Taarifa Report

Authored by Kshitiz Gupta.

Contents

Introduction **1**

Response Variable (Variable that we are trying to predict) **1**

Data Dictionary **2**

Administrative Variables **3**

Examining Missing Data Values **3**

Key Data Insights from Data Analysis **4**

ML Modeling Results **5**
Most Predictive Features 5
Least Predictive Features 25

References **26**

Introduction

In the developing world, people's health benefits and their willingness to pay for clean water are best realized when clean water infrastructure performs extremely well. Realizing this the Tanzanian Ministry of Water teamed up with Taarifa, a non-profit to host a Data Science Competition to predict water point failures in order to better maintain the country's water infrastructure. The Data Science Competition hosted on www.drivendata.com quickly became popular and data scientists from worldwide participated, exploring the use of a variety of predictive/classification modeling techniques. Throughout this report the terms 'pump' and 'waterpoint' have been used interchangeably to mean the same thing.

The training data set here consists of 59,400 Tanzanian water pumps and predictions are to be made on a test set of 14,850 waterpoints. Every water pump has 39 categorical and numerical featues (attributes) like the type of pump, organization that installed the pump, funder of the pump, location, altitude, year constructed, management method, water source, and the quality of the water delivered by the pump.

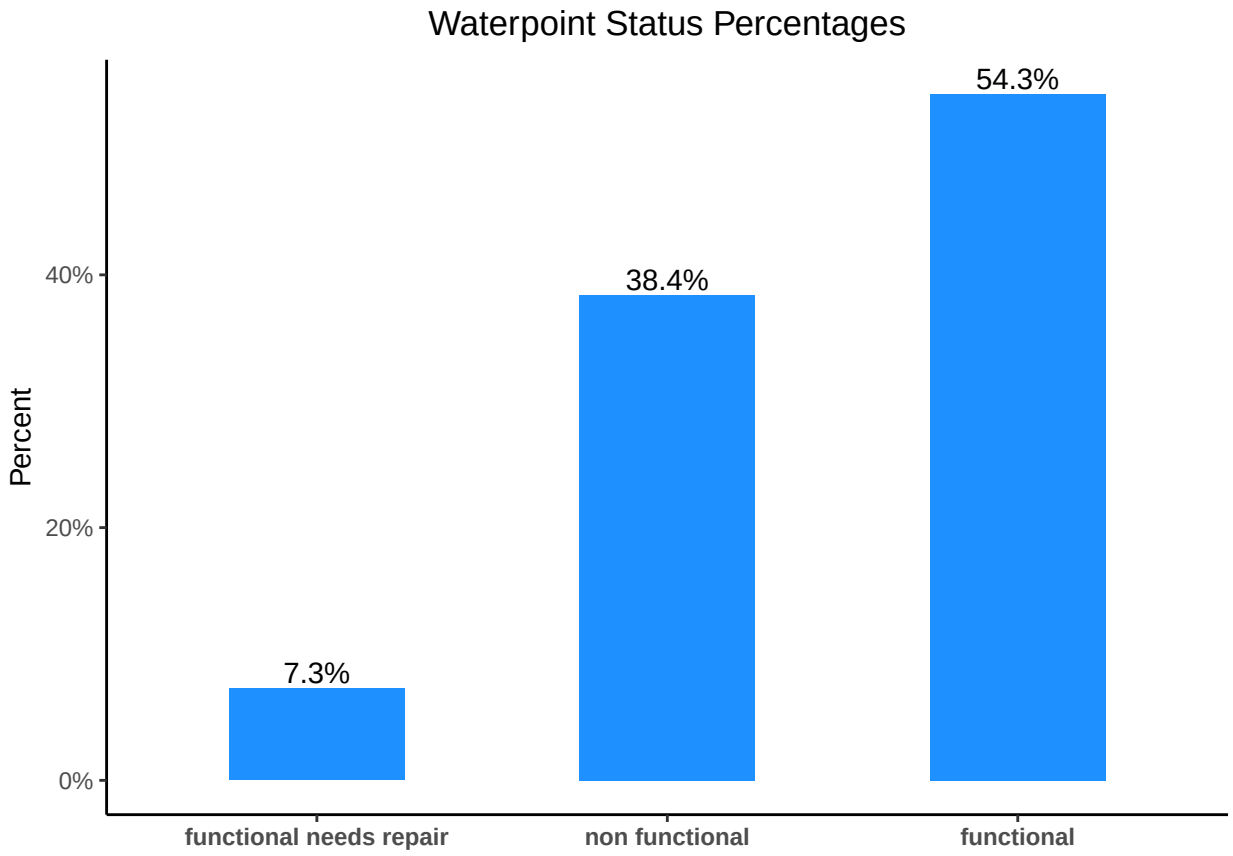
Response Variable (Variable that we are trying to predict)

The response variable `status_group` has three possible classes:

`functional`: The pump is operating and does not need any repair/maintenance

`functional needs repair`: The pump is operating but needs repair

`non functional`: The pump is not operating



As the plot tells us 54.3% of all pumps in the training data were found to be functional, 7.3% had a status of functional needs repair and 38.4% were not functioning and so we notice severe **class imbalance** in this dataset which makes highly accurate classifications of the three classes difficult.

Data Dictionary

The Data Dictionary provided in the DrivenData competition doesn't provide a comprehensive and detailed description of the variables and this is rectified in the updated data dictionary below:

Variable	Definition
id	Identification Variable (not a predictor)
date_recorded	Date of data collection by survey company
recorded_by	Name of the data collection / survey company
amount_tsh	Metric indicating total static head for pump; should be > 0
gps_height	Altitude of pump
population	Human population in immediate vicinty of pump
longitude	Longitudinal coordinate of pump
latitude	Latitudinal coordinate of pump
num_private	No definition is available for this variable
funder	Name of organization that funded installation of pump
installer	Name of organization that installed the pump
wpt_name	Name assigned to given waterpoint
basin	Name of geographic water basin where pump is located
subvillage	Name of geographic subvillage where pump is located
region	Name of geographic region in Tanzania where pump is located

Variable	Definition
region_code	Numeric ID for region variable
district_code	Numeric ID of district within a region where pump is located
lga	Tanzania-specific geographic indicator of where pump is located
ward	Name of Tanzanian geographic ward where pump is located
public_meeting	"True/False" indicator
scheme_management	Type of the organization responsible for management of pump
scheme_name	Name of organization responsible for management of pump
permit	"True/False"" indicating whether the pump has valid permit
construction_year	The year the pump was installed
extraction_type	Method of extraction used at a given pump site
extraction_type_group	Aggregation of extraction_type categories
extraction_type_class	Aggregation of extraction_type_group categories
management	Name of method employed for management of a given pump
management_group	Possibly an aggregation of management categories
payment	Categorical indicator of payment method required of pump users
payment_type	Appears to be a duplicate of payment categories
water_quality	Categorical indicator of water quality produced by pump
quality_group	Aggregation of water_quality categories
quantity	Categorical indicator of water quantity produced by pump
quantity_group	Aggregation of quantity categories
source	Type of source of the water for a given pump
source_type	Aggregation of source categories
source_type_class	Aggregation of source_type categories
waterpoint_type	The type of pump installed at a well site
waterpoint_type_group	Aggregation of waterpoint_type categories

Administrative Variables

Aside from **categorical** and **numeric variables**, the dataset also contain administrative variables which act as data management attributes within the data set. These are:

Administrative Variables	Description
id	Identification Variable (not a predictor)
date_recorded	Date of data collection by survey company
recorded_by	Name of the data collection / survey company

Examining Missing Data Values

Every non-categorical numeric variable in the dataset has significant number of invalid data values represented by zeroes:

Numeric Variables	Invalid Data Comments
amount_tsh	41,639 of 59,400 records = "0"
gps_height	20,438 of 59,400 records = "0"
population	21,381 of 59,400 records = "0"
longitude	1,812 zero values: likely invalid
latitude	1,819 values < -1: likely invalid

Numeric Variables	Invalid Data Comments
num_private	58,643 of 59,400 records = "0"

Additionally, eight categorical variables contain missing data values (as indicated by either 'NA' values, zeroes, or blank character strings):

Categorical Var.	No. of Distinct Values	Missing	Comments
funder	1898	3635	3582 NA's coinc. w installer
installer	2146	3655	3582 NA's coinc. w funder
subvillage	19288	371	
public_meeting	3	3334	valid values: TRUE/FALSE
scheme_management	13	3877	
scheme_name 2697	28166		
permit	3	3056	valid values: TRUE/FALSE
construction_year	55	20709	valid values: 1960 - 2013

More insights gained by analysis of the missing/invalid data:

- 31,587 (53.17%) of the 59,400 records in the data set have missing data values.
- 69.2% of subvillages, 52.1% of wards, 32% of lga's, and 19% of regions have no valid `amount_tsh` values.
- Four regions, namely Dodoma, Kagera, Mbeya, Tabora, were found to have all zero values for the following variables:

  1. `amount_tsh`
  2. `gps_height`
  3. `construction_year`
  4. `num_private`
  5. `population`

  These 4 regions correspond to 12,115 of the 59,400 records i.e. 20.39% of the data set including 27 of the unique lga's, 514 of the unique wards and 4644 of the unique subvillages. The 12,115 records covered by these regions contain 60% of the zero values found within the `gps_height` (12,115 / 20,438), `population` (12,115 / 21,381), and `construction_year` (12,115 / 20,709) variables.Since most of the missing values occur within regions of Dodoma, Kagera, Mbeya, Tabora, it suggests a systematic data collection issue within these 4 geographic areas and it seems that the waterpoints located within these regions were not actually visited by the surveying company (GeoData Consultants Ltd).
- 3,582 of the missing `funder` and `installer` values coincide with each other.
- 1,812 of the missing `gps_height` values coincide with missing `latitude` and `longitude` values.
- 43 wards have no valid `public_meeting` values; 75 wards have no valid `scheme_management` values; 73 wards and 3 lga's have no valid `permit` values.

Key Data Insights from Data Analysis

Analysis of data by aggregation of various variables/values provided additional insights like:

- In the`funder` and `installer` variables there are lots of instances where a single valid data value is represented by different variations. E.g. 'Oxfam' vs. 'oxfarm', 'Government of Tanzania' vs. 'Ministry of Water' vs. 'Water'

- There is no one-to-one relationship between the region and region_code variables (i.e., 21 distinct regions; 27 distinct region_codes).
- Aggregating by `extraction_type` we find an apparent chronological progression in the deployment of different pump technologies. For example, the median construction_year value for swn 80 (a type of hand pump) to be 1997 while that of the climax (a type of motor pump) is 2012.
- After aggregating by `basin` it seems that there is an apparent chronological geographic progression in the installation of these waterpoints throughout Tanzania. For example, the Lake Nyasa basin has a median construction_year value of 1980 while the Wami / Ruvu basin's median construction_year is much later (2003).
- The location of the water points seems to influence the amount of functional water points. It is hard to say whether this is due to geographical properties, like their geographical height, of the respective regions or political factors.
- Several variables contained within the data set are either duplicative or binned versions of other variables:
    1. `extraction_type_group` is a binned/aggregated version of `extraction_type`
    2. `extraction_type_class` is a binned/aggregated version of `extraction_type_group`
    3. `quality_group` is a binned/aggregated version of `water_quality`
    4. `source_type` is a binned/aggregated version of `source`
    5. `waterpoint_type_class` is a binned/aggregated version of `waterpoint_type`
    6. `payment_type` is 100% duplicative of `payment`
    7. `quantity_group` is 100% duplicative of `quantity`

ML Modeling Results

We observe that Random Forest model achieved an overall accuracy of 81.13% and so if overall accuracy across each of the three possible water pump statuses is of most importance then Random Forest should be used. However, if correctly identifying the largest number of pumps that are functional but in need of repair is the top priority, then Bootstrap Aggregation model should be adapted because even though it has a lower overall accuracy (79.91%) it can classify 'functional needs repair' pumps with greater accuracy (67.6%) than Random Forest (64.1%).

Most Predictive Features

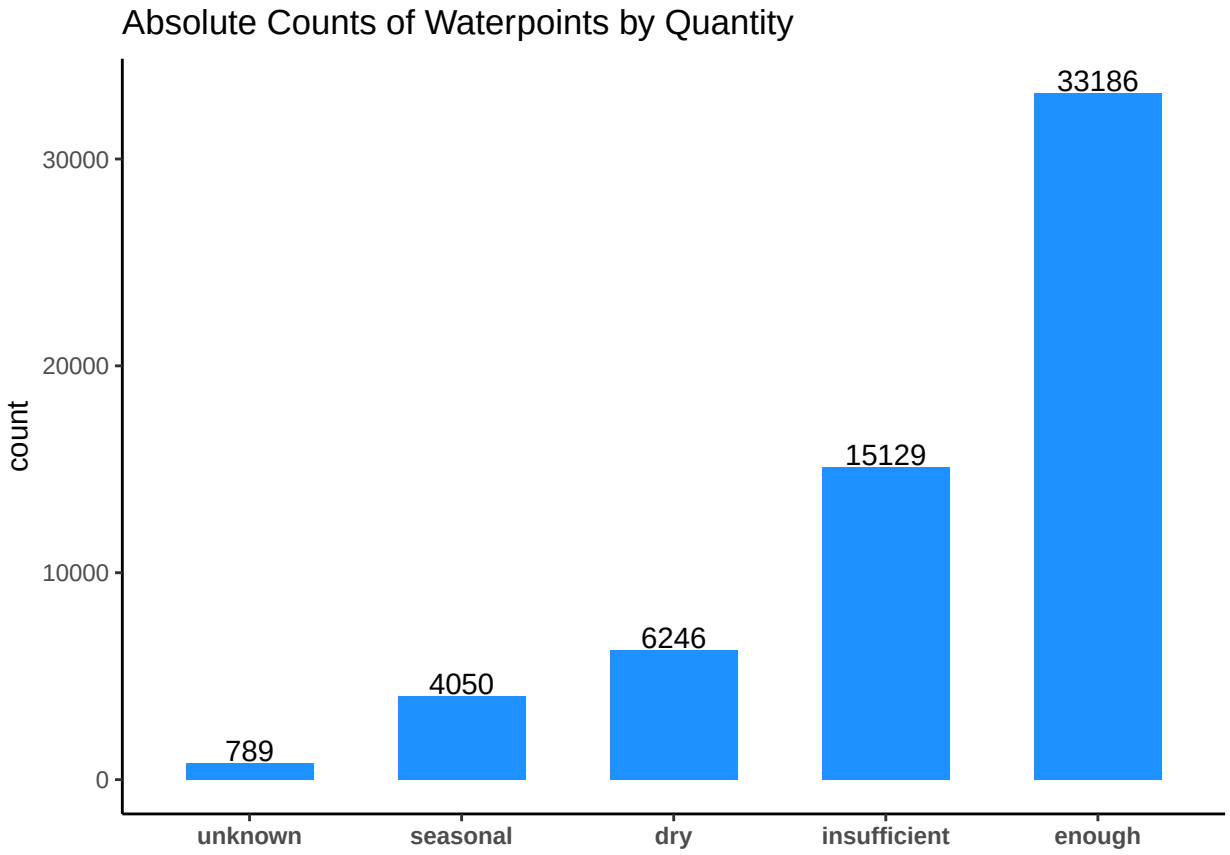
`quantity`

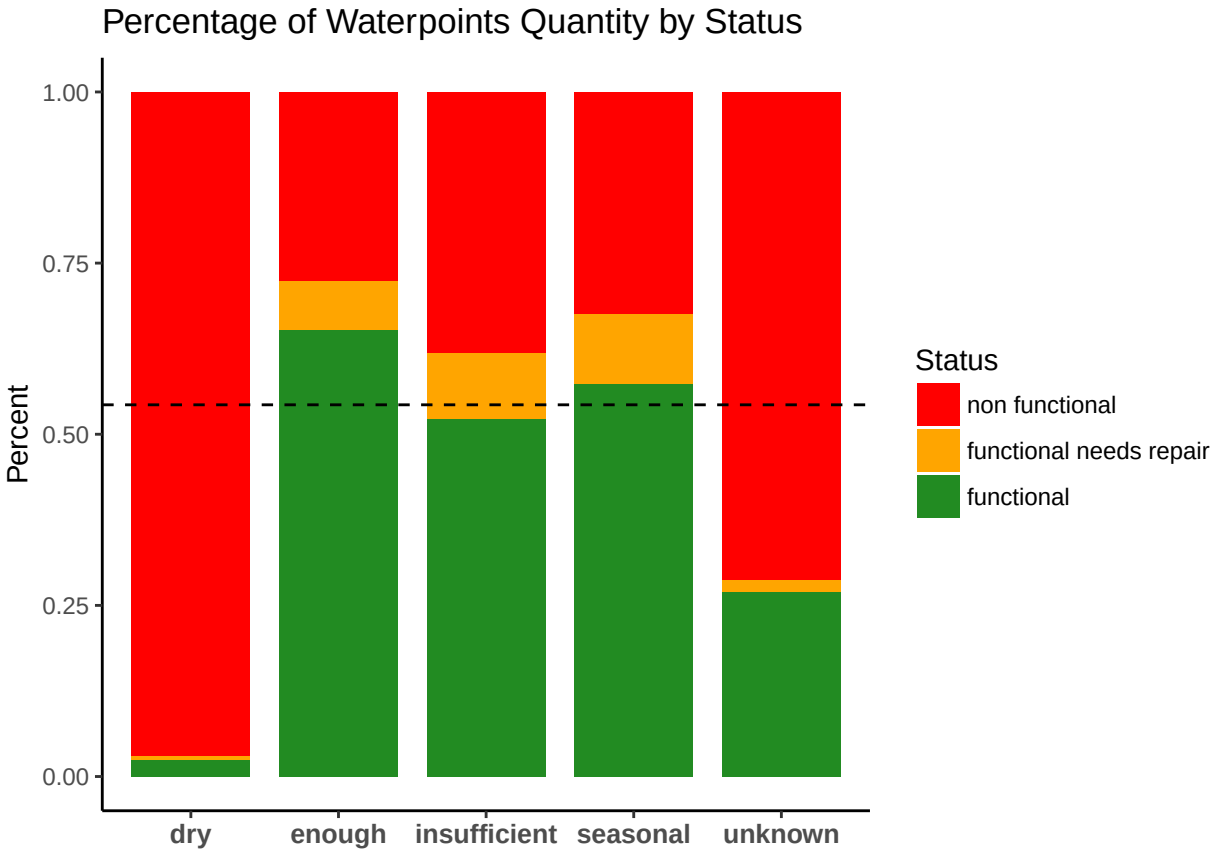
`quantity` represents the quantity of water available at the waterpoint and this column has no missing values. There are total of 5 distinct quantity categories :

```
## # A tibble: 5 x 2
##   quantity     TotalWaterPoints
##   <chr>                   <int>
## 1 enough                  33186
## 2 insufficient            15129
## 3 dry                      6246
## 4 seasonal                 4050
## 5 unknown                   789
```

In the plots shown below we observe that nearly all i.e. 91% of dry water points are non-functional. 71% of water points with unknown quantity description are non-functional as well. On the other hand, if the

quantity level is 'enough' then there is a higher chance the water point is functional, that is, 65% of water pumps with 'enough' quantity are functional.





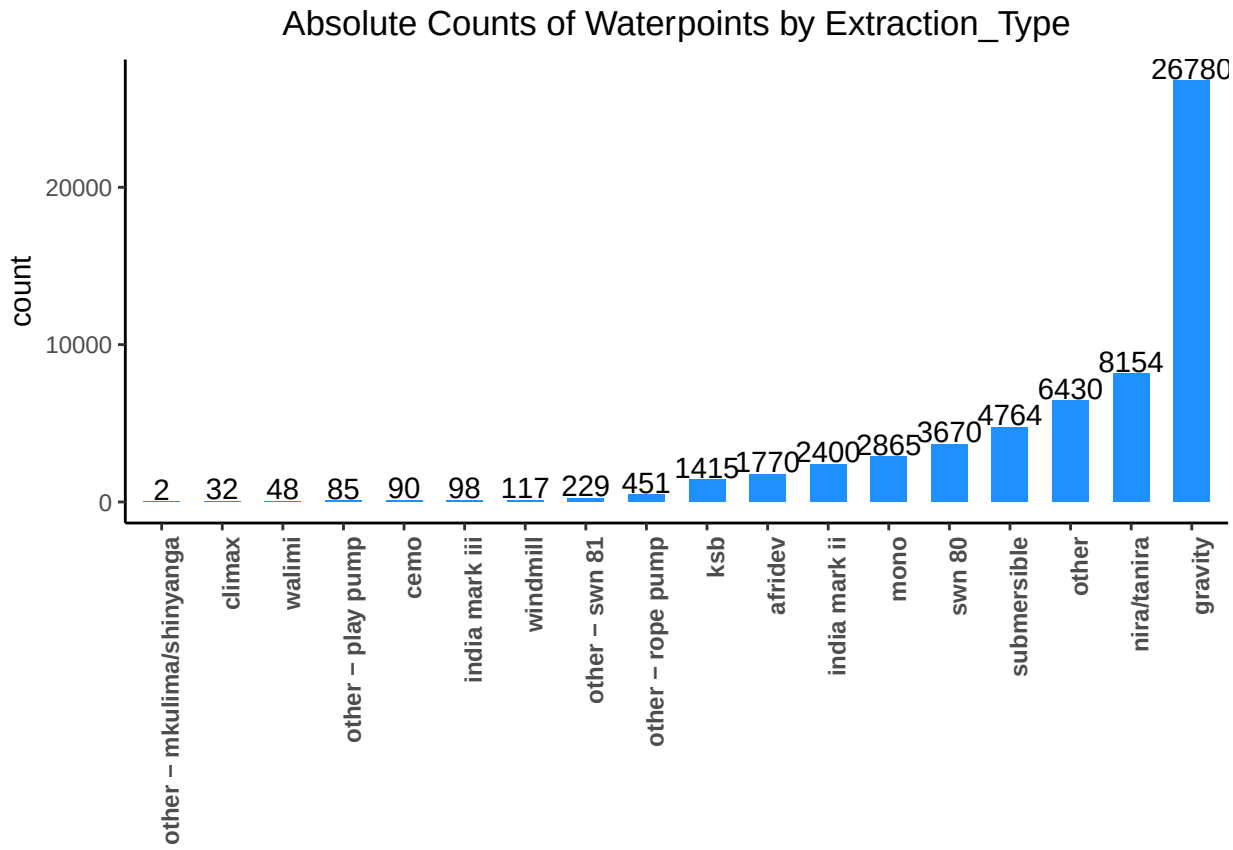
The black dashed line in the plot represents 54.3% which is the overall proportion of functional waterpoints in the dataset. Here, that proportion is used as a benchmark to compare against.

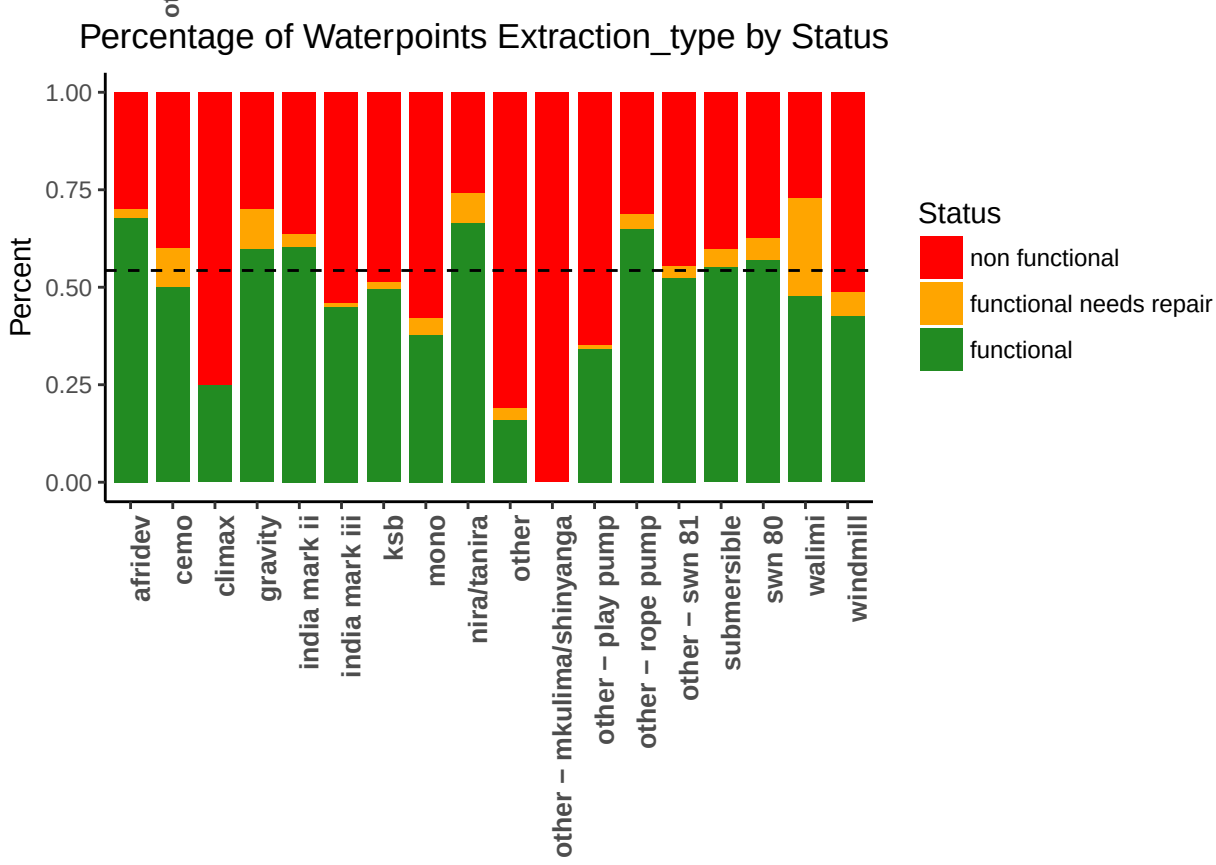
extraction_type

The extraction_type variable represents the method of extraction used by a given pump and there are no missing values for this variable. A total of 18 unique extraction_type values are found within the data set. Ten most frequently occuring extraction_types are shown below:

```
## # A tibble: 10 x 2
##    extraction_type   TotalWaterPoints
##    <chr>                        <int>
##  1 gravity                      26780
##  2 nira/tanira                   8154
##  3 other                         6430
##  4 submersible                   4764
##  5 swn 80                        3670
##  6 mono                          2865
##  7 india mark ii                 2400
##  8 afridev                       1770
##  9 ksb                           1415
## 10 other - rope pump              451
```

Plots below show that seven of the eighten extraction types have functional metrics that exceed the overall 54.3% benchmark. Of particular interest here is the gravity type since nearly 27,000 of the 59,400 total pumps rely on that approach, a far higher percentage than any of the other extraction types.





Additional features like `extraction_type_group` which is a composite version of `extraction_type` and

`extraction_type_class` which is a composite version of `extraction_type_group` (and thus a super composite version of `extraction_type`) are also available in the dataset. As the table below indicates both of these features can help us identify what broad categories do the specific extraction methods fall into.

Table 5: Granulariy of `extraction_type`

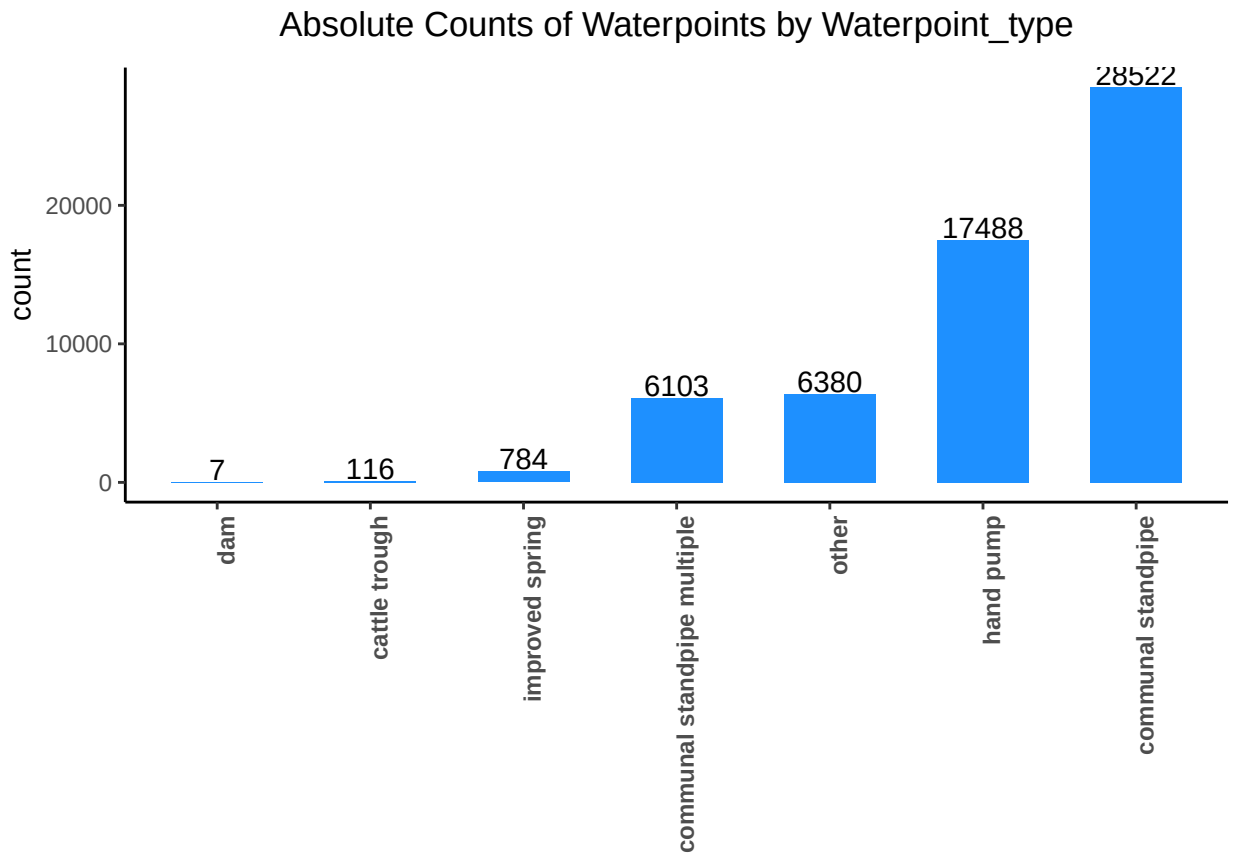
Type	Group	Class
Gravity	Gravity	Gravity
Afridev	Afridev	Hand pump
India Mark II	India Mark II	Hand pump
India Mark III	India Mark III	Hand pump
Nira/tanira	Nira/tanira	Hand pump
Swn 80	Swn 80	Hand pump
Other – play pump	Other Handpump	Hand pump
Other – Swn 81	Other Handpump	Hand pump
Walimi	Other Handpump	Hand pump
Other – mkulima / shinyanga	Other Handpump	Hand pump
Other	Other	Other
Submersible	Submersible	Submersible
KSB	Submersible	Submersible
Climax	Other Motor pump	Motor pump
Cemo	Other Motor pump	Motor pump
Mono	Mono	Motor pump
Windmill	Wind-powered	Wind-powered
Other - rope pump	Rope pump	Rope pump

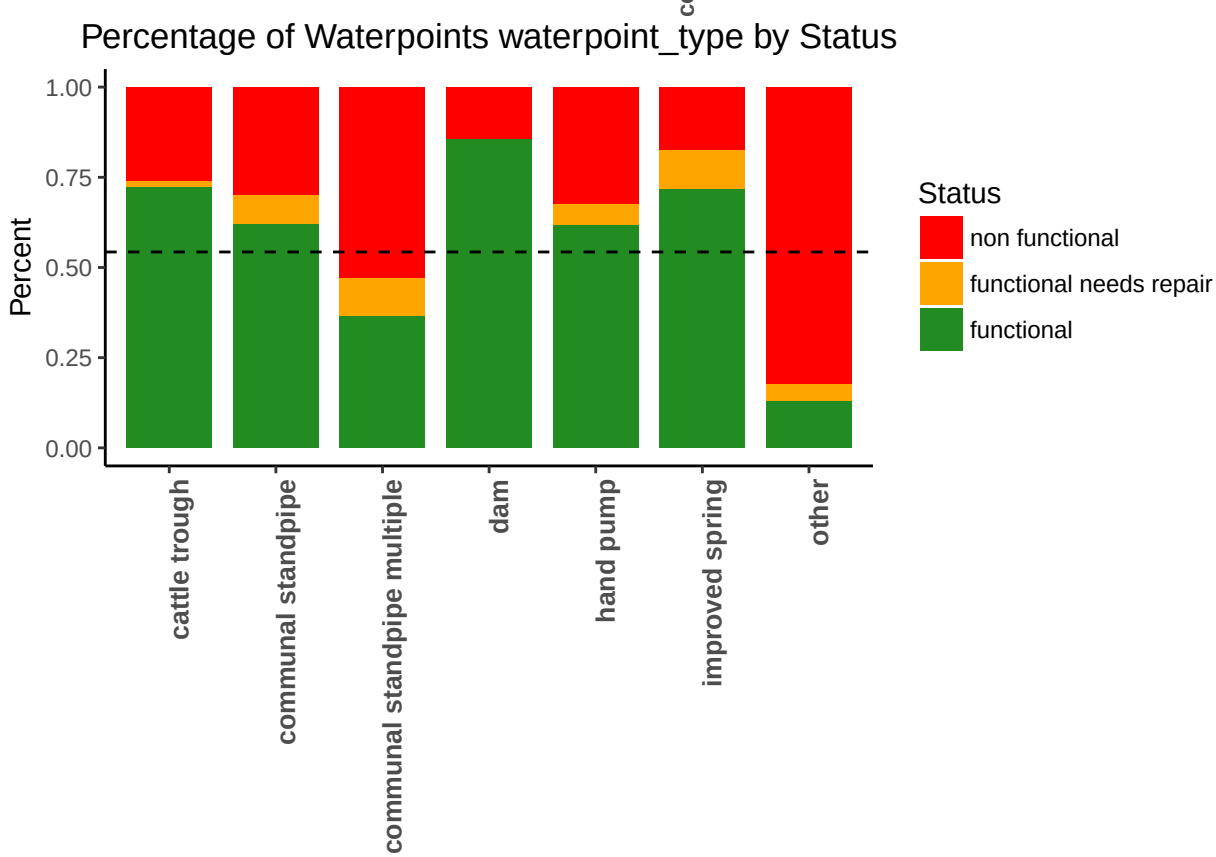
`waterpoint_type`

The waterpoint_type variable indicates the type of pump installed at a given location and all the records in the dataset have a valid value i.e no missing or invalid values. There are total 7 unique waterpoint_type values within the data set:

```
## # A tibble: 7 x 2
##   waterpoint_type             TotalWaterPoints
##   <chr>                                  <int>
## 1 communal standpipe                     28522
## 2 hand pump                              17488
## 3 other                                   6380
## 4 communal standpipe multiple             6103
## 5 improved spring                          784
## 6 cattle trough                            116
## 7 dam                                        7
```

As the plots indicate while waterpoints pumps having a dam as their waterpoint are the most likely to be functional there are only 7 pumps having a waterpoint_type of dam. Similar is the case with 'cattle trough' waterpoints, a large proportion of them are functional but very few of them exist. Interestingly, even though 'communal standpipe multiple' waterpoints are least likely to be functional, 'communal standpipe' waterpoints perform well relative to the overall functional benchmark of 54.3%. In fact, the most common waterpoint_type is communal standpipe followed by 'hand pump' waterpoints which also have a decent functional rate of 61.8%.


Absolute Counts of Waterpoints by Waterpoint_type
count
20000
10000
0
7
116
784
6103
6380
17488
28522
dam
cattle trough
improved spring
communal standpipe multiple
other
hand pump
communal standpipe


Percentage of Waterpoints waterpoint_type by Status
Percent
1.00
0.75
0.50
0.25
0.00
cattle trough
communal standpipe
communal standpipe multiple
dam
hand pump
improved spring
other
Status
non functional
functional needs repair
functional

construction_year

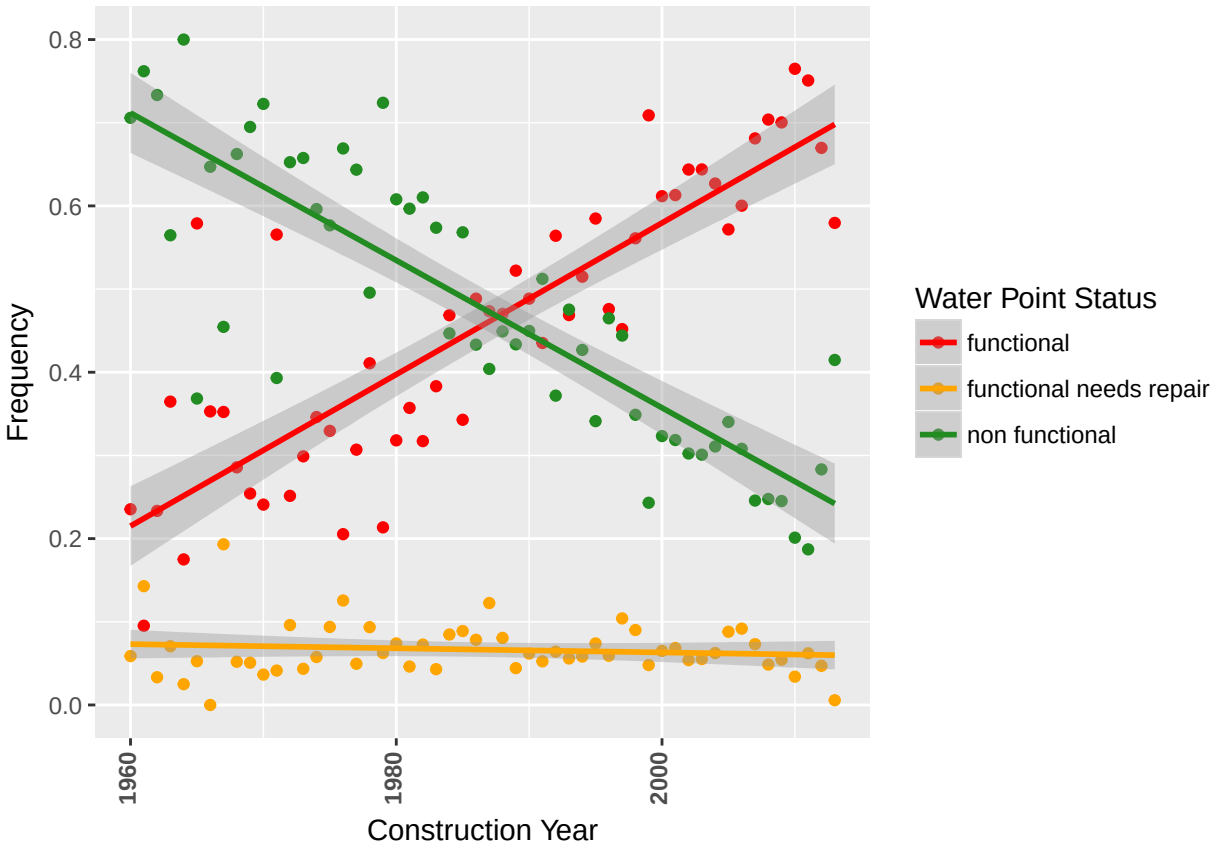
The construction_year variable represents the year in which a given pump was installed. There are 54 unique non-zero construction year values ranging from 1960 through 2013, with 20,709 missing values (encoded as "0"). The summary statistics for construction_year are shown below:

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1960    1987    2000    1997    2008    2013
```

The year values having the largest number of pump installations are shown below:

```
## # A tibble: 20 x 2
##    construction_year TotalWaterPoints
##                <int>            <int>
##  1                 0            20709
##  2              2010             2645
##  3              2008             2613
##  4              2009             2533
##  5              2000             2091
##  6              2007             1587
##  7              2006             1471
##  8              2003             1286
##  9              2011             1256
## 10              2004             1123
## 11              2012             1084
## 12              2002             1075
## 13              1978             1037
## 14              1995             1014
## 15              2005             1011
## 16              1999              979
## 17              1998              966
## 18              1990              954
## 19              1985              945
## 20              1980              811
```

To visualize the relationship between construction_year of the waterpoint and its status we can fit a linear model as shown below.



We see there is a clear linear relationship between status and construction_year for 'functional' and 'non-functional' waterpoints. That's not the case with 'functional needs repair' status, maybe because there are only very few data entries within this class and because 'functional needs repair' is a temporary state for a waterpoint.

`latitude` & `longitude`

`longitude` and `latitude` variables represent the longitude and latitude coordinates of the waterpoint. Searching for invalid values, we find 1812 zero values within the longitude variable and 1819 values less than -1 for the latitude variable and every instance of a zero longitude value corresponds to an instance of a ($<$ -1) latitude value. These values are invalid because based on Tanzania's geographic location valid entries should have a latitude between -11,73 and -1 and a longitude between 29.50 and 40.37.

Analysis of latitude and longitude tells us that pumps located between 34 and 38 degrees longitude are more likley to be functional than pumps at other longitudes, while those located between 30 and 31 degrees longitude are much more likely to have a status of functional needs repair than pumps located at other longitudes. In this way, longitude seems to be somewhat indicative of how likely a pump is to be non functional.

Using latitude and longitude we can see the distribution of three waterpoint statuses (functional, functional needs repair and non-functional) across Tanzania:

Fig 1. All Waterpoints Across Tanzania

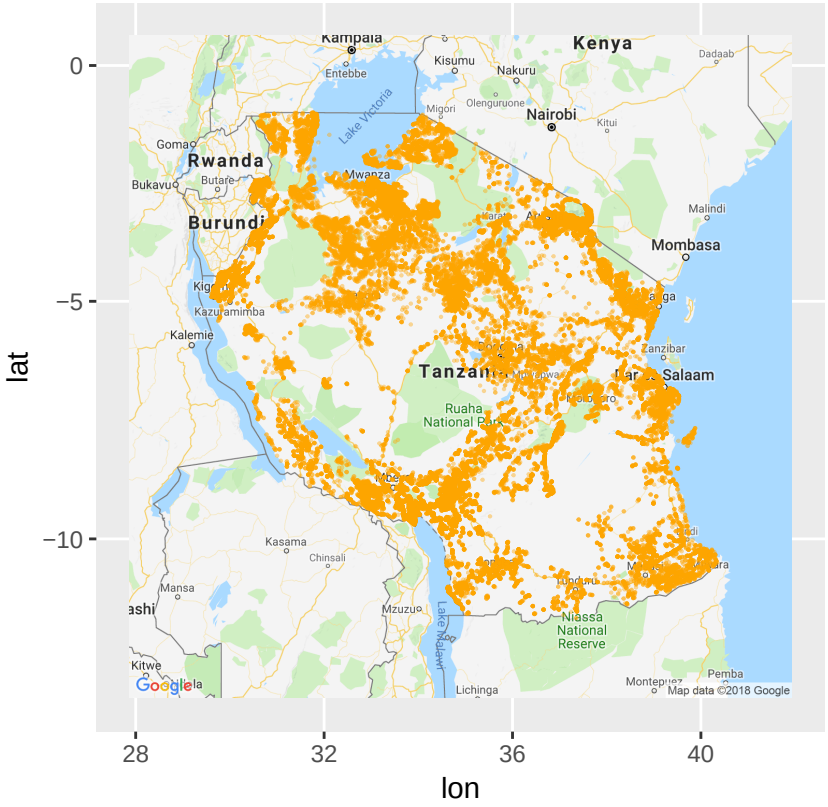


Fig 2. Functional Waterpoints Map

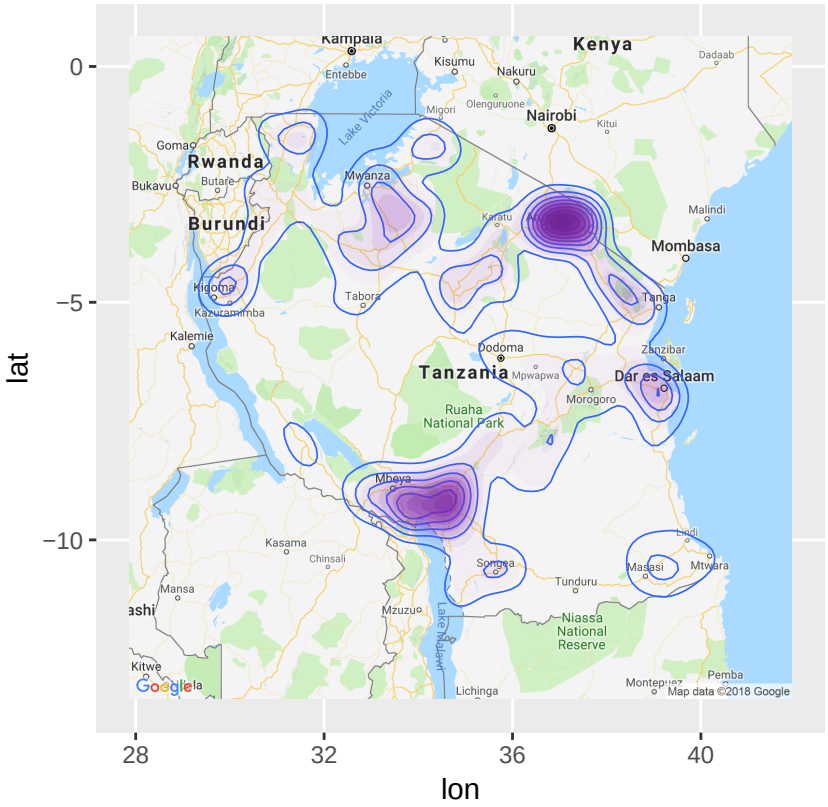


Fig 3. Functional Needs Repair Waterpoints Map

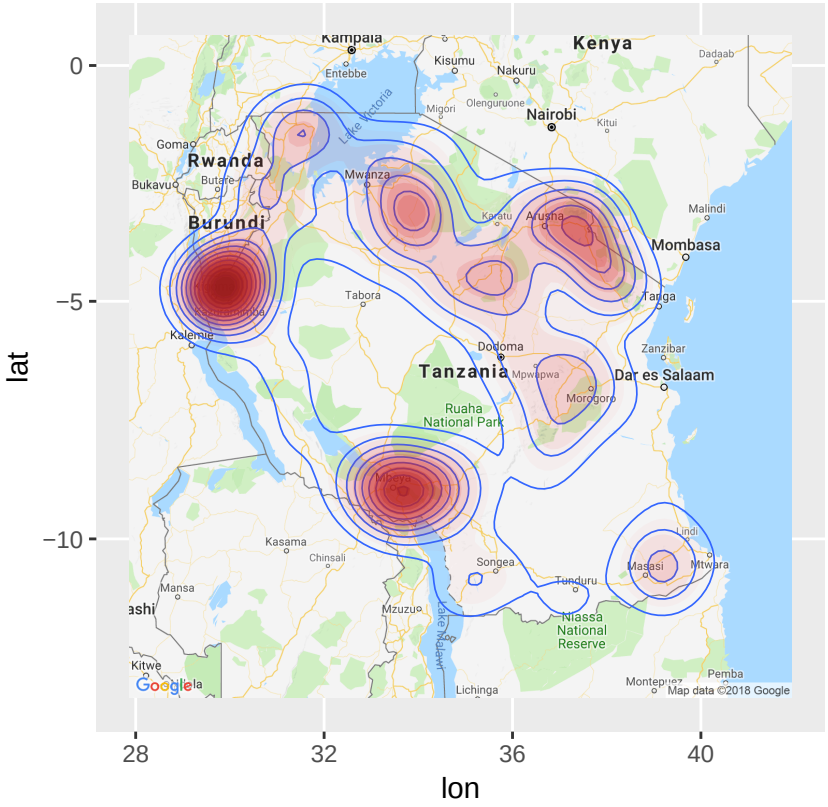
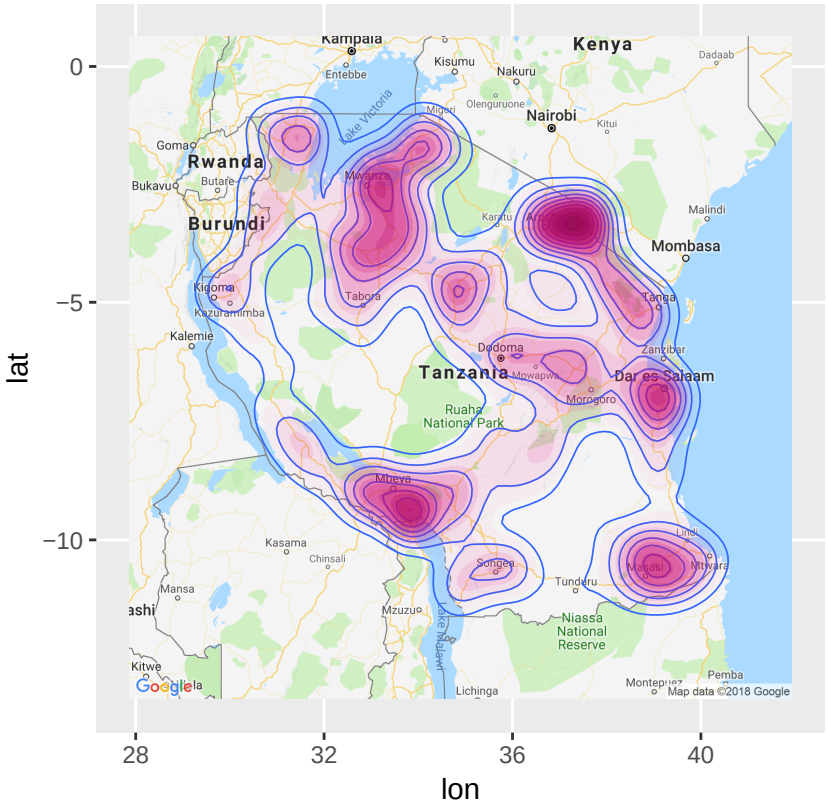


Fig 4. Non−Functional Waterpoints Map



We can notice some distinct patterns from the maps. Some areas contain more nonfunctional water points than functional. Those overlap especially with the southern coast and the area between the lakes Rukwa and Tanganyika.

As Fig. 3 and 4 suggest, the concentration for water pumps that are functional needs repair and non functional are more spread out throughout Tanzania, especially among the rural areas compared to functional.

We also see high concentration of water points that are 'functional needs repair' in the Kigoma region. We can see this map validated, because the Tanzanian Ministry of Water reported that Kigoma city suffers from extremely poor supply of water and only serves 31% of its residents' water demand with only 5 hours of running water every day[5].

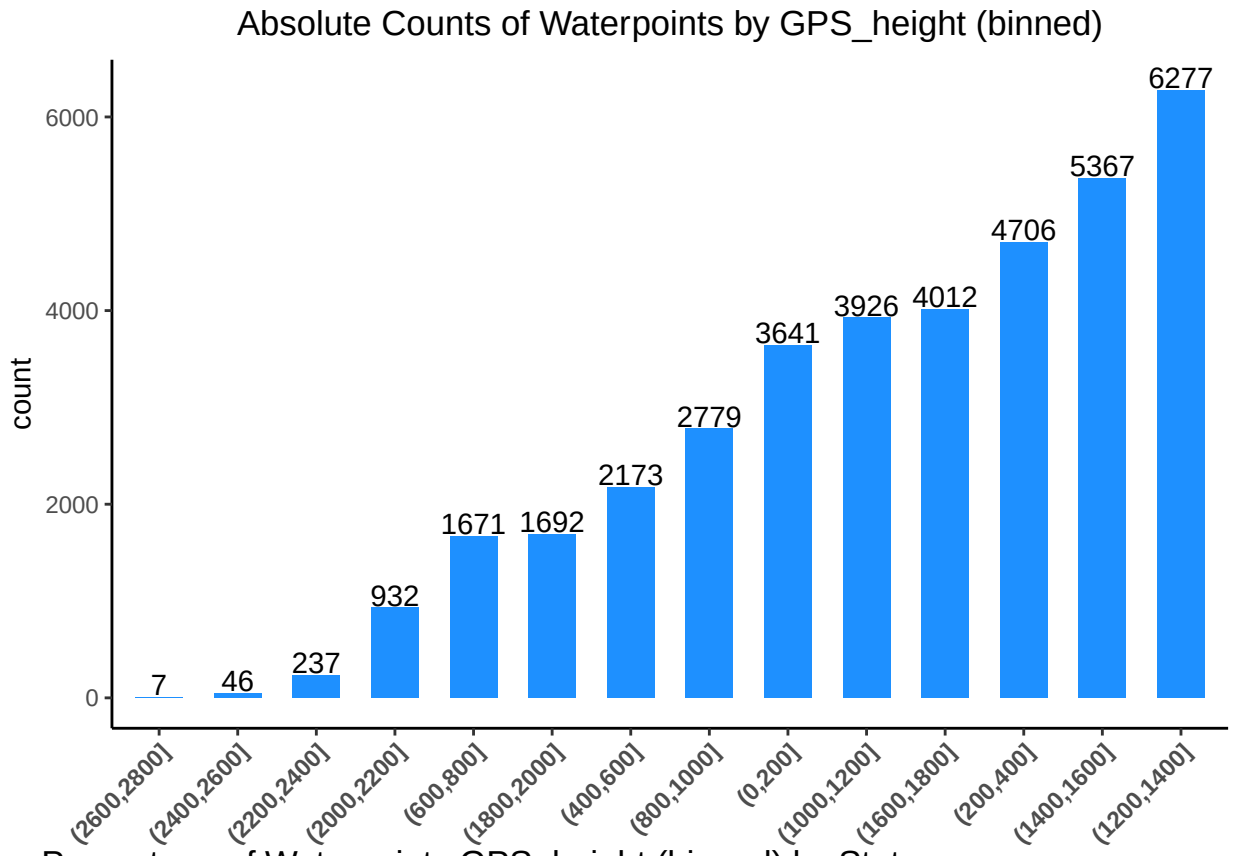
These maps suggest that water relief efforts should be concentrated in the areas where along with indadequate converage by the functional water pumps there is high concentration of 'functional needs repair' and 'non-functional' waterpoints.

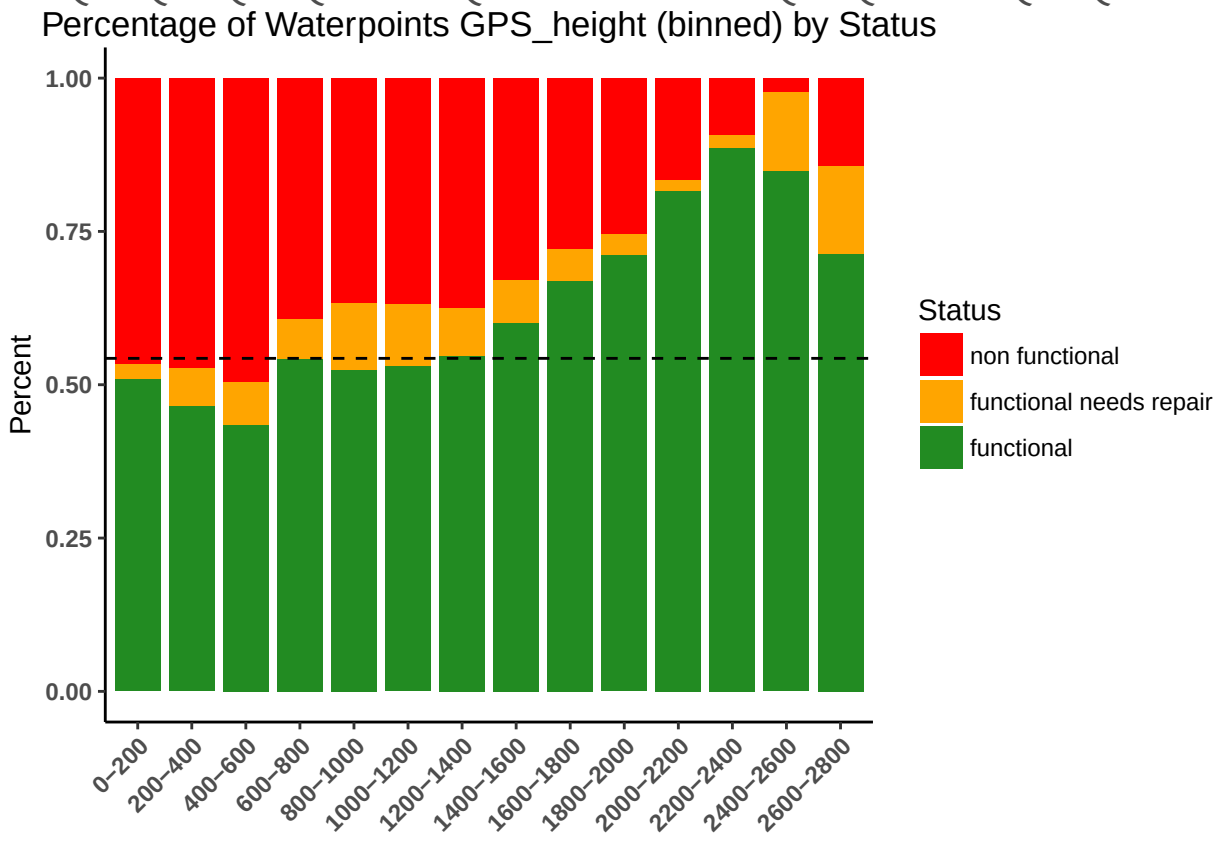
`gps_height`

The gps_height variable represents the physical altitude of a pump. A total of 20,438 data records contain a zero value for the gps_height variable. 1812 of these zero values occur when values for both latitude and longitude are apparently unknown.

Barplots for the non-zero gps_height values show that pumps located at relatively higher altitudes are more likely to be functional than are pumps found at lower altitudes. This is probably at least in part influenced by the far lower amount of water points at great heights. On the other hand, there might be confounding factors like lower usage, better climate conditions or outdoor tourism causing an increased number of functional water points.

It should also be noted that pumps located at altitudes higher than agps_height value of approximately 2500 are also the most likely to be functional needs repair, followed by those lying within the approximate range of (1000:1600).


Absolute Counts of Waterpoints by GPS_height (binned)
count
6000
4000
2000
0
7
46
237
932
1671
1692
2173
2779
3641
3926
4012
4706
5367
6277
(2600,2800]
(2400,2600]
(2200,2400]
(2000,2200]
(600,800]
(1800,2000]
(400,600]
(800,1000]
(0,200]
(1000,1200]
(1600,1800]
(200,400]
(1400,1600]
(1200,1400]

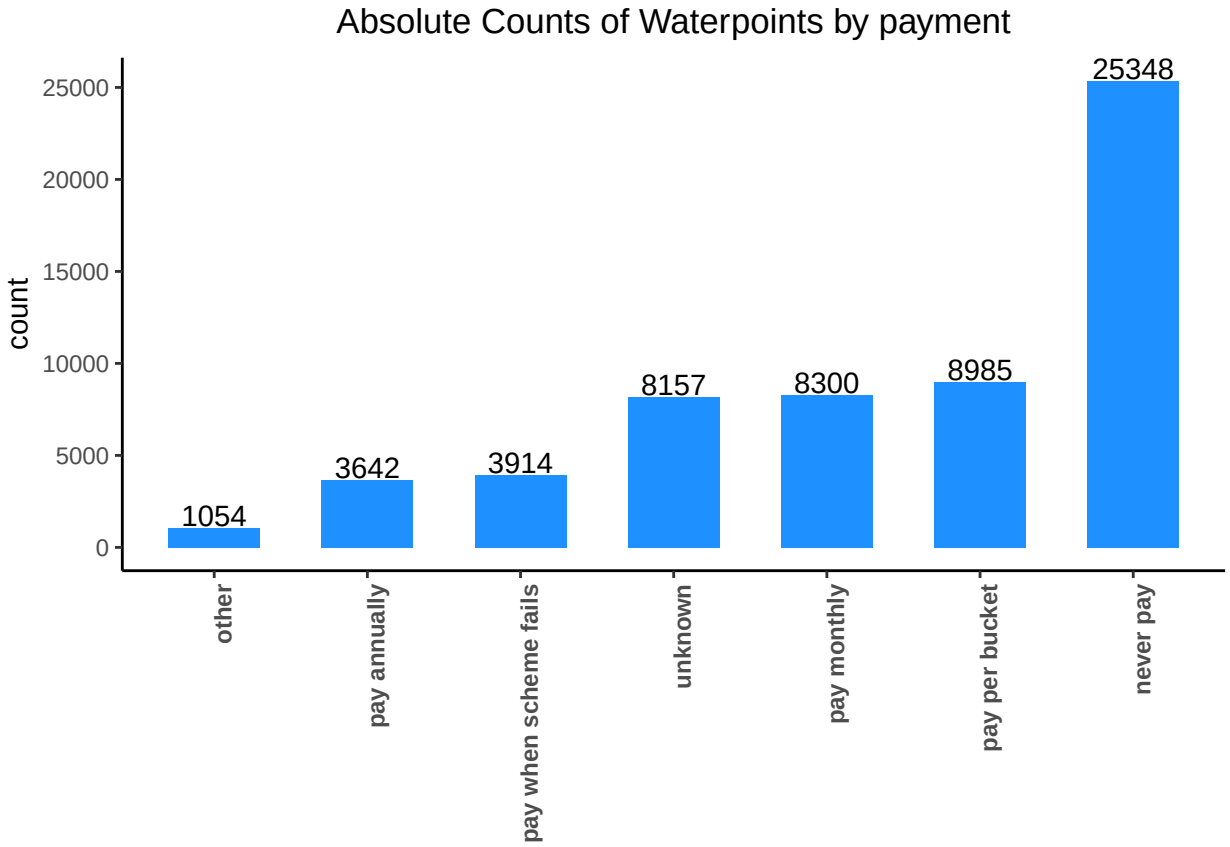

Percentage of Waterpoints GPS_height (binned) by Status
Percent
1.00
0.75
0.50
0.25
0.00
0–200
200–400
400–600
600–800
800–1000
1000–1200
1200–1400
1400–1600
1600–1800
1800–2000
2000–2200
2200–2400
2400–2600
2600–2800
Status
non functional
functional needs repair
functional

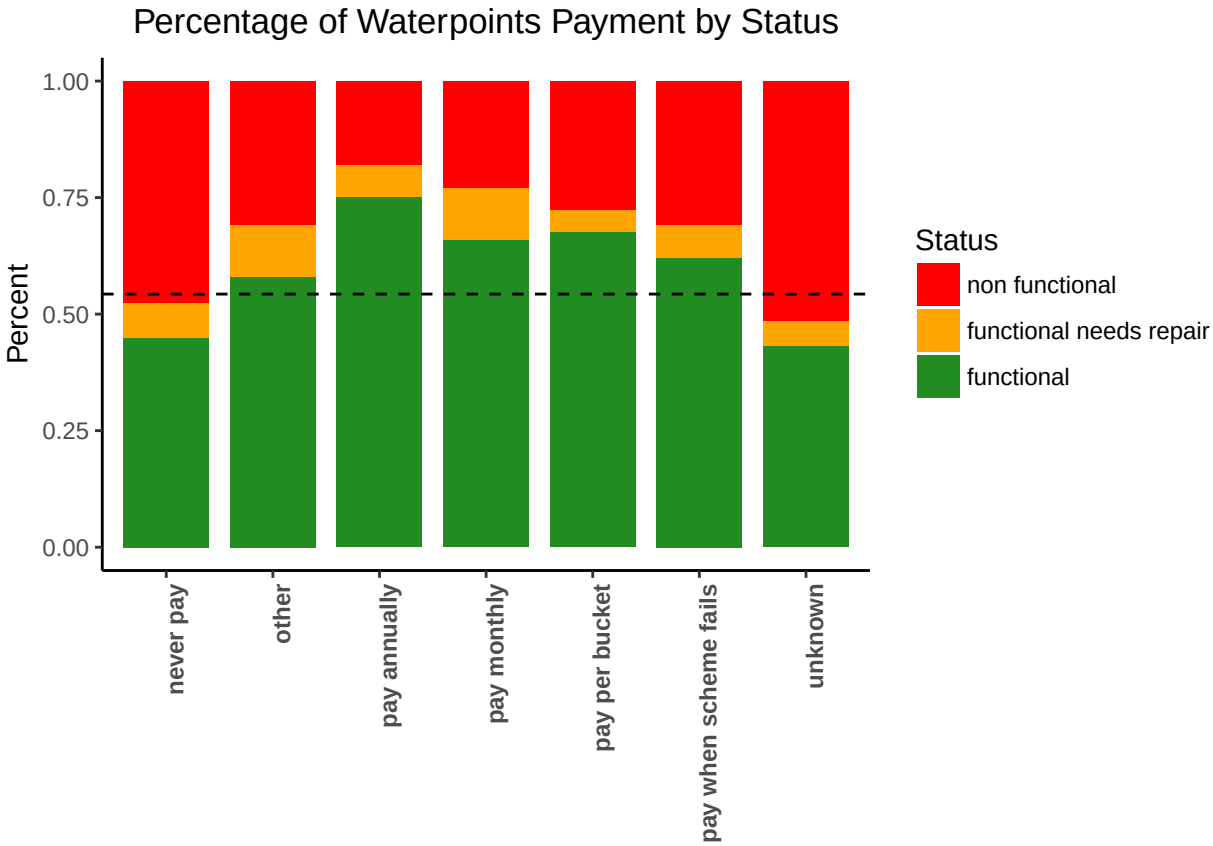
payment

The payment variable describes how water is actually paid for by users of the pump, if at all. There are 12 distinct payment values within the data set, with each record having a valid value i.e. there are no missing values. The number of pump installations per payment method is summarized in the table below.

```
## # A tibble: 7 x 2
##   payment               TotalWaterPoints
##   <chr>                            <int>
## 1 never pay                        25348
## 2 pay per bucket                    8985
## 3 pay monthly                       8300
## 4 unknown                           8157
## 5 pay when scheme fails             3914
## 6 pay annually                      3642
## 7 other                             1054
```

As shown above, more than 25,000 pumps require no payment for their use, and, unsurprisingly, as shown in the plots below, those pumps appear to be the least functional overall if unknown payment types are excluded. However, pumps that do not require payment may be located in remote areas where collection of payment is not feasible. Nevertheless, it appears reasonable to conclude that requiring users to pay for use of a pump is more likely to result in a pump remaining functional as opposed to a pump for which no payment is required.



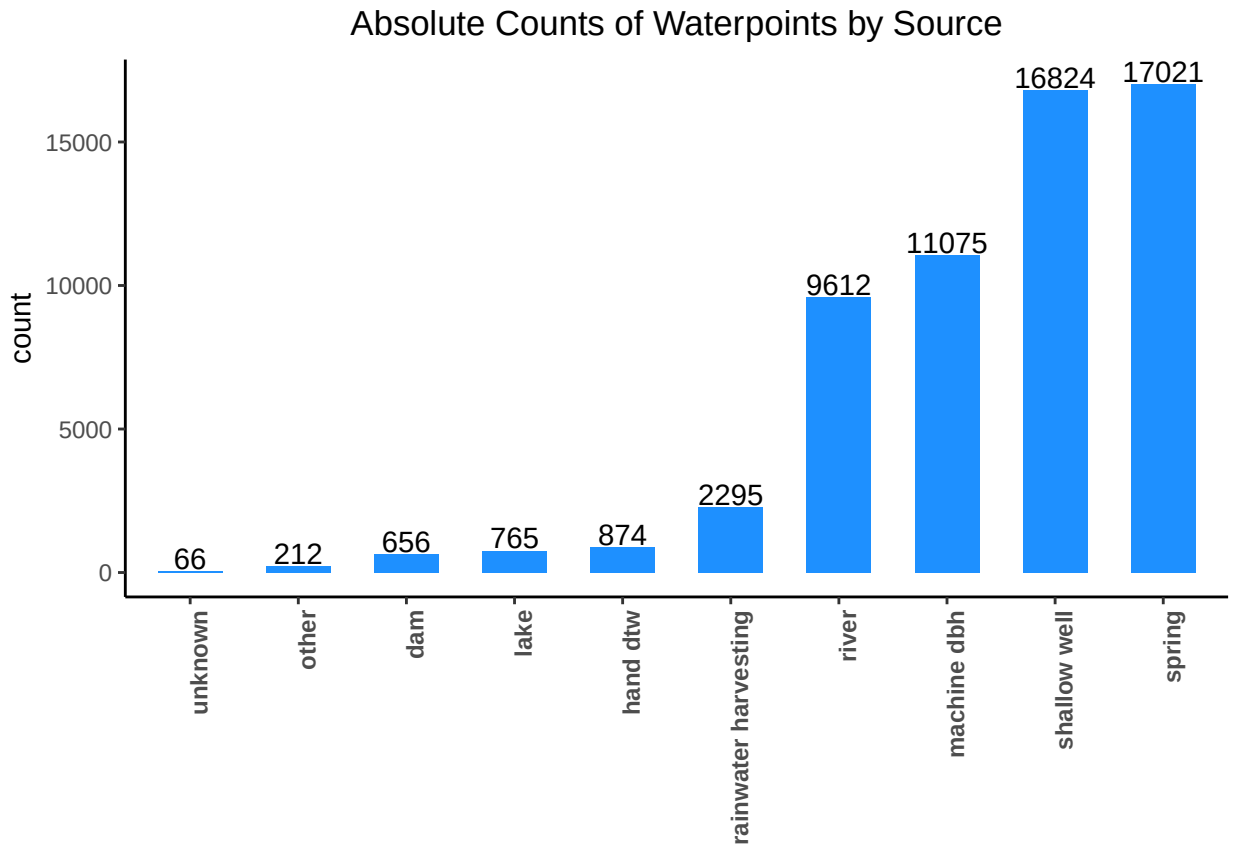


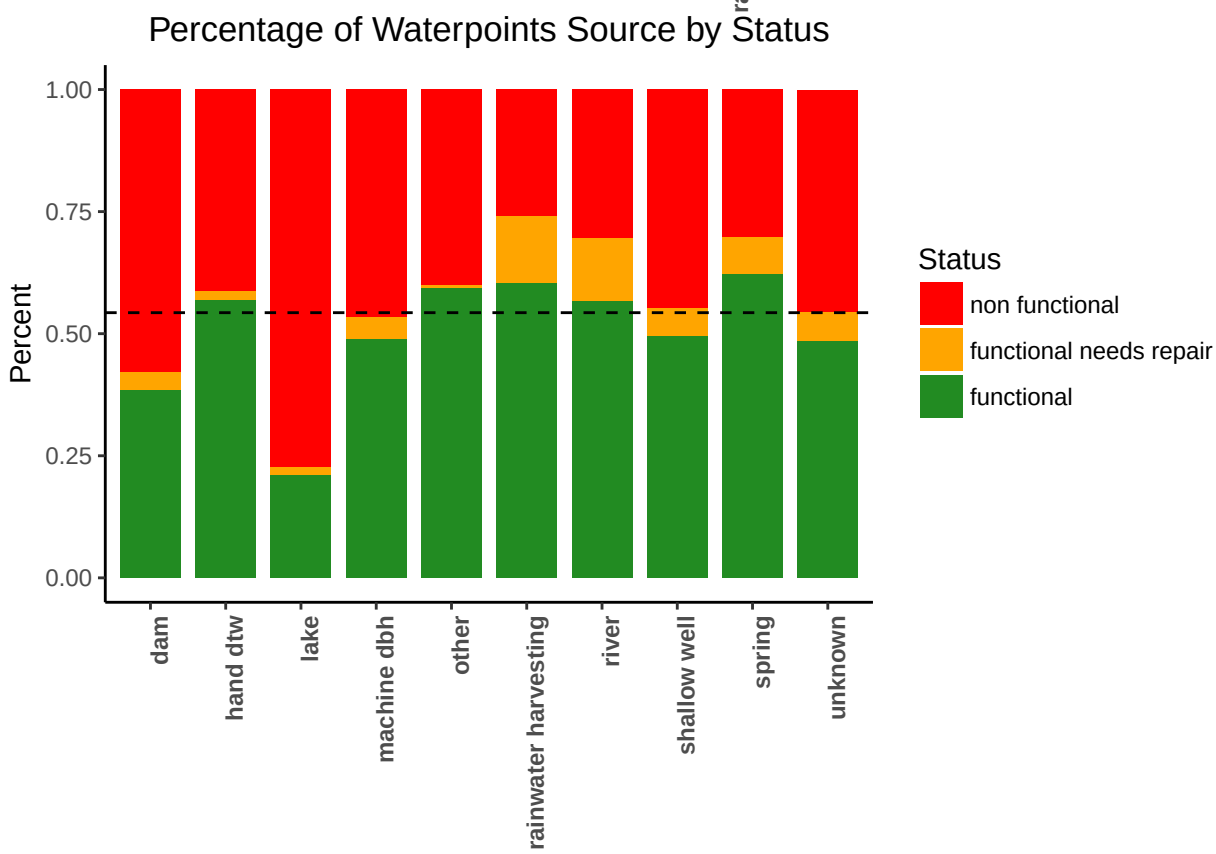
source

The source variable indicates the source of the water for a given pump. The summary statistics shown below indicate a total of 10 distinct source values within the data set, with each record having a valid value i.e. there no missing values).

```
## # A tibble: 10 x 2
##    source               TotalWaterPoints
##    <chr>                           <int>
##  1 spring                          17021
##  2 shallow well                    16824
##  3 machine dbh                     11075
##  4 river                            9612
##  5 rainwater harvesting             2295
##  6 hand dtw                          874
##  7 lake                              765
##  8 dam                               656
##  9 other                             212
## 10 unknown                            66
```

As shown below, the spring category of the source variable offers the highest percentage of functional pumps and also represents the largest source category with 17,021 pumps. The next most common source category shallow well by contrast, doesn't perform so well. While most of the lake category waterpoints are it represents only 765 of the 59,400 pumps represented in the data set. Of the other categories represented, 57.8% of are also not functioning.





Additional features like `source_type_group` which is a composite version of `source_type` and

`source_type_class` which is a composite version of `source_type_group` (and thus a super composite version of `source_type`) are also available in the dataset. As the table below indicates both of these features can help us identify what broad categories do the specific sources fall into.

Type	Group	Class
Spring	Spring	Ground water
Machine dbh	Borehole	Ground water
Hand dtw	Borehole	Ground water
Shallow well	Shallow well	Ground water
Rainwater Harvesting	Rainwater Harvesting	Surface
River/Lake	River/Lake	Surface
Dam	Dam	Surface
Other	Other	Other
Unknown	Other	Other

`region_code`

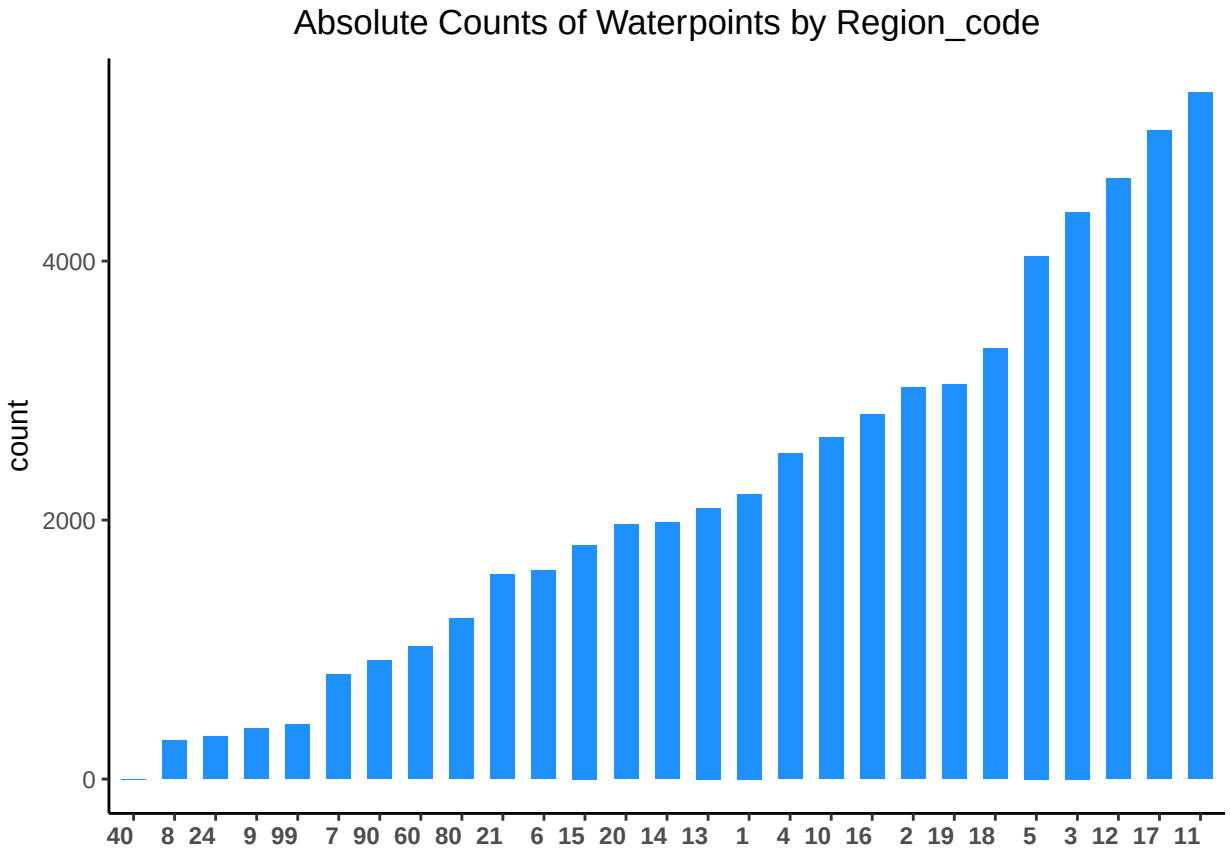
The region_code variable provides an integer code for the Tanzanian region within which a given pump is located. The summary statistics shown below indicate a total of 27 distinct region codes within the data set, with each record having a valid value. The 20 region codes having the largest number of pump installations are summarized in the table below.

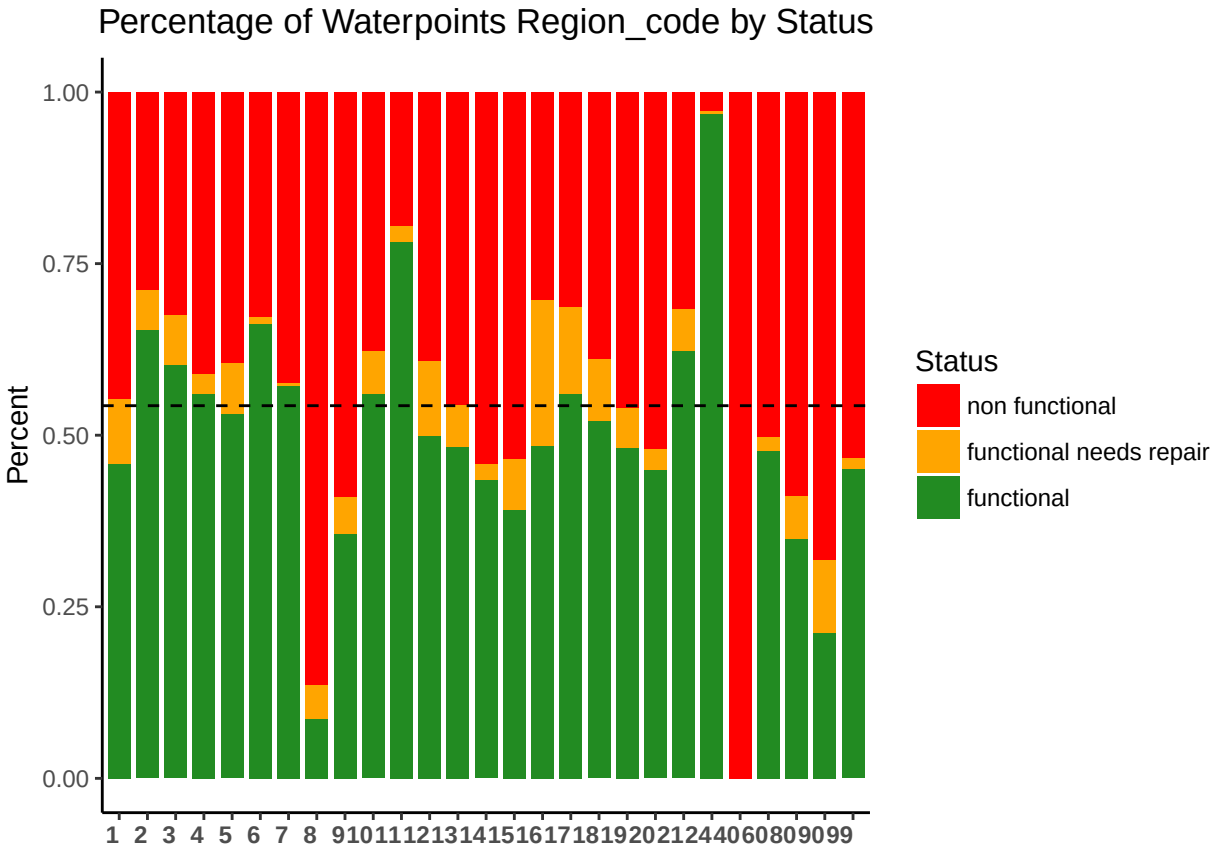
```
## # A tibble: 20 x 2
##    region_code TotalWaterPoints
##          <int>            <int>
##  1          11             5300
##  2          17             5011
##  3          12             4639
##  4           3             4379
##  5           5             4040
##  6          18             3324
##  7          19             3047
##  8           2             3024
##  9          16             2816
## 10          10             2640
## 11           4             2513
## 12           1             2201
## 13          13             2093
## 14          14             1979
## 15          20             1969
## 16          15             1808
## 17           6             1609
## 18          21             1583
## 19          80             1238
## 20          60             1025
```

It is unclear why there are 27 region_code values when there are only 20 possible values for the region variable. It is possibile that some of the region codes have either simply been entered incorrectly or were deliberately entered incorrectly due to uncertainty over the correct region_code value to apply to a given record.

Plotting the status of pumps relative to each value of the region_code variable shows even greater disparities between geographic regions for functional pumps, from a low of 8.7% in Region 8 to nearly 97% in Region 24.

Region 16 shows an unusually large 21.4% of its pumps having a status of “functional needs repair” while Region 40’s single pump (100% of the pumps in that region) is not functioning.





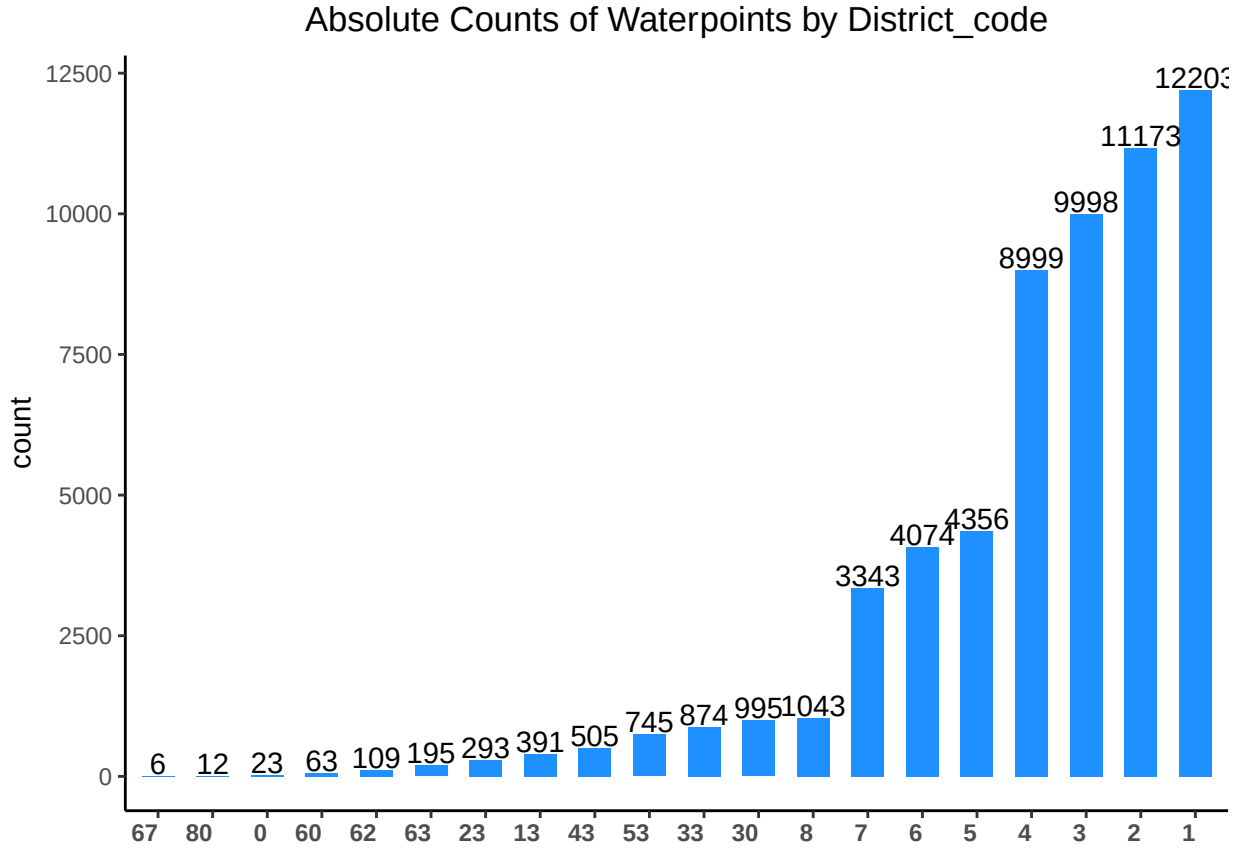
district_code

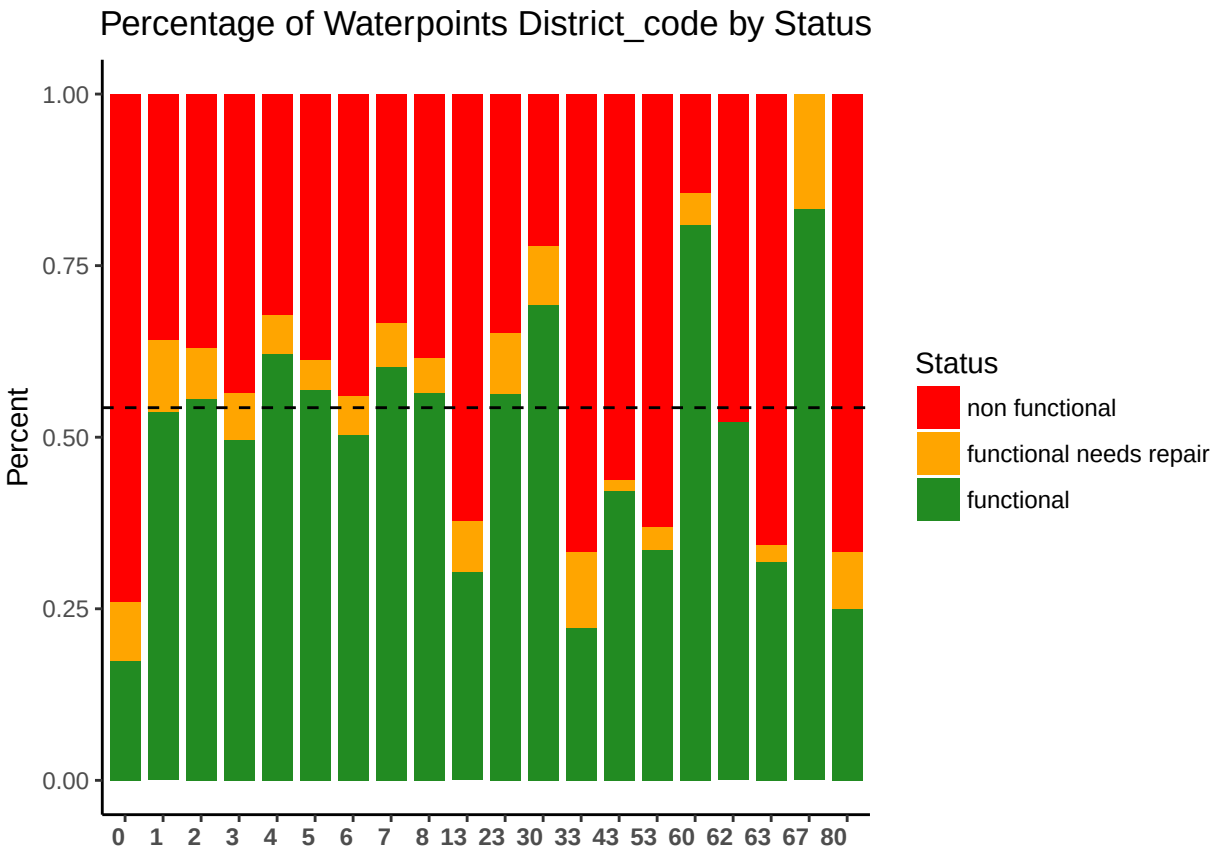
The district_code variable provides an integer coding of the Tanzanian district within which a given pump is located. The summary statistics shown below indicate a total of 20 distinct district codes within the data set, with each record having a valid value. The number of pumps per district code is summarized in the table below.

```
## # A tibble: 20 x 2
##    district_code TotalWaterPoints
##            <int>            <int>
##  1             1            12203
##  2             2            11173
##  3             3             9998
##  4             4             8999
##  5             5             4356
##  6             6             4074
##  7             7             3343
##  8             8             1043
##  9            30              995
## 10            33              874
## 11            53              745
## 12            43              505
## 13            13              391
## 14            23              293
## 15            63              195
## 16            62              109
## 17            60               63
## 18             0               23
```

```
## 19              80                12
## 20              67                 6
```

The district_code variable shows a wide disparity in the percentage of functional pumps, ranging from a low of 17.4% for district 0 to a hight of 83.3% in district 67. Only 9 of the 20 districts exceed the overall 54.3% functional pump metric. Furthermore we see that six districts' "non functional" pump percentages exceed 60%.

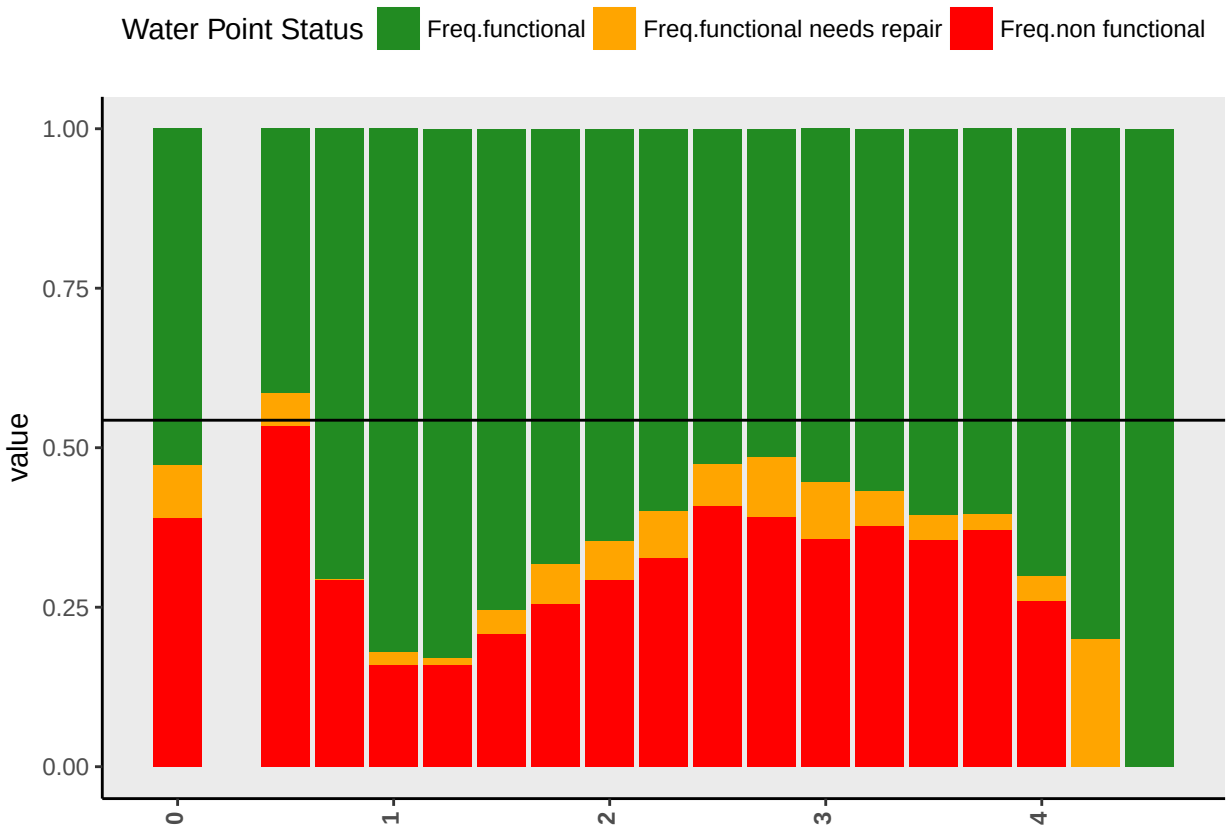





```
population
```

```
## Using label as id variables
```

```
## Warning: Removed 3 rows containing missing values (position_stack).
```



The population variable represents the human population in the area surrounding a pump. In this dataset, 21,381 data records have missing values for the population variable and these missing values can be imputed by getting the population data from the Tanzanian National Bureau of Statistics[5].

Some insights about the effect of population on status of the waterpoint:

- Most of the pumps are located in the vicinity of small villages of less than 500 people.
- For waterpoints with less than 1000 people around them there doesn't seem to be an obvious relationship between the population and the functioning status of the pump.
- In case of larger populations (> 1000) we notice that pumps located in larger population areas are somewhat more likely to have a status of `functional needs repair` than pumps located in less densely populated areas. This might be related to the high frequency usage of the pumps.
- Interestingly, pumps located in relatively higher population areas appear to be somewhat less likely to be non functional than are pumps located in areas with relatively smaller populations. Maybe large number of people leads to greater amount of collective interest in maintaing the pump.
- Finally, very high numbers of people living around a water point give apparent high frequencies of functional water points however high population waterpoints are very few and there might be confounding factors here.

Least Predictive Features

The `subvillage` variable and many of the binned `ward` and `lga` variables were found to have relatively low predictive strength and the `region` variable was found to add no value if the region_code variable was already included within a model.

References

1. Dräbing: https://github.com/tdraebing/Exploring-waterpoint-conditions-in-Tanzania
2. Topor et. al: https://rstudio-pubs-static.s3.amazonaws.com/339668_006f4906390e41cea23b3b786cc0230a.html
3. Agrawal et. al: https://www.andrew.cmu.edu/user/kdagrawa/documents/waterpump.pdf
4. Benoot: https://lib.ugent.be/fulltxt/RUG01/002/350/680/RUG01-002350680_2017_0001_AC.pdf
5. Water Sector Status Report 2009. Tanzania Ministry of Water and Irrigation. https://www.kfw-entwicklungsbank.de/migration/Entwicklungsbank-Startseite/DevelopmentFinance/About-Us/Local-Offices/Sub-Saharan-Africa/Office-Tanzania/Activities-inTanzania/Water-Sector-Status-Report-2009.pdf
6. Population Distribution of Tanzania Regions by District, Ward and Village/Mtaa. National Bureau of Statistics. http://digitallibrary.ihi.or.tz/2168/