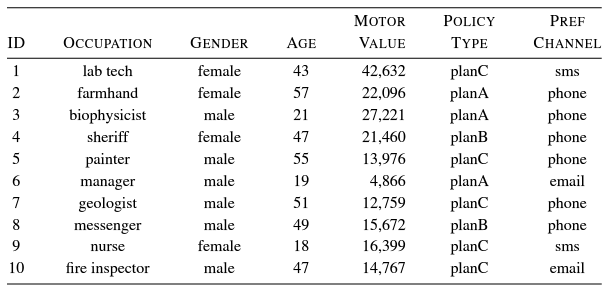
Self-Review Questions

1. The table below shows a sample of a larger dataset containing details of policy holders at an insurance company. The descriptive features (attributes) included in the table describe each policy holders’ ID, occupation, gender, age, the value of their car, the type of insurance policy they hold, and their preferred contact channel.



1. State whether each descriptive feature contains numeric, interval, ordinal, categorical, binary, or textual data. (For example, the ‘Gender’ feature is categorical data.)

‘ID’ feature contains ordinal data.

OCCUPATION’ feature contains categorical data and textual data.

‘GENDER’ feature contains categorical data and binary data.

‘AGE’ feature contains interval data.

‘VALUE’ feature contains numeric data.

‘POLICY TYPE’ feature contains categorical data.

‘PREF CHANNEL’ feature contains categorical data.

1. How many levels does each categorical and ordinal feature have? (For example, ‘Gender’ feature has 2 levels (male, female))

‘ID’ feature has 10 levels(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)

‘OCCUPATION’ feature has 10 levels (lab tech, farmhand, biophysicist, sheriff, painter, manager, geologist, messenger, nurse, fire inspector).

‘GENDER’ feature has 2 levels (male, female).

‘POLICY TYPE’ feature has 3 levels (planA, planB, planC).

‘PREF CHANNEL’ has 3 levels (sms, phone, email).

1. The table below contains sample data about the employees of an IT company.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Emp ID | Name | Year of Birth | Gender | Status | Salary |
| 100 | Smith | 1954 | M | Director | $200.000 |
| 125 | Jones | 1967 | F | Technician | $36,000 |
| 167 | Highley | 1975 | F | Senior Technician | $70,000 |
| 200 | Millins | 1987 | M | Technician | $32,000 |
| 205 | Dujevic | 1985 | M | Technician | $34,000 |
| 216 | Isovic | 1985 | F | Technician | $34,000 |
| 220 | Sun | 1986 | F | Senior Technician | $66,000 |
| 301 | Bean | 1955 | M | Deputy Director | $160,000 |

Answer the following questions.

* 1. An input data set consists of individual data objects, also known as data records, instances or examples. All data sets have the common properties: Type, Size, Dimensionality, Sparsity, Abstraction

Sum up the following characteristics of the table (as above) as a data set: type, size, dimensionality, sparsity and level of abstraction.

Type: Tubular

Size: 10

Dimensionality: 6

Sparsity: None

Level of abstraction: all of the data set

* 1. If the Salary attribute needs to be discretised into three pay bands, suggest a simple yet sensible solution for the discretisation backed with a valid argument.

If the Salary attribute needs to be discretised into three pay bands, it is as below.

Below $35,000

Between $35,000 and $69,000

Above $69,000

This due to trying to figure out the equal number of people in each band and making sure that there are equal or similar number of people in each band.

* 1. If Mr Dujevic’s salary were unknown and the unknown value needed to be imputed, what is a sensible replacement value and why?

If Mr Dujevic’s salary were unknown and the unknown value needed to be imputed, the sensible replacement value will be $33,000 because Mr Dujevic is a male, works as a technician and born in 1985. When we look for male who is born close to 1985 and a person who is born in 1985 and works as a technician, we will be able to find only Mr Millins and Ms Isovic. Therefore, the average of the salary of Mr Millins and Ms Isovic is taken as the salary of Mr Dujevic which is ($32,000 + $34,000)/2.

* 1. Among the employee records, which record can be considered as an outlier? What harm can an outlier object cause to the understanding of the data set?

Among the employee records, Smith can be considered as an outlier. This can cause the average salary of employee to be very high with is approximately $70,000. This can cause the misunderstanding in data set due to only the one salary with is very high compare to other. This affects the results of any statistical analysis or modeling that is performed on the data.

1. A team of researchers conducts research into face recognition. The team has created a database of short colour video clips of frontal face images for 50 human subjects of different genders and ethnic groups. Each video clip consists of 100~120 frames (still images) of 120x80 pixels and is stored in a separate file. Describe how to compose a table from the video clips for classification purposes. In your description, you must state clearly what the rows and columns of the table represent. Which type of data set does the table belong to?

First, we will need first look for the attributes. We can have an attribute as a Subject ID or Subject Name. Then We can have another attribute as Gender which can be stored as Binary Data Type. We can have another attribute for Ethnicity which will be stored as Textual Data Type. These will be the attributes based on the subject. We can add descriptions, notes or explanation as an attribute or an entity for the subject.

Second, we will have to look for attributes of video clip. The attributes can be considered as Number of Frames, Framerate, Count of Pixels and Name or ID of the video clip. All the data can be stored as numeric but for the name, it will be stored as Textual Data Type. We can also add descriptions, notes or explanation.

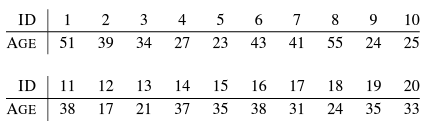
First column can be considered as Attributes with each attribute in each role.

Second column can be used for laying out data.

Third column can be determined for figuring out the data type.

The fourth column can be used as Description or Explanation for the attribute.

1. The table below shows the age of each employee at a cardboard box factory.



Based on this data calculate the following summary statistics for the AGE feature.

(This task is just to check your understanding about fundamental summary methods to describe the data set. You are not required to calculate actual numerical results. Instead, try to show how each statistical summary can be produced using the data given.)

1. Minimum, maximum and range
2. Mean and median
3. Variance and standard deviation
4. 1st quartile (25th percentile) and 3rd quartile (75th percentile)
5. Inter-quartile range
6. 12th percentile
7. Minimum: 17

Maximum: 55

Range: 38

1. Mean: 33.25

Median: 34

1. Variance: 114.22

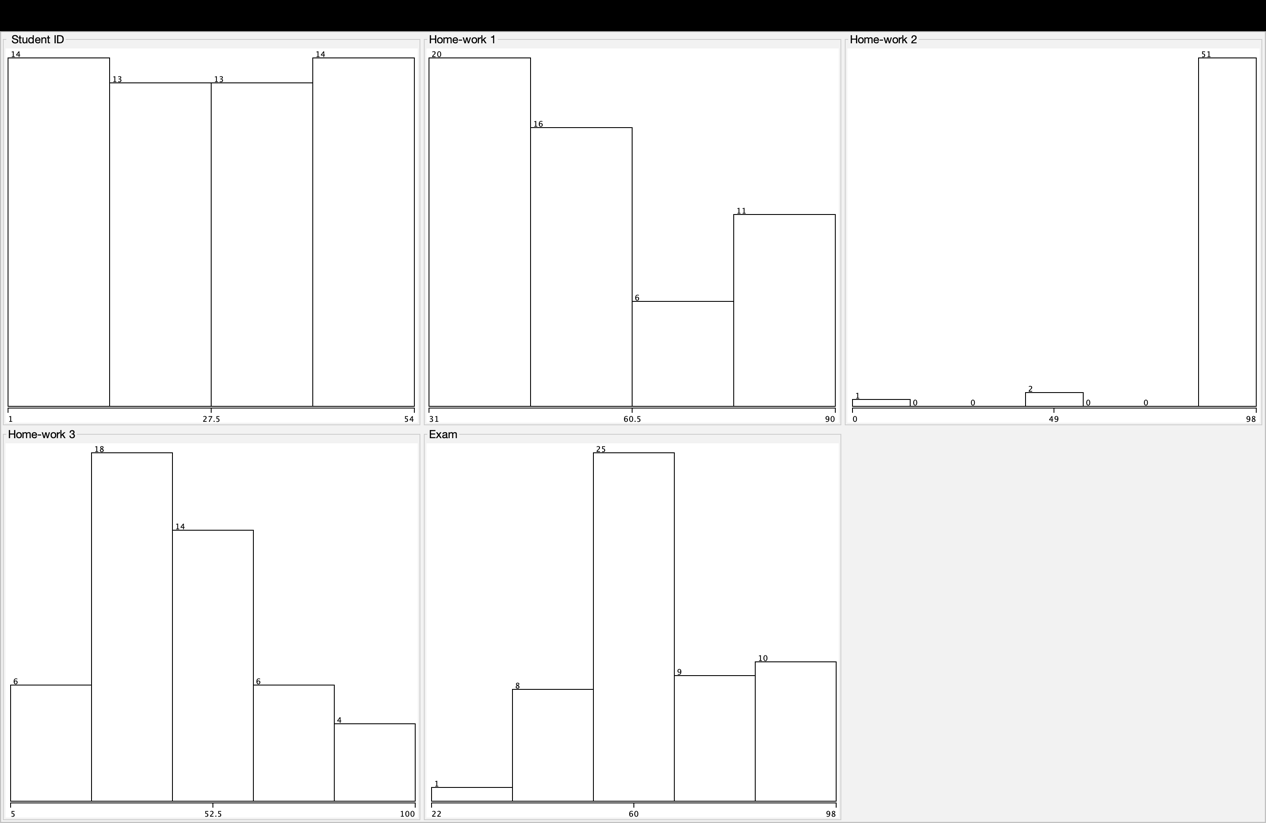
Standard deviation: 10.68

1. 1st quartile: 24

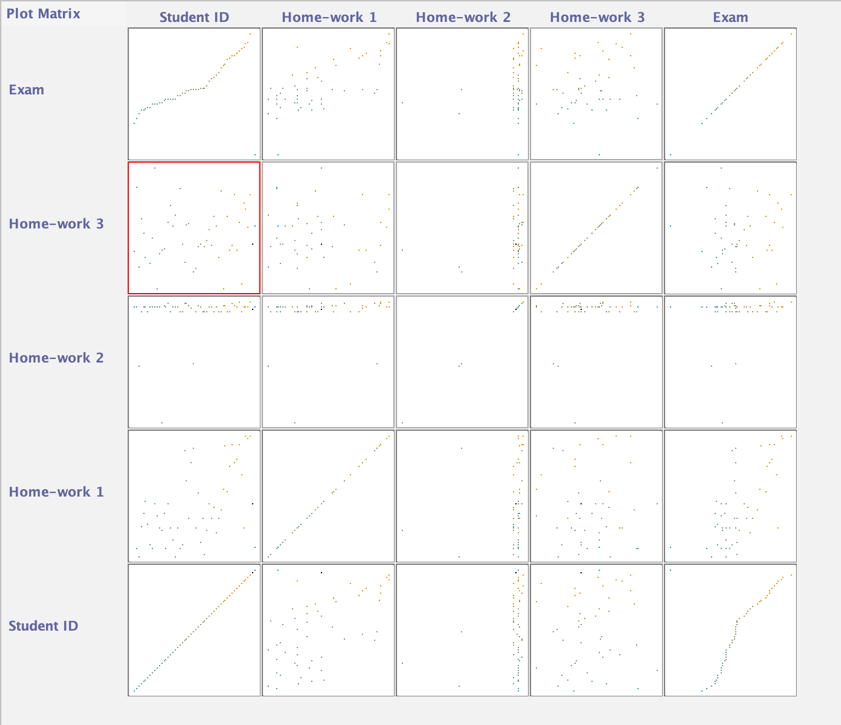
3rd quartile: 38

1. Inter-quartile range: 14
2. 12th percentile: 21.6
3. Observe the summary data for the data set and the histograms for all attributes on the Pre-process tap page. Use the Visualize tab page to view the scatter plots between the variables of the data set.

Histograms for all attributes

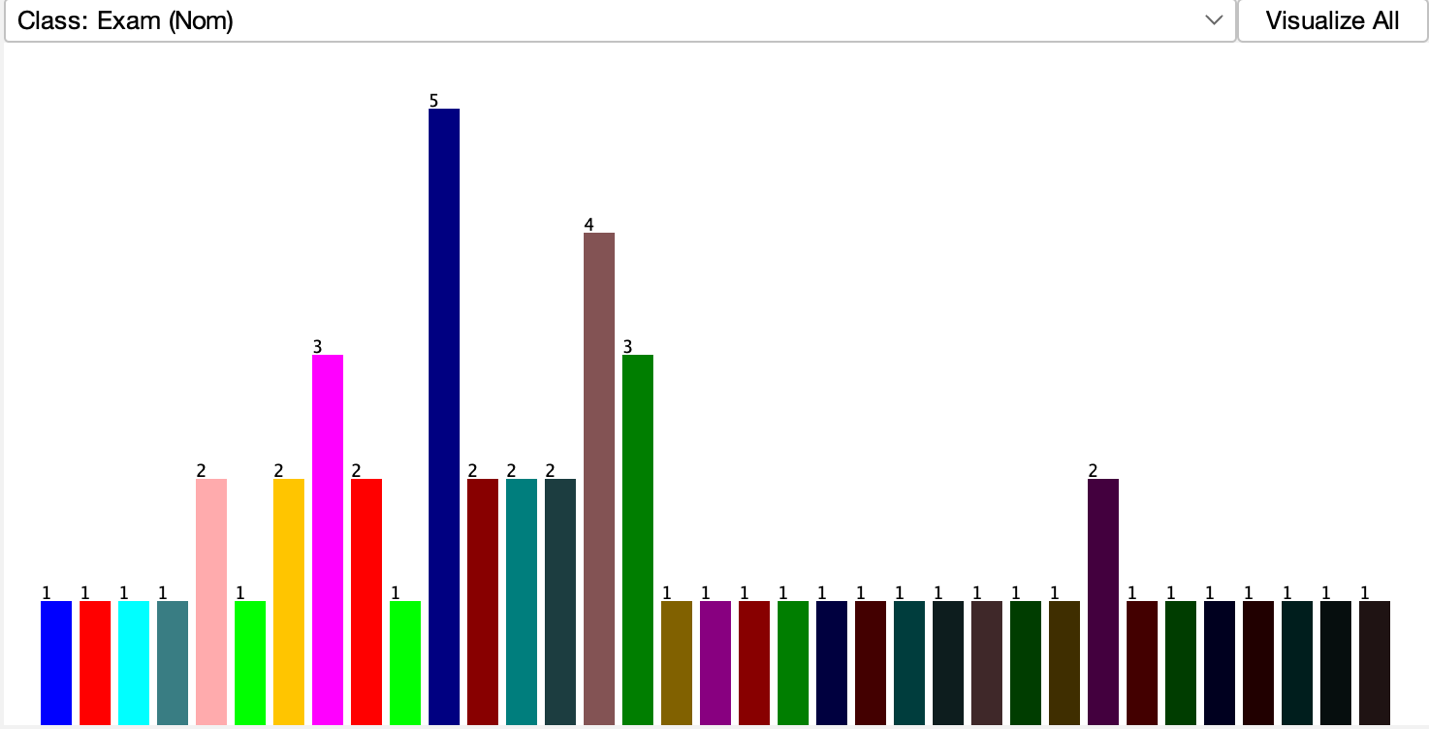


Scattered plot for all attributes

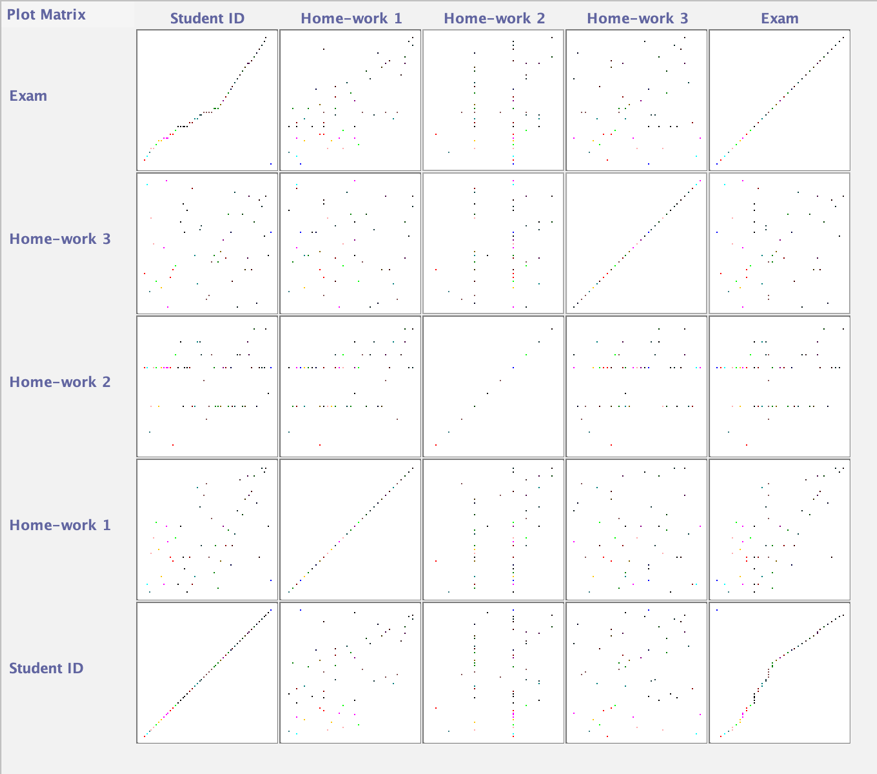


1. Apply the unsupervised Discretize filter to the exam marks.

Histograms for exam marks after discretized

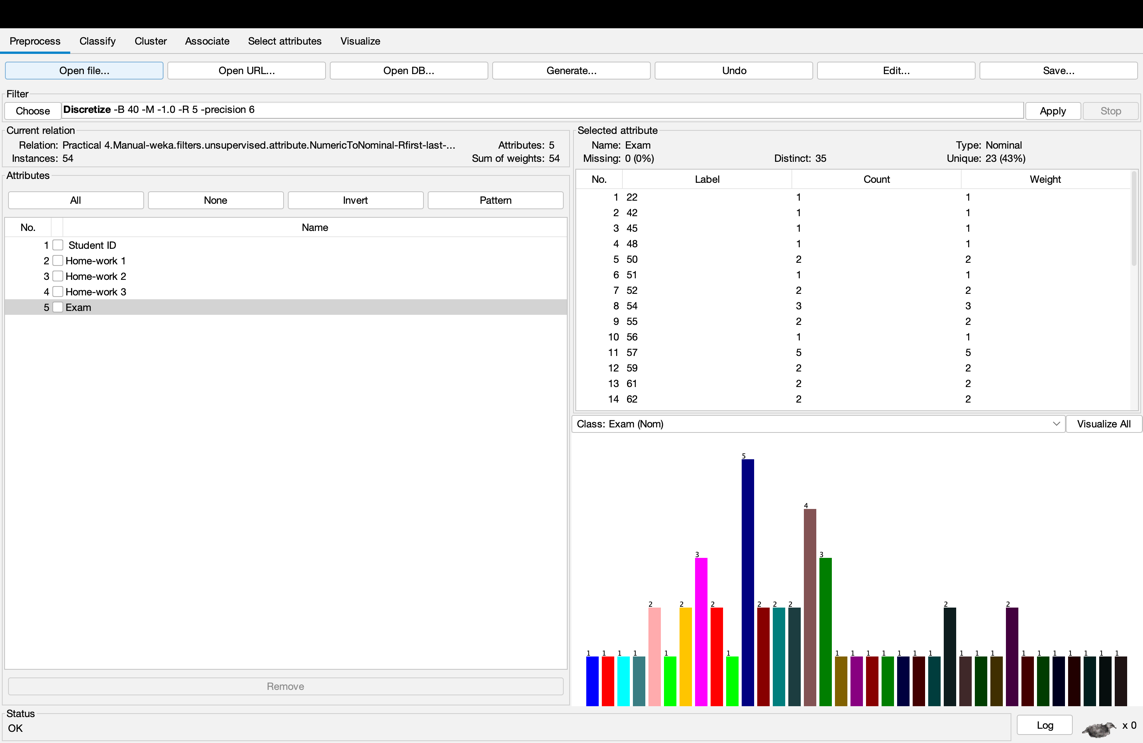


Scattered plot for all attribute after discretizing exam marks

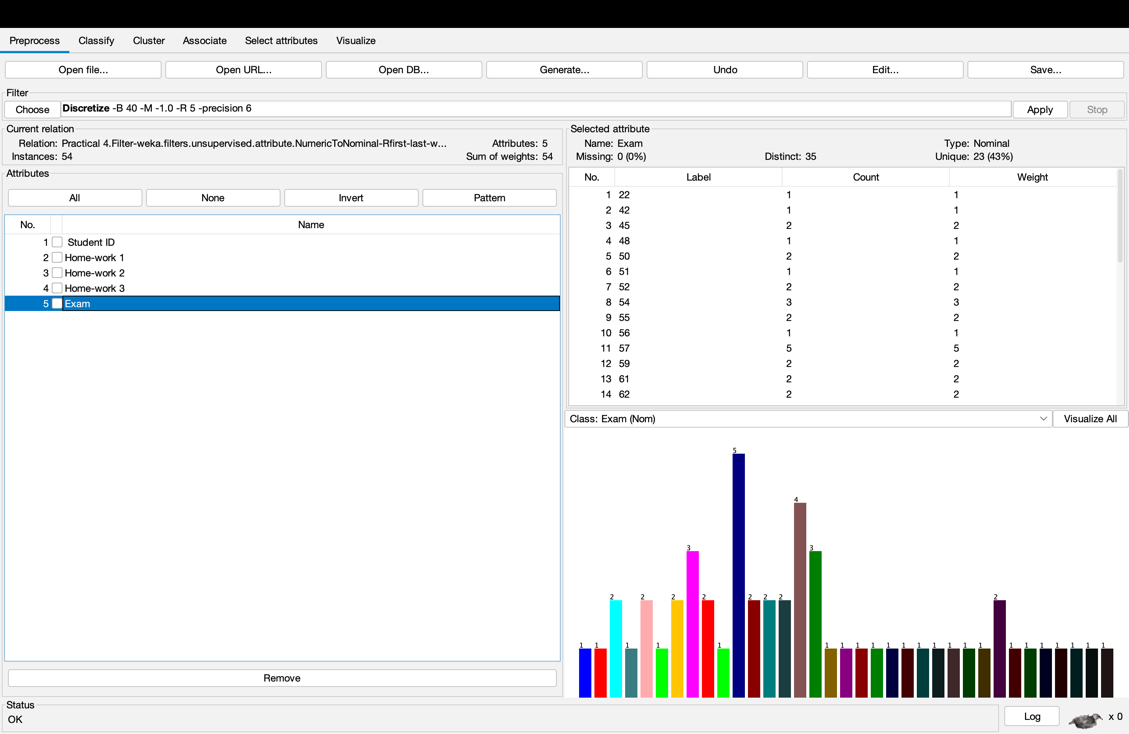


1. Practise filling in missing values in Weka both manually in the Viewer window and by using filters.

Filling in missing values in Weka manually

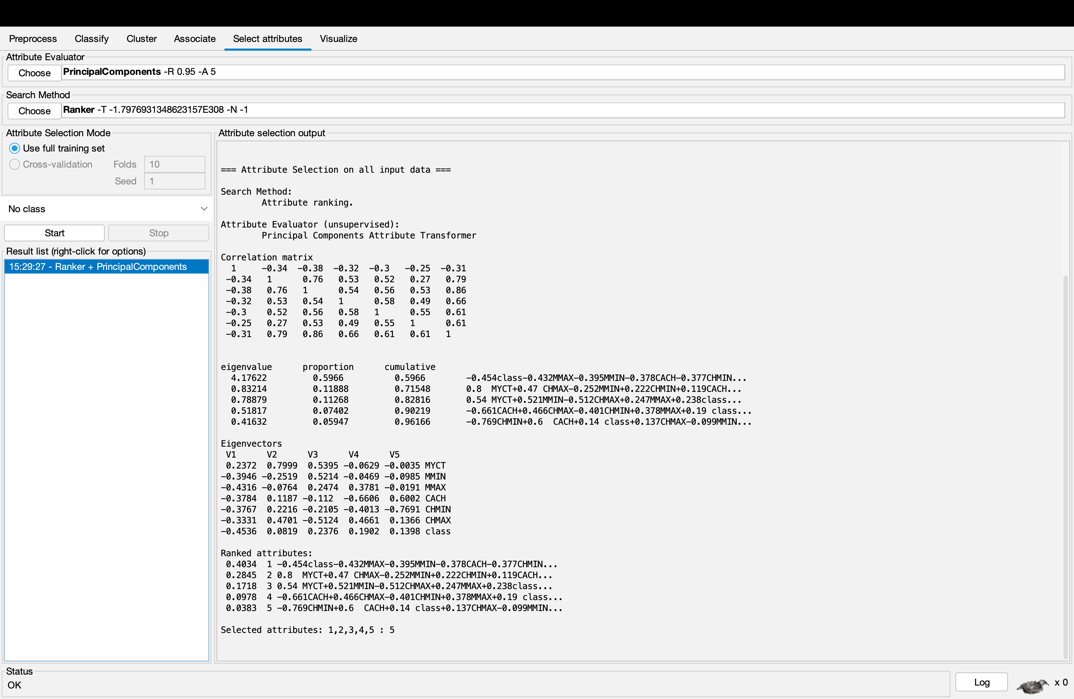


Filling in missing values in Weka by using filters



EXTRA TASK

Transformed data



Visualize transformed data

