INVESTIGATING CALIFORNIA WATER QUALITY

Angelica Victoria, Antonette Goroch, Brittni Bresee, Lorela Cabral, Adam Paganini

Motivation & Target Audience

- Clean water is a priority for CA.
- Public officials need good data to make sound decisions on behalf of their constituents.

How does this analysis contribute?

- Considers 3 substances known for negative public health effects which can be present in water: Mercury, Arsenic, Nitrates
- Negative effects can be mitigated with policy action if substances effectively identified.

RESEARCH QUESTIONS

1. What is the data picture of a single collection site over the entire historical period of collection?
2. What are the difference in surface water samples vs. well water samples?
3. What are the historic trends for the presence of Arsenic & Mercury among CA highested populated counties (LA, San Diego, Alameda, Sacramento)?
4. What are the recent trends for Nitrate among CA most populated counties?
5. What are the recent trends for Mercury among CA most populated counties?

THE DATA

Show 10 entries

Hide/Unhide Columns | Download

Showing 1 to 10 of 4,523,028 entries

Search:

county_name	sample_code	sample_date	sample_depth	sample_depth_units	parameter	result	reporting_limit
Alameda	WDIS_0719152	05/03/1967 09:00	None	Feet	Conductance	3480.00000000	1.00000000
Alameda	WDIS_0719152	05/03/1967 09:00	None	Feet	Dissolved Boron	7.70000000	0.10000000
Alameda	WDIS_0719152	05/03/1967 09:00	None	Feet	Dissolved Calcium	68.00000000	1.00000000
Alameda	WDIS_0719152	05/03/1967 09:00	None	Feet	Dissolved Chloride	758.00000000	0.10000000
Alameda	WDIS_0719152	05/03/1967 09:00	None	Feet	Dissolved Magnesium	59.00000000	0.10000000
Alameda	WDIS_0719152	05/03/1967	None	Feet	Dissolved	510.00000000	1.00000000

The Government Operations Agency sponsors data.ca.gov, a statewide open data portal run by the state of CA.

This data houses a collection of numerous state collected datasets about the essential public infrastructure like water, housing, etc.

DATA PIPELINE

CLEANING THE DATA

- Converting units
- Creating dfs by county
- Sorting by area, time, and population

```
mer_county_df = mer_data.loc[(ar_data["county_name"] == "Los Angeles") |
                             (ar_data["county_name"] == "San Diego") |
                             (ar_data["county_name"] == "Alameda") |
                             (ar_data["county_name"] == "Sacramento") ]
mer_county_df.head()
```

```
merLA_df = mer_county_df.loc[mer_county_df["county_name"] == "Los Angeles"]
merLA_avg = merLA_df.groupby("sample_date").mean()["result"]
```

STANDARDIZING AND FORMATTING TIME

- Creating one format for time
- Cleaning each year with a for loop

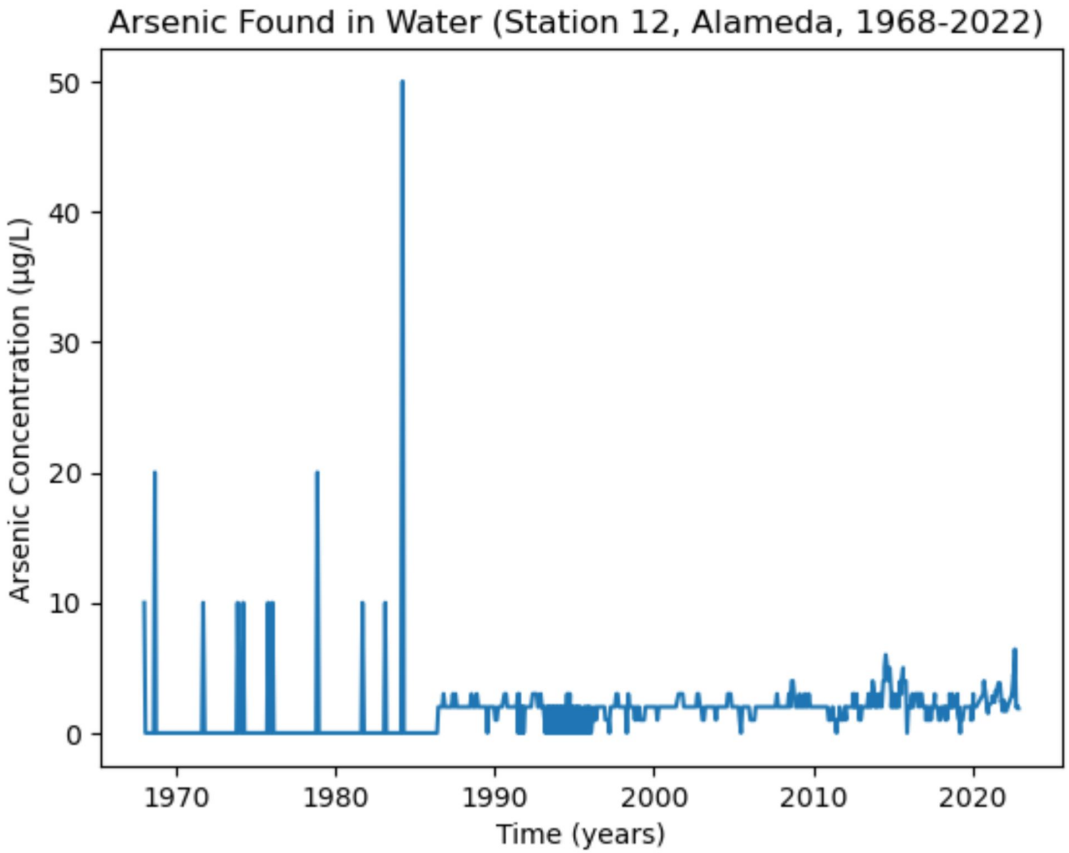
```
#Standardizing sample date in dataframe

for idx, row in ar_data.iterrows():
    year=row['sample_date'].split(' ')[0].split('/')[-1]
    if int(year)>22:
        date=row['sample_date'].split(' ')[0]
        time=row['sample_date'].split(' ')[-1]
        new_date=date.split('/')[:-1]
        new_date=new_date+[f'19{year}']
        # print('/'.join(new_date))
        new_date_str='/'.join(new_date)
    else:
        date=row['sample_date'].split(' ')[0]
        time=row['sample_date'].split(' ')[-1]
        new_date=date.split('/')[:-1]
        new_date=new_date+[f'20{year}']
        # print('/'.join(new_date))
        new_date_str='/'.join(new_date)
    ar_data.loc[idx, 'sample_date']=new_date_str+' '+time
```

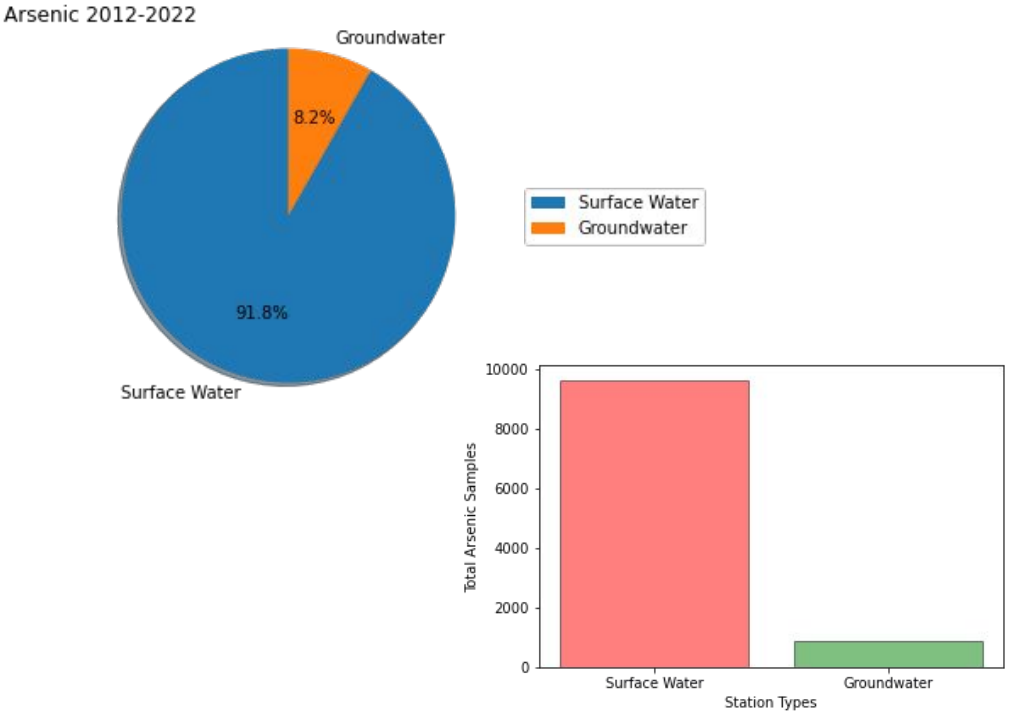
```
ar_data['sample_date'].dt.year
```

Software Used: Pandas, matplotlib, numpy, json, time, request

HISTORICAL TREND OF ARSENIC



Arsenic Samples: Surface Water vs. Ground Water



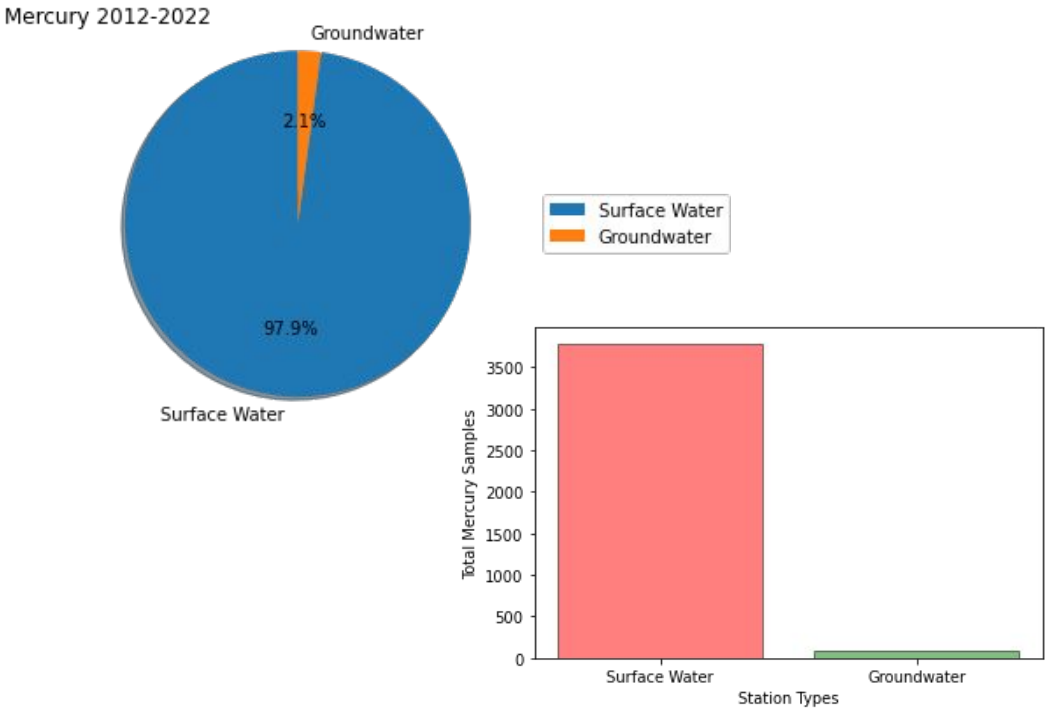
Further testing needed.

More groundwater testing targeted at sensitive areas.

Further analysis needed.

Why was ground water tested?
Where was ground water tested?
How were these tests different?

Mercury Samples: Surface Water vs. Ground Water



Further testing needed.

More groundwater testing targeted at sensitive areas.

Further analysis needed.

Why was groundwater tested?
Where was groundwater tested?
How were these tests different?

HISTORIC ARSENIC DATA BY COUNTY

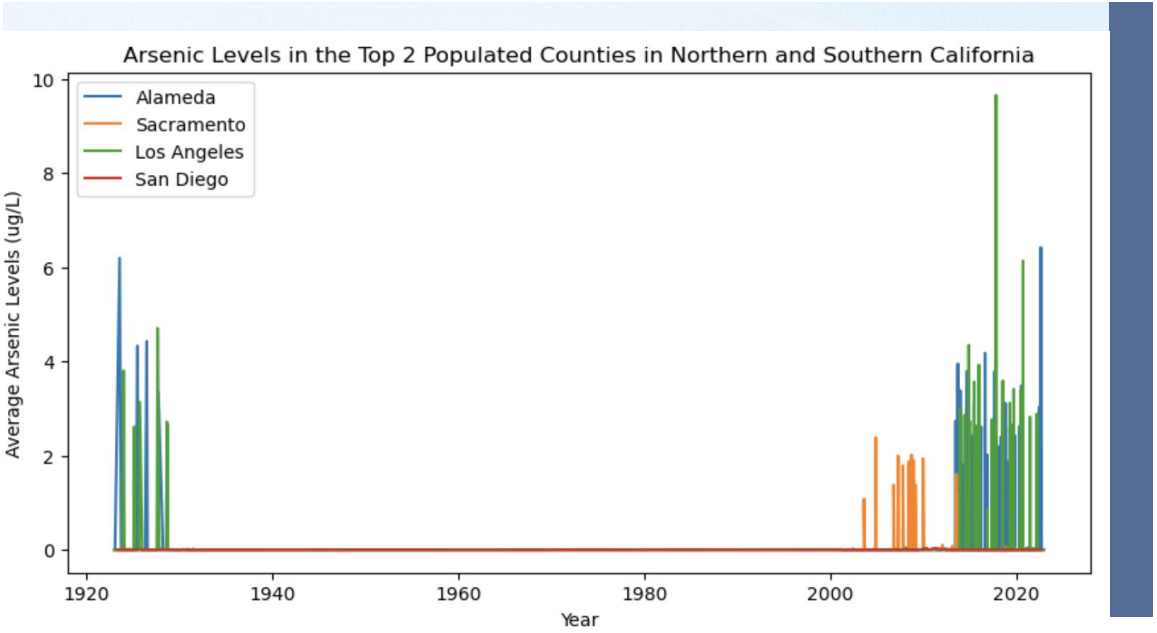


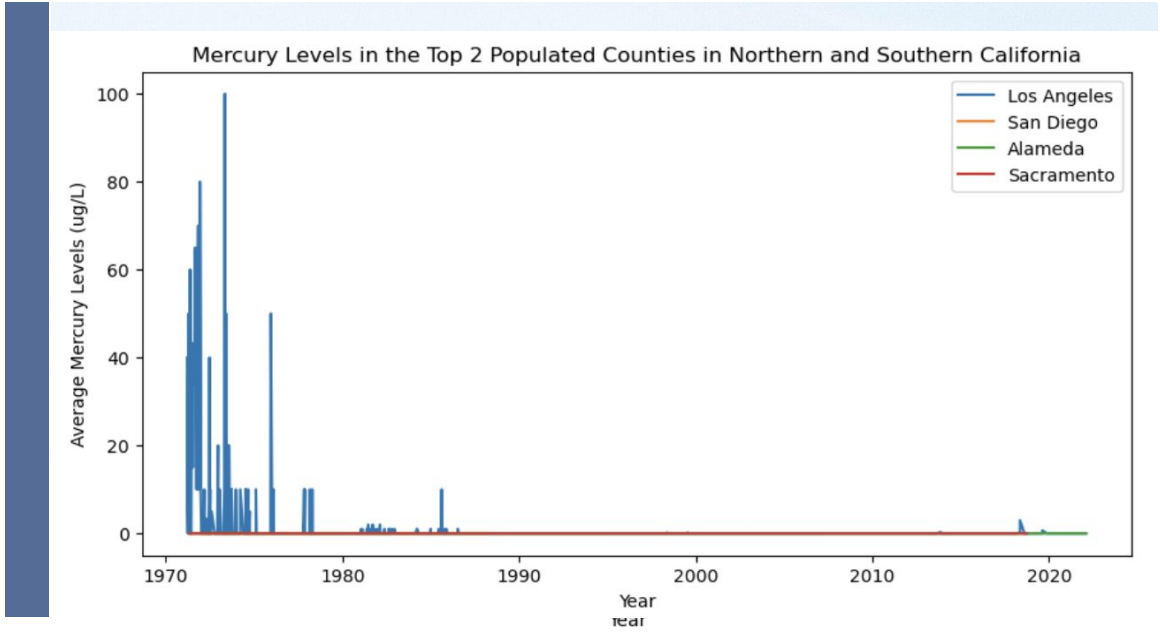
Figure 6

- No relationship between northern/southern CA arsenic levels
- Peaks in 1920's and resurgence in the 2000's
- Consider the scale:
  - CA only allows 10 ug/L of arsenic

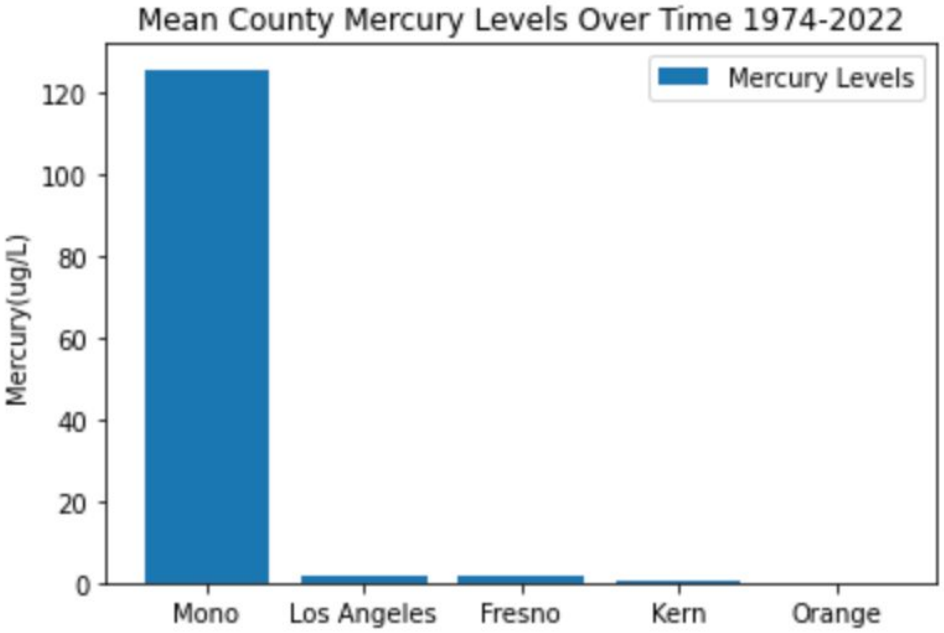
HISTORIC MERCURY DATA BY COUNTY

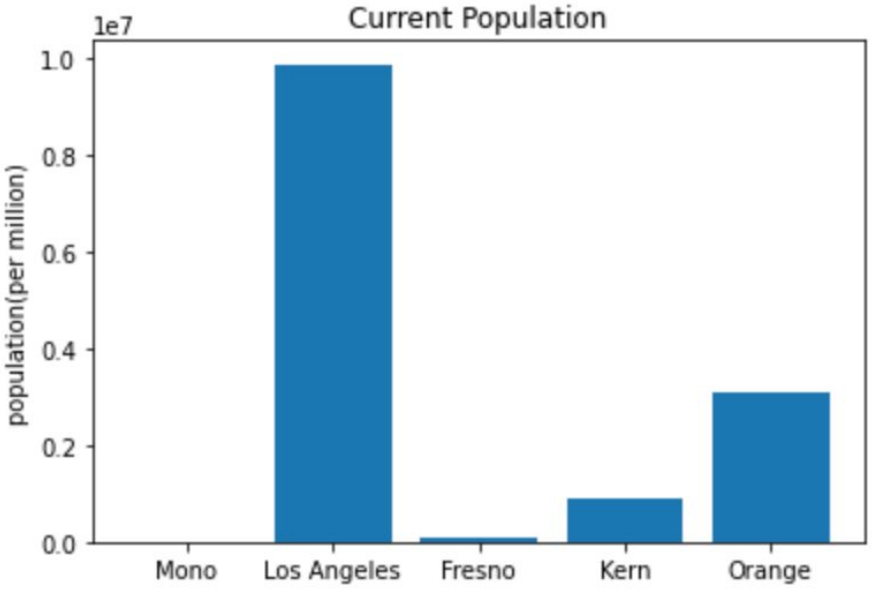
Figure 7

- Highest peaks in LA in the 1970's
- Mitigation after 1990's
- Consider the scale:
  - CA only allows 2 ug/L of mercury

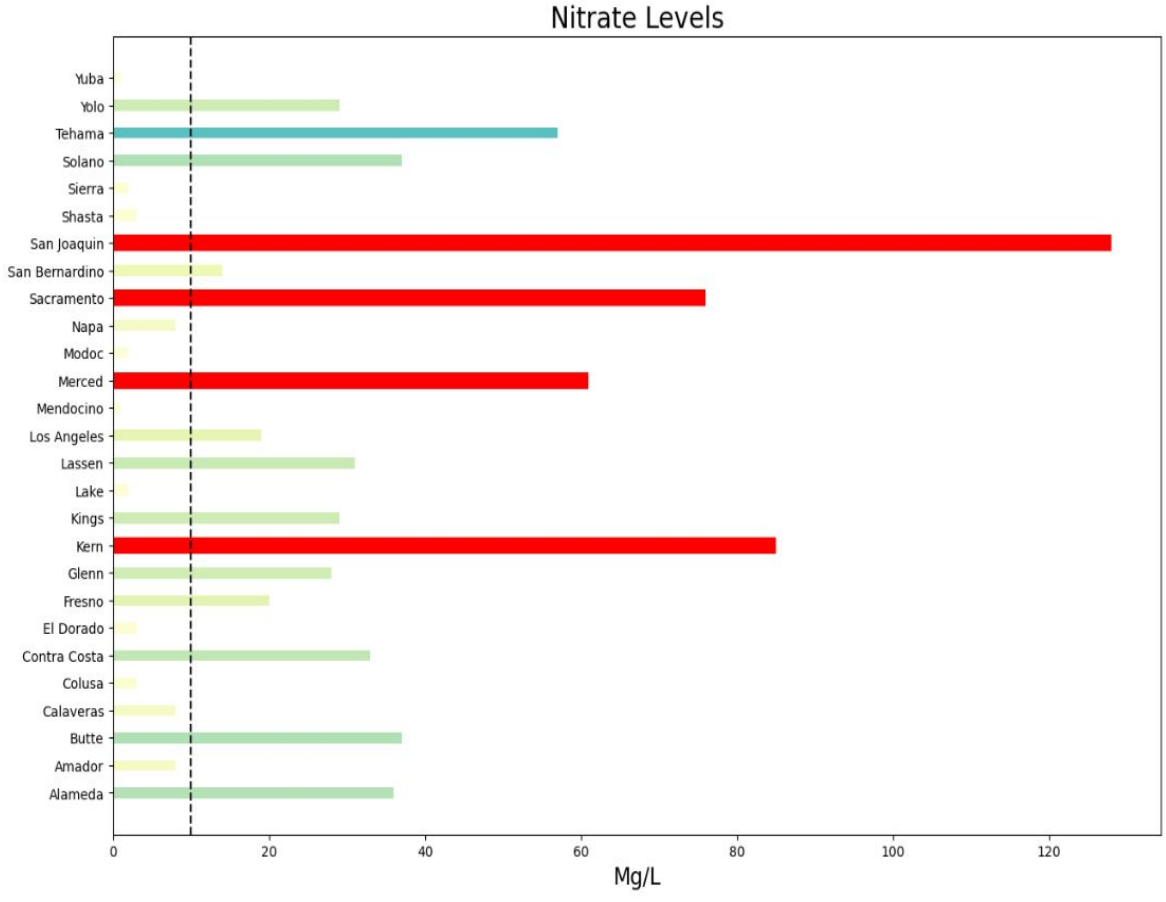


Mercury Levels Over Time





Nitrate Levels Per County Ca.



2012 - 2022

- **San Joaquin**
  population 789,410
- **Sacramento**
  population 533,321
- **Kern**
  population 917,673
- **Merced**
  population 89,308

IN SUMMARY

- Higher Arsenic & Mercury pre vs. post 1980's regulation
- Legal thresholds of Ar & Mer are violated
- Surface water prioritized in sampling
- No difference between NorCal / SoCal Arsenic levels
- Mercury concentrations high in LA, but have settled
- Nitrate levels are persistently violating legal limit, both past and present

CHALLENGES

Data Cleaning

- Formatting time
- Outliers

Inclusion of More Data

- Limits of statistical analysis

NEXT STEPS

- Industrial areas vs. farmland & population
- Water treatment plant procedures
- Toxic water levels in our home
- Reducing toxic levels in water