Data Analytics Internship Program 2024

Final Project Presentation

Project Name :Impacts of droughts and heatwaves on river water quality worldwide

Unique ID : IBM3624

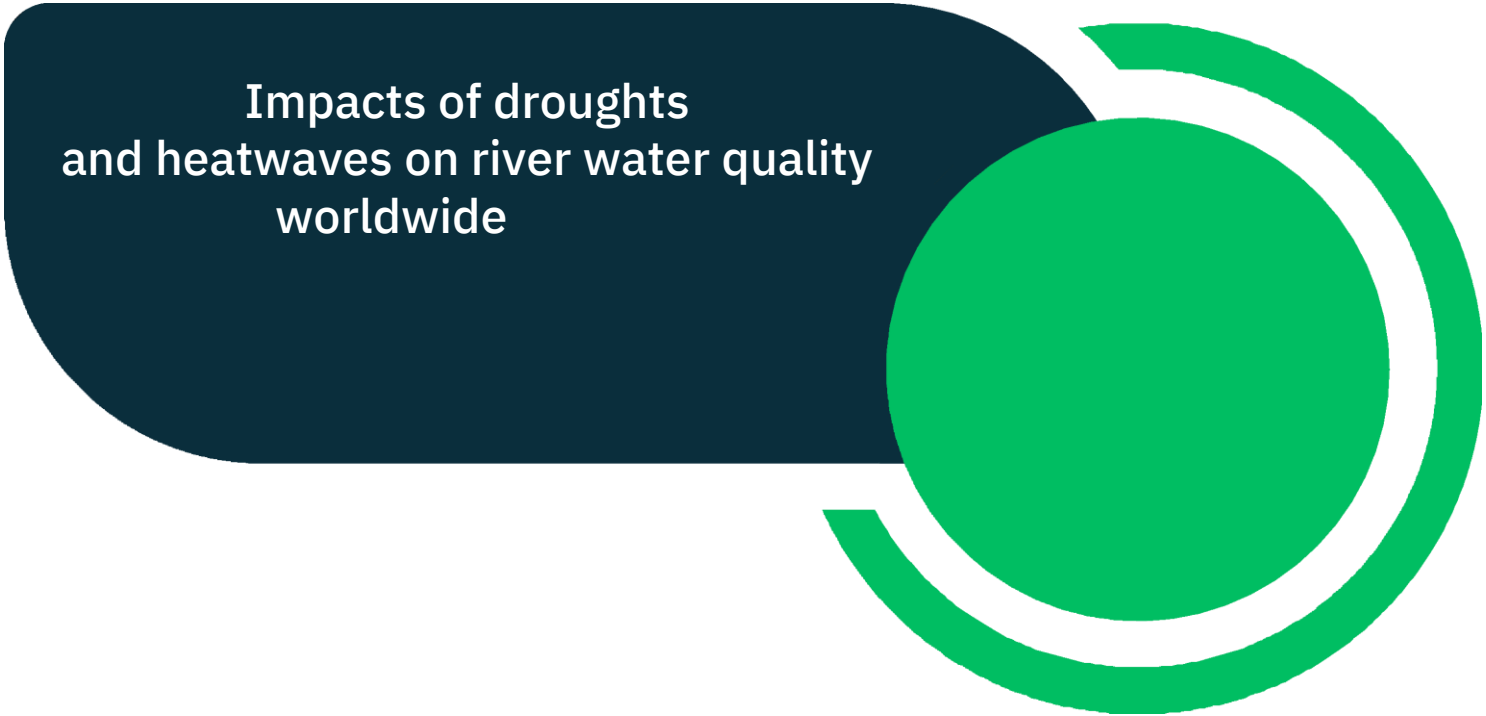
Team Name : The Codex Crew

College Name : Sushant University

Overview

PROJECT TITLE

Impacts of droughts
and heatwaves on river water quality
worldwide



TEAM DETAILS:

1.Anmol Bhatnagar
2. Ankit Tomar
3. Yash Dubal
4. Aditya Lohia
5. Hemant Pandey
6. Krishna Pathak

Objective of the project

- This project aims to leverage the power of Big Data analytics and machine learning (ML) algorithms to enhance the understanding and identification of water quality patterns.
- Collect diverse water quality data from regions of interest.
- Process and analyze data to extract insights on key parameters.
- Develop predictive models for water quality classification and environmental variable forecasting.
- Evaluate model performance and validate predictions for real-world applicability.
- Extract actionable insights for decision-making in water resource management.

Data Collection and Preprocessing

Data Extraction:

- Used pandas in Python to read CSV data.
- Imported data from Gemstat for Indus and Godavari regions in India.
- Employed pd.read_csv() function with specified delimiters.

Data Concatenation:

- Merged data from multiple sources
- Combined Indus and Godavari datasets into one DataFrame.
- Saved merged DataFrame to a new CSV for further analysis.

Data Collection and Preprocessing

Data Preproceesing:

- Removed irrelevant columns.
- Identified unique parameters.
- Created a parameter dictionary and converted it into a DataFrame.
- Combined original and parameter DataFrames for a comprehensive dataset.

Missing Data Handling:

- Imputed missing numerical values with mean.
- Replaced missing categorical data with mode (most frequent values).

Feature Engineering:

- Created new features like dew point temperature from existing data.
- Conducted PCA for dimensionality reduction.
- Utilized K-Means for clustering to uncover data patterns.

Exploratory Data Analysis

Histogram Analysis:

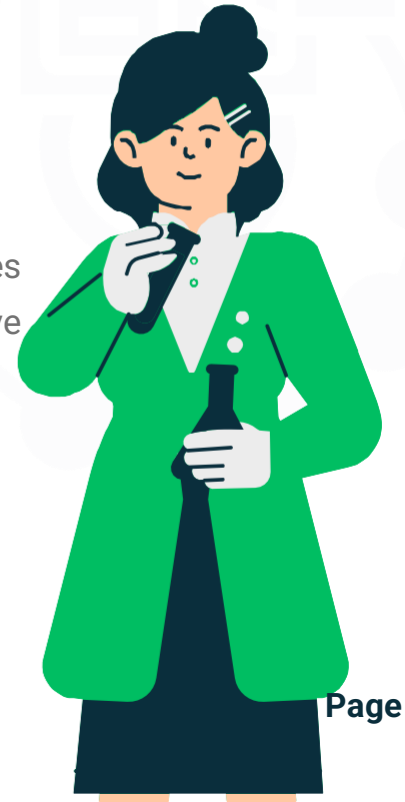
- Utilized histograms with KDE to visualize the distribution of key variables.
- Highlighted central tendency measures such as mean and median for each variable using vertical dashed lines.

Correlation Analysis:

- Employed a heatmap to visualize pairwise correlations between different variables
- Correlation coefficients were annotated on the heatmap to identify strong positive or negative relationships.

Principal Component Analysis (PCA):

- Examined the explained variance ratio to determine the number of principal components required to represent the dataset adequately.
- Visualized clusters in the PCA space to identify any inherent patterns or groupings within the data.



Exploratory Data Analysis

Box plot Analysis:

- Conducted box plot analysis for key variables such as Turbidity, Water Temperature, pH, and Rainfall.
- Identified outliers and assessed the spread and central tendency of each variable.

Time-Series Analysis:

- Plotted individual variables as a function of time to observe temporal trends.
- Analyzed how these variables fluctuated over time and identified any seasonal patterns or anomalies.

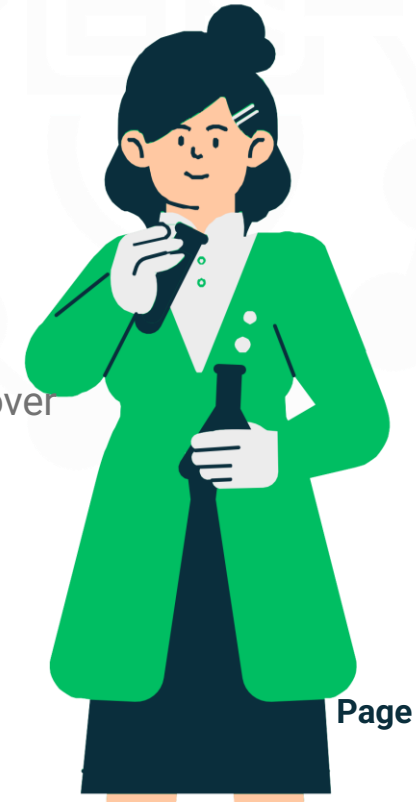
Comparative Analysis:

- Compared the distribution of variables across different categories or datasets to identify variations and trends.
- Box plots were utilized to compare the distribution of Turbidity, Water Temperature, and Rainfall.

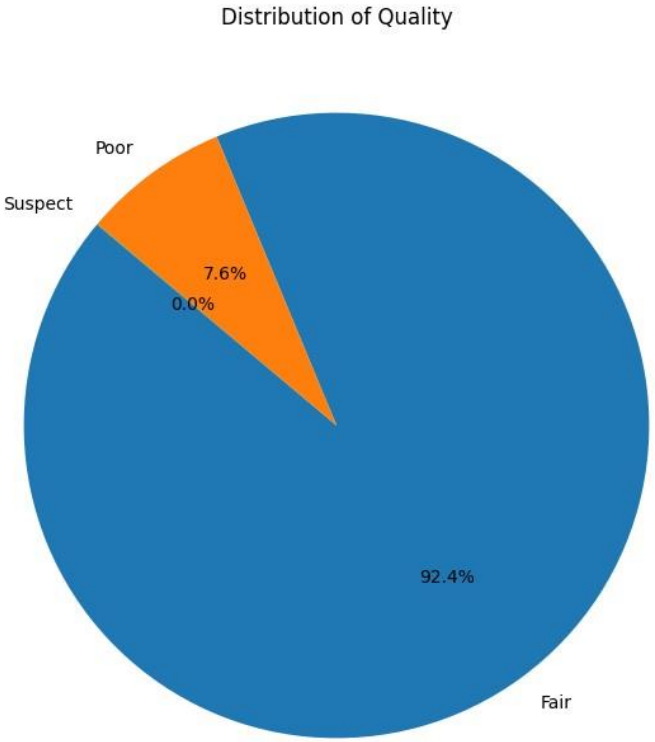
Heat Maps Generation

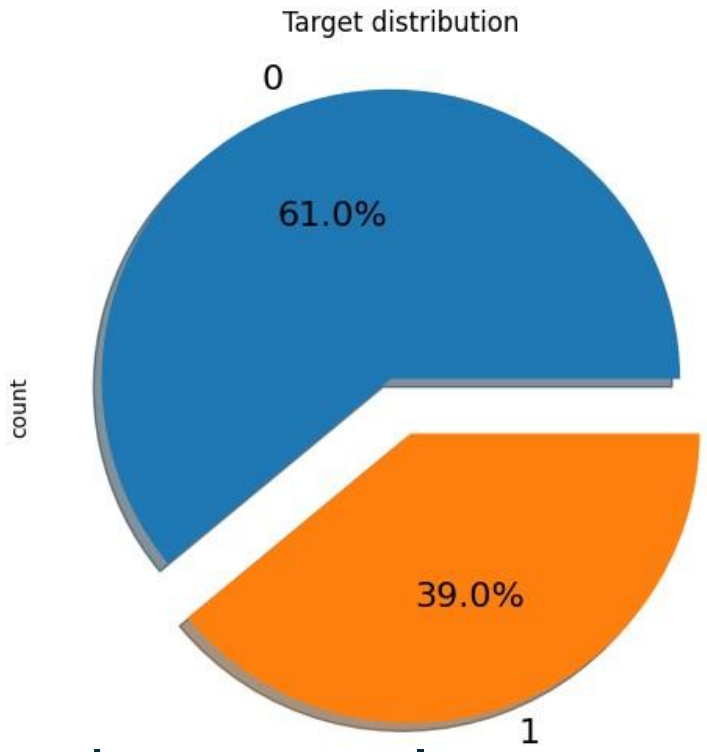
Heat Maps for generation:

- Heat maps generated for Canada and Australia.
- Coverage: 20 sites for Australia and 58 sites for Canada.
- Analyzed parameters: pH, Rainfall, Water Temperature, and Turbidity.
- Visualization techniques: Box plots, heat maps, plotting parameters over time.



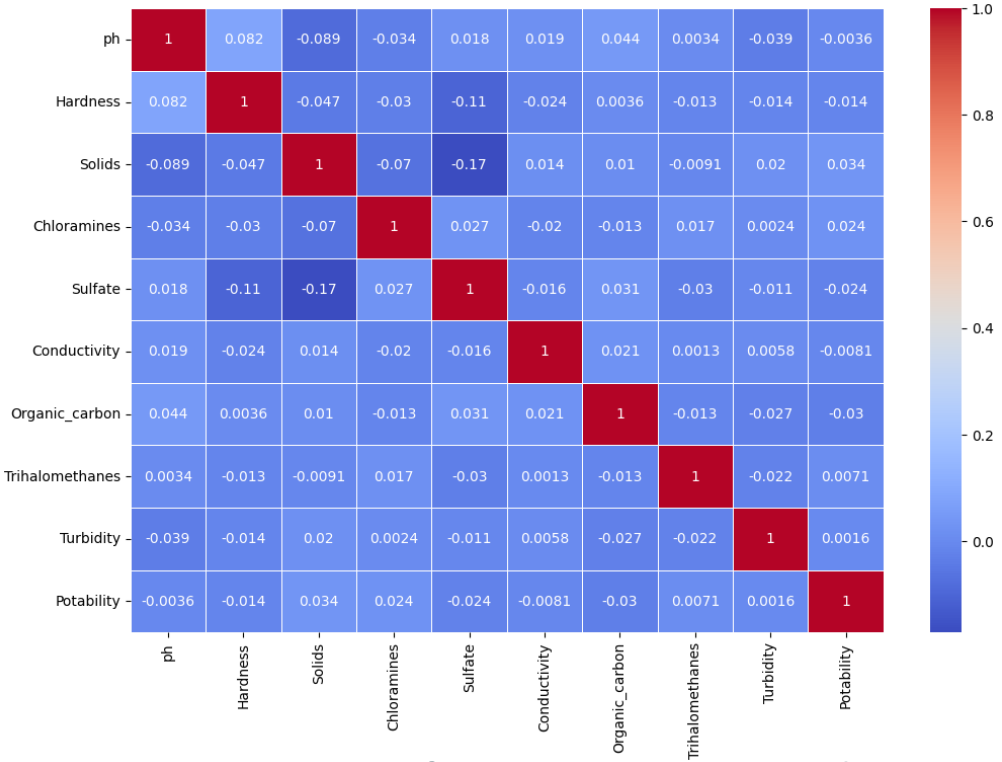
Data Analysis Results



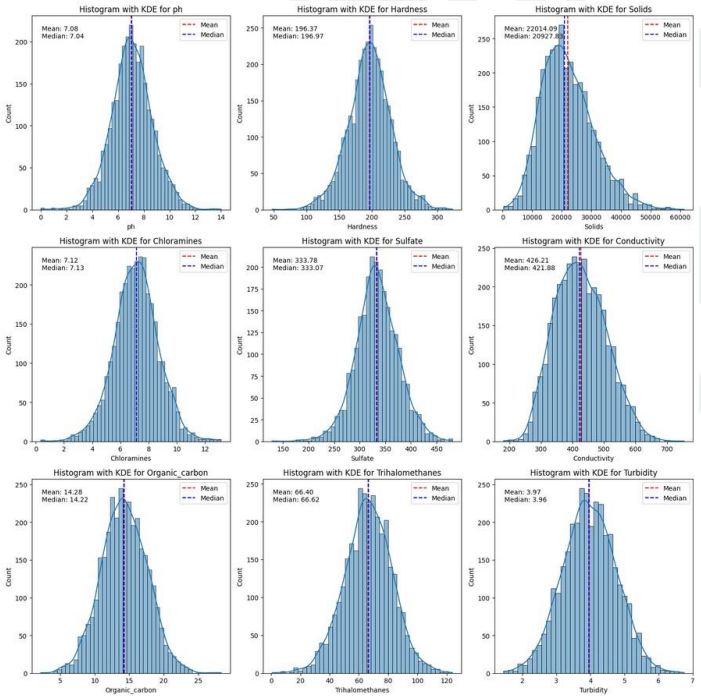


Water quality distribution Analysis

Data Analysis Results

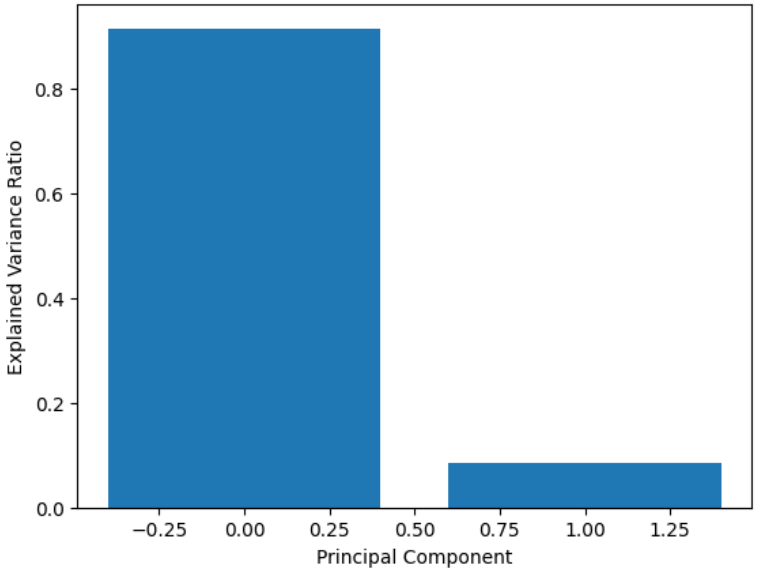


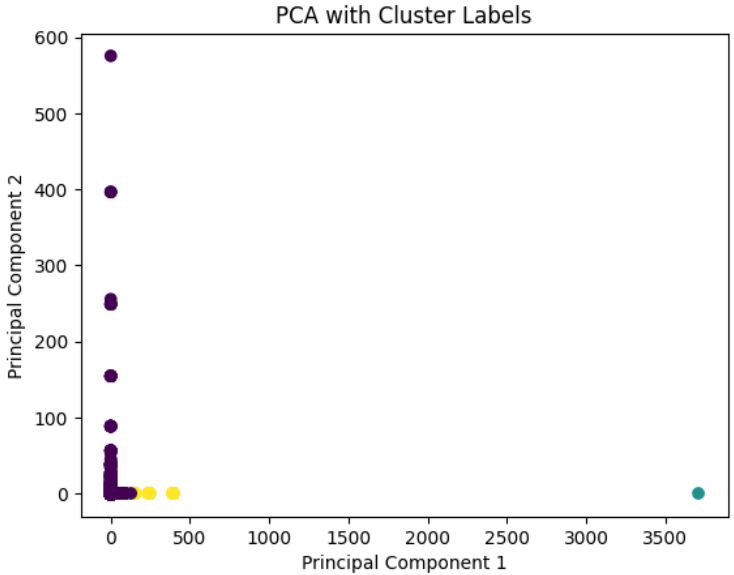
Heat Maps for Water Quality



KDE for Water Quality

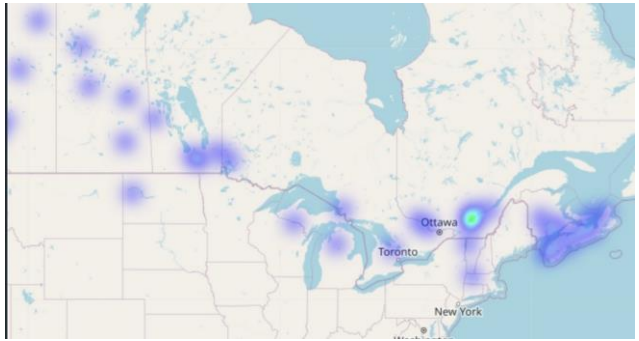
Data Analysis Results



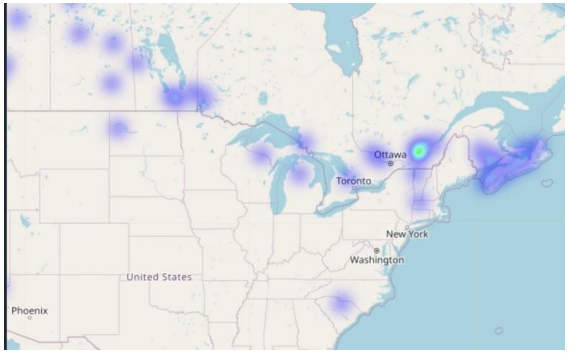


PCA Analysis

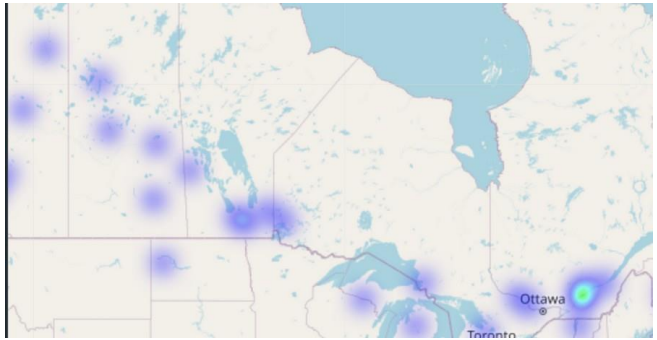
Heat Maps For Canada



PH



Temperature

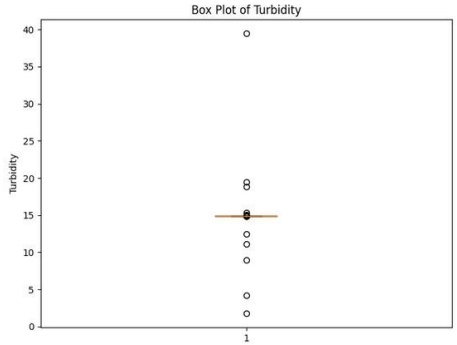


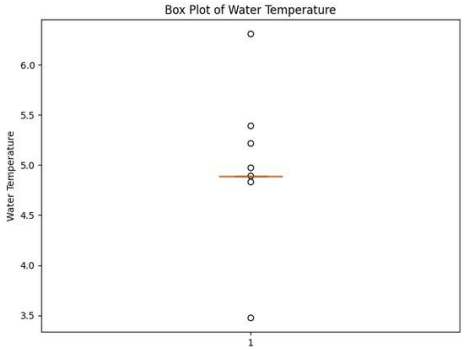
Turbidity

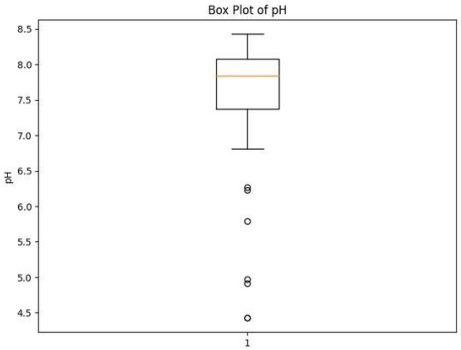
Similar Heat Maps Generated for Australia as well for 20 different Stations

Data Analysis Results

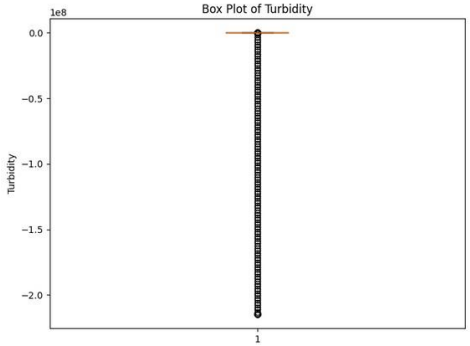
Box Plots

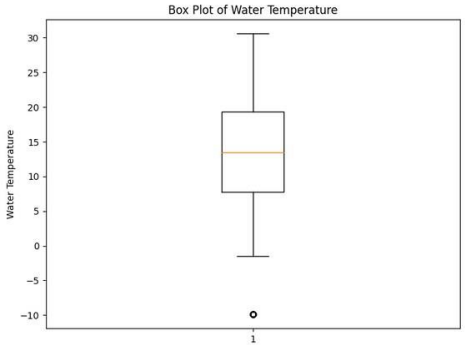


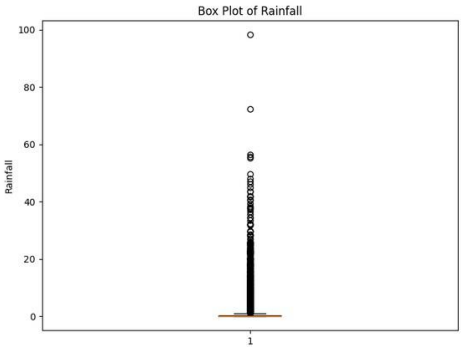




For Canada



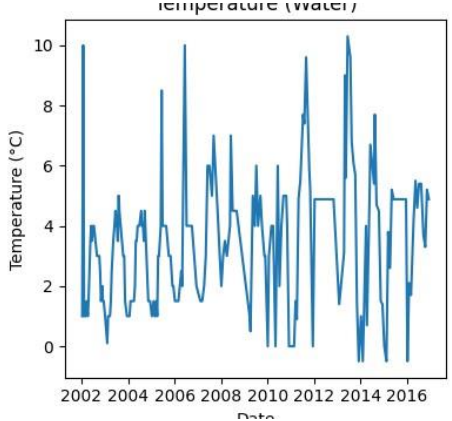


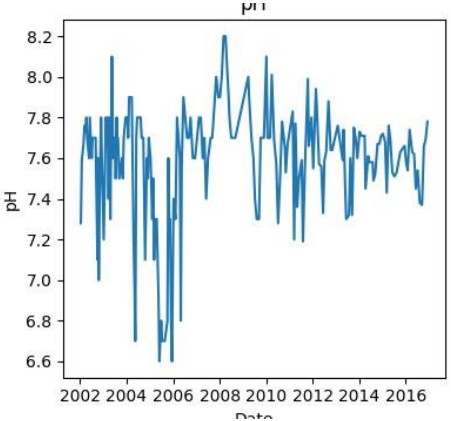


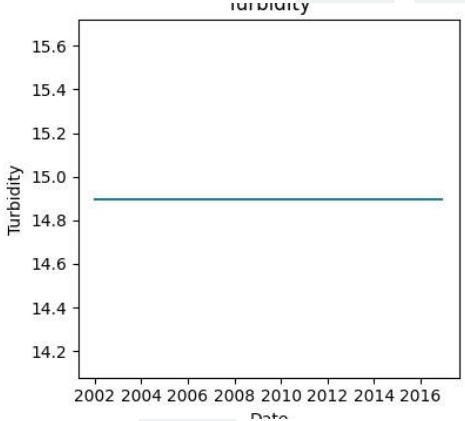
For Australia

Data Analysis Results

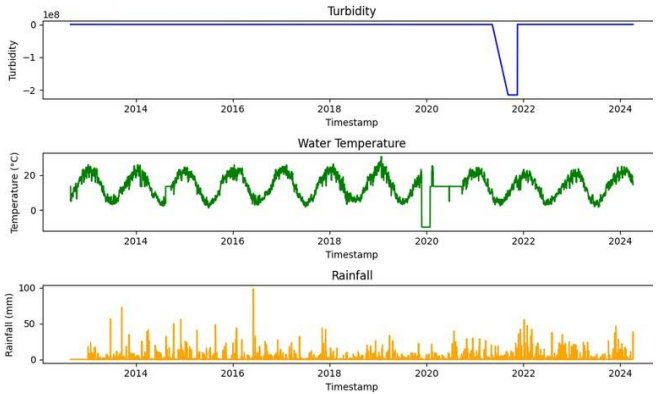
Time Series Analysis







For Canada



For Australia

Model Training and Validation

- Implemented machine learning models to predict water quality patterns based on historical data.
- Validated the models using appropriate metrics and assess their accuracy in identifying anomalies or deviations from normal water quality conditions.

Logistic Regression Model:

- Utilized a multinomial logistic regression model to predict water quality classes.
- Split the data into training and testing sets, and evaluated model performance using accuracy metrics.

```
Accuracy on known data: 0.9943743470224222
Classification Report on known data:
              precision    recall  f1-score   support

           0       0.99      0.99      0.99      4135
           1       0.99      0.99      0.99      4162
           2       1.00      1.00      1.00      4146

    accuracy                           0.99     12443
   macro avg       0.99      0.99      0.99     12443
weighted avg       0.99      0.99      0.99     12443

Predicted Quality for Unknown values saved in predicted_unknown.csv file.
```

LSTM Model:

- Organized the dataset for LSTM training, sorting it by date and scaling target values.
- Trained LSTM models for each target variable (Turbidity, Water Temperature, Rainfall) using sequential data.
- Split the dataset into training and testing sets, and assessed model performance using Mean Absolute Error (MAE).

```
Scaled input: [[7.27142917]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 86ms/step
Scaled prediction: [[0.567124]]
Inverse scaled prediction: [[-92959080.]]
Scaled input: [[33253911.35555555]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 29ms/step
Scaled prediction: [[0.58284086]]
Inverse scaled prediction: [[13.705055]]
Scaled input: [[13714698.57433809]]
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 18ms/step
Scaled prediction: [[0.02354201]]
Inverse scaled prediction: [[2.3118258]]
Predicted values: {'Turbidity': -92959080.0, 'WaterTemperature': 13.705055, 'Rainfall': 2.3118258}

Mean Absolute Error: 0.00033457
Turbidity: 15103490.890341576
WaterTemperature: 5.433072490939263
Rainfall: 3.1471335156927123
```

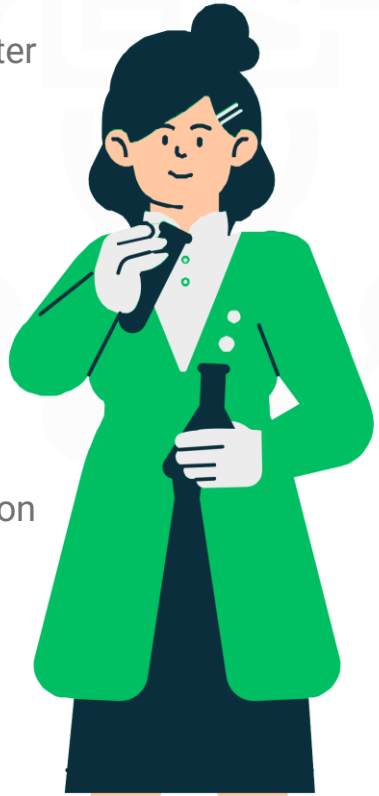
Conclusion

Insights from Analysis:

- Identified strong predictive capabilities of logistic regression model for water quality classification.
- LSTM models demonstrated effectiveness in forecasting environmental variables with low prediction errors.

Impact and Applications:

- Reliable water quality prediction models can inform decision-making for resource management and public health initiatives.
- Potential applications include early warning systems for water contamination events and optimization of water treatment processes.



Challenges faced

Data Quality and Availability:

- Ensuring the reliability and completeness of collected data from diverse sources.
- Addressing gaps and inconsistencies in the dataset for accurate analysis.

Model Generalization:

- Ensuring the developed models generalize well across different geographical regions and environmental conditions.
- Accounting for variations in data distributions and patterns that may affect model performance.

Feature Engineering and Selection

- Identifying relevant features and optimizing model inputs to enhance predictive accuracy.
- Balancing complexity and interpretability of models for practical deployment.

Future Work

- Explore advanced machine learning techniques to further enhance model accuracy and robustness.
- Integrate additional environmental factors and spatial data for comprehensive water quality assessment.

References

https://www.sciencedirect.com/science/article/pii/S0022169423015329

https://zenodo.org/records/7558906.

https://doi.pangaea.de/10.1594/PANGAEA.902360

https://gemstat.org

Data Analytics Internship Program 2024

Impacts of droughts and heatwaves on river water quality worldwide

THANK You