

Ryan Pham

May 2, 2025

DS 210

Professor Kontothanassis

Final Project Report

For my final project, I took a water pollution and disease dataset from Kaggle(link provided below), and the purpose of this was to explore the top 5 most contaminated sources of water. Drinking water, as well as sources used for other necessities or leisure, can be polluted and full of disease, which is why I wanted to explore which ones could be avoided depending on the situation. Furthermore, I intended to print out the top 5 countries where their contamination level is very high, as well as explore what is found in their water, such as nitrate, dissolved oxygen, and lead concentration. First on the Cargo.toml I made sure to use `csv = "1.3"` and the `serde` versions based on the digits we have used in previous homework assignments. Using the modules that I provided, I used the `processor` to open the CSV file using `read_data`, and although there wasn't any cleaning that was needed for this code, I have used `serde` to use `deserialize` to sort the data into a vector. Along with the `processor`, the main goal was to print out the final results as well as the model to reconstruct and compute the top countries with their contaminant level and their water concentration. The structs that were included were `WaterData` and `CountryScore`, which format the data as a float type and with the attributes for the top 5 countries that were listed in the output. Apart from the function that was reading the CSV file, there was the `topsources` function that used a `HashMap` to create a vector for the body of water and their contamination level using a for loop and adding. It would also take the average of each body of water to present the top 5 sources. Top countries would do a similar approach, where it will go

through each country in the csv file and records it into the CountryScore struct, and it would introduce the pH, turbidity, dissolved oxygen, nitrate, and lead concentration starting out at 0 and would use a for loop to add up the score. In the end it will return the score using the vector of CountryScore.

Tests:

```
(app-root) [/opt/app-root/src/.cache/finalproject2/finalproject3]
● $ cargo test
  Compiling finalproject2 v0.1.0 (/opt/app-root/src/.cache/finalproject2/finalproject3)
  Finished `test` profile [unoptimized + debuginfo] target(s) in 0.61s
  Running unittests src/main.rs (target/debug/deps/finalproject2-78e64764928180ed)

running 0 tests

test result: ok. 0 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.00s

(app-root) [/opt/app-root/src/.cache/finalproject2/finalproject3]
● $ cargo test
  Finished `test` profile [unoptimized + debuginfo] target(s) in 0.04s
  Running unittests src/main.rs (target/debug/deps/finalproject2-78e64764928180ed)

running 0 tests

test result: ok. 0 passed; 0 failed; 0 ignored; 0 measured; 0 filtered out; finished in 0.00s

(app-root) [/opt/app-root/src/.cache/finalproject2/finalproject3]
```

The purpose of this test would be to see if the contamination score would pass using random numbers we put for the pH, turbidity, dissolved oxygen, nitrate, and lead in a separate test module. I put that it would equal to 16 as the total once you subtract dissolved oxygen as this helps make water quality better.

Run/Usage Instructions:

For running the code it could be simply with cargo run and it usually will print not that long after that statement around a second or less.

```
Top 5 Most Contaminated Water Sources (by type):
```

1. Tap – Average Pollution Score: 27.42
2. River – Average Pollution Score: 26.59
3. Spring – Average Pollution Score: 25.83
4. Well – Average Pollution Score: 25.82
5. Lake – Average Pollution Score: 25.19

```
Top 5 Most Contaminated Countries (avg. metrics):
```

1. Bangladesh | pH: 7.29, Turbidity: 2.54, D0: 6.53, Nitrate: 26.21, Lead: 5.18
2. Nigeria | pH: 7.25, Turbidity: 2.41, D0: 6.21, Nitrate: 25.84, Lead: 5.14
3. Mexico | pH: 7.29, Turbidity: 2.49, D0: 6.42, Nitrate: 25.53, Lead: 5.17
4. India | pH: 7.25, Turbidity: 2.42, D0: 6.47, Nitrate: 25.42, Lead: 5.14
5. USA | pH: 7.26, Turbidity: 2.52, D0: 6.65, Nitrate: 25.57, Lead: 4.99

```
(app-root) [/opt/app-root/src/.cache/finalproject2/finalproject3]
```

```
$
```

Most of my code has been edited through ChatGPT, which was to help organize the top 5 most contaminated bodies of water and the top 5 countries with their water quality score, mainly within the processor module.