TANZANIA WATER WELLS

Business Understanding

Imagine walking about 4 miles for a few liters of questionable water. Then imagine doing that on a daily basis. Such is the situation for about 24 million Tanzanians. This water poverty has led to serious illnesses, high infant mortality, unproductive agricultural conditions, and a slumped economy.

In Tanzania, heterogeneous climate and geology contribute to significant seasonal, interannual, and geographic variability in water availability and water quality challenges. In addition, water supply challenges continue to become a huge issue due to meeting increased water demand associated with agricultural expansion and intensification and the need for improved access to domestic needs, including safe drinking water. Other key sectors fueling demand for surface and groundwater include animal husbandry, hydropower, and mining, while environmental flow requirements are also generally high due to the significant coverage of key nature reserves.

Water is an essential need of life, the Government of Tanzania, Non-Governmental Organizations, agencies, and individuals have come up with water wells to provide clean water for years. In rural areas, these water wells may be the only source of potable water and they are a lifeline for the inhabitants, but how effective are they? This project combines machine learning techniques with data visualization to point out potential causes of malfunctioning projects, identify the possible success of potential projects, and redirect funds to the places where they are in dire need and can be spent most efficiently.

Problem Statement

About half of Tanzanians face water poverty and many water points have been built to cater to the problem of water scarcity. However, these water points sometimes cease to function and therefore need an overhaul.

Our stakeholder, The government of Tanzania, has tasked The Miner League to identify patterns in nonfunctional wells. This will influence how new wells are built and help predict pumps that need repair or a complete overhaul.

Project Justification

Aside from the lucky few who reside near the great lakes of Tanzania, the majority who are living under the $1.25 poverty line and lack access to basic water supply have to trek for miles to access water, a cumbersome task that falls on women and young girls, who should otherwise be in school. This leads to a long-term decrease in robustness in future generations.

Further, most of this water is contaminated causing many waterborne diseases that cause high mortality rates among the population. With all this in mind, the Government of Tanzania aims to provide easily accessible water to the population and ensure that the technologies used are well maintained for long-term use

Objectives

Specific objectives

- To predict the condition of a waterpoint pump based on the geographical location
- To predict the condition of a waterpoint pump based on age
- To find patterns in non-functional waterpoint to influence how new water points are built
- To identify the effect of water quality on water pumps
- To identify how the extraction type affects water pump

Research questions

- Does water quality affect the functionality of a waterpoint pump?
- Does extraction type affect the functionality of a waterpoint pump?
- Does Region have an effect on functionality?

- Do different water sources have an effect on well functionality?

Business Success Criteria

- The Ministry of Water under the Government of Tanzania will use our model to:
  - identify geographical locations where a pump is likely to fail
  - successfully show how water quality affects pumping used
  - improve maintenance operations in functioning waterpoints

Project Success Criteria

- Precision Score: 65%

Data Understanding

Data Collection

The data for this project was obtained from DrivenData: Pump it Up: Data Mining the Water Table and the datasets were split into the training set values, training set labels, and testing set values. The data from the training set labels was the target variable, status group, and was merged with the training set values to make a complete data frame.

Data Description

The training data frame had a shape of 59,000 rows and 41 columns while the testing set had 14850 rows and 40 columns.

Some of the features were dropped as they were not relevant to our study and had duplicated data. The relevant features and their description were as follows:

Column	**Description**
funder	Who funded the well
gps_height	Altitude of the well
installer	Organization that installed the well
Longitude	GPS Coordinates
Latitude	GPS Coordinates
basin	Geographical water basin
region	Geographical location
population	Population around the well
age	The age of the well
decade	The decade when the well was constructed
extraction_type_class	The kind of construction the waterpoint uses
scheme_management	Who operates the waterpoint
Extraction_type_class	The kind of extraction the waterpoint uses
payment	What the water costs

water _quality	The quality of the water
quantity	The quantity of the water
source_type	The source of the water
waterpoint_type_ground	The kind of waterpoint
status _group	Condition of the well

Data Preparation

In order to use the data collected the following were performed on the data:

1. Creating a data frame of the relevant data columns
2. Changing object columns into categorical
3. Changing float column into integer

Cleaning Data

Checking for completeness, consistency, validity, and uniformity are all parts of data cleaning. We verify that every requirement is met and that there are no missing values to ensure completeness. Consistency tests for duplicates. The dataset's outliers are also examined as part of the validity process. Furthermore, uniformity verifies the accuracy of the various columns' data types.

The data cleaning process involved several steps;

- Consistency - The dataset has duplicates and were dropped.

- Completeness - some columns of the dataset was found to have no missing values and the one that had missing values their rows were dropped.
- Validity - gps_height had no outliers and the population was found to have outliers but it was left as it is because of the natural variation of the population.
- Uniformity - the dataset had accurate data types.

Data Types

The dataset contains categorical variables and integers

Assumptions

The data provided is correct and up to date.

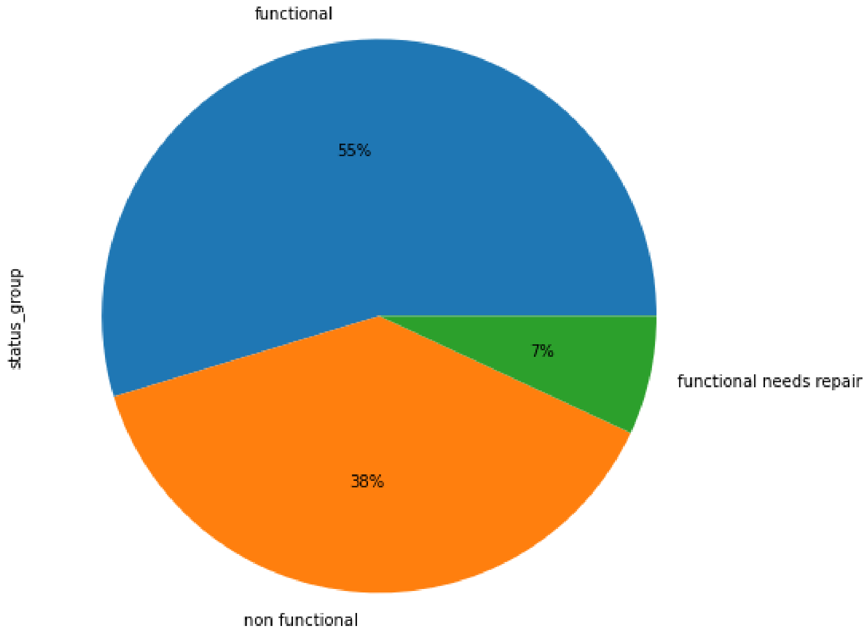
Exploratory Data Analysis Findings

The Data Science team did an Exploratory Data Analysis to monitor the trends and distribution of the data. The analysis was done on a jupyter notebook.

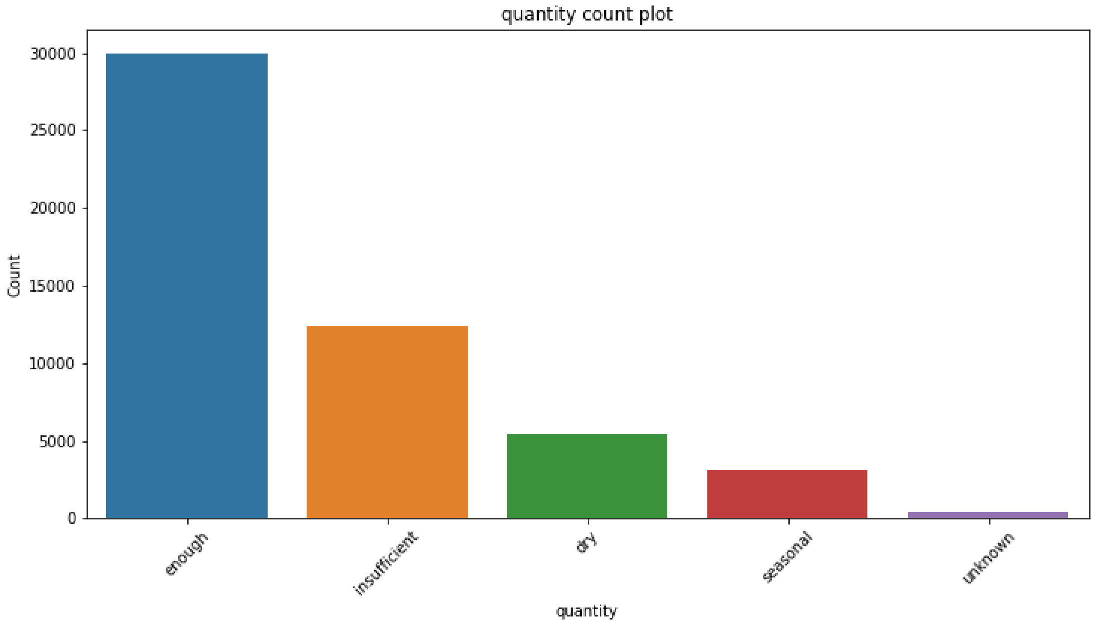
Univariate Analysis Findings

Categorical Analysis

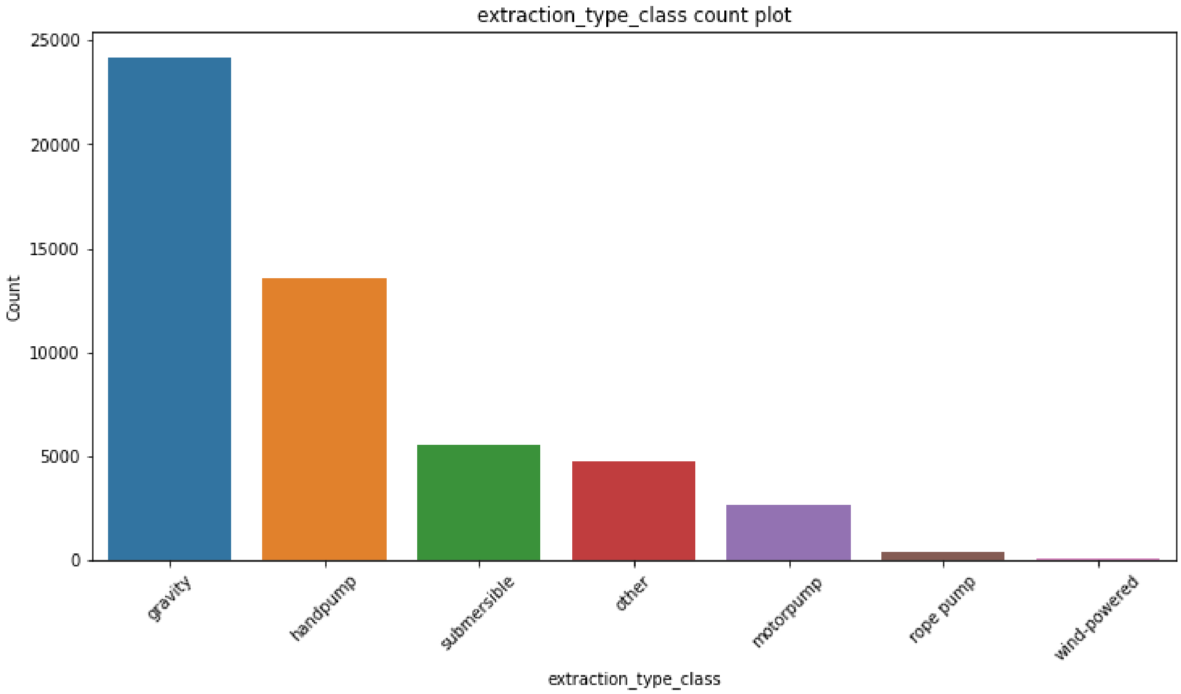
1. 55% of waterpoints are functional, 38% are non-functional, 7% are functional and need repair



2. The highest number of waterpoints have enough water, then they rest have insufficient, dry, seasonal and unknown quantities respectively.



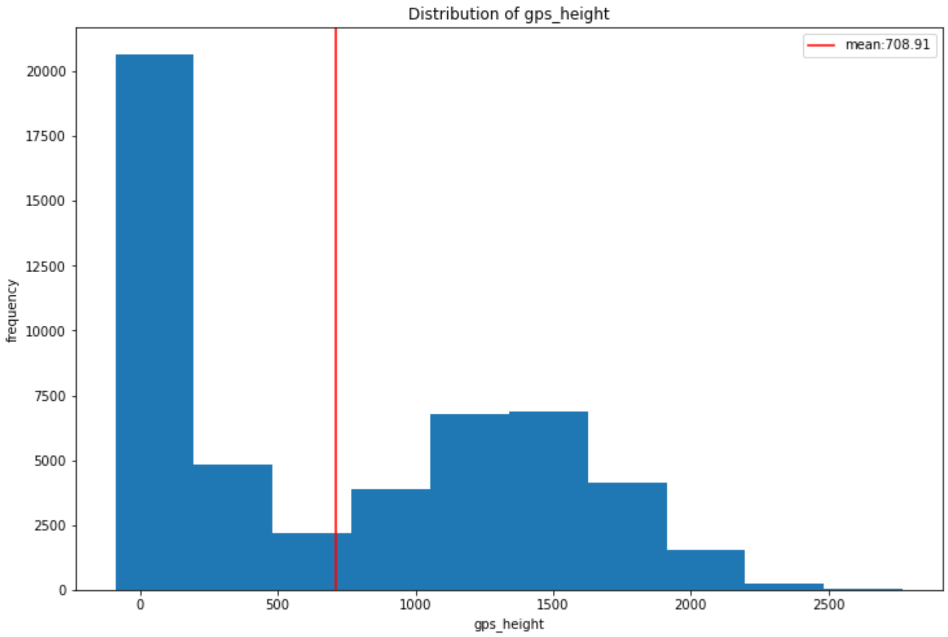
3. Most waterpoints draw their water from Pangani basin followed by Lake Victoria basin



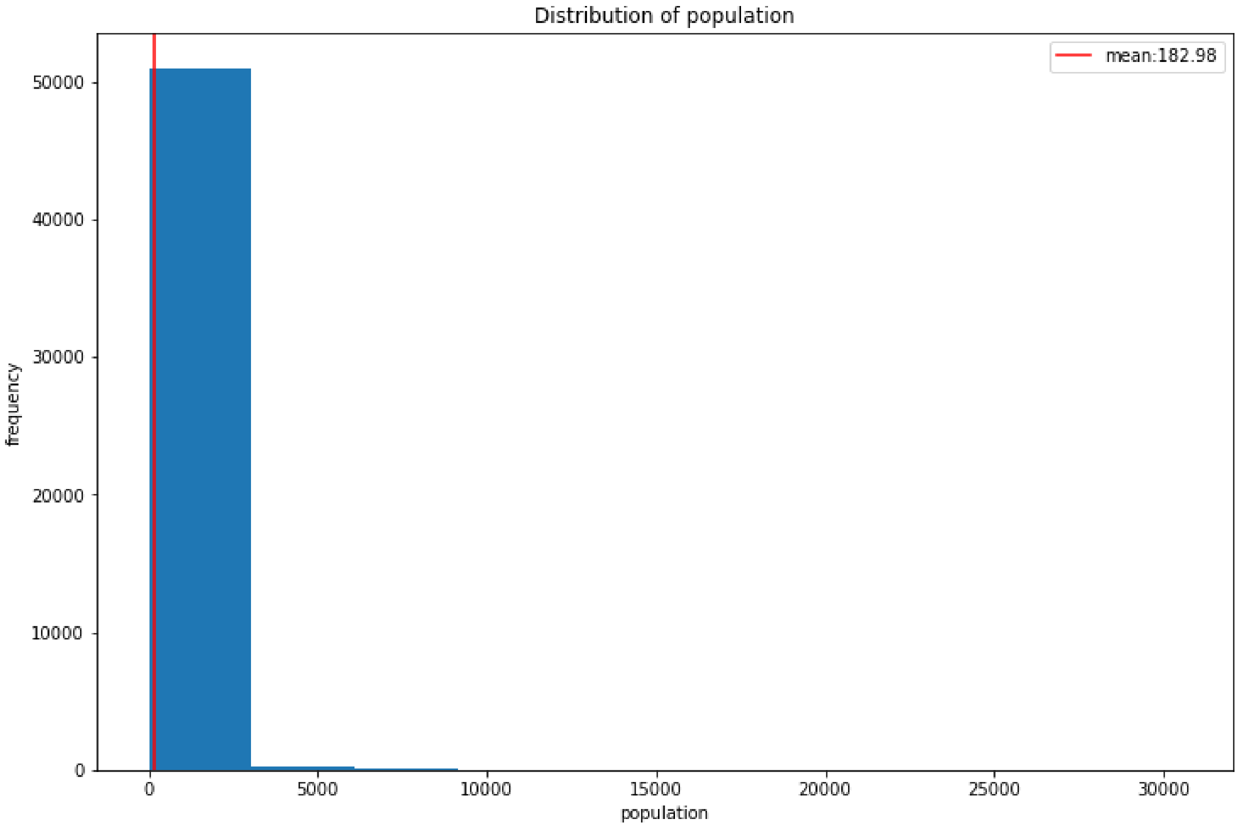
Most waterpoint pumps use gravity for extraction and handpumps extraction come in as a close second.

Numerical Analysis

1. The distribution of gps_height of the waterpoints is positively skewed

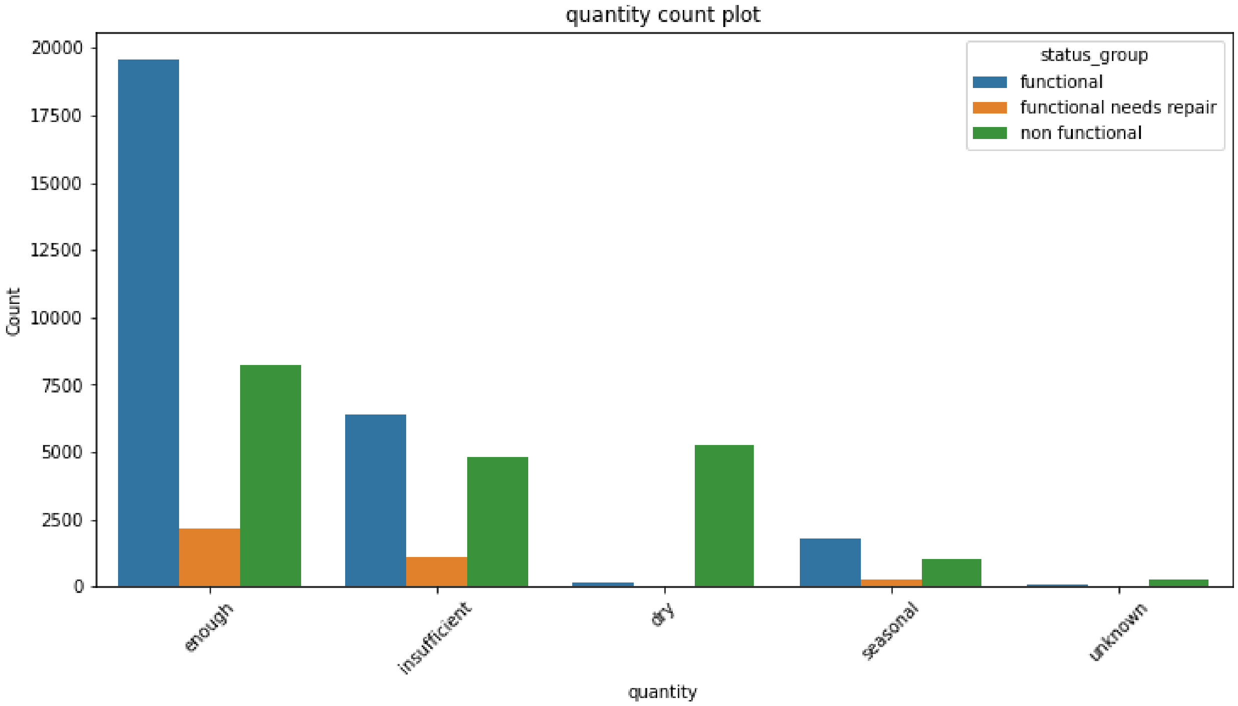


2. The distribution of the population around waterpoints is positively skewed

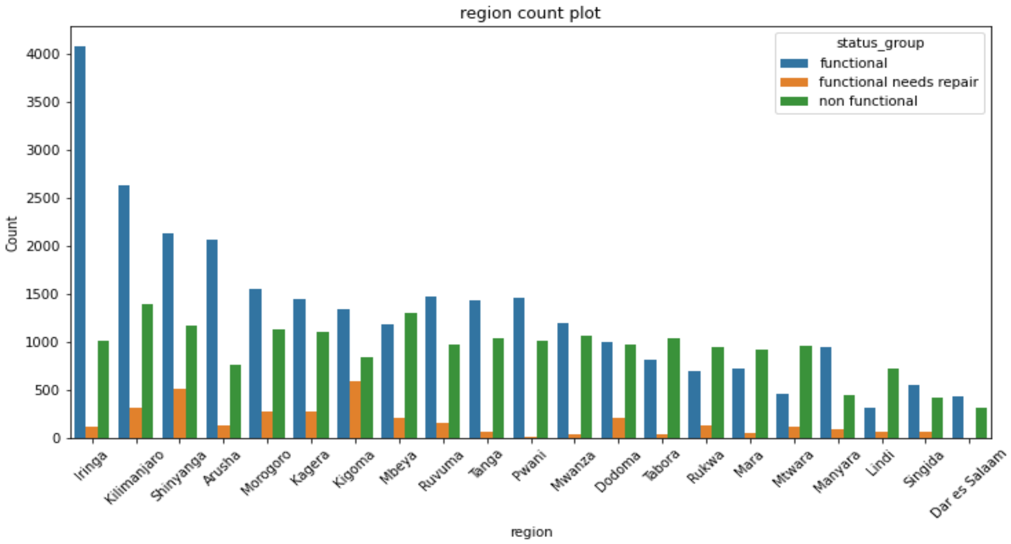


Bivariate analysis

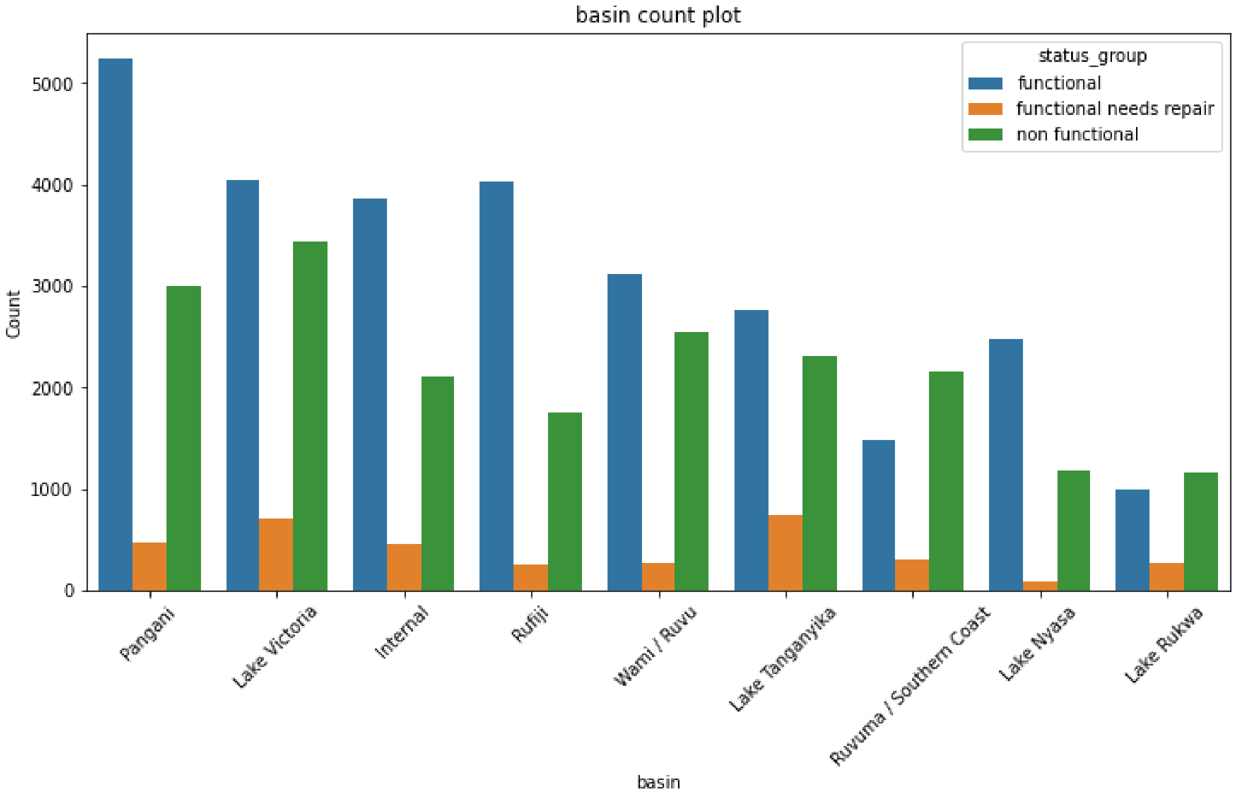
1. The waterpoints that are dry mostly consist of non functional pump and Waterpoints with enough water contribute to most functional pumps.



2. Iringa region has the most functional pumps and the Kilimanjaro region has the most nonfunctional pumps.



Most of the functional water points are located near the Pangani basin. Most non functional waterpoint pumps are located near the Lake Victoria basin while most waterpoint pumps that needs repair are located near the Lake Tanganyika region.



Feature engineering

Here we perform some manipulation of our dataset to improve our machine learning model for a better performance and greater accuracy, therefore Feature engineering was done:

- To convert construction year into decades.
- Then created a new column age
- To Change date into date time object.

Modeling

Data Preprocessing

For modeling purposes the following actions were further performed on the data:

1. Selecting columns to be fed into the model.
2. Separating the columns by their type.
3. Transforming the target variable into a numerical data type through one hot encoding

Models

The following models were considered:

1. Logistic Regression
2. Decision Tree
3. K Nearest Neighbor
4. Random Forest

Model	**Accuracy**	**Precision**	**Cross validation score**
Logistic Regression	59%	56%	59%
Decision Tree	63%	63%	63%

K Nearest Neighbor	61%	61%	61%
Random Forest	66%	66%	65%
XGBooster	65%	65%	65%

Findings

The Random Forest algorithm, having the highest precision score of all performed better than the other models and shall be used as the final model The accuracy of the model was 66% which means that it correctly identified whether a waterpoint is functional, non-functional or needs repair 66% of the time

Conclusion

- The final model was selected as the Random Forest
- Most functional waterpoint pumps are those that have soft water
- Waterpoints with enough water are the most functional
- Iringa, Kilimanjaro and Shinyanga are the top 3 regions with most functional pumps
- Kilimanjaro and Mbeya has the regions with the most non-functional pumps
- Kigoma has the highest number of waterpoints that need repair
- Water drawn from springs and shallow wells have the highest number of functional pumps
- Water drawn from Springs has the highest number of functional pumps that need repair

- Water drawn from Shallow wells and boreholes to have the highest number of nonfunctional wells
- Those waterpoints with insufficient water have very few functional pumps.

Recommendations

- The government should prioritize drawing water from springs when building the waterpoints and should not draw water sourced from shallow wells or boreholes as they spoil the pumps quicker.
- Water points with enough water should be closely monitored, as high use could lead to their failure