TOPIC

WATER QUALITY ANALYSIS

PHASE 3

- **Data Collection:** First, gather the relevant water quality data. This may include parameters like pH, turbidity, temperature, dissolved oxygen, etc.,
- **Data Cleaning:** Remove duplicates, handle missing values, and correct any obvious errors.
- **Data Transformation:** Convert data types, standardize units if necessary. Outlier Detection and Handling: Identify and deal with outliers that might affect the analysis.
- **Exploratory Data Analysis (EDA):**
    1. Descriptive Statistics: Calculate basic statistics such as mean, median, standard deviation, and quartiles for each parameter.
    2. Data Visualization: Create visualizations like histograms, box plots, and scatter plots to explore the distribution of data and relationships between variables.
    3. Correlation Analysis: Determine the correlations between different water quality parameters.

Data Collection: Obtain the water quality dataset from a reliable source such as government agencies, research organizations, or data repositories.

- ✓ **Load Data:** Import your dataset into a tool like Python (using libraries like Pandas).
- ✓ **Summary Statistics:** Compute basic statistics like mean, median, and standard deviation for numerical parameters.
- ✓ **Data Visualization:** Create various plots to visualize data distributions. For numerical parameters:

Importing libraries:

```
import pandas as pd

import numpy as np

import matplotlib.pyplot as plt

import seaborn as sns
```

Load the Data: Use a data analysis tool such as Python with libraries like Pandas to load the dataset into your environment.

main_dat = pd.read_csv("/kaggle/input/water-potability/water_potability.csv")

ks = main_dat.copy() #copy of original data set

Explore the Data: Get familiar with the dataset by examining its structure, columns, and basic statistics.

ks.head()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890456	20791.31898	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.05786	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.54173	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.41744	8.059332	356.886136	363.266516	18.436525	100.341674	4.628771	0
4	9.092223	181.101509	17978.98634	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

ks.sample()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
659	5.555353	154.300684	20503.430050	9.644997	313.470297	355.206969	18.468690	75.140362	4.536146	0
2272	8.384296	223.328185	27463.654790	6.476753	352.952803	318.042648	10.645164	64.209337	3.460998	0
2006	6.538207	214.992866	12330.406570	7.300092	389.817036	465.352665	22.089402	24.532773	3.426266	1
801	8.900865	211.306812	9592.151333	8.863272	348.437820	333.775327	18.267951	68.333170	4.518751	1
2886	NaN	206.036295	8667.720239	6.329952	353.529381	599.546019	21.118938	55.932324	4.128746	0

Ks. Shape()

```
(3276, 10)
```

Ks.columns()

```
Index(['ph', 'Hardness', 'Solids', 'Chloramines', 'Sulfate', 'Conductivity',
       'Organic_carbon', 'Trihalomethanes', 'Turbidity', 'Potability'],
      dtype='object')
```

Checking Null values:

pd.isnull(ks).sum()

```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

ks.dropna(inplace=True)

pd.isnull(ks).sum()

```
ph                0
Hardness          0
Solids            0
Chloramines       0
Sulfate           0
Conductivity      0
Organic_carbon    0
Trihalomethanes   0
Turbidity         0
Potability        0
dtype: int64
```

ks.describe()

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000
mean	7.085990	195.968072	21917.441375	7.134338	333.224672	426.526409	14.357709	66.400859	3.969729	0.403282
std	1.573337	32.635085	8642.239815	1.584820	41.205172	80.712572	3.324959	16.077109	0.780346	0.490678
min	0.227499	73.492234	320.942611	1.390871	129.000000	201.619737	2.200000	8.577013	1.450000	0.000000
25%	6.089723	176.744938	15615.665390	6.138895	307.632511	366.680307	12.124105	55.952664	3.442915	0.000000
50%	7.027297	197.191839	20933.512750	7.143907	332.232177	423.455906	14.322019	66.542198	3.968177	0.000000
75%	8.052969	216.441070	27182.587065	8.109726	359.330555	482.373169	16.683049	77.291925	4.514175	1.000000
max	14.000000	317.338124	56488.672410	13.127000	481.030642	753.342620	27.006707	124.000000	6.494749	1.000000

ks.nunique()

```
ph                2011
Hardness          2011
Solids            2011
Chloramines       2011
Sulfate           2011
Conductivity      2011
Organic_carbon    2011
Trihalomethanes   2011
Turbidity         2011
Potability           2
dtype: int64
```

ks.info()

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2011 entries, 3 to 3271
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2011 non-null   float64
 1   Hardness         2011 non-null   float64
 2   Solids           2011 non-null   float64
 3   Chloramines      2011 non-null   float64
 4   Sulfate          2011 non-null   float64
 5   Conductivity     2011 non-null   float64
 6   Organic_carbon   2011 non-null   float64
 7   Trihalomethanes  2011 non-null   float64
 8   Turbidity        2011 non-null   float64
 9   Potability       2011 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 172.8 KB
```

Handle Missing Values:

Identify missing values in the dataset. Decide how to handle missing data, which may include:

Imputation: Filling missing values with a mean, median, or mode of the column.

Deletion: Removing rows or columns with a high percentage of missing values.

Advanced techniques like regression imputation or machine learning-based imputation.

Outlier Detection and Handling:

Identify outliers using statistical methods or visualization techniques. Decide on a strategy to handle outliers, such as: Removing outliers if they are errors or anomalies. Transforming or Winsorizing the data to reduce the impact of outliers. Using robust statistical methods for analysis that are less sensitive to outliers.

- Data Scaling and Normalization: Depending on your analysis, you may need to scale or normalize the data to ensure that variables are on a similar scale.
- Feature Engineering: Create new features or derive relevant information from the existing data if needed for your analysis.
- Data Splitting: If you plan to build predictive models, split the data into training and testing sets.
- Documentation: Keep detailed records of the preprocessing steps and the rationale behind them.

Histograms to understand data distribution.

Box plots to identify outliers. Scatter plots to see relationships between parameters.

- Correlation Analysis: Calculate and visualize correlations between numerical parameters using tools like a correlation matrix or scatter plots.
- Missing Data Analysis: Identify missing values in the dataset and decide how to handle them (imputation or removal).
- Categorical Parameters: For categorical parameters, create bar plots to visualize their distribution.
- Outlier Detection: Identify and handle outliers using statistical methods or visualization.
- Data Transformation: If necessary, apply transformations like log, square root, or normalization to make the data more suitable for analysis.
- Feature Engineering: Create new features or derive insights from existing ones to help with your analysis.
- Deviation from Standards: Determine what standards or thresholds you want to apply for different parameters and analyze how your data deviates from these standards.

1.pH value:

PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.

2. Hardness:

Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.

3. Solids (Total dissolved solids - TDS):

Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. Chloramines:

Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. Sulfate:

Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. Conductivity:

Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 μS/cm.

7. Organic carbon:

Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. Trihalomethanes:

THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. Turbidity:

The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. Portability:

Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

CONCLUSION:

In this phase I've loaded the dataset to my python program and preprocessed the dataset by removing the null values, filling missing values by mean, etc., This cleaned dataset can now be used for further data exploration, machine learning, analysis and interpretation.