**EXERCISES LAB 1- 2**

**Bài tập 1: Tải mỗi bộ dữ liệu và kiểm tra các thông tin sau đây về từng bộ dữ liệu:**

**Airline**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| What’s the name of the relation: | Number of records in each dataset | How many instances does each dataset have? | How many attributes? | List attribute names and their types. |
| Airline\_passengers | **144** | **144** | **2** |  |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | What are the values that each attribute can have? | Missing: | Distinct: | Unique |
| **pasenger\_numbers** | **real** | **0 (0%)** | **118** | **93 (65%)** |
| **Date** | **Date** | **0 (0%)** | **144** | **144 (100%)** |

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

- Xác định thuộc tính lớp (nếu có). Xác định số bản ghi cho mỗi lớp.

- Vẽ Histogram.

- Hiển thị dữ liệu trong các chiều khác nhau. Điều gì xảy ra nếu thuộc tính lớp được để lại ở cột đầu tiên? So sánh các hình vẽ của 2 bộ dữ liệu (một có lớp ở cột đầu tiên, trước khi có bất kỳ xử lý nào, và một có lớp ở cột cuối cùng) khi chúng xuất hiện ở bên phải của bảng tiền xử lý: bạn có thể nhận thấy sự khác biệt nào không?

- Kiểm tra dữ liệu thông qua trình xem để chỉnh sửa (Tải lên 5 bộ dữ liệu bất kỳ).

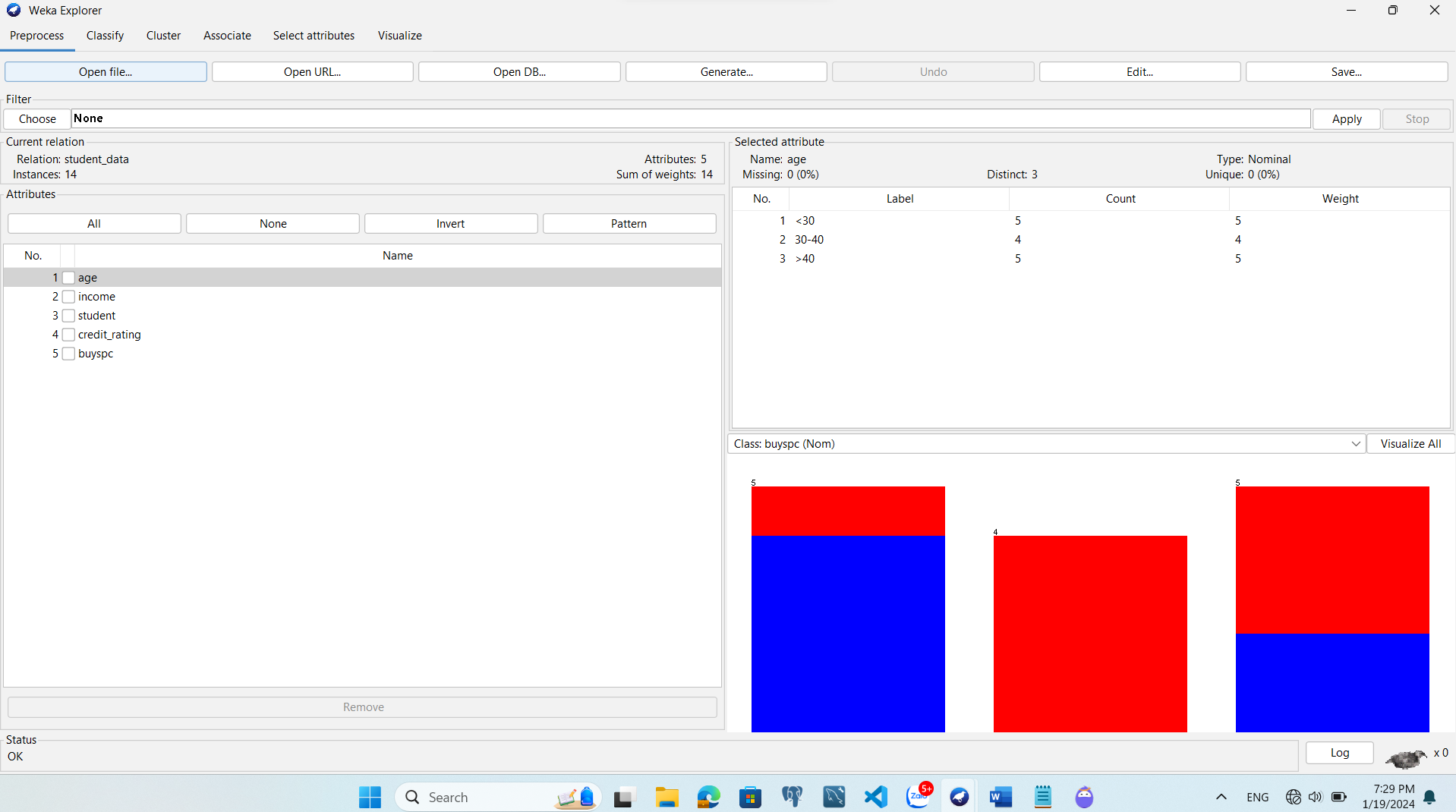
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **airline arff:** |  |  |  |  |
| What’s the name of the relation? | Airline\_passengers |  |  |  |  |
| Number of records in each dataset |  |  |  |  |  |
| How many instances does each dataset have? |  |  |  |  |  |
| How many attributes? |  |  |  |  |  |
| List attribute names and their types. |  |  |  |  |  |

Exercises 2: Create Dataset student .arff

|  |  |
| --- | --- |
| @relation student\_data  @attribute age {<30, 30-40, >40}  @attribute income {low, medium, high}  @attribute student {no, yes}  @attribute credit\_rating {fair, excellent}  @attribute buyspc {no, yes}  @data  <30,high,no,fair,no  <30,high,no,excellent,no  30-40,high,no,fair,yes  >40,medium,no,fair,yes  >40,low,yes,fair,yes  >40,low,yes,excellent,no  30-40,low,yes,excellent,yes  <30,medium,no,fair,no  <30,low,yes,fair,no  >40,medium,yes,fair,yes  <30,medium,yes,excellent,yes  30-40,medium,no,excellent,yes  30-40,high,yes,fair,yes  >40,medium,no,excellent,no |  |

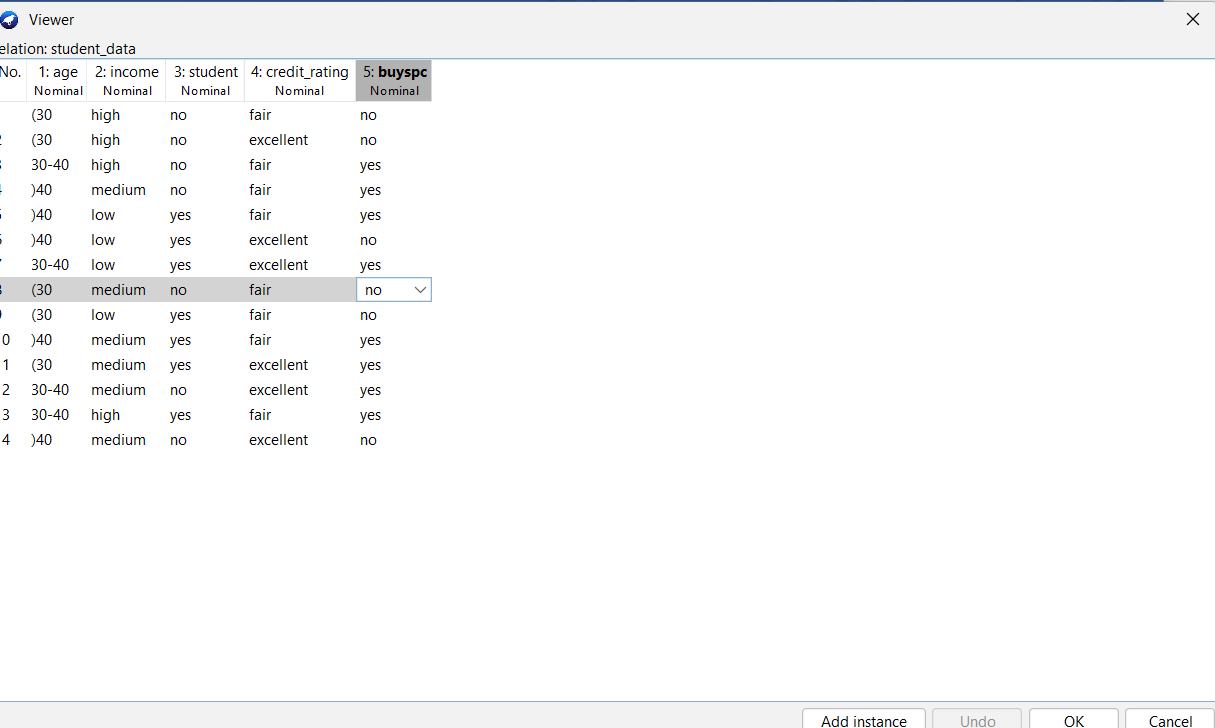
Exercises 3: Preprocessing with WEKA-Explorer on the dataset student.arff

- Loading the dataset into Explorer.



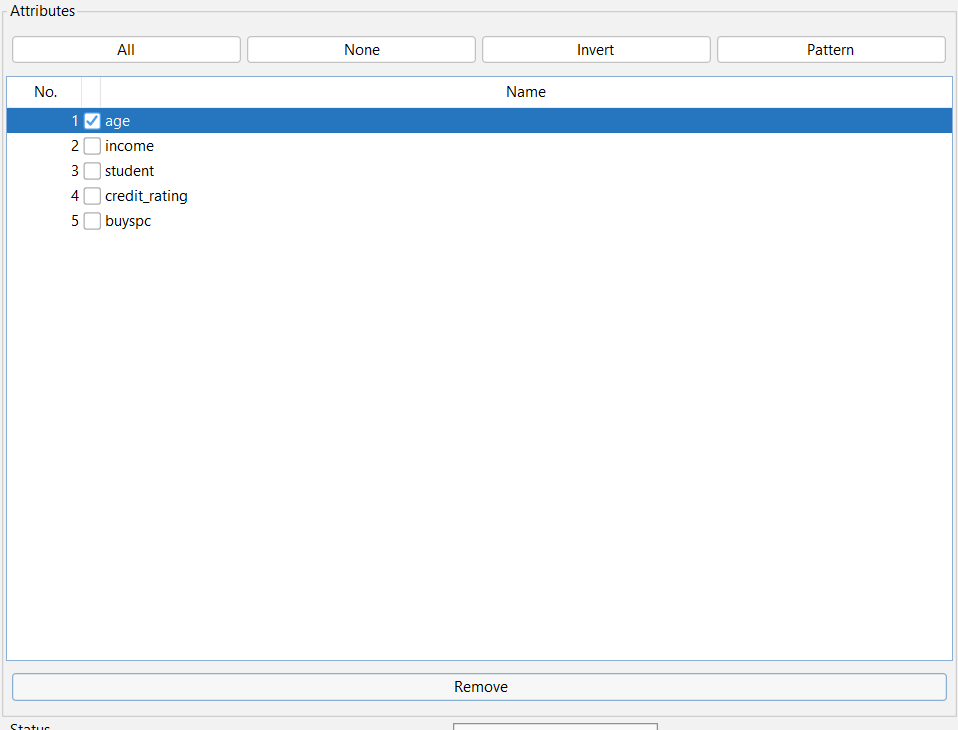
- Examming dataset with requirements in exercise 1

- What is the class value of instance number 8 in the student data?

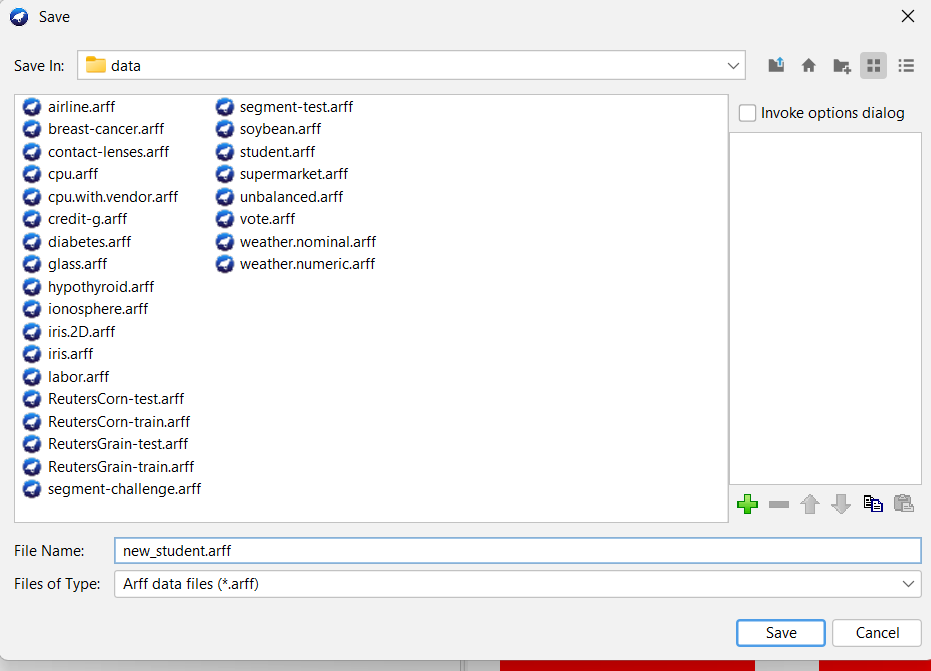


- Preprocessing data by selecting or filtering attributes.

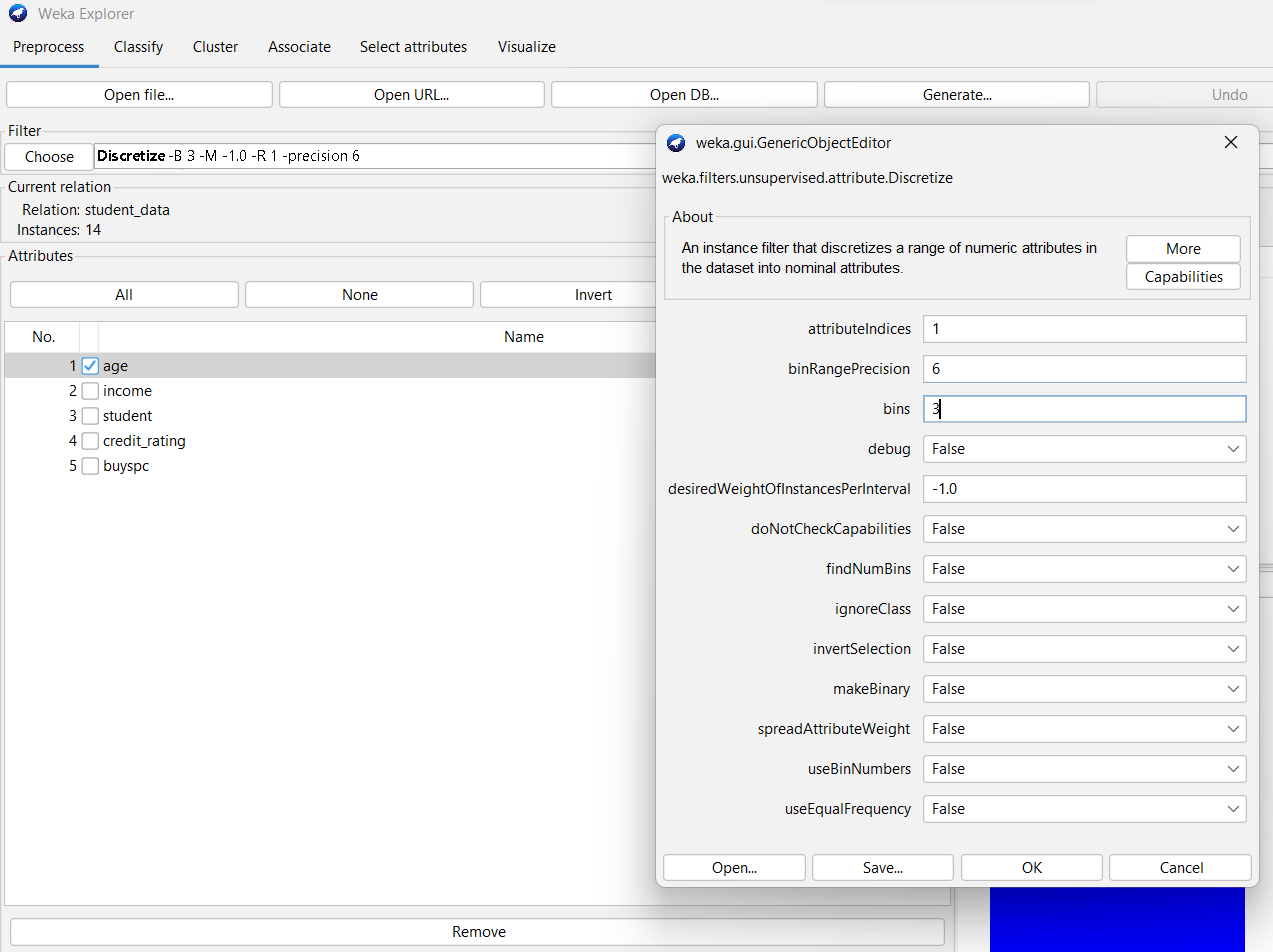
• Removing an attribute



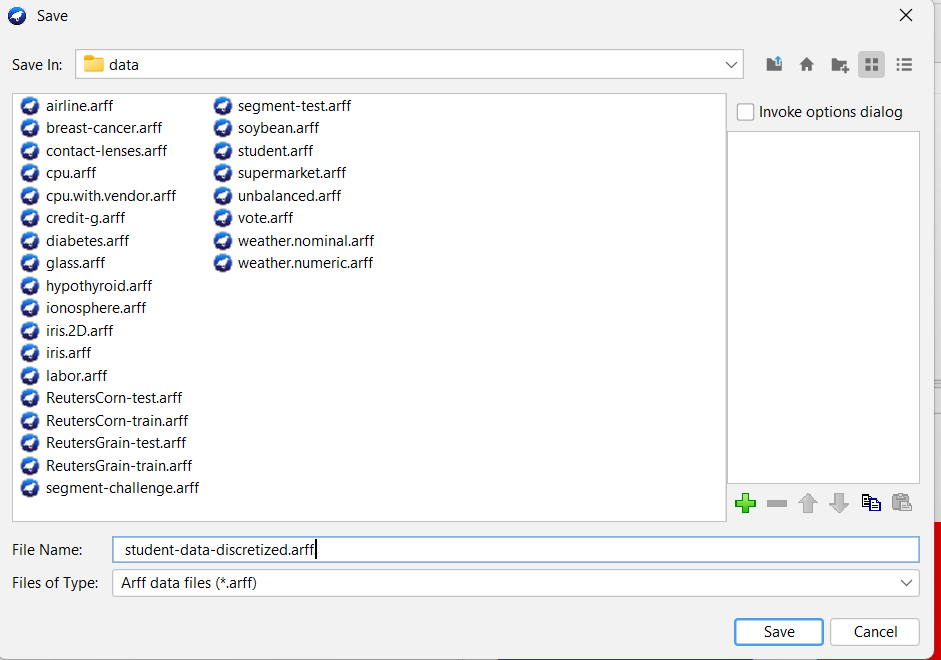
• Saving the new working relation as an arff file



• Discretizing the age attribute with the number of bins is 3



• Saving the new working relation in a file called student-data-discretized.arff



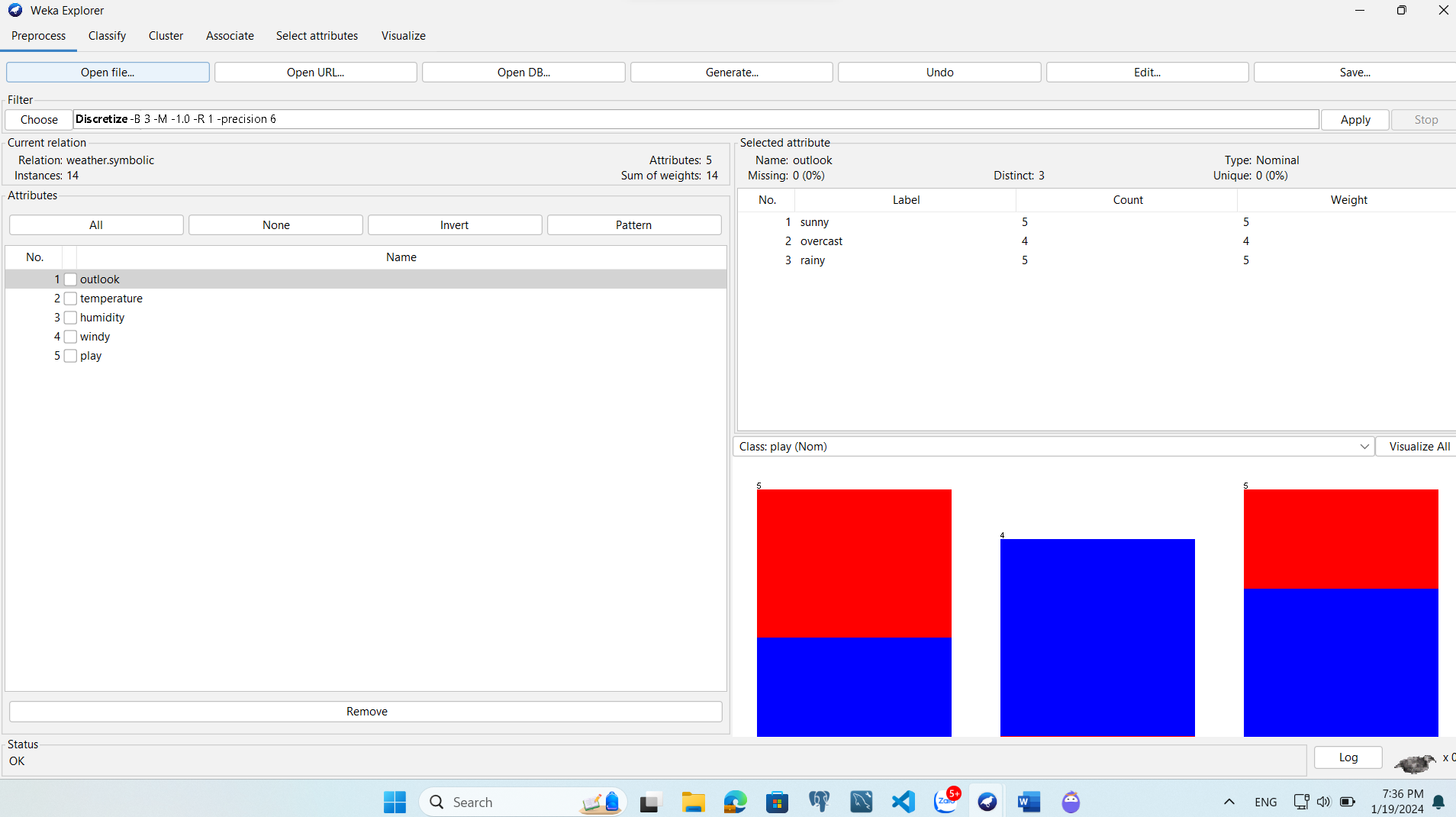
Exercises 4: Preprocessing with WEKA-Explorer on Nominal data: the Weather (nominal)

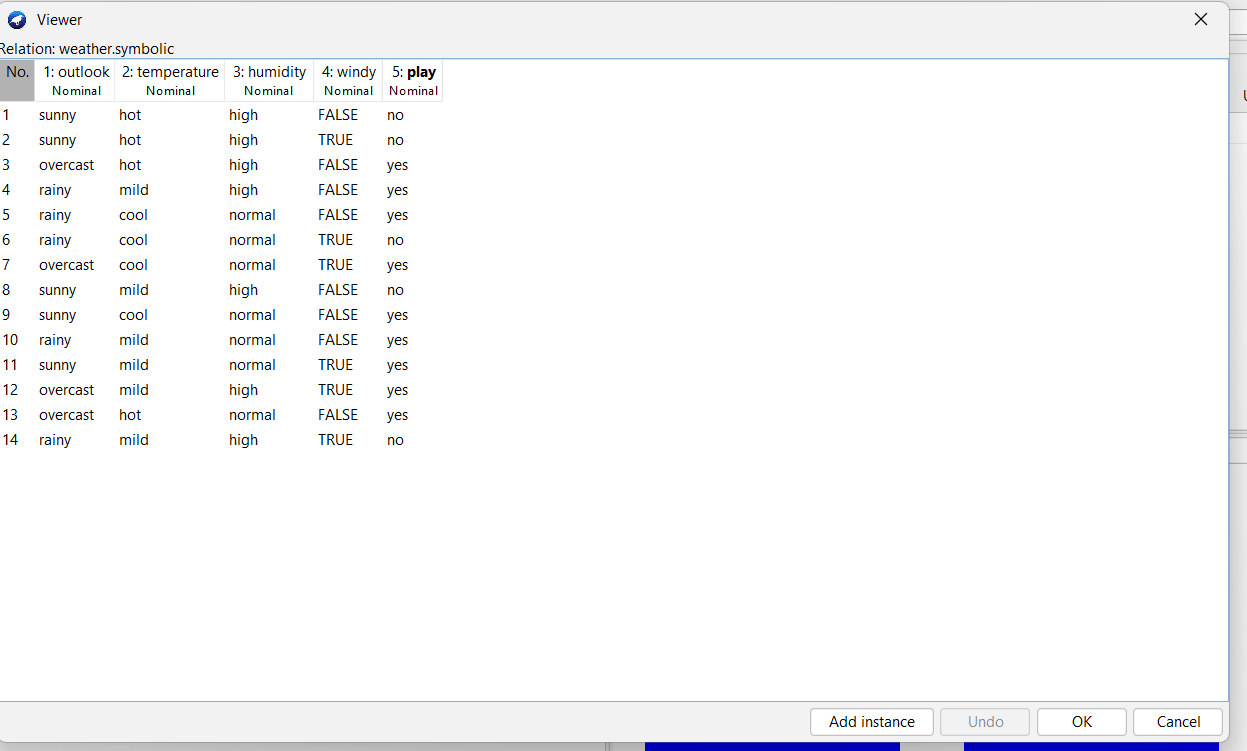
dataset.

- Loading the dataset into Explorer.

- How many instances are listed in the dataset? How many attributes? What’s the

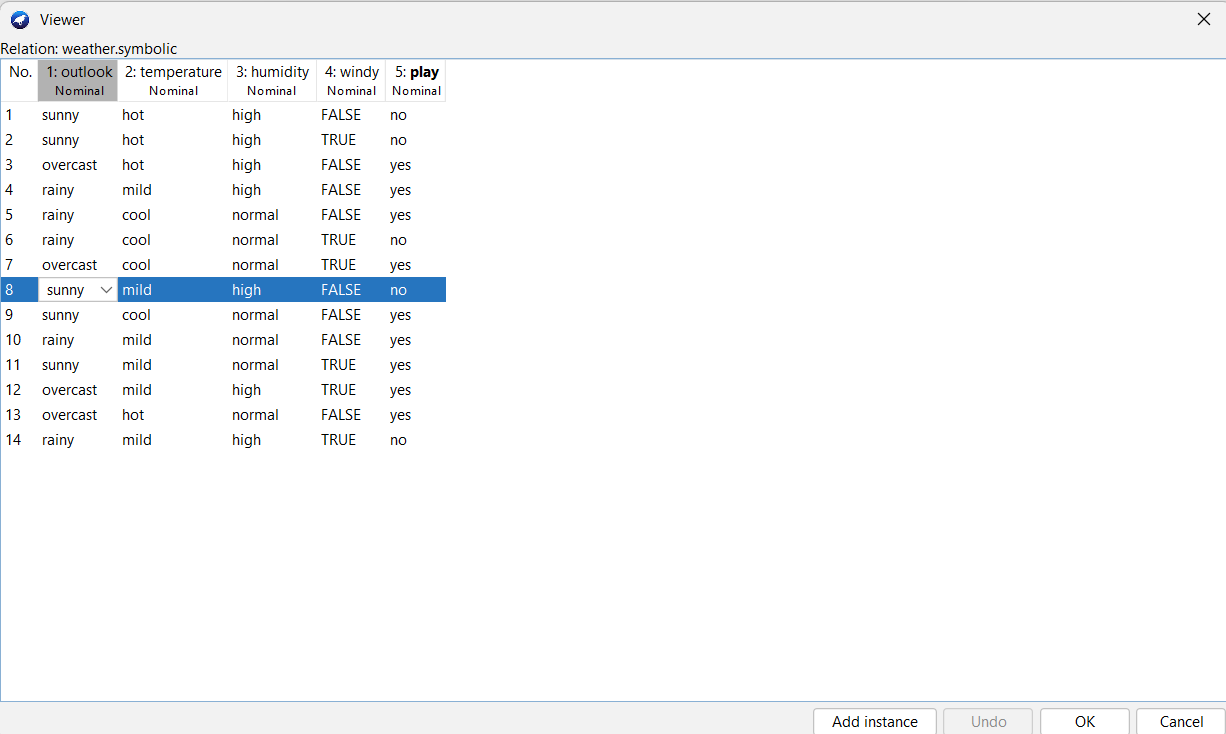
attributes names? What type of attributes are these? Which one is the class?





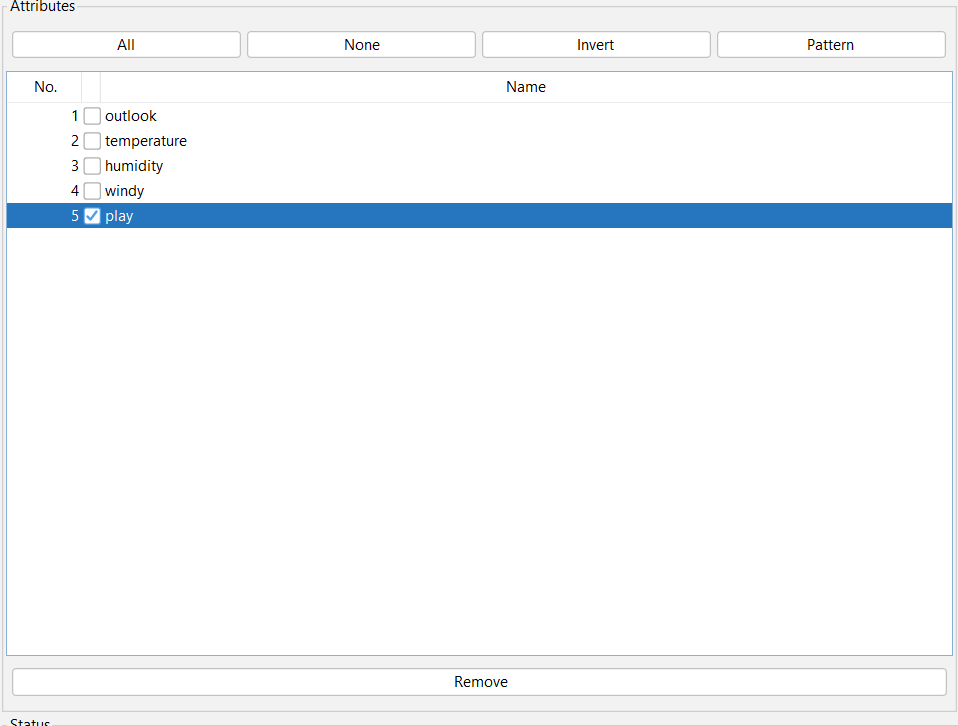
- What is the function of the first column in the Viewer window? What is the class value of

instance number 8 in the weather data?

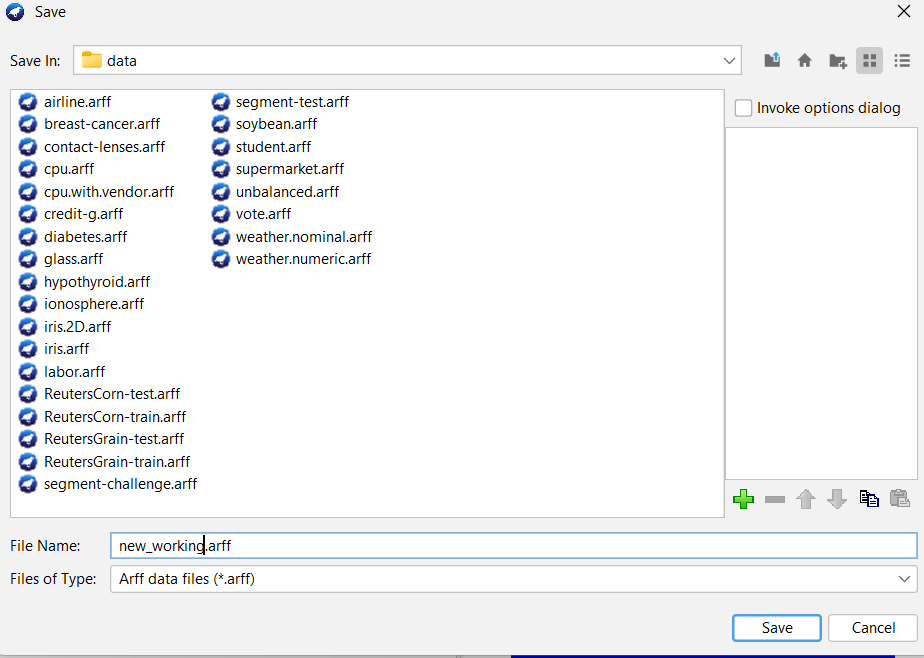


- Proprocessing data

• Removing an attribute from the dataset

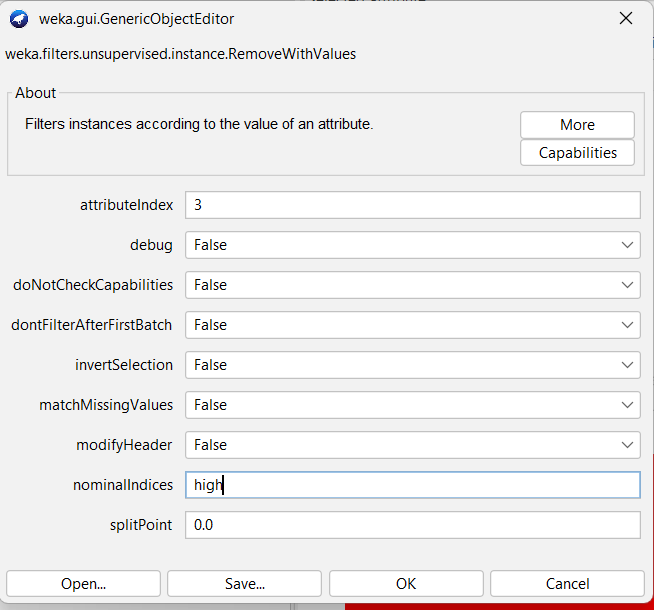


• Saving the new working relation as an arff file



• Using the filter weka.unsupervised.instance.RemoveWithValues to remove all instances

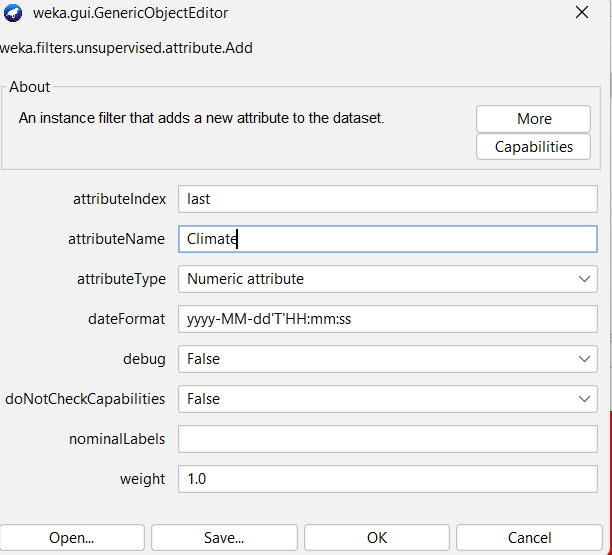
in which the humidity attribute has the value high.



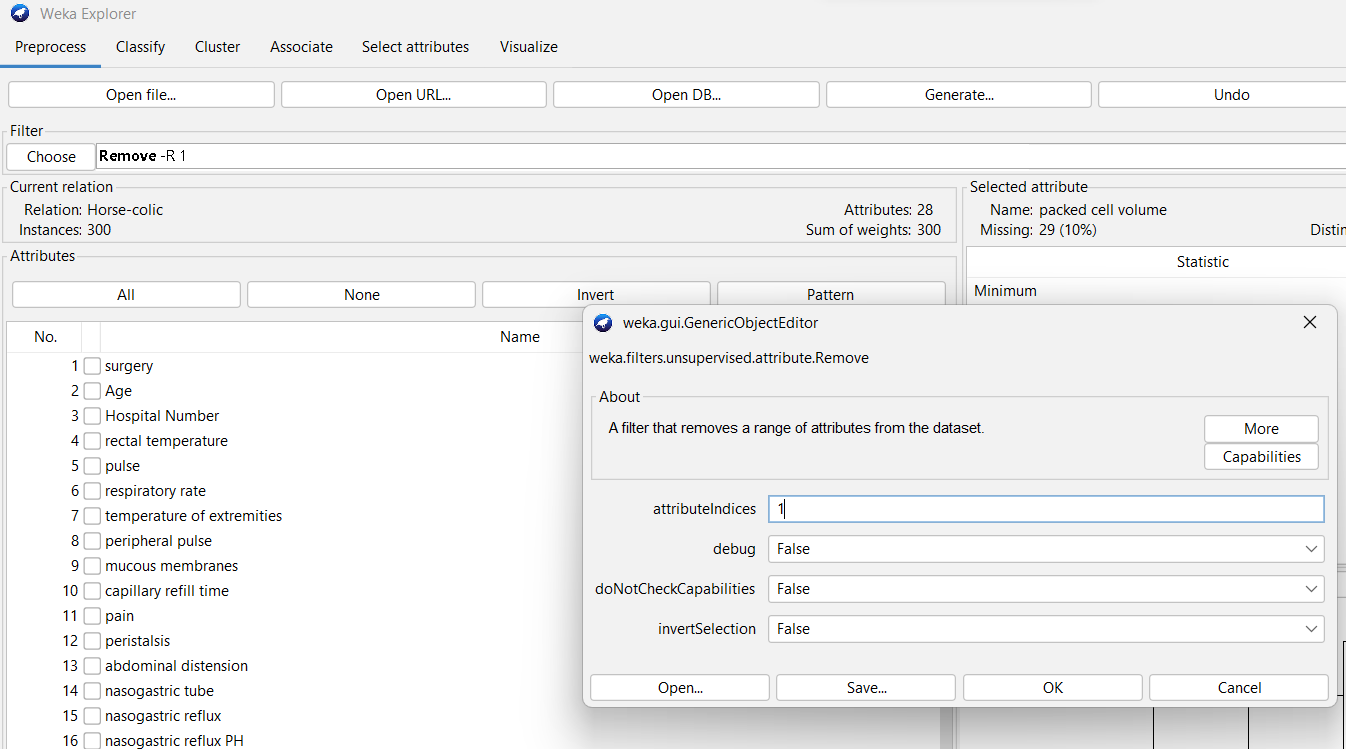
• Observing changes throught a viewer for editing

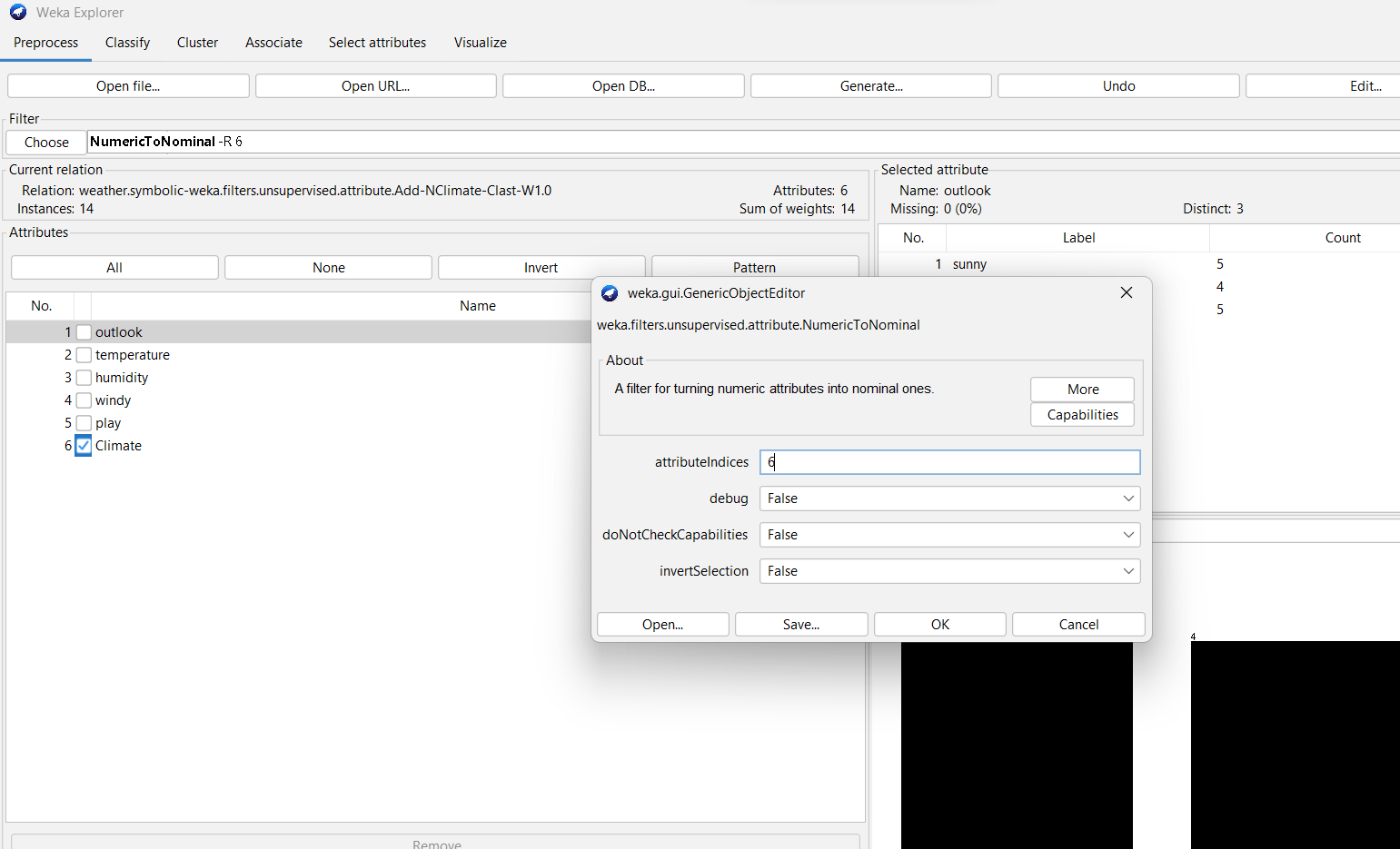
Exercises 5: Apply Preprocessing techniques to the dataset of Weather.

- Add a Climate attribute.



- Remove an attribute (select any attribute)



- Normalization: select numeric attributes to normalize

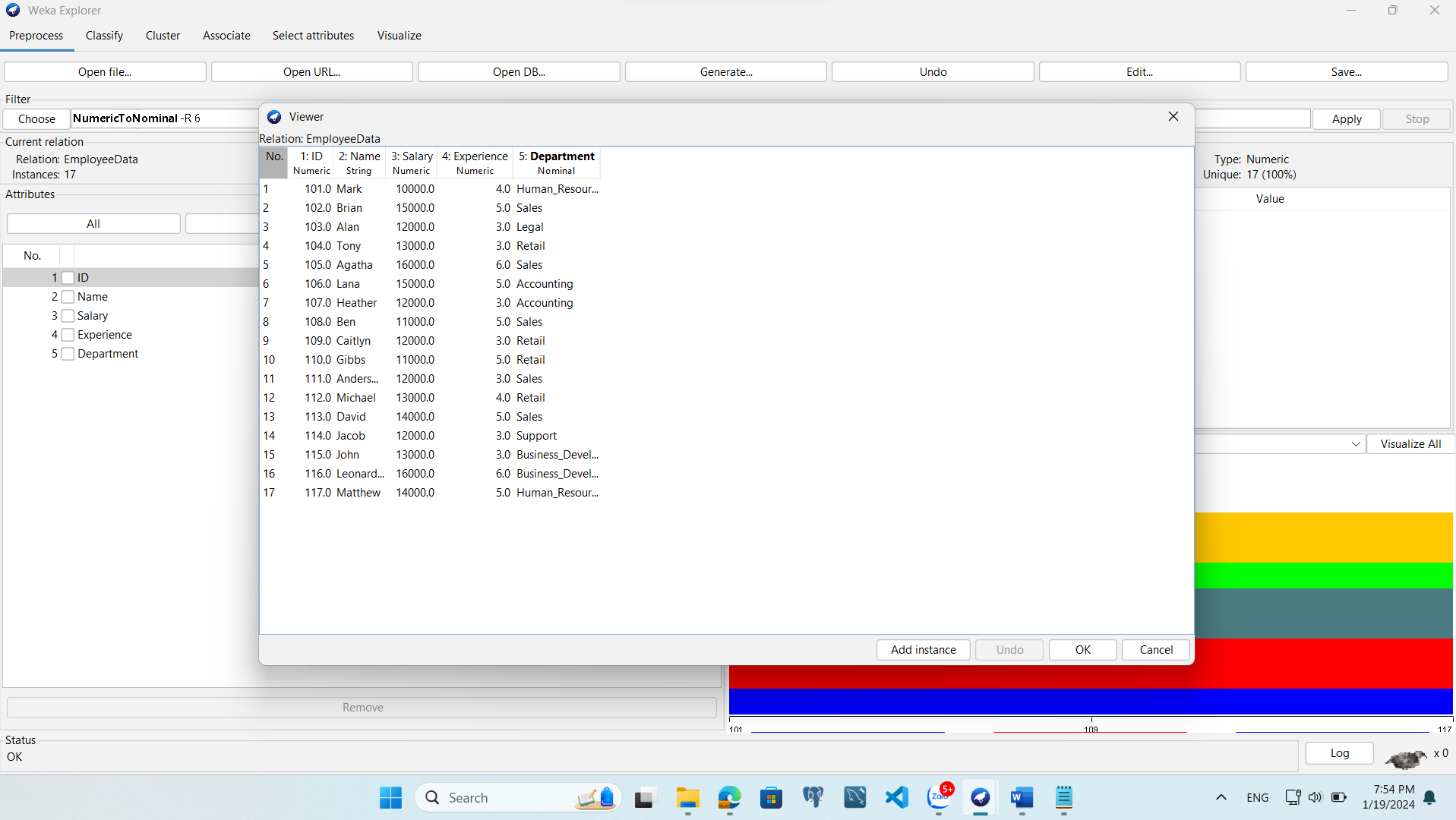
Exercises 6:

- Create a dataset Employee as:

@relation EmployeeData

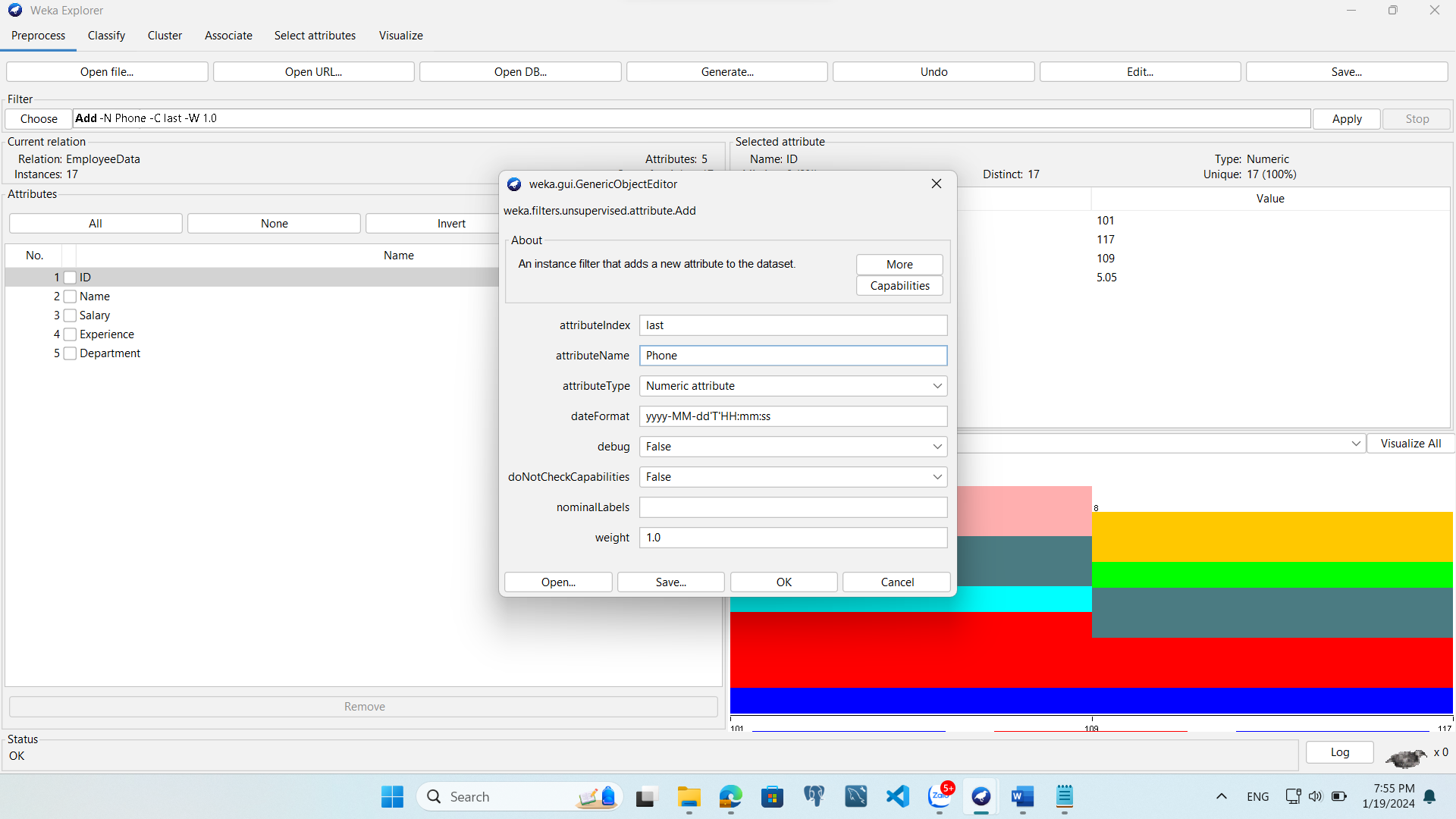
|  |
| --- |
| @attribute ID numeric  @attribute Name string  @attribute Salary numeric  @attribute Experience numeric  @attribute Department {Human\_Resources, Sales, Legal, Retail, Accounting, Support, Business\_Development} |

|  |
| --- |
| @data  101,Mark,10000,4,Human\_Resources  102,Brian,15000,5,Sales  103,Alan,12000,3,Legal  104,Tony,13000,3,Retail  105,Agatha,16000,6,Sales  106,Lana,15000,5,Accounting  107,Heather,12000,3,Accounting  108,Ben,11000,5,Sales  109,Caitlyn,12000,3,Retail  110,Gibbs,11000,5,Retail  111,Anderson,12000,3,Sales  112,Michael,13000,4,Retail  113,David,14000,5,Sales  114,Jacob,12000,3,Support  115,John,13000,3,Business\_Development  116,Leonardo,16000,6,Business\_Development  117,Matthew,14000,5,Human\_Resources |

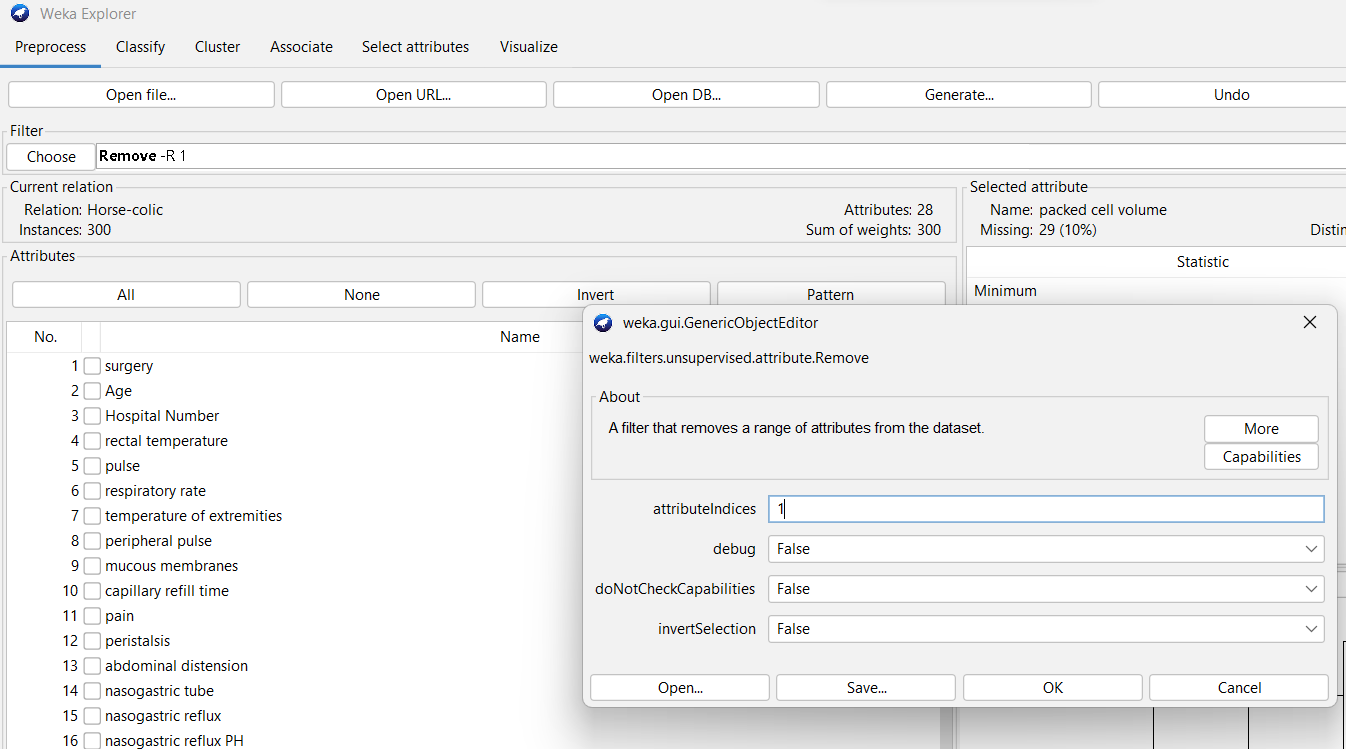


Apply Preprocessing techniques to this dataset.

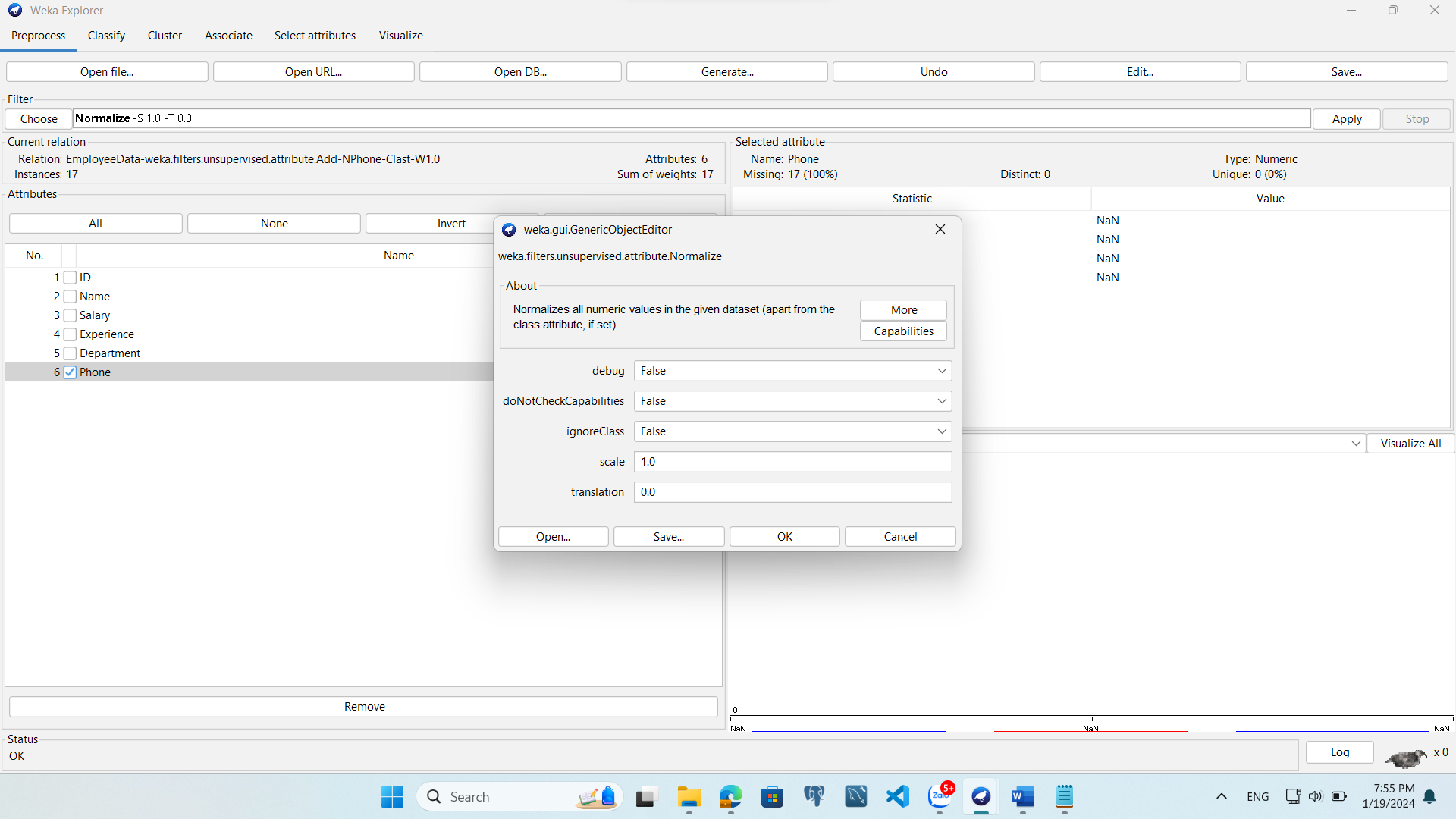
• Add a phone number attribute.



• Remove an attribute (select any attribute)



• Normalization: select numeric attributes to normalization



Exercises 7: Download and Exam dataset horse-colic

