TEAM Hamburg

Water Pump Functionality Prediction

This project develops machine learning models to predict the functionality status of water pumps across Tanzania based on a variety of features including location, water quality, management structure, and technical specifications.
This project addresses a critical infrastructure challenge: identifying which water pumps are functional, which need repairs, and which are non-functional to improve maintenance operations and ensure communities have access to clean water

- Training data set contains 59400 rows and 40 columns.
- After analyse longitude (-12,0) and latitude (29,41) for region Tanzania remained 57588 rows.
- Simple analyse of columns showed that we can gezt rid of columns:
  - ✓ waterpoint_type_group(=waterpoint_type)
  - ✓ source_class(=source_type)
  - ✓ source (=source_type, but without unknown)
  - ✓ quantity_group(=quantity)
  - ✓ quality_group (=water_quality)
  - ✓ payment_type(=payment)
  - ✓ management_group(=management)
  - ✓ extraction_type_group
  - ✓ extraction_type_class
  - ✓ scheme_name(because 48,5% are NULL)
  - ✓ recorded_by (because only one value)
  - ✓ public_meeting (92% the same value)
  - ✓ num_private(98,3% unique value)
  - ✓ wpt_name?(because its only name, no information)
  - ✓ date_recorded

```
◆ Total row: 57588

◆ Column: funder
Null values: 3624
Empty strings: 0
Unique values: 1857
-----------------------------------------
Top 10 most frequent values:
                      value  count  percent
Government Of Tanzania        8842    15.35
                       NaN    3624     6.29
                    Danida    3114     5.41
                    Hesawa    1914     3.32
                 World Bank   1345     2.34
                      Kkkt    1287     2.23
               World Vision   1224     2.13
                      Rwssp   1187     2.06
                    Unicef    1035     1.80
           District Council    843     1.46
```

Questions:

why region <> region_code (should be equal):

Analyse shows mistakes and it possible to chahge region code with most common value :

Out[58]:

	region	region_code	count
0	Arusha	2	3024
1	Arusha	24	326
2	Dar es Salaam	7	805
3	Dodoma	1	2201
4	Iringa	11	5294
5	Kagera	18	3316
6	Kigoma	16	2816
7	Kilimanjaro	3	4379
8	Lindi	8	300
9	Lindi	18	8
10	Lindi	80	1238

Questions:

What is the ideal row?

```
df_func =
pump_data_merged[pump_data_merged["status_group"] ==
"functional"]
# choose only categorical/object columns
cat_cols = df_func
.select_dtypes(include=["object"]).columns

ideal_values = {}
for col in cat_cols:
    top_value = df_func [col].mode(dropna=True)
    if not top_value.empty:
        percent = df_func
[col].value_counts(normalize=True,
dropna=True).iloc[0] * 100
        ideal_values[col] = (top_value.iloc[0],
round(percent, 2))

# show nicely
import pandas as pd
ideal_df = pd.DataFrame.from_dict(ideal_values,
orient="index", columns=["Most common value", "%
within functional"])
ideal_df.sort_values("% within functional",
ascending=False).head(15)
```

Out[69]:

	Most common value	% within functional
status_group	functional	100.00
water_quality	soft	89.15
permit	True	71.36
quantity	enough	67.29
management	vwc	64.01
scheme_management	VWC	63.11
waterpoint_type	communal standpipe	56.22
extraction_type	gravity	51.04
payment	never pay	34.93
source_type	spring	33.71
installer	DWE	30.31
basin	Pangani	17.11
region	Iringa	13.19
funder	Government Of Tanzania	12.39
lga	Njombe	6.39

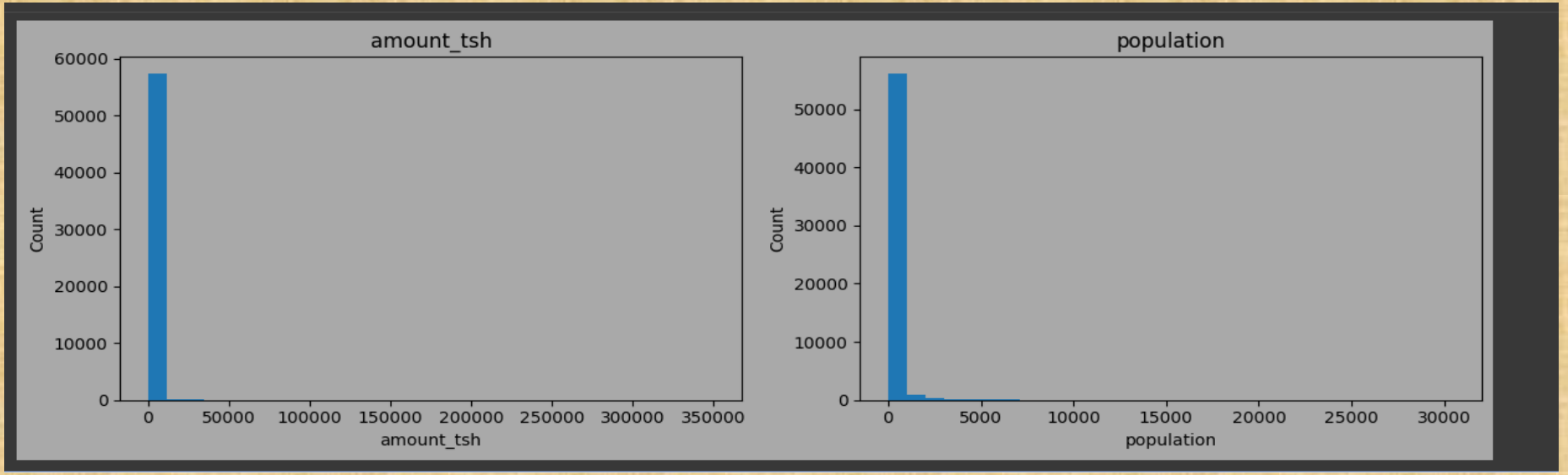
Analyse numeric columns

```
from re import I
# 6) Numeric summary

num_cols = ["amount_tsh","population"]
if num_cols:
    print("\nNumeric columns:", num_cols[:15], "…" if len(num_cols) > 15 else "")
    display(pump_data_merged[num_cols].describe().T)
```

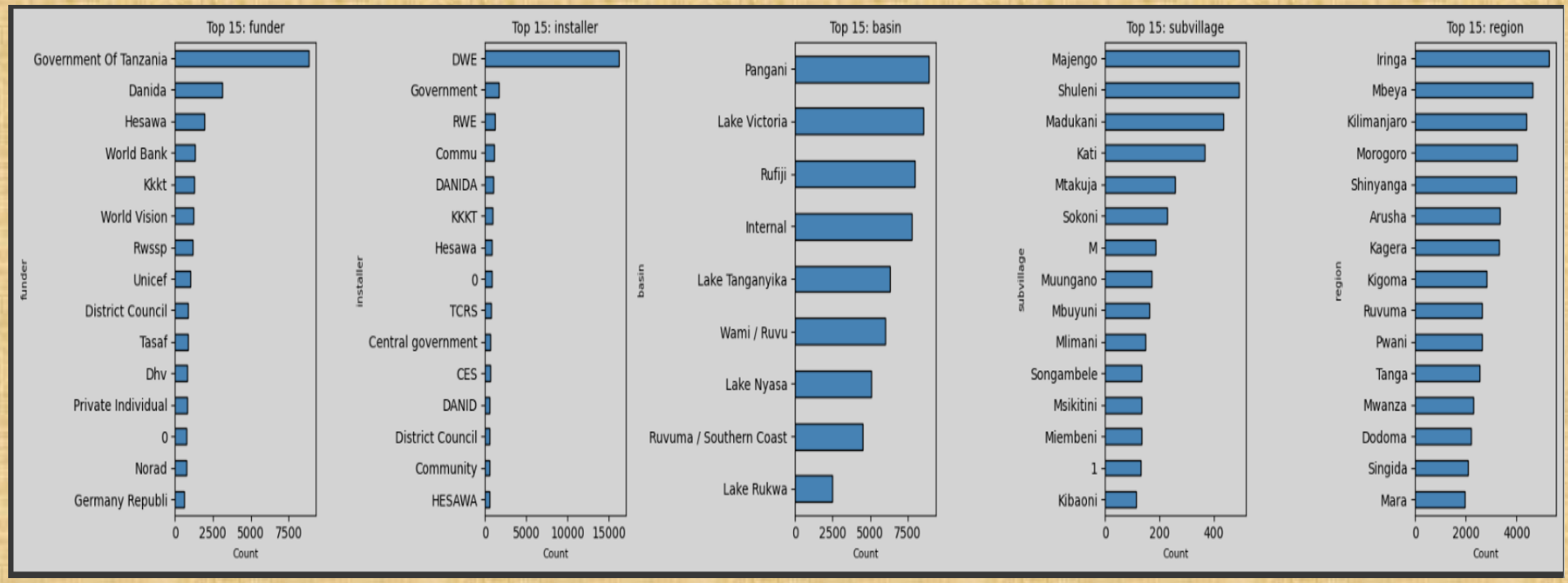
```
Numeric columns: ['amount_tsh', 'population']
```

	count	mean	std	min	25%	50%	75%	max
amount_tsh	57588.0	327.645219	3043.831403	0.0	0.0	0.0	30.0	350000.0
population	57588.0	185.570831	477.744239	0.0	0.0	35.0	230.0	30500.0



Analyse categorical columns

	n_unique
subvillage	18567
installer	2113
ward	2033
funder	1857
lga	124
region	21
extraction_type	18
management	12
scheme_management	11
basin	9
water_quality	8
source_type	7
payment	7
waterpoint_type	7
quantity	5
status_group	3
permit	2



Analyse NULL and „ „ and 0 values

	null_fraction
scheme_management	6.5%
installer	6.3%
funder	6.3%
permit	5.3%
subvillage	0.6%

- Construction_year has a lot of 0 values

```
column_summary(pump_data, "construction_year")

  🔹 Total rows: 57588
  🔹 Column: construction_year
Null values: 0
Empty strings: 0
Numeric 0 values: 18897
String '0' values: 0
'Unknown' values (any case): 0
Unique values: 55
------------------------------------------------
Top 10 most frequent values:
 value  count  percent
     0  18897    32.81
  2010   2645     4.59
  2008   2613     4.54
  2009   2533     4.40
  2000   2091     3.63
  2007   1587     2.76
  2006   1471     2.55
  2003   1286     2.23
  2011   1256     2.18
  2004   1123     1.95
```