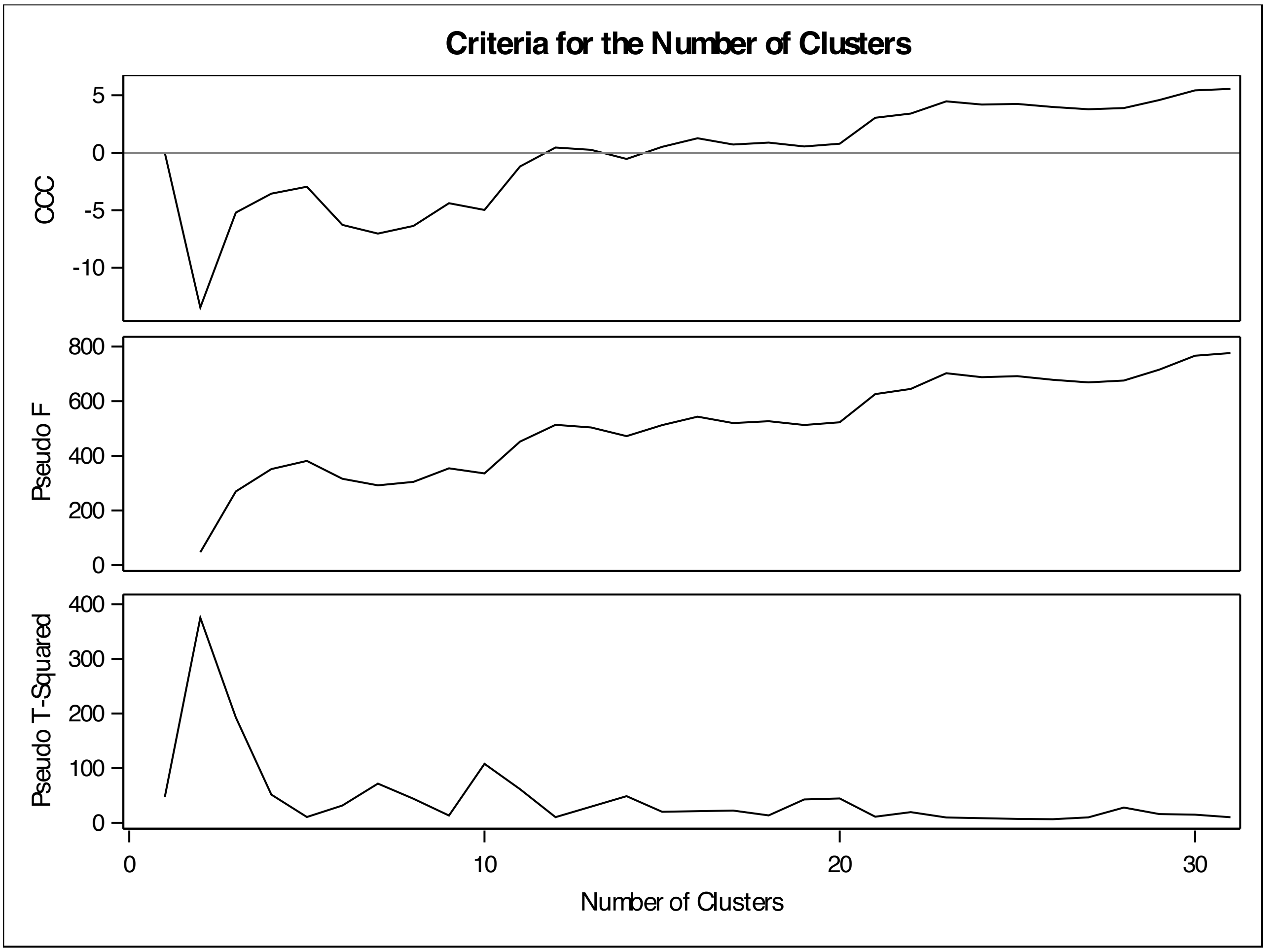
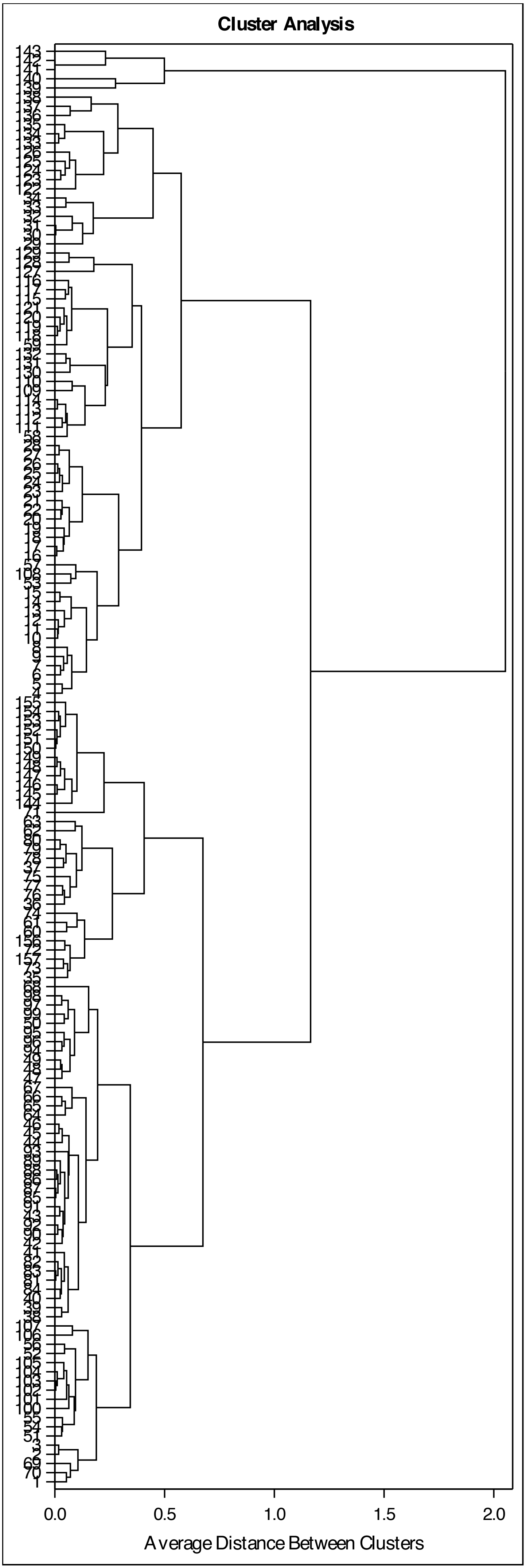
Exercise 1a):

The CLUSTER Procedure
Average Linkage Cluster Analysis




Cluster Analysis
0.0
0.5
1.0
1.5
2.0
Average Distance Between Clusters

From dendrogram plot we can see that we may have 7 distinct clusters. We can see that ccc plot might not be a valid test here; it has a lot of negative values, that indicates the existence of outliers. The Pseudo F plot has a peak at 4. The pseudo t-squared has low points at 4 and 9. But, overall it is reasonable to have 7 clusters, since we also have 7 types of species.

Exercise 1b):

The ANOVA Procedure

Dependent Variable: Weight

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	17264179.84	2877363.31	153.06	<.0001
Error	150	2819885.10	18799.23		
Corrected Total	156	20084064.94			

The ANOVA Procedure

Levene's Test for Homogeneity of Weight Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
CLUSTER	5	6.084E10	1.217E10	17.91	<.0001
Error	149	1.012E11	6.7937E8		

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Weight

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
1 - 7	500.00	125.91	874.09	***
1 - 2	645.83	356.07	935.60	***
1 - 6	823.36	556.45	1090.28	***
1 - 5	989.14	745.42	1232.87	***
1 - 3	1409.13	1166.17	1652.09	***
1 - 4	1558.20	1310.42	1805.97	***
7 - 1	-500.00	-874.09	-125.91	***
7 - 2	145.83	-188.76	480.43	
7 - 6	323.36	8.36	638.37	***
7 - 5	489.14	193.52	784.76	***
7 - 3	909.13	614.14	1204.11	***
7 - 4	1058.20	759.23	1357.16	***
2 - 1	-645.83	-935.60	-356.07	***
2 - 7	-145.83	-480.43	188.76	
2 - 6	177.53	-30.45	385.51	
2 - 5	343.31	166.07	520.55	***
2 - 3	763.29	587.11	939.48	***
2 - 4	912.36	729.59	1095.13	***
6 - 1	-823.36	-1090.28	-556.45	***
6 - 7	-323.36	-638.37	-8.36	***
6 - 2	-177.53	-385.51	30.45	
6 - 5	165.78	29.06	302.50	***
6 - 3	585.76	450.41	721.11	***
6 - 4	734.83	591.02	878.65	***
5 - 1	-989.14	-1232.87	-745.42	***
5 - 7	-489.14	-784.76	-193.52	***
5 - 2	-343.31	-520.55	-166.07	***
5 - 6	-165.78	-302.50	-29.06	***
5 - 3	419.98	339.48	500.49	***
5 - 4	569.05	475.01	663.10	***
3 - 1	-1409.13	-1652.09	-1166.17	***
3 - 7	-909.13	-1204.11	-614.14	***
3 - 2	-763.29	-939.48	-587.11	***

The ANOVA Procedure

Tukey's Studentized Range (HSD) Test for Weight

Comparisons significant at the 0.05 level are indicated by ***.				
CLUSTER Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
3 - 6	-585.76	-721.11	-450.41	***
3 - 5	-419.98	-500.49	-339.48	***
3 - 4	149.07	57.04	241.10	***
4 - 1	-1558.20	-1805.97	-1310.42	***
4 - 7	-1058.20	-1357.16	-759.23	***
4 - 2	-912.36	-1095.13	-729.59	***
4 - 6	-734.83	-878.65	-591.02	***
4 - 5	-569.05	-663.10	-475.01	***
4 - 3	-149.07	-241.10	-57.04	***

The parametric ANOVA is significant and there is an indication of at least one pair of groups that has unequal variances based on Levene’s test. And model is significant. We get significant differences in weights except cluster pairs 7 and 2, 2 and 6.

Exercise 1c):

The FREQ Procedure

Table of Species by CLUSTER								
Species	**CLUSTER**							
Frequency	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**Total**
Bream	0	6	3	0	25	0	0	34
Parkki	0	0	7	4	0	0	0	11
Perch	0	0	27	10	14	5	0	56
Pike	3	0	0	0	6	6	2	17
Roach	0	0	15	3	1	0	0	19
Smelt	0	0	0	14	0	0	0	14
Whitefish	0	0	3	0	3	0	0	6
Total	3	6	55	31	49	11	2	157

Cluster 1 contains only Pike species, Cluster 2 only Bream. Cluster 3 has mostly Perch species, and includes some Roach(15), Parkki(7), Bream and Whitefish(3). Cluster 4 contains Smelt(14), Perch(10), Parkki(4) and Roach(3). Cluster 5 contains most of Bream, some of Perch(14), Pike(6), Whitefish(3) and one Roach. Cluster 6 contains Perch and Pike species. Cluster 7 has only 2 observations from Pike. And it matches with the results from part b).

Exercise 2a):

The STEPDISC Procedure

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	Height		0.7548	76.97	<.0001	0.24517289	<.0001	0.12580452	<.0001
2	2	Length2		0.9229	297.37	<.0001	0.01889662	<.0001	0.25885632	<.0001
3	3	Length3		0.8838	187.57	<.0001	0.00219622	<.0001	0.38415545	<.0001
4	4	Width		0.5770	33.42	<.0001	0.00092904	<.0001	0.45217362	<.0001
5	5	Length1		0.2904	9.96	<.0001	0.00065925	<.0001	0.47899426	<.0001

We can see that all the variables are significant based on stepwise selection summary.

Exercise 2b):

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
468.037998	90	<.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.

The DISCRIM Procedure

Multivariate Statistics and F Approximations

S=5 M=0 N=72

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00065925	102.37	30	586	<.0001
Pillai's Trace	2.87396556	33.79	30	750	<.0001
Hotelling-Lawley Trace	49.12356467	237.05	30	378.36	<.0001
Roy's Greatest Root	39.05707956	976.43	6	150	<.0001

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

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.FISHDATA
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species								
From Species	**Bream**	**Parkki**	**Perch**	**Pike**	**Roach**	**Smelt**	**Whitefish**	**Total**
Bream	34 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	34 100.00
Parkki	0 0.00	11 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 100.00
Perch	0 0.00	0 0.00	56 100.00	0 0.00	0 0.00	0 0.00	0 0.00	56 100.00
Pike	0 0.00	0 0.00	0 0.00	17 100.00	0 0.00	0 0.00	0 0.00	17 100.00
Roach	0 0.00	0 0.00	0 0.00	0 0.00	19 100.00	0 0.00	0 0.00	19 100.00
Smelt	0 0.00	0 0.00	3 21.43	0 0.00	0 0.00	11 78.57	0 0.00	14 100.00
Whitefish	0 0.00	0 0.00	1 16.67	0 0.00	5 83.33	0 0.00	0 0.00	6 100.00
Total	34 21.66	11 7.01	60 38.22	17 10.83	24 15.29	11 7.01	0 0.00	157 100.00
Priors	0.21656	0.07006	0.35669	0.10828	0.12102	0.08917	0.03822	

Error Count Estimates for Species								
	Bream	**Parkki**	**Perch**	**Pike**	**Roach**	**Smelt**	**Whitefish**	**Total**
Rate	0.0000	0.0000	0.0000	0.0000	0.0000	0.2143	1.0000	0.0573
Priors	0.2166	0.0701	0.3567	0.1083	0.1210	0.0892	0.0382	

Test of Homogeneity of Within Covariance Matrices indicates a significant difference in covariance matrices across groups, and SAS proceeds with QDA for the model. The MANOVA tests are significant, so it should be possible to obtain some level of discrimination based on these 5 predictors.

Exercise 2c):

The frequency table indicates that 100% were correctly classified for Species Bream, Parkki, Perch, Pike and Roach. About 79% of Species Smelt were correctly classified. And we do not have any correctly classified for Species Whitefish. Consequently we have 0% of misclassified observations for Species Bream, Parkki, Perch, Pike and Roach. Nearly 21% of Species Smelt are estimated to be misclassified by the model. We also can see that 100% of Species Whitefish were misclassified. So this group can be easily confused. This yields an overall error estimate of about 5.73%.

Exercise 3a):

The STEPDISC Procedure

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	Height		0.7548	76.97	<.0001	0.24517289	<.0001	0.12580452	<.0001
2	2	Length2		0.9229	297.37	<.0001	0.01889662	<.0001	0.25885632	<.0001
3	3	Length3		0.8838	187.57	<.0001	0.00219622	<.0001	0.38415545	<.0001
4	4	Width		0.5770	33.42	<.0001	0.00092904	<.0001	0.45217362	<.0001
5	5	Weight		0.4435	19.39	<.0001	0.00051702	<.0001	0.49437664	<.0001
6	6	Length1		0.2972	10.22	<.0001	0.00036339	<.0001	0.51673267	<.0001

Again we can see that all the variables are significant based on stepwise selection summary.

Exercise 3b):

The DISCRIM Procedure

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
749.819804	126	<.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.

The DISCRIM Procedure

Multivariate Statistics and F Approximations

S=6 M=-0.5 N=71.5

Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.00036339	90.09	36	639.5	<.0001
Pillai's Trace	3.10039600	26.73	36	900	<.0001
Hotelling-Lawley Trace	52.11364485	208.02	36	410.72	<.0001
Roy's Greatest Root	39.17299881	979.32	6	150	<.0001

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

The DISCRIM Procedure
Classification Summary for Calibration Data: WORK.FISHDATA
Cross-validation Summary using Quadratic Discriminant Function

Number of Observations and Percent Classified into Species

From Species	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Bream	34 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	34 100.00
Parkki	0 0.00	11 100.00	0 0.00	0 0.00	0 0.00	0 0.00	0 0.00	11 100.00
Perch	0 0.00	0 0.00	56 100.00	0 0.00	0 0.00	0 0.00	0 0.00	56 100.00
Pike	0 0.00	0 0.00	0 0.00	17 100.00	0 0.00	0 0.00	0 0.00	17 100.00
Roach	0 0.00	0 0.00	0 0.00	0 0.00	19 100.00	0 0.00	0 0.00	19 100.00
Smelt	0 0.00	0 0.00	3 21.43	0 0.00	0 0.00	11 78.57	0 0.00	14 100.00
Whitefish	0 0.00	0 0.00	3 50.00	0 0.00	3 50.00	0 0.00	0 0.00	6 100.00
Total	34 21.66	11 7.01	62 39.49	17 10.83	22 14.01	11 7.01	0 0.00	157 100.00
Priors	0.21656	0.07006	0.35669	0.10828	0.12102	0.08917	0.03822	

Error Count Estimates for Species

	Bream	Parkki	Perch	Pike	Roach	Smelt	Whitefish	Total
Rate	0.0000	0.0000	0.0000	0.0000	0.0000	0.2143	1.0000	0.0573
Priors	0.2166	0.0701	0.3567	0.1083	0.1210	0.0892	0.0382	

Test of Homogeneity of Within Covariance Matrices indicates a significant difference in covariance matrices across groups, and SAS proceeds with QDA for the model. The MANOVA tests are significant, so it should be possible to obtain some level of discrimination based on these 6 predictors.

Exercise 3c):

The frequency table indicates that 100% were correctly classified for Species Bream, Parkki, Perch, Pike and Roach. About 79% of Species Smelt were correctly classified. And we do not have any correctly classified for Species Whitefish. Consequently we have 0% of misclassified observations for Species Bream, Parkki, Perch, Pike and Roach. Nearly 21% of Species Smelt are estimated to be misclassified by the model. We also can see that 100% of Species Whitefish were misclassified. So this group can be easily confused. We also got the singular within-class covariance matrix for Whitefish class, this might indicate that the values are identical to other group fishes(f.e. Roach). This yields an overall error estimate of about 5.73%, so two models represented the same results.