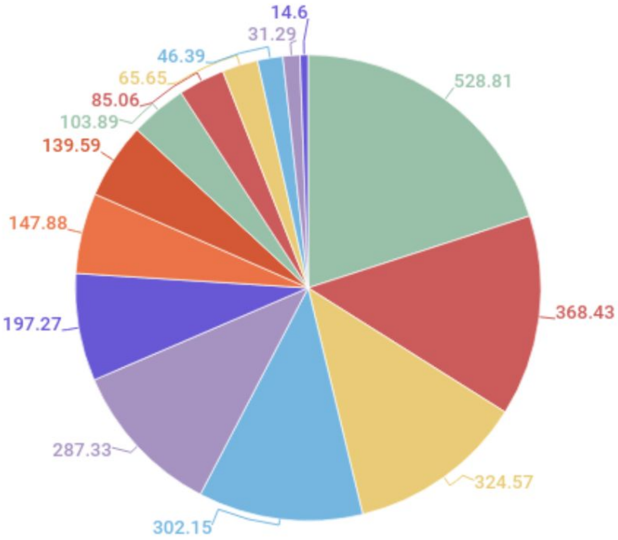
Project Topic

Chemical levels found in aquatic organism tissues in California

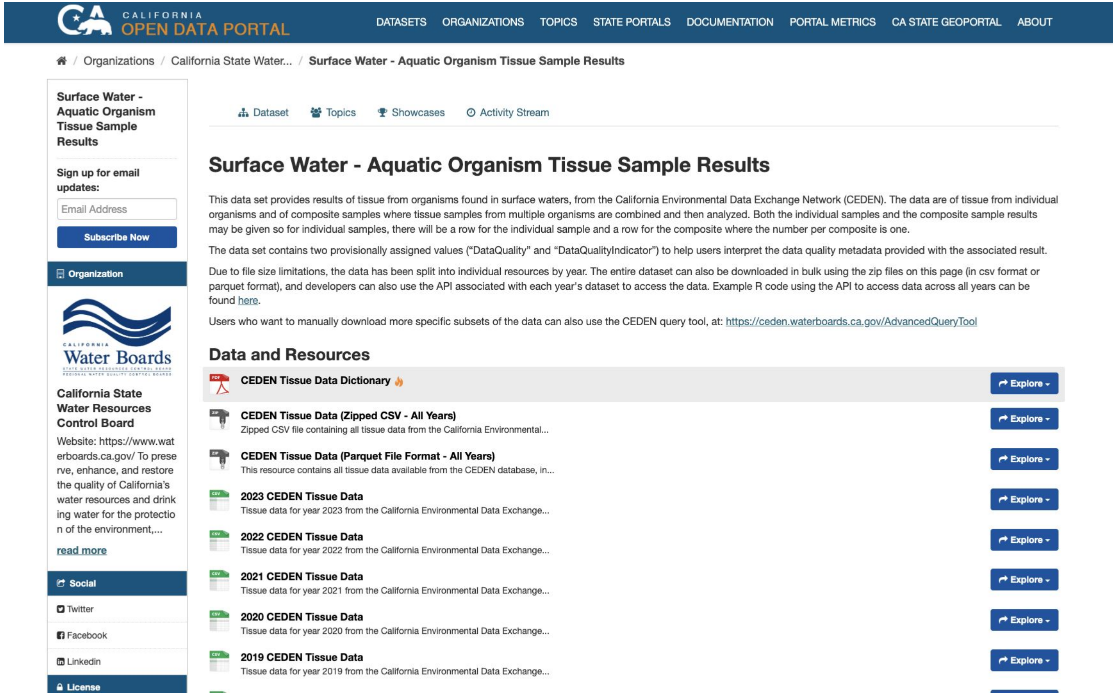
Why This Topic?

- Bioindicators are used to monitor the quality of the environment and how it changes over time
- Chemicals found in aquatic critters is an indicator of surface water quality
- Water sources are scarce in CA, constant drought
- CA is the fifth-largest supplier of food in the world, and produces 8% of America's food supply & 13% of US production value
- Agriculture contributes to 2.5 - 3% of California GDP
- Fishing contributes to 1.5% of California GDP
- Impact on seeing where environmental cleanup efforts need to be concentrated

Light blue indicates agriculture, fishing, hunting sector ratio of California GDP



Where did I get the data?



Surface Water - Aquatic Organism Tissue Sample Results

Sign up for email updates:

Email Address

California State Water Resources Control Board

Website: https://www.waterboards.ca.gov/ To preserve, enhance, and restore the quality of California's water resources and drinking water for the protection of the environment,...

read more

Surface Water - Aquatic Organism Tissue Sample Results

This data set provides results of tissue from organisms found in surface waters, from the California Environmental Data Exchange Network (CEDEN). The data are of tissue from individual organisms and of composite samples where tissue samples from multiple organisms are combined and then analyzed. Both the individual samples and the composite sample results may be given so for individual samples, there will be a row for the individual sample and a row for the composite where the number per composite is one.

The data set contains two provisionally assigned values ("DataQuality" and "DataQualityIndicator") to help users interpret the data quality metadata provided with the associated result.

Due to file size limitations, the data has been split into individual resources by year. The entire dataset can also be downloaded in bulk using the zip files on this page (in csv format or parquet format), and developers can also use the API associated with each year's dataset to access the data. Example R code using the API to access data across all years can be found here.

Users who want to manually download more specific subsets of the data can also use the CEDEN query tool, at: https://ceden.waterboards.ca.gov/AdvancedQueryTool

Data and Resources

CEDEN Tissue Data Dictionary

CEDEN Tissue Data (Zipped CSV - All Years)
Zipped CSV file containing all tissue data from the California Environmental...

CEDEN Tissue Data (Parquet File Format - All Years)
This resource contains all tissue data available from the CEDEN database, in...

2023 CEDEN Tissue Data
Tissue data for year 2023 from the California Environmental Data Exchange...

2022 CEDEN Tissue Data
Tissue data for year 2022 from the California Environmental Data Exchange...

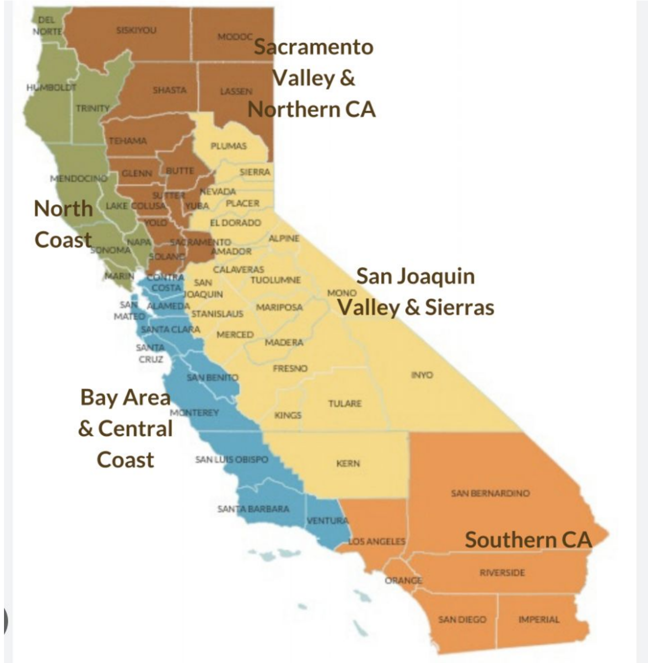
2021 CEDEN Tissue Data
Tissue data for year 2021 from the California Environmental Data Exchange...

2020 CEDEN Tissue Data
Tissue data for year 2020 from the California Environmental Data Exchange...

2019 CEDEN Tissue Data
Tissue data for year 2019 from the California Environmental Data Exchange...

https://data.ca.gov/dataset/surface-water-aquatic-organism-tissue-sample-results

Dataset - Overview



Master Dataset

- 216,797 rows
- 117 columns

20% Random Sample

- 43,349 rows
- 117 columns

Key Information:

Project Name, Species (Common Name), Chemicals (Analytes), Results of Analytes, Locations (StationName, Lat, Long), Sample specie size (length & mass), Sex, Sample Collection Date, Tissue Name

Preliminary Model Idea: Can we predict which location/region we are most likely to encounter a particular species containing elevated amounts of chemicals

Challenges

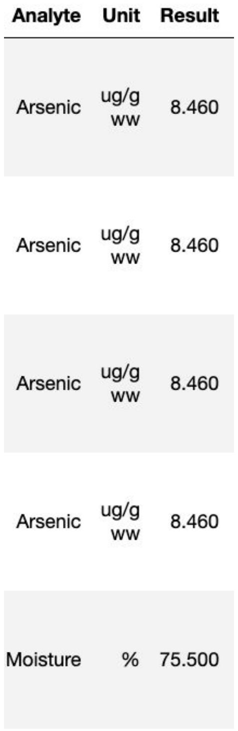
Null Values - 787,414

Over abundance of columns - 117 is a lot

Results in single column with different unit values (i.e. cm vs mm)

All analytes names under one column, and all results under another

Kernel kept crashing when trying to filter rows or concatenate



Analyte	Unit	Result
Arsenic	ug/g ww	8.460
Arsenic	ug/g ww	8.460
Arsenic	ug/g ww	8.460
Arsenic	ug/g ww	8.460
Moisture	%	75.500

Cleaning Approach

Look at column values, drop columns containing duplicate values & irrelevant informations

Pivot analytes and results - each analytes = own column w/ relevant results

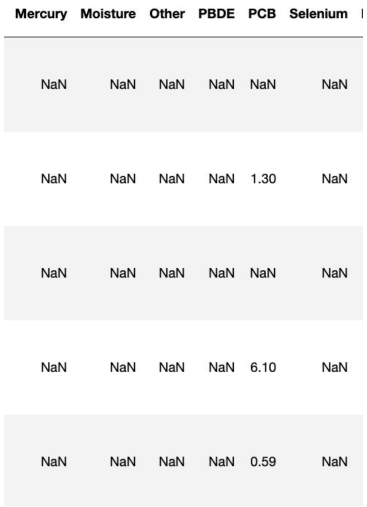
Drop duplicate rows

Remove columns w/ all nulls

Analyze numeric & categorical column nulls

- For analytes nulls, fill in w/ 0. 0 = not detected or not recorded
- Unit measures - need to analyze mean & median of each species

Group analytes, species, locations



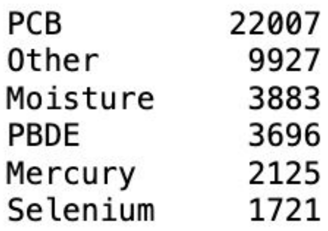
Mercury	Moisture	Other	PBDE	PCB	Selenium
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	1.30	NaN
NaN	NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	6.10	NaN
NaN	NaN	NaN	NaN	0.59	NaN

Analytes

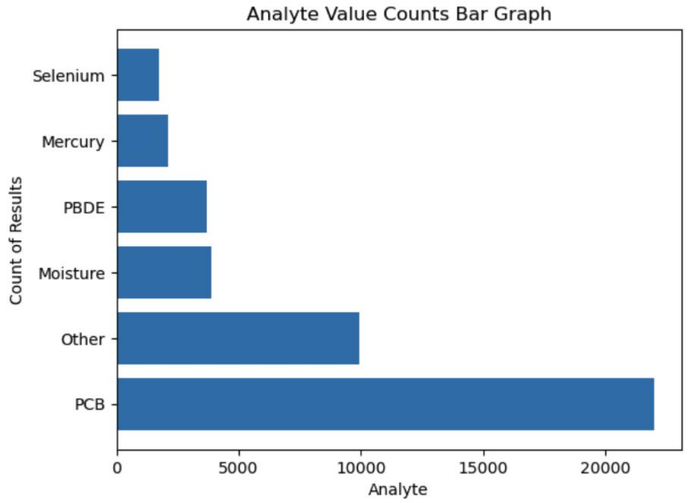
Before grouping:

644 unique values

After grouping:



```
PCB          22007
Other         9927
Moisture      3883
PBDE          3696
Mercury       2125
Selenium      1721
```



Mercury - heavy metal, bioaccumulates. Highly toxic. Can end up in our food supply chain, like tuna

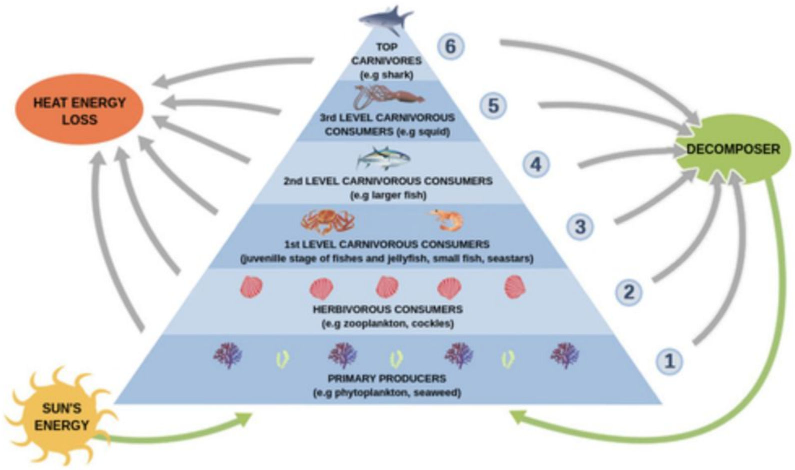
PCBs - chemical known that were used widely in building materials, paints and sealants, causes cancer

PBDEs- a group of chemicals used as flame retardants in products like electronics, furniture, and textiles, causes neuro issues

Selenium -an essential mineral that plays a key role in many bodily functions, found to protect against mercury exposures

Other - 409 chemicals, each has less than 400 samples

Species


TOP CARNIVORES (e.g shark)
6
3rd LEVEL CARNIVOROUS CONSUMERS (e.g squid)
5
2nd LEVEL CARNIVOROUS CONSUMERS (e.g larger fish)
4
1st LEVEL CARNIVOROUS CONSUMERS (juvenille stage of fishes and jellyfish, small fish, seastars)
3
HERBIVOROUS CONSUMERS (e.g zooplankton, cockles)
2
PRIMARY PRODUCERS (e.g phytoplankton, seaweed)
1
HEAT ENERGY LOSS
DECOMPOSER
SUN'S ENERGY

95 species

- From clams to sharks

Grouping approach brainstorming:

- Taxonomy
  - Channel Catfish + Flathead Catfish + White Catfish
  - Largemouth Bass + Sand Bass + Smallmouth Bass
  - Rockfish + Copper Rockfish + Yellowtail Rockfish
  - Striped Surfperch + White Surfperch + Walleye Surfperch
- Trophic levels
  - Producers
  - Herbivorous Consumers
  - 1st Level Carnivores Consumers
  - 2nd Level Carnivores Consumers
  - 3rd Level Carnivores Consumers
  - Top Carnivores

Next Steps

Finish grouping species over the weekend & calculate length/mass mean/median to fill in nulls

May need to categorize locations with different environments or regions. I.e. lakes, streams, coastal etc.

Do EDA & pre-modeling analysis to explore questions such as:

- Is there a correlation between chemical levels & the species groups?
- How does the PCB observation counts skew the data?
- Are there certain locations that shows samples with elevated levels of chemicals?
- Are there certain locations that has more species samples?
- Is there a correlation between each chemical type?
- What are the confounding factors, multicollinearity?