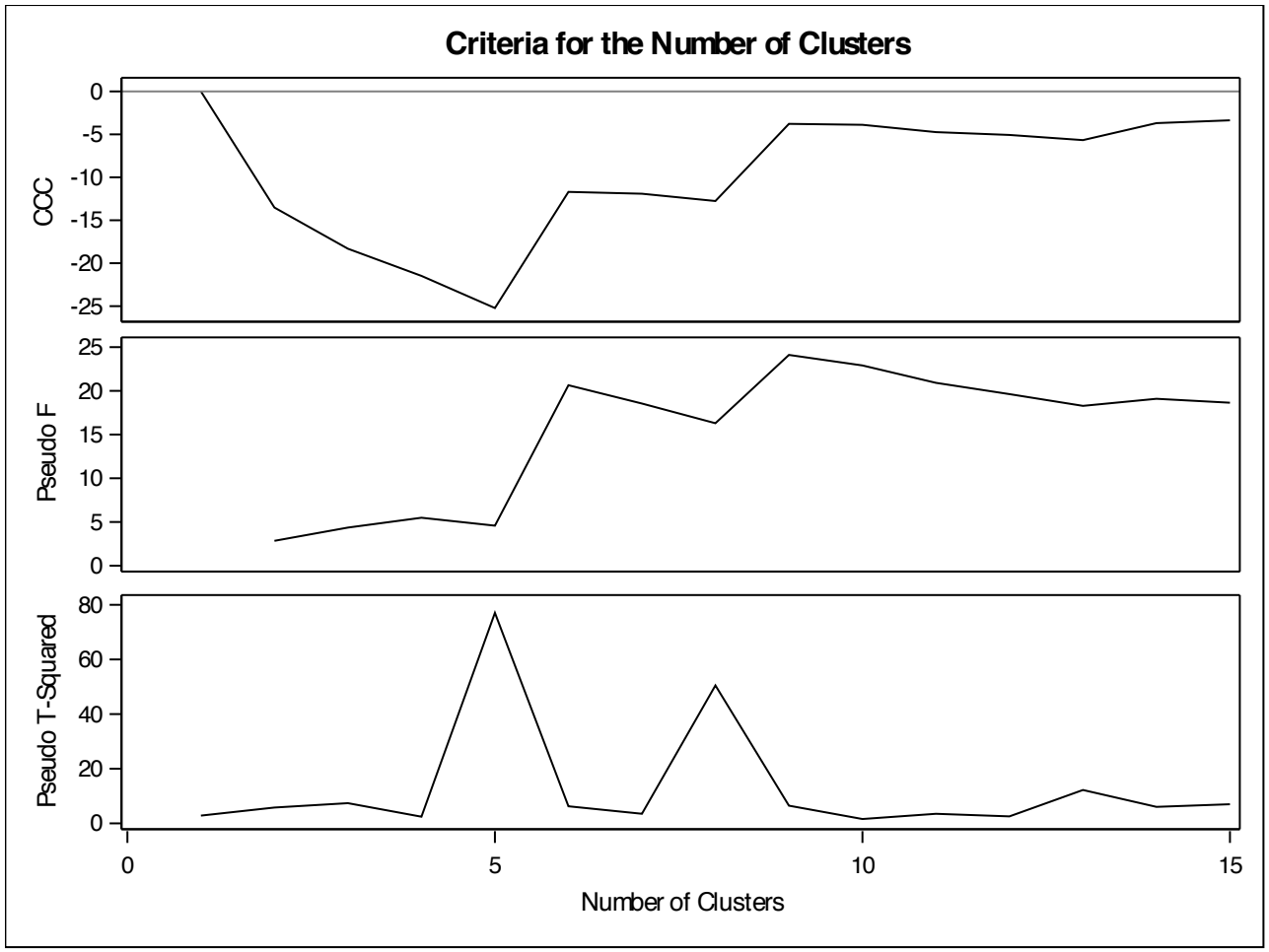
HW6

Jiangyan Feng (Netid: jf8)

Exercise 1

a) The average linkage cluster analysis results are shown as below. Based on the CCC result, the CCC is below 0 all the time. So the clustering analysis is not good. But based on the CCC, pseudo F and T-Squared plots, 9 clusters are suitable since it has high CCC, pseudo F, and low T-Squared scores.



b) The comparison between alcohol groups with clustering results is shown as below. Alcohol 1 and alcohol 3 are separated well. Although alcohol 2 is mainly clustered into cluster 2 (52 out of 71), alcohol 2 data is spread out among different clusters.

By checking the mean value of different clusters, cluster 1 is low ash, high total_phenols, high color, high hue, and high proine. Cluster 2 is high magnesium, low total_phenol, high color, low hue, and high proine. Cluster 3 is high ash, high magnesium, low total_phenol, low

color, high hue, and low proline. Since alcohol 1 and 3 are separated well by cluster 1 and 3, this shows that the difference in ash, total_phenols, color, and proine distinguish these two types of alcohols very well.

Table of CLUSTER by alcohol				
CLUSTER	**alcohol**			
Frequency	**1**	**2**	**3**	**Total**
1	56	3	0	59
2	2	52	0	54
3	0	8	46	54
4	0	0	2	2
5	0	3	0	3
6	1	2	0	3
7	0	1	0	1
8	0	1	0	1
9	0	1	0	1
Total	59	71	48	178

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	59	13.7433898	0.4834631	12.3700000	14.8300000
ash	59	1.9845763	0.7071539	1.0100000	4.0400000
alcalinity_ash	59	2.4466102	0.2199421	1.7000000	2.8700000
magnesium	59	17.0474576	2.5561525	11.2000000	25.0000000
total_phenols	59	105.4237288	11.0612766	78.0000000	132.0000000
flavanoids	59	2.8557627	0.3328570	2.2000000	3.8800000
nonflavanoid_phenols	59	3.0023729	0.3821100	2.3300000	3.9300000
proanthocyanins	59	0.2869492	0.0675751	0.1700000	0.5000000
color	59	1.9262712	0.4026545	1.2500000	2.9600000
hue	59	5.5630508	1.2091879	3.3800000	8.9000000
od280_od315	59	1.0677966	0.1227005	0.8200000	1.3600000
proline	59	3.1574576	0.3657896	2.3000000	4.0000000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	54	12.2537037	0.5150208	11.0300000	13.6700000
ash	54	1.9522222	1.0268962	0.7400000	5.8000000
alcalinity_ash	54	2.2540741	0.2580735	1.7000000	2.9200000
magnesium	54	20.5166667	2.6481571	15.5000000	28.5000000
total_phenols	54	89.2407407	7.5258929	70.0000000	108.0000000
flavanoids	54	2.2783333	0.4985517	1.3800000	3.5200000
nonflavanoid_phenols	54	2.1448148	0.5100046	1.2500000	3.7500000
proanthocyanins	54	0.3609259	0.1144184	0.1300000	0.6600000
color	54	1.6083333	0.4185183	0.7300000	2.9100000
hue	54	2.9274074	0.7632359	1.2800000	4.6800000
od280_od315	54	1.0583333	0.2064228	0.6900000	1.7100000
proline	54	2.9016667	0.3704192	2.0600000	3.6400000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	54	13.0274074	0.5609500	11.6600000	14.1600000
ash	54	3.2174074	1.1457521	0.9400000	5.6500000
alcalinity_ash	54	2.3881481	0.2015326	1.9200000	2.8600000
magnesium	54	20.8000000	2.5618390	16.0000000	27.0000000
total_phenols	54	99.4074074	10.9381081	80.0000000	123.0000000
flavanoids	54	1.6572222	0.3126781	0.9800000	2.5300000
nonflavanoid_phenols	54	0.8453704	0.3378216	0.3400000	1.5900000
proanthocyanins	54	0.4503704	0.1233948	0.1700000	0.6300000
color	54	1.0711111	0.3261882	0.4100000	1.8700000
hue	54	6.6696296	2.3261880	2.6500000	10.8000000
od280_od315	54	0.7317778	0.1645695	0.4800000	1.2500000
proline	54	1.7198148	0.2951398	1.2700000	2.5200000

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	2	13.9100000	0.6081118	13.4800000	14.3400000
ash	2	1.6750000	0.0070711	1.6700000	1.6800000
alcalinity_ash	2	2.6700000	0.0424264	2.6400000	2.7000000
magnesium	2	23.7500000	1.7677670	22.5000000	25.0000000
total_phenols	2	93.5000000	6.3639610	89.0000000	98.0000000
flavanoids	2	2.7000000	0.1414214	2.6000000	2.8000000
nonflavanoid_phenols	2	1.2050000	0.1484924	1.1000000	1.3100000
proanthocyanins	2	0.5250000	0.0070711	0.5200000	0.5300000
color	2	2.4950000	0.2899138	2.2900000	2.7000000
hue	2	12.3750000	0.8838835	11.7500000	13.0000000
od280_od315	2	0.5700000	0	0.5700000	0.5700000
proline	2	1.8700000	0.1272792	1.7800000	1.9600000

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	3	12.3366667	0.1301281	12.2100000	12.4700000
ash	3	1.2333333	0.2676440	0.9900000	1.5200000
alcalinity_ash	3	1.9666667	0.2254625	1.7500000	2.2000000
magnesium	3	16.8666667	2.1007935	14.8000000	19.0000000
total_phenols	3	149.6666667	13.0511813	136.0000000	162.0000000
flavanoids	3	2.0833333	0.3617089	1.8500000	2.5000000
nonflavanoid_phenols	3	1.8000000	0.4968903	1.2800000	2.2700000
proanthocyanins	3	0.2700000	0.1135782	0.1400000	0.3500000
color	3	2.8466667	0.3971566	2.5000000	3.2800000
hue	3	2.9500000	0.4092676	2.6000000	3.4000000
od280_od315	3	1.1666667	0.1101514	1.0600000	1.2800000
proline	3	2.6700000	0.3815757	2.3100000	3.0700000

CLUSTER=6

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	3	12.5333333	0.8434651	11.5600000	13.0500000
ash	3	1.9233333	0.2193931	1.6700000	2.0500000
alcalinity_ash	3	3.0166667	0.3608786	2.6000000	3.2300000
magnesium	3	27.8333333	2.5658007	25.0000000	30.0000000
total_phenols	3	127.3333333	10.4083300	119.0000000	139.0000000
flavanoids	3	3.0366667	0.3572581	2.6300000	3.3000000
nonflavanoid_phenols	3	3.5500000	1.3291727	2.6800000	5.0800000
proanthocyanins	3	0.3833333	0.1501111	0.2100000	0.4700000
color	3	1.9166667	0.0450925	1.8700000	1.9600000
hue	3	4.3100000	1.4680940	3.3500000	6.0000000
od280_od315	3	1.1233333	0.1900877	0.9300000	1.3100000
proline	3	3.4633333	0.2470493	3.2000000	3.6900000

CLUSTER=7

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	1	11.8100000	.	11.8100000	11.8100000
ash	1	2.1200000	.	2.1200000	2.1200000
alcalinity_ash	1	2.7400000	.	2.7400000	2.7400000
magnesium	1	21.5000000	.	21.5000000	21.5000000
total_phenols	1	134.0000000	.	134.0000000	134.0000000
flavanoids	1	1.6000000	.	1.6000000	1.6000000
nonflavanoid_phenols	1	0.9900000	.	0.9900000	0.9900000
proanthocyanins	1	0.1400000	.	0.1400000	0.1400000
color	1	1.5600000	.	1.5600000	1.5600000
hue	1	2.5000000	.	2.5000000	2.5000000
od280_od315	1	0.9500000	.	0.9500000	0.9500000
proline	1	2.2600000	.	2.2600000	2.2600000

CLUSTER=8

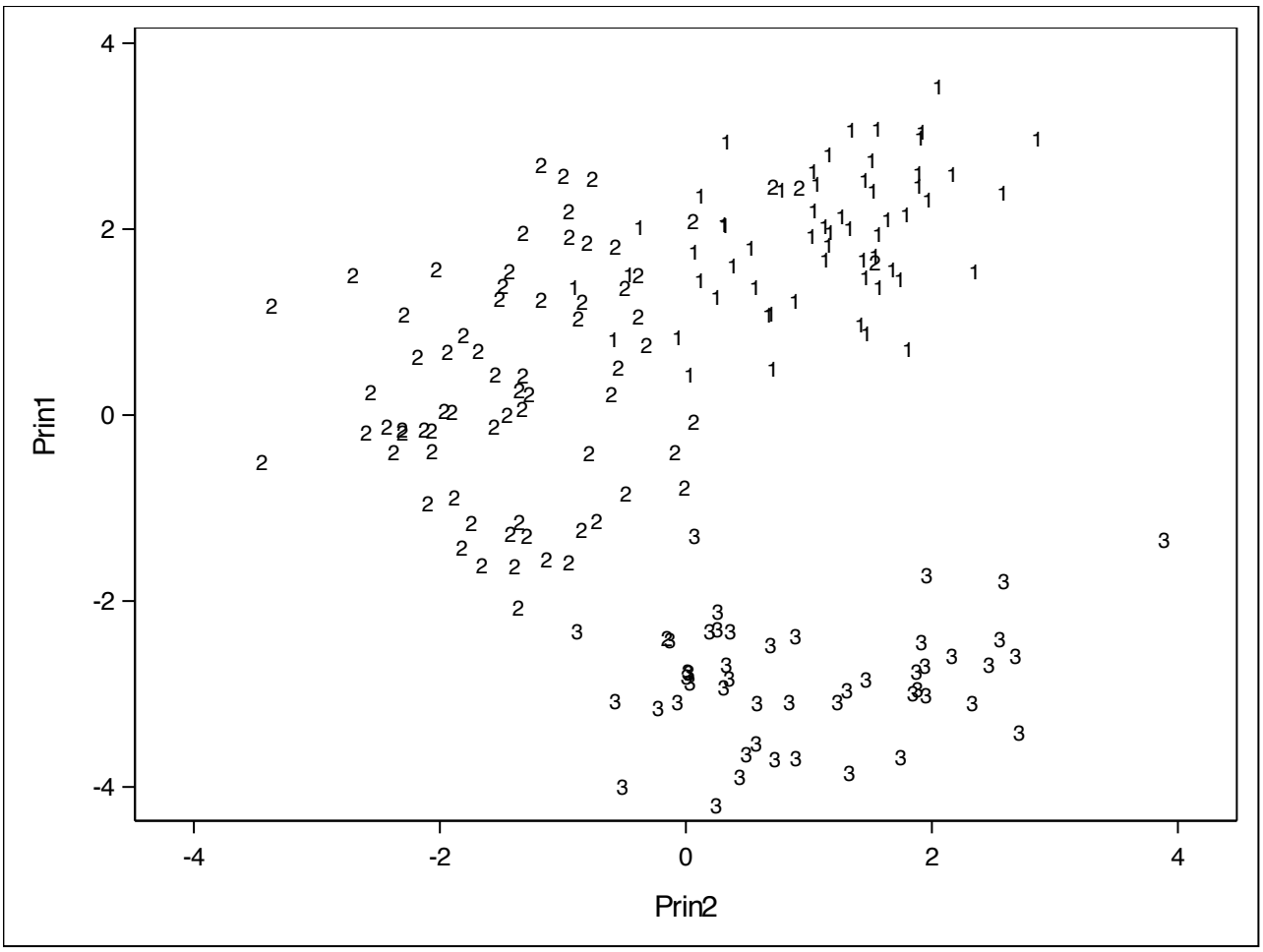
Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	1	11.4600000	.	11.4600000	11.4600000
ash	1	3.7400000	.	3.7400000	3.7400000
alcalinity_ash	1	1.8200000	.	1.8200000	1.8200000
magnesium	1	19.5000000	.	19.5000000	19.5000000
total_phenols	1	107.0000000	.	107.0000000	107.0000000
flavanoids	1	3.1800000	.	3.1800000	3.1800000
nonflavanoid_phenols	1	2.5800000	.	2.5800000	2.5800000
proanthocyanins	1	0.2400000	.	0.2400000	0.2400000
color	1	3.5800000	.	3.5800000	3.5800000
hue	1	2.9000000	.	2.9000000	2.9000000
od280_od315	1	0.7500000	.	0.7500000	0.7500000
proline	1	2.8100000	.	2.8100000	2.8100000

CLUSTER=9

Variable	N	Mean	Std Dev	Minimum	Maximum
malic_acid	1	12.3700000	.	12.3700000	12.3700000
ash	1	0.9400000	.	0.9400000	0.9400000
alcalinity_ash	1	1.3600000	.	1.3600000	1.3600000
magnesium	1	10.6000000	.	10.6000000	10.6000000
total_phenols	1	88.0000000	.	88.0000000	88.0000000
flavanoids	1	1.9800000	.	1.9800000	1.9800000
nonflavanoid_phenols	1	0.5700000	.	0.5700000	0.5700000
proanthocyanins	1	0.2800000	.	0.2800000	0.2800000
color	1	0.4200000	.	0.4200000	0.4200000
hue	1	1.9500000	.	1.9500000	1.9500000
od280_od315	1	1.0500000	.	1.0500000	1.0500000
proline	1	1.8200000	.	1.8200000	1.8200000

Exercise 2

a) The principal component analysis results are shown as below. This shows that these three types can be separated well by the two principal components. Alcohol 1 has high prin1 and high prin2. Alcohol 2 has high prin1 and low prin2. Alcohol 3 has low prin1 and high prin2.

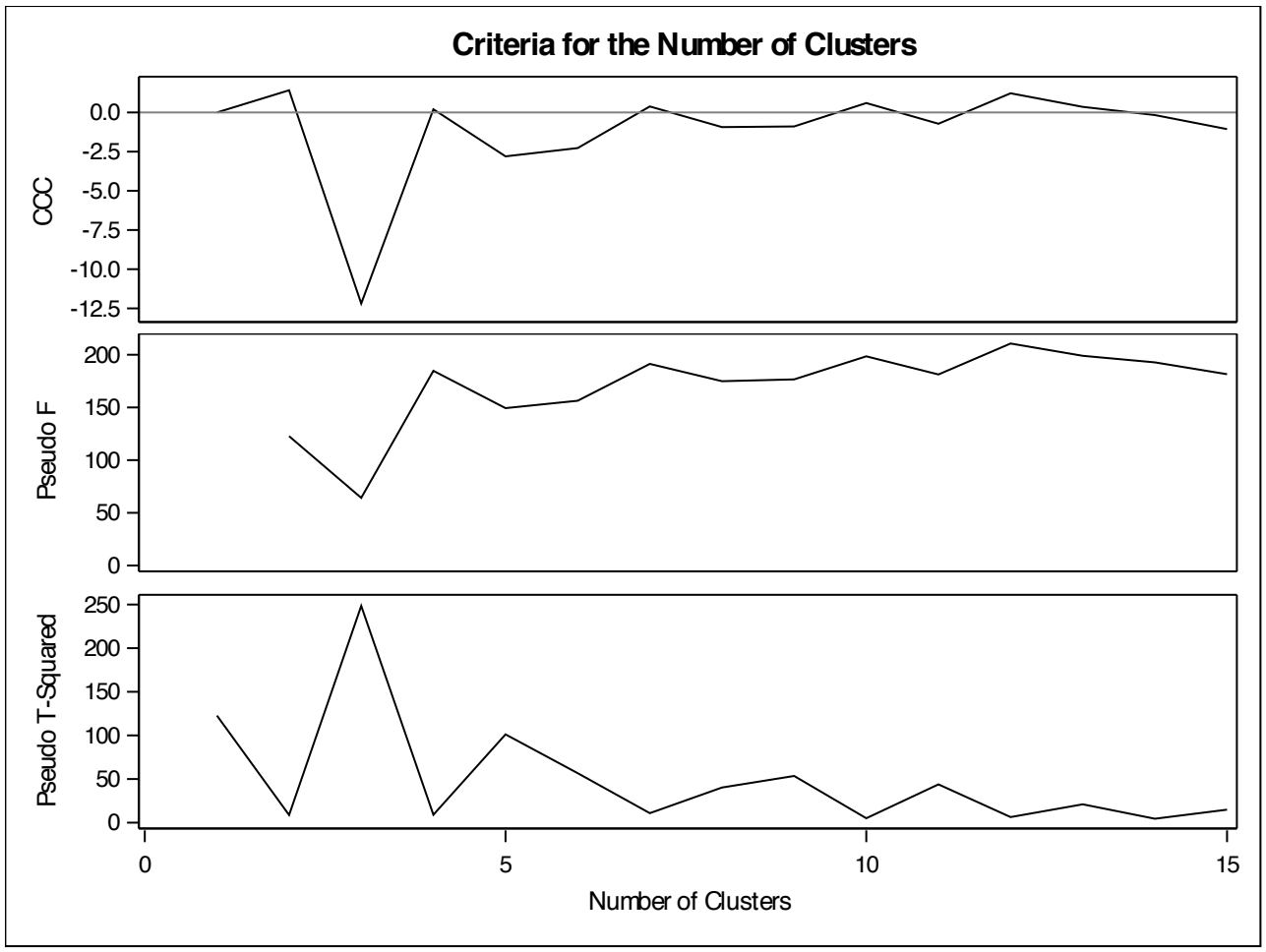


b) The average linkage cluster analysis results are shown as below. Based on the CCC, pseudo F and T-Squared plots, 4 clusters are suitable since it has high CCC, pseudo F, and low T-Squared scores.

Cluster History								
Number of Clusters	**Clusters Joined**		**Freq**	**Semipartial R-Square**	**R-Square**	**Approximate Expected R-Square**	**Cubic Clustering Criterion**	**Pseudo F Statistic**
15	CL36	CL22	19	0.0039	.940	.944	-1.1	182
14	CL26	OB56	15	0.0012	.939	.939	-.17	193
13	CL43	CL38	11	0.0032	.935	.934	0.36	199
12	CL18	CL35	27	0.0022	.933	.927	1.22	211
11	CL16	CL14	54	0.0175	.916	.920	-.72	181
10	CL29	OB178	3	0.0016	.914	.911	0.60	198
9	CL21	CL17	41	0.0209	.893	.899	-.89	177
8	CL24	CL13	30	0.0151	.878	.885	-.93	175
7	CL11	CL20	57	0.0077	.870	.867	0.38	191

Cluster History								
Number of Clusters	Clusters Joined		Freq	Semipartial R-Square	R-Square	Approximate Expected R-Square	Cubic Clustering Criterion	Pseudo F Statistic
6	CL8	CL9	71	0.0507	.820	.843	-2.3	156
5	CL12	CL15	46	0.0442	.775	.809	-2.8	149
4	CL6	CL10	74	0.0143	.761	.758	0.20	185
3	CL7	CL4	131	0.3379	.423	.674	-12	64.2
2	CL5	OB168	47	0.0123	.411	.382	1.42	123
1	CL3	CL2	178	0.4108	.000	.000	0.00	.

Cluster History					
Number of Clusters	Clusters Joined		Pseudo t-Squared	Norm RMS Distance	Tie
15	CL36	CL22	14.9	0.3608	
14	CL26	OB56	4.5	0.3623	
13	CL43	CL38	21.0	0.366	
12	CL18	CL35	6.4	0.3845	
11	CL16	CL14	43.9	0.4544	
10	CL29	OB178	5.0	0.4767	
9	CL21	CL17	53.5	0.5017	
8	CL24	CL13	40.3	0.5061	
7	CL11	CL20	10.9	0.5723	
6	CL8	CL9	56.7	0.6422	
5	CL12	CL15	101	0.6523	
4	CL6	CL10	9.0	0.7987	
3	CL7	CL4	248	1.0749	
2	CL5	OB168	8.8	1.1127	
1	CL3	CL2	123	1.2412	



c) The separation performance is shown as below. This shows all 3 types of alcohol are separated well by the first 3 clusters. Cluster 1, 2, 3 represents alcohol 3, alcohol 1, and alcohol 2, respectively. The performance is much better the cluster analysis alone, which was only able to separate alcohol 1 and 3.

Table of CLUSTER by alcohol				
CLUSTER	**alcohol**			
Frequency	**1**	**2**	**3**	**Total**
1	0	1	45	46
2	53	4	0	57
3	6	66	2	74
4	0	0	1	1
Total	59	71	48	178

Exercise 3

a) The discriminant analysis results are shown as below. P value = <.0001 for the test of homogeneity, indicating different variance. Therefore, QDA is better. According to the MANOVA test, P value = <.0001, suggesting that it is possible to discriminate between some alcohol types.

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.

Generalized Squared Distance to alcohol			
From alcohol	1	2	3
1	-20.72044	11.34703	297.50150
2	-0.98684	-12.09120	156.64235
3	144.27015	56.86134	-19.96905

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
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into alcohol				
From alcohol	**1**	**2**	**3**	**Total**
1	59 100.00	0 0.00	0 0.00	59 100.00
2	1 1.41	70 98.59	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	0.0000	0.0141	0.0000	0.0047
Priors	0.3333	0.3333	0.3333	

Total Sample Size	178	**DF Total**	177
Variables	12	**DF Within Classes**	175
Classes	3	**DF Between Classes**	2

Number of Observations Read	178
Number of Observations Used	178

Class Level Information					
alcohol	**Variable Name**	**Frequency**	**Weight**	**Proportion**	**Prior Probability**
1	1	59	59.0000	0.331461	0.333333
2	2	71	71.0000	0.398876	0.333333
3	3	48	48.0000	0.269663	0.333333

Within Covariance Matrix Information		
alcohol	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
1	12	-20.72044
2	12	-12.09120
3	12	-19.96905
Pooled	12	-13.28577

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.

Generalized Squared Distance to alcohol			
From alcohol	1	2	3
1	-20.72044	11.34703	297.50150
2	-0.98684	-12.09120	156.64235
3	144.27015	56.86134	-19.96905

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
Resubstitution Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into alcohol				
From alcohol	**1**	**2**	**3**	**Total**
1	59 100.00	0 0.00	0 0.00	59 100.00
2	1 1.41	70 98.59	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	0.0000	0.0141	0.0000	0.0047
Priors	0.3333	0.3333	0.3333	

b) The cross-validation results are shown as below. The error rates for 3 types of alcohol are 0.0339, 0.0423, and 0. This shows that alcohol 3 is the easiest to separate out. On average, the error rate is 0.0254, suggesting that the discrimination matches the groups very well. This is consistent with the analysis results from exercise 2c. In both analysis, type 3 alcohol is the easiest to separate out.

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

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	

Exercise 4

a) The stepwise selection analysis results are shown as below. In total, 9 predictors are chosen for the classification: nonflavanoid_phenols, hue, malic_acid, magnesium, alcalinity_ash, od280_od315, proline, ash, and proanthocyanins.

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

b) Similarities: In both exercise 2c and this analysis, the separation performance for 3 types of alcohol is: alcohol 3 > alcohol 2 > alcohol 1.
Dissimilarities: In this analysis, the error rate for alcohol 3 is 0, indicating that alcohol 3 can be accurately separated out. However, in exercise 2c, the error rate is 0.06.
Quality change: the average error rate decreases from 0.0254 to 0.0216. For type 2 alcohol, the error rate decreases from 0.0423 to 0.0141, suggesting a big improvement for type 2 alcohol. However, the error rate for type 1 alcohol increases from 0.0339 to 0.0508, suggesting the decrease of the quality.

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

Number of Observations and Percent Classified into alcohol				
From alcohol	**1**	**2**	**3**	**Total**
1	56 94.92	3 5.08	0 0.00	59 100.00
2	0 0.00	70 98.59	1 1.41	71 100.00
3	0 0.00	0 0.00	48 100.00	48 100.00
Total	56 31.46	73 41.01	49 27.53	178 100.00
Priors	0.33333	0.33333	0.33333	

Error Count Estimates for alcohol				
	1	**2**	**3**	**Total**
Rate	0.0508	0.0141	0.0000	0.0216
Priors	0.3333	0.3333	0.3333	