Predicting Tanzanian Well Functionality - MATH 1298 Project - Part 1

Brian Tregillis s3526783

1 September 2018

Introduction

In Tanzania access to water is an important issue. For example, only 100,000 of the city Dar Es Salaam's 3.5 million citizens have access to running water (Japan International Cooperation Agency 2008). DrivenData, a website that hosts socially oriented data competitions is providing access to data on the observed functionality of wells in Tanzania. The aim of this project is to use this data to predict the functionality of wells, and if possible to better understand the variables that contribute to well functionality. Specifically, using the data provided, the project will try and assign a status to an test data set to predict if the well is Functional, Functional but in need of repair, or Non-functional. A useful model would enable groups to understand which wells are most likely to be non-functional and provided services to improve their functionality. Additionally, the variables that are important (including some that may not be included in the final model), may be able to identify why these wells are becoming non-functinoal. For instance, there may be issues with local administration or things like geography may create barriers to access.

The provided data set has 59,400 observations with 41 variables including the response variable "status_group".

The Response Variable: Pump Status *status_group*

The variable of interest in this problem is the variable named "status_group". This indicates the observed status of the pump and has three possible values: Functional; Functional needs repair; Not Functional.

As can be seen in table 1, the majority of the 59,400 observations resulted in an outcome of Functional (32K) or Non Functional (23K). There are only 4,317 observations where the pump was found to be Functional but needing repair.

Table 1: Response Variable Observations By Category

	Observations:
functional	32259
functional needs repair	4317
non functional	22824

Predictors

There are 40 other variables in the data, table 2 details the variables.

Table 2: Data Variables In The Tanzanian Well Data

id	numeric	Observation identifier.
status_group	character	The status of the well, the response variable for this set.
amount_tsh	numeric	Total Static Head, the height that the water is raised to get to the pump.
date_recorded	date	The date that the observation was made.
funder	character	The name of the organisation that funded the well.
gps_height	numeric	Altitude of the well
installer	character	Organsisation who installed the well.
longitude	numeric	The longitude of the well's location
latitude	numeric	The latitude of the well's location
wpt_name	character	The name of the well
num_private	binary	Unknown, no description given
basin	character	Name of the water basin
subvilage	character	Sub Village, Geographic
region	character	Region, Geographic
region_code	numeric	Region Code, Geographic
district_code	numeric	District Code, Geographic
lga	character	Local Government Authority, Geographic
ward	character	Ward, Geographic
population	numeric	Population living around the well
public_meeting	binary	Unknown, assumption that it is about if the well is in a public meeting area.
recorded_by	character	Organiastion who recorded the observation
scheme_management	character	Organiastion who manages the scheme that the well is operated through
scheme_name	character	The name of the scheme that the well is operated through
permit	binary	There a permit for the well
construction_year	numeric	Year that the well was built
extraction_type	character	How water is extracted from the well, most categories
extraction_type_group	character	How water is extracted from the well, middle categories
extraction_type_class	character	How water is extracted from the well, least categories
management_group	character	The type of organisation who manages the well
payment	character	What type of payment is required to use the well
payment_type	character	What type of payment is required to use the well
water_quality	character	Quality of water that is pumped from the well, detailed
quality_group	character	Quality of water that is pumped from the well, broad
quantity	character	Categorisation of how much was is available from the well
quantity_group	character	Categorisation of how much was is available from the well
source	character	What type of water source the wells water comes from, detailed
source_type	character	What type of water source the wells water comes from, middle
source_class	character	What type of water source the wells water comes from, broad
waterpoint_type	character	How is the water accessed from the well, detailed
waterpoint_type_group	character	How is the water accessed from the well, broad

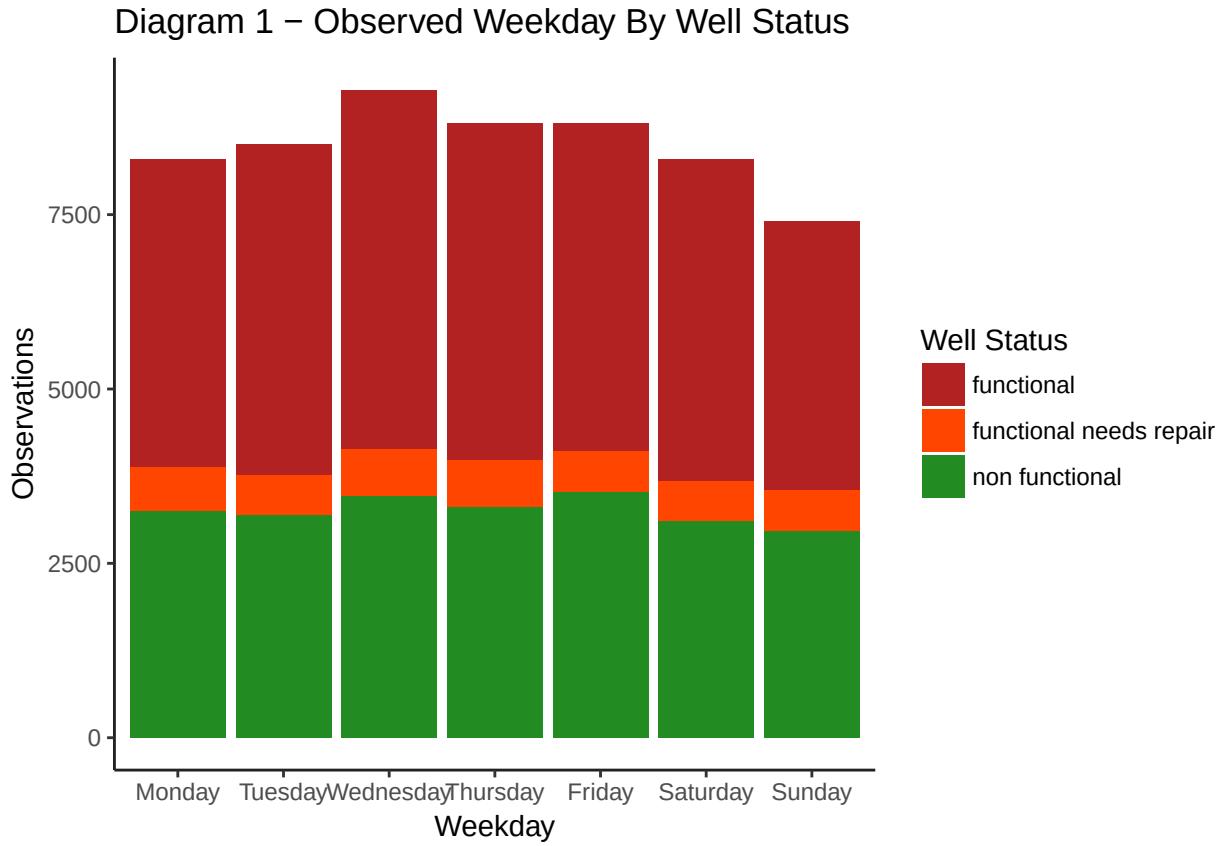
Date of observation *date_recorded*

A summary of the observation dates, shows that observations have been made from 1960 to 2013. The date field is imported as a character field, so a transformation is conducted to convert it to a POSIX date format. From this, the Year, Month and day of the week of each of the observations is extracted.

The plot of the observations by the year and month they were taken shows that there were very few observations prior to 2011. The majority of observations fall in 2011 and 2013, with a smaller number in 2012. The observations are not evenly distributed amongst the months and years. For instance, only 2012 has a sizable number of observations in October and there are very few observations in May, June, September or December in any year.

Given this inconsistent observation periods, it would be difficult to draw any conclusions about changes in the incidence of functional wells over time as it may just be a seasonal difference that is observed. Also, it is difficult to draw seasonal inference as observation practices or the quality of wells may change over time. For this reason this data was not included in the prediction set.

However, looking at the plot of the observations by weekday, there is a good distribution of observations across all of the days. There is very little difference in the spread of the Pump Status variable across the days, this may mean it will have little benefit in prediction but there is a possibility that it's interaction with other variables will make it useful.



Administrative Boundaries

Within the data set there are six variables related to geographical administrative boundaries. These are, from largest area to smallest: Region and Region Code; District and Local Government Authorities (LGA); Ward and Sub-village.

Region *region* and Region Code *region_code*

There are 21 regions. The region code variable was be dropped as it duplicates the region name, and the region name is more intuitive. The region with the most observations is Iringa with 5,294 and the least

observations were made in Dar es Salaam, 805.

Table 3: Observations By Region

Region	Observations
Iringa	5294
Shinyanga	4982
Mbeya	4639
Kilimanjaro	4379
Morogoro	4006
Arusha	3350
Kagera	3316
Mwanza	3102
Kigoma	2816
Ruvuma	2640
Pwani	2635
Tanga	2547
Dodoma	2201
Singida	2093
Mara	1969
Tabora	1959
Rukwa	1808
Mtwara	1730
Manyara	1583
Lindi	1546
Dar es Salaam	805

District *district_code* And LGA *lga*

The districts are the second largest, and the data is provided as district codes. There are 21 district codes in the data, however, they need to be combined with the regions because the district code is reused across each region. For example, all of the regions have a district 1. Therefore, these two variables were combined to create a variable called "region_district", where district 1 in the Arusha region would display as "Arusha_D1". Making this combination results in 132 districts in the data.

The data is left skewed, as can be seen in Diagram 2, with more than 65% of region-districts having less than 500 observations.

Table 4: District-Region - Observations By Unique District-Region - Summary

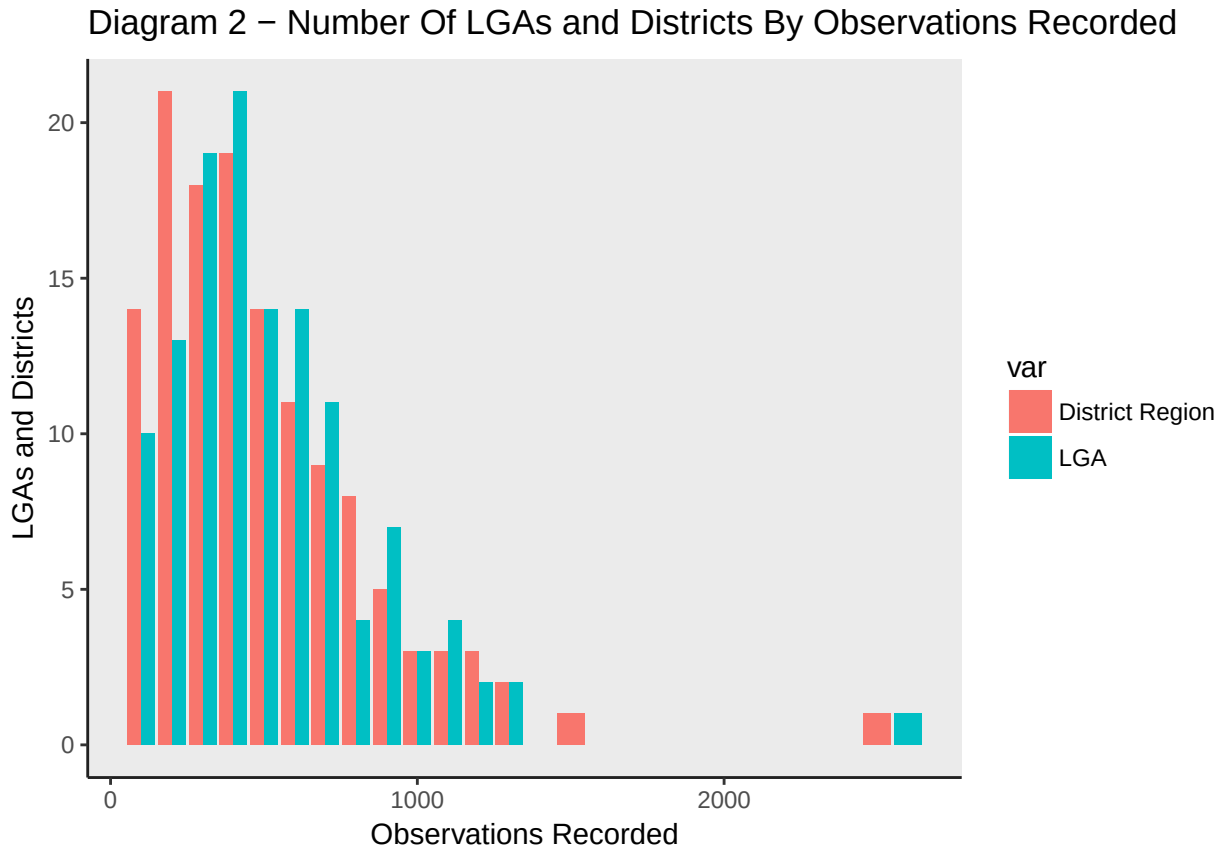
Factors	Mean Observations	Min Observations	Max Observations
132	450	1	2473

There are also 125 LGAs in the data ranging from one observation to 2503 observations. The Tanzanian 2012 Population and Housing Census International (Household Survey Network 2014) implies that the districts and LGAs are the same thing, as the districts are referred to be the LGA in this data set. Mostly LGAs and Districts are aligned in the data, however, there is not a complete match with some districts having observations from multiple LGAs and vice-versa.

Table 5: LGA - Observations By Unique LGA - Summary

Factors	Mean Observations	Min Observations	Max Observations
125	475	1	2503

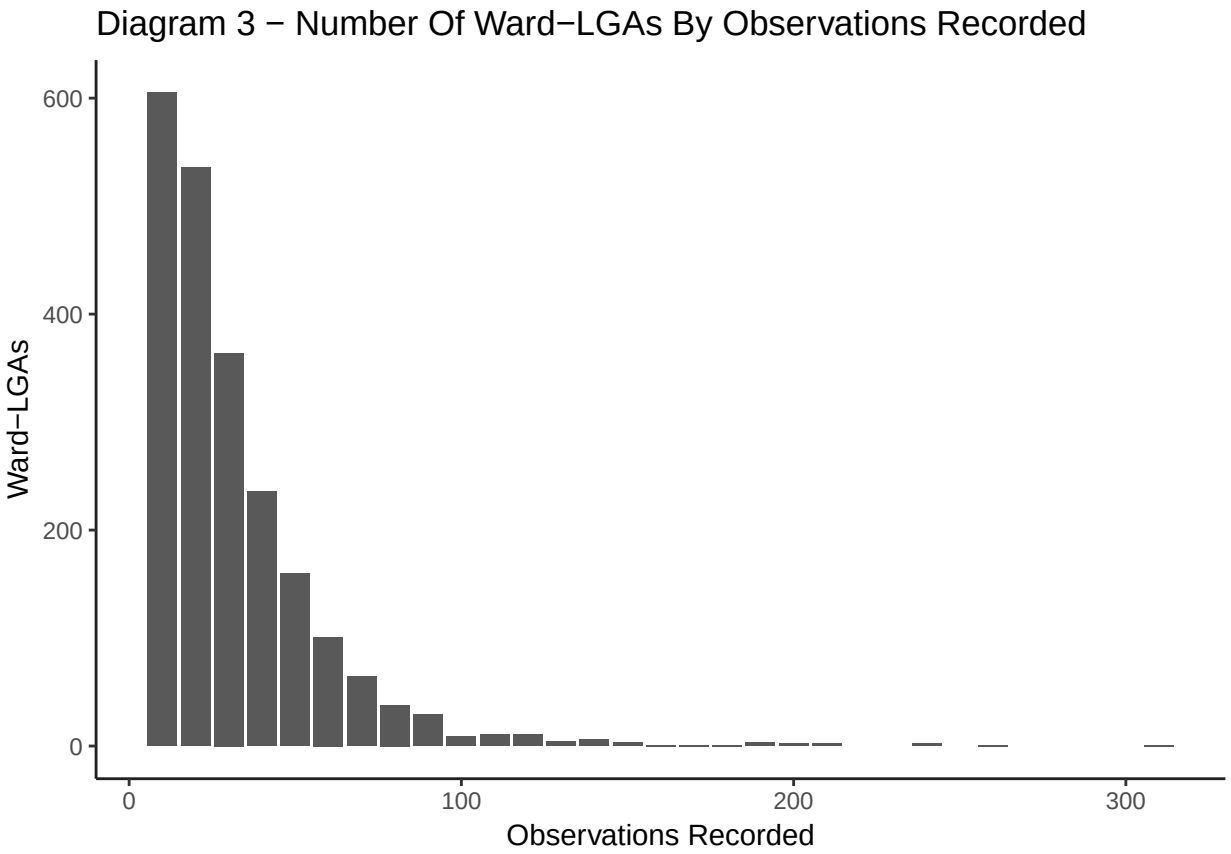
Looking at the distribution of observations by LGA (diagram 2), this is likely better distributed than the district data for predictive purposes as there are not as many low observation areas. Therefore considering the similarity between the two and the better distribution of LGA, LGA was retained while district-region and district_code were dropped from this analysis.



Ward *ward*

There are 2092 wards in the data ranging from one observation to 307. The naming is clean, but there are 88 ward names that are found within multiple LGAs. This is not a data entry issue as, for instance, "Majengo" is a ward in four LGAs (Wikipedia 2013) To clean this, a new variable ward_lga is created which combines the ward and lga data. If the ward in an observation is "Majengo" and the LGA is "Dodoma", the new variable will be named "Majengo_Dodoma". The variable "ward" is then redundant and was removed.

For this new "ward_lga" variable, there are 2191 combinations with the same range of observations as the ward data. In diagram 3 it can be seen that the data has a long right tail with small number (14) of wards having between 150 and 307.



To try and create a variable with more equal distribution, a new variable was created where the default value is the name of the Region followed by LGA followed by the Ward name e.g. for the Majengo ward in the Dodoma LGA in the Dodoma Region the value is "Dodoma_Dodoma_Majengo". However, in the case that there are less than 150 observations in one ward, then the Ward value becomes "_OTHER" name e.g. "Dodoma_Dodoma_OTHER". And, if the LGA has less than 150 observations the LGA becomes OTHER and is rolled up to the region level i.e. if the Dodoma ward had less than 150 observations it would become "Dodoma_OTHER_OTHER" This new variable is called "admin_district". "admin_district" contains 139 unique combinations with a range from 21 observations to 1376 combinations.

Subvillage *subvillage*

The final administrative area is "sub-village", there are 19,288 options. However, there are data quality issues with numerical names, other non-alpha characters and single letter names. There are also duplicated sub-village names among wards. The majority of sub-villages will have one or only couple of observations and will not be useful for prediction, however, they will be useful in imputing other missing features due to their relatively small geographic coverage i.e. missing GPS coordinates and altitudes, so the a rough data clean was conducted.

Table 6: Subvillage-Ward - Observations By Unique Subvillage-Ward - Summary

Factors	Mean Observations	Min Observations	Max Observations
26281	2	1	174

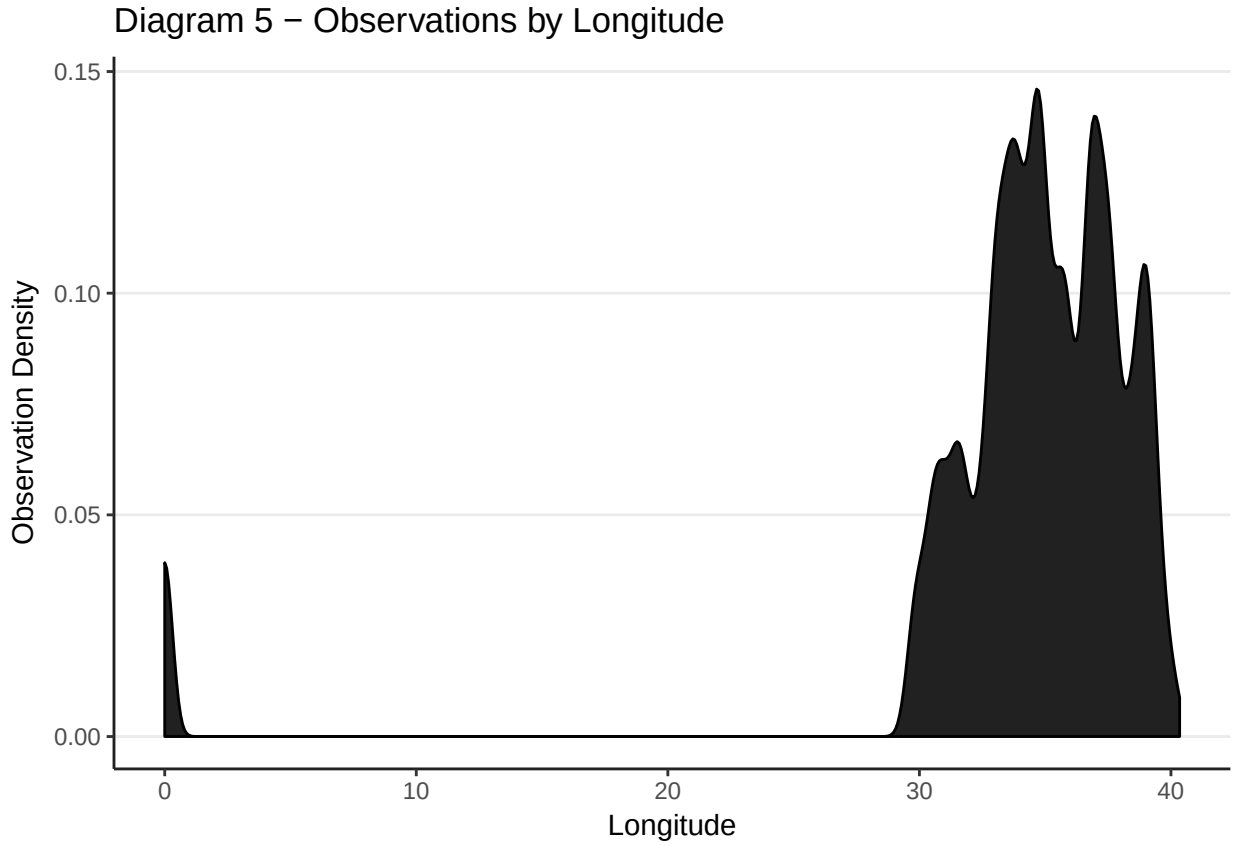
The first step was to clean out the dirty data. The data was converted to lowercase, non-alpha characters were removed, words within each string that were only one character were removed, double spaces were reduced to single spacing and any trailing or preceding spaces were removed.

Next, a new variable "sub-village_ward" was created. This is the name of the sub-village, an underscore and the name of the ward. Values that are "NA" in the previous step remain "NA", another option would have been to aggregate them under a "UNKNOWN_" ward name. They have been left as "NA" because the usefulness of this data is for imputing averages values for Longitude and Latitude. In that case it will be better to take the total average from the ward, instead of the average of the unknown villages i.e. including all sub-village locations not just the unknown ones because new "NA" observations in the same ward will not necessarily be similar to the current "NA" observations. "sub-village_ward" has 27,247 unique values. 14,692 of these have one observation, and only 260 have more than 10 observations. The "sub-village" variable was removed from the data set.

GPS Data

Latitude *latitutde* and Longitude *longitude*

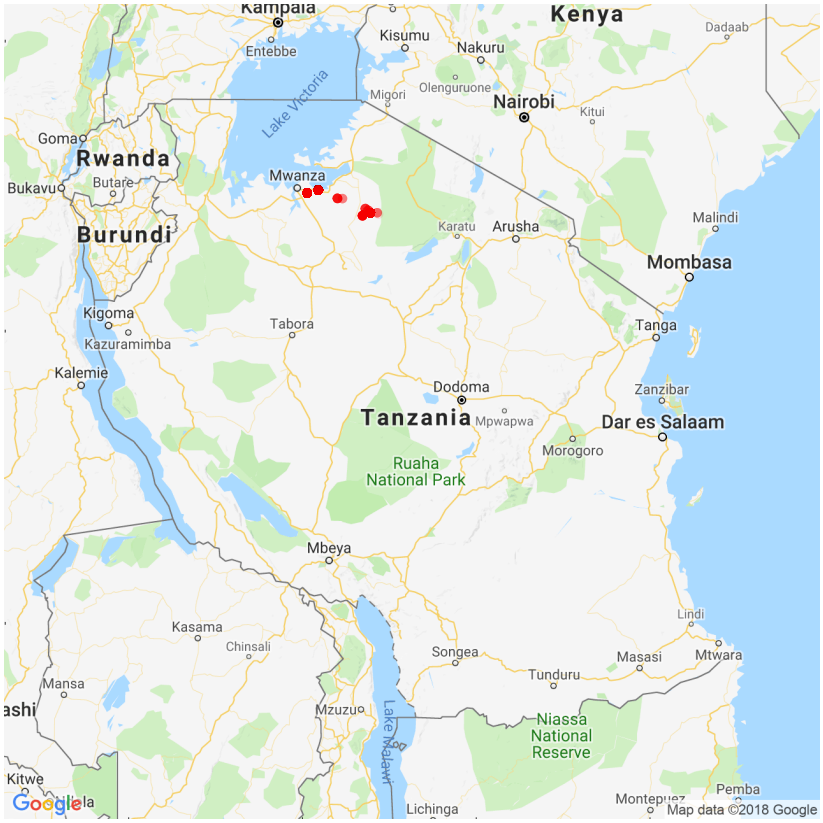
The data set includes both the latitude and longitude of the well location which could be very useful where the geographical location may influence the functionality of the well, possible factors could be remoteness affecting the availability of trade assistance, the harshness of weather on the pump or even the terrain. Looking at the longitude in a density plot (diagram 5), there is a small amount of data clustered around a longitude of 0 while the majority of the data falls between 30 and 40 degrees. Around 28 degrees longitude marks the most easterly part of Tanzania, therefore, we can conclude this is missing data.



There are 1812 records with missing longitude data and the same records have missing latitudes. Out of 125

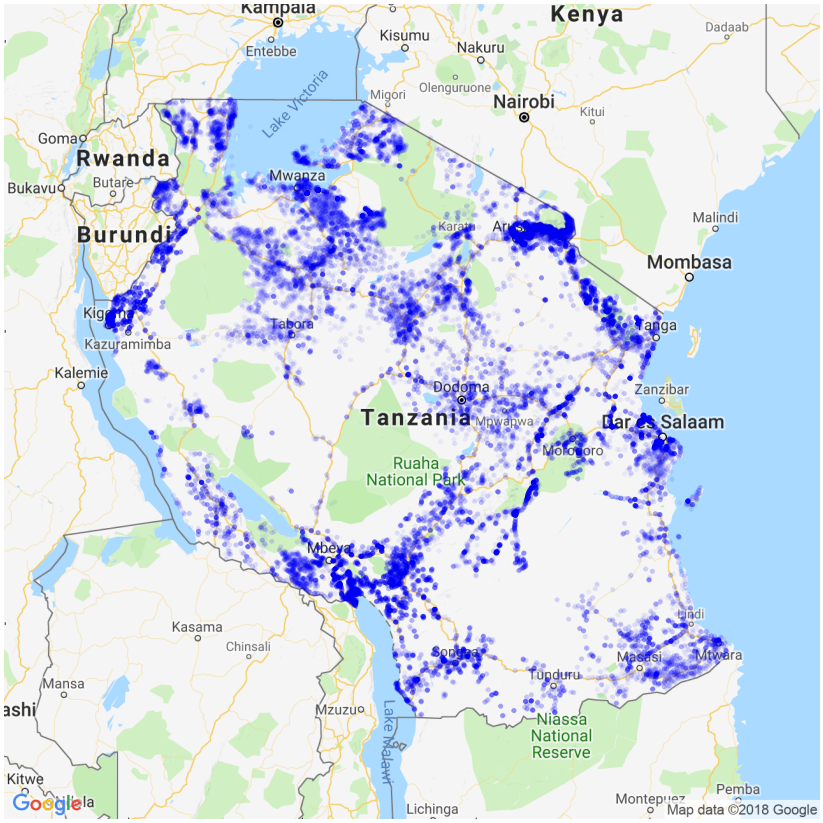
LGAs all of the missing data is from 3 distinct LGAs that are within the Mwanza and Shinyanga regions. As mentioned in the Geography section of this report, the administrative areas were used to impute GPS data for the missing records. This was done by starting with the lowest level, "sub-village_ward" and computing an average lat/lon where there is a match i.e. not an NA. Where there is no match found this process is repeated at "ward_lga", then "lga", and then "region". Of the missing data, 16 records were matched at "sub-village_ward", 175 at "ward_lga", 1133 at "lga" and the remaining 488 at "region" level.

Diagram 6 − Map Of Imputed Well GPS Coordinates



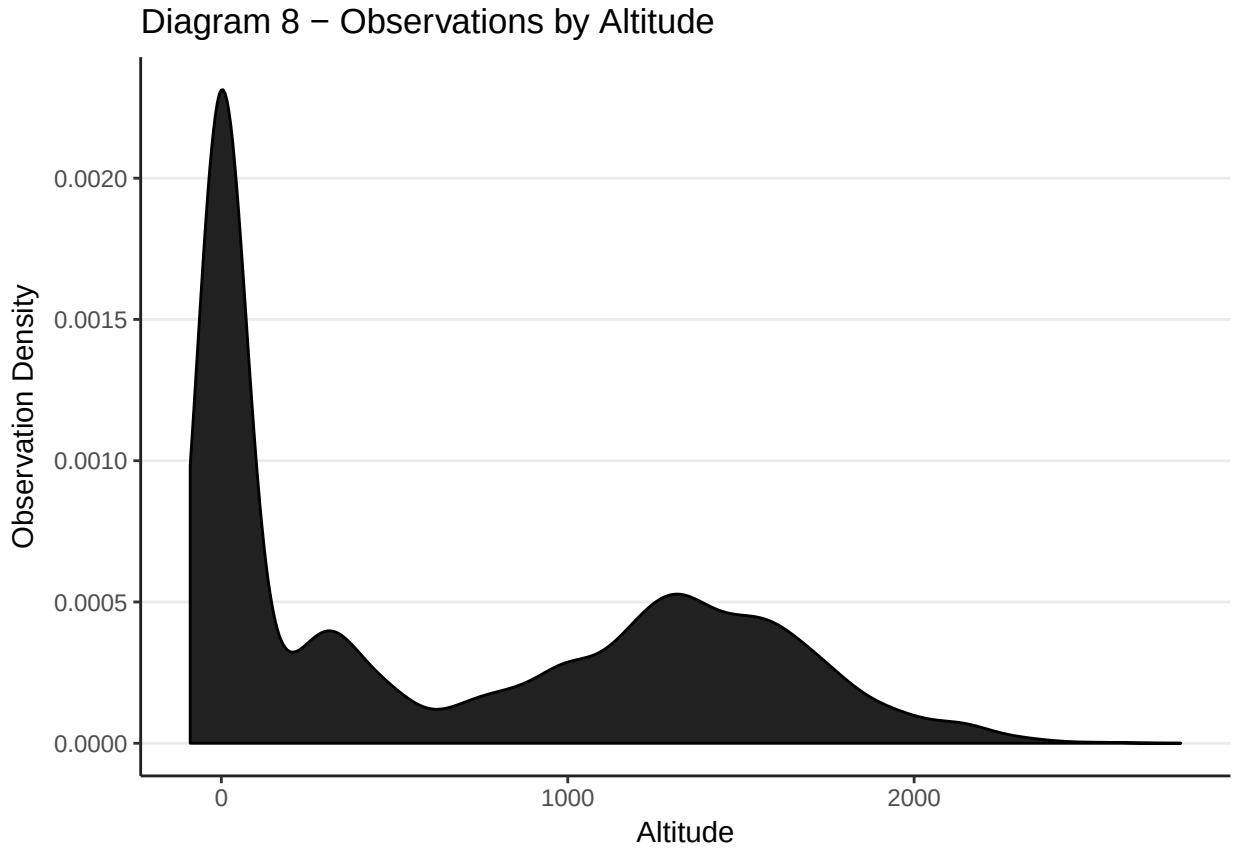
The map in diagram 6 shows the distribution of the imputed observations in red, and the location of all of the observations (including the imputed ones) is shown in diagram 7.

Diagram 7 − Map Of All Well GPS Coordinates



Altitude *gps_height*

The Altitude variable gives the height above sea water of the well. Looking at the density plot of Altitude (diagram 8), there are a large number of points around 0. There are in total 20,438 observations with an altitude of zero in the data. This is possible but more likely to represent missing data.



To test this hypothesis, 5 random observations with zero altitude and 5 with non-zero were selected. The GPS coordinates were then looked up in an online tool (What is my elevation 2018) to retrieve the altitude. The results in table 7 show that the zero altitude data does appear to be missing as the retrieved altitudes range from 1,130 to 1,432. On the other hand the difference between the two altitudes where the original was non-zero vary from -857 to 858. Using the tool also made it clear that the altitude is measured in metres, therefore the variable was renamed “altitude_metres”.

Table 7: Altitude In Data Against Online Lookup - Height Is Zero

	id	gps_height	latitude	longitude	online_height
7318	7057	0	-4.728629	32.33576	1130
41761	51498	0	-9.296508	33.91706	1359
402	24575	0	-2.620502	33.09156	1195
16141	18431	0	-9.260997	33.84013	1432
9396	9742	0	-3.690164	33.33200	1166

Table 8: Altitude In Data Against Online Lookup - Height Not Zero

	gps_height	latitude	longitude	online_height	height_diff
26473	259	-8.136332	36.67117	258	1
21784	1887	-2.661134	36.11189	1030	857
18574	293	-8.934796	36.12536	1151	-858
54786	1281	-2.073638	33.00024	1280	1
32118	1603	-4.582442	34.84708	1636	-33

Based on these outcomes, zero altitude data was treated as missing. A new altitude was imputed using a K Nearest Neighbours regression with Latitude and Longitude as the predictors. The data where the altitude and Latitude/Longitude were present in the original set was taken to build the model. A 10% test set was created and then models were built with K values of 1 to 7. Mean Squared Error (MSE) was used to measure performance and best model used K=1 with an MSE of 733, meaning an average error of 27 metres. The performance results can be seen in table 9. The model with K=1 was then used to predict the missing altitude numbers. After running the model and updating the altitude data, another sample of 5 of the predicted altitudes were compared with the online tool. The results are available in table 10 and show that there is still error in the data, but major improvement in the results.

Table 9: Results for K from KNN for Altidue data

K	MSE
1	733.4084
2	762.9838
3	847.2039
4	1024.8189
5	1142.9713
6	1349.4016
7	1521.7681

Table 10: Altitude In Data Against Online Lookup - KNN Determined

	altitude_metres	latitude	longitude	online_height	height_diff
23176	795	-8.592399	33.11487	847	-52
18093	1057	-8.683061	34.04348	1038	19
14687	2233	-8.743335	33.70469	1113	1120
54550	2484	-9.358076	33.63204	1169	1315
29446	2103	-9.490585	34.03702	484	1619

Well Construction

Installer *installer*

The installer variable gives information about who installed the well. There is a high level of data entry discrepancy (for instance there is "Losaa-Kia water supp", "Losa-kia water suppl", "Losakia water supply"). The same data cleaning rules as for sub-village were applied, namely, lowercase conversion, removal of non-alpha characters, removal of one character sub strings, double spaces reduced to single spacing and and white space was removed.

Table 11: Well Installer - Observations By Unique Installer - Summary

Factors	Mean Observations	Min Observations	Max Observations
1857	32	1	17426

After cleaning, there are 1,857 factors in the data, most containing only a few observations. However, 30% (17,406) of the observations come from a category called "DWE" which stands for District Water Engineer. There is an additional category called RWE (Rural Water Engineer) with 1206 observations. A new variable was created called "installed_water_engineer" with a 1 for all observations where the installer was RWE or DWE and a 0 for all other observations. The installer variable was dropped.

A cross tabulation of the Water Engineer variable by the well status shows very little difference between the two groups. Those not installed by a Water Engineer are very slightly more likely to have a Functional status (55.2% to 52.4%).

Table 12: Proportion Of Well Functionality Observations by Water Engineer Installer Status

	0	1
functional	55.2	52.4
functional needs repair	6.3	9.4
non functional	38.5	38.2

Funder *funder*

The source of the funding for the well is captured in the variable funder. Again, there are data entry issues although not as severe (for instance there is "Losaa-kia Water Supply" "Losakia Water Supply") and the data was cleaned as per the installer variable. For this variable, the largest number observations is 9,084 for "Government Of Tanzania". A new variable was created called "government_funded", with 1 value of 1 for all observations where "Government of Tanzania" was the observation and 0 for all other observations. A cross tabulation of this new variable shows a difference in well functionality, with 41% of Government wells being functional compared with 56.7% of the others.

Table 13: Proportion Of Well Functionality Observations by Government Funded Status

	0	1
functional	56.7	41.0
functional needs repair	7.2	7.7
non functional	36.1	51.3

There were also 22,642 records where the funder name was the same as the installer name (not including blanks) and a variable was created to capture where the installer and funder were the same. Some additional cleaning and matching rules were applied to the data before creating the variable. Firstly blank names were changed to 'unknown_funder' and 'unknown_after' to avoid false matches in the next stage, and this reduced matches to 18,119. Secondly, a fuzzy matching logic was used to account for data entry errors. The method used was Optimal String Alignment distance which returns a 'string distance'. 'String distance' is the number of deletions, insertions, substitutions and character switches to make two strings match. Funder was considered to match installer When the string distance was two or less. Two was chosen as the max string distance for matching as there was a high incidence of three letter funders and installers which would produce a score of 3 even if all characters were different. After the fuzzy matching there were 21,530 matches in total. The new variable is called "funder_installer", with a value of 1 were there is a match and 0 otherwise. The cross tabulation does not show much difference between the two groups.

The original funder variable was dropped from the data set.

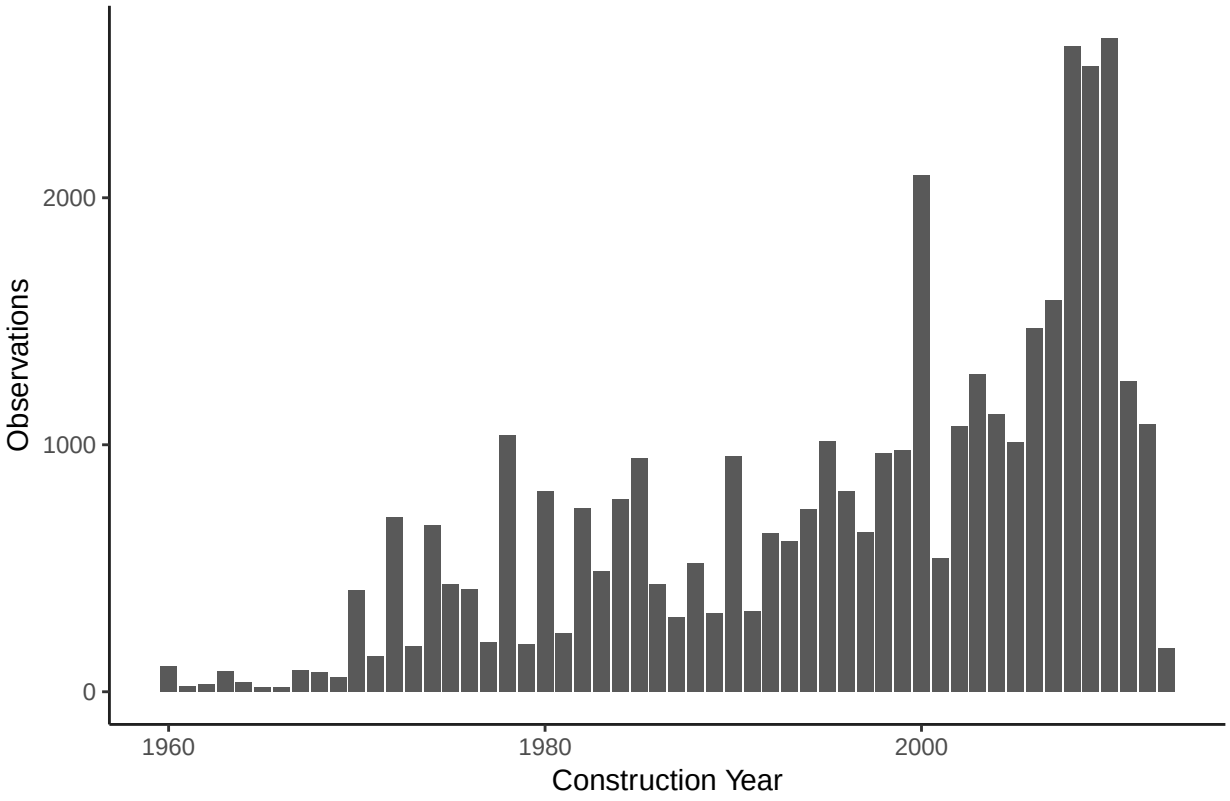
Table 14: Matched Funder and Installer Status By Observed Well Functionality (%)

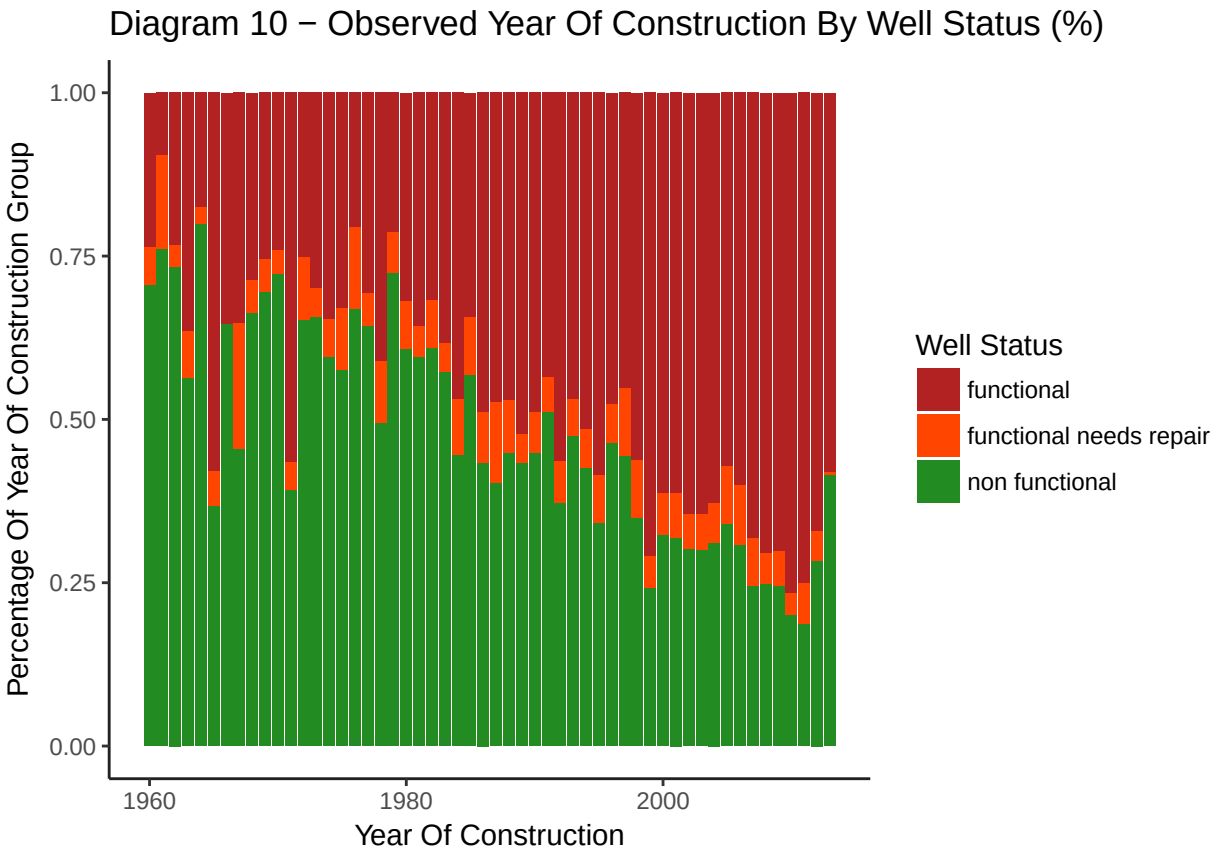
	0	1
functional	54.1	54.7
functional needs repair	7.1	7.6
non functional	38.9	37.7

Construction Year *construction_year*

The year that the well was constructed is also in the data with wells constructed between 1960 and 2013. More than a third of the data doesn't have any construction year. Diagram 11 shows that the well construction is skewed to more recent years, 17 wells observed were constructed in 1966 and 2645 were from 2010. Looking at a chart of the percentages of observations by well status, there is a clear negative linear relationship between year and non-functional wells. This indicates that it might be useful in prediction, but the data without construction year will not be able to use this.

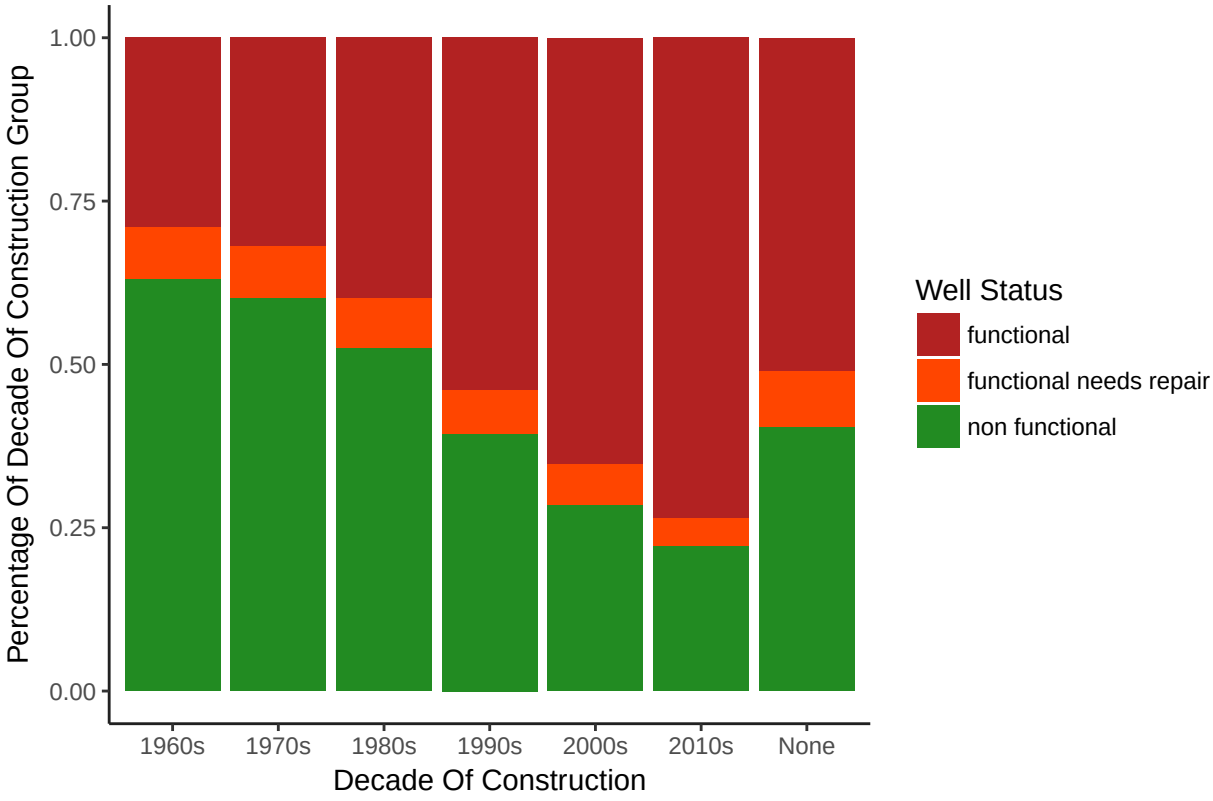
Diagram 9 − Observations by Construction Year





To amend this, the data was turned into a categorical variable, being split into decades and a 'None' category to represent missing data. This method maintains the relationship, as can be seen in diagram 12, and allows the data to be used.

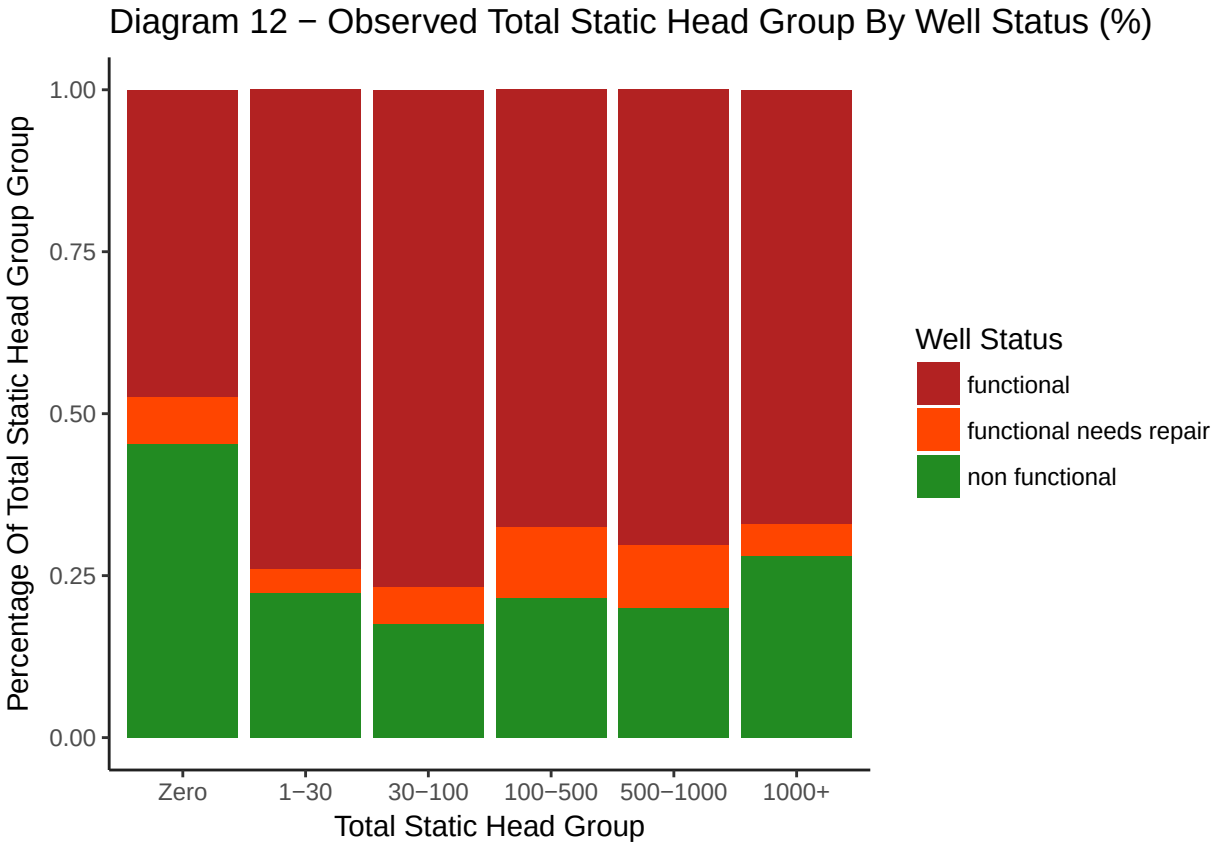
Diagram 11 − Observed Decade Of Construction By Well Status (%)



Total Static Head *amount_tsh*

Total static head is "the difference in elevation between the liquid levels of the suction and discharge" (Australian Pump Technical Handbook 1987), meaning how far vertically the water has to travel from the source. This variable has a range of zero to 350,000, with more than a third of the data having zero as the value. Without knowledge of how this variable is calculated both ends of the spectrum seem unlikely, but in an attempt to not remove data that may have value it was retained. However, an additional categorical variable was also created which coded all the 0 values together and then discretised the remaining data into three groups of equal number of observations. This variable is called "total_static_head_grp".

There does appear to be some differences which can be seen in the chart of this new variable, the zero value data has a much higher total static head than the other variables.



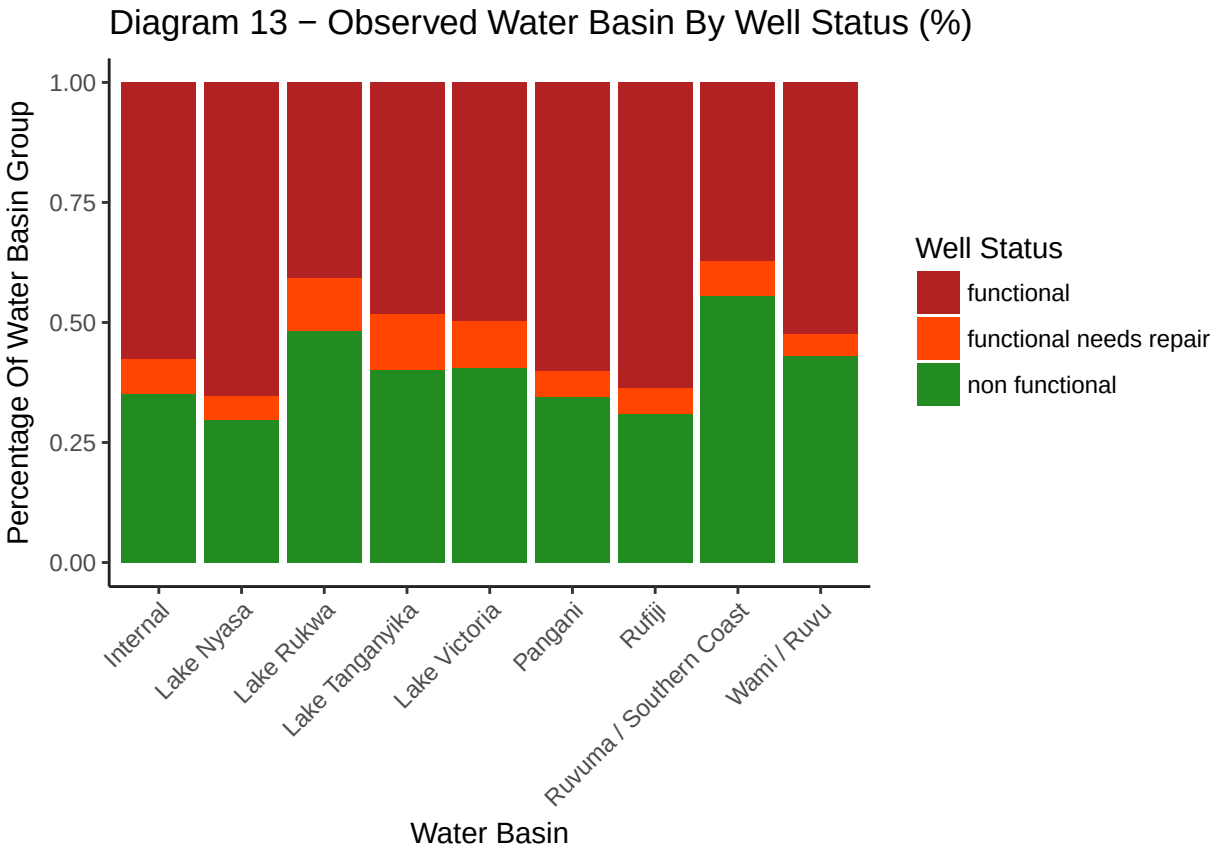
Water Source and Quality

Water Basin

The data in basin is clean and complete. There are 9 basins listed with a relatively even spread of observations across them. Lake Victoria is the basin with the most observations with 10,248 observations and Lake Rukwa has the least with 2454. The well status percentage chart by basin shows a lot of variation by basin with only 37.2% of 'Ruvuma / Southern Coast' basin wells being functional compared with 65.4% functionality for the 'Lake Nyasa' basin.

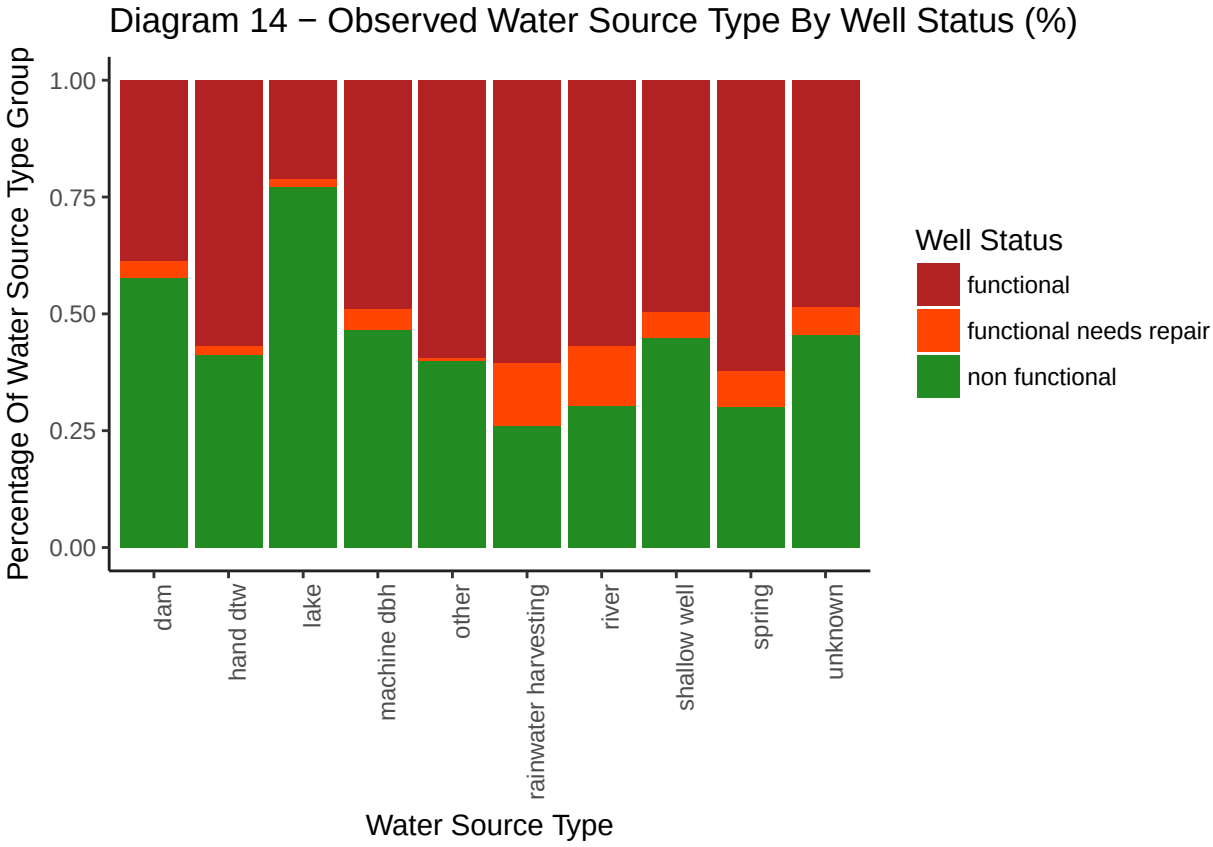
Table 15: Observations By Water Basin

Water Basin	Observations
Lake Victoria	10248
Pangani	8940
Rufiji	7976
Internal	7785
Lake Tanganyika	6432
Wami / Ruvu	5987
Lake Nyasa	5085
Ruvuma / Southern Coast	4493
Lake Rukwa	2454



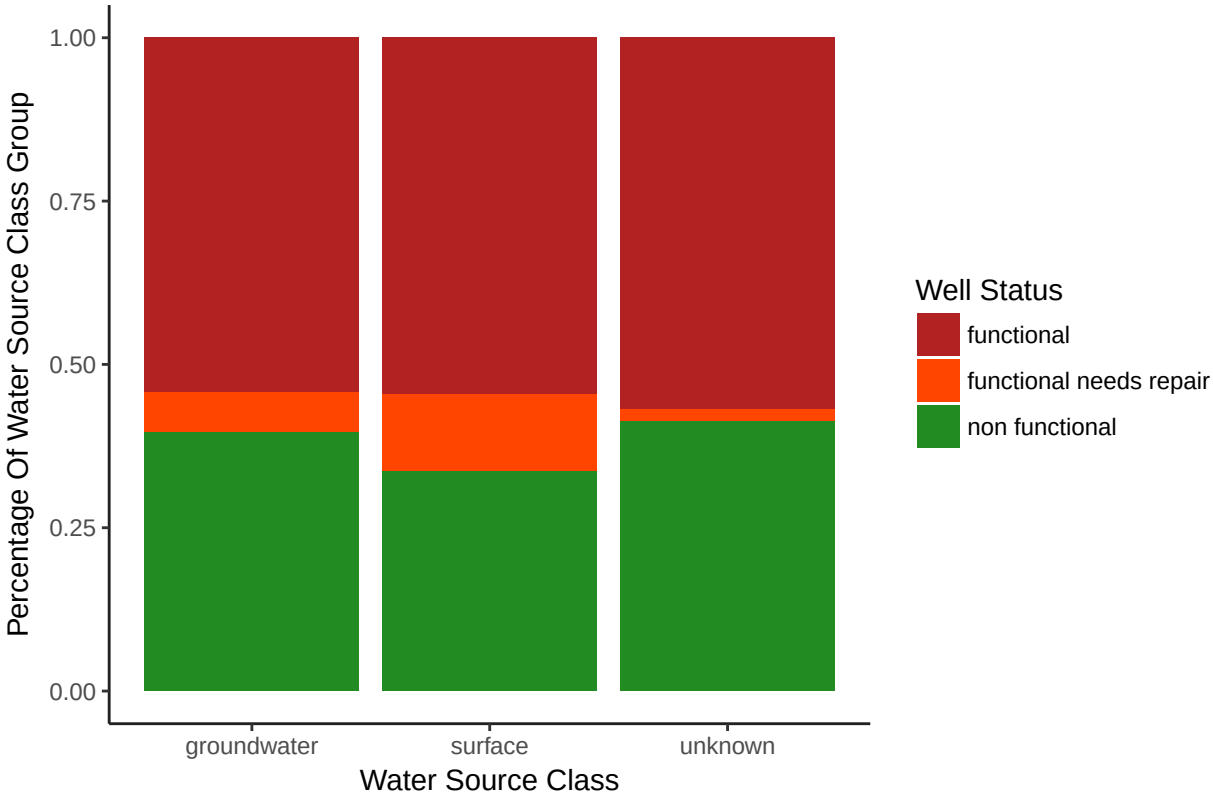
Source Of Water *source, source_type and source_class*

There are three variables that describe the source of the water. Source, source type and source class. They all contain clean data. Source is a little more detailed than source type, with boreholes split by whether they were dug with a machine or by hand, and rivers and lakes are combined in source type. Looking at a chart of source by percentage of status group, it is clear that there are differences between these split variables with Lake sourced wells especially having a much higher proportion of non-functional wells. Therefore the Source type variable was dropped.



The last source variable is an aggregation that describes if the source is from groundwater, surface or unknown.

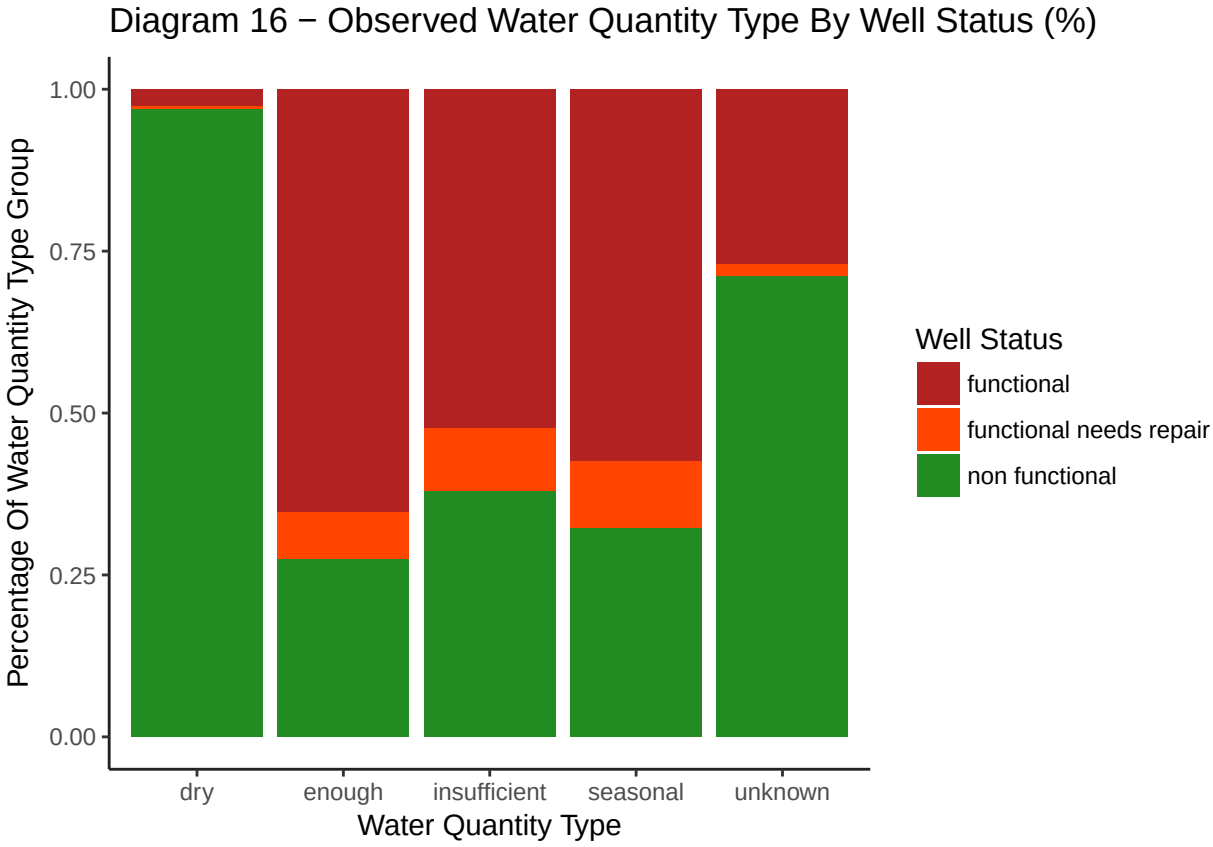
Diagram 15 − Observed Water Source Class By Well Status (%)



Quantity Of Water *quantity and quantity group*

The water quantity is described with categorical variables: dry; enough; insufficient; seasonal; or unknown. This data is duplicated across two variables quantity and quantity_group. As there is no difference, quantity_group was dropped.

A chart of the percentages of status group for each factor shows a lot of variation in well functionality, wells that have a quantity of "dry" are almost all non-functional.



Quality Of Water *water_quality and quality_group*

In the variable quality_group quality is described with categorical variables: colored; fluoride; good; milky; salty or unknown. The variable water_quality splits fluoride into fluoride and fluoride abandoned and does the same with salty, adding a salty abandoned category. Water_quality was dropped and a new variable called, quality_abandoned was created with a value of 1 for either salty_abandoned or fluoride_abandoned and 0 for all other categories.

Plots show that both variables have value in describing well functionality.

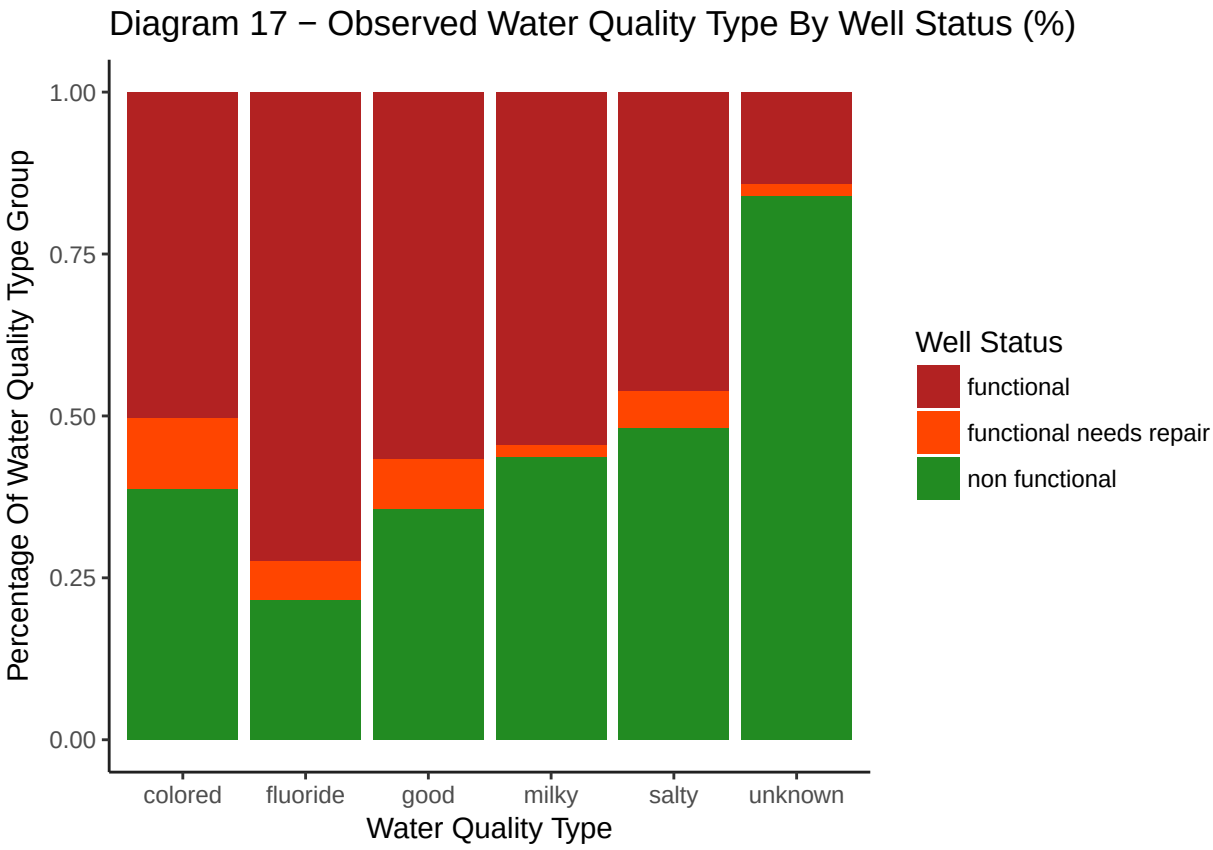
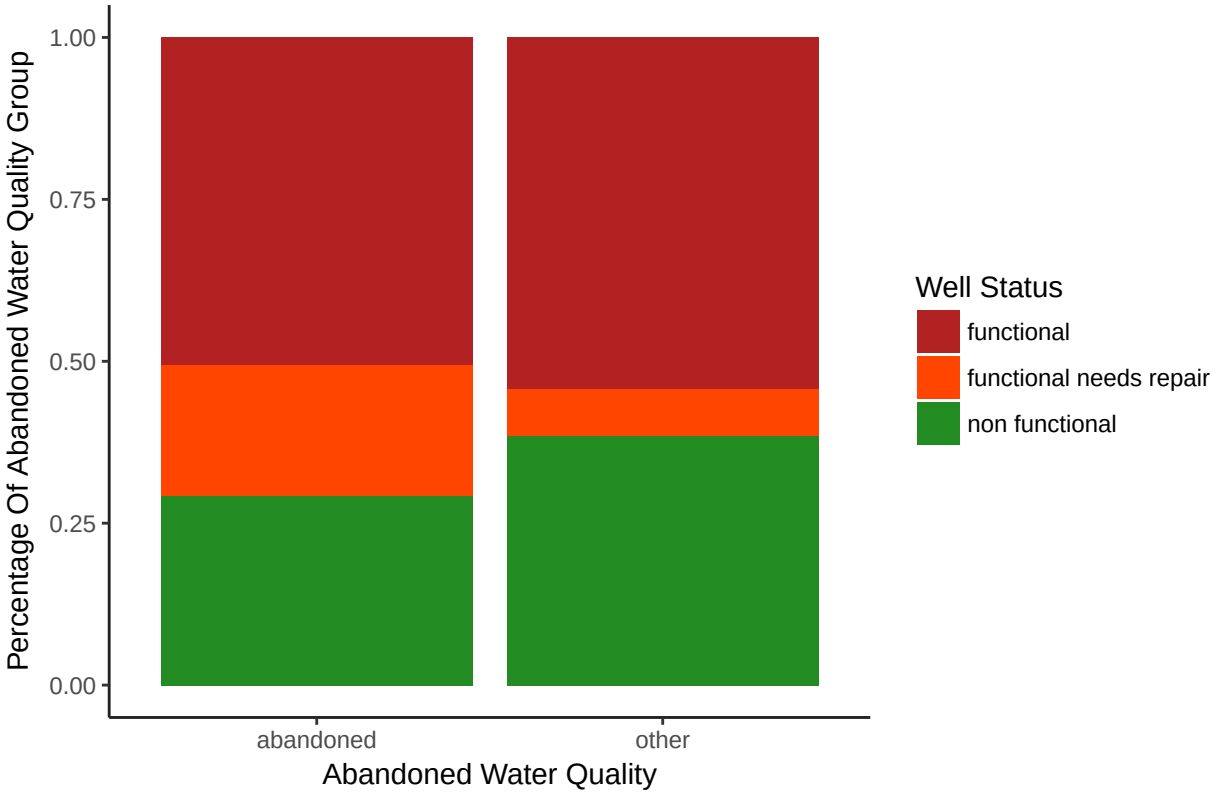


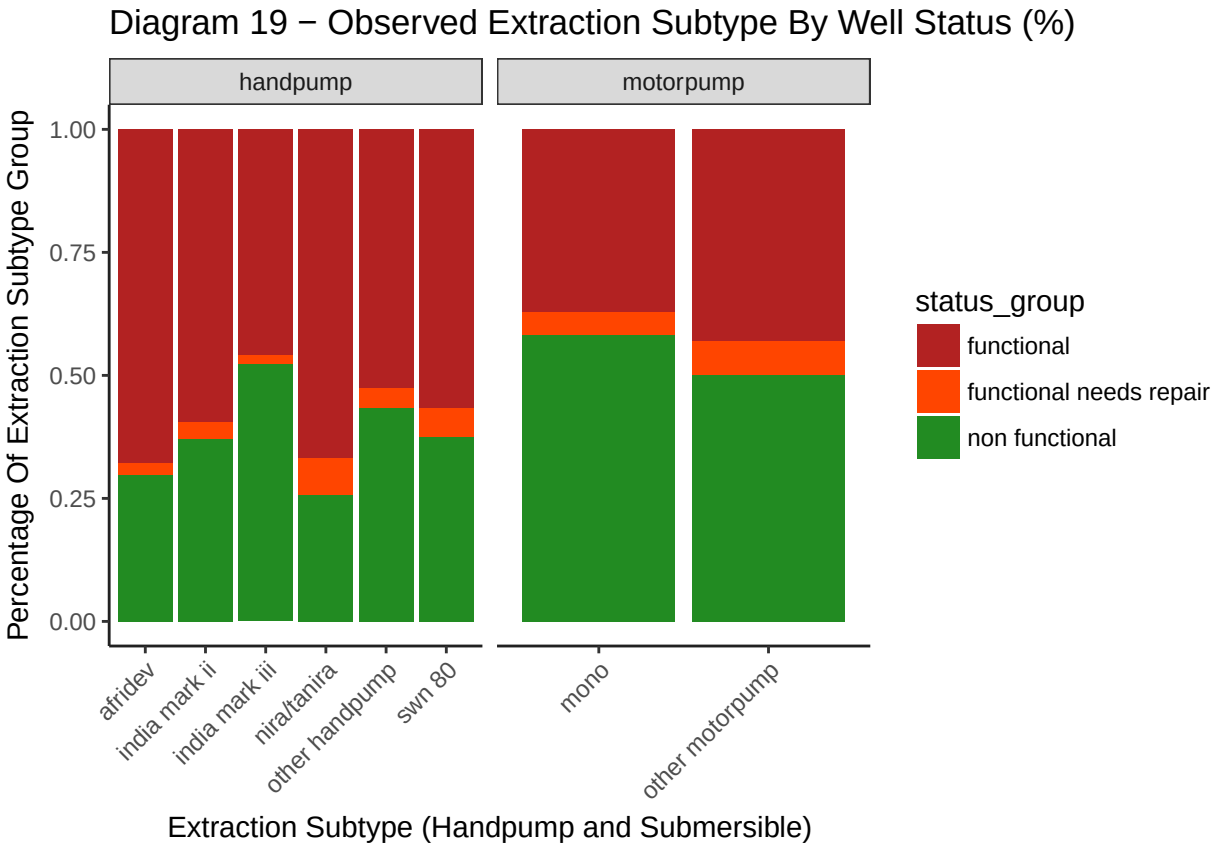
Diagram 18 − Observed Abandoned Water Quality By Well Status (%)



Well Type

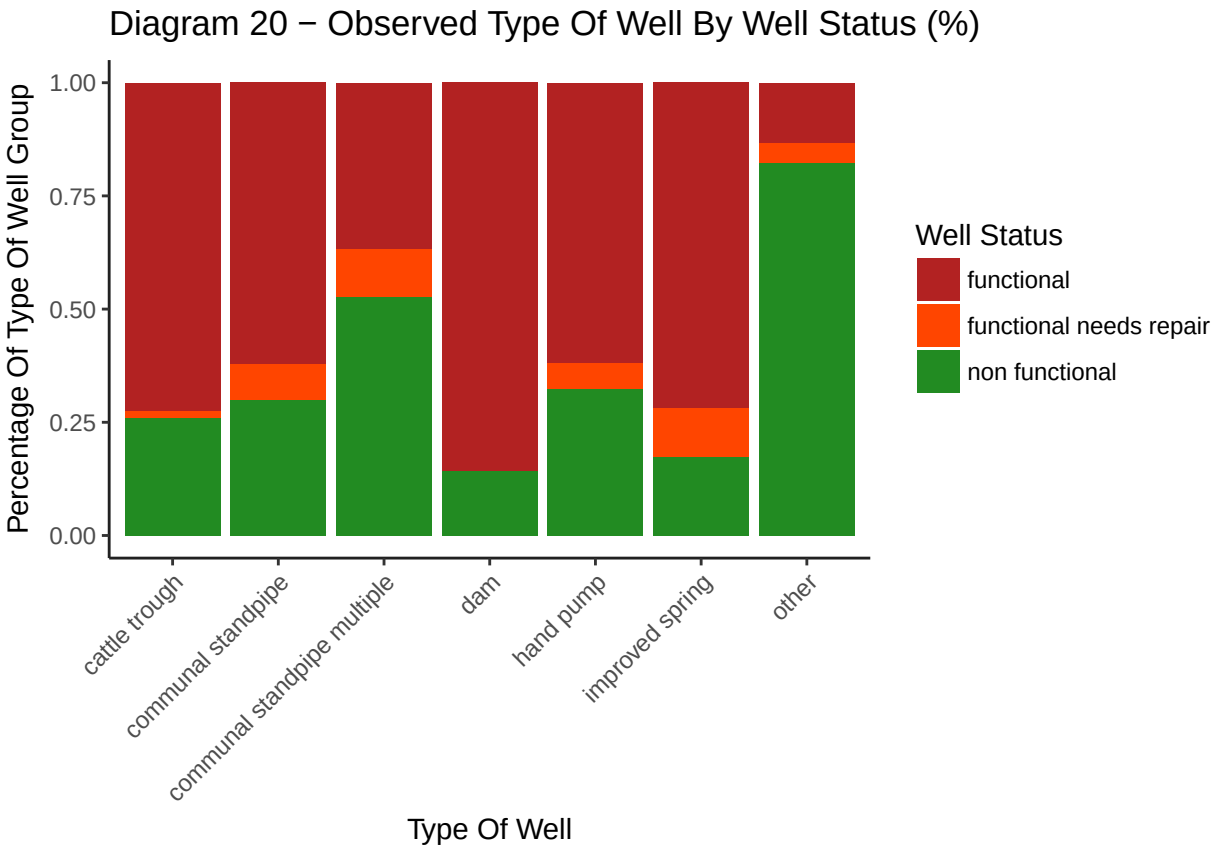
Extraction Method *extraction_type, extraction_type_group and extraction_type_class*

The extraction method describes how the water is drawn from the well. There are three variables that are available and from most to least detail they are extraction_type, extraction_type_group and extraction_type_class. Extraction_type_class describes the broad extraction methods as: gravity, handpump, motorpump, rope pump, submersible, wind-powered and other. Extra detail is provided in extraction_type_group over extraction_type_class, in the handpump category which has five factors related to brands and the in submersible category, which has two factors related to brands. In diagram 19 there is visible variation in the percentage of non functional wells across the different handpumps and motorpumps, therefore extraction_type_class was removed in favor of the more explanatory extraction_type_group. Extraction_type has only a few additional brand categories, each with only small numbers of observations. So for this reason it was also removed in favor of extraction_type_group.



Type Of Well *waterpoint_type and waterpoint_type_group*

How the well is set up is described in two variables, waterpoint_type and waterpoint_type_group. The only difference in these variables is that in waterpoint_type the communal standpipe category is split into communal standpipe and communal standpipe multiple. Both have a high number of observations, so the waterpoint_type_group variable was dropped. The categories in the remaining waterpoint_type variable are: communal standpipe; communal standpipe multiple; cattle trough; hand pump; improved spring; dam; or other.



Well Scheme *scheme_name* and Scheme Management *well_scheme_management*

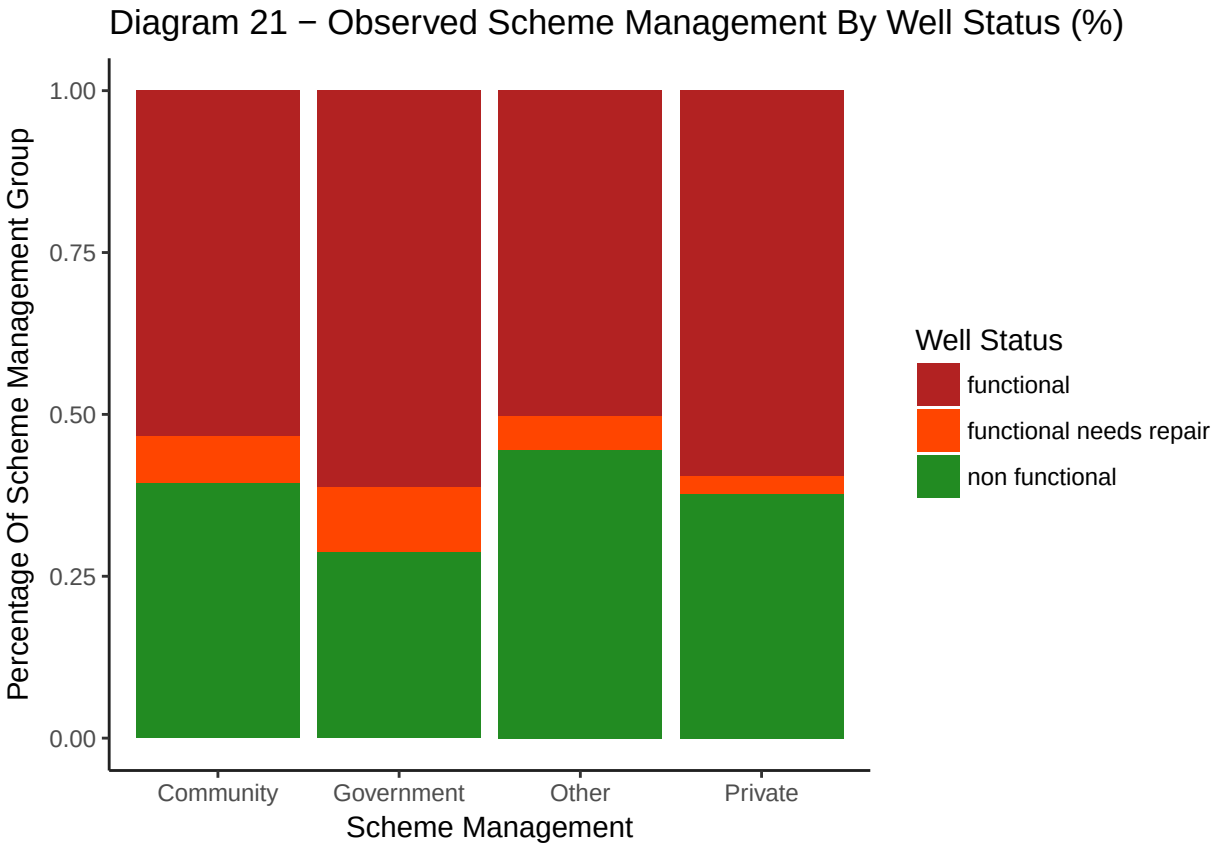
There are two variables related to the scheme that the well is built under. Well scheme names the scheme while well_scheme_management tells who the scheme owner is.

There are 13 categories of scheme management, including blank, 'None' and 'Other'. A new variable was created to grouping scheme management into "Other", "Government", "Private" and "Community". Table 16 below shows how the variable was transformed, and the new variable was called "scheme_management_broad". 'None', 'Other' and blank variables were combined into the value called 'Other'.

Table 16: Mapping Between Scheme Management and Grouped SCheme Management

scheme_management	scheme_management_broad
Other	Other
Company	Private
Private operator	Private
Trust	Community
SWC	Community
VWC	Community
WUA	Community
WUG	Community
Parastatal	Government
Water authority	Government
Water Board	Government

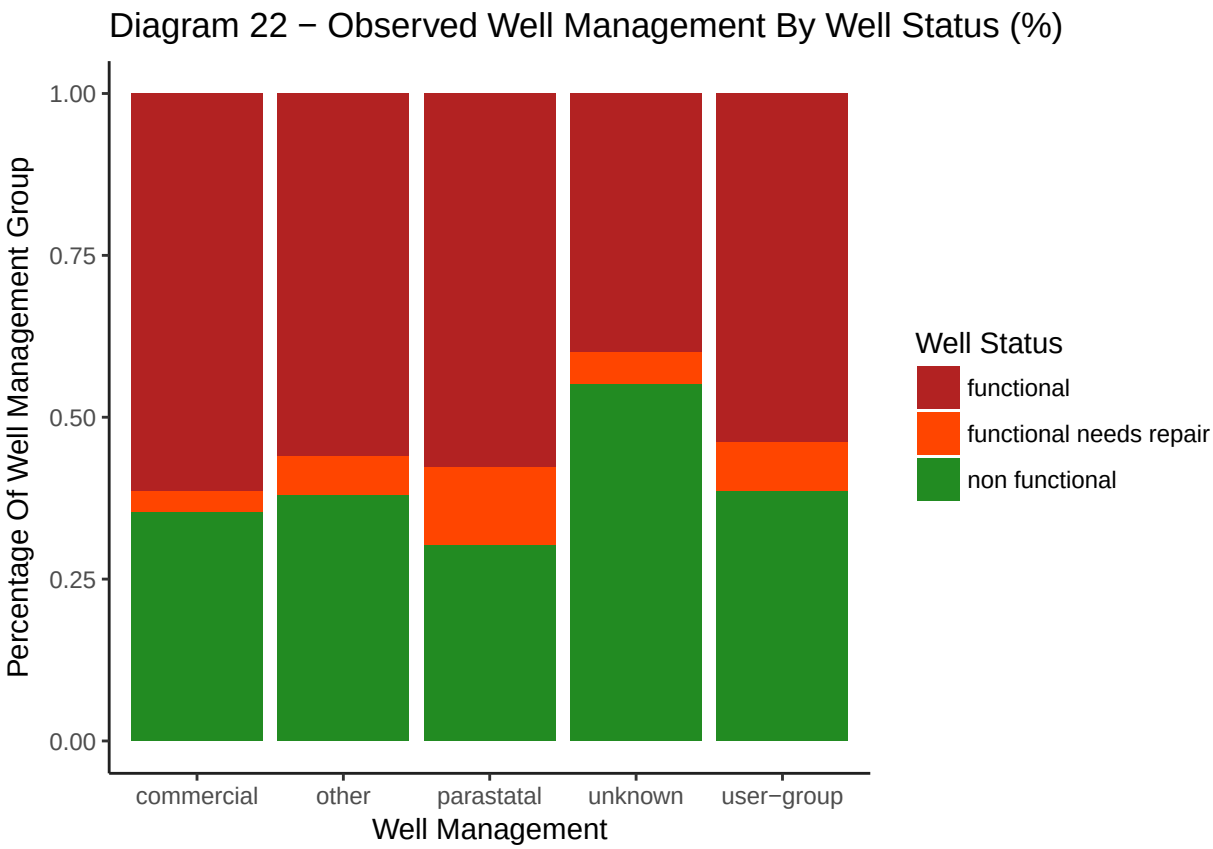
After cleaning, the scheme name has 2,468 categories. The largest category has 669 observations and most have very few observations, this variable was removed.



Well Operations

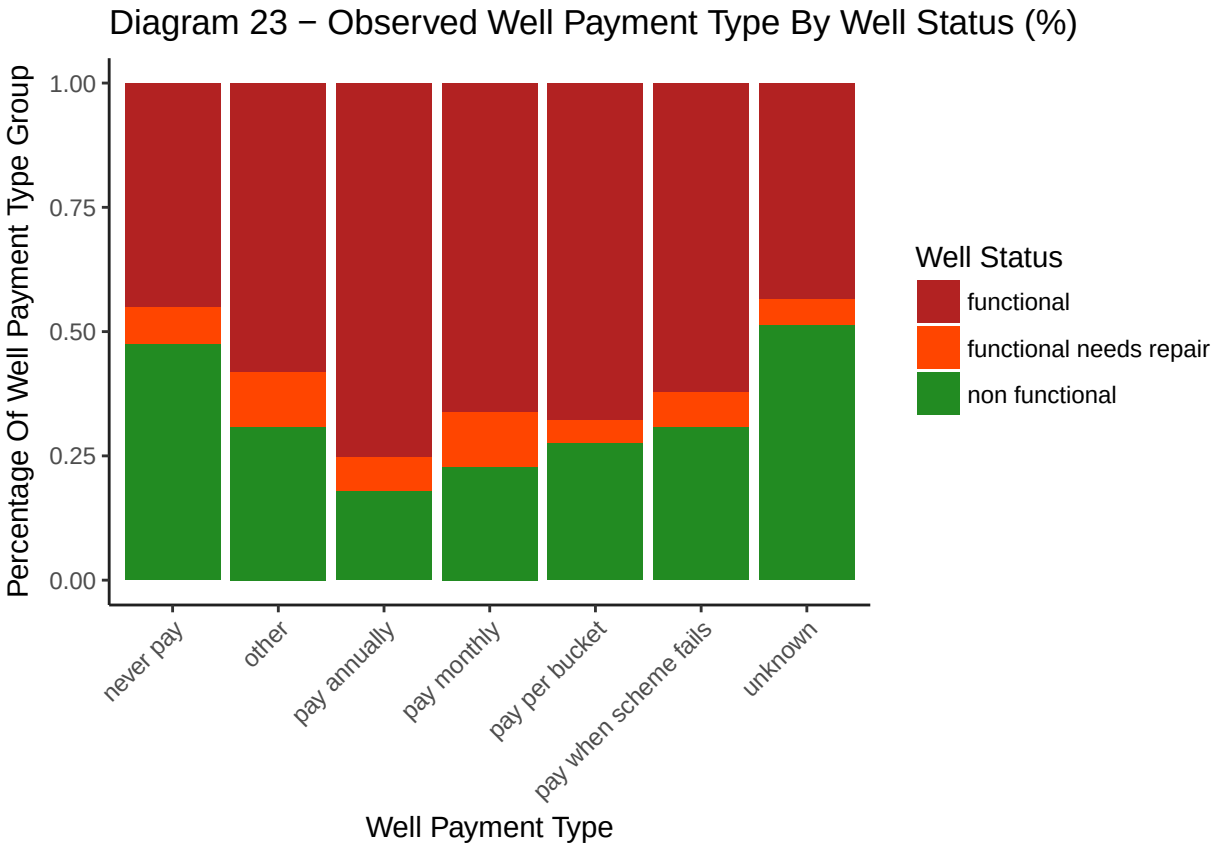
Well Management *well_management*

There is a variable related to the ongoing management of the well, it supplies five categories of well management: commercial; parastatal; user-group; unknown; or other. Wells with unknown management have the worst performance with 55% of them being non-functional.



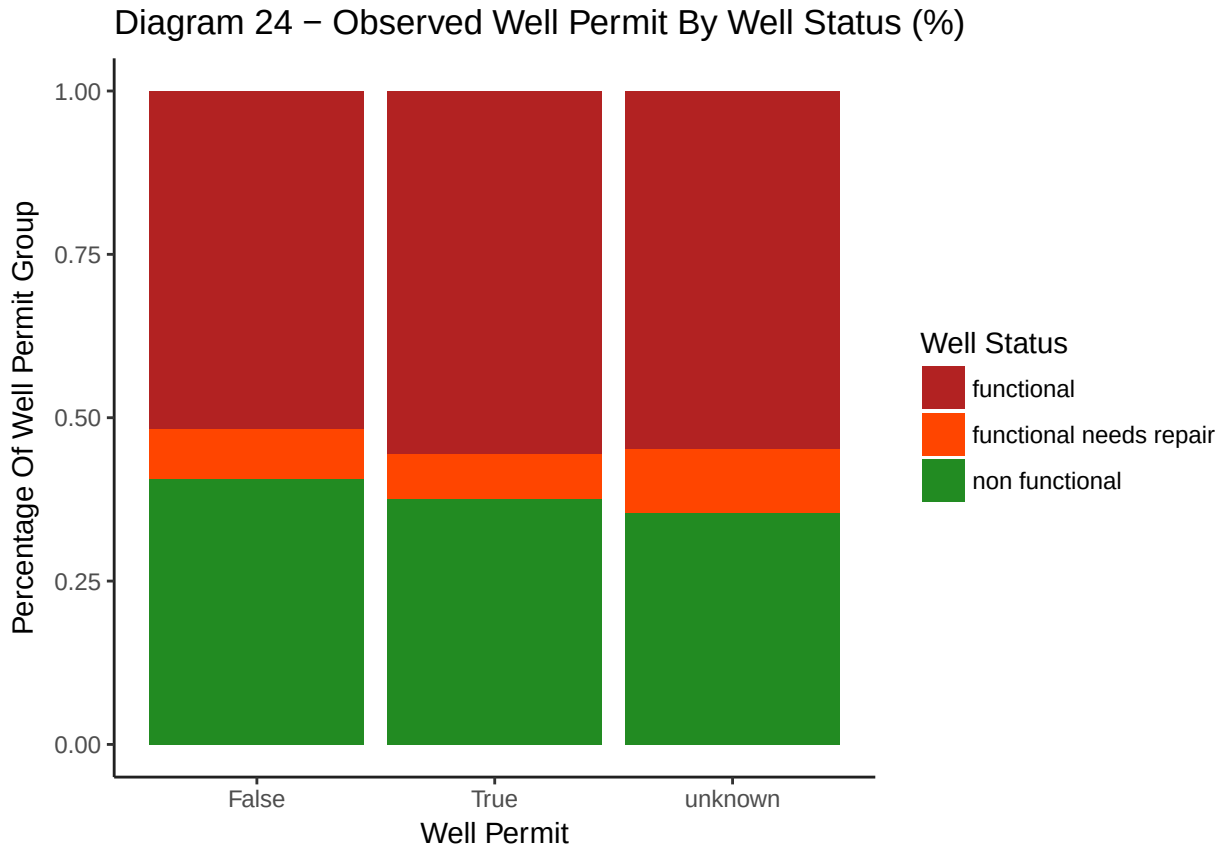
Well Payment *payment and payment_type*

The way that users pay for the well is recorded as: never pay; pay annually; pay monthly; pay per bucket; pay when scheme fails; unknown; and other. This variable is called payment. There is another variable that has the exact same data with slightly different category names called payment_type, this variable was dropped.



Permits *permit*

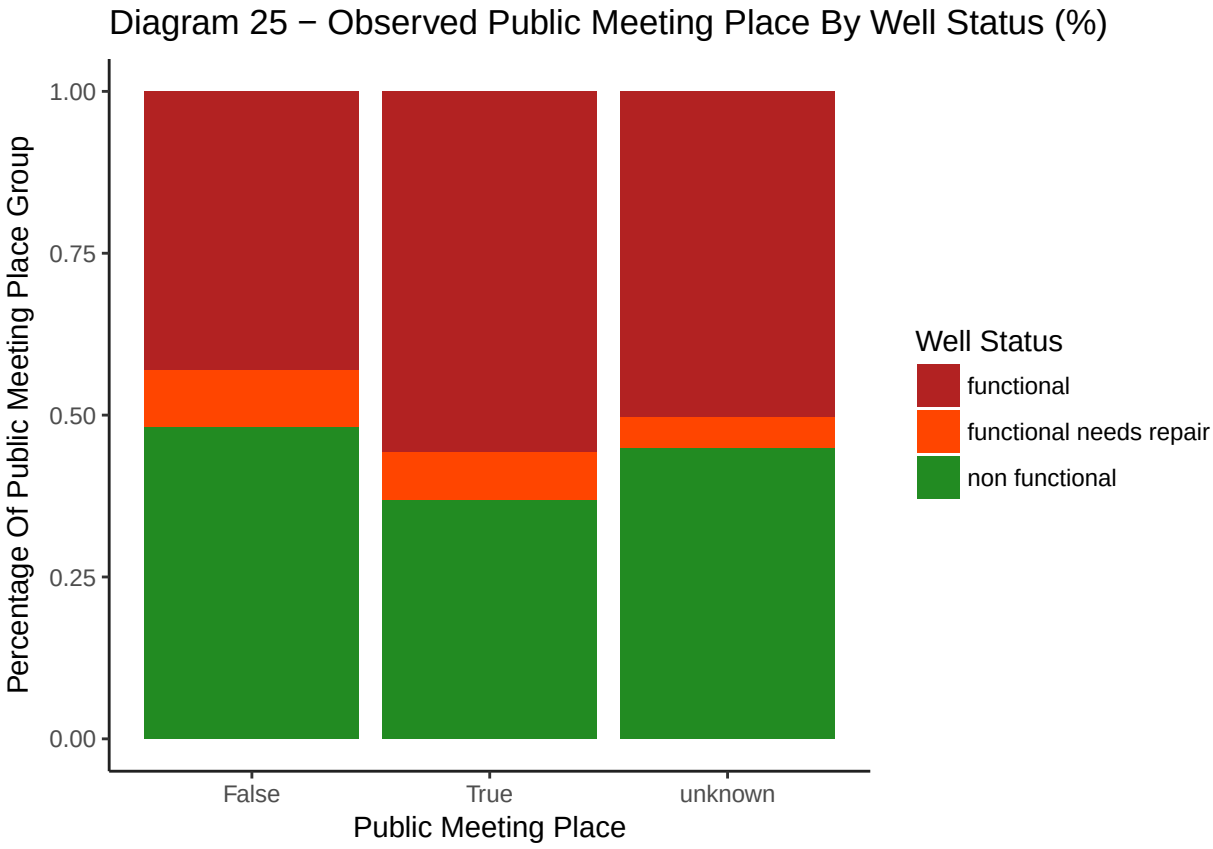
A binary variable indicates whether or not the well has a permit to operate. The 3056 observations where this was not recorded were cleaned to be classified as 'unknown', making the variable categorical.



Well Community

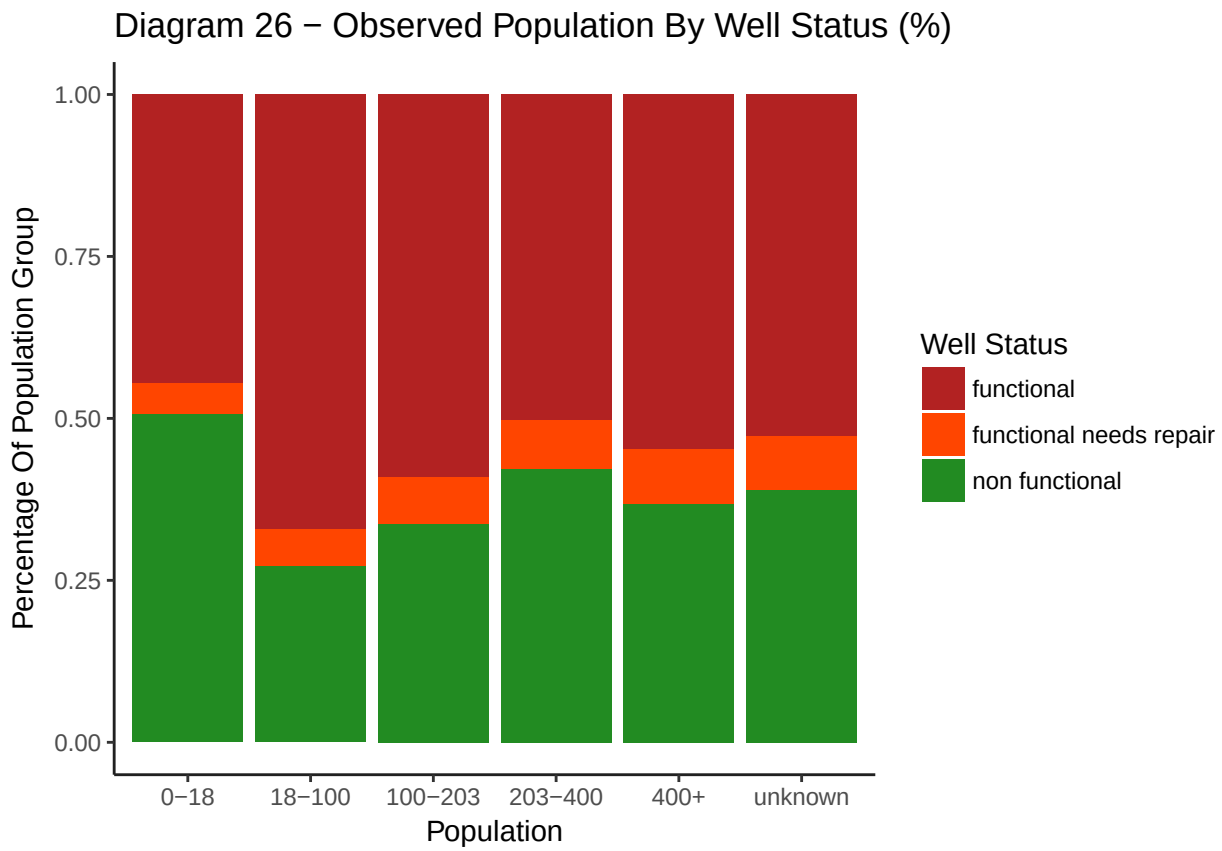
Public Meeting *public_meeting*

There is no details about the public_meeting variable. The assumption made is that it is TRUE when the well is in a public meeting place. Another option is that the well was approved at a public meeting. The variable is binary with 3334 blank variables, these are classified as 'unknown' again turning the binary variable into a categorical.



Population *population*

Population describes the number of people serviced by the well. There is a value of zero for almost one third of the observations which implies missing data. Therefore, a new variable with the existing data binned into five equal groups was created. The zero value data was labelled "unknown". The bins for the data are: 0-18; 18-100; 100-203; 203-400; 400+.



Other Data *num_private, recorded_by, wpt_name*

There were three other variables that were dropped. Num_private has no information on what the variable is; recorded_by has the same category "GeoData Consultants Ltd" for all responses; and wpt_name is the name of the specific water point name.

References

Australian Pump Technical Handbook (1987) *Pump Technical Terminology*, Accessed 08 September 2018, available: http://www.pumpapplicationengineers.com.au/files/pump-technical-terminology.pdf

DrivenData (2018) *Competition: Pump It Up: Data Mining The Water Table*, Accessed 08 September 2018, available: https://www.drivendata.org/competitions/7/pump-it-up-data-mining-the-water-table/

International Household Survey Network (2014) *Tanzania - Population and Housing Census*, Accessed 08 September 2018, available: http://catalog.ihsn.org/index.php/catalog/4618/download/58601

Japan International Cooperation Agency (2008) *Tanzania: So Much Water. So What's the Problem?*, Accessed 08 September 2018, available: https://www.jica.go.jp/english/news/focus_on/water/water_6.html

Wikipedia (2013) *Majengo - Wikipedia*, Accessed 08 September 2018, available: https://en.wikipedia.org/wiki/Majengo

What is my elevation? (2018) *What is my elevation?*, Accessed 08 September 2018, available: https://www.whatismyelevation.com/#