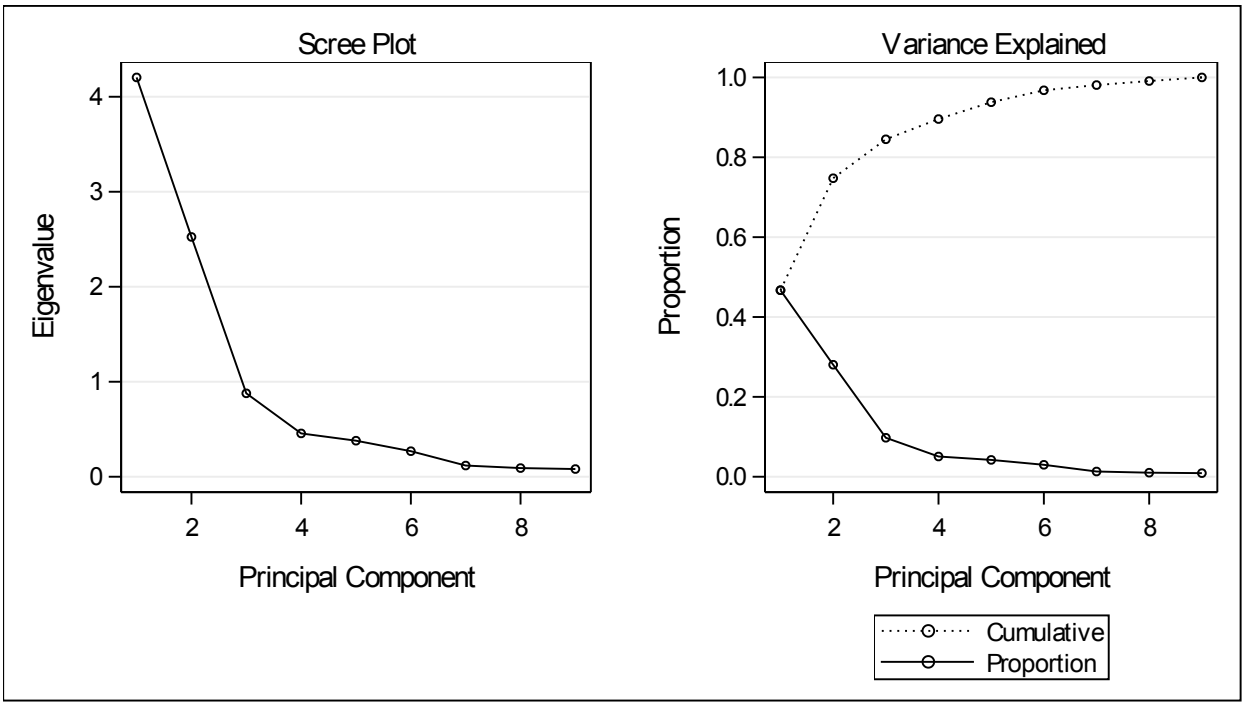
Stat 448, homework 5
Shuhui Guo

Exercise 1
a)

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



According to the above results, I would keep 3 components to retain at least 80% of the total variation from the original variables since the cumulative percentage of the eigenvalues is larger than 80% when there are three components. Based on the average eigenvalue, 2 components will be chosen because the first two eigenvalues are larger than 1. Based on the scree plot, 2 components will be chosen since the line chart becomes flatter after 2 components.

b)

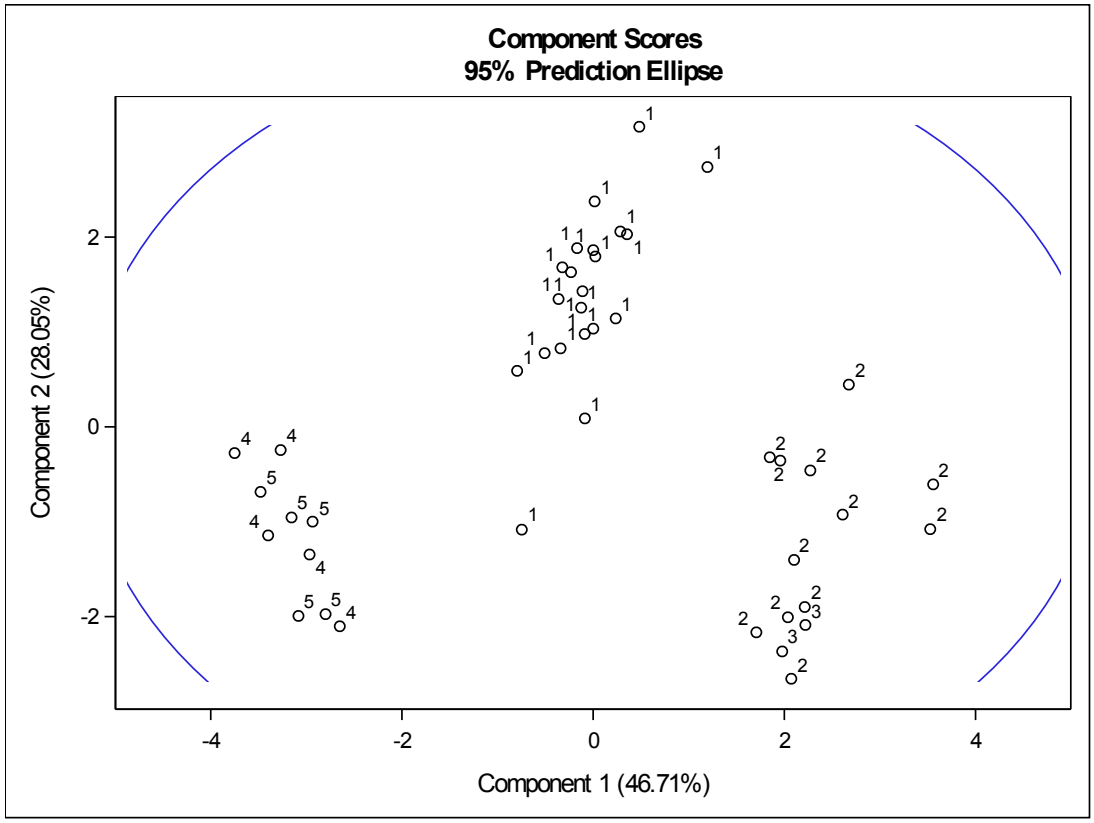
Eigenvectors									
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**	**Prin9**
Al	-.348275	0.327856	0.119016	-.033283	0.321247	0.776634	0.016917	0.219583	0.032872
Fe	0.327132	0.395211	-.264433	-.019252	0.343312	0.045868	-.244490	-.504300	-.482117
Mg	0.434611	-.189543	0.150914	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

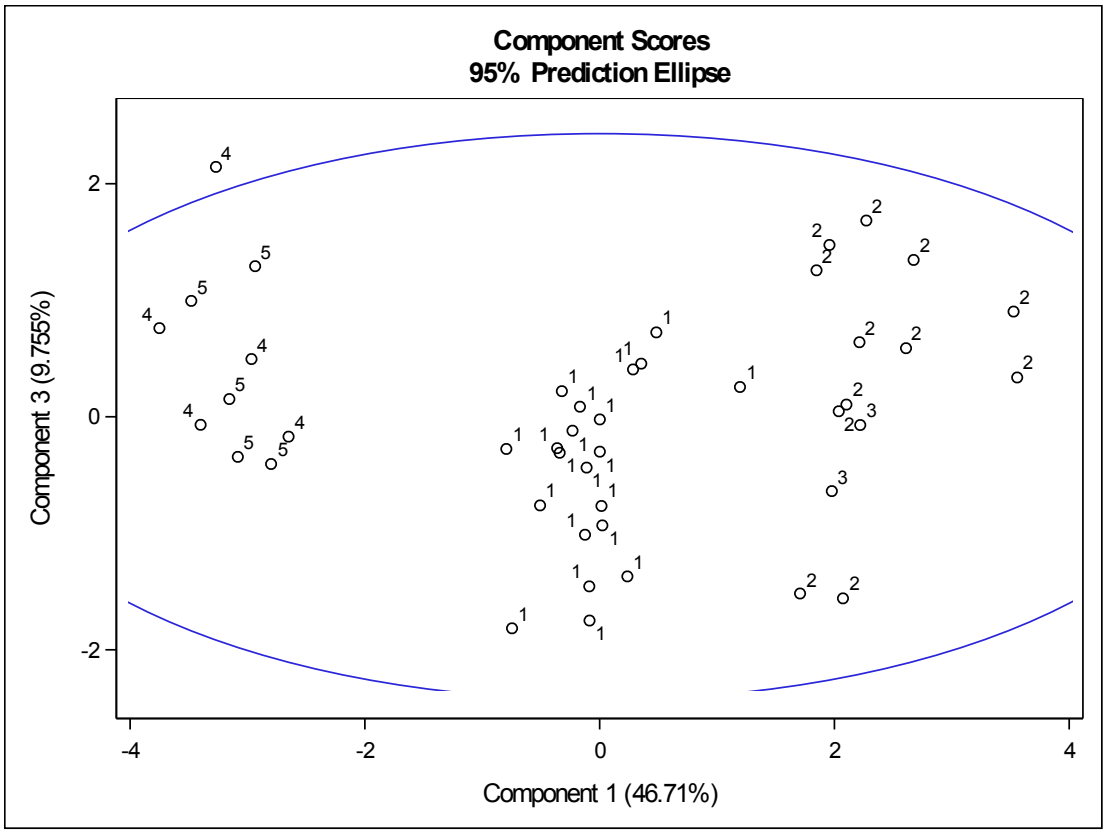
Based on the 80% criterion in part a, three components are selected. The first principle component has large positive coefficients on Mg, K, Mn. This component can be viewed as a measure of the contents of Magnesium oxide, Potassium oxide, and Manganese oxide. The first principle component increases with increasing Magnesium oxide, Potassium oxide, and Manganese oxide contents in the pottery.

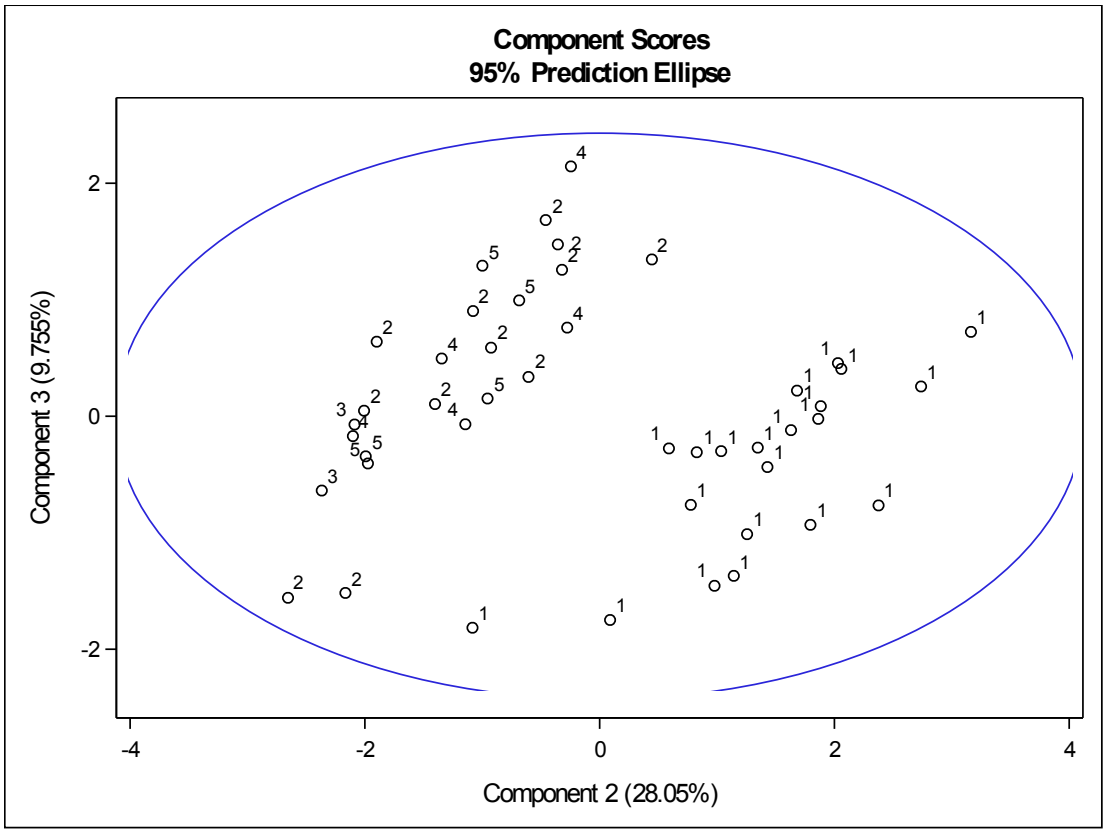
The second principle component has large positive coefficients on Ca and Na. This component can be viewed as a measure of the contents of Calcium oxide and Sodium oxide. The second principle component increases with increasing Calcium oxide and Sodium oxide contents in the pottery.

The third principle component has large positive coefficient on Ba and large negative coefficient on Ca. This component can be viewed as a measure of the contents of Barium oxide and Calcium oxide. The third principle component increases with increasing Barium oxide content and decreasing Calcium oxide content in the pottery.

c)


Component Scores
95% Prediction Ellipse
Component 2 (28.05%)
Component 1 (46.71%)
2
0
-2
-4
-2
0
2
4


Component Scores
95% Prediction Ellipse
Component 3 (9.755%)
Component 1 (46.71%)
2
0
-2
-4
-2
0
2
4



According to the above score plots, kiln site 1 has large positive value on component 2. Therefore, kiln site 1 has high contents of Calcium oxide and Sodium oxide.

Kiln sites 2 and 3 both have large positive value on component 1 and large negative value on component 2. Therefore, kiln sites 2 and 3 have high contents of Magnesium oxide, Potassium oxide, and Manganese oxide and low contents of Calcium oxide and Sodium oxide.

Kiln site 4 has large negative value on component 1 and 2, and large positive value on component 3. Therefore, kiln site 4 has low contents of Magnesium oxide, Potassium oxide, Manganese oxide, Calcium oxide, Sodium oxide and high contents of Barium oxide.

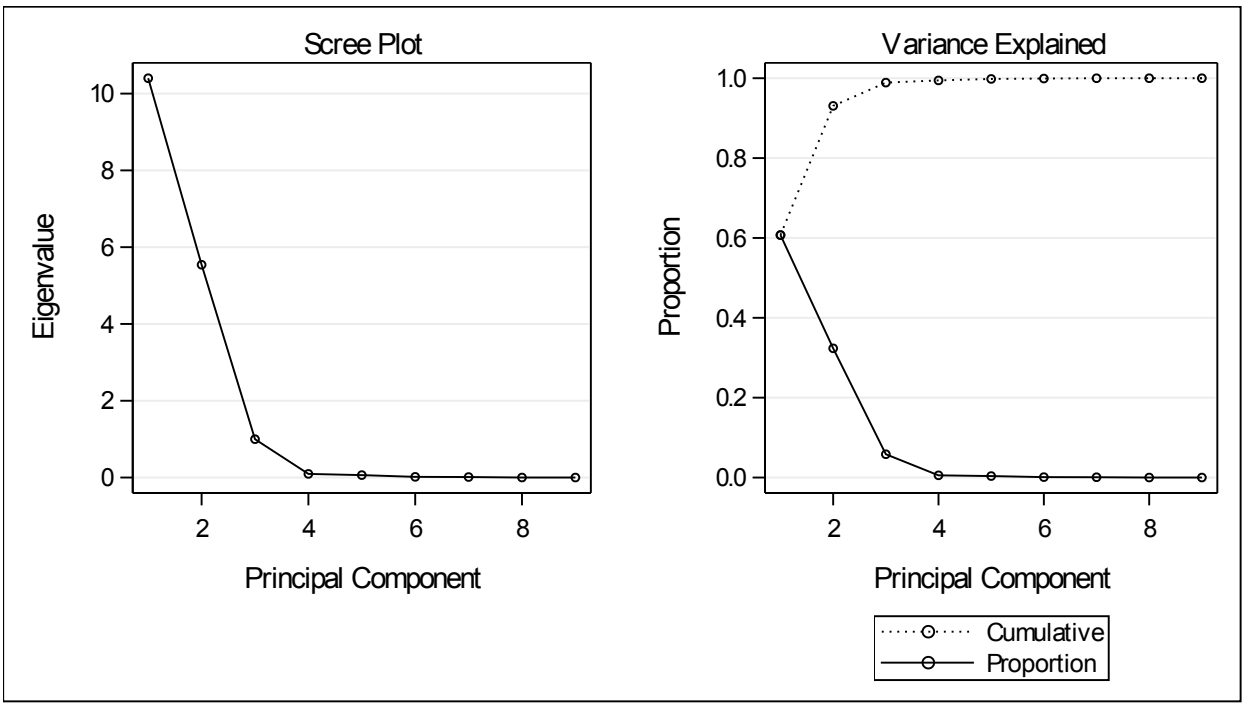
Kiln site 5 has large negative value on component 1 and 2. Therefore, kiln site 5 has low contents of Magnesium oxide, Potassium oxide, Manganese oxide, Calcium oxide, and Sodium oxide.

There are certain similarities and differences of oxide contents for various kiln sites. Kiln site 1 has high contents of Calcium oxide and Sodium oxide while kiln sites 2, 3, 4, 5 have low contents of these two types of oxide. Kiln sites 2 and 3 both have high contents of Magnesium oxide, Potassium oxide, and Manganese oxide, while kiln sites 4 and 5 have low contents of these three types of oxide. In addition, kiln site 4 has high content of Barium oxide.

Exercise 2

a)

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



According to the above results, I would keep 2 components to retain at least 80% of the total variation from the original variables since the cumulative percentage of the eigenvalues is larger than 80% when there are two components. Based on the average eigenvalue, 2 components will be chosen because the first two eigenvalues are larger than average eigenvalue. Based on the scree plot, 2 components will be chosen since the line chart becomes flatter after 2 components.

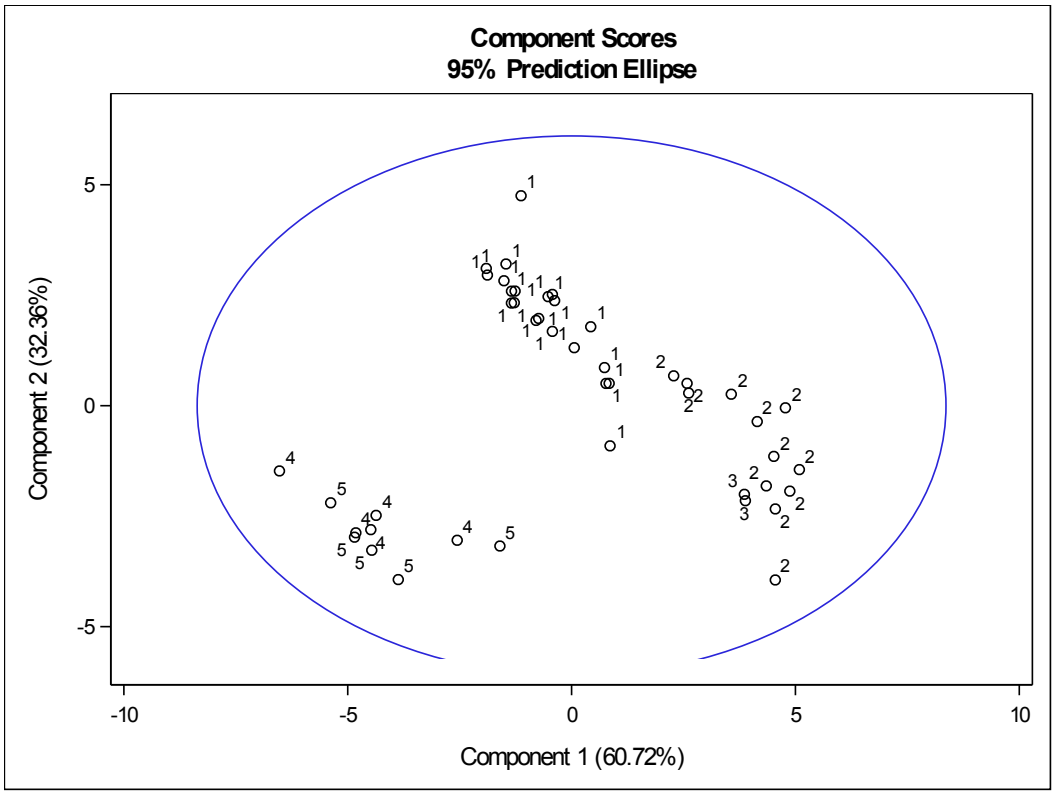
b)

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

Based on the 80% criterion, two components are selected. The first principle component has large negative coefficient on Al. This component can be viewed as a measure of the content of Aluminium oxide. The first principle component increases with decreasing Aluminium oxide content in the pottery.

The second principle component has large positive coefficient on Fe. This component can be viewed as a measure of the content of Ferrous oxide. The second principle component increases with increasing Ferrous oxide content in the pottery.

c)



According to the above score plot, kiln site 1 has large positive value on component 2. Therefore, kiln site 1 has high content of Ferrous oxide.

Kiln sites 2 and 3 both have large positive value on component 1 and large negative value on component 2. Therefore, kiln sites 2 and 3 have low contents of Aluminium oxide and Ferrous oxide.

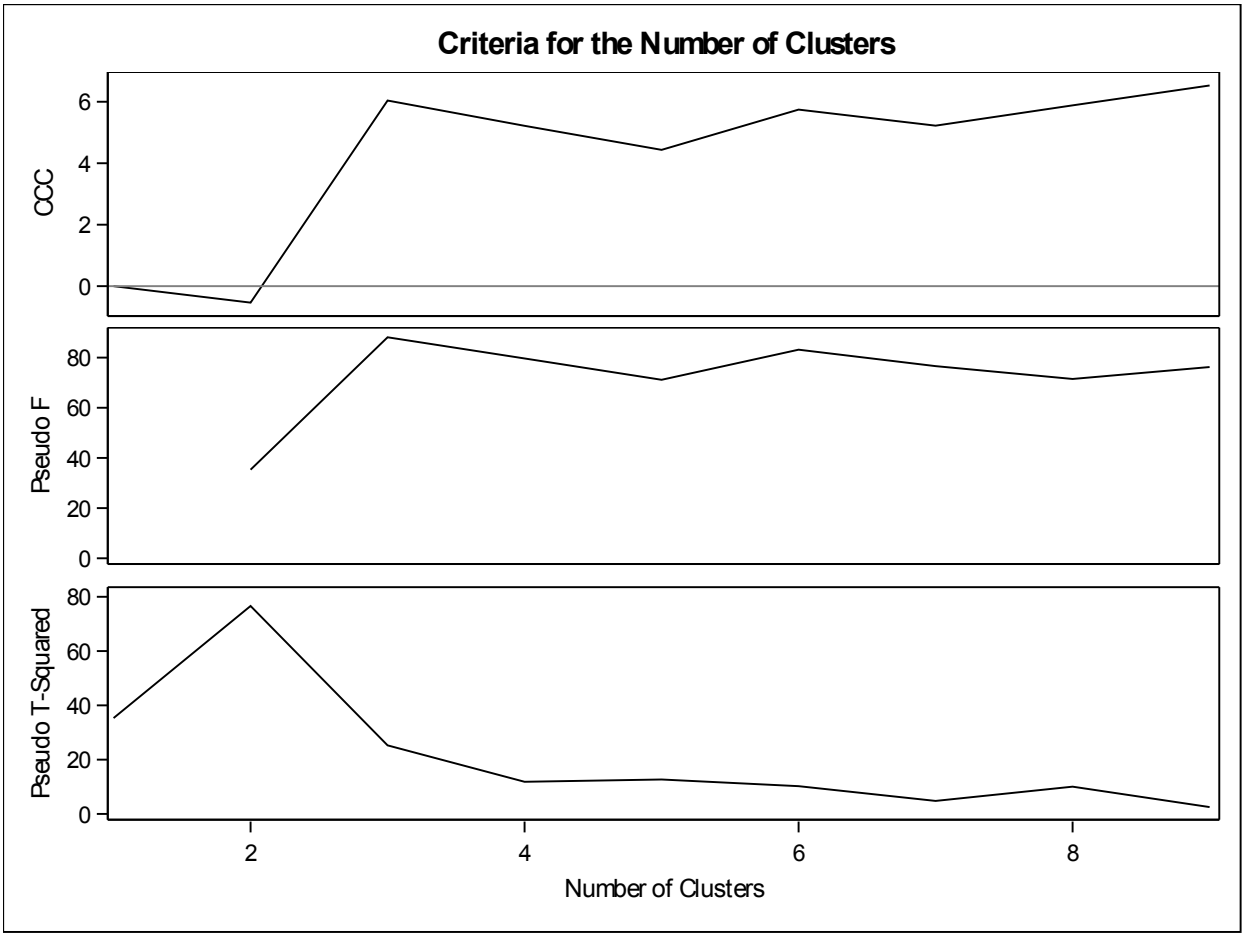
Kiln sites 4 and 5 both have large negative value on component 1 and 2. Therefore, kiln sites 4 and 5 have high content of Aluminium oxide and low content of Ferrous oxide.

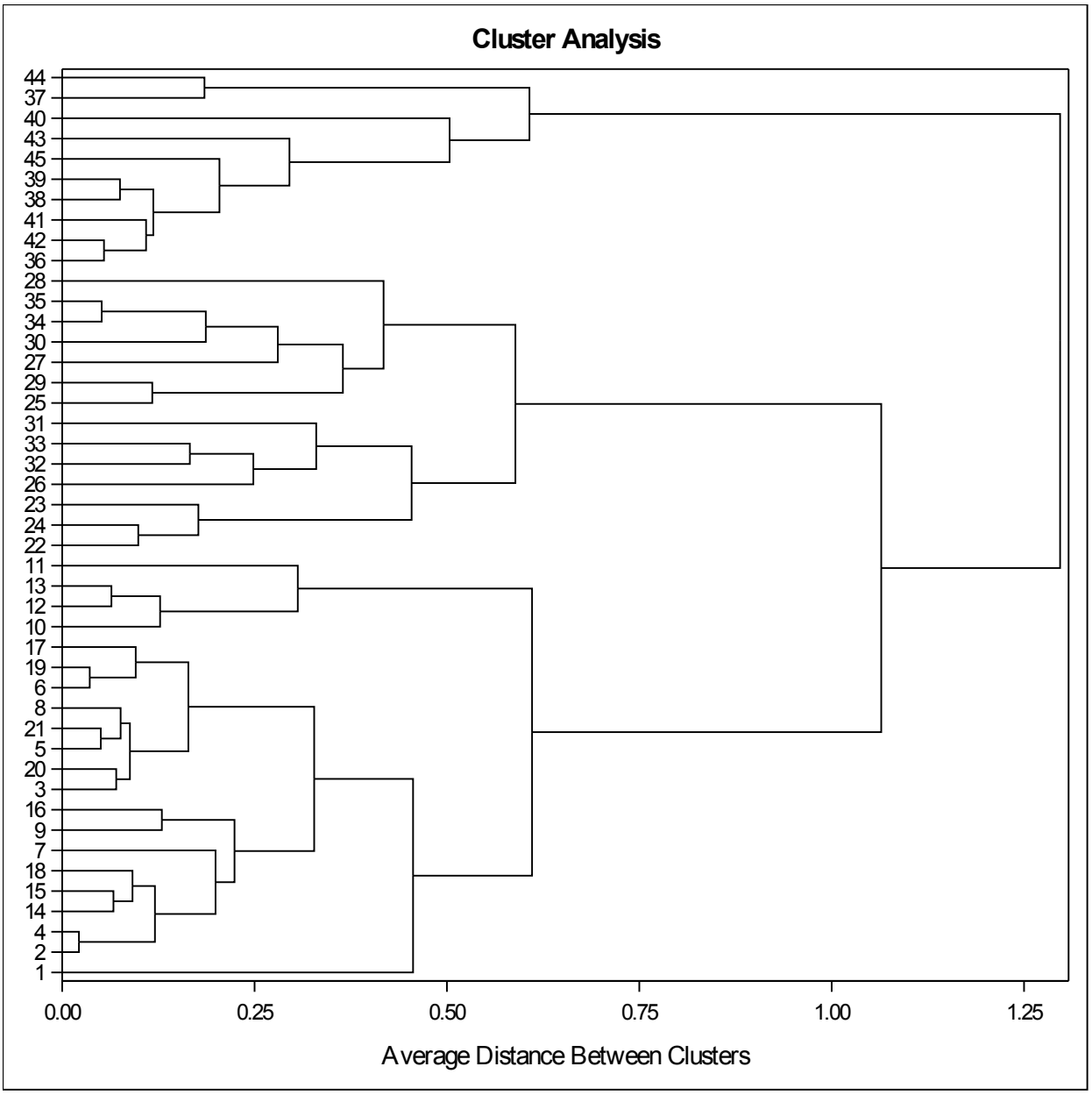
There are certain similarities and differences of oxide contents for various kiln sites. Kiln site 1 has high content of Ferrous oxide while kiln sites 2, 3, 4, 5 have low content of this type of oxide. Kiln sites 2 and 3 both have low content of Aluminium oxide, while kiln sites 4 and 5 have high content of this type of oxide.

In addition, there are certain similarities and differences between correlation-based and covariance-based results. Based on 80% criterion, there are 3 principal components in correlation-based results, while there are 2 principal components in covariance-based results. Based on the average eigenvalue, there are 2 principal components in both correlation-based and covariance-based results. Also, based on the scree plot, there are 2 principal components in both correlation-based and covariance-based results. Furthermore, the covariance-based results give extremely large coefficient to one original variable in a component because this particular variable has large covariance. Nevertheless, the correlation-based results have comparatively balanced coefficients in original variables and do not give great weight to a single variable, which is more reasonable.

Exercise 3

a)


Criteria for the Number of Clusters
CCC
6
4
2
0
Pseudo F
80
60
40
20
0
Pseudo T-Squared
80
60
40
20
0
2
4
6
8
Number of Clusters



Based on the dendrogram, 3 clusters should be chosen to ensure the clusters have a comparatively large distance. Based on the pseudo F statistic, 3, 6 clusters should be chosen. Based on the pseudo t-squared statistic, 3, 4, 7, 9 clusters should be chosen. Based on ccc statistic, 3, 6, 9 clusters should be chosen. Therefore, choosing 3 clusters is appropriate.

The means analysis on variables by cluster is as below:

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

There are oxide differences between clusters. For Aluminium oxide, the means of cluster 1 and 3 are close and much larger than the mean of cluster 2. For Ferrous oxide, the means of cluster 1 and 2 are close and much larger than the mean of cluster 3. For Magnesium oxide, Potassium oxide, and Manganese oxide, means of the three clusters are different. The mean of cluster 2 is the largest while the mean of cluster 3 is the smallest. For Calcium oxide and Sodium oxide, means of the three clusters are different. The mean of cluster 1 is the largest while the mean of cluster 3 is the smallest. For Titanium oxide, means of the three clusters are different. The mean of cluster 3 is the largest while the mean of cluster 2 is the smallest. For Barium oxide, means of the three clusters are quite close.

b)

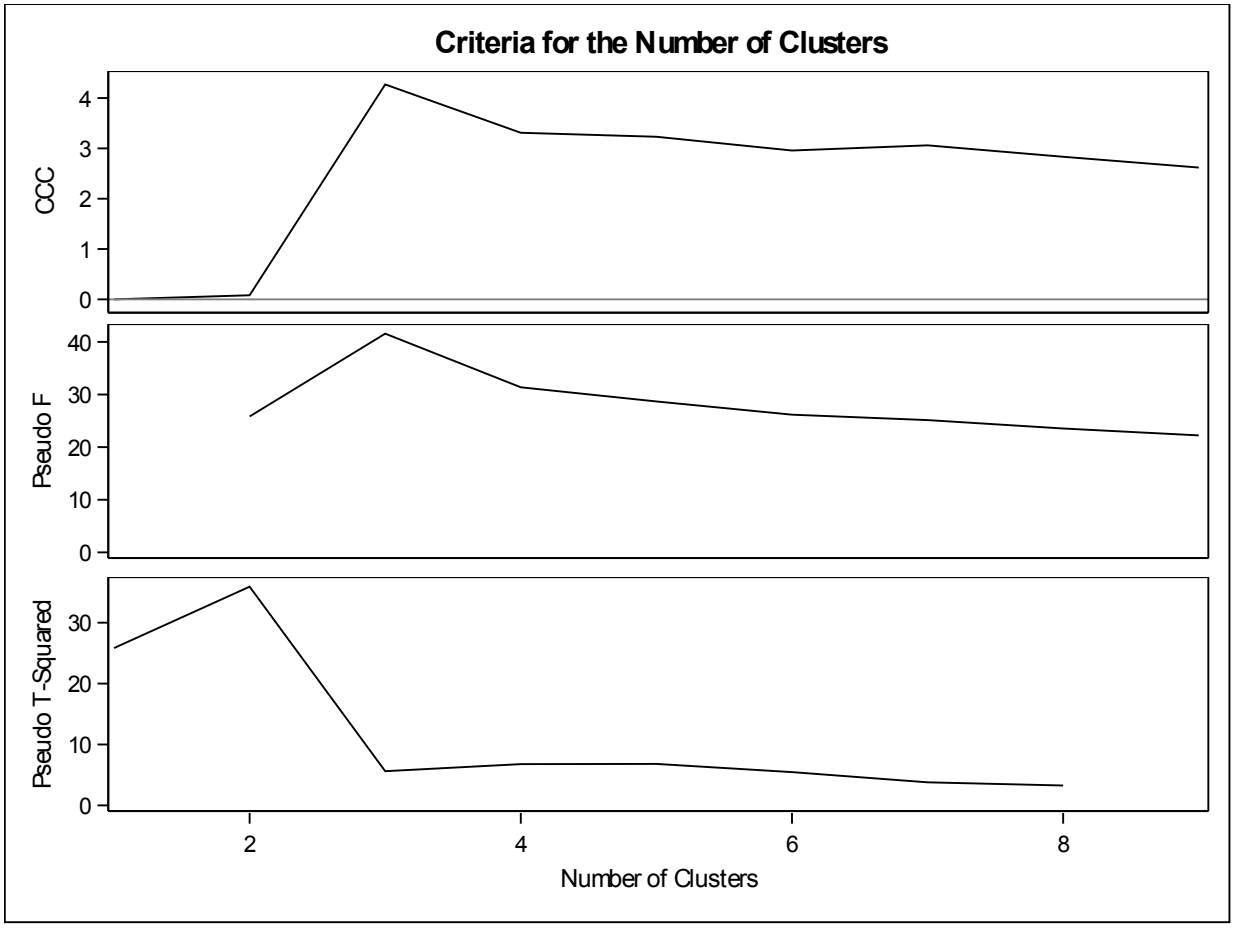
Table of CLUSTER by Kiln						
CLUSTER	**Kiln**					
Frequency	**1**	**2**	**3**	**4**	**5**	**Total**
1	21	0	0	0	0	21
2	0	12	2	0	0	14
3	0	0	0	5	5	10
Total	21	12	2	5	5	45

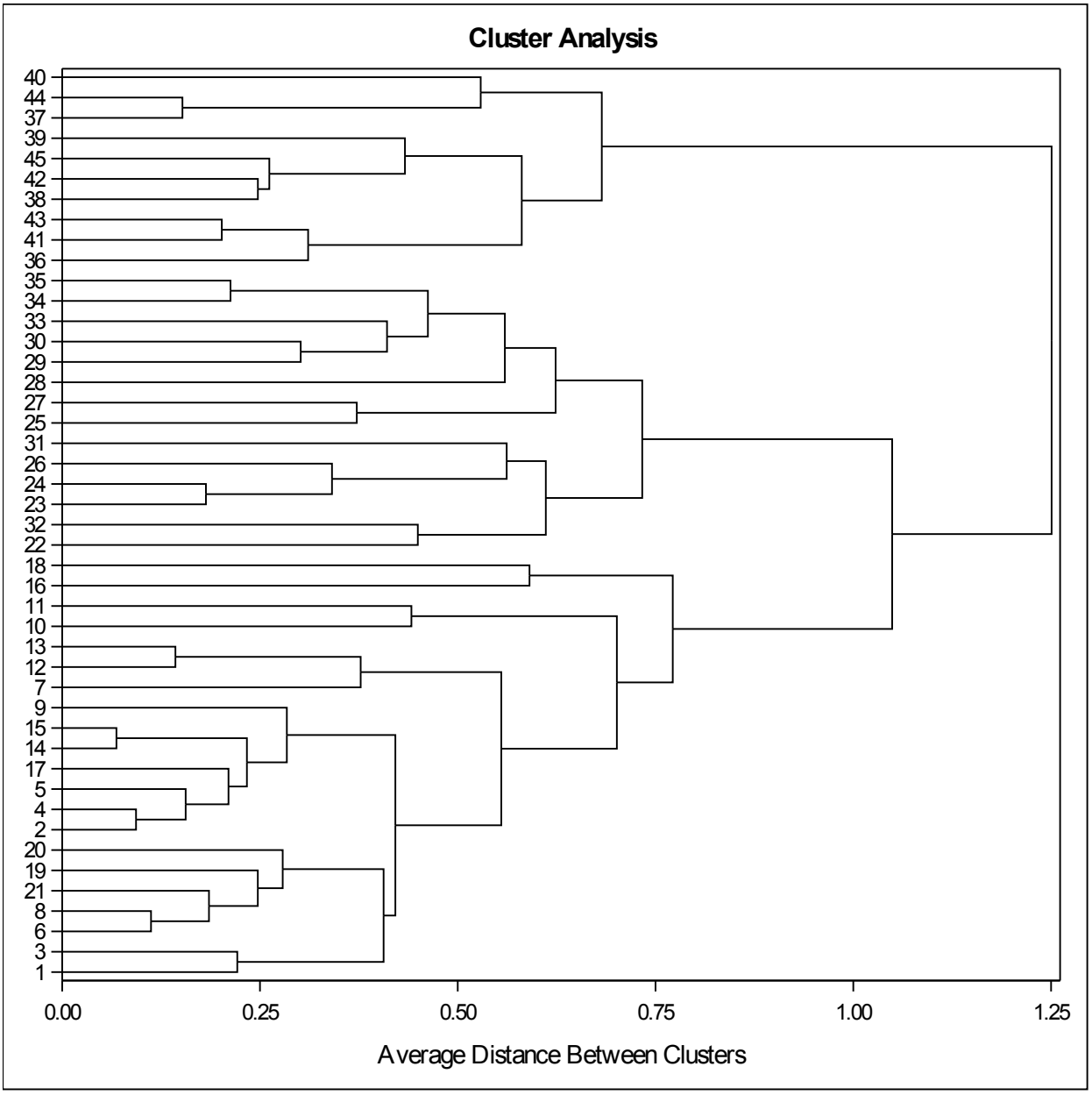
Based on the above table, kiln site 1 is separated out pretty well as cluster 1. Kiln site 2 and 3 are grouped together as cluster 2. And kiln site 4 and 5 are grouped together as cluster 3.

There are similarities with the PCA results. In PCA results, the types of oxide are significantly different between kiln site 1 and other four kiln sites. On the other hand, the types of oxide are similar in kiln site 2, 3 and kiln site 4, 5, respectively. In cluster analysis, kiln site 1 is separated out as a cluster while kiln site 2, 3 and 4, 5 are grouped as two different clusters. Therefore, the results of the two analyses are similar.

Exercise 4

a)





Based on the dendrogram, 3 clusters should be chosen to ensure the clusters have a comparatively large distance. Based on the pseudo F statistic, 3 clusters should be chosen. Based on the pseudo t-squared statistic, 3 clusters should be chosen. Based on ccc statistic, 3 clusters should be chosen. Therefore, choosing 3 clusters is appropriate.

The means analysis on variables by cluster is as below:

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

There are oxide differences between clusters. For Aluminium oxide, the means of cluster 1 and 2 are close and much larger than the mean of cluster 3. For Ferrous oxide, the means of cluster 1 and 3 are close and much larger than the mean of cluster 2. For Magnesium oxide, Potassium oxide, and Manganese oxide, means of the three clusters are different. The mean of cluster 3 is the largest while the mean of cluster 2 is the smallest. For Calcium oxide and Sodium oxide, means of the three clusters are different. The mean of cluster 1 is the largest while the mean of cluster 2 is the smallest. For Titanium oxide, means of the three clusters are different. The mean of cluster 2 is the largest while the mean of cluster 3 is the smallest. For Barium oxide, means of the three clusters are quite close.

b)

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

Based on the above table, kiln site 1 is separated out pretty well as cluster 1. Kiln site 2 and 3 are grouped together as cluster 3. And kiln site 4 and 5 are grouped together as cluster 2.

There are similarities with the PCA results. In PCA results, the types of oxide are significantly different between kiln site 1 and other four kiln sites. On the other hand, the types of oxide are similar in kiln site 2, 3 and kiln site 4, 5, respectively. In cluster analysis, kiln site 1 is separated out as a cluster while kiln site 2, 3 and 4, 5 are grouped as two different clusters. Therefore, the results of the two analyses are similar.

The results of Exercise 3 and 4 in matching the original kilns are the same. They both well separate out kiln site 1 and group together kiln site 2, 3 and 4, 5, respectively. It might because of the average linkage used in the two exercises. For this dataset, the average linkage makes the relative changes in the clustering variables be similar with the absolute changes in the clustering variables. Therefore, the two exercises give the same results.