Midterm Project 2

Jerry Chan

Introduction

In this project, we are going to inspect the Romano-British dataset. We are curious about the whether the population mean of each pottery's chemical concentrations is the same across different kiln.

Data Description

The dataset consists of chemical concentration of 48 Romano-British pottery shards.[1]

Metadata [2]

1. **No**: Integer, Number ID
2. **ID** Characters, ID
3. **Kiln**: Integer, kiln site where the pottery was found.
4. **Al2O3**: Double, concentration of aluminium trioxide
5. **Fe2O3**: Double, concentration of iron trioxide
6. **MgO**: Double, concentration of magnesium oxide
7. **CaO**: Double, concentration of calcium oxide
8. **Na2O**: Double, concentration of natrium oxide
9. **K2O**: Double, concentration of kalium oxide
10. **TiO2**: Double, concentration of titanium oxide
11. **MnO**: Double, concentration of mangan oxide
12. **BaO**: Double, concentration of barium oxide

```
library(data.table)
library(ggplot2)
```

```
## Registered S3 methods overwritten by 'ggplot2':
##   method         from
##   [.quosures     rlang
##   c.quosures     rlang
##   print.quosures rlang
```

Load Data

```
df <- read.csv('RBPottery.csv', header = TRUE)
head(df)
```

	No <int>	ID <fctr>	Kiln <int>	Al2O3 <dbl>	Fe2O3 <dbl>	MgO <dbl>	CaO <dbl>	Na2O <dbl>	K2O <dbl>
1	1	GA1	1	18.8	9.52	2.00	0.79	0.40	3.20
2	2	GA2	1	16.9	7.33	1.65	0.84	0.40	3.05
3	3	GA3	1	18.2	7.64	1.82	0.77	0.40	3.07
4	4	GA4	1	17.4	7.48	1.71	1.01	0.40	3.16

	No <int>	ID <fctr>	Kiln <int>	Al2O3 <dbl>	Fe2O3 <dbl>	MgO <dbl>	CaO <dbl>	Na2O <dbl>	K2O <dbl>
5	5	GA5	1	16.9	7.29	1.56	0.76	0.40	3.05
6	6	GB1	1	17.8	7.24	1.83	0.92	0.43	3.12

6 rows | 1-10 of 13 columns

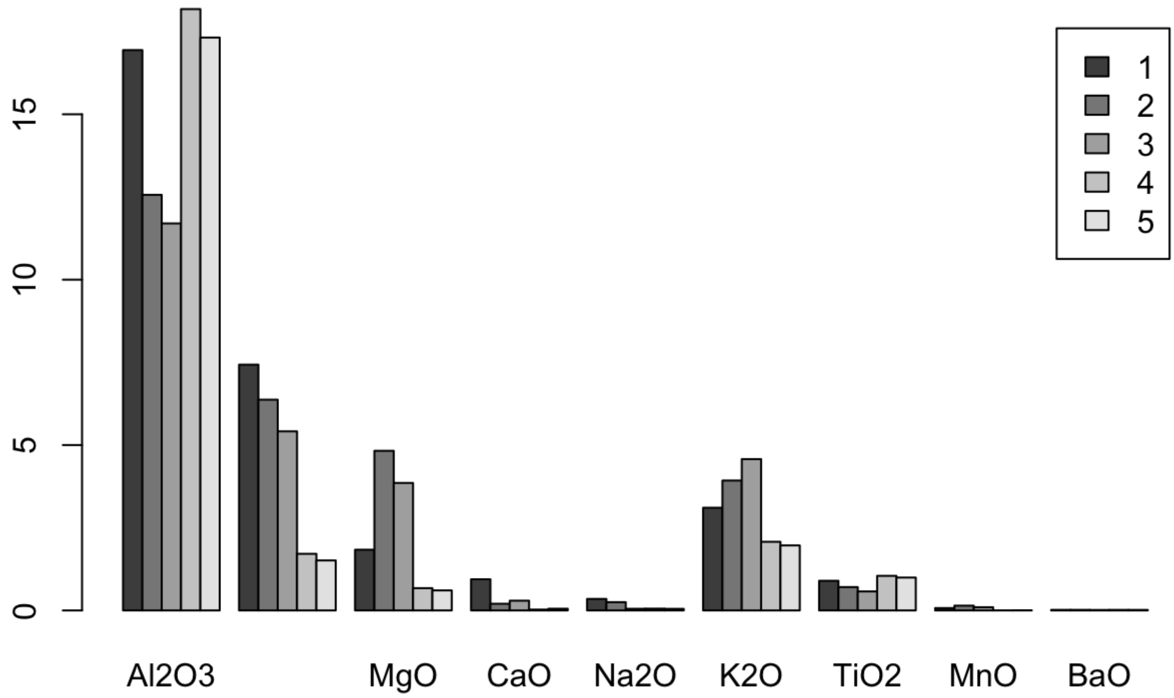
Body

0. observe the dataset and decide test method

First, we plot the means and variances for each chemicals' concentrations grouped by kiln sites.

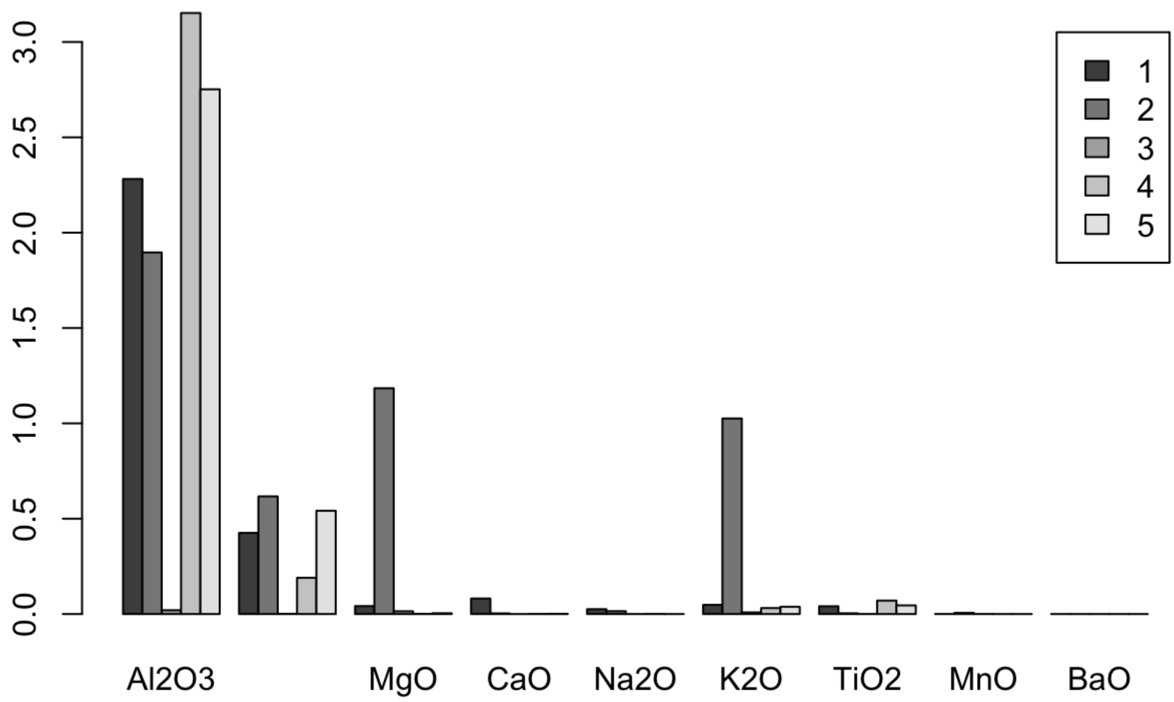
Mean:

```
agg = aggregate(df[,4:12], by=list(df$Kiln), FUN=mean)
m1 <- as.matrix(agg[-1])
row.names(m1) <- agg[,1]
barplot(m1, beside=TRUE, legend=row.names(m1))
```



Variance:

```
agg = aggregate(df[,4:12], by=list(df$Kiln), FUN=var)
m1 <- as.matrix(agg[-1])
row.names(m1) <- agg[,1]
barplot(m1, beside=TRUE, legend=row.names(m1))
```



From the visualization, we can see that the means and variance for each column vary between different kiln. To determine if the difference is statistically significant, we will perform MANOVA (Multivariate analysis of variance) in the following sections.

Null hypothesis: the means vector for each group is the same

Alternative hypothesis: the means vector for each group is not the same

Significance Level:0.05

Assumptions

1. The data from group k has common mean vector: We can assume that pottery from each site is produced by similar processes. Therefore they sould have similar chemical concentration. This assumption is satisfied.

2. Homoskedasticity: The data from all groups have common covariance matrix: From the plot above, we can see that the variances is not the same across each group. Our data doesn't seem to satisfy this assumption.

3. Independence: The observations are independently sampled: Chemical concentration from different pottery are independent. This assumption is satisfied.

4. Normality: The data are multivariate normally distributed: We can assume that chemical concentrations are fixed in different producing procedural and the errors follow normal distribution. Therefore the data are multivariate normally distributed. THis assumption is satisfied.

Test Statistic

In the lecture, 4 test statistics are introduced: Wilks's Lambda, Pillai's Trace, Hotelling-Lawley Trace, and Roy's Maximum Root. In this project, I choose to use Pillai's Trace as it's commonly consider a more robust statistic. The following section I'll implement the test following the steps on Lecture 12.[3]

1. Split the dataset by Kiln

```
df1 <- (df[df$Kiln==1,4:12])
df2 <- (df[df$Kiln==2,4:12])
df3 <- (df[df$Kiln==3,4:12])
df4 <- (df[df$Kiln==4,4:12])
df5 <- (df[df$Kiln==5,4:12])
```

2. Compute grouped sample mean and total sample mean of each column

```
m1 <- colMeans(df1)
n1 <- dim(df1)[1]
m2 <- colMeans(df2)
n2 <- dim(df2)[1]
m3 <- colMeans(df3)
n3 <- dim(df3)[1]
m4 <- colMeans(df4)
n4 <- dim(df4)[1]
m5 <- colMeans(df5)
n5 <- dim(df5)[1]
mg <- (m1*n1 + m2*n2 + m3*n3 + m4*n4 + m5*n5)/(n1 +n2+n3+n4+n5)
mg
```

```
##       Al2O3       Fe2O3         MgO         CaO        Na2O         K2O
## 15.61458333  5.82583333  2.54333333  0.51125000  0.24541667  3.18062500
##        TiO2         MnO         BaO
##  0.85333333  0.07975000  0.01672917
```

3. Compute *E*: Error Sum of Squares

```
ESS <- cov(df1)*(n1-1) + cov(df2)*(n2-1) + cov(df3)*(n3-1) + cov(df4)*(n4-1) + cov(df
5)*(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
```

4. Compute *H*: 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)
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
```

5. Pillai’s Trace

```
N <- n1+n2+n3+n4+n5
g <- 5
p <- 9
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))
p_val <- 1 - pf(pillai_stat,df1 = pillai_s*(2*pillai_m + pillai_s + 1),
       df2 = pillai_s*(2*pillai_r + pillai_s + 1))
p_val
```

```
## [1] 1.391109e-13
```

Conclusion

The p value (1.4e-13) is far below the significance level (0.05). We reject the null hypothesis. There is a significant difference among the 5 group means for these 9 variable.

---

1. Tubb, A., A. J. Parker, and G. Nickless. 1980. "The Analysis of Romano-British Pottery by Atomic Absorption Spectrophotometry". Archaeometry 22: 153-71.↩

2. : T. McElroy, "Ma189Project2" [Online]. Available: https://canvas.ucsd.edu/courses/24041/assignments/274206 (https://canvas.ucsd.edu/courses/24041/assignments/274206) [Accessed: 11-Feb-2021] #### Import Package↩

3. Tucker McElroy, "Ma189Lecture12" [Online]. Available: http://github.com/tuckermcelroy/ma189/tree/main/Lectures (http://github.com/tuckermcelroy/ma189/tree/main/Lectures)↩