Capstone Project 2 Milestone Report 1

The question to answer is:

What is the operation status of a water point in Tanzania: functional, needs repair, non functional?

Slightly more than half of the population has access to clean water in Tanzania. The operation and maintenance costs are difficult to bear for local government authorities. Tanzania receives external support from several donor agencies.

The objective of this project is to predict the operation status of water points. This will help reducing operation and maintenance cost, while improving continuity of supply.

This is a DrivenData competition.
The dataset used comes from Taarifa and the Tanzanian Ministry of Water.

The dataset contains records of 59400 water points.
Each record has the following information about the water point:

- `amount_tsh` - Total static head (amount water available to waterpoint)
- `date_recorded` - The date the row was entered
- `funder` - Who funded the well
- `gps_height` - Altitude of the well
- `installer` - Organization that installed the well
- `longitude` - GPS coordinate
- `latitude` - GPS coordinate
- `wpt_name` - Name of the waterpoint if there is one

- `num_private` -
- `basin` - Geographic water basin
- `subvillage` - Geographic location
- `region` - Geographic location
- `region_code` - Geographic location (coded)
- `district_code` - Geographic location (coded)
- `lga` - Geographic location
- `ward` - Geographic location
- `population` - Population around the well
- `public_meeting` - True/False
- `recorded_by` - Group entering this row of data
- `scheme_management` - Who operates the waterpoint
- `scheme_name` - Who operates the waterpoint
- `permit` - If the waterpoint is permitted
- `construction_year` - Year the waterpoint was constructed
- `extraction_type` - The kind of extraction the waterpoint uses
- `extraction_type_group` - The kind of extraction the waterpoint uses
- `extraction_type_class` - The kind of extraction the waterpoint uses
- `management` - How the waterpoint is managed
- `management_group` - How the waterpoint is managed
- `payment` - What the water costs
- `payment_type` - What the water costs
- `water_quality` - The quality of the water
- `quality_group` - The quality of the water
- `quantity` - The quantity of water

- `quantity_group` - The quantity of water
- `source` - The source of the water
- `source_type` - The source of the water
- `source_class` - The source of the water
- `waterpoint_type` - The kind of waterpoint
- `waterpoint_type_group` - The kind of waterpoint

The training dataset is labelled in 3 categories: functional, functional needs repair, non functional.

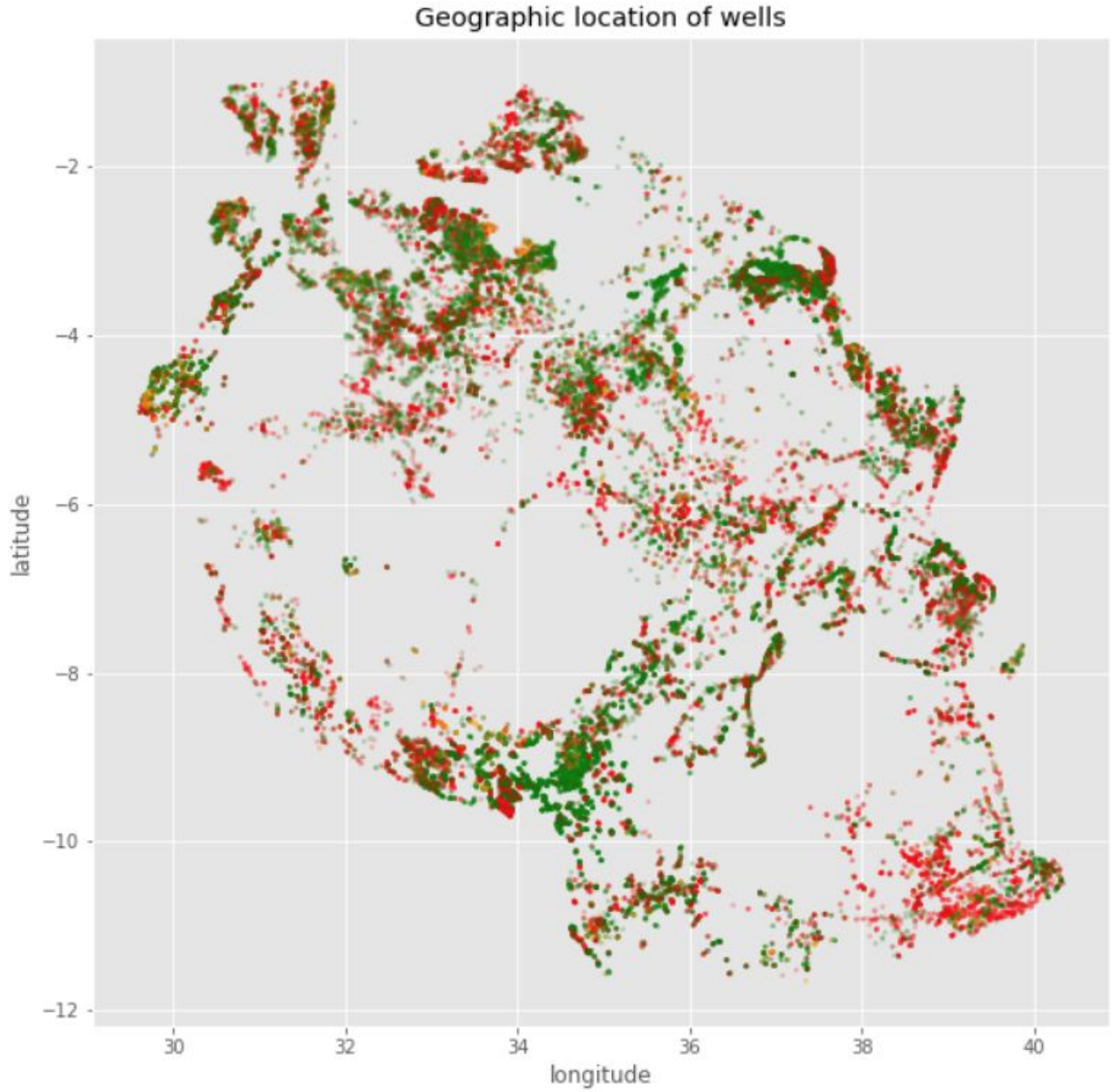
This project consists of data exploration and multiclass classification.
I plan to use one vs. rest classification technique.

The deliverables will be Jupyter Notebook, a technical report and a presentation of the project.

Data exploration and analysis

Geographic features:

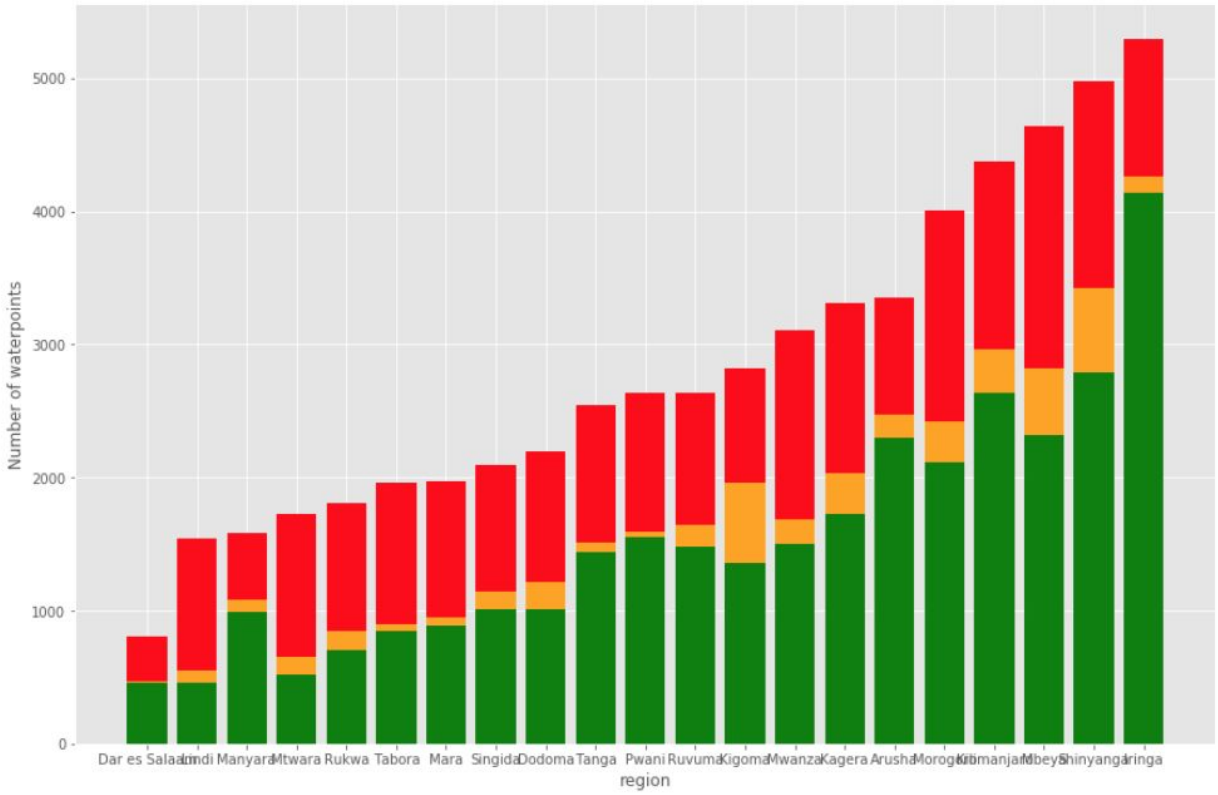
This dataset contains a lot of geographic features.
Let's explore some of them.
First, the latitude and longitude will give us an overview of the country.



The color code is green for functional, orange for functional but needs repair and red for non functional.

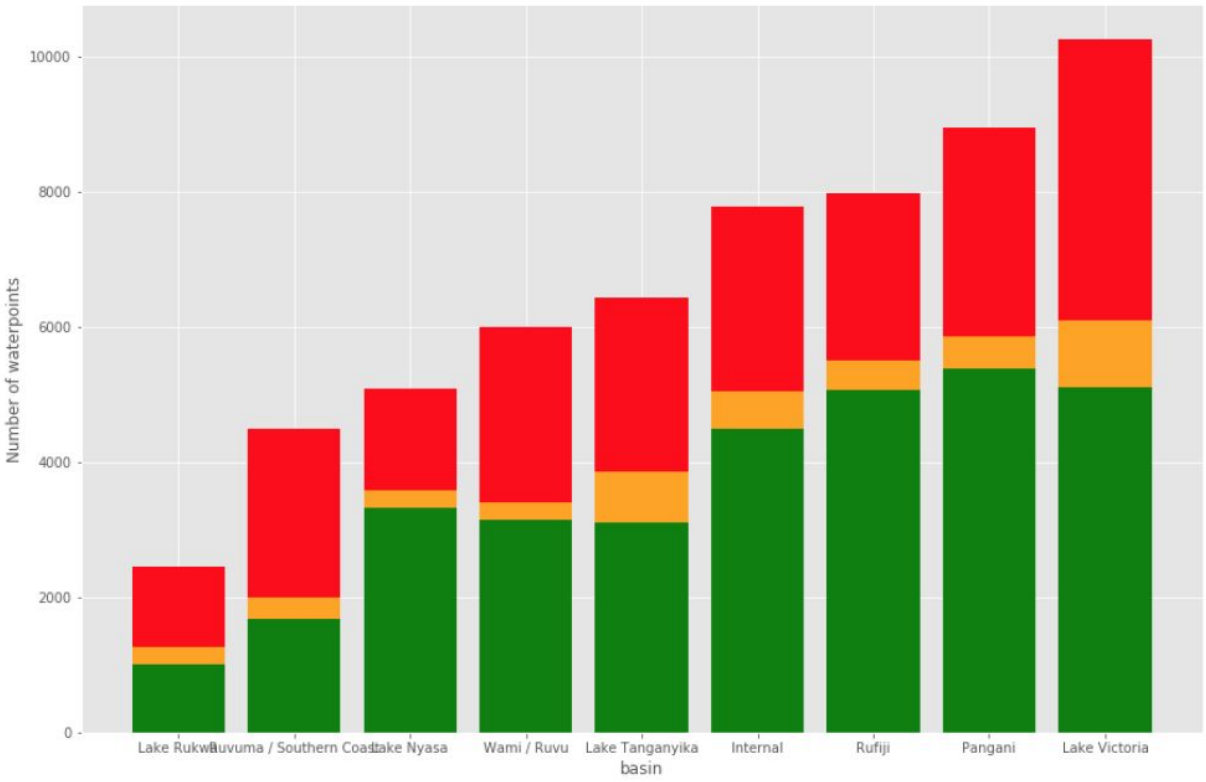
There is a majority of non functional water points at the South East of Tanzania.
There is 1812 missing coordinates that I replace with the mean of the region.

There are 21 regions in Tanzania. Here is the repartition of water points for each region:



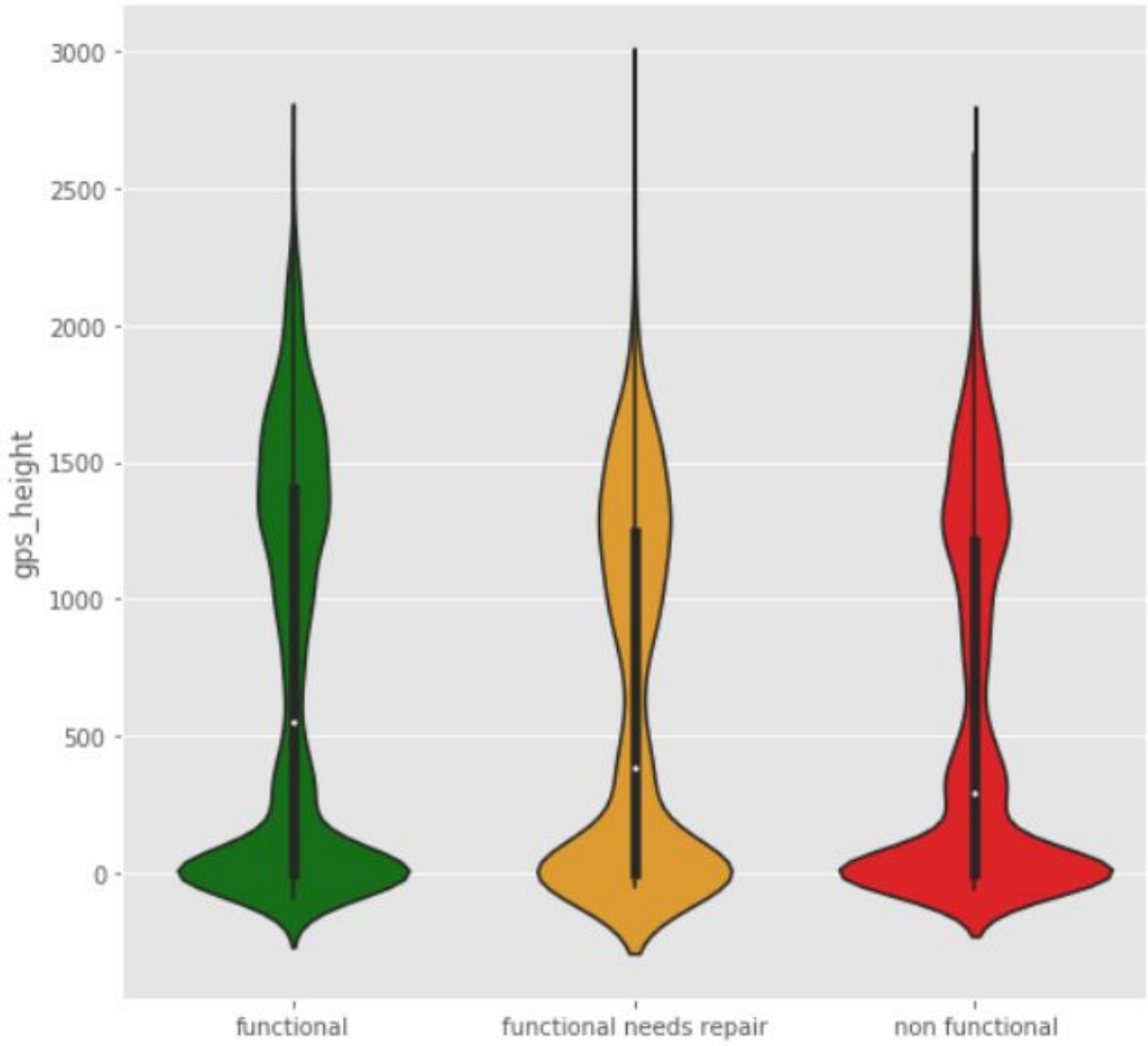
We can see that the proportions of the 3 operational status are different for the different regions. The chi-square test for independence confirms that the operational status repartition is different according to the region.

Tanzania is divided in 9 water basins, and they have a different operation status repartition as well, as shown by the plot and the chi-square test.



The other geographical features: lga, ward, district are also related to the operational status, as proved by the chi-square test for independence.
Ward as 2092 different values, so we will keep the 50 most frequent ones, and create a ‘other’ category.
We will get rid of region_code because it’s similar to region, and of subvillage because it has 19287 different values, with a lot of them referring only to one waterpoint.

The last geographic feature is gps_height:



The violins look very similar, so let's test the independence of each status against the others.

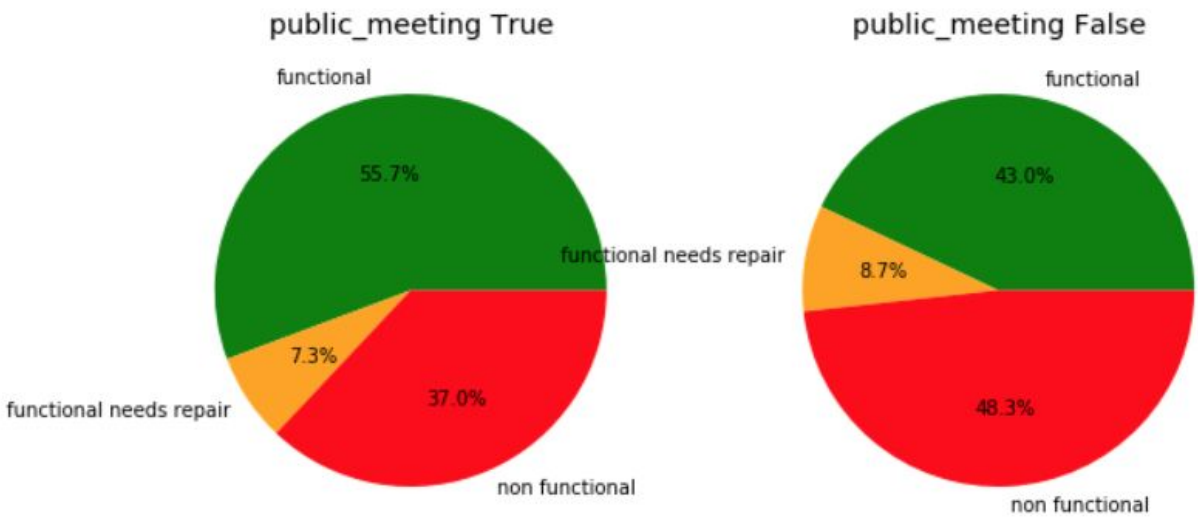
	t_stat	p_value
functional vs other	27.7151	5.42774e-168
functional needs repair vs other	-4.00602	6.18255e-05
non functional vs other	-26.2139	1.33281e-150

The low p_values show the repartition of gps_height for these groups is actually very different.

Construction features:

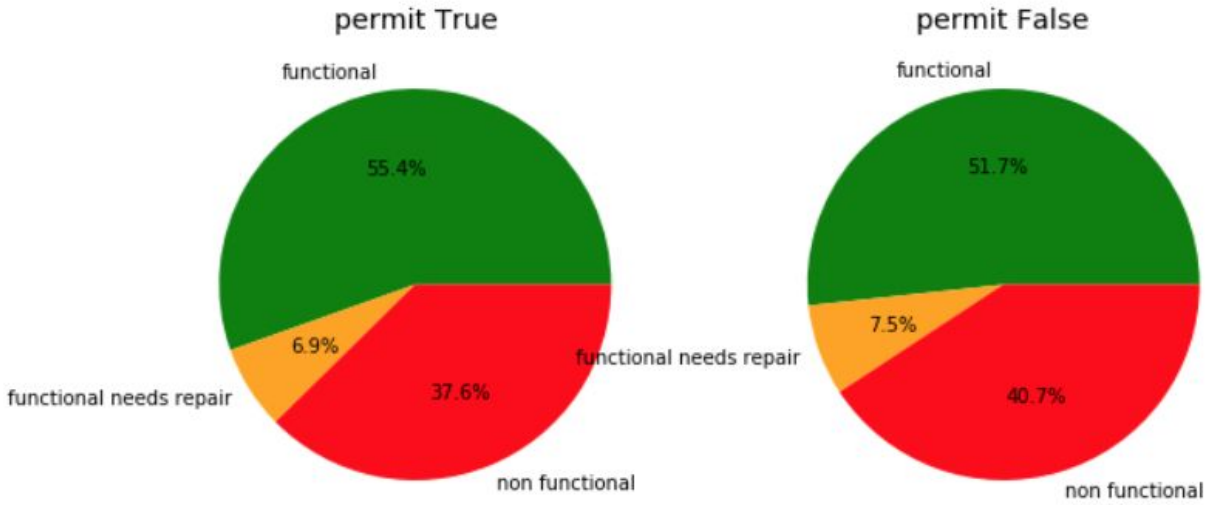
The funder and installer features are the names of the organizations that funded and installed the well. They have each about 2000 different values, with missing values and spelling mistakes. Let's fix some of the spelling mistakes, then keep the 50 most frequent organizations and name the rest of them 'other'.

56066 records report if there was a public meeting:



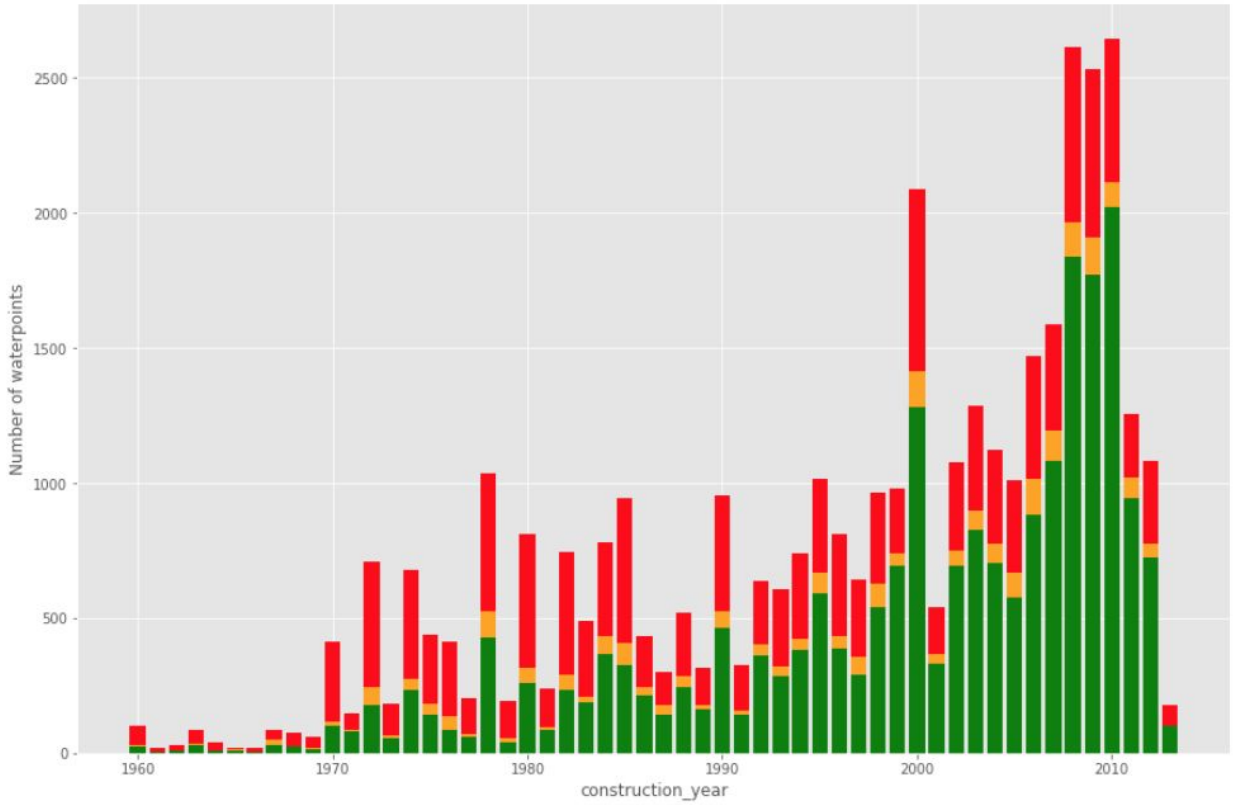
The proportion of functional water points is higher when there was a public meeting, so we will use this feature. Where it's missing, we set it to False.

56344 records report if there is a permit:



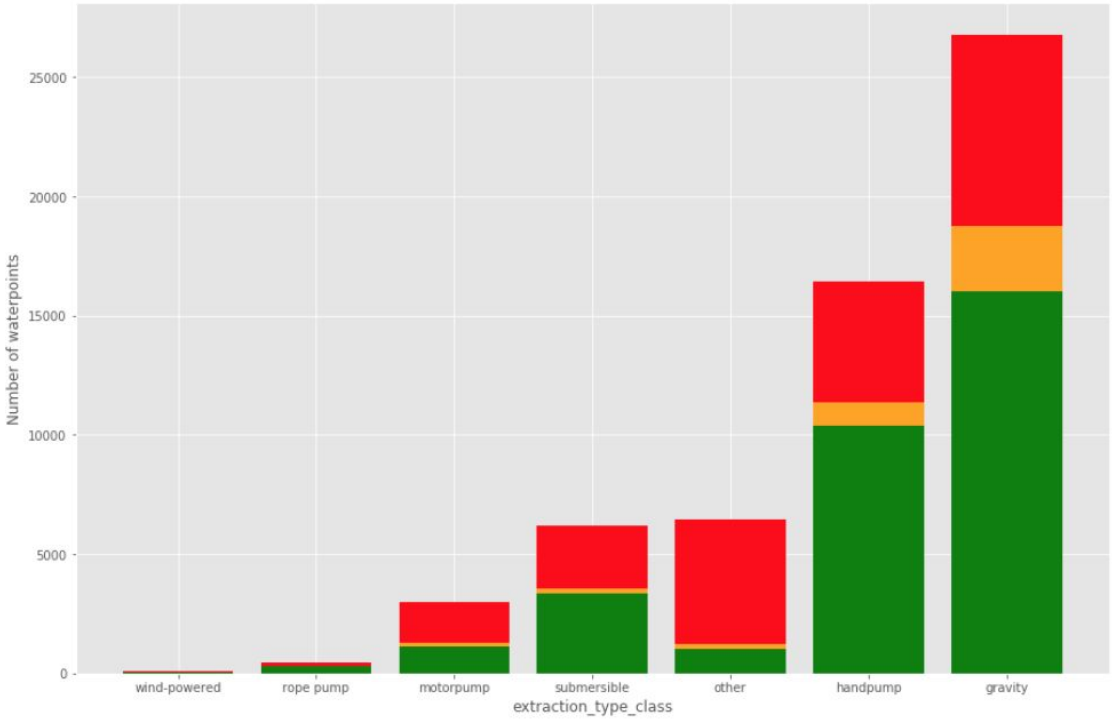
The proportions are not very different, but with the chi-square test, we can conclude that the permit and operational status are related.
Where this information is missing, we set it to False.

Here is a bar plot of the construction years:



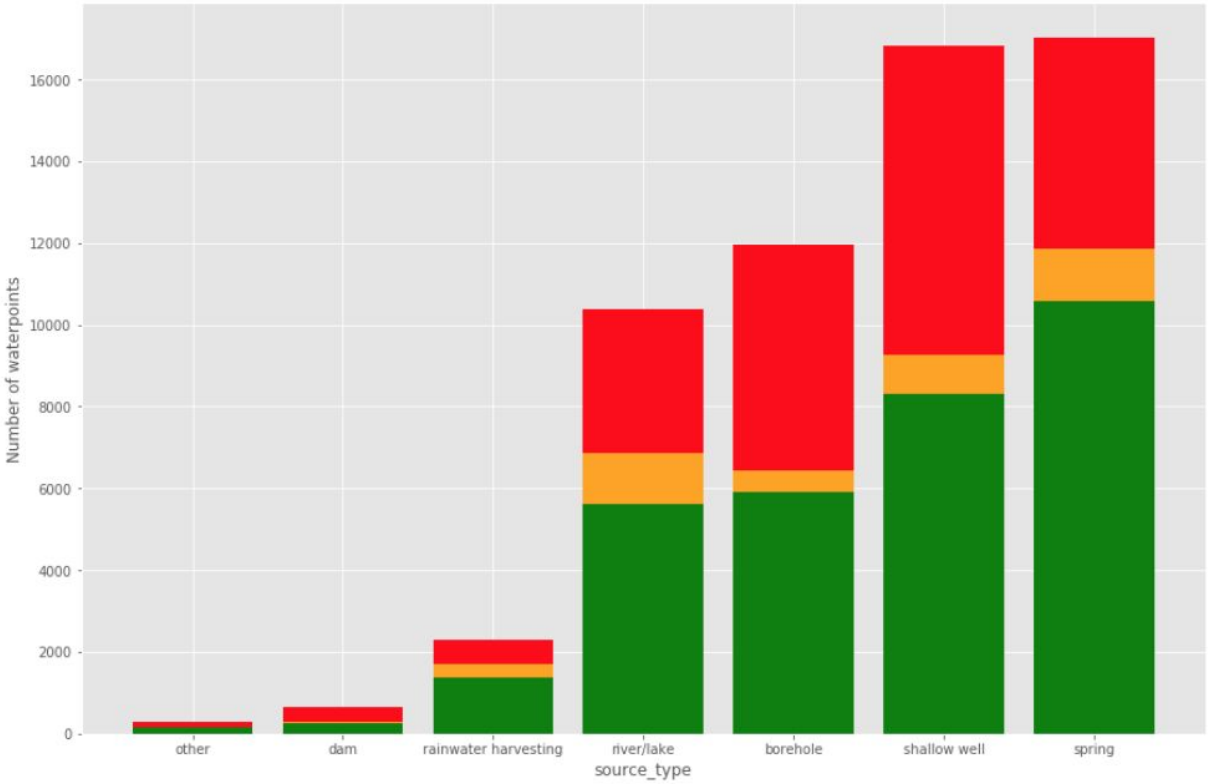
Older water points are more likely to be non functional.
Some construction years are missing. We set them to the mean of the construction years: 1997.

3 features describe the extraction type. We will keep only the extraction type class:



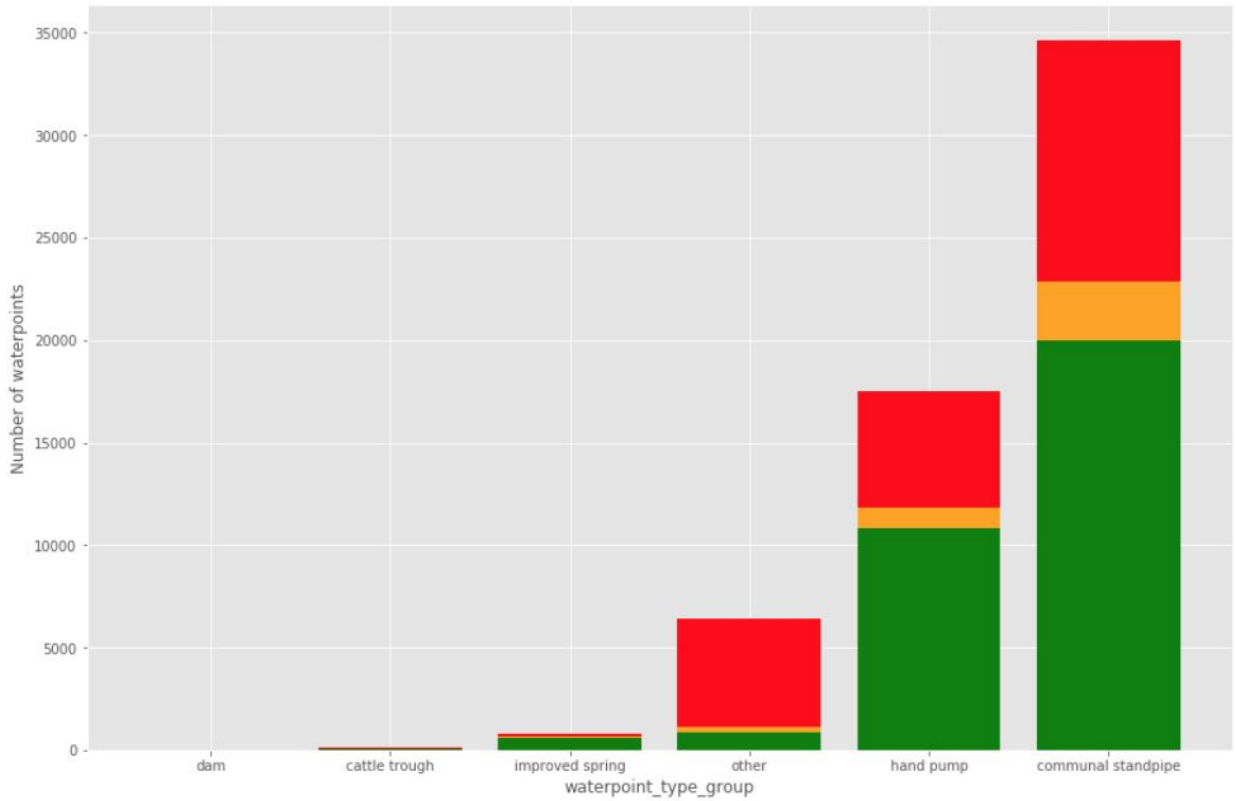
The repartition for these extraction types are very different. The chi-square test for independence shows a relation with the operational status.

3 features describe the source of the water. We will keep only the source type:



Shallow wells and boreholes are more likely to be non functional. Here again, the chi-square shows relation with the operational status.

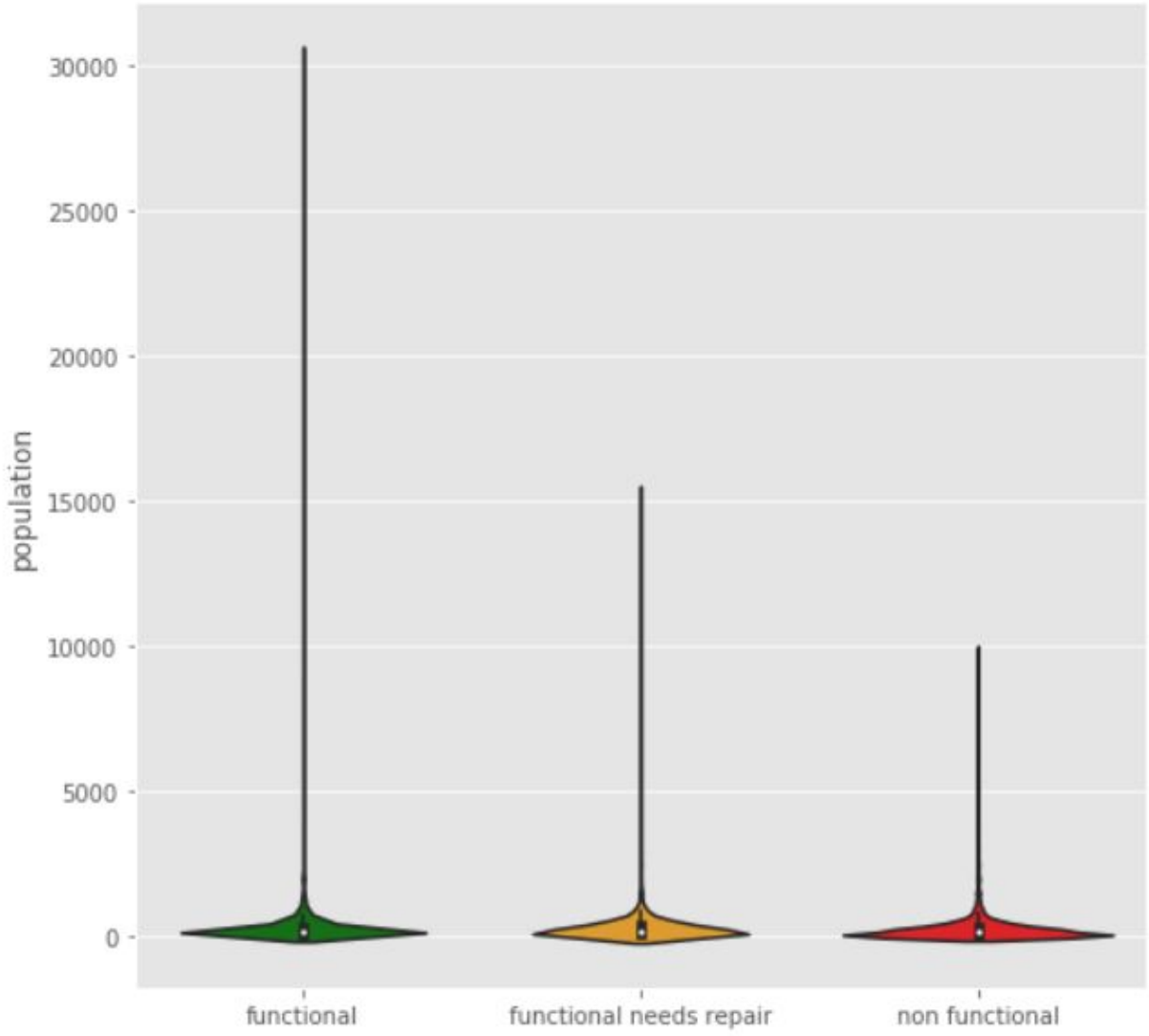
2 features describe the water point type. We will keep only waterpoint_type_group:



Chi-square test for independence shows the water point type and operational status are related.

Exploitation features:

Here is a violin plot of the population around the water points:

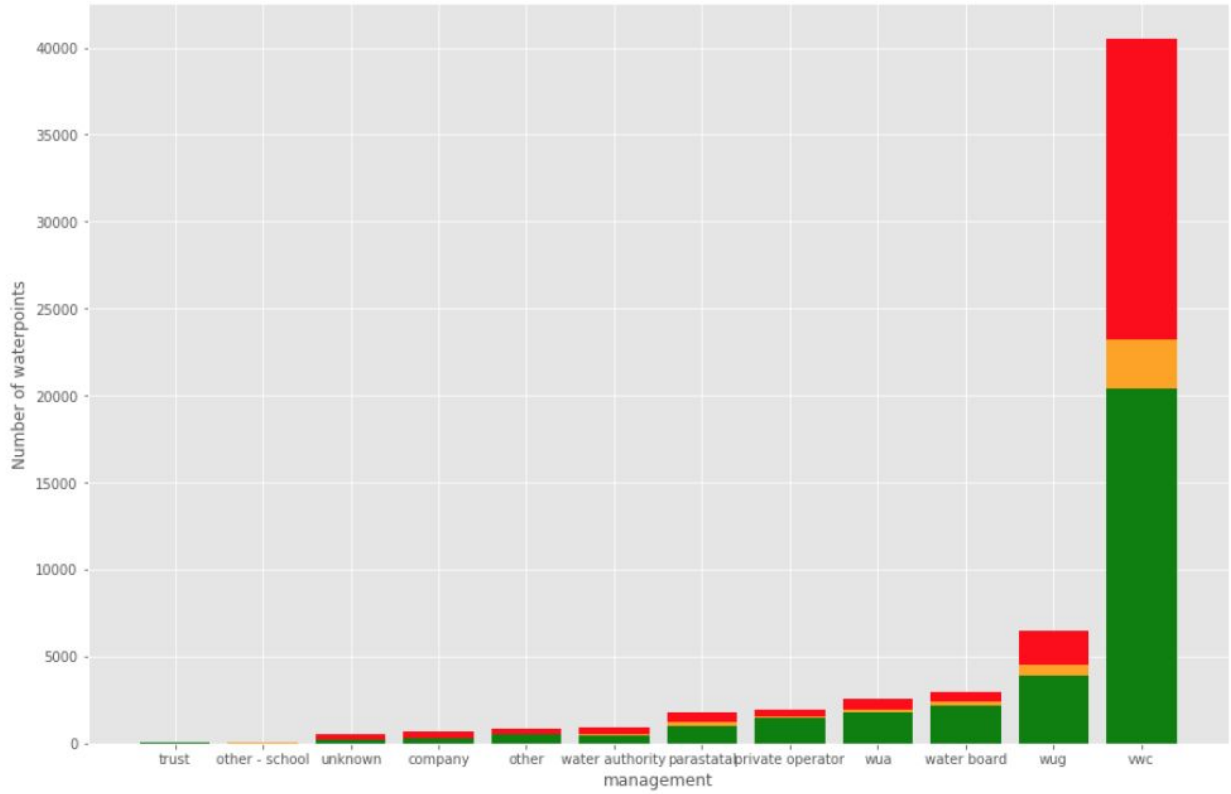


Let's test the independence of each category against the others:

	t_stat	p_value
functional vs other	2.7702	0.00560495
functional needs repair vs other	1.50508	0.132312
non functional vs other	-3.61082	0.00030562

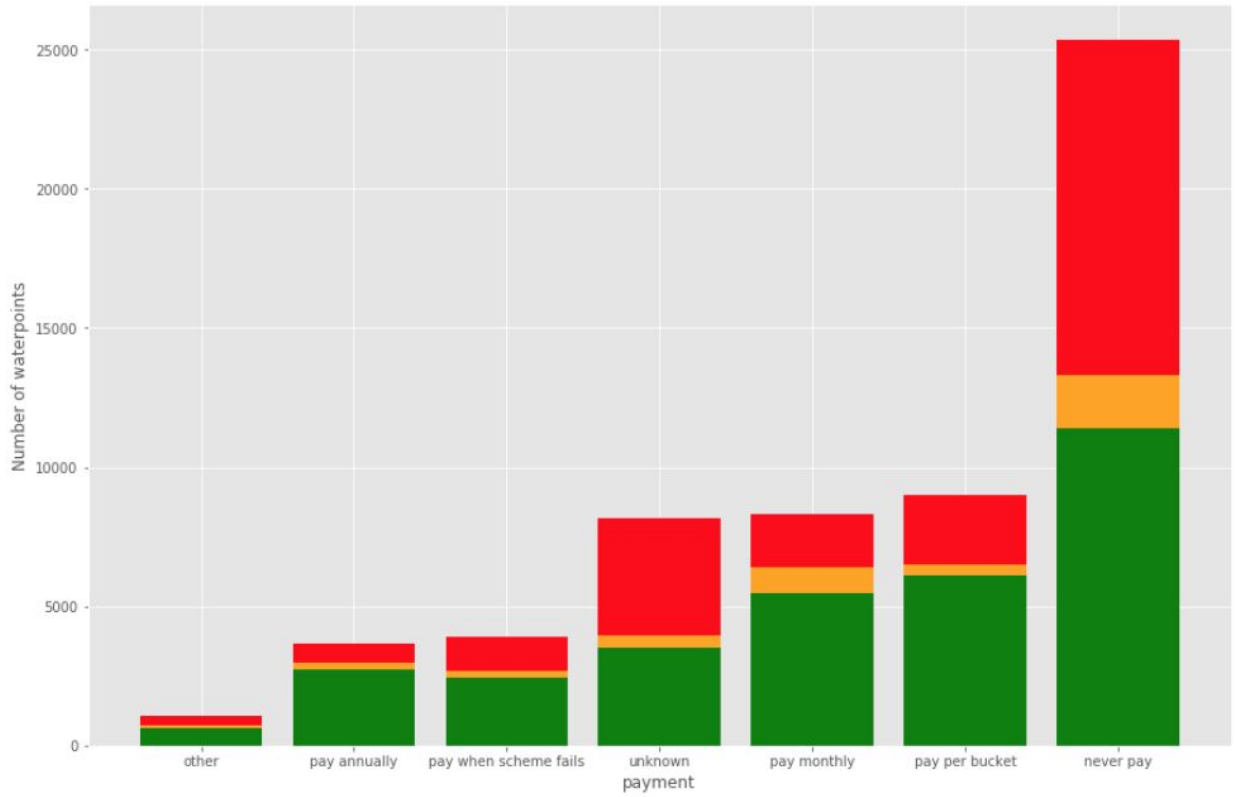
The low p_values for functional and non functional water points shows the population for these groups is very different.
The high p_value for functional needs repair shows the population for this group is not different from the rest of the waterpoints. It means it will be more difficult to predict this category accurately.
21381 values are missing, we set them to the median of the non zero values, because median is less sensitive to extreme values.

4 features describe how the water point is managed.
We drop scheme_name because half of the values are missing, scheme_management because it's redundant and has missing values, management_group because it's redundant.
We will keep only the management feature:



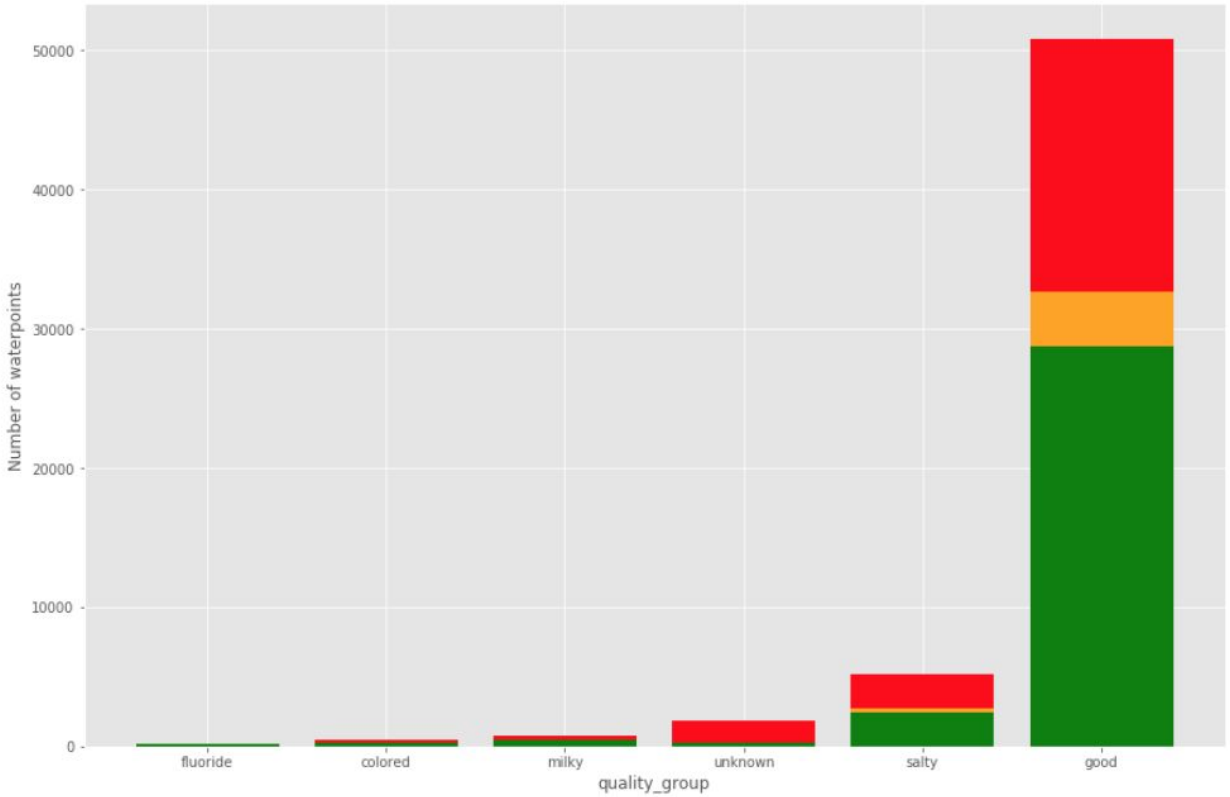
The chi-square test for independence shows that the management is related to the operational status.

2 features describe the payment type. We will keep only the payment:



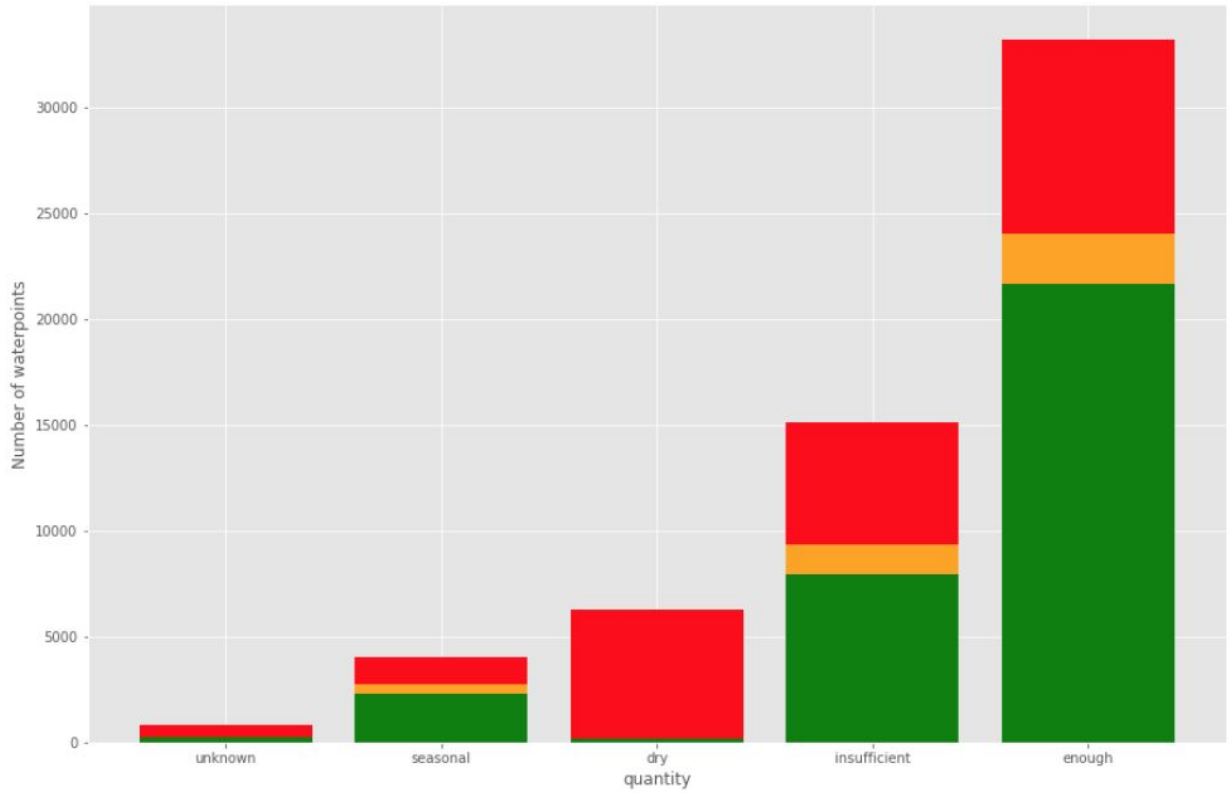
It seems that when there is payment for the water (either periodically, per bucket or on failure) the waterpoint is more likely to be functional. The chi-square test for independence confirms the relation.

2 features describe the water quality. We will keep only the quality_group, and order the categories: unknown < fluoride < salty < colored < milky < good.



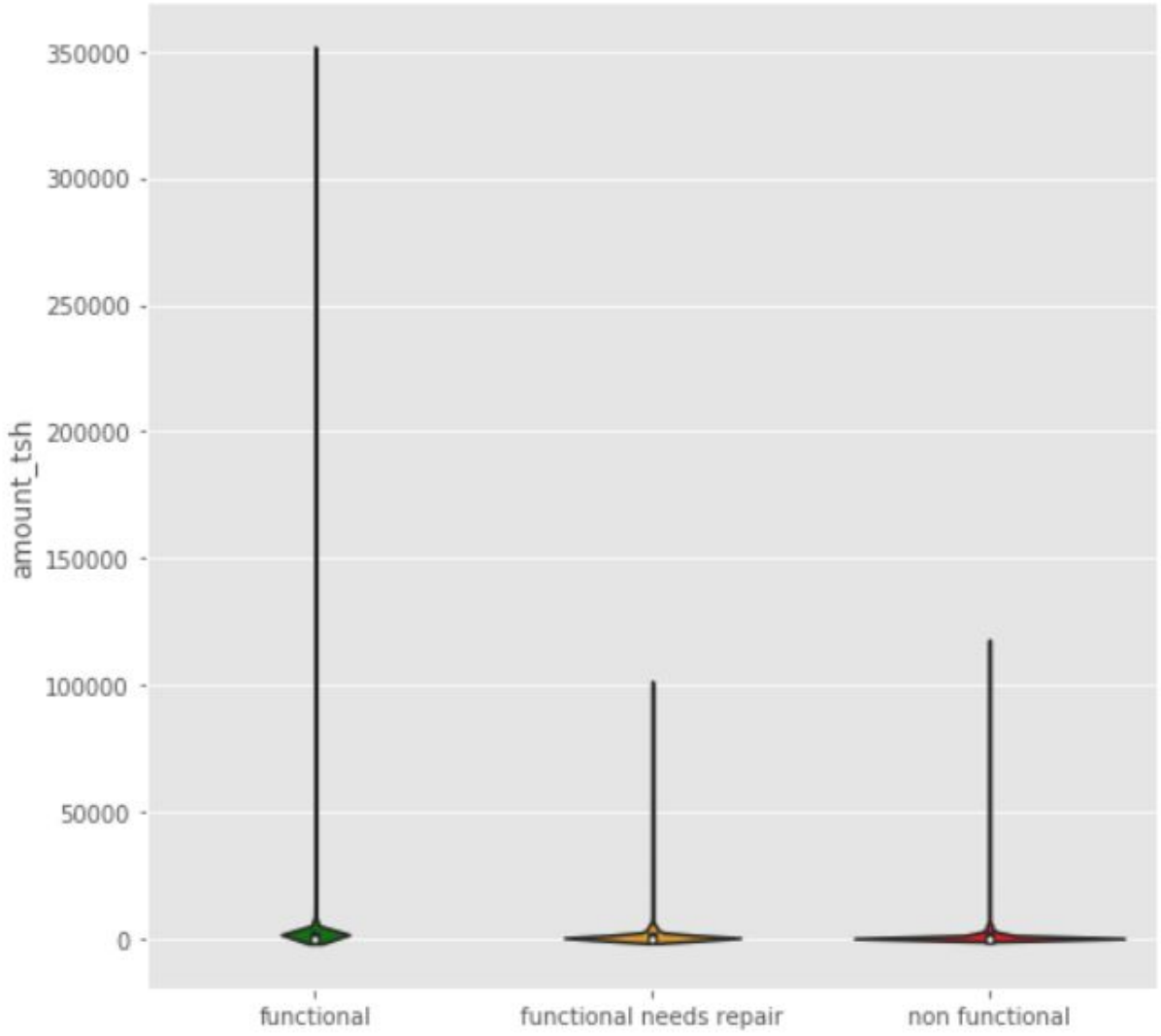
The chi-square test for independence shows the operational status is related to the water quality group.

2 features describe the water quantity. We will keep the quantity and order the categories: Unknown < dry < insufficient < seasonal < enough.



The chi-square test for independence shows the operational status is related to the water quantity.

Amount_tsh is the amount of water available to waterpoint. Most of the values are 0. When we keep only non zero values, we get this plot:

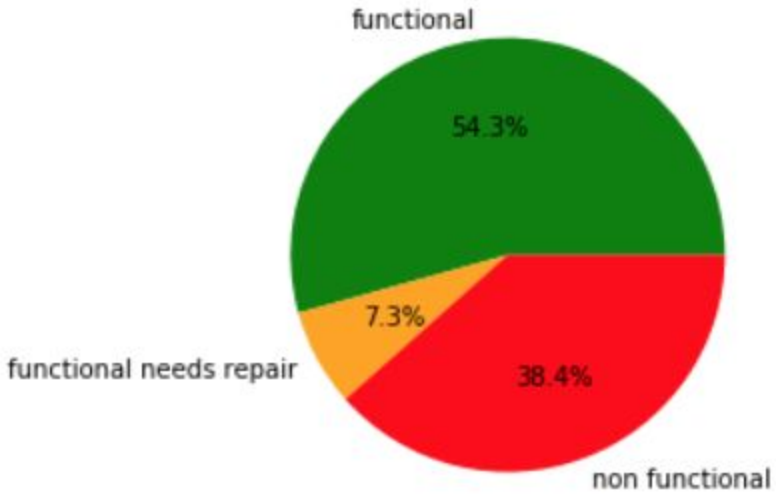


The t-test for independence of each operational status vs the others shows the functional and non functional groups are very different.
However, the p-value for functional needs repair vs the others is 0.29, which means the amount_tsh for this group is not different from the rest of the waterpoints.

We will drop the record features because they have no relation with the operational status: date_recorded, wpt_name, num_private, recorded_by.

Operational Status:

Here is the repartition of the status_group we try to predict:



We will order the categories: non functional < functional needs repair < functional.

The labels are not balanced, there is very few functional needs repair.
This means accuracy is not a good metric, and we should use another metric like log loss to evaluate the model.