Progress-report

April 4, 2025

1 Progress Report

1.1 Project Introduction

This project analyzes cancer trends in California with a focus on the potential impact of water quality, specifically PFAS contamination, on public health. It investigates whether higher PFAS levels correlate with increased cancer rates, especially in low-income communities.

The analysis uses three main data sources: 1. California Cancer Dataset (2017–2021): Over 4,500 records of county-level cancer rates by type, ethnicity, gender, and age-adjusted values. 2. California PFAS Dataset (2014–2016): More than 680,000 groundwater sampling records tracking PFAS contamination and water quality. 3. 2022 County Shapefiles: Geographic boundaries to link cancer and PFAS data to specific counties.

The goal is to identify whether environmental water contamination is associated with higher cancer incidence across California.

Notebook: https://github.com/uic-cs418/group-project-data-engineers/blob/main/Progress-report.ipynb

1.2 Changes to Scope:

The scope of the project has narrowed slightly, with a shift away from exploring all water contaminants in favor of a more focused investigation into PFAS and its direct correlation with cancer rates.

This refined approach aims to provide actionable insights into the link between water quality and public health, particularly cancer incidence, and offer policy recommendations based on the findings.

[ ]:
```
! pip install pandas matplotlib seaborn geopandas
```

[2]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
```

1.3 Data Cleaning

1.3.1 Cancer Dataset

The cancer dataset includes county-level statistics by sex and cancer type, with key fields such as AAIR, total population, and demographic percentages. It spans all California counties over

two periods: 2011–2021 (10-year) and 2017–2021 (5-year). Data is sourced from California Health Maps, ensuring faithfulness and standardized metrics. For cleaning, only 2017–2021 records were kept to align with PFAS data, and rows missing AAIR were dropped. "Both sexes" entries were removed to avoid redundancy. Where rows aggregated multiple counties, values were redistributed proportionally based on population and sex. Cancer incidents were calculated using AAIR and population, and AAIR was recomputed where needed. These steps ensured consistent granularity, scope, and structure for analysis.

[10]:
```
cancer_df = pd.read_csv("data/raw/californiahealthmaps_county_all.csv")
cancer_df.head(1)
```

[10]:
```
   AreaID Counties                                                Cities  \
0       1  Alameda  Livermore, Dublin, Berkeley, Hayward, Alameda,…

                                                 URL   Sex    Cancer Years  \
0  https://www.californiahealthmaps.org?areatype=…  Both  Prostate  05yr

    PopTot  AAIR   LCI  …  PerWhite  PerBlack  PerAsian  PerHispanic  \
0  8360006  42.4  41.2  …      32.3      11.0      34.1         22.2

   PerDocVisit  PerFOBT  PerMammo  PerMenPrev  PerCervical  PerWomenPrev
0         61.1     61.0      71.6        38.2         79.0          30.9

[1 rows x 42 columns]
```

[7]:
```
cancer_df.isna().sum()[:11]
```

[7]:
```
AreaID        0
Counties      0
Cities      702
URL           0
Sex           0
Cancer        0
Years         0
PopTot        0
AAIR        299
LCI         299
UCI         299
dtype: int64
```

[11]:
```
from cancer_data_cleaning import clean_cancer_data, split_combined_counties

cancer_df = clean_cancer_data(cancer_df)
cancer_df.head(3)
```

[11]:
```
   Counties     Sex    Cancer   PopTot  AAIR  Cancer_Incidents
1  Alameda    Male  Prostate  4150512  91.9       3814.320528
2  Alameda  Female  Prostate  4209494   0.0          0.000000
```

```
4  Alameda    Male    Breast  4150512   1.2         49.806144
```

[8]:
```
combined_counties = cancer_df['Counties'][cancer_df['Counties'].str.
↪contains(',')].unique()
combined_counties
```

[8]:
```
array(['Tulare, Fresno', 'Shasta, Tehama', 'Ventura, Kern'], dtype=object)
```

[12]:
```
cancer_df = split_combined_counties(cancer_df)
cancer_df[cancer_df['county'] == 'Tulare'].head(5)
```

[12]:

	county	Sex	Cancer	PopTot	AAIR	Cancer_Incidents
1321	Tulare	Male	Prostate	1183384.0	91.9	1087.529896
1322	Tulare	Female	Prostate	1171391.0	0.0	0.000000
1323	Tulare	Male	Breast	1183384.0	0.5	5.916920
1324	Tulare	Female	Breast	1171391.0	105.0	1229.960550
1325	Tulare	Male	Lung	1183384.0	35.4	418.917936

1.3.2 PFAS Dataset

The PFAS dataset records individual chemical measurements at California water wells from 2016–2024, including well metadata, chemical names, and result values. Each row represents one compound at a specific location and time. The data focuses on PFAS compounds like PFNA and NMEFOSAA, mostly from municipal sources, and comes from the Division of Drinking Water.

[13]:
```
pfas_df = pd.read_csv('data/raw/pfas.zip', compression='zip',␣
↪encoding='ISO-8859-1')
pfas_df.head(2).iloc[:, 4:13]
```

```
/tmp/ipykernel_99056/121182058.py:1: DtypeWarning: Columns (20,26,27,29,32,36)
have mixed types. Specify dtype option on import or set low_memory=False.
  pfas_df = pd.read_csv('data/raw/pfas.zip', compression='zip',
encoding='ISO-8859-1')
```

[13]:

	gm_chemical_vvl	gm_chemical_name	gm_result_modifier	gm_result	gm_chemical_units	gm_reporting_limit	gm_samp_collection_date	gm_latitude	gm_longitude
0	NMEFOSAA	N-Methyl perfluorooctane sulfonamidoacetic aci…	<	1.7	NG/L	1.7	2021-12-19	41.781029	-124.2006
1	PFNA	Perfluorononanoic acid (PFNA)	<	1.7	NG/L	1.7	2021-03-30	41.781029	-124.2006

PFAS data was filtered to include only measurements from 2017 to 2021, dropping rows missing chemical values (gm_result). Key fields such as chemical abbreviation, measurement (ng/L), sample date, and coordinates were extracted. Latitude and longitude were mapped to counties using

the 2022 Census Shapefile. The data was then reshaped into a wide format with each chemical as a column, missing values filled with the county's annual average, and finally merged with the cancer dataset for analysis.

[14]:
```
from pfas_data_cleaner import (fill_missing_values, filter_date_range,␣
 ↪remove_missing_results,
    extract_relevant_columns, load_county_shapefile, calculate_county_bounds,␣
 ↪get_california_counties,
    add_county_column, create_wide_format, calculate_total_pfas)

pfas_df = filter_date_range(pfas_df, '2017-01-01', '2021-12-31')
pfas_df = remove_missing_results(pfas_df)
pfas_df = extract_relevant_columns(pfas_df)

pfas_df.head(2)
```

[14]:
```
       gm_chemical_vvl  gm_result gm_samp_collection_date  gm_latitude  \
551931        NMEFOSAA       12.0              2021-12-31    35.956463
572532         PFUNDCA       11.0              2021-12-31    35.956463

        gm_longitude
551931   -120.011327
572532   -120.011327
```

[15]:
```
gdf = load_county_shapefile('data/raw/tl_2022_us_county.zip')

county_bounds = calculate_county_bounds(gdf)
california_counties = get_california_counties()
california_county_bounds = {name: county_bounds[name] for name in␣
 ↪california_counties}

# Add county information
pfas_df = add_county_column(pfas_df, california_county_bounds)
pfas_df = pfas_df[['gm_chemical_vvl', 'gm_result', 'gm_samp_collection_date',␣
 ↪'county']]

pfas_df.head(2)
```

[15]:
```
       gm_chemical_vvl  gm_result gm_samp_collection_date  county
551931        NMEFOSAA       12.0              2021-12-31  Fresno
572532         PFUNDCA       11.0              2021-12-31  Fresno
```

[16]:
```
pfas_df = create_wide_format(pfas_df)
pfas_df = fill_missing_values(pfas_df)
pfas_df = calculate_total_pfas(pfas_df)

final_df = pd.merge(pfas_df, cancer_df, on='county', how='inner')
```

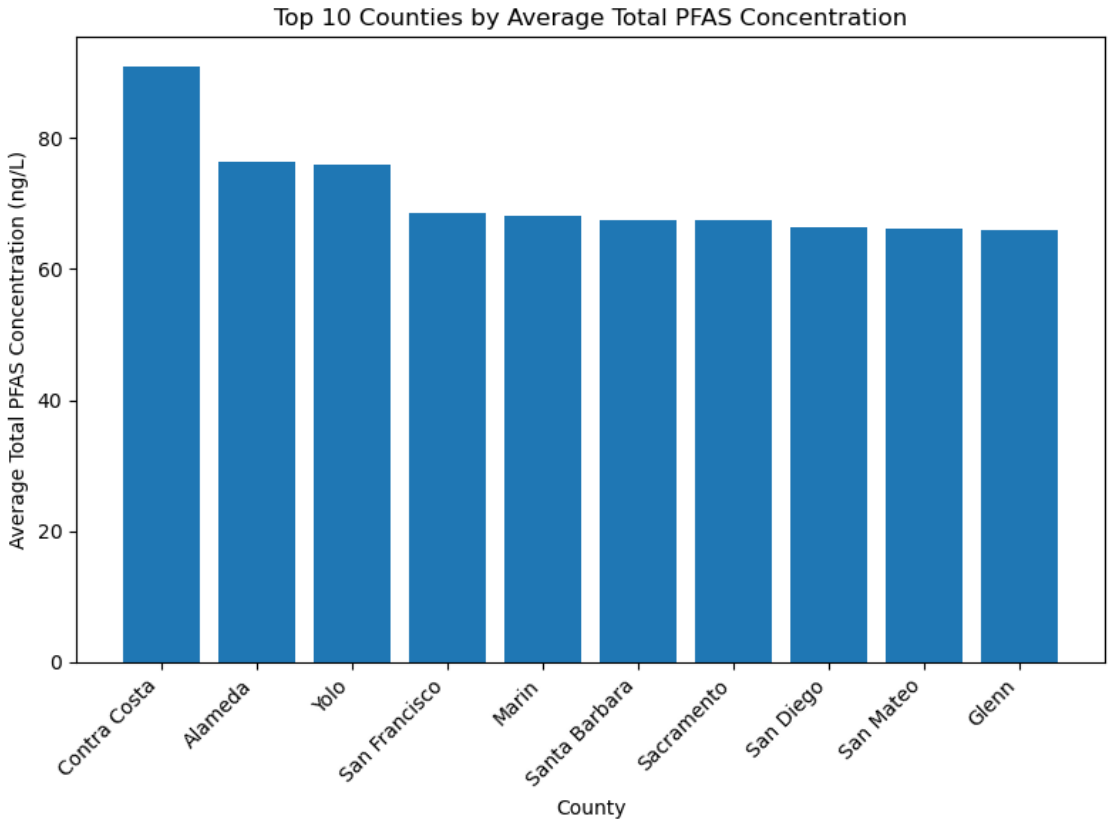
```
final_df.head(2)
```

[16]:
```
  county gm_samp_collection_date   10:2FTS  11ClPF3OUDS   3:3FTCA    4:2FTS  \
0   Yuba              2021-11-18  2.322092          2.0  4.251273  7.616667
1   Yuba              2021-11-18  2.322092          2.0  4.251273  7.616667

    5:3FTCA    6:2FTS   7:3FTCA    8:2FTS  …     PFPES  PFTEDA  PFTRIDA  \
0  3.418426  7.616667  5.346817  7.616667  …  1.922222     2.0      2.0
1  3.418426  7.616667  5.346817  7.616667  …  1.922222     2.0      2.0

   PFUNDCA  total_pfas_concentration     Sex    Cancer    PopTot   AAIR  \
0      2.0                178.339414    Male  Prostate  205837.0  116.2
1      2.0                178.339414  Female  Prostate  196337.0    0.0

   Cancer_Incidents
0        239.182594
1          0.000000

[2 rows x 46 columns]
```

1.4 Exploratory Data Analysis

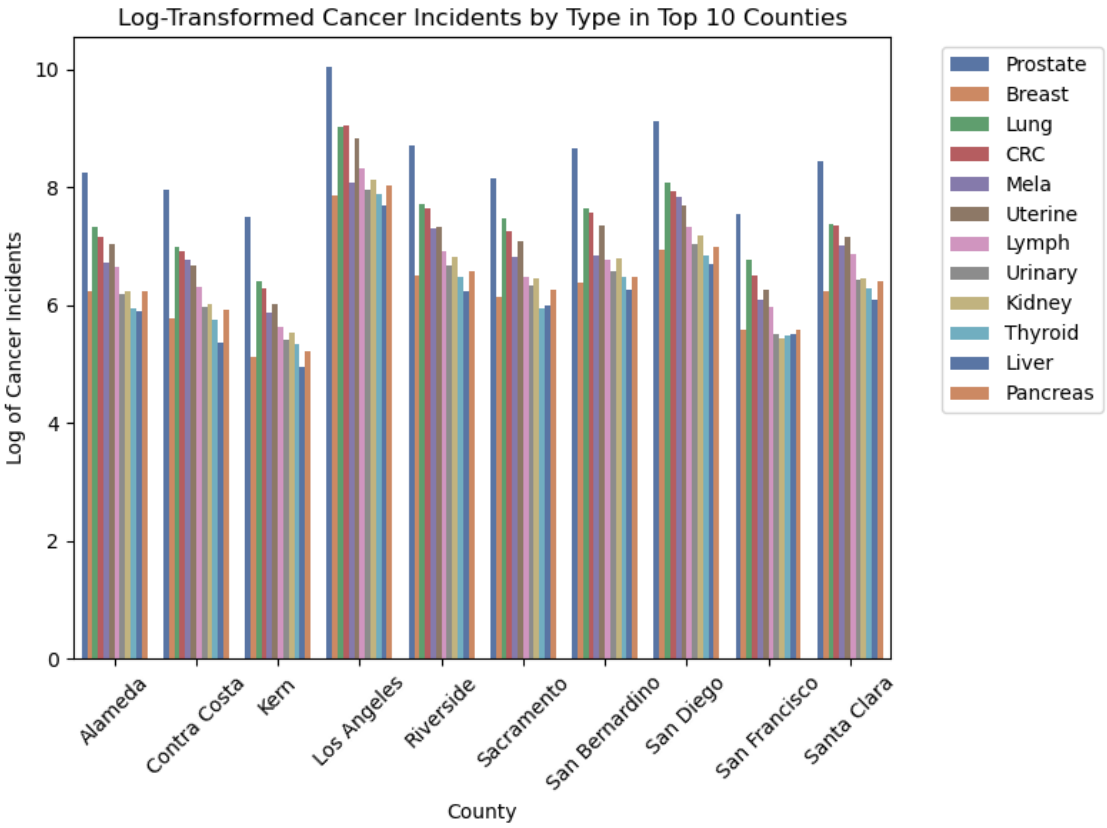
[3]:
```
df = pd.read_csv('cleaned_pfas_cancer_merged.csv')
```

[4]:
```
from visualizations import get_top_ten_cancer_counties,␣
 ↪get_top_ten_pfas_counties, hypothesis_plot
plot = get_top_ten_pfas_counties(df)

plot.show()
```



[5]:
```
plot = get_top_ten_cancer_counties(df)
plt.show()
```

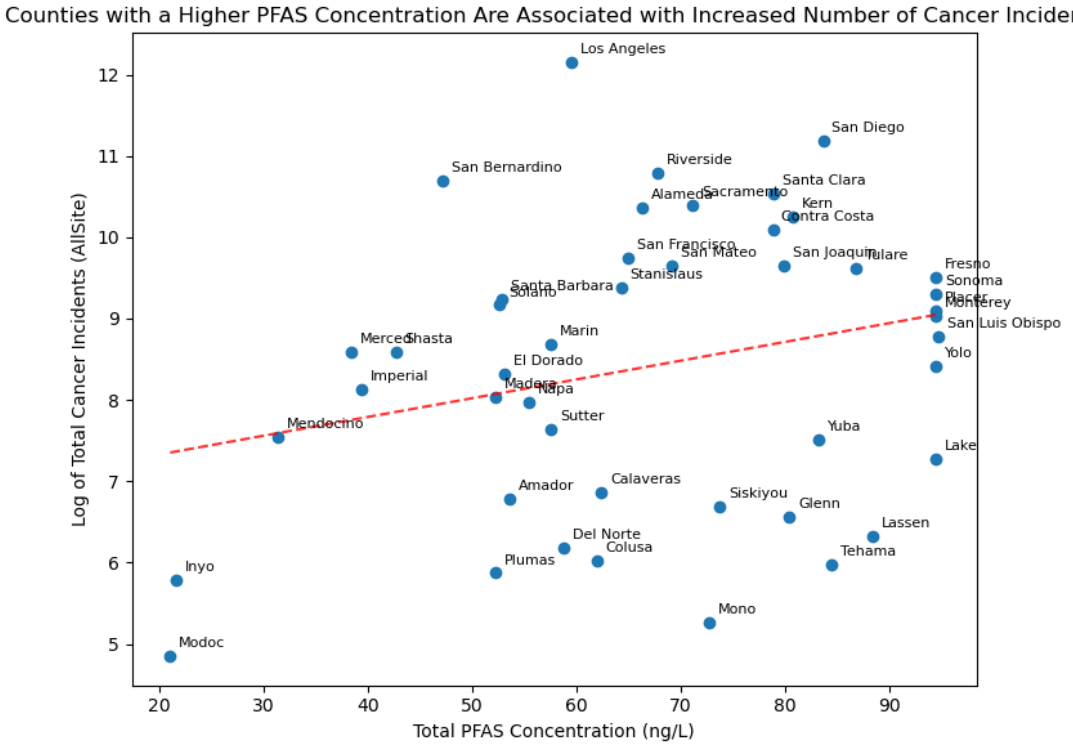
```
/home/shayan283/anaconda3/envs/cs418env/lib/python3.12/site-
packages/pandas/core/arraylike.py:399: RuntimeWarning: divide by zero
encountered in log
  result = getattr(ufunc, method)(*inputs, **kwargs)
```



1.5 PFAS vs Cancer Rate – Hypothesis Testing

This section explores whether higher PFAS levels are associated with increased cancer rates (AAIR) across California counties.

[6]:
```
plot = hypothesis_plot(df)
plot.show()
```


Counties with a Higher PFAS Concentration Are Associated with Increased Number of Cancer Incidents
Log of Total Cancer Incidents (AllSite)
Total PFAS Concentration (ng/L)
Los Angeles
San Diego
San Bernardino
Riverside
Santa Clara
Alameda
Sacramento
Kern
Contra Costa
San Francisco
San Mateo
San Joaquin
Tulare
Fresno
Sonoma
Placer
Monterey
San Luis Obispo
Yolo
Santa Barbara
Solano
Stanislaus
Marin
Merced
Shasta
El Dorado
Imperial
Madera
Napa
Sutter
Mendocino
Yuba
Lake
Amador
Calaveras
Siskiyou
Glenn
Lassen
Del Norte
Colusa
Tehama
Plumas
Inyo
Mono
Modoc

[ ]: