# PUMP IT UP

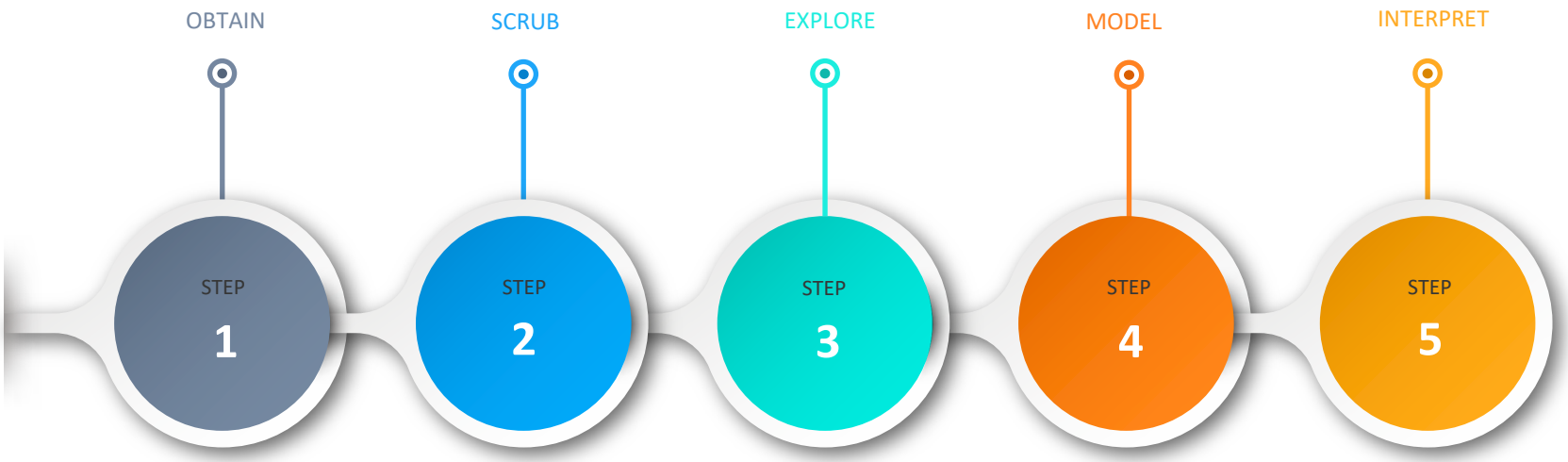
## TERNARY CLASSIFICATION MODEL

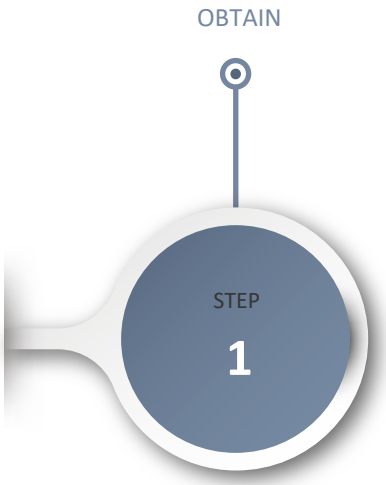
Bao Tram Duong

**Tanzania**, the largest country in East Africa, is suffering from a water crisis

- 4 million people lack access to an improved source of safe water
- 30 million people lack access to improved sanitation
- Water-borne illnesses, such as malaria and cholera, account for over half of the diseases affecting the population

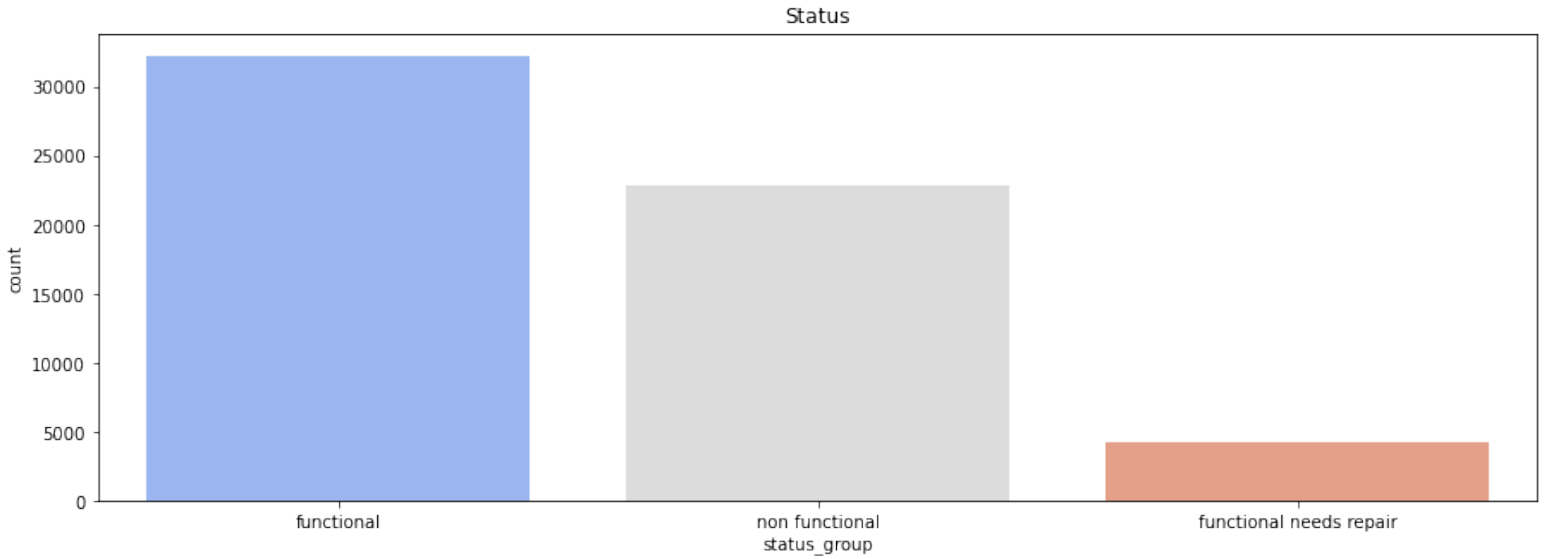
# OSEMN


OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
STEP 1
STEP 2
STEP 3
STEP 4
STEP 5



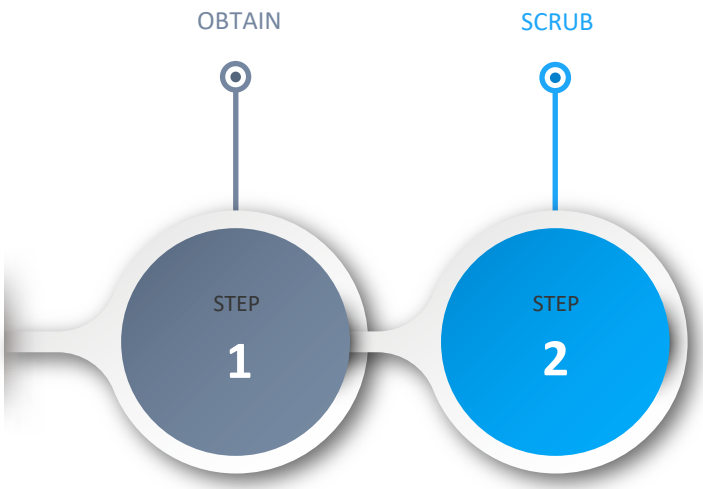

Using data provided by **The Tanzanian Water Ministry** and **Taarifa**, **DrivenData** began a competition to solve this problem by building a classification system to predict whether a given water source is working correctly.

- 59,400 water points
- 40 features

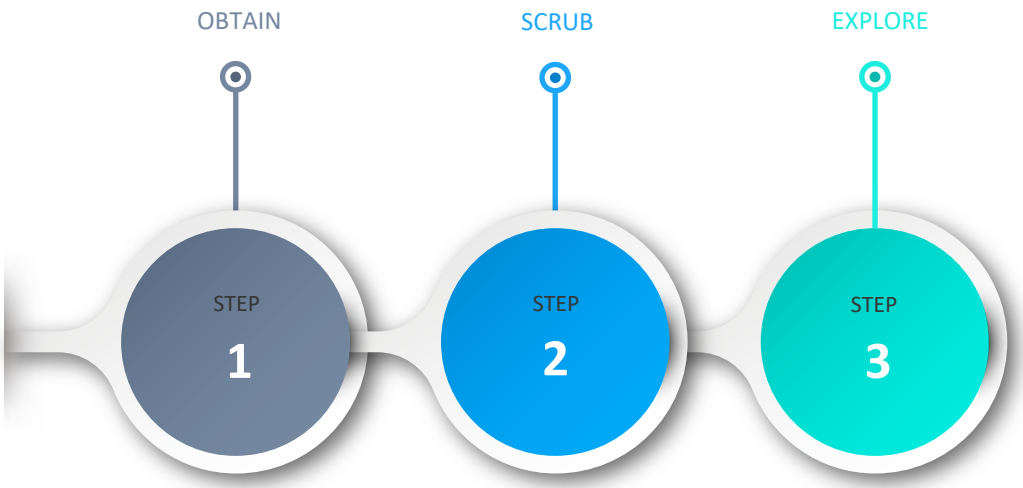

Status
30000
25000
20000
15000
10000
5000
0
count
functional
non functional
functional needs repair
status_group

**54%** **38%** **7%**

# Challenges

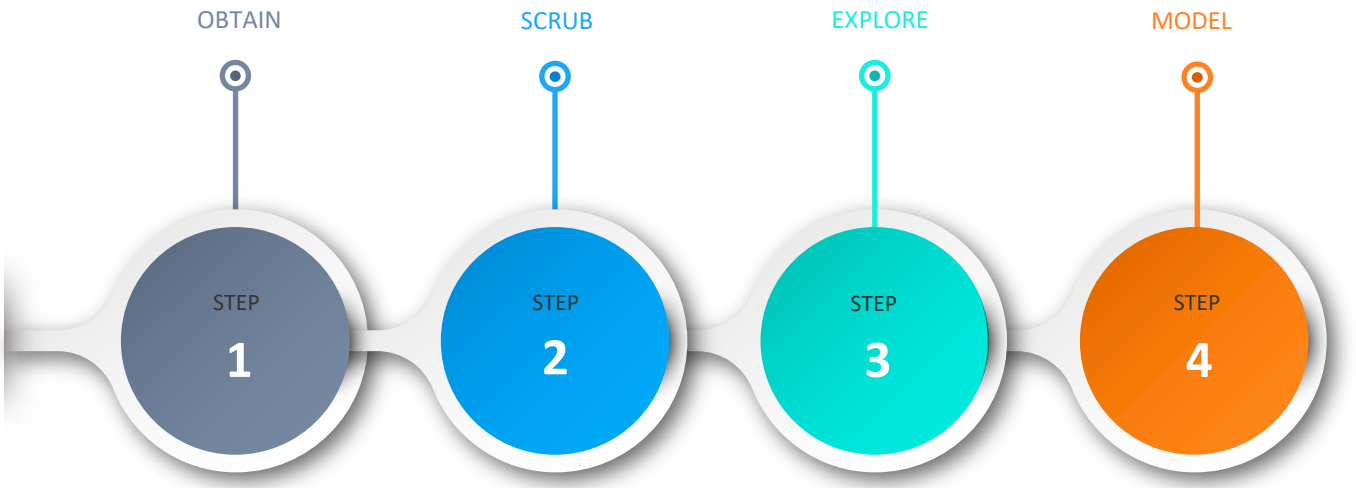



1. Replace missing values with mean, median, or classify them as ‘other’
2. Remove redundant features
3. Fix misspellings and variations
4. Select for the 20 most common values with features that have high number of unique values and categorize the rest as ‘other’
5. Correct class imbalance with SMOTE (Synthetic Minority Over-Sampling Technique)

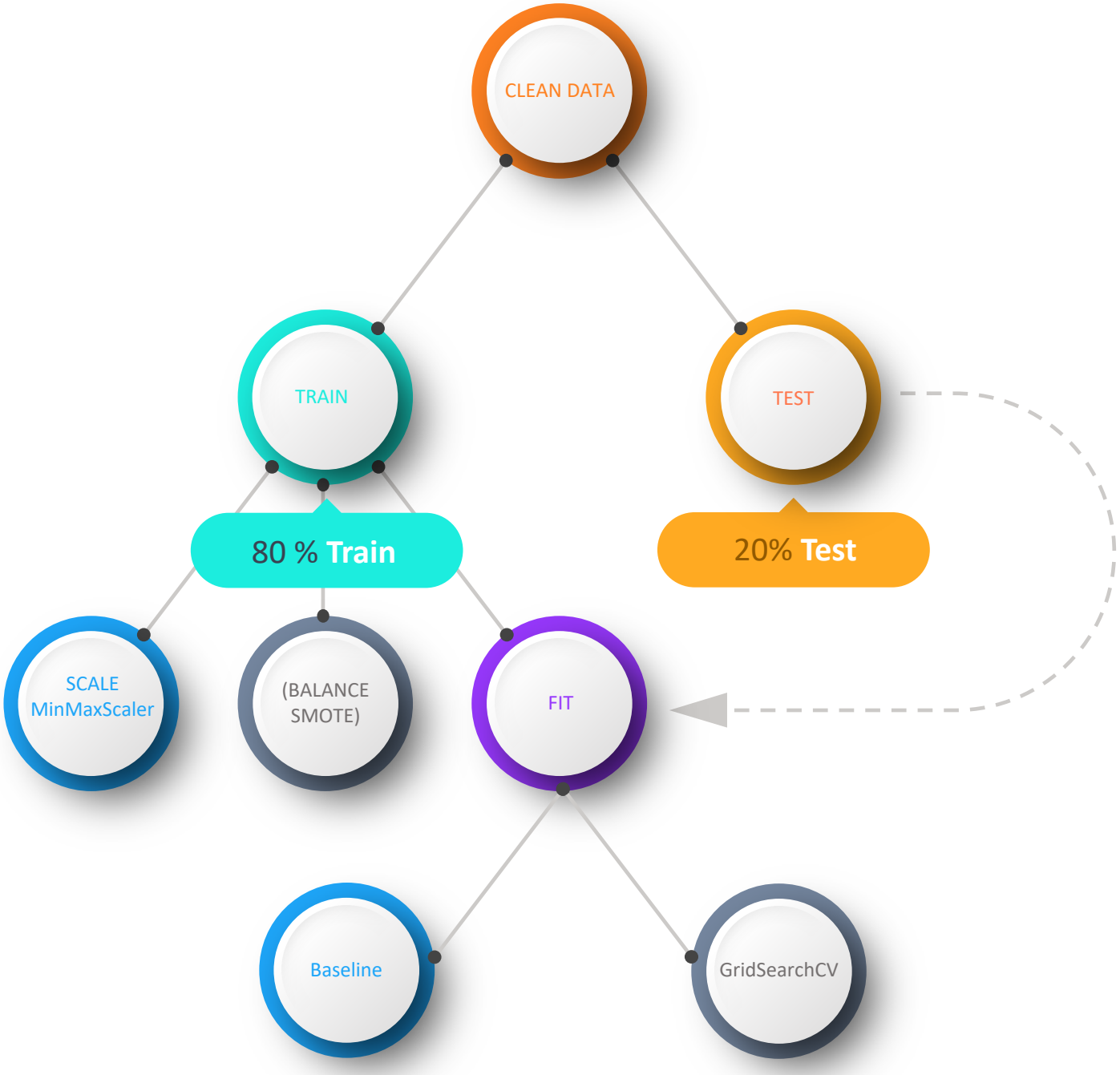

OBTAIN
SCRUB
EXPLORE
STEP
1
STEP
2
STEP
3

# Recommendations

1. Focus on sustainability: early preventative strategy rather than letting things go broken
2. Decentralized management: we need to restructure authority so that there is a system of co-responsibility between the central, regional and local levels.
3. Improved payment system:
    - A local payment system should be put in place so that the user-group can be independently responsible for their own water points
    - Direct funding from international donors to village-level should also be implemented instead of having to go through the long bureaucratic process where money get lost along the way between ministry and district level.



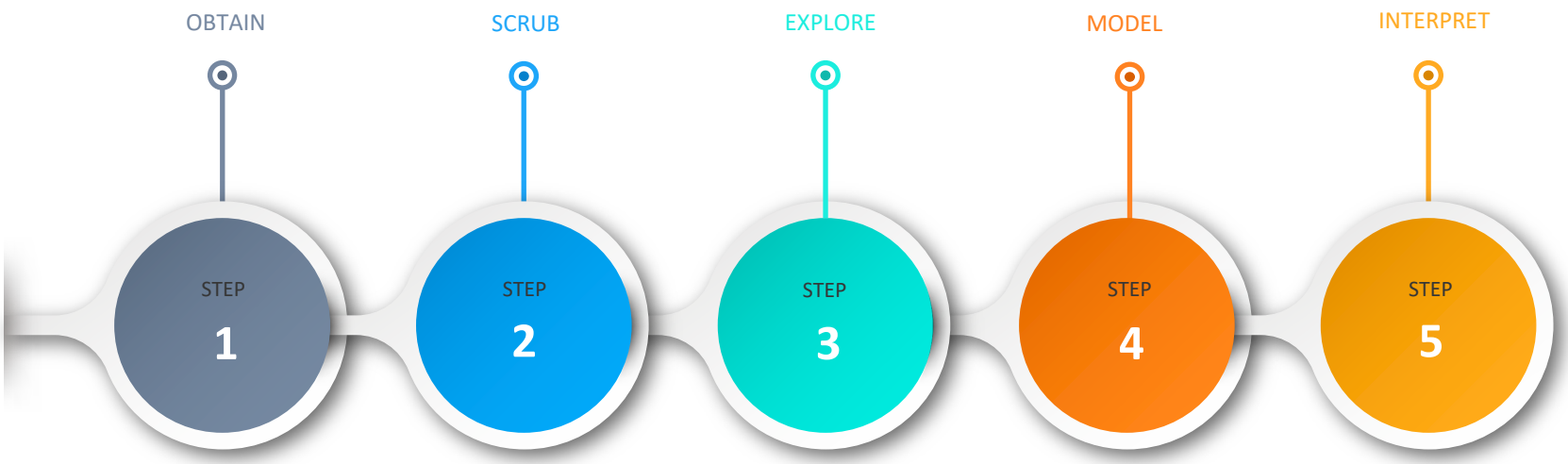

1. First, we convert the labels from strings to integers:
   - Non-functional = 0
   - Functional = 1
   - Functional-needs-repair = 2
2. Dummies encode all categorical variables


CLEAN DATA
TRAIN
TEST
80 % Train
20% Test
SCALE
MinMaxScaler
(BALANCE
SMOTE)
FIT
Baseline
GridSearchCV

# Best Model

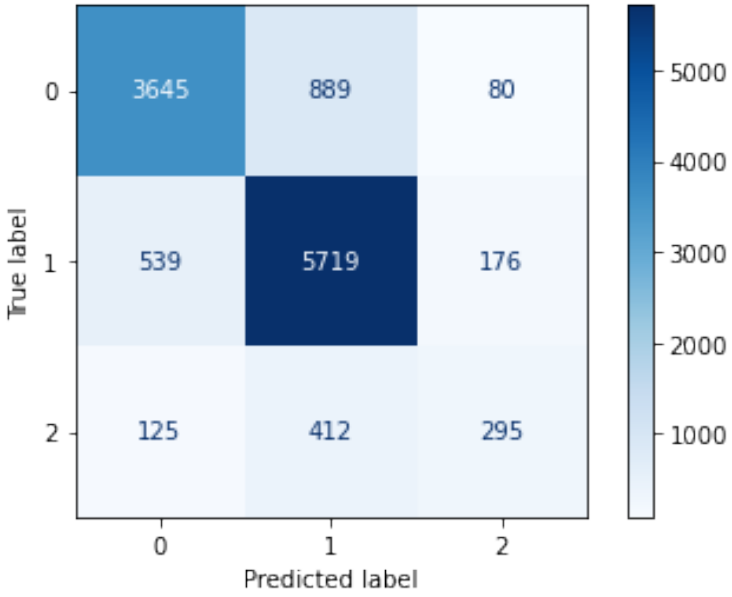
**Gradient Boosting Classifier** is a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.

The weak learner used is **Decision Tree Classifier**, which is the simplest tree-based method.


OBTAIN
SCRUB
EXPLORE
MODEL
INTERPRET
STEP 1
STEP 2
STEP 3
STEP 4
STEP 5

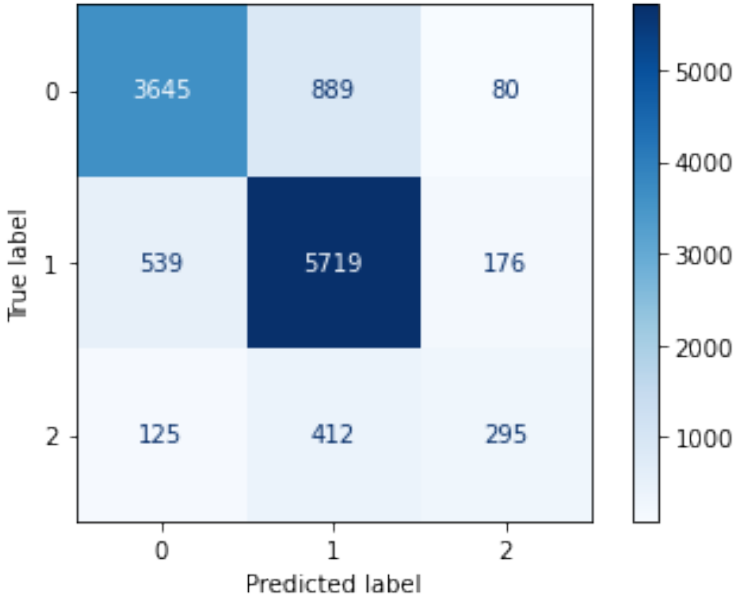
# Accuracy 81.30%

- Train accuracy: 99.33%
- Test accuracy: 81.30%



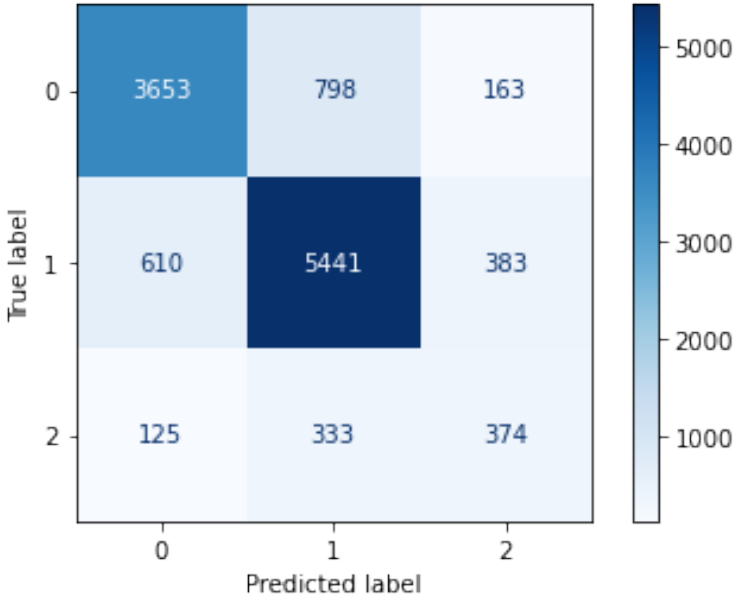

Model: Gradient Boosting

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.79 | 0.82 | 4614 |
| 1 | 0.81 | 0.89 | 0.85 | 6434 |
| 2 | 0.54 | 0.35 | 0.43 | 832 |
| accuracy | | | 0.81 | 11880 |
| macro avg | 0.73 | 0.68 | 0.70 | 11880 |
| weighted avg | 0.81 | 0.81 | 0.81 | 11880 |




Imbalanced Gradient Boost Model

**81.30%**




Balanced Gradient Boost Model

**79.86%**

# Future Work

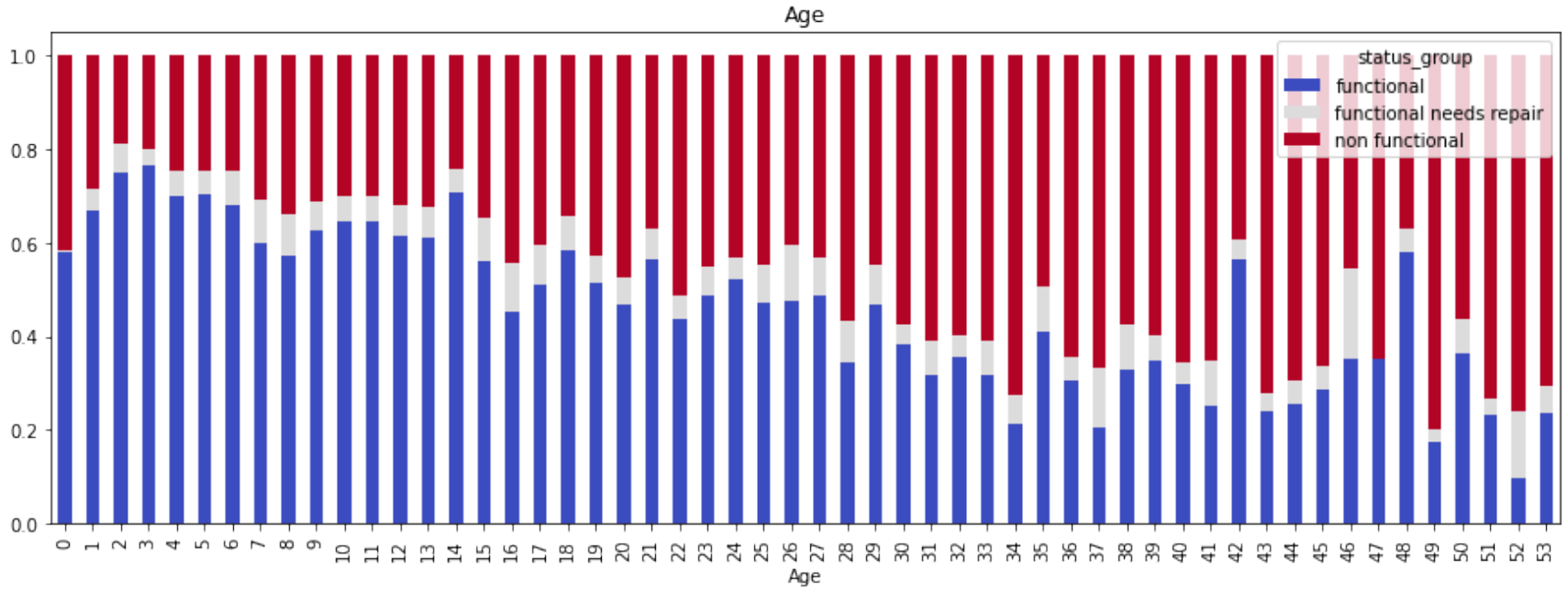
1. Since correcting class imbalance did not improve the model, we can try model stacking i.e build a binary classification between functional vs non-functional and another binary classification between functional vs. functional needs repair.
2. Try more parameters tuning with more and wider range of options
3. Work to reduce overfit while maintaining and/or improving accuracy score

THANK YOU

# APPENDIX

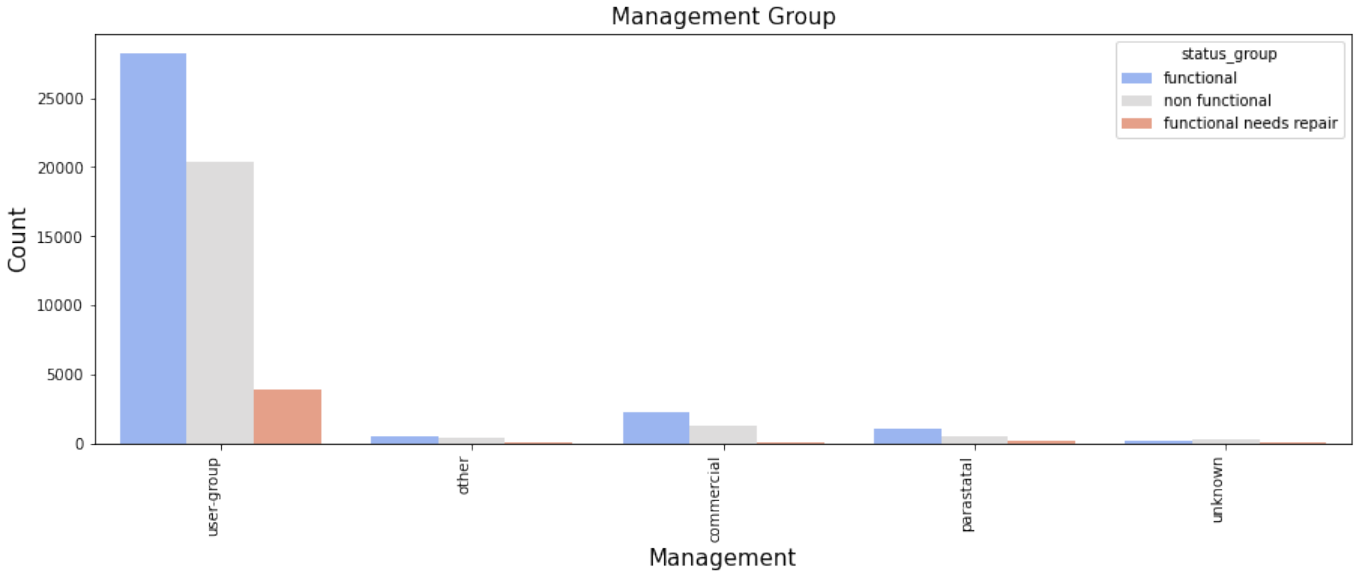
# Problems

1. Sustainability: Regardless of hundreds of millions of dollars over budget and years past the original deadline of the Water Sector Development Program (WSDP), local government and communities find themselves unable to raise the money to fix and maintain their water points.
2. Power struggle: Full responsibility for operating, maintaining and sustaining water points is done at the village level. However, disbursement of funds and report of functionality must follow a long bureaucratic process all the way from the village, to the district, and, finally, to the Ministry of Water. The problem found is not only the miscommunication but also the power struggle around roles, responsibilities, and accountability between many different levels of government.



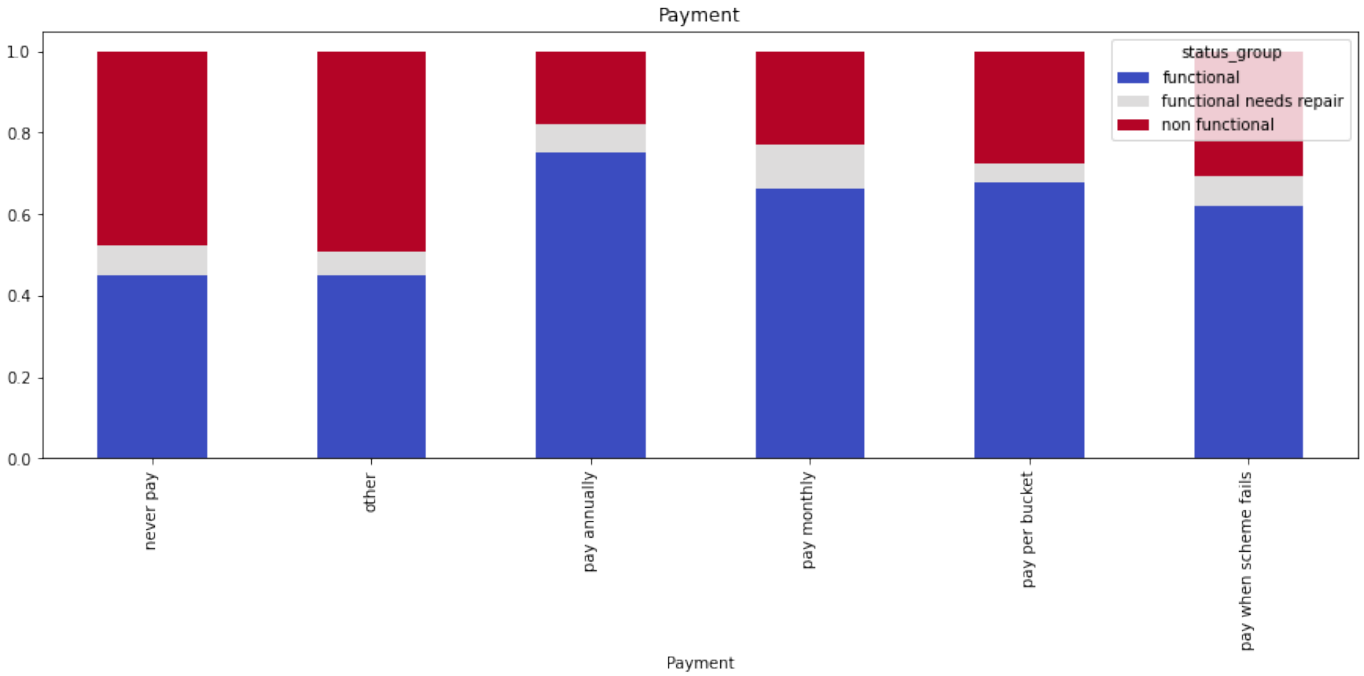

- Within the first year, 30% of water points become non-functional.
- Only 54% of water points are working 15 years after installation.

**Problem:** Regardless of hundreds of millions of dollars over budget and years past the original deadline of the Water Sector Development Programme (WSDP), local government and communities find themselves unable to raise the money to fix and maintain their water points.




- Under NAWAPO (National Water Policy ), user groups is to take the full responsibility for operating, maintaining and sustaining water points at the village level.
- Data is published by the ministry which are based on the coverage reported by district, are not reliable.
- Disbursement of funds must follow a long bureaucratic process of accountability, requiring upwards (vertical) reporting at each level of government, all the way from the village, to the district, and, finally, to the Ministry of Water.

**Problem:** Power struggle around roles, responsibilities, and accountability between different levels of government




- If the water point management charges money, the more likely that it is better maintained and kept functional.
- Regular payment, used for regular maintenance and upkeep, is a better approach for preventative treatments rather than trying to secure a large amount of fund for when the system breaks down.

# Balanced

| | Model | Accuracy | CV | Precision | Recall | F1 Score | MAE | MSE | RMSE | AUC | Bias | Variance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Decision Tree | 76.01 | 0.72 | 0.65 | 0.67 | 0.66 | 0.27 | 0.33 | 0.57 | - | 0.02 | 0.39 |
| **1** | Logistic Regression | 65.31 | 0.74 | 0.60 | 0.67 | 0.58 | 0.42 | 0.55 | 0.74 | 0.82 | 0.24 | 0.55 |
| **2** | KNN | 75.66 | 0.72 | 0.64 | 0.68 | 0.66 | 0.28 | 0.34 | 0.59 | - | 0.05 | 0.40 |
| **3** | Bagged Tree | 78.21 | 0.76 | 0.67 | 0.69 | 0.68 | 0.24 | 0.30 | 0.54 | - | 0.01 | 0.37 |
| **4** | Random Forest | 79.37 | 0.77 | 0.69 | 0.70 | 0.69 | 0.23 | 0.28 | 0.53 | - | 0.02 | 0.37 |
| **5** | Gradient Boost | 79.86 | 0.78 | 0.69 | 0.70 | 0.70 | 0.22 | 0.27 | 0.52 | - | 0.03 | 0.37 |
| **6** | ADABoost | 63.01 | 0.72 | 0.56 | 0.61 | 0.55 | 0.44 | 0.59 | 0.77 | - | 0.21 | 0.53 |
| **7** | XGBoost | 77.03 | 0.78 | 0.67 | 0.71 | 0.68 | 0.26 | 0.32 | 0.57 | - | 0.09 | 0.41 |
| **8** | SVM | 72.50 | 0.76 | 0.64 | 0.71 | 0.64 | 0.32 | 0.42 | 0.65 | - | 0.17 | 0.49 |

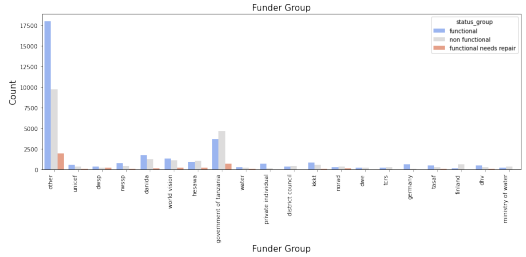
# Imbalanced

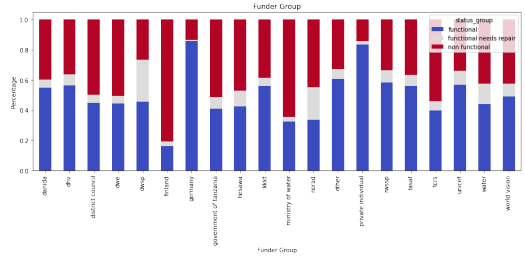
| | Model | Accuracy | CV | Precision | Recall | F1 Score | MAE | MSE | RMSE | AUC | Bias | Variance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | Imbalance Decision Tree | 75.96 | 0.72 | 0.65 | 0.66 | 0.65 | 0.27 | 0.32 | 0.57 | - | 0.01 | 0.36 |
| **1** | Imbalance Logistic Regression | 74.87 | 0.74 | 0.68 | 0.56 | 0.58 | 0.27 | 0.30 | 0.55 | 0.81 | 0.01 | 0.25 |
| **2** | Imbalance KNN | 78.96 | 0.73 | 0.70 | 0.66 | 0.68 | 0.23 | 0.27 | 0.52 | - | 0.01 | 0.32 |
| **3** | Imbalance Bagged Tree | 79.16 | 0.76 | 0.69 | 0.66 | 0.67 | 0.23 | 0.27 | 0.52 | - | -0.01 | 0.32 |
| **4** | Imbalance Random Forest | 80.37 | 0.77 | 0.71 | 0.67 | 0.69 | 0.22 | 0.26 | 0.51 | - | 0.00 | 0.32 |
| **5** | Imbalance Gradient Boost | 81.30 | 0.78 | 0.73 | 0.68 | 0.70 | 0.20 | 0.24 | 0.49 | - | 0.00 | 0.31 |
| **6** | Imbalance ADABoost | 71.80 | 0.72 | 0.60 | 0.51 | 0.50 | 0.30 | 0.32 | 0.57 | - | 0.02 | 0.22 |
| **7** | Imbalance XGBoost | 80.34 | 0.78 | 0.76 | 0.63 | 0.66 | 0.21 | 0.24 | 0.49 | - | 0.00 | 0.27 |
| **8** | Imbalance SVM | 78.54 | 0.76 | 0.75 | 0.60 | 0.63 | 0.23 | 0.26 | 0.51 | - | 0.02 | 0.25 |

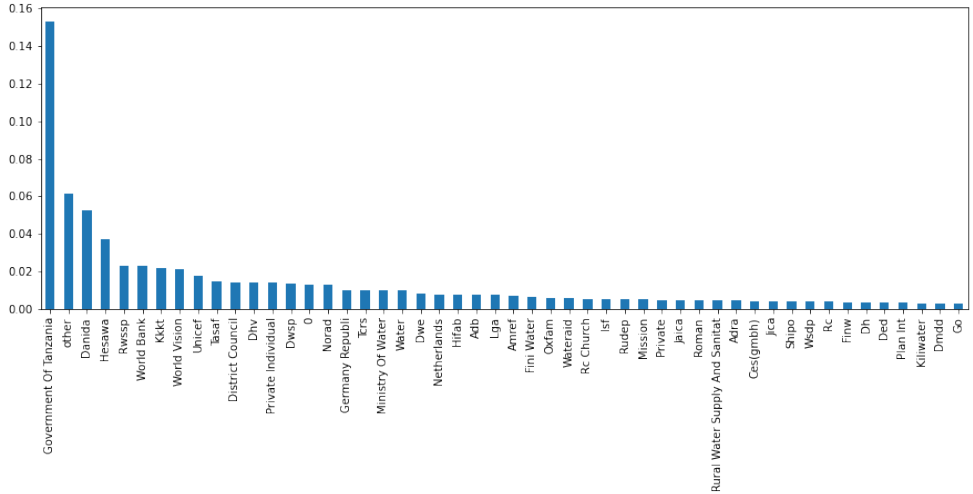
# Metrics

- **Accuracy:** Out of all the classes, how much we predicted correctly. It should be high as possible.
- **Precision:** What proportion of positive identifications was actually correct? i.e. TP / (TP + FP). It should be high as possible.
- **Recall:** What proportion of positive identifications was actually correct? i.e. TP / (TP + FN). It should be high as possible.
- **F1:** It is difficult to compare two models with low precision and high recall or vice versa. It is often convenient to combine precision and recall into a single metric called the F1 score. We want to maximize f1 score as well.
- **Cross validation:** this should be as close as possible with the model's RMSE or else we overfit our model.
- **Macro average** is the average of precision/recall/f1-score.
- **Weighted average** is just the weighted average of precision/recall/f1-score.

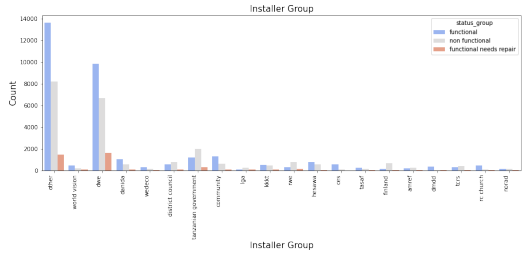
# funder

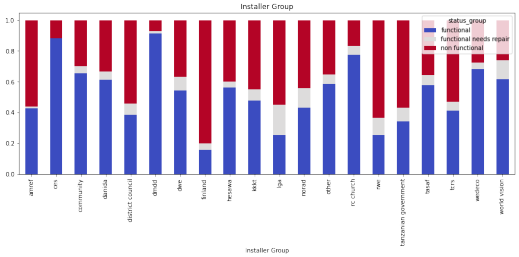

Funder Group
Count
status_group
functional
non functional
functional needs repair
other
unicef
dwsp
rwssp
danida
world vision
hesawa
government of tanzania
wateraid
private individual
district council
kkkt
norad
dwe
tcrs
germany
tasaf
finland
dhv
ministry of water
Funder Group

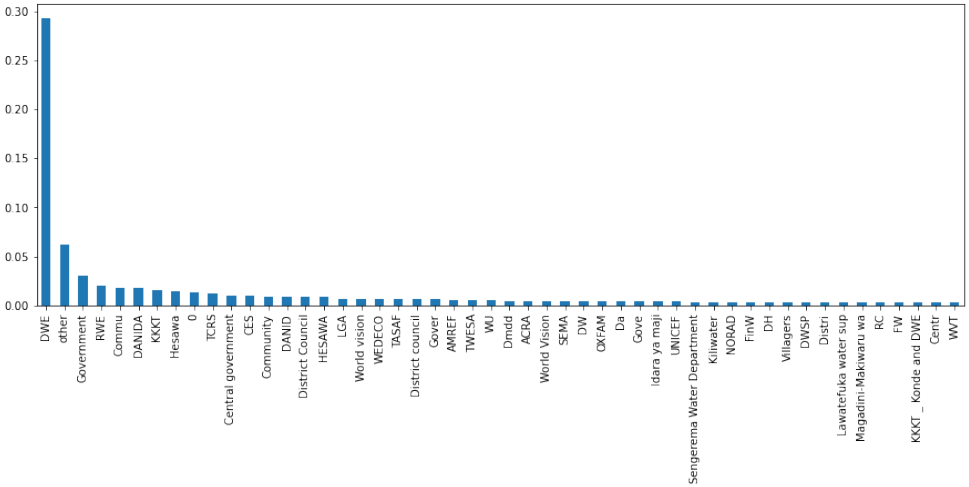

Funder Group
Percentage
status_group
functional
functional needs repair
non functional
danida
dhv
district council
dwe
dwsp
finland
germany
government of tanzania
hesawa
kkkt
ministry of water
norad
other
private individual
rwssp
tasaf
tcrs
unicef
wateraid
world vision
Funder Group


Government Of Tanzania
other
Danida
Hesawa
Rwssp
World Bank
Kkkt
World Vision
Unicef
Tasaf
District Council
Dhv
Private Individual
Dwsp
0
Norad
Germany Republi
Tcrs
Ministry Of Water
Water
Dwe
Netherlands
Hifab
Adb
Lga
Amref
Fini Water
Oxfam
Wateraid
Rc Church
Isf
Rudep
Mission
Private
Jaica
Roman
Rural Water Supply And Sanitat
Adra
Cesi(gmbh)
Jica
Shipo
Wsdp
Rc
Finw
Dh
Ded
Plan Int
Kiliwater
Dmdd
Go

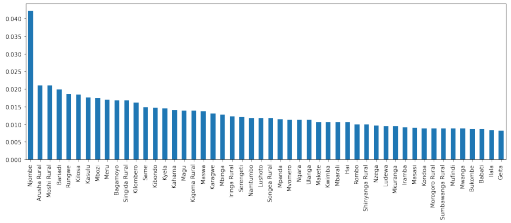
# installer

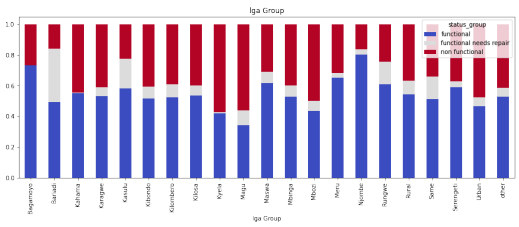

Installer Group
Count
status_group
functional
non functional
functional needs repair
Installer Group

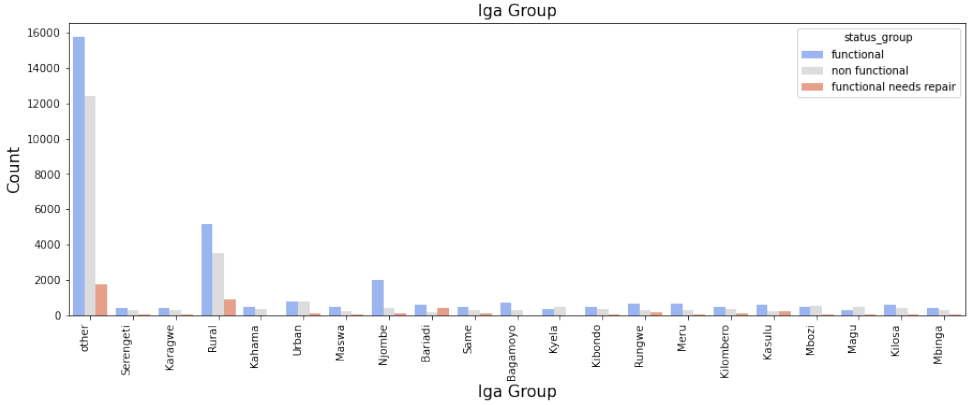

Installer Group
status_group
functional
functional needs repair
non functional
installer Group


DWE
other
Government
RWE
Commu
DANIDA
KKKT
Hesawa
0
TCRS
Central government
CES
Community
DANID
District Council
HESAWA
LGA
World vision
WEDECO
TASAF
District council
Gover
AMREF
TWESA
WU
Dmdd
ACRA
World Vision
SEMA
DW
OXFAM
Da
Gove
Idara ya maji
UNICEF
Sengerema Water Department
Kiltwater
NORAD
FinW
DH
Villagers
DWSP
Distri
Lawatefuka water sup
Magadini-Makiwaru wa
RC
FW
KKKT _ Konde and DWE
Centr
WVT

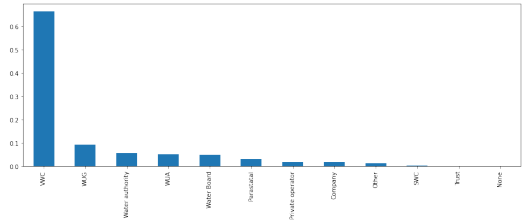
# lga

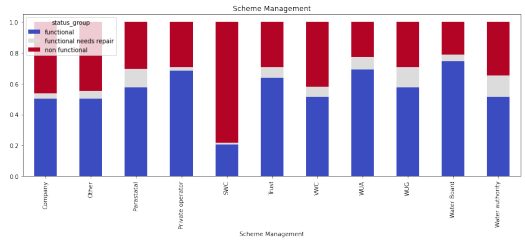


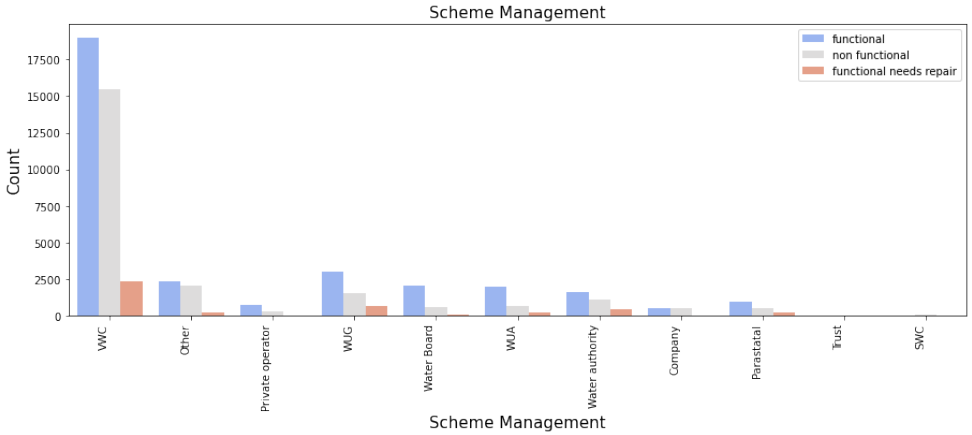

lga Group
status_group
functional
functional needs repair
non functional
lga Group


lga Group
status_group
functional
non functional
functional needs repair
Count
16000
14000
12000
10000
8000
6000
4000
2000
0
other
Serengeti
Karagwe
Rural
Kahama
Urban
Maswa
Njombe
Bariadi
Same
Bagamoyo
Kyela
Kibondo
Rungwe
Meru
Kilombero
Kasulu
Mbozi
Magu
Kilosa
Mbinga
lga Group

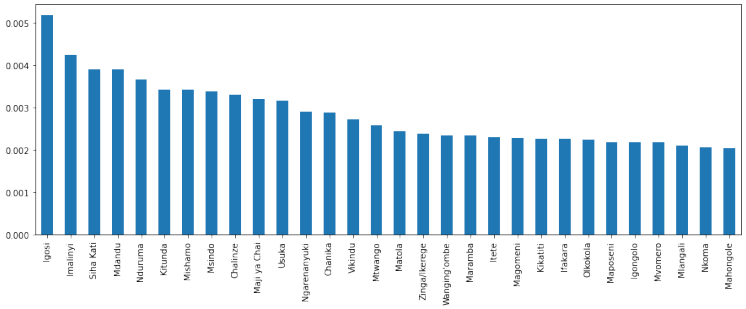
# scheme management

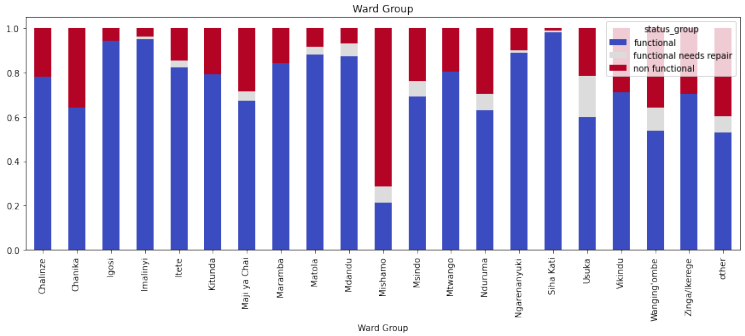


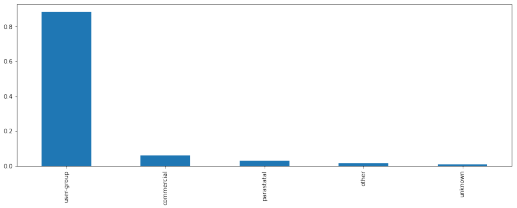



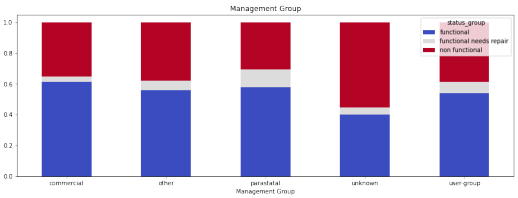
# ward

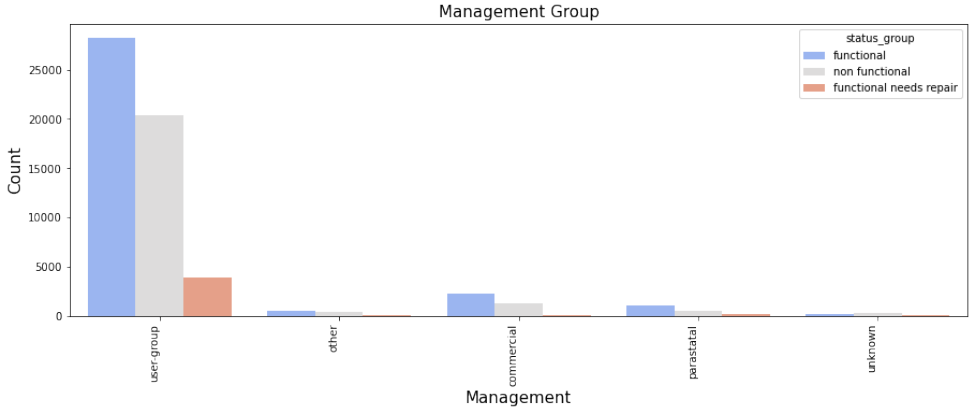

0.005
0.004
0.003
0.002
0.001
0.000
Igosi
Imalinyi
Siha Kati
Mdandu
Nduruma
Kitunda
Mishamo
Msindo
Chalinze
Maji ya Chai
Usuka
Ngarenanyuki
Chanika
Vikindu
Mtwango
Matola
Zinga/Ikerege
Wanging'ombe
Maramba
Itete
Magomeni
Kikatiti
Ifakara
Olkokola
Maposeni
Igongolo
Mvomero
Mlangali
Nkoma
Mahongole


Ward Group
1.0
0.8
0.6
0.4
0.2
0.0
status_group
functional
functional needs repair
non functional
Chalinze
Chanika
Igosi
Imalinyi
Itete
Kitunda
Maji ya Chai
Maramba
Matola
Mdandu
Mishamo
Msindo
Mtwango
Nduruma
Ngarenanyuki
Siha Kati
Usuka
Vikindu
Wanging'ombe
Zinga/Ikerege
other
Ward Group

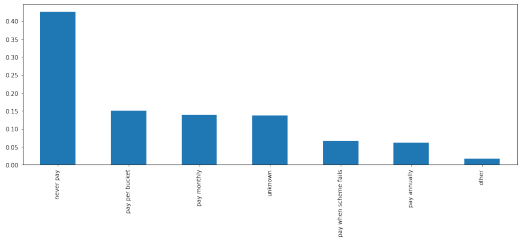
# management group

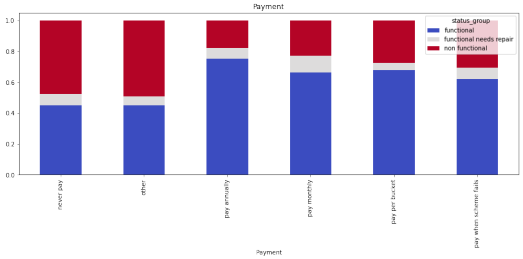


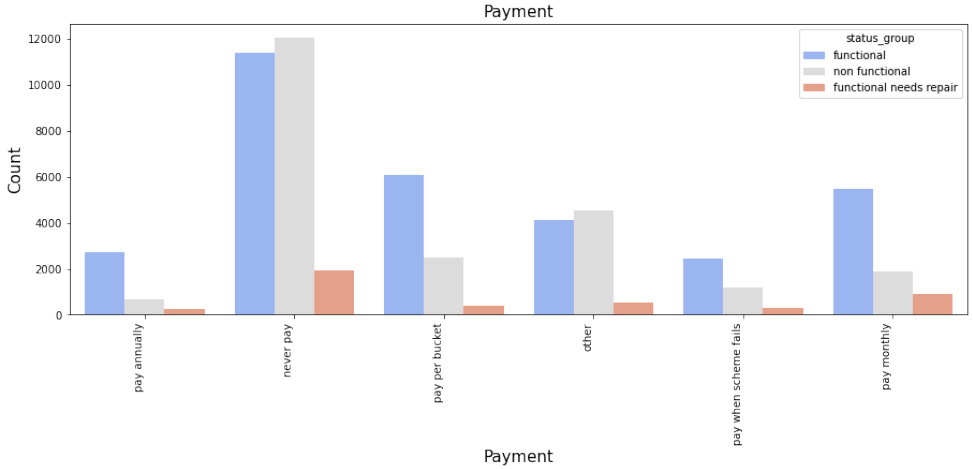



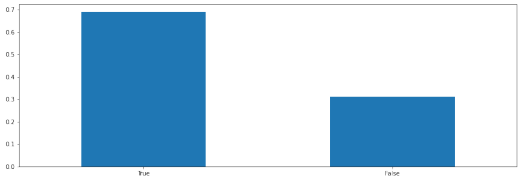
# payment

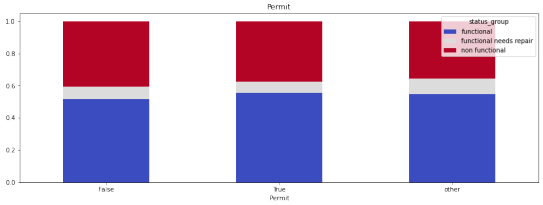

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
never pay
pay per bucket
pay monthly
unknown
pay when scheme fails
pay annually
other

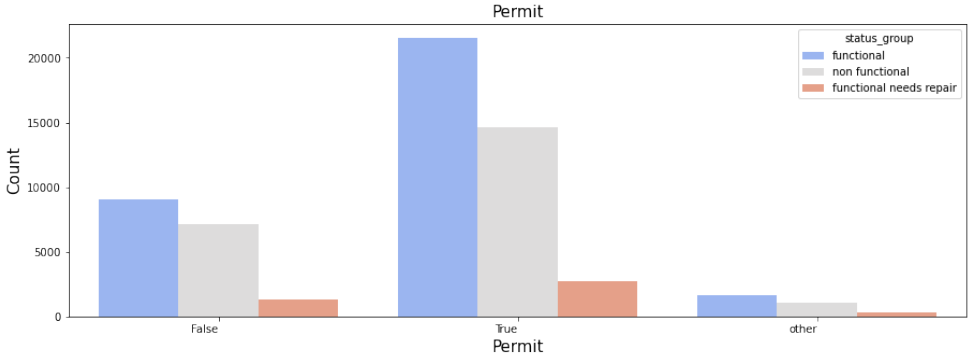

Payment
status_group
functional
functional needs repair
non functional
1.0
0.8
0.6
0.4
0.2
0.0
never pay
other
pay annually
pay monthly
pay per bucket
pay when scheme fails
Payment


Payment
status_group
functional
non functional
functional needs repair
Count
12000
10000
8000
6000
4000
2000
0
pay annually
never pay
pay per bucket
other
pay when scheme fails
pay monthly
Payment

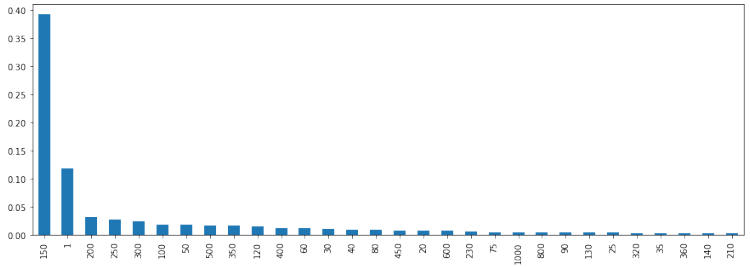
# permit

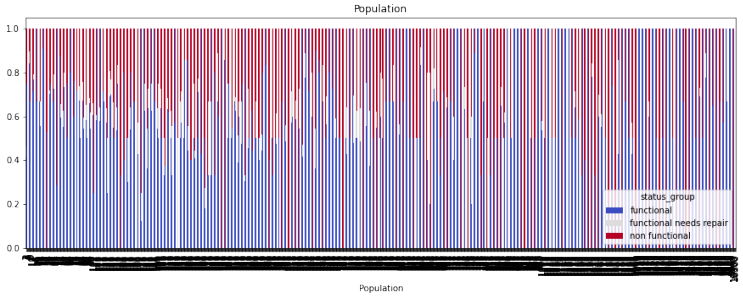

0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
True
False


Permit
status_group
functional
functional needs repair
non functional
1.0
0.8
0.6
0.4
0.2
0.0
False
True
other
Permit

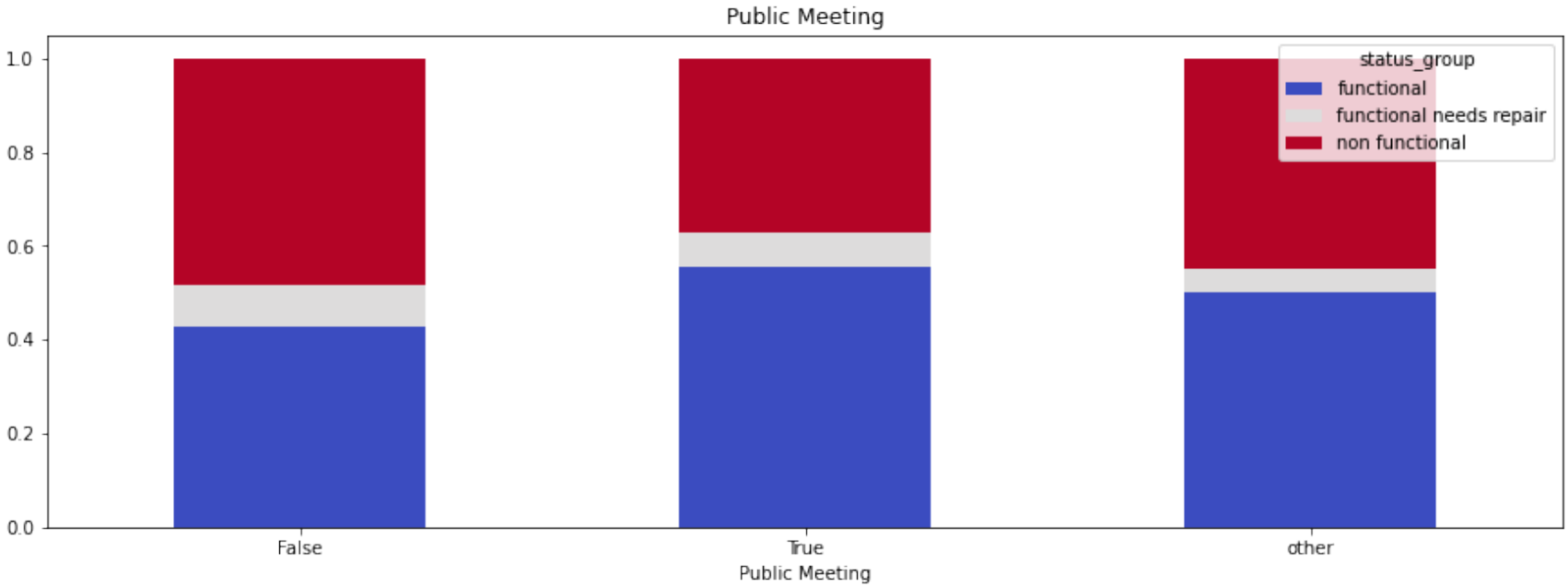

Permit
status_group
functional
non functional
functional needs repair
Count
20000
15000
10000
5000
0
False
True
other
Permit

# population

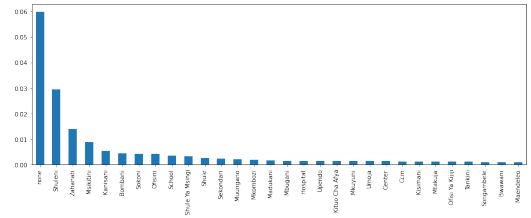

0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
150
1
200
250
300
100
50
500
350
120
400
60
30
40
80
450
20
600
230
75
1000
800
90
130
25
320
35
360
140
210

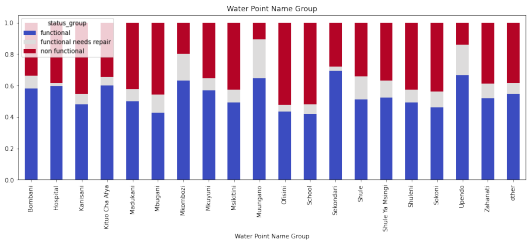

Population
1.0
0.8
0.6
0.4
0.2
0.0
status_group
functional
functional needs repair
non functional
Population

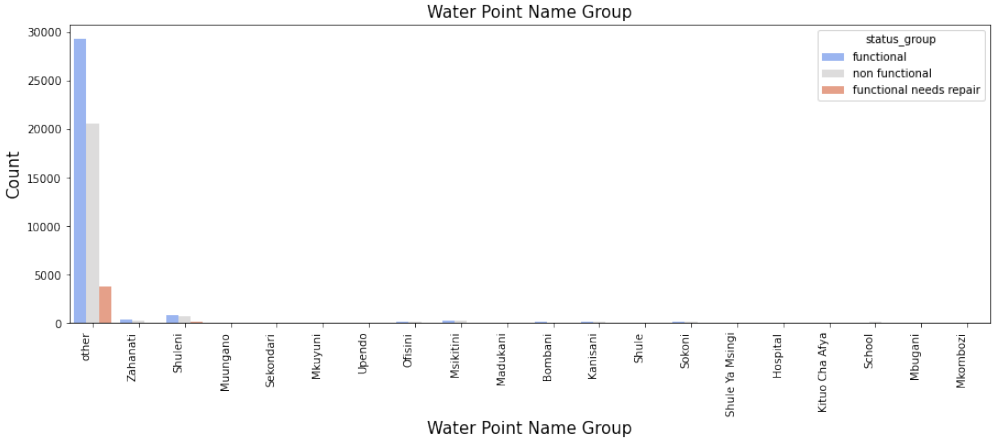
# public meeting



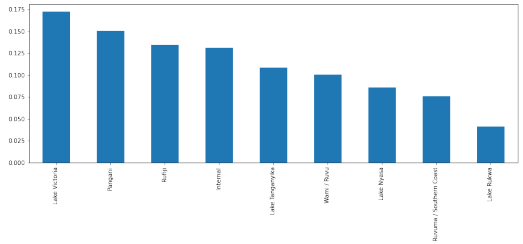
# water point names

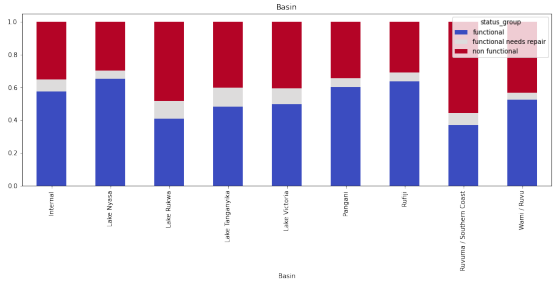


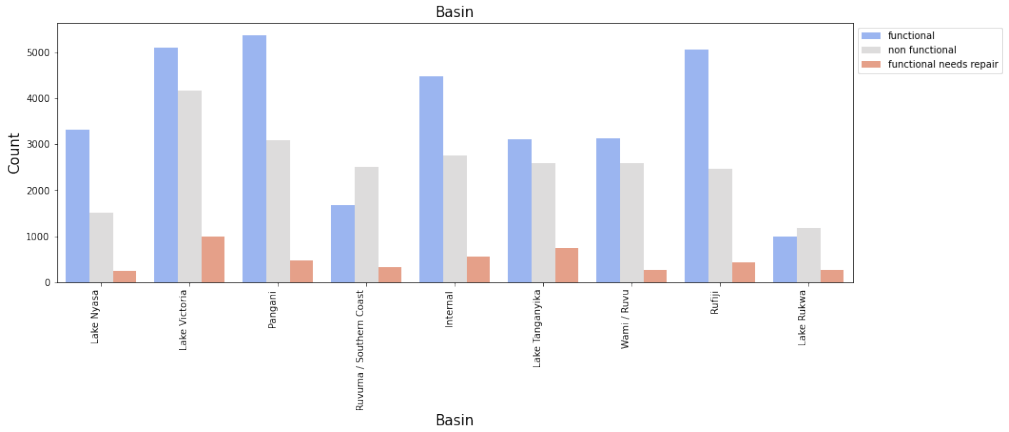


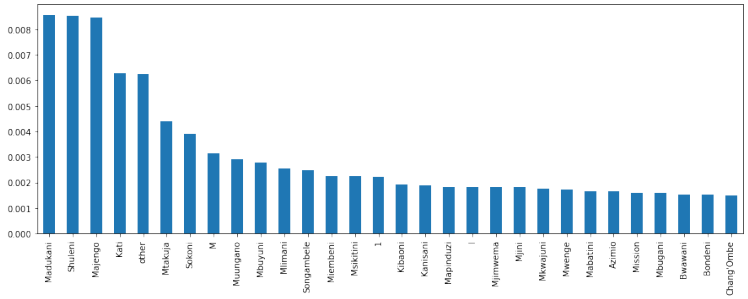
# basin

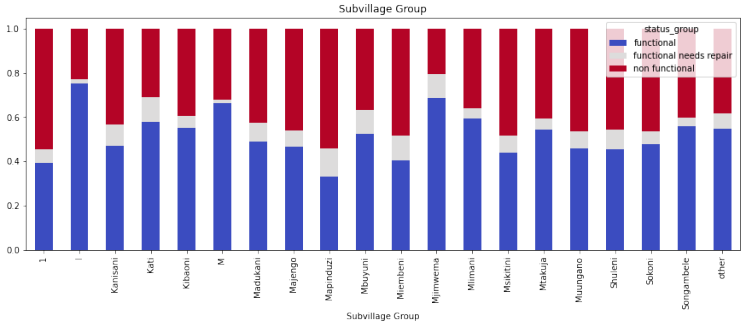



Basin
status_group
functional
functional needs repair
non functional
Basin

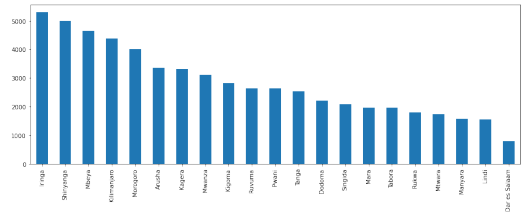

Basin
functional
non functional
functional needs repair
Count
5000
4000
3000
2000
1000
0
Lake Nyasa
Lake Victoria
Pangani
Ruvuma / Southern Coast
Internal
Lake Tanganyika
Wami / Ruvu
Rufiji
Lake Rukwa
Basin

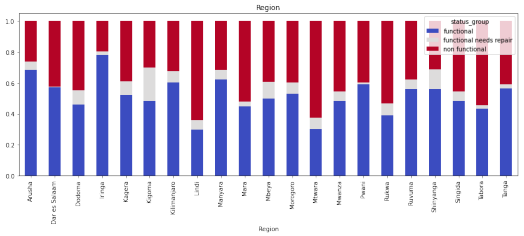
# subvillage

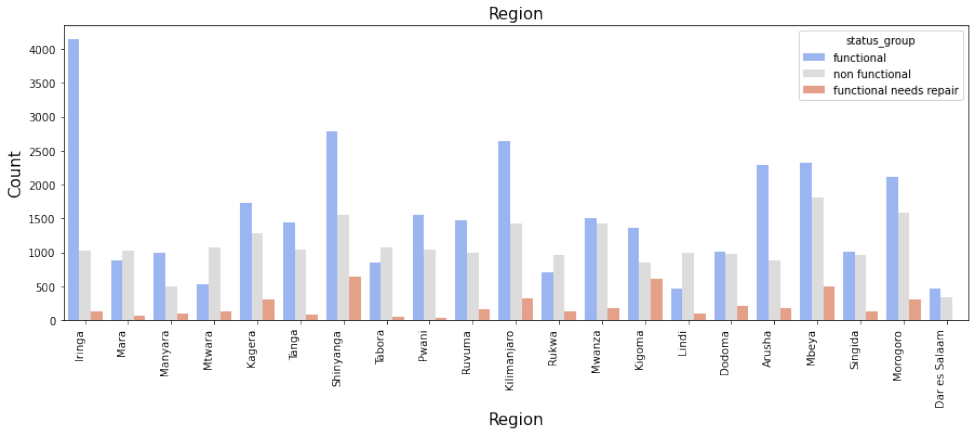

0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0.000
Madukani
Shuleni
Majengo
Kati
other
Mtakuja
Sokoni
M
Muungano
Mbuyuni
Mlimani
Songambele
Miembeni
Msikitini
1
Kibaoni
Kamsani
Mapinduzi
I
Mjimwema
Mjini
Mkwajuni
Mwenge
Mabatini
Azimio
Mission
Mbugani
Bwawani
Bondeni
Chang'Ombe


Subvillage Group
1.0
0.8
0.6
0.4
0.2
0.0
status_group
functional
functional needs repair
non functional
1
I
Kamsani
Kati
Kibaoni
M
Madukani
Majengo
Mapinduzi
Mbuyuni
Miembeni
Mjimwema
Mlimani
Msikitini
Mtakuja
Muungano
Shuleni
Sokoni
Songambele
other
Subvillage Group

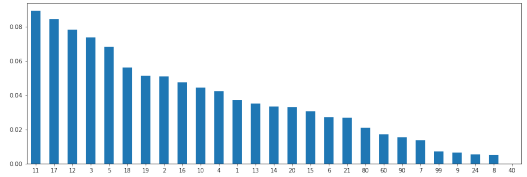
# region

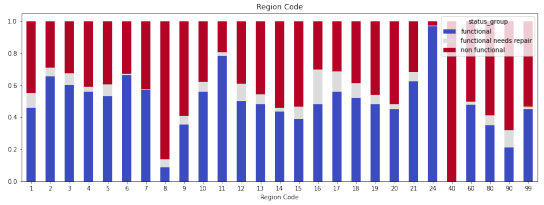


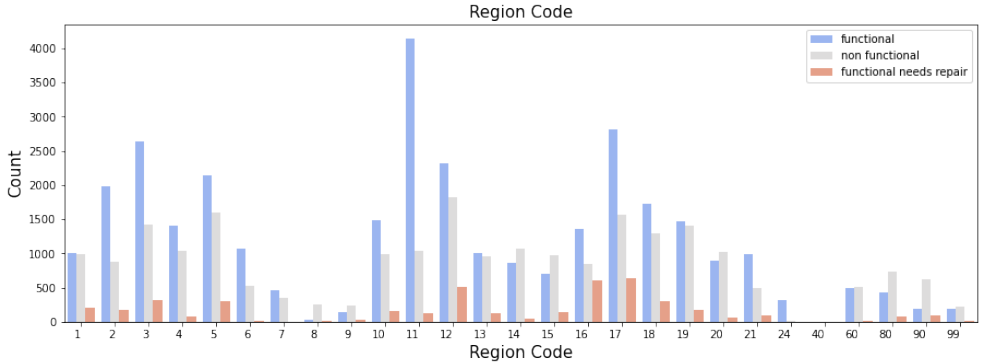

Region
status_group
functional
functional needs repair
non functional
Region


Region
status_group
functional
non functional
functional needs repair
Count
0
500
1000
1500
2000
2500
3000
3500
4000
Iringa
Mara
Manyara
Mtwara
Kagera
Tanga
Shinyanga
Tabora
Pwani
Ruvuma
Kilimanjaro
Rukwa
Mwanza
Kigoma
Lindi
Dodoma
Arusha
Mbeya
Singida
Morogoro
Dar es Salaam
Region

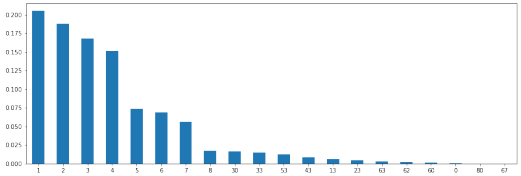
# region code

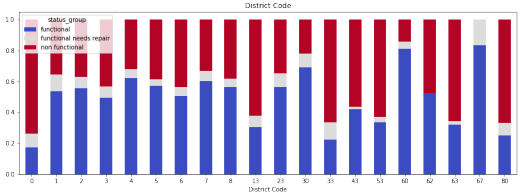


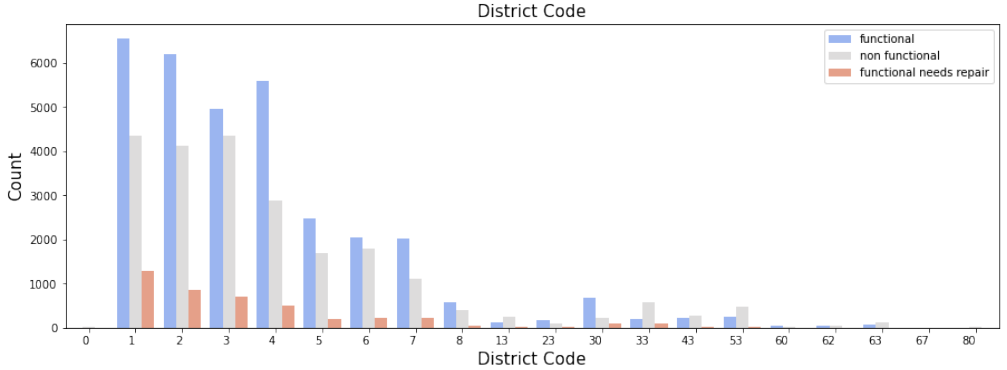

Region Code
status_group
functional
functional needs repair
non functional
Region Code


Region Code
functional
non functional
functional needs repair
Count
Region Code

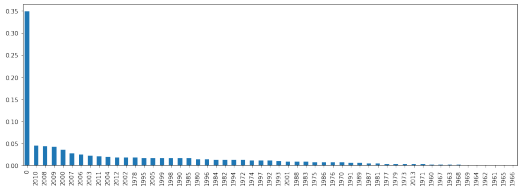
# district code

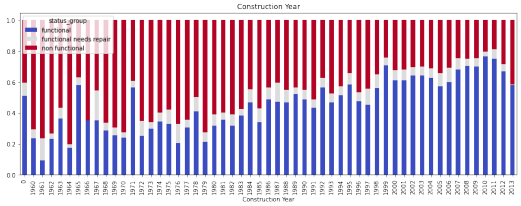


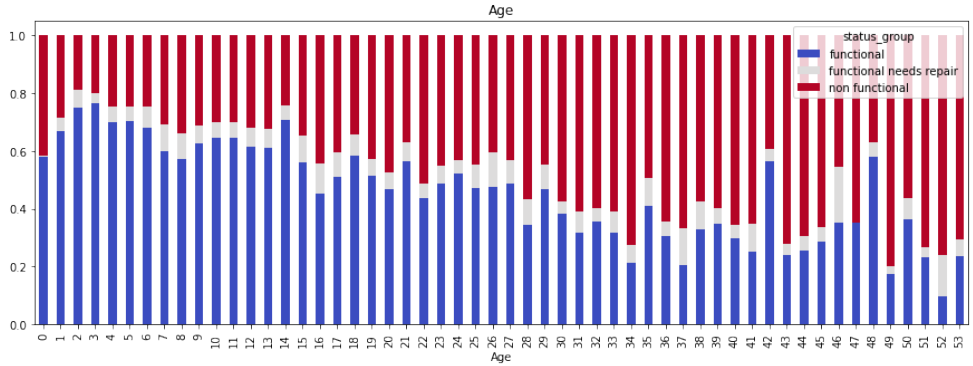


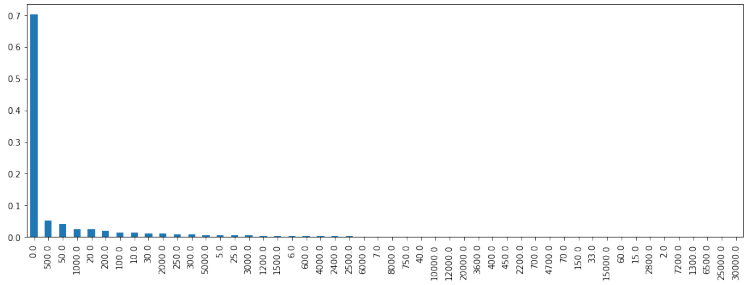
# age

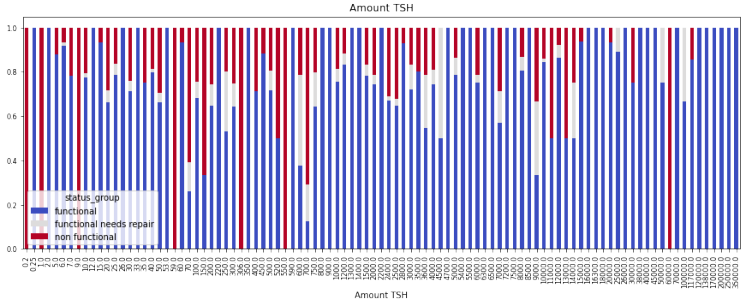



Construction Year
status_group
functional
functional needs repair
non functional
Construction Year

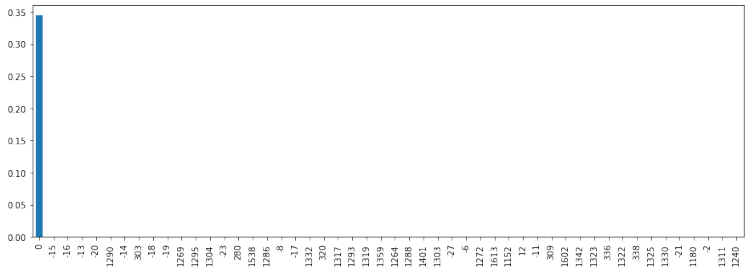

Age
status_group
functional
functional needs repair
non functional
Age

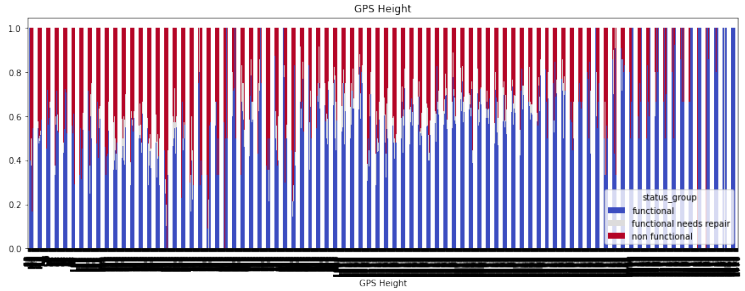
# amount tsh



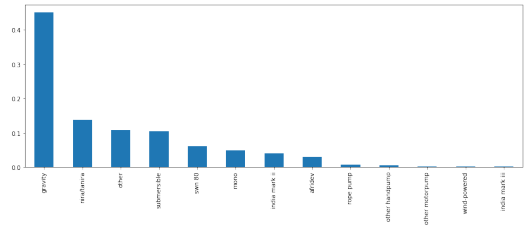

Amount TSH
status_group
functional
functional needs repair
non functional
Amount TSH

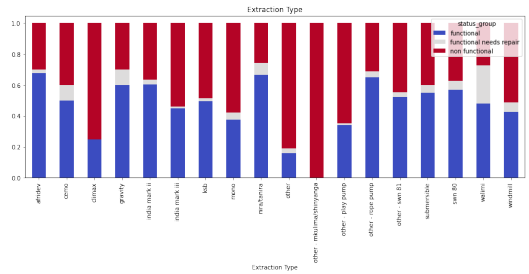
# gps height

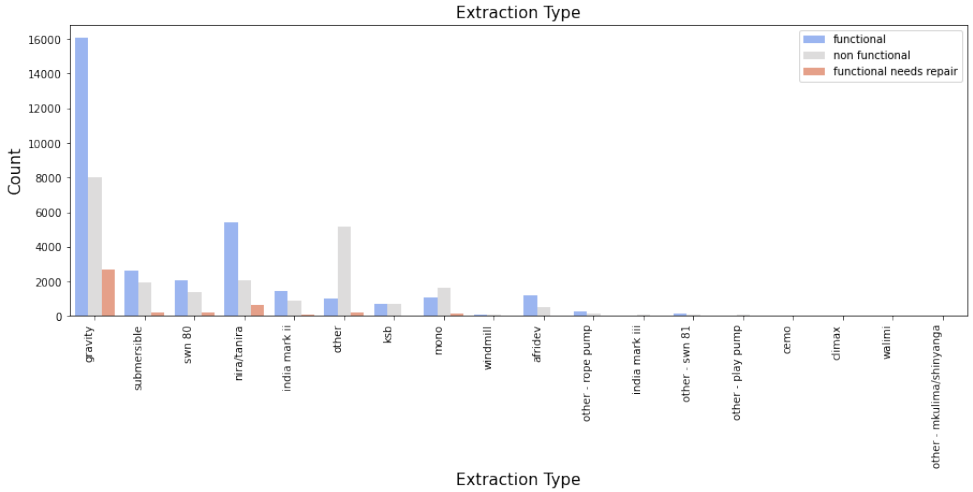

0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
-15
-16
-13
-20
1290
-14
303
-18
-19
1269
1295
1304
-23
280
1538
1286
-8
-17
1332
320
1317
1293
1319
1359
1264
1288
1401
1303
-27
-6
1272
1613
1152
12
-11
309
1602
1342
1323
336
1322
338
1325
1330
21
1180
-2
1311
1240


GPS Height
1.0
0.8
0.6
0.4
0.2
0.0
status_group
functional
functional needs repair
non functional
GPS Height

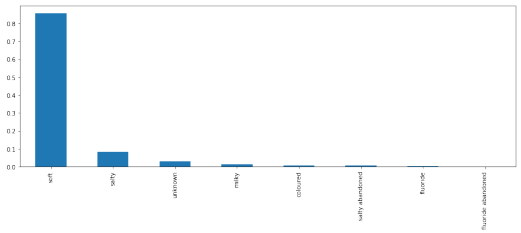
# extraction type

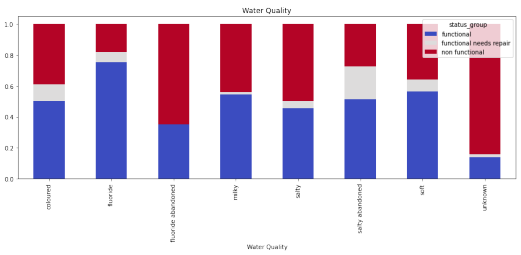


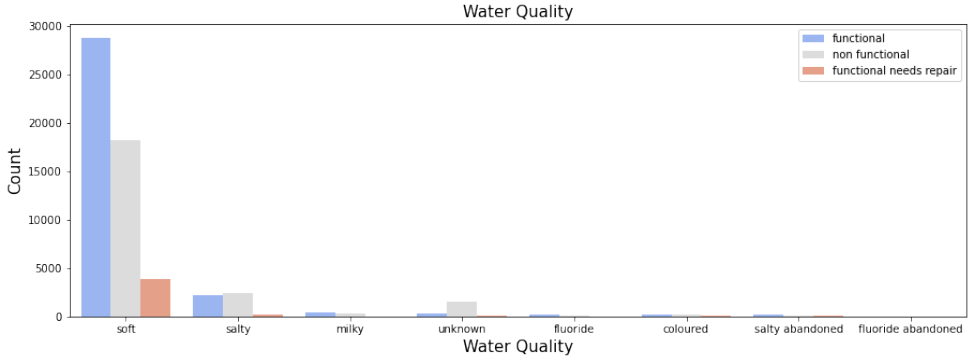


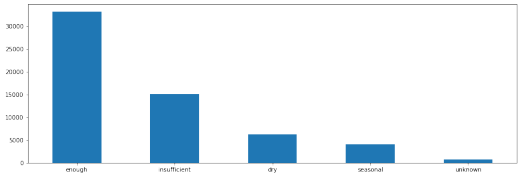
# water quality

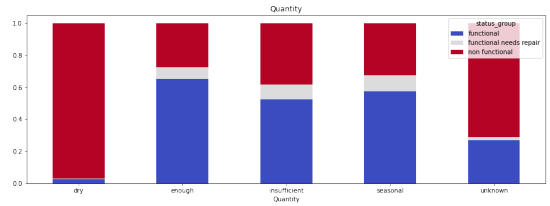


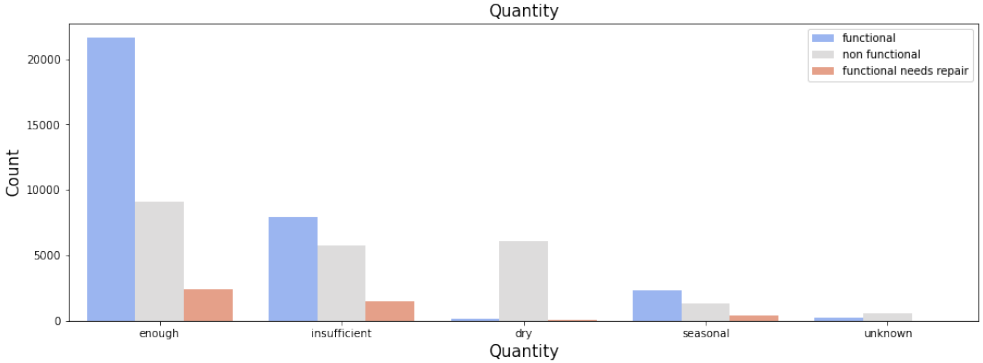



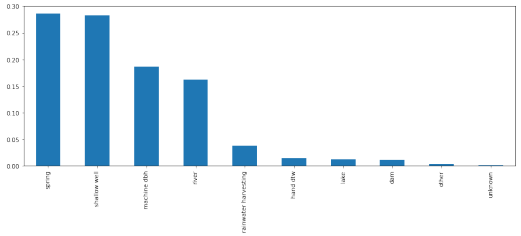
# quantity

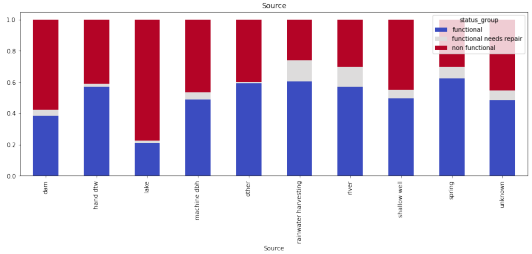

30000
25000
20000
15000
10000
5000
0
enough
insufficient
dry
seasonal
unknown

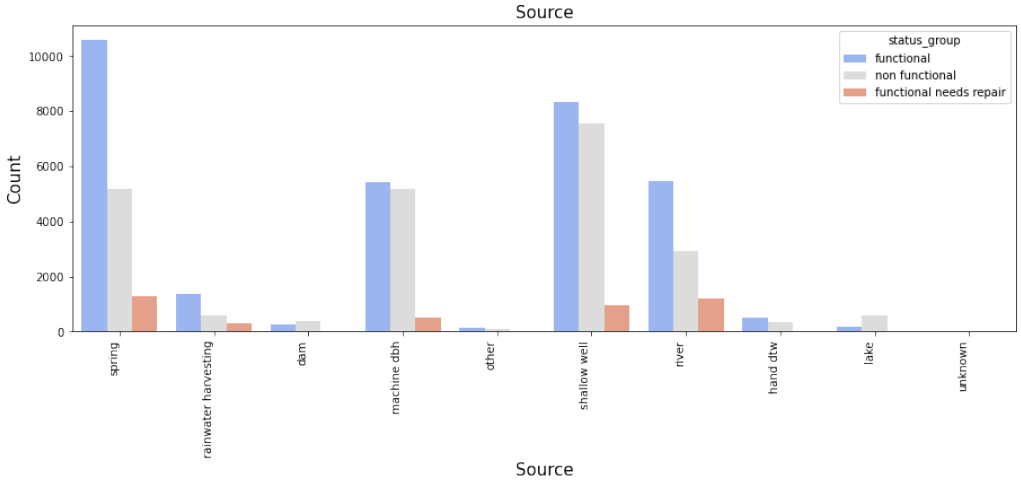

Quantity
status_group
functional
functional needs repair
non functional
1.0
0.8
0.6
0.4
0.2
0.0
dry
enough
insufficient
seasonal
unknown
Quantity


Quantity
functional
non functional
functional needs repair
Count
20000
15000
10000
5000
0
enough
insufficient
dry
seasonal
unknown
Quantity

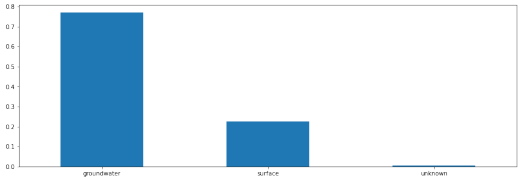
# source

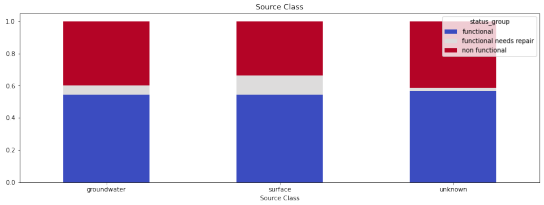

0.30
0.25
0.20
0.15
0.10
0.05
0.00
spring
shallow well
machine dbh
river
rainwater harvesting
hand dtw
lake
dam
other
unknown

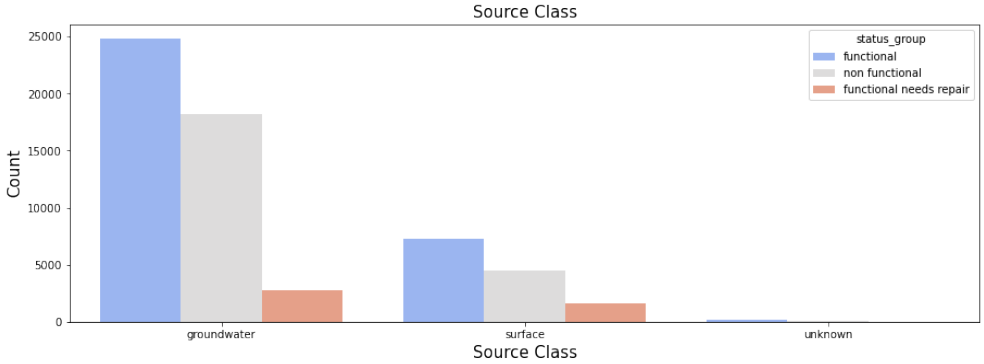

Source
status_group
functional
functional needs repair
non functional
1.0
0.8
0.6
0.4
0.2
0.0
dam
hand dtw
lake
machine dbh
other
rainwater harvesting
river
shallow well
spring
unknown
Source


Source
status_group
functional
non functional
functional needs repair
Count
10000
8000
6000
4000
2000
0
spring
rainwater harvesting
dam
machine dbh
other
shallow well
river
hand dtw
lake
unknown
Source

# source class



# water point type