**Description of the original data Where was it obtained (web page link), description of the problem on hands: target of classification, original number of examples, original number of columns meaning and kind of values, impact of missing values, etc**

The data comes from Tanzanian Ministry of Water. Contain information about the water pumps in Tanzania.

The source web page is Driven Data. It contains 4 files as you can see in the link:

* TestSetValues.csv
* TrainSetValues.csv
* TrainSetLabels.csv
* SubmissionFormat.csv

[Competition: Pump it Up: Data Mining the Water Table (drivendata.org)](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/23/)

Data set info:

* 38 columns, 59407 rows.
* 6 numerical features
  + amount\_tsh - Total static head (amount water available to waterpoint)
  + gps\_height - Altitude of the well
  + longitude - GPS coordinate
  + latitude - GPS coordinate
  + population - Population around the well
  + construction\_year - Year the waterpoint was constructed
* 32 categorical features
  + date\_recorded - The date the row was entered
  + funder - Who funded the well
  + installer - Organization that installed the well
  + wpt\_name - Name of the waterpoint if there is one
  + basin - Geographic water basin
  + subvillage - Geographic location
  + region - Geographic location
  + …

With this data you have to classificate the pumps in 3 categories: **functional, no functional, functional but need to repair.**

**Impact of missing values:**  AT LESS there have this percentage of missing values:

