**WATER QUALITY ANALYSIS**

**BATCH MEMBER**

**712521121022:PRANEESH T K**

**Phase 3 Submission Document**

**Project Title: Water Quality Analysis**

**Phase3: *Development part 1***

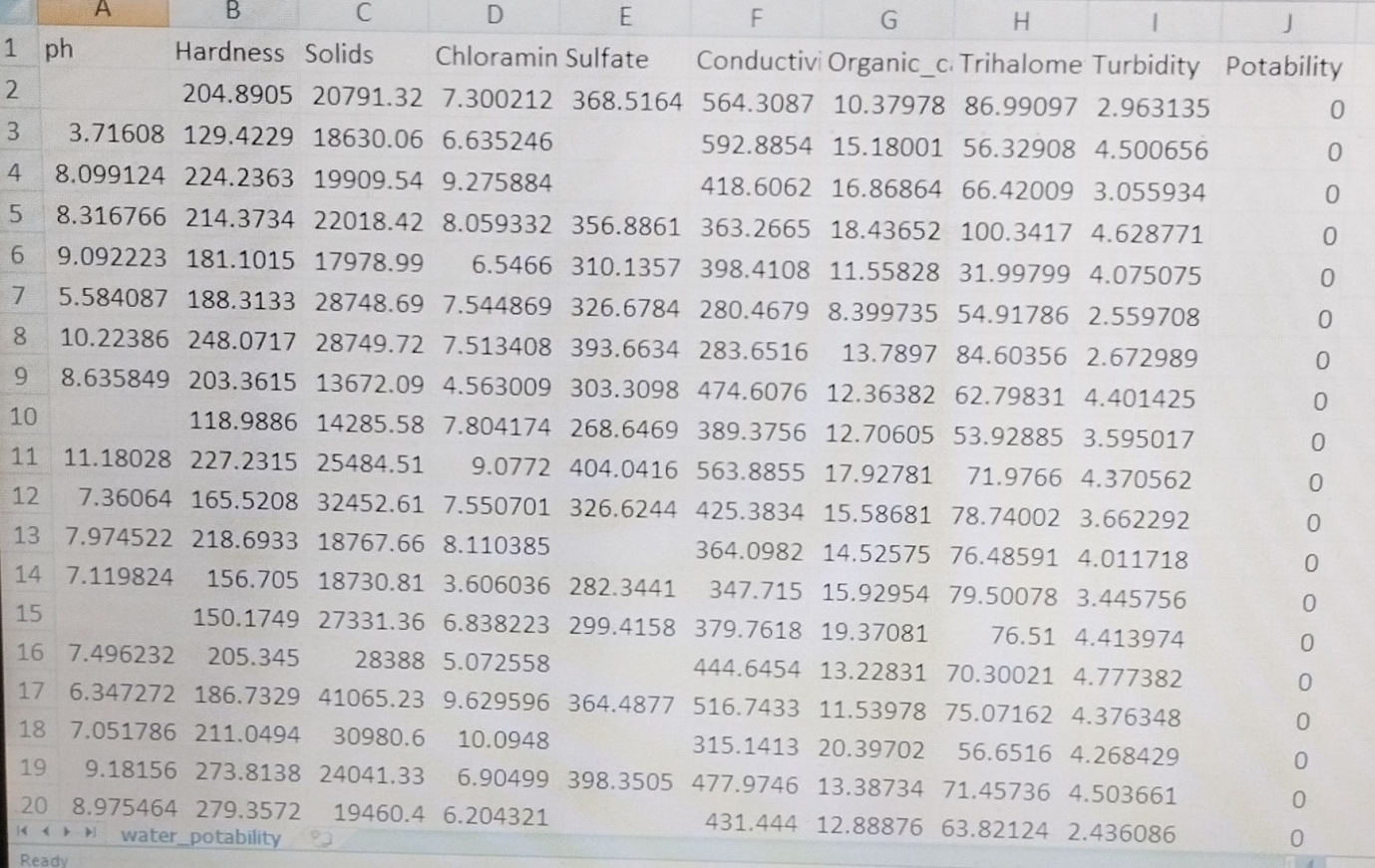
**Topic: Start Analysing the water quality by loading and pre-processing the dataset**

**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:**

3277 rows\*7 columns

**Necessary steps to follow:**

1. Import Libraries:

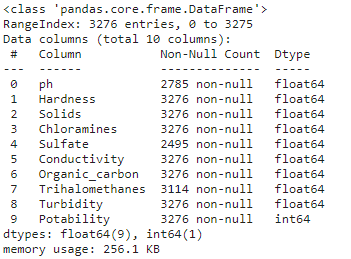
Import sys  
print(sys.version)

1. 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.



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.

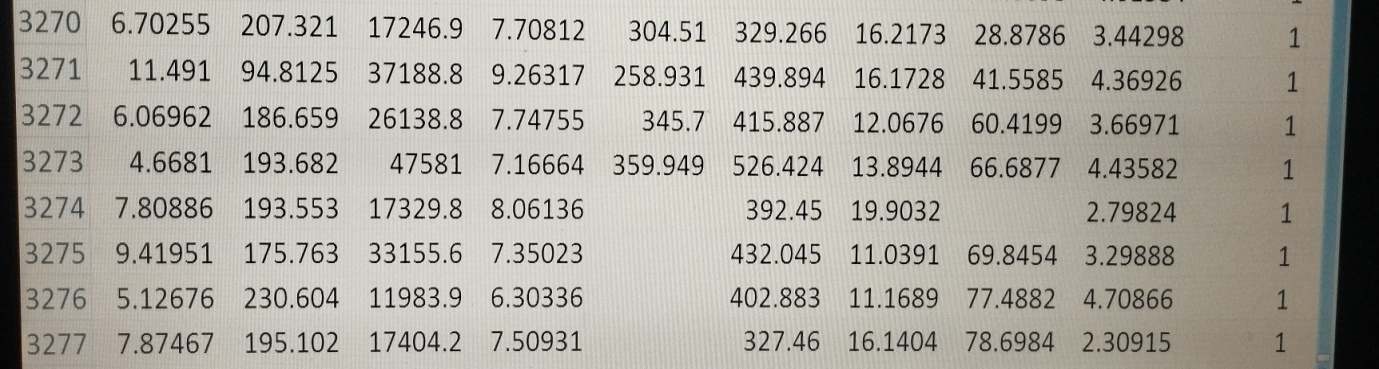
4. Temporal and Spatial Variability: Water quality can vary over time and across locations, necessitating techniques to aggregate or interpolate data for meaningful analysis.

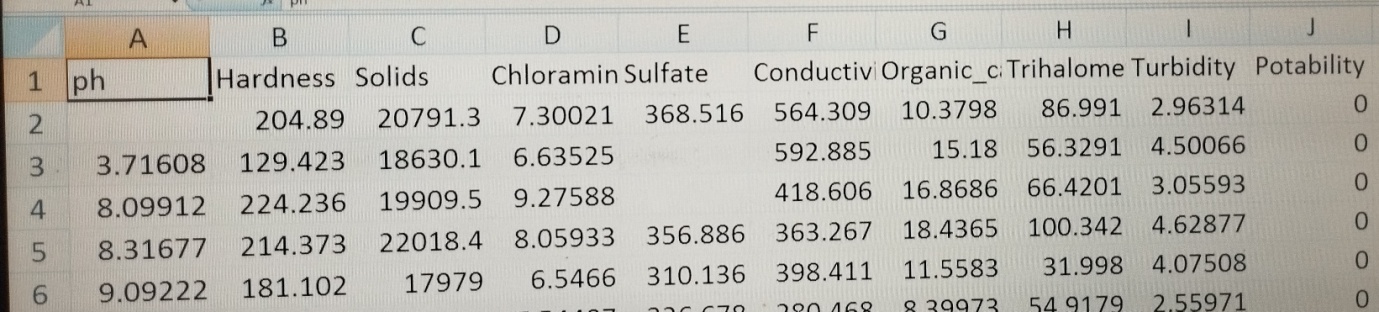
5. Data Transformation: Depending on the analysis goals, data may need various transformations, such as normalization, filtering, or feature engineering.

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:



**Visualisation and pre-processing of data:**

