Water Quality Analysis Using Data Analytics With Cognos

Phase 5 Final submission document

BATCH MEMBERS

- Yuvaraj V (2021115125)
- Shadhani TT (2021115302)
- Adnan Khan S (2021115304)
- Vishal Raj Vellaisamy (2021115120)
- Vikhram SS (2021115119)
- Sabarisrinath R (2021115335

In partial fulfillment for the award of the degree
Of

BACHELOR OF TECHNOLOGY

In

INFORMATION TECHNOLOGY

DEPARTMENT OF INFORMATION SCIENCE AND TECHNOLOGY

COLLEGE OF ENGINEERING GUINDY

ANNA UNIVERSITY, CHENNAI 600025

NOVEMBER 2023

1.Objective

1. **Assessment of Water Quality:**

   The primary objective is to assess and evaluate the quality of water in a particular area, water body, or water source. This assessment includes measuring various physical, chemical, and biological parameters to understand the current state of water quality.

2. **Compliance with Regulations:**

   Ensure that water quality meets regulatory standards and guidelines set by governmental and environmental agencies. This objective is crucial for compliance with laws and regulations governing water quality.

3. **Identification of Contaminants:**

   Detect and identify the presence of contaminants, pollutants, and potential health hazards in the water, including pathogens, heavy metals, chemicals, and other harmful substances.

4. **Monitoring Changes Over Time:**

   Track changes in water quality over time by conducting regular sampling and analysis. This is essential for understanding trends, seasonal variations, and the impact of environmental factors or human activities.

5. **Identification of Pollution Sources:**

   Identify and locate sources of pollution and contamination that may be affecting water quality. This can help in taking corrective actions to mitigate pollution.

6. **Impact Assessment:**

   Assess the impact of water quality on the ecosystem, aquatic life, and public health. Determine how changes in water quality may affect the environment and its sustainability.

7. **Risk Assessment:**

   Evaluate potential health risks associated with the consumption of water from a specific source or exposure to contaminated water. This includes assessing risks from both acute and chronic exposure.

8. **Water Treatment Design and Optimization:**

   Provide data for the design, modification, and optimization of water treatment processes to ensure that water quality meets the required standards for safe consumption.

9. **Public Health Protection:**

   Protect public health by ensuring that water is safe for human consumption. This includes addressing the presence of pathogens, chemicals, and other contaminants that can pose health risks.

2.Design Thinking Process

Design thinking is a problem-solving approach that focuses on understanding and addressing the needs and challenges of the end-users or stakeholders. When applied to a water quality analysis project, the process can help ensure that the project's objectives are aligned with the needs of the community or organization. Here's how you can apply the design thinking process to a water quality analysis project:

1. **Empathize (Understand the Stakeholders):**

    - Identify and engage with all the stakeholders involved in or affected by the water quality analysis project. This may include local communities, environmental organizations, regulatory agencies, and researchers.

    - Conduct interviews, surveys, and meetings to understand the concerns, priorities, and expectations of these stakeholders regarding water quality.

2. **Define (Problem Statement and Objectives):**

    - Based on the insights gained from the empathy phase, define a clear problem statement and objectives for the water quality analysis project. This should be framed in terms of the needs and desires of the stakeholders.

    - Consider regulatory requirements, health concerns, environmental impact, and any other relevant factors.

3. **Ideate (Generate Solutions and Approaches):**

- Brainstorm potential approaches and solutions for addressing the defined problem and achieving the project objectives.

- Encourage creativity and explore various data collection methods, sampling techniques, and analysis tools that can be applied.

4. **Prototype (Develop Data Collection Plan):**

    - Create a prototype of the data collection plan, outlining the methodology and techniques to be used. This plan should address what data to collect, where, and how often.

    - Consider sampling locations, sample size, and the parameters to be measured (e.g., physical, chemical, biological).

5. **Test (Pilot Data Collection):**

    - Implement a pilot data collection phase to test the proposed data collection plan. This will help identify any practical issues, logistical challenges, or adjustments needed in the field.

    - Gather feedback from the field teams, assess the quality of data collected, and refine the plan as necessary.

6. **Feedback and Iterate (Refine the Plan):**

    - Based on the feedback from the testing phase, make iterative improvements to the data collection plan. This may involve adjusting the sampling strategy, equipment, or protocols.

    - Continue to engage with stakeholders to ensure their input is considered in the project's design.

7. **Implement (Execute Data Collection):**

- Execute the data collection plan across the selected sampling locations and time frames. Ensure that the collection process adheres to the refined plan and that data is accurately recorded.

8. **Analyze (Data Analysis and Interpretation):**

    - Analyze the collected data to assess water quality parameters, identify trends, and evaluate compliance with standards and regulations.

    - Consider using statistical analysis, visualization, and modeling techniques to gain insights from the data.

9. **Synthesize (Generate Insights and Recommendations):**

    - Synthesize the findings from the data analysis to generate meaningful insights. Interpret the results in the context of the project objectives and stakeholders' needs.

    - Formulate recommendations for action, such as regulatory changes, remediation efforts, or public health interventions.

2.Development Phase

The development phases of a water quality analysis project typically involve a series of steps that guide the project from inception to completion. These phases help ensure a systematic approach to data collection, analysis, and decision-making. The specific phases may vary depending on the scope and goals of the project, but here is a general outline of the development phases for a water quality analysis project:

1. **Project Initiation:**
    - Define the project's goals and objectives.
    - Identify key stakeholders and establish project teams or partnerships.
    - Secure funding and resources for the project.
    - Develop a project plan outlining timelines, budgets, and deliverables.
2. **Preliminary Research and Data Collection:**
    - Gather existing information and data related to the water source or area of interest. This may include historical water quality data, regulatory requirements, and relevant research.
    - Identify potential sources of contamination or pollution.
    - Establish a baseline understanding of the water body and its surrounding environment.
3. **Sampling Design and Planning:**

- Design a sampling plan that outlines the locations, frequency, and parameters to be measured.
- Consider factors such as the number of sampling points, seasonality, and spatial distribution.
- Determine the appropriate equipment and methods for sample collection.

4. **Data Collection:**
    - Execute the sampling plan by collecting water samples according to established protocols.
    - Record metadata, such as date, time, location, and weather conditions, for each sample.
    - Ensure that all samples are properly labeled and stored.
5. **Laboratory Analysis:**
    - Transport collected water samples to a laboratory for analysis.
    - Conduct a wide range of chemical, physical, and biological tests on the samples.
    - Ensure quality control and data validation during the analysis process.
6. **Data Compilation and Management:**
    - Compile and organize the data collected from laboratory analysis.
    - Ensure that data is properly managed, stored, and documented to maintain data integrity.
7. **Data Analysis and Interpretation:**

   - Analyze the data to assess water quality parameters, identify trends, and compare results to regulatory standards or established benchmarks.

   - Interpret the findings and assess the implications for the water source or area under investigation.

8. **Report and Documentation:**

   - Prepare a comprehensive report summarizing the project's objectives, methods, findings, and recommendations.

   - Clearly communicate the results to stakeholders and decision-makers in a format that is accessible and understandable.

9. **Recommendations and Action Plans:**

   - Formulate specific recommendations based on the analysis and interpretation of the data.

   - Propose action plans to address water quality issues or make improvements.

10. **Implementation of Actions:**

    - Collaborate with relevant authorities, organizations, or communities to implement the recommended actions.

    - Execute necessary remediation efforts, regulatory changes, public health interventions, or other measures.

11. **Monitoring and Ongoing Assessment:**

    - Establish a framework for ongoing monitoring and assessment to ensure that improvements are sustained and that water quality issues are continually addressed.

    - Adapt the monitoring plan based on changes in water quality or environmental conditions.

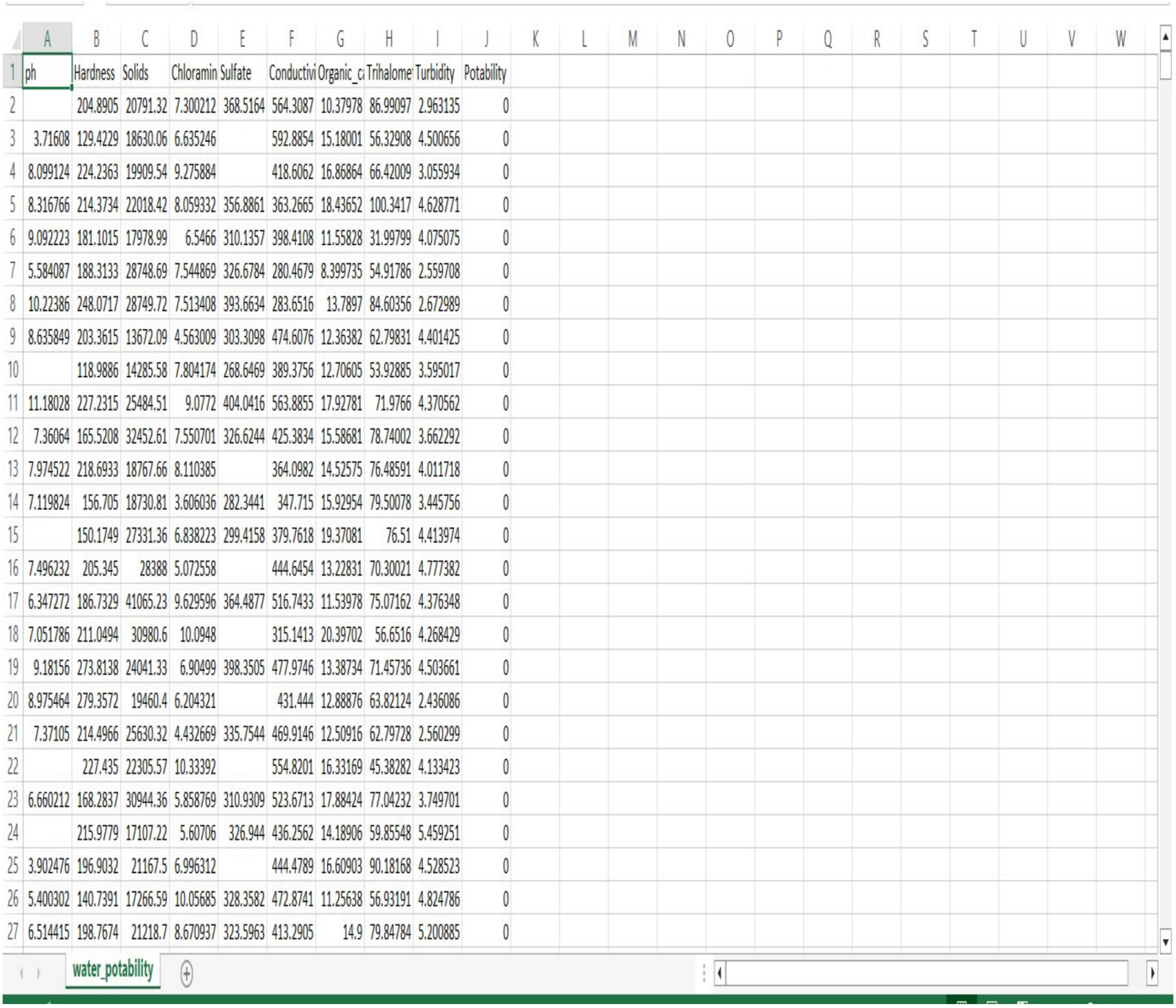
12.**Communication and Outreach:**

- Share project results and actions taken with the broader community, regulatory agencies, and other stakeholders.
- Engage in community outreach and education to raise awareness about water quality issues.

4.Working Process

We'll explore how to import essential libraries, load the housing dataset, and perform critical preprocessing steps. Data preprocessing is crucial as it helps clean, format, and prepare the data for further analysis. This includes handling missing values, encoding categorical variables, and ensuring that the data is appropriately scaled.

Given data set:



	A	B	C	D	E	F	G	H	I	J
1	ph	Hardness	Solids	Chloramin	Sulfate	Conductivi	Organic_ca	Trihalome	Turbidity	Potability
2		204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3	3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
4	8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
5	8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
6	9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
7	5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
8	10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
9	8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
10		118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11	11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
12	7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
13	7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
14	7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
15		150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
16	7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
17	6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
18	7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
19	9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
20	8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0
21	7.37105	214.4966	25630.32	4.432669	335.7544	469.9146	12.50916	62.79728	2.560299	0
22		227.435	22305.57	10.33392		554.8201	16.33169	45.38282	4.133423	0
23	6.660212	168.2837	30944.36	5.858769	310.9309	523.6713	17.88424	77.04232	3.749701	0
24		215.9779	17107.22	5.60706	326.944	436.2562	14.18906	59.85548	5.459251	0
25	3.902476	196.9032	21167.5	6.996312		444.4789	16.60903	90.18168	4.528523	0
26	5.400302	140.7391	17266.59	10.05685	328.3582	472.8741	11.25638	56.93191	4.824786	0
27	6.514415	198.7674	21218.7	8.670937	323.5963	413.2905	14.9	79.84784	5.200885	0

Necessary step to follow:

1.Import libaries

Start by importing the necessary libraries:

Program:

import

pandas as

pd import

numpy as

np

import seaborn as sb

import

matplotlib.pyplot as

plt import

plotly.express as px

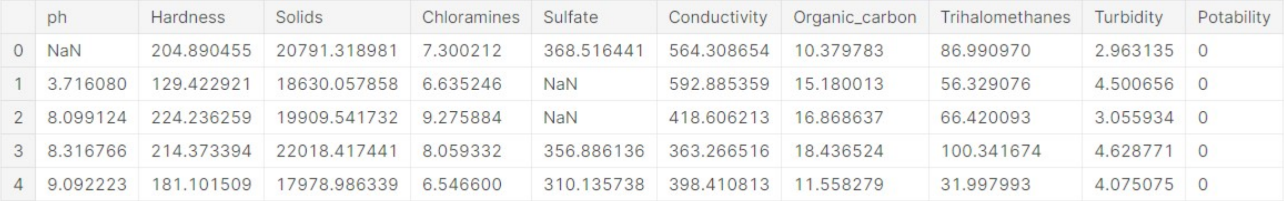
2. Read Dataset

Load your dataset into a Pandas DataFrame. You can typically find water potability datasets in CSV format, but you can adapt this code to other formats as needed.

df =

pd.read_csv('C:/Users/HP/OneDrive/Documents/water

quality analysis/water_potability.csv') df.head() output:



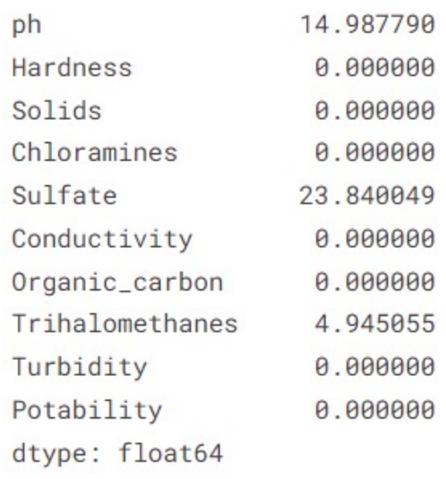
	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability
0	NaN	204.890455	20791.318981	7.300212	368.516441	564.308654	10.379783	86.990970	2.963135	0
1	3.716080	129.422921	18630.057858	6.635246	NaN	592.885359	15.180013	56.329076	4.500656	0
2	8.099124	224.236259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

3. Exploratory Data Analysis (EDA)

Perform EDA to understand your data better. This includes checking for missing values, exploring the data's statistics, andvisualizing it to identify patterns.

(df.isnull().sum()/df.shape[0])*100

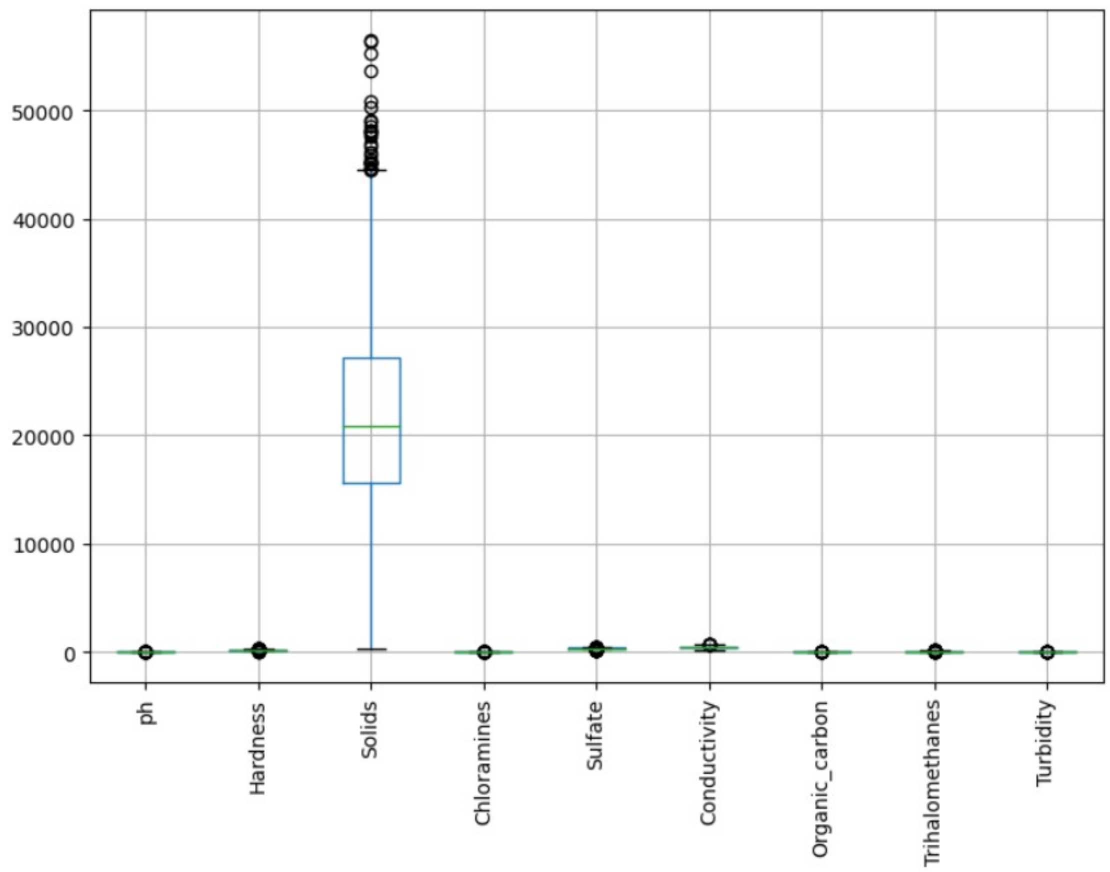
Output:



```
ph                 14.987790
Hardness            0.000000
Solids              0.000000
Chloramines         0.000000
Sulfate            23.840049
Conductivity        0.000000
Organic_carbon      0.000000
Trihalomethanes     4.945055
Turbidity           0.000000
Potability          0.000000
dtype: float64
```

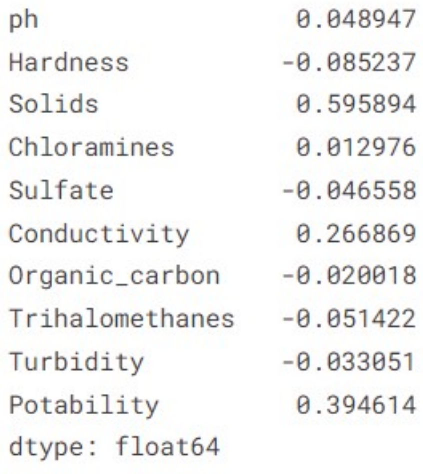
```
df.iloc[:,0:-1].boxplot()
plt.xticks(rotation=90)
plt.show()
```

Output:



df.skew()

Output:



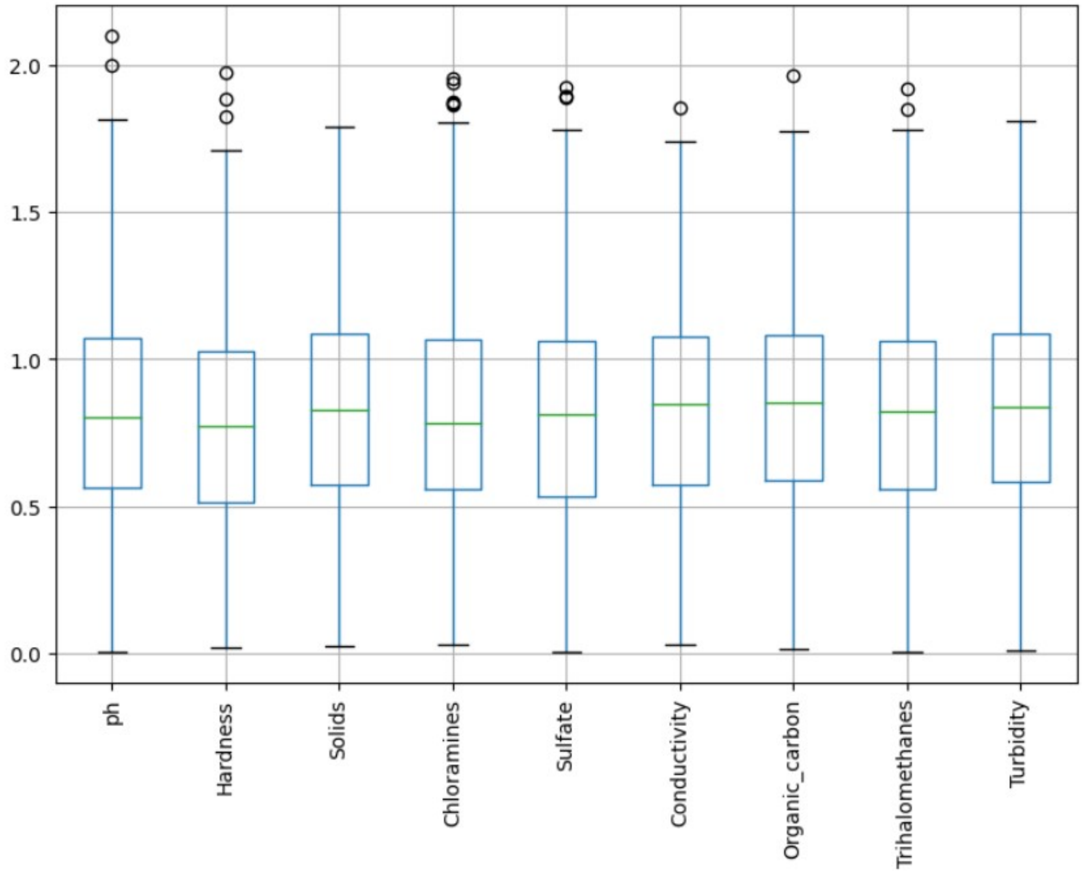
```
ph                 0.048947
Hardness          -0.085237
Solids             0.595894
Chloramines        0.012976
Sulfate           -0.046558
Conductivity       0.266869
Organic_carbon    -0.020018
Trihalomethanes   -0.051422
Turbidity         -0.033051
Potability         0.394614
dtype: float64
```

np.sqrt(df.iloc[:,0:-

1]).boxplot()

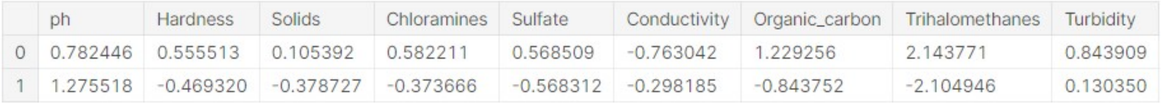
```
plt.xticks(rotation=90)

plt.show()
```

Output:



```
x = df[cols]

x.head(2)
```

Output:

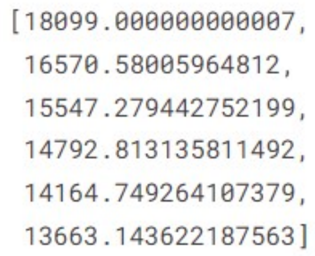


	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity
0	0.782446	0.555513	0.105392	0.582211	0.568509	-0.763042	1.229256	2.143771	0.843909
1	1.275518	-0.469320	-0.378727	-0.373666	-0.568312	-0.298185	-0.843752	-2.104946	0.130350

```
from sklearn.cluster import

KMeans max_range=7
```

```
wcss = [] for k in
range(1,max_range):
kmeans =
KMeans(n_clusters=k,rand
om_state=42)     model =
kmeans.fit(x)
wcss.append(kmeans.inerti
a_) wcss
```

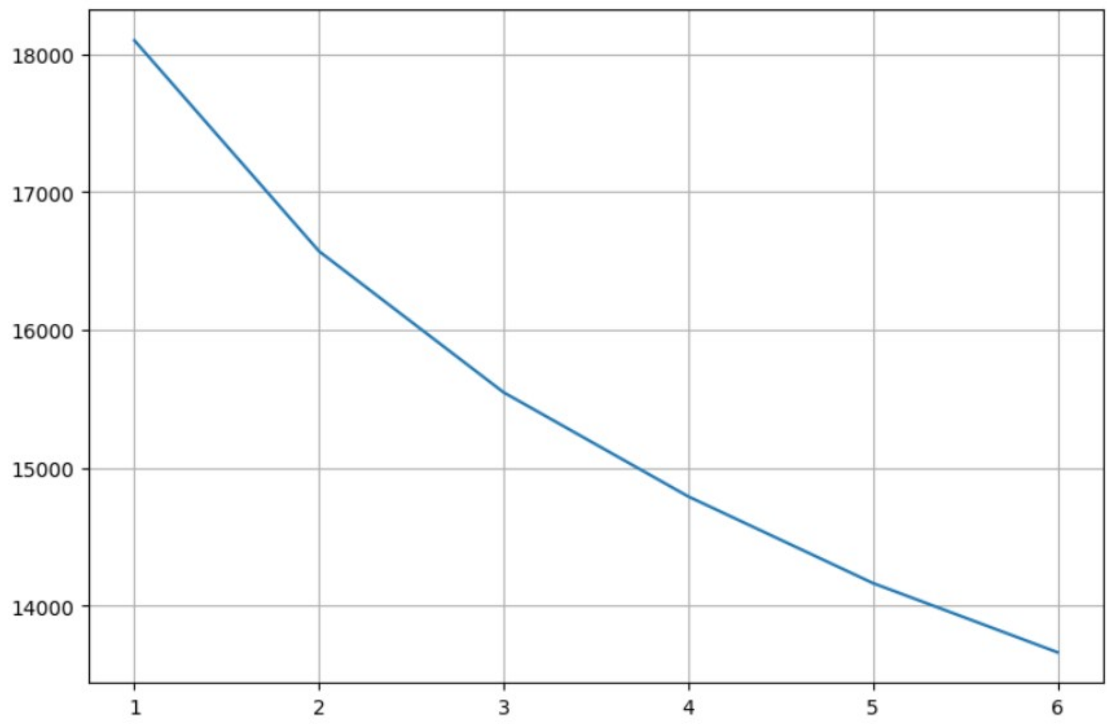
Output:



```
[18099.000000000007,
 16570.58005964812,
 15547.279442752199,
 14792.813135811492,
 14164.749264107379,
 13663.143622187563]
```

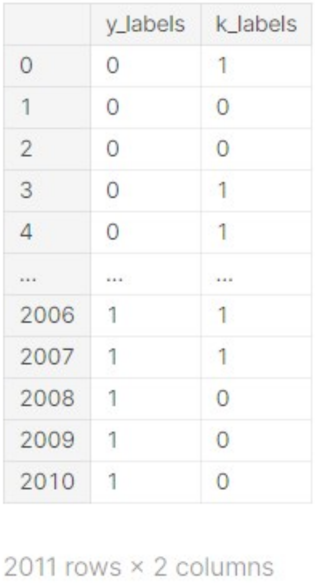
```
plt.plot(range(1,max_ran
ge),wcss) plt.grid()
plt.show()
```

Output:



```
kmeans = KMeans(n_clusters=2,random_state=42) model =

kmeans.fit(x) model.labels_

compare=pd.DataFrame({'y_labels':df['Potability'],'k_labels':mod

el.labels_

})

compare
```

Output:

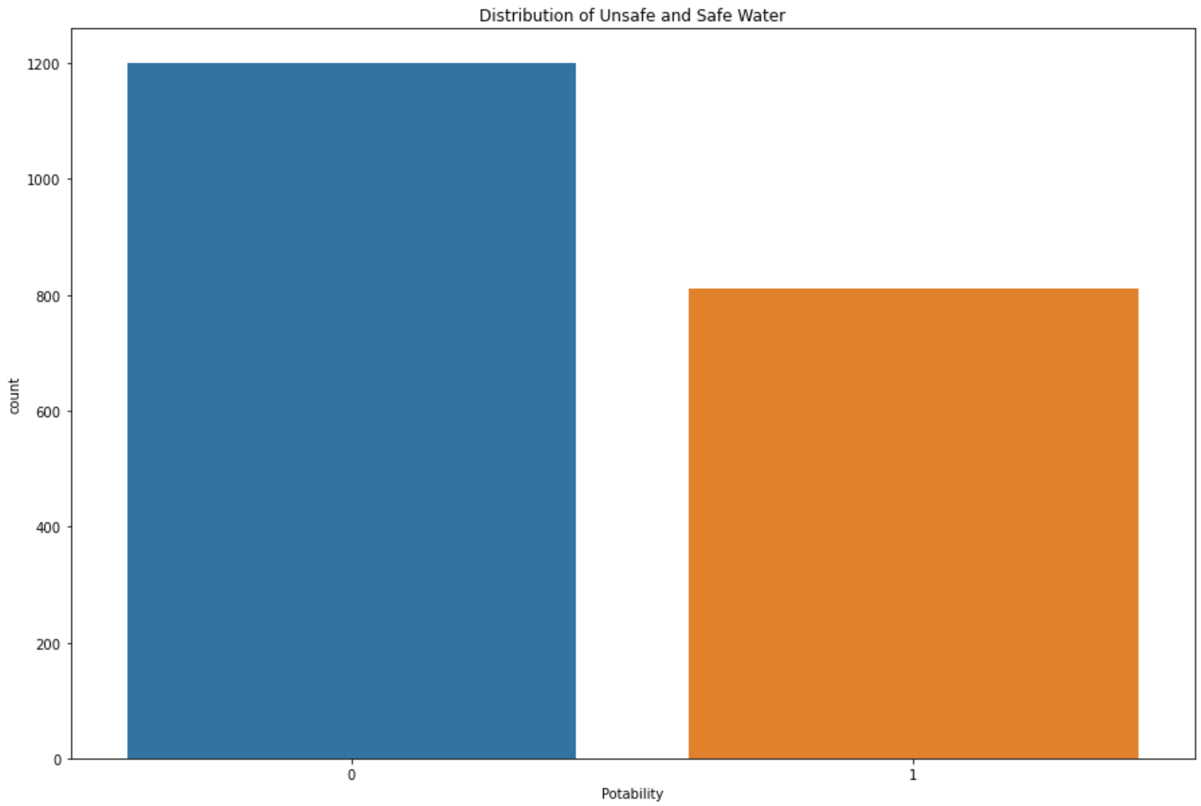


	y_labels	k_labels
0	0	1
1	0	0
2	0	0
3	0	1
4	0	1
...	...	...
2006	1	1
2007	1	1
2008	1	0
2009	1	0
2010	1	0

2011 rows × 2 columns

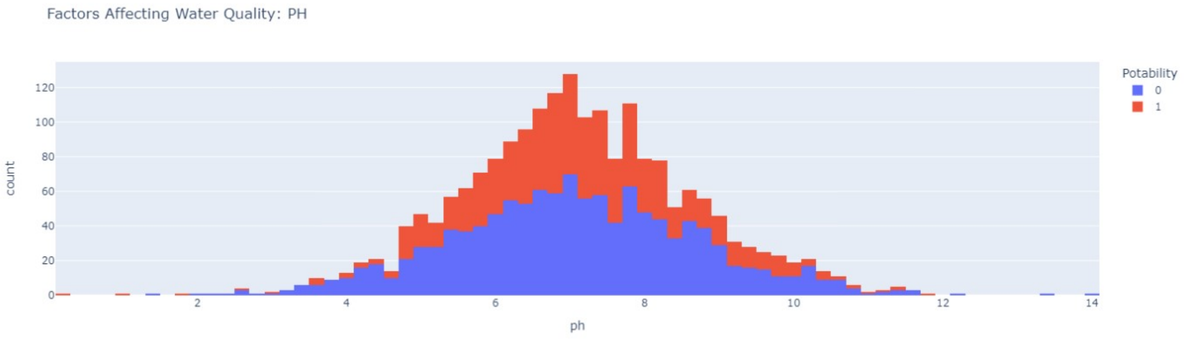
```
plt.figure(figsize=(15, 10))

sns.countplot(data.Potability)

plt.title("Distribution of Unsafe and Safe Water")

plt.show()
```

Output:



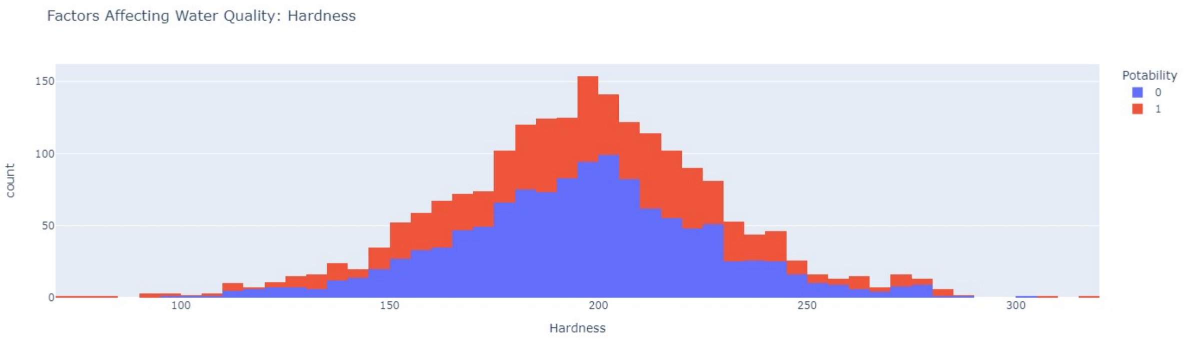
```
import   plotly.express

as px data = data

figure = px.histogram(data, x = "ph",

                    color         =         "Potability",

title=  "Factors  Affecting  Water  Quality:  PH")

figure.show()
```

Output:



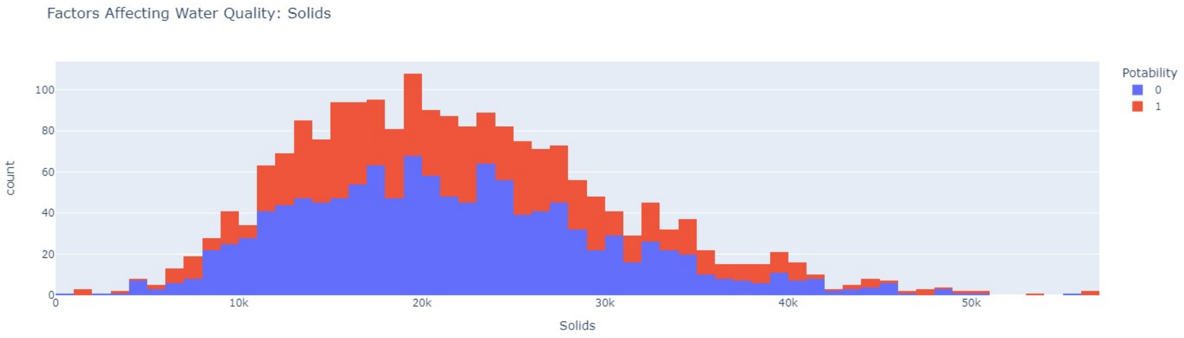
```
import   plotly.express

as px data = data

figure = px.histogram(data, x = "Hardness",

                    color = "Potability",                    title=

"Factors    Affecting    Water    Quality:    Hardness")

figure.show()
```

Output:



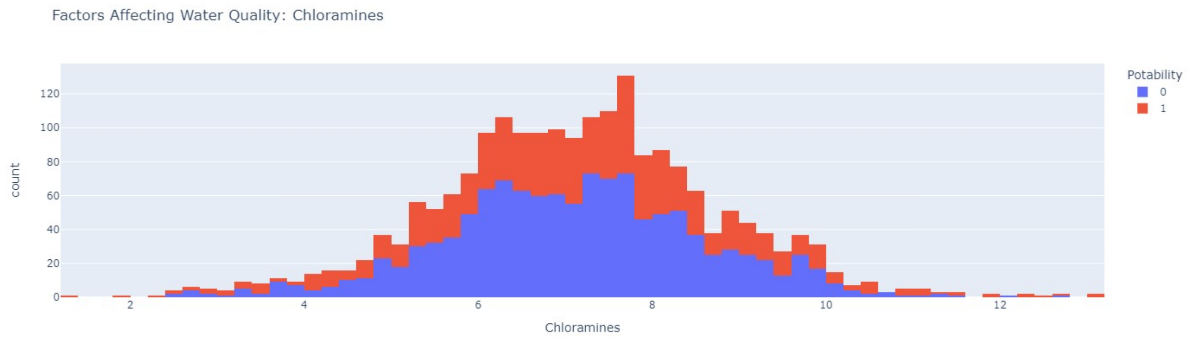
```
import plotly.express as px data

= data figure =

px.histogram(data, x =

"Solids",
color = "Potability",                    title=
"Factors     Affecting     Water     Quality:     Solids")
figure.show()
```

Output:



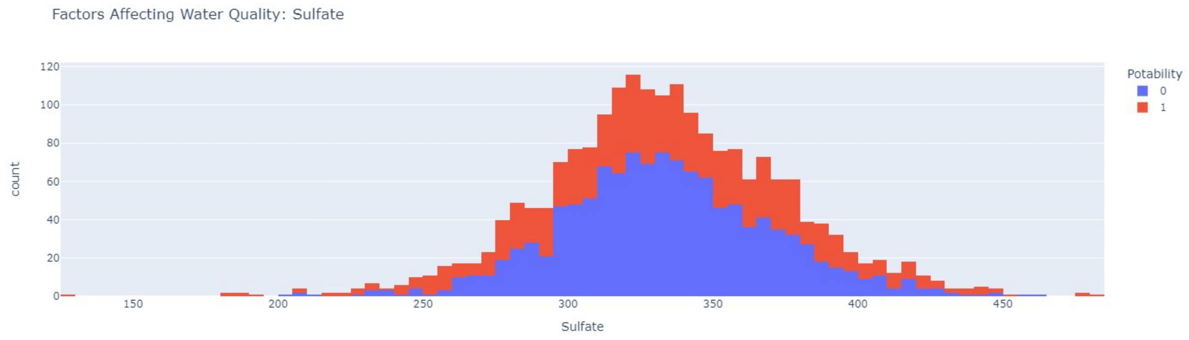
```
import    plotly.express
as px data = data
figure = px.histogram(data, x = "Chloramines",
                    color = "Potability",                              title=
"Factors     Affecting     Water     Quality:     Chloramines")
figure.show()
```

Output:



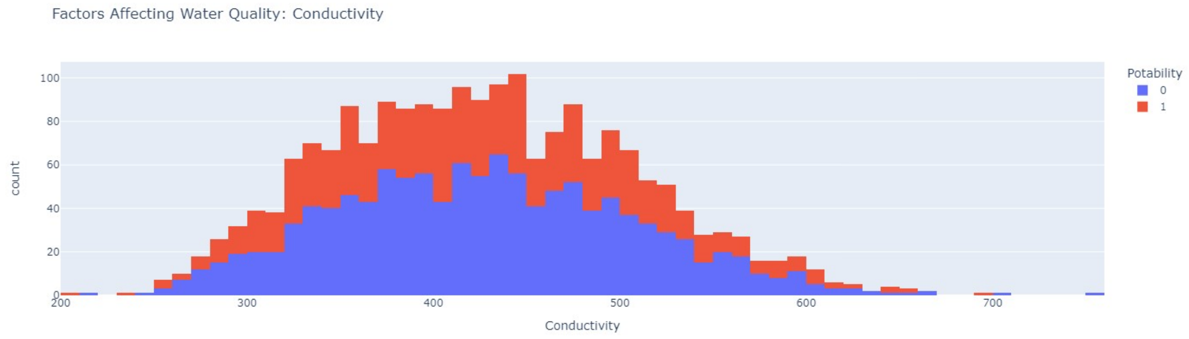
```
import plotly.express as px

data = data

figure = px.histogram(data, x = "Sulfate",

                color = "Potability",          title=

"Factors   Affecting   Water   Quality:   Sulfate")

figure.show()
```

Output:



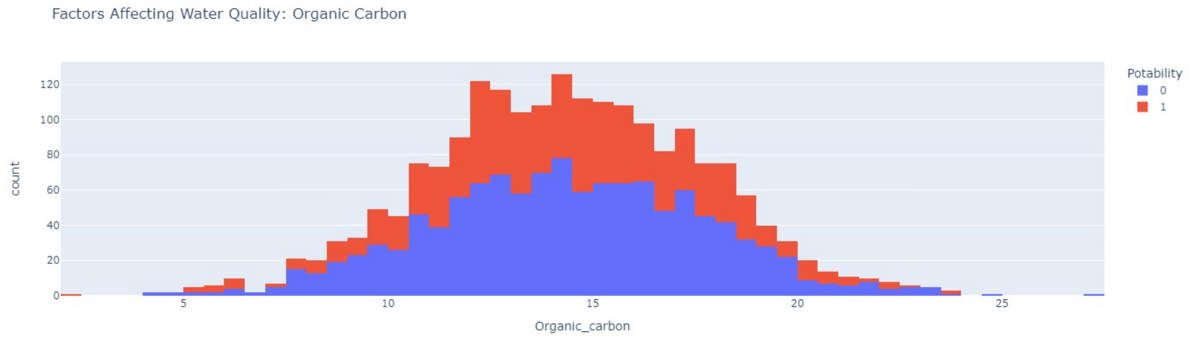
```
import   plotly.express

as px data = data

figure = px.histogram(data, x = "Conductivity",
                    color = "Potability",                          title=
"Factors   Affecting   Water   Quality:   Conductivity")
figure.show()
```

Output:



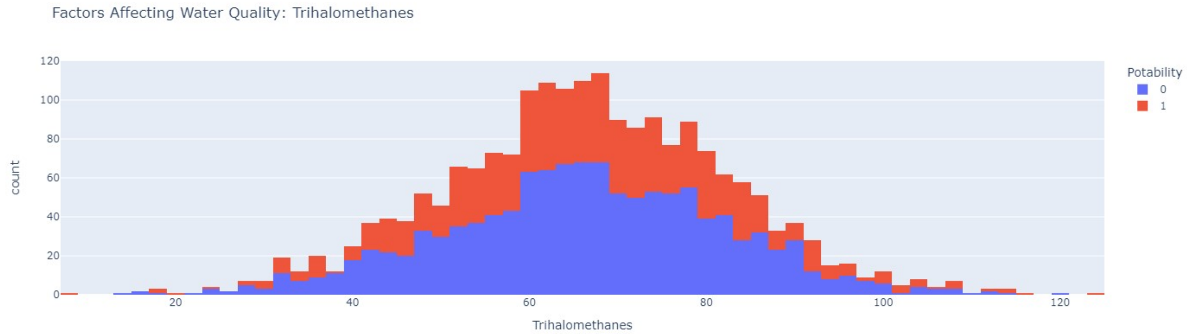
```
import   plotly.express
as px data = data
figure = px.histogram(data, x = "Organic_carbon",
                    color = "Potability",                 title= "Factors
Affecting Water Quality: Organic Carbon") figure.show()
```

Output:



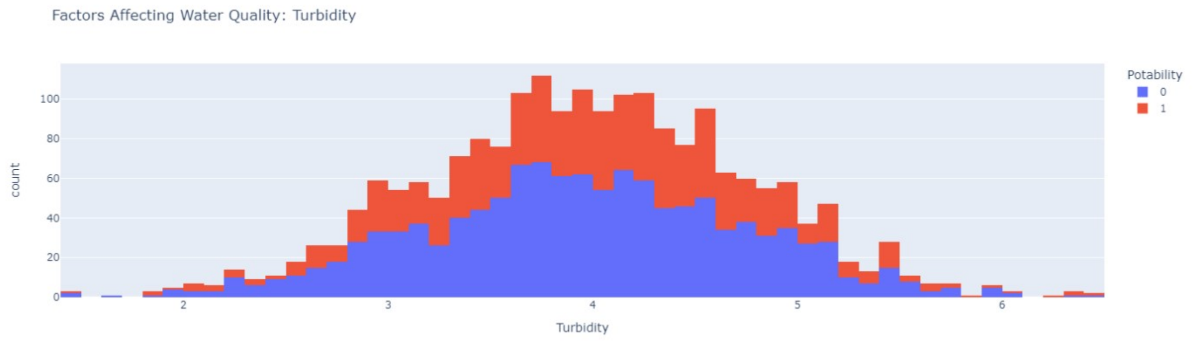
```
import    plotly.express

as px data = data

figure = px.histogram(data, x = "Trihalomethanes",

                        color = "Potability",                         title= "Factors

Affecting Water Quality: Trihalomethanes") figure.show()
```

Output:



```
 import

plotly.express as

px data = data

figure = px.histogram(data, x = "Turbidity",

                        color = "Potability",                            title=

"Factors     Affecting     Water     Quality:     Turbidity")

figure.show()
```

Output:

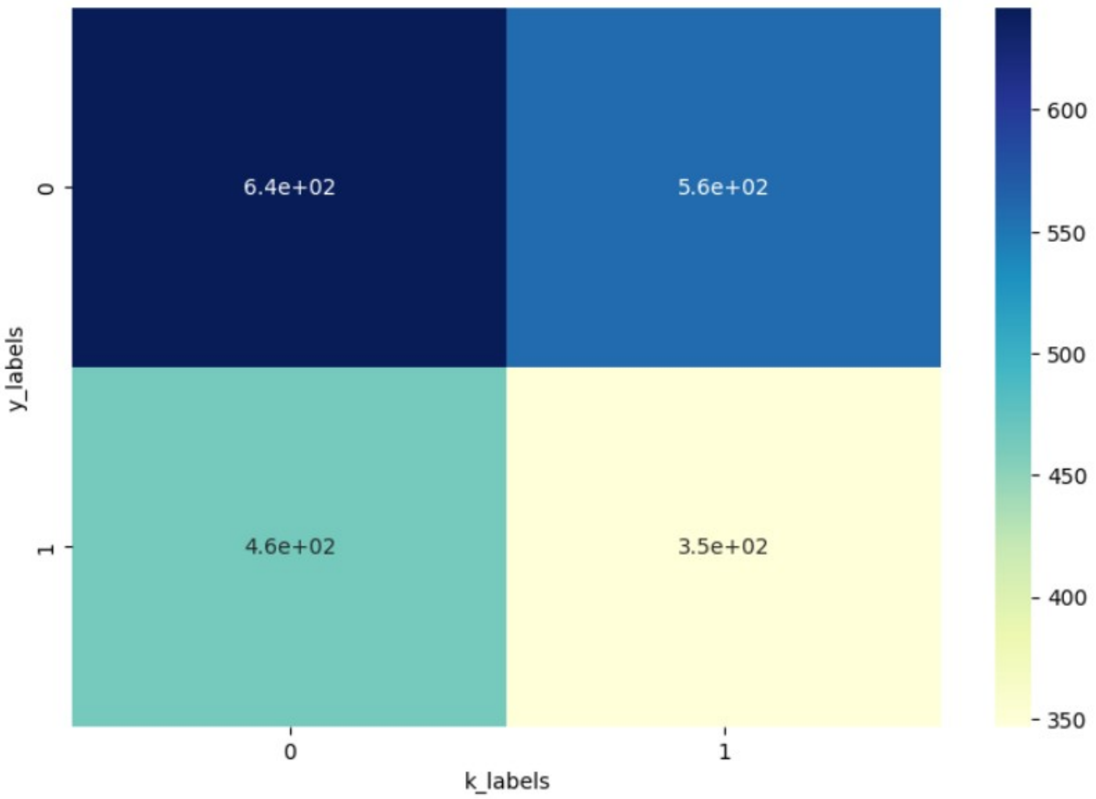


4. Split the Data

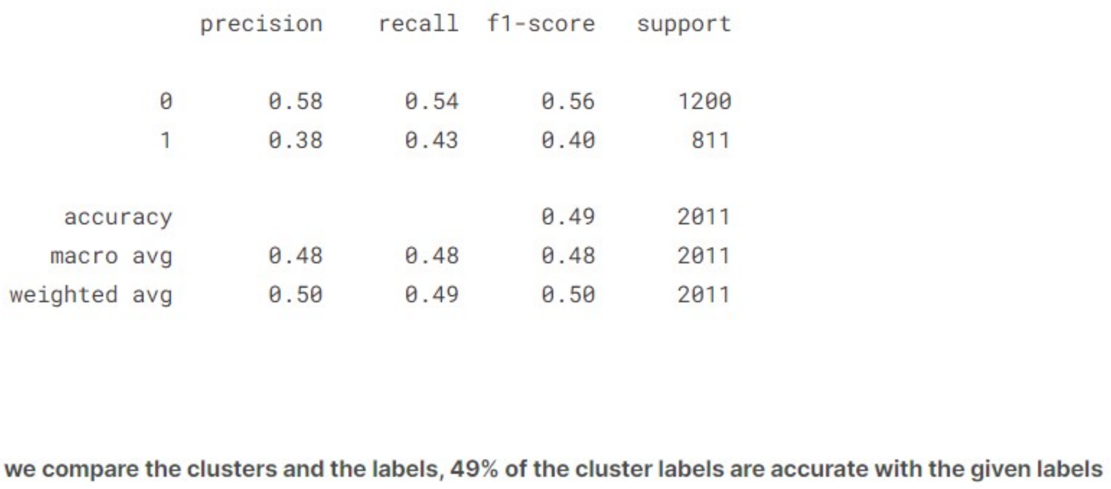
Split your dataset into training and testing sets. This helps you evaluateyour model's performance later.

```
sns.heatmap(pd.crosstab(compare['y_labels'],compare['k_labels']),
annot=T rue,cmap='YlGnBu')

plt.show()
```

Output:



```
from sklearn.metrics import accuracy_score, precision_score,
recall_score, f1_score,classification_report

accuracy_score(compare['y_labels'],compare['k_labels'])

recall_score(compare['y_labels'],compare['k_labels'])


print(classification_report(compare['y_labels'],compare['k_labels'

])) Output:
```

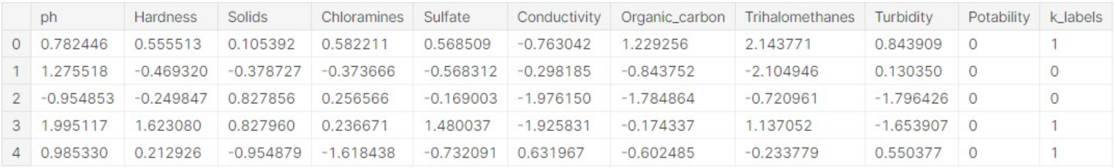


```
              precision    recall  f1-score   support

           0       0.58      0.54      0.56      1200
           1       0.38      0.43      0.40       811

    accuracy                           0.49      2011
   macro avg       0.48      0.48      0.48      2011
weighted avg       0.50      0.49      0.50      2011
```

If we compare the clusters and the labels, 49% of the cluster labels are accurate with the given labels

```
df['k_labels'] = pd.Series(model.labels_)
df.head()
```

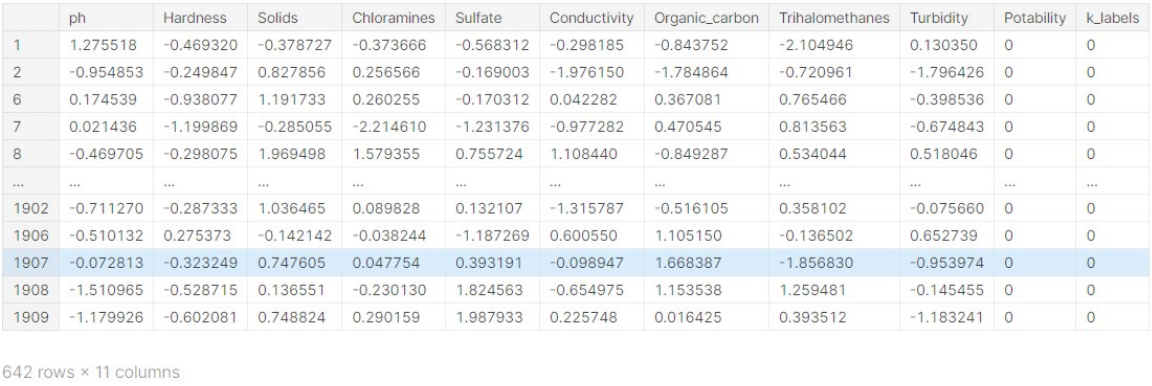
Output:



	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	k_labels
0	0.782446	0.555513	0.105392	0.582211	0.568509	-0.763042	1.229256	2.143771	0.843909	0	1
1	1.275518	-0.469320	-0.378727	-0.373666	-0.568312	-0.298185	-0.843752	-2.104946	0.130350	0	0
2	-0.954853	-0.249847	0.827856	0.256566	-0.169003	-1.976150	-1.784864	-0.720961	-1.796426	0	0
3	1.995117	1.623080	0.827960	0.236671	1.480037	-1.925831	-0.174337	1.137052	-1.653907	0	1
4	0.985330	0.212926	-0.954879	-1.618438	-0.732091	0.631967	-0.602485	-0.233779	0.550377	0	1

```
df[((df['Potability']==0)&(df['k_labels']==0))]
```

Output:

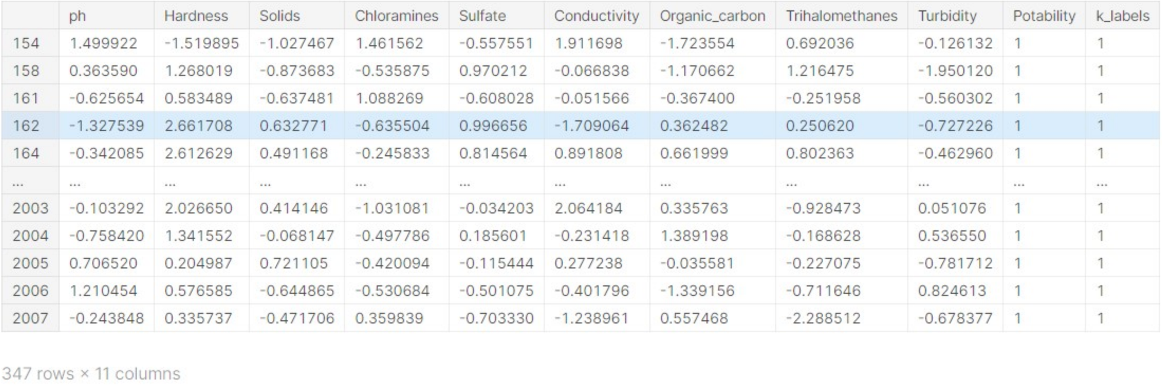


	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	k_labels
1	1.275518	-0.469320	-0.378727	-0.373666	-0.568312	-0.298185	-0.843752	-2.104946	0.130350	0	0
2	-0.954853	-0.249847	0.827856	0.256566	-0.169003	-1.976150	-1.784864	-0.720961	-1.796426	0	0
6	0.174539	-0.938077	1.191733	0.260255	-0.170312	0.042282	0.367081	0.765466	-0.398536	0	0
7	0.021436	-1.199869	-0.285055	-2.214610	-1.231376	-0.977282	0.470545	0.813563	-0.674843	0	0
8	-0.469705	-0.298075	1.969498	1.579355	0.755724	1.108440	-0.849287	0.534044	0.518046	0	0
...	...	...	...	...	...	...	...	...	...	...	...
1902	-0.711270	-0.287333	1.036465	0.089828	0.132107	-1.315787	-0.516105	0.358102	-0.075660	0	0
1906	-0.510132	0.275373	-0.142142	-0.038244	-1.187269	0.600550	1.105150	-0.136502	0.652739	0	0
1907	-0.072813	-0.323249	0.747605	0.047754	0.393191	-0.098947	1.668387	-1.856830	-0.953974	0	0
1908	-1.510965	-0.528715	0.136551	-0.230130	1.824563	-0.654975	1.153538	1.259481	-0.145455	0	0
1909	-1.179926	-0.602081	0.748824	0.290159	1.987933	0.225748	0.016425	0.393512	-1.183241	0	0

642 rows × 11 columns

```
df[((df['Potability']==1)&(df['k_labels']==1))]
```

Output:

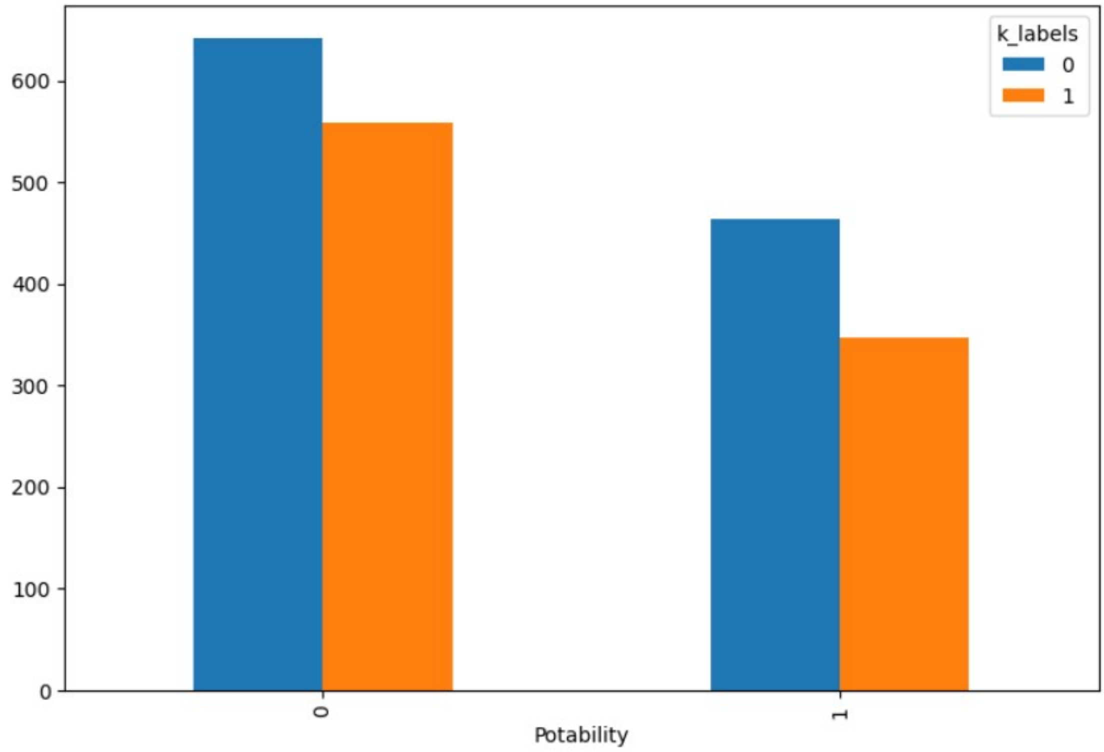


	ph	Hardness	Solids	Chloramines	Sulfate	Conductivity	Organic_carbon	Trihalomethanes	Turbidity	Potability	k_labels
154	1.499922	-1.519895	-1.027467	1.461562	-0.557551	1.911698	-1.723554	0.692036	-0.126132	1	1
158	0.363590	1.268019	-0.873683	-0.535875	0.970212	-0.066838	-1.170662	1.216475	-1.950120	1	1
161	-0.625654	0.583489	-0.637481	1.088269	-0.608028	-0.051566	-0.367400	-0.251958	-0.560302	1	1
162	-1.327539	2.661708	0.632771	-0.635504	0.996656	-1.709064	0.362482	0.250620	-0.727226	1	1
164	-0.342085	2.612629	0.491168	-0.245833	0.814564	0.891808	0.661999	0.802363	-0.462960	1	1
...	...	...	...	...	...	...	...	...	...	...	...
2003	-0.103292	2.026650	0.414146	-1.031081	-0.034203	2.064184	0.335763	-0.928473	0.051076	1	1
2004	-0.758420	1.341552	-0.068147	-0.497786	0.185601	-0.231418	1.389198	-0.168628	0.536550	1	1
2005	0.706520	0.204987	0.721105	-0.420094	-0.115444	0.277238	-0.035581	-0.227075	-0.781712	1	1
2006	1.210454	0.576585	-0.644865	-0.530684	-0.501075	-0.401796	-1.339156	-0.711646	0.824613	1	1
2007	-0.243848	0.335737	-0.471706	0.359839	-0.703330	-1.238961	0.557468	-2.288512	-0.678377	1	1

347 rows × 11 columns

```
pd.crosstab(df['Potability'],df['k_labels']).plot(ki
nd='bar') plt.show()
```

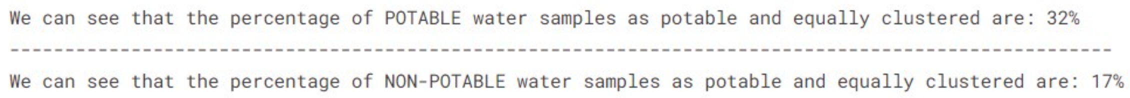
Output:



5. Final Result

```
a=(642/df.shape[0])*100

b=(347/df.shape[0])*100

print("We can see that the percentage of POTABLE water
samples as
potable and equally clustered are: {}%".format(round(a),2))

print("-"*100)

print("We can see that the percentage of NON-POTABLE water
samples as potable and equally clustered are:
{}%".format(round(b),2))
```

Output:



```
We can see that the percentage of POTABLE water samples as potable and equally clustered are: 32%
----------------------------------------------------------------------------------------------------
We can see that the percentage of NON-POTABLE water samples as potable and equally clustered are: 17%
```