Springboard Data Science Career Track

Capstone 1 Project Proposal

Vicki Brown, August 2019

What is the problem you want to solve?

The problem is to investigate the quality of the water in the San Francisco Bay. Water quality is important. It affects humans (e.g., drinking water, fishing, recreation, tourism) as well as marine life and the environment in general. Studies of changes in water quality over time help us to understand how an ecosystem changes in response to human activities and climate variability.

Who is your client and why do they care about this problem? In other words, what will your client do or decide based on your analysis that they wouldn't have done otherwise?

The client would be anyone who lives or works in the SF Bay Area, especially anyone whose vocation or avocation puts them in contact with the waters of the bay. This would include anyone with a boat, people who like to fish, people who want to swim, anyone with property near the bay, government officials concerned about tourism, and anyone interested in the environment.

I live in the San Francisco Bay area, so the quality of the water in the bay is meaningful to me, personally.

What data are you using? How will you acquire the data?

The data is being collected by the USGS. Water quality data has been collected since 1969. Phytoplankton data has been collected since 1992. Data is archived in the USGS ScienceBase catalog.

Data is readily available for download from the ScienceBase catalogue, or the SFBay USGS home page, in CSV format. (More rows and fields are available if water quality data is downloaded directly from the SFBay USGS home page.)

Links

- SFBay USGS home page
- Water quality data query page.

ScienceBase catalogue

- Water quality 1969 - 2015
- Water quality 2016, ...
- Phytoplankton 1992 - 2014
- Phytoplankton 2016...

Briefly outline how you'll solve this problem. Your approach may change later, but this is a good first step to get you thinking about a method and solution.

The water quality and phytoplankton datasets can be joined on the sampling station field. This will result in a "water quality and phytoplankton" data set with 23 features (columns). Because the samples come from different stations and different water depths, we can make various comparisons, e.g. between the same depth at different locations or different depths at the same

location. We can compare locations that are near each other to locations at greater distance. We can compare phytoplankton amounts to chlorophyll, temperature, or other values.

As the data samples are date-stamped, we can also compare results between seasons (e.g. January vs June), month to month, year to year, or across decades. We can combine comparisons by date with comparisons by location.

The SFBay USGS web site suggests:

The U.S. Geological Survey maintains a measurement program designed to describe the patterns of water quality variability in San Francisco Bay. The individual water-quality constituents (e.g. salinity or dissolved oxygen) are measured: in the vertical dimension (from the water surface to the bottom at each station), along the longitudinal dimension (from South Bay to the Sacramento River), and over time. This means that we can display different kinds of patterns of variability from the USGS data set.

Examples:

- Time Series Plots -- variability over time, at one location, of the parameters:

Chlorophyll	Temperature	Dissolved Oxygen
Salinity	Light Penetration	Suspended Particulate Matter
Ammonium	Nitrogen	Phosphorus
Silicates		

- Vertical Profiles – variability with depth, at one location and time, for the same parameters as above
- Longitudinal Sections – vertical and longitudinal variability for one sampling date
- Space and Time Contours – seasonal changes in the spatial patterns along the longitudinal axis of the Bay-Delta
- X-Y Scatter Plots – Plot and compare water quality parameter variability on any two dates.

What are your deliverables? Typically, this includes code, a paper, or a slide deck.

Deliverables will include code, documentation, and graphics in a Jupyter notebook.