TANZANIA WATER WELLS PROJECT

BRIAN NDERU

BUSINESS UNDERSTANDING

Business Overview

Tanzania faces significant challenges in providing clean water to its citizens. The existing water points in the country play a crucial role in meeting the water needs of the population.Only 30.6% of Tanzanian households use Piped water systems with almost 70% of the population using Water wells. However, many of these wells are in dire need of repair, and some have even failed entirely, thus the water crisis.

Problem Statement

Water shortage is a serious issue in Tanzania. It is therefore crucial to identify the category in which Water wells in the country are for repair either by the Ministry of Water in Tanzania, NGOs or Private Companies. Using Data from Ministry of Water in Tanzania, we will help categorize the Condition of the Water wells in the country to help them achieve their Goal

Objective

To build a classifier model for the water wells in Tanzania

Metric of Sucess

The model to have an accuracy of Above 70%

Business Questions

Who are the top 10 financiers of Water Well repairs in Tanzania

Who are the top 5 Installers of water wells in Tanzania

What is the quality of the Water in Tanzania

What is the common water source for Residents

What is the water basin from which water wells get water

What is the water wells number per region

DATA UNDERSTANDING AND PREPARATION

The dataset is obtained from The government off Tanzania

The dataset has Train and label csv which are merged to form one dataframe

The dataframe has 59400 rows and 42 columns

Some unusful columns are dropped

Missing Values

The datset has 3 columns with missing values.

The columns are installer(6.15%), funder(6.11%) and permit(5.11%)

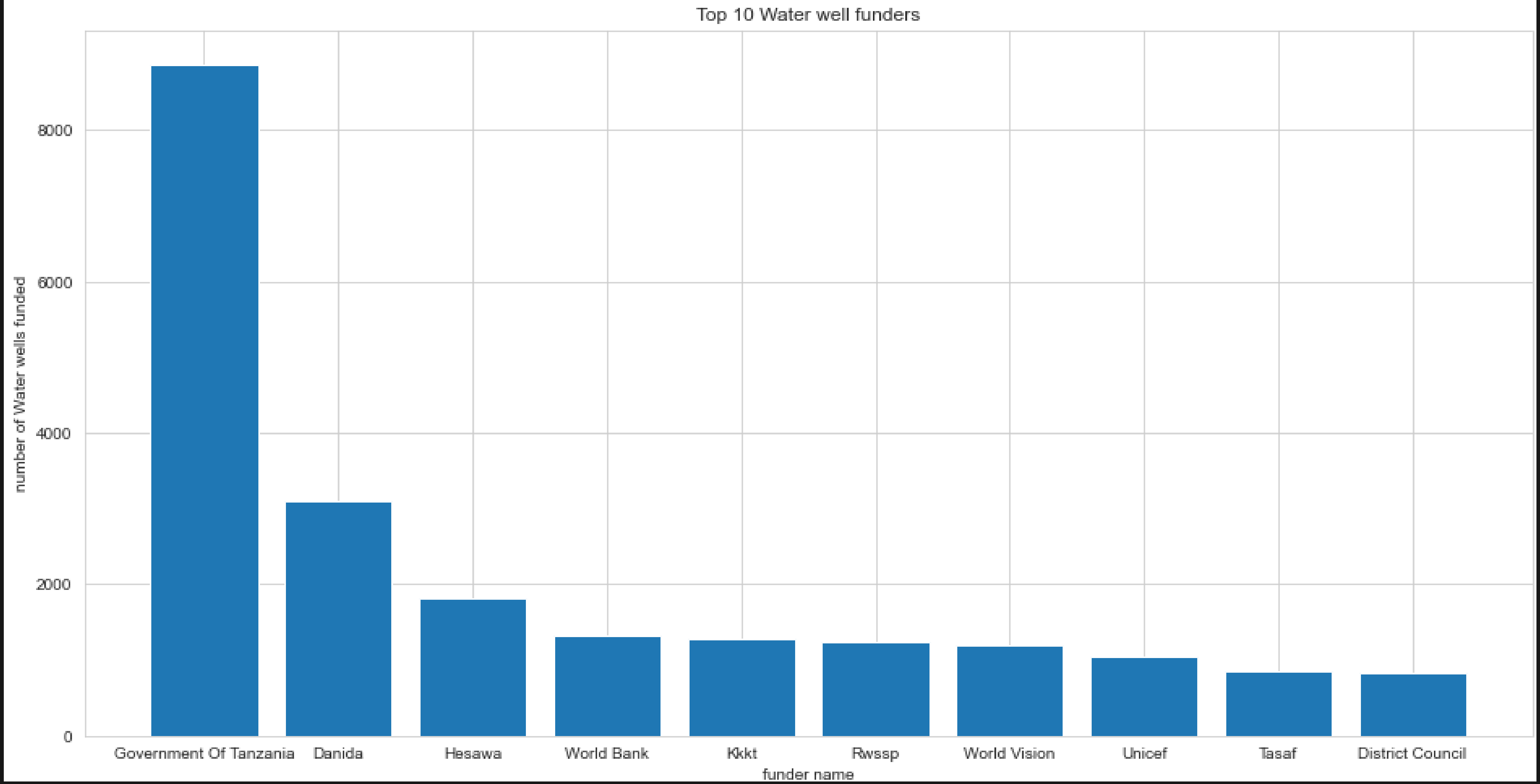
Since the percentages were low for missing values, the rows with missing values were dropped

Duplicates

There were 1180 duplicated rows which were dropped since they were a small portion

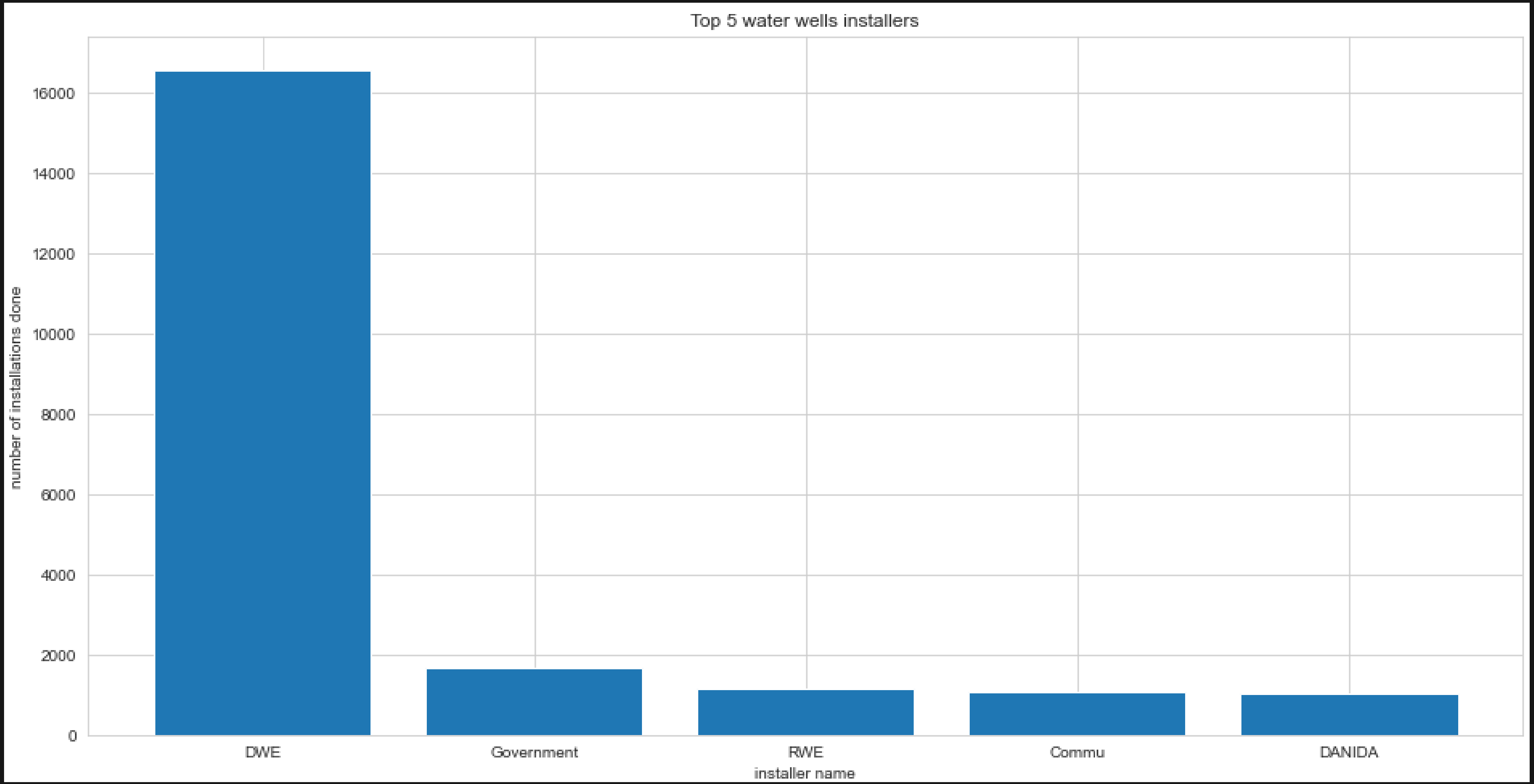
DATA EXPLORATION

Top 10 Funders



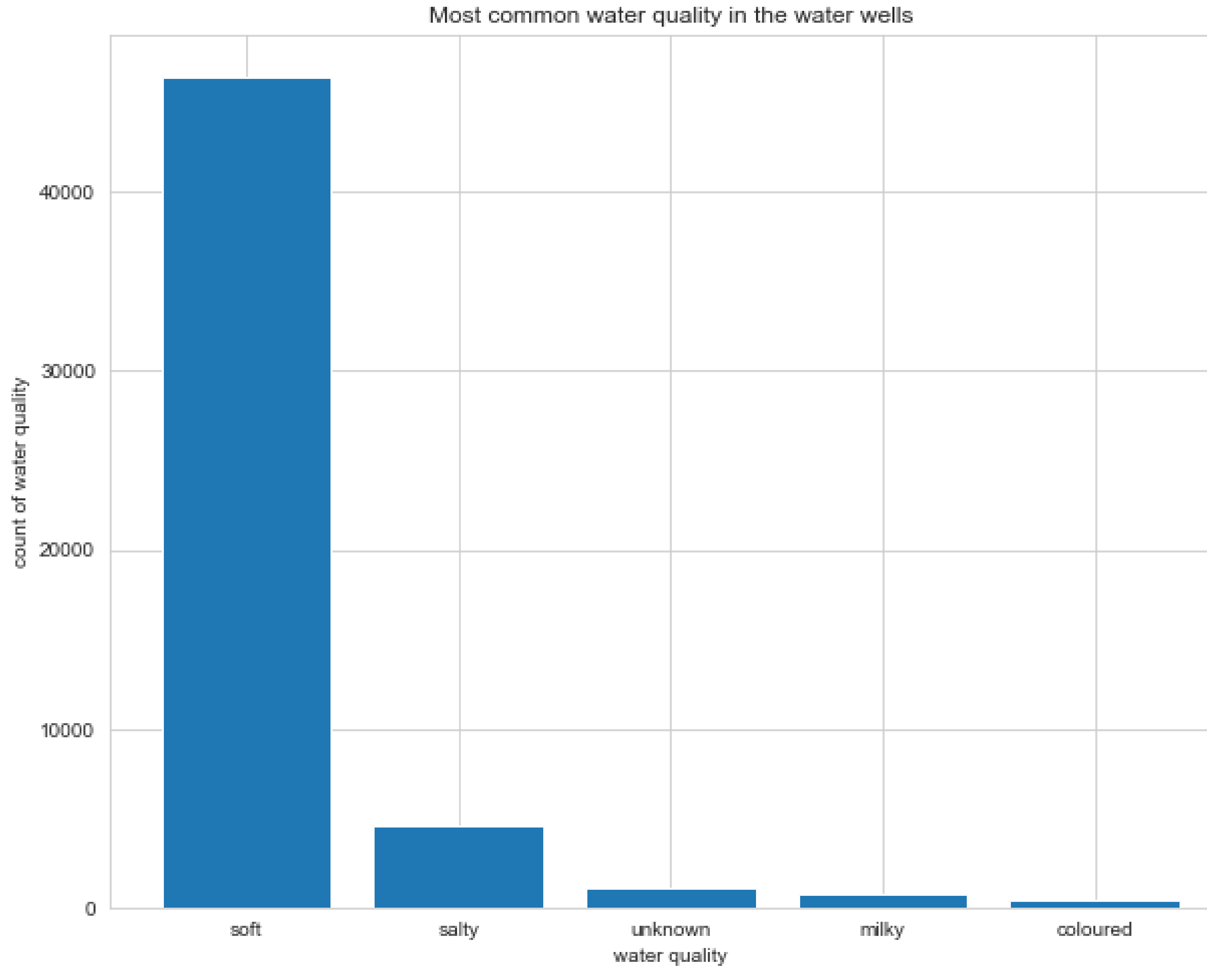
Government of Tanzania, Dandida and Hesawa are the top Funders

Top Wells Installers



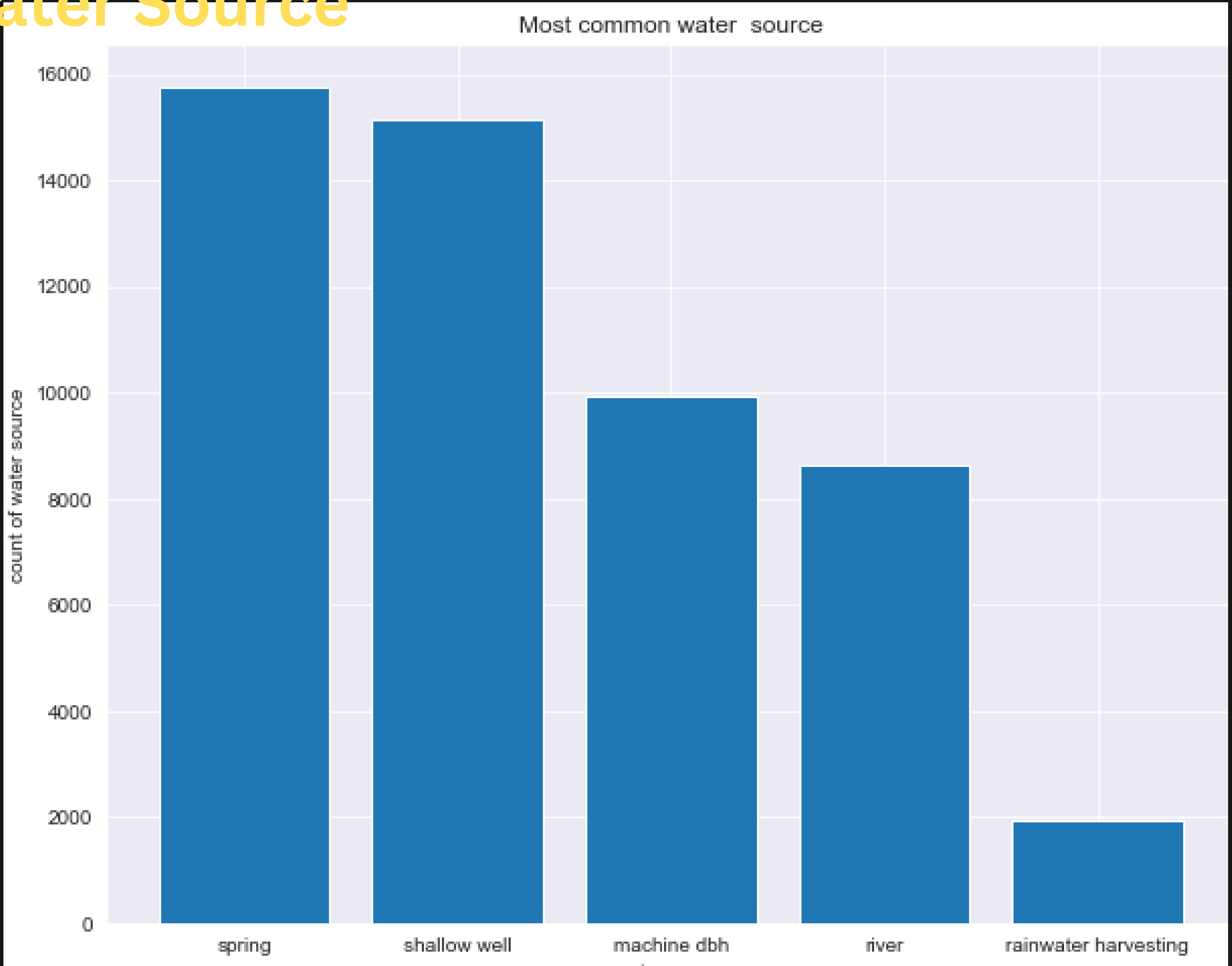
DWE are the most sought after installers for water wells with over 16,000 water wells installed

Water Quality



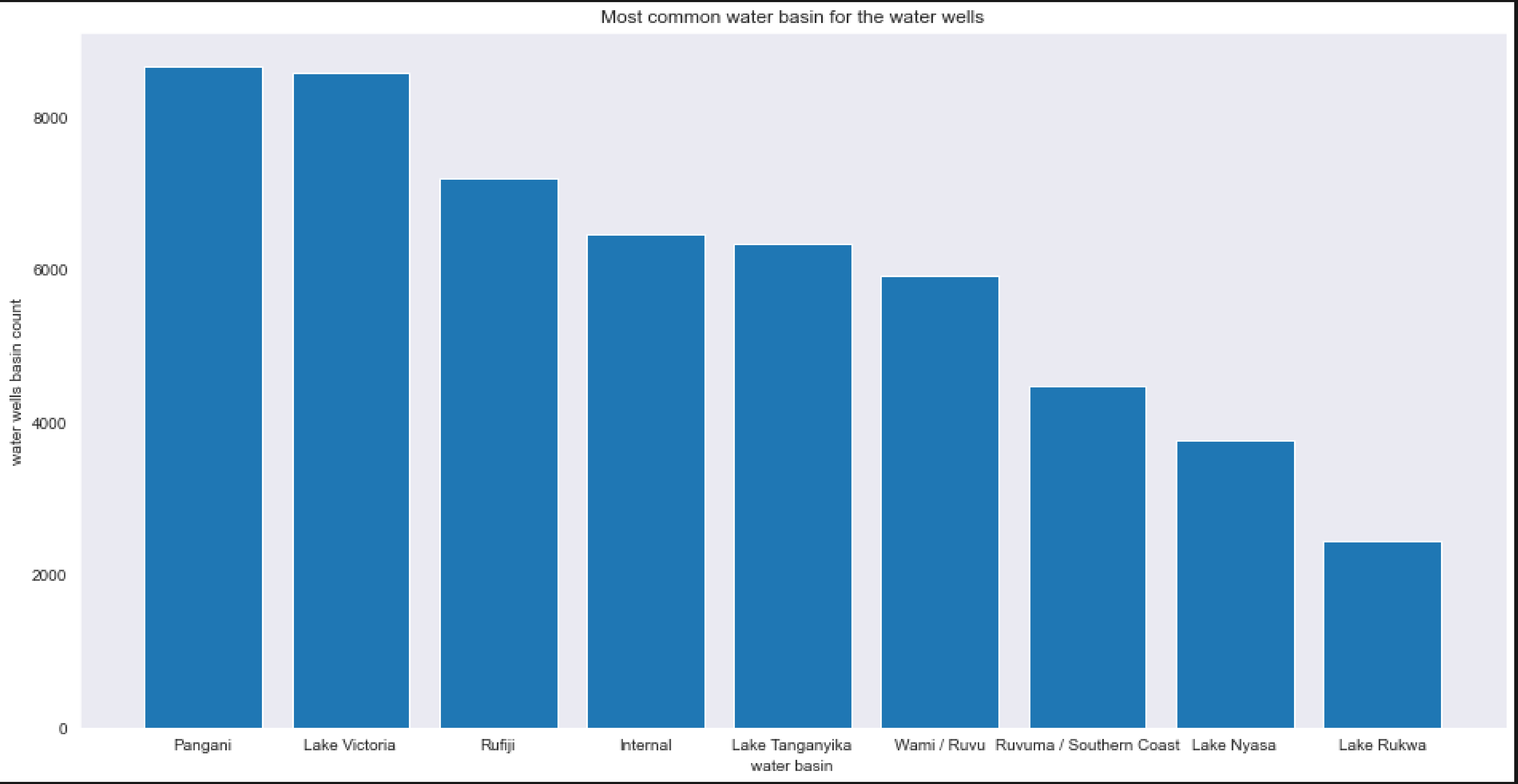
soft is most common water quality

Water Source



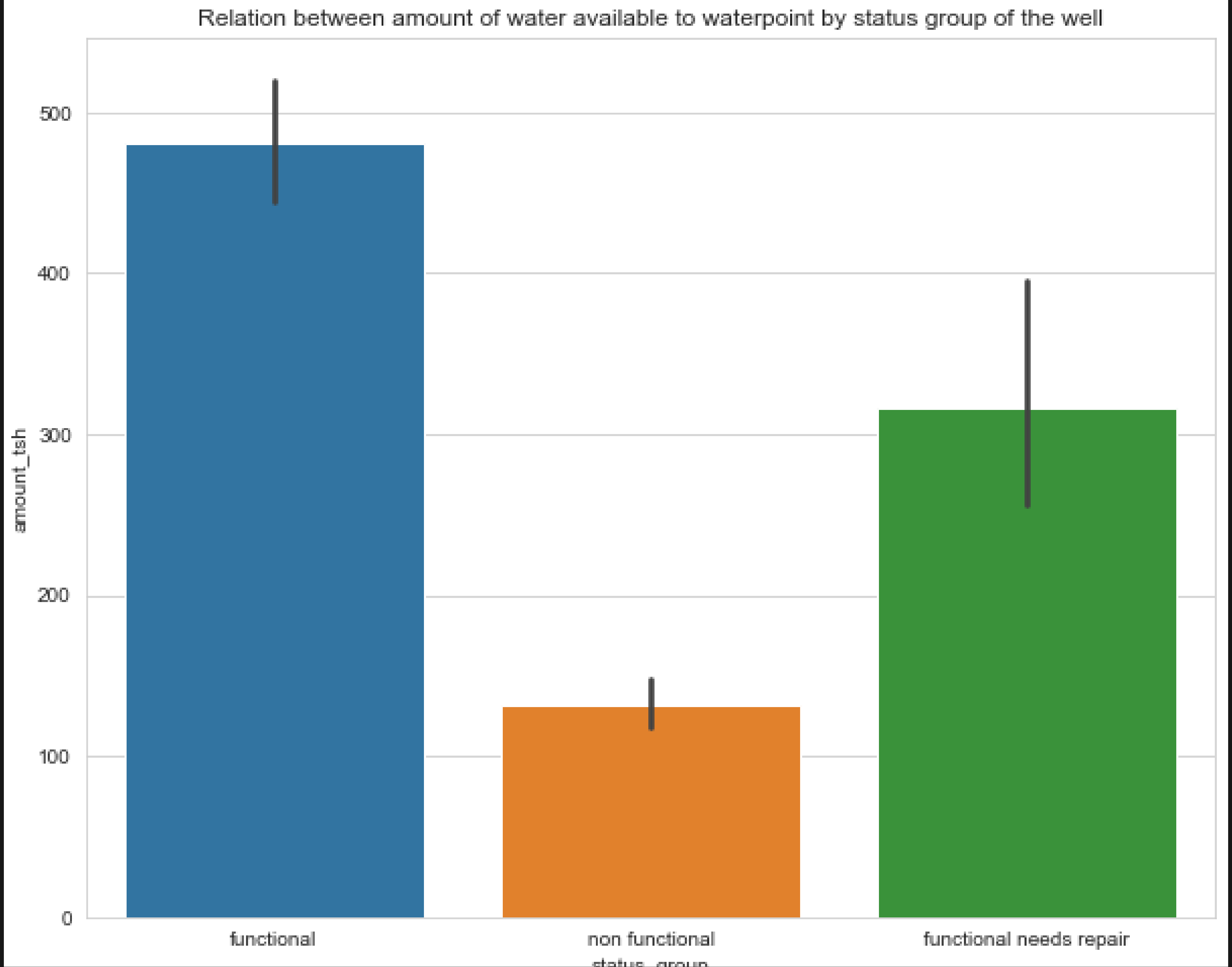
Spring and shallow well are the most common water sources in Tanzania

Water basin

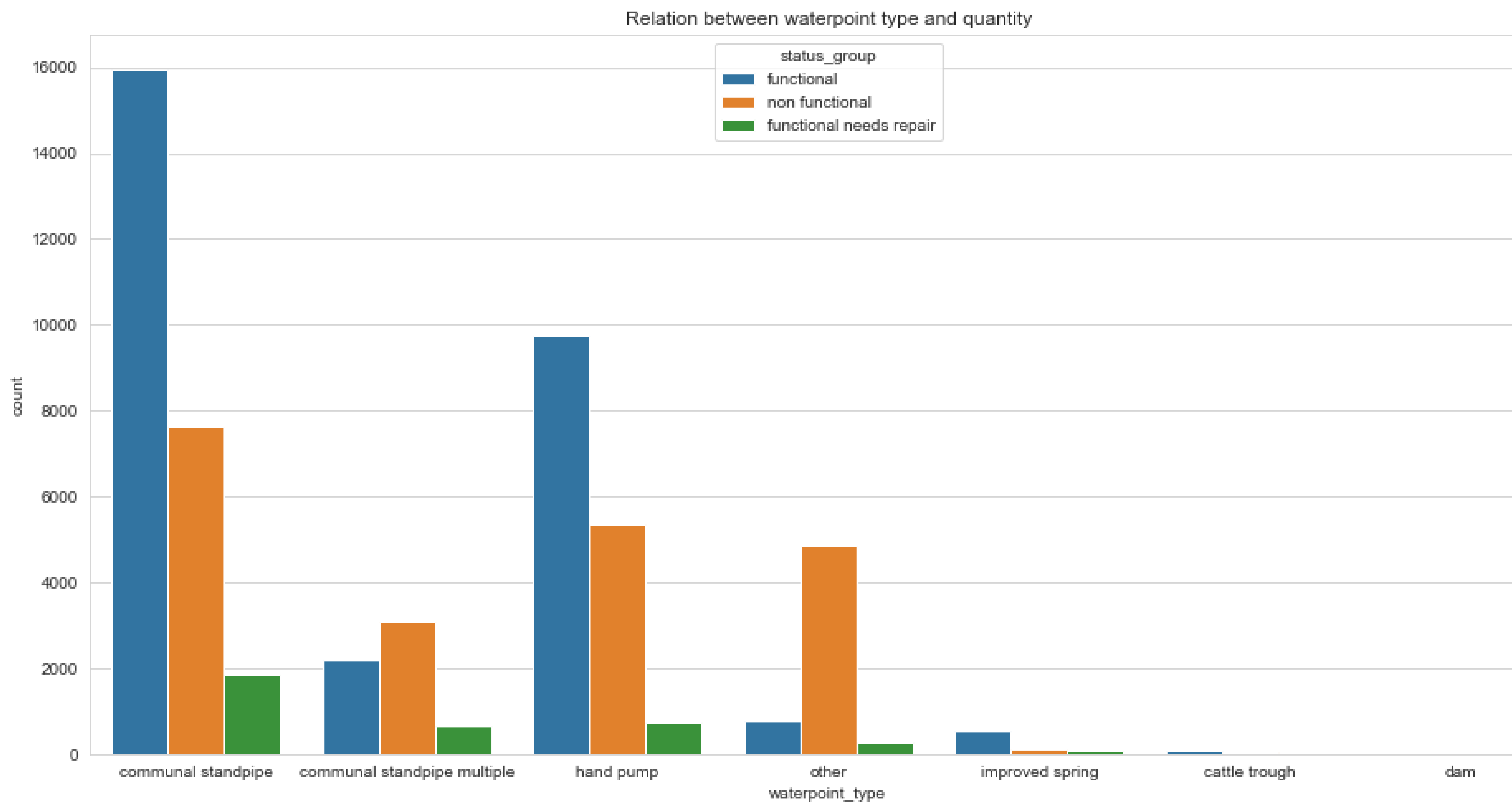


Pangani and l.Victoria are the most common water basin for Water wells

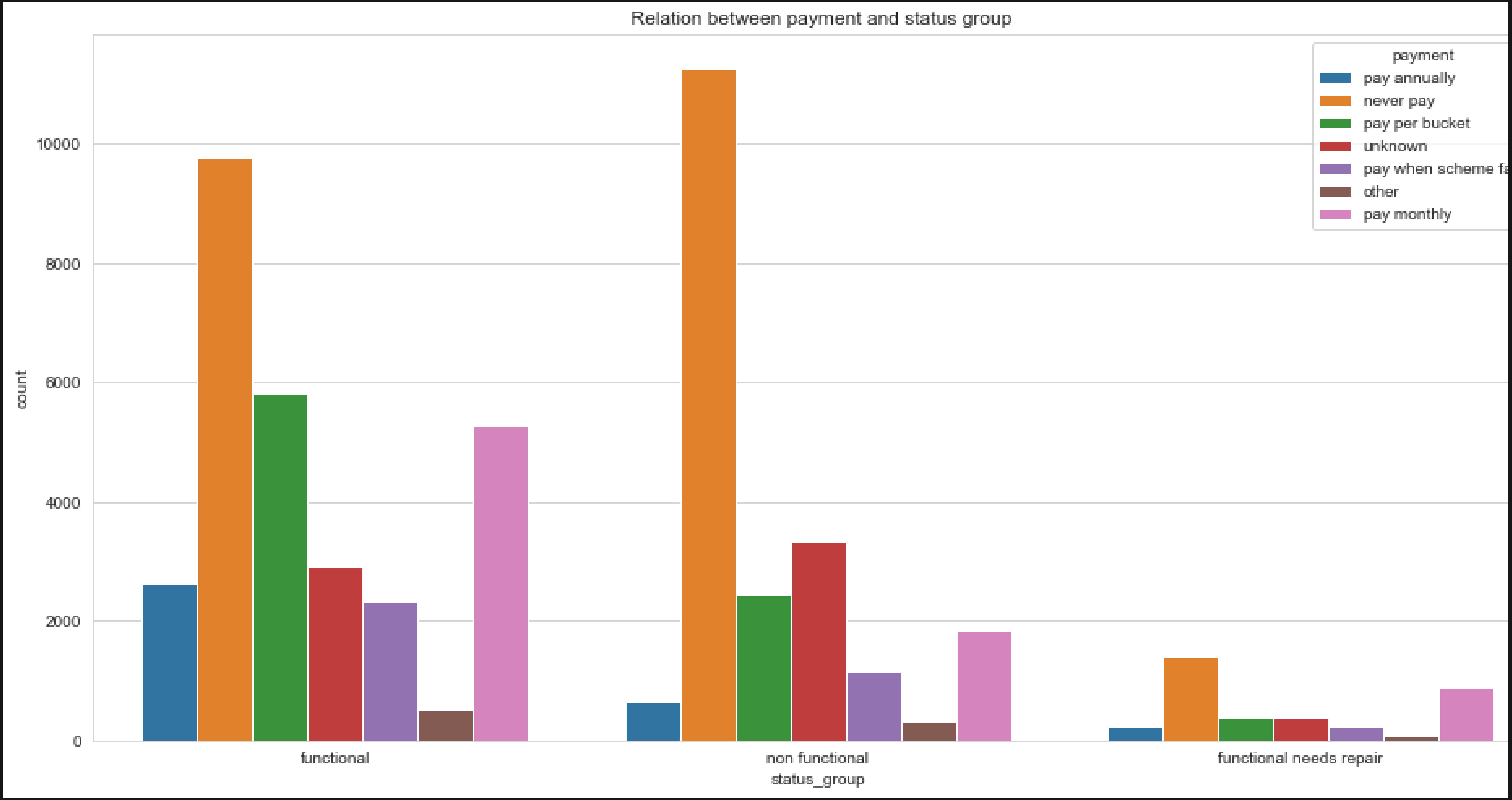
Water available to status group



Waterpoint type and quantity



Payment and status group



DATA PREPROCESSING

Categorical data were converted to numeric through One hot encoding

Scaling was conducted on the data

Label encoding was conducted on the data

SMOTE was carried out on the data to prevent class imbalance

MODELLING

Random Forest

Base model

Train Accuracy	Test Accuacy
100%	69%

Tuning n_estimators [200,300], min_sample_leaf[1], max_depth[20,30]

Tuned model

Train Accuracy	Test Accuacy
85%	67%

Logistics Model

Base model

Train Accuracy	Test Accuacy
63%	63%

Tuning	C : 01 "solver" : lbfgs, newton-cg]

Tuned model

Train Accuracy	Test Accuacy
63%	63%

K-Neighbors Model

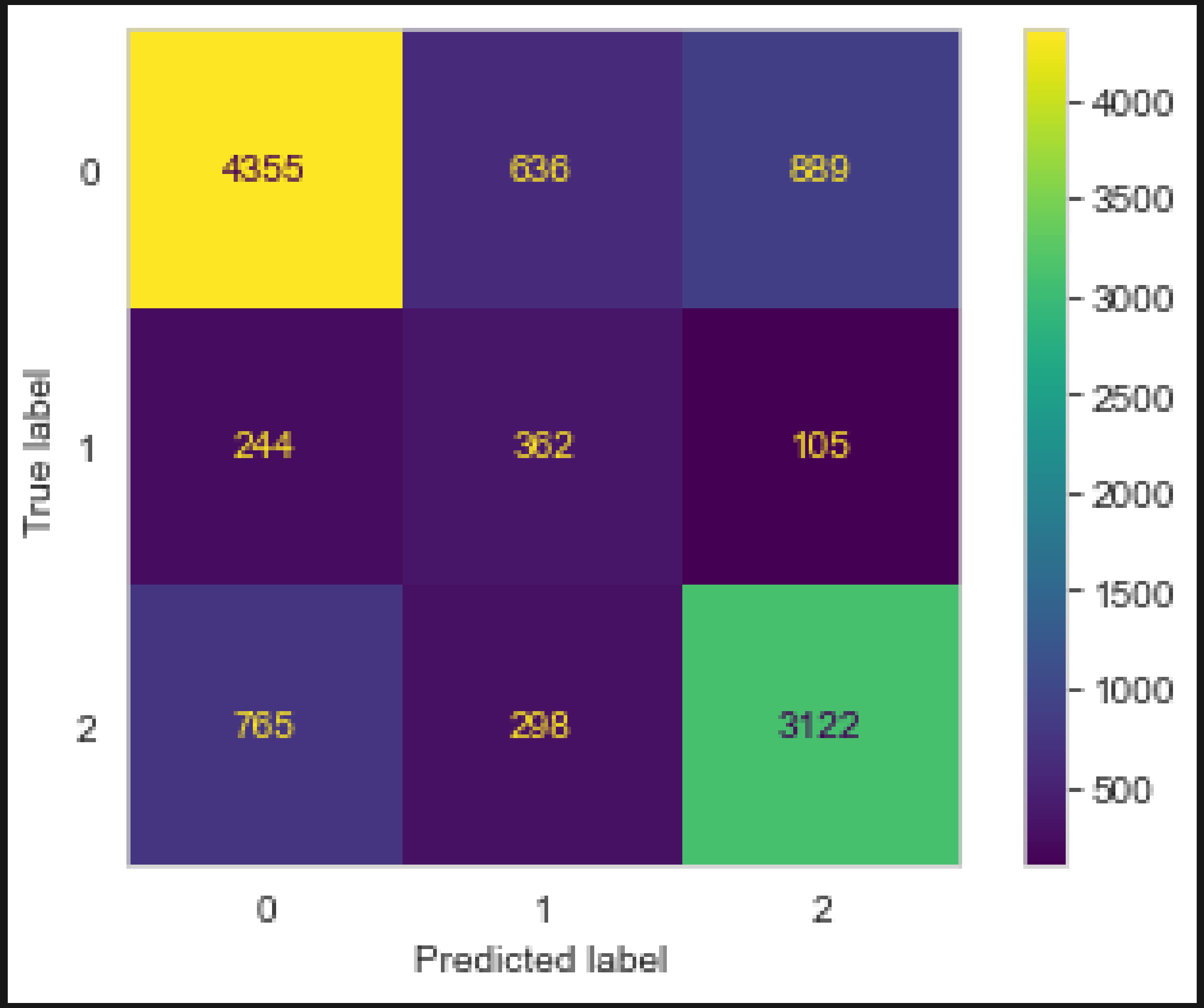
Base model	
Train Accuracy	Test Accuacy
87%	71%
Tuning	n_neighbors[5], weights : ["distance"]]
Tuned model	
Train Accuracy	Test Accuacy
82%	72%

Evaluation And Conclusion

The K Neighbors was choosen as the best model for the classification

The Model had the highest accuracy on Test Data :72%

The model also had the lowest error rate



Recommendation

The Government of Tanzania, Danida, Hesawa and world bank should be approached for financing of repairs of water wells

DWE have installed the majority of water wells are best placed for reparing the water wells

The KNN model can be used to clasify and identify faulty wells

XGboost model can be modelled as it was not used

Limitation

More paramater tuning can be done on the knn model as adding more parameters affected the run time