Water Data Preparation for CDC Data Call Spring 2025

# Start

This write-up documents the processing steps in preparing the summary water data for submission to the spring 2025 Data Call. This follows the How-to-Guide and Data Dictionary 2025. To get an understanding of names and/or rules please refer to those documents. The process was started with the unaggregated file(.xlsx) received from NDEE on 3/31/2025. That contained 12,472 rows of data. Both sampling data and PWS Inventory were validated using the Python codes. 27 of the PWS inventories had validation errors of not containing values for Latitude, Longitude, and LocationDerivationCode (included them in the PWSInventory\_2025\_errors.csv file attached to the same email. Their related values were replaced for ‘Missing’ values following HTG Guide 2025.

['PWSIDNumber', 'Year', 'AnalyteName', 'AnalyteCode', 'ConcentrationUnits', 'Concentration', 'DateSampled', 'SamplePointID', 'DetectionLimit', 'DetectionLimitUom', 'NonDetectFlag']

It contains sampling results from 2023 to 2024 as below:

| Year | Rows of data |
| --- | --- |
| 2023 | 5796 |
| 2024 | 6676 |

* **Step 1-** Adjusting AnalyteCode All data for Analytedoe 1041 removed, since not required by data call (0 rows).
* **Step 2 -** Initial validation test An initial validation against Data Dictionary rules done on columns that will be present in the final summary file. all passed validation for the following columns: ‘RowIdentifier’, ‘PWSIDNumber’, ‘Year’, ‘AnalyteCode’, ‘ConcentrationUnits’, ‘Concentration’, ‘DateSampled’]

After these intial steps number of data points bsed on each AnalyteCode are as below:

| AnalyteCode | No. Rows |
| --- | --- |
| 1038 | 4496 |
| 1005 | 1259 |
| 2050 | 1193 |
| 2039 | 1193 |
| 2987 | 1006 |
| 2984 | 1006 |
| 2950 | 682 |
| 2456 | 682 |
| 4010 | 747 |
| 4006 | 208 |

* **Step 3 -** Checked if the ConcentrationUnits for each AnalyteCode applies with the values in the Data Dictionary

All the units used are correct. Note: There is no code nitrate (1040). However, the 2024 allows using code 1038, therefore we did not change it.

* **Step 4 -** Checked if uranium needs a change of ConcentrationUnits from pci/l to ug/l
* **Step 4 -** Checking rows with NonDetectFlag. Below is the result for checking rows with a NonDetectionFlag = 1. All of them have a half LDL value for their concentrations, which follows the HTG Guide.

## # A tibble: 1 × 2  
## is\_half\_LDL number  
## <dbl> <int>  
## 1 0.5 5236

Below is the percentage of non-detects of sampling for each analyte.

## [1] "Percent of no-detect flags for each analyte:"

## # A tibble: 10 × 3  
## AnalyteCode Total percent\_no\_detect  
## <dbl> <int> <dbl>  
## 1 1005 1259 18.9   
## 2 1038 4496 8.92  
## 3 2039 1193 99.4   
## 4 2050 1193 96.1   
## 5 2456 682 8.36  
## 6 2950 682 4.11  
## 7 2984 1006 99.7   
## 8 2987 1006 98.6   
## 9 4006 208 0   
## 10 4010 747 32.4

* **Step 5 -** Averaged duplicates into one values. Checked for duplicated of same analyte sampled in the same day from same point location, and averaged them into one value. It contained repeatitions for values in columns (PWSIDNumber, Year, AnalyteCode, DateSampled, SamplePointID). Added a new column “NumSamples” that can account for these number of samples. (The highest was 12 for several analyte sampling in the same point in the same date)

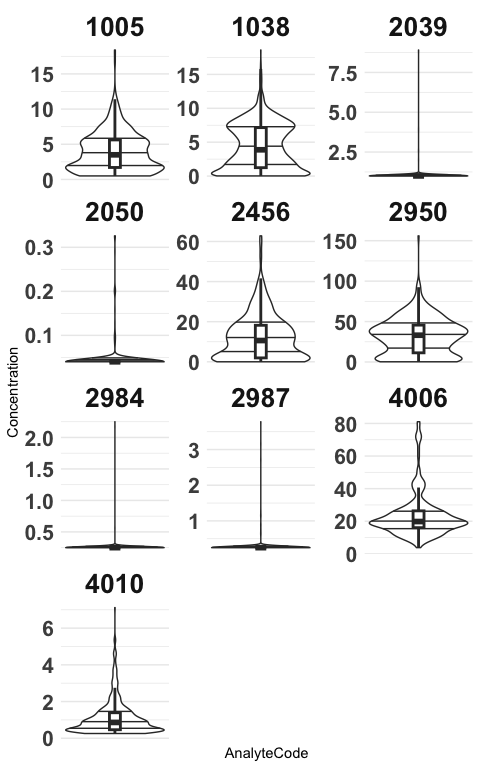
## [1] "initial rows in data: 12472"

## [1] "no. of rows after averaging same day/point/analyte: 11772"

before starting the summarization, we took a look into maximum concentration value for each analyte compared to its median over all dataset.For each maximum value, its DateSampled is also provided.

## # A tibble: 10 × 4  
## AnalyteCode maximum median DateSampled  
## <dbl> <dbl> <dbl> <date>   
## 1 1005 18.4 3.5 2023-03-20   
## 2 1038 18.6 3.88 2023-04-17   
## 3 2039 8.91 1 2023-04-12   
## 4 2050 0.326 0.04 2024-05-22   
## 5 2456 62.8 10.6 2023-03-06   
## 6 2950 156 33.1 2024-07-15   
## 7 2984 2.25 0.25 2023-05-24   
## 8 2987 3.78 0.25 2023-10-16   
## 9 4006 81.1 19.8 2024-02-20   
## 10 4010 7.12 0.862 2024-11-12

* Below, we have also provided plots for the distributions: Based on the diagrams we suggest a look into analytes 2039, 2050, 2984, and 2987 since they have very large outliers.



## Starting the Summarization

From this step we have followed the HTG guide for summarizing the data. We created three scenarios. 1- Annual averaging for 8 non-disinfecion By-products 2- Annual averaging for the two Disinfection By-products 3- Annual maximum for all 10 analytes 4- Quarterly values for Nitrate (1040, 1038) and Atrazine (2050) 5- Quarterly average for the two Disinfection By-products: TTHM(2950) and HAA5(2456)

* **Step S.1 -** Summarized annual means for the 8 anlytes following HTG 2025:

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2023 3918 5122 8 2324  
## 2 2024 4461 5986 8 2884

* **Step S.2 -** Summarized annual means for the two Disinfection By-products following HTG 2025:

..

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2023 386 674 2 16  
## 2 2024 404 690 2 12

* **Step S.3 -** Summarized annual maximums for all the analytes following HTG 2025:

…

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2023 4140 5796 10 2340  
## 2 2024 4687 6676 10 2896

* **Step S.4 -** Summarized quarterly means for all the analytes following HTG 2025: We first separatd the four analytes for which the quarterly values are required. Nitrate and Atrazine and disinfection-by-products (TTHM and HAA5). Then for each group, since the methods are different and similar to their annual averages, we calculated the average quarterly values

… \* **Step S.4.1 -** For Atrazine and Nitrate

…

## # A tibble: 8 × 5  
## SummaryTimePeriod NuLocations NuSamples Analytes Non\_detects  
## <chr> <int> <int> <int> <dbl>  
## 1 2023-1 761 777 2 160  
## 2 2023-2 641 654 2 165  
## 3 2023-3 612 633 2 138  
## 4 2023-4 588 596 2 131  
## 5 2024-1 733 743 2 180  
## 6 2024-2 810 837 2 296  
## 7 2024-3 751 789 2 300  
## 8 2024-4 650 660 2 177

* **Step S.4.1 -** For disinfection by-products

….

## # A tibble: 8 × 5  
## SummaryTimePeriod NuLocations NuSamples Analytes Non\_detects  
## <chr> <int> <int> <int> <dbl>  
## 1 2023-1 56 124 2 2  
## 2 2023-2 62 122 2 1  
## 3 2023-3 212 308 2 13  
## 4 2023-4 56 120 2 0  
## 5 2024-1 56 120 2 1  
## 6 2024-2 60 122 2 0  
## 7 2024-3 220 314 2 9  
## 8 2024-4 68 134 2 2

* **Step F -** Then the all resulted 5 tables were aggreagated into one table in conformance with HTG and Data Dictionary 2025

As the final step, we checked the data against the 8 steps in the Gateway 2022 (Appendix F of HTG)