Milestone 1: Project Proposal and Data Selection/Preparation

Step 1: Preparing for Your Proposal

Client selection and dataset

The client is Tanzania Ministry of Water. They need us to analyse the water data and predict which water pumps are faulty.

The dataset is taken from Taarifa which aggregates data from the Tanzania Ministry of Water.

Steps taken to get data

The dataset is imported from Taarifa databases. Due to messy data, the data has to be properly arranged in Excel table format.

Proper identification of features and label are done and arranged before analysis.

These are the data we are looking at:

In [1]:
```
import pandas as pd
```

In [2]:
```
df = pd.read_csv("train.csv", low_memory=False)
```

In [3]:
```
df
```

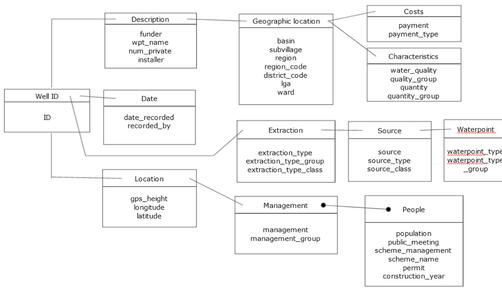
Out[3]:

	id	amount_tsh	date_recorded	funder	gps_height	installer	longitude	latitude	wpt_name	num_private	...	water_qua
0	69572	6000.0	14/3/2011	Roman	1390	Roman	34.938093	-9.856322	none	0	...	
1	8776	0.0	6/3/2013	Grumeti	1399	GRUMETI	34.698766	-2.147466	Zahanati	0	...	
2	34310	25.0	25/2/2013	Lottery Club	686	World vision	37.460664	-3.821329	Kwa Mahundi	0	...	
3	67743	0.0	28/1/2013	Unicef	263	UNICEF	38.486161	-11.155298	Zahanati Ya Nanyumbu	0	...	
4	19728	0.0	13/7/2011	Action In A	0	Artisan	31.130847	-1.825359	Shuleni	0	...	
...	...	...	...	...	...	...	...	...	...	...	...	
59395	60739	10.0	3/5/2013	Germany Republi	1210	CES	37.169807	-3.253847	Area Three Namba 27	0	...	
59396	27263	4700.0	7/5/2011	Cefa-njombe	1212	Cefa	35.249991	-9.070629	Kwa Yahona Kuvala	0	...	
59397	37057	0.0	11/4/2011	NaN	0	NaN	34.017087	-8.750434	Mashine	0	...	fluo
59398	31282	0.0	8/3/2011	Malec	0	Musa	35.861315	-6.378573	Mshoro	0	...	
59399	26348	0.0	23/3/2011	World Bank	191	World	38.104048	-6.747464	Kwa Mzee Lugawa	0	...	s

59400 rows × 41 columns

We are interested in finding status_group which is the water pump condition:

Entity Relationship Diagram



Step 2: Develop Project Proposal

Description of Project

The Tanzania Water Ministry (The Client) would like us to analyse and report findings on all water pumps conditions throughout the country. They also would like a prediction model to be developed based on data recorded. A smart understanding of which waterpoints will fail can improve maintenance operations and ensure that clean, potable water is available to communities across Tanzania.

Water is a precious resource and a daily need for all people living in Tanzania. A water pump disruption that occurs has immediate impact in using water for cooking, washing, bathing, business usage like in Food and Beverage industry. The Water Ministry views these disruptions very seriously and wants action to solve them.

But to solve them, they need to know the insights from the data they have been collected.

Questions that you want to answer with the data:

1. Are there any locations which has more breakdowns?
2. Does water management issues affect the pump operations?
3. How about water extraction, source and water points affect the pumps?

Hypothesis assumptions

1. Locations like Basin, Subvillage and Region will provide pump statuses
2. Populations density throughout Tanzania are equally distributed
3. Water source, types, quantities, quality are consistent in all water pumps

Approach to take in order to prove (or disprove) hypotheses

These are the attributes of the dataset:

Field	Description
amount_tsh	Total static head (amount water available to waterpoint)
date_recorded	The date the row was entered
funder	Who funded the well
gps_height	Altitude of the well
installer	Organization that installed the well
longitude	GPS coordinate
latitude	GPS coordinate
wpt_name	Name of the waterpoint if there is one
num_private	
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
quality_group	The quality of the water
quantity	The quantity of water
quantity_group	The quantity of water
source	The source of the water
source_type	The source of the water
source_class	The source of the water
waterpoint_type	The kind of waterpoint
waterpoint_type_group	The kind of waterpoint
functional	the waterpoint is operational and there are no repairs needed
non functional	the waterpoint is not operational

1. Location analysis will use basin, subvillage, region, region_code and district_code features. Extra info can be gleaned from gps_height, longitude and latitude.

1. Population will use population and public_meeting columns.

1. source, source_type, source_class, waterpoint_type, waterpoint_type_group columns will be explored to see any connection to pumps

We would also keen to explore any relationships for management impacts and water payments relates to water pump operations.

Metrics evaluation will be used are accuracy, precision, recall and F1 scores since this is binary problem.