# Problem Statement ¶

Access to safe drinking-water is essential to health, a basic human right and a component of effective policy for health protection. This is important as a health and development issue at a national, regional and local level. In some regions, it has been shown that investments in water supply and sanitation can yield a net economic benefit, since the reductions in adverse health effects and health care costs outweigh the costs of undertaking the interventions.

**Content** - The **water_potability.csv** file contains water quality metrics for 3276 different water bodies.

1. pH value: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
2. Hardness: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
3. Solids (Total dissolved solids - TDS): Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.
4. Chloramines: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.
5. Sulfate: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.
6. Conductivity: Pure water is not a good conductor of electric current rather’s a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 µS/cm.
7. Organic_carbon: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA < 2 mg/L as TOC in treated / drinking water, and < 4 mg/Lit in source water which is use for treatment.

8. Trihalomethanes: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.
9. Turbidity: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.
10. Potability: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

**Task 1**

- Predict if water is safe for human consumption or not.

1. ph: pH of 1. water (0 to 14).
2. Hardness: Capacity of water to precipitate soap in mg/L.
3. Solids: Total dissolved solids in ppm.
4. Chloramines: Amount of Chloramines in ppm.
5. Sulfate: Amount of Sulfates dissolved in mg/L.
6. Conductivity: Electrical conductivity of water in µS/cm.
7. Organic_carbon: Amount of organic carbon in ppm.
8. Trihalomethanes: Amount of Trihalomethanes in µg/L.
9. Turbidity: Measure of light emiting property of water in NTU.
10. Potability: Indicates if water is safe for human consumption. Potable -1 and Not potable -0

In [ ]:

# Steps :

## Import Relevant Libraries/Modules

- **Basic EDA**
  - Imbalance Data
  - Missing Data
  - Duplicate Data
  - Outliers or Anomalies
  - Data Visualization

- Feature Encoding
- Feature Selection

- **Model Building**
  - Separate your Independent and Dependent data
  - Split your data into train and test
  - Model Selection
  - Model Training
  - Model Prediction
  - Model Evaluation
  - Hyperparameter Tuning

In [ ]:

In [1]:
```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

import sklearn
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LogisticRegression
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier

import xgboost
from xgboost import XGBClassifier

from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import GridSearchCV, RandomizedSearchCV
from sklearn.metrics import confusion_matrix, accuracy_score, classification_report

import warnings
warnings.filterwarnings('ignore')
```

In [ ]:

```
```

In [2]:

```
df = pd.read_csv('water_potability.csv')
df.head()
```

Out[2]:

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbidity | Potabili |
|---|---|---|---|---|---|---|---|---|---|---|
| **0** | NaN | 204.890455 | 20791.318981 | 7.300212 | 368.516441 | 564.308654 | 10.379783 | 86.990970 | 2.963135 | |
| **1** | 3.716080 | 129.422921 | 18630.057858 | 6.635246 | NaN | 592.885359 | 15.180013 | 56.329076 | 4.500656 | |
| **2** | 8.099124 | 224.236259 | 19909.541732 | 9.275884 | NaN | 418.606213 | 16.868637 | 66.420093 | 3.055934 | |
| **3** | 8.316766 | 214.373394 | 22018.417441 | 8.059332 | 356.886136 | 363.266516 | 18.436524 | 100.341674 | 4.628771 | |
| **4** | 9.092223 | 181.101509 | 17978.986339 | 6.546600 | 310.135738 | 398.410813 | 11.558279 | 31.997993 | 4.075075 | |

In [3]:

```
df.shape
```

Out[3]:

```
(3276, 10)
```

In [ ]:

```
```

In [4]:
```
df.info()
```

```
<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
```

In [ ]:
```
```

In [5]:
```
df.describe()
```

Out[5]:

| | ph | Hardness | Solids | Chloramines | Sulfate | Conductivity | Organic_carbon | Trihalomethanes | Turbi |
|---|---|---|---|---|---|---|---|---|---|
| **count** | 2785.000000 | 3276.000000 | 3276.000000 | 3276.000000 | 2495.000000 | 3276.000000 | 3276.000000 | 3114.000000 | 3276.000 |
| **mean** | 7.080795 | 196.369496 | 22014.092526 | 7.122277 | 333.775777 | 426.205111 | 14.284970 | 66.396293 | 3.966 |
| **std** | 1.594320 | 32.879761 | 8768.570828 | 1.583085 | 41.416840 | 80.824064 | 3.308162 | 16.175008 | 0.780 |
| **min** | 0.000000 | 47.432000 | 320.942611 | 0.352000 | 129.000000 | 181.483754 | 2.200000 | 0.738000 | 1.450 |
| **25%** | 6.093092 | 176.850538 | 15666.690297 | 6.127421 | 307.699498 | 365.734414 | 12.065801 | 55.844536 | 3.439 |
| **50%** | 7.036752 | 196.967627 | 20927.833607 | 7.130299 | 333.073546 | 421.884968 | 14.218338 | 66.622485 | 3.955 |
| **75%** | 8.062066 | 216.667456 | 27332.762127 | 8.114887 | 359.950170 | 481.792304 | 16.557652 | 77.337473 | 4.500 |
| **max** | 14.000000 | 323.124000 | 61227.196008 | 13.127000 | 481.030642 | 753.342620 | 28.300000 | 124.000000 | 6.739 |

In [ ]:

# Imbalance Dataset

In [6]:
```
df['Potability']
```

Out[6]:
```
0       0
1       0
2       0
3       0
4       0
       ..
3271    1
3272    1
3273    1
3274    1
3275    1
Name: Potability, Length: 3276, dtype: int64
```

In [7]:
```
df['Potability'].value_counts()
```

Out[7]:
```
0    1998
1    1278
Name: Potability, dtype: int64
```

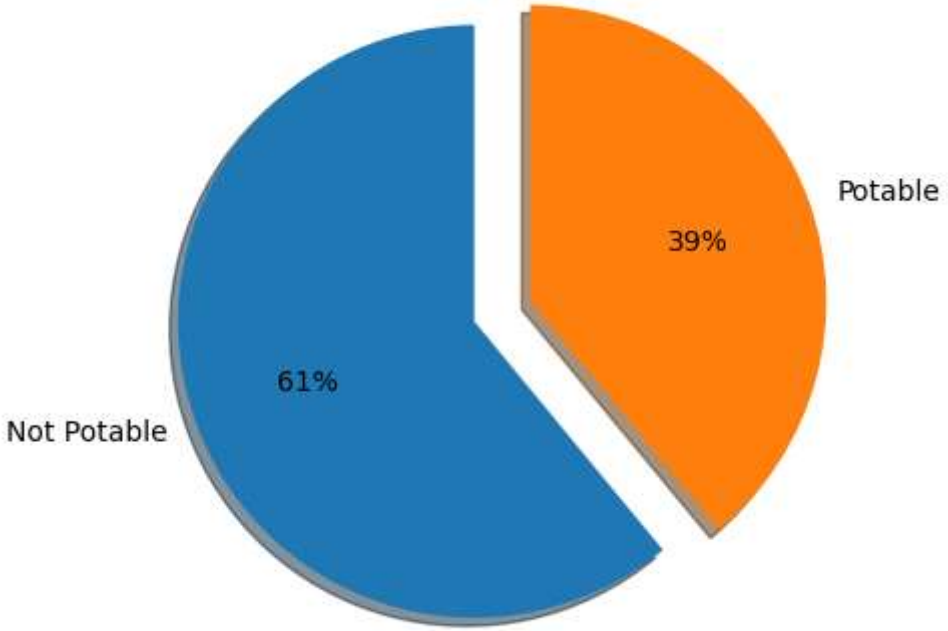
In [14]:
```
print(round(len(df[df['Potability'] == 0])/len(df)*100, 2))
print(round(len(df[df['Potability'] == 1])/len(df)*100, 2))
```

```
60.99
39.01
```

In [ ]:

In [21]:
```
plt.pie(df['Potability'].value_counts(), autopct='%1.0f%%', labels=['Not Potable', 'Potable'], startangle=
        shadow=True, explode=[0, 0.2])


plt.show()
```




In [ ]:
```

```

# Missing Data

In [22]:
```
df.isnull().sum()
```

Out[22]:
```
ph                491
Hardness            0
Solids              0
Chloramines         0
Sulfate           781
Conductivity        0
Organic_carbon      0
Trihalomethanes   162
Turbidity           0
Potability          0
dtype: int64
```

In [26]:
```
round(df.isnull().mean()*100, 2)
```

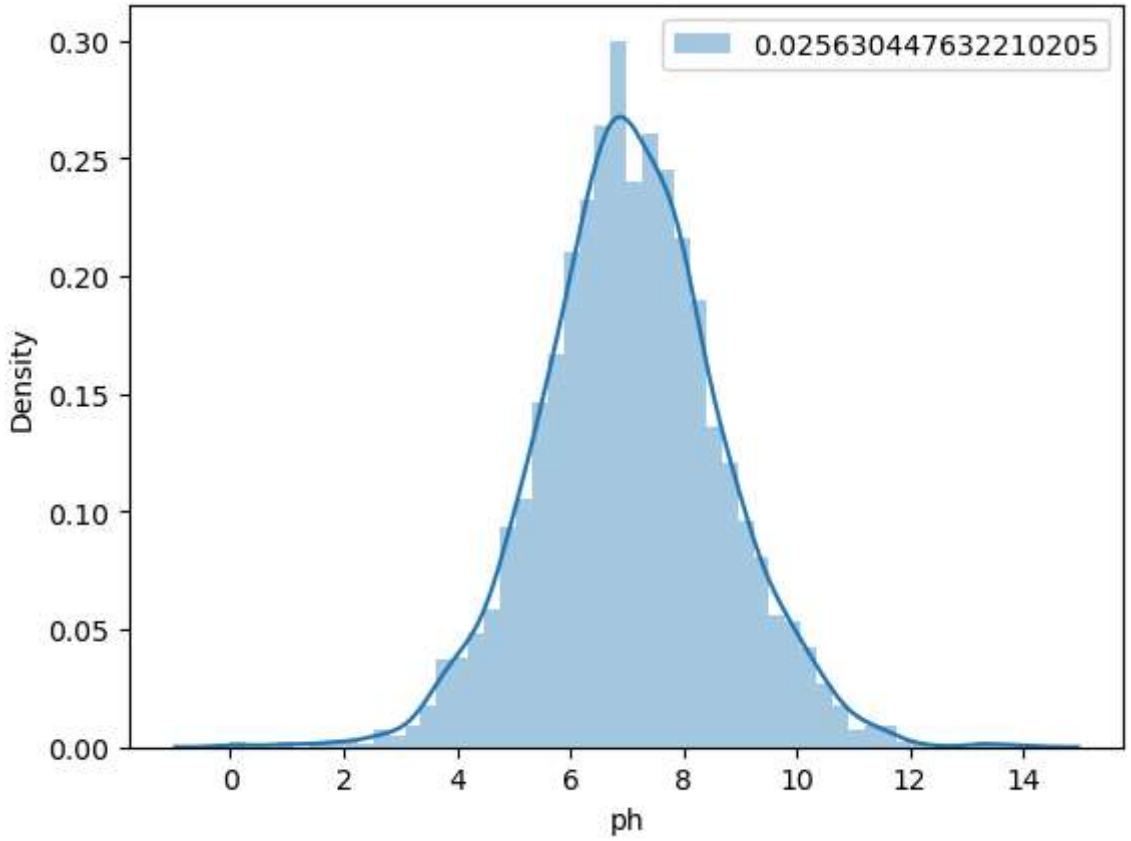
Out[26]:
```
ph                14.99
Hardness           0.00
Solids             0.00
Chloramines        0.00
Sulfate           23.84
Conductivity       0.00
Organic_carbon     0.00
Trihalomethanes    4.95
Turbidity          0.00
Potability         0.00
dtype: float64
```

In [27]:
```
df['ph'].skew()
```

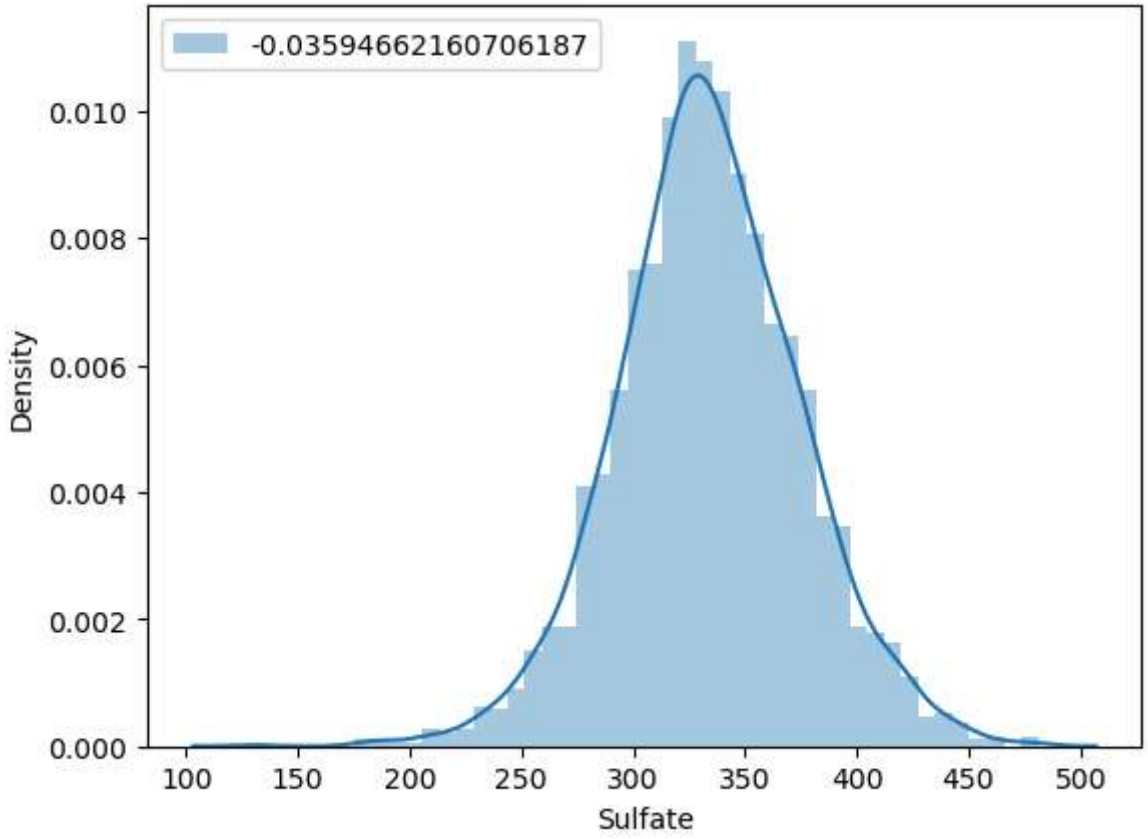
Out[27]:
```
0.025630447632210205
```

In [33]:
```
for i in df:
    if df[i].isnull().sum() > 0:
        print(i)
        sns.distplot(df[i], label= df[i].skew())
        plt.legend()
        plt.show()
```

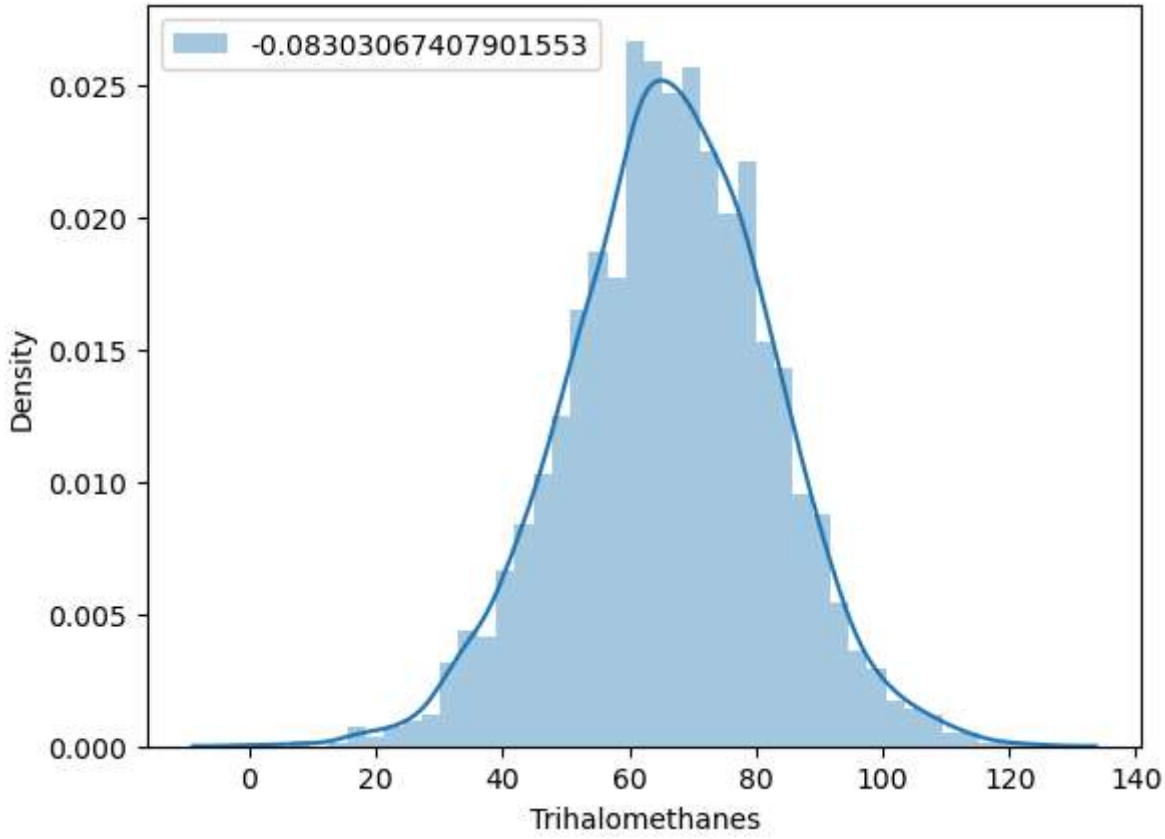
ph




Sulfate




Trihalomethanes


-0.08303067407901553
0.025
0.020
0.015
0.010
0.005
0.000
Density
0
20
40
60
80
100
120
140
Trihalomethanes

In [ ]:

In [37]:
```
for i in df:
    if df[i].isnull().sum() > 0:
        print(i)
        print(df[i].mean())
        print(df[i].median())
        print('\n')
```

```
ph
7.080794504276835
7.036752103833548


Sulfate
333.7757766108135
333.073545745888


Trihalomethanes
66.39629294676803
66.62248509808484
```

In [ ]:
```

```

In [41]:
```
for i in df:
    if df[i].isnull().sum() > 0:
        print(f' {i} : {df[i].mean()}')
        df[i].fillna(df[i].mean(), inplace=True)
```

```
 ph : 7.080794504276835
 Sulfate : 333.7757766108135
 Trihalomethanes : 66.39629294676803
```

In [ ]:
```

```

In [42]:
```
df.isnull().sum()
```

Out[42]:
```
ph                 0
Hardness           0
Solids             0
Chloramines        0
Sulfate            0
Conductivity       0
Organic_carbon     0
Trihalomethanes    0
Turbidity          0
Potability         0
dtype: int64
```

In [ ]:

In [ ]:

In [ ]:

In [ ]:
```
Basic EDA

Imbalance Data
Missing Data
Duplicate Data
Outliers or Anomalies
Data Visualization
Feature Encoding
Feature Selection
```

In [ ]:

In [ ]:

In [ ]:

In [ ]: