PHASE:3

Water Quality Analysis

TEAM MEMBERS:

NAME: Archana R

REG NO:721221104007

NAME: Illakiya R K

REG NO:721221104020

NAME: LitheeshKumar G

REG NO:7212211040032

NAME: Mohammed Harshad N

REG NO:721221104038

NAME: Nathish T

REG NO:721221104041

Introduction:

Water quality analysis is a critical aspect of environmental science and public health, aiming to assess the safety and health of water sources for various purposes, such as drinking, agriculture, industrial use, and aquatic ecosystems. Data analytics plays a vital role in this field by enabling the collection, processing, interpretation, and visualization of data related to water quality. It helps researchers, environmentalists, and policymakers make informed decisions about managing water resources and protecting public health.

Data Collection:

- Water quality data is collected from various sources, including rivers, lakes, reservoirs, groundwater, and treatment facilities.
- Data may include measurements of physical, chemical, and biological parameters, such as such temperature, PH, turbidity, dissolved oxygen, nutrients, heavy metals, and microbial contaminants.
- Sensors, monitoring stations, and sampling methods are used to collect data over time, providing a detailed picture of water quality dynamics.

DatasetLink: https://www.kaggle.com/datasets/adityakadiwal/water-potability

Data Preprocessing and Cleaning:

• Clean the collected data to ensure its quality and accuracy.

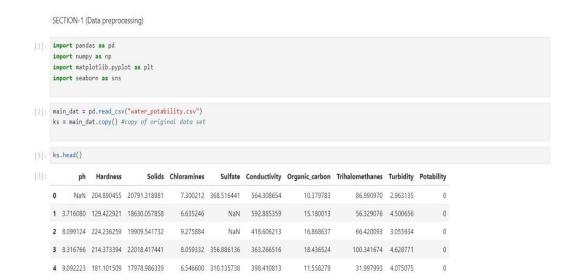
#importing data set

import pandas as pd

import numpy as np

```
import matplotlib.pyplot as plt
import seaborn as sns
main_dat = pd.read_csv("water_potability.csv")
ks = main_dat.copy() #copy of original data set
ks.head()
```

OUTPUT:



ks.sample(5)

ks.shape

ks.columns

OUTPUT:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
1018 6.0	.013161	218.843256	21573.747571	9.295852	321.168313	444.276635	14.744347	62.443239	3.455623	0
248 6.5	.581878	272.982745	37169.444404	8.114731	416.083481	351.476839	15.129334	79.261026	4.201663	0
1998 7.5	544306	211.051146	34359.400797	8.166793	365.812313	447.520655	18.553478	60.162746	3.714096	1
2227	NaN	159.832881	23917.190146	6.781576	369.223852	472.927194	13.891834	85.758645	2.857687	0
2484 6.6	653650	172.584512	34816.444538	8.289307	293.611048	389.471149	15.872474	67.976869	4.871406	0
ks.shape (3276, 1 ks.colum	10)									

pd.isnull(ks).sum()

ks.dropna(inplace=True)

pd.isnull(ks).sum()

OUTPUT:

```
[7]: pd.isnull(ks).sum()
[7]: ph
                      491
     Hardness
                      0
                    0
     Solids
     Chloramines
                       0
                    781
     Sulfate
     Conductivity
     Organic_carbon
     Trihalomethanes 162
                     0
     Turbidity
     Potability
     dtype: int64
[8]: ks.dropna(inplace=True)
     pd.isnull(ks).sum()
[8]: ph
     Hardness
                     0
     Solids
     Chloramines
     Sulfate
     Conductivity
     Organic_carbon
     Trihalomethanes
     Turbidity
                     0
     Potability
     dtype: int64
```

ks.describe()

ks.nunique()

OUTPUT:

	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
count	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000	2011.000000
mean	7.085990	195.968072	21917.441374	7.134338	333.224672	426.526409	14.357709	66.400859	3.969729	0.403282
std	1.573337	32.635085	8642.239815	1.584820	41.205172	80.712572	3.324959	16.077109	0.780346	0.490678
min	0.227499	73.492234	320.942611	1.390871	129.000000	201.619737	2.200000	8.577013	1.450000	0.000000
25%	6.089723	176.744938	15615.665390	6.138895	307.632511	366.680307	12.124105	55.952664	3.442915	0.000000
50%	7.027297	197.191839	20933.512750	7.143907	332.232177	423.455906	14.322019	66.542198	3.968177	0.000000
75%	8.052969	216.441070	27182.587067	8.109726	359.330555	482.373169	16.683049	77.291925	4.514175	1.000000
max	14.000000	317.338124	56488.672413	13.127000	481.030642	753.342620	27.006707	124.000000	6.494749	1.000000
ks.nu	nique()									
ph Hardness Solids		2011 2011								
Chloramines Sulfate		2011 2011 2011								
Organ	ctivity ic_carbon lomethanes	2011 2011 2011								
Turbi		2011								
Potab	ility : int64	2								

ks.info()

ks.dtypes

OUTPUT:

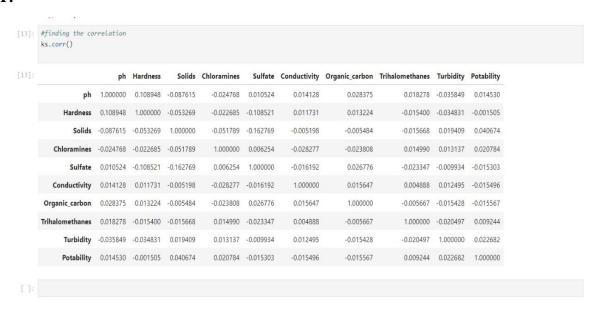
```
[11]: ks.info()
                                                                                                                                                                                          回个少去早前
         <class 'pandas.core.frame.DataFrame'>
        Index: 2011 entries, 3 to 3271
Data columns (total 10 columns):

# Column Non-Null Count Dtype
         # Column
          0 ph
1 Hardness
                                      2011 non-null
2011 non-null
                                                             float64
float64
          2 Solids
3 Chloramines
4 Sulfate
                                       2011 non-null
                                                             float64
                                       2011 non-null
                                                             float64
                                      2011 non-null
2011 non-null
                                                             float64
           5 Conductivity
                                                             float64
          6 Organic_carbon 2011 non-null
7 Trihalomethanes 2011 non-null
8 Turbidity 2011 non-null
                                                             float64
float64
                                       2011 non-null
         dtypes: float64(9), int64(1) memory usage: 172.8 KB
[12]: ks.dtypes
[12]: ph
         Hardness
Solids
Chloramines
                                   float64
float64
                                   float64
         Sulfate
Conductivity
                                   float64
                                   float64
         Organic_carbon
                                   float64
          Trihalomethanes
                                   float64
int64
         Turbidity
Potability
         dtype: object
```

#finding the correlation

ks.corr()

OUTPUT:



Project Conclusion:

In conclusion, data analytics is an indispensable tool in the field of water quality analysis, enabling us to gain valuable insights into the health and safety of water sources. By harnessing the power of data analytics, we can make informed decisions that impact public health, environmental sustainability, and resource management







