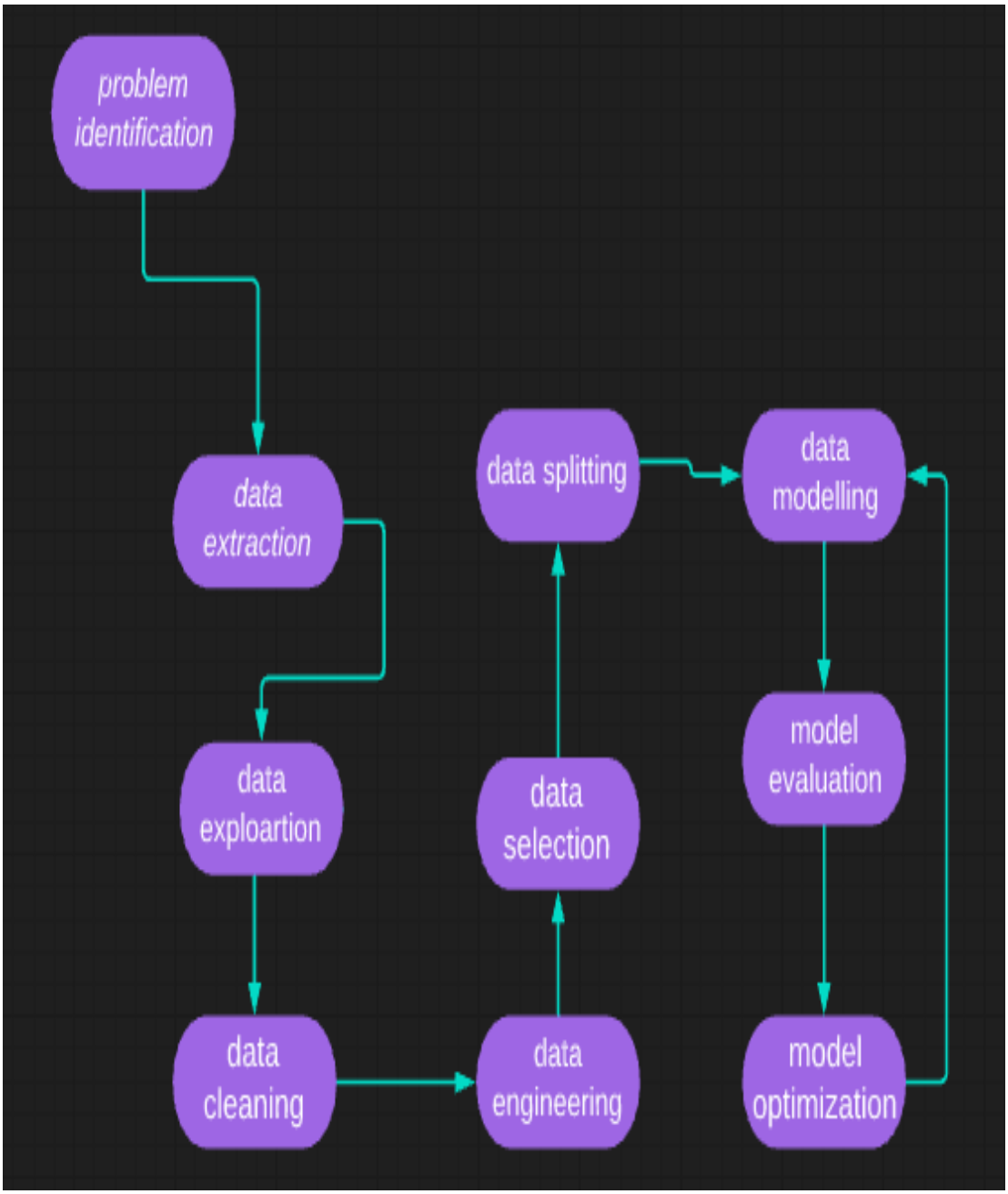
Water quality analysis with python programming

BATCH MEMBER 812121104020:J Jaasim



Introduction: Water Quality Analysis

- ➢ Analysing water quality is one of the key topics of machine learning research.
- ➢ In order to train a machine learning model that can determine if a certain water sample is safe or unsafe for eating, we must first understand all the parameters that impact water potability.
- ➢ This process is also known as water potability analysis.
- ➢ We'll be utilising a Kaggle dataset that includes information on all of the key elements that have an impact on the potability of water for the water quality analysis challenge.
- ➢ Before building a model using machine learning to predict whether the water specimen is acceptable or unsafe for eating.
- ➢ we must first quickly examine each characteristic of this dataset because all of the elements that determine water quality are crucial.

About dataset

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. Potability:

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

DATASET

# ph	# Hardness	# Solids	# Chloramines	# Sulfate	# Conductivi...	# Organic_c...
8.7572573974489 91	208.19148044285 727	21536.224687445 414	4.9151818545431 86	317.88250849783 786	404.71779915644 53	13.768323386423 37
	168.38843877429 533	27492.387306587 81	7.0462247976964 13	299.82847791882 16	383.79501999832 584	16.182066493674 37
7.8896318980194 1	188.45761589158 31	12013.550628764 531	5.2123146028653 52	247.20802604764 31	605.22012435080 56	9.6113487400113 38
6.6524888988557 87	145.81817191988 34	19871.788448385 862	4.9618663801915 02	288.85219173685 15	545.97499376248 3	10.942824256813 132
9.1471976553369 21	211.71414177764 134	11928.618835646 286	7.2387947693374	339.75191888079 38	527.70891001384 07	18.275312189276 65
10.568744636218 196	181.89336556155 354	21783.651833363 374	6.9912599962381 34	348.39837835517 297	456.55648212234 38	16.482835366963 91
7.4842548377274 66	260.89217257337 195	38616.615148853 696	9.3791335405876 24		404.67877732816 267	15.934408387128 878
8.5208065729703 83	238.33511245550 966	28779.650011033 6	8.2828864645096 09	381.64932287189 46	481.31880268961 31	6.0163366092713 67
4.9994138107969 19	198.28705014936 828	24323.865903845 946	7.2381641268628 36	324.89380784225 57	405.33848282759 83	8.2365575831053 83

Here's a list of tools and software commonly used in the process:

Water quality analysis is essential for monitoring and maintaining the safety and health of water sources. Various tools and software are commonly used in this field to collect, analyze, and interpret water quality data. Here is a list of tools and software commonly used in water quality analysis:

1. Water Quality Testing Kits:

- pH meters: Measure the acidity or alkalinity of water.
- Conductivity meters: Determine the water's electrical conductivity, which is related to ion concentration.
- Turbidity meters: Assess water clarity and the presence of suspended particles.
- Dissolved oxygen (DO) meters: Measure the concentration of oxygen dissolved in water, vital for aquatic life.
- Total dissolved solids (TDS) meters: Quantify the concentration of dissolved solids in water.
- Colorimeters: Measure specific chemical parameters by analyzing color changes in water samples.
- Test strips: Quick and simple tests for specific parameters like chlorine, hardness, and nitrate.

2. Water Sampling Equipment:

- Water samplers: Collect representative water samples from various depths and locations.
- Bailer and bailer pumps: Used for collecting groundwater samples from wells.
- Sediment samplers: Gather samples from the sediment at the bottom of water bodies.

3. Laboratory Analysis Software:

- Environmental Data Management Systems (EDMS): Manage, store, and analyze water quality data.
- LIMS (Laboratory Information Management System): Software for managing laboratory workflows and data.
- GIS (Geographic Information System) software: Used to map and analyze spatial water quality data.
- Statistical analysis software (e.g., R, MATLAB, or SPSS): Perform statistical tests on water quality data.
- Database management systems (e.g., SQL, Microsoft Access): Store and retrieve water quality data.

4. Water Quality Analysis Software:

- Aquatic modeling software (e.g., AQUATOX, CE-QUAL-W2): Predict and model water quality changes.
- HSPF (Hydrological Simulation Program - Fortran): Simulate hydrologic and water quality processes in watersheds.
- SWMM (Storm Water Management Model): Model stormwater runoff and quality in urban areas.
- Water quality modeling and simulation software (e.g., QUAL2K, WASP): Assess and predict water quality in rivers and streams.

5. Remote Sensing and Monitoring Tools:

- Satellite imagery and remote sensors: Collect data on water quality parameters from space.
- Water quality monitoring buoys and probes: Collect real-time data from water bodies.
- Telemetry systems: Transmit data from remote monitoring stations to a central location.

6. Data Visualization and Analysis Tools:

- Spreadsheet software (e.g., Microsoft Excel): Organize and visualize data using charts and graphs.
- Data visualization tools (e.g., Tableau, Power BI): Create interactive visualizations of water quality data.
- Geographic Information Systems (GIS) software: Visualize spatial water quality data through maps and layers.

7. Field Data Collection Apps:

- Mobile apps for data collection and analysis, such as EQuIS, AquaCrop, and Water Quality Field App.

These tools and software are used by environmental scientists, water resource managers, and other professionals to assess and maintain the quality of water in various environments, from rivers and lakes to groundwater and coastal areas.

DESIGN THINKING AND PRESENT IN FORM OF DOCUMENT

Design thinking is a problem-solving approach that focuses on understanding the needs of the users and developing innovative solutions to meet those needs. When applied to water quality analysis, it can lead to more effective and user-centric methods for monitoring and managing water quality. Here's a design thinking approach to water quality analysis presented in a simplified form:

1. Empathize:

- Understand the stakeholders: Identify the various stakeholders involved in water quality analysis, including scientists, government agencies, environmentalists, and the general public.
- Listen to their needs and concerns: Conduct interviews, surveys, and workshops to gather insights into what users require from water quality analysis.
- Observe the current methods and tools: Examine existing practices for collecting and analyzing water quality data.

2. Define:

- Define the problem: Clearly articulate the challenges and limitations in existing water quality analysis methods. This could include issues like data accuracy, accessibility, or timeliness.
- Create user personas: Develop profiles of different user groups, highlighting their specific needs and pain points.

3. Ideate:

- Brainstorm solutions: Organize collaborative sessions to generate creative ideas for improving water quality analysis.
- Encourage diverse perspectives: Invite participants from various backgrounds to bring fresh insights to the ideation process.
- Use design thinking tools like mind maps, brainstorming sessions, and ideation workshops to generate a wide range of potential solutions.

4. Prototype:

- Build prototypes: Create mock-ups or prototypes of the proposed solutions, whether they are new testing devices, data visualization tools, or monitoring systems.
- Test and iterate: Gather feedback from users and refine the prototypes based on their input. Iteration is a key part of the design thinking process.

5. Test:

- Conduct user testing: Invite users to interact with the prototypes and provide feedback on their usability and effectiveness.
- Evaluate the solutions: Assess the prototypes against predefined criteria, considering factors like accuracy, accessibility, cost-effectiveness, and ease of use.

6. Implement:

- Develop a comprehensive plan for implementing the chosen solution.
- Collaborate with stakeholders to ensure a smooth transition to the new water quality analysis approach.

7. Monitor and Improve:

- Continuously monitor the implemented solution to identify areas for improvement.
- Gather feedback from users and make iterative adjustments as necessary to optimize the system.

This design thinking approach to water quality analysis is a user-centered and iterative process that aims to address the specific needs and challenges of various stakeholders involved in water quality monitoring and management. By focusing on empathy, defining problems, ideation, prototyping, testing, and ongoing improvement, the process can lead to more effective and user-friendly solutions for maintaining and protecting water quality.

Design thinking is a problem-solving approach that focuses on understanding the needs of the users and developing innovative solutions to meet those needs. When applied to water quality analysis, it can lead to more effective and user-centric methods for monitoring and managing water quality.

DESIGN INTO INNOVATION

Design thinking can be a powerful approach to drive innovation in water quality analysis. By applying the principles of design thinking, you can develop innovative solutions that address the specific needs and challenges of water quality analysis. Here's how to apply design thinking to foster innovation in this field:

1. Empathize:
   - Understand the stakeholders: Begin by empathizing with all the stakeholders involved in water quality analysis, including scientists, environmentalists, regulatory agencies, and the public. Gain insights into their needs and concerns.
   - User research: Conduct in-depth interviews, surveys, and observations to gather data on how users interact with current water quality analysis methods and what issues they encounter.
   - Field visits: Visit water testing sites, laboratories, and monitoring stations to see the challenges faced by professionals on the ground.

2. Define:
   - Define the innovation challenge: Clearly articulate the specific challenges and limitations in the current water quality analysis methods. Understand the pain points and opportunities for improvement.
   - Identify user pain points: Create user personas and journey maps to visualize the user experience and highlight areas where innovation is most needed.

3. Ideate:
   - Brainstorm for innovation: Organize creative brainstorming sessions with cross-functional teams to generate a wide range of innovative ideas for water quality analysis.
   - Encourage interdisciplinary collaboration: Bring together experts from various fields such as technology, data science, environmental science, and design to explore novel approaches.

- Think beyond existing solutions: Challenge conventional thinking and explore unconventional ideas that could revolutionize water quality analysis.

4. Prototype:
- Create prototypes: Develop prototypes or mock-ups of the most promising innovative solutions. These could be new testing devices, data visualization tools, real-time monitoring systems, or data analysis algorithms.
- Low-cost experimentation: Use rapid prototyping and low-cost materials to test ideas quickly and efficiently.
- User feedback: Involve users in the testing of prototypes to ensure their practicality and user-friendliness.

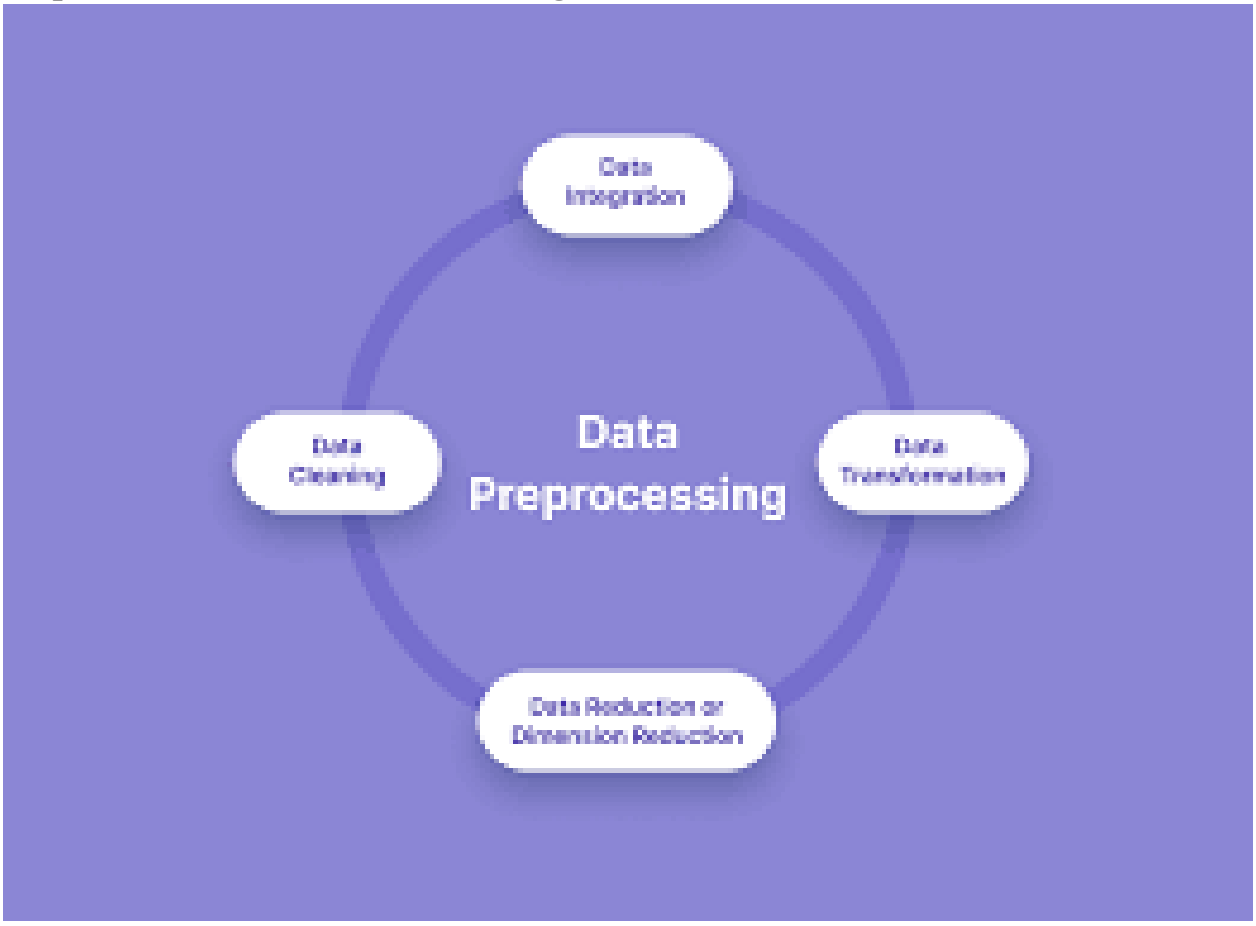
5. Test:
- Gather user feedback: Conduct usability testing and gather feedback from users and stakeholders on the prototypes.
- Iterate based on feedback: Make iterative improvements to the prototypes based on the feedback received during testing.

6. Implement:
- Develop an implementation plan: Once a refined innovative solution is identified, create a comprehensive plan for implementing it within the water quality analysis process.
- Collaboration and integration: Collaborate with relevant organizations and agencies to ensure a smooth integration of the innovative solution into existing water quality monitoring systems.

7. Monitor and Improve:
- Continuously monitor the implemented innovation to assess its performance and effectiveness.

- Gather ongoing feedback from users and stakeholders to make further improvements and refinements as needed.

By applying design thinking to water quality analysis, you can encourage innovative thinking and develop solutions that are not only technically advanced but also user-centric and more effective in addressing the challenges of water quality monitoring and management. This approach can lead to more accurate, efficient, and environmentally responsible methods for ensuring clean and safe water resources.



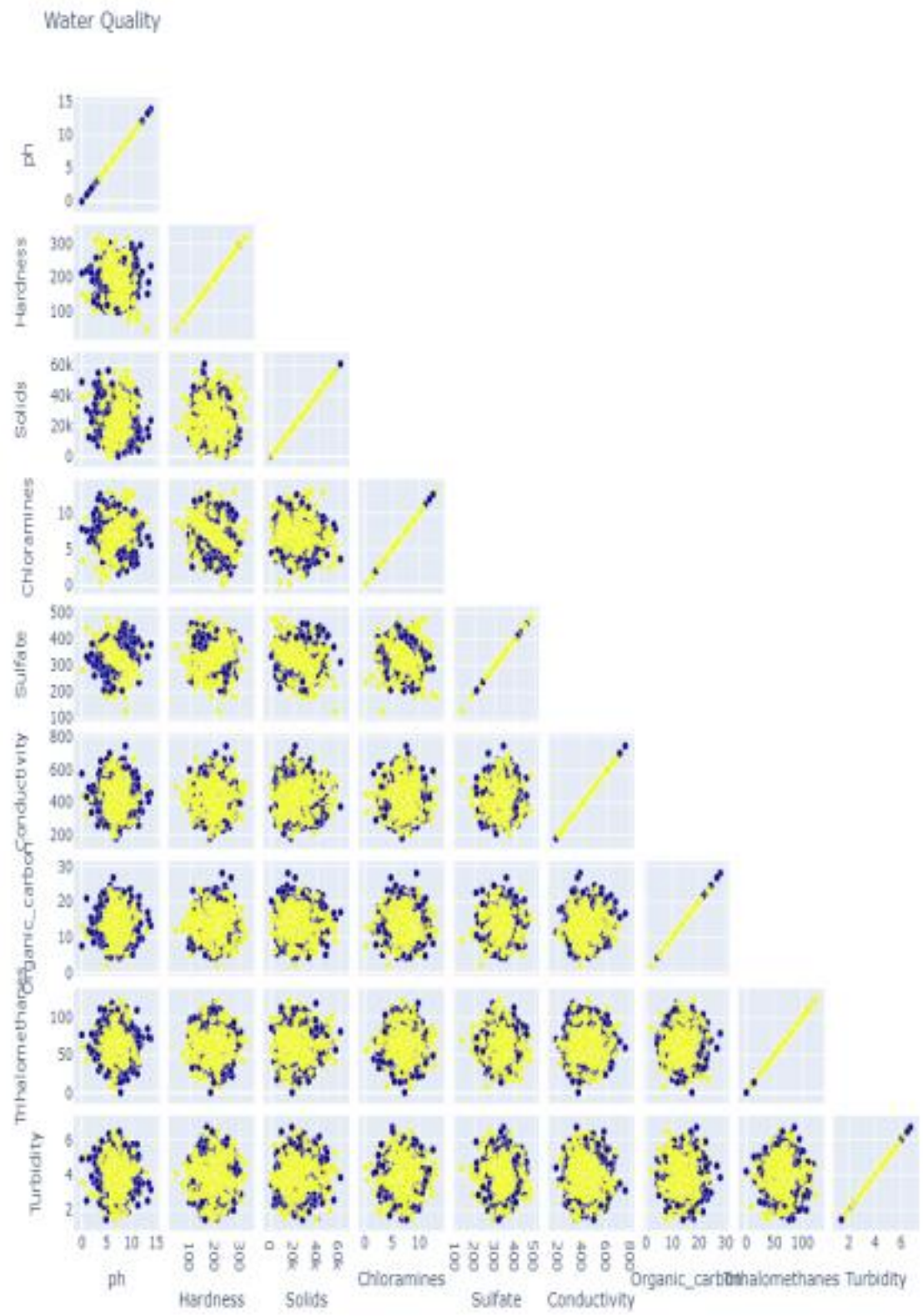
PYHON PROGRAM

Program:

```
import plotly.graph_objs as go
index_vals = data['Potability'].astype('category').cat.codes

fig = go.Figure(data=go.Splom(
dimensions=[dict(label='ph',
values=data['ph']),
dict(label='Hardness',
values=data['Hardness']),
dict(label='Solids',
values=data['Solids']),
dict(label='Chloramines',
values=data['Chloramines'])
dict(label='Sulfate',
values=data['Sulfate']),
dict(label='Conductivity',
values=data['Conductivity']),
dict(label='Organic_carbon',
values=data['Organic_carbon']),
dict(label='Trihalomethanes',
values=data['Trihalomethanes']),

dict(label='Turbidity',values=data['Turbidity'])],
showupperhalf=False,
text=data['Potability'],
marker=dict(color=index_vals,
showscale=False,
```

```
line_color='white', line_width=0.5)
))
fig.update_layout(
title='Water Quality',
width=1000,
height=1000,
)

fig.show().
```

OUTPUT


Water Quality
ph
Hardness
Solids
Chloramines
Sulfate
Conductivity
Organic_carbon
Trihalomethanes
Turbidity

Steps for Water Quality Analysis

MODULES

Selection of Parameters

The parameters of water quality are selected entirely according to the need for a specific use of that water. Some examples are:

Drinking: As per WHO/CPCB Standards

Irrigation:

pH Conductivity
Sodium & Potassium Nutrients
Specific compounds

Industries: As per specific requirement **Domestic Consumption:** As per BIS Standards **Water Bodies:** As per CPCB guidelines

1. **Selection of Methods**

The methods of water quality analysis are selected according to the requirement. The factors playing key role for the selection of methods are:

i. Volume and number of sample to be analyzed
ii. Cost of analysis
iii. Precision required
iv. Promptness of the analysis as required

Precision and Accuracy of Method Selected as per Requirement

What precision and accuracy to be maintained against a particular method is decided according to the objective of the monitoring. The factors influencing this decision includes:

- Budget of Monitoring System
- Parameters to be Monitored
- Use of the Water

Chain–of–Custody Procedures

Properly designed and executed chain-of-custody forms will ensure sample integrity from collection to data reporting. This includes the ability to trace possession and handling of the sample from the time of collection through analysis and final disposition. This process is referred to as "chain-of- custody" and is required to demonstrate sample control

when the data are to be used for regulation or litigation. Where litigation is not involved, chain-of-custody procedures are useful for routine control of samples.

A sample is considered to be under a person's custody if it is in the individual's physical possession, in the individual's sight, secured and tamper-proofed by that individual, or secured in an area restricted to authorized personnel. The following procedures summarize the major aspects of chain- of-custody:

v. **Sample Labels:** Labels are used to prevent sample misidentification as well as to identify the collector, if required. In other words, labeling ensures the responsibility and accountability of the collector.

vi. **Sample Seals:** Sample seals are used to detect unauthorized tampering with samples up to the time of analysis. So, it is essential to seal a sample before leaving the custody of the collector. Sealing must be done in such a way as one have to break the seal to access the sample.

vii. **Field Log Book:** All the useful information related to a field survey or sampling should be recorded in a Log Book. At least the following data should be in the log book:

a. Purpose of sampling
b. Location of sampling point
c. Name and address of field contact
d. Producer of material being sampled and address, if different from location
e. Type of sample
f. Method, date, and time of preservation.

viii. **Sample Analysis Request Sheet:** The sample analysis request sheet accompanies samples to the laboratory. The collector completes the field portion of such a form that includes most of the pertinent information noted in the log book. The laboratory portion of such a form is to be completed by laboratory personnel and includes: name of

person receiving the sample, laboratory sample number, date of sample receipt, condition of each sample (i.e., if it is cold or warm, whether the container is full or not, color, if more than one phase is present, etc.) and determinations to be performed.

ix. **Sample Delivery to the Laboratory:** Sample(s) should be delivered to laboratory as soon as possible after collection, typically within 2 days. Where shorter sample holding times are required, special arrangements must be made to insure timely delivery to the laboratory. Where samples are shipped by a commercial carrier, the waybill number to be included in the sample custody

documentation. Samples must be accompanied by a complete chain-of-custody record and a sample analysis request sheet.

x. **Receipt and Logging of Sample:** In the laboratory, the sample custodian inspects the condition and seal of the sample and reconciles label information and seal against the chain-of-custody record before the sample is accepted for analysis. After acceptance, the custodian assigns a laboratory number, logs sample in the laboratory log book and/or computerized laboratory information management system, and stores it in a secured storage room or cabinet or refrigerator at the specified temperature until it is assigned to an analyst.

xi. **Assignment of Sample for Analysis:** The laboratory supervisor usually assigns the sample for analysis. Once the sample is in the laboratory, the supervisor or analyst is responsible for its care and custody.

xii. **Disposal:** Samples are held for the prescribed amount and duration for the project or until the data have been reviewed and accepted. Samples are disposed usually after documentation. However, disposal must be in accordance with approved methods.

Proper Sampling

Proper sampling is a vital condition for correct measurement of water quality parameters. Even if advanced techniques and sophisticated tools are used, the parameters can give an incorrect image of the actual scenario due to improper sampling. The proper sampling should fulfill the following criteria:

xiii. **Representative:** The data must represent the wastewater or water body being sampled. So, the following factors must be well planned for proper sampling:

g. Process of Sampling
h. Sampling size/volume
i. Number of Sampling Locations

j. Number of Samples
k. Type of Samples
l. Time Intervals

During sampling, these factors must also be taken care of:

- Choosing of proper sampling container
- Avoiding contamination
- Ensure the personal safety of the collector

xiv. **Reproducible:** The data obtained must be reproducible by others following the same sampling and analytical protocols.

xv. **Defensible:** Documentation must be available to validate the sampling procedures. The data must have a known degree of accuracy and precision.

xvi. **Useful:** The data can be used to meet the objectives of the monitoring plan.

Proper Labeling

Proper labeling prevents sample misidentification and ensures the responsibility and accountability of the collector. The sample container should be labeled properly, preferably by attaching an appropriately inscribed tag or label. Alternatively, the bottle can be labeled directly with a water- proof marker. Barcode labels are also available nowadays.

Information on the sample container or the tag should include at least:

xvii. Sample code number (identifying location)
xviii. Date and time of sampling
xix. Source and type of sample
xx. Pre-treatment or preservation carried out on the sample

xxi. Any special notes for the analyst
xxii. Sampler's name

Preservation

Usually a delay occurs between the collection and analysis of a sample. The characteristics of the sample can be changed

during this period. Therefore proper preservation is required in the way to laboratory after collection, and in the laboratory upto when analysis starts.

Complete and unequivocal preservation of samples, whether domestic wastewater, industrial wastes, or natural waters, is a practical impossibility because complete stability for every constituent never can be achieved. At best, preservation techniques only retard chemical (especially, hydrolysis of constituents) and biological changes that inevitably continue after sample collection.

No single method of preservation is entirely satisfactory; the preservative is chosen with due regard to the determinations to be made. Preservation methods are limited to pH control, chemical addition, the use of amber and opaque bottles, refrigeration, filtration, and freezing.

Analysis

The samples, after reaching laboratory, are analyzed, according to the requisite parameters, following standard methods and protocols.

Reporting

The ultimate procedure of water analysis is to prepare a proper report against the submitted requisition. The report must be authenticated before

handing over the authority. All data should be kept in the laboratory log and preferably in laboratory database.

An alternative way to present the overall quality of water is to express it in the form of Water Quality Index (WQI). WQI is a concise numerical representation of overall water quality of a water body, which is convenient to interpret and used widely. WQI expresses the overall quality of water with a single digit, instead of many digits for all the WQP. Thus, it is readily conceivable for common people.

Complete and unequivocal preservation of samples, whether domestic wastewater, industrial wastes, or natural waters, is a practical impossibility because complete stability for every constituent never can be achieved. At best, preservation techniques only retard chemical (especially, hydrolysis of constituents) and biological changes that inevitably continue after sample collection.

No single method of preservation is entirely satisfactory; the preservative is chosen with due regard to the determinations to be made. Preservation methods are limited to pH control, chemical addition, the use of amber and opaque bottles, refrigeration, filtration, and freezing.

BUILD LOADING AND PREPROCESSING THE DATASET

Loading and preprocessing a dataset are critical steps in preparing data for analysis or machine learning tasks. These steps ensure that the data is

clean, structured, and ready for use. Below, I'll provide an outline of how to load and preprocess a dataset:

1. Loading the Dataset:

a. Data Source:

- Identify the source of your dataset. It can be a CSV file, Excel spreadsheet, SQL database, web API, or any other data repository.

b. Data Retrieval:

- Use appropriate libraries or tools to retrieve the data from the source. For example, you might use Python libraries like Pandas for CSV files, SQLAlchemy for databases, or requests for web APIs.

c. Data Inspection:

- Load a small portion of the dataset to inspect its structure. Use functions like `head()` or `sample()` to get an overview of the data.

2. Data Preprocessing:

a.Handling Missing Data:

- Identify and handle missing values. You can choose to remove rows with missing data, impute missing values, or use advanced techniques like interpolation.

b. Data Cleaning:

- Address any data anomalies, such as outliers or incorrect values. This may involve data transformations or removal of inconsistent data.

c. Data Conversion:

- Convert data types if needed. For example, convert date strings to datetime objects, or numerical values stored as strings to numeric data types.

d. Encoding Categorical Variables:

- Convert categorical variables into a numerical format using techniques like one-hot encoding or label encoding.

e. Feature Selection:

- If your dataset contains a large number of features, consider selecting relevant features to reduce dimensionality and improve model efficiency.

f. Scaling/Normalization:

- Scale or normalize the features if your machine learning algorithm relies on feature scaling. Common techniques include min-max scaling or z-score normalization.

g. Data Splitting:

- Split the dataset into training, validation, and test sets to evaluate model performance. Common splits are 70-30 or 80-20 for training and testing, respectively.

h. Data Balancing:

- If the dataset is imbalanced (e.g., in classification tasks), consider applying techniques like oversampling, undersampling, or using a combination of both to address class imbalance.

3. Feature Engineering:

a. Creating New Features:

- Generate new features based on domain knowledge or data insights. These features can enhance the predictive power of the model.

b. Dimension Reduction:

- Apply techniques like Principal Component Analysis (PCA) or t-SNE to reduce the dimensionality of the dataset while preserving essential information.

4. Data Visualization:

- Use data visualization libraries like Matplotlib, Seaborn, or Plotly to explore the dataset and gain insights. Visualizations can help identify trends, patterns, and potential issues in the data.

5. Save Processed Data:

- After preprocessing, save the cleaned and structured dataset to a file for future use. Common formats include CSV, Excel, or databases.

6. Documentation:

- Maintain clear and well-documented records of the preprocessing steps and transformations performed on the dataset. This documentation is essential for transparency and reproducibility.

7. Version Control:

- If working in a collaborative environment, consider using version control systems like Git to track changes to the dataset and preprocessing code.

Remember that the specific preprocessing steps will vary depending on the nature of the dataset and the analysis or machine learning task at hand. It's crucial to tailor the preprocessing steps to the characteristics of your data and the requirements of your project.

3.BUILD LOADING AND PREPROCESSING THE DATASET

1. Data Collection:
Obtain a dataset that contains information about houses and their corresponding prices. This dataset can be obtained from sources like real estate websites, government records, or other reliable data providers.
2. Load the Dataset:
Import relevant libraries, such as pandas for data manipulation and numpy for numerical operations. Load the dataset into a pandas DataFrame for easy data handling. You can use pd.read_csv() for CSV files or other appropriate functions for different file formats.

PROGRAM

```
import pandas as pd
import matplotlib.pyplot as plt

# Sample water quality data (pH and temperature)
data = {
```

```
    'pH': [7.2, 7.0, 6.8, 7.5, 7.2, 7.1, 7.3, 7.0, 6.9, 7.2],
    'Temperature (°C)': [25, 24, 26, 23, 24, 25, 25, 24, 26, 23]
}

# Create a Pandas DataFrame from the sample data
df = pd.DataFrame(data)

# Basic statistics
statistics = df.describe()

# Data visualization
plt.figure(figsize=(12, 5))
 # Histogram of pH data
plt.subplot(1, 2, 1)
plt.hist(df['pH'], bins=5, edgecolor='k')
plt.title('pH Data Histogram')
plt.xlabel('pH Value')
plt.ylabel('Frequency')
 # Box plot of temperature data
plt.subplot(1, 2, 2)
plt.boxplot(df['Temperature (°C)'])
plt.title('Temperature Data Box Plot')
 plt.tight_layout()
 # Display statistics and plots
print("Basic Statistics:")
print(statistics)
 plt.show()
```

OUTPUT

Basic Statistics:

	pH	Temperature (°C)
count	10.00000	10.000000

mean	7.15000	24.900000
std	0.17953	1.490712
min	6.80000	23.000000
25%	7.02500	23.750000
50%	7.15000	25.000000
75%	7.22500	25.750000
max	7.50000	26.000000

4.PERFORMING DIFFERENT ACTIVITIES LIKE

FEATURE ENGINEERING, MODEL TRAINING,

EVALUATION.

1. Feature Engineering:

As mentioned earlier, feature engineering is crucial. It involves creating new features or transforming existing ones to provide meaningful information for your model. Extracting information from textual descriptions (e.g., presence of keywords like "pool" or "granite countertops"). Calculating distances to key locations (e.g., schools, parks) if you have location data.

Data Preprocessing & Visualisation:

Continue data preprocessing by handling any remaining missing values or outliers based on insights from your data exploration.

DATASET

For this piece of analysis, the Water Quality dataset has been taken from Kaggle[1].

INPUT 1

```
import sys
print(sys.version) # displays the version of python installed.
```

Understanding the data

Firstly, we need to understand the data that we are working with. As the file format is a csv file, the standard pandas import statement using read_csv will be used.

INPUT2

```
# Import the dataset for review as a DataFrame
df = pd.read_csv("../input/water-potability/water_potability.csv")

# Review the first five observations
df.head()
```

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259	19909.541732	9.275884	NaN	418.606213	16.868637	66.420093	3.055934	0
3	8.316766	214.373394	22018.417441	8.059332	356.886136	363.266516	18.436524	100.341674	4.628771	0
4	9.092223	181.101509	17978.986339	6.546600	310.135738	398.410813	11.558279	31.997993	4.075075	0

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3276 entries, 0 to 3275
Data columns (total 10 columns):
 #   Column           Non-Null Count  Dtype
---  ------           --------------  -----
 0   ph               2785 non-null   float64
 1   Hardness         3276 non-null   float64
 2   Solids           3276 non-null   float64
 3   Chloramines      3276 non-null   float64
 4   Sulfate          2495 non-null   float64
 5   Conductivity     3276 non-null   float64
 6   Organic_carbon   3276 non-null   float64
 7   Trihalomethanes  3114 non-null   float64
 8   Turbidity        3276 non-null   float64
 9   Potability       3276 non-null   int64
dtypes: float64(9), int64(1)
memory usage: 256.1 KB
```

IN 3:

```
# Shape of the DataFrame - shows tuple of (#Rows, #Columns)
print(df.shape)
# Find the number of rows within a DataFrame
print(len(df))
# Extracting information from the shape tuple
print(f'Number of rows: {df.shape[0]} \nNumber of columns: {df.shape[1]}')
```

```
(3276, 10)
3276
Number of rows: 3276
Number of columns: 10
```

Summary statistics

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2785.000000	3276.000000	3276.000000	3276.000000	2495.000000	3276.000000	3276.000000	3114.000000	3276.000000	3276.000000
mean	7.080795	196.369496	22014.092526	7.122277	333.775777	426.205111	14.284970	66.396293	3.966786	0.390110
std	1.594320	32.879761	8768.570828	1.583085	41.416840	80.824064	3.308162	16.175008	0.780382	0.487849
min	0.000000	47.432000	320.942611	0.352000	129.000000	181.483754	2.200000	0.738000	1.450000	0.000000
25%	6.093092	176.850538	15666.690297	6.127421	307.699498	365.734414	12.065801	55.844536	3.439711	0.000000
50%	7.036752	196.967627	20927.833607	7.130299	333.073546	421.884968	14.218338	66.622485	3.955028	0.000000
75%	8.062066	216.667456	27332.762127	8.114887	359.950170	481.792304	16.557652	77.337473	4.500320	1.000000
max	14.000000	323.124000	61227.196008	13.127000	481.030642	753.342620	28.300000	124.000000	6.739000	1.000000

```
# Transpose the summary details - easier to review larger number of
features
df.desribe().T
```

	count	mean	std	min	25%	50%	75%	max
ph	2785.0	7.080795	1.594320	0.000000	6.093092	7.036752	8.062066	14.000000
Hardness	3276.0	196.369496	32.879761	47.432000	176.850538	196.967627	216.667456	323.124000
Solids	3276.0	22014.092526	8768.570828	320.942611	15666.690297	20927.833607	27332.762127	61227.196008
Chloramines	3276.0	7.122277	1.583085	0.352000	6.127421	7.130299	8.114887	13.127000
Sulfate	2495.0	333.775777	41.416840	129.000000	307.699498	333.073546	359.950170	481.030642
Conductivity	3276.0	426.205111	80.824064	181.483754	365.734414	421.884968	481.792304	753.342620
Organic_carbon	3276.0	14.284970	3.308162	2.200000	12.065801	14.218338	16.557652	28.300000
Trihalomethanes	3114.0	66.396293	16.175008	0.738000	55.844536	66.622485	77.337473	124.000000
Turbidity	3276.0	3.966786	0.780382	1.450000	3.439711	3.955028	4.500320	6.739000
Potability	3276.0	0.390110	0.487849	0.000000	0.000000	0.000000	1.000000	1.000000

Missing values

As discussed earlier from the metadata and summary statistics there are a number of missing values within the DataFrame. To confirm if this is correct we can apply the code block below.

```
# Check for the missing values by column
df.isnull().sum()
```

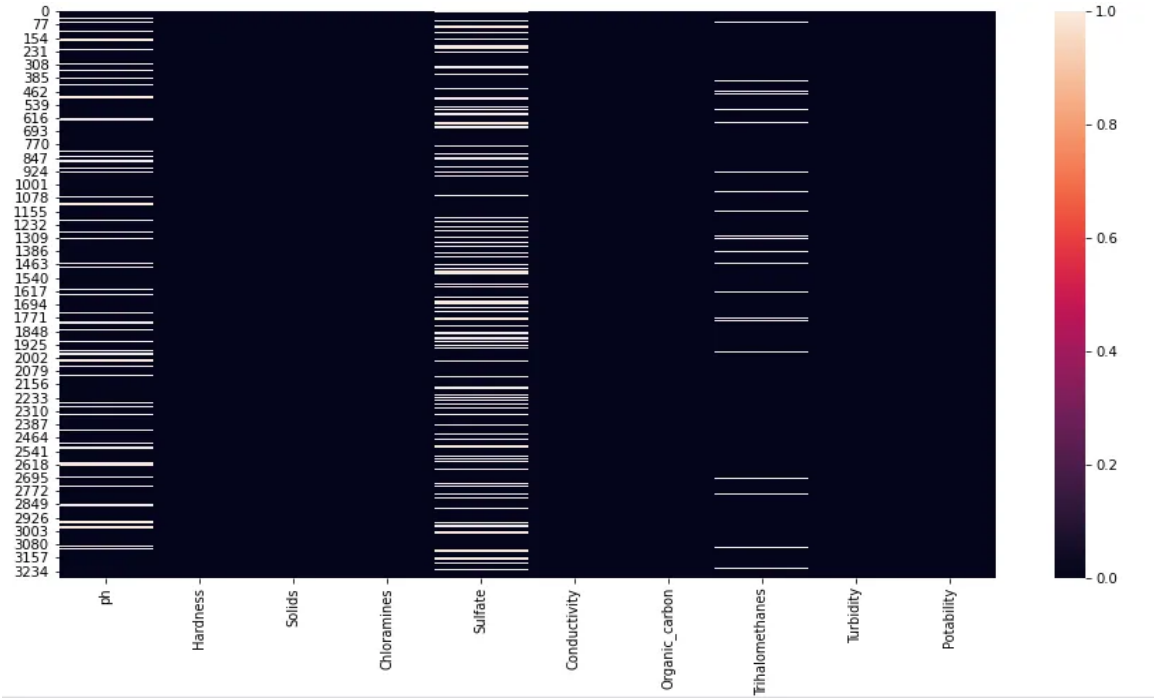
```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

```
# Proportion of missing values by column
def isnull_prop(df):
 total_rows = df.shape[0]
 missing_val_dict = {}
 for col in df.columns:
 missing_val_dict[col] = [df[col].isnull().sum(), (df[col].isnull().sum()
/ total_rows)]
 return missing_val_dict

# Apply the missing value method
null_dict = isnull_prop(df)
print(null_dict.items())
```

OUTPUT

```
dict_items([('ph', [491, 0.14987789987789987]), ('Hardness', [0, 0.0]), ('Solids', [0, 0.0]),
[162, 0.04945054945054945]), ('Turbidity', [0, 0.0]), ('Potability', [0, 0.0])])
```

```
# Display missing values using a heatmap to understand any patterns
plt.figure(figsize=(15,8))
sns.heatmap(df.isnull());
```



ADVANTAGES

Protection of Human Health:

- Monitoring and analyzing water quality help ensure that drinking water is safe and free from contaminants, reducing the risk of waterborne diseases and health problems.

Environmental Protection:

- Water quality analysis helps safeguard aquatic ecosystems, as it provides insights into the health of rivers, lakes, oceans, and other water bodies. This protects the habitats of aquatic species and maintains biodiversity.

Regulatory Compliance:

- Many countries have regulations and standards for water quality. Analyzing water quality allows authorities to ensure compliance with these regulations and take corrective actions when necessary.

Early Detection of Contamination:

- Water quality analysis can detect contaminants and pollutants in water sources at an early stage, allowing for timely intervention and pollution control.

Resource Management:

- Effective water quality analysis helps in the sustainable management of water resources, ensuring the availability of clean water for various purposes, such as agriculture, industry, and domestic use.

Monitoring and Trend Analysis:

- Regular monitoring of water quality enables the tracking of long-term trends and the assessment of the impact of human activities on water bodies.

Scientific Research:

- Water quality analysis is fundamental for scientific research related to hydrology, limnology, ecology, and environmental science, enabling researchers to study complex aquatic systems.

Data-Driven Decision-Making:

- The data generated from water quality analysis inform decision-makers in governmental agencies, industries, and communities, helping them make informed choices about water management and pollution control.

Emergency Response:

- Water quality analysis can provide early warnings of potential water quality issues, such as harmful algal blooms or chemical spills, allowing for timely emergency responses to mitigate risks.

Public Awareness and Education: Information from water quality analysis can be used to educate the public about the importance of clean water, fostering awareness and responsible water use.

Quality Control in Industries:

- Various industries rely on high-quality water for their processes. Water quality analysis helps industries maintain the quality of their water supply, leading to improved product quality and efficiency.

Sustainable Development:

- By analyzing water quality, it's possible to promote sustainable development and protect ecosystems, supporting the United Nations Sustainable Development Goal of ensuring clean water and sanitation for all.

Agriculture and Aquaculture:

- Water quality analysis is essential for managing water resources in agriculture and aquaculture, ensuring optimal conditions for crop growth and aquaculture operations.

Infrastructure Maintenance:

- Analysis of water quality helps in identifying potential issues with water supply and distribution infrastructure, ensuring its integrity and reliability.

Community Well-Being:

- Access to clean and safe water enhances the quality of life for communities by reducing the prevalence of water-related diseases and ensuring reliable water sources.

water quality analysis is crucial for safeguarding human health, protecting the environment, and supporting sustainable water resource management. Its many advantages contribute to safer, healthier, and more sustainable communities and ecosystems.

DISADVANTAGES

Complexity and Cost:

- Water quality analysis can be complex and costly, especially when a wide range of parameters and contaminants need to be monitored. The cost of instrumentation, sample collection, laboratory testing, and data analysis can be significant.

Time-Consuming:

- Conducting comprehensive water quality analysis can be time-consuming, as it often involves collecting and processing numerous samples and running multiple tests. This can lead to delays in obtaining results.

Limited Coverage:

- Water quality analysis is typically conducted at specific monitoring sites, which may not represent the entire water body. This can lead to limited spatial coverage and potential data gaps in large and complex aquatic systems.

Data Variability:

- Water quality can vary over time and space due to natural factors (e.g., weather, seasonality) and human activities (e.g., pollution events). This variability can make it challenging to capture the true state of water quality with occasional sampling.

Sampling Errors:

- Errors in sample collection, handling, and transportation can introduce inaccuracies in the analysis. Proper training and quality control measures are required to minimize sampling errors.

Instrumentation and Calibration:

- Water quality monitoring instruments require regular calibration and maintenance to ensure accurate and reliable measurements. Neglecting calibration can lead to measurement errors.

Data Interpretation:

- Interpreting water quality data can be complex, especially for non-experts. Understanding the significance of parameters and trends may require specialized knowledge.

Limited Resources:

- Many regions, especially in developing countries, lack the resources and infrastructure for comprehensive water quality analysis. This can result in inadequate monitoring and potentially adverse environmental and health effects.

Legal and Regulatory Challenges:

- Adherence to water quality regulations and standards can be challenging for industries and municipalities, leading to legal and compliance issues.

Privacy and Security:

- In cases where water quality data is collected and stored electronically, concerns related to data privacy and cybersecurity can arise

Data Management:

- The management of large volumes of water quality data can be challenging. Proper storage, retrieval, and sharing of data are important but may not always be well-organized.

Lack of Real-Time Data:

- Many water quality monitoring programs provide data with a delay, which may not be sufficient for addressing certain pollution incidents or immediate health risks.

Public Perception:

- Public concerns or misconceptions related to water quality can lead to unwarranted fears and skepticism about the safety of water sources.

Conclusion

Throughout this article, we have aimed to review the early stages of an EDA assessment. Metadata on the imported data was initially reviewed to display early insights. A deeper dive into the summary statistics allowed us to focus on the missing values. Finally, we were able to review the histogram of the pH variable to ensure that the variable followed external expectations.A follow-up article will continue the journey and seek to develop models that aim to predict water quality. Classification Machine Learning techniques will be used to provide baseline models.

PREPARED BY

M.DIVAKAR