Water Pump Presentation

For Tanzania Ministry of Water

Agenda

- Project Proposal
- Initial Hypotheses
- Project Approaches
- Data Insights
- Insights Discussion
- Recommendations and Actions

Project Proposal for Tanzania Ministry of Water

- Pinpoint any location/areas that has water pump breakdowns
- Discover which management types affecting pump maintenances
- Examine water extraction techniques, water sources and water points for relationships
- Create a machine learning model for prediction

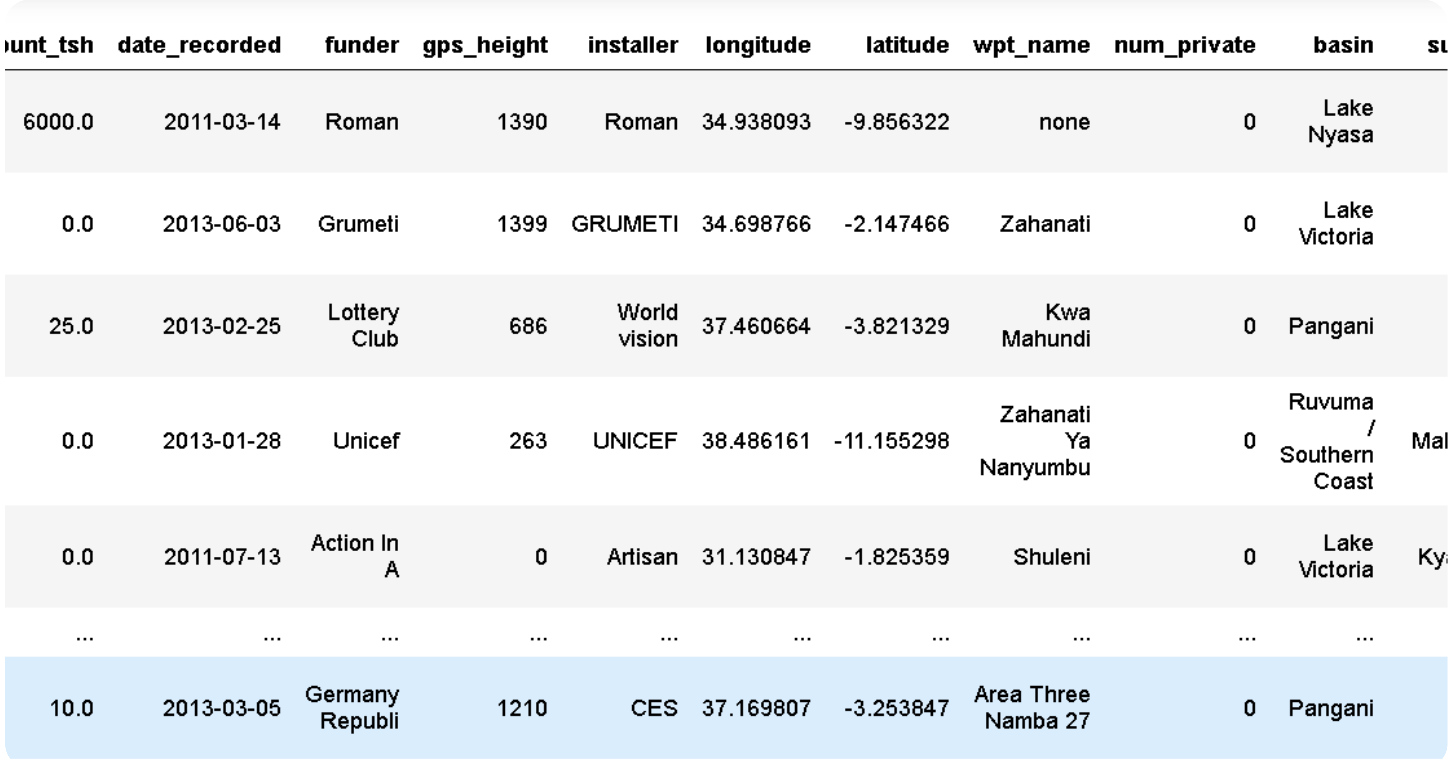
Initial Hypotheses

- Geographic locations such as Basin, Subvillage and Region will provide pump statuses
- Populations density throughout Tanzania are equally distributed
- Water source, types, quantity and quality are consistent in all water pumps

Project Approaches

- Location analysis will use basin, sub village, region, region code and district code features. Extra information can be gleaned from gps height, longitude and latitude.
- Population will use population and public meeting columns
- Source, source type, source class, waterpoint type, waterpoint type group columns will be explored to see any connection to pumps
- Metrics will be used are accuracy, precision, recall and F1 scores since this is binary outcome

Dataset Overview



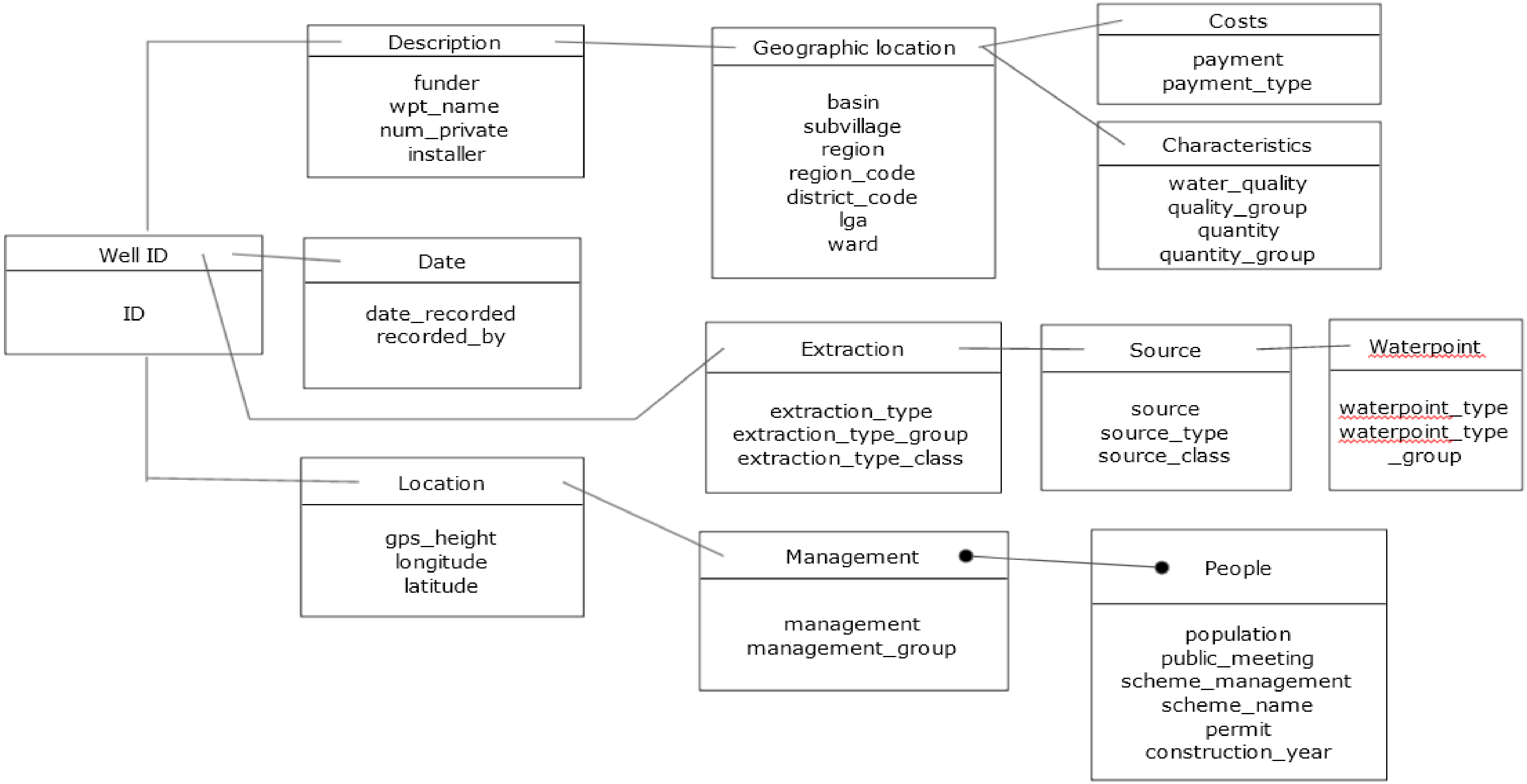
unt_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	basin	su
6000.0	2011-03-14	Roman	1390	Roman	34.938093	-9.856322	none	0	Lake Nyasa	
0.0	2013-06-03	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	Lake Victoria	
25.0	2013-02-25	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	Pangani	
0.0	2013-01-28	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	Ruvuma / Southern Coast	Mal
0.0	2011-07-13	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	Lake Victoria	Ky
...	...	...	...	...	...	...	...	...	...	
10.0	2013-03-05	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	0	Pangani	

Dataset Size:

59400 rows

41 columns

Entity Relationship Diagram



Description
funder
wpt_name
num_private
installer

Geographic location
basin
subvillage
region
region_code
district_code
lga
ward

Costs
payment
payment_type

Characteristics
water_quality
quality_group
quantity
quantity_group

Well ID
ID

Date
date_recorded
recorded_by

Extraction
extraction_type
extraction_type_group
extraction_type_class

Source
source
source_type
source_class

Waterpoint
waterpoint_type
waterpoint_type
_group

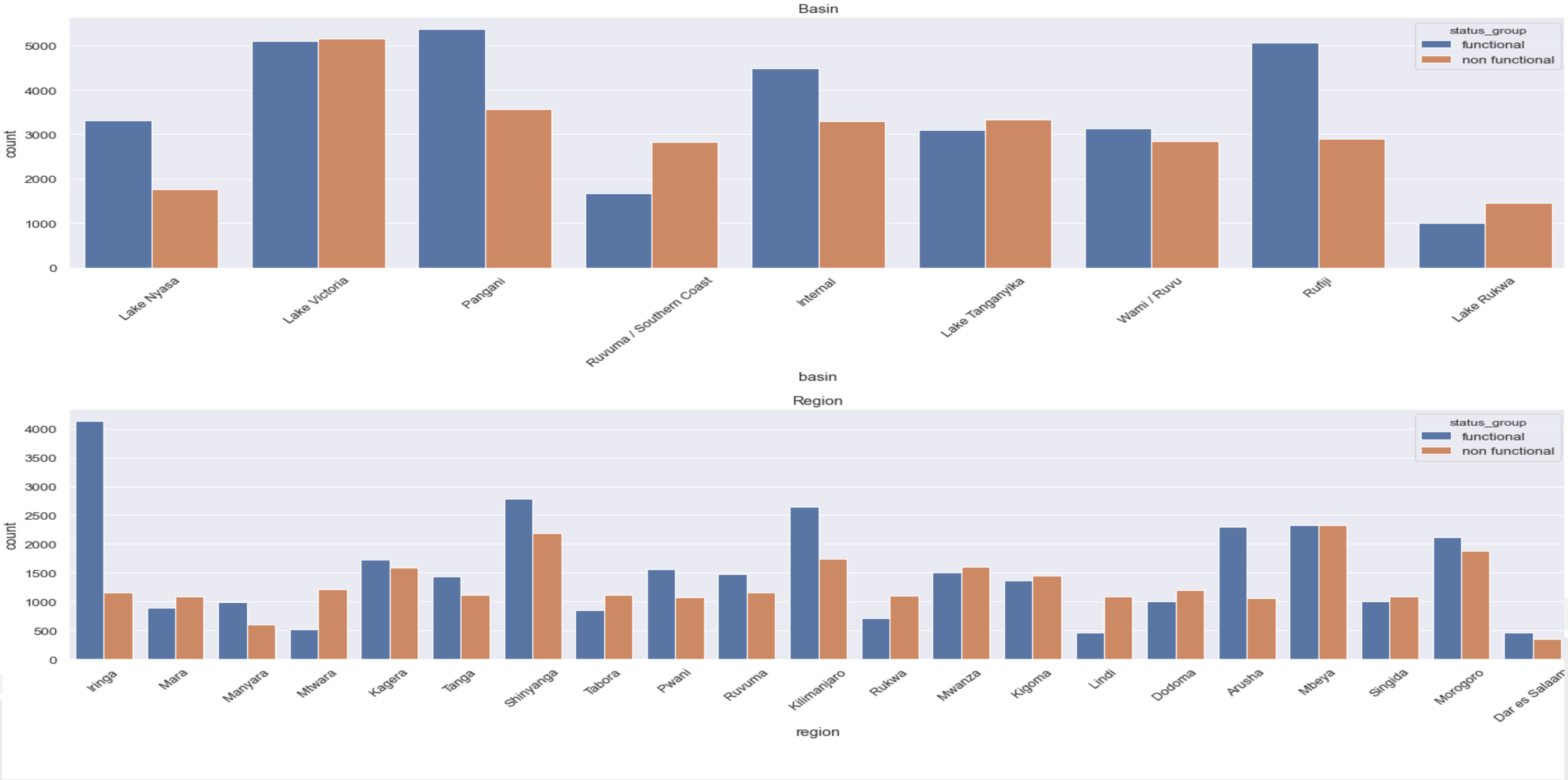
Location
gps_height
longitude
latitude

Management
management
management_group

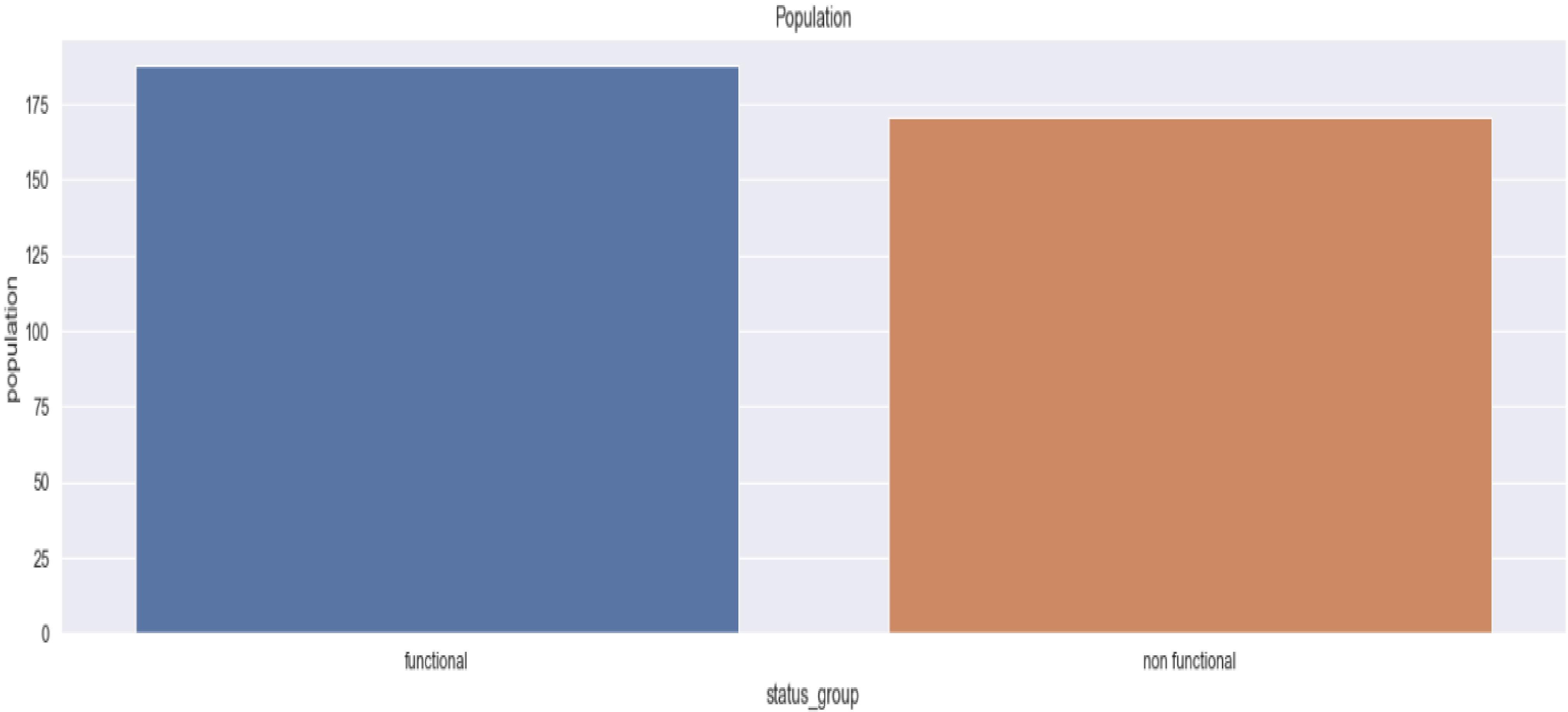
People
population
public_meeting
scheme_management
scheme_name
permit
construction_year

Insights Discovered

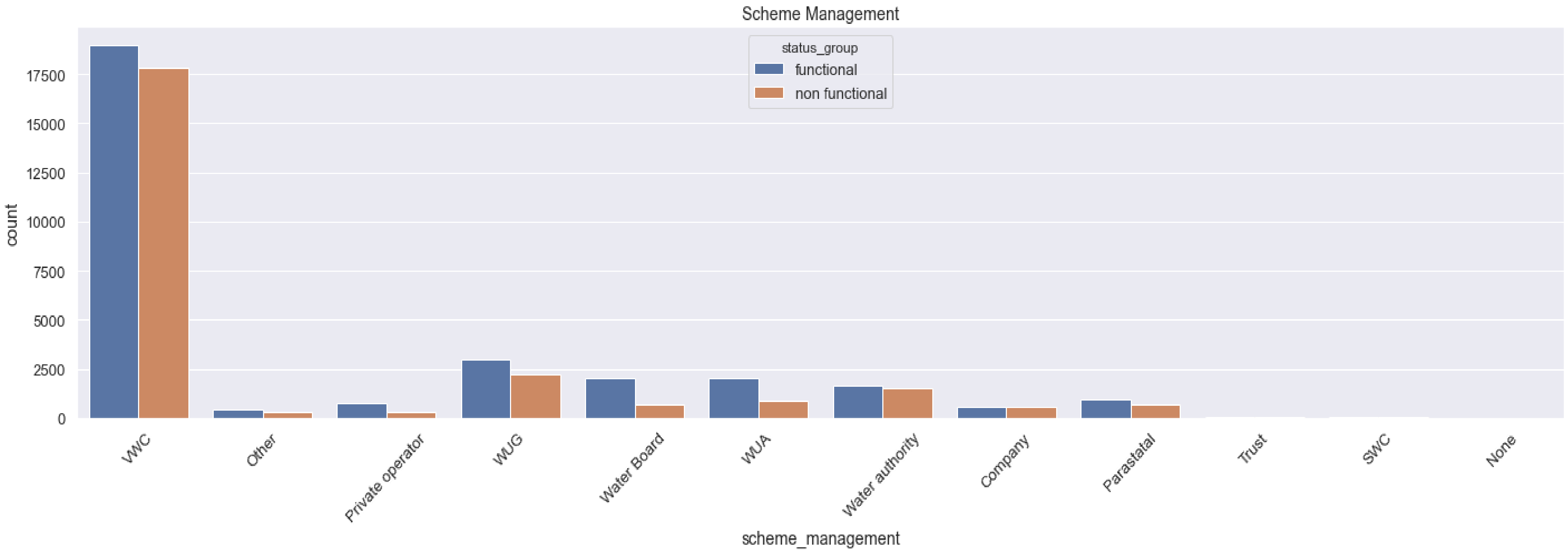
Pump waters varies per region or districts



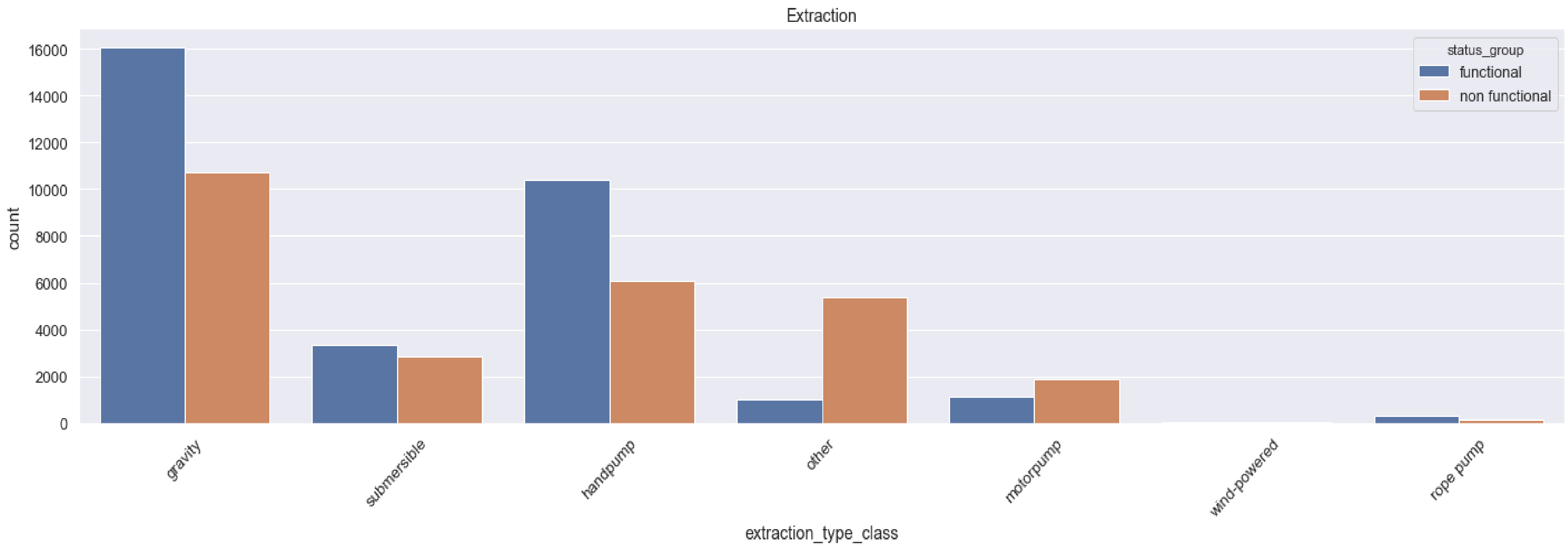
Population has not much effect on pumps



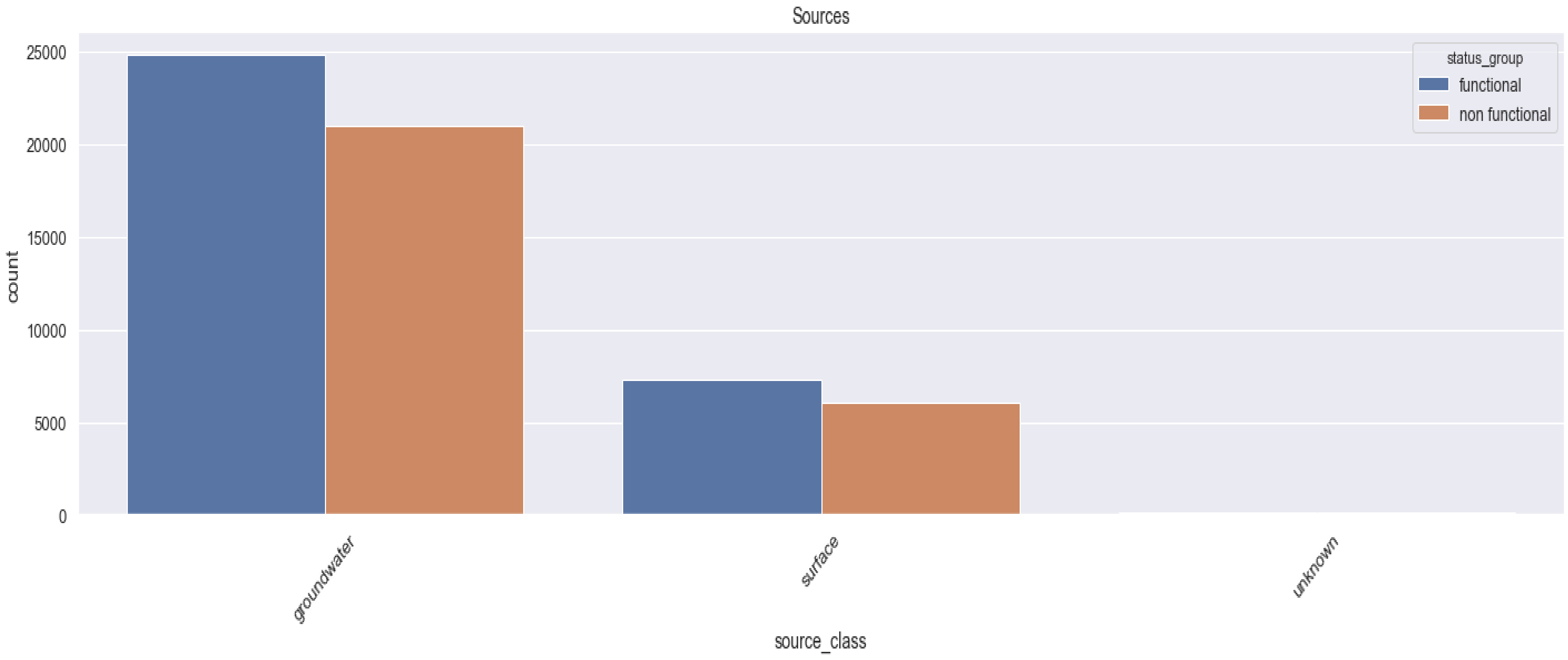
Management types are dominant on one party



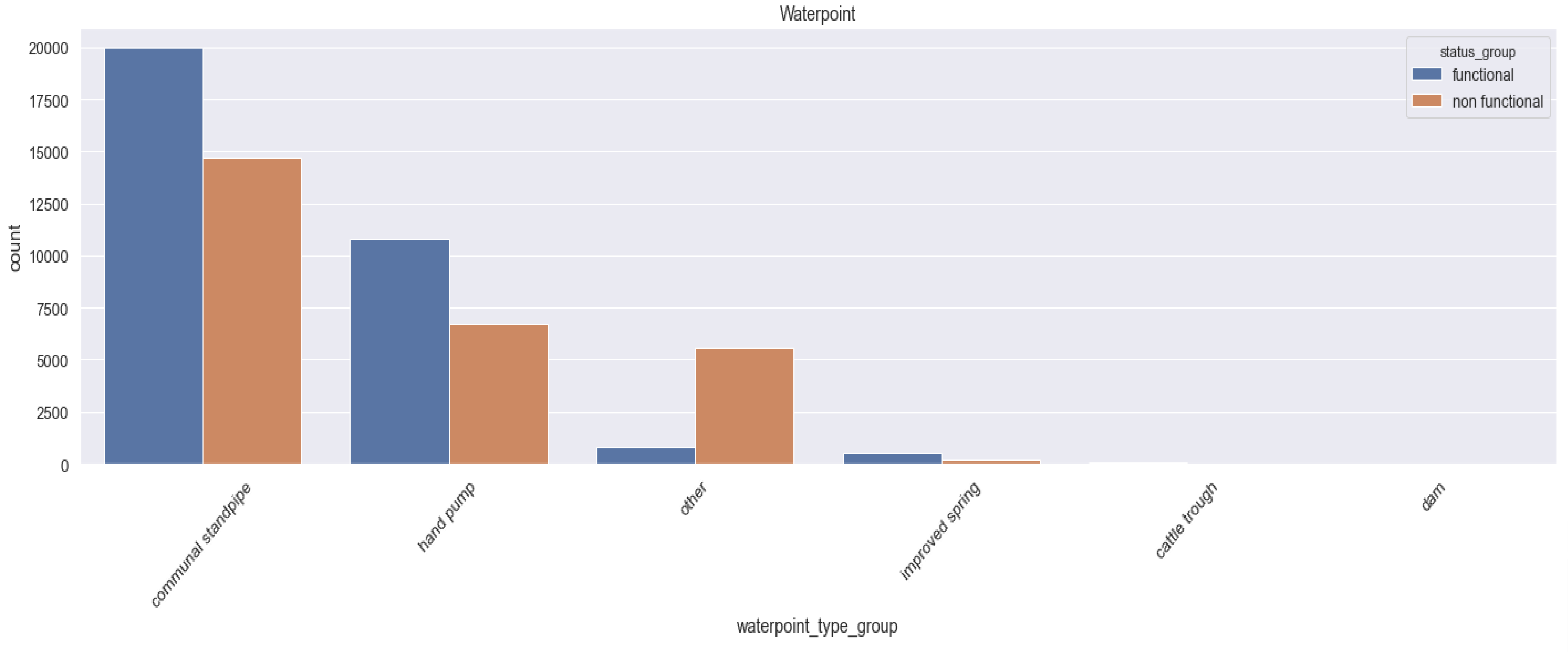
Gravity extraction is main technique used



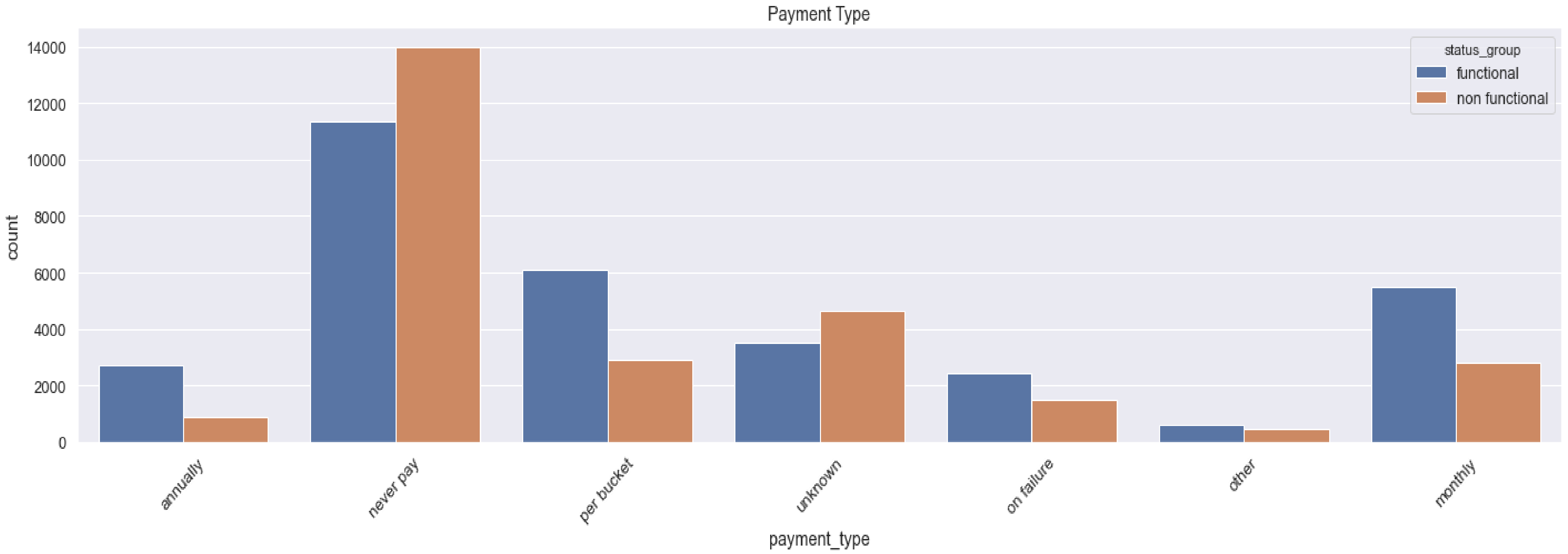
Ground Water sources are most common



Standpipe types are most common



No water payment disrupts water pumps



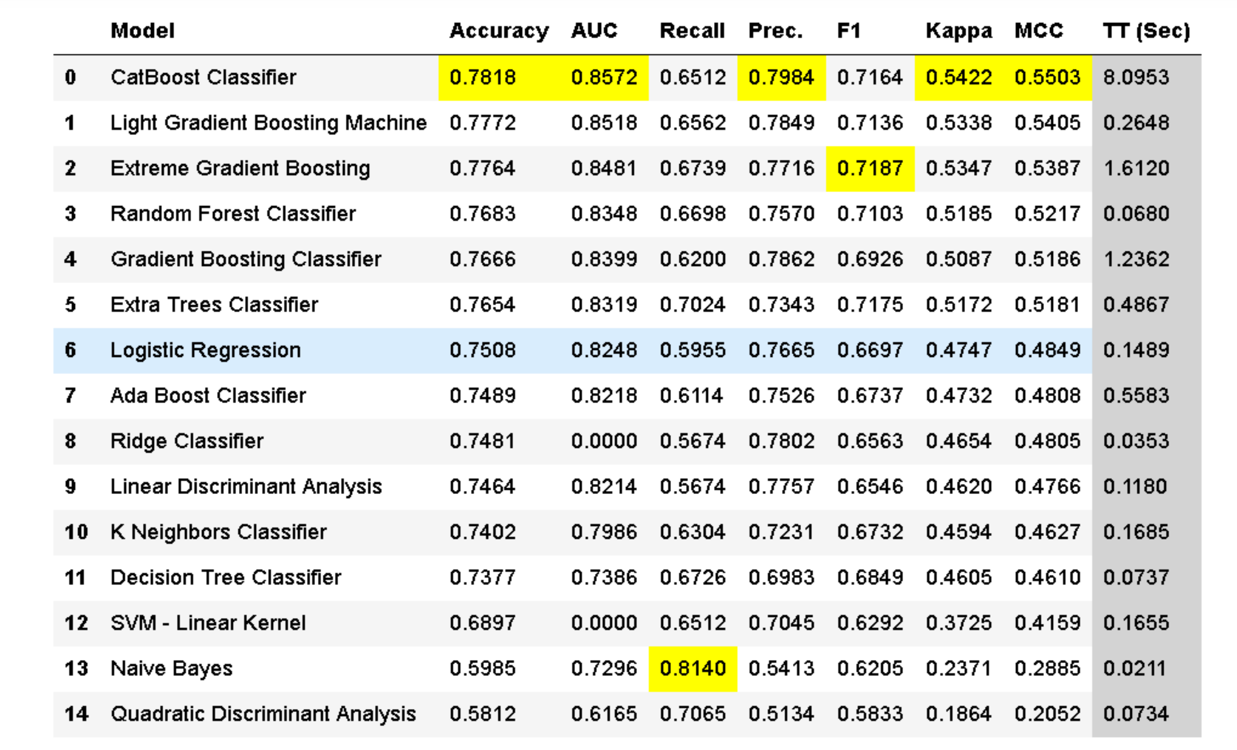
Insights Discussion

Comparing with initial hypothesis

- Basins, Regions and Districts has influence on water pump operations.
- Population are more or less similar in numbers and did not much affect water pump operations, hypothesis is rejected
- Extraction of water, sources and waterpoints are inconsistent, hence the hypothesis don’t stand

Model and Metrics

- We use XGBoost Model for prediction.



	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
0	CatBoost Classifier	0.7818	0.8572	0.6512	0.7984	0.7164	0.5422	0.5503	8.0953
1	Light Gradient Boosting Machine	0.7772	0.8518	0.6562	0.7849	0.7136	0.5338	0.5405	0.2648
2	Extreme Gradient Boosting	0.7764	0.8481	0.6739	0.7716	0.7187	0.5347	0.5387	1.6120
3	Random Forest Classifier	0.7683	0.8348	0.6698	0.7570	0.7103	0.5185	0.5217	0.0680
4	Gradient Boosting Classifier	0.7666	0.8399	0.6200	0.7862	0.6926	0.5087	0.5186	1.2362
5	Extra Trees Classifier	0.7654	0.8319	0.7024	0.7343	0.7175	0.5172	0.5181	0.4867
6	Logistic Regression	0.7508	0.8248	0.5955	0.7665	0.6697	0.4747	0.4849	0.1489
7	Ada Boost Classifier	0.7489	0.8218	0.6114	0.7526	0.6737	0.4732	0.4808	0.5583
8	Ridge Classifier	0.7481	0.0000	0.5674	0.7802	0.6563	0.4654	0.4805	0.0353
9	Linear Discriminant Analysis	0.7464	0.8214	0.5674	0.7757	0.6546	0.4620	0.4766	0.1180
10	K Neighbors Classifier	0.7402	0.7986	0.6304	0.7231	0.6732	0.4594	0.4627	0.1685
11	Decision Tree Classifier	0.7377	0.7386	0.6726	0.6983	0.6849	0.4605	0.4610	0.0737
12	SVM - Linear Kernel	0.6897	0.0000	0.6512	0.7045	0.6292	0.3725	0.4159	0.1655
13	Naive Bayes	0.5985	0.7296	0.8140	0.5413	0.6205	0.2371	0.2885	0.0211
14	Quadratic Discriminant Analysis	0.5812	0.6165	0.7065	0.5134	0.5833	0.1864	0.2052	0.0734

Metrics	Functional	Non Functional
Precision	0.76	0.80
Recall	0.87	0.65
F1-Score	0.81	0.72
Accuracy	78%	
AUC	0.85	

Recommendations and Actions

Summary

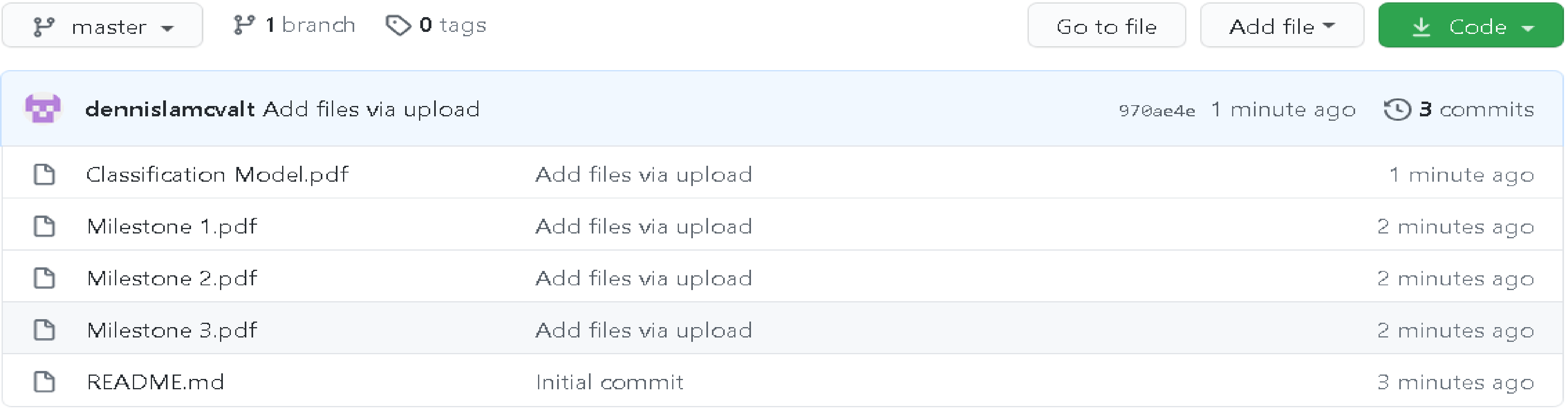
- We have created the project proposal and establish initial hypothesis.
- Dataset is explored and added visualizations for clarity
- A recommended machine learning model was created to predict water pump operations in Tanzania

Recommended Actions

- Areas which has sparse water pumps may need to be increased for consumption
- Revamp or restructure water pump management companies
- Explore other methods of water extraction
- Possible to include other water sources?
- Conversion to waterpipes for easy distribution if possible
- Water pricing revision to allow affordable payment

Thank you

Appendix: PDF Reports Download



dennislamcvalt Add files via upload 970ae4e 1 minute ago **3** commits

Classification Model.pdf	Add files via upload	1 minute ago
Milestone 1.pdf	Add files via upload	2 minutes ago
Milestone 2.pdf	Add files via upload	2 minutes ago
Milestone 3.pdf	Add files via upload	2 minutes ago
README.md	Initial commit	3 minutes ago

GitHub Link:
https://github.com/dennislamcvalt/SQLforDataScienceCapstone