Water Quality Analysis in Local Water Bodies Across the U.S.

Introduction

Water quality serves as a vital indicator of environmental health, directly influencing ecosystems, public health, and economic sustainability. This project investigates the factors affecting water quality in U.S. water bodies, focusing on parameters like dissolved oxygen, pH, and temperature. By analyzing seasonal trends and geographic variations using automated data pipelines, the project aims to provide actionable insights for environmental monitoring and conservation.

Main Question

What are the key factors influencing water quality in local water bodies, and how do they vary across different locations and seasons?

Data Sources

1. National Aquatic Resource Surveys (NLA)

- **URL:** NLA Metadata
- **Download Link:** Dataset in CSV Format
- **Reason for Choosing:** Comprehensive indicators of aquatic health, including dissolved oxygen and population estimates, essential for studying seasonal and spatial trends.
- **Structure and Quality:** Tabular data with consistent formatting; minimal missing values in non-critical fields.
- **License and Permissions:**
  Licensed under the CC0 1.0 Universal Public Domain Dedication, allowing unrestricted use, modification, and distribution.
  Obligation: Proper attribution will credit the EPA in outputs.

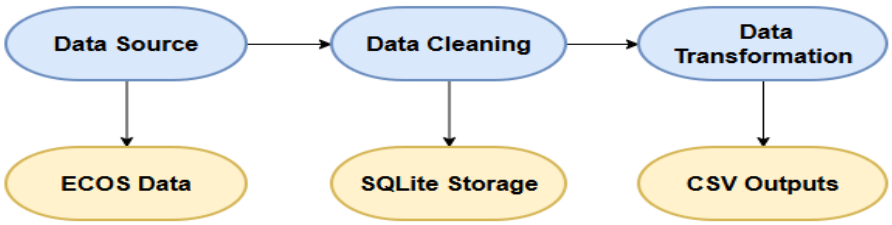
2. U.S. Government's Open Data Portal (ECOS)

- **URL:** Water Quality Dataset on Data.gov
- **Download Link:** Dataset in CSV Format
- **Reason for Choosing:** Fine-grained measurements (e.g., pH, salinity, temperature) and widespread geographic coverage to complement the NLA dataset.
- **Structure and Quality:** Structured tabular data, requiring minor cleaning for standardization.
- **License and Permissions:**
  Licensed under CC0 1.0 Universal Public Domain Dedication, permitting unrestricted use.
  Obligation: Acknowledgment practices will ensure proper credit to the source.

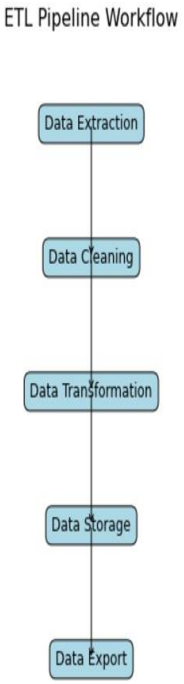
Methodology

ETL Pipeline Design: The project uses an automated ETL (Extract, Transform, Load) pipeline implemented in Python to efficiently process data.

Basic Workflow Diagram:



Pipeline Workflow



1. **Extraction:**
   Data is fetched using `requests` and saved locally. File integrity is validated before proceeding. (Tools: `requests`, Python scripts)

2. **Cleaning:**

   Columns are renamed for consistency, missing values are imputed, and irrelevant fields are removed. (Tools: `Pandas`, Python scripts)

3. **Transformation:**
   Seasonal aggregation and new metric derivations (e.g., water quality indices) are applied. (Tools: `Pandas`, Python scripts)

4. **Data storage:**
   Cleaned and transformed data is stored in an SQLite database using the `to_sql` function in the `Pandas` library. The database organizes data into tables (e.g., `nla_condition` and `ecos_data`) for efficient querying. (Tools: `SQLite` library)

5. **Data Export:**
   Processed data is stored in SQLite for querying and exported as CSV for further analysis. (Tools: `SQLite`, `Pandas`, Python scripts)

Problems Encountered and Solved

1. **Inconsistent Column Names:**
   Solution: Implemented a standardization process during the cleaning phase to unify column names by stripping whitespace and renaming columns to a consistent format.

2. **Missing Values:**
   Solution: Applied median substitution for numerical fields to impute missing values, ensuring minimal impact on data integrity. For non-critical fields, default placeholders were used.

3. **Data Overlap and Redundancy:**
   Solution: Developed a schema matching process to accurately merge datasets based on unique spatial and temporal identifiers, eliminating duplicates and ensuring data integrity.

4. **Dynamic File Downloads:**
   Solution: Utilized regular expressions within the Python scripts to dynamically identify and process the correct files, ensuring seamless data extraction regardless of filename changes.

5. **Storage Errors:**
   Solution: Implemented robust error handling and data type validation before loading data into SQLite, ensuring compatibility and preventing insertion errors.

Results and Limitations

Output Data	**Limitations**
• **Structure:** Final dataset includes key indicators (e.g., pH, salinity, temperature) across locations and seasons. • **Format:** Available in CSV for visualization and SQLite for querying. • **Quality Assurance:** Enhanced consistency and accuracy through rigorous validation.	• **Temporal Gaps:** Irregular dataset updates may limit temporal trends analysis. • **Geographic Bias:** Overrepresentation of urban or heavily monitored regions. • **Measurement Standards:** Methodological differences across datasets may affect comparability.

Future Improvements

- Real-time data integration for enhanced temporal accuracy.
- Expansion of geographic scope to include underrepresented areas.
- Advanced visualizations, such as interactive dashboards, to communicate insights effectively

Reflection	
Strengths	**Challenges**
• Fully automated pipeline ensures reproducibility. • Complementary datasets enhance comprehensiveness of analysis. • Rigorous cleaning and validation improve data quality.	• Future enhancements should address temporal and geographic limitations. • Real-time monitoring integration could improve analytical robustness.

Conclusion

The implementation of an automated ETL pipeline enables comprehensive analysis of water quality data. By integrating and processing datasets, this project provides valuable insights into environmental health across U.S. water bodies. Future developments will focus on addressing identified limitations and expanding the scope of the analysis.