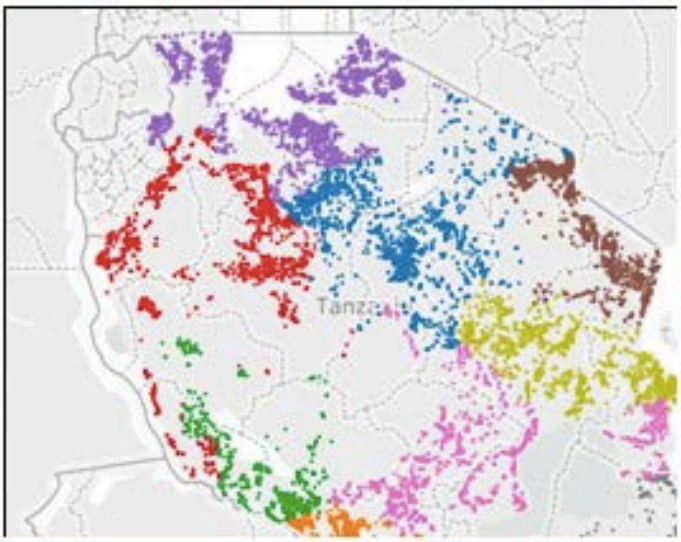
Pump
It
Up



Predicting The Operating Condition of Tanzanian Wells

Technical Analysis Supplement

Tanzanian Water Ministry

November 15, 2016

Contents

Appendix A - Description of the Data

The data to be used for this project is obtained from the *DrivenData* competition website *Pump it Up: Data Mining the Water Table* (https://www.drivendata.org/competitions/7/). Each of the 59,400 records has a unique identification number, 39 attributes to serve as predictor variables, and a response variable.

Data	Description
Data Files	2
Observations	59,400
Fields	32
Categorical Predictors	29
Numeric Predictors	6
Date Value predictor	1
Response Variable Predictor	1 with three levels

Predictor Variables:

Categorical Predictors

The data set provided contains 32 categorical attribute fields. The categorical fields include information such as geographic classifications; funder, installer, and management of the water pump; extraction type; payment type; water quality; and water quantity. Each field is reviewed as a part of the initial data quality check. Based on the review, three fields are dropped from the analysis – one field has the same value for each record, and two fields are redundant, such that the values they contain are identical to data values provided in other fields. This leaves 29 potential categorical predictors. A review of data values for each predictor indicates that both missing data and erroneous data will pose a challenge in this analysis.

Numeric Predictors

The available data set includes six numeric predictors. Predictors include Water Amount, GPS Height, Latitude, Longitude, Population, and Construction Year. Based on our review, missing values will also present a challenge in leveraging these variables in modeling. It is hypothesized that the fact that information is missing may be predictive.

Date Value Predictor

One attribute, Date Recorded, is a date value. Dates represent the years 2002, 2004, 2011, 2012, and 2013. Only one record is from 2002, which may suggest an error.

Response Variable:

The response variable is a categorical variable with three levels. Over half of the records indicate that the water pump is *functional*, with 38% identified as *non-functional* and 7% identified as *functional needs repair.*

Appendix B - Overview of the Data

As an initial step in preparing for model construction we study the data available for analysis to understand whether or not each field may offer predictive value, whether there are missing or erroneous values, and whether transformations may be appropriate. We begin with a review of categorical predictors.

Categorical Predictors

The original data source contains 32 categorical attribute fields. Each field is reviewed below. Based on the review, three fields should be omitted from the analysis: Reported By has the same value for each record; Payment Type is identical to Payment; Quantity Group is identical to Quantity. This leaves 29 potential categorical predictors. Additionally, four categories of attributes are provided in varying levels of granularity (Extraction Type Class, Extraction Type Group, and Extraction Type represent the same attribute; Water Quality and Quality Group represent the same attribute; Source, Source Type, and Source Class represent the same attribute; Waterpoint Type and Waterpoint Type Group represent the same attribute). We plan to identify a single predictor from each of these categories that performs best in each model

Funder - Who funded the well: 1,896 levels; 4,412 blank or 0 (missing) values; *Government of Tanzania* is the most common source of funding (9,084 records). Both missing data and high cardinality must be addressed in order to include this predictor.

Installer - Organization that installed the well: 2,143 valid levels; 4,435 blank, 0, or – (missing) values; DWE is the most common organization (17,405 records). Similar to Funder, missing data and high cardinality must be addressed in order to include this predictor.

Wpt Name - Name of the waterpoint if there is one: 37,398 records have names; 3,565 records have the value *none*. We view the name of the waterpoint as of little predictive value, so this field will be excluded from analysis.

Num Private – no description available: 58,643 records have a value of *0*; 64 other levels are represented. With 99% of the records having the same value, we also view this field as having little predictive value and will exclude from analysis.

Basin - Geographic water basin: 9 classes; no missing data; relatively balanced distribution.

Basin	Count	%
Internal	7,785	13%
Lake Nyasa	5,085	9%
Lake Rukwa	2,454	4%
Lake Tanganyika	6,432	11%
Lake Victoria	10,248	17%
Pangani	8,940	15%
Rufiji	7,976	13%
Ruvuma / Southern Coast	4,493	8%
Wami / Ruvu	5,987	10%
Grand Total	**59,400**	**100%**

Region - Geographic location: 21 classes; no missing data; Iringa has the largest proportion of records; Dar es Salaam the smallest.

Region	Count	%
Arusha	3,350	6%
Dar es Salaam	805	1%
Dodoma	2,201	4%
Iringa	5,294	9%
Kagera	3,316	6%
Kigoma	2,816	5%
Kilimanjaro	4,379	7%
Lindi	1,546	3%
Manyara	1,583	3%
Mara	1,969	3%
Mbeya	4,639	8%
Morogoro	4,006	7%
Mtwara	1,730	3%
Mwanza	3,102	5%
Pwani	2,635	4%
Rukwa	1,808	3%
Ruvuma	2,640	4%
Shinyanga	4,982	8%
Singida	2,093	4%
Tabora	1,959	3%
Tanga	2,547	4%
Grand Total	**59,400**	**100%**

Region Code - Geographic location (coded): 27 classes; no missing data. It appears that some of the codes are errors (for example, Mtwara is mapped to Region Code 9, 90, and 99). With this observation we view this field as one that should be excluded from analysis.

District Code - Geographic location (coded): 20 levels; 23 missing values; most risks are in 4 districts. In preparation for modeling we will impute missing data and combine smaller volume district codes.

District Code	Count	%
0	23	0%
1	12,203	21%
2	11,173	19%
3	9,998	17%
4	8,999	15%
5	4,356	7%
6	4,074	7%
7	3,343	6%
8	1,043	2%
13	391	1%
23	293	0%
30	995	2%
33	874	1%
43	505	1%
53	745	1%
60	63	0%
62	109	0%
63	195	0%
67	6	0%
80	12	0%
Grand Total	**59,400**	**100%**

Lga - Geographic location: 125 levels for Tanzanian districts. High cardinality will need to be addressed to include in modeling.

Ward - Geographic location: 2,092 levels; no missing values. High cardinality will need to be addressed to include in modeling.

Subvillage - Geographic location: 19,287 levels; 371 missing values. With so many levels, it is unlikely that this field will be useful for modeling.

Public Meeting - True/False predictor with 3,334 missing values. Imputation will be required.

Public Meeting	Count	%
TRUE	51,011	86%
FALSE	5,055	9%
MISSING	3,334	6%
Grand Total	**59,400**	**100%**

Recorded_by - Group entering this row of data: This is the same value for all records. It will not be included in modeling.

Scheme Management - Who operates the waterpoint (management); 12 levels; VWC is most common operator; 3,877 missing values. Imputation will be required.

Scheme Management	Count	%
Company	1,061	2%
None	1	0%
Other	766	1%
Parastatal	1,680	3%
Private operator	1,063	2%
SWC	97	0%
Trust	72	0%
VWC	36,793	62%
Water authority	3,153	5%
Water Board	2,748	5%
WUA	2,883	5%
WUG	5,206	9%
Missing	3,877	7%
Grand Total	**59,400**	**100%**

Scheme Name - Who operates the waterpoint (name): 2,696 levels; 28,166 missing values; based on visual inspection there appear to be data-entry errors. With the extent of missing data and errors this field will not be included in modeling.

Permit - If the waterpoint is permitted: TRUE/FALSE predictor with 3,056 missing values. Missing values will be imputed.

Permit	Count	%
TRUE	38,852	65%
FALSE	17,492	29%
MISSING	3,056	5%
Grand Total	**59,400**	**100%**

Extraction Type - The kind of extraction the waterpoint uses: 18 levels; largest class is *gravity*.

Extraction Type	Count	%
afridev	1,770	3%
cemo	90	0%
climax	32	0%
gravity	26,780	45%
india mark ii	2,400	4%
india mark iii	98	0%
ksb	1,415	2%
mono	2,865	5%
nira/tanira	8,154	14%
other	6,430	11%
other - mkulima/shinyanga	2	0%
other - play pump	85	0%
other - rope pump	451	1%
other - swn 81	229	0%
submersible	4,764	8%
swn 80	3,670	6%
walimi	48	0%
windmill	117	0%
Grand Total	**59,400**	**100%**

Extraction Type Group - The kind of extraction the waterpoint uses: 13 levels; combines some of the Extraction Types (e.g. *other motorpump = cemo + climax*).

Extraction Type Group	Count	%
afridev	1,770	3%
gravity	26,780	45%
india mark ii	2,400	4%
india mark iii	98	0%
mono	2,865	5%
nira/tanira	8,154	14%
other	6,430	11%
other handpump	364	1%
other motorpump	122	0%
rope pump	451	1%
submersible	6,179	10%
swn 80	3,670	6%
wind-powered	117	0%
Grand Total	**59,400**	100%

Extraction Type Class - The kind of extraction the waterpoint uses; 7 levels that combine levels of Extraction Type Group (e.g. *motorpump = mono and other motorpump*)

Extraction Type Class	Count	%
gravity	26,780	45%
handpump	16,456	28%
motorpump	2,987	5%
other	6,430	11%
rope pump	451	1%
submersible	6,179	10%
wind-powered	117	0%
Grand Total	**59,400**	**100%**

Management - How the waterpoint is managed: 12 levels; mostly *vwc*; no missing data, but 561 unknown.

Management	Count	%
company	685	1%
other	844	1%
other - school	99	0%
parastatal	1,768	3%
private operator	1,971	3%
trust	78	0%
unknown	561	1%
vwc	40,507	68%
water authority	904	2%
water board	2,933	5%
wua	2,535	4%
wug	6,515	11%
Grand Total	**59,400**	**100%**

Management Group - How the waterpoint is managed: 5 levels; no missing data, but 561 unknown; combines levels of Management (e.g. *user-group = vwc+wua+wug+water board*).

Management	Count	%
commercial	3,638	6%
other	943	2%
parastatal	1,768	3%
unknown	561	1%
user-group	52,490	88%
Grand Total	**59,400**	100%

Payment - What the water costs; 7 levels; no missing data, but there is an unknown level (8,157 records); identical to Payment Type (we will only want to include one of these two variables in the modeling process); largest class is *never pay*.

Payment	Count	%
annually	3,642	6%
monthly	8,300	14%
never pay	25,348	43%
on failure	3,914	7%
other	1,054	2%
per bucket	8,985	15%
unknown	8,157	14%
Grand Total	**59,400**	**100%**

Payment Type - What the water costs: Redundant (identical to Payment except level names are slightly different).

Water Quality - The quality of the water: 8 levels; no missing data, but 1,876 in class *unknown*; mostly *soft water*.

Water Quality	Count	%
coloured	490	1%
fluoride	200	0%
fluoride abandoned	17	0%
milky	804	1%
salty	4,856	8%
salty abandoned	339	1%
soft	50,818	86%
unknown	1,876	3%
Grand Total	**59,400**	**100%**

Quality Group - The quality of the water: 6 levels; no missing data, but 1,876 in class *unknown*; mostly *good*. Note that these levels are identical to Water Quality except *good = soft*; *fluoride = fluoride + fluoride abandoned*; *salty = salty + salty abandoned.*

Quality Group	Count	%
colored	490	1%
fluoride	217	0%
good	50,818	86%
milky	804	1%
salty	5,195	9%
unknown	1,876	3%
Grand Total	**59,400**	**100%**

Quantity - The quantity of water: 5 levels; no missing values, but 789 *unknown*. Note that these values are identical to Quantity Group (need to omit one of these two predictors).

Quantity	Count	%
dry	6,246	11%
enough	33,186	56%
insufficient	15,129	25%
seasonal	4,050	7%
unknown	789	1%
Grand Total	**59,400**	**100%**

Quantity Group - The quantity of water: Identical to Quantity.

Source - The source of the water: 10 levels; no missing values, but 66 classified as *unknown.*

Source	Count	%
dam	656	1%
hand dtw	874	1%
lake	765	1%
machine dbh	11,075	19%
other	212	0%
rainwater harvesting	2,295	4%
river	9,612	16%
shallow well	16,824	28%
spring	17,021	29%
unknown	66	0%
Grand Total	**59,400**	**100%**

Source Type - The source of the water: 7 levels; no missing values; combines levels of Source (*borehole = hand dtw + machine dbw*; *other = other + unknown*; *river/lake = river + lake*).

Source Type	Count	%
borehole	11,949	20%
dam	656	1%
other	278	0%
rainwater harvesting	2,295	4%
river/lake	10,377	17%
shallow well	16,824	28%
spring	17,021	29%
Grand Total	**59,400**	**100%**

Source Class - The source of the water: 3 levels; no missing values; combines levels of Source Type (*groundwater = shallow well + borehole + spring*; *surface = dam + river / lake + rainwater harvesting*).

Source Class	Count	%
groundwater	45,794	77%
surface	13,328	22%
unknown	278	0%
Grand Total	**59,400**	**100%**

Waterpoint Type - The kind of waterpoint: 7 levels; no missing values; mostly communal standpipes.

Waterpoint Type	Count	%
cattle trough	116	0%
communal standpipe	28,522	48%
communal standpipe multiple	6,103	10%
dam	7	0%
hand pump	17,488	29%
improved spring	784	1%
other	6,380	11%
Grand Total	**59,400**	**100%**

Waterpoint Type Group - The kind of waterpoint: 6 levels; no missing values; same as Waterpoint Type except *communal standpipe* and *communal standpipe multiple* are combined.

Waterpoint Type Group	Count	%
cattle trough	116	0%
communal standpipe	34,625	58%
dam	7	0%
hand pump	17,488	29%
improved spring	784	1%
other	6,380	11%
Grand Total	**59,400**	**100%**

Date Value Predictors

One attribute, Date Recorded, is a date value. There are 356 dates from the years 2002, 2004, 2011, 2012, and 2013. Only one record is from 2002, which may suggest an error. We will extract year from this field.

Numeric Predictors

The available data includes six numeric predictors. Based on an initial review missing values will present a challenge in leveraging these variables in modeling. It is hypothesized that the fact that information is missing will be predictive of the response variable.

Amount TSH - Total static head (amount water available to waterpoint): Non-zero values range from 0.2 to 350,000 (units unknown); 41,639 records have a value of *0*, which likely indicates missing data. With 70% of the records having missing values, we will likely exclude this field from models.

GPS Height - Altitude of the well: Values range from -90 to 2,770, with 20,438 records having a value of *0*, which may indicate missing values. We will attempt imputation, however, the extent of missing data may pose a problem for this predictor as well.

Longitude - GPS coordinate: Valid values range from 29.607 to 40.345; 1,812 records have a value of *0*, which implies a missing value, based on Tanzania's geographical boundaries. We will impute missing values prior to modeling.

Latitude - GPS coordinate: Valid values range from -11.649 to -0.998; 1,812 records have a value of *0*, which implies a missing value, based on Tanzania's geographical boundaries. We will impute missing values prior to modeling.

Population - Population around the well: Values range from 1 to 30,500. 21,281 records have a value of *0*, which likely indicate missing values. We will attempt imputation, however, the extent of missing data may pose a problem for this predictor.

Construction Year - Year the waterpoint was constructed: Values range from 190 to 2013. 20,709 records have a value of *0*, which indicates missing data. Similar to GPS Height and Population, we will impute missing values, however, the predictive value may be limited.

Response Variable

The response variable is a categorical variable with three levels. Over half of the records indicate that the water pump is *functional*, with 38% identified as *non-functional* and 7% identified as *functional needs repair*.

Response Level	Count	%
functional	32,259	54%
functional needs repair	4,317	7%
non functional	22,824	38%
Grand Total	**59,400**	**100%**

Appendix C - Description of Transformation of Data

Based on our data review, we identified the need to perform three types of transformations. First, we must perform imputation for missing data. Second, we need to address the issue of high cardinality for several of the predictor variables. Finally, we need to standardize the data prior to modeling.

Data Imputation

We reviewed potential data imputation techniques and identified two viable options – MICE and missForest imputation.

MICE (multivariate imputation by chained equations) is a methodology that can handle different data types at the same time (e.g. categorical and numeric) and can be used when multiple predictor variables have missing values. The MICE algorithm applies a regression strategy to estimate missing values for one predictor at a time based on values of the other predictors. This process is repeated until convergence is achieved.

The R package "missForest" is also suitable for imputing mixed type data. This non-parametric approach uses a random forest based on actual observed values to predict the missing values. An advantage to this approach is that it can be performed using parallel processing to expedite processing time.

To determine which imputation methodology to implement, we performed both the MICE and missForest imputations on the predictor variables, trained a random forest model on each dataset, and evaluated results using a multivariate ROC technique to obtain an AUC (area under the curve) metric. Results were very similar, however, the dataset with missing variables predicted using missForest had slightly more favorable performance. As such, we continued our analysis with the dataset using the missForest imputations.

High Cardinality

We noted in our data overview that a number of potential categorical predictors had high cardinality, so many levels modeling options may not have been able to process the data. To address this challenge we applied a transformation to several of the predictors to combine levels with smaller numbers of observations. Specifically, we grouped levels with less than X% of the observations to achieve a suitable number of levels. Predictors transformed using this process included Funder, Installer, LGA, District Code, and Scheme Management.

Data Standardization

Once we addressed missing data and high cardinality, we also standardized our data, as some of the proposed modeling options are sensitive to scale. For numeric data we centered the values by subtracting the sample mean, and then we scaled the data by dividing by the standard deviation. For categorical predictors we used dummy coding to have an indicator for each level of the variable.

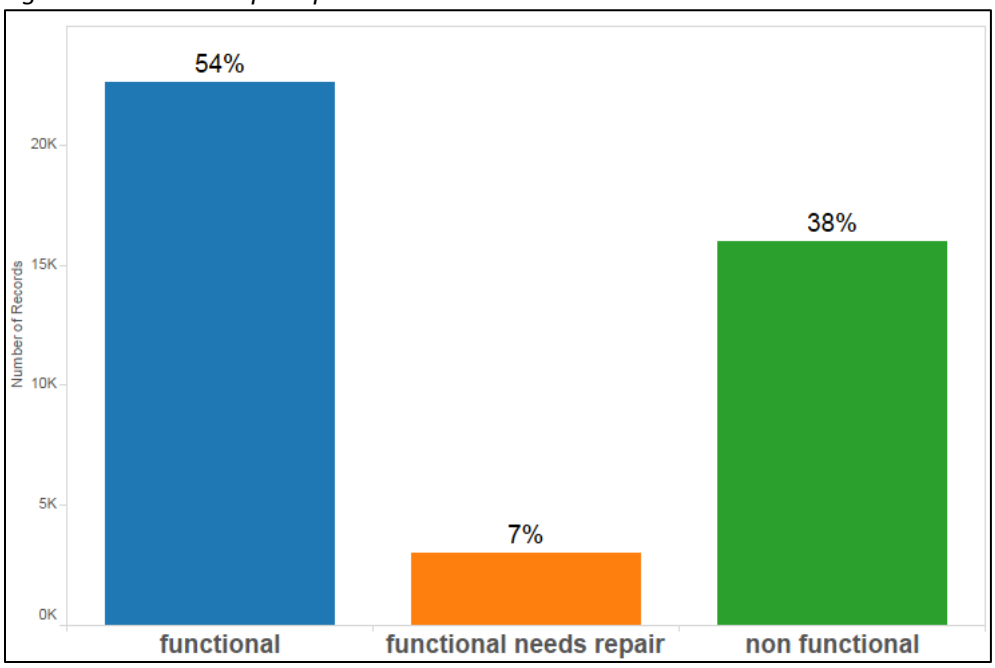
Appendix D - Exploratory Data Analysis

With the data quality check complete, we partition the data using a 70/30 split to create a training data set (41,581 records) for model construction and a test data to be used for model validation (17,819 records). We use the "set seed" command to create reproducible data sets.
For this report, the exploratory data analysis (EDA) was enhanced through the creation of an EDA dashboard using Tableau business intelligence software. The dashboard compares all predictor variables with the response variable and can be accessed and manipulated by all team members. This allows each team member to perform their own EDA and choose the variables they would like to analyze. All of the figures included in this section of the report were created using the dashboard.

Response Variable

We begin the exploratory data analysis with an assessment of the response variable *status group.* Within the training data set, 54% of the wells are considered functional, 7% functional needs repair and 38% non-functional.

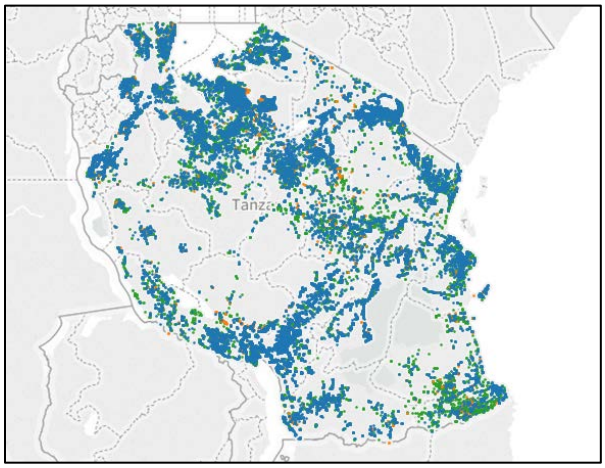
Figure 1. Status Group Proportions



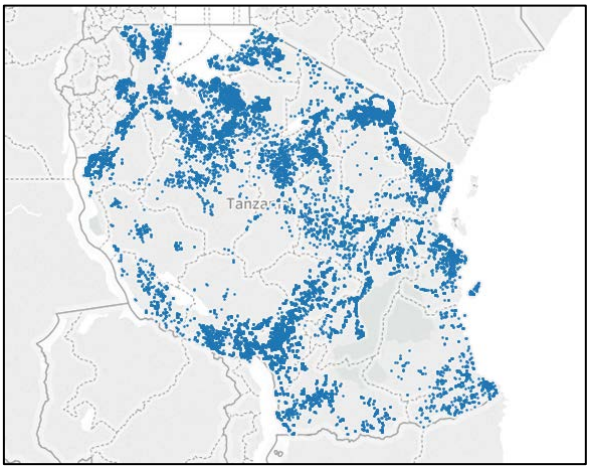
Understanding the geographic location of the wells is very important when it comes to efficiently deploying teams to fix the wells. To further the teams' understanding of the location of the wells as well as investigate the geographic distribution of the wells by *status group,* well maps were created (Figure 2). As seen in Figure 2, wells within each st*atus group* are distributed across the country.

Figure 2. Well Maps by Status Group

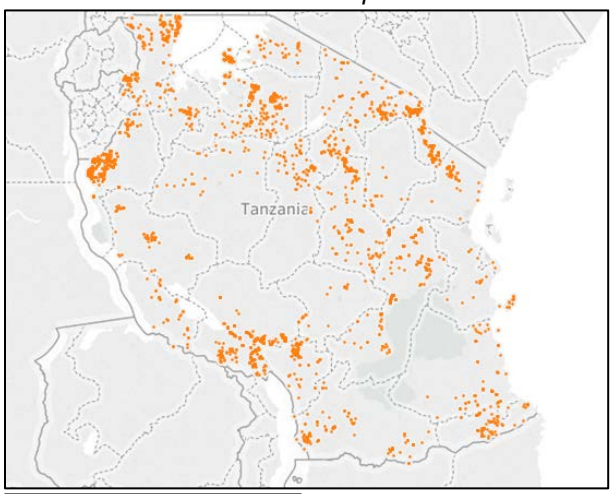
All Status Groups



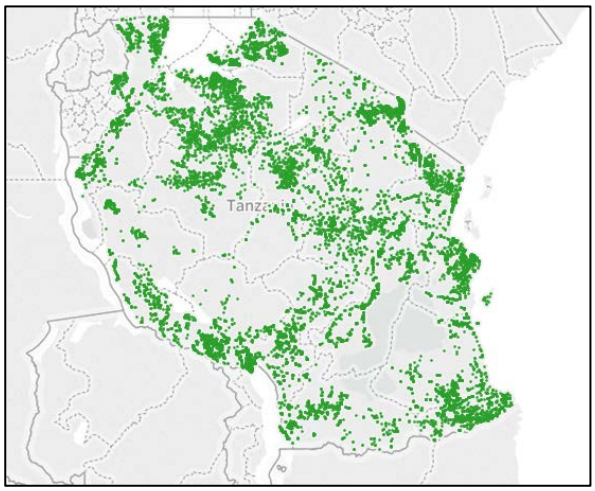
Functional Wells

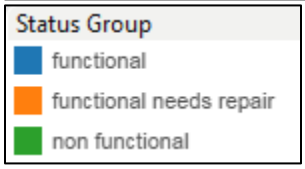


Functional Needs Repair Wells



Non Functional Wells



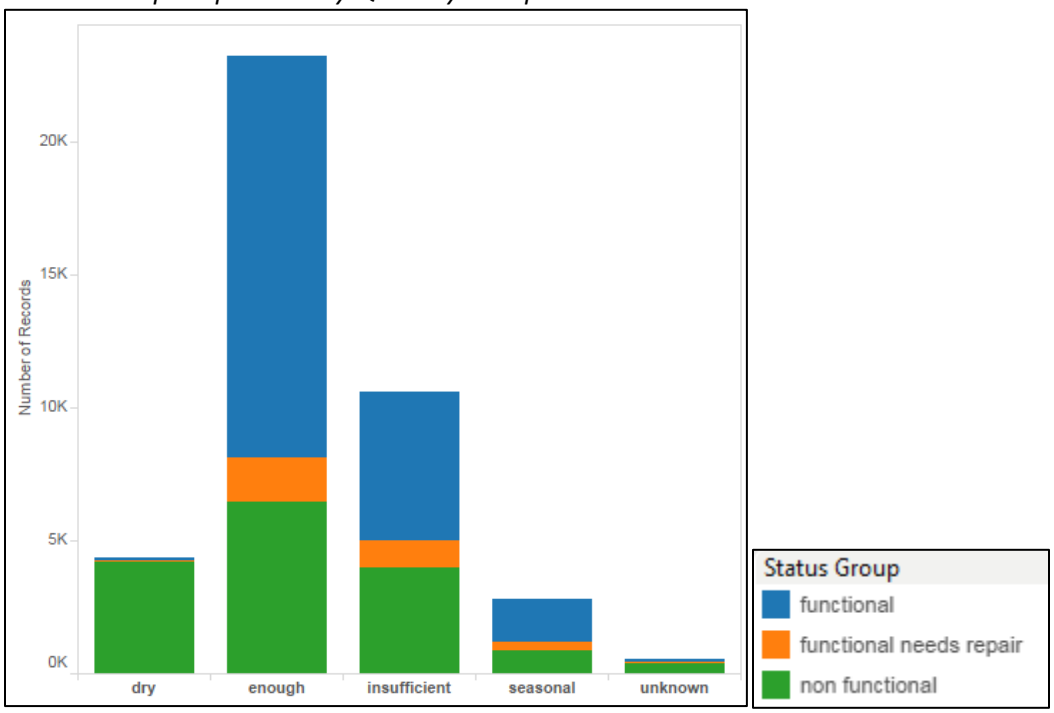


That dashboard containing the wells maps can be accessed via the link below.
https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-WellsMap/WellMap

Categorical Variables

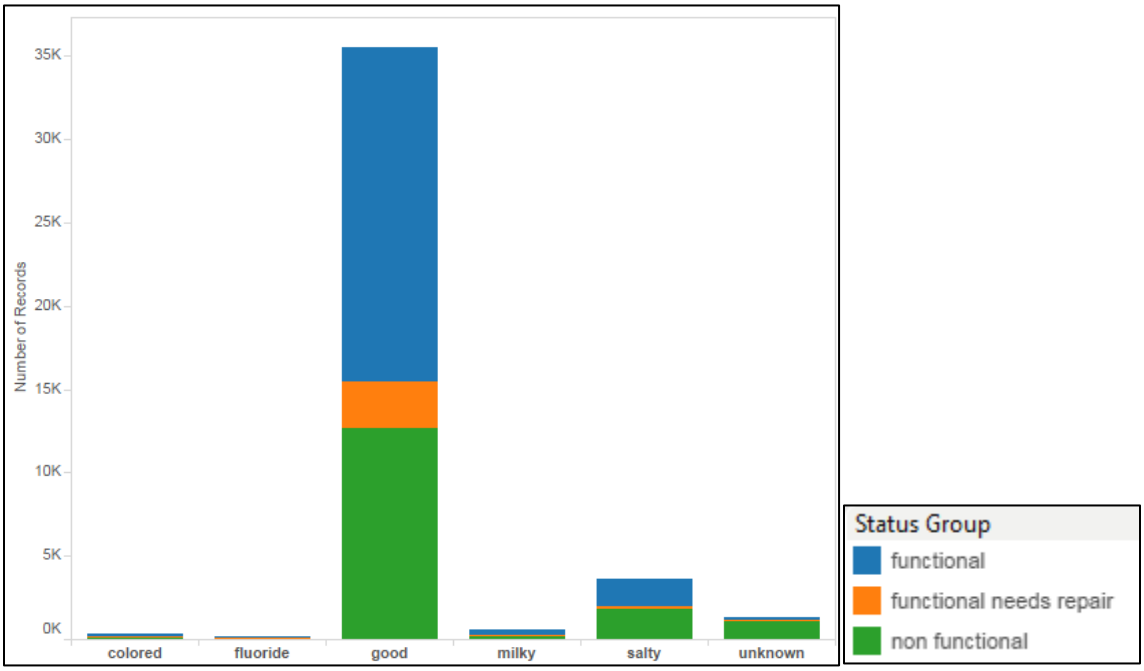
The next step in the exploratory data analysis is an assessment of how select categorical predictors interact with the response variable *status group.* The dashboard with the categorical variables analysis can be accessed via the link below.
https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-InteractiveEDA/InteractiveEDA
The first predictor we review is *quantity group*, which classifies available water amount. We observe an intuitive pattern showing that the "dry" level has predominantly "non-functional" water pumps; the "enough" level has the greatest proportion of "functional" pumps (Figure 3). Based on this exhibit, the distribution of *status group* varies meaningfully by level of *quantity group*

Figure 3. Status Group Proportions by Quantity Group Level



The distribution of the predictor *quality group* demonstrates that the most common water quality is "good" followed by "salty." The distribution of water pumps by *status group* within these classes suggests that the greatest proportion of water pumps within the "good" class are "functional," while the majority of water pumps in the "unknown" class are "non-functional" (Figure 4).

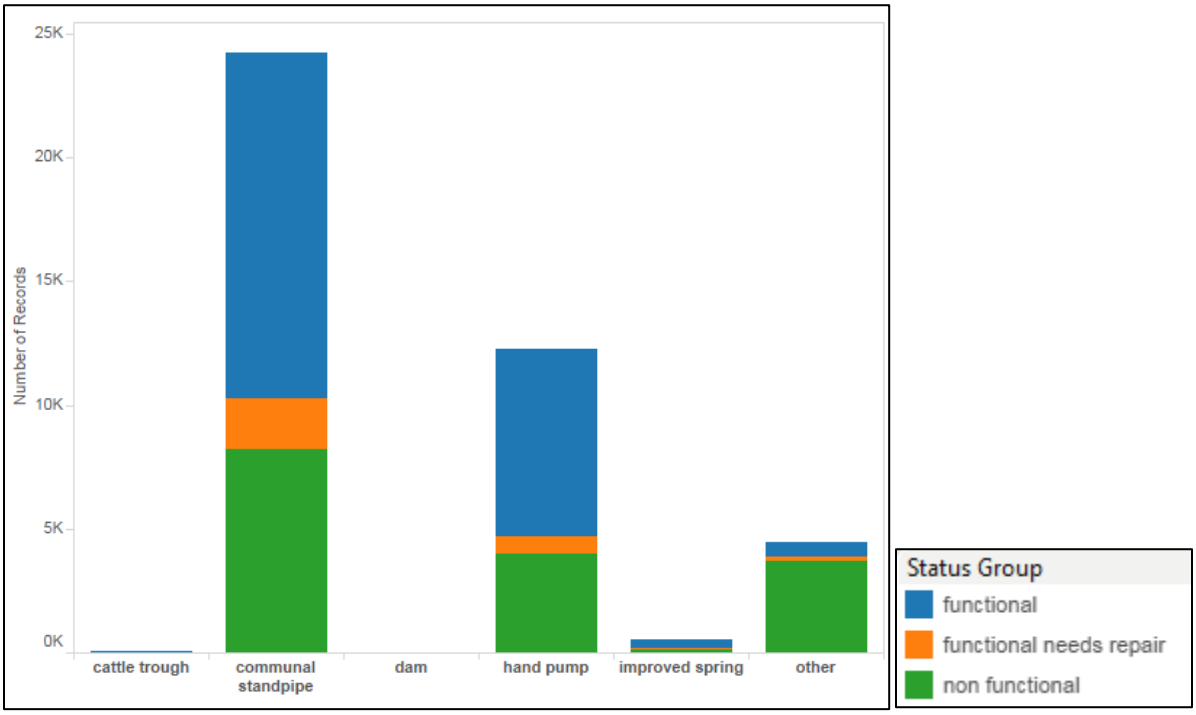
Figure 4. Status Group Proportions by Quality Group



An analysis of the predictor *waterpoint type group* reveals relatively similar proportions of water pump

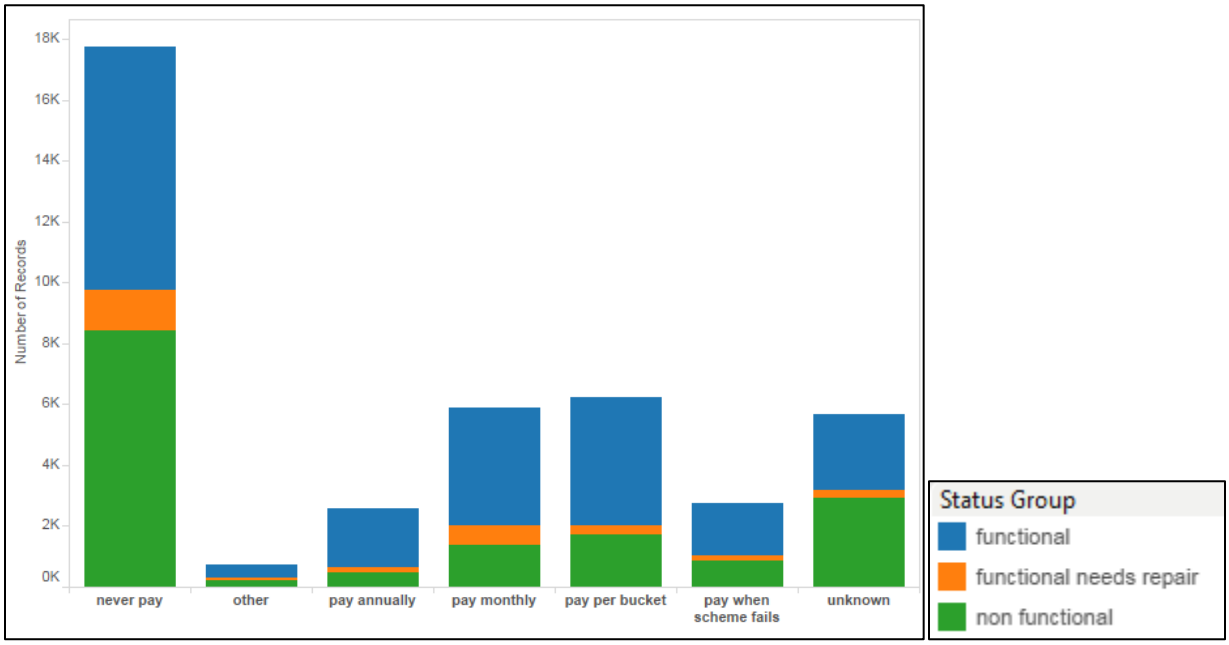
status group for “communal standpipe” and “hand pump,” while “other” types of water pumps are primarily “non-functional” (Figure 5).

Figure 5. Status Group Proportions by Waterpoint Type Group



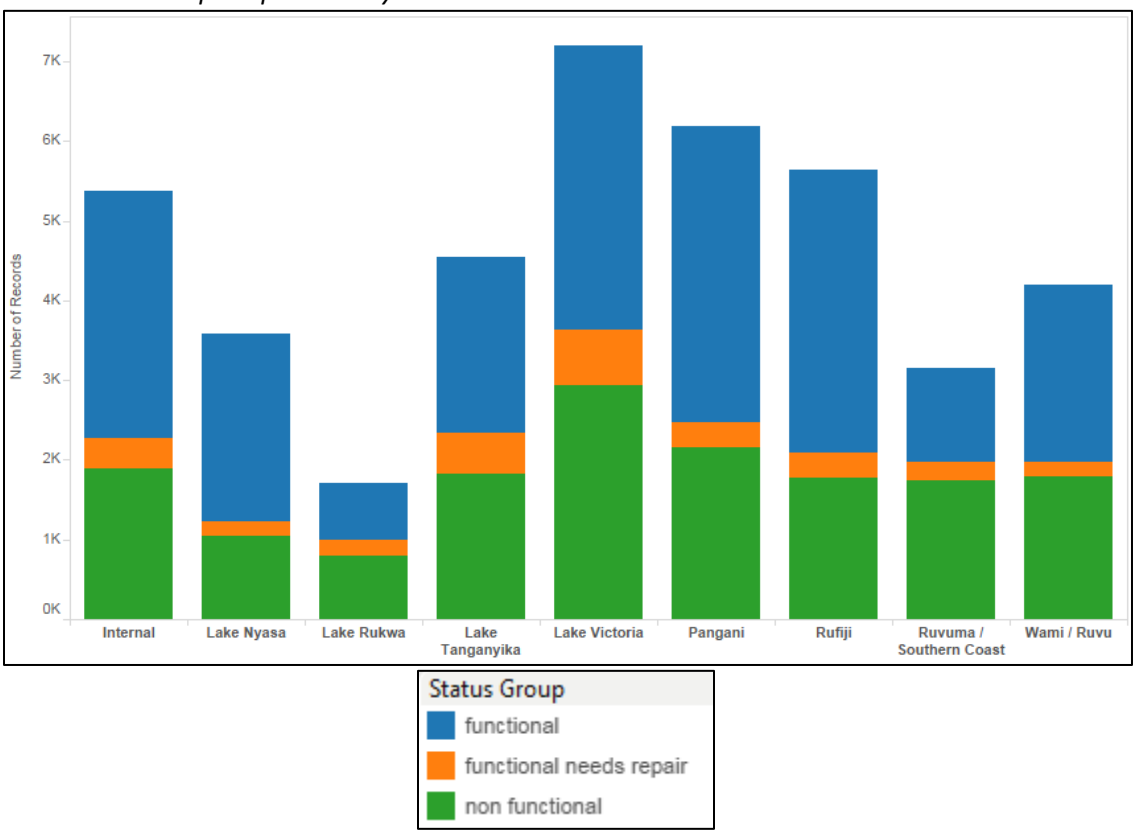
The functionality of water pumps appears to vary by *payment* class as well, with “never pay” and “unknown” having higher levels of “non-functional” pumps compared to the paying classes (Figure 6).

Figure 6. Status Group Proportions by Payment Class



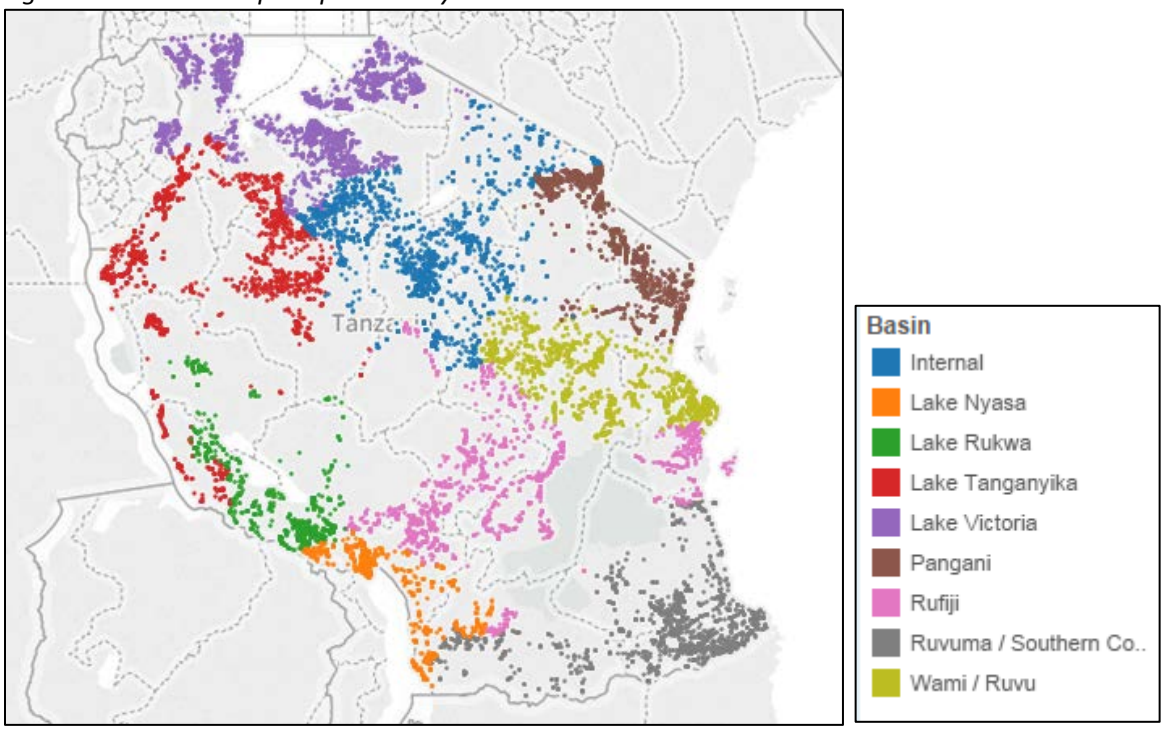
Water pump status appears to be somewhat similar across the various geographical *basins*, however, there may be a tendency for Lake Victoria to have a higher portion of water pumps that need repair, while the Ruvama / Southern Coast basin appears to have a greater proportion of "non-functional" water pumps (Figure 7).

Figure 7. Status Group Proportions by Basin



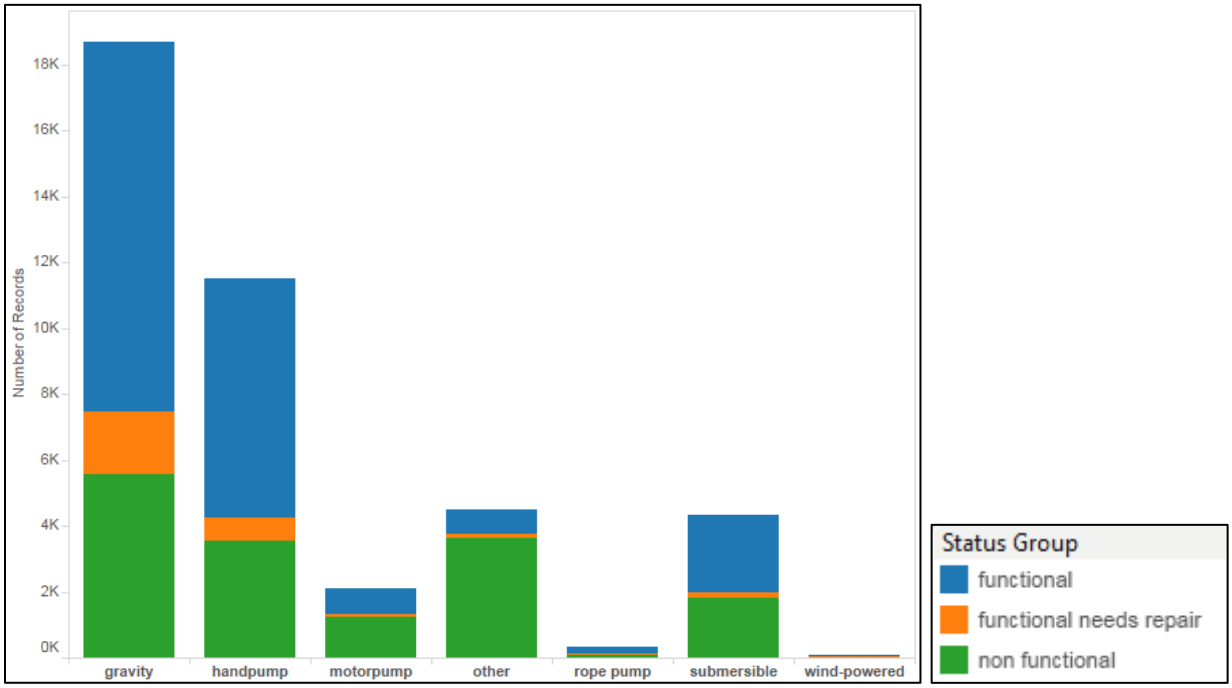
The figure (Figure 8) below shows the geographic location of each basin.

Figure 8. Status Group Proportions by Basin



A review of water pump status by *extraction type class* suggests that pumps that use "gravity" have a relatively higher portion of pumps that are in need of repair when compared with other types. The extraction type "other" has disproportionately more "non-functional" water pumps.

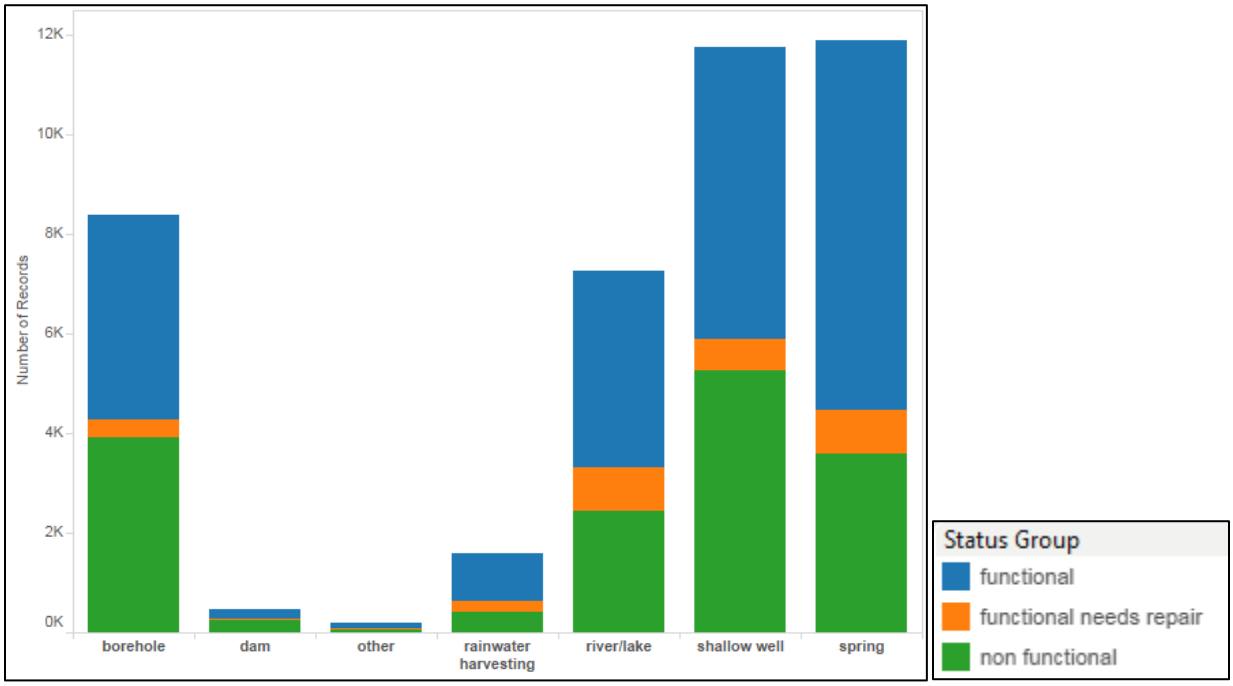
Figure 9. Status Group Proportions by Extraction Type Class



An assessment of water pump *status group* by *source type* suggests that a disproportionately large number of "borehole" and "shallow well" water pumps are "non-functional." The water pumps with the

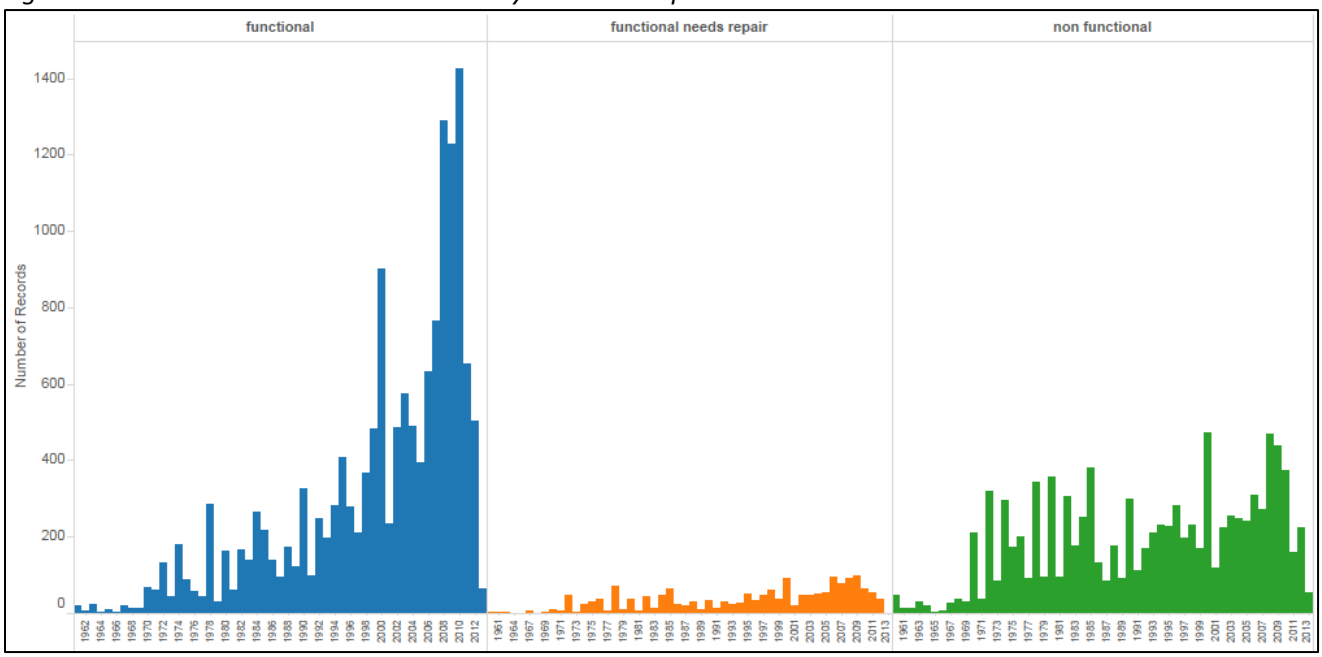
"spring" *source type* have a more favorable distribution, such that water pumps are more likely to be "functional" or "functional in need of repair" (Figure 10).

Figure 10. Status Group Proportions by Source Type



The distribution of *construction year* appears to vary by water pump *status group* such that "functional" water pumps are more likely to be more recently constructed, whereas "non functional" and "functional needs repair" pumps are less likely to have been constructed after the year 2000 (Figure 11).

Figure 11. Construction Year Distribution by Status Group



Appendix E – Data Visualization Uses and Functionality

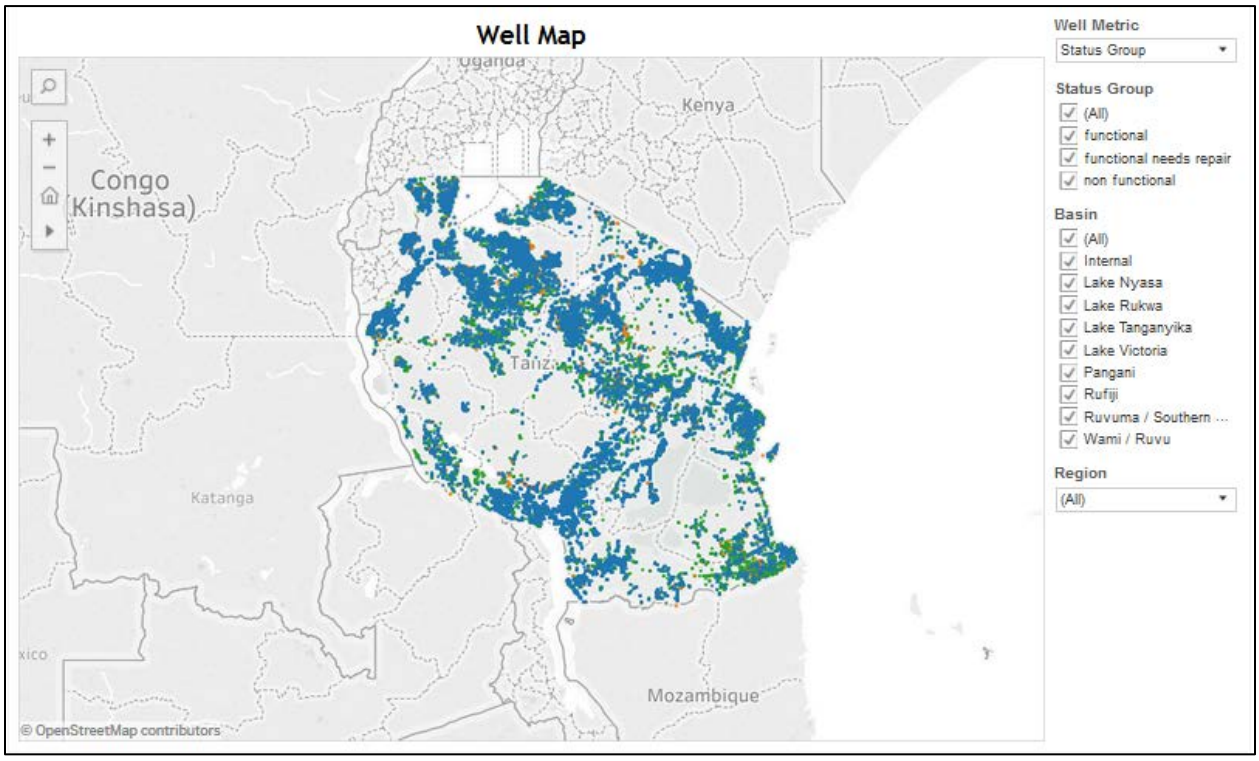
Exploratory Well Map

Link: https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-WellsMap/WellMap

Filter Options:

- Status Group
- Basin
- Region

Use: Understanding the geographic location of the wells as it relates to the functionality. Used to investigate the geographic distribution of the wells by status group.



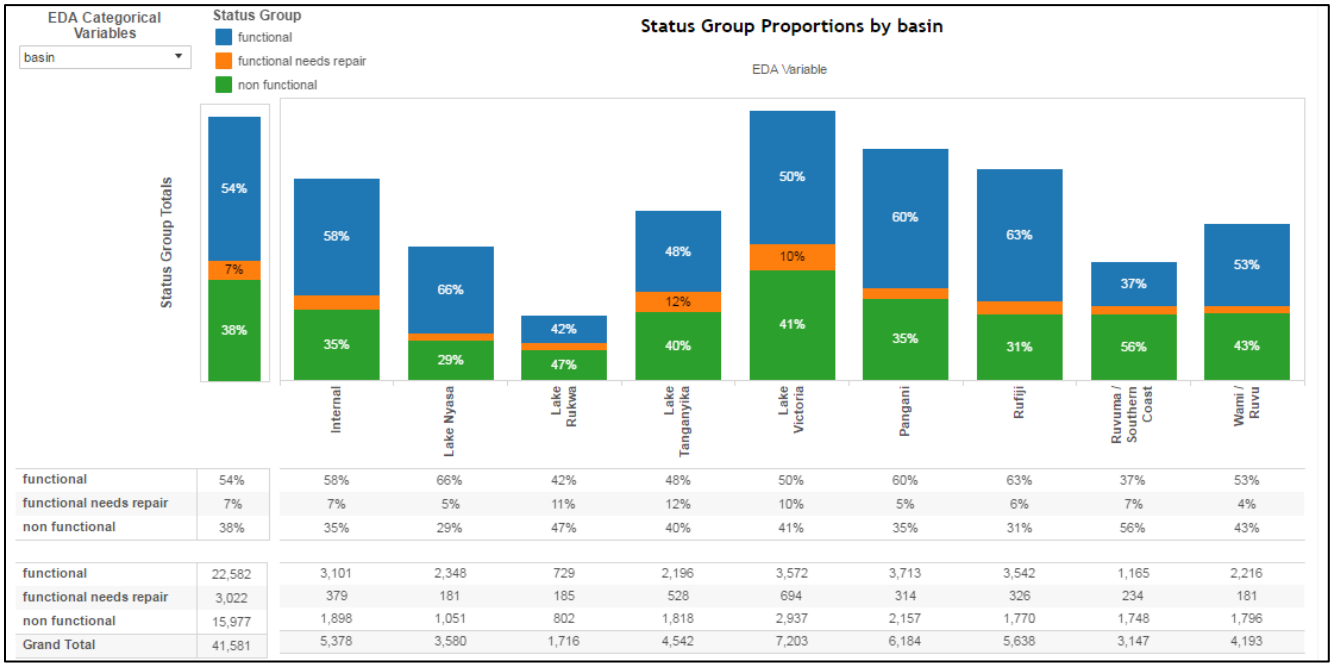
Interactive EDA

Link: https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-InteractiveEDA/InteractiveEDA

Filter Options:

- Categorical Variable

Use: Enhances EDA by allowing each team member to perform their own EDA and choose the variables they would like to analyze. The dashboard compares all predictor variables with the response variable and can be accessed and manipulated by all team members.



		Internal	Lake Nyasa	Lake Rukwa	Lake Tanganyika	Lake Victoria	Pangani	Rufiji	Ruvuma / Southern Coast	Wami / Ruvu
functional	54%	58%	66%	42%	48%	50%	60%	63%	37%	53%
functional needs repair	7%	7%	5%	11%	12%	10%	5%	6%	7%	4%
non functional	38%	35%	29%	47%	40%	41%	35%	31%	56%	43%
functional	22,582	3,101	2,348	729	2,196	3,572	3,713	3,542	1,165	2,216
functional needs repair	3,022	379	181	185	528	694	314	326	234	181
non functional	15,977	1,898	1,051	802	1,818	2,937	2,157	1,770	1,748	1,796
Grand Total	41,581	5,378	3,580	1,716	4,542	7,203	6,184	5,638	3,147	4,193

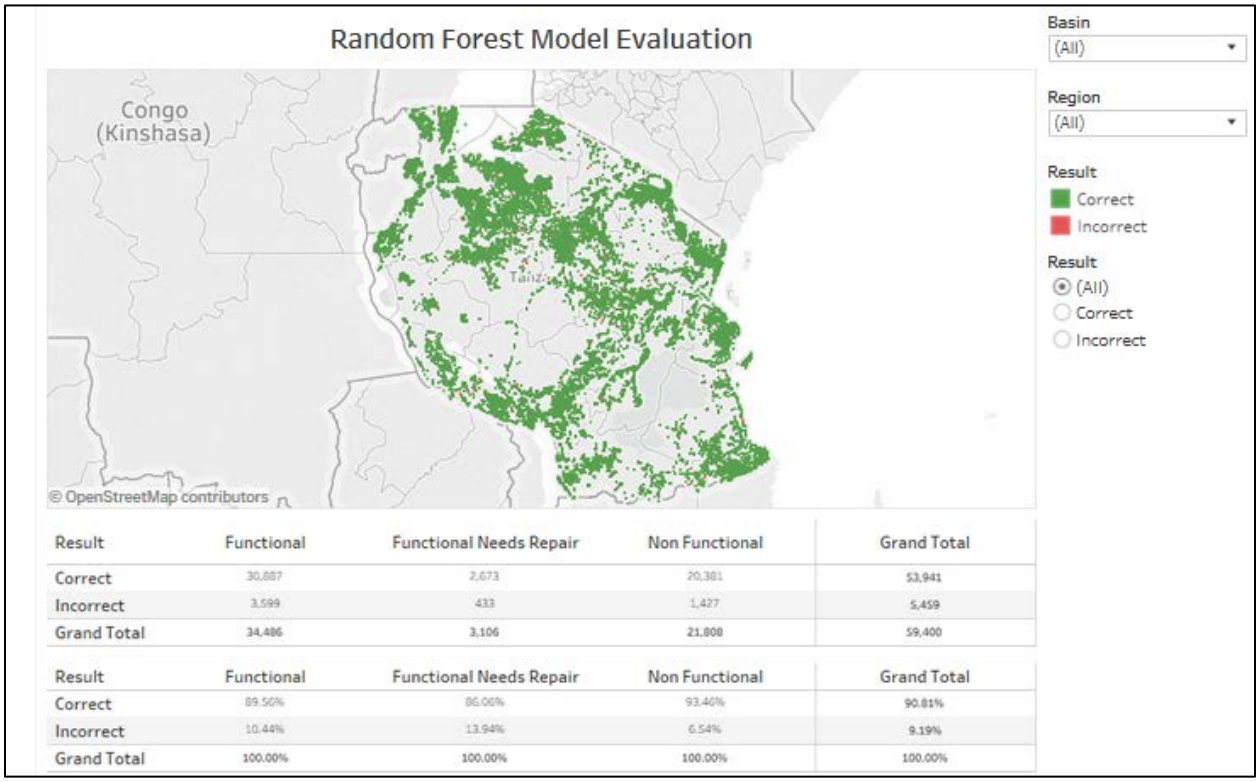
Random Forest Model Evaluation Map

Link: https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-RandomForestModelEvaluation/RandomForestModelEvaluation

Filter Options:

- Basin
- Region
- Result

Use: Evaluate the recommended predictive model based on geography. Identify areas that had higher or lower than average prediction success.



Result	Functional	Functional Needs Repair	Non Functional	Grand Total
Correct	30,887	2,673	20,381	53,941
Incorrect	3,599	433	1,427	5,459
Grand Total	34,486	3,106	21,808	59,400

Result	Functional	Functional Needs Repair	Non Functional	Grand Total
Correct	89.56%	86.06%	93.46%	90.81%
Incorrect	10.44%	13.94%	6.54%	9.19%
Grand Total	100.00%	100.00%	100.00%	100.00%

Well Maintenance Team Deployment Map

Link: https://public.tableau.com/profile/rob.herold#!/vizhome/PumpItUp-WellMaintenanceTeamDeploymentMap/WellMaintenanceTeamDeploymentMap

Filter Options:

- Well Status
- Basin
- Region

Use: Identify the geographic location of wells that are functional but in are in need of repair and the non functional wells in order to send out well maintenance teams.