Math 189 Project 2

Yilan Guo

2/14/2021

Introduction

In this project, to practice the material we learned up to lecture 12, I will exam on Romano-British Pottery dataset and determine whether there is a significant difference among the 5 group means on each sites for these 9 chemical variables. Through the statistical analysis, I would also discuss about the model (Multivariate Analysis of Variance) assumptions, and calculate four unique test statistics.

Data and Citation

The data, **Romano-British Pottery dataset**, comes from math189 class github repo[1] RBPottery.csv, accessed on February 14, 2021. The dataset contains measurements on pottery shards that were collected from five sites (indicated by using 1,2,3,4, and 5) in the British Isles. The original dataset contains 48 observations and 12 features; for the purpose of the project, **I would only keep 10 columns**: Kiln column to indicate the specific sites and 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

Analysis

To determine if there is a significant difference among the 5 group means (on each sites) for these 9 chemical variables, I choose **multivariate Analysis of Variance model** which allows us to compare mean vectors/nine means together among multiple populations (sites).

Compared to other test:

1. Univariate ANOVA: Because the univariate ANOVA only considers one univariate mean among multiple population and in this project our objective is to compare mean vectors of nine chemical variables on each site, this is not the best choice.

[1] repo: Math189 *RBPottery.cs*, adopted from: Tubb, A., A. J. Parker, and G. Nickless. 1980. “The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry”. Archaeometry 22: 153-71.

2. Two sample hotelling test: Since two sample hotelling test only tests on two population and here we have five populations(site), so this model is not our choice.

Null Hypothesis: There is no significant difference among the 5 group means for these 9 chemical variables and all the mean vectors equal to each other.

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

Alternative Hypothesis: at least one mean vectors are not equal to each other and there is at least one difference in means among the 5 group means for these 9 chemical variables.

$$H_a : \mu_j^{(k)} \neq \mu_j^{(h)}$$

for some variable j, and for some groups k and h.

Alpha = 0.05

Read Data

Here, we read the data and separate the data into new datasets based on site number (Kiln).

```
pottery <- read.csv("~/Downloads/MATH_189/ma189/Data/RBPottery.csv")[,-c(1,2)]
head(pottery)
```

```
##   Kiln Al2O3 Fe2O3  MgO  CaO Na2O  K2O TiO2   MnO   BaO
## 1    1  18.8  9.52 2.00 0.79 0.40 3.20 1.01 0.077 0.015
## 2    1  16.9  7.33 1.65 0.84 0.40 3.05 0.99 0.067 0.018
## 3    1  18.2  7.64 1.82 0.77 0.40 3.07 0.98 0.087 0.014
## 4    1  17.4  7.48 1.71 1.01 0.40 3.16 0.03 0.084 0.017
## 5    1  16.9  7.29 1.56 0.76 0.40 3.05 1.00 0.063 0.019
## 6    1  17.8  7.24 1.83 0.92 0.43 3.12 0.93 0.061 0.019
```

Exploratory Data Analysis

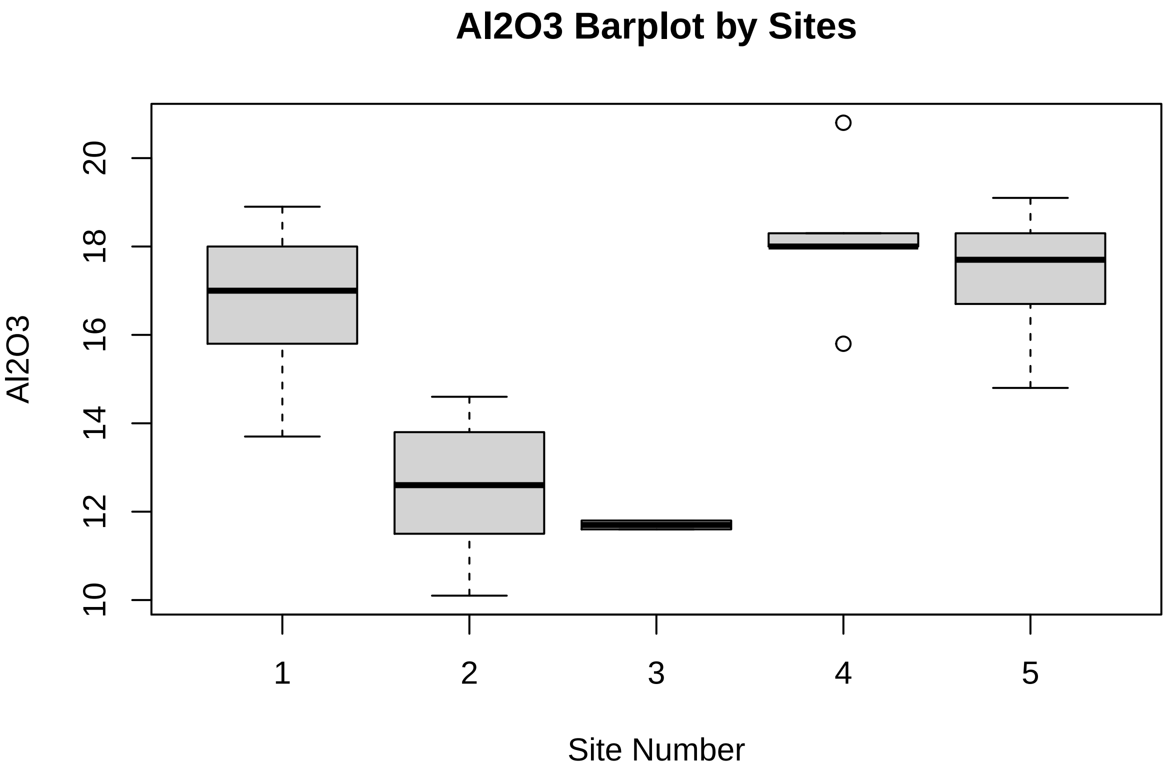
- Calulate the chemical variable means across the sites. Because we assume each observation is iid, the mean and variance would be unbiased estimators.

```
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,]
means = NULL
means <- rbind(means,colMeans(pot_glou))
means <- rbind(means,colMeans(pot_llan))
means <- rbind(means,colMeans(pot_cald))
means <- rbind(means,colMeans(pot_is))
means <- rbind(means,colMeans(pot_ar))
means
```

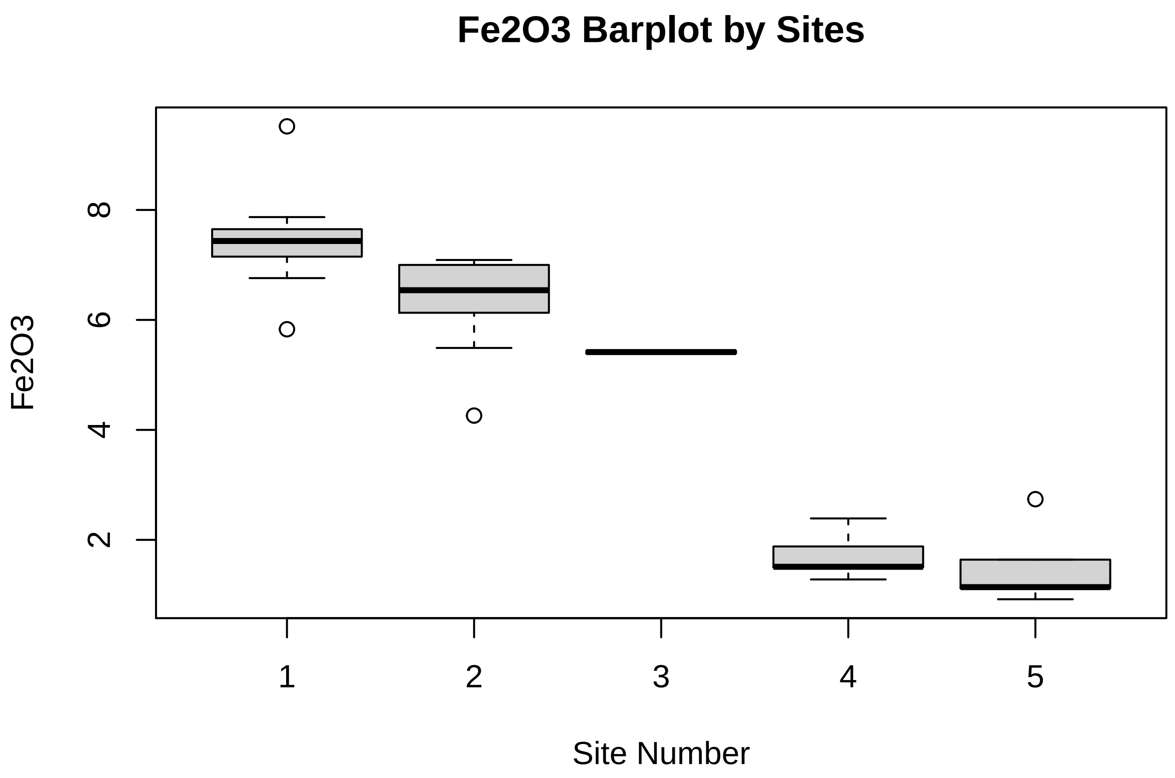
```
##      Kiln    Al2O3    Fe2O3      MgO       CaO      Na2O      K2O      TiO2
## [1,]    1 16.94091 7.430909 1.836364 0.9422727 0.3481818 3.105455 0.8963636
## [2,]    2 12.56429 6.372143 4.826429 0.2021429 0.2507143 3.927857 0.7064286
## [3,]    3 11.70000 5.415000 3.855000 0.2950000 0.0500000 4.575000 0.5750000
## [4,]    4 18.18000 1.712000 0.674000 0.0260000 0.0540000 2.076000 1.0460000
## [5,]    5 17.32000 1.512000 0.606000 0.0520000 0.0480000 1.966000 0.9940000
##             MnO        BaO
## [1,] 0.07172727 0.01713636
## [2,] 0.14450000 0.01700000
## [3,] 0.09750000 0.01400000
## [4,] 0.00220000 0.01640000
## [5,] 0.00420000 0.01560000
```

- Use boxplot to display the distribution of the chemical variable across the sites.

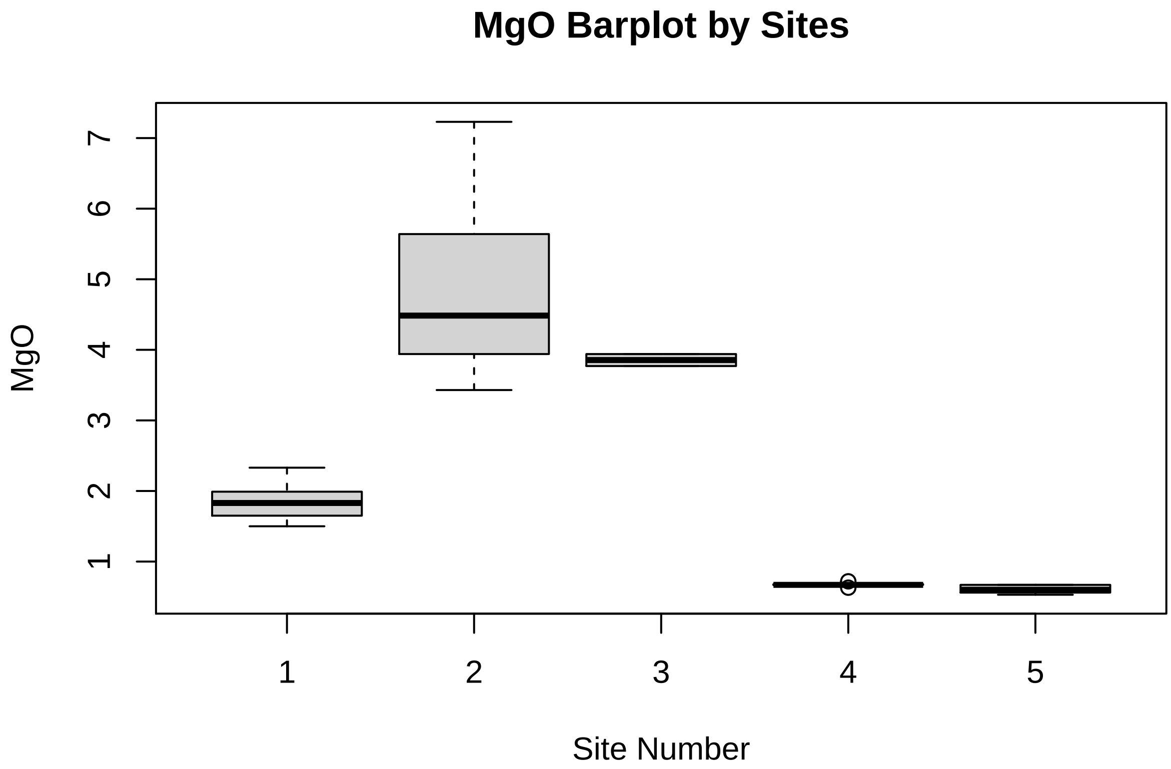
```
boxplot(Al2O3~Kiln,data=pottery, main="Al2O3 Barplot by Sites",
   xlab="Site Number", ylab="Al2O3")
```



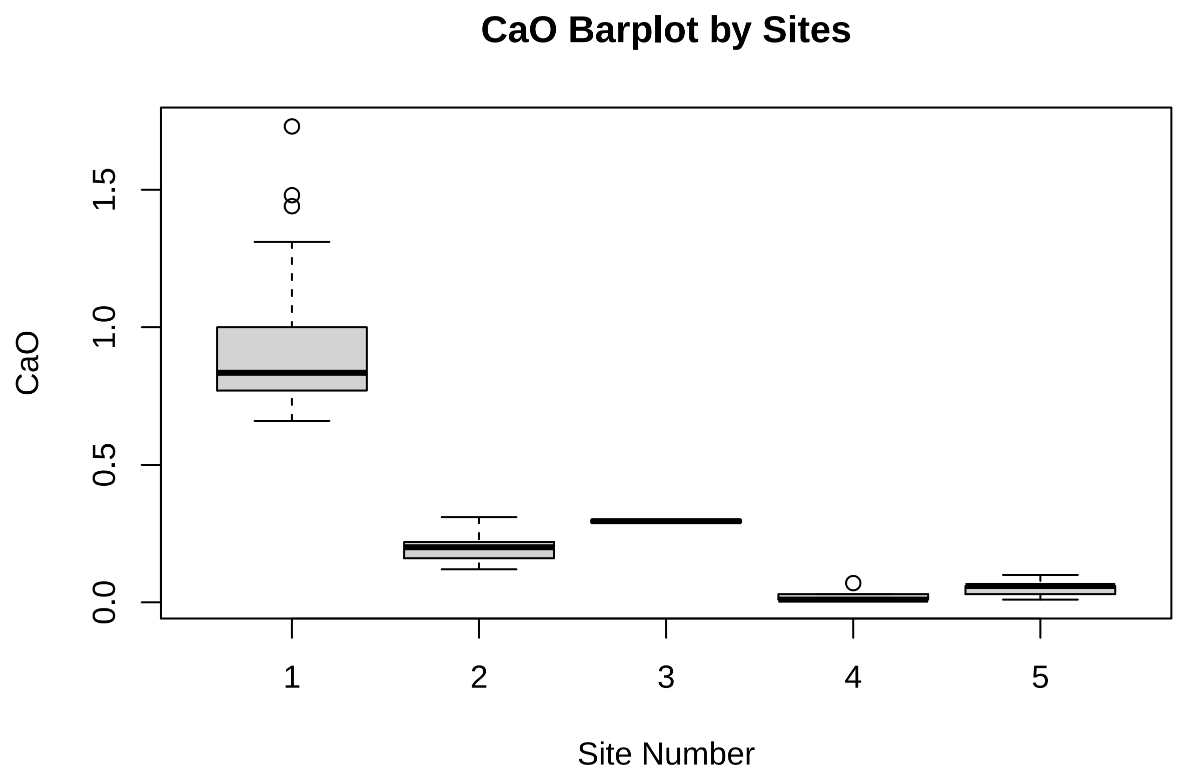
```
boxplot(Fe2O3~Kiln,data=pottery, main="Fe2O3 Barplot by Sites",
   xlab="Site Number", ylab="Fe2O3")
```



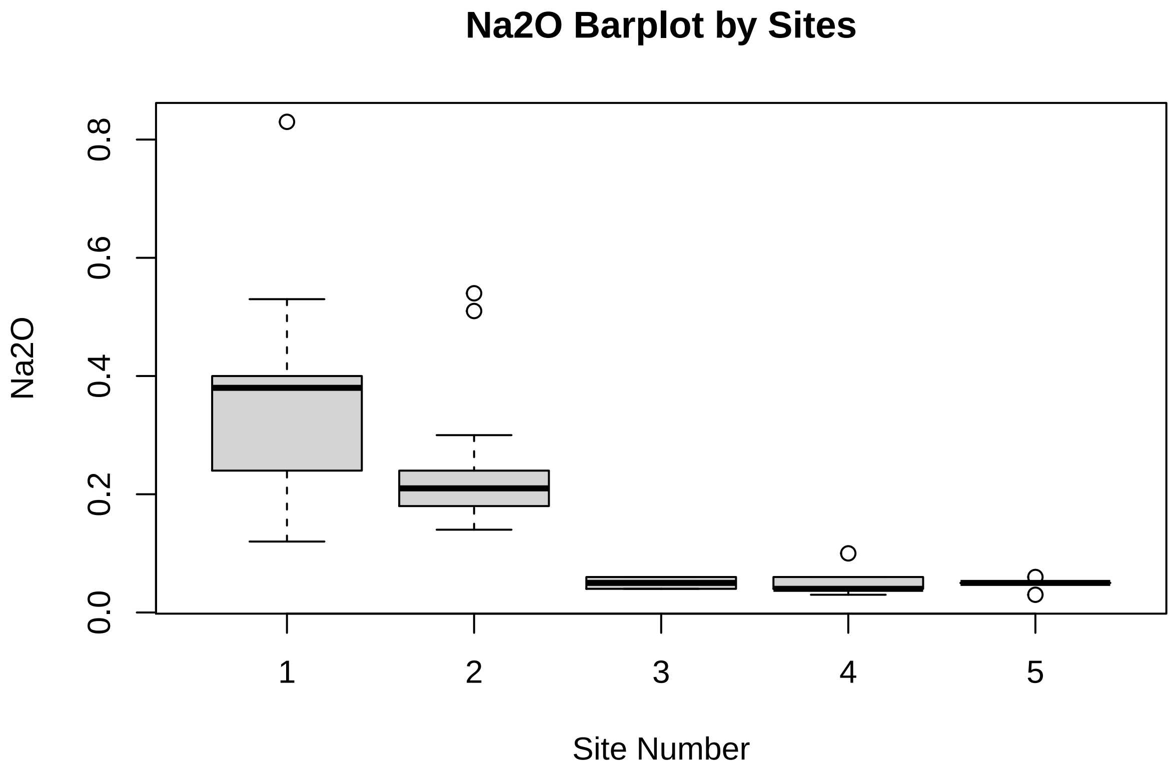
```
boxplot(MgO~Kiln,data=pottery, main="MgO Barplot by Sites",
   xlab="Site Number", ylab="MgO")
```



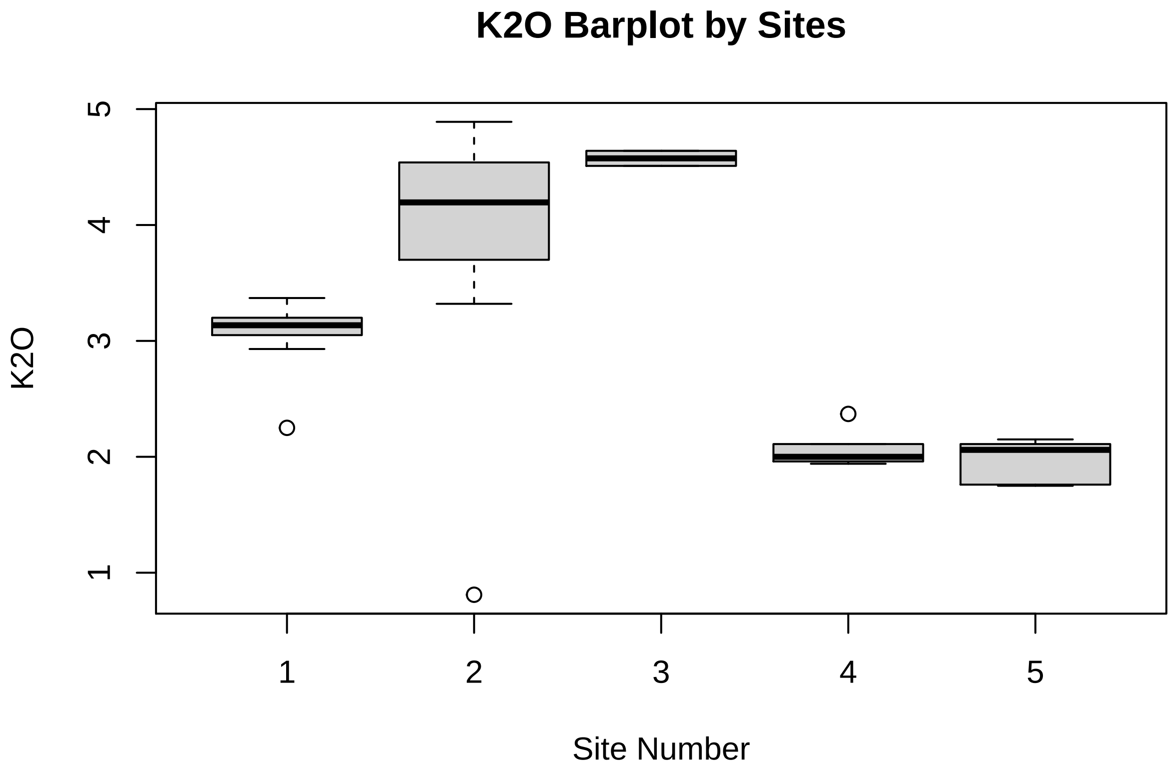
```
boxplot(CaO~Kiln,data=pottery, main="CaO Barplot by Sites",
   xlab="Site Number", ylab="CaO")
```



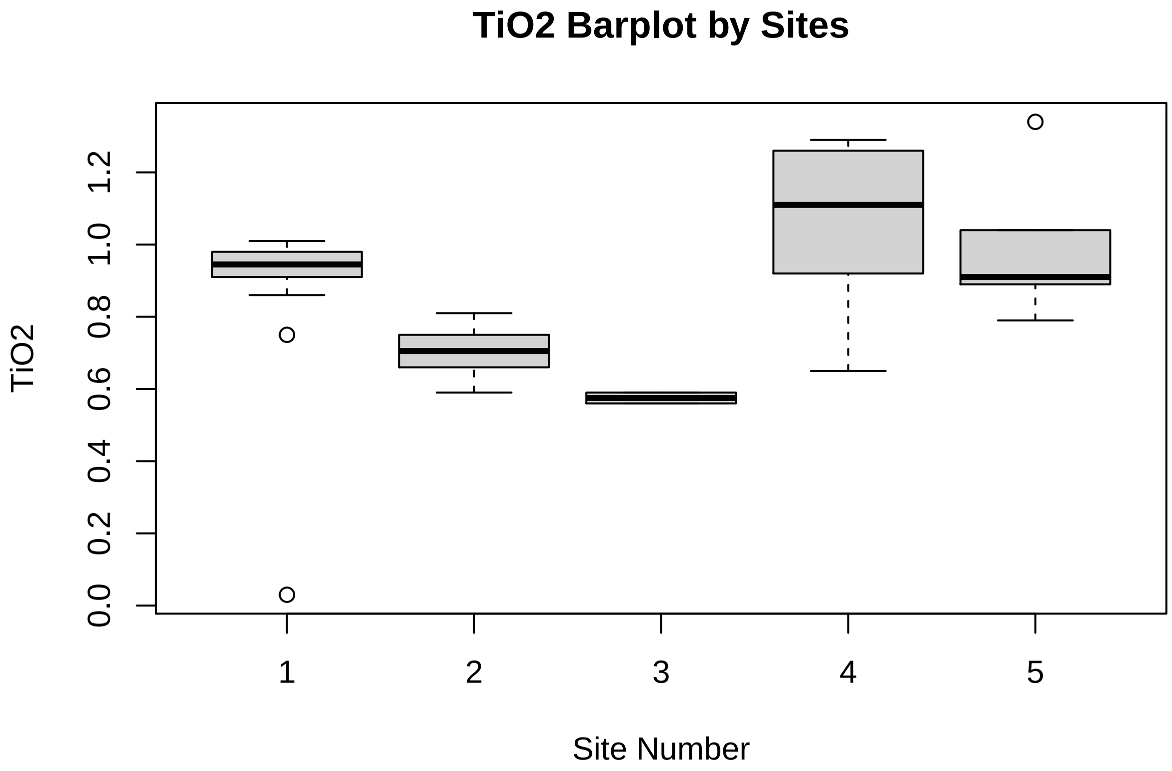
```
boxplot(Na2O~Kiln,data=pottery, main="Na2O Barplot by Sites",
   xlab="Site Number", ylab="Na2O")
```



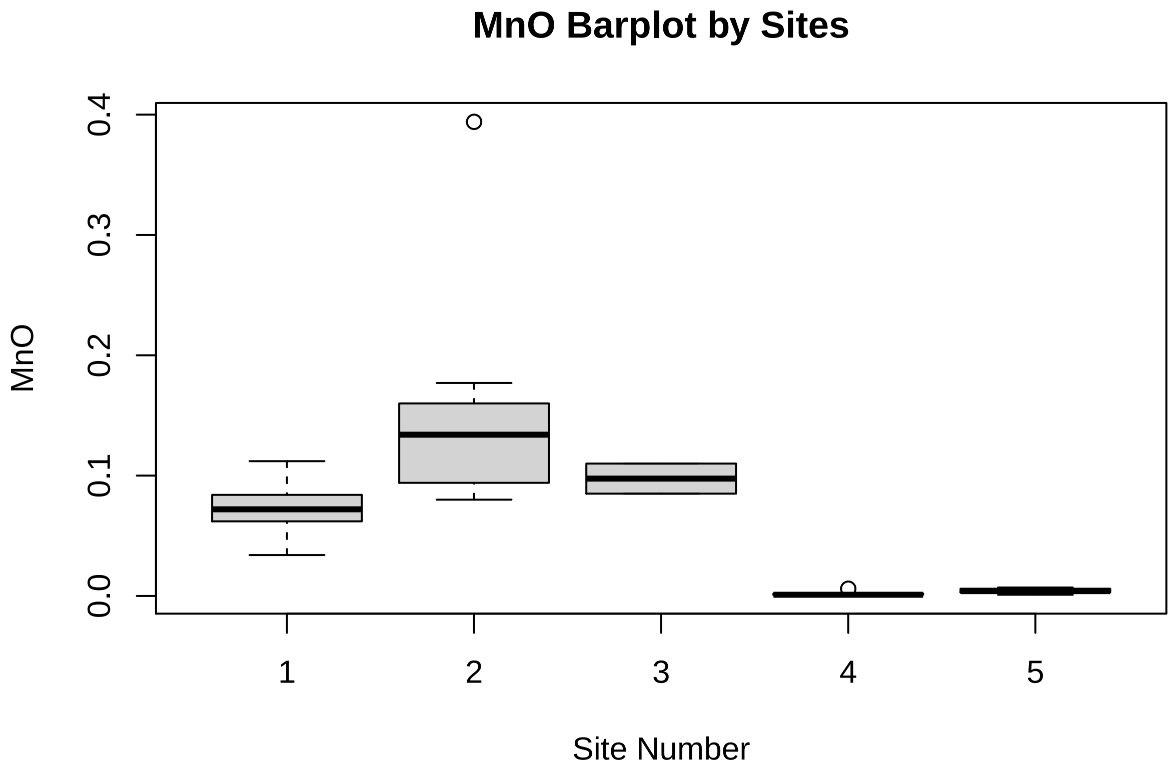
```
boxplot(K2O~Kiln,data=pottery, main="K2O Barplot by Sites",
   xlab="Site Number", ylab="K2O")
```



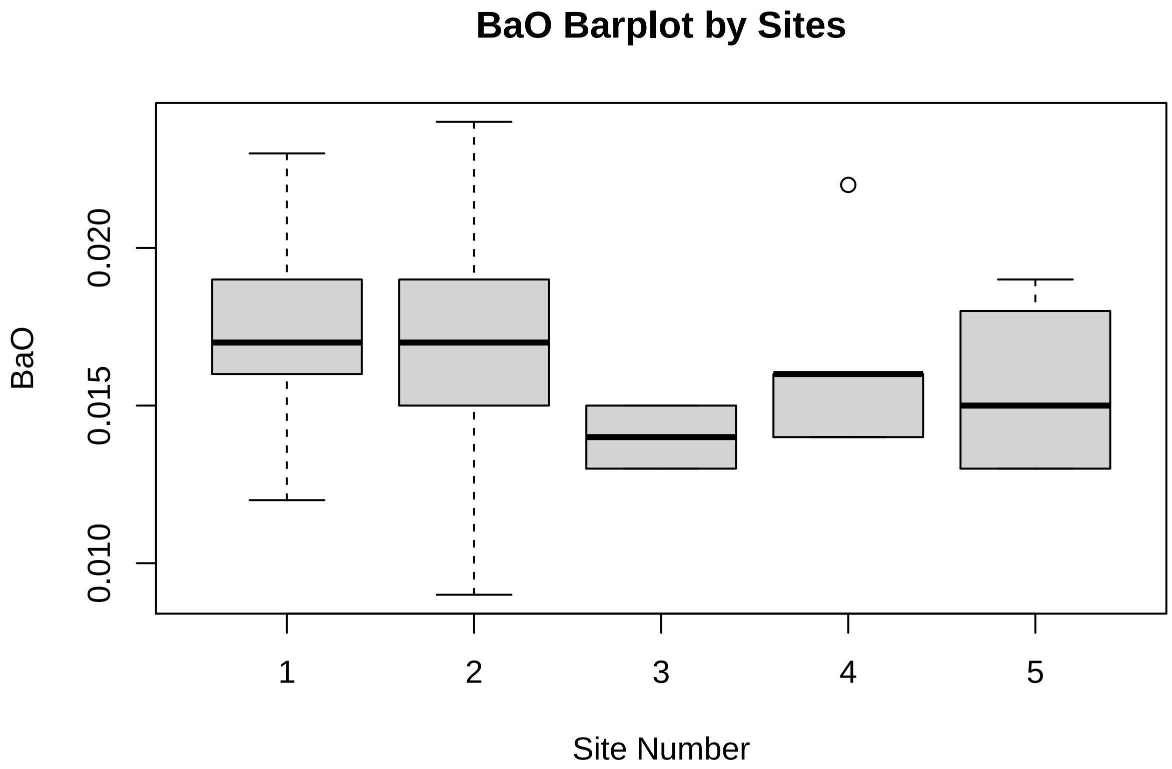
```r
boxplot(TiO2~Kiln,data=pottery, main="TiO2 Barplot by Sites",
   xlab="Site Number", ylab="TiO2")
```



```r
boxplot(MnO~Kiln,data=pottery, main="MnO Barplot by Sites",
   xlab="Site Number", ylab="MnO")
```



```
boxplot(BaO~Kiln,data=pottery, main="BaO Barplot by Sites",
   xlab="Site Number", ylab="BaO")
```



Assumptions for the MANOVA:

- The data from group k has **common mean vector** $\underline{\mu}^{(k)}$, i.e.,

  $$\mathbb{E}[x_{ij}^{(k)}] = \underline{\mu}_j^{(k)}.$$

  (The m components of the vector correspond to the m variables.) We assume this assumption hold since we believe each observation on each site is a random variable that is iid.

- **Homoskedasticity**: The data from all groups have common covariance matrix $\mathbf{\Sigma}$, i.e.,

  $$\mathbf{\Sigma} = \mathrm{Cov}[\underline{x}_i^{(k)}, \underline{x}_i^{(k)}]$$

  for any record i, and the matrix does not depend on k (the group index).

To test equal covariance, I conducted a Box's M-test, which tests for homogeneity of covariance matrices (the variances in each group are roughly equal) using the data obtained from multivariate normal chemical variable data according to one classification factor – site. The test is based on the chi-square approximation.

```
library(biotools)
```

```
## Loading required package: MASS
```

```
## ---
## biotools version 4.0
```

```
##
```

```r
boxM(data = pottery[,2:10],group = pottery[,1])
```

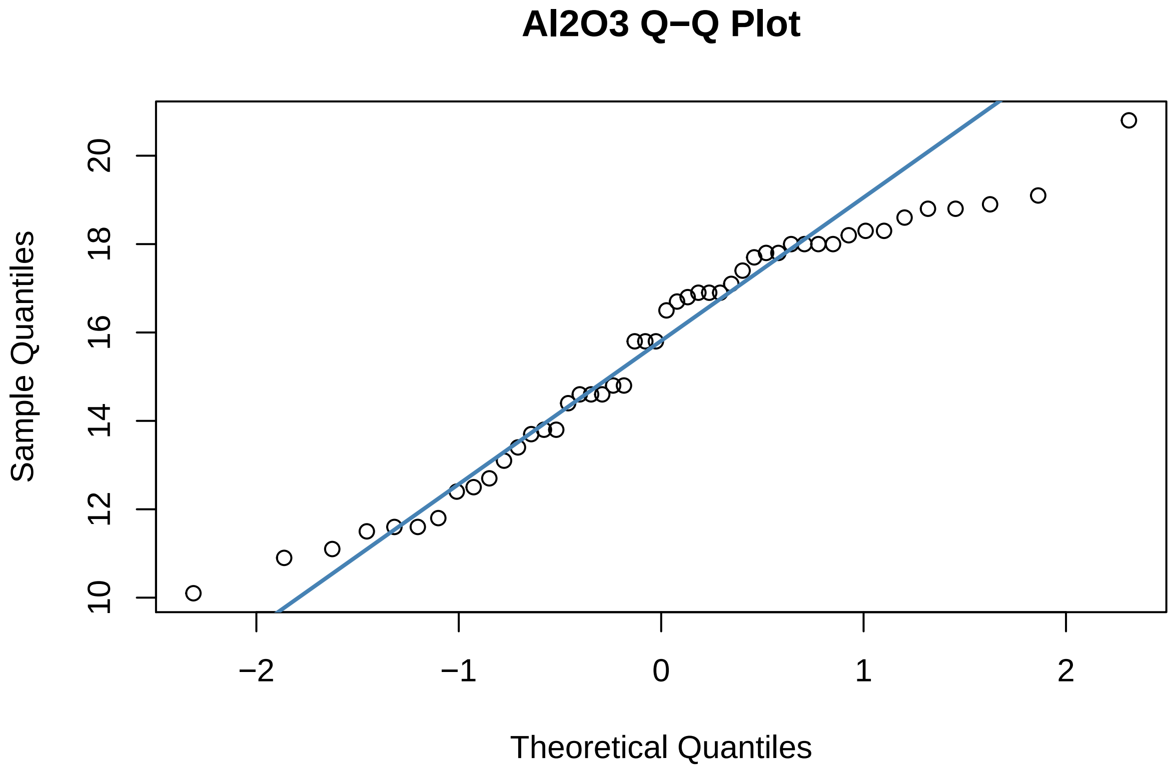
```
## Warning in boxM(data = pottery[, 2:10], group = pottery[, 1]): there are one or
## more levels with less observations than variables!

##
##  Box's M-test for Homogeneity of Covariance Matrices
##
## data:  pottery[, 2:10]
## Chi-Sq (approx.) = -577.02, df = 180, p-value = 1
```

Because the p-value is much greater than 0.05, we fail to reject the null which claims that the variances in each group are roughly equal. However, because the sample size in site 3, 4, and 5 are extremly small (smaller than the observations), the result might not be accurate.

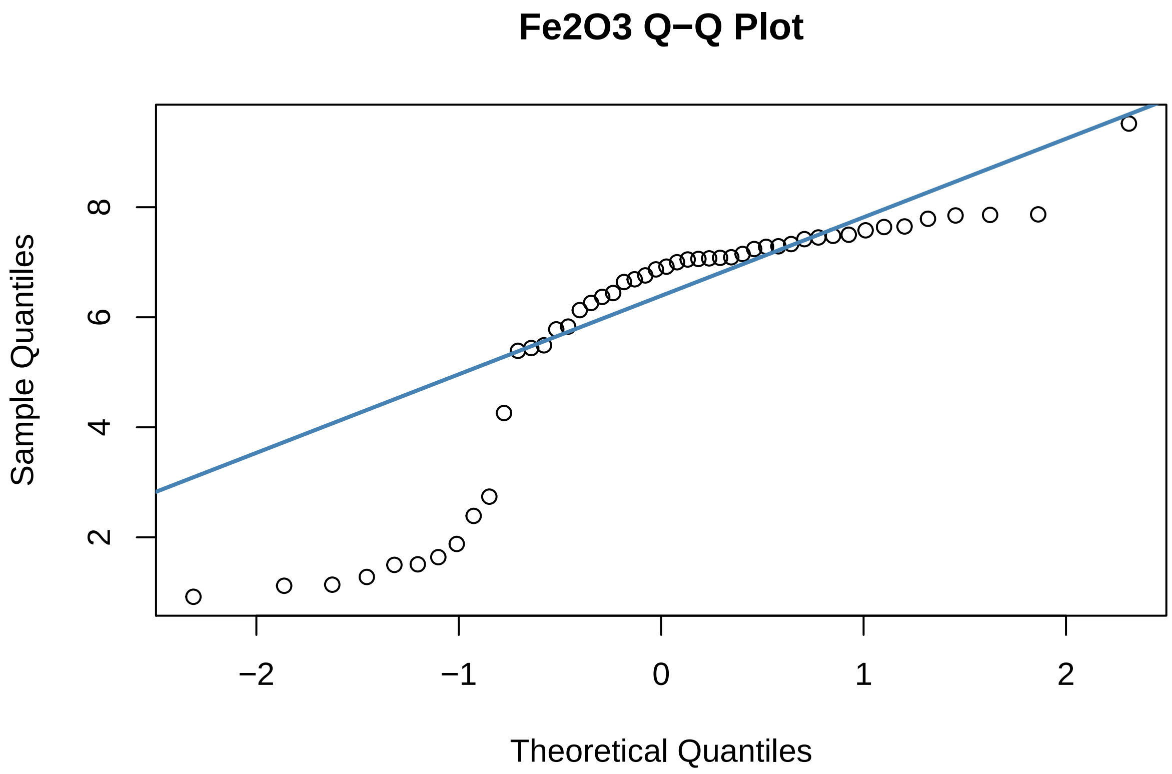
- **Independence**: The observations are independently sampled. For this project, we assume that that each observation is **i.i.d**.
- **Normality**: The data are multivariate normally distributed.

We first test univariate normality; if univariate normality is violated, then multivariate normality assumption is highly likely to be violated as well.

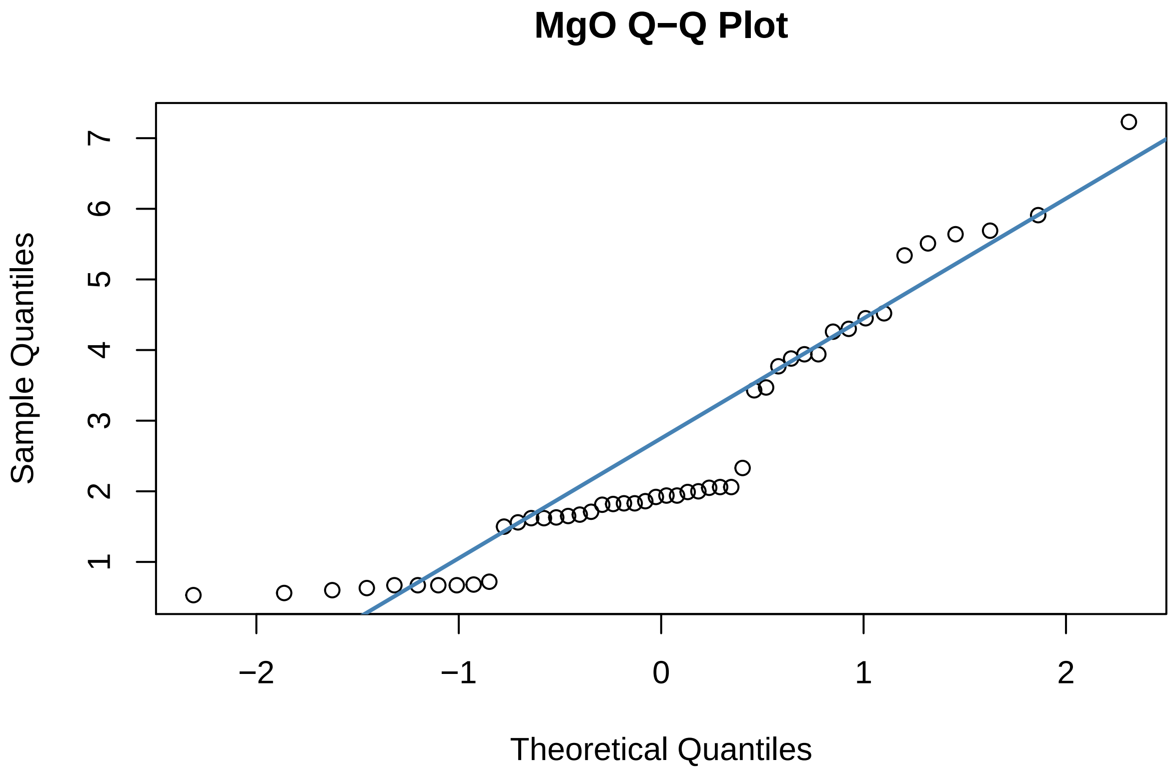
```r
qqnorm(pottery$Al2O3, pch = 1,main = "Al2O3 Q-Q Plot")
qqline(pottery$Al2O3, col = "steelblue", lwd = 2)
```



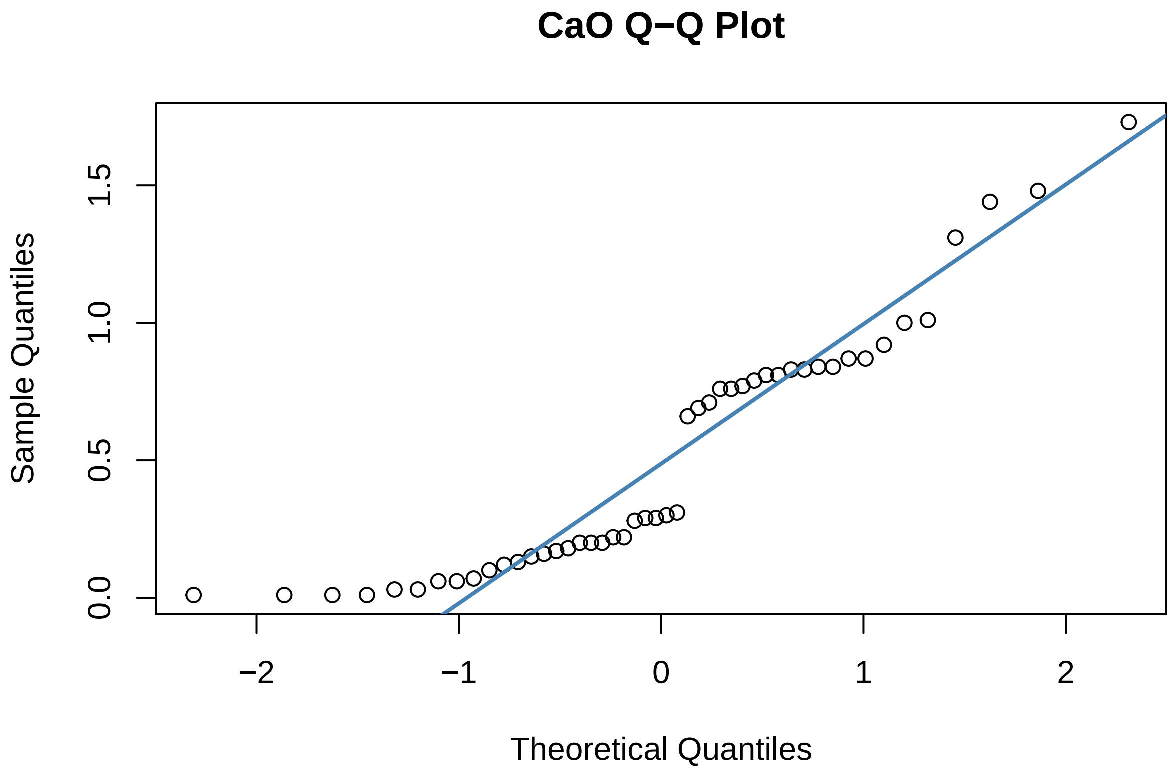
```r
qqnorm(pottery$Fe2O3, pch = 1,main = "Fe2O3 Q-Q Plot")
qqline(pottery$Fe2O3, col = "steelblue", lwd = 2)
```



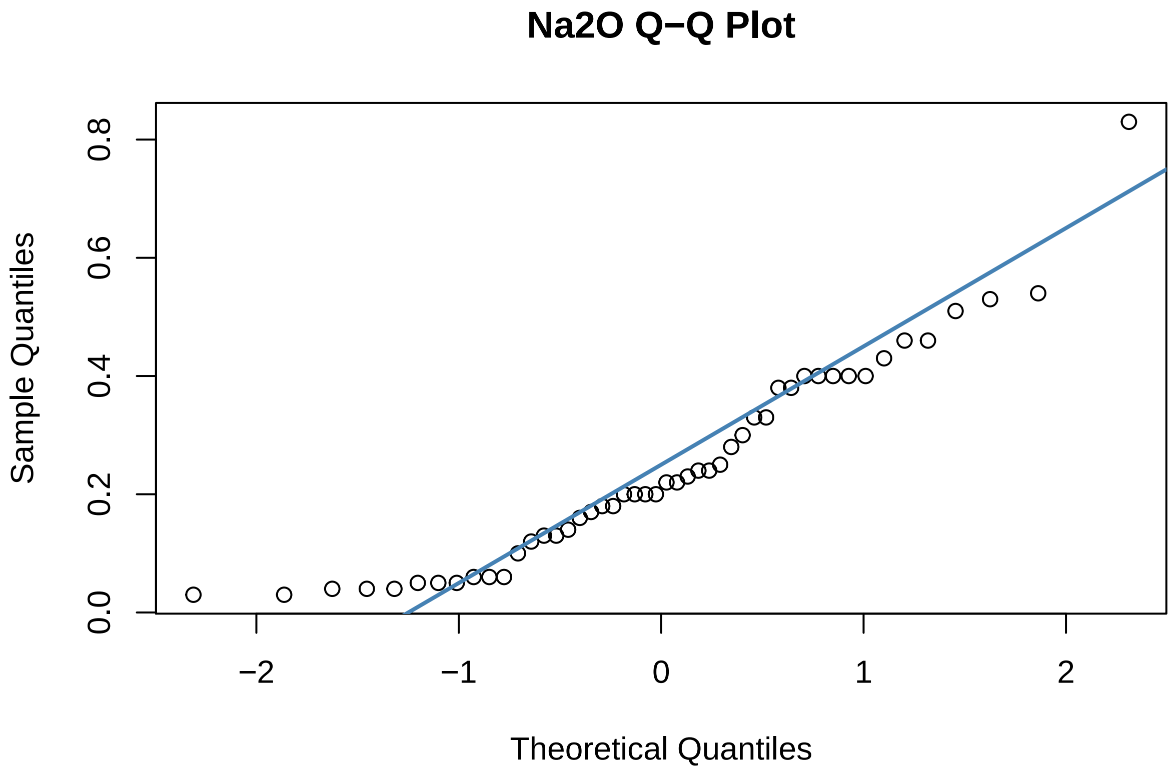
```r
qqnorm(pottery$MgO, pch = 1, main = "MgO Q-Q Plot")
qqline(pottery$MgO, col = "steelblue", lwd = 2)
```



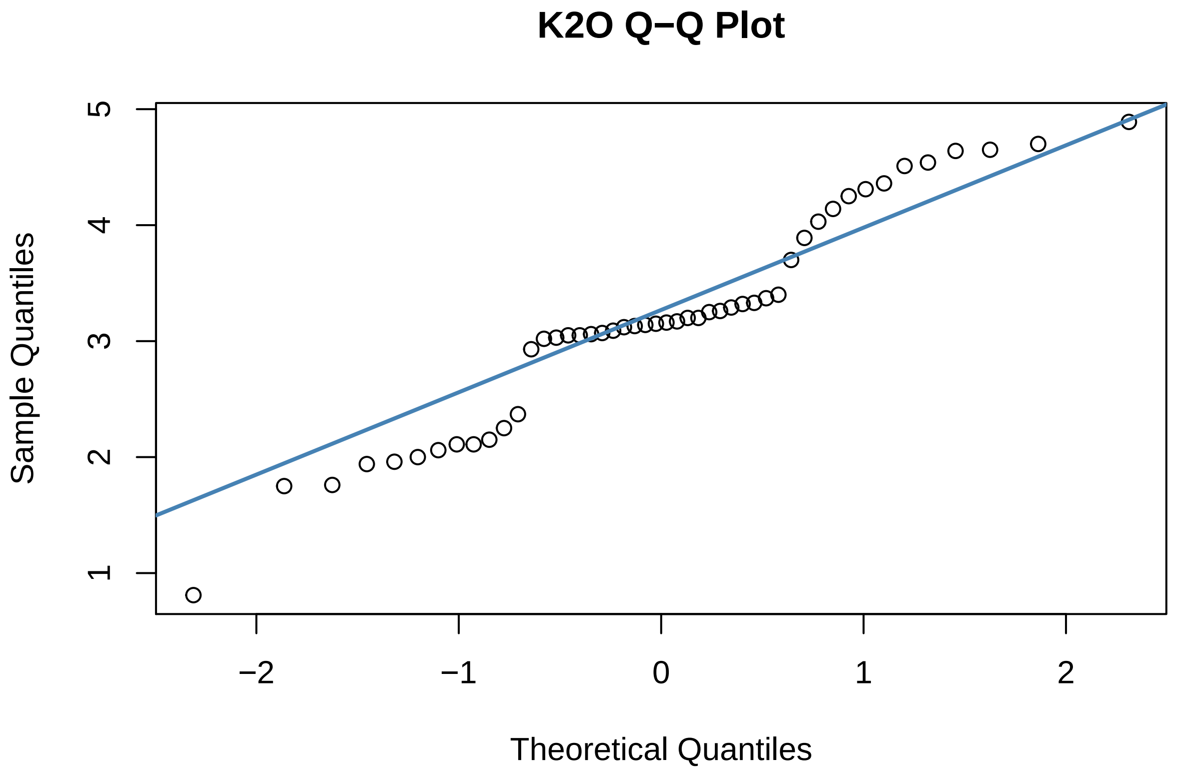
```r
qqnorm(pottery$CaO, pch = 1, main = "CaO Q-Q Plot")
qqline(pottery$CaO, col = "steelblue", lwd = 2)
```



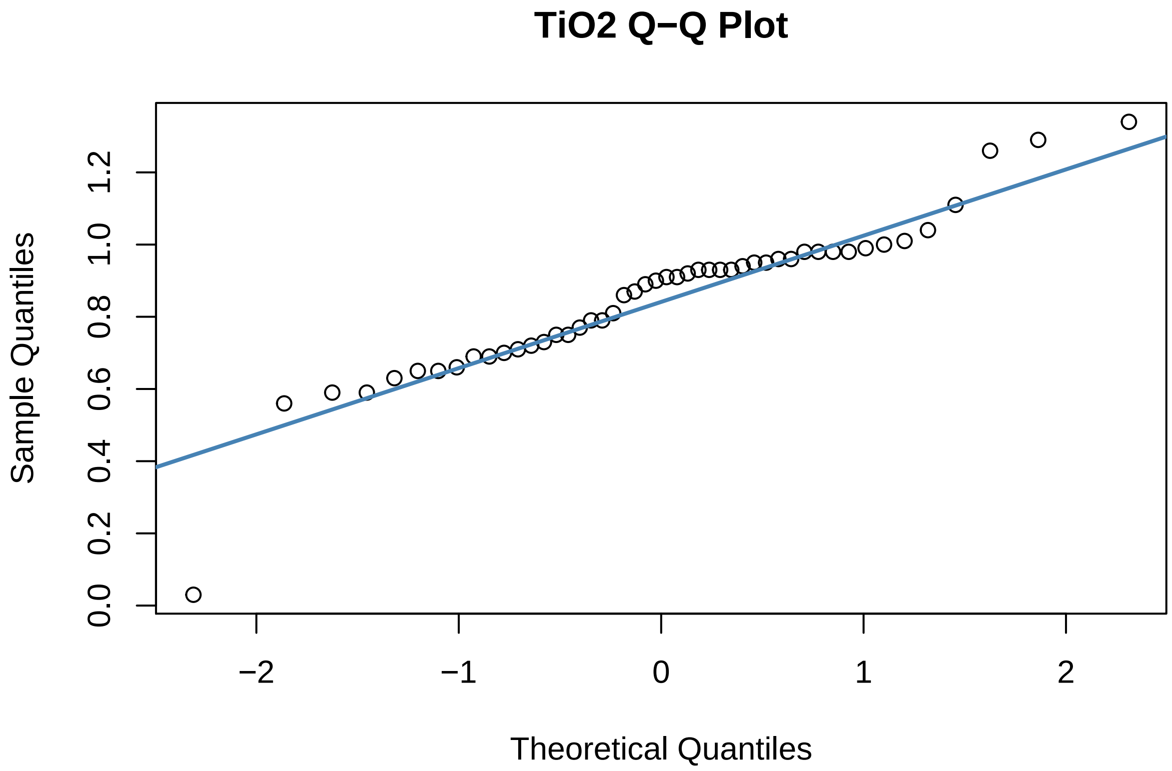
```
qqnorm(pottery$Na2O, pch = 1,main = "Na2O Q-Q Plot")
qqline(pottery$Na2O, col = "steelblue", lwd = 2)
```



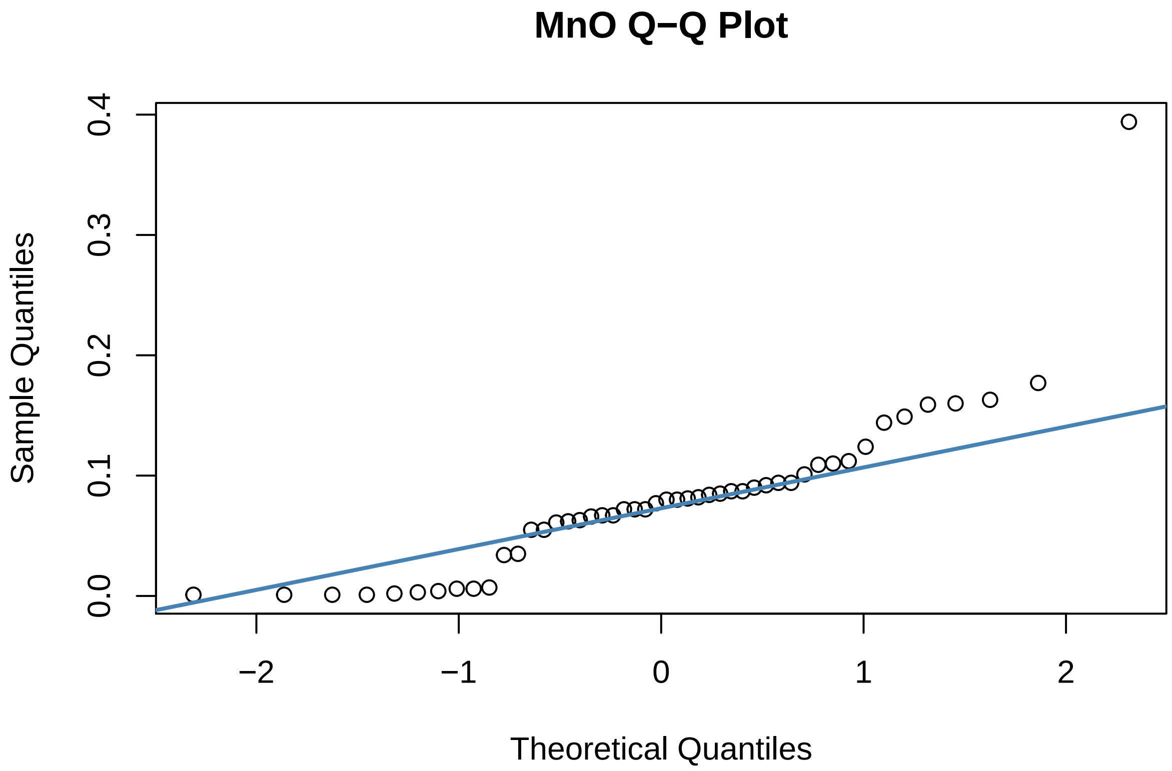
```
qqnorm(pottery$K2O, pch = 1,main = "K2O Q-Q Plot")
qqline(pottery$K2O, col = "steelblue", lwd = 2)
```



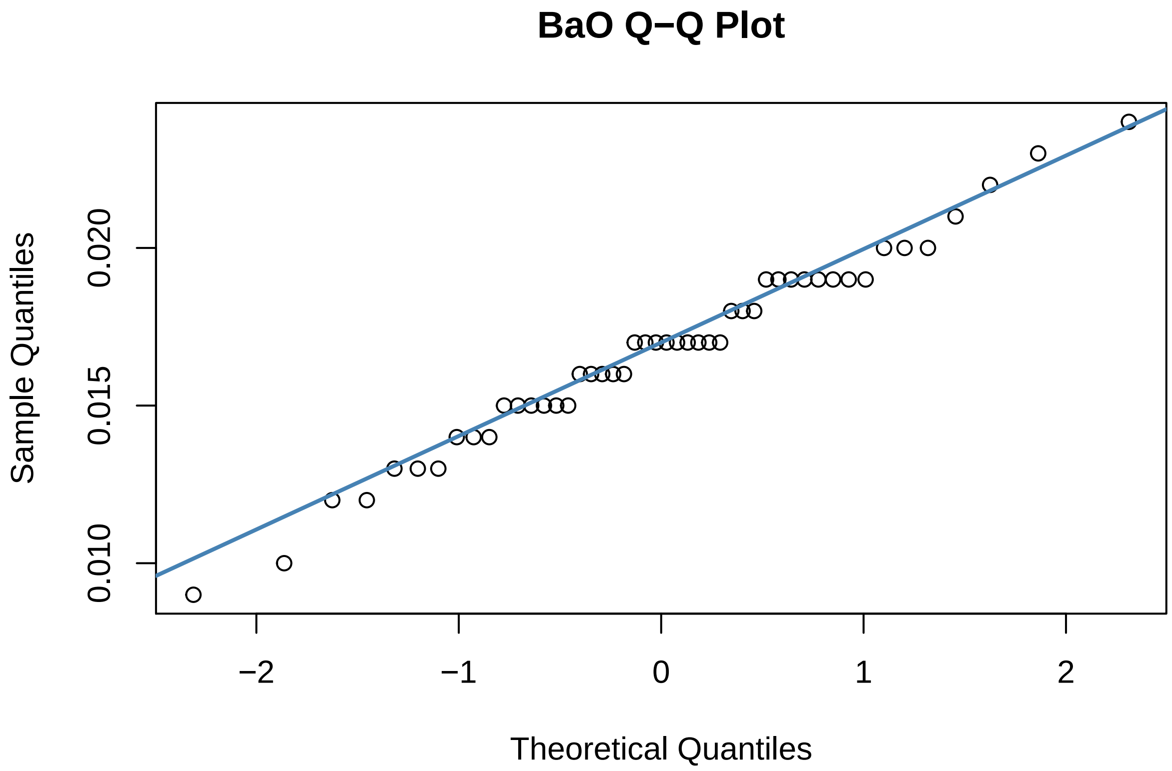
```
qqnorm(pottery$TiO2, pch = 1,main = "TiO2 Q-Q Plot")
qqline(pottery$TiO2, col = "steelblue", lwd = 2)
```



```
qqnorm(pottery$MnO, pch = 1,main = "MnO Q-Q Plot")
qqline(pottery$MnO, col = "steelblue", lwd = 2)
```



```
qqnorm(pottery$BaO, pch = 1,main = "BaO Q-Q Plot")
qqline(pottery$BaO, col = "steelblue", lwd = 2)
```



As the Q-Q plot shown, almost every chemical variables roughly follow a normal distribution, except for the Fe2O3. However, since we only verify this condition by visualization (not a formal test), we could not conclude that the normality assumption is violated; if it is violated, then the model would not be reliable anymore.

Statistical Analysis

calculating E and H In order to calculate the test statistics, we need to calculate the **E** (Error Sum of Squares and Cross Products Matrix) and **H** (Hypothesis Sum of Squares and Cross Products Matrix).

```
# Group: kiln 1
x1 <- pottery[pottery$Kiln==1,-1]
m1 <- colMeans(x1)
n1 <- dim(x1)[1]
# Group: kiln 2
x2 <- pottery[pottery$Kiln==2,-1]
m2 <- colMeans(x2)
n2 <- dim(x2)[1]
# Group: kiln 3
x3 <- pottery[pottery$Kiln==3,-1]
m3 <- colMeans(x3)
n3 <- dim(x3)[1]
# Group: kiln 4
x4 <- pottery[pottery$Kiln==4,-1]
m4 <- colMeans(x4)
n4 <- dim(x4)[1]
```

```r
# Group: kiln 5
x5 <- pottery[pottery$Kiln==5,-1]
m5 <- colMeans(x5)
n5 <- dim(x5)[1]
# Grand Mean
mg <- (m1*n1 + m2*n2 + m3*n3 + m4*n4 + m5*n5)/(n1+n2+n3+n4+n5)
```

Calculate E using:

$$\mathbf{E} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right)\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right)'.$$

```r
ESS <- cov(x1)*(n1-1) + cov(x2)*(n2-1) + cov(x3)*(n3-1) + cov(x4)*(n4-1) + cov(x5)*(n5-1)
ESS
```

```
##             Al2O3       Fe2O3          MgO          CaO          Na2O
## Al2O3 96.20132468 21.11225325  5.506287013 -2.096574026  0.569593506
## Fe2O3 21.11225325 19.88942753  2.157729870 -0.685039740  0.918994935
## MgO    5.50628701  2.15772987 16.303520519  0.274558961  0.090970260
## CaO   -2.09657403 -0.68503974  0.274558961  1.760672078 -0.025830519
## Na2O   0.56959351  0.91899494  0.090970260 -0.025830519  0.735820130
## K2O   10.55401948  4.50978519  5.888079221  0.248701558  0.560279610
## TiO2   0.96768701  1.99152987  0.041040519 -0.120881039  0.062710260
## MnO    0.37119545  0.26490145 -0.131911818  0.009635636  0.059562091
## BaO    0.07495727  0.02567727 -0.007025091  0.004785182  0.004963455
##               K2O         TiO2          MnO           BaO
## Al2O3 10.55401948  0.967687013  0.371195455  0.0749572727
## Fe2O3  4.50978519  1.991529870  0.264901455  0.0256772727
## MgO    5.88807922  0.041040519 -0.131911818 -0.0070250909
## CaO    0.24870156 -0.120881039  0.009635636  0.0047851818
## Na2O   0.56027961  0.062710260  0.059562091  0.0049634545
## K2O   14.63247117  0.321679221  0.104890727  0.0100536364
## TiO2   0.32167922  1.368520519  0.015238182  0.0037669091
## MnO    0.10489073  0.015238182  0.089093964  0.0030718182
## BaO    0.01005364  0.003766909  0.003071818  0.0004249909
```

Calculate H using:

$$\mathbf{H} = \sum_{k=1}^{g} n_k\left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)\left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)'.$$

```r
HSS <- n1*(m1 - mg)%*%t(m1 - mg) + n2*(m2 - mg) %*% t(m2 - mg) + n3*(m3 - mg) %*% t(m3 - mg) +
  n4*(m4 - mg) %*% t(m4 - mg) + n5*(m5 - mg) %*% t(m5 - mg)
HSS
```

```
##                Al2O3        Fe2O3           MgO          CaO         Na2O
##   [1,]  2.470585e+02 -62.83133658 -1.688936e+02 17.329699026  0.163614827
##   [2,] -6.283134e+01 238.85773913  7.165714e+01 32.920489740 12.025288398
##   [3,] -1.688936e+02  71.65713680  1.236503e+02 -8.167158961  1.759763074
##   [4,]  1.732970e+01  32.92048974 -8.167159e+00  7.750252922  1.953805519
##   [5,]  1.636148e-01  12.02528840  1.759763e+00  1.953805519  0.687171537
##   [6,] -6.954646e+01  50.83463981  5.080132e+01  0.919660942  1.596657890
##   [7,]  1.337898e+01  -6.37246320 -9.258374e+00  0.373681039 -0.128076926
##  [8,] -4.777120e+00   3.42203855  3.697632e+00 -0.002320636  0.128522909
##  [9,]  7.832311e-03   0.04981856  9.178424e-03  0.007261068  0.003436962
##                 K2O          TiO2           MnO          BaO
##  [1,] -69.546456981 13.3789796537 -4.7771204545 7.832311e-03
##  [2,]  50.834639805 -6.3724632035  3.4220385455 4.981856e-02
##  [3,]  50.801320779 -9.2583738528  3.6976318182 9.178424e-03
##  [4,]   0.919660942  0.3736810390 -0.0023206364 7.261068e-03
##  [5,]   1.596657890 -0.1280769264  0.1285229091 3.436962e-03
##  [6,]  25.307410081 -4.3025792208  1.6272767727 3.224489e-03
##  [7,]  -4.302579221  0.7823461472 -0.2784881818 2.364242e-04
##  [8,]   1.627276773 -0.2784881818  0.1193510364 6.309318e-04
##  [9,]   0.003224489  0.0002364242  0.0006309318 2.648826e-05
```

Four Test Statistics

- In the table below, we list the calculated test statistics, associated F-statistics, and their p-values.

```r
library(rootWishart)
N <- n1+n2+n3+n4+n5
g <- 5
p <- 9
output <- NULL
```

- **Wilks's Lambda (Ratio of Determinants)**

$$\Lambda = \frac{\det \mathbf{E}}{\det \mathbf{T}} = \frac{\det \mathbf{E}}{\det (\mathbf{E}+\mathbf{H})}.$$

```r
# 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 (Trace of Ratio)**

$$V = \operatorname{tr}\left[\mathbf{H}(\mathbf{H}+\mathbf{E})^{-1}\right],$$

```r
# 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,df2 = pillai_s*(2*pillai_m + pillai_s + 1),
      df1 = pillai_s*(2*pillai_r + pillai_s + 1))))
```

- **Hotelling-Lawley Trace (Trace of Ratio)**

$$U = \text{tr}\left[\mathbf{HE}^{-1}\right],$$

```
# 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's Maximum Root (Largest Eigenvalue of Ratio)**

$$R = \lambda_m \left[\mathbf{HE}^{-1}\right],$$

Because the eigenvalues contains complex numbers, we use Re() function to convert them into real number.

```
# 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             0.001388136     17.6757748 0.000000e+00
## Pillai            2.226843233      5.3025357 7.585866e-08
## Hotelling-Lawley 44.728161807     41.9995682 0.000000e+00
## Roy              28.628211091      0.9662484 2.220446e-16
```

```
# Total output table with four test statistics
colnames(output) <- c("Statistic","Test Statistic","P-value")
rownames(output) <- c("Wilks","Pillai","Hotelling-Lawley","Roy")
output
```

```
##                    Statistic Test Statistic      P-value
## Wilks             0.001388136     17.6757748 0.000000e+00
## Pillai            2.226843233      5.3025357 7.585866e-08
## Hotelling-Lawley 44.728161807     41.9995682 0.000000e+00
## Roy              28.628211091      0.9662484 2.220446e-16
```

Since all p-values are smaller than 0.05, we reject the null and favorite the alternatives – there are at least one mean vector that is significantly different from others.

Conclusion

In this project, I compared the similarities and differences of two sample hotelling test, ANOVA, and MANOVA tests and decides to use Multivariate Analysis of Variance to exam whether there is a significant difference among the 5 group means on each sites for these 9 chemical variables. Since all four test statistics indicated that the p-values (Wilks:0, Pillai:7.585866e-08, Hotelling-Lawley:0, Roy:2.220446e-16) are much smaller than 0.05, we have sufficient evidence to **reject the null hypothesis** and **favorite the alternatives**, which claims that there is at least one significant means difference among the 5 group means for these 9 chemical variables and at least one mean vectors that are not equal to ohters.

However, I do recognize there are a few inefficiencies in my project. For instance, I only utilize Q-Q plot to visualize normality assumption and **did not test the assumption formally**. As a result,even though the normality assumptions for Fe2O3 that seem to be violated, it is not statistically sufficient to conclude the failure of our test; these assumptions, **especially for normality assumption, need to be further validate**.

Also, because some of the **data sizes are too small** (kiln3 sample size:2, kiln4 sample size:5, and kiln5 sample size:5) compared with other populations, it is more likely reject the null hypothesis in our covariance boxM test and lead an inaccurate result when we test our assumptions. Small sample size with large variance would also lead our test more likely to reject the null, therefore we should consider more on the stringent test statistics (Pillai).

Notice: Due to the possible violations on MANOVA assumptions, the test might not be reliable.