R Notebook: O-well3

Josh Pickel [Joshua.Pickel]

Data Summary

The source of the data is from an article written in 2013 called "Groundwater Quality of Coastal Aquifer Systems in the Eastern Coast of the Gulf of Aqaba, Saudi Arabia" written by Ahmed A Al-Taani, Awni T Batayneh, Saad Mogren, Yousef Habous Nazzal. The purpose of the article is to assess shallow groundwater quality along the Gulf of Aqaba. It also aims to determine the status of saline intrusions, the cause of groundwater contamination and the potential for contamination. After modifying the data for table1, it contains 21 columns which represent the well number, different dissolved metals in the 23 wells analyzed in the article, longigtude, latitude, altitude and geological composition. The values are rounded to the nearest 100th for the different metals. The first column is the well number, the 17th, 18th and 19th columns are longitude, latitude and altitude respectively. Columns 20 and 21 are the distance in meters to the nearest fault, and the geological composition of the well respectively. For table2, it contains 26 columns which represent the well number, different physical and chemical parameters for the 23 wells analyzed longitude, latitude, altitude and geological composition. The first column is the well number, the 22nd, 23rd and 24th columns are longitude, latitude and altitude respectively. Columns 25 and 26 are the distance in meters to the nearest fault, and the geological composition of the well respectively.

Data Dictionary

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

UNIT OF MEASUREMENT: ug/L

pH: The pH of the water
Eh: redox measurement of the water (oxidation-reduction)
TDS: Total Dissolved Solids
Ca: Calcium
K: Potassium
Mg: Magnesium
Na: Sodium

HCO3: Bicarbonate
Cl: Chlorine
SO4: Sulfate
NO3: Nitrate
F: Flourine
PO4: Phosphate
TH: Thorium
TA: Tantalum
TS: Tennessine
SS: Stainless Steel
COD: Chemical Oxygen Demand
BOD: Biological Oxygen Demand
DO: Dissolves Oxygen

UNITS: mV mg/L

Reading Data

Accessing .RDS remotely

To read in the data remotely, I used the github.path specifying where the .rds file is located on my github repository, and read it in using the tidyverse package. I saved it as a variable so I can access the data frames stored in the 'wells.rds' list.

Merging Chemistry and Metals Data Frames

To merge the two data frames, I used the 'merge' function to join the two tables where they share attributes. So for the 'chemistry' and 'metals' data frames, this would be 'well', 'latitude', 'longitude', 'altitude.ft', 'fault' and 'geology'. The result is one data frame that contains the shared attributes, as well as the values for chemistry and metals for each well. So there is now one data frame with all the relevant information needed for analyses for each well number. I then saved the data frame as a '.txt' file with a pipe separated delimiter named 'saudi.txt'

Overall Basics

Summary

Looking at the summary, we can see there are several attributes with high variance. I wrote my own functions to compute the various summary statistics reported here, and wrote them such that I can call on each function individually and get the result with the attribute name. Being the variance seems to be quite different from attribute to attribute, I would like to visualize the variances.

```
##                      well          latitude          longitude      altitude.ft
## Variance               46 0.105834753346056 0.0150670288908024 695520.300395257
## SD       6.78232998312527 0.325322537408732   0.12274782641987 833.978597084636
## Max                    23         29.336922          35.218209             3168
## Min                     1         28.438361          34.786616              257
## Mode                    1         28.438361          34.786616              257
## Mean                   12  28.7524256521739   34.9670935652174 1415.86956521739
## Median                 12         28.603468          34.976893             1245
```

```
## Q1                    11         28.602789          34.969706             1167
## Q3                    17         28.651337          35.017044             1815
##                    fault                 pH                 Eh              TDS
## Variance  1955271.6916996 0.0580237154150197 70.3517786561265 4765538.45059289
## SD       1398.31029878908  0.240881122994351 8.38759671515783 2183.01132626308
## Max                  6100                7.8                394            10018
## Min                   238                  7                355              406
## Mode                  238                7.4                377              406
## Mean     1491.65217391304   7.44347826086956 378.478260869565 2342.21739130435
## Median                900                7.5                378             1578
## Q1                    850                7.4                378             1471
## Q3                   1865                7.6                382             2142
##                       Ca                  K                 Mg               Na
## Variance 25488.8577075099 121.573122529644 775.897233201581 920063.794466403
## SD       159.652302543715 11.0260202489223 27.8549319367609 959.199559250526
## Max                   900                39                133             3879
## Min                   214                 3                 12               64
## Mode                  214                10                 29              109
## Mean     381.304347826087 16.8695652173913 54.4782608695652 686.391304347826
## Median                330                17                 58              272
## Q1                    316                16                 56              267
## Q3                    456                21                 68              447
##                      HCO3                Cl                SO4              NO3
## Variance 1096.62450592885 2957570.32806324 102398.150197628 76.6916996047431
## SD       33.1153213170106 1719.75879938532 319.997109670741 8.75737972253933
## Max                   226              7455               1402               48
## Min                   110               213                 92                7
## Mode                  128               710                120               40
## Mean     158.521739130435 1461.65217391304 448.826086956522 39.6521739130435
## Median                159               745                341               42
## Q1                    153               710                331               41
## Q3                    171              1349                606               44
##                         F                 PO4               TH
## Variance 0.00786561264822135 0.00565217391304348 182589.604743083
## SD        0.0886882892394557  0.0751809411556112 427.305048815343
## Max                      1.1                 0.4             2478
## Min                      0.8                 0.1              646
## Mode                       1                 0.1              646
## Mean       0.982608695652174   0.126086956521739 1176.82608695652
## Median                     1                 0.1             1078
## Q1                         1                 0.1             1033
## Q3                         1                 0.1             1376
##                       TA                  TS                SS               COD
## Variance 660.770750988142 6194086.74703557  95438.533596838 0.159683794466403
## SD       25.7054615011702 2488.79222657006 308.931276495013 0.399604547604758
## Max                   185               12722              2704               1.6
## Min                    90                1698              1292               0.2
## Mode                  105                1698              1292               0.4
## Mean      131.95652173913 3944.73913043478 1602.52173913043 0.817391304347826
## Median                130                3144              1505               0.8
## Q1                    130                2976              1501               0.8
## Q3                    140                3762              1620               1.2
##                       BOD                 DO                As                B
## Variance 0.237707509881423 0.181462450592885 0.248577075098814 22.9516996047431
```

```
## SD       0.487552571402739 0.425984096643156 0.498575044600925 4.79079321247986
## Max                      2               7.5               2.2              22.8
## Min                    0.4                 6               0.1               4.4
## Mode                   1.9                 7               0.3                 7
## Mean      1.49565217391304  6.83478260869565 0.630434782608696 10.8521739130435
## Median                 1.7               6.9               0.5               9.4
## Q1                     1.6               6.9               0.5               9.3
## Q3                     1.9                 7               0.8              13.2
##                         Ba                Be                Cd
## Variance 786.776126482213 0.00509881422924901 0.00272727272727273
## SD       28.0495298798788  0.0714059817469728  0.0522232967867094
## Max                 100.8                 0.5                 0.4
## Min                   3.3                 0.2                 0.1
## Mode                  7.1                 0.4                 0.3
## Mean     28.3391304347826   0.365217391304348                 0.3
## Median               16.7                 0.4                 0.3
## Q1                   13.9                 0.4                 0.3
## Q3                   41.2                 0.4                 0.3
##                        Co                Cr                Cu
## Variance 0.0163241106719368 0.194189723320158 0.0124110671936759
## SD        0.127765843134763  0.44066963058527  0.111404969340133
## Max                     0.7               2.1               0.9
## Min                     0.1               0.1               0.5
## Mode                    0.1               0.4               0.6
## Mean      0.178260869565217 0.665217391304348  0.617391304347826
## Median                  0.2               0.5               0.6
## Q1                      0.1               0.5               0.6
## Q3                      0.2               0.7               0.7
##                         Fe                Hg                Mn                Mo
## Variance 374.961778656126 238.841106719368  1.10837944664032 35.6640316205534
## SD       19.3639298350342 15.4544850033693   1.0527960137844 5.97193700741672
## Max                    93             58.3               5.2                31
## Min                   0.2              0.1               0.1                11
## Mode                  0.5              0.1               0.1                11
## Mean     8.07826086956522 10.7739130434783 0.473913043478261 18.8695652173913
## Median                 2              3.2               0.2                18
## Q1                     2              2.3               0.2                18
## Q3                     4             18.7               0.3                23
##                         Pb                Se                Zn
## Variance 0.499367588932806 0.839881422924901 2.32494071146245
## SD       0.706659457541471 0.916450447610181  1.5247756265964
## Max                      3                 3               6.5
## Min                    0.1               0.1               0.1
## Mode                   2.2               1.5               0.3
## Mean      2.01304347826087  1.25217391304348  1.5304347826087
## Median                 2.2               0.9               1.1
## Q1                     2.1               0.9               0.9
## Q3                     2.4               1.8               1.6
```

Variance

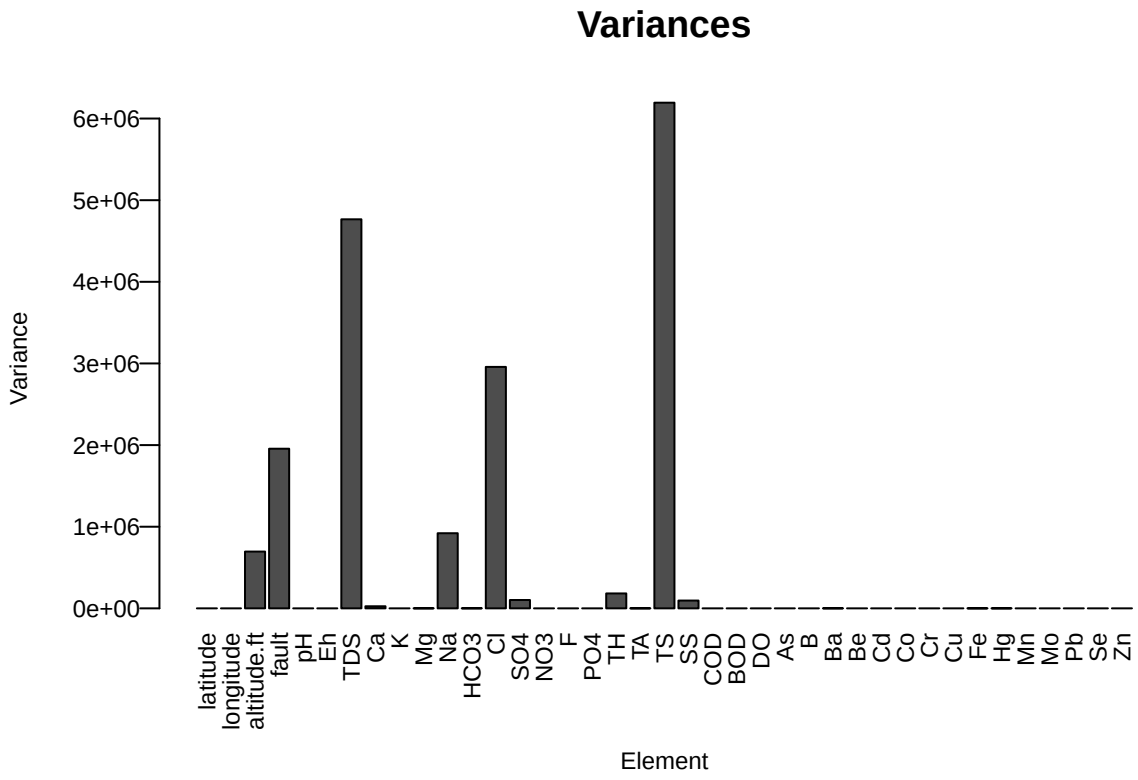

Figure 1: Barplot of Variances

Looking at the variances for each attribute, we can see there are 6 attributes that seem to vary a lot, and 4 that have a high enough variance worth looking into. (See Figure 1). Altitude.ft, fault, TDS, NA, CL, and TS have very high variation. CA, SO4, TH, and SS have lower, but still "high" variation. # Deeper Analysis

Looking at Geology

I now want to look at how the values of TDS, TS and Cl differ between the different rock formations. ### TDS

Being there are some elements with high variation, specifically TDS, I want to look and see if the "geology" attribute seems to have an effect on TDS. To do this, I will group the different "geology" values and see how they compare. (I realized I had some typos in my "geology" vector for o-well1, so I changed that). To look at the differences, I will make a boxplot using ggplot to see how different TDS is between the geology groups.

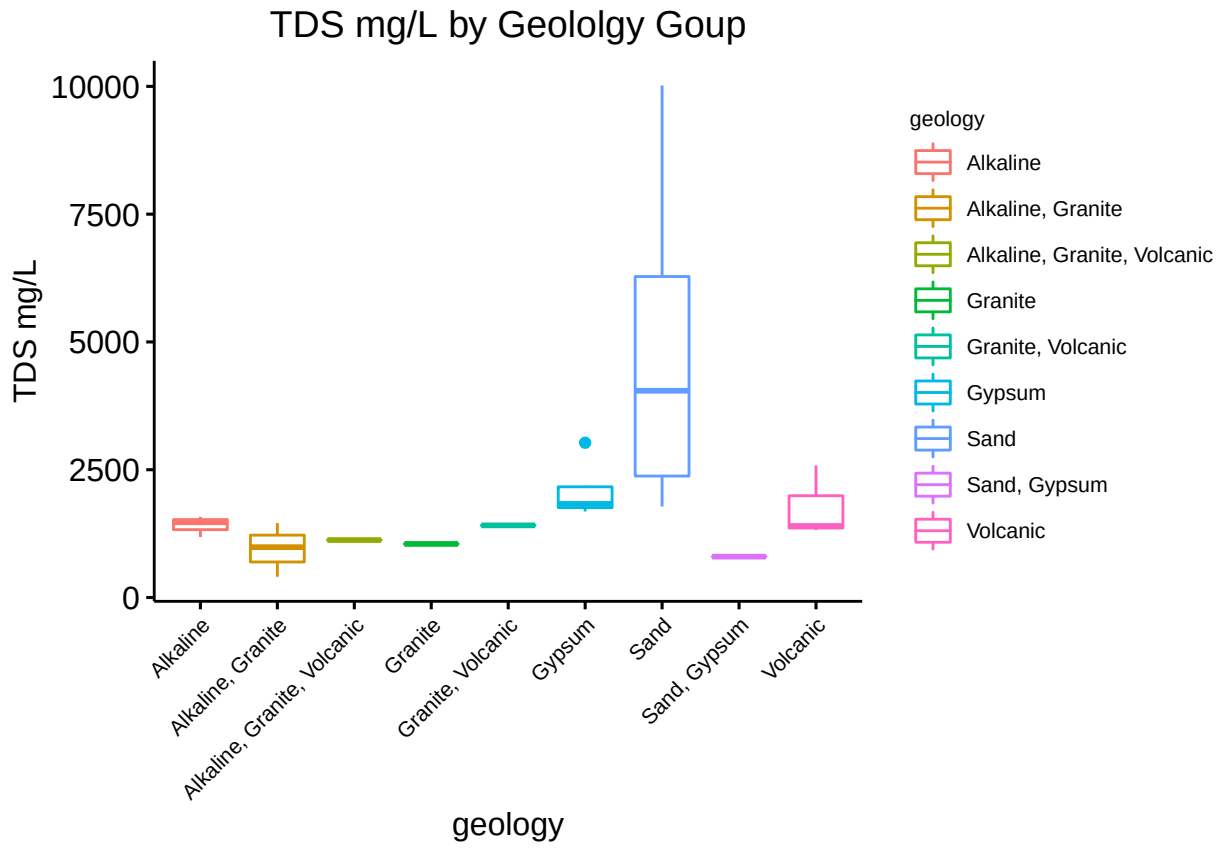

Figure 2: Geology Group Boxplot for TDS

Looking at this boxplot (Figure 2) we can see that wells that have a Sand rock formation have higher values of TDS than any other rock formation group. We can also see that wells with a Gypsum rock formation have an outlier. Wells that have a Sand rock formation have 50% of their data between approximately 2,500 mg/L and 6,250 mg/L.. It is also apparent that wells with a Sand rock formation have a fairly large spread of TDS values when compared to other rock formations.

TS

Being TS has the highest variance, I want to look and see if the "geology" attribute seems to have an effect on TS. I will use the same geology groups created when analyzing TDS, and see what the results show.

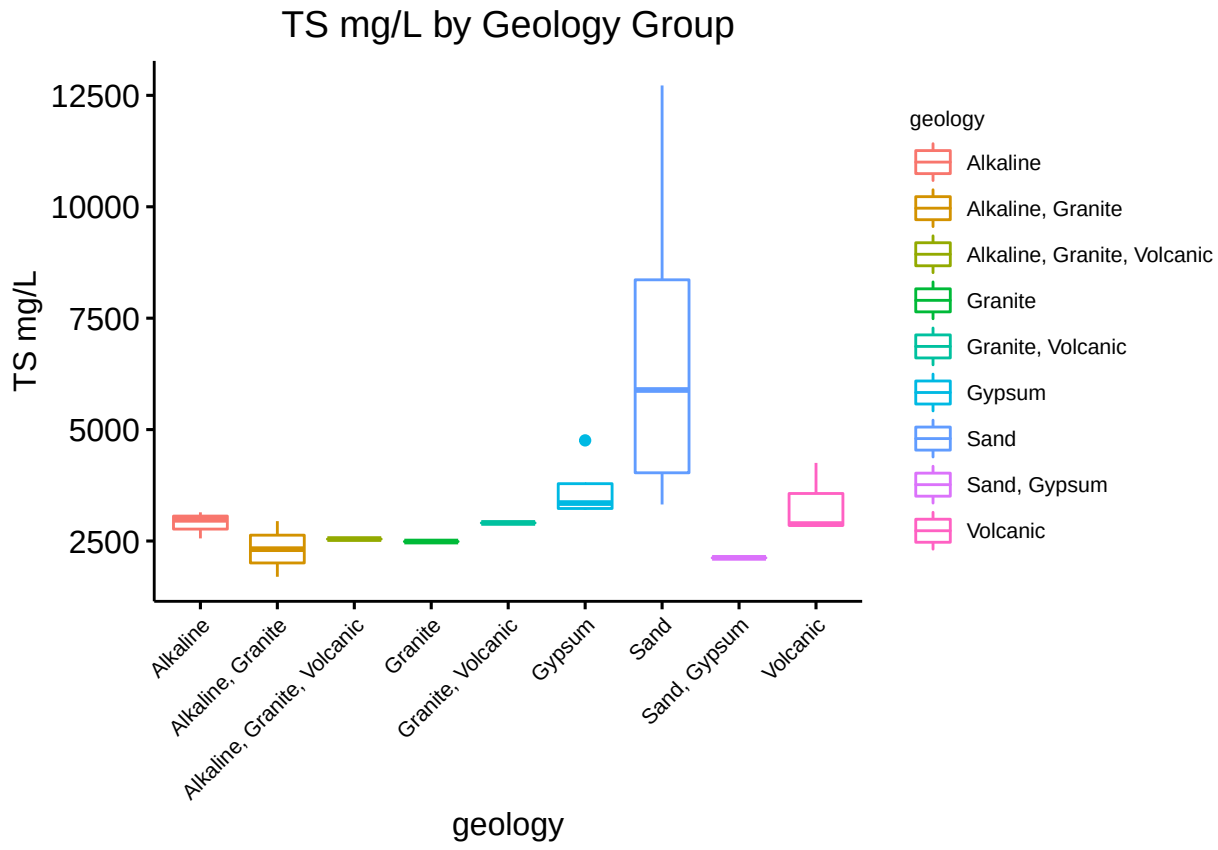

Figure 3: Geology Group Boxplot for TS

Interestingly, the two boxplots look identical (Figure 2 and Figure 3). Wells that have an Alkaline/Granite rock formation have a lower median in TS than TDS, but again, wells with Sand rock formations have a higher amount of TS when compared to the other formations. Although wells with a Sand rock formation see 50% of their TS values between 3,750 mg/L and approximately 7,800 mg/L.

Cl

I am now curious if Cl is higher with Sand rock formations than other formations. Again, using the same geology factors created above, I am going to make a boxplot to look at Cl levels between different rock formations.

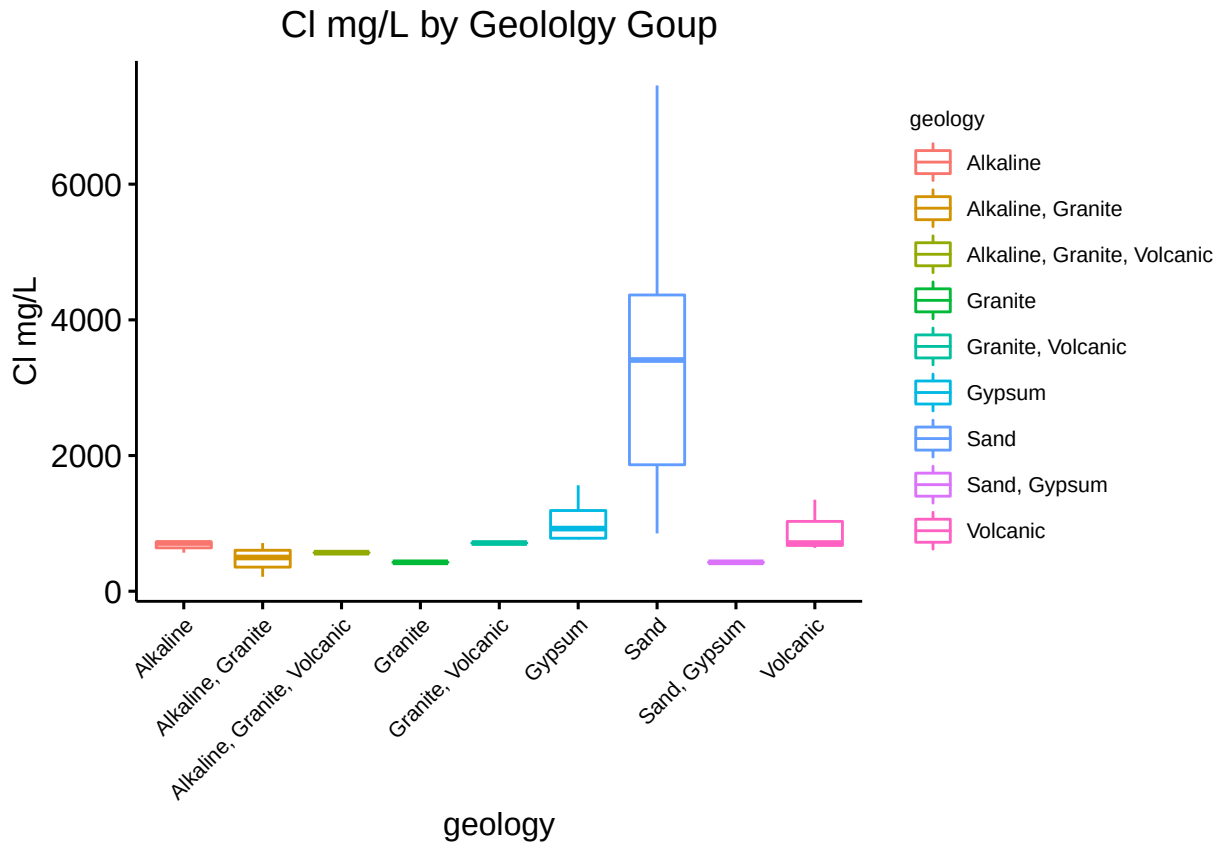

Figure 4: Geology Group Boxplot for Cl

Once again, wells with a Sand rock formation have far higher amounts of Cl than any other rock formations (See Figure 4). This is leading me to believe the high variation in some of the attributes might be related to the type of rock formation the well is. Now I am interesting in seeing if there are any differences in the 3 well groups for TDS, Ts, and Cl.

Difference Between Well Groups

I am now going to look at the three different well groups, and see if their mean values for TDS, TS, and Cl are different. To do this, I created a separate data frame adding a group column and grouped the wells by referencing the pdf. Group 1 is wells 1-17, group 2 is wells 18-22 and group 3 is only well 23. I then created the group a factor, and looked at the difference in TDS, Ts and Cl values.

Making Groups

Looking at TDS

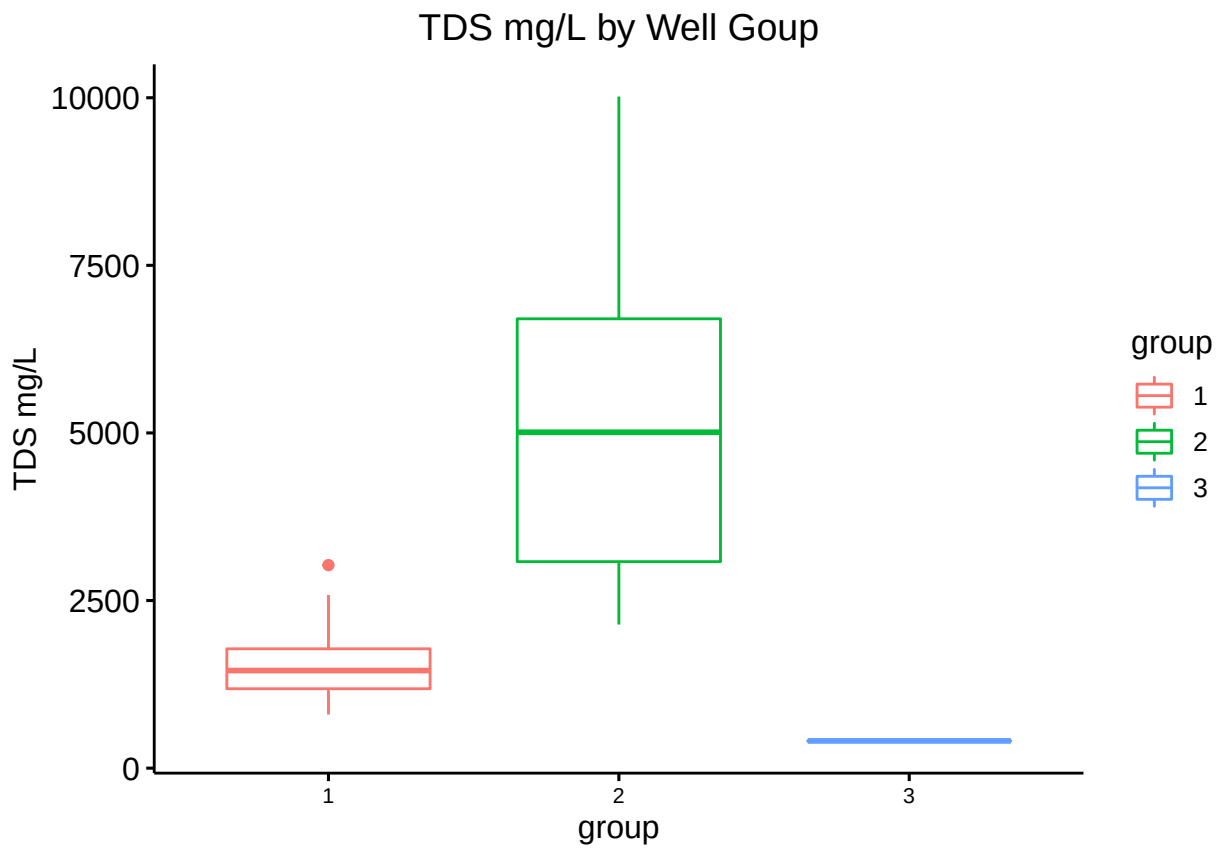

Figure 5: Well Group Boxplot for TDS

It is apparent that group 2 has 50% of its TDS values being greater than both group 1 and group 3. Group 3 is obviously one well, and it a bit lower than group 2 (See Figure 5). But there seems to be a pretty big difference in TDS values from group 1 and group 2. 50% of group 2's TDS values are between approximately 3,500 mg/L and 6,500 mg/L. While most of group 1's TDS values are below approximately 2,200 mg/L.

Looking at TS

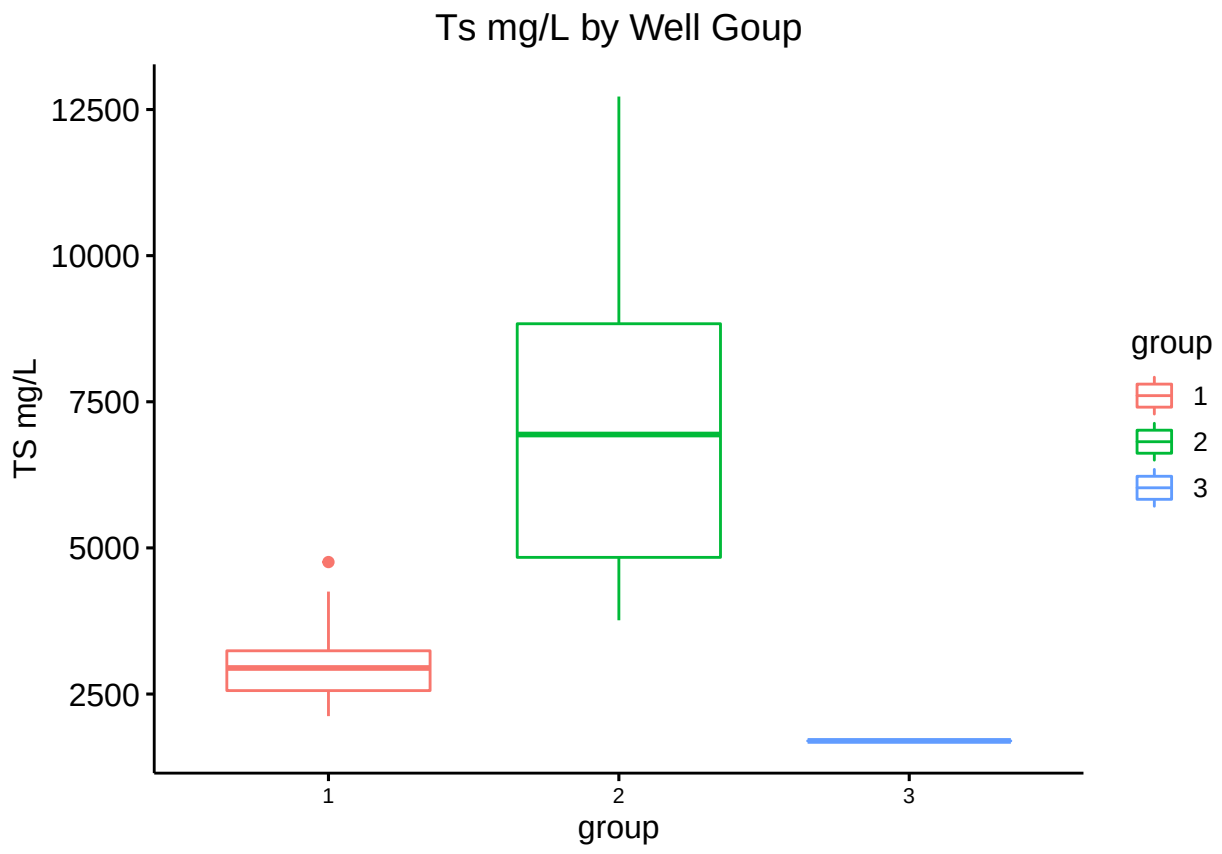

Figure 6: Well Group Boxplot for TS

Once again, we can see that group 2 has higher TS values than group 1 or group 3 (See Figure 6). 50% of the TS values for group 2 are between approximately 4,900 mg/L and 9,000 mg/L. This is far higher than group 1 whose TS values are between approximately 2,500 and 3,200 mg/L.

Looking at Cl

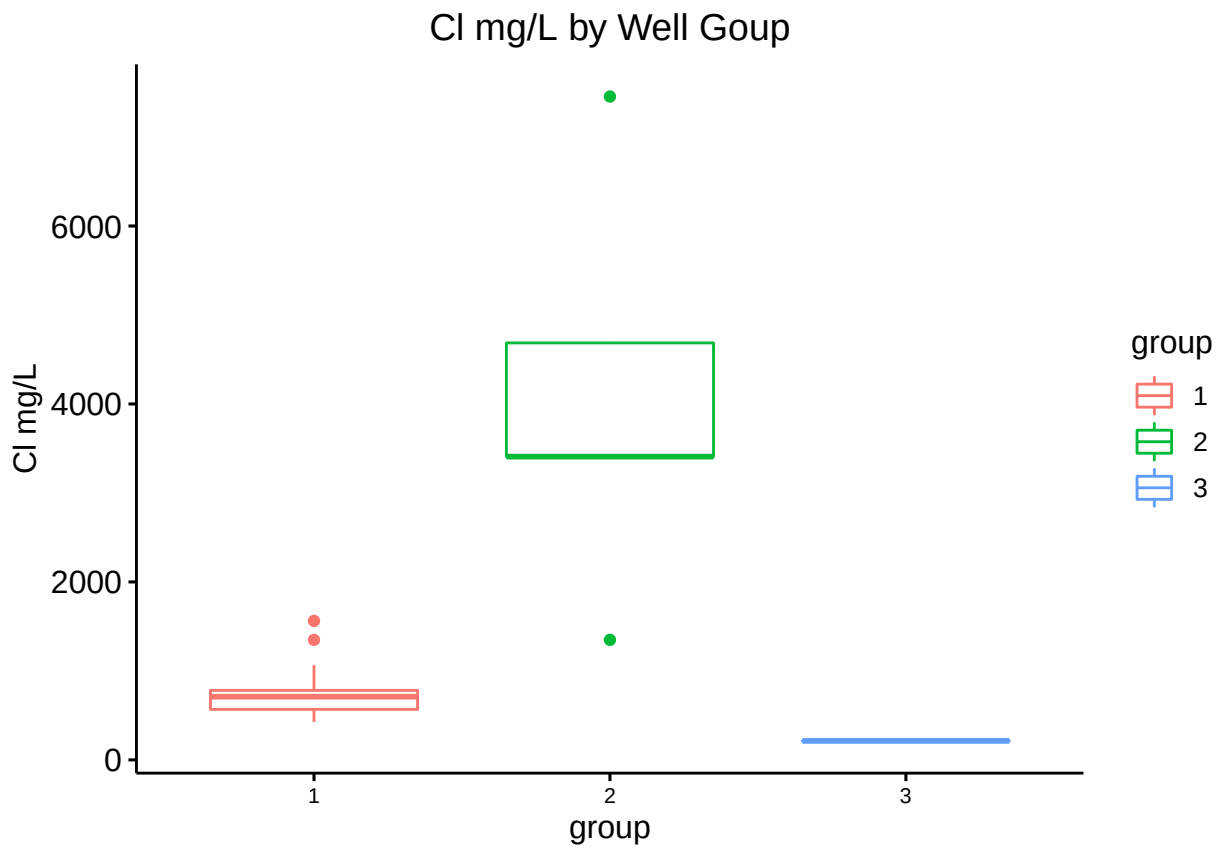

Figure 7: Well Group Boxplot for Cl

So this is really interesting to me. 50% of group 2's values are between two outliers, one high outlier and one low outlier. But still, group 2 has higher Cl values then either group 1 or group 3 (See Figure 7). Looking at the data frame, I guess this isn't too surprising, ass all of group 2's wells have a sand rock formation. So it seems that having a Sand rock formation has an effect on the TDS, TS and Cl levels in a well.

Looking at Altitude

Now I wanted to look and see how different altitude is between the different rock formations.

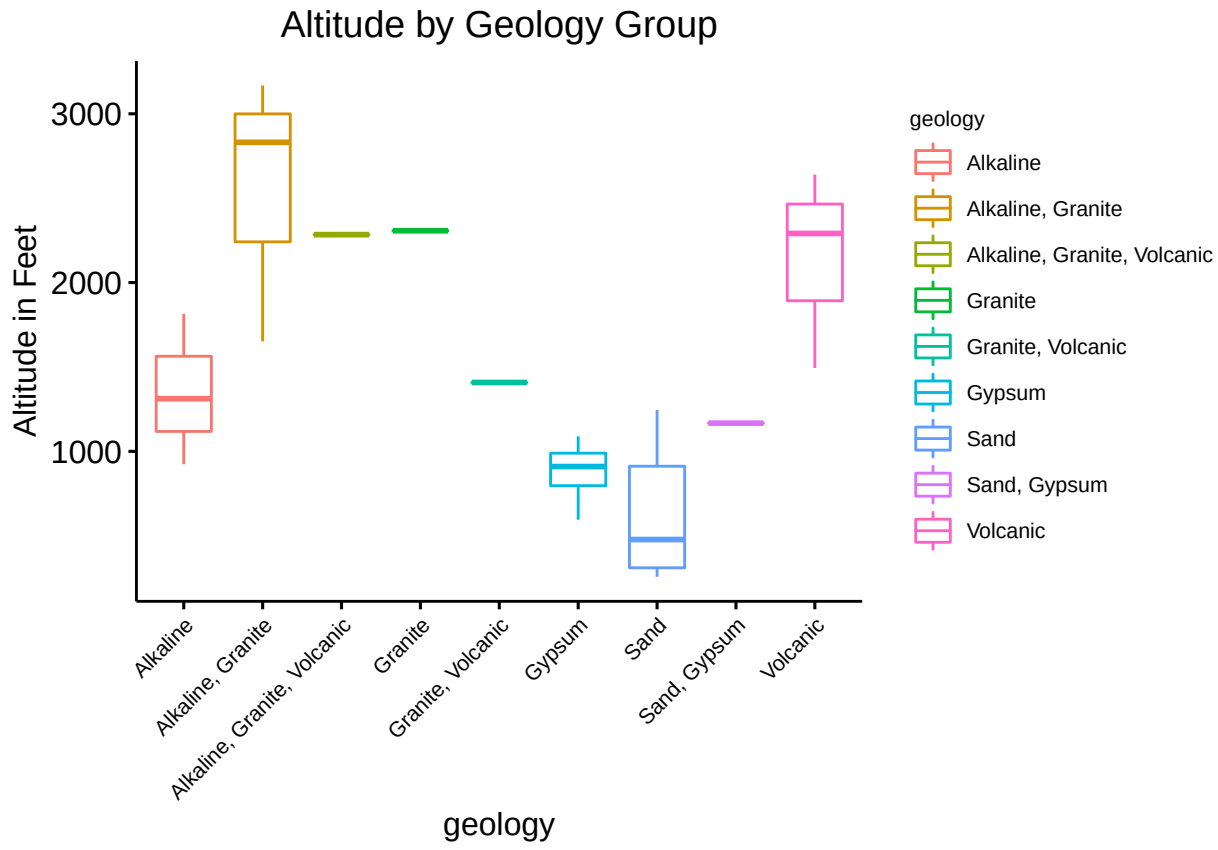

Figure 8: Well Geology Group Boxplot for Altitude

I found this to be very interesting. Wells that are a sand rock formation tend to have lower altitudes than the other rock formations (See Figure 8). So this might mean that TDS, TS and Cl might not be higher in Sand rock formations because of the sand, but maybe it's the altitude, and maybe the altitude has something to do with sand formations.

Checking Means Between Groups

I now want to implement a test to determine if there is any significant difference between the TDS, TS, and CL values between the different rock formations.

Using Kruskal-Wallis to Test Different Mean-Ranks of TDS, TS, and Cl Between Rock Formations

I used the Kruskal-Wallis test, because I do not know what distribution the data falls under, and I am not so familiar with how to test for distributions. I did some research and decided to play it safe for now and read about the tests as time permits. Kruskal-Wallis allows me to do an experiment that I would normally do using a one-way ANOVA, when the assumptions of the one-way ANOVA fail. http://www.biostathandbook.com/kruskalwallis.html

TDS

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TDS by geology
## Kruskal-Wallis chi-squared = 17.313, df = 8, p-value = 0.02701
```

We can see that because the p-value of the Kruskal-Wallis test is low, then we can assume there is a difference in TDS mean ranks between the different rock formations. However, we do not yet know which groups are different, which will be investigated later in the notebook.

TS

```
##
##  Kruskal-Wallis rank sum test
##
## data:  TS by geology
## Kruskal-Wallis chi-squared = 17.486, df = 8, p-value = 0.02543
```

We can see that because the p-value of the Kruskal-Wallis test is low, then we can assume there is a difference in TS mean ranks between the different rock formations. However, we do not yet know which groups are different, which will be investigated later in the notebook.

Cl

```
##
##  Kruskal-Wallis rank sum test
##
## data:  Cl by geology
## Kruskal-Wallis chi-squared = 17.977, df = 8, p-value = 0.0214
```

We can see that because the p-value of the Kruskal-Wallis test is low, then we can assume there is a difference in Cl mean ranks between the different rock formations. However, we do not yet know which groups are different, which will be investigated later in the notebook.

I find it very interesting that indeed, there does seem to be a difference in these 3 attributes that have high variances between the different rock formations. This could mean that the rock formations do indeed have an impact on the levels of TDS, TS and Cl in the wells.

Looking at Which Groups are Different for TDS, TS and Cl Between Different Rock Formations

To determine which groups are different when it comes to TDS, TS, and Cl, I used the conover test, which is the non parametric form of Levene's test for equality of variance and will provide some insight on which groups are statistically different. https://en.wikipedia.org/wiki/Squared_ranks_test#:~:text=In%20statistics%2C%20the%20Conover%20squared,test%20for%20equality%20of%20variance.&text=The%20s

Looking at TDS

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.3134, df = 8, p-value = 0.03
##
##
##                            Comparison of x by group
##                              (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |   Alkaline   Alkaline   Alkaline    Granite   Granite,     Gypsum
## ---------+-----------------------------------------------------------------
## Alkaline |   1.560529
##          |     0.1492
##          |
## Alkaline |   1.029897  -0.073564
##          |     0.2404     0.4712
##          |
##  Granite |   1.250589   0.147128   0.180194
##          |     0.1812     0.4686     0.4689
##          |
## Granite, |   0.147128  -0.956333  -0.720777  -0.900972
##          |     0.4552     0.2557     0.3219     0.2650
##          |
##   Gypsum |  -2.029736  -3.698012  -2.450247  -2.678177  -1.538527
##          |     0.0928    0.0143*     0.0459     0.0360     0.1385
##          |
##     Sand |  -3.543824  -5.345768  -3.420983  -3.656913  -2.477264  -1.480448
##          |    0.0117*    0.0019*    0.0124*    0.0117*     0.0479     0.1448
##          |
## Sand, Gy |   1.691973   0.588512   0.540583   0.360388   1.261361   3.134036
##          |     0.1353     0.3636     0.3707     0.4072     0.1864    0.0188*
##          |
## Volcanic |  -0.416141  -1.976671  -1.324153  -1.544845  -0.441384   1.584862
##          |     0.3969     0.0876     0.1771     0.1447     0.3994     0.1522
## Col Mean-|
## Row Mean |       Sand   Sand, Gy
## ---------+----------------------
## Sand, Gy |   4.128773
##          |    0.0092*
##          |
## Volcanic |   3.063305  -1.986230
##          |    0.0190*     0.0927
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

We can see after running the conover test, that there are specific rock formation combinations that stick out when analyzing TDS measurements. Those combinations are;

Sand and Alkaline
Sand and Alkaline, Granite
Sand and Granite, Alkaline, Volcanic
Sand and Granite Sand, Gypsum and Gypsum Sand, Gypsum and Alkaline Volcanic and Alkaline

So we can see that Sand seems to be different than most of the groups, indicating that sand might indeed have something to do with increased TDS levels.

Looking at TS

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.4855, df = 8, p-value = 0.03
##
##
##                            Comparison of x by group
##                              (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |   Alkaline   Alkaline   Alkaline    Granite   Granite,     Gypsum
## ---------+------------------------------------------------------------------
## Alkaline |   1.589996
##          |     0.1420
##          |
## Alkaline |   1.049344  -0.074953
##          |     0.2338     0.4707
##          |
##  Granite |   1.274203   0.149906   0.183597
##          |     0.1748     0.4675     0.4674
##          |
## Granite, |   0.149906  -0.974391  -0.734388  -0.917985
##          |     0.4541     0.2494     0.3166     0.2590
##          |
##   Gypsum |  -1.983074  -3.682852  -2.438455  -2.670689  -1.509520
##          |     0.0866    0.0111*     0.0469     0.0366     0.1381
##          |
##     Sand |  -3.671940  -5.507910  -3.525645  -3.766030  -2.564105  -1.676003
##          |    0.0090*    0.0014*    0.0101*    0.0125*     0.0405     0.1304
##          |
## Sand, Gy |   1.723922   0.599625   0.550791   0.367194   1.285179   3.135157
##          |     0.1281     0.3589     0.3665     0.4044     0.1797    0.0164*
##          |
## Volcanic |  -0.423999  -2.013995  -1.349157  -1.574016  -0.449719   1.529800
##          |     0.3937     0.0881     0.1703     0.1378     0.3959     0.1405
## Col Mean-|
## Row Mean |       Sand   Sand, Gy
## ---------+----------------------
## Sand, Gy |   4.246800
##          |    0.0073*
##          |
## Volcanic |   3.182348  -2.023735
##          |    0.0171*     0.0938
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

We can see after running the conover test, that there are specific rock formation combinations that stick out when analyzing TDS measurements. Those combinations are;

Gypsum and Alkaline, Granite
Sand and Alkaline
Sand and Alkaline, Granite
Sand and Granite, Alkaline, Volcanic
Sand and Granite Sand, Gypsum and Gypsum
Sand, Gypsum and Alkaline
Volcanic and Alkaline

We can see that these results are very similar to that of TDS, and it is appearing that Sand may be affecting the amount of TDS and TS in the wells that are on sand rock formations.

Looking at Cl

```
##   Kruskal-Wallis rank sum test
##
## data: x and group
## Kruskal-Wallis chi-squared = 17.9765, df = 8, p-value = 0.02
##
##
##                            Comparison of x by group
##                              (Benjamini-Hochberg)
## Col Mean-|
## Row Mean |   Alkaline   Alkaline   Alkaline    Granite   Granite,     Gypsum
## ---------+-----------------------------------------------------------------
## Alkaline |   1.408750
##          |     0.1549
##          |
## Alkaline |   0.836754  -0.159381
##          |     0.2885     0.4636
##          |
##  Granite |   1.553973   0.557836   0.585606
##          |     0.1425     0.3295     0.3405
##          |
## Granite, |  -0.119536  -1.115673  -0.780808  -1.366415
##          |     0.4662     0.2040     0.2879     0.1450
##          |
##   Gypsum |  -2.349386  -3.855404  -2.469133  -3.209873  -1.481480
##          |     0.0510    0.0063*     0.0442    0.0126*     0.1446
##          |
##     Sand |  -4.196846  -5.823530  -3.642009  -4.408748  -2.619690  -1.817582
##          |    0.0040*    0.0008*    0.0080*    0.0036*     0.0363     0.1019
##          |
## Sand, Gy |   1.553973   0.557836   0.585606   0.000000   1.366415   3.209873
##          |     0.1509     0.3401     0.3522     0.5000     0.1513    0.0142*
##          |
## Volcanic |  -0.788900  -2.197650  -1.394591  -2.111809  -0.438300   1.506017
##          |     0.2956     0.0627     0.1513     0.0638     0.3643     0.1462
## Col Mean-|
## Row Mean |       Sand   Sand, Gy
## ---------+----------------------
## Sand, Gy |   4.408748
##          |    0.0054*
##          |
## Volcanic |   3.285902  -2.111809
```

```
##           |    0.0139*     0.0683
##
## alpha = 0.05
## Reject Ho if p <= alpha/2
```

We can see after running the conover test, that there are specific rock formation combinations that stick out when analyzing TDS measurements. Those combinations are;

Gypsum and Alkaline Granite
Gypsum and Granite
Sand and Alkaline
Sand and Alkaline, Granite
Sand and Granite, Alkaline, Volcanic
Sand and Granite
Sand, Gypsum and Gypsum
Sand, Gypsum and Alkaline
Volcanic and Alkaline

Once again the results are similar for Cl. Sand seems to be different than most other groups, indicating TDS, TS and Cl levels may be affected by Sand.

Correlations

I want to now look at the correlation matrix between TDS, TS, Cl and altitude, longitude, latitude and fault. I am curious how correlated TDS, TS and Cl are to these attributes, aside from all of the other elements and metals.

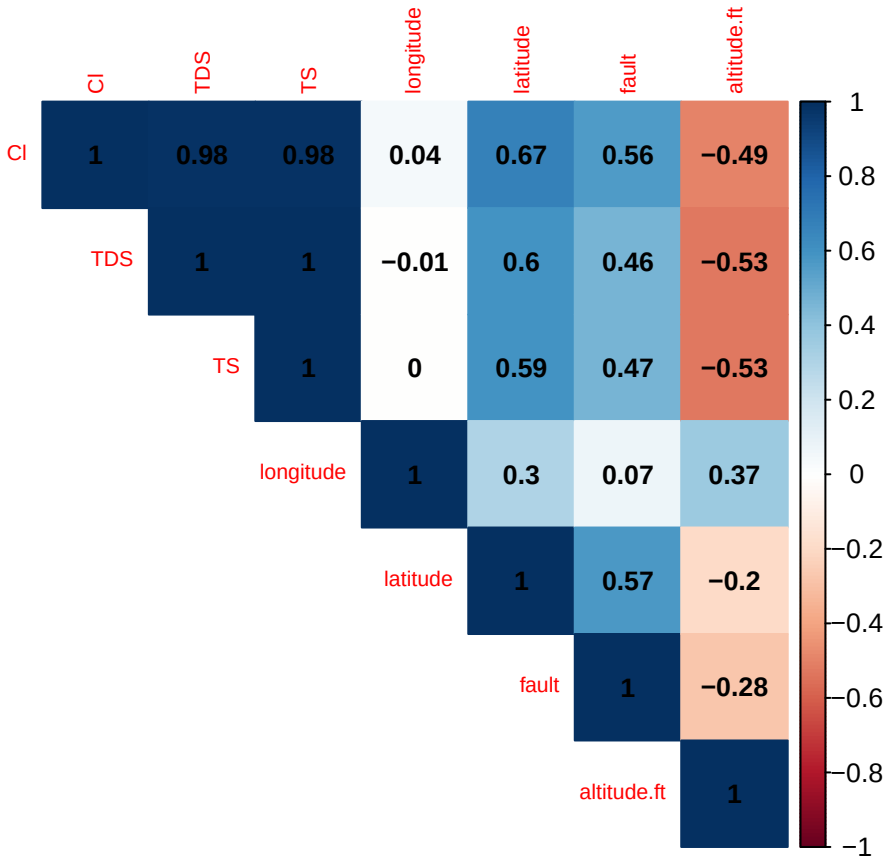

Figure 9: Correlation Matrix Plot for Numeric Attributes

Upon inspection of Figure 9, we can see that the previous analyses make some sense. TDS and TL have a perfect positive correlation, meaning if one of them goes up, the other goes up the same amount. Cl and TDS have a near perfect positive correlation, and so does Cl and TS. It is interesting that all of these 3 elements have some of the highest variation in their amounts, yet they are all nearly perfectly correlated. Being all but one of the wells that have a sand rock formation are in the same group which showed to have higher amounts of TDS, TS and Cl, it is not so surprising that TDS, TS and Cl have a fairly high correlation (.59-.67) between latitude. What is interesting is there is a fairly high correlation between TDS, TS and Cl and fault. This would indicate that as the distance from a fault increases, so does the concentration of these elements. What I find really interesting is that TDS, TS and Cl have a fairly strong negative correlation between altitude. So this does add to the previous though that maybe altitude does have something to do with the concentration of these elements. As altitude goes up, the levels of TDS, TS and Cl go down, meaning lower elevation wells will see higher concentrations. We did wee earlier that most of the wells with Sand rock formations had higher amounts of these elements, and they also were at lower elevations. So I'm, still on the fence about whether it is the sand or elevation, or maybe a combination of both. I will analyze further as we go on.

Adding Distance to Red Sea

To add the distance from each well to the Red Sea, I used the provided .kml file and used the "measure" tool to measure the distance in meters from each well to the Red Sea. I then recorded these measurements into a vector and added the vecotr to the 'big.df' data frame as a column named "seaDist.m"

Adding Wilcox Data

Visualizing correlation between Distance to Red Sea and Ph

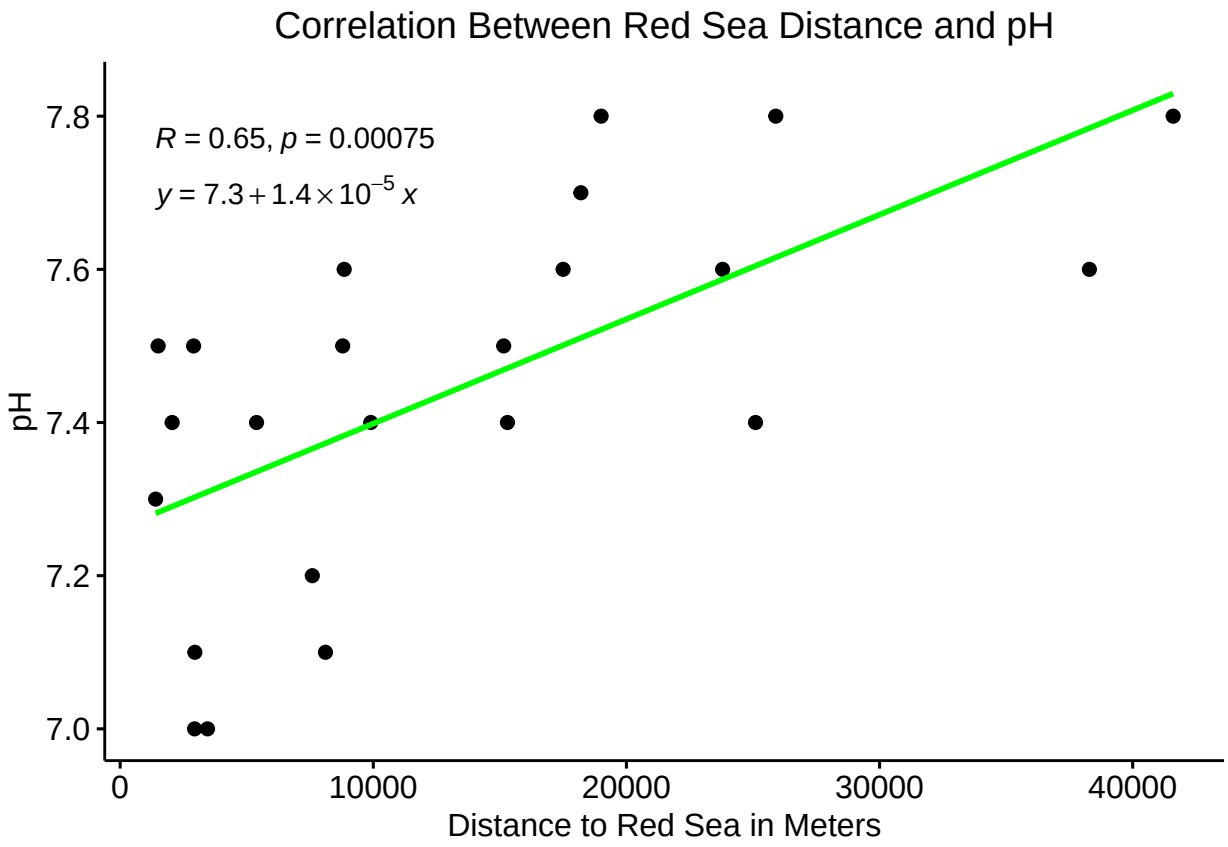

Figure 10: Correlation Between Red Sea Distance and pH

As we can see from the above graphic, the correlation between the distance to the red sea and the pH level for the wells, there is a moderate positive correlation (R = 0.65). I added the regression line to the graphic to show that the correlation is positive. This would indicate that as the distance to the Red Sea increases, the pH level of the well will generally increase.

PH, Altitude and Fault Distance Correlations

Correlation Matrix

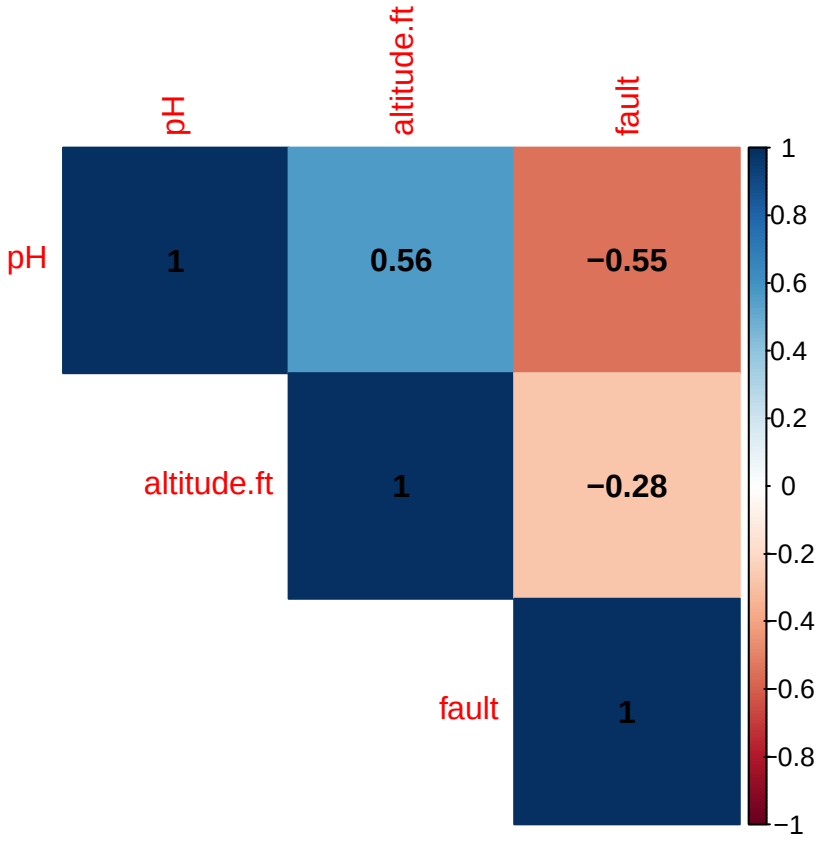

Figure 11: Correlation Matrix for pH, Altitude and Fault

We can see from the correlation matrix that pH and the distance from a fault have a moderate negative correlation. This would indicate that as the distance to a fault increases, the pH level tends to decrease, and as distance to a fault decreases, pH tends to increase. The same is true if pH increases, distance to a fault tends to decrease and as pH decreases, distance to a fault tends to increase. We can also see that pH and altitude have a moderate positive correlation. This would indicate that as altitude increases, pH tends to increase, and as pH increases, so altitude tends to increase. Also apparent is a weak negative correlation between altitude and fault. This indicates that as altitude increases, distance to a fault tends to decrease and as altitude decreases, distance to a fault tends to increase. The same is true if distance to a fault increases, altitude tends to decreases and as distance to a fault decreases, altitude tends to increase.

Boxplot for Altitude and pH

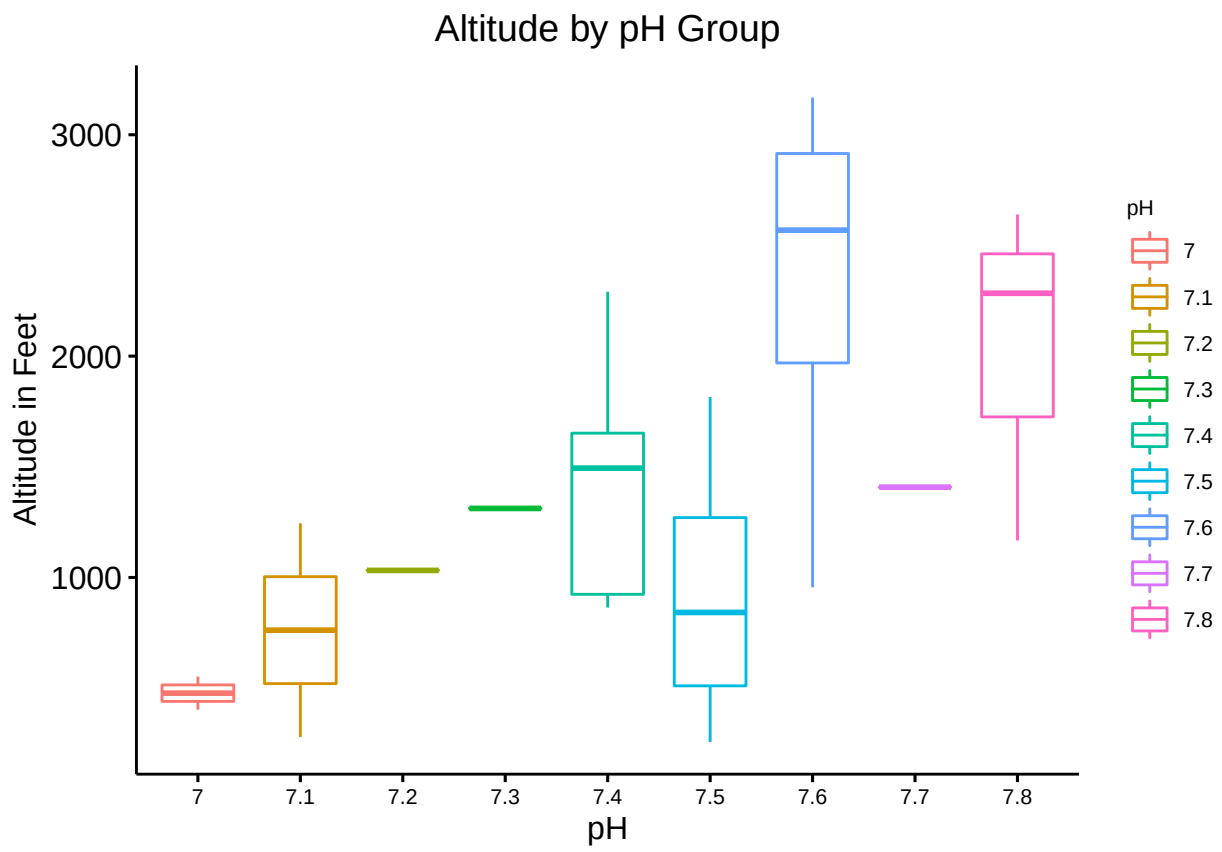

Figure 12: Boxplot of Altitude by pH Group

We can see form the boxplot that the highest pH from the data, 7.8, sees 50% of its occurrence at an altitude between approximately 1700 feet and 2500 feet. We can also see that a pH of 7, the lowest pH in the data, sees the majority of its occurrences at an altitude below 500 feet, which is lower than the majority of occurrences of any other pH value. Overall, we can see the moderate positive correlation between pH and altitude, as we can see that as pH increases, altitude generally increases.

Latitude, Geology and pH

Latitude and Geology Boxplot

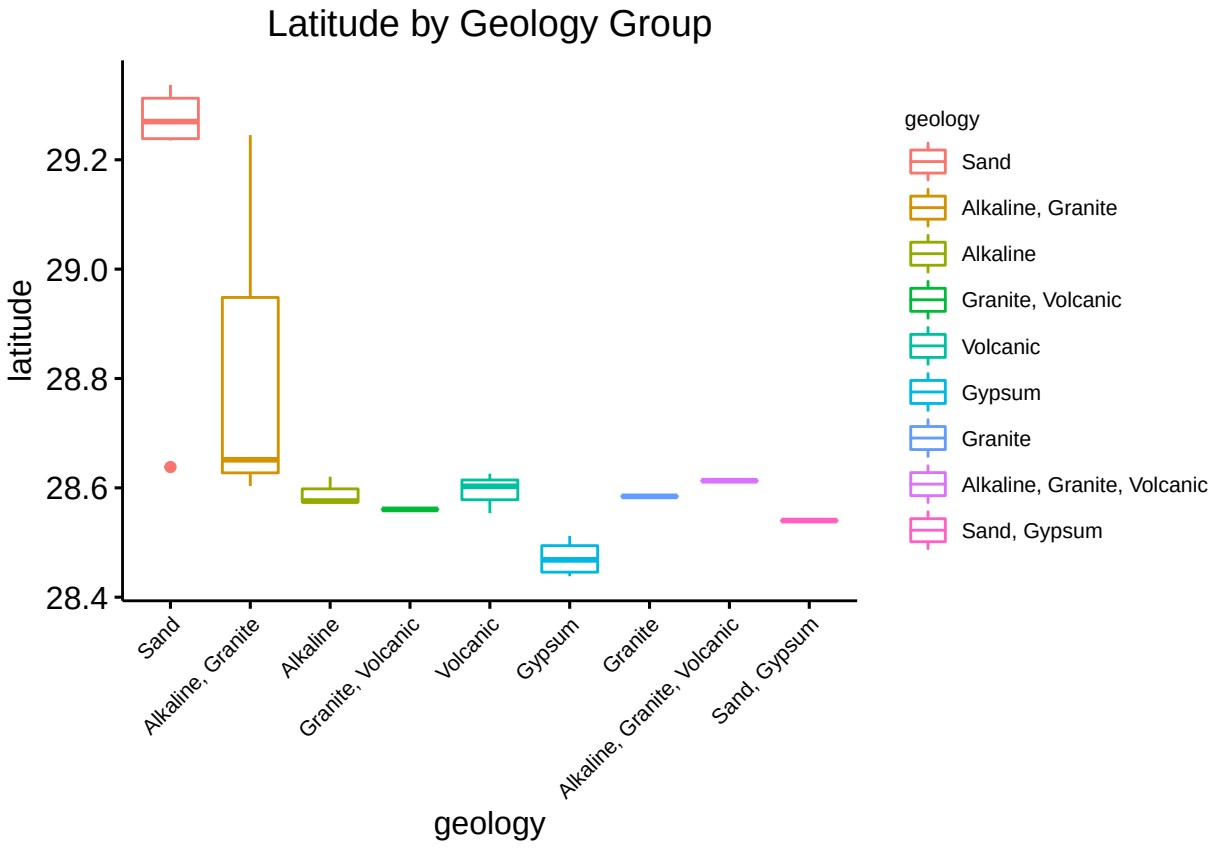

Figure 13: Boxplot of Latitude by Geology Group

We can see looking at the boxplot that 50% of the latitude values for wells with a sand rock formation are higher than the other rock formations. We can also see that wells with an alkaline, granite rock formation have a bigger spread of latitude values when compared to other rock formations. Gypsum has lower latitude values when compared to any other rock formation. This probably has to do with where the rock formations are located saudi Arabia.

Latitude and Geology Density Plots

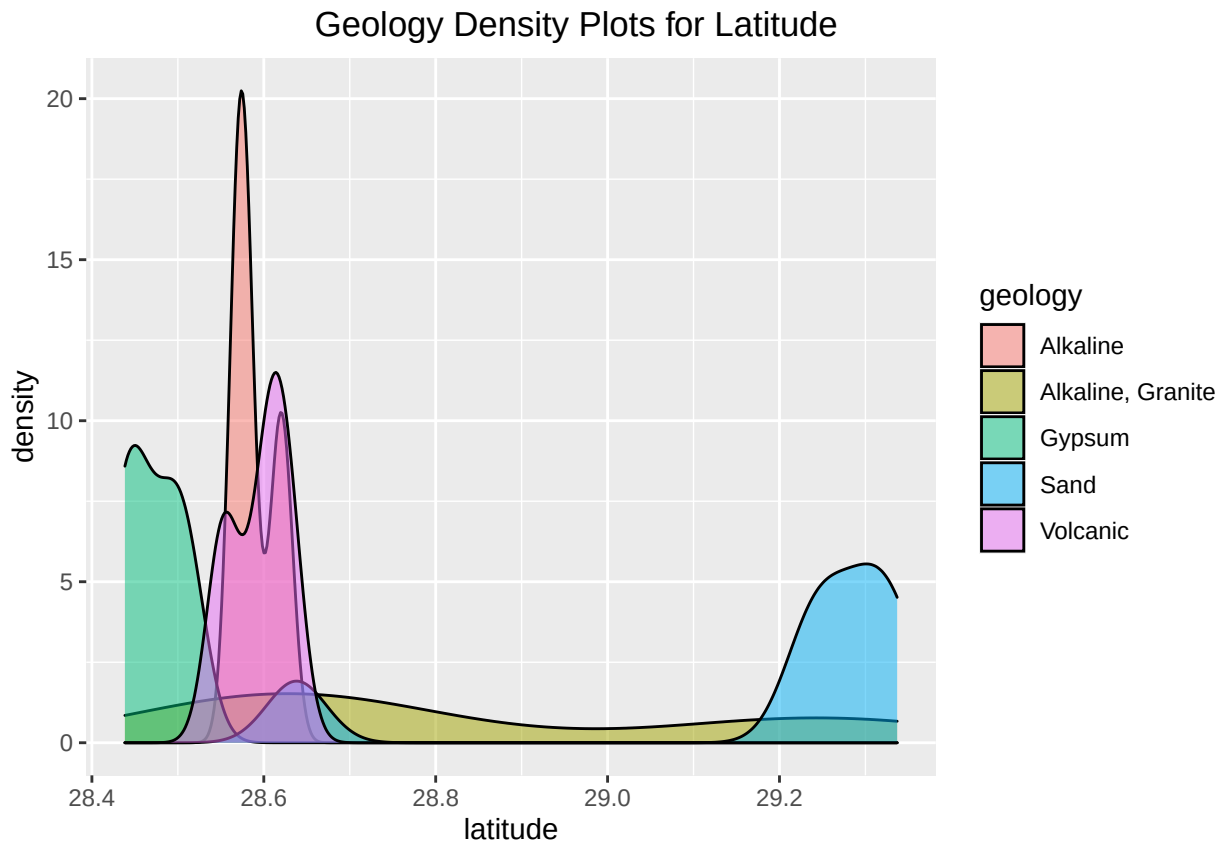

Figure 14: Density Plots of Latitude by Geology Group

We can see looking at the distribution plots for latitude by geology group that the distributions are quite different. Again, sand seems to have more values at a higher latitude value than the other rock formations, and we can see the lower latitude values for gypsum as well. This also shows the spread in latitude values for the alkaline, granite rock formation indicated in the previous boxplot. NOTE: only five of the rock formations are plotted, as ggplot removes samples with lower than two observations. .

#Latitude and pH Boxplot

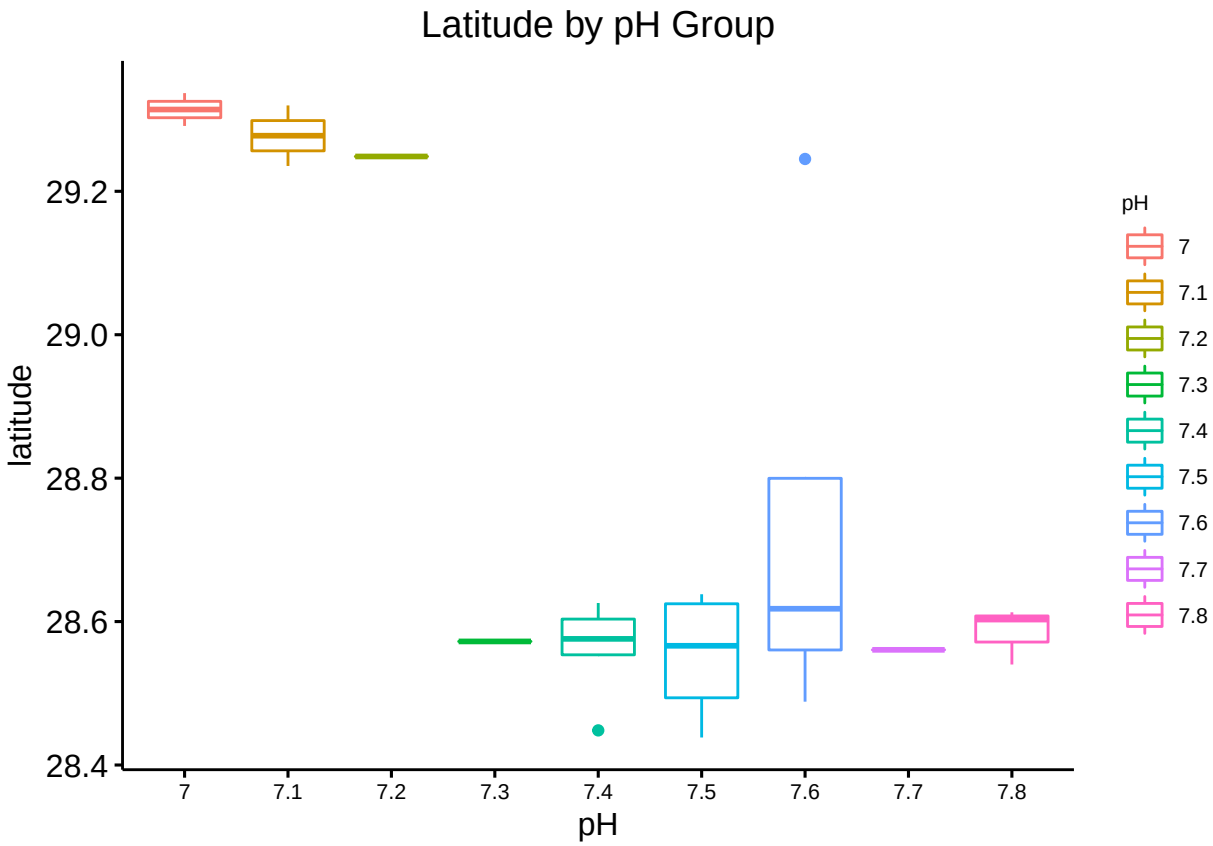

Figure 15: Boxplot of Latitude by pH Group

We can see from the above boxplot that there is a fairly large difference in latitude values with wells that have a pH of 7 to 7.2 and 7.3 to 7.8. There is some differences in latitude values for wells with a pH of 7.3 to 7.8, particularly with wells that have a pH of 7.5 and 7.6. Wells with a pH of 7.6 have a larger spread of latitude values then wells with any other pH value. It does appear that wells with a pH value below 7.3 tend to have a higher latitude value.

Latitude and pH Density Plots

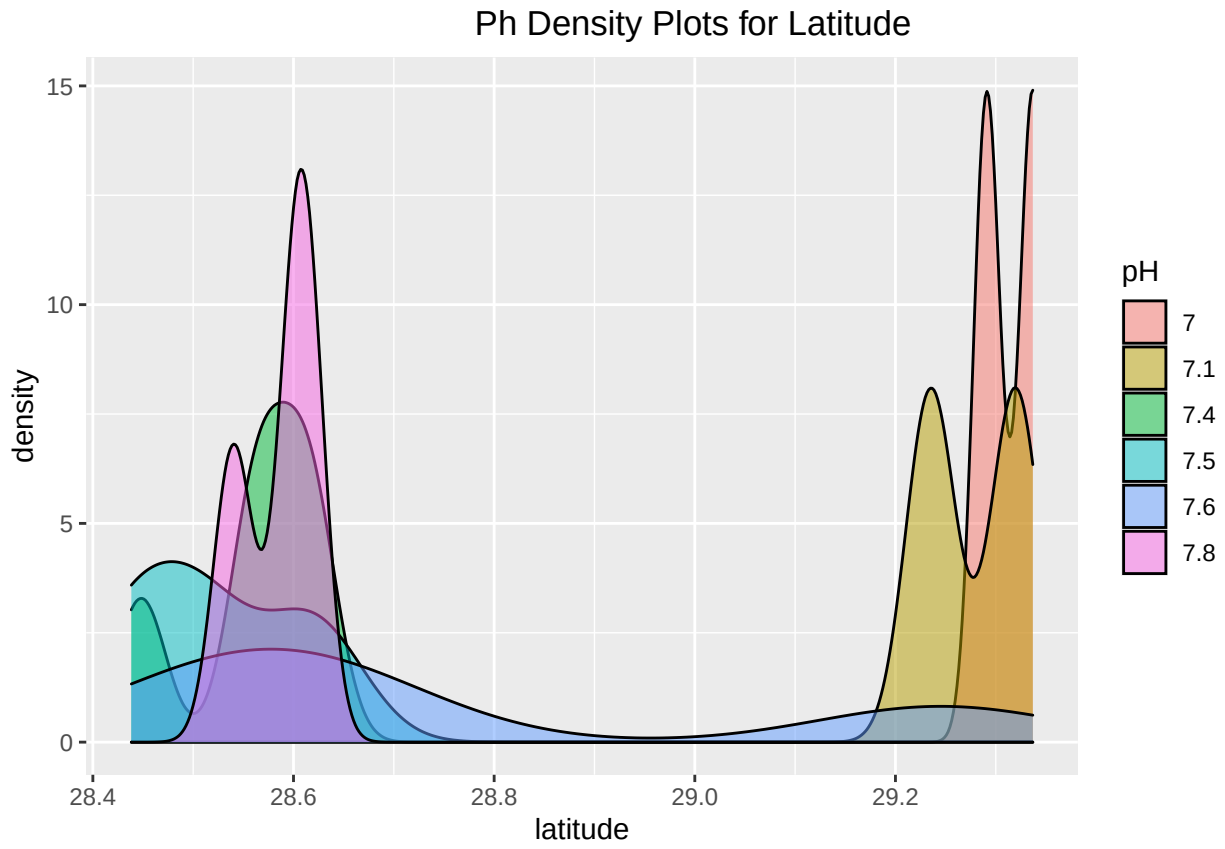

Figure 16: Density Plots of Latitude by pH Group

Looking at the density plot above, we can see that the distributions of latitude values between the different pH groups is quite different. Although, we can see two of the pH groups, 7, 7.1 have a higher distribution of latitude above 29.1, which was also seen in the boxplot. Also evident is the 4 pH values, 7.4, 7.5, 7.6, 7.8, having higher distributions of latitude below 28.7, with pH of 7.6 being spread, which was also seen in the boxplot.

Altitude and ph

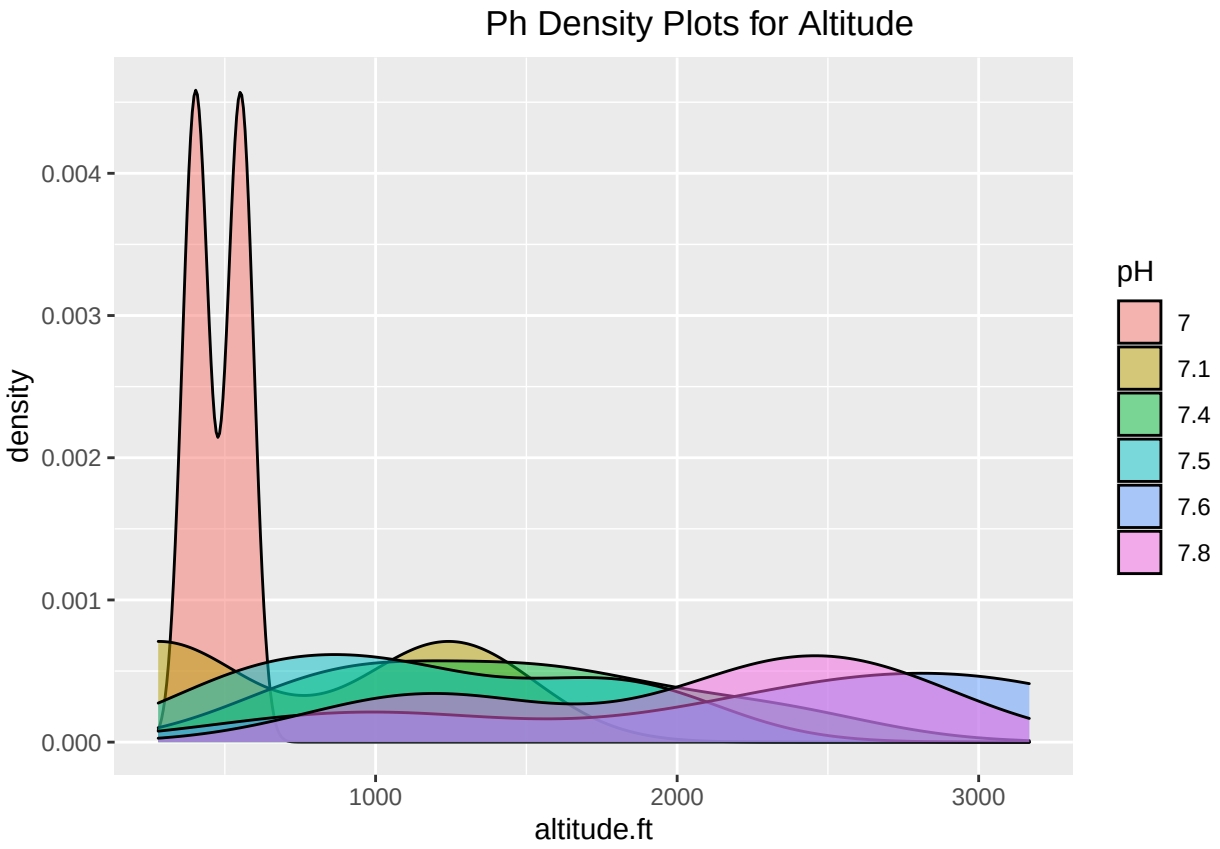

Figure 17: pH Density Plots for Altitude

Looking at the above density plot, we can see that pH 7 has a vastly different distribution than the other pH groups. We can see that a pH of 7 has a distribution of mostly below 1000 ft in altitude. The other pH groups have a fairly large spread in their data. This was indicated in the boxplot done previously in this report.