Analysis Report

This report is structured as follows.

Contents

Outlier Evaluation

In this analysis, removing outliers was not deemed crucial due to the flexibility of the Gamma GLMM in accommodating skewed data distributions. However, to improve the robustness of the model and reduce the influence of highly extreme values, two particularly high outliers were excluded from the analysis. Specifically, the top two values, representing "slo partners" with a PFAS level of 11.104 and "j m sims water supply" with a PFAS level of 3.873, were removed, as these values exceeded the typical range and could disproportionately impact the model parameters.

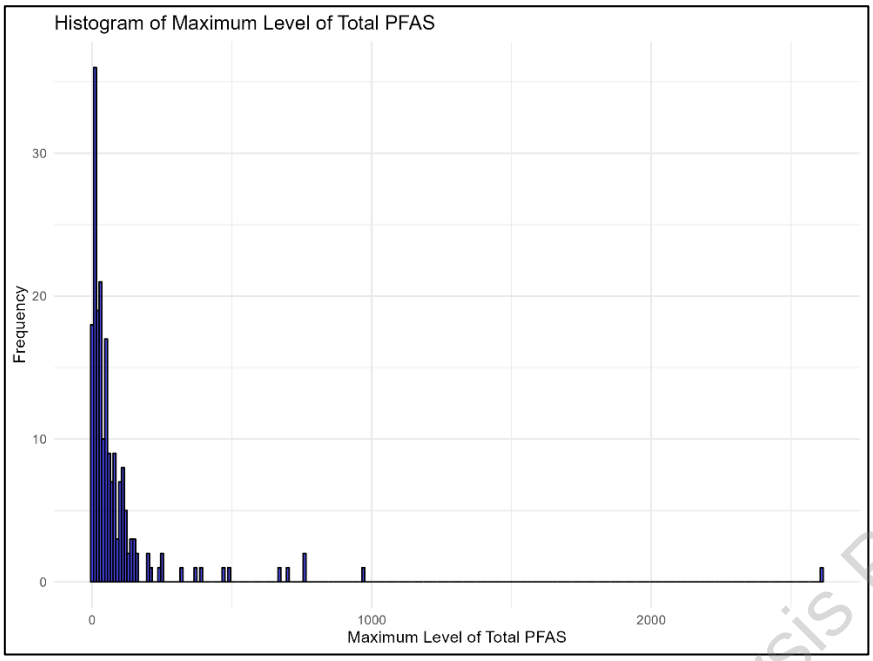
Public.Water.System	PFAS
slo partners	11.104
j m sims water supply	3.873
city of corona	2.947
strasbaugh, inc	2.947
yorba linda water district	2.699
noll properties	2.527
eastern municipal water district	1.760
chino basin desalter authority - desalter 2	1.680
cal-am water company - duarte	1.327
monrovia water department	1.213

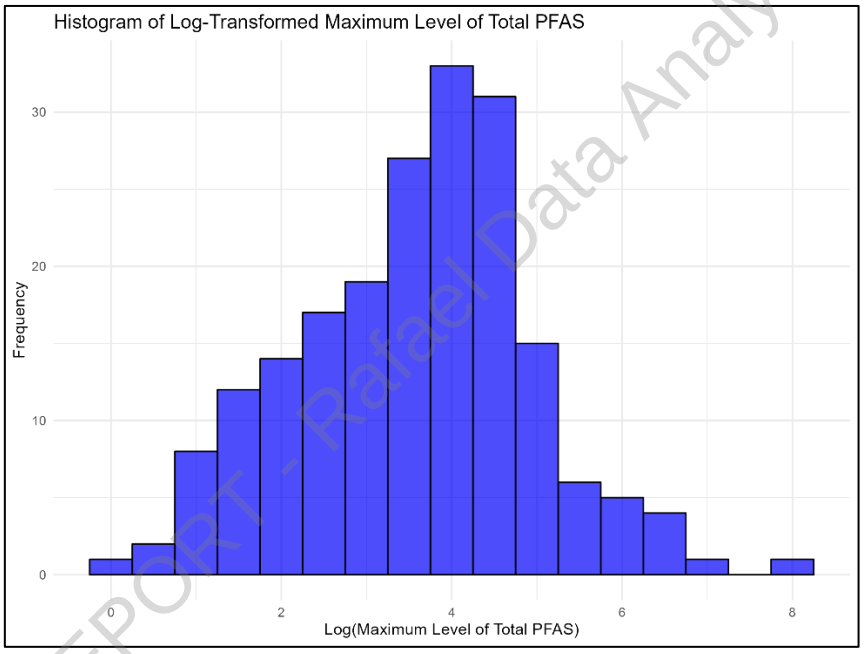
Data Distribution

In this analysis, the comparison of maximum PFAS levels across multiple counties was conducted while accounting for potential non-normality in the data. Initially, the distribution of the response variable, "Maximum Level of Total PFAS," was explored through a histogram, revealing significant right-skewness (Figure 1). This skewness was characterized by a mean PFAS level of 75.657 and a median of 41, highlighting the presence of extreme values that inflated the mean. The data also exhibited high variability, with a standard deviation of 121.588, and substantial skewness (3.782) and kurtosis (19.090), indicating a long-tailed distribution.

Variable	Mean	Median	SEM	SD	Skewness	Kurtosis
Maximum.Level.of.Total.PFAS	75.657	41	8.730	121.588	3.782	19.090

A log transformation was considered to normalize the data; however, the log-transformed variable did not sufficiently improve the normality of the distribution, as seen in Figure 2. Despite the reduction in skewness, the transformed data retained some degree of non-normality, leading to the exploration of alternative modeling methods.





A Generalized Linear Model (GLM) with a Gamma distribution was selected, as this method is well-suited for continuous, positively-skewed data. The Gamma distribution, paired with a log link function, enabled the modeling of multiplicative effects of the predictor variable, "County," on the response variable. This approach allowed for comparisons of mean PFAS levels across counties while accounting for the skewed nature of the data. Post-hoc pairwise comparisons were performed using the emmeans package, and specific differences between counties were identified. These results were saved in tidy dataframes for reporting, along with the estimated marginal means for each county. However, further analysis for cities and public water systems was not possible due to low representativity in the data, which hindered robust statistical comparison.

Descriptive Statistics by City

The table presents descriptive statistics for PFAS levels by city, ranked by the median PFAS levels. The highest median PFAS levels were observed in Corona (761.000), Yorba Linda (705.000), and Homeland (492.000). These cities had considerably higher PFAS concentrations compared to other locations. Duarte and Monrovia followed with median levels of 394.000 and 368.000, respectively. The substantial variation in PFAS concentrations across cities suggests notable geographic differences in contamination levels, warranting further investigation into the factors contributing to these elevated values.

City	N	Median	Mean	SEM	SD
corona	1	761.000	761.000		5.657
yorba linda	1	705.000	705.000		
homeland	1	492.000	492.000		105.147
duarte	1	394.000	394.000		16.971
monrovia	1	368.000	368.000		
atascadero	3	290.000	122.000	188.354	
riverside	3	249.333	249.000	129.615	
east los angeles	1	244.000	244.000		326.239
valencia	1	215.000	215.000		
san juan capistrano	1	202.000	202.000		24.520
palmdale	2	160.700	160.700	157.300	
san luis obispo	11	146.773	97.000	65.023	
garden grove	1	146.000	146.000		7.778
villa park	1	139.000	139.000		1.414
lancaster	1	134.000	134.000		2.121
lake elsinore	1	125.000	125.000		
commerce	1	116.000	116.000		
santa clarita	4	114.500	114.500	36.434	
tustin	1	106.000	106.000		
carson	1	105.000	105.000		
montebello	2	102.500	102.500	17.500	
lakeland	1	102.000	102.000		26.870
santa ana	1	102.000	102.000		
whittier	3	99.333	108.000	9.171	30.052
monterey park	1	98.000	98.000		
pico rivera	2	94.500	94.500	2.500	21.213
downey	2	91.500	91.500	19.500	
norwalk	3	91.333	107.000	18.747	
orange	4	89.750	89.500	33.368	
fallbrook	1	89.000	89.000		
bell gardens	2	84.000	84.000	1.000	
jurupa valley	1	84.000	84.000		
anaheim	2	83.650	83.650	74.350	
upland	1	83.000	83.000		27.577
helendale	1	76.000	76.000		
placentia	1	75.000	75.000		
el monte	5	67.400	54.000	25.934	57.990
kernville	2	66.500	66.500	40.500	
irwindale	1	64.000	64.000		
la crescenta	1	63.000	63.000		

City	N	Median	Mean	SEM	SD
south gate	2	62.500	62.500	8.500	34.648
lake arrowhead	1	60.000	60.000		
norco	2	59.250	59.250	55.750	
victorville	2	58.500	58.500	1.500	
cudahy	1	57.000	57.000		
arlington	1	55.000	55.000		
camarillo	2	52.000	52.000	19.000	
los osos	2	51.500	51.500	7.500	
altadena	1	50.000	50.000		9.192
hesperia	1	50.000	50.000		
ontario	2	49.500	49.500	19.500	
oro grande	2	49.000	49.000	5.000	
rialto	2	47.000	47.000	5.000	
west covina	1	46.000	46.000		57.276
fullerton	2	45.500	45.500	24.500	
moreno valley	1	44.000	44.000		
lynwood	2	40.500	40.500	13.500	
bakersfield	4	39.450	50.500	12.260	
colton	1	38.000	38.000		
barstow	2	36.500	36.500	5.500	
oxnard	8	35.050	21.000	16.164	
covina	1	35.000	35.000		
irvine	1	35.000	35.000		
temecula	2	35.000	35.000	21.000	
claremont	2	34.000	34.000	15.000	
pomona	1	33.000	33.000		10.607
redlands	1	32.000	32.000		19.092
adelanto	2	31.000	31.000	4.000	
arroyo grande	1	31.000	31.000		24.749
solvang	1	31.000	31.000		
beverly hills	1	30.000	30.000		
yucaipa	1	29.000	29.000		
arcadia	2	28.000	28.000	12.000	78.842
castaic	2	27.750	27.750	21.250	32.470
perris	2	26.700	26.700	20.300	
huntington park	2	26.500	26.500	6.500	
compton	1	26.000	26.000		27.577
bellflower	2	23.500	23.500	1.500	66.735
calimesa	1	23.000	23.000		7.071
nuevo	1	23.000	23.000		45.720
artesia	1	22.000	22.000		222.456
san bernardino	3	20.500	16.000	11.906	
azusa	1	20.000	20.000		
crestline	1	20.000	20.000		28.709
warner springs	1	20.000	20.000		3.536
baldwin park	1	19.000	19.000		
willowbrook	3	17.267	12.000	10.218	
encinitas	1	17.000	17.000		
wildomar	1	15.000	15.000		
paramount	1	14.000	14.000		7.071
santa paula	1	14.000	14.000		224.500
big bear lake	1	13.000	13.000		
burbank	1	13.000	13.000		20.622

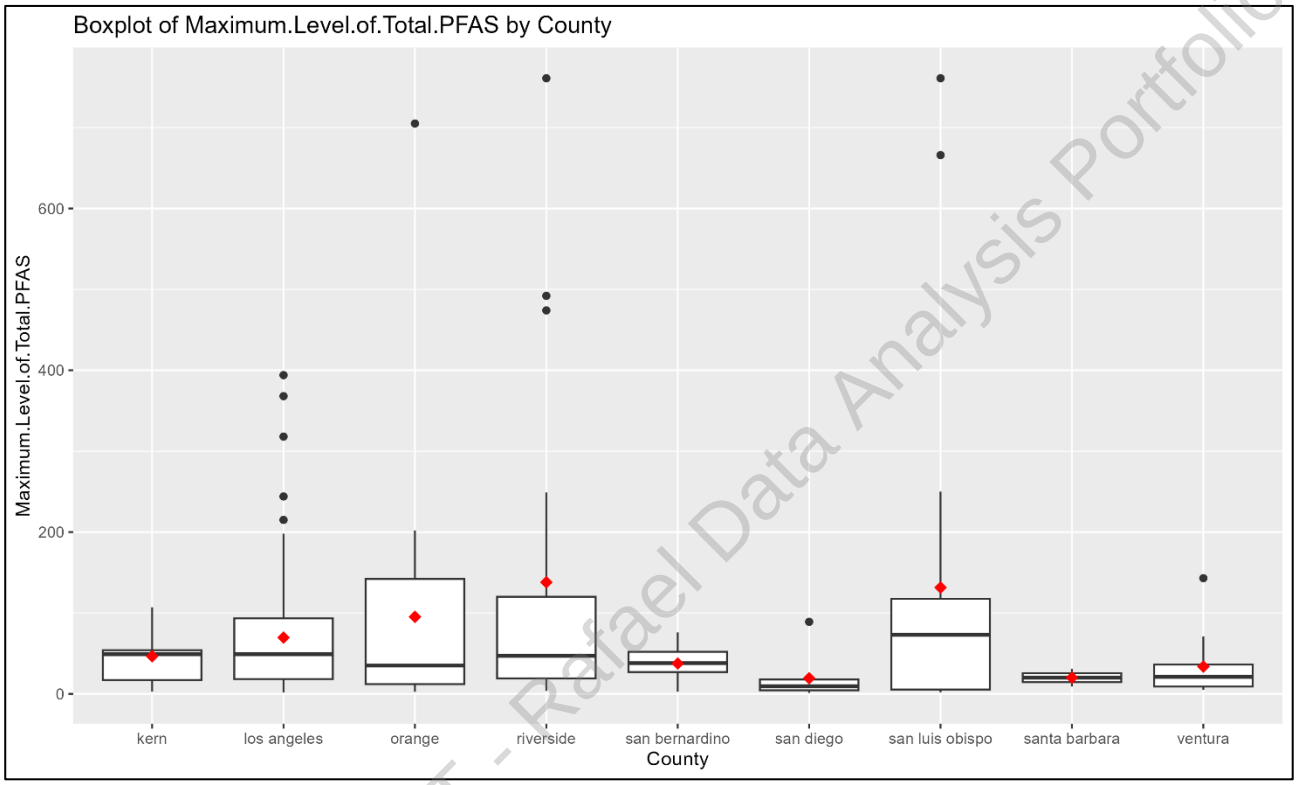
City	N	Median	Mean	SEM	SD
westminster	1	13.000	13.000		8.415
wofford heights	2	12.550	12.550	4.450	
buena park	1	12.000	12.000		215.655
huntington beach	1	12.000	12.000		
industry	1	11.000	11.000		
newport beach	1	10.000	10.000		72.867
rowland heights	1	10.000	10.000		
big bear	1	9.500	9.500		
santa maria	1	9.000	9.000		
leona valley	1	8.300	8.300		
banning	1	8.000	8.000		
san diego	2	7.050	7.050	5.950	12.021
glendale	1	6.900	6.900		29.698
glendora	1	6.700	6.700		0.919
lake forest	1	6.700	6.700		
cerritos	1	6.600	6.600		
ventura	1	6.100	6.100		
laguna beach	1	5.800	5.800		
santee	1	5.800	5.800		
san miguel	1	5.400	5.400		2.121
loma linda	1	5.000	5.000		
oceanside	1	4.800	4.800		
torrance	2	4.450	4.450	0.650	
vernon	1	4.300	4.300		
rancho santa fe	1	3.100	3.100		15.885
shandon	1	2.700	2.700		
fountain valley	1	2.400	2.400		17.698
lake hughes	1	2.200	2.200		6.293
paso robles	1	2.000	2.000		
la verne	1	1.900	1.900		

Descriptive Statistics by County

The boxplot presents the distribution of PFAS levels across various counties, revealing substantial differences in the maximum levels of total PFAS. The highest mean PFAS concentration was observed in Riverside County (137.889), followed by San Luis Obispo County (131.430) and Orange County (95.183). Kern and Los Angeles Counties had similar median values of 49.000, though Los Angeles displayed a wider spread of values, indicating more variability. The counties with the lowest median PFAS levels were San Diego (9.400) and Santa Barbara (20.000). Riverside and San Luis Obispo also had some of the most extreme outliers, as indicated by the points above 600 in the boxplot. The wide variability in these counties suggests a need for further investigation into local environmental or regulatory factors influencing these elevated levels.

Rank	County	N	Mean	Median	SEM	SD
1	san luis obispo	20	131.430	73.000	46.683	37.927

2	kern	9	46.433	49.000	12.642	78.855
3	los angeles	76	69.559	49.000	9.045	146.969
4	riverside	19	137.889	47.000	48.215	210.164
5	san bernardino	25	37.480	38.000	3.952	19.760
6	orange	23	95.183	35.000	30.645	29.007
7	ventura	12	33.708	21.000	11.411	208.774
8	santa barbara	2	20.000	20.000	11.000	15.556
9	san diego	8	19.225	9.400	10.255	39.529



GLMM Results

Generalized Linear Mixed Model (GLMM) results were used to compare PFAS levels across different counties using Los Angeles as the reference category. The model output reveals significant differences in PFAS concentrations between Los Angeles and several other counties.

The intercept, representing the expected log-transformed PFAS level for Los Angeles County, was estimated at 4.242 (SE = 0.151, $p < .001$). Riverside County had significantly higher PFAS levels compared to Los Angeles (B = 0.684, SE = 0.338, t = 2.026, p = .044). Similarly, San Luis Obispo County showed a significant increase in PFAS levels relative to Los Angeles (B = 1.401, SE = 0.319, t = 4.392, $p < .001$). In contrast, San Diego and San Bernardino Counties exhibited significantly lower PFAS levels compared to Los Angeles, with San Diego having the largest negative effect (B = -1.286, SE = 0.490, t = -2.627, p = .009) and San Bernardino showing a smaller but significant decrease (B = -0.618, SE = 0.304, t = -2.036, p = .043).

Other counties, such as Kern, Orange, Santa Barbara, and Ventura, did not show statistically significant differences in PFAS levels when compared to Los Angeles ($p > .05$). Orange County's estimate ($B = 0.314$, $SE = 0.313$) was close to significance, but the effect remained non-significant ($p = .318$).

Term	B	SE	t	p
(Intercept)	4.242	0.151	28.079	0.000
County - kern	-0.404	0.464	-0.870	0.385
County - orange	0.314	0.313	1.001	0.318
County - riverside	0.684	0.338	2.026	0.044
County - san bernardino	-0.618	0.304	-2.036	0.043
County - san diego	-1.286	0.490	-2.627	0.009
County - san luis obispo	1.401	0.319	4.392	0.000
County - santa barbara	-1.246	0.944	-1.321	0.188
County - ventura	-0.724	0.409	-1.771	0.078

Estimated Marginal Means

The estimated marginal means (EM Means) further illustrate the differences in PFAS concentrations across counties. Los Angeles had an EM Mean of 4.242 ($SE = 0.151$), while Riverside (EM Mean = 4.926) and San Luis Obispo (EM Mean = 5.643) had notably higher levels. On the lower end, San Diego (EM Mean = 2.956) and Santa Barbara (EM Mean = 2.996) were among the counties with the lowest predicted PFAS levels.

County	EM Mean	SE	df	lower.CL	upper.CL
los angeles	4.242	0.151	187	3.944	4.540
kern	3.838	0.439	187	2.972	4.704
orange	4.556	0.275	187	4.014	5.098
riverside	4.926	0.302	187	4.330	5.523
san bernardino	3.624	0.263	187	3.104	4.143
san diego	2.956	0.466	187	2.038	3.875
san luis obispo	5.643	0.281	187	5.089	6.197
santa barbara	2.996	0.931	187	1.158	4.833
ventura	3.518	0.380	187	2.768	4.268

Pairwise Comparisons

The pairwise comparisons between counties for PFAS levels, using Los Angeles as the reference category, revealed several significant differences. The comparison between Los Angeles and San Luis Obispo showed a statistically significant difference, with San Luis Obispo having much higher PFAS levels (B = -1.401, SE = 0.319, t = -4.392, p = 0.001). Similarly, San Bernardino also exhibited significantly lower PFAS levels compared to San Luis Obispo (B = -2.019, SE = 0.385, t = -5.244, p < 0.001). The difference between San Diego and San Luis Obispo was similarly large and significant (B = -2.687, SE = 0.544, t = -4.941, p < 0.001).

Conversely, most comparisons between Los Angeles and other counties, such as Kern, Orange, Riverside, and Ventura, did not reach statistical significance after adjusting for multiple comparisons (p > 0.05). For instance, the difference between Los Angeles and Riverside, although notable in the initial model (B = -0.684, SE = 0.338, t = -2.026), did not retain significance in the pairwise comparison (p = 1.000).

Notably, the pairwise comparison between Riverside and San Diego revealed a significant difference, with Riverside exhibiting higher PFAS levels (B = 1.970, SE = 0.555, t = 3.549, p = 0.018). Other significant differences were observed between Riverside and San Bernardino (B = 1.303, SE = 0.401, t = 3.250, p = 0.049), indicating that San Bernardino had lower PFAS levels than Riverside.

These pairwise comparisons illustrate the geographical variability in PFAS levels, with significant differences between counties like San Luis Obispo, San Diego, and Riverside compared to other regions.

Constrast	B	SE	df	t ratio	p
los angeles - kern	0.404	0.464	187.000	0.870	1.000
los angeles - orange	-0.314	0.313	187.000	-1.001	1.000
los angeles - riverside	-0.684	0.338	187.000	-2.026	1.000
los angeles - san bernardino	0.618	0.304	187.000	2.036	1.000
los angeles - san diego	1.286	0.490	187.000	2.627	0.336
los angeles - san luis obispo	-1.401	0.319	187.000	-4.392	0.001
los angeles - santa barbara	1.246	0.944	187.000	1.321	1.000
los angeles - ventura	0.724	0.409	187.000	1.771	1.000
kern - orange	-0.718	0.518	187.000	-1.386	1.000
kern - riverside	-1.088	0.533	187.000	-2.042	1.000
kern - san bernardino	0.214	0.512	187.000	0.418	1.000
kern - san diego	0.882	0.640	187.000	1.378	1.000
kern - san luis obispo	-1.805	0.521	187.000	-3.463	0.024
kern - santa barbara	0.842	1.030	187.000	0.818	1.000
kern - ventura	0.320	0.581	187.000	0.551	1.000
orange - riverside	-0.371	0.408	187.000	-0.908	1.000
orange - san bernardino	0.932	0.381	187.000	2.449	0.549

Constrast	B	SE	df	t ratio	p
orange - san diego	1.600	0.541	187.000	2.959	0.126
orange - san luis obispo	-1.087	0.393	187.000	-2.767	0.224
orange - santa barbara	1.560	0.971	187.000	1.607	1.000
orange - ventura	1.038	0.469	187.000	2.213	1.000
riverside - san bernardino	1.303	0.401	187.000	3.250	0.049
riverside - san diego	1.970	0.555	187.000	3.549	0.018
riverside - san luis obispo	-0.716	0.412	187.000	-1.737	1.000
riverside - santa barbara	1.931	0.979	187.000	1.972	1.000
riverside - ventura	1.409	0.486	187.000	2.901	0.150
san bernardino - san diego	0.668	0.535	187.000	1.248	1.000
san bernardino - san luis obispo	-2.019	0.385	187.000	-5.244	0.000
san bernardino - santa barbara	0.628	0.968	187.000	0.649	1.000
san bernardino - ventura	0.106	0.463	187.000	0.229	1.000
san diego - san luis obispo	-2.687	0.544	187.000	-4.941	0.000
san diego - santa barbara	-0.040	1.041	187.000	-0.038	1.000
san diego - ventura	-0.562	0.601	187.000	-0.934	1.000
san luis obispo - santa barbara	2.647	0.973	187.000	2.721	0.256
san luis obispo - ventura	2.125	0.473	187.000	4.496	0.000
santa barbara - ventura	-0.522	1.006	187.000	-0.519	1.000