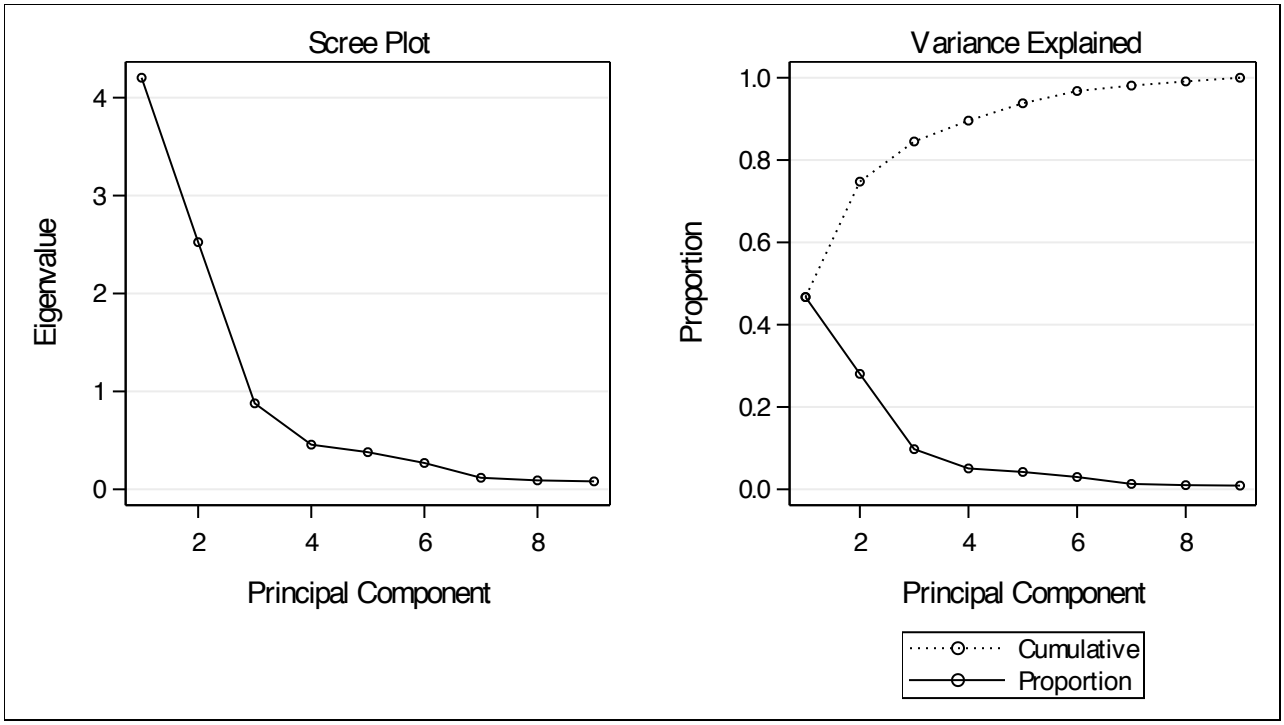
STAT 448
HW5
JINRAN YANG

Problem 1a

The PRINCOMP Procedure

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.20350337	1.67913381	0.4671	0.4671
2	2.52436956	1.64642759	0.2805	0.7475
3	0.87794197	0.42190900	0.0975	0.8451
4	0.45603296	0.07667086	0.0507	0.8958
5	0.37936211	0.11046823	0.0422	0.9379
6	0.26889387	0.15109598	0.0299	0.9678
7	0.11779790	0.02667251	0.0131	0.9809
8	0.09112539	0.01015252	0.0101	0.9910
9	0.08097287		0.0090	1.0000

Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
Al	-.348275	0.327856	0.119016	-.033283	0.321247	0.776634	0.016917	0.219583	0.032872
Fe	0.327132	0.395211	-.264433	-.019252	0.343312	0.045868	-.244490	-.504300	-.482117
Mg	0.434611	-.189543	0.150914	0.055441	0.280789	0.010166	0.444126	0.490206	-.482537
Ca	0.064293	0.501050	-.477908	-.498002	-.065421	-.225888	0.171981	0.393425	0.169549
Na	0.216930	0.455874	-.007046	0.574745	-.533385	0.156247	0.321284	-.045999	0.022040
K	0.456364	-.018368	0.102101	-.036773	0.389624	0.079710	0.307399	-.285211	0.667547
Ti	-.340213	0.300728	0.089586	0.493411	0.491170	-.520837	0.005927	0.147234	0.090027
Mn	0.455251	0.087533	0.140205	0.153209	-.023697	0.047862	-.717466	0.429307	0.200099
Ba	0.018539	0.378263	0.791569	-.385785	-.133038	-.198187	0.024560	-.113736	-.103176



According to the table ***Eigenvalues of the Correlation Matrix***, we need to keep 3 components (84.51%) to retain at least 80% of the total variation from the original variables.
Based on the average eigenvalue, according to the table ***Eigenvalues of the Correlation Matrix***, we notice that there are two variables whose eigenvalue is larger than 1 so that we should choose 2 components.
Based on scree plot, we should choose 4 components, since 4 is an elbow point in the graph (which means the curve after 4 tends to be flat which imply that including these points will not result in big change).

Problem 1b

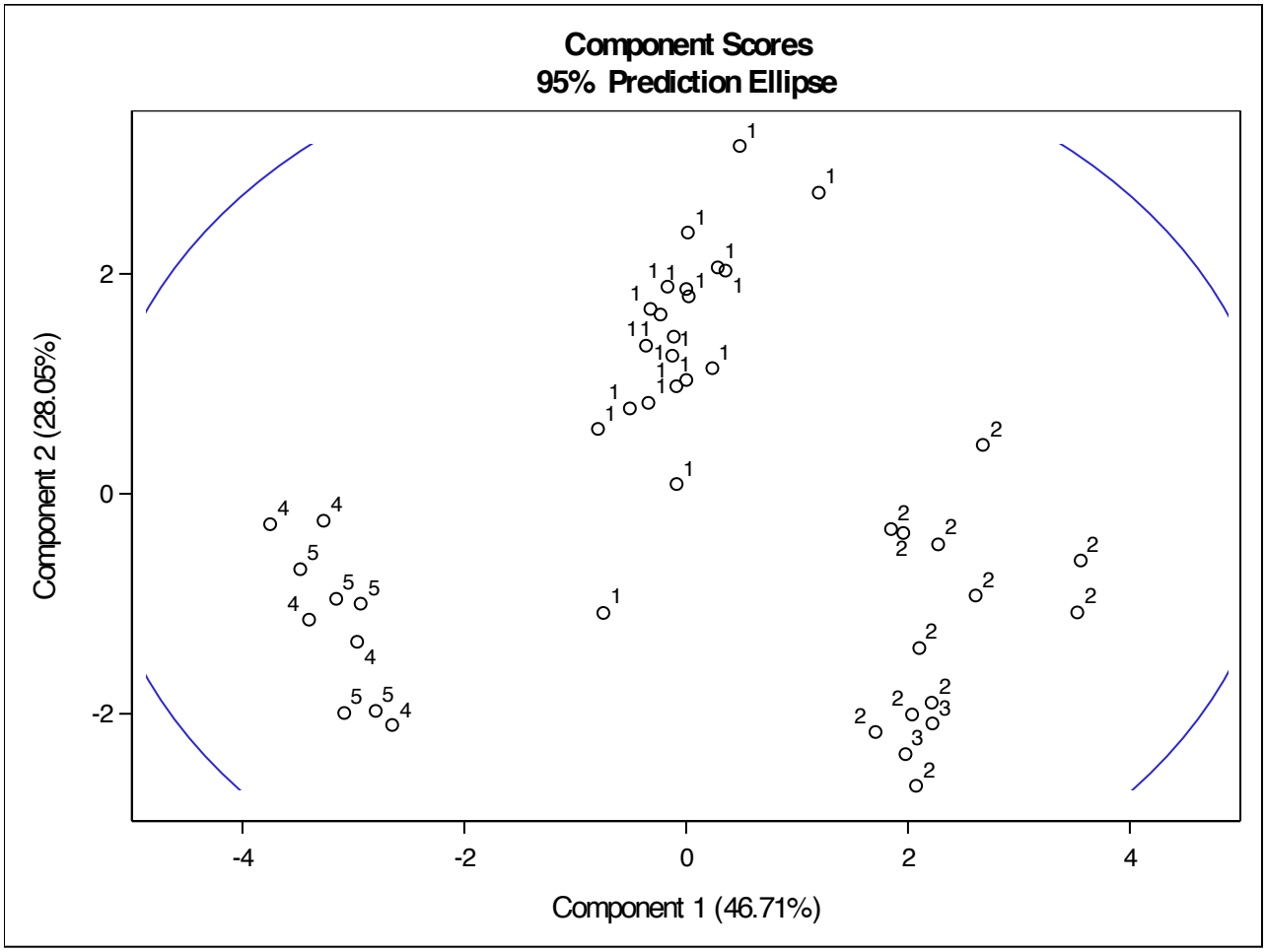
Features of component 1: the corresponding eigenvector value of K, Mn, Mg, and Fe oxide are positive and relatively large, but that of Al and Ti are negative and its absolute value is relatively larger which imply that the **contrast** between the amount of the sum of Fe, Mg, K, Mn oxide and the amount of the sum of Al and Ti oxide.

Features of component 2: the corresponding eigenvector value of Fe, Ca, Na and Ba oxide are positive and relatively large which imply the amount of the sum of Fe, Ca, Na and Ba oxide generally.

Features of component 3: the corresponding eigenvector value of Ba oxide is positive and **really** large, but that of Ca is negative and its absolute value is relatively larger which imply the **contrast** between the amount of Ba oxide and Ca oxide.

The MEANS Procedure

Problem 1c



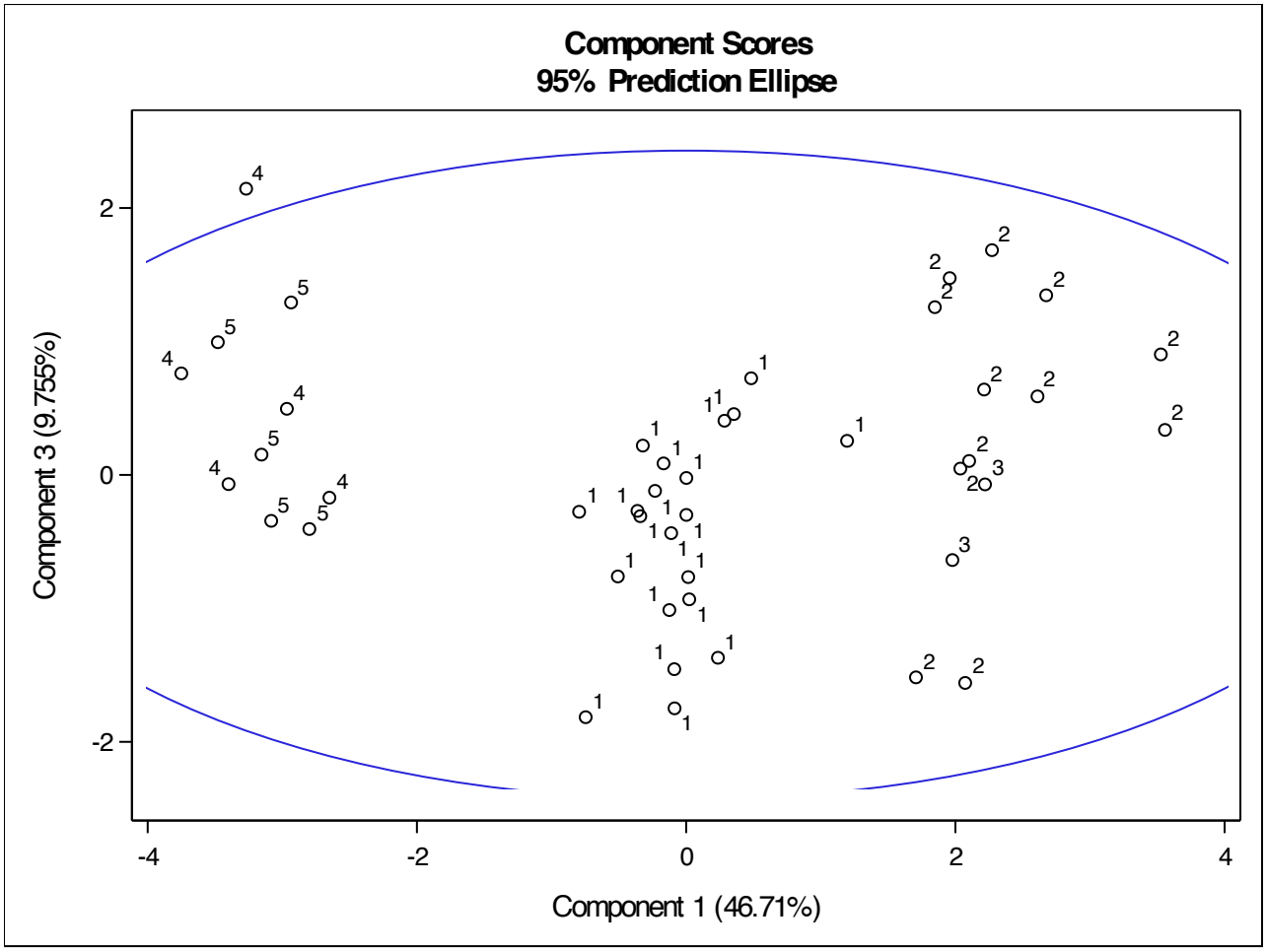
From the plot above, we can find that for potteries found in kiln site 4 and 5 all have a negative score of component 1 and its absolute value is large. According to the interpretation of component 1 we just made, we can make a conclusion that, for the pottery found in kiln site 4 and 5, its usually contain a **larger** amount of Al or Ti oxide compared to the value of sum of Fe, Mg, K, Mn oxide.

Similarly, we can find that for potteries found in kiln site 2 and 3 all have positive and large score of component 1. According to the interpretation of component 1 we just made, we can make a conclusion that, for the pottery found in kiln site 2 and 3, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain is **larger** than the value of the sum of Al and Ti oxide.

For potteries found in kiln site 1, its score for component 1 is close to 0 so that these pottery's value of sum of Fe, Mg, K, Mn oxide is **equal to** the value of the sum of Al and Ti oxide.

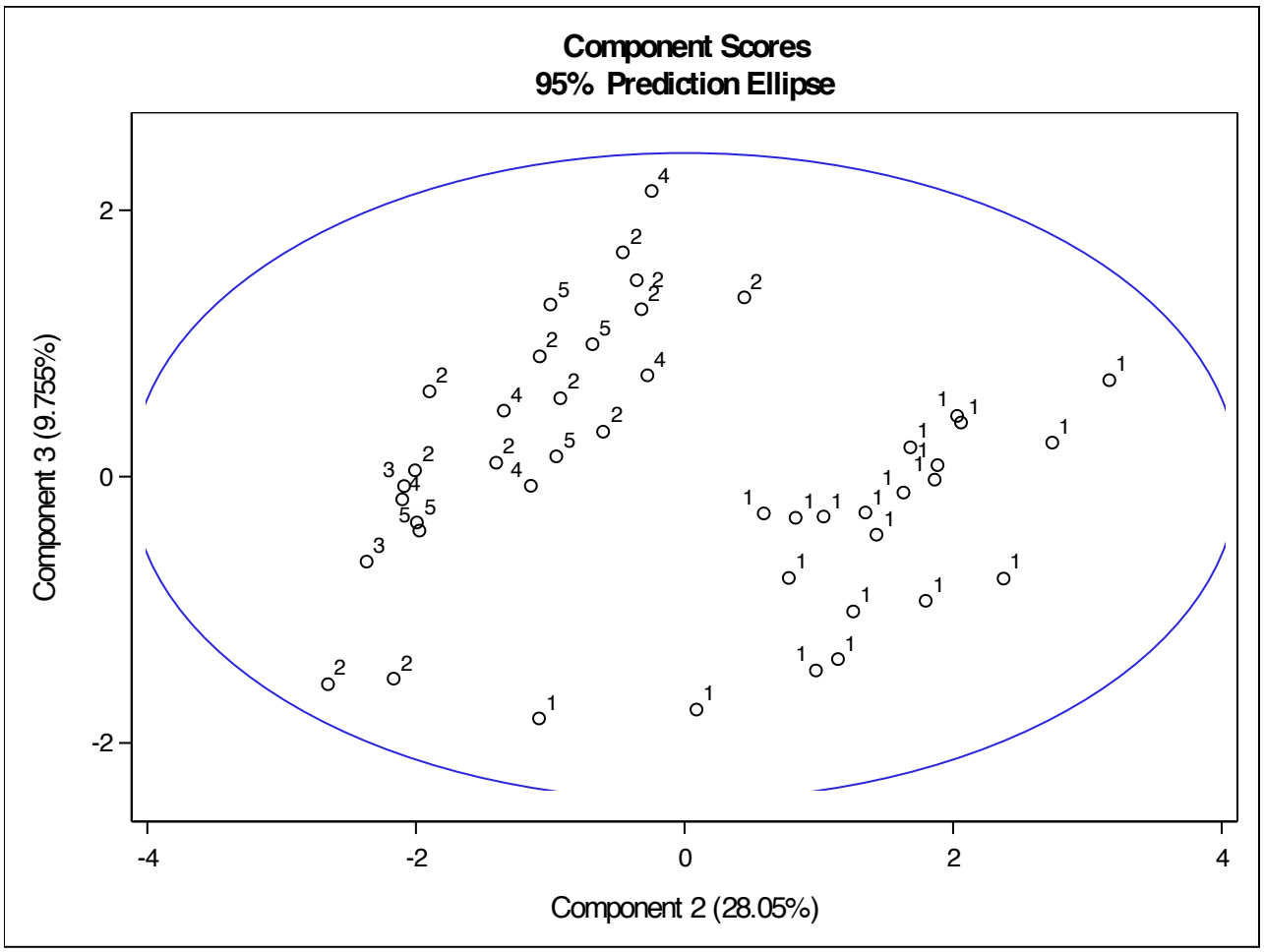
For the interpretation of the score component 2, we can find that these points are all randomly distributed either between 0 and -2 (for potteries found in kiln site 2,3,4 and 5) or between 0 and 2

(for potteries found in kiln site 1). So, for **partially** potteries found in kiln site 2,3,4 and 5, the Mg oxide these potteries contain should be **larger than** the value of the sum of the Fe, Ca, Na, Ba oxide in general.
So, for **partially** potteries found in kiln site 1, the Mg oxide these potteries contain should be **smaller than** the value of the sum of the Fe, Ca, Na, Ba oxide general.



From the plot above, we can find that the score of component 3 for potteries found in any kiln is randomly distributed in a range, so that it is hard to interpret.

The MEANS Procedure



From the plot above, there is no obvious cluster so that it is hard to interpret.

In conclusion, for potteries found in kiln 1, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **equal to** the value of the sum of Al and Ti oxide these potteries contain. Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 1) contain should generally **larger** than the Mg oxide these potteries contain.

For potteries found in kiln 2 and 3, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **larger** than the value of the sum of Al and Ti oxide these potteries contain. Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 2,3) contain should generally **smaller** than the Mg oxide these potteries contain.

For potteries found in kiln 4 and 5, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **smaller** than the value of the sum of Al and Ti oxide these potteries contain. Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 4,5) contain should generally **smaller** than the Mg oxide these potteries contain.

The MEANS Procedure

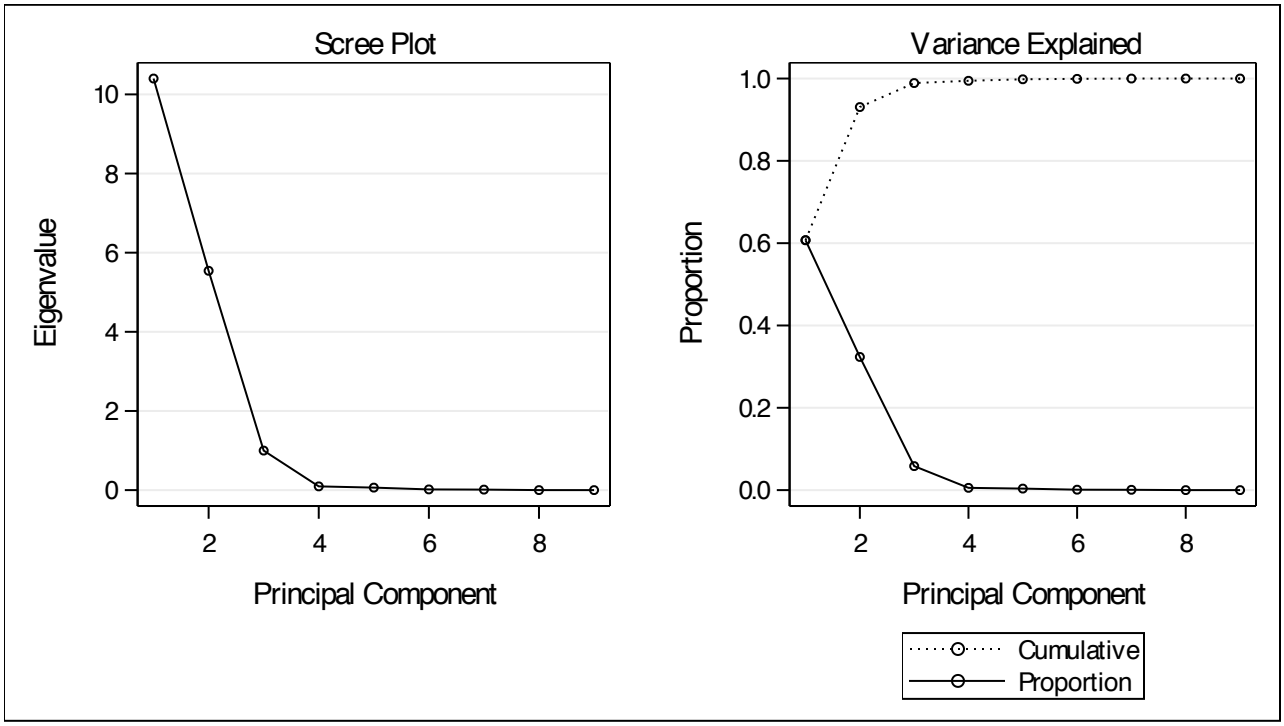
Problem 2a

Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	10.4003191	4.8582150	0.6072	0.6072
2	5.5421041	4.5433501	0.3236	0.9307
3	0.9987540	0.9044789	0.0583	0.9890
4	0.0942752	0.0311993	0.0055	0.9945
5	0.0630759	0.0456610	0.0037	0.9982
6	0.0174149	0.0046641	0.0010	0.9992
7	0.0127508	0.0124129	0.0007	1.0000
8	0.0003379	0.0003323	0.0000	1.0000
9	0.0000057		0.0000	1.0000

Total Variance	17.129037622

Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
Al	-.754921	0.457401	0.468852	-.000690	0.020942	-.010792	-.022321	0.001610	-.000574
Fe	0.383341	0.871114	-.226594	-.085104	-.178456	-.060509	-.009696	-.004595	0.000696
Mg	0.480292	-.018666	0.788149	-.367291	0.107417	0.017780	0.031592	-.006147	0.000061
Ca	-.000278	0.143937	-.178934	-.100764	0.962925	0.026159	0.095221	0.011491	-.001748
Na	0.013294	0.048210	-.019712	-.010552	-.003110	0.936317	-.336926	-.082300	-.002664
K	0.224974	0.090509	0.273167	0.919688	0.127060	0.028709	0.059468	-.015586	-.001112
Ti	-.039195	0.017961	-.023619	-.039876	-.112937	0.336282	0.932387	0.028116	-.004021
Mn	0.011697	0.006384	0.008423	0.013111	-.006438	0.067662	-.054222	0.995221	-.039445
Ba	-.000140	0.000462	0.000555	0.001257	0.001239	0.006634	0.000938	0.039189	0.999208

The MEANS Procedure



Similarities and Differences：

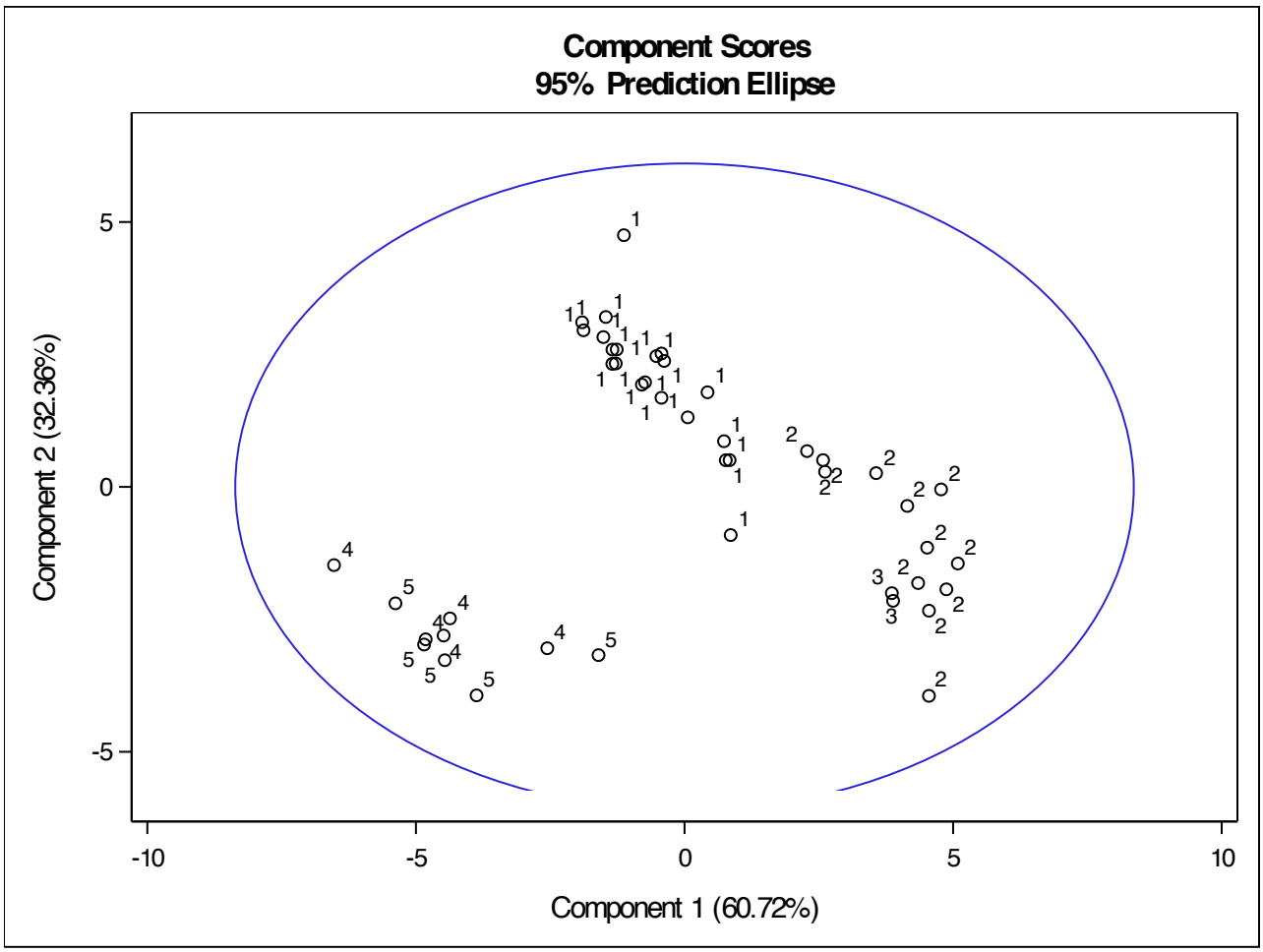
2(a)

Using the ***Correlation Matrix***, we need to keep 3 components (84.51%) to retain at least 80% of the total variation from the original variables. But using the ***Covariance Matrix***, according to the table ***Eigenvalues of the Covariance Matrix***, we just need to keep 2 components (93.07%) to retain at least 80% of the total variation from the original variables.

Using the ***Covariance Matrix,*** the average eigenvalue is 17.13/9 =1.90, so based ***on the average eigenvalue of the Covariance Matrix,*** we notice that there are two variables whose eigenvalue is larger than 1.90 so that we should choose 2 components. And the result is the same as using the ***Correlation Matrix.***

Using the ***Covariance Matrix***, based on scree plot, we should choose 4 components, since 4 is an elbow point in the graph (which means the curve after 4 tends to be flat which imply that including these points will not result in big change). This agree with result of using the ***Correlation Matrix.***

The MEANS Procedure



2(b)

Using the ***Covariance Matrix,*** we draw the conclusions as follow (the same procedures we use to analyze the pca result by using correlation matrix):

For potteries found in kiln 1, the value of the sum of Fe, Mg oxide these potteries contain should generally **equal to** the value of Al oxide these potteries contain. Besides, part of the potteries found in kiln 1 should contain some Al and Fe oxide.

For potteries found in kiln 2 and 3, the value of the sum of Fe, Mg oxide these potteries contain should generally **larger** than value of Al oxide these potteries contain. Besides, part of the potteries found in kiln 2 and 3 should contain some Mg oxide.

For potteries found in kiln 4 and 5, the value of the sum of Fe, Mg oxide these potteries contain should generally **smaller** than value of Al oxide these potteries contain. Besides, part of the potteries found in kiln 4 and 5 should contain some Mg oxide.

Similarities:

(1) Both find that potteries found in kiln site 2, 3, 4 and 5 contain some Mg oxide.
(2) There is relationship between the value of the amount of the Fe and Mg oxide and the value of the amount of Al and Ti oxide for potteries in kiln 1,2,3,4,5.

Differences:

(1) ***Correlation Matrix***
for potteries found in kiln 1, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **equal to** the value of the sum of Al and Ti oxide these potteries contain.

Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 1 contain should generally **larger** than the Mg oxide these potteries contain.

Covariance Matrix
For potteries found in kiln 1, the value of the sum of Fe, Mg oxide these potteries contain should generally **equal to** the value of Al oxide these potteries contain.

Besides, part of the potteries found in kiln 1 should contain some Al and Fe oxide.

(2) ***Correlation Matrix***
For potteries found in kiln 2 and 3, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally larger than the value of the sum of Al and Ti oxide these potteries contain.

Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 2,3) contain should generally smaller than the **Mg** oxide these potteries contain.

Covariance Matrix
For potteries found in kiln 2 and 3, the value of the sum of Fe, Mg oxide these potteries contain should generally larger than value of Al oxide these potteries contain.

Besides, part of the potteries found in kiln 2 and 3 should contain some **Mg** oxide.

(3) ***Correlation Matrix***
For potteries found in kiln 4 and 5, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **smaller** than the value of the sum of Al and Ti oxide these potteries contain.
Besides, the value of the sum of Fe, Ca, Na, Ba oxide these potteries (part of the potteries found in kiln 4,5) contain should generally **smaller** than the **Mg** oxide these potteries contain.

The MEANS Procedure

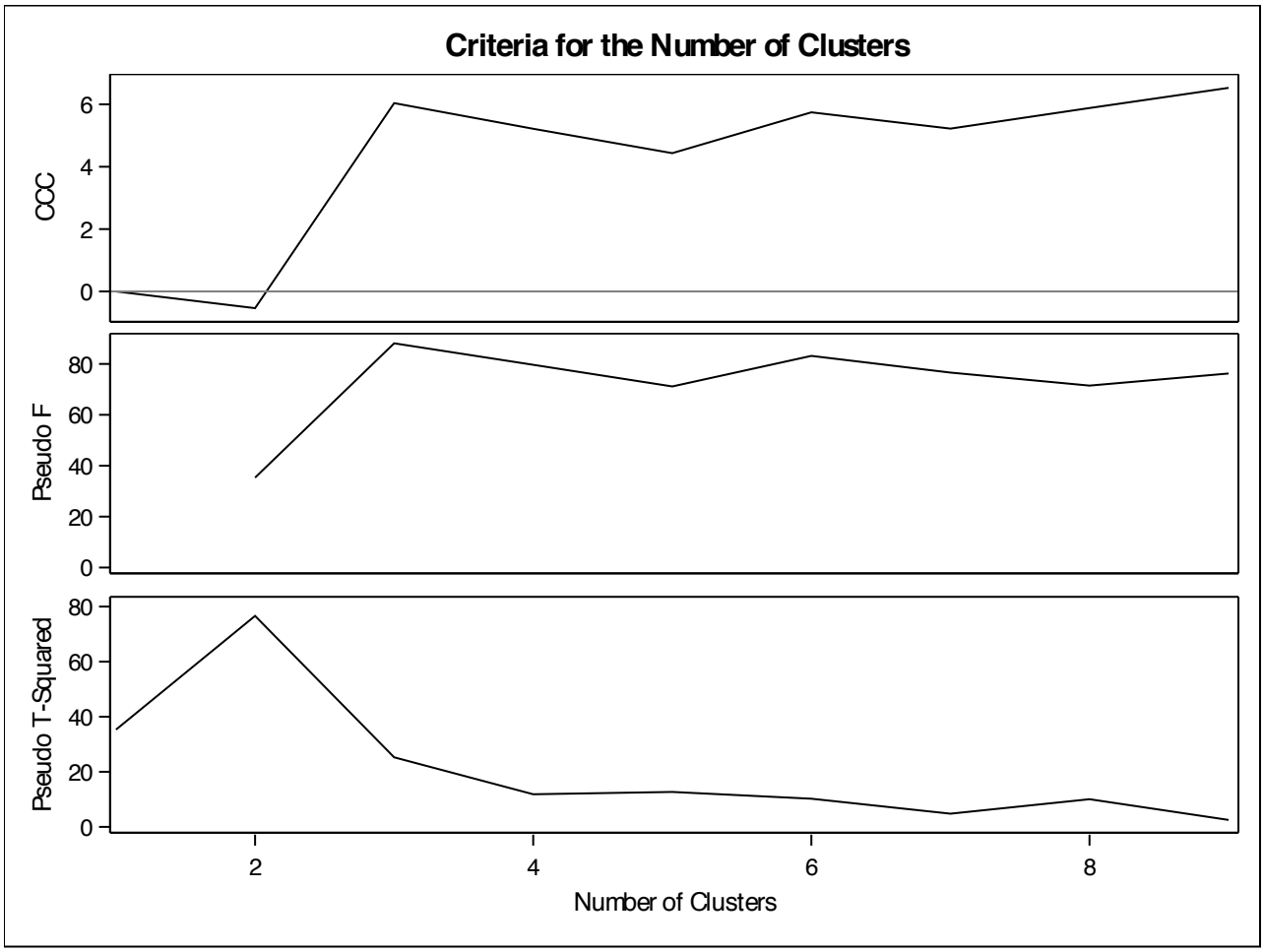
Covariance Matrix

For potteries found in kiln 4 and 5, the value of the sum of Fe, Mg oxide these potteries contain should generally smaller than value of Al oxide these potteries contain.

Besides, part of the potteries found in kiln 4 and 5 should contain some **Mg** oxide.

Problem 3

3(a)



Since higher ccc and pseudo F values indicate better clustering, and lower pseudo t^2 values indicate better clustering, here we choose 3 clusters.

The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	21	16.9190476	1.5442212	13.7000000	18.9000000
Fe	21	7.4285714	0.6684331	5.8300000	9.5200000
Mg	21	1.8423810	0.2070243	1.5000000	2.3300000
Ca	21	0.9390476	0.2919230	0.6600000	1.7300000
Na	21	0.3461905	0.1634771	0.1200000	0.8300000
K	21	3.1028571	0.2247697	2.2500000	3.3700000
Ti	21	0.9376190	0.0585581	0.7500000	1.0100000
Mn	21	0.0711429	0.0186636	0.0340000	0.1120000
Ba	21	0.0171429	0.0026511	0.0120000	0.0230000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	14	12.4357143	1.4118221	10.1000000	14.6000000
Fe	14	6.2078571	0.8490916	4.2600000	7.0900000
Mg	14	4.7778571	1.1209967	3.4300000	7.2300000
Ca	14	0.2142857	0.0673355	0.1200000	0.3100000
Na	14	0.2257143	0.1430822	0.0400000	0.5400000
K	14	4.1878571	0.4735330	3.3200000	4.8900000
Ti	14	0.6828571	0.0756946	0.5600000	0.8100000
Mn	14	0.1176429	0.0315512	0.0800000	0.1630000
Ba	14	0.0159286	0.0034965	0.0090000	0.0210000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	10	17.7500000	1.6820953	14.8000000	20.8000000
Fe	10	1.6120000	0.5799579	0.9200000	2.7400000
Mg	10	0.6400000	0.0594418	0.5300000	0.7200000
Ca	10	0.0390000	0.0317805	0.0100000	0.1000000
Na	10	0.0510000	0.0202485	0.0300000	0.1000000
K	10	2.0210000	0.1850195	1.7500000	2.3700000
Ti	10	1.0200000	0.2285704	0.6500000	1.3400000
Mn	10	0.0032000	0.0023944	0.0010000	0.0070000
Ba	10	0.0160000	0.0029059	0.0130000	0.0220000

According to the tables above, potteries in cluster 1 generally contain **very high** level of **Al oxide**, some Fe, K and Mg oxide. And very little any other components.
Potteries in cluster 2 generally contain **high level of Al oxide,** quite some Fe, K and Mg oxide. And very little any other components.
Majority component of potteries in cluster 3 is **Al oxide,** and there is also some Fe, K oxide. And barely no any other components.

The FREQ Procedure

Table of CLUSTER by Kiln						
CLUSTER	**Kiln**					
Frequency	**1**	**2**	**3**	**4**	**5**	**Total**
1	21	0	0	0	0	21
2	0	12	2	0	0	14
3	0	0	0	5	5	10
Total	21	12	2	5	5	45

According to the table above, we can find that cluster analysis does a good job on matching the original kilns although it doesn't spilt the potteries in different kiln thoroughly. For cluster 1, it does a good job. All potteries in cluster 1 found in kiln 1. However, for cluster 2, it contains potteries found in kiln 2 and 3. And as for cluster 3, it contains potteries found in kiln 4 and 5.

Similarities with what observed in the PCA results:

(1) Both find that potteries found in kiln site 2, 3, 4 and 5 contain some Mg oxide.
(2) There is relationship between the value of the amount of the Fe and Mg oxide and the value of the amount of Al and Ti oxide for potteries in kiln 1,2,3,4,5.

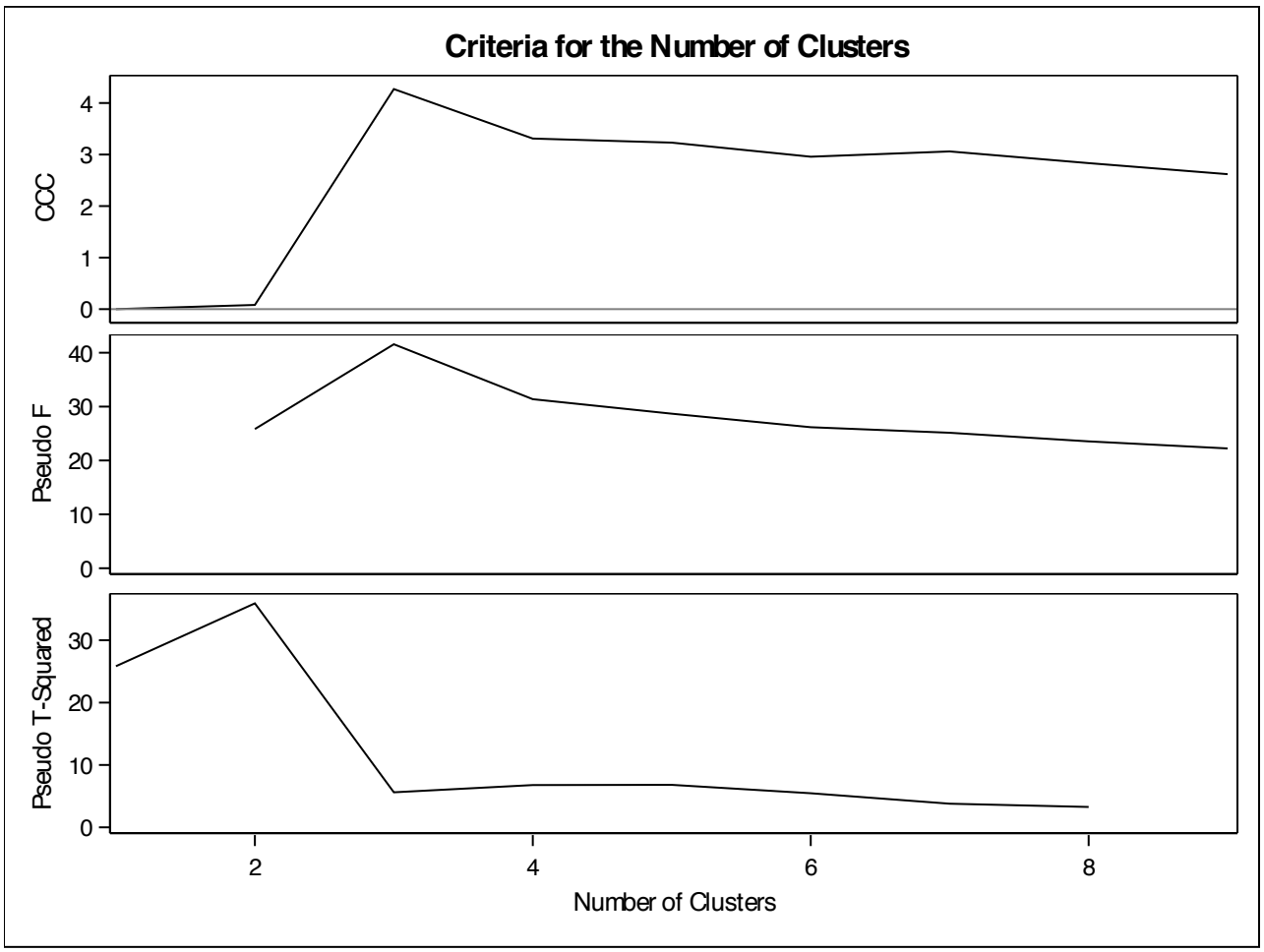
More specifically, for potteries found in kiln 1, the value of the sum of Fe, Mg, K oxide these potteries contain should generally **equal to** the value of Al oxide these potteries contain.

For potteries found in kiln 2 and 3, the value of the sum of Fe, Mg, K, Mn oxide these potteries contain should generally **larger than** the value of the sum of Al and Ti oxide these potteries contain.

For potteries found in kiln 4 and 5, the value of the sum of Fe, Mg oxide these potteries contain should generally **smaller than** value of Al oxide these potteries contain.

Problem 4

The data have been standardized to mean 0 and variance 1



Since higher ***ccc*** and ***pseudo F*** values indicate better clustering, and lower ***pseudo t^2*** values indicate better clustering, here we choose 3 clusters.

The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	21	16.9190476	1.5442212	13.7000000	18.9000000
Fe	21	7.4285714	0.6684331	5.8300000	9.5200000
Mg	21	1.8423810	0.2070243	1.5000000	2.3300000
Ca	21	0.9390476	0.2919230	0.6600000	1.7300000
Na	21	0.3461905	0.1634771	0.1200000	0.8300000
K	21	3.1028571	0.2247697	2.2500000	3.3700000
Ti	21	0.9376190	0.0585581	0.7500000	1.0100000
Mn	21	0.0711429	0.0186636	0.0340000	0.1120000
Ba	21	0.0171429	0.0026511	0.0120000	0.0230000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	10	17.7500000	1.6820953	14.8000000	20.8000000
Fe	10	1.6120000	0.5799579	0.9200000	2.7400000
Mg	10	0.6400000	0.0594418	0.5300000	0.7200000
Ca	10	0.0390000	0.0317805	0.0100000	0.1000000
Na	10	0.0510000	0.0202485	0.0300000	0.1000000
K	10	2.0210000	0.1850195	1.7500000	2.3700000
Ti	10	1.0200000	0.2285704	0.6500000	1.3400000
Mn	10	0.0032000	0.0023944	0.0010000	0.0070000
Ba	10	0.0160000	0.0029059	0.0130000	0.0220000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
Al	14	12.4357143	1.4118221	10.1000000	14.6000000
Fe	14	6.2078571	0.8490916	4.2600000	7.0900000
Mg	14	4.7778571	1.1209967	3.4300000	7.2300000
Ca	14	0.2142857	0.0673355	0.1200000	0.3100000
Na	14	0.2257143	0.1430822	0.0400000	0.5400000
K	14	4.1878571	0.4735330	3.3200000	4.8900000
Ti	14	0.6828571	0.0756946	0.5600000	0.8100000
Mn	14	0.1176429	0.0315512	0.0800000	0.1630000
Ba	14	0.0159286	0.0034965	0.0090000	0.0210000

The MEANS Procedure

Table of CLUSTER by Kiln						
CLUSTER	**Kiln**					
Frequency	**1**	**2**	**3**	**4**	**5**	**Total**
1	21	0	0	0	0	21
2	0	0	0	5	5	10
3	0	12	2	0	0	14
Total	21	12	2	5	5	45

As we can see the result are relatively the same as the result of Question 3.

Theoretically, the result of the Question 4 should do a better job of matching the original kilns, because after standardizing variables, every variable is considered to weight equally rather than one or two variables with high absolute value dominate the result. However, in this question, both of them perform exactly the same.