Water Data Preparation for CDC Data Call Spring 2024

# Start

This write-up documents the processing steps in preparing the summary water data for submission to the spring 2024 Data Call. This follows the How-to-Guide and Data Dictionary 2024. 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 5/14/2024. That contained 11,761 rows of data. Both sampling data and PWS Inventory were validated using the Python codes. 28 of the PWS inventories had validation errors of not containing values for Latitude, Longitude, and LocationDerivationCode (included them in the PWSInventory\_2024\_errors.csv file attached to the same email. This is similar to last year, when two sets of inventory datasets were provided and the earliest file had all the data. I therefore used that earlier inventory file from last to ONLY replace the missed field values for the inventories.

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

It contains sampling results from 2022 to 2023 as below:

| Year | Rows of data |
| --- | --- |
| 2022 | 5968 |
| 2023 | 5793 |

* **Step 1-** Adjusting AnalyteCode All data for Analytedoe 1041 removed, since not required by data call (17 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 | 4340 |
| 1005 | 1175 |
| 2050 | 1110 |
| 2039 | 1110 |
| 2987 | 1016 |
| 2984 | 1016 |
| 2950 | 696 |
| 2456 | 694 |
| 4010 | 433 |
| 4006 | 167 |

* **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. Most of them have a half LDL value for their concentrations (refer to HTG Guide), and 74 have no value for Concentration LDL.

## # A tibble: 1 × 2  
## is\_half\_LDL number  
## <dbl> <int>  
## 1 0.5 4965

“2. All samples with results below the detection limit must have a non-zero and positive detection limit value provided. If a detection limit is not available from the source data, grantees are expected to estimate the detection limit from available data and/or provide a standard detection limit number. Guidance for determining detection limits when one is not provided is available in Appendix C.” We therefore, followed Appendix C to calculate the DL for zero DL values (6,792 which is about 58% of all data!)

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 1175 19.2   
## 2 1038 4340 8.92  
## 3 2039 1110 99.1   
## 4 2050 1110 93.7   
## 5 2456 694 9.94  
## 6 2950 696 5.75  
## 7 2984 1016 99.4   
## 8 2987 1016 99.0   
## 9 4006 167 0   
## 10 4010 433 36.3

* **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 28 for one analyte sampling in the same point in the same date)

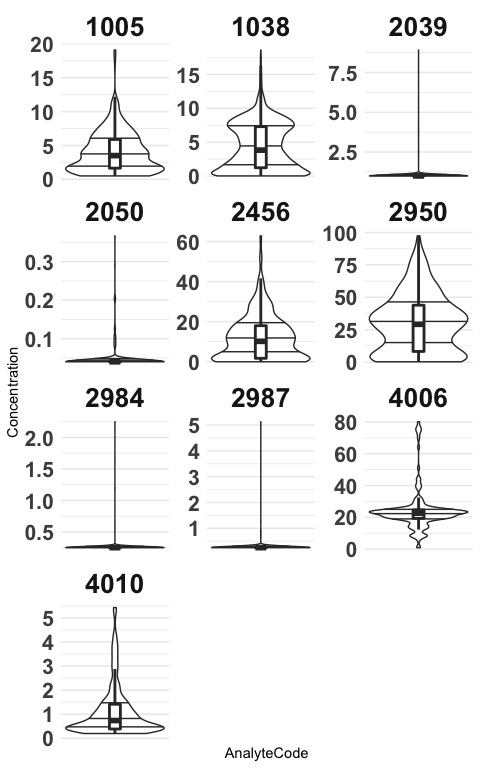
## [1] "initial rows in data: 11757"

## [1] "no. of rows after averaging same day/point/analyte: 11063"

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: 12 × 4  
## AnalyteCode maximum median DateSampled  
## <dbl> <dbl> <dbl> <date>   
## 1 1005 19.1 3.5 2022-06-27   
## 2 1005 19.1 3.5 2022-06-27   
## 3 1038 18.6 3.83 2023-04-17   
## 4 2039 8.91 1 2023-04-12   
## 5 2050 0.367 0.04 2022-01-19   
## 6 2050 0.367 0.04 2022-01-19   
## 7 2456 62.8 10.2 2023-03-06   
## 8 2950 97.4 29.2 2023-08-10   
## 9 2984 2.25 0.25 2023-05-24   
## 10 2987 5.13 0.25 2022-10-05   
## 11 4006 80.2 22.6 2022-08-01   
## 12 4010 5.43 0.736 2022-05-09

* 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 2023:

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2022 4075 5248 8 2601  
## 2 2023 3915 5119 8 2324

* **Step S.2 -** Summarized annual means for the two Disinfection By-products following HTG 2023:

..

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2022 432 716 2 24  
## 2 2023 386 674 2 16

* **Step S.3 -** Summarized annual maximums for all the analytes following HTG 2023:

…

## # A tibble: 2 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2022 4350 5964 10 2625  
## 2 2023 4137 5793 10 2340

* **Step S.4 -** Summarized quarterly means for all the analytes following HTG 2023: 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 2022-1 813 835 2 233  
## 2 2022-2 638 651 2 170  
## 3 2022-3 612 629 2 195  
## 4 2022-4 661 678 2 235  
## 5 2023-1 761 777 2 160  
## 6 2023-2 641 654 2 165  
## 7 2023-3 609 630 2 138  
## 8 2023-4 588 596 2 131

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

….

## # A tibble: 8 × 5  
## SummaryTimePeriod NuLocations NuSamples Analytes Non\_detects  
## <chr> <int> <int> <int> <dbl>  
## 1 2022-1 52 114 2 0  
## 2 2022-2 55 117 2 0  
## 3 2022-3 264 358 2 23  
## 4 2022-4 61 127 2 1  
## 5 2023-1 56 124 2 2  
## 6 2023-2 62 122 2 1  
## 7 2023-3 212 308 2 13  
## 8 2023-4 56 120 2 0

* **Step F -** Then the all resulted 5 tables were aggreagated into one table in conformance with HTG and Data Dictionary 2024

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