



Tanzanian Water Wells

E d w a r d C h e n g



Our Mission

Clean water, basic toilets and good hygiene practices are essential for the survival and development of children. For children under five, water- and sanitation-related diseases are one of the leading causes of death. Every day, over 800 children die from preventable diseases caused by poor water, and a lack of sanitation and hygiene. UNICEF's water, sanitation and hygiene (WASH) team works in over 100 countries worldwide to improve water and sanitation services, as well as basic hygiene practices. Last year, UNICEF's efforts provided nearly 14 million people with clean water and over 11 million with basic toilets. Our current mission is to predict the condition of water wells in Tanzania, as they are facing a massive water crisis. Here are some alarming statistics about the Tanzanian water crisis.

01 24 Million

People are impacted by the United Republic of Tanzania's water crisis; that's almost half the population of Tanzania

02 43%

Of the Tanzanian population are relying on water that does not meet their standard

03 17%

Of the Tanzanian population have no place to wash their hands. Without hand washing at critical times, such as after using the bathroom, people are more prone to get sick.

01

Obtain the Data

- Import the `Tanzanian_data_set.csv` and `Tanzanian_labels.csv`

02

Scrub the Data

- Cast appropriate data types to predictor variables
- Clean up NA values
- Filter data set to exclude insignificant values
- Convert important categorical variables into numerical ones to use for modelling stage

03

Explore the Data

- Explore relationships between categorical and continuous predictor variables with the target variable 'well condition'
- Observe for any patterns/trends, which indicates that a certain variable may be useful for the final model

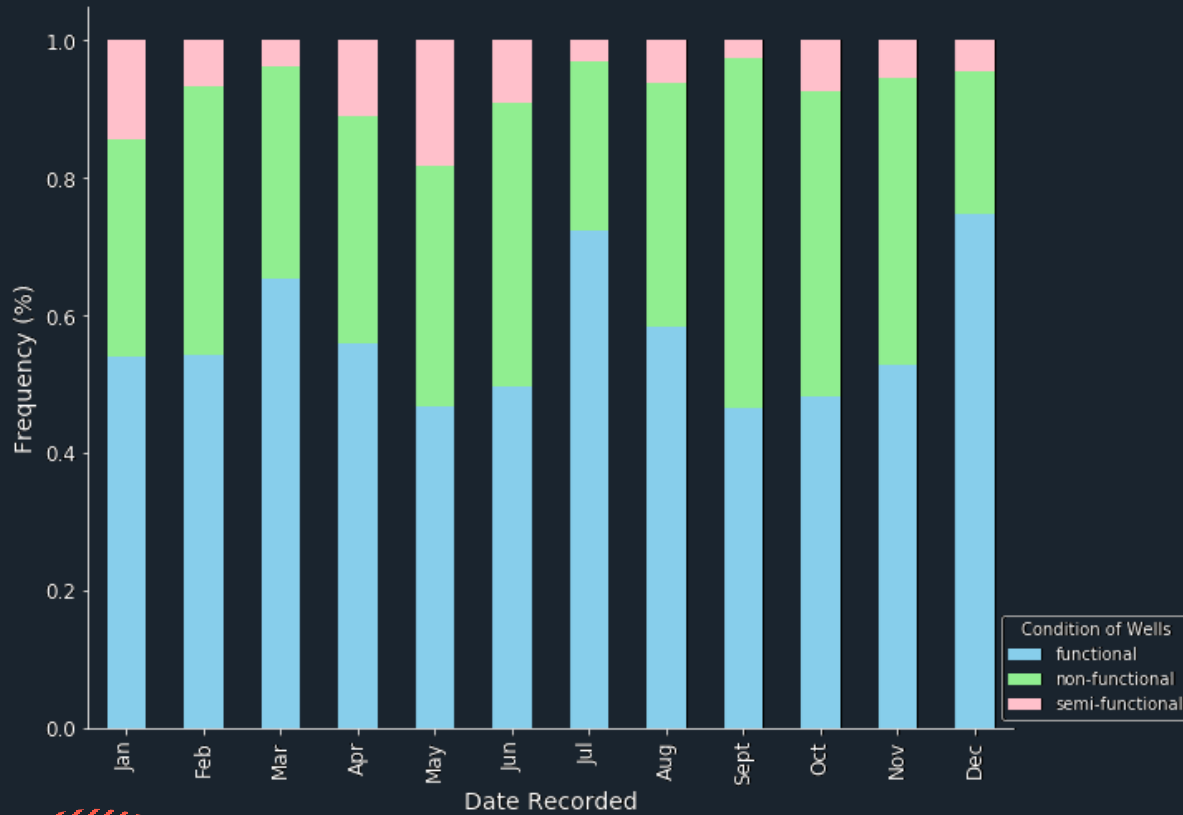
04

Model the Data

- Run 5 different machine learning algorithms:
 - Logistic Regression
 - Decision Tree
 - Random Forest
 - Ada boosting
 - Gradient boosting
- Select best performing algorithm and perform Randomized Search

Distribution of Well Conditions

By Date Recorded

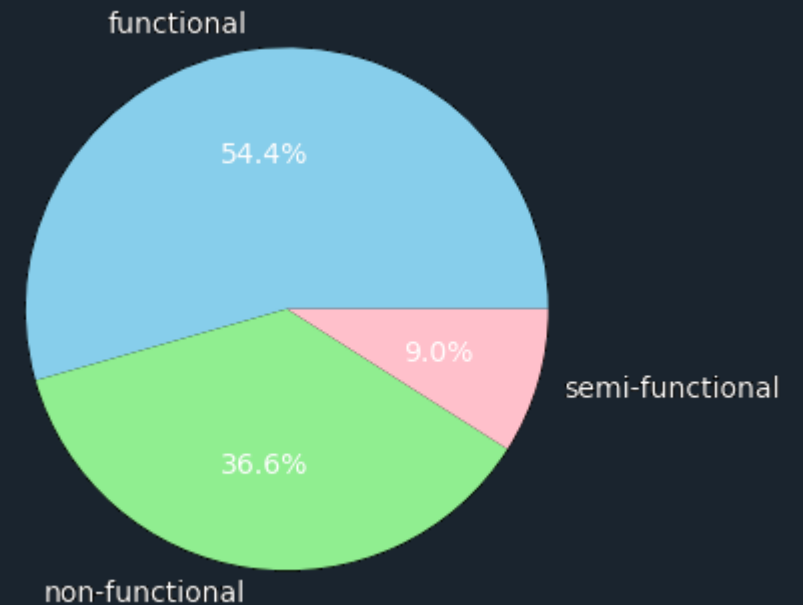


54.4%

Of the wells are functional

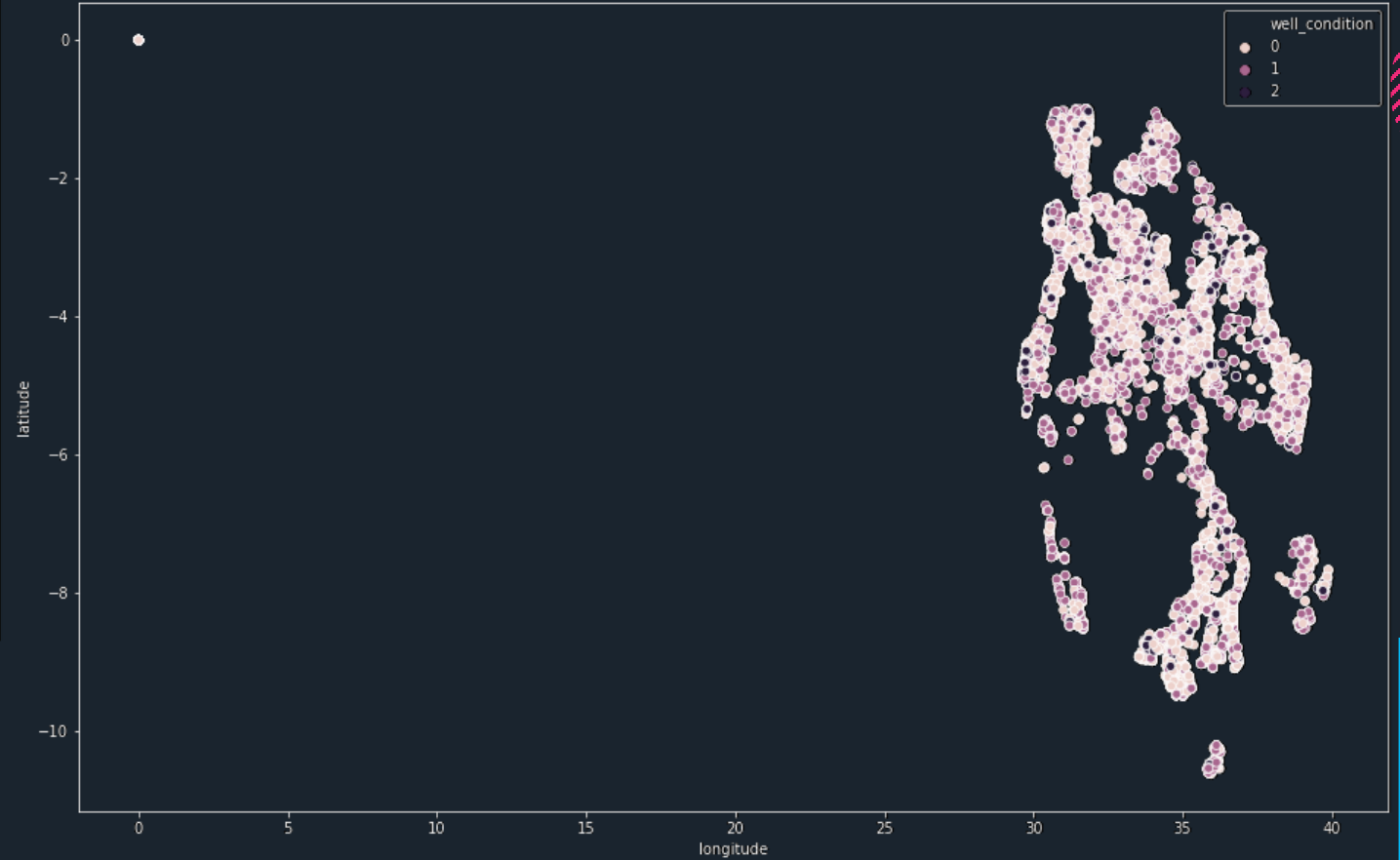
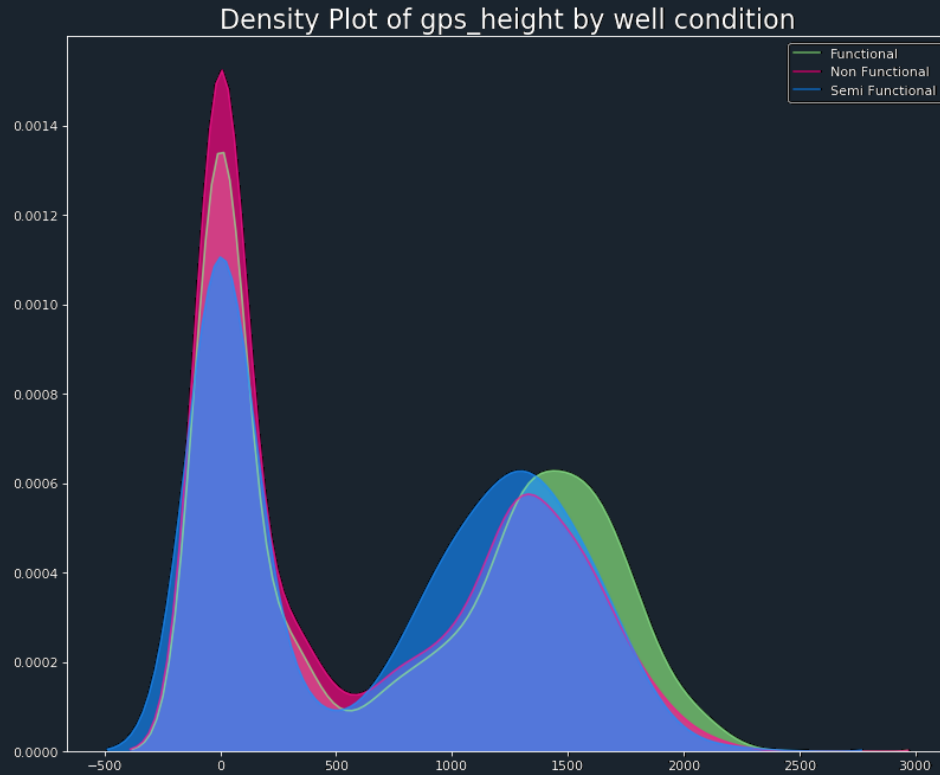
36.6%

Of the wells are non-functional



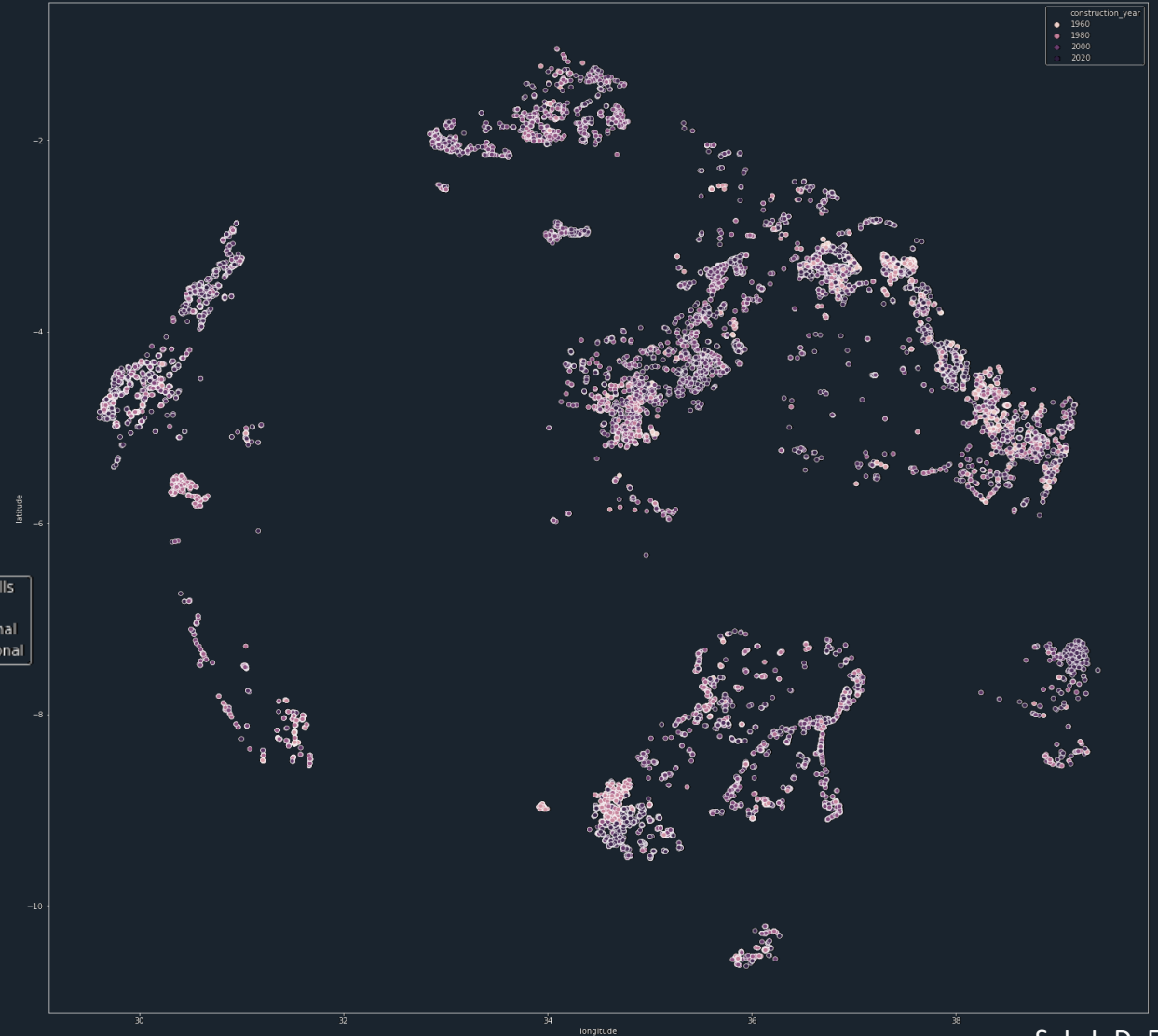
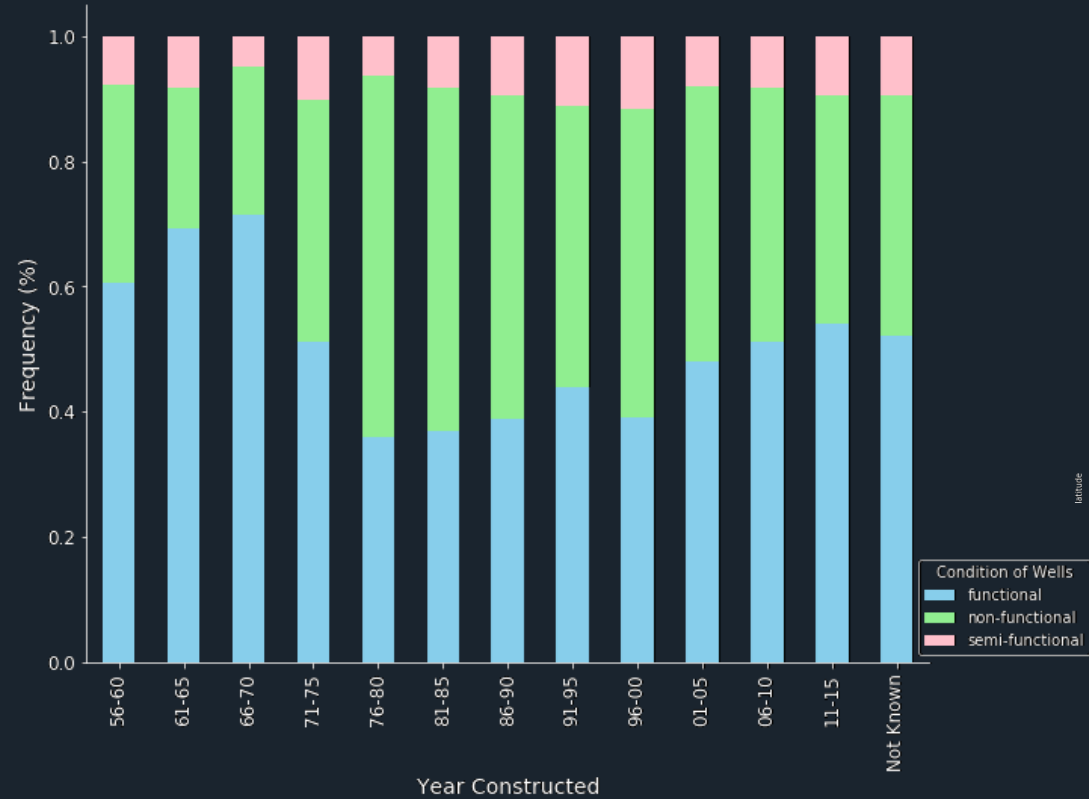
Locations of Wells

And the distribution of its altitude



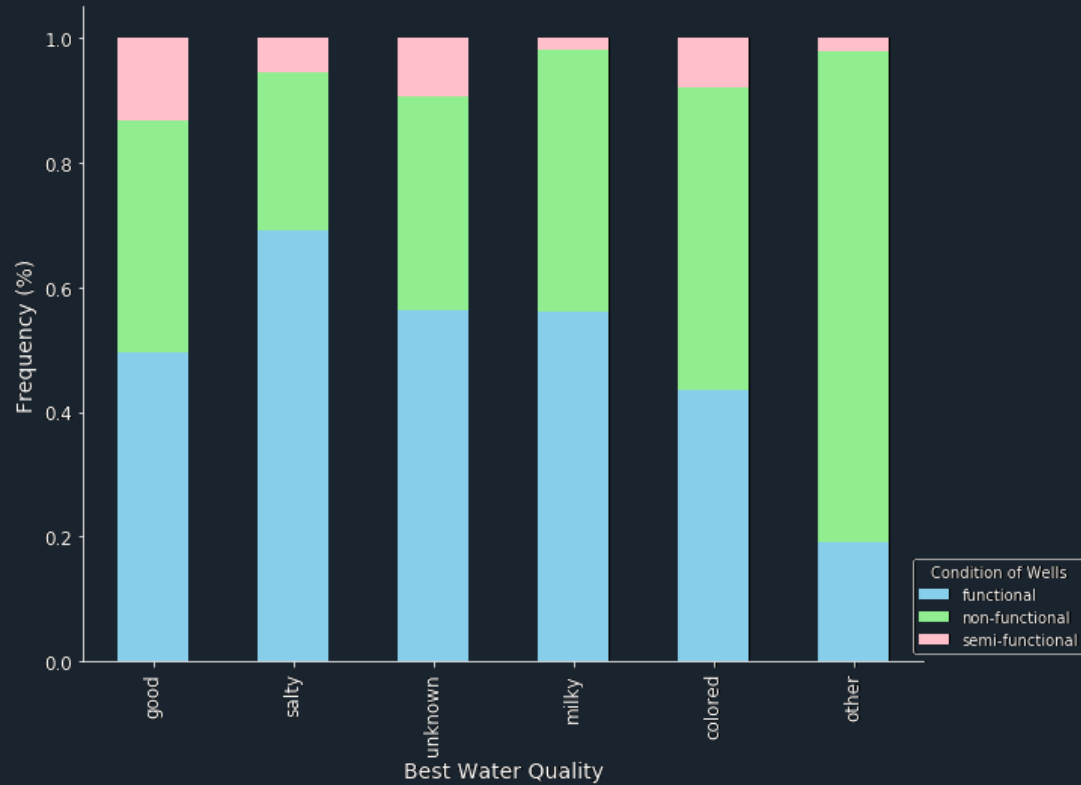
Construction Years of Wells

By Well Condition



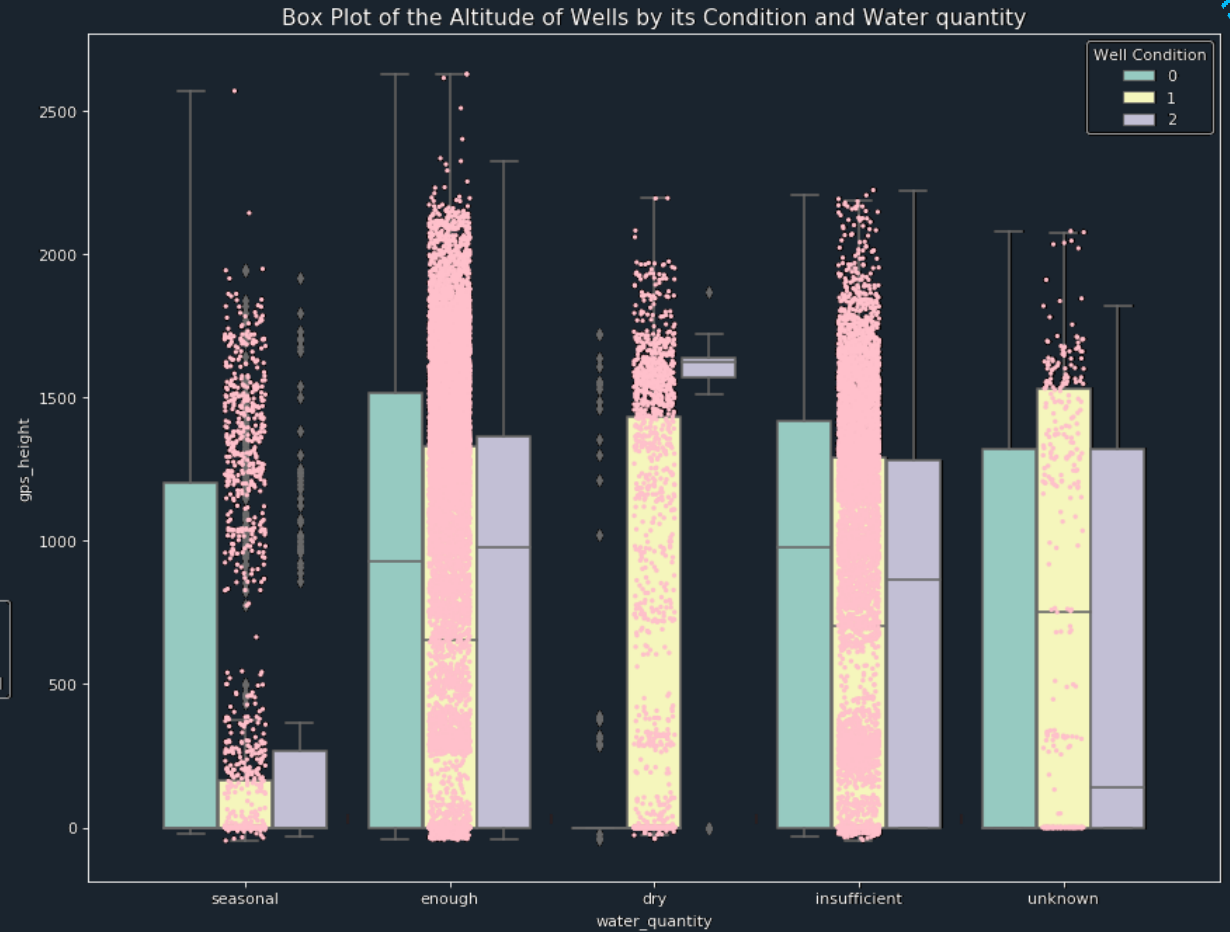
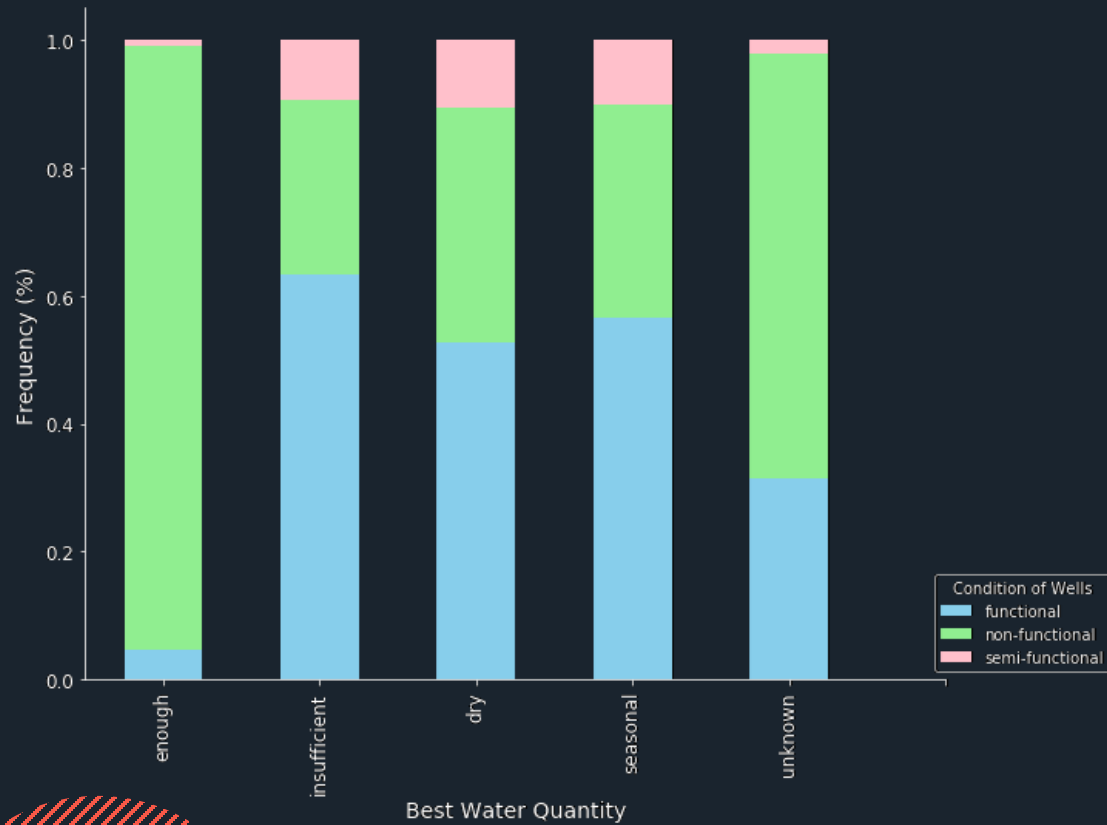
Best Water Quality

By Well Condition



Best Water Quantity

By Well Condition



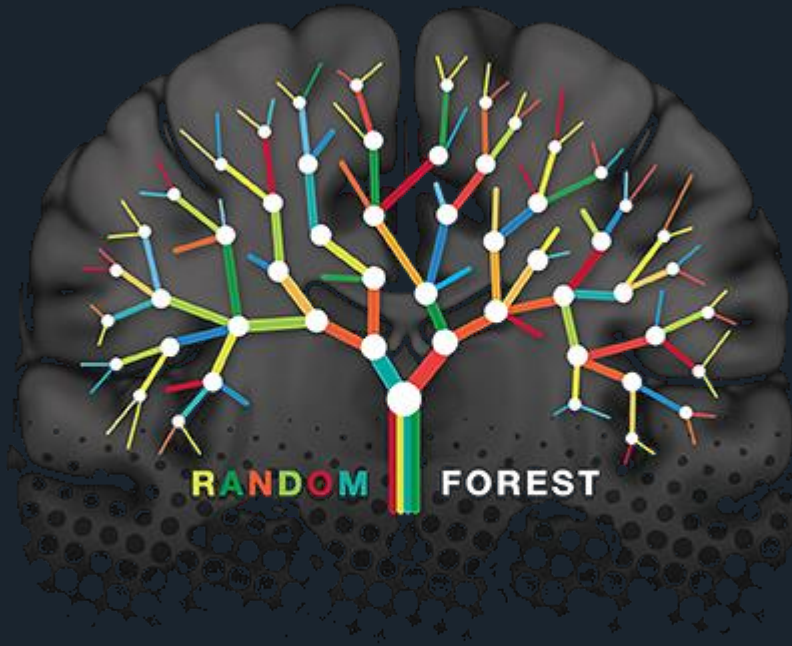
Model Selection Algorithms

Logistic Regression, Decision Tree, Random Forest, Ada Boosting, Gradient Boosting

Best Performing Model: **Random Forest**

70.5% Accuracy Score

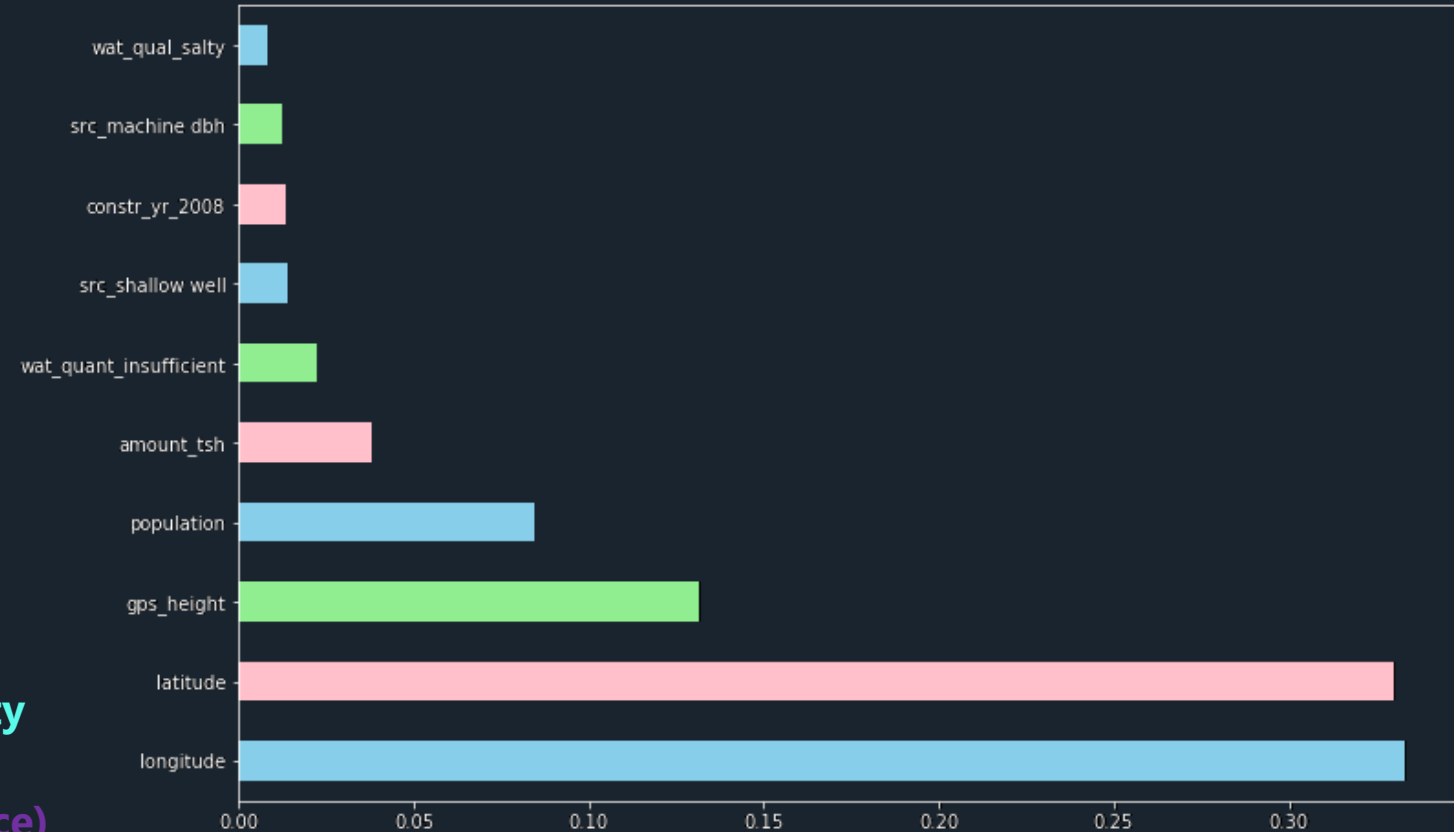
70.0% F1 Score



Recommendations

Top 10 Features

- 01 Longitude
- 02 Latitude
- 03 Altitude of the Wells
- 04 Population
- 05 Total Static Head
- 06 Insufficient Water Quantity
- 07 Shallow Well (Water Source)
- 08 Wells Constructed in 2006-2010
- 09 Machine Dbh (Water Source)
- 10 Salty Water Quality



Future Work

- Trying different data cleaning methods
- Having more time to play around with combining unused features together
- Collect more data to train the model
- Transforming more categorical variables into numerical ones for modelling purposes
- Making predictions with another modelling algorithm such as XGBoost



Thank You

Flatiron School Online Data Science Bootcamp