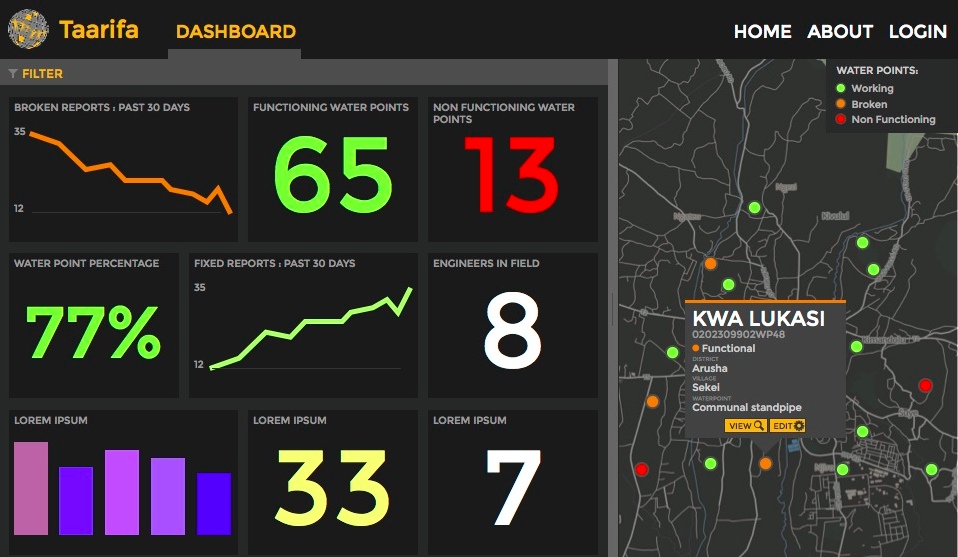
Problem description

This is where you'll find all of the documentation about this dataset and the problem we are trying to solve. For this competition, there are three subsections to the problem description:

* **Features**
* [List of features](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/#features_list)
* [Example of features](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/#features_eg)
* **Labels**
* [List of labels](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/#labels_list)
* **Submission Format**
* [Format example](https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/page/25/#sub_values)

The features in this dataset



Your goal is to predict the operating condition of a waterpoint for each record in the dataset. You are provided the following set of information about the waterpoints:

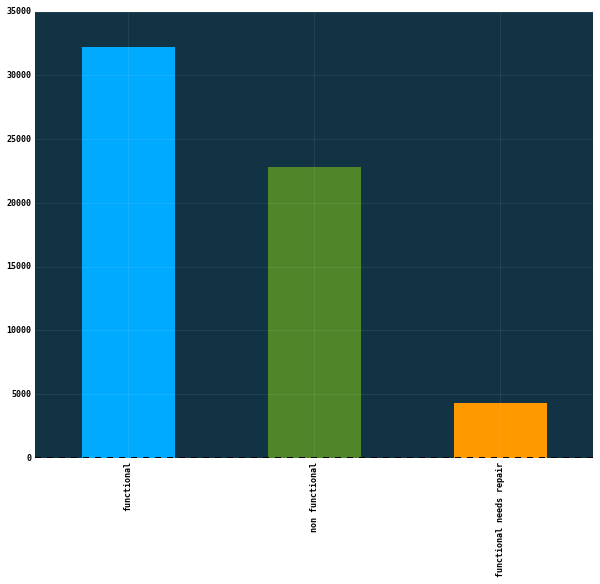
* amount\_tsh - Total static head (amount water available to waterpoint)
* date\_recorded - The date the row was entered
* funder - Who funded the well
* gps\_height - Altitude of the well
* installer - Organization that installed the well
* longitude - GPS coordinate
* latitude - GPS coordinate
* wpt\_name - Name of the waterpoint if there is one
* num\_private -
* basin - Geographic water basin
* subvillage - Geographic location
* region - Geographic location
* region\_code - Geographic location (coded)
* district\_code - Geographic location (coded)
* lga - Geographic location
* ward - Geographic location
* population - Population around the well
* public\_meeting - True/False
* recorded\_by - Group entering this row of data
* scheme\_management - Who operates the waterpoint
* scheme\_name - Who operates the waterpoint
* permit - If the waterpoint is permitted
* construction\_year - Year the waterpoint was constructed
* extraction\_type - The kind of extraction the waterpoint uses
* extraction\_type\_group - The kind of extraction the waterpoint uses
* extraction\_type\_class - The kind of extraction the waterpoint uses
* management - How the waterpoint is managed
* management\_group - How the waterpoint is managed
* payment - What the water costs
* payment\_type - What the water costs
* water\_quality - The quality of the water
* quality\_group - The quality of the water
* quantity - The quantity of water
* quantity\_group - The quantity of water
* source - The source of the water
* source\_type - The source of the water
* source\_class - The source of the water
* waterpoint\_type - The kind of waterpoint
* waterpoint\_type\_group - The kind of waterpoint

Feature data example

For example, a single row in the dataset might have these values:

|  |  |
| --- | --- |
| **amount\_tsh** | 300.0 |
| **date\_recorded** | 2013-02-26 |
| **funder** | Germany Republi |
| **gps\_height** | 1335 |
| **installer** | CES |
| **longitude** | 37.2029845 |
| **latitude** | -3.22870286 |
| **wpt\_name** | Kwaa Hassan Ismail |
| **num\_private** | 0 |
| **basin** | Pangani |
| **subvillage** | Bwani |
| **region** | Kilimanjaro |
| **region\_code** | 3 |
| **district\_code** | 5 |
| **lga** | Hai |
| **ward** | Machame Uroki |
| **population** | 25 |
| **public\_meeting** | True |
| **recorded\_by** | GeoData Consultants Ltd |
| **scheme\_management** | Water Board |
| **scheme\_name** | Uroki-Bomang'ombe water sup |
| **permit** | True |
| **construction\_year** | 1995 |
| **extraction\_type** | gravity |
| **extraction\_type\_group** | gravity |
| **extraction\_type\_class** | gravity |
| **management** | water board |
| **management\_group** | user-group |
| **payment** | other |
| **payment\_type** | other |
| **water\_quality** | soft |
| **quality\_group** | good |
| **quantity** | enough |
| **quantity\_group** | enough |
| **source** | spring |
| **source\_type** | spring |
| **source\_class** | groundwater |
| **waterpoint\_type** | communal standpipe |
| **waterpoint\_type\_group** | communal standpipe |

The labels in this dataset



Distribution of Labels

The labels in this dataset are simple. There are three possible values:

* functional - the waterpoint is operational and there are no repairs needed
* functional needs repair - the waterpoint is operational, but needs repairs
* non functional - the waterpoint is not operational

Submission format

The format for the submission file is simply the row id and the predicted label (for an example, see SubmissionFormat.csv on the data download page.

For example, if you just predicted that all the waterpoints were functional you would have the following predictions:

| **id** | **status\_group** |
| --- | --- |
| **50785** | functional |
| **51630** | functional |
| **17168** | functional |
| **45559** | functional |
| **49871** | functional |

Your .csv file that you submit would look like:

id,status\_group

50785,functional

51630,functional

17168,functional

45559,functional

...

Good luck!