STAT 448 HW #5

Tianqi Wu

2018/ 10/31

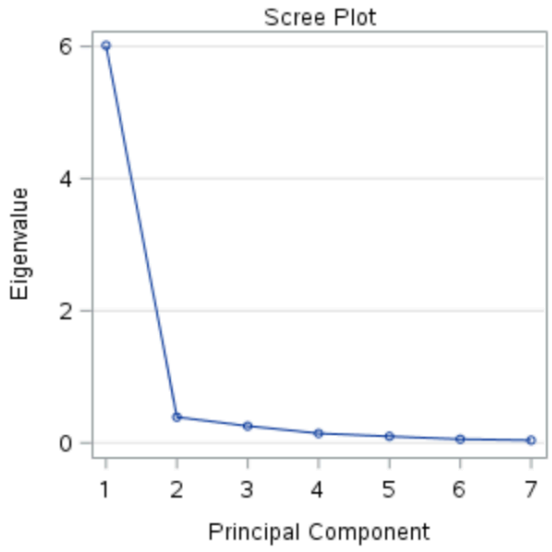
Problem 1

Here is the code that produces correlation-based PCA:

```
proc princomp data=crime2012;
id state;
run;
```

From the result, 1 component(85.93% of variation) needs to be kept to retain at least 70% of the total variation from the original variables. For average eigenvalue, 1 component would be chosen since only one component has eigenvalue greater than 1. Looking at the scree plot, we should also keep 1 component since the elbow occurs at component 2. The results for three methods agree that we should only keep one component.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.01505522	5.62352768	0.8593	0.8593
2	0.39152754	0.13485690	0.0559	0.9152
3	0.25667064	0.11364820	0.0367	0.9519
4	0.14302244	0.04568207	0.0204	0.9723
5	0.09734037	0.04012355	0.0139	0.9862
6	0.05721682	0.01804985	0.0082	0.9944
7	0.03916697		0.0056	1.0000



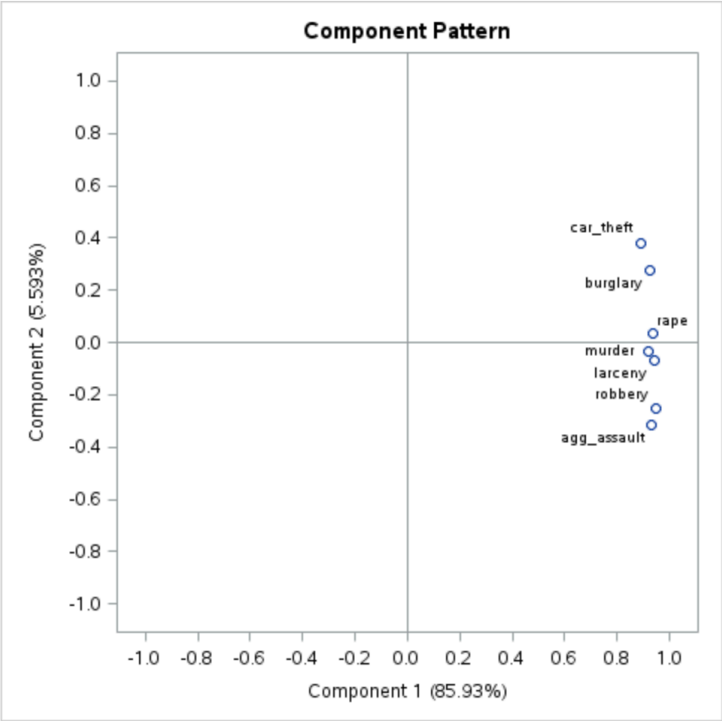
Problem 2

Here is the code that produces correlation-based PCA:

```
proc princomp data=crime2012 plots(ncomp=2)=all;
id state;
run;
```

From the component pattern plot, we can see that all the variables are positive for component 1.

From the table, the first eigenvector shows approximately equal loadings on all variables. Hence, component 1 probably describes overall averaging crime rate.



Eigenvectors

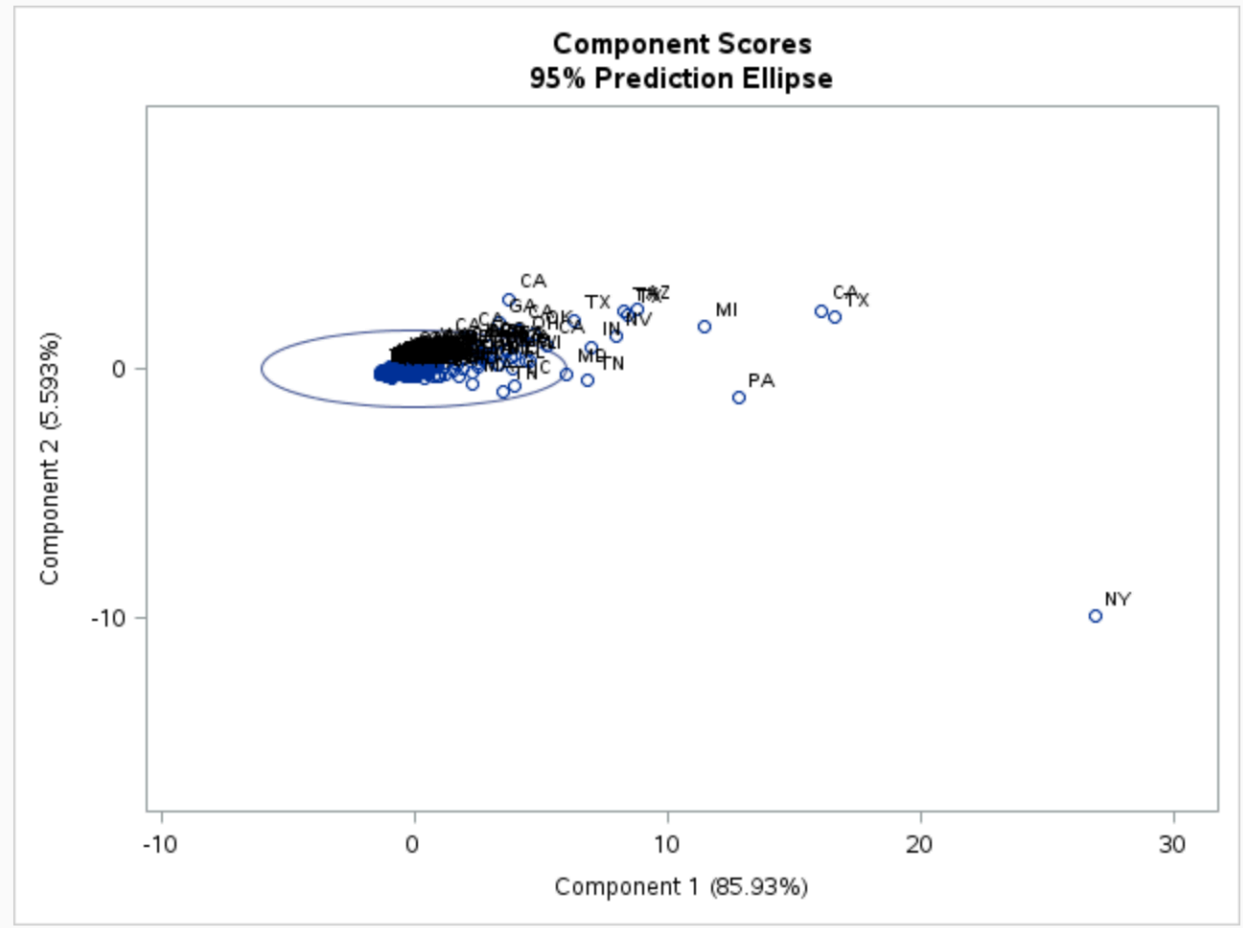
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
agg_assault	0.380661	-.500178	-.033149	0.126922	-.042987	0.512367	0.568633
burglary	0.376251	0.443684	-.298066	0.037973	-.644082	0.323483	-.227620
car_theft	0.362186	0.609607	0.296901	0.342866	0.497975	0.097284	0.184527
larceny	0.384036	-.112234	-.505661	0.328942	0.009475	-.673606	0.148960
murder	0.374988	-.057826	0.678931	-.265855	-.392133	-.394425	0.122779
rape	0.381305	0.059490	-.279655	-.794956	0.372602	0.026212	-.037244
robbery	0.385825	-.402326	0.164499	0.232499	0.206497	0.115636	-.743064

Problem 3

Here is the code that produces correlation-based PCA:

```
proc princomp data=crime2012 plots(ncomp=2)=score(ellipse);
id state;
run;
```

From the score plot, there are a lot of extreme points. NY, TX, CA, PA, MI appear to be the most extreme ones. Hence, these states have the highest crime rates and their order is as following: NY(New York)> TX(Houston) > CA(Los Ange) > PA(Philadel) > MI(Detroit), where agency is included in ().



Problem 4

Here is the code that produces covariance-based PCA:

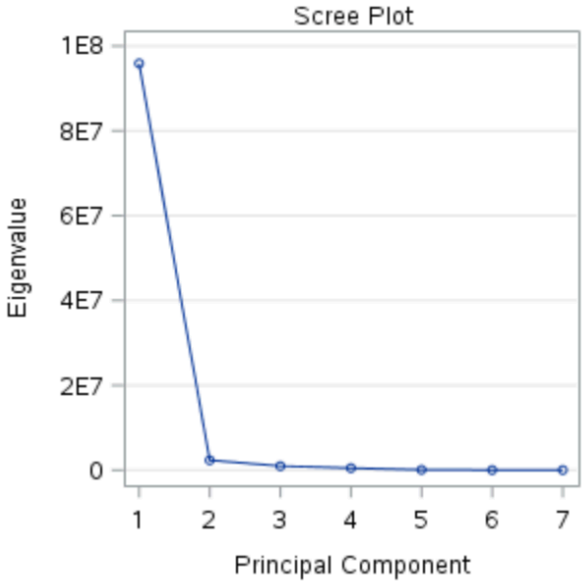
```
proc princomp data=crime2012 cov;
id state;
run;
```

From the result, 1 component(96.09% of variation) needs to be kept to retain at least 70% of the total variation from the original variables. The total variance is 99763819.959 and the average eigenvalue is 99763819.959/7 = 14251974.28. Hence, we should only choose 1 component since there is only one component with eigenvalue greater than the average eigenvalue. Looking at the scree plot, we should also keep 1 component since the elbow occurs at component 2. The results for three methods agree that we should only keep one component.

Total Variance	99763819.959

Eigenvalues of the Covariance Matrix

	Eigenvalue	Difference	Proportion	Cumulative
1	95861869.8	93544579.4	0.9609	0.9609
2	2317290.5	1338868.0	0.0232	0.9841
3	978422.4	472924.1	0.0098	0.9939
4	505498.4	407429.2	0.0051	0.9990
5	98069.1	95658.0	0.0010	1.0000
6	2411.1	2152.5	0.0000	1.0000
7	258.6		0.0000	1.0000



Problem 5

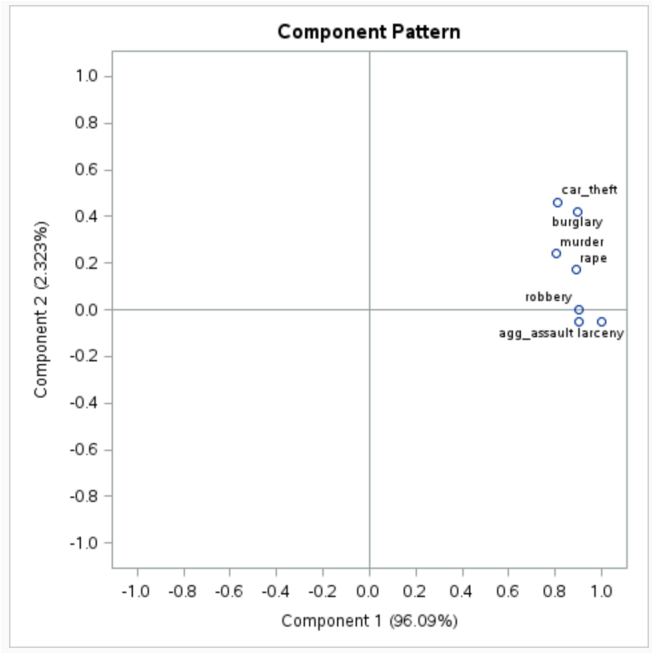
Here is the code that produces covariance-based PCA:

```
proc princomp data=crime2012 plots(ncomp=2)=all cov;
id state;
run;
```

From the component pattern plot, we can see that all the variables are positive for component 1.

From the table, the first eigenvector has much higher positive loading on larceny. Hence,

component 1 mainly describes the crime rate for larceny.



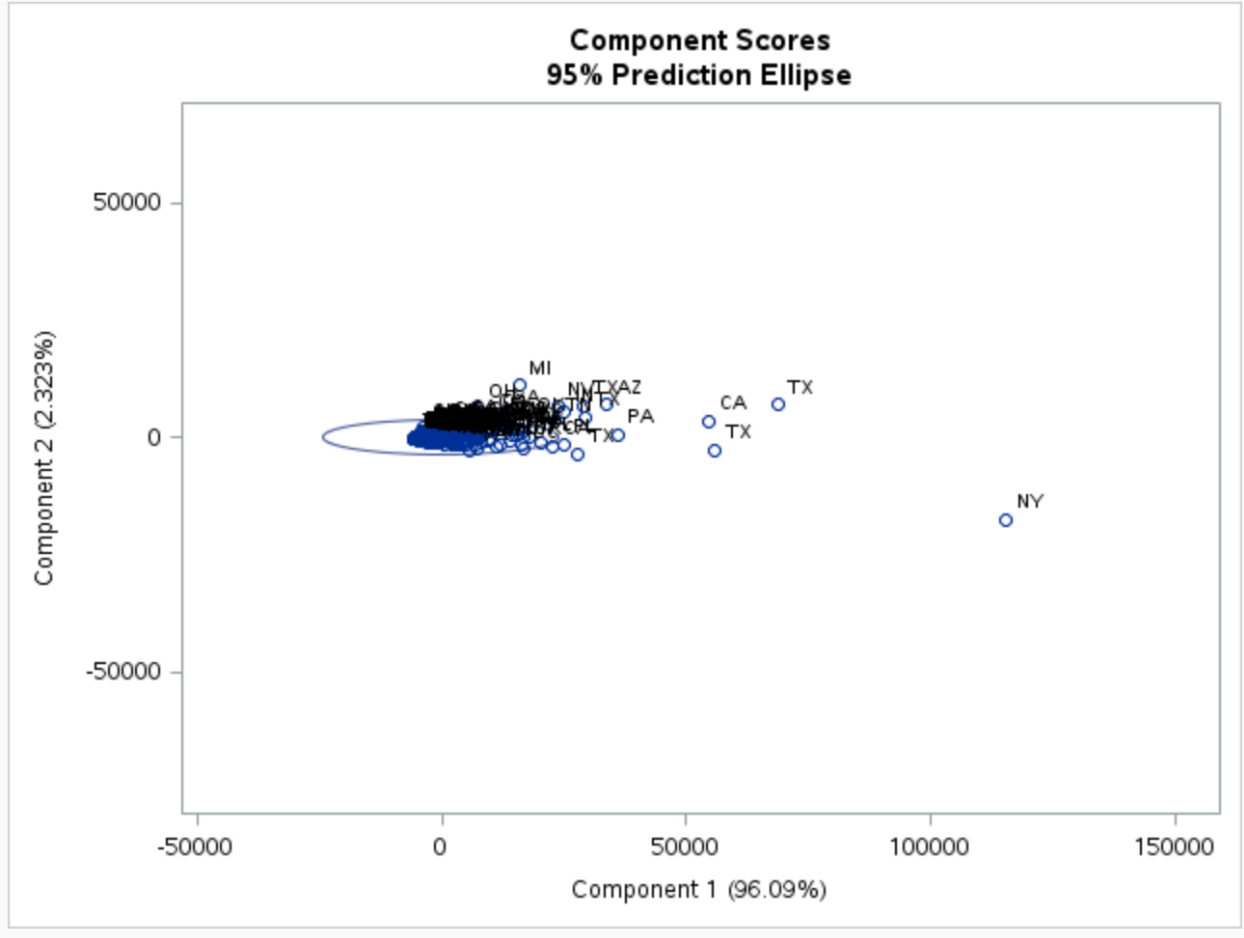
Eigenvectors							
	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
agg_assault	0.179023	-.064335	0.790398	-.288536	-.505533	-.015382	-.005540
burglary	0.267971	0.809039	-.126935	-.500944	0.079855	-.014359	-.002510
car_theft	0.137787	0.503583	0.181204	0.810403	-.194197	-.008079	-.007307
larceny	0.928268	-.295781	-.213564	0.071702	-.008411	-.002515	0.002129
murder	0.003506	0.006756	0.015987	0.003736	0.010526	0.058509	0.998067
rape	0.010975	0.013564	0.015096	-.004574	-.001347	0.997989	-.058845
robbery	0.123867	-.002242	0.529368	0.062234	0.836758	-.008953	-.017432

Problem 6

Here is the code that produces covariance-based PCA:

```
proc princomp data=crime2012 plots(ncomp=2)=all cov;
id state;
run;
```

From the score plot, there are a lot of extreme points. NY, TX, CA, PA appear to be the most extreme ones. Hence, these states have the highest crime rates of larceny and their order is as following: NY(New York) > TX(Hoston) > TX(San Anto) >CA(Los Ange) > PA(Philadel), where agency is included in ().



Problem 7

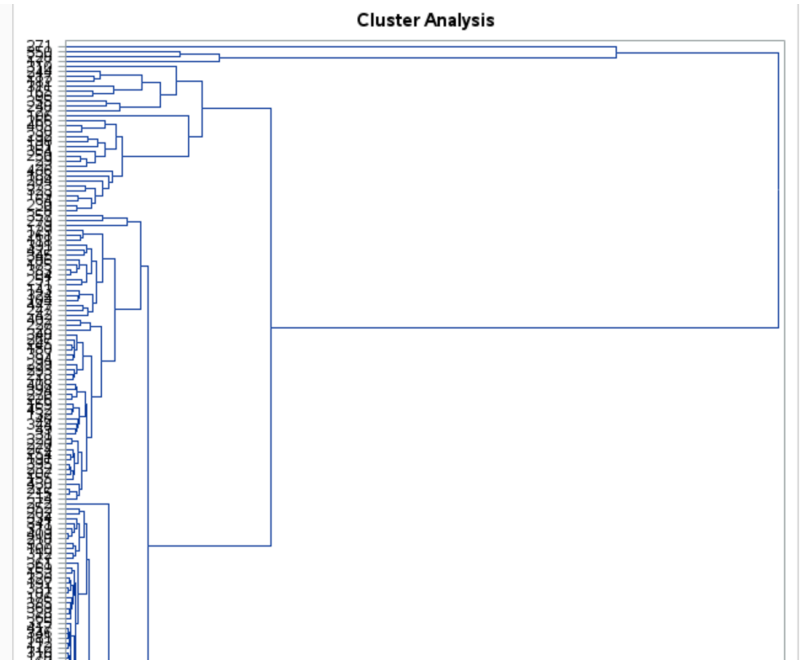
Both correlation-based and covariance-based PCA keep the same number of principal component of 1. However, the first component describes the overall averaging crime rate for correlation-based PCA and the first component captures mostly the crime rate of larceny for covariance-based one. Both of the score plots show that NY(New York), TX(Hoston), CA(Los Ange) and PA(Philadel) are the states with highest overall crime rate and crime rate of larceny.

Problem 8

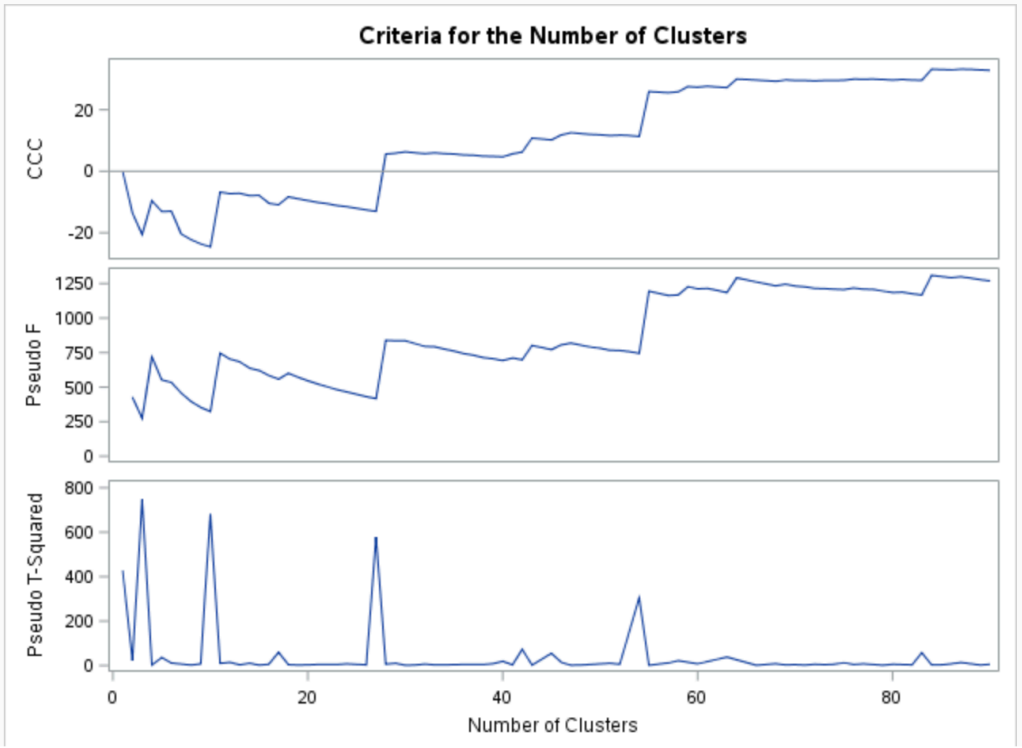
Here is the code that produces the cluster analysis:

```
proc cluster data = crime2012 method = average
ccc pseudo outtree = crime2012_tree plots = all
PLOTS(MAXPOINTS=500);
var agg_assault--robbery;
copy state;
run;
```

Part of the dendrogram is shown and 4 clusters may be chosen based on the graph. Also, 4 clusters are chosen since it gives reasonably high CCC(-9.5) and pseudo F(720) and low pseudo(2.0) t^2 statistics.



5	CL6	CL8	30	0.0249	.832	.924	-13	553	36.3	1.0778	
4	TX	CL7	3	0.0037	.828	.902	-9.5	720	2.0	1.2107	
3	CL10	CL5	447	0.2785	.550	.856	-21	274	747	1.6234	
2	CL4	NY	4	0.0616	.488	.722	-14	428	22.2	4.3481	
1	CL3	CL2	451	0.4883	.000	.000	0.00	.	428	5.6317	



Problem 9

Here is the code that produces the cluster analysis:

```
proc tree data = crime2012_tree out = crime2012_tree1 n =
4;
copy agg_assault--robbery state;
run;

proc sort data = crime2012_tree1;
by cluster;
run;

proc print data=crime2012_tree1;
run;

proc means data = crime2012_tree1;
var agg_assault--robbery;
by cluster;
run;
```

In all of 4 clusters, the count of larceny is the largest with greatest spread. The counts of murder and rape are the smallest with least spread. The order of the counts of crimes is as following: larceny > burglary > car_theft > agg_assault > robbery > rape > murder. Looking across the clusters, the order of overall crime counts is as following: cluster 4 > cluster 3 > cluster 2 > cluster 1. From the table, we can see that cluster 4 only contains state NY and cluster 3 only contains state TX and CA. Hence, those observations may be outliers with much higher crime rates.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	417	531.5947242	562.4661155	0	3732.00
burglary	417	1482.58	1259.67	182.0000000	9740.00
car_theft	417	595.9280576	855.1320802	7.0000000	8759.00
larceny	417	4098.71	3215.94	483.0000000	15534.00
murder	417	11.1702638	18.0204819	0	193.0000000
rape	417	57.0743405	59.5073261	1.0000000	403.0000000
robbery	417	279.4292566	419.7820639	3.0000000	4338.00

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	30	3997.23	2040.77	1044.00	9341.00
burglary	30	8756.90	3864.51	3519.00	17912.00
car_theft	30	4183.13	2279.98	1046.00	11500.00
larceny	30	23184.97	6437.22	15088.00	38592.00
murder	30	94.8333333	83.5716019	20.0000000	386.0000000
rape	30	307.2666667	172.8710768	42.0000000	880.0000000
robbery	30	2435.17	1556.34	681.0000000	7984.00

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	3	8037.67	3460.21	4441.00	11343.00
burglary	3	19562.00	6131.64	15668.00	26630.00
car_theft	3	11507.00	4563.85	6367.00	15084.00
larceny	3	61539.00	6037.20	56006.00	67978.00
murder	3	201.6666667	105.8363517	89.0000000	299.0000000
rape	3	716.6666667	198.6059751	549.0000000	936.0000000
robbery	3	6744.00	4230.98	1864.00	9385.00

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	1	31211.00	.	31211.00	31211.00
burglary	1	18635.00	.	18635.00	18635.00
car_theft	1	8190.00	.	8190.00	8190.00
larceny	1	115935.00	.	115935.00	115935.00
murder	1	419.0000000	.	419.0000000	419.0000000
rape	1	1162.00	.	1162.00	1162.00
robbery	1	20201.00	.	20201.00	20201.00

453	OB106	9341	13488	11500	15968	386	441	4843	MI	2	CL5
454	OB228	8329	16388	15084	56006	299	936	8983	CA	3	CL4
455	OB350	4441	15668	6367	60633	89	549	1864	TX	3	CL4
456	OB177	11343	26630	13070	67978	217	665	9385	TX	3	CL4
457	OB271	31211	18635	8190	115935	419	1162	20201	NY	4	OB271

Problem 10

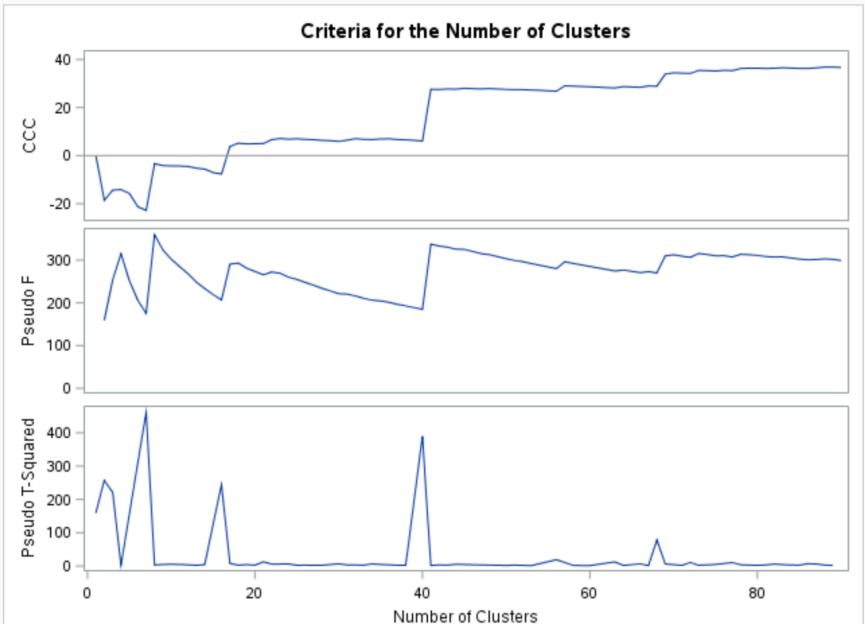
Here is the code that produces the cluster analysis:

```
proc cluster data = crime2012 method = average
ccc pseudo outtree = crime2012_tree2 plots = all std
PLOTS(MAXPOINTS=500);
var agg_assault--robbery;
copy state;
run;
proc tree data = crime2012_tree2 out = crime2012_tree3 n =
4;
copy agg_assault--robbery state;
run;
proc sort data = crime2012_tree3;
by cluster;
run;
proc print data=crime2012_tree3;
run;
proc means data = crime2012_tree3;
var agg_assault--robbery;
by cluster;
run;
```

Part of the dendrogram is shown and 4 clusters may be chosen based on the graph. 4 clusters are chosen since it gives reasonably high CCC(-14) and pseudo F(315) and low pseudo(2.8) t^2 statistics. The number of clusters chosen is the same as the original variables. However, CCC and pseudo F is smaller than the original variables and pseudo t^2 is larger than the original variables. Overall, we observe higher CCC and pseudo F and lower pseudo t^2 statistics with the standardized variables.

Also, the standardized variables cluster TX, CA ,MI , PA as cluster 3 and NY as cluster 4. There are more observations clustered as cluster 1. Overall, the variables tend to have larger mean and greater spread with standardization.

5	OB106	OB310	2	0.0057	.693	.831	-16	252	.	1.6084
4	CL5	CL6	4	0.0141	.679	.808	-14	315	2.8	2.0709
3	CL7	CL8	446	0.1480	.531	.766	-14	254	221	2.3008
2	CL3	CL4	450	0.2691	.262	.646	-19	159	257	4.1047
1	CL2	OB271	451	0.2619	.000	.000	0.00	.	159	7.7087



CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	439	660.0045558	818.2414878	0	5453.00
burglary	439	1744.95	1721.12	182.0000000	9854.00
car_theft	439	731.5466970	1070.71	7.0000000	8759.00
larceny	439	4947.74	4960.57	483.0000000	33913.00
murder	439	13.9544419	23.2994742	0	218.0000000
rape	439	65.8496583	72.5273697	1.0000000	403.0000000
robbery	439	354.7813212	559.2764794	3.0000000	4338.00

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	7	5470.29	1478.89	3647.00	7572.00
burglary	7	15012.43	1731.10	12575.00	17912.00
car_theft	7	5901.86	1550.28	2969.00	7187.00
larceny	7	34103.43	12258.89	25522.00	60633.00
murder	7	107.0000000	30.0721355	76.0000000	154.0000000
rape	7	476.8571429	103.0540773	295.0000000	596.0000000
robbery	7	3266.43	746.1357913	1864.00	4093.00

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	4	9417.75	1350.93	8329.00	11343.00
burglary	4	17127.50	6591.42	12004.00	26630.00
car_theft	4	11513.75	3710.75	6401.00	15084.00
larceny	4	44636.00	22601.70	15968.00	67978.00
murder	4	308.2500000	70.6511382	217.0000000	386.0000000
rape	4	730.5000000	225.5962470	441.0000000	936.0000000
robbery	4	7798.75	2056.65	4843.00	9385.00

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
agg_assault	1	31211.00	.	31211.00	31211.00
burglary	1	18635.00	.	18635.00	18635.00
car_theft	1	8190.00	.	8190.00	8190.00
larceny	1	115935.00	.	115935.00	115935.00
murder	1	419.0000000	.	419.0000000	419.0000000
rape	1	1162.00	.	1162.00	1162.00
robbery	1	20201.00	.	20201.00	20201.00

453	OB177	11343	26630	13070	67978	217	665	9385	TX	3	CL4
454	OB228	8329	16388	15084	56006	299	936	8983	CA	3	CL4
455	OB106	9341	13488	11500	15968	386	441	4843	MI	3	CL4
456	OB310	8658	12004	6401	38592	331	880	7984	PA	3	CL4
457	OB271	31211	18635	8190	115935	419	1162	20201	NY	4	OB271