Pump it Up: Data Mining the Water Table

HOSTED BY DRIVENDATA

HOME (/COMPETITIONS/7/PAGE/23/)

ABOUT TAARIFA (/COMPETITIONS/7/PAGE/24/)

PROBLEM DESCRIPTION (/COMPETITIONS/7/PAGE/25/)

6
MONTHS LEFT

Glory!

LEADERBOARD (/COMPETITIONS/7/LEADERBOARD/)

DATA DOWNLOAD (/COMPETITIONS/7/DATA/)

SUBMISSIONS (/COMPETITIONS/7/SUBMISSIONS/)

TEAM (/COMPETITIONS/7/TEAM/)

DISCUSSION (HTTPS://COMMUNITY.DRIVENDATA.ORG/C/PUMP-IT-UP-DATA-MINING-THE-WATER-TABLE) 12

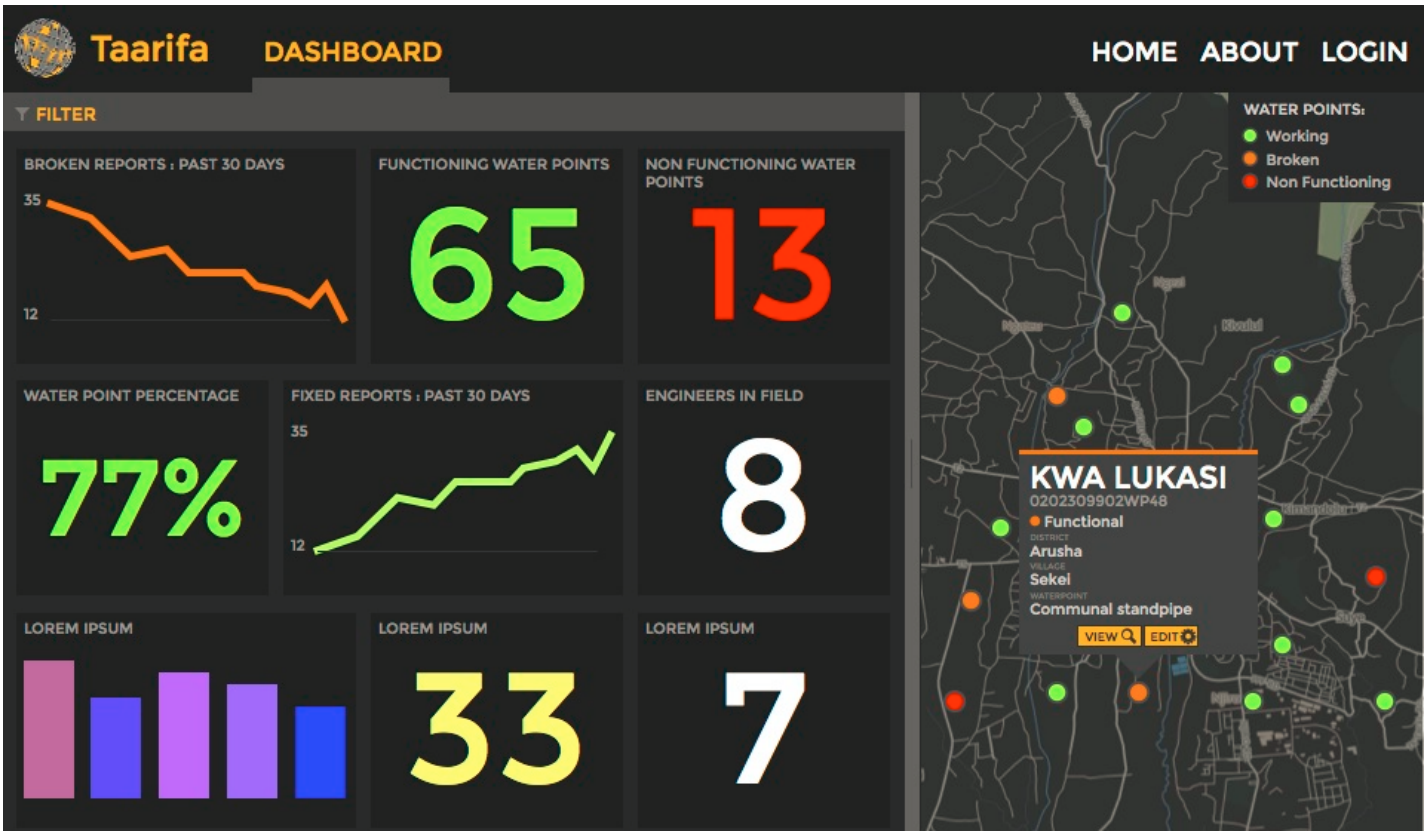
OFFICIAL RULES (/COMPETITIONS/7/RULES/)

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	**Labels**	**Submission Format**
List of features	List of labels	Format example
Example of features		

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

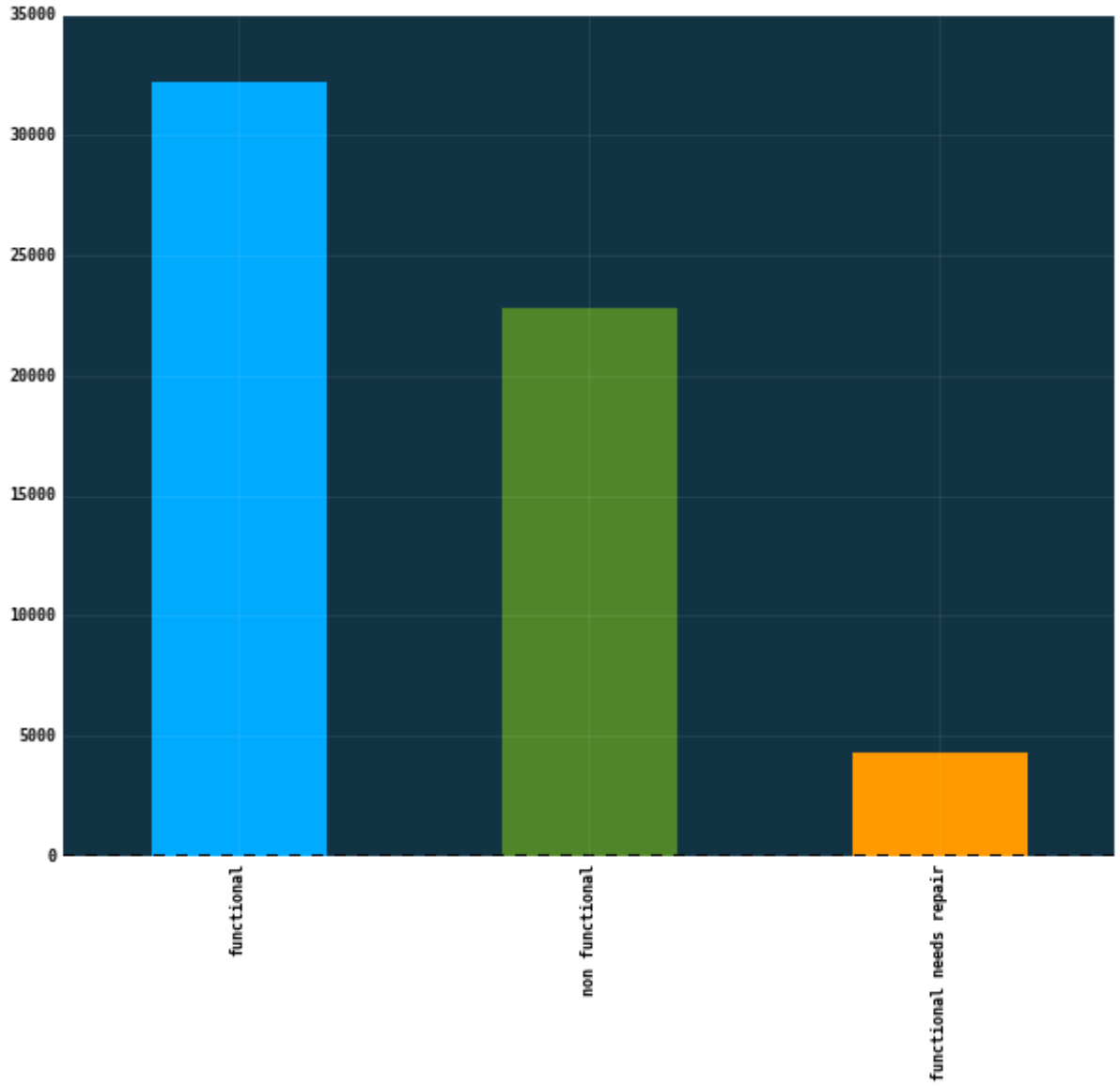
Feature data example

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 (/competitions/7/data/).

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!

Good luck and enjoy this problem! If you have any questions you can always visit the user forum (http://community.drivendata.org/)!

Denver, CO 80204