CSCI 485 (Machine Learning)

Sami Al-Qusus

Spring 2019 - Assignment 1
Submit deadline: 11:30, 31 January 2019, Thursday

Back-story:
An analytics consultant at an insurance company has collected a set of data that will be used to train a model to predict the best communications channel to use to contact a potential customer with an offer of a new insurance product. The full data set contains 5,200 instances and can be accessed at (http://csci.viu.ca/~liuh/485/assignments/A1-data.csv).

Task:
Explaining the process of generating a data quality and preliminary data exploration report for the following data set:

- AGE: the customer's age
- GENDER: the customer's gender (*male* or *female*)
- LOC: the customer's location (*rural* or *urban*)
- OCC: the customer's occupation
- MOTOR_INS: whether the customer holds a motor insurance policy with the company (*yes* or *no*)
- MOTOR_VALUE: the value of the car on the motor policy
- HEALTH_INS: whether the customer holds a health insurance policy with the company (*yes* or *no*)
- HEALTH_TYPE: the type of the health insurance policy (*PlanA*, *PlanB*, or *PlanC*)
- DEPS_ADULTS: how many dependent adults are included on the health insurance policy
- DEPS_KIDS: how many dependent children are included on the health insurance policy
- PREF_CHANNEL: the customer's preferred contact channel (*email*, *phone*, or *sms*)

Report:

- Id:
  - Attribute type: Numeric
  - Data item count: 5200
  - Percentage of missing values: 0%
  - Cardinality (number of unique values): 5200

- AGE:
  - Attribute type: Numeric
  - Data item count: 5200
  - Percentage of missing values: 0%
  - Cardinality (number of unique values): 56
  - For continuous typed attribute
    - Minimum value: 20
    - First quarter value: 33.75
    - Mean: 47.958
    - Median: 47.5
    - Third quarter value: 61.25
    - Maximum: 75
    - Standard deviation: 16.265
  - Outliers/Errors: no
- GENDER:
  - Attribute type: Categorical (Nominal)
  - Data item count: 5200-47=5153
  - Percentage of missing values: 1%
  - Cardinality (number of unique values): 3
  - Statistics
    - Mode: male
    - Mode frequency: 2565
    - Mode percentage: 49.78
    - Second mode: female
    - Second mode frequency: 2534
    - Second mode percentage: 49.18
  - Outliers/Errors:
    - "rural" label is an outlier it only occurs 54 times (54/5153*100=1.05%) and from my basic knowledge on gender "rural" is not a gender type therefore its probably an error.
- LOC:
  - Attribute type: Categorical (Nominal)
  - Data item count: 5200-118=5082
  - Percentage of missing values: 2%
  - Cardinality (number of unique values): 3
  - Statistics
    - Mode: urban
    - Mode frequency: 2565
    - Mode percentage: 50.47
    - Second mode: rural
    - Second mode frequency: 2516
    - Second mode percentage: 49.51

    - Outliers/Errors:
        - "female" label is an outlier it only occurs 1 time (1/5083*100~=0%) and from my basic knowledge on location "female" is not a location type therefore it an error.
- OCC:
    - Attribute type: Categorical (Nominal)
    - Data item count: 5200-303=4897
    - Percentage of missing values: 6%
    - Cardinality (number of unique values): 20
    - Statistics
        - Mode: Courier
        - Mode frequency: 279
        - Mode percentage: 5.70
        - Second mode: Doctor
        - Second mode frequency: 275
        - Second mode percentage: 5.62
    - Outliers/Errors: no
- MOTOR_INS:
    - Attribute type: Categorical (Nominal/Binary)
    - Data item count: 5200-168=5032
    - Percentage of missing values: 3%
    - Cardinality (number of unique values): 2
    - Statistics
        - Mode: yes
        - Mode frequency: 3991
        - Mode percentage: 79.31
        - Second mode: no
        - Second mode frequency: 1041
        - Second mode percentage: 20.69
    - Outliers/Errors: no
- MOTOR_VALUE:
    - Attribute type: Numeric
    - Data item count: 5200-312=4888
    - Percentage of missing values: 6%
    - Cardinality (number of unique values): 4573
    - For continuous typed attribute
        - Minimum value: 3000
        - First quarter value: 124,418
        - Mean: 22768.656
        - Median: 245,836
        - Third quarter value: 365,754
        - Maximum: 119,918
        - Standard deviation: 19705.785

    - Outliers/Errors:
        - Maybe some outliers, visually it looks like there are because in the third and fourth quarter the combined motors are a couple of percent compared to all the motors, but from my basic world knowledge only a few individuals drive cars valued over $245,836 so might not be errors but needs to be looked at closely by an expert to make sure.
- HEALTH_INS:
    - Attribute type: Categorical (Nominal/Binary)
    - Data item count: 5200-99=5101
    - Percentage of missing values: 2%
    - Cardinality (number of unique values): 2
    - Statistics
        - Mode: yes
        - Mode frequency: 3957
        - Mode percentage: 77.57
        - Second mode: no
        - Second mode frequency: 1144
        - Second mode percentage: 22.43
    - Outliers/Errors: no
- HEALTH_TYPE:
    - Attribute type: Categorical (Nominal)
    - Data item count: 5200-1243=3957
    - Percentage of missing values: 24%
    - Cardinality (number of unique values): 1
    - Statistics
        - Mode: PlanA
        - Mode frequency: 3957
        - Mode percentage: 100
        - Second mode: n/a
        - Second mode frequency: n/a
        - Second mode percentage: n/a
    - Outliers/Errors: no
- DEPS_ADULTS:
    - Attribute type: Numeric
    - Data item count: 5200-1243=3957
    - Percentage of missing values: 24%
    - Cardinality (number of unique values): 3
    - For continuous typed attribute
        - Minimum value: 0
        - First quarter value: 0.5
        - Mean: 0.783
        - Median: 1
        - Third quarter value: 1.5
        - Maximum: 2
        - Standard deviation: 0.628

    - Outliers/Errors: no
- DEPS_KIDS:
    - Attribute type: Numeric
    - Data item count: 5200-1243=3957
    - Percentage of missing values: 24%
    - Cardinality (number of unique values): 4
    - For continuous typed attribute
        - Minimum value: 0
        - First quarter value: 0.75
        - Mean: 1.502
        - Median: 1.5
        - Third quarter value: 2.25
        - Maximum: 3
        - Standard deviation: 0.893
    - Outliers/Errors: no
- PREF_CHANNEL:
    - Attribute type: Categorical (Nominal)
    - Data item count: 5200
    - Percentage of missing values: 0%
    - Cardinality (number of unique values): 3
    - Statistics
        - Mode: email
        - Mode frequency: 1761
        - Mode percentage: 33.87
        - Second mode: sms
        - Second mode frequency: 1731
        - Third mode percentage: 33.29
        - Third mode: sms
        - Third mode frequency: 1708
        - Third mode percentage: 32.85
    - Outliers/Errors: no

Findings:

- Any information that can be easily observed from the data:
    - Visually looks like all people have almost equal preference for communications channel regardless of age, gender, location, occupation, motor_ins, motor_value, health_ins, health_type, Deps_adults, or Deps_kids.
- Any relationship among the attributes that can be easily detected:
    - They are almost all even distributed/related nothing that stands out.

Procedures:

1. Download the data set file A1-data.csv from http://csci.viu.ca/~liuh/485/assignments/A1-data.csv

2. Open the Weka Knowledge Explorer application
    a. Click on open file and select A1-data.csv
    b. Then click on the attributes one by one to visually explore the data.
    c. Select visualize all to see relationships between attributes.
        i. Notice
            1. The outliers: look for off numbers.
            2. Visually looks like all people have almost same preference percentage for communications channel regardless of attributes.
            3. From the relationship between attributes they are almost all even distributed/related nothing that stands out.
3. First find the count of instances in table through SQL with statement:
    a. SELECT COUNT(*) FROM tabe_name;
    b. We get 5200 so we can use that to compare others for missing values later after finding the count.
4. Repeat step 3 for all attributes but replace * with attribute name.
    a. This way you get the count.
5. Now subtract the count found in step 4 from 5200 you get the missing values for the given attribute. (Do that for all attributes)
    a. Divide the missing number by 5200 x 100 to get percentage.
6. Repeat step 3 for all attributes but replace * with DISTINCT and attribute name.
    a. This way you get the Cardinality.
7. Finding statistical info for categorical typed attributes:

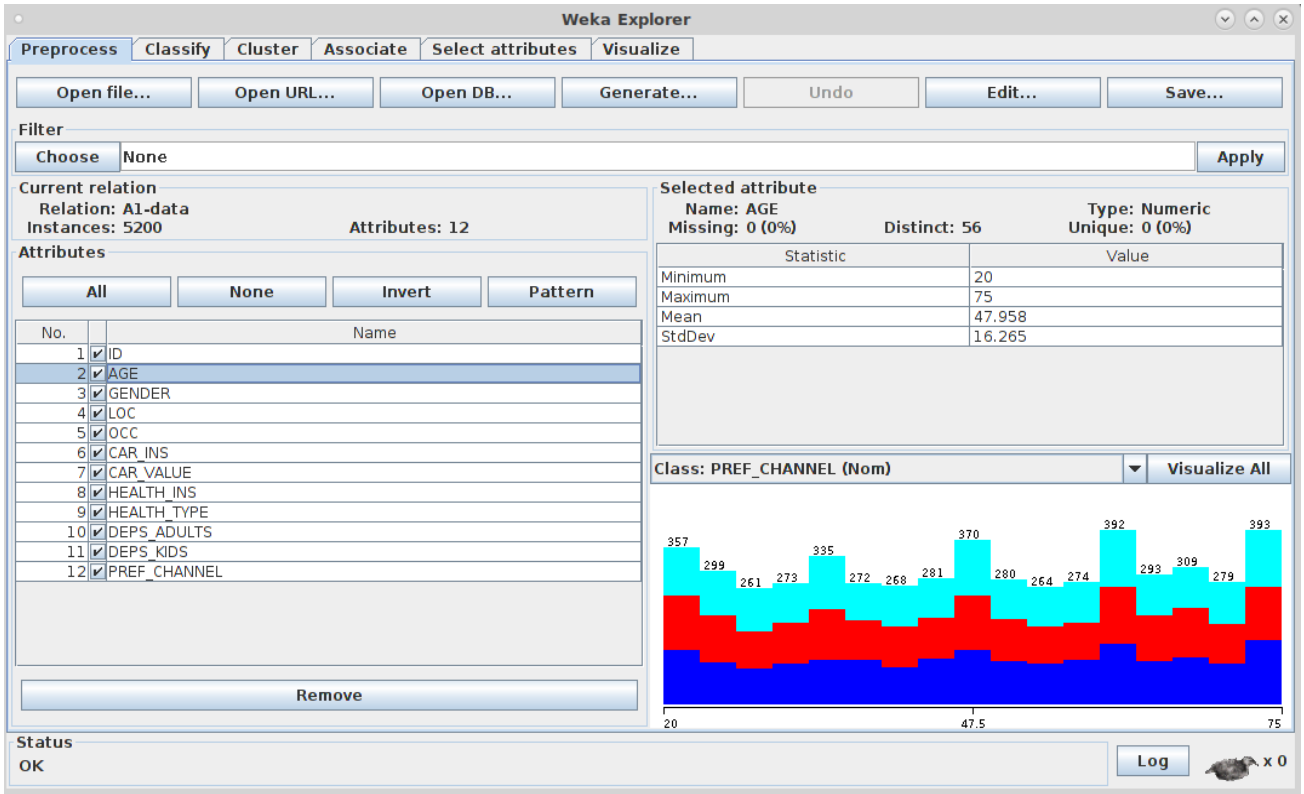
    a. Mode: to find mode
        1. Select attribute_type, count(*) from tabe_name group attribute_type order by count(*) DESC
        2. First one would be the mode one.
        3. Second is mode two.
    b. Mode frequency: from step 1 in the frequency is the count.
    c. Mode percentage: to find mode percentage, take the frequency divide it by the data item count * 100.

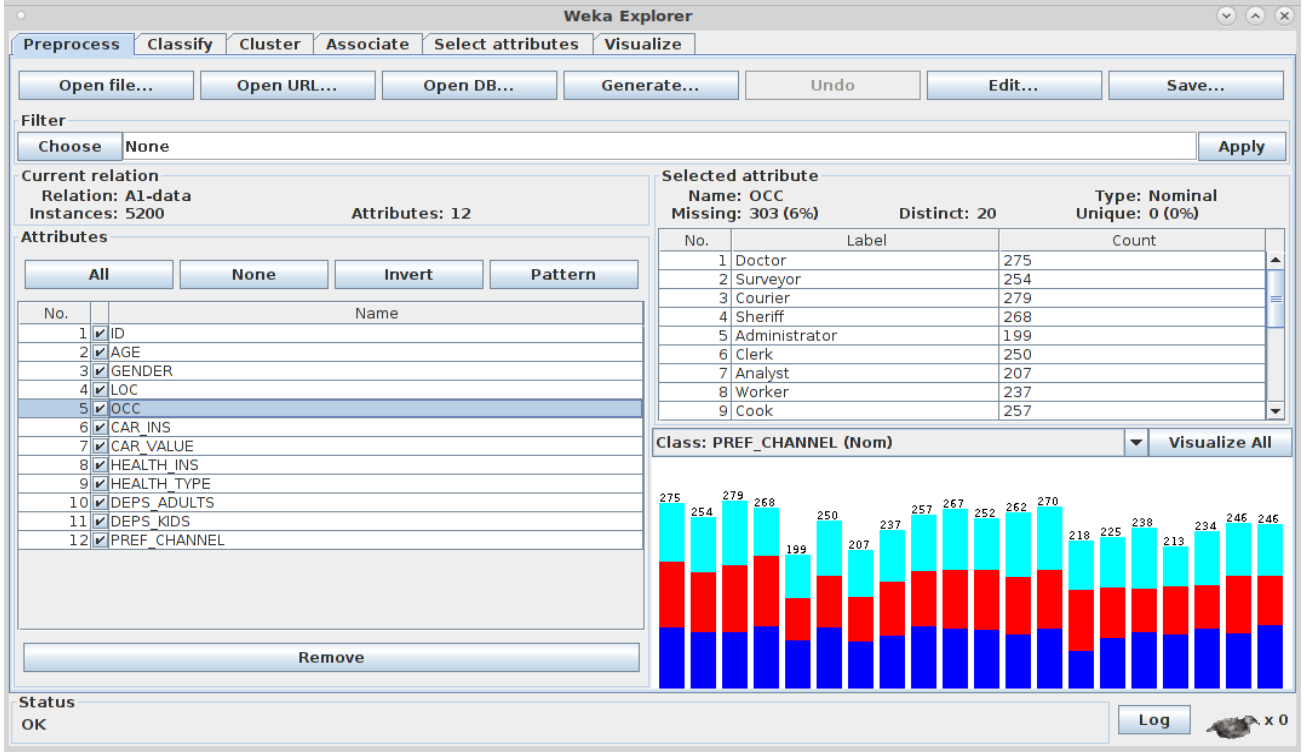
8. Finding statistical info for continuous (numeric) typed attributes:

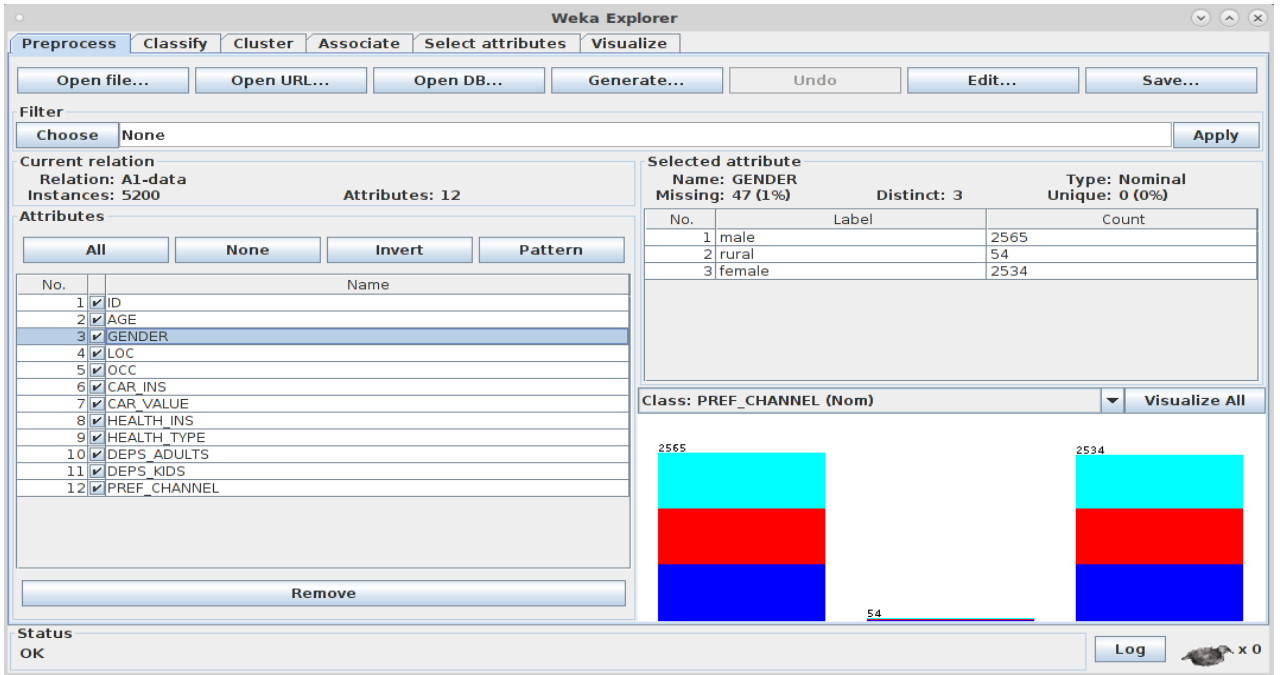
    a. Min and Max values: to find min and max value
        1. Put data on excel
        2. Select the attribute cell range you want to sort.
        3. Select the Data tab on the Ribbon, then click the Sort command.
        4. Then Select Orders from sort by, select values from sort on, and largest to smallest from Order.
        5. The first value on the list will be the maximum and last will be the minimum.

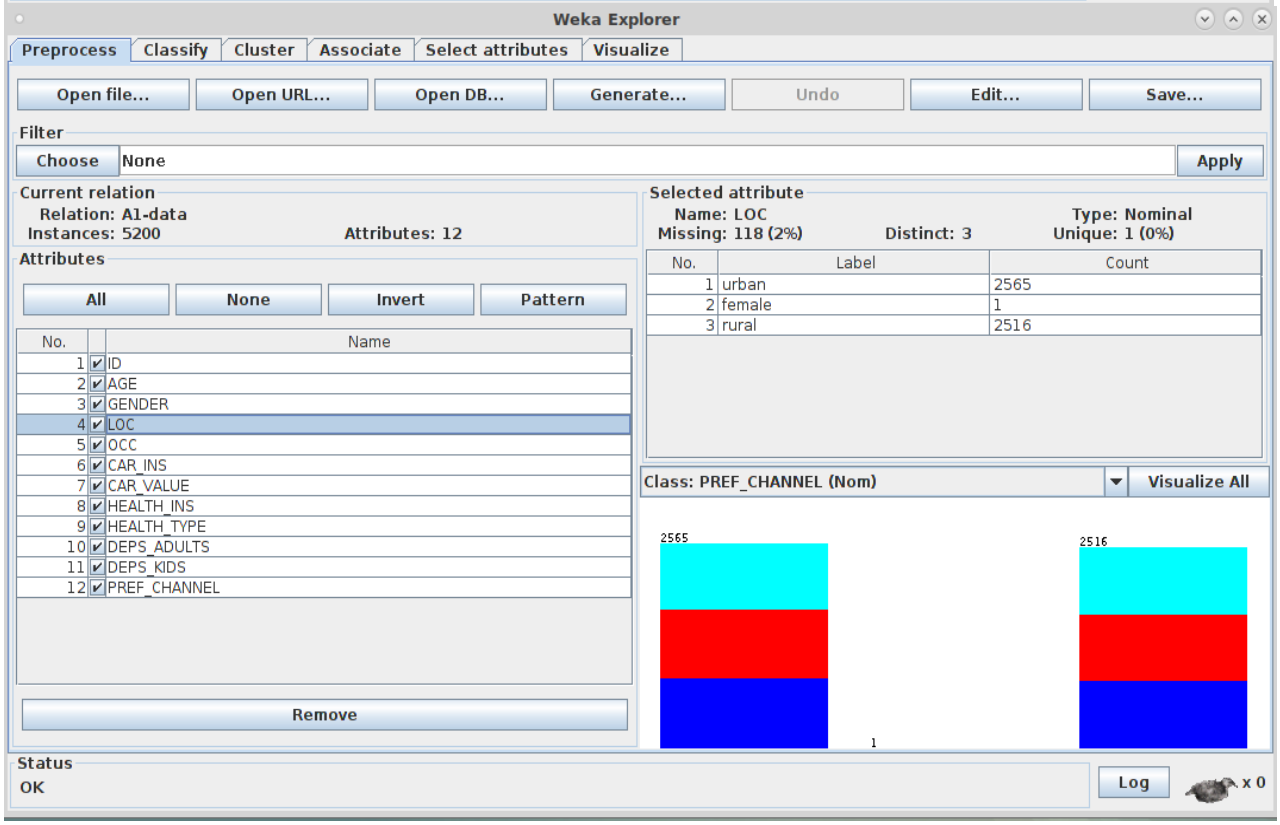
b. Median: to find the median add the max and min and divide by 2.
c. First quarter value: to find the first quarter add the median and min and divide by 2.
d. Third quarter value: to find the third quarter add the median and max and divide by 2.
e. Mean: to find mean type =AVERAGE(function and select all data
f. Standard deviation: to find standard deviation type =STDEV(function and select all data.

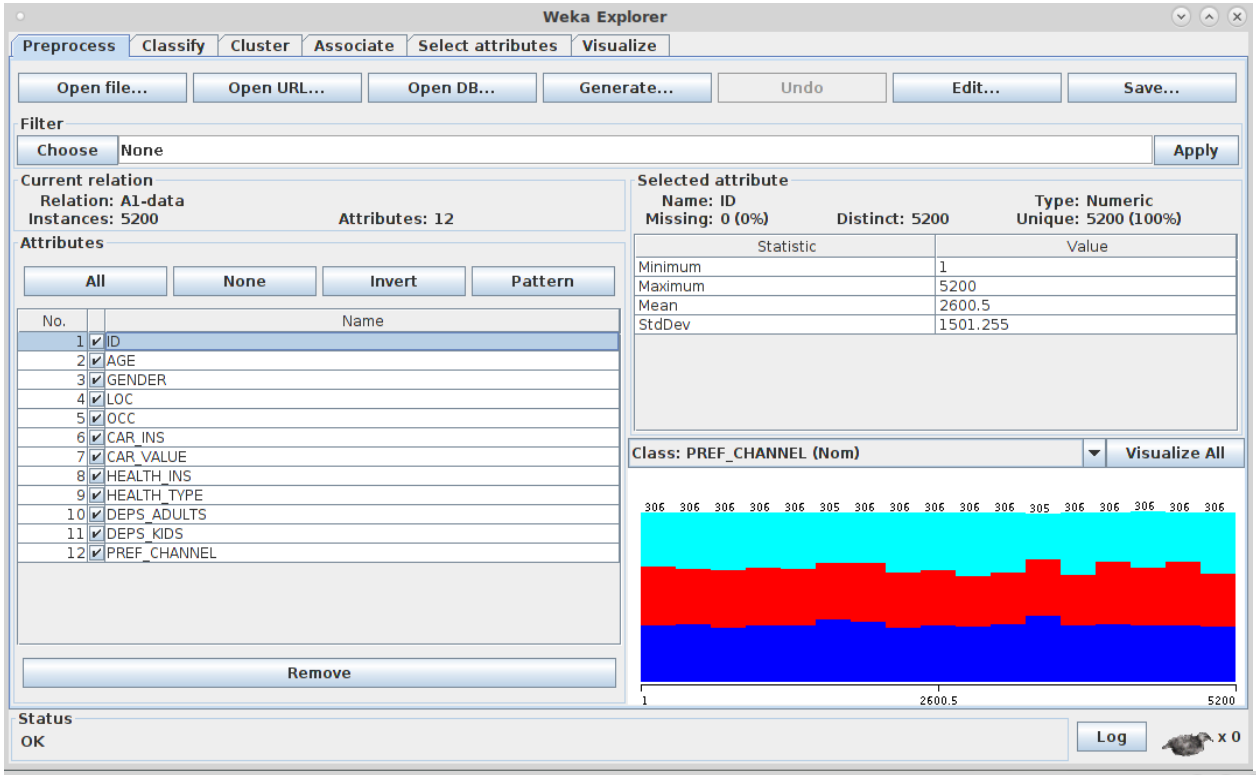
Appendix

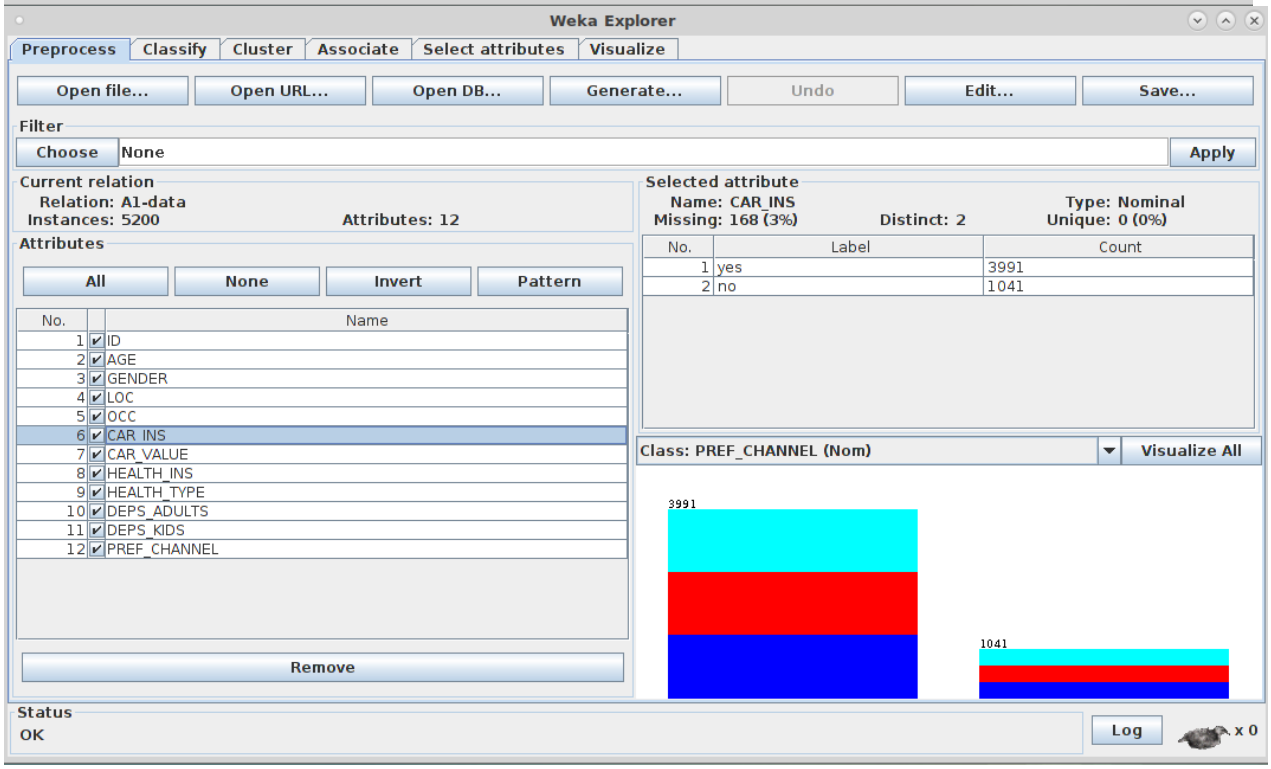

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: AGE
Missing: 0 (0%)
Distinct: 56
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 20
Maximum 75
Mean 47.958
StdDev 16.265
Class: PREF_CHANNEL (Nom)
Visualize All
357 299 261 273 335 272 268 281 370 280 264 274 392 293 309 279 393
20
47.5
75
Status
OK
Log
x 0

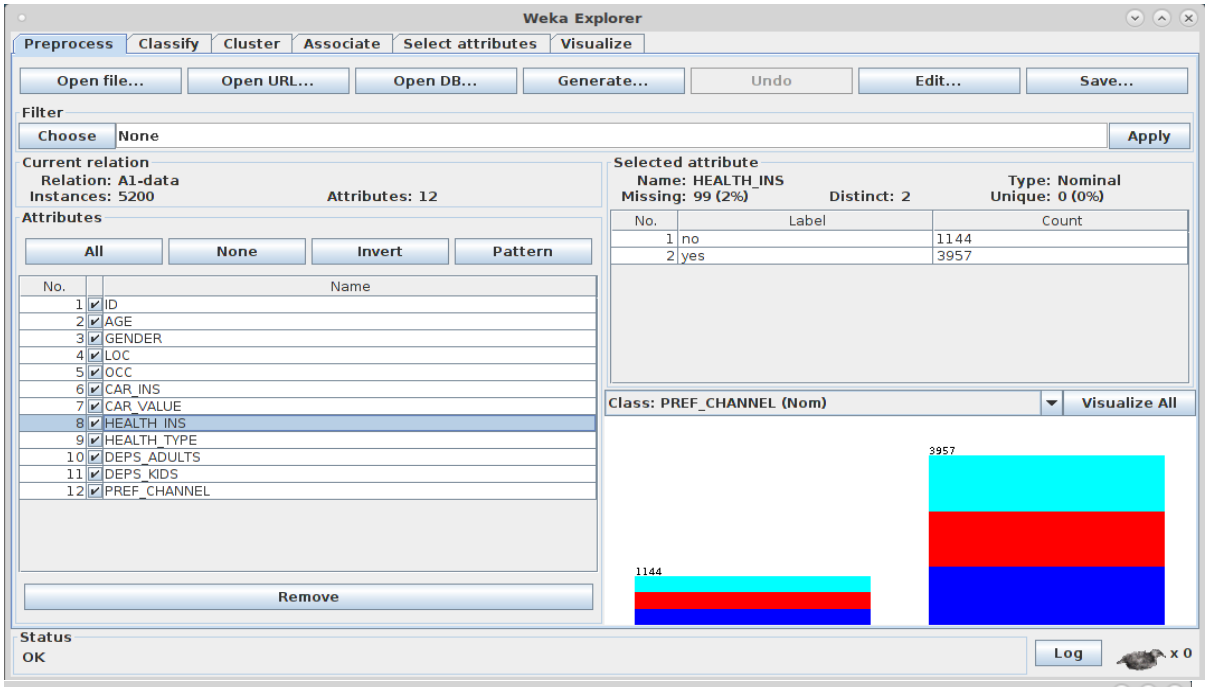

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: OCC
Missing: 303 (6%)
Distinct: 20
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 Doctor 275
2 Surveyor 254
3 Courier 279
4 Sheriff 268
5 Administrator 199
6 Clerk 250
7 Analyst 207
8 Worker 237
9 Cook 257
Class: PREF_CHANNEL (Nom)
Visualize All
275 254 279 268 199 250 207 237 257 267 252 262 270 218 225 238 213 234 246 246
Status
OK
Log
x 0

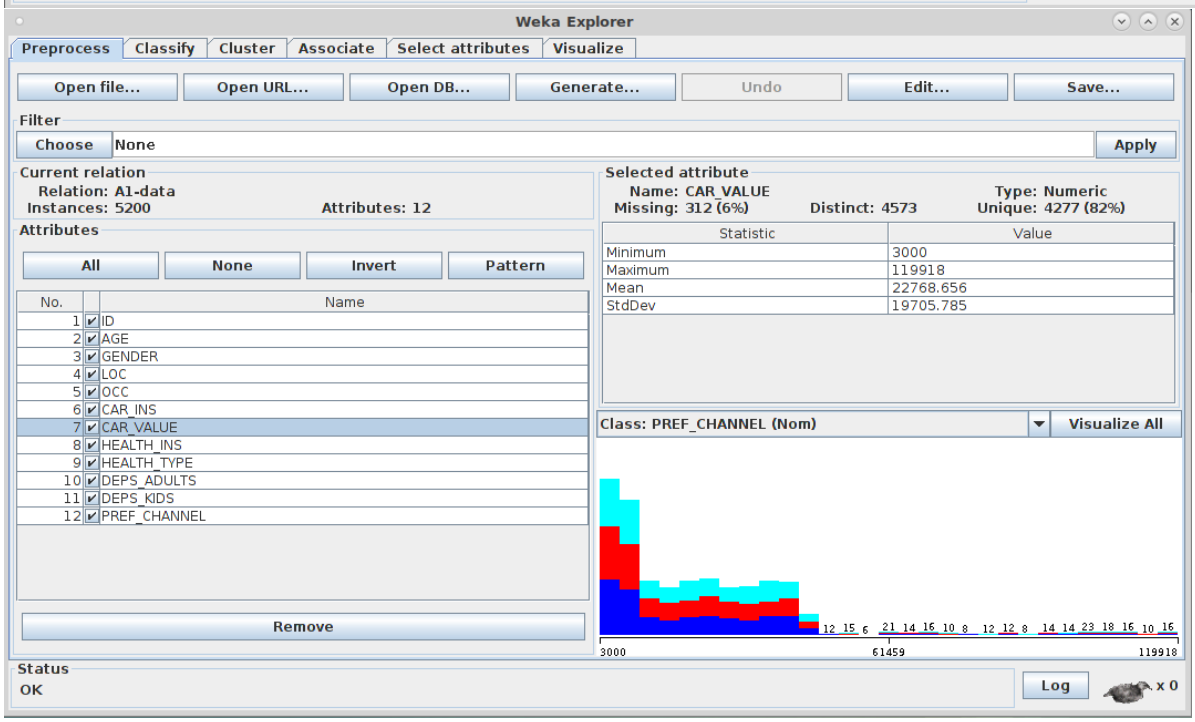

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: GENDER
Missing: 47 (1%)
Distinct: 3
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 male 2565
2 rural 54
3 female 2534
Class: PREF_CHANNEL (Nom)
Visualize All
2565
54
2534
Status
OK
Log
x 0

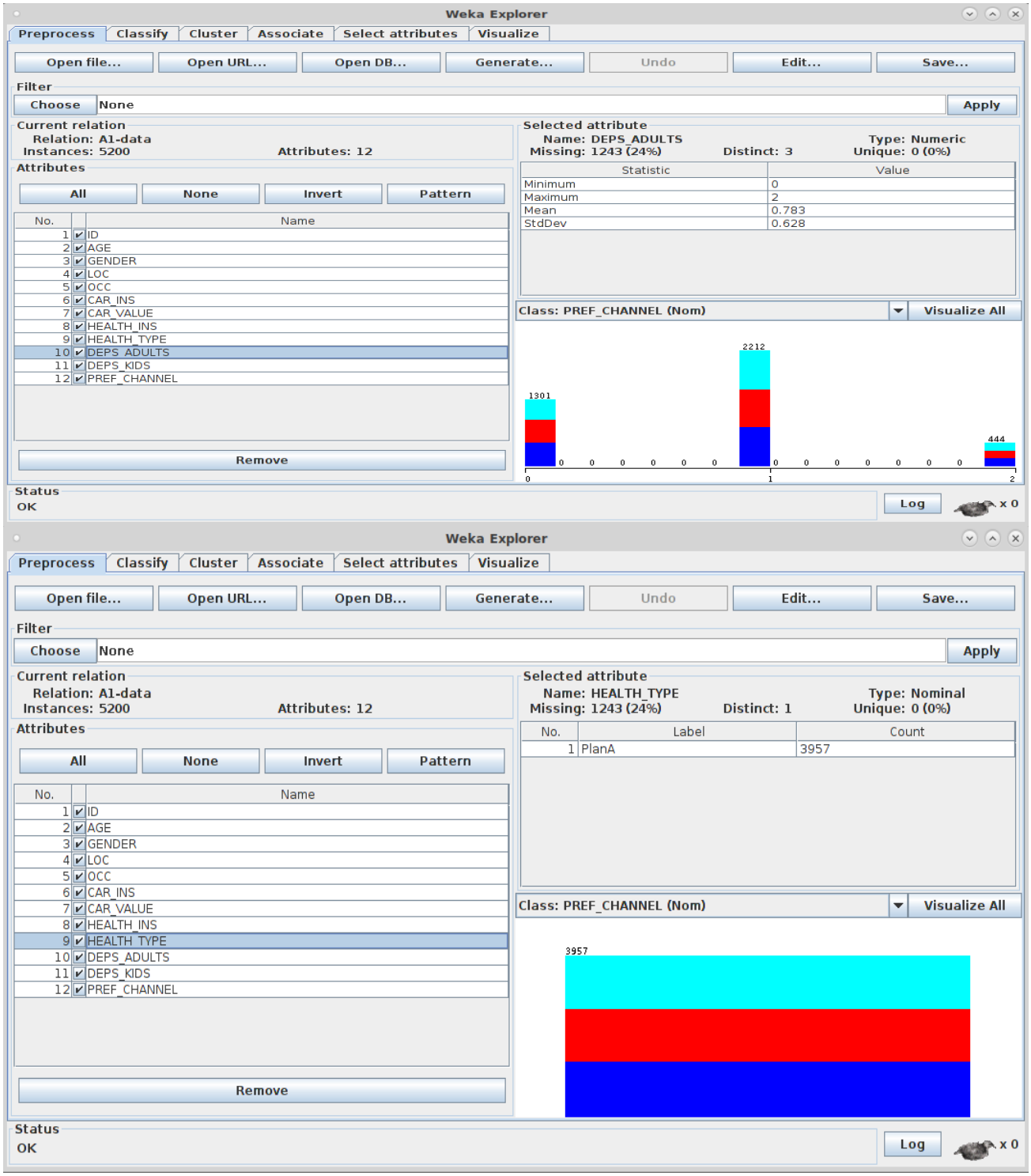

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: LOC
Missing: 118 (2%)
Distinct: 3
Type: Nominal
Unique: 1 (0%)
No.
Label
Count
1 urban 2565
2 female 1
3 rural 2516
Class: PREF_CHANNEL (Nom)
Visualize All
2565
1
2516
Status
OK
Log
x 0

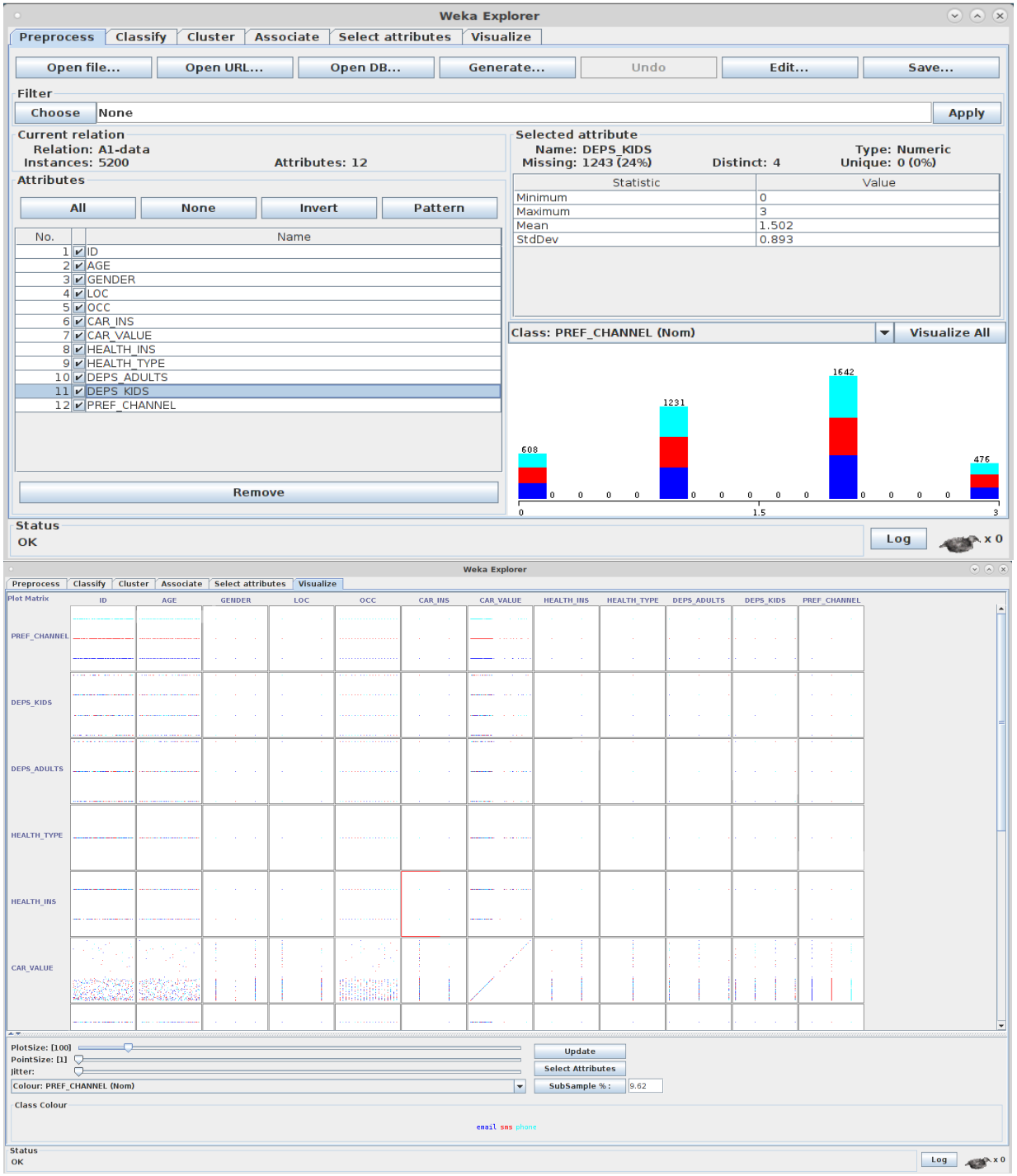

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: ID
Missing: 0 (0%)
Distinct: 5200
Type: Numeric
Unique: 5200 (100%)
Statistic
Value
Minimum 1
Maximum 5200
Mean 2600.5
StdDev 1501.255
Class: PREF_CHANNEL (Nom)
Visualize All
306 306 306 306 306 305 306 306 306 306 306 305 306 306 306 306 306
1
2600.5
5200
Status
OK
Log
x 0

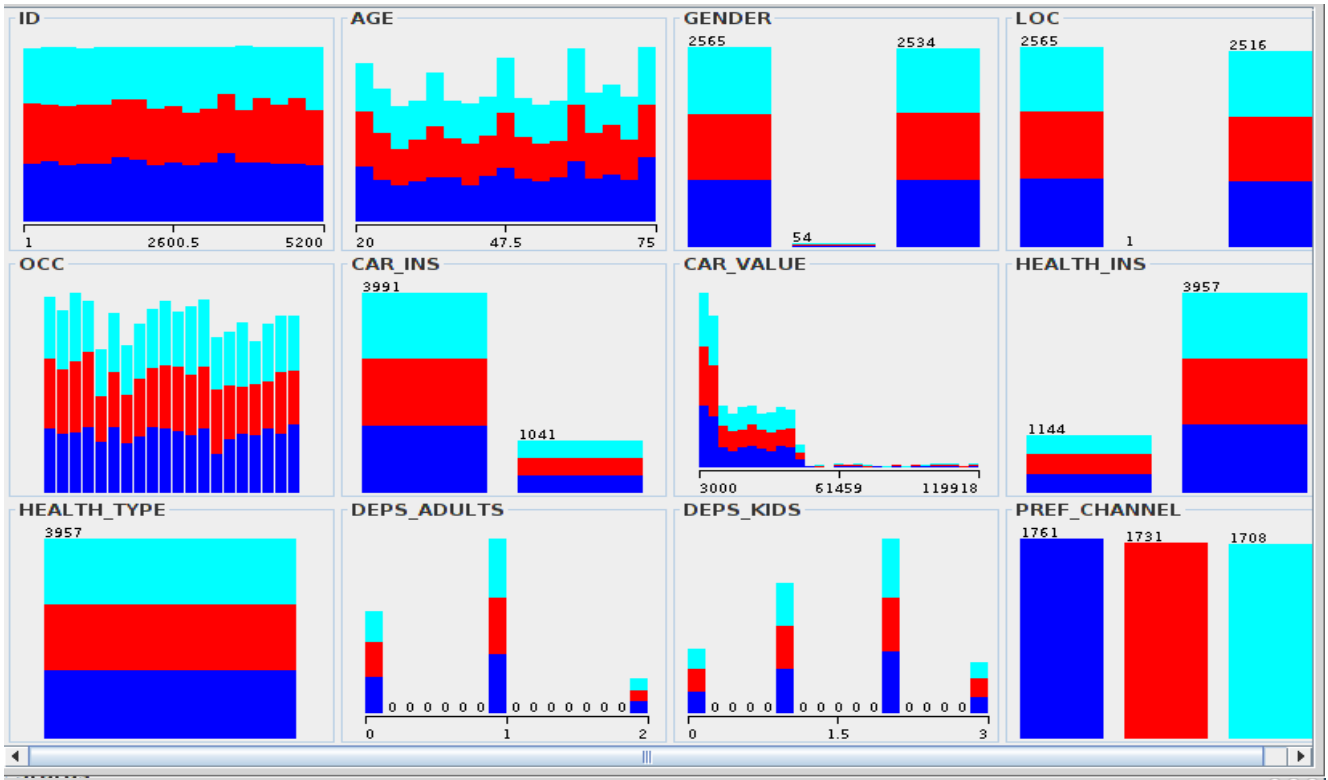

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: CAR_INS
Missing: 168 (3%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 yes 3991
2 no 1041
Class: PREF_CHANNEL (Nom)
Visualize All
3991
1041
Status
OK
Log
x 0

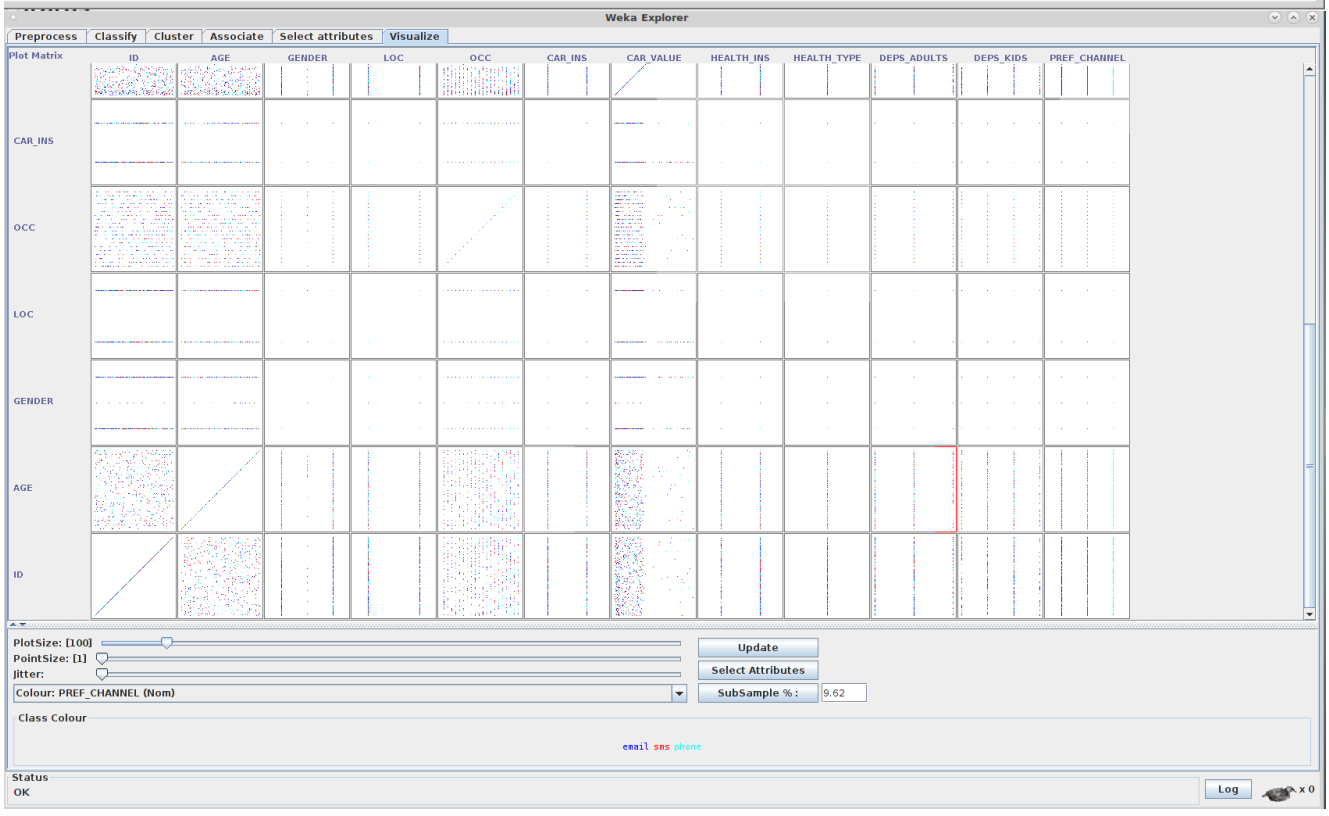

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: HEALTH_INS
Missing: 99 (2%)
Distinct: 2
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 no 1144
2 yes 3957
Class: PREF_CHANNEL (Nom)
Visualize All
3957
1144
Status
OK
Log
x 0

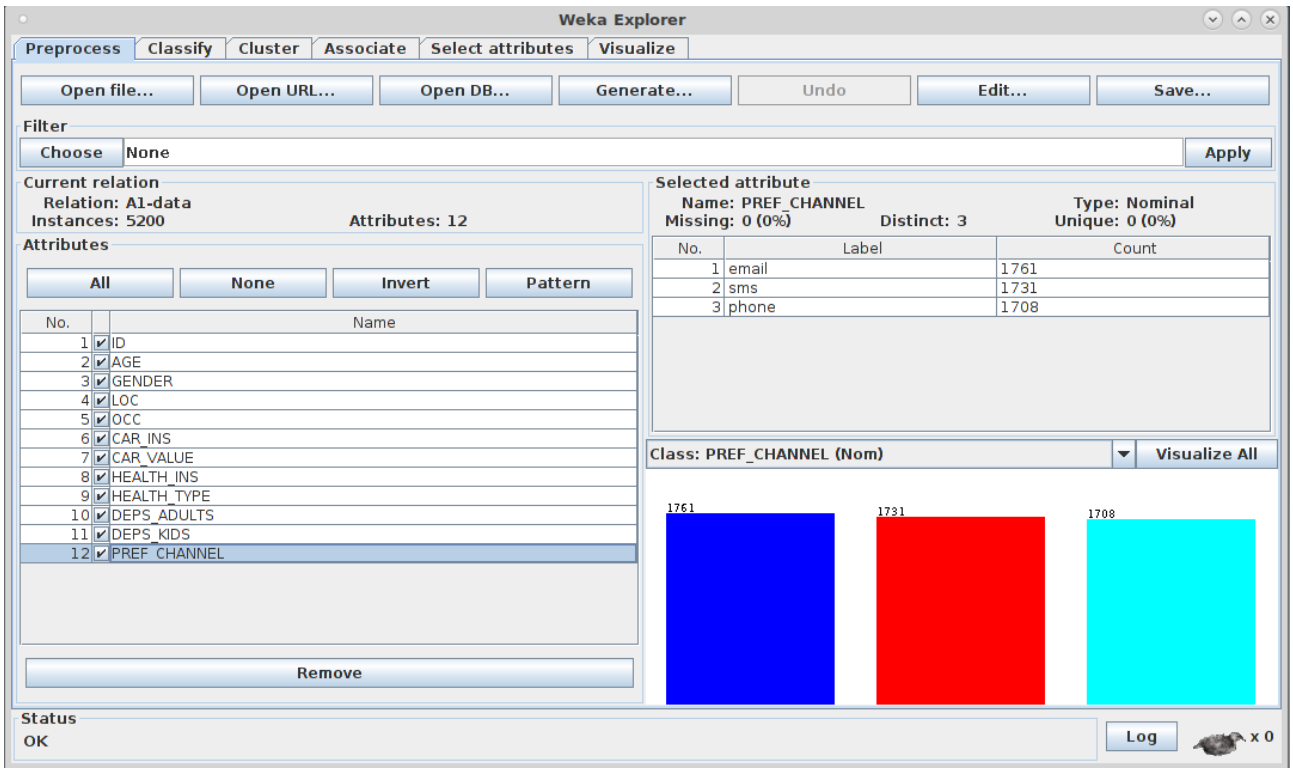

Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: CAR_VALUE
Missing: 312 (6%)
Distinct: 4573
Type: Numeric
Unique: 4277 (82%)
Statistic
Value
Minimum 3000
Maximum 119918
Mean 22768.656
StdDev 19705.785
Class: PREF_CHANNEL (Nom)
Visualize All
12 15 6 21 14 16 10 8 12 12 8 14 14 23 18 16 10 16
3000
61459
119918
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
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: DEPS_ADULTS
Missing: 1243 (24%)
Distinct: 3
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 0
Maximum 2
Mean 0.783
StdDev 0.628
Class: PREF_CHANNEL (Nom)
Visualize All
1301
2212
444
0
1
2
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
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: HEALTH_TYPE
Missing: 1243 (24%)
Distinct: 1
Type: Nominal
Unique: 0 (0%)
No.
Label
Count
1 PlanA 3957
Class: PREF_CHANNEL (Nom)
Visualize All
3957
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
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: DEPS_KIDS
Missing: 1243 (24%)
Distinct: 4
Type: Numeric
Unique: 0 (0%)
Statistic
Value
Minimum 0
Maximum 3
Mean 1.502
StdDev 0.893
Class: PREF_CHANNEL (Nom)
Visualize All
608
1231
1642
476
0
1.5
3
Status
OK
Log
x 0
Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
ID
AGE
GENDER
LOC
OCC
CAR_INS
CAR_VALUE
HEALTH_INS
HEALTH_TYPE
DEPS_ADULTS
DEPS_KIDS
PREF_CHANNEL
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: PREF_CHANNEL (Nom)
SubSample %:
9.62
Class Colour
email sms phone
Status
OK
Log
x 0


ID
1
2600.5
5200
AGE
20
47.5
75
GENDER
2565
54
2534
LOC
2565
1
2516
OCC
CAR_INS
3991
1041
CAR_VALUE
3000
61459
119918
HEALTH_INS
1144
3957
HEALTH_TYPE
3957
DEPS_ADULTS
0
1
2
DEPS_KIDS
0
1.5
3
PREF_CHANNEL
1761
1731
1708


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Plot Matrix
ID
AGE
GENDER
LOC
OCC
CAR_INS
CAR_VALUE
HEALTH_INS
HEALTH_TYPE
DEPS_ADULTS
DEPS_KIDS
PREF_CHANNEL
PlotSize: [100]
PointSize: [1]
Jitter:
Update
Select Attributes
Colour: PREF_CHANNEL (Nom)
SubSample % :
9.62
Class Colour
email sms phone
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
None
Apply
Current relation
Relation: A1-data
Instances: 5200
Attributes: 12
Attributes
All
None
Invert
Pattern
No.
Name
1 ID
2 AGE
3 GENDER
4 LOC
5 OCC
6 CAR_INS
7 CAR_VALUE
8 HEALTH_INS
9 HEALTH_TYPE
10 DEPS_ADULTS
11 DEPS_KIDS
12 PREF_CHANNEL
Remove
Selected attribute
Name: PREF_CHANNEL
Type: Nominal
Missing: 0 (0%)
Distinct: 3
Unique: 0 (0%)
No.
Label
Count
1 email 1761
2 sms 1731
3 phone 1708
Class: PREF_CHANNEL (Nom)
Visualize All
1761
1731
1708
Status
OK
Log
x 0