**NAME:GENOVIYA D**

**DEPT/YEAR:ECE/III**

**CLG:2129-ST.JOSEPH COLLEGE OF ENGINEERING**

**REG N0:212921106019**

**PHASE 4:WATER QUALITY ANALYSIS**

Section 1: Data Loading and Exploration  
Section 2: Data Visualization  
Section 3: Initial Data Processing  
Section 4: Feature Selection

Section 1: Data Loading and Exploration  
In this section, we’ll load a dataset from a CSV file using pandas and explore its properties and basic statistics.

First, we import the necessary libraries:

import pandas as pd  
import numpy as np  
import matplotlib.pyplot as plt  
import seaborn as sns  
from sklearn.feature\_selection import SelectKBest, f\_regression

Then, we load the dataset from the CSV file using the pandas.read\_csv() function. We also define column names based on data description files provided by the UCI Machine Learning Repository.

# Load the data set from a CSV file  
df = pd.read\_csv(‘water\_data.csv’, names=[‘station’, ‘year’, ‘month’, ‘day’, ‘temp’, ‘do’, ‘ph’, ‘cond’, ‘turb’, ‘po4’, ‘no2’, ‘no3’, ‘nh4’, ‘o2sat’, ‘bod5’, ‘cod’, ‘ss’, ‘cl’, ‘nh3n’, ‘o2mg’, ‘coli\_fecais’, ‘coli\_totais’, ‘enterococos’, ‘clostridios’, ‘cfu\_100ml\_1’, ‘cfu\_100ml\_2’, ‘cfu\_100ml\_3’, ‘cfu\_100ml\_4’, ‘cfu\_100ml\_5’, ‘cfu\_100ml\_6’, ‘cfu\_100ml\_7’, ‘cfu\_100ml\_8’, ‘cfu\_100ml\_9’, ‘cfu\_100ml\_10’,’cfu\_100ml\_11',’cfu\_100ml\_12',’cfu\_100ml\_13',’level’])

# Display the first five rows of the data frame  
df.head()

Section 2: Data Visualization  
In this section, we will use matplotlib and seaborn to plot and examine the distribution and correlation of variables in a data set.

First, we’ll create a histogram for each variable using the seaborn.histplot() function. The histogram is a graphical representation of the frequency distribution of a variable. It shows how many observations fell into the trash or the different value intervals.

We’ll also use the seaborn.set() function to set the plot style and size, and the plt.tight\_layout() function to set the spacing between plots.

# Set the style and size of the plots  
sns.set(style=’whitegrid’, font\_scale=1.2, rc={‘figure.figsize’:(15, 25)})

# Create a histogram for each variable  
plt.figure()  
for i, col in enumerate(df.columns):  
 plt.subplot(10, 4, i+1)  
 sns.histplot(df[col], kde=False)  
 plt.xlabel(col)  
plt.tight\_layout()  
plt.show()