**PHASE-3: WATER QUALITY ANALYSIS**

**CHECKING THE MISSING VALUES**

import pandas as pd

#This library used to read and write a data frame ,

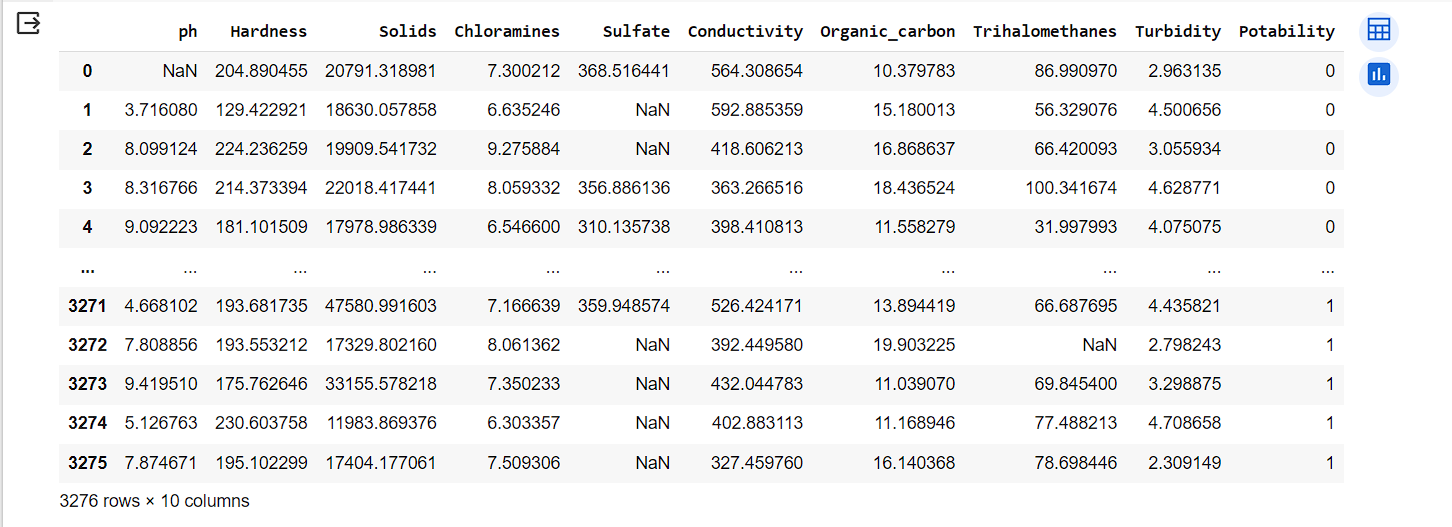
df = pd.read\_csv('/content/water\_potability.csv')

#upload the dataset using the pandas library

df

#printing the data set

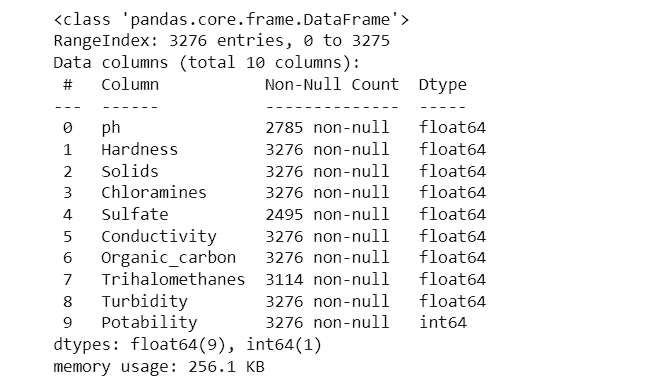
**OUTPUT:**



df.info()

#checking the datatype of the dataset

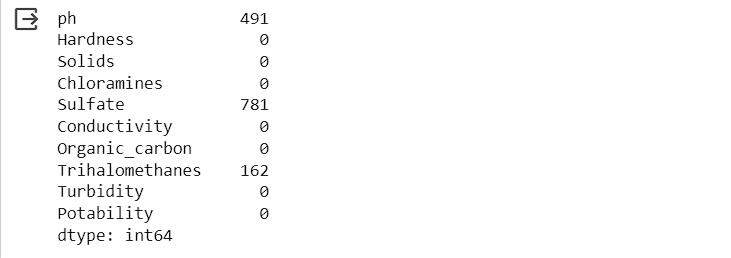
**OUTPUT:**



df.isnull().sum()

#checking the missing values in the dataset

**OUTPUT:**



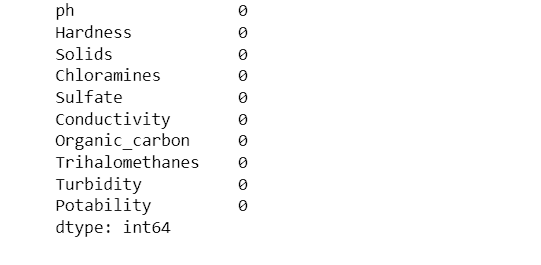
df.fillna(9.5,inplace=True)

#Replace the missing values using fillna() command with the common value 9.5

df.isnull().sum()

#again checking for the missing values

**OUTPUT:**



Exploratory Analysis:

**#Importing the libraries**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.metrics import r2\_score, mean\_squared\_error, mean\_absolute\_error

**#upload the data set using the library**

df = pd.read\_csv("/content/water\_potability.csv",

encoding="ISO-8859-1")

df = df.replace("NA", np.nan)

df = df.dropna()

df = df.astype(float)

df = df[df["Conductivity"] > 0]

X = df[["ph", "Hardness", "Solids", "Chloramines", "Sulfate", "Conductivity"]]

y = df["Turbidity"]

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state=42)

model = LinearRegression()

model.fit(X\_train, y\_train)

y\_pred = model.predict(X\_test)

r2 = r2\_score(y\_test, y\_pred)

mse = mean\_squared\_error(y\_test, y\_pred)

mae = mean\_absolute\_error(y\_test, y\_pred)

print(f"R-squared: {r2:.3f}")

print(f"Mean Squared Error: {mse:.3f}")

print(f"Mean Absolute Error: {mae:.3f}")

**#ploting the data for scatter plot**

plt.scatter(y\_test, y\_pred)

plt.xlabel("Actual WQI")

plt.ylabel("Predicted WQI")

plt.title("Linear Regression Model for Water Quality Prediction")

plt.show()

**#OUTPUT:**

