Water Pumps

Capstone Project Two
Springboard - School of Data

David Clark
May, 2021

Outline

- Present background and business problem.
- Discuss data acquisition and cleaning.
- Exploratory data analysis.
- Modeling.
- Final results and future directions.

Background

- Tanzania
  - Developing country.
  - Access to water is very important.
  - Water distributed using pumps located through out country.
  - Difficult to efficiently monitor pump working status.

- Business problem: How can the government of Tanzania improve water pump maintenance by knowing the pump functional status in advance?

Data Acquisition

- Data provided by:
  - Taarifa.
  - Tanzanian Ministry of Water.

- Data characteristics:
  - 59,400 rows.
  - 39 columns.
  - Data types: strings, numerical, boolean, and dates.

- Data wrangling:
  - Missing values.
  - Data and variable types.
  - Duplicate values.

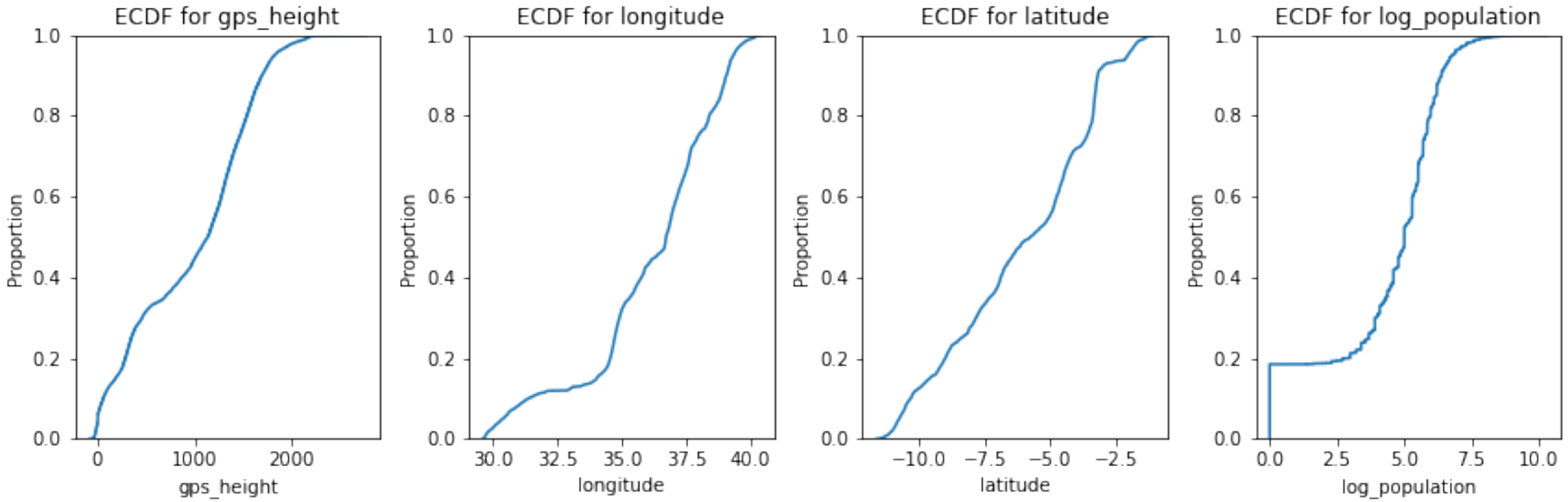
Missing Values

- Removed columns that contained 50% or more missing values.
- Deleted rows with missing populations.
  - 30% of data.
- Set missing boolean values to majority boolean value for that column.
- Set empty values for columns with empty values to None.

Uninformative Features

- Redundancy
  - Similar features captured same information.
    - Selected more general features for modeling.
    - Removed the rest
- Irrelevant columns.
  - Removed columns that were not related to outcome variable.

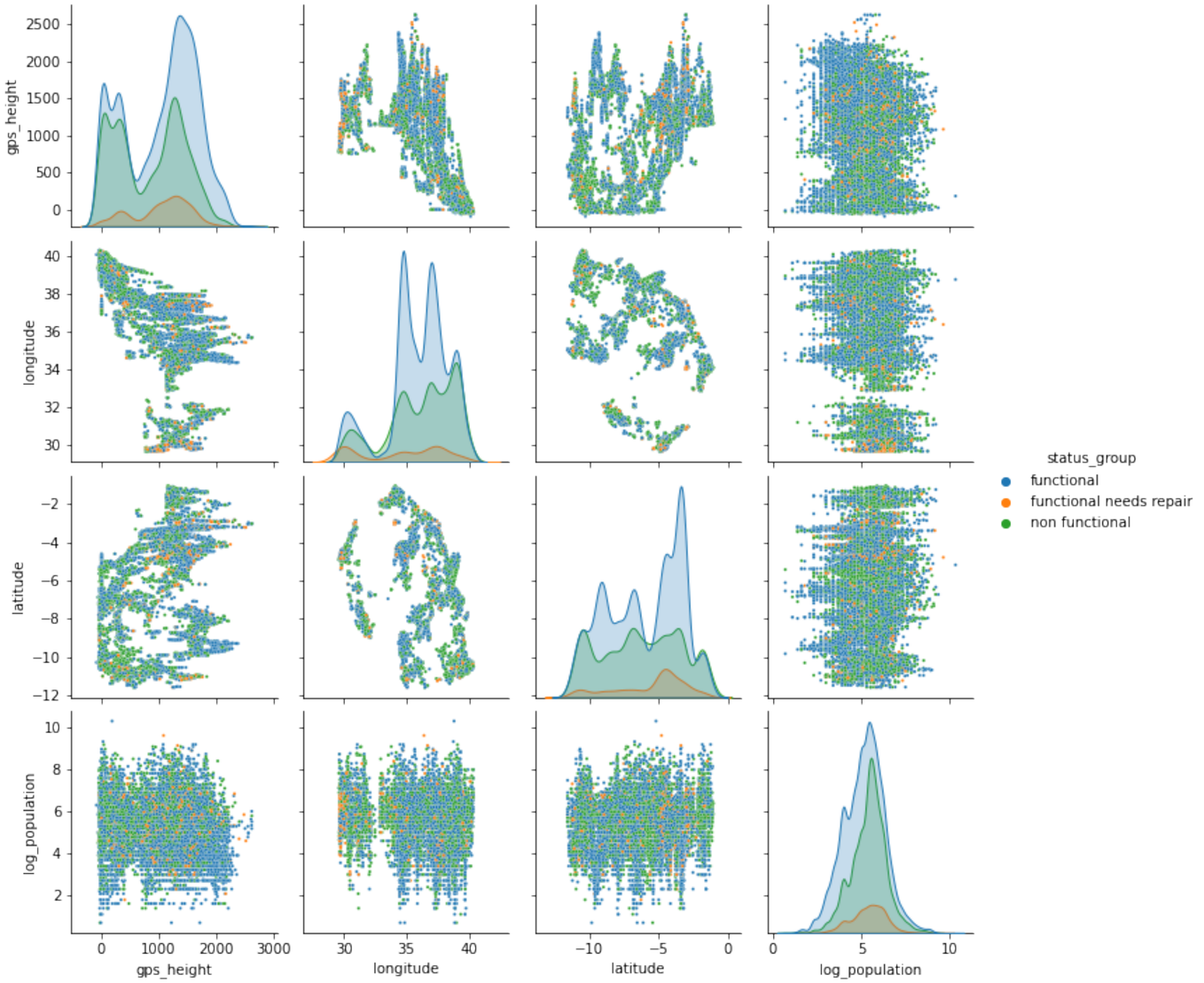
Numerical Variables: ECDFs



- Convert population to log space.
- Most variables appear uniform.
- Log(population) looks like a normal distribution.

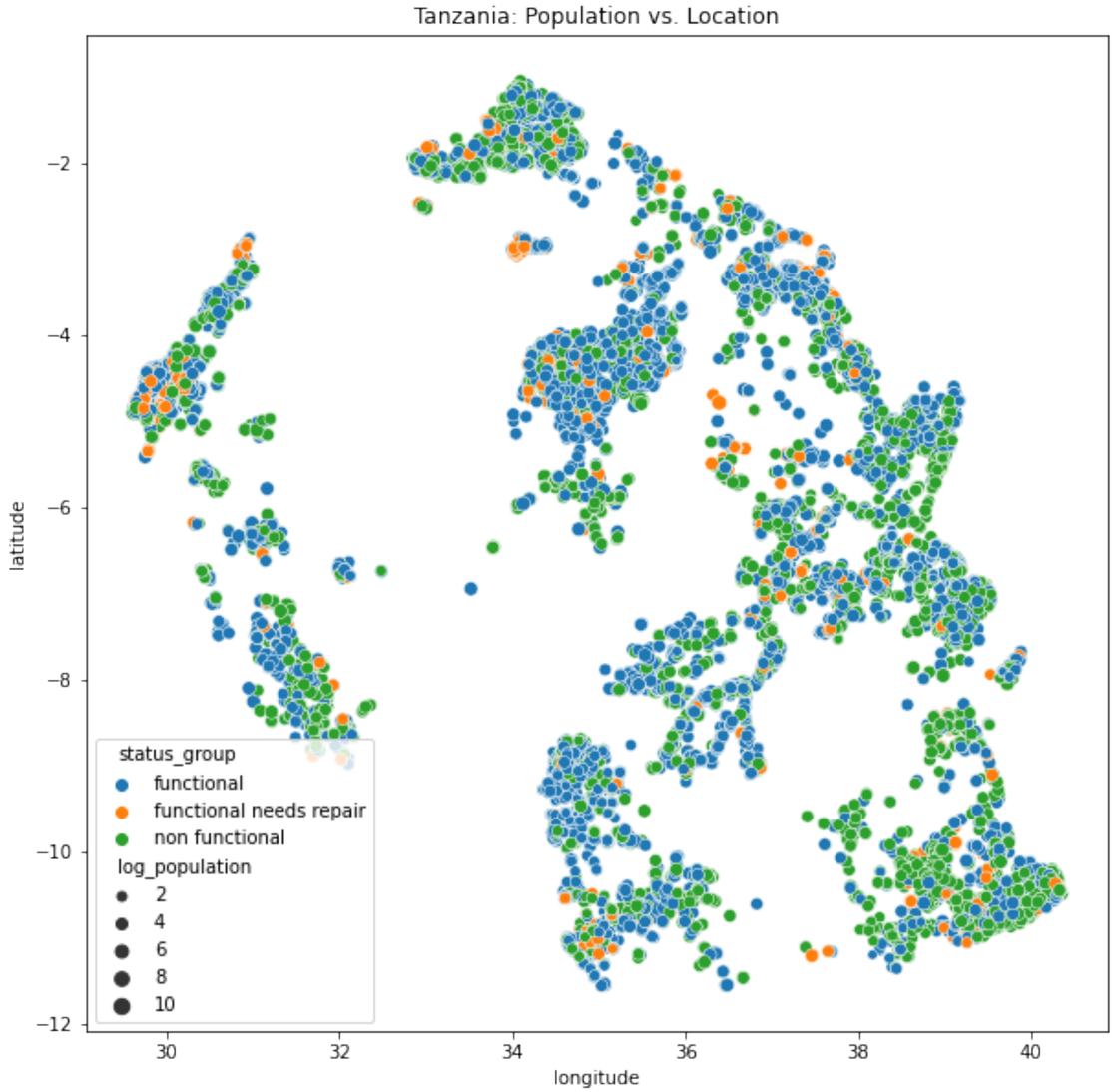
Numerical Variables: Pair Plots

- No obvious linear trend between any of the variables.
- Longitude-latitude plots outline country boundaries.
  - Pumps well distributed in country.



Numerical Variables: Geographical location and Population

- Functional pumps concentrate along middle of country.
- Non-functional pumps are in East and West.
- No obvious trend between population size and functional status.



Categorial Variables

- Large range in categories for each variable.
- Pumps lacking payments were more often non-functional.
- One-hot encoded categorical variables before modeling.

Baseline Model

Table 1: Baseline Model Results

	Train			Test		
	precision	**recall**	**support**	**precision**	**recall**	**support**
functional	0.76	0.91	12482	0.77	0.91	5349
functional needs repair	0.63	0.15	1505	0.63	0.15	645
non functional	0.78	0.65	7309	0.79	0.66	3133

- Train-test split
  - 70% train.
  - 30% test.
- Scaled data to values between 0 and 1.
- Chose logistic regression.
  - Good at solving classification problems.

Extended Modeling

Table 2: Final Results

	base_line		logreg_over		logreg_under		rf_over		rf_under		xgb_over		xgb_under		
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	class count
functional	0.768	0.911	0.846	0.652	0.839	0.638	0.835	0.867	0.851	0.649	0.825	0.867	0.853	0.646	5349
functional needs repair	0.632	0.152	0.226	0.741	0.218	0.726	0.467	0.416	0.260	0.757	0.463	0.388	0.236	0.758	645
non functional	0.788	0.659	0.733	0.674	0.718	0.668	0.813	0.778	0.716	0.725	0.803	0.759	0.738	0.706	3133

- Recall improved compared with baseline model for all three models.
- Best model:
  - XGBoost.
  - Undersampling.
  - Recall score of 0.76 for minority class.
- Recall still lower than 0.80.

Extended Modeling - 2 Classes

Table 3: Final Results - 2 Classes

	logreg_over		logreg_under		rf_over		rf_under		xgb_over		xgb_under		
	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	precision	recall	class count
faulty	0.73	0.74	0.73	0.74	0.81	0.76	0.76	0.80	0.79	0.75	0.75	0.78	3778
functional	0.81	0.81	0.82	0.80	0.84	0.87	0.85	0.83	0.83	0.86	0.84	0.82	5349

- Best model:
  - Random Forest
  - Undersampling
  - Recall = 0.80 for minority class

Conclusions

- Classification problem to predict water pump status.

- Heavy class imbalance between majority and minority classes
  - Explored under and over sampling techniques.
  - Combined two minority classes into one class, *faulty*.

- Best recall score = 0.80 on minority class
  - Two-classes, *functional* and *faulty.*
  - Random Forest model
  - Undersampling

Recommendations to Client

- Use best fit model to predict water pump status in Tanzania
  - Fewer in person trips will be needed to monitor water pumps.
  - Money and time will be saved.

Future Work

- Try additional sampling methods to counter class imbalance.

- Perform a more detailed exploration of hyperparameter values.
  - A more granular search could lead to more refined models with better results.

- Try a different train/test split.
  - A smaller split could give the models more data to train on, which could possibly improve the predictions.