San Francisco Bay Water Quality

Problem statement

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.

The client is 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.

Description of the dataset

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

USGS Contact: Tara Schraga

Approach

I downloaded the water quality data in three sets, then merged the sets. Phytoplankton datasets are separate and were joined to the rest of the data using the sampling station field.

This resulted 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.

Data Wrangling

Water Quality Data

Data Acquisition

Access

No API is available. I used the "Expert Query" form, requesting data in three chunks, saved as CSV files:

- Julian Date < 1999001
- 1999001 < Julian Date < 2009001
- Julian Date > 2009001

Note: Water quality data is also available for download from ScienceBase; however, that archive includes fewer parameters and is not as up to date as the database at `sfbay.wr.usgs.gov`.

Data Review

Water Quality files have two header rows; the second row shows units of measure.

```
Date, Time, Station Number, Distance from 36, Depth, Discrete Chlorophyll,
Chlorophyll a/a+PHA, Fluorescence, Calculated Chlorophyll, Discrete Oxygen, Oxygen
Electrode Output, Oxygen Saturation %, Calculated Oxygen, Discrete SPM, Optical
Backscatter, Calculated SPM, Measured Extinction Coefficient, Calculated Extinction
Coefficient, Salinity, Temperature, Sigma-t, Nitrite, Nitrate + Nitrite, Ammonium,
Phosphate, Silicate
```

```
MM/DD/YYYY, 24 hr., , [km], [meters], [mg/m3], , [volts], [mg/m3], [mg/L], [volts],
, [mg/L], [mg/L], [volts], [mg/L], [per meter], [per meter], [psu], [°C], [kg/m3],
[µM], [µM], [µM], [µM], [µM]
```

I created a Python dictionary to hold the units, then changed the column headers to one row.

Cleaning

What kind of cleaning steps were performed?

- Combine multiple datasets.
- Convert Date and Time columns to DateTime.
- Remove columns that are not useful.

Combine multiple datasets

The data was imported in three files. These files have identical columns, so they were easily concatenated into one DataFrame and exported to a single file.

Convert Date and Time columns to DateTime

The initial dataset had a Date column and a Time column. I would rather have a single DateTime column.

There were a few issues to get past

- The initial Date column is type `string` (e.g., `M/D/YYYY`), no leading zeroes on day or month, but possibly a leading space. Conveniently, `pd.to_datetime` is able to convert this to DateTime format without trouble.
- The initial Time column is type `int`, no leading zeroes on the hour. To concatenate this to the Date column, I need it to be type `string`, 0-padded.

Once I had two strings, I concatenated them into a new DateTime column and converted that to Datetime format.

Remove Columns that are not useful

Optical Backscatter

According to the data dictionary, due to sensor changes and gain differences, this value is only comparable within cruises and may not be comparable between cruises.

Discrete vs Calculated values: chlorophyll, SPM, and O2

The USGS contact suggested that I should ignore "discrete" values for chlorophyll, SPM, and oxygen, using the "calculated" values instead. The latter values are calculated using linear regression between the discrete values and other measurements.

A quick look confirmed that summary statistics for these calculated values match those for the comparable discrete values.

I removed the "discrete" columns from the dataset.

Time

I have a `DateTime` column, so I no longer need the `Time` column. I removed it as well.

Missing Values

Some of the water quality parameters (particularly the dissolved "nutrients") are not checked for every sample. However, there were no columns that were entirely missing data and no obvious overlap of columns.

Further investigation indicated that nutrient (e.g., nitrate, phosphate, ammonium) samples are typically only taken near the surface (e.g., 1 - 2 m depth). This was confirmed by the USGS contact for the dataset, who told me that we can assume dissolved nutrient values from surface to bottom are generally the same. "The Bay is well mixed."

Solution: These values are not "missing".

Given that nutrients are sampled for only a subset of records (primarily shallower depths), I created a separate DataFrame in which these samples are present.

Outliers

I calculated summary statistics per sampling station for each remaining column in the dataset. The statistics included `zscore`. I also plotted values in scatter plots by date.

There were a few apparent outliers for several parameters at most stations. However, at this point in the analysis, I do not anticipate that these will cause problems. If I need to remove them at a future date, I will easily be able to locate them.

Station Location Information

Data Acquisition

Location data for "standard" stations is available from ScienceBase.

However, more complete location data is available in tables at sfbay.wr.usgs.gov. These tables include the general location of each station (by geographical landmark) as well as data for "non-standard" stations which are sampled less often.

These tables were copied, pasted into a spreadsheet, then exported as CSV. Header fields were edited to remove newlines and several fields were modified to remove artifacts before exporting to CSV format.

Data Review

The Station Locations file is CSV format with one header row and 5 columns.

```
Station Number, General Location, North Latitude, West Longitude, Depth MLW (meters)
```

Cleaning

I did some cleaning of the data in Numbers (Mac OS spreadsheet app) before importing into Jupyter Notebooks. I removed:

- newlines in the column headers
- non-ASCII characters in the data
- extraneous quotation marks
- an asterix in a station number field

After importing, I converted the Station numbers to categories (rather than strings) so that I could set the sort order geographically from northernmost to southernmost station through the bay.

Missing Data

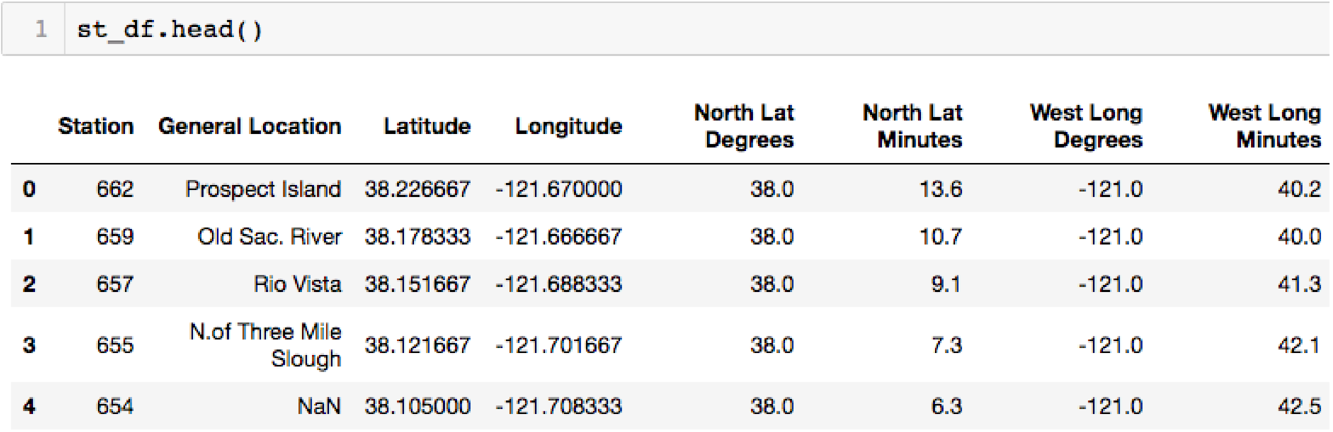
Many records in the table did not include geographic degrees; instad, they inherited this information from the station on the row above. I wanted to fill in this data so that ach ow was complete in itself and did not rely on other rows for information.

I created four new columns, extracting the degrees and minutes from the previously existing `North Latitude` and `West Longitude` columns:

- North Lat Degrees
- North Lat Minutes
- West Long Degrees
- West Long Minutes

I then filled in missing degrees using the most recent non-null value above. When the new columns were filled, I dropped the original `North Latitude` and `West Longitude` columns.

Next, I created new `Latitude` and `Longitude` columns which contain degrees in decimal format (preferred by some mapping applications.) Finally, I reordered the DataFrame columns.



```
st_df.head()
```

	Station	General Location	Latitude	Longitude	North Lat Degrees	North Lat Minutes	West Long Degrees	West Long Minutes
0	662	Prospect Island	38.226667	-121.670000	38.0	13.6	-121.0	40.2
1	659	Old Sac. River	38.178333	-121.666667	38.0	10.7	-121.0	40.0
2	657	Rio Vista	38.151667	-121.688333	38.0	9.1	-121.0	41.3
3	655	N.of Three Mile Slough	38.121667	-121.701667	38.0	7.3	-121.0	42.1
4	654	NaN	38.105000	-121.708333	38.0	6.3	-121.0	42.5

Phytoplankton Data

Data Acquisition

Phytoplankton data was downloaded in three files from ScienceBase. * ScienceBase: Phytoplankton, 1992-2015 * ScienceBase: Phytoplankton, 2016...

The apparent overlap in dates for Phytoplankton files, `Phytoplankton_San_Francisco_Bay_1992_2014.csv` and `Phytoplankton_San_Francisco_Bay_2014_2016.csv`, is not an error. Data from the first three months of 2014 is included in two separate files. Possible redundant rows from 2014 will need to be investigated.

Data Review

All files are formatted as CSV (comma-separated values). Phytoplankton files have one header row.

Column headers include units of measure. Units were removed from the headers and saved in a Python dictionary.

```
Taxonomic Identification, Phylum or Class, Date,Station Number, Depth (m), Actual Count, Density (cells/mL), Biovolume (cubic micrometers/mL), Cell Volume (cubic micrometers/cell)
```

Note: The Actual Count parameter was not included before 2014. The recommendation from USGS is to not use this field, as it is essentially "for internal use only".

Cleaning

File Encoding

When reading in the Phytoplankton data, I got an error due to file encoding.

Once I understood the file encoding a bit better, I fixed the encoding of the CSV files on disk to be UTF8.

Combine multiple datasets

One of the imported Phytoplankton datasets had one fewer columns than the other two. Thus, a simple concatenation deos not work. Setting sort=True aligned the columns-in-common, allowing concatenation with the missing column not causing problems.

Possibly redundant records in phytoplankton data for 2014

From the filenames, I suspected there might be overlap for part of 2014. An "overlapping" row would match on every field except for Actual Count, which would be either a number or NaN.

I examined some of the 2014 data and did not see any clear overlap.

DateTime values

The format of the Date column is such that *pandas* already sees it as a DateTime.

Reorder Columns

Due to the need to sort columns before concatenation, the data columns were now in alphabetical order, left to right. Before saving, I re-ordered the columns, moving Date, Station number, and Depth to the front.

Missing Data and Non-useful Columns

The only missing data was in the Actual Count column. This value was not part of the dataset before 2014.

Actual Count is described as the "Number of phytoplankton cells counted in the sample". However, the sample size is not provided.

Actual Count values seem to be at odds with Density values. Density is described as the "Number of phytoplankton cells per milliliter of water. Most "actual counts" are very small; the mean is 25 cells. But Density is much larger; that mean is 2803.

I asked my USGS contact, Tara Schraga. Her recommendation is that I ignore the Actual Count field.

- It's primarily a "notation" field; someone saw these cells under a microscope.
- The sample size is indeterminable.
- It's unrelated to the rest of the data.

I removed this column.

Tara also recommended removing the Density and Cell Volume columns, concentrating instead on `Biovolume`. I removed these two columns.

Outliers

I am not particularly worried about outliers in this dataset as the only numeric value is a calculated value.

Consolidate Biovolume and Merge with Water Quality

The original phytoplankton data had multiple entries for each date and station, separating the phytoplankton by genus and species. Per recommendation from Tara Schraga at USGS, I consolidated this information for each {date, station, depth} combination, producing a simple sum total of biovolume.

I merged the resulting column into the WQ DataFrame.

Statistical Investigation

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?

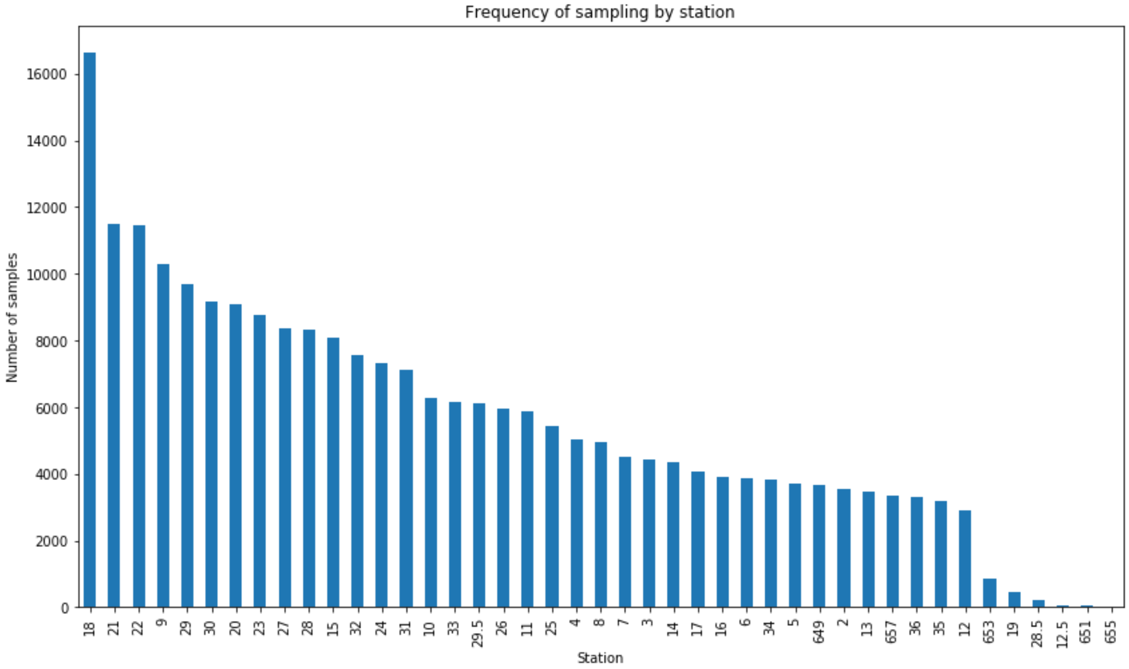
```
Number of depths sampled: 99

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?

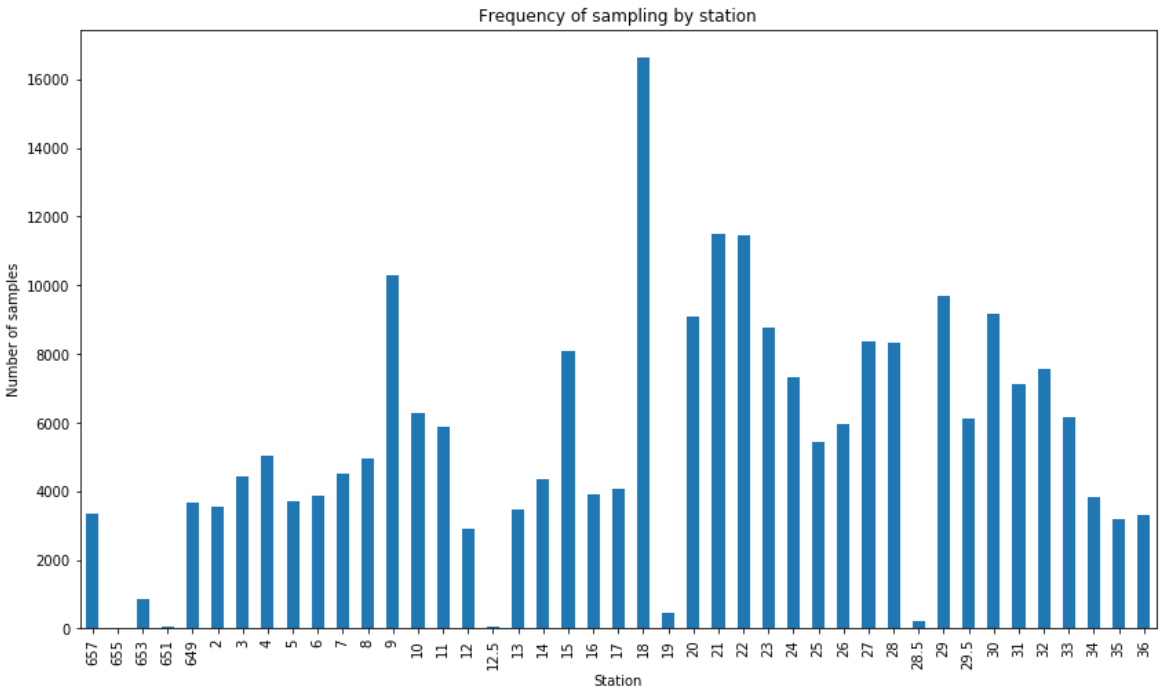
We know the data covers 40 years, but samples were not taken every day.

```
Number of days sampling occurred: 1172

Earliest date seen: 1969-04-10

Latest date seen: 2019-06-04

Days elapsed between these dates: 18317
```

Station sampling frequency



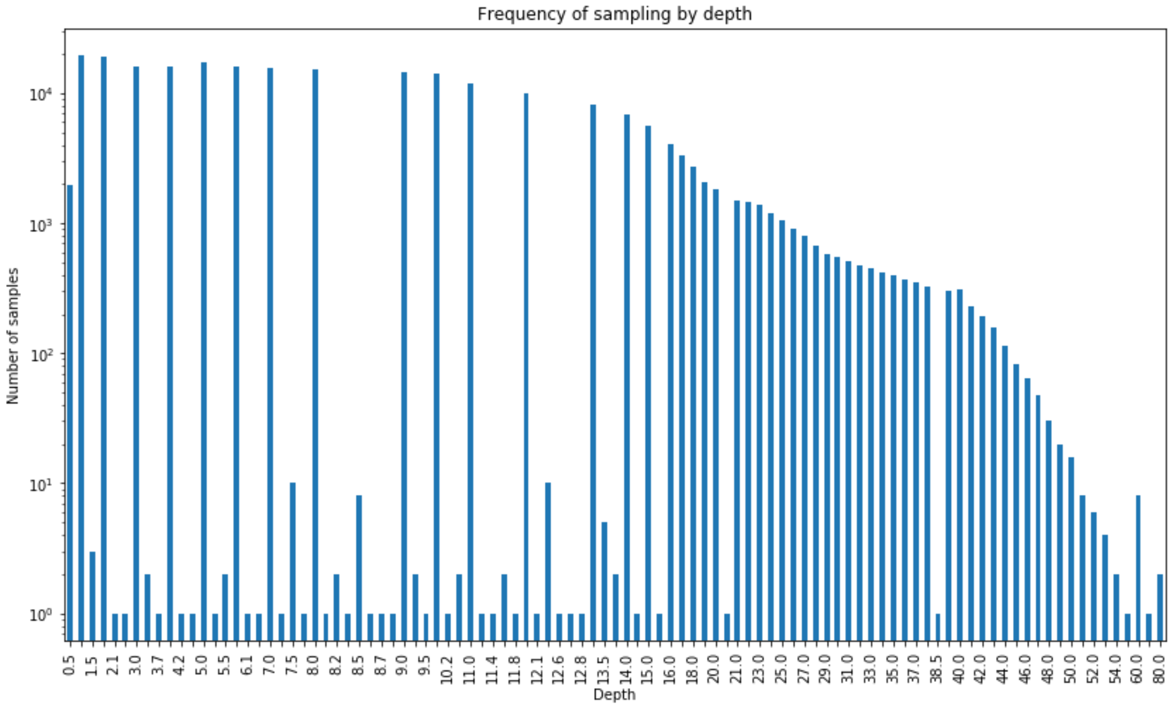
Station sampling frequency plot, ordered by frequency, highest to lowest.



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

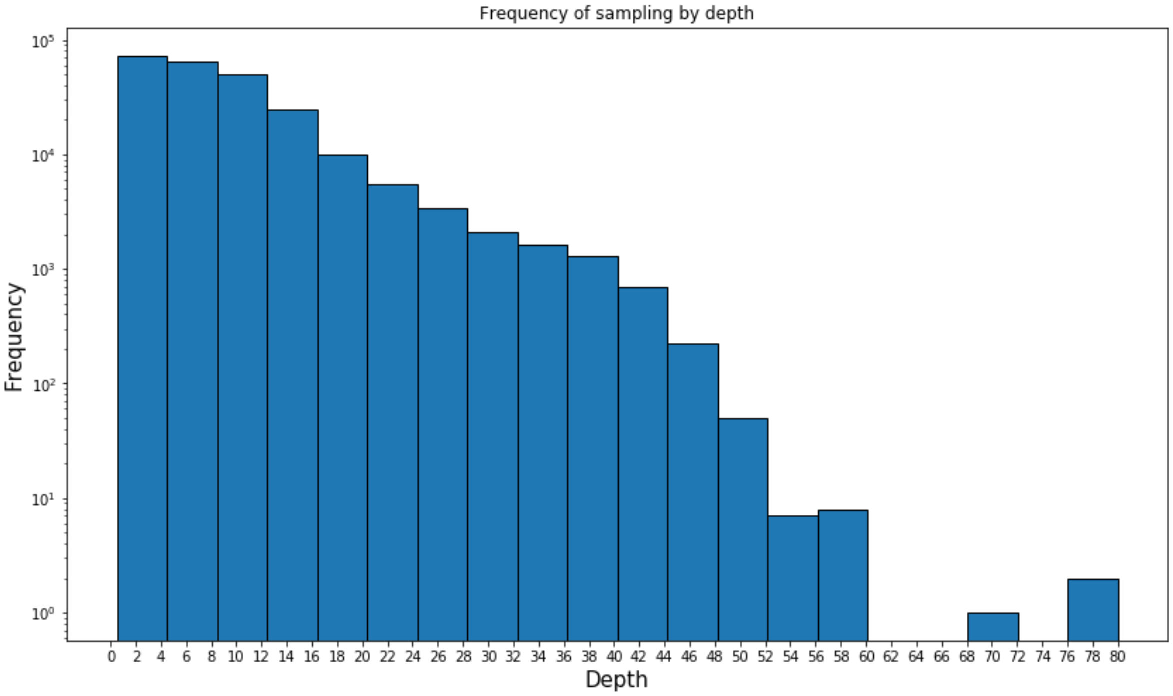
Depth sampling frequency

Depth frequency, bar plot.

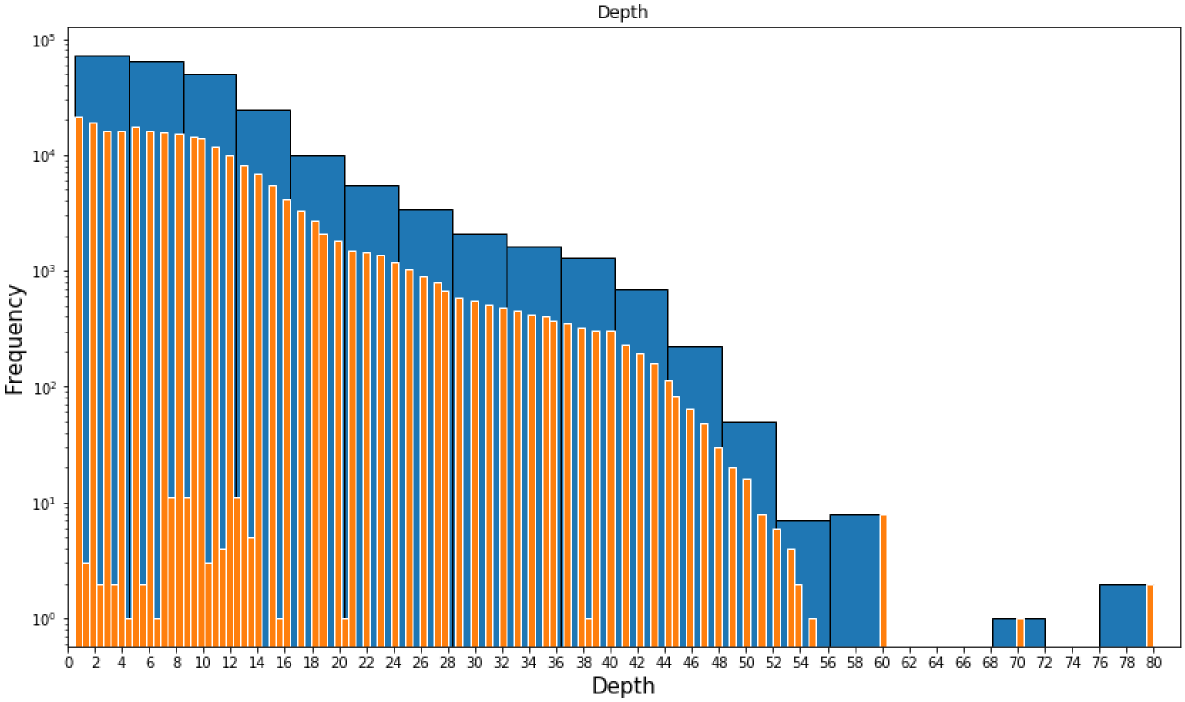


While 99 different depths were recorded, many were sampled less than 100 times in 40 years. Many of the shallower depths can be aggregated.

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.

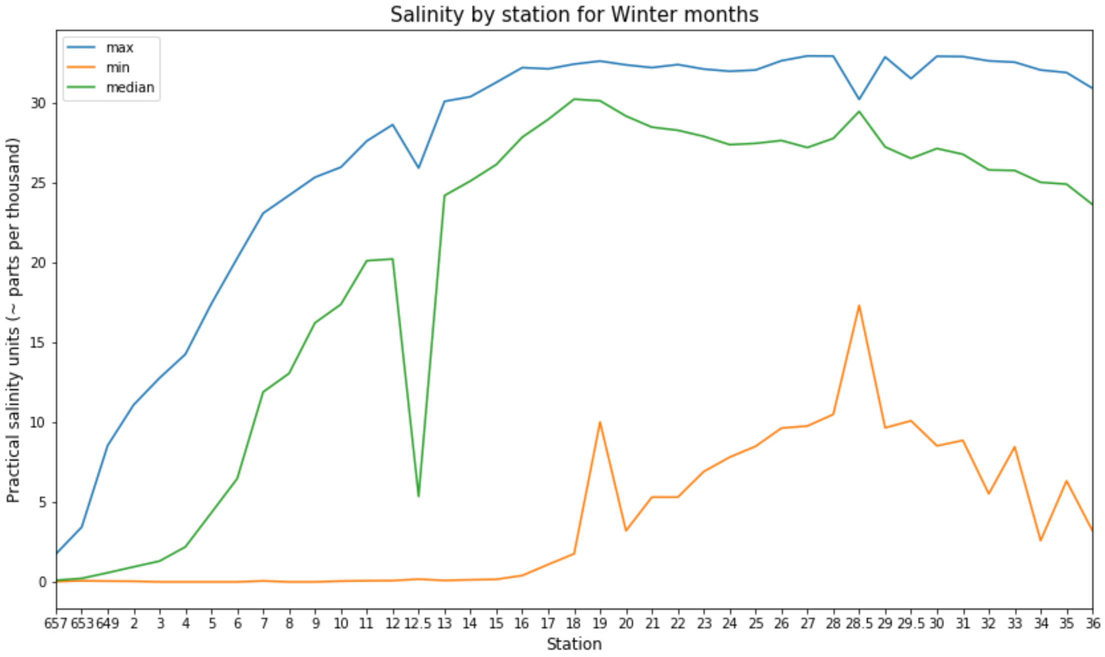


Possible trends (e.g. high, low, increasing, decreasing, anomalies)

I looked 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.

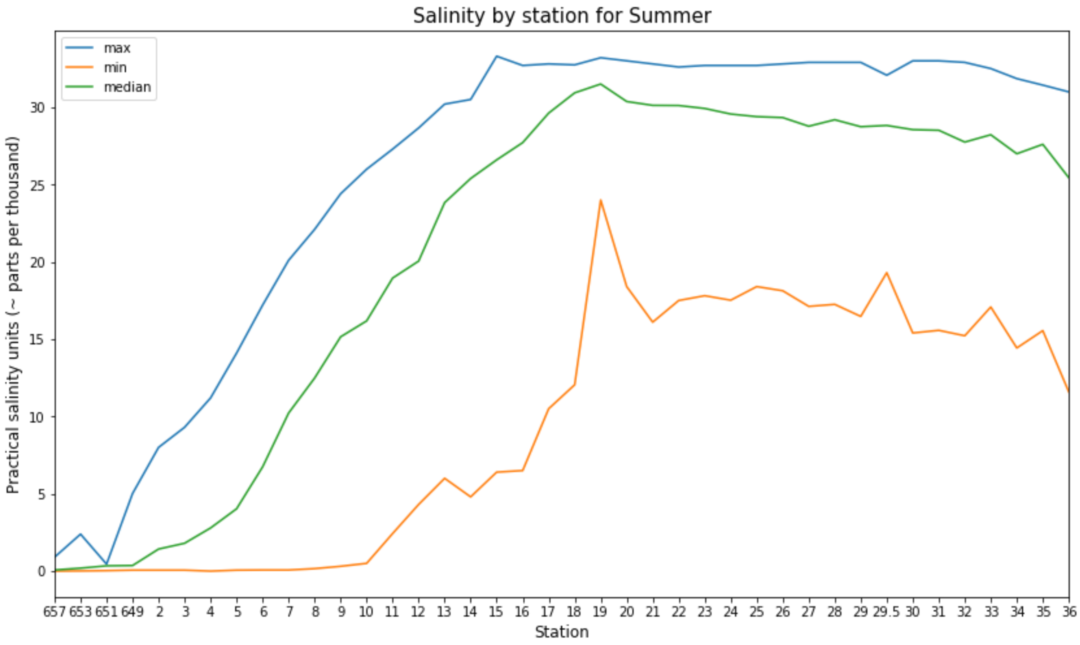
- Salinity graph, winter months:



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 graph, summer months:



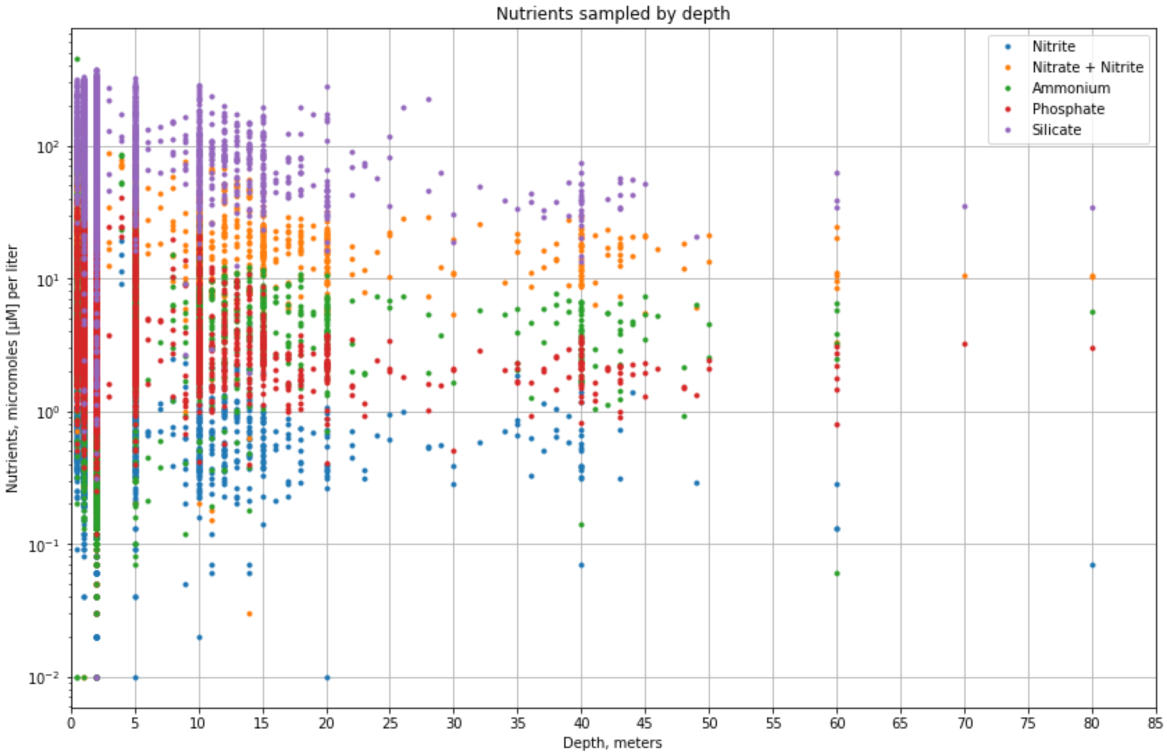
Salinity is, as I would have guessed, lowest in the Sacramento River. It climbs steadily as the water heads through Suisun Bay and San Pablo Bay and peaks in the Central Bay at the Golden Gate. Salinity then remains high going south toward San Jose.

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.

Nutrients by Depth

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

- Nutrients by depth:



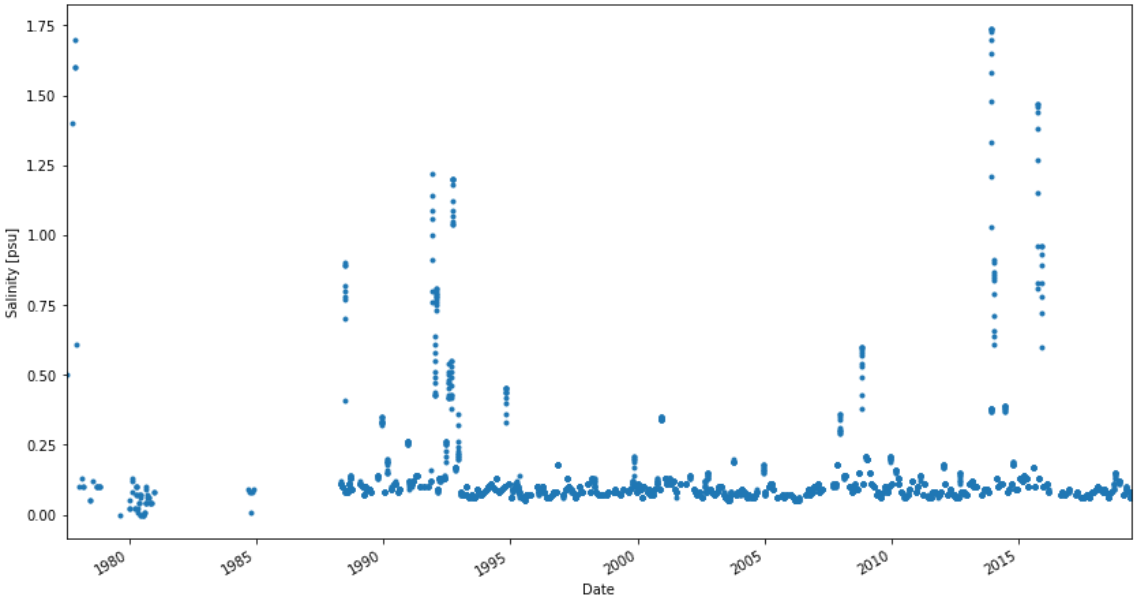
Time-series Plot

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).

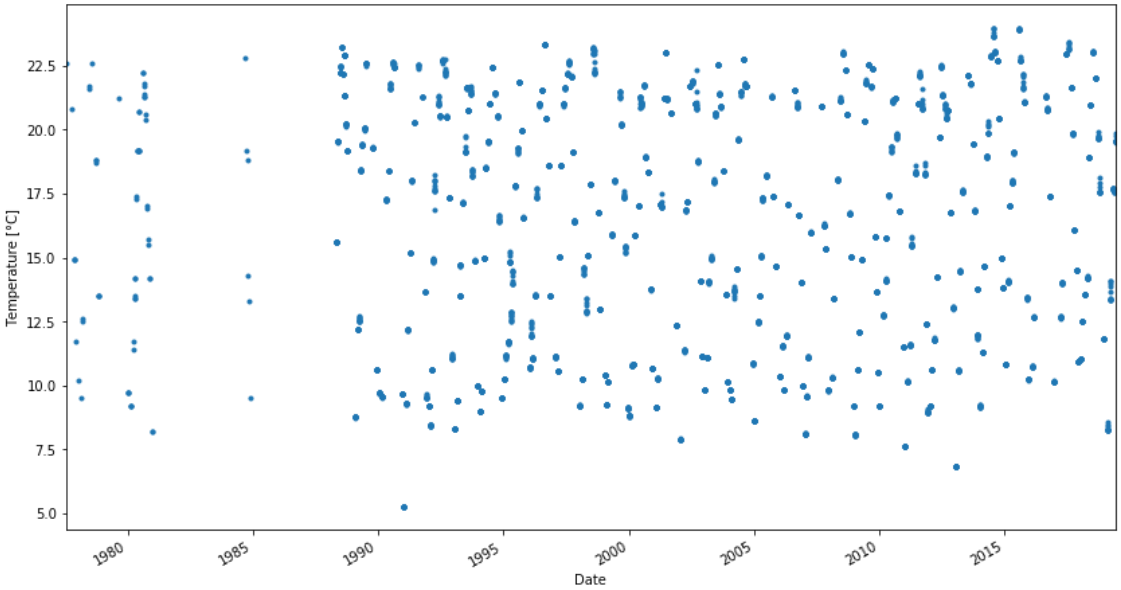
Note that the scale for salinity for station 657 is very different from the scale used for stations 18 and 36. If we used the same (0,30) scale for all three, no changes in salinity would be visible for station 657, where all values are between 0 and 2 ppt.

Station 657

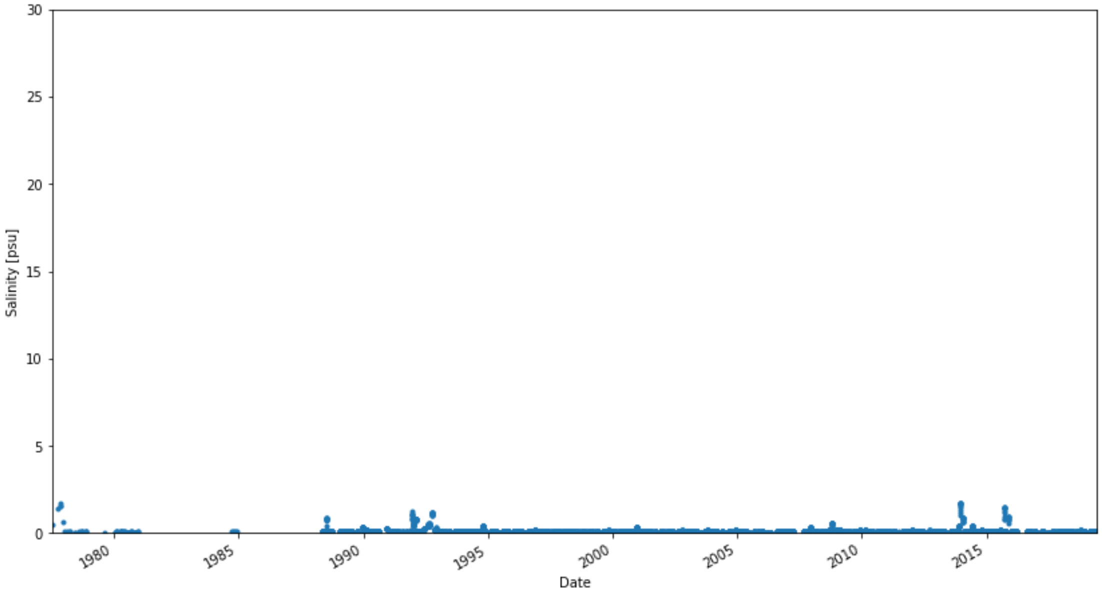
- Station 657 salinity by year:



- Station 657 temperature by year:

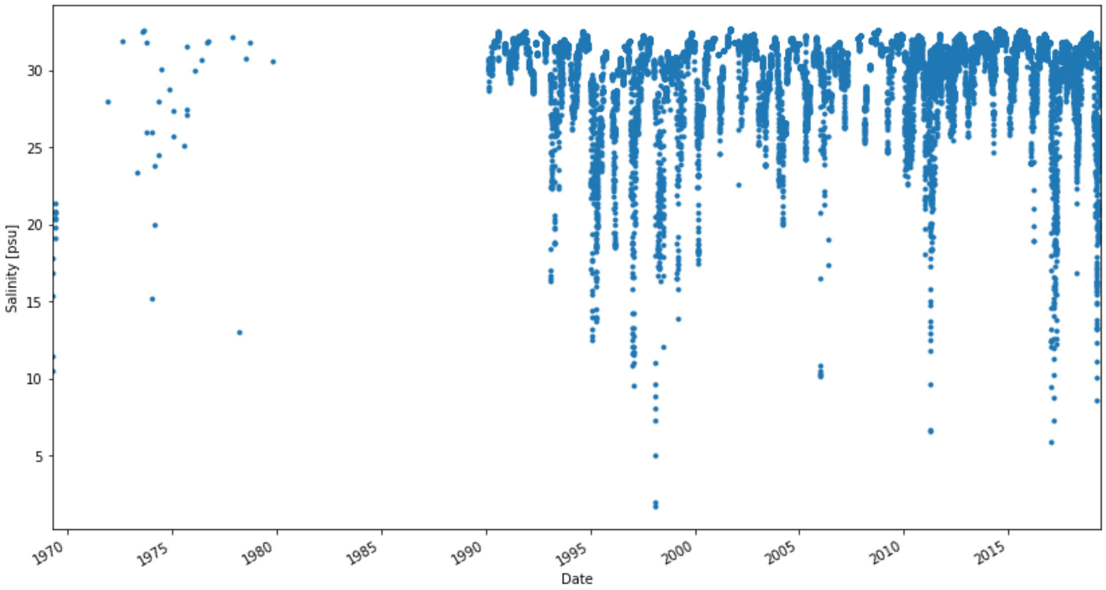


- Station 657 salinity using the same scale used for stations 18 and 36



Station 18

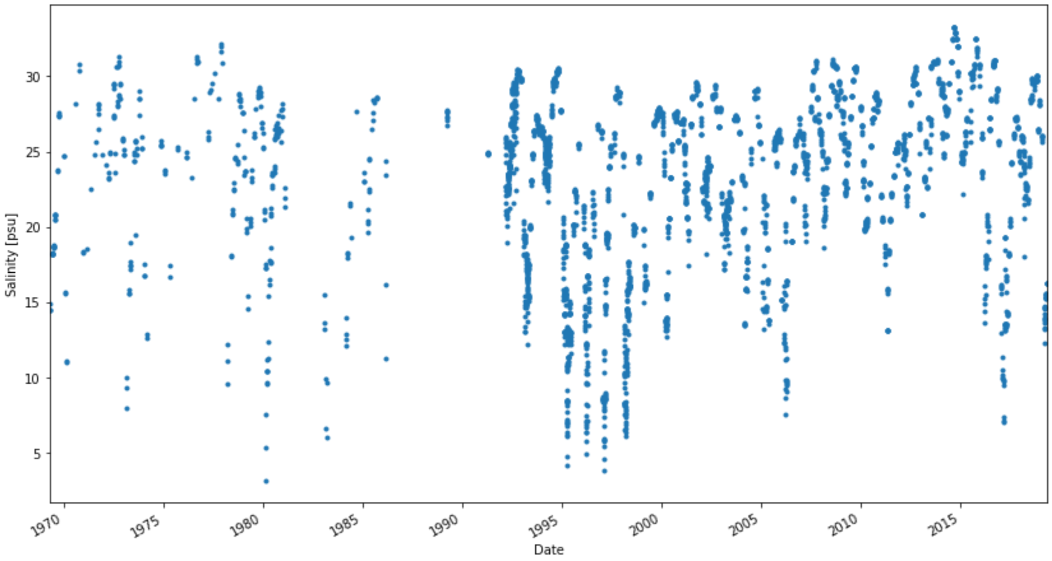
- Station 18 salinity by year:

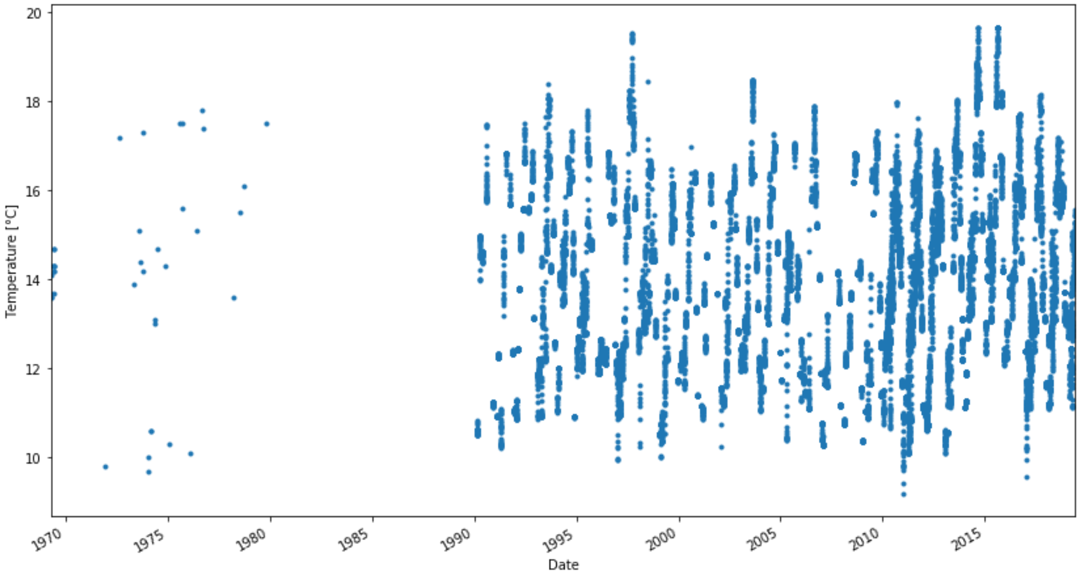


- Station 18 temperature by year:

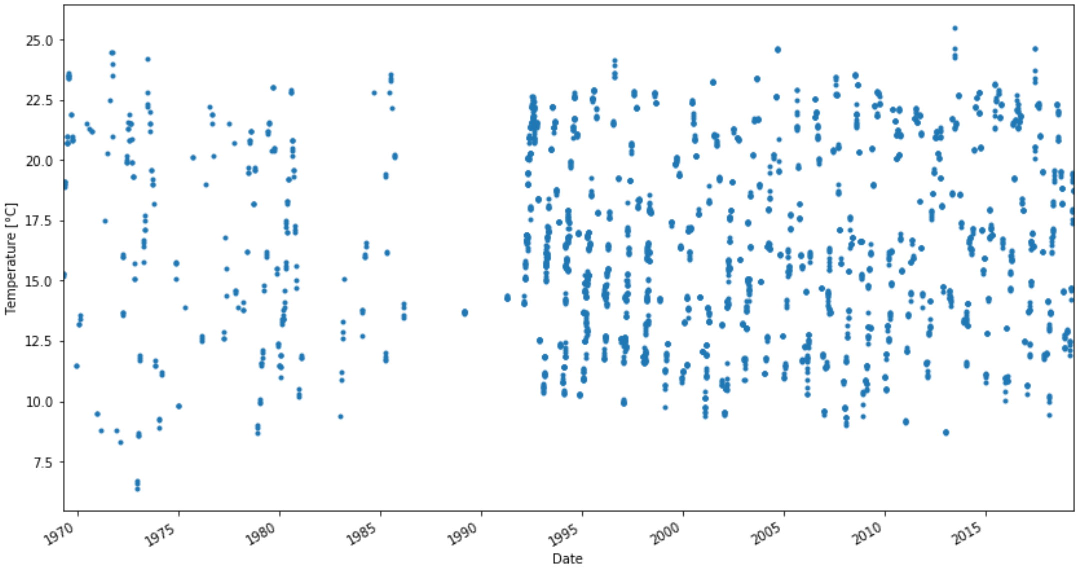
Station 36

- Station 36 salinity by year:





- Station 36 temperature by year:



Comparisons

How does chlorophyll compare to oxygen?

How does biovolume compare to

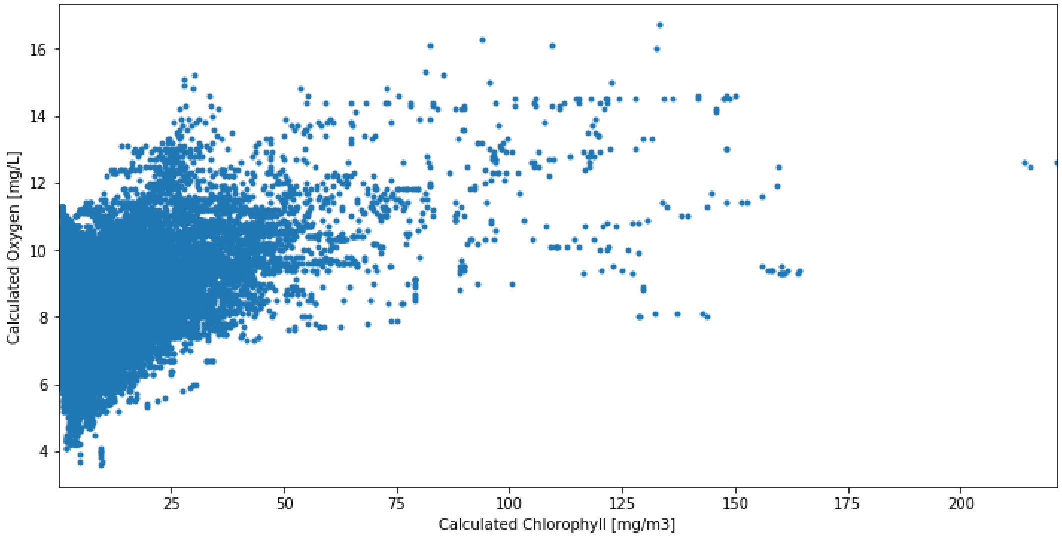
- Chlorophyll
- Oxygen
- Nutrients

Hypothesis:

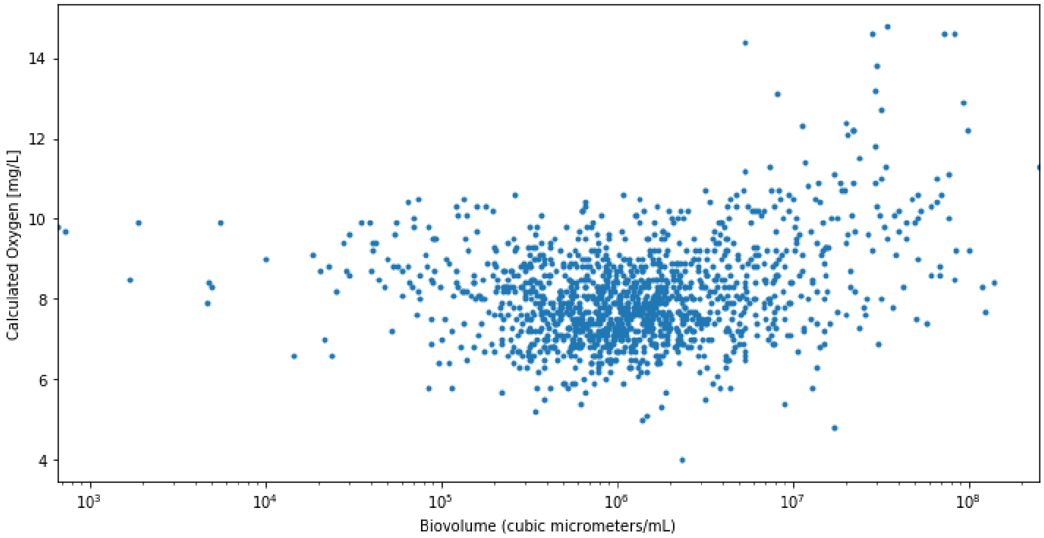
- O2 should increase as chlorophyll increases
- Chlorophyll should increase as biovolume increases
- All nutrients should decrease as biovolume increases (consumed)

Oxygen vs. Chlorophyll and Biovolume

- O2 vs chlorophyll:

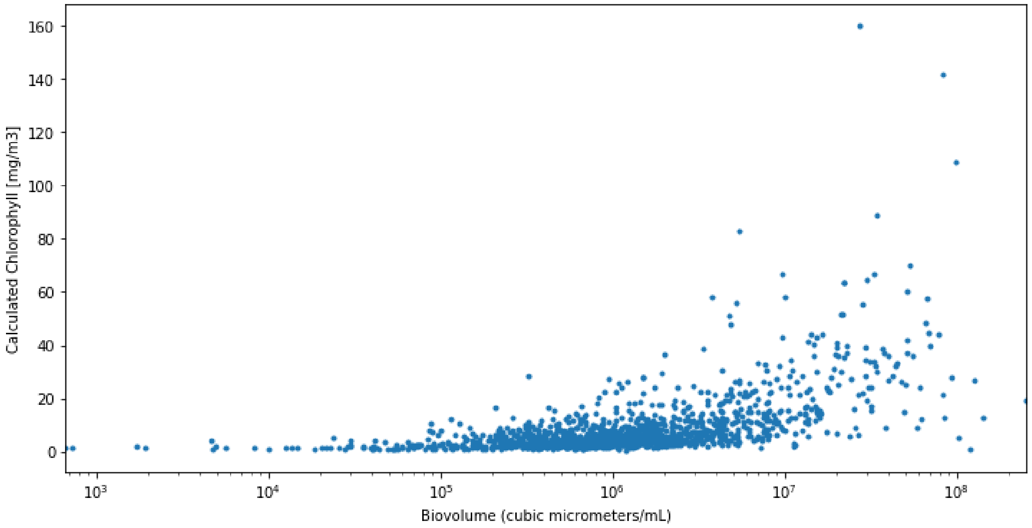


- O2 vs. biovolume:

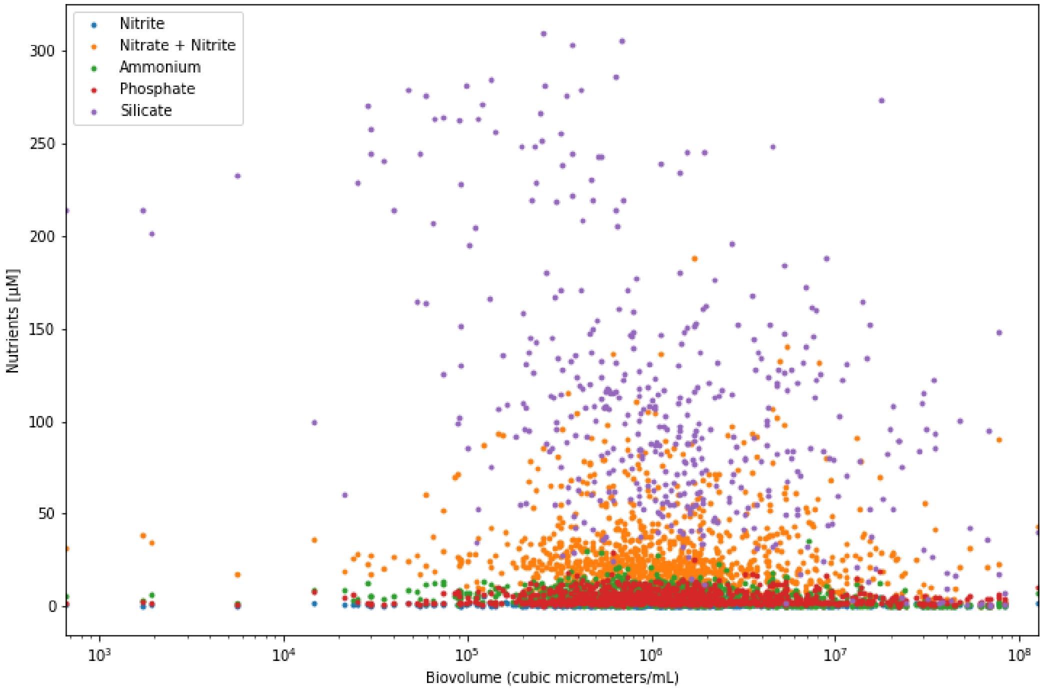


Biovolume vs. Chlorophyll and Nutrients

- 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.

Difference of Means

I know that salinity is very low where the Sacramento River meets the Bay and highest at the Golden Gate where the Bay meets the ocean, but how does mean salinity compare between stations south of the Gate? They're similar; are they the same?

Also, mean temperature appears to be very similar throughout the Bay. Is it statistically the same?

Mean Salinity

Null Hypothesis: Mean salinity is essentially the same for the stations nearest the Golden Gate and to the south.

Alternative Hypothesis: Mean salinity differs between the station groups at the Golden Gate, in San Mateo County, and Santa Clara county.

I calculated mean salinity for station groups 0 - 4.

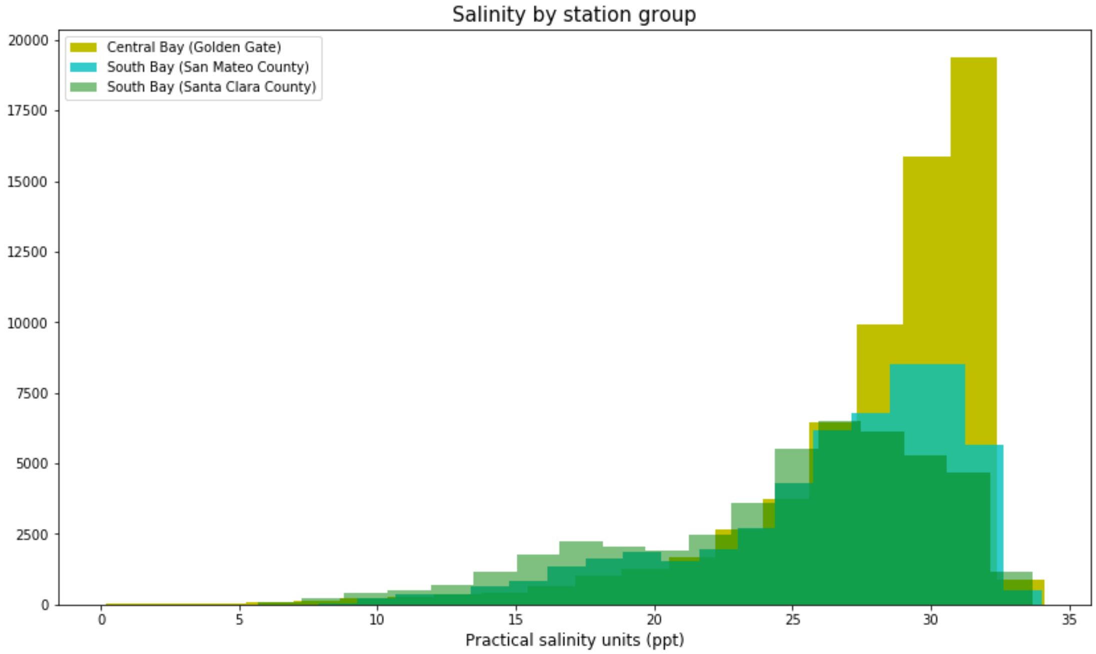
Mean Salinity

Station group	Mean salinity (practical salinity units)
0	24.78
1	26.46
2	28.06
3	17.31
4	5.78

Station groups 3 and 4 are clearly not similar. I investigated groups 0 - 2 further.

First I plotted the salinity for these station groups.

- Salinity plot by station group:



Then I checked normality and variance. The distributions are not normal (there's a tail on one side for each group) and the variances are not homogenous, but I tried a t-est anyway, using `equal_var=False`.

```
T-test result for station group 2, Central Bay (Golden Gate)

   vs group 1, South Bay (San Mateo County)

58.941584001098654

p: 0.0


T-test result for station group 1, South Bay (San Mateo County)

   vs group 0, South Bay (Santa Clara County)

50.57729469105698

p: 0.0
```

Since p < alpha (0.05) in both cases, we can say that the null hypothesis is rejected and suggest that there is a difference in the mean salinity between station groups 0, 1, and 2 - the two South Bay station groups and the Central Bay station group at the Golden Gate.

Mean Temperature

Null Hypothesis: Mean temperature is essentially the same for most station groups.

Alternative Hypothesis: Mean temperature differs between station groups.

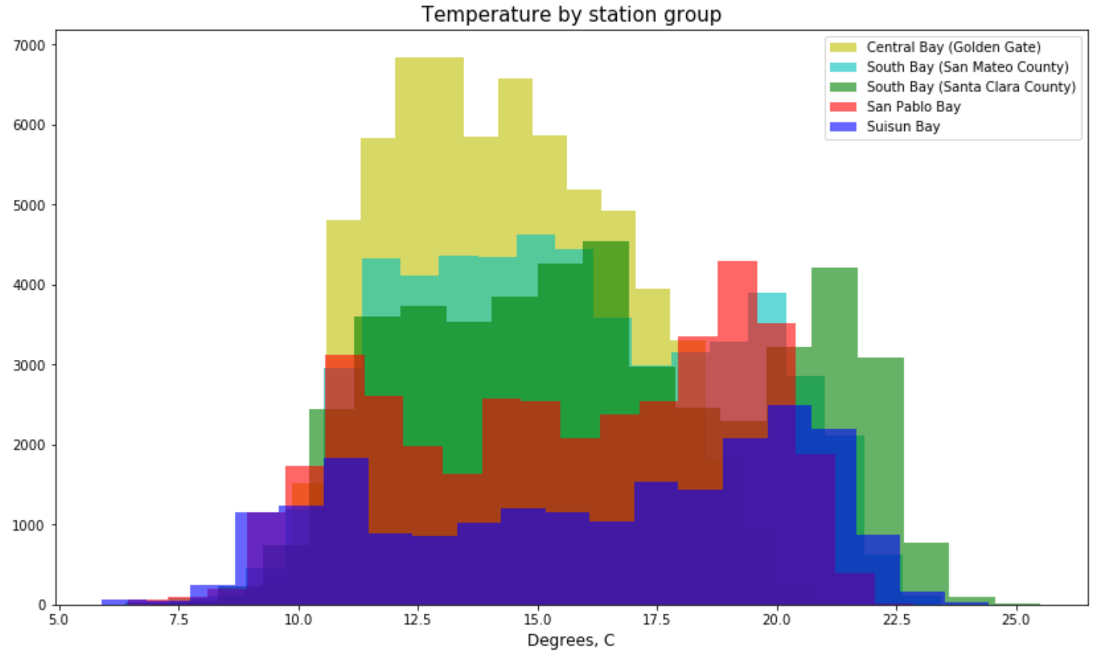
Mean Temperature

Station group	**Mean temperature (degrees C)**

0 South Bay (Santa Clara County)	16.37
1 South Bay (San Mateo County)	15.67
2 Central Bay (Golden Gate)	14.36
3 San Pablo Bay	15.67
4 Suisun Bay	16.2

First I plotted the temperature for these station groups.

- Temperature plot by station group:



They overlap. The means could be the same.

Because five stations appear to have similar means, I decided to try a test that can be used for more than two sample groups.

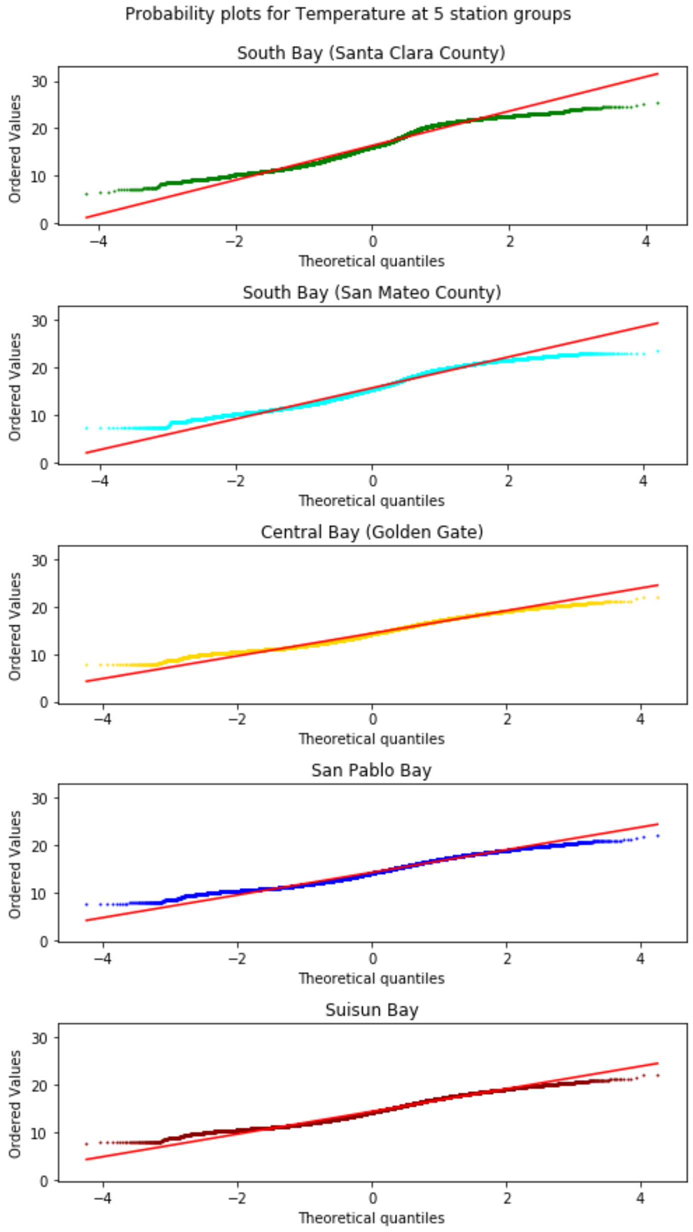
The OneWay ANOVA test can be used to compare the means from three or more groups. This test has important assumptions that must be satisfied in order for the associated p-value to be valid.

- The samples are independent.
- Each sample is from a normally distributed population.
- The population standard deviations of the groups are all equal. This property is known as homoscedasticity.

If these assumptions are not true for a given set of data, it may still be possible to use the Kruskal-Wallis H-test although with some loss of power.
The samples are independent (or as independent as they can be, all coming from the same large body of water).

Next, I used probability plots to see if the samples are normally distributed. They appear to fairly close.

- Probability plots:



Next, I checked the standard deviations. They're not very close.

Standard Deviations

Station group	Standard deviation for temperature
0 South Bay (Santa Clara County)	3.69
1 South Bay (San Mateo County)	3.29
2 Central Bay (Golden Gate)	2.40
3 San Pablo Bay	3.54
4 Suisun Bay	4.19

However, I did read that "if group sizes are equal, the F-statistic is robust to violation of the equal standard deviations requirement", so I checked sample size next.

Sample Size

Station group	Number ofSamples
0 South Bay (Santa Clara County)	46050
1 South Bay (San Mateo County)	53610
2 Central Bay (Golden Gate)	64894
3 San Pablo Bay	38167
4 Suisun Bay	21528

The sizes don't look close enough, so I tried the Kruskal-Wallis H-test.

For station groups 0 - 4, the p-value returned was 0.0. We cannot reject the null hypothesis; there is a significant difference in mean temperature. However, the test does not tell us where the difference is.

I ran the test again, removing station group 2 (lowest mean temperature, 14.36 degrees) from the set. This time, the p-value was *ver slightly* larger than 0.

I ran the test one more time, using only station groups 1 (South Bay (San Mateo County)) and 3 (San Pablo Bay). This time, the p-value returned was 0.0727, indicating that the mean temperature of these two sections of the bay is not statistically different.

Correlations

Next, I turned to looking for correlations between variables.

Biovolume

I tested for a correlation between:

- phytoplankton biovolume and chlorophyll
- chlorophyll and oxygen
- all nutrients and phytoplankton biovolume

I found small positive correlations between phytoplankton biovolume vs. chlorophyll (Pearsons correlation: 0.481) and biovolume vs. oxygen (Pearsons correlation: 0.333).

I was surprised to see that any correlations between biovolume and nutrients were weak, if present. The best, for biovolume vs silicate was (Pearsons) -0.308. Perhaps the quantity of dissolved nutrients in the water is sufficiently large that phytoplankton consumption has only a small effect.

Depth

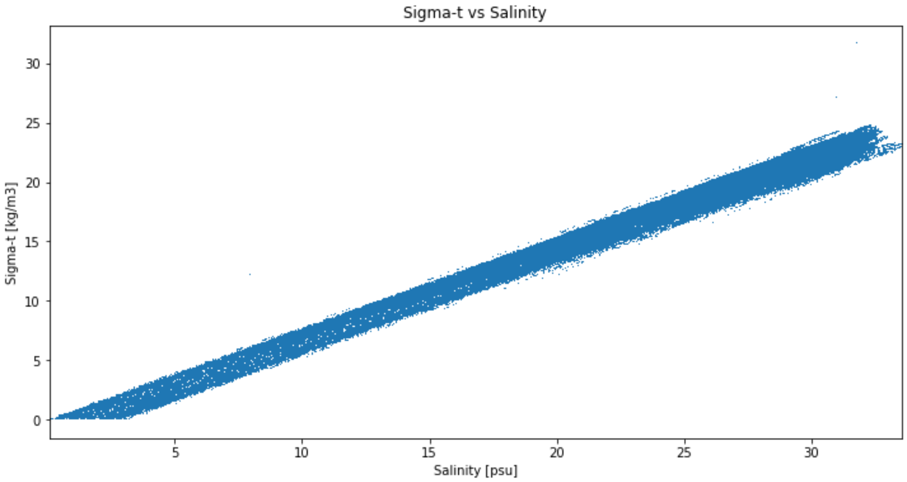
I tested for any correlation between depth and temperature or depth and suspended particulate matter (SPM). I found no correlation.

Sigma-t

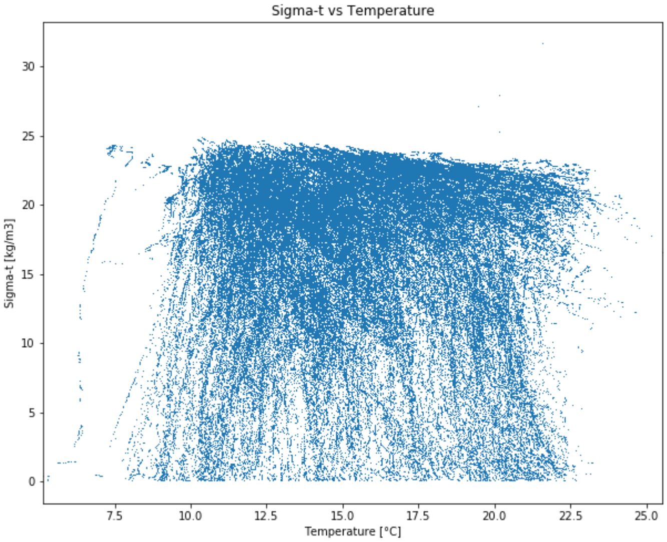
I then tested for a correlation between Sigma-t and salinity or temperature. Sigma-t is defined as "a measure of the density of the water, which is calculated as a function of salinity and temperature. Density increases with increasing salinity and decreasing temperature."

There should be a positive correlation between Sigma-t and salinity. There should be a negative correlation between Sigma-t and temperature. I found a very clear correlation between Sigma-t and salinity, but I did not find any correlation to temperature. Perhaps the San Francisco Bay never gets cold enough to show this.

- Salinity vs Sigma-t:



- Temperature vs Sigma-t:



In Summary

The dataset is easy to work with and the pieces were easy to combine. With forty years of samples and thirty-six stations, tere's a lot of data to examine.

I didn't have to worry much about missing values. In most cases, missing valus at a given depth can be assumed to be approximately the same as measured values at other depths for the same station and date/time.

I did not find the correlations I hoped to find beween phytoplankton biovolume and oxygen, chlorophyll, or nutrients. Perhaps there just isn't enough data for this comparison to be meaningful.

Statistically, I can conclude that:

- Mean salinity varies significantly between sections (groups of stations) across the Bay.
- Mean salinity varies significantly between sections (groups of stations) across the Bay. However, the mean temperatures of San Pablo Bay and the upper end of the South Bay (in San Mateo County) are significantly similar.

Sigma-t is strongly correlated to salinity. Other than this, I found very few significant correlations between features.