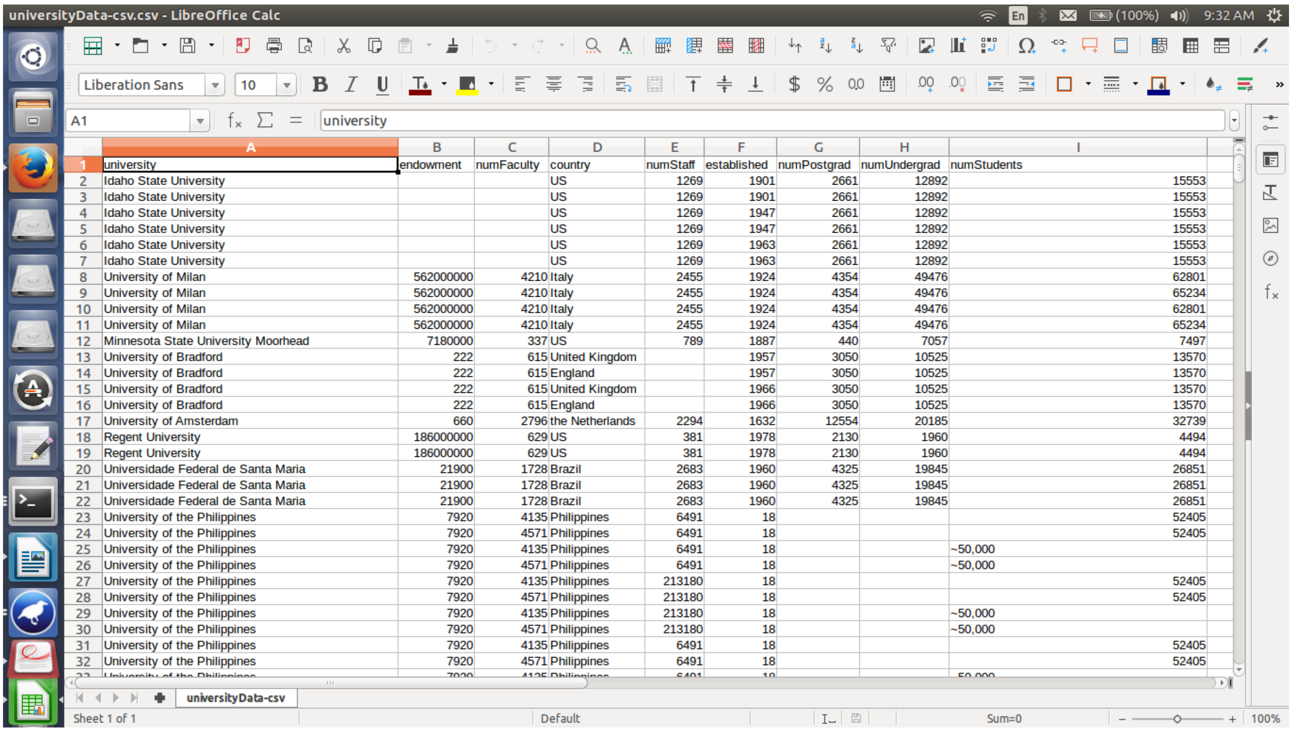
AMMANAMANCHI SAI KARTHIK

B150310CS

Data Set

The size of the data set has been reduced(10000 rows) and the data set is converted into proper CSV fomat and loaded into WEKA which uses ARFF format through the inbuilt converter.



	A	B	C	D	E	F	G	H	I
1	university	endowment	numFaculty	country	numStaff	established	numPostgrad	numUndergrad	numStudents
2	Idaho State University			US	1269	1901	2661	12892	15553
3	Idaho State University			US	1269	1901	2661	12892	15553
4	Idaho State University			US	1269	1947	2661	12892	15553
5	Idaho State University			US	1269	1947	2661	12892	15553
6	Idaho State University			US	1269	1963	2661	12892	15553
7	Idaho State University			US	1269	1963	2661	12892	15553
8	University of Milan	562000000	4210	Italy	2455	1924	4354	49476	62801
9	University of Milan	562000000	4210	Italy	2455	1924	4354	49476	65234
10	University of Milan	562000000	4210	Italy	2455	1924	4354	49476	62801
11	University of Milan	562000000	4210	Italy	2455	1924	4354	49476	65234
12	Minnesota State University Moorhead	7180000	337	US	789	1887	440	7057	7497
13	University of Bradford	222	615	United Kingdom		1957	3050	10525	13570
14	University of Bradford	222	615	England		1957	3050	10525	13570
15	University of Bradford	222	615	United Kingdom		1966	3050	10525	13570
16	University of Bradford	222	615	England		1966	3050	10525	13570
17	University of Amsterdam	660	2796	the Netherlands	2294	1632	12554	20185	32739
18	Regent University	186000000	629	US	381	1978	2130	1960	4494
19	Regent University	186000000	629	US	381	1978	2130	1960	4494
20	Universidade Federal de Santa Maria	21900	1728	Brazil	2683	1960	4325	19845	26851
21	Universidade Federal de Santa Maria	21900	1728	Brazil	2683	1960	4325	19845	26851
22	Universidade Federal de Santa Maria	21900	1728	Brazil	2683	1960	4325	19845	26851
23	University of the Philippines	7920	4135	Philippines	6491	18			52405
24	University of the Philippines	7920	4571	Philippines	6491	18			52405
25	University of the Philippines	7920	4135	Philippines	6491	18			~50,000
26	University of the Philippines	7920	4571	Philippines	6491	18			~50,000
27	University of the Philippines	7920	4135	Philippines	213180	18			52405
28	University of the Philippines	7920	4571	Philippines	213180	18			52405
29	University of the Philippines	7920	4135	Philippines	213180	18			~50,000
30	University of the Philippines	7920	4571	Philippines	213180	18			~50,000
31	University of the Philippines	7920	4135	Philippines	6491	18			52405
32	University of the Philippines	7920	4571	Philippines	6491	18			52405

Format of Submission

The initial data set used is included as a .csv file and the final modified data set is availiable in .xls format.

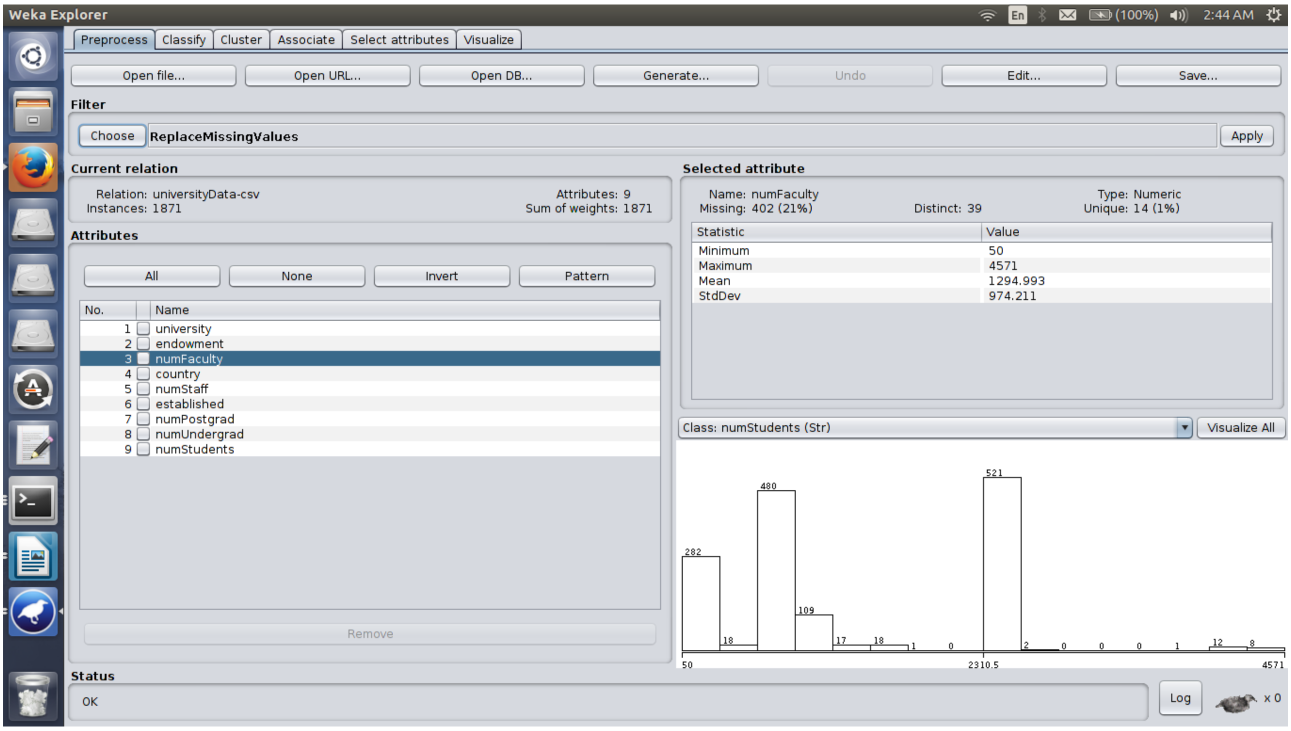
Raw Dataset: universityData-csv.csv

Link: https://drive.google.com/open?id=0B5bh5JNC6R13RktBR2pDOEJMX0U

Pre-Processed Dataset: karthik_b150310cs_dm.xls

Link: https://drive.google.com/open?id=0B5bh5JNC6R13NGw4WFlfMldjQ00

Missing values



Filter: ReplaceMissingValues

Relation: universityData-csv
Instances: 1871
Attributes: 9
Sum of weights: 1871

No.	Name
1	university
2	endowment
3	numFaculty
4	country
5	numStaff
6	established
7	numPostgrad
8	numUndergrad
9	numStudents

Name: numFaculty
Missing: 402 (21%)
Distinct: 39
Type: Numeric
Unique: 14 (1%)

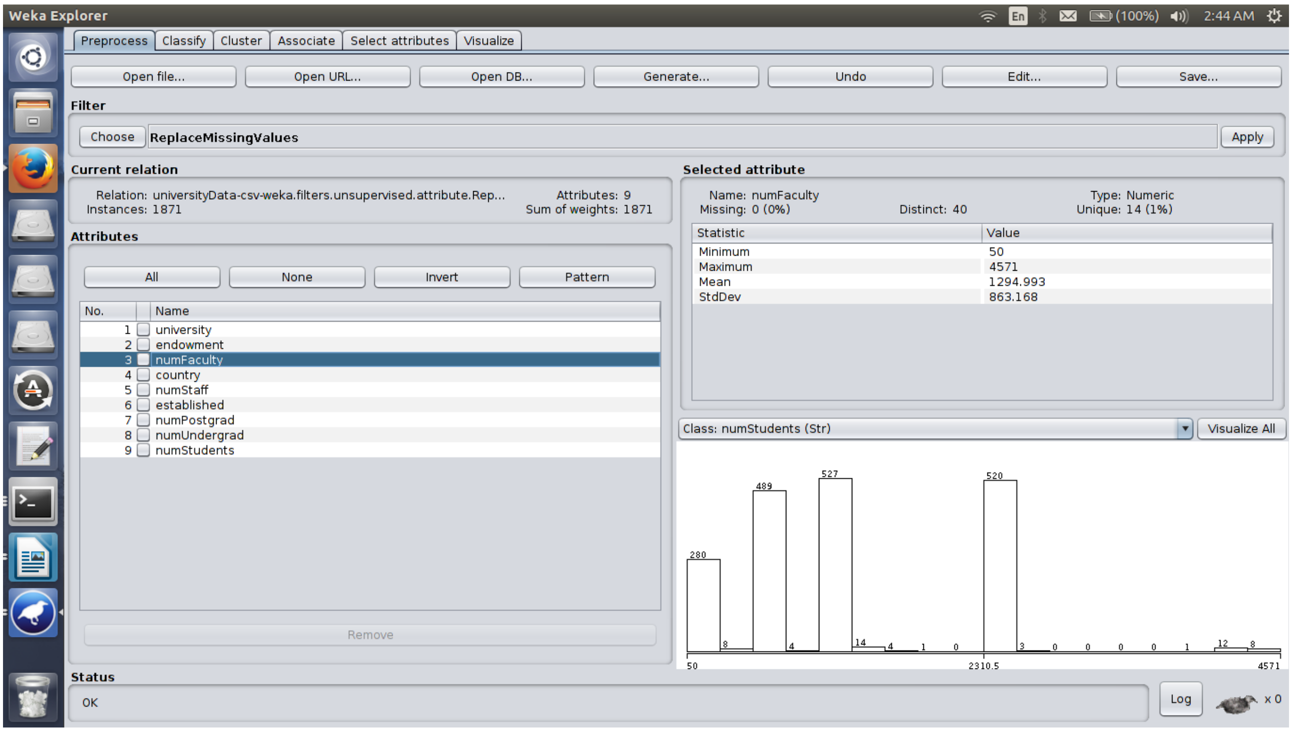
Statistic	Value
Minimum	50
Maximum	4571
Mean	1294.993
StdDev	974.211

Class: numStudents (Str)

Status: OK

We can replace all the missing values by going to **choose->unsupervised attribute filter->replace missing value** .

It automatically replaces the missing values in the data with the mean ot the mode of the dataset.



Filter: ReplaceMissingValues

Relation: universityData-csv-weka.filters.unsupervised.attribute.Rep...
Instances: 1871
Attributes: 9
Sum of weights: 1871

No.	Name
1	university
2	endowment
3	numFaculty
4	country
5	numStaff
6	established
7	numPostgrad
8	numUndergrad
9	numStudents

Name: numFaculty
Missing: 0 (0%)
Distinct: 40
Type: Numeric
Unique: 14 (1%)

Statistic	Value
Minimum	50
Maximum	4571
Mean	1294.993
StdDev	863.168

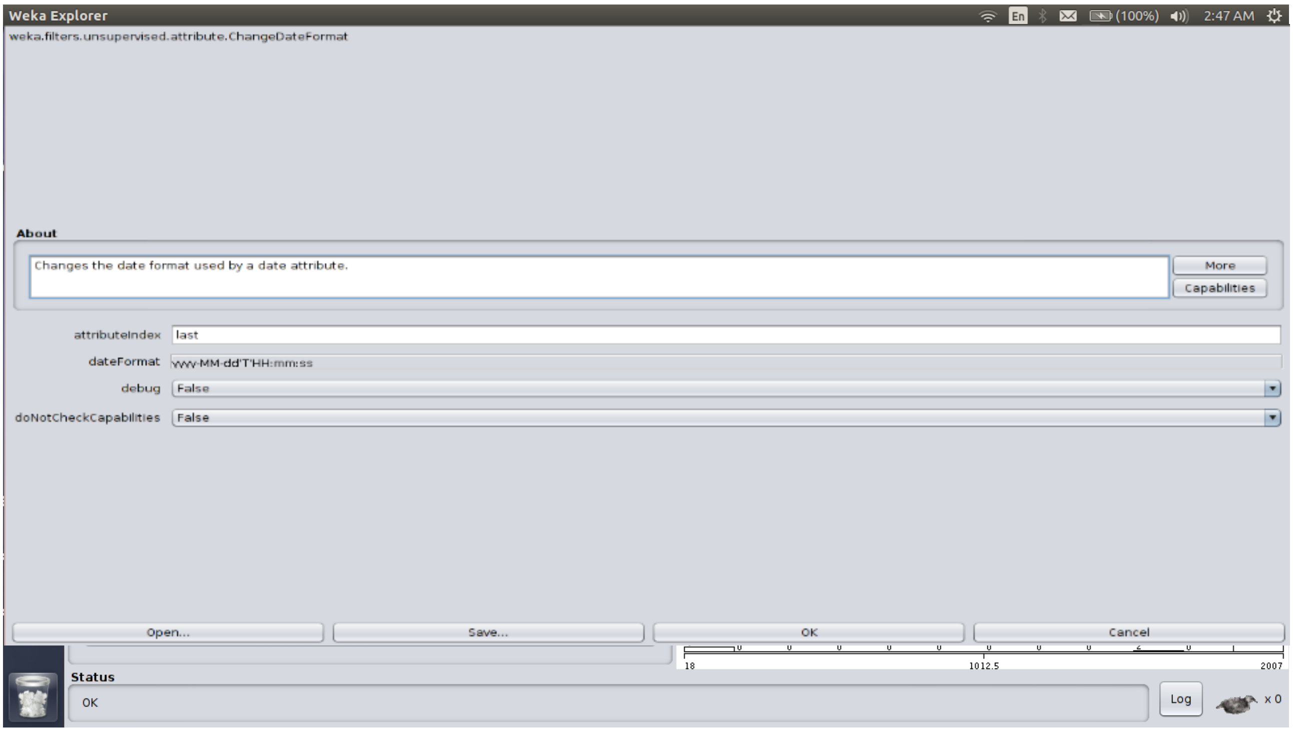
Class: numStudents (Str)

Status: OK

CHANGING DATE FORMAT

In our data we have an attribute violation data which is of the form MM/dd/yyyy.The dataset is clean regarding dates.

We can change date format by choosing ChangeDateFormat from unsupervised folder.



weka.filters.unsupervised.attribute.ChangeDateFormat

About

Changes the date format used by a date attribute.

attributeIndex: last

dateFormat: yyyy-MM-dd'T'HH:mm:ss

debug: False

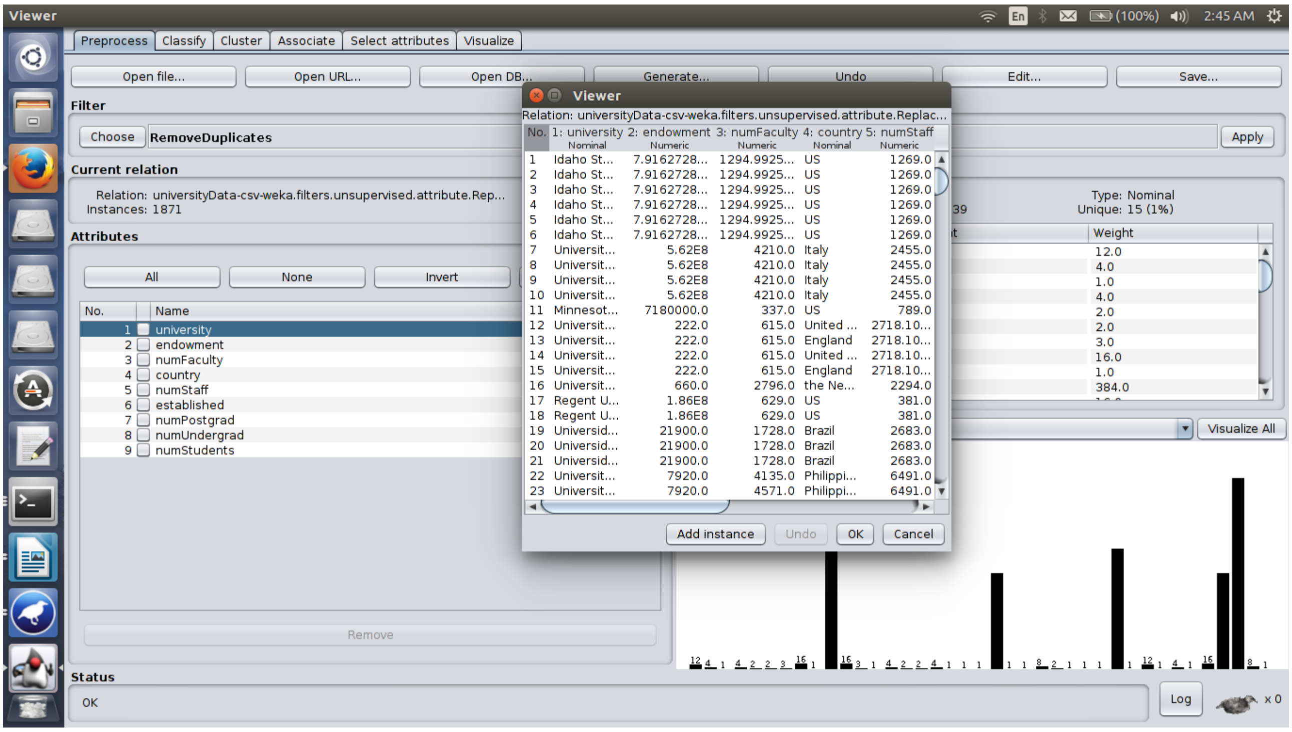
doNotCheckCapabilities: False

Status

OK

REMOVING DUPLICATE DATA

To remove duplicate data we click on choose->filter->remove Duplicate Data under unsupervised category.



Filter: RemoveDuplicates

Current relation

Relation: universityData-csv-weka.filters.unsupervised.attribute.Rep...
Instances: 1871

Attributes

No.	Name
1	university
2	endowment
3	numFaculty
4	country
5	numStaff
6	established
7	numPostgrad
8	numUndergrad
9	numStudents

Type: Nominal
Unique: 15 (1%)

Viewer

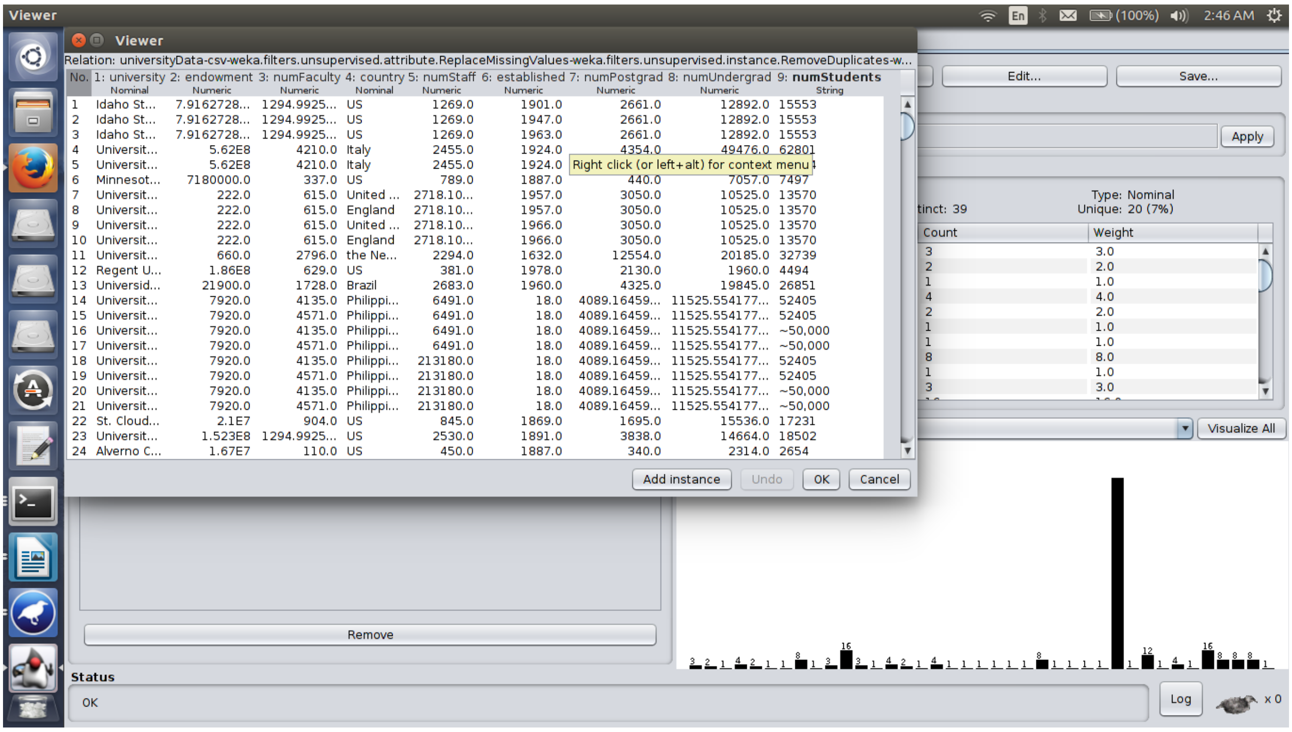
Relation: universityData-csv-weka.filters.unsupervised.attribute.Replac...

No.	1: university Nominal	2: endowment Numeric	3: numFaculty Numeric	4: country Nominal	5: numStaff Numeric
1	Idaho St...	7.9162728...	1294.9925...	US	1269.0
2	Idaho St...	7.9162728...	1294.9925...	US	1269.0
3	Idaho St...	7.9162728...	1294.9925...	US	1269.0
4	Idaho St...	7.9162728...	1294.9925...	US	1269.0
5	Idaho St...	7.9162728...	1294.9925...	US	1269.0
6	Idaho St...	7.9162728...	1294.9925...	US	1269.0
7	Universit...	5.62E8	4210.0	Italy	2455.0
8	Universit...	5.62E8	4210.0	Italy	2455.0
9	Universit...	5.62E8	4210.0	Italy	2455.0
10	Universit...	5.62E8	4210.0	Italy	2455.0
11	Minnesot...	7180000.0	337.0	US	789.0
12	Universit...	222.0	615.0	United ...	2718.10...
13	Universit...	222.0	615.0	England	2718.10...
14	Universit...	222.0	615.0	United ...	2718.10...
15	Universit...	222.0	615.0	England	2718.10...
16	Universit...	660.0	2796.0	the Ne...	2294.0
17	Regent U...	1.86E8	629.0	US	381.0
18	Regent U...	1.86E8	629.0	US	381.0
19	Universid...	21900.0	1728.0	Brazil	2683.0
20	Universid...	21900.0	1728.0	Brazil	2683.0
21	Universid...	21900.0	1728.0	Brazil	2683.0
22	Universit...	7920.0	4135.0	Philippi...	6491.0
23	Universit...	7920.0	4571.0	Philippi...	6491.0

Status

OK

Before removing missing values.



Relation: universityData-csv-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.instance.RemoveDuplicates-w...

No.	1: university Nominal	2: endowment Numeric	3: numFaculty Numeric	4: country Nominal	5: numStaff Numeric	6: established Numeric	7: numPostgrad Numeric	8: numUndergrad Numeric	9: **numStudents** String
1	Idaho St...	7.9162728...	1294.9925...	US	1269.0	1901.0	2661.0	12892.0	15553
2	Idaho St...	7.9162728...	1294.9925...	US	1269.0	1947.0	2661.0	12892.0	15553
3	Idaho St...	7.9162728...	1294.9925...	US	1269.0	1963.0	2661.0	12892.0	15553
4	Universit...	5.62E8	4210.0	Italy	2455.0	1924.0	4354.0	49476.0	62801
5	Universit...	5.62E8	4210.0	Italy	2455.0	1924.0			4
6	Minnesot...	7180000.0	337.0	US	789.0	1887.0	440.0	7057.0	7497
7	Universit...	222.0	615.0	United ...	2718.10...	1957.0	3050.0	10525.0	13570
8	Universit...	222.0	615.0	England	2718.10...	1957.0	3050.0	10525.0	13570
9	Universit...	222.0	615.0	United ...	2718.10...	1966.0	3050.0	10525.0	13570
10	Universit...	222.0	615.0	England	2718.10...	1966.0	3050.0	10525.0	13570
11	Universit...	660.0	2796.0	the Ne...	2294.0	1632.0	12554.0	20185.0	32739
12	Regent U...	1.86E8	629.0	US	381.0	1978.0	2130.0	1960.0	4494
13	Universid...	21900.0	1728.0	Brazil	2683.0	1960.0	4325.0	19845.0	26851
14	Universit...	7920.0	4135.0	Philippi...	6491.0	18.0	4089.16459...	11525.554177...	52405
15	Universit...	7920.0	4571.0	Philippi...	6491.0	18.0	4089.16459...	11525.554177...	52405
16	Universit...	7920.0	4135.0	Philippi...	6491.0	18.0	4089.16459...	11525.554177...	~50,000
17	Universit...	7920.0	4571.0	Philippi...	6491.0	18.0	4089.16459...	11525.554177...	~50,000
18	Universit...	7920.0	4135.0	Philippi...	213180.0	18.0	4089.16459...	11525.554177...	52405
19	Universit...	7920.0	4571.0	Philippi...	213180.0	18.0	4089.16459...	11525.554177...	52405
20	Universit...	7920.0	4135.0	Philippi...	213180.0	18.0	4089.16459...	11525.554177...	~50,000
21	Universit...	7920.0	4571.0	Philippi...	213180.0	18.0	4089.16459...	11525.554177...	~50,000
22	St. Cloud...	2.1E7	904.0	US	845.0	1869.0	1695.0	15536.0	17231
23	Universit...	1.523E8	1294.9925...	US	2530.0	1891.0	3838.0	14664.0	18502
24	Alverno C...	1.67E7	110.0	US	450.0	1887.0	340.0	2314.0	2654

Right click (or left+alt) for context menu

tinct: 39

Type: Nominal
Unique: 20 (7%)

Count	Weight
3	3.0
2	2.0
1	1.0
4	4.0
2	2.0
1	1.0
1	1.0
8	8.0
1	1.0
3	3.0

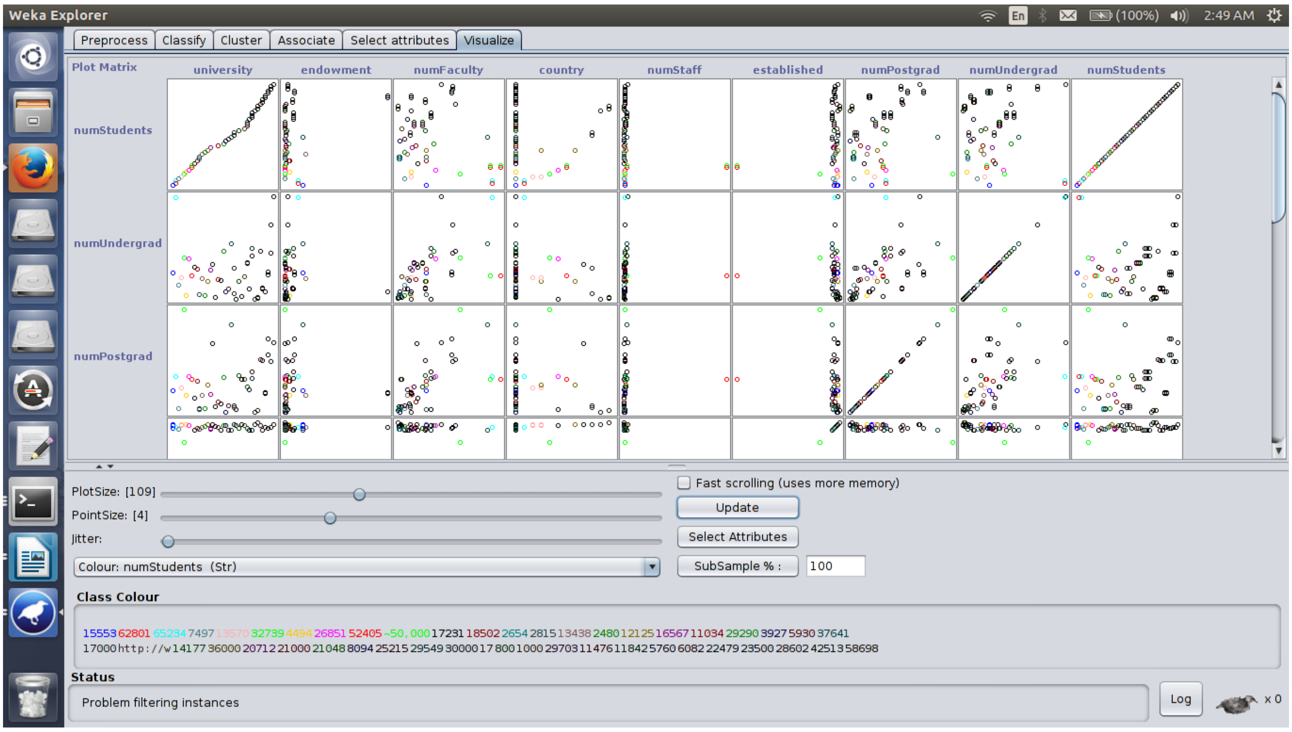
Status
OK

After removing missing values.

Here two instances of a university are same as some of the mmissing values are replaced by means and mode which might come out to be different for differrent instances.

SCATTERPLOT MATRIX

When attributes are numeric we can create a scatter plot of one attribute against another. This is useful as it can highlight any patterns in the relationship between the attributes, such as positive or negative correlations.So,Weka provides us Visualize tab for this purpose.



PlotSize: [109]

PointSize: [4]

Jitter:

Colour: numStudents (Str)

□ Fast scrolling (uses more memory)

SubSample % : 100

Class Colour

15553 62801 65234 7497 13570 32739 4494 26851 52405 ~50, 000 17231 18502 2654 2815 13438 2480 12125 16567 11034 29290 3927 5930 37641
17000 http://w 14177 36000 20712 21000 21048 8094 25215 29549 30000 17 800 1000 29703 11476 11842 5760 6082 22479 23500 28602 42513 58698

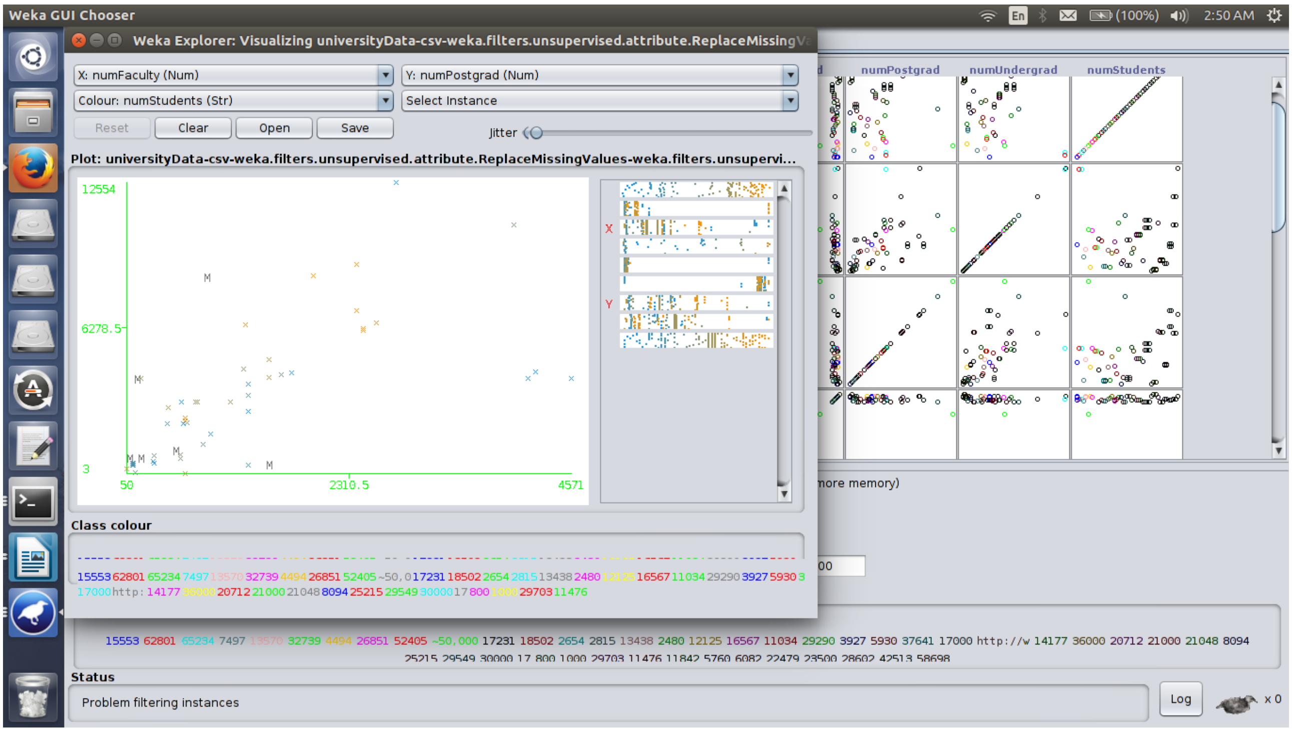
Status

Problem filtering instances

The dots in the scatter plots are colored by their class value.Clicking on a plot will give you a new window with the plot .All combinations of attributes are plotted in a systematic way.

We can also see that each plot appears twice, first in the top left triangle and again in the bottom right triangle with the axes flipped. We can also see a series of plots starting in the bottom left and continuing to the top right where each attribute is plotted against itself.

Controls at the bottom of the screen. These let us increase the size of the plots, increase the size of the dots and add jitter.



X: numFaculty (Num)

Y: numPostgrad (Num)

Colour: numStudents (Str)

Select Instance

Plot: universityData-csv-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervi...

Class colour

15553 62801 65234 7497 13570 32739 4494 26851 52405 ~50, 0 17231 18502 2654 2815 13438 2480 12125 16567 11034 29290 3927 5930 3
17000 http: 14177 36000 20712 21000 21048 8094 25215 29549 30000 17 800 1000 29703 11476

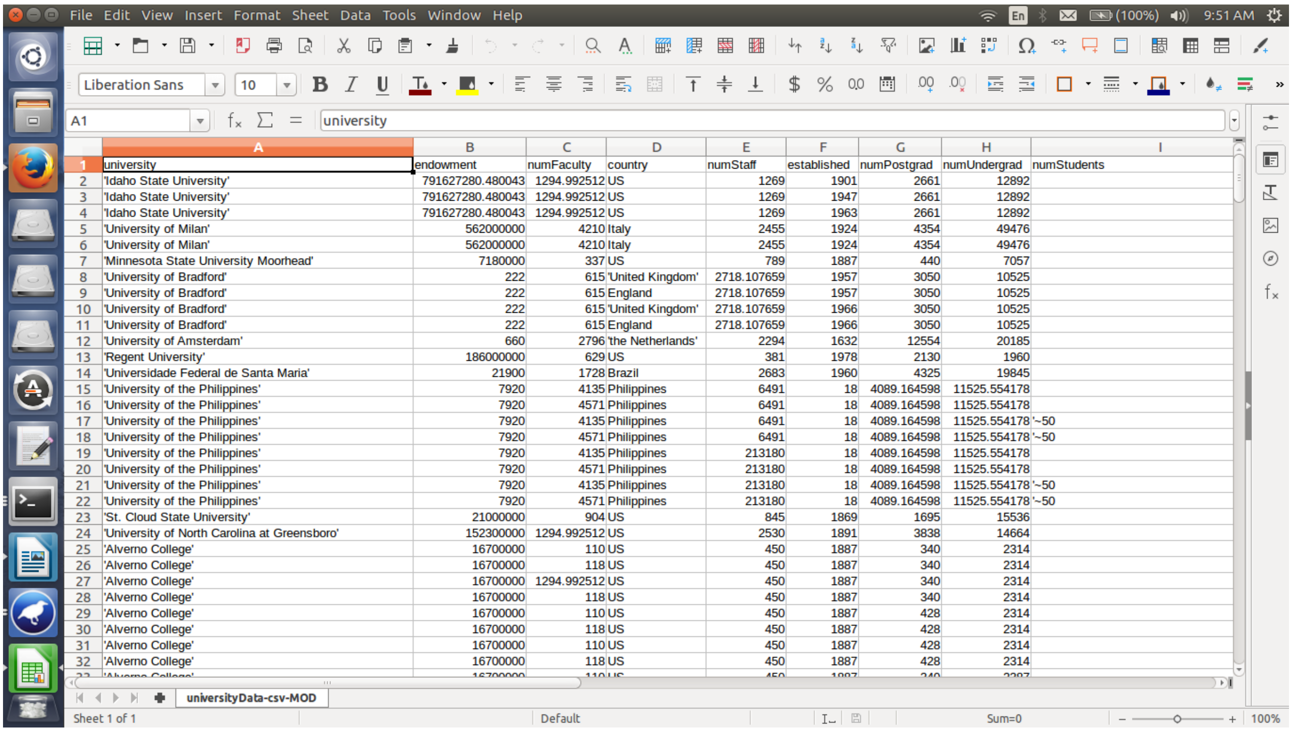
15553 62801 65234 7497 13570 32739 4494 26851 52405 ~50, 000 17231 18502 2654 2815 13438 2480 12125 16567 11034 29290 3927 5930 37641 17000 http://w 14177 36000 20712 21000 21048 8094
25215 29549 30000 17 800 1000 29703 11476 11842 5760 6082 22479 23500 28602 42513 58698

Status

Problem filtering instances

EXPLORE DATA

Right-click the model, choose "Visualize classifier errors" and Save the data (including the prediction).



	A	B	C	D	E	F	G	H	I
1	university	endowment	numFaculty	country	numStaff	established	numPostgrad	numUndergrad	numStudents
2	'Idaho State University'	791627280.480043	1294.992512	US	1269	1901	2661	12892	
3	'Idaho State University'	791627280.480043	1294.992512	US	1269	1947	2661	12892	
4	'Idaho State University'	791627280.480043	1294.992512	US	1269	1963	2661	12892	
5	'University of Milan'	562000000	4210	Italy	2455	1924	4354	49476	
6	'University of Milan'	562000000	4210	Italy	2455	1924	4354	49476	
7	'Minnesota State University Moorhead'	7180000	337	US	789	1887	440	7057	
8	'University of Bradford'	222	615	'United Kingdom'	2718.107659	1957	3050	10525	
9	'University of Bradford'	222	615	England	2718.107659	1957	3050	10525	
10	'University of Bradford'	222	615	'United Kingdom'	2718.107659	1966	3050	10525	
11	'University of Bradford'	222	615	England	2718.107659	1966	3050	10525	
12	'University of Amsterdam'	660	2796	'the Netherlands'	2294	1632	12554	20185	
13	'Regent University'	186000000	629	US	381	1978	2130	1960	
14	'Universidade Federal de Santa Maria'	21900	1728	Brazil	2683	1960	4325	19845	
15	'University of the Philippines'	7920	4135	Philippines	6491	18	4089.164598	11525.554178	
16	'University of the Philippines'	7920	4571	Philippines	6491	18	4089.164598	11525.554178	
17	'University of the Philippines'	7920	4135	Philippines	6491	18	4089.164598	11525.554178	'~50
18	'University of the Philippines'	7920	4571	Philippines	6491	18	4089.164598	11525.554178	'~50
19	'University of the Philippines'	7920	4135	Philippines	213180	18	4089.164598	11525.554178	
20	'University of the Philippines'	7920	4571	Philippines	213180	18	4089.164598	11525.554178	
21	'University of the Philippines'	7920	4135	Philippines	213180	18	4089.164598	11525.554178	'~50
22	'University of the Philippines'	7920	4571	Philippines	213180	18	4089.164598	11525.554178	'~50
23	'St. Cloud State University'	21000000	904	US	845	1869	1695	15536	
24	'University of North Carolina at Greensboro'	152300000	1294.992512	US	2530	1891	3838	14664	
25	'Alverno College'	16700000	110	US	450	1887	340	2314	
26	'Alverno College'	16700000	118	US	450	1887	340	2314	
27	'Alverno College'	16700000	1294.992512	US	450	1887	340	2314	
28	'Alverno College'	16700000	118	US	450	1887	340	2314	
29	'Alverno College'	16700000	110	US	450	1887	428	2314	
30	'Alverno College'	16700000	118	US	450	1887	428	2314	
31	'Alverno College'	16700000	110	US	450	1887	428	2314	
32	'Alverno College'	16700000	118	US	450	1887	428	2314	

OpenRefine does not have the option/pre-defined algorithm to fill in missing values.

Advantages of open refine over weka :

1. In open refine we can edit a lot of instances at the same time by using the facet option but in weka it is not possible we can edit only one instance at a time.

2. Exporting the data from open refine to excel is very easy as it provided a direct option but in weka the file has to be saved in some format and then it has to be loaded onto excel.

3. quick, interactive, filter facets which allow for easy browsing of instances/rows which match a variety of filters

4.complete provenance/undo history of all modifications

5. wide variety of input & output formats including both file formats and online repositories like Google Spreadsheets & Fusion Tables.

Advantages of weka over open refine:

1. Weka contains a large collection of data algorithms.

2. removal of attributes can be done easily in weka rather than open refine.

3. Weka provides a lot of attribute filter options such as discritisation,principal components,binarization,normalization etc..

4. Weka also provides a lot of attribute selection options such as correlation attribute evaluation,gainratio attribute evaluation,principal components etc..

5.Weka contains various data transformation techniques such as nominal to binary Nominal to string etc..

6. Missing values can be replaced for all attributes at once in weka but in open refine it has to be done separately for each attribute.