Jal Jeevan Mission Water Quality
Data Preprocessing

July 26, 2024

Need for Preprocessing

The water quality data from 2009 to 2019 is collected from the NRDWP website: Format E3- Habitation wise laboratory testing Reports, and data from 2021 to 2024 is collected from the JJM website in Format L2: JJM Lab Testing Data. This difference in Water Quality Data data sources lead to issues with format consistency and irregular column naming structures, thereby reducing the efficiency of the data analysis process.

To address these challenges, following steps are done.

- **Created two new columns**: 'Contaminants' and 'Exceeded Contaminants'.

  For NRDWP data the 'Contaminants' and 'Exceeded Contaminants' column was populated with data from : `['AbovePMandatory', 'BelowPMandatory', 'AbovePEmerging', 'BelowPEmerging']`.

  For JJM data the 'Contaminants' and 'Exceeded Contaminants' column was populated with data from : `['Tested chemical parameters', 'Tested bacteriological parameters']`.

- **Organized data in the new columns**:

  The 'Contaminants' column included all the tested pollutants, while the 'Exceeded Contaminants' column contained only those pollutants that exceeded the permissible limit. The structure of pollutant names and quantities was kept uniform across all data for consistency.

- **Mapped pollutant** names to standard names to ensure uniformity and avoid spelling issues across the dataset.

- **Implemented a uniform column naming structure**:

  The NRDWP naming convention follows the format `StateName, DistrictName, BlockName,`
  `PanchayatName, VillageName, HabitationName`. In contrast, the naming convention in JJM files is `State name, District name, Block name, Gram panchayat, Village name`. The NRDWP naming convention was applied to JJM data to maintain uniformity.

Mapping used to Standarise all the Pollutant names

Standard Name	**Variants**
Alkalinity	Total Alkalinity (as Calcium Carbonate) (mg/l)
Aluminium	Aluminium, Aluminum (As Al) (mg/l)
Arsenic	Arsenic, Total Arsenic (As As) (mg/l)
Cadmium	Cadmium, Cadmium (As Cd) (mg/l)
Calcium	Calcium, Calcium (as Ca) (mg/l)
Chloride	Chloride, Chloride (as Cl) (mg/l)
Chromium	Chromium, Total Chromium (As Cr) (mg/l)
Copper	Copper, Copper (As Cu) (mg/l)
E.Coli	E. coli (CFU/100 ml), E.coli (MPN/100 ml)
Fluoride	Fluoride, Fluoride (as F) (mg/l)
Hardness	Total Hardness (As CaCO3) (mg/l)
Iron	Iron, Iron (As Fe) (mg/l)
Lead	Lead, Lead (as Pb) (mg/l)
Magnesium	Magnesium, Magnesium (As Mg) (mg/l)
Manganese	Manganese, Manganese (As Mn) (mg/l)
Mercury	Mercury, Mercury (As Hg) (mg/l)
Nickel	Nickel, Nickel (As Ni) (mg/l)
Nitrate	Nitrate, Nitrate (as NO3) (mg/l)
Selenium	Selenium, Selenium (As Se) (mg/l)

Standard Name	**Variants**
Sulphates	Sulphates, Sulphate (as SO4) (mg/l)
TDS	TDS, Total dissolved solids (mg/l), TDS (mg/l)
Total Coliform	Total coliform (CFU/100 ml), Total Coliform
Zinc	Zinc, Zinc (As Zn) (mg/l)
pH	pH, pH (NA)
Turbidity	Turbidity, Turbidity (NTU)
Acidity	Acidity, Acidity (NA)
Ammonia	Ammonia, Ammonia (as Total Ammonia- N) (mg/l)
Bacteriological	Bacteriological (H2S test) (NA)
Barium	Barium, Barium (as Ba) (mg/l)
Bicarbonate	Bicarbonate, Bicarbonate (mg/l)
Boron	Boron, Boron (as B) (mg/l)
Butachlor	Butachlor, Butachlor (µg/l)
Carbonate	Carbonate, Carbonate (mg/l)
Chloramine	Chloramine, Chloramine (mg/l)
Colour	Colour, Colour (Hazen units)
Ehion	Ehion, Ehion (µg/l)
Endosuplhan	Endosuplhan, Endosuplhan (µg/l)
Malathion	Malathion, Malathion (µg/l)
Molybdenum	Molybdenum, Molybdenum (As Mo) (mg/l)
Monocrotophos	Monocrotophos, Monocrotophos (µg/l)

Standard Name	**Variants**
Nitrite	Nitrite, Nitrite (mg/l)
Odour	Odour, Odour (NA)
Phorate	Phorate, Phorate (µg/l)
Phosphate	Phosphate, Phosphate (mg/l)
Potassium	Potassium, Potassium (mg/l)
Silicate	Silicate, Silicate (mg/l)
Silver	Silver, Silver (as Ag) (mg/l)
Sodium	Sodium, Sodium (mg/l)
Suspended Solids	Suspended Solids, Suspended Solids (mg/l)
Taste	Taste, Taste (NA)
Temperature	Temperature, Temperature (NA)
Total Solids	Total Solids, Total Solid (mg/l)
Toxicity	Toxicity, Toxicity (NA)

Permissible limits for All pollutants

Standard Name	P Limit
Alkalinity	600
Aluminium	0.2
Ammonia	0.5
Arsenic	0.01
Barium	0.7
Cadmium	0.003
Calcium	200
Chloride	1000
Chloride	1000
Colour	15
Copper	1.5
Cyanide	0.05
E.Coli	0
Fluoride	1.5
Hardness	600
Iron	1
Lead	0.01
Magnesium	100

Standard Name	P Limit
Malathion	190
Manganese	0.3
Mercury	0.001
Mineral oil	0.5
Molybdenum	0.07
Monocrotophos	1
Nickel	0.02
Nitrate	45
Phorate	2
pH	8.5
Selenium	0.01
Silver	0.1
Sulphates	400
Sulphide	0.05
TDS	2000
Turbidity	5
Uranium	0.03
Zinc	15