5/5/2022

Andres Almaraz & David Luis Hiraldo-Panchana

Client: Fred Sieling

PFAS/Teflon

# Team Members

## Andres Almaraz

Data Analyzer

Visualization

Project Scribe/Documenter



## David Luis Hiraldo-Panchana

Data Modeler

Visualization

Team Manager & Client Liaison

# Abstract

Our goal was to find trends that will allow us to find PFAS/Teflon in US and GA. It is a manufactured chemical that can cause serious health issues if it enters a person's blood system. Teflon which can contain a chemical of PFAS and can be found on materials like nonstick pan or water repellent clothing. The chemical can never fully go away since it stays wherever it is left whether it is in the air, water, or someone’s body. In addition to our goal, we want to be able to prevent any harm to people from PFAS at any site and find the offenders causing it.

## Technologies Used

|  |  |
| --- | --- |
| Python | Programming Language |
| GitHub | Storage |
| Jupyter Notebook | Statistics |
| Tableau | Visuals |
| Deepnote | Collaboration Notebook |
| Bootstrap | Web development tool |

## Hypotheses

We as a team determined four hypothesis that we can be test based on the data we had from the datasets. They were split between the two of us with Andres overseeing the first two hypotheses while David overseeing the last two.

1. Is there an indicator in finding PFAS in the water?
2. If there is a rise of people exposed to PFAS across the US
3. A correlation between number of violations and the number of sites visited within those counties in Georgia
4. That there is a facility that is causing the rise of PFAS in Georgia

# Iterations

## First Iteration

For the first iteration we focused our main effort on finding effective datasets, we landed with six datasets could be used for the hypotheses. At the time the hypotheses used to be only focused on Georgia but that changed later. A plan was made for the flow of our project. The datasets we had were then in the process of cleaning and explored for the hypotheses at that time.

## Second Iteration

The second iteration focused more on supporting our hypothesis with visualizations. One major issue however was found which was that Georgia did not have enough data to evaluate the hypotheses, so we shifted our first two hypothesis into looking at the US. Two of our original datasets were dropped by this point since they did not contain enough information and were too complicated to understand. We then transferred from Google Colab which was our original collab notebook to Deepnote. Deepnote was able to hold our dataset better and allowed us to work together more efficiently. Changes from and to the datasets allowed us to create clearer visuals which gave better information. An example being that one of the datasets were reformatted into months to make it clearer which helped with seeing the varying growth of the chemicals.

## Third Iteration

The third iteration was implementing algorithms into our datasets while also creating a visual presentation for the CREATE and client. A website was made, and updates were done to our notebook and GitHub with the results we made from the semester. A few issues were discovered when making the results was that PFAS is a group of chemicals. Originally, they were all grouped together but we realized that everyone in the group could have varying results, so PFAS was broken down even further to better test our first hypothesis. We were able to make better results because of that.

# Implemented Features

* Linear regression was completed with every part of PFAS with non-PFAS chemicals
* The facilities, violations, and populations in Georgia was done with K means clustering
* Multiple interactive plots were done and provided a visual understanding of the hypotheses

# Future

One of our original goals was to set up a predictive model that could tell how much PFAS could be found in the water systems. If this was to be continued by us or another team, that should be implemented. We feel that more algorithms would have had to be used to complete that goal. Another option to do is to use more current datasets since regulations have changed from when the dataset was released in 2016 and more variations of PFAS is being recorded then before. Due to time, we were unable to connect the datasets to see if there was a connection between the factories and the PFAS that was found. This is one of the last objectives that can be done if picked up since we found that there was a connection between the datasets.

# Known Issues

Our data required us to restructure or reformat it multiple times to find the results we were looking for. There is a chance that we may have overlooked a key part because of that.

One issue we also had was that there was information from datasets we found but did not understand since we do are not chemist. We missed data that could have been used for testing.

# Results

## Is there an indicator in finding PFAS in the water?

Chart, scatter chart

Description automatically generated

What we found we assessing this hypothesis was that out of all the types of the recorded PFAS there was one that had a correlation. That was PFOA which can sometimes be found in Teflon in tiny amounts. Then the chemicals seen to correlate was vanadium, chromium-6, and chlorate and the only ones to have a R2 above 0.6. The closer the value is to one, the better the fit or connection between variables. That leads us to say that there an indicator to reveal or use for predicting one of the types of PFAS.

Chart, scatter chart

Description automatically generated Chart, scatter chart

Description automatically generated

## If there is a rise of people exposed to PFAS across the US

Chart, bar chart

Description automatically generated

This was the most we were able to find out when we were evaluating this hypothesis. Though we did obtain a result, we feel the difference is too insignificant and there is not enough information to truly tell if there has been a change. We learned this could be because when PFAS is found in one’s body, it takes four years to go down half it is size. Due to that, we would need more years to assess this hypothesis.

## A correlation between number of violations and the number of sites visited within those counties in Georgia

Using the K means scatter plot we were able to find that there is a relationship between facilities, violations, and populations in Georgia. We also discovered that lower populations equal more facilities.

## That there is a facility that is causing the rise of PFAS in Georgia

Chart

Description automatically generated

When researching this hypothesis, we were able to find that there were two offenders for most cases of PFAS in Georgia. The two being Dalton County and Bartow County. We had learned from our client that Dalton County has a rug manufacturer that was known for being a part of the cost of this.

# Conclusion

Even though we were able to find results for our hypotheses, there is much more that could have been done with it. Like mentioned before our data is older with older regulations which have then changed significantly. With newer data we would be able to better find indications and tell more people are being contaminated. However, we were still able to find correlations for PFAS that could be used to prevent any from entering the water which was the major focus for us.

# Links

[GitHub](https://github.com/GGC-DSA/pfas)

[Deepnote](https://deepnote.com/workspace/andres-al-71a4-641a0af0-55f4-45b5-b3cd-3b9eadbe9894/project/Capstone-1c411a1a-39b3-482a-9f41-b53f24acaa96/%2FCapstone.ipynb)

[Website Preview](https://youtu.be/xrL_Gv2wUks)