Data Warehousing and Data Mining

Tutorial 3

Student Details

Name : Krunal Rank
Adm. No. : U18CO081

Suppose that the data for analysis includes the attribute age. The age values for the data tuples are (in increasing order) 13, 15, 16, 16, 19, X, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, Y, 45, 46, 52, 70.

Explain imputation using KNN. Find the value of the X and Y for above data using the same.

K Nearest Neighbours Imputation uses the existing records to determine the missing value. For a given record with a missing value, the algorithm tries to find K records that are nearest to the given record based on the common non missing values of attributes. The value that is used to determine how near two records are can be either the Euclidean distance or Manhattan distance. In Euclidean distance, the attribute values of records are subtracted and their Root Mean Square values are used. For the above data,
With K = 1, X = 19, Y = 36
With K = 2, X = 19.5, Y = 40.5

Explain missing value replacement (imputation) with an example. Also, find the value of the X and Y for above data using median value replacement.

Missing Value Replacement is a technique that is used to detect missing values based on other attribute values and existing records with known values for that particular attribute.
Missing Values can be replaced using following techniques:

- Replacing using Measures of Central Tendencies
- Replacing using Regression models or Trees
- Replacing using a constant value
- Replacing using a distinct value
- Replacing using a random value

For given data, using Median Replacement technique, X = Y = x_{13} = 25

Suppose that the data for analysis includes the attribute Colour such as red, green, blue, black, red, green, green, green, blue, red, black, white.

Explain Most frequent value replacement with example. Also, find the value of the X and Y for above data using Most frequent value replacement.

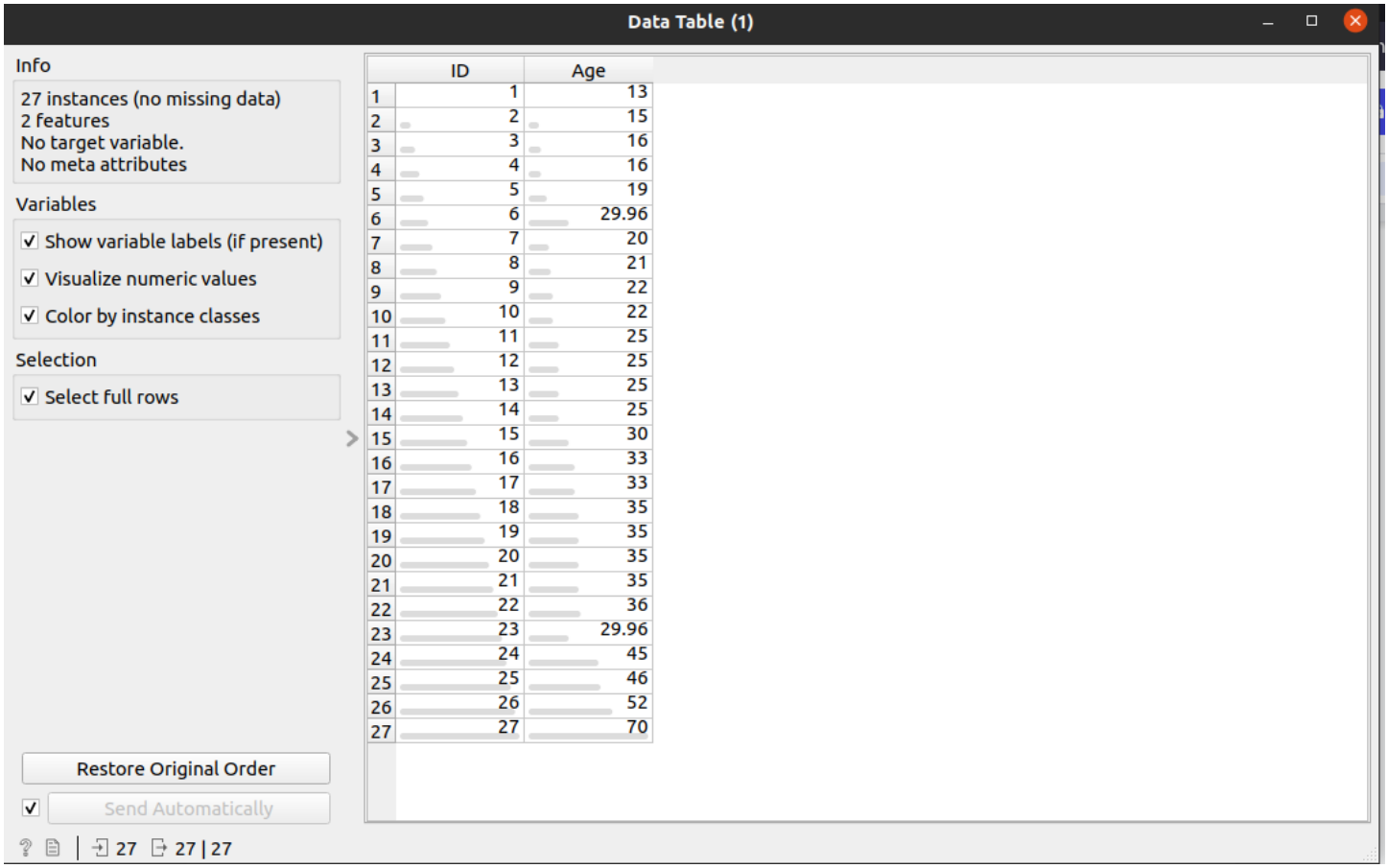
For Most Frequent Value Replacement, we use Mode as Missing Value Replacement. Hence, X = Y = Green

Explain global constant replacement with an example. Also, find the value of the X and Y for above data using global constant value replacement (Consider Red as global constant).

For Global Constant Value Replacement, we use Global Constant as Missing Value Replacement. Hence, X = Y = Red

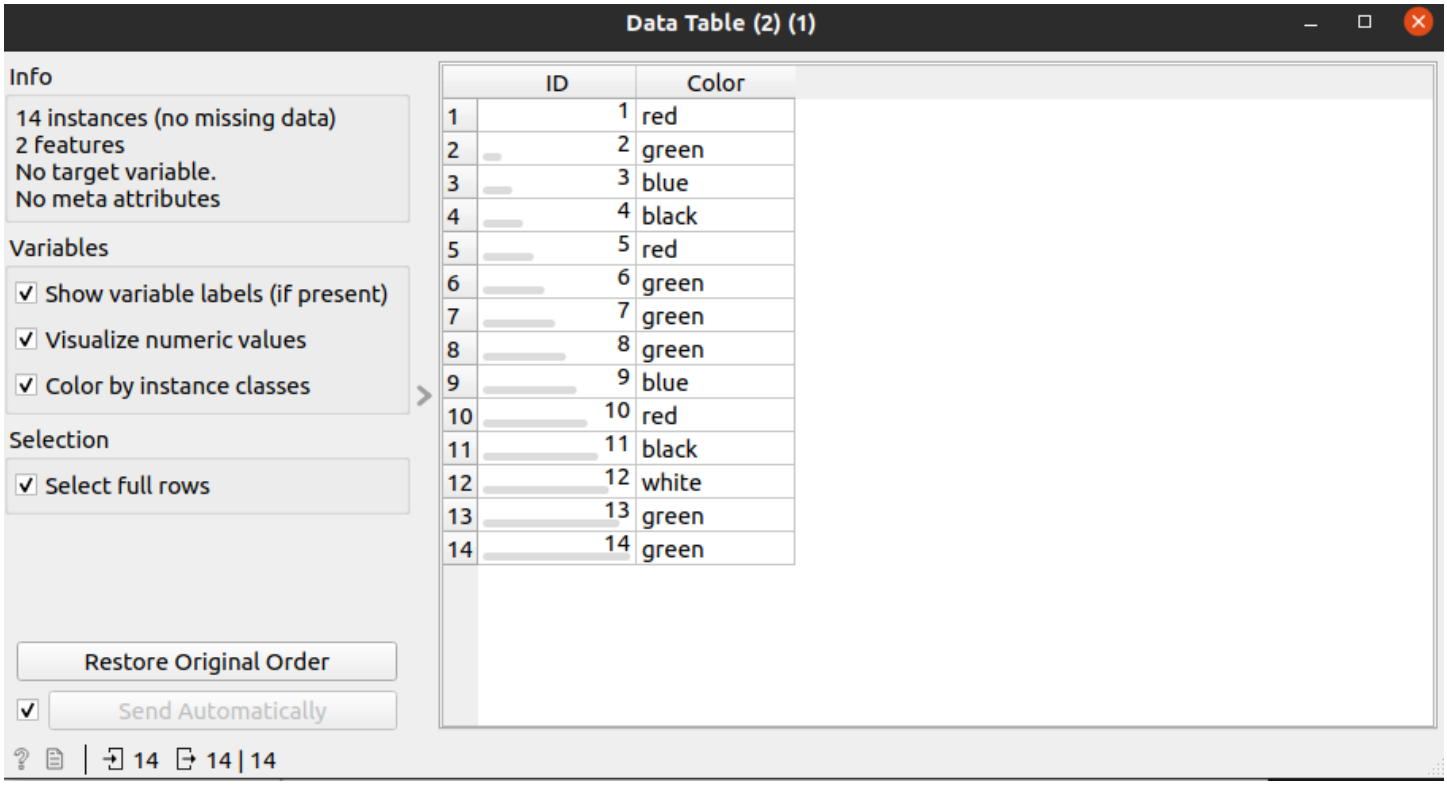
Analyse above techniques using Orange Data Mining framework.

Replacing using Mean Value



	ID	Age
1	1	13
2	2	15
3	3	16
4	4	16
5	5	19
6	6	29.96
7	7	20
8	8	21
9	9	22
10	10	22
11	11	25
12	12	25
13	13	25
14	14	25
15	15	30
16	16	33
17	17	33
18	18	35
19	19	35
20	20	35
21	21	35
22	22	36
23	23	29.96
24	24	45
25	25	46
26	26	52
27	27	70

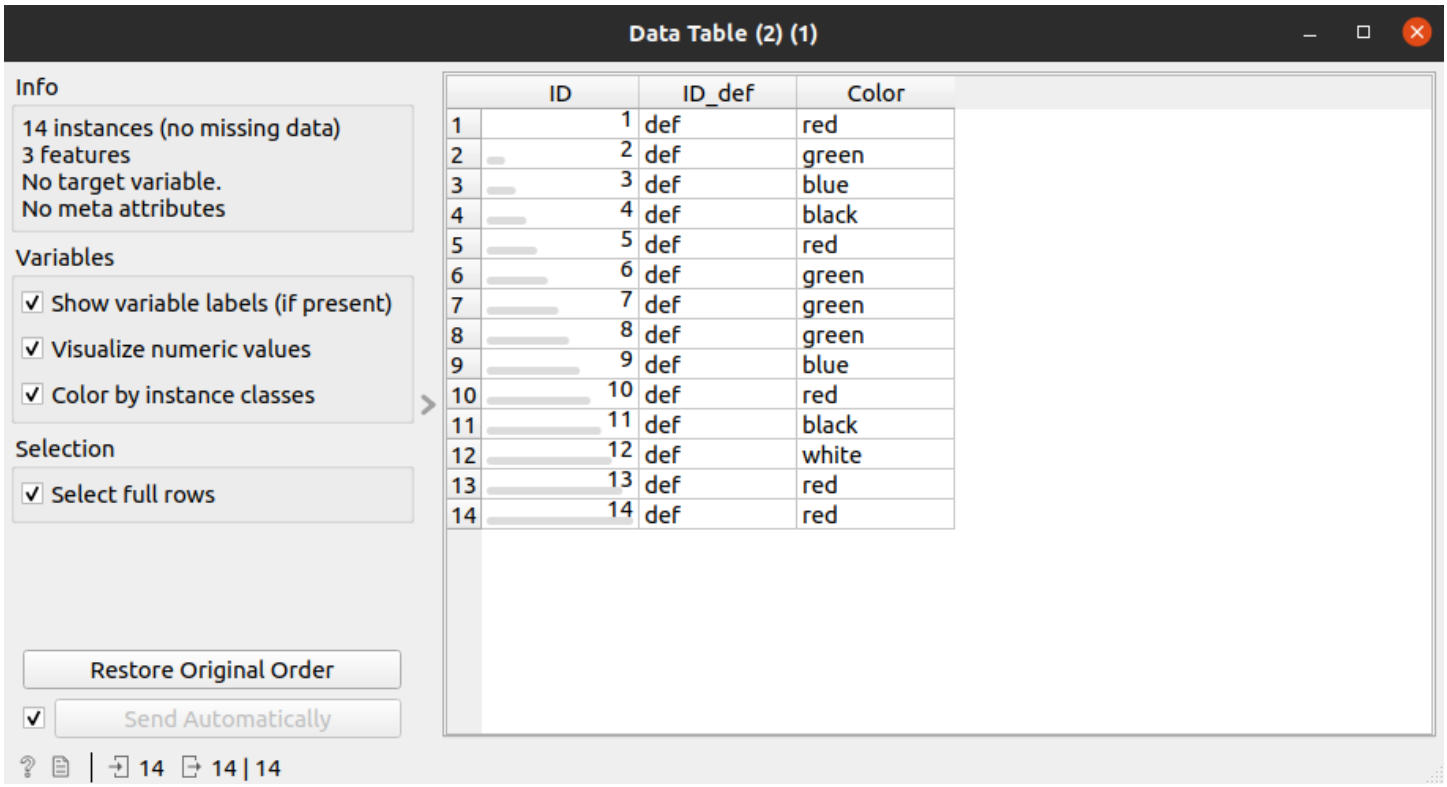
Replacing using Mode Value



	ID	Color
1	1	red
2	2	green
3	3	blue
4	4	black
5	5	red
6	6	green
7	7	green
8	8	green
9	9	blue
10	10	red
11	11	black
12	12	white
13	13	green
14	14	green

Note : Last 2 values were missing

Replacing using Global Constant



	ID	ID_def	Color
1	1	def	red
2	2	def	green
3	3	def	blue
4	4	def	black
5	5	def	red
6	6	def	green
7	7	def	green
8	8	def	green
9	9	def	blue
10	10	def	red
11	11	def	black
12	12	def	white
13	13	def	red
14	14	def	red

Note : Last 2 values were missing, Global Constant = Red