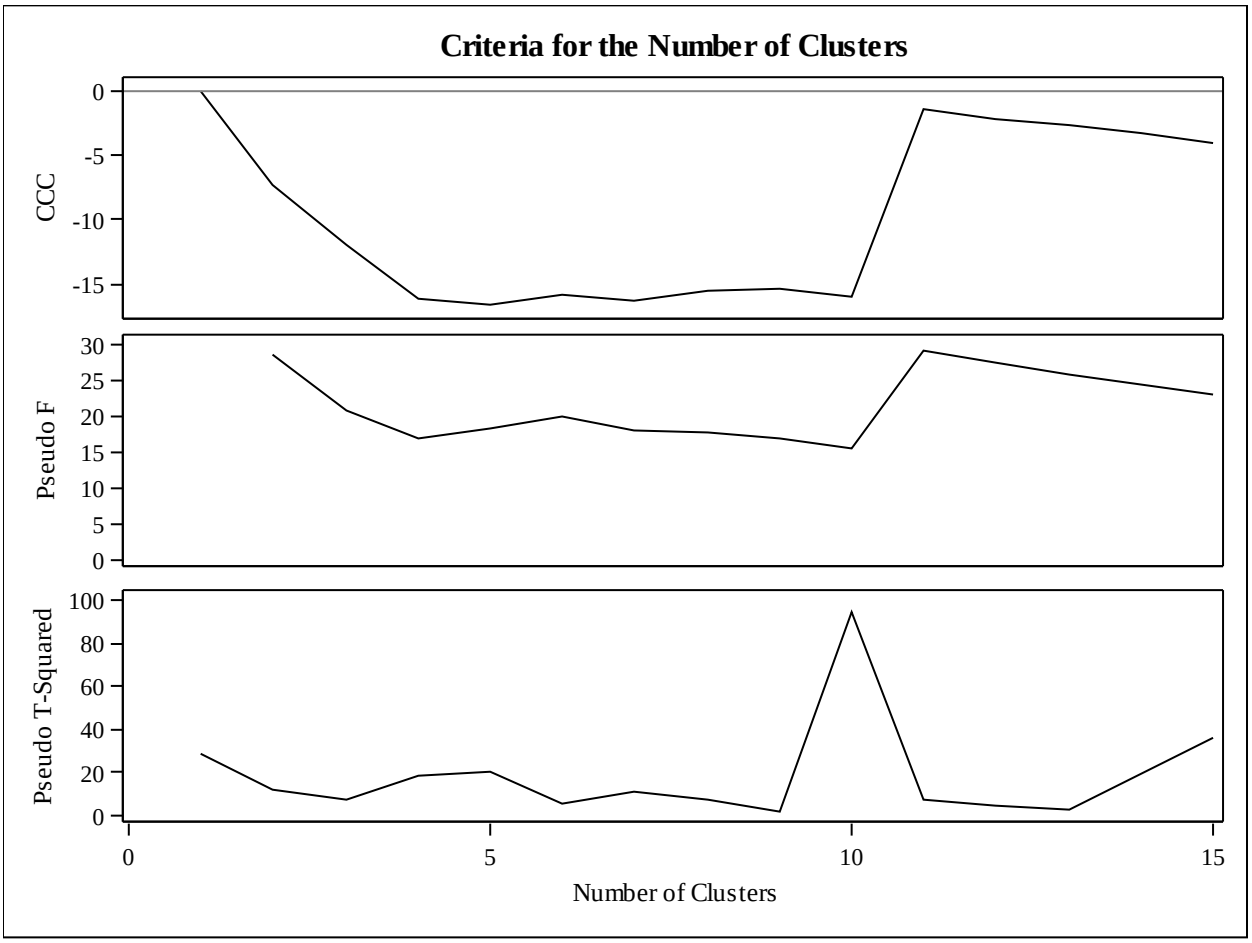
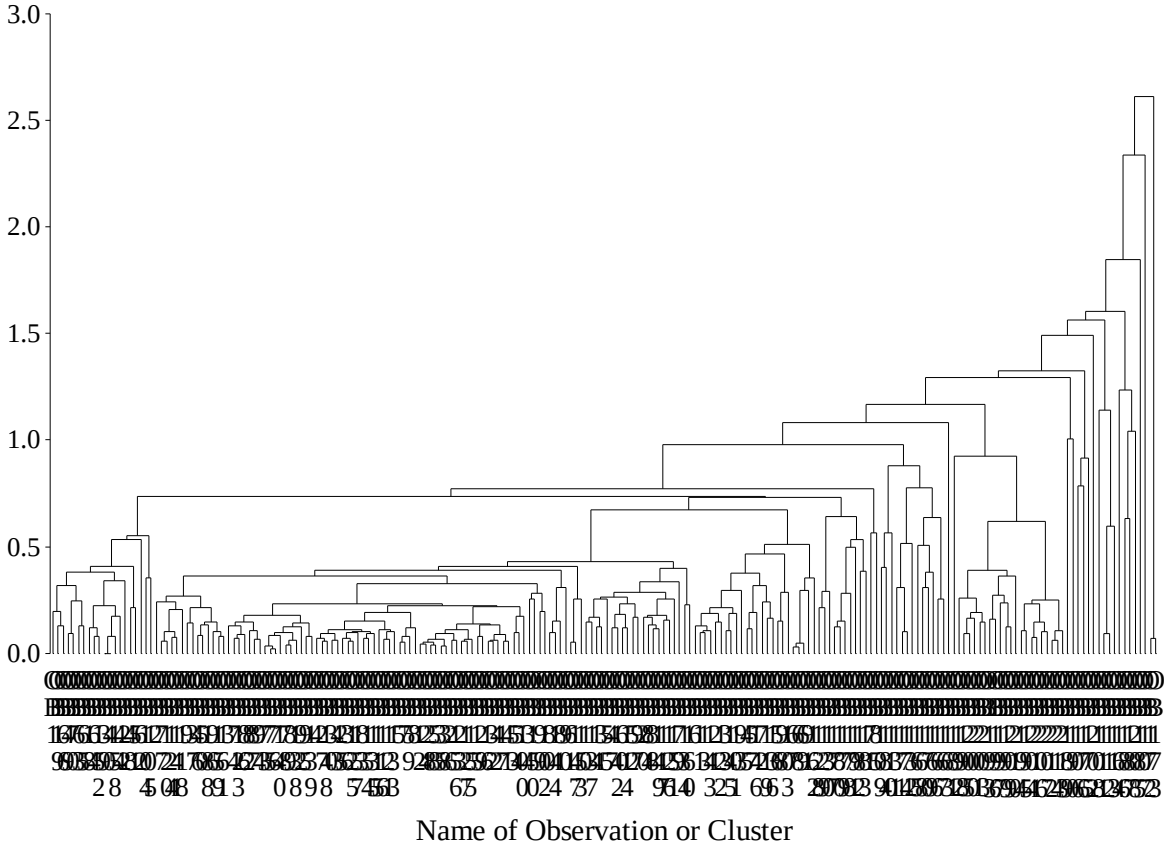
Exercise 1

1a) To choose the optimal number of clusters, we see CCC, Pseudo F, and Pseudo t^2 statistics. First, the CCC plot has a peak at 1 and 11, and the Pseudo F plot shows a peak at around 2 and 11. In terms of Pseudo t^2, we find deep points at 3, 6, 9, and 11. The three criteria commonly say 11 clusters will be the best choice. However, note that CCC shows negative values for each number of clusters. It implies that it is hard to find clear separation among clusters in our dataset. As the next reference, the dendrogram shows that somewhere around 11 clusters could be okay, but still all clusters are close to each other. Based on diagnostics, we choose 11 clusters to be the number of clusters and continue further analysis, but negative CCC values and dendrogram imply that this dataset do not show clear separation among clusters.



Exercise 1



1b) A frequency table gives the cross-cluster frequencies between 11 clusters and the original 4 glass groups. We can see that three of the glass groups, buildingwindow, glassware, and vehiclewindow, are mostly grouped to cluster 1. The glass type headlamp is mostly in cluster 2. All other clusters include a few from each glass types. Thus, we can say that 11 clusters we choose from part (a) do not match with the glass types very well. However, we can at least infer that buildingwindow, glassware, vehiclewindow have similar chemical composition (which would be a feature of glasses in cluster 1). At the same time, headlamps would have a different chemical composition compared to other three.

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

Exercise 2

2a) From the frequency table in Exercise 1 (b), we see that only cluster 1 and 2 have more than 5 observations. So we will perform ANOVA for refractive index as a function of two clusters. Indeed, it would be an analysis to compare the means of refractive index between two groups, cluster 1 and 2. First, we see that Levene's test for homogeneity of variance has p-value larger than .05. It implies that we can continue ANOVA and its output is valid. Second, we see that model is significant with p-value less than .001, thus we can conclude that two groups have significant difference in refractive index. Lastly, ANOVA model can explain 7.65% of total variation in refractive index, which seems a bit low.

Dependent Variable: RI

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			

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

Source	DF	Anova SS	Mean Square	F Value	Pr > F
CLUSTER	1	0.00007956	0.00007956	16.09	<.0001

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		

Tukey's Studentized Range (HSD) Test for RI

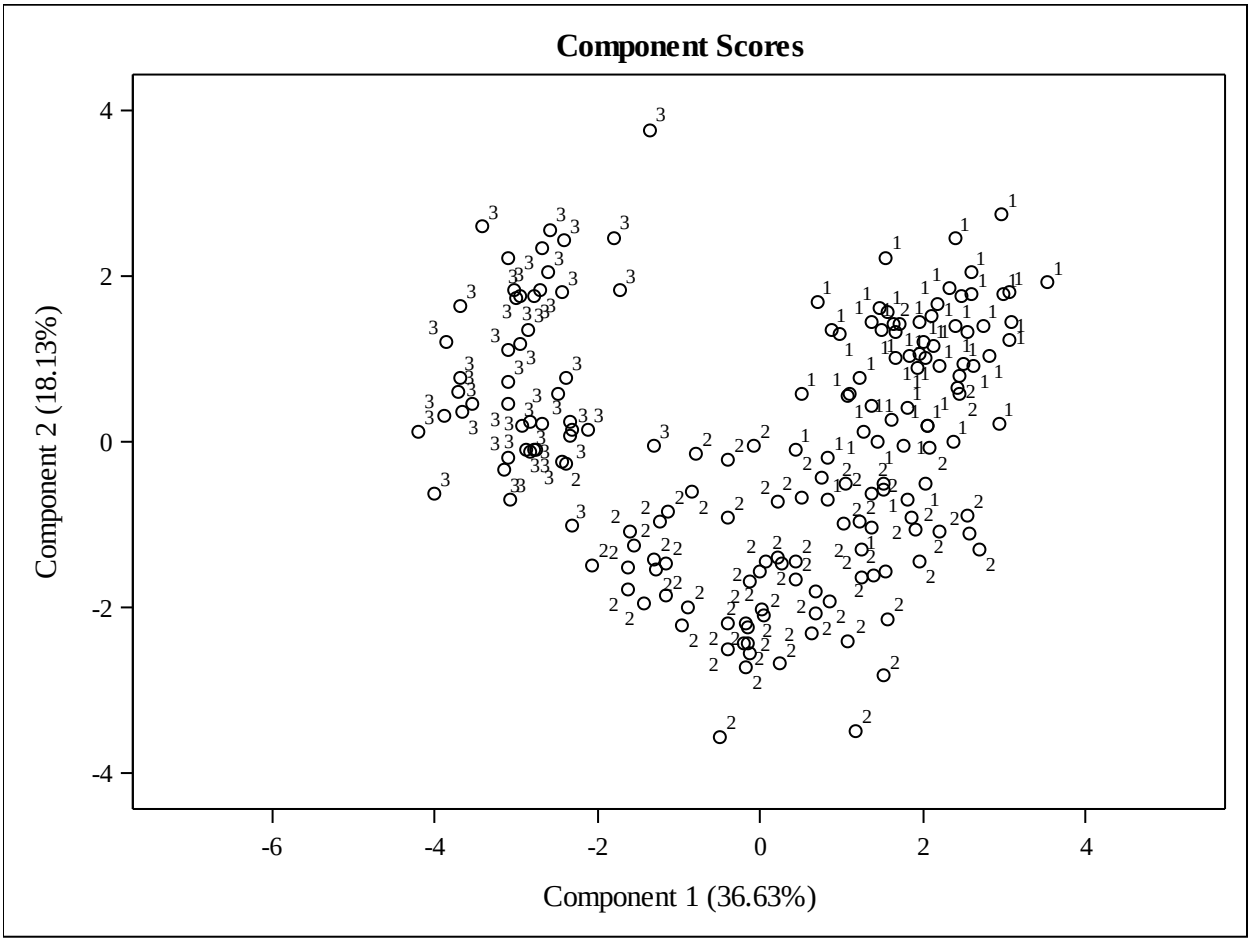
Exercise 2

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	***

2b) Tukey's pairwise comparison result shows that there is significant difference between cluster 1 and 2 in terms of refractive index, and cluster 1 has higher mean than cluster 2 has. However, as we see in part (a), its R^2 is pretty low, less than 10 %. It implies that this model is not very useful to predict refractive index.

Exercise 3

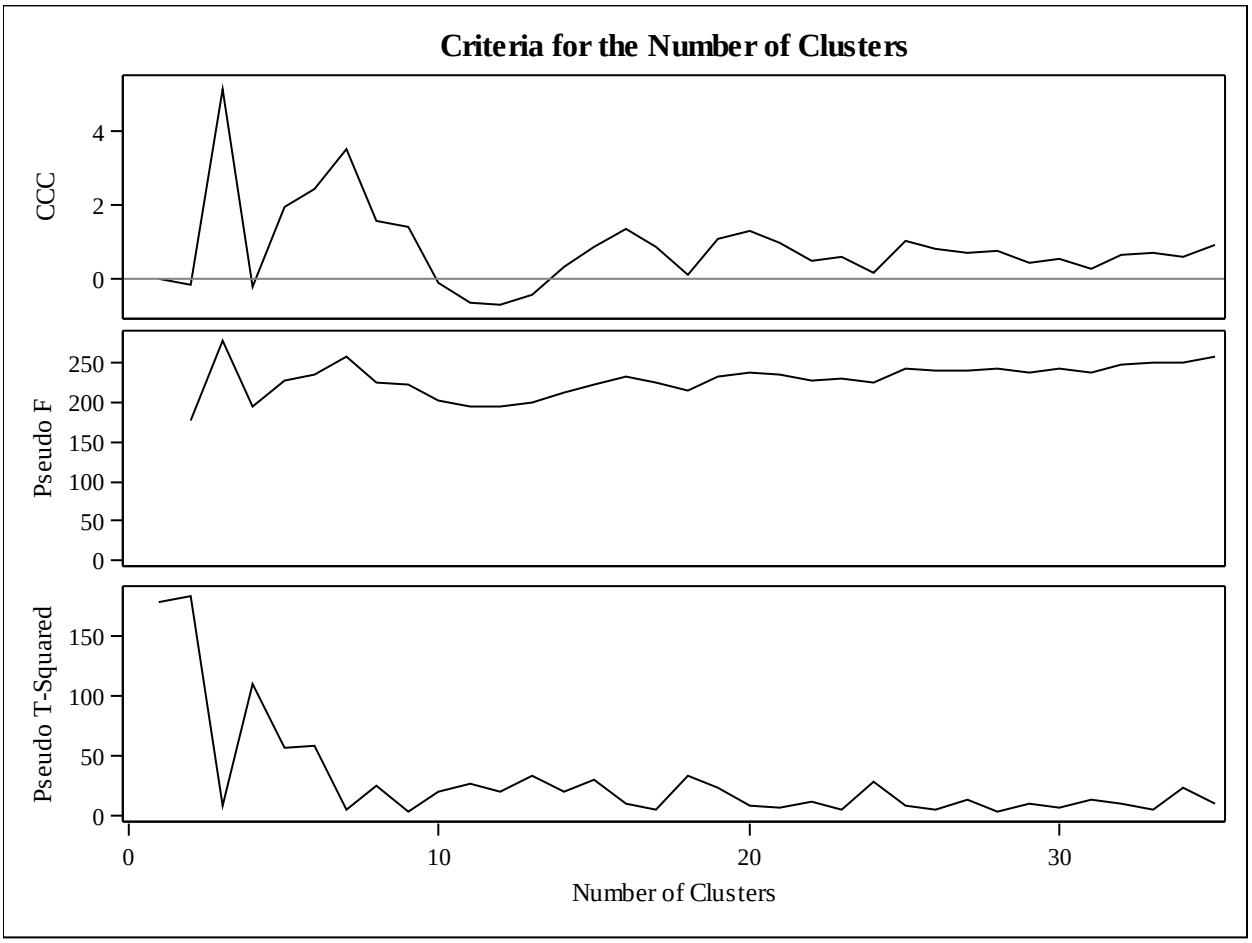
3a) We start by performing a principal component analysis on the wine characteristics and keeping the first 2 components.



The score plot shows noticeable difference in the location of the alcohol types, indicating noticeable differences in the the most prominant features of the alcohols. Alcohol type 1 tends to have higher values for both component 1 and component 2. Alcohol type 2 tends to have roughly average values for component 1 and lower than average values for component 2. Alcohol 3 tends to have lower than avergae values for component 1 and higher than average values for component 2.

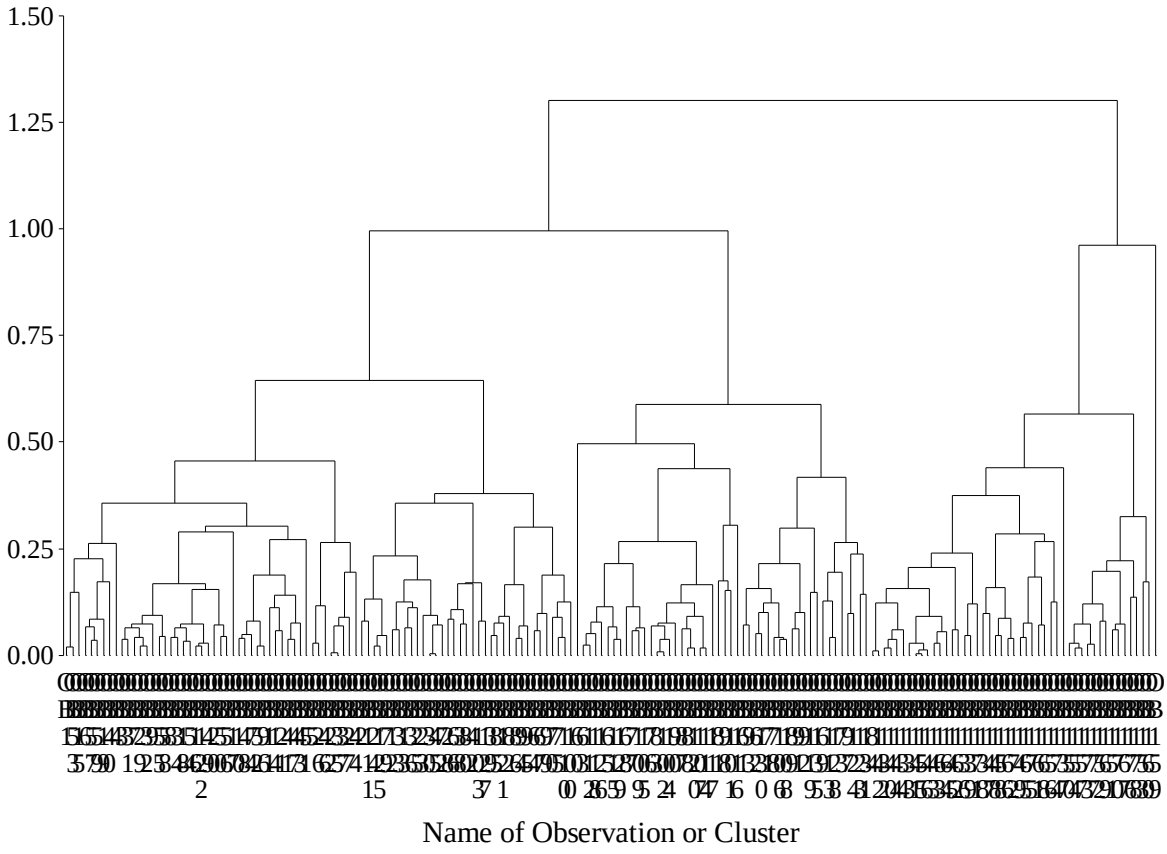
3b) Next we cluster the observations based on these two principal components using average linkage.

Exercise 3



The CCC and pseudo F plots both show high points at 3 clusters, and pseudo t-squared has a low point at 3 with a big jump at 2. These all suggest 3 as the best choice of number of clusters. In the dendrogram, we might consider 4 but notice that one of the 4 clusters would only have 1 observation. Three clusters omitting that isolated point would be a good choice based on the dendrogram.

Exercise 3



3c) Frequency analysis will help determine how well the alcohol types are separated by these clusters.

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

Alcohol 3 is very well separated based on the frequency table. Nearly all of the alcohol 3 observations are in cluster 1. Only two alcohol 3 observations are in another cluster, and only 1 observation from a different alcohol type is in cluster 1.

Separation of alcohol types 1 and 2 are not very good. Most of alcohol 2 is in cluster 3, but roughly a third of alcohol 2 observations are grouped with all of the alcohol 1 observations in cluster 2.

Exercise 4

4a) Results for a stepwise selection follow. Based on these results, we will want to use all of our predictors for a discriminant analysis for alcohol type.

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

Step	**Number In**	**Entered**	**Removed**	**Average Squared Canonical Correlation**	**Pr > ASCC**
1	1	nonflavanoid_phenols		0.36388775	<.0001
2	2	hue		0.62136638	<.0001
3	3	malic_acid		0.73590105	<.0001
4	4	magnesium		0.75251993	<.0001
5	5	alcalinity_ash		0.78878774	<.0001
6	6	od280_od315		0.79933202	<.0001
7	7	proline		0.80706733	<.0001
8	8	ash		0.81176624	<.0001
9	9	proanthocyanins		0.81553790	<.0001

Exercise 4

4b) Results for discrimination based on all predictors follow. There is a noticeable, but not huge, difference in the number of observations from each group. We will use proportional priors, but equal priors could also be reasonable here.

The Chi-square test is also highly significant, indicating that we need to account for differences of covariance across groups, and use quadratic discriminant analysis. The MANOVA tests are all highly significant, indicating that there is noticeable difference in values for at least some of the predictors for at least some of the groups. Therefore, we should stand a chance of discriminating between some groups based on some predictors in our model.

Test of Homogeneity of Within Covariance Matrices

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

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

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.					

Classification Summary for Calibration Data: WORK.WINE
Cross-validation Summary using Quadratic Discriminant Function

Exercise 4

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	0.33146	0.39888	0.26966	

Error Count Estimates for alcohol				
	1	**2**	**3**	**Total**
Rate	0.0339	0.0423	0.0000	0.0281
Priors	0.3315	0.3989	0.2697	

4c) A look at the cross-validation error estimates shows that this model worked quite well, with an overall error rate under 3% and all individual error rates under 5%. All of the type 3 alcohols are correctly classified, only two type 1 alcohols are misclassified as type 2 and three type 2 misclassified as type 1.

The classification is much more successful than clustering based on the first two principal components at identifying the three groups. This should not be too surprising given that the discrimination used all variables in the data rather than just the two most prominent underlying features, and discriminant analysis trains based on known groups while clustering finds groups without using any knowledge of existing classifications.