ENGMAUKA / ml_project_phase3

Type / to search

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

ml_project_phase3 / index.ipynb

ENGMAUKA Conclusion fb3034c · 21 minutes ago History

Preview Code Blame 4288 lines (4288 loc) · 1.32 MB Raw

1. Overview

Tanzania, a developing nation, faces challenges in supplying its more than 57,000,000 inhabitants with water that is safe for drinking. It is extremely difficult for people to find clean, sanitary water if they do not reside close to one of the three large lakes that border the country, as one-third of the country is arid to semi-arid. Consequently, Tanzanians rely heavily on groundwater as their primary supply of water;

According to the Sustainable Development Goals (SDG) standards, just 61% of Tanzanian households presently have access to a basic water supply, 32% to basic sanitation, and 48% to basic hygiene. As a direct result, Tanzania has had to deal with mortality and illness, with the poor and vulnerable, women, and children bearing the brunt of this burden. Inadequate WASH services are thought to be the cause of 31,000 fatalities annually in Tanzania, accounting for almost 10% of avoidable deaths. These deaths also cost the country more than $2.4 billion annually in lost productivity and additional medical expenses.

The nation already has a large number of water points (stations), but some of them require maintenance, while others have completely failed.

The project aims to builds classification model, using an iterative approach, to predict the condition of water wells in Tanzania. The dataset for modelling was obtained from data provided by Taarifa and the Tanzanian Ministry of water.

2. Business and Data Understanding

2.1 Business Problem

Victoria Inc. has been procured by the Government of Tanzania on a consultancy basis to study the severe water crisis experienced by the country and propose a data driven solution to clean water accessibility. Victoria Inc. is tasked with coming up with a model that predicts the operating condition of the water points. This model will assist the government to:

- Prioritize maintenance and repairs based on operating status;
- Understand the failure rate of the water points;
- Optimize allocation of resources to restore the water points.

The objective of Victoria Inc is to:

- Develop a predictive model for classifying water points;
- Identify factors that affect water points functionality;

Proposed Solution:

- Develop a machine learning classification model with an accuracy score of 80%.

Performance Metrics:

- Acurracy
- Precision
- Recall
- F1-score

2.2 Data Understanding

2.2.1 Data Source

The dataset employed in the study was downloaded from https://www.drivendata.org/competitions/7/data/

2.2.2 Dataset Features

The following features about the water points are provided:

- *amount_tsh* - Total static head (amount water available to waterpoint)
- *date_recorded* - The date the row was entered
- *funder* - Who funded the well
- *gps_height* - Altitude of the well
- *installer* - Organization that installed the well
- *longitude* - GPS coordinate
- *latitude* - GPS coordinate
- *wpt_name* - Name of the waterpoint if there is one

- *num_private* -
- *basin* - Geographic water basin
- *subvillage* - Geographic location
- *region* - Geographic location
- *region_code *- Geographic location (coded)
- *district_code* - Geographic location (coded)
- *lga* - Geographic location
- *ward* - Geographic location
- *population* - Population around the well
- *public_meeting* - True/False
- *recorded_by *- Group entering this row of data
- *scheme_management* - Who operates the waterpoint
- *scheme_name* - Who operates the waterpoint
- *permit* - If the waterpoint is permitted
- *construction_year* - Year the waterpoint was constructed
- *extraction_type* - The kind of extraction the waterpoint uses
- *extraction_type_group* - The kind of extraction the waterpoint uses
- *extraction_type_class* - The kind of extraction the waterpoint uses
- *management* - How the waterpoint is managed
- *management_group* - How the waterpoint is managed
- *payment* - What the water costs
- *payment_type* - What the water costs
- *water_quality* - The quality of the water
- *quality_group* - The quality of the water
- *quantity* - The quantity of water
- *quantity_group* - The quantity of water
- *source* - The source of the water
- *source_type* - The source of the water
- *source_class* - The source of the water
- *waterpoint_type* - The kind of waterpoint
- *waterpoint_type_group* - The kind of waterpoint

The target variable has three possible values:

- **functional** - the waterpoint is operational and there are no repairs needed
- **functional needs repair** - the waterpoint is operational, but needs repairs
- **non functional** - the waterpoint is not operational

2.3 Methodology

The adopted structure for the project was CRISP-DM that entails undertaking Business Understanding; Data Understanding; Data Preparation; Data Cleaning and Explatory Data Analysis(EDA); Modelling; Conclusion and Recommendations.

3.Data Cleaning and EDA

3.1 Data Cleaning & Preparation

Importing packages

In [110…

```
#importing standard packages
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
%matplotlib inline
from textwrap import fill

from sklearn.model_selection import train_test_split, cross_val_score,RepeatedStratifiedKFold, GridSearchCV
import  warnings
warnings . filterwarnings("ignore")
from sklearn.pipeline import Pipeline

#classification models
from sklearn.dummy import DummyClassifier
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

#classification metrics
```

```
from sklearn.metrics import plot_confusion_matrix
from sklearn.metrics import accuracy_score, f1_score, precision_score, recall_score
from sklearn.metrics import roc_curve, roc_auc_score, auc
from sklearn.metrics import classification_report, ConfusionMatrixDisplay



#scalers
from sklearn.impute import SimpleImputer
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import LabelBinarizer, label_binarize

#dummies
from sklearn.preprocessing import OneHotEncoder
```

In [3]:

```
# loading the training data set
data_train_set = pd.read_csv("data/training_set_values.csv", index_col="id")
data_train_set
```

Out[3]:

id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	basin
69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	Lake Nyasa
8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	Lake Victoria
34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	Pangani
67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	Ruvuma / Southern Coast
19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	Lake Victoria
...	...	...	...	...	...	...	...	...	...	...
60739	10.0	2013-05-03	Germany	1210	CES	37.169807	-3.253847	Area Three	0	Pangani

			Republi					Namba 27		
27263	4700.0	2011-05-07	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	0	Rufiji
37057	0.0	2011-04-11	NaN	0	NaN	34.017087	-8.750434	Mashine	0	Rufiji
31282	0.0	2011-03-08	Malec	0	Musa	35.861315	-6.378573	Mshoro	0	Rufiji
26348	0.0	2011-03-23	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa	0	Wami / Ruvu

59400 rows × 39 columns

In [4]:

```python
# importing the test data set
data_test_set = pd.read_csv("data/test_set_values.csv", index_col="id")
data_test_set
```

Out[4]:

	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	b
id										
50785	0.0	2013-02-04	Dmdd	1996	DMDD	35.290799	-4.059696	Dinamu Secondary School	0	Inte
51630	0.0	2013-02-04	Government Of Tanzania	1569	DWE	36.656709	-3.309214	Kimnyak	0	Pan
17168	0.0	2013-02-01	NaN	1567	NaN	34.767863	-5.004344	Puma Secondary	0	Inte
45559	0.0	2013-01-22	Finn Water	267	FINN WATER	38.058046	-9.418672	Kwa Mzee Pange	0	Ruv Sout C
49871	500.0	2013-03-27	Bruder	1260	BRUDER	35.006123	-10.950412	Kwa Mzee Turuka	0	Ruv Sout C

...	...	...	...	...	...	...	...	...	...	
39307	0.0	2011-02-24	Danida	34	Da	38.852669	-6.582841	Kwambwezi	0	Wa F
18990	1000.0	2011-03-21	Hiap	0	HIAP	37.451633	-5.350428	Bonde La Mkondoa	0	Pan
28749	0.0	2013-03-04	NaN	1476	NaN	34.739804	-4.585587	Bwawani	0	Inte
33492	0.0	2013-02-18	Germany	998	DWE	35.432732	-10.584159	Kwa John	0	N
68707	0.0	2013-02-13	Government Of Tanzania	481	Government	34.765054	-11.226012	Kwa Mzee Chagala	0	N

14850 rows × 39 columns

In [5]:

```
# importing the training set labels
data_train_labels = pd.read_csv("data/training_set_labels.csv", index_col="id")
data_train_labels
```

Out[5]:

	status_group
id	
69572	functional
8776	functional
34310	functional
67743	non functional
19728	functional
...	...
60739	functional
27263	functional
37057	functional

31282	functional
26348	functional

59400 rows × 1 columns

In [6]:

```
#merging the training dataset with the train labels
data = pd.merge(data_train_labels, data_train_set, on="id", how="inner")
data
```

Out[6]:

	status_group	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_priva
id										
69572	functional	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	
8776	functional	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	
34310	functional	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	
67743	non functional	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	
19728	functional	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	
...	...	...	...	...	...	...	...	...	...	
60739	functional	10.0	2013-05-03	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	
27263	functional	4700.0	2011-05-07	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	
37057	functional	0.0	2011-04-11	NaN	0	NaN	34.017087	-8.750434	Mashine	
31282	functional	0.0	2011-03-08	Malec	0	Musa	35.861315	-6.378573	Mshoro	

26348	functional	0.0	2011-03-23	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa

59400 rows × 40 columns

In [7]:

```python
# creating new index for the merged dataset
data.reset_index(inplace=True)
data
```

Out[7]:

	id	status_group	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	...
0	69572	functional	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	..
1	8776	functional	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	..
2	34310	functional	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	..
3	67743	non functional	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	..
4	19728	functional	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	..
...	...	...	...	...	...	...	...	...	...	...	..
59395	60739	functional	10.0	2013-05-03	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	..
59396	27263	functional	4700.0	2011-05-07	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	..
59397	37057	functional	0.0	2011-04-11	NaN	0	NaN	34.017087	-8.750434	Mashine	..
59398	31282	functional	0.0	2011-03-08	Malec	0	Musa	35.861315	-6.378573	Mshoro	..

59399	26348	functional	0.0	2011-03-23	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa	..

59400 rows × 41 columns

In [8]:

```python
# checking the datatypes
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 41 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   id                     59400 non-null  int64  
 1   status_group           59400 non-null  object 
 2   amount_tsh             59400 non-null  float64
 3   date_recorded          59400 non-null  object 
 4   funder                 55765 non-null  object 
 5   gps_height             59400 non-null  int64  
 6   installer              55745 non-null  object 
 7   longitude              59400 non-null  float64
 8   latitude               59400 non-null  float64
 9   wpt_name               59400 non-null  object 
 10  num_private            59400 non-null  int64  
 11  basin                  59400 non-null  object 
 12  subvillage             59029 non-null  object 
 13  region                 59400 non-null  object 
 14  region_code            59400 non-null  int64  
 15  district_code          59400 non-null  int64  
 16  lga                    59400 non-null  object 
 17  ward                   59400 non-null  object 
 18  population             59400 non-null  int64  
 19  public_meeting         56066 non-null  object 
 20  recorded_by            59400 non-null  object 
 21  scheme_management      55523 non-null  object 
 22  scheme_name            31234 non-null  object 
 23  permit                 56344 non-null  object 
 24  construction_year      59400 non-null  int64  
 25  extraction_type        59400 non-null  object 
 26  extraction_type_group  59400 non-null  object 
 27  extraction_type_class  59400 non-null  object 
 28  management             59400 non-null  object 
```

```
 29  management_group       59400 non-null  object
 30  payment                59400 non-null  object
 31  payment_type           59400 non-null  object
 32  water_quality          59400 non-null  object
 33  quality_group          59400 non-null  object
 34  quantity               59400 non-null  object
 35  quantity_group         59400 non-null  object
 36  source                 59400 non-null  object
 37  source_type            59400 non-null  object
 38  source_class           59400 non-null  object
 39  waterpoint_type        59400 non-null  object
 40  waterpoint_type_group  59400 non-null  object
dtypes: float64(3), int64(7), object(31)
memory usage: 18.6+ MB
```

The data set is made up of 41 columns and 59,400 rows. The data frame features has 3 float datatype; 7 integer datatype and 31 object datatype.

In [9]:
```
#checking for duplicates
data.duplicated().sum()
```

Out[9]:
```
0
```

The data set had zero duplicates

In [10]:
```
# checking null values
data.isna().sum()
```

Out[10]:
```
id                          0
status_group                0
amount_tsh                  0
date_recorded               0
funder                   3635
gps_height                  0
installer                3655
longitude                   0
latitude                    0
wpt_name                    0
num_private                 0
basin                       0
subvillage                371
```

```
subvillage                 371
region                       0
region_code                  0
district_code                0
lga                          0
ward                         0
population                   0
public_meeting            3334
recorded_by                  0
scheme_management         3877
scheme_name              28166
permit                    3056
construction_year            0
extraction_type              0
extraction_type_group        0
extraction_type_class        0
management                   0
management_group             0
payment                      0
payment_type                 0
water_quality                0
quality_group                0
quantity                     0
quantity_group               0
source                       0
source_type                  0
source_class                 0
waterpoint_type              0
waterpoint_type_group        0
dtype: int64
```

The columns "funder", "installer", "subvillage", "public_meeting", "scheme_management", "scheme_name", and "permit" had 3635, 3655, 371, 3334, 3877, 28166 and 3056 missing (null) values respectively. Scheme_name has the highest number of missing values.

In [11]:

```
for col in data.columns:
    print(data[col].value_counts())
```

```
69572    1
27851    1
6924     1
61097    1
48517    1
        ..
59036    1
```

```
59036    1
56446    1
3855     1
52786    1
26348    1
Name: id, Length: 59400, dtype: int64
functional                 32259
non functional             22824
functional needs repair     4317
Name: status_group, dtype: int64
0.0         41639
500.0        3102
50.0         2472
1000.0       1488
20.0         1463
            ...
6300.0          1
120000.0        1
138000.0        1
350000.0        1
59.0            1
Name: amount_tsh, Length: 98, dtype: int64
2011-03-15    572
2011-03-17    558
2013-02-03    546
2011-03-14    520
2011-03-16    513
             ...
2011-09-11      1
2011-08-31      1
2011-09-21      1
2011-08-30      1
2013-12-01      1
Name: date_recorded, Length: 356, dtype: int64
Government Of Tanzania    9084
Danida                    3114
Hesawa                    2202
Rwssp                     1374
World Bank                1349
                          ...
Rarymond Ekura               1
Justine Marwa                1
Municipal Council            1
Afdp                         1
Samlo                        1
```

```
Name: funder, Length: 1897, dtype: int64
 0       20438
-15         60
-16         55
-13         55
 1290       52
         ...
 2378        1
-54          1
 2057        1
 2332        1
 2366        1
Name: gps_height, Length: 2428, dtype: int64
DWE                17402
Government          1825
RWE                 1206
Commu               1060
DANIDA              1050
                   ...
Wizara  ya maji        1
TWESS                  1
Nasan workers          1
R                      1
SELEPTA                1
Name: installer, Length: 2145, dtype: int64
0.000000     1812
37.375717       2
38.340501       2
39.086183       2
33.005032       2
             ...
35.885754       1
36.626541       1
37.333530       1
38.970078       1
38.104048       1
Name: longitude, Length: 57516, dtype: int64
-2.000000e-08    1812
-6.985842e+00       2
-6.980220e+00       2
-2.476680e+00       2
-6.978263e+00       2
                 ...
-3.287619e+00       1
```

```
-8.234989e+00      1
-3.268579e+00      1
-1.146053e+01      1
-6.747464e+00      1
Name: latitude, Length: 57517, dtype: int64
none                       3563
Shuleni                    1748
Zahanati                    830
Msikitini                   535
Kanisani                    323
                           ...
Kwa Medadi                    1
Kwa Kubembeni                 1
Shule Ya Msingi Milanzi       1
Funua                         1
Kwa Mzee Lugawa               1
Name: wpt_name, Length: 37400, dtype: int64
0       58643
6          81
1          73
5          46
8          46
        ...
42          1
23          1
136         1
698         1
1402        1
Name: num_private, Length: 65, dtype: int64
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
Name: basin, dtype: int64
Madukani       508
Shuleni        506
Majengo        502
Kati           373
Mtakuja        262
```

```
               ...
Kipompo          1
Chanyamilima     1
Ikalime          1
Kemagaka         1
Kikatanyemba     1
Name: subvillage, Length: 19287, dtype: int64
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
Name: region, dtype: int64
11    5300
17    5011
12    4639
3     4379
5     4040
18    3324
19    3047
2     3024
16    2816
10    2640
4     2513
1     2201
13    2093
14    1979
20    1969
```

```
15    1808
6     1609
21    1583
80    1238
60    1025
90     917
7      805
99     423
9      390
24     326
8      300
40       1
Name: region_code, dtype: int64
1     12203
2     11173
3      9998
4      8999
5      4356
6      4074
7      3343
8      1043
30      995
33      874
53      745
43      505
13      391
23      293
63      195
62      109
60       63
0        23
80       12
67        6
Name: district_code, dtype: int64
Njombe          2503
Arusha Rural    1252
Moshi Rural     1251
Bariadi         1177
Rungwe          1106
                ...
Moshi Urban       79
Kigoma Urban      71
Arusha Urban      63
Lindi Urban       21
Nyamagana          1
```

```
Nyamagana           1
Name: lga, Length: 125, dtype: int64
Igosi             307
Imalinyi          252
Siha Kati         232
Mdandu            231
Nduruma           217
                 ... 
Uchindile           1
Thawi               1
Uwanja wa Ndege     1
Izia                1
Kinungu             1
Name: ward, Length: 2092, dtype: int64
0       21381
1        7025
200      1940
150      1892
250      1681
        ...  
6330        1
5030        1
656         1
948         1
788         1
Name: population, Length: 1049, dtype: int64
True     51011
False     5055
Name: public_meeting, dtype: int64
GeoData Consultants Ltd    59400
Name: recorded_by, dtype: int64
VWC                 36793
WUG                  5206
Water authority      3153
WUA                  2883
Water Board          2748
Parastatal           1680
Private operator     1063
Company              1061
Other                 766
SWC                    97
Trust                  72
None                    1
Name: scheme_management, dtype: int64
K                        682
```

```
None                      644
Borehole                  546
Chalinze wate             405
M                         400
                         ...
Mradi wa maji Vijini        1
Villagers                   1
Magundi water supply        1
Saadani Chumv               1
Mtawanya                    1
Name: scheme_name, Length: 2696, dtype: int64
True     38852
False    17492
Name: permit, dtype: int64
0       20709
2010     2645
2008     2613
2009     2533
2000     2091
2007     1587
2006     1471
2003     1286
2011     1256
2004     1123
2012     1084
2002     1075
1978     1037
1995     1014
2005     1011
1999      979
1998      966
1990      954
1985      945
1980      811
1996      811
1984      779
1982      744
1994      738
1972      708
1974      676
1997      644
1992      640
1993      608
2001      540
```

```
1988     521
1983     488
1975     437
1986     434
1976     414
1970     411
1991     324
1989     316
1987     302
1981     238
1977     202
1979     192
1973     184
2013     176
1971     145
1960     102
1967      88
1963      85
1968      77
1969      59
1964      40
1962      30
1961      21
1965      19
1966      17
Name: construction_year, dtype: int64
gravity                        26780
nira/tanira                     8154
other                           6430
submersible                     4764
swn 80                          3670
mono                            2865
india mark ii                   2400
afridev                         1770
ksb                             1415
other - rope pump                451
other - swn 81                   229
windmill                         117
india mark iii                    98
cemo                              90
other - play pump                 85
walimi                            48
climax                            32
other - mkulima/shinyanga          2
```

```
Name: extraction_type, dtype: int64
gravity           26780
nira/tanira        8154
other              6430
submersible        6179
swn 80             3670
mono               2865
india mark ii      2400
afridev            1770
rope pump           451
other handpump      364
other motorpump     122
wind-powered        117
india mark iii       98
Name: extraction_type_group, dtype: int64
gravity        26780
handpump       16456
other           6430
submersible     6179
motorpump       2987
rope pump        451
wind-powered     117
Name: extraction_type_class, dtype: int64
vwc                 40507
wug                  6515
water board          2933
wua                  2535
private operator     1971
parastatal           1768
water authority       904
other                 844
company               685
unknown               561
other - school         99
trust                  78
Name: management, dtype: int64
user-group    52490
commercial     3638
parastatal     1768
other           943
unknown         561
Name: management_group, dtype: int64
never pay                25348
pay per bucket            8985
pay monthly               8300
```

```
pay monthly                8300
unknown                    8157
pay when scheme fails      3914
pay annually               3642
other                      1054
Name: payment, dtype: int64
never pay     25348
per bucket     8985
monthly        8300
unknown        8157
on failure     3914
annually       3642
other          1054
Name: payment_type, dtype: int64
soft                  50818
salty                  4856
unknown                1876
milky                   804
coloured                490
salty abandoned         339
fluoride                200
fluoride abandoned       17
Name: water_quality, dtype: int64
good        50818
salty        5195
unknown      1876
milky         804
colored       490
fluoride      217
Name: quality_group, dtype: int64
enough          33186
insufficient    15129
dry              6246
seasonal         4050
unknown           789
Name: quantity, dtype: int64
enough          33186
insufficient    15129
dry              6246
seasonal         4050
unknown           789
Name: quantity_group, dtype: int64
spring                  17021
shallow well            16824
machine dbh             11075
```

```
river                   9612
rainwater harvesting    2295
hand dtw                 874
lake                     765
dam                      656
other                    212
unknown                   66
Name: source, dtype: int64
spring                 17021
shallow well           16824
borehole               11949
river/lake             10377
rainwater harvesting    2295
dam                      656
other                    278
Name: source_type, dtype: int64
groundwater    45794
surface        13328
unknown          278
Name: source_class, dtype: int64
communal standpipe             28522
hand pump                      17488
other                           6380
communal standpipe multiple     6103
improved spring                  784
cattle trough                    116
dam                                7
Name: waterpoint_type, dtype: int64
communal standpipe    34625
hand pump             17488
other                  6380
improved spring         784
cattle trough           116
dam                       7
Name: waterpoint_type_group, dtype: int64
```

The following columns contain either similar or duplicated data, therefore, in order to avoid multicollinearity one or both of the columns will be dropped:

- *scheme_management *and *management*;
- *extraction_type*, *extraction_type_group* and *extraction_type_class*;
- *payment* and *payment_type*;
- *water_quality and quality_group*

- *water_quality* and *quality_group*
- *quantity* and *quantity_group*
- *source* and *source_type*
- *waterpoint_type* and *waterpoint_type_group*

Similarly, the following columns will either be transformed or dropped:

- columns to be dropped:
  - *id* as its just an index identifier, *num_private* since its not defined and therefore its relevance is not clear; *recorded_by* since it has the same value throughout the dataset;
  - *population*, *amount_tsh*, *construction_year*, *longitude*, *latitude*, and *gps_height* have 0 entered on most of their rows. The rows with 0 will be dropped.
- columns to be transformed or feature engineered:
  - *permit* and *public_meeting* are boolean;
  - *wpt_name*, *scheme_name*, *subvillage*, *ward*, *date_recorded*, *funder* are categorical variables with their unique values in integers. Therefore, they will require further analysis and probable use of dummy variables.

In [12]:

```
# columns to be dropped
drop_cols = ["scheme_management", "extraction_type", "extraction_type_group",
             "payment", "water_quality", "source_type", "waterpoint_type_group",
             "id", "num_private", "recorded_by", "quantity", "public_meeting",
             "wpt_name", "scheme_name", "subvillage", "ward", "date_recorded",
             "funder", "district_code", "lga", "region_code", "ward", "management_group",
             "longitude", "latitude", "gps_height", "source"]

#dropping the columns:
data = data.drop(drop_cols, axis=1)

# checking the information of the new dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   status_group           59400 non-null  object 
 1   amount_tsh             59400 non-null  float64
 2   installer              55745 non-null  object 
```

```
 3   basin                  59400 non-null  object
 4   region                 59400 non-null  object
 5   population             59400 non-null  int64
 6   permit                 56344 non-null  object
 7   construction_year      59400 non-null  int64
 8   extraction_type_class  59400 non-null  object
 9   management             59400 non-null  object
 10  payment_type           59400 non-null  object
 11  quality_group          59400 non-null  object
 12  quantity_group         59400 non-null  object
 13  source_class           59400 non-null  object
 14  waterpoint_type        59400 non-null  object
dtypes: float64(1), int64(2), object(12)
memory usage: 6.8+ MB
```

The new dataframe has 15 columns and 59400 rows. Its only columns "installer" and "permit" that have missing values. The dataframe is made up of three datatypes: 1 column of type float, 2 columns with type integer and 12 columns with type object.

In [13]:

```python
# dropping the null values
data = data.dropna()

# confirming no null values are remaining
data.isna().sum()
```

Out[13]:

```
status_group             0
amount_tsh               0
installer                0
basin                    0
region                   0
population               0
permit                   0
construction_year        0
extraction_type_class    0
management               0
payment_type             0
quality_group            0
quantity_group           0
source_class             0
waterpoint_type          0
dtype: int64
```

In [14]:

```python
# converting permit column to integers
```

```
data["permit"] = data["permit"].astype(int)

# checking datatypes of the dataset
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55102 entries, 0 to 59399
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   status_group           55102 non-null  object 
 1   amount_tsh             55102 non-null  float64
 2   installer              55102 non-null  object 
 3   basin                  55102 non-null  object 
 4   region                 55102 non-null  object 
 5   population             55102 non-null  int64  
 6   permit                 55102 non-null  int32  
 7   construction_year      55102 non-null  int64  
 8   extraction_type_class  55102 non-null  object 
 9   management             55102 non-null  object 
 10  payment_type           55102 non-null  object 
 11  quality_group          55102 non-null  object 
 12  quantity_group         55102 non-null  object 
 13  source_class           55102 non-null  object 
 14  waterpoint_type        55102 non-null  object 
dtypes: float64(1), int32(1), int64(2), object(11)
memory usage: 6.5+ MB
```

The dataframe shape has been reduced to 15 columns by 55102 rows.

In [15]:
```
# checking unique values after initial data cleaning
data.nunique()
```

Out[15]:
```
status_group                3
amount_tsh                 95
installer                2056
basin                       9
region                     21
population               1026
permit                      2
construction_year          55
extraction_type_class       7
management                 12
```

```
management                12
payment_type               7
quality_group              6
quantity_group             5
source_class               3
waterpoint_type            7
dtype: int64
```

Categorical variable "installer", requires further investigation since it still has 2056 unique variables which would overwhelm the model. Population is an integer and therefore having different unique values is not problematic.

In [16]:

```python
# viewing the installer column data values
print(data["installer"].unique().tolist())
```

```
['Roman', 'GRUMETI', 'World vision', 'UNICEF', 'Artisan', 'DWE', 'DWSP', 'Water Aid', 'Private', 'DANIDA', 'Lawate
fuka water sup', 'WEDECO', 'Danid', 'TWE', 'ISF', 'Kilolo Star', 'District council', 'Water', 'WU', 'Not known',
'Central government', 'CEFA', 'Commu', 'Accra', 'World Vision', 'LGA', 'MUWSA', 'KKKT _ Konde and DWE', 'Governmen
t', 'Olgilai village community', 'KKKT', 'RWE', 'Adra /Community', 'SEMA', 'SHIPO', 'HESAWA', 'ACRA', 'Community',
'IFAD', 'Sengerema Water Department', 'HE', 'ISF and TACARE', 'Kokeni', 'DA', 'Adra', 'ALLYS', 'AICT', 'KIUMA', 'C
ES', 'District Counci', 'Ruthe', 'Adra/Community', 'Tulawaka Gold Mine', 'KKT C', 'Water board', 'LOCAL CONTRACT',
'LIPS', 'TASAF', 'World', '0', 'SW', 'Shipo', 'Fini water', 'Kanisa', 'OXFARM', 'VILLAGE COUNCIL Orpha', 'Villager
s', 'Idara ya maji', 'FPCT', 'WVT', 'Ir', 'DANID', 'Angli', 'secondary school', 'Amref', 'JBG', 'DADIS', 'Internat
ional Aid Services', 'RW', 'Dmdd', 'TCRS', 'RC Church', 'WATER AID', 'JICA', 'Gwasco L', 'AF', 'AMREF', 'wananch
i', 'FW', 'Central Government', 'MWE &', 'Gove', 'TDFT', 'RWE/DWE', 'Central govt', 'World Bank', 'TWESA', 'Nora
d', 'Hans', 'FinW', 'FIN WATER', 'OXFAM', 'Plan Internationa', 'District Council', 'RWEDWE', 'Fini Water', 'ANGL
I', 'CDT', 'RC CHURCH', 'North', 'Oikos E .Africa', 'SHAWASA', 'UN', 'NORAD', 'Save the rain', 'John gemuta co',
'TLC', 'RC Churc', 'Plan Int', 'Phase', 'LVIA', 'Rhobi', 'Hesawa', 'Makonde water population', 'RWE/ Community',
'Is', 'KILI WATER', 'RDDC', 'FINN WATER', 'FINI WATER', 'DHV', 'Kamama', 'DDCA', 'Victoria company', 'RWSSP', 'C
e', 'KYASHA ENTERPR', 'ERETO', 'REDESO', 'Villa', 'Priva', 'KUWAIT', 'Mw', 'Magadini-Makiwaru wa', 'Dr. Matomola',
'Af', 'RCchurch/CEFA', 'Tardo', 'GOVERNMENT', 'Individuals', 'Chamavita', 'GEN', 'Missi', 'Safari Roya', 'DAWASC
O', 'Gover', 'Mission', 'DWE/', 'Halmashauri ya wilaya sikonge', 'Ki', 'Rhoda', 'HAPA SINGIDA', 'Consulting Engine
er', 'Karugendo', 'Co', 'Marafip', 'COSMOS ENG LTD', 'World banks', 'WFP', 'Tanz', 'Handeni Trunk Main(', 'SIMBA C
O', 'Local technician', 'Village', 'Centr', 'CONS', 'DW', 'DCT', 'District water department', 'Sabodo', 'MLADE',
'I.E.C', 'LWI', 'Kiliflora', 'ICS', 'T. N. karugendo', 'DED', 'Kuwait', 'ADP', 'JUIN CO', 'TPP', 'GOVER', 'CIPRO/G
overnment', 'MWE', 'MTUWASA', 'Unisef', 'REGIONAL WATER ENGINEER ARUSHA', 'IDARA', 'Wizara ya maji', 'Tasaf and Lg
a', 'JAICA', 'KKKT-Dioces ya Pare', 'Onesm', 'Te', 'MTN', 'HESAWS', 'Islamic', 'Local', 'KTA C', 'RC', 'Killflora
/Community', 'Distri', 'Maji block', 'CALTAZ KAHAMA', 'GOVERNME', 'Omar Ally', 'HAM', 'QUWKWIN', 'ADRA', 'DO', 'D
H', 'RC Ch', 'SAXON BUILDING CONTRACTOR', 'Bokera W', 'Bulyahunlu Gold Mine', 'MBIUWASA', 'ADRA /Government', 'The
Isla', 'Rotary club', 'YELL LTD', 'KIMKUM', 'Tanesco', 'CJEJOW CONSTRUCTION', 'Victoria', 'TLTC', 'Wachina', 'WE',
'HSW', 'Communit', 'Kibaha Town Council', 'Dr. Matobola', 'Go', 'DWR', 'Huches', 'WATERAID', 'Maswi company', 'Kil
iwater', 'TA', 'wanan', 'MEM', 'Region water Department', 'Jeica', 'Ndanda missions', 'District Water Department',
'MSF/TACARE', 'Fathe', 'DARDO', 'Wa', 'MSIKIT', 'Regional Water', 'D', 'VILLAGE COUNCIL', 'RDC', 'TLC/John Majal
a', 'Kilwa company', 'Local  technician', 'TASSAF', 'VWC', 'PIDP', 'TAN PLANT LTD', 'Japan Government', 'Kata', 'G
```

```
TZ', 'ISF/Government', 'KUWASA', 'Hydrotec', 'Pr', 'Ch', 'Jaica', 'Taboma/Community', 'P', 'Ubung', 'Chur', 'BESAD
A', 'Action Contre La Faim', 'Wanjoda', 'CBHCC', 'HW/RC', 'Sumbaw', 'CCEC', 'Nice', 'CCT', 'World Vission', 'Inte
r', 'DMMD', 'WORLD BANK', 'AQUA BLUES ANGELS', 'MACK DONALD CONTRACTOR', 'Water Aid /sema', 'Henure Dema', 'Kirde
p', 'ADRA/Government', 'Kilwater', 'Da', 'Villi', 'KOYI', 'AD', 'Arab community', 'District water depar', 'HOLLAN
D', 'RC church/Central Gover', 'Active MKM', 'GEOTAN', 'LENCH', 'NCAA', 'CHINA HENAN CONSTUCTION', 'Kaembe', 'Ma',
'FinWater', 'Kuamu', 'Adra/ Community', 'Locall technician', 'UKILIG', 'Mbunge', 'The desk and chair foundat', 'DU
WAS', 'Diwani', 'Biore', 'Water aid /sema', 'KKKT CHURCH', 'EA', 'Halmashauri ya manispa tabora', 'ML appro', 'SHY
BUILDERS', 'Finwater', 'JIKA', 'Orien', 'DMDD', 'DWE}', 'CDTF', 'KAEMP', 'TUWASA', 'MARAFIP', 'MDRDP', 'Jeshi la w
okovu', 'kuwait', 'MBOMA', 'Grobal resource alliance', 'Village Council', 'Shamte Said', 'AUWASA', 'WSDP', 'COUN',
'KIDP', 'Mombo urban water s', 'TRIDEP', 'Wananchi', 'Martha Emanuel', 'St', 'GIDA contractor', 'WD and ID', 'Pade
p', 'Po', 'Village Counil', 'MINISTRY OF WATER', 'Ga', 'K', 'Swiss If', 'Miziriol', 'Yasini Selemani', 'DBSPE', 'E
uropean Union', 'H', 'TPP TRUSTMOSHI', 'Atisan', 'Jika', 'ISF/TACARE', 'Oikos E.Africa', 'Hydom Luthelani', 'Kalum
bwa', 'ILCT', 'MS', 'RUVUMA BASIN', 'Gold star', 'Mi', 'Mzungu Paul', 'Kanisa katoliki', 'Caltas', 'RED CROSS', 'W
orld bank', 'Losaa-Kia water supp', 'Jica', 'PET', 'Finland Government', 'GAICA', 'Institution', 'TCRS/TLC', 'Loli
ondo Parish', 'GACHUMA CONSTRUCTION', 'Diocese of Geita', 'Villages', 'Total landcare', 'VICTORIA DRILL CO', 'U.S.
A', 'VTECOS', 'COW', 'Vill', 'Contr', 'Wadeco', 'KIM KIM CONSTRUCTION', 'Msabi', 'VC', 'CMSR', 'Ko', 'Roman Cathol
ic Rulenge Diocese', 'Shule', 'W', 'inkinda', 'Africa Amini Alama', 'Consultant', 'L', 'Moroil', 'Sekei village co
mmunity', 'US Embassy', 'PIT COOPERATION  LTD', 'Do', 'world', 'Government /TCRS', 'UNHCR', 'DESK C', 'Dr.Matomol
a', 'FOLAC', 'Village govt', 'BSF', 'Roman Cathoric Same', 'RWE/Community', 'Mileniam project', 'Ncaa', 'Africa Is
lamic Agency Tanzania', 'Max Mbise', 'DADS', 'Institutional', 'SOWASA', 'CCPK', 'AUSTRALIA', 'not known', 'Kalago
enterprises Co.Ltd', 'Roman Catholic', 'NANRA contractor', 'No', 'ADP Busangi', 'TSRC', 'SOLIDAME', 'Barry A. Murp
hy', 'Tanzania Government', 'WILLIAMSON DIAMOND LTD', 'TAG', 'The I', 'Total Landcare', 'CENTRAL GOVERNMENT', 'Ara
bs Community', 'Secondary school', 'Water Aid/Sema', 'Jiks', 'Konoike', 'ABASIA', 'LAMP', 'SINGIDA YETU', 'RWSP',
'MDALA Contractor', 'Netherlands', 'DWT', 'TCRS /CARE', 'Makonde', 'Japan', 'Milenium', 'Goldstar', 'District COUN
CIL', 'MUWASA', 'Green', 'Kigoma municipal', 'KINAPA', 'CHINA HENAN CONTRACTOR', 'Musa', 'TANAPA', 'Ministry of wa
ter engineer', 'EFG', 'MASWI', 'Roman Cathoric -Kilomeni', 'Mbozi Secondary School', 'TASAF/DMDD', 'MWS', 'Roman c
atholic', 'Shekhe', 'Rished', 'KONOIKE', 'Pata', 'TAHEA', 'Luthe', 'Kalta', 'Pentecost church', 'Amboni Plantatio
n', 'Municipal', 'Sekondari', 'Kalitasi', 'HOTELS AND LOGGS TZ LTD', 'DISTRICT COUNCIL', 'Germany', 'Orphanage',
'WWF', 'W.B', 'IDYDC', 'SIA Ltd', 'WINAM  CONSTRUCTION', 'RIDEP', 'NORA', 'SCHOOL', 'Village community', 'Britis
h', 'Msuba', 'Villaers', 'TLC/Thimotheo Masunga', 'WB', 'Council', 'DAK', 'COCANE', 'WINAMU CO', 'Ubalozi wa Marek
ani', 'Conce', 'BGM', 'DMK', 'Mviwa', 'KA', 'MGM', 'AIMGOLD', 'YEBE CHIKOMESH', 'Omari Mzee', 'Camartec', 'Total l
and care', 'DASP', 'Islamic Agency Tanzania', 'Tanz Egypt technical coopera', 'Village Govt', 'local technician',
'TAWASA', 'WATER  AID', 'AAR', 'MSF', 'Di', 'Mackd', 'MAMAD', 'PADEP', 'Fabia', 'CONCERN', 'ITALI', 'Water aid/sem
a', 'Save the rain USA', 'Plan Tanzania', 'Roman Church', 'Singasinga', 'RC/Mission', 'In', 'V', 'Korogwe water wo
rks', 'PCI', 'Atlas', 'DWE /TASSAF', 'Local te', 'World Division', 'Gwaseco', 'Kambi Migoko', 'AI', 'Nyakilangany
i', 'DEE', 'MANYARA CONSTRUCTION', 'Rotte', 'KMCL', 'LINDALA CO', 'Government /Community', 'CCPS', 'SI', 'Rundu ma
n', 'Water Aid/sema', 'Naishu construction co. ltd', 'WOULD BANK', 'Mark', 'Cosmo', 'Halmashauri', 'Concern /gover
nment', 'Quick win project', 'Mh Kapuya', 'Halmashauri ya wilaya', 'Edward', 'COMMU', 'Baric', 'Consuting Enginee
r', 'FiNI WATER', 'CPRO', 'Jicks', 'Wahidi', 'Mohamed Ally', 'ASDP', 'CITIZEN ENGINE', 'KADP', 'Dar es salaam Tech
nician', 'Halmashauli', 'ACORD', 'MA', 'Water  Aid/Sema', 'RC church/CEFA', 'Wedeco', 'DWE/Ubalozi wa Marekani',
'VIFAFI', 'WORLD VISION', 'Cosmos Engineering', 'OLDONYOLENGAI', 'NYAKILANGANI CO', 'Village Community', 'MINJING
U', 'EL', 'Songa', 'Consultant and DWE', 'AC', 'Gain', 'DASIP', 'TANROAD', 'Tasaf', 'Wasso', 'Teonas Wambura', 'Mg
aya Masese', 'TUKWALE ENTERP', 'Sao', 'MWAKI CONTRACTOR', 'VIEN CONSTRUCTION', 'DADS/village community', 'Africar
e', 'Mosque', 'Chiko', 'central government', 'VITECOS', 'IN', 'Msikiti', 'Word Bank', 'Kwamdulu estate', 'SEMA Con
sultant', 'Concern', 'Belgiam Government', 'Wanan', 'Exaud Msambwa', 'Niger', 'MWANZA', 'SONGAS', 'MINISTRYOF WATE
```

```
sultant', 'Concern', 'Belgiam Government', 'Wahan', 'Exaud Msambwa', 'Niger', 'MWANZA', 'SONGAS', 'MINISTRYOF WATE
R', 'COMMUNITY', 'Zaburi and neighbors', 'NDM', 'Killflora/ Community', 'PART', 'secondary', "lion's club", 'luthe
ran church', 'Mileniam', 'Canada na Tanzania', 'FRANKFURT', 'GOVERM', 'Kuji foundation', 'Mamvua Kakungu', 'Rusumo
Game reserve', 'MTUWASA and Community', 'UMOJA DRILLING', 'KkKT', 'Mzinga A', 'RE', 'SUA', 'RUNDAGA', 'RWE /Commun
ity', 'Wo', 'Happy watoto foundation', 'GDP', 'ViLLAGE COUNCIL', 'MBULU DISTRICT COUNCIL', 'Maliasili', 'Roman C
a', 'NZILA', 'AFRICAN DEVELOPMENT FOUNDATION', 'FPTC', 'KARUMBA BIULDING COMPANY LTD', 'Kalugendo', 'Village Gover
nment', 'Tabraki', 'MASWI DRILLING', 'Ikela Wa', 'Shallow well', 'WEDECO/WESSONS', 'CIPRO/CARE/TCRS', 'Care intern
ational', 'Wasso contractors', 'villagers', 'Mwanga town water authority', 'Jumanne Siabo', 'Mama Kalage', 'Hind
u', 'Rural', 'TANAP', 'Makonde water supply', 'villigers', 'Bingo foundation Germany', 'Ilwilo community', 'St p
h', 'WDECO', 'LIVI', 'Pet Corporation Ltd', 'DWE & LWI', 'LC', 'KKKT Leguruki', 'HIAP', 'DIMON', 'Italy governmen
t', 'MASWI DRILL', 'WVC', 'TACRI', 'Hasnein Murij', 'Faudh Tamimu', 'Free Pentecoste Church of Tanz', 'Summit for
water/Community', 'Sanje Wa', 'Makundya', 'JANDU PLUMBER CO', 'Individual', 'OLA', 'RC C', 'TREDEP', 'Consultant E
ngineer', 'AQUA WEL', 'Cental Government', 'Nyanza road', 'Kizenga', 'KKT', 'HAPA', 'Oikos E. Africa', 'Ramadhani
Nyambizi', 'Mdala Contractor', 'DENISH', 'Mkuyu', 'GOVERN', 'GACHUMA GINERY', 'Resolute', 'Morrov', 'Serikali ya k
ijiji', 'Counc', 'Igolola community', 'S', 'NYAKILANGANI CONSTRUCTION', 'RDWS', 'Said Omari', 'AFRICA MUSLIM', 'IA
DO', 'W/', 'Ngiresi village community', 'UDC/Sema', 'AMP contractor', 'rc ch', 'QWICKWIN', 'TLC/Samora', 'Oikos E.
Afrika', 'Ruangwa contractor', 'HAYDOM LUTHERAN HOSPITAL', 'VICFISH LTD', 'Lindi contractor', 'RC CH', 'Kilomber',
'Pet Coporation Ltd', 'Afroz Ismail', 'Ja', 'commu', 'Sisal Estste Hale', 'KOREA', 'CVS Miss', 'Songas', 'Living w
ater international', 'Kajima', 'Missio', 'UAACC', 'GERMANY MISSIONARY', 'MI', 'Rips', 'LVA Ltd', 'BUKUMB', 'Taas
i', 'STAMPERS', 'Meru Concrete', 'WIZARA', 'MLAKI CO', 'Segera Estate', 'WADECO', 'Hospi', 'Cebtral Government',
'local  technician', 'Siza Mayengo', 'SAXON', 'Greec', 'KASHERE', 'GURUMETI SAGITA CO', 'China', 'MP', 'Islam', 'w
ater board', 'AMP Contract', 'Thomasi busigaye', 'Local technitian', 'SIJM', 'KKKT Ndrumangeni', 'YUMBAKA ENGINEER
ING', 'Usambala sisters', 'KOBERG Contractor', 'hesawa', 'Water Authority', 'Mr Chi', 'Hearts helping hands.Inc.',
'IDEA', 'Selous G', 'SULEMAN IDD', 'Pump entecostal Sweeden', 'Nyabarongo Kegoro', 'Canop', 'QUIK', 'DADP', 'Kanis
ani', 'CARTAS', 'Mzung', 'wizara ya maji', 'VILLAGE COUNCIL .ODA', 'CG', 'Caltus', 'Cons', 'malola', 'DCCA', 'Wate
r Project Mbawala chini', 'Unicef', 'Totoland care', 'Maswi drilling co ltd', 'NDDP', 'KMT', 'NGINIL', 'RC churc
h', 'VILLAG', 'Local technical tec', 'Cultus', 'T', 'Hery', 'OBC', 'RUDEP', 'RWE Community', 'Nyamongo Gold minin
g', 'Redep', 'Norani', 'Mahita', '-', 'Villag', 'germany', 'KARUMBA BIULDIN', 'AIXOS', 'Selikali', 'DDP', 'Village
government', 'Zacharia MTN', 'Africa', 'PAD', 'KASHWA', 'TWENDE PAMOJA', 'Uhai wa mama na mtoto', 'OLOMOLOKI', 'Ar
dhi water well', 'Distric Water Department', 'Conta', 'SHUWASA', 'Makori', 'Sangea District Coun', 'CHINA', 'Briti
sh colonial government', 'Maendeleo ya jamii', 'CARITAS', 'Taes', 'KWIKWIZ', 'SEMA CO LTD', 'SENAPA', 'REGWA COMPA
NY OF EGYPT', 'COBASHEC', 'AQUA Wat', 'Dr.Matobola', 'Central basin', 'Mamlaka ya maji ngara', 'PRF', 'Church', 'M
agadini Makiwaru wa', 'Mpang', 'KAYEMPU LTD', 'TRACHOMA', 'FURAHIA TRADING', 'HESAW', 'Moravian', 'Samsoni', 'MD',
'GURUMETI SAGITA', 'Songea District Coun', 'Cast', 'N.P.R.', 'Panone', 'Hemed Abdallah', 'Lawate fuka water su',
'St Gasper', 'Ha', 'MMG GOLD MINE', 'P.N.R.', 'Nandra Construction', 'Mchuk', 'African Realief Committe of Ku', 'S
COTT', 'D$L', 'Vi', 'JLH CO LTD', 'Msiki', 'Namungo', 'Nassor Fehed', 'TWESA /Community', 'DBFPE', 'EF', 'Serikal
i', 'Mgaya Mwita', 'Clause workers', 'MLAKI  CO', 'Busoga trust', 'mzee mabena', 'SAUWASA', 'NORAD/', 'BR', 'local
technitian', 'Comunity', 'Brad', 'Tanganyika Basin', 'MORNING CONSTRUCTION', 'Healt', 'Governme', 'Roma', 'KUMKU
M', 'PNR co', 'Muslims', 'Paffec', 'Tansi', 'CRAELIUS', 'APM', 'Zao water spring X', 'TASA', 'CSPD', 'DALDO', 'VIF
AF', 'MTC', 'TCRS Kibondo', 'Howard and humfrey consultant', 'RUDEP/', 'LUNGWE', 'Dhinu', 'AIC KI', 'Mataro', 'FIN
I Water', 'Mombo urban water', 'REDAP', 'Kagulo', 'TMP', 'Nimrod Mkono[mb]', 'Red Cross', 'SHULE', 'Maro', 'WEDEK
O', 'MSABI', 'UN ONE', 'BRA', 'MasjId Takuar', 'SUWASA', 'TWIG', 'Tanzania Egypt Technical Co Op', 'TCRS.TLC', 'Lu
theran', 'TASF', 'RC CATHORIC', 'TASAF and Comunity', 'world banks', 'Eliza', 'MAKE ENGINEERING', 'Halmashauri wil
aya', 'EFAM', 'TASAF/', 'Mgaya', 'Grail Mission Kiseki bar', 'RUDE', 'local technical tec', 'Lga', 'JHL CO LTD',
'Ansnani Murij', 'LIUWASSA', 'USTAWI', 'GERMAN', 'NSSF', 'Cefa', 'Kilol', 'Judge Mchome', 'Milenia', 'AMP Contract
```

```
s', 'Masjid', 'MSIKITI', 'Government /SDA', 'FARM-AFRICA', 'Mama Agnes Kagimbo', 'UNIVERSAL CONSTRUCTION', 'BFFS',
'KYELA_MOROGORO', 'Msikitini', 'HDV', 'Shelisheli commission', 'JALCA', 'Mungaya', 'AIC', 'Boni', 'TGTS', 'TCRS /G
overnment', 'Adam mualuaka', 'unknown', 'IRC', 'ADB', 'Enyueti', 'Regina group', 'Seram', 'ADAP', 'JI', 'Laizer',
'Salehe', 'MASAI LAND', 'HPA', 'People P', 'Oikos E Africa', 'AQUARMAN DRILLERS', 'W.D &', 'Wafidh', 'Kitukuni wat
er supply', 'Rural Drinking Water Supply', 'Morovian church', 'WOYEGE', 'ATIGH BUILDINGS', 'Muslimu Society(Shi
a)', 'Pankrasi', 'IREVEA SISTER', 'IUCN', 'KDC', 'Morovian Church', 'JAPAN EMBASSY', 'INDIVIDUALS', 'Dwe', 'Italia
n government', 'Marti', 'R.C', 'GREINAKER', 'Totoland', 'Bahresa', 'Mwalimu  Muhenza', 'CEFA/rc church', 'PRINCE M
EDIUM SCHOOL', 'Kaluwike', 'TCRS /TWESA', 'MAKAMA CONSTRUCTION', 'Tanzania', 'Seleman Masoud', 'RC mission', 'Babu
Sajini', 'W.D. and I.D.', 'Word divisio', 'Ardhi Instute', 'DIOCESE OF MOUNT KILIMANJARO', 'VICKFI', 'JUIN', 'Sing
ida General  Supplies  Ltd', 'ISF / TASAFF', 'DE', 'Ubalozi wa Marekani /DWE', 'Muwaza', 'Chacha Issame', 'Villege
Council', 'Kilimarondo Parish', 'NYAKILANGANI', 'Tempo', 'Danda', 'SIMAVI', 'RURAL WATER SUPPLY', 'Rotary Club',
'COYI', 'Yakwetu Contractor', 'CGI', 'Ta', 'BRUDER', 'TRUST', 'shule', 'Said Hashim', 'TLC/Nyengesa Masanja', 'Mio
mb', 'Staford Higima', 'CF Builders', 'Waheke', "LION'S", 'Icf', 'private', 'Ardhi Water Wells', 'TWESA/ Communit
y', 'Private individuals', 'MISHENI', 'MASWI DRILLING CO. LTD', 'GGM', 'SPAR DRILLING', 'John kiminda co', 'Missio
nary', 'FAUSTINE', 'Gwasco', 'ms', 'District  Council', 'Engin', 'PMO', 'Village council', 'NIRAD', 'COWI', 'LGCD
G', 'EGYPT REGWA', 'Athumani  Janguo', 'Geita Goldmain', 'GREINEKER', 'Tareto', 'Teresa Munyama', 'KKKT DME', 'Mas
wi', 'MAJI MUGUMU', 'Issa Mohamedi Tumwanga', 'Morovian', 'Fin water', 'American', 'Anglican Church', 'BENGUKA',
'William Acles', 'Jackson Makore', 'MISSION', 'Mr Kwi', 'Hanja Lt', 'ABDUL', 'Mwanamisi Ally', 'KAWINGA', 'Tanap
a', 'ODA', 'ACTION AID', 'Juma Maro', 'W.C.S', "Okong'o", 'Water Department', 'VITECOS INVEST', 'local fundi', 'LE
I', 'Water department', 'St Elizabeth Majengo', 'JSICA', 'Calvary connection', 'TANZANIAN GOVERNMENT', 'Local tech
nical', 'Stephano', 'JAWABU', 'J. Mc', 'OLS', 'wasab', 'Domnik', 'Nyakaho Mwita', 'G.D&I.D', 'Daniel', 'Ifakara',
'EMAYO', 'Amec', 'Rotery c', 'DAR ES SALAAM ROUND TABLE', 'VICTORIA DRILL', 'Rc', 'BKHWS', 'Njula', 'Nampopanga',
'marafip', 'Mzee Waziri Tajari', 'TUMAINI FUND', 'Fin Water', 'BEMANDA', 'Nassan workers', 'Consulting engineer',
'Quick win project /Council', 'Piscop', 'Farm Africa', 'is', 'FINLAND', 'Mwl. Nyerere sec. school', 'DWE&', 'Rps',
'Idara ya Maji', 'Church Of Disciples', 'COEK', 'nandra Construction', 'Marumbo Community', 'Ju', 'VW', 'PATUU',
'TANAS', 'Nyeisa', 'Mwakabalula', 'TASSAF /TCRS', 'Machibya', 'MDRD_', 'WORDL BANK', 'ANGLIKANA CHURCH', 'DDCA C
O', 'Mianz', 'Desk and chair foundation', 'ELCT', 'Ungan', 'Eastmeru medium School', 'EMANDA', 'ENO', 'Losa-kia wa
ter suppl', 'Secondary', 'Wilson', 'KILANGANI CO', 'Africa Muslim Agenc', 'WINAM CO', 'Ar', 'Mbozi District Counci
l', 'Village  Council', 'Villagerd', 'LUWASSA', 'Raymond Ekura', 'JAICA CO', 'J LH CO LTD', 'Bobby', 'Municipal Co
uncil', 'ACTIVE TANK CO LTD', 'Quik', 'Concen', 'Tom', 'Howard and Humfrey Consultants', 'Zuber Mihungo', 'Mwl. Ny
erere sec.school', 'SIDA', 'Efarm', 'NG', 'TANZAKESHO', 'RWI', 'ACTIVE TANK CO', 'Mzee Omari', 'Msig', 'Overland H
igh School', 'KAGERA MINE', 'DWE/TASSAF', 'Adam Kea', 'Rashid Mahongwe', 'NAFCO', 'Belgij', 'Kalitesi', 'Water use
r Group', 'MW', 'harison', 'MIDA', 'Plan International', 'Makuru', 'MSIGWA', 'Singida yetu', 'MINISTRY OF EDUCATIO
N', 'Centra govt', 'HESAWZ', 'CONCE', 'B.A.P', 'R', 'Nasan workers', 'TWESS', 'Wizara  ya maji', 'Water Hu', 'KK',
'CIP', 'Monmali', 'DW$', 'KARUMBA BIULDING CONTRACTOR', 'Maji Tech', 'DSPU', 'Nu', 'AFRICA', 'CCP', 'Upendo Grou
p', 'GRUMETI SINGITA', 'WA', 'Insititutiona', 'kanisa', 'Colonial Government', 'TUKWARE ENTERP', 'ANGRIKANA', 'chu
rch', 'Anglican church', 'TASAFcitizen and LGA', 'SHIP', 'Zingibali Secondary', 'KAEM', 'Tajiri Jumbe Lila', 'SAXO
N  BUILDING CONTRACTOR', 'Ngelepo group', 'VILLAGERS', 'Nduku village', 'Amadi', 'Jafary Mbaga', 'Sa', 'Water hu',
'Luleka', "TLC/Seleman Mang'ombe", 'Lutheran Church', 'Railway', 'Laramatak', 'TASAF and MMEM', 'DSV', 'WUA', 'Sal
eh Zaharani', 'HESAWQ', 'Action Contre la Faim', 'KIDIJAS', 'Mwalimu  Muhenzi', 'Heri mission', 'Africaone', 'Misr
i Government', 'Gtz', 'GLOBAL RESOURCE CONSTRUCTION', 'GERMAN MISSIONSRY', 'Total land Care', 'Tanzanian Governmen
t', 'LOMOLOKI', 'Halmashauri ya mburu', 'UMOJA DRILLING CONSTRUCTION', 'Mayiro', 'K/Primary', 'DANIDA CO', 'TSCR',
'Mohamad Masanga', 'EWE', 'VILLAGER', 'SCHOO', 'Atlas Company', 'Got', 'CIPRO', 'Sacso', 'NMDC INDIA', 'NSC', 'Wat
er  Aid/Maji tech', 'Hussein Ayubu', 'Government and Community', 'COMMUNITY BANK', 'Villager', 'TASAF 1', 'Ns', 'M
```

```
kulima', 'Baadela', 'SAIDI CO', 'SOLIDARM', 'Filber', 'Runduman', 'Tanza', 'DSP', 'Rotary club Australia', 'J mal
Abdallah', 'Rotar', 'Anglikana', 'Private owned', 'KARUMBA BUILDING COMPANY LTD', 'Maswi Company', 'Kuwaiti', 'MAC
K DONALD CO LTD', 'DASSIP', 'Yoroko mwalongo', 'Subvillage', 'Holili water supply', 'MTAMBO', 'Nyabweta', 'UDC/sem
a', 'Misana george', 'Livi', 'Moyowosi', 'DAWASA', 'Pet  Corporation  Ltd', 'Mtwara Technician', 'ISSAA KANYANGE',
'Megis', 'MANDIA CONSTRUCTION', 'Recoda', 'USAID', 'African Muslims Age', 'Serikari', 'RO', 'BGSS', "NGO'S", 'ANSW
AR', 'DMDD/SOLIDER', 'Word bank', 'KOICA', 'Team Rafiki', "TAG Patmo's", 'Water use Group', 'KKKT Kilinga', 'O',
'Buguba', 'Babu Sajin', 'Sh', 'wananchi technicians', 'Cetral government /RC', 'DANNY', 'Indi', 'Billy Phillips',
'Wamissionari wa kikatoriki', 'Kapelo', 'Water authority', 'local', 'JANDU PLUMBER  CO', 'PNR Da', 'Tanz/Egypt tec
hnical coopera', 'Masjid Nnre', 'Ahmad', 'Dydrotec', 'Red cross', 'DANIAD', 'Private Technician', 'JACKSON MAHAMB
O', 'Unknown', 'Rombo Dalta', 'Jeshi la wokovu [cida]', 'Mwita Machoa', "Lion's", 'WDP', 'GRA', 'SDP', 'Pentecosta
l church', 'KISIRIRI ADP', 'Kitiangare village community', 'Bridge north', 'Mtewe', 'ONESM', 'DFID', 'Ox', 'Water
AID', 'Water /sema', 'Presadom', 'ir', 'Seif Ndago', 'AQUA Wel', 'Linda', 'Inves', 'Dawasco', 'Sweeden', 'KEREBUK
A', 'School Adminstrarion', 'Friend from UN', 'Mombo urban  water', 'Kijiji', 'Village Technician', 'Building work
s Company Ltd', 'F', 'VTTP', 'Zao', 'TECH SUPPORT BEST CO', 'CG/RC', 'MH Kapuya', 'E ETO', 'Ardhi Water well', 'Ju
ma Makulilo', 'Others', 'Tabora Municipal Council', 'Friedkin conservation fund', 'Mh.chiza', 'NGO', 'IS', 'CARTAS
Tanzania', 'Grobal resource  alliance', 'COEW', 'CHELA', 'Mosqure', 'DESK a', 'KKKT Canal', 'RC MISSIONARY', 'Mako
ye', 'Bingo foundation', 'WB / District Council', 'Lindi rural water department', 'KILL WATER', 'Active KMK', 'Arr
ian', 'FIDA', 'Mzee Yassin Naya', 'Amboni plantation', 'TANEDAPS Society', 'KOWI', 'MAISHULE', 'GRA TZ MUSOMA', 'H
emed Abdalkah', 'Hamisi Fidia', 'Socie', 'UNDP', 'SUNAMCO', 'Jimmy', 'Hesewa', 'QUICKWINS', 'ambwene mwaikeke', 'R
EGWA', 'Mpango wa Mwisa', 'DODDEA', 'Marijan Ally Dadi', 'UNICRF', 'plan Int', 'world vision', 'Concern/Governmen
t', 'Oldadai village community', 'Emmanuel Kiswagala', 'ENGINEERS WITHOUT BORDER', 'ABDALA', 'Shule ya msingi', 'W
EEPERS', 'Goldwill  foundation', 'TWESA/Community', 'Mu', 'LWI &CENTRAL GOVERNMENT', 'Obadia', 'MAZI INVESTMENT',
'Benjamin', 'Muham', 'Company', 'mchina', 'Townsh', 'ABD', 'Abdallah Ally Wazir', 'Hospital', 'NYAKILANGANI  CO',
'CJEJOW', 'Lions club kilimanjaro', 'George', 'BioRe', 'SOLIDERM', 'Makonde water Population', 'WASHIMA', 'Naishu
Construction Co. ltd', 'WORLD NK', 'MCHOME', 'SSU', 'mwakalinga', 'Samwel', 'KKKT Katiti juu', 'Romam', 'Nathal Ha
madi', 'Pet corporation Ltd', 'Marke', 'Cathoric', 'Bonite Bottles Ltd', 'SDA CHURCH', 'Kigwa', 'DADS/Village comm
unity', 'Luali Kaima', 'Mama Hamisa', 'METHODIST CHURCH', 'DIWANI', 'George mtoto', 'DW E', 'JWTZ', 'Wajerumani',
'Mama joela', 'TLC/community', 'World Visiin', 'Napupanga', 'MOSQUE', 'morovian church', 'desk and chair foundatio
n', 'GRUMET', 'Q-sem Ltd', 'Motiba Manyanya', 'MECO', 'Neemia mission', "Rashid Seng'ombe", 'Msagin', 'Vodacom',
'Altai Co. ltd', 'Chuo', 'ZINDUKA', 'MUSLIMEHEFEN INTERNATIONAL', 'SERONERA', 'Roman Cathoric -Same', 'Richard M.K
yore', 'Maseka community', 'Makala', 'Wamisionari wa Kikatoriki', 'HAAM', 'Ubalozi wa Japani', 'Al Ha', 'Latifu',
'Islamic community', 'Halmashauri/Quick win project', 'Kiliwater r', 'MWL NGASSA', 'RunduMan', "Lion's club", 'Cou
n', 'Foreigne', 'Wasso companies', 'Mbozi Hospital', 'Building works engineering Ltd', 'Kanamama', 'sengerema Wate
r Department', 'CJEJ0', 'Masele Nzengula', 'Care  international', 'KKKT MAREU', 'TGT', 'British government', 'Mini
stry of water', 'TRC', 'Magani', 'CHONJA CHARLES', 'WAMBA', 'Hesawz', 'CHRISTIAN OUTRICH', 'KC', 'District Communi
ty j', 'ROMAN CATHOLIC', 'COCU', 'Robert Mosi', "Ng'omango", 'salamu kita', 'INDIVIDUAL', 'Kassim', 'Seff Mtambo',
'Halmashauri ya manispa  tabora', 'Patrick Nyanzwi', 'MAJ MUGUMU', 'MKON CONSTRUCTION', 'BESADO', 'Embasy of Japan
in Tanzania', 'MASWI CO', 'School', 'FAO', 'KIBO', 'Adam', 'Ilolangulu water supply', 'Steven Nyangarika', 'Jere
m', 'People from Egypt', 'Tumaini fund', 'Kinga', 'Yohanis Mgaya', 'HASHI', 'Elina', 'CHRISTAN OUTRICH', 'NJOONJO
O', 'RC Msufi', 'Chacha', 'PWD', 'Action Aid', 'lusajo', 'Frida mokeki', 'Salum Tambalizeni', 'Primo', 'VILLAGE WA
TER COMMISSION', 'villager', 'Shingida yetu', 'TANCRO', 'TAIPO', 'TCRS TWESA', 'Angrikana', 'HAIDOMU LUTHERAN CHUR
CH', 'Kibo potry', 'FRESH WATER PLC ENGLAND', 'Mashaka M', 'Safe Rescue Ltd', 'Rhobi Wamburs', 'Great Lakes', 'Mke
to', 'LGQ', 'UN Habitat', 'SIMBA', 'kegocha', 'Egypt Government', 'Perusi Bhoke', 'DADS/village Community', 'FinWa
te', 'Li', 'Village Office', 'Handeni Trunk Main', 'Village community members', 'Aartisa', 'RC CHURCH BROTHER', 'Q
```

```
UKWIN', 'Mwigicho', 'Private person', 'Localtechnician', 'WATER', 'RWE/TCRS', 'MWAKI CONTRACTO', 'Scholastica Pank
rasi', 'DEW', 'Zao water spring', 'Member of Perliament Ahmed Ali', 'AGRICAN', 'Tadeo', 'Nchagwa', 'Theo', 'Paskal
i', 'Matiiti', 'Shule ya sekondari Ipuli', 'Dokta Mwandulami', 'Adrs', 'Mara inter product', 'Tanzania/ Egypt', 'W
ater users Group', 'CRISTAN OUTRICH', 'IDC', 'Water boards', 'IRAN GOVERN', 'Rotary Club of Chico and Moshi', 'KYA
SHA ENTREPR', 'Indiv', 'JUINE CO', 'Malec', 'Kagunguli Secondary', 'BOAZI', 'AQAL', 'Males', 'C', 'TBL', 'TCRS/DW
E', 'Kando', 'Egypt Technical Co Operation', 'MBWAMBO', 'PIUS CHARLES', 'VWT', 'REGWA Company', 'St Magreth Churc
h', 'Schoo', 'KOBERG', 'MASWI COMPANY', 'Quick win/halmashauri', 'GD&ID', 'Taees', 'Governmen', 'Shule ya msingi u
fala', 'HesaWa', 'Mzee Smith', 'Mkuluku', 'TCRS/village community', 'Kamata project', 'Chama cha Ushirika', 'Clave
r', 'WAMA', 'Gerald', 'Colonial government', 'Pentecosta', 'Jeshi la Wokovu', 'Mrish', 'WINAM CONSTRUCTION', 'MSJI
MUGUMU', 'DANNIDA', 'WBK', 'SAXON BUILDING CONTRACTORS', 'MP Mloka', 'India', 'rc church', 'Government/TCRS', 'Ang
lica Church', 'Noshadi', 'GLOBAL RESOURCE CO', 'MACK DONALD CONTRSCTOR', 'Makusa', 'Rotary Club of USA and Moshi',
'TANCAN', 'CHANI', 'DW#', 'Athumani Issa', 'O &', 'Msudi', 'WATER  AIDS', 'SINGIDA  YETU', 'Mzee Salum Bakari Daru
s', 'Crety', 'Mahemba', 'RWET/WESA', 'upper Ruvu', 'CHINA Co.', 'Nampapanga', 'Compa', 'KDPA', 'kw', 'Mwamvita Raj
abu', 'TINA/Africare', 'Rotary club kitchener', 'ICF/TWESA', 'Mwamama', 'Mbwiro', 'KU', 'Morovi', 'FLORESTA', 'Wor
d', 'TCRS/TWESA', 'GEOCHAINA', 'Kiwanda cha Ngozi', 'Carmatech', 'go', 'Nyamingu subvillage', 'plan int', 'Sister
makulata', 'DDSA', 'SHIPO CONSTRUCTORS', 'Christopher', 'Matogoro', 'RC Mi', 'LUKE SAMARAS LTD', 'ICAP', 'Winkyen
s', 'CIPRO/CARE', 'Deogra', 'Mr Sau', 'Muhindi', 'Sumry', 'BAPTIST CHURCH OF TANZANIA', 'A.D.B', 'OMARY MONA', 'Wa
ter aid', 'Samweli', 'UYOGE', 'Waitaliano', 'TASSAF/ TCRS', 'WUS', 'Robert kampala', 'TASAF/TLC', 'Elias Mahemba',
'CHENI', 'TMN', 'DWEB', 'ter', 'TLC/Jenus Malecha', 'Prof. Saluati', 'Ardhi and PET Companies', 'Morrovian', 'Tara
ngire park', 'Not kno', 'DV', 'Region water', 'DHV Moro', 'VILLAGE', 'FPCT Church', 'RC Njoro', 'Kindoroko water p
roject', 'Mama Kapwapwa', 'Birage', 'Kikom', 'UDEA', 'Mambe', 'Mwita Mahiti', 'DANIDS', 'H4CCP', 'Anglikan', 'Vill
age water committee', 'TLC/Sorri', 'Africaone Ltd', 'BALYEH', 'EMANDA BUILDERS', 'Unknown Installer', 'SADIKI KANG
ELO', 'MSUKWA CONSTRUCTION COMPANY', 'mwakifuna', 'Tanzania government', 'Emmanuel kitaponda', 'mbeje', 'Wizra ya
maji na egypt', 'AFRICAN REFLECTIONS FOUNDATION', 'MBIUSA', 'Maji tech Construction', 'PRIV', 'DBSP', 'Manyota pri
mary School', 'Ramadhani M. Mvugalo', 'Naishu construction co.ltd', 'REGWA COMPANY OF EGPTY', 'Masese', 'TLC/Emman
uel Kasoga', 'LIZAD', 'WDE', 'MKONG CONSTRUCTION', 'TANGA CEMENT', 'p', 'WW', 'Tanload', 'Othod', 'BOMA SAVING',
'MBULI CO', 'Saidi Halfani', 'MKONGO BUILDING CONTRACTOR', 'Prima', 'Sua', 'Government /World Vision', 'Nyamwanj
i', 'Privat', 'RC .Church', 'Bao', 'ALIA', 'Madra', 'Anglican', 'EGYPT', 'DAWE', 'SUMO', 'Sophia Wazir', 'Kuweit',
'BABTEST', 'TCRS /DWE', 'Charlotte Well', 'Hanja', 'Jumuhia', 'JAPAN', 'Village water attendant', 'George mtoto co
mpany', 'Mwananchi Engineeri', 'REDEP', 'Leopad Abeid', 'TCRS/ TASSAF', 'Nyamasagi', 'maji mugumu', 'College', 'RE
SOLUTE MINING', 'RC Mis', 'RC MISSION', 'UMOJA DRILLING CONTRACTOR', 'Ambrose', 'BATIST CHURCH', 'KURRP', 'Water A
id/DWE', 'Rudri', 'Alex moyela', 'Centra Government', 'JANDU', 'CH', 'Aqual', 'DWW', 'CHANDE CO', 'Village local c
ontractor', 'Yasini', 'School Adm9nstrarion', 'BIORE', 'The Co', 'Jacks', 'sengerema water Department', 'Rc Missio
n', 'Jumaa', 'ESAWA', 'Kanisa la TAG', 'WSSP', 'MOSES', 'Pentekoste', 'UMOJA DRILLING CONTRUCTO', 'Amari', 'MKONGO
CONSTRUCTION', 'Kkkt', 'Simon Lusambi', 'Chinese', 'Moshono ADP', 'DESK A', 'LDEP', 'NYAHALE', 'THREE WAY GERMAN',
'LOLMOLOKI', 'Internal Drainage Basin', 'Water Solution', 'Regwa Company', 'VICF', 'Aqwaman Drilling', 'Hilfe Fur
Bruder', 'S.P.C Pre-primary School', 'MTUI', 'Omar Rafael', 'Mwita Lucas', 'Hamis Makombo', 'BUMABU', 'Manyovu Agr
iculture Institute', 'Gerald Mila', 'Natio', 'Region Water Department', 'Simango Kihengu', 'M', 'HEESAW', 'Goldmai
n', 'M and P', 'MASU COMPANY', 'Africa M', 'Rombo delta', 'MIAB', 'GETDSC00', 'Private company', 'TZ as', 'Luka',
'Jumuiya', 'Arisan', 'Makanya Sisal Estate', 'maendeleo ya jamii', 'Rural Drinkung Water Supply', 'WWF/', 'john sk
wese', 'peter', 'CHURC', 'Enyuati', 'Mr Luo', 'Noshad', 'NDRDP', 'Ongan', 'Nyabibuye Islamic center', 'STABEX', 'P
ori la akiba kigosi', 'Britain', 'Losakia water supply', 'FILEX MUGANGA', 'Local l technician', 'MANGO TREE', 'Ril
ayo water project', 'Sent Tho', 'UPM', 'Magul', 'Magoma ADP', 'Swalehe Rajabu', 'Kidika', 'TCRS/ TWESA', 'Kahema',
'Missionaries', 'GRUMENTI', 'Buruba', 'PRIVATE INSTITUTIONS', 'Kauzeni', 'Paul', 'Juma', 'TCRS a', 'Hasawa', 'TWES
```

```
A/JAMII', 'Joseph nkunda', 'Halimashauri', 'Mr Kas', 'Upendo primary School', 'SIPDO', 'TOVE', 'Nerthlands', 'CA
P', 'Cida', 'TASSAF/TCRS', 'DWE/Anglican church', 'VIFAI', 'Dina', 'brown', 'SELEPTA']
```

In [17]:

```
#exporting installer column details to a file
data['installer'].to_csv('data/column_details.csv', index=False)
```

In [18]:

```
# correcting typos on the installer column

data["installer"] = data["installer"].replace(to_replace=(
    'World vision', 'World Vision', 'world vision', 'World Visiin', 'World Division',
    'WORLD VISION'), value="World Vision")

data["installer"] = data["installer"].replace(to_replace=(
    'Central government', 'Government', 'Central Government','GOVERNMENT', 'Tanzania Government',
    'CENTRAL GOVERNMENT', 'Government /Community', 'Concern /government', 'central government',
    'Cental Government','Cebtral Government', 'TANZANIAN GOVERNMENT', 'Tanzanian Government',
    'Government and Community', 'Cetral government /RC', 'Tanzania government', 'Centra Government',
    'GOVERNME', 'GOVER', 'Gover','Central govt', 'Gove', 'GOVERM', 'GOVERN',
    'Centra govt', 'Cetral government', 'Governmen', 'TZ as', 'Jumuiya', 'Jumuhia', ),
                                             value="Central Government")

data["installer"] = data["installer"].replace(to_replace=(
    'World Bank', 'World banks', 'WORLD BANK', 'World bank','WOULD BANK', 'Word Bank',
    'world banks', 'WORDL BANK', 'Word bank', ),
                                             value="World Bank")

data["installer"] = data["installer"].replace(to_replace=('UNICRF', 'Unicef', 'Unisef'),
                                             value="UNICEF")

data["installer"] = data["installer"].replace(to_replace=(
    'JICA', 'JIKA', 'Jika', 'Jica', 'Jiks', 'JAICA','Jaica', 'JAICA CO', 'Jeica', 'GAICA'),
                                             value="JICA")

data["installer"] = data["installer"].replace(to_replace=(
    'Olgilai village community', 'VILLAGE COUNCIL Orpha', 'Villagers', 'Villa',
    'Villages', 'Vill', 'Village', 'VILLAGE COUNCIL', 'Villi', 'Village Council',
    'Village Counil', 'Sekei village community', 'Village govt', 'Village community',
    'Villaers', 'Village Govt', 'ViLLAGE COUNCIL', 'villagers', 'Village Government',
     'villigers', 'VILLAGE COUNCIL .ODA','VILLAG', 'Villag', 'Villege Council',
     'Villagerd','VILLAGERS', 'Nduku village','Subvillage', 'Kitiangare village community',
     'Village Technician', 'Oldadai village community', 'VILLAGE WATER COMMISSION',
     'villager','Village Office', 'Village community members', 'Nyamingu subvillage',
     'Village water committee', 'Village water attendant', 'Village local contractor'),
```

```
                                        value="VILLAGE")

data["installer"] = data["installer"].replace(to_replace="0", value="Unknown")

data["installer"] = data["installer"].replace(to_replace=(
    'Adra/ Community',  'Arab community', 'Taboma/Community','Communit', 'Killflora /Community',
    'RWE/ Community', 'Olgilai village community', 'Commu', 'Sekei village community', 'COMMU',
    'COMMUNITY', 'Ilwilo community', 'Igolola community', 'Comunity', 'Marumbo Community',
    'Maseka community','Islamic community', 'District Community j' ),
                                        value="COMMUNITY")

data["installer"] = data["installer"].replace(to_replace=(
    'Fini water', 'FinW', 'FIN WATER','FINN WATER', 'FINI WATER', 'FinWater',
    'FiNI WATER', 'FINI Water', 'Fin water', 'FinWate', ),
                                        value="FINI WATER")

data["installer"] = data["installer"].replace(to_replace=(
    'District council','District Counci' 'District Council','District water department',
    'Distri', 'District Water Department', 'District water depar', 'District COUNCIL',
    'DISTRICT COUNCIL', 'MBULU DISTRICT COUNCIL', 'Sangea District Coun',
    'Songea District Coun', 'Mbozi District Council'),
                                        value="DISTRICT COUNCIL")

data["installer"] = data["installer"].replace(to_replace=(
    'RC Church', 'RC CHURCH','RC Churc', 'RCchurch/CEFA', 'Chur',
    'RC church/Central Gover', 'KKKT CHURCH', 'Pentecost church', 'Roman Church',
    'RC church/CEFA', 'lutheran church', 'Free Pentecoste Church of Tanz', 'RC C',
    'Church', 'Morovian church', 'CEFA/rc church', 'Anglican Church', 'Church Of Disciples',
    'ANGLIKANA CHURCH', 'ANGRIKANA', 'church', 'Anglican church', 'kanisa',
    'Lutheran Church', 'Pentecostal church','Jeshi la wokovu [cida]', 'METHODIST CHURCH',
    'CHURCH', 'morovian church', 'Angrikana', 'HAIDOMU LUTHERAN CHURCH','RC CHURCH BROTHER',
    'St Magreth Church', 'Pentecosta','rc church', 'Anglica Church', 'RC Mi',
    'BAPTIST CHURCH OF TANZANIA', 'FPCT Church', 'RC .Church','BATIST CHURCH', 'CHURC',
    'DWE/Anglican church', 'RC Churc','RCchurch/CEFA', 'RC', 'RC Ch','HW/RC' 'RC CH',
    'rc ch','RC CATHORIC', 'RC mission',  'Church Of Disciples', 'CG/RC', 'RC MISSIONARY', 'RC Msufi',
    'Rc Mis','Jeshi la Wokovu'),
                                        value="CHURCH")

data["installer"] = data["installer"].replace(to_replace=(
    'DANIDA', 'Danid', 'DANID', 'DANIDA CO','DANIAD', 'DANIDS' ),
                                        value="DANIDA")

data["installer"] = data["installer"].replace(to_replace=(
    'HESAWA', 'Hesawa', 'HESAWS',  'hesawa', 'HESAW', 'HESAWZ', 'HESAWQ', 'Hesawz', 'HesaWa',   ),
```

```
                                              value="HESAWA")

data["installer"] = data["installer"].replace(to_replace=(
    'KKKT _ Konde and DWE','KKKT','KKKT-Dioces ya Pare', 'KkKT', 'KKKT Leguruki',
    'KKKT Ndrumangeni','KKKT DME', 'KKKT Kilinga', 'KKKT Canal', 'KKKT Katiti juu',
    'KKKT MAREU', 'Kkkt'),
                                              value="KKKT")
# viewing the top 30 installer values after initial cleaning
data["installer"].value_counts().head(30)
```

Out[18]:

```
DWE                   17361
Central Government     3587
DANIDA                 1676
HESAWA                 1225
RWE                    1203
COMMUNITY              1162
KKKT                   1091
DISTRICT COUNCIL        817
Unknown                 780
TCRS                    702
World Vision            671
CHURCH                  647
CES                     610
FINI WATER              572
Community               552
District Council        546
VILLAGE                 514
JICA                    427
LGA                     408
WEDECO                  397
TASAF                   390
UNICEF                  333
TWESA                   316
AMREF                   313
WU                      301
Dmdd                    286
ACRA                    277
SEMA                    249
DW                      246
OXFAM                   234
Name: installer, dtype: int64
```

In [19]:

```
# converting all the categories to "OTHERS" and keeping only the top 30 categories for installer column
```

```
top_30_installer = data["installer"].value_counts(normalize=True).nlargest(30).index

data["installer"]= [x if x in top_30_installer else "OTHERS" for x in data["installer"]]
data["installer"].value_counts()
```

Out[19]:

```
DWE                   17361
OTHERS                17209
Central Government     3587
DANIDA                 1676
HESAWA                 1225
RWE                    1203
COMMUNITY              1162
KKKT                   1091
DISTRICT COUNCIL        817
Unknown                 780
TCRS                    702
World Vision            671
CHURCH                  647
CES                     610
FINI WATER              572
Community               552
District Council        546
VILLAGE                 514
JICA                    427
LGA                     408
WEDECO                  397
TASAF                   390
UNICEF                  333
TWESA                   316
AMREF                   313
WU                      301
Dmdd                    286
ACRA                    277
SEMA                    249
DW                      246
OXFAM                   234
Name: installer, dtype: int64
```

Cleaning Numerical Data *Construction Year*

In [20]:

```
data["construction_year"].value_counts().sort_index(ascending=True)
```

Out[20]:

```
0        18392
```

```
Out[20]: 0       [illegible]
         1960       45
         1961       20
         1962       29
         1963       84
         1964       40
         1965       19
         1966       17
         1967       83
         1968       68
         1969       59
         1970      310
         1971      145
         1972      705
         1973      183
         1974      675
         1975      437
         1976      411
         1977      199
         1978     1027
         1979      192
         1980      647
         1981      237
         1982      741
         1983      487
         1984      777
         1985      941
         1986      431
         1987      297
         1988      520
         1989      316
         1990      666
         1991      322
         1992      632
         1993      595
         1994      703
         1995      978
         1996      766
         1997      612
         1998      921
         1999      950
         2000     1565
         2001      530
         2002     1064
         2003     1276
```

```
2004     1107
2005      983
2006     1447
2007     1557
2008     2568
2009     2490
2010     2427
2011     1211
2012     1025
2013      173
Name: construction_year, dtype: int64
```

There are a lot of entries (18392) with "0" construction year. This could imply their year of construction is unknown or its a natural water point e.g. natural springs. Replacing all years having "0" values with "1960" (the minimum year) so as to assist with modelling and visualization:

In [21]:

```
data["construction_year"].replace(to_replace=0, value=1960, inplace=True)
```

In [22]:

```
# checking the statistics of numerical variables
data.describe()
```

Out[22]:

	amount_tsh	population	permit	construction_year
count	55102.000000	55102.000000	55102.000000	55102.000000
mean	326.595438	182.670556	0.693169	1984.575551
std	2670.687601	467.570627	0.461183	20.147098
min	0.000000	0.000000	0.000000	1960.000000
25%	0.000000	0.000000	0.000000	1960.000000
50%	0.000000	35.000000	1.000000	1987.000000
75%	30.000000	230.000000	1.000000	2005.000000
max	250000.000000	30500.000000	1.000000	2013.000000

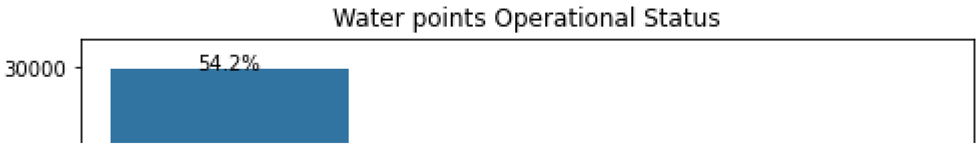
3.2 EDA

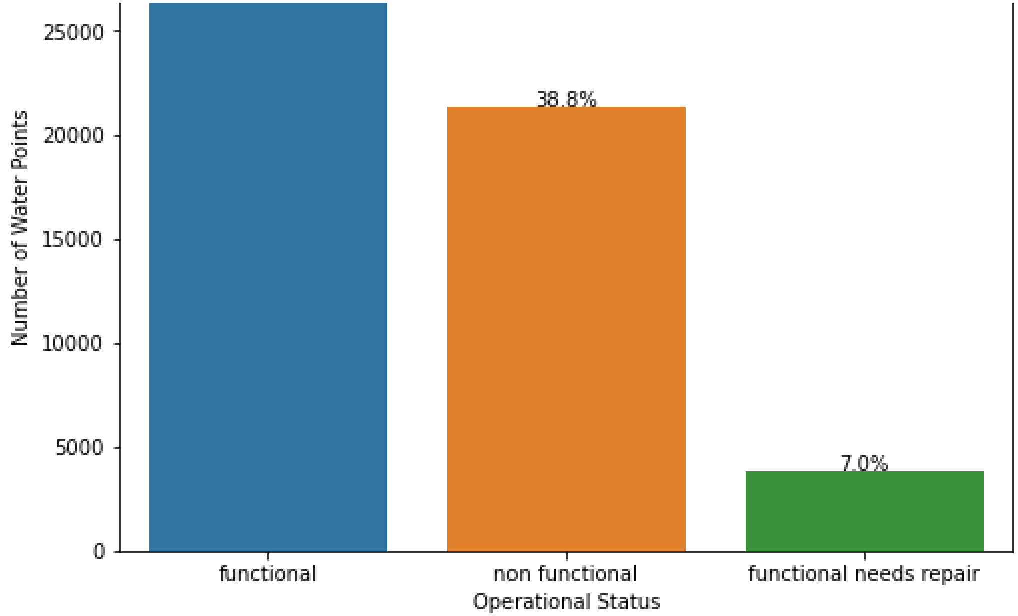
Water Points Functionality

In [23]:
```
# finding out the operational status of each water point:
well_grouping=data["status_group"].value_counts()
well_grouping
```

Out[23]:
```
functional                 29885
non functional             21381
functional needs repair     3836
Name: status_group, dtype: int64
```

In [24]:
```
#visualizing the distribution of waterpoints based on status
#creating the seaborn count plot
fig, ax = plt.subplots(figsize=(8, 6))
ax = sns.countplot(x="status_group", data=data, ax=ax)


#calculating the total
total = len(data)

#adding percentage annotation on each bar
for p in ax.patches:
    height = p.get_height()  # Get the height of each bar
    percentage = (height / total) * 100  # Calculate the percentage
    ax.text(p.get_x() + p.get_width() / 2, height + 0.1, f'{percentage:.1f}%', ha='center')


#labeling the graph
plt.ylabel("Number of Water Points")
plt.xlabel("Operational Status")
plt.title("Water points Operational Status")
plt.show()

#saving the plot as jpeg
fig.savefig("images/functionality_plot.jpeg", format="jpeg", dpi=300)
```



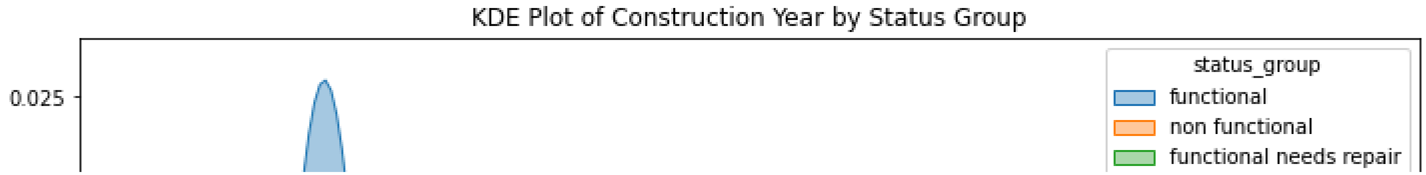


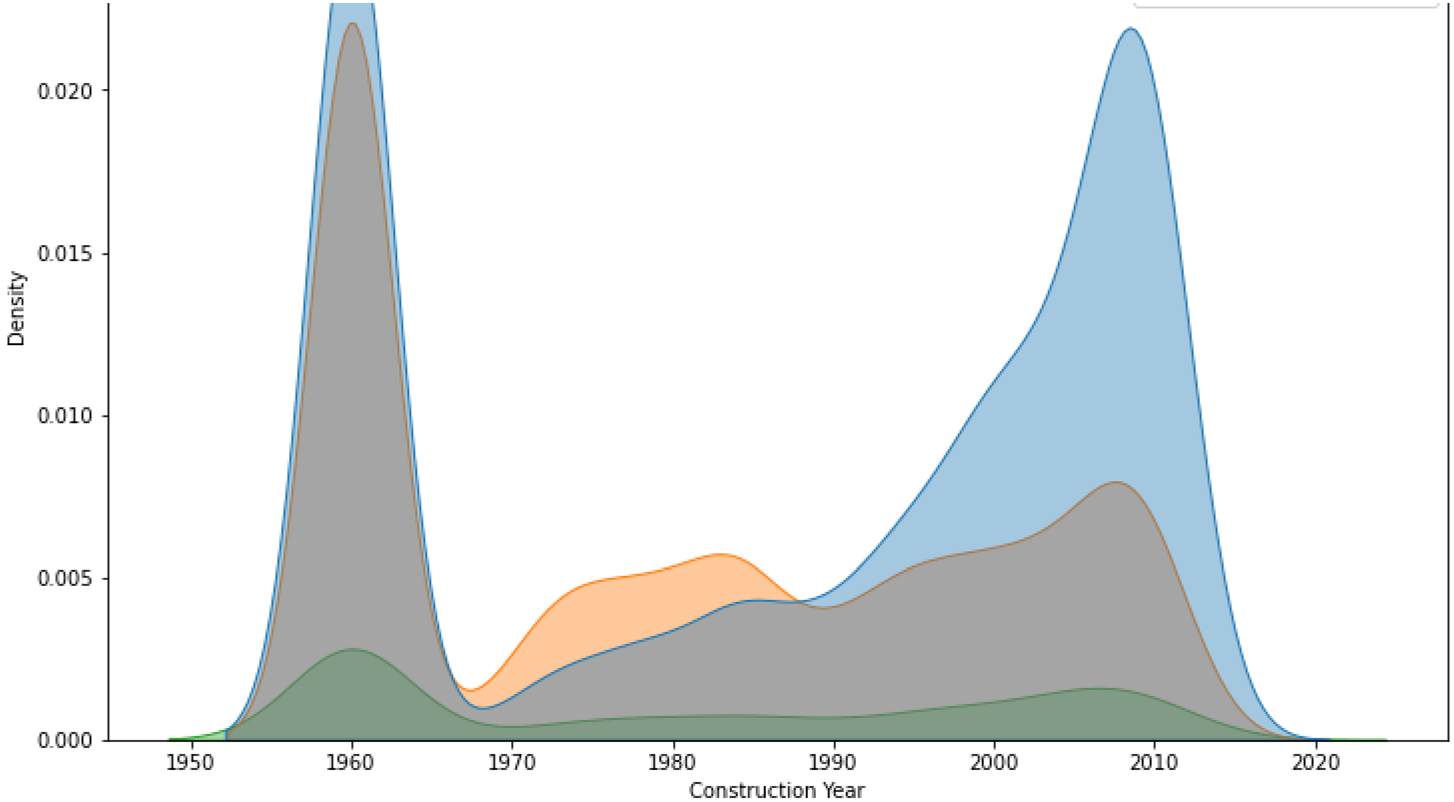
54.2% of the water points are functional, 38.8% are non functional while the remaininig 7% are functional but needs repair.

Construction Year for Water Points

In [25]:

```
#plotting the water points based on the construction year
fig, ax = plt.subplots(figsize=(12, 8))

sns.kdeplot(data=data, x="construction_year", hue="status_group", fill=True, alpha=0.4, ax=ax)

# Add labels
plt.xlabel("Construction Year")
plt.ylabel("Density")
plt.title("KDE Plot of Construction Year by Status Group")

# Show the plot
plt.show()
```





There has been a gradual increase in the number of functional water points from about 1990.

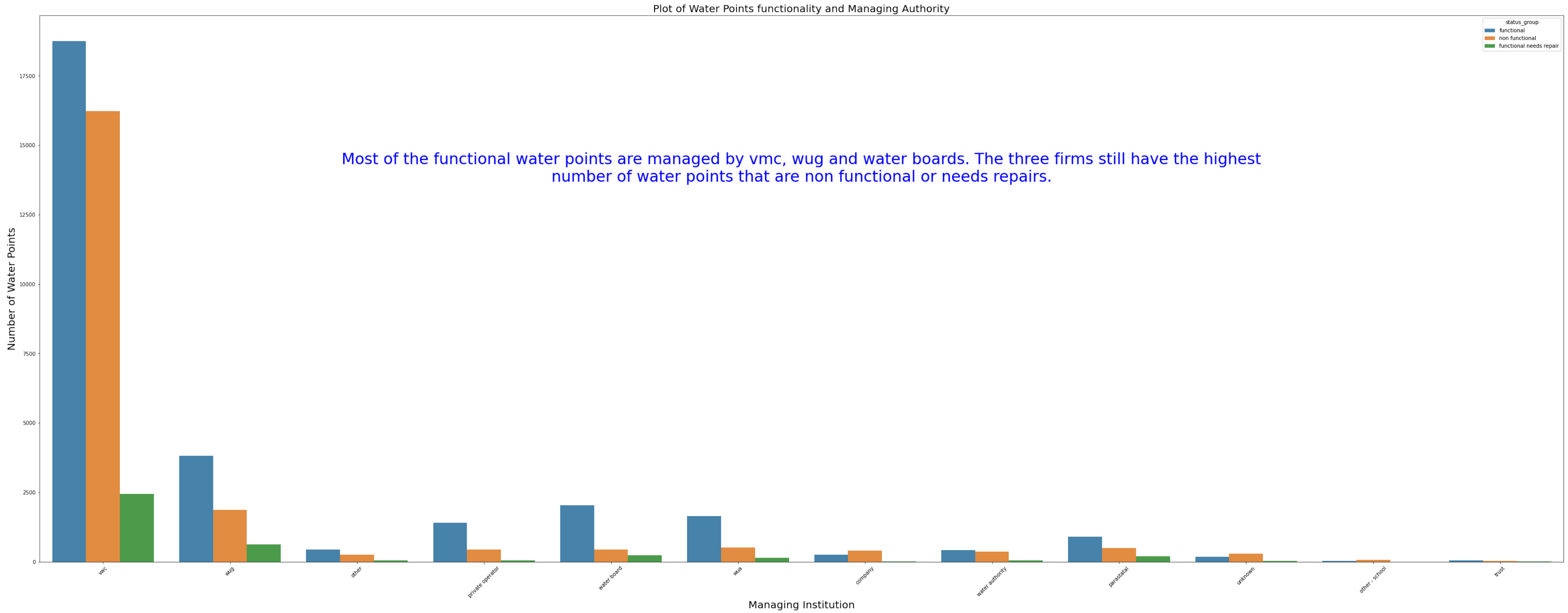
In [26]:

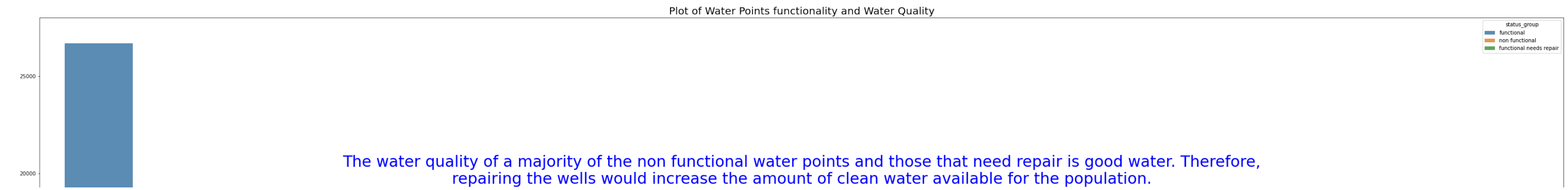
```
# create subplots
fig, axes = plt.subplots(9, figsize=(42,150),  constrained_layout=True)

# Water Points functionality and Managing Authority plot
sns.countplot(data=data, x="management", hue="status_group", alpha=0.9, ax=axes[0])

axes[0].set_xlabel("Managing Institution", fontsize=20)
axes[0].set_ylabel("Number of Water Points", fontsize=20)
axes[0].tick_params(axis="x", rotation=45, labelright=False)
axes[0].set_title("Plot of Water Points functionality and Managing Authority", fontsize=20)
text= ("Most of the functional water points are managed by vmc, wug and water boards. The three firms still have
wrapped_text = fill(text, width=120)
axes[0].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[0].transAxes, ha="center", wrap=Tr
fig.savefig("images/management_plot.jpeg", format="jpeg", dpi=300) #saving the plot
#plotting the water quality along the water points
```

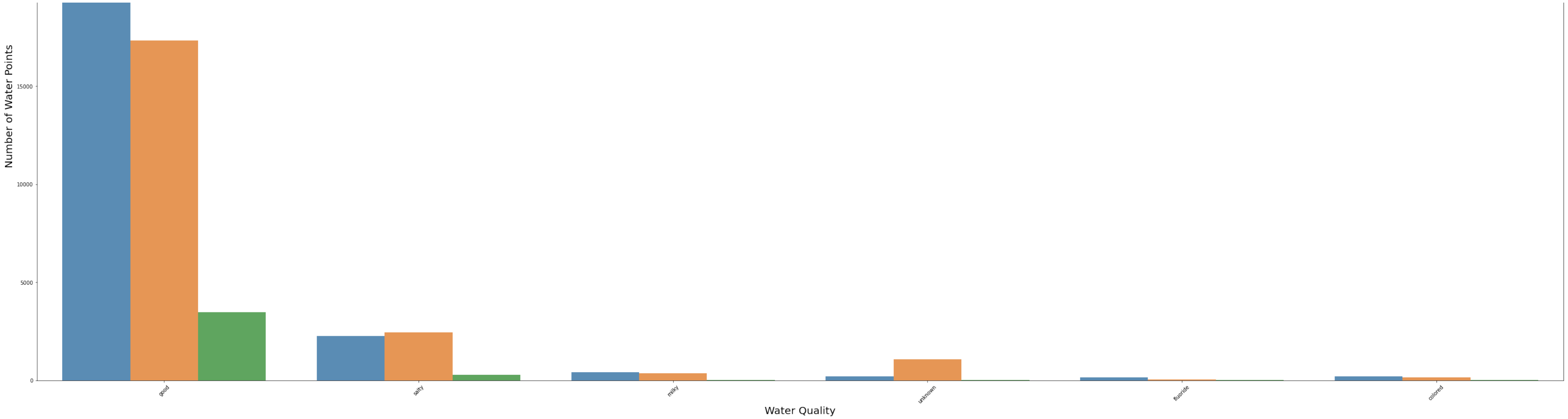
```
sns.countplot(data=data, x="quality_group", hue="status_group", alpha=0.8, ax=axes[1])

axes[1].set_xlabel("Water Quality", fontsize=20)
axes[1].set_ylabel("Number of Water Points", fontsize=20)
axes[1].tick_params(axis="x", rotation=45, labelright=False)
axes[1].set_title("Plot of Water Points functionality and Water Quality", fontsize=20)
text= ("The water quality of a majority of the non functional water points and those that need repair is good wat
wrapped_text = fill(text, width=120)
axes[1].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[1].transAxes, ha="center", wrap=Tr
fig.savefig("images/quality_plot.jpeg", format="jpeg", dpi=300)

#plotting the extraction type along the water points

sns.countplot(data=data, x="extraction_type_class", hue="status_group", alpha=0.8, ax=axes[2])

axes[2].set_xlabel("Extraction Type", fontsize=20)
axes[2].set_ylabel("Number of Water Points", fontsize=20)
axes[2].tick_params(axis="x", rotation=45, labelright=False)
axes[2].set_title("Plot of Water Points functionality and Extraction Type", fontsize=20)
plt.grid(True)
text= ("The highest functional water point are gravity operated, followed by hand pumps and submersible pumps. Wa
wrapped_text = fill(text, width=120)
axes[2].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[2].transAxes, ha="center", wrap=Tr
fig.savefig("images/extractiontype_plot.jpeg", format="jpeg", dpi=300)

#plotting the source of water along the water points

sns.countplot(data=data, x="source_class", hue="status_group",  alpha=0.9, ax=axes[3])

axes[3].set_xlabel("Water Source", fontsize=20)
axes[3].set_ylabel("Number of Water Points", fontsize=20)
axes[3].tick_params(axis="x", rotation=45, labelright=False)
axes[3].set_title("Plot of Water Points functionality and Water Source", fontsize=20)
text= ("A majority of the water points in Tanzania are ground water followed by surface run offs.")
wrapped_text = fill(text, width=120)
axes[3].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[3].transAxes, ha="center", wrap=Tr

#plotting the quantity of water along the water points

sns.countplot(data=data, x="quantity_group", hue="status_group", alpha=0.9, ax=axes[4])

axes[4].set_xlabel("Water Quantity", fontsize=20)
axes[4].set_ylabel("Number of Water Points", fontsize=20)
axes[4].tick_params(axis="x", rotation=45, labelright=False)
```

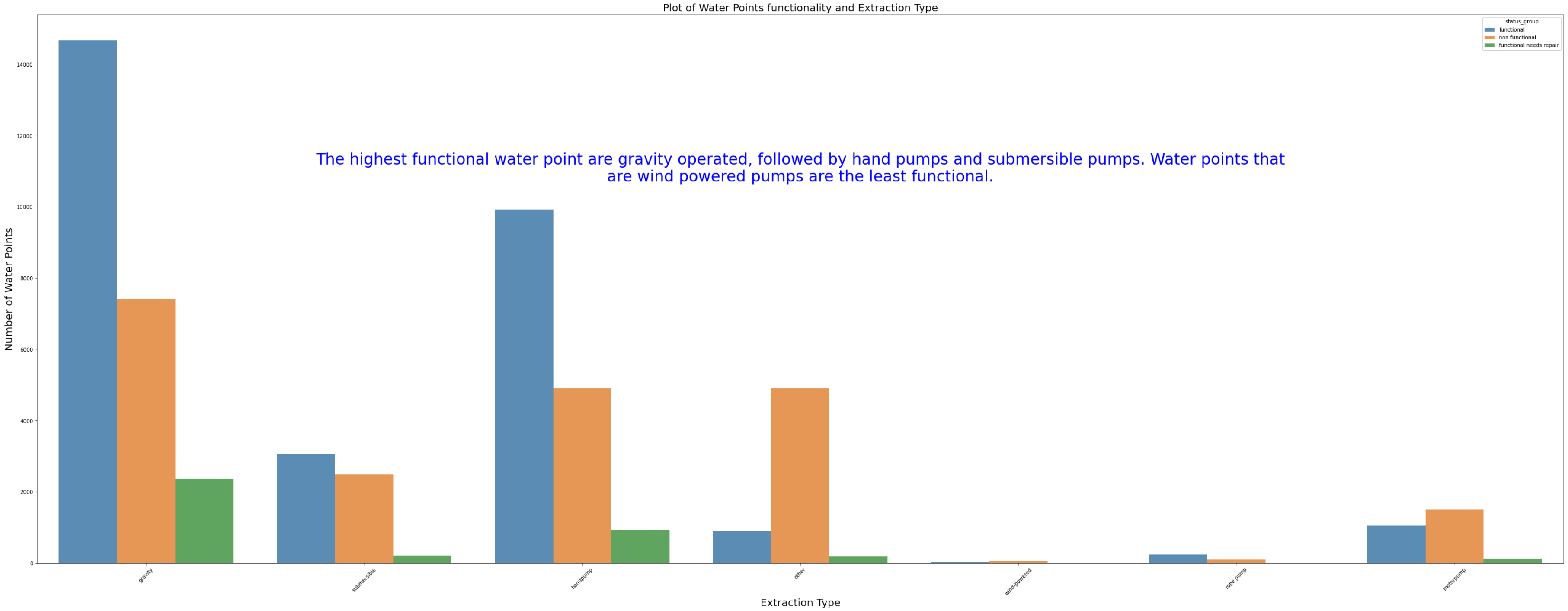
```
axes[4].tick_params(axis="x", rotation=45, labelright=False)
axes[4].set_title("Plot of Water Points functionality and Quantity", fontsize=20)
text= ("The dry water points form a considerable part of the nonfunctional water points. Similarly, most of the f
wrapped_text = fill(text, width=120)
axes[4].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[4].transAxes, ha="center", wrap=Tr
fig.savefig("images/quantity_plot.jpeg", format="jpeg", dpi=300)

#plotting the payment type along the water points

sns.countplot(data=data, x="payment_type", hue="status_group", alpha=0.9, ax=axes[5])

axes[5].set_xlabel("Payment type", fontsize=20)
axes[5].set_ylabel("Number of Water Points", fontsize=20)
axes[5].tick_params(axis="x", rotation=45, labelright=False)
axes[5].set_title("Plot of Water Points functionality and Payment type", fontsize=20)
plt.grid(True)
text= ("Most of the non functional water points are not paid for. It is also interesting to note that most of the
wrapped_text = fill(text, width=120)
axes[5].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[5].transAxes, ha="center", wrap=Tr

#plotting the Water Basins along the water points

sns.countplot(data=data, x="basin", hue="status_group", alpha=0.9, ax=axes[6])

axes[6].set_xlabel("Water Basins", fontsize=20)
axes[6].set_ylabel("Number of Water Points", fontsize=20)
axes[6].tick_params(axis="x", rotation=45, labelright=False)
axes[6].set_title("Plot of Water Basin and Water Points functionality", fontsize=20)
text= ("There are a high number of functional water points in the following basins: Rufiji, Wami/Ruvu, Lake Tanga
wrapped_text = fill(text, width=120)
axes[6].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[6].transAxes, ha="center", wrap=Tr

#plotting the permit status along the water points

sns.countplot(data=data, hue="permit", x="status_group", alpha=0.9, ax=axes[7])

axes[7].set_ylabel("Permits", fontsize=20)
axes[7].set_xlabel("Number of Water Points", fontsize=20)
axes[7].tick_params(axis="x", rotation=45, labelright=False)
axes[7].set_title("Plot of Permit and Water Points functionality", fontsize=20)
plt.grid(True)
fig.savefig("images/permits_plot.jpeg", format="jpeg", dpi=300)


#plotting the Geographical locations along the water points
```

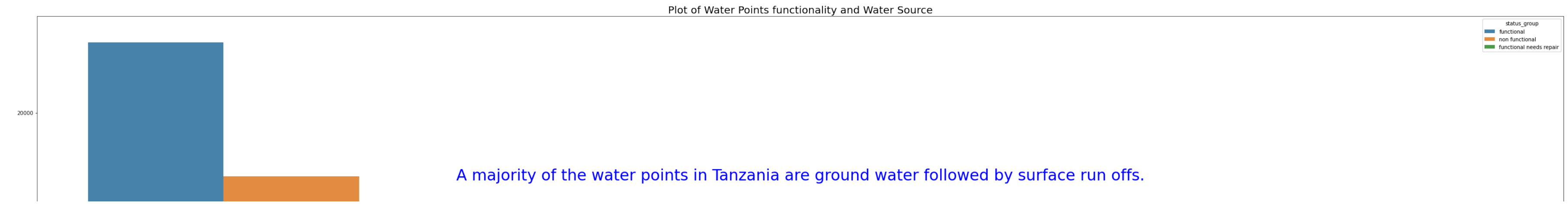
```
#plotting the Geographical locations along the water points

sns.countplot(data=data, x="region", hue="status_group", alpha=0.9, ax=axes[8])

axes[8].set_xlabel("Geographical Location", fontsize=20)
axes[8].set_ylabel("Number of Water Points", fontsize=20)
axes[8].tick_params(axis="x", rotation=45, labelright=False)
axes[8].set_title("Plot of Water Points functionality and Regions", fontsize=20)
text= ("There are a high number of functional water points compared to non-functional ones in the following regic
wrapped_text = fill(text, width=120)
axes[8].text(0.5, 0.7, wrapped_text, fontsize=30, color="blue", transform=axes[8].transAxes, ha="center", wrap=Tr


# show plots
plt.show()
```

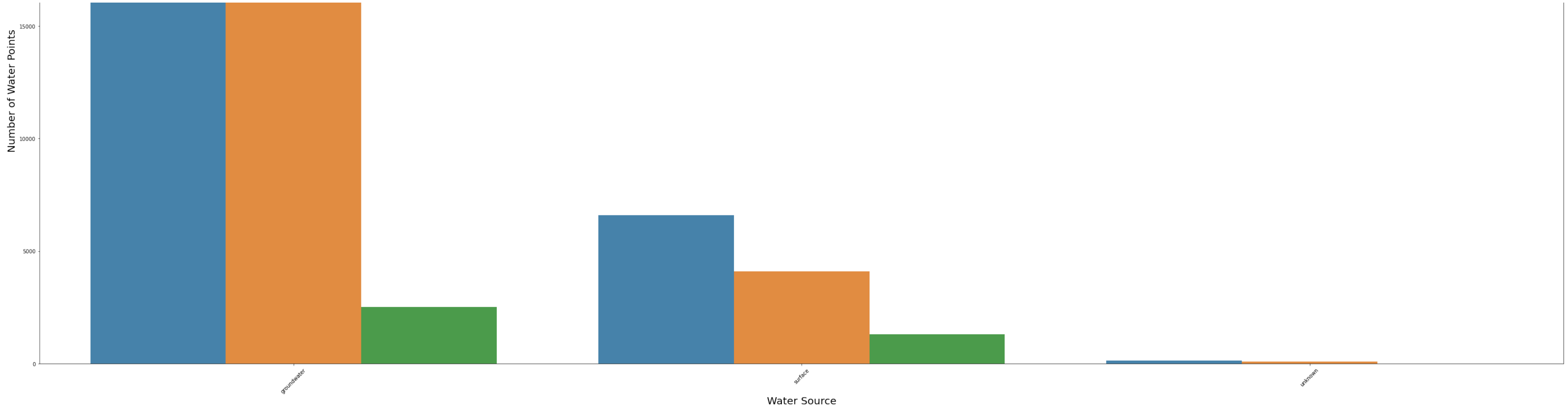


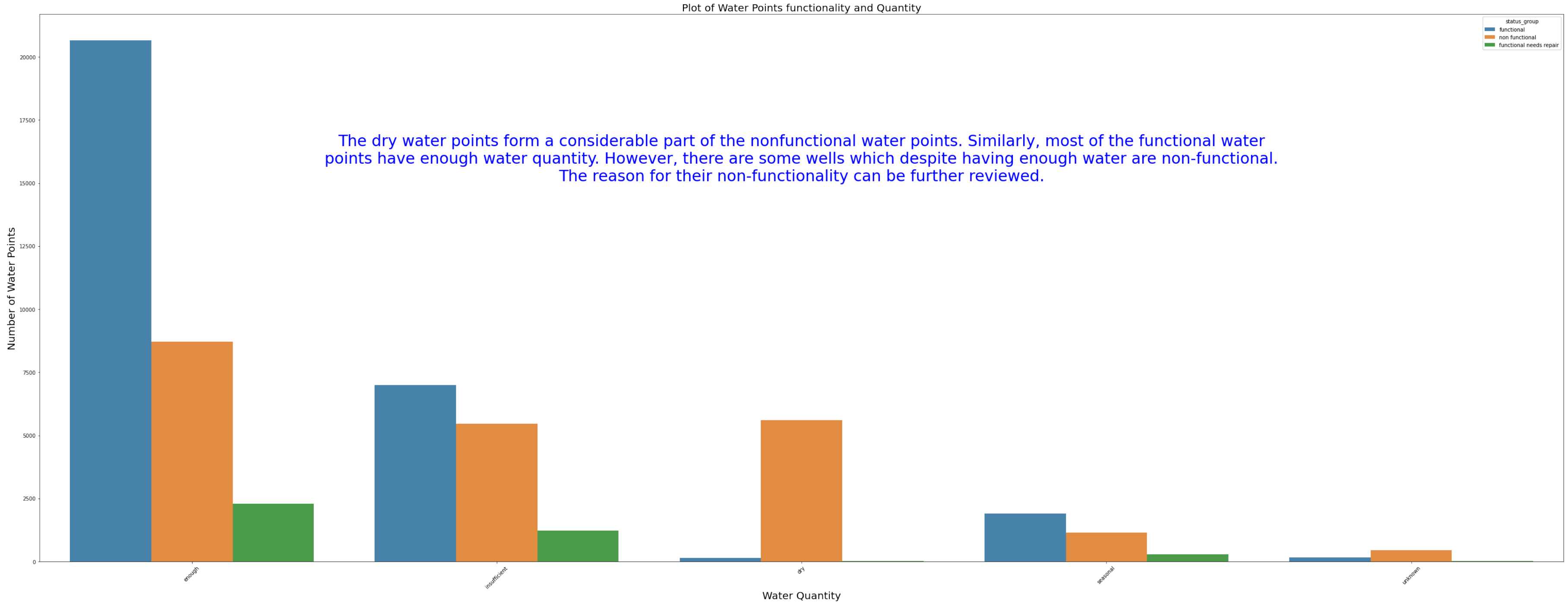


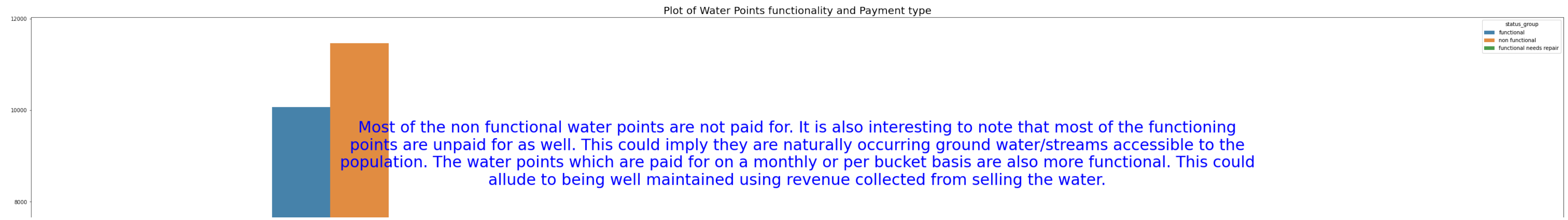

Number of Water Points
15000
10000
5000
0
good
salty
milky
unknown
fluoride
colored
Water Quality

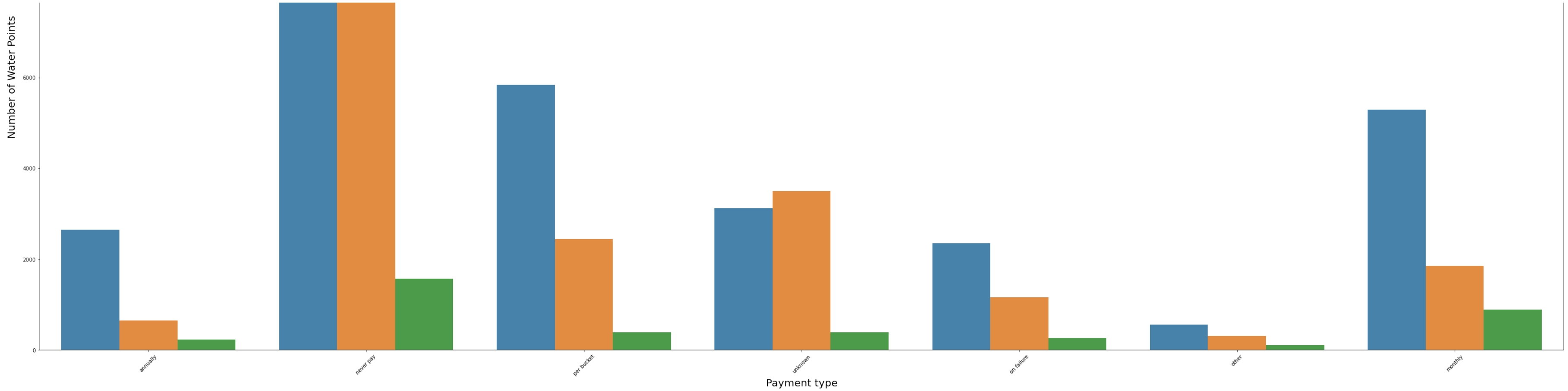

Plot of Water Points functionality and Extraction Type
status_group
functional
non functional
functional needs repair
Number of Water Points
14000
12000
10000
8000
6000
4000
2000
0
The highest functional water point are gravity operated, followed by hand pumps and submersible pumps. Water points that are wind powered pumps are the least functional.
gravity
submersible
handpump
other
wind-powered
rope pump
motorpump
Extraction Type

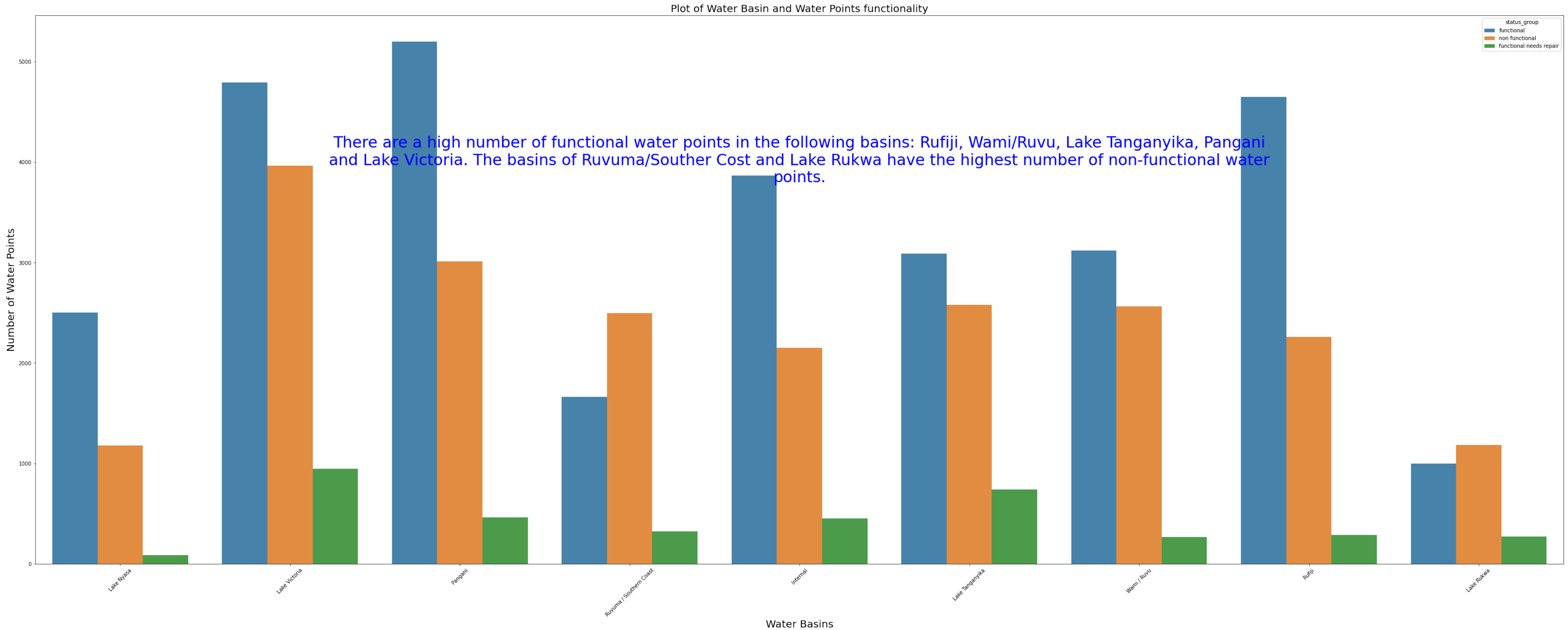

Plot of Water Points functionality and Water Source
status_group
functional
non functional
functional needs repair
20000
A majority of the water points in Tanzania are ground water followed by surface run offs.

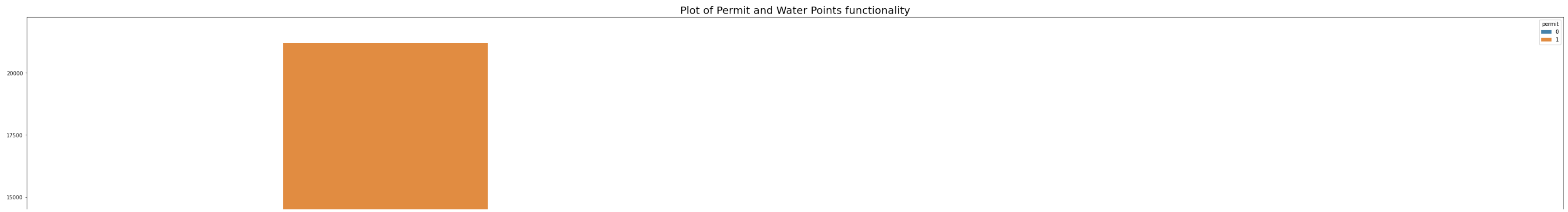

Number of Water Points
groundwater
surface
unknown
Water Source

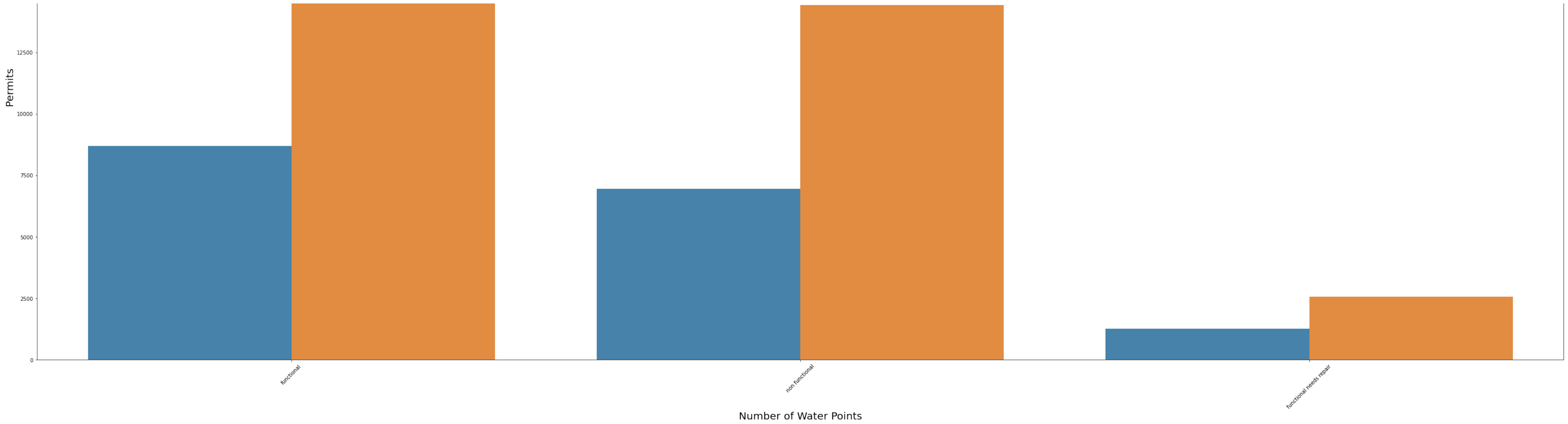

Plot of Water Points functionality and Quantity
status_group
functional
non functional
functional needs repair
The dry water points form a considerable part of the nonfunctional water points. Similarly, most of the functional water points have enough water quantity. However, there are some wells which despite having enough water are non-functional. The reason for their non-functionality can be further reviewed.
Number of Water Points
enough
insufficient
dry
seasonal
unknown
Water Quantity

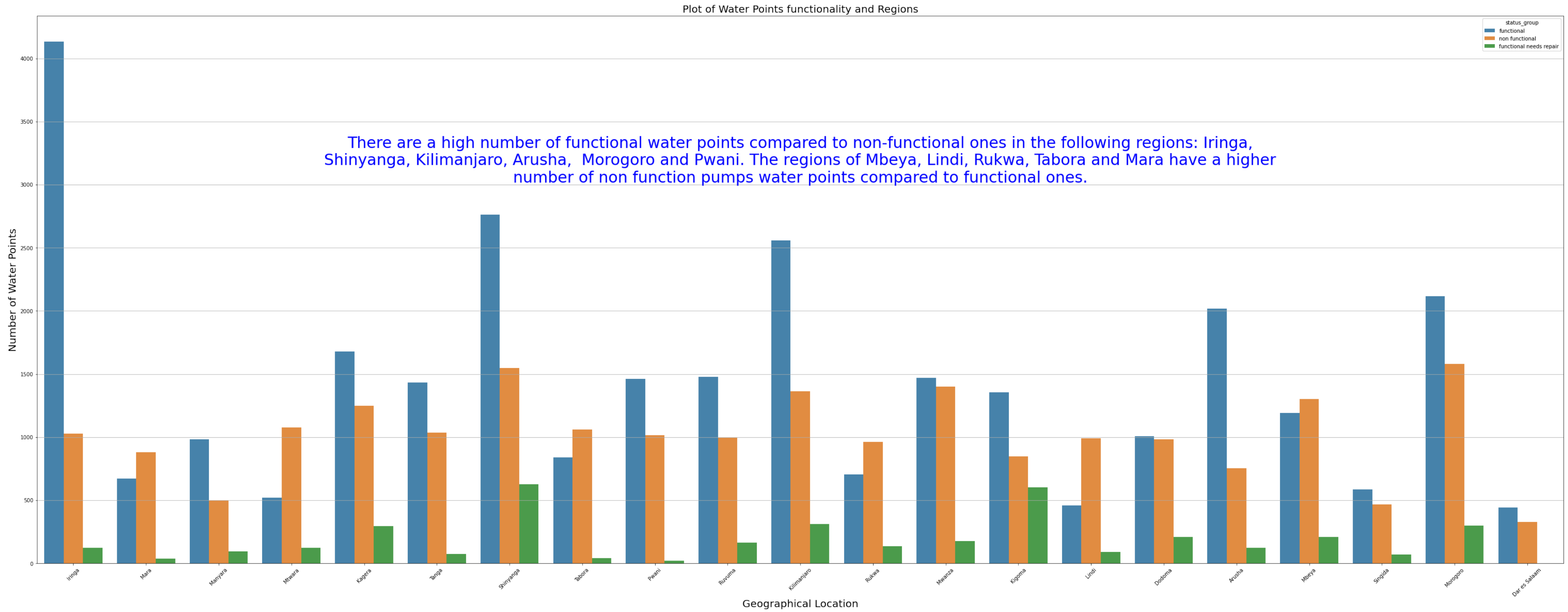

Plot of Water Points functionality and Payment type
status_group
functional
non functional
functional needs repair
Most of the non functional water points are not paid for. It is also interesting to note that most of the functioning points are unpaid for as well. This could imply they are naturally occurring ground water/streams accessible to the population. The water points which are paid for on a monthly or per bucket basis are also more functional. This could allude to being well maintained using revenue collected from selling the water.


Number of Water Points
Payment type
annually
never pay
per bucket
unknown
on failure
other
monthly


Plot of Water Basin and Water Points functionality
status_group
functional
non functional
functional needs repair
There are a high number of functional water points in the following basins: Rufiji, Wami/Ruvu, Lake Tanganyika, Pangani and Lake Victoria. The basins of Ruvuma/Souther Cost and Lake Rukwa have the highest number of non-functional water points.
Number of Water Points
Water Basins
Lake Nyasa
Lake Victoria
Pangani
Ruvuma / Southern Coast
Internal
Lake Tanganyika
Wami / Ruvu
Rufiji
Lake Rukwa


Plot of Permit and Water Points functionality
permit
0
1





4. Modelling

4.1. Data Preprocessing

Start off by creating dummy variables for categorical columns and performing train test split.

In [27]:
```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 55102 entries, 0 to 59399
Data columns (total 15 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   status_group           55102 non-null  object 
 1   amount_tsh             55102 non-null  float64
 2   installer              55102 non-null  object 
 3   basin                  55102 non-null  object 
 4   region                 55102 non-null  object 
 5   population             55102 non-null  int64  
 6   permit                 55102 non-null  int32  
 7   construction_year      55102 non-null  int64  
 8   extraction_type_class  55102 non-null  object 
 9   management             55102 non-null  object 
 10  payment_type           55102 non-null  object 
 11  quality_group          55102 non-null  object 
 12  quantity_group         55102 non-null  object 
 13  source_class           55102 non-null  object 
 14  waterpoint_type        55102 non-null  object 
dtypes: float64(1), int32(1), int64(2), object(11)
memory usage: 6.5+ MB
```

Creating Dummy Variables

In [28]:
```
# create a list of categorical and numeric columns
cat_col = ["installer","basin", "region", "extraction_type_class", "management", "payment_type", "quality_group",
num_col = ["amount_tsh", "population", "permit", "construction_year"]
```

In [29]:
```
#creating the dummies

dummy_data = pd.get_dummies(data, columns=cat_col, drop_first=True)
dummy_data.shape
```

Out[29]: (55102, 103)

Separating the Target Variable and Performing the Train Test Split

In [30]:

```
#target variable
y = dummy_data["status_group"]

#predictor variables
X = dummy_data.drop(["status_group"], axis=1)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.25, random_state=42)
```

Model Statistics

Recall will be the main metric used to track model performance. However, accuracy recall, auc and f1 score will also be computed so as to provide more details about the model using sklearn's classification_report( ) function.

In [31]:

```
def model_performance(trained_model, X, y_pred,  y_true):
    #defining the target variable names
    target_var_names= ["non_functional","functional_need_repair", "functional"]
    #print classification report
    print(classification_report(y_true, y_pred, target_names=target_var_names))

    #plotting the confusion matrix
    return plot_confusion_matrix(trained_model, X, y_true, display_labels = target_var_names, cmap=plt.cm.Blues)
    # showing the plot
    plt.show()
```

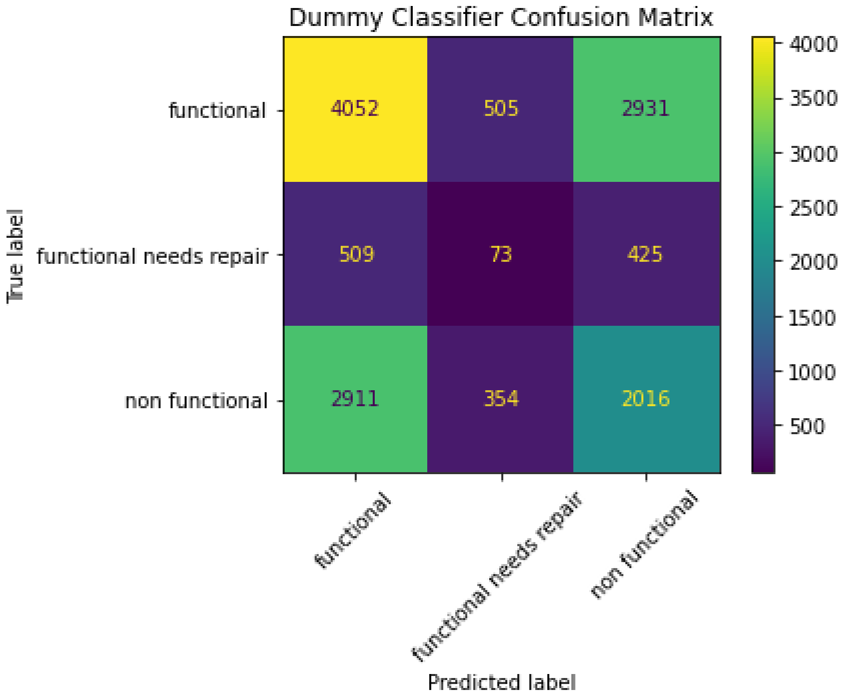
4.2. Dummy Classifier Model

In [32]:

```
# Initialize and fit the DummyClassifier with "stratified" strategy
dummy_model=DummyClassifier(random_state=42, strategy="stratified")
dummy_model.fit(X_train, y_train)

#making predictions
y_pred = dummy_model.predict(X_test)

#calculate the scores (recall and accuracy score)
recall = recall_score(y_test, y_pred, average="weighted")
print (f"Dummy Classifier Weighted Recall: {recall:.3f}")
```

```
accuracy = accuracy_score(y_test, y_pred, )
print (f"Dummy Classifier Accuracy Score: {accuracy:.3f}")

# set the plot size
plt.figure(figsize=(14,8))

# plot the confusion matrix
plot_confusion_matrix(dummy_model, X_test, y_test)

plt.xticks(rotation=45)
plt.title("Dummy Classifier Confusion Matrix")

plt.show()
```

```
Dummy Classifier Weighted Recall: 0.446
Dummy Classifier Accuracy Score: 0.446
<Figure size 1008x576 with 0 Axes>
```

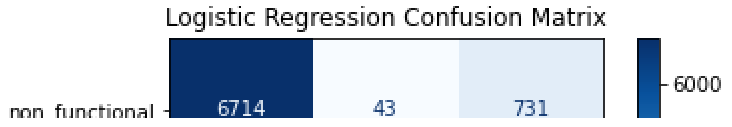


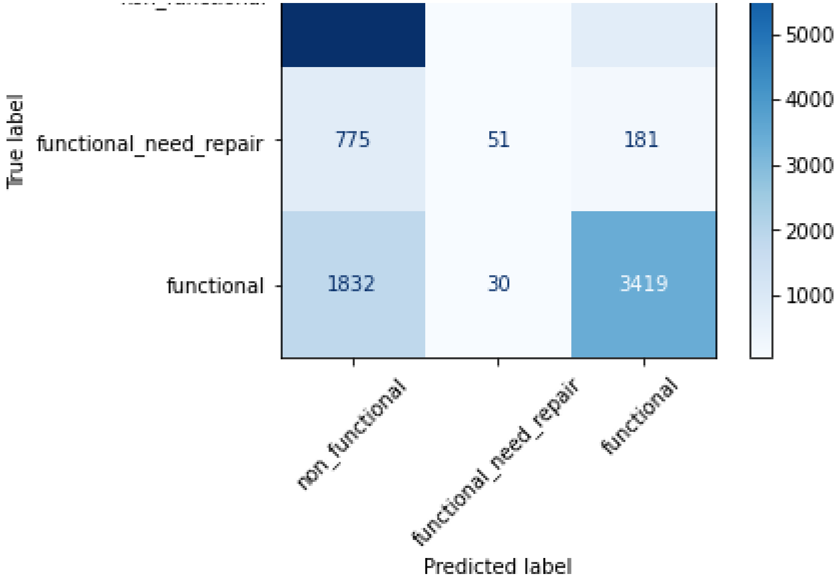
The baseline model performed poorly with a recall score and accuracy score of 44.6%. TOur data is very imbalanced which explains the base model performance of close to 50%.

4.3. Logistic Regression

In [47]:
```
#Make pipe
pipe_logreg = Pipeline([
    ('stdscaler', StandardScaler()), #standard scaler step
    ('logreg' , LogisticRegression()) # logistic regression step
])
#Fit the pipelne on the trainiing data:
pipe_logreg.fit( X_train, y_train)

#making predictions
test_pred_logreg = pipe_logreg.predict(X_test)

# set the plot size
plt.figure(figsize=(24, 46))

# evaluating the model
print("Logistic Regression Test data model score:")
logreg_score = model_performance(pipe_logreg, X_test, test_pred_logreg, y_test)

plt.xticks(rotation=45)
plt.title("Logistic Regression Confusion Matrix")

plt.show()
```

```
Logistic Regression Test data model score:
                         precision    recall  f1-score   support

         non_functional       0.72      0.90      0.80      7488
functional_need_repair       0.41      0.05      0.09      1007
             functional       0.79      0.65      0.71      5281

               accuracy                           0.74     13776
              macro avg       0.64      0.53      0.53     13776
           weighted avg       0.72      0.74      0.71     13776

<Figure size 1728x3312 with 0 Axes>
```





The logistic regression model improved over the dummy model with an accuracy score of 74% compared to 44.6%. The model struggled to predict the functional but need repairs water points with a precision score of 41%. This could likely be due class imbalances originating from the available dataset. The functional class had the highest precision at 79% while the non function class had the highest recall and f1_score of 90% and 80% respectively.

4.4. Decision Tree Model

In [ ]:

```
# Initialize the decision tree classifier
dec_tree = DecisionTreeClassifier(random_state=42)

# hyperparameter grid to tune
dec_tree_grid = {
    "criterion": ["entropy", "gini"],
    "max_depth": [5, 15, 30, 45, 60, None],
    "min_samples_split": [1, 2, 3, 5, 15, 25, 38, 45],
    "min_impurity_decrease": [0.0, 0.1, 0.2, 0.3, 0.4],
    }

# performing a grid search with cross validation
dec_tree_grid_search = GridSearchCV(estimator=dec_tree, param_grid=dec_tree_grid, cv=5, n_jobs=-1)
```

```
# fit the GridSearchCV
dec_tree_grid_search.fit(X_train, y_train)

# print the best parameters found by GridSearchCV
print(f"Best parameters are: {dec_tree_grid_search.best_params_}")

# Evaluate the best model on the test data
print(f"Best estimator score: {dec_tree_grid_search.best_estimator_.score(X_test, y_test):.3f}")
```

```
Best parameters are: {'criterion': 'gini', 'max_depth': 30, 'min_impurity_decrease': 0.0, 'min_samples_split': 38}
Best estimator score: 0.757
```

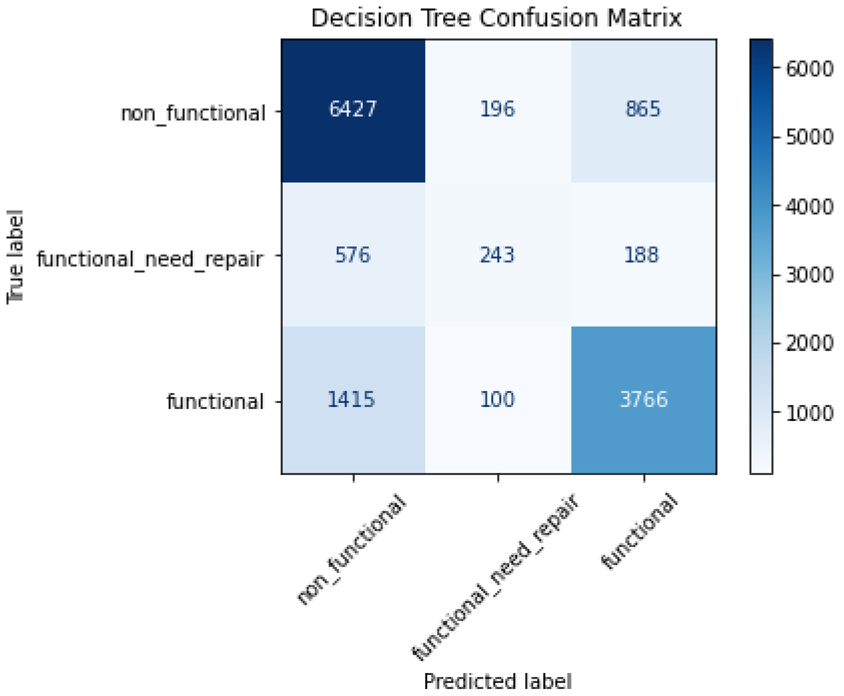
In [ ]:

```
# making the pipeline
pipe_dectree = Pipeline([
    ('stdscaler', StandardScaler()), #standard scaler step
    ('dec_tree' , DecisionTreeClassifier(
        criterion="gini", max_depth=30, min_impurity_decrease=0.0, min_samples_split=38)
     ) # decision tree step
])
#Fit the pipeline on the training data:
pipe_dectree.fit( X_train, y_train)

#making predictions
test_pred_dectree = pipe_dectree.predict(X_test)

# set the plot size
plt.figure(figsize=(24,24))

# evaluating the model
print("Decision Tree Test data model score:")
dectree_score = model_performance(pipe_dectree, X_test, test_pred_dectree, y_test)

plt.xticks(rotation=45)
plt.title("Decision Tree Confusion Matrix")

plt.show()
```

```
Decision Tree Test data model score:
                        precision    recall  f1-score   support

        non_functional       0.76      0.86      0.81      7488
functional_need_repair       0.45      0.24      0.31      1007
            functional       0.78      0.71      0.75      5281
```

```
    accuracy                           0.76     13776
   macro avg       0.67      0.60      0.62     13776
weighted avg       0.75      0.76      0.75     13776
```

```
<Figure size 1728x1728 with 0 Axes>
```



The decision tree model accuracy score improved to 76% compared to the dummy model accuracy score of 44.6% and the logistic regression accuracy of 74%. Like the logistic regression model, the model struggled to predict the functional but need repairs water points with a precision score of 45% and recall of 24%. This could likely be due class imbalances originating from the available dataset. The functional class had the highest precision at 78%.

In [54]:

```
#Predict on training and test set using the decision tree classifier
dectree_train_preds= pipe_dectree.predict(X_train)
dectree_test_preds = pipe_dectree.predict(X_test)

#Accuracy of training and test set
train_accuracy =accuracy_score(y_train, dectree_train_preds)
test_accuracy = accuracy_score(y_test, dectree_test_preds)

print(f'Training Accuracy:{train_accuracy:.3f}')
```

```
print(f'Validation Accuracy {test_accuracy:.3f}')
```

```
Training Accuracy:0.820
Validation Accuracy 0.758
```

The decision tree model is highly overfitting evident from the high training accuracy of 82% and relatively lower test accuracy of 75.8%. This means the model has learned the training data too well, including its noise and details, but is not generalizing well to unseen data.

4.5 Random Forest

In [ ]:

```
# Initialize the RandomForest classifier
rforest = RandomForestClassifier(random_state=42, n_estimators=100, max_depth=10)
rforest.fit(X_train,y_train)

#Evaluate on folds using cross validation
rforest_fold_score= cross_val_score(estimator=rforest, X=X_train, y=y_train, cv=5)
print(f"RandomForest Average Cross-Validation fold score: {np.mean(rforest_fold_score):.3f}")

#Evaluate on test set
rforest_test_score = rforest.score(X_test, y_test)
print(f"RandomForest test set score: {rforest_test_score:.3f}")


# hyperparameter grid for Random Forest
rforest_grid = {
    "n_estimators": [50, 100, 200],
    "criterion": ["entropy", "gini"],
    "max_depth": [15, 30, None],
    "min_impurity_decrease": [0.0, 0.01, 0.1],
    "max_features": ["sqrt", "log2"],
    }

# performing a grid search with cross validation
rforest_grid_search = GridSearchCV(
    estimator=RandomForestClassifier(random_state=42),
    param_grid=rforest_grid,
    cv=5,
    n_jobs=-1
    )
```

```python
# fit the GridSearchCV
rforest_grid_search.fit(X_train, y_train)

# print the best parameters found by GridSearchCV
print(f"Random Forest Best parameters: {rforest_grid_search.best_params_}")

# Evaluate the best model on the test data
print(f"Optimized Random Forest Test Set Score: {rforest_grid_search.best_estimator_.score(X_test, y_test):.3f}")
```

```
RandomForest Average Cross-Validation fold score: 0.743
RandomForest test set score: 0.743
Random Forest Best parameters: {'criterion': 'gini', 'max_depth': 30, 'max_features': 'log2', 'min_impurity_decrea
se': 0.0, 'n_estimators': 200}
Optimized Random Forest Test Set Score: 0.778
```

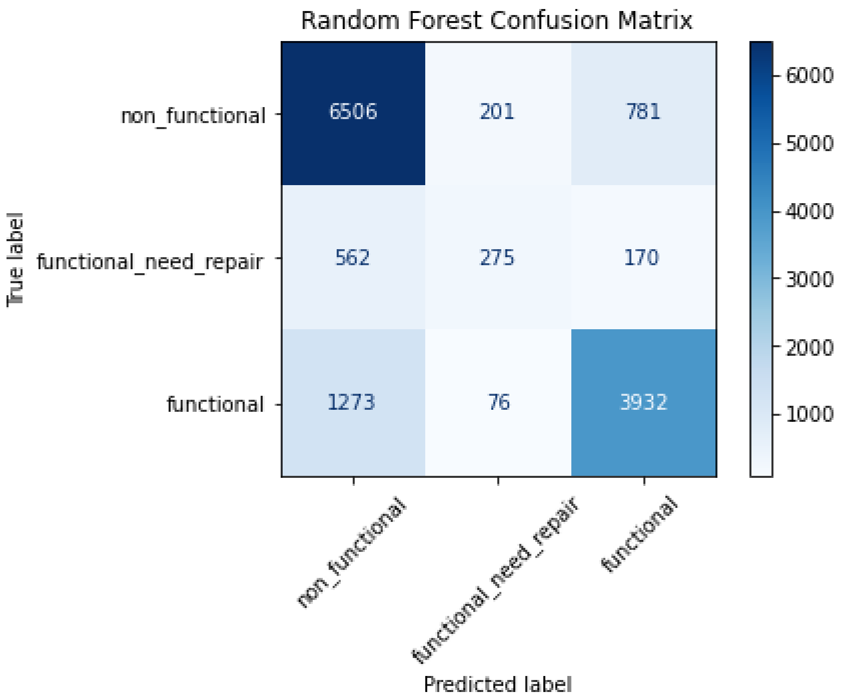
In [78]:

```python
# making the pipeline
pipe_rforest = Pipeline([
    ('stdscaler', StandardScaler()), #standard scaler step
    ('rforest' , RandomForestClassifier(
        bootstrap = True,
        criterion="gini",
        max_depth=30,
        min_impurity_decrease=0.0,
        max_features = "log2",
        n_estimators=200)
     ) # random forest step
])
#Fit the pipeline on the training data:
pipe_rforest.fit( X_train, y_train)

#making predictions
test_pred_rforest = pipe_rforest.predict(X_test)

# set the plot size
plt.figure(figsize=(24,24))

# evaluating the model
print("Random Forest Test data model score:")
rforest_score = model_performance(pipe_rforest, X_test, test_pred_rforest, y_test)

plt.xticks(rotation=45)
plt.title("Random Forest Confusion Matrix")
```

```
plt.show()
```

```
Random Forest Test data model score:
                      precision    recall  f1-score   support

      non_functional       0.78      0.87      0.82      7488
functional_need_repair       0.50      0.27      0.35      1007
          functional       0.81      0.74      0.77      5281

            accuracy                           0.78     13776
           macro avg       0.69      0.63      0.65     13776
        weighted avg       0.77      0.78      0.77     13776
```

<Figure size 1728x1728 with 0 Axes>



The random forest model accuracy score improved to 78% compared to the dummy model accuracy score of 44.6%, the decision tree model of 76% and the logistic regression accuracy of 74%. Prediction of the functional but need repairs water points improved slightly with a precision score of 50% and recall of 27%. The functional class had the highest precision at 81%.

In [79]:

```
#Predict on training and test set using the decision tree classifier
rforest_train_preds= pipe_rforest.predict(X_train)
```

```
rforest_train_preds= pipe_rforest.predict(X_train)
rforest_test_preds = pipe_rforest.predict(X_test)

#Accuracy of training and test set
train_accuracy = accuracy_score(y_train, rforest_train_preds)
test_accuracy = accuracy_score(y_test, rforest_test_preds)

print(f'Training Accuracy:{train_accuracy:.3f}')
print(f'Validation Accuracy {test_accuracy:.3f}')
```

```
Training Accuracy:0.916
Validation Accuracy 0.778
```

Running the GridSearch with the RandomForestPipeline, our baseline accuracy was once again improved to 81% precision for the functional class over the Decision Tree model at 78%. The model is still over fitting the training data, as the training accuracy is 91.6% and the validation accuracy is 77.8%. However, this is our best performing model so far.

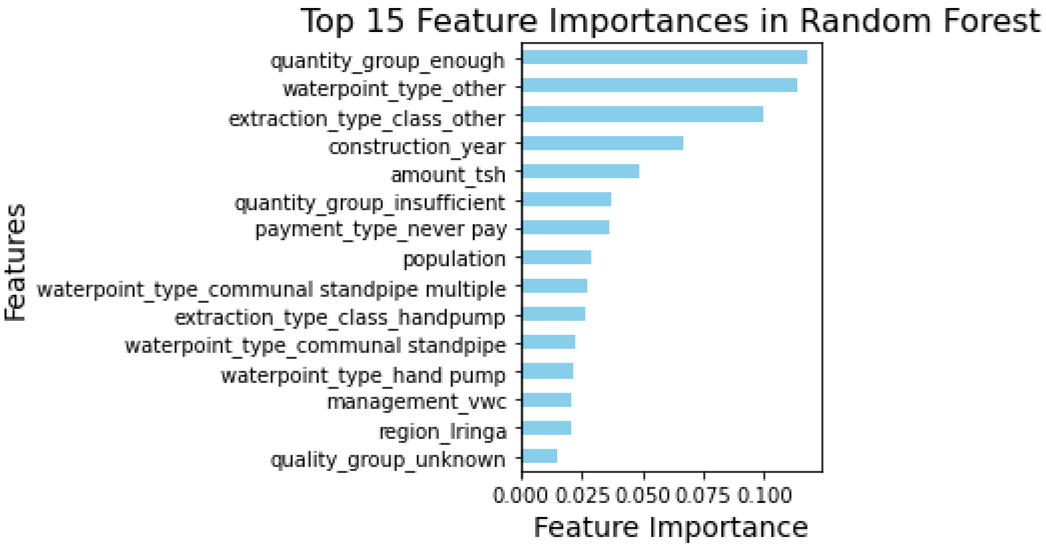
4.5.1 Random Forest Feature Importance

In [99]:

```
# Extracting  feature importances from the Random Forest Model
feature_importances=pd.DataFrame({
    "feature": X_train.columns,
    "importance":rforest.feature_importances_
    })
# sort feature importances in descending order
feature_importances=feature_importances.sort_values(by="importance", ascending=False)

# Plot the top 15 features
plt.figure(figsize=(24,8))
feature_importances.head(15).plot(
    kind='barh',
    x='feature',
    y='importance',
    legend=False,
    color='skyblue'
)

# Add labels and title
plt.xlabel('Feature Importance', fontsize=14)
plt.ylabel('Features', fontsize=14)
plt.title('Top 15 Feature Importances in Random Forest', fontsize=16)
plt.gca().invert_yaxis()  # Invert y-axis to show the highest importance on top
```

```
plt.tight_layout()
plt.show()
```

<Figure size 1728x576 with 0 Axes>



The random forest model shows quantity_group_enough, waterpoint_type_other, extraction_type_class_other, construction_year and amount_tsh as being hte most important features to the model.

5. Conclusion and Recommendation

Random Forests was the best performing model with Decision Tree being the second best model. The poor performance of the Logistic Regression models indicate that the data is not easily separable. The Random forest model performs with an 78% testing accuracy and precision for the functional class at 81%. It also had the highest f1 score of any model at 82%.

The main source of water for Tanzania is ground water. There are a high number of functional water points Iringa, Shinyanga, Kilimanjaro, Arusha, Morogoro and Pwani regions. The regions of Mbeya, Lindi, Rukwa, Tabora and Mara have a higher number of non function pumps water points. Therefore, more resources should be allocated to these areas as the situation is critical. There is a cluster of functional but need repair water points in Lake Victoria, Southern Coast, Lake Rukwa, Pnangani and Lake Tanganyika basins. These should be addressed to prevent failure which can be more expensive to repair.

The Random Forest model showed that the most important features are quantity of water (enough), water point type and extraction type for the waterpoint. There are over 8,000 waterpoints that have enough water in them but are non functional. These are recommended as high priority class to address. Wells with no fees are more likely to be non functional. Payment provides incentive and means to keep wells functional. Water points managed by VMC, WUG and Water Board have a lower rate of pump failure. The three organizations can be used for case studies on good water point management practices. Investigate why these installers have