Analysis of Water Quality in the Puget Sound Region

~ Saloni Jain

Water Quality Analysis in the Puget Sound Region

Overview:

- *Introduction to Water Quality:* Understand the basics and importance.
- *Key Water Parameters:* Learn what measures define water quality.
- *Data Overview:* Overview of the dataset we used.
- *Data Cleaning:* How we prepared the data for analysis.
- *Exploratory Data Analysis:* Key findings from our initial data exploration.
- *Predictive Model:* Understanding our approach to forecasting water quality.
- *Conclusions:* What we've learned and next steps.

Introduction to Water Quality (Background information)

- ***Definition:*** Water quality refers to the chemical, physical, biological, and radiological characteristics of water.
- ***Importance:*** Essential for public health, ecosystem health, and economic well-being.
- ***Challenges:*** Pollution from industrial, agricultural, and domestic sources.

Business Question

How can we effectively identify, monitor, and improve the critical areas impacting water quality in the Puget Sound region to ensure sustainable environmental health and public safety?

Key Water Quality Parameters

Parameters Overview:

pH:

- Indicates water's acidity or alkalinity.
- Essential for maintaining marine life balance.
- In our dataset: Varied pH levels across regions indicate areas possibly affected by industrial waste or natural runoff.

Dissolved Oxygen (DO):

- Vital for aquatic organisms' survival.
- Low DO levels can lead to unhealthy or dead aquatic zones.
- In our dataset: Fluctuations in DO reflect areas of concern, particularly in enclosed or densely populated areas.

Turbidity:

- Measures water clarity and sediment presence.
- High turbidity can block sunlight, affecting plant and animal life.
- In our dataset: Varied levels suggest areas of erosion, runoff, or pollution.

Contaminants:

- Includes chemicals, metals, and biological toxins.
- Impacts on health of ecosystem and water safety.

These parameters are vital indicators of water quality health in the Puget Sound.

About the Puget Sound Water Quality Dataset

Dataset Origin:

- Source: Extracted from King County's public data repository.(King County. (2024). Water quality. Data.gov. https://catalog.data.gov/dataset/water-quality)
- Scope: Focuses on diverse water quality metrics across the Puget Sound.
- Objective: Assess water quality conditions at various locations and times.

Dataset Composition:

- Records: Over 1.8 million entries.
- Variables: 25 different variables, including 'Depth', 'Area', 'Parameter', and 'Value'.
- Types of Data:
  - Numerical: Depth (m), Value, MDL, RDL, etc.
  - Categorical: Site Type, Area, Method, Data Source, etc.

Data Structure (Sample Entries):

- Identifiers: Sample ID, Grab ID, Profile ID, Sample Number.
- Time-Stamped: Collect DateTime.
- Measurements: Depth (m), Value (various parameters like temperature, pH).
- Site Information: Site Type, Area, Locator, Site.
- Analysis Details: MDL, RDL, Method, Date Analyzed.

Data Cleaning and Preparation

- **Challenges Encountered:** Significant missing values and irrelevant information across several columns such as 'Grab ID', 'Depth (m)', and 'Text Value'.
- **Cleaning Actions:** Eliminated columns with high percentages of missing data or irrelevance. Applied strategies such as median imputation for numerical gaps and mode imputation for categorical discrepancies.
- **Outcome:** Reduced dataset complexity to 13 focused attributes, enhancing clarity and analysis efficiency.

Pre-Cleanup Missing Data Overview

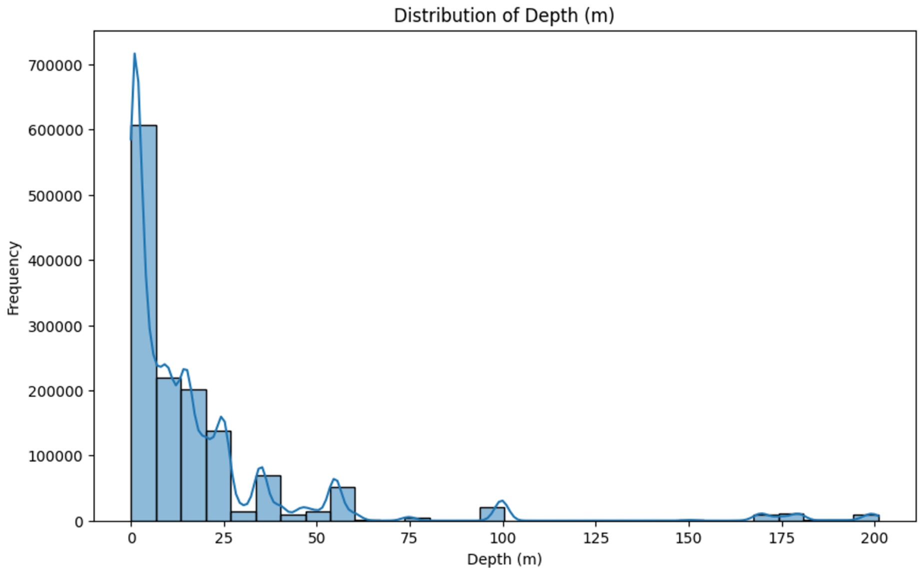
Data Aspect	Missing Values (%)
Sampling Details	
Grab ID, Depth	24.18
Measurement Info	
Value	8.75
Area, Units	Up to 0.07
Quality & Lab Checks	
Lab Qualifier, Text Value	~86.44
MDL, RDL	~46.05
Documentation & Records	
Sample Info, Steward Note	~99.43
Replicates, Replicate Of	~99.80
Analysis Details	
Method, Date Analyzed	Up to 41.19
Basic Information	
Sample ID, Site Type, etc.	0.00

Post-Cleanup Data Variables Overview

Retained Variables	Missing Values (%)
Profile ID	0.0
Depth (m)	0.0
Site Type	0.0
Area	0.0
Locator	0.0
Site	0.0
Parameter	0.0
Value	0.0
Units	0.0
QualityId	0.0
MDL	0.0
RDL	0.0
Method	0.0

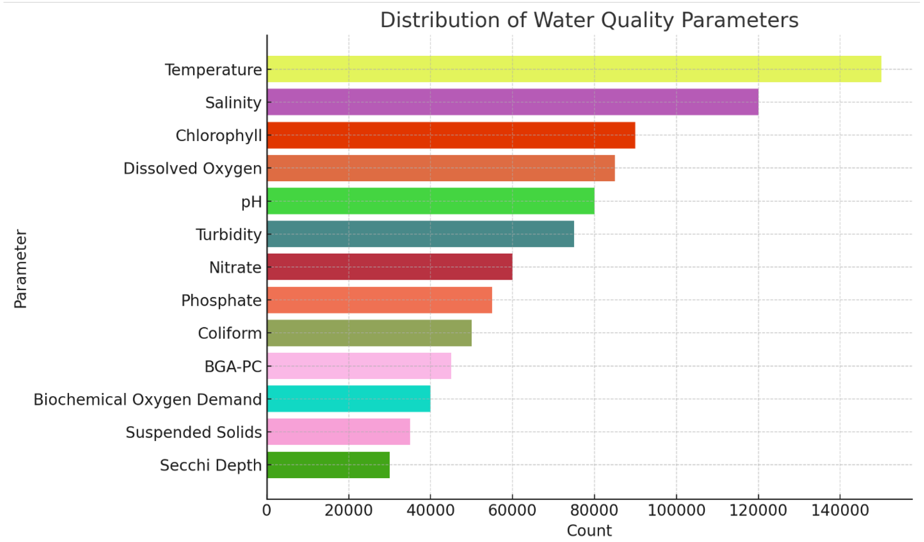
Exploratory Data Analysis (EDA)

The "**Depth" variable** indicates the water sampling depth, showing a bias towards shallower areas which might overlook deeper water conditions.



Distribution of Water Quality Parameters:

- **Test Distribution**: This chart shows how often different water tests, like for temperature or salinity, are done.
- **Focus Areas**: There's a big focus on checking temperature and salinity to understand how weather and sea water mixing affect the region.
- **Aquatic Health**: Lots of oxygen tests are done because oxygen is crucial for fish and other sea life.
- **Purposeful Testing**: The variety in test frequencies shows that monitoring is specifically designed to meet Puget Sound's unique environmental needs and rules.



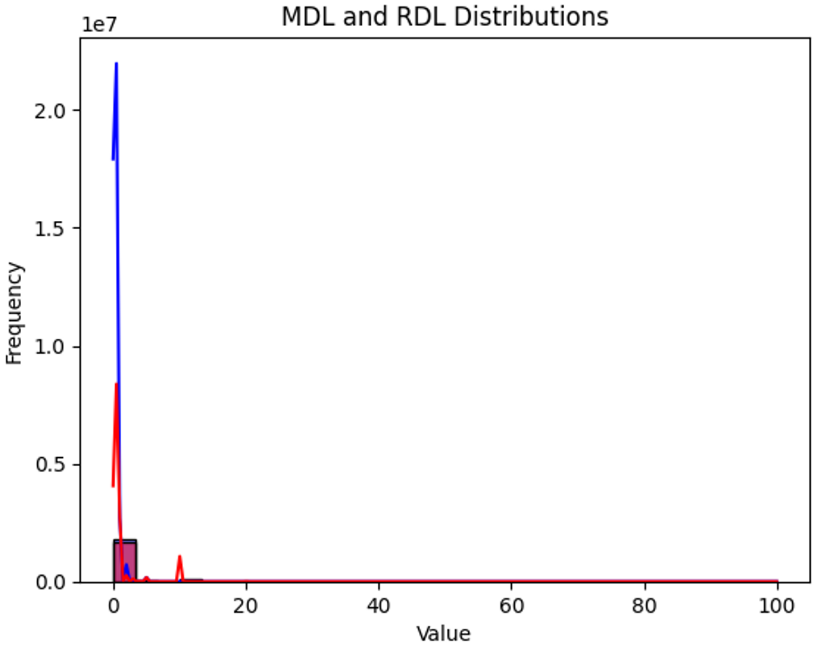
Variability in Detection Limits for Water Quality Testing

Detection Limits Simplified:

- MDL (Method Detection Limit) and RDL (Reporting Detection Limit) measure how well our tests can find small amounts of pollutants in water.
- A low MDL means the test is really good at finding even tiny amounts of pollution, which is important for catching harmful substances early.
- A higher RDL shows that some pollutants can't be detected or reported unless they're present in larger amounts.

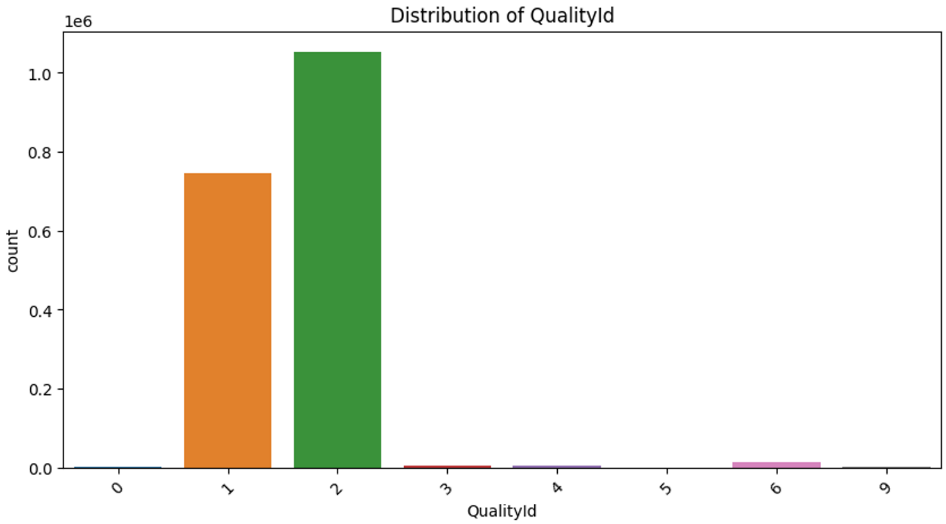
Why This Matters:

- These differences in detection show us the strengths and limits of our current water tests.
- They point out the importance of updating and improving our testing methods to monitor water quality more effectively.



Water Quality Categorization (Distribution of QualityId):

- **Main Findings**: Most water tests fall into two common 'QualityId' categories, showing that these are the usual water conditions in the area.
- **Stability vs. Extremes**: The lack of extreme values (very poor or excellent) might mean the water is generally stable, but we're not seeing the full picture.
- **Action Needed**: We should do more tests in areas where we don't have much data, especially to catch and understand rare but serious problems.

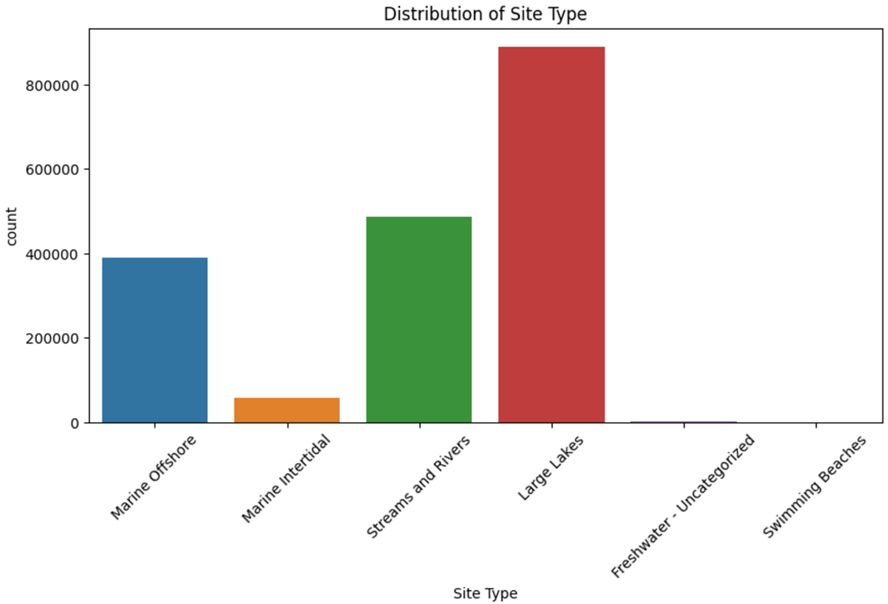


Sampling Site Type Distribution:

Main Sampling Locations: Concentrated in Large Lakes, Streams & Rivers, and Marine Offshore areas.

Why These Areas: Indicates targeted monitoring due to potential for higher pollution or significant ecological value.

Areas with Less Data: Fewer samples from Marine Intertidal and Swimming Beaches suggest possible gaps in data or lower perceived risks, which could miss critical environmental or health issues.



Water Quality Prediction Using Random Forest

Why Random Forest?

- Adapts well to complex data, capturing diverse environmental impacts on water quality.
- Handles imbalanced datasets, crucial given our QualityId skew towards classes 1 & 2.
- Offers insights into feature importance, aiding targeted environmental strategies.

Alternatives Considered:

- Linear models, less capable in handling non-linear relationships.Neural networks, resource-intensive and risk of overfitting.
- Single decision trees, prone to overfitting, less robust than Random Forest.

Performance Insights:

- High Overall Accuracy: Achieved **~98.62%,** indicating strong predictive capability.
- Varied Class Performance: Excellent for majority classes (1 & 2); challenges in minority classes (0, 3, 4, 5) reflect data imbalance.

Class	Precision	Recall	F1-Score	Support
0	0.79	0.40	0.53	390
1	0.98	1.00	0.99	223,545
2	0.99	0.99	0.99	315,669
3	0.60	0.13	0.21	1,684
4	0.94	0.47	0.63	1,035
5	0.00	0.00	0.00	6
6	0.81	0.55	0.66	3,906
9	0.88	0.55	0.68	806
Total/Average	**0.98**	**0.99**	**0.98**	**547,041**

Conclusions-Insights & Actions

Key Conclusions:

- Geographic & Seasonal Effects: Urban areas and changing seasons majorly influence water quality.
- Data Imbalance: Need to better predict lesser-seen water quality scenarios.
- Comprehensive Approach: Must consider all environmental, social, and economic factors for full understanding.

Policy & Management Implications:

- Focused Cleanup: Prioritize efforts in identified pollution-prone areas.
- Adaptive Monitoring: Adjust monitoring strategies with seasonal water quality changes.
- Health & Safety: Maintain water standards to protect public health, especially in recreational zones.

Looking Ahead:

- Broaden Data Collection: Target less studied areas for improved insights and model precision.
- Improve Predictions: Investigate new modeling approaches to handle rare conditions better.
- Cross-Sector Collaboration: Unite various stakeholders for a unified water quality strategy.

Thank You