UNIVERSIDAD DE LIMA
ESCUELA UNIVERSITARIA DE INGENIERÍA
CARRERA DE INGENIERÍA DE SISTEMAS

Taller 4: Exploración y Limpieza de Datos

Sección 701

Curso: SISTEMAS DE INTELIGENCIA EMPRESARIAL

Profesor: LUIS ARMANDO RAYGADA VARGAS

Integrantes:

AGÜERO ESTRELLA AARON FERNANDO	**20131537**
ESPINOZA GRADOS ROSSY	**20131797**
MANDUJANO NIMA RICARDO JESUS	**20100661**
PALOMARES HUERTA IVAN ENRIQUE	**20133037**
ROJAS HINOSTROZA JASON RAFAEL	**20133136**
UGAZ BURGA CARLOS GABRIEL	**20131349**

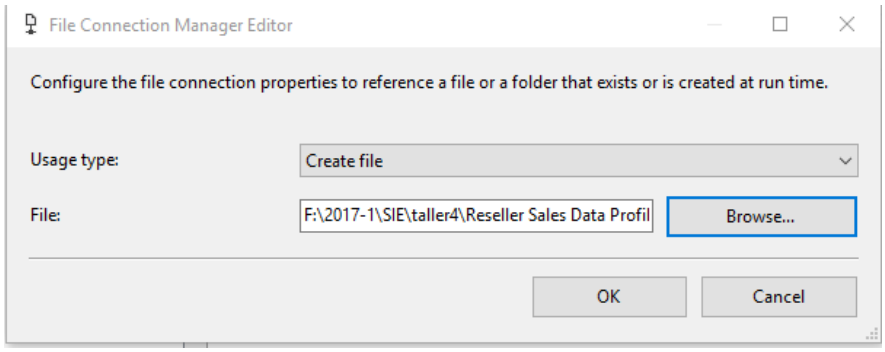
Lima – Perú

Junio 2017

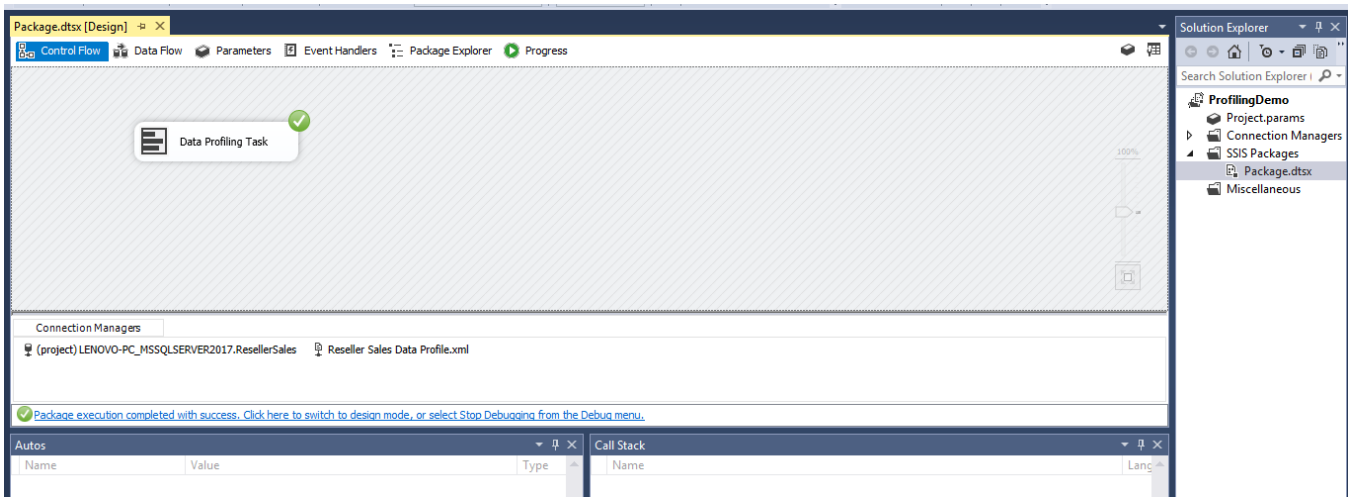
PARTE SQL SERVER 2014:

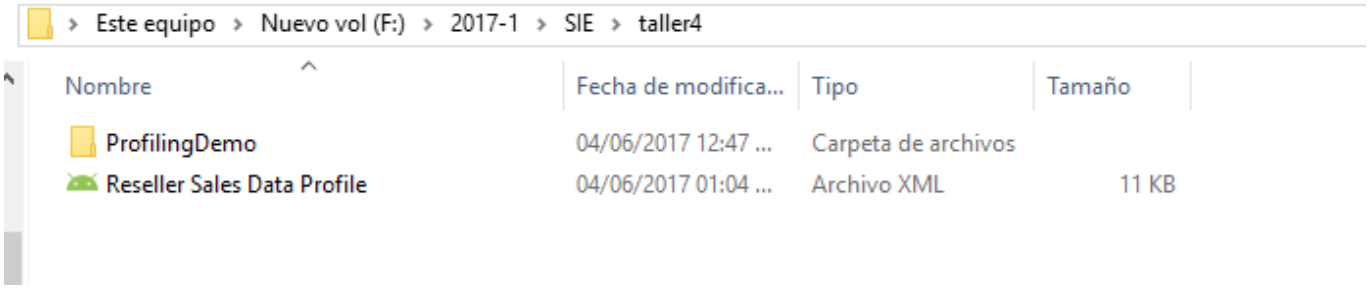
Ver los resultados de los perfiles **Value Inclusion Profile Request y Column Null Ratio Profile Request,** determinar y escribir que información muestran estos perfiles. En base a la información de la calidad de estos datos ¿Que decisiones tomaría? (se puede ayudar de consultas SQL para tomar más información).

Creación del archivo Reseller Sales Data Profile.xml:

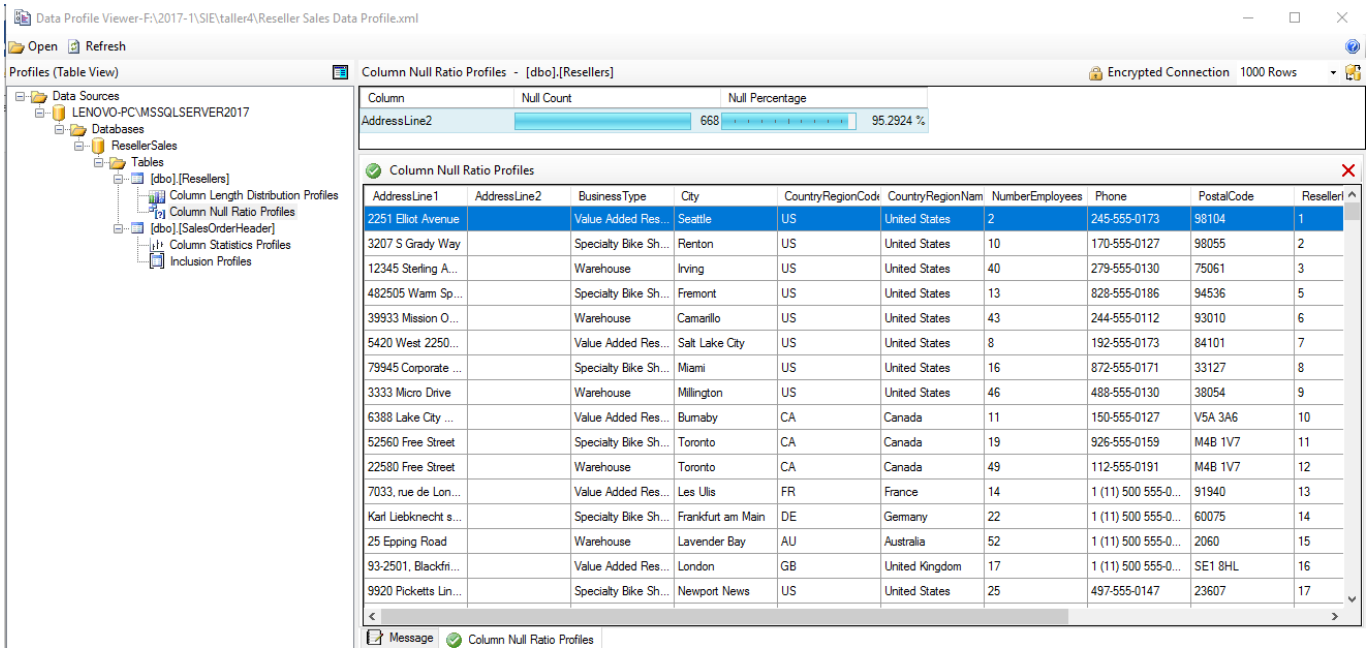


Debugging ejecutado correctamente: Data Profiling Task



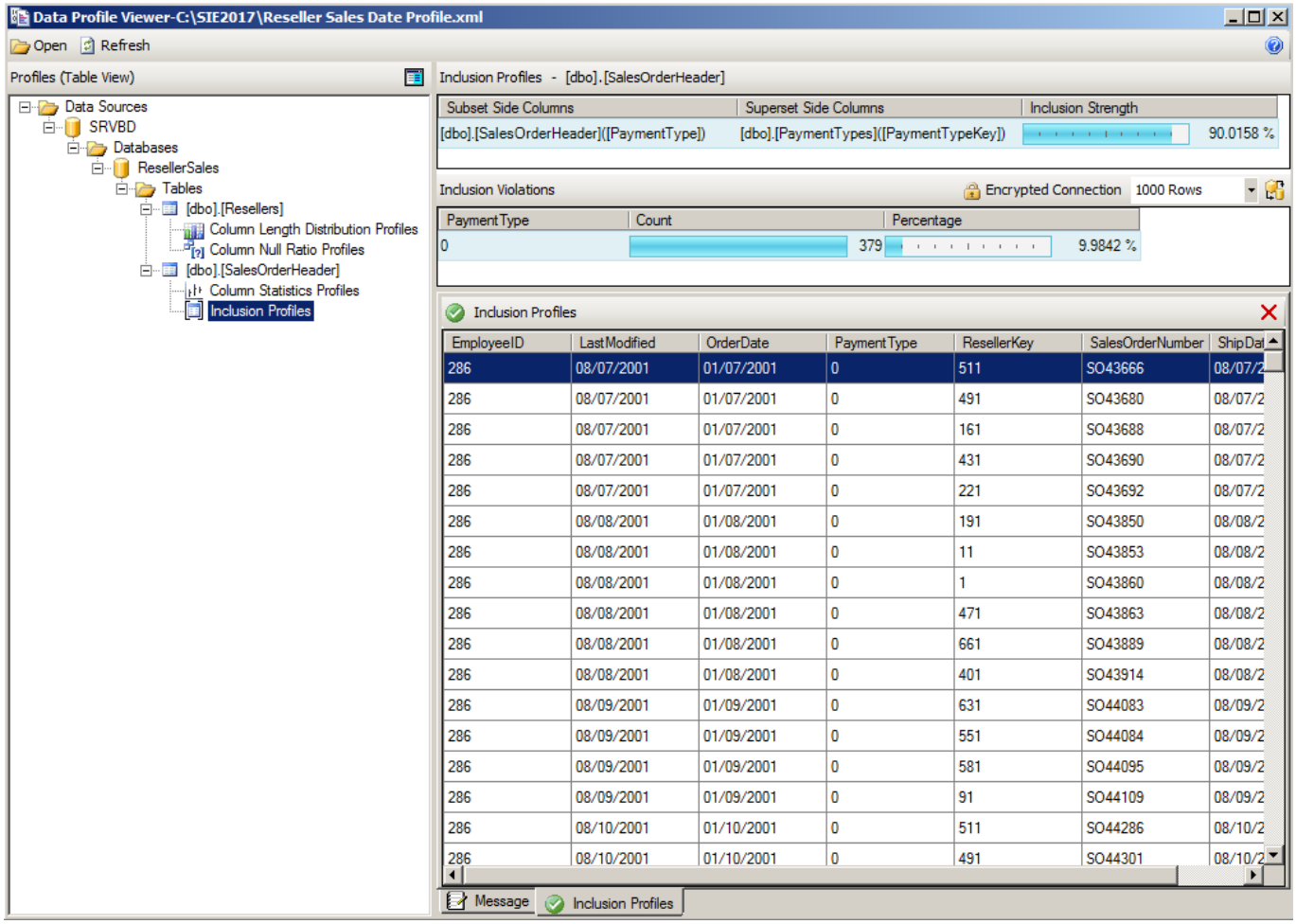


Análisis de resultados:



En esta parte se visualiza que el 95.29% de la información de la columna AddressLine2 en la tabla Resellers es nula.

Para solucionar este inconveniente, se haría una transformación de columna para indicar “No tiene” en caso de que el registro tenga valor nulo en el campo AddressLine2. Por ello, se aplicaría Derived Column con la función ReplaceNull para el campo AddressLine2.

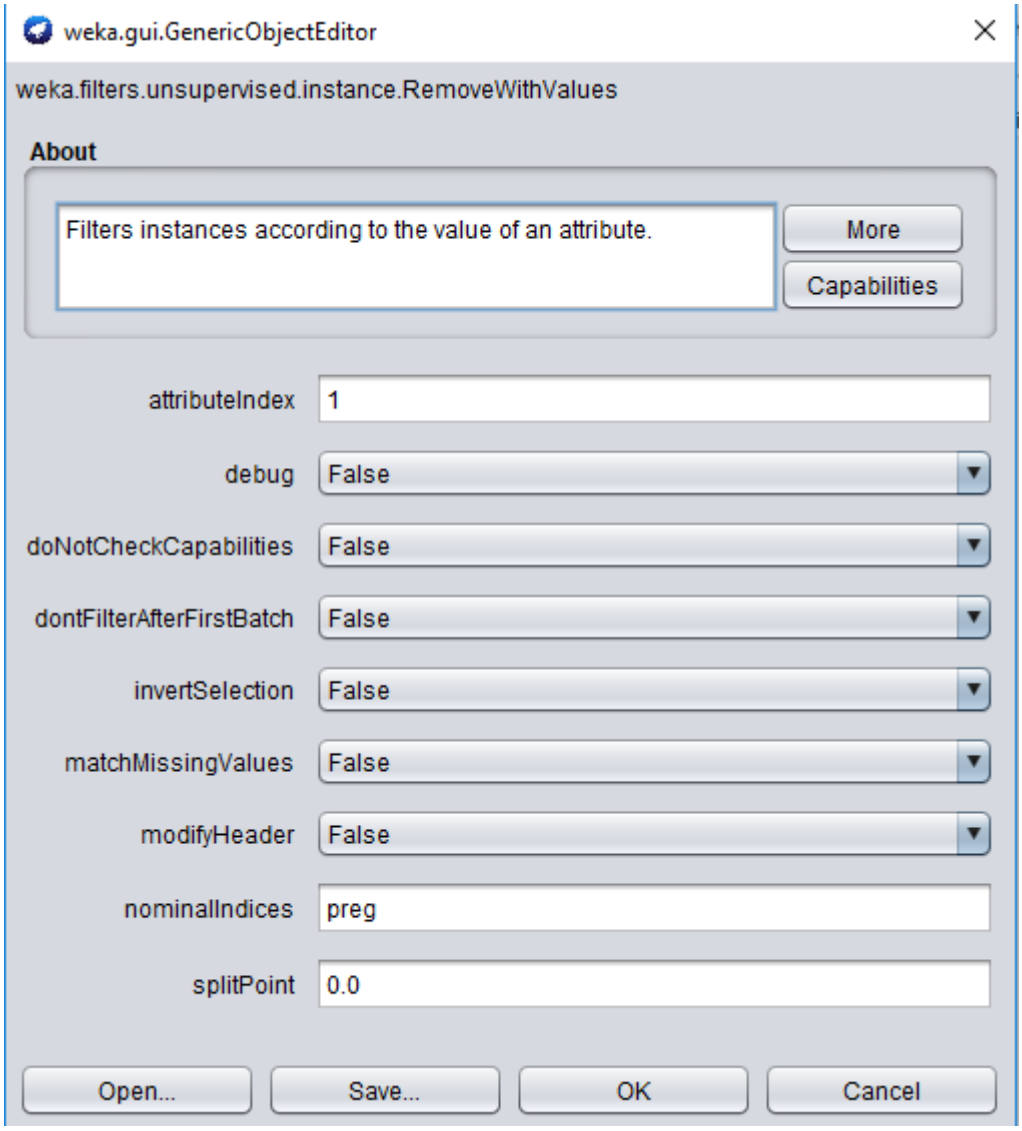


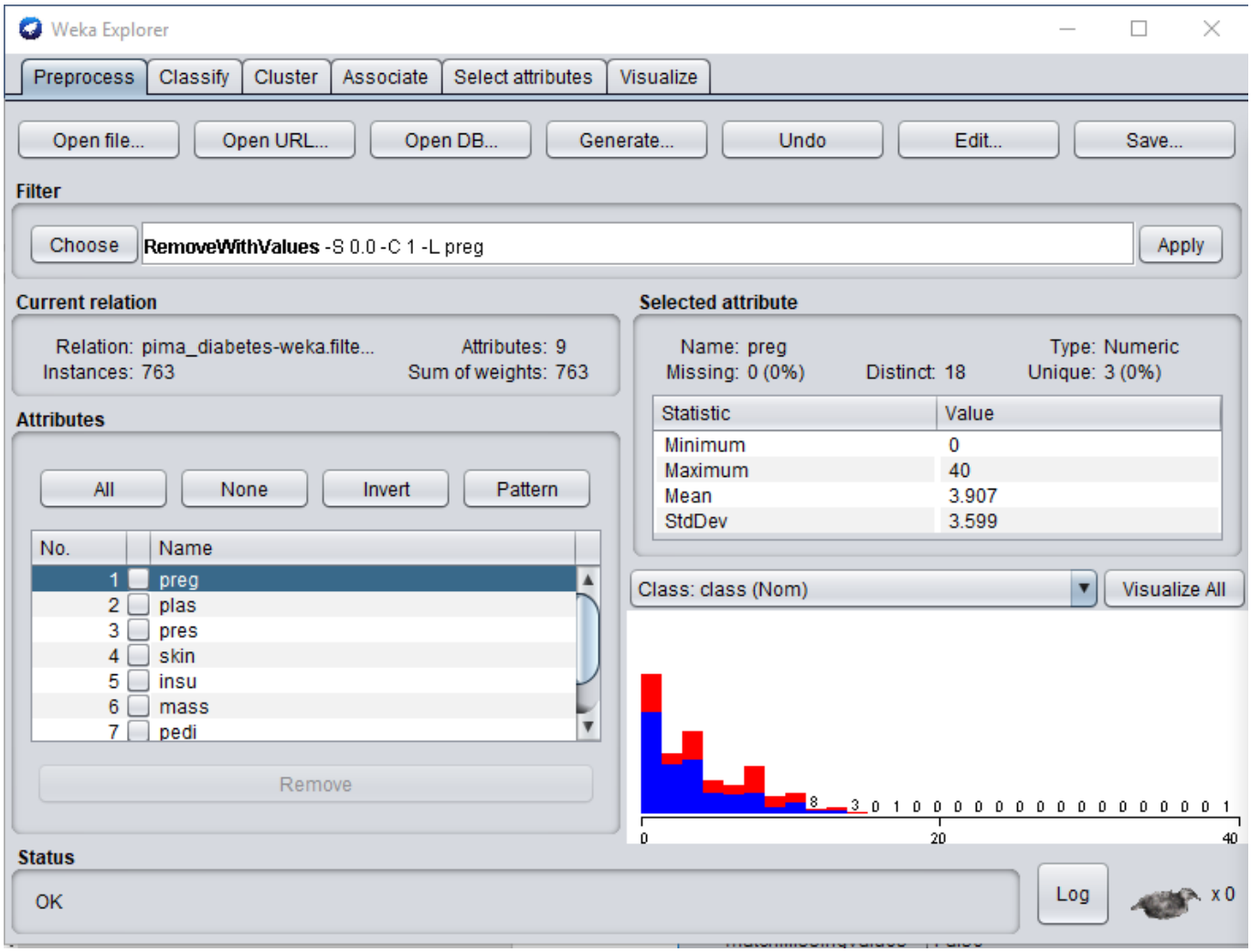
En esta imagen se visualiza que el 90.01% de los datos de la columna PaymentType de la tabla SalesOrderHeader hace referencia a la columna PaymentTypeKey de la tabla PaymentTypes. Sin embargo, el 9.98% restante de la columna PaymentType no hace referencia a la columna de la tabla PaymentTypes. Para ello se realizaría una función Lookup con una consulta de inner joins que obtenga solamente los registros que hagan referencia a la tabla de PaymentTypes.

PARTE WEKA – DIABETES:

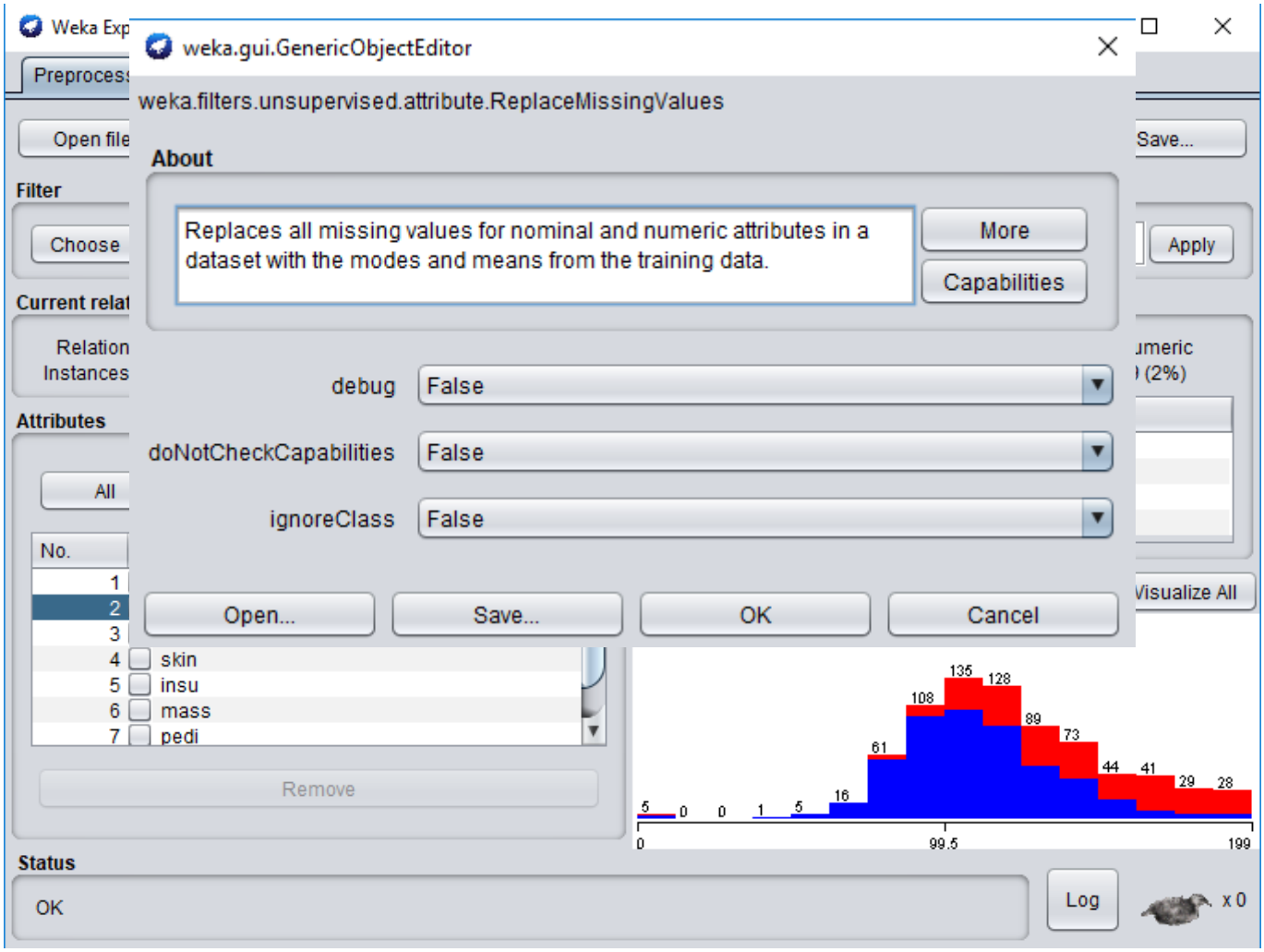
Primero, se separó en 6 rangos los valores del atributo edad.

1. Se eliminó los registros que contenían el valor -1 como número de embarazos.




Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Open file...
Open URL...
Open DB...
Generate...
Undo
Edit...
Save...
Filter
Choose
RemoveWithValues -S 0.0 -C 1 -L preg
Apply
Current relation
Relation: pima_diabetes-weka.filte...
Attributes: 9
Instances: 763
Sum of weights: 763
Selected attribute
Name: preg
Type: Numeric
Missing: 0 (0%)
Distinct: 18
Unique: 3 (0%)
Statistic
Value
Minimum
0
Maximum
40
Mean
3.907
StdDev
3.599
Attributes
All
None
Invert
Pattern
No.
Name
1 preg
2 plas
3 pres
4 skin
5 insu
6 mass
7 pedi
Remove
Class: class (Nom)
Visualize All
8 3 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
0
20
40
Status
OK
Log
x 0

2. Se reemplazó los valores vacíos por la moda y la media pues el campo plasma es un tipo numérico. Además, al utilizar estas medidas de tendencia central, la alteración de la variabilidad es mínima.



8. Al tener los 6 rangos de edad, se reemplazaron los rótulos que indicaban valores infinitos. Los cambios fueron los siguientes: ‘(0-31]’ en vez de ‘(-inf-31]’ y ‘(71-130]’ en reemplazo de ‘(71-inf)’.