**Water Quality Analysis Using Data Analytics With Cognos**

**Phase 5 Final submission document**

BATCH MEMBERS

Submitted by

Yuvaraj V 2021115125

Shadhani TT 2021115302

Adnan Khan S 2021115304

Vishal Raj Vellaisamy 2021115120

Vikhram SS 2021115119

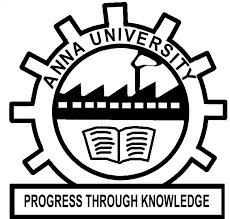
Sabarisrinath R 2021115335

***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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.

1. **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.

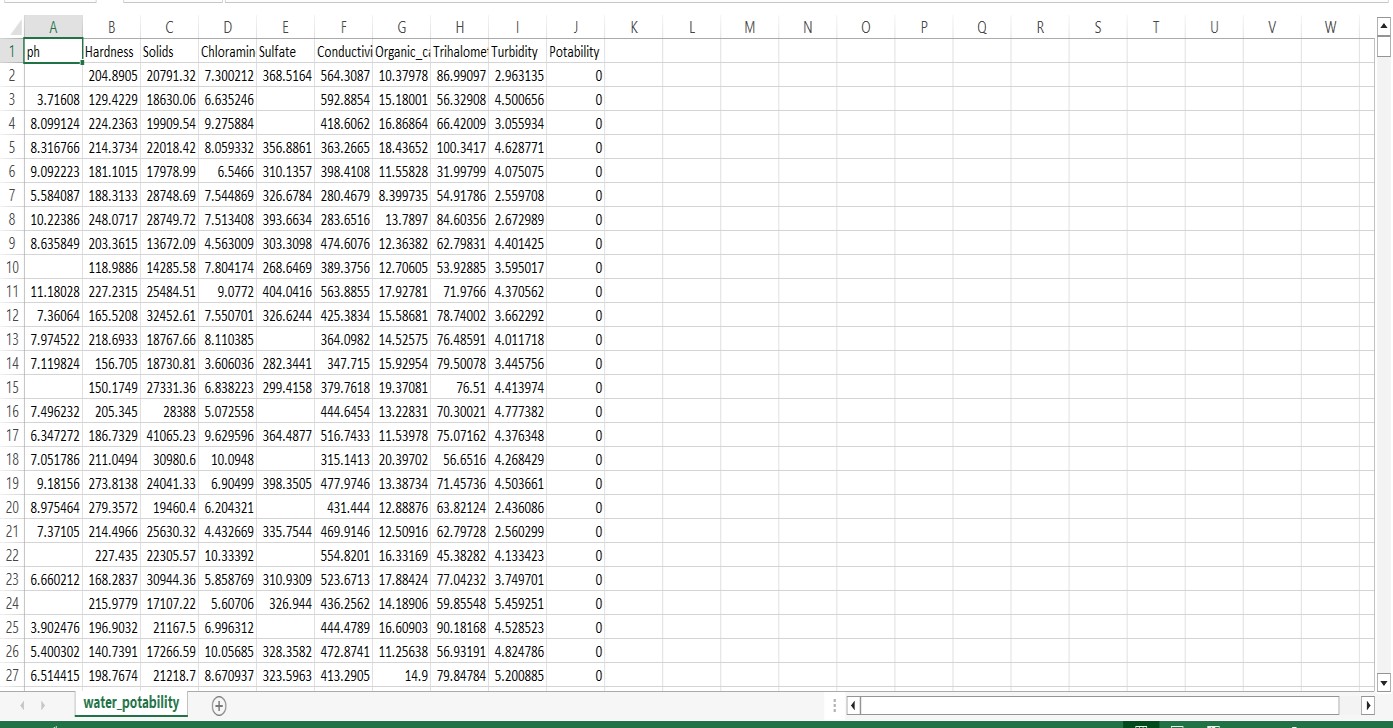
12.**Communication and Outreach:**

* + 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:



Necessary step to follow:

1.Import libaries

Start by importing the necessary libraries:

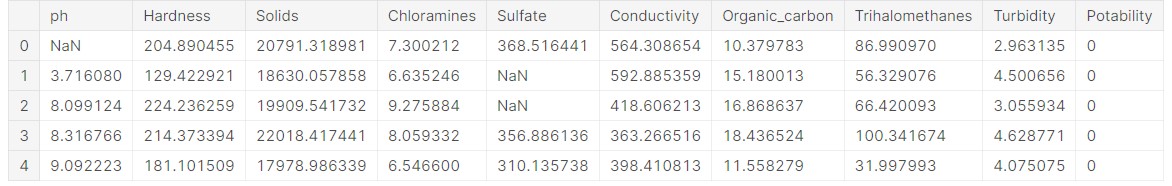
Program: import pandas as pd import numpy as np

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.

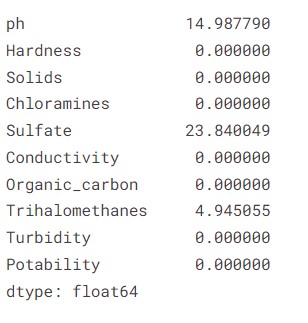
df = pd.read\_csv('C:/Users/HP/OneDrive/Documents/water quality analysis/water\_potability.csv') df.head() output:



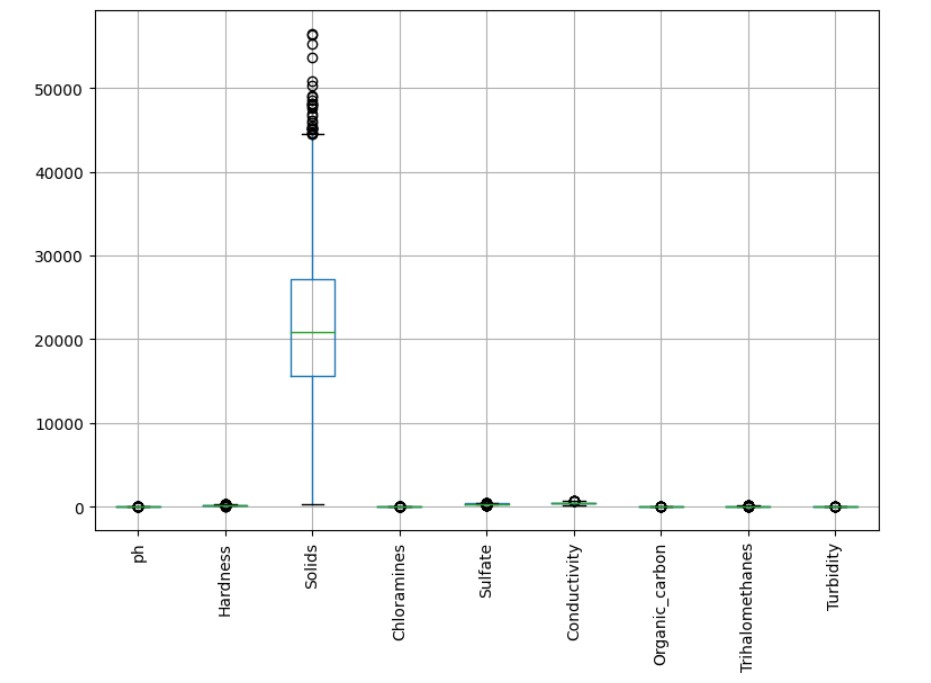
# Exploratory Data Analysis (EDA)

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:



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



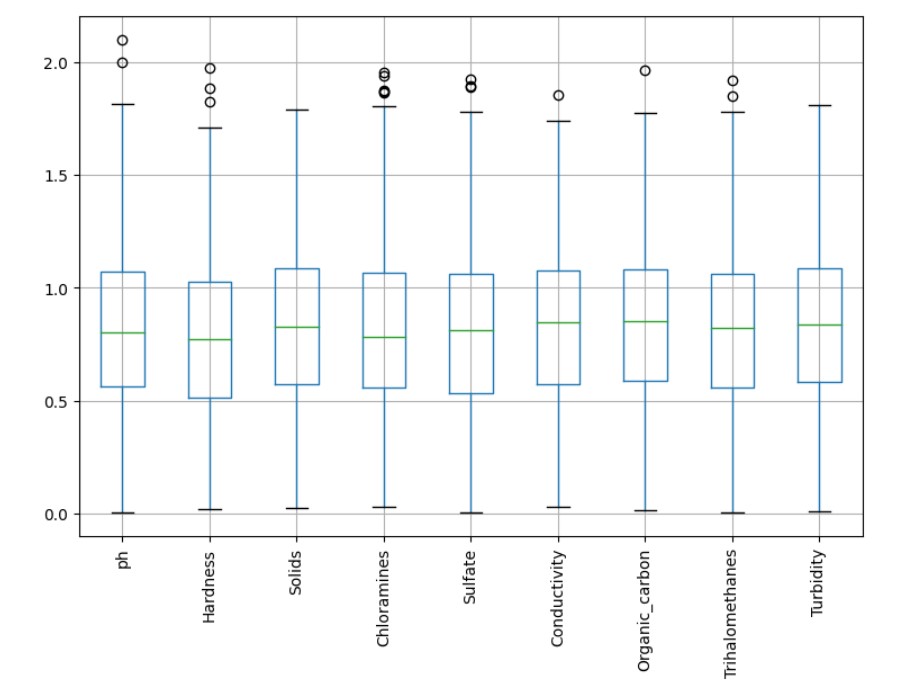
df.skew()

Output:



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

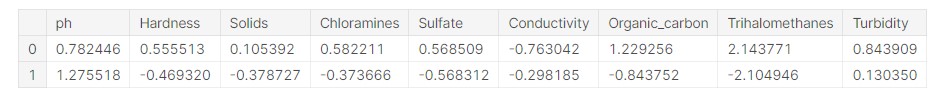
Output:



x = df[cols]

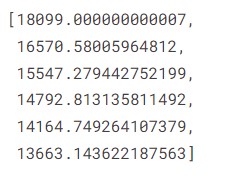
x.head(2)

Output:



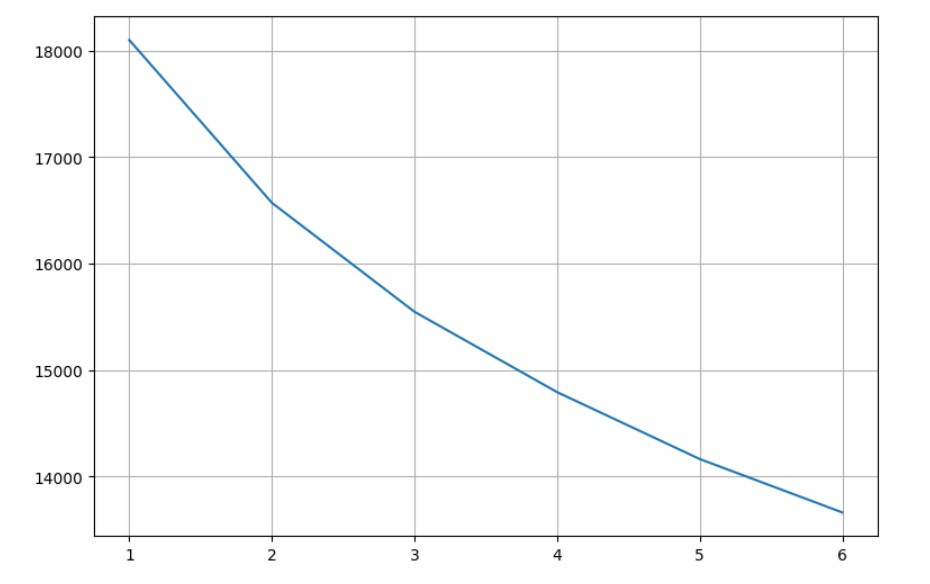
from sklearn.cluster import KMeans max\_range=7 wcss = [] for k **in** range(1,max\_range): kmeans = KMeans(n\_clusters=k,random\_state=42) model = kmeans.fit(x) wcss.append(kmeans.inertia\_) wcss

Output:



plt.plot(range(1,max\_range),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':model.labels\_

})

compare

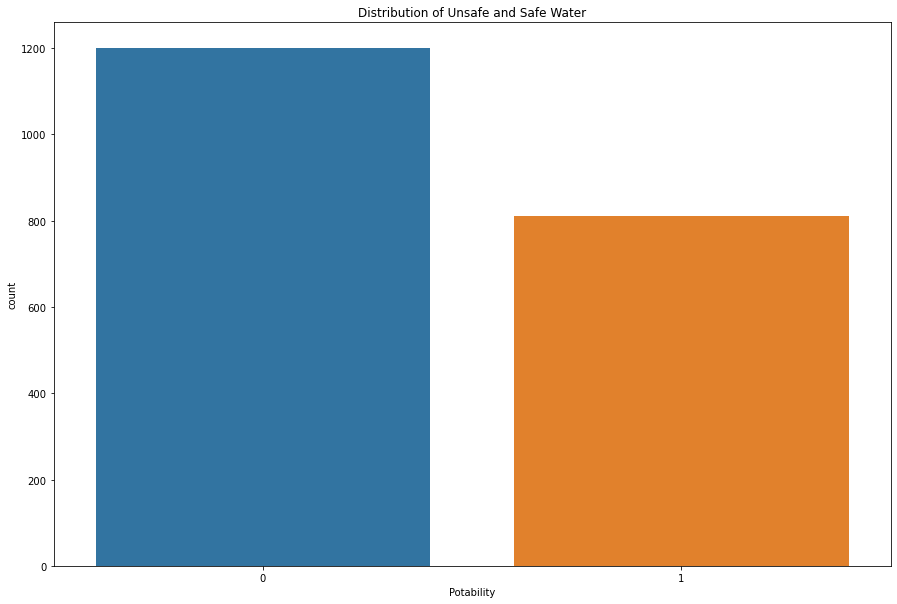
Output:



plt.figure(figsize=(15, 10)) sns.countplot(data.Potability)

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

Output:

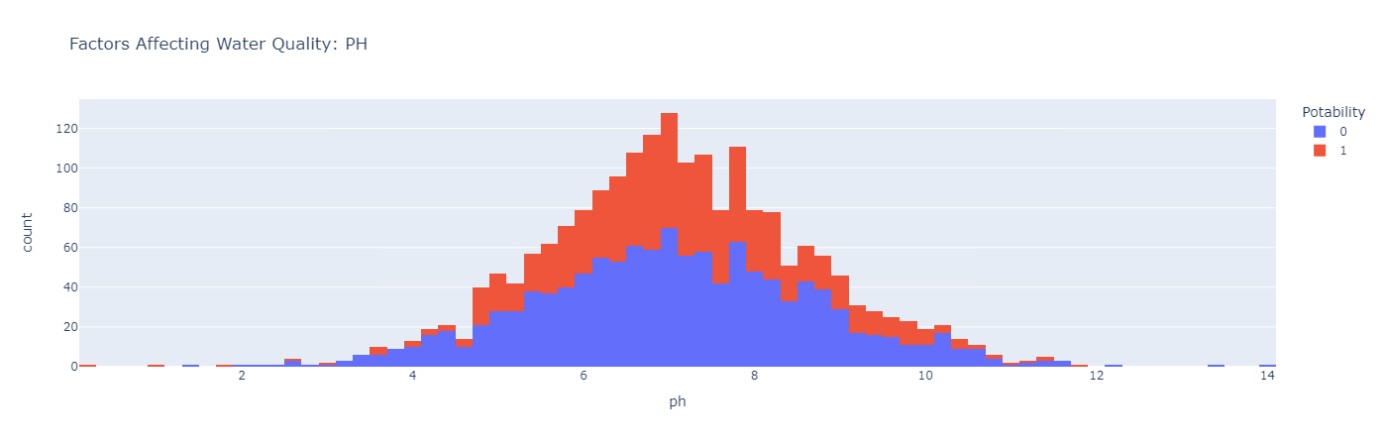


import plotly.express as px data = data

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

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

Output:

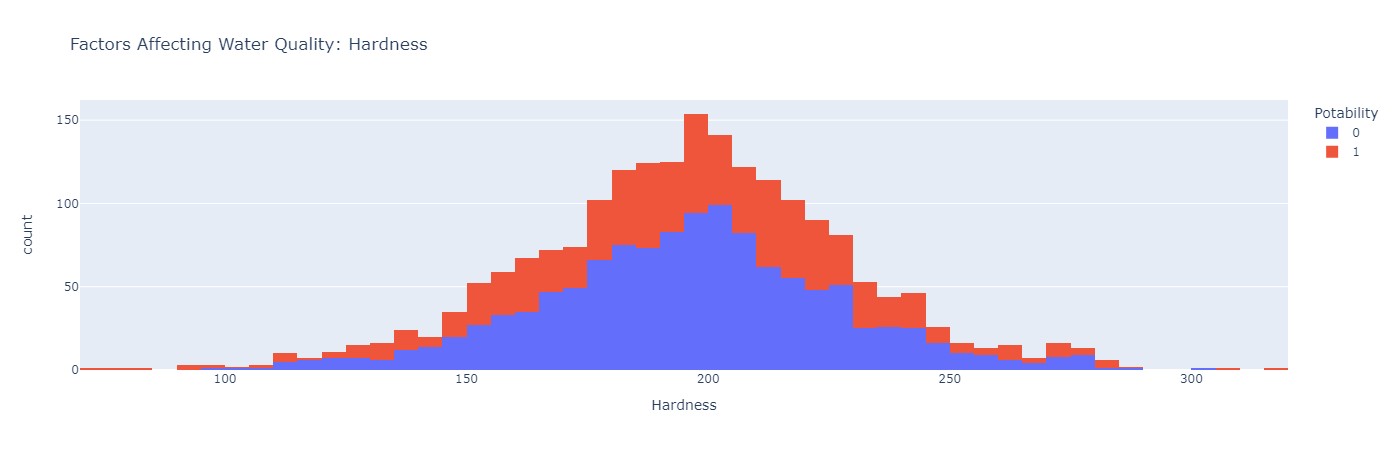


import plotly.express as px data = data

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

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

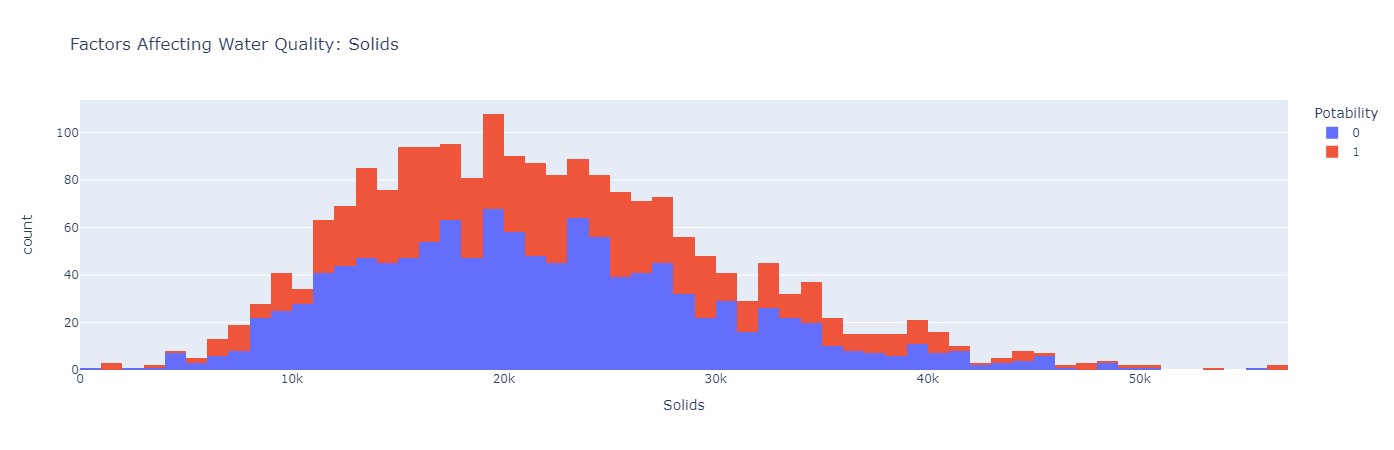
Output:



import plotly.express as px data = data figure = px.histogram(data, x = "Solids",

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

Output:

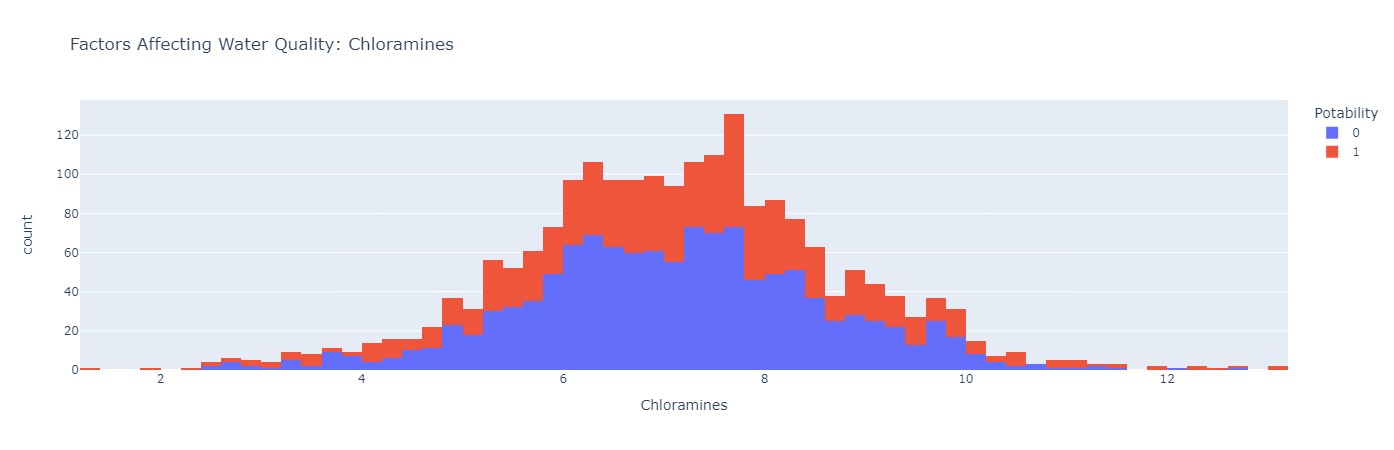


import plotly.express as px data = data

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

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

Output:



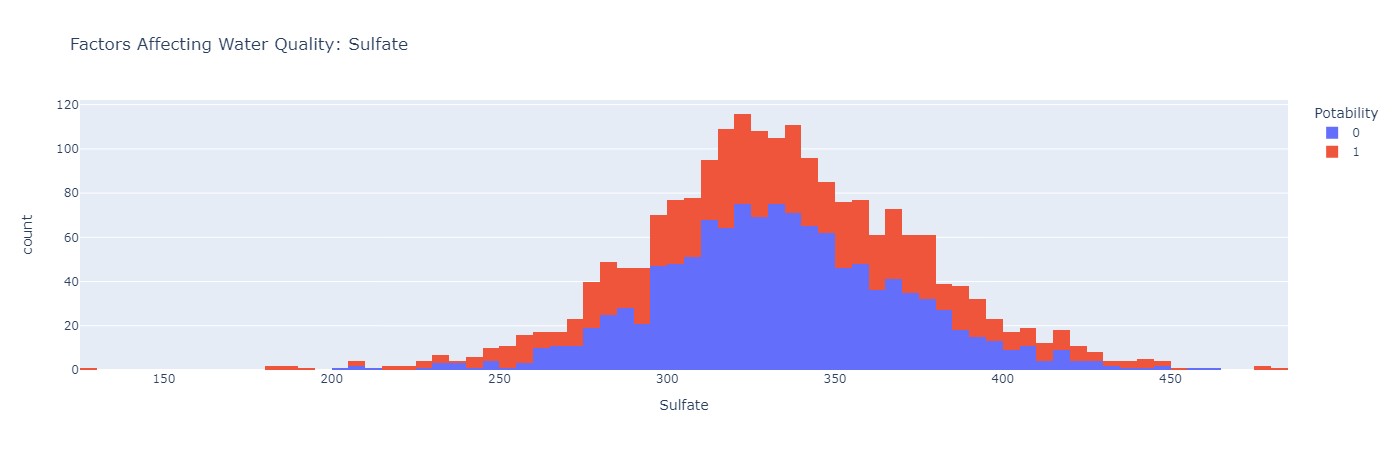
import plotly.express as px

data = data

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

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

Output:

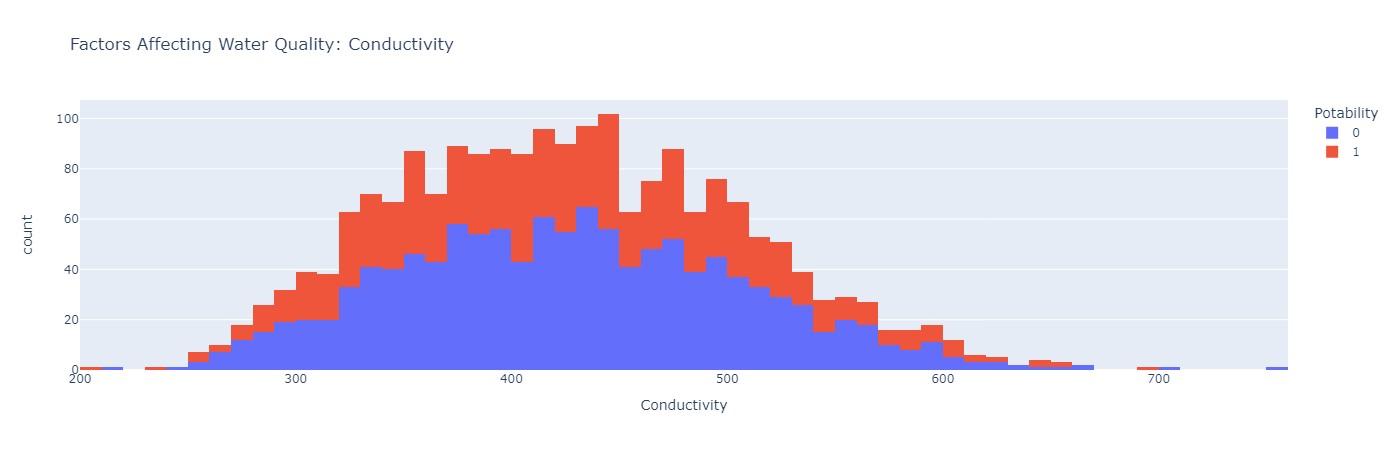


import plotly.express as px data = data

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

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

Output:

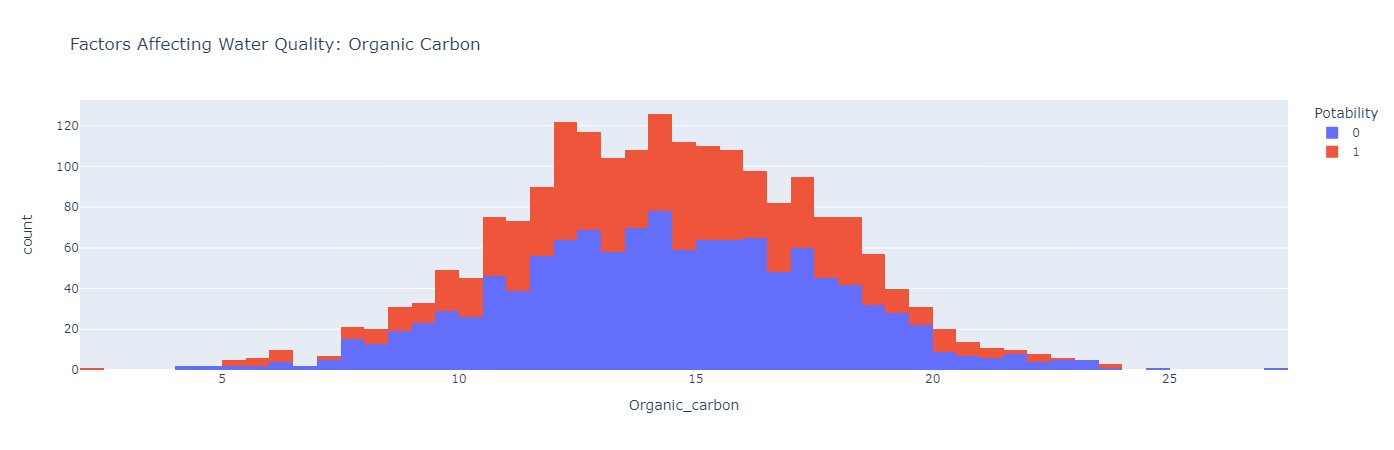


import plotly.express as px data = data

figure = px.histogram(data, x = "Organic\_carbon",

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

Output:

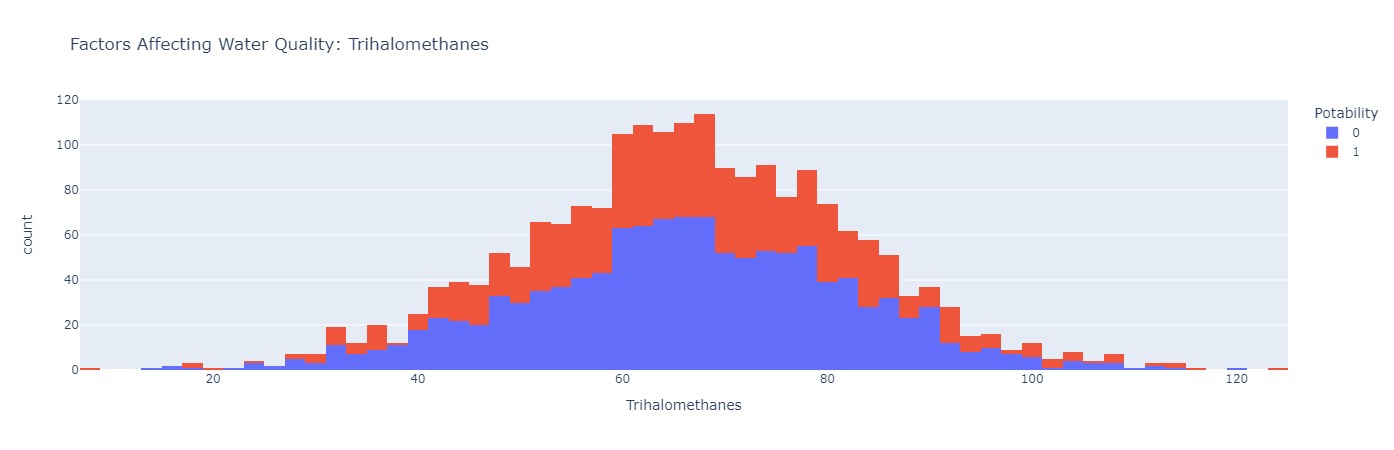


import plotly.express as px data = data

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

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

Output:

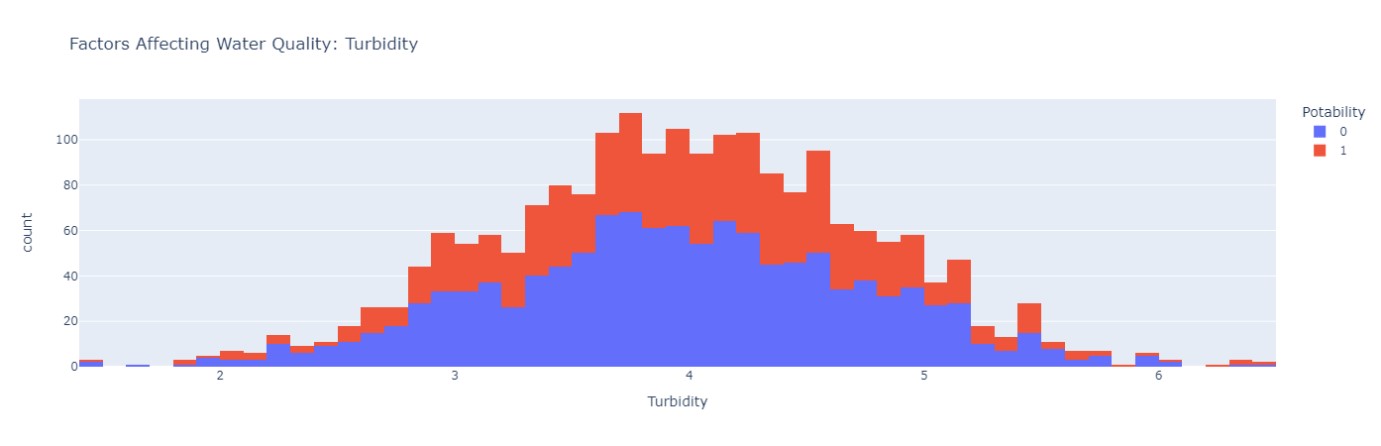


import plotly.express as px data = data

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

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

Output:

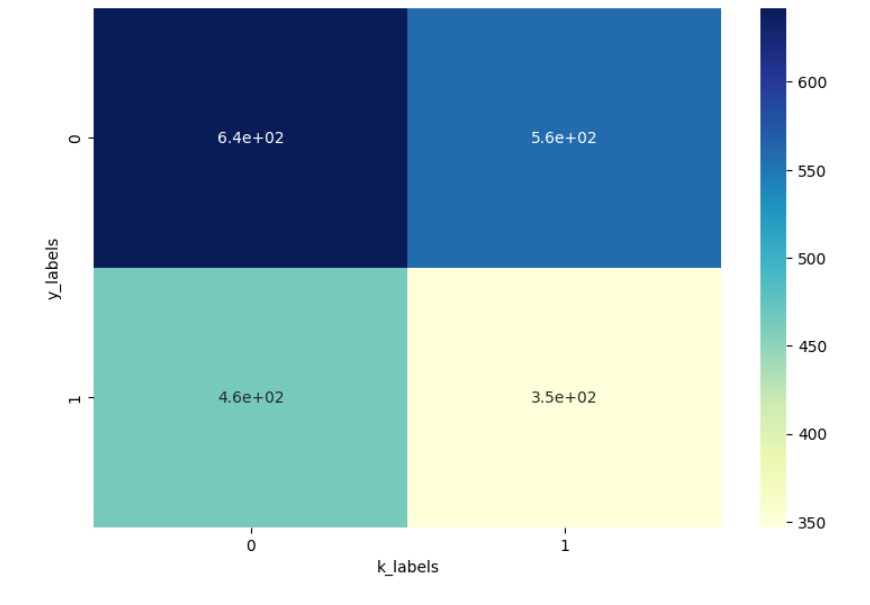


# Split the Data

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()

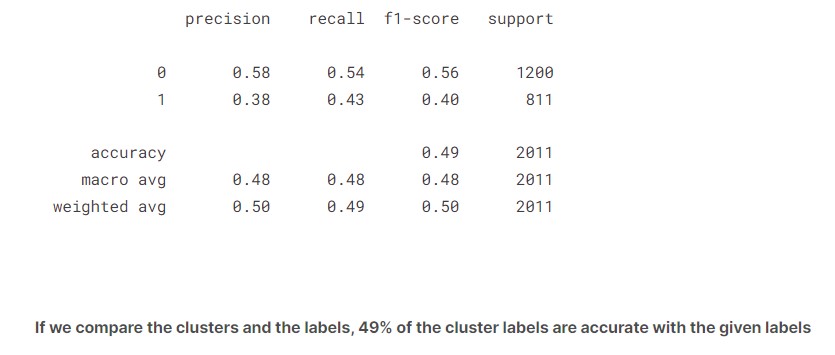
Output:



from sklearn.metrics import accuracy\_score, precision\_score,

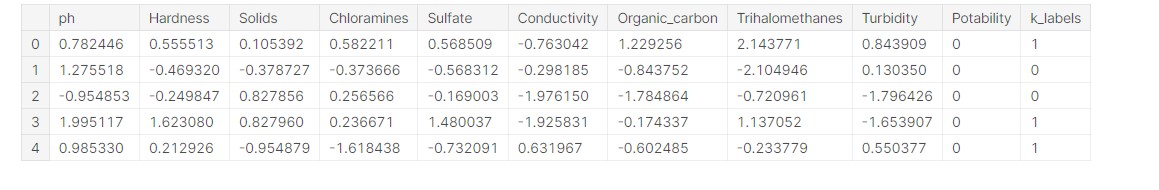
recall\_score, f1\_score,classification\_report accuracy\_score(compare['y\_labels'],compare['k\_labels']) recall\_score(compare['y\_labels'],compare['k\_labels'])

print(classification\_report(compare['y\_labels'],compare['k\_labels'])) Output:

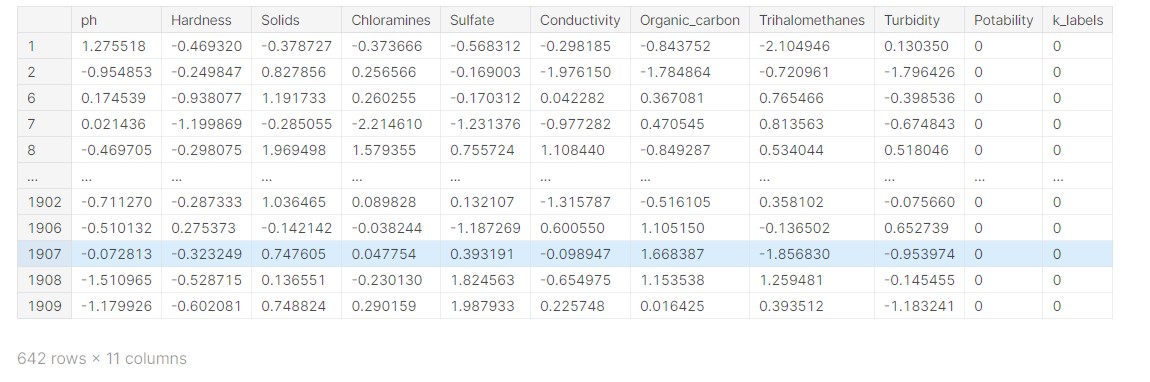


df['k\_labels'] = pd.Series(model.labels\_) df.head()

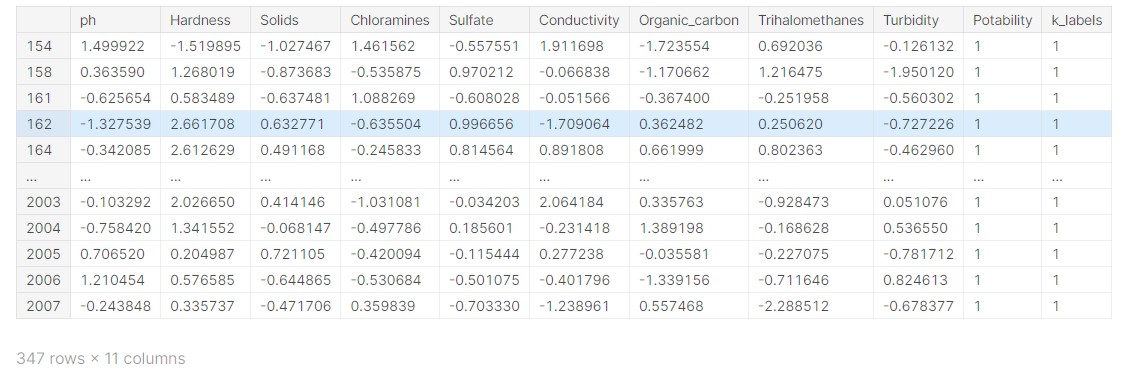
Output:



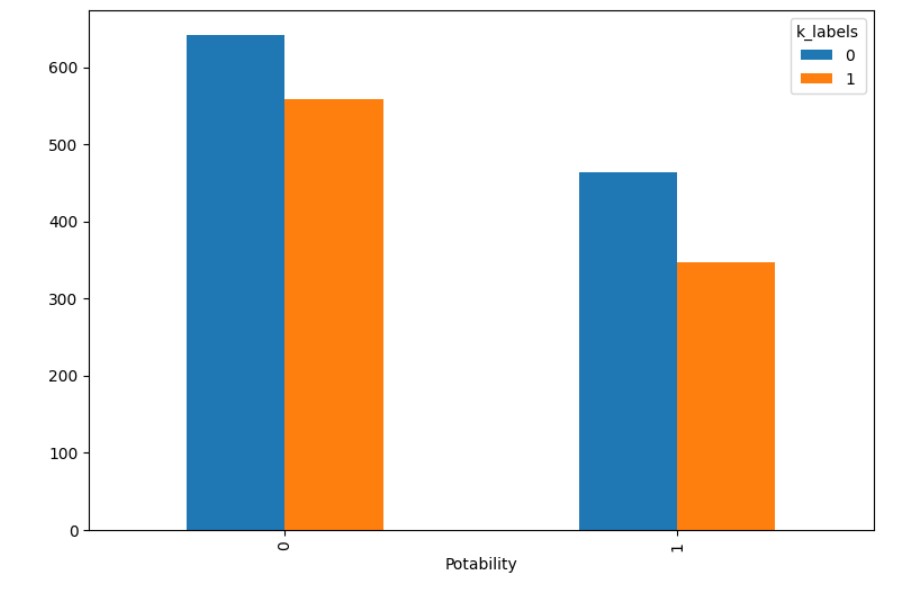
df[((df['Potability']==0)&(df['k\_labels']==0))] Output:



df[((df['Potability']==1)&(df['k\_labels']==1))] Output:



pd.crosstab(df['Potability'],df['k\_labels']).plot(kind='bar') plt.show() Output:



# 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:

