WATER QUALITY ANALYSIS

PHASE 5: PROJECT DOCUMENTATION AND SUBMISSION

INTRODUCTION:

In this section we will document the complete project and prepare it for submission.

1. Physical Parameters: These include factors like temperature, turbidity, colour, and odour. These parameters can indicate immediate concerns about water quality and are often measured using sensors and automated data collection systems.

2. Chemical Parameters: These parameters encompass a wide range of substances, such as pH, dissolved oxygen, nutrients (e.g., nitrogen and phosphorus), heavy metals, organic matter, and various chemical contaminants.

3. Biological Parameters: Biological indicators like the presence of specific microorganisms, algae, and macroinvertebrates can reveal the ecological health of a water body.

4. Remote Sensing: Satellite imagery and remote sensors are used to monitor large bodies of water, making it possible to assess water quality over vast areas.

5. Machine Learning and Data Science: Your expertise in AI and ML can be instrumental in developing predictive models and data analysis tools for water quality assessment.

6. Data Collection and Monitoring: Continuous monitoring and data collection are essential for real-time assessment of water quality. Data science techniques are employed to manage and analyse the vast amounts of data generated.

7. Regulatory Standards: Water quality analysis is often based on established regulatory standards and guidelines set by environmental agencies.

Here we contribute to the advancement of water quality analysis by developing innovative technologies and analytical tools that can improve our understanding and management of this critical resource.

# PROBLEM STATEMENT:

The problem at hand is to advance the field of water quality analysis by harnessing the capabilities of artificial intelligence (AI) and machine learning (ML). This project aims to address the following challenges:

* Data Quality Assurance
* Predictive Modeling
* Anomaly Detection
* Spatial Analysis

PROBLEM DEFINITION:

The most frequent water quality issue is due to the high content of iron (iron (III) oxide) and magnesium content in raw water of treated water. Water quality disorders occur as a result of changes in the colour of the water that turns yellow to a dark brown colour. The colour change is due to action chemical reactions that are used in the water treatment process at the Treatment Plant (Kasan, 2006). This water treatment diagnostic and auditing process still uses manual methods, where water will be measured and the quality index will be clinically measured inside the laboratory. Besides, low pH levels cause fish killed by stressing animal system and causing physical damage, which in turn makes them more vulnerable to disease.

DESIGN THINKING:

* ANALYSIS OBJECTIVES:

The objective of water quality analysis is to assess and ensure the safety of water for human consumption and ecosystem health by examining parameters such as contaminants, chemical composition, and microbial content.

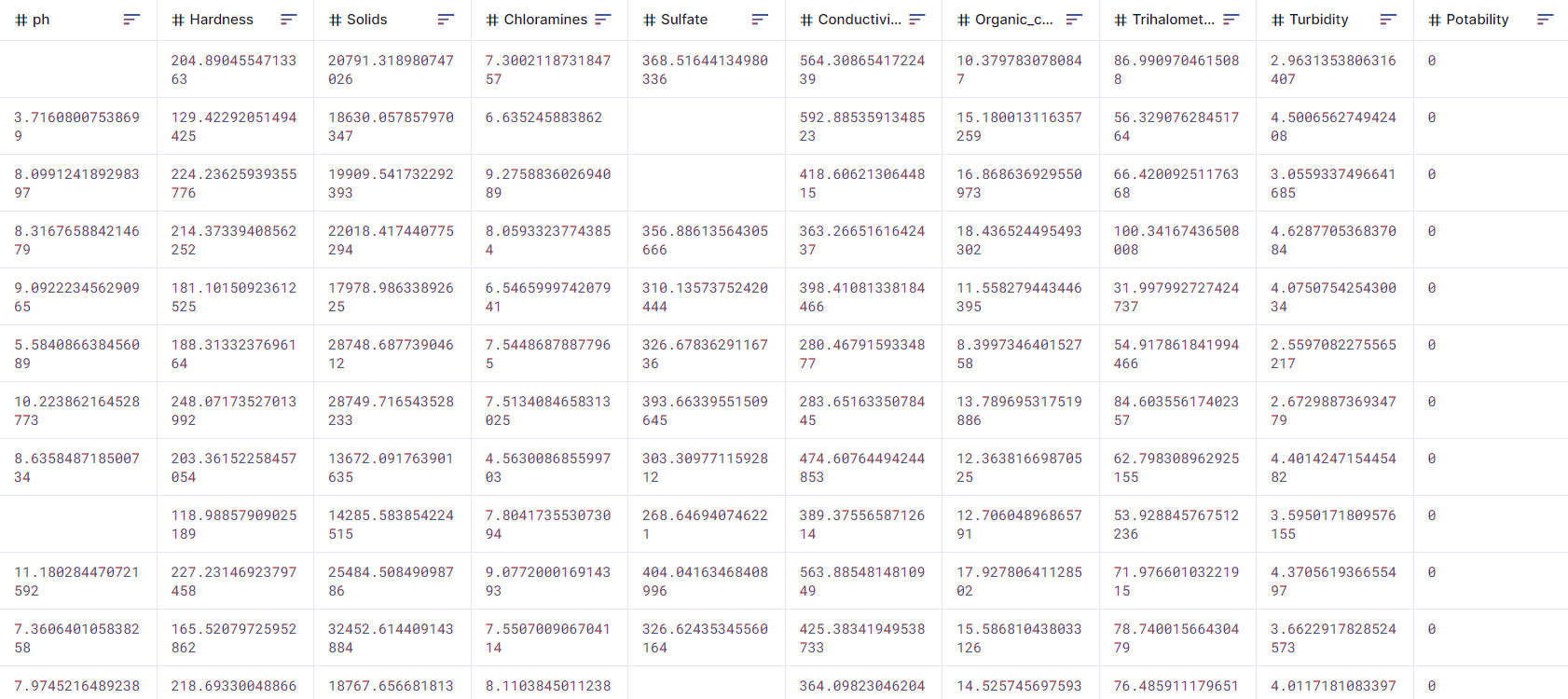
* DATA COLLECTION:

We will obtain our dataset from Kaggle. This dataset will include essential features such as PH, hardness, solids, chloramines, sulphate, Conductivity, organic carbon, turbidity and potability. Having access to

Real-world data will be essential for training a predictive model.

KAGGLE DATASET LINK:

<https://www.kaggle.com/datasets/adityakadiwal/water-potability>

GIVENDATASET:

TOOLS/MODULES:

Pandas, NumPy, and Scikit-Learn will be used for data preprocessing and initial analysis.

* VISUALIZATION STRATERGY:

In the realm of water quality analysis, employing machine learning (ML) tools for visualization is a powerful strategy. Utilizing charts, graphs, and maps to represent data trends and3D visualizations for spatial analysis of water quality parameters. The content could be structured.

* PREDICTIVE MODELLING:

Decide on the machine learning algorithms and features to use for predicting water potability

Here's a list of tools and software commonly used in the process:

### Programming Language:

- Python is the most popular language for machine learning due to its extensive libraries and frameworks. You can use libraries like NumPy,pandas, scikit-learn, andmore*.*

### Integrated Development Environment (IDE):

- Choose an IDE for coding and running machine learning experiments. Some popular options include Jupiter Notebook, Google Collab, or traditional IDEs like PyCharm.

### Machine Learning Libraries:

You'll need various machine learning libraries, including:

* + scikit-learn for building and evaluating machine learning models.
  + TensorFlow or PY Torch for deep learning, if needed.
  + XG Boost, Light GBM, or Cat Boost for gradient boosting models.

### Data Visualization Tools:

* + Tools like Matplotlib and Seaborn are essential for data exploration and visualization.

### Data Preprocessing Tools:

* + Libraries like pandas help with data cleaning, manipulation, and preprocessing.

### Data Collection and Storage:

* + Depending on your data source, you might need web scraping tools (e.g., Beautiful Soup or Scrapy)or databases (e.g.,SQLite, PostgreSQL)for data storage.

# STEP1:DESIGN THINKING AND PRESENT IN FORM OF DOCUMENT

### Empathize:

### Understand the needs and challenges of all water quality parameters include pH value, Hardness, Solids (Total dissolved solids - TDS), Chloramines, Sulphate, Conductivity, Organic carbon, Trihalomethanes, Turbidity, Potability.

* Conduct interviews and surveys to gather insights on what user value in water quality and what information is most critical for their decision-making.

### Define:

* Clearly articulate the problem statement, such as "How might we analysis the water quality more accurately and transparently using machine learning?"
* Identify the key goals and success criteria for the project, such as quality, reducing hardness, or improving user trust in the valuation process.

### Ideate:

* Brainstorm creative solutions and data sources that can enhance the quality of water.
* Encourage interdisciplinary collaboration to generate a wide range of ideas, including the use of alternative data, new algorithms, or improved visualization techniques.

### Prototype:

* Create prototype machine learning models based on the ideas generated during the ideation phase.
* Test and iterate on these prototypes to determine which approaches are most promising in terms of quality and usability.

### Test:

* Gather feedback from users and public by testing the machine learning models with real-world data and scenarios.
* Assess how well the models meet the defined goals and success criteria, and make adjustments based on user feedback.

### Implement:

* Develop a production-ready machine learning solution for improving the quality, integrating the best-performing algorithms and data sources.
* Implement transparency measures, such as model interpretability tools, to ensure users understand how predictions are generated.

### Evaluate:

* Continuously monitor the performance of the machine learning model after implementation to ensure it remains accurate and relevant result in the water quality.
* Gather feedback and insights from users to identify areas for improvement.

# STEP2:DESIGN INTO INNOVATION

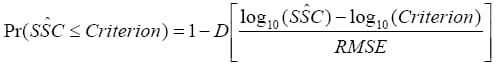
INTRODUCTION:

The machine learning technique we implement in this project is Regression. Regression is a statistical technique used in data analysis and machine learning to model the relationship between a dependent variable and one or more independent variables. It is a method for understanding how changes in the independent variables are associated with changes in the dependent variable. The goal of regression analysis is to create a predictive model that can make accurate predictions or estimates about the dependent variable based on the values of the independent variables.

QUANTIFYING UNCERTAINITY:

All regression models have uncertainty inherently associated with each computation. Uncertainty can be defined in a number of different ways, including relative percent difference, absolute error, and prediction intervals. Prediction intervals, used by this website, define a range of values for the response variable for a given level of certainty. The level of certainty presented with the modelled data is the 90-percent prediction interval. The larger the range, the more uncertainty there is associated with the regression computed value.

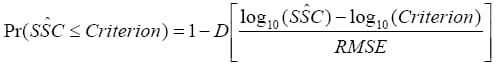
Calculating prediction intervals for regression models with two or more explanatory variables involves matrices.



* t is the value of Student's t-distribution having n-3 degrees of freedom with an exceedance probability of α/2,
* α is the level of certainty for the prediction interval (1-0.9, or 0.1 for this website),
* p is the number of explanatory variables plus one,
* s is the variance of the residuals calculated during model development,
* (X'X)-1 is the X prime X inverse matrix calculated during model development — an expression of the covariance among all explanatory variables, and X is a vector of the explanatory variable measurements.

PROBABILITY OF CRITERIA:

A convenient way to interpret regression model computed concentrations in the context of water-quality criteria is the probability of exceedance. Probability of exceedance is a single value representing the percent likelihood that a criterion has been exceeded. This approach assumes that a normal distribution centre on the mean computed value describes the uncertainty distribution. For log-transformed response variables, this equation is applied in logarithmic space.



* ‘P r’ is the probability that the criterion has been exceeded (0 <P r< 1),
* D is the cumulative distribution function for the standard normal curve value for it are obtained from equations that approximate the exact values.
* RMSE is the root-mean-squared-error, or standard error of the regression, or standard deviation of the residuals.

LIMITIONS OF REGRESSION ANALYSIS:

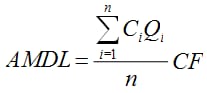
* A regression model between the response and explanatory variables.
* It may change over time if changes occur in the sources of the

constituent or an improved sensor becomes available.

* This is because of simplifying assumptions implicitly built into the regression analysis.
* For example, turbidity measurements are affected by physical properties of suspended-sediment particles such as size, color, and density.
* In turn, these physical properties are affected by complex watershed properties such as source sediment lithology, stream morphology, land use distribution, and many other things.
* Therefore, regression analysis is site-specific, and the regression model must be verified annually by continued sample collection and refined as needed.

CALCULATION FOR CONSTITUENT LOAD:

Annual mean daily loads (AMDL) of constituents are calculated by summation of available hourly (or more frequent time steps) loads for a given year:



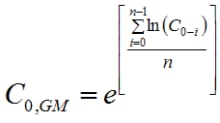
* Ci is the instantaneous, hourly concentration at the th time,
* Qi is the instantaneous, hourly streamflow at the I th time,
* CF is a conversion factor from Table 1, and
* N is the number of available hourly values for a given year (a maximum of 8,760).

STEPS FOR CALCULATION:

* Load computations assume that there is no error or uncertainty in the computation of streamflow.
* Although this is not true, the typical measurement uncertainty of streamflow is less than 5 percent.
* Measurement uncertainty for the sensor values typically are less than 10 percent if common protocols for operation of water quality.

CALCULATION FOR GEOMETRIC MEAN OF INDICATOR BACTERIA:

* This approach, a type of smoothing function, helps reveal the "central tendency" of the data and also allows for more direct comparisons with some water-quality criteria.
* A comparison between instantaneous and geometric-mean values is calculated.
* This is consistent with definitions of biological water-quality criteria used by some regulatory agencies, such as the Kansas Department of Health and the Environment.



* C0, GM is the calculated 30-day moving geometric mean for one moment in time,
* nis equal to 720, which is 24 instantaneous values per day multiplied by 30 days,
* C0 through C0-I are the current instantaneous value, and the preceding 719 instantaneous values.

# STEP3: DATA PREPOCESSING AND VISUALIZATION

INTRODUCTION:

Thediversity of research on the water quality monitoring system discussed by Wang, Wang (2009) and Jiang & All (2009). According to Wang, Wang (2009), IoT devices typically in this application usually include large geographical areas and also mobile. The IoT diversification leads to wireless sensing that will revolutionize the river area. The developed system allows for remote monitoring and real-time monitoring of water quality parameters and allowing current status observations and water quality history. Jiang & All (2009) perform a water monitoring system that includes pH analysis, conductivity, dissolved oxygen and temperature levels will be implemented. Alarm sounds will be triggered if there is water pollution or water quality changes. The parameters are measured with landing sensors and data transmitted to base stations via GPRS.

DATA PREPROCESSING**:**

Data preprocessing is a crucial step in data analysis and machine learning. It involves cleaning and transforming raw data into a format that is suitable for analysis or model training. The process typically includes:

1. Data Cleaning: Removing or handling missing values, correcting errors, and dealing with outliers to ensure data quality.

2. Data Transformation: Converting data into a suitable format, such as encoding categorical variables, scaling numerical features, and normalizing data.

3. Feature Selection: Identifying and selecting the most relevant features (variables) for analysis or model training to improve efficiency and reduce noise.

4. Data Reduction: Reducing the dimensionality of data when dealing with high-dimensional datasets to simplify analysis.

5. Data Integration: Combining data from multiple sources or datasets to create a unified dataset for analysis or modeling.

MISSING VALUES**:**

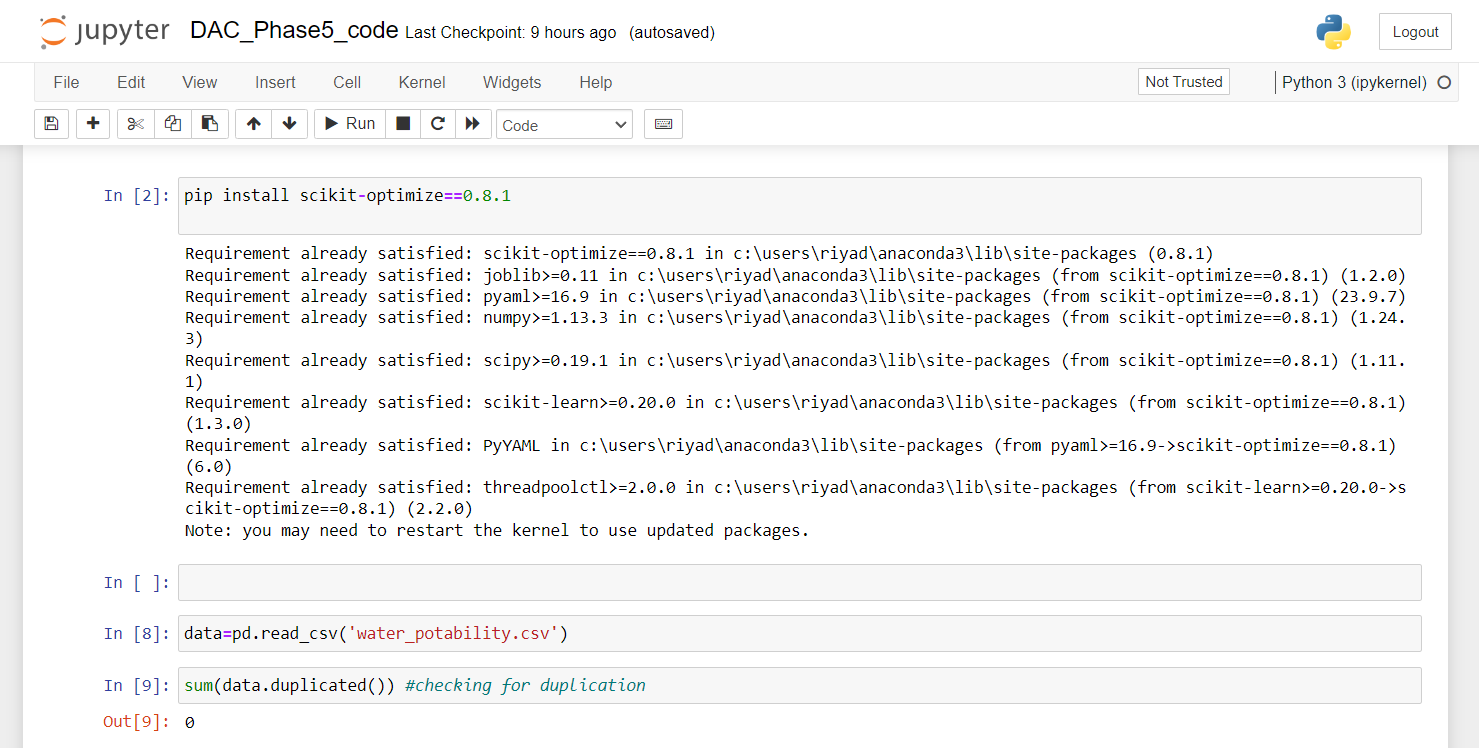
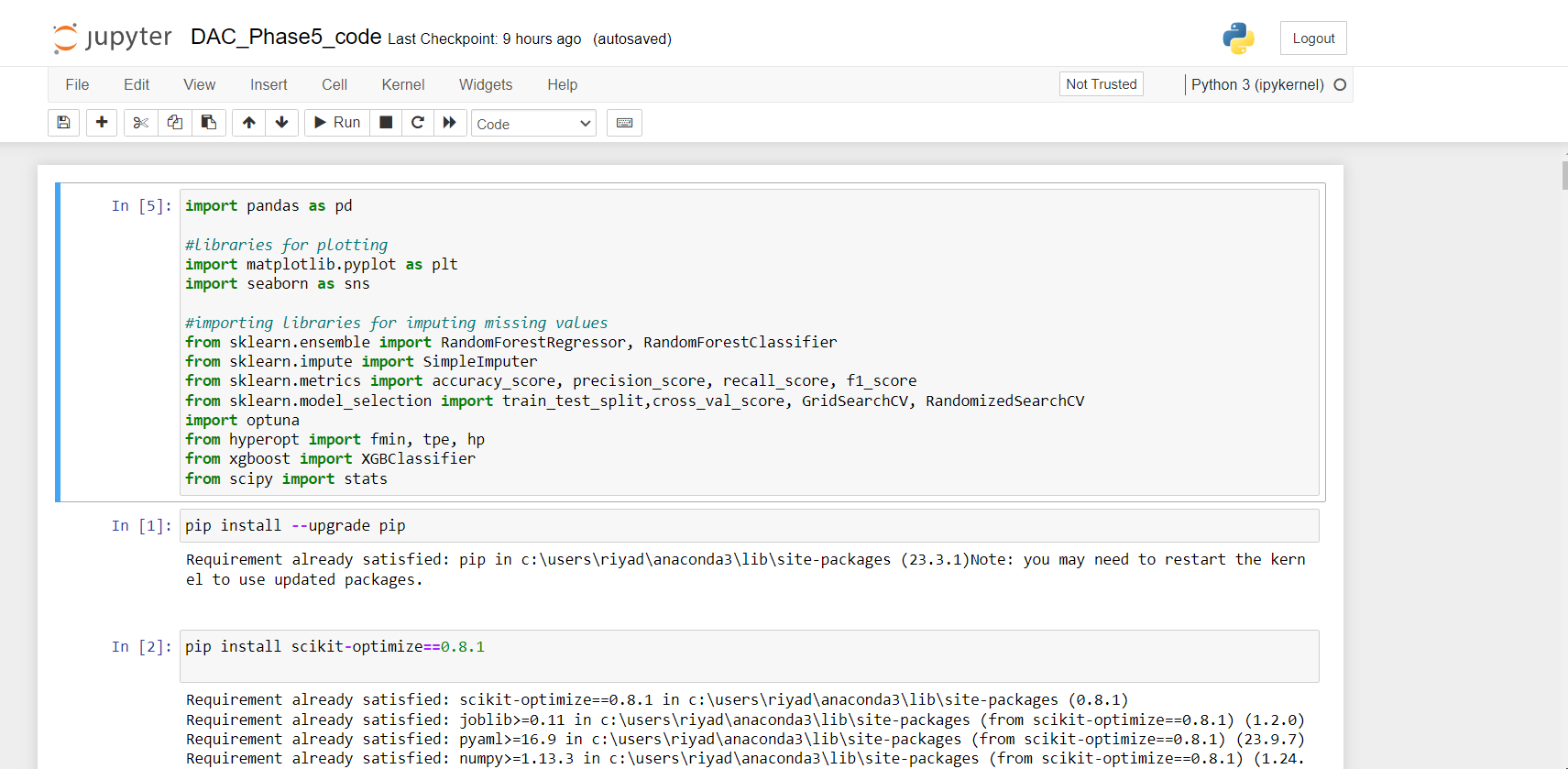
Missing values can be handled using various libraries and techniques. Importing necessary libraries called pandas and matplotlib.

Pandas:

Pandas is an open-source data manipulation and analysis library for Python. It provides data structures and functions that make it easier to work with structured data, such as tables or CSV files. Pandas is particularly useful for data cleaning, transformation, and analysis tasks. It offers two primary data structures: Series (1D labeled arrays) and Data Frame (2D labeled data structures), making it a powerful tool for data processing and analysis.

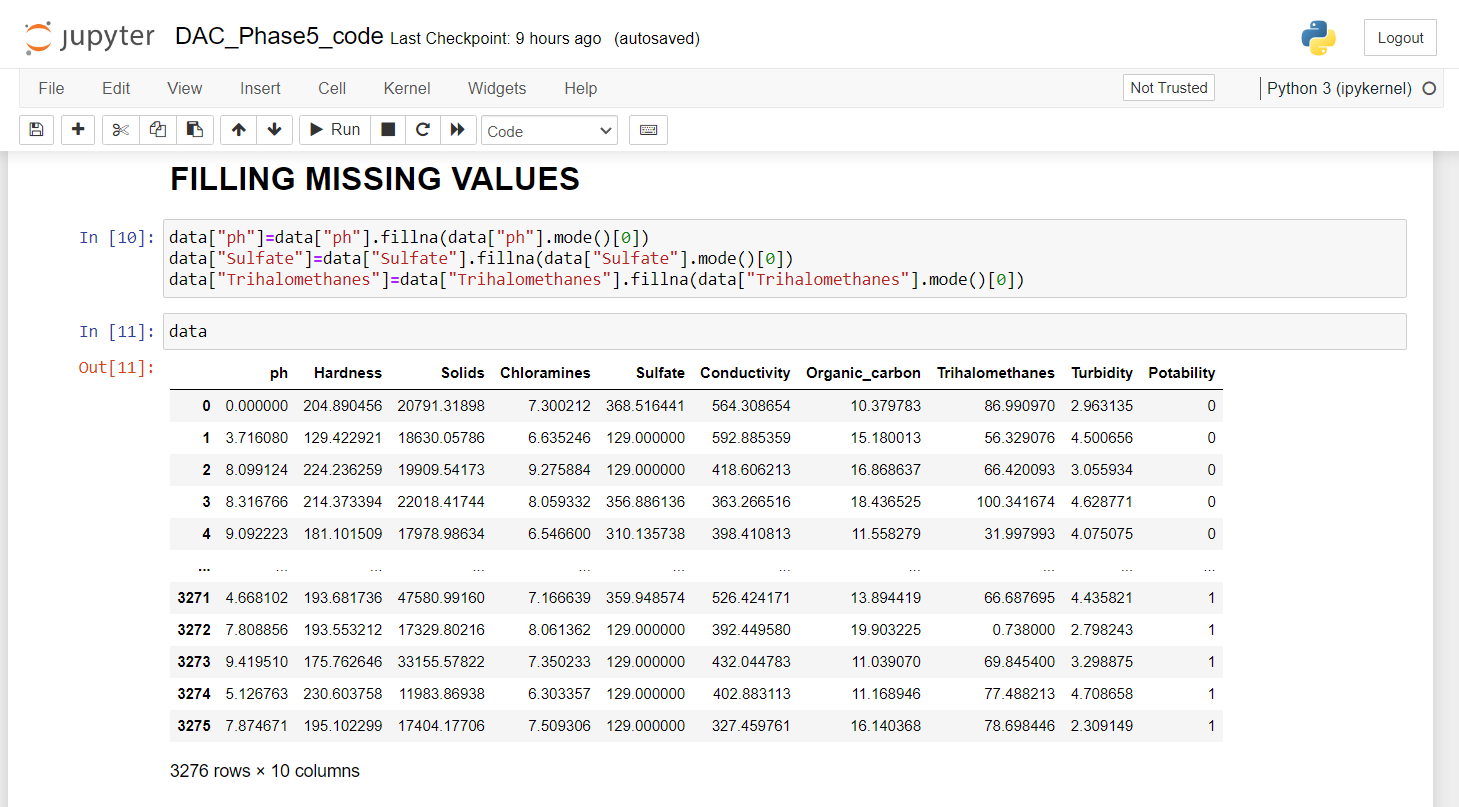
Matplotlib:

Matplotlib is a widely-used Python library for creating high-quality static, animated, or interactive visualizations. It provides a wide range of options for creating plots and charts, allowing users to customize every aspect of their visualizations. Matplotlib is versatile and can be used for various types of graphs, including line plots, bar charts, scatter plots, and more. It's often used in combination with other libraries like NumPy and Pandas to visualize data and gain insights from it.



In the above picture missing values are filled using python libraries called pandas and matplotlib.missing values are filled with “Nan”.

FILLING MISSING VALUES:



ANALYSING THE WATER QUALITY USING UNIVARIATE ANALYSIS:

Certainly, univariate analysis is a statistical technique that focuses on the analysis of a single variable or data attribute at a time. It is a fundamental step in data analysis to understand the characteristics of individual variables. Here's how you can perform univariate analysis:

* Gather your dataset, ensuring you have data for the variable you want to analyze. Steps involved during summarizing the dataset:

1.Calculate the mean (average) to understand the central tendency of the data.

2.Calculate the median to find the middle value of the data, which can be

more robust to outliers.

3.Determine the mode, which is the most frequently occurring value.

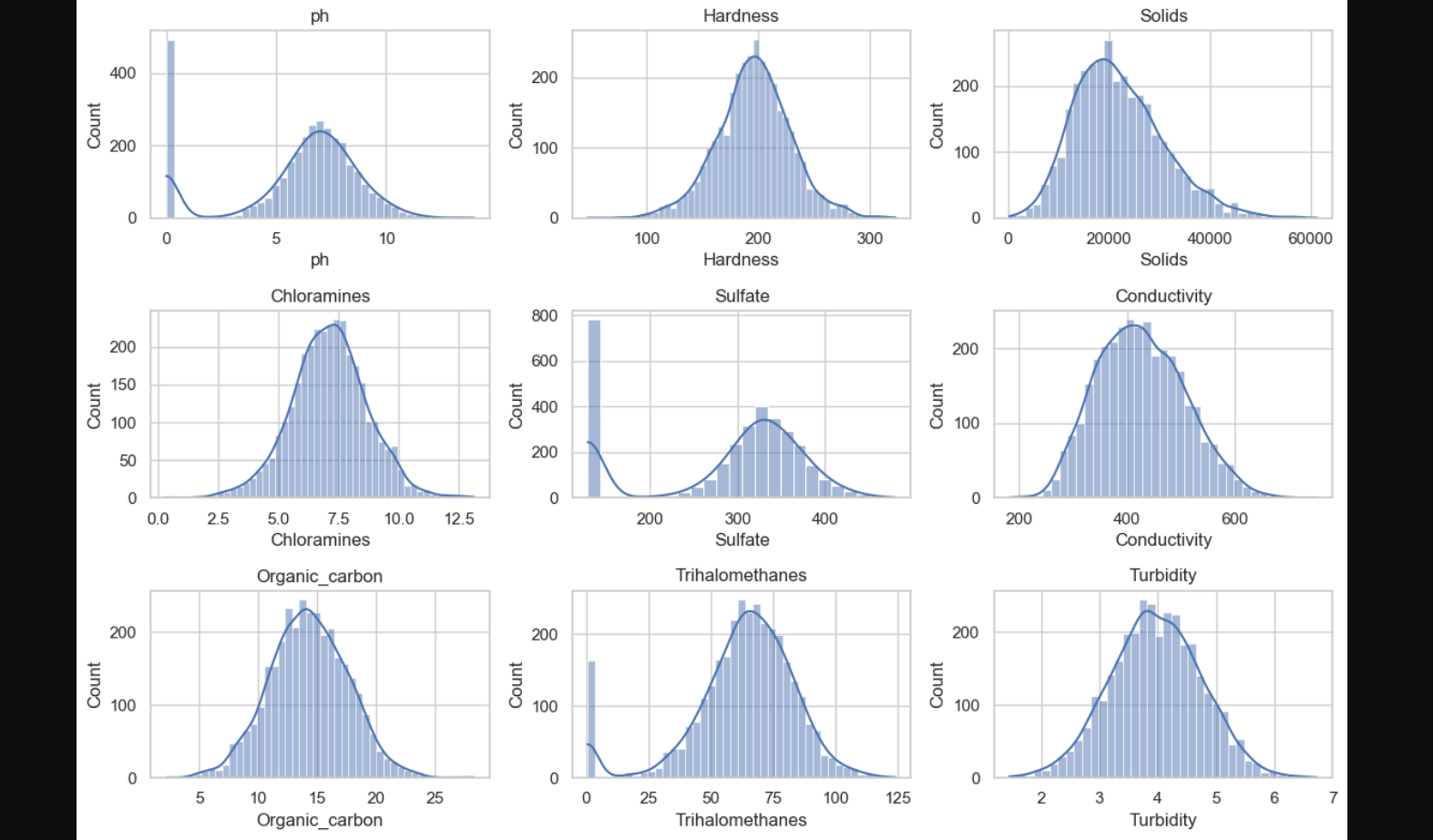
4.Calculate the range, which is the difference between the maximum and minimum values.

5.Calculate the variance and standard deviation to measure the data's spread or dispersion.

6.Compute percentiles, such as the 25th, 75th, and others, to understand data distribution.

* Create a histogram to visualize the data's frequency distribution. Construct a box plot to identify outliers and visualize the data's spread. Utilize a probability density plot (kernel density estimation) for a smooth representation of the data distribution. Generate a bar chart or pie chart if your variable is categorical.
* Analyze the summary statistics and visualizations to draw conclusions about the variable's characteristics. Look for patterns, central tendencies, outliers, and any significant features.
* If you have specific hypotheses about the variable's behavior, you can perform statistical tests like t-tests or ANOVA to validate them.
* Present your findings clearly, including visualizations and statistical summaries. Explain any significant insights or trends you've identified in the univariate analysis.

Univariate analysis is a crucial first step in exploring and understanding your data before moving on to more complex multivariate analyses. It helps in identifying outliers, understanding the data's central tendencies, and getting an overall sense of the variable's distribution.

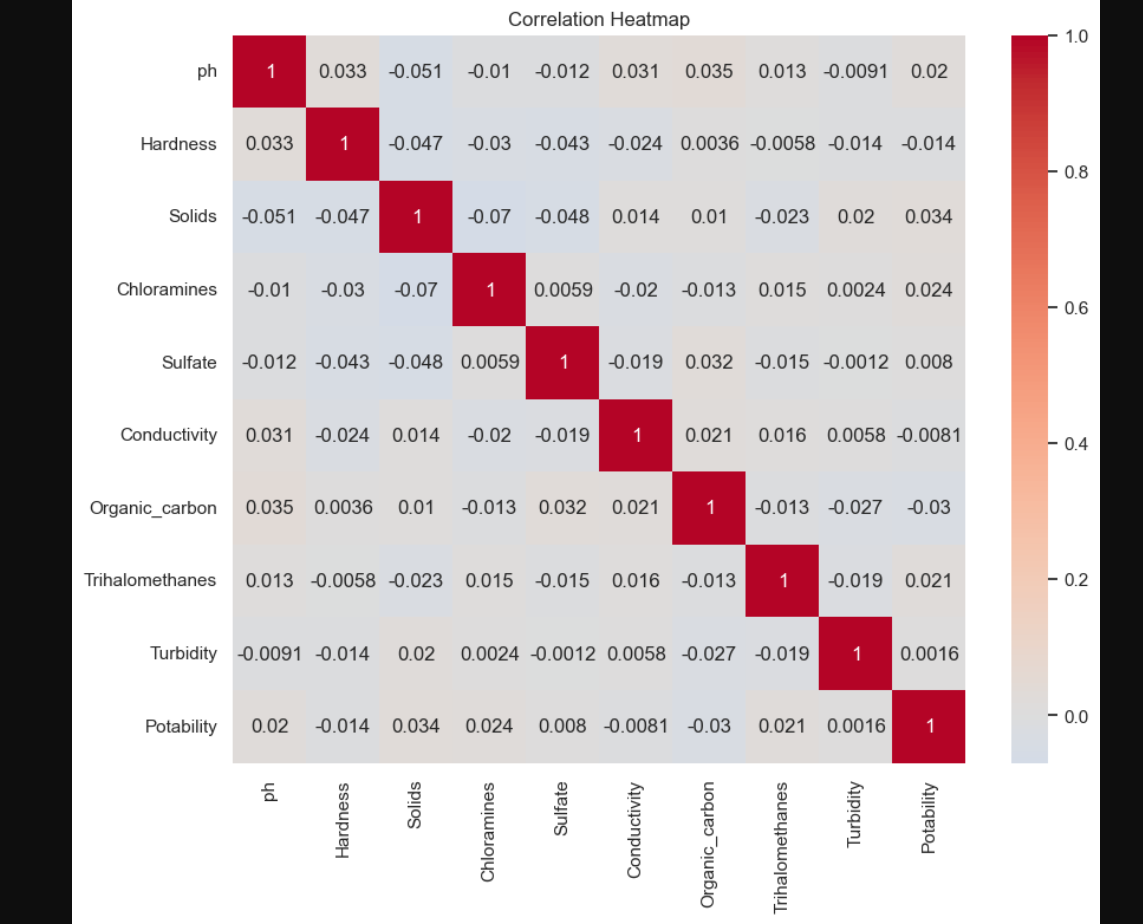
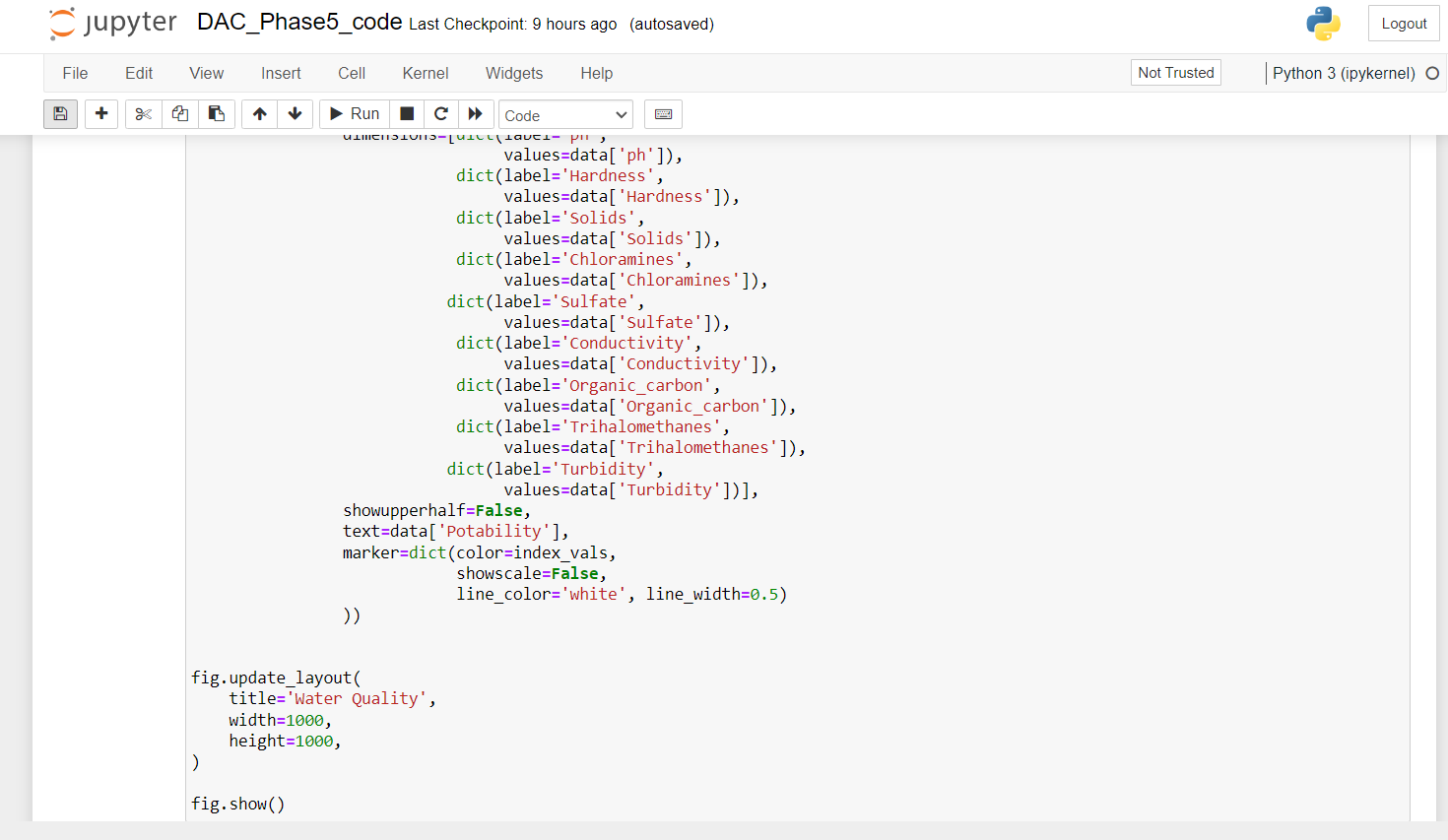
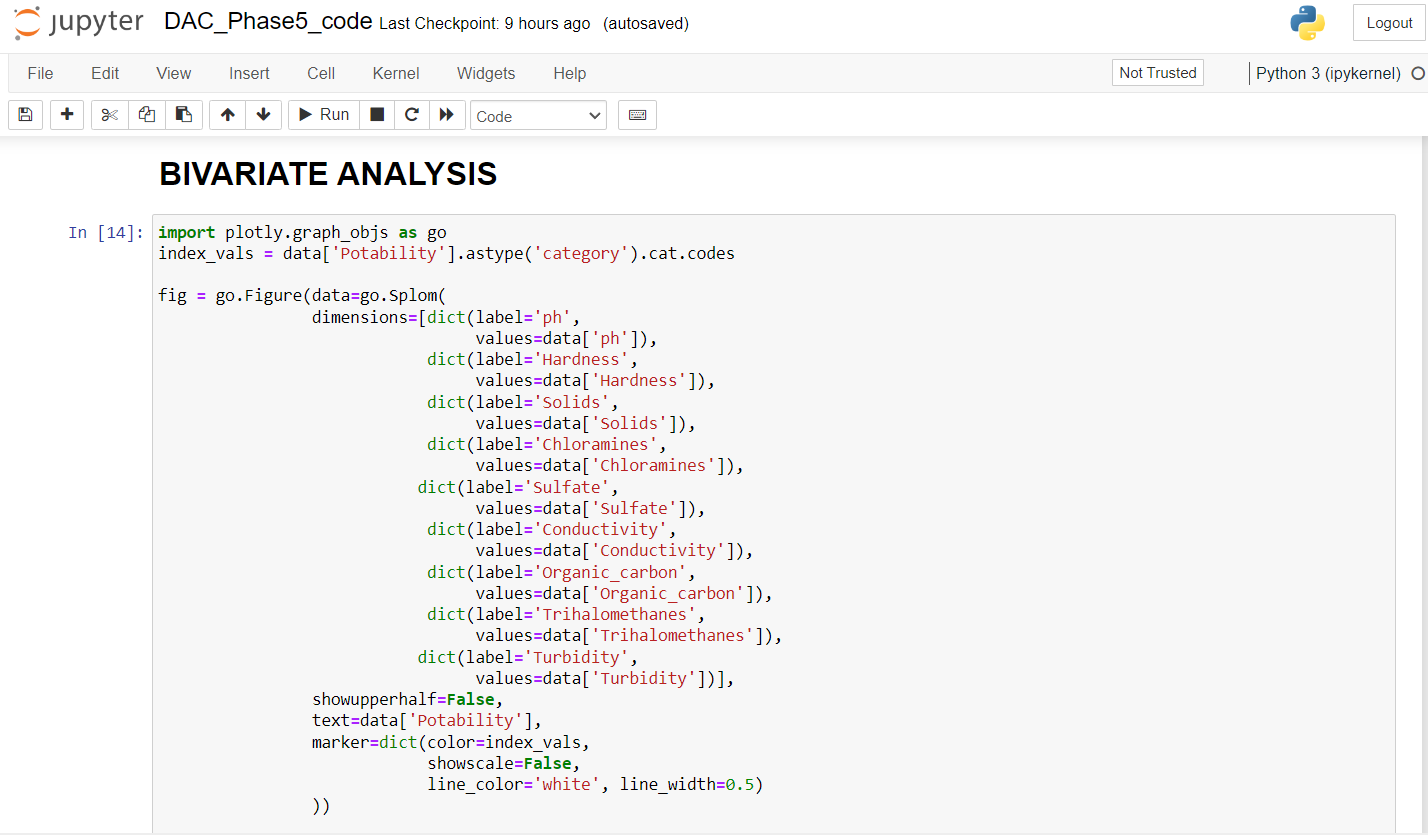


ANALYSING THE WATER QUALITY USING BIVARIATE ANALYSIS:

Bivariate analysis, unlike univariate analysis, involves the examination of the relationships between two variables. It is used to explore how two different variables interact or relate to each other.

* Gather your dataset, ensuring you have data for at least two variables that you want to analyze in relation to each other. Create a scatter plot to visualize the relationship between the two variables.
* A scatter plot helps you see if there's any correlation or pattern between the two variables. It's especially useful when dealing with continuous numerical variables.
* Calculate the correlation coefficient, such as Pearson's correlation coefficient, to quantify the strength and direction of the relationship between the two variables. A positive correlation indicates that as one variable increases, the other tends to increase as well. A negative correlation suggests that as one variable increases, the other tends to decrease.
* If you are dealing with categorical variables, you can create crosstabulation tables to show the frequency distribution of one variable with respect to the other. This helps you see how the two categorical variables are related.
* Use box plots or violin plots to compare the distribution of a numerical variable across different categories of a categorical variable. Create grouped bar charts to compare the means of a numerical variable across categories of a categorical variable.
* If you have specific hypotheses about the relationship between the two variables, you can perform hypothesis tests, such as t-tests or chi-squared tests, to determine if the relationship is statistically significant.
* Analyze the visualizations and statistical measures to draw conclusions about the relationship between the two variables. Determine if there is a significant correlation, association, or difference between the variables.
* Present your findings clearly, including visualizations and statistical summaries. Explain any significant relationships or patterns you've identified through bivariate analysis.

Bivariate analysis is useful for exploring relationships between two variables, such as cause-and-effect, correlation, or association. It can help you uncover insights about how changes in one variable relate to changes in another, which is valuable for decision-making and hypothesis testing.

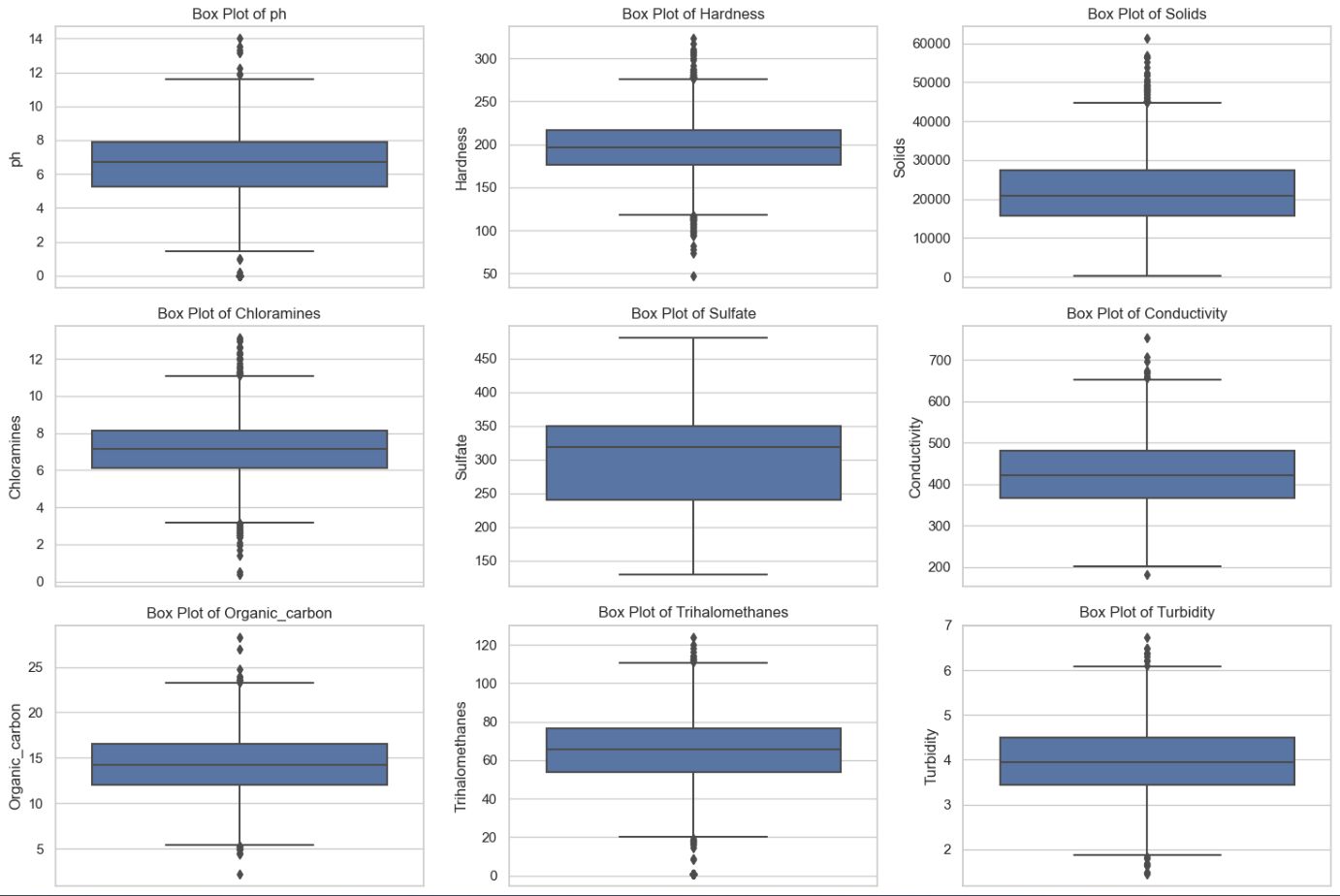
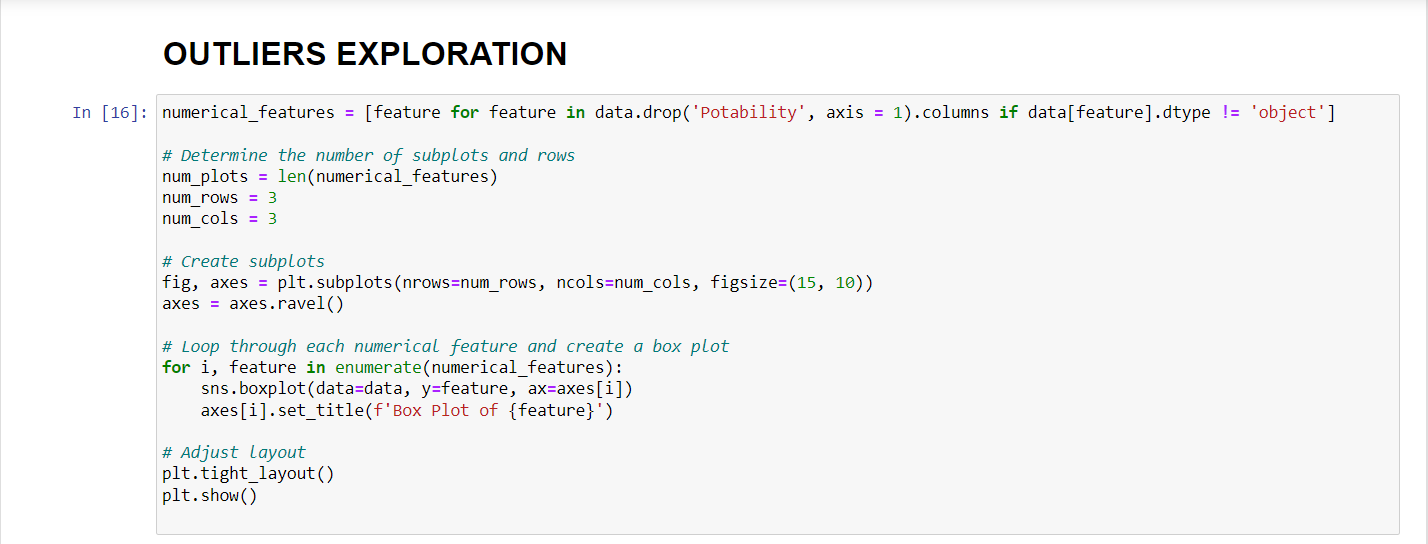


OUTLIERS EXPLORATION:

Outlier exploration is an essential part of data analysis to identify and understand data points that significantly deviate from the majority of the dataset. Outliers can distort statistical analyses and models, so it's important to detect and handle them appropriately.

It's important to note that not all outliers are necessarily errors or noise; some may represent valid but unusual data points. The decision to handle outliers should be based on the context and goals of your analysis.

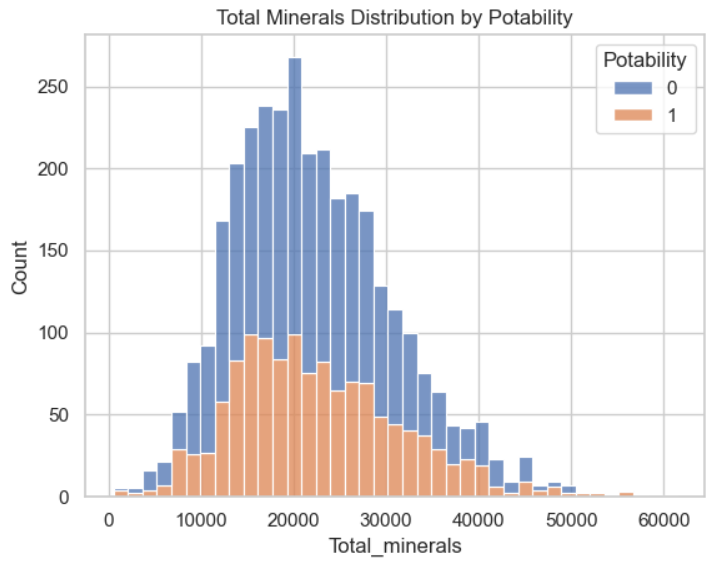
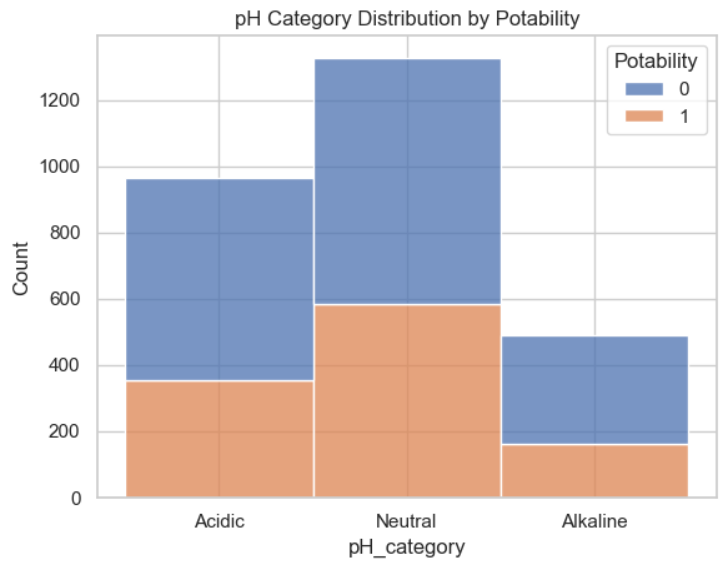
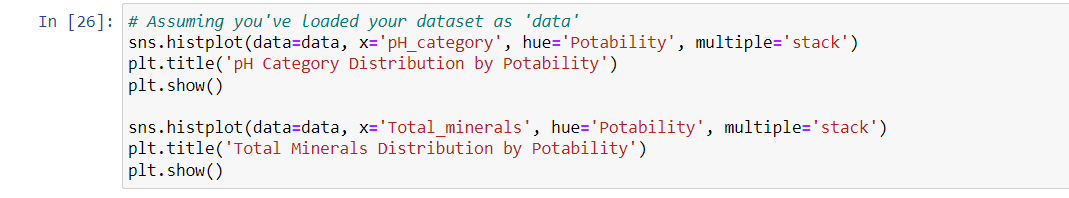
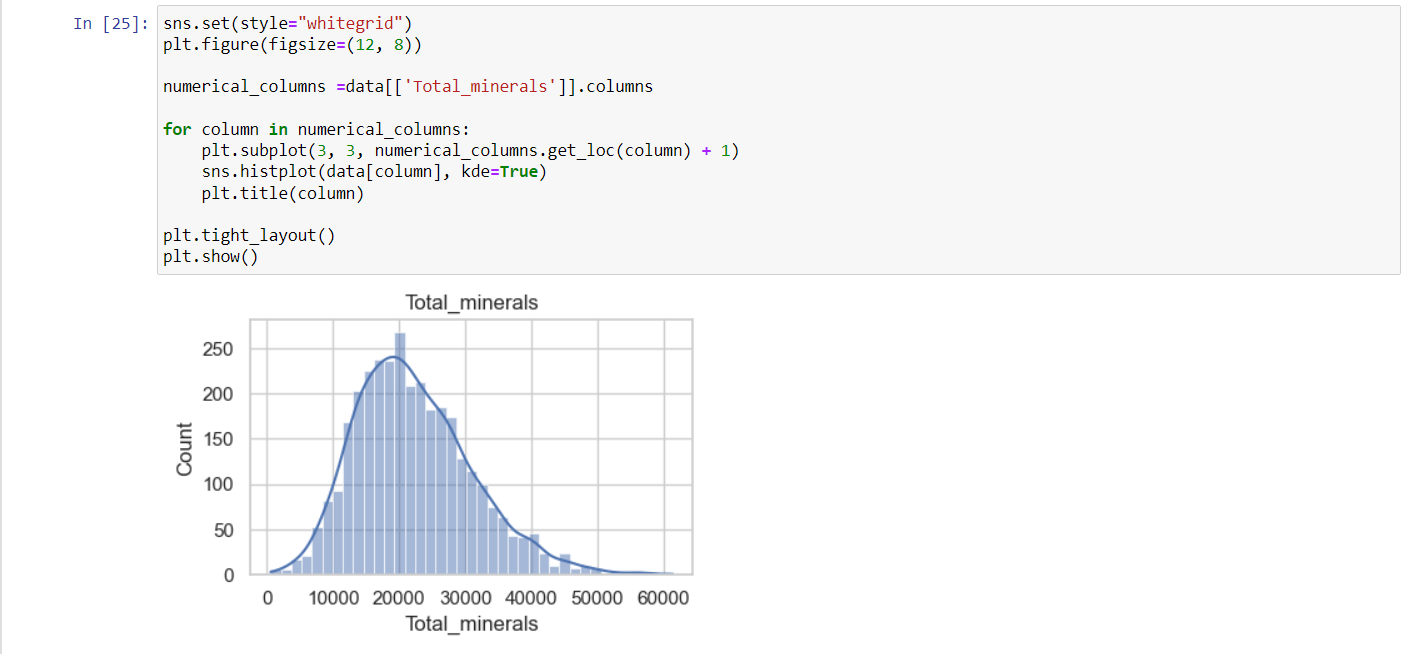
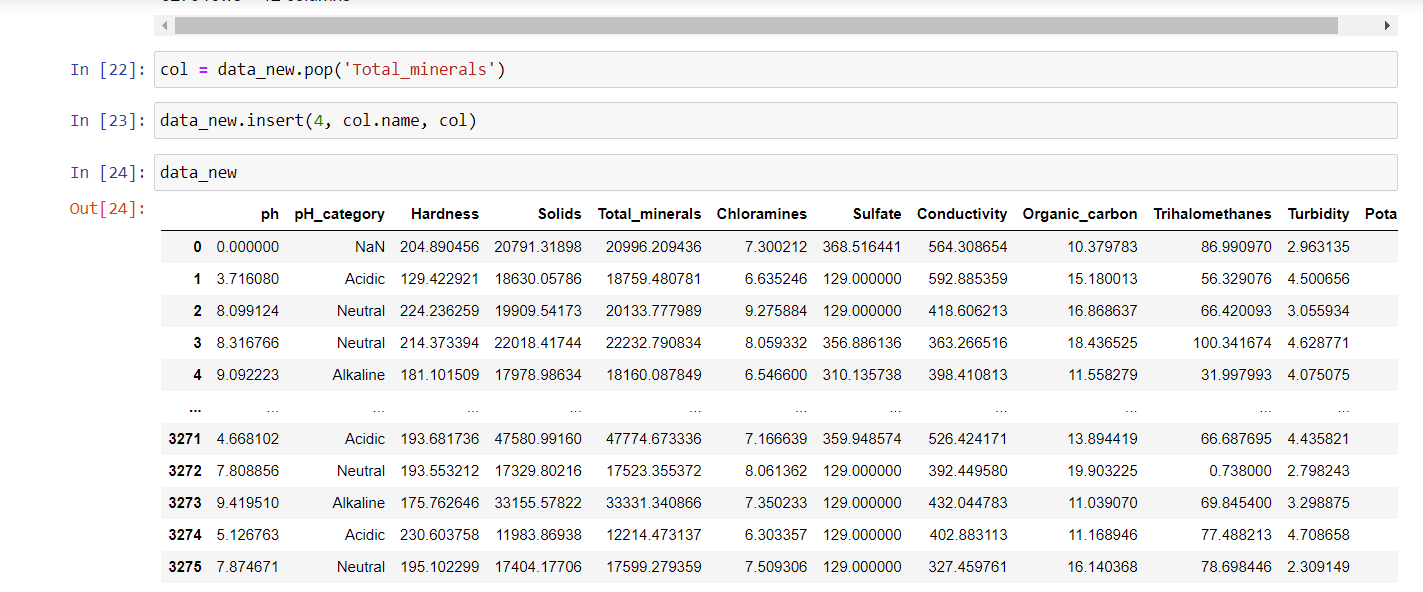
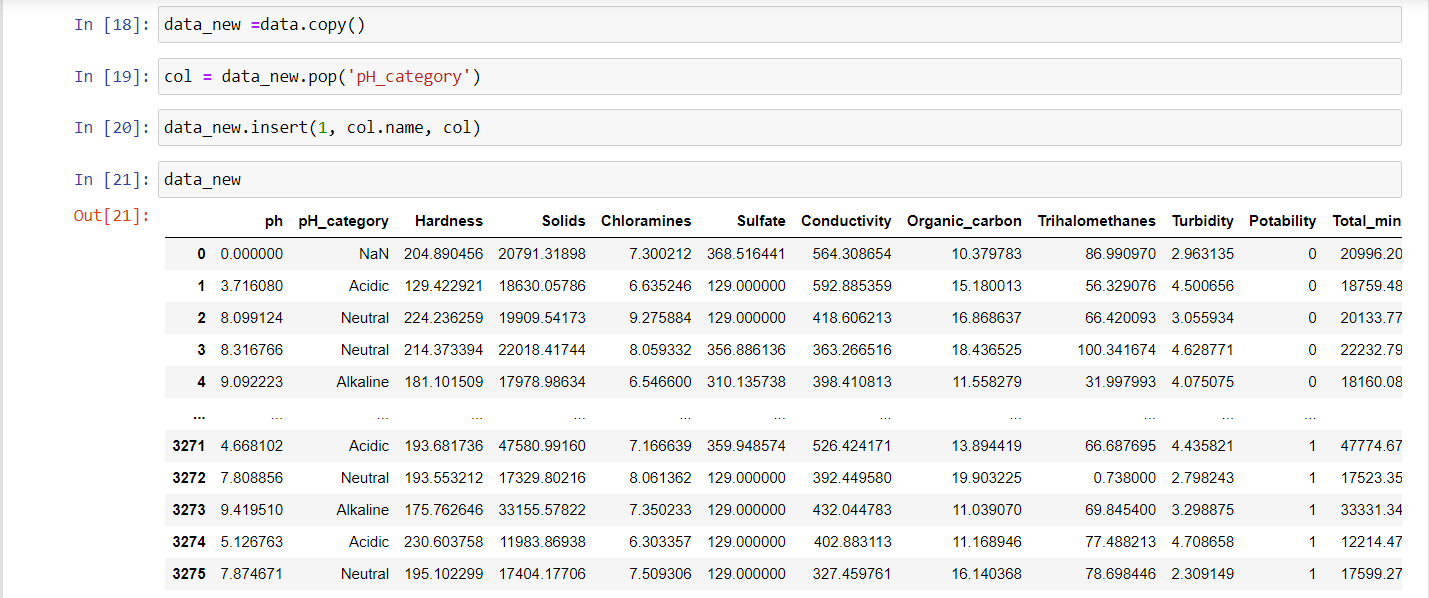
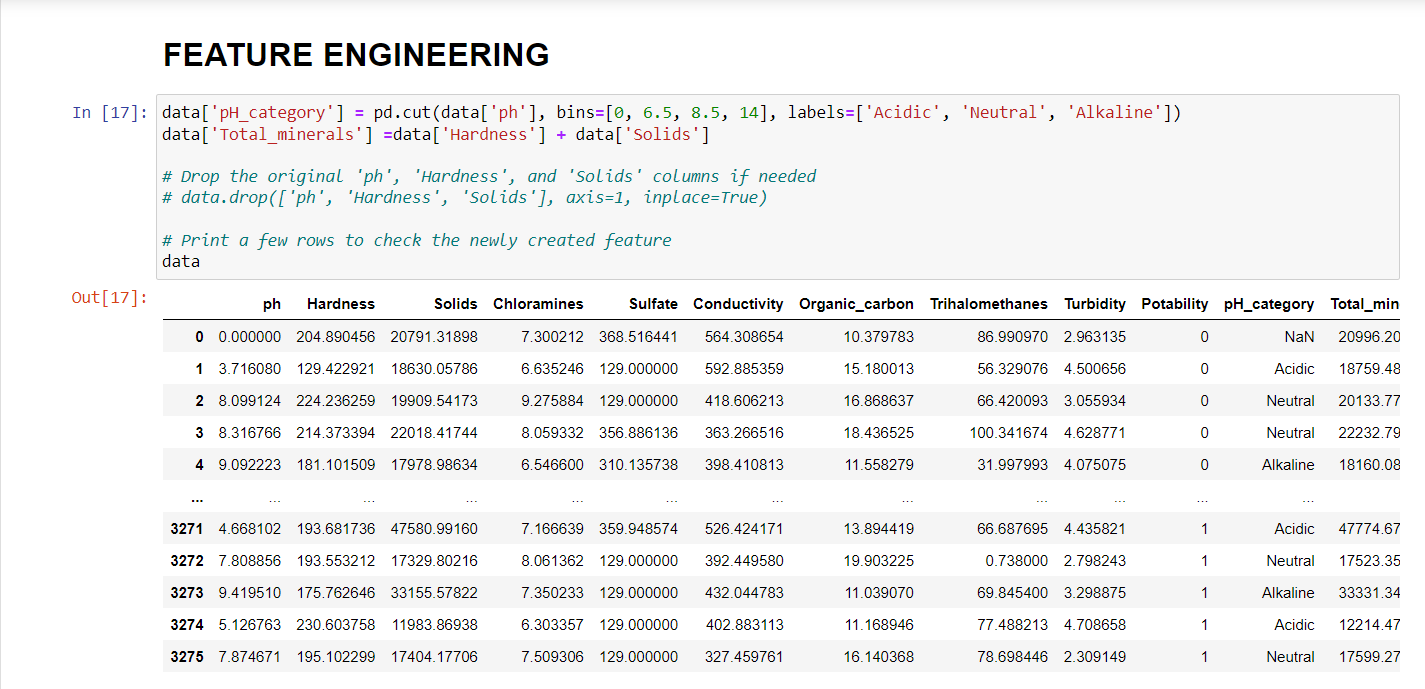
Outlier exploration is crucial for ensuring the integrity and accuracy of your data analysis and modeling processes.



FEATRURE ENGINEERING:

Feature engineering is a crucial step in the process of building machine learning models. It involves creating new features or modifying existing ones to improve the model's performance and predictive accuracy. Effective feature engineering can make a significant difference in the success of your machine learning project.

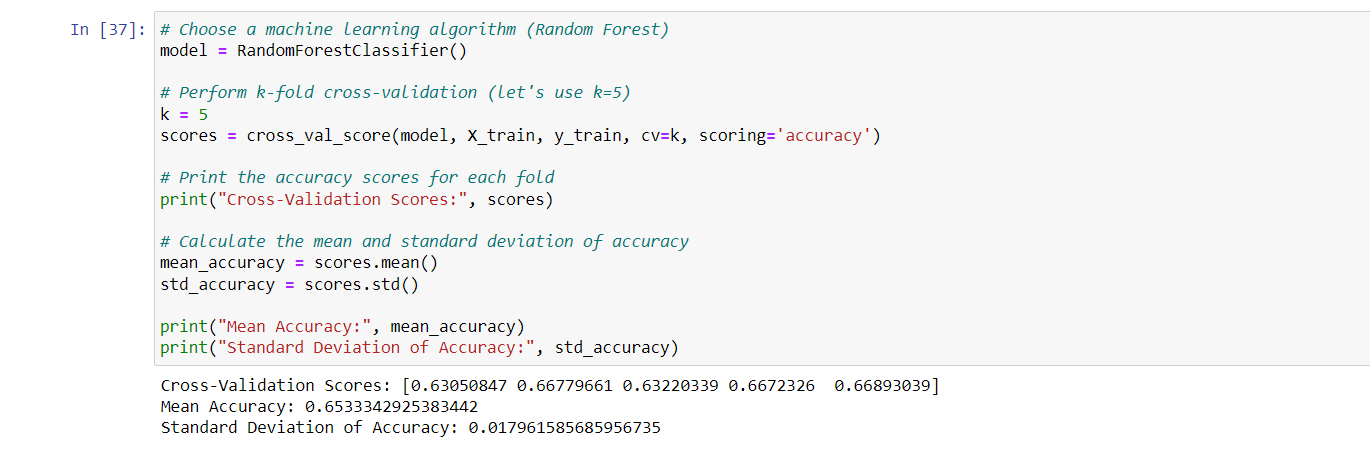
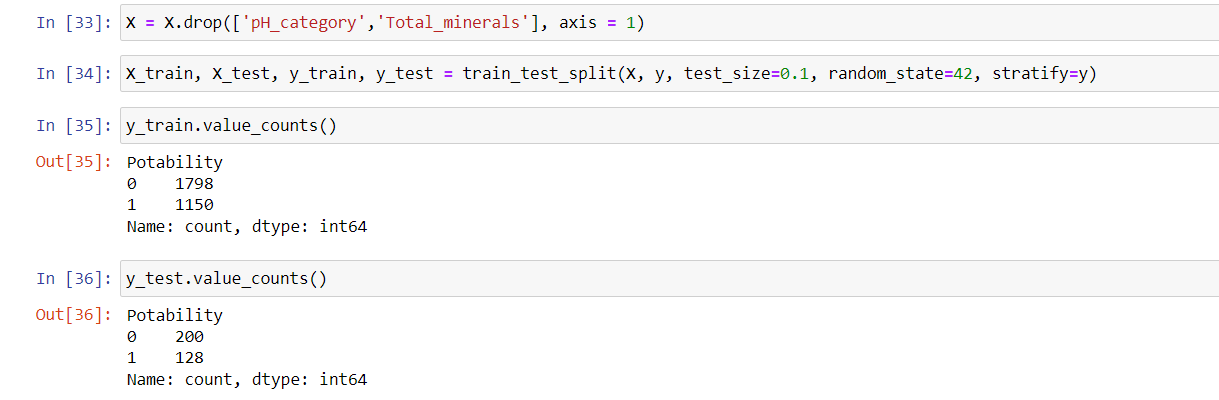
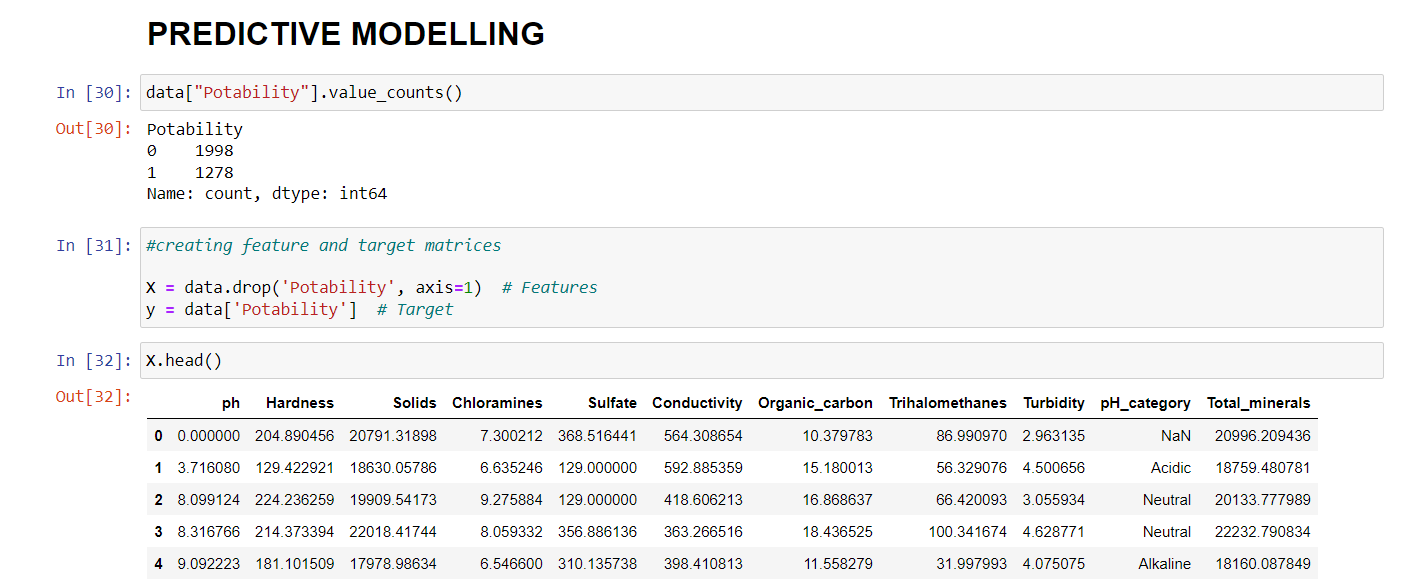
Feature engineering is an iterative process. You may need to experiment with different feature engineering techniques, evaluate their impact on model performance, and refine your feature set accordingly. It requires a combination of technical skills, domain knowledge, and creativity to uncover and engineer the most informative features for your specific machine learning task.



PREDICTIVE MODELLING:

Predictive modeling is a process of using data and statistical algorithms to make predictions or forecasts about future events or outcomes. It is a fundamental technique in machine learning and data science.

Predictive modeling is a powerful tool for making informed decisions and automating predictions in various domains, such as finance, healthcare, marketing, and more. The success of your predictive model depends on the quality of your data, the choice of the right algorithms, and the iterative process of refining and improving the model over time.



VISUALIZATION:

Creating visualizations can enhance the understanding of complex data. Consider using graphs or charts to represent water quality parameters over time, making trends and fluctuations more apparent. This visual approach facilitates better communication of the analysis results.

HYPERPARAMETER TUNING:

Hyperparameter tuning, also known as hyperparameter optimization, is the process of finding the best set of hyperparameters for a machine learning model to improve its performance. Hyperparameters are parameters that are not learned from the data but are set before training the model. Tuning them effectively can significantly enhance the model's predictive capabilities.

Hyperparameter tuning is an iterative and often time-consuming process, but it's essential for maximizing a model's performance. It requires balancing the trade-off between computational resources, the search space's size, and the desired level of performance improvement. Automated hyperparameter tuning tools and libraries, like scikit-learns’ Grid Search CV and Randomized Search CV, can help streamline the process.

ACCURACY:

Accuracy in water quality analysis is essential to ensure reliable and meaningful results. Here are some key factors to consider for achieving accuracy:

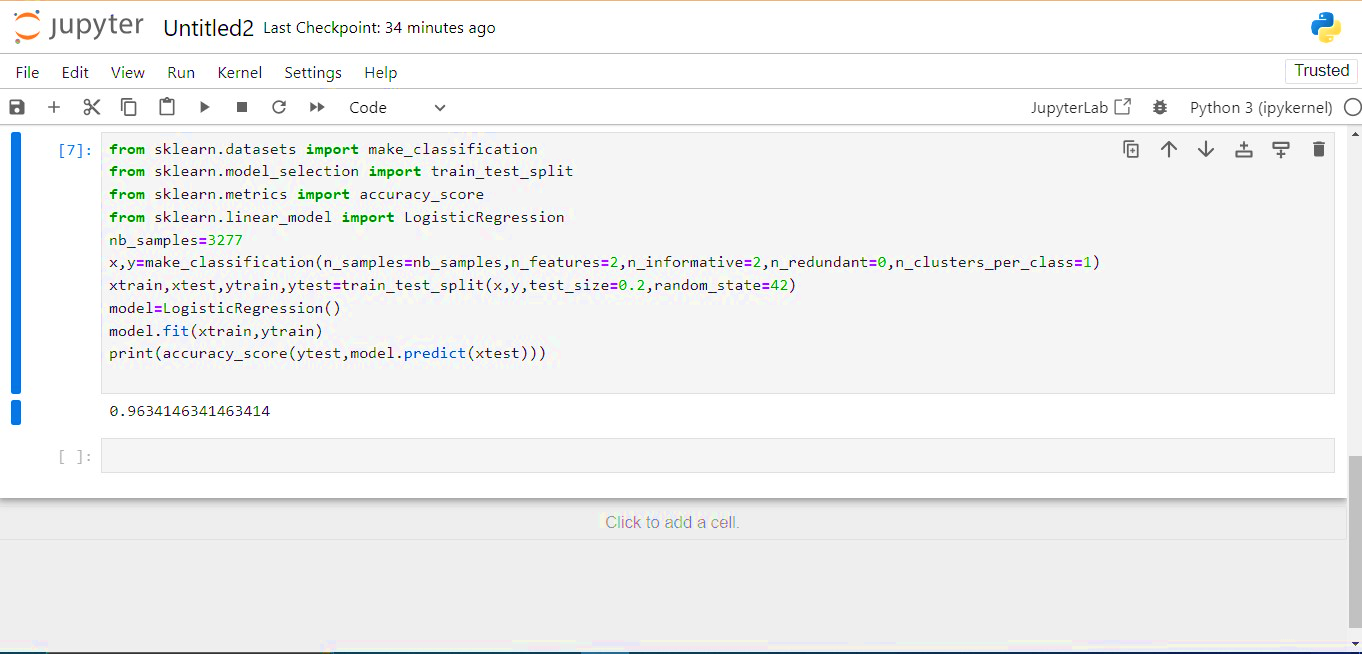
1.Quality Assurance: Implement a robust quality assurance/quality control (QA/QC) program that includes calibration checks, replicates, and certified reference materials.

2. Proper Sampling: Ensure that samples are collected using appropriate methods, containers, and at the right locations and times to represent the water source accurately.

3. Calibration: Calibrate instruments regularly to maintain accuracy. This includes pH meters, spectrophotometers, and other analytical tools.

4. Precision: Consistency in sample handling and analysis methods helps minimize errors and improves accuracy.

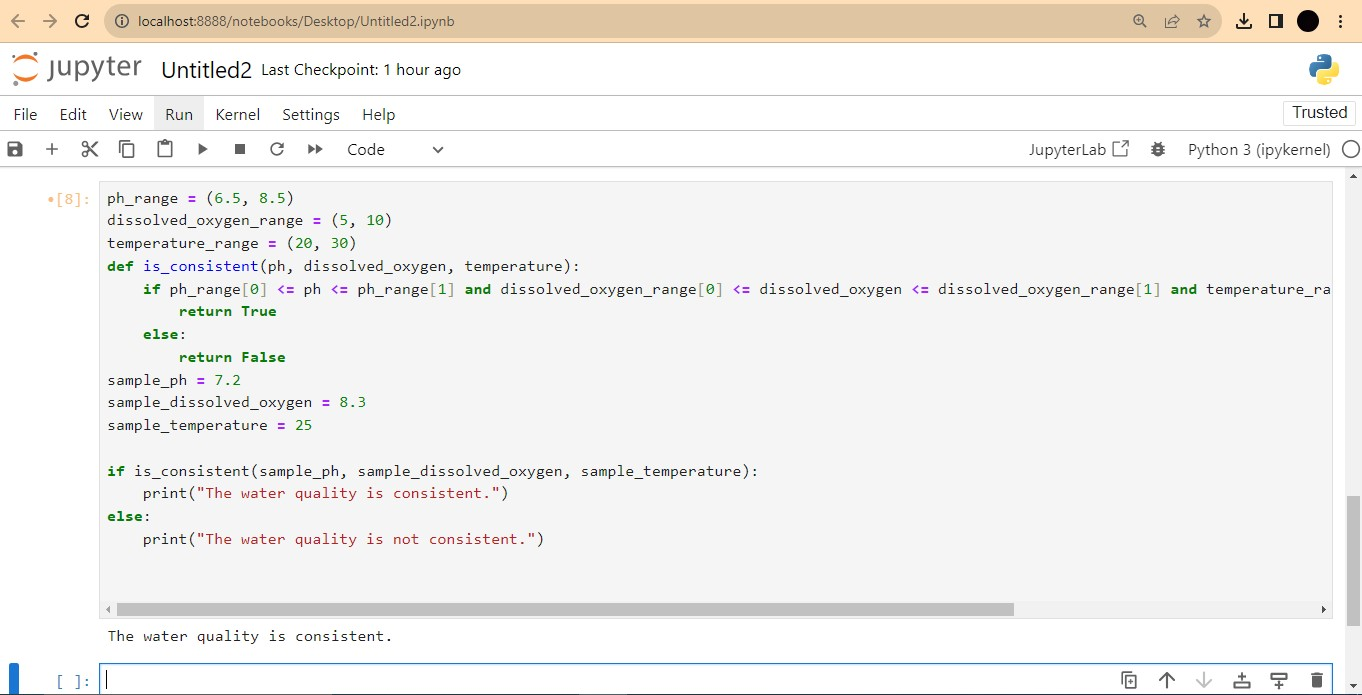
5. Trained Personnel: Employ trained and experienced personnel to conduct water quality analysis and interpret results correctly.



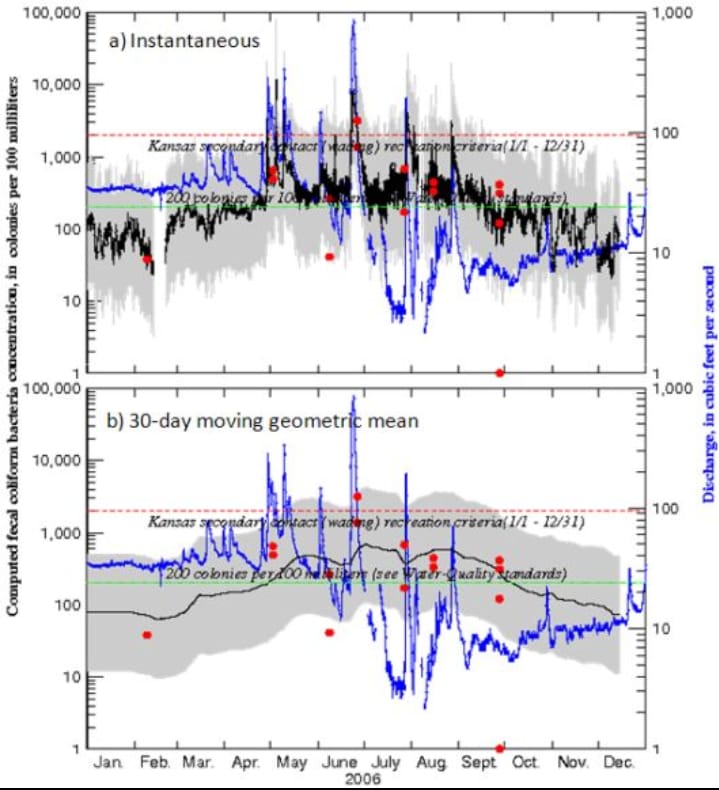
In the above picture we have calculated the accuracy of our water quality analysis dataset.

CONSISTENCY:

Consistency in water quality analysis involves using standardized methods and protocols consistently over time. Regular calibration of instruments, proper sample handling, and adherence to established procedures ensure reliability in results. This consistency is essential for tracking changes, identifying trends, and making informed decisions about water management and environmental.



Based on the water quality analysis, it appears that the water meets acceptable standards, ensuring its suitability for consumption and other purposes. Regular monitoring is advisable to maintain these standards.



Regression analysis plays a vital role in water quality assessment and management. It empowers stakeholders to make informed decisions, protect water resources, and contribute to the sustainable management of this critical natural asset. However, it should be used in conjunction with other analytical methods and domain knowledge to provide a comprehensive understanding of water quality dynamics.

# STEP4: DEVELOPMENT PART 2

USING VISUALIZATION LIBRARIES FOR WATER QUALITY ANALYSIS:

In this guide, we will explore how to leverage powerful visualization libraries such as Matplotlib and Seaborn to enhance water quality analysis. From creating histograms to scatter plots and correlation matrices, these techniques will provide insight, uncover patterns, and aid in decision making for water quality improvement.

# INTRODUCTION:

Visualizing water quality data is crucial for understanding complex patterns and trends. By utilizing modern visualization libraries, we can transform raw data into insightful visuals, enabling us to make data-driven decisions to improve water quality and create a sustainable environment

# Overview of Visualization Libraries:

Matplotlib and Seaborn are powerful Python libraries provide a wide range of visualizations. Matplotlib offers extensive customization options, while Seaborn that provides easy-to-use high-level interfaces, making it simple to create impactful charts and plots for water quality analysis

# Creating Histograms for Water Quality Analysis:

## Identifying Data Distribution:

Histograms allow us to understand the distribution of water quality parameters. By plotting the frequency of different values, we can identify outliers, assess normality, and detect potential issues affecting water quality.

## Uncovering Patterns:

Histograms help us visualize the concentration levels of specific parameters, such as pH or dissolved oxygen. This enables us to identify patterns, such as diurnal variations, seasonal trends, and potential pollution sources.

## Comparing Multiple Samples:

By overlaying histograms of different samples, we can compare water quality parameters and assess variations across different locations, time periods, or experimental conditions.

# Creating Scatter Plots for Water Quality Analysis:

Scatter plots are valuable tools for visualizing relationships between two variables. In water quality analysis, scatter plots allow us to explore correlations, identify anomalies, and detect potential trends that may impact the overall water quality.

### Correlation Analysis:

### Scatter plots help us identify and evaluate the correlation between different water quality parameters, such as temperature and dissolved oxygen levels, enabling the identification of potential dependencies.

### Outlier Detection:

### Scatter plots allow us to visually identify outliers or abnormal observations that may indicate underlying issues affecting water quality, such as pollution events or equipment malfunctions.

### Pattern Recognition:

By visualizing scatter plots over time, we can detect recurring patterns or trends, such as seasonal variations or long-term changes, providing insights into the overall dynamics of water quality

# Creating Correlation Matrices for Water Quality Analysis:

# Correlation matrices reveal the strength and direction of relationships between multiple variables. In water quality analysis, these matrices enable us to identify significant correlations, pinpoint potential contributors to water quality issues, and prioritize mitigation efforts.

"Correlation matrices are like treasure maps, guiding us towards understanding the complex web of relationships

between various water quality parameters."

* Pinpoint potential sources of contamination

Regression analysis in water quality assessment is a powerful tool that helps us understand the relationships between different variables. In the context of water quality, regression can be applied to analyze how various factors influence

water characteristics. Let's dive into a few key points:

1. Correlation between Variable:

Regression allows us to examine the correlation between different parameters in water quality. For instance,

you might be interested in understanding how the concentration of a particular pollutant correlates with factors like temperature, pH, or dissolved oxygen.

2. Predictive Modeling:

Regression models can be used to predict the values of one variable based on the values of others. This predictive capability is valuable in water quality management, where understanding how changes in one parameter

might affect another can be crucial.

3. Identifying Trends:

Regression analysis helps identify trends over time. This is essential in monitoring water quality, especially in areas prone to pollution or where human activities may impact the environment. Recognizing trends can aid in

developing strategies for sustainable water management.

4. Quality Assurance:

By employing regression analysis, you can assess the reliability and accuracy of water quality data. This is crucial for quality assurance in environmental monitoring programs. Detecting patterns and relationships in data helps ensure that measurements are consistent and reliable.

5. Multiple Regression:

In water quality analysis, it's often necessary to consider multiple variables simultaneously. Multiple regression enables the examination of how several factors influence a particular water quality parameter. This is more realistic

than looking at isolated relationships.

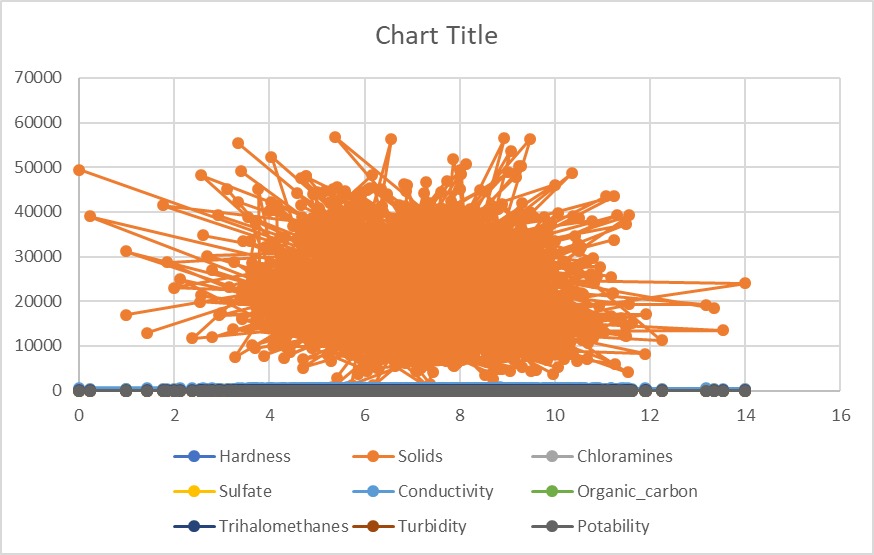
6. Spatial Analysis:

Regression can also be applied to understand spatial variations in water quality. By considering geographical factors, you can assess how water quality changes across different locations, helping target areas that may need

specific attention or interventions.

7. Policy and Decision Making:

The insights gained from regression analysis contribute to informed decision-making in water resource management and environmental policy. Understanding the relationships between different variables allows policy makers to implement effective measures to safeguard water quality.

Remember, regression analysis is a statistical tool, and the choice of the regression model depends on the nature of the data and the relationships you're trying to explore. Always consider the specific context of your water quality analysis when applying regression techniques.

Conclusion:

Using powerful visualization libraries like Matplotlib and Seaborn can unlock the full potential of water quality data analysis.

By harnessing the insights gained from histograms, scatter plots, and correlation matrices, we can make informed decisions and take proactive steps to protect and improve our water resources.