**Water Quality Data Science Project**

**Background**

The safe to swim program was developed to let the public know which public waters are safe for recreational use. This program is based on testing the waters regularly for certain bacteria and chemical markers. The five markers of interest for this project are: E. Coli, Enterococci, Total Coliform, Fecal Coliform, and HF183. Water Quality Objectives (WQO’s) are maximum amounts of the specific markers that can be present in a given body of water to be considered safe for use. These bodies of water are tested regularly to determine whether they meet the requirements. The sources for WQO’s and testing requirements can be found in the following sources:

* Ocean Plan: <https://www.waterboards.ca.gov/water_issues/programs/ocean/docs/oceanplan2019.pdf>
* Basin Plan: <https://www.waterboards.ca.gov/sandiego/water_issues/programs/basin_plan>
* ISWEBE Plan: <https://www.waterboards.ca.gov/plans_policies/docs/bacteria.pdf>

The below table summarizes these findings (please refer to above documents for units)



There is a large amount of overlap in the numbers and methods present; however, there exists some questions about the calculations and conditions:

* From what start date should the calculations be done?
* How do we incorporate wet/dry weather data

Although these and many other questions still abound, the key take aways are:

* One aspect of determining recreational water safety is by testing the water regularly for certain markers, including: E. Coli, Enterococci, Total Coliform, Fecal Coliform, and HF183.
* WQO’s and testing requirements are outlined in the various water quality plans
* Certain measurements are used to determine whether there is an exceedance of these markers: single sample maximums, geometric/arithmetic means, and statistical threshold values.
* If results indicate these WQO’s are not being met, then the water is not safe for recreational use.

**Work done**

The CEDEN database is the main source of truth for these chemical tests. CEDEN has a lot of columns stored within, but the main columns of interest for this analysis are the following:

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Column Type** | **Column Description** |
| Station Code | String | Unique code name for station where |
| Program | String |  |
| Project | String |  |
| TargetLatitude | Float | Station latitude |
| TargetLongitude | Float | Station longitude |
| DW\_AnalyteName | String | Name of maker |
| Unit | String | Unit of measurement |
| SampleDate | DateTime | Date sample was taken |
| MethodName | String | Sampling method |
| RL | Float |  |
| Result | Float | Numerical result of testing |
| ResultQualCode | String |  |
| SSM\_WQO | Float | Single Sample maximum Water quality objective |
| SSM\_Exceedance | String | Describes quantitative relationship between WQO and Result |
| STV\_WQO | Float | Statistical Threshold Value Water quality objective |
| STV | Float |  |
| STV\_Exceedance | String | Describes quantitative relationship between WQO and STV |
| GM\_WQO | Float | Geometric mean Water quality objective |
| GM\_30 | Float | 30 day rolling geo mean |
| GM\_42 | Float | 42 day rolling geo mean |
| GM\_30\_Exceedance | String | Describes quantitative relationship between WQO and GM\_30 |
| GM\_42\_Exceedance | String | Describes quantitative relationship between WQO and GM\_42 |

The top rows without the grey fill come directly from CEDEN; the bottom, grey filled rows are to be calculated by the scripts developed for this project.

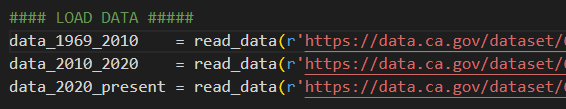
The scripts were broken into four parts:

* data\_transformation\_historical
* ca\_open\_data\_api\_2020
* data\_transformation\_update
* geo\_mean

The first script, data\_transformation\_historical, requires the URL’s for the csv’s of the three databases as given on the California Open Data Source: <https://data.ca.gov/dataset/surface-water-fecal-indicator-bacteria-results>:

****

These are then pasted into the script:



These are then processed by:

* Changing the datatypes
* Calculating the geometric means and statistical value threshold
* Mapping the correct exceedance logic

These are then combined into one csv file to be saved on the network. From then on, the data\_transformation\_update script is then run on a weekly (? to be determined) basis. This will call ca\_open\_data\_api\_2020, which handles the API calls and queries the new data. The data\_transformation\_update script then does the same data transformations, recalculates all statistical values, and appends this data to the csv. This csv will be used with some dashboarding software (PowerBI, Tableau, etc.) for use by R9 staff, both for visualization and to download data of interest.

As of June 2022, all the scripts were written in Python 3.9.4. The required packages are

* Requests
* Pandas
* Numpy

Future work to be done (in order of importance):

* Import into PowerBI
* Add geographic boundaries (currently running for entire dataset/state/?)
* Calculate STV (need to refer to documentation, one possibly methodology is found at the following link: <https://producesafetyalliance.cornell.edu/sites/producesafetyalliance.cornell.edu/files/shared/documents/2017%20GM%20STV%20Worksheet%20v1.0.pdf>)
* Add HF183 (possibly found in <https://data.ca.gov/dataset/surface-water-chemistry-results> )
* If IT unwilling to schedule weekly runs: Convert everything to R
* Add wet/dry weather data (?)