PFAS/Teflon where is it?

By: David Luis Hiraldo-Panchana and Andres Almaraz

Client: Fred Sieling

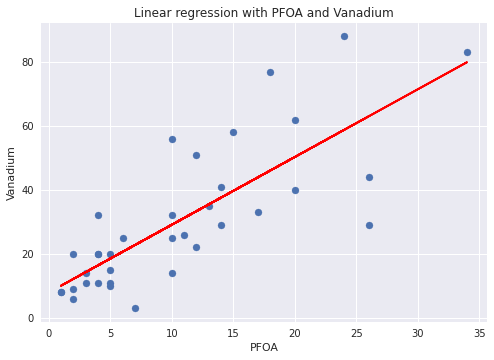
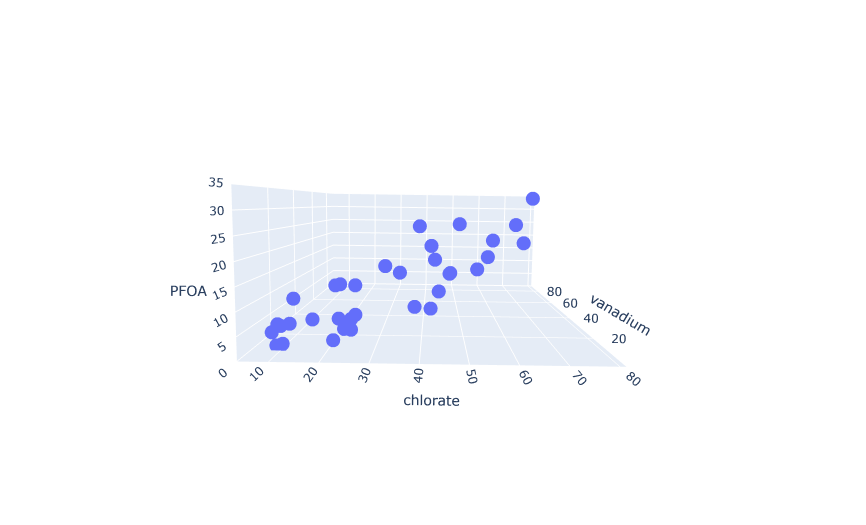
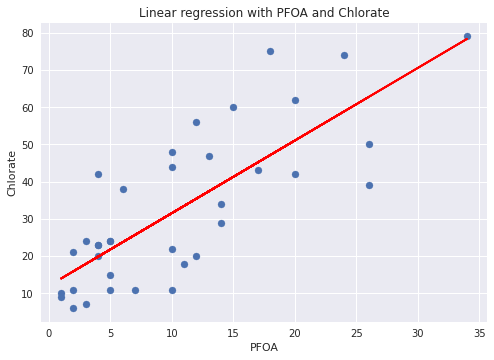
The goal of our data analytics capstone project is to see any trends or anything that will allow us to find where pfas/Teflon in GA have been and will be. This will allow us to show any future sites that may have pfas in the water system. We are using previously found data to find any correlation to possible sites that may have been overlooked. Our data sets have two primary focus areas: those being the actual chemical compound of Teflon and the factories. We will be pooling our data to make an interactive website that will locate any future sites. We can do this by tracking the factories that cause the most pollution, and also revisiting the sites that had a null value for the water containment level. We as a team have decided to take certain roles within the project, David focused around being the data modeler and visualizations a data modeler worked on implementing the databases and working with the data structures. Andres focused on being the data analyzer which focused on putting in the algorithms for our data bases while also working on visualizations as well, visualizations focused on implementing visuals do the data.

The first iteration of our project we focused our main effort on finding an effective data set, we landed with four data sets that we felt that we could use in our project to prove some of the hypotheses that we came up with. The first that we are testing is the correlation between the number of violations and the number of sites visited within those counties. Another hypothesis can be, does population affect the number of violations that occurred on site? Third hypothesis is to see if there are chemicals from dumping that hints the future exposure of pfas into the water. The last one is to see if there is a growth in people exposed to pfas in contaminated locations. We also discussed what technologies we are going to use in the data. We decided that we would use Jira, GitHub, Jupyter for notebook, Google Collab, and Tableau. We also discussed the flow chart of our project, and we also added our GitHub. We did however run into some problems with our dataset. We noticed some null values and it was locked into a time set which is 2016.

The second iteration of our project we made some changes to our GitHub, we then changed from google collab to Deep note because deep note ran our dataset easier and allowed us to combine both of our notebooks with less hassle. The second iteration focused more on supporting our hypothesis with visualizations. We implemented more graphs to show the progress we made within our notebooks. Some of the issues we ran into were that some of our data was formatted poorly but it was reformatted to be clearer.

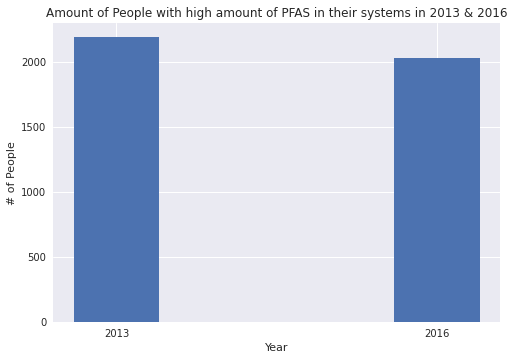
The last phase of our project is wrapping up everything, we implemented algorithms into our datasets, David focused on clustering while Andres focused on linear regression with his data set. We also had to begin updating our GitHub with everything we changed and start compiling everything we needed to make our final report. We also are finishing up our website that will have an interactive map and our poster which will be ready for our presentation.

Main results with hypothesis we found some correlation with certain chemicals and pfas, also there seemed to be a negative correlation with the exposure of people that was contaminated with pfas but that could be a problem with our data since the data only calculated pfas within the system of someone that had levels above .2 which recently was found out that it is dangerous to have .1 in your system. We also found a small connection with the population and the number of facilities that the county had.

So to add onto the paragraph above we listed the hypothesis now we will show you our findings, The first question we asked was, Is there a is an indicator that can be used in finding PFAS in the water? And we found Chlorate, Vanadium, Chromium-6 are indicators of PFOA which is a byproduct of Teflon. The graphs the support said hypothesis are the graphs 1,2,3.

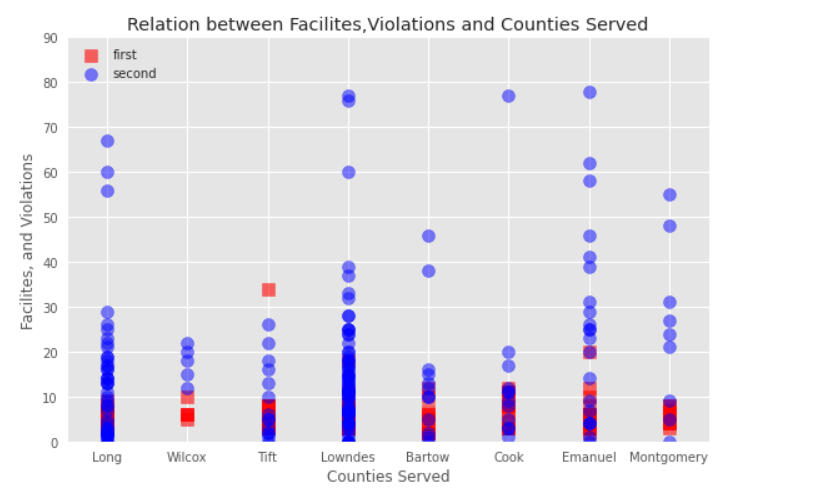
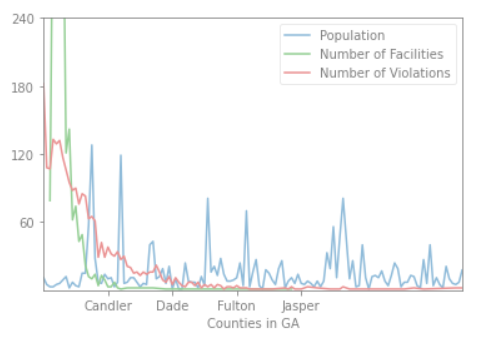
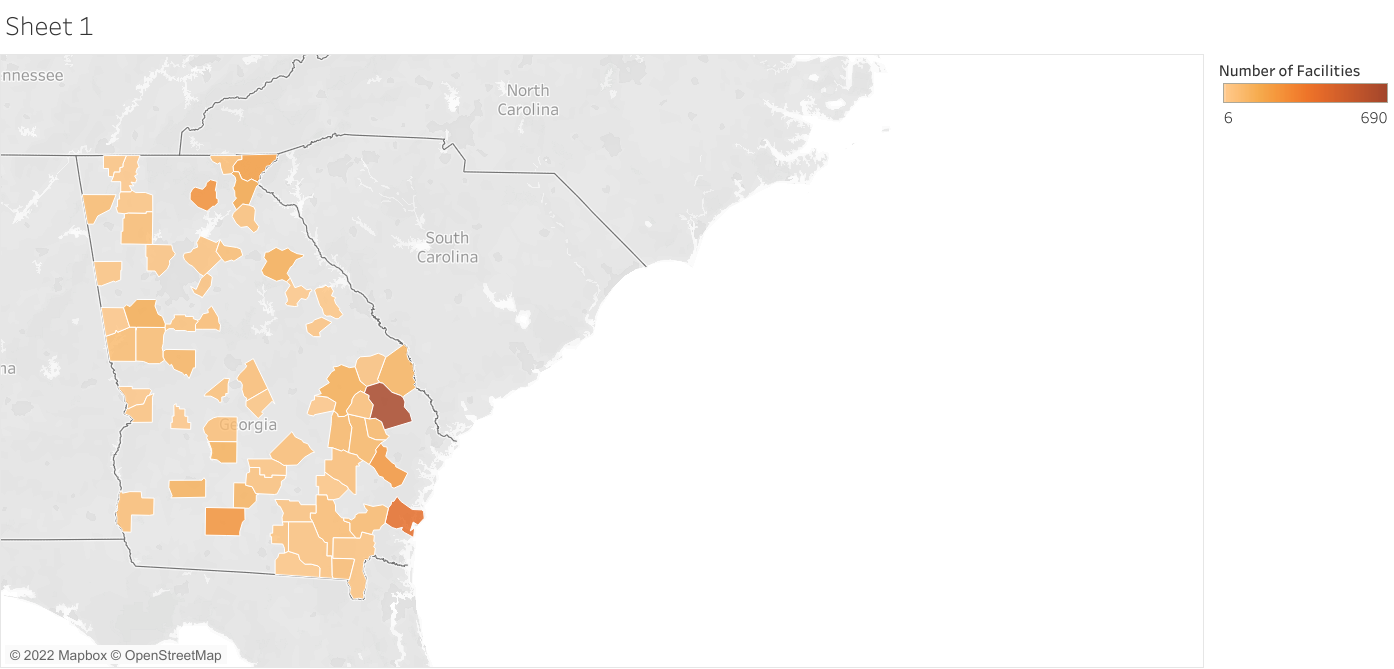
Graph 1 Graph 2 Graph 3

We used linear regression to see if there was correlation with the chemicals shown and pfas/Teflon which the graph does show a fairly strong link between the chemicals and Teflon. Our second hypothesis was asking if people were getting exposed more to Teflon/pfas? Our conclusion was that with the information we had at our disposal we couldn’t draw a conclusion support by graph 4. We couldn’t come to a conclusion since we didn’t have an up to date picture it was hard to determine. Teflon is a forever chemical which is a chemical that shrinks its size by half which would eventually lead people to believe that it is out of your system when in reality it hasn’t left.



Graph 4

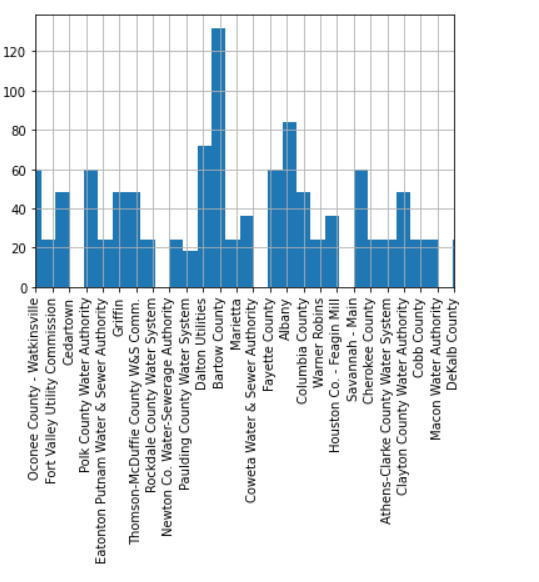
The next hypothesis focuses on Georgia specifically. The question asked was, Is there a relationship between facilities, violations, and populations in Georgia? Which the conclusion we found was yes There is a relationship between facilities, violations, and populations in Georgia: Lower populations equals more facilities. Support by our graphs 5,6,7.



Graph 5 Graph 6 Graph 7

These Graph show some of the relationship between facilities and population, I found by digging deeper into my dataset that counties within Georgia that had a low population size had a lot more facilities stationed there. Graph 7 I used K means clustering to see if there was any relationship with facilities and population which there was a strong case as well. Graph 5 just shows where the facilities are located within Georgia. This may however may not be the most accurate picture since its not the most up to date.

Our last hypothesis was asking Where did PFAS occur most in Georgia in 2016? Which we concluded that the two most active pfas cases occurred in Dalton and Bartow County. Which is supported by Graph 8.



Graph 8

I

With technologies we used such as deep note which allowed us to run our datasets, we found that some of our data had correlations with some other attributes. We used bootstrap to run our website which allows us to present our findings with our interactive maps. Another program that we used to help compile our visualizations is lucid press. We used Jupyter notebook just to run our datasets. We used GitHub as a central hub to put all of our data in one consolidated spot.

In the future we implemented 4 out of our 5 hypotheses, we could revisit our fifth hypothesis because as it stands our 5 hypothesis were not fruitful and didn't give us any usable results. If we had more time, we could find more up to date datasets which would provide us a better picture for our current times. If we found better data as well, hopefully we wouldn’t have as many null values that were put into the data set. If we had more time, we could look at other states more in depth giving each state a deep dive into which county is producing the most Teflon in those respective states. With more time we could produce different hypothesis that could provide different results. We could dig deeper in other states and focus on their counties as well and found out if they had any similarities to the facilities in Georgia We could dig deeper with more algorithms as well to find more correlation with the things we found.

