Pump it Up: Data Mining the Water Table

Peder Norr
May 23, 2021

Outline

- Business Problem
- Data
- Methods
- Findings
- Results
- Conclusions

Business Problem

- Tanzanian Ministry of Water wants to improve water pump maintenance operations
- Needs a way to better predict functionality status of water pumps
- Needs to determine what characteristics might indicate a non functional pump in the future

Data & Methods

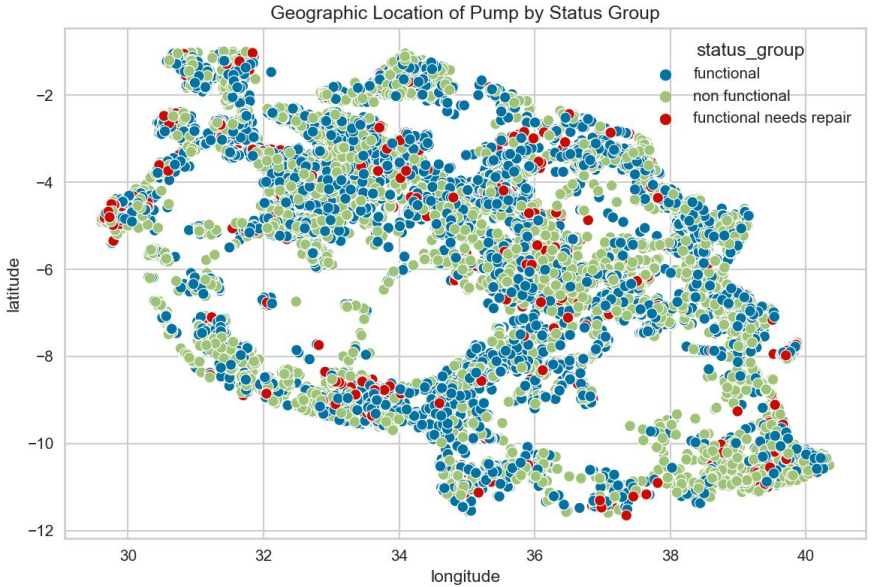
- Data sourced from Taarifa and DrivenData competition site
- Dataset contains 41 variables describing pump functionality status (the target variable), pump geographic location, what kind of pump is operating, when it was installed, how it is managed, etc.
- Dataset encompassed 59400 pumps from 2011-2013
- Created random forest classifier model
- Resulting model had an accuracy of 80%, meaning it could accurately predict the status of a given pump 80% of the time

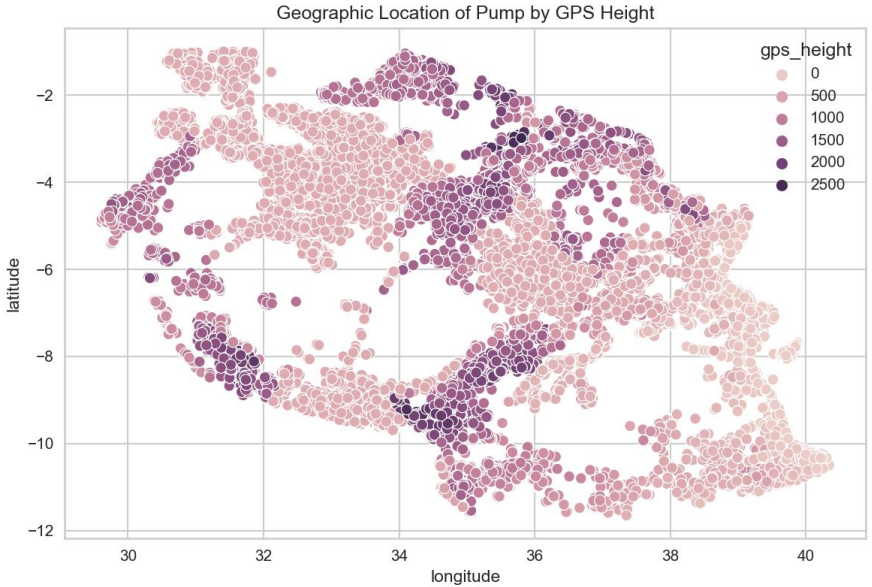
Findings

Random forest classifier model analysis of Tanzanian water pump data identifies actions the Ministry of Water can take to improve pump maintenance efficiency:

- Target pumps in higher altitude
- Target pumps with lower water quantity
- Target pumps in lower population areas
- Target older pumps

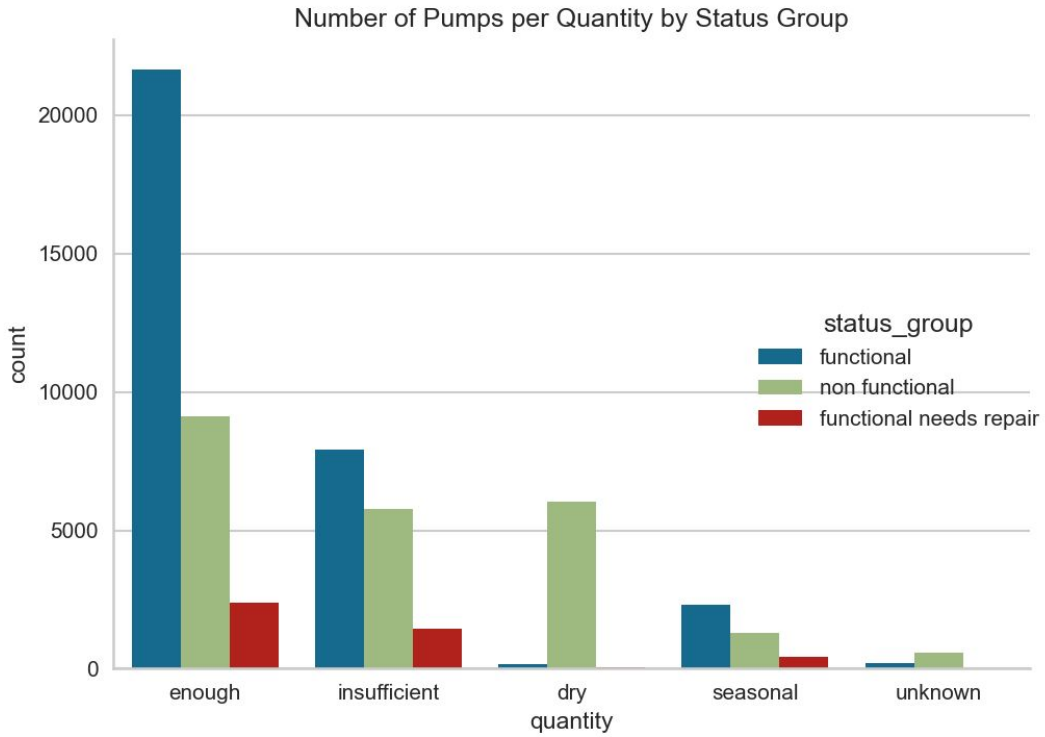
Results





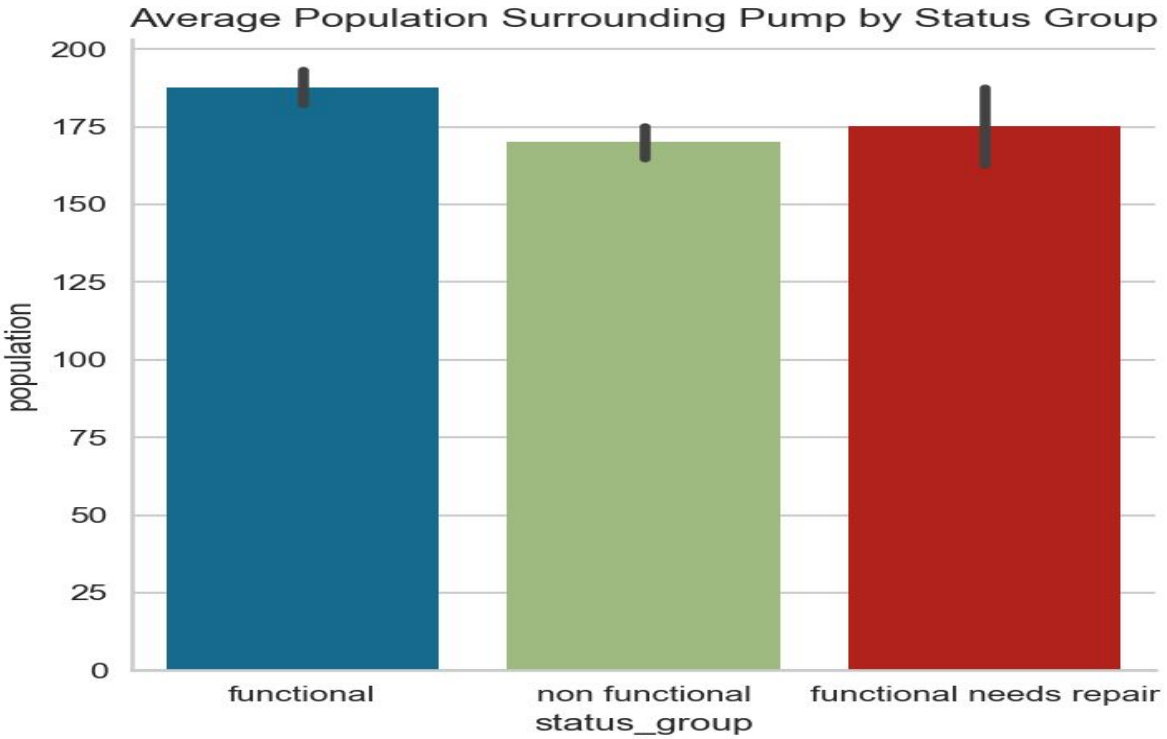
- The figures above show pump location sorted by pump status and altitude. Non functional status and higher altitude seem to correlate.

Results



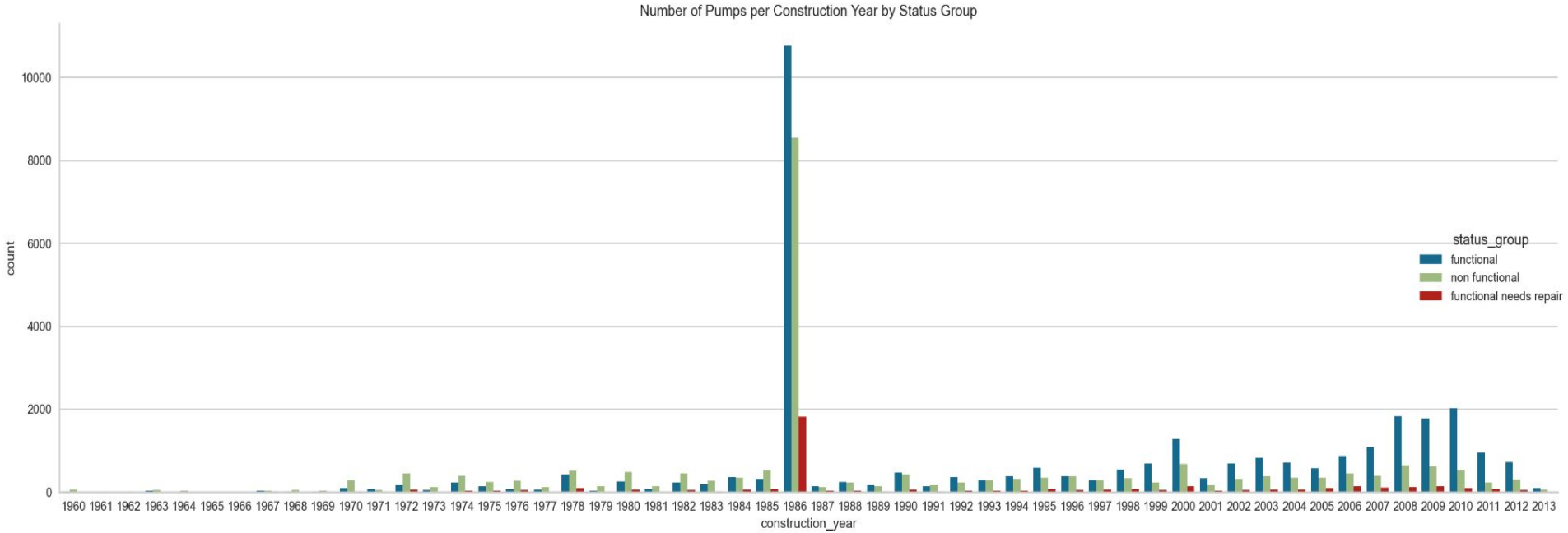
- Pumps with lower water quantities may be more likely to be non functional or needing repair.

Results



- Pumps in lower population areas may be more likely to be non functional or needing repair.

Results



- Older pumps may be more likely to be non functional or needing repair.

Conclusions

- **Location**: The Ministry should focus resources on higher altitude pumps.
- **Quantity**: The Ministry should focus resources on pumps with low quantities of water.
- **Population**: The Ministry of Water should focus resources on low population areas, as they may not be receiving enough.
- **Construction Year**: The Ministry should focus resources on modernizing older pumps

Next Steps

- The model and analysis are not complete solutions
- Model still struggles with identifying ‘functional needs repair’ pumps
- Model is slightly over fit
- Scrub data further, create more features
- Use XGBoost, LightGBM, or Catboost model to improve accuracy, reduce overfitting, and reduce computation time

Thank You!

Email: norr.peder@gmail.com
GitHub: @pederknorr
LinkedIn: linkedin.com/in/pedernorr/