WATER QUALITY

ANALYSIS

BATCH MEMBER

712521121028:SIVADHARMAN.R

Phase 3 Submission Document

Project Title: Water Quality Analysis

Phase3: ***Development part 1***

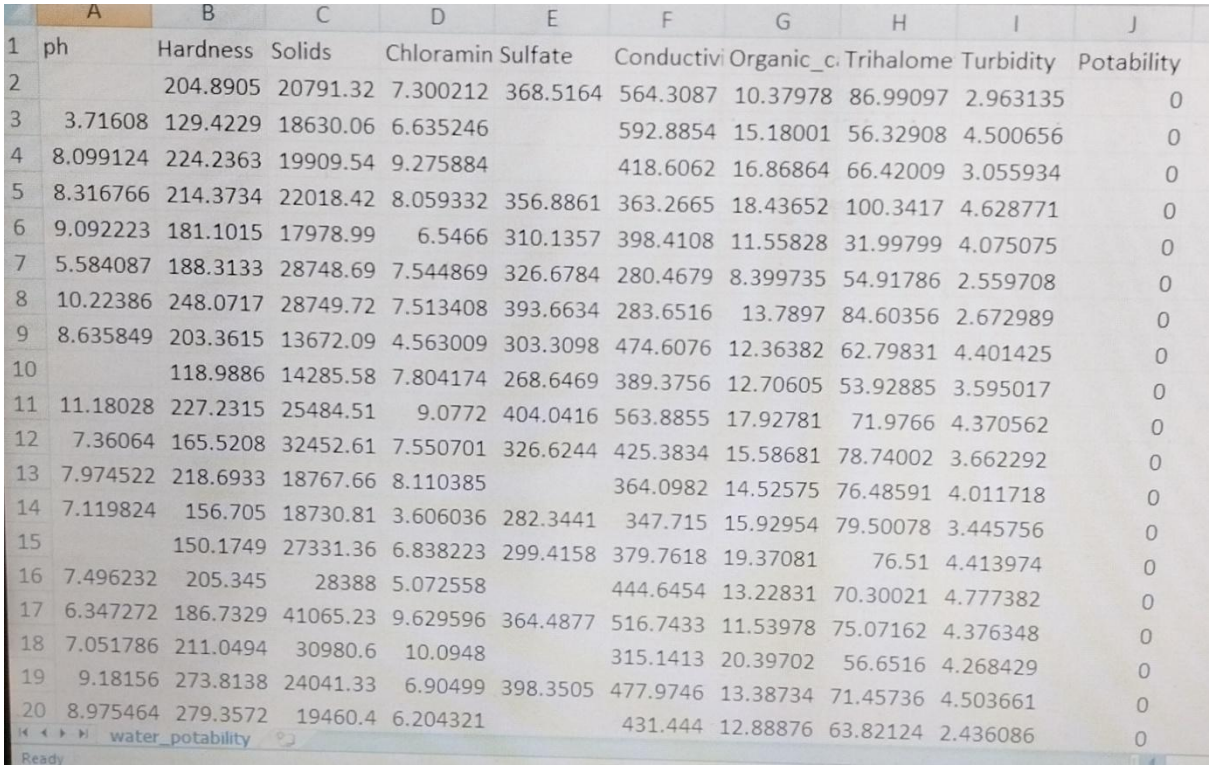
Topic: Start Analysing the water quality by loading and pre-processing the dataset

Water Quality Analysis

Introduction:

Water quality analyzers are used for monitoring process chemistry including water quality, providing process optimization and control. Water quality parameters are of three types – physical, chemical and biological – and are tested or monitored according to the desired water parameters. Water quality parameters often sampled or monitored include pH, ORP, conductivity, dissolved oxygen, chlorine, salinity, ozone, and corrosion rate. However water monitoring may also include measurement of chlorophyll, blue-green algae, ammonia nitrogen, nitrate, fluoride ions, or laboratory parameters such as BOD, COD and TOC.

Given Data Set:



	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramin	Sulfate	Conductivi	Organic_c	Trihalome	Turbidity	Potability
2		204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3	3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
4	8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
5	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
6	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
7	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
8	10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
9	8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
10		118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11	11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
12	7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
13	7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
14	7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
15		150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
16	7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
17	6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
18	7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
19	9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
20	8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0

water_potability

Ready

3277 rows*7 columns

Necessary steps to follow:

1. Import Libraries:

```
Import sys

print(sys.version)
```

2. 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.

```
# Import the dataset for review as a Data Frame
df = pd.read_csv("../input/water-portability/water_portability.csv")

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

Having imported the data, the code assigns the variable df with the Data Frame output results from the pandas method.

```
<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
```

Output: Provides an overview of the features and details of memory usage

```
# 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]}'
```

When calling an attribute in Python such as shape, there will be no parenthesis required. An attribute is a data result that can be accessed by both a class and its object. Earlier we reviewed a method which is a function that is contained within a class. For further insights on the smaller details a deep dive into how Python class statements function would be required. However, we can continue with the code that is used and show that with output 1.3 a number of values have been displayed.

Output:

No of rows:3276

No of columns:10

Challenge involved in loading and pre-processing of water quality analysis

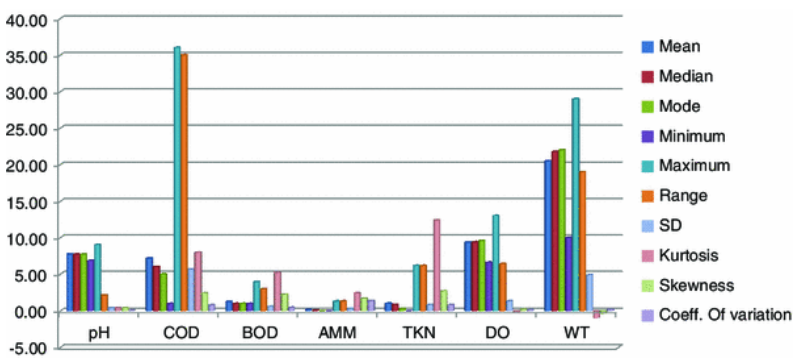
1.Data Sources: Water quality data can come from multiple sources, such as sensors, lab tests, or manual measurements, each with its own format and quality issues. Combining and standardizing these sources can be complex.
2. Missing Data: Incomplete or missing data points are common in water quality datasets. Deciding how to handle missing values, like imputation or removal, can impact the quality of analysis.
3. Data Volume: Large datasets with high temporal and spatial resolution can be challenging to manage and process efficiently, requiring specialized tools and hardware.
4. Temporal and Spatial Variability: Water quality can vary over time and across locations, necessitating techniques to aggregate or interpolate data for meaningful analysis.

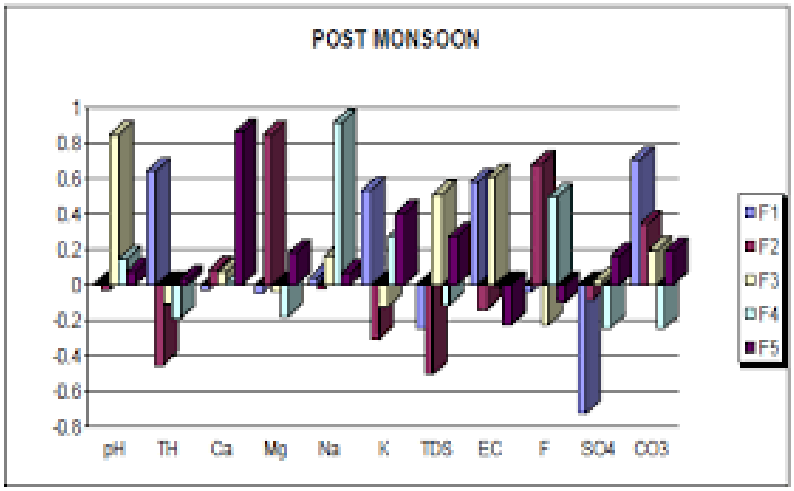
5. Data Transformation: Depending on the analysis goals, data may need various transformations, such as normalization, filtering, or feature engineering.
6. Data Exploration: Understanding the dataset's characteristics and patterns is crucial but can be time-consuming, especially with large datasets.
7. Tools and Software: Using appropriate software and tools for data manipulation, analysis, and visualization is important, and it may require a learning curve.
8. Automation and Scalability: For ongoing monitoring, setting up automated pipelines and scalable solutions is essential to handle continuous data streams.

Output Dataset:

	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramin	Sulfate	Conductiv	Organic_c	Trihalome	Turbidity	Potability
2		204.89	20791.3	7.30021	368.516	564.309	10.3798	86.991	2.96314	0
3	3.71608	129.423	18630.1	6.63525		592.885	15.18	56.3291	4.50066	0
4	8.09912	224.236	19909.5	9.27588		418.606	16.8686	66.4201	3.05593	0
5	8.31677	214.373	22018.4	8.05933	356.886	363.267	18.4365	100.342	4.62877	0
6	9.09222	181.102	17979	6.5466	310.136	398.411	11.5583	31.998	4.07508	0
3270	6.70255	207.321	17246.9	7.70812	304.51	329.266	16.2173	28.8786	3.44298	1
3271	11.491	94.8125	37188.8	9.26317	258.931	439.894	16.1728	41.5585	4.36926	1
3272	6.06962	186.659	26138.8	7.74755	345.7	415.887	12.0676	60.4199	3.66971	1
3273	4.6681	193.682	47581	7.16664	359.949	526.424	13.8944	66.6877	4.43582	1
3274	7.80886	193.553	17329.8	8.06136		392.45	19.9032		2.79824	1
3275	9.41951	175.763	33155.6	7.35023		432.045	11.0391	69.8454	3.29888	1
3276	5.12676	230.604	11983.9	6.30336		402.883	11.1689	77.4882	4.70866	1
3277	7.87467	195.102	17404.2	7.50931		327.46	16.1404	78.6984	2.30915	1

Visualisation and pre-processing of data:




POST MONSOON
1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
pH
TH
Ca
Mg
Na
K
TDS
EC
F
SO4
CO3
F1
F2
F3
F4
F5