Classification Models to Identify Water Pump Functionality in Tanzania

Overview

This project consists of classification models to identify water pump functionality in Tanzania. Limited access to safe water is a major health risk. These models are potential tools for government agencies or non-governmental organizations to identify areas with limited access to clean water. Through an iterative modeling process, we produced a model that predicted pump functionality with 85% accuracy. By identifying non-functional water pumps, organizations can divert resources to areas in need of assistance and improve water access and health in Tanzania.

Business Understanding

Millions of people in Tanzania lack access to safe water. This results in paying high prices for water from vendors or collecting water from unsafe natural sources. In order to combat this problem, resources must be allocated to fix non-functioning water distribution points. Age is an important metric in predicting the condition of distribution points. Older pumps and engine systems are more likely to fail than newer ones. In many cases age related data is not available. The goal of this analysis is to build the model that can predict the condition of waterpoints based on their other features such as regional factors, installer, type of pump, population and others

Data Understanding

The data was sourced from the Taarifa waterpoint dashboard, which aggregates data from the Tanzania Ministry of Water. The information collected was recorded by GeoData Consultants Ltd. There are 59,400 rows and 40 columns in the "water_well_train_data.csv".

Our target data is stored in "water_well_train_labels.csv". There are 59,400 rows and 2 columns in this csv file. The two columns in this csv file are 'id' and 'status_group'. The 'id' column aligns with the 'id' column in the "water_well_train_data.csv" file. Our target column is 'status_group' which consists of three values describing the status of a water pump: "functional", "functional needs repair", and "non-functional".

Data Cleaning and EDA

In [1]:

```
#import statements
import pandas as pd
import numpy as np

#data visualization
import matplotlib.pyplot as ply
import seaborn as sns

#sci-kit learn
import sklearn
from sklearn.model_selection import train_test_split, cross_validate, cross_val_score, Gr
idSearchCV
from sklearn.preprocessing import FunctionTransformer, OneHotEncoder, StandardScaler
from sklearn.metrics import accuracy_score, plot_confusion_matrix
from sklearn.dummy import DummyClassifier

from sklearn.neighbors import KNeighborsClassifier, NearestNeighbors
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

from sklearn.pipeline import Pipeline
from sklearn.compose import ColumnTransformer
```

In [2]:

```
#import train data
#DO NOT LOOK AT TEST DATA UNTIL VALIDATION
df_train = pd.read_csv('./data/water_well_train_data.csv')
```

In [3]:

```
df_train.head()
```

Out[3]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	pay
0	69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	I
3	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	

5 rows × 40 columns

In [4]:

```
# info of train data
df_train.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 40 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   id                     59400 non-null  int64  
 1   amount_tsh             59400 non-null  float64
 2   date_recorded          59400 non-null  object 
 3   funder                 55765 non-null  object 
 4   gps_height             59400 non-null  int64  
 5   installer              55745 non-null  object 
 6   longitude              59400 non-null  float64
 7   latitude               59400 non-null  float64
 8   wpt_name               59400 non-null  object 
 9   num_private            59400 non-null  int64  
 10  basin                  59400 non-null  object 
 11  subvillage             59029 non-null  object 
 12  region                 59400 non-null  object 
 13  region_code            59400 non-null  int64  
 14  district_code          59400 non-null  int64  
 15  lga                    59400 non-null  object 
 16  ward                   59400 non-null  object 
 17  population             59400 non-null  int64  
 18  public_meeting         56066 non-null  object 
 19  recorded_by            59400 non-null  object 
 20  scheme_management      55523 non-null  object 
 21  scheme_name            31234 non-null  object 
 22  permit                 56344 non-null  object 
 23  construction_year      59400 non-null  int64  
 24   extraction_type        59400 non-null  object
 25   extraction_type_group  59400 non-null  object
 26   extraction_type_class  59400 non-null  object
 27   management             59400 non-null  object
 28   management_group       59400 non-null  object
 29   payment                59400 non-null  object
 30   payment_type           59400 non-null  object
 31   water_quality          59400 non-null  object
 32   quality_group          59400 non-null  object
 33   quantity               59400 non-null  object
 34   quantity_group         59400 non-null  object
 35   source                 59400 non-null  object
 36   source_type            59400 non-null  object
 37   source_class           59400 non-null  object
 38   waterpoint_type        59400 non-null  object
 39   waterpoint_type_group  59400 non-null  object
dtypes: float64(3), int64(7), object(30)
memory usage: 18.1+ MB
```

The data contains 59,400 rows and 40 columns.

Our target column is stored in a separate csv file.

In [5]:

```
#import target information
df_label = pd.read_csv('./data/water_well_train_labels.csv')
```

In [6]:

```
df_label.shape
```

Out[6]:

```
(59400, 2)
```

In [7]:

```
df_label.head()
```

Out[7]:

	id	status_group
0	**69572**	**functional**
1	**8776**	**functional**
2	**34310**	**functional**
3	**67743**	**non functional**
4	**19728**	**functional**

In [8]:

```
df_label.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 2 columns):
 #   Column        Non-Null Count  Dtype 
---  ------        --------------  ----- 
 0   id            59400 non-null  int64 
 1   status_group  59400 non-null  object
dtypes: int64(1), object(1)
memory usage: 928.2+ KB
```

In [9]:

```
df_label['status_group'].value_counts()
```

Out[9]:

```
functional                 32259
non functional             22824
functional needs repair     4317
Name: status_group, dtype: int64
```

In [10]:

```
df_label['status_group'].value_counts(normalize = True)
```

Out[10]:

```
functional                 0.543081
non functional             0.384242
functional needs repair    0.072677
Name: status_group, dtype: float64
```

There are three target classifications: functional (54%), non-functional (38%), and function needs repair (7%). We combined the 'status_group' dataframe with the train_data dataframe.

In [11]:

```
#combine train and label dataframes prior to cleaning to address any dropped rows
df = df_train.join(other = df_label, rsuffix = '_label')
```

In [12]:

```
df.head()
```

Out[12]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
0	69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	
3	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	

5 rows × 42 columns

A quick confirmation to see that the columns in the dataframe are properly aligned. The id values from each dataframe should match.

In [13]:

```
#check that id columns align
df[df['id'] != df['id_label']]
```

Out[13]:

id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	quality_group

0 rows × 42 columns

Now that our features and target were in the same dataframe, we could begin exploring and cleaning the data.

In [14]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 42 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   id                     59400 non-null  int64  
 1   amount_tsh             59400 non-null  float64
 2   date_recorded          59400 non-null  object 
 3   funder                 55765 non-null  object 
 4   gps_height             59400 non-null  int64  
 5   installer              55745 non-null  object 
 6   longitude              59400 non-null  float64
 7   latitude               59400 non-null  float64
 8   wpt_name               59400 non-null  object 
 9   num_private            59400 non-null  int64  
 10  basin                  59400 non-null  object 
 11  subvillage             59029 non-null  object 
 12  region                 59400 non-null  object 
 13  region_code            59400 non-null  int64  
 14  district_code          59400 non-null  int64  
 15  lga                    59400 non-null  object 
 16  ward                   59400 non-null  object 
 17  population             59400 non-null  int64  
 18  public_meeting         56066 non-null  object 
 19  recorded_by            59400 non-null  object 
 20  scheme_management      55523 non-null  object 
 21  scheme_name            31234 non-null  object 
 22  permit                 56344 non-null  object 
 23  construction_year      59400 non-null  int64  
 24  extraction_type        59400 non-null  object 
 25  extraction_type_group  59400 non-null  object 
 26  extraction_type_class  59400 non-null  object 
 27  management             59400 non-null  object 
 28  management_group       59400 non-null  object 
 29  payment                59400 non-null  object 
 30  payment_type           59400 non-null  object 
 31  water_quality          59400 non-null  object 
 32  quality_group          59400 non-null  object 
 33  quantity               59400 non-null  object 
 34  quantity_group         59400 non-null  object 
 35  source                 59400 non-null  object 
 36  source_type            59400 non-null  object 
 37  source_class           59400 non-null  object 
 38  waterpoint_type        59400 non-null  object 
 39  waterpoint_type_group  59400 non-null  object 
 40  id_label               59400 non-null  int64  
 41  status_group           59400 non-null  object 
dtypes: float64(3), int64(8), object(31)
memory usage: 19.0+ MB
```

We have 42 columns: the majority of them were object type with a few numerical.

Columns with nulls:

- **funder**
- **installer**
- **subvillage**
- **public_meeting**
- **scheme_management**
- **scheme_name**
- **permit**

In [15]:

```
df.describe()
```

Out[15]:

	id	amount_tsh	gps_height	longitude	latitude	num_private	region_code	district_code	
count	59400.000000	59400.000000	59400.000000	59400.000000	5.940000e+04	59400.000000	59400.000000	59400.000000	59
mean	37115.131768	317.650385	668.297239	34.077427	-5.706033e+00	0.474141	15.297003	5.629747	
std	21453.128371	2997.574558	693.116350	6.567432	2.946019e+00	12.236230	17.587406	9.633649	
min	0.000000	0.000000	-90.000000	0.000000	-1.164944e+01	0.000000	1.000000	0.000000	
25%	18519.750000	0.000000	0.000000	33.090347	-8.540621e+00	0.000000	5.000000	2.000000	
50%	37061.500000	0.000000	369.000000	34.908743	-5.021597e+00	0.000000	12.000000	3.000000	
75%	55656.500000	20.000000	1319.250000	37.178387	-3.326156e+00	0.000000	17.000000	5.000000	
max	74247.000000	350000.000000	2770.000000	40.345193	-2.000000e-08	1776.000000	99.000000	80.000000	30

Dealing with Nulls

We looked at every column with null values to determine how to address those missing values.

funder

In [16]:

```
#who funded the well
print(df['funder'].isna().sum())
df['funder'].value_counts()
```

```
3635
```

Out[16]:

```
Government Of Tanzania        9084
Danida                        3114
Hesawa                        2202
Rwssp                         1374
World Bank                    1349
                              ...
Pentekoste                       1
Usambala Sister                  1
Rumaki                           1
Friedkin Conservation Fund       1
Rashid Seng'ombe                 1
Name: funder, Length: 1897, dtype: int64
```

There were 3,635 nulls in 'funder'.

In [17]:

```
unique_funder = list(df['funder'].unique())
print(len(unique_funder))
unique_funder
```

```
1898
```

Out[17]:

```
['Roman',
 'Grumeti',
 'Lottery Club',
 'Unicef',
 'Action In A',
 'Mkinga Distric Coun',
 'Dwsp',
 'Rwssp',
 'Wateraid',
 'Isingiro Ho',
 'Private',
 'Danida',
 'World Vision',
 'Lawatefuka Water Supply',
 'Biore',
 'Rudep',
 'Hesawa',
 'Twe',
 'Isf',
 'African Development Bank',
 'Government Of Tanzania',
 'Sobodo',
 'Water',
 'Private Individual',
 'Undp',
 nan,
 'Not Known',
 'Kirde',
 'Cefa',
 'Ces(gmbh)',
 'European Union',
 'Lga',
 'District Council',
 'Muwsa',
 'Dwe/norad',
 'Kkkt_makwale',
 'Sawaka',
 'Ces (gmbh)',
 'Olgilai Village Community',
 'Kkkt',
 'Roman Catholic',
 'Norad',
 'Adra',
 'Sema',
 'Piusi',
 'Dwe',
 'Rc Church',
 'Swisland/ Mount Meru Flowers',
 'Ifad',
 'Swedish',
 'Idc',
 'He',
 'Isf/tacare',
 'Jica',
 'Mzee Sh',
 'Aict',
 'Tcrs',
 'Kiuma',
 'Germany Republi',
 'Netherlands',
 'Ruthe',
 'Tulawaka Gold Mine',
 'Nethalan',
 'Tasaf',
 'Concern World Wide',
 'Wfp',
 'Lips',
 'Sida',
 'World Bank',
 'Tanza',
 '0',
 'Sw',
 'Shipo',
 'Fini Water',
 'Kanisa'
```

```
 'Kahisa',
 'Oxfarm',
 'Village Council',
 'Hesawz',
 'Shanta',
 'Fpct',
 'Wvt',
 'Dhv',
 'Ir',
 'Oikos E.Afrika',
 'Anglican Church',
 'Peters',
 'Donor',
 'Secondary Schoo',
 'Amref',
 'Ministry Of Water',
 'Adb',
 'Jbg',
 'Dadis',
 'International Aid Services',
 'Germany',
 'Kibaha Town Council',
 'Dsdp',
 'Dfid',
 'Rural Water Supply And Sanitat',
 'Af',
 'Wananchi',
 'Fw',
 'No',
 'Dct',
 'Africare',
 'Norad /government',
 'British Colonial Government',
 'Co',
 'Ridep',
 'Tassaf',
 'Hans',
 'Socie',
 'Finw',
 'Fin Water',
 'Oxfam',
 'Plan International',
 'African Muslim Agency',
 'Go',
 'Cdtf',
 'Shawasa',
 'Un',
 'Awf',
 'Commu',
 'Community',
 'Save The Rain Usa',
 'Kibara Foundation',
 'Tlc',
 'Rc Churc',
 'Plan Int',
 'W.B',
 'Lvia',
 'Songea District Council',
 'Hifab',
 'Rc Ch',
 'Makonde Water Population',
 'Snv',
 'Government/ Community',
 'National Rural',
 'Is',
 'Giz',
 'Cspd',
 'Medicine',
 'Wsdp',
 'Unice/ Cspd',
 'Finn Water',
 'Kamama',
 'Villagers'
```

```
    'Villagers',
    'Ereto',
    'Abasia',
    'Unhcr',
    'Ebaha',
    'Kuwait',
    'Magadini-makiwaru Water',
    'Mh An',
    'Kaemp',
    'African Relie',
    'Rcchurch/cefa',
    'Norad/ Kidep',
    'Private Owned',
    'Tardo',
    'Insututional',
    'Sabemo',
    'Missi',
    'Dmdd',
    'Dhv\\norp',
    'Mission',
    'Ru',
    'Halmashauri Ya Wilaya Sikonge',
    'Japan',
    'Simone',
    'Ki',
    'Peace Cope',
    'Finland',
    'Marafip',
    'Ta',
    'Su-ki Jang',
    'Tado',
    'Tanzania',
    'Il',
    'Bank',
    'Ded',
    'Irc',
    'Sabodo',
    'Soda',
    'I.E.C',
    'Drdp Ngo',
    'Lwi',
    'Maxavella',
    'Ics',
    'African',
    'Bilila',
    'Tpp',
    'Cipro/government',
    'Tabora Municipal Council',
    'Salim Ahmed Salim',
    'Eu/acra',
    'Kadres Ngo',
    'Regional Water Engineer Arusha',
    'Quickwi',
    'Dhv Moro',
    'Hewasa',
    'Tasaf And Lga',
    'Jaica',
    'Village Res',
    'Kkkt-dioces Ya Pare',
    'Aic',
    'Solidarm',
    'Christan Outrich',
    'Kanisa La Menonite',
    'Islamic',
    'Rc',
    'Killflora',
    'Bread For The Wor',
    'Wua',
    'Mac',
    'Caltaz Kahama',
    'Mianz',
    'Dw',
    'Makanyushini'
```

```
 'Makapuchini',
 'Tasaf Ii',
 'Omar Ally',
 'Md',
 'Mitema',
 'Ham',
 'Quwkwin',
 'Do',
 'Dh',
 'Bokera W',
 'Bulyahunlu Gold Mine',
 'Mbiuwasa',
 'The Isla',
 'Rotary Club',
 'Muslims',
 'Care International',
 'Kimkuma',
 'Tanesco',
 'Mbozi District Council',
 'Dasip',
 'Tltc',
 'Sdg',
 'Hsw',
 'Mwaya Mn',
 'Resolute Mining',
 'Tz Japan',
 'Roman Cathoric Same',
 'Concern',
 'Caritas',
 'Conce',
 'Huches',
 'Wamarekani',
 'Devon Aid Korogwe',
 'Kiliwater',
 'Lamp',
 'Bsf',
 'Mem',
 'Jeica',
 'Father Bonifasi',
 'Bgm',
 'Lcgd',
 'Karadea Ngo',
 'Msf/tacare',
 'Fathe',
 'Unice',
 'Mdc',
 'Dasp',
 'Songea Municipal Counci',
 'Tasae',
 'Water User As',
 'Msikiti',
 'Cct',
 'Islamic Found',
 'Tgrs',
 'Unicef/ Csp',
 'Jimbo Fund',
 'Tlc/john Majala',
 'Magoma Adp',
 'Vwc',
 'Pidp',
 'Japan Government',
 'Kata',
 'De',
 'Acra',
 'Gtz',
 'Isf/government',
 'Kuwasa',
 'China Government',
 'Taboma',
 'P',
 'Kingupira S',
 'Churc',
 'Walakala'
```

```
    'Walokole',
    'Mkinga  Distric Cou',
    'Cafod',
    'Hw/rc',
    'Sumbawanga Munici',
    'Tacare',
    'Urt',
    'Camavita',
    'Member Of Parliament',
    'Dmmd',
    'Aqua Blues Angels',
    'Water Aid /sema',
    'Kirdep',
    'Cc Motor Day 2010',
    'Kilwater',
    'Ndrdp',
    'Hez',
    'Nethe',
    'Denat',
    'Kibo Brewaries',
    'Arab Community',
    'Elct',
    'Adp',
    'Priva',
    'Holland',
    'Rc Church/centr',
    'Cocen',
    'Wfp/tnt',
    'Lench Taramai',
    'Ncaa',
    'Mzee Don',
    'World Vision/ Kkkt',
    'Finwater',
    'Kuamu',
    'Dwssp',
    'Musilim Agency',
    'Ukiligu',
    'Wamakapuchini',
    'Mbunge',
    'The Desk And Chair Foundat',
    'Duwas',
    'Diwani',
    'Kkkt Church',
    'Ea',
    'Halmashauri Ya Manispa Tabora',
    'Finidagermantanzania Govt',
    'Bahewasa',
    'Jika',
    'Asb',
    'Qwiqwi',
    'Pmo',
    'Tuwasa',
    'Irish Ai',
    'Mdrdp',
    'Jeshi La Wokovu',
    'Government /tassaf',
    'Mboma',
    'People From Japan',
    'Kilindi District Co',
    'Shamte Said',
    'Auwasa',
    'Kidp',
    'Tridep',
    'St',
    'Wd And Id',
    'Serikali',
    'Kanisa Katoliki',
    'Po',
    'Ga',
    'Cocern',
    'Finida German Tanzania Govt',
    'National Rural And Hfa',
    'K'
    'K',
    'Idara Ya Maji',
    'Moslem Foundation',
    'Swiss If',
    'Miziriol',
    'Yasini Selemani',
    'Dbspe',
    'H',
    'A/co Germany',
    'Oikos E.Africa/european Union',
    'Hydom Luthelani',
    'Ilct',
    'Peter Tesha',
    'Ms',
    'Mzungu Paul',
    'Caltas',
    'Red Cross',
    'Losaa-kia Water Supply',
    'Tassaf I',
    'Kanisa Katoliki Lolovoni',
    'Finland Government',
    'Gaica',
    'Institution',
    'Tcrs.Tlc',
    'Magereza',
    'Loliondo Parish',
    'Diocese Of Geita',
    'Total Landcare',
    'U.S.A',
    'Tdft',
    'Parastatal',
    'Rished',
    'Dwt',
    'The People Of Japan',
    'Kcu',
    'Abd',
    'Village Government',
    'Msabi',
    'Vc',
    'Cmsr',
    'Konoike',
    'Roman Catholic Rulenge Diocese',
    'Bened',
    'Shule',
    'W',
    'Partage',
    'Inkinda',
    'Robert Loyal',
    'Africa Amini Alama',
    'Imf',
    'L',
    'Moroil',
    'Sekei Village Community',
    'Us Embassy',
    'Missionaries',
    'Tcrs /government',
    'Desk And Chair Foundation',
    'Ms-danish',
    'Wsdp & Sdg',
    'Roman Cathoric-same',
    'Cefa-njombe',
    'Aar',
    'Village Govt',
    'Farm Africa',
    'Mheza Distric Counc',
    'Chamavita',
    'Mileniam Project',
    'Undp/ilo',
    'Dads',
    'Institutional',
    'Sowasa',
    'Ccpk',
    'Tasafu'
```

```
 'Tasalu',
 'Government/ World Bank',
 'Luthe',
 'Wirara Ya Maji',
 'Mzee Mkungata',
 'Rada',
 'Twesa',
 'Plan Internatio',
 'Solidame',
 'Rwsso',
 'Williamson Diamond Ltd',
 'Tag',
 'Dar Al Ber',
 'Watu Wa Ujerumani',
 'Dwe/bamboo Projec',
 'Danida /government',
 'Semaki K',
 'Arabs Community',
 'Water Aid/sema',
 'District Rural Project',
 'Gen',
 'Redep',
 'Kiwanda Cha Samaki',
 'Singida Yetu',
 'Rwsp',
 'Moravian',
 'Sema S',
 'Cbhi',
 'Tcrs /care',
 'Makonde',
 'Millenium',
 'Swisland/mount Meru Flowers',
 'Kigoma Municipal',
 'Kinapa',
 'People Of Japan',
 'Kijij',
 'Wfp/tnt/usaid',
 'Tanapa',
 'Efg',
 'Local',
 'Kyariga',
 'Tanzakesho',
 'Roman Cathoric -kilomeni',
 'World Vision/adra',
 'Mbozi Secondary School',
 'Tasaf/dmdd',
 'Mws',
 'Shekhe',
 'Pataji',
 'Tahea',
 'Kalta',
 'Pentecosta Church',
 'Sekondari',
 'Kyela Council',
 'Kalitasi',
 'Quick Wins',
 'Lowasa',
 'Hotels And Loggs Tz Ltd',
 'Cobashec',
 'Orphanage',
 'Adf',
 'Wwf',
 'Idydc',
 'Cper',
 'School',
 'Ilo',
 'Olumuro',
 'Villaers',
 'Tlc/thimotheo Masunga',
 'Dak',
 'Kidep',
 'Ubalozi Wa Marekani',
 'Dmk Anglican',
 'Dmk Anglican',
 'Franc',
 'Ka',
 'Mgm',
 'Aimgold',
 'Mzee Omari',
 'Petro Patrice',
 'Camartec',
 'Loliondo Secondary',
 'Islamic Agency Tanzania',
 'Tanz Egypt Technical Cooper',
 'Safari Roya',
 'Koica',
 'Rdc',
 'Total Land Care',
 'Pad',
 'Msf',
 'Mamad',
 'Padep',
 'One Un',
 'Fabia',
 'Lake Tanganyika',
 'Italy',
 'Solar Villa',
 'Roman Church',
 'Singasinga',
 'Rc/mission',
 'In',
 'Adp Mombo',
 'Pci',
 'Norad/ Tassaf Ii',
 'I Wash',
 'Bs',
 'Kambi Migoko',
 'Ai',
 'Sauwasa',
 'Icdp',
 'Rotte',
 'Dhv/gove',
 'Kmcl',
 'Ccps',
 'Si',
 'Rundu Man',
 'Serikari',
 'Undp/aict',
 'Hdv',
 'Halmashauri',
 'Concern /govern',
 'Quick Win Project /council',
 'Mh Kapuya',
 'Halmashauri Ya Wilaya',
 'Baric',
 'Cpro',
 'Getekwe',
 'Gain',
 'Wahidi',
 'Asdp',
 'Kadp',
 'Aco/germany',
 'Majengo Prima',
 'Hortanzia',
 'Quick',
 'Hasnan Murig (mbunge)',
 'Ikeuchi Towels Japan',
 'Halmashauli',
 'Acord',
 'Menon',
 'Wate Aid/sema',
 'Dwe/ubalozi Wa Marekani',
 'Vifafi',
 'Cdg',
 'Kwasenenge Group',
 'Dad/rwssp'
```

```
    'Ded/rwssp',
    'Oldonyolengai',
    'None',
    'Village Community',
    'Minjingu',
    'El',
    'D',
    'Songas',
    'Mi',
    'Action Aid',
    'Tanroad',
    'Lake Tanganyika Basin',
    'Pwc',
    'Teonas Wambura',
    'Mgaya Masese',
    'Stantons',
    'Sao H',
    'Ukida',
    'Taasaf',
    'Mwita Kichere',
    'Lwf',
    'Mosque',
    'Peter Ngereka',
    'Svn',
    'Investor',
    "Ju-sarang Church' And Bugango",
    'Lgcdg',
    'Action Contre La Faim',
    'Kwamdulu Estate',
    'Quick Wins Scheme',
    'Cpps',
    'Belgian Government',
    'Cmcr',
    'Care Int',
    'Mavuno Ngo',
    'Niger',
    'Mwanza',
    'Zaburi And Neig',
    'Women For Partnership',
    'Artisan',
    'Sisa',
    'Cdcg',
    'Ndm',
    'Secondary',
    'Da Unoperaio Siciliano',
    'Town Council',
    'Lions Club',
    'Lutheran Church',
    'Shirika La Kinamama Na Watot',
    'Pangadeco',
    'Uyoge',
    'Canada',
    'Frankfurt',
    'Redet',
    'Rural Water Department',
    'Buptist',
    'Unp/aict',
    'Timothy Shindika',
    'Village Office',
    'Lotary Club',
    'Hesaw',
    'Malec',
    'Kuji Foundation',
    'Mamvua Kakungu',
    'Rusumo Game Reserve',
    'Mtuwasa And Community',
    'W.D.&.I.',
    'Act Mara',
    'Sda',
    'Mzinga A',
    'Vgovernment',
    'Re',
    'Lesgin',
    'Loocip',
    'Sua',
    'Brdp',
    'Hamref',
    'Happy Watoto Foundation',
    'Gdp',
    'Lgdcg',
    'Jgb',
    'Mfuko Wa Jimbo',
    'Doddea',
    'Maliasili',
    'Roman Ca',
    'Tcrst',
    'Holla',
    'African Development Foundation',
    'Fptc - Pent',
    'Makona',
    'Oxfam Gb',
    'African 2000 Network',
    'Netherland',
    'Tabraki',
    'Balo',
    'Dadp',
    'Ikela Wa',
    'Rotary I',
    'Rwssp/wsdp',
    'Christian Outrich',
    'Cipro/care/tcrs',
    'Italian',
    'Kome Parish',
    'Mwanga Town Water Authority',
    'Jumanne Siabo',
    'Hindu',
    'Rural',
    'H/w',
    'Tanap',
    'Roman Cathoric Church',
    'Rombo Dalta',
    'Ilwilo Community',
    'Un/wfp',
    'St Ph',
    'Lwiji Italy',
    'Livin',
    'Cg',
    'Hhesawa',
    'Lwi & Central Government',
    'Lc',
    'Kkkt Leguruki',
    'Tanzania Compasion',
    'Louise Elucas Sala',
    'Hiap',
    'Cpps Mission',
    'Matyenye',
    'Dimon',
    'Italy Government',
    'Tag Church Ub',
    'Aic Church',
    'Wvc',
    'Lgcbg',
    'Tacri',
    'Chai Wazir',
    'Hasnein Murij',
    'Rural Water Supply And Sanita',
    'Simba Lodge',
    'Free Pentecoste Church Of Tanz',
    'Summit For Water',
    'Sanje Wa',
    'Makundya',
    'Uhai Wa Mama Na Mtoto',
    'Ola',
    'Ba As',
    'Tredep',
    'Nyanza Road',
    'Nyanza Road',
    'Cgc',
    'Swidish',
    'Kizenga',
    'Hapa',
    'Ramadhani Nyambizi',
    'Denish',
    'Mkuyu',
    'Ras',
    'Mwinjuma Mzee',
    'Gachuma Ginery',
    'Resolute',
    'Morovian',
    'Water Board',
    'Kigoma Municipal Council',
    'Mafwimbo',
    'Pentecostal',
    'Rocci Ross',
    'Igolola Community',
    'Pancrasi',
    'S',
    'Rdws',
    'Said Omari',
    'Ngiresi Village Community',
    'Kilomber',
    'Sharifa Athuman',
    'Qwickwin',
    'Mwita Muremi',
    'Mbwana Omari',
    'Tlc/samora',
    'Mmem',
    'Haydom Lutheran Hospital',
    'Vicfish Ltd',
    'Afroz Ismail',
    'Sisal Estste Hale',
    'Eu',
    'Korea',
    'Cvs Miss',
    'Moradi',
    'Living Water International',
    'Kajima',
    'Uaacc',
    'Germany Misionary',
    'Rips',
    'France',
    'Bukumbi',
    'Rhobi',
    'Kiwanda Cha Tangawizi',
    'Ten Degree Hotel',
    'Wssp',
    'Meru Concrete',
    'Gg',
    'Wizara',
    'Segera Estate',
    'Hospital',
    'Dmk',
    'Siza Mayengo',
    'Greec',
    'Makli',
    'Mp',
    'Islam',
    'Dassip',
    'Rvemp',
    'Adp Bungu',
    'Thomasi Busigaye',
    'Sijm',
    'W.D & I.',
    'British Tanza',
    'Kkkt Ndrumangeni',
    'Tag Church',
    'Council',
    'Usambala Sister',
    'Hearts Helping Hands Inc.',
    'Hearts Helping Hands.Inc.',
    'Idea',
    'Filo',
    'Qwekwin',
    'Selous G',
    'Pentecostal Hagana Sweeden',
    'Ester Ndege',
    'Oikos E .Africa/european Union',
    'Nyabarongo Kegoro',
    'Quik',
    'Ringo',
    'Kanisani',
    'Wfp/usaid/tnt',
    'Village Council/ Haydom Luther',
    'Fpct Church',
    'Mzung',
    'Kwikwiz',
    'Kanisa La Mitume',
    'Iom',
    'Oda',
    'Caltus',
    'Gt',
    'Malola',
    'Water Project Mbawala Chini',
    'Totoland Care',
    'Nddp',
    'Kmt',
    'Anjuman E Seifee',
    'Nginila',
    'Usa Embassy',
    'Village',
    'Pdi',
    'T',
    'Hery',
    'Obc',
    'Nyamongo Gold Mining',
    'Women Fo Partnership',
    'Sister Francis',
    'Norani',
    'Mahita',
    'Kalebejo Parish',
    'Aixos',
    'Government',
    'Wrssp',
    'Ddp',
    'Game Division',
    'Rudep /dwe',
    'Kashwas',
    'Twende Pamoja',
    'Gwitembe',
    'Makori',
    'Sangea District Council',
    'Unicef/central',
    'Africa 2000 Network/undp',
    'Mmanya Abdallah',
    'Snv Ltd',
    'Taes',
    'Canada Aid',
    'Senapa',
    'Regwa Company Of Egypt',
    'Water Se',
    'Mamlaka Ya Maji Ngara',
    'Wama',
    'Prf',
    'Church',
    'Magadini Makiwaru Water',
    'Kayempu Ltd',
    'Trachoma',
    'Seleman Rashid',
    'Afriican Reli',
    'Tassaf Ii',
    'Samsoni',
    'Quick Wings',
    'Quick Wings',
    'Ngos',
    'Kurrp Ki',
    'Cast',
    'Rudep/norad',
    'Kwa Mzee Waziri',
    'Panone',
    'Lawate Fuka Water Suppl',
    'St Gasper',
    'Wug And Ded',
    'Pr',
    'Mmg Gold Mine',
    'Nordic',
    'Mchukwi Hos',
    'Dwst',
    'Serikaru',
    'African Realief Committe Of Ku',
    'Fao',
    'Scott',
    'Mzungu',
    'Vttp',
    'Vi',
    'Irish Government',
    'Namungo Miners',
    'Nassor Fehed',
    'Dbfpe',
    'Clause',
    'Busoga Trust',
    'Mzee Mabena',
    'Br',
    'Brad',
    'Koico',
    'Healt',
    'Ro',
    'Jeshi Lawokovu',
    'Paffect Mwanaindi',
    'Tansi',
    'Craelius',
    'Apm[africa Precious Metals Lt',
    'Zao Water Spring X',
    'Shinyanga Shallow Wells',
    'Cipro/care',
    'Vifaf',
    'Mtc',
    'Lungwe',
    'Dhinu',
    'Aic Kij',
    'Mataro',
    'Dagida',
    'Redap',
    'Nwssp',
    'Lench',
    'Wanakijiji',
    'Nk',
    'Nimrodi Mkono[mb]',
    'Maro',
    'Professor Ben Ohio University',
    'Rafael Michael',
    'Tdrs',
    'Bra',
    'Suwasa',
    'Twig',
    'Tanzania Egypt Technical Co Op',
    'Lifetime',
    'Comunedi Roma',
    'Unhcr/danida',
    'Bread Of The Worl',
    'Lutheran',
    'Tasf',
    'Rc Cathoric',
    'Halmashauri Wil',
    'Mgaya',
    'Grail Mission Kiseki Bar',
    'Grail Mission Kiseki Bar',
    'Answeer Muslim Grou',
    'John Gileth',
    'Care/dwe',
    'Liuwassa',
    'Ustawi',
    'Nssf',
    'Kilol',
    'Nado',
    'Judge Mchome',
    'Minis',
    'Milenia',
    'Water User Group',
    'Opec',
    'Government /sda',
    'Farm-africa',
    'Bffs',
    'Kyela-morogoro',
    'Ggm',
    'Msikitini',
    'Kwik',
    'Shelisheli Commission',
    'Mungaya',
    'Baptist Church',
    'Tgts',
    'Unknown',
    'Ndorobo Tours',
    'Zaben',
    'Serikali Ya Kijiji',
    'Enyueti',
    'Watu Wa Marekani',
    'Regina Group',
    'Snv-swash',
    'Seram',
    'Lcdg',
    'Adap',
    'Laizer',
    'African Barrick Gold',
    'Salehe',
    'Jumanne',
    'Masai Land',
    'Jipa',
    'S. Kumar',
    'Hpa',
    'Mp Mzeru',
    'W.D &',
    'Wafidhi Wa Ziwa T',
    'Matimbwa Sec',
    "Lee Kang Pyung's Family",
    'Rwsssp',
    'Rural Drinking Water Supply',
    'Mhoranzi',
    'Woyege',
    'Quick Win Project',
    'Muslimu Society(shia)',
    'Morovian Church',
    'Grazie Franco Lucchini',
    'Pankrasi',
    'Irevea Sister Water',
    'Unesco',
    'Iucn',
    'Kdc',
    ...]
```

In [18]:

```
df[df['funder'] == '0']
```

Out[18]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
145	[illegible]	50.0	2013-03-21	0	12	0	[illegible]	-6.892593	[illegible]	0	...	
152	60983	0.0	2013-03-16	0	-15	0	39.527114	-6.988748	Msikitini	0	...	
393	39749	0.0	2013-03-18	0	28	0	39.159887	-6.902548	Kwa Chambuso	0	...	
417	15832	50.0	2013-03-22	0	30	0	39.178404	-6.938013	Ccm Kivule	0	...	
428	50233	0.0	2013-03-12	0	30	0	39.178849	-6.973206	Ofisi Ya Kata	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
59237	2138	59.0	2013-03-19	0	81	0	39.119109	-6.898919	Kata	0	...	
59243	3396	50.0	2013-03-16	0	-20	0	39.524021	-6.984802	Kwa Mariwala	0	...	
59276	62818	50.0	2013-03-21	0	18	0	39.183790	-6.897566	Kwa Mkunduge	0	...	
59351	55322	50.0	2013-03-18	0	-19	0	39.534599	-7.088183	Kwa China	0	...	
59387	26640	100.0	2013-03-12	0	25	0	39.176480	-6.957098	Kwa Maliba	0	...	

777 rows × 42 columns

We saw in the value_counts that there was already a 'Not Known' value so the nulls were changed to 'Not Known'.

In [19]:

```
#replace nulls in 'funder' col with "Not Known"
df['funder'].fillna(value="Not Known", inplace=True)
```

A value of '0' was also treated a a null. These values seemed out-of-place in the column and didn't seem to describe a particular organization or individual responsible for funding.

In [20]:

```
df['funder'] = df['funder'].replace(to_replace={'0':'Not Known'}, value=None)
```

In [21]:

```
#confirm replacement of nulls
df['funder'].isna().sum()
```

Out[21]:

```
0
```

In [22]:

```
df[df['funder'] == '0']
```

Out[22]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	quality_group

0 rows × 42 columns

We've confirmed removal of null values and '0' values from the 'funder' column.

installer

In [23]:

```
#organization that installed the well
df['installer'].value_counts()
```

Out[23]:

```
DWE                   17402
Government             1825
RWE                    1206
Commu                  1060
DANIDA                 1050
                     ...
Ta                        1
Zao                       1
Marumbo Community         1
BATIST CHURCH             1
Regina group              1
Name: installer, Length: 2145, dtype: int64
```

Installer had 2,154 unique values which could be an issue if we planned on including this as a categorical feature in our modeling.

In [24]:

```
df['installer'].isna().sum()
```

Out[24]:

```
3655
```

There were 3,655 nulls in 'installer'. This value matched the number of nulls we saw in 'funder', which could suggest that the same rows had missing values.

In [25]:

```
list(df['installer'].unique())
```

Out[25]:

```
['Roman',
 'GRUMETI',
 'World vision',
 'UNICEF',
 'Artisan',
 'DWE',
 'DWSP',
 'Water Aid',
 'Private',
 'DANIDA',
 'Lawatefuka water sup',
 'WEDECO',
 'Danid',
 'TWE',
 'ISF',
 'Kilolo Star',
 'District council',
 'Water',
 'WU',
 nan,
 'Not known',
 'Central government',
 'CEFA',
 'Commu',
 'Accra',
 'World Vision',
 'LGA',
 'MUWSA',
 'KKKT _ Konde and DWE',
 'Government',
 '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',
 'CES',
 'District Counci',
 'Ruthe',
 'Adra/Community',
 'Tulawaka Gold Mine',
 'KKT C',
 'Hesawa',
 'Water board',
 'LOCAL CONTRACT',
 'WFP',
 'LIPS',
 'TASAF',
 'World',
 '0',
 'SW',
 'Shipo',
 'Fini water',
 'Kanisa',
 'OXFARM',
 'VILLAGE COUNCIL Orpha',
 'Villagers',
 'Idara ya maji',
 'FPCT',
 'WVT',
 'Ir',
 'DANID',
 'Angli',
 'secondary school',
 'Amref',
 'JBG',
 'DADIS',
 'International Aid Services',
 'RW',
 'Dmdd',
 'TCRS',
 'RC Church',
 'WATER AID',
 'JICA',
 'Gwasco L',
 'AF',
 'AMREF',
 'wananchi',
 'FW',
 'Central Government',
 'MWE &',
 'Gove',
 'RC CHURCH',
 'TDFT',
 'RWE/DWE',
 'Central govt',
 'World Bank',
 'TWESA',
 'Norad',
 'Hans',
 'FinW',
 'FIN WATER',
 'OXFAM',
 'Plan Internationa',
 'District Council',
 'RWEDWE',
 'Fini Water',
 'ANGLI',
 'CDT',
 'North',
 'Oikos E .Africa',
 'SHAWASA',
 'UN',
 'NORAD',
 'Save the rain',
 'John gemuta co',
 'TLC',
 'RC Churc',
 'Plan Int',
 'Phase',
 'LVIA',
 'Rhobi',
 'Makonde water population',
 'RWE/ Community',
 'Is',
 'KILI WATER',
 'RDDC',
 'FINN WATER',
 'FINI WATER',
 'DHV',
 'Kamama',
 'DDCA',
 'Victoria company',
 'RWSSP',
 'Ce',
 'KYASHA ENTERPR',
 'ERETO',
 'REDESO',
 'Villa',
 'Priva',
 'KUWAIT',
 'Mw',
 'Magadini-Makiwaru wa',
 'Dr. Matomola',
 'Af',
 'RCchurch/CEFA',
 'Tardo',
 'GOVERNMENT',
 'Individuals',
 'Chamavita',
 'GEN',
 'Missi',
 'Safari Roya',
 'DAWASCO',
 'Gover',
 'Mission',
 'DWE/',
 'Halmashauri ya wilaya sikonge',
 'Ki',
 'Rhoda',
 'HAPA SINGIDA',
 'Consulting Engineer',
 'Karugendo',
 'Co',
 'Marafip',
 'COSMOS ENG LTD',
 'World banks',
 'Tanz',
 'Handeni Trunk Main(',
 'SIMBA CO',
 'Local technician',
 'Village',
 'Centr',
 'CONS',
 'DW',
 'DCT',
 'IRC',
 'District water department',
 'Sabodo',
 'MLADE',
 'I.E.C',
 'LWI',
 'Kiliflora',
 'ICS',
 'T. N. karugendo',
 'DED',
 'Kuwait',
 'ADP',
 'JUIN CO',
 'BILILA',
 'TPP',
 'GOVER',
 'CIPRO/Government',
 'MWE',
 'MTUWASA',
 'Unisef',
 'REGIONAL WATER ENGINEER ARUSHA',
 'IDARA',
 'Wizara ya maji',
 'Tasaf and Lga',
 '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',
 'DH',
 'RC Ch',
 'SAXON BUILDING CONTRACTOR',
 'Bokera W',
 'Bulyahunlu Gold Mine',
 'MBIUWASA',
 'ADRA /Government',
 'The Isla',
 'Rotary club',
 'YELL LTD',
 'Care international',
 'KIMKUM',
 'Tanesco',
 'CJEJOW CONSTRUCTION',
 'Victoria',
 'TLTC',
 'Wachina',
 'WE',
 'HSW',
 'Communit',
 'Kibaha Town Council',
 'Dr. Matobola',
 'Go',
 'DWR',
 'Huches',
 'WATERAID',
 'Maswi company',
 'Kiliwater',
 '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 Majala',
 'Kilwa company',
 'Local  technician',
 'TASSAF',
 'VWC',
 'PIDP',
 'TAN PLANT LTD',
 'Japan Government',
 'Kata',
 'GTZ',
 'ISF/Government',
 'KUWASA',
 'Hydrotec',
 'Pr',
 'Ch',
 'Jaica',
 'Taboma/Community',
 'P',
 'Ubung',
 'Chur',
 'BESADA',
 'Action Contre La Faim',
 'Wanjoda',
 'CBHCC',
 'HW/RC',
 'Sumbaw',
 'CCEC',
 'Nice',
 'CCT',
 'World Vission',
 'Inter',
 'DMMD',
 'WORLD BANK',
 'AQUA BLUES ANGELS',
 'MACK DONALD CONTRACTOR',
 'Water Aid /sema',
 'Henure Dema',
 'Kirdep',
 'ADRA/Government',
 'Kilwater',
 'Da',
 'Villi',
 'KOYI',
 'AD',
 'Arab community',
 'District water depar',
 'HOLLAND',
 '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',
 'DUWAS',
 '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 wokovu',
 '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',
 'Padep',
 'Po',
 'Village Counil',
 'MINISTRY OF WATER',
 'Ga',
 'K',
 'Swiss If',
 'Miziriol',
 'Yasini Selemani',
 'DBSPE',
 'European Union',
 'H',
 'TPP TRUSTMOSHI',
 'Atisan',
 'Jika',
 'ISF/TACARE',
 'Oikos E.Africa',
 'Hydom Luthelani',
 'Kalumbwa',
 'ILCT',
 'MS',
 'RUVUMA BASIN',
 'Gold star',
 'Mi',
 'Mzungu Paul',
 'Kanisa katoliki',
 'Caltas',
 'RED CROSS',
 'World bank',
 'Losaa-Kia water supp',
 'Jica',
 'PET',
 'Finland Government',
 'GAICA',
 'Institution',
 'TCRS/TLC',
 'Loliondo Parish',
 'GACHUMA CONSTRUCTION',
 'Diocese of Geita',
 'Villages',
 'MSABI',
 'Total landcare',
 'VICTORIA DRILL CO',
 'U.S.A',
 'VTECOS',
 'COW',
 'Vill',
 'Contr',
 'Wadeco',
 'KIM KIM CONSTRUCTION',
 'Msabi',
 'VC',
 'CMSR',
 'Ko',
 'Roman Catholic Rulenge Diocese',
 'Shule',
 'W',
 'inkinda',
 'Africa Amini Alama',
 'Consultant',
 'L',
 'Moroil',
 'Sekei village community',
 'US Embassy',
 'PIT COOPERATION  LTD',
 'Do',
 'world',
 'Government /TCRS',
 'UNHCR',
 'DESK C',
 'Dr.Matomola',
 'FOLAC',
 'Village govt',
 'BSF',
 'Roman Cathoric Same',
 'RWE/Community',
 'Mileniam project',
 'ACTIVE TANK CO',
 'Ncaa',
 'Africa Islamic Agency Tanzania',
 'Max Mbise',
 'DADS',
 'Institutional',
 'SOWASA',
 'CCPK',
 'AUSTRALIA',
 'not known',
 'Kalago enterprises Co.Ltd',
 'Roman Catholic',
 'NANRA contractor',
 'WORLD VISION',
 'No',
 'ADP Busangi',
 'TSRC',
 'SOLIDAME',
 'Barry A. Murphy',
 'Tanzania Government',
 'WILLIAMSON DIAMOND LTD',
 'TAG',
 'The I',
 'Total Landcare',
 'CENTRAL GOVERNMENT',
 'Arabs Community',
 'Secondary school',
 'Water Aid/Sema',
 'Jiks',
 'Konoike',
 'ABASIA',
 'LAMP',
 'SINGIDA YETU',
 'RWSP',
 'MDALA Contractor',
 'Netherlands',
 'DWT',
 'TCRS /CARE',
 'Makonde',
 'Japan',
 'Milenium',
 'Goldstar',
 'District COUNCIL',
 'MUWASA',
 'Green',
 'Kigoma municipal',
 'KINAPA',
 'CHINA HENAN CONTRACTOR',
 'Musa',
 'TANAPA',
 'Ministry of water engineer',
 'EFG',
 'MASWI',
 'Kyariga',
 'Roman Cathoric -Kilomeni',
 'Mbozi Secondary School',
 'TASAF/DMDD',
 'MWS',
 'Roman catholic',
 'Shekhe',
 'Rished',
 'KONOIKE',
 'Pata',
 'TAHEA',
 'Luthe',
 'Kalta',
 'Pentecost church',
 'Amboni Plantation',
 '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',
 'British',
 'Msuba',
 'Villaers',
 'TLC/Thimotheo Masunga',
 'WB',
 'Council',
 'DAK',
 'COCANE',
 'WINAMU CO',
 'Ubalozi wa Marekani',
 'Conce',
 'BGM',
 'DMK',
 'Mviwa',
 'KA',
 'MGM',
 'AIMGOLD',
 'YEBE CHIKOMESH',
 'Omari Mzee',
 'Petro Patrice',
 'Camartec',
 'Total land care',
 'Wasso companies',
 '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/sema',
 'Save the rain USA',
 'Plan Tanzania',
 'Roman Church',
 'Singasinga',
 'RC/Mission',
 'In',
 'V',
 'Korogwe water works',
 'PCI',
 'Atlas',
 'DWE /TASSAF',
 'Local te',
 'World Division',
 'Gwaseco',
 'Kambi Migoko',
 'AI',
 'SAUWASA',
 'Nyakilanganyi',
 'DEE',
 'MANYARA CONSTRUCTION',
 'Rotte',
 'KMCL',
 'LINDALA CO',
 'Government /Community',
 'CCPS',
 'SI',
 'Rundu man',
 'Water Aid/sema',
 'Naishu construction co. ltd',
 'WOULD BANK',
 'Mark',
 'Cosmo',
 'Halmashauri',
 'Concern /government',
 'Quick win project',
 'Mh Kapuya',
 'Halmashauri ya wilaya',
 'Edward',
 'COMMU',
 'Baric',
 'Consuting Engineer',
 'JANDU PLUMBER CO',
 'FiNI WATER',
 'CPRO',
 'Getekwe',
 'Jicks',
 'Wahidi',
 'Mohamed Ally',
 'ASDP',
 'CITIZEN ENGINE',
 'KADP',
 'Dar es salaam Technician',
 'Halmashauli',
 'ACORD',
 'MA',
 'Water  Aid/Sema',
 'RC church/CEFA',
 'Wedeco',
 'DWE/Ubalozi wa Marekani',
 'VIFAFI',
 'Kwasenenge Group',
 'Cosmos Engineering',
 'OLDONYOLENGAI',
 'NYAKILANGANI CO',
 'Village Community',
 'MINJINGU',
 'EL',
 'Songa',
 'Consultant and DWE',
 'AC',
 'Gain',
 'DASIP',
 'TANROAD',
 'Tasaf',
 'Wasso',
 'Teonas Wambura',
 'Mgaya Masese',
 'TUKWALE ENTERP',
 'Sao',
 'MWAKI CONTRACTOR',
 'VIEN CONSTRUCTION',
 'mwita kichere',
 'DADS/village community',
 'Africare',
 'Mosque',
 'Chiko',
 'central government',
 'VITECOS',
 'IN',
 'Msikiti',
 'Word Bank',
 'Kwamdulu estate',
 'SEMA Consultant',
 'Concern',
 'Belgiam Government',
 'Wanan',
 'Exaud Msambwa',
 'Niger',
 'MWANZA',
 'SONGAS',
 'MINISTRYOF WATER',
 'COMMUNITY',
 'Zaburi and neighbors',
 'NDM',
 'Killflora/ Community',
 'PART',
 'secondary',
 "lion's club",
 'lutheran church',
 'Mileniam',
 'UYOGE',
 'Christina Magoge',
 'Canada na Tanzania',
 'FRANKFURT',
 'GOVERM',
 'Kuji foundation',
 'Mamvua Kakungu',
 'Rusumo Game reserve',
 'MTUWASA and Community',
 'ACT MARA',
 'UMOJA DRILLING',
 'KkKT',
 'SDA',
 'Mzinga A',
 'RE',
 'LOOCIP',
 'SUA',
 'RUNDAGA',
 'RWE /Community',
 'Wo',
 'Happy watoto foundation',
 'GDP',
 'ViLLAGE COUNCIL',
 'MBULU DISTRICT COUNCIL',
 'Maliasili',
 'Roman Ca',
 'NZILA',
 'stansilaus',
 'AFRICAN DEVELOPMENT FOUNDATION',
 'FPTC',
 'KARUMBA BIULDING COMPANY LTD',
 'Kalugendo',
 'Village Government',
 'Tabraki',
 'MASWI DRILLING',
 'Ikela Wa',
 'Shallow well',
 'WEDECO/WESSONS',
 'CIPRO/CARE/TCRS',
 'Wasso contractors',
 'villagers',
 'Mwanga town water authority',
 'Jumanne Siabo',
 'Mama Kalage',
 'Hindu',
 'Rural',
 'TANAP',
 'Makonde water supply',
 'villigers',
 'Bingo foundation Germany',
 'Ilwilo community',
 'St ph',
 'WDECO',
 'LIVI',
 'Pet Corporation Ltd',
 'DWE & LWI',
 'LC',
 'KKKT Leguruki',
 'HIAP',
 'Matyenye',
 'DIMON',
 'Italy government',
 'MASWI DRILL',
 'WVC',
 'TACRI',
 'Hasnein Murij',
 'SIMBA LODGE',
 'Faudh Tamimu',
 'Free Pentecoste Church of Tanz',
 'Summit for water/Community',
 'Sanje Wa',
 'Makundya',
 'Individual',
 'OLA',
 'RC C',
 'TREDEP',
 'Consultant Engineer',
 '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 kijiji',
 'Counc',
 'Igolola community',
 'S',
 'NYAKILANGANI CONSTRUCTION',
 'RDWS',
 'Said Omari',
 'AFRICA MUSLIM',
 'IADO',
 'W/',
 'Ngiresi village community',
 'UDC/Sema',
 'AMP contractor',
 'rc ch',
 'QWICKWIN',
 'Mwita Muremi',
 '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 water international',
 'Kajima',
 'Missio',
 'UAACC',
 'GERMANY MISSIONARY',
 'MI',
 'Rips',
 'LVA Ltd',
 'BUKUMB',
 'Taasi',
 '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',
 'water board',
 'AMP Contract',
 'Thomasi busigaye',
 'Local technitian',
 'SIJM',
 'KKKT Ndrumangeni',
 'YUMBAKA ENGINEERING',
 'TAG CHURCH',
 'Usambala sisters',
 'KOBERG Contractor',
 'hesawa',
 'Water Authority',
 'Mr Chi',
 'Hearts helping hands.Inc.',
 'IDEA',
 'Selous G',
 'SULEMAN IDD',
 'Pump entecostal Sweeden',
 'Ester Ndege',
 'Nyabarongo Kegoro',
 'Canop',
 'QUIK',
 'DADP',
 'Kanisani',
 'CARTAS',
 'Mzung',
 'wizara ya maji',
 'VILLAGE COUNCIL .ODA',
 'CG',
 'Caltus',
 'Cons',
 'ISSAC MOLLEL',
 'malola',
 'DCCA',
 'Juma Maro',
 'Water Project Mbawala chini',
 'Unicef',
 'Totoland care',
 'Maswi drilling co ltd',
 'NDDP',
 'KMT',
 'NGINIL',
 'Serengeti District concil',
 'RC church',
 'VILLAG',
 'Local technical tec',
 'Cultus',
 'T',
 'Hery',
 'OBC',
 'RUDEP',
 'RWE Community',
 'Nyamongo Gold mining',
 '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',
 'Ardhi water well',
 'Distric Water Department',
 'gwitembe',
 'Conta',
 'HOWARD HUMFREYS',
 'SHUWASA',
 'JANDU PLUMBER  CO',
 'Makori',
 'Sangea District Coun',
 'CHINA',
 'British colonial government',
 'Maendeleo ya jamii',
 'CARITAS',
 'Taes',
 'KWIKWIZ',
 'SEMA CO LTD',
 'SENAPA',
 'REGWA COMPANY OF EGYPT',
 'COBASHEC',
 'AQUA Wat',
 'Dr.Matobola',
 'Central basin',
 'Mamlaka ya maji ngara',
 'PRF',
 'Church',
 'Magadini 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',
 'WINNIN SPIRIT CO',
 'Ha',
 'MMG GOLD MINE',
 'P.N.R.',
 'Nandra Construction',
 'Mchuk',
 'African Realief Committe of Ku',
 'SCOTT',
 'D$L',
 'Mzungu',
 'Vi',
 'JLH CO LTD',
 'Msiki',
 'Namungo',
 'Nassor Fehed',
 'TWESA /Community',
 'DBFPE',
 'EF',
 'Serikali',
 'Mgaya Mwita',
 'Clause workers',
 'MLAKI  CO',
 'Busoga trust',
 'mzee mabena',
 'NORAD/',
 'BR',
 'local technitian',
 'Comunity',
 'Brad',
 'Tanganyika Basin',
 'MORNING CONSTRUCTION',
 'Healt',
 'Governme',
 'Roma',
 'KUMKUM',
 'PNR co',
 'Muslims',
 'Paffec',
 'Tansi',
 'CRAELIUS',
 'APM',
 'Zao water spring X',
 'TASA',
 'CSPD',
 'CIPRO/CARE',
 'DALDO',
 'VIFAF',
 'MTC',
 'TCRS Kibondo',
 'Howard and humfrey consultant',
 'RUDEP/',
 'LUNGWE',
 'Dhinu',
 'AIC KI',
 'Mataro',
 'FINI Water',
 'Mombo urban water',
 'REDAP',
 'Kagulo',
 'TMP',
 ...]
```

Again, we saw there was a 'Not known' value in 'installer', which we decided to use to replace nulls.

In [26]:

```python
df['installer'].fillna(value='Not known', inplace=True)
```

In [27]:

```python
df['installer'].isna().sum()
```

Out[27]:

```
0
```

We noticed significant overlap in values for 'funder' and 'installer'. From our business understanding, this made sense as organizations who were funding water pumps were also likely to be responsible for their installation.

In [28]:

```python
#check cases where the installer is not also the funder
df[df['funder'] != df['installer']]
```

Out[28]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...
1	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...
2	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...
3	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-	Zahanati Ya	0	

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...
4	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...
5	9944	20.0	2011-03-13	Mkinga Distric Coun	0	DWE	39.172796	-4.765587	Tajiri	0	...
...	...	...	...	...	...	...	...	...	...	...	...
59395	60739	10.0	2013-05-03	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	0	...
59396	27263	4700.0	2011-05-07	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	0	...
59397	37057	0.0	2011-04-11	Not Known	0	Not known	34.017087	-8.750434	Mashine	0	...
59398	31282	0.0	2011-03-08	Malec	0	Musa	35.861315	-6.378573	Mshoro	0	...
59399	26348	0.0	2011-03-23	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa	0	...

54481 rows × 42 columns

In [29]:

```
#investiating funder and installer relationship with World Bank as a specific value case
df[df['installer']=='World Bank']
```

Out[29]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
280	9474	250.0	2013-02-12	World Bank	1343	World Bank	30.017098	-4.332583	Jeshini	0	...	
304	43345	0.0	2012-10-23	World Bank	0	World Bank	33.430917	-4.389084	Shule Ya Msingi	0	...	
2071	28588	0.0	2012-10-23	World Bank	0	World Bank	33.436073	-4.421944	Mwanza Road	0	...	
5699	74077	0.0	2012-10-18	World Bank	0	World Bank	33.435268	-4.671744	Kwa Fupe	0	...	
6327	44441	0.0	2012-10-12	World Bank	0	World Bank	33.150261	-3.705625	Kwa Maraba	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
58348	54574	0.0	2012-10-23	World Bank	0	World Bank	33.442457	-4.426592	Kilabuni	0	...	
58434	54545	250.0	2013-02-12	World Bank	1311	World Bank	30.017351	-4.323850	Mission	0	...	
58681	6335	250.0	2013-02-12	World Bank	1306	World Bank	30.017716	-4.308701	Kwa Thomas	0	...	
58691	67819	0.0	2012-10-12	World Bank	0	World Bank	33.142605	-3.699442	Kwa Nyamizi Maswa	0	...	
58803	44356	0.0	2013-02-06	World Bank	1697	World Bank	29.784471	-4.445344	Kwa Esrom	0	...	

95 rows × 42 columns

In [30]:

```
df[(df['installer']=='World') & (df['funder']!='World Bank')]
```

Out[30]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
33726	63910	0.0	2011-04-02	Nethe	141	World	38.204463	-6.870355	Kwa Kiwele	0	...	

1 rows × 42 columns

In [31]:

```
df[df['installer'] == 'Not known']
```

Out[31]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	c
34	41583	0.0	2011-02-23	Not Known	-41	Not known	39.812912	-7.889986	Msikitini Wa Ijumaa	0	...	
35	57355	0.0	2013-03-28	Not Known	1546	Not known	36.618699	-3.293003	Sekondari	0	...	
43	19282	0.0	2013-01-15	Not Known	1642	Not known	34.967789	-4.628921	Mvae Primary	0	...	
47	13620	0.0	2011-07-27	Not Known	0	Not known	33.540607	-9.172905	Mahakamani	0	...	
65	51072	0.0	2013-02-09	Not Known	1415	Not known	34.621598	-5.173136	Nyambi	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
59357	46563	0.0	2013-02-19	Not Known	1635	Not known	34.971841	-5.098362	Shabani	0	...	
59366	55232	0.0	2013-02-02	Not Known	1541	Not known	34.765729	-5.027725	Joshoni	0	...	
59370	14796	200.0	2013-01-29	Not Known	1154	Not known	30.058731	-4.902633	Village Office	0	...	
59376	34716	0.0	2013-02-03	Not Known	1581	Not known	34.821039	-5.076258	Nasingo	0	...	
59397	37057	0.0	2011-04-11	Not Known	0	Not known	34.017087	-8.750434	Mashine	0	...	

3672 rows × 42 columns

In [32]:

```
df['installer'].value_counts(normalize=True)[:20]
```

Out[32]:

```
DWE                     0.292963
Not known               0.061818
Government              0.030724
RWE                     0.020303
Commu                   0.017845
DANIDA                  0.017677
KKKT                    0.015118
Hesawa                  0.014141
0                       0.013081
TCRS                    0.011902
Central government      0.010471
CES                     0.010269
```

```
Community               0.009310
DANID                   0.009293
District Council        0.009276
HESAWA                  0.009074
LGA                     0.006869
World vision            0.006869
WEDECO                  0.006684
TASAF                   0.006667
Name: installer, dtype: float64
```

As with the 'funder' column, we treated values of '0' as nulls and replaced them with 'Not known'.

In [33]:

```
df['installer'] = df['installer'].replace(to_replace={'0':'Not known'}, value=None)
```

In [34]:

```
df['installer'].value_counts(normalize=True)[:20]
```

Out[34]:

```
DWE                     0.292963
Not known               0.074899
Government              0.030724
RWE                     0.020303
Commu                   0.017845
DANIDA                  0.017677
KKKT                    0.015118
Hesawa                  0.014141
TCRS                    0.011902
Central government      0.010471
CES                     0.010269
Community               0.009310
DANID                   0.009293
District Council        0.009276
HESAWA                  0.009074
World vision            0.006869
LGA                     0.006869
WEDECO                  0.006684
TASAF                   0.006667
District council        0.006599
Name: installer, dtype: float64
```

In [35]:

```
df[df['installer'] == "0"]
```

Out[35]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	quality_group

0 rows × 42 columns

subvillage

In [36]:

```
#geographic location
df['subvillage'].isna().sum()
```

Out[36]:

```
371
```

In [37]:

```
df['subvillage'].value_counts()
```

Out[37]:

```
Madukani        508
Shuleni         506
Majengo         502
Kati            373
Mtakuja         262
               ...
Masaladi          1
Mgodi             1
Ipwasi            1
Tobo              1
Bombambili 2      1
Name: subvillage, Length: 19287, dtype: int64
```

There were 3,635 nulls in 'subvillage'. This column has over 19,287 unique values.

In [38]:

```
list(df['subvillage'].unique())
```

Out[38]:

```
['Mnyusi B',
 'Nyamara',
 'Majengo',
 'Mahakamani',
 'Kyanyamisa',
 'Moa/Mwereme',
 'Ishinabulandi',
 'Nyawishi Center',
 'Imalauduki',
 'Mkonomre',
 'Mizugo',
 'Ngondombwito',
 'Nkilifa',
 'Omarini',
 'Mwabasabi',
 'Tunzi',
 'Kidudumo',
 'Yeriko',
 'Center',
 'Manyanya',
 'Ibabachegu',
 'Mkanivega',
 'Mkonga Juu',
 'Msasa',
 'Kitereni',
 'Shuleni',
 'Chakahaya',
 'Kiyao',
 'Merali',
 'Karume',
 'Kudipera',
 'Mosheni',
 'Lupanga A',
 'Kilombero B',
 'Afya',
 'Ndanganyika',
 'Baura',
 'Mwanzala',
 'Nyabwai B',
 'Reli B',
 'Kilunduwe',
 'Sokoni',
 'Mwarufyu',
 'Marurani Juu',
 'Isenegeja',
 'Kachulu',
 'Mpandapanda',
 'Mlandege Juu',
 'Ikanga',
 'Msaranga Street',
 'Maporomoko',
 'Isimba',
 'Kagoye B',
 'Dhobi Street',
 'Msufini',
 'Soyekiutu',
 'Ants B',
 "Izimbya 'A'",
 'Nairobi',
 'Nkaloi',
 'Kiganza Centre',
 'Ulkusare',
 'Mzimba',
 'Mfumbu',
 'Mahalule',
 'Kihanga',
 'Iponda',
 'Kisoro',
 'Mtakuja',
 'Lembuka',
 'Mapinduzi',
 'Kalimungoma',
 'Umkituri',
 'Usita',
 'Komoro',
 'Kulasi Majengo',
 'Butimba',
 'Ikovo',
 'Dongo',
 'Moivaro',
 'Bushoma',
 'Shule',
 'Bulyahilu Center B',
 'Mlanda B',
 'Kasharunga',
 'Magwila',
 'Juhudi',
 'Kilundo',
 'Mwena',
 'Mara B',
 'Mushasha',
 'Kitobo',
 'Kishiha',
 'Misasi C',
 'Msewo Mwaazi',
 'Madukani A',
 'Kiruku Mchangamweupe',
 'Mashine',
 'Mlima Ndabaneze',
 'Nguvumali',
 'Tema',
 'Kwa Nyange',
 'Malula',
 'Maendeleo',
 'Kidete',
 'Ipuguso',
 'Mseseweni',
 'Kiwawa',
 'Nyansalala',
 'Vimetu',
 "Chang'Ombe",
 'Kalamila',
 'Mingo',
 'Kabarongo',
 'Mwebebonda',
 'Nendebe',
 'Ihela Shu',
 'Ikongora',
 'Majengo B',
 'Ewerendeke',
 'Busekele',
 'Kitega Uchumi',
 'Mosi',
 'Utengule',
 'Butondolo',
 'Isera',
 'Nyarutembo',
 'Msumbiji',
 'Wihanga',
 'Igalula A',
 'Mgwashi',
 'Ligelango',
 'Ipumpila',
 'Mkopwe',
 'Nyakafundikwa A',
 'Wichamoyo',
 'Mikoroshini',
 'Mbuyuni',
 'Nyakahunga',
 'Kipogoro',
 'Kikundi Kati',
 'Mtaa Wa Kitunda Kati',
 'Busiiko',
 'Usetule Kati',
 'Mpakani',
 'Mkanyageni',
 'Buhemba',
 'Usafwa',
 'Madrasa',
 'Madukani',
 'Ruhororo',
 'Kijiweni',
 'Kwa Karoli',
 'Nangumbu',
 'Mahongole A',
 'Mianga',
 'Kati',
 'Kagembe B',
 'Matela A',
 'Londoni',
 'Hospital',
 'Kagwila',
 'Ilula',
 "Ng'Uni",
 'Mwinuko',
 'Ichese',
 'Sekondari',
 'Katindiuka A',
 'Mbigili',
 'Malangilisho',
 'Migungani',
 'Mahaha',
 'Nzovu',
 'Kwegole',
 'Igavilo',
 'Kigamboni',
 'Uzunguni',
 'Jimbo Mjini',
 'Itimbwi',
 'Nyakalembe',
 'Nyalwela',
 'Kanisani',
 'Siuyu',
 'J',
 'I',
 'Chalinze Mzee A',
 'Sadani',
 'Mjini',
 'Tariso',
 'Kitakura',
 'Stooni',
 'Bulilwa',
 'Nyamanawa',
 'Azimio',
 'Manzese',
 'Mumigezi',
 'Bulibata',
 'Msia Kati',
 'Patandi',
 'Kudibona',
 'Qanqali',
 'Sosola',
 'Barabara 5',
 'Isanga',
 'Nasuro B',
 'Bagamoyo',
 'Kijijini',
 'Zobogo',
 'Olama',
 'Kawawa',
 'Ngujini',
 'Kumsenga',
 'Kidenge',
 'Mughanga',
 'Ishilanga',
 'Nyambazu',
 'Makwei Mvuleni',
 'Songambele',
 'Ccm',
 'Olobeshi',
 'Kachibijo',
 'Zahanati',
 'Kanga A',
 'Pongwe Kati',
 'Ofsini',
 'Kibehe A',
 'Abdujumbe',
 'Mtimbwani A',
 'Uyamba',
 'Ntungamo',
 'Madibila',
 'Madago',
 'Lole',
 'Kimogola',
 'Kilimanihewa',
 "Maring'A Juu",
 'Nyamatala',
 'Kinyunyu',
 'Nyandekwa',
 'Ujamaa',
 'Njia Panda A',
 'Kipundu Kati',
 'Bulifani',
 'Usanguni',
 'Chizomoche',
 'Uborwa',
 'Nyerer',
 'Magomeni',
 'Muungano',
 'Tazama',
 'Ilima',
 'Usega',
 'Ipenya',
 'Chembeli',
 'Kinani',
 'Dakusi',
 'Magaoni',
 'Mbale',
 'Fuzi',
 'Rulemba',
 'Kabigwa',
 'Bukurwa',
 'Tarakwa',
 'Mnazi Mmoja',
 'Nyalwashi',
 'Mringeni',
 'Mahenge',
 'King`Ombe',
 'Nkoni B',
 'Kauzeni',
 'Ilkirumuni',
 'Juu',
 'Kiwalaa',
 'Makulu',
 'Chiwindi',
 'Mwidea',
 'Iluma',
 'Bombambili 2',
 'Mundalu',
 'Nyanza',
 'Ijanija',
 'Lubele',
 'Barazani',
 'Imalabupina',
 'Mpuga',
 'Sirigadi',
 'Ukunda',
 'Godawni',
 'Magoye',
 'Kiimo',
 'Manoro',
 'Majengo Mapya',
 'Kange',
 'Chihano',
 'Miembeni',
 'Ikulu',
 'Mkundi',
 'Sunzula Madukani',
 'Pozamoyo',
 'Mtanga',
 'Katapulo',
 'Msimba B',
 'Likalo',
 'Muongozo',
 'Mitalula',
 'Nambigili',
 'Bugoro Asili',
 'Senani',
 'Mwenge',
 'Ilulu A',
 'Tazara',
 'Kanyezi Kati',
 'Itobanilo A',
 'Koloa',
 nan,
 'Lebusha',
 'Endasak',
 'Koniko B',
 "Mang'Ada",
 'Luwai',
 'Mungi Juu',
 'Kikondeni',
 'Nyati',
 'Sanze',
 'Ufantamie',
 'Mkombozi',
 'Kibaoni',
 'Miziro',
 'Itete',
 'Orro',
 'Mtarudi',
 'Ibonde',
 'Uswahilini',
 'Tabirugu',
 'Kidatu B',
 'Leganga',
 'Hospitali',
 'Taifa Road',
 'Kituma',
 'Ndundu',
 'Mfalanyaki',
 'Nyampakupwani',
 'Misufini',
 'Bwawani B',
 'Nyasa',
 'Kariwa Kaskazini',
 'Sabato',
 'Ifungira',
 'Lulanga',
 'Jangwani A',
 'Kigangama',
 'Kiwanja',
 'Pulugumwa',
 'Chulasitu',
 'Mtaa Wa Mji Mpya',
 'Lyasongoro',
 'Mheta',
 'M',
 'Kilimahewa',
 'Msikitini',
 'Sautimoja',
 'Nazareti',
 'Kasama',
 "Matale Ng'O",
 'Makungu',
 'Mwamigundui',
 'Ilangi',
 'Doma B',
 'Napulu',
 'Katambo',
 'Miogoni',
 'Mikaragata',
 'Kiverenge',
 'Luchili',
 'Mtaa Wa Kivule',
 'Kilosa Juu',
 'Karamsingi',
 'Naibile',
 'Kidimu',
 'Mkwajuni',
 'Kilimbili',
 'Tawingo',
 'Mlongo',
 'Mtaa Yangeyange',
 'Muyuni',
 'Iligiti',
 'Senta',
 'Mbugani',
 'Rau Ya Kati',
 'Lekako',
 'Uzogore',
 'Kimbogo',
 'Musibuka',
 'Mapambano',
 'Mpuje',
 'General Tyre',
 'Kalihanya',
 'Mkwanyule',
 'Uzunguni B',
 'Mwabadimi',
 'Msunuzi',
 'Landani',
 'Madina',
 'Saza Kati',
 'Mvinila',
 'Maziwa',
 'Mwaniriri',
 'Kapolo',
 'Uchiliwala',
 'Chomboko',
 'Butwale',
 'Gombelesa',
 'Fukayosi',
 'Kikomero',
 'Ruvumela',
 'Simbalo',
 'Munge',
 'Makungani',
 'Ukombozi',
 'Mtimui',
 'Mtapenda C',
 'Ilobashi',
 'Mikindani',
 'Nyankende',
 'Kingwande',
 'Ngwramu',
 'Mlembea',
 'Iyumbu',
 'Larkaria',
 'Ikongoigare',
 'Jumuiya',
 'Sakei',
 'Sangatini',
 'Bondeni',
 'Magengeni',
 'Matadi A',
 'Mchangani',
 'Mlale B',
 'Ulaukya',
 'Mangu',
 'Sisikwasisi',
 'S',
 'Keko',
 'Chambeo',
 'Iyembela B',
 'Ndemanilwa',
 'Jihu A',
 'Mfuruwashe',
 'Migombani',
 'Masaini',
 'Tyula',
 'Sabore',
 'Mwinyi',
 'Kiduguda A',
 'Bunukangoma',
 'Sabasaba',
 'Muvwa',
 'M/Kati',
 'Muraiweni',
 'Igumija',
 'Chavakaa',
 'Maanga',
 'Nkungulu',
 'Ulinzi',
 'Kolandoto',
 'Magomeni B',
 'Kwemkangala',
 'Lumumba',
 'Nyerere',
 'Buganda',
 'Chiraga',
 'Gengeni',
 'Gua D',
 'Kariakoo',
 'Damaygwa',
 'Mikongeni',
 'Matwalani',
 'Matangini A',
 'Miyomboni',
 'Matamba Juu',
 'Ulete',
 'Utsewa',
 'Onya',
 'Getamoki',
 "Makugulu 'B'",
 'Matunda',
 'Mtoghoo',
 'Mkawaganga',
 'Ndaushei',
 'Lukula',
 'Mbwawa Shule',
 'Chabura',
 'Lituta',
 'Unyanyembe',
 'Sotele B',
 'Ngelura',
 'Kaseni',
 'Madamba',
 'Midibwi',
 'Mtaa Wa Kichangani',
 'Mpanda',
 'Ukiwayuyu',
 'Tankini',
 'Itiyeja',
 'Busulwa',
 'Bitale A',
 'Vikuge',
 'Matarau',
 'Malimka',
 'Mjimwema',
 'Kibururu B',
 'Magagai 2',
 'Sole',
 'Kisutu',
 'Gumba',
 'Iteka',
 'Sanga',
 'Makingi A',
 'Katandala',
 'Kizerui',
 'Mkombola',
 'Jamnono',
 'Nyamizoka',
 'Kwiriba',
 'Mgaraganza',
 'Sokomoko',
 'Mondelo',
 'Mifugo',
 'Lyandu',
 'Wangama',
 'Minyinga',
 'Nyanhiga',
 'Bukene Mjini',
 'Itekesha',
 'Chimbuko',
 'Msanzi Kati',
 'Tobora',
 'Nanyala Kati',
 'Majenje',
 'Itubula',
 'Udushi B',
 'Selemembe',
 'Dodoma',
 'Kipangule',
 'Kwevumo',
 'Nhundya',
 'Sambala',
 'Mwanakibwengo',
 'Mwabambasi',
 'Mzingezinge',
 'Mgulu Wa Ndege',
 'Siboti',
 'Masuguru Shule',
 'Jambe',
 'Bukuba A',
 'Godown',
 'Kouri',
 'Dege',
 'Mashariki',
 'Uwasi',
 'Sahoni',
 'Okaseni Chini',
 'Baukani B',
 'Mwerera',
 'Mwanakalenge',
 'Makwale Ofisin',
 'Ghana',
 'Kidai Pwita',
 'Kibengele',
 'Ihwa',
 'Buyogwa',
 'Nyambemba Juu',
 'Ngulumbi',
 'Mangara',
 'Mnazimmoja',
 'Linyare',
 'Kidatu A',
 'Kilimani',
 'Parokiani',
 'Kwa Philipo',
 'Bunogwz',
 'Mwamuze',
 'Nanga Kati',
 "Kwemng'Weng'We",
 'Mchombe A',
 'Mpui',
 'Misri',
 'Ikulungilo',
 'Minazi Mikinda',
 'Gulioni',
 'Malamba',
 'Nyasubi',
 'Kabale',
 'Maleta',
 'Mwandu',
 'Runyogoza',
 'Kurui A',
 'Nyamafurila',
 'Budekwa Kati',
 'Mishale',
 'Maswele',
 'Lyamalagwa',
 'Masange',
 'Gubali',
 'Katente Namba Moja',
 'Ofisini',
 'Magatini',
 'Sange Kijijini',
 'Mwabalimi',
 'Utegule',
 'Ikiligano',
 'Bunonga',
 'Terema',
 'Olomitu',
 'Luhovelo',
 'Inbumba',
 'Mbulani',
 'Kakinga',
 'Malwilo',
 'Matoroka',
 'Songwe Mjini',
 'Kwemianga',
 'Mhunda',
 'Sechuni',
 'Pala',
 'Mwamala',
 'Nachiungo',
 'Mria',
 'Mwamabu B',
 'Kichangani',
 'Miteja',
 'Wichamike',
 'Jilangamili',
 'Uyole',
 'Sabasaba Street',
 'National Park',
 'Galani',
 'Mkanawalo',
 'Nzonze',
 'Gibise',
 'Kyembo',
 'Mwemberadu',
 'Nyashimba',
 'Mwamakumbi A',
 'Dodoma B',
 'Noho',
 'Chibombo',
 'Isundambwa',
 'Kongei',
 'Ulindiwa',
 'Kakola',
 'Ilembo Kati',
 'Mbae Mashariki',
 'Nyamagala',
 'Gezaulole',
 'Maembe',
 'Miwaleni',
 'Kikukuru',
 'Garbapi',
 'Nyashimo',
 'Dihimba',
 'Imbambasi',
 'Mkumba',
 'Harsha Kubwa',
 'Mijelejele',
 'Igumangobo',
 'Idetemya',
 'Msumbiji A',
 'Bangwe',
 'Matare',
 'Kisiwani',
 'Kanamalenga',
 'Kota',
 'Kilimia',
 'Zaburi',
 'Motomoto',
 "Kayanga 'A'",
 'Lihami',
 'Liwoyola',
 'Mwanona',
 'Relini',
 'Wala',
 'Mkora',
 'Magungu',
 'Legezamwendo',
 'Misheni',
 "Mkorin'Ga",
 'Mkupuka',
 'Welezo',
 'Kenyana B',
 'Tuyombo',
 'Sabasita',
 'Stahabu Azimio',
 'Arauyo',
 'Kusini',
 'Mdeme',
 'Kiluluma',
 'Kisimani',
 'Mtaa Wa Kimwani',
 'Zagana',
 'Mumasiga',
 'Mtongani',
 'Nakwene',
 'Kahunda 3',
 'Tulieni',
 'Kilolero',
 "Nyamateng'Enga",
 'Mnyambe',
 'Mabuki',
 'Minazini',
 'Mgubulile',
 'Mkoroshini',
 'Kidimwi',
 'Loko',
 'Ikula',
 'Shinga Juu',
 'Omundeba',
 'Kihesa',
 'Kiondeni',
 'Kigangama A',
 'Ikungi',
 'Isanzike',
 'Mbangamao',
 'Kifaru',
 'Kilida Kona',
 'Kisambare',
 'Mseko',
 'Maringo',
 'Ipogolo',
 'Nyaniga',
 'Ungindoni',
 'Mbugani A Mashariki',
 'Chakitalagu B',
 'Togo',
 'Oloshonokie',
 'Bumburyahe',
 'Mpuguso',
 'Mpiji',
 'Anna Abdalla',
 'Mivumoni',
 'Mnyakatu',
 'Nyarutunga',
 'Bwawani',
 'Mtaa Wa Mzinga',
 'Vigulu',
 'Ibale',
 'Rutenge B',
 'Barabara',
 'Mwabashinda',
 'Msikiti',
 "Gitianga 'B'",
 'Vugwama Kati',
 'Mwandege',
 'Shugului',
 'Kishale',
 'Buyagaa',
 'Tungamaa',
 'Kinepa',
 'Nyakabanga',
 'Mugongo',
 'Kasange Kati',
 'Manyire',
 'Mendai',
 'Mtukula',
 'Ngoingo',
 'Iponyabukoli',
 'Estate',
 'Kisesa',
 'Ilala',
 'Nkuninkana',
 'Lwanilo',
 'Sole 3',
 'Mwankali',
 "Ng'Ambo",
 'Kigero',
 'Nkwambati',
 'Olorieni',
 'Masanko',
 'Ibanza',
 'Jimboni',
 "Mng'Ende",
 'Isengo A',
 'Kwazuberi',
 'Bwemengo',
 'Mwitikio',
 'Ukinga',
 'Nyange',
 'Tengeru',
 'Bwizo',
 'Mjini A',
 'Kibagwe',
 'Kimwangoko B',
 'Kerenero',
 'Tuwemacho',
 'Kamambande',
 'Nyigumba',
 'Vigoda',
 'K',
 'Mwabukikya',
 'Kishoju 1',
 'Mbwawa Mkoleni',
 'Majani Mapana',
 'Maseleka',
 'Ihanda',
 'Ichimbo',
 'Mukasesero',
 'Isupilo',
 'Matepwende',
 'Mwabadimu',
 'Nyasura A',
 'Nyamamba',
 'Likere',
 'Qanya',
 'Mlimani',
 'Shabaha',
 'Usafa',
 'Songosongo',
 'Gomapembe',
 'Mtunge',
 'Basutu Ndogo',
 'Jaira',
 'Kikonya',
 'Busubi',
 'Mwamagaka',
 'Mboguiyola',
 'Uramboni',
 'Izunya',
 'Ngudulugulu',
 'Kitonto',
 'Kiwanda',
 'Maweni',
 'K/Center',
 'Sambandole',
 'Birubiru',
 'Puslukunya',
 'Kakukuru',
 'Iyogelo',
 'Irangi',
 'Mawangala',
 'Kagembe A',
 'Wandeni',
 'Mkameni',
 'Ndandalo Mpaka',
 'Msolwa Kati',
 'Kwigena',
 'Kakumbo',
 'Wigelekelo Mashariki',
 'Kitabaga',
 'Makorongoni',
 'Kitakata',
 'Mwamasasi',
 'Ikenywa',
 'Alagwa',
 'Ngereka B',
 'Mwaike',
 'Simbani',
 'Kamana',
 'Mwatandai',
 'Baraza A',
 'Imalangi',
 'Sarama',
 'Tuleane',
 'Changalawe',
 'Luguma',
 'Mampando A',
 'Kinela',
 'Makaoni',
 'Lutumbi',
 'Msumbi',
 'Mohogoigwa',
 'Igongolo Kati',
 'Sokoine',
 'Lugwashi',
 'Mbuguru',
 'Mwalo',
 "Munmunang'I",
 'Kondiki',
 'Mapogoro B',
 'Manyete',
 'Nyakinyo',
 'Kinondo',
 'Maloya',
 'Mabashura',
 'Malolo',
 'Kagarama',
 'Mungonya',
 'Mkyashi Kati',
 'Kotela',
 'Bariadi',
 'Kamena',
 'Madale Kati',
 'Ipenza',
 'Nandutu',
 'Ntanganyika',
 'Manjore',
 'Lungerengere',
 'Shighati',
 'Mkese',
 'Ukondamoyo',
 'Arombe',
 'Kilanga',
 'Mkutano',
 'Badagi',
 'Nyamateke',
 'Iwawa',
 "Mfuma Ng'Ombe",
 'Kitopeni',
 'Nyamikalango',
 'Njiwa Kati',
 'Mchomoro',
 'Headquarters',
 'Longululu',
 'Monederer',
 'Hengeni',
 'Nambunju',
 'Mtaa Wa Kigezi',
 'Itunda A',
 'Magwira',
 'Usalama',
 'Kagulembela',
 'Msfuriko C',
 'Stand B',
 'Mumirama',
 'Mwabayanda',
 'Ngalu',
 'Lusind',
 'Nkaya',
 'Kiga',
 'Zawiani',
 'Pemba',
 'Lugala',
 'Tumaini',
 'Kitichi',
 'Chumo B',
 'Mkabogo',
 'Marumbo',
 'Kabagunda',
 'Isembuka',
 'Njiapanda',
 'Njoro',
 'Mahata',
 'Chabula',
 'Mrau',
 'Mwakasumve',
 'Igumo Kati',
 'Malungo',
 'Nyakatakala B',
 'Mwanghalanga B',
 'Nyabirere',
 'Makongoga',
 'Rusesa',
 'Bukumbulwa',
 'Endashagwe',
 'Kikaranga B',
 'Mbuga',
 'Ibambula',
 'Balanga',
 'Migera',
 'Ilongovoto',
 'Elimu',
 'Mtola',
 'Buyumba',
 'Uhuru B',
 'Mwabakali',
 'Mwambalizi',
 'Kimawe',
 'Bwiti',
 'Station',
 'Guanaya',
 'Mhandu',
 'Uhuru',
 'Mayadi',
 'Imalambeo',
 'Tanya',
 'Kitunda',
 'Mwasubi',
 'Tundu La Leo',
 'Mkuya',
 'Bwelu A',
 'Kitundu',
 'Yaniko B',
 'Sandege',
 'Kituntu A',
 'Idwele',
 'Omukigando',
 'Mbalizi',
 'Ilelamhina Shuleni',
 'Kaija',
 'Busabaga',
 'Kizinga A',
 'Ghalani',
 'Lyabasura',
 'Uparo',
 'Kabasonge',
 'Wiligwamabu',
 'Sinai',
 'Mihogoni',
 'Gongoti',
 'Mwegerezi',
 'Mawanda',
 ...]
```

In [39]:

```
df[df['subvillage'] == 'Unknown']
```

Out[39]:

id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	quality_group

0 rows × 42 columns

Unlike our previous cases, there wasn't a clear value to change nulls to. We decided to create a new 'Unknown' value to replace nulls rather than dropping these rows.

In [40]:

```
df['subvillage'].fillna(value='Unknown', inplace=True)
```

In [41]:

```
df['subvillage'].isna().sum()
```

Out[41]:

```
0
```

public_meeting

In [42]:

```
df['public_meeting'].isna().sum()
```

Out[42]:

```
3334
```

In [43]:

```
df['public_meeting'].value_counts()
```

Out[43]:

```
True      51011
False      5055
Name: public_meeting, dtype: int64
```

There were 3,334 nulls in 'public_meeting'. Given the large imbalance between True and False value counts, we decided that nulls should match the majority class, True.

In [44]:

```
df['public_meeting'].fillna(value=True, inplace=True)
```

In [45]:

```
df['public_meeting'].isna().sum()
```

Out[45]:

```
0
```

scheme_management

In [46]:

```
#who operates the waterpoint (organization/category)
df['scheme_management'].isna().sum()
```

Out[46]:

```
3877
```

There were 3,877 nulls in 'scheme_management'.

In [47]:

```
df['scheme_management'].value_counts()
```

Out[47]:

```
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
```

In [48]:

```
#investigating the one value_count of None
df[df['scheme_management'] == "None"]
```

Out[48]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qu
23603	23849	50.0	2013-03-18	Not Known	-11	Not known	39.431194	-7.100783	Kwa Nyamtawa	0	...	

1 rows × 42 columns

In [49]:

```
df[df['funder'] == '0']
```

Out[49]:

id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	quality_group

0 rows × 42 columns

Since there was already an instance of 'None' in this column, we changed all nulls to 'None'.

In [50]:

```
df['scheme_management'].fillna(value='None', inplace=True)
```

In [51]:

```
df['scheme_management'].isna().sum()
```

Out[51]:

```
0
```

scheme_name

This column had a large number of nulls: 28,166. Since the values appeared to be individual names with great diversity, we decided this column was unusuable for our modeling. However, we didn't drop it from the dataframe since our modeling process uses a pipeline for this purpose.

In [52]:

```
#who operates the waterpoint
#how is this information different from scheme_management?
df['scheme_name'].value_counts()
```

Out[52]:

```
K                          682
None                       644
Borehole                   546
Chalinze wate              405
M                          400
                          ... 
Nkwe                         1
BUWADA                       1
BL Bonifas Kingu             1
BL Losei                     1
DMK:Anglican church          1
Name: scheme_name, Length: 2696, dtype: int64
```

In [53]:

```
df['scheme_name'].isna().sum()
```

Out[53]:

```
28166
```

In [54]:

```
unique_scheme = list(df['scheme_name'].unique())
print(len(unique_scheme))
unique_scheme
```

```
2697
```

Out[54]:

```
['Roman',
 nan,
 'Nyumba ya mungu pipe scheme',
 'Zingibali',
 'BL Bondeni',
 'None',
 "wanging'ombe water supply s",
 'Makanj',
 'Kidabu',
 'Mashangwi',
 'Quick wins Program',
 'Komaka mandaka',
 'Sobodo Borehole Scheme',
 'Sobodo Borehole Scheme',
 'Kitukuni water supply',
 'BL Mwakikoti',
 'Chalinze wate',
 'Mae pipeline',
 'UNDP',
 'Ngana water supplied scheme',
 'Itun',
 'Bomala',
 'Kirua kahe pumping water trust',
 'Misiwa',
 'Mtwango water supplied sche',
 'K',
 'Machumba estate pipe line',
 'SHIMASA',
 'wangama water supply scheme',
 'Shirimatunda Water Supply',
 'Laela group water Supp',
 'Makwale water supplied sche',
 "Nyang'hwale",
 'Njoro Water Supply',
 'Kirua kahe gravity water supply trust',
 'Olgilai pipe line',
 'Mabula mountains spr',
 'Mkongoro One',
 'Maambreni gravity water supply',
 'Mwando water supply',
 'M',
 'Kaisho/Isingiro w',
 "wanging'ombe supply scheme",
 'Tove Mtwango gravity Scheme',
 'Tengeru gravity water supply',
 'Kulasi water supply',
 'Malemb',
 'Tawa',
 'Loruvani gravity water supply',
 'HESAWA',
 'B',
 'Gawa',
 'Chankele/Bubango water project',
 'Bagamoyo wate',
 'Kijiji',
 'Donge',
 'Mradi wa maji wa mpitimbi',
 "Mwang'hosha Nyamalogo",
 'Shallow well',
 'Mbati Water Supply',
 'Losaa-Kia water supply',
 'Kwa Nyange water supply',
 'Kiwele',
 'N',
 'Tuvaila gravity water supply',
 'Bumi',
 'Kan',
 'TASAF',
 'Ihela',
 'Kabindi Water Supply',
 'Shagai streem',
 'Distri',
 'LIPS Borehole Scheme',
 'Kanga water supplied scheme',
 'Mkam',
 'Vugiro',
 'Kiro',
 'Government Borehole Scheme',
 'Maji ya kutega',
 'Londoni water supply',
 'Njalamatata water gravity scheme',
 'Mtowisa water suply',
 'Msanzi Water Supply Sc',
 'Amani spring',
 'Meseke',
 'Kisimiri gravity water supply',
 'Kisimiri gravity water supply',
 'S',
 'Hinju water supply',
 'Hempanga water supply',
 'EKTM 3 water supply',
 'Kasahunga pipe scheme',
 'Kashanda spring source',
 'Matuli/Mdaula',
 'Vyama vya watumia maji',
 'Sinyanga water supplied sch',
 'Coffee curing-Kahe pipeline',
 'Olkokola pipe line',
 'Kilotweni water supply',
 'Kabali',
 'Kazilankanda Water Supply',
 'MWS',
 'Dimamba',
 'Ma',
 'Luga',
 'Borehole',
 'Mahuni',
 'Jihoro',
 'Ufinga river',
 'FW',
 'Mtiro pipeline',
 'Ipepo',
 'Chikuyu water supply',
 'Chiz',
 'Mkon',
 'Mradi wa maji wa sikonge',
 'Kigaga gravity water supply',
 'DANIDA',
 'Mtimbira',
 'Deep well',
 'Kindoroko water supply',
 'Kabuye',
 'Tungu water piped scheme',
 'Saitero olosaita pipe line',
 'Nyanza water project',
 'Otaruni water supply',
 'Kimanda',
 'Ilesi gravity water supply',
 'Mowasu',
 'Mbokomu east',
 'Nzas',
 'World Bank',
 'Libango water use group scheme',
 'Kalangalala',
 'Mkongoro Two',
 'Government',
 'Mkoy',
 'CDTF',
 'Baba',
 'imalilo water supply scheme',
 'Isongo w',
 'Saga',
 'Nabaiye pipe line',
 'Una mkolowoni',
 "Nang'awanga water supply",
 'Mirerani pipe scheme',
 'Kanenge',
 'Tanzania flowers pipe line',
 'Tingatinga Ngerayani water',
 'Mlan',
 'Ukange',
 'Community',
 'Kyonza',
 'Ikela Wa',
 'Chal',
 'Pand',
 'Kifaru water Supply',
 'Huru mawela water project',
 'Kayiwasu',
 'Raviwasa',
 'Mradi wa maji wa kilagano',
 'Olumulo pipe line',
 'Jongoj',
 'mtwango water supply scheme',
 'Marangu baraza',
 'Marieni Makanya water supply',
 'Nyo',
 'Mkunya',
 'Sinyanga  water supplied sc',
 'kaleng',
 'Nzi',
 'Ruwini water supply',
 'CDG',
 'Uchira water users association',
 'TM part Three',
 'Sokeni pipeline',
 'Lyamungo umbwe water supply',
 'Mamire water supply',
 'Coffee-curing pipeline',
 'Mwaya Mn',
 'Kit',
 'Mtanga water supply',
 'Kidahwe water project',
 'Mwamanota water piped scheme',
 'Fukayosi Wate',
 'Kasurua water supply',
 'Munge water scheme',
 'Majimingi',
 'REDESO',
 'Shagayo forest',
 'Ngabav',
 'IKTM 2 water supply',
 'Olchoronyokye water projec',
 'Nabai pipe line',
 'Mtit',
 'Mt',
 'BL Matadi A',
 'Kitend',
 'Mchangani',
 'Namwinyu Water Supply',
 'matembwe water supply schem',
 'Jihu piped water Scheme',
 'Kandika water supply',
 'Mradi wa maji Komuge',
 'Nduruma pipe line',
 'Chovora',
 'Jaira water supply',
 'TASSAF',
 'Kibohelo  forest',
 "Uroki-Bomang'ombe water sup",
 'Lima gravity water scheme',
 'Mangamba forest',
 'World Bank Water Project',
 'Migoli',
 'Tangeni',
 'Mlimba W',
 'Water from DAWASCO',
 'Muwimb',
 'Tandal',
 'Onya water  supply',
 'Nyazwa',
 'mbigil',
 'Mradi wa maji wa mahanje',
 'Hakwe water supply',
 'Ninga hydram water scheme',
 'Hemb',
 'Upper Ruvu',
 'Mradi wa maji wa ntend',
 'Hingilili',
 'Chig',
 'Iw',
 'Saseni',
 'Saseni',
 'RWSSP',
 'Mgaraganza water project',
 'Msanzi water Supply',
 'Handeni Trunk Main(H',
 'Mazinde ngua water',
 'Kwa',
 'Mtandao wa Mabomba',
 'Kwagwisembeza',
 'Nkenja',
 'Likawage Water supply',
 'Nakombo',
 'Nyafisi',
 'Nyanga/Kalege',
 'Mandawa',
 'Chanjare water supply',
 'Sabodo Borehole Scheme',
 'Maambreni gravity water supply breni',
 'Sakale water supply',
 'Kurui water supply',
 'I',
 'Masa',
 'Foki',
 'Lupali',
 'Lake Victoria basin',
 'Songota pipe line',
 'Mfumbi',
 'KATORO PUMPING SCHEME',
 'Rain water',
 'Nandembo Water Supply',
 'Mapinduzi',
 'Vulue water supply',
 'Huru materuni water supply',
 'Mwigumbi piped scheme',
 'Machame water supply',
 'Nduguti pipes water supply',
 'Milola Water Scheme',
 'Mpwa',
 'Rwamgurusi water',
 'Kongei water supply',
 'Mradi wa maji wa kakola',
 'Mbae',
 'Dihimba water supply',
 'MANGISA',
 'It',
 'Maweni water supply',
 'Shengui forest',
 'Mwanona',
 'Mtum',
 'JAICA Borehole Scheme',
 'Fark',
 'Ilawa',
 'Kisanja water supply',
 'ILOL',
 'Maji Coast(Ruvu)',
 'World banks',
 'AIC',
 'RW',
 'Mradi wa maji Shirati',
 'Mitema',
 'Ilolo',
 'Cham',
 'Mvuh',
 'Magati gravity water',
 "Nyang'",
 'Matund',
 'MAKOGA WATER SUPPLY',
 'BUUDER',
 'Mto China',
 'Kibola gravity water supply',
 'Mseko water supply',
 'Kiboelo forest',
 'Naisinyia pipe scheme',
 'Naisinyia pipe scheme',
 'Nya',
 'Kand',
 'Ipululu water supply',
 'Irole',
 'NCHULOWAIBALE WATER SUPPLY SCHEME',
 'Ru',
 'ADRA',
 'Ilente streem',
 'Bagamoyo Wate',
 'TM part Four',
 'RUMWAMCH',
 'Magang',
 'Luchelengwa',
 'Mang`ula',
 'upper Ruvu',
 'Losaa Kia water supply',
 'G',
 'BRUDER',
 'Kiumba water supply',
 'Inyonga water supply',
 'manyunyu water supply schem',
 'Mgandazi',
 'Tanesco water supply',
 'Kaviwaso',
 'Marera-Lole pipeline',
 'Songosongo Water supply',
 'Jidowaso',
 'Boza water supply',
 'Njia panda Piped water Scheme',
 'Katoro',
 'Seje',
 'Mbuo mkunwa water supply',
 'Upper Ruvu Ba',
 'Kishoju Water  Su',
 'Mkomazi water supply',
 'Centra',
 'Njog',
 'Kagongo water project',
 'Buku',
 'Mtobo',
 'Hesawa',
 'Twendembele Water Supply',
 'Anglic',
 'Maga',
 'Ga',
 'Igongolo gravity water sche',
 'Ngulu water supply',
 'Matai group water Supp',
 'Mradi wa maji wa businde',
 'LIPS Water Scheme',
 'Mradi wa maji wa peramiho',
 'lugalo',
 'Kiverenge water supply',
 'Mradi wa maji wa Kipanga',
 'Mradi wa maji vijijini',
 'Mkabenga spring source',
 'itulahumba water supply sch',
 'Marangu west',
 'Itete wa',
 'Ngelen',
 'Kashutakasimba',
 'Monduli pipe line',
 'OMBASI',
 'Lipumburu Water',
 'Mzizima Water',
 'Maleng',
 'Mkangoru spring source',
 'Maya',
 'Magoma water supply',
 'Chun',
 'Ikuna gravity water project',
 'Old keni water supply',
 'Old Keni water supply',
 'Isanga',
 'Nyachenda',
 'Busi',
 'Mpute',
 'Lembeni water supply',
 'none',
 'Michenga',
 'Mkalamo water supply',
 'Water board',
 'Mwande',
 'Kizi',
 'Mad',
 'Chongera water su',
 'Central basin',
 'Ikonda',
 'Kisaika  pipeline',
 'Magula mountains spr',
 'Mroroma',
 'Mbwinji/MWS',
 'Kirachi water supply',
 'Robanda pumping scheme',
 'Rain harvest',
 'Luholo',
 'Igoj',
 'Mamsera water supply',
 'Libango water scheme',
 'Baga 2 streem',
 'Upuge water supply',
 'Nyamabale spring source',
 'Lemanyata pipe line',
 'Hogo',
 'Nagoma water supply',
 'Mahida mawanda water supply',
 'Kamatendeli spring source',
 'Nselembwe water supply',
 'Kimara Water Supply',
 'Nala',
 'Tove mtwango',
 'Kwawameku water supply',
 'Itigi water supply',
 'Kiremeta water supply',
 'Mkom',
 'Luwumb',
 'TPRI pipe line',
 'Upami gravity water scheme',
 'BL Kirishi',
 'Mnyawi water supply',
 'Kilimatinde water supply',
 'Mradi wa maji wa gungu',
 'Kitunguli water supply',
 'Kirwa Keni water supply',
 'Mtan',
 'Likamba mindeu pipe line',
 'Kib',
 'Hedaru kati water supply',
 'Kihinda water sup',
 'Matai group  Water Sup',
 'Maramba gravity spri',
 'Utende water supply',
 'Igalula water supply',
 'Maho',
 'Rika',
 'Iduo',
 'BL Nkini',
 'Olkokola mwandet pipe line',
 'Kimasaki gravity water supply',
 'imalinyi water supply schem',
 'Mkongoro one',
 'Sofi Maj',
 'Murubila spring source',
 'Nkwenda water sup',
 'Igosi',
 'Igosi',
 'Mradi wa maji wa senga',
 'Lake Victoria pipe scheme',
 'Manu',
 'Onya water supply',
 'Handeni water supply',
 'Chif',
 'Mradi wa maji wa matimila',
 'Chambogo forest',
 'OLD RUBALE WATER SUPPLY SCHEME',
 'Gamowaso',
 'Nyamitoko  water',
 'Mradi wa maji wa Kiloleli',
 'AIC kahunda',
 'Kabagendera water',
 'Kole',
 'Ikete',
 'Mrao water supply',
 'Holili water supply',
 'U',
 'Shirimatunda water Supply',
 'Rofa',
 'Murinyina spring source',
 'Lufuo',
 'Mnerongongo',
 'Kifaru/kituri water supply',
 'Mradi wa maji wa pito',
 'Ngwarwa water scheme',
 'Maliwa',
 'Mnyuzi water supply',
 'MONGAHAI RIVER',
 'Makamba kwa lukonge',
 'Iyen',
 'isoliwaya water supply sche',
 'Mazinde water supply',
 'Tuta',
 'Ihowanja',
 'LENCH',
 'Orumekeke water scheme',
 'Timbolo sambasha TPRI pipe line',
 'Kasangezi',
 'Lendanai pipe scheme',
 'Ugabwa',
 'Ekenywa pipe scheme',
 'Mindutulieni',
 'Lyamungo-Umbwe water supply',
 "Seela Sing'isi gravity water supply",
 'Doroto water supply',
 'Ki',
 'Njengwa water supply',
 'Wasa',
 'Kwevumo streem',
 'Kinyinya gravity water supply',
 'Kise',
 'Idegen',
 'Mivumoni borehole',
 'Uhekule',
 'Matekwe',
 'Mbunge',
 'Yongoma',
 'Nzug',
 'Kiamachini water supply',
 'RURAL WATER SUPPLY',
 'Mafi Mountains',
 'Marambo',
 'Chip',
 'Sali Wat',
 'Nzihi',
 'Chawi water supply',
 'Makidi water supply',
 'W',
 'Mradi wa mkombozi',
 'Tyeme water supply',
 'Iyeme water supply',
 'Maswa Water supply program',
 'no scheme',
 'Dindimo Water Supply',
 'Ifunda',
 'Mradi wa maji wa kanyenye',
 'Marua msahatie water supply',
 'Mradi wa maji wa litisha',
 'Idodi',
 'Ngamanga water supplied sch',
 'Mkutimango water supply',
 'Kitere water supply',
 'Sanje Wa',
 'Nyabibuye gravity water supply',
 'Nkunga',
 'Mkunya/MWS',
 'Uso',
 'ADP Simbo',
 'Nyakasanda gravity water supply',
 'Mshewa Water Supply',
 'MMILUKI',
 'AUWASA pipe scheme',
 'Nyaruyoba/Kasaka gravity water supply',
 'Nkuuny gravity water supply',
 'Mombo water supply',
 'shallow well',
 'WSDP',
 'Ihum',
 'Nkul',
 'Maja',
 'Kidaba',
 'Mradi wa maji Nyanduga',
 'Morongo',
 'Masaseni water supply',
 'Majonanga',
 'BUWASA',
 'Nasula gravity water supply',
 'Lake Victoria',
 'Tamp',
 'Janda',
 'Bulong',
 'Maland',
 'Mradi wa maji matendo',
 'Kw',
 'L',
 'Matamb',
 'Namahimba Water gravity scheme',
 'Mkuzu forest',
 'Mradi wa maji wa Ipole',
 'Kitwechembogo Water Supply',
 'Kibohelo streem',
 'Msolwa U',
 'Luko',
 'Kirwa  water supply',
 'Tove',
 'Mananga-himo pipeline',
 'Ilindi',
 'Endawasu',
 'Kuri',
 'Pahi',
 'Ikuy',
 'Maun',
 'Dingidingi water supply',
 'Kakonko /Mbizi gravity water supply',
 'Moronga',
 'Zepalama',
 'Jumuiya ya watumia maji kilema kusini',
 'Mradi wa maji wa kabila',
 'Murufiti',
 'Kong`hoi',
 'Malemeu gravity water supply',
 'Sasani',
 'Mlomboza forest',
 'Mlomboza forest',
 'Akheri gravity water supply',
 'Naroko pipe line',
 'Uru shimbwe',
 'Oljoro Namba 5 pipe scheme',
 'Nyamtukuza',
 'Mradi wa Maji mwanzi',
 'Mi',
 'Maboga',
 'Nyamugarika',
 'Nyangao Water Supply',
 'BL Kashashi',
 'BL Tindigani',
 'Mv',
 'Tank refu Mtakuja',
 'Kichananga gravity water supply',
 'Malama',
 'Ukwama',
 'Olkimo water project',
 'Finwater Supply',
 'Wangingombe gravity Scheme',
 'LVEMP',
 'Nansio Water Supply',
 'Rau kariwa karikachi',
 'Kwamalima water supply',
 'Moniki water scheme',
 'Kinyik',
 'Bangata water project',
 'Murutunguru Water Supply',
 'Mradi wa maji wa kasimbu',
 'Vianzi Water Supply',
 'Kidege forest',
 'Shaba water supply',
 'Ipuli pipes water supply',
 'Heivu water supply',
 'Kidia kilemapunda',
 'water supply at Nyakasungwa',
 'Kiranjeranje Water supply',
 'Tumb',
 'Divue ri',
 'Mich',
 'Maut',
 'Nyaluhande',
 'Tutu',
 'Ubaa water supply',
 'Mwamashele water piped scheme',
 'Changalikwa water su',
 'KARUKEKERE WATER ENVIRONMENT SANITATION SCHEME',
 'Gelailumbwa water project',
 'Nyaruyoba/ Kasaka gravity water',
 'Diburuma',
 'Saranda water supply',
 'Kipumbwi water suppl',
 'Kwitaba spring source',
 'Nyakaiga  water s',
 'Muriti Water Supply',
 'Mashangwe',
 'Kipara',
 'Mlal',
 'ngamanga water supplied sch',
 'Tangawizi Water Supply',
 'TM part Six',
 'Du',
 'Kazuramimba water project',
 'Bang',
 'Gonjaugu water supply',
 'Kambi ya chokaa pipe scheme',
 'Oldonyowas maji salama',
 'Nyantamba',
 'Nyasovu',
 'New keni water supply',
 'Makang',
 'Kiindogolo',
 'Kilindogolo',
 'Sekei pipe line',
 'US Embassy Borehole Scheme',
 'Ilambila water supply',
 'Ussoke mlimani water supply',
 'Kilumb',
 'Mvaji ri',
 'Kima',
 'Sopa water supply',
 'Mkulu water supply',
 'Churu water supply',
 'Mradi wa maji Vijijni',
 'Kwakoa water supply',
 'IWEKULE WATER SUPPLY',
 'Tamb',
 'Chiola',
 'Kenswa',
 'Suwasa water supply',
 'Mlinga streem',
 'Muun',
 'Kagunguli Water Supply',
 'Ballaa pipe line',
 'BL Kandashi',
 'Chibula',
 'Namikango',
 'BL Lekirumuni',
 'Kumubila and mukalinzi  spring source',
 'Kihoro',
 'Gallapo water supply',
 'Intinka Water Suppy',
 'Ngulu water Supply',
 'Ihongo',
 'Nsololo water supply',
 'Segese pipe scheme',
 'Manyoni water supply',
 'Ndapo',
 'Ikondo electrical water sch',
 'Magoto piped Water suplly',
 'Machame Aleni water supply',
 'Mambreni gravity water supply',
 'Ilas',
 'Dong',
 'KASHWASA',
 'Nyakagera water s',
 'Kindoroko  water supply',
 'Ushiri water supply',
 'Sumayan gravit water supply',
 'Image',
 'Nameqhwadiba',
 "Mang'ora juu water supply",
 'Shengena water supply',
 'Kalesha water supply',
 'BL Ngarenairobi',
 'Chikola water supply',
 'Maswa Water supplier supply program',
 'Mitema/MWS',
 'Mashami water supply',
 'Msitu wa tembo pipe scheme',
 'IRAMBA NAMHULA WATER SCHEME',
 'Kileo Water Supply',
 'Mushori',
 'Gyewasu',
 'New keni A water suply',
 'Mradi wa maji wa ruvuma',
 'Kigw',
 'Kyamara gravity water supply',
 'Nger',
 'Minyinya Piped water supply',
 'MigelegeleWater supply',
 'Samb',
 'Ivalal',
 'Mwera water supply',
 'Losinoni maji salama',
 'Losinoni maji salama',
 'Namajani',
 'Mradi wa maji wa misha',
 'Makelele water suppl',
 'Soya',
 'Ntom',
 'Olmolog water project',
 'RADA',
 'Mradi wa maji wa kizwi',
 'Bulale Water Supply',
 'Kamb',
 'Kido',
 'Kwemkuna',
 'Mwadui piped scheme',
 'Misha water suppl',
 'Usungi',
 'Mradi wa maji wa chemchem',
 'Mzia water supply',
 'Nywelo streem',
 'Mradi wa maji wa matogolo',
 'Msakangoto',
 'Nyamno',
 'Mtikanga gravity Scheme',
 'Minaki Water Supply',
 'Mtam',
 'Mawande gravity Scheme',
 'Mshinde',
 'Mzizima Water Supply',
 "Lerang'wa water supply",
 'Muhalala water supply',
 'Kiwawa water supply',
 'GEN Borehole Scheme',
 'Maji mingi',
 'Magadini pipe scheme',
 'Machoneni',
 'Kasahunga piped scheme',
 'D',
 'Magoto piped Water  suplly',
 'Mkotokuyana',
 'Makingo water supply',
 'Luwasu water supply',
 'Komolo pipe scheme',
 'Mrike water supply',
 'Olkungabo gravity water supply',
 'Riftvalley Project water supply',
 'Maji ya Chai gravity water supply',
 'MAKOGA',
 'Kigongoi gravity wat',
 'Kuna',
 'Nhobola piped scheme',
 'LAMP water Supplying',
 'Mahuta Scheme',
 'Itunun',
 'Msemembo water supply',
 'Mkalama Water supply',
 'Msaginya',
 'Utaruni pipeline',
 'Mateng',
 'Iwumbu',
 'KINAPA',
 'Bujara water supp',
 'Mradi wa maji wa chanj',
 'Mwabalebi village water pipe scheme',
 'BL Serengeti',
 'WAUSA',
 'World food program',
 'Kitumb',
 'Ndulam',
 'Mradi wa maji wa wino',
 'Mradi wa maji  Kadas',
 "Kiparang'anda Water Supply",
 'Windmili system',
 'Rura',
 'Kura',
 'mazinde water  suppl',
 'Pote',
 'Mradi wa maji wa isevya',
 'Me',
 'Senashida water supply',
 'ENDAYAYA WATER SUPPLY',
 'Pamila water project',
 'Kiholo',
 'Kaserebuka Water Supply',
 'KIBUWA',
 'Mradi wa maji shirati',
 'Mazinde ngua water s',
 'World banks water supplying',
 'Lema water supplied scheme',
 'Rural water supply',
 'Mradi wa maji wa magagura',
 'Mradi wa maji wa mbinga mh',
 'Mgama',
 'Mpun',
 'Koronani Borehole',
 'Mlowa',
 'Madiha',
 'Maili Sita',
 'Lufumb',
 'Kizingu',
 'Vuje Water Supply',
 'Lowasa pipe scheme',
 'Chanyangabwa wate',
 'Sanjaranda water supply',
 'Mradi wa Mission',
 'Kuut',
 'WD and ID',
 'Ruz',
 'Mallama',
 'Kasota',
 'Ibiki gravity water scheme',
 'Mbawala chini water supply',
 'Ulanda',
 'Chipuputa',
 'Wekule',
 'A',
 'Nrashu gravity water supply',
 'Kiba',
 'Ipande water supply',
 'Lima gravity water supply',
 'Bere',
 'Ngasamo Water Supply',
 'Kwa Sondro water supply',
 'Kastamu water supply',
 'sakalenga water supply sche',
 'Kandawale Water supply',
 'Maleut',
 'Kway',
 'Mwazye water supply',
 'Olumuro pipe scheme',
 'Atta',
 'Mongwa r',
 'Ms',
 "Ng'au",
 'Utanzi',
 'Mukabuye gravity water supply',
 'Jongoo',
 'Kabingo/kiobela gravity  water supply',
 'Handeni water suply',
 'Mbakwe water supply',
 'Mwangwe',
 'Solo',
 'BL Kimaroroni',
 'BL Mkombozi',
 'Mission',
 'Tank fupi Mwenge',
 'Kino',
 'Kiho',
 'Kamwanga Erikaswa water pr',
 'Kwamazandu water sup',
 'Kalenge',
 'Mpal',
 'Rondo Water Supply',
 'Mvango Water Supply',
 'Kishiha',
 'Gumb',
 'BL Makiwaru',
 'Murukoli spring source',
 'Nyakonga piped Water  suplly',
 'Mwendakulima pipe',
 'Loliondo secondary scheme',
 'Mradi wa maji wa mjimwema',
 'Pipe scheme',
 'Orkesumeti pipe scheme',
 'mtikanga supply scheme',
 'Shidere mrimasha water supply',
 'Sokoni II pipe line',
 'Mahuta',
 'Kadashi Water Supply',
 'Chimbendenga',
 'Idif',
 'Kwemishiwi water sup',
 'Mshi',
 'Riv',
 'Maranje water supply',
 'Rural water supply &sanitation  program',
 'Mwongolele water project',
 'Mott',
 'Lutende Scheme',
 'ngamanga  water supplied sc',
 'Tubu',
 'Many',
 'Kasanda solar pumping water supply',
 'Mkasale/Mkotamo Water Supply',
 'kidewa',
 'Chil',
 'Mufisi',
 'Mvum',
 'Mlingotini wa',
 'Ruvu Juu',
 'Isikizya water supply',
 'Mgowel',
 'Mche',
 'JUWAMASU',
 'DED',
 'Luka',
 'Tiflo masaki branch line',
 'Ikola water supply',
 'Misigyo pipelines',
 'Kilimarondo',
 'Mang`o',
 'Mba',
 'Mradi wa maji wa matan',
 'Makiba pumping water supply',
 'Chib',
 'Longido water Supply',
 'Bwiti gravity water',
 'Rain water harvesting',
 'P',
 'Mahurunga water supply',
 "Lerang'wa water",
 'Merali Line',
 'Jumuhiya ya watumia maji',
 'image water supply scheme',
 'Welela Shallow well',
 'Kumsasa  spring source',
 'Gale water supply',
 'Tingi water supply',
 'Mgun',
 'UNUA piped scheme',
 'uNoA piped scheme',
 'Ilolangulu water supply',
 'Mwembe Water Supply',
 'Kidu',
 'Isapul',
 'Mbimbi water gravity scheme',
 'Mtangashari',
 'Kishuro water sup',
 'Arisi/himo',
 'Zigi',
 'Ndimira water supply',
 'Saibala gravity water supply',
 'Lihimalyao water supply',
 'Mzinga r',
 'Utweve',
 'Kagenyi water sup',
 'BL Kitahemo',
 'Bwambo Water Supply',
 'Njomlole water gravity scheme',
 'Sauwasa water supply',
 'Tagame',
 'Matala pipeline',
 'Gedamar water supply',
 'Marera water supply',
 'Kilesi water supply',
 'Maka',
 'Kolo',
 'Kiroriko water supply',
 'Malambo water scheme',
 'Muungano',
 'BL Nshere Juu',
 'Mkata ri',
 'Vunta water supply',
 'Luwamakaa branch line',
 'Msjimingi',
 'Kitowo',
 'BL Zahanati',
 'Ichonde',
 'Kashishi water supply',
 'EKTM 2 water suply',
 'Nyamasenene Water Supply',
 'Rumashi gravity water supply',
 'BL Sanya Hoi',
 'Mangola pipe scheme',
 'Chol',
 'Kaguruka Water Supply',
 'Mradi wa maji Kowak',
 'Ibih',
 'Mban',
 'TM part Three water supply',
 'Kabaheshi pring source',
 'Makiyui stream',
 ...]
```

We replaced the nulls with 'None'. Although there is a strong case to drop this column (which now contains 28,810 'None' values), we left it in the dataframe.

In [55]:

```
df['scheme_name'].fillna(value='None', inplace = True)
```

In [56]:

```
df['scheme_name'].isna().sum()
```

Out[56]:

```
0
```

In [57]:

```
df['scheme_name'].value_counts()
```

```
df['scheme_name'].value_counts()
```

Out[57]:

```
None                       28810
K                            682
Borehole                     546
Chalinze wate                405
M                            400
                           ...  
Nkwe                           1
BUWADA                         1
BL Bonifas Kingu               1
BL Losei                       1
DMK:Anglican church            1
Name: scheme_name, Length: 2696, dtype: int64
```

Permit

In [58]:

```
df['permit'].isna().sum()
```

Out[58]:

```
3056
```

There were 3,056 nulls in 'permit'.

We assumed that if a permit status is unknown, there is no permit. The nulls were changed to False.

In [59]:

```
#if the waterpoint is permitted
df['permit'].value_counts()
```

Out[59]:

```
True     38852
False    17492
Name: permit, dtype: int64
```

In [60]:

```
df['permit'].isna().sum()
```

Out[60]:

```
3056
```

In [61]:

```
df['permit'].fillna(value=False, inplace=True)
```

In [62]:

```
df['permit'].isna().sum()
```

Out[62]:

```
0
```

Column Exploration

Now that null values have been treated, we continued our exploration of the data by examining columns individually.

In [63]:

```
df.info()
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 59400 entries, 0 to 59399
Data columns (total 42 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   id                     59400 non-null  int64  
 1   amount_tsh             59400 non-null  float64
 2   date_recorded          59400 non-null  object 
 3   funder                 59400 non-null  object 
 4   gps_height             59400 non-null  int64  
 5   installer              59400 non-null  object 
 6   longitude              59400 non-null  float64
 7   latitude               59400 non-null  float64
 8   wpt_name               59400 non-null  object 
 9   num_private            59400 non-null  int64  
 10  basin                  59400 non-null  object 
 11  subvillage             59400 non-null  object 
 12  region                 59400 non-null  object 
 13  region_code            59400 non-null  int64  
 14  district_code          59400 non-null  int64  
 15  lga                    59400 non-null  object 
 16  ward                   59400 non-null  object 
 17  population             59400 non-null  int64  
 18  public_meeting         59400 non-null  bool   
 19  recorded_by            59400 non-null  object 
 20  scheme_management      59400 non-null  object 
 21  scheme_name            59400 non-null  object 
 22  permit                 59400 non-null  bool   
 23  construction_year      59400 non-null  int64  
 24  extraction_type        59400 non-null  object 
 25  extraction_type_group  59400 non-null  object 
 26  extraction_type_class  59400 non-null  object 
 27  management             59400 non-null  object 
 28  management_group       59400 non-null  object 
 29  payment                59400 non-null  object 
 30  payment_type           59400 non-null  object 
 31  water_quality          59400 non-null  object 
 32  quality_group          59400 non-null  object 
 33  quantity               59400 non-null  object 
 34  quantity_group         59400 non-null  object 
 35  source                 59400 non-null  object 
 36  source_type            59400 non-null  object 
 37  source_class           59400 non-null  object 
 38  waterpoint_type        59400 non-null  object 
 39  waterpoint_type_group  59400 non-null  object 
 40  id_label               59400 non-null  int64  
 41  status_group           59400 non-null  object 
dtypes: bool(2), float64(3), int64(8), object(29)
memory usage: 18.2+ MB
```

In [64]:

```
#total static head (amount of water available to waterpoint)
df['amount_tsh'].value_counts()
```

Out[64]:

```
0.0         41639
500.0        3102
50.0         2472
1000.0       1488
20.0         1463
            ...  
8500.0          1
6300.0          1
220.0           1
138000.0        1
12.0            1
Name: amount_tsh, Length: 98, dtype: int64
```

In [65]:

```
df[df['amount_tsh'] == 0.0]['status_group'].value_counts()
```

Out[65]:

```
functional                 19706
non functional             18885
functional needs repair     3048
Name: status_group, dtype: int64
```

In [66]:

```
df['date_recorded']
```

Out[66]:

```
0        2011-03-14
1        2013-03-06
2        2013-02-25
3        2013-01-28
4        2011-07-13
            ...
59395    2013-05-03
59396    2011-05-07
59397    2011-04-11
59398    2011-03-08
59399    2011-03-23
Name: date_recorded, Length: 59400, dtype: object
```

In [67]:

```
sorted(df['date_recorded'].tolist(), reverse = True )
#2002 to 2013
```

Out[67]:

```
['2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
 '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-03',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-02',
    '2013-12-01',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-03',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-11-02',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-03',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-10-02',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-03',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-09-02',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
    '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 '2013-08-03',
 ...]
```

'date_recorded' is stored in yy-mm-dd format. We thought about using this column along with 'construction_year' to determine pump age. However, 'construction_year' has too many 0 values for us to construct a meaningful age feature.

In [68]:

```
df['construction_year'].value_counts()
```

Out[68]:

```
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
```

```
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
```

In [69]:

```
#altitude of the well
df['gps_height']
```

Out[69]:

```
0        1390
1        1399
2         686
3         263
4           0
         ...
59395    1210
59396    1212
59397       0
59398       0
59399     191
Name: gps_height, Length: 59400, dtype: int64
```

In [70]:

```
df['longitude'].value_counts()
```

Out[70]:

```
0.000000     1812
37.540901       2
33.010510       2
39.093484       2
32.972719       2
             ...
37.579803       1
33.196490       1
34.017119       1
33.788326       1
30.163579       1
Name: longitude, Length: 57516, dtype: int64
```

In [71]:

```
#remove filler values in longitude and latitude
# dropping longitude of 0
df = df.loc[df["longitude"] != 0]
df.longitude.value_counts()
```

Out[71]:

```
33.090347     2
```

```
33.090347    2
32.982698    2
37.297680    2
33.010510    2
39.093484    2
            ..
37.579803    1
33.196490    1
34.017119    1
33.788326    1
35.005922    1
Name: longitude, Length: 57515, dtype: int64
```

In [72]:

```
df['latitude'].value_counts()
```

Out[72]:

```
-2.496459    2
-6.964258    2
-6.981884    2
-7.175174    2
-7.104625    2
            ..
-5.726001    1
-9.646831    1
-8.124530    1
-2.535985    1
-2.598965    1
Name: latitude, Length: 57516, dtype: int64
```

Both latitude and longitude seem to have placeholder values. These values are 0.000000 for 'longitude' and -2.000000e-08 for 'latitude'. There are 1,812 of these placeholders.

In [73]:

```
#name of the waterpoint
df['wpt_name']
```

Out[73]:

```
0                         none
1                     Zahanati
2                  Kwa Mahundi
3         Zahanati Ya Nanyumbu
4                      Shuleni
                 ...
59395      Area Three Namba 27
59396        Kwa Yahona Kuvala
59397                  Mashine
59398                   Mshoro
59399          Kwa Mzee Lugawa
Name: wpt_name, Length: 57588, dtype: object
```

In [74]:

```
df['wpt_name'].value_counts()
```

Out[74]:

```
none                    3492
Shuleni                 1734
Zahanati                 814
Msikitini                533
Kanisani                 322
                        ...
Kwa Kadungu                1
Nyanguruguru               1
Kwa Piguli                 1
Kalakabanga                1
Kwa Mtunga Lyimo           1
Name: wpt_name, Length: 36720, dtype: int64
```

```
Name: wpt_name, Length: 30720, dtype: int64
```

We thought 'wpt_name' is probably not a useful feature for classification, similar to 'scheme_name'. The many unique values would limit out model's performance with 37,400 unique values in this column. Some of these are subvillage names (like Shuleni), explaning why some values have much higher value counts than others.

In [75]:

```
#no context
df['num_private'].value_counts()
```

Out[75]:

```
0      56831
6         81
1         73
5         46
8         46
       ...
180        1
213        1
23         1
55         1
94         1
Name: num_private, Length: 65, dtype: int64
```

There was no context available for 'num_private', not even from the data source. We decided this was another column that would go unused in modeling since we were unable to identify its meaning. Additionally, the large value counts imbalance was worrying without any business understanding of these values.

In [76]:

```
#Geographic water basin
df['basin'].value_counts()
```

Out[76]:

```
Pangani                    8940
Lake Victoria              8535
Rufiji                     7976
Internal                   7785
Lake Tanganyika            6333
Wami / Ruvu                5987
Lake Nyasa                 5085
Ruvuma / Southern Coast    4493
Lake Rukwa                 2454
Name: basin, dtype: int64
```

In [77]:

```
#Geographic location
df['subvillage'].value_counts()
```

Out[77]:

```
Majengo                 494
Shuleni                 492
Madukani                435
Unknown                 371
Kati                    366
                       ...
Igodimwa                  1
Jengemwanama              1
Buyoga A                  1
Ilolangulu Busenda B      1
Bombambili 2              1
Name: subvillage, Length: 18568, dtype: int64
```

In [78]:

```
#geographic location
```

```
df['region'].value_counts()
```

Out[78]:

```
Iringa           5294
Mbeya            4639
Kilimanjaro      4379
Morogoro         4006
Shinyanga        3977
Arusha           3350
Kagera           3316
Kigoma           2816
Ruvuma           2640
Pwani            2635
Tanga            2547
Mwanza           2295
Dodoma           2201
Singida          2093
Mara             1969
Tabora           1959
Rukwa            1808
Mtwara           1730
Manyara          1583
Lindi            1546
Dar es Salaam     805
Name: region, dtype: int64
```

Both 'region_code' and 'district_code' are integer types. However, as codes, it's unlikely that there is any meaning to the order or progression of values. Therefore, we decided to treat them as categorical variables and converted them to strings.

In [79]:

```
#geographic location (coded)
df['region_code'].value_counts()
```

Out[79]:

```
11    5297
12    4639
3     4379
5     4040
17    3954
18    3324
2     3024
16    2816
10    2640
4     2513
19    2295
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
```

In [80]:

```
#region code is type int, but best presented as a categorical feature
#convert to type string
```

```
df['region_code'] = df['region_code'].astype(str)
```

In [81]:

```
df['region_code']
```

Out[81]:

```
0        11
1        20
2        21
3        90
4        18
         ..
59395     3
59396    11
59397    12
59398     1
59399     5
Name: region_code, Length: 57588, dtype: object
```

In [82]:

```
#geographic location (coded)
df['district_code'].value_counts()
```

Out[82]:

```
1     11146
2     10909
3      9998
4      8996
5      4356
6      3586
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
```

In [83]:

```
#district code is type int, but best presented as a categorical feature
#convert to type string
df['district_code'] = df['district_code'].astype(str)
```

In [84]:

```
df['district_code']
```

Out[84]:

```
0         5
1         2
2         4
3        63
4         1
         ..
59395     5
59396     4
59397     7
59398     4
59399     2
```

```
Name: district_code, Length: 57588, dtype: object
```

Based on some research and information lookup, 'lga' seems to refer to cities or areas of cities.

In [85]:

```
#geographic location (city?)
df['lga'].value_counts()
```

Out[85]:

```
Njombe          2503
Arusha Rural    1252
Moshi Rural     1251
Rungwe          1106
Kilosa          1094
                ... 
Moshi Urban       79
Kigoma Urban      71
Arusha Urban      63
Lindi Urban       21
Nyamagana          1
Name: lga, Length: 124, dtype: int64
```

In [86]:

```
#geographic location (ward?)
df['ward'].value_counts()
```

Out[86]:

```
Igosi              307
Imalinyi           252
Siha Kati          232
Mdandu             231
Nduruma            217
                  ... 
Nsemulwa             1
Ifinga               1
Kihangimahuka        1
Themi                1
Mawenzi              1
Name: ward, Length: 2033, dtype: int64
```

The 'population' column has a large number of '0' values for the population around the pump.

In [87]:

```
#population around the well
df['population'].value_counts()
```

Out[87]:

```
0       19569
1        7025
200      1940
150      1892
250      1681
        ... 
3241        1
1960        1
1685        1
2248        1
1439        1
Name: population, Length: 1049, dtype: int64
```

In [88]:

```
df['public_meeting'].value_counts()
```

Out[88]:

```
True      52713
False      4875
Name: public_meeting, dtype: int64
```

The 'recorded_by' column told us that the data was collected by a group called GeoData Consultants Ltd. This fact was useful for our data understanding but inessential to our modeling.

In [89]:

```
#feature unimportant for modeling
df['recorded_by'].value_counts()
```

Out[89]:

```
GeoData Consultants Ltd    57588
Name: recorded_by, dtype: int64
```

As mentioned earlier, there were numerous 0 values in 'construction_year'. 20,709 values represents about one-third of our total amount of data.

In [90]:

```
#how do we treat 0's in construction year?
#age as an ordinal encoded variable to properly treat 0's?
df['construction_year'].value_counts()
```

Out[90]:

```
0       18897
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
```

In [91]:

```
df[df['construction_year'] != 0]['construction_year'].describe()
#contruction years range from 1960 to 2013, with 20709 values of 0
```

Out[91]:

```
count    38691.000000
mean      1996.814686
std         12.472045
min       1960.000000
25%       1987.000000
50%       2000.000000
75%       2008.000000
max       2013.000000
Name: construction_year, dtype: float64
```

From the known construction years, we saw a range from 1960 to 2013. This gave us an idea of the vastness of this data set.

Similar Columns

The following section identifies groups of columns that all contain similar information. Within these groups, some columns are more granular than others. For modeling, we only wanted to use a single column for each group so that we wouldn't have multiple features all encapsulating essentially the same information.

In [92]:

```
df['extraction_type'].value_counts()
```

Out[92]:

```
gravity                         26696
nira/tanira                      7361
other                            6160
submersible                      4688
swn 80                           3448
mono                             2817
india mark ii                    2284
afridev                          1659
ksb                              1358
other - rope pump                 451
other - swn 81                    229
windmill                          117
india mark iii                     91
cemo                               90
other - play pump                  85
climax                             32
walimi                             20
other - mkulima/shinyanga           2
Name: extraction_type, dtype: int64
```

In [93]:

```
df['extraction_type_group'].value_counts()
```

Out[93]:

```
gravity              26696
nira/tanira           7361
other                 6160
submersible           6046
swn 80                3448
mono                  2817
india mark ii         2284
afridev               1659
rope pump              451
other handpump         336
other motorpump        122
wind-powered           117
india mark iii          91
Name: extraction_type_group, dtype: int64
```

In [94]:

```
df['extraction_type_class'].value_counts()
```

Out[94]:

```
gravity          26696
handpump         15179
other             6160
submersible       6046
motorpump         2939
rope pump          451
wind-powered       117
Name: extraction_type_class, dtype: int64
```

'extraction_type' is the most granular, and 'extraction_type_class' is the most broad.

In [95]:

```
#how does this differ from scheme_management
df['management'].value_counts()
```

Out[95]:

```
vwc                 39746
wug                  5556
water board          2932
wua                  2533
private operator     1970
parastatal           1696
water authority       902
other                 840
company               685
unknown               551
other - school         99
trust                  78
Name: management, dtype: int64
```

In [96]:

```
df['scheme_management'].value_counts()
```

Out[96]:

```
VWC                 36143
WUG                  4249
None                 3751
Water authority      3151
WUA                  2882
Water Board          2747
Parastatal           1607
Private operator     1063
Company              1061
Other                 765
SWC                    97
```

```
Trust                     72
Name: scheme_management, dtype: int64
```

In [97]:

```
df['management_group'].value_counts()
```

Out[97]:

```
user-group    50767
commercial     3635
parastatal     1696
other           939
unknown         551
Name: management_group, dtype: int64
```

'management' is the most granular, and 'management_group' is the most broad. Capitalization was also inconsistent between columns.

In [98]:

```
df['payment'].value_counts()
```

Out[98]:

```
never pay                24380
pay per bucket            8953
pay monthly               8229
unknown                   7654
pay when scheme fails     3843
pay annually              3626
other                      903
Name: payment, dtype: int64
```

In [99]:

```
df['payment_type'].value_counts()
```

Out[99]:

```
never pay     24380
per bucket     8953
monthly        8229
unknown        7654
on failure     3843
annually       3626
other           903
Name: payment_type, dtype: int64
```

The value counts for 'payment' and 'payment_type' aligned perfectly. They relayed the same information with differently named categories.

In [100]:

```
df['water_quality'].value_counts()
```

Out[100]:

```
soft                  49431
salty                  4772
unknown                1661
milky                   803
coloured                479
salty abandoned         228
fluoride                199
fluoride abandoned       15
Name: water_quality, dtype: int64
```

In [101]:

```
df['quality_group'].value_counts()
```

Out[101]:

```
good        49431
salty        5000
unknown      1661
milky         803
colored       479
fluoride      214
Name: quality_group, dtype: int64
```

'water_quality' and 'quality_group' convey the same information with some slight differences in values. water_quality is more granular.

In [102]:

```
df['quantity'].value_counts()
```

Out[102]:

```
enough          32260
insufficient    14564
dry              5990
seasonal         4001
unknown           773
Name: quantity, dtype: int64
```

In [103]:

```
df['quantity_group'].value_counts()
```

Out[103]:

```
enough          32260
insufficient    14564
dry              5990
seasonal         4001
unknown           773
Name: quantity_group, dtype: int64
```

'quantity' and 'quantity_group' had identical value counts and category names.

In [104]:

```
df['source'].value_counts()
```

Out[104]:

```
spring                  17006
shallow well            15499
machine dbh             10826
river                    9612
rainwater harvesting     2218
hand dtw                  873
dam                       649
lake                      639
other                     202
unknown                    64
Name: source, dtype: int64
```

In [105]:

```
df['source_type'].value_counts()
```

Out[105]:

```
spring                  17006
shallow well            15499
borehole                11699
river/lake              10251
rainwater harvesting     2218
dam                       649
other                     266
other                     266
Name: source_type, dtype: int64
```

In [106]:

```
df['source_class'].value_counts()
```

Out[106]:

```
groundwater    44204
surface        13118
unknown          266
Name: source_class, dtype: int64
```

'source' is the most granular, and 'source_class' is the most broad.

In [107]:

```
df['waterpoint_type'].value_counts()
```

Out[107]:

```
communal standpipe             28375
hand pump                      16181
other                           6167
communal standpipe multiple     5959
improved spring                  783
cattle trough                    116
dam                                7
Name: waterpoint_type, dtype: int64
```

In [108]:

```
df['waterpoint_type_group'].value_counts()
```

Out[108]:

```
communal standpipe    34334
hand pump             16181
other                  6167
improved spring         783
cattle trough           116
dam                       7
Name: waterpoint_type_group, dtype: int64
```

waterpoint_type and waterpoint_type_group are almost identical except the communal standpipe multiple group in 'waterpoint_type' is included in 'communal standpipe' in 'waterpoint_type_group'

In [109]:

```
df['id'].value_counts()
```

Out[109]:

```
2047     1
20959    1
4759     1
661      1
2708     1
        ..
62836    1
52595    1
50546    1
56689    1
0        1
Name: id, Length: 57588, dtype: int64
```

In [110]:

```
df['id_label'].value_counts()
```

Out[110]:

```
2047     1
20959    1
4759     1
661      1
2708     1
        ..
62836    1
52595    1
50546    1
56689    1
0        1
Name: id_label, Length: 57588, dtype: int64
```

We confirmed that no duplicate 'id' appeared in the dataset

In [111]:

```
df.head()
```

Out[111]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
0	69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	
3	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	

5 rows × 42 columns

In [112]:

```
#FINAL NULLS CHECK
df.isna().sum()
```

Out[112]:

```
id                         0
amount_tsh                 0
date_recorded              0
funder                     0
gps_height                 0
installer                  0
longitude                  0
latitude                   0
wpt_name                   0
num_private                0
basin                      0
subvillage                 0
region                     0
region_code                0
district_code              0
lga                        0
ward                       0
population                 0
public_meeting             0
recorded_by                0
scheme_management          0
scheme_name                0
permit                     0
```

```
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
id_label                 0
status_group             0
dtype: int64
```

Our cleaning removed all nulls in the dataset.

After exploring the columns, we adjusted some values that appeared as placeholders. We also identified columns that continued similar or identical information and determined how we would only implement one column from every similar group into our model.

In [113]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 57588 entries, 0 to 59399
Data columns (total 42 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   id                     57588 non-null  int64  
 1   amount_tsh             57588 non-null  float64
 2   date_recorded          57588 non-null  object 
 3   funder                 57588 non-null  object 
 4   gps_height             57588 non-null  int64  
 5   installer              57588 non-null  object 
 6   longitude              57588 non-null  float64
 7   latitude               57588 non-null  float64
 8   wpt_name               57588 non-null  object 
 9   num_private            57588 non-null  int64  
 10  basin                  57588 non-null  object 
 11  subvillage             57588 non-null  object 
 12  region                 57588 non-null  object 
 13  region_code            57588 non-null  object 
 14  district_code          57588 non-null  object 
 15  lga                    57588 non-null  object 
 16  ward                   57588 non-null  object 
 17  population             57588 non-null  int64  
 18  public_meeting         57588 non-null  bool   
 19  recorded_by            57588 non-null  object 
 20  scheme_management      57588 non-null  object 
 21  scheme_name            57588 non-null  object 
 22  permit                 57588 non-null  bool   
 23  construction_year      57588 non-null  int64  
 24  extraction_type        57588 non-null  object 
 25  extraction_type_group  57588 non-null  object 
 26  extraction_type_class  57588 non-null  object 
 27  management             57588 non-null  object 
 28  management_group       57588 non-null  object 
 29  payment                57588 non-null  object 
 30  payment_type           57588 non-null  object 
 31  water_quality          57588 non-null  object 
 32  quality_group          57588 non-null  object 
 33  quantity               57588 non-null  object 
 34  quantity_group         57588 non-null  object 
 35   source                 57588 non-null  object
 36   source_type            57588 non-null  object
 37   source_class           57588 non-null  object
 38   waterpoint_type        57588 non-null  object
 39   waterpoint_type_group  57588 non-null  object
 40   id_label               57588 non-null  int64
 41   status_group           57588 non-null  object
dtypes: bool(2), float64(3), int64(6), object(31)
memory usage: 18.1+ MB
```

In [114]:

```
df.head()
```

Out[114]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	qua
0	69572	6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	2013-03-06	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	
3	67743	0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	

5 rows × 42 columns

We saved the dataframe containing all fo the cleaned data and the 'target_group' into a separate CSV file available in the data folder of this project's repository. This file was imported into our modeling notebook.

In [115]:

```
#SAVE CLEAN DATA OFF TO CSV FILE FOR IMPORT INTO OTHER NOTEBOOKS
#index= False
#df.to_csv('./data/water_well_train_clean.csv', index_label=False)
```

Please proceed to the modeling notebook.