**WATER QUALITY ANALYSIS**

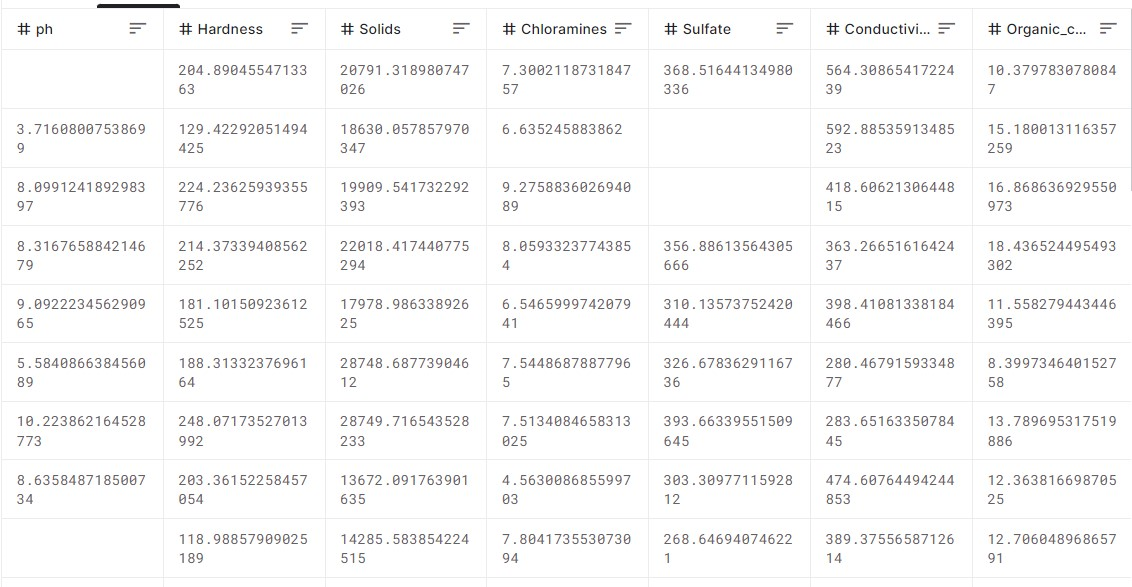
**PHASE 3: Development Part 1**

Introduction

Water is a fundamental resource for all living organisms and plays a crucial role in maintaining ecological balance and supporting human activities. Ensuring the availability of clean and safe water is essential for human health, agricultural productivity, industrial processes, and the overall well-being of communities. Water quality analysis is a scientific process used to assess the physical, chemical, and biological characteristics of water. By understanding these parameters, scientists, environmentalists, and policymakers can make informed decisions regarding water treatment, conservation, and environmental protection.

In this phase of the project we will be preprocessing the given dataset and perform EDA.

Given DataSet



3200 rows and 10 columns

Preprocessing the given DataSet

Data processing is used to transform the given data into a format that is most effective before using the dataset, in data processing we remove all the null values and all the outliners.

Removing the null values

The first step of data processing is to remove all the null values ,to remove the null values perform the following steps

Step1:

Save your data set with extension csv

Now to import your data set use the following coad

Import pandas as pd

dataset=pd.read(“DataSet.csv”)

#now the dataset is imported

#to print the dataset

dataset.head()

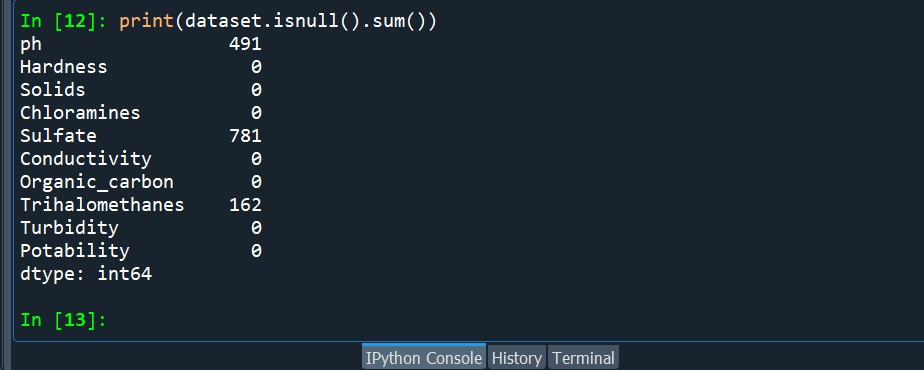
Step2

#Now to find the columns with nullvalues

dataset.columns

dataset.describe

print(dataset.isnull().sum())



Step3:

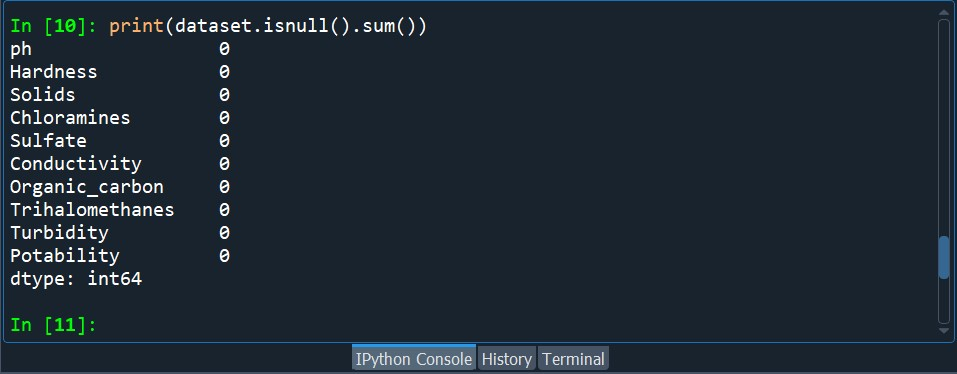
As we can see columns ph , Sulfate ,Trihalomethanes contains null values, so to replace the null values type the following code

dataset.ph=dataset.ph.fillna(“unknown”)

dataset.Sulfate=dataset.Sulfate.fillna(“unknown”)

dataset.Trihalomethanes=dataset.Trihalomethanes.fillna(“unknown”)

print(dataset.isnull().sum())



As you can see all the null values has been removed by replacing it with the word “unknown”.

Removing outliners

The next step of datapreprocessing is to remove the outliners,

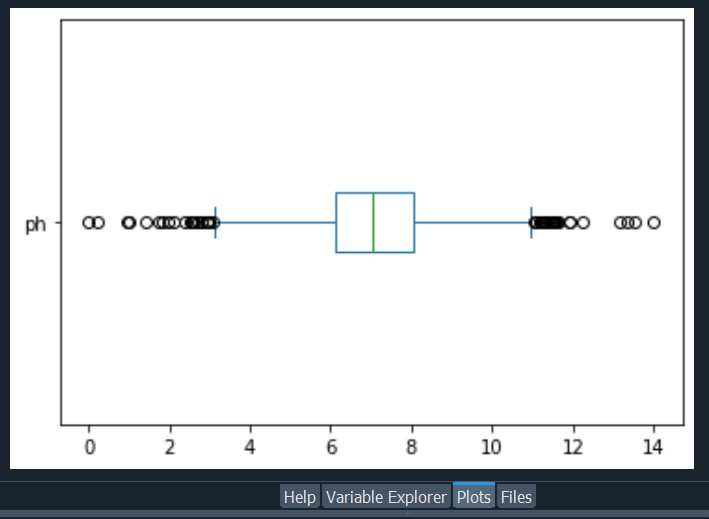
Outliners are the values in a data set which are greatly different from other values ,either too large or either too small.

First to remove the outliners we have to find the outliners , to find the outliners write the following code

Import matplotlib.pyplot as plt

dataset[‘ph’].plot(kind=’box’);

#here ph is one of the colums in the given dataset



Here the circles visible outside are the outliners

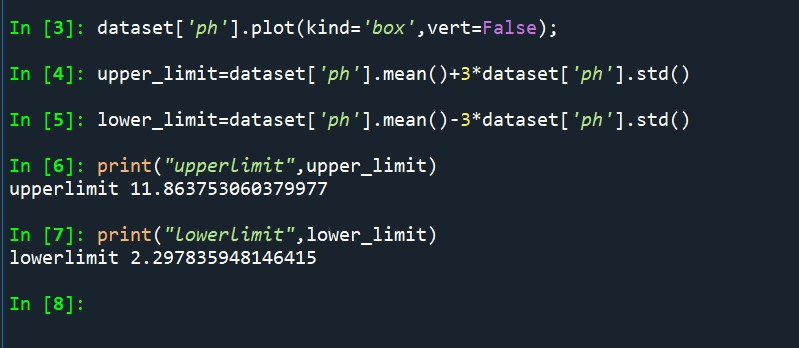
Now to find the upper limit and the lower limit follow the code given below:

upper\_limit=dataset[‘ph’].mean()+3\*dataset[‘ph’].std()

lower\_limit=dataset[‘ph’].mean()-3\*dataset[‘ph’].std()

print(“upperlimit”,upper\_limit)

print(“lowerlimit”,lower\_limit)



Now after finding the upper and the lower limit we have to remove the outliners to remove the outliners perform the following codes

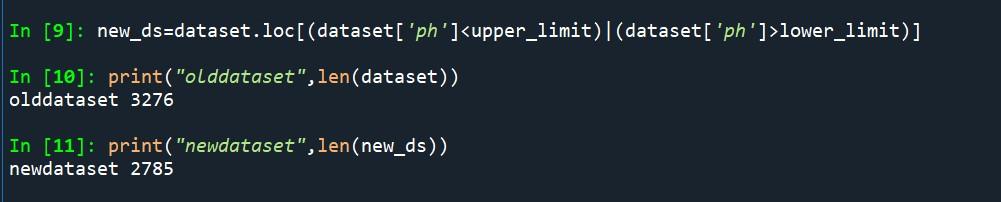
#finding the location

dataset.loc[(dataset[‘ph’]>upper\_limit)|(dataset[‘ph’]<lower\_limit)]

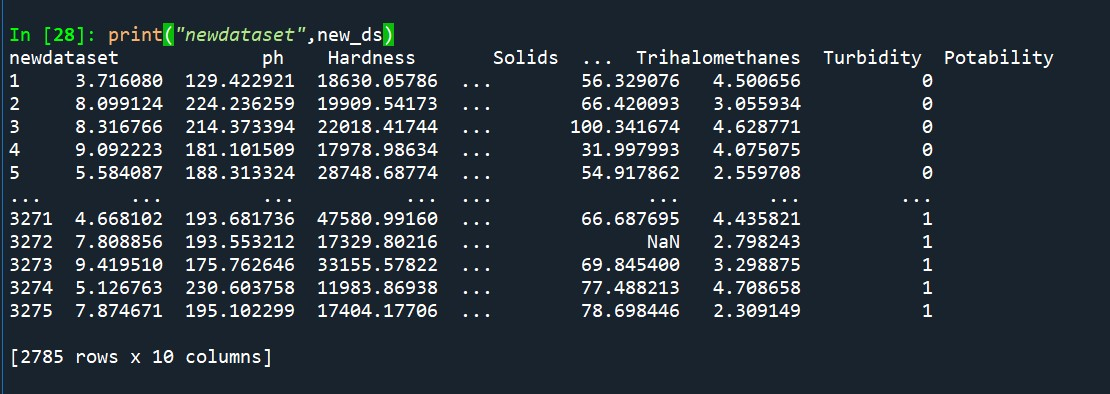
new\_dataset=dataset.loc[(dataset[‘ph’]<upper\_limit)|(dataset[‘ph’]>lower\_limit)]

print(“old dataset”,len(dataset))

print(“new dataset”,len(new\_dataset))



As you can see the outliners has been removed from the database



Now we have removed all the null values and outliners from the dataset.

Performing Exploratory Data Analysis(EDA)

Exploratory Data Analysis (EDA) refers to the method of studying and exploring record sets to apprehend their predominant traits, discover patterns, locate outliers, and identify relationships between variables. EDA is normally carried out as a preliminary step before undertaking extra formal statistical analyses or modelling. Now we are going to conduct EDA to visualize parameter distributions, correlations, and potential deviations from

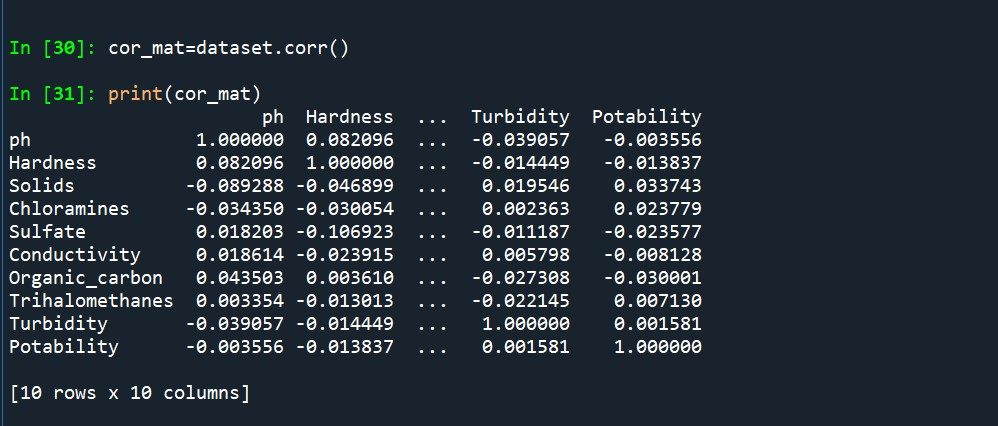
standards.

Correlation

Correlation is used to measure the relationship between two variables , to find the correlation perform the following steps

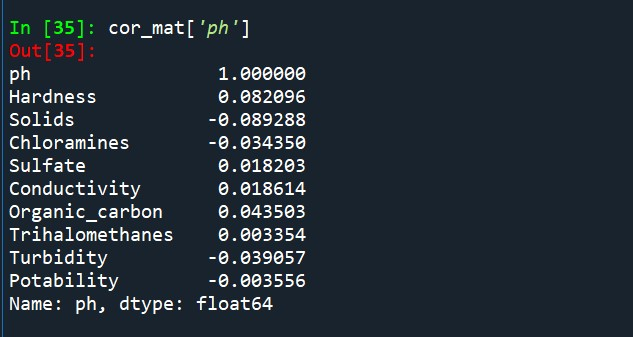
corr\_matr=dataset.corr()

print(corr\_matr)



The correlation of a column can be found by

corr\_matr[‘ph’]



Standard Deviation

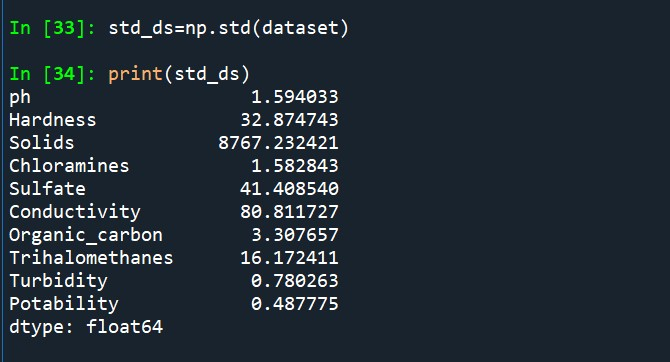
**Standard deviation** is the measure of the dispersion of statistical data. **Standard deviation formula** is used to find the deviation of the data value from the mean value i.e. it is used to find the dispersion of all the values in a data set with respect to the mean value.

To find the standard deviation of the given dataset perform the following code

Import numpy as np

std=np.std(dataset)

print(std)



Conclusion

In this phase of the project we have preprocessed the given dataset by removing all the null values and the outliners and we have also performed Exploratory Data Analysis(EDA) to visualise parameter distribution , correlation and potential deviation from the standards.