Math 189: Midterm Project 2 Solution

Introduction

In this second midterm project for Math 189 the Romano-British Pottery dataset is analyzed. We wish to know whether there is a significant difference among the 5 group means for the 9 variables in the dataset. This will be examined using techniques learned from the first 12 lectures of the course.

The dataset was provided in the course materials (https://github.com/tuckermcelroy/ma189), and contains measurements on pottery shards that were collected from five sites in the British Isles. These sites correspond to five values of the *Kiln* variable:

1. G: Gloucester
2. L: Llanedeyrn
3. C: Caldicot
4. I: Isle Thorns
5. A: Ashley Rails

The dataset contains 48 observations on 9 chemical variables:

1. Al2O3: aluminium trioxide
2. Fe2O3: iron trioxide
3. MgO: magnesium oxide
4. CaO: calcium oxide
5. Na2O: natrium oxide
6. K2O: kalium oxide
7. TiO2: titanium oxide
8. MnO: mangan oxide
9. BaO: barium oxide

Each of the nine variables might differ across the sites, and we will investigate whether these differences are significant.

Analysis

A First Peek

We begin by reading in the data, and taking a look. There are 9 variables, besides an index, an identifier, and the Kiln variable.

```
pottery <- read.csv("RBPottery.csv")
colnames(pottery) <- c("No", "ID", "Kiln", "Al", "Fe", "Mg", "Ca", "Na", "K2O", "TiO2", "MnO", "BaO")
head(pottery)
```

```
##   No  ID Kiln   Al   Fe   Mg   Ca   Na  K2O TiO2   MnO   BaO
## 1  1 GA1    1 18.8 9.52 2.00 0.79 0.40 3.20 1.01 0.077 0.015
## 2  2 GA2    1 16.9 7.33 1.65 0.84 0.40 3.05 0.99 0.067 0.018
## 3  3 GA3    1 18.2 7.64 1.82 0.77 0.40 3.07 0.98 0.087 0.014
## 4  4 GA4    1 17.4 7.48 1.71 1.01 0.40 3.16 0.03 0.084 0.017
## 5  5 GA5    1 16.9 7.29 1.56 0.76 0.40 3.05 1.00 0.063 0.019
## 6  6 GB1    1 17.8 7.24 1.83 0.92 0.43 3.12 0.93 0.061 0.019
```

We separate the data into five sub datasets according to the Kiln sites.

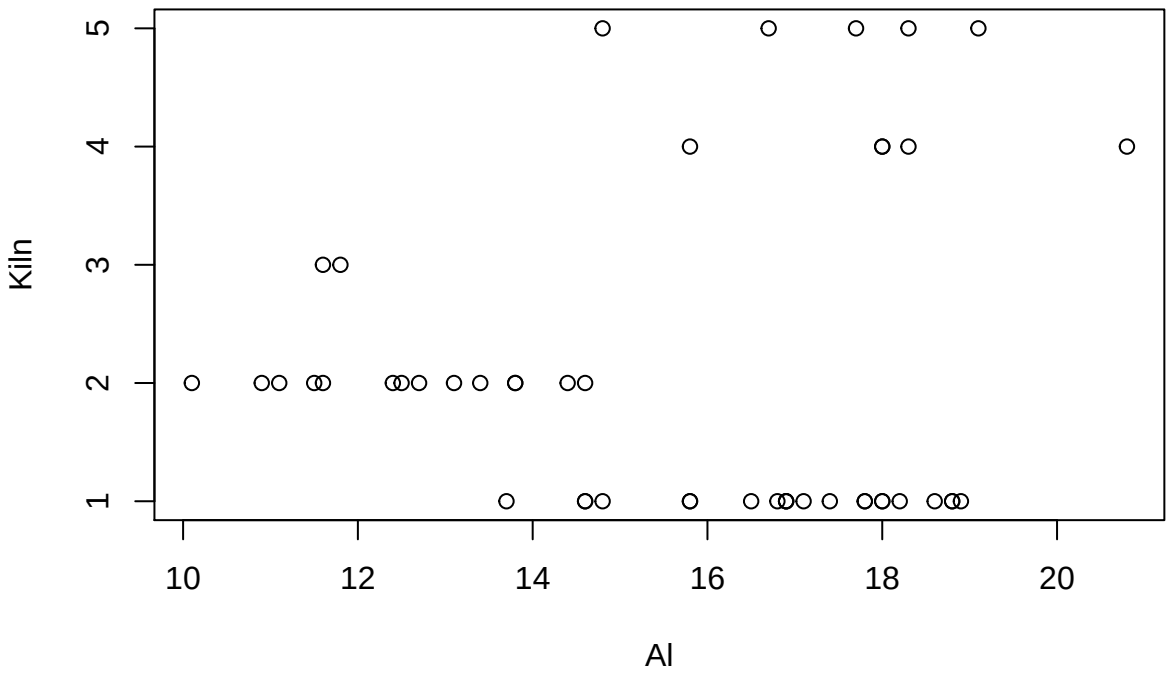
```
pot_glou <- pottery[pottery$Kiln==1,]
pot_llan <- pottery[pottery$Kiln==2,]
pot_cald <- pottery[pottery$Kiln==3,]
pot_is <- pottery[pottery$Kiln==4,]
pot_ar <- pottery[pottery$Kiln==5,]
```

Note that the third Kiln site, for Caldicot, only has 2 observations. Such a small sample size renders our conclusions somewhat dubious, and we return to this concern below.

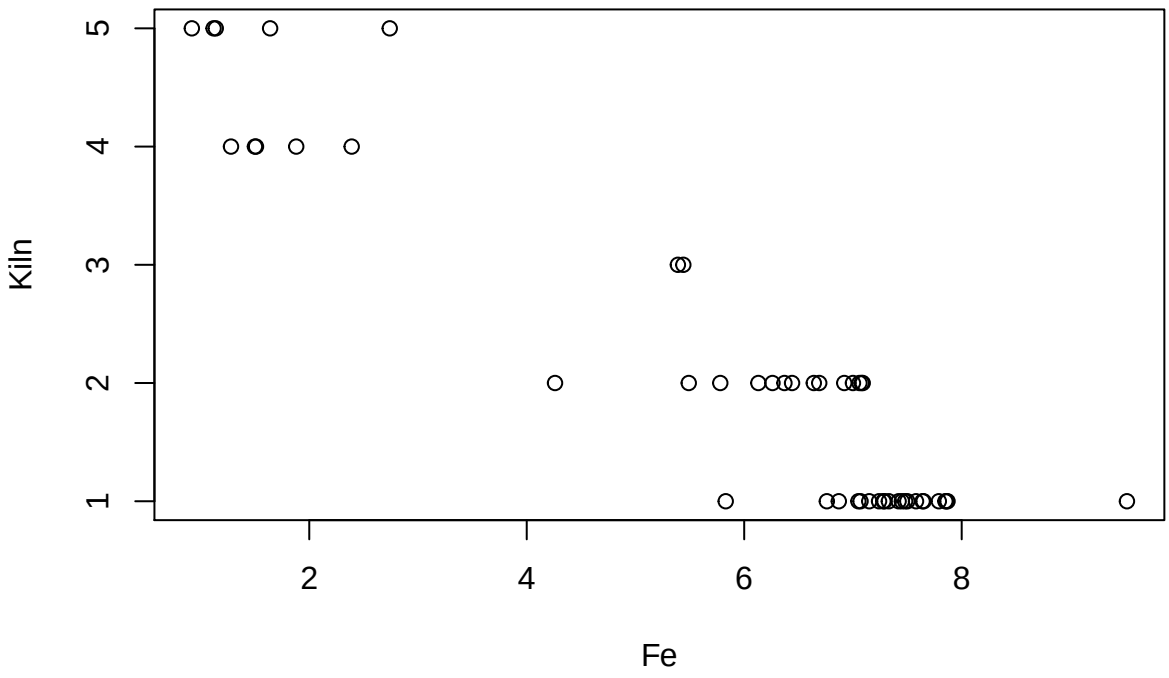
Visualization

Before formulating our test hypothesis, we just examine the data. We are not principally concerned with correlations between the variables, but rather want to see if Kiln site has an impact. So we choose to plot each of the nine variables separately, each by Kiln. We could do this with a histogram, but the sample sizes are small so a scatterplot will suffice.

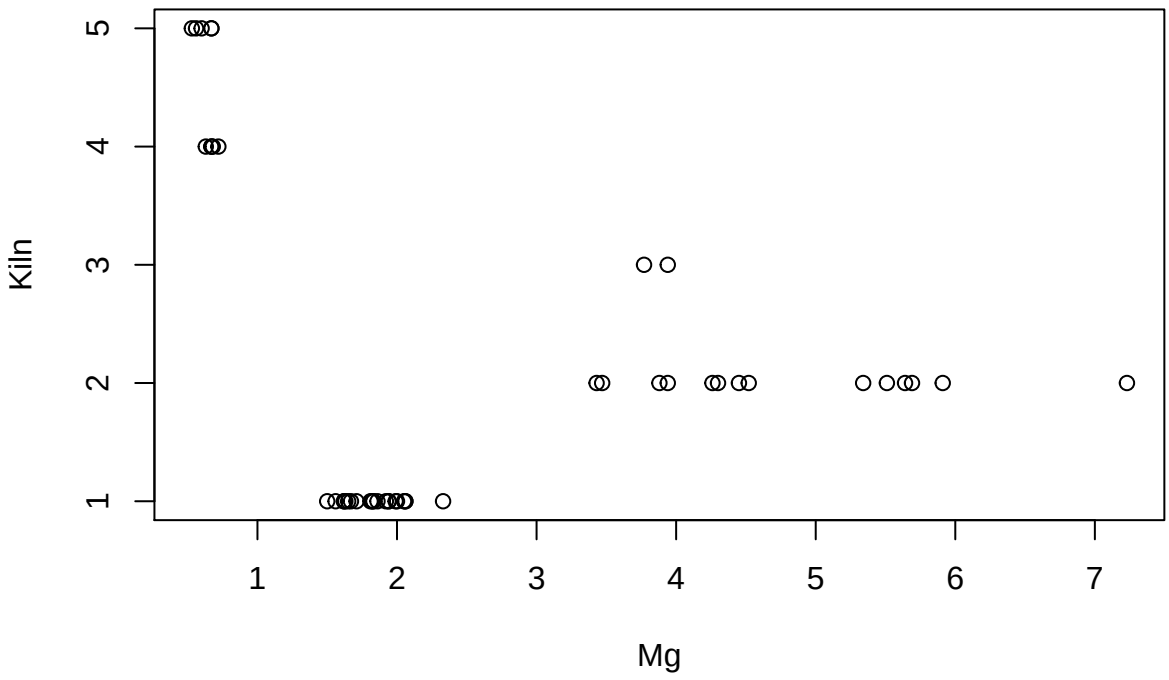
```
plot(pottery$Al,pottery$Kiln, xlab="Al",ylab="Kiln")
```



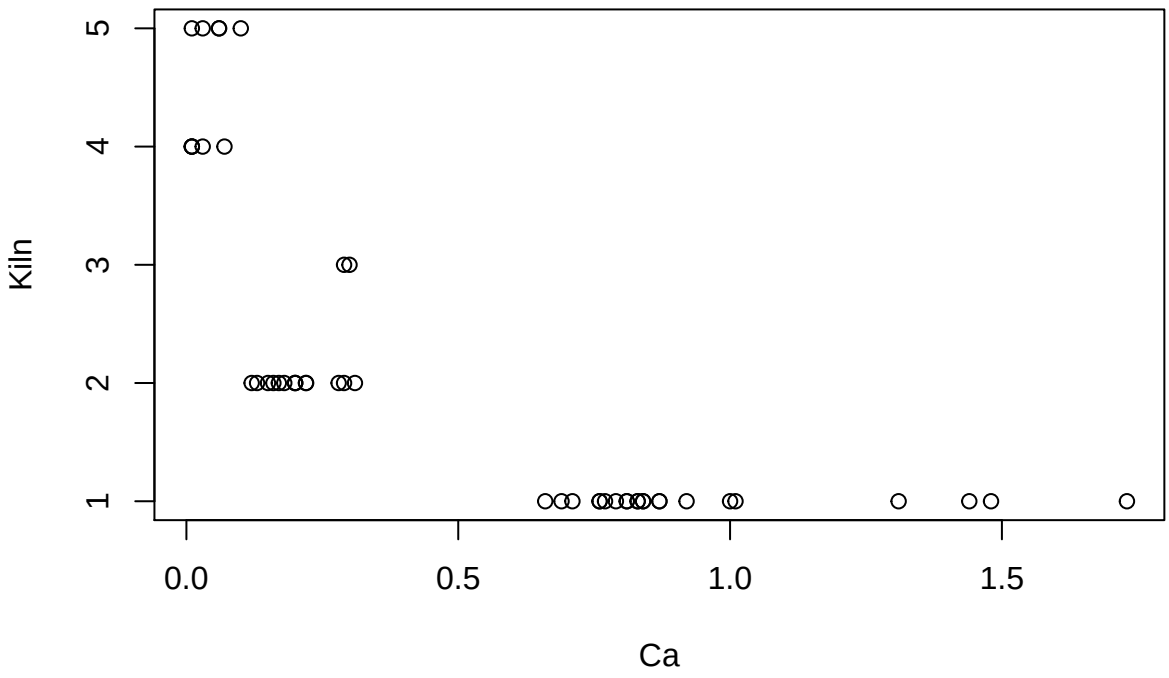

```
plot(pottery$Fe,pottery$Kiln, xlab="Fe",ylab="Kiln")
```



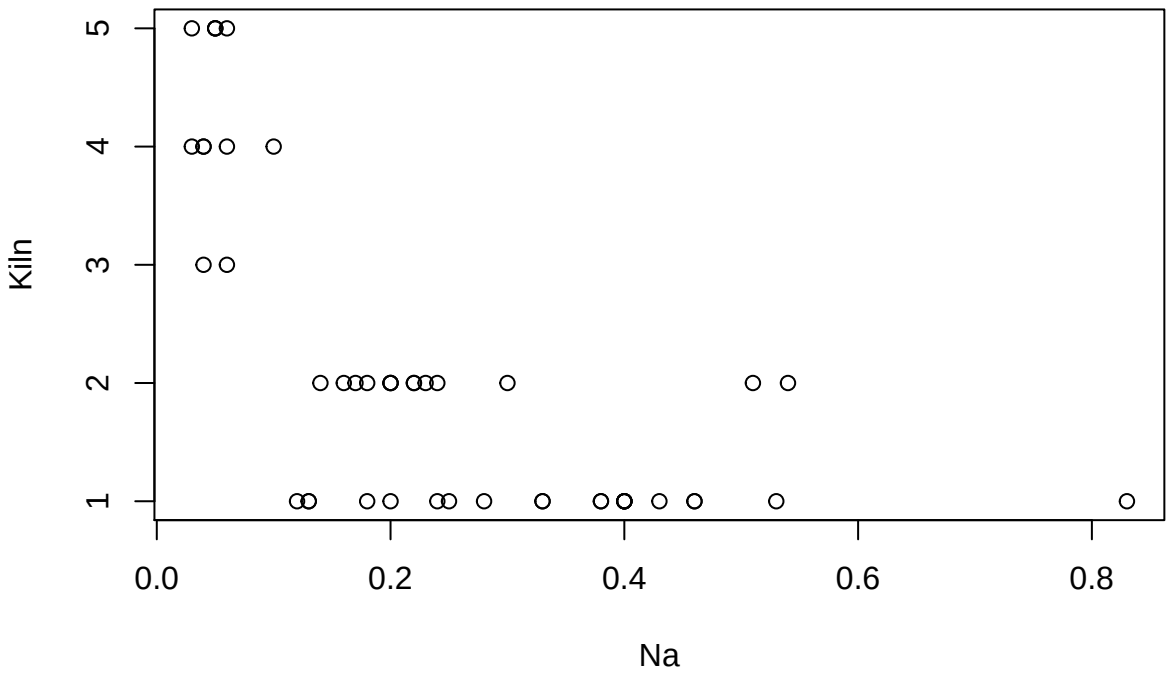

```
plot(pottery$Mg,pottery$Kiln, xlab="Mg",ylab="Kiln")
```



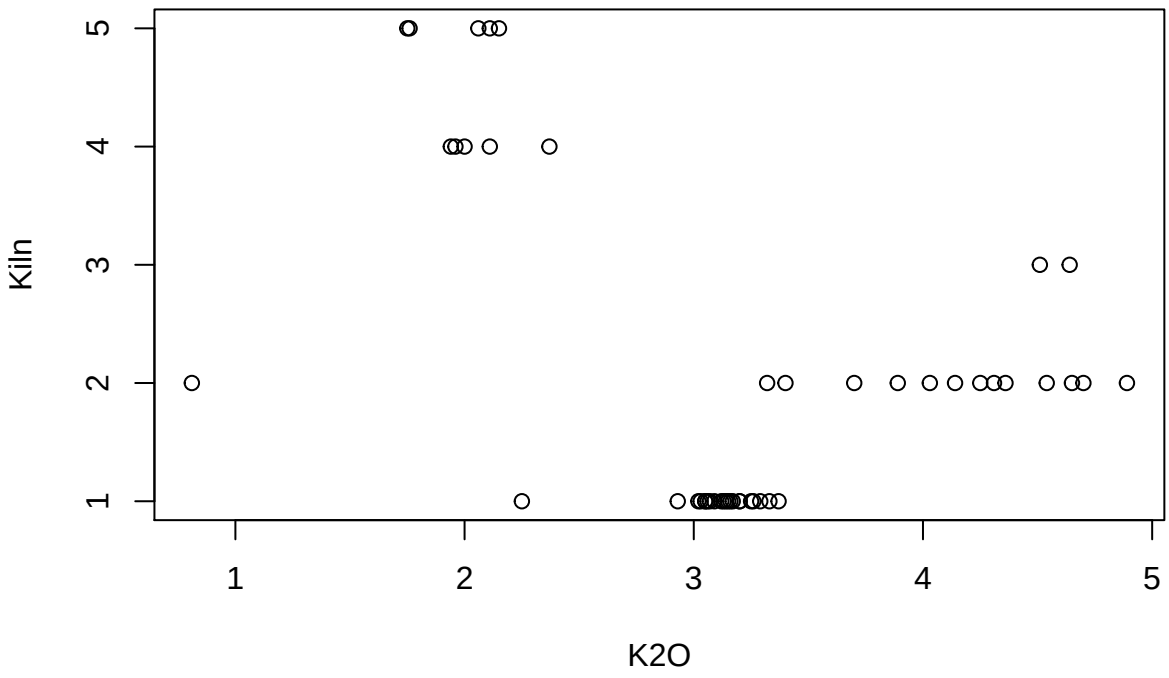

```
plot(pottery$Ca,pottery$Kiln, xlab="Ca",ylab="Kiln")
```



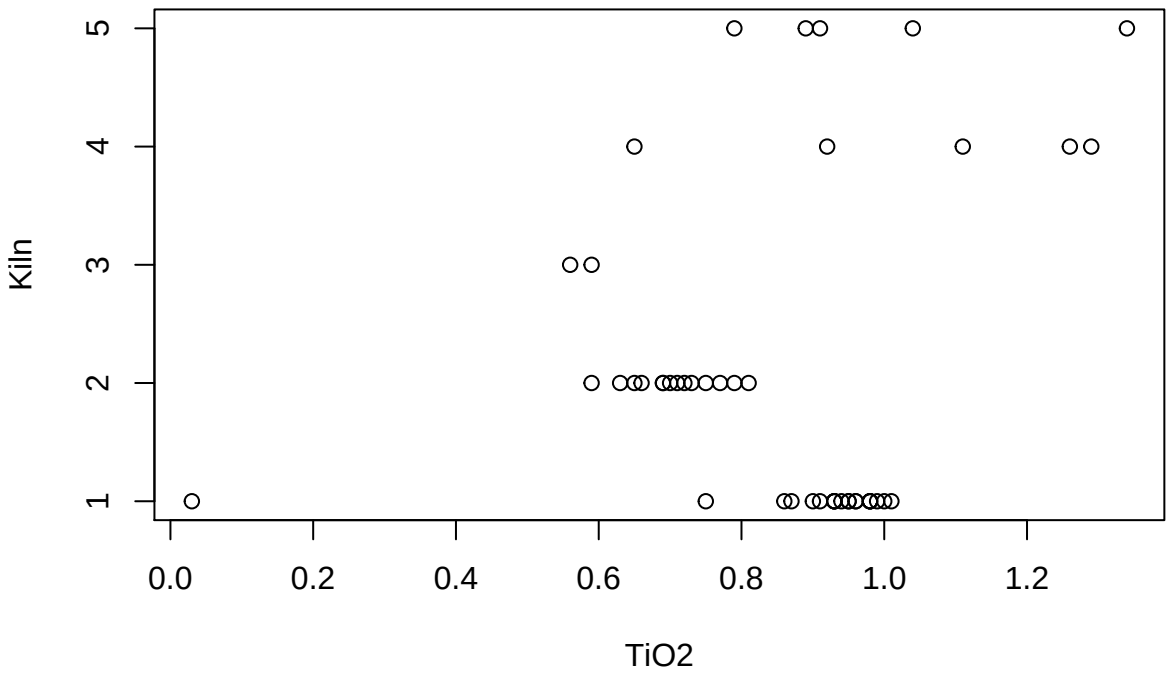

```
plot(pottery$Na,pottery$Kiln, xlab="Na",ylab="Kiln")
```



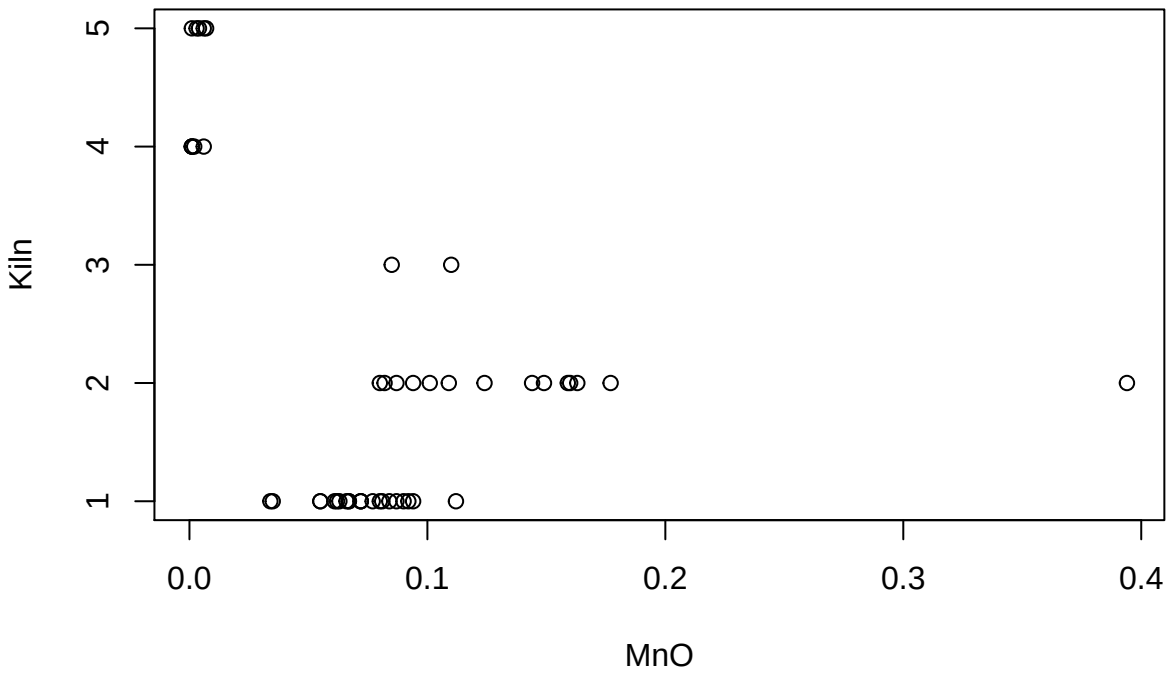

```
plot(pottery$K2O,pottery$Kiln, xlab="K2O",ylab="Kiln")
```



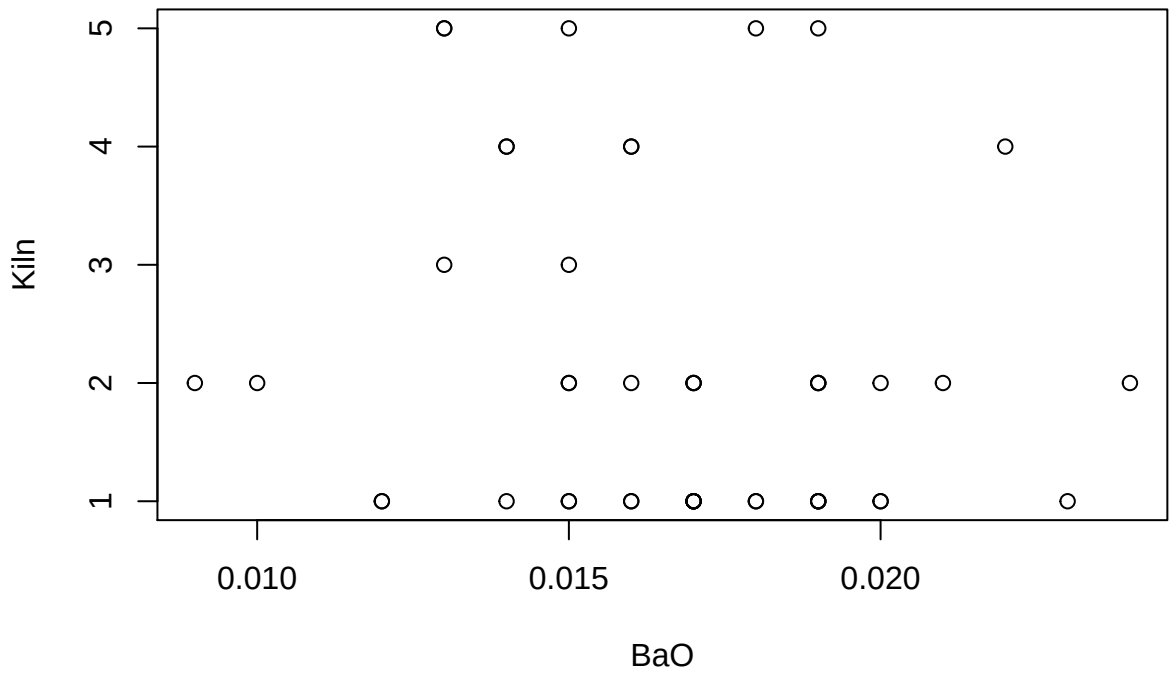

```
plot(pottery$TiO2,pottery$Kiln, xlab="TiO2",ylab="Kiln")
```




```
plot(pottery$MnO,pottery$Kiln, xlab="MnO",ylab="Kiln")
```




```
plot(pottery$BaO,pottery$Kiln, xlab="BaO",ylab="Kiln")
```



For Fe, Mg, and Ca there seems to be some clear clustering of records by Kiln site, whereas for BaO it is hard to discriminate site. Both the center and the dispersion of the records will play a key role in discerning whether we can discriminate between sites. For instance, in Mg the within-group variability is very small for Kiln sites 1, 4, and 5, which makes the separation somewhat easier. The large variability in BaO is opposite, making it harder to discern sites.

Formulating the Null Hypothesis

Our notation is $X_{ij}^{(k)}$ for the ith record of the jth variable ($1 \leq j \leq 9$) in the kth population ($1 \leq k \leq 5$). The population mean vector is $\underline{\mu}^{(k)}$, where $\mu_j^{(k)} = \mathbb{E}[X_{ij}^{(k)}]$. Assuming homoscedasticity (i.e., that the variance-covariance matrix of $\underline{x}^{(k)}$ does not depend on k), in addition to independence and normality, we can use the ANOVA/MANOVA framework to test

$$H_0 : \underline{\mu}^{(1)} = \underline{\mu}^{(2)} = \underline{\mu}^{(3)} = \underline{\mu}^{(4)} = \underline{\mu}^{(5)}.$$

The alternative is that any of these five vectors are unequal to one of the others. Recall that two vectors are unequal if any of their nine components are not the same. We will use MANOVA to do the testing.

We could also consider 9 separate null hypotheses, one for each variable. Then we could use ANOVA for each of the 9 cases. However, this would involve us in a multiple testing situation, since the overall question of interest is whether *any* of the variables are different across groups. Then we would need to combine the 9 ANOVA analyses using FWER or FDR methodology. Instead we use MANOVA, which automatically combines the 9 individual tests by taking into account cross-correlations between variables.

However, there may be value to doing an ANOVA *after* finishing the MANOVA. Once we have determined *whether* there are significant differences, we may want to know *which* variables are really the most important.

Discussing Assumptions

Because sample sizes are so small, there seems to be little value in judging the normality of each variable. On scientific grounds, normality seems a reasonable distribution to use, given that chemical concentrations cannot be arbitrarily large in a given ceramic shard. The independence assumption seems more shaky, given that we don't know if the shards from a particular site were collected from different jars or from the same object. Moreover, which shards were accessible to archaeologists (and which ones survived through the eons) may be correlated with particular attributes. The best we can say is that independence is an unverified working assumption, and its dubiousness should temper the reliability of our final results accordingly.

Assuming the records have common distribution within a site is not really an assumption, but more like a statistical axiom: without it, very little inference can be done. For the homoskedastic assumption, we form estimates of the variance-covariance matrix for each of the five groups, and make numerical comparisons; more sophisticated testing is beyong the scope of the course.

```
var_glou <- var(pot_glou[,4:12])
round(var_glou,digits=6)
```

```
##             Al        Fe        Mg        Ca        Na      K2O      TiO2
## Al    2.281580  0.668199  0.233251 -0.104907 -0.000922 0.194814  0.045156
## Fe    0.668199  0.425647  0.077651 -0.025231  0.040583 0.077742  0.025142
## Mg    0.233251  0.077651  0.041615 -0.007658  0.003017 0.025026  0.011443
## Ca   -0.104907 -0.025231 -0.007658  0.081390 -0.001710 0.007711 -0.004872
## Na   -0.000922  0.040583  0.003017 -0.001710  0.025549 0.010201  0.001455
## K2O   0.194814  0.077742  0.025026  0.007711  0.010201 0.048264  0.007054
## TiO2  0.045156  0.025142  0.011443 -0.004872  0.001455 0.007054  0.040710
## MnO   0.007107  0.006545  0.001417  0.000632  0.002239 0.001793 -0.000040
## BaO   0.000994  0.000527  0.000125  0.000220  0.000222 0.000202  0.000063
##            MnO      BaO
## Al    0.007107 0.000994
## Fe    0.006545 0.000527
## Mg    0.001417 0.000125
## Ca    0.000632 0.000220
## Na    0.002239 0.000222
## K2O   0.001793 0.000202
## TiO2 -0.000040 0.000063
## MnO   0.000339 0.000024
## BaO   0.000024 0.000007
```

```
var_llan <- var(pot_llan[,4:12])
round(var_llan,digits=6)
```

```
##             Al        Fe        Mg        Ca       Na      K2O      TiO2
## Al    1.896319  0.944236  0.034324 -0.017841 0.036181 0.454302  0.076170
## Fe    0.944236  0.617110  0.033947 -0.010197 0.006598 0.212736  0.040993
## Mg    0.034324  0.033947  1.184225  0.033093 0.002072 0.407576 -0.018029
## Ca   -0.017841 -0.010197  0.033093  0.003387 0.000675 0.005743 -0.001484
## Na    0.036181  0.006598  0.002072  0.000675 0.015038 0.024817  0.001626
## K2O   0.454302  0.212736  0.407576  0.005743 0.024817 1.025587  0.005246
## TiO2  0.076170  0.040993 -0.018029 -0.001484 0.001626 0.005246  0.003825
## MnO   0.016465  0.009835 -0.012292 -0.000279 0.000953 0.005367  0.001210
## BaO   0.003162  0.001022 -0.000776  0.000000 0.000028 0.000442  0.000115
##            MnO       BaO
```

```
## Al    0.016465  0.003162
## Fe    0.009835  0.001022
## Mg   -0.012292 -0.000776
## Ca   -0.000279  0.000000
## Na    0.000953  0.000028
## K2O   0.005367  0.000442
## TiO2  0.001210  0.000115
## MnO   0.006278  0.000192
## BaO   0.000192  0.000016
```

```
var_cald <- var(pot_cald[,4:12])
round(var_cald,digits=6)
```

```
##           Al        Fe        Mg        Ca       Na       K2O      TiO2
## Al    0.0200  0.005000  0.017000  0.001000 -0.00200  0.013000  0.003000
## Fe    0.0050  0.001250  0.004250  0.000250 -0.00050  0.003250  0.000750
## Mg    0.0170  0.004250  0.014450  0.000850 -0.00170  0.011050  0.002550
## Ca    0.0010  0.000250  0.000850  0.000050 -0.00010  0.000650  0.000150
## Na   -0.0020 -0.000500 -0.001700 -0.000100  0.00020 -0.001300 -0.000300
## K2O   0.0130  0.003250  0.011050  0.000650 -0.00130  0.008450  0.001950
## TiO2  0.0030  0.000750  0.002550  0.000150 -0.00030  0.001950  0.000450
## MnO  -0.0025 -0.000625 -0.002125 -0.000125  0.00025 -0.001625 -0.000375
## BaO  -0.0002 -0.000050 -0.000170 -0.000010  0.00002 -0.000130 -0.000030
##            MnO      BaO
## Al   -0.002500 -2.0e-04
## Fe   -0.000625 -5.0e-05
## Mg   -0.002125 -1.7e-04
## Ca   -0.000125 -1.0e-05
## Na    0.000250  2.0e-05
## K2O  -0.001625 -1.3e-04
## TiO2 -0.000375 -3.0e-05
## MnO   0.000312  2.5e-05
## BaO   0.000025  2.0e-06
```

```
var_is <- var(pot_is[,4:12])
round(var_is,digits=6)
```

```
##            Al        Fe        Mg        Ca        Na       K2O      TiO2
## Al    3.15200 -0.546700  0.056100  0.039900  0.038100  0.271900 -0.014100
## Fe   -0.54670  0.190070 -0.008885 -0.005190 -0.003010 -0.030365  0.082685
## Mg    0.05610 -0.008885  0.001030  0.000670  0.000680  0.004845  0.000395
## Ca    0.03990 -0.005190  0.000670  0.000680  0.000570  0.003830  0.001230
## Na    0.03810 -0.003010  0.000680  0.000570  0.000780  0.004870  0.003570
## K2O   0.27190 -0.030365  0.004845  0.003830  0.004870  0.031330  0.016630
## TiO2 -0.01410  0.082685  0.000395  0.001230  0.003570  0.016630  0.070530
## MnO   0.00043  0.000159  0.000019 -0.000009 -0.000006 -0.000021  0.000134
## BaO   0.00086  0.000129  0.000033 -0.000018 -0.000002  0.000012  0.000172
##            MnO       BaO
## Al    0.000430  0.000860
## Fe    0.000159  0.000129
## Mg    0.000019  0.000033
## Ca   -0.000009 -0.000018
## Na   -0.000006 -0.000002
```

```
## K2O  -0.000021  0.000012
## TiO2  0.000134  0.000172
## MnO   0.000005  0.000007
## BaO   0.000007  0.000011
```

```
var_ar <- var(pot_ar[,4:12])
round(var_ar,digits=6)
```

```
##             Al        Fe        Mg        Ca        Na       K2O      TiO2
## Al    2.752000 -0.753300 -0.019900  0.044450 -0.007950 -0.135900 -0.229350
## Fe   -0.753300  0.541720  0.029260 -0.000530 -0.001620  0.057460  0.149790
## Mg   -0.019900  0.029260  0.004030  0.000410 -0.000085  0.008405  0.007745
## Ca    0.044450 -0.000530  0.000410  0.001170 -0.000220 -0.000965 -0.001085
## Na   -0.007950 -0.001620 -0.000085 -0.000220  0.000120  0.001315 -0.000740
## K2O  -0.135900  0.057460  0.008405 -0.000965  0.001315  0.038130  0.009220
## TiO2 -0.229350  0.149790  0.007745 -0.001085 -0.000740  0.009220  0.045330
## MnO   0.002170 -0.000103  0.000046  0.000040 -0.000022 -0.000207  0.000049
## BaO   0.002435  0.000216  0.000120  0.000061 -0.000018  0.000038  0.000074
##             MnO       BaO
## Al    0.002170  0.002435
## Fe   -0.000103  0.000216
## Mg    0.000046  0.000120
## Ca    0.000040  0.000061
## Na   -0.000022 -0.000018
## K2O  -0.000207  0.000038
## TiO2  0.000049  0.000074
## MnO   0.000006  0.000006
## BaO   0.000006  0.000008
```

The numbers come out fairly different, so that homoskedasticity might not seem plausible. However, samples are small which means that there will be more variability in the estimates of these variance-covariance matrices. So we shall proceed with these assumptions, but with one exception: we decide to excise the third Kiln site (Caldicot) completely. With only 2 observations, it seems very weak to include this data: if there really are significant differences among the four other sites, the presence of Caldicot will hardly make this more apparent. Conversely, if there was no significant discrepancy based on the other four sites, but the addition of Caldicot were to alter those results, then we would still be skeptical.

MANOVA

We now proceed with MANOVA on 9 variables for Kilns 1, 2, 4, and 5. We choose to adapt the R code from class, because it is easy to do and is transparent.

- First obtain the group means and the grand mean. We print these out to take a look.

```
pot <- NULL
pot <- rbind(pot,pot_glou)
pot <- rbind(pot,pot_llan)
pot <- rbind(pot,pot_is)
pot <- rbind(pot,pot_ar)

# Group: kiln 1
x1 <- pot[pot$Kiln==1,4:12]
```

```
m1 <- colMeans(x1)
n1 <- dim(x1)[1]
# Group: kiln 2
x2 <- pot[pot$Kiln==2,4:12]
m2 <- colMeans(x2)
n2 <- dim(x2)[1]
# Group: kiln 4
x4 <- pot[pot$Kiln==4,4:12]
m4 <- colMeans(x4)
n4 <- dim(x4)[1]
# Group: kiln 5
x5 <- pot[pot$Kiln==5,4:12]
m5 <- colMeans(x5)
n5 <- dim(x5)[1]
# Grand Mean
mg <- (m1*n1 + m2*n2 + m4*n4 + m5*n5)/(n1+n2+n4+n5)
all_means <- rbind(m1,m2,m4,m5,mg)
print(all_means)
```

```
##          Al       Fe       Mg        Ca        Na      K2O      TiO2        MnO
## m1 16.94091 7.430909 1.836364 0.9422727 0.3481818 3.105455 0.8963636 0.07172727
## m2 12.56429 6.372143 4.826429 0.2021429 0.2507143 3.927857 0.7064286 0.14450000
## m4 18.18000 1.712000 0.674000 0.0260000 0.0540000 2.076000 1.0460000 0.00220000
## m5 17.32000 1.512000 0.606000 0.0520000 0.0480000 1.966000 0.9940000 0.00420000
## mg 15.78478 5.843696 2.486304 0.5206522 0.2539130 3.120000 0.8654348 0.07897826
##           BaO
## m1 0.01713636
## m2 0.01700000
## m4 0.01640000
## m5 0.01560000
## mg 0.01684783
```

- Next, get the ESS and HSS matrices. We print these out to take a look.

```
ESS <- cov(x1)*(n1-1) + cov(x2)*(n2-1) + cov(x4)*(n4-1) + cov(x5)*(n5-1)
HSS <- n1*(m1 - mg) %*% t(m1 - mg) + n2*(m2 - mg) %*% t(m2 - mg) +
  n4*(m4 - mg) %*% t(m4 - mg) + n5*(m5 - mg) %*% t(m5 - mg)
round(ESS,digits=4)
```

```
##           Al      Fe      Mg      Ca      Na     K2O    TiO2     MnO     BaO
## Al   96.1813 21.1073  5.4893 -2.0976  0.5716 10.5410  0.9647  0.3737  0.0752
## Fe   21.1073 19.8882  2.1535 -0.6853  0.9195  4.5065  1.9908  0.2655  0.0257
## Mg    5.4893  2.1535 16.2891  0.2737  0.0927  5.8770  0.0385 -0.1298 -0.0069
## Ca   -2.0976 -0.6853  0.2737  1.7606 -0.0257  0.2481 -0.1210  0.0098  0.0048
## Na    0.5716  0.9195  0.0927 -0.0257  0.7356  0.5616  0.0630  0.0593  0.0049
## K2O  10.5410  4.5065  5.8770  0.2481  0.5616 14.6240  0.3197  0.1065  0.0102
## TiO2  0.9647  1.9908  0.0385 -0.1210  0.0630  0.3197  1.3681  0.0156  0.0038
## MnO   0.3737  0.2655 -0.1298  0.0098  0.0593  0.1065  0.0156  0.0888  0.0030
## BaO   0.0752  0.0257 -0.0069  0.0048  0.0049  0.0102  0.0038  0.0030  0.0004
```

```
round(HSS,digits=4)
```

```
##              Al       Fe        Mg      Ca      Na      K2O    TiO2     MnO
##  [1,]  215.0780 -66.1877 -158.1779 15.5630 -1.4329 -58.1550 11.1051 -4.6321
##  [2,]  -66.1877 238.5055   72.7817 32.7351 11.8577  52.0302 -6.6111  3.4373
##  [3,] -158.1779  72.7817  120.0598 -7.5752  2.2947  46.9844 -8.4965  3.6490
##  [4,]   15.5630  32.7351   -7.5752  7.6527  1.8656   1.5489  0.2481  0.0057
##  [5,]   -1.4329  11.8577    2.2947  1.8656  0.6075   2.1653 -0.2416  0.1358
##  [6,]  -58.1550  52.0302   46.9844  1.5489  2.1653  21.2498 -3.4926  1.5756
##  [7,]   11.1051  -6.6111   -8.4965  0.2481 -0.2416  -3.4926  0.6207 -0.2682
##  [8,]   -4.6321   3.4373    3.6490  0.0057  0.1358   1.5756 -0.2682  0.1187
##  [9,]   -0.0145   0.0475    0.0166  0.0060  0.0023   0.0112 -0.0013  0.0007
##           BaO
##  [1,] -0.0145
##  [2,]  0.0475
##  [3,]  0.0166
##  [4,]  0.0060
##  [5,]  0.0023
##  [6,]  0.0112
##  [7,] -0.0013
##  [8,]  0.0007
##  [9,]  0.0000
```

- We decide to examine all four statistics discussed in class. Note that for the Roy maximum root statistic we wrap the eigenvalue output with the Re() function, since the imaginary parts are all zeroes anyways for a symmetric non-negative definite matrix.

```
library(rootWishart)
N <- n1+n2+n4+n5
g <- 4
p <- 9
output <- NULL

# Wilks Lambda
wilks <- det(ESS)/det(ESS + HSS)
wilk_f <- ((N - g) - (p - g + 2)/2)
wilk_xi <- 1
if((p^2 + (g-1)^2 - 5) > 0)
{
  wilk_xi <- sqrt((p^2*(g-1)^2 - 4)/(p^2 + (g-1)^2 - 5))
}
wilk_omega <- (p*(g-1)-2 )/2
wilks_stat <- (wilk_f*wilk_xi - wilk_omega)*
  (1 - wilks^(1/wilk_xi))/(p*(g-1)*wilks^(1/wilk_xi))
output <- rbind(output,c(wilks,wilks_stat,
  1 - pf(wilks_stat,df1 = p*(g-1), df2 = (wilk_f*wilk_xi - wilk_omega))))

# Pillai's Trace
pillai <- sum(diag(HSS %*% solve(ESS + HSS)))
pillai_s <- min(p,g-1)
pillai_m <- (abs(p-g+1)-1)/2
pillai_r <- (N-g-p-1)/2
pillai_stat <- (2*pillai_r + pillai_s + 1)*pillai/
```

```
  ((2*pillai_m + pillai_s + 1)*(pillai_s - pillai))
output <- rbind(output,c(pillai,pillai_stat,
  1 - pf(pillai_stat,df1 = pillai_s*(2*pillai_m + pillai_s + 1),
       df2 = pillai_s*(2*pillai_r + pillai_s + 1))))

# Hotelling-Lawley
hotel <- sum(diag(HSS %*% solve(ESS)))
hotel_b <- (N-p-2)*(N-g-1)/((N-g-p-3)*(N-g-p))
hotel_df1 <- p*(g-1)
hotel_df2 <- 4 + (hotel_df1 + 2)/(hotel_b - 1)
hotel_c <- hotel_df1*(hotel_df2 - 2)/(hotel_df2*(N-g-p-1))
hotel_stat <- hotel/hotel_c
output <- rbind(output,c(hotel,hotel_stat,
  1 - pf(hotel_stat,df1 = hotel_df1,df2 = hotel_df2)))

# Roy
roy <- max(Re(eigen(HSS %*% solve(ESS))$values))
roy_stat <- roy/(1+roy)
output <- rbind(output,c(roy,roy_stat,
  1 - doubleWishart(roy_stat,p=p,m=N-g,n=g-1)))
```

```
## Using multiprecision
```

```
colnames(output) <- c("Statistic","Test Statistic","P-value")
rownames(output) <- c("Wilks","Pillai","Hotelling-Lawley","Roy")
output
```

```
##                     Statistic Test Statistic      P-value
## Wilks             0.002109065     26.8210655 0.000000e+00
## Pillai            1.923387071      7.1460671 4.607426e-14
## Hotelling-Lawley 43.124652314     52.6474734 0.000000e+00
## Roy              28.197529917      0.9657505 4.440892e-16
```

- The four statistics provide a similar message: H_0 is rejected in the strongest possible terms. Indeed, the population means differ across the four sites.

ANOVA

At this point our project is done. But we may want to push further: which variables seem to be driving the discrepancy across sites? We could consider dropping out variables until some of the four test statistics are no longer significant. Instead we do something that is less laborious but still insightful: 9 ANOVA analyses.

- To do this, recall the test statistics are obtained from the diagonal entries of the ESS and HSS matrices that we've already computed.

```
df1 <- 4-1
df2 <- N-4
F_stats <- (df2/df1) * diag(HSS)/diag(ESS)
1 - pf(F_stats,df1=df1,df2=df2)
```

```
##           Al           Fe           Mg           Ca           Na          K2O
## 8.602430e-11 0.000000e+00 0.000000e+00 2.442491e-15 1.174614e-05 2.692834e-08
##         TiO2          MnO          BaO
## 1.193501e-03 7.338810e-08 7.805878e-01
```

- Judging by the p-values, TiO2 and BaO are not as compelling (the latter is not significant at a .05 level), but Fe and Mg are very good *predictors* for Kiln.

Summary

We have found very strong evidence of a discrepancy between the sites on the basis of the 9 variables. All four MANOVA statistics (Wilks' Lambda, Pillai's Trace, Hotelling-Lawley, and Roy's Maximum Root) rejected the null hypothesis that the population mean vectors were equal. These findings are tempered by some noted limitations of the data: small sample sizes (we even eliminated the Caldicot Kiln site data, because of the paucity) and inability to verify the independence, normality, or homoskedasticity assumptions. Yet, given the available tools from the course, these findings seem reasonable for a preliminary exploration. Moreover, of the 9 variables it seems that Fe and Mg are very powerful for discriminating between sites; however, we don't know which sites (of the four considered). Therefore, future research could explore which variables are best for predicting discrepancies between particular sites. This could be treated as a classification problem.