Exercise 1a):

The PRINCOMP Procedure

Observations	47
Variables	11

Simple Statistics							
	Age	**Ed**	**Ex0**	**LF**	**M**	**N**	**NW**
Mean	138.5744681	105.6382979	85.00000000	561.1914894	983.0212766	36.61702128	101.1276596
StD	12.5676339	11.1869985	29.71897359	40.4118140	29.4673654	38.07118801	102.8288187

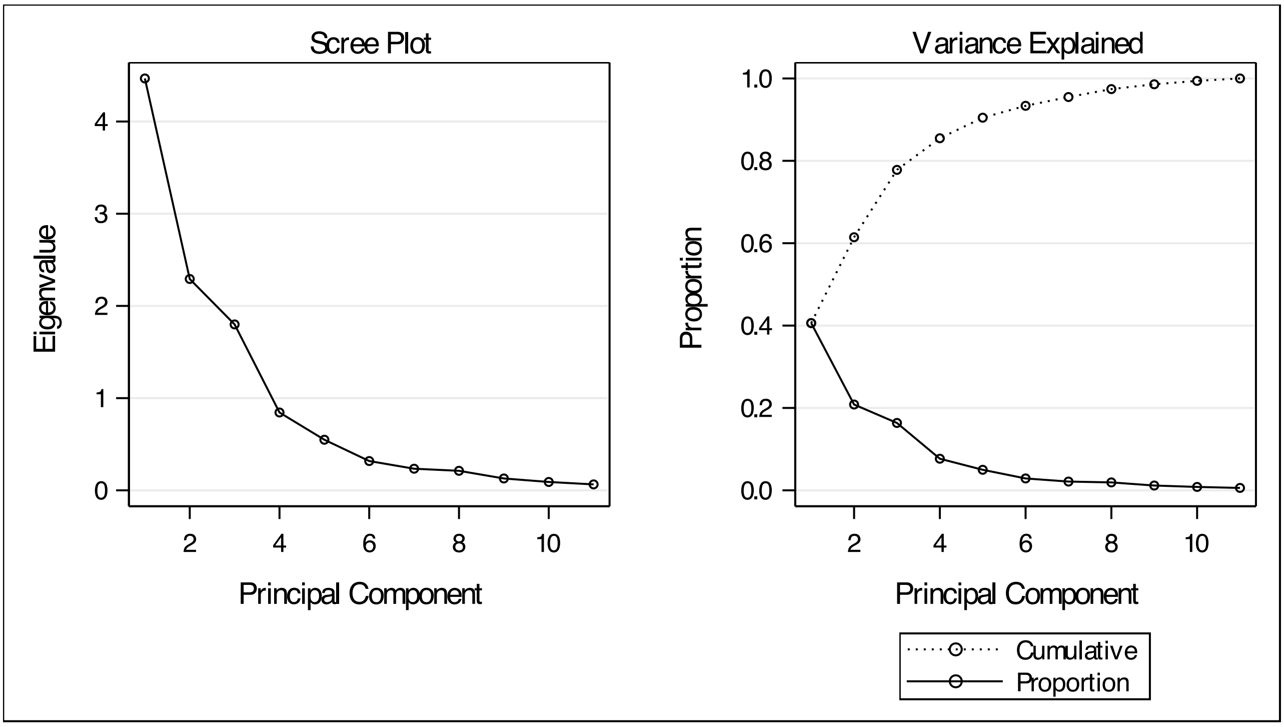
Simple Statistics				
	U1	**U2**	**W**	**X**
Mean	95.46808511	33.97872340	525.3829787	194.0000000
StD	18.02878262	8.44544992	96.4909442	39.8960606

Correlation Matrix											
	Age	**Ed**	**Ex0**	**LF**	**M**	**N**	**NW**	**U1**	**U2**	**W**	**X**
Age	1.0000	-.5302	-.5057	-.1609	-.0287	-.2806	0.5932	-.2244	-.2448	-.6701	0.6392
Ed	-.5302	1.0000	0.4830	0.5612	0.4369	-.0172	-.6649	0.0181	-.2157	0.7360	-.7687
Ex0	-.5057	0.4830	1.0000	0.1215	0.0338	0.5263	-.2137	-.0437	0.1851	0.7872	-.6305
LF	-.1609	0.5612	0.1215	1.0000	0.5136	-.1237	-.3412	-.2294	-.4208	0.2946	-.2699
M	-.0287	0.4369	0.0338	0.5136	1.0000	-.4106	-.3273	0.3519	-.0187	0.1796	-.1671
N	-.2806	-.0172	0.5263	-.1237	-.4106	1.0000	0.0952	-.0381	0.2704	0.3083	-.1263
NW	0.5932	-.6649	-.2137	-.3412	-.3273	0.0952	1.0000	-.1565	0.0809	-.5901	0.6773
U1	-.2244	0.0181	-.0437	-.2294	0.3519	-.0381	-.1565	1.0000	0.7459	0.0449	-.0638
U2	-.2448	-.2157	0.1851	-.4208	-.0187	0.2704	0.0809	0.7459	1.0000	0.0921	0.0157
W	-.6701	0.7360	0.7872	0.2946	0.1796	0.3083	-.5901	0.0449	0.0921	1.0000	-.8840
X	0.6392	-.7687	-.6305	-.2699	-.1671	-.1263	0.6773	-.0638	0.0157	-.8840	1.0000

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.46699135	2.17472428	0.4061	0.4061
2	2.29226707	0.49290960	0.2084	0.6145
3	1.79935747	0.95506483	0.1636	0.7781
4	0.84429264	0.29537868	0.0768	0.8548
5	0.54891395	0.23116974	0.0499	0.9047
6	0.31774422	0.08323077	0.0289	0.9336
7	0.23451344	0.02270468	0.0213	0.9549
8	0.21180877	0.08302014	0.0193	0.9742
9	0.12878863	0.03795696	0.0117	0.9859

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
10	0.09083167	0.02634089	0.0083	0.9941
11	0.06449078		0.0059	1.0000

Eigenvectors											
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7	Prin8	Prin9	Prin10	Prin11
Age	-.360083	-.202052	-.023821	0.280333	0.437093	0.604564	0.198909	-.347243	0.165242	-.063955	0.047705
Ed	0.412649	-.201179	0.021906	0.024890	0.069270	0.120185	0.440819	0.485458	0.558524	0.148122	-.070897
Ex0	0.333346	0.245893	-.245985	0.375312	0.354161	-.166331	-.319844	0.015903	0.219065	-.568308	0.020907
LF	0.216510	-.432770	-.019098	0.422942	-.501450	-.198616	0.278542	-.417126	-.027027	-.204431	0.039818
M	0.161412	-.305691	0.477670	0.453784	0.119909	0.087151	-.463341	0.199381	-.211443	0.281507	-.214580
N	0.096028	0.423702	-.367382	0.334714	-.449058	0.524315	-.028367	0.164744	-.151113	0.144061	-.127119
NW	-.355977	0.130789	-.162528	0.465262	0.240490	-.442682	0.416851	0.288919	-.267347	0.174650	-.016473
U1	0.054270	0.299353	0.626794	0.044940	-.071474	0.187414	0.299433	0.184749	-.232219	-.419871	0.343962
U2	0.010990	0.529168	0.376451	0.142102	-.029661	-.177801	0.085052	-.432598	0.398736	0.282410	-.303873
W	0.439177	0.103624	-.104899	0.042769	0.229726	-.003169	0.021190	-.247690	-.154656	0.465774	0.654298
X	-.429102	-.022454	0.054178	0.206375	-.307636	-.087908	-.315643	0.191087	0.485978	0.075559	0.538001



We performed a PCA on all the possible predictors. We can see that we could keep first four components to retain 80% of the total variation from the original variables. We also could keep first three components(have eigenvalues greater than 1) based on the average eigenvalue(=1), and scree plot says says that we could keep first four components, since the plot becomes very flat starting at 5 for me.

Exercise 1b):

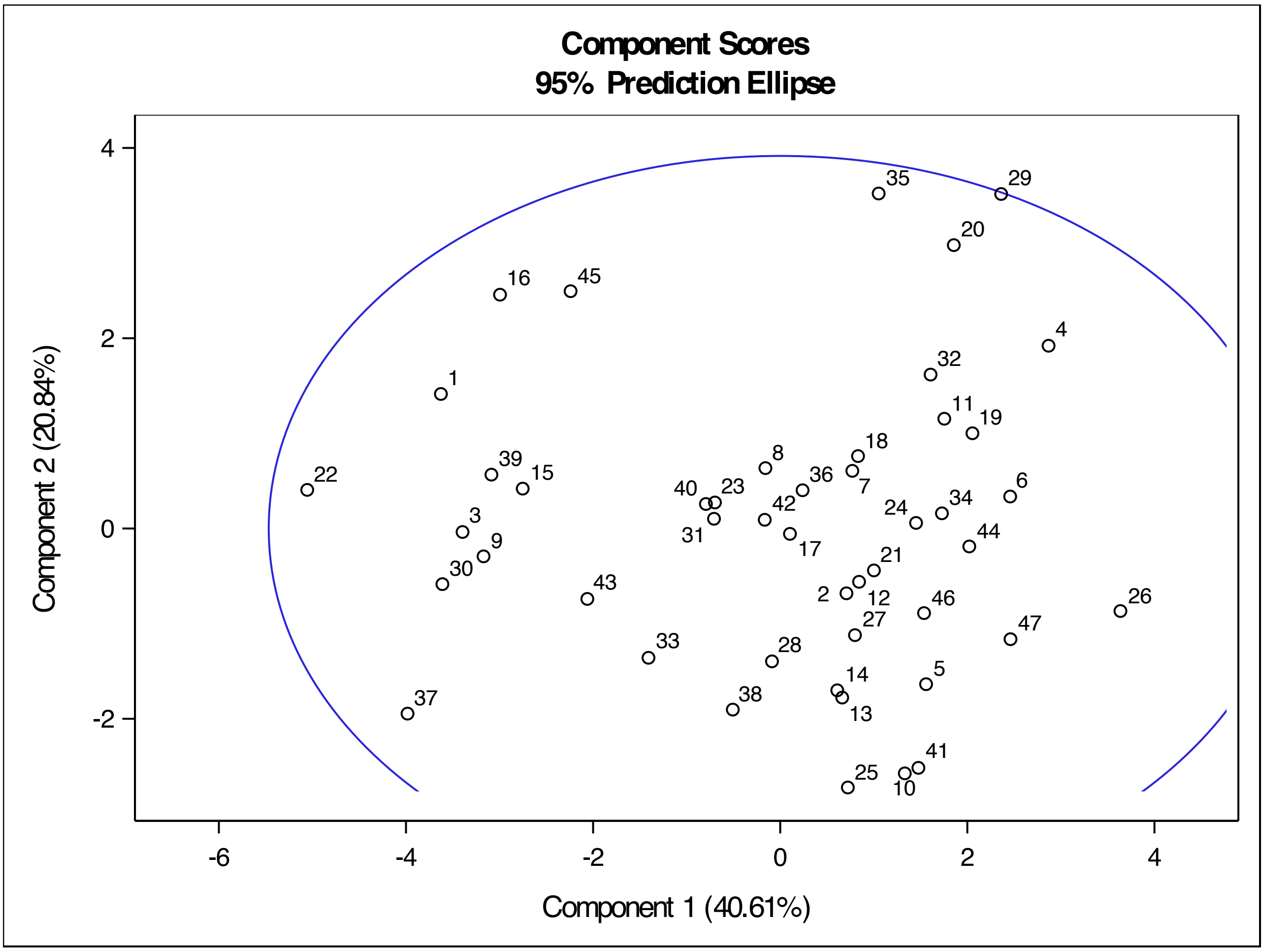
For component 1, on the positive side, W(wealth) and Ed(education level) have a little more impact than Ex0(police expenditure), LF(labor force), M(# of males). But none of them have important relationship to principal component 1. On the negative side, X(income inequality) has more impact than Age and NW(# of non-whites). But none of them are highly important.

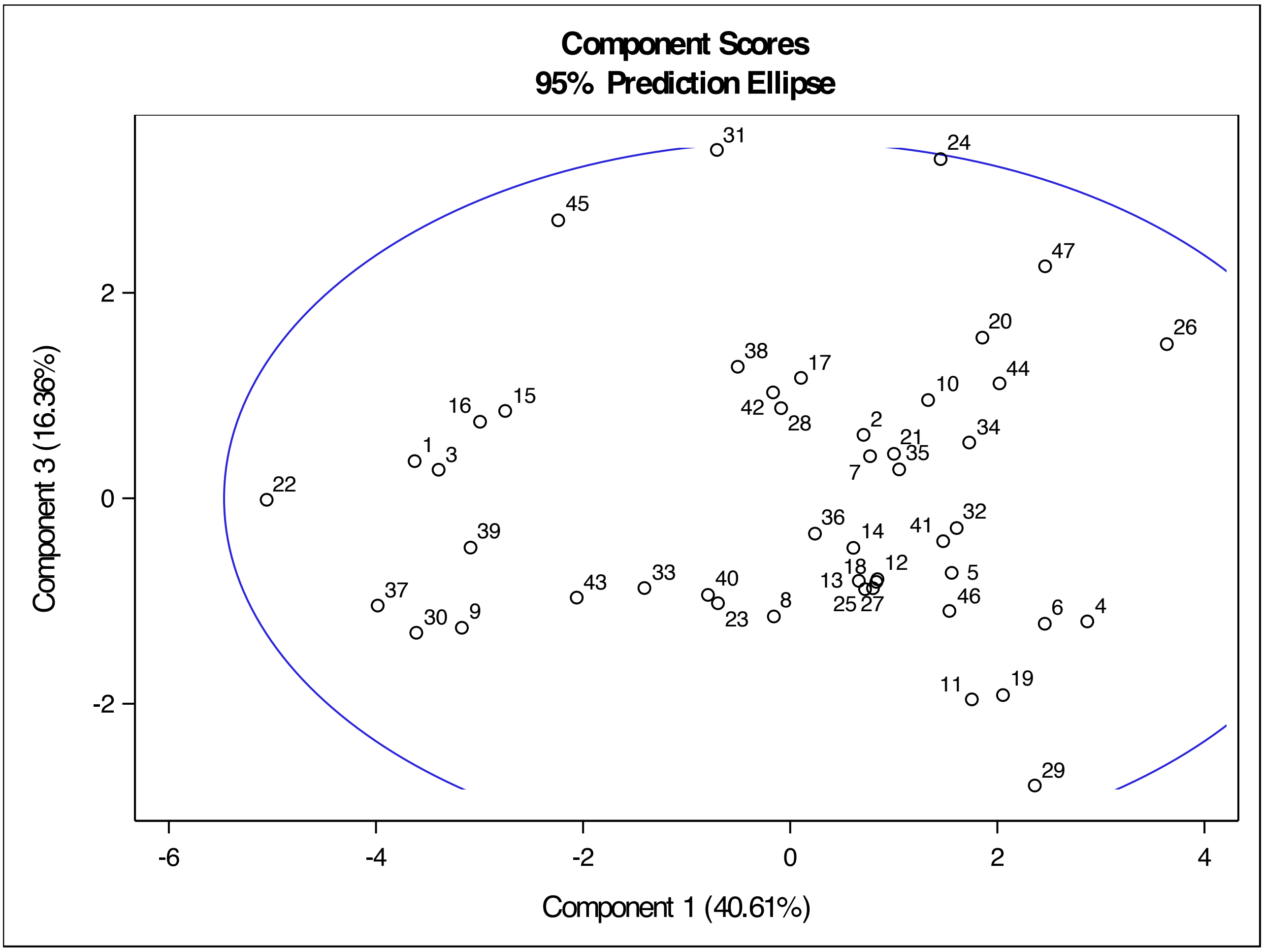
For component 2, on the positive side, U2 and N(state population size) have a little more impact than U1 and Ex0. On the negative side, LF and M are more important.

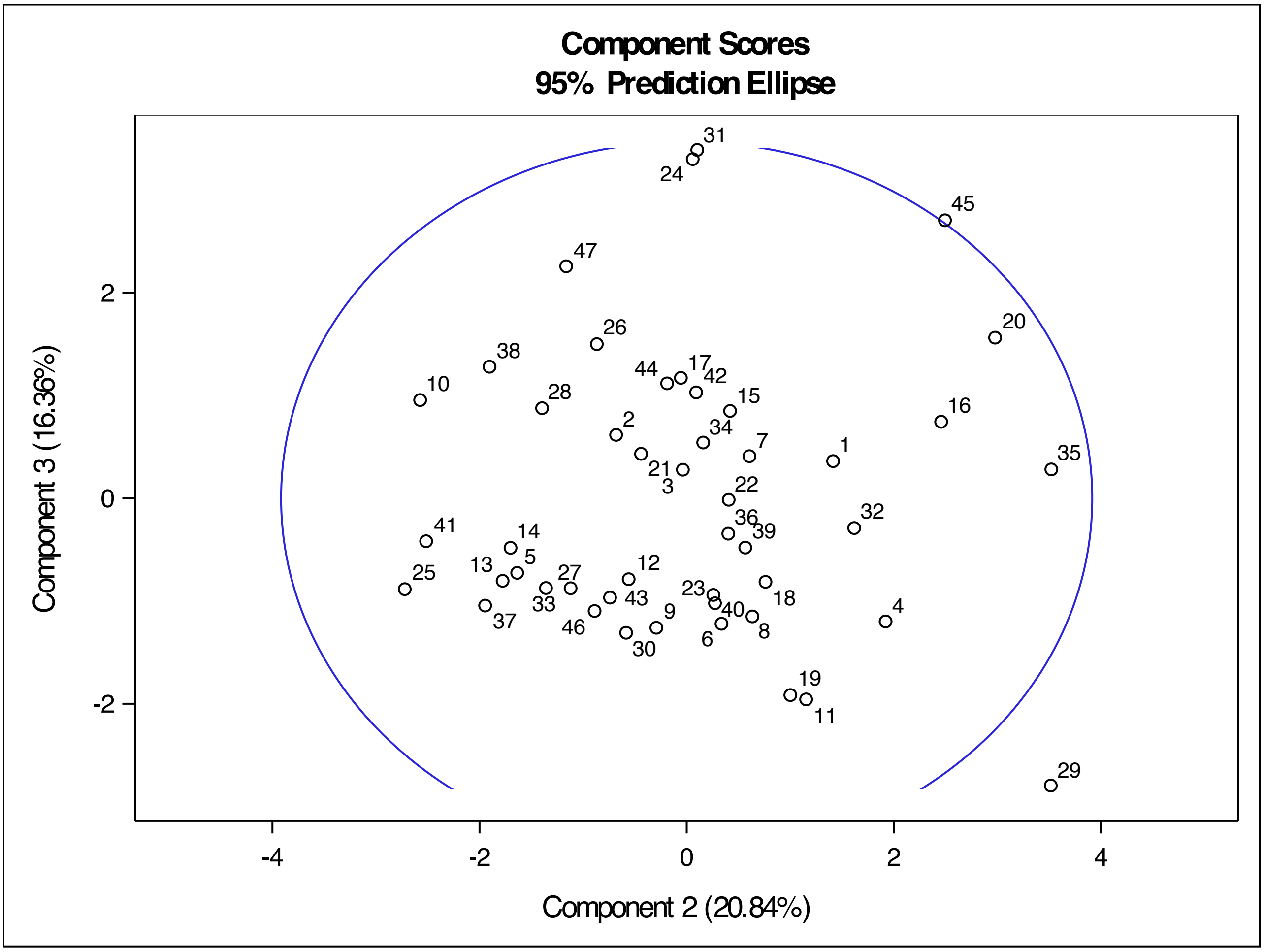
For component 3, on the positive side, U1 has more impact than M and U2. On the negative side, N and X0 have a similar impact.

For component 4, we have only positive coefficients. M, LF, Ex0 and NW have similar impact on principal component 4. If one of them goes up, principal component 4 gets higher.

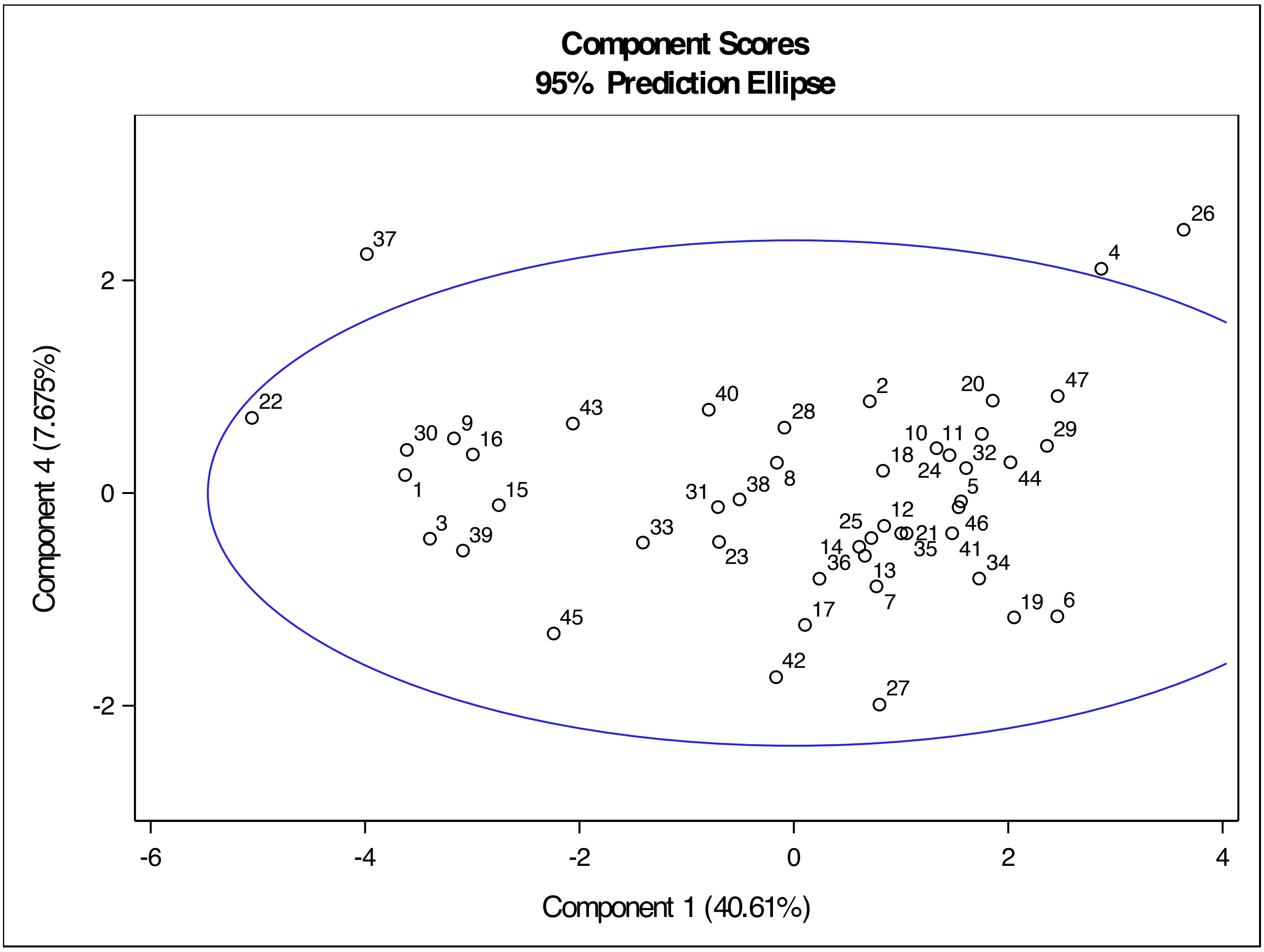
Exercise 1c):


Component Scores
95% Prediction Ellipse
Component 2 (20.84%)
Component 1 (40.61%)
4
2
0
-2
-6
-4
-2
0
2
4
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

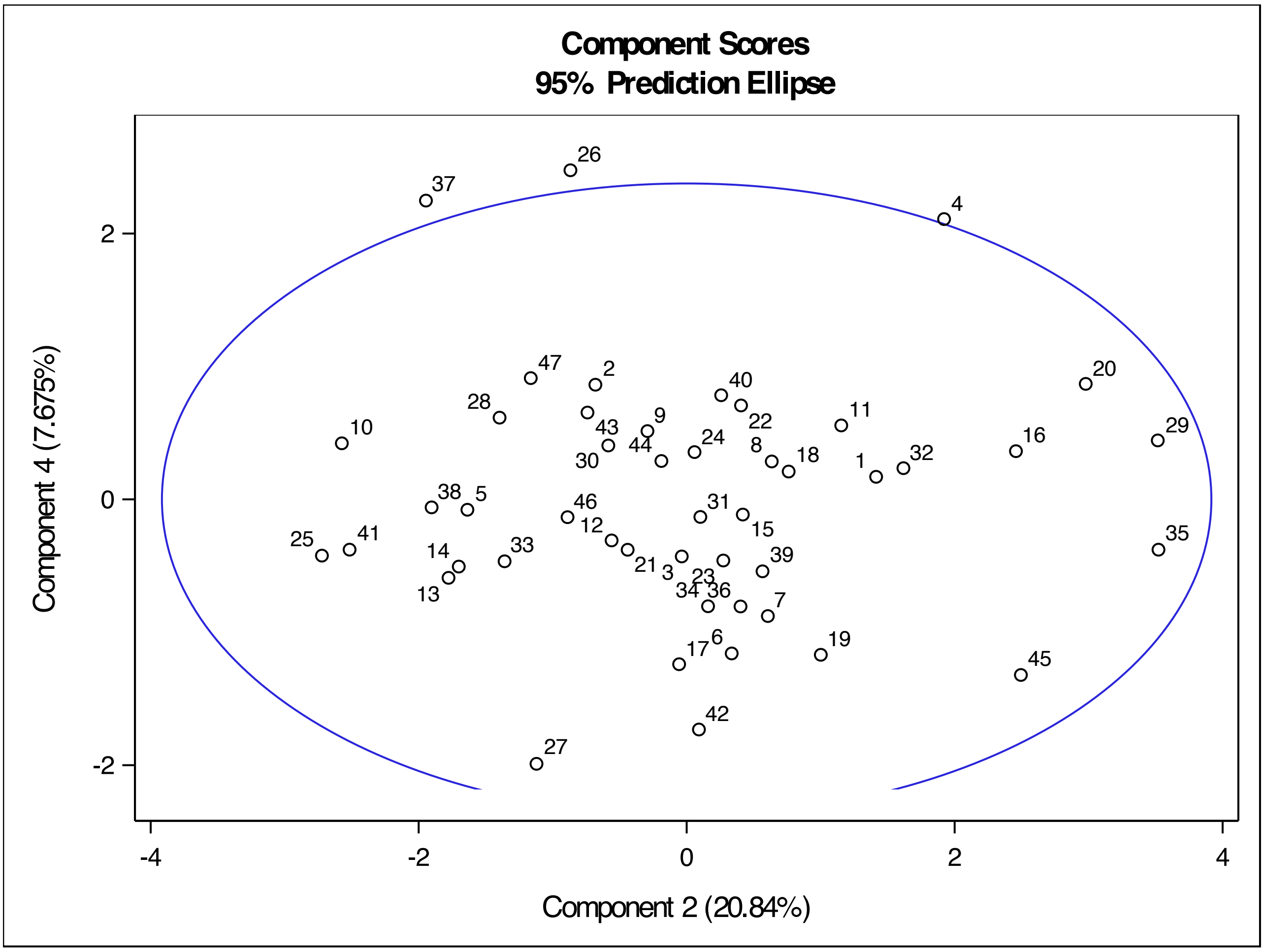

Component Scores
95% Prediction Ellipse
Component 3 (16.36%)
Component 1 (40.61%)
2
0
-2
-6
-4
-2
0
2
4
31
24
45
47
20
26
38
17
44
10
16
15
42
28
2
34
1
3
21
35
7
22
36
32
39
14
41
12
18
5
37
43
33
40
13
8
25
27
46
30
9
23
6
4
11
19
29

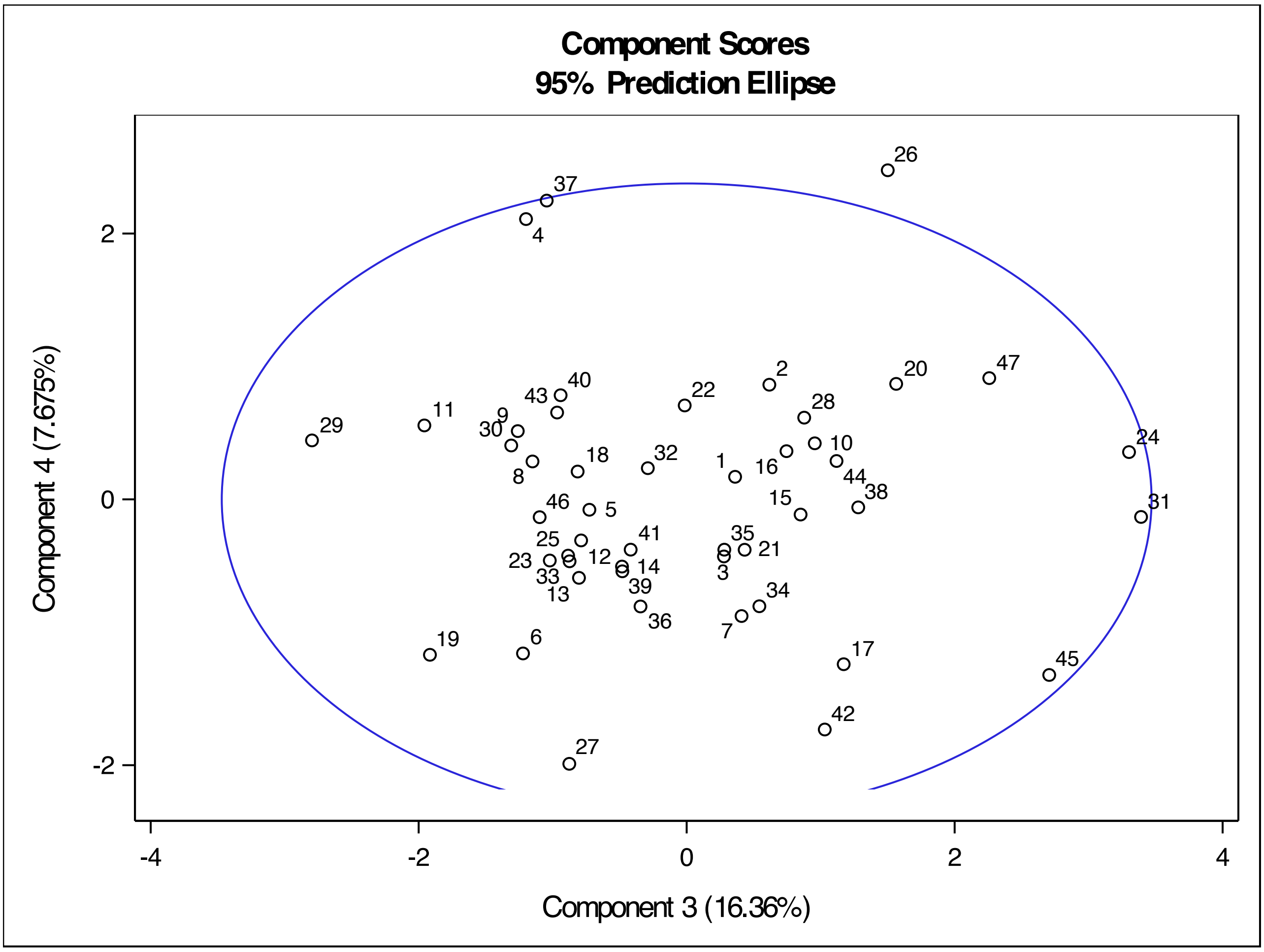


Component Scores
95% Prediction Ellipse
Component 3 (16.36%)
Component 2 (20.84%)
-4
-2
0
2
4
2
0
-2
31
24
45
47
20
26
38
10
28
44
17
42
15
16
2
34
7
1
35
21
3
22
32
41
14
36
39
5
13
25
27
12
33
43
23
18
37
46
9
40
8
4
30
6
19
11
29


Component Scores
95% Prediction Ellipse
Component 4 (7.675%)
Component 1 (40.61%)
-6
-4
-2
0
2
4
2
0
-2
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47

The PRINCOMP Procedure





The first plot shows that states 35, 20 and 29 have higher crime rates in general. And state 37 has lowest. From the second plot we can see that states 45, 31 and 24 have higher values in component 3. While in the third plot, we see that states 24 and 31 have also high values in component 3, but near average in component 2. State 29 has the lowest value in component 3 and has higher value component 2. In the fourth plot, we see that state 37, 4 and 26 have higher values in component 4. But state 37 has the lowest values for component 1 and states 4 and 26 have higher values for component 1. In the fifth plot states 37 and 26 have higher values in component 4 and states 20, 29, 35 higher in component 2. State 4 has higher rate in general. Sixth plot shows that states 4, 37, 26 have higher crime rates for component 4 and states 24, 31 have higher rates for component 3.

Exercise 2a):

The PRINCOMP Procedure

Observations	47
Variables	11

Simple Statistics							
	Age	**Ed**	**Ex0**	**LF**	**M**	**N**	**NW**
Mean	138.5744681	105.6382979	85.00000000	561.1914894	983.0212766	36.61702128	101.1276596
StD	12.5676339	11.1869985	29.71897359	40.4118140	29.4673654	38.07118801	102.8288187

Simple Statistics				
	U1	**U2**	**W**	**X**
Mean	95.46808511	33.97872340	525.3829787	194.0000000
StD	18.02878262	8.44544992	96.4909442	39.8960606

Covariance Matrix							
	Age	**Ed**	**Ex0**	**LF**	**M**	**N**	**NW**
Age	157.94542	-74.54857	-188.89130	-81.74283	-10.62118	-134.27521	766.59898
Ed	-74.54857	125.14894	160.56522	253.70120	144.02960	-7.33719	-764.84413
Ex0	-188.89130	160.56522	883.21739	145.91304	29.56522	595.45652	-653.08696
LF	-81.74283	253.70120	145.91304	1633.11471	611.56105	-190.27290	-1417.91628
M	-10.62118	144.02960	29.56522	611.56105	868.32562	-460.66559	-991.76364
N	-134.27521	-7.33719	595.45652	-190.27290	-460.66559	1449.41536	372.50648
NW	766.59898	-764.84413	-653.08696	-1417.91628	-991.76364	372.50648	10573.76596
U1	-50.83996	3.65125	-23.41304	-167.13506	186.94635	-26.16466	-290.03932
U2	-25.98751	-20.37743	46.45652	-143.60453	-4.65171	86.94820	70.26364
W	-812.55088	794.46762	2257.45652	1148.88159	510.68733	1132.41073	-5855.07169
X	320.50000	-343.06522	-747.56522	-435.13043	-196.43478	-191.82609	2778.65217

Covariance Matrix				
	U1	**U2**	**W**	**X**
Age	-50.83996	-25.98751	-812.55088	320.50000
Ed	3.65125	-20.37743	794.46762	-343.06522
Ex0	-23.41304	46.45652	2257.45652	-747.56522
LF	-167.13506	-143.60453	1148.88159	-435.13043
M	186.94635	-4.65171	510.68733	-196.43478
N	-26.16466	86.94820	1132.41073	-191.82609
NW	-290.03932	70.26364	-5855.07169	2778.65217
U1	325.03700	113.57539	78.03423	-45.91304
U2	113.57539	71.32562	75.03006	5.28261

The PRINCOMP Procedure

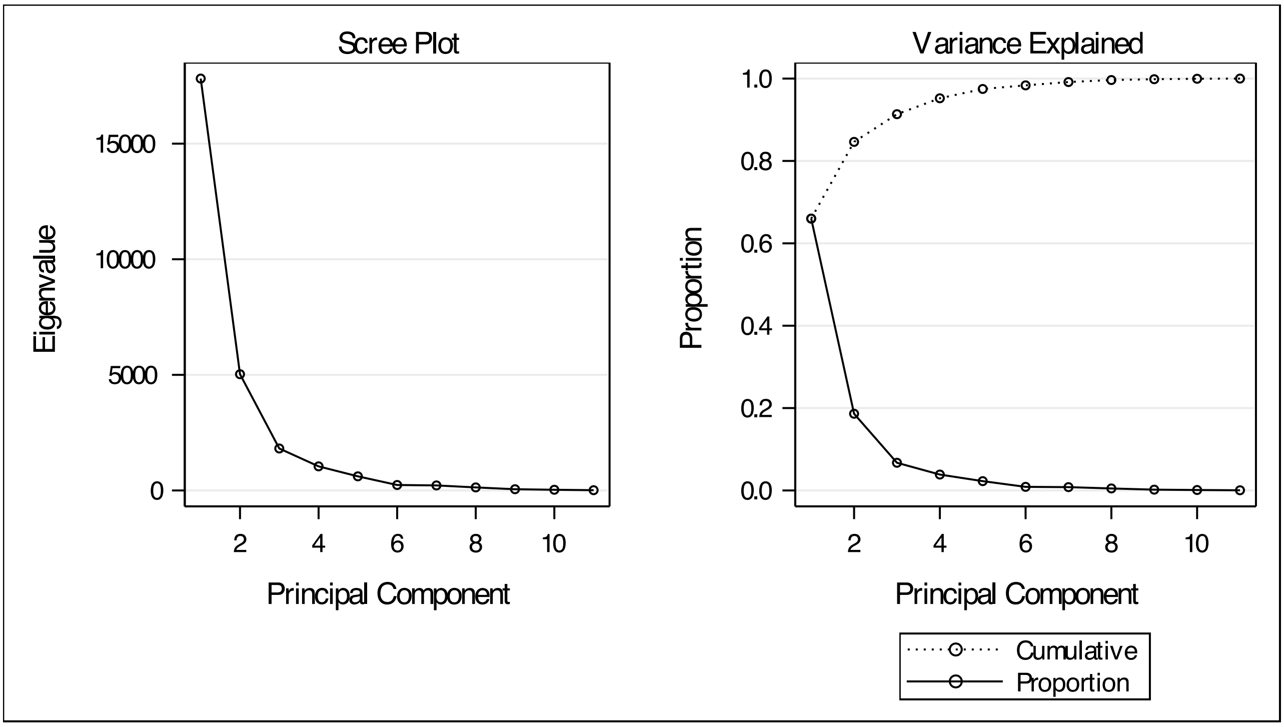
Covariance Matrix				
	U1	**U2**	**W**	**X**
W	78.03423	75.03006	9310.50231	-3403.04348
X	-45.91304	5.28261	-3403.04348	1591.69565

Total Variance	26989.493987

Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	17813.6423	12786.3209	0.6600	0.6600
2	5027.3214	3212.3413	0.1863	0.8463
3	1814.9801	772.8587	0.0672	0.9135
4	1042.1214	431.5081	0.0386	0.9522
5	610.6133	374.3341	0.0226	0.9748
6	236.2792	18.6489	0.0088	0.9835
7	217.6303	85.9636	0.0081	0.9916
8	131.6667	78.3670	0.0049	0.9965
9	53.2997	23.1296	0.0020	0.9984
10	30.1701	18.4004	0.0011	0.9996
11	11.7697		0.0004	1.0000

Eigenvectors											
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**	**Prin9**	**Prin10**	**Prin11**
Age	-.066773	-.022074	0.055475	-.030043	-.035107	0.113536	-.070427	-.413522	0.883506	0.129021	0.061333
Ed	0.067385	0.003307	0.084723	-.000591	-.014361	-.078415	-.085480	-.042748	-.174040	0.904764	0.351985
Ex0	0.128054	0.285116	-.003663	0.052509	0.141915	0.310785	-.715726	0.489327	0.156746	0.033306	-.073694
LF	0.117411	-.084716	0.752092	0.533170	-.135361	-.177699	0.100278	0.229490	0.118405	-.048506	-.011472
M	0.066817	-.125546	0.444409	-.198342	0.666035	0.104547	-.221007	-.417963	-.217547	-.125524	0.042207
N	0.034286	0.330353	-.363113	0.729070	0.336752	-.158376	-.019723	-.291905	-.045830	-.005662	0.009066
NW	-.682176	0.650578	0.265959	-.151145	-.029342	-.107219	0.045715	-.022939	-.052581	0.007784	-.002559
U1	0.014474	-.033491	-.075003	-.184701	0.541863	-.455184	0.259584	0.392193	0.276874	0.205150	-.343767
U2	-.000395	0.028379	-.072403	-.056234	0.207548	-.101175	0.123999	0.277081	0.145372	-.281779	0.863015
W	0.647310	0.588453	0.114513	-.200486	-.030831	0.190212	0.371730	-.069522	0.022700	0.001565	-.027199
X	-.265693	-.121175	-.004809	0.202167	0.254147	0.742751	0.439486	0.198029	0.006241	0.154780	-.034652

The PRINCOMP Procedure



We now performed a covariance-based PCA on all the possible predictors. We can see that we could keep first two components to retain 80% of the total variation from the original variables. We also could keep first two components(have eigenvalues greater than 2453.590362) based on the average eigenvalue(=2453.590362), and scree plot says says that we could keep first two components, since the plot becomes very flat starting at 3 for me.

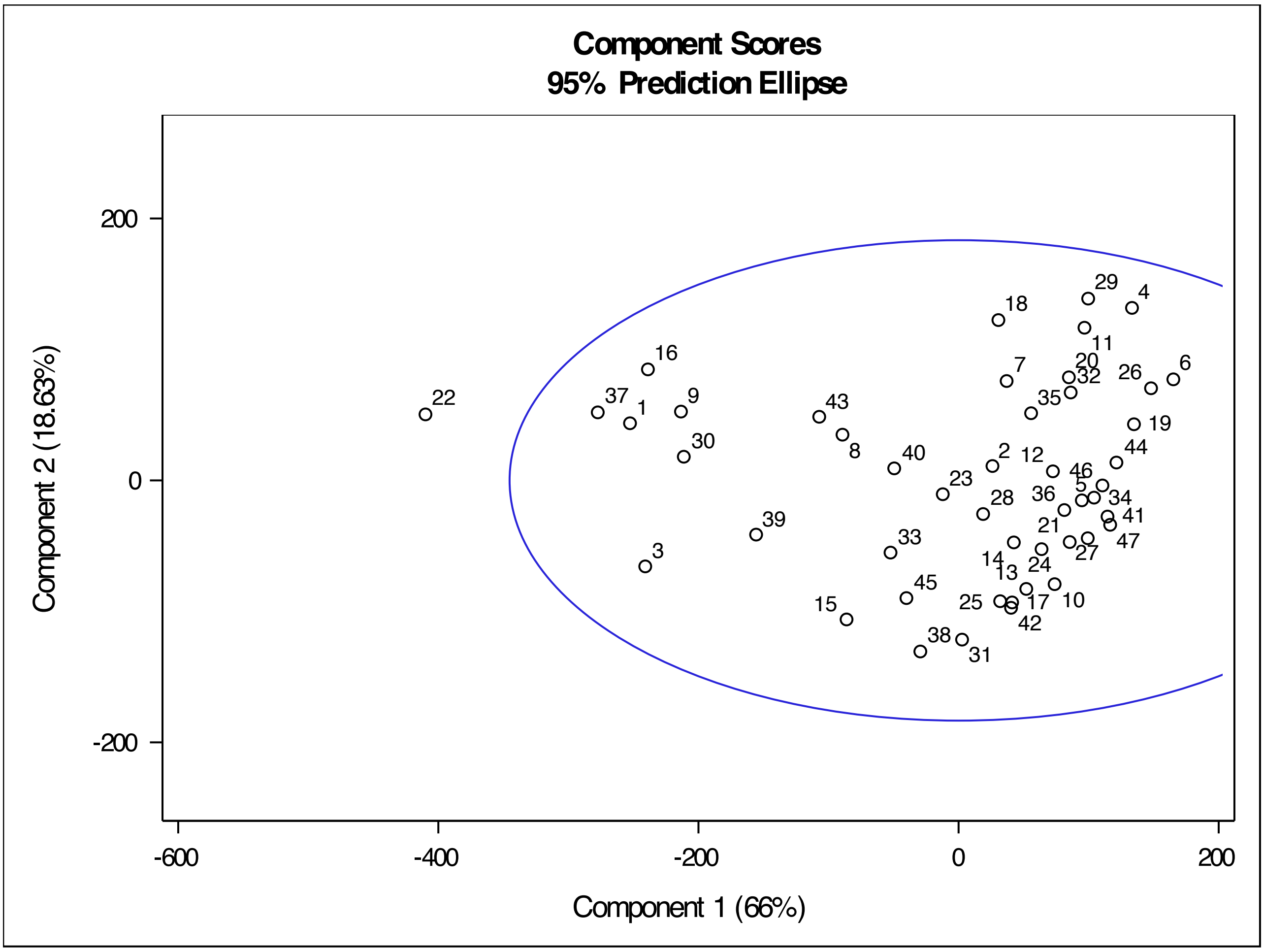
Exercise 2b):

For component 1, on the positive side, W(wealth) has the highest impact and NW(# of non-whites) has the lowest impact.

For component 2, on the positive side, W and NW have the highest impact. On the negative side, M and X have the similar coefficient, however none of them are highly important to principal component 2.

Exercise 2c):

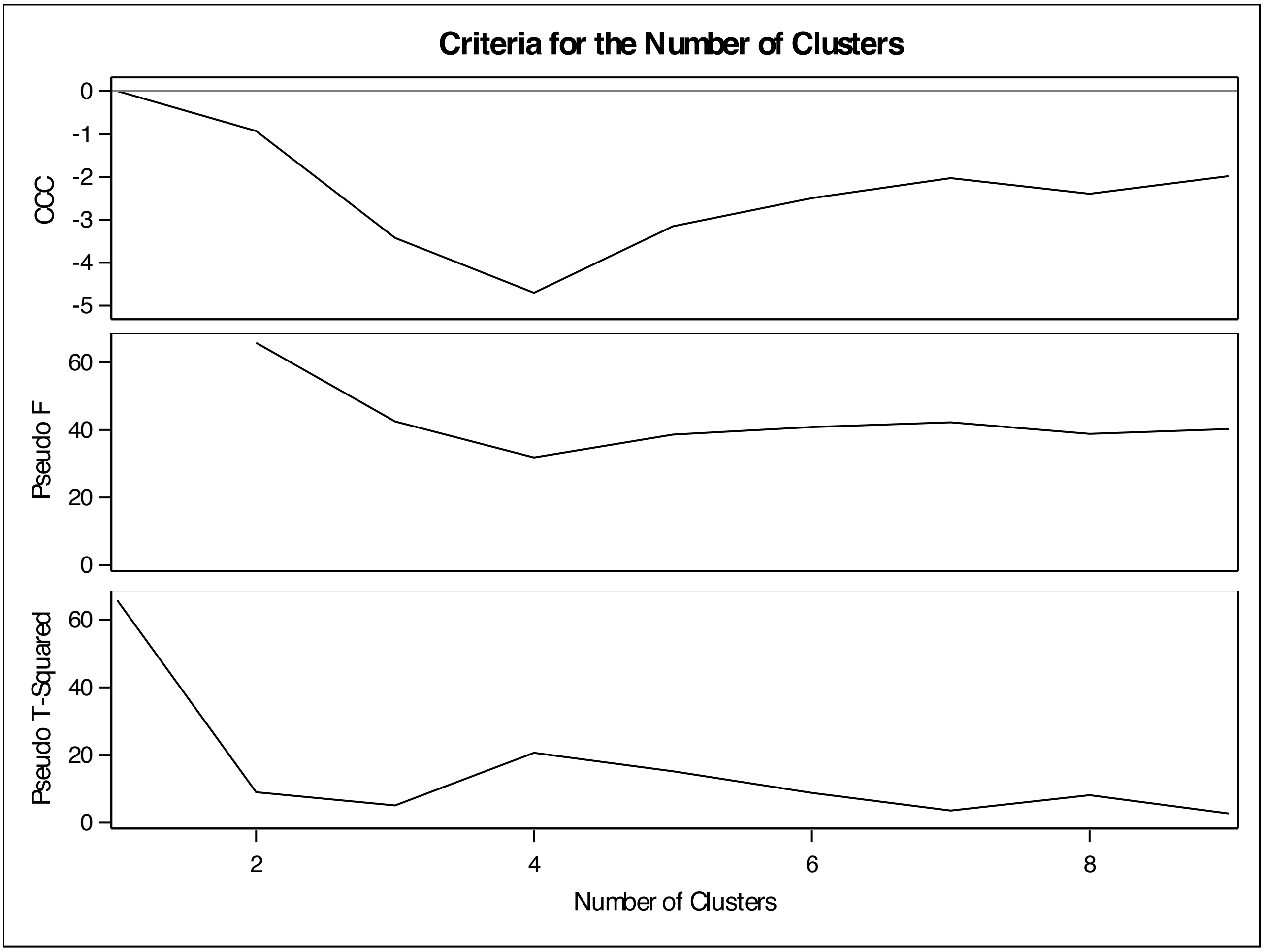
The PRINCOMP Procedure



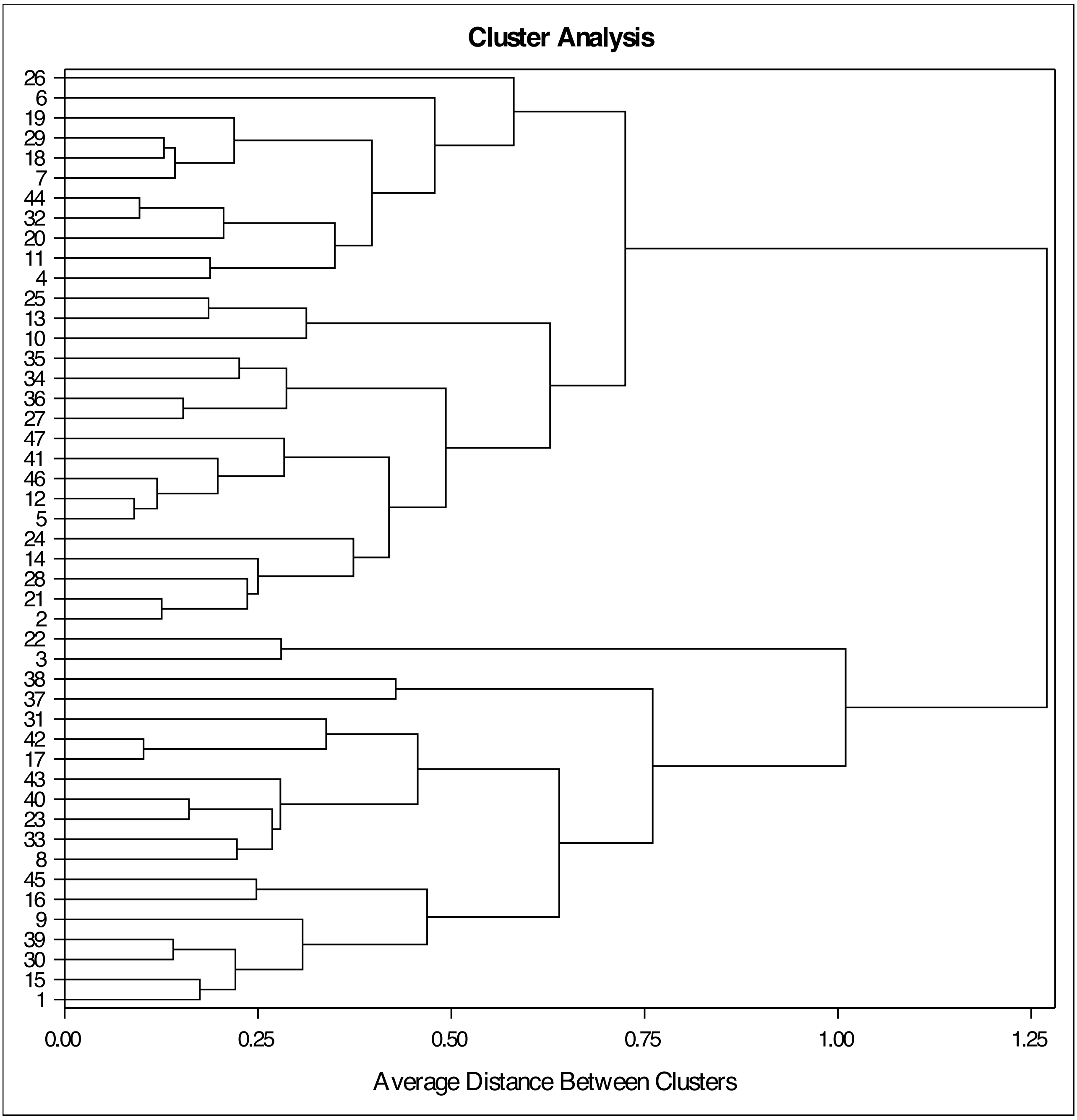
The score plot shows that state 22 has the lowest crime rate for component 1. While it is slighly above average for component 2.
In the correlation results, we had chosen 4 principal components, while in the covariance case, we had only 2 components to retain 80%. This is because some of the variables will have more impact than others.

Exercise 3a):

The CLUSTER Procedure
Average Linkage Cluster Analysis



The CLUSTER Procedure
Average Linkage Cluster Analysis



From dendrogram plot we can see that we may have 2 distinct clusters. From the plot CCC, it is   suggested we have higher value if there are 2 clusters. The same holds for Pseudo F plot, peak at 2. The pseudo t-squared has low points at 2, 3 and 7. So, overall we can say that 2 clusters might be a good choice.

Exercise 3b):

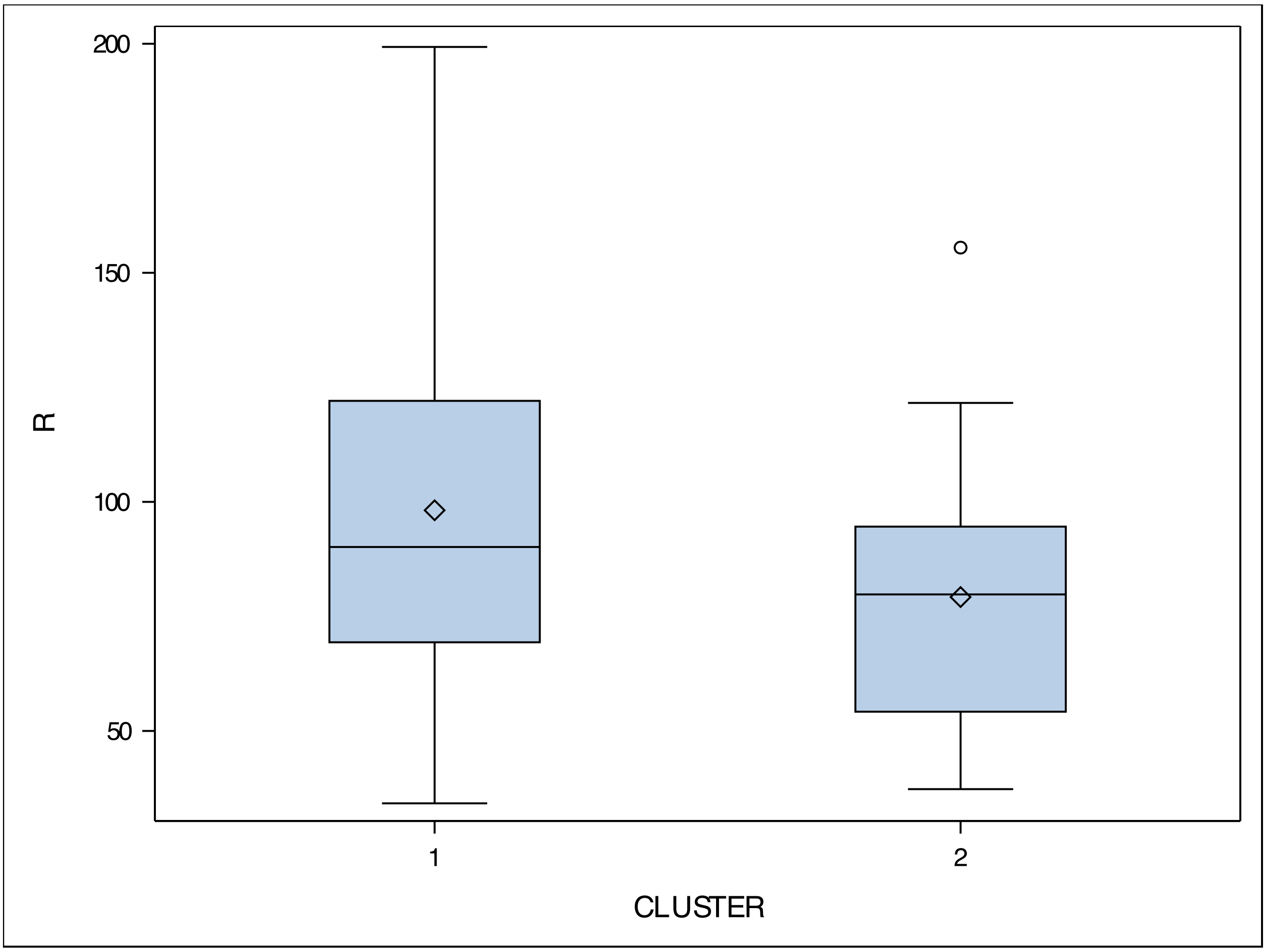
The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
LF	28	576.6071429	36.0856381	519.0000000	641.0000000
U1	28	95.5714286	17.3705340	70.0000000	142.0000000
U2	28	33.6071429	9.0690327	20.0000000	58.0000000
W	28	590.1785714	51.8795282	486.0000000	689.0000000
X	28	168.2500000	20.7981926	126.0000000	215.0000000

CLUSTER=2

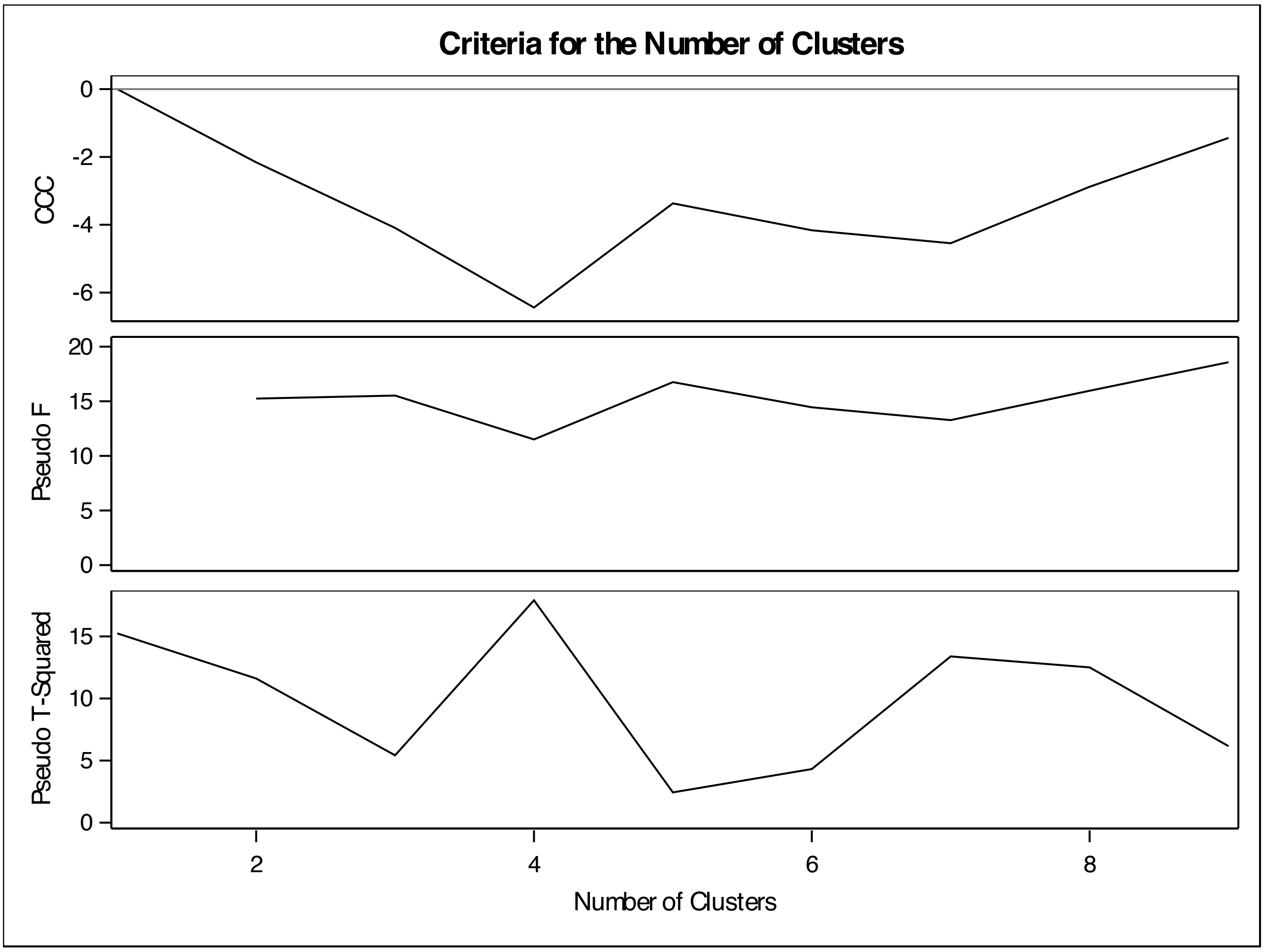
Variable	N	Mean	Std Dev	Minimum	Maximum
LF	19	538.4736842	36.1361137	480.0000000	638.0000000
U1	19	95.3157895	19.4423725	72.0000000	135.0000000
U2	19	34.5263158	7.6403056	24.0000000	53.0000000
W	19	429.8947368	60.0063837	288.0000000	513.0000000
X	19	231.9473684	29.4514170	166.0000000	276.0000000



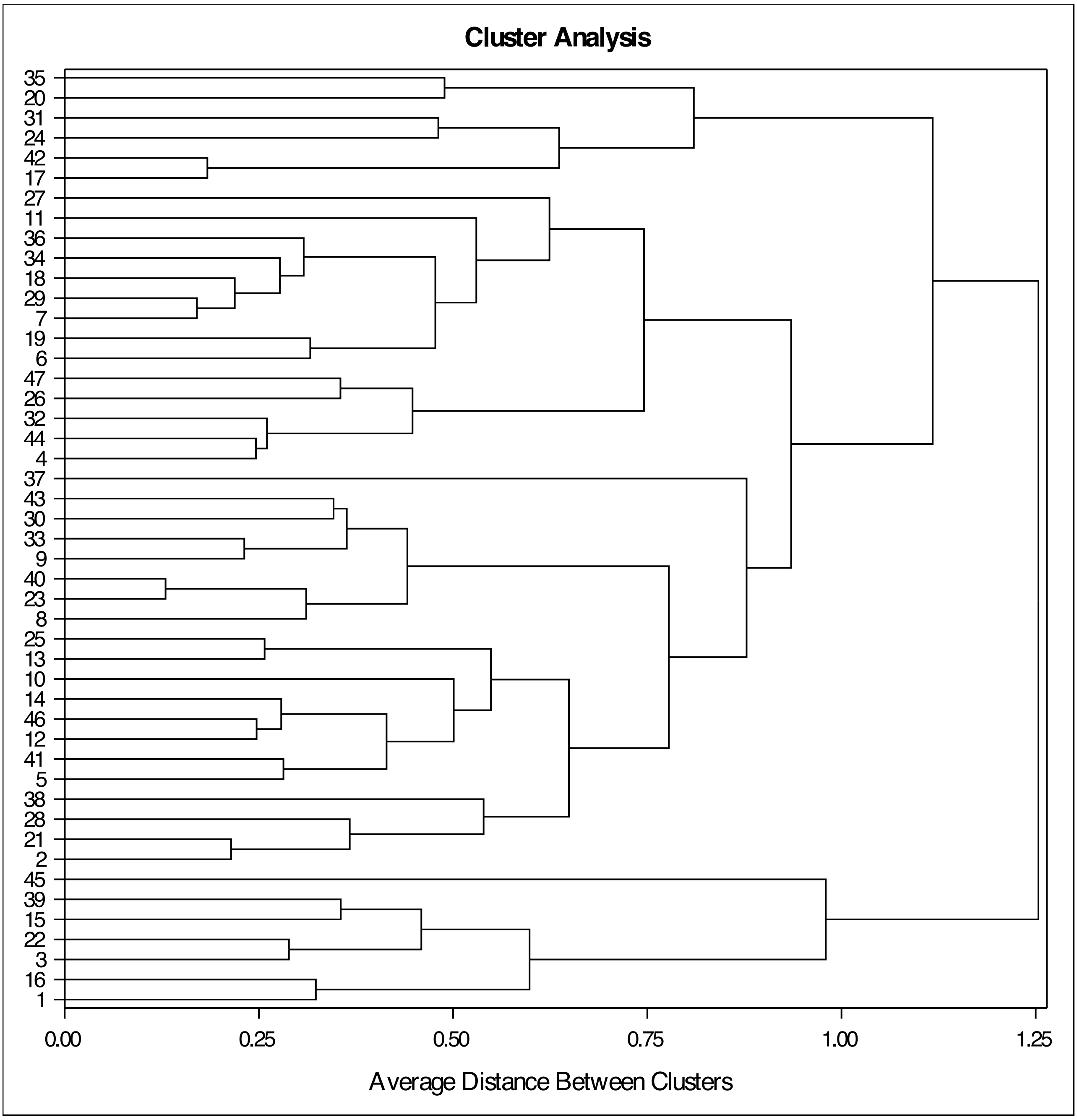
Cluster 1 includes states that tend to have higher LF(labor force) than Cluster 2. U1 is almost identical in both clusters. U2 in cluster 2 is slightly higher than in cluster 1. W(wealth) is higher for cluster 1 than for cluster 2. Cluster 2 consists of states that have higher X(income inequality) than Cluster 1. We have 28 observations in cluster 1 and 19 in cluster 2. States in Cluster 1 tend to have higher crime rates than in Cluster 2.

Exercise 4a):

The CLUSTER Procedure
Average Linkage Cluster Analysis



The CLUSTER Procedure
Average Linkage Cluster Analysis



From dendrogram plot we can see that we may have 4 or 5 distinct clusters. From the plot CCC it is suggested we have higher value if there are 3 and 5 clusters. The same holds for Pseudo F plot, peak at 3 or 5. The pseudo t-squared has low points at 3, 5 and 6. So, overall we can say that 5 clusters might be a good choice.

Exercise 4b):

The MEANS Procedure

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
LF	20	584.6500000	35.2991352	521.0000000	641.0000000
U1	20	84.4000000	10.5899655	70.0000000	103.0000000
U2	20	27.4000000	4.5814270	20.0000000	36.0000000
W	20	504.7500000	63.8970183	382.0000000	593.0000000
X	20	205.2000000	28.7046668	144.0000000	254.0000000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
LF	14	560.2857143	35.7457444	519.0000000	631.0000000
U1	14	94.2142857	12.0586935	77.0000000	113.0000000
U2	14	35.7857143	5.2795479	22.0000000	43.0000000
W	14	624.7857143	40.0464703	559.0000000	689.0000000
X	14	155.7142857	13.5897505	126.0000000	170.0000000

CLUSTER=3

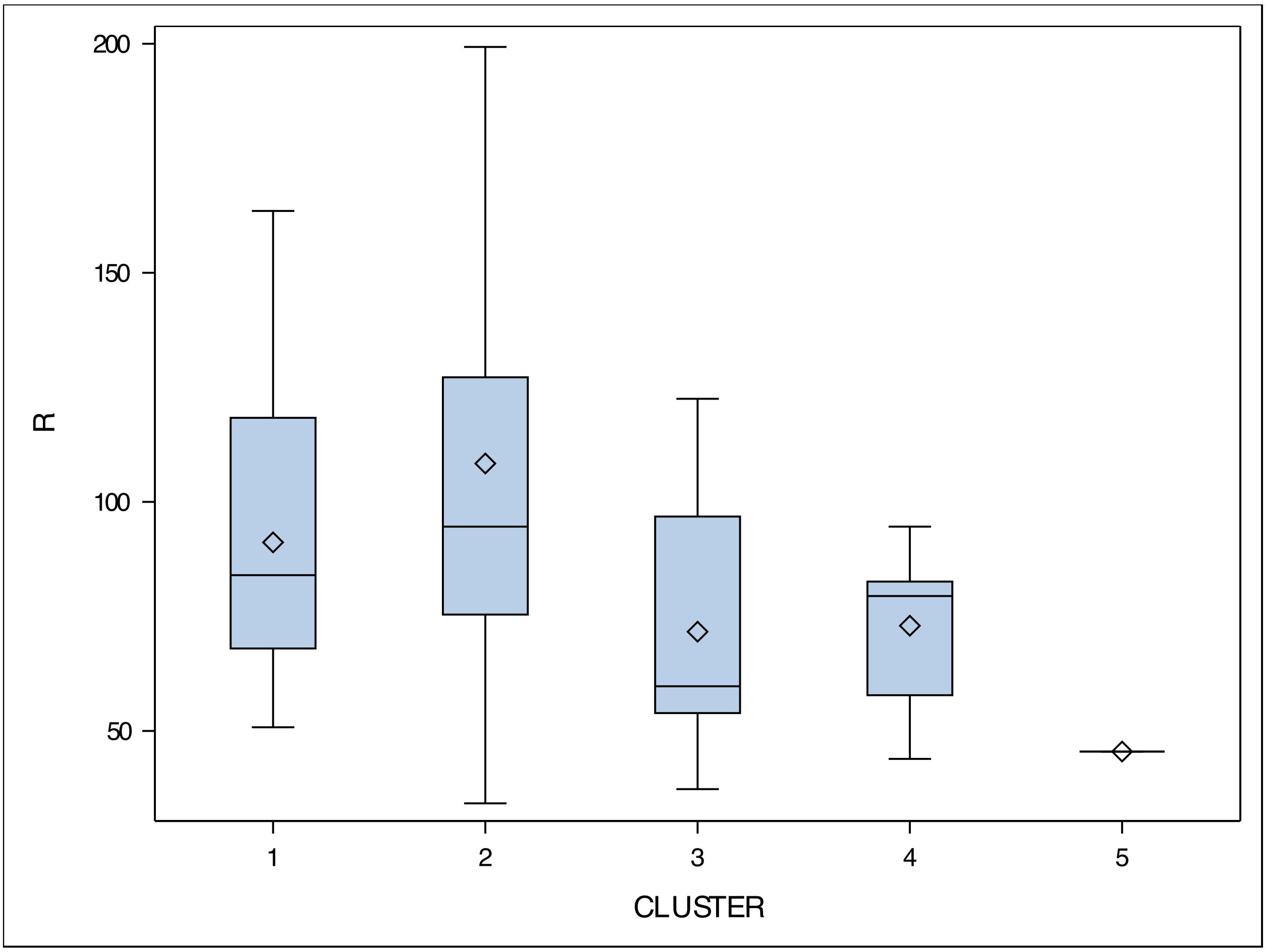
Variable	N	Mean	Std Dev	Minimum	Maximum
LF	6	543.6666667	21.5561283	523.0000000	574.0000000
U1	6	125.3333333	13.1097928	107.0000000	142.0000000
U2	6	43.6666667	8.7330789	35.0000000	58.0000000
W	6	527.8333333	64.0481590	453.0000000	626.0000000
X	6	172.6666667	14.6241809	158.0000000	200.0000000

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
LF	6	516.1666667	13.4077092	497.0000000	533.0000000
U1	6	98.8333333	11.1070548	86.0000000	116.0000000
U2	6	38.8333333	5.6715665	33.0000000	47.0000000
W	6	371.1666667	54.9451241	288.0000000	427.0000000
X	6	258.1666667	10.9802854	247.0000000	276.0000000

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
LF	1	480.0000000	.	480.0000000	480.0000000
U1	1	135.0000000	.	135.0000000	135.0000000
U2	1	53.0000000	.	53.0000000	53.0000000
W	1	457.0000000	.	457.0000000	457.0000000
X	1	249.0000000	.	249.0000000	249.0000000



Cluster 1 includes states that tend to have higher values of LF. Cluster 5 has higher U1 values. However note that it has only one observation, so we might merge this cluster with another one. Cluster 5 also has higher U2 value as well. Cluster 2 includes states which have higher values of W. Cluster 4 consists of states that have high value of X. Cluster 1 has 20 observations, Cluster 2 has 14, Cluster 3 and 4 have both 6, and Cluster 5 only one observation. States in Cluster 2 have higher crime rates in general comparing to the states in other clusters.

Regarding the comparison, we can see that using standardized variables has led us to have more clusters, with one cluster having only one observation.