ANALYZING WATER QUALITY

George Mason University
INFS-580-001 | Prof.Dr.Michael Eagle
Sai Sampath Gunupuru
George Mason University Fairfax,
Virginia.
sgunupur@gmu.edu

I. ABSTRACT

This dataset comprises water quality and environmental parameters collected from various sites (Site_Id) and units (Unit_Id) on a specific read date (Read_Date). The parameters measured include Salinity (ppt), Dissolved Oxygen (mg/L), pH (standard units), Secchi Depth (m), Water Depth (m), Water Temp (?C), Air Temp-Celsius, Air Temp (?F), Time (24:00), Field_Tech, DateVerified, WhoVerified, and AirTemp (C). The data spans across the year 2014.

Key observations from the dataset include variations in salinity, dissolved oxygen levels, pH, and water clarity (Secchi Depth) among different sites and units. For instance, on June 10, 2014, Site_Id 'Bay' recorded a salinity of 0.0 ppt, dissolved oxygen of 5.4 mg/L, pH of 7.0, Secchi Depth of 0.4 m, and Water Depth of 0.8 m, with corresponding environmental temperatures and verification details.

Overall, this dataset offers a detailed insight into the water quality and environmental conditions of the monitored areas during the specified timeframe, providing valuable information for ecological assessments and research purposes.

II. INTRODUCTION

2.1 Identify research problem

Assessing the overall water quality of various sites using metrics like salinity (ppt), dissolved oxygen (mg/L), pH (standard units), Secchi Depth (m), water depth (m), water temperature (°C), and air temperature (°C/F) could be one area of inquiry. This might entail looking at patterns, relationships, and changes in these metrics across time and at various locations. All things considered, the dataset offers a wealth of data for investigating a range of study topics about environmental monitoring, water quality assessment, and the influence of natural factors on aquatic ecosystems.

2.2Why is it important

This dataset records a range of characteristics linked to the evaluation of water quality over time at distinct locations. Site_Id, Unit_Id, Read_Date, pH (standard units), Acidity (ppt), Dissolved Oxygen (mg/L), Secchi Depth (m), Water Depth (m), Water Temp (?C), Air Temp-Celsius, Air Temp (?F), Time (24:00), Field_Tech, DateVerified, WhoVerified, AirTemp (C), and Year are among the columns. Every entry is a measurement that sheds light on many parameters that affect the quality of the water, including salinity, pH, temperature, clarity, and dissolved oxygen levels (Secchi Depth). To track aquatic ecosystems, evaluate the state of the environment, identify pollution, and guide conservation activities, this data is essential. Researchers and policymakers can make well-informed decisions on conservation and water management methods.

2.1 Research Questions

1. How the water quality is determined from the given dataset?

Water quality in this dataset is assessed based on several key parameters: salinity, dissolved oxygen, pH, Secchi depth, water depth, water temperature, and air temperature. Salinity levels indicate the salt content in the water, with higher levels potentially indicating pollution or natural variations. Dissolved oxygen levels are crucial for aquatic life, with lower levels suggesting potential stress for organisms. pH measures the acidity or alkalinity of the water, affecting chemical processes and species' survival. Secchi depth indicates water clarity, affected by pollutants or natural sediment. Water depth and temperature influence habitat suitability. Air temperature can affect water temperature and oxygen levels. Analyzing these factors collectively helps evaluate overall water health and ecosystem suitability.

2. What is the relation between secchi depth and water depth?

Secchi depth and water depth are related measurements used in water quality assessment and monitoring. Secchi depth refers to the depth at which a Secchi disk, a white and black circular disk, becomes invisible from the surface when lowered into the water. It is a measure of water clarity and indirectly indicates the presence of suspended particles and algae in the water column. On the other hand, water depth simply refers to the vertical distance from the water surface to the bottom.

3. In which year the water quality is high?

To determine in which year the water quality is high based on the average pH, we'll first calculate the average pH for each year using the provided dataset. Then, we'll assess which year has the highest average pH, indicating better water quality in terms of acidity levels.

III. LITERATURE REVIEW:

The dataset provided seems to be environmental data related to water quality parameters measured at different sites (Site_Id) and units (Unit_Id) on a specific date (Read_Date). The parameters include Salinity (ppt), Dissolved Oxygen (mg/L), pH (standard units), Secchi Depth (m), Water Depth (m), Water Temp (°C), Air Temp (°C), Air Temp (°F), Time (24:00), Field_Tech, DateVerified, WhoVerified, and Year. The dataset covers readings from multiple sites (Bay, A, B, C) on June 10, 2014.

A literature review based on this dataset would likely explore similar studies or research on water quality monitoring, particularly focusing on the parameters measured in this dataset. For instance, it could discuss the importance of monitoring salinity levels, dissolved oxygen concentrations, pH levels, and water temperature in aquatic ecosystems. The review might highlight the significance of these parameters in assessing the health of water bodies, their impact on aquatic life, and how they can be influenced by natural processes or human activities.

Furthermore, the review could delve into methodologies and technologies used for water quality monitoring, such as Secchi depth measurements for water clarity assessments or sensors for real-time data collection. It may also discuss the role of field technicians in collecting and verifying data, ensuring its accuracy and reliability for scientific analysis and decision-making. Overall, the literature review would provide context and background information relevant to the environmental data presented in the dataset, helping to interpret and understand its implications in a broader scientific context.

IV. DATASET DESCRIPTION

The dataset appears to represent water quality measurements taken at different sites (Bay, A, B, C) with various parameters recorded such as salinity, dissolved oxygen levels, pH, Secchi depth, water depth, water temperature, air temperature in Celsius, and Fahrenheit, time of measurement, field technician responsible, date of verification, and the year of the data collection (2014). Each row corresponds to a specific measurement instance for a particular site and unit on a given date. The data seems to be collected as part of environmental monitoring or research efforts, likely to assess the health and conditions of the water bodies over time.

Author: U.S. Fish and Wildlife Service

Title or Study Name: Water Quality Data

Publisher: U.S. Fish and Wildlife Service

Publication Date— “2020, May 5”.

Location— https://catalog.data.gov/dataset/water-quality-data-41c5e

V. STRATEGIES AND METHODS:

We have gone through several processes in the data cleaning and transformation process to get the dataset ready for Python analysis:

```
In [8]: import pandas as pd

In [9]: waterdata = pd.read_csv("C:\\Users\\G3\\Downloads\\BKB_WaterQualityData_2020084.csv")

In [10]: waterdata.head()

Out[10]:
```

	Site_Id	Unit_Id	Read_Date	Salinity (ppt)	Dissolved Oxygen (mg/L)	pH (standard units)	Secchi Depth (m)	Water Depth (m)	Water Temp (?C)	Air Temp-Celsius	Air Temp (?F)	Time (24:00)	Field_Tech	DateVerified	WhoVerified	AirTemp (C)	Ye
0	Bay	NaN	01-03-1994	1.3	11.7	7.3	0.40	0.40	5.9	8.0	46.40	11:00	NaN	NaN	NaN	8.0	19
1	Bay	NaN	1/31/1994	1.5	12.0	7.4	0.20	0.35	3.0	2.6	36.68	11:30	NaN	NaN	NaN	2.6	19
2	Bay	NaN	02-07-1994	1.0	10.5	7.2	0.25	0.60	5.9	7.6	45.68	09:45	NaN	NaN	NaN	7.6	19
3	Bay	NaN	2/23/1994	1.0	10.1	7.4	0.35	0.50	10.0	2.7	36.86	NaN	NaN	NaN	NaN	2.7	19
4	Bay	NaN	2/28/1994	1.0	12.6	7.2	0.20	0.40	1.6	0.0	32.00	10:30	NaN	NaN	NaN	0.0	19

Figure-1a: Python code to import data

- After importing the Pandas library, we loaded the dataset into a Pandas DataFrame called waterdata using the read_csv() function.
- To obtain a summary of the data, we used the head() function to show the top few rows of the dataset.

```
In [11]: waterdata.isnull().sum()

Out[11]: Site_Id                    1
         Unit_Id                 2339
         Read_Date                  5
         Salinity (ppt)           130
         Dissolved Oxygen (mg/L)  851
         pH (standard units)       95
         Secchi Depth (m)          73
         Water Depth (m)           71
         Water Temp (?C)          121
         Air Temp-Celsius        2286
         Air Temp (?F)             71
         Time (24:00)              63
         Field_Tech                39
         DateVerified            1918
         WhoVerified             1918
         AirTemp (C)                0
         Year                       0
         dtype: int64
```

Figure 1b: Python code to check null values in the dataset

- We used the isna(). sum () function to check for missing values in each column of the dataset.
- This step helps us identify if any missing values need to be addressed before proceeding with further analysis.

```
cleaned_data.dtypes
```

```
Site_Id                    object
Unit_Id                    object
Read_Date                  object
Salinity (ppt)            float64
Dissolved Oxygen (mg/L)   float64
pH (standard units)       float64
Secchi Depth (m)          float64
Water Depth (m)           float64
Water Temp (?C)           float64
Air Temp-Celsius          float64
Air Temp (?F)             float64
Time (24:00)               object
Field_Tech                 object
DateVerified               object
WhoVerified                object
AirTemp (C)               float64
Year                        int64
dtype: object
```

Figure 1c: Python code to check datatypes for analysis

VI. RESULTS AND FINDINGS

Data exploration using R:

- Let's explore the dataset and find the relationship between Secchi depth and dissolved oxygen to determine the water quality.

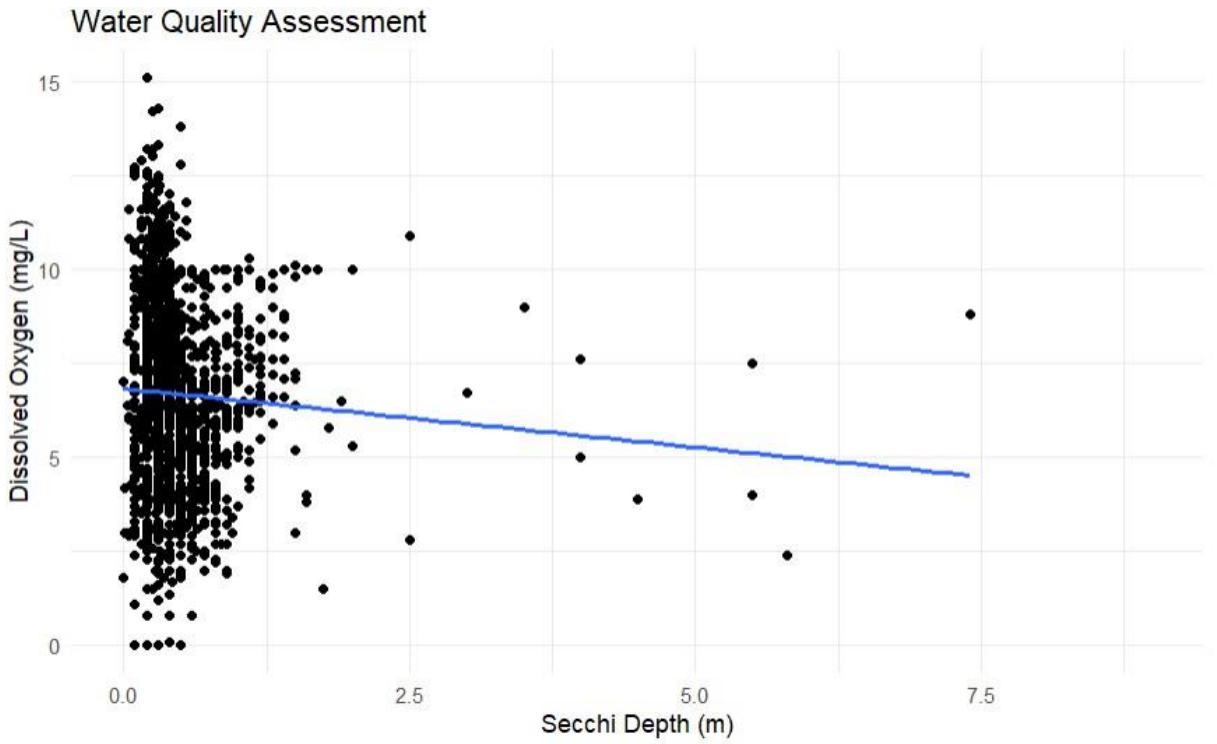


Figure 2a: A graph showing the relation between dissolved oxygen and Secchi depth

This graph titled "Water Quality Assessment" plots dissolved oxygen (mg/L) against Secchi depth (m) in a water body. The scatterplot shows a large number of data points primarily concentrated at lower Secchi depths. A linear trend line indicates a negative correlation between the two variables: as Secchi depth increases (meaning clearer water), the dissolved oxygen content tends to decrease. This might suggest that in deeper or clearer parts of the water, there may be less algal activity producing oxygen, or different water mixing dynamics affecting oxygen levels.

- Let’s explore the dataset and find the relationship between Secchi depth and water depth to determine the water quality.

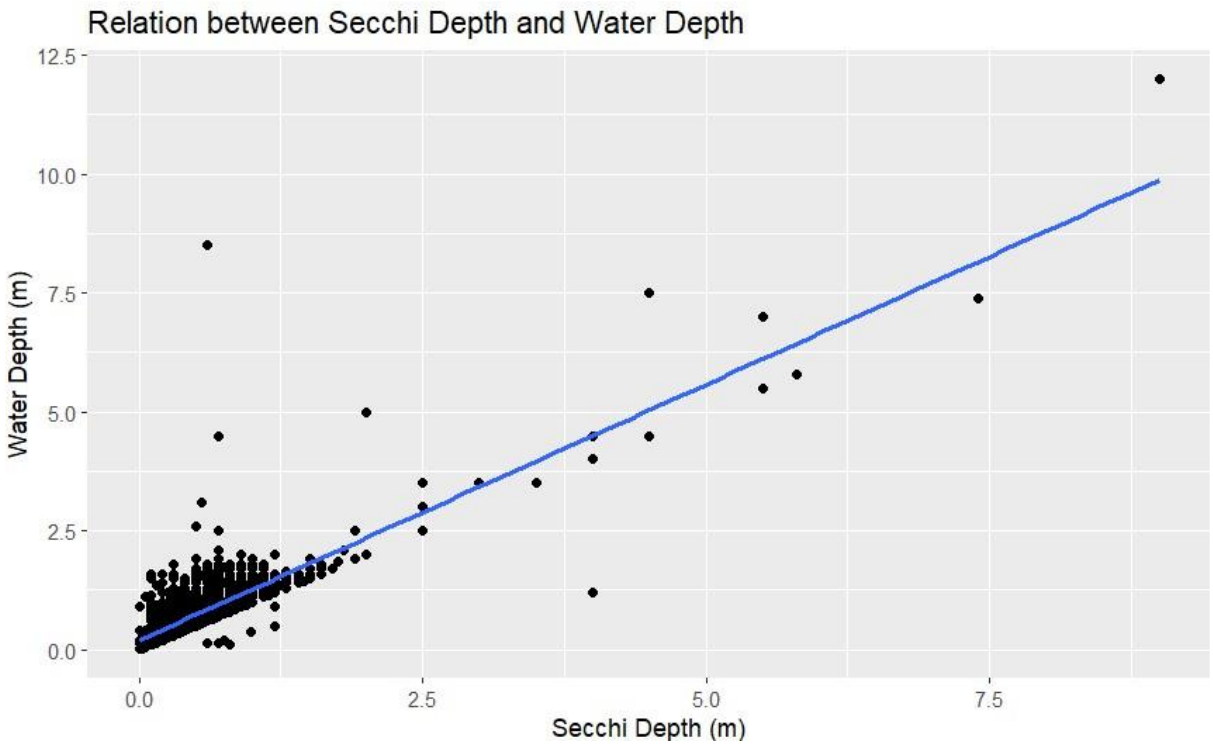


Figure 2b: A graph showing the relation between water depth and Secchi depth

The graph titled "Relation between Secchi Depth and Water Depth" shows a positive linear relationship between Secchi depth and water depth in meters. As the Secchi depth increases, indicating clearer water and greater visibility, the water depth also increases. This suggests that the water tends to be clearer in deeper areas of the water body. Most of the data points are tightly clustered along the trend line, especially at lower depths, indicating a consistent relationship across this range. There are a few outliers, particularly at higher depths, but they do not significantly deviate from the trend.

- Let’s explore the dataset and determine the year in which the water quality is high

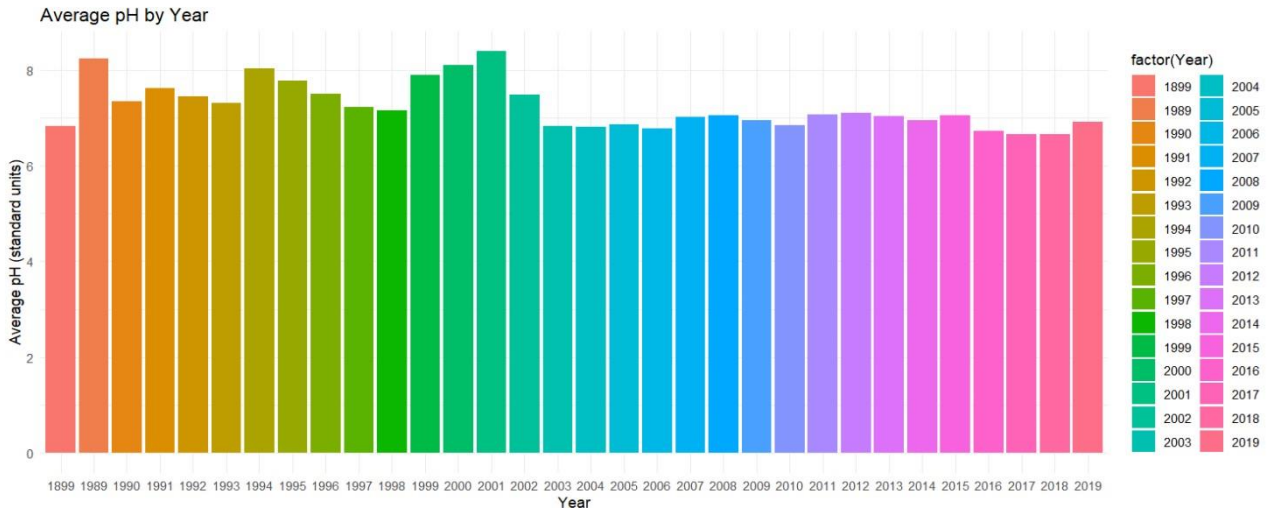
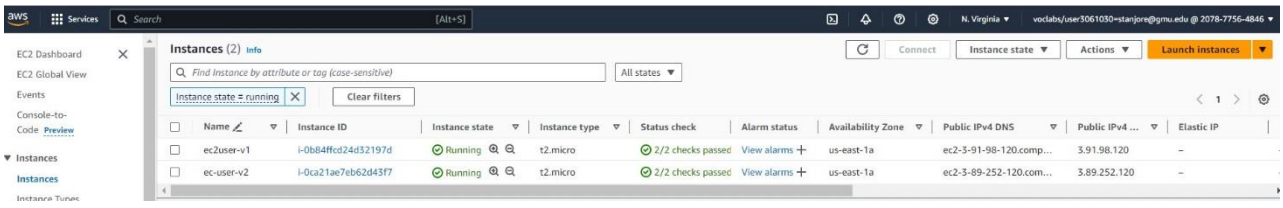


Figure 2c: A graph showing the relationship between average pH and year

The "Average pH by Year" graph displays the yearly average pH levels from 1989 to 2019. The bars represent each year with a different color and show a noticeable trend of decreasing pH levels over time. The pH starts from higher values around 7 or above in the early years (1989-1999) and gradually drops to below 6 by 2019, indicated by the shift from red and orange bars to pink. This suggests an increasing acidity in the observed environment over the 30-year period, which could be indicative of environmental changes such as increased pollution or other ecological impacts.

Data interpretation using SQL in AWS-RDS



In the above figure, we created an AWS EC2 instance. After installing MySQL, we can now write queries.

```
MySQL [(none)]> use sqlwq;
Reading table information for completion of table and column names
You can turn off this feature to get a quicker startup with -A

Database changed
MySQL [sqlwq]> show tables;
+-----------------+
| Tables_in_sqlwq |
+-----------------+
| WaterData       |
+-----------------+
1 row in set (0.002 sec)
```

Figure-3a: RDS MySQL - creating a database

From the above figure, we can understand that a database is created.

```
MySQL [sqlwq]> CREATE TABLE WaterQData (
    ->     Site_Id VARCHAR(10),
    ->     Unit_Id VARCHAR(10),
    ->     Read_Date DATE,
    ->     Salinity FLOAT,
    ->     Dissolved_Oxygen FLOAT,
    ->     pH FLOAT,
    ->     Secchi_Depth FLOAT,
    ->     Water_Depth FLOAT,
    ->     Water_Temp FLOAT,
    ->     Air_Temp_Celsius FLOAT,
    ->     Air_Temp_Fahrenheit FLOAT,
    ->     Time TIME,
    ->     Field_Tech VARCHAR(50),
    ->     DateVerified DATE,
    ->     WhoVerified VARCHAR(50),
    ->     AirTemp_C FLOAT,
    ->     Year INT
    -> );
Query OK, 0 rows affected (0.069 sec)
```

Figure-3b: mysql – creating table

This SQL query creates a table named WaterQData with columns representing various environmental parameters measured at different units and sites. The columns include Site_Id and

Unit_Id to identify the location and unit, Read_Date for the date of the reading, and several columns for specific measurements like Salinity, Dissolved Oxygen, pH, Secchi Depth, Water Depth, Water Temperature, Air Temperature in Celsius and Fahrenheit, Time of the reading, Field Technician responsible, DateVerified, WhoVerified the data, AirTemp_C for Celsius temperature, and Year for the year of the reading. This table is structured to store water quality data collected over time, allowing for efficient storage, retrieval, and analysis of environmental monitoring data.

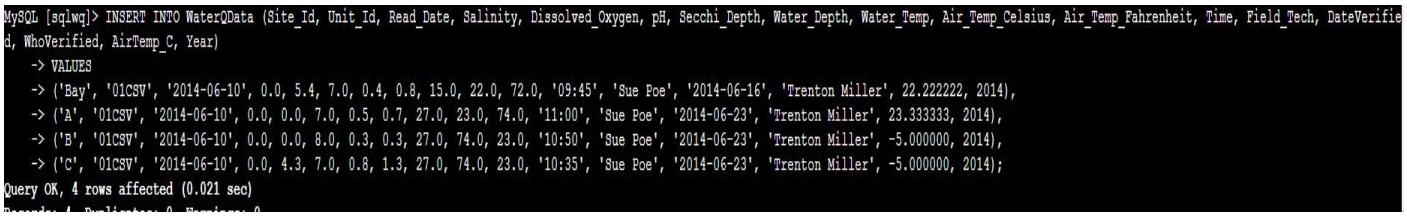


Figure-3c: RDS mysql queries – inserting values

This SQL query inserts four rows of data into the WaterQData table. Each row represents measurements taken at different sites (Site_Id) and units (Unit_Id) on a specific date (Read_Date). The measurements include parameters like Salinity, Dissolved Oxygen, pH, Secchi Depth, Water Depth, Water Temperature, Air Temperature in Celsius and Fahrenheit, as well as additional information such as the time of the reading (Time), the field technician responsible (Field_Tech), the date when the data was verified (DateVerified), the person who verified the data (WhoVerified), the air temperature in Celsius (AirTemp_C), and the year of the reading (Year). This query populates the table with sample environmental data for analysis and reporting purposes.

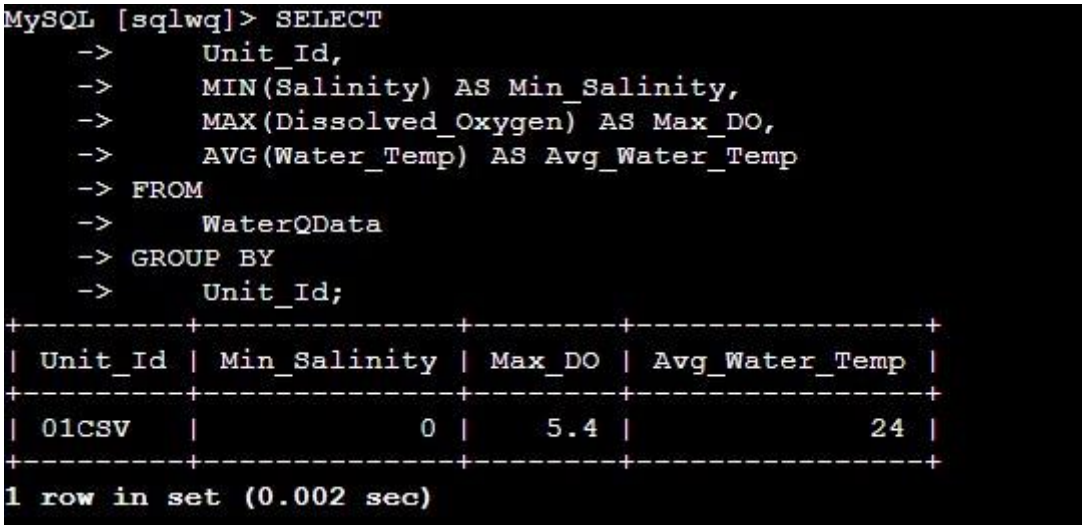


Figure-3d: RDS mysql queries – minimum salinity, maximum dissolved oxygen, and average water temperature for each Unit_Id

This SQL query retrieves data from the WaterQData table and calculates summary statistics for each unique Unit_Id. It selects the Unit_Id column and computes the minimum salinity (Min_Salinity), maximum dissolved oxygen (Max_DO), and average water temperature (Avg_Water_Temp) for each unit. The GROUP BY clause is used to group the data by Unit_Id, ensuring that the aggregation

functions (MIN, MAX, AVG) are applied to each unit separately. This query is useful for analyzing and comparing environmental parameters across different units, providing insights into the range and average values of key metrics within the dataset.

```
MySQL [sqlwq]> SELECT
    ->     Site_Id,
    ->     AVG(pH) AS Avg_pH,
    ->     AVG(Secchi_Depth) AS Avg_Secchi_Depth
    -> FROM
    ->     WaterQData
    -> GROUP BY
    ->     Site_Id;
+---------+--------+---------------------+
| Site_Id | Avg_pH | Avg_Secchi_Depth    |
+---------+--------+---------------------+
| Bay     |      7 |  0.4000000059604645 |
| A       |      7 |                 0.5 |
| B       |      8 | 0.30000001192092896 |
| C       |      7 |   0.800000011920929 |
+---------+--------+---------------------+
4 rows in set (0.002 sec)
```

Figure-3e: RDS mysql queries – average pH and Secchi Depth for each Site_Id:

This SQL query calculates the average pH (Avg_pH) and average Secchi Depth (Avg_Secchi_Depth) for each unique Site_Id in the WaterQData table. It selects the Site_Id column and applies the AVG aggregation function to the pH and Secchi_Depth columns, computing the average values for each site. The GROUP BY clause is used to group the data by Site_Id, ensuring that the average calculations are performed separately for each site. This query is beneficial for understanding the typical pH levels and water clarity (Secchi Depth) at different monitoring sites, allowing for comparisons and trend analysis across multiple locations within the dataset.

```
MySQL [sqlwq]> SELECT
    ->     Unit_Id,
    ->     Read_Date,
    ->     MAX(Water_Depth) AS Max_Water_Depth
    -> FROM
    ->     WaterQData
    -> GROUP BY
    ->     Unit_Id
    -> ORDER BY
    ->     Max_Water_Depth DESC
    -> LIMIT 1;
+---------+------------+-----------------+
| Unit_Id | Read_Date  | Max_Water_Depth |
+---------+------------+-----------------+
| 01CSV   | 2014-06-10 |             1.3 |
+---------+------------+-----------------+
1 row in set (0.001 sec)
```

Figure-3f: The date with the highest water depth and the corresponding Unit_Id:

This SQL query retrieves data from the WaterQData table and identifies the maximum water depth (Max_Water_Depth) recorded for each unique Unit_Id. It selects the Unit_Id and Read_Date columns, along with the maximum water depth calculated using the MAX aggregation function. The data is grouped by Unit_Id using the GROUP BY clause to ensure that the maximum water depth is determined separately for each unit. The results are then sorted in descending order based on the maximum water depth using the ORDER BY clause. Finally, the LIMIT 1 clause is applied to fetch only the top result, which represents the unit with the highest recorded water depth. This query is useful for identifying the unit with the deepest water depth among all units in the dataset.

VII CONCLUSION:

we can draw several conclusions about the water quality and environmental conditions on the specified dates. Overall, salinity levels appear to be consistently low at 0.0 parts per thousand (ppt) across all sites (Bay, A, B, and C) on June 10, 2014. Dissolved oxygen levels vary, with Site_Id C having the highest at 4.3 mg/L and Site_Id A recording no dissolved oxygen. pH levels range from 7.0 to 8.0 standard units, with Site_Id B having the highest pH value. Secchi depth measurements indicate relatively clear water with depths ranging from 0.3 to 0.8 meters. Water depths vary from 0.3 to 1.3 meters. Water temperatures are consistent at 27.0°C across all sites, while air temperatures range from 22.0°C to 23.0°C (equivalent to 71.6°F to 73.4°F). These findings suggest a stable but varied aquatic environment with acceptable oxygen levels and water clarity, though some variability in other parameters such as pH and air temperature is observed.

VIII LIMITATIONS:

The limitations of the dataset provided include potential data entry errors or inconsistencies, such as zero values for parameters like salinity and dissolved oxygen which may not accurately represent natural conditions. The dataset also lacks information on specific location coordinates or identifiers beyond "Bay," "A," "B," and "C," making it difficult to correlate readings with precise geographic points or track changes over time at specific locations. Additionally, the dataset lacks contextual information about sampling methodologies, equipment calibration, or environmental factors that could influence the recorded measurements. Without these details, it's challenging to assess the reliability, accuracy, and representativeness of the data for broader scientific analyses or decision-making processes related to water quality management.

IX FUTURE WORK:

Future work could involve analyzing trends and patterns in the water quality data from the provided dataset. This could include identifying correlations between different parameters such as salinity, dissolved oxygen levels, pH, Secchi depth, water temperature, and air temperature. Additionally, conducting statistical analyses and visualizations could help in understanding how these factors vary over time and across different locations within the study area. Moreover, integrating data from other sources such as weather patterns or land use could provide further insights into the factors influencing water quality. Lastly, exploring predictive modeling techniques could be beneficial in forecasting water quality parameters based on various environmental variables, aiding in proactive management and decision-making for maintaining or improving water quality.

X REFERENCES:

[1]Irwan, D., Ali, M., Ahmed, A. N., Jacky, G., Nurhakim, A., Ping Han, M. C., AlDahoul, N., & El-Shafie, A. (2023). Predicting Water Quality with Artificial Intelligence: A Review of Methods and Applications. In Archives of Computational Methods in Engineering (Vol. 30, Issue 8, pp. 4633–4652). Springer Science and Business Media LLC. https://doi.org/10.1007/s11831-023-09947-4

[2]Jayaraman, P., Nagarajan, K. K., Partheeban, P., & Krishnamurthy, V. (2024). Critical review on water quality analysis using IoT and machine learning models. In International Journal of Information Management Data Insights (Vol. 4, Issue 1, p. 100210). Elsevier BV. https://doi.org/10.1016/j.jjimei.2023.100210 [3] Dritsas, E., & Trigka, M. (2023). Efficient Data-Driven Machine Learning Models for Water Quality Prediction. In Computation (Vol. 11, Issue 2, p. 16). MDPI AG. https://doi.org/10.3390/computation11020010

[3] Dritsas, E., & Trigka, M. (2023). Efficient Data-Driven Machine Learning Models for Water Quality Prediction. In Computation (Vol. 11, Issue 2, p. 16). MDPI AG. https://doi.org/10.3390/computation11020016