Data Mining of PFAS Contamination in US Drinking Water: Correlation with Socioeconomic Factors and Regional Disparities

Arwyn Lewis
University of Colorado, Boulder

Chenchen Yuan
University of Colorado, Boulder

Copeland Laris
University of Colorado, Boulder

Moritz Knodler
University of Colorado, Boulder

ABSTRACT

This project investigates the presence and distribution of per- and polyfluoroalkyl substances (PFAS) in drinking water sources. PFAS are synthetic chemicals widely used in industrial and consumer products that have been linked to environmental and health risks. We collect, clean, and explore two datasets: water quality data from the Environmental Protection Agency (EPA) and a poverty dataset from the U.S. Census Bureau. The datasets are assessed for missing values, outliers, merged, and analyzed to uncover patterns between PFAS contamination and regional socioeconomic indicators. We apply multiple machine learning and data mining techniques, including classification, clustering, regression, and frequent pattern mining. Our findings show notable correlations between income levels and PFAS exceedances, highlight common PFAS co-occurrence patterns, and identify regional contamination clusters.

CCS CONCEPTS

Applied computing → Environmental sciences; Computing methodologies → Supervised learning; Unsupervised learning; Feature selection; Mathematics of computing → Regression; Cluster

KEYWORDS

PFAS, water contamination, environmental justice, poverty, classification, clustering, apriori, machine learning, geographic data

1 INTRODUCTION

1.1 Water Quality

The water quality dataset was obtained from the US Environmental Protection Agency (EPA). UCMR, or the Unregulated Contaminant Monitoring Rule, is how the EPA collects data for contaminants that might be in drinking water but do not have regulatory standards under the Safe Drinking Water Act or National Primary Drinking Water Regulations. The UCMR program was developed as a way to track these contaminants every five years. However for PFOA and PFOS, the EPA has set maximum contaminant levels in drinking water (MCL) at 4.0 ppt and a MCL of 10 ppt for PFHxS and PFNA. There are multiple files that contain water quality data points for PFAS contamination over the years throughout the United States. Only UCMR 5, which contains PFAS data for 2023-2025, and UCMR 3, which contains PFAS data from 2013 to 2015 were used. The raw data contained a lot of information, including the PWS (Public Water System) names, IDs, and sizes; Facility names, IDs, and water types; Sample types, sample collection dates, contaminants, units, methods, result values, state, region, and more (Figure 1).

The data was collected through a script that outputs the zip file from a given website. This dataset is relevant to our research questions since it contains the PFAS levels that were sampled in our water sources across the country. We are interested in determining how PFAS contamination varies with other parameters such as geography, time, poverty, state boundaries, water sources, etc. This will allow us to further analyze contamination patterns and identify at-risk areas and populations.

	PWSID	PWSName	Size	FacilityID	FacilityName	FacilityWaterType	SamplePointID	SamplePointName	SamplePointType
0	010106001	Mashantucket Pequot Water System	L	6	MPTN WTP	GU	D11	WTP EPTDS	EP
1	010106001	Mashantucket Pequot Water System	L	6	MPTN WTP	GU	D11	WTP EPTDS	EP
2	010106001	Mashantucket Pequot Water System	L	6	MPTN WTP	GU	D11	WTP EPTDS	EP
3	010106001	Mashantucket Pequot Water System	L	6	MPTN WTP	GU	D11	WTP EPTDS	EP
4	010106001	Mashantucket Pequot Water System	L	6	MPTN WTP	GU	D11	WTP EPTDS	EP

Figure 1: Part of the EPA raw data

1.2 Poverty

The poverty dataset was obtained from the US Census website and provides small area income and poverty estimates (SAIPE) of income and poverty statistics from states. The data aims to provide estimates of income and poverty for the administration of federal programs. The raw datasets have parameters such as (state) names, median income, child poverty counts, child poverty rates, overall poverty counts and rates, corresponding years, etc (Figure 2).

The poverty data was collected through an API call from census.gov. This dataset is relevant to our research questions since we are interested in determining if PFAs contamination has any correlation with poverty levels, as well as region. Once both sets of data files were acquired, they were uploaded to Github.

	NAME	COUNTY	SAEMHI_PT	SAEPOV0_17_PT	SAEPOVRT0_17_PT	SAEPOVALL_PT	SAEPOVRTALL_PT	SAEPOVU_0_17
0	Autauga County	001	26898	2071	20.2	4956	14.3	None
1	Baldwin County	003	24043	4838	18.6	13031	13.2	None
2	Barbour County	005	18673	2826	38.2	6601	26.2	None
3	Bibb County	007	19604	1206	25.0	3133	18.9	None
4	Blount County	009	24035	1972	19.4	5296	13.4	None

Figure 2: Part of the raw Census poverty data

1.3 PWS County Data

The poverty and EPA data only had geographical data on the state level in addition to the PWS names. We used Google's Places API to fetch county names from the PWS name, and exported that data as a seperate CSV file.

2 DATA PREPARATION AND PREPROCESSING

2.1 Water Quality

Columns not needed for our analysis were dropped (e.g. FacilityID, PWSID). The MRL had NA values because certain contaminants do not have a minimum reporting level. This value does not have any health implications - it is just the lowest value that labs can report. To make it clear that this value has not been set by governing and research bodies yet, these NAs are replaced with -1. The NAs in column AnalyticalResultValue will be replaced with 0 since they are lower than the minimum value labs need to report, and the concentration is then functionally zero.

```
Out[8]:
PWSID                          0
PWSName                        0
Size                           0
FacilityID                    60
FacilityName                  30
FacilityWaterType              0
SamplePointID                  0
SamplePointName                0
SamplePointType                0
AssociatedFacilityID     2503268
AssociatedSamplePointID  2503268
CollectionDate                 0
SampleID                       0
Contaminant                    0
MRL                       190359
Units                          0
MethodID                       0
AnalyticalResultsSign          0
AnalyticalResultValue    2850960
SampleEventCode                0
MonitoringRequirement          0
Region                         0
State                          0
UCMR1SampleType          3363695
dtype: int64
```

Figure 3: NAs in the original dataset

An overview of our data is presented in Figure 4 and 5. There are 65 unique states, which is because different tribal territories are included where instead of a state, they just have their EPA region designation (01, 02, etc).

```
📌 Dataset Shape (Rows, Columns):
(3363695, 16)

📌 Column Names:
['PWSName', 'Size', 'FacilityName', 'FacilityWaterType', 'SamplePointName', 'SamplePointType', 'CollectionDate', 'SampleID',
'Contaminant', 'MRL', 'Units', 'AnalyticalResultsSign', 'AnalyticalResultValue', 'MonitoringRequirement', 'Region', 'State']

📌 General Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 3363695 entries, 0 to 3363694
Data columns (total 16 columns):
 #   Column                 Dtype
---  ------                 -----
 0   PWSName                object
 1   Size                   object
 2   FacilityName           object
 3   FacilityWaterType      object
 4   SamplePointName        object
 5   SamplePointType        object
 6   CollectionDate         object
 7   SampleID               object
 8   Contaminant            object
 9   MRL                    float64
 10  Units                  object
 11  AnalyticalResultsSign  object
 12  AnalyticalResultValue  float64
 13  MonitoringRequirement  object
 14  Region                 int64
 15  State                  object
dtypes: float64(2), int64(1), object(13)
memory usage: 410.6+ MB
```

Figure 4: Data Overview (1)

As our study focuseed on PFAS only, we filtered the dataframe accordingly. We converted the CollectionDate to a datetime format and added columns for Year and Month for easier retrieval and analysis later on.

We also calculated which samples had a measured value over or under the MRL, or minimum required level. This is the minimum level labs are required to report to the EPA. While the MRL does not have any health indications, knowing that some contaminants are not at the MRL while some are over is still useful. The number of values that are above or at/below the MRL are counted, and for those that are above, the relative contamination level is calculated by taking the recorded value divided by the MRL. Any contaminant with a measured level at or below the MRL was filled with zeros instead of NAs, since the levels were so low that they functionally zero.

```
📌 Unique Value Count per Column:
PWSName                  16507
Size                         2
FacilityName             25794
FacilityWaterType            4
SamplePointName          50316
SamplePointType              3
CollectionDate            2617
SampleID                395782
Contaminant                 90
MRL                         33
Units                        3
AnalyticalResultsSign        2
AnalyticalResultValue    51938
MonitoringRequirement        3
Region                      10
State                       65
dtype: int64

📌 Summary Statistics for Categorical Columns:
                              PWSName     Size         FacilityName  \
count                         3363695  3363695              3363695
unique                          16507        2                25794
top     Suffolk County Water Authority        L  Distribution System
freq                            24119  2606676               267367

       FacilityWaterType              SamplePointName SamplePointType  \
count            3363695                      3363695          3363695
unique                 4                        50316                3
top                   GW  Entry Point to Dist. System               EP
freq             1834004                       406383          2983025

       CollectionDate SampleID Contaminant    Units AnalyticalResultsSign  \
count         3363695  3363695     3363695  3363695                3363695
unique           2617   395782          90        3                      2
top       11/14/2023    EP001       PFHpA     µg/L                      <
freq             8522       91       82296  3358475                2850960

       MonitoringRequirement    State
count                3363695  3363695
unique                     3       65
top                       AM       CA
freq                 3275905   504509

📌 Number of Duplicate Rows:
0
```

Figure 5: Data Overview (2)

2.2 Poverty Data

Missing values, duplicates, or any outliers are assessed along with updating the dataset to include more interpretable column names and only include relevant years in relation to the other dataset used (EPA data). We also assessed the data for completeness, consistency, and usability. The original dataset had their own naming convention for the columns (Figure 6). The column names were renamed to be more consistent - for example. Redundant columns, e.g. NAME (full state name) and STATE (numbers of states), were removed.

	NAME	Median_Income	Child_Poverty_Count	Child_Poverty_Rate	Poverty_Count	Poverty_Rate	All_Child_Poverty_Count
1020	Alabama	42882	300649	27.4	889091	18.9	1096240.0
1021	Alaska	70058	25030	13.6	72643	10.1	184278.0
1022	Arizona	48504	422435	26.6	1206948	18.6	1590036.0
1023	Arkansas	40605	197129	28.3	557399	19.4	695853.0
1024	California	60185	2119056	23.5	6328064	16.8	9030489.0
1025	Colorado	58942	204866	16.8	665351	12.9	1222544.0
1026	Connecticut	67262	112225	14.5	373387	10.7	775350.0
1027	Delaware	58244	38254	19.1	115774	12.9	200641.0
1028	District of Columbia	66326	32071	29.3	115096	18.8	109630.0

Figure 6: Clean Poverty Data from Census.gov

2.3 PWS County Data

Any rows where the PWS name did not pull a county name was removed - that means that some PWS's won't be included in the county-wide analysis of PFAS contamination, but the dataset is so large that removing a few rows would not impact the final analysis. After cleaning, the data was exported

to a CSV file and then uploaded to Github so that it can be pulled for future analysis (Figure 7).

	State	PWSName	County
0	PR	GURABO URBANO	Rincón
1	GA	CUMMING	Forsyth County
2	IA	ASBURY MUNICIPAL WATER SYSTEM	Dubuque County
3	OK	MUSKOGEE	Muskogee County
4	TX	Lake Cities Municipal Utility Authority	Denton County

Figure 7: Cleaned County Data

2.4 Combined Dataset

The poverty and water quality dataset were matched by state and year. The PWS county data was then merged with the dataset (Figure 8).

	County	PWSName	Size	FacilityName	FacilityWaterType	SamplePointName	SamplePointType	CollectionDate
463166	Williamson County	HERRIN	L	Rend Lake Intercity Water Interconnection	SW	Maintenance Shed (Utility Room Sink)	EP	2024-08-20
995954	Las Piedras	LAS PIEDRAS HUMACAO	L	PF LAS PIEDRAS HUMACAO	SW	PF LAS PIEDRAS HUMACAO	EP	2024-04-18
980069	York County	STEWARTSTOWN BORO WATER AUTH	S	Well House - 5	GW	Well 5 EP	EP	2023-09-19
1070877	Larimer County	CANYON SPRINGS WATER	S	WELL 1	GW	Sample Tap @ 133 WATER OAK LN, CANYON LAKE	EP	2024-07-17
1435537	Chickasaw County	City of Houston	S	2nd Avenue Well	GW	#1	EP	2014-05-07
1458544	Camden County	NJ American Water Co. - Raritan	L	Hummocks Station Plant	GW	EPTDS from Hummocks Station Plant	EP	2015-09-04
974275	Dauphin County	SUEZ MECHANICSBURG	L	Yellow Breeches TP	SW	Yellow Breeches Finished Water	EP	2023-02-08

Figure 8: Combined dataset from EPA, Census, and PWS County data

3 EDA AND VISUALIZATION

3.1 Poverty Data

Figure 9: Heat map of poverty data

The correlation table and the heat map show relatively strong (above −0.70) negative correlations between: Child_Poverty_Rate and Median_Income, and Poverty_Rate and Median_Income. This makes sense, as it suggests higher income areas have lower poverty rates.

There are relatively strong (above 0.70) positive correlations between: Child_Poverty_Count and All_Poverty_Count, Child_ Poverty_Count and All_Child_Poverty_Count, Poverty_Rate and Child_Poverty_Rate, Poverty_Count and Child_Poverty_Count and All_Child_Poverty _Count and All_Poverty_Count. These correlations also make sense given the economic expectations, since it is suggesting that child poverty and general poverty rise and fall together.

The poverty rate across states tends to fluctuate, with Nevada having the highest mean poverty rate, and North Carolina having the lowest. In general most other states appear to have a poverty rate around 13.0, but there is no clear pattern between the state and the poverty rate present.

From the Q-Q plots, we can determine that the Child_Poverty_Rate and the Poverty_Rate variables are generally normally distributed, but both have some skewness in the left tail, indicating that both of these features likely have a slight right skew.

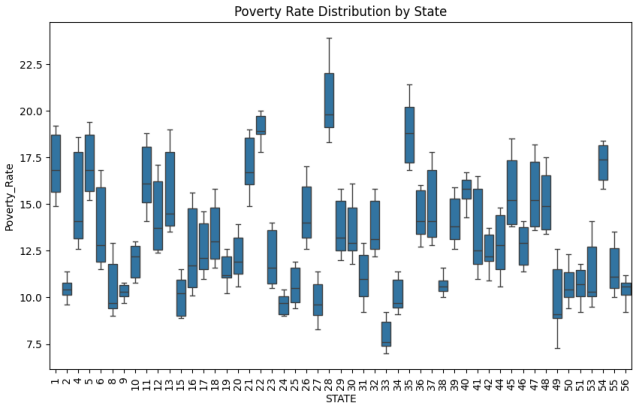

Figure 10: Combined dataset from EPA, Census, and PWS County data

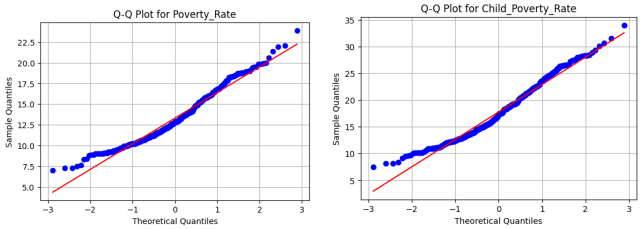

Figure 11: Diagnostic Plots for Poverty Rates

3.2 EPA Data

For our further analysis we focused on data restricted to year 2023, as most poverty and EPA data was available in this year. Figure 12 shows a rough overview of the samples in this timeframe by facility water type. Most samples were taken from groundwater (left chart). While for PFAS contaminated samples the split is similar (right chart), the higher ratio of surface water samples indicates that these samples have a higher PFAS contamination ratio vs. ground water (samples with any detected amount of PFAS, right chart).

In the following we distinguish between two major types of contamination levels:

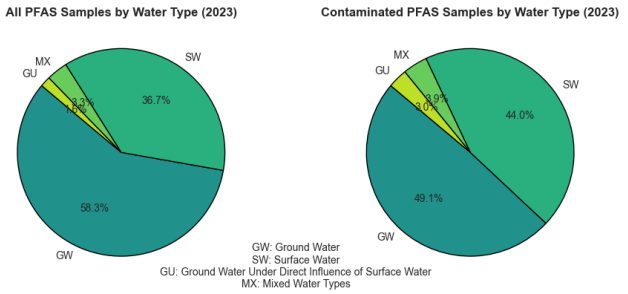

Figure 12: PFAS Contaminated by Facility Water Type

- **MRL**, or the minimum reporting level, is the lowest concentration of contaminant that can be reported to the EPA for UCMR substances. The MRL is essentially the reporting threshold, and does not have health implications.
- **MCL** is the maximum contaminant level of a contaminant allowed in drinking water and does have health implications. These are legally enforceable levels, and anything above is considered illegal and unsafe. Samples can have concentrations above the MRL but still be considered to be within the legal limit per its MCL.

For a more meaningful analysis we grouped our data by US states (excluding US territories). Figure 13 displays both the percentage of samples exceeding both MRL and MCL by state. We notice that the MCL is always below or equal the MRL level. The states with most samples exceeding MRL (and MCL) levels are Delaware 9.6% (7.7%), New Jersey 6.4% (5.2%) and Florida with 6.0% (4.9%) respectively.

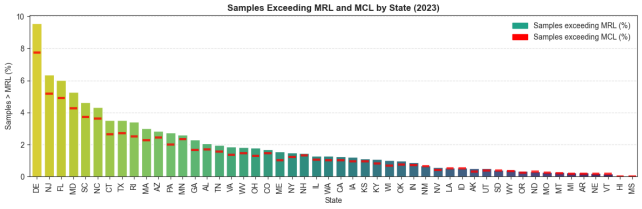

Figure 13: PFAS Contaminated Samples exceeding MRL and MCL by State

Figure 14 compares the MRL (left) and MCL (right) sample distributions in more detail. For a comprehensive analysis we display strip plots, violin plots and bar plots. The data shows only a few outliers and appears to be normally or exponentially distributed. The distributions of MRL and MCL exceedance per state are quite similar, with a mean of approximately 2% of all samples being contaminated beyond the thresholds. Since MCL provides a more direct measure of potential health impacts, our further analysis and correlations will focus primarily on this parameter.

While the EPA data aggregated by state is easy to interpret, the data points were not granular enough for a further correlation analysis with the poverty data. Our data set was lacking any other location data besides the state information. We used a script to first fetch geo-coordinates of the water facilities the samples were taken at and then converted these into county names. The final data set resulted in aggregated data for approximately 1500 counties scattered across the US (including poverty and EPA data).

Figure 15 shows the county data plotted across a map of the US (excluding Hawaii and Alaska). Each circle corresponds to sample data aggregated per county. The color scale and circle size indicated the percentage of samples

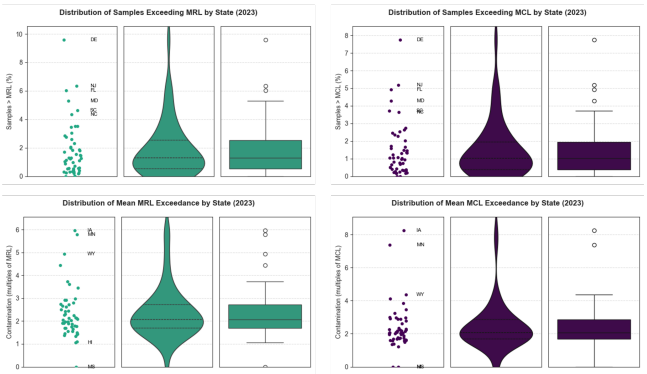

Figure 14: Distributions of Water Samples exceeding MRL (and MCL) by state

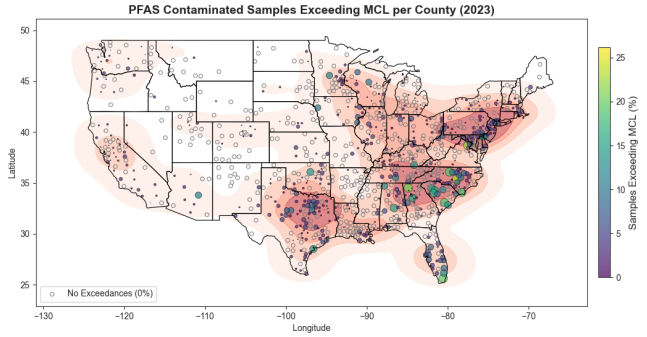

Figure 15: PFAS Contaminated Samples Excedding MCL per County (2023)

exceeding MCL thresholds for each county. Counties with water samples but without any exceedance of MRL thresholds are plotted as empty circles. A KDE heatmap is overlaid for better clarity.

We can distinguish three larger hotspots with particularly high PFAS water contamination, centered around New Jersey, Fort Worth / Dallas TX, and the region stretching from North Carolina to Georgia. The regions in the mid-west and west, including California, show a remarkably less pronounced water contamination. A more detailed analysis and causal relationships will be explored in our final report.

4 METHODS AND MODEL IMPLEMENTATION

4.1 Classification

For the following prediction classification models, the model will attempt to predict whether a water sample will exceed the maximum contaminant level (MCL). The MCL is the established threshold determining whether water can be delivered to users of the public water system. In-short exceeding this level would likely result in adverse health impacts for users. Models will be fed basic information on when and where the sample was collected then asked to predict whether the sample exceeds the MCL.

For the following analysis adjustments to the original data frame were made to ensure model compatibility - including one-hot encoding of categorical variables.

The data was then split into test and training sets - the training set containing 80% of the data. The training sets were then under-sampled to provide

adequate examples of observations which both exceed and don't exceed the MCL level.

To determine the best model for this binary classification problem three models were created, and their performance was evaluated against one another: decision tree, naive bayes, and support vector machines.

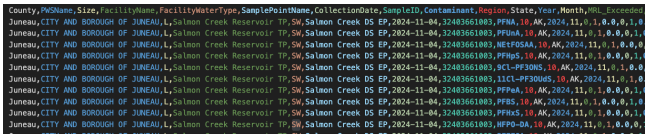

Figure 16: Snippet of Preprocessed Dataset

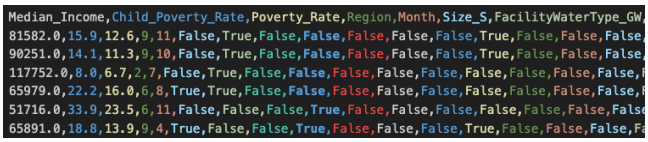

Figure 17: Snippet of Processed Dataset for Classification

4.1.1 Decision Tree. The decision tree model was included due to its high interpretability and strong predictive capabilities for binary classifications. For this model and gini criterion was implemented and tree depth was limited to 5 and the minimum samples split was set to 1% to maximize model performance while preventing overfitting.

```
Accuracy: 0.8486267639634172
                precision    recall  f1-score   support

Not Exceed MCL       1.00      0.85      0.92    142560
    Exceed MCL       0.08      0.77      0.14      2426
```

Figure 18: Decision Tree Performance Metrics

As evidenced by the produced confusion matrix (Figure 19, below) and the above performance statistics the tree performed quite well in predicting whether a sample will exceed the MCL. Exhibiting a recall score of 77% on the exceeded MCL label. However, the tree tends to over predict an observation will exceed the threshold. Given the context of the problem, this false-positive is preferred to the alternative false-negative. Further tests can be run on contaminated water, but if water is cleared to be safe it could have catastrophic impacts on the public.

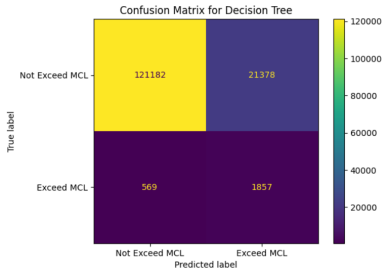

Figure 19: Decision Tree Confusion Matrix

```
Accuracy: 0.7955043935276509
                precision    recall  f1-score   support

Not Exceed MCL       0.99      0.80      0.88    142560
    Exceed MCL       0.06      0.72      0.10      2426
```

Figure 20: Multinomial Naive Bayes Performance Metrics

4.1.2 Multinomial Naive Bayes. The naive bayes was included due to its ability to identify feature interactions and its high interpretability lending itself to classifications. For this naive bayes model a multinomial model was implemented to best manage count and frequency data.

As evidenced by the produced confusion matrix (Figure 21, below) and the above performance statistics the multinomial naive bayes model also performed adequately in the context of MCL prediction. However, the NB model over assumes that an observation will exceed the threshold. Additionally, mislabeling about 1/4 of observations that exceed the MCL. These inaccuracies are likely due to the model's assumption of independence between features.

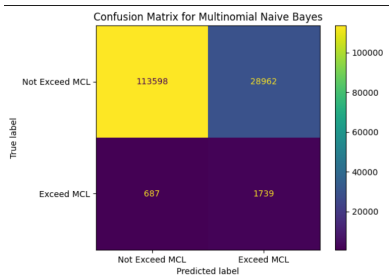

Figure 21: Multinomial Naive Bayes Confusion Matrix

4.1.3 Support Vector Machine. The support vector machine was included due to the powerful and flexible nature of separating data for classification. For this model an rbf kernel was implemented as the data should not be assumed to be linearly separable. However, it is assumed to be separated in higher dimensional space. Furthemore, a gamma parameter of 1 was implemented to maximize model performance while reducing the risk of overfitting.

```
Accuracy: 0.9007076545321617
                precision    recall  f1-score   support

Not Exceed MCL       0.99      0.90      0.95    142560
    Exceed MCL       0.11      0.72      0.20      2426
```

Figure 22: RBF SVM Performance Metrics

As evidenced by the above performance statistics the rbf svm performed very well in these metrics, boasting an accuracy score of 90%. However, the model recall on the exceed MCL label is only 72%. As with the other models, this RBF SVM also exhibits a high false-positive rate.

As evidenced by the above performance statistics the rbf svm performed very well in these metrics, boasting an accuracy score of 90%. However, the model recall on the exceeded MCL label is only 72%. As with the other models, this RBF SVM also exhibits a high false-positive rate.

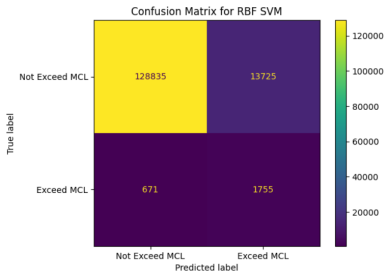

Figure 23: RBF SVM Confusion Matrix

4.1.4 Model Comparison. In addition to assessing the confusion matrices and performance statistics for each of the three created models the following ROC-AUC and PR comparison curves were also created.

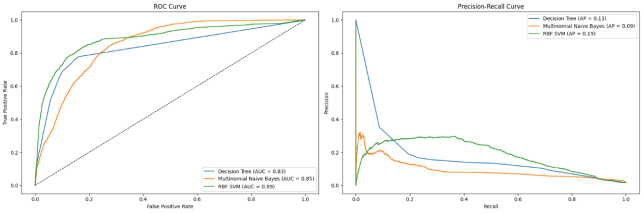

Figure 24: ROC and PR curves for classification models

It is clear from nearly all of the metrics that the rbf support vector machine model does the best job at predicting whether an observation will exceed the MCL. This model boasts the highest AUC (0.89), AP(0.19), accuracy (.9), and F1 scores (.95, .20). Given this performance this will likely be the model implemented. However, given the context of the problem and the high consequence of false negatives, the decision tree model should also be considered given that it has the highest recall (.77) for the exceeded MCL category.

4.1.5 Fairness Audit. After the creation of the above classification models, we performed a fairness audit to examine if the best performing model - the SVM - was under-performing with specific subgroups. Specifically, we explored how the model performed with groups exhibiting different poverty rates. The SVM performance was assessed when predicting a response for an observations from the lowest quartile for poverty rate (<0.093) compared with predictions for observations from the highest quartile (>0.156). We also then ran a fairness metric test for the highest poverty quartile vs. all other groups. We are specifically concerned that the model will under-perform for higher poverty groups.

Upon initial examination of the fairness test and performance metrics from the subgroup predictions (above) there does not seem to be any major red flags. In fact, the high-poor group exhibits a higher accuracy rate than the low-poor group. However, the recall for high-poverty rate observations that exceed the MCL is far worse (0.69) than the same recall rate for the low-poverty group (0.78). In other-words, while the model has a high accuracy rate with the high-poverty group, it does a significantly worse job at identifying harmful water samples when compared to the low-poverty group.

```
High Poverty Group:
Accuracy: 0.9372156505914467
                precision    recall  f1-score   support

Not Exceed MCL       1.00      0.94      0.97     35824
    Exceed MCL       0.13      0.69      0.21       443
```

```
Low Poverty Group:
Accuracy: 0.8542225950782998
                precision    recall  f1-score   support

Not Exceed MCL       0.99      0.86      0.92     34962
    Exceed MCL       0.11      0.78      0.19       798
```

Figure 25: Poverty Fairness Comparison: Model Performance Metrics

```
# create list of contaminants by samples
transactions_samples=df_exceed.groupby("SampleVolnName")["Contaminant"].apply(list).tolist()
transactions_samples
```

```
[['PFHxA',
  'PFOA',
  'PFBA',
  'PFHpA',
  'PFPeA',
  'PFDoA',
  'PFTeA',
  'PFHxA',
  'PFHpA',
  'PFPeA'],
 ['PFPeA', 'PFOS', 'PFOA', 'PFHxA', 'PFBS', 'PFHpS'],
 ['PFBA'],
 ['PFBA', 'PFBA'],
 ['PFBS'],
 ['PFBA',
  'PFPeA',
  'PFHxA',
```

Figure 26: Transformed data for apriori()

4.2 Frequent Pattern Mining - Apriori

Apriori was used to determine which contaminants tend to occur together. PFAS does not refer to just one compound - it encompasses a wide range of chemicals. We want to know which PFAS, specifically, tend to appear together. Apriori from the package Apyori was used. We tried to use Mlxtend's apriori() as well, but the runtime was significantly longer than that of Apyori. Not only do we want to know which contaminants appear together more often, more specifically we want to know which contaminants that exceed their MCL appear more often. PFAS chemicals are so widespread at this point that looking through the entire dataset would not only take a long time, but wouldn't provide that much useful information. A new dataframe was made containing only contaminants that exceeded their MCL as those are more of concern and high priority for remediation. The dataset first had to be transformed to a format that the apriori() method could accept (Figure 26).

Apriori was then run on those 'transactions' with specified minimum support, confidence, and lift values that have been tweaked to get a manageable amount of results. The results were converted to an array where the support, confidence, and lift values for each itemset were extracted and then everything was reformatted back to a dataframe. Duplicates were removed, and we only filtered for 3+ itemsets as a higher number of items can give us more insight about their relationships.

4.2.1 Sample-Level. Apriori for co-occurring contaminants at the sample level yielded the six results seen in Figure 27 and **??**. Since there are a lot of samples, the lower support and confidence levels of around 20% are expected. However, a lift of 1 is a bit concerning as that means that these co-occurrences have the same probability of happening by chance or independently. The sample-level may be too localized to see any major commonalities, as water quality will differ across the country and individual samples would not capture all those differences.

	itemsets	support	confidence	lift
49	[PFHxA, PFBA, PFPeA]	0.199803	0.199803	1.0
70	[PFOA, PFHxA, PFPeA]	0.188958	0.188958	1.0
56	[PFBS, PFHxA, PFPeA]	0.184029	0.184029	1.0
77	[PFOS, PFHxA, PFPeA]	0.172198	0.172198	1.0
84	[PFOA, PFOS, PFPeA]	0.166283	0.166283	1.0
63	[PFOA, PFOS, PFHxA]	0.154124	0.154124	1.0

Figure 27: Sample-level apriori results

4.2.2 Facility-Level. Based on the results at the facility-level (Figure 28), it seems that the lifts for these itemsets are over 2, which means that they are strongly associated; these contaminants are twice as likely to occur together than by sheer chance. There are a total of 15 itemsets, with 3 sets having

lifts over 2.5 (Figure 29). They also have a fairly high support value for such a large dataset/large number of facilities. The confidence values are solidly between 25-50%, which for 4-itemsets and the large amounts of PFAS types, are fairly high compared to the sample-level results.

	itemsets	support	confidence	lift
176	[PFOS, PFHxA, PFHxS, PFPeA]	0.101181	0.271028	2.543704
70	[PFBS, PFHxA, PFBA, PFPeA]	0.139560	0.373832	2.537153
104	[PFOA, PFHxA, PFBA, PFPeA]	0.132045	0.353702	2.519876
114	[PFOS, PFHxA, PFBA, PFPeA]	0.122383	0.327822	2.487706
60	[PFBS, PFOS, PFHxA, PFBA]	0.101449	0.308824	2.437874
50	[PFBS, PFOA, PFHxA, PFBA]	0.107622	0.327614	2.398214
186	[PFOA, PFOS, PFHxA, PFPeA]	0.168009	0.450036	2.388652
90	[PFBS, PFOS, PFBA, PFPeA]	0.103060	0.313725	2.380736
98	[PFOA, PFOS, PFHxA, PFBA]	0.107085	0.286844	2.364559
80	[PFBS, PFOA, PFBA, PFPeA]	0.107622	0.327614	2.334018
130	[PFBS, PFOA, PFHxA, PFOS]	0.134192	0.408497	2.327307
166	[PFBS, PFOA, PFPeA, PFOS]	0.140097	0.426471	2.263575
154	[PFBS, PFOS, PFHxA, PFPeA]	0.143586	0.437092	2.249452
142	[PFBS, PFOA, PFHxA, PFPeA]	0.151637	0.461601	2.087289
124	[PFOA, PFOS, PFBA, PFPeA]	0.108964	0.247109	2.037009

Figure 28: Facility-level apriori results

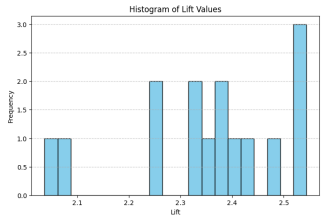

Figure 29: Lift histogram - Facility-level

Unfortunately, because there are so many facilities across the US, finding which facilities match these itemsets isn't very helpful. The results below (Figure 30) show the sheer number of facilities and often how non-descriptive they are.

Figure 30: Facilities that have the common contaminant occurrences

4.2.3 County-Level. The county-level apriori results are shown in Figure 31 - there are 21 3-itemsets, all with lifts above 2, which is quite high - these contaminants are very positively associated (Figure 32). The confidences here are even higher than that at the facility level, and the support values are still respectable for how big the dataset is. These associations make sense, as they are comprised of long-chain PFAS and their short-chain derivatives and substitutes. The PFOA and PFOS might be degrading over time into these other variants.

The counties with these common occurrence sets were found (Figure 33) and then mapped out for better visualization (Figure 34). Like with data visualization, only the continental US was looked at here. Hawaii, Alaska, tribal nations, and other US territories have been excluded for this study. Future studies should include these communities.

This map shows counties that have the most common sets of contaminants that exceed their MCL, and contamination is widespread. The main spots that don't have the common contaminant occurrences are in the sparsely populated, dry desert area of the country (western Montana, Nevada, etc.).

	itemsets	support	confidence	lift
44	[PFHxA, PFPeA, PFHpA]	0.198872	0.411079	2.023992
38	[PFOA, PFHxA, PFHpA]	0.196051	0.891026	2.458121
60	[PFOA, PFPeA, PFHpA]	0.193230	0.878205	2.323311
56	[PFOA, PFOS, PFHpA]	0.190409	0.865385	2.207042
2	[PFHxA, PFBA, PFHpA]	0.188999	0.858974	2.198602
42	[PFOS, PFHxA, PFHpA]	0.188999	0.858974	2.315638
62	[PFOS, PFPeA, PFHpA]	0.187588	0.852564	2.174345
8	[PFOA, PFBA, PFHpA]	0.187588	0.852564	2.379795
14	[PFBA, PFHpA, PFPeA]	0.186178	0.846154	2.013165
30	[PFBS, PFPeA, PFHpA]	0.184767	0.839744	2.096402
16	[PFBS, PFHxA, PFHpA]	0.184767	0.839744	2.229881
12	[PFOS, PFBA, PFHpA]	0.183357	0.833333	2.196406
28	[PFBS, PFOS, PFHpA]	0.181946	0.826923	2.204092
24	[PFBS, PFOA, PFHpA]	0.180536	0.820513	2.246114
0	[PFBS, PFBA, PFHpA]	0.177715	0.807692	2.169143
32	[PFHxA, PFHxS, PFHpA]	0.163611	0.743590	2.522513
52	[PFHxS, PFHpA, PFPeA]	0.163611	0.743590	2.429517
46	[PFOA, PFHxS, PFHpA]	0.157969	0.717949	2.597070
48	[PFOS, PFHxS, PFHpA]	0.156559	0.711538	2.222382
20	[PFBS, PFHxS, PFHpA]	0.155148	0.705128	2.380647
4	[PFBA, PFHpA, PFHxS]	0.155148	0.705128	2.499679

Figure 31: County level apriori results

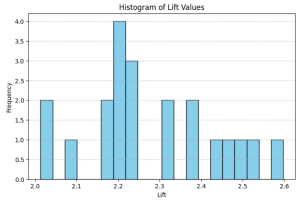

Figure 32: Lift histogram - County-level

Figure 33: Counties that have the common contaminant occurrences

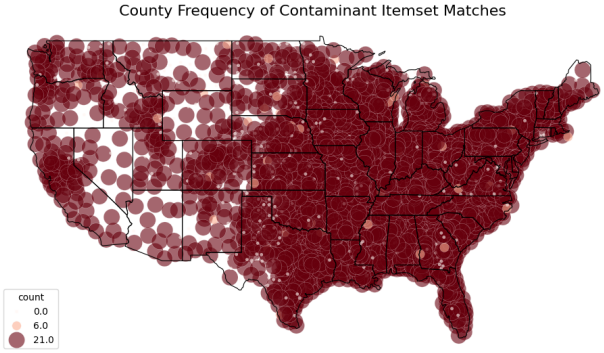

Figure 34: Map of counties in the continental US of contaminant itemset matches

Some counties in the Southeast and Mid Atlantic have less matching itemsets, though there are definitely industrial activities and population centers there. However, that does not mean that the water there isn't contaminated - the waterways could have other PFAS as well that wasn't picked up by the apriori model.

4.2.4 State-level. We looked at 4-itemsets here to avoid seeing the same few contaminants in different permutations. There are 10 4-itemsets here, all with lifts or association values that are positive, but not as high as that of the county level or facility level model. The support and confidence levels are higher, however, which might be because statewide data has more noise, and county/facility level analyses have more specific patterns that the model picked up on. For the state level, we're lumping together a lot of data that might behave differently when looked at in more detail. This is a typical Modifiable Area Unit Problem in mapping where how boundaries/control areas are set affects the results.

	itemsets	support	confidence	lift
78	[PFBS, PFHxS, PFPeA, PFHxA]	0.816327	0.975610	1.195122
116	[PFHxS, PFOS, PFOA, PFHxA]	0.816327	0.975610	1.195122
128	[PFHxS, PFPeA, PFOA, PFHxA]	0.816327	0.975610	1.195122
142	[PFHxS, PFOS, PFPeA, PFHxA]	0.836735	1.000000	1.195122
68	[PFBS, PFHxS, PFOS, PFHxA]	0.816327	0.975610	1.165973
90	[PFBS, PFOS, PFPeA, PFHxA]	0.816327	0.975610	1.165973
154	[PFOA, PFOS, PFPeA, PFHxA]	0.816327	0.975610	1.165973
106	[PFBS, PFOA, PFOS, PFPeA]	0.816327	0.952381	1.138211
166	[PFHxS, PFOS, PFOA, PFPeA]	0.816327	0.952381	1.138211
100	[PFBS, PFHxS, PFOS, PFPeA]	0.816327	0.930233	1.111741

Figure 35: State level apriori results (4-itemsets)

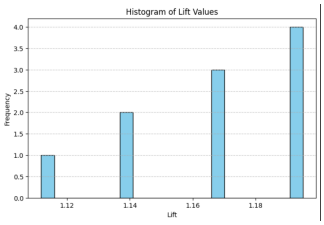

Figure 36: Lift histogram - state-level

Based on this apriori model, it seems that (PFHxS, PFHxA, PFBS, PFPeA) and various permutations of those compounds are more likely to appear together than if they were independent, as their lift values are all greater than 1 in the facility, county, and state-level models. If we were to look at 5-itemsets, it seems that PFHxS, PFPeA, PFHxA, and PFOS all tend to occur together - the only two that aren't appearing together as much are PFBS and PFOA in a 5-itemset. PFBS and PFOA do occur together in various permutations in 4-itemsets, however. These 6 contaminants seem to be the main PFAS that contaminate our water (Figure 35). We then mapped these itemsets back to the states that have them. All of these contaminants are either byproducts of one another, are currently produced, or were very prolific when PFAS were first introduced (Figure 37).

```
Itemset ['PFBS', 'PFHxS', 'PFPeA', 'PFHxA'] found in states: ['AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'MA',
Itemset ['PFHxS', 'PFOS', 'PFOA', 'PFHxA'] found in states: ['AK', 'AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY',
Itemset ['PFHxS', 'PFPeA', 'PFOA', 'PFHxA'] found in states: ['AK', 'AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY',
Itemset ['PFHxS', 'PFOS', 'PFPeA', 'PFHxA'] found in states: ['AK', 'AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY',
Itemset ['PFBS', 'PFHxS', 'PFOS', 'PFHxA'] found in states: ['AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'MA',
Itemset ['PFBS', 'PFOS', 'PFPeA', 'PFHxA'] found in states: ['AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'MA',
Itemset ['PFOA', 'PFOS', 'PFPeA', 'PFHxA'] found in states: ['AK', 'AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY',
Itemset ['PFBS', 'PFOA', 'PFOS', 'PFPeA'] found in states: ['AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'LA',
Itemset ['PFHxS', 'PFOS', 'PFOA', 'PFPeA'] found in states: ['AK', 'AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY',
Itemset ['PFBS', 'PFHxS', 'PFOS', 'PFPeA'] found in states: ['AL', 'AZ', 'CA', 'CO', 'CT', 'DE', 'FL', 'GA', 'IA', 'ID', 'IL', 'IN', 'KS', 'KY', 'MA',
```

Figure 37: States that have the common contaminant occurrences

This state-level map (Figure 38) shows that the most common PFAS contaminant combinations are occurring basically throughout the continental United States, which more or less matches the county map. However, while the county map shows that some counties in Wyoming, Montana, Louisiana, etc. still have these itemsets, the state level map shows that Wyoming and Louisiana barely have any, and Montana, Nebraska, Arkansas, Mississippi,

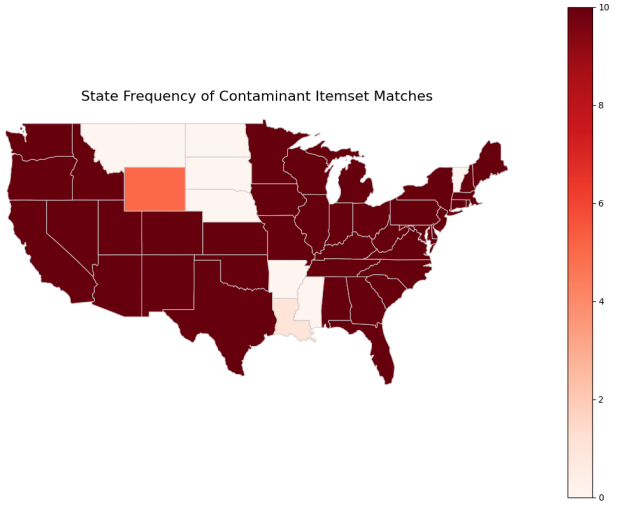

Figure 38: Map of states in the continental US of contaminant itemset match

North Dakota, and South Dakota do not have these combinations at all. Wyoming, Nebraska, Montana, and the Dakotas are states with relatively low populations, which perhaps is reflected by the PFAS levels in their water.

Louisiana had unexpected results since there are a lot of oil and chemical plants. We would've expected LA to also have those sets of contaminants. Similarly, we did not expect Arkansas and Mississippi to not have any of those sets either. These three states have industry, have a lot of shared waterways, and are near the Mississippi River in a location that is downstream from many other states. Even if they don't produce those PFAS chemicals specifically, we expected that runoff would bring the contaminants there anyways. We checked the dataset to see if the dataset was imbalanced and found that Colorado, a state with all the combinations of the contaminants, did have more samples that exceeded the MCL than Arkansas and Louisiana, for example. However, when checking the total number of occurrences for those states, Louisiana had even more samples than Colorado - they just had less that were above the MCL.

Some potential reasons for this might be that the PFAS that do occur in states like Louisiana, Montana, etc. are not the popular co-occurring ones that we found. There are a lot of other more obscure PFAS compounds that are in the dataset. Another reason is that there can also be differences in sampling methodology between states over the years. The EPA seems to only sample every decade or so, and a lot can and has changed over time in environmental policy and protection.

4.2.5 Model Comparison. Diving deeper into the identified compounds, out of the 29 PFAS in the dataset, the following 6 appeared the most in water samples: PFHxS is used in surfactants, protective coatings, and metal plating agents. While its production is on the decline, it can also result from PFOS production or from the degradation of other PFAS chemicals. PFBA and PFBS also have typical PFAS properties such as being used in nonstick and stain resistant products, firefighting foam, etc. They are also a breakdown product of other PFAS. PFBS is still being produced by 3M as a safer alternative to PFOS [3][4]. PFHpA, PFHxA, PFBS, and PFPeA are labeled as "safer" alternatives to PFOA and PFOS. However, these short chain PFAS still induced adverse effects in mice [5]. PFOA and PFOS are long (C8) PFAS and are the "traditional" PFAS that have been linked to numerous health and environmental issues. It seems that overall, the "safer" short chain PFAS appear together more often.

The best model here would be the county-level model as it more accurately portrays the association of the contaminants and is detailed enough to show differences within a state while not getting too into the weeds (like with the facility level model). It has high lift values to show the positive associations, and high confidence and support values while being able to be visualized geospatially to show all the hotspots in the country. This model and map would be more effective for policy making than the state, which is too generalized, and the facility one, which is too localized.

4.3 Clustering

For the clustering models, we looked to answer the following question: can we group facilities based on water system characteristics, contaminant presence, and local socioeconomic factors - and do the clusters reveal patterns of contamination risk or environmental injustice. A KMeans clustering algorithm was implemented to help answer this question.

4.3.1 Preprocessing. To preprocess the data for the KMeans clustering algorithm, we first read in the dataset and selected the variables of interest for this specific research question. In this case, we are curious about the facilities, contamination levels, regions, and socioeconomic factors (poverty data). Next, rows with missing information for any of these features were dropped, and a smaller sample of the dataset was used in the final model for faster processing. Next the categorical variables FacilityWaterType, Contaminant, and County were label encoded to adhere to the model, and the features were split into a training/test set and standardized.

	County	PWSName	Size	FacilityName	FacilityWaterType	SamplePointName	CollectionDate	SampleID	Contaminant	Region	...	Month
0	Juneau	CITY AND BOROUGH OF JUNEAU	L	Salmon Creek Reservoir TP	SW	Salmon Creek DS EP	2024-11-04	32403661003	PFNA	10	...	11
1	Juneau	CITY AND BOROUGH OF JUNEAU	L	Salmon Creek Reservoir TP	SW	Salmon Creek DS EP	2024-11-04	32403661003	PFUnA	10	...	11
2	Juneau	CITY AND BOROUGH OF JUNEAU	L	Salmon Creek Reservoir TP	SW	Salmon Creek DS EP	2024-11-04	32403661003	NEtFOSAA	10	...	11
3	Juneau	CITY AND BOROUGH OF JUNEAU	L	Salmon Creek Reservoir TP	SW	Salmon Creek DS EP	2024-11-04	32403661003	PFHpS	10	...	11
4	Juneau	CITY AND BOROUGH OF JUNEAU	L	Salmon Creek Reservoir TP	SW	Salmon Creek DS EP	2024-11-04	32403661003	9Cl-PF3ONS	10	...	11

5 rows × 22 columns

Figure 39: Snippet of Preprocessed Dataset

	FacilityWaterType	Contaminant	Relative_MCL	Median_Income	Child_Poverty_Rate	Poverty_Rate	MCL_Exceeded	County
422898	3	16	0.0	107032.0	6.6	6.6	0	295
362598	1	22	0.0	71753.0	16.9	13.5	0	405
722341	1	15	0.0	60268.0	23.2	19.4	0	754
342882	1	4	0.0	72065.0	12.3	10.5	0	731
1495346	1	15	0.0	56365.0	13.7	10.6	0	480

Figure 40: Snippet of Processed Dataset

4.3.2 KMeans Clustering. We want to use the elbow method on the data to determine the number of clusters that would be beneficial, and gather the most amount of information from the data. We can determine this by looking at the elbow plot, and finding the point where the line starts to flatten, indicating that the amount of clusters at that point is sufficient for explaining the information in the dataset.

Looking at the plot above, we can see that the line starts to flatten around 20 clusters. However, for the sake of interpretability, less clusters are used (k=15) as this value still encompasses much of the data and allows for more interpretable results. KMeans clustering with k=20 was tested, but did not change much of the outcome and did not provide much more information to the outcome.

Next the KMeans model is tested, and output is interpreted by the size of each cluster, and as well as the means of the variables of interest in each of the clusters. This data was able to give us some significant insights about, specifically, the socioeconomic relationship between the relative MCL and poverty/income rates.

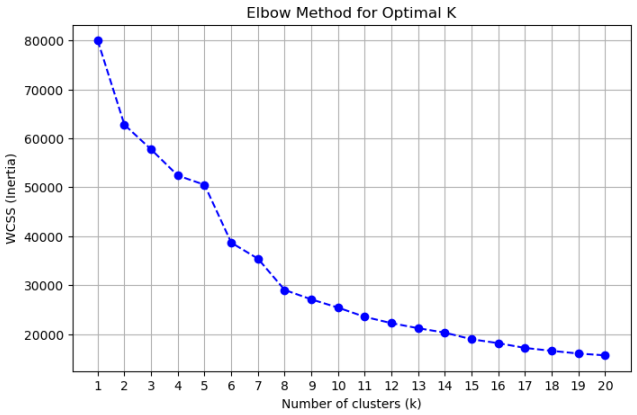

Figure 41: Elbow Method for Optimal K Value

Cluster	FacilityWaterType	Contaminant	Relative_MCL	Median_Income	Child_Poverty_Rate	Poverty_Rate	MCL_Exceeded	County
0	2.959849	4.929737	0.001161	81887.089084	14.240652	11.318946	0.0	561.616060
1	1.005875	18.977673	0.001146	53721.297297	24.594125	17.790364	0.0	751.823737
2	2.976562	18.609375	0.006758	81366.018750	12.887031	10.328437	0.0	214.417188
3	1.739130	17.408696	1.827391	89705.826087	14.273043	11.117391	1.0	542.217391
4	0.977891	21.221939	0.002105	79228.082483	13.681122	10.859439	0.0	756.721939
5	2.986190	16.376929	0.003412	53797.252640	24.067425	17.419171	0.0	518.680747
6	1.000000	13.473753	0.000984	88575.950131	10.323491	8.770079	0.0	240.346457
7	1.357143	18.357143	10.017857	63343.000000	20.585714	14.821429	1.0	407.071429
8	1.484211	16.389474	0.000000	42392.602105	36.752632	26.422947	0.0	485.720000
9	1.000000	23.500000	46.750000	80842.500000	15.400000	10.300000	1.0	469.500000
10	1.010526	6.442105	0.001391	60830.524812	22.584962	16.813684	0.0	415.362406
11	1.300455	13.616085	0.003718	122233.698027	7.282398	6.719575	0.0	786.330804
12	2.931507	20.276256	0.004737	80334.105023	13.756050	10.952854	0.0	800.771689
13	1.016393	6.816393	0.001885	79405.397814	14.748634	11.591694	0.0	778.759563
14	1.009732	20.467153	0.001125	60650.266423	19.788443	14.769100	0.0	258.891727

Figure 42: KMeans Model Output

After looking at these clusters and the means of each feature - there are some interesting observations related to income, poverty levels, and MCL levels. Cluster 9 only had two samples, so these are likely outliers. However, this cluster had an extremely high relative MCL of around 46.75, and also a higher median income. This *could* mean that contamination isn't strictly tied to poverty, but this is also a broad generalization and is likely more of a rare case. Cluster 7 also had a high relative MCL around 10.01. This cluster, similarly to cluster 8, also had a relatively high median income, indicating that poverty is not strictly tied to clean water, but this is rare in this dataset/clusters. Cluster 3 has 115 samples showing a more significant impact, and has a relative MCL value of 1.83 (not as high as for clusters 8 and 3, but large enough to exceed the MCL level). It also has a more moderate median income and poverty level, thus indicating relatively unclean water for middle class incomes. Cluster 11 is also interesting to point out, since it has a very high median income and very low poverty rate, while also maintaining one of the lowest MCL levels. This is likely the most "safe" water cluster. Cluster 5, on the other hand, had a high poverty rate and low median income. The MCL levels did not exceed the threshold, but were still relatively high.

Next, we want to evaluate the performance of this model, using the Silhouette Score and Davies-Bouldin Index metrics. Ideally, a Silhouette score (which measures how similar a point is to its own cluster) should be above 0.5 to be considered well-clustered, and the Davies-Bouldin index (which measures the average similarity of each cluster, compared to the cluster next to it) should be low.

The Silhouette Score and the Davies-Bouldin score indicate that the model isn't ideal. The Silhouette score is low, but indicated at least a moderate amount of well-clustering. The Davies-Bouldin score is also relatively high, indicating that there may be more similarity between clusters than what

```
from sklearn.metrics import silhouette_score, davies_bouldin_score

# y_kmeans holds the cluster labels
silhouette = silhouette_score(X_scaled, y_kmeans)
davies_bouldin = davies_bouldin_score(X_scaled, y_kmeans)

print(f"Silhouette Score: {silhouette}")
print(f"Davies-Bouldin Index: {davies_bouldin}")
```

✓ 0.5s

```
Silhouette Score: 0.22731248705602644
Davies-Bouldin Index: 1.0313603863443017
```

Figure 43: Code snippet Clustering Performance Metrics

would be expected. Because of this, it may be worth implementing a different model. It should be noted that KMeans clustering assumes spherical, equally sized clusters, which are relatively uncommon in environmental or socioeconomic datasets.

4.3.3 KMeans Visualizations. To answer the two other parts of the question of interest for clustering (impact of water type and the impact of region) the clusters were grouped and plotted first by the water type to assess any interesting patterns. To streamline interpretability, the clusters mentioned above of interest relating to income metrics (clusters 9, 7, 3, 11, and 5) are what are analyzed more closely in this section.

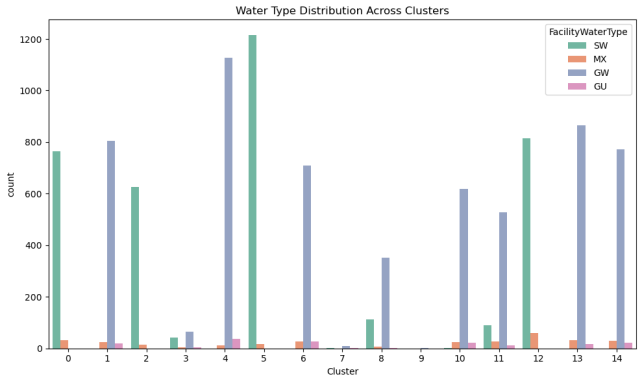

Figure 44: Water Type Distribution Across Clusters

From the plot above, it should be noted that clusters 9, 7, and 3 have ground water as the highest influence of water type. These clusters are the ones that also had the highest MCL contamination levels, as well as moderately low income and moderately high poverty values. The other clusters of interest, 12 and 5, which had some of the highest income rates as well as low MCL contamination, both have surface water as the most significant water type. Thus, there may be a small relationship between the type of water and the contamination level, which also has a relation to income or poverty.

4.4 Regression Model

4.4.1 Model Selection. We selected linear regression for this analysis because it is well-suited and straightforward for exploring quantitative relationships between continuous variables. In our case these are socionomic indicators, i.e. median income and poverty rate, and water contamination metrics. For the latter we selected the mean relative MCL and the ration of MCL samples, because the MCL data (threshold deemed to be save for health) has more real life explanatory value compared to MRL (detection threshold). Given the structure of our dataset, which contains clean numeric variables and a modest number of observations, linear regression provides an interpretable baseline to assess whether linear trends exist between

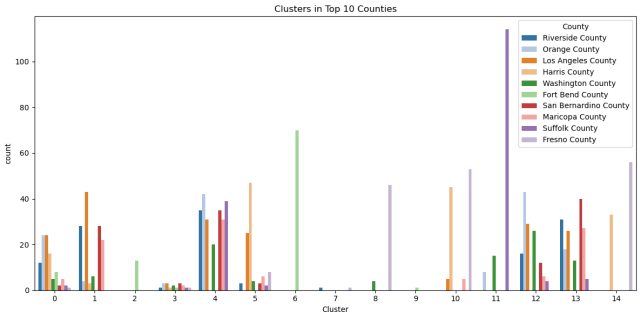

Figure 45: Water Type Distribution Across Clusters

these factors. It is also a widely accepted model for hypothesis testing in socio-environmental research contexts.

4.4.2 Data Preprocessing. Figure 46 shows the initial data set (columns, count, type), the pre-processing steps undertaken and the resulting dataset used for modelling linear regression. A data snippet of the resulting data is shown in Figure 47, the initial dataset is the one as described in the previous sections. The data size was vastly reduced.

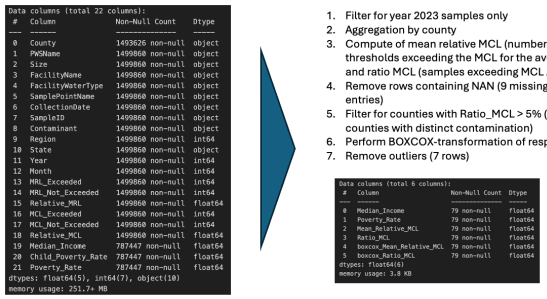

Figure 46: Data pre-processing steps taken prior data modeling

Median_Income	Poverty_Rate	Mean_Relative_MCL	Ratio_MCL	boxcox_Mean_Relative_MCL	boxcox_Ratio_MCL
54379	16.117642	0.205309	6.490943	0.716734	0.923167
52082	6.333927	2.199652	13.180087	0.823243	0.259966
112347	24.770528	0.230662	4.088826	1.031328	0.120861
113075	21.079308	0.498741	14.741578	1.104136	0.485138
93132	14.818420	1.904263	3.540076	0.480377	0.483248

Figure 47: Data snippet of finally processed data

4.4.3 Model Assumptions. Linear regression assumes that the relationship between the independent and dependent variables is approximately linear, that residuals are normally distributed, and that errors have constant variance (homoscedasticity). Additionally we look at influential outliers, as these can influence our model significantly.

Normal distribution of residuals (Boxcox transformed). Figure 48 shows histograms and corresponding q-q-plots for the aggregated county data. To better analyze the distributions we transformed them into log-scale. However this showed heteroscadicity issues later on. A transformation that yielded much better results was using a BoxCar transformation.

The top row shows the distribution of the mean relative MCL contamination (i.e. by how many MCL thresholds a typical sample is contaminated) and the bottom row the ratio of contaminated samples (as defined by simply exceeding the MCL threshold). Comparing the distributions both before

and after the transformation, we conclude that a boxcar transformation describes our data much (even though with spread out tails, as would be expected).

The data shown here is prior removal of any outliers.

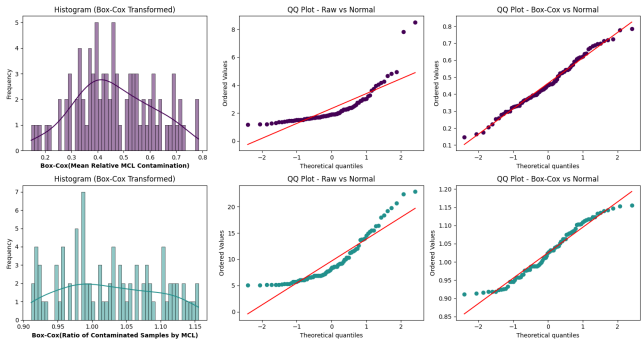

Figure 48: BOXCOX-Histograms and corresponding q-q-plots for normal distributions for MCL County data

Outliers (Cook's Distance). The Cook's Distance plots in Figure 49 for each of the four Box-Cox transformed regression models reveal a number of data points with high influence on model estimates. The red dashed line represents the threshold for potential influence. Most observations fall well below this threshold, indicating limited undue influence on the regression models. This suggests that while the overall models are robust, a few individual counties may disproportionately affect the slope and intercept. These influential observations were identified and removed to improve the model's generalizability and reduce the risk of overfitting.

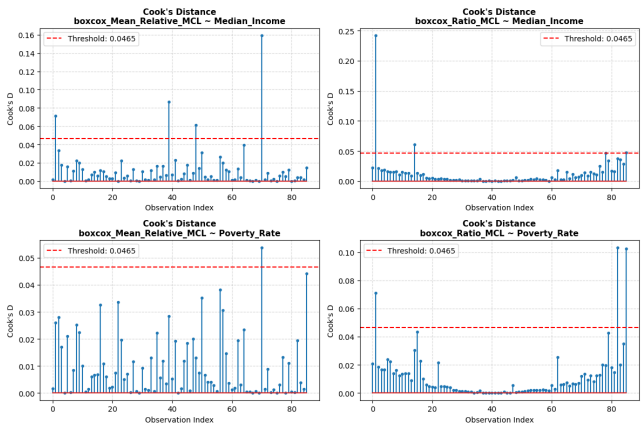

Figure 49: Outlier detection using Cook's distance on all for modelling plots

Linearity and homoscedastic assumptions. The residual vs. fitted value plots for all four linear regression models (using Box-Cox transformed target variables) show a significant improvement in meeting model assumptions. In each plot, residuals appear randomly scattered around the zero line without clear patterns, indicating that the assumption of linearity is reasonably satisfied. Additionally, the spread of residuals remains fairly constant across fitted values, which supports homoscedasticity.

Minor deviations in the boxcox_Mean_Relative_MCL Poverty_Rate plot suggest slight heteroscedasticity, but not to a degree that would seriously compromise model validity. Overall, the BoxCox transformation has improved the model fit tremendously and residual structure, supporting that using linear regression for this analysis is appropriate.

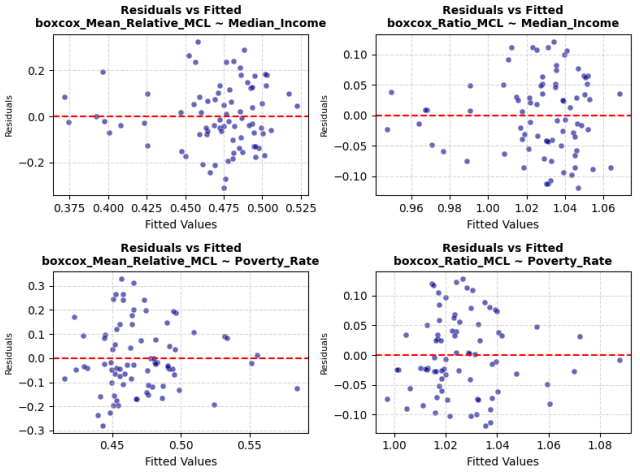

Figure 50: Residuals vs Fitted plots for all four models

4.4.4 Model Analysis. Figure 51 (left) presents mean relative contamination (in units of MCL) at the top and the percentage of samples exceeding the MCL threshold at the bottom. While the data exhibits high variance and a relatively low R^2, two key trends emerge: counties with higher median incomes tend to have fewer samples exceeding the MCL threshold and lower mean contamination levels. Conversely, counties with higher poverty rates show a greater incidence of contaminated samples and higher mean water contamination.

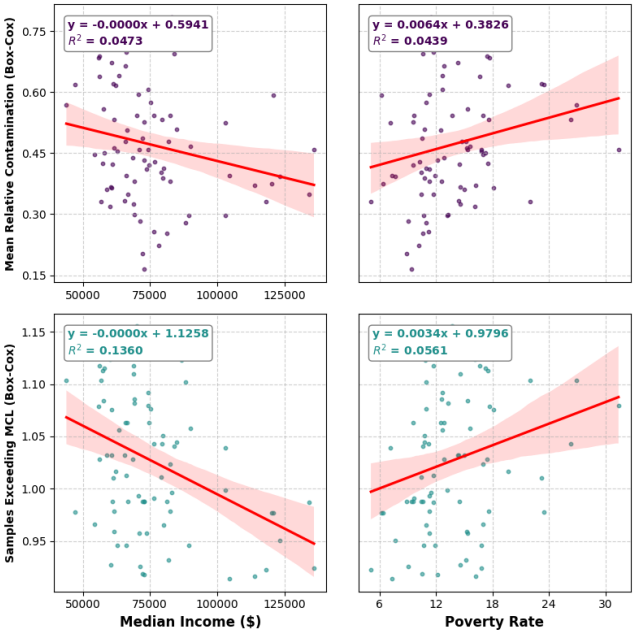

Figure 51: Linear regression between Contamination and Socioeconomic Factors

	Predictor (X)	Response (Y)	MSE	RMSE	R^2
0	Median_Income	boxcox_Mean_Relative_MCL	0.0194	0.1394	0.0473
1	Median_Income	boxcox_Ratio_MCL	0.0039	0.0628	0.1360
2	Poverty_Rate	boxcox_Mean_Relative_MCL	0.0195	0.1397	0.0439
3	Poverty_Rate	boxcox_Ratio_MCL	0.0043	0.0656	0.0561

Figure 52: Evaluation metrics corresponding to the plots of Figure 51

In Figure 52 the MSE, RMSE, and R^2 evaluation metrics are shown for all models. The highest R^2 value (0.1360) is observed for the model predicting boxcox_Ratio_MCL from Median_Income, suggesting that 13.6% of the variability in contamination exceedance (Ratio_MCL) can be explained by median income (lowest MSE and RMSE). The negative slope in this model indicates that higher median income is moderately associated with fewer samples exceeding MCL thresholds, hinting at possible environmental or infrastructural inequities.

In contrast, the relationship between Median_Income and boxcox _Mean_ Relative_MCL yields an R^2 of just 0.0473, indicating a weaker relationship where only about 4.7% of variation is explained. However, the trend still suggests a slight decrease in average contamination with increasing income.

For Poverty Rate, both models show weaker associations. The model for boxcox_Ratio_MCL has an R^2 of 0.0561, and the model for boxcox_Mean_ Relative_MCL shows the weakest of all at 0.0439. Both still show positive slopes, suggesting that higher poverty is mildly linked with more contamination and more frequent MCL exceedances, although the explanatory power is limited.

5 RESULTS AND DISCUSSION

For KMeans clustering, visualizing the clusters by region provides somewhat less information than by the water type, but still provides us interesting results. Clusters 9, 7, and 3 all had their highest County rate in California or Texas counties. The top counties for these clusters were Fort Bend (Texas), Riverside and Fresno (California) and Orange, Los Angeles, and San Bernardino (California). Clusters 11 and 5 vary more, with cluster 11 being primarily in Washington (Minnesota) and cluster 5 in Harris (Texas). Cluster 11 is particularly interesting here because it is the only one not in Texas or California, but rather Minnesota. This was also the cluster with surface water being the primary water type and one of the moderately high incomes.

For apriori, the main takeaway from this analysis is how widespread the long chain PFOA/PFOS and their derivatives are. Even as they degrade, they still exist in the water system in different forms. The county vs state level maps show that the state level map perhaps generalized the contamination patterns too much. There is still contamination in all those counties, but when clumped together, that contamination pattern gets diluted. Therefore, the county-level model provides the best insight into contamination patterns.

While the linear regression models show statistically weak but directionally consistent relationships, the negative correlation with income and positive correlation with poverty in contamination metrics reinforce environmental justice concerns — that economically disadvantaged areas are more vulnerable to water contamination. Despite the rather low R^2 values, these patterns may warrant further investigation using nonlinear models or incorporating additional contextual variables like infrastructure age, proximity to industrial activity, or regulation strength. The use of Box-Cox transformations also significantly improved model normality and residual behavior, enabling more reliable inference from these regression models. This is underlined by the low MSE and RMSE values for all linear regression models shown.

Conversely, when population statistics were examined in conjunction with other readily available information regarding sample water source, supervised machine learning models were able to extrapolate clear patterns. Specifically, a support vector machine trained using a radial basis function kernel – to account for highly nonlinear relationships between features – was able to predict the presence of hazardous water quality at a rate of over 0.90 accuracy. However, we found that during the fairness audit our model under-performed when identifying water samples that exceeded the MCL level in high-poverty areas.

6 CONCLUSION AND FUTURE WORK

Based on our analysis, it appears that PFAS contamination is still a major issue in much of the water sources across the continental US, whether it be the legacy long-chain compounds or its short-chain derivatives. All variants are detrimental to human health, and more severe contamination patterns are associated with lower income, highlighting a major environmental injustice. However, given the ability of predictive machine learning models to determine the likelihood of hazardous water samples, areas of greatest concern can be easily identified. This comes with the caveat that areas of greatest concern are not as easily identifiable in high-poverty regions, as a result models should be further enhanced by the incorporation of ensemble learning methods to reduce individual model biases. It will also be of paramount importance to track and understand water quality shifts that coincide with changing political legislation. Furthermore, continued research should be conducted to include modeling and analysis done for Hawaii, Alaska, tribal nations, and other US territories.

ACKNOWLEDGEMENTS

We gratefully acknowledge Alfonso G. Bastias, Ph.D., the U.S. Environmental Protection Agency (EPA), and the U.S. Census Bureau for their valuable support and publicly available data, which made this research possible. Public funding for research remains a vital foundation for scientific advancement.

REFERENCES

[1] U.S. Environmental Protection Agency. 2024. *Learn About the Unregulated Contaminant Monitoring Rule.* Retrieved April 2025 from https://www.epa.gov/dwucmr/learn-about-unregulated-contaminant-monitoring-rule
[2] U.S. Environmental Protection Agency. 2024. *General Overview Webinar Presentation: PFAS National Primary Drinking Water Regulation.* Retrieved April 2025 from https://www.epa.gov/system/files/documents/2024-04/general-overview-webinar-presentation-final-pfas-ndpwr.pdf
[3] Minnesota Department of Health. 2024. *Perfluorobutanoic Acid (PFBA) Information Sheet.* Retrieved April 2025 from https://www.health.state.mn.us/communities/environment/risk/docs/guidance/gw/pfbainfo.pdf
[4] Minnesota Department of Health. 2024. *Perfluorobutanesulfonic Acid (PFBS) Information Sheet.* Retrieved April 2025 from https://www.health.state.mn.us/communities/environment/risk/docs/guidance/gw/pfbsinfo.pdf
[5] Centers for Disease Control and Prevention (CDC). 2024. *Systemic toxicity induced by topical application of a perfluorinated compound in rodents.* Retrieved April 2025 from https://data.cdc.gov/National-Institute-for-Occupational-Safety-and-Hea/Systemic-toxicity-induced-by-topical-application-o/x7xw-nitb/about_data