ReDI School of Digital Integration

DataDrip

Exploratory Data Analysis (Sprint 1)

Shakeel Ahmed
Amey Chitnis

Table of Contents

Introduction

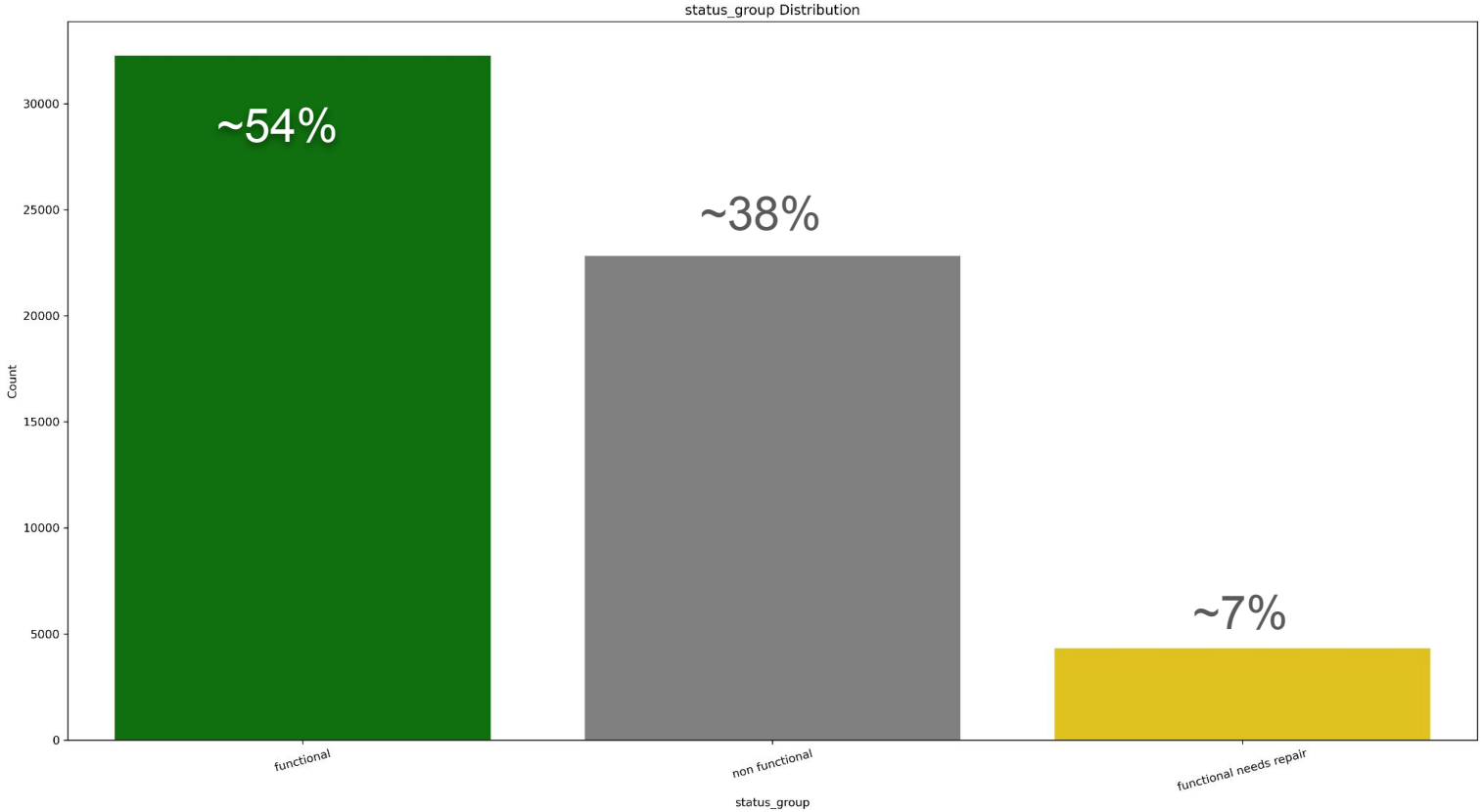
59000 water pumps in Tanzania

Features- geolocation, management, water and temporal information

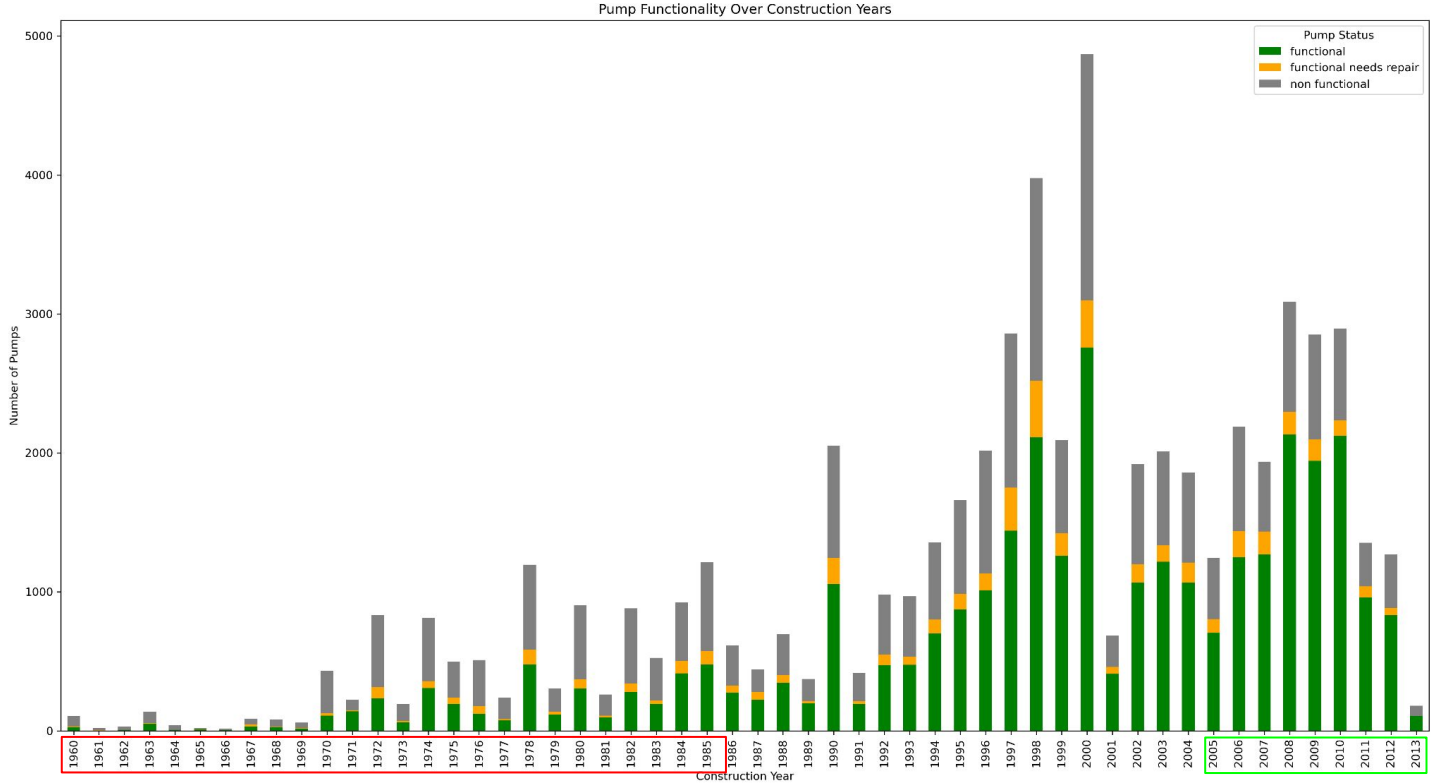
Functional status- Working, not working, needs repair

Several missing values: NaN, 0s and 1s

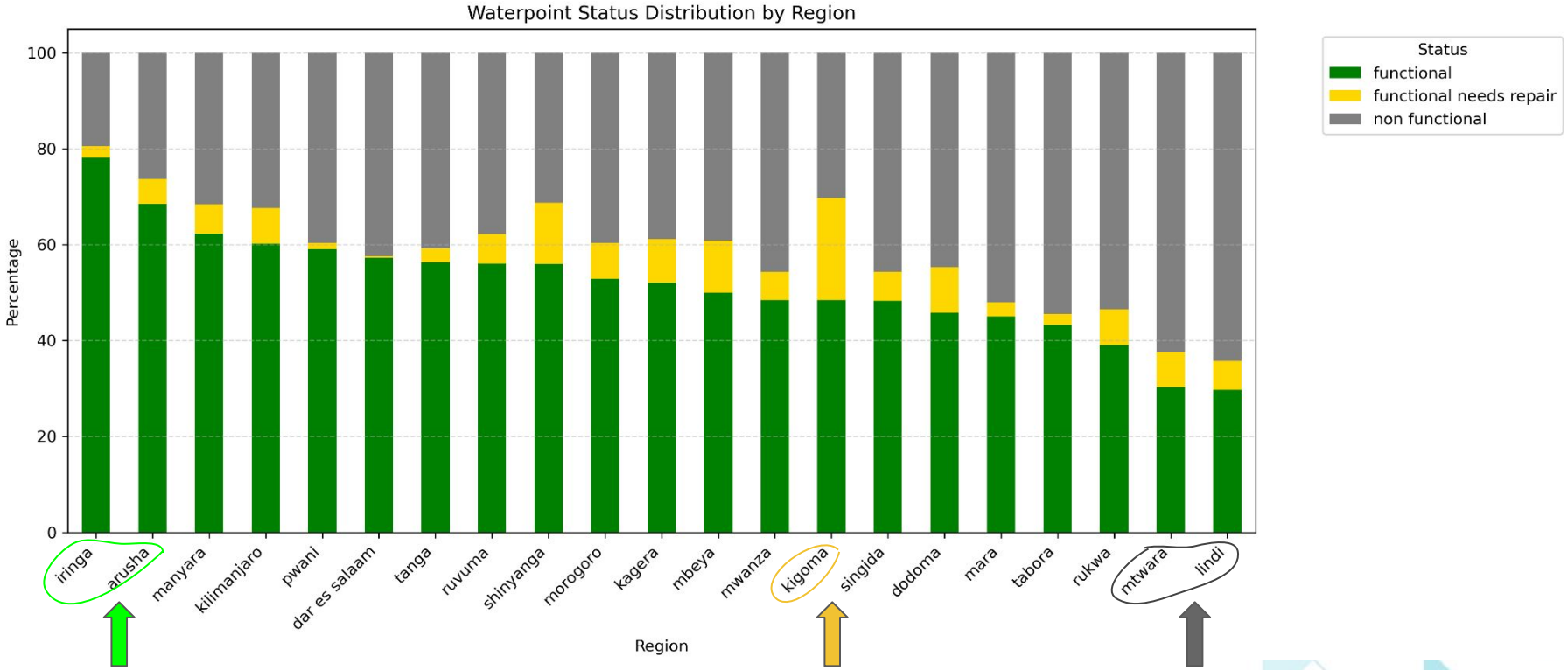
Dataset Imbalance!



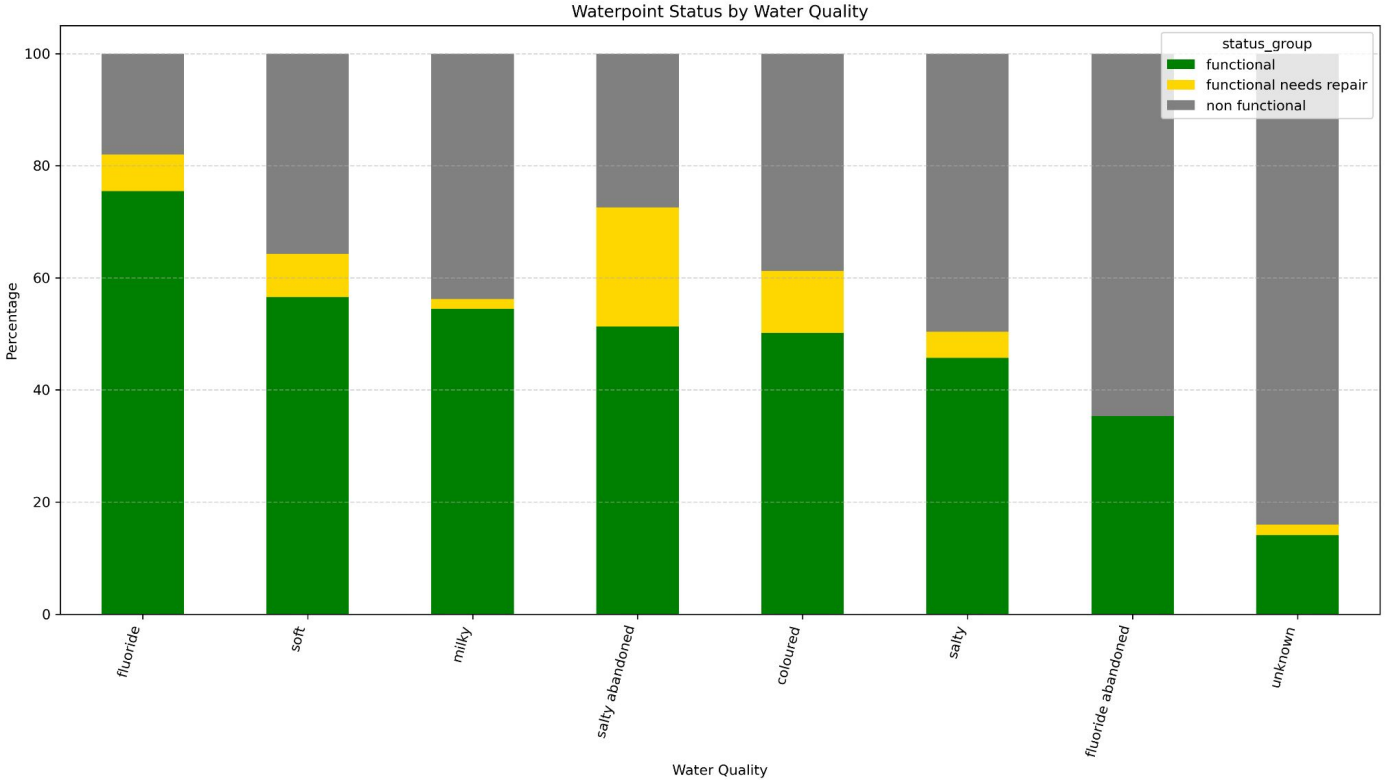
Construction year



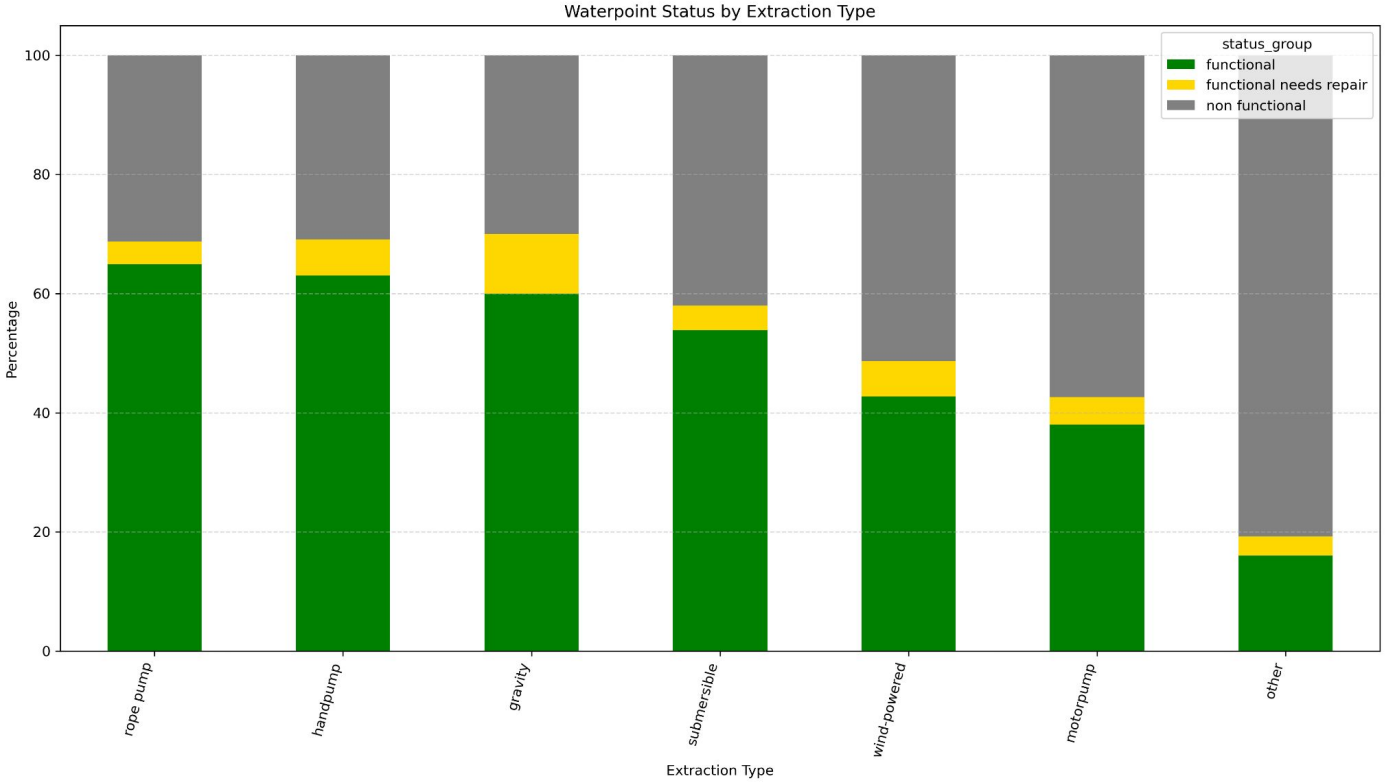
Functional pumps per region



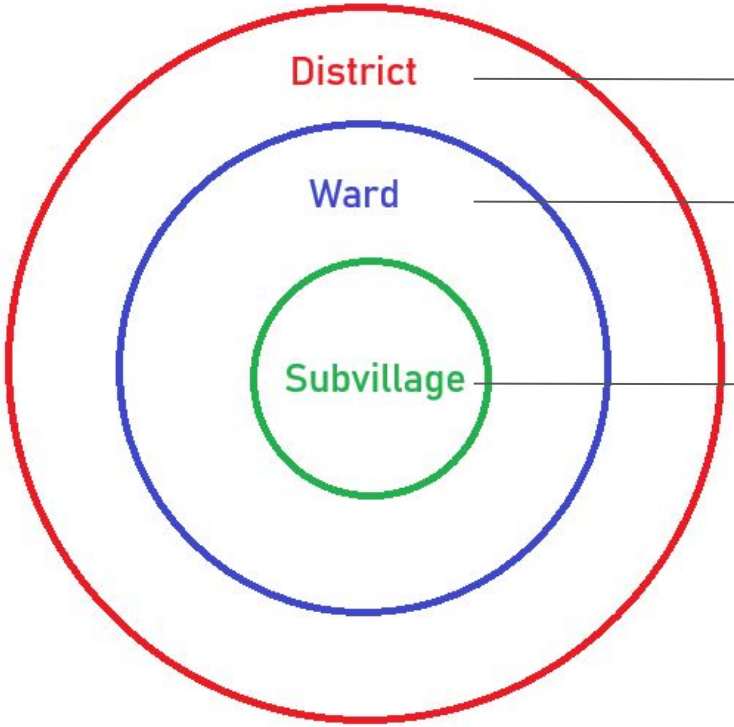
Water Quality vs Functionality



Extraction Type



“Group fill” missing values



	district_code	latitude	longitude
4	1	-1.825359	31.130847
9	1	-1.257051	30.626991
21	1	-1.541205	30.878919

	ward	latitude	longitude
19118	bumera	-1.327684	34.280274
42088	bumera	-1.326142	34.276339
46626	bumera	-1.326913	34.278307

Check for NaNs;
If NaNs present, move to next group (outer circle)

	subvillage	latitude	longitude
304	barazani	-4.389084	33.430917
5782	barazani	-4.405503	33.43444
31403	barazani	-4.397294	33.432679

Columns selected for initial training

Col. No	Column Name Description	Processing
2	date_recorded - The date the row was entered	outliers removed (31 records) 2004 30 2002 1 Year extracted from date, Ordinal encoding is used
4	gps_height - Altitude of the well	Boxplot is used to check for outlier, No outlier MinMax normalization is used
6	longitude - GPS coordinate	Outlier removed using IQR formula and replaced with median MinMax Normalization
7	latitude - GPS coordinate	No outlier but there is a gap between -1 and 0 in th box plot (1819 values close to 0) MinMax Normalization is used
10	basin - Geographic water basin	09 unique values OneHotEncoding is used
12/13	Region/region_code - Geographic location (coded)	21/27 unique values I used one hot encoding

Columns selected for initial training

Col. No	Column Name Description	Processing
14/15	district_code/lga - Geographic location	20/125 unique values I used one hot encoding
17	Population	Outliers replaced with mean
18	public_meeting - True/False	3334 NaN values replaced with string Unknown Now we have three categories,Used Onehot encoding
22	permit - If the waterpoint is permitted	3055 NaN values replaced with string Unknown Now we have three categories,Used Onehot encoding
23	construction_year - Year the waterpoint was constructed	33% values are zero, mean its unknown replaced with median of non zero values
24	extraction_type - The kind of extraction the waterpoint uses	extraction type,group and class has 18,13 and 7 categories respectively, extraction type is selected,1 hot encoding is used

Columns selected for initial training

Col. No	Column Name Description	Processing
27	management - How the waterpoint is managed	management has 12 categories, group has 5 categories management column is selected, 1 hot encoding is used
29	payment - What the water costs	Both payment and payment type have the same 7 unique values Payment column selected and,1 hot encoding is used
31	water_quality - The quality of the water	water quality has 8 unique values, quality_group has 6 unique values water quality is selected,1 hot encoding is used
33	quantity - The quantity of water	quantity and quantity_group columns are same with 5 categories I chose qunatity, 1 hot encoding is sued
35	source - The source of the water	source has 10 unique values,source_type has 7 unique values source is selected,1 hot encoding is used
37	source_class - The source of the water	3 unique categories 1 hot encoding is used
38	waterpoint_type - The kind of waterpoint	water point type and group has 7 and 6 unique values, water point type is selected,1 hot encoding is used

Columns dropped for initial training

Col. No	Column Name Description	Reason
1	amount_tsh - Total static head (amount water available to waterpoint)	50 % values are zero
3	funder - Who funded the well	1895 Unique values.3535 NaN values. Highest contribution of pumps (around 8000) from Govt of Tanzania
5	installer - Organization that installed the well	2143 Unique values, 3653 NaN values. Higest contribution from DWE (around 17000)
8	wpt_name-water point nam	37381 unique values,2 Nan values
9	num_private -	70 % values are zero
11	subvillage - Geographic location	19281 unique values, 371 Nan values
16	ward- Geographic location	2092 unique values, No NaN values

Columns dropped for initial training

Col. No	Column Name Description	Reason
19	recorded_by - Group entering this row of data	All values are same
20	scheme_management - Who operates the waterpoin	2695 unique values,3874 NaN values
21	scheme_name - Who operates the waterpoint	50% values are NaN (count=28790)
25/26	extraction_type_group / extraction_type_group -The kind of extraction the waterpoint uses	13/7 categories..already selected extraction_type with 18 categories
28	management_group - How the waterpoint is managed	5 unique values in management group. Selected management column with 12 categories
30	payment_type - What the water costs	07 unique values..almost same as selected payment column
32,34	Quality group/quantity group	water quality/quantity columns are selected
36,39	Source type, water point type group	Source/waterpoint_type columns are selected

Conclusion

Plenty of missing data that needs to be handled

Imbalance in data

Some features tend to have interesting relation to target variable