

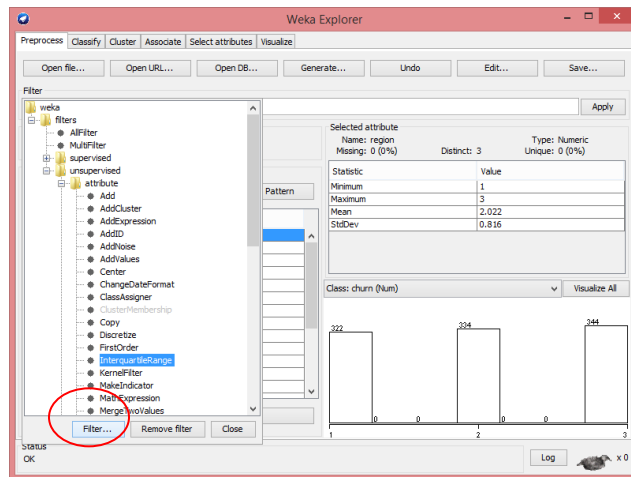
Lab Exercise One

Data Preprocessing with WEKA Explorer

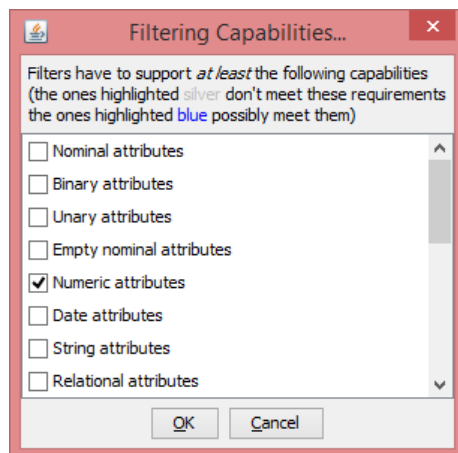
Using Filters to handle outliers and extreme values

Unsupervised Attribute Filter – InterquartileRange: This filter adds new attributes that indicate whether the values of instances can be considered **outliers or extreme** values.

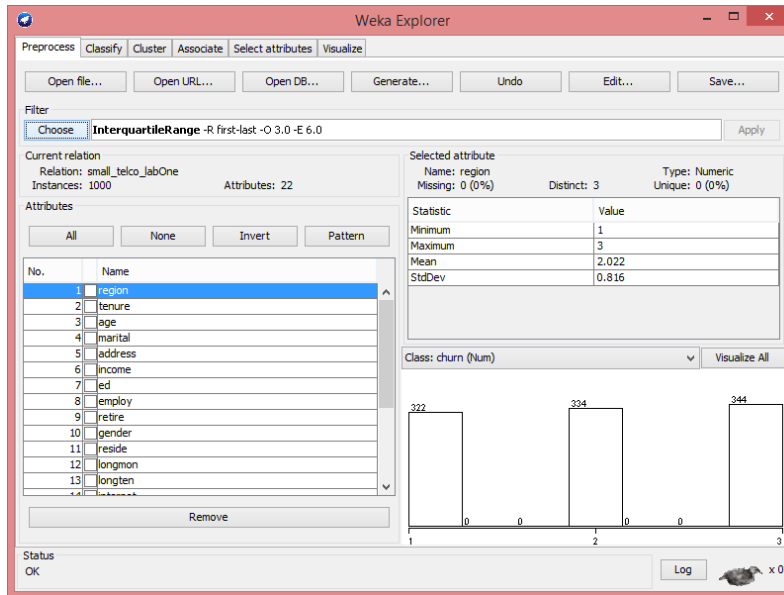
1. Open the dataset – **small_telco_labOne**. Perform the replacing missing values step with the filter – **ReplaceMissingValues**. Please pay attention that there are total **22 attributes** in the dataset.
2. Then Click **Choose** button under **Filter**. Click **Filter** button at the bottom of the drop-down window.



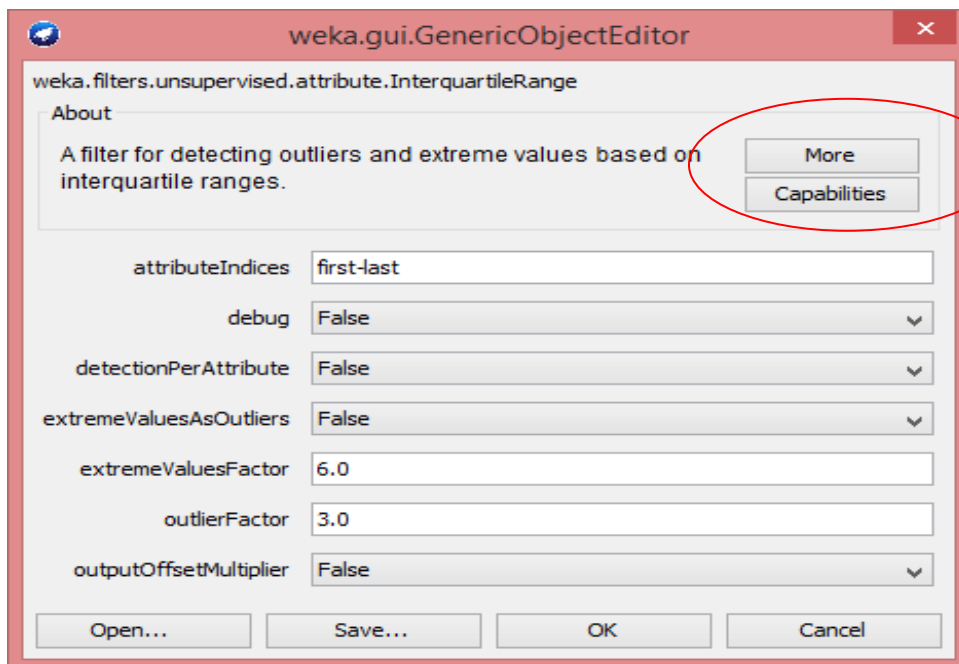
3. A window called Filtering Capabilities opens. This window shows what kind of attributes that filters support. Make sure that only **Numeric Attributes** and **Numeric Class** are checked. Click **OK**.

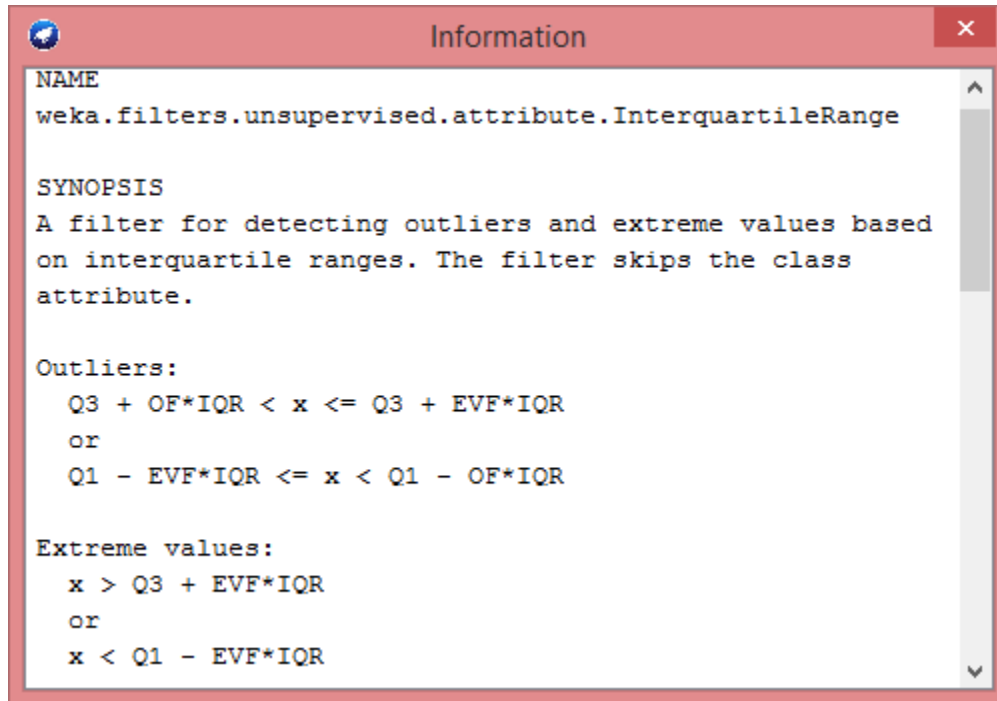


4. Choose **InterquartileRange** filter from the drop down list of **unsupervised attribute** filter list.



5. Left-click the box of the filter, the properties window shows. Click **More** button to show more information about this filter. The factors are used to define extreme values and outlier.





- Click **Apply** button at the end of the filter box. You will find two extra attributes are generated. These two attributes flag an instance as an outlier or extreme if any of its attribute values are deemed outliers or extreme.

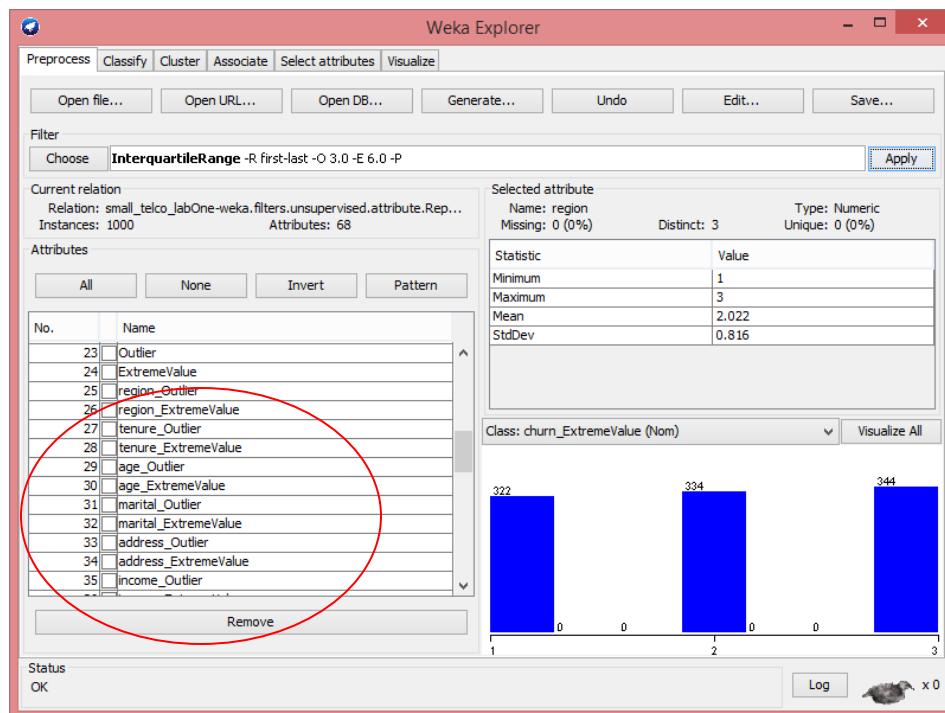
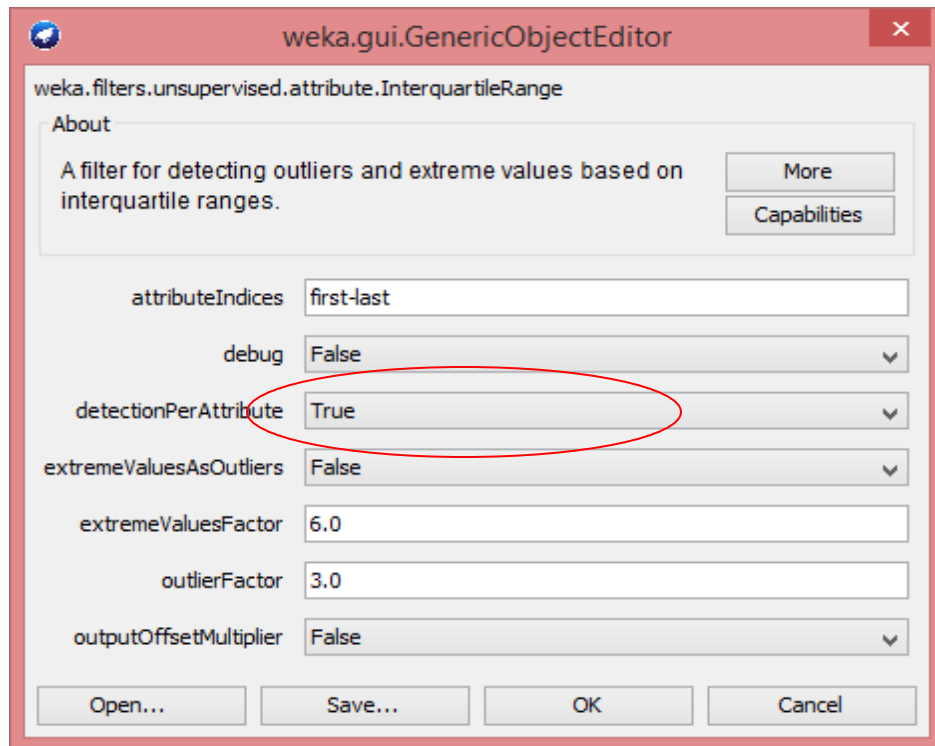
Viewer

Relation: small_telco_labOne-weka.filters.unsupervised.attribute.ReplaceMissingValues-wek...

g	logequi	logcard	logwire	lninc	custcat	churn	Outlier	ExtremeValue
ic	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Nominal	Nominal
...	3.568...	2.014...	3.598...	4.158...	1.0	1.0	no	no
...	3.568...	2.72458	3.575...	4.912...	4.0	1.0	no	yes
...	3.568...	3.409...	3.598...	4.75359	3.0	0.0	no	no
...	3.568...	2.854...	3.598...	3.496...	1.0	1.0	no	no
55	3.568...	2.854...	3.598...	3.401...	3.0	0.0	no	no
81	3.568...	2.60269	3.598...	4.356...	3.0	0.0	no	no
...	3.568...	2.169...	3.598...	2.944...	2.0	1.0	no	no
...	3.914...	3.146...	4.172...	4.330...	4.0	0.0	no	yes
...	3.568...	2.484...	3.598...	5.111...	3.0	0.0	no	no
...	3.568...	2.80336	3.598...	4.276...	2.0	0.0	no	no
...	3.263...	2.854...	3.598...	4.828...	1.0	1.0	no	yes
...	3.568...	3.167...	3.598...	4.382...	3.0	0.0	no	no
...	3.568...	3.731...	3.598...	3.610...	1.0	0.0	no	no
...	3.843...	2.854...	4.111...	4.744...	4.0	1.0	no	yes
...	3.568...	2.854...	3.598...	3.218...	1.0	0.0	no	no
...	3.409...	2.420...	3.598...	4.317...	2.0	0.0	no	yes
92	3.443...	3.401...	3.598...	5.087...	3.0	0.0	no	yes
...	3.568...	2.854...	3.598...	3.89182	3.0	0.0	no	no
79	3.568...	2.854...	3.598...	2.995...	1.0	0.0	no	no
...	3.873...	3.188...	3.64545	4.343...	4.0	1.0	no	yes
...	3.520...	2.854...	2.928...	2.772...	2.0	1.0	no	yes
...	3.568...	3.091...	3.598...	4.787...	1.0	0.0	no	no
...	3.903...	3.286...	3.939...	4.615...	4.0	0.0	no	yes

Undo OK Cancel

7. If we change the option for **InterquartileRange** filter, detectionPerAttribute from False to True, an outlier-extreme indicator pair for each attribute is generated.



8. You could click each generated attribute to check the outlier and extreme values for original attribute. Remove those attribute indicator without outlier or extreme values with **Remove** button.

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **NumericCleaner** -min -1.7976931348623157E308 -min-default -1.7976931348623157E308 -max 1.7976931348623157E308 -max: Apply

Current relation: Relation: small_telco_labOne-weka.filters.unsupervised.attribute.Rep... Instances: 1000 Attributes: 62

Attributes: All None Invert Pattern

No.	Name
50	longten_ExtremeValue
51	<input checked="" type="checkbox"/> internet_Outlier
52	<input checked="" type="checkbox"/> internet_ExtremeValue
53	<input checked="" type="checkbox"/> ebill_Outlier
54	<input checked="" type="checkbox"/> ebill_ExtremeValue
55	<input checked="" type="checkbox"/> loglong_Outlier
56	<input checked="" type="checkbox"/> loglong_ExtremeValue
57	<input checked="" type="checkbox"/> logequi_Outlier
58	<input type="checkbox"/> logequi_ExtremeValue
59	<input type="checkbox"/> logcard_Outlier
60	<input type="checkbox"/> logcard_ExtremeValue
61	<input type="checkbox"/> logwire_Outlier
62	<input type="checkbox"/> logwire_ExtremeValue

Remove

Selected attribute: Name: internet_Outlier Type: Nominal Missing: 0 (0%) Distinct: 1 Unique: 0 (0%)

No.	Label	Count
1	no	1000
2	yes	0

Class: logwire_ExtremeValue (Nom) Visualize All

Status: OK Log x 0

Weka Explorer

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

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **NumericCleaner** -min -1.7976931348623157E308 -min-default -1.7976931348623157E308 -max 1.7976931348623157E308 -max: Apply

Current relation: Relation: small_telco_labOne-weka.filters.unsupervised.attribute.Rep... Instances: 1000 Attributes: 34

Attributes: All None Invert Pattern

No.	Name
22	churn
23	Outlier
24	ExtremeValue
25	income_Outlier
26	income_ExtremeValue
27	retire_ExtremeValue
28	longmon_Outlier
29	longmon_ExtremeValue
30	longten_Outlier
31	longten_ExtremeValue
32	logequi_ExtremeValue
33	logcard_Outlier
34	<input checked="" type="checkbox"/> logwire_ExtremeValue

Remove

Selected attribute: Name: logwire_ExtremeValue Type: Nominal Missing: 0 (0%) Distinct: 2 Unique: 0 (0%)

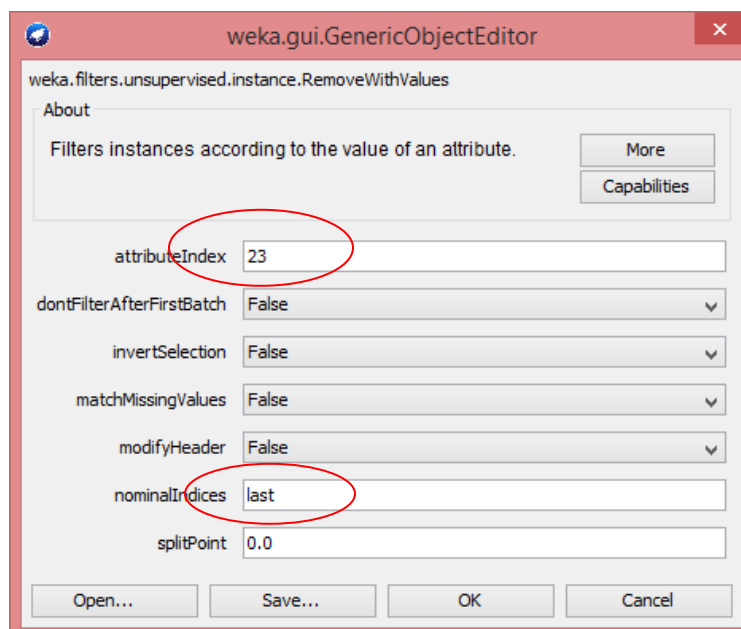
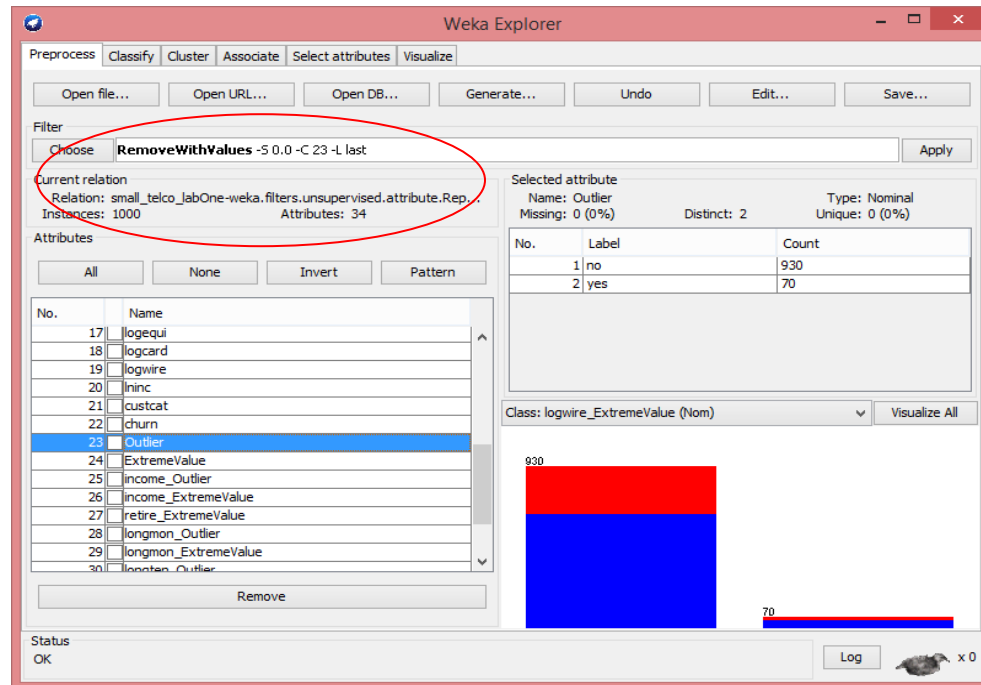
No.	Label	Count
1	no	704
2	yes	296

Class: logwire_ExtremeValue (Nom) Visualize All

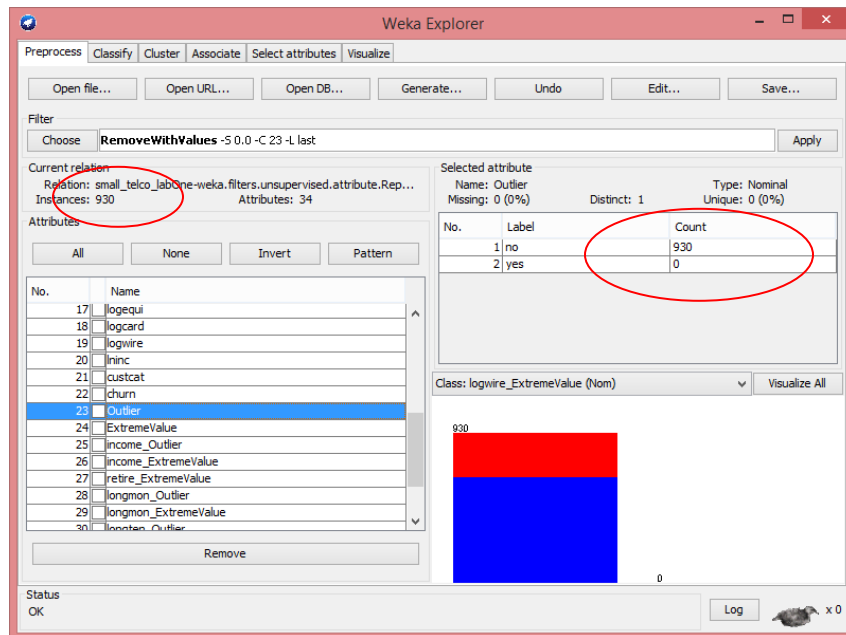
Status: OK Log x 0

Unsupervised Instance Filter – RemoveWithValues: This filter removes instances according to the values of an attribute.

1. After we find out which instances having outliers or extreme values, we could remove those instances with outliers completely from the dataset. Choose **RemoveWithValues** from the drop-down list of **unsupervised instance** Filter. Then left-click the box of the filter. Since **outlier attribute** is indexed as 23 and “yes” value is the **last** nominal value of this attribute, change the options of the filter accordingly.



2. Then click **Apply** after confirming the changes. 70 instances are removed from the dataset and Outlier attribute has no Yes values.



3. You could also remove instances according to the outlier-attribute-pair indicators in the same way.