Predicting Well Functionality: Limitations of Survey Data

Hubert Luo and Amanda Wu

October 13, 2018

Introduction

Machine learning has previously been used to identify wells that might not be working, allowing for quicker repair of non-functional wells. Previous work in this space has focused on creating models using survey data, for example from information such as the source or type of the well. Using new datasets from Taarifa and geospatial data, we improve upon the existing work completed by Topor et. al on predicting well functionality, with a focus on Tanzania. We looked at seven categorical features and three numerical features across 122,076 wells and found that the most important features in predicting well functionality were water quantity, distance to the nearest road, extraction type, and construction year. When we combined this survey data with just a single non-survey-based feature, the distance to the nearest road, there was an improvement in the random forest model's performance from 77.49% accuracy to 78.48% accuracy, illustrating the potential for data unrestricted to surveys in improving prediction abilities.

Data Summary

The total number of features examined was ten, with seven categorical features and three continuous features. There was also one response variable, which was categorical, the status of the well. These were all the unique features which could be applicable to wells in any country, so features such as the state the well was located were excluded. There was a total of 122,076 unique wells in the combined dataset.

Categorical Variables

The table below summarizes the categorical variables: water quantity, extraction type, waterpoint type, payment type, source, water quality, management type, and well status in the tables below. Note the last attribute in each row, specified (other), includes wells of known attribute that do not fall into one of the attributes listed above.

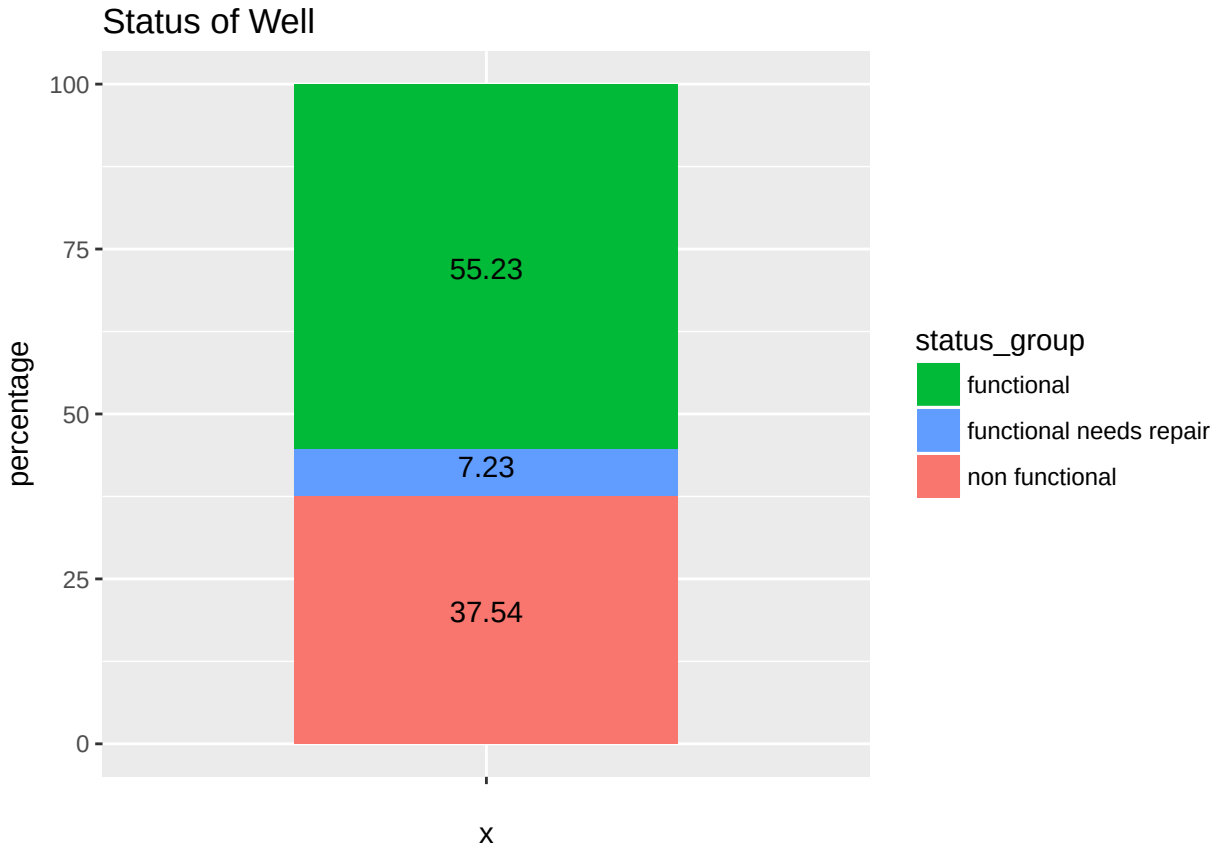
quantity	extraction_type	waterpoint_type	payment
dry :11574	gravity :55362	unknown :62676	annually : 7371
enough :68955	nira/tanira:17236	communal standpipe :28522	monthly :17143
insufficient:31572	other :13097	hand pump :17488	never pay :52234
seasonal : 8333	submersible:11230	other : 6380	on failure: 7902
unknown : 1642	swn 80 : 7541	communal standpipe multiple: 6103	other : 2261
NA	mono : 5311	improved spring : 784	per bucket:17769
NA	(Other) :12299	(Other) : 123	unknown :17396

source_type	quality_group	management_group	status_group
borehole :23473	colored : 1012	user-group :52490	functional :67422
dam : 1400	fluoride: 463	vwc :41571	functional needs repair: 8829
other : 588	good :104552	wug : 7099	non functional :45825
rainwater harvesting: 4863	milky : 1676	parastatal : 3724	NA

source_type	quality_group	management_group	status_group
river/lake :21658	salty : 10505	commercial : 3638	NA
shallow well :35050	unknown : 3868	water board: 3430	NA
spring :35044	NA	(Other) :10124	NA

Summary of Well Status

As mentioned in the introduction, each well has 3 possible statuses, as summarized below. 37.5% of wells are non-functional, while 7.23% are functional and in need of repair and the remaining wells (55.2%) are functional without the need of repair. Figure 1.1 demonstrates the overall well distribution that fall into the categories of either functional, non-functional, or functional but in need of repair.



Continuous Variables

The tables and graphs below describe two continuous variables from the Taarifa survey data (population and construction year), and one newly computed variable, the distance to the nearest road from a well, that was computed from a road dataset using latitude and longitude. The data is right skewed, with a long right tail as the majority of wells have surrounding populations of less than 1,000 while some have populations that far exceed 1,000 and reach as high as 30,500. The standard deviation is also extremely large at 573.64 even though 75% of the data is less than or equal to 320, demonstrating the significant effect the extremely high population values have on our data.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    40.0   150.0   283.7   320.0 30500.0
```

Therefore, we binned the population data into 10 bins of size 100, ranging from 0 to 1,000 and putting all data points exceeding 1,000 into the final bin of 900 to 1,000 in order to reduce the disproportionately large affect these extreme values played in our analysis. The final summary counts and a histogram are displayed below.

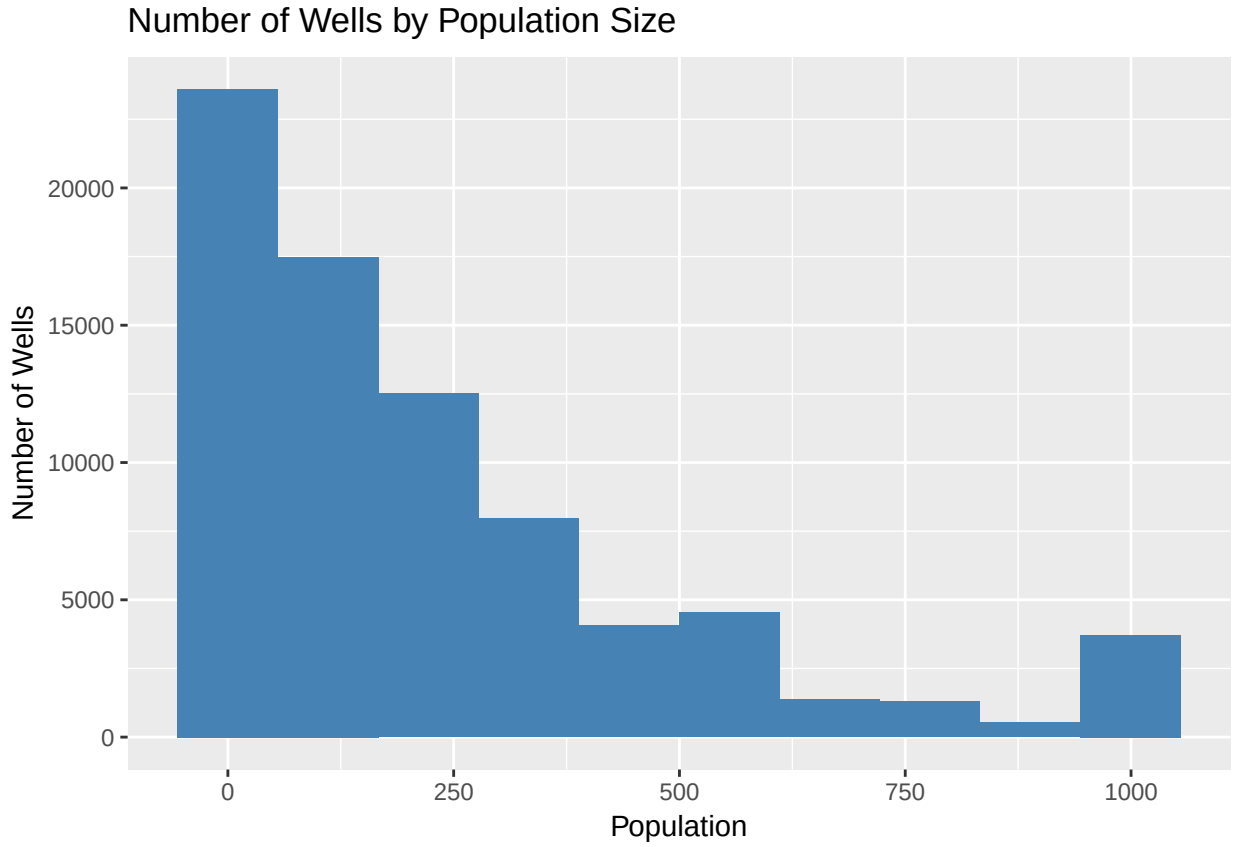


Table 3: Number of Wells by Population Size

	Number of Wells
(0,100]	32180
(100,200]	14047
(200,300]	10855
(300,400]	6185
(400,500]	4404
(500,600]	2422
(600,700]	1348
(700,800]	1275
(800,900]	605
(900,1e+03]	3765

The wells ranged from being constructed in 1960 to 2014, with a median of 2000.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1960    1988    2000    1997    2008    2014
```

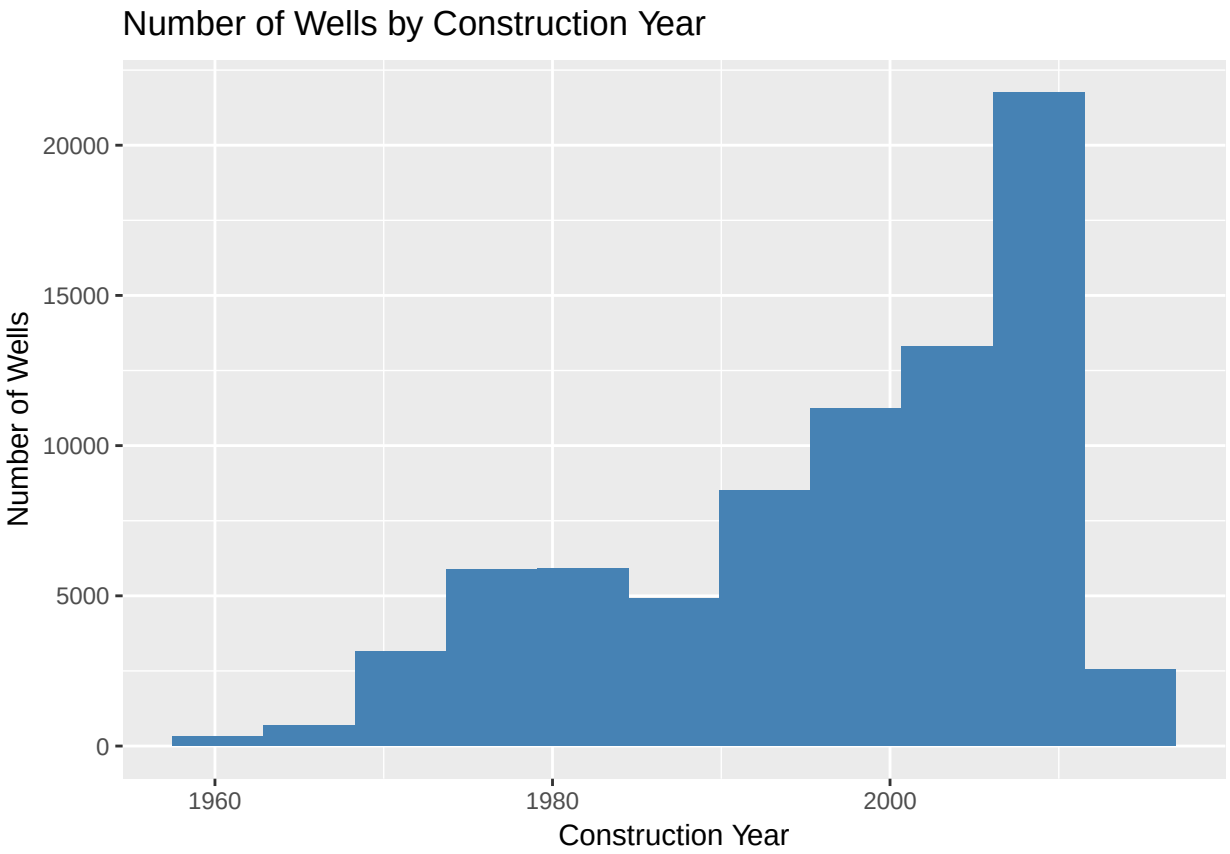


Table 4: Number of Wells by Construction Year

	Number of Wells
(1960,1965]	410
(1965,1970]	1355
(1970,1975]	4430
(1975,1980]	5245
(1980,1985]	6149
(1985,1990]	4954
(1990,1995]	6578
(1995,2000]	11234
(2000,2005]	10285
(2005,2010]	22259
(2010,2015]	5069
NA's	216

The data is extremely right-skewed, as the mean of 54.72km exceeds even that of the 75th percentile of 7.88km. As a result, we capped the distances at 50km and assigned all distances greater than 50km a value of 50km.

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.000    0.470    2.474   54.720    7.470 3337.890
```

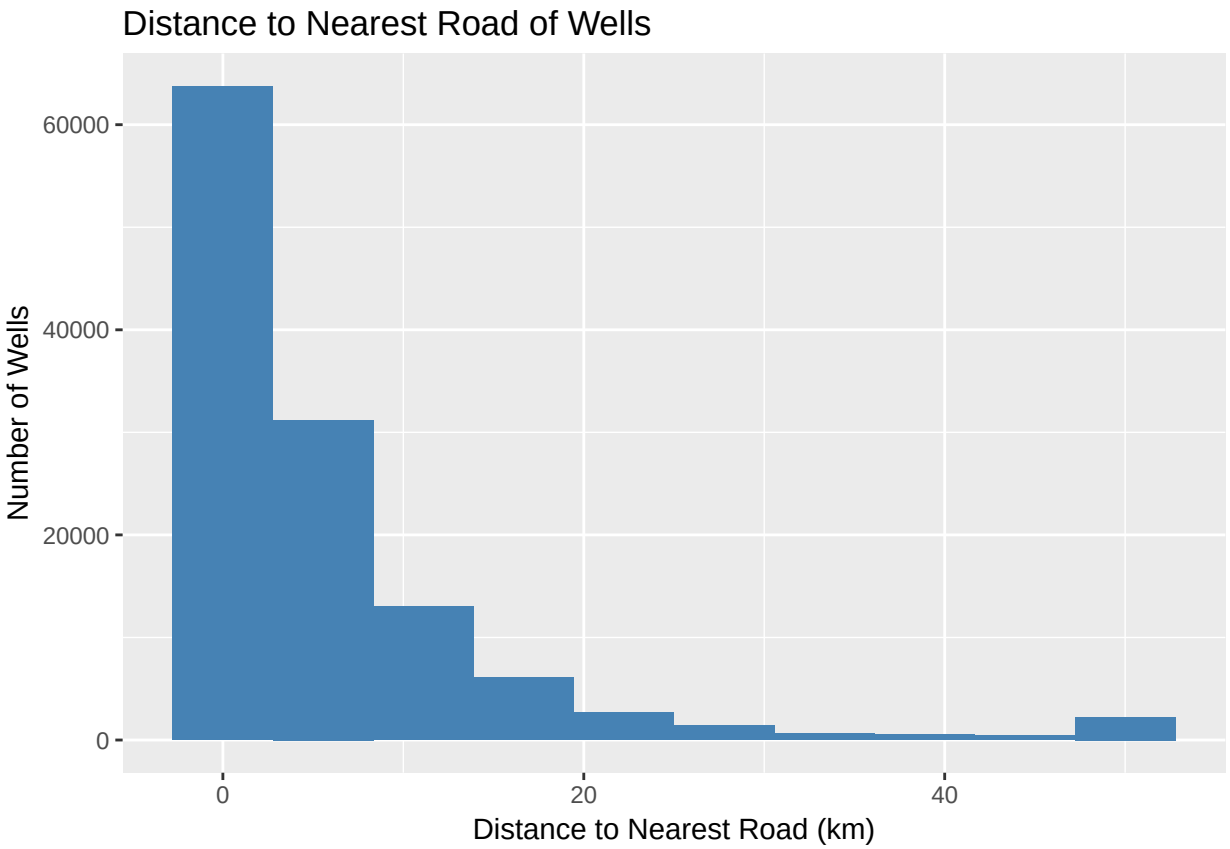


Table 5: Distance to Nearest Road for Wells

	Number of Wells
(0,5]	79295
(5,10]	20379
(10,15]	9693
(15,20]	4824
(20,25]	2347
(25,30]	1334
(30,35]	584
(35,40]	548
(40,45]	494
(45,50]	2338
NA's	240

Missing Data

Most of our unknown categorical data is missing at random, so keeping these values in our dataset should not create bias. However, water quality and water quantity have a large number of missing values as well as a higher-than-average non-functioning well fraction. Thus, there is evidence that missing values associated with these 3 categories need to be investigated further or handled through classification later.

A table summary of the percentage of missing data for each variable is listed below. Note that waterpoint type for all of the data from the Taarifa API was missing so it had the highest percentage of missing values at 51.34%. Population and construction year had missing values around 35%, while the rest of the variables

were below 15%. Source, well distance to nearest road, extraction type, management, and water quantity had the lowest percentages of missing values (all below 1.5%).

For missing continuous data such as population, year, and distance, we replaced the missing values with the median, as the large number of outliers meant that the median was more meaningful than the mean for our data.

Table 6: Percentage of Missing Values for Datasets

	Overall	Dataset 1	Dataset2
quantity	1.35	1.33	1.36
extraction_type	0.00	NA	0.01
waterpoint_type	51.34	NA	100.00
construction_year	35.95	34.86	0.00
payment	14.25	13.73	14.74
source	0.00	0.00	0.00
population	36.85	35.99	37.63
quality	3.17	3.16	3.18
management	0.93	0.94	0.92
distance_to_nearest_road	0.00	0.00	0.00

Comparing Datasets

We compared the coordinates of the wells in the two datasets to identify wells which were present in both datasets. We found 10,249 overlapping wells, resulting in a total of 122,076 unique wells in the merged dataset.

38.42% of the wells in the first dataset were non-functional, compared to 37.7% in the second dataset. From the table of missing values above, we see that the largest difference was in waterpoint type as mentioned earlier (missing for Dataset 2, all present for Dataset 1). In addition, population, payment, and construction year were the features that had a higher proportion of missing values in the second dataset than the first one.

Feature Selection

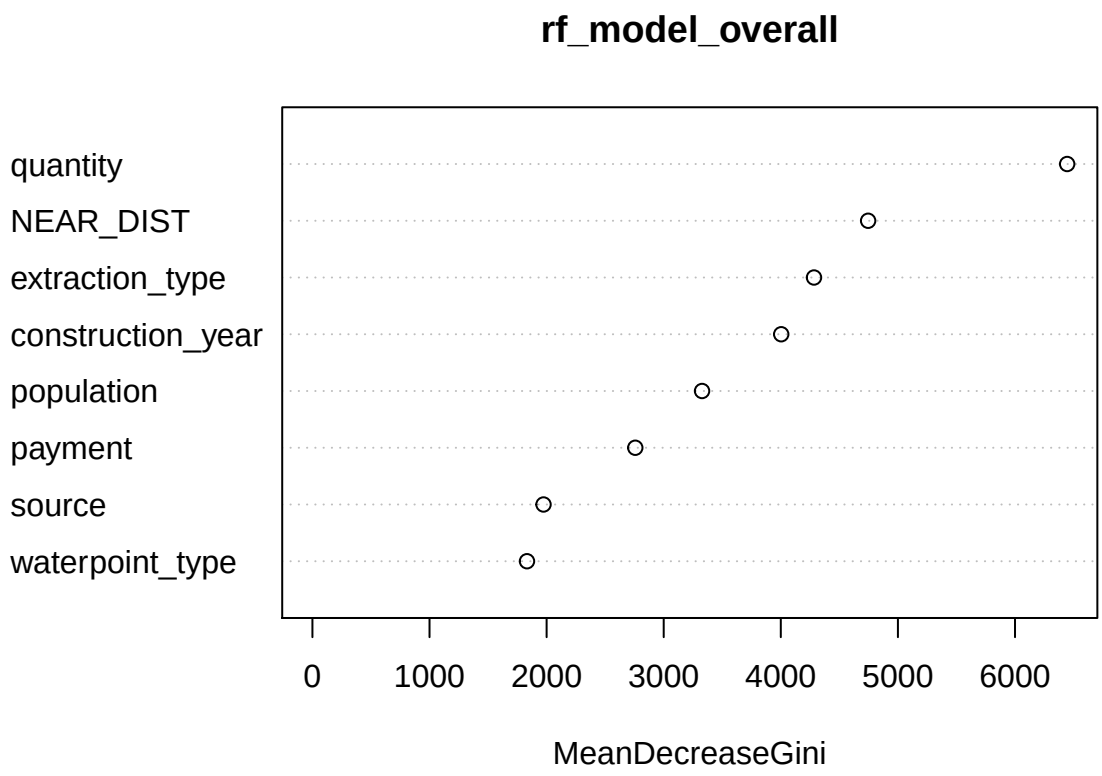
We first narrowed down the list of features to only include those that appeared in both datasets and were also meaningful for future analysis on different datasets. For example, we dropped region and latitude/longitude as any conclusions provided by these features from Tanzania were not broadly applicable to different situations. An exception was made for waterpoint type - we decided to investigate it as a feature even though it wasn't in both datasets as previous analysis by Topor et. al [1] had identified it as an important predictive feature.

To analyze the importance of features, we used random forest methods to evaluate the importance of each feature by its mean decrease in Gini coefficient [2]. The results are outlined in the table and graph below.

Table 7: Feature Importance by Mean Decrease in Gini

	Overall
quantity	6445.423
waterpoint_type	1832.474
extraction_type	4283.320
construction_year	4003.432
payment	2757.132
source	1973.522
population	3327.451

	Overall
NEAR_DIST	4745.759



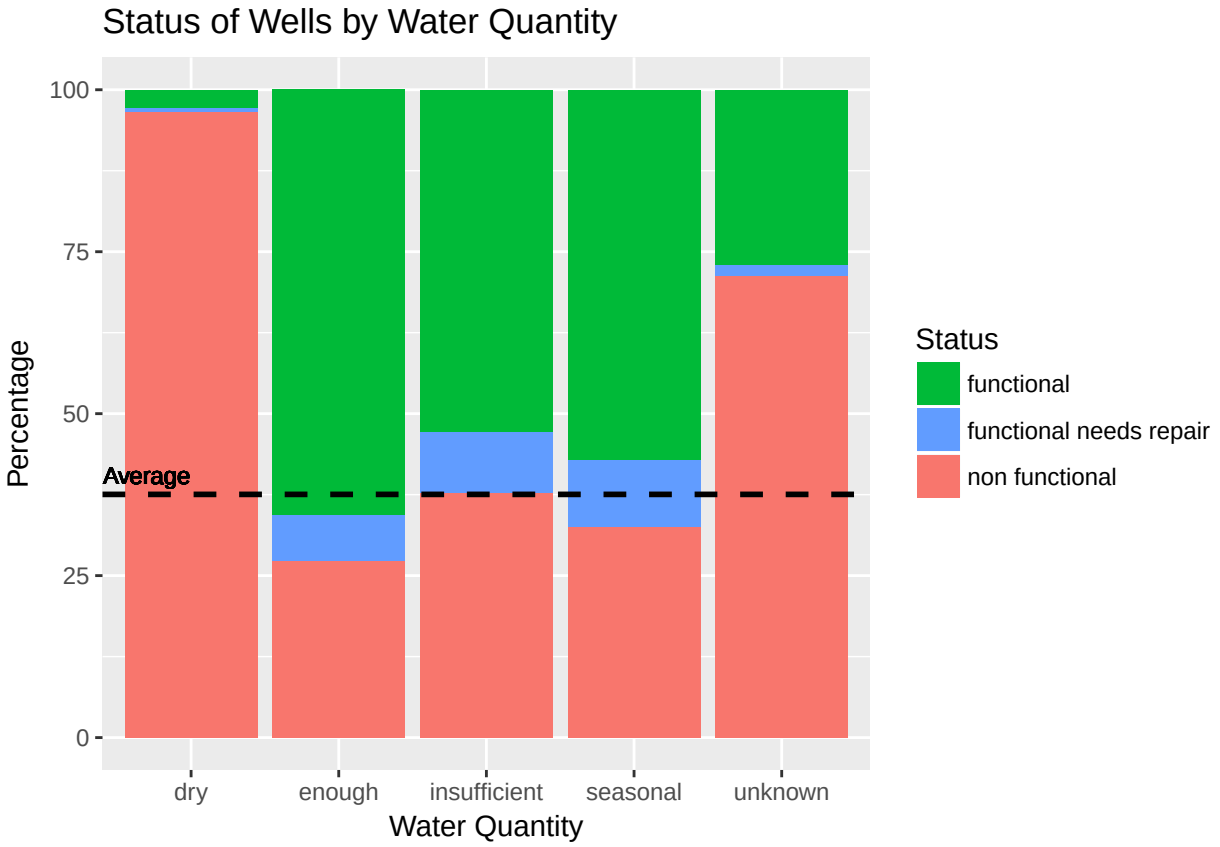
As seen above, the most important feature was quantity, followed by distance to the nearest road, extraction type, and construction year. It is interesting to note that distance to the nearest road, a feature absent in existing models in this field, was found to be the second most important feature, demonstrating the strong potential for further work using alternative, non-survey data.

Note that this list of the most important features is intuitive. For example, if a well is old and not producing sufficient water, it is likely to be non-functional. Due to the limitations of conducting a one-time field survey, we only have data from a single point in time. As a result, we do not know whether any changes have occurred to the well since it was last surveyed. Concrete action is therefore difficult to recommend. More timely and regular measurements of the well would increase the amount of data and further improve the accuracy. Although traditional field surveys are limited by the number of people that physically visit wells, sensor-based technologies could provide real-time measurement of well functionality and usage.

Predictive Features

We then examined each of these features in further depth below. The black dotted line on the graphs marks the overall average percentage of non-functional wells (38.02%). Red bars that extend above this dotted line demonstrate a higher than average likelihood to be non-functional, and red bars which are below the dotted line are less likely to be non-functional.

1. Water Quantity



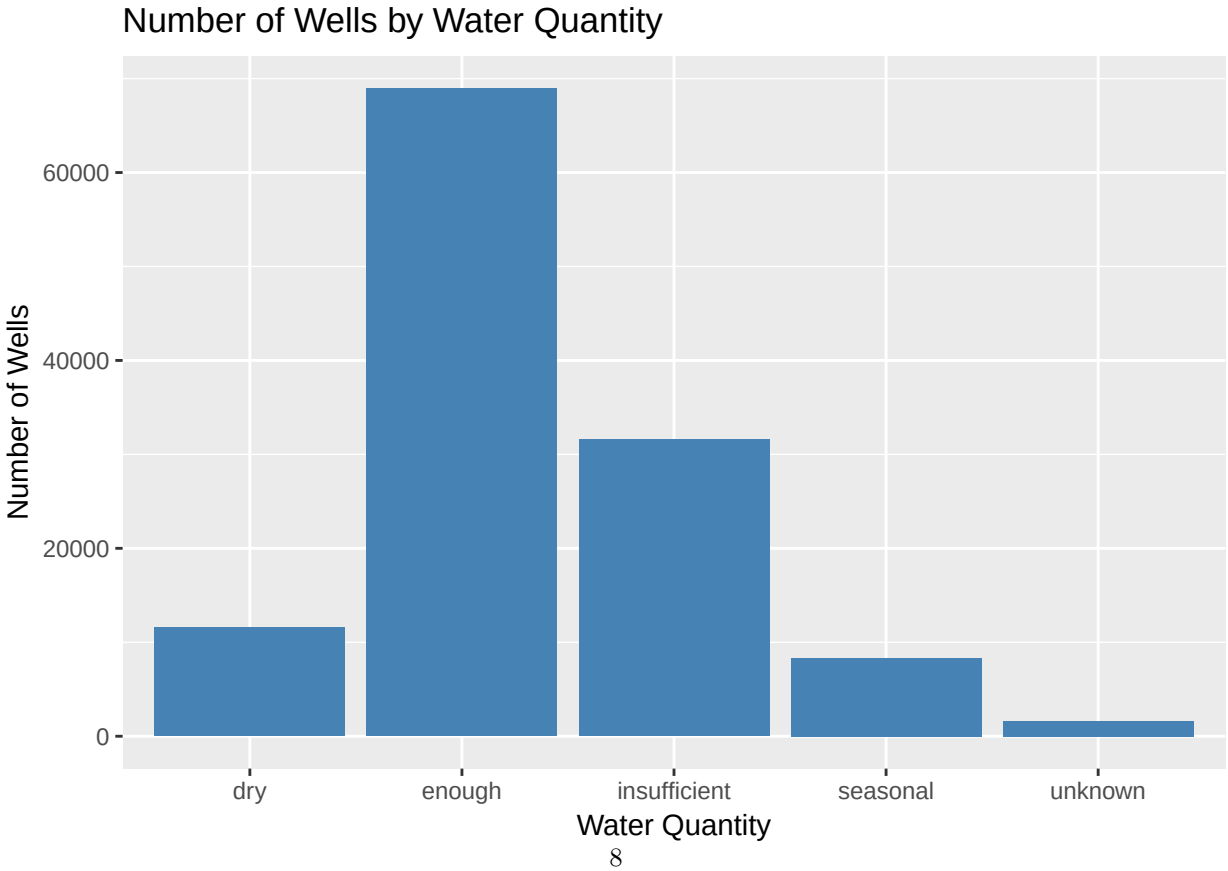
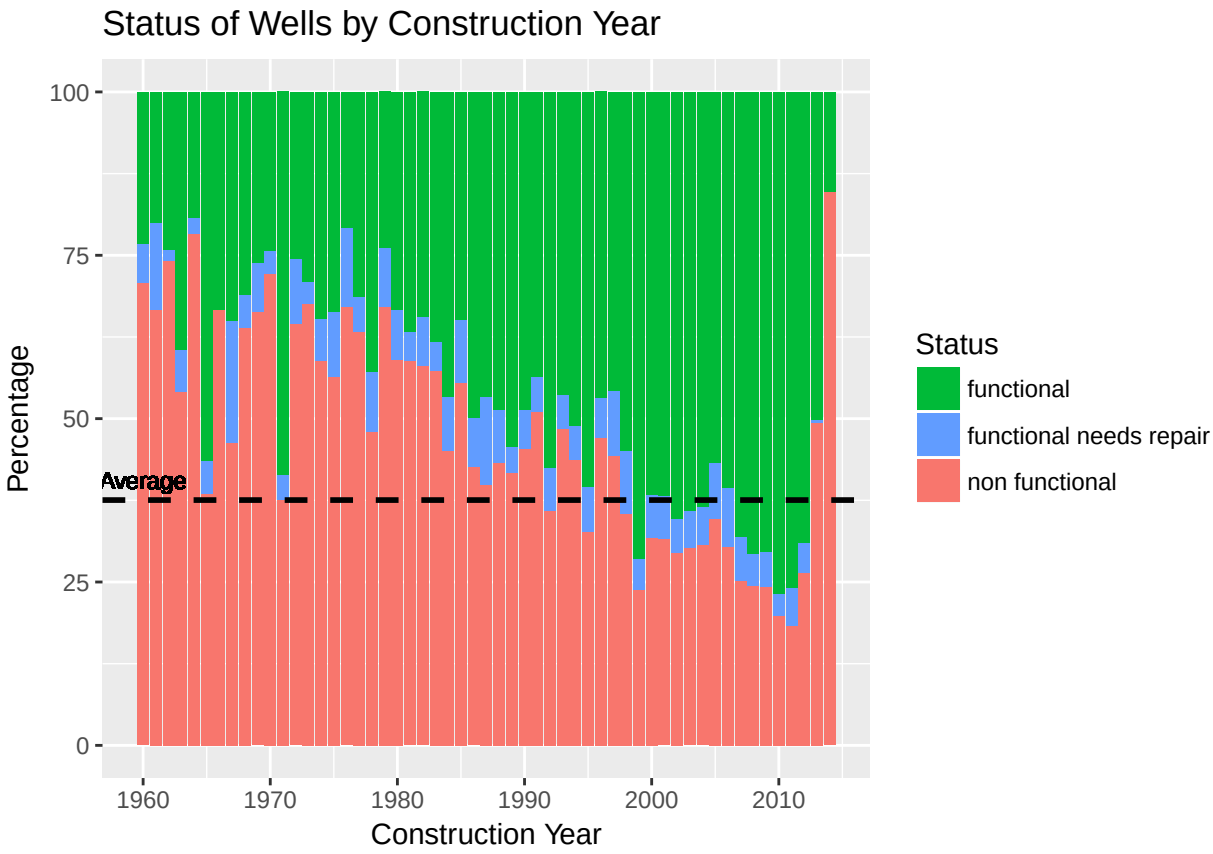


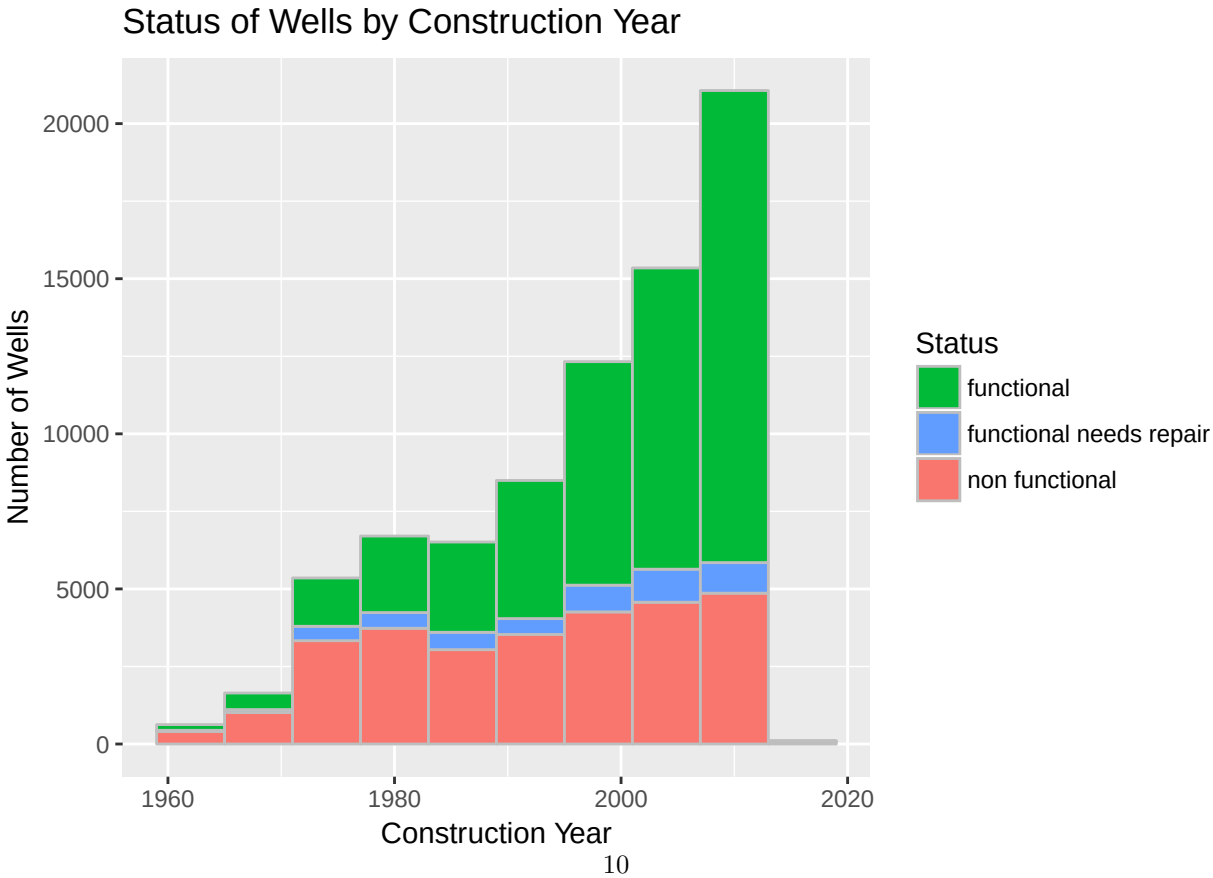
Table 8: Percentage of Non-Functional Wells by Water Quantity

Water Quantity	Number of Non-Functional Wells	Percentage of Non-Functional Wells
dry	11183	96.62
unknown	1169	71.19
insufficient	11945	37.83
seasonal	2714	32.57
enough	18814	27.28

It is noteworthy that 96.6% of dry wells are non-functional, while just 27.3% of wells with enough water quantity are non-functional. Clearly, wells with dry or insufficient water quantity are much more likely to be non-functional than those with seasonal or enough water quantity, and thus water quantity is one of the most significant features we can use to predict whether a well is functional or not.

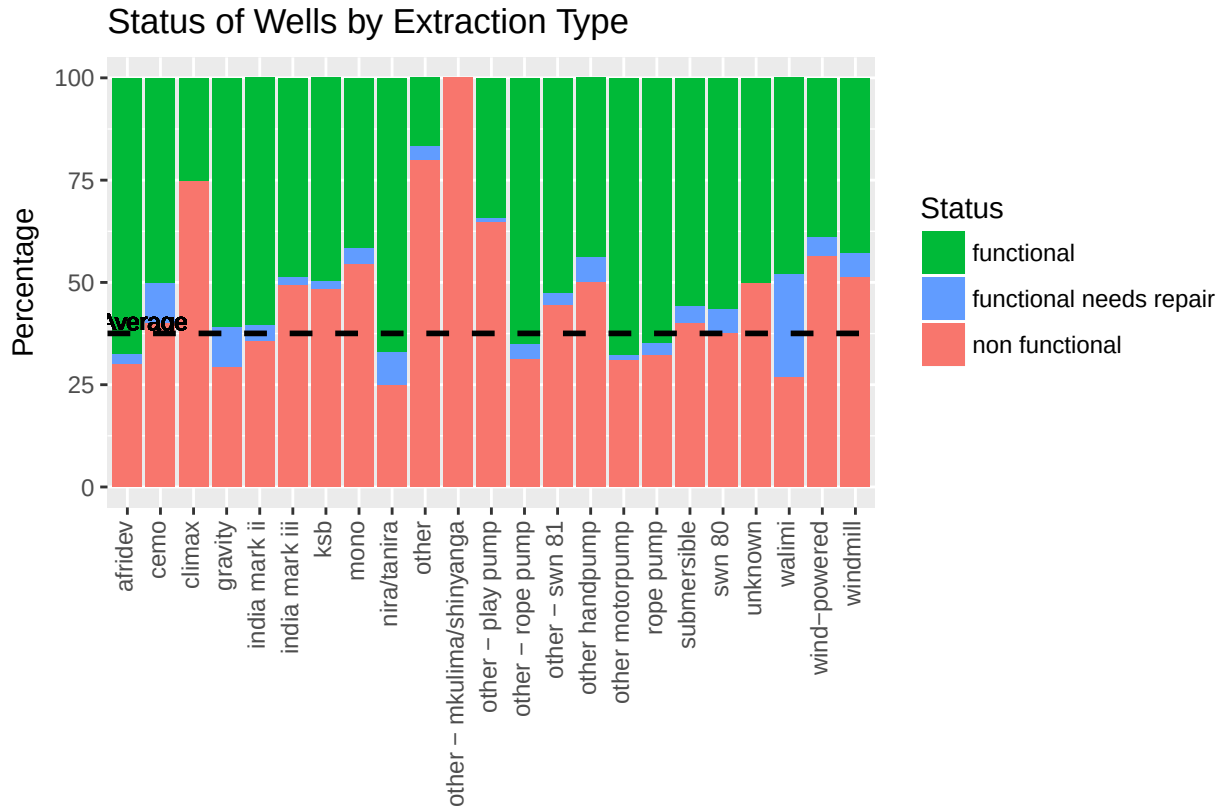
2. Construction Year





Wells that are constructed later are generally more likely to be functional.. However, only looking at the construction year does not take into account when the well has been last serviced, which would also be important in determining whether a well is functional.

3. Extraction Type



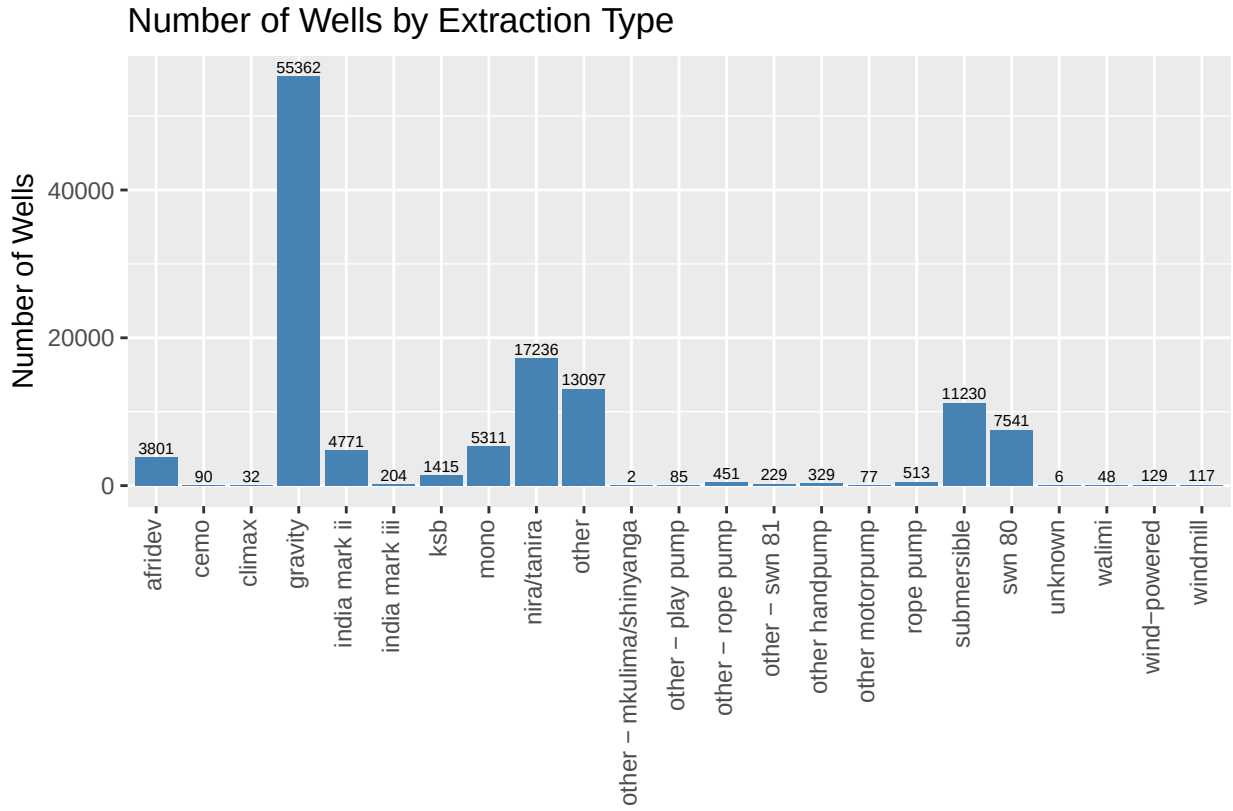


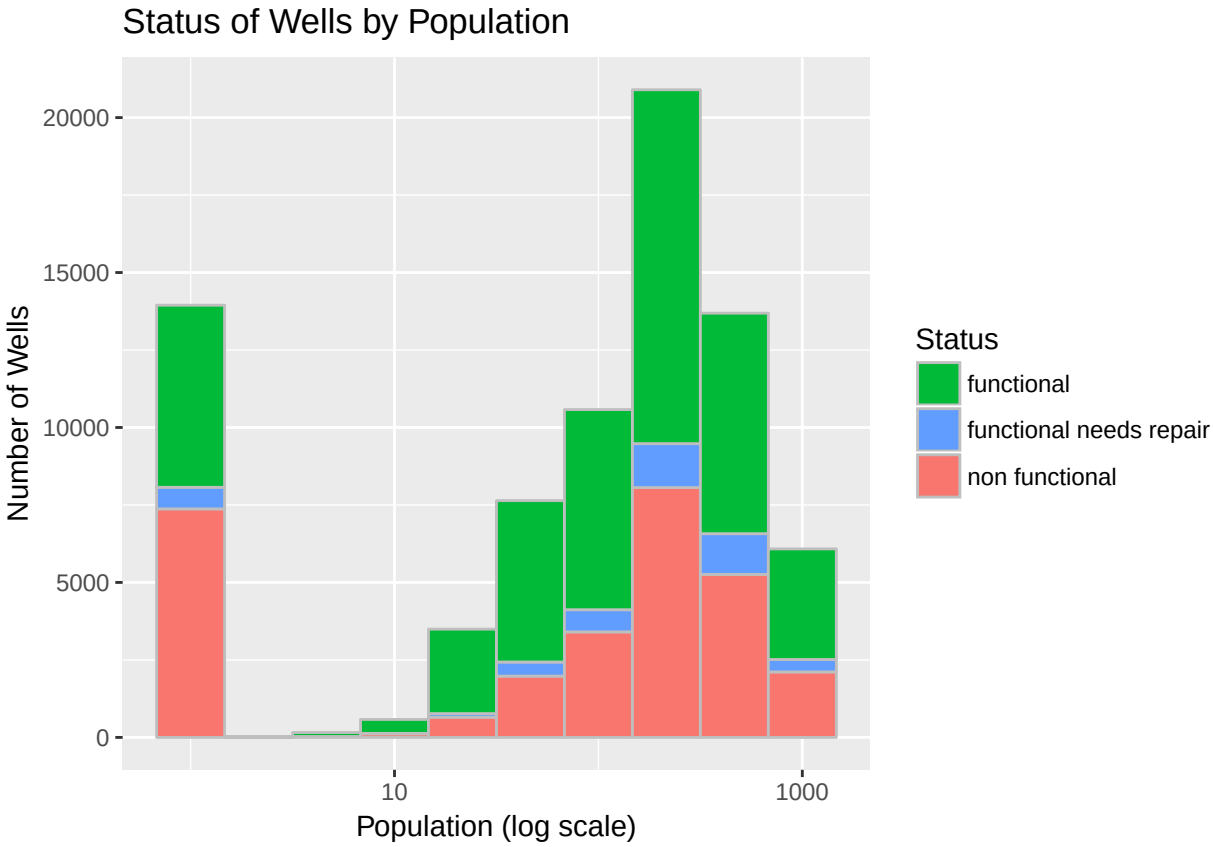
Table 9: Percentage of Non-Functional Wells by Extraction Type

Extraction Type	Number of Non-Functional Wells	Percentage of Non-Functional Wells
other - mkulima/shinyanga	2	100.00
other	10469	79.93
climax	24	75.00
other - play pump	55	64.71
wind-powered	73	56.59
mono	2902	54.64
windmill	60	51.28
other handpump	165	50.15
unknown	3	50.00
india mark iii	101	49.51
ksb	686	48.48
other - swn 81	102	44.54
submersible	4518	40.23
cemo	36	40.00
swn 80	2842	37.69
india mark ii	1713	35.90
rope pump	166	32.36
other - rope pump	141	31.26
other motorpump	24	31.17
afridev	1145	30.12
gravity	16284	29.41
walimi	13	27.08
nira/tanira	4301	24.95

Gravity pumps are the most reliable extraction method, with a non-functional rate of just 29.4%. The afridev handpump also has a below-average non-functional rate at 30.1%, while other handpumps such as the India Mark II and SWN80 are around or just under the average non-functional rate at 35.9% and 37.7%, respectively.

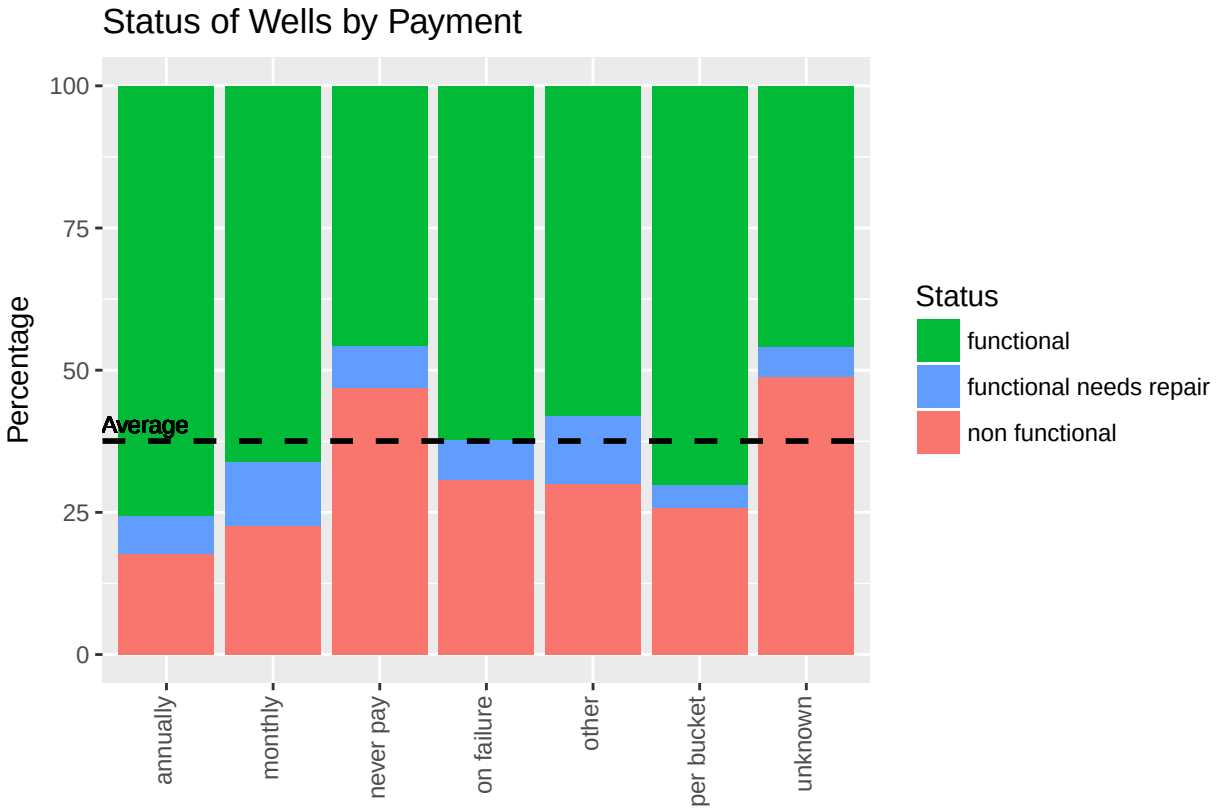
The results of the failure rates could imply that pumps with more complicated extraction types, such as mono (motor pump), climax (motor pump), ksb (submersible) and wind-powered, all of which have non-functional rates around 50% or higher, are more prone to failure as they may be less resilient to poor weather and not as conducive to maintenance conditions. It is noteworthy that wells with an 'other' extraction type have an extremely high (79.9%) non-functional rate. A possible explanation for this is that "other" extraction methods are less common and do not have standardized maintenance protocols.

4. Population



Note this only includes 64.33% of the dataset because the rest are missing this factor. In general, wells with a very low surrounding population of between 0 and 10 (note this does *not* include the missing population values of 0) have an extremely high failure rate over 50% while the rest of the wells demonstrate somewhat similar failure rates between 30 and 40%. Somewhat surprisingly, there is no clear evidence of a link between higher population and increasing well non-functionality. This may be due to the fact that their heavy usage is compensated for by better maintenance and resources in larger population areas. In addition, it may be that people naturally settle in large population areas close to natural bodies of water such as a river or spring, which was shown in the section above to have a lower non-functional rate than wells with other sources.

5. Payment


Status of Wells by Payment
Percentage
100
75
50
25
0
Average
annually
monthly
never pay
on failure
other
per bucket
unknown
Status
functional
functional needs repair
non functional

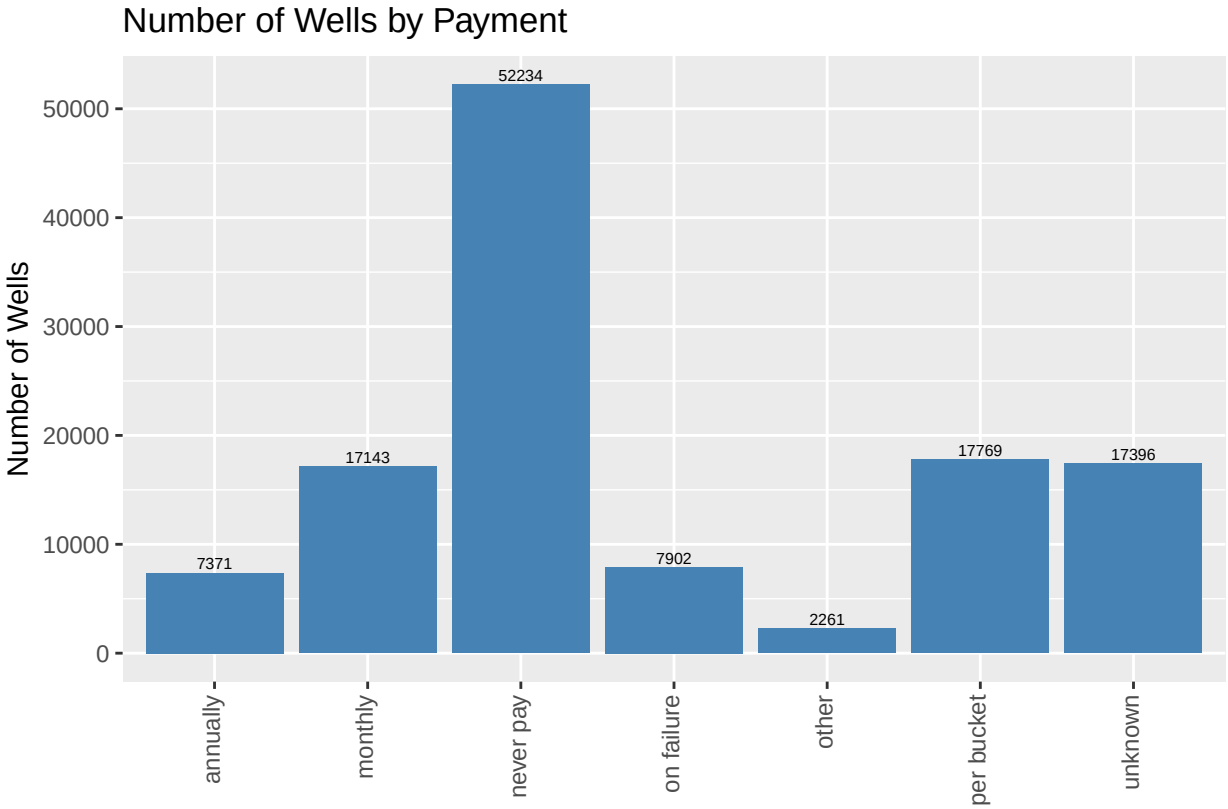
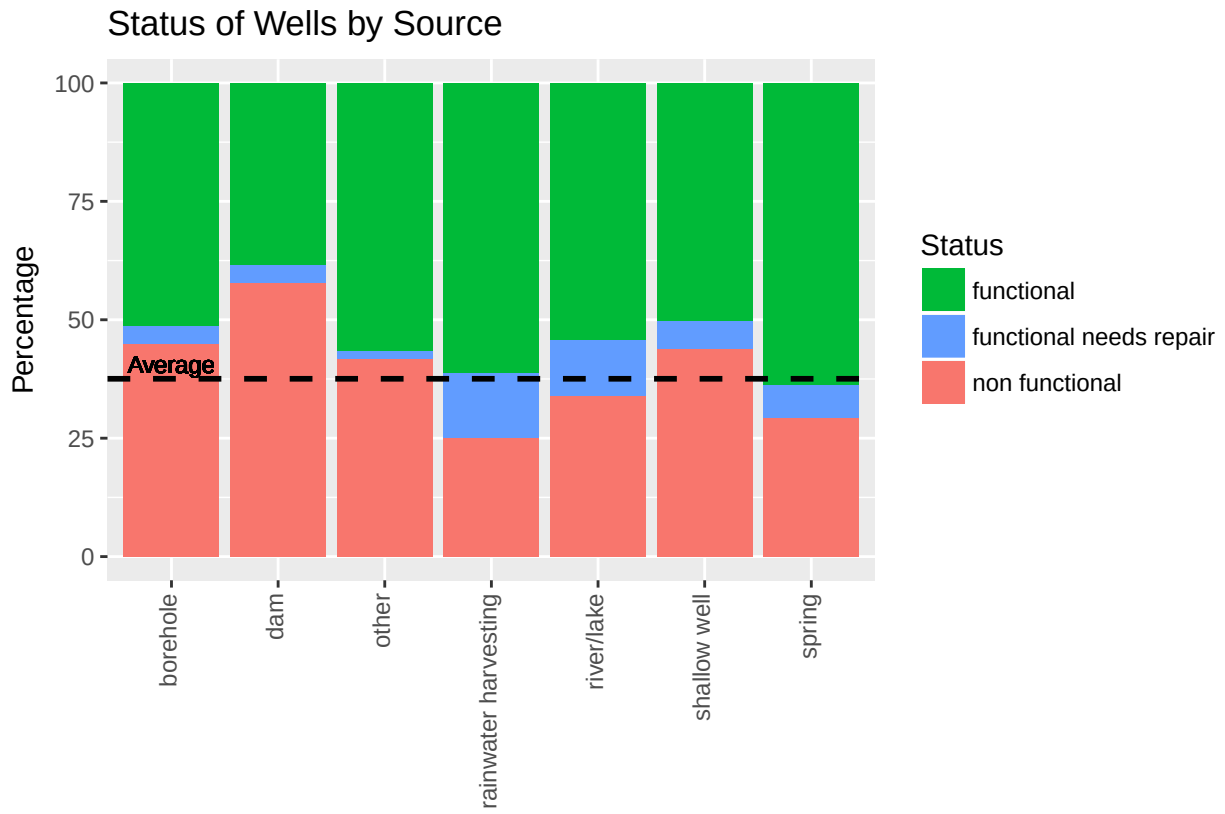

Number of Wells by Payment
Number of Wells
50000
40000
30000
20000
10000
0
7371
17143
52234
7902
2261
17769
17396
annually
monthly
never pay
on failure
other
per bucket
unknown

Table 10: Percentage of Non-Functional Wells by Payment

Payment	Number of Non-Functional Wells	Percentage of Non-Functional Wells
unknown	8479	48.74
never pay	24475	46.86
on failure	2421	30.64
other	680	30.08
per bucket	4590	25.83
monthly	3875	22.60
annually	1305	17.70

Wells with a known payment method are more likely to be functional - wells with a payment type of 'never pay' have a 48.9% non-functional rate, much higher than the overall average of 37.5%. Those with annual payments have the lowest non-functional rate at 17.8%, followed by those with monthly payments at 22.6% and then per bucket payments at 25.8%.

6. Source


Status of Wells by Source
Percentage
100
75
50
25
0
Average
borehole
dam
other
rainwater harvesting
river/lake
shallow well
spring
Status
functional
functional needs repair
non functional

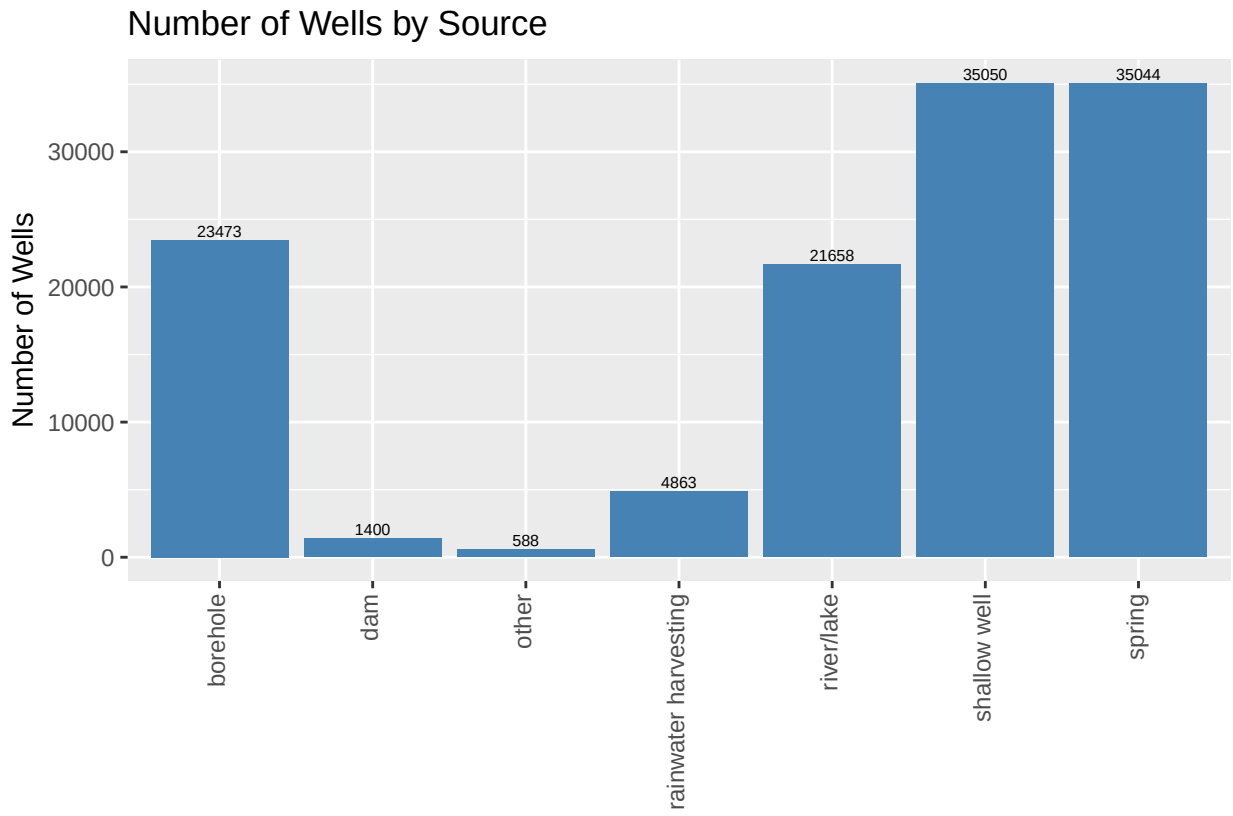
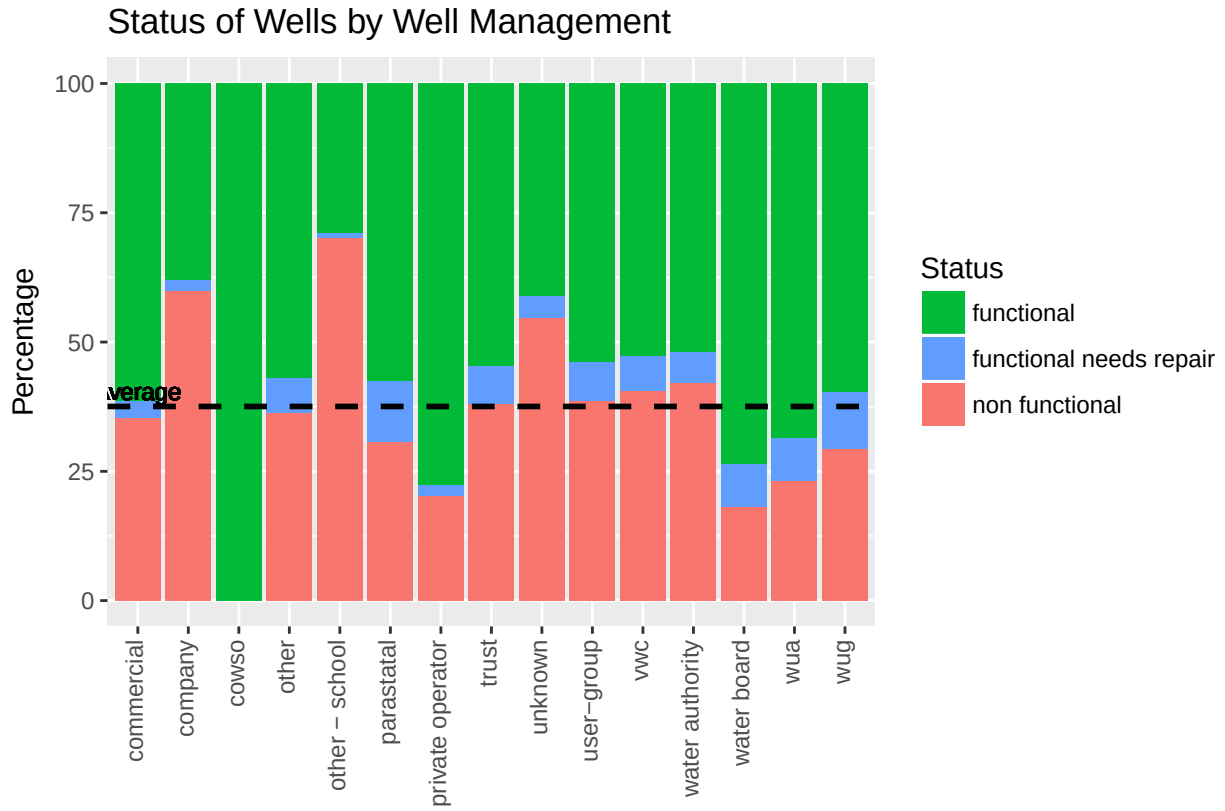

Number of Wells by Source
Number of Wells
30000
20000
10000
0
23473
1400
588
4863
21658
35050
35044
borehole
dam
other
rainwater harvesting
river/lake
shallow well
spring

Table 11: Percentage of Non-Functional Wells by Source

Source	Number of Non-Functional Wells	Percentage of Non-Functional Wells
dam	810	57.86
borehole	10552	44.95
shallow well	15358	43.82
other	246	41.84
river/lake	7358	33.97
spring	10276	29.32
rainwater harvesting	1225	25.19

Wells with a dam source have the highest non-functional rate at 57.9%, followed by borehole wells at 44.9%. Wells that have a rainwater, spring, or river/lake source are less likely to experience failure with non-functional rates of 25.2%, 29.3% and 34.0% respectively.

7. Well Management



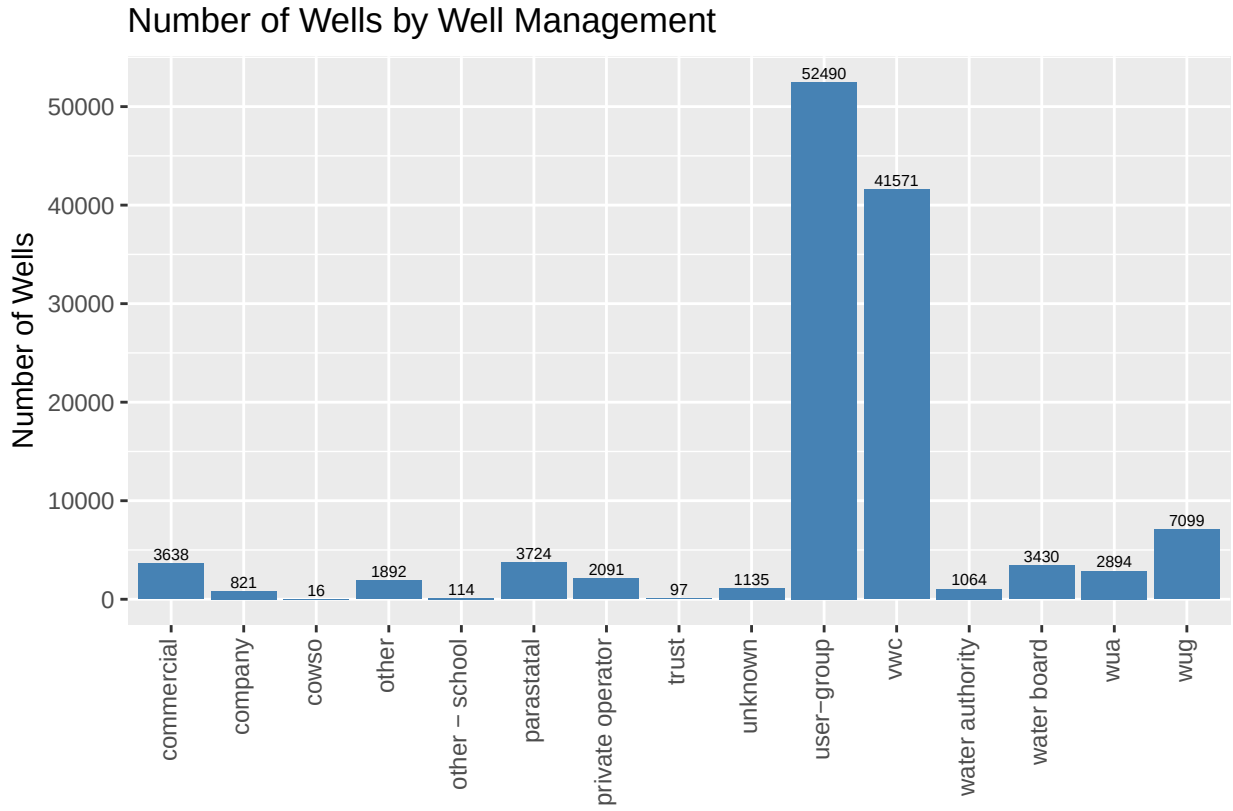
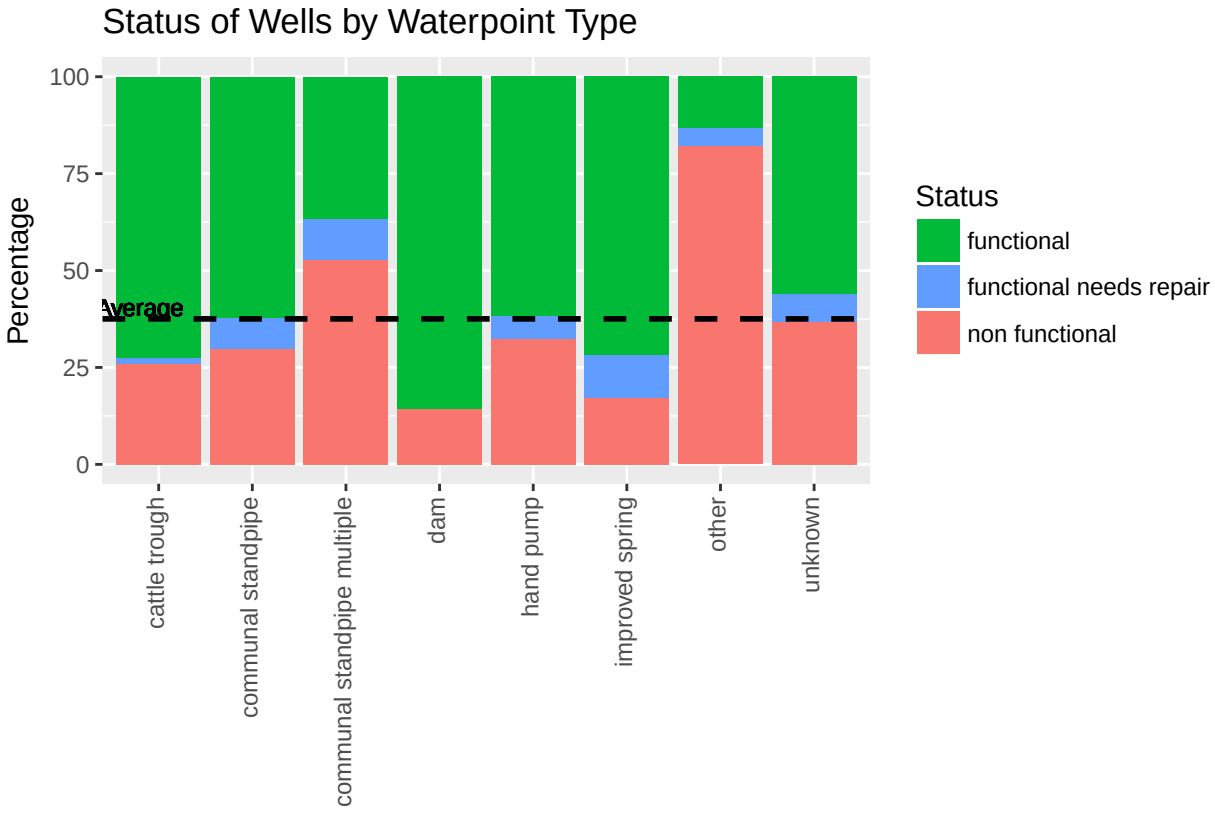


Table 12: Percentage of Non-Functional Wells by Well Management

Well Management	Number of Non-Functional Wells	Percentage of Non-Functional Wells
other - school	80	70.18
company	492	59.93
unknown	621	54.71
water authority	448	42.11
vwc	16899	40.65
user-group	20332	38.73
trust	37	38.14
other	687	36.31
commercial	1286	35.35
parastatal	1145	30.75
wug	2084	29.36
wua	672	23.22
private operator	424	20.28
water board	618	18.02

Wells managed by schools saw the highest non-functional rate of 70.2%, although such management types were rare. Most of the wells were managed by user-groups or vwc's, which had similar non-functional rates around the average for the entire dataset. Private operators and water boards had the lowest non-functional rates at 20.3% and 18.0%, although again the number of wells with such management types is much smaller than those managed by user-groups or vwc's.

8. Waterpoint Type



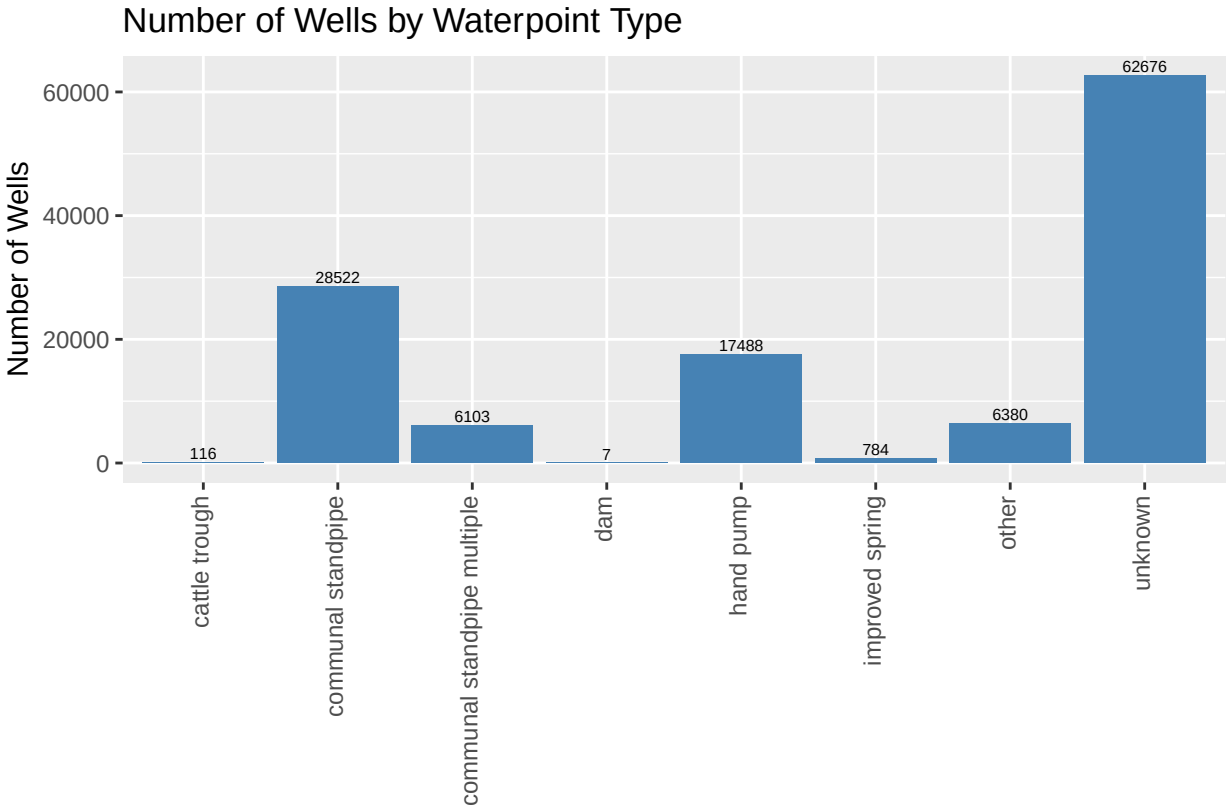
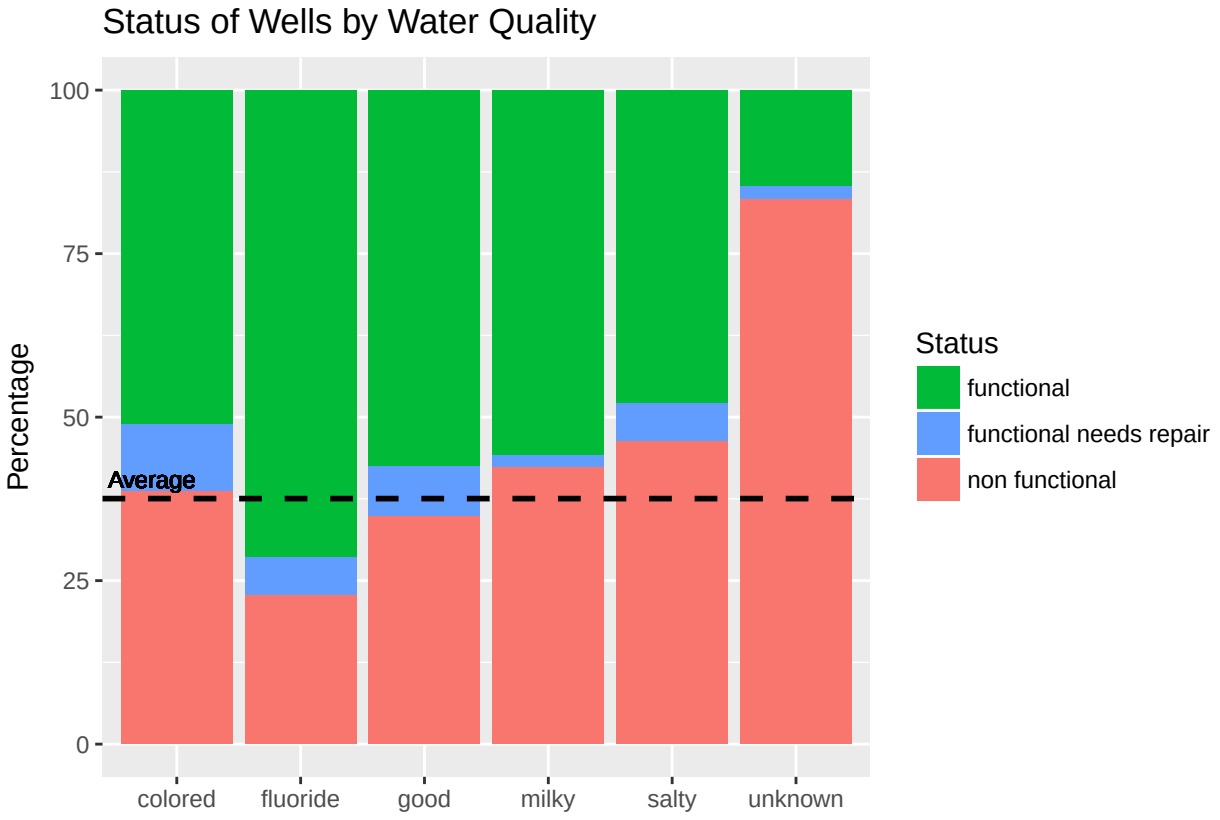


Table 13: Percentage of Non-Functional Wells by Waterpoint Type

Waterpoint Type	Number of Non-Functional Wells	Percentage of Non-Functional Wells
other	5247	82.24
communal standpipe multiple	3220	52.76
unknown	23001	36.70
hand pump	5654	32.33
communal standpipe	8536	29.93
cattle trough	30	25.86
improved spring	136	17.35
dam	1	14.29

The findings in this section come with the caveat that over half of the dataset did not report the waterpoint type, so most of the data falls into the unknown category. Just 29.9% of communal standpipes are non-functional, which is significantly less than the overall average of 38%. Likewise, hand pumps are below average at 32.3%. Multiple communal standpipes have a high failure rate of 52.8% while wells marked 'other' have an extremely high 82.2% non-functional rate. However, this may be due to the fact that it is more difficult to determine the type of a non-functional well and therefore non-functional wells are more likely to be listed as other, which inflates the non-functional rate.

9. Water Quality



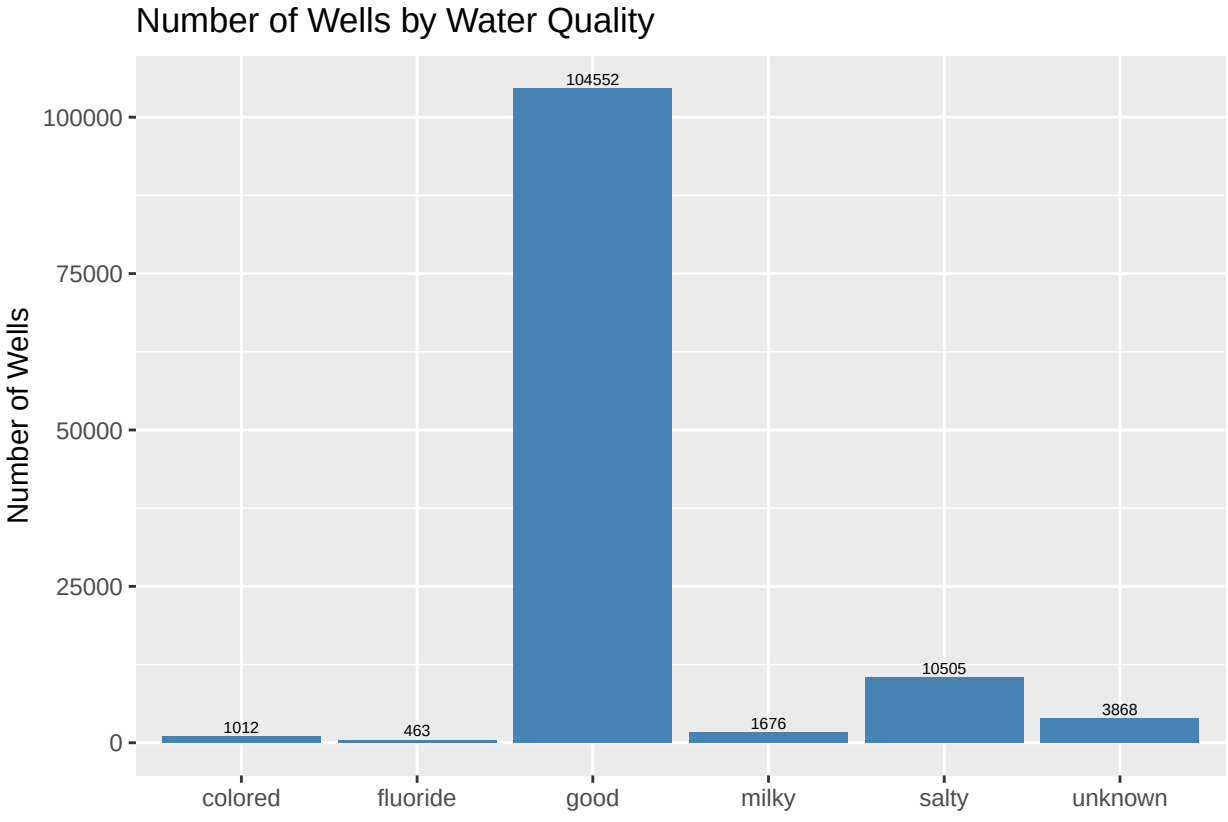
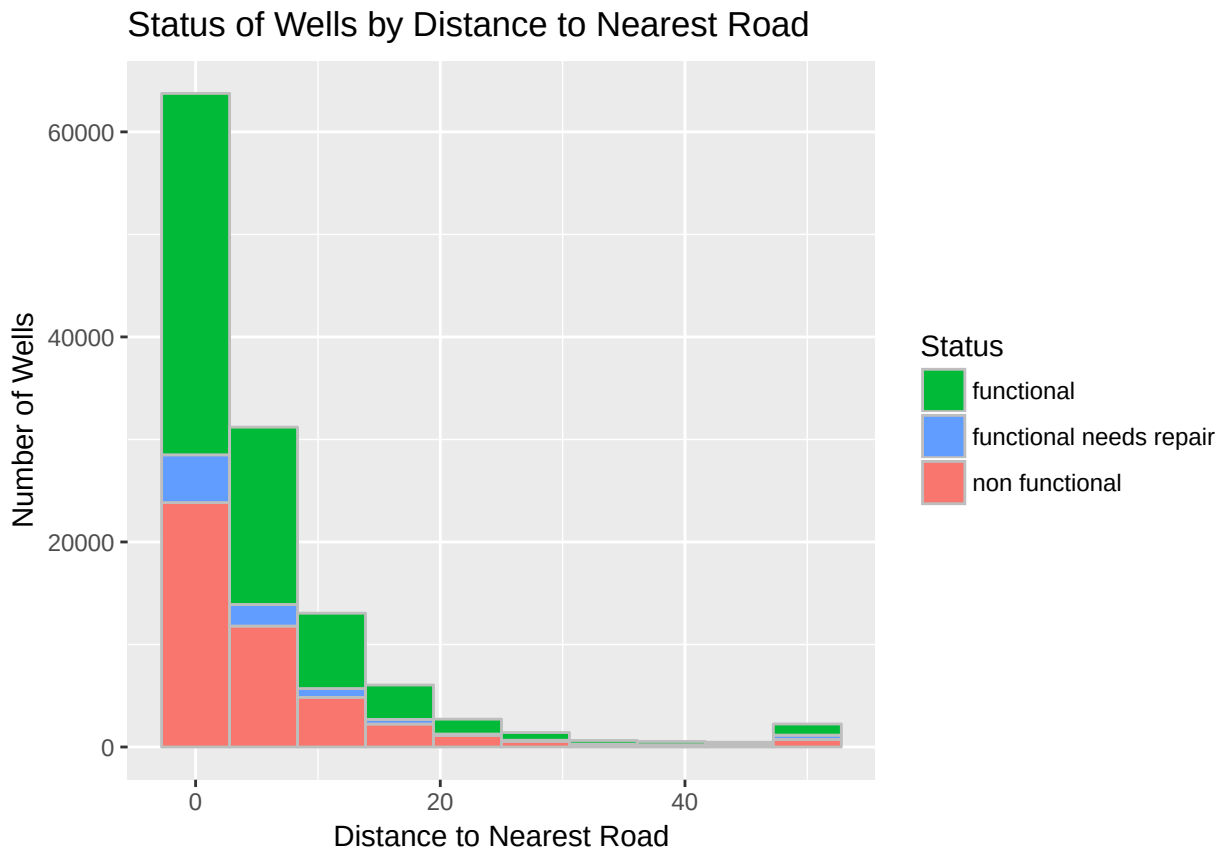


Table 14: Percentage of Non-Functional Wells by Water Quality

Water Quality	Number of Non-Functional Wells	Percentage of Non-Functional Wells
unknown	3227	83.43
salty	4864	46.30
milky	711	42.42
colored	391	38.64
good	36526	34.94
fluoride	106	22.89

The vast majority of wells reported good water quality, and their non-functional rate was slightly lower than the overall average at 34.9%. The wells with the lowest non-functional rates had fluoride water quality and just 22.9% of such wells were non-functional. All other water qualities observed were correlated with higher non-functional rates, with those of unknown water quality topping the list at 83% - this is likely because wells which are non-functional are much more likely to not have any water and thus making it more difficult to determine water quality. Out of wells with a known water quality, salty wells had the highest non-functional rate at 46.3%, followed by milky and coloured water at 42.4% and 38.6% respectively.

10. Distance to Nearest Road



Wells close to roads saw a slightly lower than average nonfunctional rate, but there was not a large difference for wells observed that were somewhat close to a road. However, for wells located farther away, especially noticeable for those located about 50km or more from the nearest road, the non-functional rate increased.

Model Results

We first split the combined dataset (removing overlapping locations) into training and testing sets using a 75-25 split, with 75% of the wells in the data going into the training set and 25% into the testing set. Two different Random Forest models were then trained on part of the combined dataset, with the first one without the distance to the nearest road feature and the second one with the distance to the nearest road feature. Each model was then applied to parts of the first dataset, the second dataset, and the combined dataset that were in the testing dataset.

Dataset 1 Results

Both models performed relatively well on testing data from just the first dataset (14,715 wells). The accuracy of the model without the distance feature was 77.51% and the accuracy with the distance feature was 78.86%. This meant the model's accuracy increased by 1.35% by including just one non-survey-based feature from the road data.

Dataset 2 Results

Both models performed slightly worse on testing data from just the second dataset (15,804 wells). The accuracy of the model without the distance feature was 77.45% and the accuracy with the distance feature was 78.06%. Although the model's accuracy after including the distance feature did not increase as much as it did in the first dataset, there was still an increase of 0.61% after including a single non-survey-based feature from the road data.The accuracy on the second dataset was lower than that on the first likely due the higher proportion of missing values in the second dataset than the first.

Combined Dataset Results

On the combined testing data (30,519 wells), the accuracy of the model without the distance feature was 77.49% and the accuracy with the distance feature was 78.48%, an increase of 0.99% after including one non-survey-based feature. The accuracy on the combined dataset was lower than that on the first likely due the higher proportion of missing values in the second dataset, and thus the combined dataset, than the first. The results are summarized in the table below:

Table 15: Summary of Model Performance

Dataset	Accuracy.without.Distance.Feature	Accuracy.with.Distance.Feature	Change.in.Accuracy
1	77.51	78.86	1.35
2	77.45	78.06	0.61
Combined	77.49	78.48	0.99

Since our model only uses reproducible features that would apply to wells in other countries besides Tanzania, we did not use features such as province, latitude, and longitude. As a result, we would expect an accuracy lower than the 81.1% accuracy found by Topor et. al [1] in their analysis as they considered all features. From the confusion matrix, we can see that our model was better at correctly predicting whether a well was non-functional rather than whether it was in one of the two other categories.

```
## Confusion Matrix and Statistics
##
##
## predicted_status     0     1     2
##                0 15708  1702  3493
```

```
##                1    59   208    45
##                2   996   273  8035
##
## Overall Statistics
##
##                Accuracy : 0.7848
##                  95% CI : (0.7801, 0.7894)
##     No Information Rate : 0.5493
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.5759
##  Mcnemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##                      Class: 0 Class: 1 Class: 2
## Sensitivity            0.9371 0.095282   0.6943
## Specificity            0.6223 0.996330   0.9330
## Pos Pred Value         0.7515 0.666667   0.8636
## Neg Pred Value         0.8903 0.934618   0.8332
## Prevalence             0.5493 0.071529   0.3792
## Detection Rate         0.5147 0.006815   0.2633
## Detection Prevalence   0.6849 0.010223   0.3049
## Balanced Accuracy      0.7797 0.545806   0.8137
```

Conclusion

There are several limitations inherent to using survey data when predicting well failures. The data collection process is costly and time-consuming due to the need for a human to visit each well in-person. This limits the frequency of collection by the amount of money and the number of trained personnel that are available. The data collected in this report was ultimately of low quality with numerous missing values and concerns about the accuracy and consistency of reported survey features. Regardless of the quality of data collected, however, the problem persists that survey data collection only accounts for one specific point in time. The most important feature to predict functionality is water quantity, and since the functional status of a well can change abruptly, the validity of a single data snapshot is insufficient.

A solution to this problem lies in having continuous data across an interval of time. Remote sensors will allow us to efficiently monitor wells and identify when a well is in need of repair, for example when the flow rate drops precipitously. It will then be possible to present technicians a concrete list of wells which are in need of repair with real-time information and updates, improving efficiency of repair efforts and ultimately boost the reliability of wells.

References

1. Topor et. al *Predicting Tanzanian Water Pump Maintenance Needs*
2. Liaw and Wiener CRAN-R randomForest Package Manual

Appendix: Data Dictionary

- amount_tsh - Total static head (amount water available to waterpoint)
- date_recorded - The date the row was entered

- funder - Who funded the well
- gps_height - Altitude of the well
- installer - Organization that installed the well
- longitude - GPS coordinate
- latitude - GPS coordinate
- wpt_name - Name of the waterpoint if there is one
- basin - Geographic water basin
- subvillage - Geographic location
- region - Geographic location
- region_code - Geographic location (coded)
- district_code - Geographic location (coded)
- lga - Geographic location
- ward - Geographic location
- population - Population around the well
- public_meeting - True/False
- recorded_by - Group entering this row of data
- scheme_management - Who operates the waterpoint
- scheme_name - Who operates the waterpoint
- permit - If the waterpoint is permitted
- construction_year - Year the waterpoint was constructed
- extraction_type - The kind of extraction the waterpoint uses
- extraction_type_group - The kind of extraction the waterpoint uses
- extraction_type_class - The kind of extraction the waterpoint uses
- management - How the waterpoint is managed
- management_group - How the waterpoint is managed
- payment - What the water costs
- payment_type - What the water costs
- water_quality - The quality of the water
- quality_group - The quality of the water
- quantity - The quantity of water
- quantity_group - The quantity of water
- source - The source of the water
- source_type - The source of the water
- source_class - The source of the water
- waterpoint_type - The kind of waterpoint
- waterpoint_type_group - The kind of waterpoint
- NEAR_DIST - Distance to nearest road in kilometers