# Math 189: Midterm Project 2

Xiangyu Wei

February 14, 2021

## Introduction

The Romano-British Pottery dataset contains measurements on pottery shards that were collected from 5 sites in the British Isles. In this study, I aim to explore whether there is a significant difference among the 5 kiln sites for the 9 chemicals that are reported in the dataset. More specifically, I will first compare the sample means and the histogram distributions to check for preliminary visual differences. Then, I would apply MANOVA and ANOVA to test whether the differences I observe through the means are significant. Testings and discussions on whether the assumptions of MANOVA and ANOVA are met will also be done, since it has impact on the interpretation of the results.

## Tasks & Analysis

### Dataset

The Romano-British Pottery dataset contains 48 observations on 9 chemical variables (Al2O3: aluminium trioxide; Fe2O3: iron trioxide; MgO: magnesium oxide; CaO: calcium oxide; Na2O: natrium oxide; K2O: kalium oxide; TiO2: titanium oxide; MnO: mangan oxide; BaO: barium oxide), along with the Kiln column indicating at which kiln site the pottery was found. The kiln sites come from (1=Gloucester (G), 2=Llanedeyrn (L), 3=Caldicot (C), 4=Islands Thorns (I), 5=Ashley Rails (A)). The numbers are the percentage of metal oxide in the pottery determined by atomic absorption spectrophotometry. The dataset is 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. And I extracted the dataset from https://github.com/tuckermcelroy/ma189/blob/main/Data/RBPottery.csv at 2021-02-13 10:25:22 PST.

#### Load & Clean Dataset

```
# load full dataset
library(tidyverse)
pottery <- read.csv('Data/RBPottery.csv')[,3:12]
# recode kiln sites
pottery$Kiln_sum <- pottery$Kiln
pottery$Kiln <- pottery$Kiln %>%
  recode("1" = "Gloucester (G)", "2" = "Llanedeyrn (L)", "3" = "Caldicot (C)",
         "4" = "Islands Thorns (I)", "5" = "Ashley Rails (A)")
pottery$Kiln_sum <- pottery$Kiln_sum %>%
  recode("1" = "Gloucester", "2" = "Llanedeyrn", "3" = "Caldicot",
         "4" = "Islands Thorns", "5" = "Ashley Rails")
```

```r
# look at selected data
pottery[c(1,23,37,39,44),c(1,2,5,7,8,10)]
```

```
##                  Kiln Al2O3  CaO  K2O TiO2   BaO
## 1      Gloucester (G)  18.8 0.79 3.20 1.01 0.015
## 23     Llanedeyrn (L)  14.4 0.15 4.25 0.79 0.019
## 37       Caldicot (C)  11.6 0.29 4.51 0.56 0.015
## 39 Islands Thorns (I)  18.3 0.03 1.96 0.65 0.014
## 44   Ashley Rails (A)  17.7 0.06 2.06 0.79 0.013
```

## Method

To preliminarily check whether the means differ across 5 kiln sites for each chemicals, I calculate the sample means and compare the distributions (using histogram and boxplot) of each percentage chemical compounds for each kiln site. Then, in order to see whether any pair of sites differ significantly on any chemicals, I justify the assumptions and use a MANOVA. From the results of MANOVA, I want to see for which chemical(s) do at least one pair of sites differ significantly, I use univariate ANOVA and conduct 9 different univariate tests. Assumptions of ANOVA are also fully discussed.

## Analysis: Significant difference among the 5 group means for 9 variables?

Our main analysis is to see if there is a significant difference for the 9 different chemicals among the 5 different kiln sites. In order to do so, we will apply the above method to check the relationship.

### Preliminary Check

**Sample Mean** The simplest way to compare the group mean difference is to calculate the sample mean directly among the 5 kiln sites for the 9 different chemical compounds. In order to do so, I group each chemical by the kiln sites using `group_by()` for `mean()` calculation.

```r
pottery %>%
  group_by(Kiln_sum) %>%
  summarise_if(is.numeric, .funs = c(mean="mean")) %>%
  pivot_longer(cols = -Kiln_sum) %>%
  pivot_wider(names_from = Kiln_sum, values_from = value) %>%
  column_to_rownames("name")
```

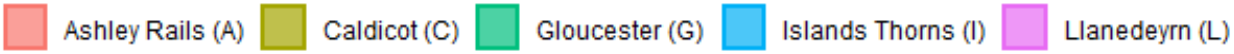
```
##            Ashley Rails Caldicot  Gloucester Islands Thorns Llanedeyrn
## Al2O3_mean      17.3200  11.7000 16.94090909        18.1800 12.5642857
## Fe2O3_mean       1.5120   5.4150  7.43090909         1.7120  6.3721429
## MgO_mean         0.6060   3.8550  1.83636364         0.6740  4.8264286
## CaO_mean         0.0520   0.2950  0.94227273         0.0260  0.2021429
## Na2O_mean        0.0480   0.0500  0.34818182         0.0540  0.2507143
## K2O_mean         1.9660   4.5750  3.10545455         2.0760  3.9278571
## TiO2_mean        0.9940   0.5750  0.89636364         1.0460  0.7064286
## MnO_mean         0.0042   0.0975  0.07172727         0.0022  0.1445000
## BaO_mean         0.0156   0.0140  0.01713636         0.0164  0.0170000
```

Each row here represents each chemical compound mean, and each column represents each kiln site. Seen from the summary above, the means of each chemical seem to differ across those 5 kiln sites. More specifically, the means seem to differ a lot for Al2O3, Fe2O3, MgO, CaO, & K2O; and differ somewhat for Na2O, TiO2, MnO, & BaO.

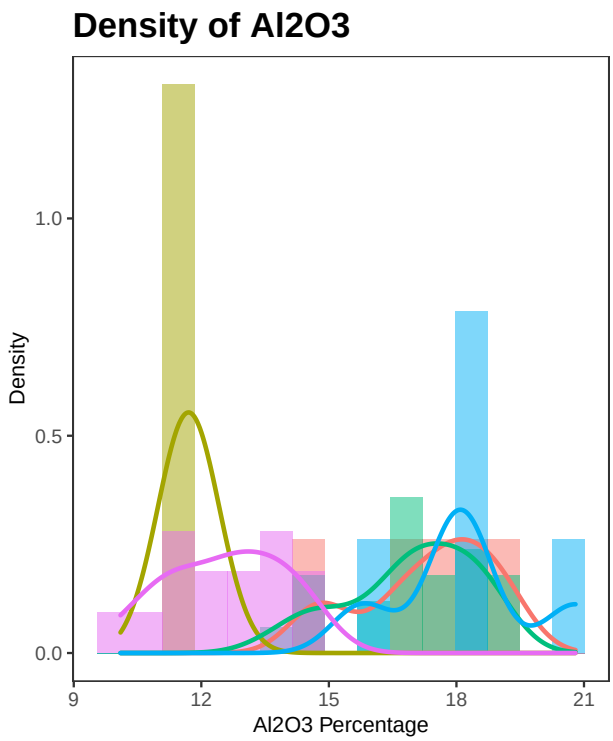
**Histogram & Boxplot** To better visualize the differences, it's best that we plot the histogram & boxplot for each chemical compound by 5 different kiln sites. I choose histogram (with density) because it would better tell us the central tendency and spread of each distribution, and I choose boxplot because it would make is easier for us to compare the central tendencies (note that only medians are returned here in the graph). The following is the function used to plot histograms and boxplots.

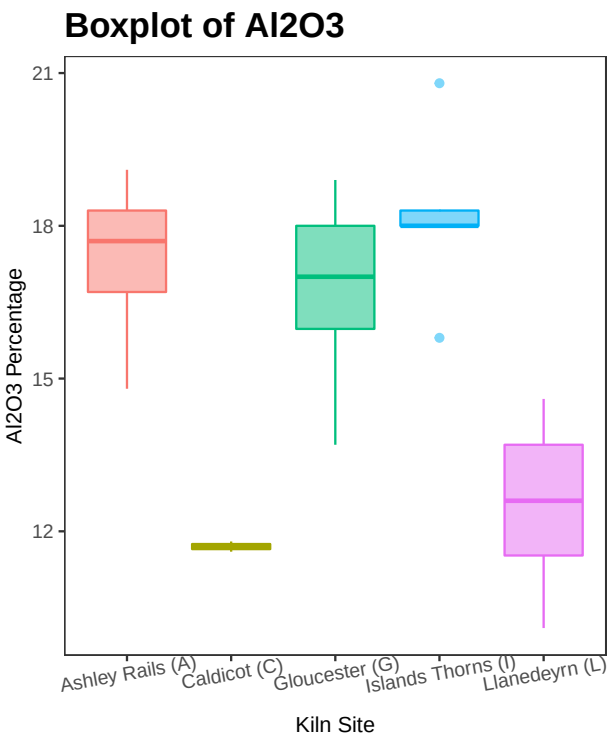
```
graphing <- function(colName){
  bin_width = round(diff(range(pottery[[colName]]))/15,5)
  # plot histogram with density
  hist <- pottery %>%
    ggplot(aes(x = .data[[colName]])) +
    geom_histogram(aes(y=..density.., fill = Kiln),
                   position = 'identity', bins = 15, alpha = 0.5) +
    geom_density(aes(color = Kiln), position = 'identity',
                 bw = bin_width, size = 1, alpha = 0.4)+
    labs(
      x = paste(colName, "Percentage"),
      y = "Density",
      title = paste("Density of", colName)
    ) +
    theme_bw() +
    theme(legend.position = "none",
          panel.grid = element_blank(),
          plot.title = element_text(size=16, face='bold'),
          axis.title = element_text(size=10))

  # plot boxplot
  box <- pottery %>%
    ggplot(aes(x = Kiln, y = .data[[colName]],
               color = Kiln, fill = Kiln)) +
    geom_boxplot(alpha = 0.5) +
    labs(
      x = "Kiln Site",
      y = paste(colName, "Percentage"),
      title = paste("Boxplot of", colName)
    ) +
    theme_bw() +
    theme(legend.position = "none",
          panel.grid = element_blank(),
          plot.title = element_text(size=16, face='bold'),
          axis.title = element_text(size=10),
          axis.text.x = element_text(angle = 10))

  # return the graphs
  gridExtra::grid.arrange(hist, box, nrow = 1)
}
```

Each color below in graphs represents a kiln site as shown in the picture below.

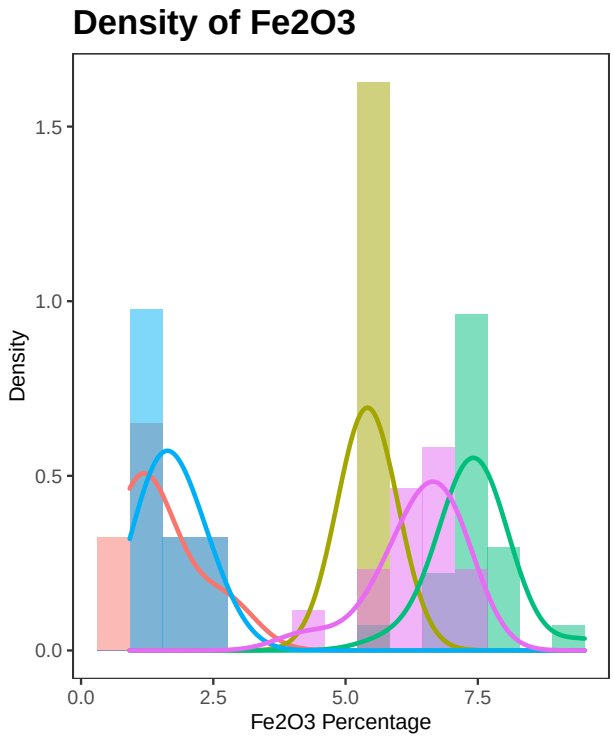


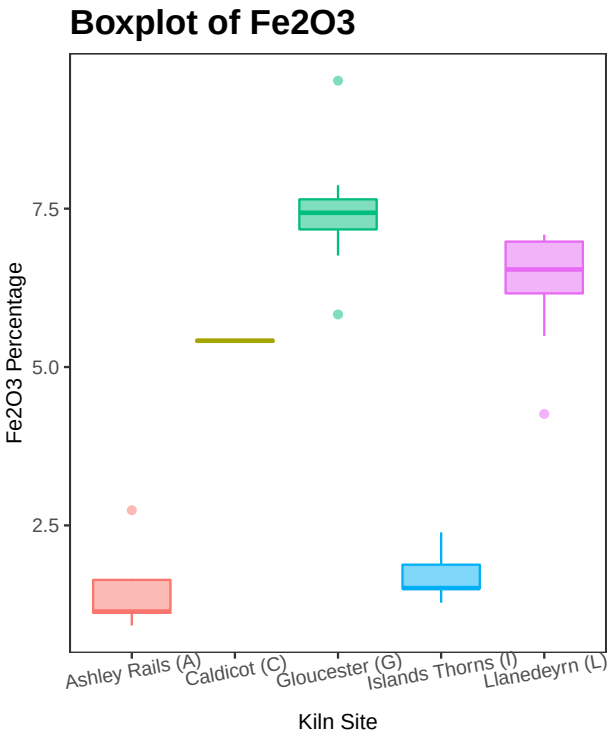
```
# histogram & boxplot for Al2O3
graphing("Al2O3")
```



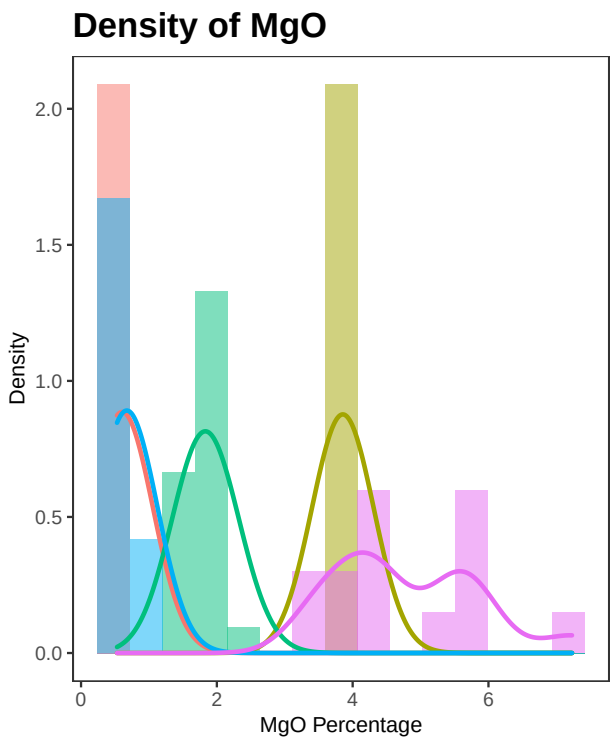



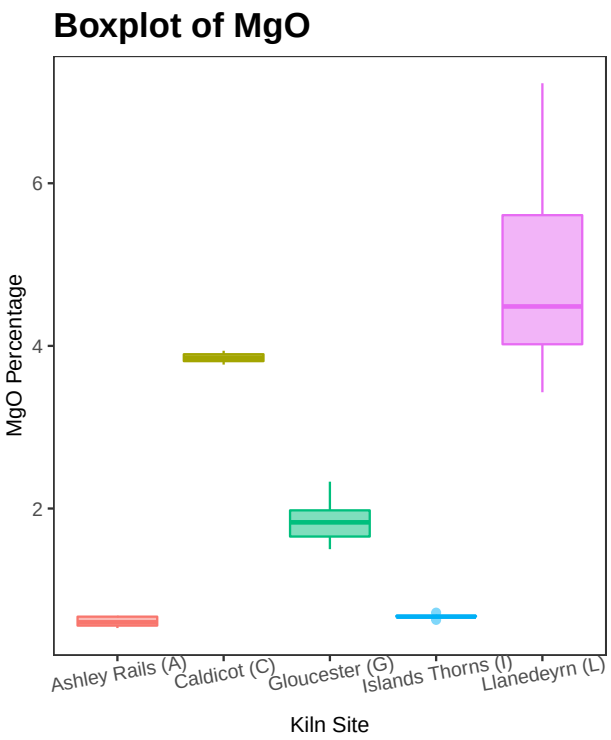

```
# histogram & boxplot for Fe2O3
graphing("Fe2O3")
```



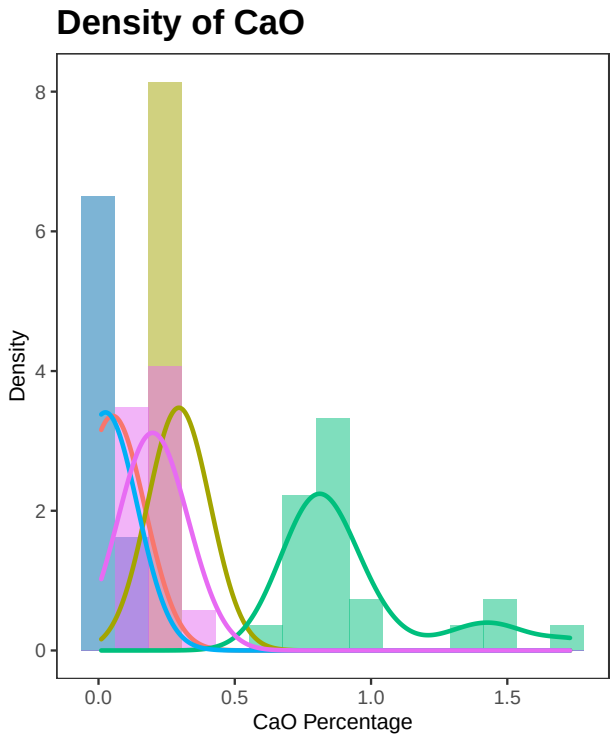



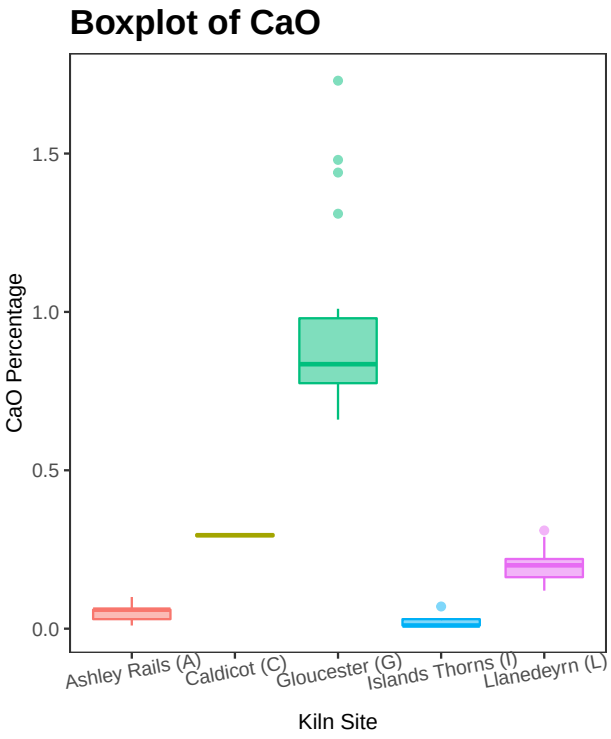
```
# histogram & boxplot for MgO
graphing("MgO")
```



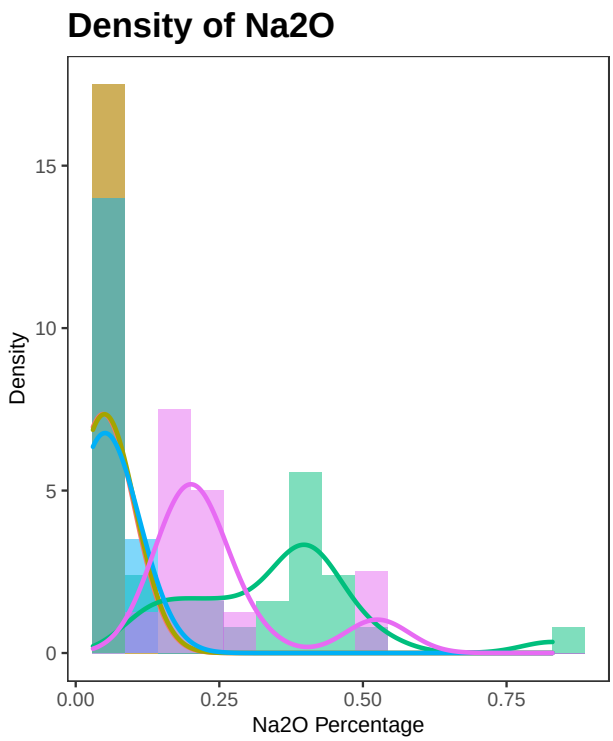



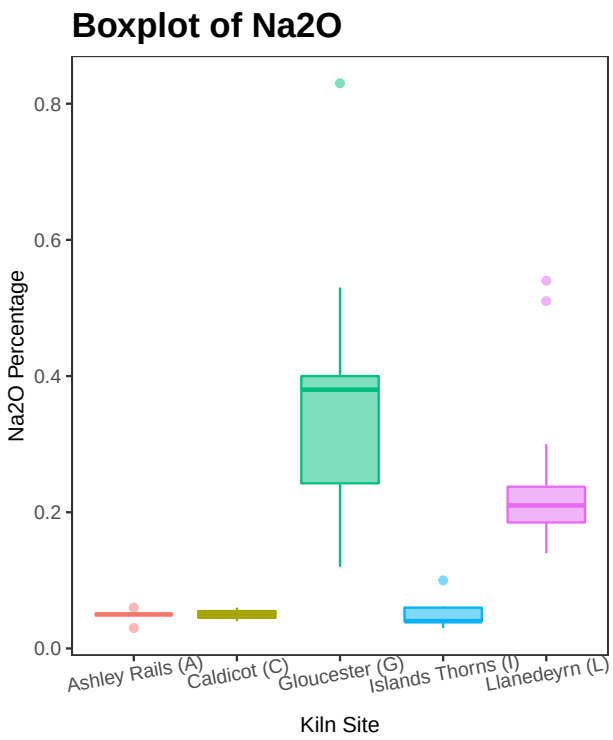

```
# histogram & boxplot for CaO
graphing("CaO")
```



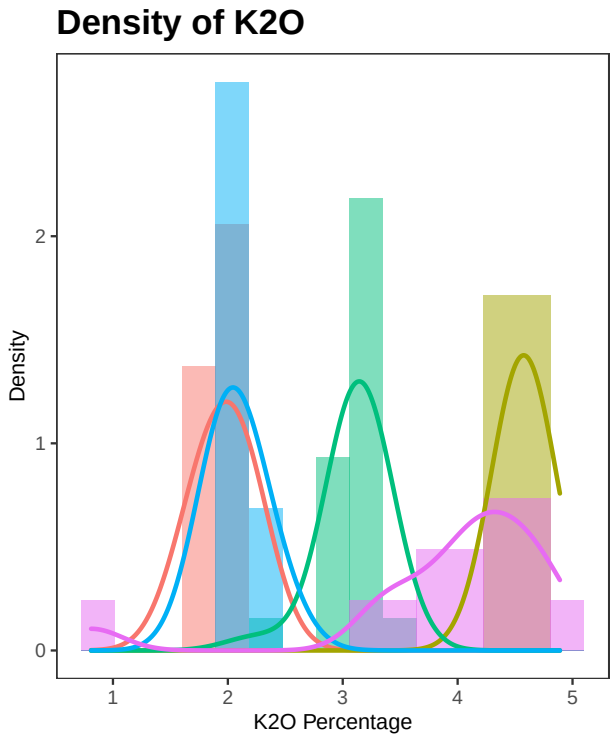



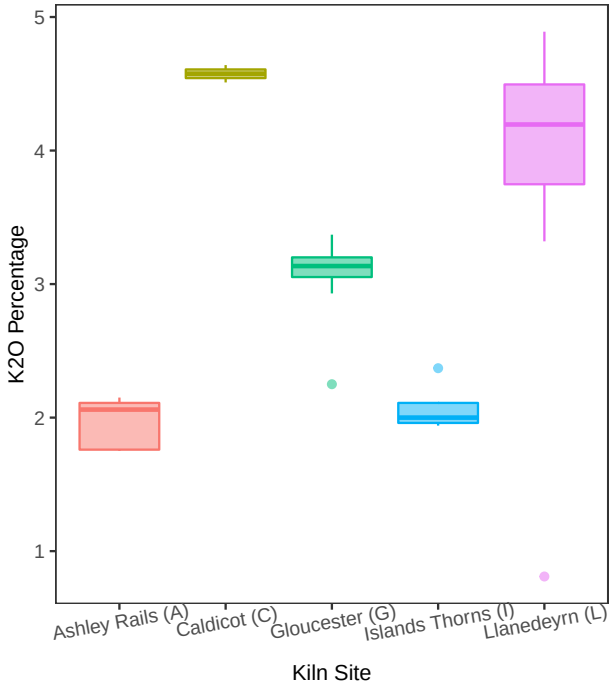
```
# histogram & boxplot for Na2O
graphing("Na2O")
```



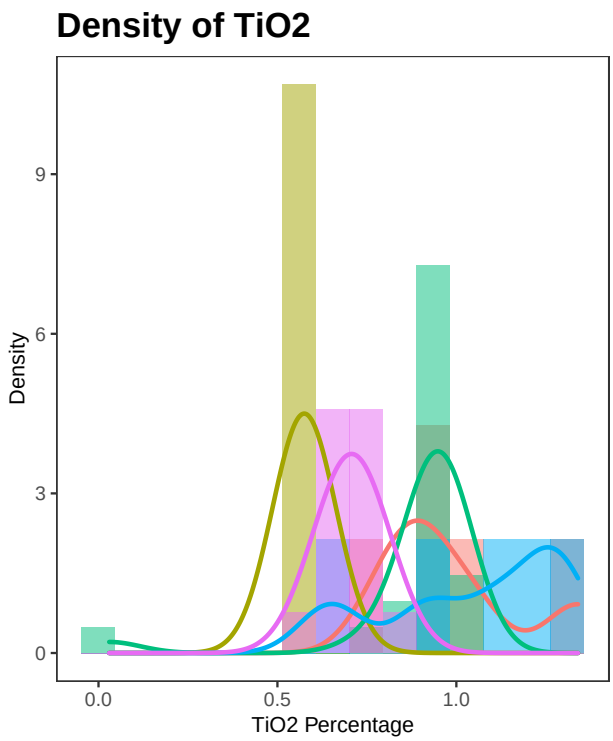



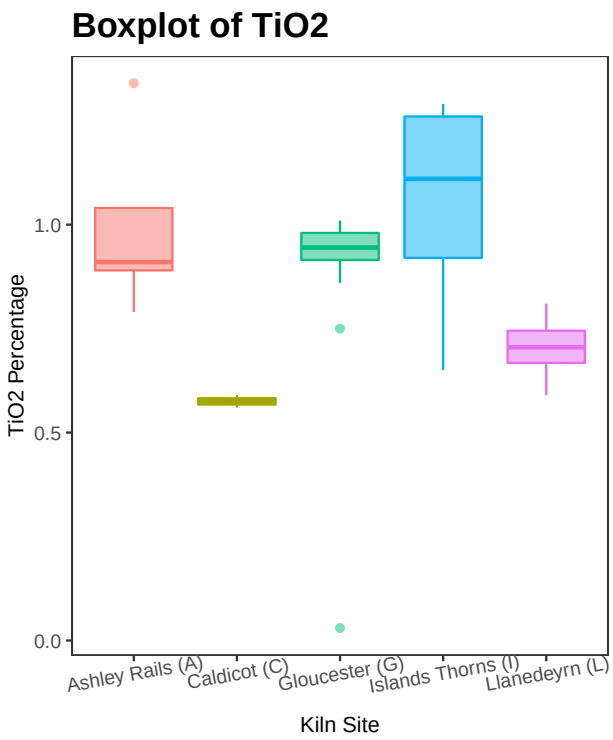

```
# histogram & boxplot for K2O
graphing("K2O")
```



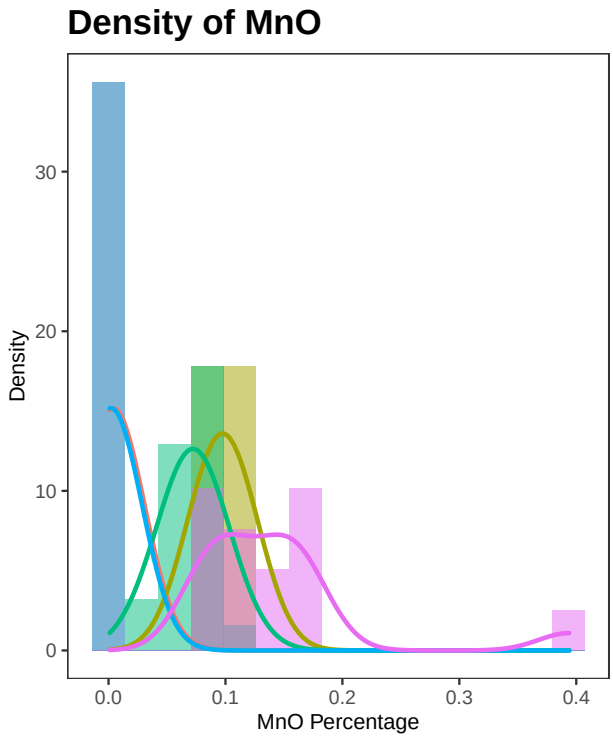



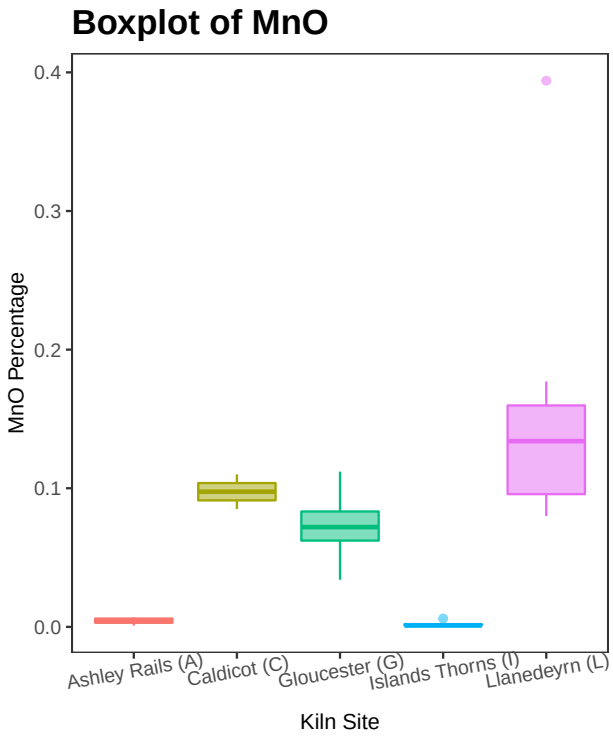
```
# histogram & boxplot for TiO2
graphing("TiO2")
```



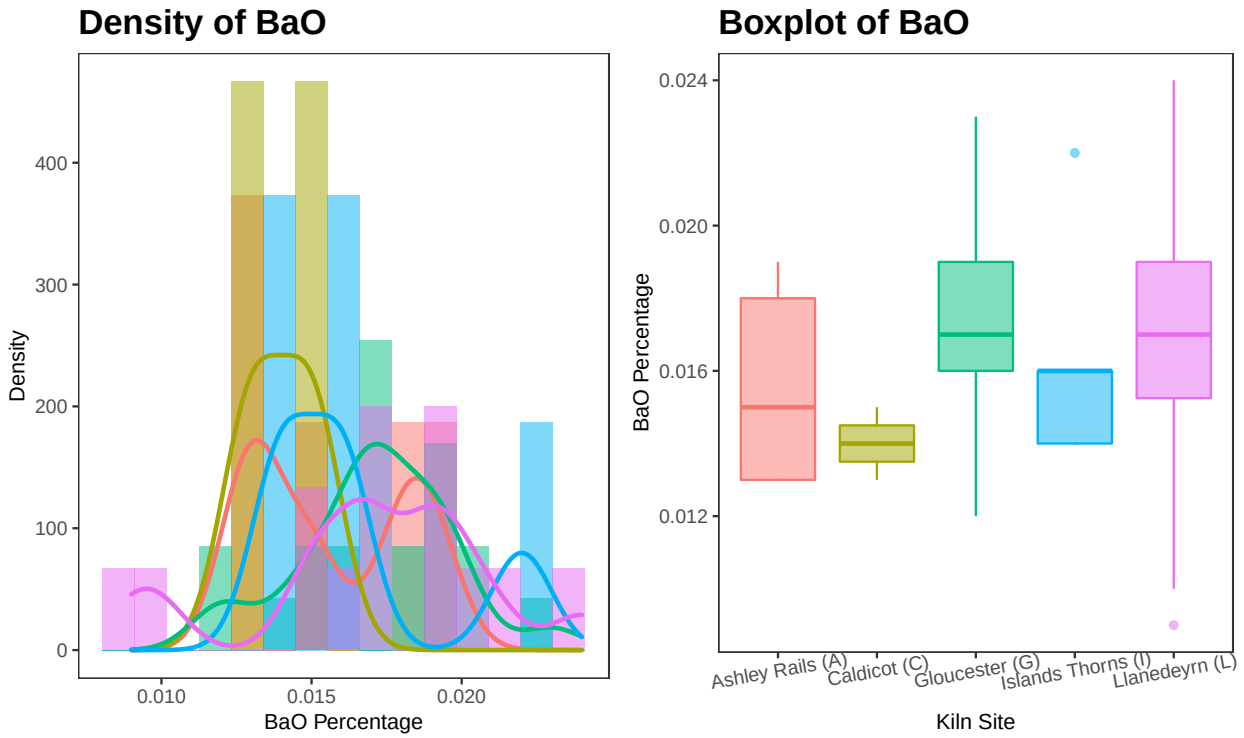




```
# histogram & boxplot for MnO
graphing("MnO")
```



```
# histogram & boxplot for BaO
graphing("BaO")
```




From the histograms and boxplots above, it seems to tell us the same conclusion that the means differ a lot for Al2O3, Fe2O3, MgO, CaO, & K2O; and differ somewhat for Na2O, TiO2, MnO, & BaO. More specifically, for Al2O3, site A G I seem to group together while site C L seem to cluster. For Fe2O3, site C G L seem to group together while site A I seem to cluster. For MgO, site A I seem to group together, site C L seem to cluster, but site G seems to be on its own. For CaO, site A C I L seem to cluster while site G seems to be on its own. For Na2O, site A C I seem to group together while site G and site L seem to be on its own separately. For K2O, site A I seem to group together, site C L seem to cluster, but site G seem to be on its own (like MgO). For TiO2, site A G I seem to group together while site C L seem to cluster (like AL2O3, only that the differences are not as big). For MnO, site A I seem to group together while site C G L seem to cluster. For BaO, site A C G I L seem to be around the same percentage.

### MANOVA Analysis

To see whether the differences we observe above is significant, I first decide to perform a Multivariate ANOVA (MANOVA) for all 9 different chemical compounds altogether. The null hypothesis and the alternative hypothesis for the MANOVA are as follows:

$$H_0 : \underline{\mu}^{(1)} = \ldots = \underline{\mu}^{(5)} \quad \text{versus} \quad H_a : \mu_j^{(k)} \neq \mu_j^{(h)}$$

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

**MANOVA Assumptions** Before we move on to perform MANOVA, I want to see if our dataset met the assumptions of MANOVA model. The four assumptions of MANOVA are as follows:

1. The data from group $k$ has common mean vector $\underline{\mu}^{(k)}$.

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

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

   for any record $i$, and the matrix does not depend on $k$ (the group index).
3. Independence: The observations are independently sampled.
4. Normality: The data are multivariate normally distributed.

I assume that criterion 1 (common mean) and criterion 3 (independence) are met based on the description of the dataset, where the percentage metal oxide in pottery specimens were collected using atomic absorption spectrophotometry (*leading to independent data collection*) on 5 different kiln site (*leading to common mean for each group*).

**Normality:** To test for criterion 4 (multivariate normality), I use the multivariate Shapiro-Wilk's method `mshapiro_test()`.

```
pottery %>%
  select(Al2O3, Fe2O3, MgO, CaO, Na2O, K2O, TiO2, MnO, BaO) %>%
  rstatix::mshapiro_test()
```

```
## # A tibble: 1 x 2
##   statistic  p.value
##       <dbl>    <dbl>
## 1     0.529 3.50e-11
```

The multivariate Shapiro-Wilk's test for normality above ($p < 0.001$) suggests that the data are NOT multivariate normally distributed. Even though I would still continue to conduct the MANOVA because MANOVA is relatively robust to not normal data, I would need to keep in mind that the data is not multivariate normal, and the results might not be as reliable.

**Homoskedasticity:** Note that since the data failed the multivariate normality check, the statistical testing for homoskedasticity may not be reliable anymore since homogeneity tests for covariances (e.g. Box's M-test) are based on the fact that the data is multivariate normally distributed. Thus, I only assumed that the covariance would be homogeneous across each group just to continue my analysis. But I would keep in mind that the dataset might have heteroskedastic problem, thus rendering an unreliable result.

### MANOVA Statistics

After checking the assumptions, I would get started to test the hypothesis using MANOVA.

```
# Group: Gloucester (G)
x1 <- pottery[pottery$Kiln == "Gloucester (G)",2:10]
m1 <- colMeans(x1)
n1 <- dim(x1)[1]
# Group: kiln 2
x2 <- pottery[pottery$Kiln == "Llanedeyrn (L)",2:10]
m2 <- colMeans(x2)
n2 <- dim(x2)[1]
# Group: kiln 3
x3 <- pottery[pottery$Kiln == "Caldicot (C)",2:10]
m3 <- colMeans(x3)
n3 <- dim(x3)[1]
# Group: kiln 4
```

```r
x4 <- pottery[pottery$Kiln == "Islands Thorns (I)",2:10]
m4 <- colMeans(x4)
n4 <- dim(x4)[1]
# Group: kiln 5
x5 <- pottery[pottery$Kiln == "Ashley Rails (A)",2:10]
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)
# Error sum of squares
ESS <- cov(x1)*(n1-1) + cov(x2)*(n2-1) +
  cov(x3)*(n3-1)+ cov(x4)*(n4-1) + cov(x5)*(n5-1)
# Hypothesis sum of squares
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)

# MANOVA calculations
N <- n1+n2+n3+n4+n5
g <- 5 # 5 kiln sites
p <- 9 # 9 chemicals
MANOVA.df <- data.frame(name = c("Wilks", "Pillai", "Hotelling-Lawley", "Roy"),
                        statistic = rep(NA, 4), test.statistic = rep(NA, 4),
                        p.value = rep(NA, 4))

# 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))
wilks_pval <- 1 - pf(wilks_stat, df1 = p*(g-1), df2 = wilk_f*wilk_xi - wilk_omega)
MANOVA.df[MANOVA.df$name == "Wilks",2:4] <- c(wilks, wilks_stat, wilks_pval)

# 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))
pillai_pval <- 1 - pf(pillai_stat,df1 = pillai_s*(2*pillai_m + pillai_s + 1),
                      df2 = pillai_s*(2*pillai_r + pillai_s + 1))
MANOVA.df[MANOVA.df$name == "Pillai",2:4] <- c(pillai, pillai_stat, pillai_pval)

# Hotelling-Lawley
hotel <- sum(diag(HSS %*% solve(ESS)))
hotel_b <- (N-p-2)*(N-g-1)/((N-g-p-3)*(N-g-p))
```

```r
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
hotel_pval <- 1 - pf(hotel_stat, df1 = hotel_df1, df2 = hotel_df2)
MANOVA.df[MANOVA.df$name == "Hotelling-Lawley",2:4] <- c(hotel, hotel_stat, hotel_pval)

# Roy
roy <- max(Re(eigen(HSS %*% solve(ESS))$values))
roy_stat <- roy/(1+roy)
roy_pval <- 1 - rootWishart::doubleWishart(roy_stat,p=p,m=N-g,n=g-1)
MANOVA.df[MANOVA.df$name == "Roy",2:4] <- c(roy, roy_stat, roy_pval)

# print MANOVA table
MANOVA.df
```

```
##                name    statistic test.statistic      p.value
## 1             Wilks  0.001388136     17.6757748 0.000000e+00
## 2            Pillai  2.226843233      5.3025357 1.391109e-13
## 3 Hotelling-Lawley 44.728161807     41.9995682 0.000000e+00
## 4               Roy 28.628211091      0.9662484 2.220446e-16
```

The four MANOVA methods: Wilks's Lambda, Pillai's Trace, Hotelling-Lawley Trace, & Roy's Maximum Root, all have p-values smaller than 0.001. Thus, the analysis imply that we would reject the null hypothesis and leaning towards the alternative that there is at least one pair of kiln sites that is different on at least one chemical.

### ANOVA Analysis

Since MANOVA shows that there is at least one pair of sites that is different on at least on chemical compound, to see for which chemical compound(s) do at least one pair of kiln sites differ significantly, I then decide to perform a univariate ANOVA for 9 different chemical compounds separately. The null hypothesis and the alternative hypothesis for each univariate ANOVA are as follows:

$$H_0 : \mu^{(1)} = \ldots = \mu^{(5)} \quad H_a : \mu^{(i)} \neq \mu^{(j)} \text{ for some } i \neq j.$$

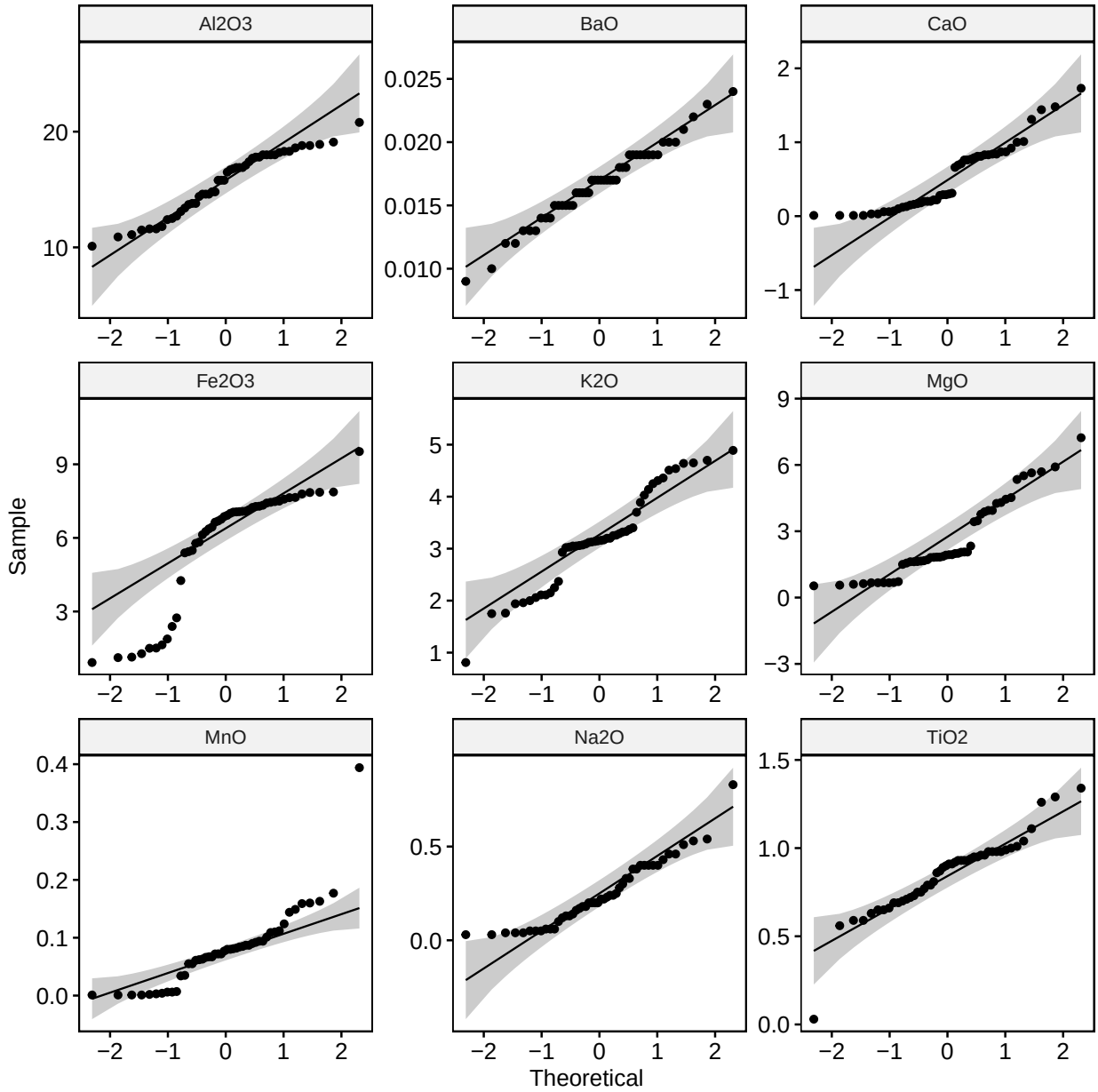
**ANOVA Assumptions** Before we move on to perform 9 univariate ANOVAs, I want to see if our dataset met the assumptions of ANOVA model. The four assumptions of ANOVA are as follows:

1. The data from group $k$ has common mean $\mu_k$.
2. Homoskedasticity: the data from all groups have common variance $\sigma^2$.
3. Independence: the observations are independently sampled.
4. Normality: the data are normally distributed.

Again, for the same reason as MANOVA, I assume that criterion 1 (common mean) and criterion 3 (independence) are met based on the description of the dataset.

**Normality:** To test for criterion 4 (univariate normality), I plot the Q-Q plot using `ggqqplot()` from `ggpubr` package, and test the normality using the Shapiro-Wilk's method `shapiro.test()`. Since we are doing multiple univariate testings, I also corrected the p-values using Bonferroni correction, which means the corrected significant level $\alpha_{correct} = \frac{0.05}{9} \approx 0.0056$.

```r
# corrected alpha level
alpha <- 0.05
bonf.alpha <- alpha / 9

# pivotted dataframe for plotting
pottery_long <- pottery[-11] %>%
  pivot_longer(c(Al2O3, Fe2O3, MgO, CaO, Na2O, K2O, TiO2, MnO, BaO),
               names_to = "chemicals", values_to = "percentage")

# QQ plot
ggpubr::ggqqplot(pottery_long, x = "percentage", facet.by = c("chemicals"), scales = "free")
```



```r
# Shapiro-Wilk's method
pottery_long[-1] %>%
  group_by(chemicals) %>%
  summarise_all(.funs = funs(statistic = shapiro.test(.)$statistic,
                             p.value = shapiro.test(.)$p.value,
                             significant = p.value < bonf.alpha))
```

```
## # A tibble: 9 x 4
##   chemicals statistic    p.value significant
##   <chr>         <dbl>      <dbl> <lgl>
## 1 Al2O3         0.948 0.0322     FALSE
## 2 BaO           0.983 0.716      FALSE
## 3 CaO           0.881 0.000166   TRUE
## 4 Fe2O3         0.799 0.00000121 TRUE
## 5 K2O           0.951 0.0445     FALSE
## 6 MgO           0.880 0.000150   TRUE
## 7 MnO           0.805 0.00000168 TRUE
## 8 Na2O          0.920 0.00300    TRUE
## 9 TiO2          0.915 0.00199    TRUE
```

From the Q-Q plot we can see that Al2O3, BaO, Na2O, TiO2 seem to be normally distributed as the pattern seem to fall along a straight line. However, from the Shapiro-Wilk's test for normality, after Bonferroni correction, we can only see that Al2O3 ($p = 0.0322$), BaO ($p = 0.716$), & K2O ($p = 0.0445$) are not significant, implying that they are normally distributed.

Even though I would still conduct the 9 univariate ANOVA tests because ANOVA is also relatively robust to not normal data, I would only be confident about the results I get from Al2O3, Bao, & K2O (potentially also Na2O & TiO2) since these are the only chemicals that pass the normality check. I would keep my mind of other chemicals that do not pass the normality check while interpreting the ANVOA results, thus might not be as reliable.

**Homoskedasticity:** To test for criterion 2 (homogeneity for variance), I perform the studentized Breusch-Pagan (BP) test using `bptest()` from `lmtest` package. Studentized BP test is performed because the asymptotic power of the original BT test is extremely sensitive to the skewness of the distribution of the data, while the studentized BP test is more robust since it would give asymptotically correct significance levels [1]. Since we are doing multiple univariate testings, I also corrected the p-values using Bonferroni correction, which means the corrected significant level $\alpha_{correct} = \frac{0.05}{9} \approx 0.0056$.

```r
# create output dataframe
homosked.anova <- data.frame(chemicals = c("Al2O3", "Fe2O3", "MgO", "CaO",
                                           "Na2O", "K2O", "TiO2", "MnO", "BaO"),
                             BP.pvalue = rep(NA, 9), BP.sig = rep(NA, 9))

# loop through to create model, and get NCV test & BP test p.values
for (ind in 1:9){
  chem <- homosked.anova$chemicals[ind]
  mdl <- lm(data = pottery, as.formula(paste(chem, "~", "Kiln")))
  homosked.anova$BP.pvalue[ind] <- lmtest::bptest(mdl)$p.value[[1]]
  homosked.anova$BP.sig[ind] <- lmtest::bptest(mdl)$p.value[[1]] < bonf.alpha
}
homosked.anova
```

[1] Koenker, Roger. "A note on studentizing a test for heteroscedasticity." Journal of econometrics 17.1 (1981): 107-112.

```
##   chemicals  BP.pvalue BP.sig
## 1     Al2O3 0.74473084  FALSE
## 2     Fe2O3 0.88038077  FALSE
## 3       MgO 0.00756616  FALSE
## 4       CaO 0.16275480  FALSE
## 5      Na2O 0.49875921  FALSE
## 6       K2O 0.36668084  FALSE
## 7      TiO2 0.84166242  FALSE
## 8       MnO 0.40893856  FALSE
## 9       BaO 0.41797001  FALSE
```

According to the studentized BP test, all chemical compounds show homoskedasticity. Combinging the tests for normality, I would then be more confident when interpreting the ANOVA output for Al2O3, BaO, & K20, than for other chemicals since they violate the normality assumption.

### ANOVA Statistics

After checking the assumptions, I would get started to test the hypothesis using ANOVA. Since we are doing multiple univariate testings, I also corrected the p-values using Bonferroni correction, which means the corrected significant level $\alpha_{correct} = \frac{0.05}{9} \approx 0.0056$. The following is a function used to calculate the univariate ANOVA statistics for each chemicals.

```
anova.stats <- function(colName){
  x <- pottery[[colName]]
  n1 <- length(x[pottery$Kiln == "Gloucester (G)"])
  n2 <- length(x[pottery$Kiln == "Llanedeyrn (L)"])
  n3 <- length(x[pottery$Kiln == "Caldicot (C)"])
  n4 <- length(x[pottery$Kiln == "Islands Thorns (I)"])
  n5 <- length(x[pottery$Kiln == "Ashley Rails (A)"])
  N <- n1+n2+n3+n4+n5 # total number
  m1 <- mean(x[pottery$Kiln == "Gloucester (G)"])
  m2 <- mean(x[pottery$Kiln == "Llanedeyrn (L)"])
  m3 <- mean(x[pottery$Kiln == "Caldicot (C)"])
  m4 <- mean(x[pottery$Kiln == "Islands Thorns (I)"])
  m5 <- mean(x[pottery$Kiln == "Ashley Rails (A)"])
  mg <- mean(x) # grand mean
  ESS <- sum((x[pottery$Kiln == "Gloucester (G)"] - m1)^2) +
    sum((x[pottery$Kiln == "Llanedeyrn (L)"] - m2)^2) +
    sum((x[pottery$Kiln == "Caldicot (C)"] - m3)^2) +
    sum((x[pottery$Kiln == "Islands Thorns (I)"] - m4)^2) +
    sum((x[pottery$Kiln == "Ashley Rails (A)"] - m5)^2) # error SS
  TSS <- n1*(m1-mg)^2 + n2*(m2-mg)^2 + n3*(m3-mg)^2 +
    n4*(m4-mg)^2 + n5*(m5-mg)^2 # total SS
  F.stats <- TSS/ESS * (N-5)/(5-1) # F stats
  p.val <- pf(F.stats, df1 = 5-1, df2 = N-5, lower.tail = F) # p.val
  list("df1"=5-1, "df2"=N-5, "F stats"=F.stats,
       "p value"=p.val, "significance"=p.val<bonf.alpha)
}
```

Then I loop through all 9 chemicals to calculate the ANOVA statistics.

```r
# create output dataframe
ANOVA.df <- data.frame(chemicals = c("Al2O3", "Fe2O3", "MgO", "CaO",
                                     "Na2O", "K2O", "TiO2", "MnO", "BaO"),
                       df1 = rep(NA, 9), df2 = rep(NA, 9), `F stats` = rep(NA, 9),
                       `p value` = rep(NA, 9), `significance` = rep(NA, 9))

# loop through to get ANOVA statistic
for (chem in ANOVA.df$chemicals){
  ANOVA.df[ANOVA.df$chemicals == chem,2:6] <- anova.stats(chem)
}
ANOVA.df
```

```
##   chemicals df1 df2     F.stats      p.value significance
## 1     Al2O3   4  43  27.6075047 2.184291e-11         TRUE
## 2     Fe2O3   4  43 129.0997788 2.304646e-23         TRUE
## 3       MgO   4  43  81.5309319 1.683857e-19         TRUE
## 4       CaO   4  43  47.3201228 3.295616e-15         TRUE
## 5      Na2O   4  43  10.0392660 7.906371e-06         TRUE
## 6       K2O   4  43  18.5925299 6.153570e-09         TRUE
## 7      TiO2   4  43   6.1454841 5.292941e-04         TRUE
## 8       MnO   4  43  14.4007920 1.540328e-07         TRUE
## 9       BaO   4  43   0.6700114 6.163530e-01        FALSE
```

From the above analysis, we can see that all chemicals (Al2O3 ($F(4, 43) = 27.6075, p < 0.001$), Fe2O3 ($F(4, 43) = 129.0998, p < 0.001$), MgO ($F(4, 43) = 81.5309, p < 0.001$), CaO ($F(4, 43) = 47.3201, p < 0.001$), Na2O ($F(4, 43) = 10.0393, p < 0.001$), K2O ($F(4, 43) = 18.5925, p < 0.001$), TiO2 ($F(4, 43) = 6.1455, p < 0.001$), MnO ($F(4, 43) = 14.4008, p < 0.001$)) except BaO ($F(4, 43) = 0.67, p = 0.6164$). This means that for all chemicals except BaO, the percentage of each chemical compound differ significantly among the 5 kiln sites for at least one pair of sites.

## Conclusion

Assuming that criterion 1 (common mean), criterion 2 (covariance homoskedasticity) and criterion 3 (independence) are met for MANOVA, I conducted a multivariate normality check and found that the dataset failed the criterion of being normally distributed. Keeping this in mind, the analysis of MANOVA tells me that there's at least one pair of kiln sites that differ significantly for at least one chemical compounds.

Assuming that criterion 1 (common mean) and criterion 3 (independence) are met for ANOVA, I conducted a univariate normality check and a variance homoskedasticity check. I found that only Al2O3 ($p = 0.0322$), BaO ($p = 0.716$) and K2O ($p = 0.0445$) passed the normality check, while all the chemical compounds pass the homoskedasticity check. Keeping this in mind, the analysis of ANOVA tells me that for all chemicals except BaO, the percentage of each chemical differ significantly for at least one pair of kiln sites.

## Discussion

The ANOVA analysis makes sense after seeing the results for MANOVA because having 8 out of 9 chemicals significantly differ for at least one pair of sites (in ANOVA) would produce a result like the output of the MANOVA analysis. Also, for most of the time, ANOVA and MANOVA analysis are pretty robust to outliers or violations of homoskedasticity, which makes the results more convincing.

However, I have to point out that the sample size of the dataset is really small (n = 48), and that the number of observations are imbalanced among the 5 groups. Group 1 (Gloucester) has 22 observations,

which is almost 50% of the data, while Group 3 (Caldicot) only has 2 observations. Thus, even given the robustness of ANOVA and MANOVA, we can't guarantee that the data would be representative of the actual population, nor can we guarantee that the analysis wouldn't be biased towards Group 1 (Gloucester). The small sample size might also contribute the non-normality of groups, which creates further issues.

Also, for the MANOVA analysis, there's a complete violation of multivariate normality, thus the results might not be as reliable as it would be even though when comparing with the results of ANOVA it makes sense. The results from ANOVA might also be contaminated by non-normal chemicals (such as Fe2O3, MgO, CaO, Na2O, TiO2, & MnO).

Therefore, in the future, we would need to 1) collect more data; 2) balance the dataset so that each group/site would have the same/similar number of observations to compare with each other.

# Appendix

## Sample Mean

The sample mean is an unbiased descriptive statistic of the population mean defined as the following:

$$\underline{\overline{x}} = \begin{bmatrix} \overline{x}_1 \\ \overline{x}_2 \\ \vdots \\ \overline{x}_p \end{bmatrix} = \begin{bmatrix} n^{-1} \sum_{i=1}^{n} x_{i1} \\ n^{-1} \sum_{i=1}^{n} x_{i2} \\ \vdots \\ n^{-1} \sum_{i=1}^{n} x_{ip} \end{bmatrix}.$$

## (Univariate) Analysis of Variance (ANOVA)

Analysis of variance (ANOVA), developed by British statistician Ronald Fisher, is a set of statistical tools used to investigate the differences among population means given their samples. It is useful for testing three or more group means for statistical significance.

ANOVA partitions the total sum of squares into the hypothesis sum of squares and the error sum of squares:

$$\begin{aligned} \mathrm{SS}_{total} &= \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ik} - \overline{x})^2 && (1) \\ &= \sum_{k=1}^{g} \sum_{i=1}^{n_k} [(x_{ik} - \overline{x}_k) - (\overline{x}_k - \overline{x})]^2 && (2) \\ &= \sum_{k=1}^{g} \sum_{i=1}^{n_k} (x_{ik} - \overline{x}_k)^2 + \sum_{k=1}^{g} n_k (\overline{x}_k - \overline{x})^2. && (3) \end{aligned}$$

When the null hypothesis is true, we would expect the between-group variability (hypothesis sum of squares) to be small and within-group variability (error sum of squares) to be large

## (Multivariate) Analysis of Variance (MANOVA)

MANOVA allows us to investigate the differences among population means given their samples across multiple variables. Like ANOVA, it also partitions the total sum of squares into error sum of squares (+ cross products)

and hypothesis sum of squares (+ cross products):

$$\begin{aligned}\mathbf{T} &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\underline{x}_i^{(k)} - \overline{\underline{x}}\right)\left(\underline{x}_i^{(k)} - \overline{\underline{x}}\right)' \\ &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}\left\{\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right) + \left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)\right\}\left\{\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right) + \left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)\right\}' \\ &= \sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right)\left(\underline{x}_i^{(k)} - \overline{\underline{x}}^{(k)}\right)' + \sum_{k=1}^{g} n_k\left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)\left(\overline{\underline{x}}^{(k)} - \overline{\underline{x}}\right)'.\end{aligned}$$

When the null hypothesis is true, we would expect the Hypothesis Sum of Squares and Cross Products matrix **H** to be smaller relative to the Error Sum of Squares and Cross Products matrix. How "large" or "small" is determined by four types of tests: Wilks's Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Maximum Root.

## Quantile-quantile plot (Q-Q plot)

Q-Q plot is a graphical method for comparing two probability distributions by plotting the quantiles of the variables against each other. It is helpful for us to assess whether some datasets plausibly came from some theoretical distribution such as normal or exponential. Comparing the plotted data with the normal line, if the Q-Q plot form a line around the straight line in the graph, then the distribution would be considered normal.

## Shapiro-Wilk's method

Shapiro-Wilk's test was developed by Samuel Sanford Shapiro and Martin Wilk in 1965. It is a test for normality in frequentist statistics that tests the null hypothesis that the sample came from a normal distribution.

## Breusch-Pagan test

Breusch–Pagan test was developed by Trevor Breusch and Adrian Pagan in 1979. It is a test for homoskedesticity in linear regression model that tests the null hypothesis that the sample is homoskedestic.

## Bonferroni Correction

A popular method to correct FWER is called the Bonferroni Correction, where the significance level, $\alpha$, is divided by the number of significance tests conducted, m:

$$\alpha_{bonf} = \frac{\alpha}{m}$$

Note that the Bonferroni Correction required the assumption that the significance tests be independent to reach a FWER of $\alpha$.