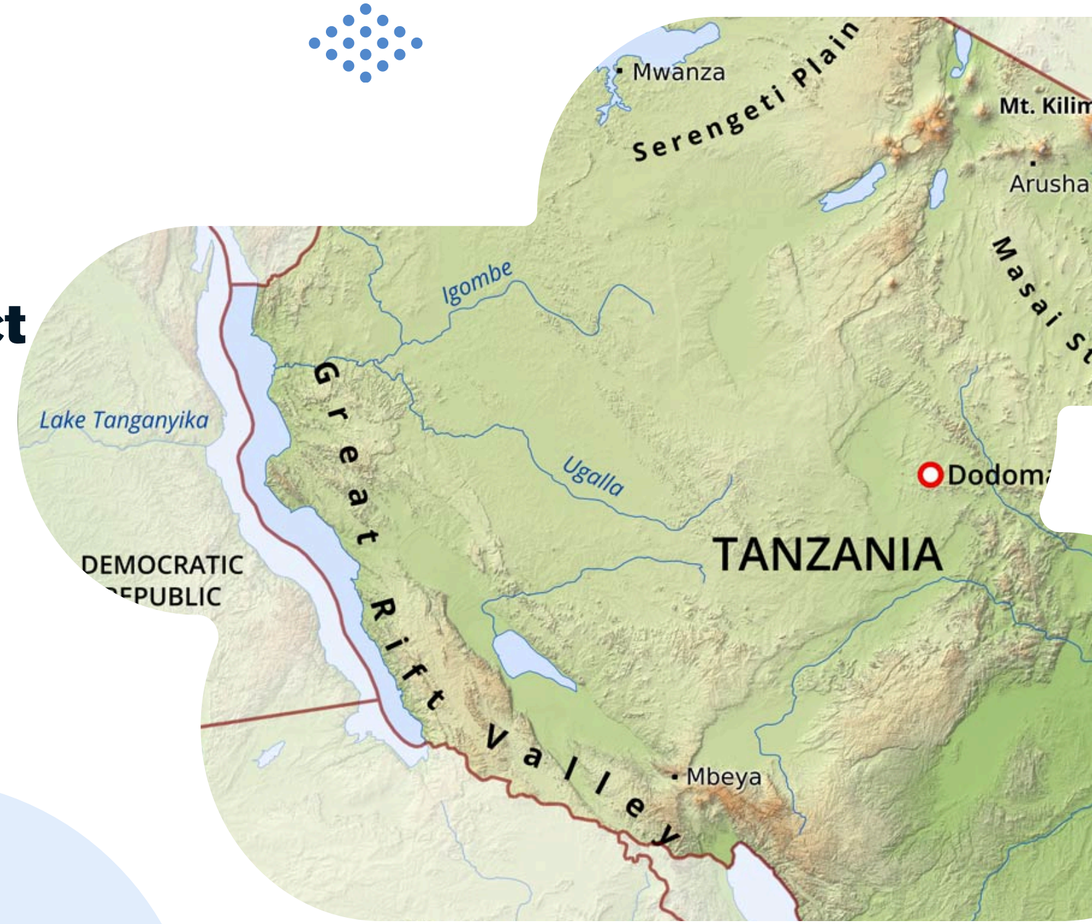
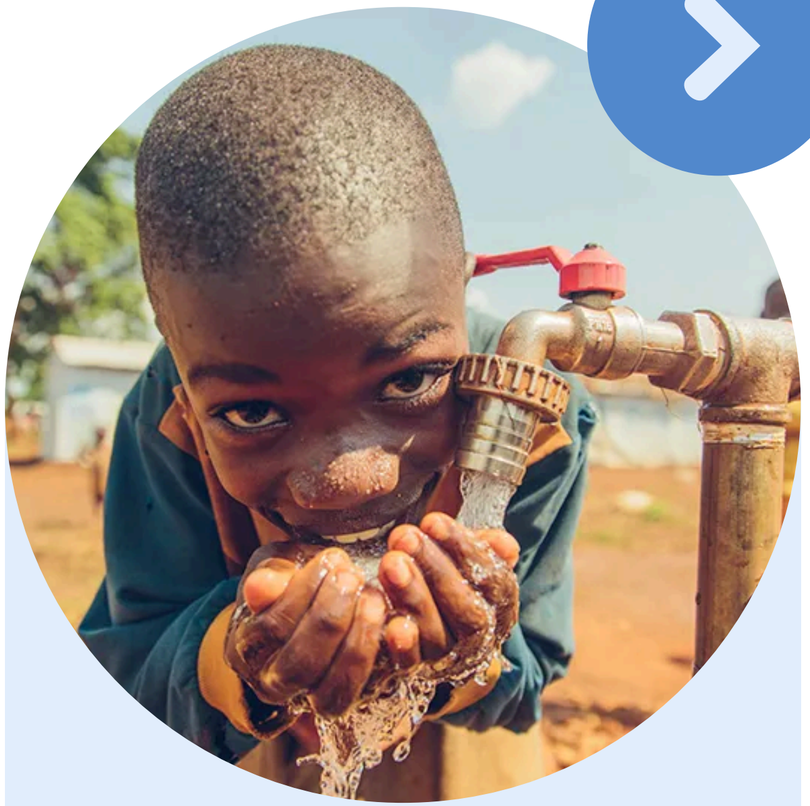
ReDI School of Digital Integration

HYDROLOGIC

Using Machine learning models to predict the functionality status of water pumps across Tanzania

Subaye Opoku-Acquah
Mariam Farda
Farzaneh Gerami





Why This Project Matters

Reliable access to clean water is essential for public health and socio-economic development, yet over 80% of Tanzania's population still lacks safe water. A major cause is the high number of non-functional or poorly maintained water pumps, which are difficult and costly to inspect manually across such a wide and rural landscape. To address this challenge, our project aims to analyze water pump data and build machine learning models that can predict the functionality status of each pump. By doing so, we hope to optimize maintenance strategies, improve the allocation of resources, and ultimately enhance community access to clean and safe water — contributing to the global effort toward achieving sustainable water and sanitation for all.

Sprint 1 - EDA

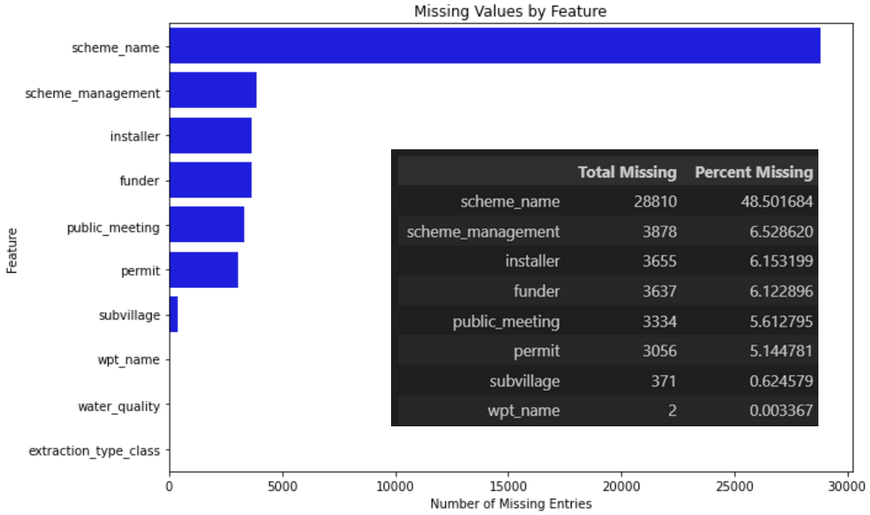
In this sprint, our goal was to understand the water pump dataset, clean it, and prepare it for modeling. We focused on uncovering patterns in pump functionality and identifying impactful features.

Data at a Glance

Checked data dimensions, types and missing values

Basic distribution chart of functionality status

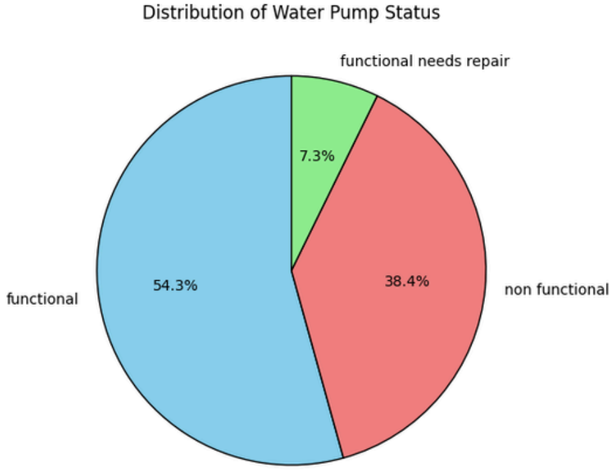
Overview of relation between columns



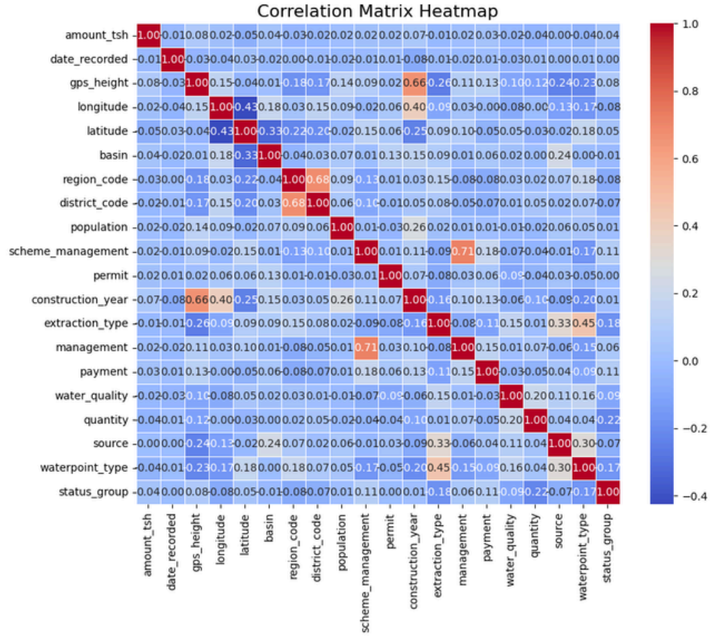
	Total Missing	Percent Missing
scheme_name	28810	48.501684
scheme_management	3878	6.528620
installer	3655	6.153199
funder	3637	6.122896
public_meeting	3334	5.612795
permit	3056	5.144781
subvillage	371	0.624579
wpt_name	2	0.003367

Training Shape: (59400, 41)
Test Shape: (14850, 40)

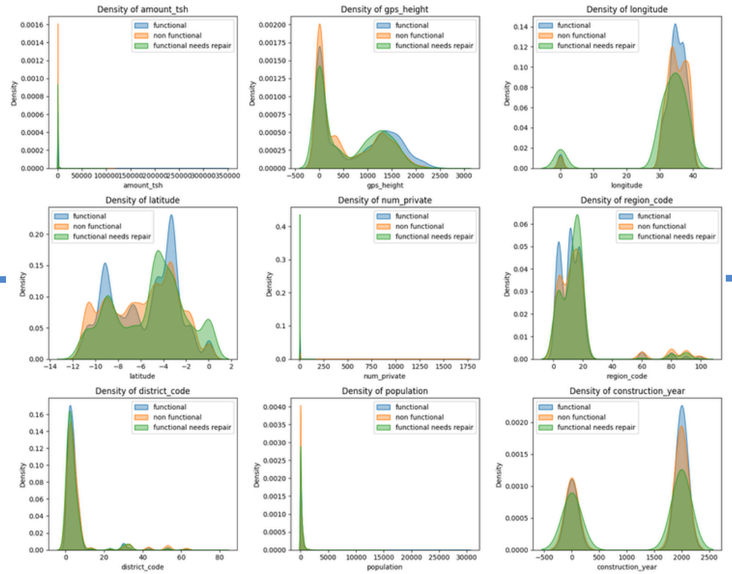
Total missing values: 46743
Percentage: 1.92%



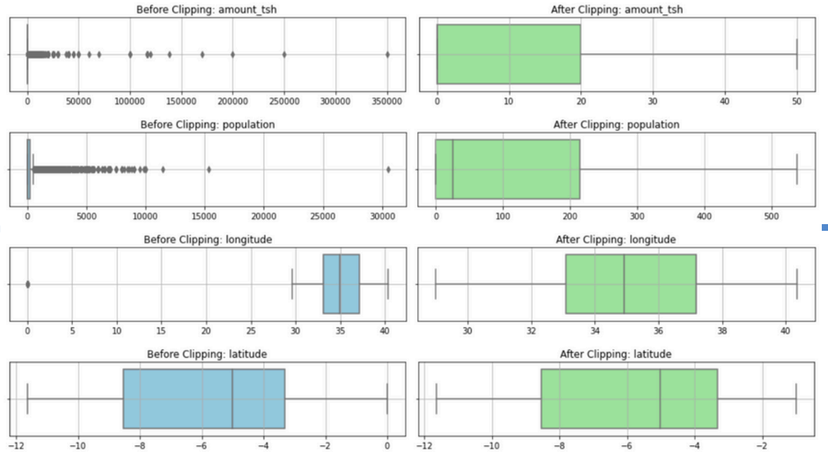
Features: 39 input features and 1 target
The target variable `status_group` has 3 classes

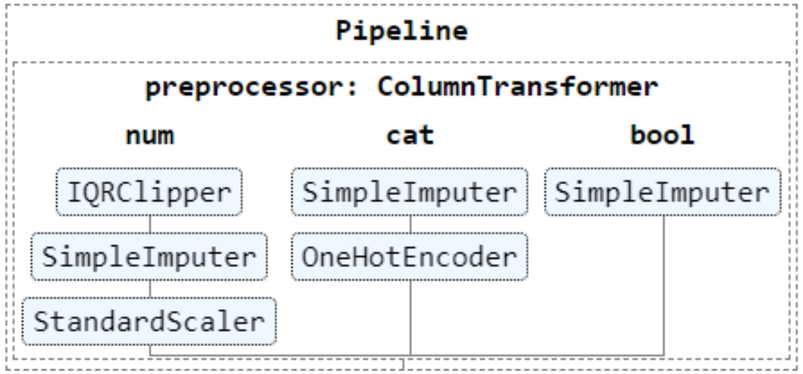


Data Cleaning and Preprocessing



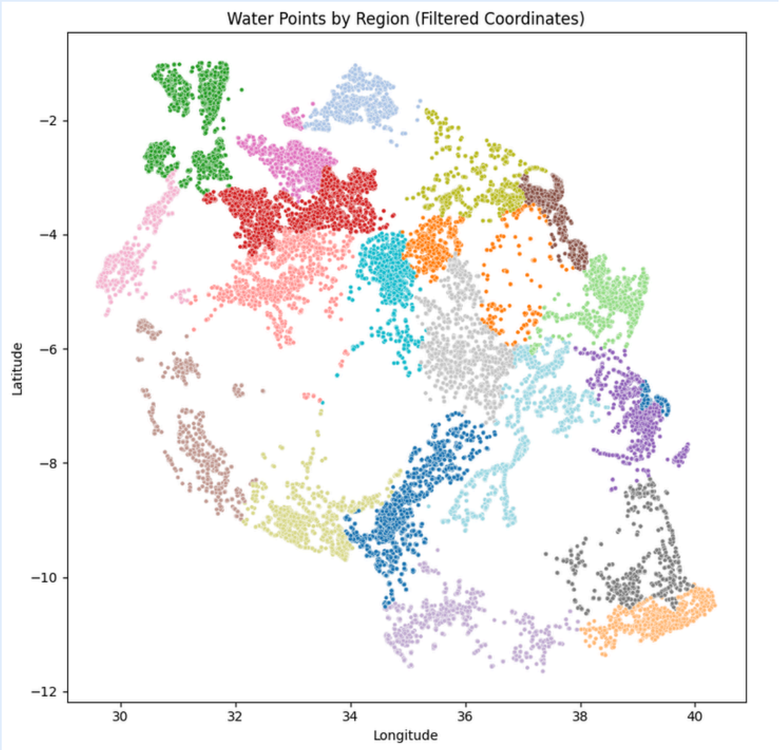
Distribution of numerical columns by functionality

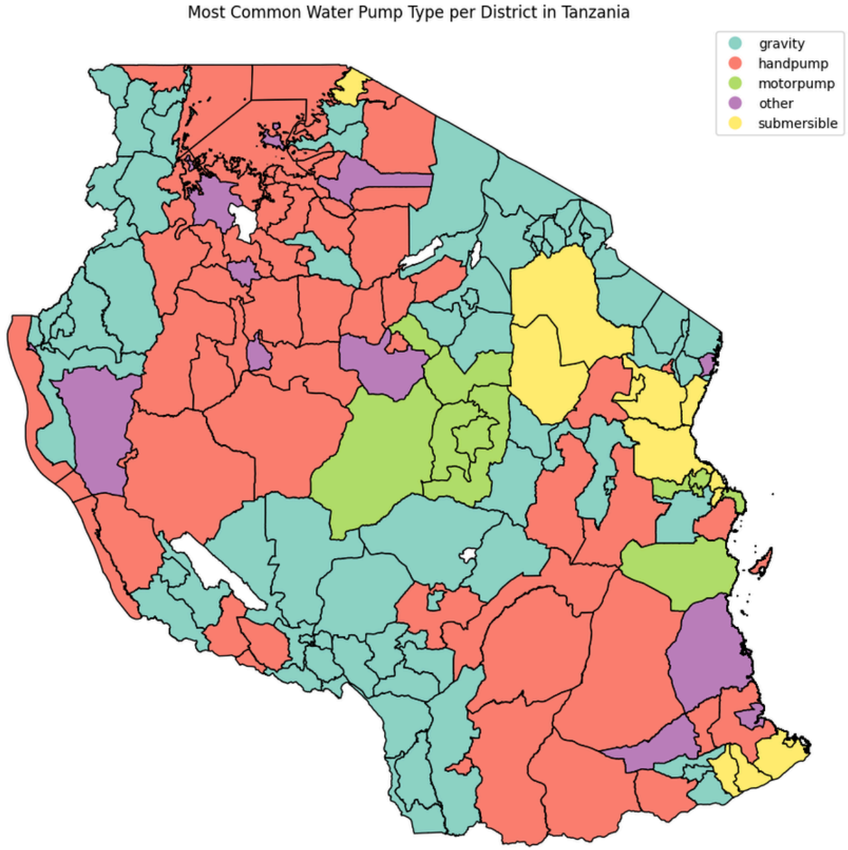


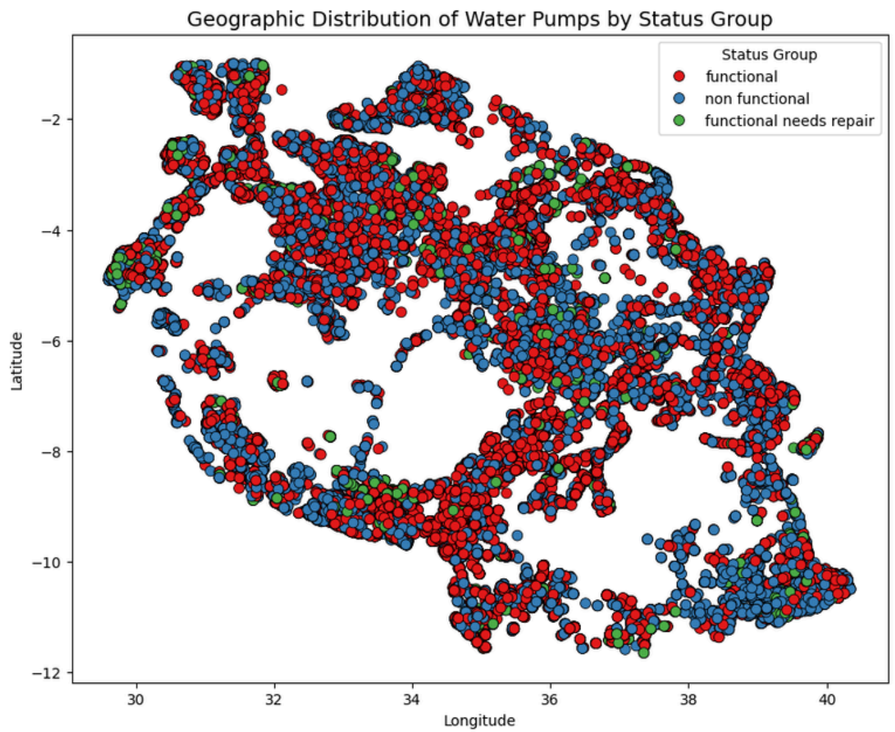


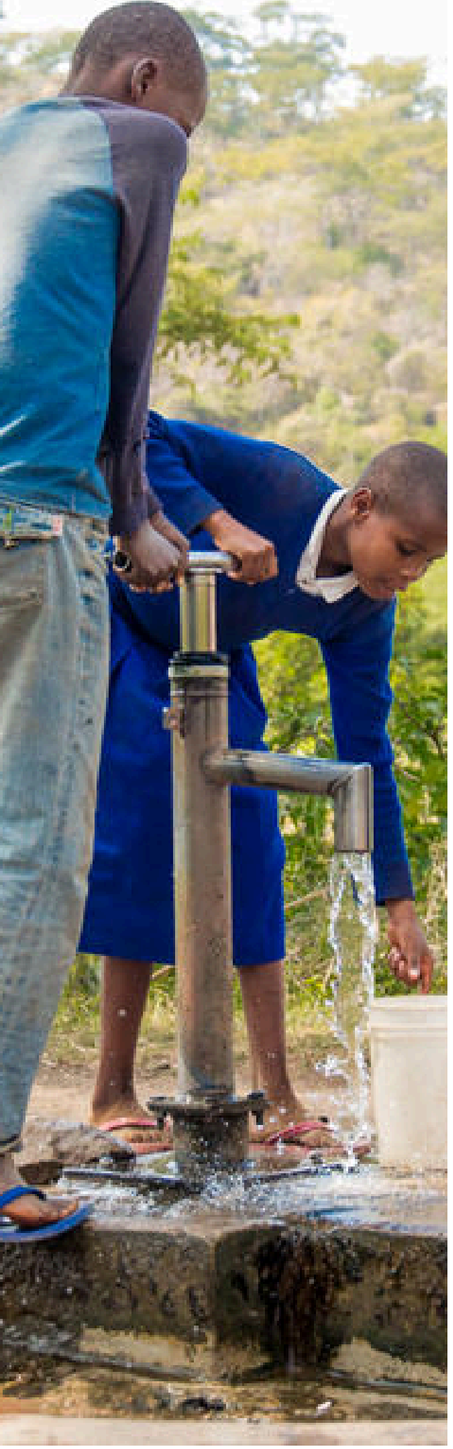
Applied One-Hot Encode Low Cardinality Columns
Applied Frequency Encode High Cardinality Columns
Handle Outliers in Numerical Features
Normalize / Standardize Numerical Features to make sure all the numerical features are on the same scale
Build a Data Cleaning Pipeline

Geospatial Analysis









Key Insights

- **Data Quality issues:**

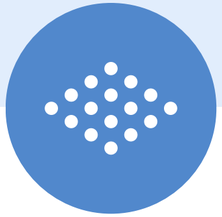
1. Missing or incomplete values in: 'ward', 'funder', and 'installer'
2. Nearly 50% of entries in 'construction_year' had placeholder value 0
3. Zero (0) used in place of missing values (e.g., NaN) in numeric fields
4. Some GPS coordinates fall outside of Tanzania’s borders

- **Geospatial Insights**

1. Moderate positive correlation between elevation (gps_height) and pump functionality
2. Pumps at lower elevations were more likely to be non-functional
3. Elevation may reflect environmental or accessibility challenges affecting pump performance

- **Feature Analysis**

1. 'construction_year' is a strong predictor of functionality
2. Pumps built after 2000 are significantly more likely to be functional
3. Infrastructure age is a key factor for model prediction and prioritization



THANK YOU!