**README: Alaska Lake Water Quality Data Processing**

**Overview**

This dataset contains scripts and final datasets for processing water quality data for lakes in Alaska. The work integrates data from multiple sources, links it to geospatial references, and summarizes it at both site and lake levels.

**Included Files**

**Scripts**

1. **cleanData\_WQP.py**:
	* Cleans water quality data from the Water Quality Portal (WQP).
	* Input: WQP\_resultphyschem\_\*\_forProcessing.csv (user-supplied).
	* Output: WQP\_Data\_Cleaned.csv (intermediate).
2. **appendCleanedData.py**:
	* Compiles cleaned data from multiple sources and removes outliers.
	* Input: Cleaned datasets from various sources.
	* Output: appendedData\_Cleaned.csv (intermediate).
3. **createLakePoints.py**:
	* Creates geospatial reference points for lakes using public datasets.
	* Input: TIGER/Line Shapefile, National Hydrography Dataset (NHD), Geographic Names Information System (GNIS)
	* Output: lakePoints.gpkg (intermediate).
4. **processGeospatialSiteData.py**:
	* Links site data to lake references using geospatial matching.
	* Input: lakePoints.gpkg, WQP station data, and non-WQP site data (user-supplied).
	* Outputs:
		+ sitesProcessed.csv (tabular format).
		+ sitesProcessed.gpkg (geospatial format).
5. **summarizeWaterQualityData.py**:
	* Summarizes water quality data by site and lake.
	* Inputs:
		+ appendedData\_Cleaned.csv (intermediate).
		+ sitesProcessed.gpkg (processed site data).
		+ lakePoints.gpkg (geospatial lake references).
	* Outputs:
		+ waterQualitySamples.csv (sample-level data).
		+ waterQualityBySiteDate\_long.csv (site-level summary in long format).
		+ waterQualityByLake\_wide.csv (lake-level summary in wide format).
		+ waterQualityByLake\_wide.gpkg (geospatial version of the lake-level summary).

**Workflow**

**1. Data Cleaning**

* **Script**: cleanData\_WQP.py
* Cleans water quality data from WQP, standardizes units, and filters relevant characteristics.
* **Script**: appendCleanedData.py
* Combines cleaned datasets into a unified format and removes outliers for geospatial processing.

**2. Geospatial Processing**

* **Script**: createLakePoints.py
* Generates geospatial reference points for Alaska lakes from GIS datasets.
* **Script**: processGeospatialSiteData.py
* Links water quality sampling sites to lakes using tiered geospatial matching (by name, intersection, proximity).

**3. Data Summarization**

* **Script**: summarizeWaterQualityData.py
* Summarizes water quality data at site and lake levels.
* Outputs final datasets in tabular and geospatial formats.

**Key Outputs**

1. **Sample-Level Data**:
	* waterQualitySamples.csv.
2. **Site-Level Data**:
	* waterQualityBySiteDate\_long.csv (long format).
	* sitesProcessed.csv (processed site data).
3. **Lake-Level Data**:
	* waterQualityByLake\_wide.csv (tabular).
	* waterQualityByLake\_wide.gpkg (geospatial).

**Data Sources and References**

This analysis uses publicly available geospatial and water quality datasets. Key data sources include:

1. **Water Quality Portal (WQP)**
	* **Description**: Provides water quality monitoring data collected by multiple agencies and organizations in the United States.
	* **Reference**: Water Quality Portal. (Accessed 2024). Water Quality Data for the Nation. U.S. Geological Survey, U.S. Environmental Protection Agency, and National Water Quality Monitoring Council. Available at: <https://www.waterqualitydata.us/>.
2. **TIGER/Line Shapefiles**
	* **Description**: Spatial extracts from the Census Bureau's database, including boundaries and geographic features.
	* **Reference**: U.S. Census Bureau, Department of Commerce (2023). TIGER/Line Shapefile, Current, Nation, U.S., State and Equivalent Entities. <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html>
3. **Geographic Names Information System (GNIS)**
	* **Description**: Database of federally recognized geographic names, including lake names and GNIS IDs.
	* **Reference**: U.S. Geological Survey (2024). Geographic Names Information System (GNIS) Full Model National (published 20240201) GeoPackage. <https://apps.nationalmap.gov/downloader/>
4. **National Hydrography Dataset (NHD)**
	* **Description**: Geospatial dataset of surface water features in the United States, including lakes, rivers, and reservoirs.
	* **Reference**: U.S. Geological Survey (2019). National Hydrography Dataset (ver. NHD 20191017 for Alaska State or Territory FileGDB 10.1 Model Version 2.2.1). <https://www.usgs.gov/national-hydrography>

**Usage Instructions**

**Prerequisites**

* **Software**: Python 3.8+ with the following libraries:
	+ pandas, geopandas, shapely, numpy
* **Data Sources** (user-supplied):
	+ WQP data: <https://www.waterqualitydata.us/>
	+ TIGER/Line Shapefiles: <https://www.census.gov/geographies/mapping-files/time-series/geo/tiger-line-file.html>
	+ GNIS Gazetteer: <https://apps.nationalmap.gov/downloader/>
	+ NHD: <https://www.usgs.gov/national-hydrography>

**Running the Scripts**

1. **Prepare Input Files**:
	* Download and format raw data as specified in the scripts.
	* Place files in appropriate directories or update file paths in the scripts.
2. **Execute Scripts**:
	* Run scripts sequentially in the workflow order.
	* Customize file paths to match your system’s directory structure.

**Contact**

For questions or issues, contact:

* **Name**: Dana Brown
* **Email**: dana.brown@alaska.edu
* **Institution**: University of Alaska Fairbanks