Department of Mathematics
University of Calabria

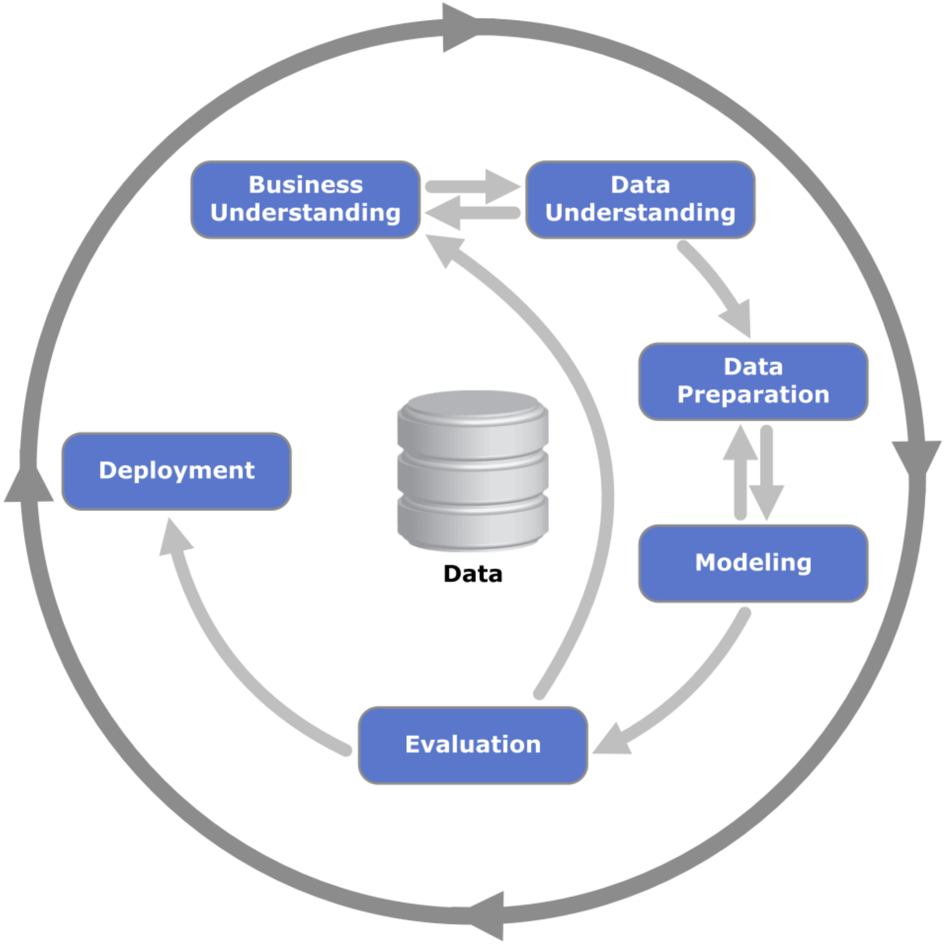
Data Warehouse and Data Mining

Module II – Data Mining

Data Preparation

Ph.D. Ettore Ritacco

The Knowledge Discovery Process (CRISP-DM)



Business Understanding

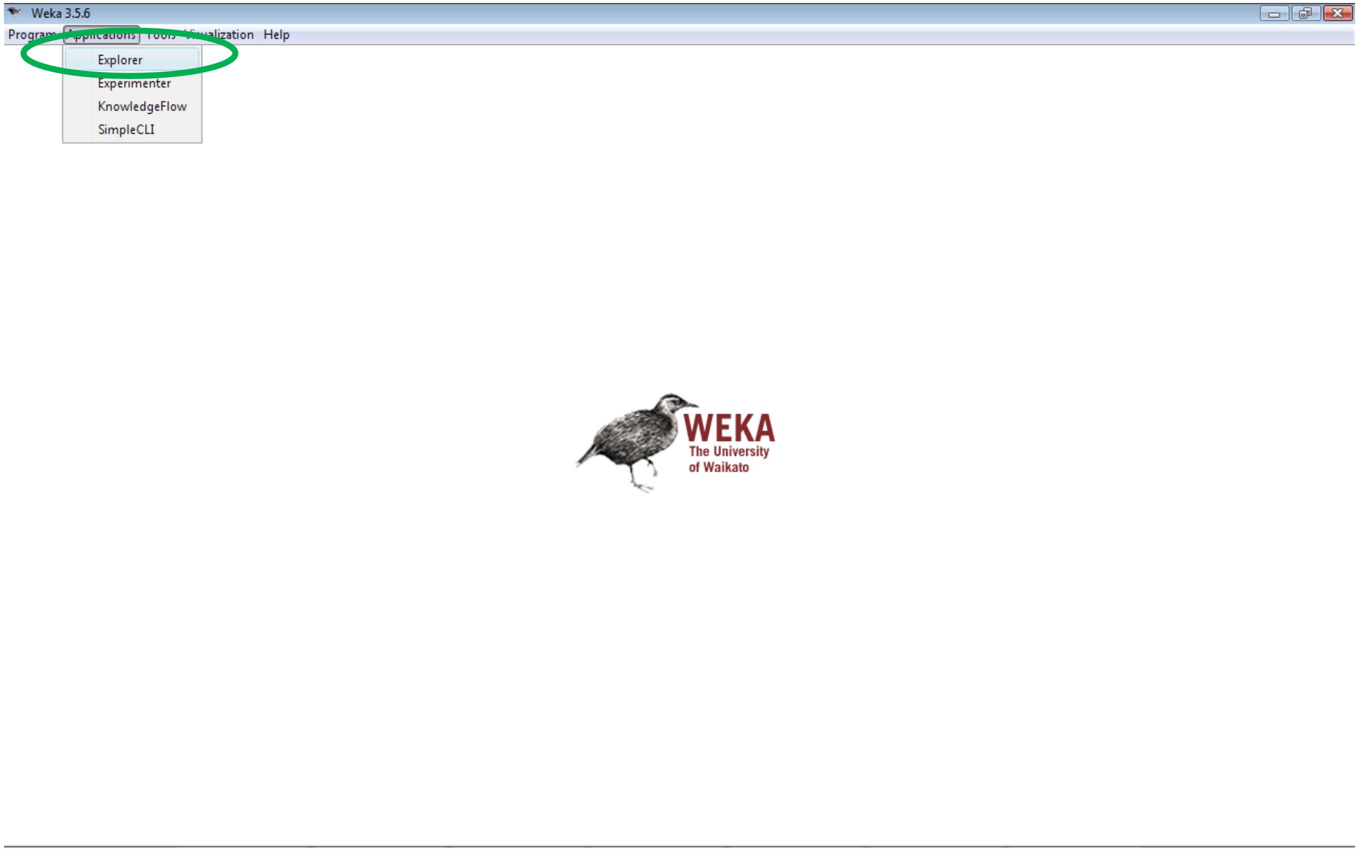
- You are a doctor of a medical division, who collected some data from your patients
- All of them contracted the same disease
- The therapy consists in 5 different and exclusive cures, according to the patient condition
- Does an automatic procedure, for the cure assignment, exist?

Data Understanding

Attribute	Description
Instance_number	Incremental tuple ID (INTEGER)
ID	Patient’s ID (INTEGER)
Age	Patient’s age (INTEGER)
Sex	Patient’s gender: F or M
BP	Blood Pressure: HIGH, NORMAL or LOW
Cholesterol	Concentration of cholesterol in the blood: NORMAL or HIGH
Na	Concentration of sodium in the blood (REAL)
K	Concentration of potassium in the blood (REAL)
Drug	The chosen cure: drugY, drugC, drugX, drugA, drugB

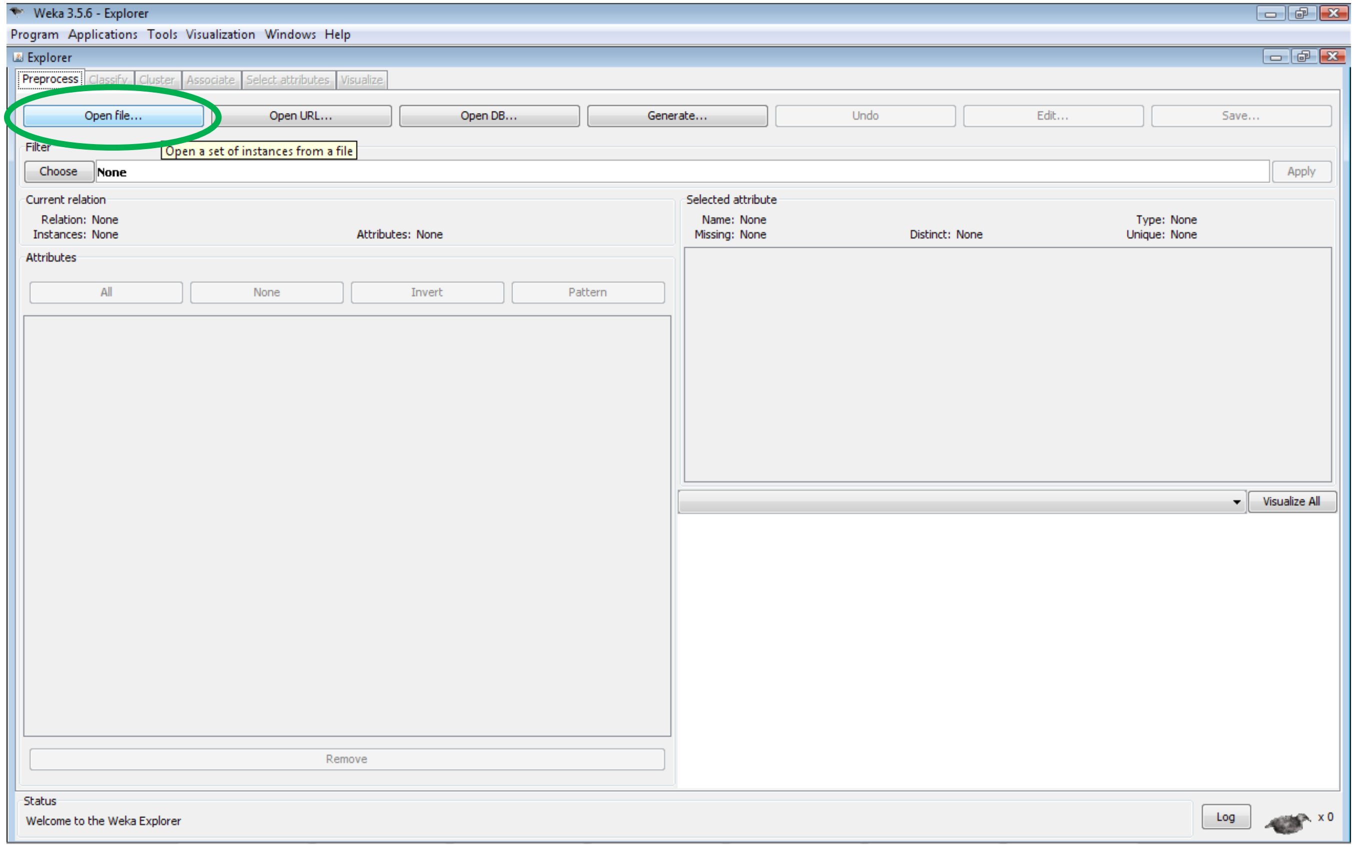
Data Understanding

- Data acquisition in Weka:



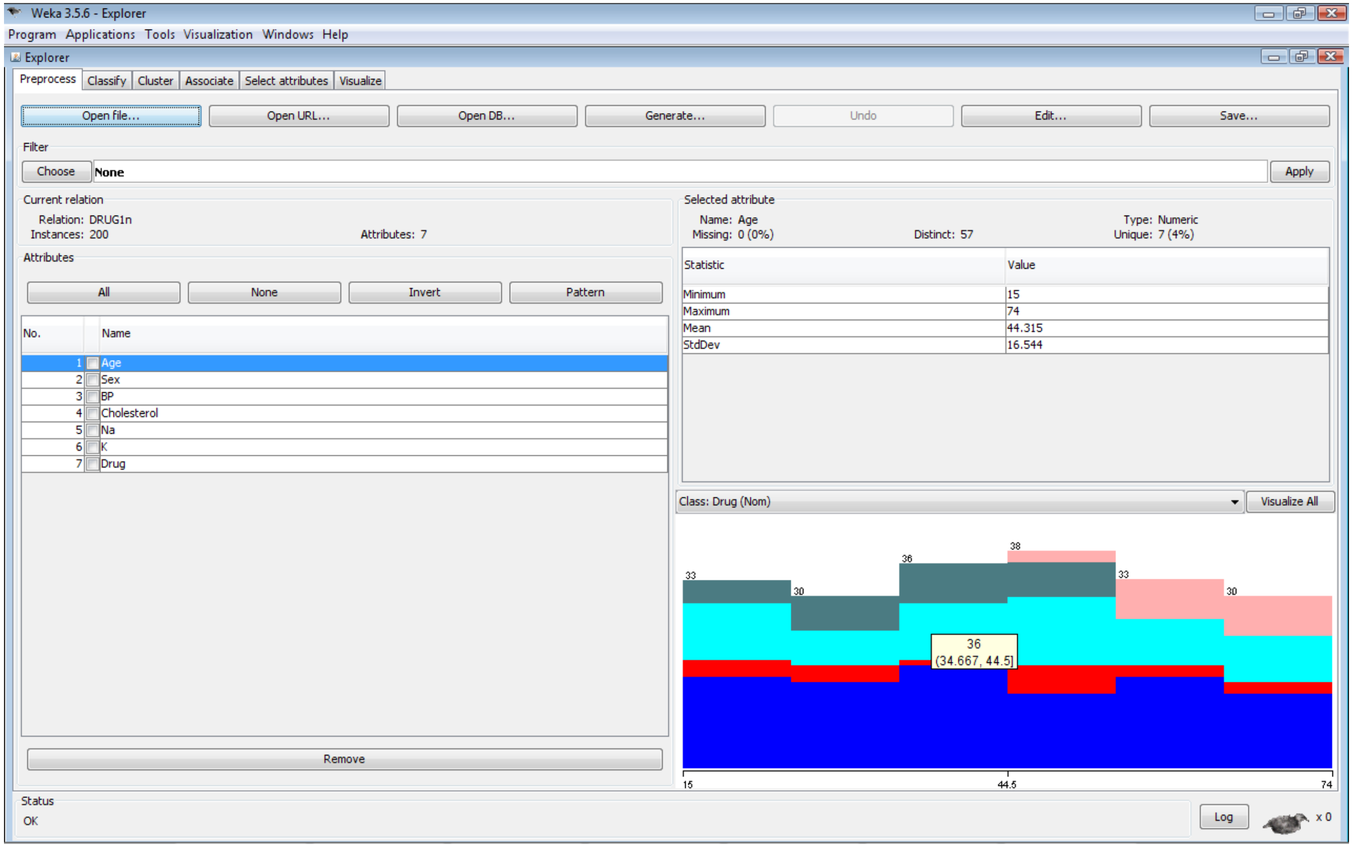
Data Understanding

- Data acquisition in Weka:



Data Understanding

- Data acquisition in Weka:


Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose
Apply
All None Invert Pattern
Remove
Class: Drug (Nom)
Visualize All
33
30
36
38
33
30
36
[34.667, 44.5]
15
44.5
74
Log
x 0

Filter: None

Current relation
Relation: DRUG1n
Instances: 200
Attributes: 7

Selected attribute
Name: Age
Type: Numeric
Missing: 0 (0%)
Distinct: 57
Unique: 7 (4%)

Attributes

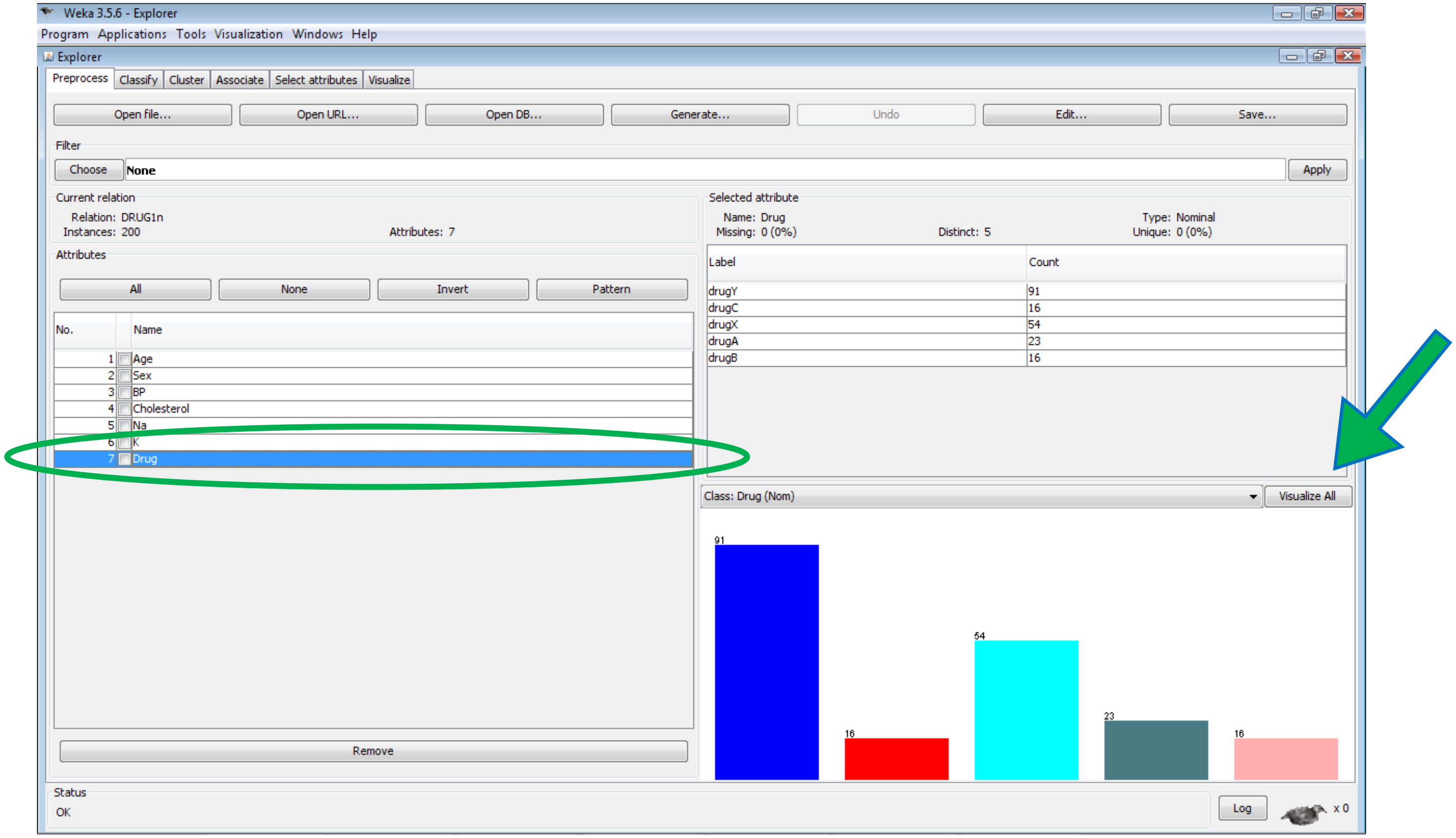
No.	Name
1	□ Age
2	□ Sex
3	□ BP
4	□ Cholesterol
5	□ Na
6	□ K
7	□ Drug

Statistic	Value
Minimum	15
Maximum	74
Mean	44.315
StdDev	16.544

Status
OK

Data Understanding

- Target attribute (the class attribute): Drug



Filter: None

Current relation
Relation: DRUG1n
Instances: 200
Attributes: 7

Attributes

No.		Name
1	□	Age
2	□	Sex
3	□	BP
4	□	Cholesterol
5	□	Na
6	□	K
7	□	Drug

Selected attribute
Name: Drug
Type: Nominal
Missing: 0 (0%)
Distinct: 5
Unique: 0 (0%)

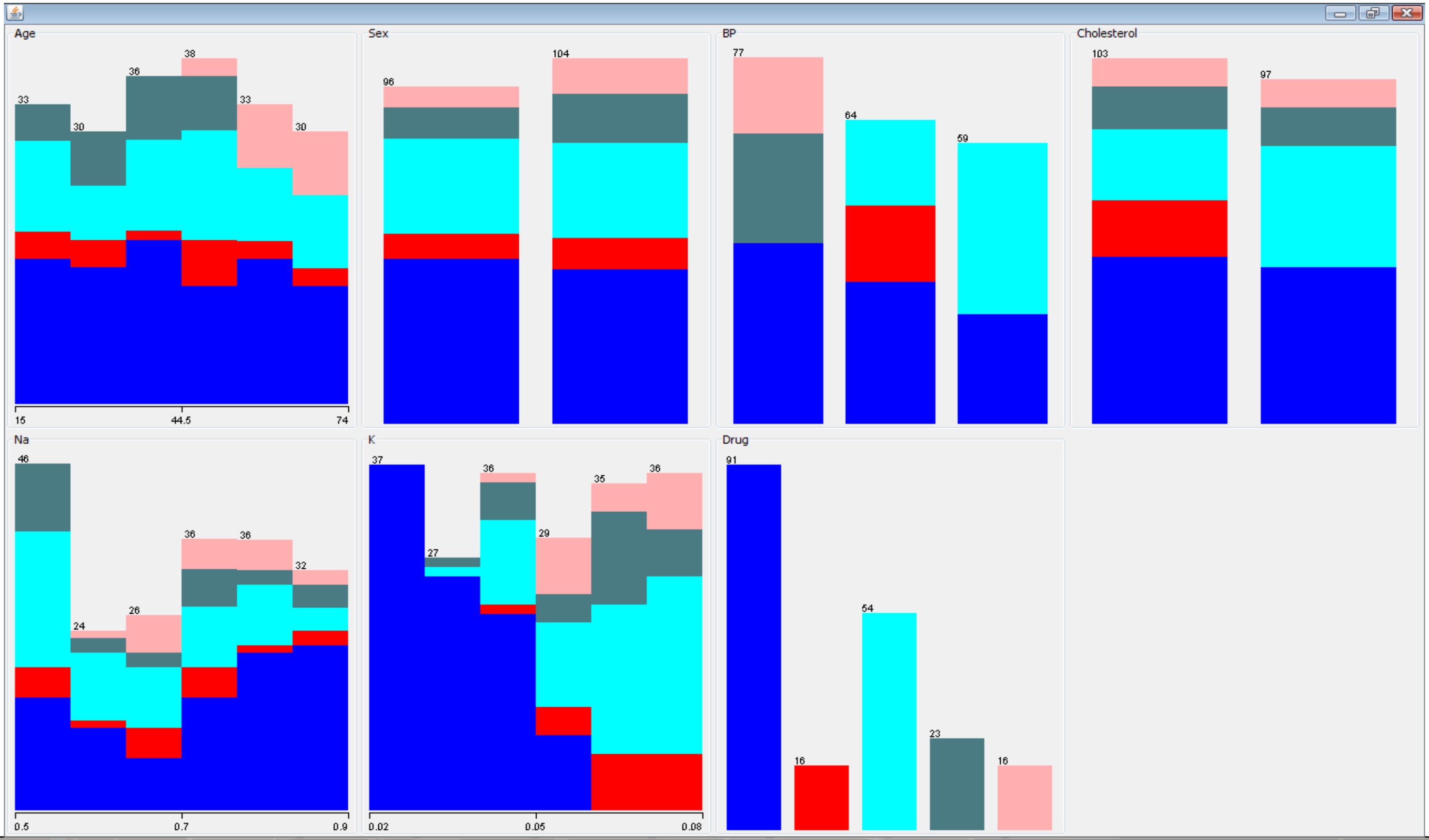
Label	Count
drugY	91
drugC	16
drugX	54
drugA	23
drugB	16

Class: Drug (Nom)

Status
OK

Data Understanding

- Distributions: comments?

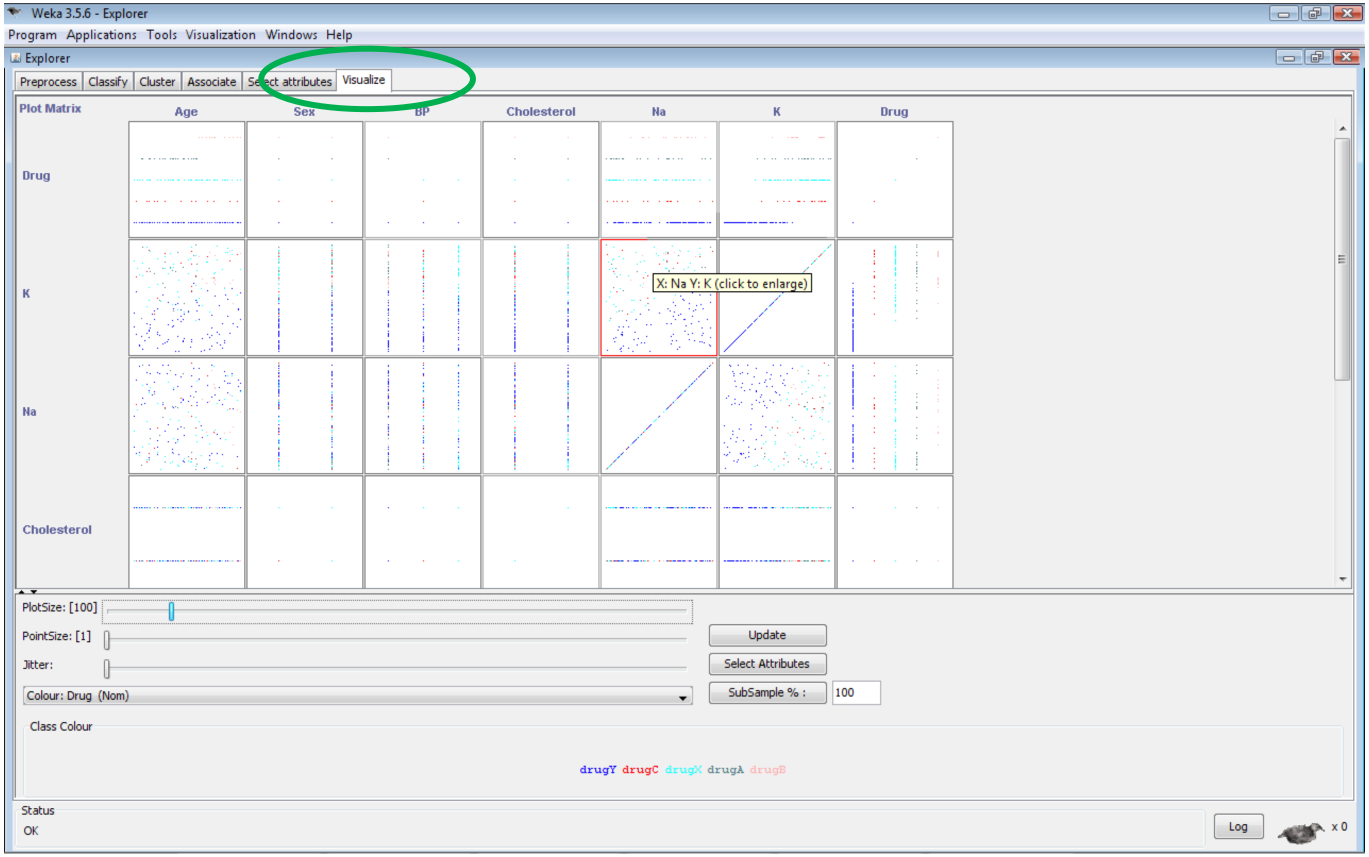


Data Understanding

- When age is greater then (about) 44 years, the class drugB starts to appear
- The sex and the concentration of sodium seem to not affect the class distribution
- Some cures are provided according to the values of blood pressure and cholesterol
- When the concentration of potassium is lower than (about) 0.05 the dominant class is drugY
- Data are unbalanced:
  - There is one dominant class: drugY (the blue one)

Data Understanding

- Scatter plot



PlotSize: [100]

PointSize: [1]

Jitter:

Colour: Drug (Nom)

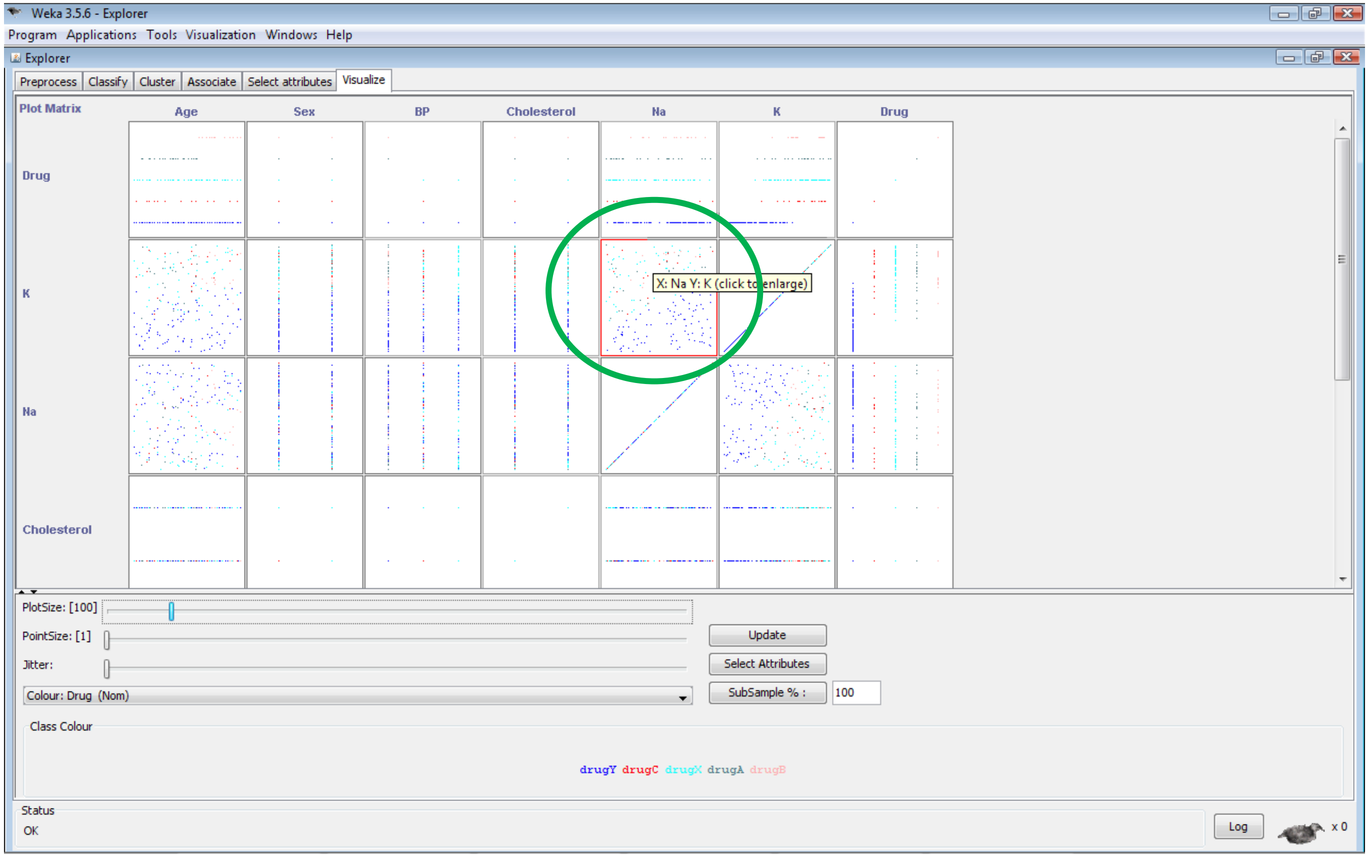
SubSample % : 100

Status

OK

Data Understanding

- Scatter plot


Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Plot Matrix
Age
Sex
BP
Cholesterol
Na
K
Drug
Drug
K
Na
Cholesterol
X: Na Y: K (click to enlarge)
Update
Select Attributes
Class Colour
drugY drugC drugX drugA drugB
Log
x 0

PlotSize: [100]

PointSize: [1]

Jitter:

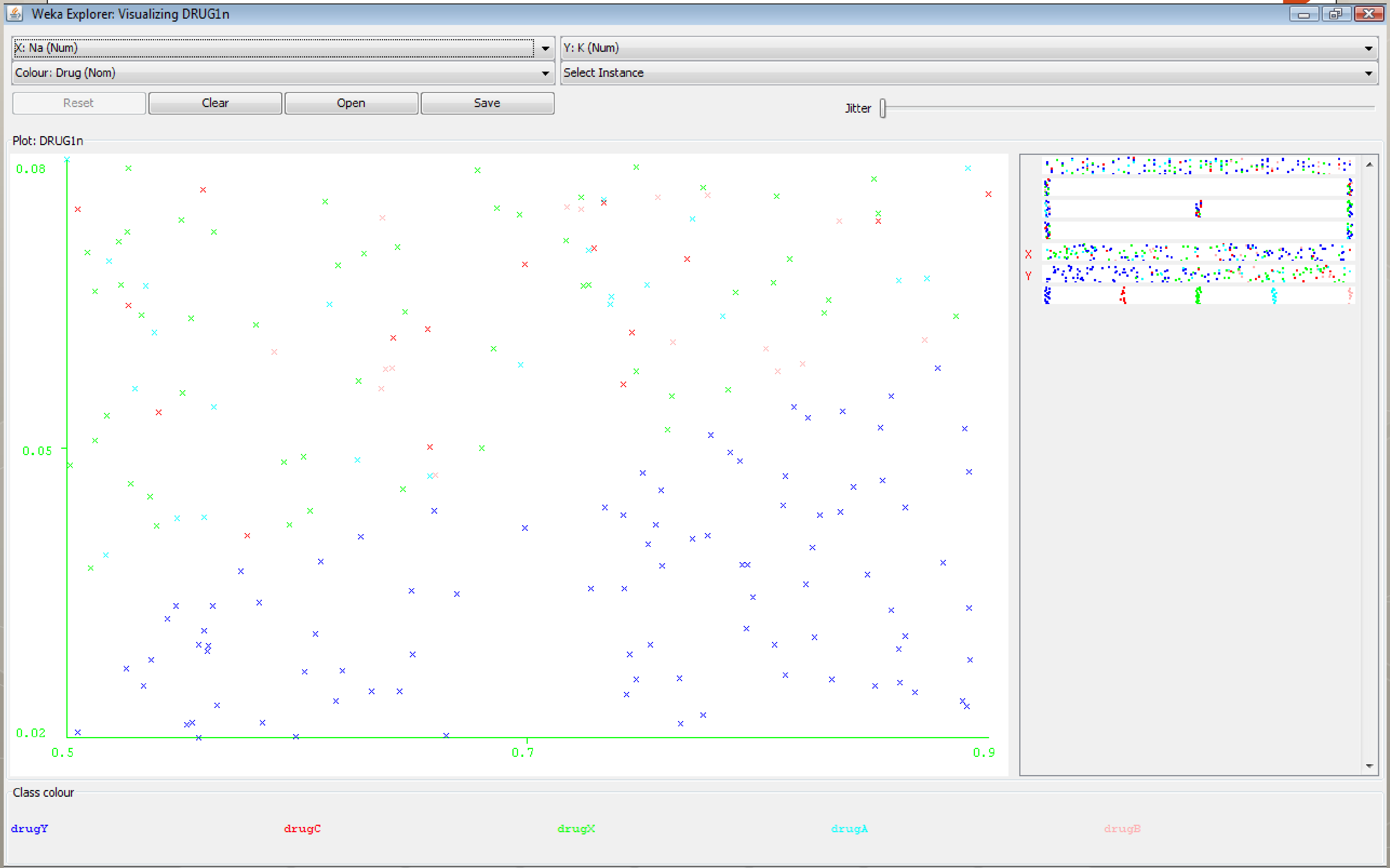
Colour: Drug (Nom)

SubSample % : 100

Status

OK

Data Understanding



Weka Explorer: Visualizing DRUG1n

X: Na (Num)

Y: K (Num)

Colour: Drug (Nom)

Select Instance

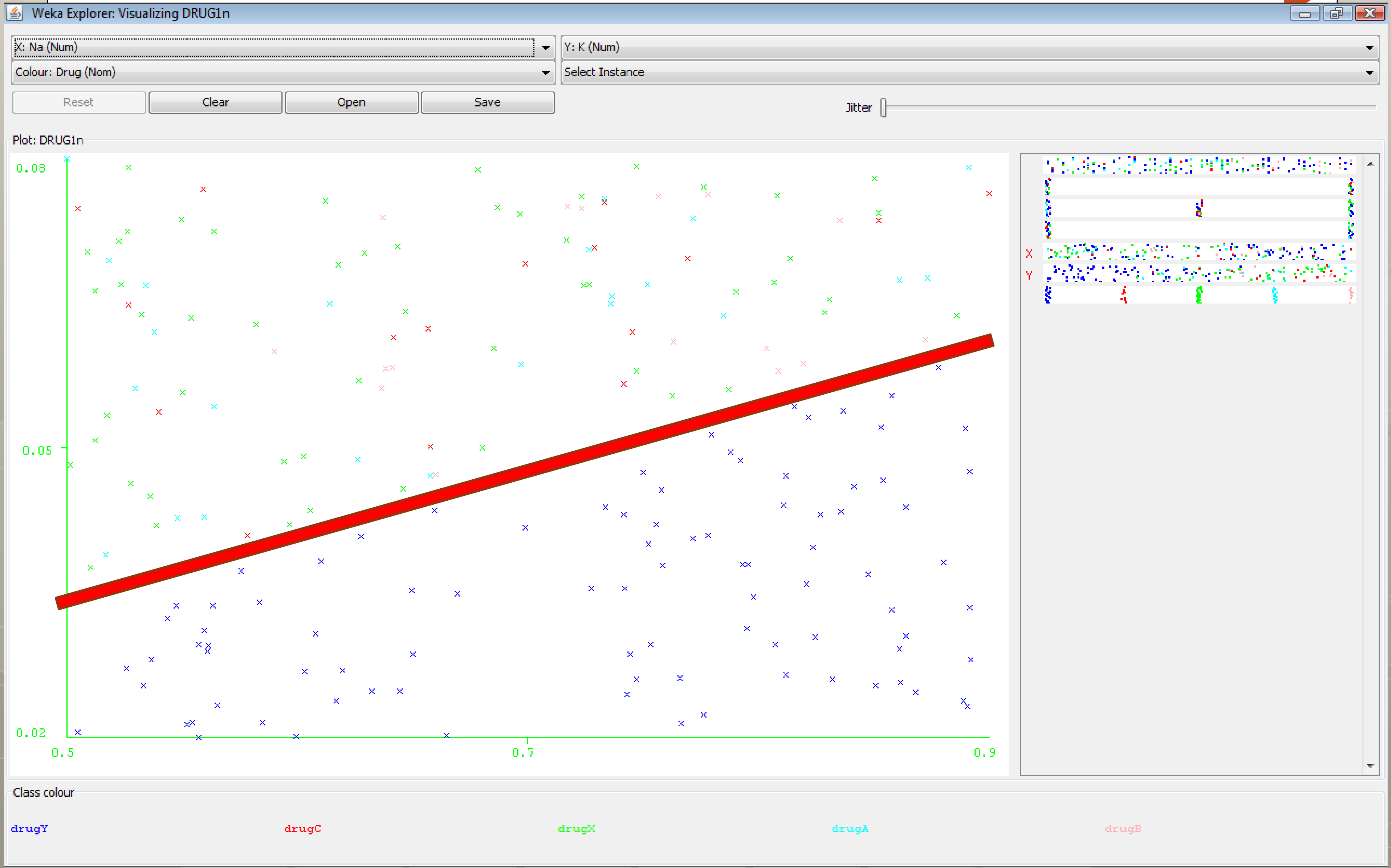
Data Understanding

- Comments?

Data Understanding

- A hyperplane can clearly separate one class (drugY, the blue one) from the others!

Data Understanding



Weka Explorer: Visualizing DRUG1n

X: Na (Num)

Y: K (Num)

Colour: Drug (Nom)

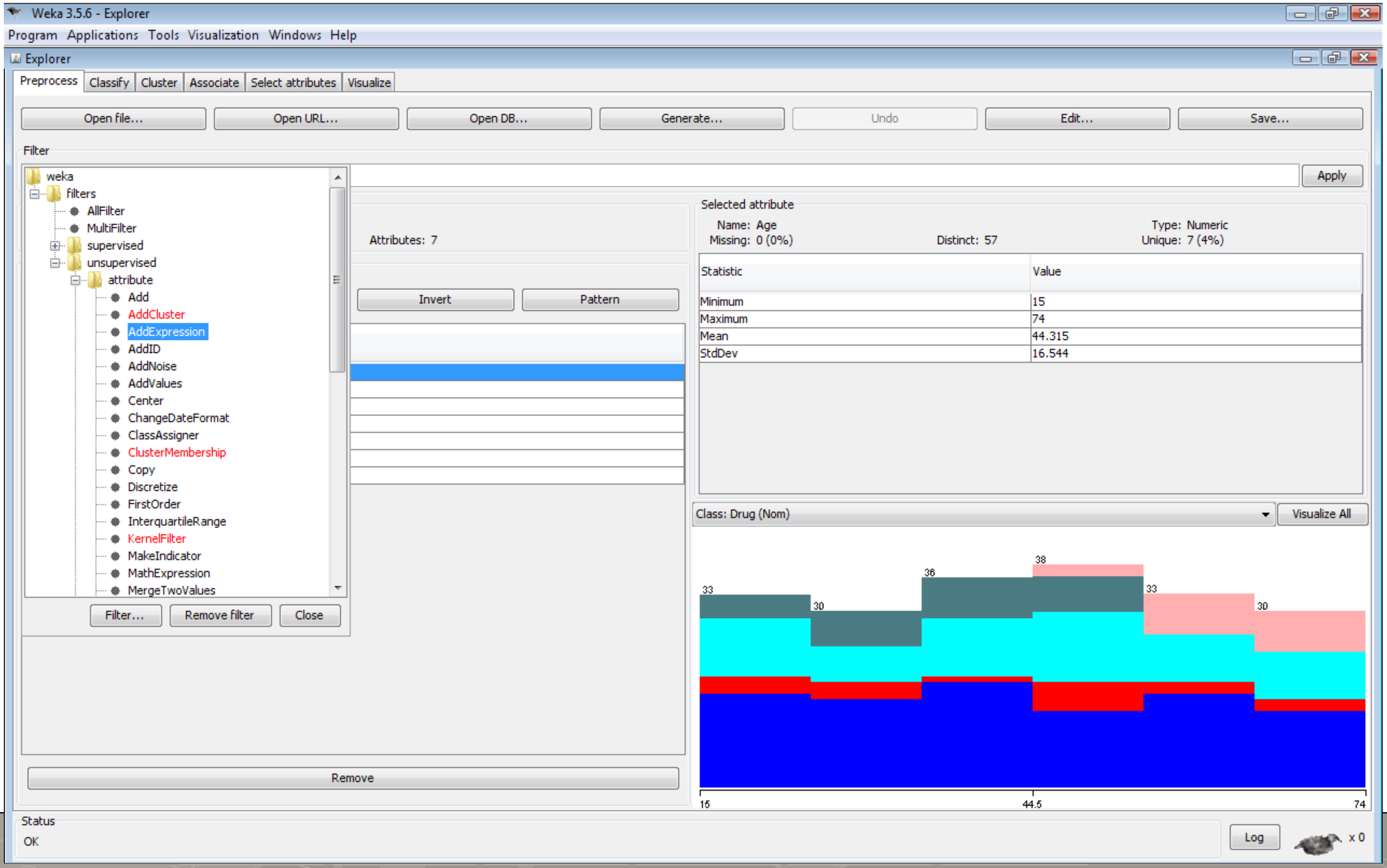
Select Instance

Data Preparation

- Hyperplane:
  - K = m * Na + q
- Let’s ignore the constant *q*, then
  - m = Na / K
- We can create a new attribute Na_su_K (Na out of K, in *italian*)

Data Preparation

- Choose the AddExpession filter



Attributes: 7

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 57
Type: Numeric
Unique: 7 (4%)

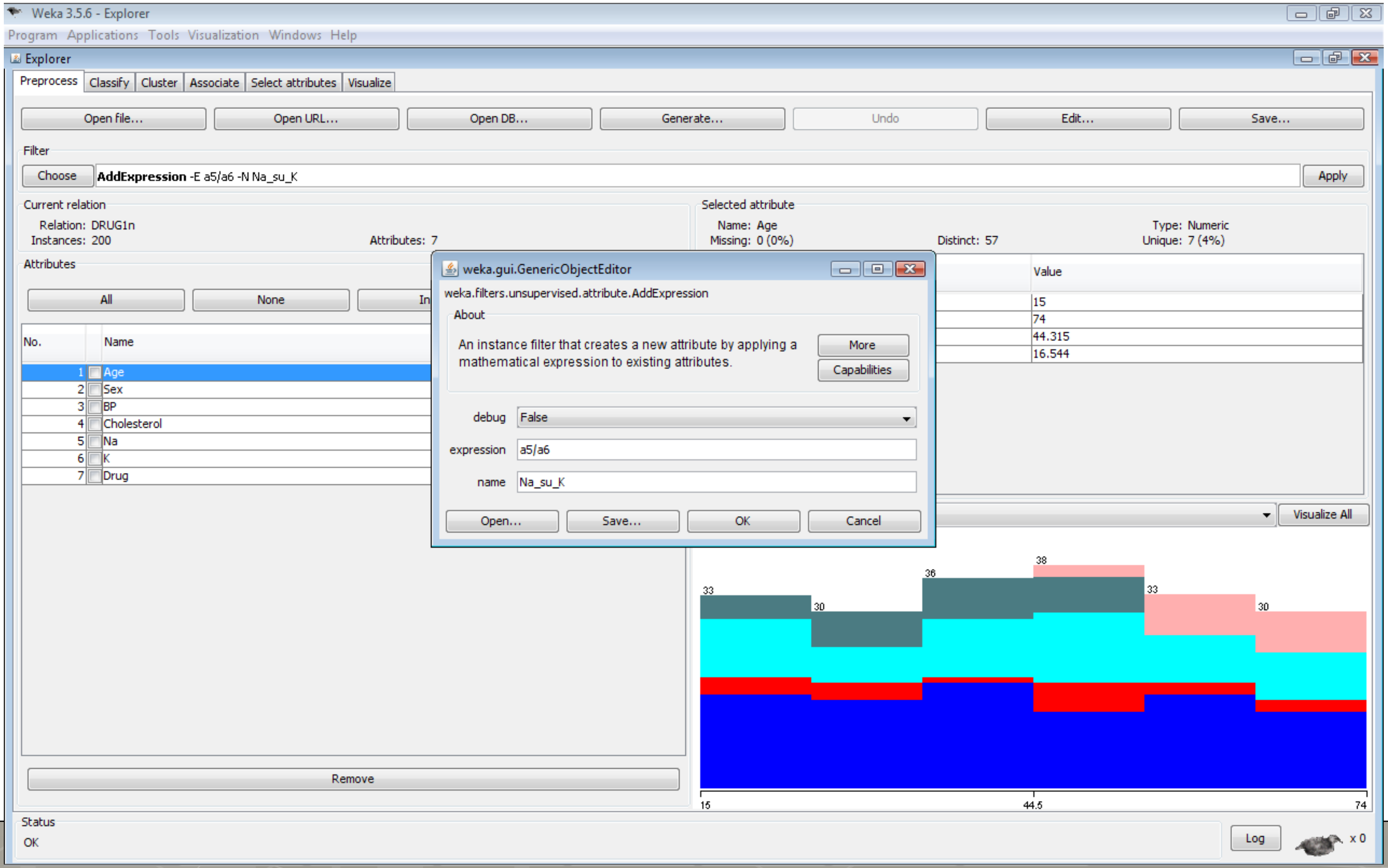
Statistic	Value
Minimum	15
Maximum	74
Mean	44.315
StdDev	16.544

Class: Drug (Nom)

OK

Data Preparation

- Configure the filter


Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Choose
Apply
All None In
Remove
Visualize All
More
Capabilities
Open... Save... OK Cancel
33
30
36
38
33
30
15
44.5
74
Log
x 0

Filter

AddExpression -E a5/a6 -N Na_su_K

Current relation

Relation: DRUG1n
Instances: 200
Attributes: 7

Selected attribute

Name: Age
Missing: 0 (0%)
Distinct: 57
Type: Numeric
Unique: 7 (4%)

Attributes

No.	Name
1	□ Age
2	□ Sex
3	□ BP
4	□ Cholesterol
5	□ Na
6	□ K
7	□ Drug

Value
15
74
44.315
16.544

weka.gui.GenericObjectEditor

weka.filters.unsupervised.attribute.AddExpression

About

An instance filter that creates a new attribute by applying a mathematical expression to existing attributes.

debug: False

expression: a5/a6

name: Na_su_K

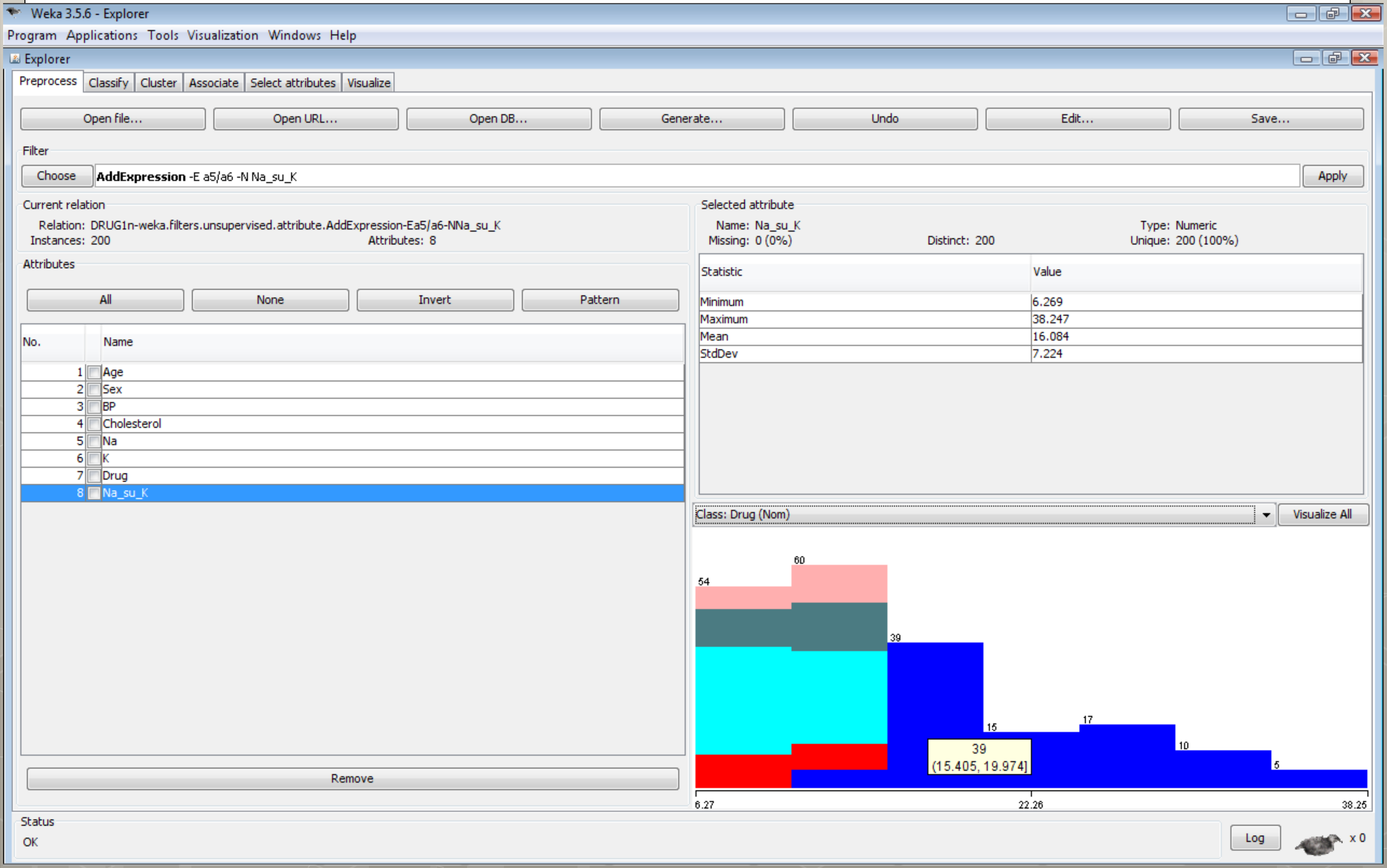
Status

OK

Data Preparation

- This filter adds a new attribute to the table, applaying a mathematical expression to the existing attributes (field *expression*). The new attribute will take name from the field *name*.
- Click on *more* and *capabilities* for further information about the filter

Data Preparation



Weka 3.5.6 - Explorer
Program Applications Tools Visualization Windows Help
Explorer
Preprocess Classify Cluster Associate Select attributes Visualize
Open file... Open URL... Open DB... Generate... Undo Edit... Save...
Filter
Choose
Apply
All None Invert Pattern
Remove
Visualize All
Log
x 0

Filter: **AddExpression** -E a5/a6 -N Na_su_K

Current relation

Relation: DRUG1n-weka.filters.unsupervised.attribute.AddExpression-Ea5/a6-NNa_su_K
Instances: 200 Attributes: 8

Attributes

No.	Name
1	□ Age
2	□ Sex
3	□ BP
4	□ Cholesterol
5	□ Na
6	□ K
7	□ Drug
8	□ Na_su_K

Selected attribute

Name: Na_su_K Type: Numeric
Missing: 0 (0%) Distinct: 200 Unique: 200 (100%)

Statistic	Value
Minimum	6.269
Maximum	38.247
Mean	16.084
StdDev	7.224

Class: Drug (Nom)

Status
OK

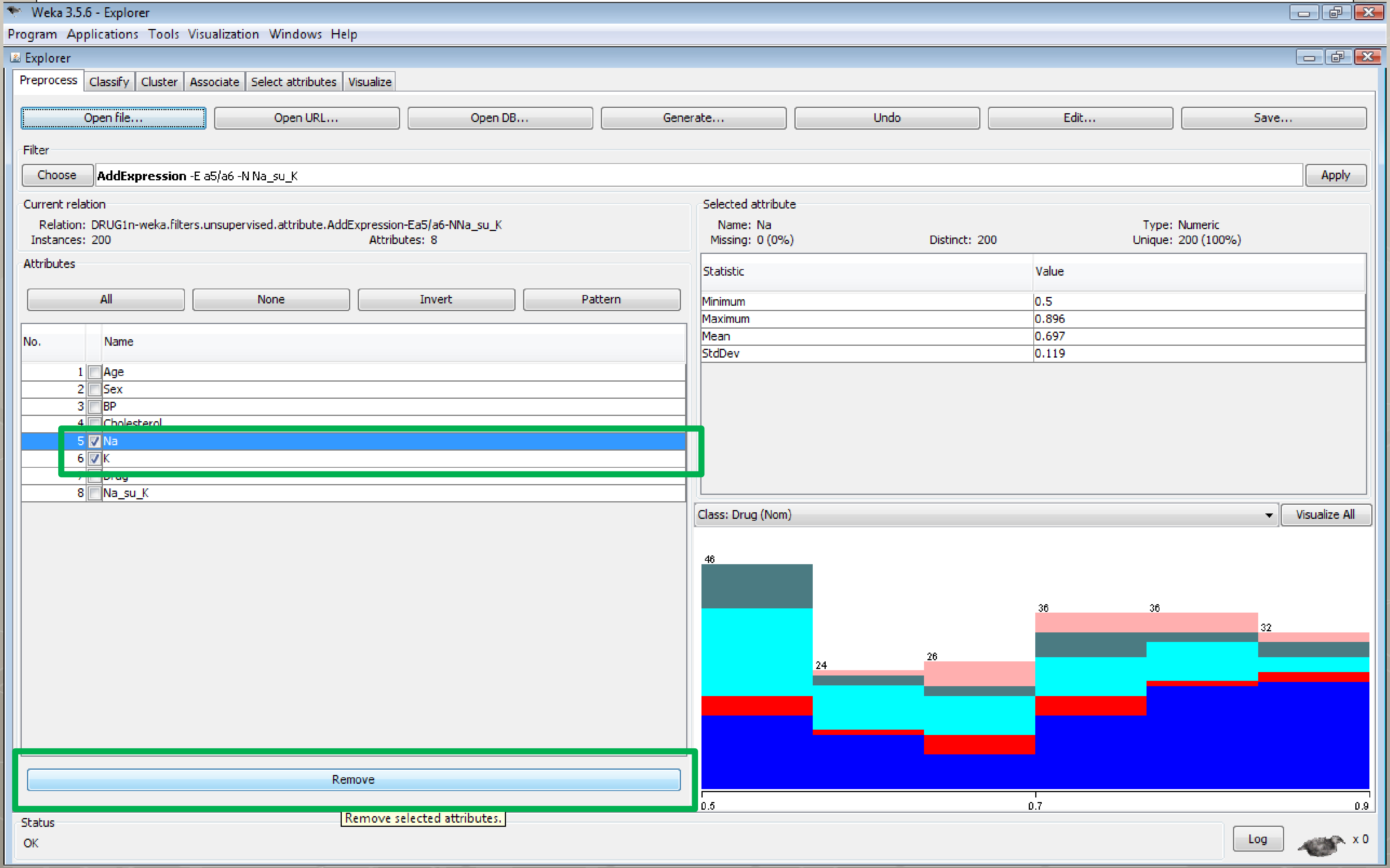
Data Preparation

- Comments?

Data Preparation

- The attribute Na and K now are redundant. Let’s remove them

Data Preparation



Filter

AddExpression -E a5/a6 -N Na_su_K

Current relation

Relation: DRUG1n-weka.filters.unsupervised.attribute.AddExpression-Ea5/a6-NNa_su_K
Instances: 200 Attributes: 8

Attributes

No.		Name
1	□	Age
2	□	Sex
3	□	BP
4	□	Cholesterol
5	☑	Na
6	☑	K
7	□	Drug
8	□	Na_su_K

Selected attribute

Name: Na Type: Numeric
Missing: 0 (0%) Distinct: 200 Unique: 200 (100%)

Statistic	Value
Minimum	0.5
Maximum	0.896
Mean	0.697
StdDev	0.119

Class: Drug (Nom)

Remove selected attributes.

Status

OK

Modeling

- Data seems ready to be used in the modeling phase