4/12/24
Dr. Strange
DS2500
Team 125

Ethan Kurtz kurtz.et@northeastern.edu | Anzar Chowdhury chowdhury.an@northeastern.edu |

Brandon Gunasti gunasti.b@northeastern.edu

Public Water System Contamination Prediction

Problem Statement and Background:

Managing the quality of water traveling through public water systems is not an easy task. There are currently rules and regulations in place to ensure the quality of these water sources. One of these regulations is the Safe Drinking Water Act, which was passed by Congress in 1974. Under this act, the Environmental Protection Agency sets standards for drinking water quality and monitors the states, local authorities, and water suppliers who maintain these standards. The act outlines treatment requirements for over 90 different drinking water contaminants, as well as maximum levels for the contaminants in drinking water that must be monitored.[1] In order to make sure that the water quality is in line with the Safe Drinking Water Act the Environmental Protection Agency and States review and evaluate analytical results of water samples collected by public water systems. Using the data from the reports it can be ensured that drinking water standards are being met. When results show that a contaminant level exceeds the standard, the state and Environmental Protection Agency work with the public water system to take steps to remove contaminants and notify consumers to reduce harm to the public.[2] This process in itself is

[1] "Drinking Water Standards and Regulations," Centers for Disease Control and Prevention, August 10, 2022, https://www.cdc.gov/healthywater/drinking/public/regulations.html#:~:text=Under%20the%20SDWA%2C%20EPA%20sets,contaminants%20in%20public%20drinking%20water.
[2] "Safe Water Drinking Act Monitoring," EPA, accessed April 3, 2024, https://www.epa.gov/compliance/safe-drinking-water-act-sdwa-compliance-monitoring.

time-consuming and takes a lot of government resources. In addition to the Safe Drinking Water Act, there is also the Fifth Unregulated Contaminant Monitoring Rule.[3] This rule is much more recent, published on December 21st, 2021. This rule posits that once every 5 years, the Environmental Protection Agency must publish a list of unregulated contaminants to be monitored by public water systems. This sample collection is required for 30 chemical contaminants between 2023 and 2025 using analytical methods developed by the Environmental Protection Agency. This act is meant to inform the Safe Water Drinking Act better and update its list of monitored contaminants.[4]

This process is both time-consuming and resource-draining. The data collection takes place quarterly over 5 years, too long considering it determines the safety of the water the public drinks in their day-to-day lives. A better solution to this could save both government time and money, freeing up precious government resources to focus on other issues. In recent news, the U.S. government has also allocated an additional $1 billion to states for public water testing specifically to look at the "cancer-causing chemicals" that occur in water. These chemicals are called PFAs and are extremely harmful to people.[5] They account for 29 of the 30 chemicals that are currently being tested under the Fifth Unregulated Contaminant Monitoring Rule. With this extra push from the government to improve monitoring of these chemicals, it is more important now than ever to use data science approaches to better tackle this problem in a cost and time-efficient manner.

[3] https://www.epa.gov/dwucmr/fifth-unregulated-contaminant-monitoring-rule
[4] Ibid
[5]https://www.reuters.com/world/us/us-sets-first-standard-curb-forever-chemicals-drinking-water-2024-04-10/

Introduction to Data:

The data we are using comes directly from the Environmental Protection Agency collected under the Fifth Unregulated Contaminant Monitoring Rule. The most recent large update to the data took place in January of 2024 and will be updated quarterly until 2026. Current data only makes up 24% of the total data planned to be collected. The data is also updated regularly based on further review by analytical laboratories, public water systems, states, and the EPA. The data more specifically contains a list of different water systems and corresponding contaminant tests. As 30 contaminants are being tested each system is listed that number of times. It is once again important to note that these contaminants aren't required to be tested by the Safe Drinking Water Act, but are instead candidates to be added. There are no ethical or privacy concerns about how the data was collected, as all was done in line with government regulations and following government acts.

Data Science Approaches:

Parsing and cleaning data

The methodology for our approach initially involved downloading the aforementioned EPA-sponsored chemical data and implementing a pandas data frame to parse and structure the data. This separated each variable located within the data, such as each state, each contaminant, and the corresponding results of the tests to determine if each contaminant was found to be present in a water source. Next, to clean the data and simplify the variables that we worked with, we proceeded to drop several columns within the data that mostly comprised administrative IDs that had no relevance to our calculations. This included AssociatedFacilityID,

AssociatedSamplePointID, and 'UCMR1SampleType. For simplicity, we then created an additional column that directly states whether or not each test was able to detect a contaminant.

Visualizations

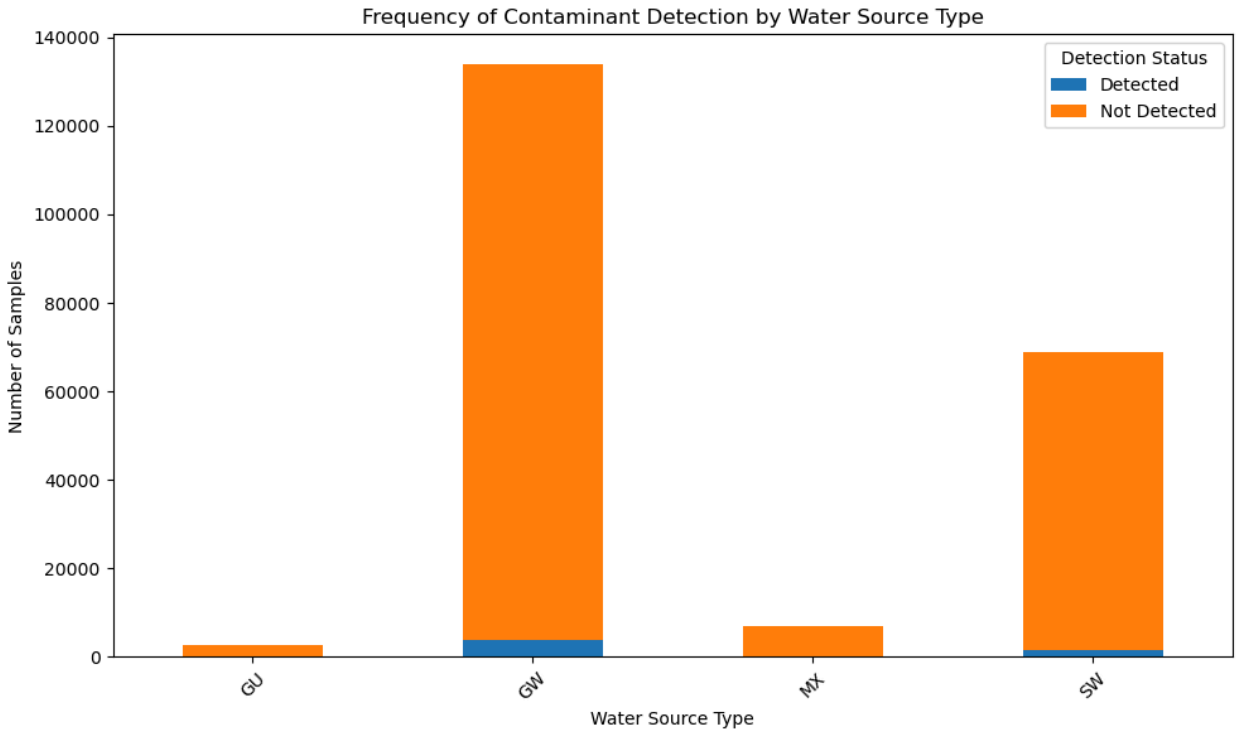


Figure: Frequency of Contaminant Detection by Water Source Type (GU: Groundwater under Surface Water, GW: Groundwater, MX: Mixed Sources, SW: Surface Water)

We used the plt.pyplot library to create a visualization the demonstrated the frequency of contaminant detection by water source. Each bar within the bar graph represented the total number of tests conducted on a given type of water source, regardless of all other factors. The share of the tests that resulted in positive detections are colored blue and the tests with negative results are colored orange. The four types of water sources include groundwater under the direct influence of surface water, groundwater, mixed water, and surface water. The Y-axis represents the number of tests, or samplings done on each water source.

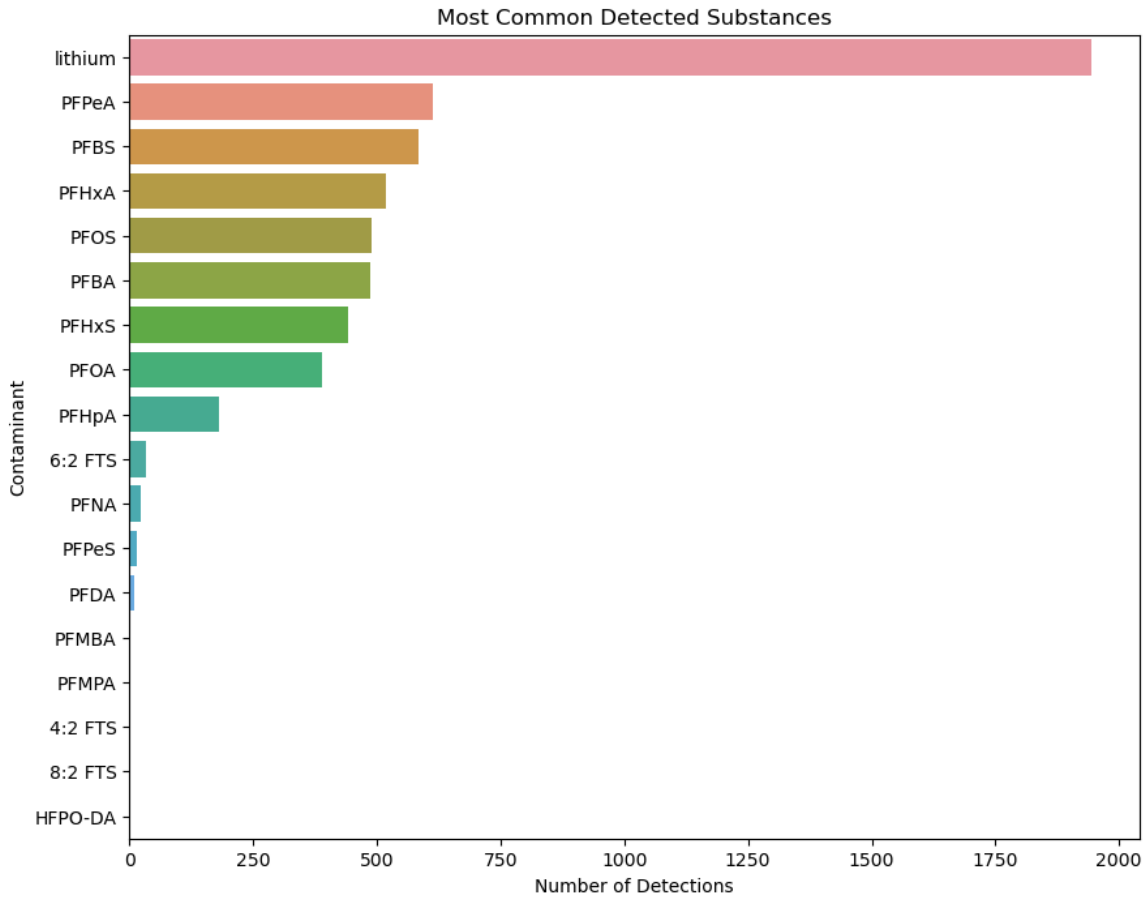


Figure: Most Common Detected Substances (all except Lithium are PFAs)

Next, we used the plt.pyplot library to create a bar plot that visualized the number of detections for the contaminants. We did this by obtaining the counts for the contaminants, as well as utilizing the index, or labels assigned to them in the function, sns.barplot(x=contaminant_counts.values, y=contaminant_counts.index).

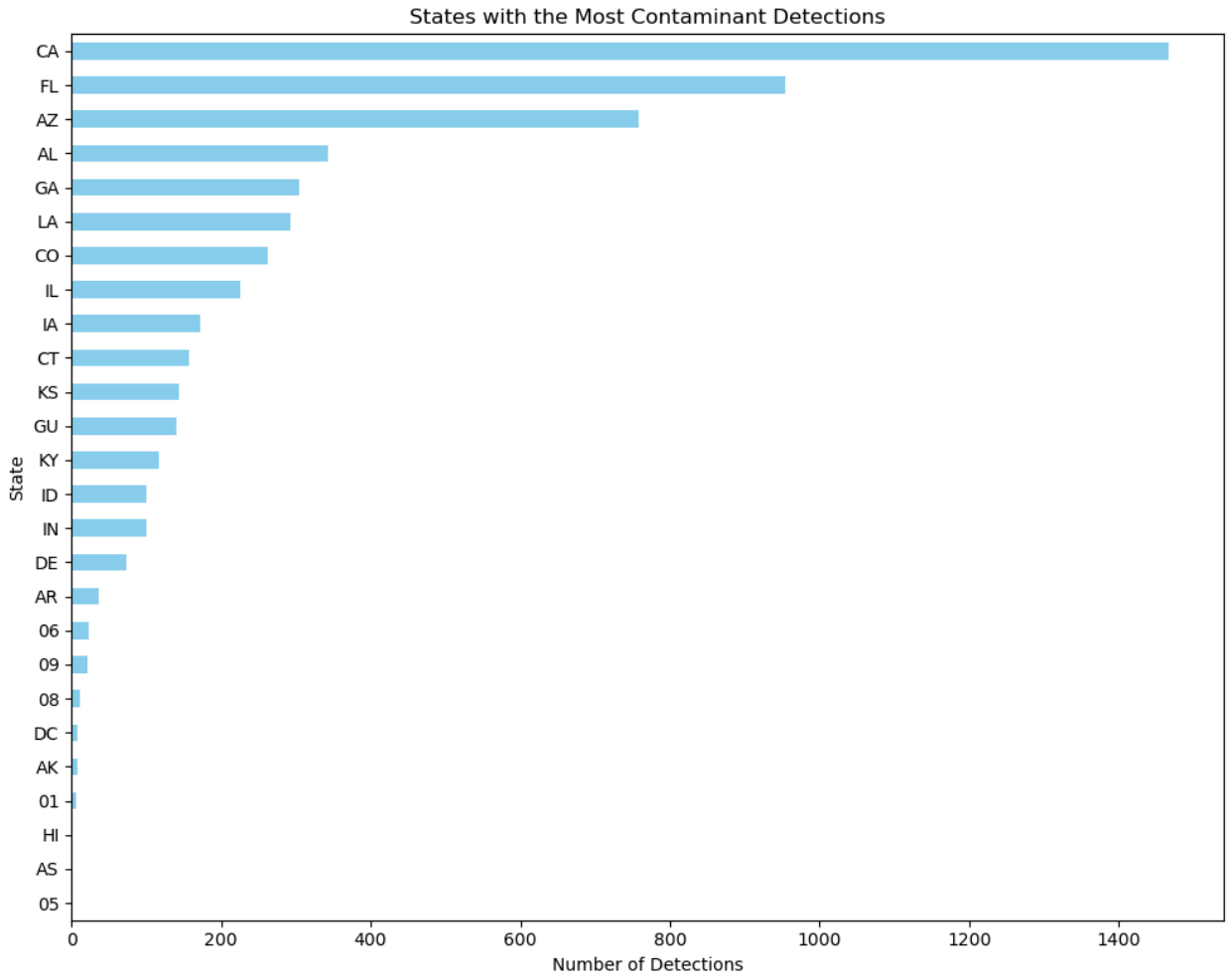


Figure: States with most detections (the numbers refer to Tribal PWSs which are tied to an EPA region)

Using a very similar methodology, we then visualized the states with the highest quantities of contaminant detections by utilizing a state_counts variable that comprised of detected_df['State'].value_counts().

Logistic Regression:

To fulfill our aim to develop a model that can predict whether a water sample from a particular public water system (PWS) will contain a specific contaminant at levels above the minimum reporting level (MRL), we used LogisticRegression for classification. Given that we want to find whether a specific contaminant will be detected, we will have to run the model for every unique contaminant. Additionally, given that we predicted a binary outcome, that is, whether a specific contaminant is detected or not LogisticRegression is an appropriate algorithm to use. Additionally, logistic regression not only classifies outcomes but also provides probabilities for the predictions. This is particularly useful in risk assessment and decision-making processes where understanding the likelihood of contamination is essential. Logistic Regression also offers straightforward easily interpretable results which are valuable for stakeholders who need to understand the factors influencing water quality.

We then created a binary column for each unique contaminant that had a 1 if that specific contaminant was detected. As this made our previous detection column redundant, we proceeded to drop it from our data table. We added a column for collection month to account for seasonal differences, as the concentration of particulates within bodies of water are constantly influenced by seasonal factors that may skew the data. We then dropped instances where the contaminant was detected less than 100 times to ensure that the logistic regression model was train on a significant amount of data for each contaminant to ensure that the model produced clear and conclusive results without being influenced by extremely minimal outliers. Finally we created a function that calculated the logistic regression with our x parameter being the featured dataframe and the y parameter being the target series. We encoded the categorical features that we wanted to use as variables to apply for their given correlations to predicted contaminants. This included

['FacilityWaterType', 'Region', 'State', 'SamplePointType', 'CollectionMonth'] and split the dataset into both training and test sets. Finally, we calculated a confusion matrix using the code confusion_matrix(y_test, y_pred) to determine if our predictions were correct, or incorrect.

***Results**:*

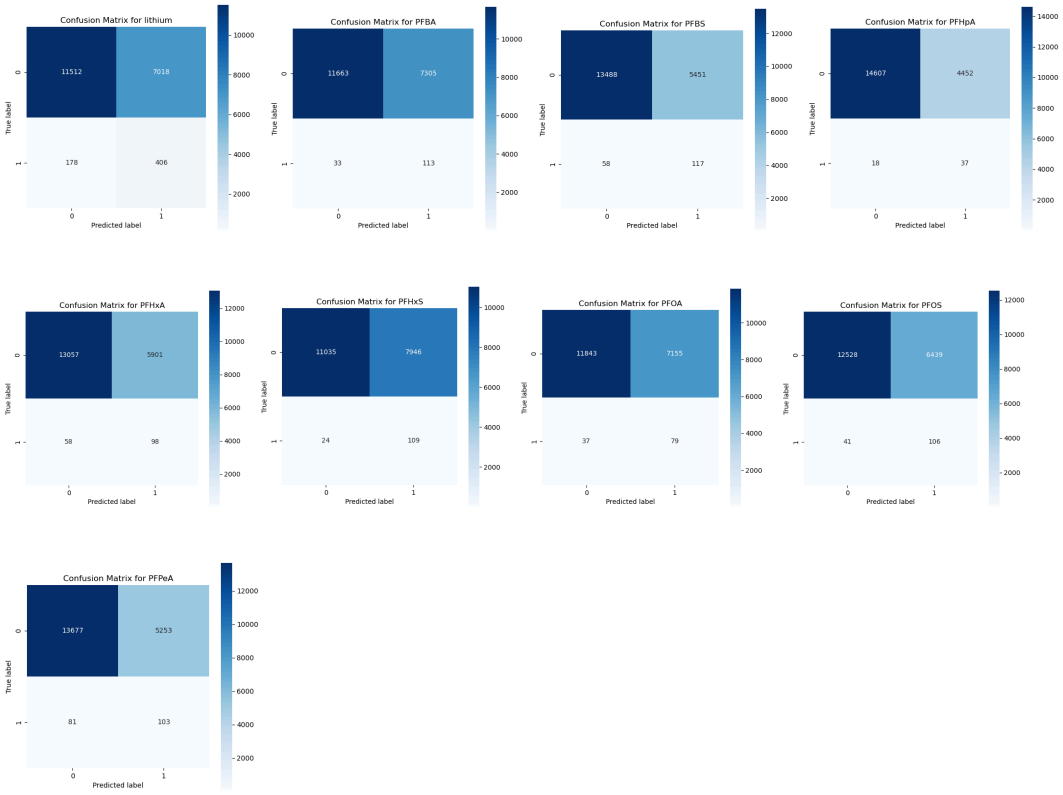


Fig: confusion matrices for all the contaminants

Given that we had run a model for each of the contaminants present, we will be giving an overview regarding the performance metrics across all contaminants.

The support values indicate a significant class imbalance for all contaminants, with the majority class (0, or not detected) having much larger sample sizes compared to the minority class (1, or detected).

For the majority class (0), precision is consistently high across all models, indicating that when the model predicts a sample does not contain the contaminant, it is usually correct. For the minority class (1), precision is very low, showing that there are many false positives — instances where the model incorrectly predicts the presence of the contaminant. Recall varies across the models for both classes. For the majority class, it ranges from around 0.58 to 0.77, suggesting that the model misses some true negatives. For the minority class, the recall is higher, suggesting the model is better at identifying most of the true positives, but this is likely at the cost of incorrectly classifying many negatives as positives (low precision).

The F1-score is relatively high for the majority class due to the high precision and moderate recall. For the minority class, the F1-score is very low across all contaminants. This indicates a poor balance between precision and recall — the model struggles to predict the presence of contaminants accurately.

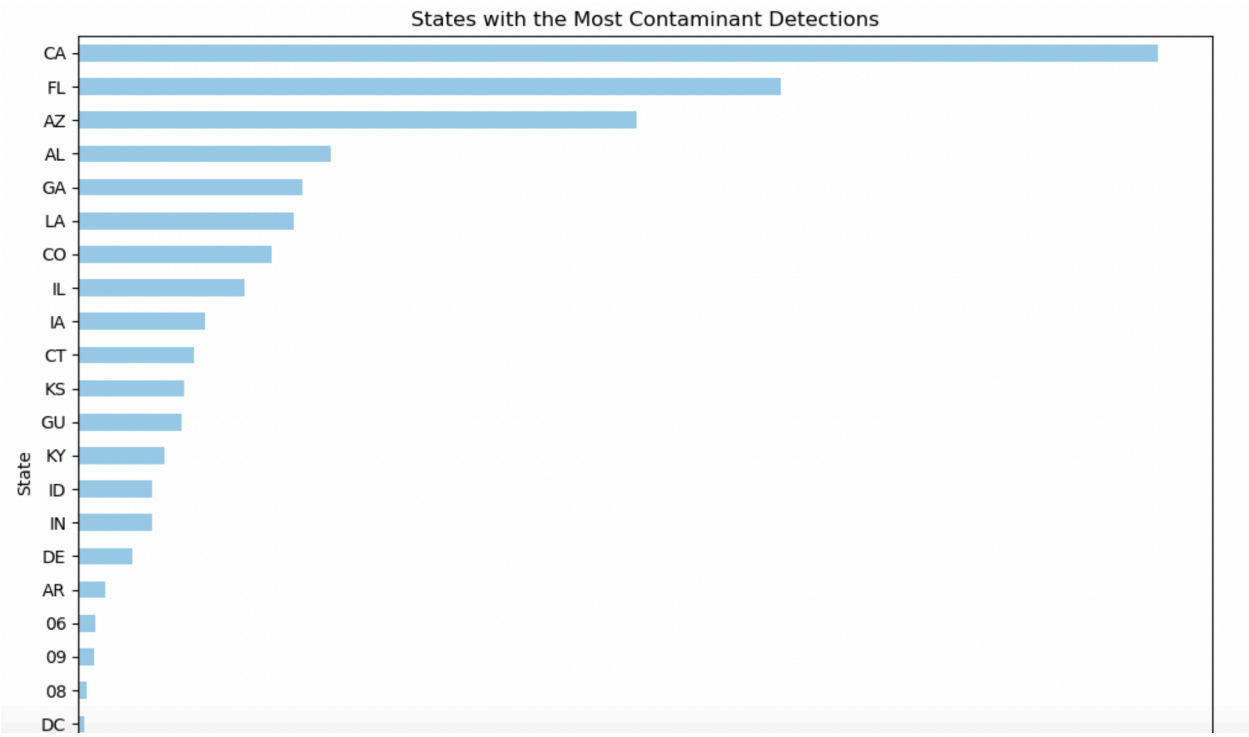
The overall accuracy ranges from 0.62-0.77. Meaning that 0.62-0.77% of the the predicitons were accurately guessed by the model for their corresponding contaminants. Accuracy alone doesn't provide a full picture of the model's performance due to the dominance of the majority class.

Macro averages do not account for class imbalance and treat all classes equally, resulting in much lower average precision, recall, and F1-scores compared to the weighted averages, which consider the support of each class. Weighted averages are higher because they reflect the dominance of the majority class in the dataset, making it the most important metric for the models. (The model uses a 'balanced' class weight to address the class imbalance, which improves recall for the minority class at the expense of its precision.)

Future Work:

For future work, we would like to work with an updated version of the data set. Due to the fact that there are so few contaminants detected in the data set the model has difficulty predicting when there will be contaminants detected. Working with an updated and more complete data set would help regression be more accurate. It would also be useful to check how EPA funding impacts the detection rate of contaminants. It would help understand where funds may be being misallocated, and which states might need more funding. Furthermore, if we were given access to the breakdown of the way the budget is being used and conducted a cross analysis with the contamination rates of water systems it would give insight to how to better use government funds at a more specific level.

Figures:


States with the Most Contaminant Detections
State
CA
FL
AZ
AL
GA
LA
CO
IL
IA
CT
KS
GU
KY
ID
IN
DE
AR
06
09
08
DC