
Data Sources

Primary Dataset: Kaggle Water Quality dataset found here:
https://www.kaggle.com/adityakadiwal/water-potability

Secondary Dataset (for cross-referencing and possibly filling in missing data):
Environmental Protection Agency Water Quality Data found here:
https://www.epa.gov/waterdata/water-quality-data-download#portal

---

Data Dictionary

Attribute Name	Attribute Type	Max Field Size	Constraint
id	Int	255	Primary Key
ph	Float	64	
Hardness	Float	64	Not null
Solids	Float	64	Not null
Chloramines	Float	64	Not null
Sulfate	Float	64	
Conductivity	Float	64	Not null
Organic_carbon	Float	64	Not null
Trihalomethanes	Float	64	
Turbidity	Float	64	Not null
Potability	Bool	1	Not null

Data Point Description

Every row in the Water Quality Datasets has the attributes shown above, with the exception of some rows having missing data as not all attributes are required. Below is a description of each attribute along with the preferred range of values for drinking water

as recommended by the World Health Organization (WHO). Note, however, that the observed range of values (the possible values) is not always equivalent to the recommended range.

1. **Id**: The water body key (not a reference to geologic coordinates, just a series of unique and incrementing values to distinguish the water bodies). Ranges from 1 to 3276 for these datasets, as the physical and chemical characteristics of 3276 different water bodies were measured.
2. **ph**: Acid-base balance of water. The WHO has designated a permissible range of pH from 6.5 to 8.5.
3. **Hardness**: Capacity of water to precipitate soap in mg/L. The WHO has recommended a hardness no higher than 180 mg/L.
4. **Solids**: Total dissolved solids in ppm. The desired limit for TDS is 500 ppm and the maximum limit is 1000 ppm.
5. **Chloramines**: Amount of Chloramines in ppm. Chlorine levels up to 4 ppm are considered safe in drinking water.
6. **Sulfate**: Amount of Sulfates dissolved in mg/L. Sulfate in drinking water has a maximum contaminant level of 250 mg/L, though this is mostly for aesthetic effects like taste and odor.
7. **Conductivity**: Electrical conductivity of water in µS/cm. According to WHO standards, EC value should not exceed 400 µS/cm.
8. **Organic_carbon**: Amount of organic carbon in ppm. The desired limit for OC in drinking water is 4 mg/L.
9. **Trihalomethanes**: Amount of Trihalomethanes in µg/L. THM levels up to 80 ppm or 80000 µg/L is considered safe in drinking water.
10. **Turbidity**: Measure of light emitting property of water in NTU. The WHO recommended turbidity value is 5.00 NTU.
11. **Potability**: Indicates if water is safe for human consumption. Potable = 1 and Not potable = 0.