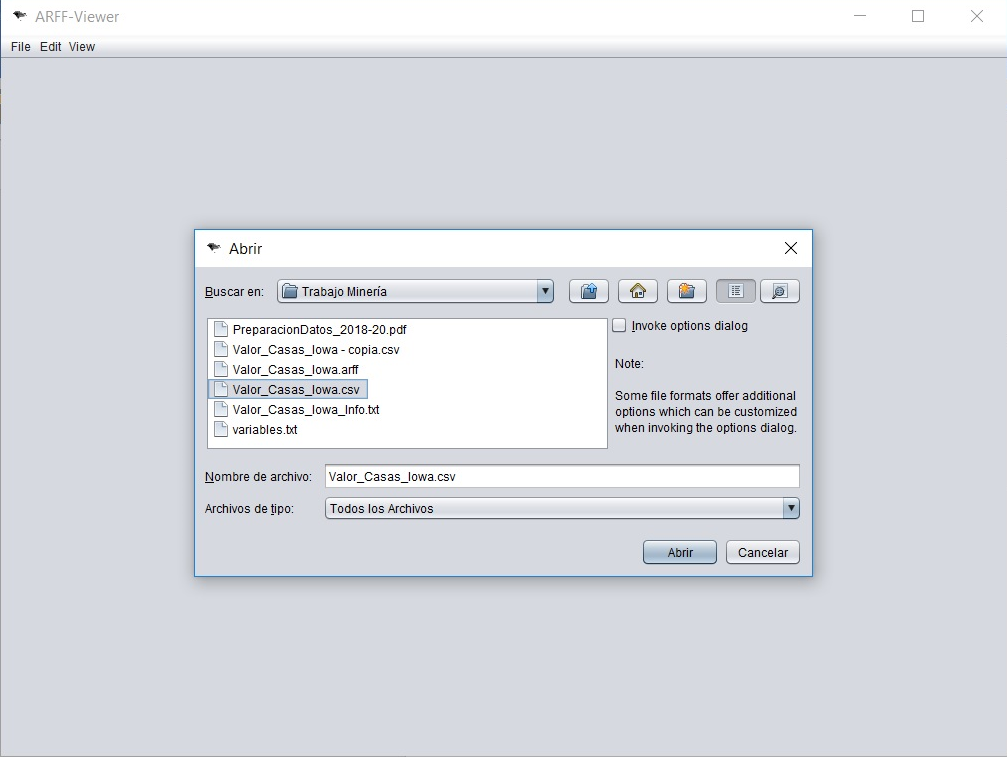
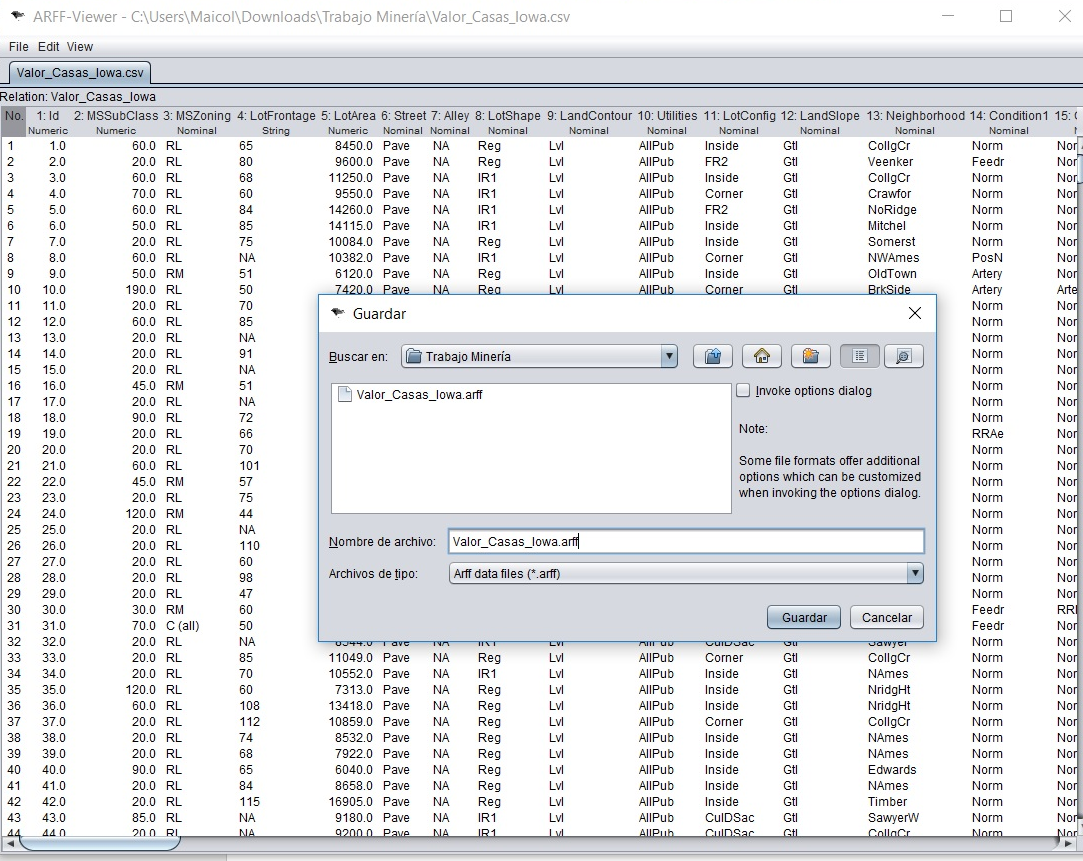
**PREPARACIÓN DE DATOS EN WEKA**

**Dataset:** Iowa Housing

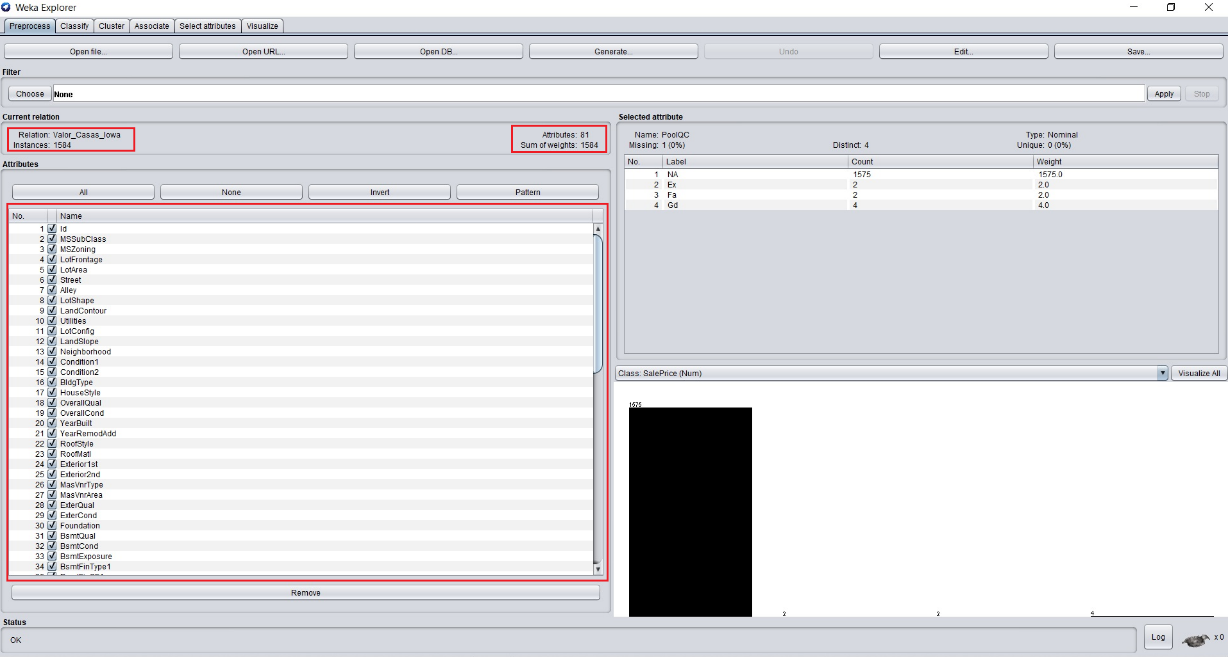
1. Inicialmente, se cargó el archivo .csv en el visualizador de archivos arff (tools/arffviewer) en WEKA, ya que se tuvo problemas para abrirlo directamente en el explorador:



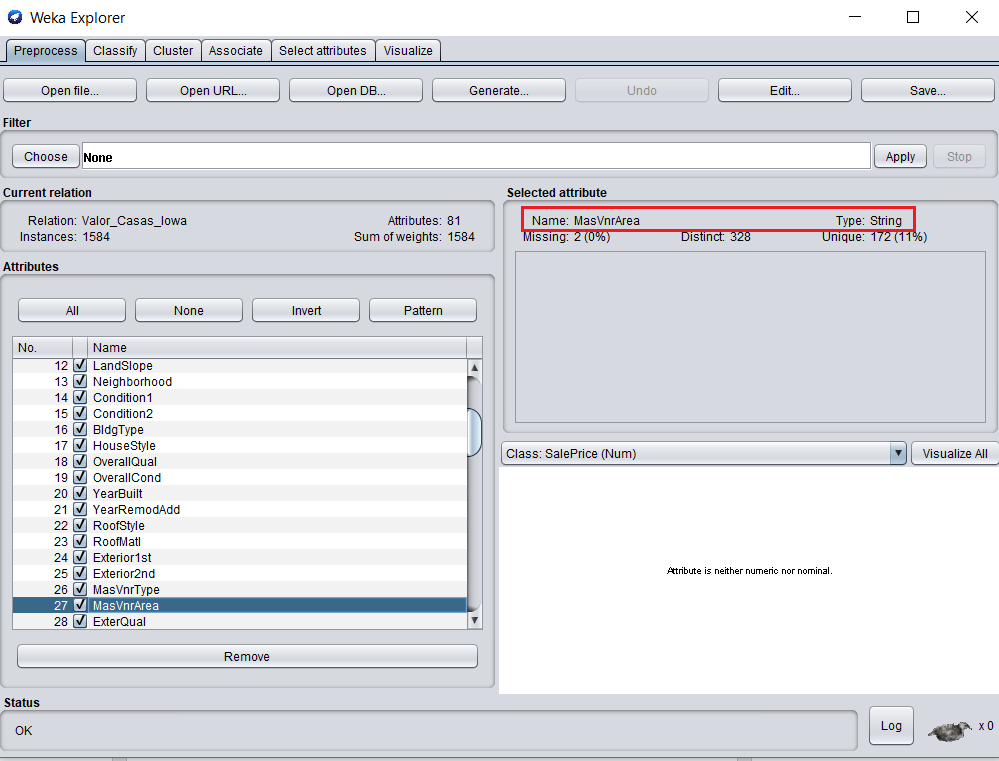
Una vez cargado el archivo, se exporta el mismo como archivo con extensión .arff para poder ser abierto directamente con el explorador de WEKA:



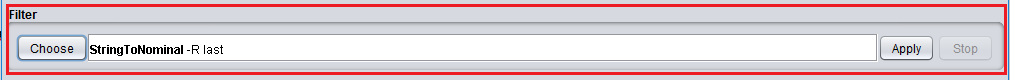
1. Se abre el archivo .arff en el explorador de WEKA y se observan las características principales del dataset:

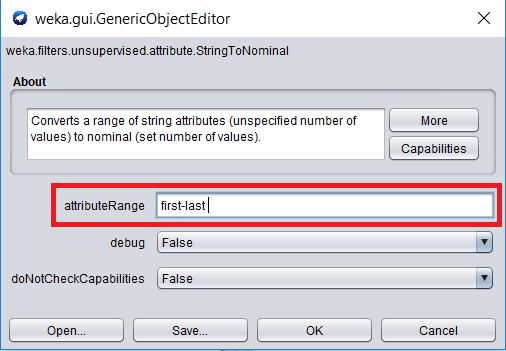


1. Al hacer un análisis de los atributos, se encuentra que en su mayoría son de naturaleza numérica y nominal, sin embargo, hay algunos atributos que son catalogados como string.

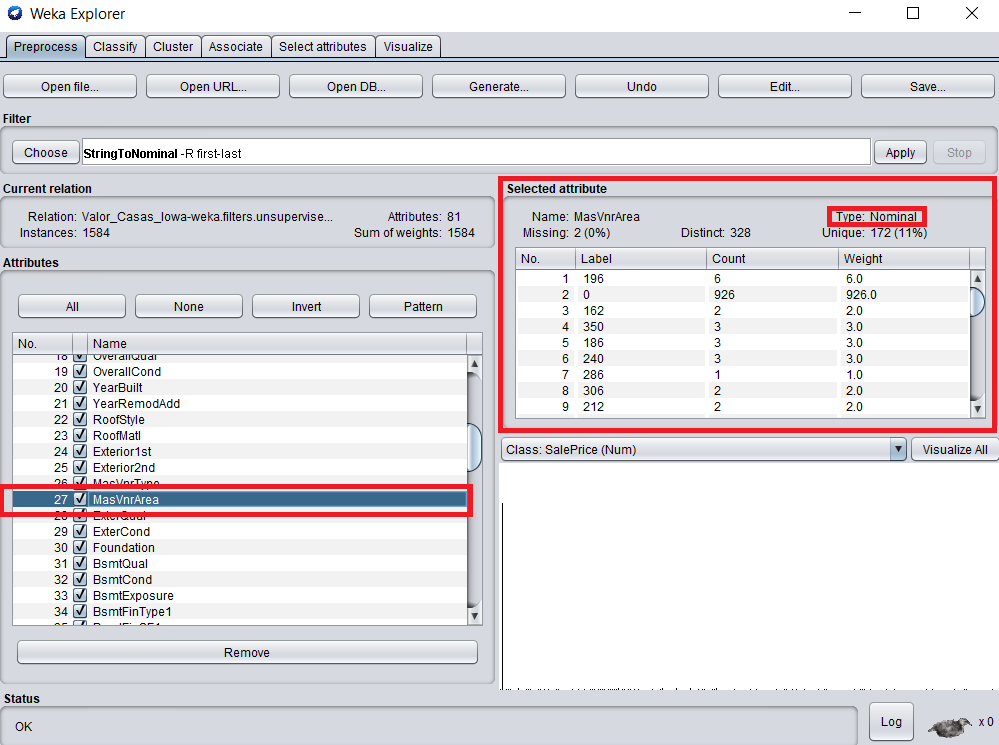


Se aplica el filtro correspondiente para transformar estos atributos en tipo nominal (categórico): filters/unsupervised/attribute/stringtonominal

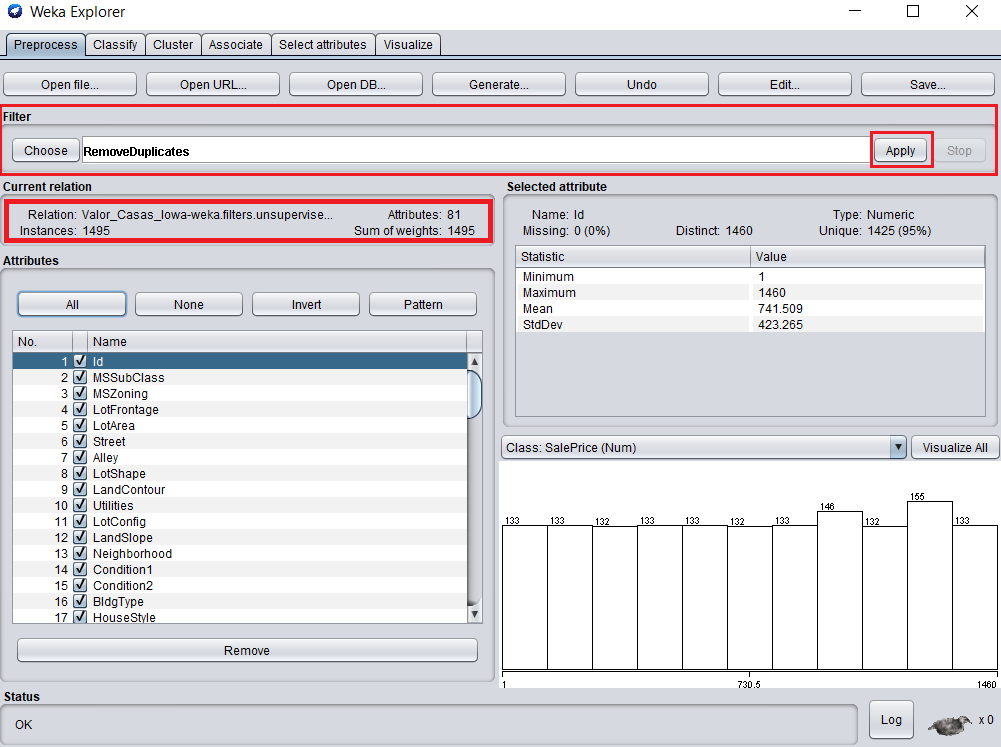




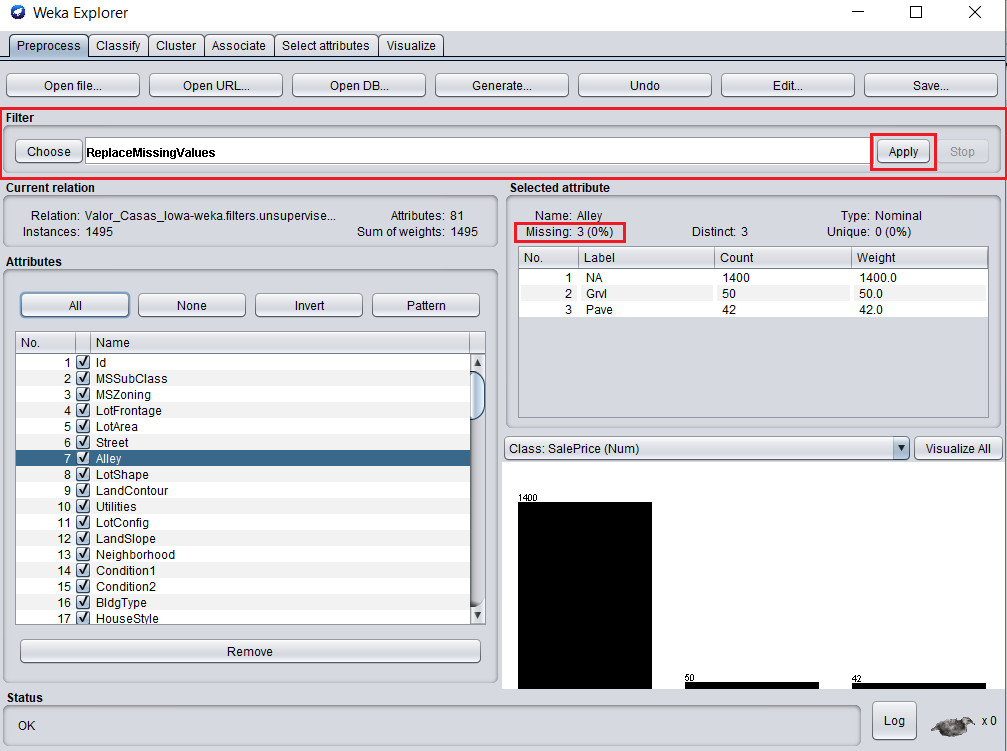
Se observa que el filtro fue aplicado exitosamente:



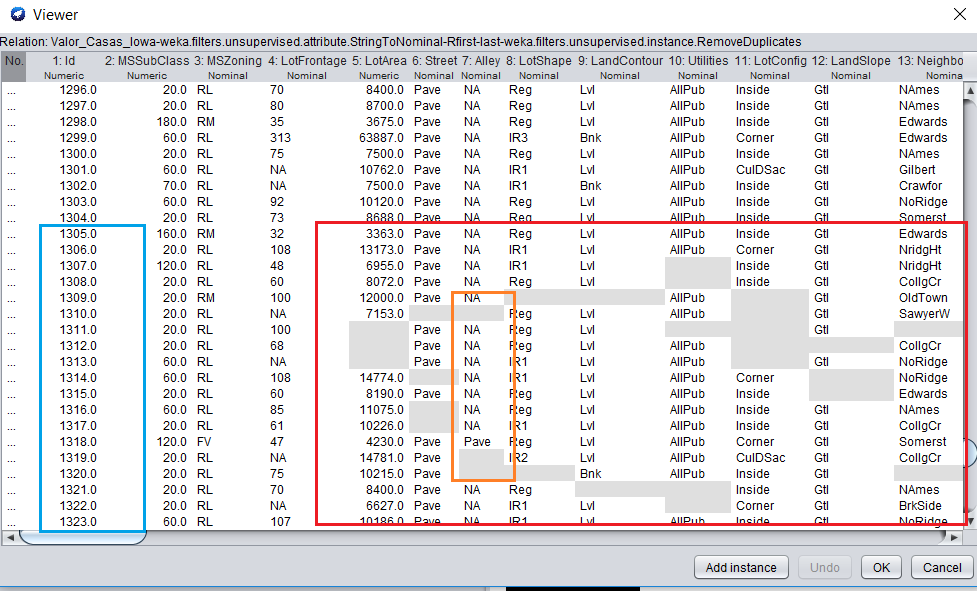
1. Luego de esto se procede a remover las instancias duplicadas. Para esto se accede al filtro /filters/unsupervised/instance/Removeduplicates. Se observa que la cantidad de datos se reduce



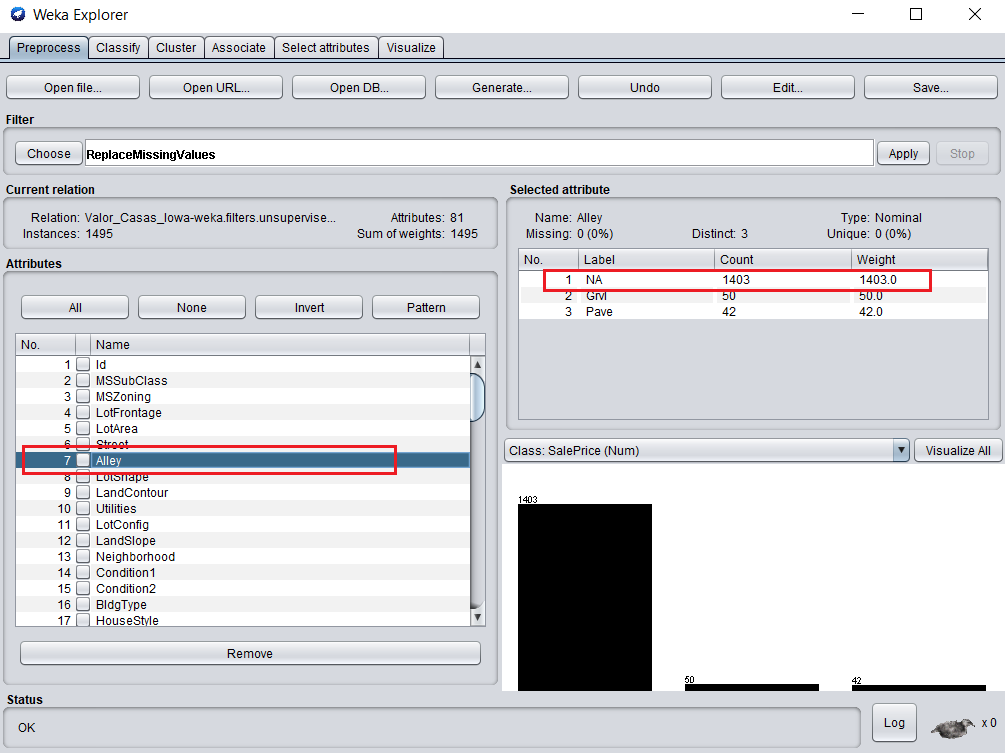
1. Se procede a aplicar el filtro para reemplazar valores faltantes: filters/unsupervised/instance/ReplaceMissingValues. Este filtro reemplaza los valores faltantes con el valor más común de cada atributo (nominal) o con el promedio (numeric). Para comprobar, tomamos como referencia el atributo Alley, que tiene 3 valores faltantes.



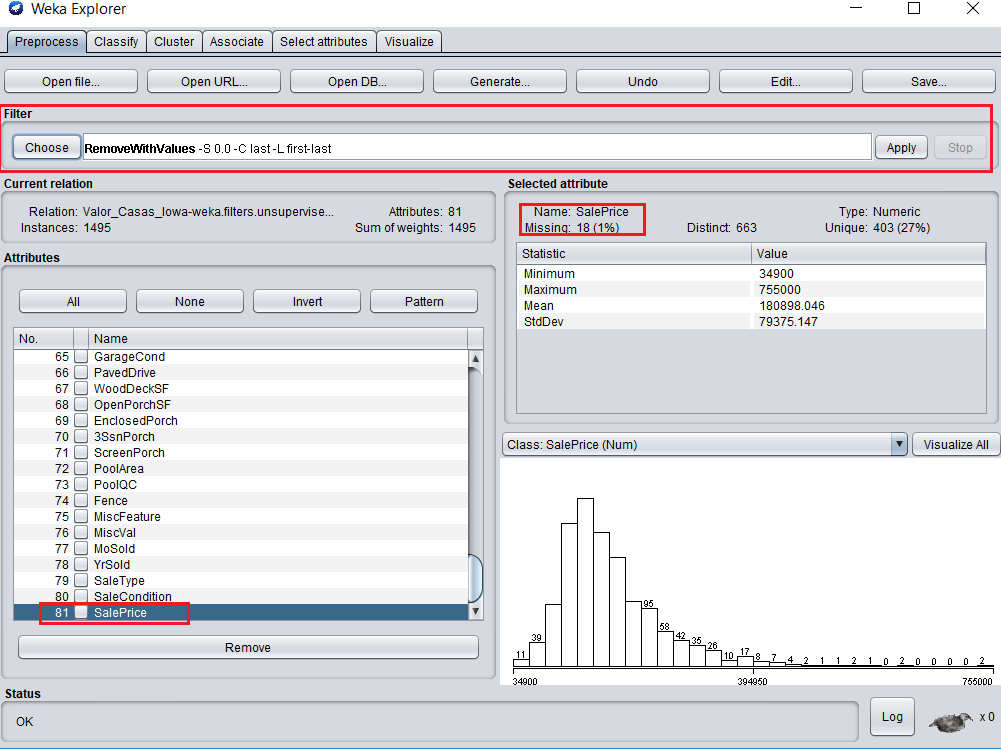
En el editor, se observan las instancias en las cuales se encuentran los datos faltantes (notar que en esta zona hay una buena cantidad de datos faltantes en otros atributos)

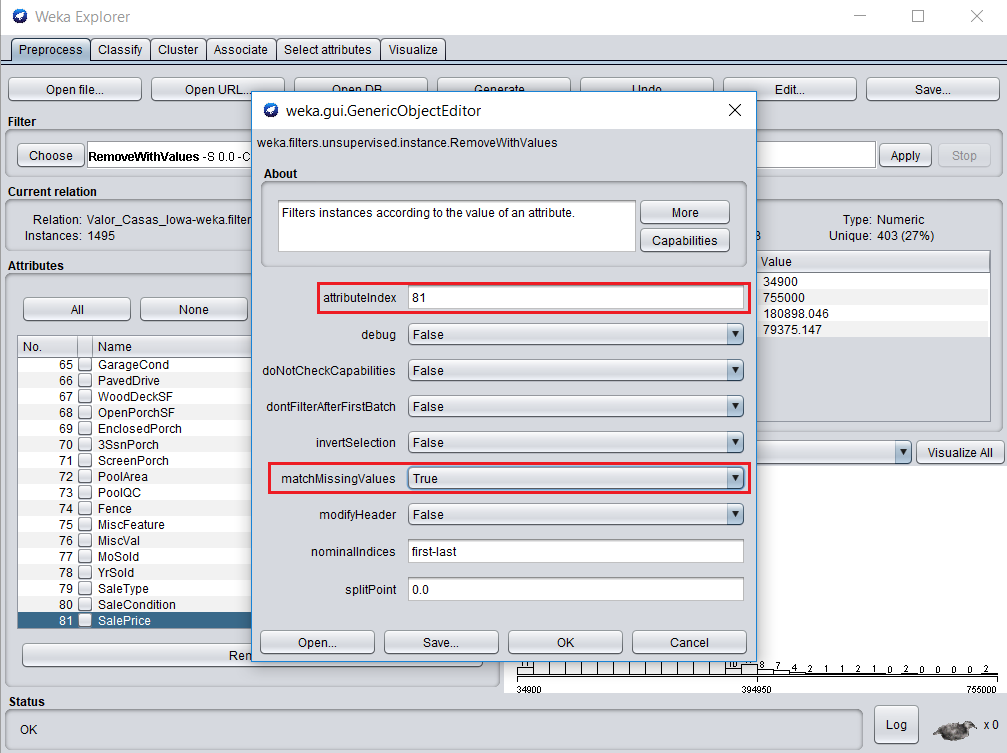
-

Se aplica el filtro y se comprueban los resultados. La variable Alley, ahora no tiene valores faltantes y estos fueron reemplazados por NA (el label más común). Un proceso similar ocurrió con los demás atributos que contenían valores faltantes.

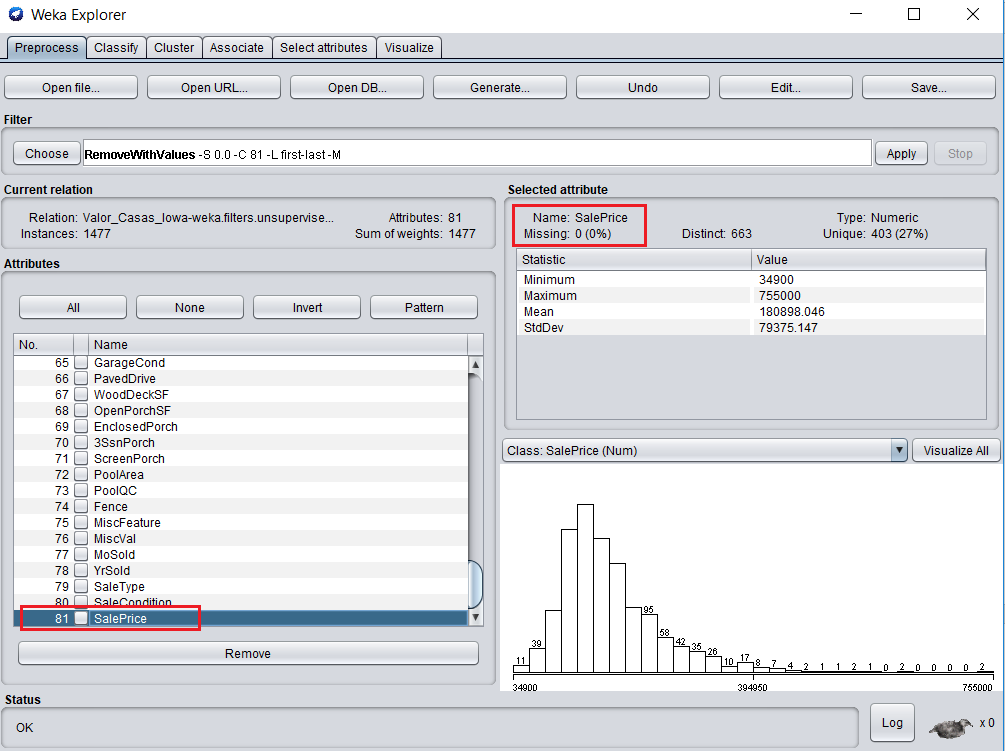


1. Se observa que el atributo SalePrice aún contiene valores faltantes, por tanto, se procede a aplicar el filtro para remover valores faltantes: filters/unsupervised/instance/Removewithvlues y se configura de la siguiente manera



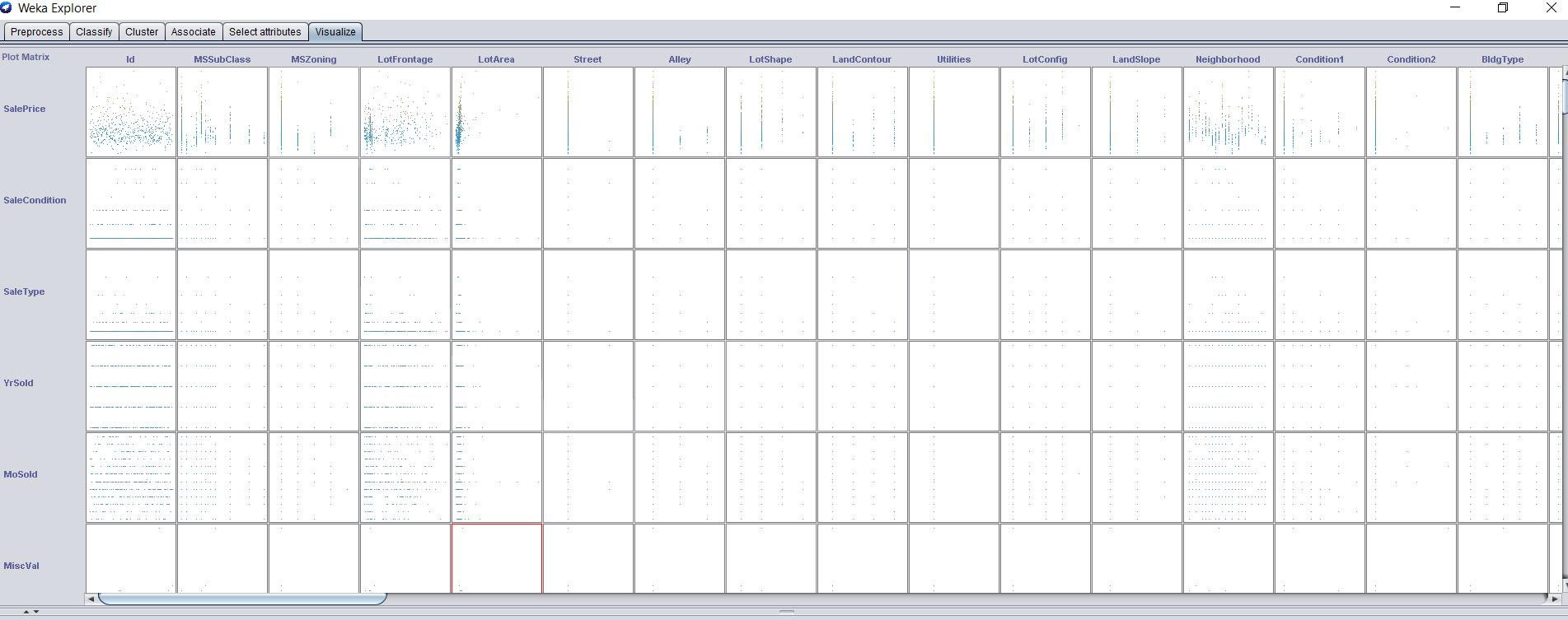


Una vez aplicado el filtro, se observa que el atributo SalePrice ya no tiene datos faltantes:



De esta manera, el dataset ya no contiene registros duplicados ni valores faltantes.

1. En la pestaña Visualize, se pueden ver gráficas de correlación de todas las matrices.



También es posible observar histogramas para todos los atributos:



1. CONCLUSIONES

* El entorno gráfico de WEKA es amigable y permite acceder rápidamente a información valiosa acerca de loas diferentes instancias o atributos
* Algunos filtros de la herramienta son limitados en la medida en que no permiten aplicarse a rangos específicos de variables o atributos
* Como no es un lenguaje de programación, no es tan versátil como Python y esto es una desventaja para realizar procesos de preparación. Por ejemplo para remover datos duplicados, resultó ser más eficiente Python en la medida en que el usuario es quien controla el funcionamiento casi que a su antojo.