Week 2 Problem Set

Daniel Crownover

2023-02-06

Questions:

- (1) Methane levels near fracking sites vs. Methane levels - far from fracking sites for ALL observations
- (2) Methane levels near fracking sites vs. Methane levels far from fracking sites for valley observations— only filter
- (3) Methane levels near fracking sites vs. Methane levels far from fracking sites for upland observations– only filter
- (4) Methane levels in the valley vs. Methane levels in the upland

Introduction: In this Problem Set, we delve into the analysis of water quality data extracted from a study conducted by Molofsky et al. (2013), focusing on methane levels in 1,701 drinking wells situated in Susquehanna County, Pennsylvania. The study aims to investigate whether methane levels in drinking water wells exhibit variations based on their proximity to fracking sites. Fracking, a prevalent method for extracting natural gas, has raised concerns regarding its potential impact on groundwater quality. The dataset segregates the drinking wells into two primary categories: those within a 1 km radius of a fracking site and those situated farther away. Additionally, the wells are further classified based on their geographical location, distinguishing between those in valley regions and those in upland areas. The study by Molofsky et al. (2013) serves as the foundation for our examination of water quality dynamics in the context of fracking activities. In this problem set, our objective is to employ R programming to conduct comparison of means tests, encompassing both parametric and non-parametric analyses. By leveraging statistical techniques, we endeavor to elucidate potential disparities in methane levels between wells near fracking sites and those distant from such operations. Through meticulous analysis and interpretation of the data, we aim to contribute insights into the broader discourse surrounding fracking and its ramifications on water quality. Prior to embarking on the problem set, it is imperative to familiarize oneself with the seminal work of Molofsky et al. (2013) to contextualize the dataset and its implications effectively. The ensuing discussion and analyses will be presented in a structured manner, utilizing tables to encapsulate our findings succinctly. In addition to the narrative report, the submission necessitates the provision of an .rmd file containing the executed R code, facilitating transparency and reproducibility in our analytical endeavors.

Data Description: The box plots in Figures 2.1.1 and 2.1.2 revealed numerous outliers in methane levels near and far from fracking sites, observed in both upland and valley locations. Shapiro-Wilks Test results (Figure 2.1.3) demonstrated strong evidence against normal distribution for both near (p = 7.093e-35) and far (p = 1.11e-59) proximities, a conclusion supported by visual inspection using ggplot (Figure 2.1.3). The mean methane levels for far wells were 684.25 g/L (shown in 2.2.4 and 2.2.5), with a median methane level in both locations at far samples of 0.6 g/L. The standard deviation of far wells was 3132.928 g/L, with an interquartile range (IQR) of 15.83 g/L for far wells in both upland and valley locations. Similarly, the mean methane levels for near wells were 1225.604.111 g/L (shown in 2.2.4 and 2.2.5), with a median methane level for near wells in both locations was 5.9 g/L. The standard deviation of near wells was 4086.95 g/L, with an

IQR of 25.575 g/L for near wells in the valley and upland. Levine's test shown in 2.1.6 (p =.59) suggested equal variances across far and near groups in the valley, while the parametric t-test shown in 2.1.7 (p = 0.65) failed to find a significant difference in mean methane levels. Cohen's d test (effect size = -0.03041743) indicated minimal differences between near and far proximity means.

The box plots in Figures 2.2.1 and 2.2.2 revealed numerous outliers in methane levels near and far from fracking sites, observed in Valley locations. Shapiro-Wilks Test results (Figure 2.2.3) demonstrated strong evidence against normal distribution for both near (p = 2.148e-27) and far (p = 7.97e-44) proximities, a conclusion supported by visual inspection using ggplot (Figure 2.2.3). The mean methane levels for far wells were 1186.406 g/L (shown in 2.2.4 and 2.2.5), with a median methane level in the valley at far samples of 1.3 g/L. The standard deviation of far wells was 4058.772 g/L, with an interquartile range (IQR) of 25.8 g/L for far wells in the valley. Similarly, the mean methane levels for near wells were 1225.604.111 g/L (shown in 2.2.4 and 2.2.5), with a median methane level for near wells in the valley at 19.0 g/L. The standard deviation of near wells was 5172.061 g/L, with an IQR of 25.48 g/L for near wells in the valley. Levine's test shown in figure 2.2.6 (p =.9155) suggested equal variances across far and near groups in the valley, while the parametric t-test shown in figure 2.2.7 (p = 0.922) failed to find a significant difference in mean methane levels. Cohen's d test shown in figure 2.2.8 (effect size = -0.008431705) indicated minimal differences between near and far proximity means.

The box plots in Figures 2.3.1 and 2.3.2 revealed numerous outliers in methane levels near and far from fracking sites, observed in Upland locations. Shapiro-Wilks Test results (Figure 2.3.3) demonstrated strong evidence against normal distribution for both near (p =7.46e-24)and far (p = 2.28e-49) proximities, a conclusion supported by visual inspection using ggplot (Figure 2.3.3). The mean methane levels for far wells were 209.731 g/L (shown in 2.3.4 and 2.3.5), with a median methane level in the upland at far samples of 1.3 g/L. The standard deviation of far wells was 1753.102 g/L, with an interquartile range (IQR) of 2.25 g/L for far wells in the upland. Similarly, the mean methane levels for near wells were 133.880.111 g/L (shown in 2.3.4 and 2.3.5), with a median methane level for near wells in the upland at 1.4 g/L. The standard deviation of near wells was 799.431 g/L, with an IQR of 25.69 g/L for near wells in the upland. Levine's test (shown in figure 2.3.4) (p =.63) suggested equal variances across far and near groups in the upland, while the parametric t-test shown in figure 2.3.7 (p = 0.434) failed to find a significant difference in mean methane levels. Cohen's d shown in figure 2.3.8 test (effect size = 0.0055672) indicated minimal differences between near and far proximity means.

The box plots in Figures 2.4.1 and 2.4.2 revealed numerous outliers in methane levels in Upland and Valley sample locations. Shapiro-Wilks Test results (Figure 2.4.3) demonstrated strong evidence against normal distribution for both upland (p = 1.953928e-52)and valley (p = 8.197769e-4) locations, a conclusion supported by visual inspection using ggplot (Figure 2.4.3). The mean methane levels for upland location was 198.208 g/L (shown in 2.4.4 and 2.4.5), with a median methane at upland location of 0.47 g/L. The standard deviation of upland locations was 1644.111 g/L, with an interquartile range (IQR) of 3.565 g/L for far wells in the upland. Similarly, the mean methane levels for the valley location was 1195.243 g/L (shown in 2.4.4 and 2.4.5), with a median methane level for Valley at 1.8 g/L. The standard deviation of upland location was 4331.548 g/L, with an IQR of 25.780 g/L for the location of the upland. Levine's test (shown in figure 2.4.6) (p =.63 for upland) and (0.91 upland for valley) suggested equal variances across groups in the upland, while the parametric t-test shown in figure 2.3.7 (p = 0.43 UPLAND) AND P = 0.92 VALLEY. this failed to find a significant difference in mean methane levels. Cohen's d shown in figure 2.4.8 test (effect size = 0.0055672 upland) and indicated minimal differences between upland and valley location means.

The Wilcox test for question (1) provided a p-value of 1.64e-10 as shown in Figure 3.1.1. The effect size for this test was 0.159 as shown in figure 3.1.2. The Wilcox test for question (2) provided a p-value of .00145 as shown in Figure 3.2.1. The effect size for this test was 0.108 as shown in figure 3.2.2.The Wilcox test for question (3) provided a p-value of 1.05e-06 as shown in Figure 3.3.1. The effect size for this test was 0.168 as shown in figure 3.3.2. The wilcox test for question (4) provided a p-value of 1.05e-06 as shown in Figure 3.4.1. The effect size for this test was .16 as shown in figure 3.4.2.

For question (1)the log transformations revealed a levine test(Figure 4.1.3) p value of (p = 0.83) representing equal variance in mean values across Near and Far proximities across all sites/loactions as shown in figure 4.1.3. The Shapiro test indicated a (p = 1.187122e-35) for far observations across all sites and (p= 9.47e-14)

for near proximities across the 4 sites. These va;ues demonstrated strong evidence against normal distribution for both far a near proximities across all sites, a conclusion supported by visual inspection using histogram (Figure 4.1.2)

For question (2) the log transformations revealed a levine test (Figure 4.2.3)p value of (p = 0.0028) representing varied variance in mean values across Near and Far proximities across the valley location as shown in figure 4.2.3. The Shapiro test(figure 4.2.4) indicated a (p = 1.547412e-23) for far observations across the valley and (p = 1.54149) for near proximities across the valley. This demonstrated strong evidence against normal distribution for both far a near proximities across all sites, a conclusion supported by visual inspection using histogram (Figure 4.2.2)

For question (3)The log transformations revealed a leivine test (figure 4.3.3)p value of (p=.005582372) representing varied variance in mean values across Near and Far proximities across the upland location as shown in figure 4.2.3. The Shapiro test(figure 4.3.4) indicated a (p = 2.381698e-27) for far observations across the upland and (1.061215e-06) for near proximities across upland. This demonstrated strong evidence against normal distribution for both far a near proximities across all sites, a conclusion supported by visual inspection using histogram (Figure 4.3.2)

For question (4)The log transformations revealed a leivine test (figure 4.4.3)p value of (p=.0.00558) representing varied(not equal) variance in mean values across upland and valley locations as shown in figure 4.4.3. The Shapiro test(figure 4.4.4) indicated a (p = 6.92e-28) for upland observations and (p = 1.08e-24) for valley locations across upland. This demonstrated strong evidence against normal distribution for both upland a valley locations for both far and near proximities, a conclusion supported by visual inspection using histogram (Figure 4.4.2).

Discussion and Statistical Analysis: Our preliminary parametric tests indicated a failure to meet the assumptions necessary for conducting a t-test. Despite this, when we proceeded with the t-test, contrary to the preliminary results, we obtained parametric t-test p-values below the critical value of 0.05 for all four questions posed. Consequently, we did not find a statistically significant difference between the mean levels of methane for the far and near groups or between the valley and upland regions. In response to the ineffectiveness of the parametric t-test, we pursued non-parametric testing and log transformations for further analysis and hypothesis testing. The Wilcoxon test, conducted for all four questions in the non-parametric hypothesis testing, suggested that the observed data were highly improbable under the null hypothesis, indicating that the data still exhibited an abnormal kurtosis distribution. Furthermore, log-transformed data across all four sites yielded very small p-values for the Shapiro test. Consequently, we concluded that we do not possess sufficient evidence to assert a significant difference between the mean methane levels among the far and near groups or between the valley and upland regions. Even post-normalization, the data remained highly skewed and demonstrated an abnormal distribution.

Conclusion: Based on our analysis, we conclude that the methods used in our study, including parametric and non-parametric tests, were unable to provide and lack sufficient and conclusive evidence regarding variations in mean methane levels among drinking wells near fracking sites compared to those distant from such operations or based on relative location. Despite attempts to address data distribution challenges through parametric and non-parametric and logarithmic approaches, including log transformations, the results remained inconclusive due to persistent abnormal distributions and skewness. Our preliminary parametric tests failed to meet underlying assumptions. Subsequent non-parametric analyses and transformations did not yield evidence to support statistically significant differences in methane levels between the near and far groups, or between valley and upland regions. The findings of our study echo the concerns raised by Molofsky et al. (2013) regarding the potential impacts of fracking on groundwater quality. However, our inability to discern clear patterns underscores the complexity of assessing methane levels in the context of fracking activities.

Key:

- Number inside parentheses (#) corresponds to one of the comparison groups (1, 2, 3, or 4).

Appendix:

- Figure 1.1: Download the Water Quality Data
- 2.0.0: Parametric t-test and Preliminary Test
- 2.1.0 (1)
- Figure 2.1.1: Create a Box Plot (1)
- Figure 2.1.2: 2.1 Identify Outliers by Groups (1)
- Figure 2.1.3: 2.1 Check Normality by Groups (1)
- Figure 2.1.4: 2.1 Summary Statistics ~ Mean and Interquartile Range ~ (1)
- Figure 2.1.5: 2.1 Summary Statistics ~ Mean Methane and Median Methane ~ (1)
- Figure 2.1.6: 2.1 Levine Test (1)
- Figure 2.1.7: 2.1 Parametric t-test (1)
- Figure 2.1.8: 2.1 Cohens D test (1)
- Figure 2.1.9: 2.1 Report Results (1)
- 2.2.0 (2)
- Figure 2.2.1: 2.2 Create a Box Plot (2)
- Figure 2.2.2: 2.2 Identify Outliers by Groups (2)
- Figure 2.2.3: 2.2 Check Normality by Groups (2)
- Figure 2.2.4: 2.2 Summary Statistics ~ Mean and Interquartile Range ~ (2)
- Figure 2.2.5: 2.2 Summary Statistics ~ Mean Methane and Median Methane ~ (2)
- Figure 2.2.6: 2.2 Levine Test (2)
- Figure 2.2.7: 2.2 Parametric t-test (2)
- Figure 2.2.8: 2.2 Cohens D test (2)
- Figure 2.2.9: 2.2 Report Results (2)
- 2.3.0 (3)
- Figure 2.3.1: 2.3 Create a Box Plot (3)
- Figure 2.3.2: 2.3 Identify Outliers by Groups (3)
- Figure 2.3.3: 2.3 Check Normality by Groups (3)
- Figure 2.3.4: 2.3 Levine test (3)
- Figure 2.3.5: 2.3 Summary Statistics ~ Mean and Interquartile Range ~ (3)
- Figure 2.3.6: 2.3 Summary Statistics ~ Mean Methane and Median Methane ~ (3)
- Figure 2.3.7: 2.3 Parametric t-test (3)
- Figure 2.3.8: 2.3 Cohens D test (3)
- Figure 2.3.9: 2.3 Report Results (3)
- 2.4.0
- Figure 2.4.1: 2.4 Create a Box Plot (4)
- Figure 2.4.2: 2.4 Identify Outliers by Groups (4)
- Figure 2.4.3: 2.4 Check Normality by Groups (4)
- Figure 2.4.4: 2.4 Summary Statistics ~ Mean and Interquartile Range ~ (4)
- Figure 2.4.5: 2.4 Summary Statistics ~ Mean Methane and Median Methane ~ (4)
- Figure 2.4.6: 2.4 Levine test (4)
- Figure 2.4.7: 2.4 Parametric t-test (4)
- Figure 2.4.8: 2.4 Cohens D test (4)
- Figure 2.4.9: 2.4 Report Results (4)
- 3.0.0: Non-Parametric t-test and Preliminary test
- 3.1.0
- Figure 3.1.1: 3.1 Compute the Wilcox test (1)
- Figure 3.1.2: 3.1 Calculate Effect Size (1)
- Figure 3.2.1: 3.2 Compute the Wilcox test (2)
- Figure 3.2.2: 3.2 Calculate Effect Size (2)
- Figure 3.3.1: 3.3 Compute the Wilcox test (3)

- Figure 3.3.2: 3.3 Calculate Effect Size (3)
- Figure 3.4.1: 3.4 Compute the Wilcox test (4)
- Figure 3.4.2: 3.4 Calculate Effect Size (4)
- 4.0.0: Log Transformations and t-test
- Figure 4.1.1: 4.1 Log Transforming Data (1)
- Figure 4.1.2: 4.1 Histogram of Logged Data (1)
- Figure 4.1.3: 4.1 Levine test (1)
- Figure 4.1.4: 4.1 Shapiro and Normality Test on Log Value (1)
- Figure 4.2.1: 4.2 Log Transforming Data (2)
- Figure 4.2.2: 4.2 Histogram of Logged Data (2)
- Figure 4.2.3: 4.2 Levine test (2)
- Figure 4.2.4: 4.2 Shapiro and Normality Test on Log Value (2)
- Figure 4.3.1: 4.3 Log Transforming Data (3)
- Figure 4.3.2: 4.3 Histogram of Logged Data (3)
- Figure 4.3.3: 4.3 Levine test (3)
- Figure 4.3.4: 4.3 Shapiro and Normality Test on Log Value (3)
- Figure 4.4.1: 4.4 Log Transforming Data (4)
- Figure 4.4.2: 4.4 Histogram of Logged Data (4)
- Figure 4.4.3: 4.4 Levine test (4)
- Figure 4.4.4: 4.4 Shapiro and Normality Test on Log Value (4)

Figure 1.1: Download the Water Quality Data

```
##   ID methane dl location proximity
## 1  1 2.6e+01  1   Valley       Far
## 2  2 1.7e+04  0   Valley       Far
## 3  3 2.6e+01  1   Upland       Far
## 4  4 2.6e+01  1   Upland       Far
## 5  5 1.1e-01  0   Upland      Near
## 6  6 1.1e-01  0   Valley       Far
```

2.0.0: Parametric t-test and Preliminary Test

2.1.0 (1)

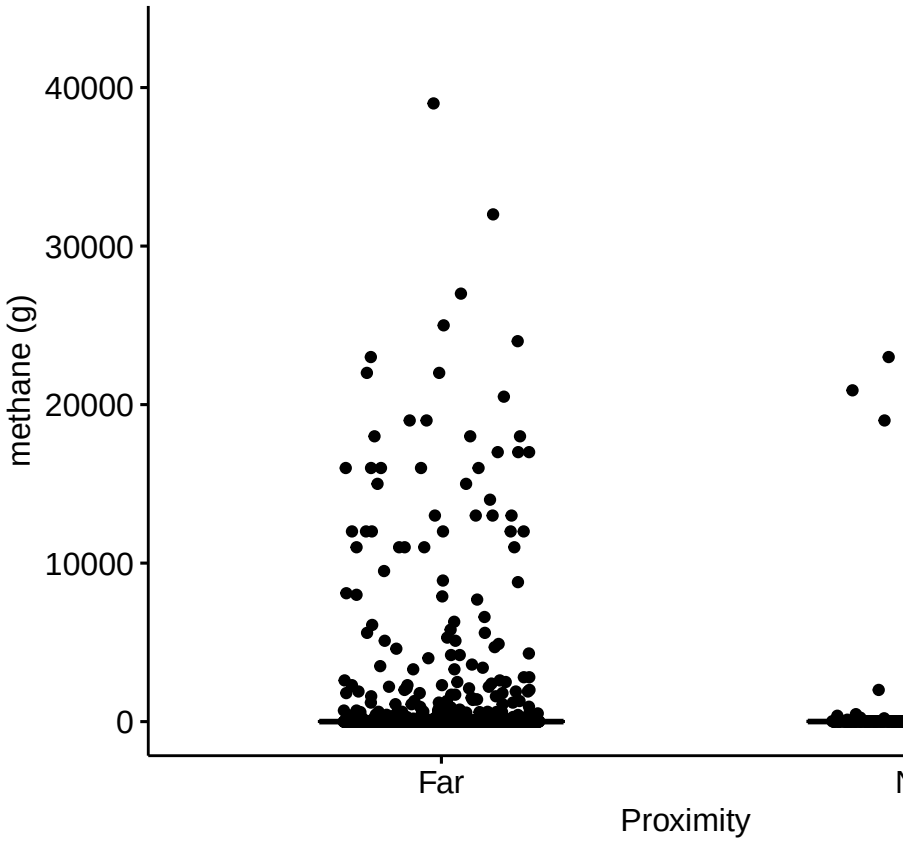


Figure 2.1.1: 2.1 Create a Box Plot (1)

Figure 2.1.2: 2.1 Identify Outliers by Groups (1)

Figure 2.1.3: 2.1 Check Normality by Groups (1)

proximity	ID	methane	dl	location	is.outlier	is.extreme
Far	2	17000.0	0	Valley	TRUE	TRUE
Far	13	1400.0	0	Valley	TRUE	TRUE
Far	16	1300.0	0	Valley	TRUE	TRUE
Far	18	230.0	0	Upland	TRUE	TRUE
Far	29	68.0	0	Upland	TRUE	TRUE
Far	32	18000.0	0	Valley	TRUE	TRUE
Far	56	9500.0	0	Valley	TRUE	TRUE
Far	63	140.0	0	Valley	TRUE	TRUE
Far	67	610.0	0	Valley	TRUE	TRUE
Far	78	440.0	0	Valley	TRUE	TRUE
Far	88	17000.0	0	Valley	TRUE	TRUE
Far	97	610.0	0	Upland	TRUE	TRUE
Far	99	13000.0	0	Valley	TRUE	TRUE
Far	100	25000.0	0	Valley	TRUE	TRUE
Far	101	27000.0	0	Valley	TRUE	TRUE
Far	103	22000.0	0	Valley	TRUE	TRUE
Far	106	24000.0	0	Valley	TRUE	TRUE
Far	112	85.0	0	Valley	TRUE	TRUE
Far	114	4300.0	0	Upland	TRUE	TRUE
Far	115	6600.0	0	Valley	TRUE	TRUE
Far	117	4200.0	0	Valley	TRUE	TRUE
Far	124	430.0	0	Upland	TRUE	TRUE
Far	127	140.0	0	Valley	TRUE	TRUE
Far	130	16000.0	0	Valley	TRUE	TRUE
Far	143	16000.0	0	Upland	TRUE	TRUE
Far	144	85.0	0	Valley	TRUE	TRUE
Far	152	120.0	0	Valley	TRUE	TRUE
Far	157	5100.0	0	Valley	TRUE	TRUE
Far	171	59.0	0	Upland	TRUE	FALSE
Far	187	1100.0	0	Valley	TRUE	TRUE
Far	198	50.0	0	Valley	TRUE	FALSE
Far	199	180.0	0	Upland	TRUE	TRUE
Far	264	290.0	0	Valley	TRUE	TRUE
Far	268	1100.0	0	Valley	TRUE	TRUE
Far	274	440.0	0	Upland	TRUE	TRUE
Far	300	185.0	0	Valley	TRUE	TRUE
Far	392	40.0	0	Upland	TRUE	FALSE
Far	404	170.0	0	Upland	TRUE	TRUE
Far	426	15000.0	0	Valley	TRUE	TRUE
Far	438	1200.0	0	Valley	TRUE	TRUE
Far	445	290.0	0	Upland	TRUE	TRUE
Far	493	39000.0	0	Valley	TRUE	TRUE
Far	511	213.0	0	Upland	TRUE	TRUE
Far	518	120.0	0	Valley	TRUE	TRUE
Far	523	11000.0	0	Valley	TRUE	TRUE
Far	540	160.0	0	Upland	TRUE	TRUE
Far	555	5300.0	0	Upland	TRUE	TRUE
Far	557	95.0	0	Valley	TRUE	TRUE
Far	558	330.0	0	Valley	TRUE	TRUE
Far	568	1800.0	0	Valley	TRUE	TRUE
Far	569	5800.0	0	Valley	TRUE	TRUE
Far	573	160.0	0	Valley	TRUE	TRUE
Far	575	2800.0	0	Valley	TRUE	TRUE
Far	579	3300.0	0	Valley	TRUE	TRUE
Far	580	670.0	0	Valley	TRUE	TRUE
Far	590	40.0	0	Valley	TRUE	FALSE
Far	596	1900.0	0	Upland	TRUE	TRUE

proximity	variable	statistic	p
Far	methane	0.2298413	0
Near	methane	0.1954787	0

proximity	variable	n	mean	sd
Far	methane	1379	684.257	3132.928
Near	methane	322	795.017	4086.957

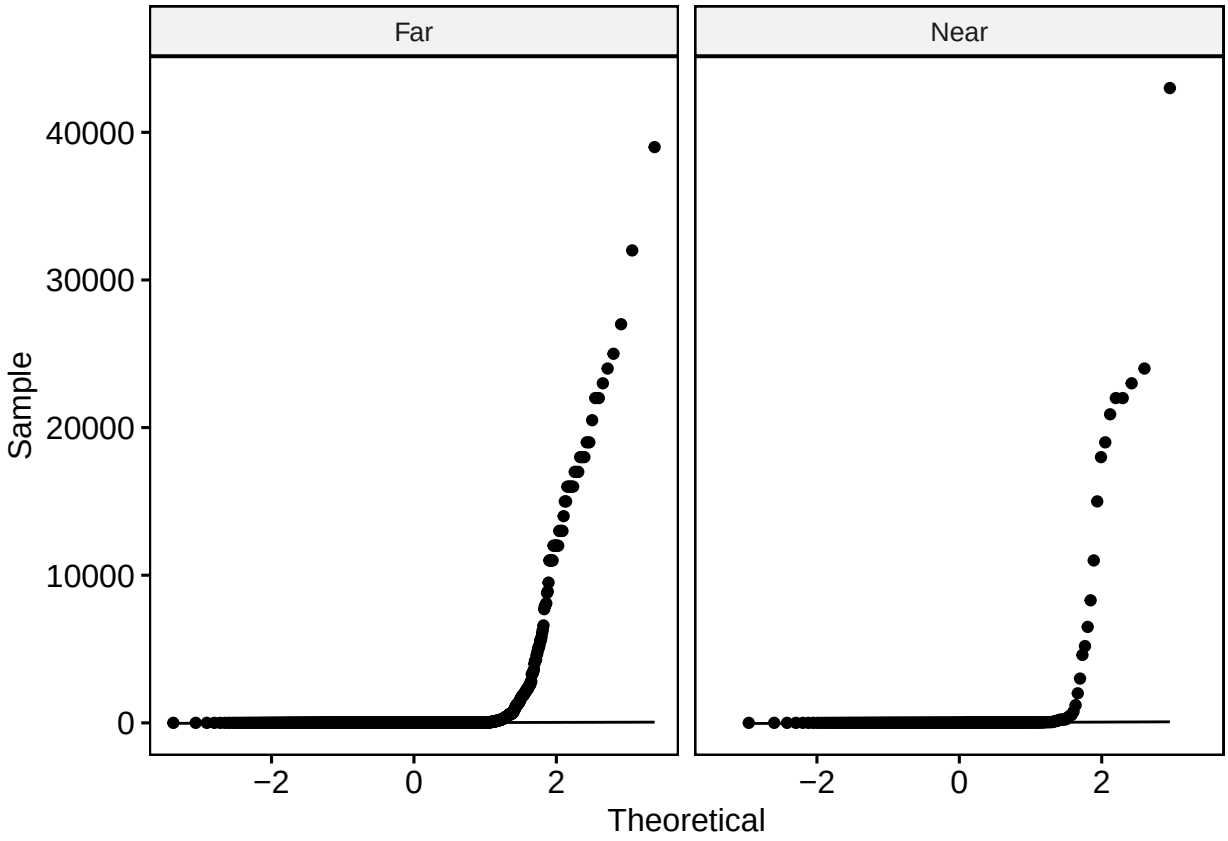


Figure 2.1.4: 2.1 Summary Statistics ~ Mean and Interquartile Range ~ (1)

Figure 2.1.5: 2.1 Summary Statistics ~ Mean Methane and Median Methane ~ (1)

Figure 2.1.6: 2.1 Levine Test (1)

Figure 2.1.7: 2.1 Parametric t-test(1)

```
## # A tibble: 1 x 16
##   estimate estimate1 estimate2 .y.     group1 group2    n1    n2 statistic     p
##      <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>  <int> <int>     <dbl> <dbl>
## 1    -111.      684.      795. methane Far    Near    1379   322    -0.456 0.649
## # i 6 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
## #   alternative <chr>, p.signif <chr>
```

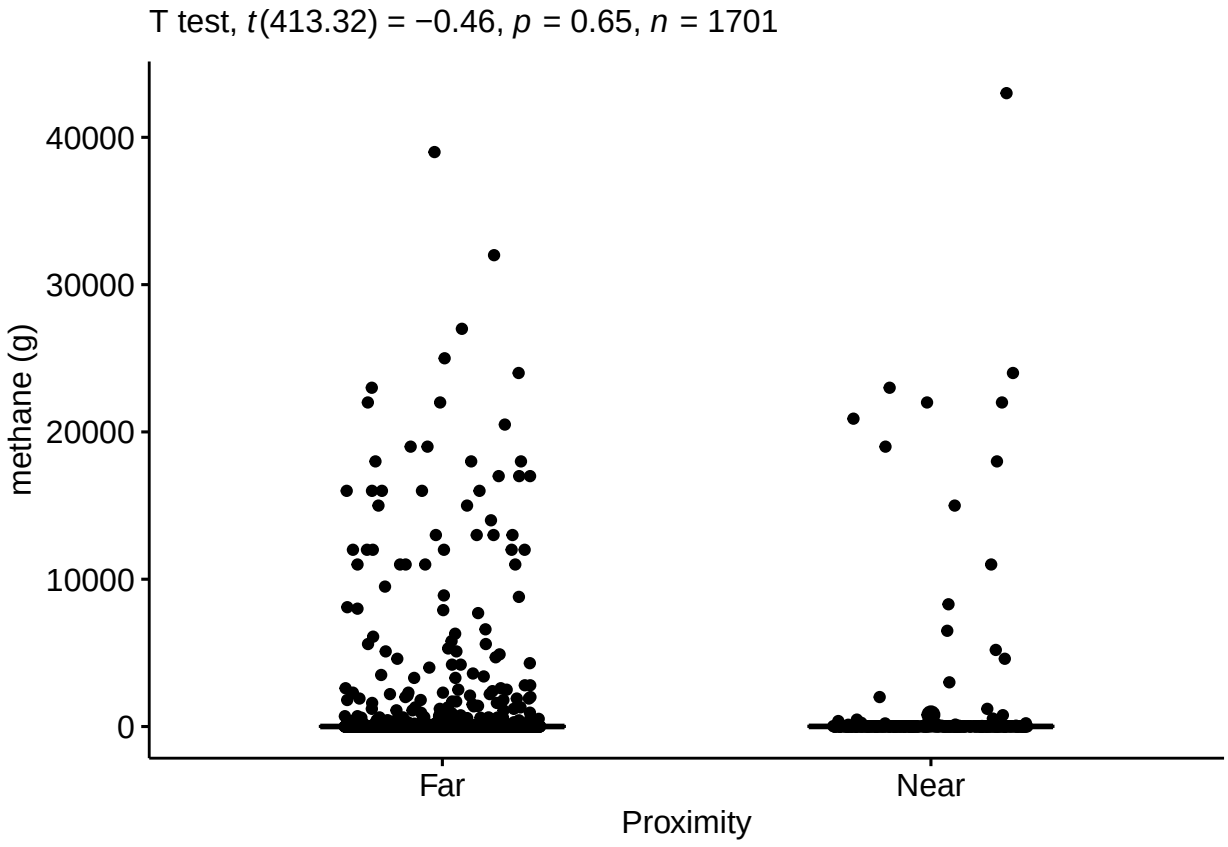
proximity	variable	n	median	iqr
Far	methane	1379	0.6	15.830
Near	methane	322	5.9	25.575

Proximity of All	Mean Methane	Median Methane
Far	684.26	0.6
Near	795.02	5.9

df1	df2	statistic	p
1	1699	0.2846576	0.5937344

Figure 2.1.8: 2.1 Cohens D test (1)

Figure 2.1.9: 2.1 Report Results (1)



2.2.0 (2)

Figure 2.2.1: 2.2 Create a Box Plot (2)

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	-0.0304174	1379	322	negligible

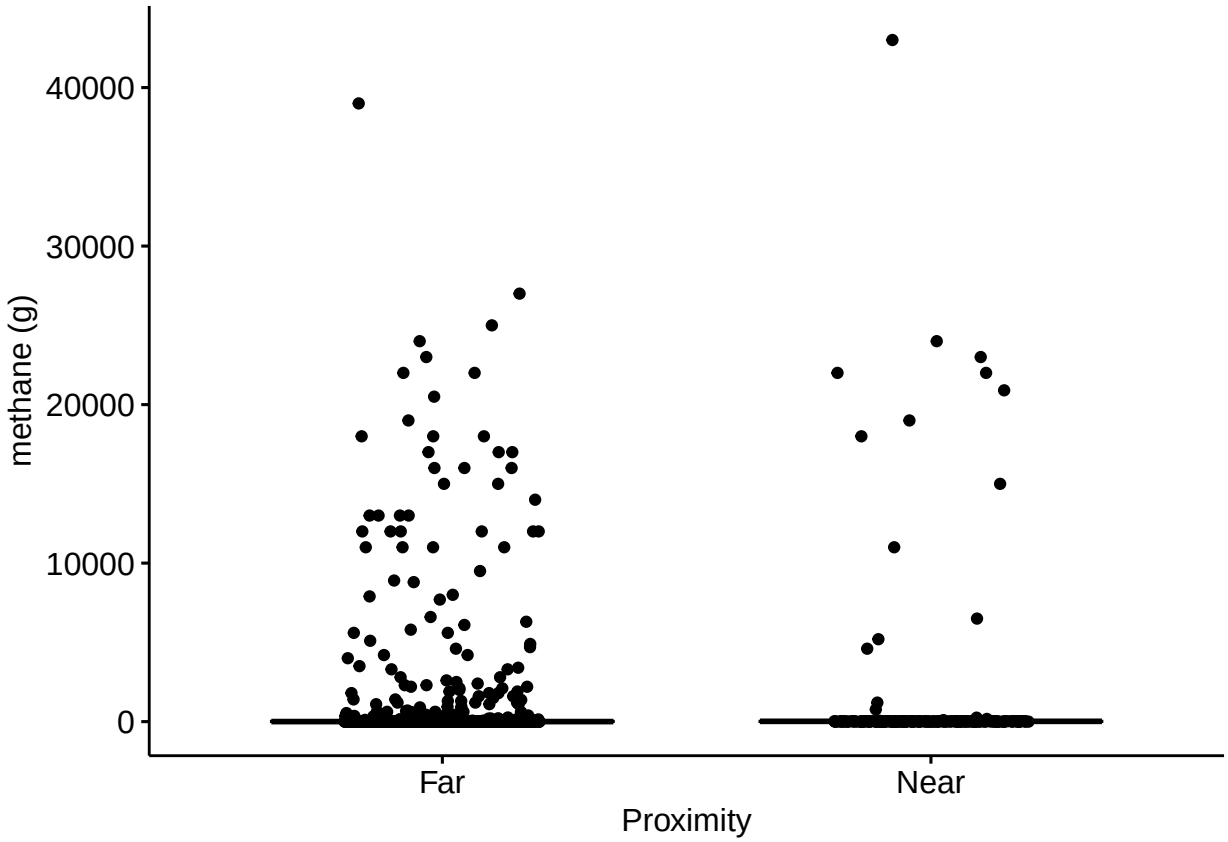


Figure 2.2.2: 2.2 Identify Outliers by Groups (2)

Figure 2.2.3: 2.2 Check Normality by Groups (2)

proximity	ID	methane	dl	location	is.outlier	is.extreme
Far	2	17000	0	Valley	TRUE	TRUE
Far	13	1400	0	Valley	TRUE	TRUE
Far	16	1300	0	Valley	TRUE	TRUE
Far	32	18000	0	Valley	TRUE	TRUE
Far	56	9500	0	Valley	TRUE	TRUE
Far	63	140	0	Valley	TRUE	TRUE
Far	67	610	0	Valley	TRUE	TRUE
Far	78	440	0	Valley	TRUE	TRUE
Far	88	17000	0	Valley	TRUE	TRUE
Far	99	13000	0	Valley	TRUE	TRUE
Far	100	25000	0	Valley	TRUE	TRUE
Far	101	27000	0	Valley	TRUE	TRUE
Far	103	22000	0	Valley	TRUE	TRUE
Far	106	24000	0	Valley	TRUE	TRUE
Far	112	85	0	Valley	TRUE	FALSE
Far	115	6600	0	Valley	TRUE	TRUE
Far	117	4200	0	Valley	TRUE	TRUE
Far	127	140	0	Valley	TRUE	TRUE
Far	130	16000	0	Valley	TRUE	TRUE
Far	144	85	0	Valley	TRUE	FALSE
Far	152	120	0	Valley	TRUE	TRUE
Far	157	5100	0	Valley	TRUE	TRUE
Far	187	1100	0	Valley	TRUE	TRUE
Far	264	290	0	Valley	TRUE	TRUE
Far	268	1100	0	Valley	TRUE	TRUE
Far	300	185	0	Valley	TRUE	TRUE
Far	426	15000	0	Valley	TRUE	TRUE
Far	438	1200	0	Valley	TRUE	TRUE
Far	493	39000	0	Valley	TRUE	TRUE
Far	518	120	0	Valley	TRUE	TRUE
Far	523	11000	0	Valley	TRUE	TRUE
Far	557	95	0	Valley	TRUE	FALSE
Far	558	330	0	Valley	TRUE	TRUE
Far	568	1800	0	Valley	TRUE	TRUE
Far	569	5800	0	Valley	TRUE	TRUE
Far	573	160	0	Valley	TRUE	TRUE
Far	575	2800	0	Valley	TRUE	TRUE
Far	579	3300	0	Valley	TRUE	TRUE
Far	580	670	0	Valley	TRUE	TRUE
Far	617	420	0	Valley	TRUE	TRUE
Far	622	1200	0	Valley	TRUE	TRUE
Far	625	2300	0	Valley	TRUE	TRUE
Far	627	260	0	Valley	TRUE	TRUE
Far	648	600	0	Valley	TRUE	TRUE
Far	685	150	0	Valley	TRUE	TRUE
Far	695	11000	0	Valley	TRUE	TRUE
Far	698	210	0	Valley	TRUE	TRUE
Far	700	3300	0	Valley	TRUE	TRUE
Far	711	140	0	Valley	TRUE	TRUE
Far	714	73	0	Valley	TRUE	FALSE
Far	723	2200	0	Valley	TRUE	TRUE
Far	726	2400	0	Valley	TRUE	TRUE
Far	738	94	0	Valley	TRUE	FALSE
Far	740	700	0	Valley	TRUE	TRUE
Far	746	1300	0	Valley	TRUE	TRUE
Far	747	210	0	Valley	TRUE	TRUE
Far	766	12000	0	Valley	TRUE	TRUE

proximity	variable	statistic	p
Far	methane	0.3308850	0
Near	methane	0.2533037	0

proximity	variable	n	mean	sd
Far	methane	670	1186.406	4058.772
Near	methane	195	1225.604	5172.061

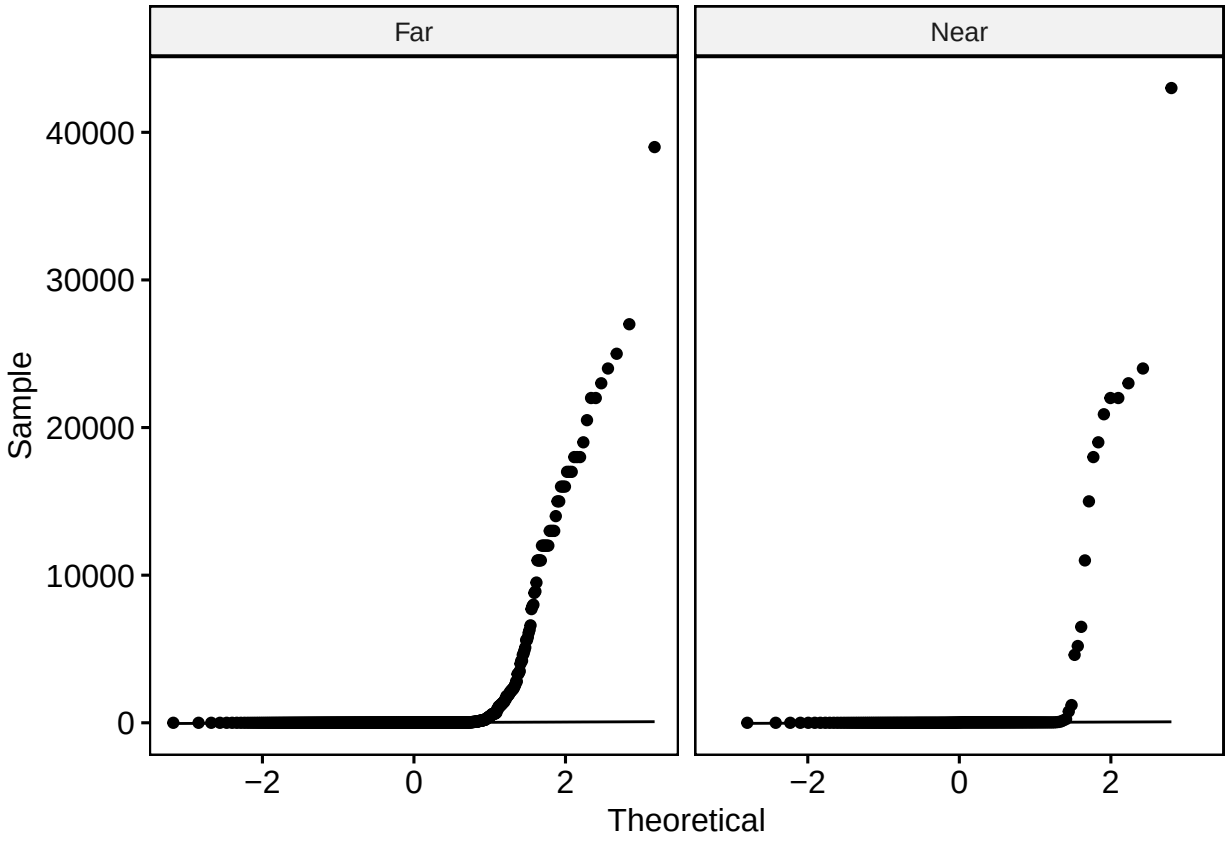


Figure 2.2.4: 2.2 Summary Statistics ~ Mean and Interquartile Range ~ (2)

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v forcats   1.0.0     v stringr   1.5.0
## v lubridate 1.9.2     v tibble    3.2.1
## v purrr     1.0.1     v tidyr     1.3.0
## v readr     2.1.4
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x rstatix::filter()        masks dplyr::filter(), stats::filter()
## x kableExtra::group_rows() masks dplyr::group_rows()
## x dplyr::lag()             masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

proximity	variable	n	median	iqr
Far	methane	670	1.3	25.808
Near	methane	195	19.0	25.480

Proximity in the Valley	Mean Methane	Median Methane
Far	1186.41	1.3
Near	1225.60	19.0

df1	df2	statistic	p
1	863	0.011246	0.9155695

Figure 2.2.5: 2.2 Summary Statistics ~ Mean Methane and Median Methane ~ (2)

Figure 2.2.6: 2.2 Levine Test (2)

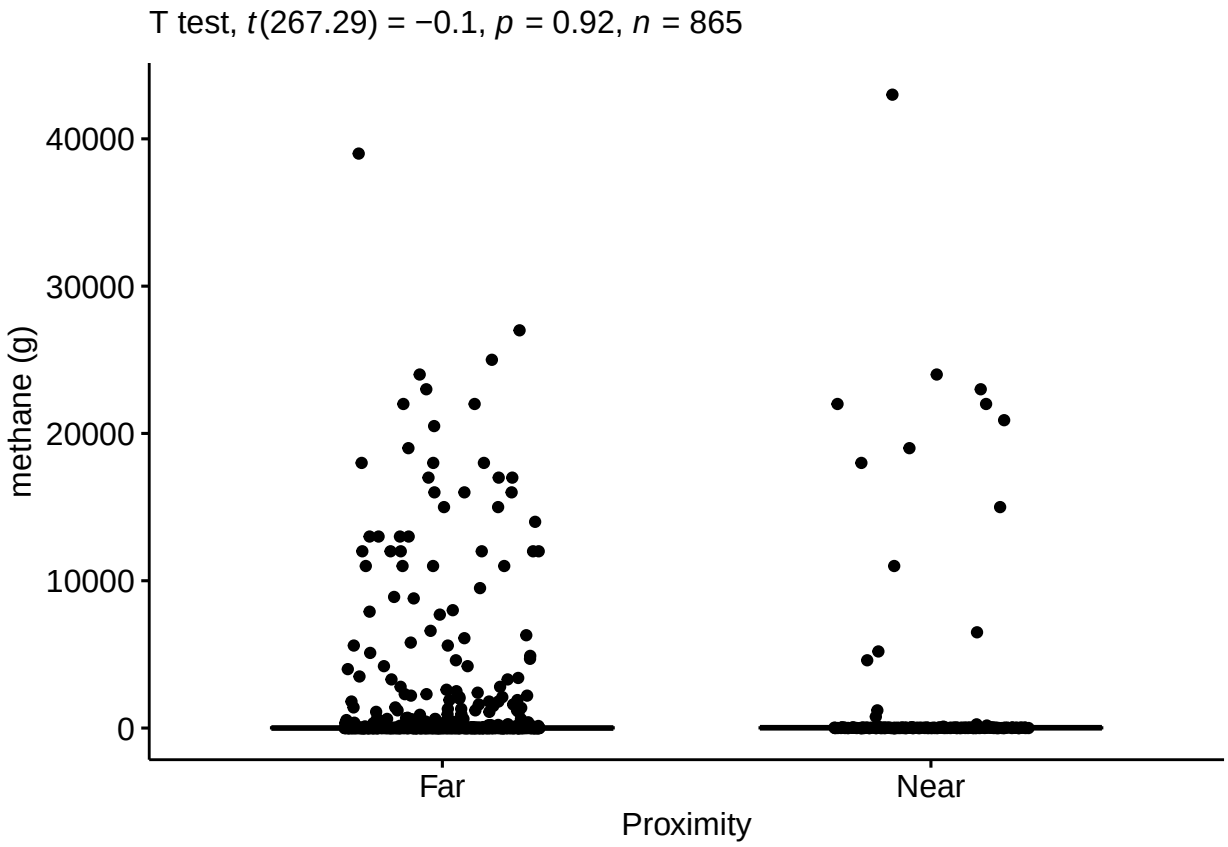
Figure 2.2.7: 2.2 Parametric t-test (2)

```
## # A tibble: 1 x 16
##   estimate estimate1 estimate2 .y.     group1 group2    n1    n2 statistic     p
##      <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>  <int> <int>     <dbl> <dbl>
## 1    -39.2     1186.     1226. methane Far    Near     670   195   -0.0975 0.922
## # i 6 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
## #   alternative <chr>, p.signif <chr>
```

Figure 2.2.8: 2.2 Cohens D test (2)

Figure 2.2.9: 2.2 Report Results (2)

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	-0.0084317	670	195	negligible



2.3.0 (3)

Figure 2.3.1: 2.3 Create a Box Plot (3)

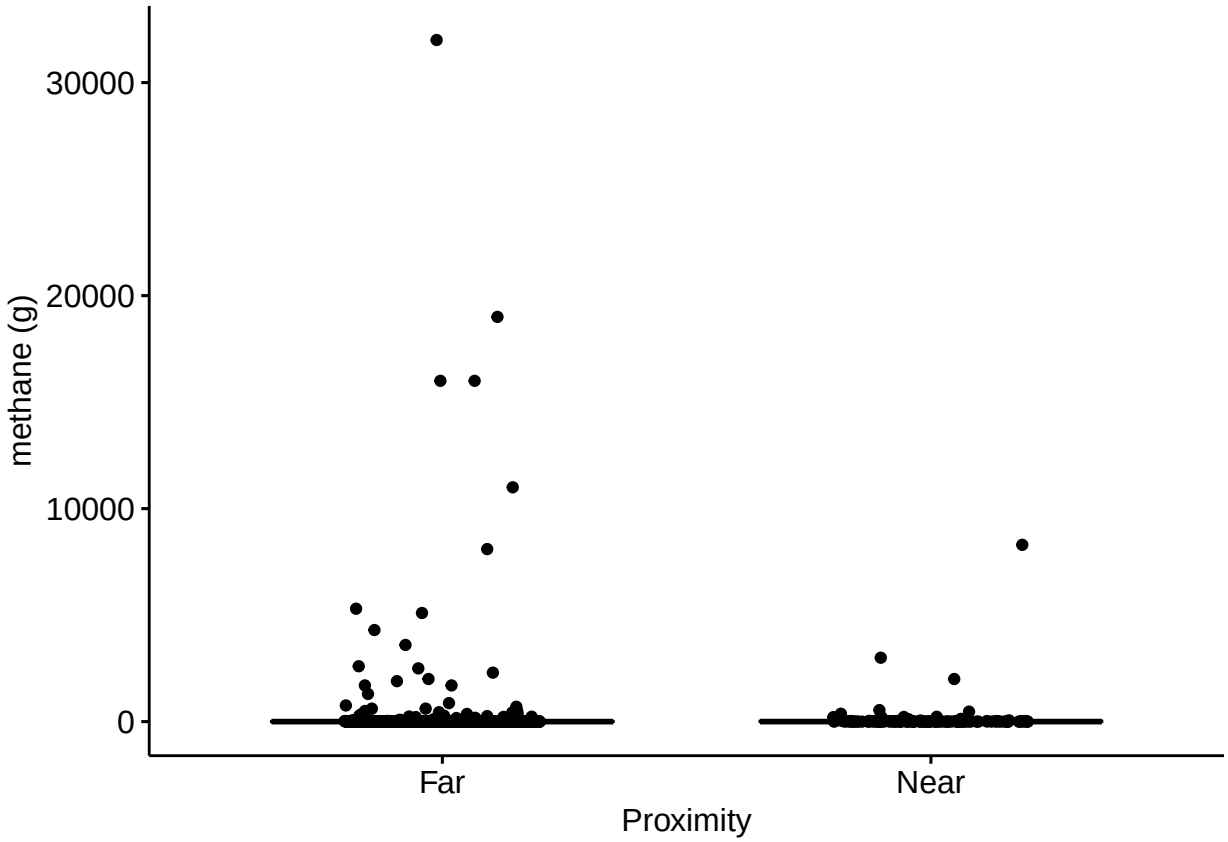


Figure 2.3.2: 2.3 Identify Outliers by Groups (3)

Figure 2.3.3 2.3 Check Normality by Groups (3)

proximity	ID	methane	dl	location	is.outlier	is.extreme
Far	3	26.0	1	Upland	TRUE	TRUE
Far	4	26.0	1	Upland	TRUE	TRUE
Far	17	28.0	0	Upland	TRUE	TRUE
Far	18	230.0	0	Upland	TRUE	TRUE
Far	29	68.0	0	Upland	TRUE	TRUE
Far	31	26.0	1	Upland	TRUE	TRUE
Far	97	610.0	0	Upland	TRUE	TRUE
Far	114	4300.0	0	Upland	TRUE	TRUE
Far	124	430.0	0	Upland	TRUE	TRUE
Far	143	16000.0	0	Upland	TRUE	TRUE
Far	165	8.1	0	Upland	TRUE	FALSE
Far	171	59.0	0	Upland	TRUE	TRUE
Far	172	17.0	0	Upland	TRUE	TRUE
Far	199	180.0	0	Upland	TRUE	TRUE
Far	274	440.0	0	Upland	TRUE	TRUE
Far	324	26.0	1	Upland	TRUE	TRUE
Far	325	26.0	1	Upland	TRUE	TRUE
Far	328	26.0	1	Upland	TRUE	TRUE
Far	330	26.0	1	Upland	TRUE	TRUE
Far	392	40.0	0	Upland	TRUE	TRUE
Far	404	170.0	0	Upland	TRUE	TRUE
Far	445	290.0	0	Upland	TRUE	TRUE
Far	460	26.0	1	Upland	TRUE	TRUE
Far	462	26.0	1	Upland	TRUE	TRUE
Far	463	26.0	1	Upland	TRUE	TRUE
Far	464	26.0	1	Upland	TRUE	TRUE
Far	473	26.0	1	Upland	TRUE	TRUE
Far	483	26.0	1	Upland	TRUE	TRUE
Far	485	26.0	1	Upland	TRUE	TRUE
Far	489	26.0	1	Upland	TRUE	TRUE
Far	495	26.0	1	Upland	TRUE	TRUE
Far	497	26.0	1	Upland	TRUE	TRUE
Far	498	26.0	1	Upland	TRUE	TRUE
Far	507	26.0	1	Upland	TRUE	TRUE
Far	509	26.0	1	Upland	TRUE	TRUE
Far	511	213.0	0	Upland	TRUE	TRUE
Far	540	160.0	0	Upland	TRUE	TRUE
Far	555	5300.0	0	Upland	TRUE	TRUE
Far	589	13.0	0	Upland	TRUE	TRUE
Far	596	1900.0	0	Upland	TRUE	TRUE
Far	614	520.0	0	Upland	TRUE	TRUE
Far	641	760.0	0	Upland	TRUE	TRUE
Far	642	2600.0	0	Upland	TRUE	TRUE
Far	659	2300.0	0	Upland	TRUE	TRUE
Far	686	1700.0	0	Upland	TRUE	TRUE
Far	697	260.0	0	Upland	TRUE	TRUE
Far	712	7.6	0	Upland	TRUE	FALSE
Far	728	5.8	0	Upland	TRUE	FALSE
Far	732	14.0	0	Upland	TRUE	TRUE
Far	735	15.0	0	Upland	TRUE	TRUE
Far	749	16.0	0	Upland	TRUE	TRUE
Far	750	500.0	0	Upland	TRUE	TRUE
Far	799	1300.0	0	Upland	TRUE	TRUE
Far	835	13.0	0	Upland	TRUE	TRUE
Far	873	110.0	0	Upland	TRUE	TRUE
Far	874	870.0	0	Upland	TRUE	TRUE
Far	928	75.0	0	Upland	TRUE	TRUE

proximity	variable	statistic	p
Far	methane	0.0974243	0
Near	methane	0.1511203	0

df1	df2	statistic	p
1	834	0.2304622	0.6313072

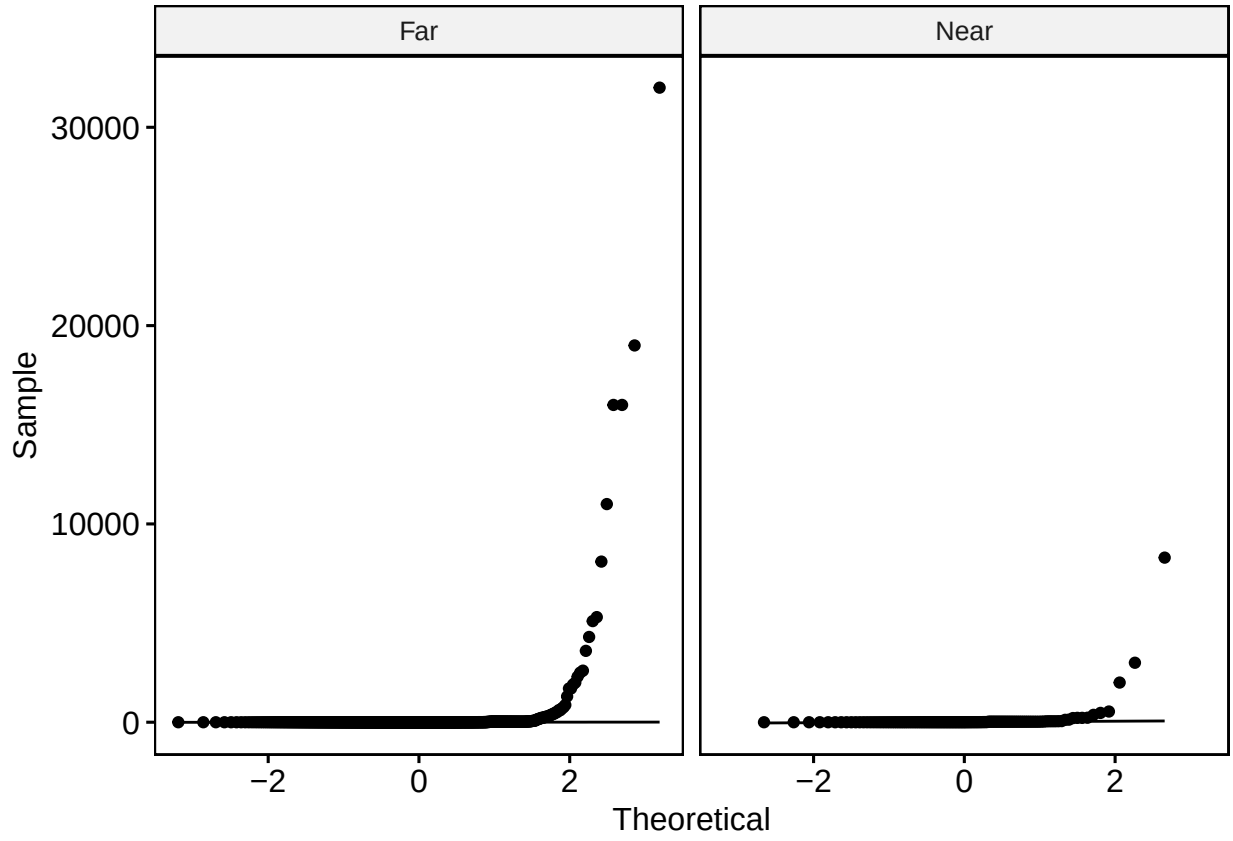


Figure 2.3.4: 2.3 Levine test (3)

Figure 2.3.5: 2.3 Summary Statistics ~ Mean and Interquartile Range ~ (3)

Figure 2.3.6: 2.3 Summary Statistics ~ Mean Methane and Median Methane ~ (3)

Figure 2.3.7: 2.3 Parametric t-test (3)

```
## # A tibble: 1 x 16
##   estimate estimate1 estimate2 .y.     group1 group2    n1    n2 statistic     p
##      <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>  <int> <int>     <dbl> <dbl>
## 1     75.9      210.      134. methane Far    Near     709   127     0.784 0.434
## # i 6 more variables: df <dbl>, conf.low <dbl>, conf.high <dbl>, method <chr>,
## #   alternative <chr>, p.signif <chr>
```

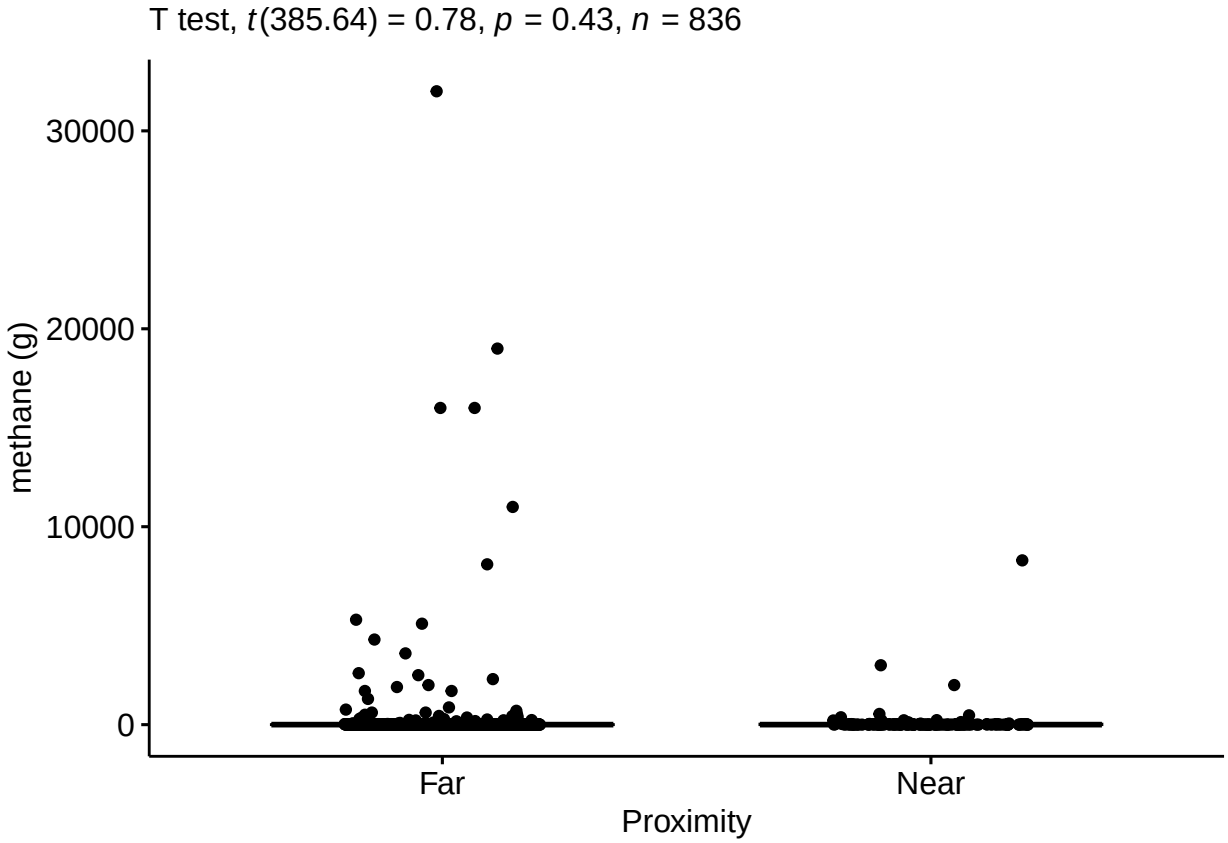
Figure 2.3.8: 2.3 Cohens D test (3)

proximity	variable	n	mean	sd
Far	methane	709	209.731	1753.102
Near	methane	127	133.880	799.431

proximity	variable	n	median	iqr
Far	methane	709	0.4	2.25
Near	methane	127	1.4	25.69

Proximity in the Upland	Mean Methane	Median Methane
Far	209.73	0.4
Near	133.88	1.4

Figure 2.3.9: 2.3 Report Results (3)



2.4.0

Figure 2.4.1: 2.4 Create a Box Plot (4)

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	0.0556729	709	127	negligible

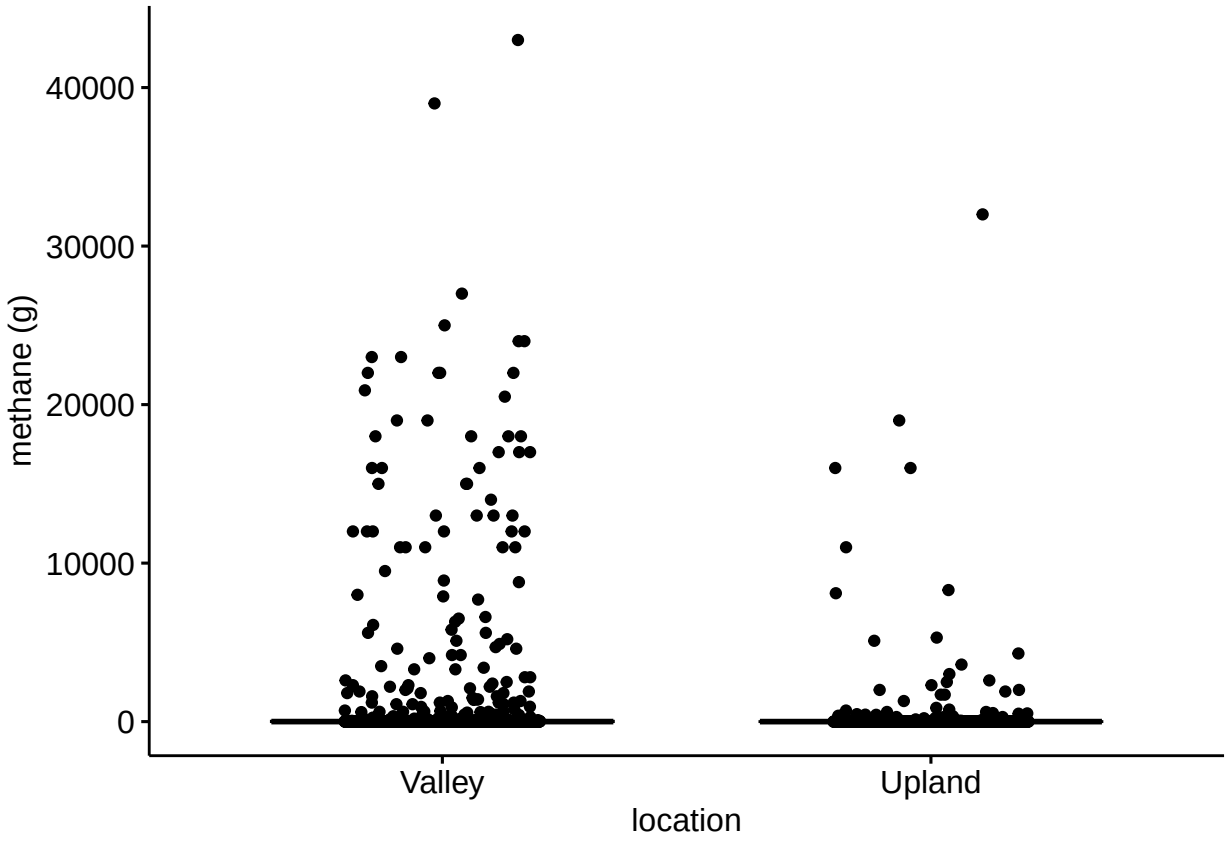


Figure 2.4.2: 2.4 Identify Outliers by Groups (4)

Figure 2.4.3: 2.4 Check Normality by Groups (4)

location	ID	methane	dl	proximity	is.outlier	is.extreme
Upland	3	26.0	1	Far	TRUE	TRUE
Upland	4	26.0	1	Far	TRUE	TRUE
Upland	17	28.0	0	Far	TRUE	TRUE
Upland	18	230.0	0	Far	TRUE	TRUE
Upland	27	57.0	0	Near	TRUE	TRUE
Upland	29	68.0	0	Far	TRUE	TRUE
Upland	31	26.0	1	Far	TRUE	TRUE
Upland	97	610.0	0	Far	TRUE	TRUE
Upland	114	4300.0	0	Far	TRUE	TRUE
Upland	124	430.0	0	Far	TRUE	TRUE
Upland	143	16000.0	0	Far	TRUE	TRUE
Upland	148	120.0	0	Near	TRUE	TRUE
Upland	170	210.0	0	Near	TRUE	TRUE
Upland	171	59.0	0	Far	TRUE	TRUE
Upland	172	17.0	0	Far	TRUE	TRUE
Upland	185	3000.0	0	Near	TRUE	TRUE
Upland	199	180.0	0	Far	TRUE	TRUE
Upland	253	27.0	0	Near	TRUE	TRUE
Upland	274	440.0	0	Far	TRUE	TRUE
Upland	299	26.0	1	Near	TRUE	TRUE
Upland	318	2000.0	0	Near	TRUE	TRUE
Upland	322	26.0	1	Near	TRUE	TRUE
Upland	324	26.0	1	Far	TRUE	TRUE
Upland	325	26.0	1	Far	TRUE	TRUE
Upland	328	26.0	1	Far	TRUE	TRUE
Upland	330	26.0	1	Far	TRUE	TRUE
Upland	331	26.0	1	Near	TRUE	TRUE
Upland	334	26.0	1	Near	TRUE	TRUE
Upland	337	26.0	1	Near	TRUE	TRUE
Upland	338	26.0	1	Near	TRUE	TRUE
Upland	347	26.0	1	Near	TRUE	TRUE
Upland	357	26.0	1	Near	TRUE	TRUE
Upland	366	26.0	1	Near	TRUE	TRUE
Upland	369	26.0	1	Near	TRUE	TRUE
Upland	392	40.0	0	Far	TRUE	TRUE
Upland	401	220.0	0	Near	TRUE	TRUE
Upland	404	170.0	0	Far	TRUE	TRUE
Upland	422	49.0	0	Near	TRUE	TRUE
Upland	445	290.0	0	Far	TRUE	TRUE
Upland	459	46.7	0	Near	TRUE	TRUE
Upland	460	26.0	1	Far	TRUE	TRUE
Upland	462	26.0	1	Far	TRUE	TRUE
Upland	463	26.0	1	Far	TRUE	TRUE
Upland	464	26.0	1	Far	TRUE	TRUE
Upland	466	26.0	1	Near	TRUE	TRUE
Upland	467	26.0	1	Near	TRUE	TRUE
Upland	473	26.0	1	Far	TRUE	TRUE
Upland	483	26.0	1	Far	TRUE	TRUE
Upland	485	26.0	1	Far	TRUE	TRUE
Upland	489	26.0	1	Far	TRUE	TRUE
Upland	495	26.0	1	Far	TRUE	TRUE
Upland	497	26.0	1	Far	TRUE	TRUE
Upland	498	26.0	1	Far	TRUE	TRUE
Upland	507	26.0	1	Far	TRUE	TRUE
Upland	509	26.0	1	Far	TRUE	TRUE
Upland	511	213.0	0	Far	TRUE	TRUE
Upland	540	160.0	0	Far	TRUE	TRUE

location	variable	statistic	p
Upland	methane	0.0977966	0
Valley	methane	0.3085224	0

location	variable	n	mean	sd
Upland	methane	836	198.208	1644.111
Valley	methane	865	1195.243	4331.548

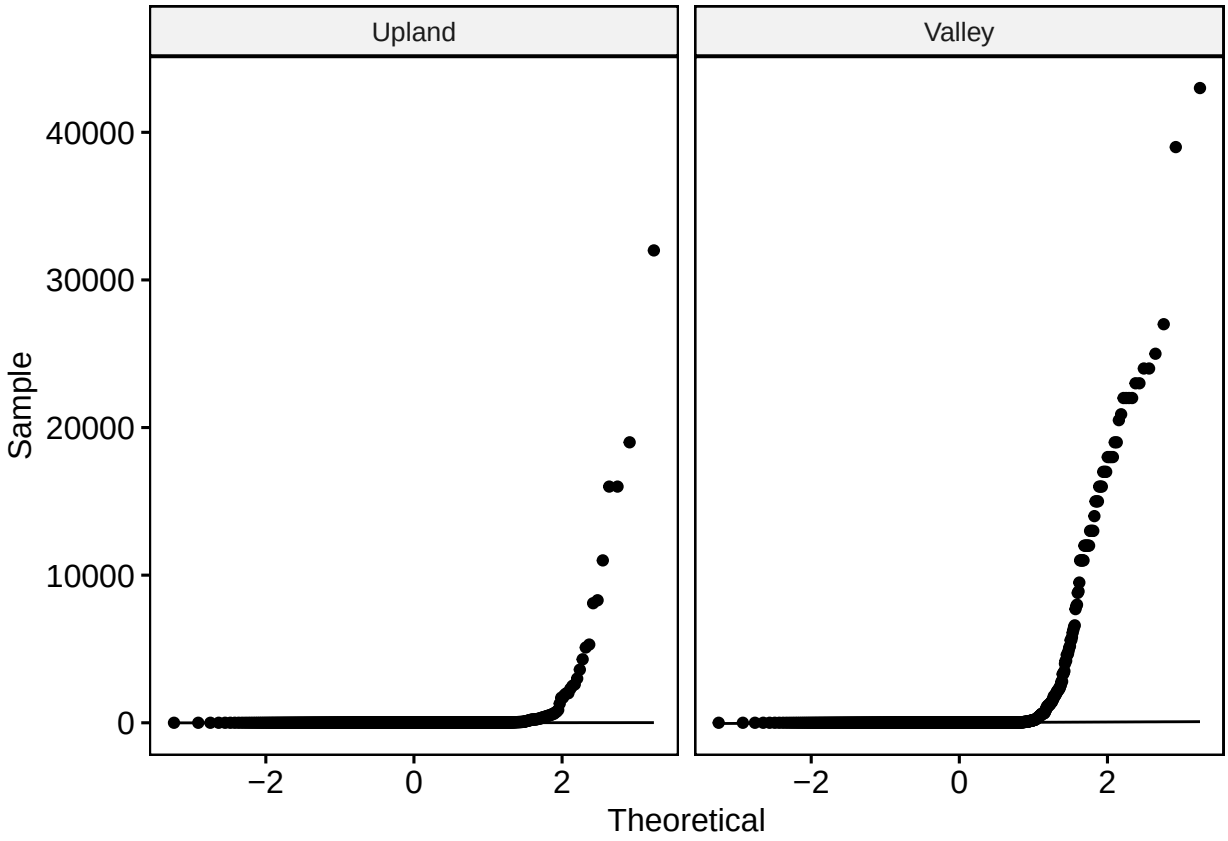


Figure 2.4.4: 2.4 Summary Statistics ~ Mean and Interquartile Range ~ (4)

Figure 2.4.5: 2.4 Summary Statistics ~ Mean Methane and Median Methane ~ (4)

Figure 2.4.6: 2.4 Levine test (4)

Figure 2.4.7: 2.4 Parametric t-test (4)

```
## # A tibble: 2 x 17
##   location estimate estimate1 estimate2 .y.     group1 group2    n1    n2
##   <chr>       <dbl>     <dbl>     <dbl> <chr>   <chr>  <chr>  <int> <int>
## 1 Upland       75.9      210.      134. methane Far    Near     709   127
## 2 Valley      -39.2     1186.     1226. methane Far    Near     670   195
## # i 8 more variables: statistic <dbl>, p <dbl>, df <dbl>, conf.low <dbl>,
```

location	variable	n	median	iqr
Upland	methane	836	0.47	3.565
Valley	methane	865	1.80	25.780

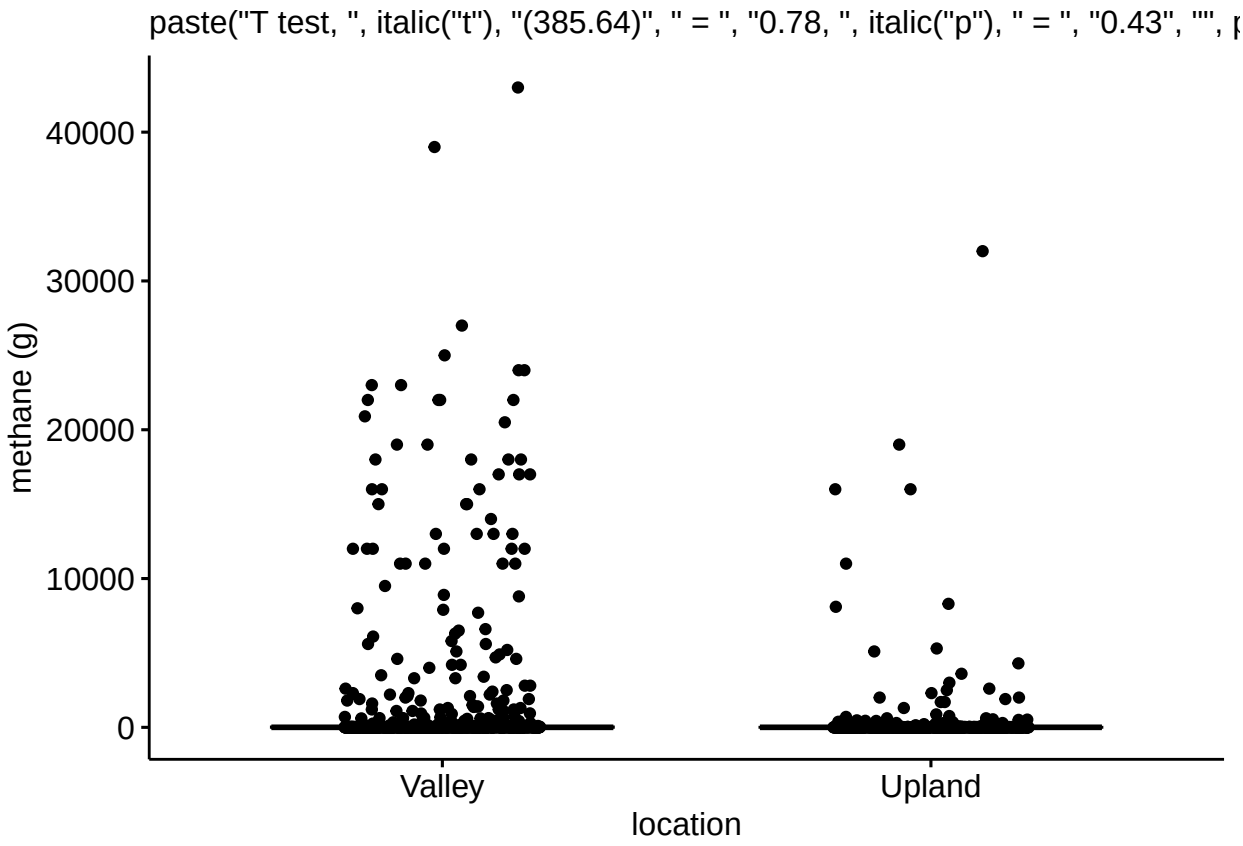
Location	Mean Methane	Median Methane
Upland	198.21	0.47
Valley	1195.24	1.80

location	df1	df2	statistic	p
Upland	1	834	0.2304622	0.6313072
Valley	1	863	0.0112460	0.9155695

```
## #   conf.high <dbl>, method <chr>, alternative <chr>, p.signif <chr>
```

Figure 2.4.8: 2.4 Cohens D test (4)

Figure 2.4.8: 2.4 Report Results (4)



3.0.0: 3.0.0: Non-Parametric t-test and Preliminary test

Figure 3.1.1: 3.1 Compute the Wilcox test (1)

Figure 3.1.2: 3.1 Calculate Effect Size (1)

Figure 3.2.1: 3.2 Compute the Wilcox test (2)

Figure 3.2.2: 3.2 Calculate Effect Size (2)

Figure 3.3.1: 3.3 Compute the Wilcox test (3)

Figure 3.3.2: 3.3 Calculate Effect Size (3)

Figure 3.4.1: 3.4 Compute the Wilcox test (4)

Figure 3.4.2: 3.4 Calculate Effect Size (4)

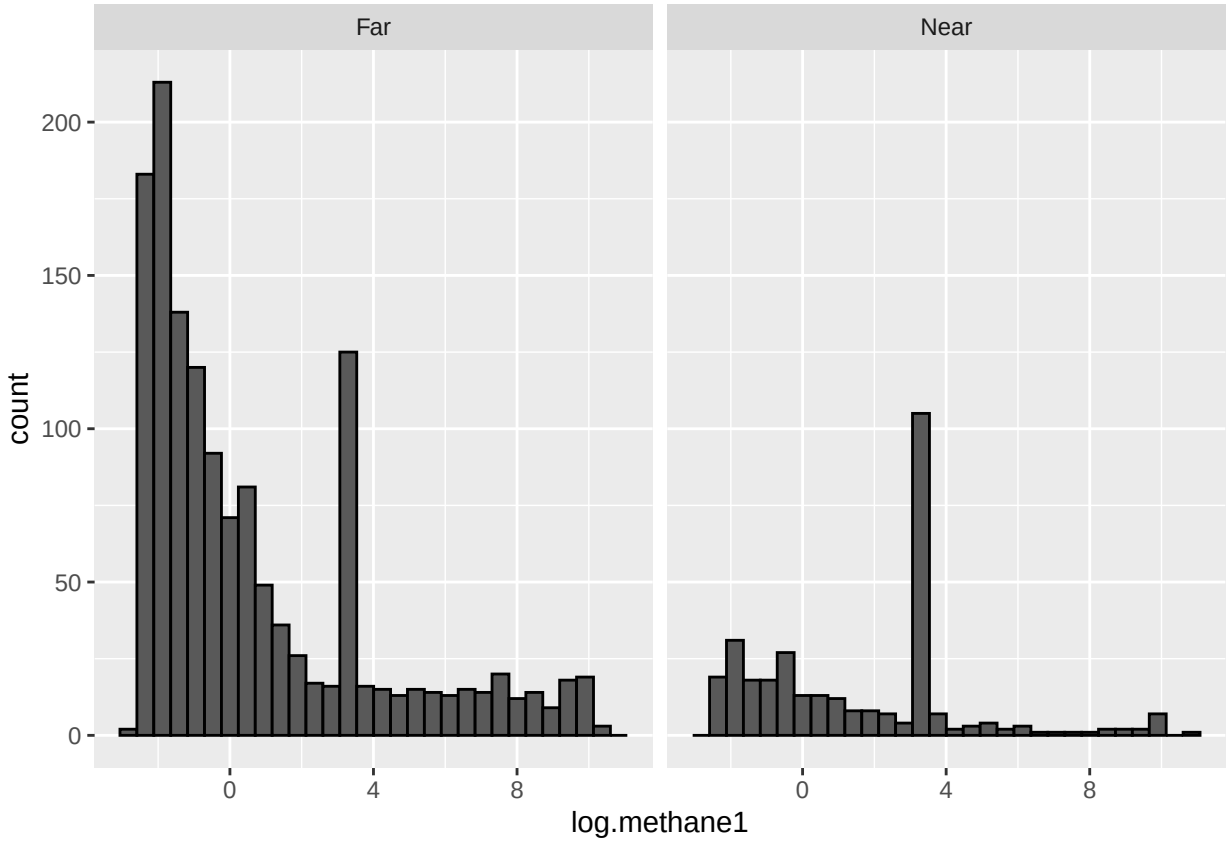
4.0.0: Log Transformations and t-test

.y.	group1	group2	n1	n2	statistic	p	p.signif
methane	Far	Near	1379	322	171369	0	****

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	0.154974	1379	322	small

Figure 4.1.2: 4.1 Histogram of Logged Data (1)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

.y.	group1	group2	n1	n2	statistic	p	p.signif
methane	Far	Near	670	195	55564.5	0.00145	**

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	0.1082596	670	195	small

.y.	group1	group2	n1	n2	statistic	p	p.signif
methane	Far	Near	709	127	32805	1e-06	****

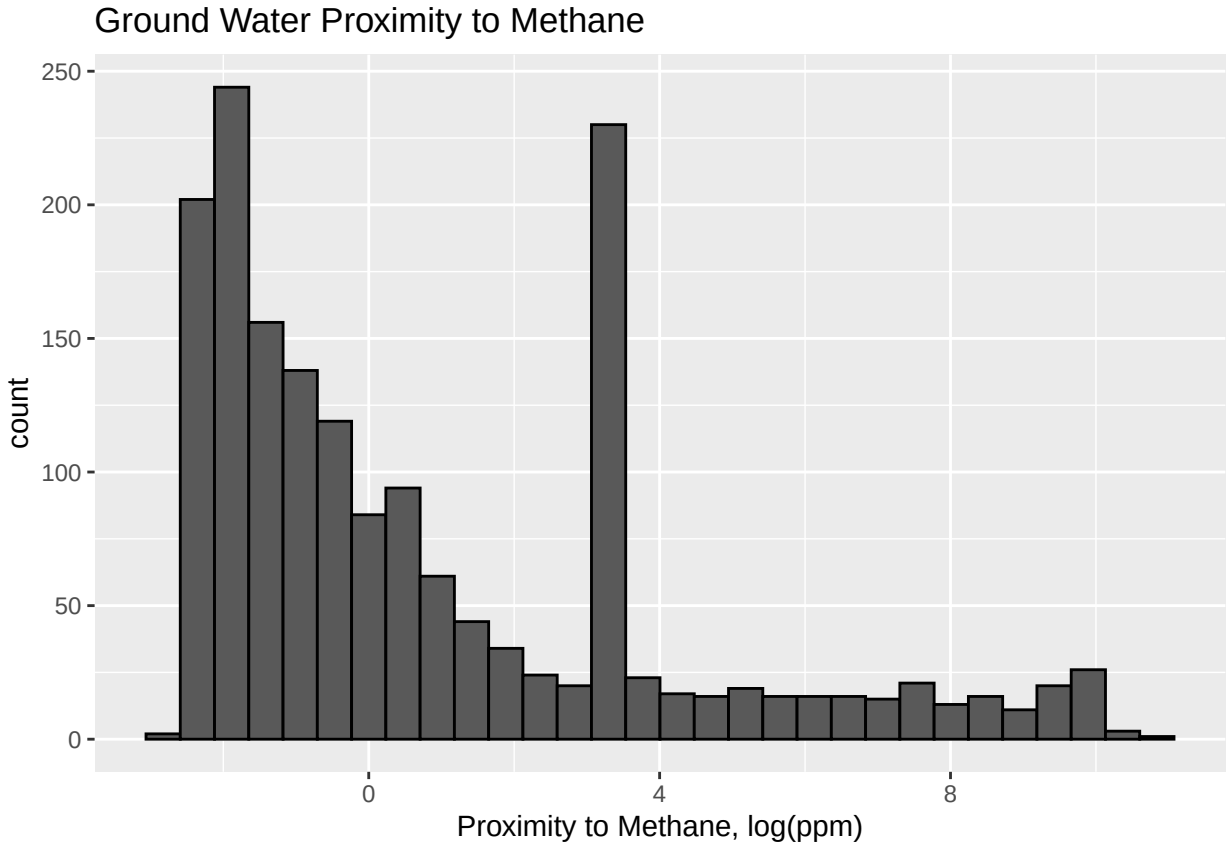


Figure 4.1.3: 4.1 Levine test (1)

Figure 4.1.4: 4.1 Shapiro and Normality Test on Log Value (1)

```
## # A tibble: 2 x 4
##   proximity variable     statistic        p
##   <chr>     <chr>            <dbl>    <dbl>
## 1 Far       log.methane1     0.834 1.19e-35
## 2 Near      log.methane1     0.900 9.47e-14
```

Figure 4.2.1: 4.2 Log Transforming Data (2)

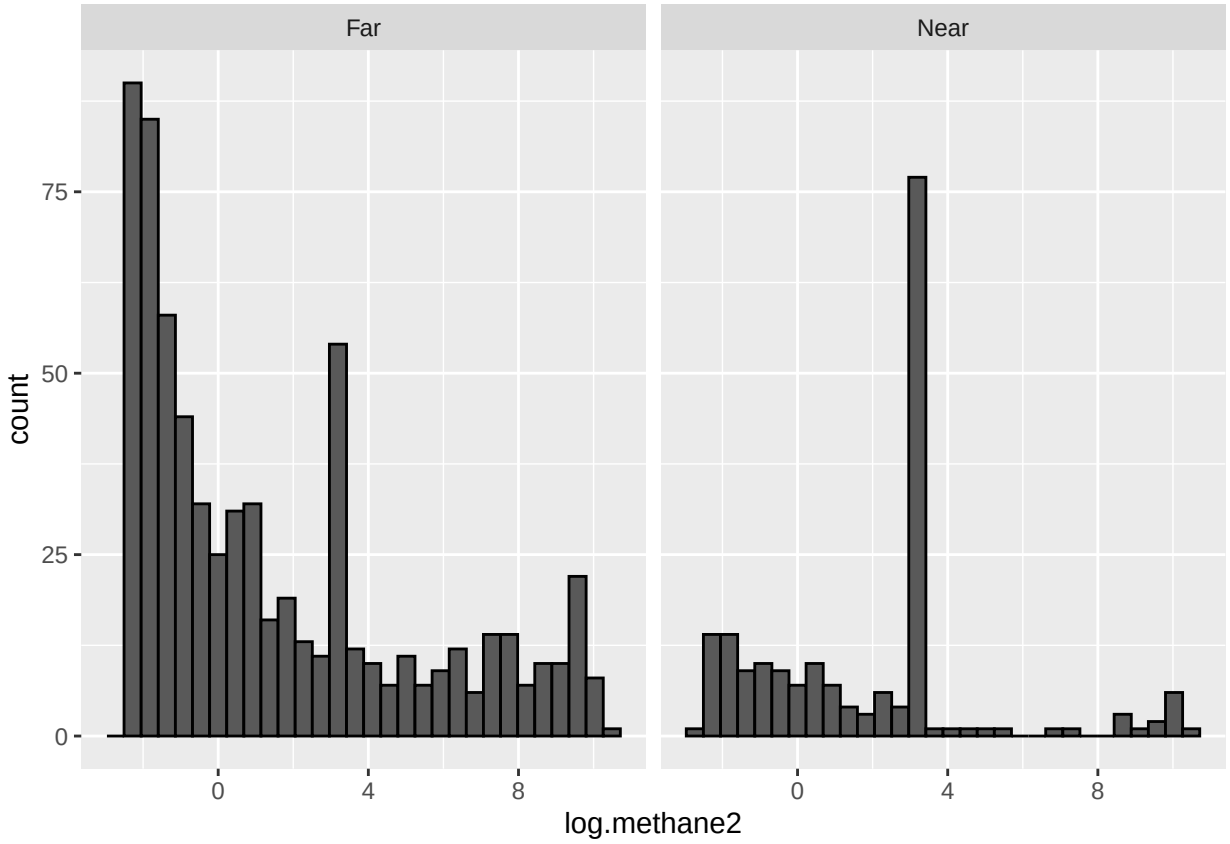
Figure 4.2.2: 4.2 Histogram of Logged Data (2)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

.y.	group1	group2	effsize	n1	n2	magnitude
methane	Far	Near	0.1688332	709	127	small

location	.y.	group1	group2	n1	n2	statistic	p	p.signif
Upland	methane	Far	Near	709	127	32805.0	0.000001	****
Valley	methane	Far	Near	670	195	55564.5	0.001450	**

.y.	group1	group2	effsize	location	n1	n2	magnitude
methane	Far	Near	0.1688332	Upland	709	127	small
methane	Far	Near	0.1082596	Valley	670	195	small




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

df1	df2	statistic	p
1	1699	0.0450635	0.8319127

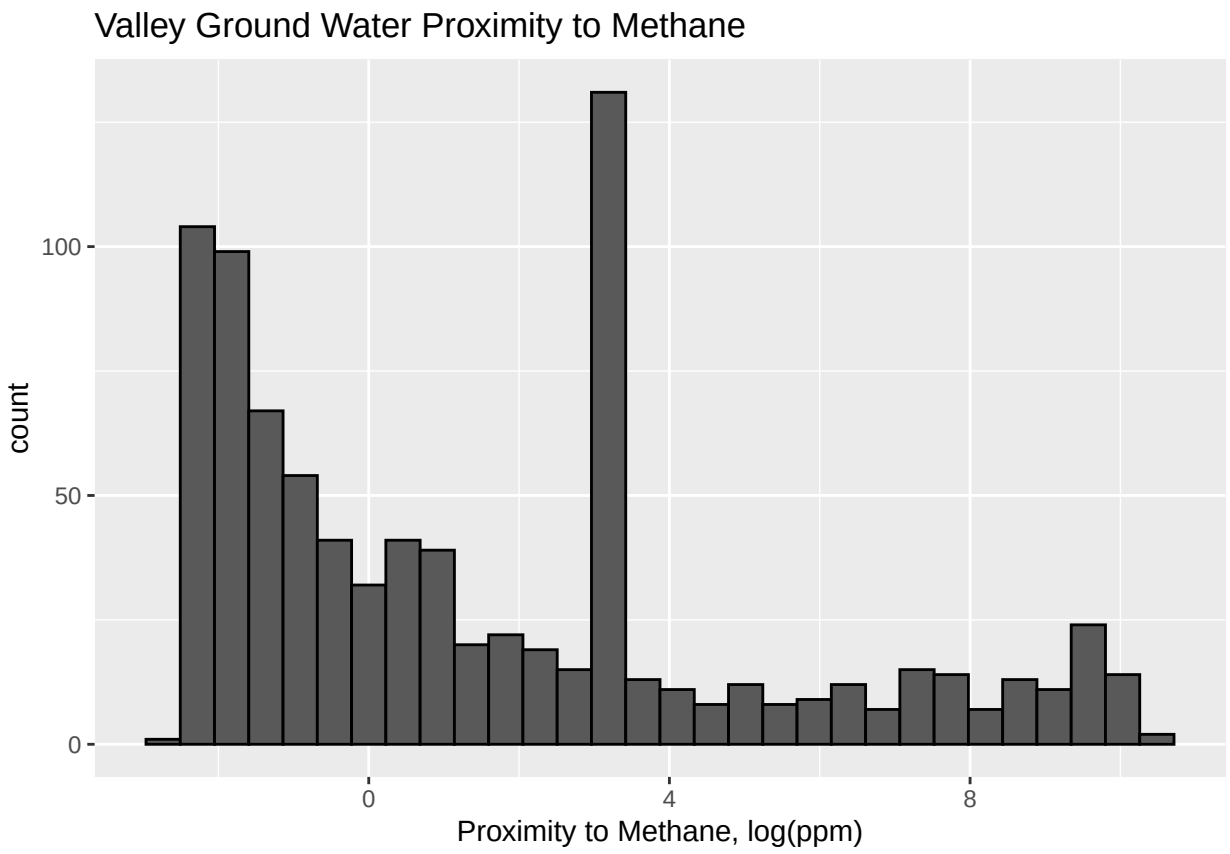


Figure 4.2.3: 4.2 Levine test (2)

```
## # A tibble: 1 x 4
##     df1   df2 statistic       p
##   <int> <int>     <dbl>   <dbl>
## 1     1   863      8.92 0.00290
```

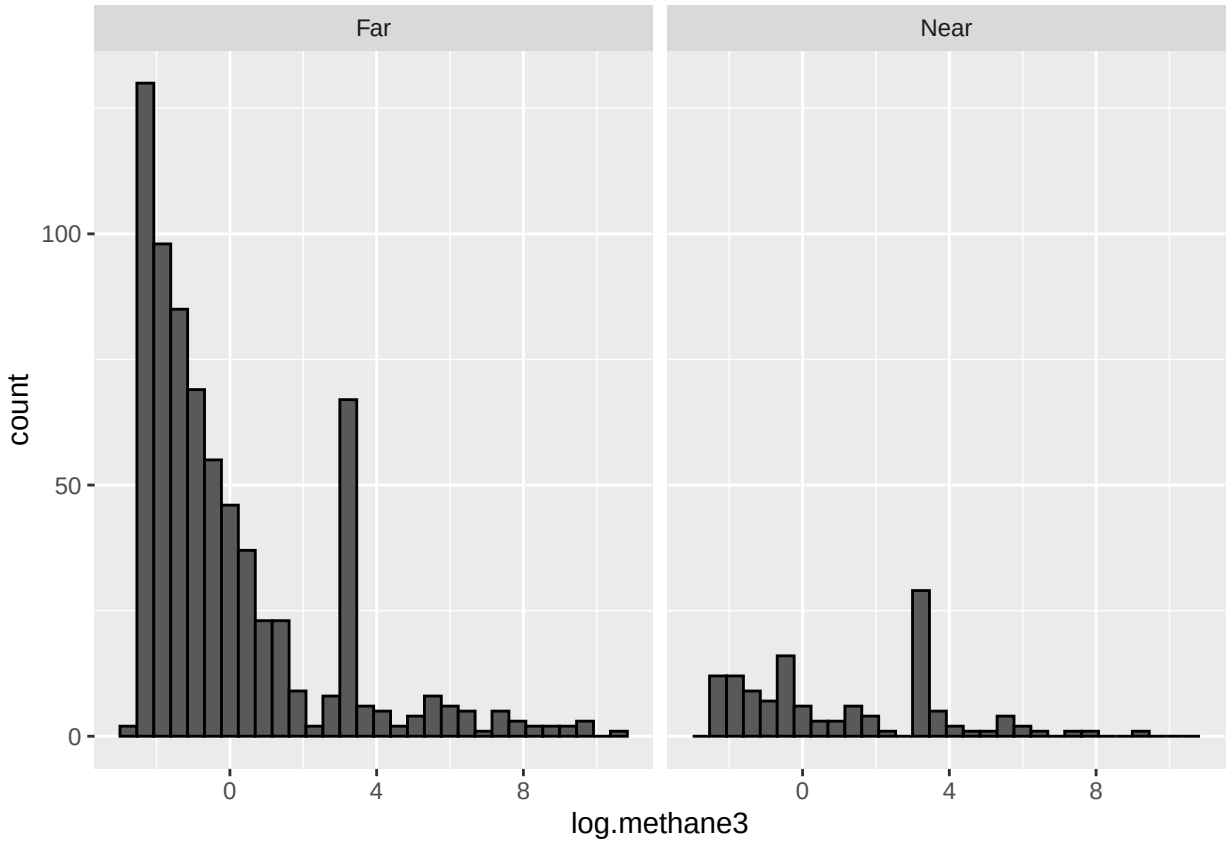
Figure 4.2.4: 4.2 Shapiro and Normality Test on Log Value (2)

```
## # A tibble: 2 x 4
##   proximity variable     statistic        p
##   <chr>     <chr>            <dbl>    <dbl>
## 1 Far       log.methane2     0.865 1.55e-23
## 2 Near      log.methane2     0.877 1.53e-11
```

Figure 4.3.1: 4.3 log transforming data for Uplnd (3)

Figure 4.3.2: 4.3 Histogram of logged data for Upland (3)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

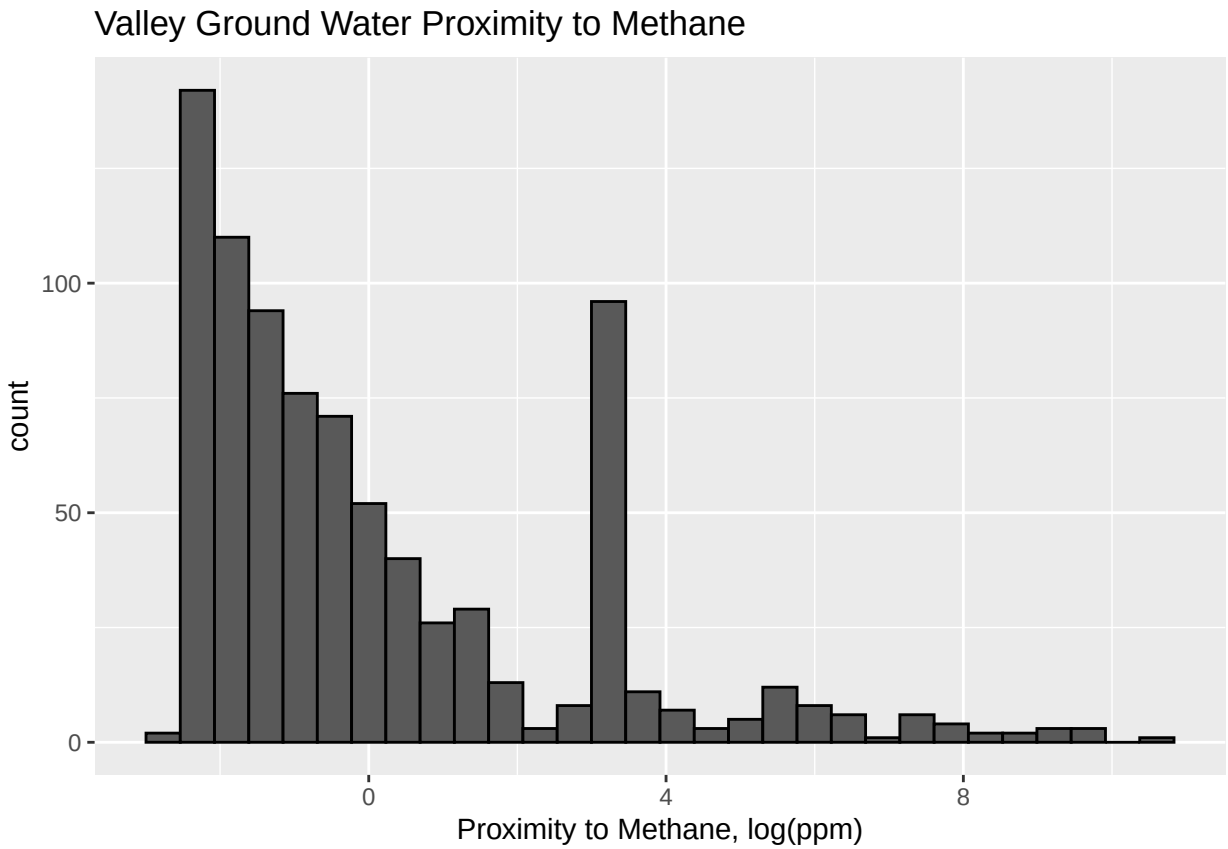


Figure 4.3.3: 4.3 Levine test (3)

```
## # A tibble: 1 x 4
##     df1   df2 statistic       p
##   <int> <int>     <dbl>   <dbl>
## 1     1   834      7.72 0.00558
```

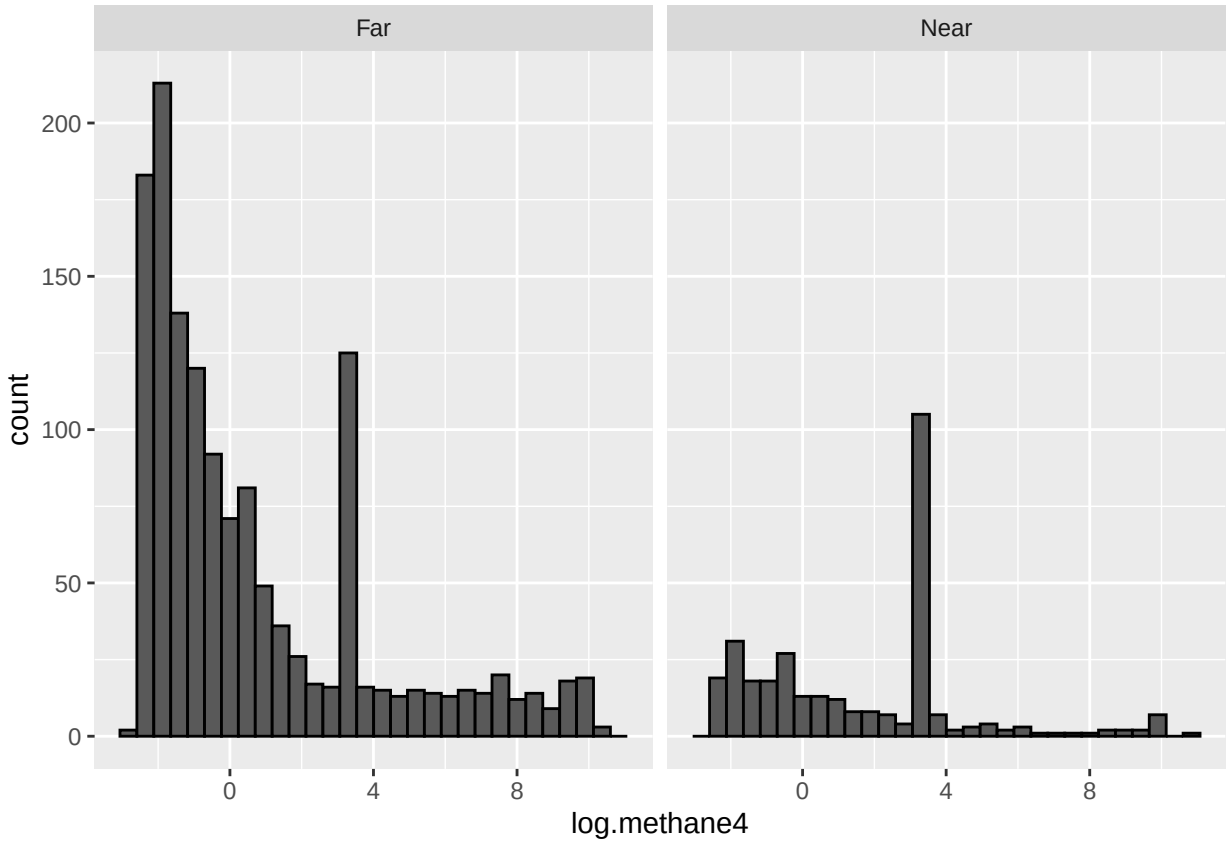
Figure 4.3.4: 4.2 Shapiro and Normality Test on Log Value (3)

```
## # A tibble: 2 x 4
##   proximity variable     statistic        p
##   <chr>     <chr>            <dbl>    <dbl>
## 1 Far       log.methane3     0.821 2.38e-27
## 2 Near      log.methane3     0.918 1.06e- 6
```

Figure 4.4.1: 4.4 log transforming data for Valley (4)

Figure 4.4.2: 4.4 Histogram of logged data for Valley (4)

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

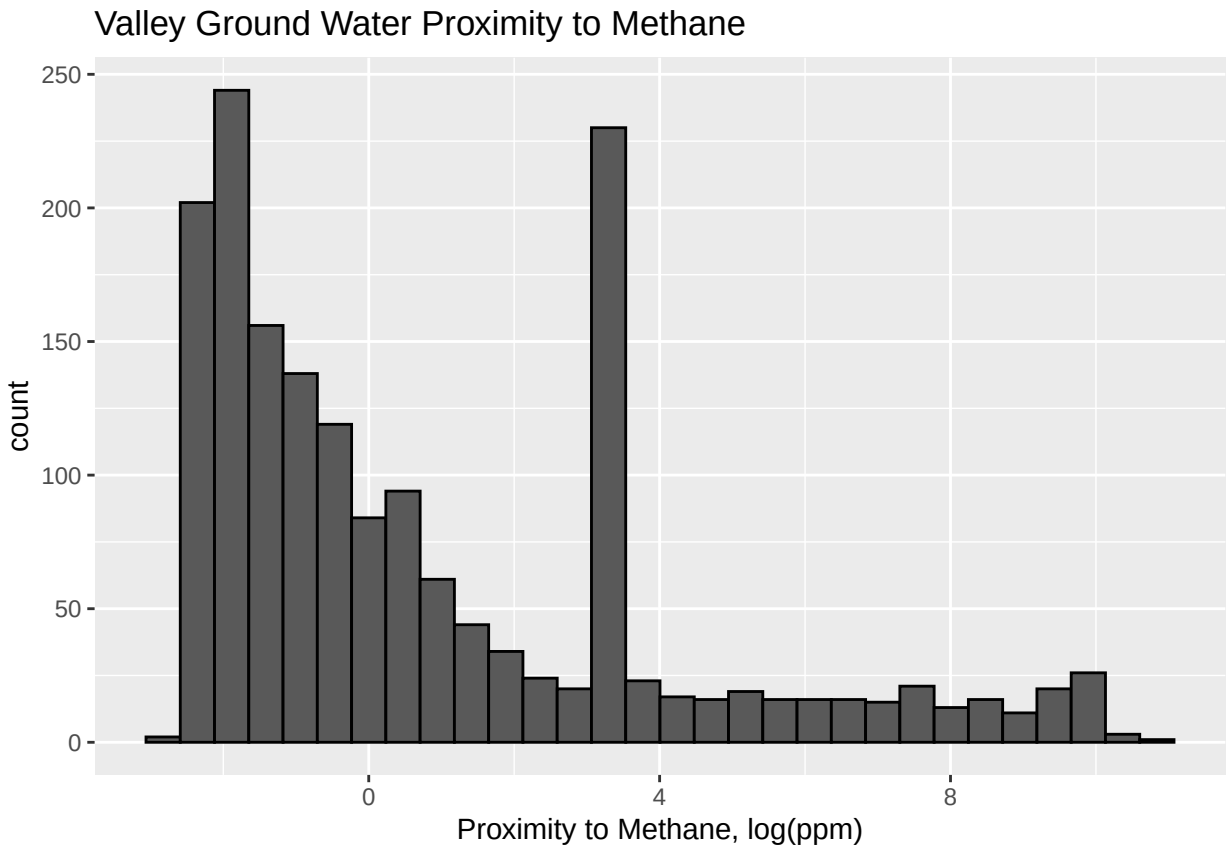


Figure 4.4.3: 4.4 Levine test (4)

```
## # A tibble: 2 x 5
##   location   df1   df2 statistic       p
##   <chr>    <int> <int>     <dbl>   <dbl>
## 1 Upland       1   834      7.72 0.00558
## 2 Valley       1   863      8.92 0.00290
```

Figure 4.4.4: 4.4 Shapiro and Normality Test on Log Value (4)

```
## # A tibble: 2 x 4
##   location variable     statistic        p
##   <chr>    <chr>            <dbl>    <dbl>
## 1 Upland   log.methane4     0.844 6.92e-28
## 2 Valley   log.methane4     0.886 1.08e-24
```