Final Project

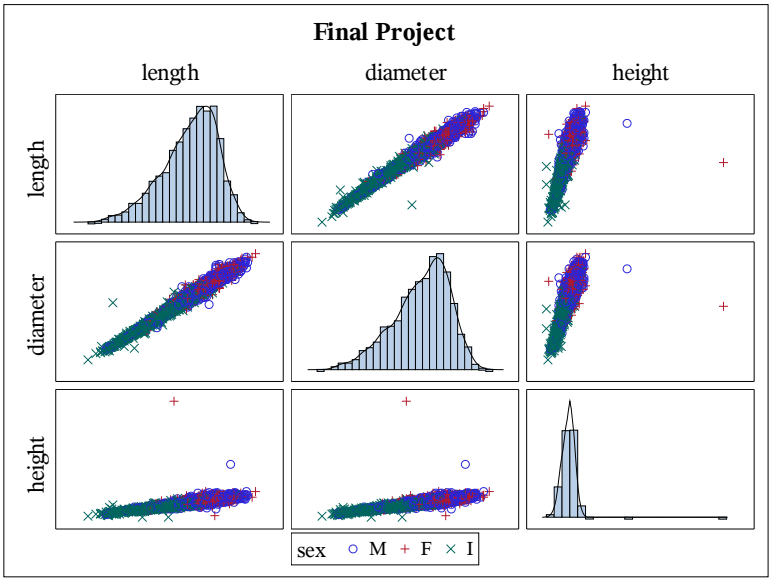
Yilun Fu

Introduction

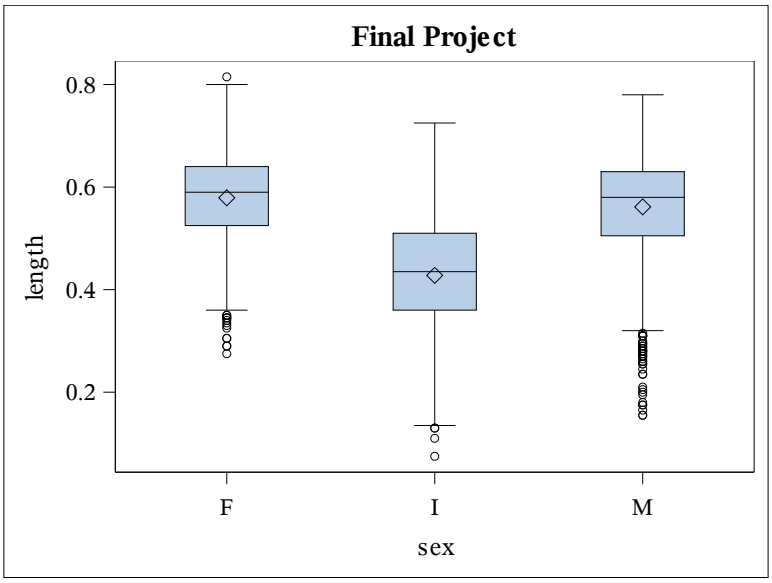
The data is based on the abalone data set from UCI's Machine Learning Database with totally 4177 rows. It contains some physical measurements about abalone like 'Sex', 'Length', 'Diameter', 'Height', 'whole_weight', 'meat_weight', 'gut_weight' and 'shell_weight'. There is also one variable named 'Rings' which +1.5 gives the age in years. 'Rings' is a integer and 'Sex' is a nominal variable, the rest are all continuous variables.

Question 1

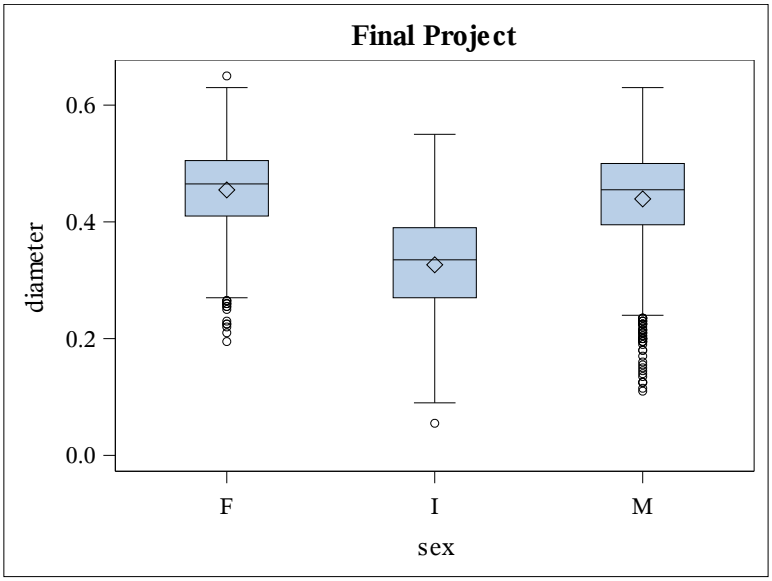
This is a general descriptive overview of the sizes and weights of abalone by sex.



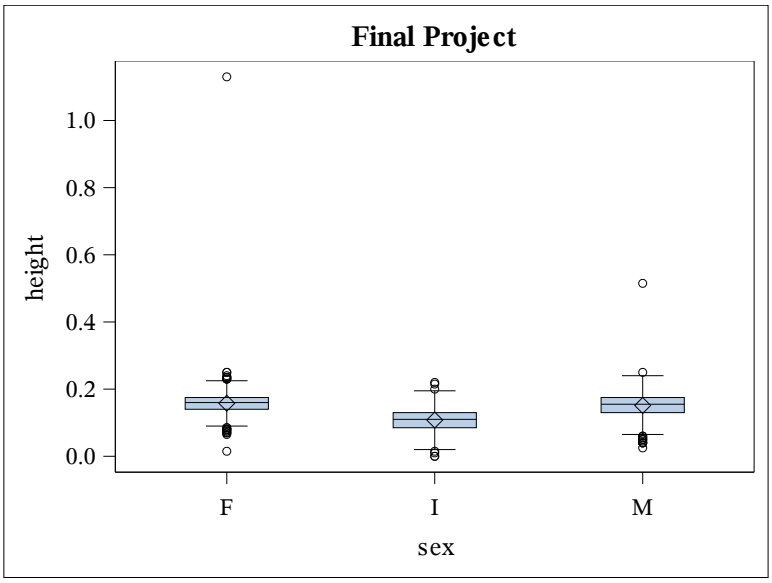
It is easy to notice that infant has minimum length, diameter and height compared to female or male.



From this box plot of variable 'length', there is no such big difference between male and female. However, infant has noticeable relatively low length.

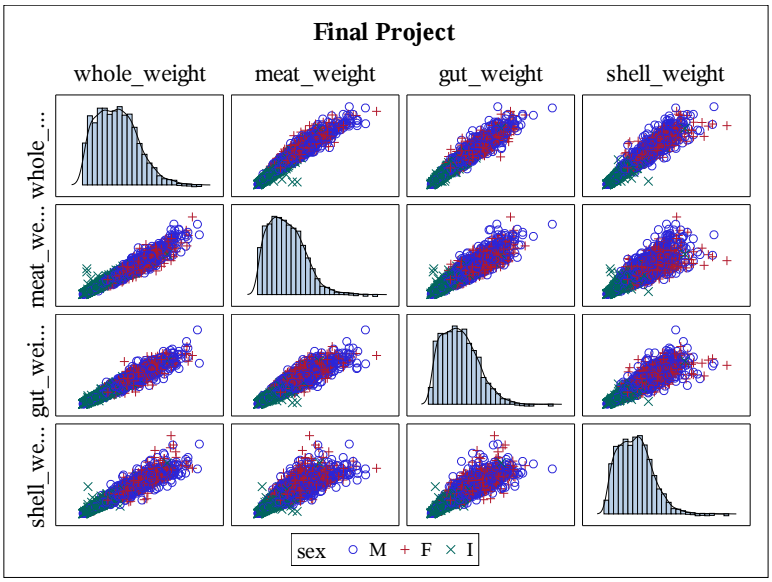


As for diameter, just like length and there is no such big difference between male and female. Maybe the diameter of female is slightly than that of male. But infant has noticeable relatively low diameter compared to female or male.

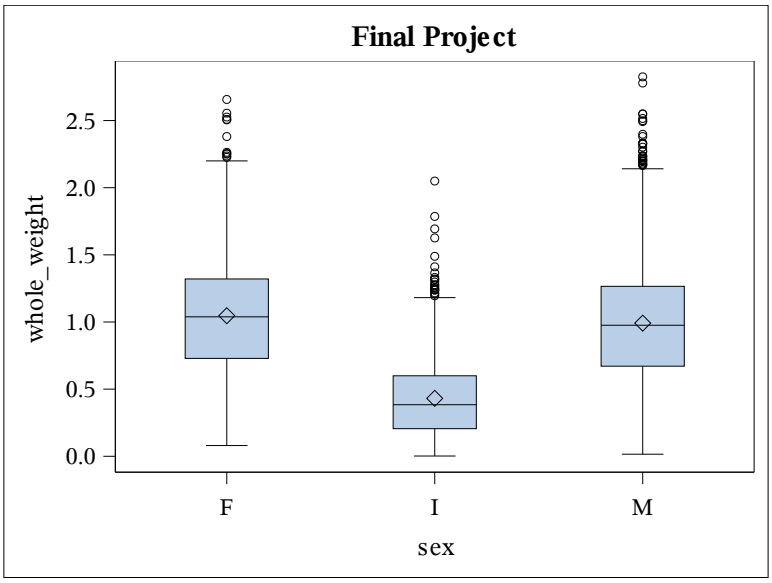


Similar conclusion for height, there is no such big difference between male and female. But infant has noticeable relatively low height compared to female or male.

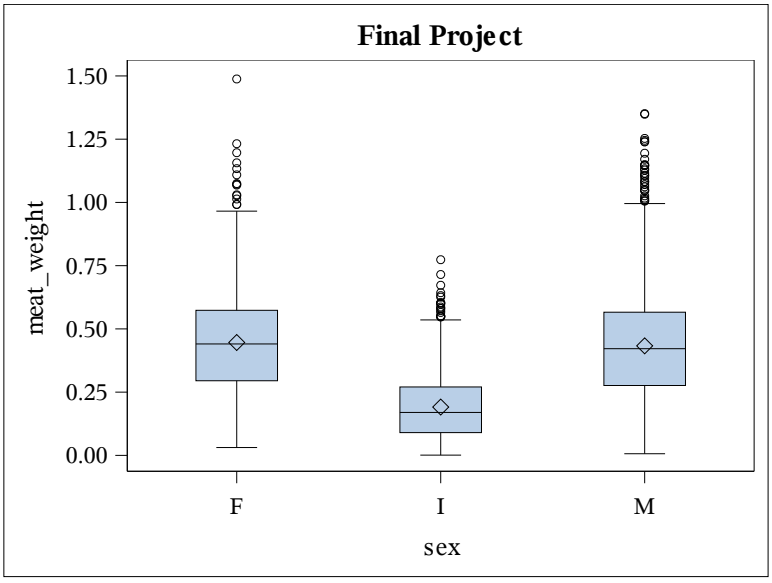
We can conclude that noticeable difference in sizes between infant and female or male can be observed. While there is nearly no difference between female and male.



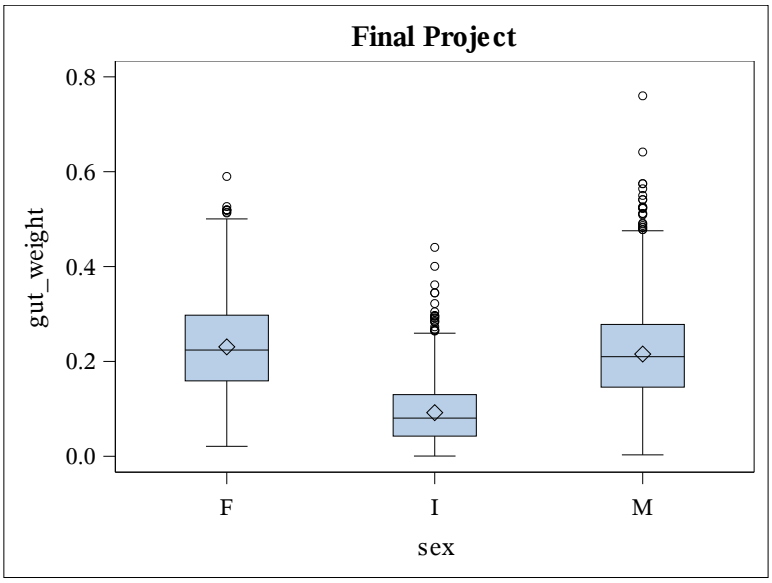
It can be noticed that infant has minimum weights compared to female or male.



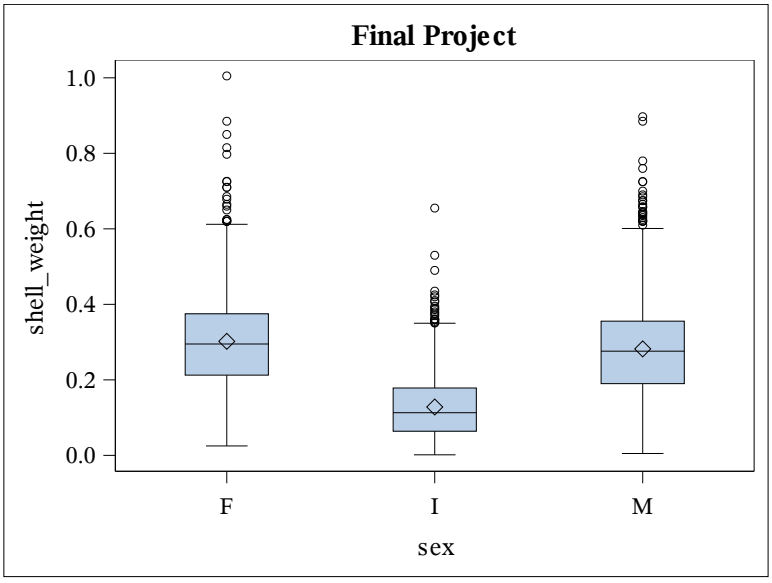
From the boxplot of whole_weight, infant has the lowest height while difference between female and male cannot be easily detected.



It can be noticed that infant has lower meat weight compared to female or male.



It can be noticed that infant has lower gut weight compared to female or male. Male abalones have some extreme large values.



Just like other weights, infant has lower shell weight compared to female or male.

In conclusion, there are noticeable differences in sizes (e.g. lengths, widths, heights) and weights for the three sexes (female, male and infant).

Question 2

In this part, the goal is to identify female abalone by using trained models. As indicated from the question, the wholesaler's supplier wants to identify females based on measurable quantities that can be quickly obtained without hurting the abalone. So length, diameter, height and whole weight

will be our possible predictors.

At the first stage, we do not consider separate infants and adults. Therefore, this is a classification problem with 'sex' as response variable. Let's do the Discriminant Analysis first. Here use the proportional priors.

Test of Homogeneity of Within Covariance Matrices

Chi-Square	DF	Pr > ChiSq
3199.760493	20	<.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=0.5 N=2084.5					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.66000551	240.78	8	8342	<.0001
Pillai's Trace	0.34083580	214.26	8	8344	<.0001
Hotelling-Lawley Trace	0.51386418	267.88	8	5956.2	<.0001
Roy's Greatest Root	0.51137148	533.36	4	4172	<.0001
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

Since the hypothesis test reveals that the variances in each group are not equal, we will use QDA rather than LDA. The subsequent statistics, e.g., Wilk's Lambda, and their p-values suggest that some amount of discrimination will be possible by some of the independent variables.

Number of Observations and Percent Classified into sex				
From sex	**F**	**I**	**M**	**Total**
F	233 17.83	322 24.64	752 57.54	1307 100.00
I	50 3.73	1163 86.66	129 9.61	1342 100.00
M	236 15.45	450 29.45	842 55.10	1528 100.00
Total	519 12.43	1935 46.33	1723 41.25	4177 100.00
Priors	0.3129	0.32128	0.36581	

Error Count Estimates for sex				
	F	**I**	**M**	**Total**
Rate	0.8217	0.1334	0.4490	0.4642
Priors	0.3129	0.3213	0.3658	

The cross-validation predictions match the female pretty bad with probability 0.8217. 57.54 percent of females classified into males. The total error rate is 0.4642.

Error Count Estimates for sex				
	F	**I**	**M**	**Total**
Rate	0.7399	0.1230	0.5249	0.4626
Priors	0.3333	0.3333	0.3333	

Use equal prior probability 0.333333 instead and the result looks like a little better with overall error rate 0.4626. However, the misclassification rate of female is still high with 0.7399 percentage.

Since we do not need to separate male and infant, so we can consider modeling adults only since the supplier believes they can keep infants and adults separate. As before, try Discriminant Analysis with proportional priors.

Chi-Square	DF	Pr > ChiSq
654.469173	10	<.0001

Multivariate Statistics and Exact F Statistics					
S=1 M=1 N=1414					
Statistic	**Value**	**F Value**	**Num DF**	**Den DF**	**Pr > F**
Wilks' Lambda	0.98303921	12.21	4	2830	<.0001
Pillai's Trace	0.01696079	12.21	4	2830	<.0001
Hotelling-Lawley Trace	0.01725342	12.21	4	2830	<.0001
Roy's Greatest Root	0.01725342	12.21	4	2830	<.0001

Number of Observations and Percent Classified into sex			
From sex	**F**	**M**	**Total**
F	235 17.98	1072 82.02	1307 100.00
M	239 15.64	1289 84.36	1528 100.00
Total	474 16.72	2361 83.28	2835 100.00
Priors	0.46102	0.53898	

Error Count Estimates for sex			
	F	**M**	**Total**
Rate	0.8202	0.1564	0.4624
Priors	0.4610	0.5390	

Since the hypothesis test reveals that the variances in each group are not equal, we will use QDA rather than LDA. The subsequent statistics, e.g., Wilk's Lambda, and their p-values suggest that some amount of discrimination will be possible by some of the independent variables.

The misclassification rate of female still very high with value 0.8202, over 80 percent are wrongly classified into male.

Consider to use stepwise discrimination to determine the best predictors for a discriminant analysis based on variables in the data set.

Stepwise Selection Summary								
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	diameter		0.0095	27.28	<.0001	0.99046221	<.0001
2	2	whole_weight		0.0053	15.21	<.0001	0.98517016	<.0001
3	3	height		0.0021	6.01	0.0143	0.98308304	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	diameter		0.00953779	<.0001
2	2	whole_weight		0.01482984	<.0001
3	3	height		0.01691696	<.0001

After the stepwise procedure, we only need to keep diameter, whole_weight and height to effectively classify abalones between female and male.

Error Count Estimates for sex			
	F	M	Total
Rate	0.8386	0.1374	0.4607
Priors	0.4610	0.5390	

We still use the QDA rather than LDA to perform the Discriminant Analysis. With one variable less, the error rate doesn't look better.

Let's try logistic model instead to check if that model performs better.

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.8282	0.4377	41.7558	<.0001
length	1	0.9332	2.0001	0.2177	0.6408
diameter	1	5.4159	2.3955	5.1114	0.0238
height	1	5.2361	2.1861	5.7369	0.0166
whole_weight	1	-1.0720	0.2414	19.7267	<.0001

As we have seen before, the parameter length is not significant with p value 0.6408. Let's use stepwise selection with default significance levels to choose the best set of explanatory variables for predicting the probability of female.

Summary of Stepwise Selection							
	Effect						
Step	**Entered**	**Removed**	**DF**	**Number In**	**Score Chi-Square**	**Wald Chi-Square**	**Pr > ChiSq**
1	**diameter**		1	1	27.0396		<.0001
2	**whole_weight**		1	2	15.7828		<.0001
3	**height**		1	3	6.3112		0.0120

Analysis of Maximum Likelihood Estimates					
Parameter	**DF**	**Estimate**	**Standard Error**	**Wald Chi-Square**	**Pr > ChiSq**
Intercept	1	-2.7431	0.3969	47.7665	<.0001
diameter	1	6.3505	1.3152	23.3147	<.0001
height	1	5.2019	2.1842	5.6720	0.0172
whole_weight	1	-1.0380	0.2297	20.4113	<.0001

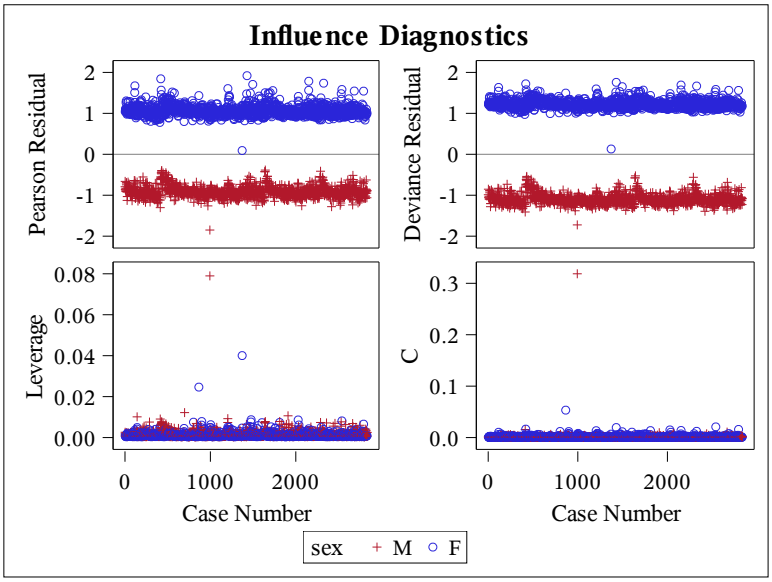
Odds Ratio Estimates			
Effect	**Point Estimate**	**95% Wald Confidence Limits**	
diameter	572.784	43.500	>999.999
height	181.617	2.512	>999.999
whole_weight	0.354	0.226	0.556

Diameter, whole_weight and height are kept based on their p values are all significant and indicating the coeffients are different from 0 at a .05 level.

We can also note that the parameters of diameter and height are positive, so increases in any of these two predictors will increase the log odds (and consequently increase the odds) of being female. On the other hand, decreasing whole_weight will lead to decrease the log odds of being female.

As for the odds ratios, the confidence intervals are pretty wide for diameter and height.
Specifically, an increase of 1 of diameter would correspond to an expected multiplicative increase of 572.784 in the odds of being female with a 95% confidence interval of (43.500, >999.999), a unit increase in height would correspond to an expected multiplicative increase of 181.617 with a 95% confidence interval of (2.512, >999.999), and an increase of whole_weight would correspond to an expected multiplicative increase of 0.354 with a 95% confidence interval of (0.226, 0.556).

In addition to the default results for the final model, diagnostic plots are included here.



With binary logistic regression, the females are underestimated by the model and the males overestimated.

Table of sex by _INTO_			
sex	_INTO_(Formatted Value of the Predicted Response)		
Frequency Percent Row Pct	F	M	Total
F	384 13.54 29.38	923 32.56 70.62	1307 46.10
M	350 12.35 22.91	1178 41.55 77.09	1528 53.90
Total	734 25.89	2101 74.11	2835 100.00

A frequency table is useful here to compare the observed sex values to the predicted groups and the _INTO_ values which represent the category the observation would have been placed into based on the model. Only 29.38 percent of females are correctly classified in to female. 70.62 percent of females are wrongly classified into male. Therefore, the model does not classify very well for the intended purpose.

Question 3

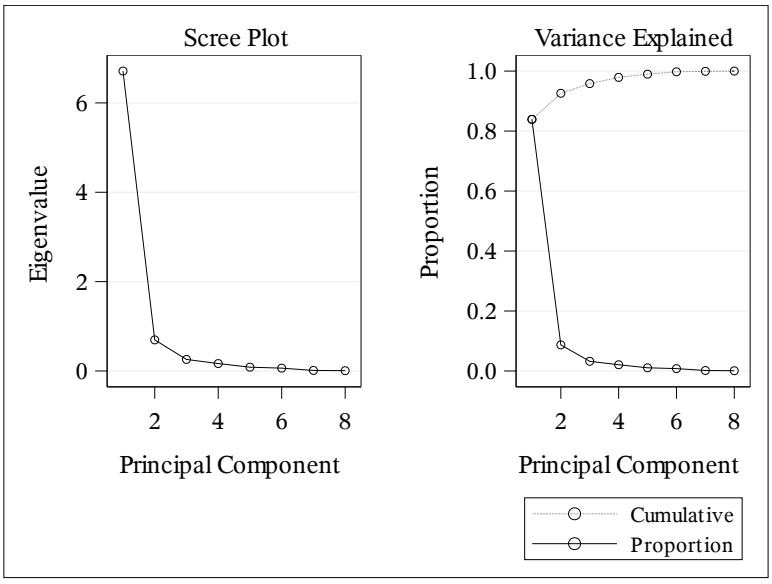
The wholesaler believes that the infant abalone are very different in size and weight characteristics from female and male abalone, but that female and male abalone are not very different in size and weight.

In this part we want to distinguish between the three sexes based on dimensions and weights. Lat's try PCA first. Perform a correlation-based principal components analysis.

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	6.71243915	6.01682618	0.8391	0.8391
2	0.69561297	0.43716985	0.0870	0.9260
3	0.25844312	0.09245328	0.0323	0.9583
4	0.16598984	0.08104018	0.0207	0.9791
5	0.08494966	0.02147690	0.0106	0.9897

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
6	0.06347277	0.05077860	0.0079	0.9976
7	0.01269416	0.00629584	0.0016	0.9992
8	0.00639832		0.0008	1.0000

Eigenvectors								
	Prin1	**Prin2**	**Prin3**	**Prin4**	**Prin5**	**Prin6**	**Prin7**	**Prin8**
length	0.372139	-.068283	0.031070	-.604054	0.011125	-.047497	0.698826	-.016349
diameter	0.373094	-.040048	0.041005	-.587595	-.057911	-.023375	-.712985	0.000219
height	0.340027	0.070463	0.899706	0.256777	0.056729	0.026691	0.008614	-.002688
whole_weight	0.378307	-.137346	-.206194	0.241849	-.015656	0.117255	-.008331	-.850264
meat_weight	0.362454	-.298840	-.208286	0.183246	0.398525	0.624893	-.009282	0.391101
gut_weight	0.368558	-.172979	-.197380	0.265221	0.309821	-.765844	-.027346	0.204179
shell_weight	0.370758	0.045400	-.161574	0.244192	-.830564	0.032832	0.047395	0.285624
rings	0.242713	0.921204	-.192144	0.043310	0.220026	0.068196	0.008421	0.023370



To retain 80% of the variation in the original variables, we would need to keep the first 1 principal component. The average eigenvalue is 1, so we would also choose 1 component based on the average eigenvalue criterion. The scree plot becomes fairly flat after the third component, so we would choose 3 based on this criterion. A case could also be made for one component based on

the scree plot given the large drop from the first to the second component and then much smaller drop from the second to the third. We choose to have 1 component.

Looking at the first component, all of the coefficients are positive. All parameters in size and weight are around 0.36 or 0.37, so this is basically a weighted average of the measurements. The measurements with largest coefficients have the most impact, but all have a positive relationship.

In this case, PCA will not help us to distinguish between the three sexes. Let's try stepwise discriminant analysis.

Stepwise Selection Summary								
Step	Number In	Entered	Removed	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda
1	1	diameter		0.3225	993.61	<.0001	0.67746216	<.0001
2	2	gut_weight		0.0180	38.30	<.0001	0.66525126	<.0001
3	3	meat_weight		0.0122	25.76	<.0001	0.65713555	<.0001
4	4	height		0.0071	14.94	<.0001	0.65246027	<.0001
5	5	length		0.0046	9.60	<.0001	0.64946914	<.0001
6	6	whole_weight		0.0022	4.56	0.0105	0.64805142	<.0001

Step	Number In	Entered	Removed	Average Squared Canonical Correlation	Pr > ASCC
1	1	diameter		0.16126892	<.0001
2	2	gut_weight		0.16740699	<.0001
3	3	meat_weight		0.17279130	<.0001
4	4	height		0.17517552	<.0001
5	5	length		0.17668413	<.0001
6	6	whole_weight		0.17740277	<.0001

The stepwise procedure keep the predictors of diameter, gut_weight, meat_weight, height, length and whole_weight. Only remove Shell weight.

Number of Observations and Percent Classified into sex				
From sex	**F**	**I**	**M**	**Total**
F	311 23.79	349 26.70	647 49.50	1307 100.00
I	45 3.35	1162 86.59	135 10.06	1342 100.00
M	284 18.59	486 31.81	758 49.61	1528 100.00
Total	640 15.32	1997 47.81	1540 36.87	4177 100.00
Priors	0.3129	0.32128	0.36581	

Error Count Estimates for sex				
	F	**I**	**M**	**Total**
Rate	0.7621	0.1341	0.5039	0.4659
Priors	0.3129	0.3213	0.3658	

Based on the Test of Homogeneity of Within Covariance Matrices, we use QDA in this case. 23.79 percent of females are correctly classified. 86.59 percent of infants are correctly classified which is a pretty high rate. 49.61 percent of the males are correctly classified, not that bad. As we can see, the error rate of infants is the lowest, and it's much better hard to well classify female or male. Among the three, females classification performs the worst.

Question 4

In this part, we want to understand how final meat weight is related to dimensions, whole weight, age(rings), and sex. Let's try general linear model first.

	length		diameter		height		whole_weight		rings	
	Mean	N	Mean	N	Mean	N	Mean	N	Mean	N
sex										
F	0.58	1307	0.45	1307	0.16	1307	1.05	1307	11.13	1307
I	0.43	1342	0.33	1342	0.11	1342	0.43	1342	7.89	1342
M	0.56	1528	0.44	1528	0.15	1528	0.99	1528	10.71	1528

From the table above, clearly, there is quite a bit of variation in the cell means. However, we only have one categorical predictor, so we can still the normal means(cell means) rather than least squares means.

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	196.6492645	28.0927521	12881.5	<.0001
Error	4169	9.0920274	0.0021809		
Corrected Total	4176	205.7412919			

R-Square	Coeff Var	Root MSE	meat_weight Mean
0.955808	12.99498	0.046700	0.359367

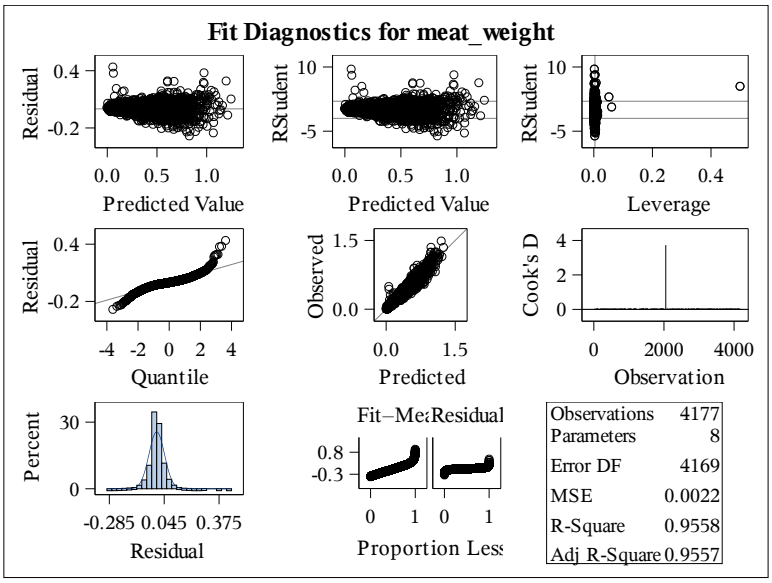
Source	DF	Type I SS	Mean Square	F Value	Pr > F
length	1	165.8786980	165.8786980	76061.0	<.0001
diameter	1	0.3947952	0.3947952	181.03	<.0001
height	1	0.5038488	0.5038488	231.03	<.0001
whole_weight	1	26.9728269	26.9728269	12367.9	<.0001
rings	1	2.8429121	2.8429121	1303.57	<.0001
sex	2	0.0561835	0.0280918	12.88	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
length	1	0.07026979	0.07026979	32.22	<.0001
diameter	1	0.00994627	0.00994627	4.56	0.0328
height	1	0.02381151	0.02381151	10.92	0.0010

Source	DF	Type III SS	Mean Square	F Value	Pr > F
whole_weight	1	26.64522375	26.64522375	12217.7	<.0001
rings	1	2.76567336	2.76567336	1268.15	<.0001
sex	2	0.05618354	0.02809177	12.88	<.0001

This model works great with a very small p value based on F statistic and R Square is 0.955808, which is pretty good. All of the predictors we considered perform great with small p value based on both of the Type I and Type III error.

As for the fit diagnostics plots. It looks like the residuals start to spread out as predicted value increases. Therefore the constant variance assumption may be violated. It's possible that a log-transformation or square root of meat weight would reduce the heteroscedasticity.



As for the leverage plot, there is one point that has large leverage. So we can consider to remove that specific data point. Also, we notice there is one point with a large cooks distance. All other points look great. The quantile plot is not too far from straight except some points at the two sides and the histogram is not too far from bell-shaped, so the normality assumption should be fine.

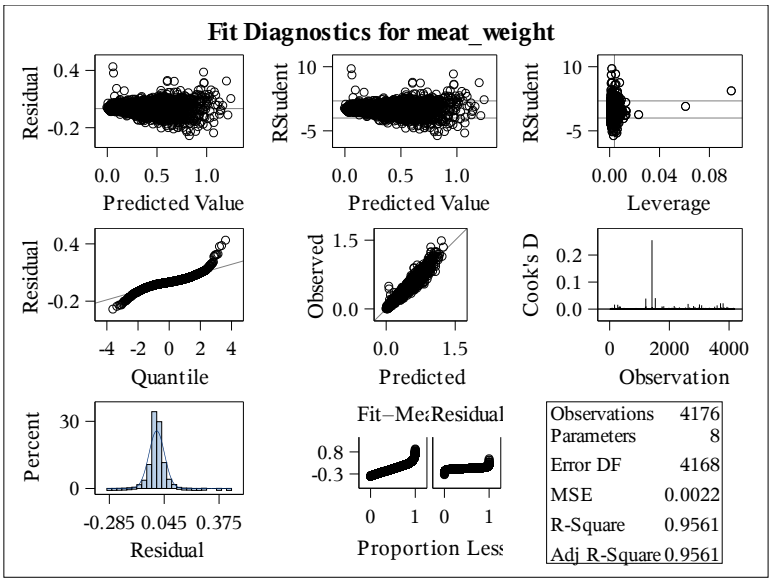
Obs	sex	length	diameter	height	whole_weight	meat_weight
2052	F	0.455	0.355	1.13	0.594	0.332

Obs	gut_weight	shell_weight	rings	cd
2052	0.116	0.1335	8	3.72993

We get to know that observation 2052 is a high influential point. Remove that and re-try the

model.

R-Square	Coeff Var	Root MSE	meat_weight Mean
0.956126	12.94955	0.046537	0.359374



R square is a little bit higher with 0.956126. But the predictor of diameter becomes not significant. So remove this predictor and re-fit the model.

Source	DF	Type I SS	Mean Square	F Value	Pr > F
length	1	165.8855430	165.8855430	76569.8	<.0001
height	1	1.0298573	1.0298573	475.36	<.0001
whole_weight	1	27.0535473	27.0535473	12487.4	<.0001
rings	1	2.6822165	2.6822165	1238.06	<.0001
sex	2	0.0574010	0.0287005	13.25	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
length	1	0.20652118	0.20652118	95.33	<.0001
height	1	0.09238998	0.09238998	42.65	<.0001
whole_weight	1	26.36897741	26.36897741	12171.5	<.0001
rings	1	2.60311148	2.60311148	1201.55	<.0001
sex	2	0.05740102	0.02870051	13.25	<.0001

Parameter		Estimate		Standard Error	t Value	Pr > |t|
Intercept		0.0349067744	B	0.00609904	5.72	<.0001
length		0.1710339050		0.01751763	9.76	<.0001
height		-.2971234933		0.04549881	-6.53	<.0001
whole_weight		0.4561999196		0.00413508	110.32	<.0001
rings		-.0098363017		0.00028377	-34.66	<.0001
sex	F	-.0090559849	B	0.00176064	-5.14	<.0001
sex	I	-.0041173767	B	0.00203966	-2.02	0.0436
sex	M	0.0000000000	B	.	.	.

Now all of the predictors look great with high significance. And the R square is 0.956100.
The expected meat weight will increase 0.17 if length increases 1 unit and 0.456 if whole_weight increases 1 unit. But the expected meat weight decreases -.297 for 1 unit increase in height and -.0098 for 1 unit increase in rings. As for sex, the expected meat weight will decrease -.0091 of being female than male. And decrease -.0041 of being infant than male for abalones.

Let's consider gamma model with log link. And check if that model works better.

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	4169	252.1568	0.0605
Scaled Deviance	4169	4218.5969	1.0119
Pearson Chi-Square	4169	326.2966	0.0783
Scaled Pearson X2	4169	5458.9587	1.3094
Log Likelihood		5307.3822	
Full Log Likelihood		5307.3822	
AIC (smaller is better)		-10596.7644	
AICC (smaller is better)		-10596.7212	
BIC (smaller is better)		-10539.7282	

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.6640	0.0350	-4.7327	-4.5954	17732.7	<.0001
length		1	4.5453	0.1881	4.1765	4.9140	583.71	<.0001
diameter		1	2.5323	0.2307	2.0802	2.9845	120.48	<.0001
height		1	1.5853	0.2099	1.1738	1.9967	57.03	<.0001
whole_weight		1	-0.0682	0.0227	-0.1127	-0.0237	9.04	0.0026
rings		1	-0.0159	0.0015	-0.0188	-0.0130	115.52	<.0001
sex	F	1	-0.0252	0.0092	-0.0433	-0.0071	7.43	0.0064
sex	I	1	-0.0318	0.0107	-0.0527	-0.0109	8.85	0.0029
sex	M	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	16.7301	0.3625	16.0345	17.4558		

LR Statistics For Type 1 Analysis				
Source	2*LogLikelihood	DF	Chi-Square	Pr > ChiSq
Intercept	1188.1618			
length	10331.0268	1	9142.86	<.0001
diameter	10449.9310	1	118.90	<.0001
height	10487.2400	1	37.31	<.0001
whole_weight	10494.0898	1	6.85	0.0089
rings	10602.8060	1	108.72	<.0001
sex	10614.7644	2	11.96	0.0025

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
length	1	516.79	<.0001
diameter	1	122.13	<.0001
height	1	69.98	<.0001
whole_weight	1	8.99	0.0027

LR Statistics For Type 3 Analysis			
Source	**DF**	**Chi-Square**	**Pr > ChiSq**
rings	1	111.76	<.0001
sex	2	11.96	0.0025

As we mentioned, this is a gamma model with log link. The Scaled Deviance is roughly 1 so no need to consider overdispersion problem. The type 1 and type 3 analyses tell us that the length, diameter, height, whole_weight, rings and sex all have a significant relationship with meat weight. Since all six predictors are significant, this is the final model.

Specifically, we would expect meat weight to be multiplied by exp(4.5453) if length increase 1 unit. And meat weight to be multiplied by exp(2.5323) if diameter increase 1 unit. And meat weight to be multiplied by exp(1.5853) if height increase 1 unit. But the expected meat weight will increase exp(-0.0682) which is less than 1 if whole_weight increases 1 unit. And increase exp(-0.0159) if rings increases 1 unit. What's more, we would expect meat weight to increase exp(-0.0252) of being female than male. And exp(-0.0318) of being female compared to male.

However, we will still use the general linear model since it has lower AIC and BIC.

Question 5

For this part, we want to group abalone based on size and weight characteristics and check whether those groups are consistent with the sexes in some way. Also, we can consider using subsets or transformations of the possible predictors to improve grouping if necessary.

To do cluster analysis, let's use univariate analysis to identify extreme observations first.

Variable: length

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.075	237	0.775	2091
0.110	239	0.780	1210
0.130	2115	0.780	3716
0.130	238	0.800	2335
0.135	1987	0.815	1429

Variable: diameter

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.055	237	0.625	1208
0.090	239	0.630	1210
0.095	2115	0.630	1764
0.100	720	0.630	2335
0.100	238	0.650	1429

Variable: height

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.000	3997	0.250	1429
0.000	1258	0.250	1764
0.010	237	0.250	2180
0.015	2170	0.515	1418
0.015	1175	1.130	2052

Variable: whole_weight

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.0020	237	2.5500	166
0.0080	239	2.5550	1052
0.0105	2115	2.6570	1210
0.0130	238	2.7795	1764
0.0140	1430	2.8255	892

Variable: meat_weight

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.0010	237	1.2455	3714
0.0025	239	1.2530	2812
0.0045	720	1.3485	1764
0.0045	238	1.3510	1529
0.0050	3900	1.4880	1210

Variable: gut_weight

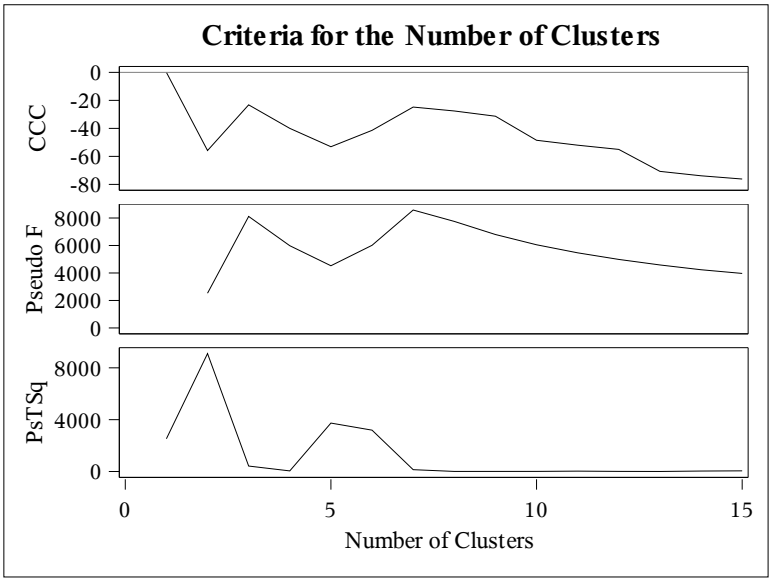
Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.0005	3523	0.5745	3716
0.0005	237	0.5750	3428
0.0020	239	0.5900	2335
0.0025	1430	0.6415	1763
0.0025	695	0.7600	1764

Variable: shell_weight

Extreme Observations			
Lowest		**Highest**	
Value	**Obs**	**Value**	**Obs**
0.0015	237	0.850	167
0.0030	239	0.885	2109
0.0035	2115	0.885	2162
0.0040	1430	0.897	892
0.0040	238	1.005	164

There are 2052 observations have extreme value 1.130 in height. So these data definitly should be moved. As for whole_weight, there are 237 extreme observations with value 0.0020 which should be also removed. And we also got 164 extreme observations with value 1.005 for shell_weight. After removing the extreme observations, we can start to do clustering analysis. Since all three

sizes measurements are labeled in mm, but in grams for four weight measurements. So standardize these measurements would be a good idea. Let's use complete linkage and obtain ccc values for number of clusters here.



From the CCC, and pseudo F and t-squared plots, 3 looks like a good choice for the number of clusters based on each criterion. We see peaks at clusters for the CCC and pseudo F statistics, and we see a pretty big jump from 2 clusters to 3 clusters for the pseudo t-squared statistic.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
length	1757	0.4115253	0.0874267	0.1100000	0.5650000
diameter	1757	0.3144991	0.0711703	0.0900000	0.4650000
height	1757	0.1050996	0.0266453	0	0.1850000
whole_weight	1757	0.3706881	0.1849616	0.0080000	0.7105000
meat_weight	1757	0.1597442	0.0836448	0.0025000	0.4950000
gut_weight	1757	0.0809727	0.0430446	0.000500000	0.2270000
shell_weight	1757	0.1124308	0.0573982	0.0030000	0.3505000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
length	2017	0.5893133	0.0457647	0.4050000	0.7200000
diameter	2017	0.4623500	0.0382524	0.3100000	0.5850000
height	2017	0.1583019	0.0205004	0.0150000	0.2500000
whole_weight	2017	1.0392362	0.2180424	0.6355000	1.5270000
meat_weight	2017	0.4498753	0.1116563	0.1650000	0.8000000
gut_weight	2017	0.2271807	0.0571881	0.0950000	0.4405000
shell_weight	2017	0.2990982	0.0703476	0.0995000	0.5800000

CLUSTER=3

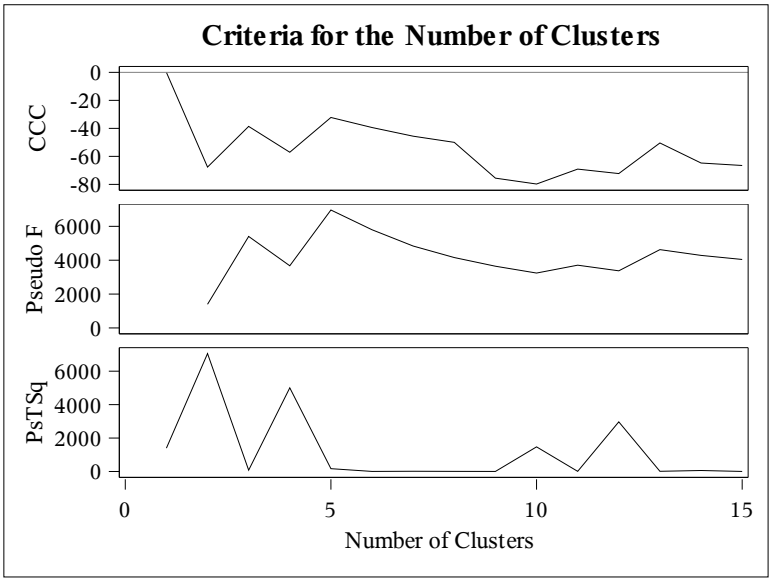
Variable	N	Mean	Std Dev	Minimum	Maximum
length	400	0.6894125	0.0374931	0.5600000	0.8150000
diameter	400	0.5440375	0.0321223	0.4200000	0.6500000
height	400	0.1936375	0.0247561	0.1300000	0.5150000
whole_weight	400	1.7787012	0.2596017	1.3480000	2.8255000
meat_weight	400	0.7800650	0.1643357	0.4125000	1.4880000
gut_weight	400	0.3833313	0.0706116	0.1920000	0.7600000
shell_weight	400	0.4890863	0.1008094	0.1780000	0.8970000

Comparing the clusters, cluster1 has the smallest mean value of length, diameter and height. Cluster3 has all of these three size measurements with the largest value. The same conclusions can be made for weights, that cluster3 has the largest mean values and then cluster2 and cluster1 the least.

Table of CLUSTER by sex				
CLUSTER	**sex**			
Frequency **Percent** **Col Pct**	**F**	**I**	**M**	**Total**
1	283 6.78 21.69	1087 26.04 81.06	387 9.27 25.33	1757 42.09
2	826 19.79 63.30	249 5.97 18.57	942 22.57 61.65	2017 48.32
3	196 4.70 15.02	5 0.12 0.37	199 4.77 13.02	400 9.58
Total	1305 31.26	1341 32.13	1528 36.61	4174 100.00

We can see most of the females(63.30 percent) are in cluster1. 81.06 percent of infants are in cluster1, 18.57 percent in cluster2 and nearly none of them are in cluster3. As for males, 25.33 percent are in cluster1, 61.65 percent in cluster2 and 13.02 percent in cluster3. So cluster3 has similar number of females and males.

From the previous analysis. All the three measurements in size do not perform that obvious difference among clusters. So let's take the exponential transformation of these three predictors.



From the CCC, and pseudo F and t-squared plots, 5 looks like a good choice for the number of

clusters based on each criterion.

CLUSTER=1

Variable	N	Mean	Std Dev	Minimum	Maximum
exp_length	1387	1.4758980	0.1149704	1.1162781	1.6736385
exp_diameter	1387	1.3444856	0.0862736	1.0941743	1.5920142
exp_height	1387	1.1035408	0.0273271	1.0000000	1.1972174
whole_weight	1387	0.3059643	0.1517764	0.0080000	0.6485000
meat_weight	1387	0.1314993	0.0690757	0.0025000	0.4950000
gut_weight	1387	0.0664355	0.0349029	0.000500000	0.1805000
shell_weight	1387	0.0937441	0.0481953	0.0030000	0.3505000

CLUSTER=2

Variable	N	Mean	Std Dev	Minimum	Maximum
exp_length	1393	1.7212772	0.0613851	1.4993025	1.8870221
exp_diameter	1393	1.5276723	0.0467287	1.3634251	1.6904588
exp_height	1393	1.1543234	0.0206017	1.0887171	1.2649088
whole_weight	1393	0.7930147	0.1406375	0.4750000	1.1185000
meat_weight	1393	0.3408683	0.0756202	0.1580000	0.5950000
gut_weight	1393	0.1724311	0.0373289	0.0850000	0.2840000
shell_weight	1393	0.2335761	0.0509354	0.0995000	0.3900000

CLUSTER=3

Variable	N	Mean	Std Dev	Minimum	Maximum
exp_length	1190	1.8814199	0.0621610	1.6652912	2.0959355
exp_diameter	1190	1.6442843	0.0476130	1.4405140	1.8039884
exp_height	1190	1.1878581	0.0231373	1.0151131	1.2840254
whole_weight	1190	1.2853487	0.1823332	0.9700000	1.8060000
meat_weight	1190	0.5581718	0.0986234	0.2885000	0.9600000
gut_weight	1190	0.2827983	0.0540838	0.1255000	0.4875000
shell_weight	1190	0.3640109	0.0737342	0.1550000	0.8150000

CLUSTER=4

Variable	N	Mean	Std Dev	Minimum	Maximum
exp_length	193	2.0298649	0.0658698	1.7506725	2.2591757
exp_diameter	193	1.7460567	0.0484022	1.5219616	1.9155408
exp_height	193	1.2213303	0.0387795	1.1676580	1.6736385
whole_weight	193	1.9266813	0.1915226	1.6465000	2.4990000
meat_weight	193	0.8582772	0.1473132	0.5415000	1.3510000
gut_weight	193	0.4090104	0.0624738	0.1965000	0.5640000
shell_weight	193	0.5220337	0.0930352	0.3410000	0.8850000

CLUSTER=5

Variable	N	Mean	Std Dev	Minimum	Maximum
exp_length	11	2.1349713	0.0522320	2.0647311	2.2255409
exp_diameter	11	1.8415659	0.0345860	1.7682671	1.8776106
exp_height	11	1.2383730	0.0234584	1.2092496	1.2840254
whole_weight	11	2.5875000	0.1154987	2.4925000	2.8255000
meat_weight	11	1.1821818	0.1442306	0.9330000	1.4880000
gut_weight	11	0.5466818	0.0969065	0.4190000	0.7600000
shell_weight	11	0.6487727	0.0991061	0.5235000	0.8970000

Exp_length and exp_diameter perform much better and become easier to distinguish, but the differences in exp_height are still small among 5 clusters. Just like the previous clustering version, all of our predictors tend to have larger mean value from cluster1 to cluster5 one by one.

Table of CLUSTER by sex				
CLUSTER	**sex**			
Frequency **Percent** **Col Pct**	**F**	**I**	**M**	**Total**
1	179 4.29 13.72	930 22.28 69.35	278 6.66 18.19	1387 33.23
2	464 11.12 35.56	367 8.79 27.37	562 13.46 36.78	1393 33.37
3	567 13.58 43.45	42 1.01 3.13	581 13.92 38.02	1190 28.51
4	90 2.16 6.90	2 0.05 0.15	101 2.42 6.61	193 4.62
5	5 0.12 0.38	0 0.00 0.00	6 0.14 0.39	11 0.26
Total	1305 31.26	1341 32.13	1528 36.61	4174 100.00

35.56 and 43.45 percent of females are grouped into cluster2 and cluster3 respectively. 69.35 percent of infants are grouped into cluster1. 36.78 and 38.02 percent of males are grouped into cluster2 and cluster3. Also remember they have larger values of measurements if they are in a group with larger cluster number. However, it seems like using the exponential transformation for 3 predictors in size does not improve the clustering.