ReDI School of
Digital Integration

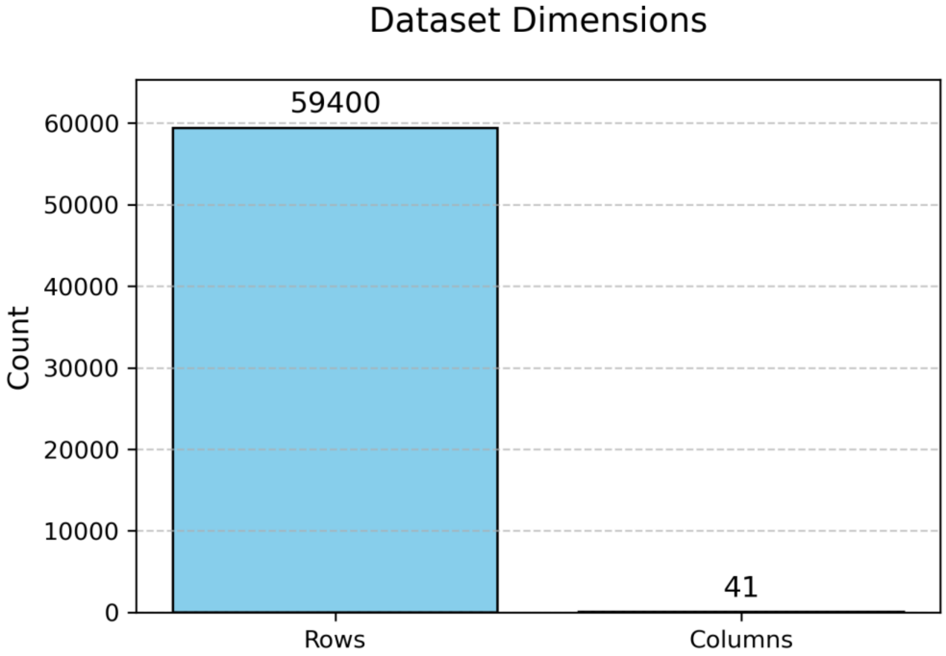
Sidra Bibi, Ariful Islam

Water Pump

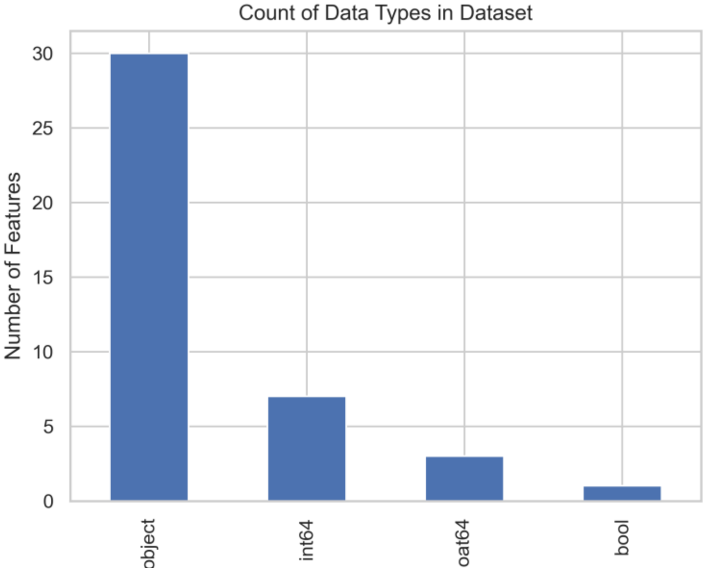
16.4.2025

We use tech to connect human potential and
opportunity with dignity & humility

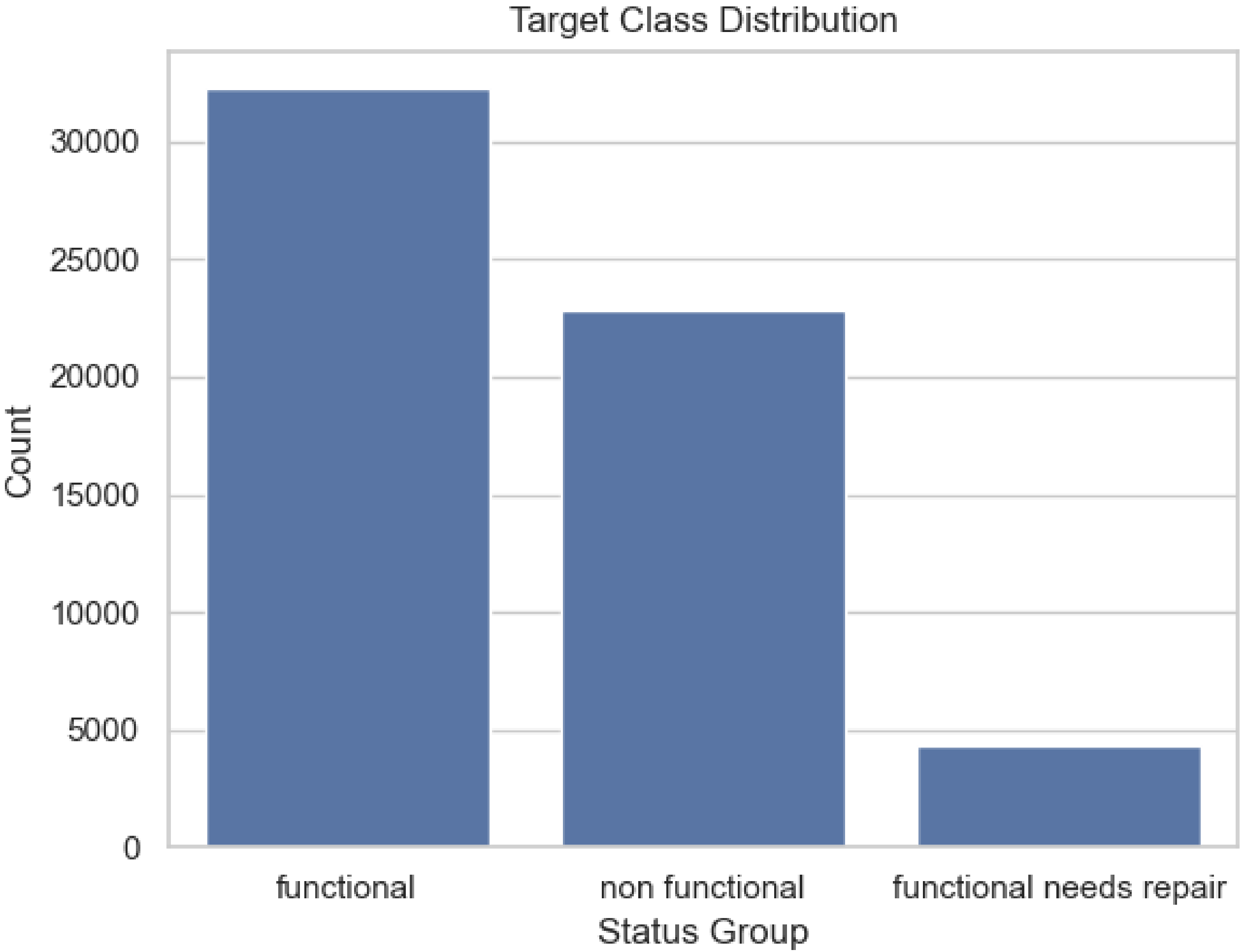
Data Types



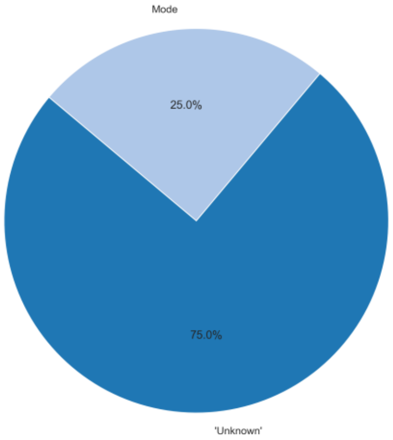
Dataset Dimensions
Count
60000
50000
40000
30000
20000
10000
0
59400
41
Rows
Columns



Count of Data Types in Dataset
Number of Features
30
25
20
15
10
5
0
object
int64
oat64
bool

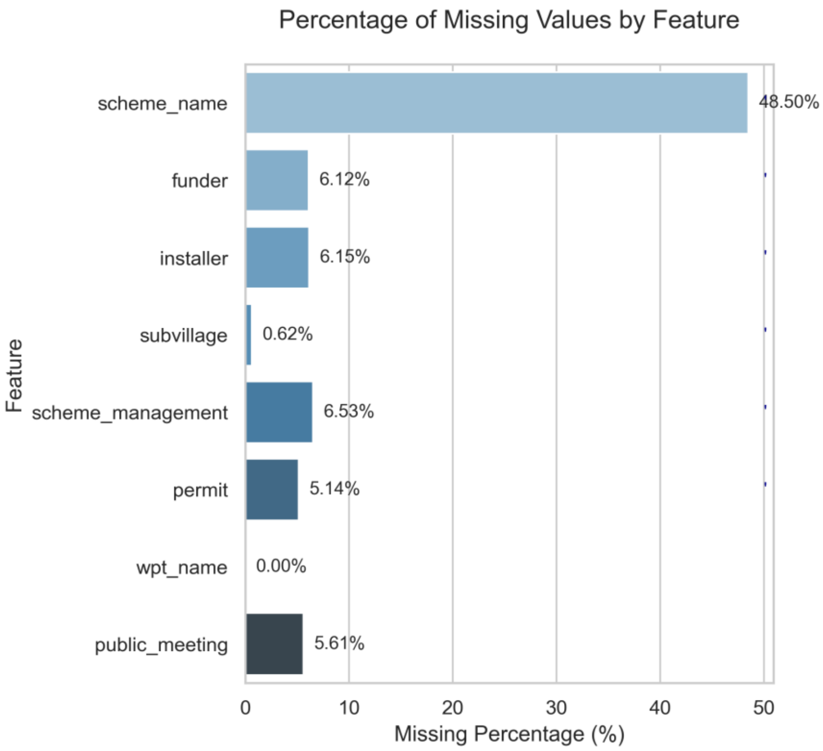

Target Class Distribution
Count
30000
25000
20000
15000
10000
5000
0
functional
non functional
functional needs repair
Status Group

Missing values and imputation



Imputation Strategies Used

- 'Unknown': Used for high missingness (e.g., scheme_name: 48.5%) or high cardinality
- Mode: Used for low missingness (e.g., wpt_name: 0.003%) or binary features



Percentage of Missing Values by Feature

Unknown' (48.5% missing, high cardinality)

Unknown' (6.1% missing, 1896 unique values)

Unknown' (6.2% missing, 2145 unique values)

Unknown' (0.6% missing, 19287 unique values)

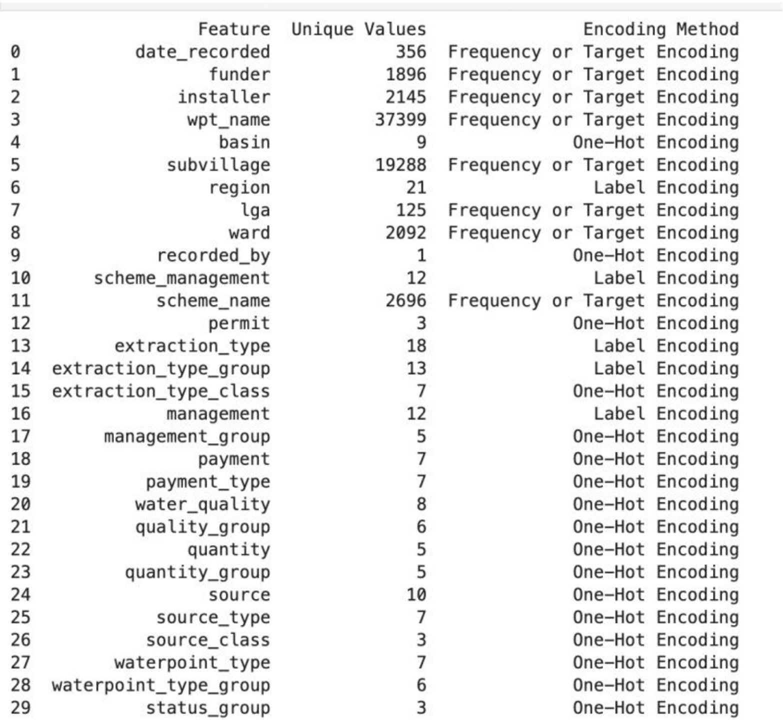
Unknown' (6.5% missing, 11 unique values)

Unknown' (5.1% missing, binary but significant missingness)

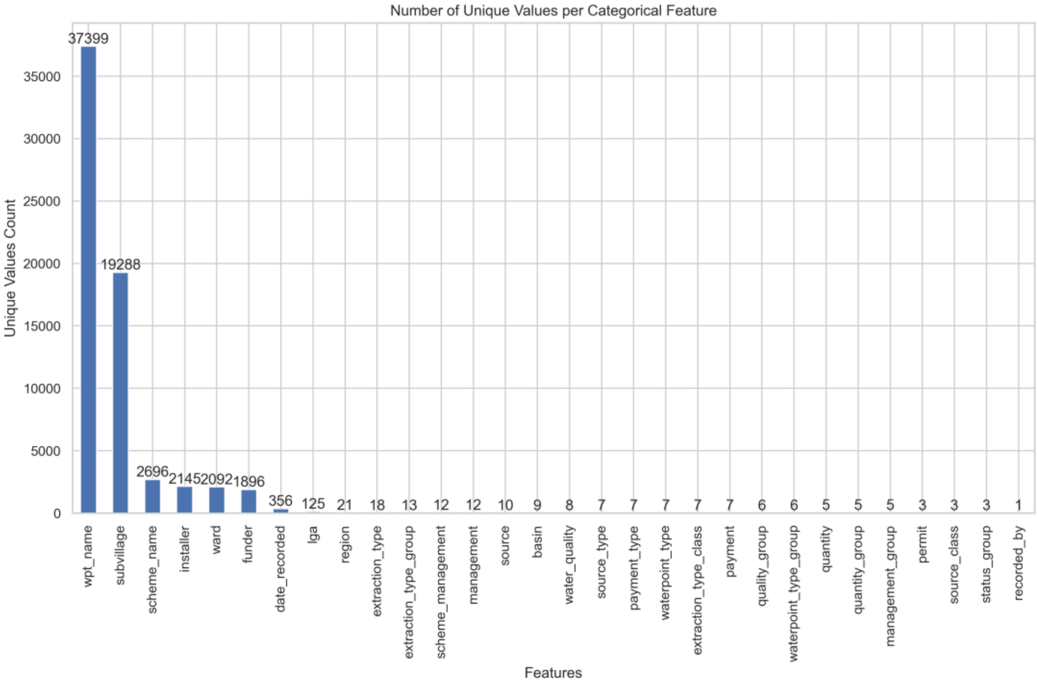
Mode (0.003% missing, minimal impact)

Mode (5.6% missing, binary feature)

Proposed Encoding Techniques

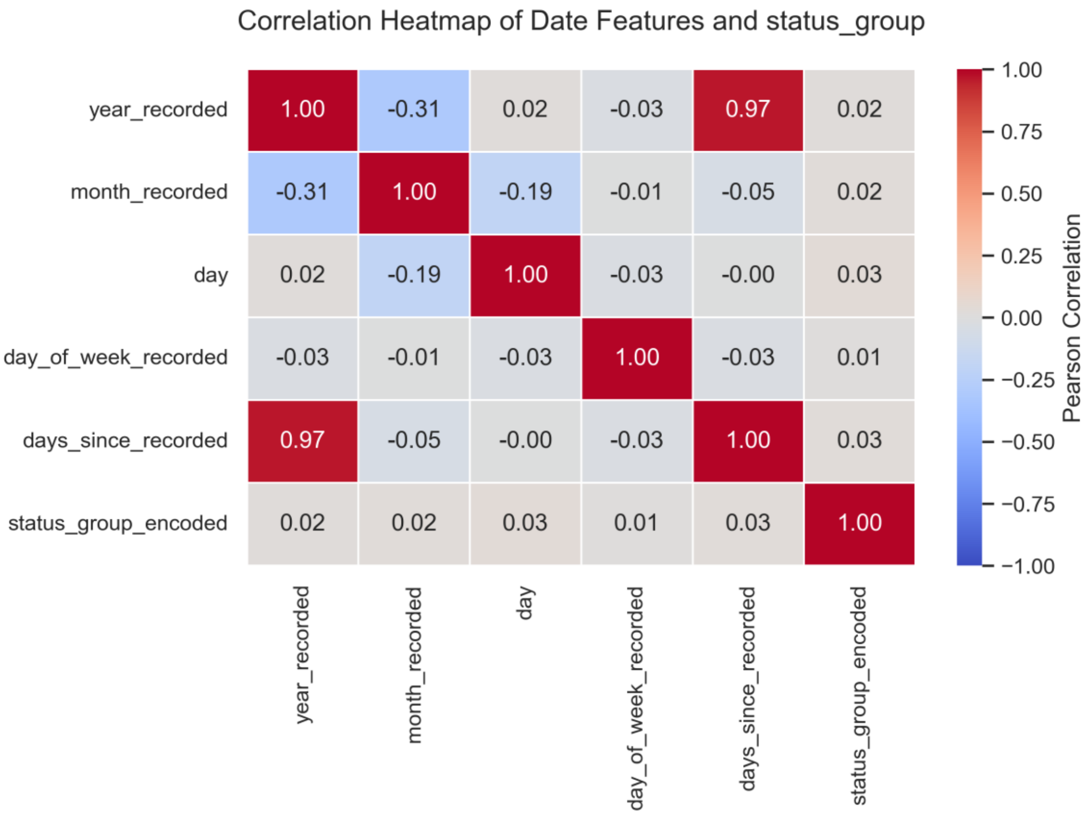


```
                  Feature  Unique Values             Encoding Method
0           date_recorded            356  Frequency or Target Encoding
1                  funder           1896  Frequency or Target Encoding
2               installer           2145  Frequency or Target Encoding
3                wpt_name          37399  Frequency or Target Encoding
4                   basin              9              One-Hot Encoding
5              subvillage          19288  Frequency or Target Encoding
6                  region             21                Label Encoding
7                     lga            125  Frequency or Target Encoding
8                    ward           2092  Frequency or Target Encoding
9             recorded_by              1              One-Hot Encoding
10      scheme_management             12                Label Encoding
11            scheme_name           2696  Frequency or Target Encoding
12                 permit              3              One-Hot Encoding
13        extraction_type             18                Label Encoding
14  extraction_type_group             13                Label Encoding
15  extraction_type_class              7              One-Hot Encoding
16             management             12                Label Encoding
17       management_group              5              One-Hot Encoding
18                payment              7              One-Hot Encoding
19           payment_type              7              One-Hot Encoding
20          water_quality              8              One-Hot Encoding
21          quality_group              6              One-Hot Encoding
22               quantity              5              One-Hot Encoding
23         quantity_group              5              One-Hot Encoding
24                 source             10              One-Hot Encoding
25            source_type              7              One-Hot Encoding
26           source_class              3              One-Hot Encoding
27         waterpoint_type              7              One-Hot Encoding
28   waterpoint_type_group              6              One-Hot Encoding
29           status_group              3              One-Hot Encoding
```



Features Extraction from Date_recorded

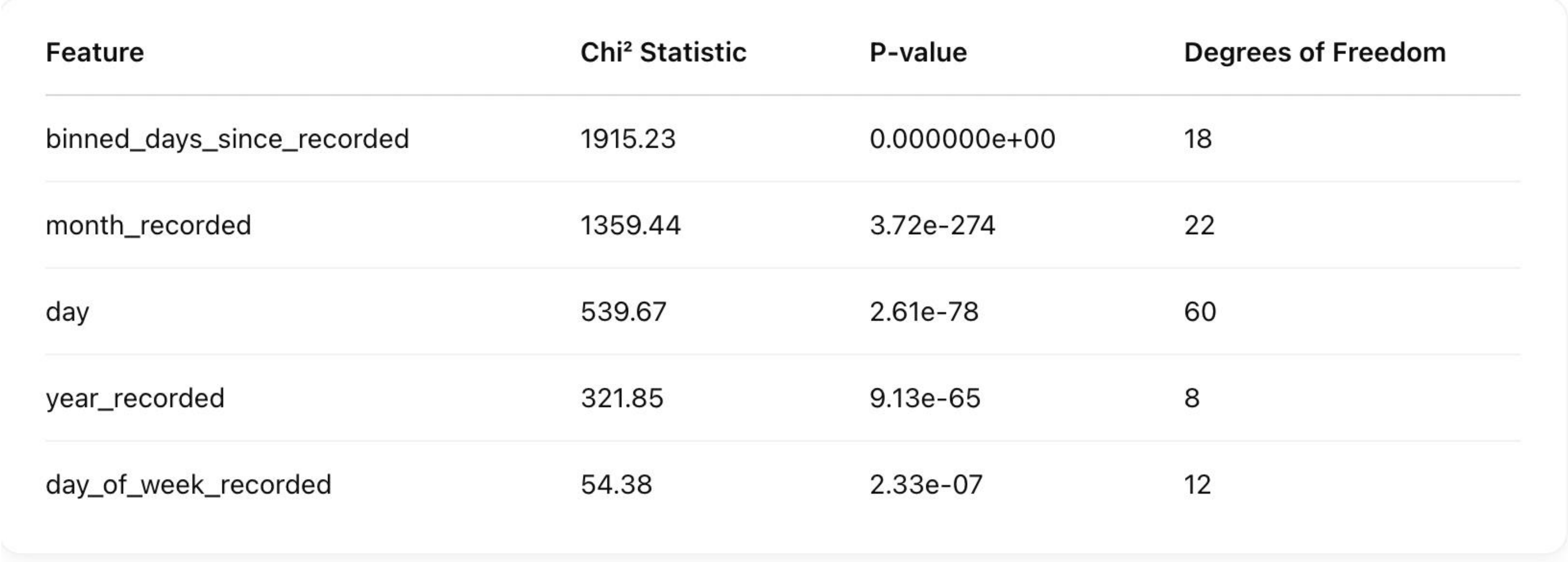
- year_recorded
- month_recored
- day
- day_of_week_recorded
- days_since_recorded



Correlation Heatmap of Date Features and status_group

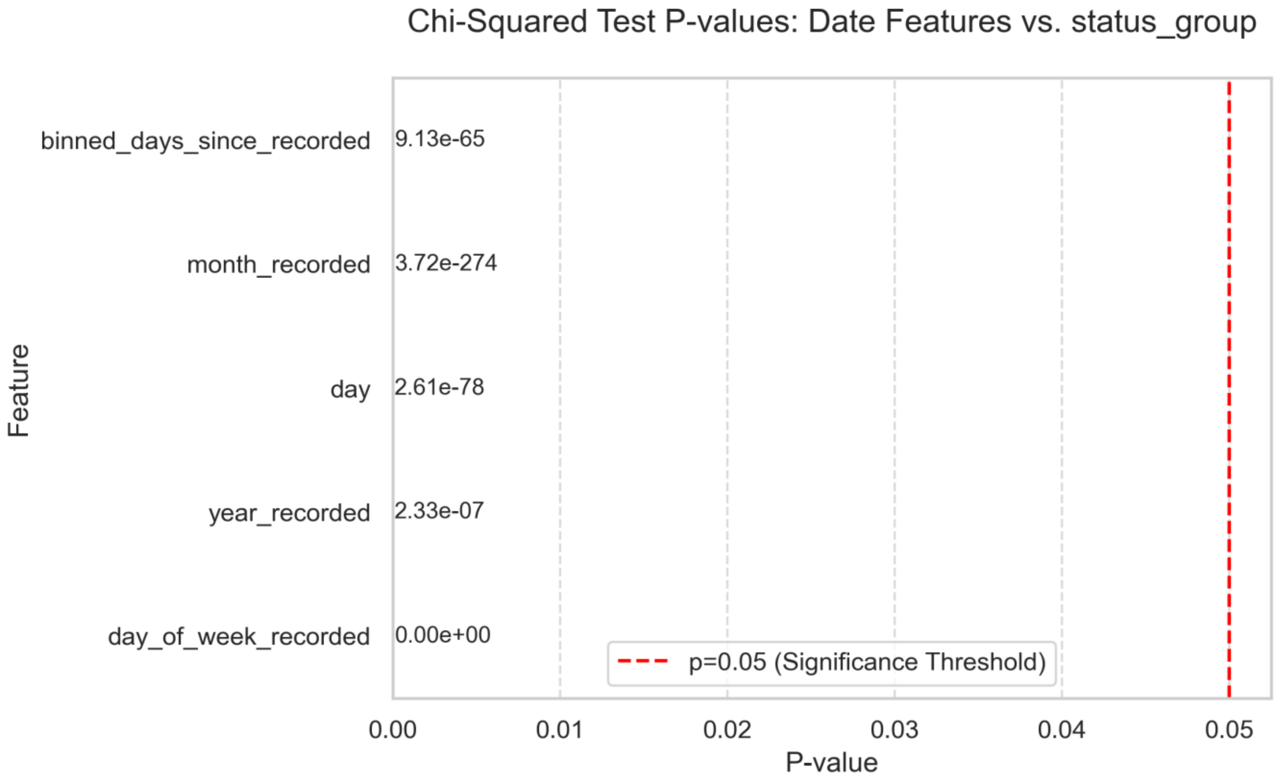
	year_recorded	month_recorded	day	day_of_week_recorded	days_since_recorded	status_group_encoded
year_recorded	1.00	-0.31	0.02	-0.03	0.97	0.02
month_recorded	-0.31	1.00	-0.19	-0.01	-0.05	0.02
day	0.02	-0.19	1.00	-0.03	-0.00	0.03
day_of_week_recorded	-0.03	-0.01	-0.03	1.00	-0.03	0.01
days_since_recorded	0.97	-0.05	-0.00	-0.03	1.00	0.03
status_group_encoded	0.02	0.02	0.03	0.01	0.03	1.00

Date_record Features Chi-squared

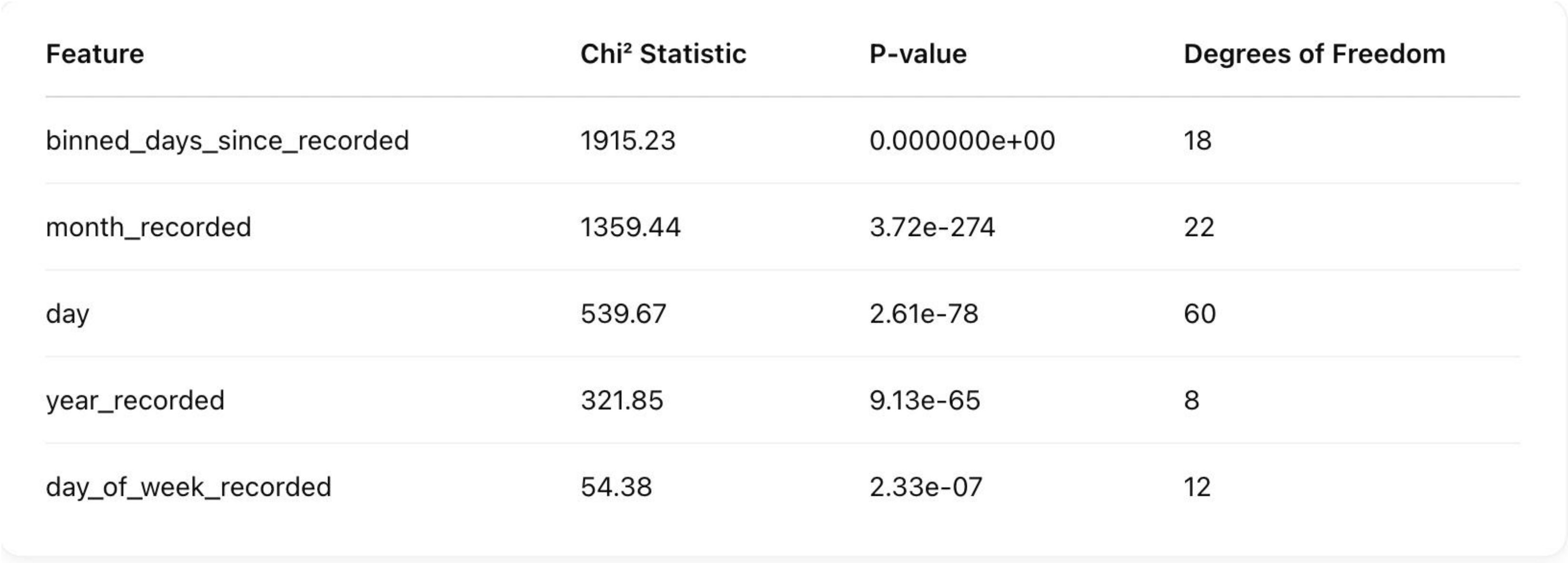


Feature	Chi² Statistic	P-value	Degrees of Freedom
binned_days_since_recorded	1915.23	0.000000e+00	18
month_recorded	1359.44	3.72e-274	22
day	539.67	2.61e-78	60
year_recorded	321.85	9.13e-65	8
day_of_week_recorded	54.38	2.33e-07	12

Ch2-P values

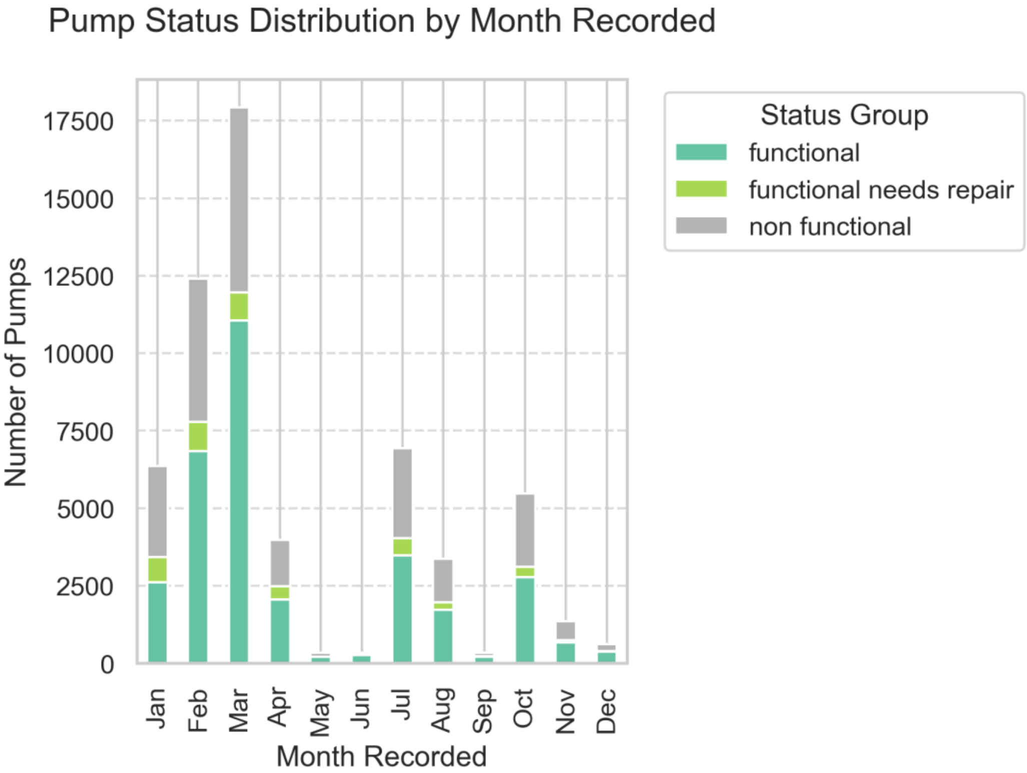

Chi-Squared Test P-values: Date Features vs. status_group
Feature
binned_days_since_recorded
month_recorded
day
year_recorded
day_of_week_recorded
9.13e-65
3.72e-274
2.61e-78
2.33e-07
0.00e+00
p=0.05 (Significance Threshold)
0.00
0.01
0.02
0.03
0.04
0.05
P-value

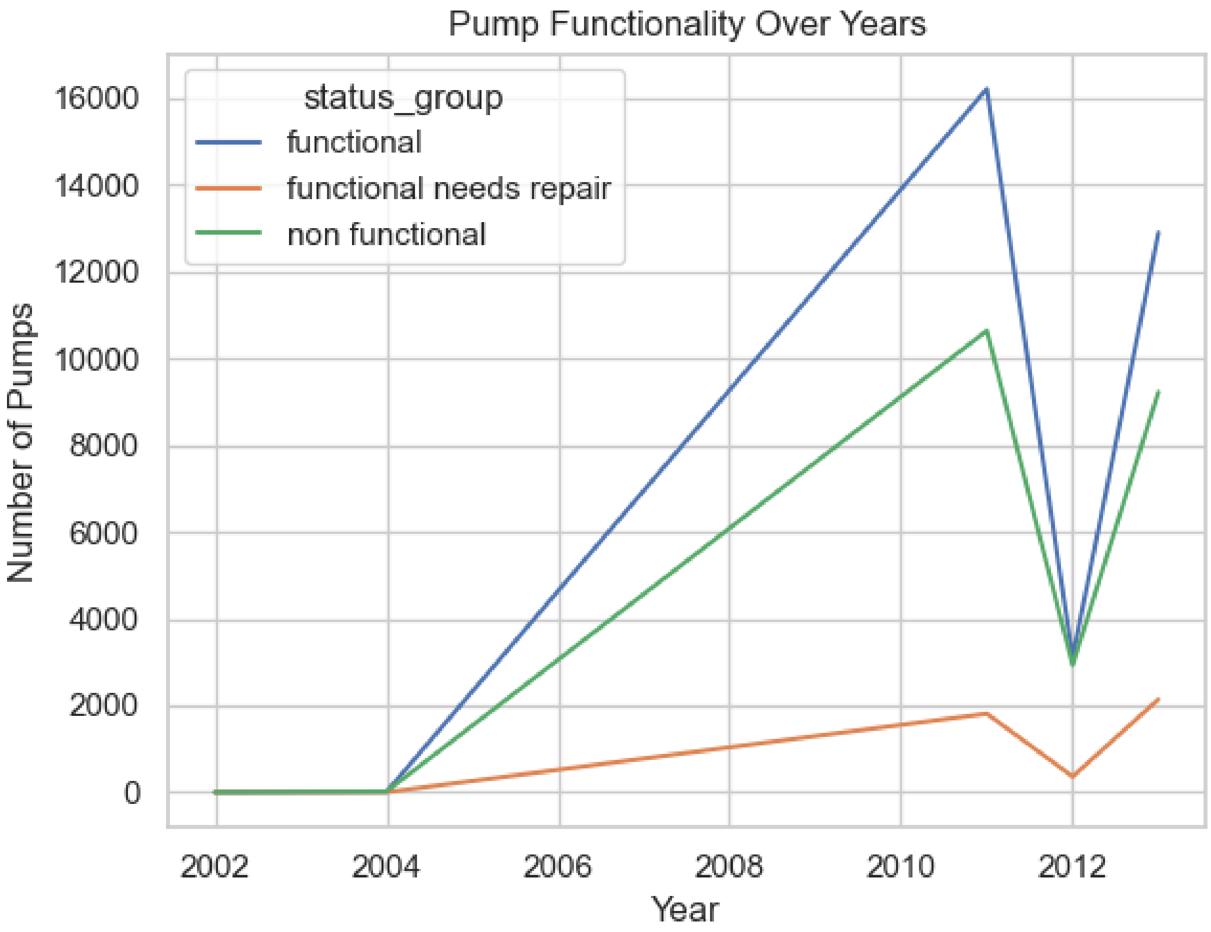
Chi-Squared value



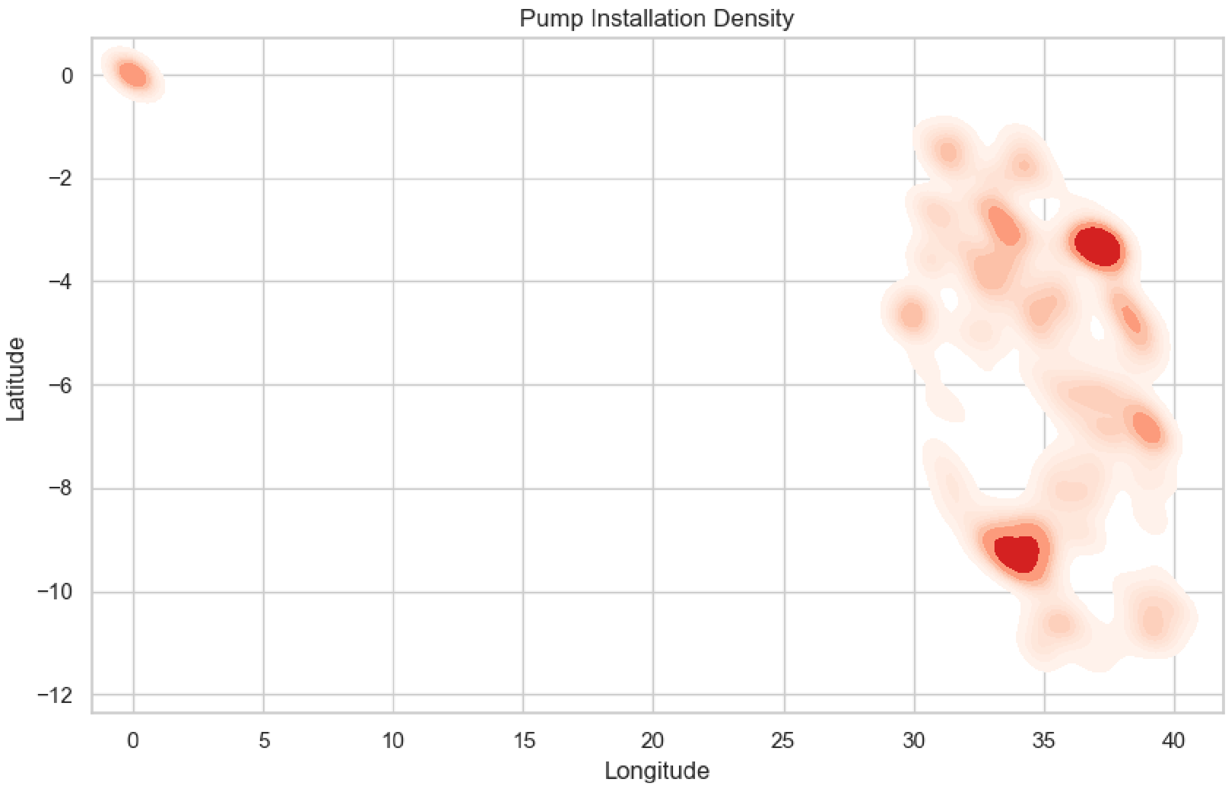
Feature	Chi^2 Statistic	P-value	Degrees of Freedom
binned_days_since_recorded	1915.23	0.000000e+00	18
month_recorded	1359.44	3.72e-274	22
day	539.67	2.61e-78	60
year_recorded	321.85	9.13e-65	8
day_of_week_recorded	54.38	2.33e-07	12

Monthly status

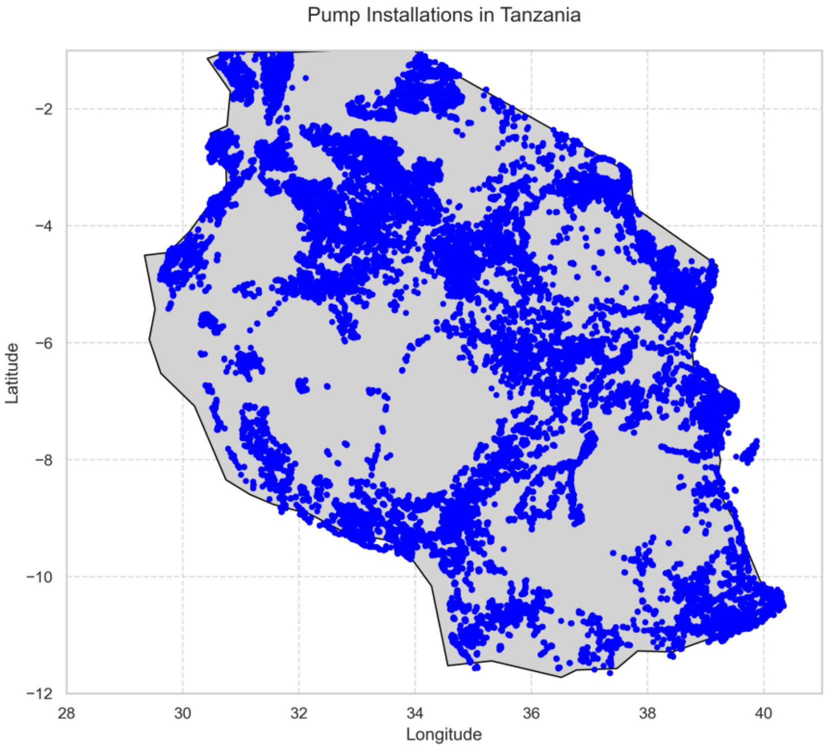



Pump Functionality Over Years
status_group
functional
functional needs repair
non functional
Number of Pumps
0
2000
4000
6000
8000
10000
12000
14000
16000
2002
2004
2006
2008
2010
2012
Year

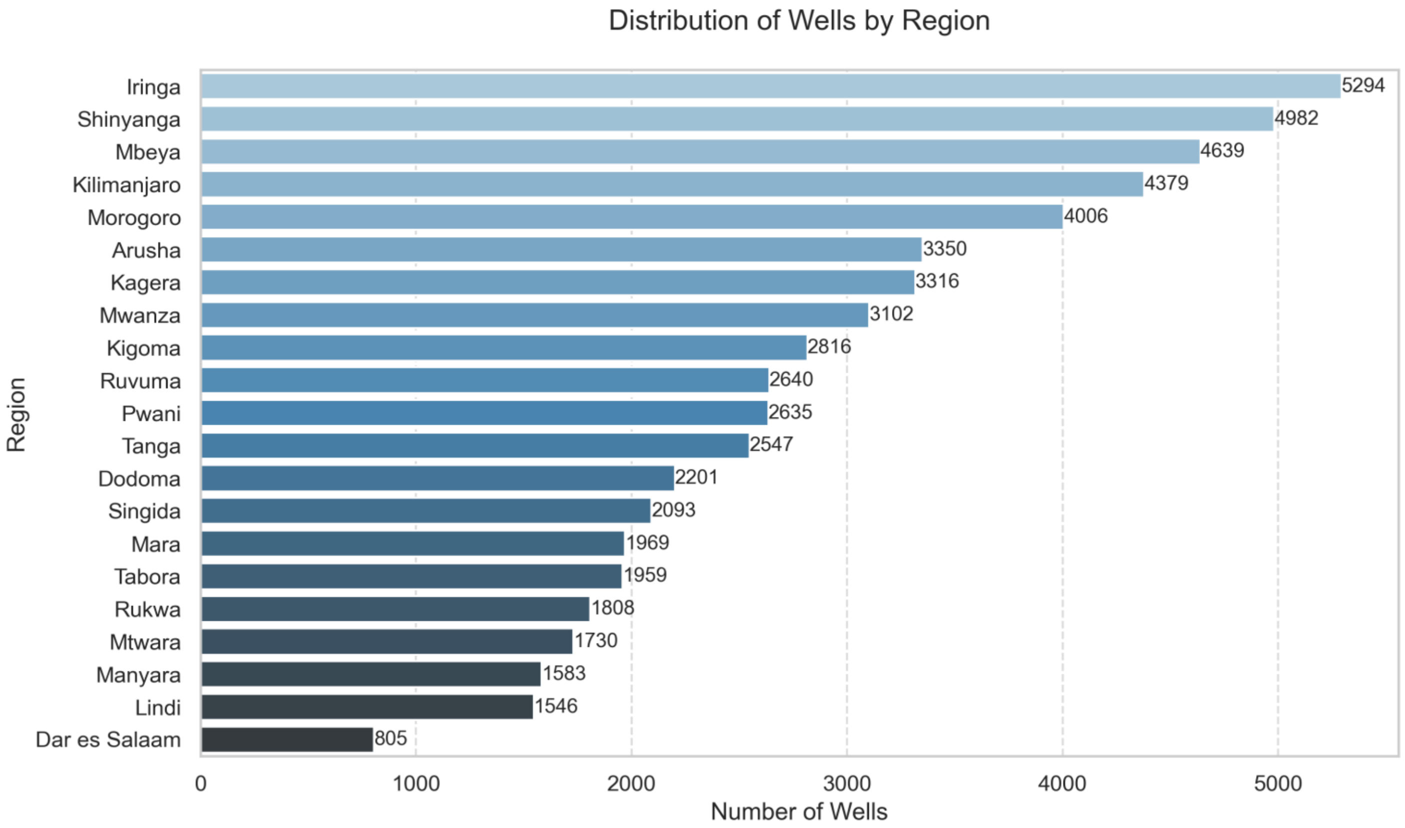
Pump installation Density



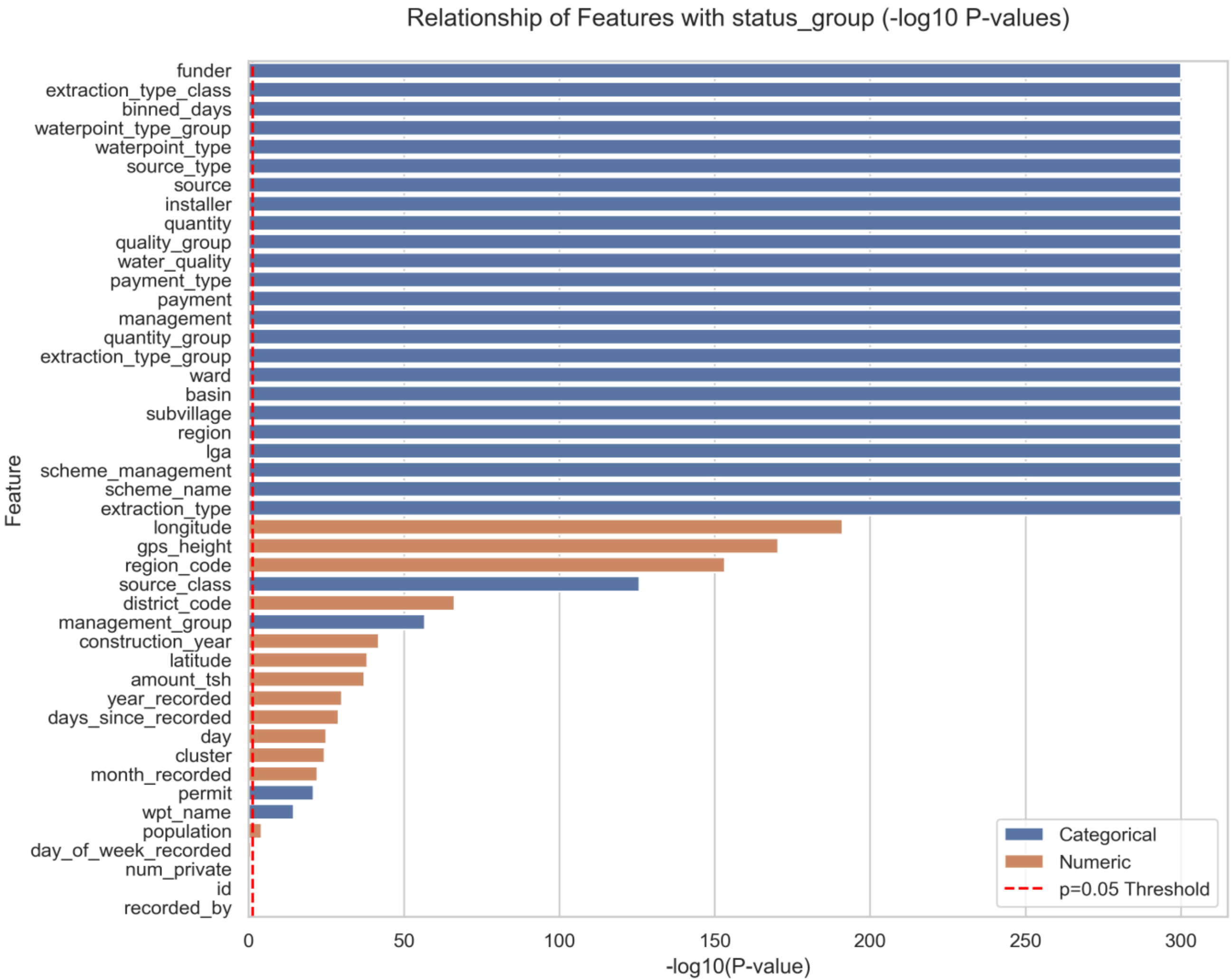
Water Pump in Tanzania



Distrinution water pump based on region



- ❖ Chi-squared Test
- ❖ ANOVA Test



Thank you!