**Data Mining**

**Mohammad Khattab**

A1) Identify the type of each attribute

|  |  |  |
| --- | --- | --- |
| Number | Attribute Name | Attribute Type |
| 1 | User\_id | Numeric |
| 2 | Gender | Nominal(symmetric) |
| 3 | Age | Numeric(ratio) |
| 4 | Marital\_status | Nominal(symmetric) binary |
| 5 | Website\_Activity | Nominal(ordinal) |
| 6 | Browed\_Electronics\_12Mo | Nominal(Asymmetric)binary |
| 7 | Bought\_Electronics\_12Mo | Nominal(Asymmetric)binary |
| 8 | Bought\_Digital\_Media\_18Mo | Nominal(Asymmetric)binary |
| 9 | Bought\_Digital\_Books | Nominal(Asymmetric)binary |
| 10 | Payment\_method | Nominal |

\*\*\* Because we are talking about a selling system, there are some attributes that are similar in importance (symmetric), and this means that I do not care if the seller or buyer is male or female or single or married \*\*\*

A2) Identify the values of the summarizing properties for each attribute

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attribute Name | Location | Max | Min | Range | Variance |
| User\_id | First att |  |  |  |  |
| Gender | Second att |  |  | 2 |  |
| Age | Third att | 70 | 17 | 70-17=53 | 199.98 |
| Marital\_status | Fourth att |  |  | 2 |  |
| Website\_Activity | Fifth att |  |  | 3 |  |
| Browed\_Electronics\_12Mo | Sixth att |  |  | 2 |  |
| Bought\_Electronics\_12Mo | Seventh att |  |  | 2 |  |
| Bought\_Digital\_Media\_18Mo | Eight att |  |  | 2 |  |
| Bought\_Digital\_Books | Nineth att |  |  | 2 |  |
| Payment\_method | Tenth att |  |  | 4 |  |

\*\*\*\*The range is the different between the highest and lowest value \*\*\*\*\*

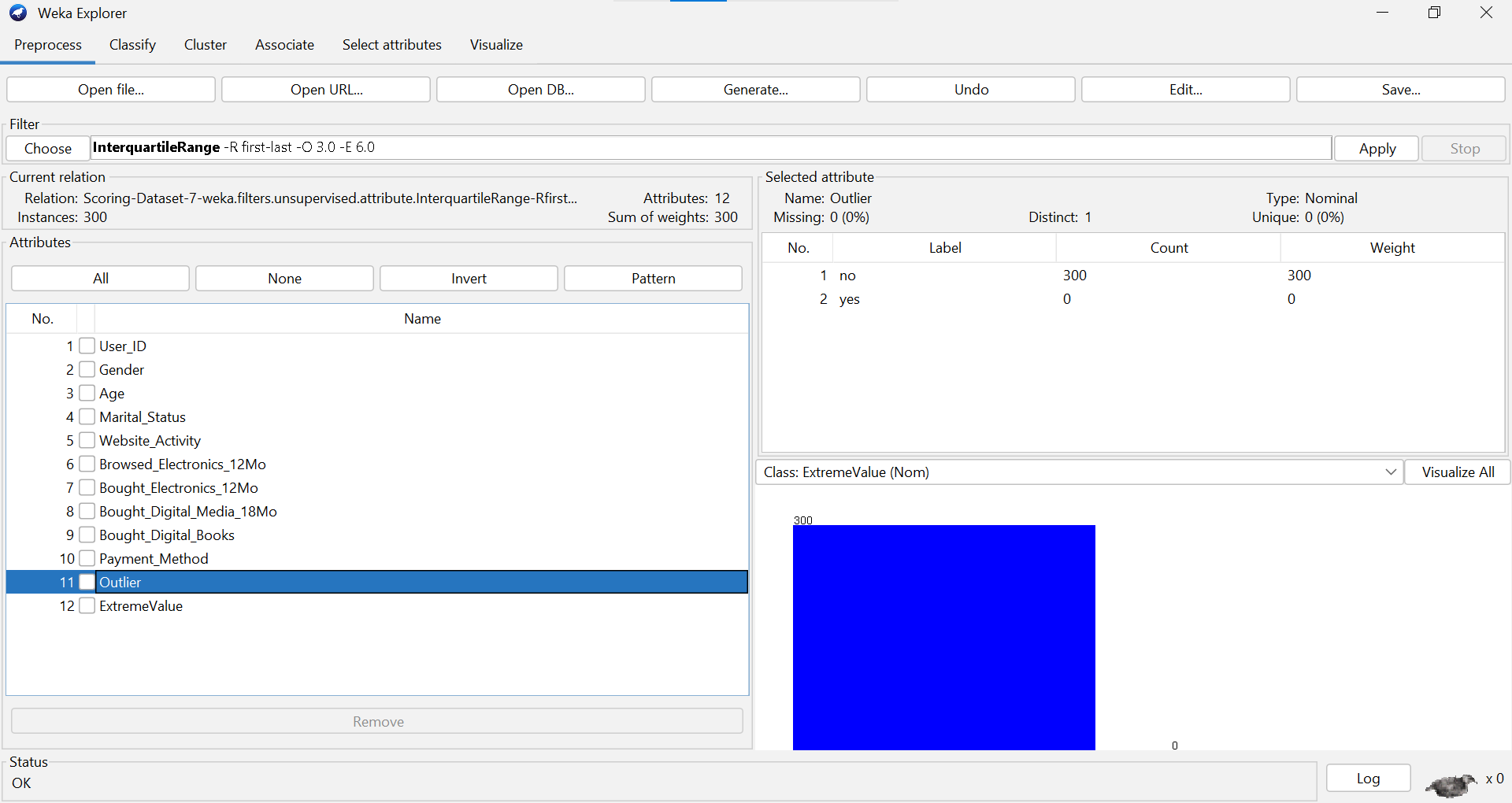
|  |  |  |  |
| --- | --- | --- | --- |
| Attribute Name | Mode | Median | Mean |
| User\_id |  |  |  |
| Gender | M |  |  |
| Age | 51 | 49 | 46.1 |
| Marital\_status | S |  |  |
| Website\_Activity | Seldom |  |  |
| Browed\_Electronics\_12Mo | Yes |  |  |
| Bought\_Electronics\_12Mo | No |  |  |
| Bought\_Digital\_Media\_18Mo | Yes |  |  |
| Bought\_Digital\_Books | No |  |  |
| Payment\_method | WebsitAccount |  |  |

\*\*\* The Mode means the most frequent value in this attribute \*\*\*

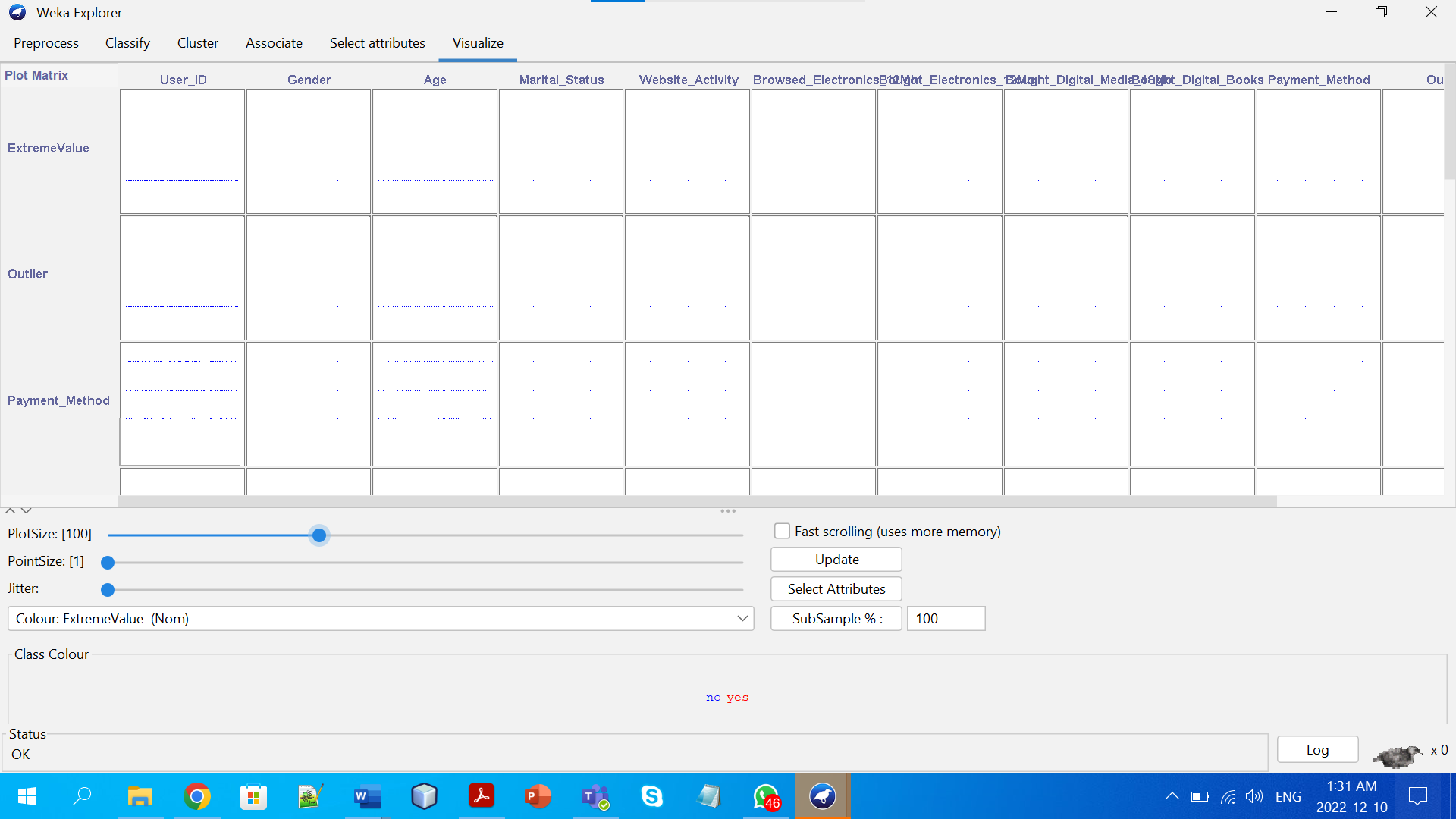
|  |  |
| --- | --- |
| Attribute Name | The frequency |
| User\_id |  |
| Gender | M:155/F:145 |
| Age | 17:6/ 18:3/19:2/20:0/21:4/22:3/23:3/24:2/25:4/26:6/27:3/  28:6/29:6/30:6/31:7/32:4/33:4/34:7/35:5/36:9/37:8/38:3/ 40:4/41:1/42:4/43:8/44:9/45:16/46:9/47:14/48:6/49:7/  50:8/51:6/52:8/53:8/54:6/55:4/56:9/57:9/58:10/59:3/  60:1/61:3/62:9/63:4/64:11/65:9/66:3/67:8/68:3/69:3/70:1 |
| Marital\_status | M:148/S:152 |
| Website\_Activity | Frequent:23/Regular:83/seldom:194 |
| Browed\_Electronics\_12Mo | Yes:289/No:11 |
| Bought\_Electronics\_12Mo | Yes:128/No:172 |
| Bought\_Digital\_Media\_18Mo | Yes:227/No:73 |
| Bought\_Digital\_Books | Yes:127/No:173 |
| Payment\_method | Monthiybilling:43/credit card:38/website account:119/bank transfer:100 |

A3) When there isn't an outlier the label assigned to yes

When there is an outlier the label assigned to no.



Examples of patterns in scatter plot (the number of scatter plots is zero in this data set (all relationships are Uncorrelated), because most of the data is of type nominal.)



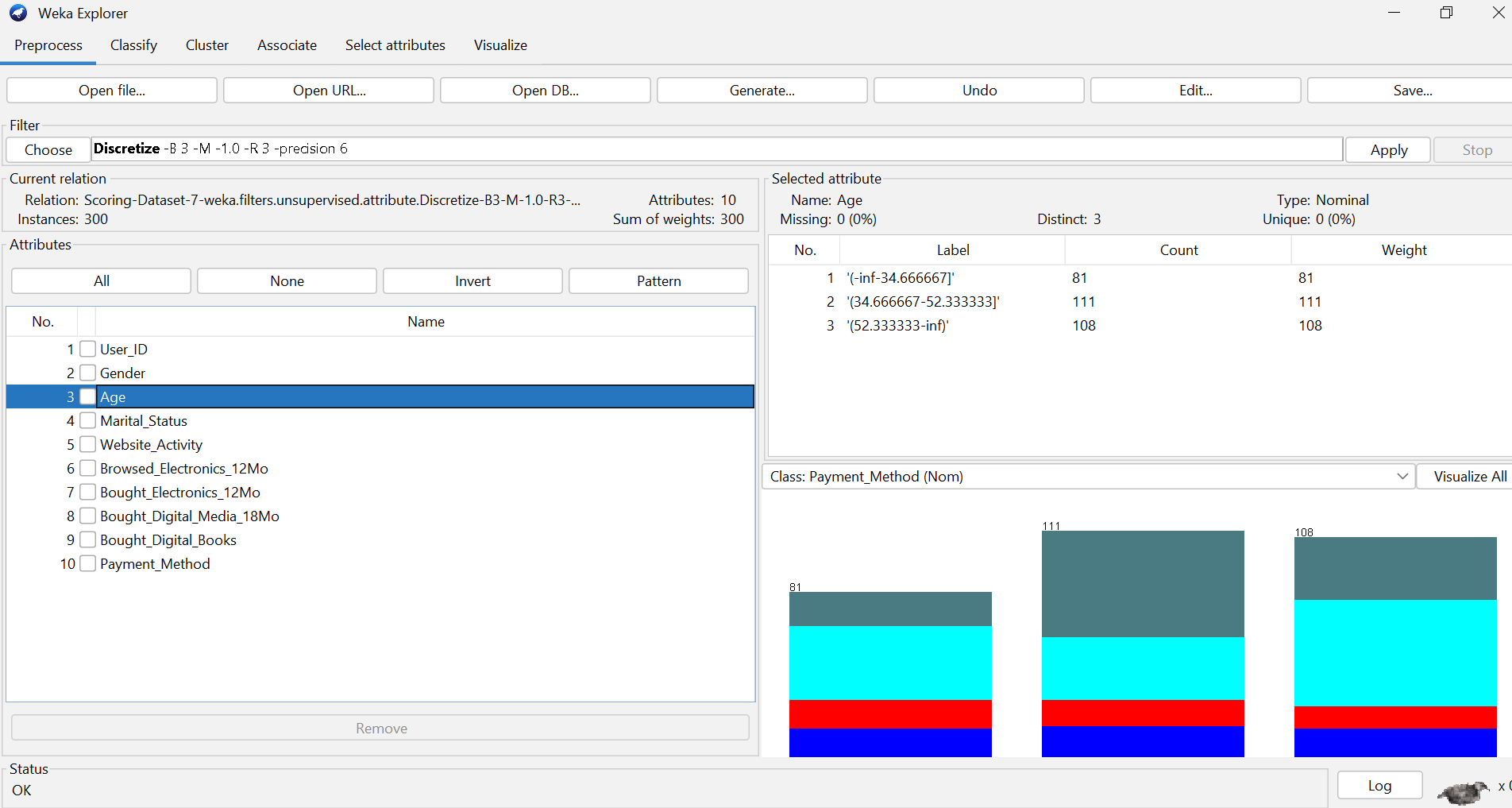
صورة تحتوي على منضدة

تم إنشاء الوصف تلقائياً

B1)Use the following binning techniques to smooth the value of the “Age” attribute:

1. Equi-width binning(3 bins):

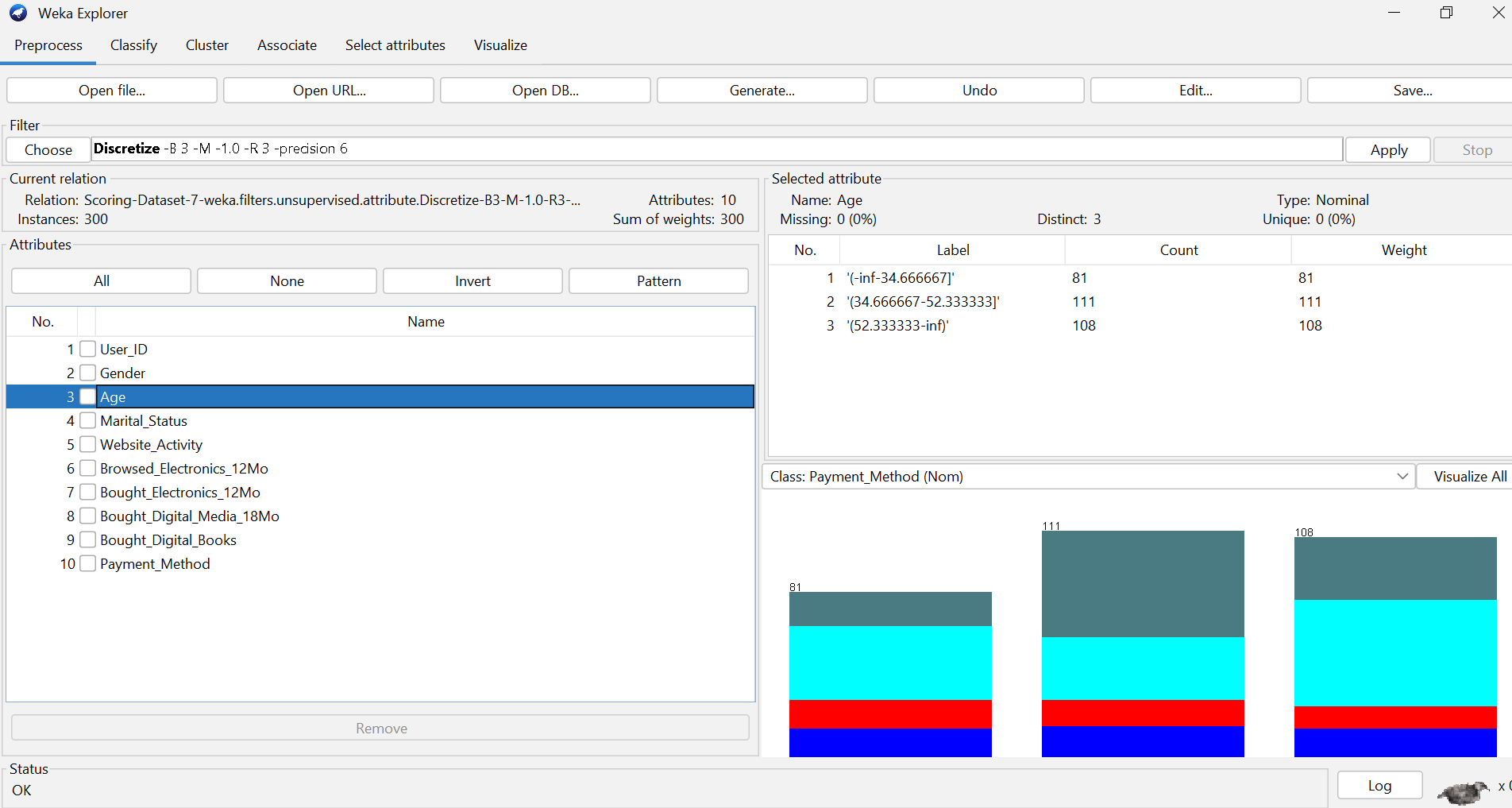
As🡺max=x,min=y,🡺width=(x-y)/3



\*as we see after applying the simple equation above this is what will appear 🡺3bins (width binning).

b)equi-depth binning(3 bins):

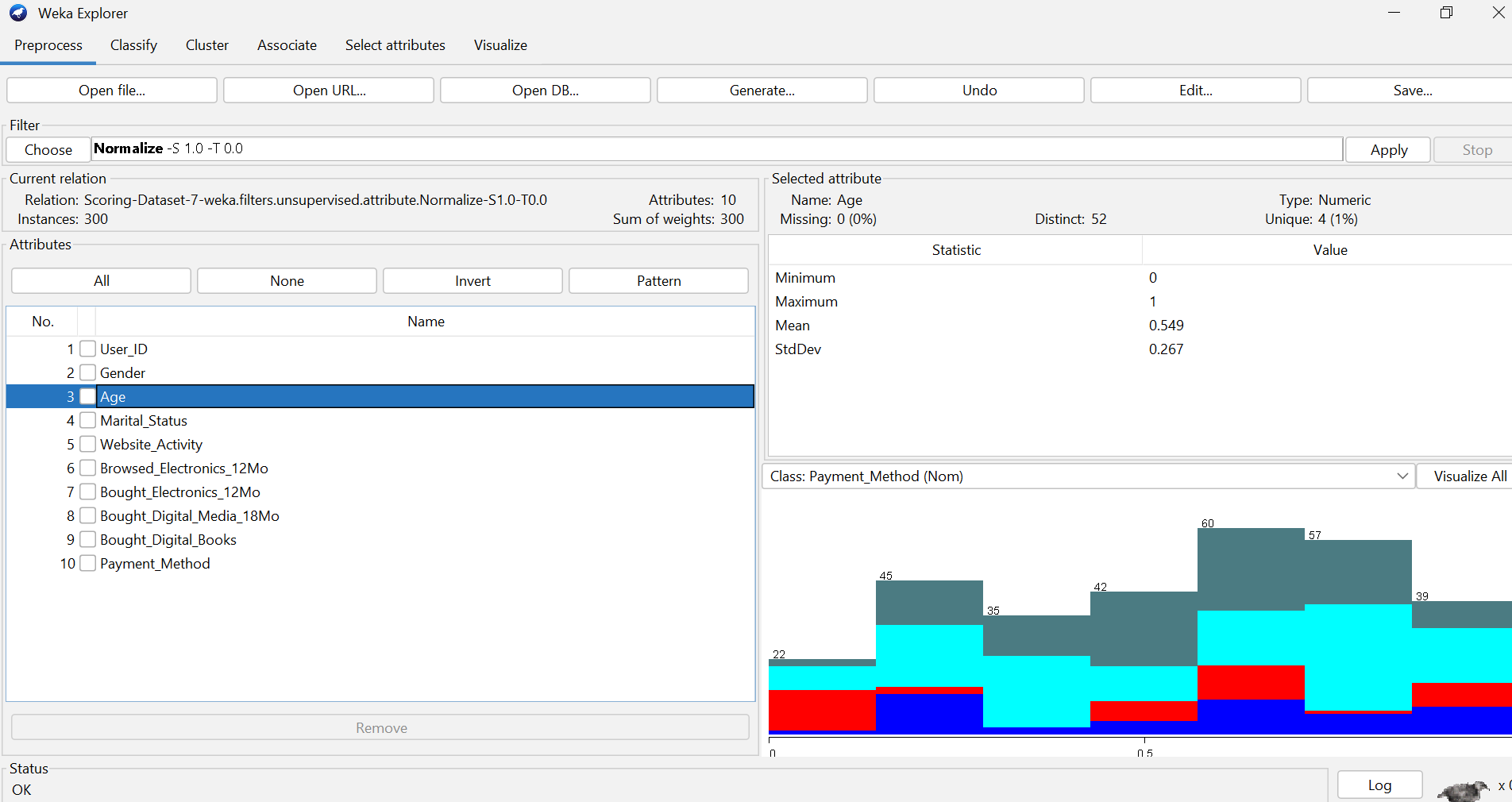
number of values=300, 🡺depth🡺300/3=100



\*after applying the equation we will see the 3 bins in depth.

B2)Use the following techniques to normalize the”Age” attribute:

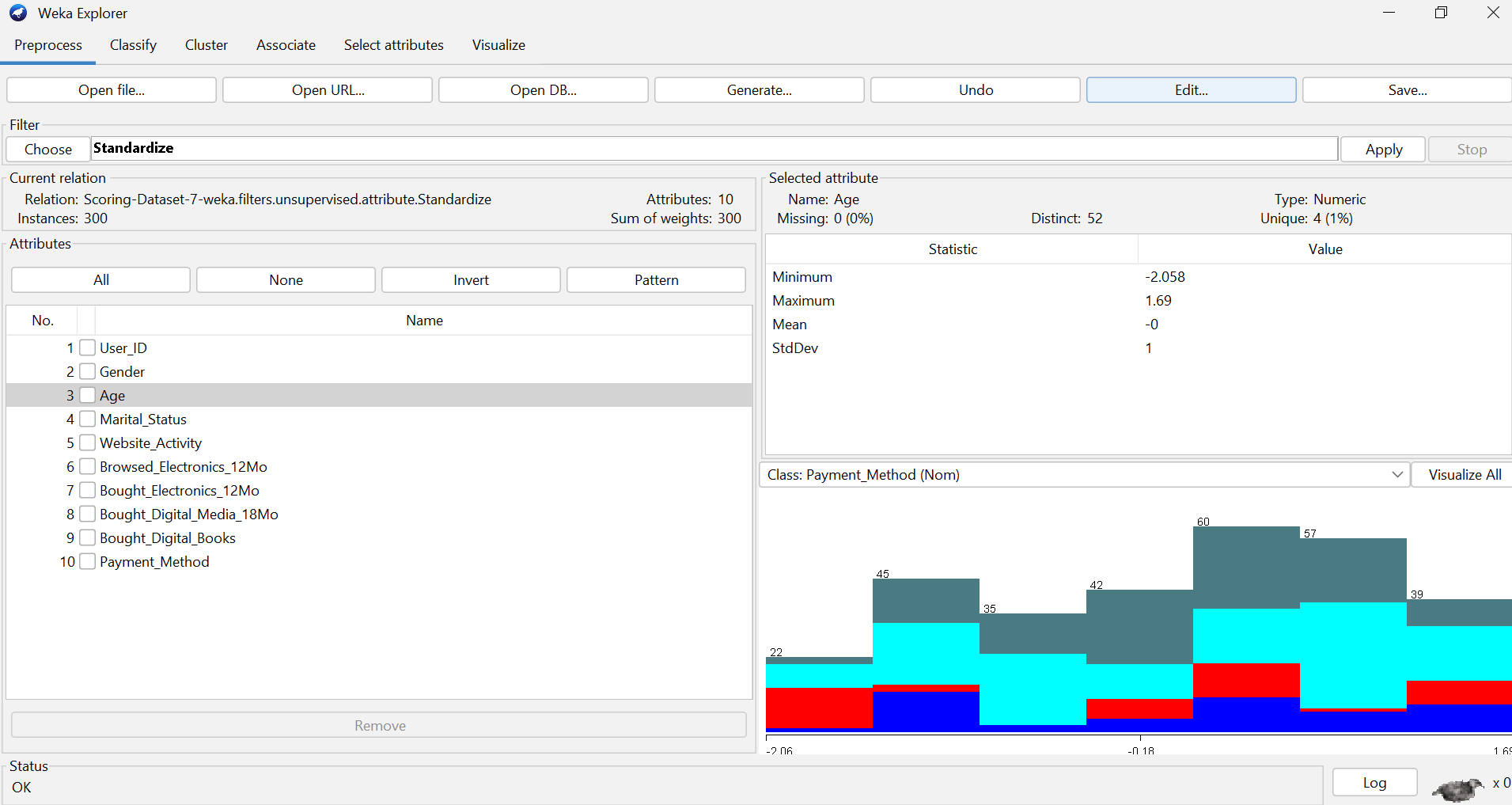
a)min-max normalization to transform the values onto the range[0.0-1.0]:



\*this is the max and min values after divided it between(0,1) interval.

b)Z-score normalization to transform the values:

\*)(x-mean)/std



\*after applying the z-score equation this is the values we will achieve.

B3)Discretise the Age attribute into the following categories:

1)Teenager=1-16:

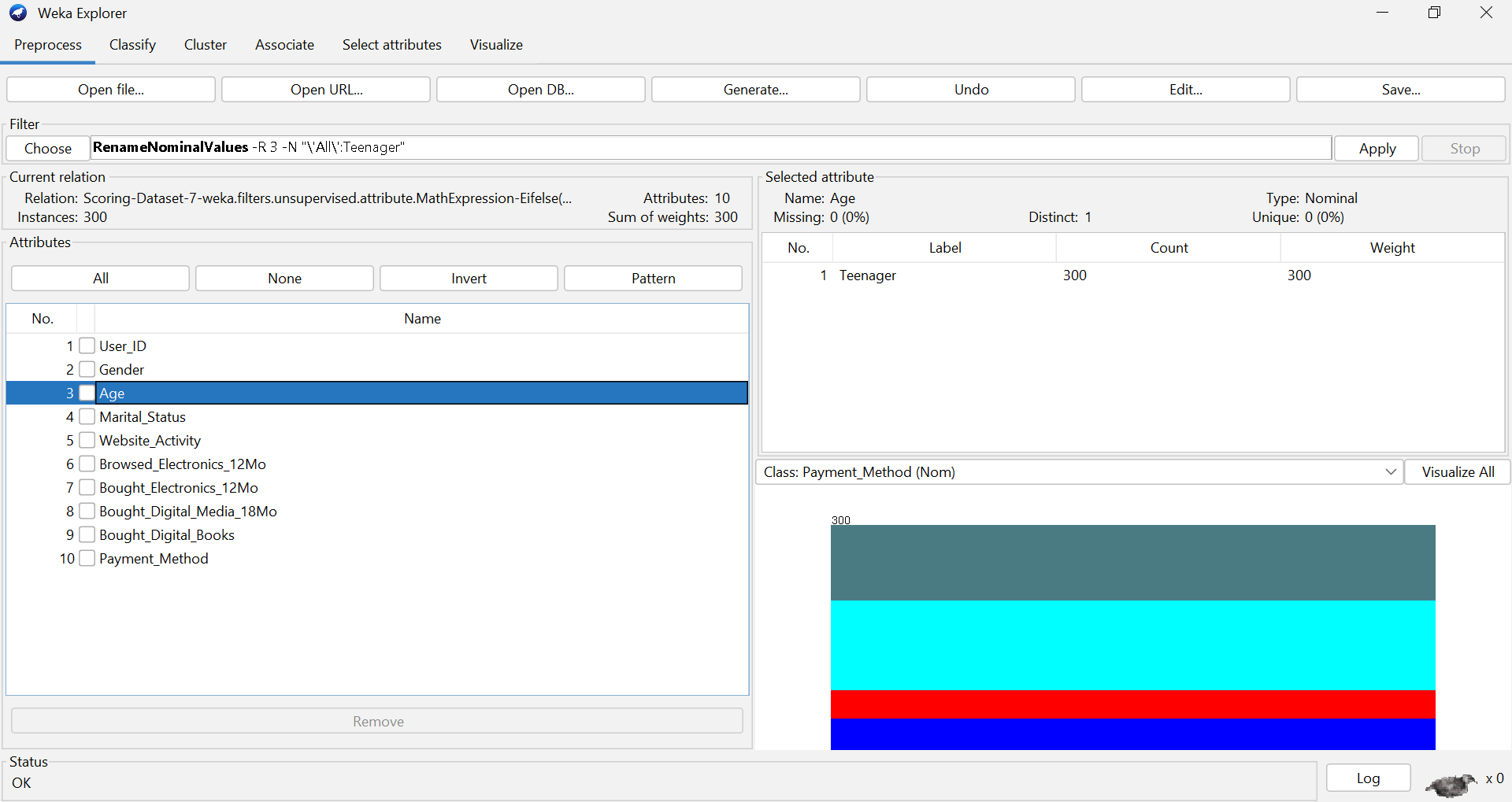
2)young=17-35:

3)Mid\_Age=36-55:

4)Mature=56-70:

5)Old=71+:

1)Teenager=1-16:



\*after dividing the age into age from 1 to 16 the result is the all values 🡺this indicates that is no one his age between this interval.

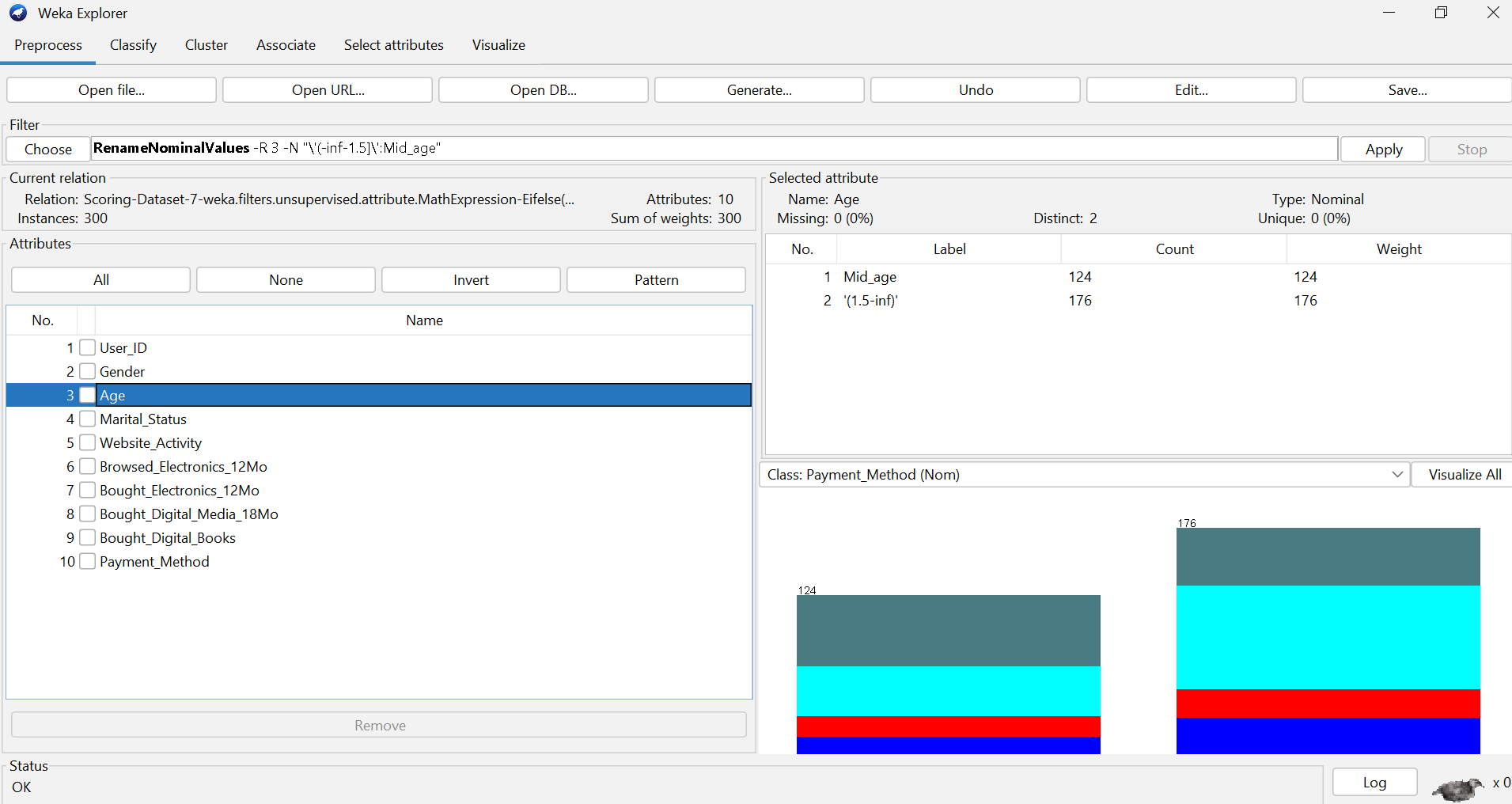
2)young=17-35:

Graphical user interface, application, Word

Description automatically generated

\*)after dividing the age into age between 17 and 35 the result is 85 persons his ages between this interval.

3)Mid\_Age=36-55:



\*)after dividing the age into age between 36 and 55 the result is 124 persons his ages between this interval.

4)Mature=56-70:

Graphical user interface, application

Description automatically generated

\*)after dividing the age into age between 56 and 70 the result is 91 persons his ages between this interval.

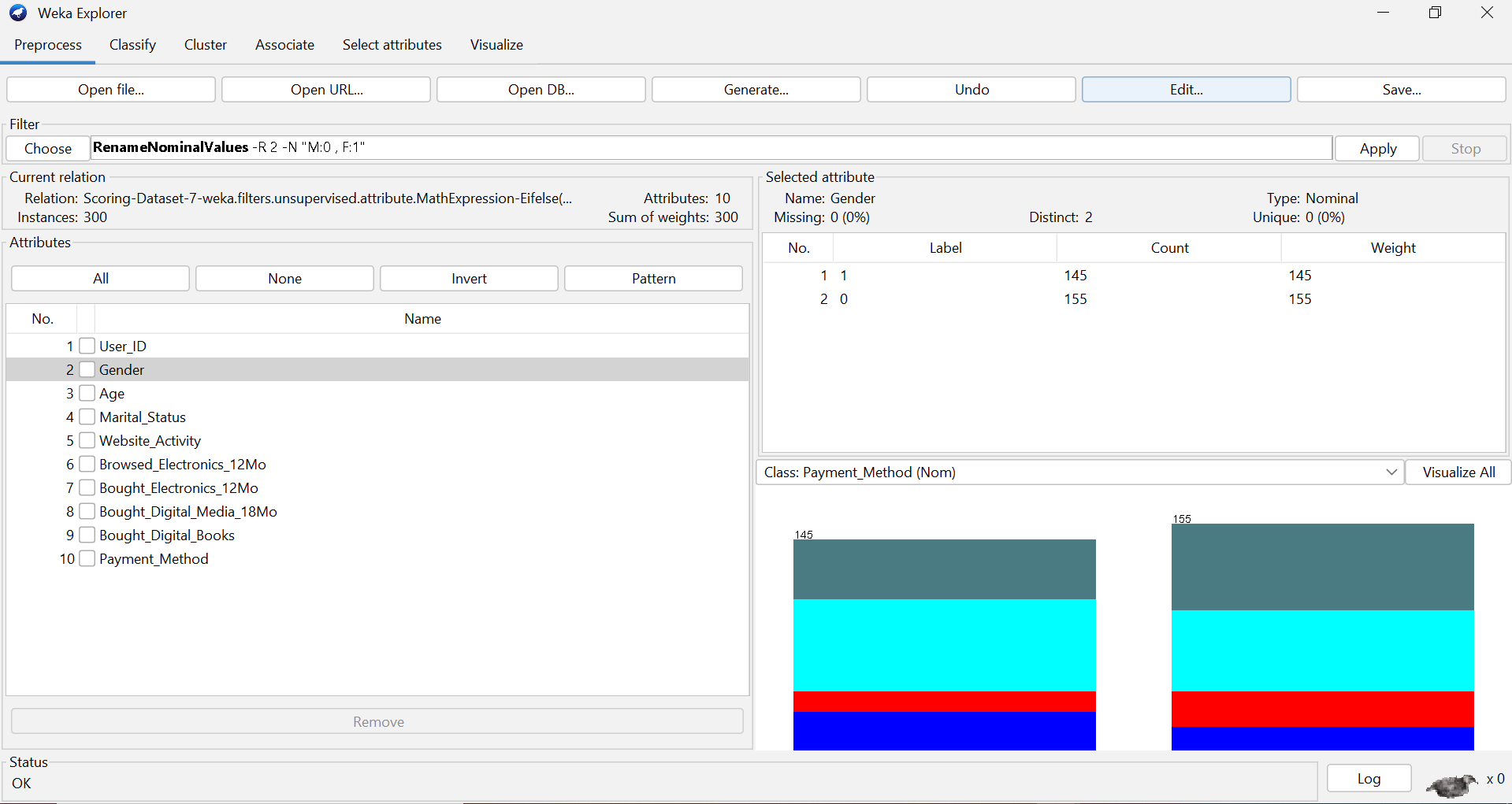
5)Old=71+:

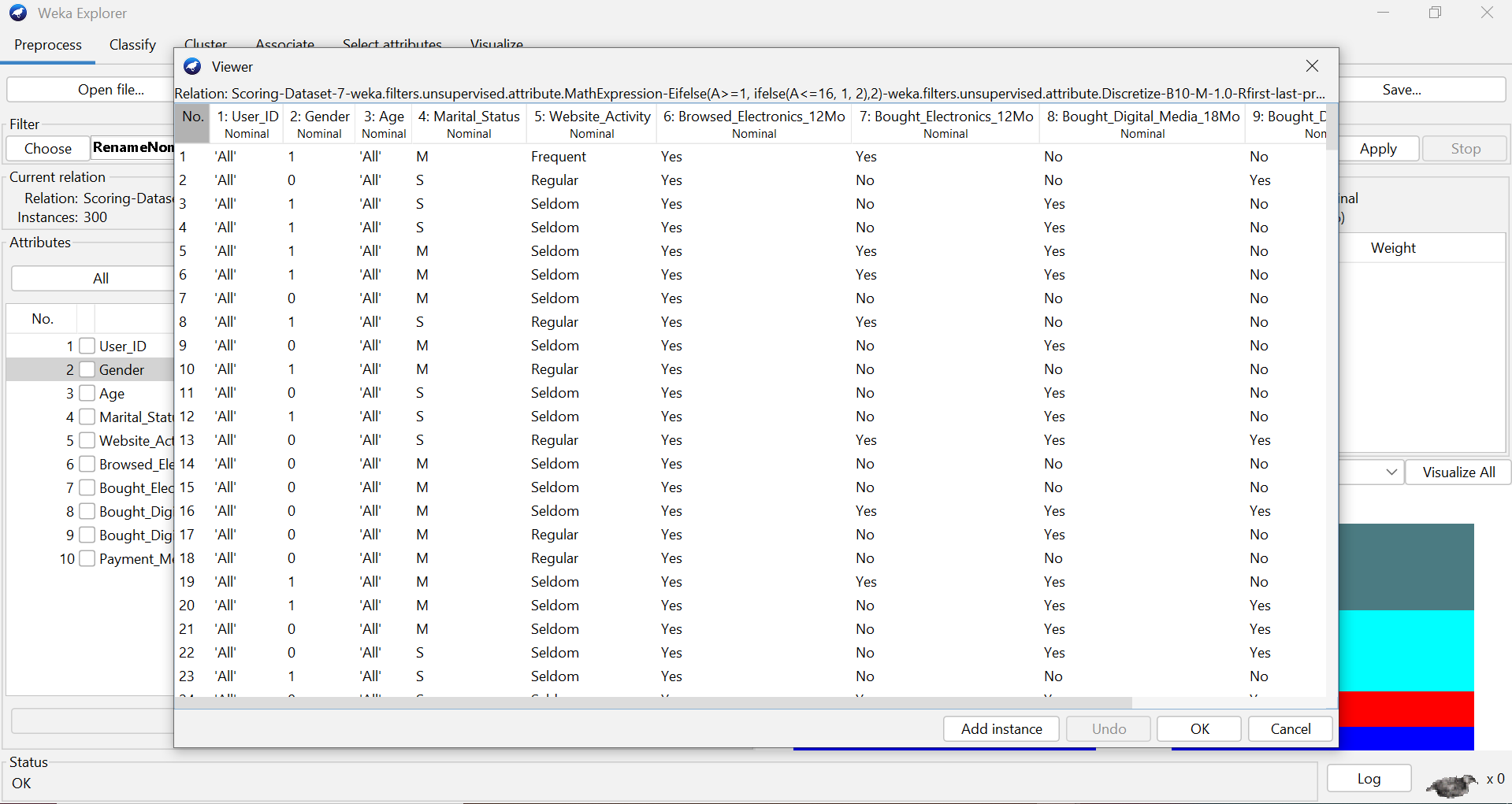
Graphical user interface, application

Description automatically generated

\*after dividing the age into ages more than 71 the result is the all values 🡺this indicates that is no one his age more than 71 years.

B4)Convert the “Gender” variable into binary variables[with values “0” or”1”]:





\*this is the result after converting each male to 1 and each female to 0.

C.1 : Extracting and Evaluating possible associations

Graphical user interface, text

Description automatically generated

C.2:Explanation Of Associations rules

First rule :

If Professional = No and Support\_Group = No ,Then Political=No

The confidence of this rule 0.94

# of objects that contain left side = 1167

# of objects that contain all rule = 1102

Confidence = Appearance of all rule / Appearance of left side 🡺 1102/1167 = 0.94

Second rule :

If Professional = No ,Then Political = No

The confidence of this rule 0.93

# of objects that contain left side = 1341

# of objects that contain all rule = 1245

Confidence = Appearance of all rule / Appearance of left side 🡺 1245/1341 = 0.93

Third rule :

If Social\_Club = No and Support\_Group = No ,Then Political = No

The confidence of this rule 0.92

# of objects that contain left side = 1410

# of objects that contain all rule = 1300

Confidence = Appearance of all rule / Appearance of left side 🡺 1300/1410 = 0.92