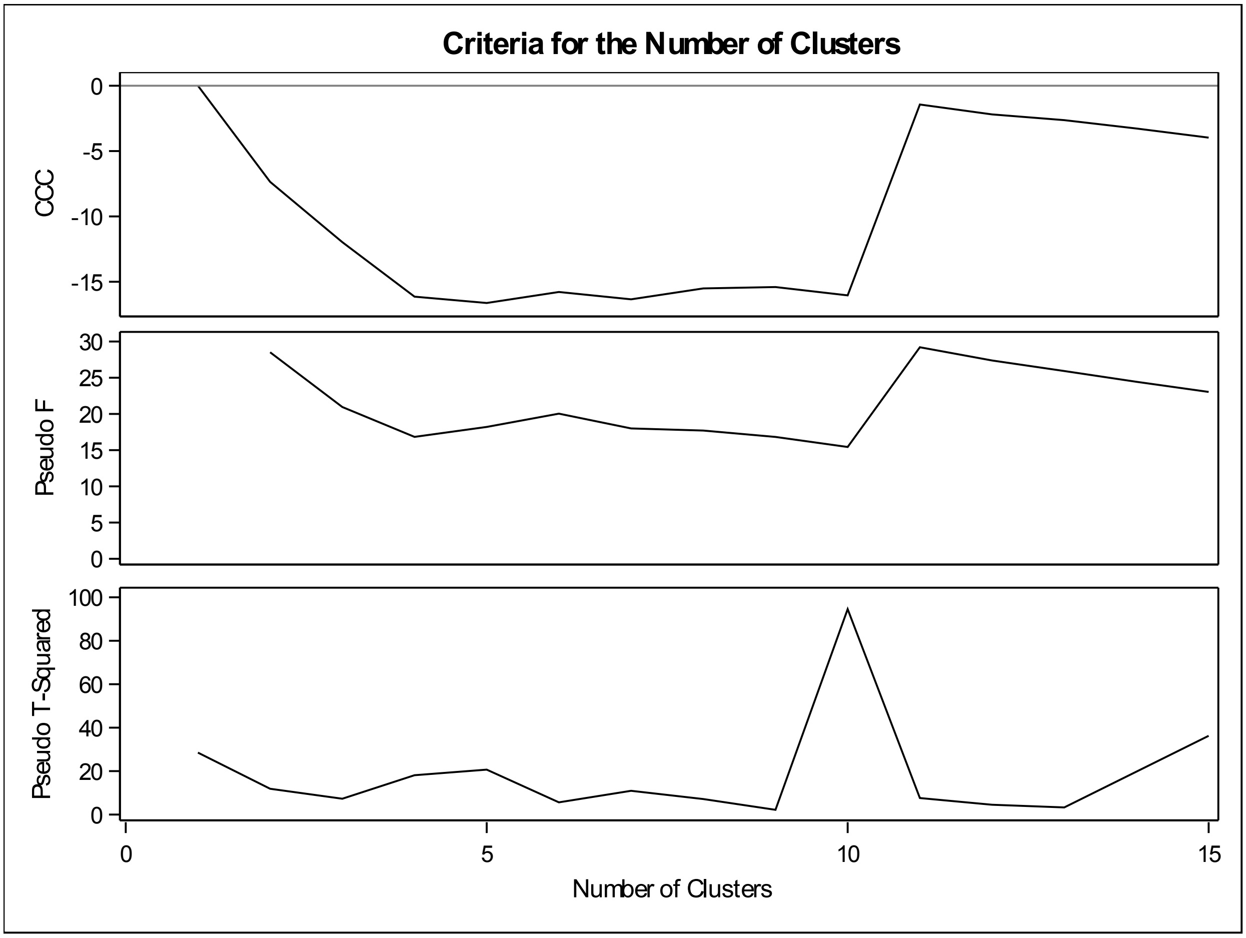
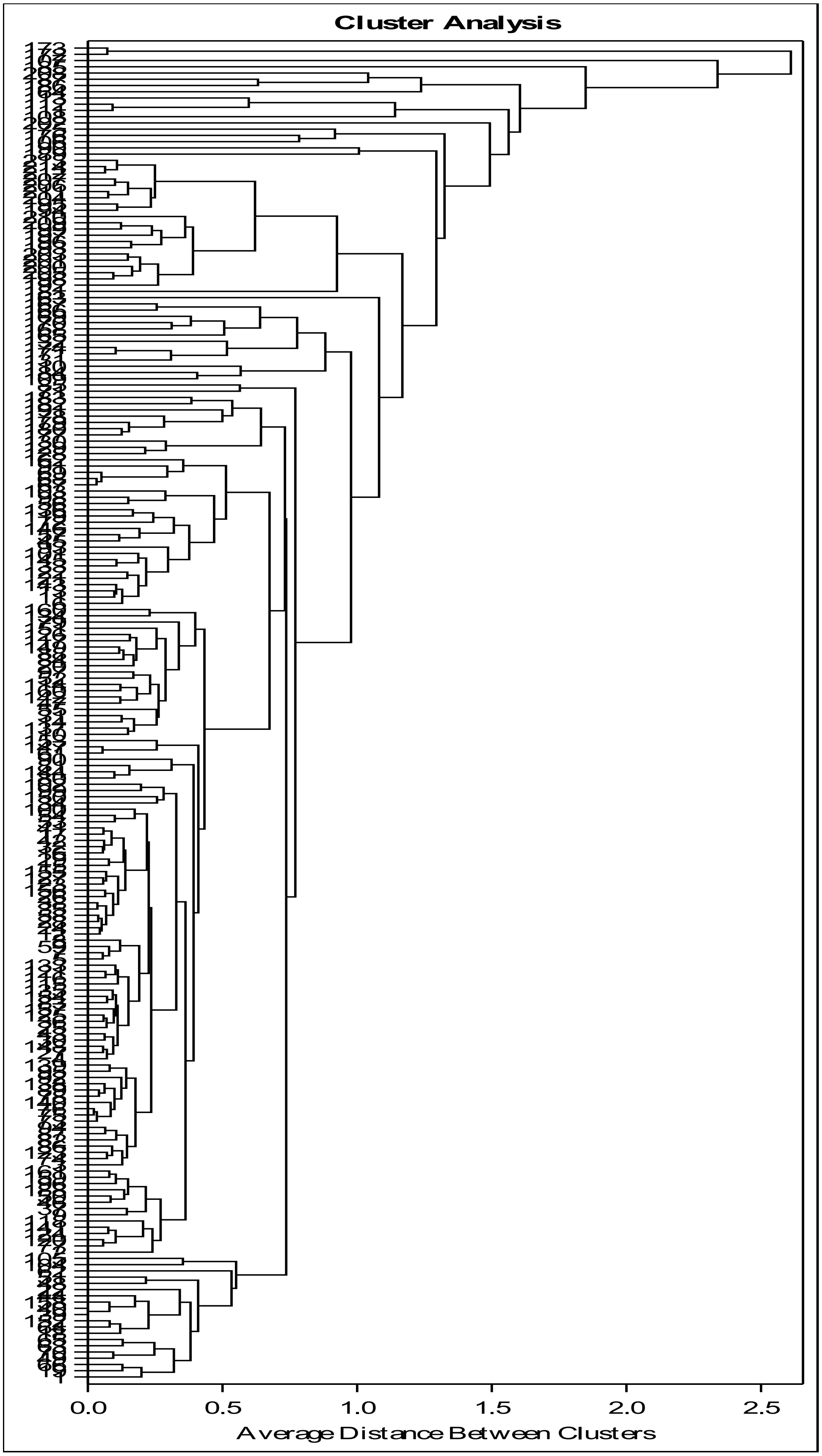
Stat 448, homework 6
Shuhui Guo

Exercise 1

a)




Cluster Analysis
0.0
0.5
1.0
1.5
2.0
2.5
Average Distance Between Clusters

Based on the pseudo F statistic and CCC statistic, there are peaks at clusters 11, thus 11 clusters should be chosen. Based on the pseudo t-squared statistic, there is a pretty big jump from 10 clusters to 11 clusters, thus 11 clusters should be chosen. The dendrogram also indicates 11 as a good choice for the number of clusters. Therefore, choosing 11 clusters is appropriate.

b)

Table of CLUSTER by groupedtype					
CLUSTER	**groupedtype**				
Frequency	**buildingwindow**	**glassware**	**headlamps**	**vehiclewindow**	**Total**
1	140	15	2	17	174
2	0	1	21	0	22
3	0	2	0	0	2
4	4	0	0	0	4
5	0	0	3	0	3
6	1	2	0	0	3
7	0	0	2	0	2
8	0	1	0	0	1
9	0	0	1	0	1
10	0	1	0	0	1
11	1	0	0	0	1
Total	146	22	29	17	214

Based on the above table, Headlamps is mostly separated out in cluster 2. Building window, glassware, and vehicle window types are mostly grouped together in cluster 1. The clustering does not match glass types very well. Based on the results, building window, glassware, and vehicle window types have similar chemical composition.

Exercise 2

a)

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	1	0.00007956	0.00007956	16.09	<.0001
Error	194	0.00095959	0.00000495		
Corrected Total	195	0.00103916			

The above table is the results of F test, which tests whether all parameters are zero in the ANOVA model. The test statistic follows an F distribution with degree of freedom 1 and 194 under the null hypothesis. The p-value is less than 0.05. Therefore under the significant level of 5%, the null hypothesis that all parameters are zero is rejected. We can conclude that not all parameters are zero and the model is significant.

R-Square	Coeff Var	Root MSE	RI Mean
0.076566	0.146499	0.002224	1.518130

The R-square is 0.076566, which indicates that about 7.6566% of variation of refractive index can be described by the model.

Levene's Test for Homogeneity of RI Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
CLUSTER	1	2.4E-10	2.4E-10	2.37	0.1253
Error	194	1.964E-8	1.01E-10		

The above table is the results of Levene’s test, which tests the homogeneity of variance. The test statistic follows an F distribution with degree of freedom 1 and 194 under the null hypothesis. The p-value is greater than 0.05. Therefore under the significant level of 5%, the null hypothesis which is homogeneity of refractive index variance cannot be rejected. We can conclude that the refractive index variable has homogeneous variance.

b)

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
1 - 2	0.0020184	0.0010258	0.0030109	***
2 - 1	-0.0020184	-0.0030109	-0.0010258	***

According to the above table, cluster 1 has significantly greater refractive index than cluster 2. Based on the results, refractive index is associated with glass type. In cluster 1, building window, glassware, and vehicle window types have similar chemical composition. The three types have higher refractive index than the cluster with headlamps.

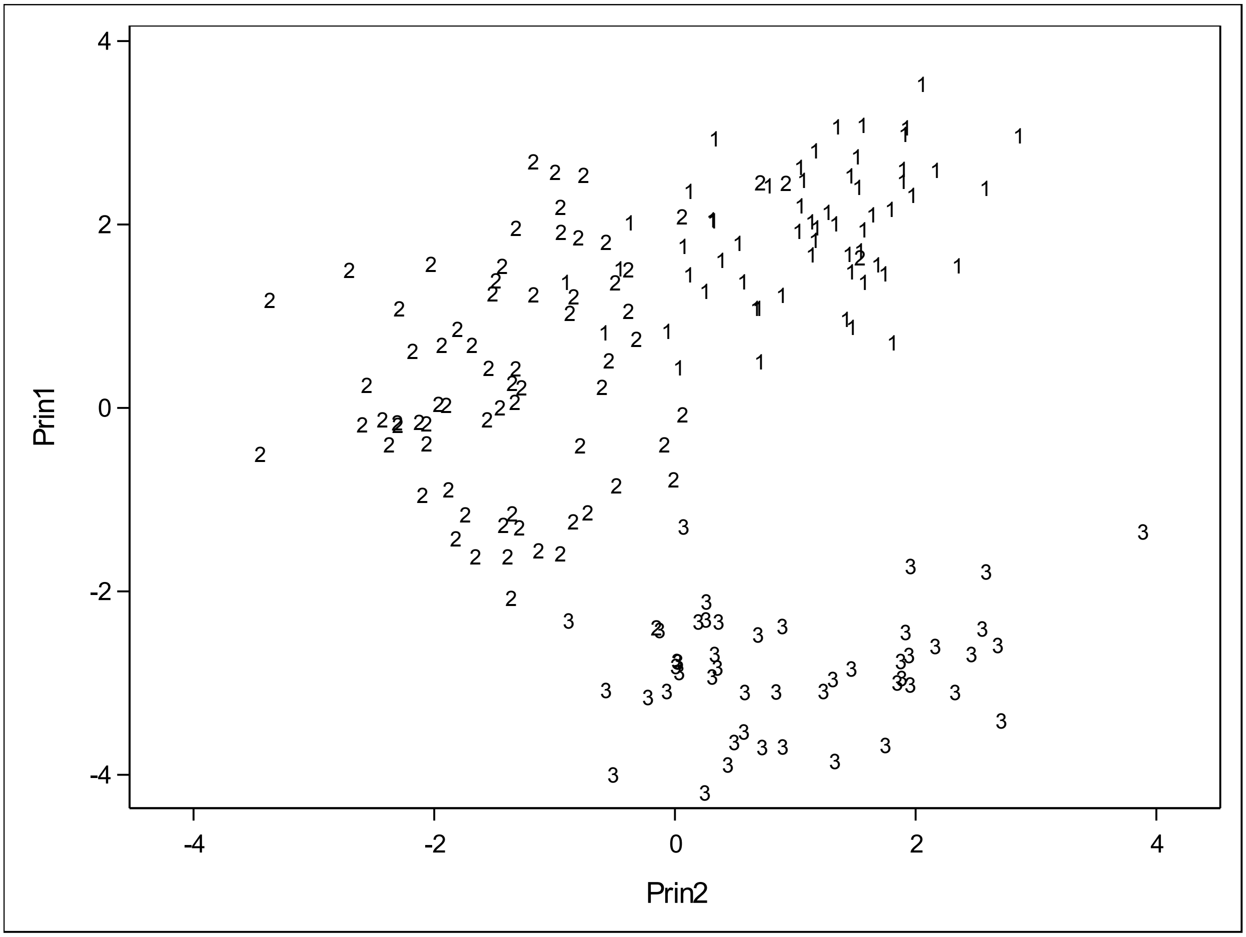
Although the mean differences are significant, the model is not much useful for predicting refractive index because the variation explained by this model is too small.

Exercise 3

a)

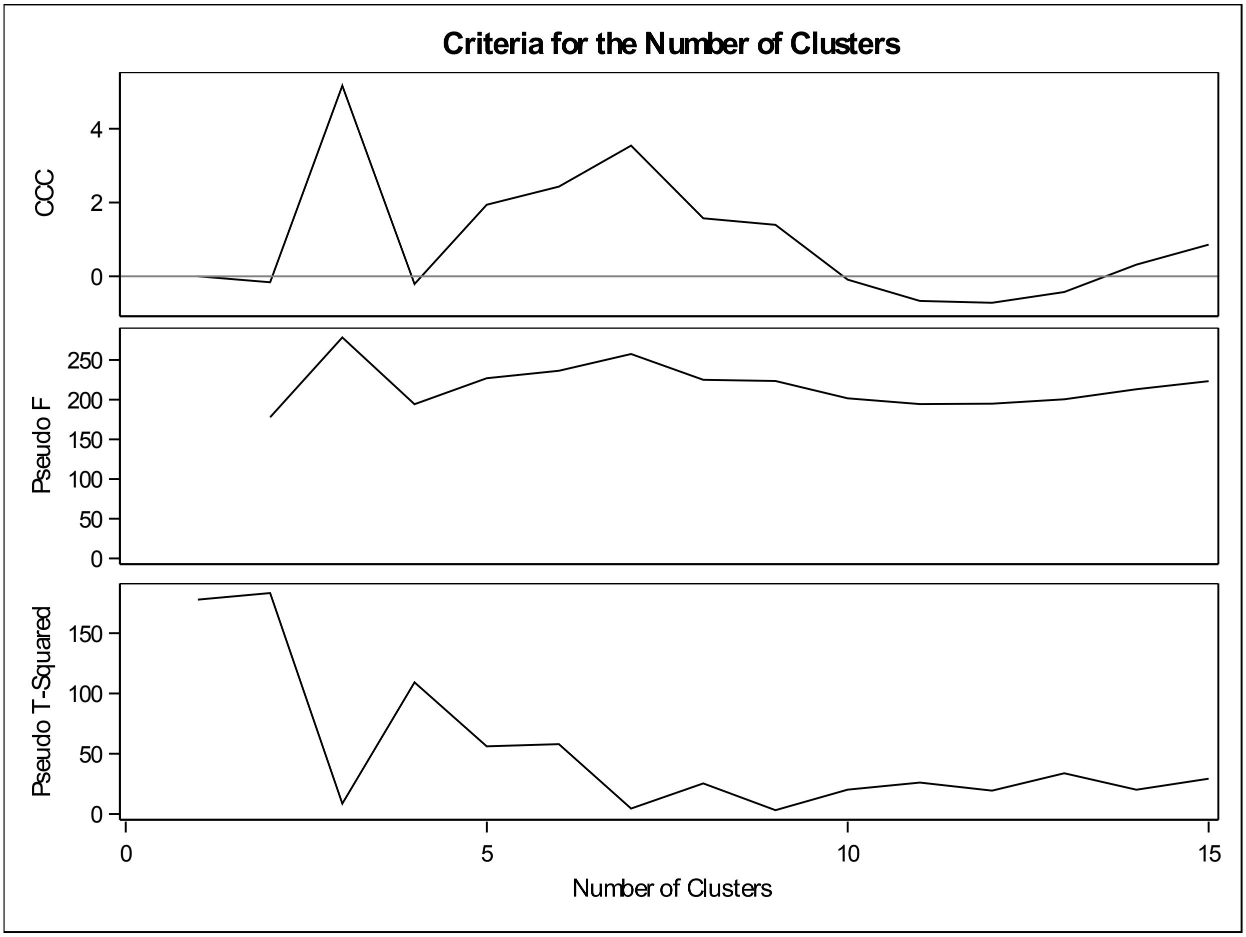
Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.39595632	2.22065955	0.3663	0.3663
2	2.17529677		0.1813	0.5476

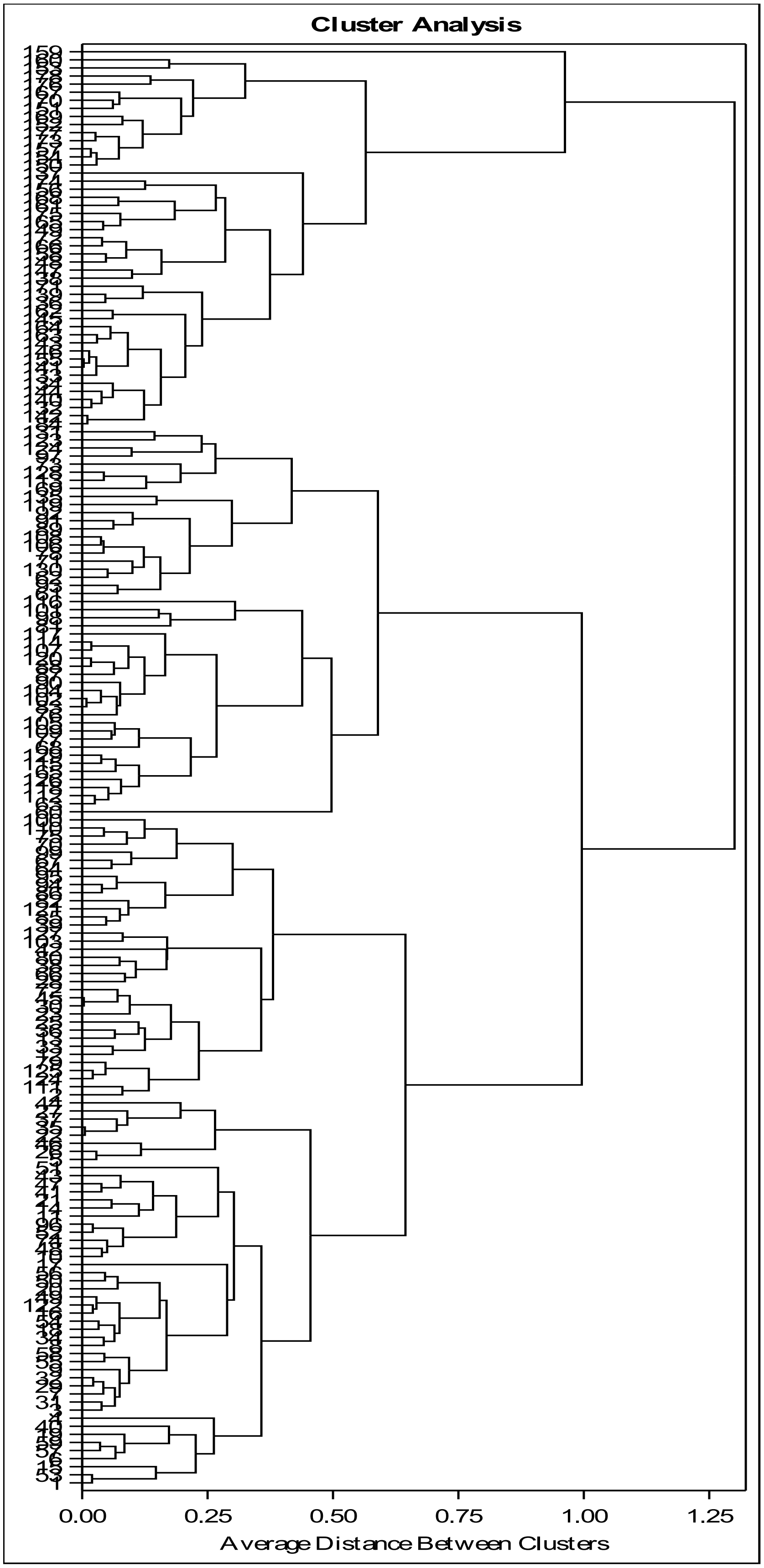
Eigenvectors		
	Prin1	**Prin2**
malic_acid	0.091117	0.513423
ash	-.270461	0.212854
alcalinity_ash	-.032286	0.368823
magnesium	-.234041	0.000809
total_phenols	0.115882	0.360852
flavanoids	0.401759	0.174439
nonflavanoid_phenols	0.437057	0.104214
proanthocyanins	-.311970	-.041107
color	0.324574	0.146536
hue	-.147665	0.537132
od280_od315	0.328153	-.256709
proline	0.405845	-.077350



According to the above scatter plot, the different alcohols are separated well. Alcohol 1 has positive values both in PC 1 and PC 2. Alcohol 2 has values around 0 in PC 1 and negative values in PC 2. And alcohol 3 has negative values in PC 1 and positive values in PC 2.

b)




Cluster Analysis
0.00
0.25
0.50
0.75
1.00
1.25
Average Distance Between Clusters

Based on the pseudo F statistic, 3, 7, 9 clusters should be chosen. Based on the pseudo t-squared statistic, 3, 5, 7, 9, 12, 14 clusters should be chosen. Based on CCC statistic, 3, 7 clusters should be chosen. The dendrogram also indicates 3 as a good choice for the number of clusters. Therefore, choosing 3 clusters is appropriate.

c)

Table of CLUSTER by alcohol				
CLUSTER	**alcohol**			
Frequency	**1**	**2**	**3**	**Total**
1	0	1	46	47
2	59	24	0	83
3	0	46	2	48
Total	59	71	48	178

Based on the above table, all observations of alcohol 1 and a small number of observations of alcohol 2 are grouped together in cluster 2. A large number of observations of alcohol 2 are separated in cluster 3. And alcohol 3 is mostly separated out in cluster 1. The separation performance is OK but not very well.

Exercise 4

a)

Stepwise Selection Summary										
Step	**Number In**	**Entered**	**Removed**	**Partial R-Square**	**F Value**	**Pr > F**	**Wilks' Lambda**	**Pr < Lambda**	**Average Squared Canonical Correlation**	**Pr > ASCC**
1	1	nonflavanoid_phenols		0.7278	233.93	<.0001	0.27222451	<.0001	0.36388775	<.0001
2	2	hue		0.6235	144.08	<.0001	0.10249051	<.0001	0.62136638	<.0001
3	3	malic_acid		0.4006	57.80	<.0001	0.06143622	<.0001	0.73590105	<.0001
4	4	magnesium		0.1532	15.55	<.0001	0.05202633	<.0001	0.75251993	<.0001
5	5	alcalinity_ash		0.2131	23.15	<.0001	0.04094029	<.0001	0.78878774	<.0001
6	6	od280_od315		0.1172	11.29	<.0001	0.03614114	<.0001	0.79933202	<.0001
7	7	proline		0.1037	9.78	<.0001	0.03239310	<.0001	0.80706733	<.0001
8	8	ash		0.0552	4.91	0.0085	0.03060568	<.0001	0.81176624	<.0001
9	9	proanthocyanins		0.0374	3.24	0.0415	0.02946112	<.0001	0.81553790	<.0001

According to the above table, there are 9 predictors selected. They are nonflavanoid_phenols, hue, malic_acid, magnesium, alcalinity_ash, od280_od315, proline, ash, and proanthocyanins.

b)

Chi-Square	DF	Pr > ChiSq
597.189174	156	<.0001

The above table is the results of the test for homogeneity of within covariance matrices. The test statistic follows a chi-square distribution with degree of freedom 156 under the null hypothesis. The p-value is less than 0.05. Therefore under the significant level of 5%, the null hypothesis which is homogeneity of within covariance matrices is rejected. We can conclude that QDA is more appropriate.

Multivariate Statistics and F Approximations

S=2 M=4.5 N=81

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02832411	67.54	24	328	<.0001
Pillai's Trace	1.63745462	62.10	24	330	<.0001
Hotelling-Lawley Trace	10.79988562	73.42	24	280.09	<.0001
Roy's Greatest Root	7.77768507	106.94	12	165	<.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

According to the above table, we should reject the null hypothesis and conclude that this classification model is meaningful.

c)

Number of Observations and Percent Classified into alcohol

From alcohol	1	2	3	Total
1	57 96.61	2 3.39	0 0.00	59 100.00
2	3 4.23	68 95.77	0 0.00	71 100.00
3	0 0.00	0 0.00	48 100.00	48 100.00
Total	60 33.71	70 39.33	48 26.97	178 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for alcohol

	1	2	3	Total
Rate	0.0339	0.0423	0.0000	0.0254
Priors	0.3333	0.3333	0.3333	

The cross-validation error is 0.0254, which is low. It seems that the discrimination matches the groups well. Alcohol 1 and 2 are mostly classified to their relative groups. And alcohol 3 is classified pretty well to group 3.

There are similarities between the classification results and the cluster frequency analysis results from Exercise 3. The alcohols are both separated into three groups and alcohol 3 is mostly separated out in a group. There are also dissimilarities between the classification results and the cluster frequency analysis results from Exercise 3. The classification results better separate the alcohols than the cluster frequency analysis. In the cluster frequency analysis results from Exercise 3, alcohol 1 and some observations of alcohol 2 are grouped together, while in the classification results, alcohol 1 and 2 are mostly classified into two groups.