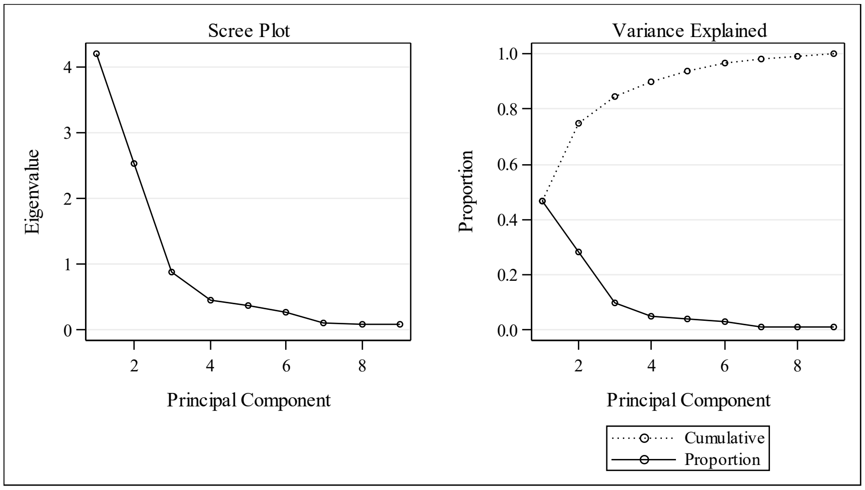
ST448 HW5 Solution

Exercise1

(a) We would check three criteria to decide the number of PCs to keep. First, to retain 80% of the variation in the original variables, we need to keep the first 3 principal components. Second, we would choose 2 components based on the average eigenvalue criterion. This is correlation-based PCA, so average eigenvalue is 1. Lastly, the scree plot becomes fairly flat after the third component, so we would choose 3 based on scree plot.

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



(b) Based on the 80% criterion, we chose 3 components.

The relatively large positive loadings in PC 1 (highlighted with green) are for K, Mn, Mg, and Fe, and the dominant negative values (highlighted with dark pink) are for Al and Ti. Therefore, PC 1 contrasts the first 4 oxides with Al and Ti oxides. It implies that pots with large positive PC 1 score would tend to have more K, Mn, Mg, and Fe oxides and less of the Al and Ti oxides, and the opposite would be true for large negative values.

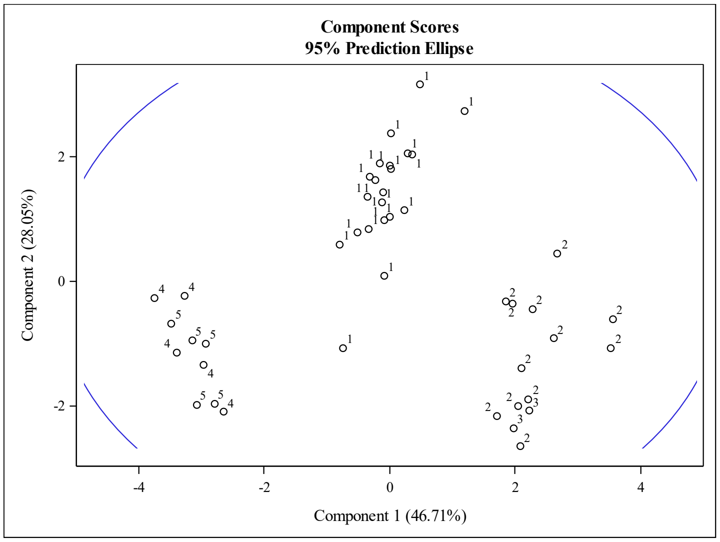
For PC 2, large positive values are for Ca, Na, Fe, Ba, Al, and Ti oxides and the most prominent negative value is for Mg oxide. This indicates that pots with large negative PC 2 score would tend to have much more Mg oxide relative to the other 6 oxides having large positive coefficients. Pots with large positive PC 2 score would have much less Mg oxide relative to the other 6 oxides.

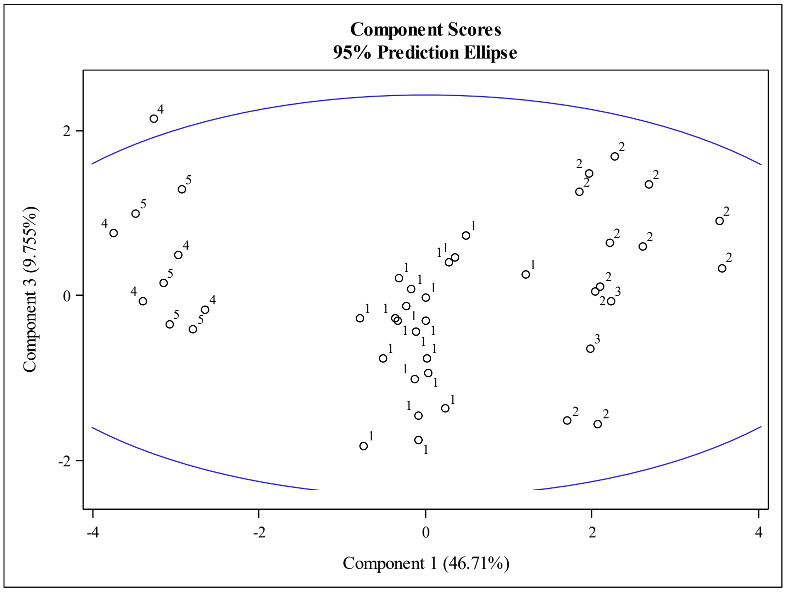
The PC 3 is mostly picking up Ba and Ca oxides and represents contrast between the two. Large positive PC 3 values would tend to indicate a greater amount of Ba oxide relative Ca oxide levels. Large negative values would indicate less Ba oxide relative to Ca oxides.

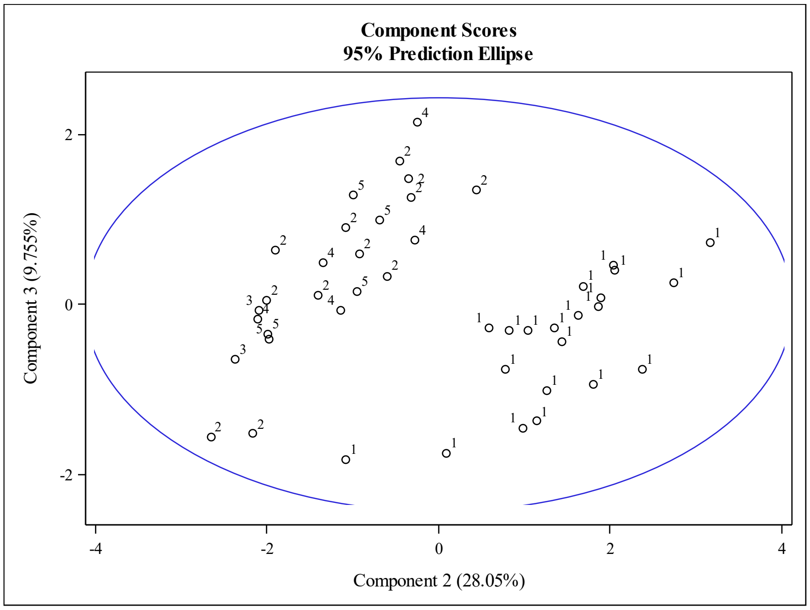
Eigenvectors									
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9
Al	-.348275	0.327856	0.119016	-.033283	0.321247	0.776634	0.016917	0.219583	0.032872
Fe	0.327132	0.395211	-.264433	-.019252	0.343312	0.045868	-.244490	-.504300	-.482117
Mg	0.434611	-.189543	0.150914	0.055441	0.280789	0.010166	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

(c) First we see PC 1 vs. PC 2 score plot. In terms of PC 1, we see that for component 1 kiln 1 has values right around 0, kilns 2 and 3 have positive values, and kilns 4 and 5 have negative values. This indicates that kiln 1 pots have an average contrast of K, Mn, Mg, and Fe to Al and Ti oxides. Pots from kilns 2 and 3 tend to be clustered together in terms of PC1, and it implies that they have lower amounts of Al and Ti oxides relative to the other 4 oxides. Lastly, pots from kilns 4 and 5 tend to higher levels of Al and Ti oxides relative to the other 4 oxides listed. In terms of PC 2, we see that kiln 1 tends to have higher values and the other four kilns have lower values of PC 2. This indicates lower levels of Mg oxide in kiln 1 pots and higher levels of Mg oxide in the other four kilns relative to Ca, Na, Fe, Ba, Al, and Ti oxides.

For component 3, the separation is less noticeable. In most cases, kiln 2 and 4 tend to have more Ba oxide, though there are two kiln 2 pots with noticeably less Ba oxide compared to Ca and Fe oxides than the average pot. Kiln 1 seems to have a slight downward shift, so on average there seems to be slightly more Ca oxide compared to Ba oxide. As mentioned, there is a quite a bit of spread and that spread covers the positive and negative range, so this particular contrast would not be very useful for distinguishing between the kilns.


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

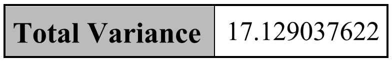

Component Scores
95% Prediction Ellipse
Component 3 (9.755%)
Component 1 (46.71%)
2
0
-2
-4
-2
0
2
4


Component Scores
95% Prediction Ellipse
Component 3 (9.755%)
Component 2 (28.05%)
2
0
-2
-4
-2
0
2
4

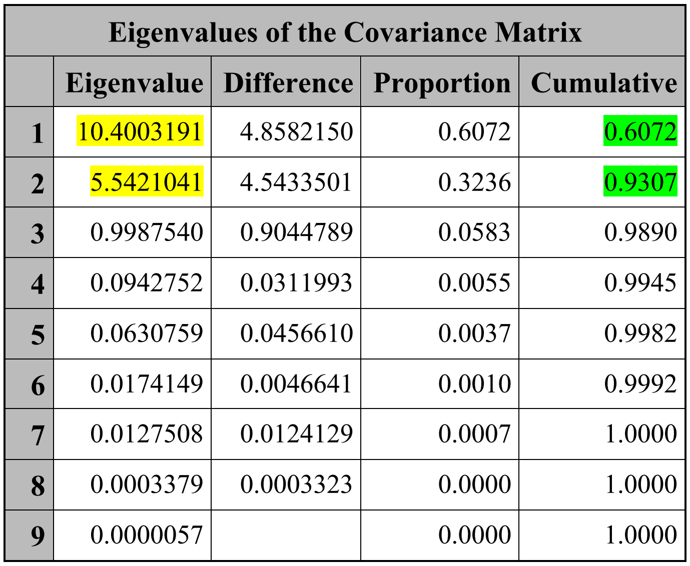
Exercise 2

(a) We perform covariance-based PCA and repeat the analysis. In this case, variables are not rescaled so the magnitude of variation for each variable will come into play.

Again, we need to decide the number of PCs to keep based on the three criteria. First, to retain 80% of the variation in the original data, we would need to keep the first two components. Second, the averaged eigenvalue is the total variation divided by the number of variables, which is 17.13/9=1.9. Thus, we keep two PCs that have eigenvalues larger than the average. Third, the scree plot becomes almost flat after the third component, so this would indicate 3 components should be kept. An argument could also be made that, relative to the magnitudes of the first two components, the scree plot is fairly flat starting at the third component, in which case 2 would be a good choice for number of components.

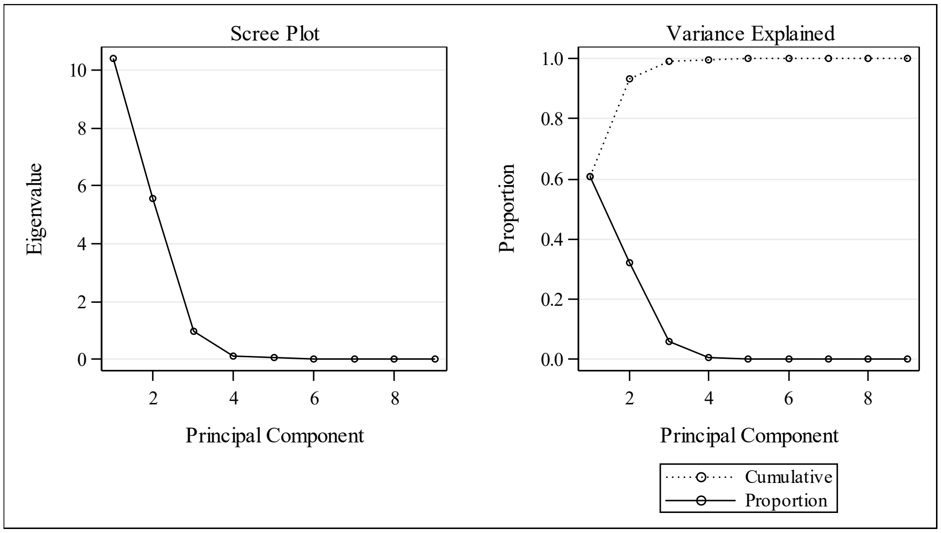


Total Variance	17.129037622



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



(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

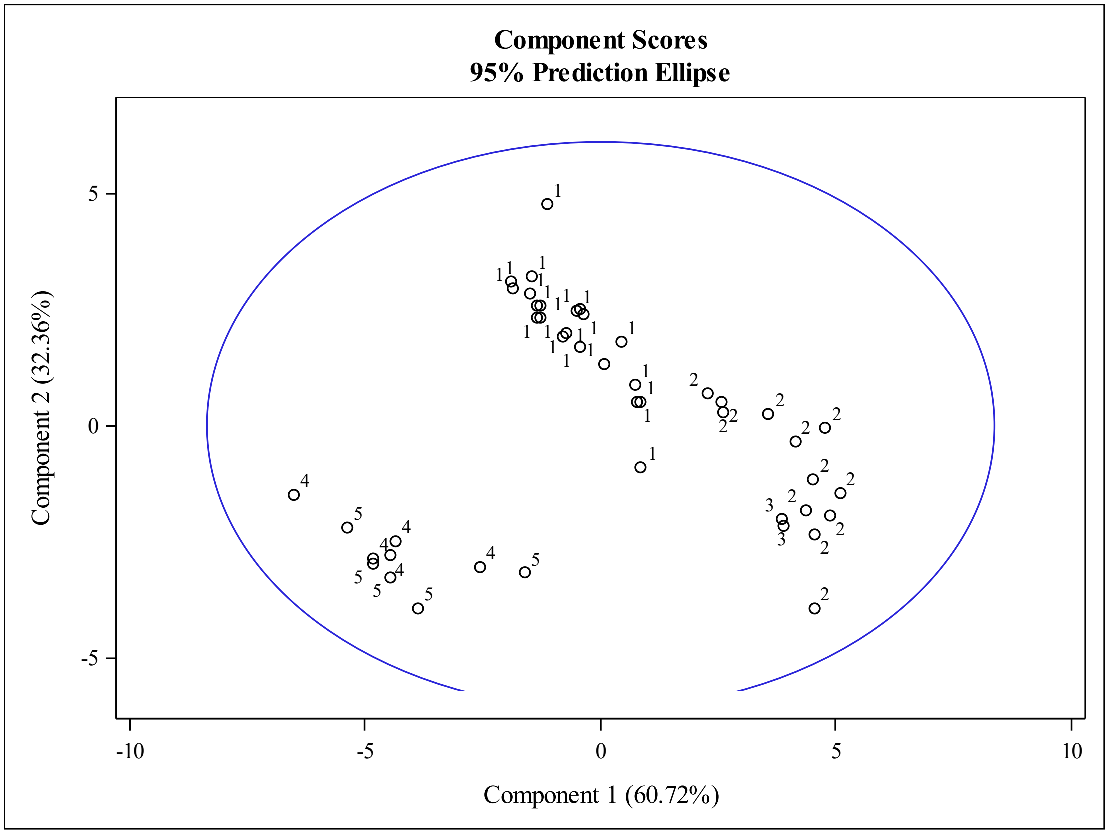
Based on 80% we chose the first two components.

The first component has large positive values for Mg, Fe and the negative value for Al oxide. This indicates that pots with large negative PC 1 score tend to have more Al oxide relative to Mg and Fe. The second PC is mostly picking up on Al and Fe oxides. Larger values of principal component 2 would tend to indicate more of those oxides and smaller values would indicate less of those two types of oxides.

(c) From the PC 1 vs. PC 2 score plot, we can see that kiln 1 pots tend to have average PC 1 values and positive principal component 2 values. Kiln 2 and 3 pots tend to have positive PC 1 values and negative PC 2 values. Kiln 4 and 5 pots have negative values for both components.

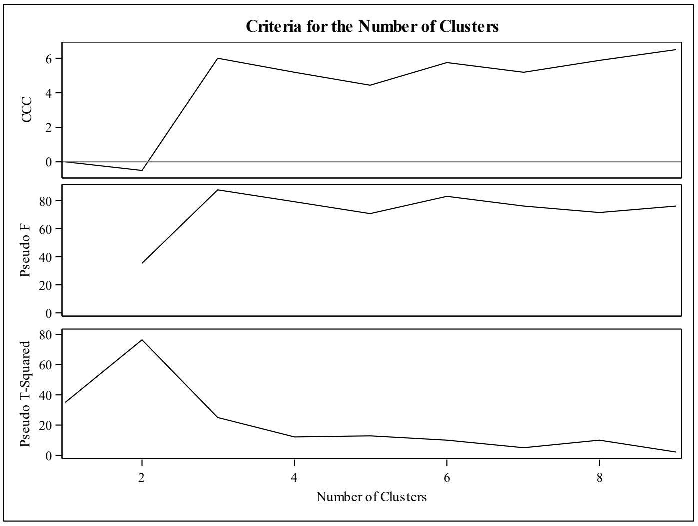
For the interpretation, first we see PC 2 scores. We can infer that pots from kiln 2,3,4 and 5 have relatively lower level of Al and Fe oxides. Based on PC1 scores, we can find further information that, pots from kiln 2 and 3 would have relatively higher level of Mg compared to pots from kiln 4 and 5, and pots from kiln site 1 have a roughly typical contrast of Al to FE and Mg with some pots having a slightly greater amount of Al.

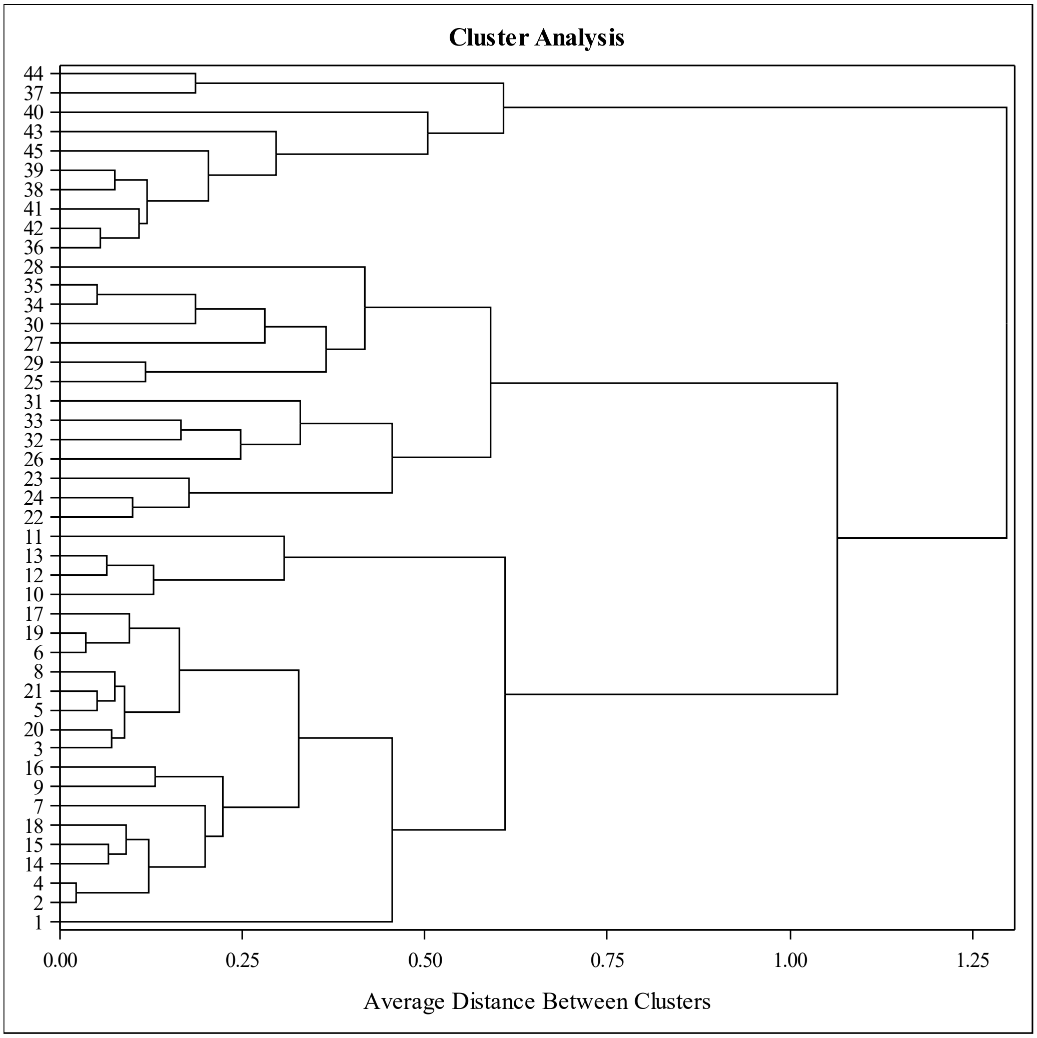
Comparing correlation and covariance results, we see that the same groupings of kilns are found. We also see that only 2 components are needed to explain 80% of the oxide variation in the covariance case, while 3 were needed in the correlation-based analysis. The oxides with higher variance are prominent in the covariance-based result, while oxides with very small variances have very little impact in the covariance based analysis.



Exercise3

(a) From the CCC, and pseudo F and t-squared plots, 3 looks like a good choice for the number of clusters based on each criterion. We see peaks at clusters 3 for the CCC and pseudo F statistics, and we see a pretty big jump from 3 clusters to 2 clusters for the pseudo t-squared statistic. The dendrogram also indicates 3 as a good choice for the number of clusters.



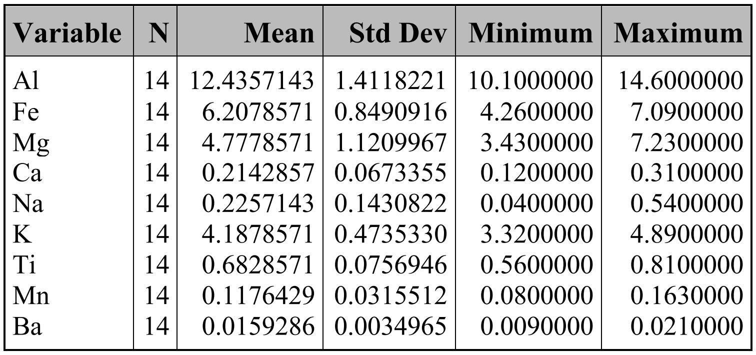


Comparing the clusters, there are higher values of Al oxide in clusters 1 and 3, higher values of Fe oxide in clusters 1 and 2, higher values of Mg oxide in cluster 2, higher values of Ca oxide in cluster 1, and lower values of K oxide in cluster 3.

CLUSTER=1

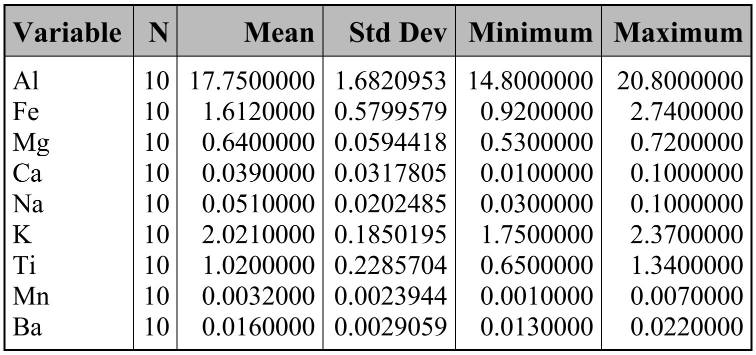
Variable	N	Mean	Std Dev	Minimum	Maximum
Al	21	16.9190476	1.5442212	13.7000000	18.9000000
Fe	21	7.4285714	0.6684331	5.8300000	9.5200000
Mg	21	1.8423810	0.2070243	1.5000000	2.3300000
Ca	21	0.9390476	0.2919230	0.6600000	1.7300000
Na	21	0.3461905	0.1634771	0.1200000	0.8300000
K	21	3.1028571	0.2247697	2.2500000	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

(b) A frequency analysis of the clusters and kilns shows that cluster 1 matches to kiln 1, cluster 2 contains the pots from kilns 2 and 3, and cluster 3 contains the pots from kilns 4 and 5. The clustering matches kiln groups very well, and is consistent with the groupings we saw in the score plots in the PCA analyses. The differences in characteristics across clusters also matches with the differences noted in the principal component values for pots from each kiln.

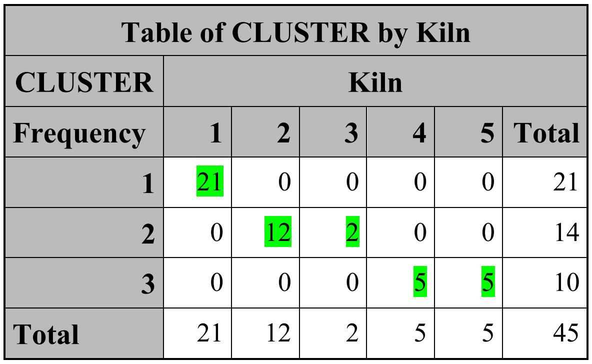
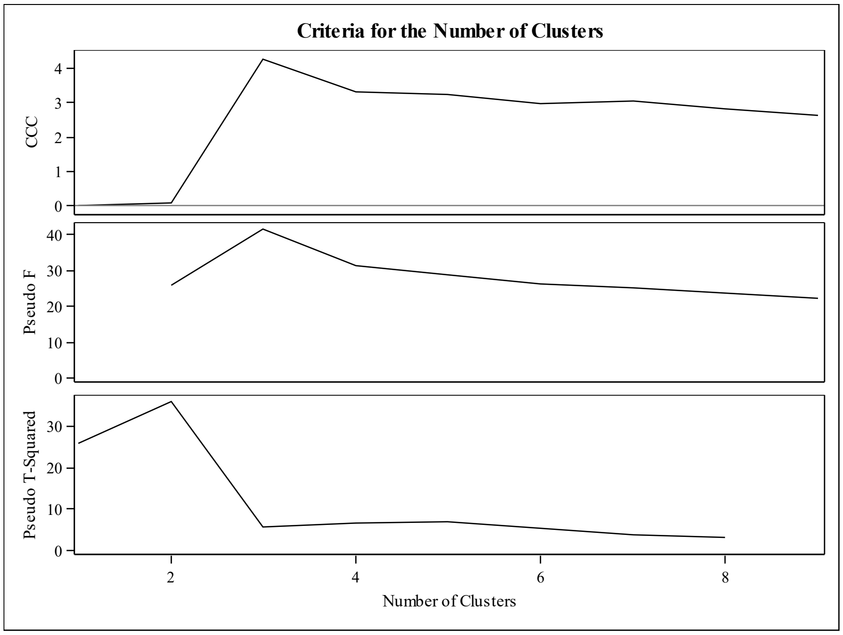
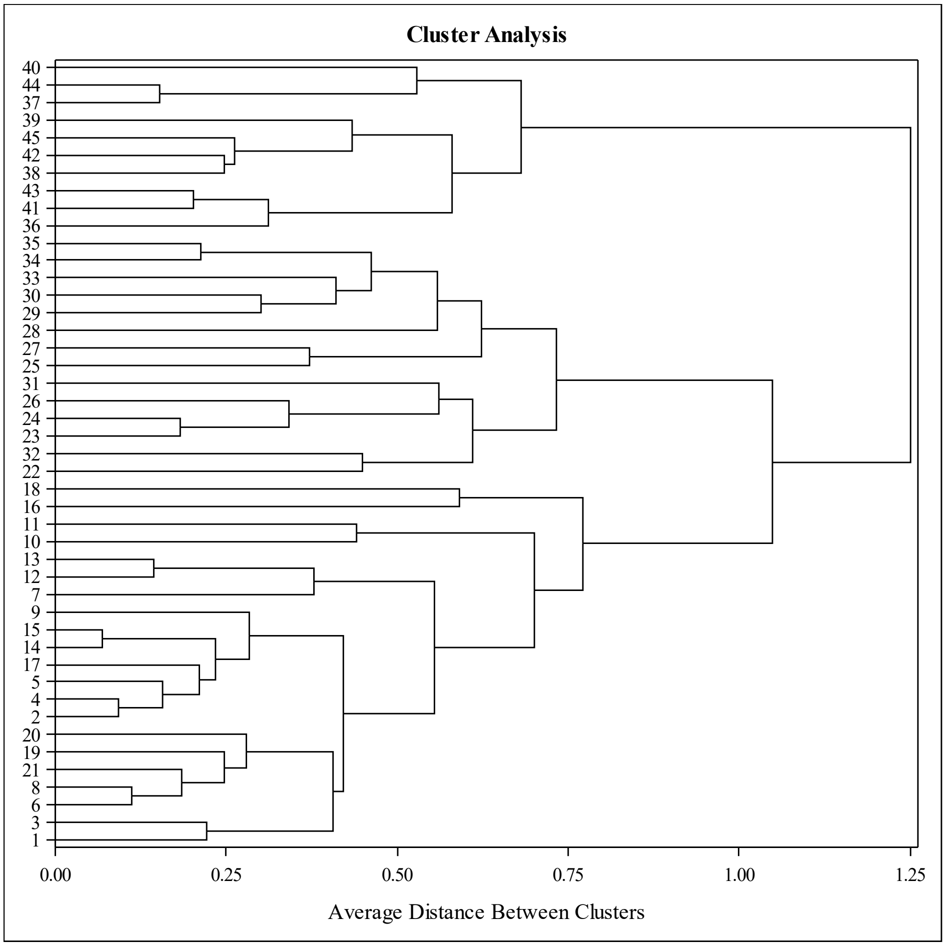


Table of CLUSTER by Kiln						
CLUSTER	Kiln					
Frequency	1	2	3	4	5	Total
1	21	0	0	0	0	21
2	0	12	2	0	0	14
3	0	0	0	5	5	10
Total	21	12	2	5	5	45

Exercise4

(a) We repeat the analysis with standardized variables, we again choose 3 clusters based on the peaks in the CCC and pseudo F statistics and the large jump from 3 to 2 clusters for the pseudo t-squared statistics. The choice is more obvious this time as the CCC and pseudo F statistics now are smaller for more than 3 clusters. The dendrogram also indicates 3 as a good choice for the number of clusters. The dendrogram is fairly similar to the one we saw in the unstandardized case in exercise 3.


Criteria for the Number of Clusters
CCC
0
1
2
3
4
Pseudo F
0
10
20
30
40
Pseudo T-Squared
0
10
20
30
2
4
6
8
Number of Clusters


Cluster Analysis
40
44
37
39
45
42
38
43
41
36
35
34
33
30
29
28
27
25
31
26
24
23
32
22
18
16
11
10
13
12
7
9
15
14
17
5
4
2
20
19
21
8
6
3
1
0.00
0.25
0.50
0.75
1.00
1.25
Average Distance Between Clusters

Comparing the clusters, the clusters are fairly similar to those in exercise 3. In fact they are the same clusters with the change in the cluster numbers. Cluster 3 here is cluster 2 in the previous analysis and cluster 2 here was cluster 3 in the previous analysis. The characteristics of the clusters are therefore the same as before, except with comments about cluster 3 now being for cluster 2 and vice versa. In general, cluster analyses based on original scale and standardized variables do not have the same result. This is a special example.

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

(b) Cluster 1 corresponds to kiln 1, cluster 2 corresponds to kilns 4 and 5, and cluster 3 corresponds to kilns 4 and 5. As before, the differences of cluster features are consistent with the principal component results in the earlier exercises. The standardized and unstandardized cluster analysis results match the original kilns equally well because they give us the same grouping of observations in this case.

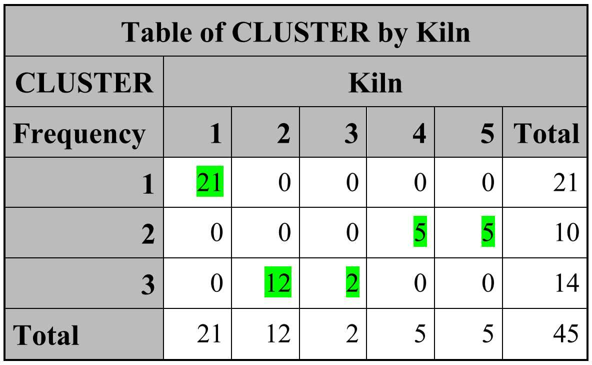


Table of CLUSTER by Kiln

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