Water Date Preparation for CDC Data call Spring 2023

# Start

This is to document the processing steps in preparing the summary water data for submission to spring 2023 Data Call. This follows the How-to-Guide and Data Dictionary 2023. To get undrestanding about names and/or rules please refer to those documents. The process was started with the unaggregated file(.xlsx) received from NDEE on 4/4/2023. That contained 68,108 rows of data with the following columns:

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

It contains sampling results from 2012 to 2021 as below:

| Year | Rows of data |
| --- | --- |
| 2012 | 9069 |
| 2018 | 7491 |
| 2015 | 7140 |
| 2013 | 6970 |
| 2016 | 6460 |
| 2021 | 6444 |
| 2019 | 6396 |
| 2017 | 6292 |
| 2020 | 5972 |
| 2014 | 5869 |

* **Step 1-** Adjusting AnalyteCode following the instructions all code values of 1038 were changed into 1040 (21,928 rows). Also 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 |
| --- | --- |
| 1040 | 21945 |
| 2050 | 7684 |
| 2039 | 7684 |
| 2987 | 6518 |
| 2984 | 6518 |
| 1005 | 6268 |
| 2950 | 3457 |
| 2456 | 3416 |
| 4010 | 3093 |
| 4006 | 1503 |

* **Step 3 -** Checked if the ConcentrationUnits for each AnalyteCode applies with the values in the Data Dictionary
* **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: 2 × 2  
## is\_half\_LDL number  
## <dbl> <int>  
## 1 0.5 31886  
## 2 NaN 74

We checked for the applied LDL and since some of them do not match to the values in HTG, we cahnged all the concentrations for these rows (31,886) into 0, so in the next step will be checked and changed into appropriate half LDL as suggested by HTG 2023.

* **Step 5 -** Followed Appendix G of HTG 2023, checked the Concentration values against thir Lowest Detection Limit (LDL) and replaced with the suggested values where lower

## [1] "number of revised below LDL value: 32180"

* **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)

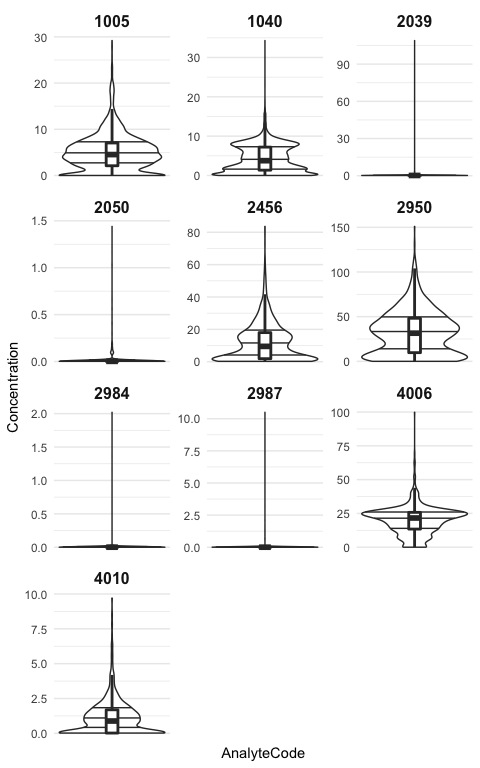
## [1] "initial rows in data: 68086"

## [1] "no. of rows after averaging same day/point/analyte: 64144"

before starting the summarization, we took a look into maximum concentration value for each analyte compared to its median over all dataset (from 2012 to 2021)

## # A tibble: 15 × 4  
## AnalyteCode maximum median DateSampled  
## <dbl> <dbl> <dbl> <date>   
## 1 1005 29.2 4.53 2013-04-15   
## 2 1040 34.3 3.76 2016-12-13   
## 3 2039 109 0.23 2014-10-10   
## 4 2050 1.44 0.002 2014-05-21   
## 5 2050 1.44 0.002 2014-05-21   
## 6 2456 83.6 9.49 2015-07-14   
## 7 2950 151 31.5 2016-07-11   
## 8 2984 2.02 0.001 2020-04-28   
## 9 2984 2.02 0.001 2020-04-28   
## 10 2987 10.5 0.005 2014-10-07   
## 11 4006 99.8 21.6 2012-09-10   
## 12 4010 9.7 0.884 2013-05-09   
## 13 4010 9.7 0.884 2013-05-09   
## 14 4010 9.7 0.884 2013-05-09   
## 15 4010 9.7 0.884 2013-05-09

===> The maximum value for analyte 2039 (DEHP) seems too large compared to its median. May be good to double-check. last column above shows the sampling date for the maximum value. Below, we have also provided plots for the distributions:



## 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) 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: 10 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2012 5496 8490 8 4136  
## 2 2013 4596 6370 8 3344  
## 3 2014 3895 5245 8 2692  
## 4 2015 4516 6319 8 3667  
## 5 2016 4463 5664 8 3102  
## 6 2017 4134 5592 8 2820  
## 7 2018 4962 6808 8 3597  
## 8 2019 4282 5662 8 2885  
## 9 2020 3887 5276 8 2536  
## 10 2021 4349 5787 8 3024

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

..

## # A tibble: 10 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2012 244 579 2 32  
## 2 2013 306 598 2 45  
## 3 2014 282 623 2 18  
## 4 2015 384 820 2 55  
## 5 2016 408 796 2 62  
## 6 2017 368 700 2 44  
## 7 2018 376 682 2 24  
## 8 2019 445 734 2 49  
## 9 2020 398 687 2 22  
## 10 2021 374 654 2 26

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

…

## # A tibble: 10 × 5  
## Year NuLocations NuSamples Analytes Non\_detects  
## <dbl> <int> <int> <int> <dbl>  
## 1 2012 5615 9069 10 4168  
## 2 2013 4771 6968 10 3389  
## 3 2014 4022 5868 10 2710  
## 4 2015 4719 7139 10 3722  
## 5 2016 4716 6460 10 3164  
## 6 2017 4336 6292 10 2864  
## 7 2018 5170 7490 10 3621  
## 8 2019 4558 6396 10 2934  
## 9 2020 4103 5963 10 2558  
## 10 2021 4561 6441 10 3050

* **Step S.4 -** Summarized annual maximums 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: 40 × 5  
## SummaryTimePeriod NuLocations NuSamples Analytes Non\_detects  
## <chr> <int> <int> <int> <dbl>  
## 1 2012-1 978 989 2 312  
## 2 2012-2 1202 1206 2 426  
## 3 2012-3 1035 1050 2 390  
## 4 2012-4 708 715 2 137  
## 5 2013-1 999 1007 2 265  
## 6 2013-2 862 874 2 250  
## 7 2013-3 709 717 2 226  
## 8 2013-4 670 677 2 211  
## 9 2014-1 792 801 2 159  
## 10 2014-2 682 691 2 167  
## # … with 30 more rows  
## # ℹ Use `print(n = ...)` to see more rows

Here we only provided the first 10 rows of total 40.

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

….

## # A tibble: 40 × 5  
## SummaryTimePeriod NuLocations NuSamples Analytes Non\_detects  
## <chr> <int> <int> <int> <dbl>  
## 1 2012-1 39 96 2 4  
## 2 2012-2 45 137 2 3  
## 3 2012-3 119 242 2 24  
## 4 2012-4 41 104 2 1  
## 5 2013-1 53 114 2 7  
## 6 2013-2 43 106 2 2  
## 7 2013-3 165 274 2 34  
## 8 2013-4 45 104 2 2  
## 9 2014-1 50 112 2 3  
## 10 2014-2 57 122 2 3  
## # … with 30 more rows  
## # ℹ Use `print(n = ...)` to see more rows

Here we only provided the first 10 rows of total 40.

* **Step F -** Then the all resulted 5 tables were aggreagated into one table in conformance with HTG and Data Dictionary 2023

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