Tanzanian water pumps status analysis

Flatiron School
Vadym Byesyedin
Spring 2021

This analysis was done to better understand and find ways to improve water pumps operation and consequently quality of life in Tanzania.

Analysis was performed through EDA using data from Taarifa and the Tanzanian Ministry of Water.

Next water points features will be checked against pumps status:
 - Year built
 - Payment type
 - Extraction type
 - Height above sea level
 - Geographical location

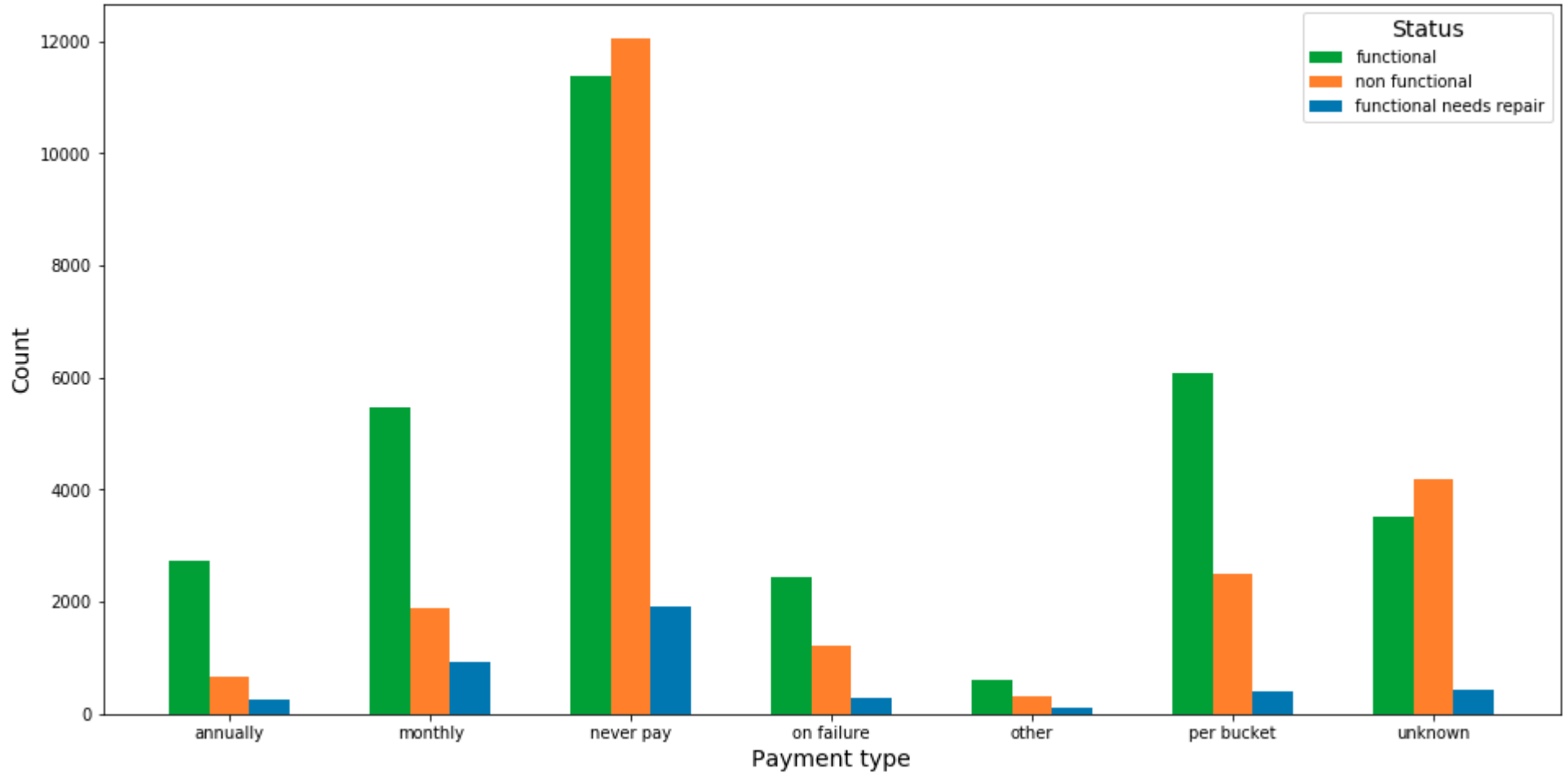Number of waterpoints by operating condition and construction year
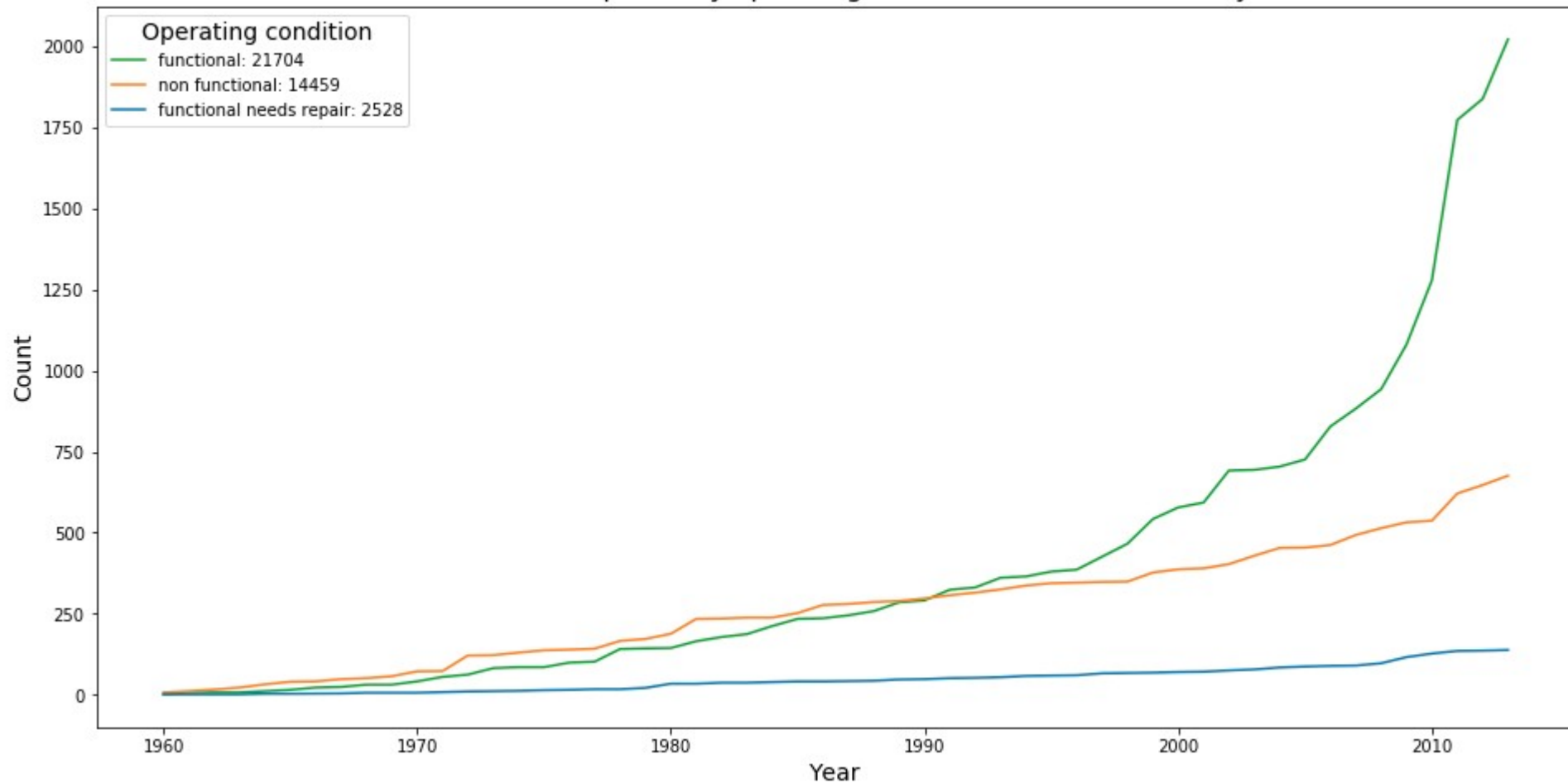
After 1990 number of functional water pumps became more than non-functional.
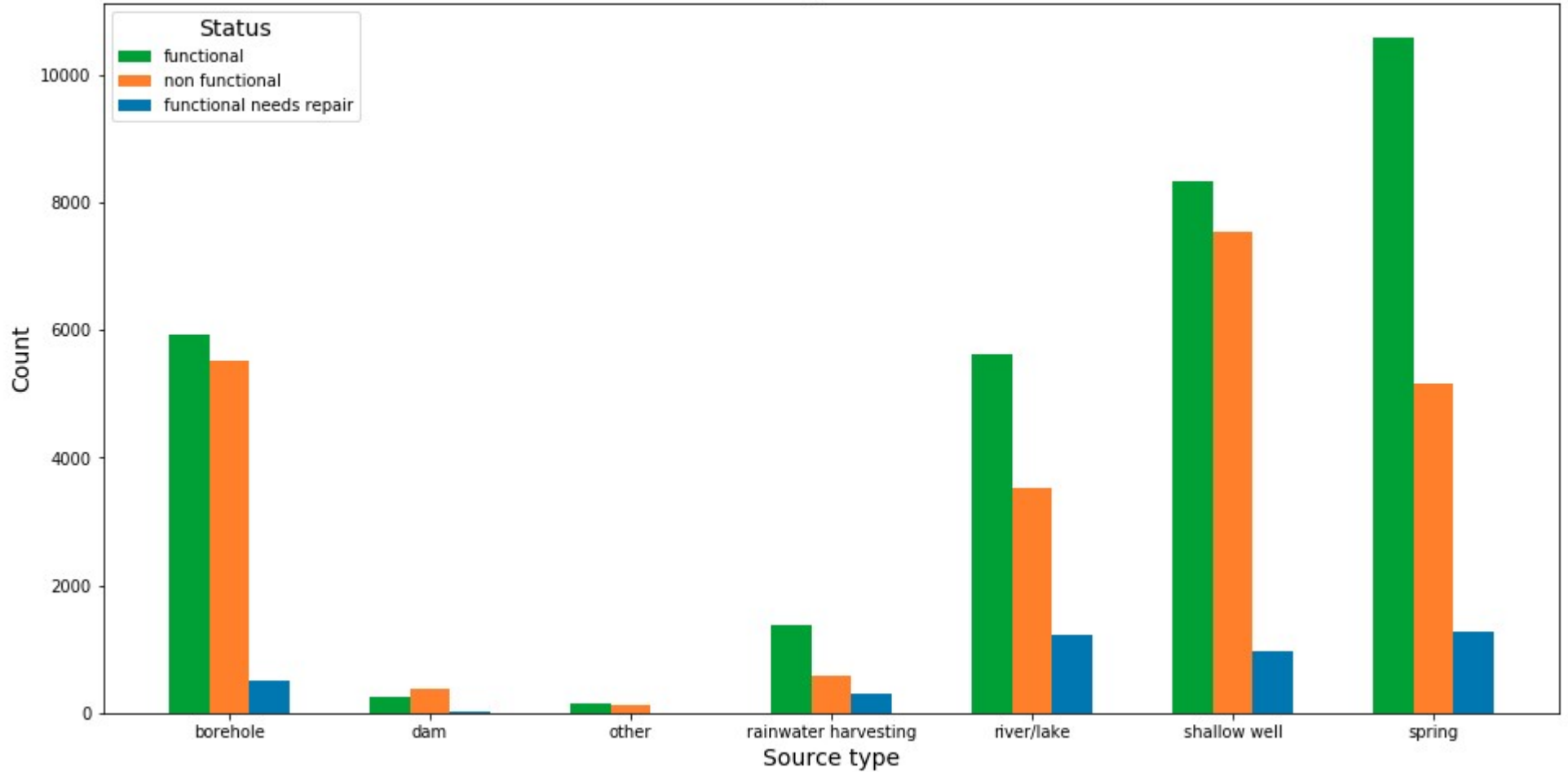After 2005 number of water pumps increased dramatically.
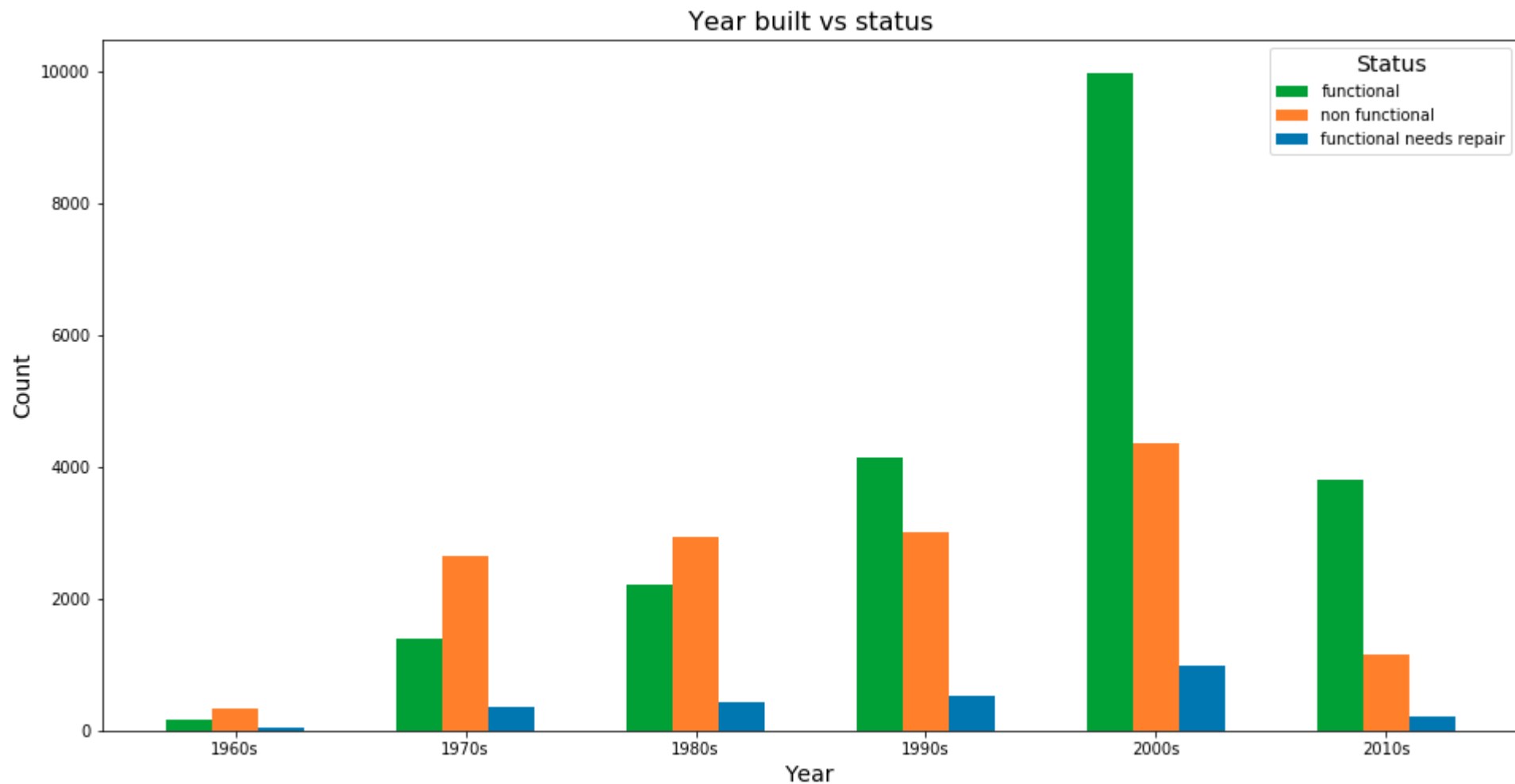
Above histogram confirms that a significant number of water points was built between 2000 and 2010

Payment type vs status

There is dependence between payment type and pump status.
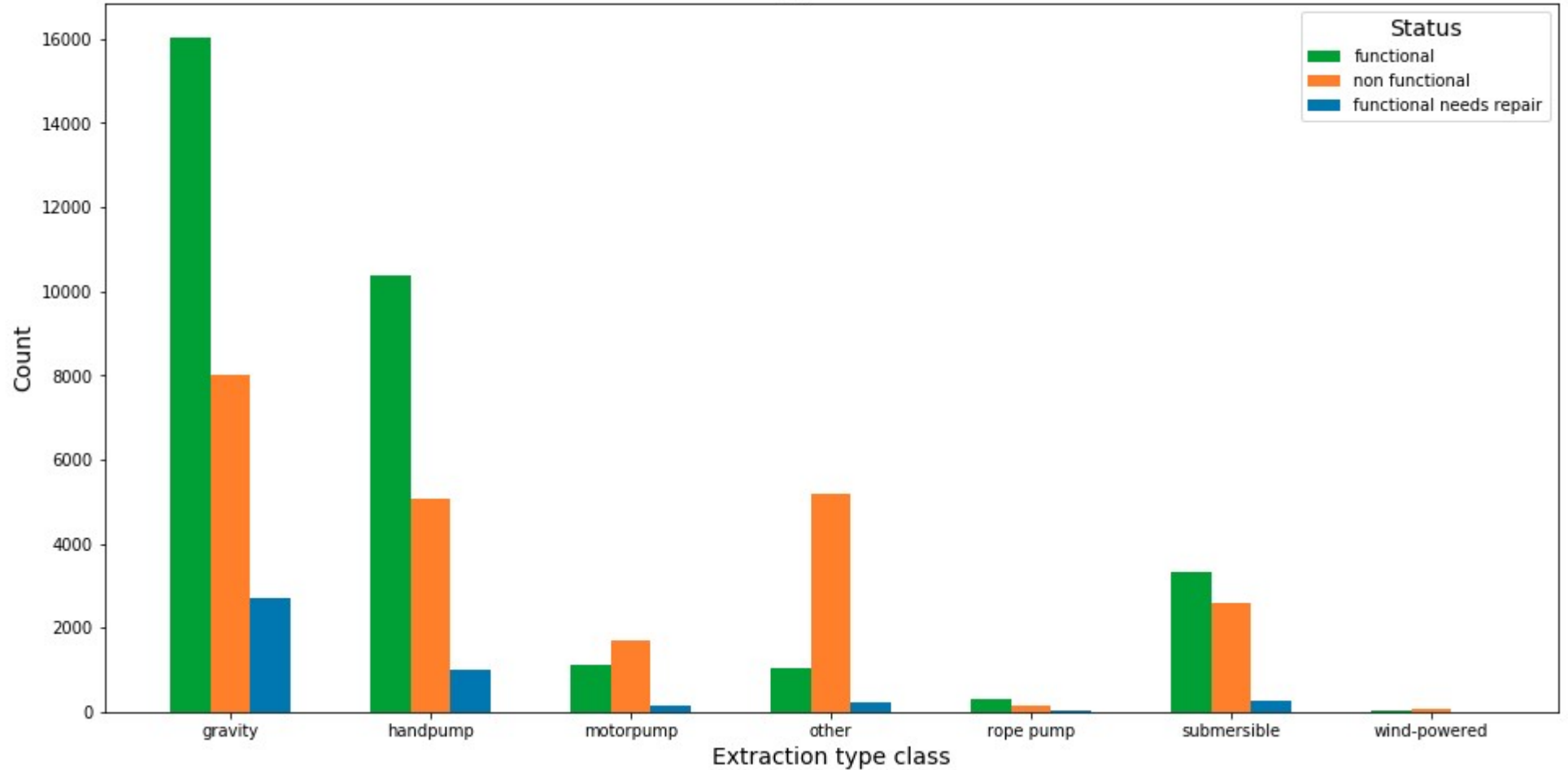'never pay' and 'unknown' types have more non functional pumps then functional.

Source type vs status
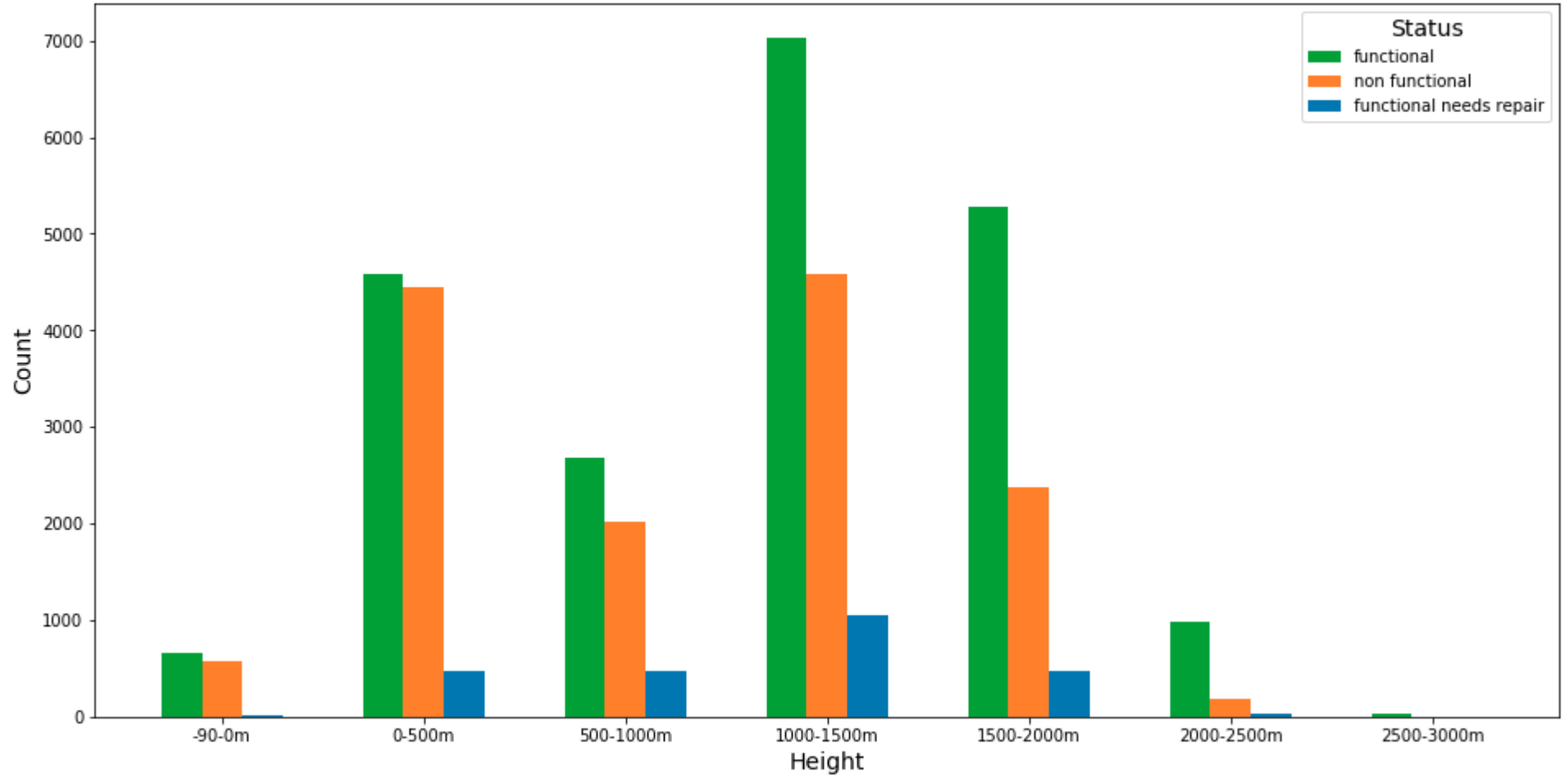
'dam' source type has more non functional water pumps.
'spring' source type has best functional to non functional ratio.
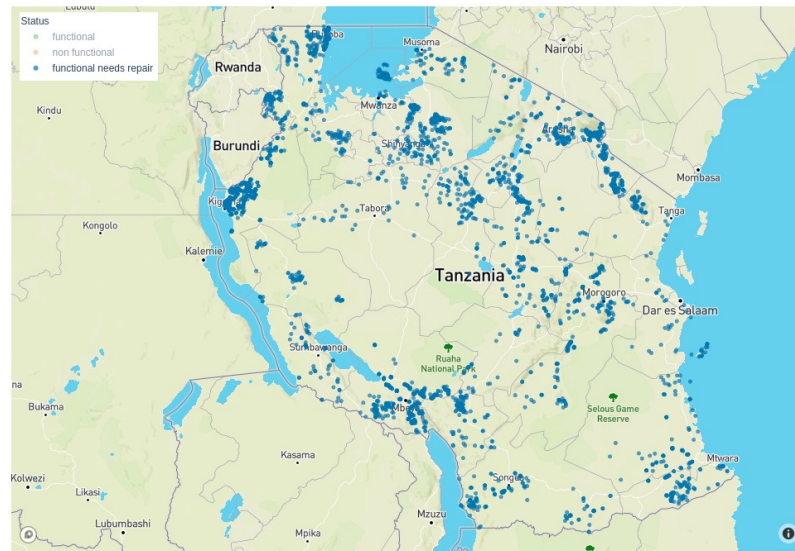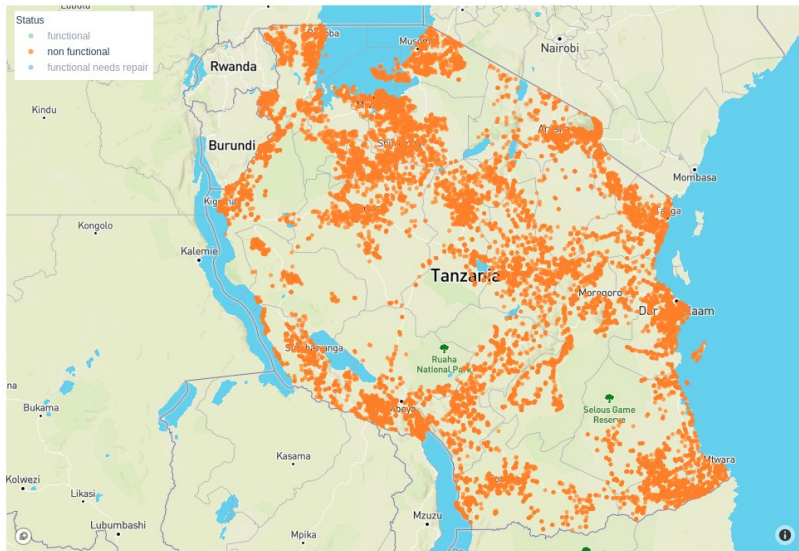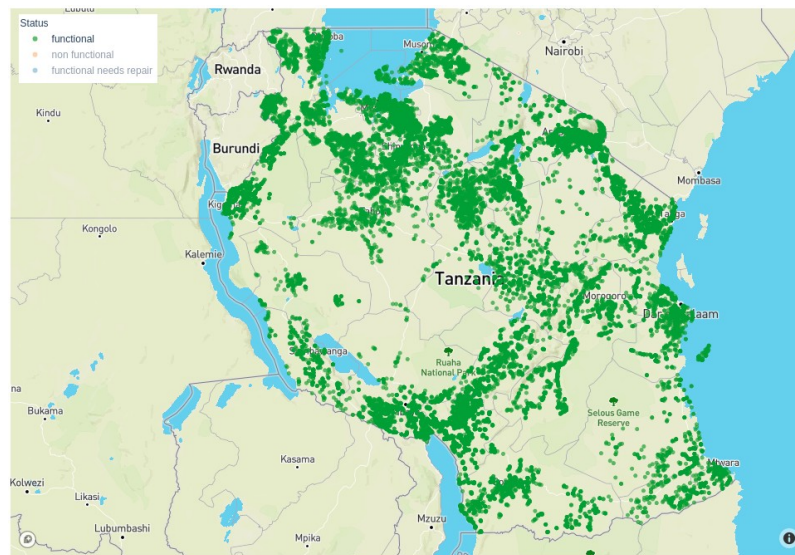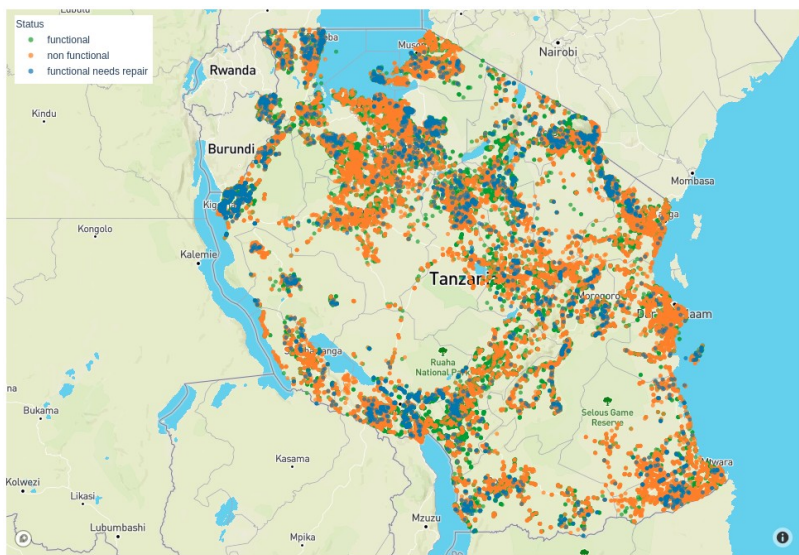
Extraction type class vs status

Most frequent and reliable are 'gravity' and 'handpump' extraction types.
'motorpump' and 'wind-powered' have more non functional than functional pumps.

Height vs status

There is no visible dependency between height above sea level and status of water point

No visible dependency between location and status of water point

Conclusion:

1. Water point year built greatly affects pumps status.

2. Water points with monthly and per bucket payment type have the highest functional to non functional ratio.

3. Spring source type has much higher ratio of functional to non functional water pumps in comparison to the rest.

4. Gravity and handpump are most frequent and most reliable extraction type.

5. There is no visible dependency neither between height above sea level nor geographical location and status of water pump.

Tuned Random Forest Classifier performs best of all with highest accuracy score **0.802**

Model's confusion matrix
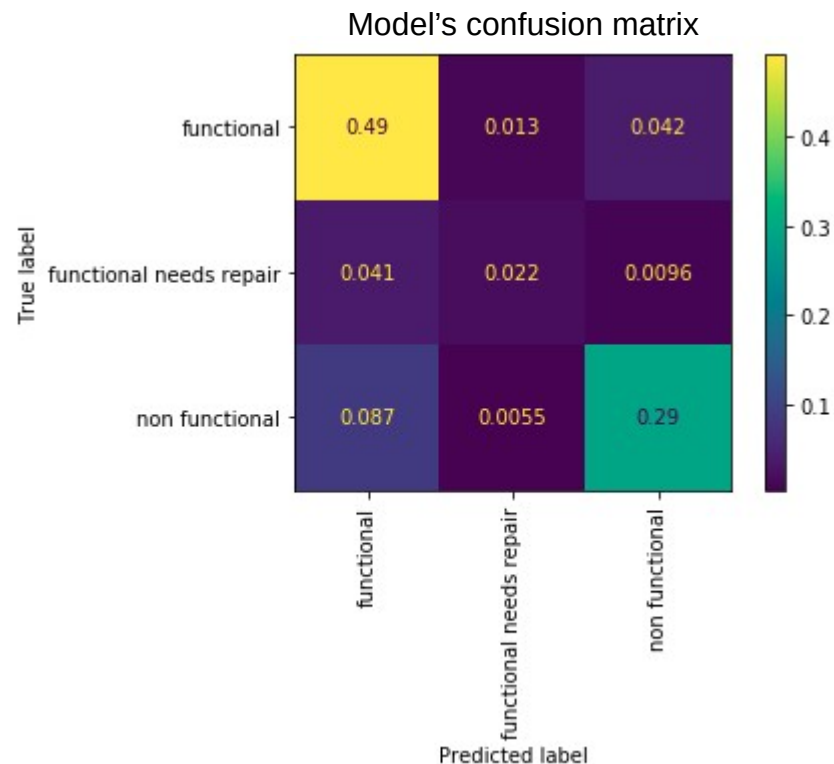
	Classifier	Train score	Test score
2	RandomForestClassifier	0.936	0.802
1	XGBClassifier	0.916	0.801
0	BaggingClassifier	0.891	0.795

Future work:

In the future, analysis of impact of population on pump status,
dependency between basin and water quality,
scheme management and status can be done.

Thank you!