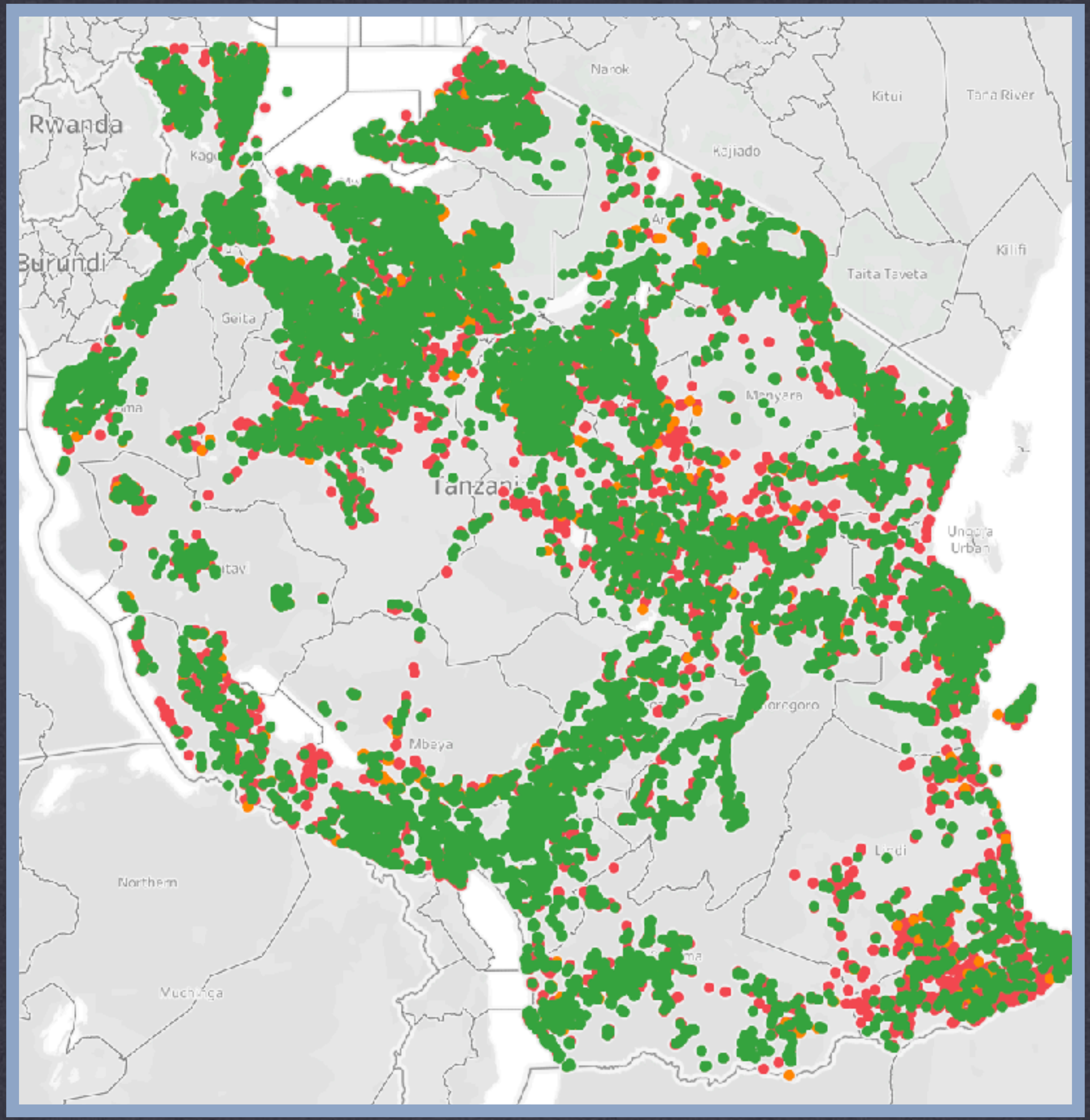
PREDICTING THE FUNCTIONALITY OF WATER PUMPS IN TANZANIA

AUTHOR: **ALBERTO HERNANDEZ**

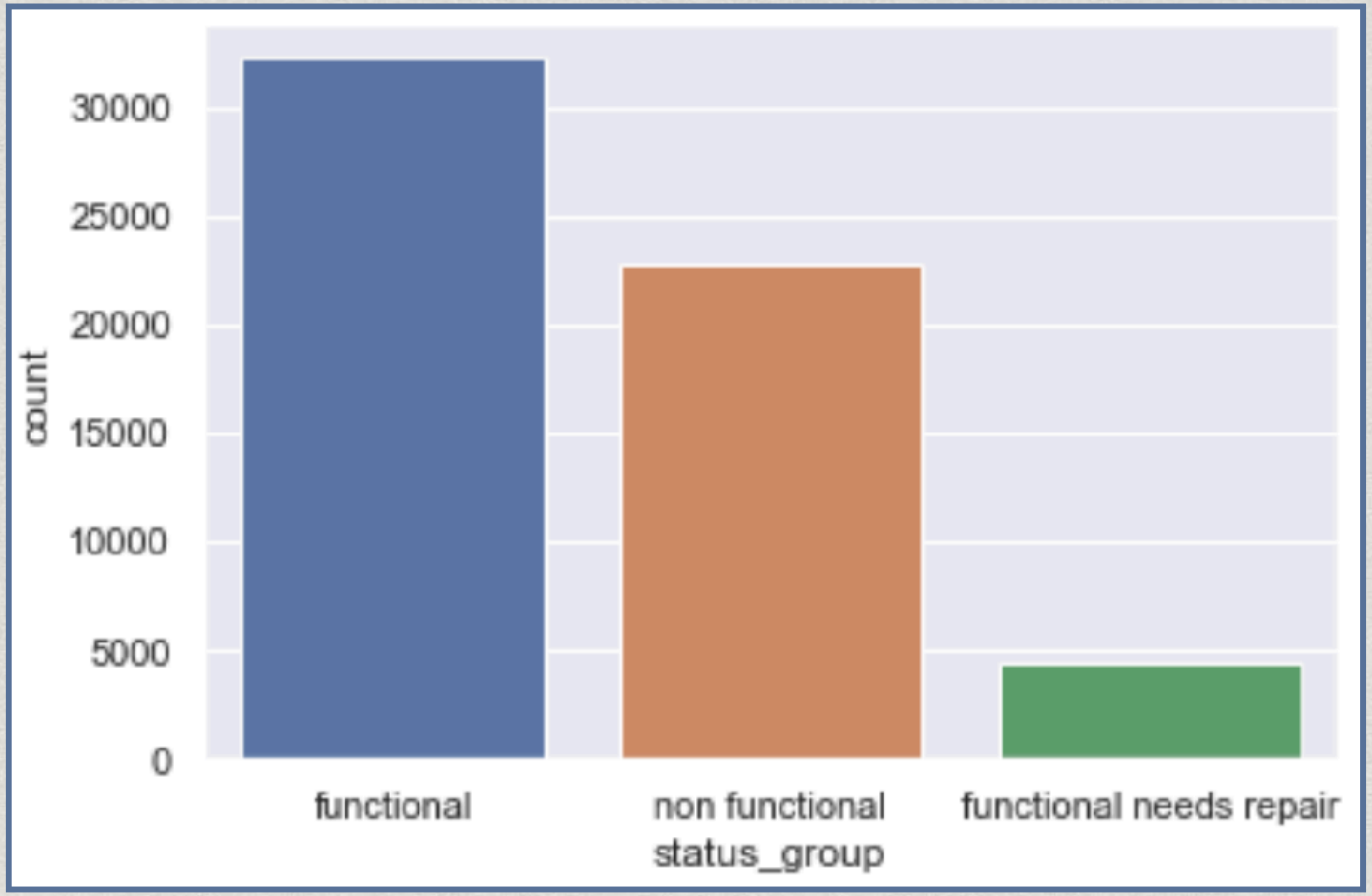
COURSE: **MSDS 692 - PRACTICUM I**



Problem Description

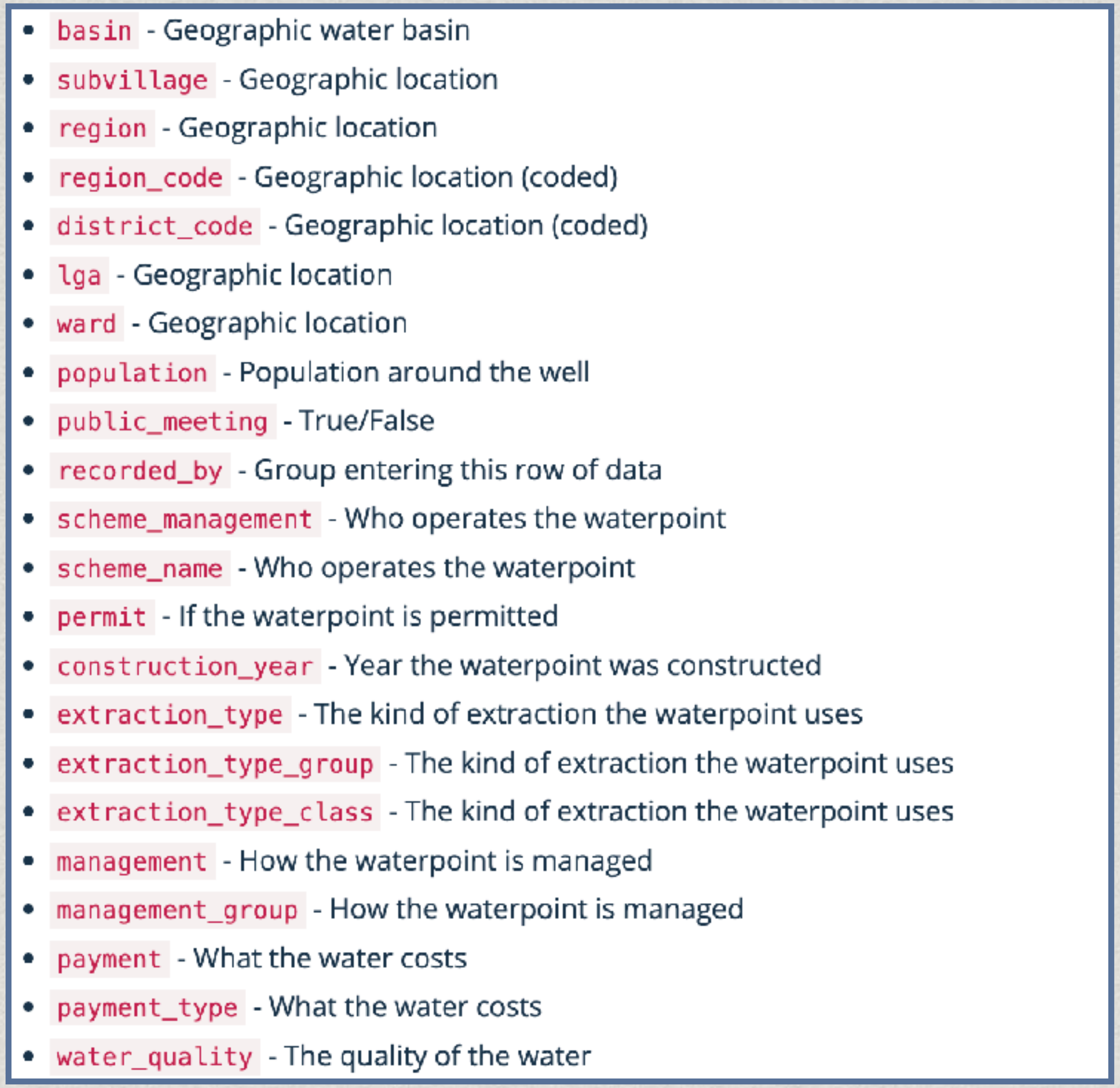
Motivation: A functional water pump ensures that a community has access to a reliable source of potable water

Goal: Identify the causes of water pump failure, and predict which water-points will fail so that maintenance operations can be improved.



Data Description

- Two datasets were obtained from *DrivenData*
- The first dataset contains 59,400 entries and 40 attributes most of which are categorical
- The second dataset matches each pump with an ID, and classifies each pump as either functional, functional but needs repair, and non functional



- `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

Project Overview

This data science task involves classification using supervised machine learning, and data visualization.

1. Data Cleaning

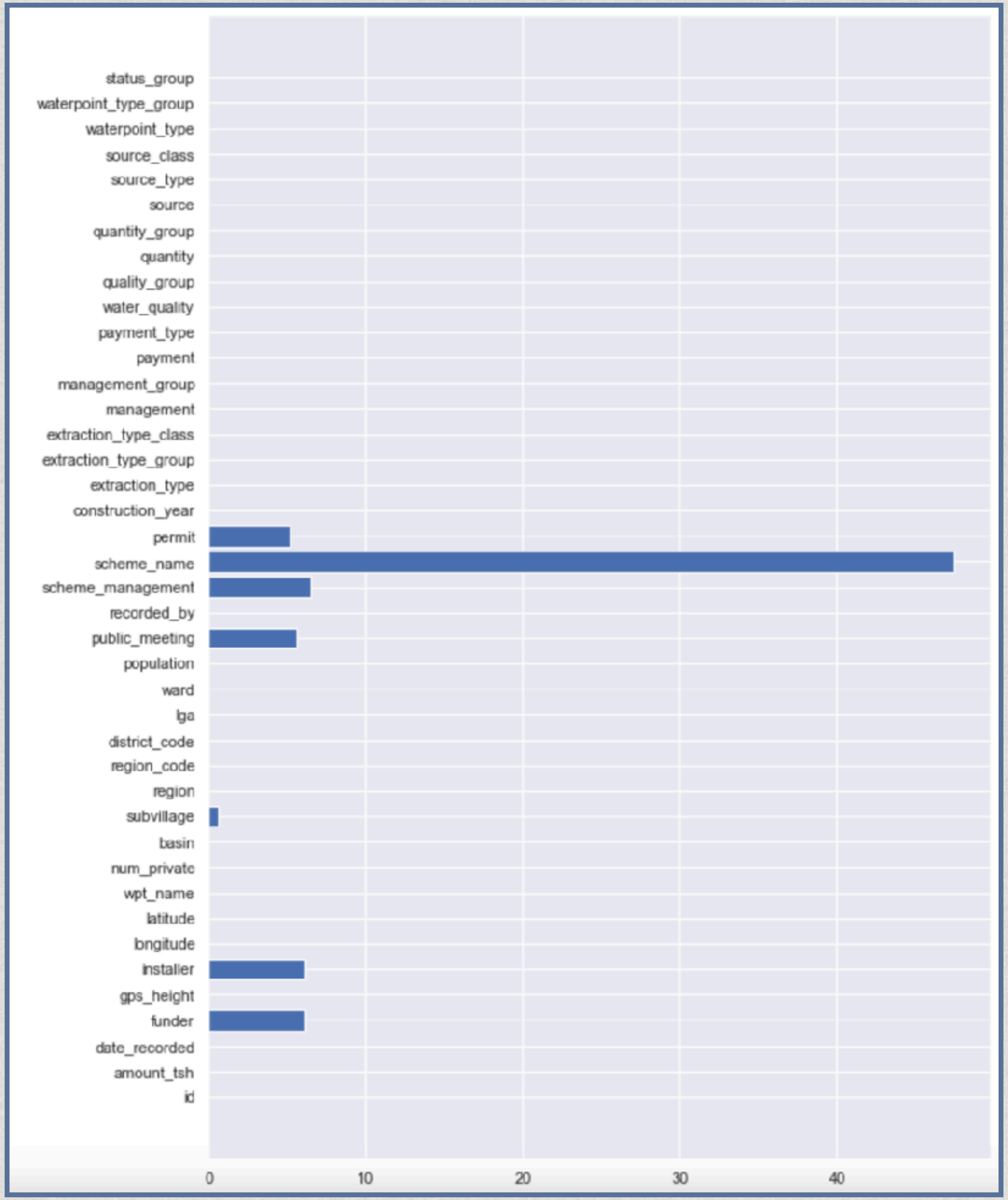
2. EDA

3. Build Models

4. Analysis

Data Cleaning

- Seven columns with missing values
- Redundant features were reduced to one
- High cardinality addressed by CatBoost encoder
- KNNImputer was used to fill missing values



Exploratory Data Analysis

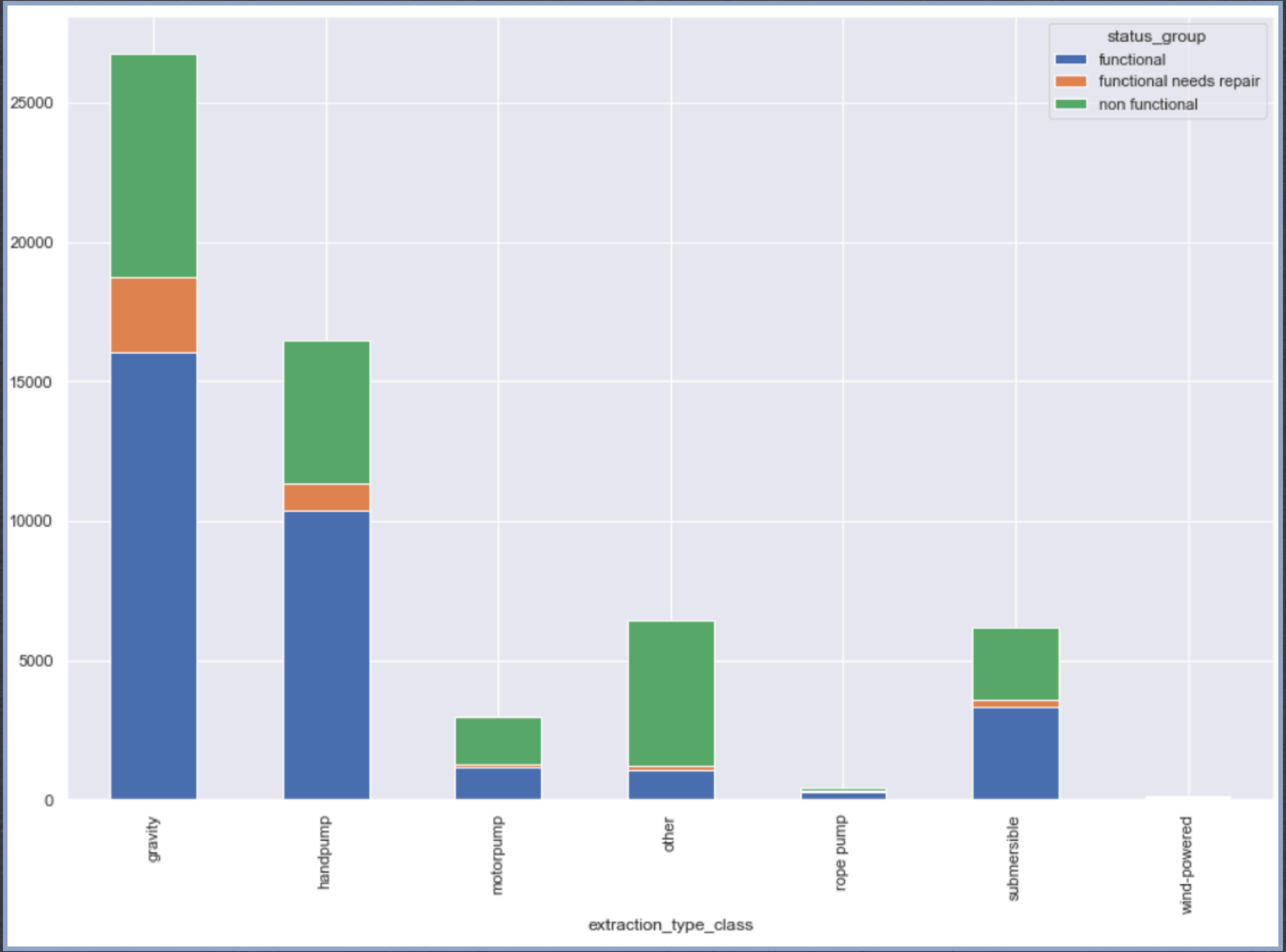
Notable Relationships between Features and Water Pump Functionality:

1. Extraction Type
2. Basin
3. Year of Construction
4. Installer

EXTRACTION TYPE

DEFINITION:

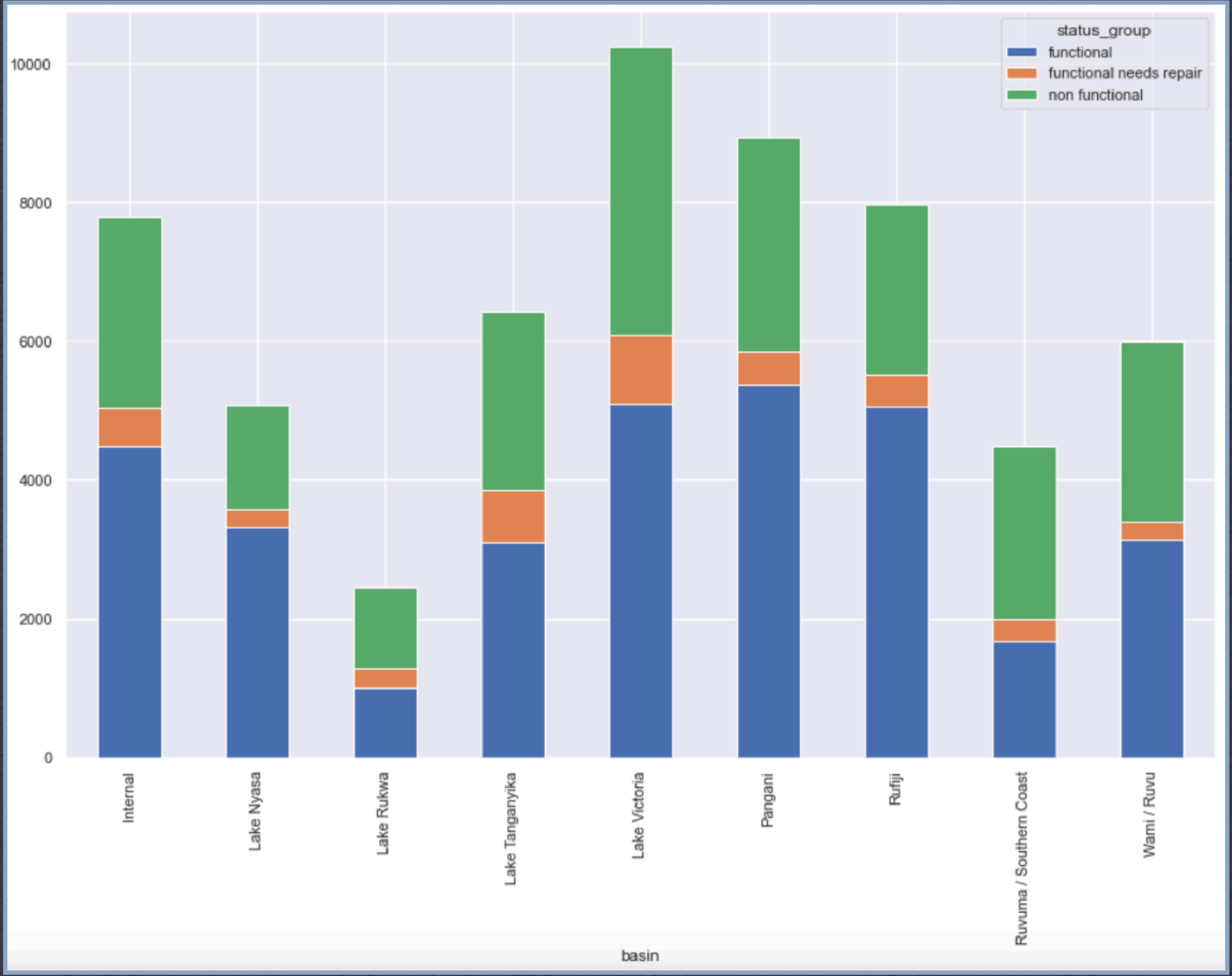
Refers to the method the water point uses to extract water



BASIN

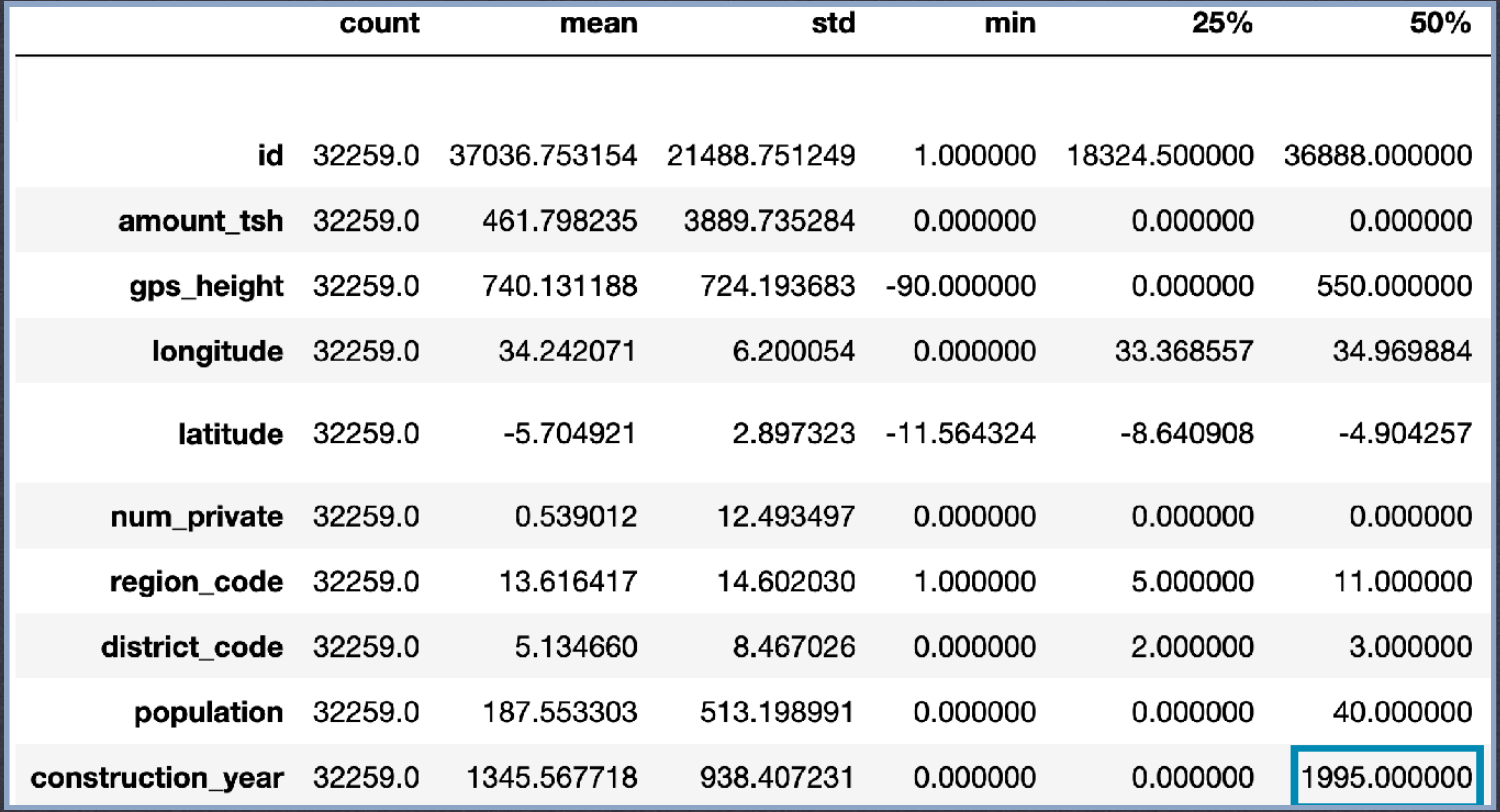
DEFINITION:

A basin is a depression, or dip, in the Earth's surface, which in this case, if filled with water



YEAR OF CONSTRUCTION

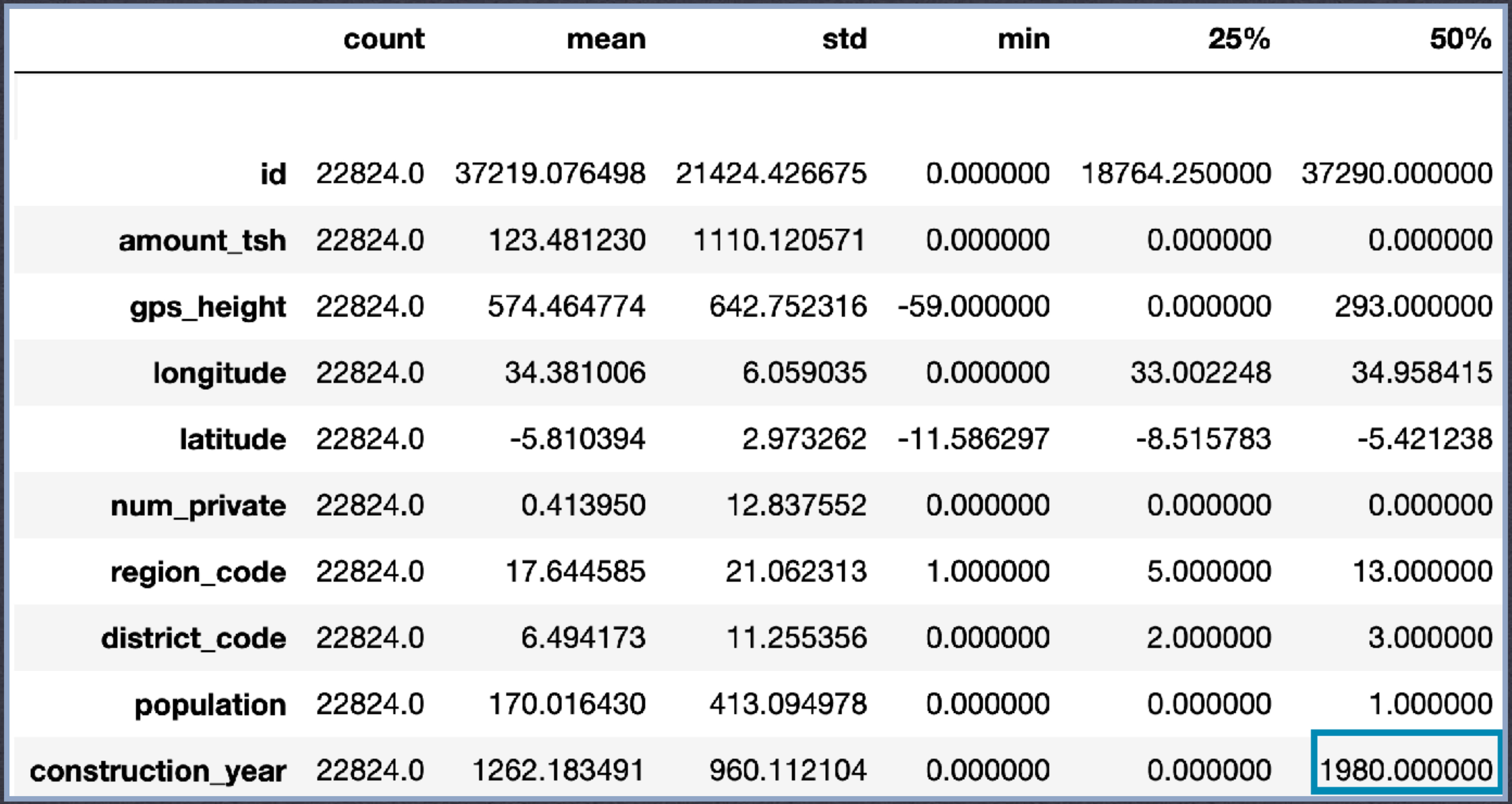
Functional Water Pumps



	count	mean	std	min	25%	50%
id	32259.0	37036.753154	21488.751249	1.000000	18324.500000	36888.000000
amount_tsh	32259.0	461.798235	3889.735284	0.000000	0.000000	0.000000
gps_height	32259.0	740.131188	724.193683	-90.000000	0.000000	550.000000
longitude	32259.0	34.242071	6.200054	0.000000	33.368557	34.969884
latitude	32259.0	-5.704921	2.897323	-11.564324	-8.640908	-4.904257
num_private	32259.0	0.539012	12.493497	0.000000	0.000000	0.000000
region_code	32259.0	13.616417	14.602030	1.000000	5.000000	11.000000
district_code	32259.0	5.134660	8.467026	0.000000	2.000000	3.000000
population	32259.0	187.553303	513.198991	0.000000	0.000000	40.000000
construction_year	32259.0	1345.567718	938.407231	0.000000	0.000000	1995.000000

YEAR OF CONSTRUCTION

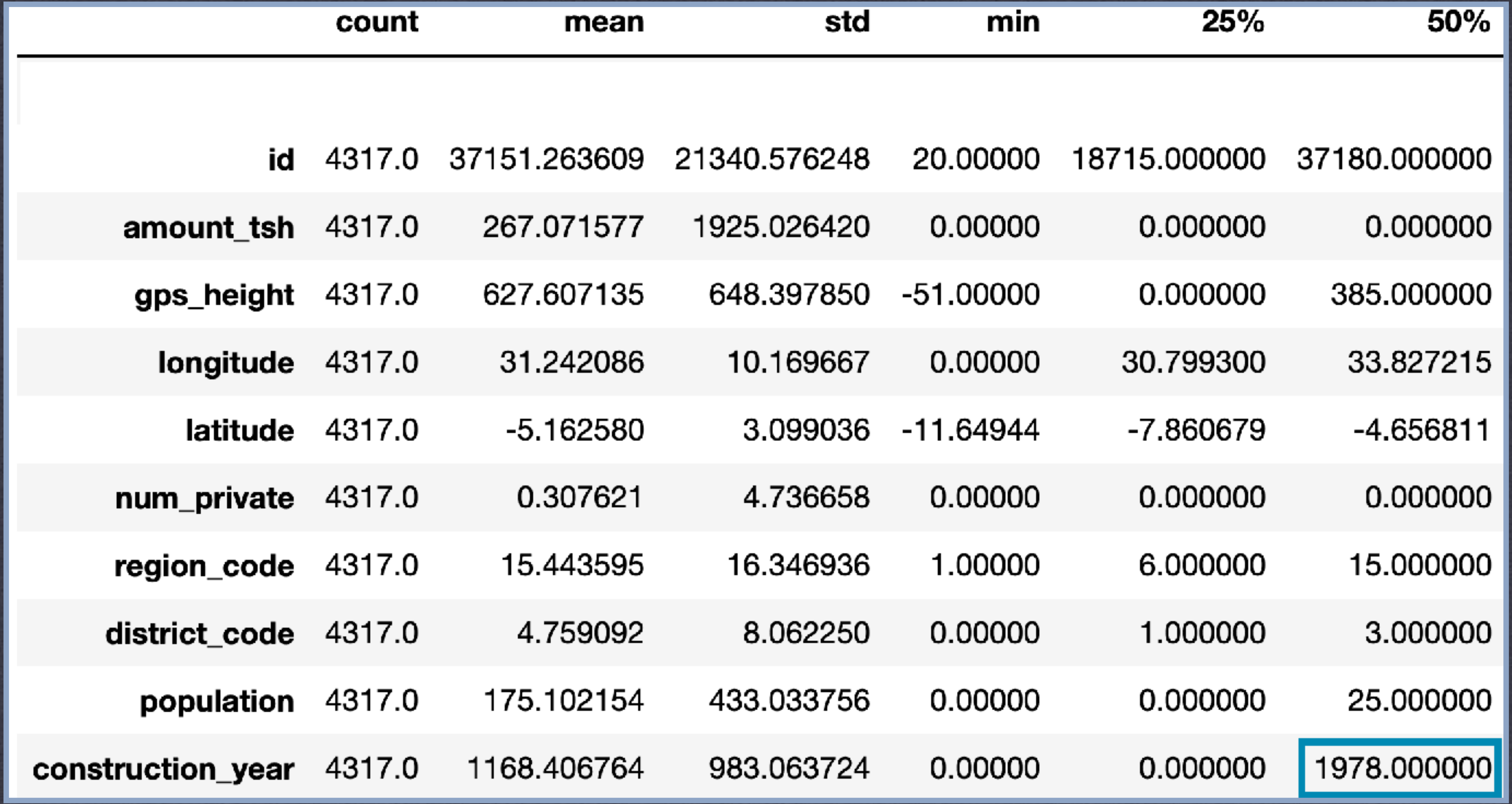
Non-functional Water Pumps



	count	mean	std	min	25%	50%
id	22824.0	37219.076498	21424.426675	0.000000	18764.250000	37290.000000
amount_tsh	22824.0	123.481230	1110.120571	0.000000	0.000000	0.000000
gps_height	22824.0	574.464774	642.752316	-59.000000	0.000000	293.000000
longitude	22824.0	34.381006	6.059035	0.000000	33.002248	34.958415
latitude	22824.0	-5.810394	2.973262	-11.586297	-8.515783	-5.421238
num_private	22824.0	0.413950	12.837552	0.000000	0.000000	0.000000
region_code	22824.0	17.644585	21.062313	1.000000	5.000000	13.000000
district_code	22824.0	6.494173	11.255356	0.000000	2.000000	3.000000
population	22824.0	170.016430	413.094978	0.000000	0.000000	1.000000
construction_year	22824.0	1262.183491	960.112104	0.000000	0.000000	1980.000000

YEAR OF CONSTRUCTION

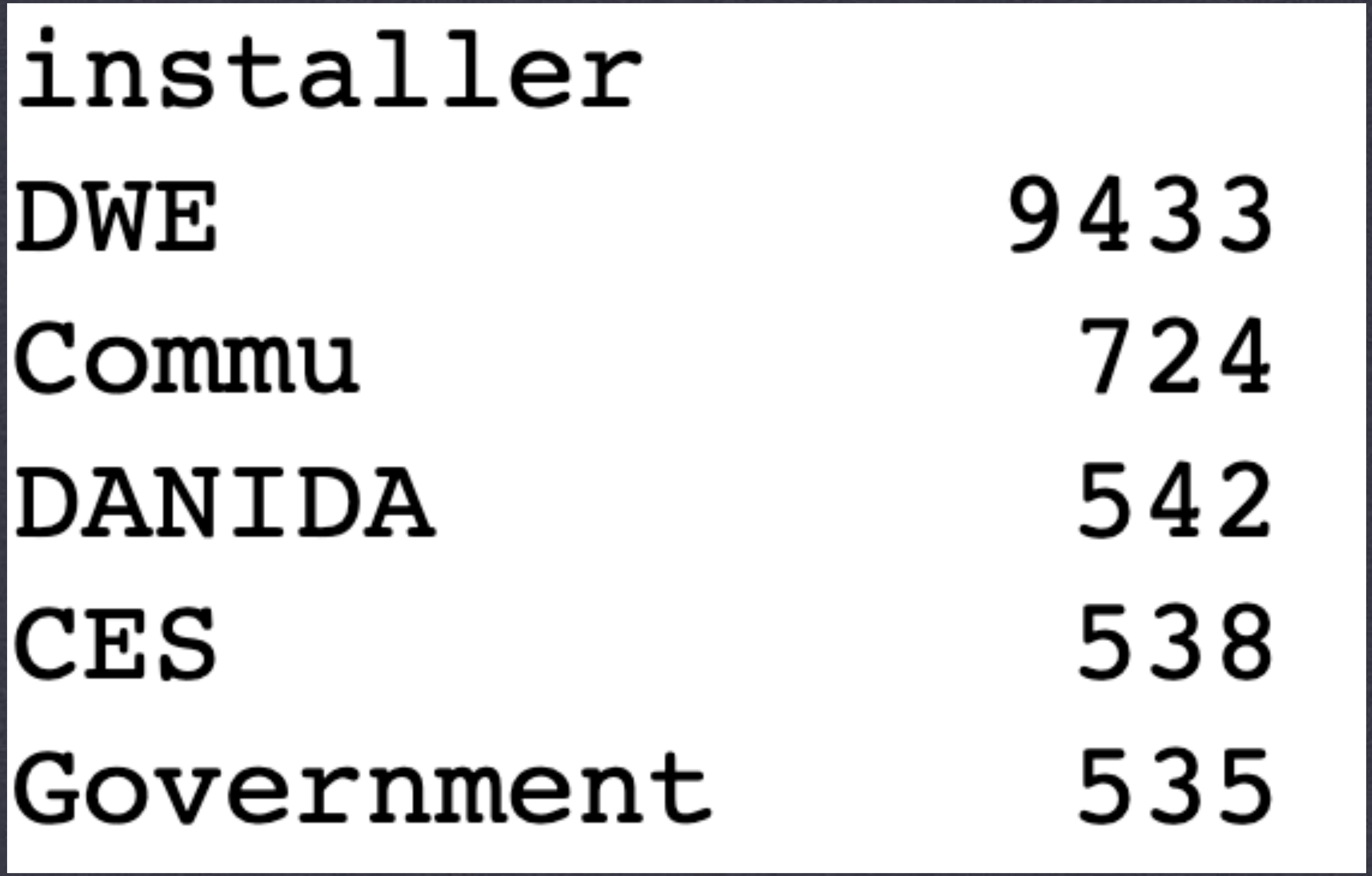
Functional but Needs Repair Water Pumps



	count	mean	std	min	25%	50%
id	4317.0	37151.263609	21340.576248	20.00000	18715.000000	37180.000000
amount_tsh	4317.0	267.071577	1925.026420	0.00000	0.000000	0.000000
gps_height	4317.0	627.607135	648.397850	-51.00000	0.000000	385.000000
longitude	4317.0	31.242086	10.169667	0.00000	30.799300	33.827215
latitude	4317.0	-5.162580	3.099036	-11.64944	-7.860679	-4.656811
num_private	4317.0	0.307621	4.736658	0.00000	0.000000	0.000000
region_code	4317.0	15.443595	16.346936	1.00000	6.000000	15.000000
district_code	4317.0	4.759092	8.062250	0.00000	1.000000	3.000000
population	4317.0	175.102154	433.033756	0.00000	0.000000	25.000000
construction_year	4317.0	1168.406764	983.063724	0.00000	0.000000	1978.000000

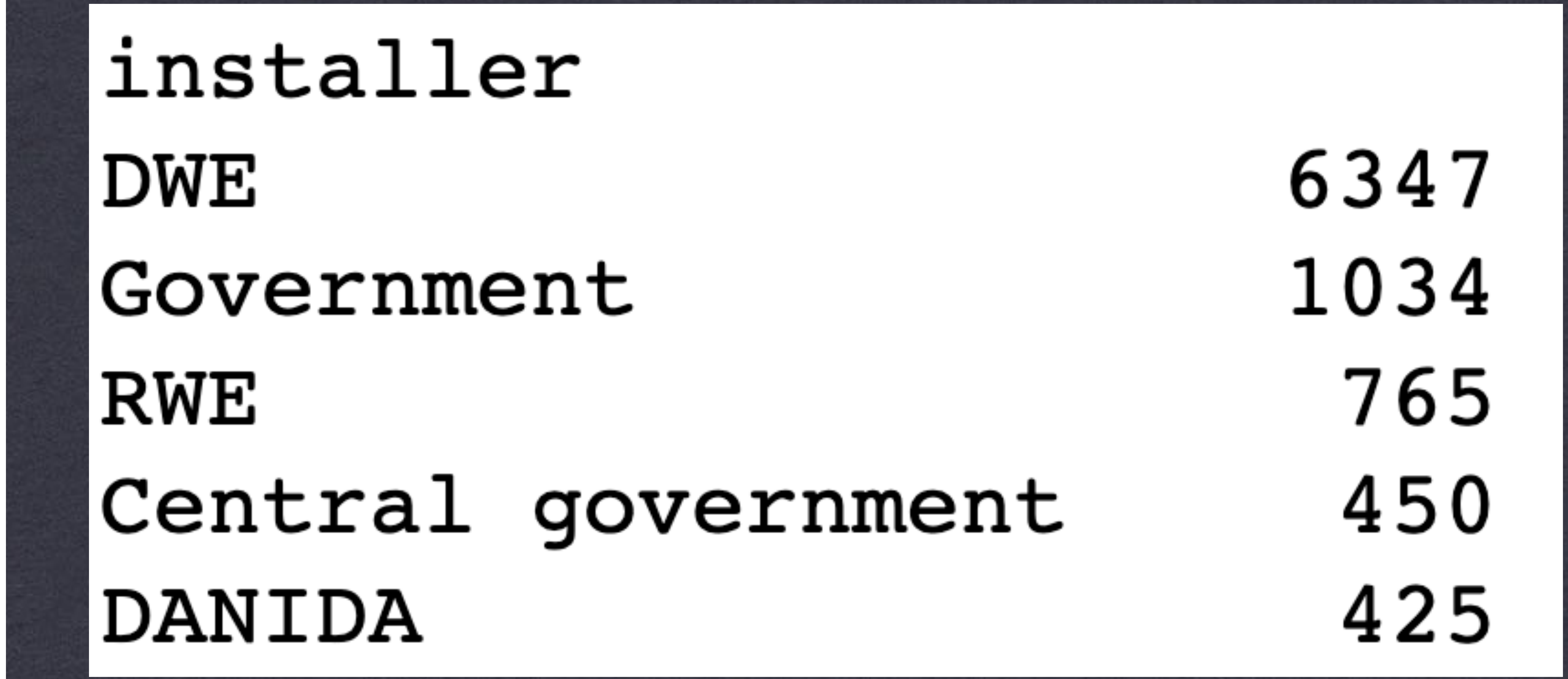
INSTALLER

Functional Water Pumps



```
installer
DWE                9433
Commu               724
DANIDA              542
CES                 538
Government          535
```

Non-Functional Water Pumps

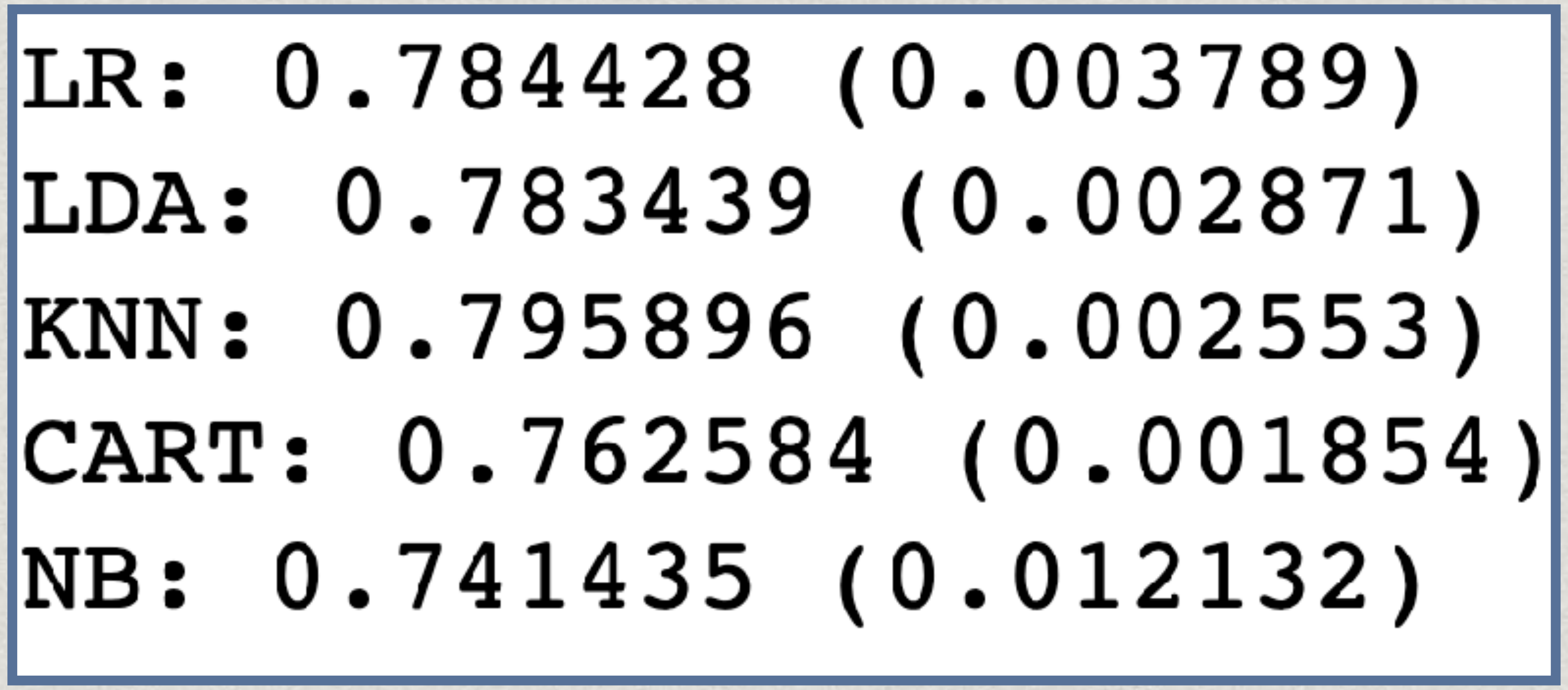


```
installer
DWE                          6347
Government                   1034
RWE                           765
Central government            450
DANIDA                        425
```

Classification Using Supervised Learning

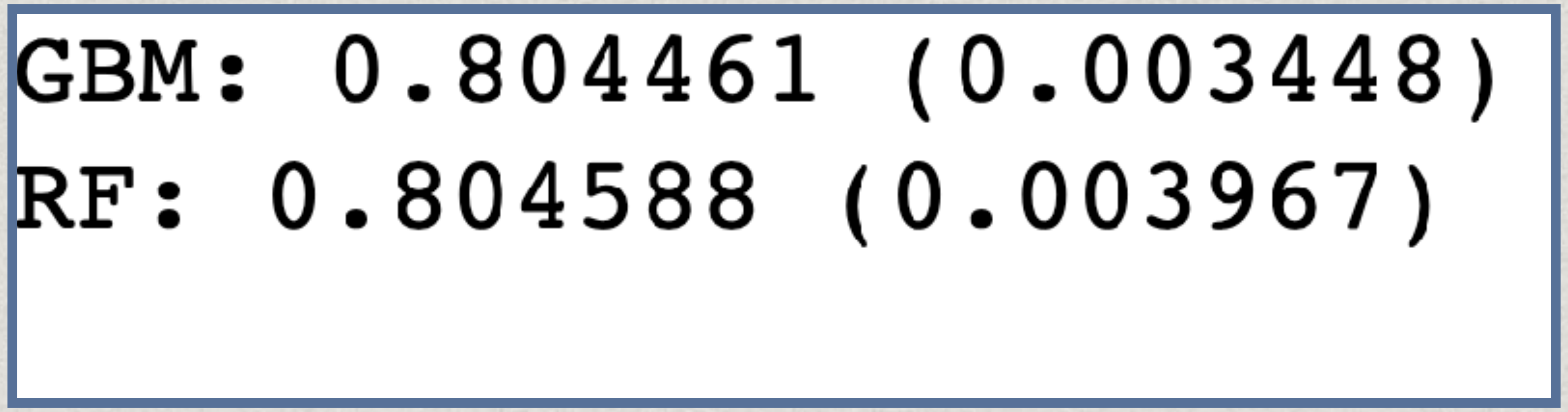
- Multi-class classification was accomplished by implementing linear, non-linear, and ensemble methods.

Linear & Non-Linear Models



```
LR: 0.784428 (0.003789)
LDA: 0.783439 (0.002871)
KNN: 0.795896 (0.002553)
CART: 0.762584 (0.001854)
NB: 0.741435 (0.012132)
```

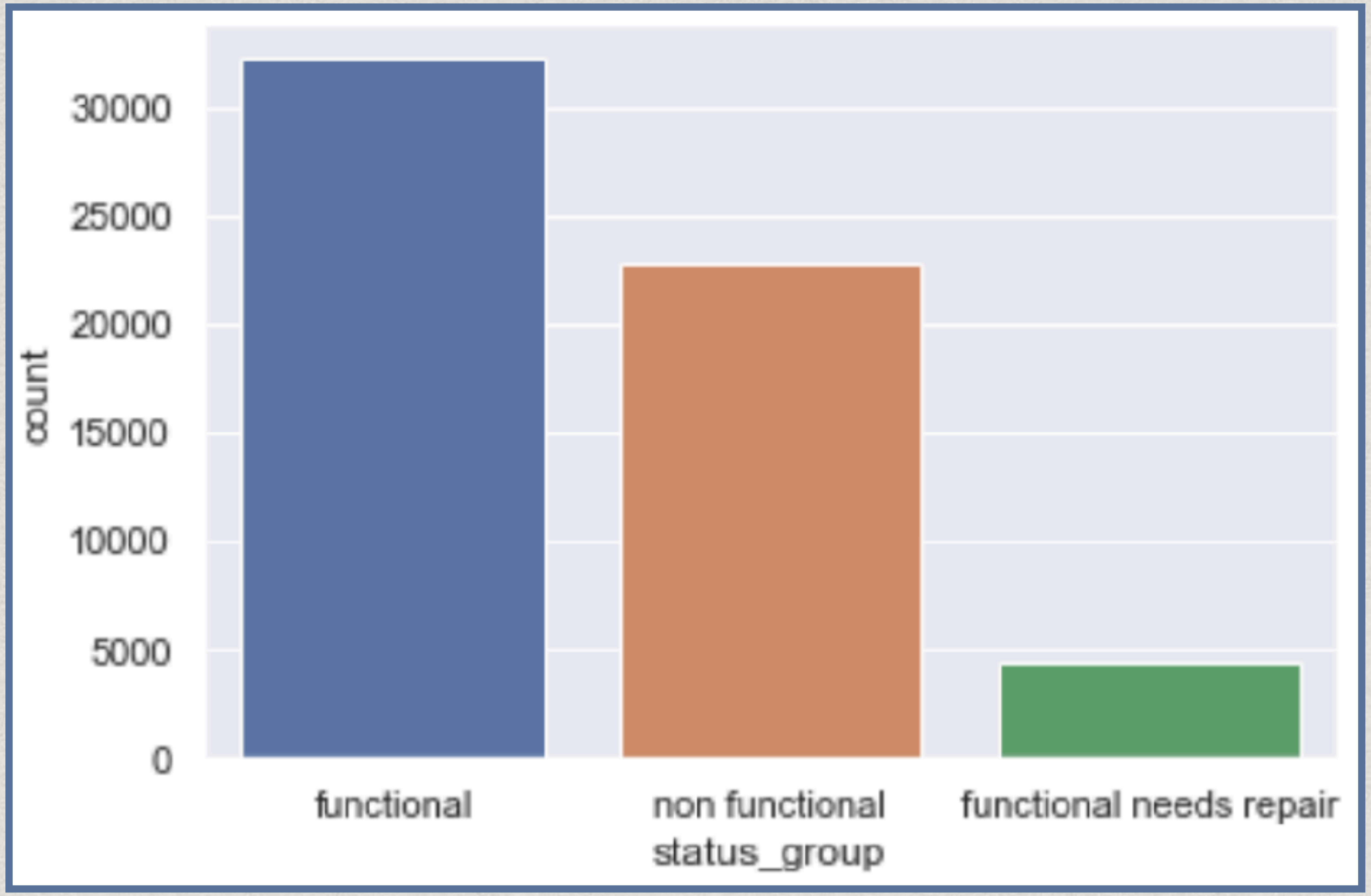
Ensemble Methods



```
GBM: 0.804461 (0.003448)
RF: 0.804588 (0.003967)
```

Balancing the Dataset

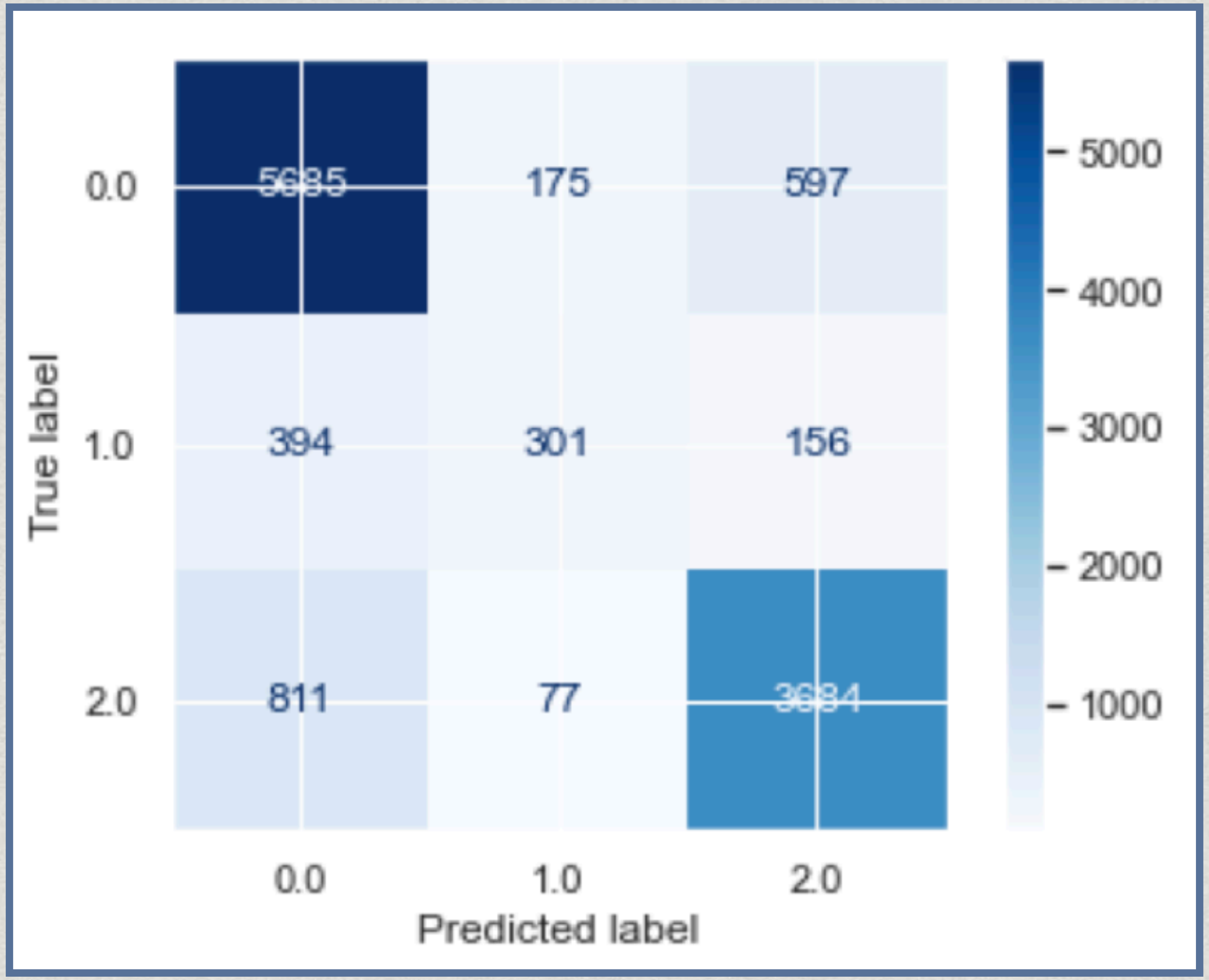
- Approach: a balanced dataset was created by taking a sample from each class equal to the size of the underrepresented class



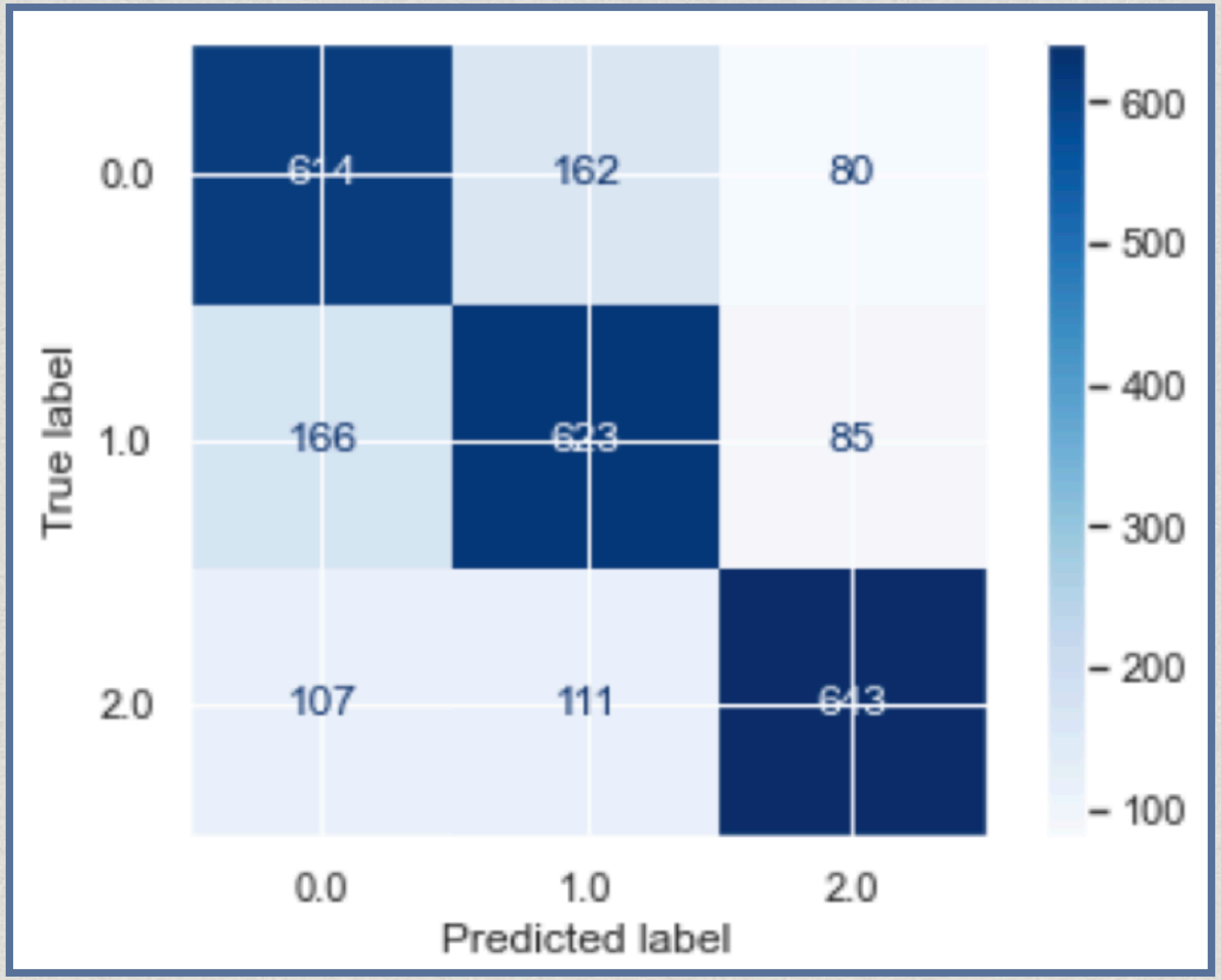
True Positive Rate

TPR = TP / TP + FN

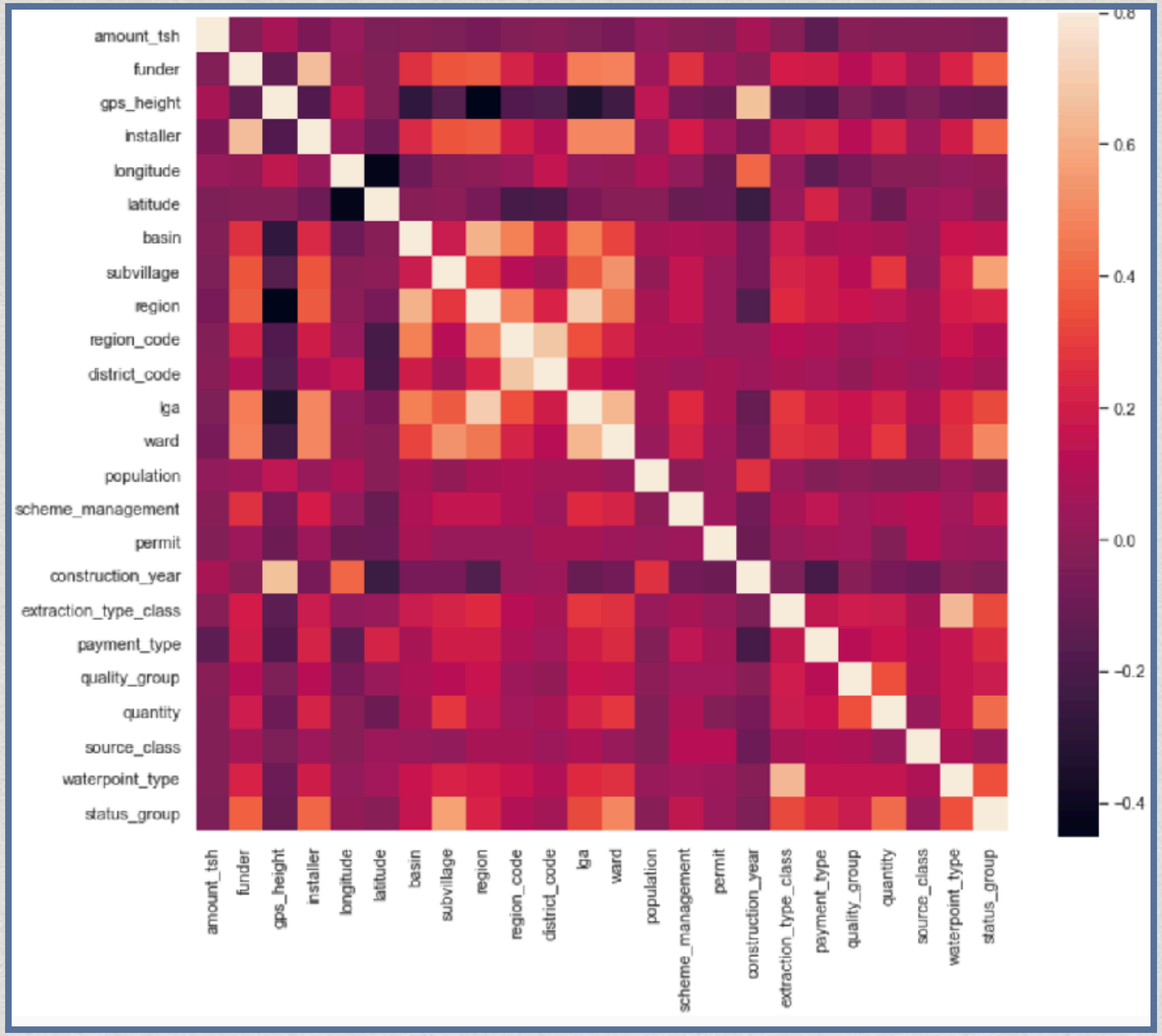
Unbalanced

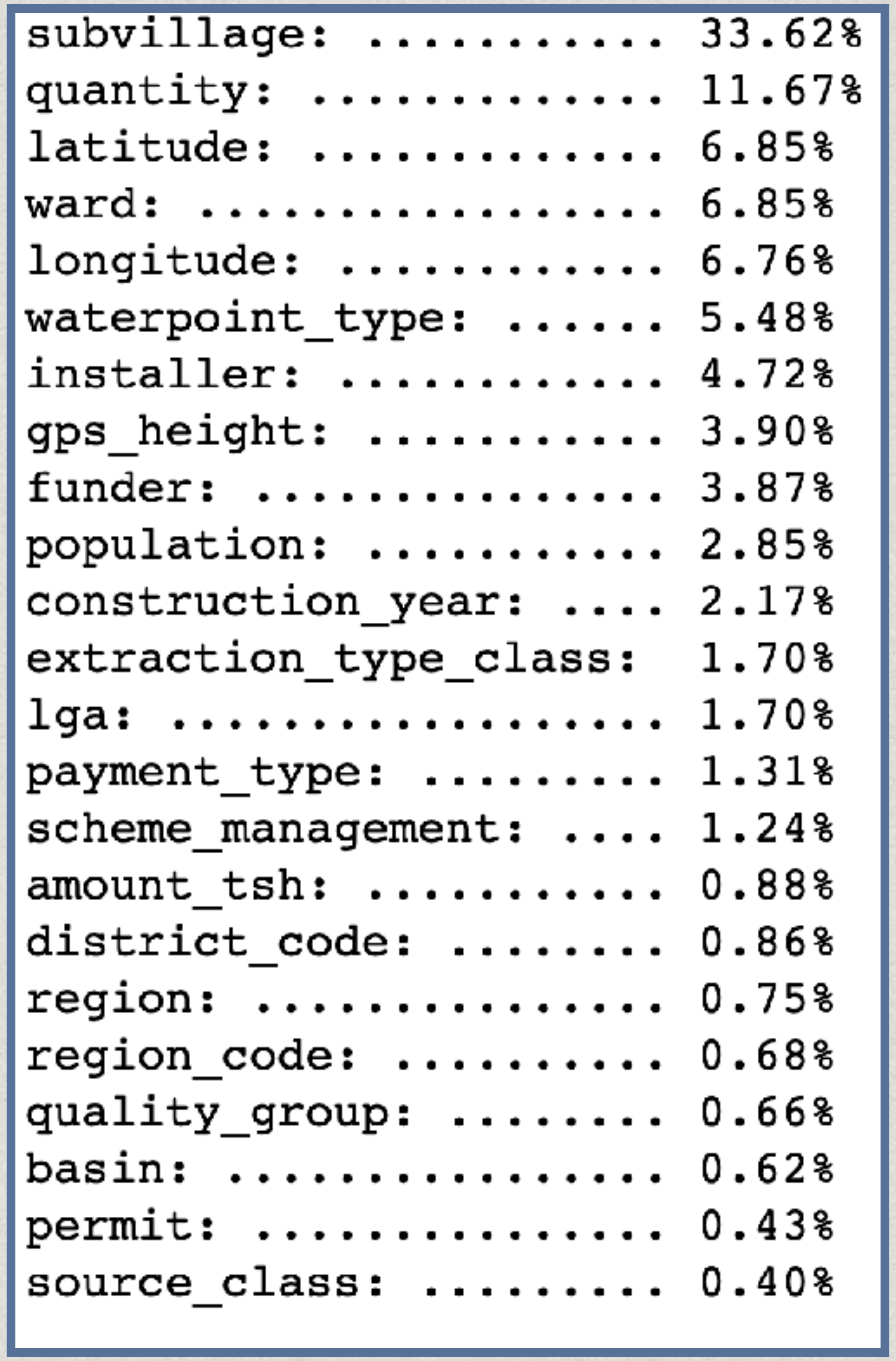


Balanced



Analysis





```
subvillage: ........... 33.62%
quantity: ............. 11.67%
latitude: ............. 6.85%
ward: ................. 6.85%
longitude: ............ 6.76%
waterpoint_type: ...... 5.48%
installer: ............ 4.72%
gps_height: ........... 3.90%
funder: ............... 3.87%
population: ........... 2.85%
construction_year: .... 2.17%
extraction_type_class:  1.70%
lga: .................. 1.70%
payment_type: ......... 1.31%
scheme_management: .... 1.24%
amount_tsh: ........... 0.88%
district_code: ........ 0.86%
region: ............... 0.75%
region_code: .......... 0.68%
quality_group: ........ 0.66%
basin: ................ 0.62%
permit: ............... 0.43%
source_class: ......... 0.40%
```

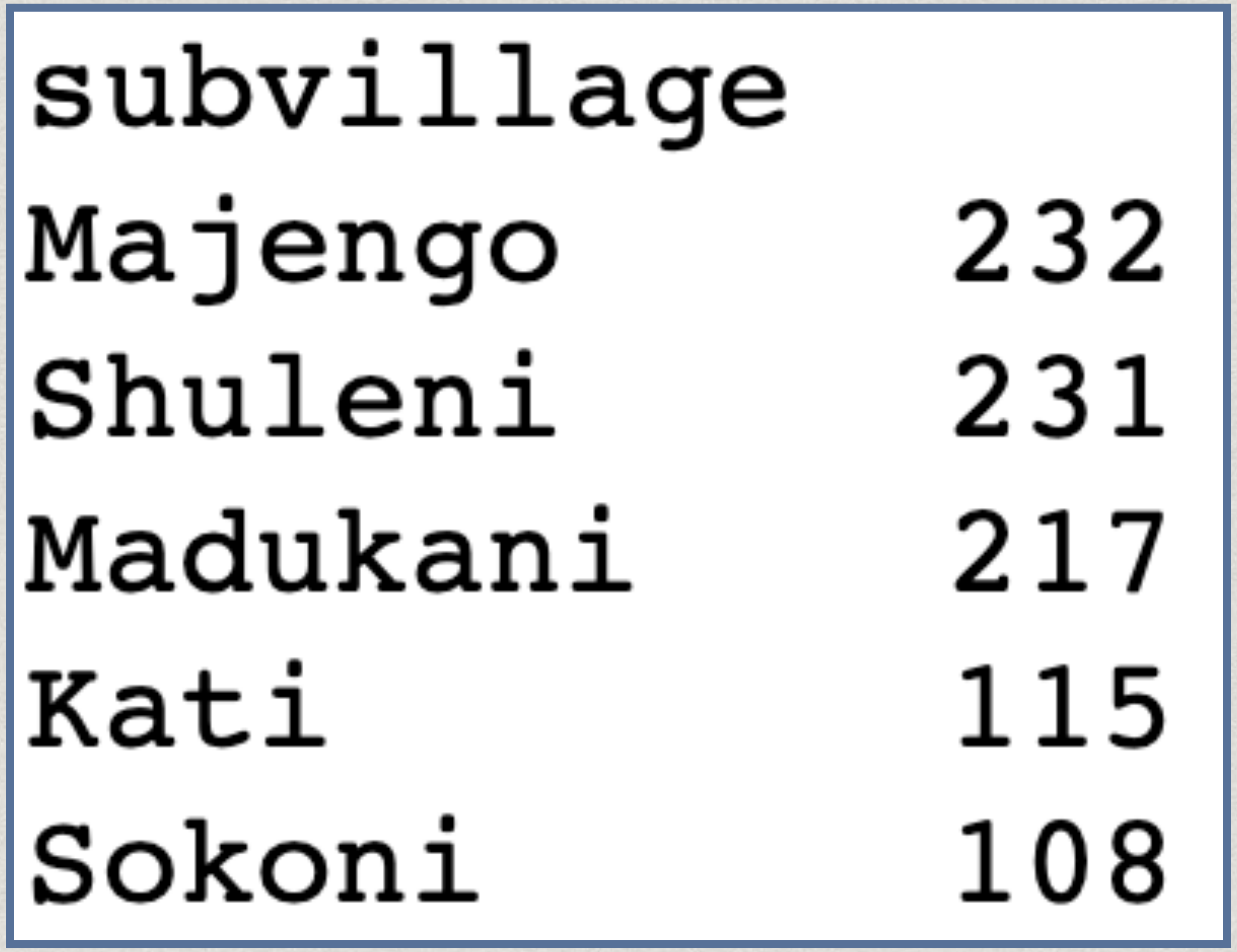
Findings

The results obtained from a Random Forest Regressor, a Correlation Matrix, and EDA demonstrate a correlation between functionality and:

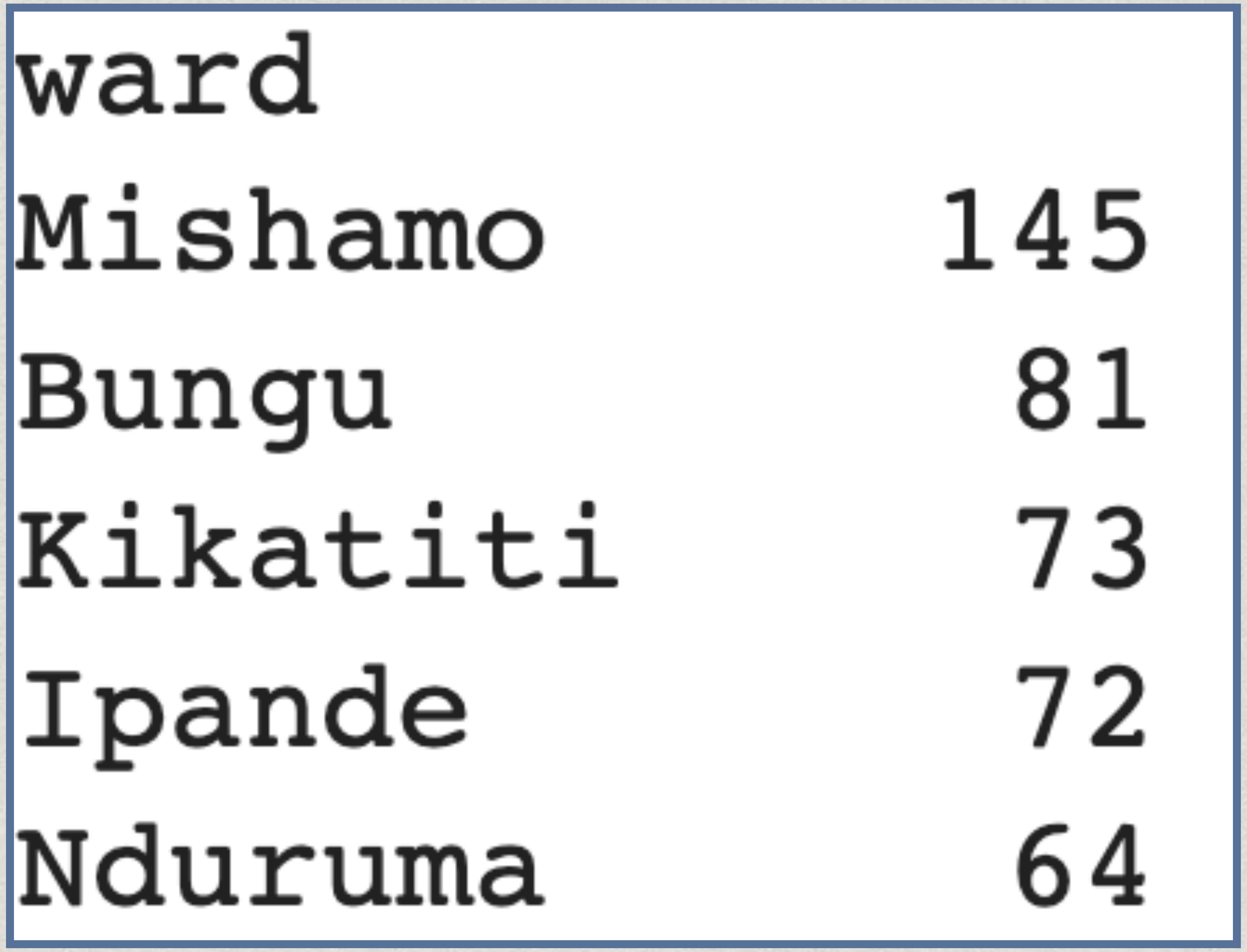
1. Sub-village and Ward
2. Extraction Type
3. Year of Construction
4. Installer
5. Basin

Sub-village and Ward

There are sub villages and wards where the number of non-functional water pumps is high



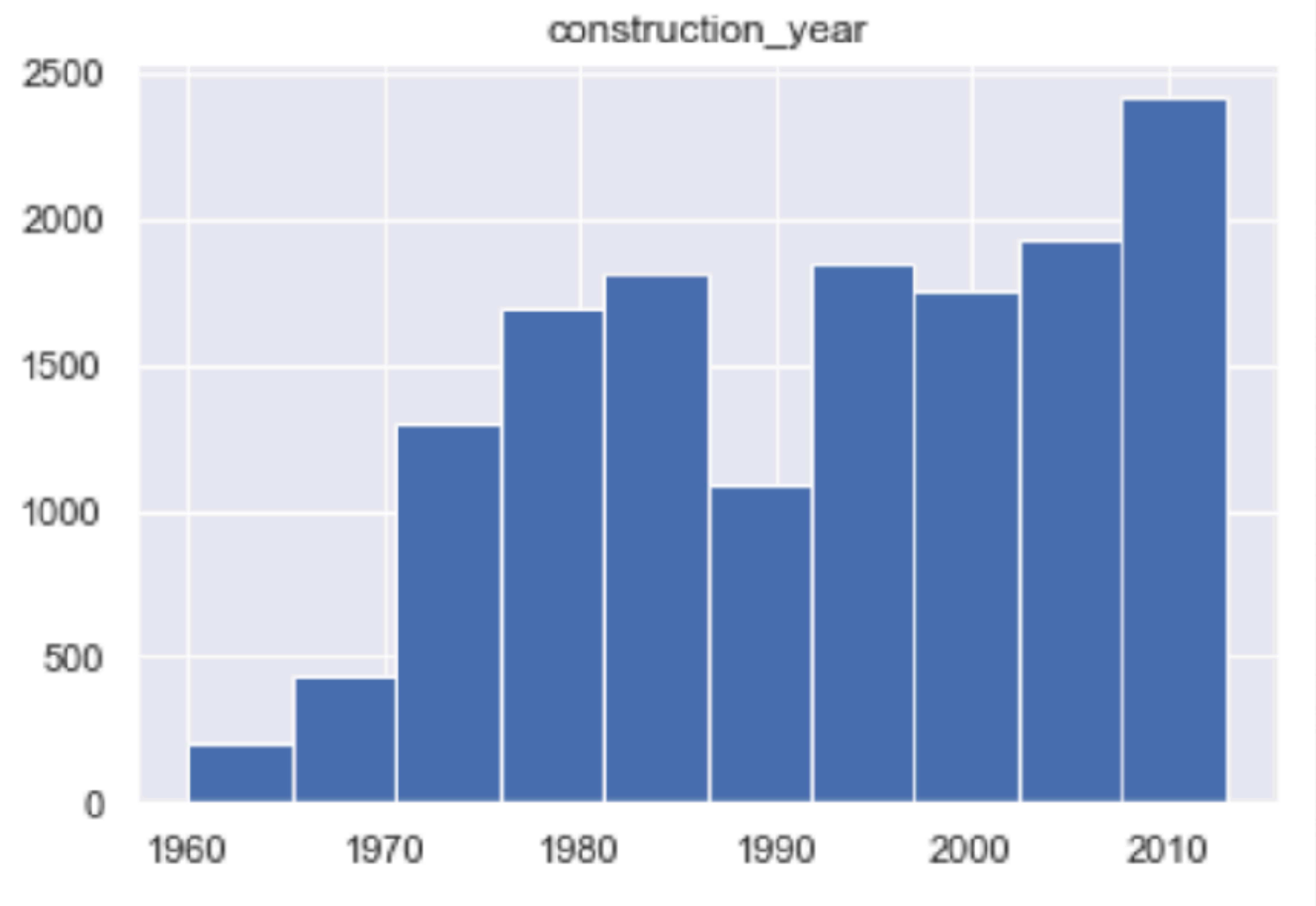
```
subvillage
Majengo       232
Shuleni       231
Madukani      217
Kati          115
Sokoni        108
```



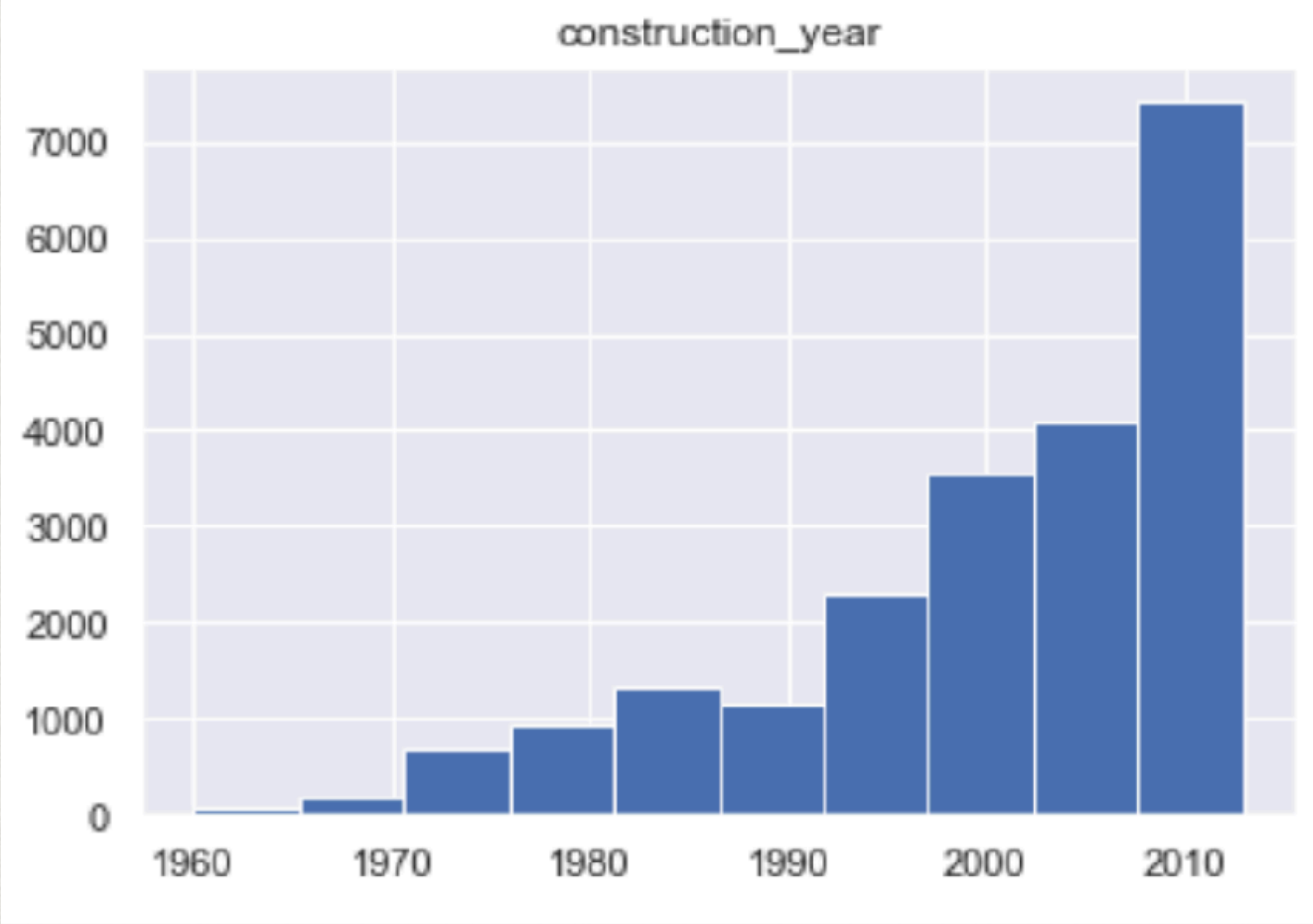
```
ward
Mishamo       145
Bungu          81
Kikatiti       73
Ipande         72
Nduruma        64
```

Year of Construction

Non-Functional



Functional



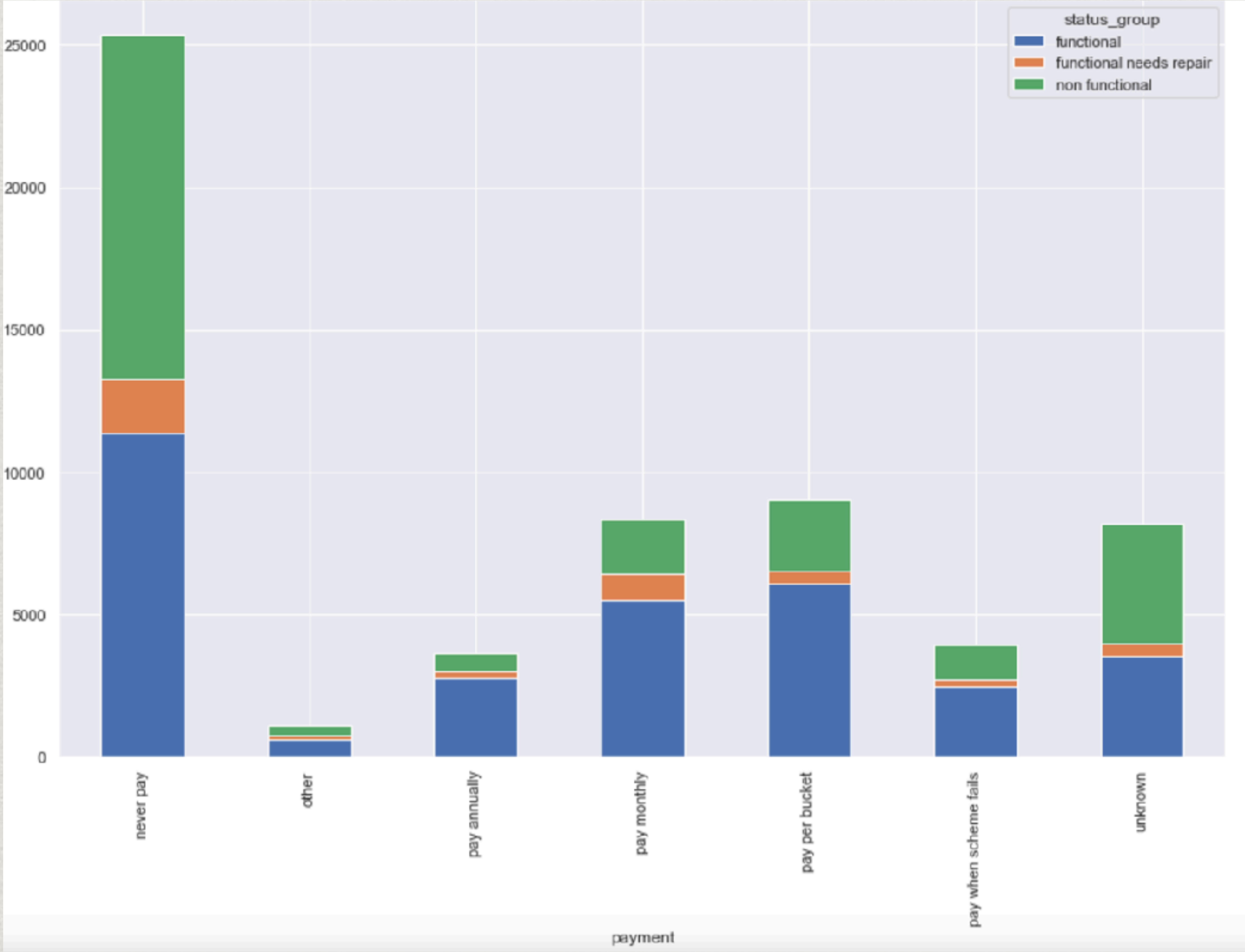
CONCLUSION

- Most water pumps in Tanzania are functional
- There are wards and sub villages with a high concentration of non-functional water pumps
- Factors that determine functionality include:
  - Year of construction
  - Basin
  - Installer
  - Extraction type

Payment Type

DEFINITION:

What the water costs



THE END