Capstone 1 - SF Bay Water Quality

Data Storytelling Report

This is a report on some of the descriptive and visualization methods used in my first capstone project.

Data storytelling is an important skill in data science. Doing exploratory data analysis early also allows you to understand emerging themes and share early results with your clients, before moving onto deeper and more complex data analysis.

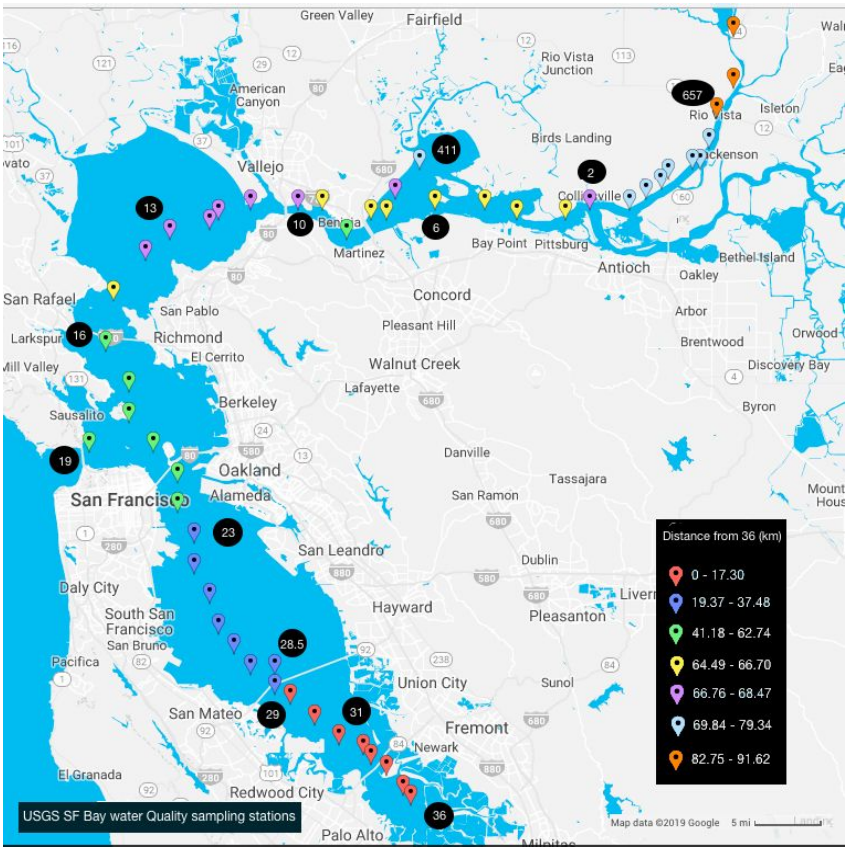
Tasks

Ask the following questions and look for the answers using code and plots:

- Can you count something interesting?
- Can you make a bar plot or a histogram?
- Can you find trends (e.g. high, low, increasing, decreasing, anomalies)?
- Can you make a scatterplot?
- Can you make a time-series plot?
- Can you compare two related quantities?
- Looking at the plots, what are some insights you can make? Do you see any correlations? Is there a hypothesis you'd like to investigate further? What other questions do the insights lead you to ask?

Sampling Stations

Map of sampling stations. Distance from station 36 has been computed. For additional detail, see the interactive version of the map. (map generated using BatchGeo)



Can you count something interesting?

Were some stations sampled more often than others?

```
Number of stations sampled: 43
median sampling frequency: 4957.0
```

Lowest sampling frequency:

```
station     frequency
655             32
```

Highest sampling frequency:

```
station     frequency
18           16622
```

Stations are definitely not sampled at the same frequency.

How many different depths were sampled?

```
Number of depths sampled: 99
shallowest depth 0.5 meters; frequency 1962
deepest depth 80.0 meters; frequency 2
median depth 7.0 meters
```

How many actual days of sampling?

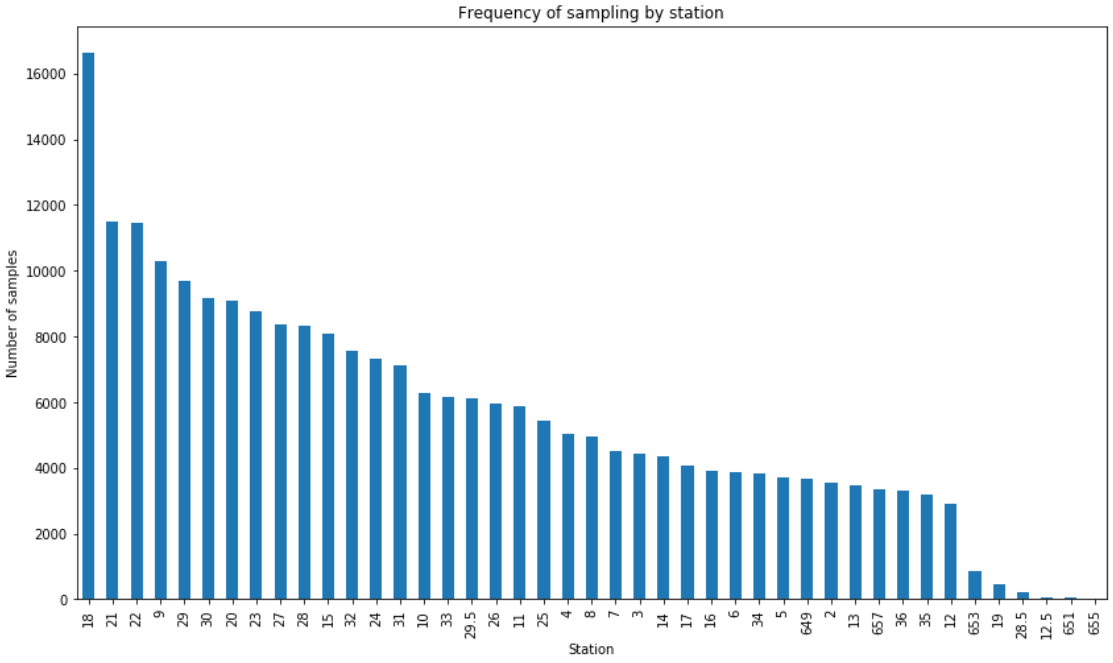
We know the data covers 40 years, but samples were not taken every day.

```
Number of days sampling occurred: 1172
Earliest date seen: 1969-04-10
Latest date seen: 2019-06-04
Days elapsed between these dates: 18317
```

Can you make a bar plot or a histogram?

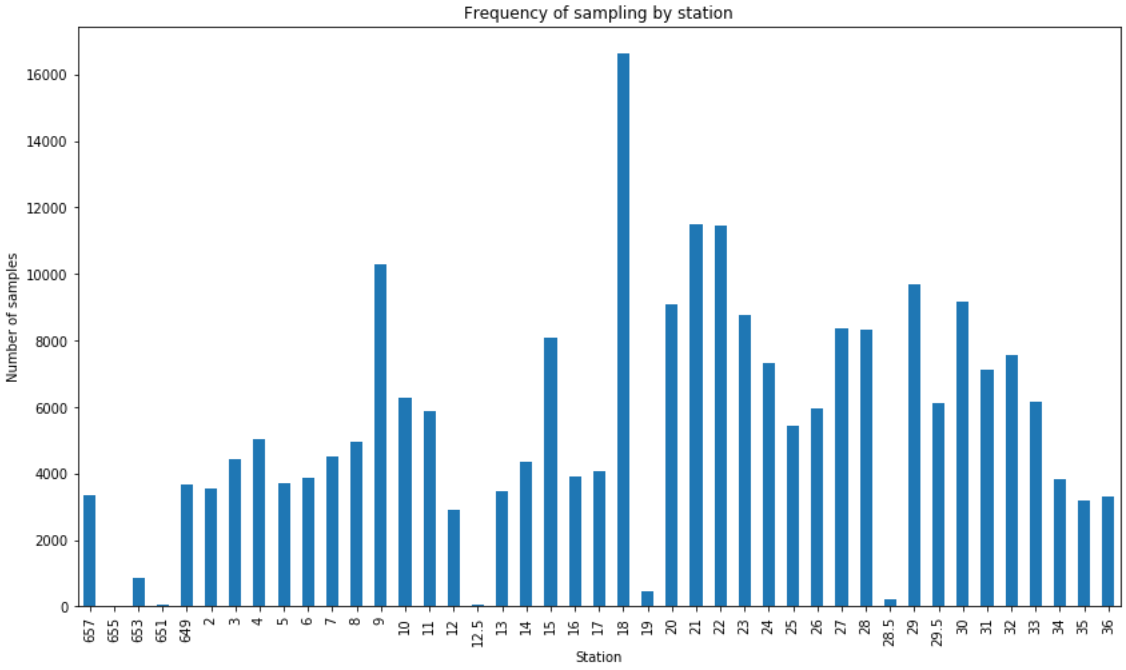
Station sampling frequency

Station sampling frequency plot, ordered by frequency, highest to lowes



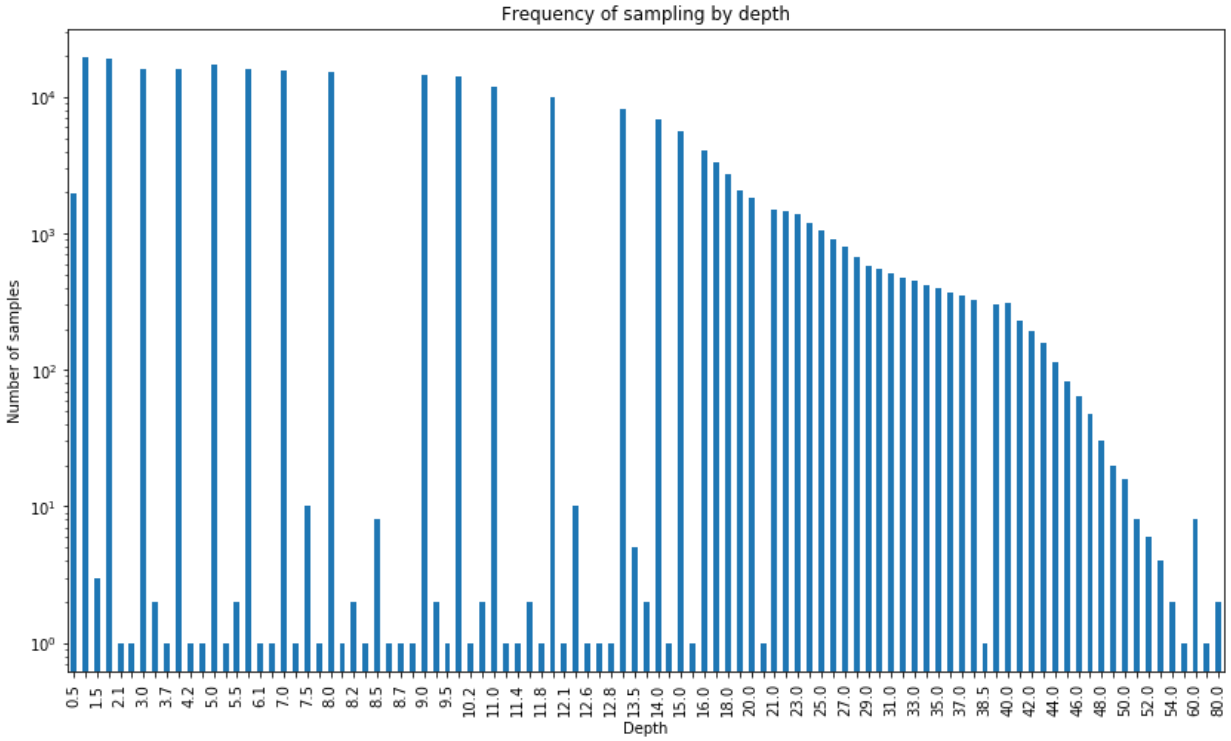
t.

Station sampling frequency plot, ordered by station number, north to south.



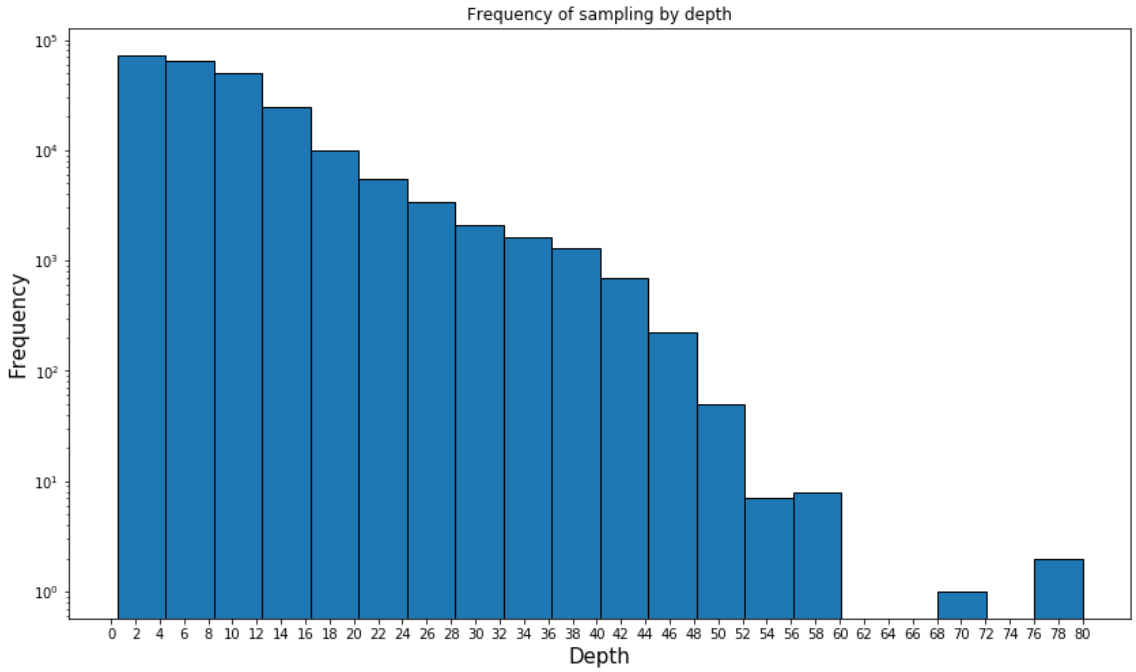
Depth sampling frequency

Depth frequency, bar plot.

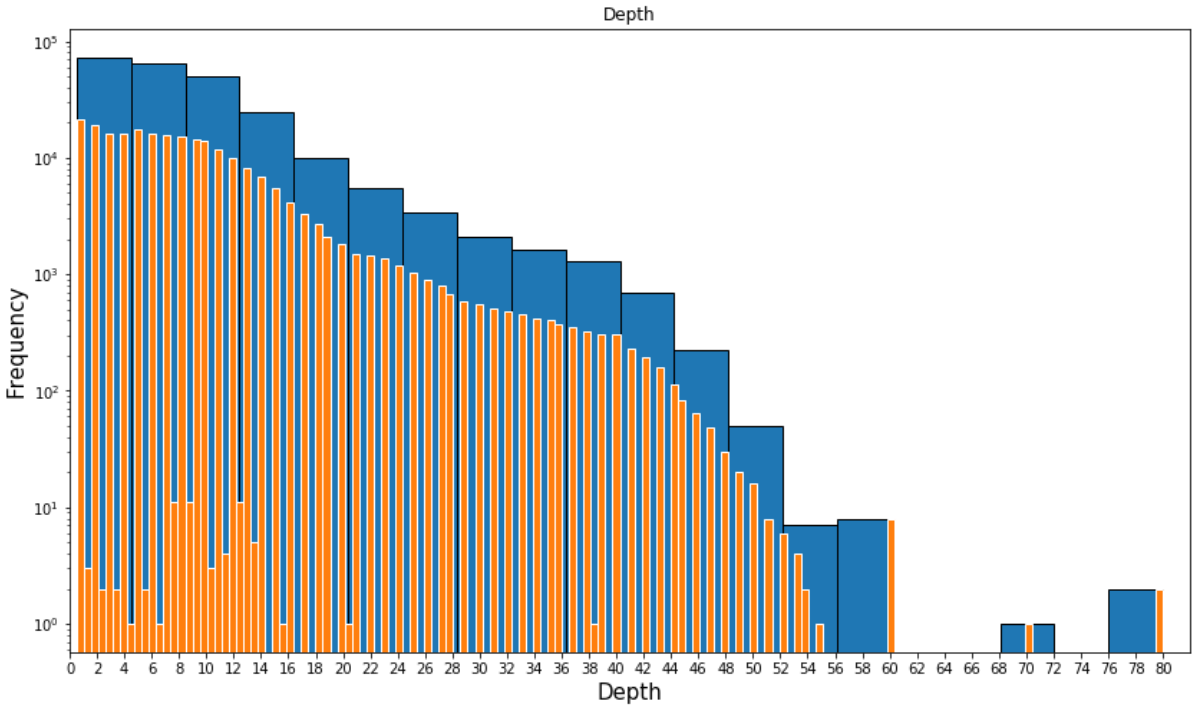


While 99 different depths were recorded, many were sampled less than 100 times in 40 years. Many of the shallower depths can be aggregated.

Depth frequency histogram.

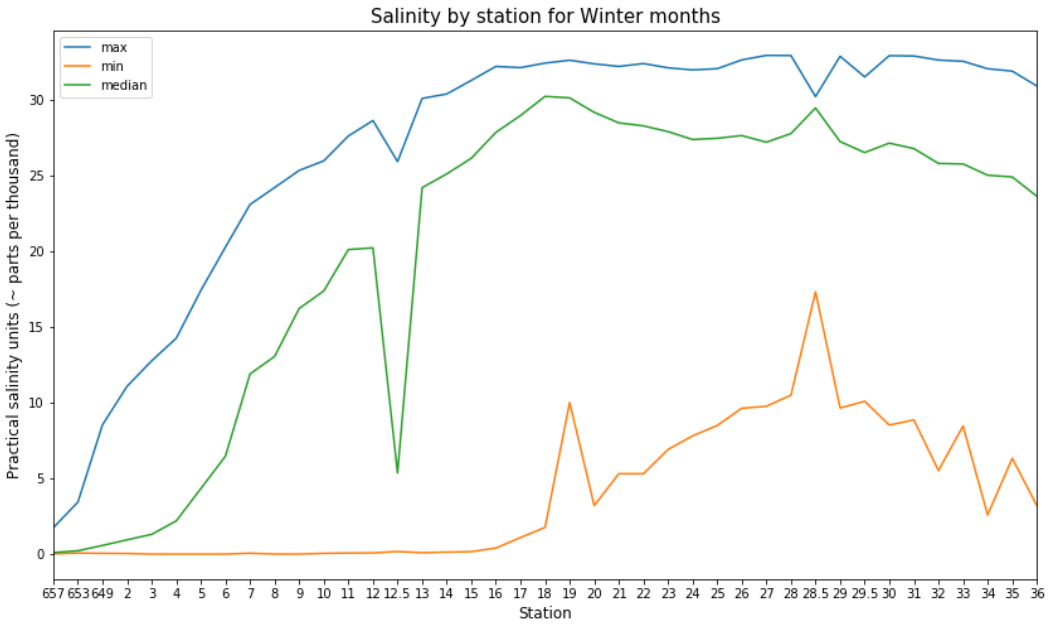


I'd like to overlay the bar chart and the histogram. Conveniently, I can simulate the bar chart using a histogram with many bins, then overlay two histograms on the same plot.



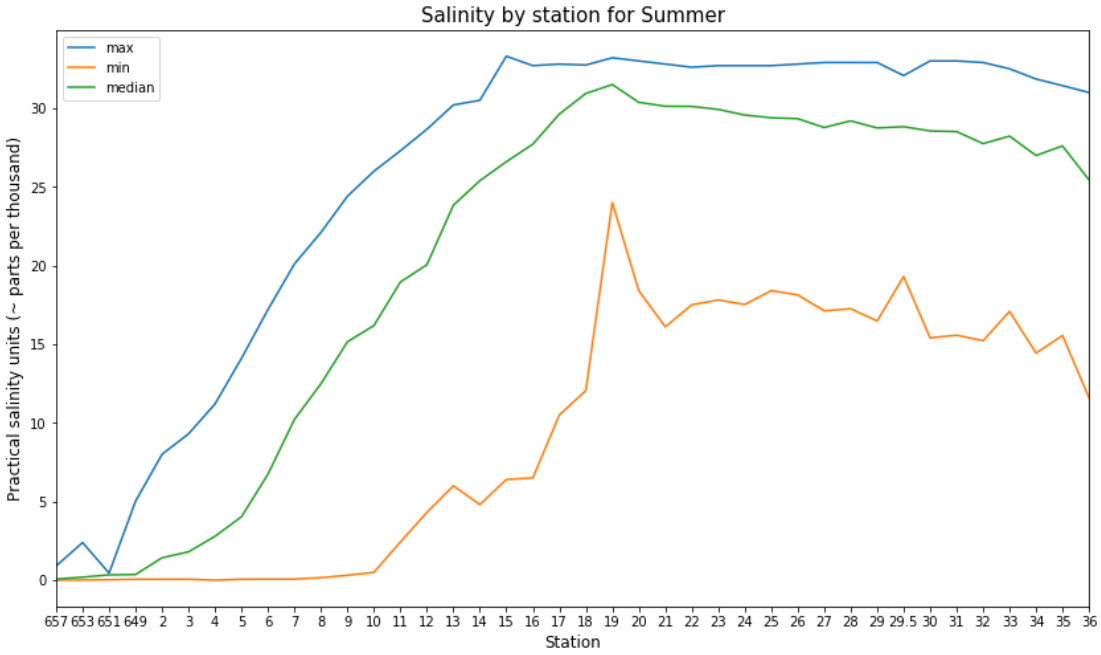
Can you find trends (e.g. high, low, increasing, decreasing, anomalies)?

Let's look at salinity differences by station. Start by extracting only records for winter months, (December, January, February). For each station, calculate statistics on salinity across all winter samples, regardless of depth.



There is a large anomaly at station 12.5 and a smaller one at 28.5. (Note that when I only checked January samples, the station 28.5 anomaly was much larger; the max, min, and median values were almost identical for 16 samples during all Januarys.) These stations warrant further investigation.

Now, let's plot salinity by station for the summer months (June, July, August).

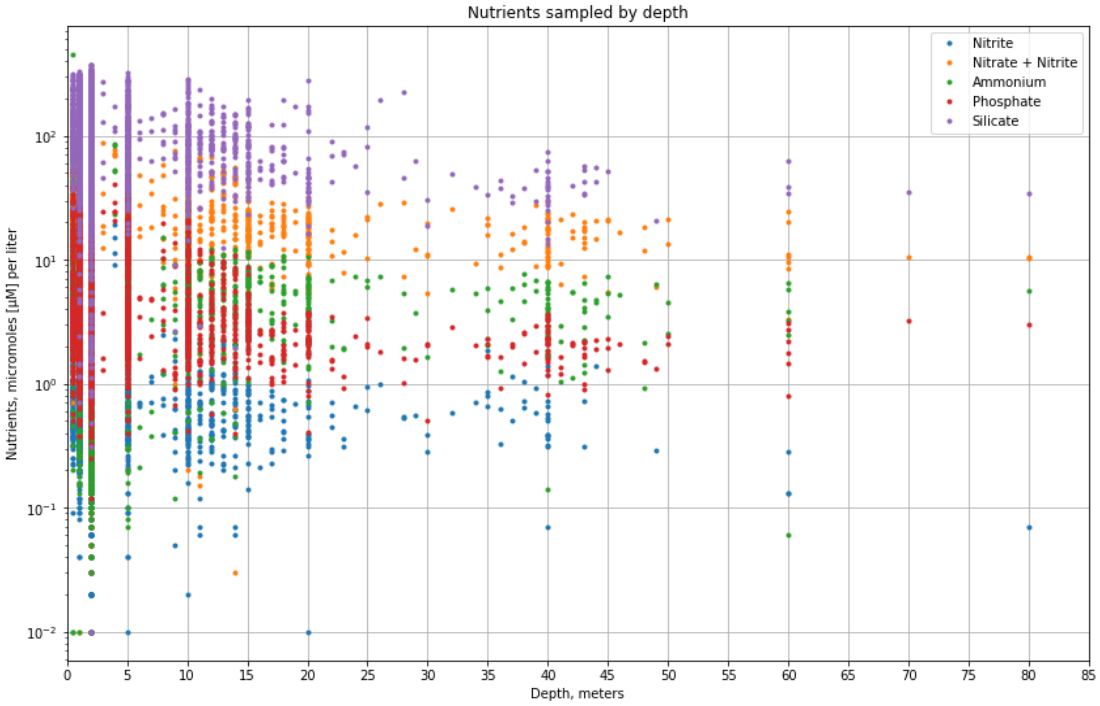


Salinity is, as I would have guessed, lowest in the Sacramento River. It climbs steadily as the water heads through Suisun Bay and San Pablo Bay and peaks in the Central Bay at the Golden Gate. Salinity then remains high going south toward San Jose.

Maximum salinity doesn't seem to change much between January and July. Minimum salinity is much lower in January, probably due to rain and storms. However, minimum salinity is always higher at stations 18, 19, and 20, just inside the Golden Gate.

Can you make a scatterplot?

During the data wrangling phase, I made a scatterplot of nutrients, sampled by depth.



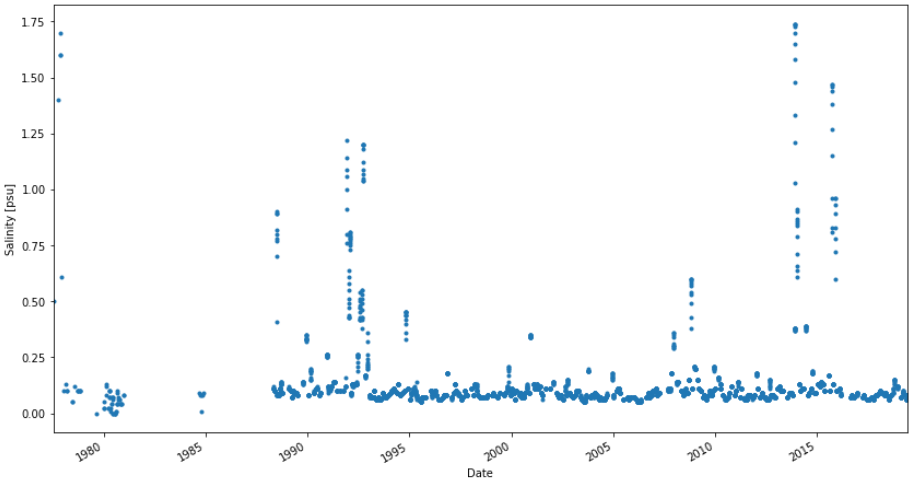
Can you make a time-series plot?

For three stations, plot salinity and temperature across all years sampled. The stations chosen are 657 (at the Sacramento River), 18 (at the mouth of the Golden Gate), and 36 (at San Jose, the southern-most station).

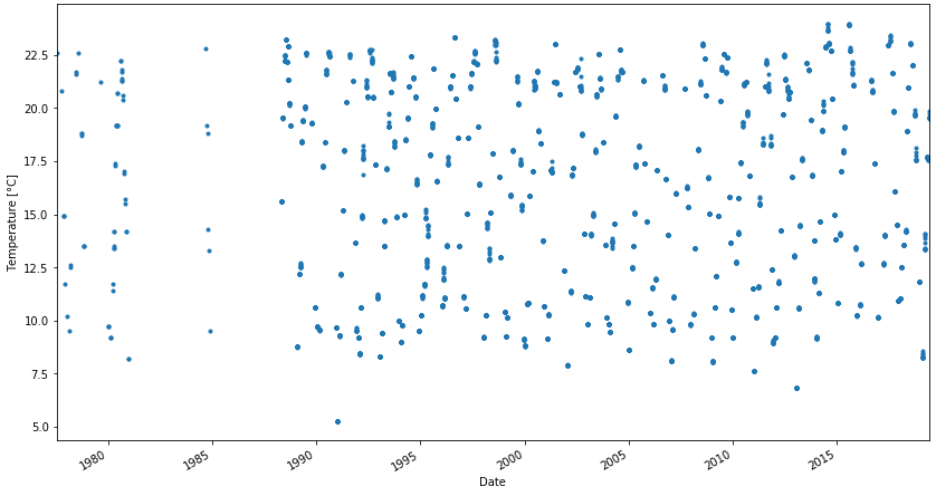
Note that the scale for salinity for station 657 is very different from the scale used for stations 18 and 36. If we used the same (0,30) scale for all three, no changes in salinity would be visible for station 657, where all values are between 0 and 2 ppt.

Station 657

Station 657 salinity by year:

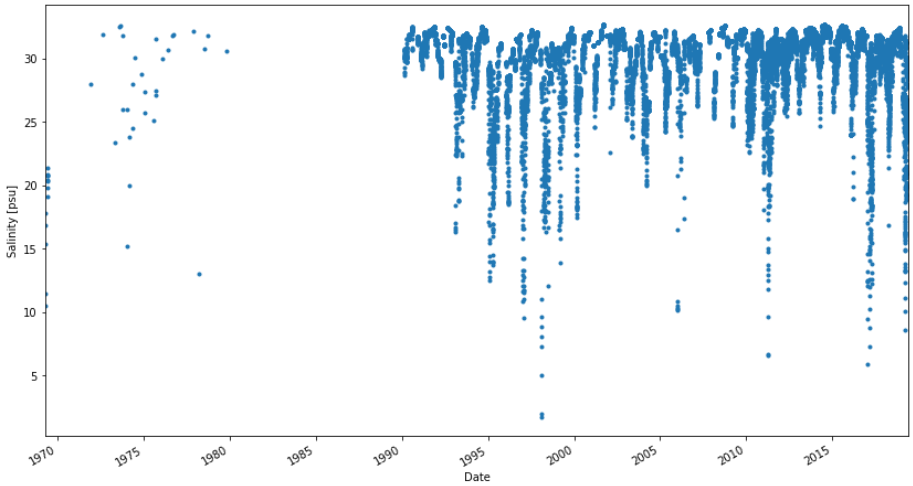


Station 657 temperature by year:

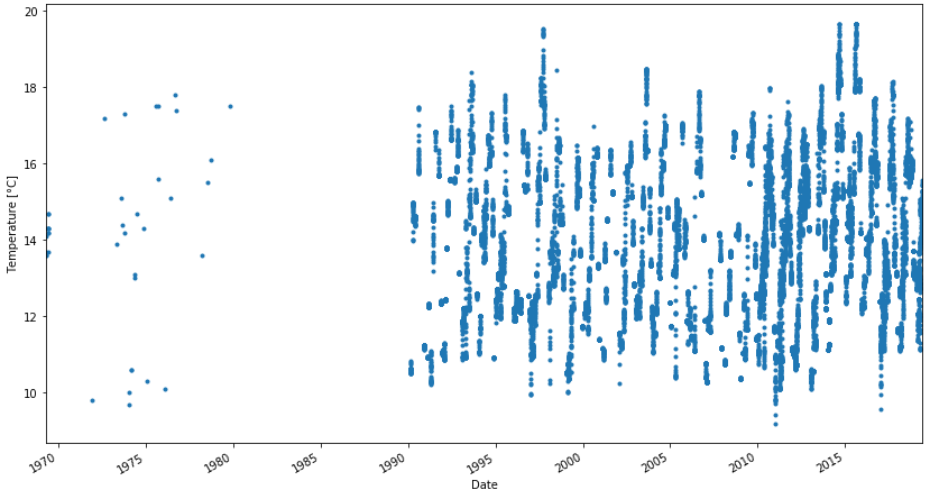


Station 18

Station 18 salinity by year:

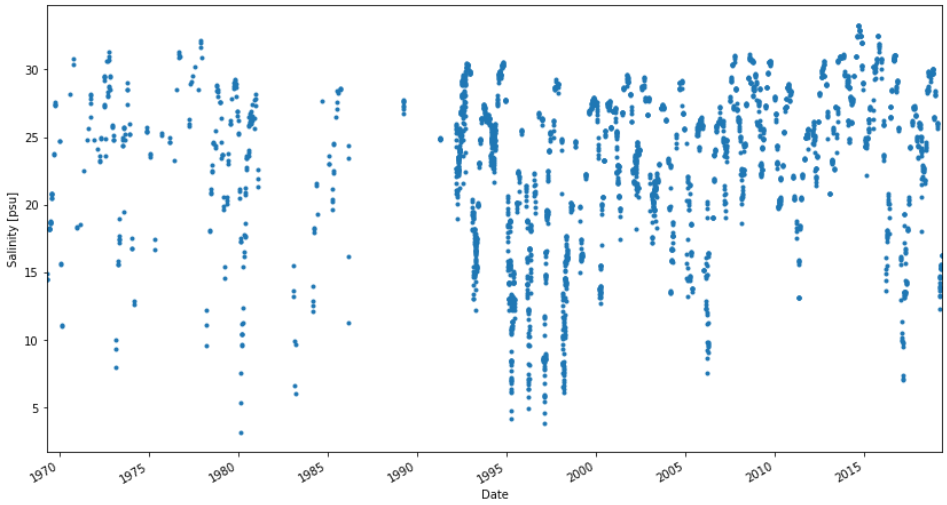


Station 18 temperature by year:

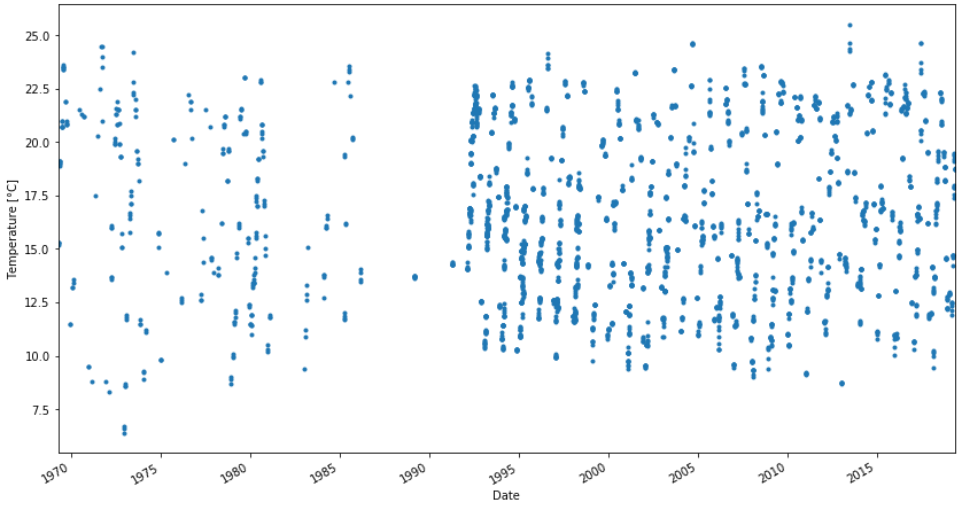


Station 36

Station 36 salinity by year:



Station 36 temperature by year:



Can you compare two related quantities?

How does chlorophyll compare to oxygen?

How does biovolume compare to

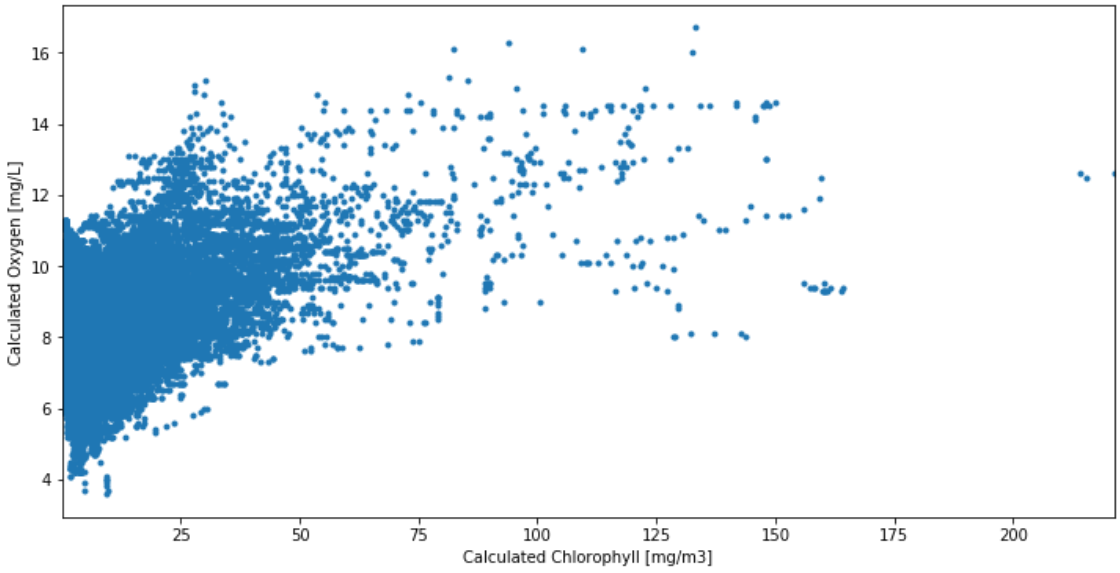
- Chlorophyll
- Oxygen
- Nutrients

Hypothesis:

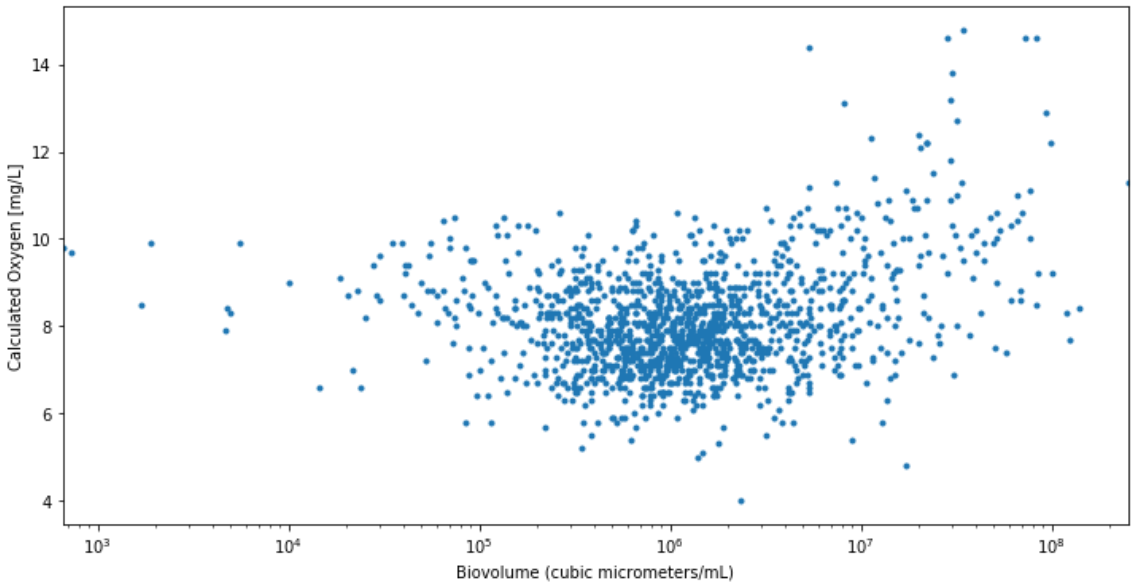
- O2 should increase as chlorophyll increases
- Chlorophyll should increase as biovolume increases
- All nutrients should decrease as biovolume increases (consumed)

Oxygen vs. Chlorophyll and Biovolume

O2 vs chlorophyll:

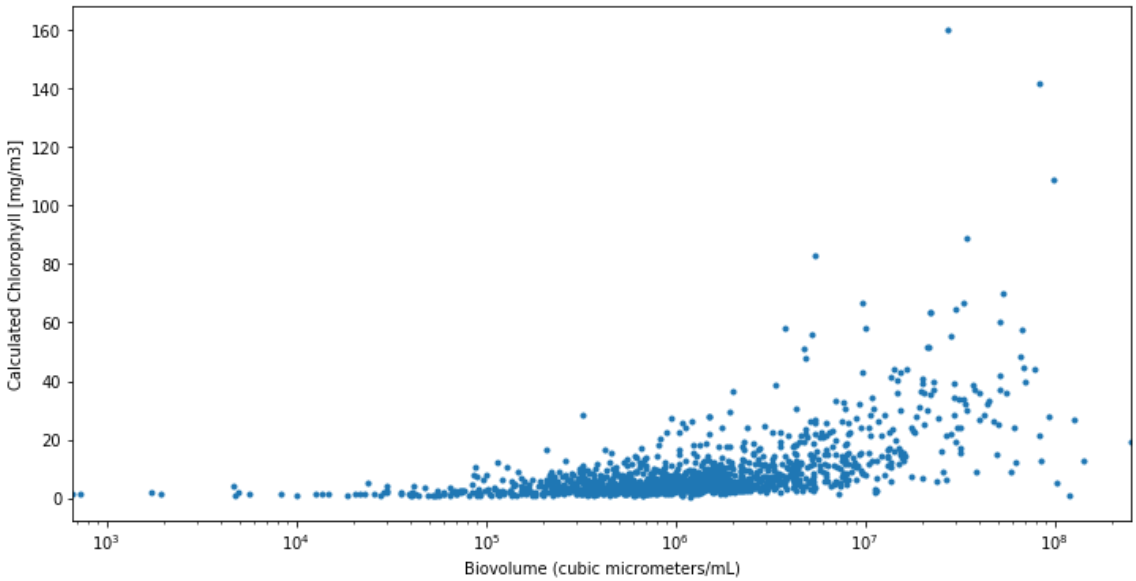


O2 vs. biovolume:

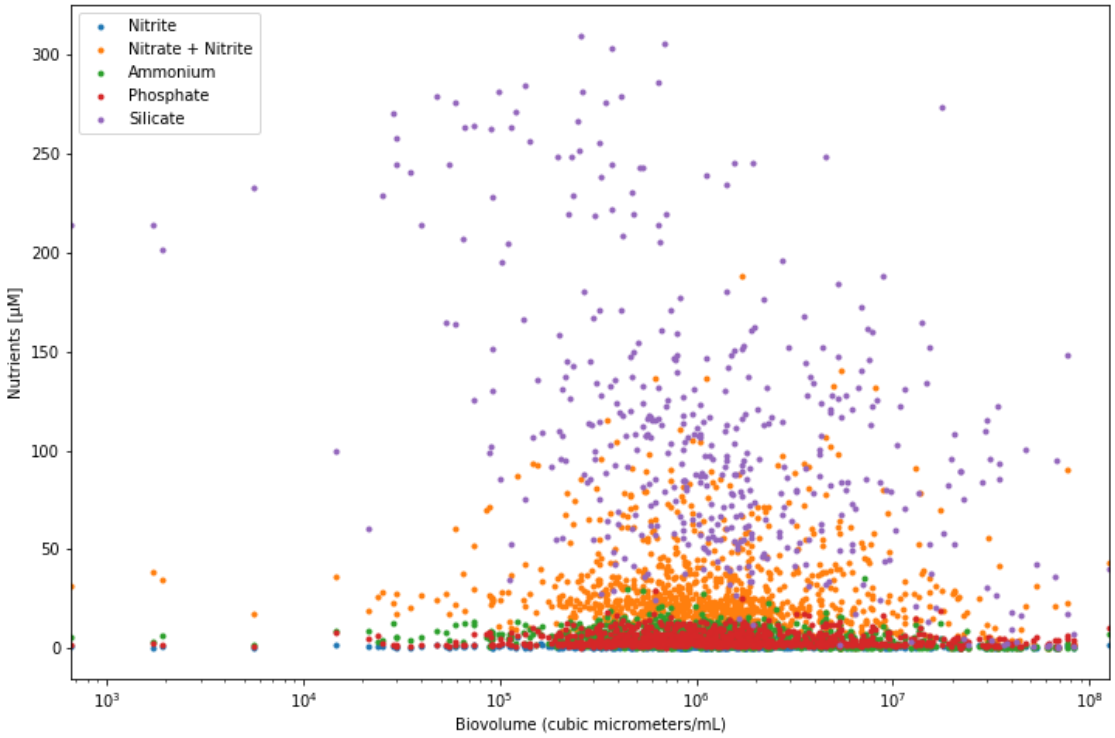


Biovolume vs. Chlorophyll and Nutrients

Biovolume vs. Chlorophyll:



Biovolume vs. Nutrients



Looking at the plots, what are some insights you can make? Do you see any correlations?

Salinity

Salinity is lowest in the Sacramento River. It climbs steadily as the water heads through Suisun Bay and San Pablo Bay and peaks in the Central Bay at the Golden Gate. Salinity then remains high south to San Jose.

Maximum salinity doesn't change much between winter and summer. Minimum salinity, however, is much lower in the winter months, probably due to rain and storms.

Minimum salinity is always highest at stations 18, 19, and 20, just inside the Golden Gate.

Changes by year

The plots for three stations over 40 years show that temperature varies considerably, but more likely by depth and month than by station. I could dig deeper into this.

Salinity differences are very clear between these three stations and the differences are consistent from year to year.

Phytoplankton

As nutrients increase, phytoplankton biovolume increases as well. However, as biovolume continues to increase, nutrients begin to decrease as they are consumed.

Similarly, as phytoplankton increase, so do chlorophyll and calculated oxygen.

Is there a hypothesis you'd like to investigate further? What other questions do the insights lead you to ask?

I should look more closely at stations 12.5 and 28.5. Salinity shows an unexpected dip during the winter months. Do any other parameters have unexpected levels compared to stations nearby?

Does biovolume vary more obviously with one of the nutrients vs any of the others? In summer months vs winter? At stations closer to the river or further south? I can also look at biovolume compared to salinity, temperature, or O2.

I would like to do more station by station plots, as the USGS has done in their *Nature* article.

I would like to compare biovolume or O2 to SPM (suspended particulate matter) and also determine SPM by station. Are parts of the Bay "murkier" than others? How does the extinction coefficient (light attenuation) relate to SPM? To biovolume?