Program – 4

AIM: Explain missing value treatment and outlier treatment in WEKA on sample dataset.

Introduction and Theory

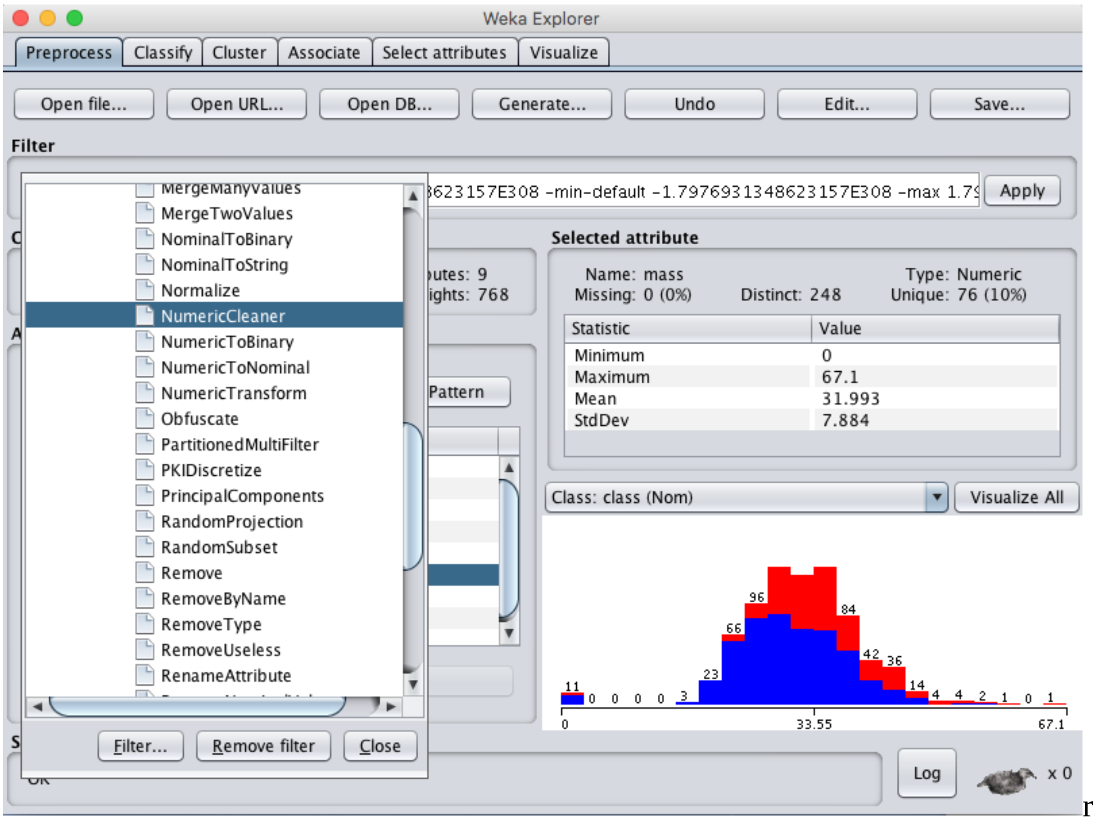
Data is rarely clean and often you can have corrupt or missing values. It is important to identify, mark and handle missing data when developing machine learning models in order to get the very best performance.

In most cases it is normal to find some values missing in the data, in other cases there may be some data elements whose values don’t match with any other data points, such values are called outliers, normally any value outside of 1.5 quantiles of the data mean is taken as an outlier.

Procedure

Mark Missing Values

1. Open the Weka Explorer.
2. Load the Pima Indians onset of diabetes dataset.
3. Click the “Choose” button for the Filter and select NumericalCleaner, it us under unsupervized.attribute.NumericalCleaner



623157E308 -min-default -1.7976931348623157E308 -max 1.79

utes: 9
ights: 768

Selected attribute

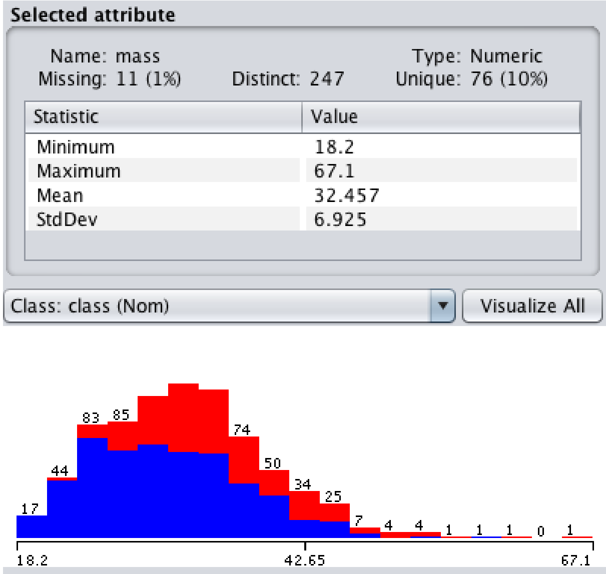
Name: mass Type: Numeric
Missing: 0 (0%) Distinct: 248 Unique: 76 (10%)

Statistic	Value
Minimum	0
Maximum	67.1
Mean	31.993
StdDev	7.884

r.

4. Click on the filter to configure it.

5. Set the attributeIndicies to 6, the index of the mass attribute.
6. Set minThreshold to 0.1E-8 (close to zero), which is the minimum value allowed for the attribute.
7. Set minDefault to NaN, which is unknown and will replace values below the threshold.
8. Click the "OK" button on the filter configuration.
9. Click the "Apply" button to apply the filter.



Selected attribute

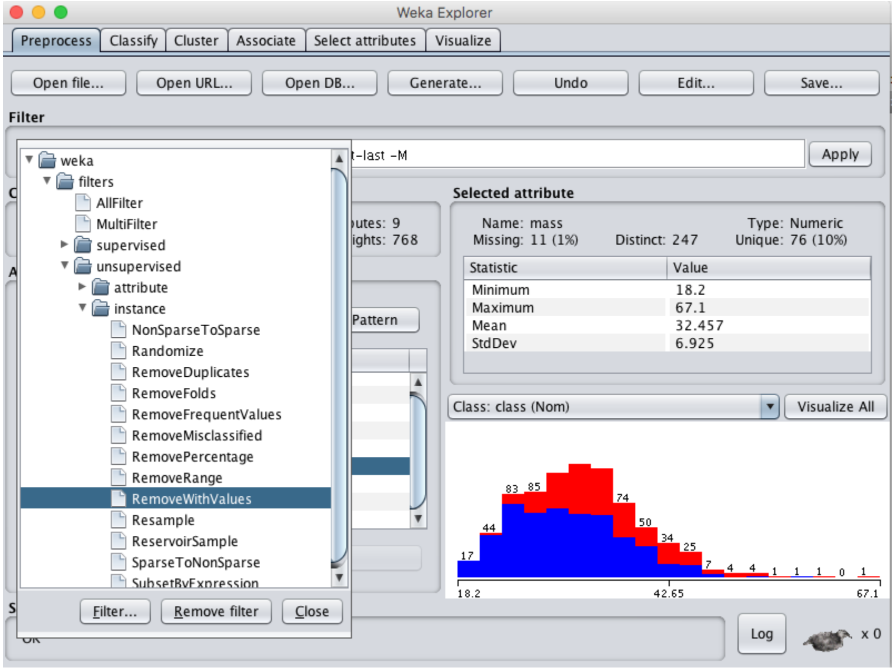
Name: mass
Type: Numeric
Missing: 11 (1%)
Distinct: 247
Unique: 76 (10%)

Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925

Class: class (Nom)

Removing missing values

1. Click the "Choose" button for the Filter and select RemoveWithValues, it us under unsupervized.instance.RemoveWithValues.



-last -M

utes: 9
ights: 768

Selected attribute

Name: mass
Type: Numeric
Missing: 11 (1%)
Distinct: 247
Unique: 76 (10%)

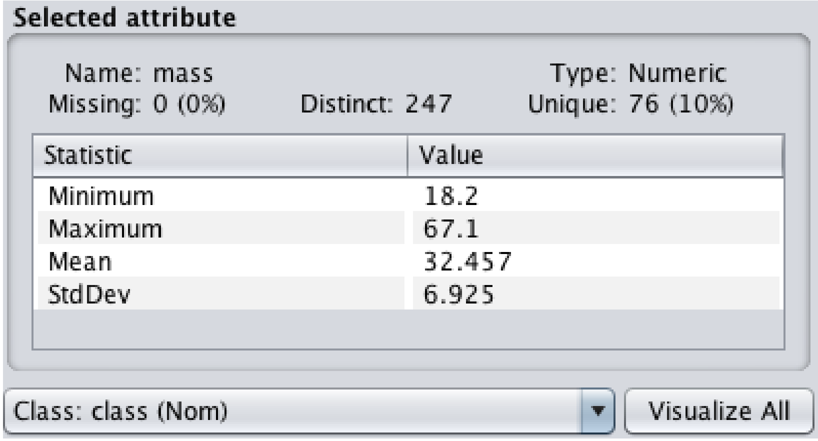
Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925

Class: class (Nom)

2. Click on the filter to configure it.

Program – 4

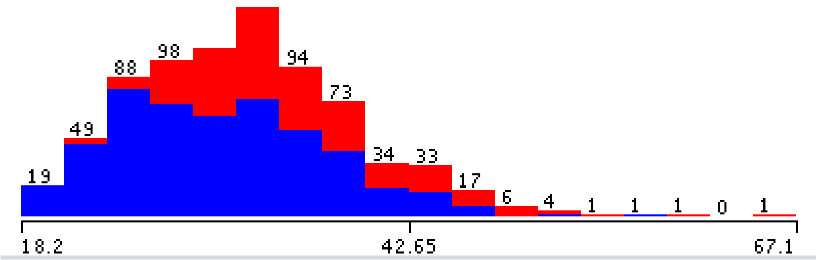
3. Set the *attributeIndicies* to 6, the index of the mass attribute.
4. Set *matchMissingValues* to “True”.
5. Click the “OK” button to use the configuration for the filter.
6. Click the “Apply” button to apply the filter.



Selected attribute

Name: mass
Missing: 0 (0%)
Distinct: 247
Type: Numeric
Unique: 76 (10%)

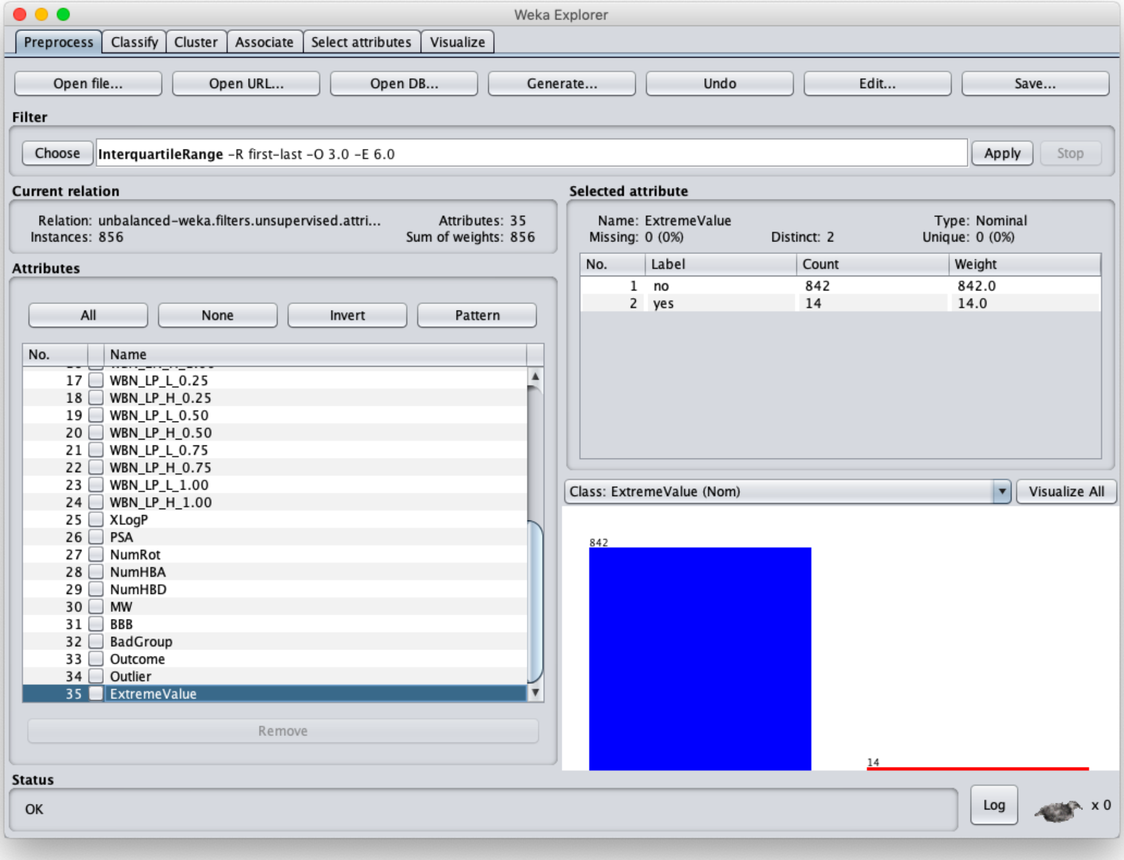
Statistic	Value
Minimum	18.2
Maximum	67.1
Mean	32.457
StdDev	6.925



Outlier treatment

1. Open the Weka Explorer.
2. Load the Unbalanced dataset.
3. Click the “Choose” button for the Filter and select InterquantileRange, it us under unsupervized.attribute.InterquantileRange
4. Click on the filter to configure it.
5. Click the “OK” button on the filter configuration.
6. Click the “Apply” button to apply the filter.

Program – 4



Filter

InterquartileRange -R first-last -O 3.0 -E 6.0

Current relation

Relation: unbalanced-weka.filters.unsupervised.attri...   Attributes: 35
Instances: 856   Sum of weights: 856

Attributes

No.	Name
17	WBN_LP_L_0.25
18	WBN_LP_H_0.25
19	WBN_LP_L_0.50
20	WBN_LP_H_0.50
21	WBN_LP_L_0.75
22	WBN_LP_H_0.75
23	WBN_LP_L_1.00
24	WBN_LP_H_1.00
25	XLogP
26	PSA
27	NumRot
28	NumHBA
29	NumHBD
30	MW
31	BBB
32	BadGroup
33	Outcome
34	Outlier
35	ExtremeValue

Selected attribute

Name: ExtremeValue   Type: Nominal
Missing: 0 (0%)   Distinct: 2   Unique: 0 (0%)

No.	Label	Count	Weight
1	no	842	842.0
2	yes	14	14.0

Class: ExtremeValue (Nom)

Status

OK

Findings and Learnings:

1. Missing value treatment and outlier removal are very important steps to be performed before data analysis and statistical modeling.
2. Weka Software makes it easy and simple to visualize data and apply appropriate treatments to remove the anomalies.