Experiment 3: Data Processing Techniques on Data Set

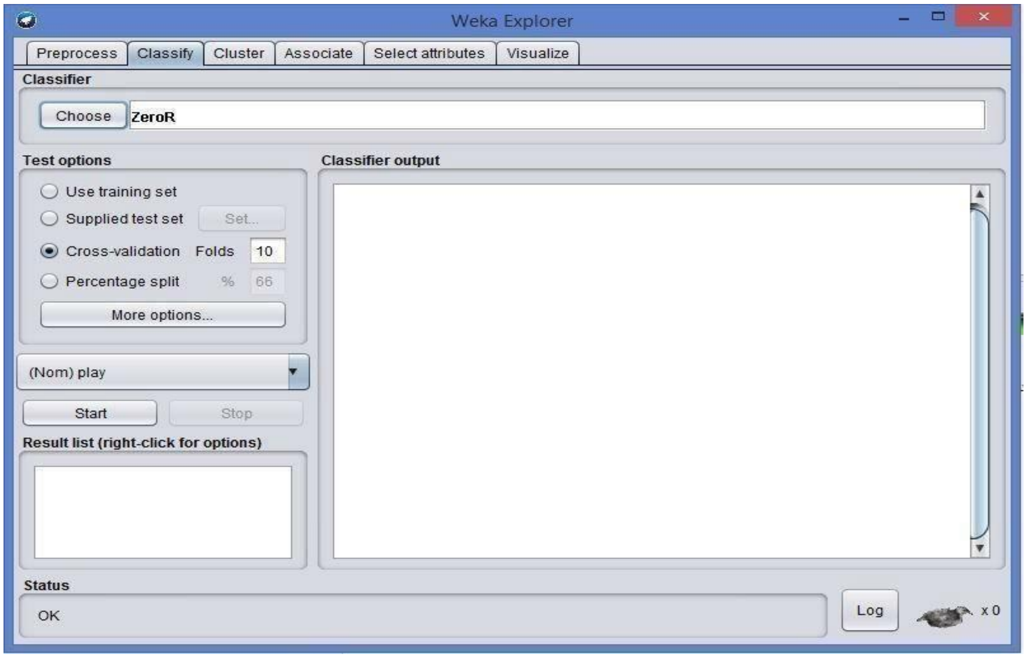
Aim: 3a) Pre-process a given dataset based on Attribute selection

To search through all possible combinations of attributes in the data and find which subset of attributes works best for prediction, make sure that you set up attribute evaluator to „Cfs Subset Val‟ and a search method to „Best First‟. The evaluator will determine what method to use toassign a worth to each subset of attributes. The search method will determine what style of search to perform. The options that you can set for selection in the „Attribute Selection Mode‟ fig no: 3.2

1. **Use full training set.** The worth of the attribute subset is determined using the full set of training data.

2 **Cross-validation.** The worth of the attribute subset is determined by a process of cross-validation. The „Fold‟ and „Seed‟ fields set the number of folds to use and the random seed used when shuffling the data.

Specify which attribute to treat as the class in the drop-down box below the test options. Once all the test options are set, you can start the attribute selection process by clicking on „Start‟ button.



Classifier: ZeroR

● Cross-validation Folds 10

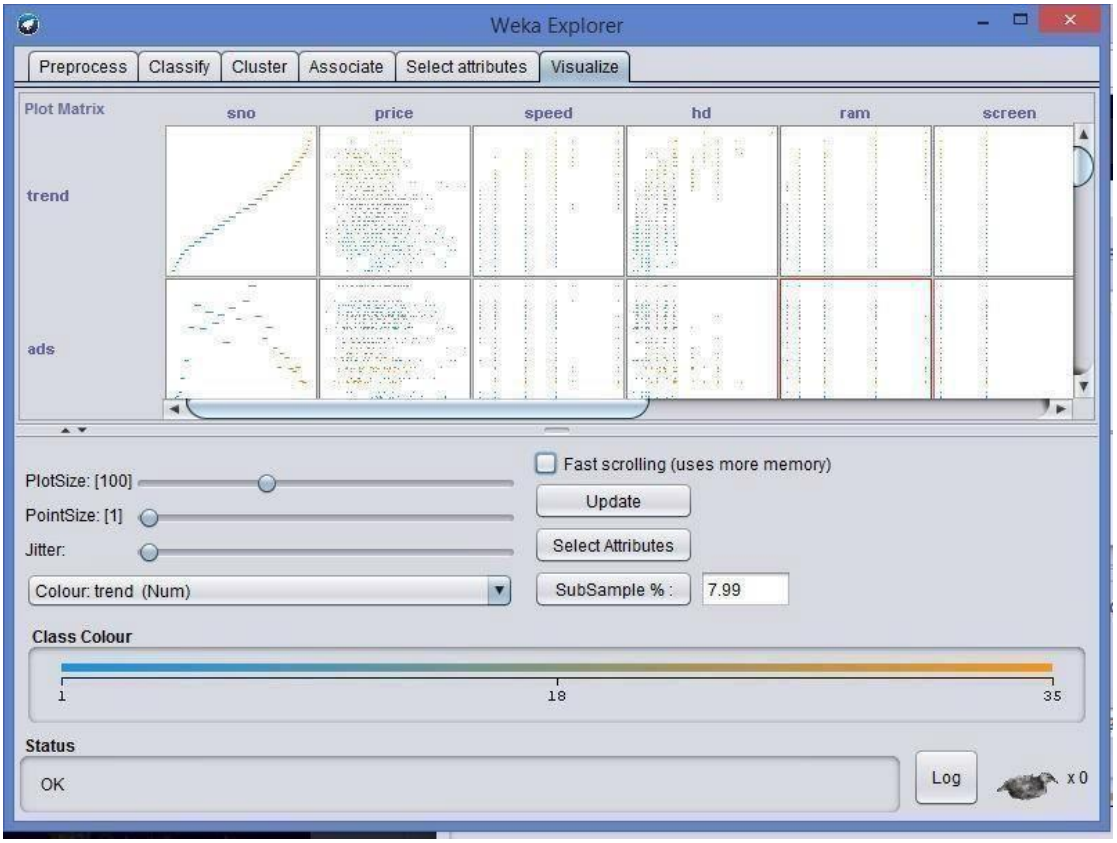
(Nom) play

Status: OK

Fig: 3.1 Choosing Cross validation

When it is finished, the results of selection are shown on the right part of the window and entry is added to the „Result list‟.

2. Visualizing Results



☐ Fast scrolling (uses more memory)

PlotSize: [100]

PointSize: [1]

Jitter:

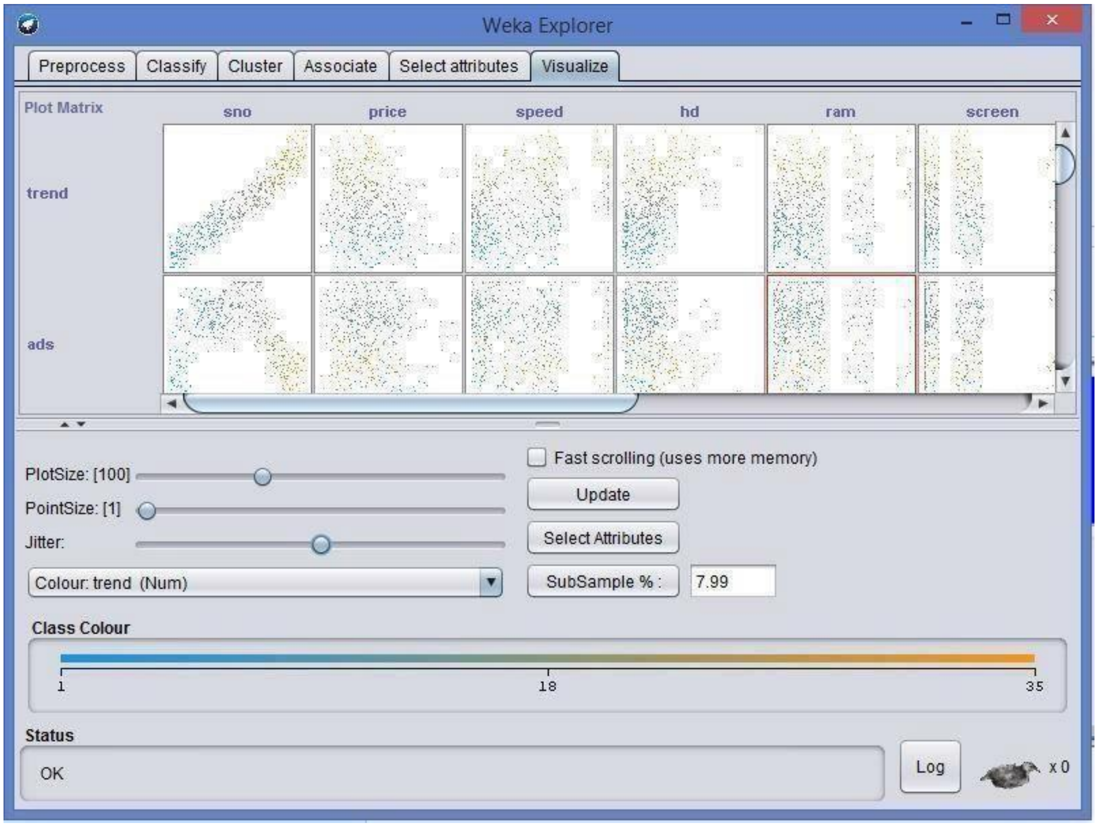
Colour: trend (Num)

SubSample % : 7.99

Status: OK

Fig: 3.2 Data Visualization

WEKA‟s visualization allows you to visualize a 2-D plot of the current working relation. Visualization is very useful in practice; it helps to determine difficulty of the learning problem. WEKA can visualize single attributes (1-d) and pairs of attributes (2-d), rotate 3-d visualizations (Xgobi-style). WEKA has “Jitter” option to deal with nominal attributes and to detect “hidden” data points.



Weka Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Plot Matrix

PlotSize: [100]

PointSize: [1]

Jitter:

Colour: trend (Num)

☐ Fast scrolling (uses more memory)

Update

Select Attributes

SubSample % : 7.99

Class Colour

Status

OK

Log

x 0

Fig 3.3: Preprocessing with jitter

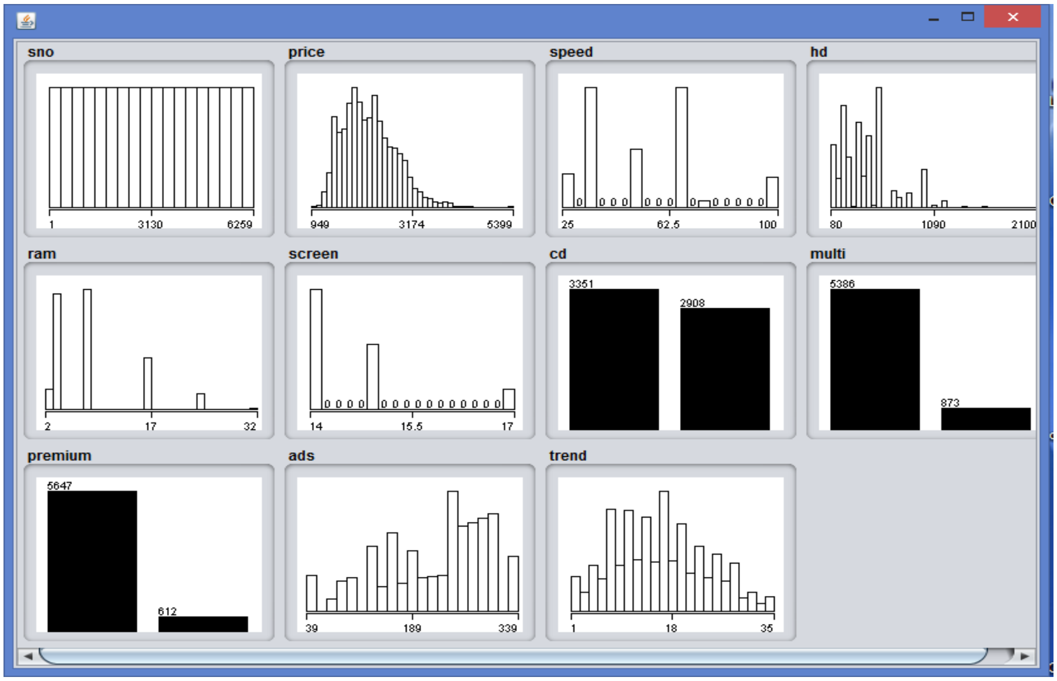


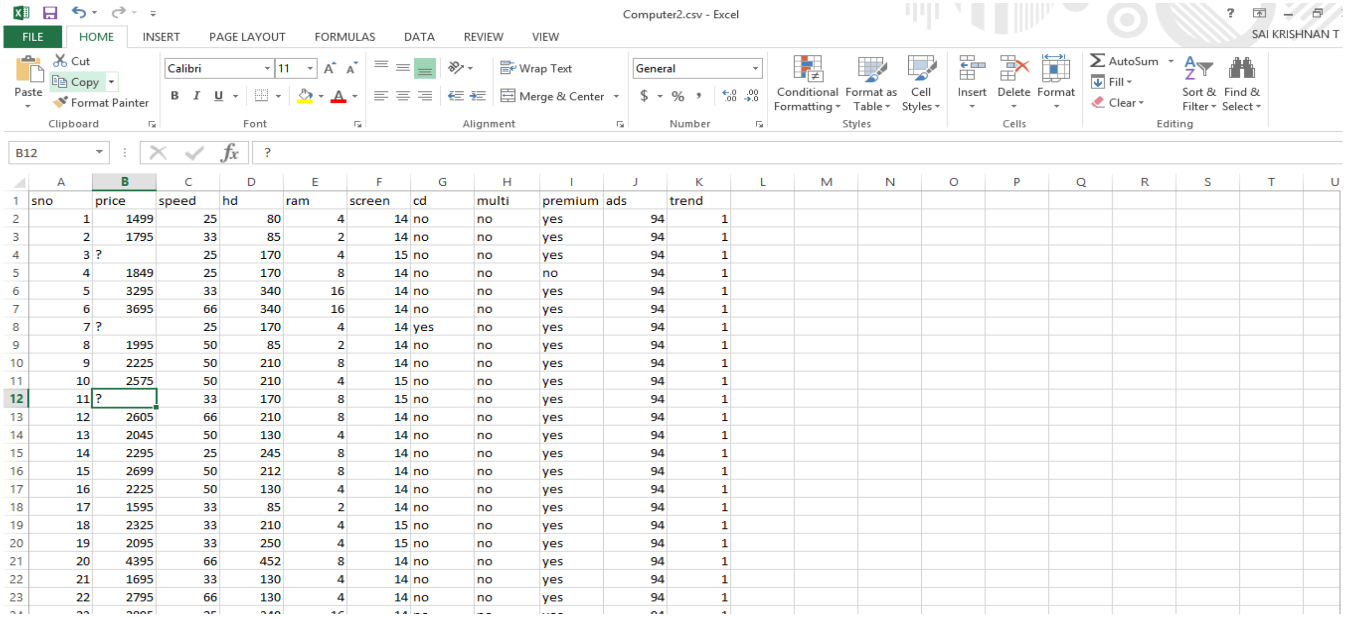
Fig: 3.3 Data visualization

Exercise

1. Explain data preprocessing steps for heart disease dataset.

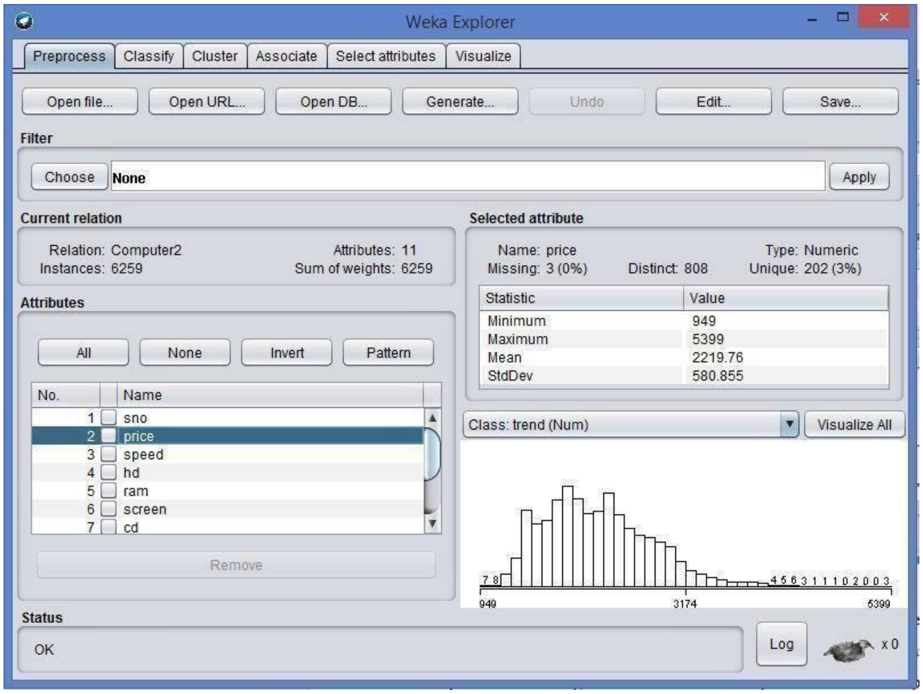
Aim: B. Pre-process a given dataset based on Handling Missing Values

Process: Replacing Missing Attribute Values by the Attribute Mean. This method is used for data sets with numerical attributes. An example of such a data set is presented in fig no: 3.4



	A	B	C	D	E	F	G	H	I	J	K
1	sno	price	speed	hd	ram	screen	cd	multi	premium	ads	trend
2	1	1499	25	80	4	14	no	no	yes	94	1
3	2	1795	33	85	2	14	no	no	yes	94	1
4	3	?	25	170	4	15	no	no	yes	94	1
5	4	1849	25	170	8	14	no	no	no	94	1
6	5	3295	33	340	16	14	no	no	yes	94	1
7	6	3695	66	340	16	14	no	no	yes	94	1
8	7	?	25	170	4	14	yes	no	yes	94	1
9	8	1995	50	85	2	14	no	no	yes	94	1
10	9	2225	50	210	8	14	no	no	yes	94	1
11	10	2575	50	210	4	15	no	no	yes	94	1
12	11	?	33	170	8	15	no	no	yes	94	1
13	12	2605	66	210	8	14	no	no	yes	94	1
14	13	2045	50	130	4	14	no	no	yes	94	1
15	14	2295	25	245	8	14	no	no	yes	94	1
16	15	2699	50	212	8	14	no	no	yes	94	1
17	16	2225	50	130	4	14	no	no	yes	94	1
18	17	1595	33	85	2	14	no	no	yes	94	1
19	18	2325	33	210	4	15	no	no	yes	94	1
20	19	2095	33	250	4	15	no	no	yes	94	1
21	20	4395	66	452	8	14	no	no	yes	94	1
22	21	1695	33	130	4	14	no	no	yes	94	1
23	22	2795	66	130	4	14	no	no	yes	94	1

Fig: 3.4 Missing values



Filter: None

Relation: Computer2
Instances: 6259
Attributes: 11
Sum of weights: 6259

No.	Name
1	sno
2	price
3	speed
4	hd
5	ram
6	screen
7	cd

Name: price
Missing: 3 (0%)
Distinct: 808
Type: Numeric
Unique: 202 (3%)

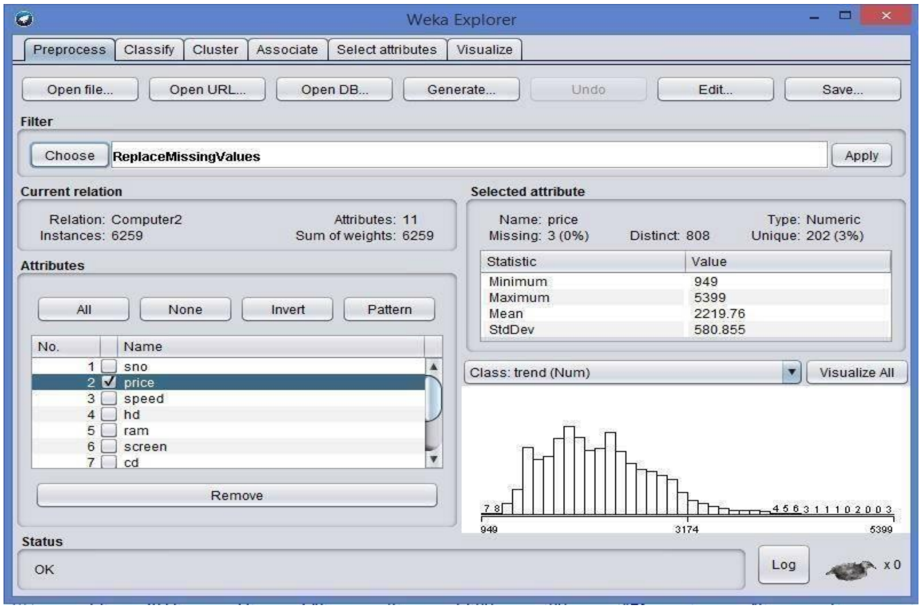
Statistic	Value
Minimum	949
Maximum	5399
Mean	2219.76
StdDev	580.855

Class: trend (Num)

Status: OK

Fig: 3.5 Choosing a dataset

In this method, every missing attribute value for a numerical attribute is replaced by the arithmetic mean of known attribute values. In Fig, the mean of known attribute values for Temperature is 99.2, hence all missing attribute values for Temperature should be replaced by The table with missing attribute values replaced by the mean is presented in fig. For symbolic attributes Headache and Nausea, missing attribute values were replaced using the most common value of the Replace Missing Values.



Filter: ReplaceMissingValues

Current relation

Relation: Computer2 Attributes: 11
Instances: 6259 Sum of weights: 6259

Selected attribute

Name: price Type: Numeric
Missing: 3 (0%) Distinct: 808 Unique: 202 (3%)

Statistic	Value
Minimum	949
Maximum	5399
Mean	2219.76
StdDev	580.855

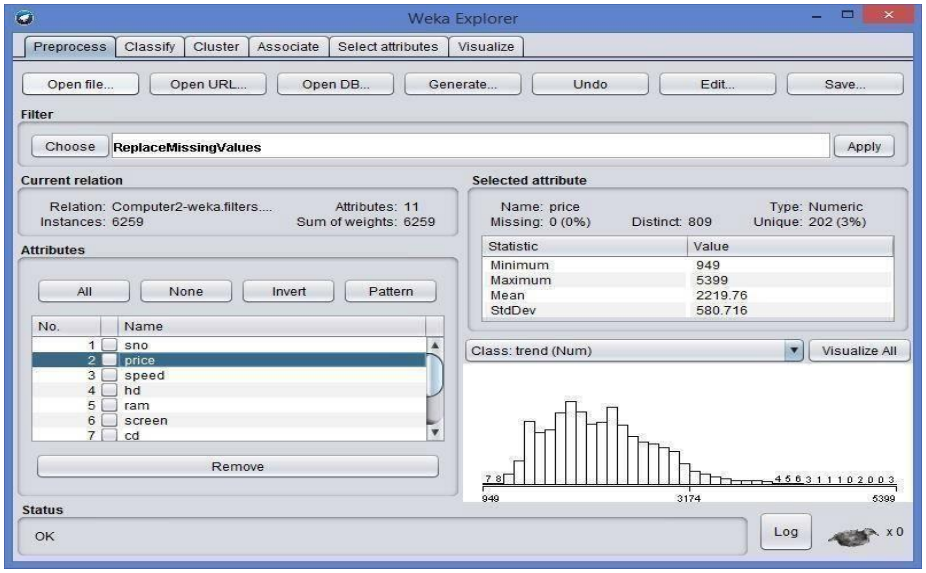
Attributes

No.		Name
1	□	sno
2	☑	price
3	□	speed
4	□	hd
5	□	ram
6	□	screen
7	□	cd

Class: trend (Num)

Status

OK



Filter: ReplaceMissingValues

Current relation

Relation: Computer2-weka.filters.... Attributes: 11
Instances: 6259 Sum of weights: 6259

Selected attribute

Name: price Type: Numeric
Missing: 0 (0%) Distinct: 809 Unique: 202 (3%)

Statistic	Value
Minimum	949
Maximum	5399
Mean	2219.76
StdDev	580.716

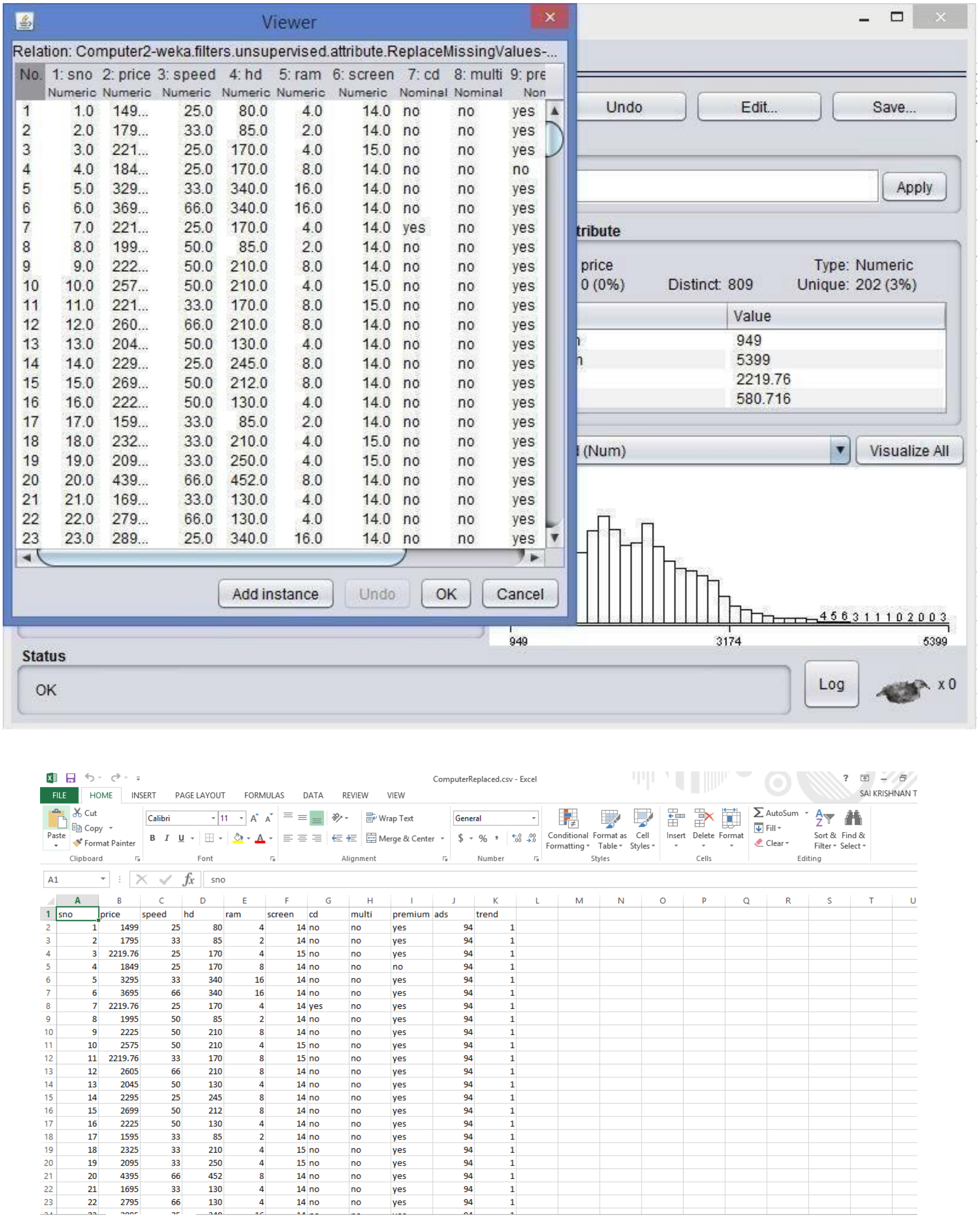
Attributes

No.		Name
1	□	sno
2	□	price
3	□	speed
4	□	hd
5	□	ram
6	□	screen
7	□	cd

Class: trend (Num)

Status

OK



Viewer

Relation: Computer2-weka.filters.unsupervised.attribute.ReplaceMissingValues-...

No.	1: sno Numeric	2: price Numeric	3: speed Numeric	4: hd Numeric	5: ram Numeric	6: screen Numeric	7: cd Nominal	8: multi Nominal	9: pre Non
1	1.0	149...	25.0	80.0	4.0	14.0	no	no	yes
2	2.0	179...	33.0	85.0	2.0	14.0	no	no	yes
3	3.0	221...	25.0	170.0	4.0	15.0	no	no	yes
4	4.0	184...	25.0	170.0	8.0	14.0	no	no	no
5	5.0	329...	33.0	340.0	16.0	14.0	no	no	yes
6	6.0	369...	66.0	340.0	16.0	14.0	no	no	yes
7	7.0	221...	25.0	170.0	4.0	14.0	yes	no	yes
8	8.0	199...	50.0	85.0	2.0	14.0	no	no	yes
9	9.0	222...	50.0	210.0	8.0	14.0	no	no	yes
10	10.0	257...	50.0	210.0	4.0	15.0	no	no	yes
11	11.0	221...	33.0	170.0	8.0	15.0	no	no	yes
12	12.0	260...	66.0	210.0	8.0	14.0	no	no	yes
13	13.0	204...	50.0	130.0	4.0	14.0	no	no	yes
14	14.0	229...	25.0	245.0	8.0	14.0	no	no	yes
15	15.0	269...	50.0	212.0	8.0	14.0	no	no	yes
16	16.0	222...	50.0	130.0	4.0	14.0	no	no	yes
17	17.0	159...	33.0	85.0	2.0	14.0	no	no	yes
18	18.0	232...	33.0	210.0	4.0	15.0	no	no	yes
19	19.0	209...	33.0	250.0	4.0	15.0	no	no	yes
20	20.0	439...	66.0	452.0	8.0	14.0	no	no	yes
21	21.0	169...	33.0	130.0	4.0	14.0	no	no	yes
22	22.0	279...	66.0	130.0	4.0	14.0	no	no	yes
23	23.0	289...	25.0	340.0	16.0	14.0	no	no	yes

price — Type: Numeric

0 (0%) — Distinct: 809 — Unique: 202 (3%)

	Value
n	949
n	5399
	2219.76
	580.716

OK

A1 — sno

	A	B	C	D	E	F	G	H	I	J	K
1	sno	price	speed	hd	ram	screen	cd	multi	premium	ads	trend
2	1	1499	25	80	4	14	no	no	yes	94	1
3	2	1795	33	85	2	14	no	no	yes	94	1
4	3	2219.76	25	170	4	15	no	no	yes	94	1
5	4	1849	25	170	8	14	no	no	no	94	1
6	5	3295	33	340	16	14	no	no	yes	94	1
7	6	3695	66	340	16	14	no	no	yes	94	1
8	7	2219.76	25	170	4	14	yes	no	yes	94	1
9	8	1995	50	85	2	14	no	no	yes	94	1
10	9	2225	50	210	8	14	no	no	yes	94	1
11	10	2575	50	210	4	15	no	no	yes	94	1
12	11	2219.76	33	170	8	15	no	no	yes	94	1
13	12	2605	66	210	8	14	no	no	yes	94	1
14	13	2045	50	130	4	14	no	no	yes	94	1
15	14	2295	25	245	8	14	no	no	yes	94	1
16	15	2699	50	212	8	14	no	no	yes	94	1
17	16	2225	50	130	4	14	no	no	yes	94	1
18	17	1595	33	85	2	14	no	no	yes	94	1
19	18	2325	33	210	4	15	no	no	yes	94	1
20	19	2095	33	250	4	15	no	no	yes	94	1
21	20	4395	66	452	8	14	no	no	yes	94	1
22	21	1695	33	130	4	14	no	no	yes	94	1
23	22	2795	66	130	4	14	no	no	yes	94	1

Fig: 3.6 Replaced values

Exercise

1. Create your own dataset having missing values included.