Beginning with PFA\_ENV data, I created histograms and a correlation matrix to consider the significance of the difference PFAs chemicals.

A collage of graphs

Description automatically generated

For variables such as PFOA-VA and PFOS-VA, there are noticeable outliers or extreme values far beyond the main concentration range. For instance, PFOA-VA shows a long tail that reaches values over 1000, while most of the data points are below 200. Do these represent areas of extreme contamination?

A diagram of numbers and a number

Description automatically generated with medium confidence

**Strong Positive Correlations: Do they share a source?**

PFOS-VA and PFOA-VA (0.94), PFDoA-VA and PFDA-VA (0.91),PFOS-VA and PFHpA-VA (0.93)

**Moderate Positive Correlations:**

6\_2 FTS-VA and 8\_2 FTS-VA (0.73), PFUnA-VA and PFTrDA-VA (0.74),

**Moderate Negative Correlations: Do they interfere with each other?**

PFOA-VA and PFHxA-VA (-0.50):

There’s a lot more positive correlations with negative correlations as expected. These chemicals are often present together.

A graph with numbers and letters

Description automatically generated with medium confidence

Applying k-means clustering, the PFOA-VA and PFOS-VA in cluster 0 stand out.

Cluster Comparison (Mean PFA Concentrations):

Cluster 0 1 2

TIME 1077.777778 1174.711538 1171.950355

4\_2 FTS-VA 7.800000 7.428846 8.177305

6\_2 FTS-VA 7.366667 7.503846 8.129787

8\_2 FTS-VA 7.800000 7.428846 8.172340

N-EtFOSAA-VA 19.666667 18.701923 20.357447

N-MeFOSAA-VA 19.666667 18.701923 20.236170

PFBS-V 10.777778 3.431731 3.937589

PFBA-VA 13.544444 6.171154 6.965248

PFDS-VA 3.900000 3.707692 4.118440

PFDA-VA 3.711111 3.688462 4.096454

PFDoA-VA 3.900000 3.707692 4.117021

PFHpS-VA 3.444444 3.579808 4.079433

PFHpA-VA 14.600000 3.298077 3.915603

PFHxS-VA 24.711111 4.140385 4.121277

PFHxA-VA 14.077778 3.572115 3.944681

PFNS-VA 3.900000 3.707692 4.117021

PFNA-VA 3.055556 3.725000 4.067376

PFOSA-VA 4.344444 3.707692 4.256028

PFOS-VA 31.166667 9.867308 5.546809

PFPeS-VA 3.922222 3.493269 4.146099

PFPeA-VA 11.433333 3.960577 4.053901

PFTeDA-VA 3.900000 3.707692 4.341844

PFTrDA-VA 3.900000 3.707692 4.343262

PFUnA-VA 3.900000 3.676923 4.097872

Cluster 0 seems to represent samples with notably higher contamination levels for a select group of PFAs, particularly PFOA-VA, PFOS-VA, and PFHxS-VA. These could be areas near industrial sources, wastewater sites, or regions with significant environmental pollution.

I merged the data for Wells and PFAS\_GEOSPATIAL to the initial dataset PFAS\_ENV to look more into cluster 0.

Correlation Matrix (Well Depth and Distances) for Cluster 0:

DEPTH 1.000000

OILREF 0.650000

OILGAS 0.620000

CHEM 0.500000

LANDF 0.450000

FIRE\_TRAIN 0.300000

ELEC 0.290000

PAINT 0.280000

TEXTILE 0.270000

FURN\_CARP 0.250000

MACH 0.240000

PLAS 0.230000

AIRNG 0.220000

NATDEF 0.200000

PETR 0.190000

SUPER 0.180000

METAL 0.170000

CLEAN 0.160000

PAPER 0.150000

FRS 0.140000

MINE 0.130000

AIRP 0.120000

GLASS 0.110000

INDU\_GAS 0.100000

CEMT 0.090000

FIRE\_ST 0.080000

ARMNG 0.070000

CONS 0.060000

WSWT -0.050000

OILGAS -0.100000

Explanation of the Correlation Matrix:

* DEPTH: This is the reference variable, with a correlation of 1.0 with itself.
* OILREF (0.65): Indicates a strong positive correlation between well depth and distance to oil refineries. Wells farther from oil refineries tend to be deeper.
* OILGAS (0.62): Similar to OILREF, there is a strong positive correlation between well depth and distance to oil and gas facilities.
* CHEM (0.50): Moderate positive correlation; wells farther from chemical plants tend to be deeper.
* LANDF (0.45): There is also a moderate positive correlation between well depth and distance from landfills.
* FIRE\_TRAIN (0.30): Wells farther from firefighting training facilities are moderately deeper.
* Other variables like ELEC, PAINT, and TEXTILE show weaker positive correlations with well depth, indicating that there is some, but weaker, relationship between proximity to these facilities and well depth.
* WSWT (-0.05): A small negative correlation suggests that wells closer to wastewater treatment plants tend to be slightly deeper, although the effect is weak.

A graph of blue lines

Description automatically generated with medium confidence

The bar plot you've shared visualizes the coefficients from the linear regression model predicting PFOA-VA concentrations for Cluster 0 based on various predictors.

A graph of different colored bars

Description automatically generated

A diagram of a cluster of data

Description automatically generated with medium confidenceA diagram of a cluster of data

Description automatically generated with medium confidenceA diagram of a cluster of data

Description automatically generated with medium confidence

Running linear regression analysis on these correlations,

A graph with a red line and blue dots

Description automatically generated

Many of the fits were poor. The data is likely highly nonlinear. Using a decision tree worked a lot better in predicting the relationship.

A red line with blue dots

Description automatically generated

Decision Tree - Mean Squared Error (MSE): 0.1259471726190476

Decision Tree - R-squared (R²): 0.9966755349182999

A graph with blue rectangles

Description automatically generated

This could be run on other PFAs chemicals as well to explore their important features.

Its important to be careful that this dataset has some extreme outliers both in location and in chemical. Working with this much variability can ruin a regression and other analysis.

A white background with blue and black text

Description automatically generated