O-Well Analysis

Read in data

We begin by reading in a .RDS file that contains a list of data frames that we will use during our analysis of well water in Gulf of Aqaba, Saudi Arabia.

Summary of data

In March 2012 ground water samples were collected from twenty-three different wells along the coastal area of Gulf of Aqaba, Saudi Arabia. Most of the well are privately owned, dug in swallow aquifers and are located in relative close proximity tot eh east coast, except for well 23 which is a deeply dug well.

> "Results of dissolved metals and physicochemical properties of groundwater samples are presented in Table 1 and Table 2. Metal contents in groundwater samples were low throughout the sampling wells and they are within the range listed for waters suitable for drinking water (WHO, 2008). High concentrations of these metals have been found in the adjacent soil samples and geologic units (Table 3 and Table 4). This suggests that the primary source of dissolved metals to groundwater is not probably metals leached from the surrounding rocks and soils, but rather released from aquifer materials (water-rock interaction). It may also suggest that groundwater aquifer is not significantly recharged from surface runoff or the recharge rate from surface water is low or negligible. This is consistent with the low and erratic annual precipitation rate occurred in the region." - (Journal of Applied Science and Agriculture,2013)

At the end of this document I have provided a data dictionary to help understand some of the abbreviations used in this data set.

Create Factors for Analysis

I decided to turn the geology column in the two data sets into a factor data type. This will help later when we compare the mean values of an element across the different geology groups.

Descriptive Statistics

Below is the summary statistics of the two data sets metals and chemistry. The summary statistics shows the min, max, first and third quartiles, the mean and the median of the the metals and elements. You can also see the different group under the "geology" section. These groups will be used later in the analysis to compare mean values of a metal or chemistry data point across different groups.

Metals Summary Statistics

Below is the summary statistics of the metals data set and includes the different geology groups.

```
##        As                B                Ba               Be        
##  Min.   :0.1000   Min.   : 4.40   Min.   :  3.30   Min.   :0.2000  
##  1st Qu.:0.3000   1st Qu.: 7.00   1st Qu.:  8.75   1st Qu.:0.3000  
##  Median :0.5000   Median : 9.40   Median : 16.70   Median :0.4000  
```

```
##  Mean   :0.6304   Mean   :10.85   Mean   : 28.34   Mean   :0.3652
##  3rd Qu.:0.8000   3rd Qu.:14.20   3rd Qu.: 42.15   3rd Qu.:0.4000
##  Max.   :2.2000   Max.   :22.80   Max.   :100.80   Max.   :0.5000
##
##        Cd             Co               Cr               Cu
##  Min.   :0.1   Min.   :0.1000   Min.   :0.1000   Min.   :0.5000
##  1st Qu.:0.3   1st Qu.:0.1000   1st Qu.:0.4000   1st Qu.:0.5000
##  Median :0.3   Median :0.2000   Median :0.5000   Median :0.6000
##  Mean   :0.3   Mean   :0.1783   Mean   :0.6652   Mean   :0.6174
##  3rd Qu.:0.3   3rd Qu.:0.2000   3rd Qu.:0.7500   3rd Qu.:0.7000
##  Max.   :0.4   Max.   :0.7000   Max.   :2.1000   Max.   :0.9000
##
##        Fe               Hg              Mn               Mo
##  Min.   : 0.200   Min.   : 0.10   Min.   :0.1000   Min.   :11.00
##  1st Qu.: 0.750   1st Qu.: 1.10   1st Qu.:0.1000   1st Qu.:13.50
##  Median : 2.000   Median : 3.20   Median :0.2000   Median :18.00
##  Mean   : 8.078   Mean   :10.77   Mean   :0.4739   Mean   :18.87
##  3rd Qu.: 4.250   3rd Qu.:18.95   3rd Qu.:0.3000   3rd Qu.:23.50
##  Max.   :93.000   Max.   :58.30   Max.   :5.2000   Max.   :31.00
##
##        Pb              Se              Zn          latitude       longitude
##  Min.   :0.100   Min.   :0.100   Min.   :0.10   Min.   :28.44   Min.   :34.79
##  1st Qu.:1.800   1st Qu.:0.450   1st Qu.:0.50   1st Qu.:28.56   1st Qu.:34.88
##  Median :2.200   Median :0.900   Median :1.10   Median :28.60   Median :34.98
##  Mean   :2.013   Mean   :1.252   Mean   :1.53   Mean   :28.75   Mean   :34.97
##  3rd Qu.:2.450   3rd Qu.:2.000   3rd Qu.:1.80   3rd Qu.:28.94   3rd Qu.:35.02
##  Max.   :3.000   Max.   :3.000   Max.   :6.50   Max.   :29.34   Max.   :35.22
##
##      fault                   geology
##  Min.   : 493   Alkaline          :4
##  1st Qu.:1982   Alkaline,Granite  :2
##  Median :3313   Alkaline,Volcanic :1
##  Mean   :3356   Granite           :2
##  3rd Qu.:4856   Gypsum            :5
##  Max.   :6009   Sand              :6
##                 Volcanic          :3
```

Chemistry summary statistics

Below is the summary statistics of the chemistry data set with the "well" column removed.

```
##        pH              Eh             TDS              Ca
##  Min.   :7.000   Min.   :355.0   Min.   :  406   Min.   :214.0
##  1st Qu.:7.350   1st Qu.:376.5   1st Qu.: 1252   1st Qu.:265.5
##  Median :7.500   Median :378.0   Median : 1578   Median :330.0
##  Mean   :7.443   Mean   :378.5   Mean   : 2342   Mean   :381.3
##  3rd Qu.:7.600   3rd Qu.:382.0   3rd Qu.: 2362   3rd Qu.:458.5
##  Max.   :7.800   Max.   :394.0   Max.   :10018   Max.   :900.0
##
##        K               Mg               Na              HCO3
##  Min.   : 3.00   Min.   : 12.00   Min.   :  64.0   Min.   :110.0
##  1st Qu.: 7.50   1st Qu.: 32.00   1st Qu.: 199.0   1st Qu.:131.0
##  Median :17.00   Median : 58.00   Median : 272.0   Median :159.0
##  Mean   :16.87   Mean   : 54.48   Mean   : 686.4   Mean   :158.5
##  3rd Qu.:22.50   3rd Qu.: 70.00   3rd Qu.: 464.5   3rd Qu.:177.0
##  Max.   :39.00   Max.   :133.00   Max.   :3879.0   Max.   :226.0
##
##        Cl              SO4              NO3              F
##  Min.   : 213.0   Min.   :  92.0   Min.   : 7.00   Min.   :0.8000
##  1st Qu.: 603.5   1st Qu.: 250.0   1st Qu.:39.50   1st Qu.:1.0000
##  Median : 745.0   Median : 341.0   Median :42.00   Median :1.0000
##  Mean   :1461.7   Mean   : 448.8   Mean   :39.65   Mean   :0.9826
##  3rd Qu.:1349.0   3rd Qu.: 609.0   3rd Qu.:44.50   3rd Qu.:1.0000
##  Max.   :7455.0   Max.   :1402.0   Max.   :48.00   Max.   :1.1000
##
##       PO4              TH              TA              TS
##  Min.   :0.1000   Min.   : 646.0   Min.   : 90.0   Min.   : 1698
##  1st Qu.:0.1000   1st Qu.: 868.5   1st Qu.:112.5   1st Qu.: 2690
##  Median :0.1000   Median :1078.0   Median :130.0   Median : 3144
##  Mean   :0.1261   Mean   :1176.8   Mean   :132.0   Mean   : 3945
##  3rd Qu.:0.1000   3rd Qu.:1388.0   3rd Qu.:145.0   3rd Qu.: 4008
##  Max.   :0.4000   Max.   :2478.0   Max.   :185.0   Max.   :12722
##
##        SS             COD              BOD              DO
##  Min.   :1292   Min.   :0.2000   Min.   :0.400   Min.   :6.000
##  1st Qu.:1448   1st Qu.:0.4000   1st Qu.:1.400   1st Qu.:6.650
##  Median :1505   Median :0.8000   Median :1.700   Median :6.900
##  Mean   :1603   Mean   :0.8174   Mean   :1.496   Mean   :6.835
##  3rd Qu.:1645   3rd Qu.:1.2000   3rd Qu.:1.900   3rd Qu.:7.050
##  Max.   :2704   Max.   :1.6000   Max.   :2.000   Max.   :7.500
##
##     latitude       longitude         fault                    geology
##  Min.   :28.44   Min.   :34.79   Min.   : 493   Alkaline          :4
##  1st Qu.:28.56   1st Qu.:34.88   1st Qu.:1982   Alkaline,Granite  :2
##  Median :28.60   Median :34.98   Median :3313   Alkaline,Volcanic:1
##  Mean   :28.75   Mean   :34.97   Mean   :3356   Granite           :2
##  3rd Qu.:28.94   3rd Qu.:35.02   3rd Qu.:4856   Gypsum            :5
##  Max.   :29.34   Max.   :35.22   Max.   :6009   Sand              :6
##                                                 Volcanic          :3
```

Observing Metals data

I decided to normalize the metals data set and plot a box plot for each metal. This allows us to easily compare all metals together in one graphic. The reason I normalized is to reduce the distortion of the different values of the metals this allows us to easily compare metals side by side. We can see from (figure 1.) that we have several outliers.

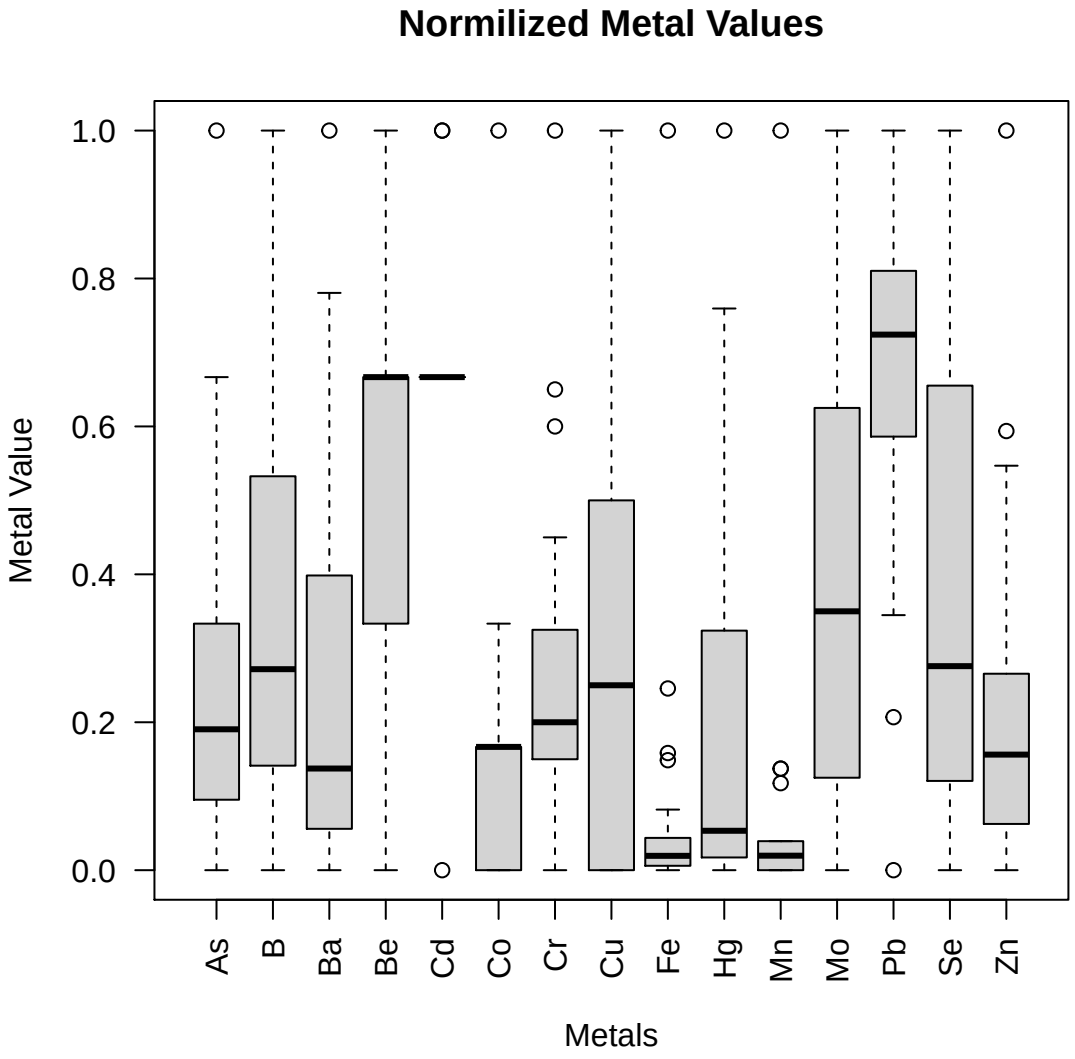

Figure 1: Normilized boxplot of all metals.

Observing Chemistry data

Here is the chemistry data set normalized for easier visualization see (figure 2.).

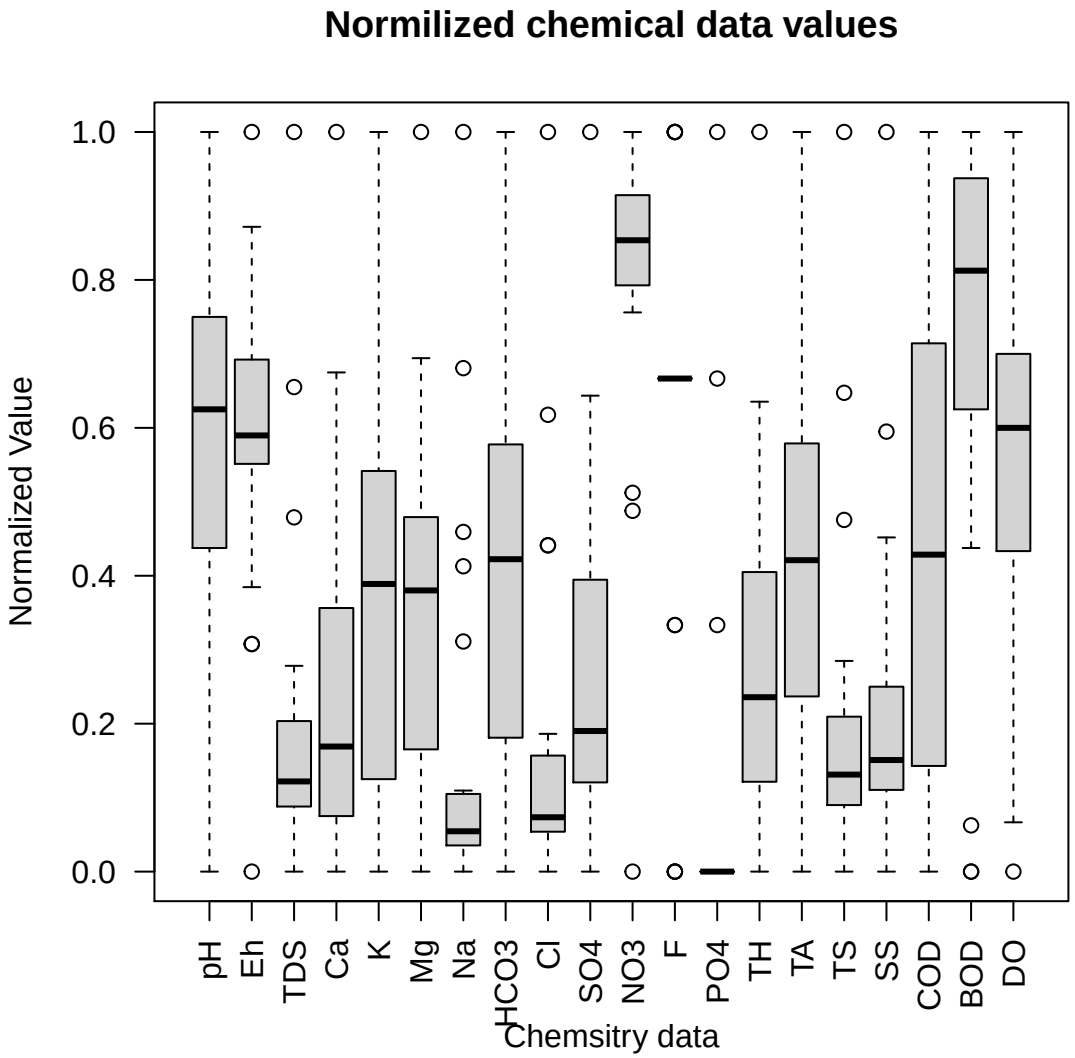

Figure 2: Normilized boxplot of all chemistry.

Lead Analysis

When looking at (figure 2.) I found it interesting that element Pb also known as lead has a higher median value than the other metals. Upon this observation I wanted to see if there is a statistical difference in lead values among the different geology groups.

Lead Analysis Visualization

To see if we have a statistical difference between groups I first decided to visualize the lead values across the different groups using a box plot. Here each box plot represents a different geology group and the body of the box plot represents the lead values from the metals data set. After visual inspection of (figure 3.) we can see that there does seem to be a difference in lead levels across the different groups. To confirm this hypothesis we will use either the one way ANOVA or the Kruskal-Wallis test.

Lets first take a look at the summary statistics of lead across the groups. Then we can visualize this data using the box plots.

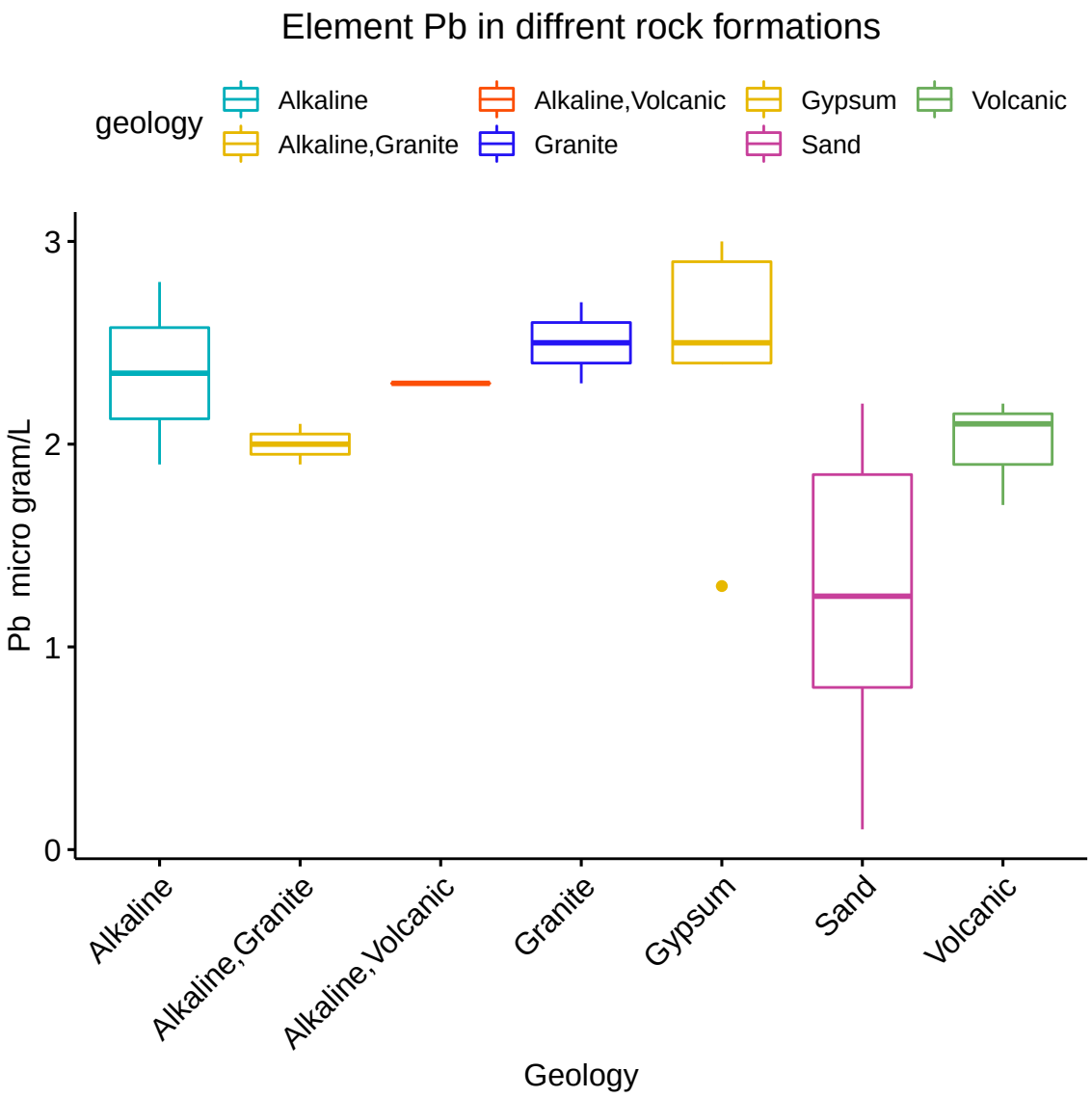

Figure 3: lead values across the diffrent geology groups

One-way ANOVA

A one-way ANOVA test is an extension of independent two-samples t-test for comparing means in a situation where there are more than two groups. In a one-way ANOVA, the data is organized into several groups base on one single grouping variable (also called *factor* variable).

In order to determine if we can use the parametric one-way ANOVA test we need to determine if the data conforms to the one-way ANOVA assumptions.

The Assumptions are:

- The observations are obtained independently and randomly from the population defined by the factor

levels.

- The data of each factor level are normally distributed.
- These normal populations have a common variance. (Levene's test can be used to check this.)

Since the groups are independent we are good for the first assumption. Next we will check for normality.

Checking for Normality

To check for normality we can use the shapiro-test as well as a density plot. One issue with the shapiro test is that you need at least three samples to perform the test. Since some of our groups have less than three samples we can not assume normality for all groups. This means we have failed the second assumption of the one-way ANOVA test. Since we can not fully determine if the data is normally distributed across all groups we will use the non parametric equivalent Kruskal-Wallis test just to be safe. However we can still look at the density plot just for fun.

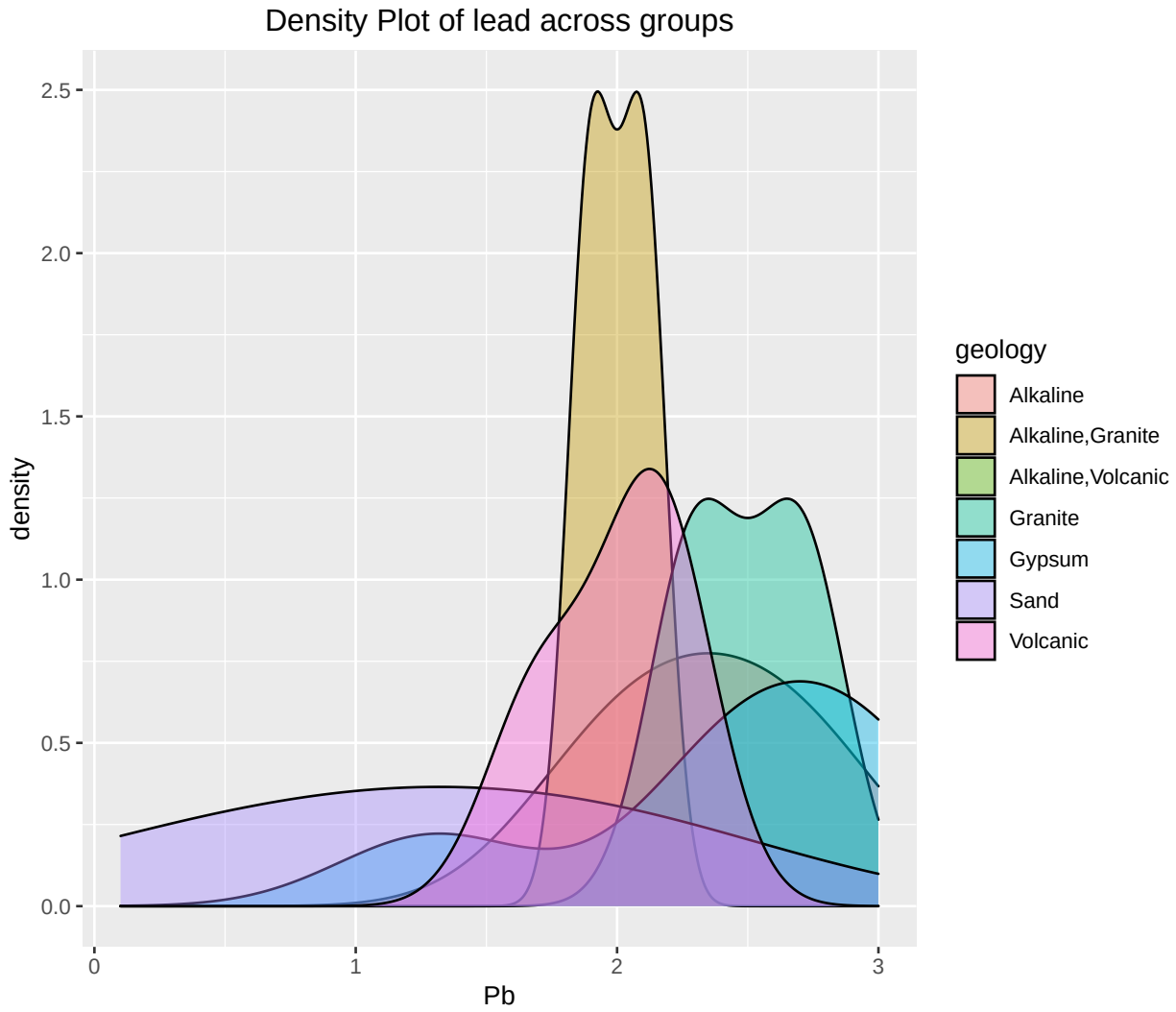

Figure 4: Density plot of lead value across all geology groups

Lead Density Plot

While looking at the density plot (figure 4.) we can see most of the groups follow a normal distribution but since we can not confirm through a second test we will use the Kruskal-Wallis test.

Kruskal-Wallis test on Lead (Pb)

The Kruskal-Wallis test is a non-parametric alternative to the one-way ANOVA test, which extends the two-samples Wilcoxon test in the situation where there are more than two groups. It is recommended when the assumptions of one-way ANOVA test are not met. The null hypothesis of the Kruskal-Wallis test is that there is no significant difference between the groups.

Interpreting the Kruskal-Wallis test on lead (Pb)

After running the Kruskal-Wallis test we can see that the p-value of the test is not less than the alpha value of 0.05. This means we fail to reject the null hypothesis indicating that there is no significant difference of lead levels across the different geology groups.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  metals$Pb by geology
## Kruskal-Wallis chi-squared = 11.349, df = 6, p-value = 0.07816
```

Kruskal-Wallis test across all metals

I decided it would be interesting to see if there are any metals that are statistically significant across the different groups. As we look at the results we can see that no metal has a small enough p-value to be significant.

```
##     As      B     Ba     Be     Cd     Co     Cr     Cu     Fe     Hg     Mn
## 0.9571 0.1011 0.1174 0.2770 0.1901 0.4892 0.3320 0.1086 0.6627 0.3979 0.4672
##     Mo     Pb     Se     Zn
## 0.9384 0.0782 0.2372 0.4440
```

pH Analysis

Looking back at (figure 2.) I found it interesting that pH had a higher median than some of the other chemical data points. I decided to ask the same question as I did with lead. Does the mean pH differ from the various geology groups. To help answer the question we will use the Kruskal-Wallis test as we did with the lead data. We are doing this since we would fail the same assumptions needed to use the one-way ANOVA test.

```
##
##  Kruskal-Wallis rank sum test
##
## data:  chemistry$pH by geology
## Kruskal-Wallis chi-squared = 14.413, df = 6, p-value = 0.02534
```

Interpruting the Kruskal-Wallis test on pH

After performing the Kruskal-Wallis test on pH we can see that the p-value of the test is lower than the alpha value of 0.05 which means we reject the null hypothesis. By rejecting the null hypothesis we can state that there is a statistically significant difference between the mean pH across the various groups. However, the Kruskal-Wallis test does not tell us which groups are different. To determine which groups are different we need to use the Conover-Iman test.

Conover-Iman test

The Conover-Iman test performs a pairwise comparison based on Conover-Iman t-test statistic of rank differences. The Conover-Iman test is strictly valid if and only if the corresponding Kruskal-Willis null hypothesis is rejected.

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 14.4134, df = 6, p-value = 0.03
##
##
##                            Comparison of x by group
##                              (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |   Alkaline   Alkaline   Alkaline    Granite     Gypsum       Sand
## ---------+-----------------------------------------------------------------
## Alkaline |  -2.032081
##          |     0.0776
##          |
## Alkaline |  -2.445821  -0.795820
##          |     0.0462     0.2704
##          |
##  Granite |  -2.344709  -0.270743   0.574759
##          |     0.0484     0.3950     0.3169
##          |
##   Gypsum |  -1.848491   0.621313   1.364290   0.944913
##          |     0.0969     0.3168     0.1674     0.2354
##          |
##     Sand |   1.551906   3.382237   3.459107   3.713829   3.702140
##          |     0.1339    0.0100*    0.0113*    0.0198*    0.0102*
##          |
## Volcanic |  -1.122541   0.988616   1.625664   1.285200   0.523966  -2.629173
##          |     0.2086     0.2363     0.1297     0.1753     0.3189     0.0383
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

Interpruting the Conover-Iman pairwise comparison

The above results show a pairwise matrix between the different geology groups. The top number is the t-test statistic and the bottom number is the p-value. Unfortunately the display of the Conover test does not properly display the names at the top and right side since our group names are too long or because they are comma delimited. To better understand the order of the names we reference our geology levels.

```
## [1] "Alkaline"          "Alkaline,Granite"  "Alkaline,Volcanic"
## [4] "Granite"           "Gypsum"            "Sand"
## [7] "Volcanic"
```

The results from the matrix show that pH is statistically different between Sand and the groups Alkaline_Granite, Alkaline_Volcanic, Granite, and Gypsum. We can visualize the results using a box plot (figure 5.).

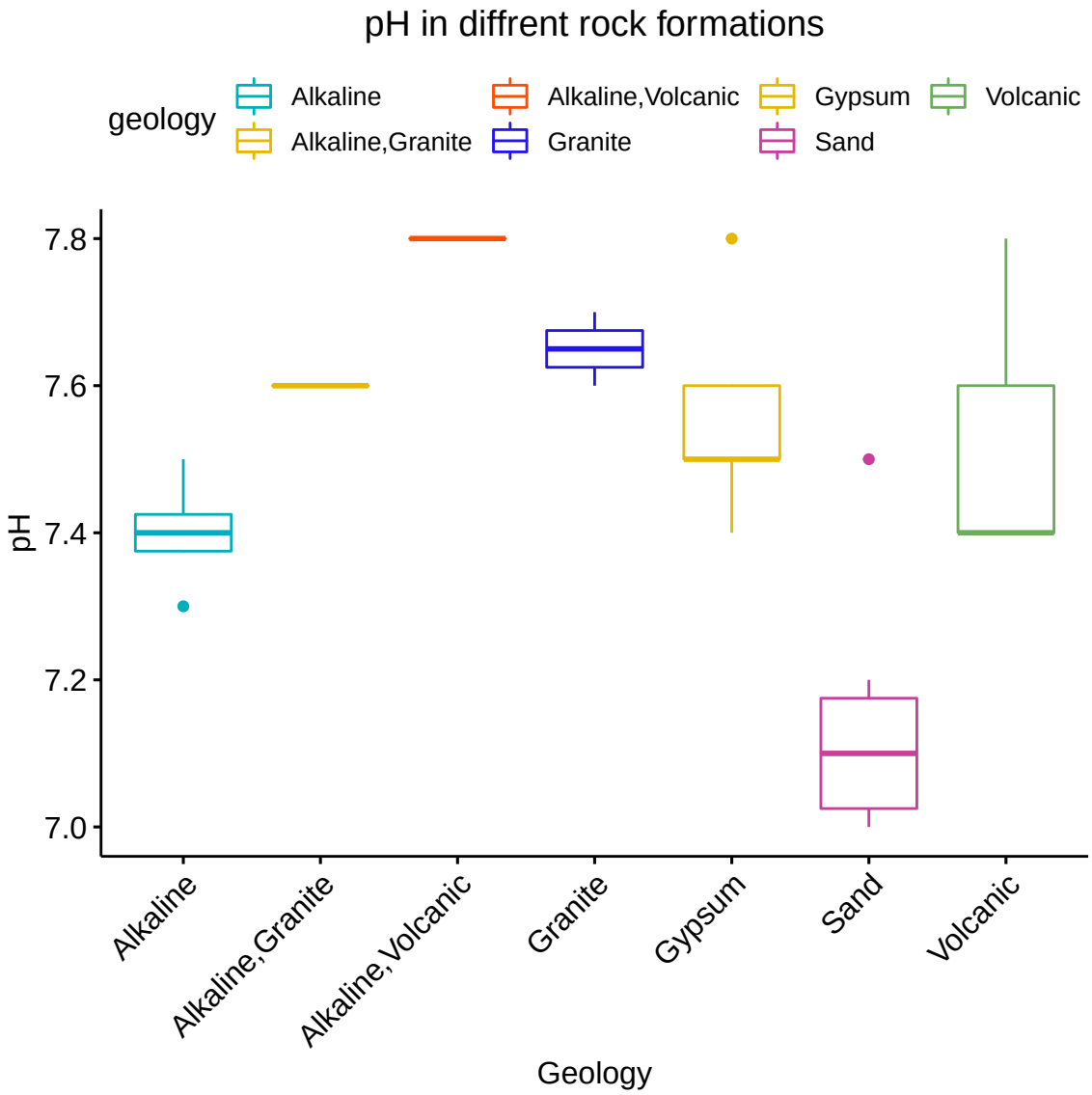

Figure 5: pH values across the diffrent geology groups

Kruskal-Wallis test across all chemistry data points

I decided it would be interesting to see if there are any other chemistry data points that are statistically significant across the different groups. After running the code below we can see that some data points have a small enough p-value to be significant. These data points include pH, TDS, Na, Cl, TS, SS, and latitude. These data points would be interesting to investigate in the future.

```
##          pH          Eh         TDS          Ca           K          Mg
## 0.025344597 0.065890165 0.022902776 0.107992084 0.112560973 0.152918391
##          Na        HCO3          Cl         SO4         NO3           F
## 0.011351209 0.522394845 0.018827850 0.164283253 0.494115116 0.232123029
##         PO4          TH          TA          TS          SS         COD
## 0.157597117 0.145817895 0.809208053 0.020491905 0.031477137 0.169937425
##         BOD          DO    latitude   longitude       fault
## 0.143658425 0.708290834 0.003365011 0.133250117 0.095933530
```

Correlation of chemsitry data

I though I would include a correlation plot to easily visualize the correlation between the data. Below is a correlation plot for the metals data as well as the chemistry data.

Metals Correlation

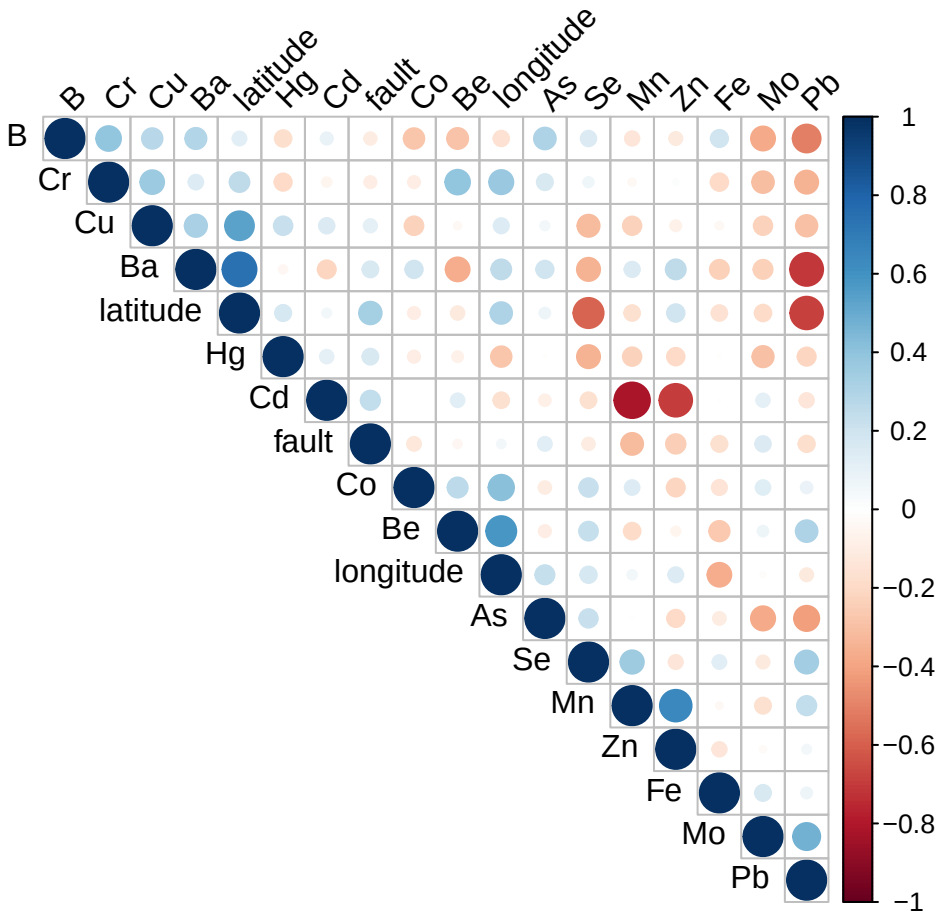

Figure 6: Correlation matrix for the metals dataset

Chemistry Correlation

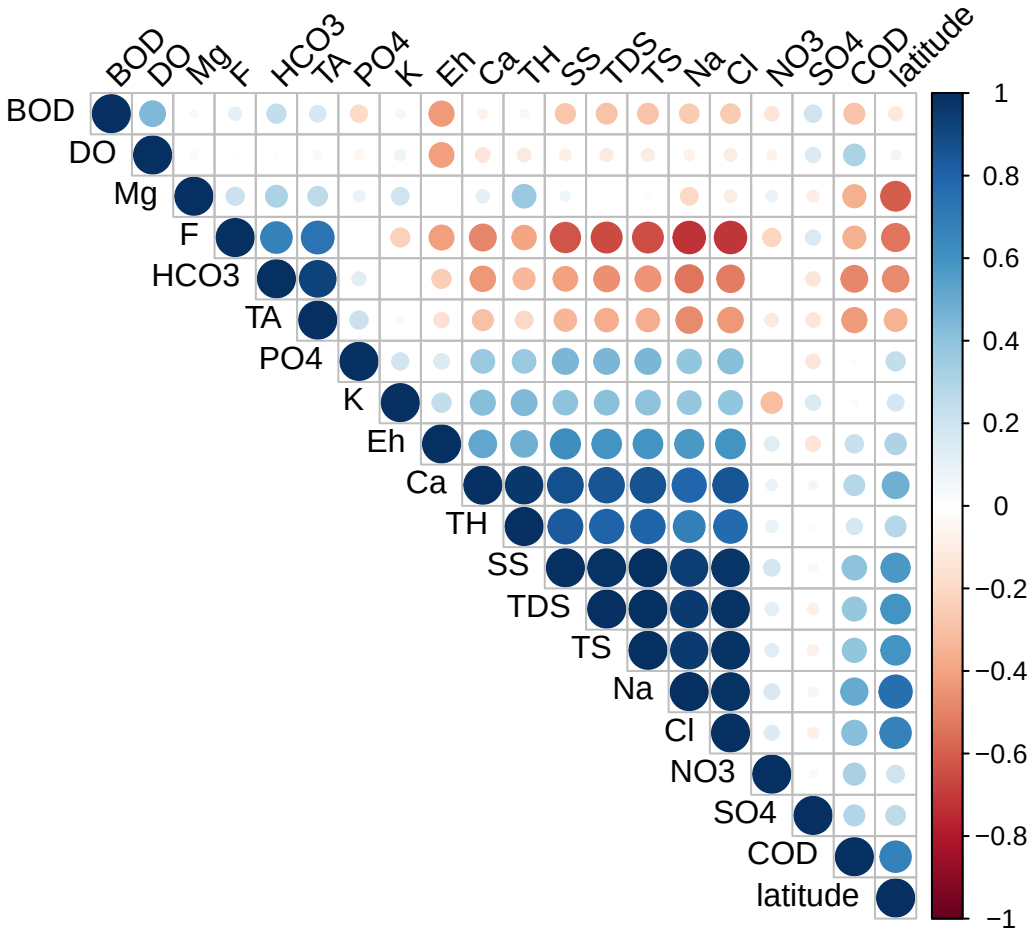

Figure 7: Correlation matrix for the chemistry dataset

Conclusion

After all the analysis we can conclude that the mean lead levels in the well water do not differ based on the geological rock formations. We can also conclude that the mean pH levels do differ across the various geological rock formations. In fact the mean pH value in sand is different from the rock formations Alkaline-Granite, Alkaline-Volcanic, Granite, and Gypsum. We can investigate further by looking at the correlation matrix to see if any linear relationship occur between the various data points. We could also perform principle component analysis to determine which factor are acting as the predictor variables to help explain why the pH level in Sand formations differ from the other rock formations.

Data Dictionary

The following dictionary has been made to better understand the columns names of the two tables. Each value in the table is expressed in μg\L (micro grams).

Metals

As - Arsenic
B - Boron
Ba - Barium
Be - Beryllium
Cd - Cadmium
Co - Cobalt
Cr - Chromium
Cu - Copper Fe - Iron
Hg - Mercury
Mn - Manganese
Mo - Molybdenum
Pb - Lead
Se - selenium
Zn - Zinc

Chemistry

well - The well number
pH - The ph of the well water
Eh - The redox of the well
TDS - Total dissolved solids
Ca - Calcium
K - Potassium
Mg - Magnesium
Na - Sodium
HCO3 - Bicarbonate
Cl - Chlorine
SO4 - Sulfate
NO3 - Nitrate
F - Fluorine Po4 - Phosphate
TH - Total Hardness
TA - Total Alkalinity
TS - Total Solids
SS - Suspended Solids
COD - Chemical Oxygen Demand
BOD - Biological Oxygen Demand
DO - Dissolved Oxygen