February 9, 2018

**Homework 1**

Sanchit Singhal

CE 395R – Data Mining with Carlos Caldas

Spring 2018

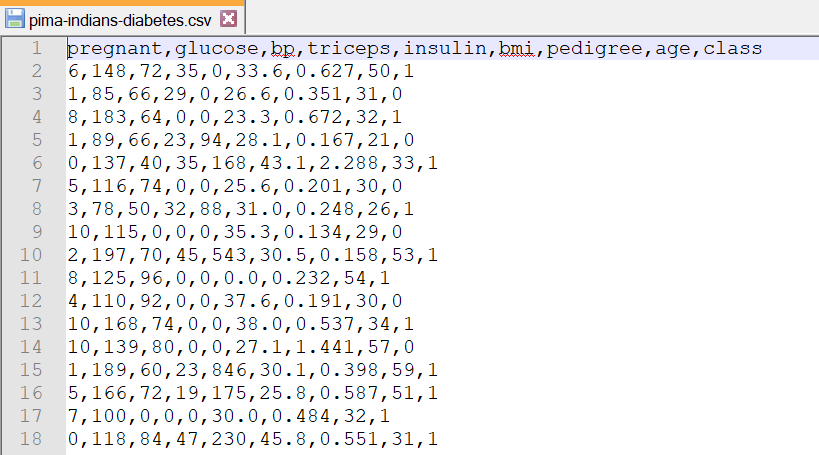
University of Texas at Austin

**Pima Indians Diabetes Database**

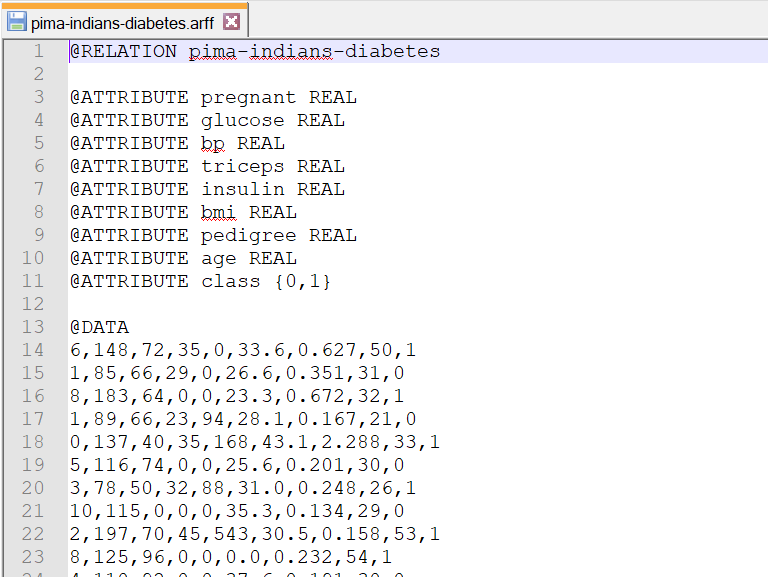
1. **Data Selected**

* Number of Attributes = 8 + class (all numeric)
  + Number of times pregnant
  + Plasma glucose centration
  + Diastolic blood pressure (mm Hg)
  + Triceps skin fold thickness (mm)
  + Serum insulin (mu U/ml)
  + Body Mass Index (kg/m^2)
  + Diabetes pedigree
  + Age (years)
  + Class variable (0 or 1) – Diabetes
* Number of Instances = 768

*I downloaded this dataset in .data and .name file types. I first converted this information into .csv:*

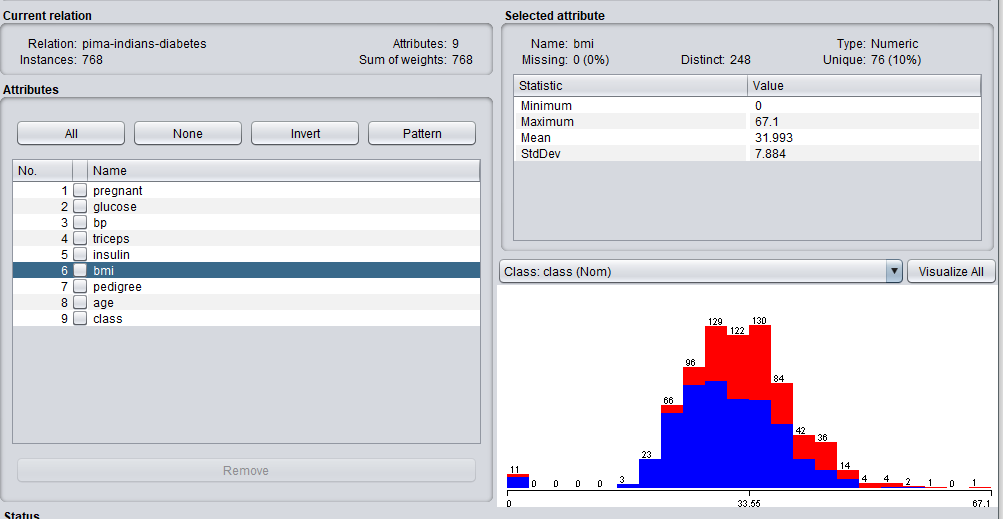


*From my csv file I converted the file into .arff for ease of use with Weka 3.8:*



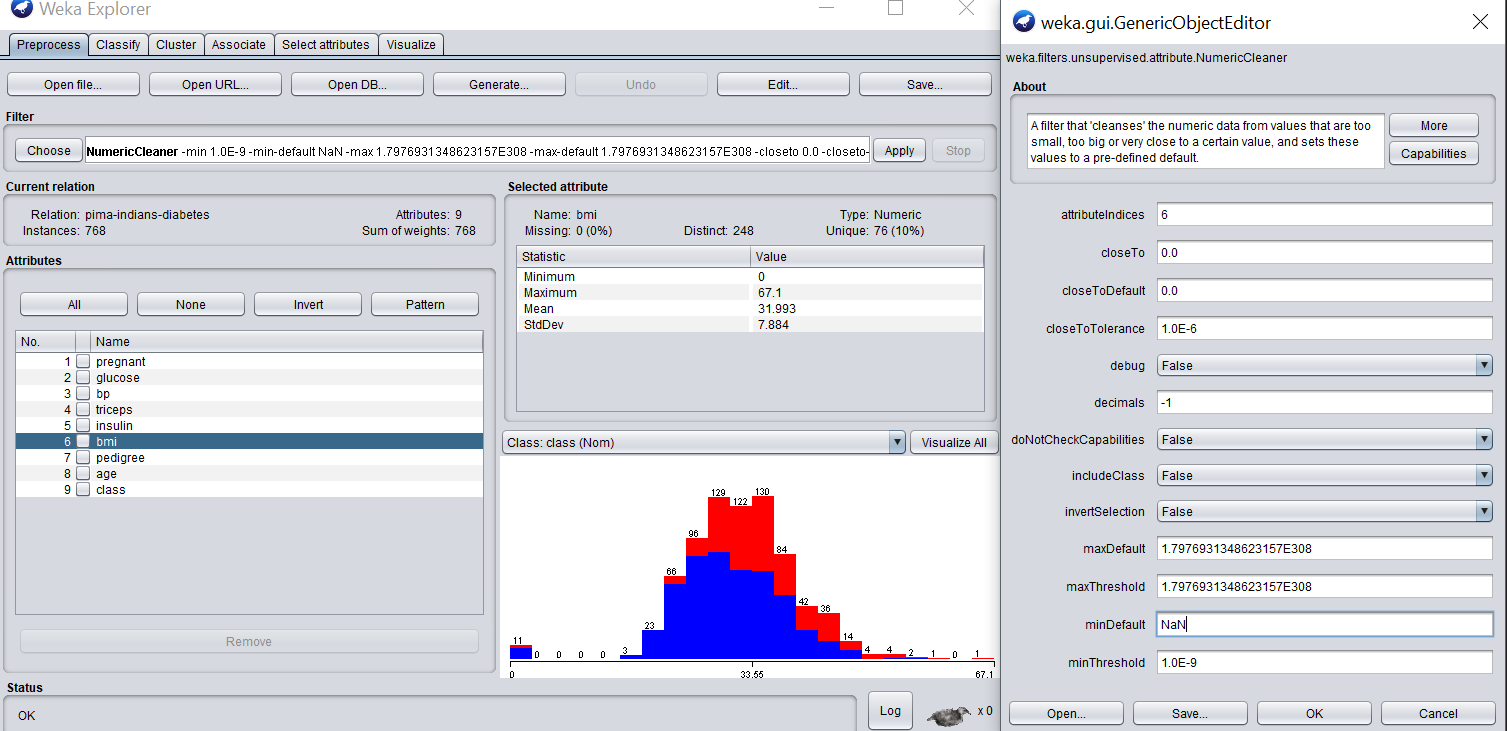
1. **Data Quality Problem Identified**

*After reviewing the data in the Explorer, I realized that the BMI attribute had a minimum of 0 and there were 11 such instances. This makes no sense as nobody can have zero mass. This tells me that this is actually just missing data:*

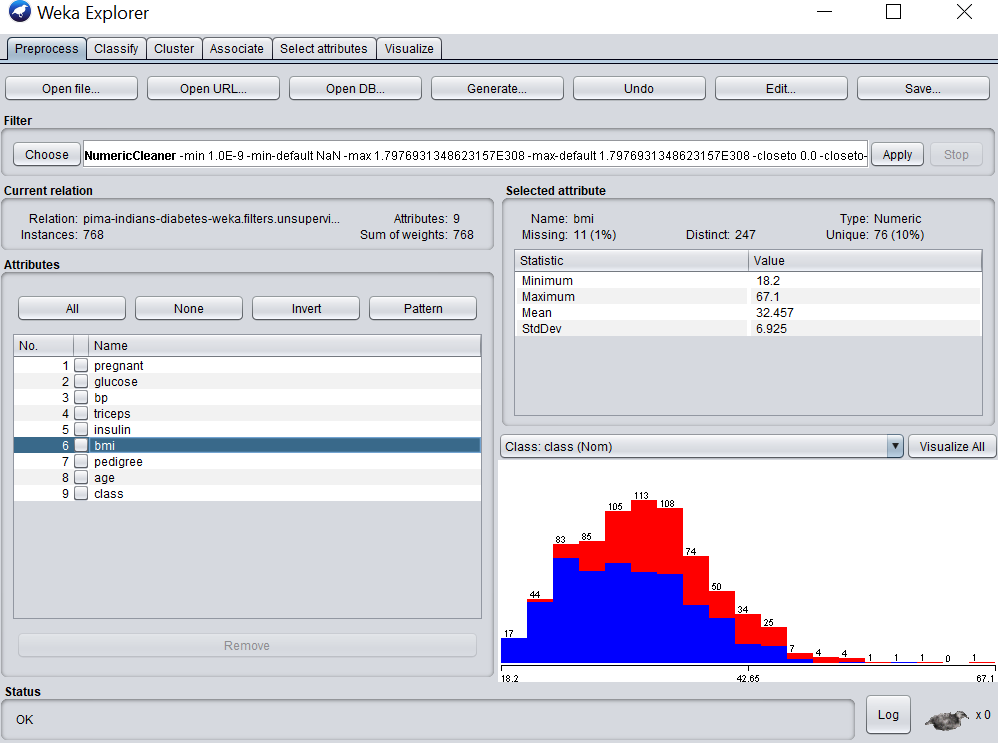


1. **Proposed Solution**

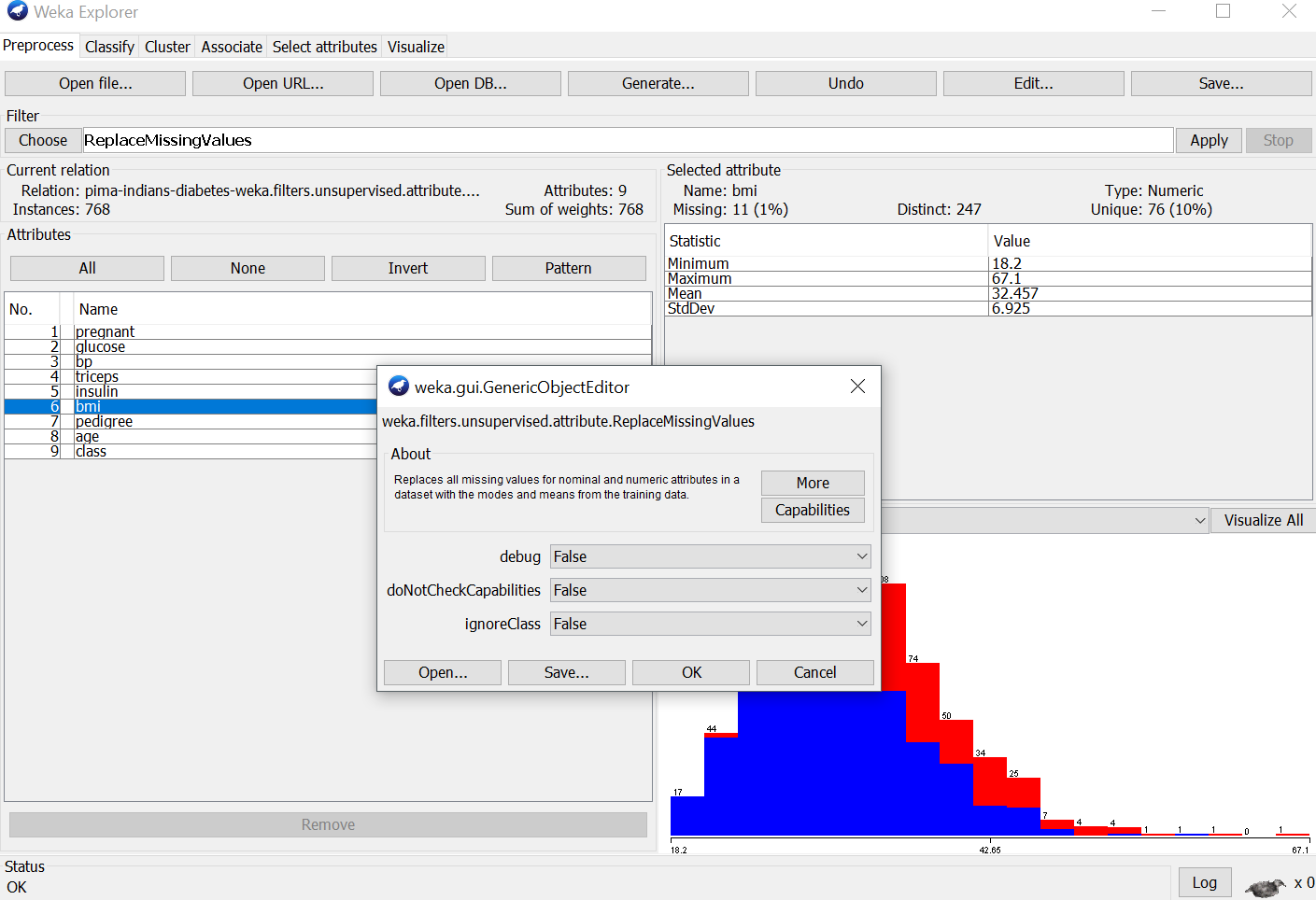
*The first step to clean this data would be to identify these instances as having missing values. I did this using the unsupervised.attribute.NumericalCleaner filter. I set the attributeIndicies to 6 to specify the bmi attribute and changed the minThreshold to slightly above 0. The minDefault was changed to NaN as that is unknown:*



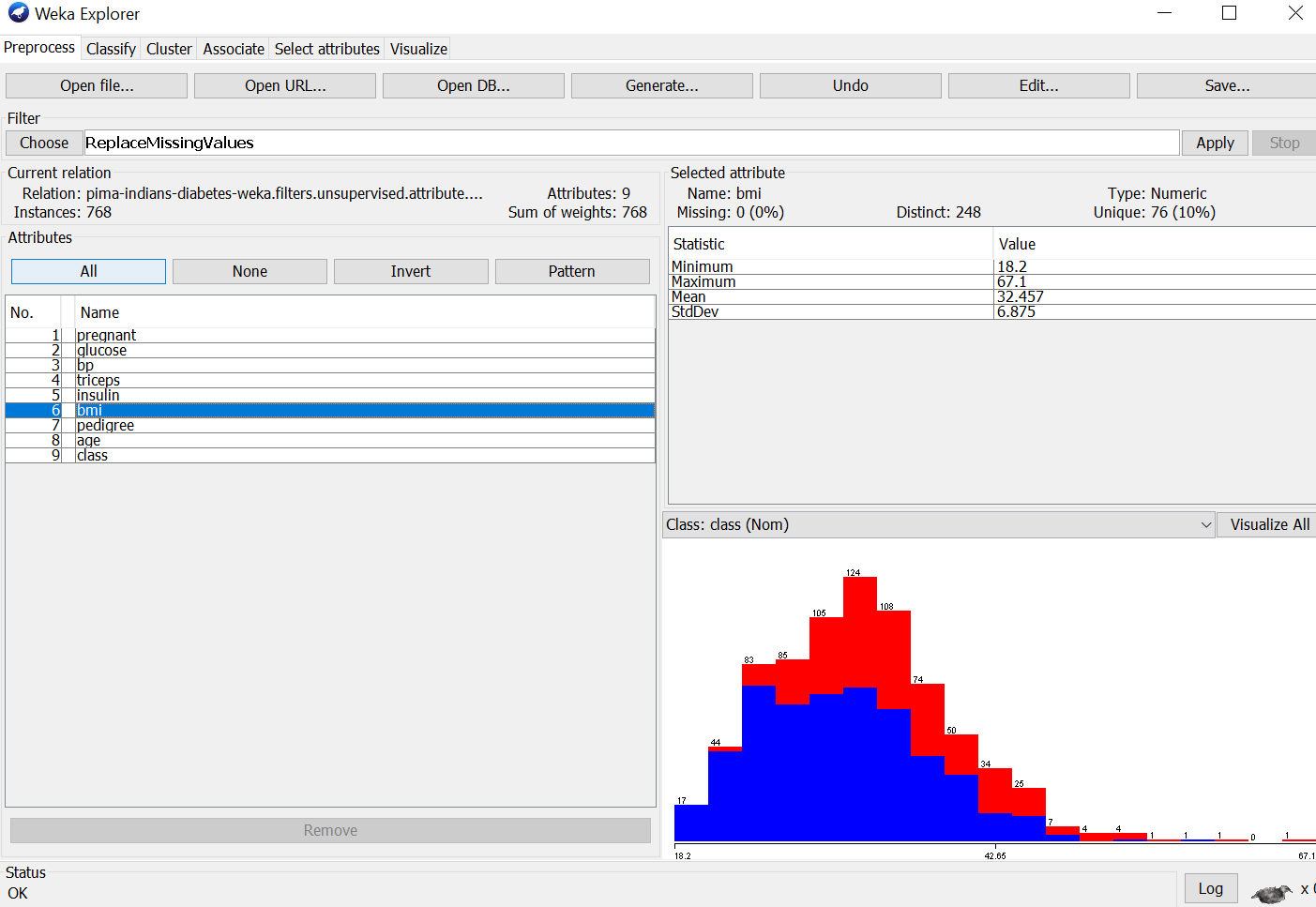
*After applying the filter, I was able to identify 11 attribute values as missing which were formally set to 0 :*



*Now that I have the missing data identified, I must decide on a strategy to handle them. Two obvious methods would be to either remove them or impute them with the mean. Because we are dealing with population data, I think it would make sense to replace the missing values with the mean of the bmi attribute. This will ensure the results will not be affected much by the artificial values we impute. I did this using the unsupervised.attribute.ReplaceMissingValues:*



*After applying the filter, I was able impute the mean for the missing values for bmi attribute:*



*I exported this cleaned dataset into a new CSV.*

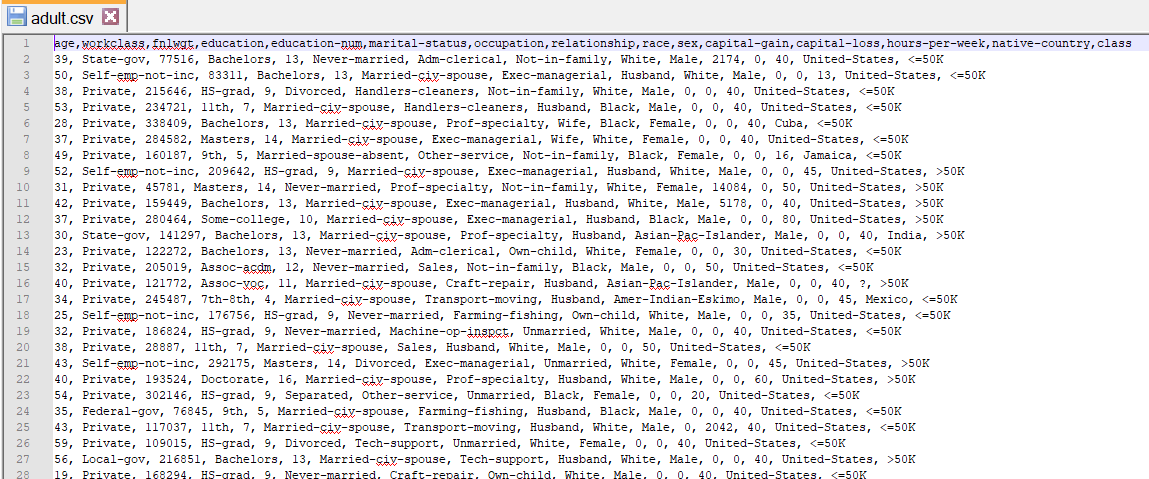
**Adult Income Database**

1. **Data Selected**

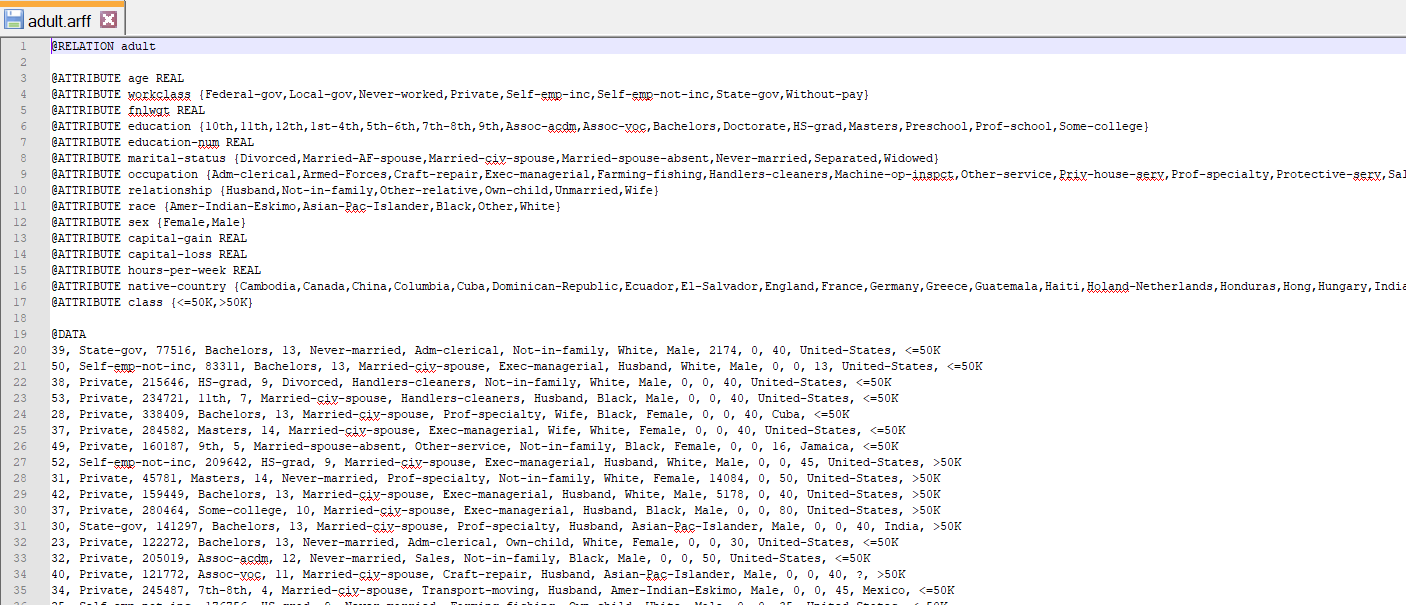
* Number of Attributes = 14 + class
  + Age (years)
  + Workclass (categorical)
  + Final Weight (continuous)
  + Education (categorical)
  + Education-num (continuous)
  + Martial-status (categorical)
  + Occupation (categorical)
  + Relationship (categorical)
  + Race (categorical)
  + Sex (categorical)
  + Capital-gain (continuous)
  + Capital-loss (continuous)
  + Hours-per-week (continuous)
  + Native-country (categorical)
  + Class (>50K, <=50K)

1. Number of Instances = 32561

*I downloaded this dataset in .data and .name file types. I first converted this information into .csv:*

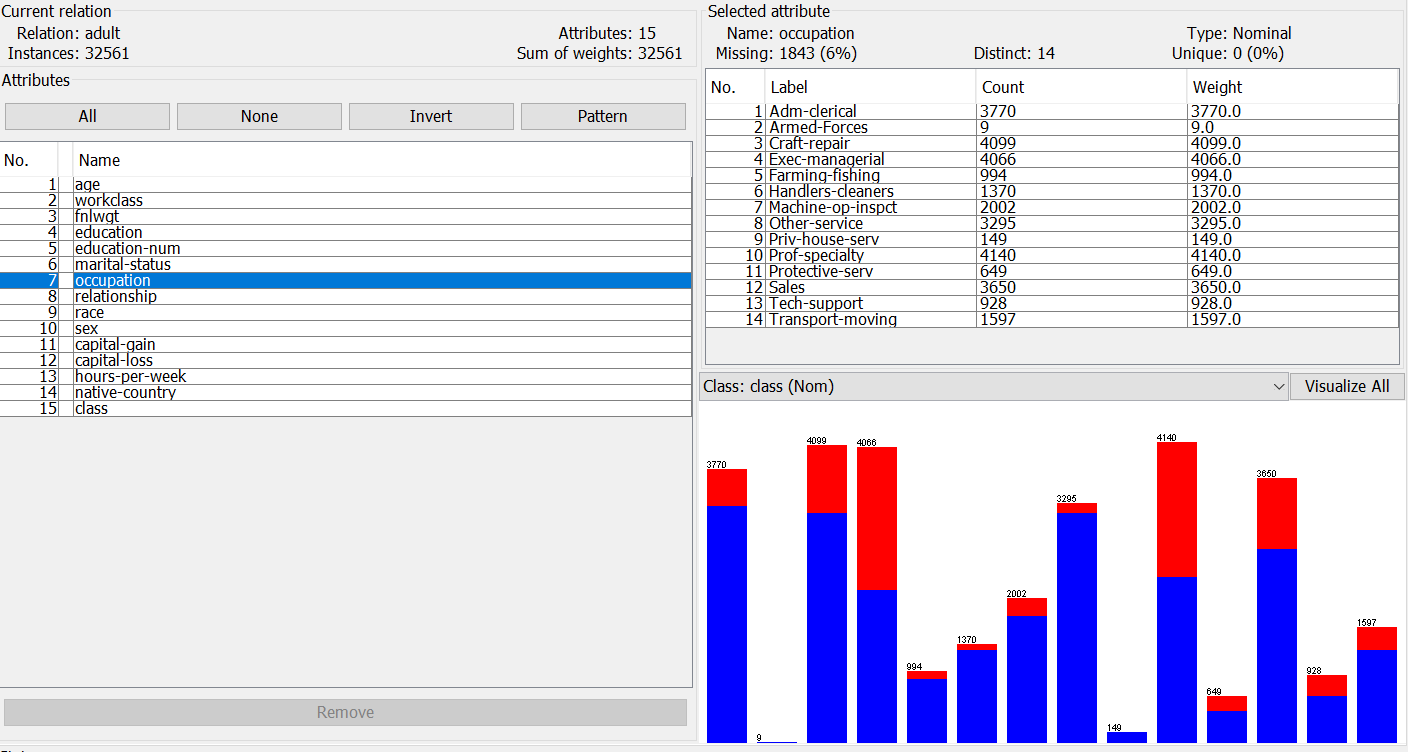


*From my csv file I converted the file into .arff for ease of use with Weka 3.8:*



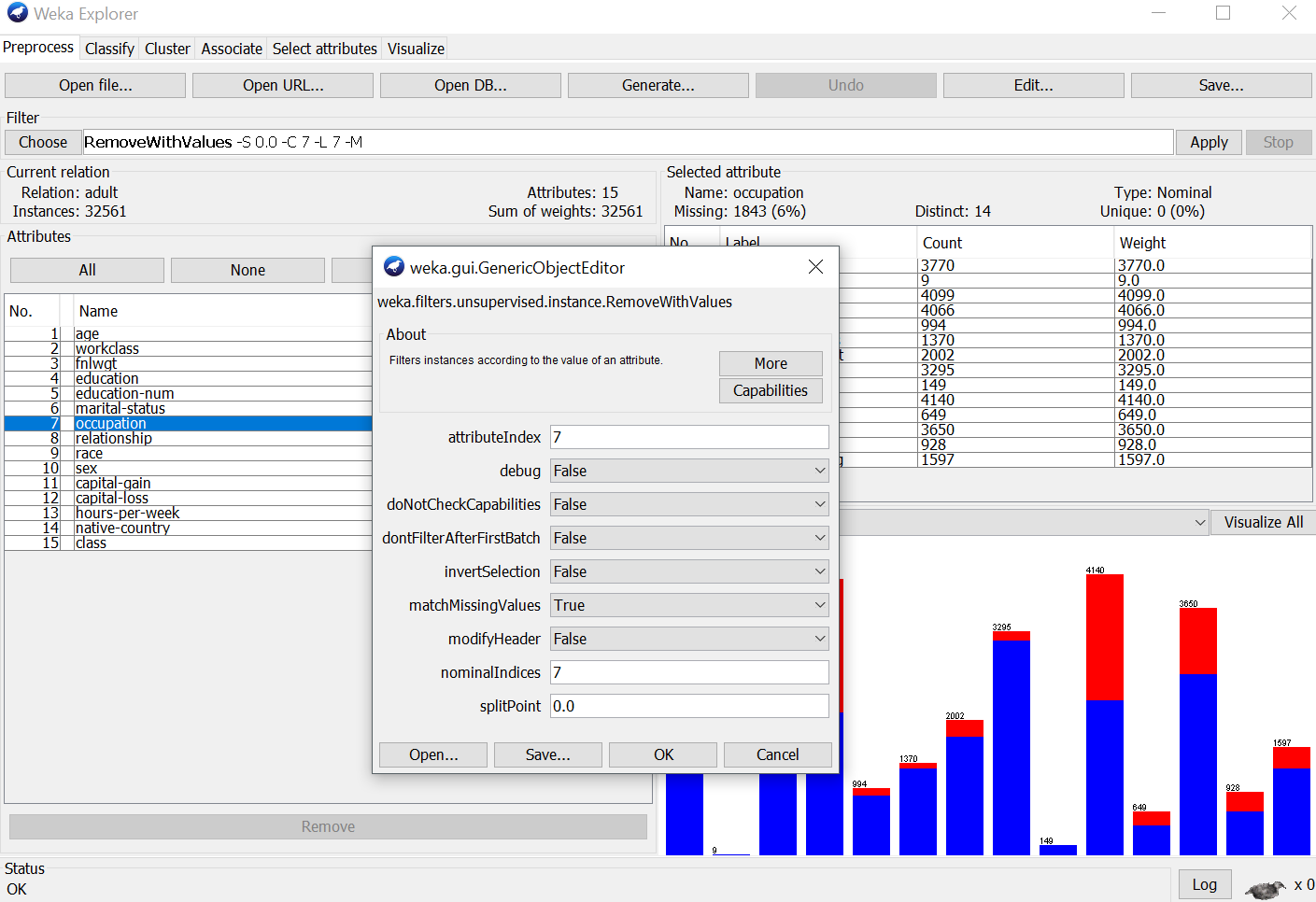
1. **Data Quality Problem Identified**

*After reviewing the data in the Explorer, I realized that the occupation attribute has 1842 missing values which is roughly 6% of the total data set:*

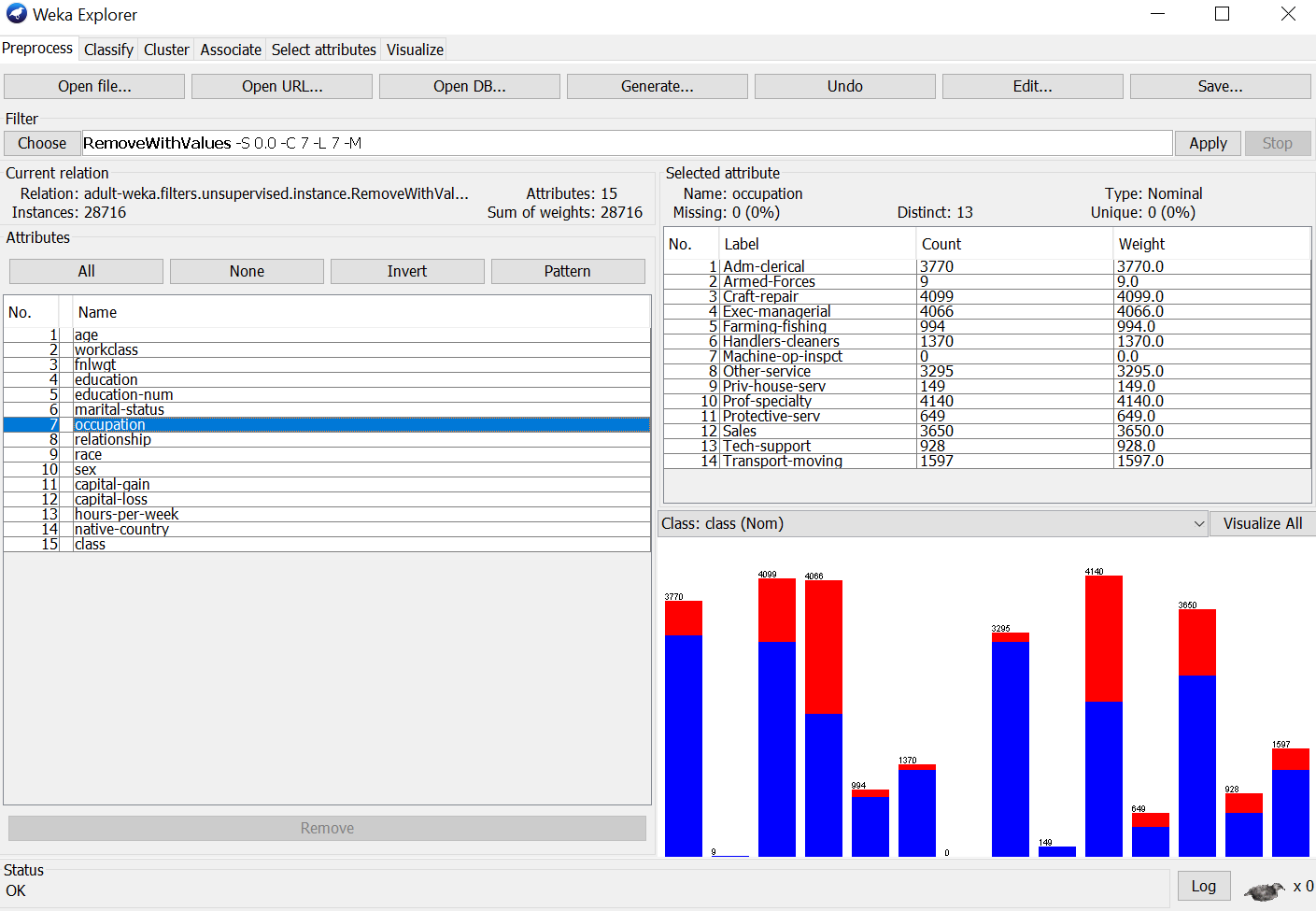


1. **Proposed Solution**

*Because this is a nominal attribute, it does not make sense to try and impute them like the other database and therefore I decided to simply remove these records. I did this using the unsupervised.instance.RemoveWithValues filter. I set the attributeIndicies and nomialIndicies to 7, the index of the occupation attribute. I marked ‘True’ for the matchMissingValues option as we have the instances marked in the dataset already:*



*After applying the filter, I was able to get rid of the missing values for occupation attribute:*



*Like before, I exported this clean dataset into a new csv file. Obviously, this new file will contain 1842 fewer records than the original database since we have removed some records.*