Question 2

2a).

non_dreaming	dreaming	MLE	gestation	log_body	log_brain
Min.: 2.100	Min. :0.500	Min.: 2.00	Min.: 12.0	Min.:-5.2983	Min.:-1.966
1st Qu.: 6.300	1st Qu.:0.900	1st Qu.: 4.85	1st Qu.: 33.0	1st Qu.:-1.4412	1st Qu.: 1.104
Median:9.100	Median :1.500	Median : 9.80	Median : 68.0	Median:0.5306	Median : 2.434
Mean:8.985	Mean :1.951	Mean :15.25	Mean :116.4	Mean: 0.5308	Mean : 2.515
3rd Qu.:11.000	3rd Qu.:2.500	3rd Qu.:24.00	3rd Qu.:175.0	3rd Qu.:2.1098	3rd Qu.: 4.745
Max.:17.900	Max. :6.600	Max. :50.00	Max. :392.0	Max.: 6.2557	Max. : 6.485

The above table gives the statistical summary of the numeric columns in the Animal dataset. It shows the minimum, quartile, median, average and Maximum estimates of the observations found in each numeric variable column.

	non_dreaming	dreaming	MLE	gestation	log_body	log_brain
non_dreaming	1.0000	0.5491	-0.3235	-0.5759	-0.6187	-0.5883
dreaming	0.5491	1.0000	-0.3467	-0.5710	-0.2902	-0.4097
MLE	-0.3235	-0.3467	1.0000	0.6885	0.6887	0.7449
gestation	-0.5759	-0.5710	0.6885	1.0000	0.7286	0.7617
log_body	-0.6187	-0.2902	0.6887	0.7286	1.0000	0.9477
log_brain	-0.5883	-0.4097	0.7449	0.7617	0.9477	1.0000

The correlation table above shows how correlated the numeric values are to each other, a more graphic representation of the correlation is represented on the correlation scatterplot.

The Histogram shows the distribution of each variable and also helps determine the presence of skewness in the distribution of any variable.

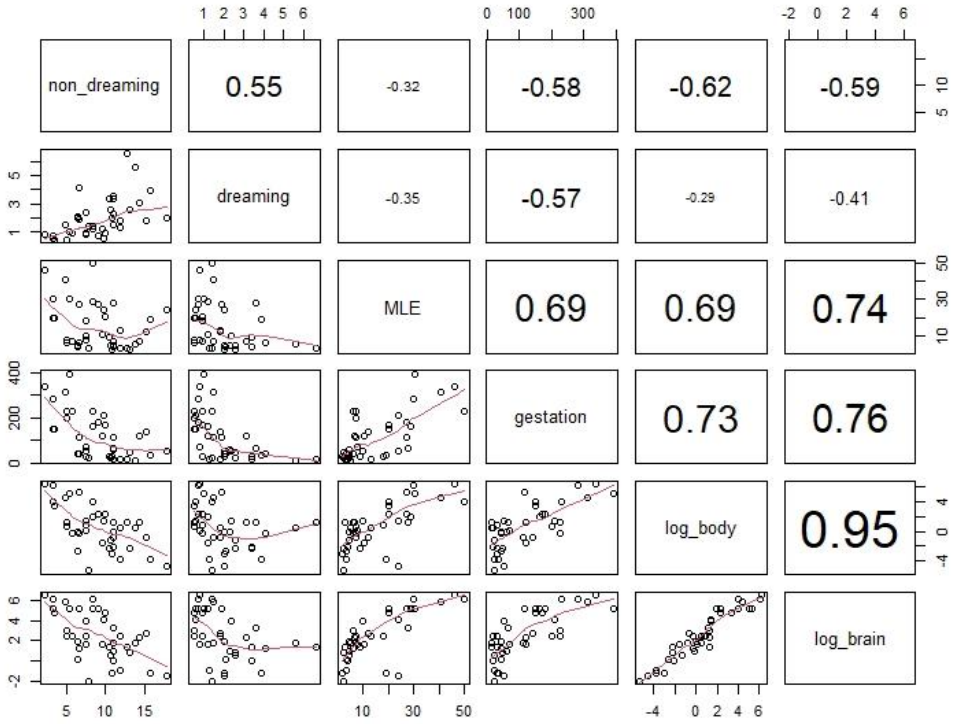


Figure 1: Scatter Correlation plot

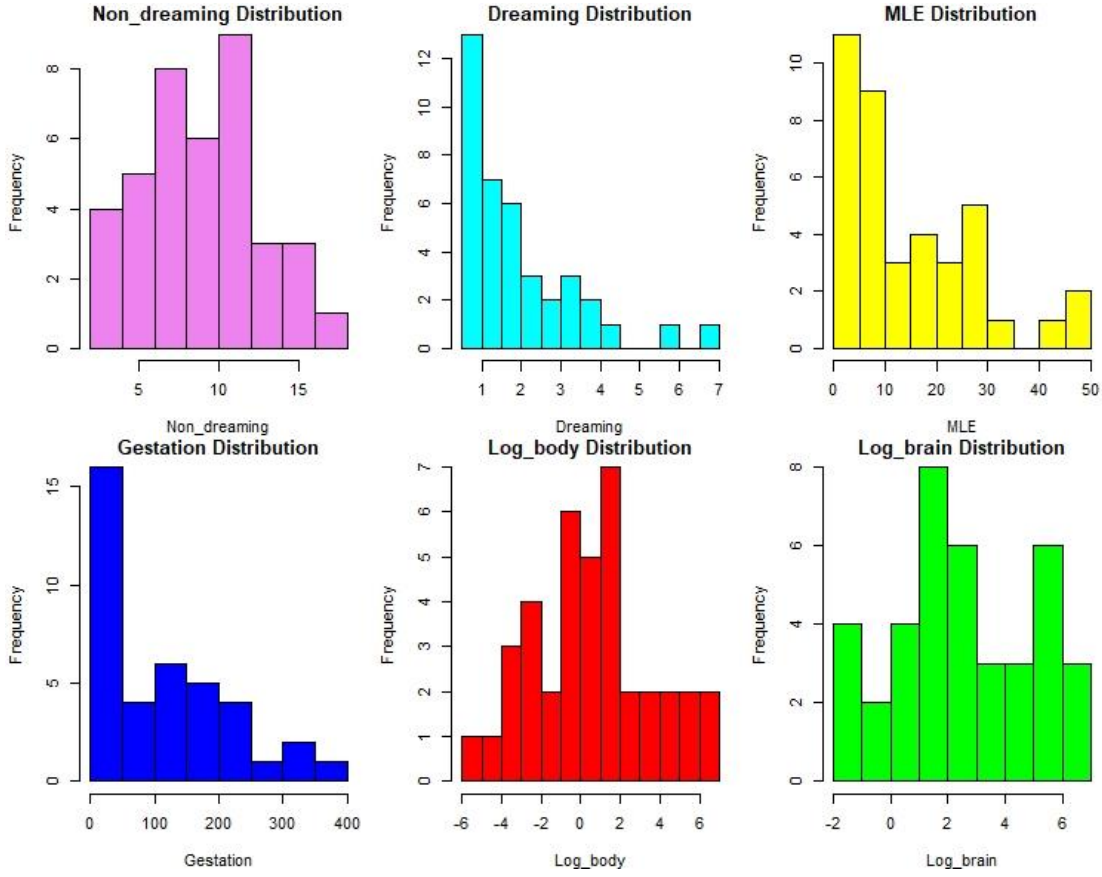


Figure 2: Histogram distribution

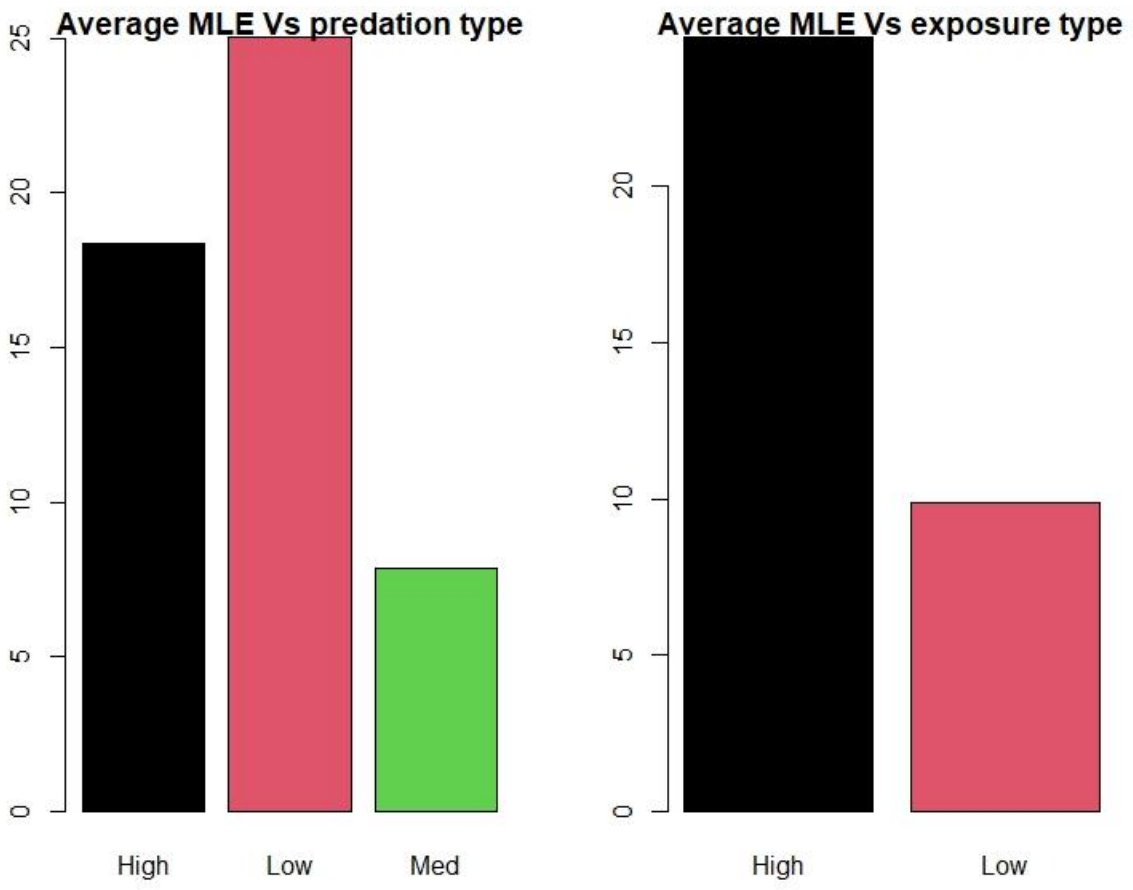


Figure 3: Bar chart - Mean MLE vs predation and exposure

Average MLE of predation groups

High	Low	Med
18.37500	25.04286	7.83125

Average MLE of exposure groups

High	Low
24.800	9.896

The bar chart above is a representation of the values presented in the table above which shows the average lifespan of each predation and exposure group respectively. It will be observed that the animal species with a low risk of being preyed upon have a longer lifespan and species with the lowest mean lifespan have a medium risk of being upon.

Likely, the exposure plot points out that species with high protection while asleep have longer lifespan compared to the other group.

It can be assumed that animal species with both low predation level and high exposure level tend to live longer than other animal species.

b.)

Call:				
lm(formula = MLE ~ non_dreaming + dreaming + gestation + predation +				
exposure + log_body + log_brain, data = data_dropna)				
Residuals:				
Min	1Q	Median	3Q	Max
-14.6468	-2.9379	-0.3064	3.5951	12.9423
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.23892	5.91229	-0.040	0.9680
non_dreaming	0.85506	0.41463	2.062	0.0479 *
dreaming	-0.34825	1.27599	-0.273	0.7868
gestation	0.03874	0.01824	2.124	0.0420 *
predationLow	10.09344	3.81653	2.645	0.0129 *
predationMed	-5.01179	3.37081	-1.487	0.1475
exposureLow	-2.30514	4.26206	-0.541	0.5926
log_body	0.83462	1.42592	0.585	0.5627
log_brain	2.08846	1.69709	1.231	0.2280

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 6.337 on 30 degrees of freedom				
Multiple R-squared: 0.8075, Adjusted R-squared: 0.7562				
F-statistic: 15.73 on 8 and 30 DF, p-value: 8.283e-09				

The linear regression equation for the fitted model is:

E(MLE) = -0.23892 + 0.85506(non_dreaming) - 0.34825(dreaming) + 0.03874(gestation) + 10.09344(predationLow) - 5.01179(predationMed) - 2.30514(exposureLow) + 0.83462(log_body) + 2.08846(log_brain)

With the above model, we are able to predict the MLE of various animal species with respect to their relative risk which is either high, low or med; and also, with respect to their protection while asleep. The model was regressed with predationHigh and exposureHigh as the base conditions for prediction.

c.)

Yes, there is multicollinearity. The VIF (variance inflation factor) is a tool to detect multicollinearity in a regressed model. Taking threshold VIF value to be "5" implies that all VIF values above 5 signify multicollinearity with another variable(s)

non_dreaming	dreaming	gestation	predationLow	predationMed	exposureLow	log_body	log_brain
2.356343	3.055576	3.227187	2.083500	2.670085	4.059917	16.339404	14.593529

The above table is the result of the VIF output and it is observed that log_brain and log_body are highly correlated; therefore, only one of the variables is required. We take out one or both of the variable to evaluate the important variable by running the regression each time a variable is removed.

	Model	**R2**	**Adj_R2**
1	original(mod)	0.807527969020935	0.756202094093185
2	without both(mod2)	0.73189957113491	0.681630740722705
3	with log_body only(mod3)	0.797811918201462	0.752156544892114
4	with log_brain only(mod4)	0.805329941508945	0.761372186365803

The table above shows the R2 and Adjusted R2 of the different models, it is observed that the initial model has a R2 and Adj R2 of 0.806 and 0.756 respectively, the model closest in R2 and adj R2 value is that consisting of all variables in the dataset except the log_body variable. It is seen that the Adj R2 increased on removal of the log_body variable from 0.756 to 0.761. Therefore, the new output from the recent regression will be:

Call:				
lm(formula = MLE ~ non_dreaming + dreaming + gestation + predation +				
exposure + log_brain, data = data_dropna)				
Residuals:				
Min	1Q	Median	3Q	Max
-14.8349	-2.9734	-0.0883	3.7072	13.1632
Coefficients:				
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1.80928	5.21236	-0.347	0.73085
non_dreaming	0.75068	0.37034	2.027	0.05133

dreaming	0.09158	1.02026	0.090	0.92905
gestation	0.04143	0.01746	2.373	0.02401 *
predationLow	9.68460	3.71206	2.609	0.01385 *
predationMed	-5.10927	3.33080	-1.534	0.13519
exposureLow	-2.70056	4.16332	-0.649	0.52134
log_brain	2.94153	0.86021	3.420	0.00178 **

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
Residual standard error: 6.269 on 31 degrees of freedom				
Multiple R-squared: 0.8053, Adjusted R-squared: 0.7614				
F-statistic: 18.32 on 7 and 31 DF, p-value: 2.161e-09				

The new regression model will be:

E(MLE) = -1.80928 + 0.75068 (non_dreaming) + 0.09158(dreaming) + 0.04143(gestation) + 9.68460(predationLow) - 5.10927(predationMed) - 2.70056 (exposureLow) + 2.94153 (log_brain)

d.) i.)

Provided other predictors are held constant, when the predation of an animal specie is high, there will be no change to the species life span therefore, it remains as it should be, but when the predation level is medium, the MLE is reduced by a value of approximately 5.1. Lastly, when the predation level is low, the Maximum lifespan of the specie increases by a unit of approximately 9.68.

ii.)

The output summary of the model indicates that the gestation variable is a significant predictor of MLE with its p-value in consideration. The MLE will increase by 0.041 for every new day (unit increase) in the gestation time provided other predictor variable in the dataset remains constant.

Question 3

a.)

'data.frame':	56 obs. of 10 variables:
$ survive:	Factor w/ 2 levels "0","1": 2 1 1 2 2 1 2 1 2 1 ...
$ length :	int 154 165 160 160 155 161 154 160 156 163 ...
$ alar :	int 41 40 45 50 43 49 45 46 47 50 ...
$ weight :	num 4.5 6.5 6.1 6.9 6.9 5.6 4.3 5.9 4.1 5.5 ...
$ lbh :	num 31 31 30 30.8 30.6 30.3 31.7 30.3 31.5 30.5 ...
$ lhum :	num 0.687 0.738 0.736 0.736 0.733 0.743 0.741 0.738 0.715 0.75 ...
$ lfem :	num 0.668 0.704 0.709 0.709 0.704 0.718 0.688 0.709 0.706 0.731 ...
$ ltibio :	num 1 1.09 1.11 1.18 1.15 ...
$ wskull :	num 0.587 0.606 0.611 0.6 0.6 0.6 0.584 0.607 0.575 0.603 ...
$ lkeel :	num 0.83 0.847 0.84 0.841 0.846 0.808 0.839 0.869 0.801 0.888 ...

The data given consists a total 56 observations and 10 different variables of which seven (7) of them are numeric columns (real numbers), 2 integer columns and a factor column which is the response variable. The response variable has only 2 levels with “1” indicating bird survived the drought or “0” meaning the birds did not survive.

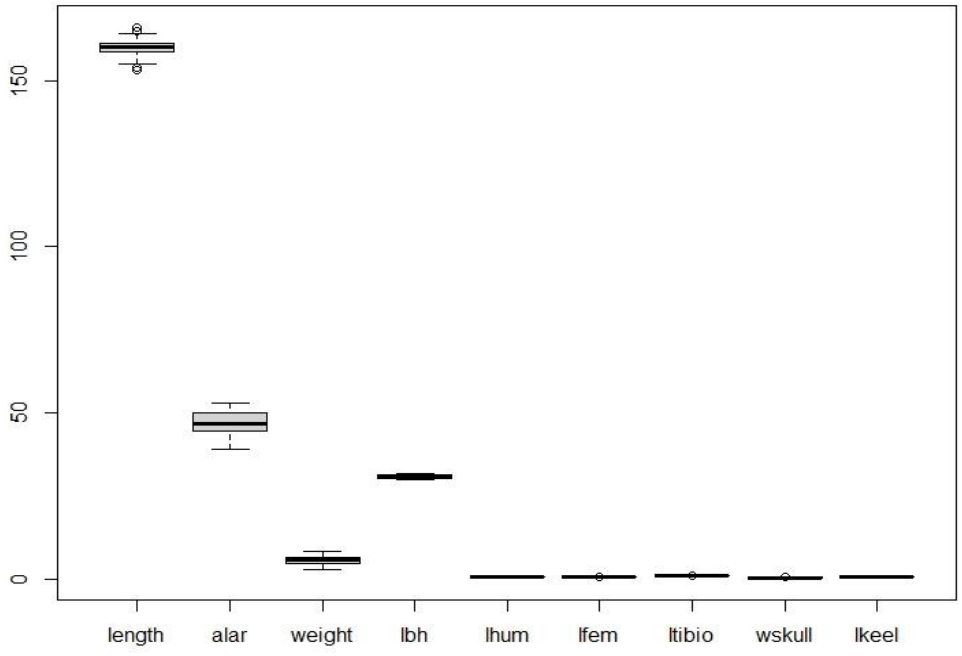


Figure 4: Boxplot to identify outliers

The boxplot in the above figure shows an insignificant total of outliers are present in the data which is visually observed in the body length distribution. The figure below shows the distribution of all the numeric variables in the dataset, and it can be seen that they some are fairly normally distributed, and some distributions are bimodal.

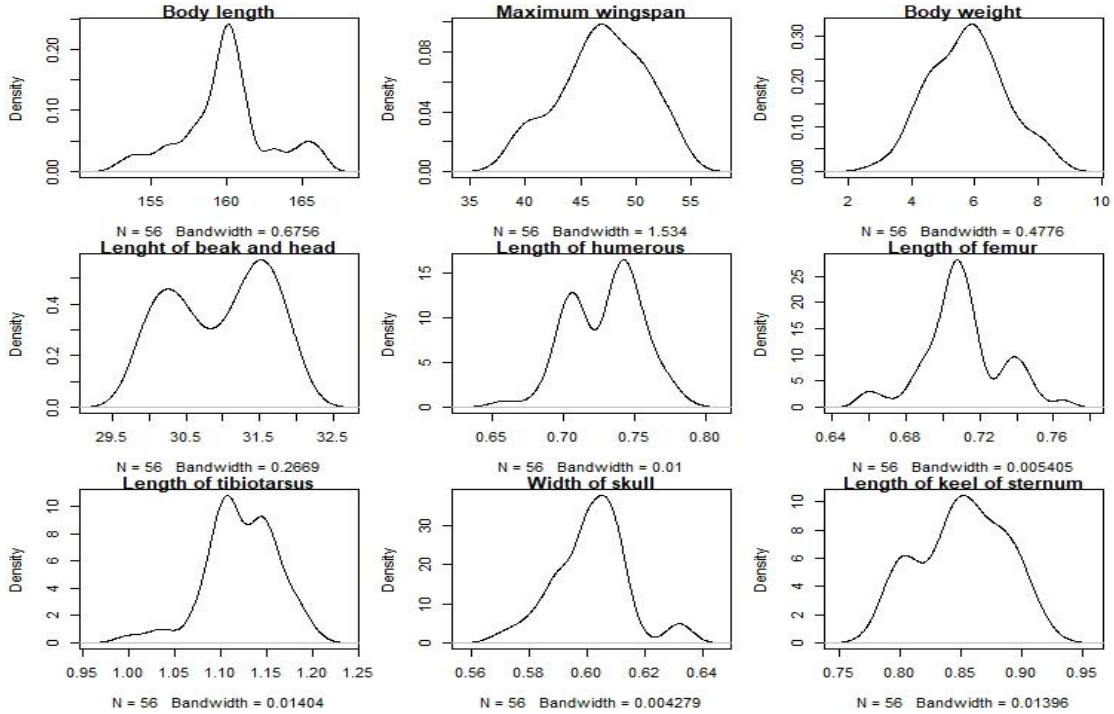


Figure 5: Variable Distribution

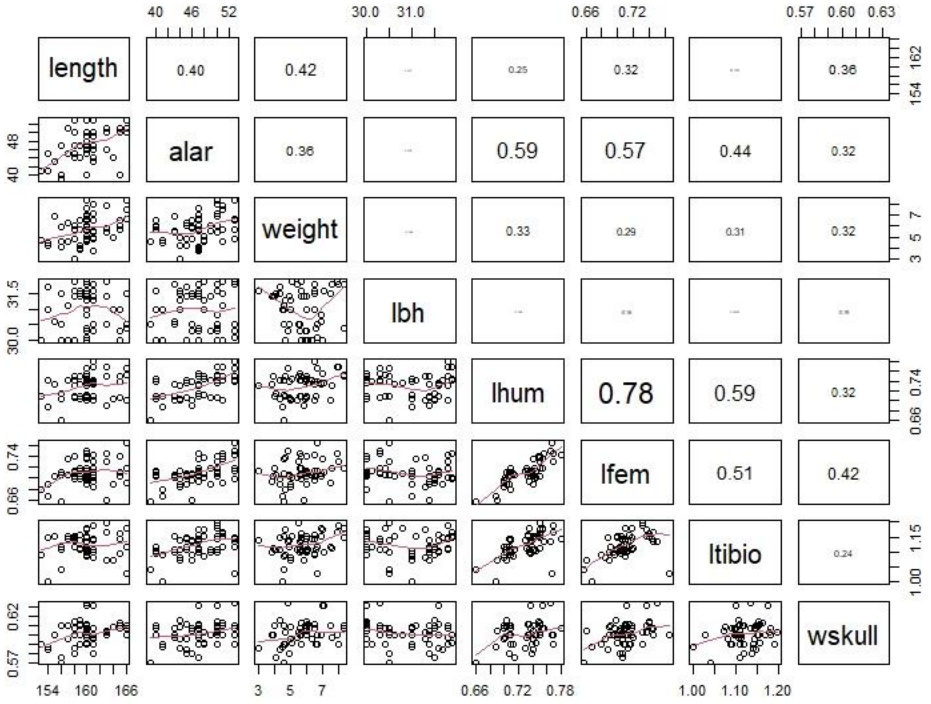


Figure 6: Correlation Plot

The correlation plot below shows how all numeric variables correlate with each other, this highest correlation is the Length of humerous and Length of femur of the bird. Also, it shows the scatter plot of each variable found in the dataset.

c.)

Call:				
glm(formula = survive ~ ., family = "binomial", data = train)				
Deviance Residuals:				
Min	1Q	Median	3Q	Max
-1.7587	-0.2355	0.0112	0.2795	2.0632
Coefficients:				
	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	126.0482	88.0138	1.432	0.1521
length	-1.8395	0.8642	-2.129	0.0333 *
alar	0.2414	0.2961	0.815	0.4150
weight	-1.3112	0.9863	-1.329	0.1837
lbh	0.8058	0.8681	0.928	0.3533
lhum	-72.9649	67.6029	-1.079	0.2804
lfem	366.8060	159.4883	2.300	0.0215 *
ltibio	-37.1304	39.5571	-0.939	0.3479
wskull	-70.8944	78.3809	-0.904	0.3657
lkeel	21.8904	21.8511	1.002	0.3164

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1				
(Dispersion parameter for binomial family taken to be 1)				
Null deviance: 50.920 on 38 degrees of freedom				
Residual deviance: 20.617 on 29 degrees of freedom				
AIC: 40.617				
Number of Fisher Scoring iterations: 8				

The above table shows the output result of performing logistic regression on the model, it shows the weight (coefficients) of each variable with respect to predicting the response variable (survive).

Below is the logistic regression equation that can be used in deriving the expected value of survive

E(survive) = 126.0482 - 1.8395(length) + 0.2414(alar) - 1.3112(weight) + 0.8058(lbh) - 72.9649(lhum) + 366.8060(lfem) - 37.1304(ltibio) - 70.8944(wskull) + 21.8904(lkeel)

	Dataset	**Accuracy**
Accuracy	Train Accuracy	0.871794871794872
Accuracy.1	Test Accuracy	0.705882352941177

The above table shows the accuracy of predicting both the training and test labels of the data using the logistic model. It is seen that the training data gives a higher accuracy when compared to the test data, this can be assumed to be an adequate result as the model is trained using the training data and therefore will easily predict the training data labels, also, the training data contains significantly a larger amount of data which is a factor that helps improve the overall accuracy of its predictions.

d.)

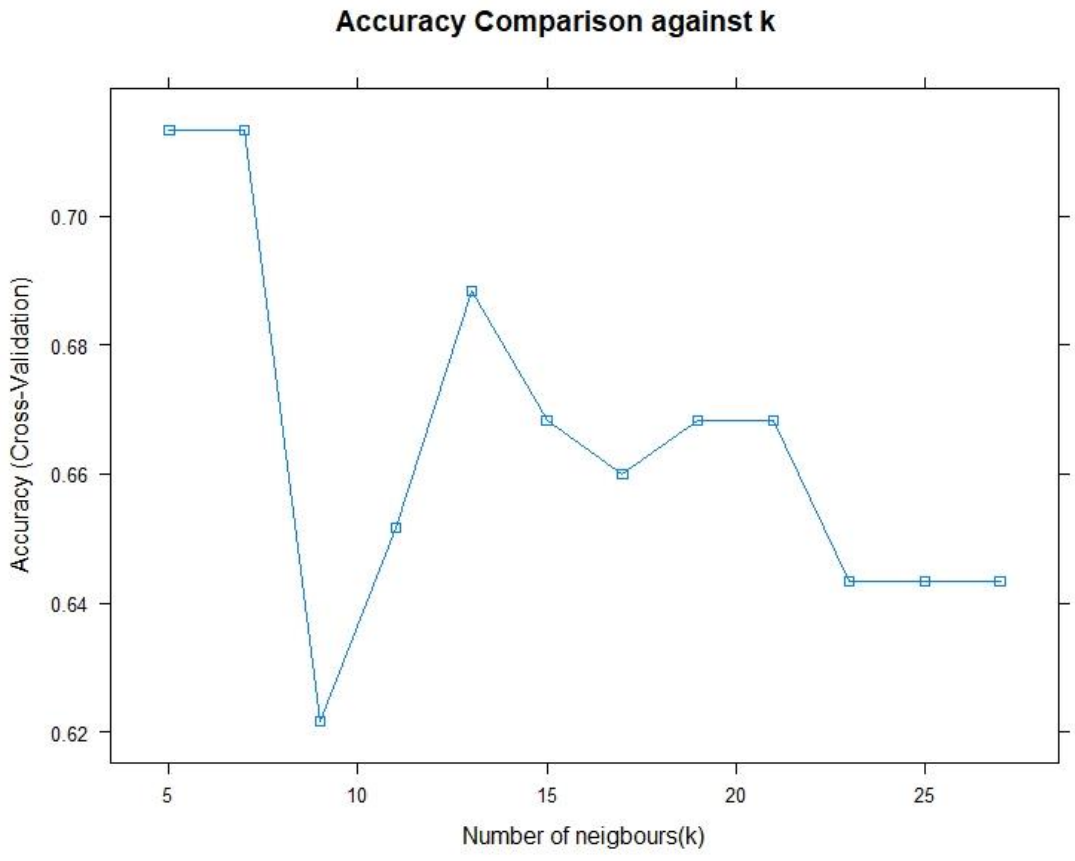
Using KNN model to predict Survival

Firstly, the entire "wren" data is split into training and test data which will then be used to train the model. The model is trained using the **"train()"** function from the r-packages, it uses the cross-validation method to control the trained data and the data are scaled from the centre. This training method loops through a range of values for k and determines the best value for k which is then used to create the model.

k-Nearest Neighbors
39 samples
9 predictor
2 classes: '0', '1'
Pre-processing: centered (9), scaled (9)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 35, 35, 36, 34, 35, 36, ...
Resampling results across tuning parameters:
k Accuracy Kappa
5 0.7133333 0.27121212
7 0.7133333 0.25454545

9 0.6216667 0.10454545
11 0.6516667 0.15000000
13 0.6883333 0.15454545
15 0.6683333 0.06666667
17 0.6600000 0.05000000
19 0.6683333 0.05000000
21 0.6683333 0.05000000
23 0.6433333 0.00000000
25 0.6433333 0.00000000
27 0.6433333 0.00000000
Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 7.

The above table shows the result of the trained data, and it will be observed that the k value which gives the highest accuracy is "7".



The plot above shows the accuracy of the iterated model created against a range of k values. Although, the k values of 1 and 2 gives a higher accuracy for the model, but a small value of k indicates that noise will an increased effect on the result. A rule of thumb states that the whole number approximate of the root of the total amount of observations should be used as the k value provided a k value is not specified. And for this data, the approximate root is 6, therefore, a k value of 7 could be acceptable for prediction.

	Dataset	**Accuracy**
Accuracy	Train Accuracy	0.769230769230769
Accuracy.1	Test Accuracy	0.705882352941177

The above table shows the accuracy of predicting both the training and test labels of the data using the KNN model. It is seen that the training accuracy is greater that the test accuracy, this can be assumed to be an adequate result as the model is trained using the training data and therefore will easily predict the training data labels also, the training data contains significantly a larger amount of data which is a factor that helps improve the overall accuracy of its predictions.

e.)

While comparing both models, it can be observed that both are capable of predicting the classes at an acceptable level, and given the size of the data, they have both performed well in predicting a small with high accuracy, but, the logistic regression tends to be more accurate in predicting its training data when compared to the KNN model. For such a small data, I would advice any researcher to use the logistic regression not only because of its better performance with the training data, but also, it is observed that the KNN model runs at a comparatively lower speed even with the small data. The logistic model also provides probability prediction and not just classification labels as found in the KNN model. In addition, the logistic model interprets coefficients. Overall, the logistic model is easier to implement, interpret and has a higher training efficiency.