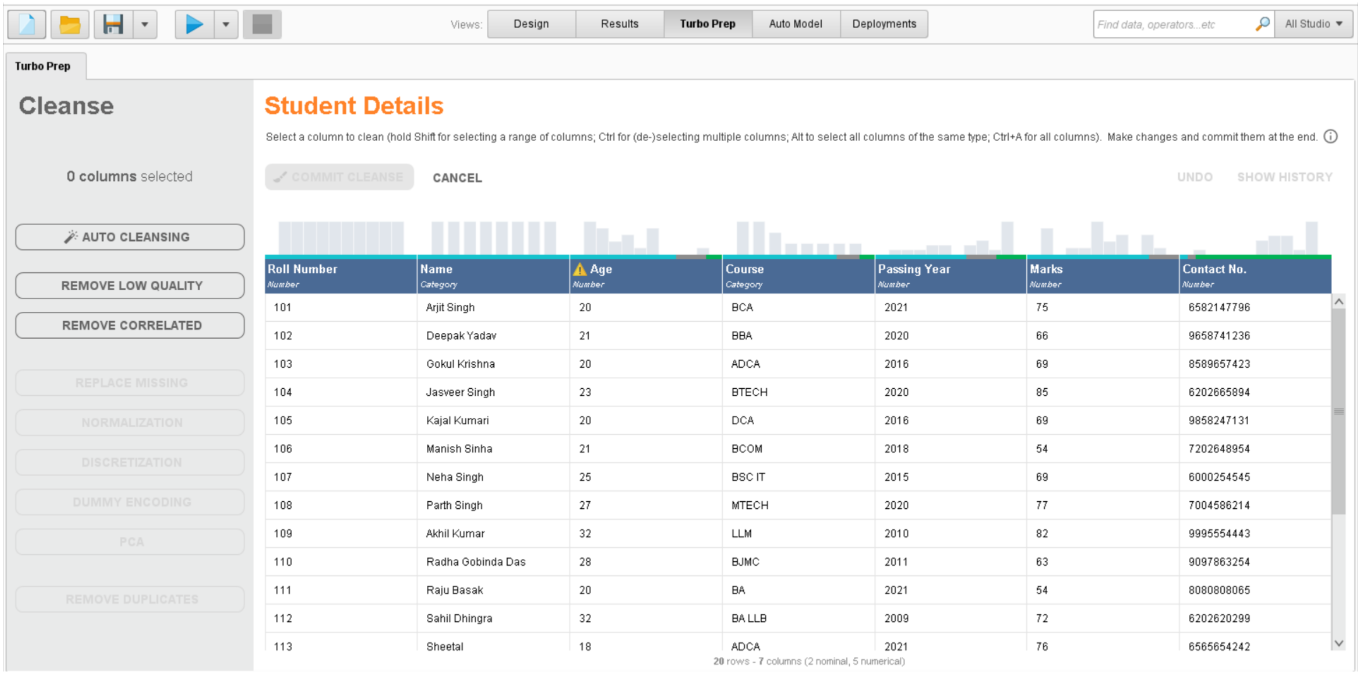
Student Details

Add new data sets on the left. Details for the selected data are shown below. You can change the data with the following actions.

Roll Number Number	me Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

This interface will come up after clicking on **CLEANSE.**



Cleanse

0 columns selected

Student Details

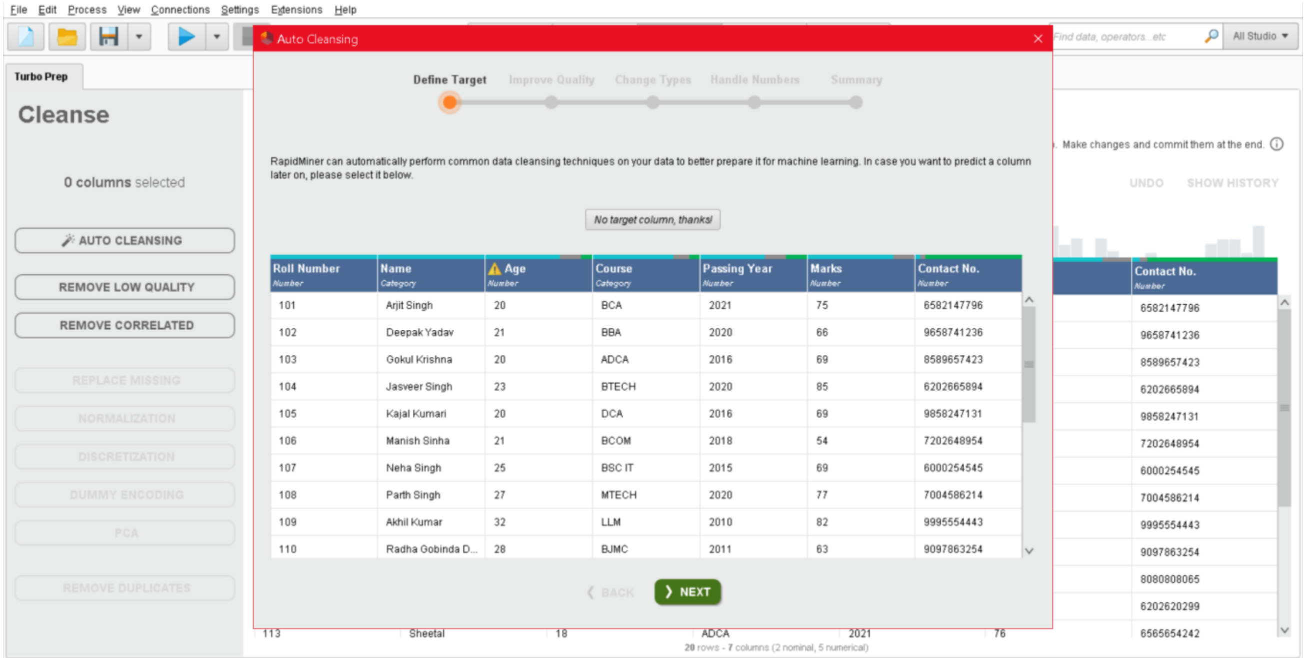
Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the end.

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda Das	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065
112	Sahil Dhingra	32	BA LLB	2009	72	6202620299
113	Sheetal	18	ADCA	2021	76	6565654242

20 rows - 7 columns (2 nominal, 5 numerical)

a) Auto Cleanse

Step 1: Click on **AUTO CLEANSING.**



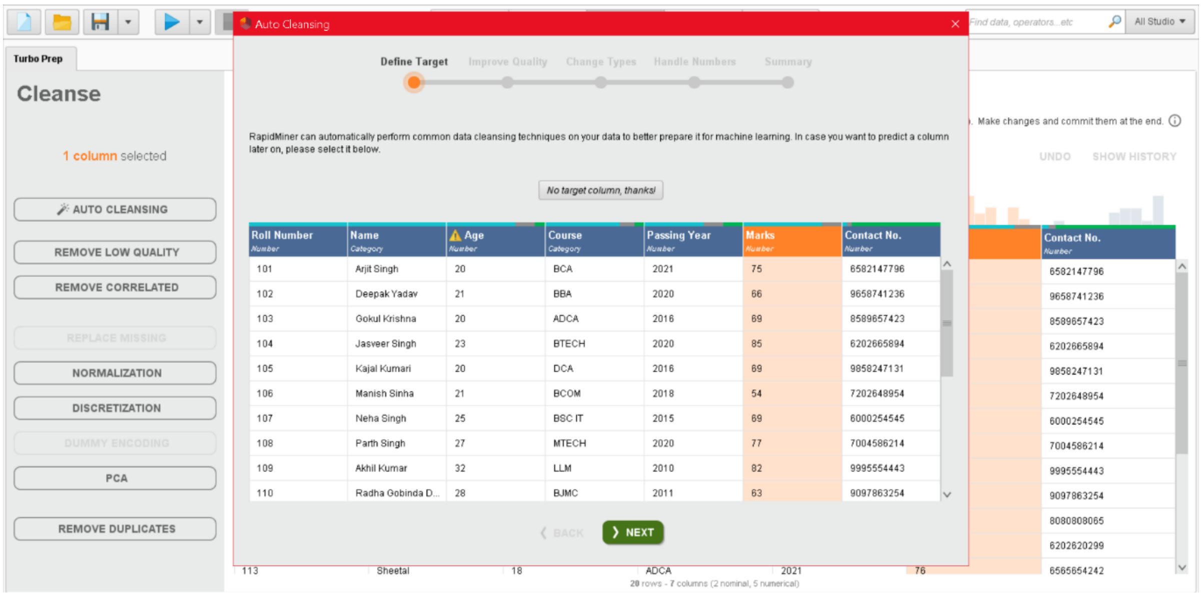
RapidMiner can automatically perform common data cleansing techniques on your data to better prepare it for machine learning. In case you want to predict a column later on, please select it below.

Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda D...	28	BJMC	2011	63	9097863254

113 Sheetal 18 ADCA 2021 76

20 rows - 7 columns (2 nominal, 5 numerical)

Step 2: Select the attribute, here I selected **Marks.**



RapidMiner can automatically perform common data cleansing techniques on your data to better prepare it for machine learning. In case you want to predict a column later on, please select it below.

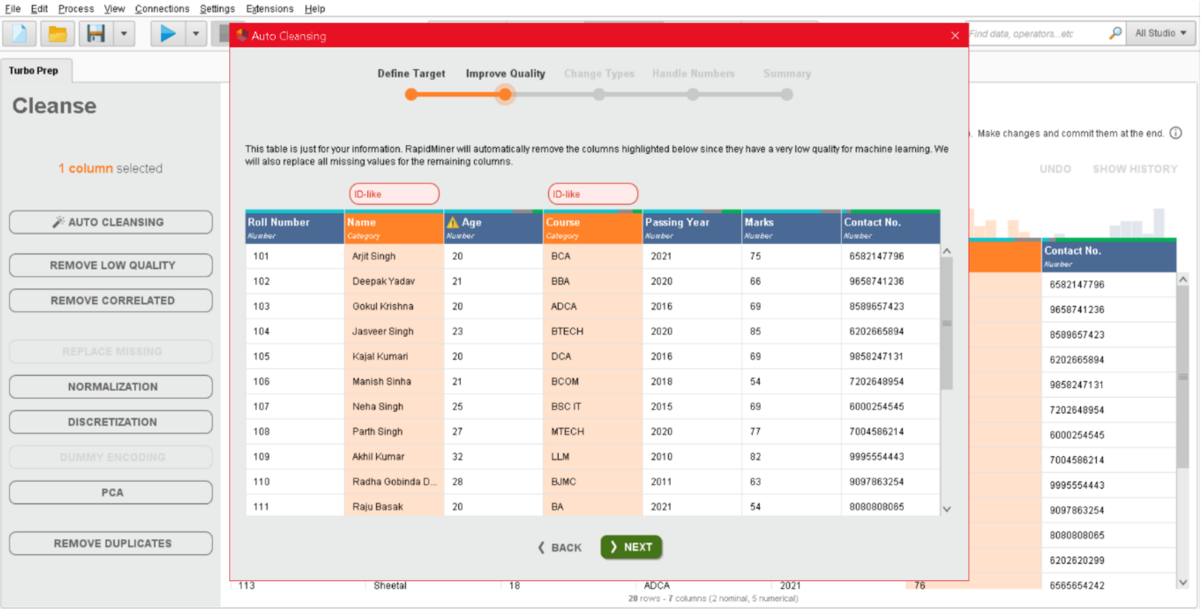
Roll Number Number	Name Category	Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda D...	28	BJMC	2011	63	9097863254

113 Sheetal 18 ADCA 2021 76

20 rows - 7 columns (2 nominal, 5 numerical)

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **03** Page No. **13**	Total Pages. **17**

Step 3: Rapid miner automatically select some attributes whose have very **low quality** and **remove** them at the last.

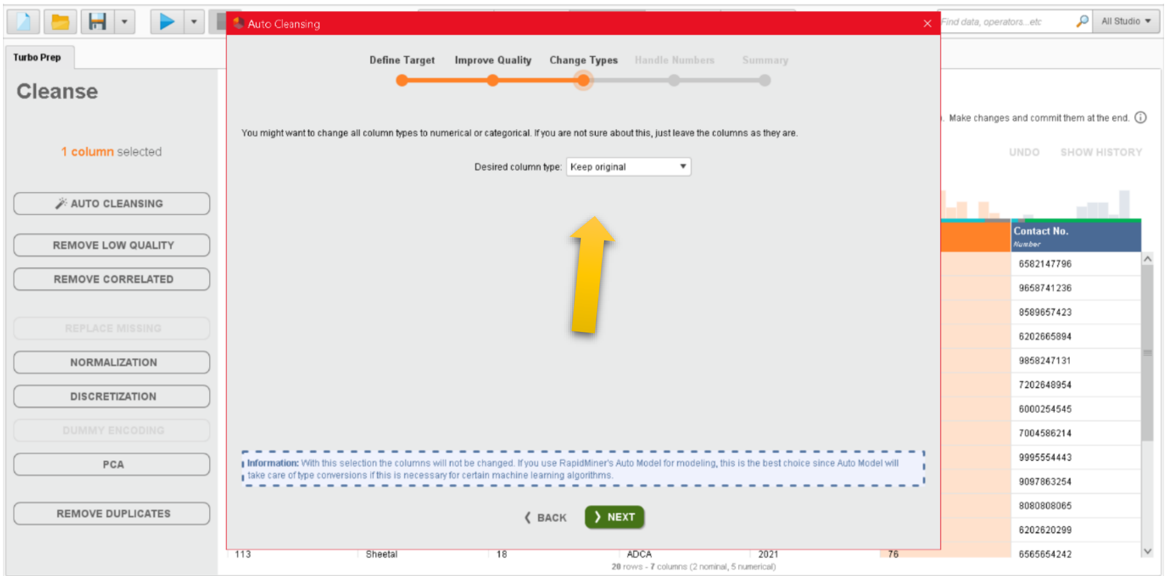


This table is just for your information. RapidMiner will automatically remove the columns highlighted below since they have a very low quality for machine learning. We will also replace all missing values for the remaining columns.

Roll Number Number	Name Category	⚠ Age Number	Course Category	Passing Year Number	Marks Number	Contact No. Number
101	Arjit Singh	20	BCA	2021	75	6582147796
102	Deepak Yadav	21	BBA	2020	66	9658741236
103	Gokul Krishna	20	ADCA	2016	69	8589657423
104	Jasveer Singh	23	BTECH	2020	85	6202665894
105	Kajal Kumari	20	DCA	2016	69	9858247131
106	Manish Sinha	21	BCOM	2018	54	7202648954
107	Neha Singh	25	BSC IT	2015	69	6000254545
108	Parth Singh	27	MTECH	2020	77	7004586214
109	Akhil Kumar	32	LLM	2010	82	9995554443
110	Radha Gobinda D...	28	BJMC	2011	63	9097863254
111	Raju Basak	20	BA	2021	54	8080808065

20 rows - 7 columns (2 nominal, 5 numerical)

Step 4: If you want to change the type of selected attribute then Select change otherwise select keep original. Here I selected **Keep Original.**



You might want to change all column types to numerical or categorical. If you are not sure about this, just leave the columns as they are.

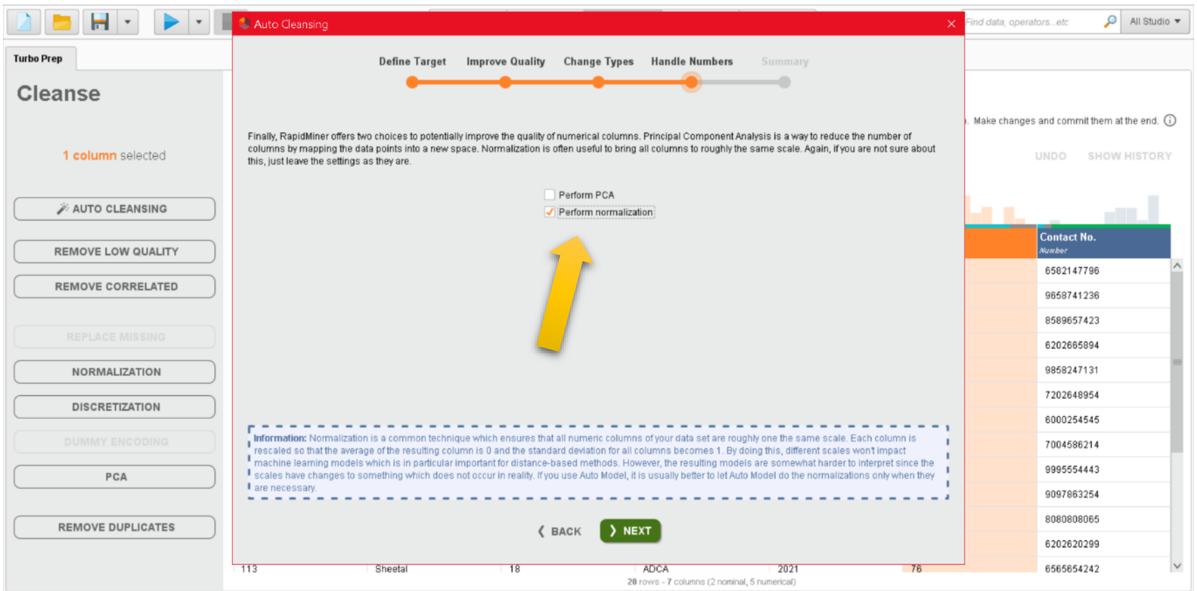
Desired column type: Keep original

Information: With this selection the columns will not be changed. If you use RapidMiner's Auto Model for modeling, this is the best choice since Auto Model will take care of type conversions if this is necessary for certain machine learning algorithms.

20 rows - 7 columns (2 nominal, 5 numerical)

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **03** Page No. **14**	Total Pages. **17**

Step 5: Now we have two options **Perform PCA/Perform normalization.** Here I selected **Perform normalization.**



Finally, RapidMiner offers two choices to potentially improve the quality of numerical columns. Principal Component Analysis is a way to reduce the number of columns by mapping the data points into a new space. Normalization is often useful to bring all columns to roughly the same scale. Again, if you are not sure about this, just leave the settings as they are.

□ Perform PCA
☑ Perform normalization

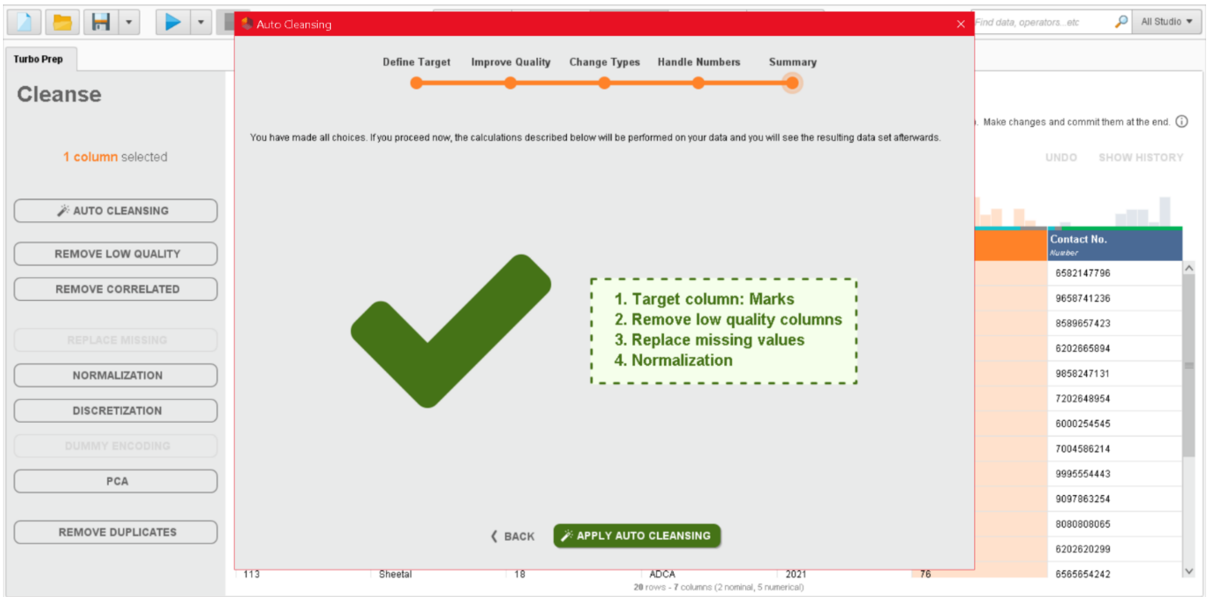
Information: Normalization is a common technique which ensures that all numeric columns of your data set are roughly one the same scale. Each column is rescaled so that the average of the resulting column is 0 and the standard deviation for all columns becomes 1. By doing this, different scales won't impact machine learning models which is in particular important for distance-based methods. However, the resulting models are somewhat harder to interpret since the scales have changes to something which does not occur in reality. If you use Auto Model, it is usually better to let Auto Model do the normalizations only when they are necessary.

Contact No. Number
6582147796
9658741236
8589657423
6202665894
9858247131
7202648954
6000254545
7004586214
9995554443
9097863254
8080808065
6202620299
6565654242

113	Sheetal	18	ADCA	2021	76

20 rows - 7 columns (2 nominal, 5 numerical)

Step 6: We can see a green tick sign means all the operations have been completed. Now click on **APPLY AUTO CLEANSING.**



You have made all choices. If you proceed now, the calculations described below will be performed on your data and you will see the resulting data set afterwards.

1. Target column: Marks
2. Remove low quality columns
3. Replace missing values
4. Normalization

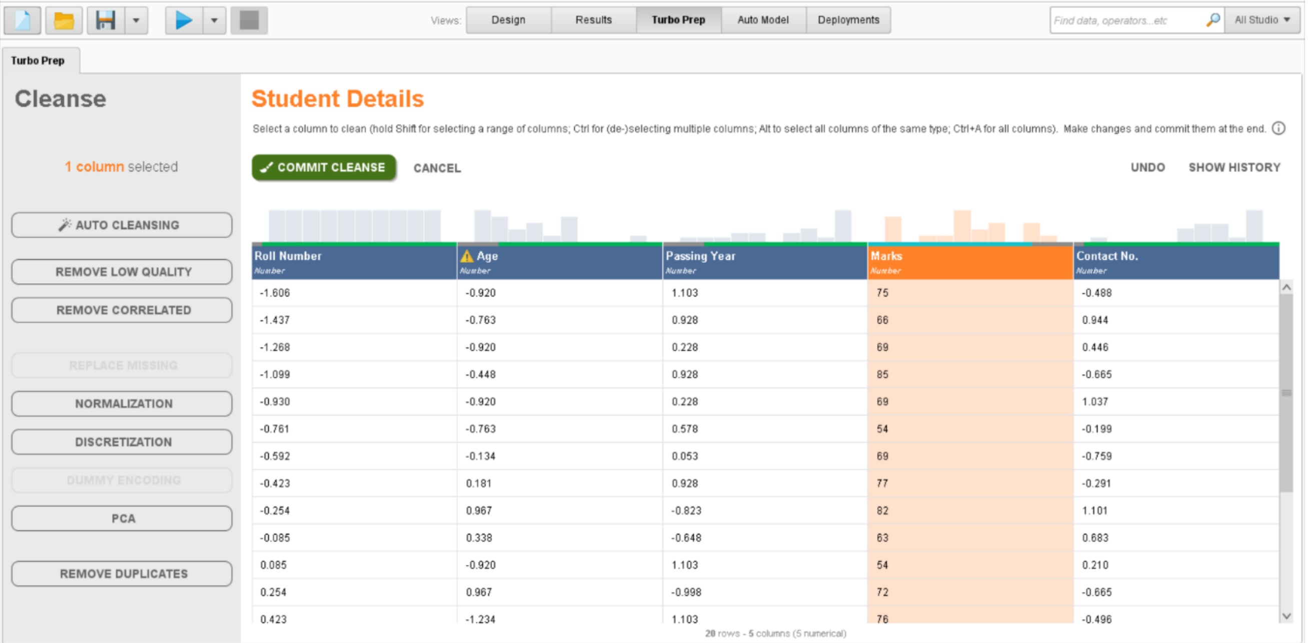
Contact No. Number
6582147796
9658741236
8589657423
6202665894
9858247131
7202648954
6000254545
7004586214
9995554443
9097863254
8080808065
6202620299
6565654242

113	Sheetal	18	ADCA	2021	76

20 rows - 7 columns (2 nominal, 5 numerical)

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**
Course Instructor: **Dr. Geeta Sharma**	
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**
Question No. **03** Page No. **15**	Total Pages. **17**

Step 7: Now we can see the **Final Result** of auto cleansing.



Student Details

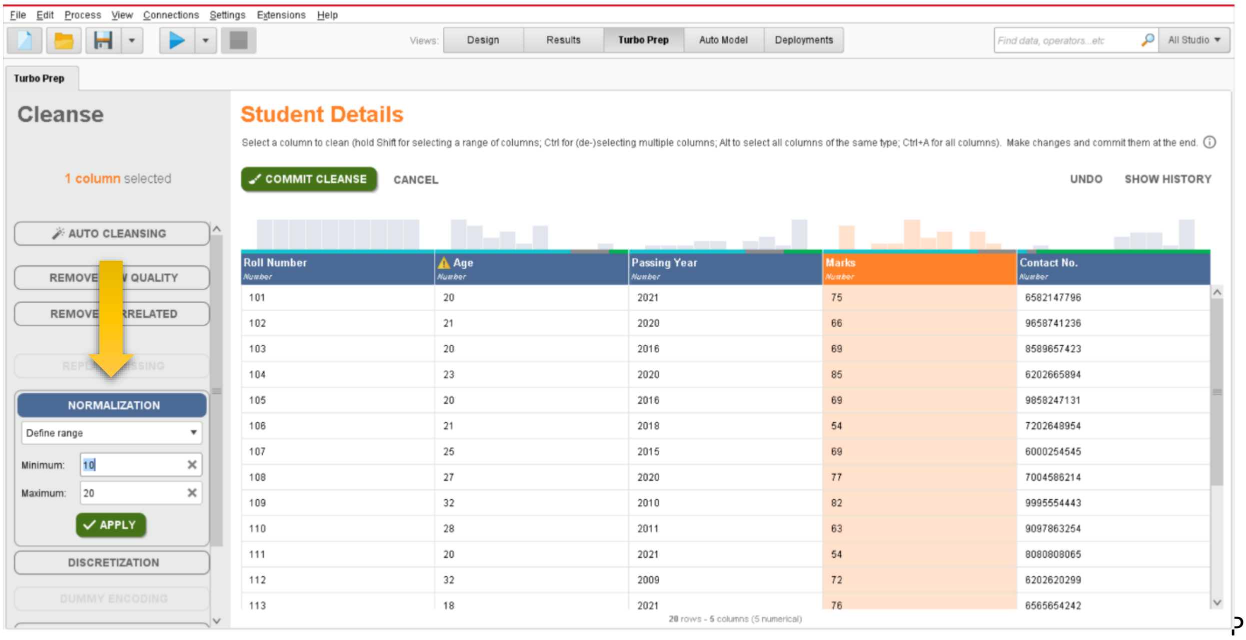
Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the end.

Roll Number Number	Age Number	Passing Year Number	Marks Number	Contact No. Number
-1.606	-0.920	1.103	75	-0.488
-1.437	-0.763	0.928	66	0.944
-1.268	-0.920	0.228	69	0.446
-1.099	-0.448	0.928	85	-0.665
-0.930	-0.920	0.228	69	1.037
-0.761	-0.763	0.578	54	-0.199
-0.592	-0.134	0.053	69	-0.759
-0.423	0.181	0.928	77	-0.291
-0.254	0.967	-0.823	82	1.101
-0.085	0.338	-0.648	63	0.683
0.085	-0.920	1.103	54	0.210
0.254	0.967	-0.998	72	-0.665
0.423	-1.234	1.103	76	-0.496

20 rows - 5 columns (5 numerical)

b) Normalization

Step 1: Select the attribute where we want to perform normalization. Here I selected **Marks.** Now **Define range.**



Define range

Minimum: 10

Maximum: 20

Student Details

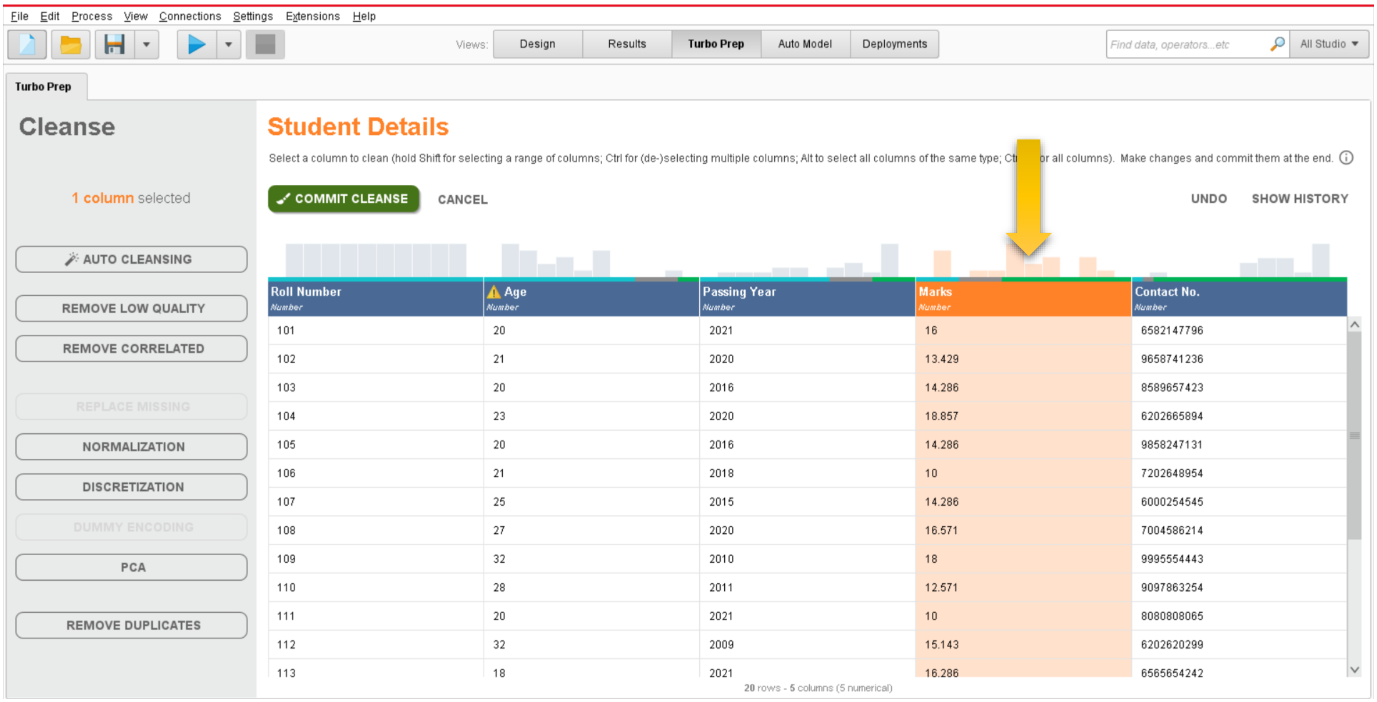
Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the end.

Roll Number Number	Age Number	Passing Year Number	Marks Number	Contact No. Number
101	20	2021	75	6582147796
102	21	2020	66	9658741236
103	20	2016	69	8589657423
104	23	2020	85	6202665894
105	20	2016	69	9858247131
106	21	2018	54	7202648954
107	25	2015	69	6000254545
108	27	2020	77	7004586214
109	32	2010	82	9995554443
110	28	2011	63	9097863254
111	20	2021	54	8080808065
112	32	2009	72	6202620299
113	18	2021	76	6565654242

20 rows - 5 columns (5 numerical)

Course Code: **CAP 447**	Course Title: **Data Warehousing and Data Mining Lab**	
Course Instructor: **Dr. Geeta Sharma**		
Student's Roll no: **RD2110B79** Name: **Atul Kumar**	Student's Reg. no: **12102801**	
Question No. **03**	Page No. **16**	Total Pages. **17**

Step 2: Now we can see marks are **Normalized** between 10-20.



Student Details

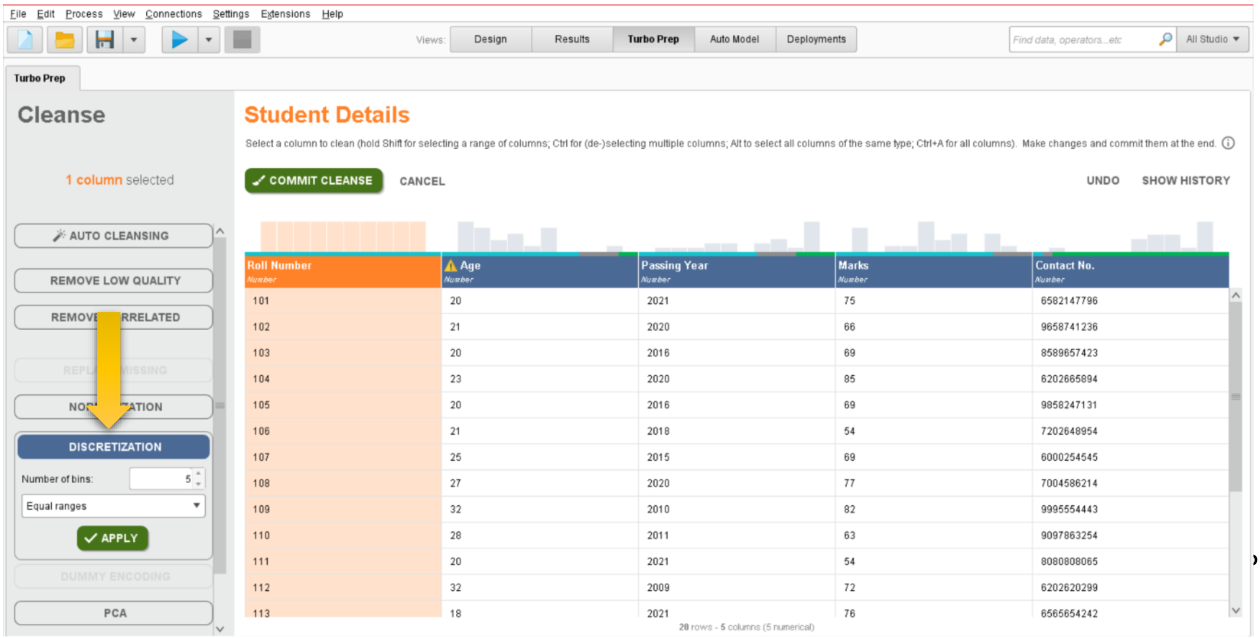
Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ct or all columns). Make changes and commit them at the end.

Roll Number Number	Age Number	Passing Year Number	Marks Number	Contact No. Number
101	20	2021	16	6582147796
102	21	2020	13.429	9658741236
103	20	2016	14.286	8589657423
104	23	2020	18.857	6202665894
105	20	2016	14.286	9858247131
106	21	2018	10	7202648954
107	25	2015	14.286	6000254545
108	27	2020	16.571	7004586214
109	32	2010	18	9995554443
110	28	2011	12.571	9097863254
111	20	2021	10	8080808065
112	32	2009	15.143	6202620299
113	18	2021	16.286	6565654242

20 rows - 5 columns (5 numerical)

c) Discretization

Step 1: Here I selected **Roll Number** for discretization. Now set the number of bins, here I write **5** and then click on **Apply.**



Number of bins: 5

Equal ranges

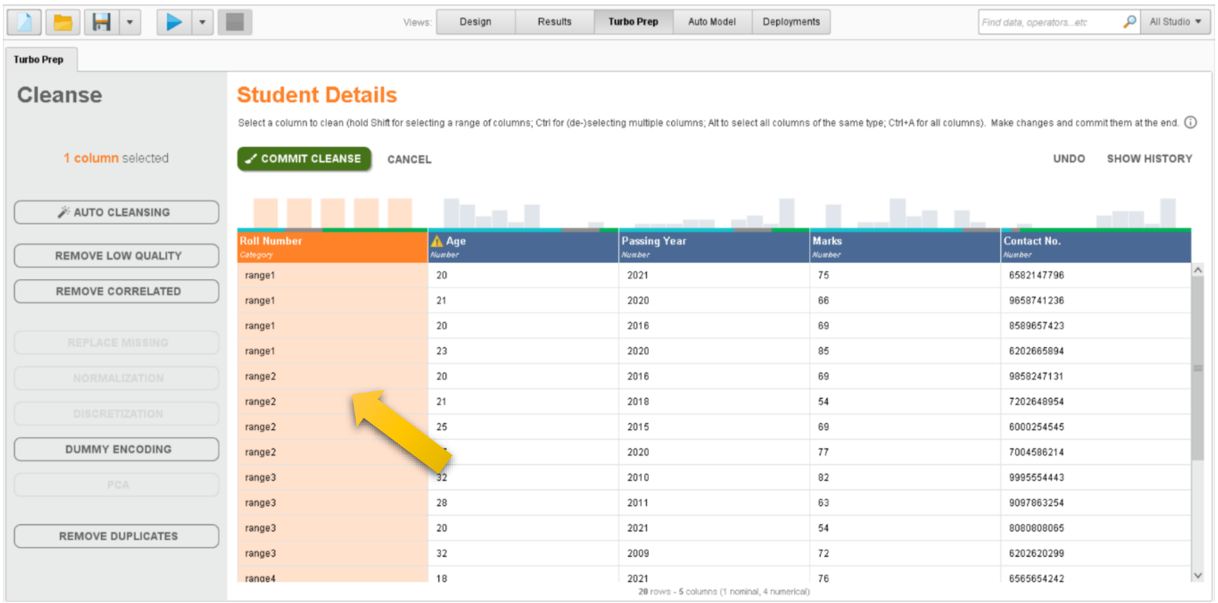
Student Details

Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the end.

Roll Number Number	Age Number	Passing Year Number	Marks Number	Contact No. Number
101	20	2021	75	6582147796
102	21	2020	66	9658741236
103	20	2016	69	8589657423
104	23	2020	85	6202665894
105	20	2016	69	9858247131
106	21	2018	54	7202648954
107	25	2015	69	6000254545
108	27	2020	77	7004586214
109	32	2010	82	9995554443
110	28	2011	63	9097863254
111	20	2021	54	8080808065
112	32	2009	72	6202620299
113	18	2021	76	6565654242

20 rows - 5 columns (5 numerical)

Step 2: Here we can see the data of Roll Number attributes has been changed to **5 equal bins (range1 to range5).**



Student Details

Select a column to clean (hold Shift for selecting a range of columns; Ctrl for (de-)selecting multiple columns; Alt to select all columns of the same type; Ctrl+A for all columns). Make changes and commit them at the end.

Roll Number Category	Age Number	Passing Year Number	Marks Number	Contact No. Number
range1	20	2021	75	6582147796
range1	21	2020	66	9658741236
range1	20	2016	69	8589657423
range1	23	2020	85	6202665894
range2	20	2016	69	9858247131
range2	21	2018	54	7202648954
range2	25	2015	69	6000254545
range2	[illegible]	2020	77	7004586214
range3	32	2010	82	9995554443
range3	28	2011	63	9097863254
range3	20	2021	54	8080808065
range3	32	2009	72	6202620299
range4	18	2021	76	6565654242

20 rows - 5 columns (1 nominal, 4 numerical)

Thank You