

WATER QUALITY ANALYSIS

Phase 3: Development part 1

INTRODUCTION

Phase 3 marks a significant milestone in our water quality analysis project as we transition from the foundational stages to the practical implementation of our data-driven solution. This phase is dedicated to loading, pre-processing the water quality dataset, and conducting exploratory data analysis (EDA). By delving into this phase, we lay the groundwork for more advanced analysis and modelling in subsequent stages.

Project Implementation - Water Quality Analysis

In Phase 3, we transition into the project's implementation phase, where we take the first steps in building the water quality analysis system. This phase primarily focuses on loading and pre-processing the water quality dataset and conducting exploratory data analysis (EDA). The primary objectives are to ensure data integrity and understand the dataset's characteristics.

Step 1: Dataset Loading

1. **Acquire the Water Quality Dataset:** Obtain the water quality dataset from reliable sources, ensuring it aligns with the project's objectives.
2. **Data Format Compatibility:** Confirm that the dataset is in a suitable format for analysis (e.g., CSV, Excel, or a structured database).
3. **Data Extraction:** Load the dataset into the chosen data analysis environment (e.g., Python using pandas, Jupyter Notebook, or any other relevant tool).

Step 2: Pre-processing

4. **Data Inspection:** Review the dataset to understand its structure, including column names, data types, and the presence of any missing values.
5. **Handling Missing Values:** Implement strategies to handle missing data, which may include imputation, removal, or interpolation, depending on the nature and extent of missing values.

6. Outlier Detection: Identify and address potential outliers in the dataset, which may affect the accuracy of the analysis.

Step 3: Exploratory Data Analysis (EDA)

7. Visualize Parameter Distributions: Create histograms, density plots, and box plots to visualize the distribution of water quality parameters such as pH, hardness, turbidity, and more.
8. Correlation Analysis: Investigate the relationships and dependencies between different water quality parameters using correlation matrices, scatterplots, and heatmaps. Identify pairs of parameters that are strongly related.
9. Deviation Detection: Compare measured values of water quality parameters (e.g., pH, hardness, turbidity) to established regulatory standards to detect any potential deviations.
10. Data Summary: Summarize the EDA findings, highlighting significant observations, trends, and potential areas of concern.

Data Exploration Tools:

- Utilize data analysis libraries such as pandas, NumPy, and matplotlib for data preprocessing and visualization.
- Python-based tools like Jupyter Notebook or Python IDEs are ideal for performing data analysis and creating visualizations.

This phase serves as the foundation for the water quality analysis project, ensuring that the data is cleaned and ready for further analysis. The insights gained from EDA will guide subsequent phases, including the development of the predictive model and continuous monitoring system.

The outcomes of this phase will be used to refine the project's objectives and further define the analysis plan. Additionally, the findings from EDA will be critical for making informed decisions on how to address deviations from water quality standards and improve potability.

PYTHON SCRIPT

Import necessary libraries

```
import pandas as pd
```

```
import matplotlib.pyplot as plt
```

```
import seaborn as sns
```

Load the water quality dataset

```
data = pd.read_csv("C:/Users/STUDENT/Downloads/dataset.csv")
```

Step 1: Data Preprocessing

Data Inspection

```
print(data.info())
```

Handling Missing Values

```
data = data.dropna() # For simplicity, remove rows with missing values. You  
may choose to impute instead.
```

Outlier Detection (Example: Using Z-Score)

```
from scipy import stats
```

```
z_scores = stats.zscore(data.drop(['Potability'], axis=1))
```

```
data = data[(z_scores < 3).all(axis=1)] # Keep data points within 3 standard  
deviations from the mean
```

```
# Step 2: Exploratory Data Analysis (EDA)
```

```
# Visualize Parameter Distributions
```

```
plt.figure(figsize=(12, 8))
```

```
for column in data.drop(['Potability'], axis=1).columns:
```

```
    sns.histplot(data[column], kde=True, label=column, alpha=0.5)
```

```
plt.xlabel('Parameter Value')
```

```
plt.ylabel('Frequency')
```

```
plt.title('Distribution of Water Quality Parameters')
```

```
plt.legend()
```

```
plt.show()
```

```
# Correlation Analysis
```

```
correlation_matrix = data.corr()
```

```
plt.figure(figsize=(10, 8))
```

```
sns.heatmap(correlation_matrix,          annot=True,          cmap='coolwarm',  
linewidths=0.5)
```

```
plt.title('Correlation Matrix')
```

```
plt.show()
```

```
# Deviation Detection (Example: Comparing pH to WHO standards)
```

```
pH_deviation = data[data['pH'] < 6.5] # Assuming WHO recommends pH > 6.5  
for potable water
```

```
print(f'Instances with pH deviation: {len(pH_deviation)}')
```

```
# Data Summary
```

```
print(data.describe())
```

```
# Your EDA insights will guide further steps in your water quality analysis project.
```

```
# Save the preprocessed dataset if needed
```

```
# data.to_csv('/path/to/preprocessed_dataset.csv', index=False)
```

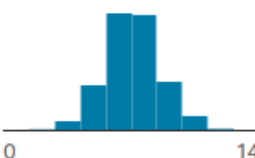
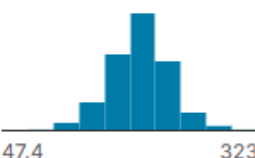
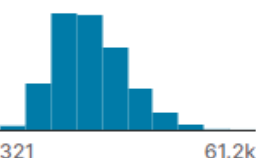
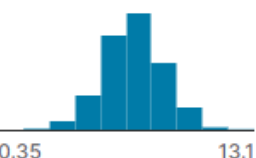
SAMPLE PLOTS

- Dataset in its raw form

ph	Hardness	Solids	Chloramir	Sulfate	Conductiv	Organic_c	Trihalome	Turbidity	Potability
	204.8905	20791.32	7.300212	368.5164	564.3087	10.37978	86.99097	2.963135	0
3.71608	129.4229	18630.06	6.635246		592.8854	15.18001	56.32908	4.500656	0
8.099124	224.2363	19909.54	9.275884		418.6062	16.86864	66.42009	3.055934	0
8.316766	214.3734	22018.42	8.059332	356.8861	363.2665	18.43652	100.3417	4.628771	0
9.092223	181.1015	17978.99	6.5466	310.1357	398.4108	11.55828	31.99799	4.075075	0
5.584087	188.3133	28748.69	7.544869	326.6784	280.4679	8.399735	54.91786	2.559708	0
10.22386	248.0717	28749.72	7.513408	393.6634	283.6516	13.7897	84.60356	2.672989	0
8.635849	203.3615	13672.09	4.563009	303.3098	474.6076	12.36382	62.79831	4.401425	0
	118.9886	14285.58	7.804174	268.6469	389.3756	12.70605	53.92885	3.595017	0
11.18028	227.2315	25484.51	9.0772	404.0416	563.8855	17.92781	71.9766	4.370562	0
7.36064	165.5208	32452.61	7.550701	326.6244	425.3834	15.58681	78.74002	3.662292	0
7.974522	218.6933	18767.66	8.110385		364.0982	14.52575	76.48591	4.011718	0
7.119824	156.705	18730.81	3.606036	282.3441	347.715	15.92954	79.50078	3.445756	0
	150.1749	27331.36	6.838223	299.4158	379.7618	19.37081	76.51	4.413974	0
7.496232	205.345	28388	5.072558		444.6454	13.22831	70.30021	4.777382	0
6.347272	186.7329	41065.23	9.629596	364.4877	516.7433	11.53978	75.07162	4.376348	0
7.051786	211.0494	30980.6	10.0948		315.1413	20.39702	56.6516	4.268429	0
9.18156	273.8138	24041.33	6.90499	398.3505	477.9746	13.38734	71.45736	4.503661	0
8.975464	279.3572	19460.4	6.204321		431.444	12.88876	63.82124	2.436086	0

- Visualization of given data

\

# ph	# Hardness	# Solids	# Chloramines
pH of water	Capacity of water to precipitate soap in mg/L	Total dissolved solids in ppm	Amount of Chloramines in ppm
			
014	47.4323	32161.2k	0.3513.1
	204.8904554713363	20791.318980747026	7.300211873184757
3.71608007538699	129.42292051494425	18630.057857970347	6.635245883862
8.099124189298397	224.23625939355776	19909.541732292393	9.275883602694089
8.316765884214679	214.37339408562252	22018.417440775294	8.05933237743854
9.092223456290965	181.10150923612525	17978.98633892625	6.546599974207941
5.584086638456089	188.3133237696164	28748.68773904612	7.54486878877965
10.223862164528773	248.07173527013992	28749.716543528233	7.5134084658313025

CONCLUSION

In Phase 3, we've taken significant strides towards ensuring water safety. We've handled data issues, identified outliers, and explored parameter relationships. This foundation positions us for the next phase, where we'll construct a predictive model for water portability. With each phase, we're advancing closer to our goal of delivering safe and clean water for all.