
Data Sources

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

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

---

Data Point Description

Each row from the Kaggle dataset above represents a water resource. While the dataset is synthetic, the attribute values of each water resource are derived from random bodies of water in the EPA water quality dataset at one point in time; thus, the values are consistent with those of real water resources. One attribute, however, was added to the dataset: the binary is_safe attribute that describes the potability of the water. While not present in any EPA water quality datasets, the values of this attribute are accurate in that they were based on whether water boiling or distillation measures were in place at the time of sample collection, where an active boiling or distillation measure would indicate that the water resource was not potable while no orders would indicate that the water was. A summary of this attribute, along with all the other attributes of the dataset, can be found in the data dictionary on the following page. One more aspect to note about the dataset is that because it is synthetic, nearly all values are present; however, data cleansing will still be necessary to some degree.

Data Dictionary

Attribute Name	Attribute Type	Max Field Size	Description
aluminium	Float	64	Concentration of aluminum (ppm)
ammonia	Float	64	Concentration of ammonium (ppm)
arsenic	Float	64	Concentration of arsenic (ppm)
barium	Float	64	Concentration of barium (ppm)
cadmium	Float	64	Concentration of cadmium (ppm)
chloramine	Float	64	Concentration of chloramine (ppm)
chromium	Float	64	Concentration of chromium (ppm)
copper	Float	64	Concentration of copper (ppm)
flouride	Float	64	Concentration of fluoride (ppm)
bacteria	Float	64	Count of bacteria colonies (CFU)
viruses	Float	64	Count of virus cultures (CFU)
lead	Float	64	Concentration of lead (ppm)
nitrates	Float	64	Concentration of nitrate (ppm)
nitrites	Float	64	Concentration of nitrite (ppm)
mercury	Float	64	Concentration of mercury (ppm)
perchlorate	Float	64	Concentration of perchlorate (ppm)
radium	Float	64	Concentration of radium (ppm)
selenium	Float	64	Concentration of selenium (ppm)
silver	Float	64	Concentration of silver (ppm)
uranium	Float	64	Concentration of uranium (ppm)
is_safe	Int	255	Nonpotable {0} or potable {1}