Project 3

Md Mominul Islam

2021-11-26

Project

Project Summary

In our project we have developed a logistic regression model with 46 ***Subterranious*** samples, 43 ***Multiplex*** samples and the 199 ***Unknown*** samples.Initially we have done formatting and cleaning of our given data set which were subdivided into three csv files and then we did some common exploratory data analysis like pairs plot, scatterplot, boxplot, density plot, correlation analysis between numeric variables etc. From our exploratory data analysis, we have drawn several conclusions about the symmetry, skewness and distribution of our 3 separate data set of Multiplex, subterranious and unknown species. After that, we Prepared a two preliminary models with two and three variables using generalized linear regression model. From our generalized linear model, We can see that the coefficients of Vole Skull Height (Scale of 0.01 mm) and Vole Skull Weight (Scale of 0.01 mm) are significant ($p < 0.05$), while the coefficient of Vole Skull Length (Scale of 0.01 mm) is non-significant considering three variables. Our model with 2 variables has AIC of 60.4188 and model with three variables has the smallest AIC(56.87). So based on AIC, we have selected the model with the lowest AIC. Using the p-value in the ANOVA output, we rejected the null hypothesis as the p-value was less than the significance level and concluded that at least two of the species parameters are different from each other. Also we have done QQ plotting of our regression model and saw that the the data was normally distributed for our model except some outliers like 24,241 and 271.After moidelling we have done cross-validation for the models and checked model accuracy. Our cross validation showed that the full model and the model based of width and height are effectively the same for the LOOCV. Furthermore, the model with just Vole Skull Width and Vole Skull Height ("WH") is better than the models with Vole Skull Length in the model. Lastly, we have predicted for the unknown species. out of 197 unknown samples, there are 82 multiplex data and 115 subterranious data. The error rate of our model is 58.37%. .

Develop a Logistic regression model from the 89 specimens that you can use to predict the group membership of the remaining 199 specimens.

Methodology

- For this project, the data set is imported from the ***Skull Data Work*** as ***skull.data***. For our convenience we have saved a copy of the original data and did our analysis on work data set.Then, we looked at the raw data and have an overall idea about the dataset like presence of potential outliers or typos. In real world, it is better to contact the researchers about these data points.
- There are 46 ***Subterranious*** samples and 43 ***Multiplex*** samples and the remaining 199 samples are ***Unknown***. It is apparent that this is a classification problem with a binary outcome. In this project we are asked to develop a logistic regression model from the 89 specimens that we can use to predict the group membership of the remaining 199 specimens.Hence, we are going to develop a logistic regression model using the ***glm*** function in base R.
- Firstly, we have done some common exploratory data analysis like pairs plot, scatterplot, boxplot, density plot, correlation analysis between numeric variables etc. After that, we Prepared a preliminary

model like logistic regression or linear regression model. After moidelling we have done cross-validation for the models and checked model accuracy. At the end, we compared different models and chose the one with the highest accuracy.Finally, With the finalized model we have done some visualizations and used the model to make predictions.

Assumptions

The following assumptions are made in the process of a building binomial logistic regression model:

- Dependent variable is binary or discrete.
- The observations must be independent of each other.
- Continuous explanatory variables follow normal Gaussian distribution.
- A linear relationship exists between the independent explanatory variables and the logit output
- No outliers that exert undue influence on the model.
- No troublesome multicollinearity among the independent variables.

Data Exploration Methodology

Notes on xlxs file

We quickly visualized the xlxs file. In the given file, We have 3 sheets of data.

First Sheet: Unknown data By looking at the ***Unknown Data***, we summarized few points

- A skipped row between vole skulls 20 and 21.
- Looks like vole skull width at 10th row maybe a typo.

Second Sheet: Subterraneous Vole Skulls By looking at the ***Subterraneous vole skulls Data***, we summarized few points

- Vole skull typo on width at 2nd row.
- Vole skull typo on length at 13th row.
- Vole skull typo on height at 33rd row.

Third Sheet : Multiplex Vole Skulls

By looking at the ***Multiplex Vole Skulls Data***, we summarized few points

- Typo on vole skull height at 8th row
- Typo on vole skull length at 39th row.
- Looks like 29 and 30 may be a stutter or the same skull entered twice.

1. **Format the data contained in the excel spreadsheet for use in R.**
2. **Perform an exploratory data analysis.**

Loading Data Set in R

We have made three csv files out of the xlxs file and named them as ***multi.csv, sub.csv and unknown.csv***. We have loaded these three csv files and named them dat.m, dat.s and dat.u respectively.

Formatting and Cleaning Multiplex data (dat.m)

We have removed typos at 8th and 39th row. We have kept 29th row so our double entry problem is solved as well.

Table 1: First 6 Rows of Multiplex data

Skull.Index	Species	Length	Height	Width
1	multiplex	2355	805	475
2	multiplex	2305	760	450

Skull.Index	Species	Length	Height	Width
3	multiplex	2388	775	460
4	multiplex	2370	766	460
5	multiplex	2470	815	475
6	multiplex	2535	838	521

```
## Multiplex data has 41 rows and 5 columns.
```

Summary Statistics of Numeric Variables of Multiplex data

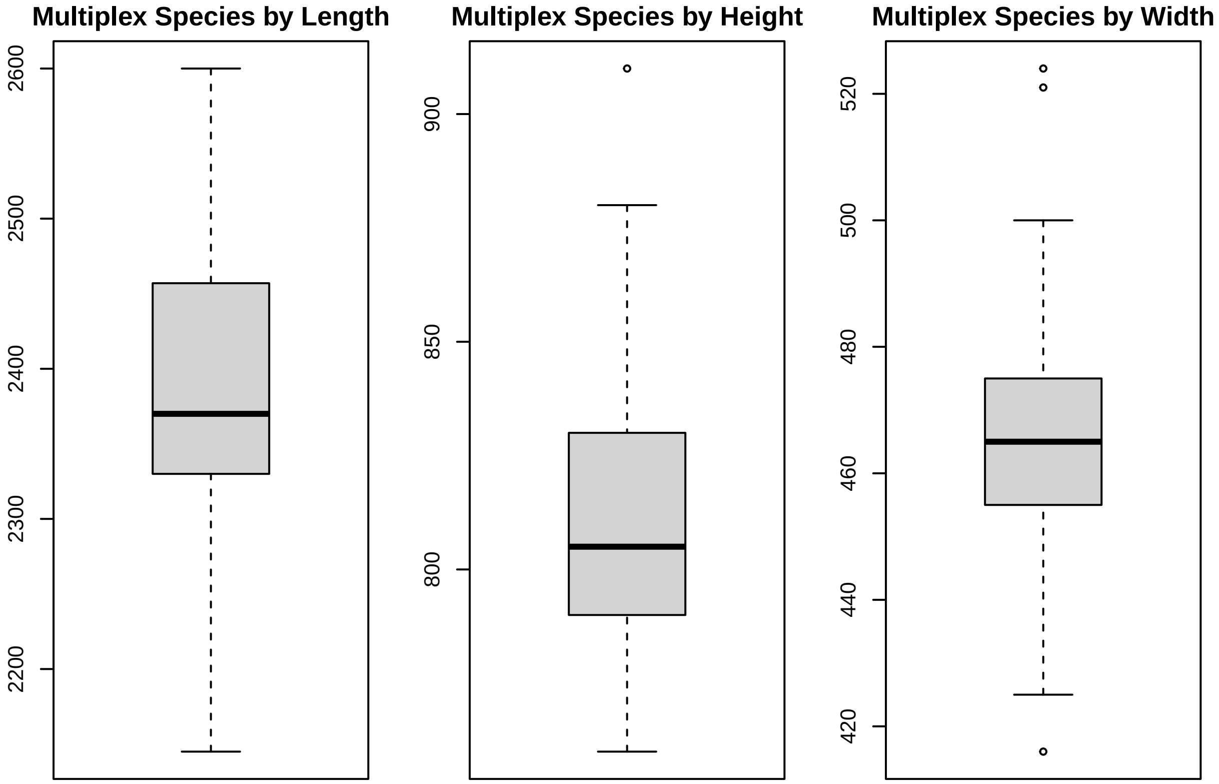
Table 2: Summary Statistics of Numeric Variables of Multiplex data

Length	Height	Width
Min. :2145	Min. :760	Min. :416.0
1st Qu.:2330	1st Qu.:790	1st Qu.:455.0
Median :2370	Median :805	Median :465.0
Mean :2386	Mean :809	Mean :466.3
3rd Qu.:2457	3rd Qu.:830	3rd Qu.:475.0
Max. :2600	Max. :910	Max. :524.0

Measures of central tendency are a class of statistics used to identify a value that falls in the middle of a set of data. From our summary statistic for the numeric variables, we can say that

- Average value of length suggests that typical length for the multiplex species would go around 2386 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (2386) and median (2370) values for the length column in multiplex data are close.
- Average value of height suggests that typical height for the multiplex species would go around 809 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (809) and median (805) values for the height column in multiplex data are close.
- Average value of width suggests that typical width for the multiplex species would go around 466.3 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (466.3) and median (465) values for the width column in multiplex data are close.

Individual Boxplot for Multiplex Species



We know that boxplot gives us a good indication of how the values in the data are spread out. The boxplot displays the center and spread of a numeric variable in a format that allows us to quickly obtain a sense of the range and skew of a variable, or compare it to other variables.

- From the boxplot depicted above, we can see some outliers (identified by asterisks (*)) in height and width column.
- The length and height data for Multiplex Species are skewed to the right as the longer part of the box are to the right (or above) the median.
- But, we can say that data is symmetric in case of width for Multiplex Species.

Individual Histogram for Multiplex Species

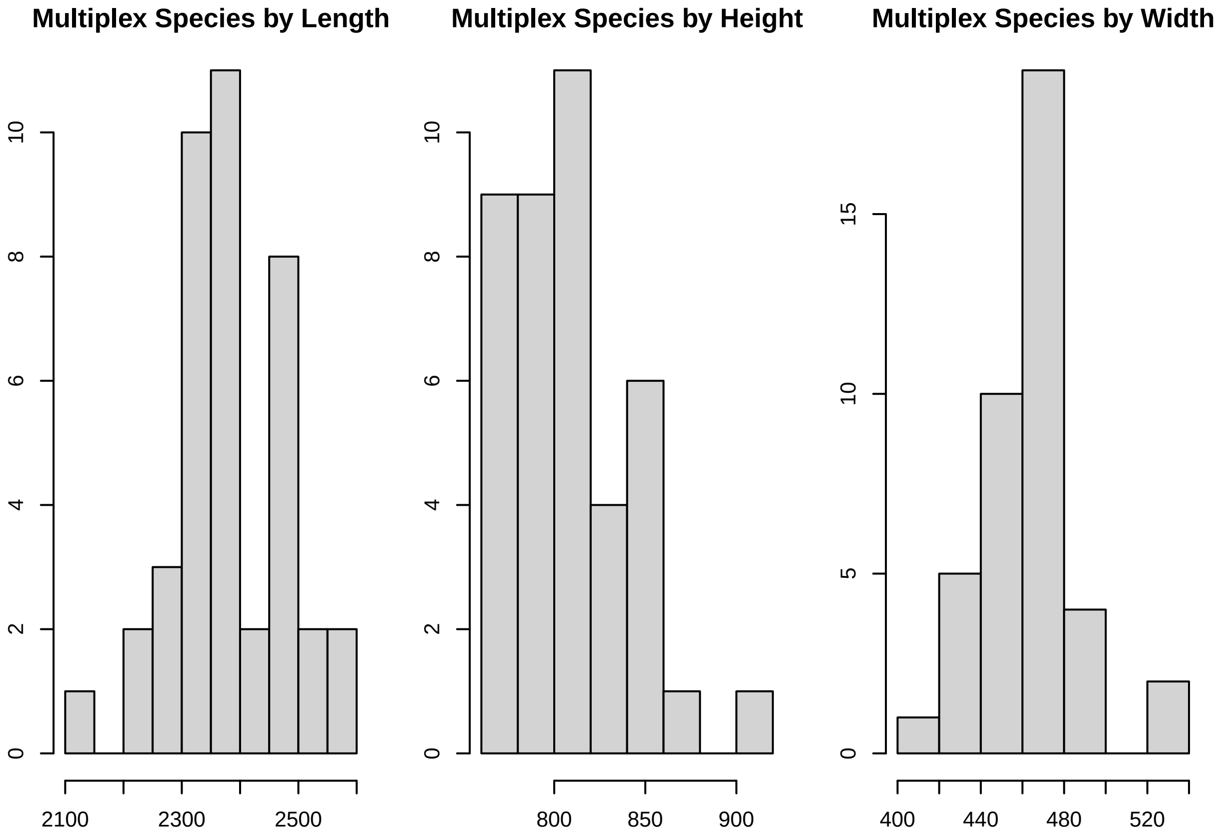


Figure 1: Histogram for Multiplex Species

From the histogram, we can say that:

- For the histogram of the length of multiplex species, the shape looks like symmetrical except some outliers at the beginning.
- The shape is right skewed for height of the multiplex species. It has a peak that is left of center. This is a unimodal data set, with the mode closer to the left of the graph and smaller than either the mean or the median. The mean (809) of right-skewed data is located to the right side of the graph and greater value than the median (805). This shape indicates that there are a number of data points, perhaps outliers, that are greater than the mode.
- The shape is symmetrical for width of the multiplex species. So we can tell that the mean, median, and mode are all the same value.Also a normally distributed data set except an outlier at the end creates a symmetric histogram that looks like a bell, leading to the common term for a normal distribution: a bell curve.

Pairs Plot for Multiplex Species

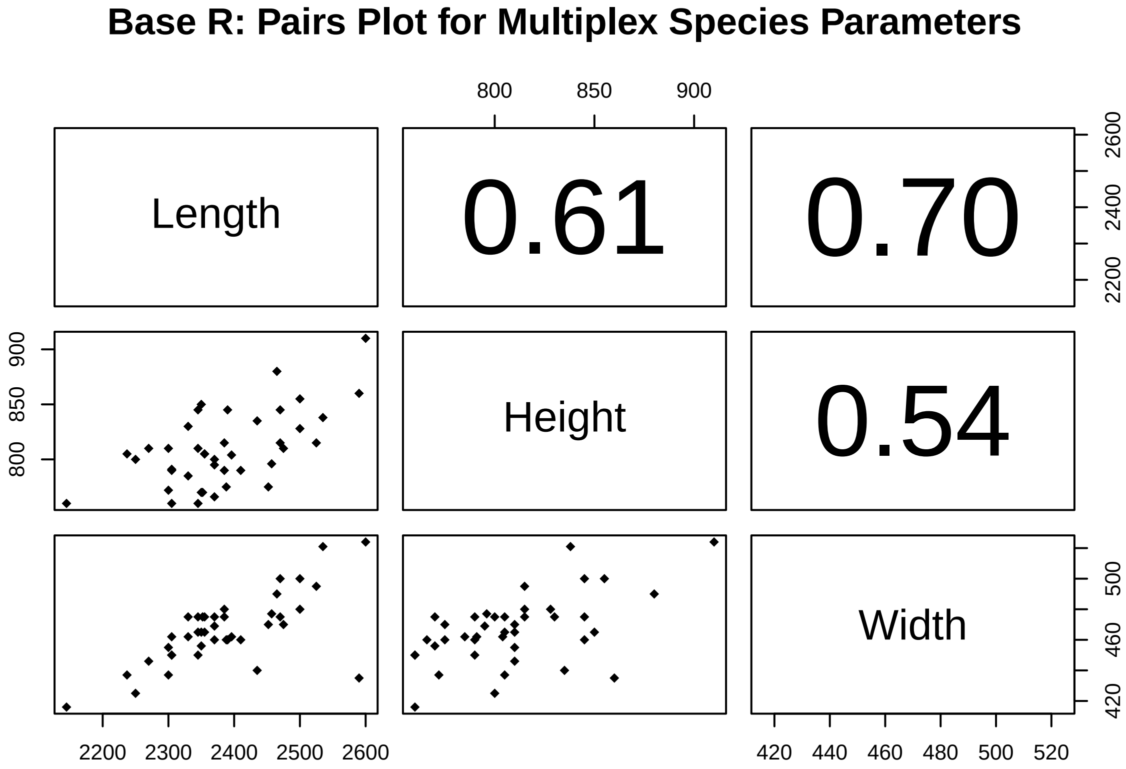


Figure 2: Pairs Plot for Multiplex Species

Table 3: Correlation Matrix for Multiplex Data

	Length	Height	Width
Length	1.0000000	0.6061855	0.6967046
Height	0.6061855	1.0000000	0.5401901
Width	0.6967046	0.5401901	1.0000000

All other boxes display a scatterplot of the relationship between each pairwise combination of variables.

- The box in the left right corner of the matrix displays a scatterplot of values for length and width. The box in the middle left displays a scatterplot of values for Length and Height, and so on.
- This single plot gives us an idea of the relationship between each pair of variables in our dataset. For example, parameter Length and width seem to be positively correlated with a value of 0.70.

Density Plot for Multiplex Species

Density plots are used to observe the distribution of a variable in a dataset. It plots the graph on a continuous interval or time-period. This is also known as Kernel density plot. A density curve gives us a good idea of

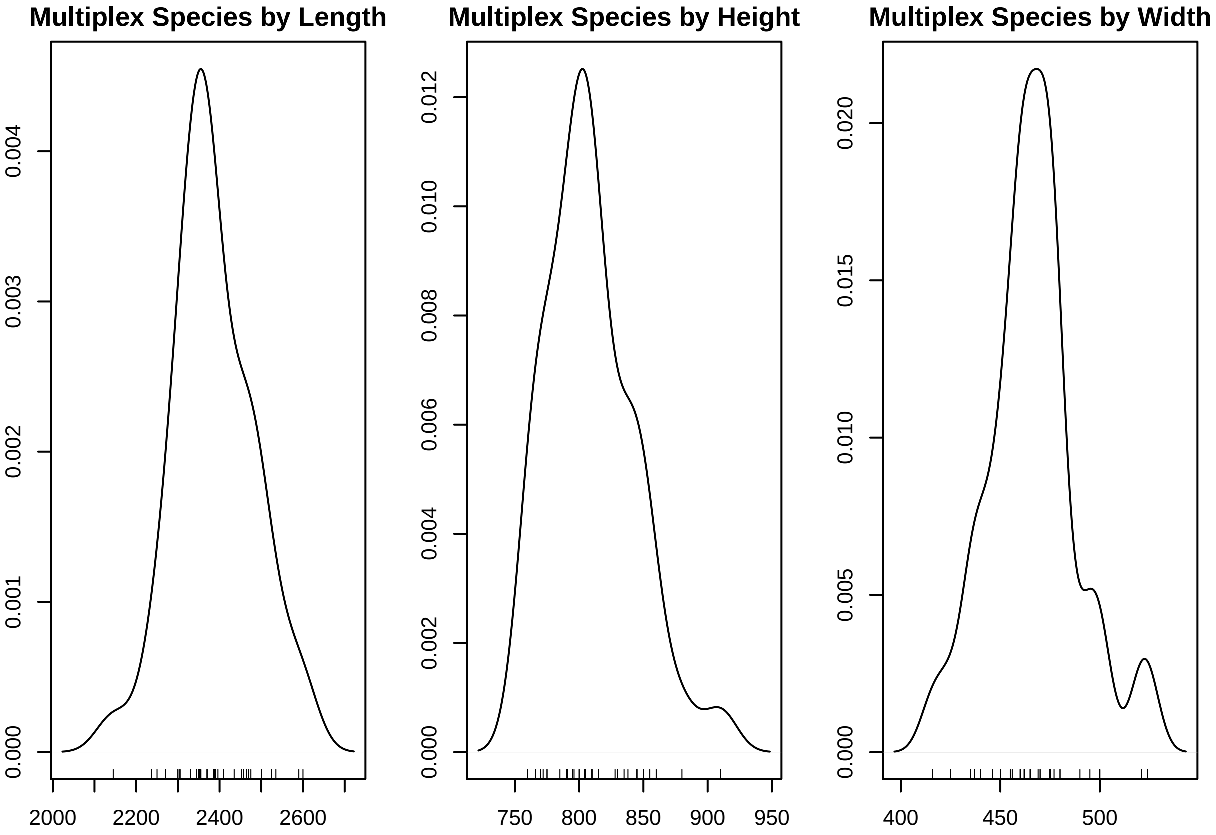


Figure 3: Desnity Plot for Multiplex Species

the “shape” of a distribution, including whether or not a distribution has one or more “peaks” of frequently occurring values and whether or not the distribution is skewed to the left or the right.

- For length of multiplex species, density curve is right skewed, then the mean is greater than the median.
- For height of multiplex species, density curve is right skewed, then the mean is greater than the median.
- For width of multiplex species, density curve has no skew, then the mean is equal to the median.

Formatting and Cleaning Subterraneous data (dat.s)

By formatting our subterranious data, we got rid of missing values and typos.

- Firstly, we have checked missing values and found missing values at 13th row. We removed the entire row.
- We have removed typos at 2nd and 13th row and also fixed vole skull 33.

Table 4: First 6 Rows of Subterranious data

	Skull.Index	Species	Length	Height	Width
1	1	subterraneus	2350	735	450
3	3	subterraneus	2170	738	415
4	4	subterraneus	2060	720	415
5	5	subterraneus	2275	785	417
6	6	subterraneus	2330	790	450
7	7	subterraneus	2260	760	426

```
## Subterranious data has 43 rows and 5 columns.
```

Summary Statistics of Numeric Variables of Subterranious data

Table 5: Summary Statistics of Numeric Variables of Subterranious data

Length	Height	Width
Min. :1965	Min. :715.0	Min. :395.0
1st Qu.:2172	1st Qu.:740.0	1st Qu.:415.0
Median :2250	Median :750.0	Median :425.0
Mean :2227	Mean :757.7	Mean :426.8
3rd Qu.:2290	3rd Qu.:775.5	3rd Qu.:433.5
Max. :2365	Max. :805.0	Max. :488.0

From our summary statistic for the numeric variables, we can say that

- Average value of length suggests that typical length for the subterranious species would go around 2227 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (2227) and median (2250) values for the length column in subterranious data are close.
- Average value of height suggests that typical height for the subterranious species would go around 757.7 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (757.7) and median (750) values for the height column in subterranious data are close.
- Average value of width suggests that typical width for the subterranious species would go around 426.8 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (426.8) and median (425) values for the width column in subterranious data are close.

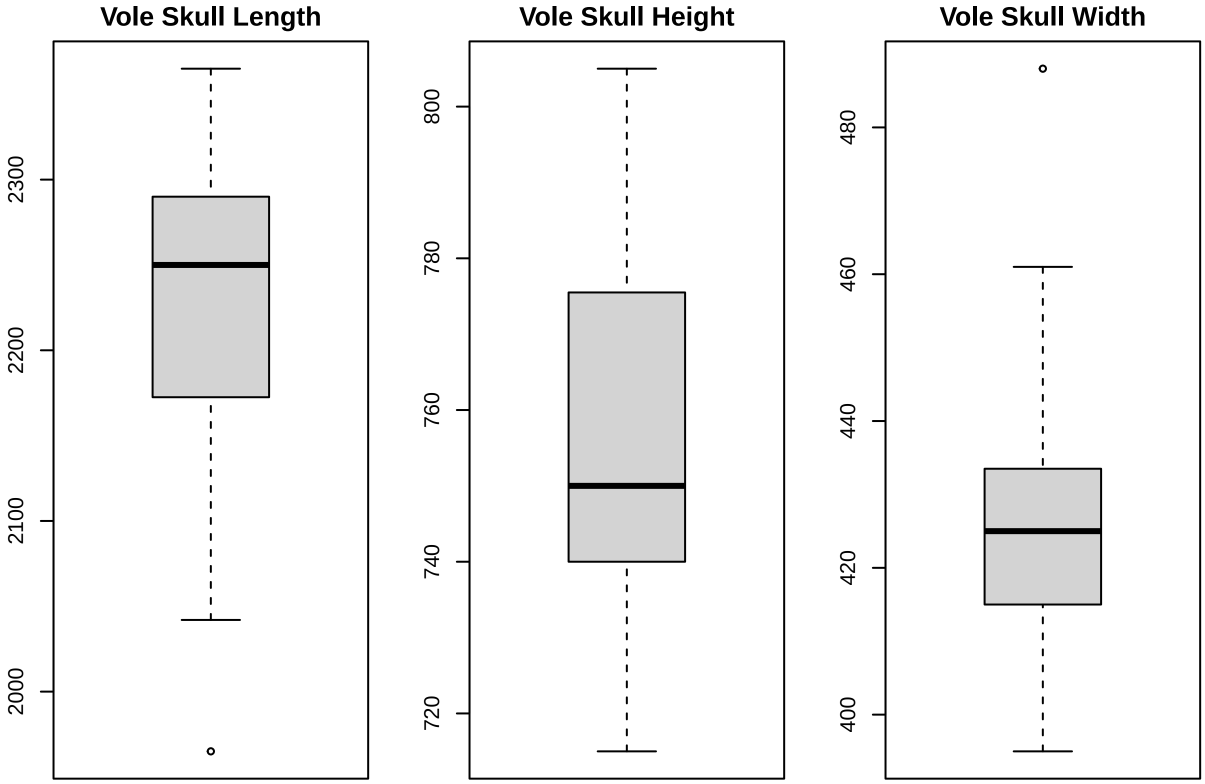


Figure 4: Boxplot for Subterranious Species

Individual Boxplot for Subterranious Species

From the boxplot depicted above, we can see an outlier (identified by asterisks (*)) in width column.

- The length data for subterranious Species are skewed to the left as the longer part of the box are to the left to the median.
- The height data for subterranious Species are skewed to the right as the longer part of the box are to the right (or above) the median.
- But, we can say that data is symmetric in case of width for subterranious Species.

Individual Histogram for Subterranious Species

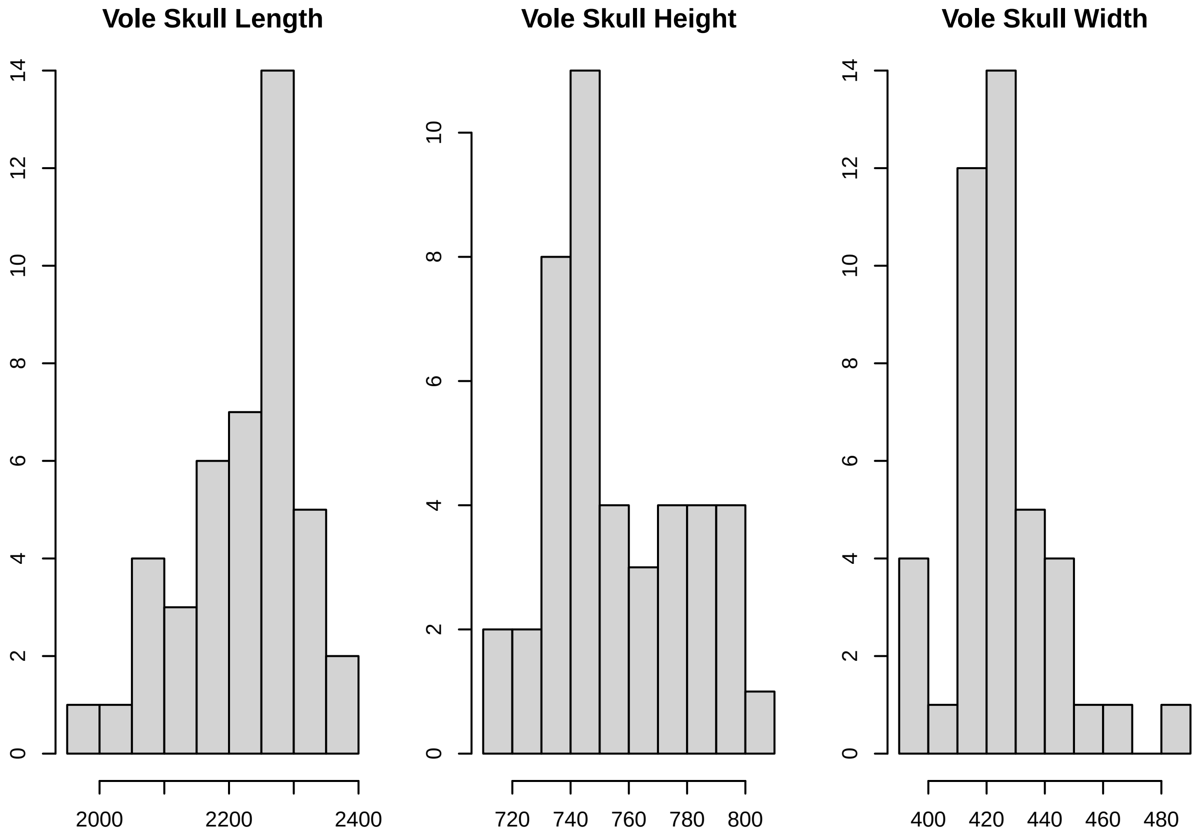


Figure 5: Histogram for Subterranious Species

From the histogram, we can say that:

- The shape is left skewed for length of the subterranious species. It has a peak that is right of center. This is a unimodal data set, with the mode closer to the right of the graph and smaller than either the mean or the median. The mean (2227) of skewed data is located to the right side of the graph and smaller value than the median (2250).
- The shape is right skewed for height of the subterranious species. It has a peak that is left of center. This is a unimodal data set,the mean (757.7) of skewed data is located to the left side of the graph and greater value than the median (750).

- For the histogram of the width of subterranious species, the shape looks like symmetrical except some outliers at the end.

Pairs Plot for Subterranious Species

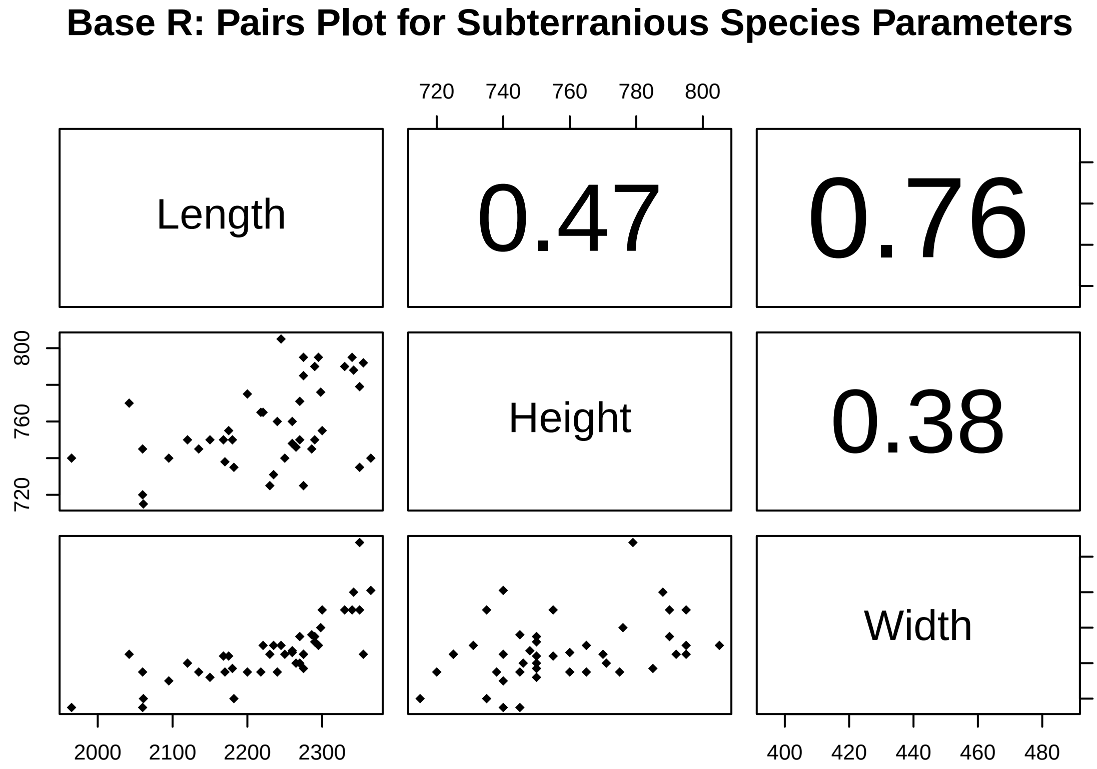


Figure 6: Pairs Plot for Subterranious Species

Table 6: Correlation Matrix for Subterranious Data

	Length	Height	Width
Length	1.0000000	0.4664476	0.7597001
Height	0.4664476	1.0000000	0.3806132
Width	0.7597001	0.3806132	1.0000000

All other boxes display a scatterplot of the relationship between each pairwise combination of variables.

- The box in the left right corner of the matrix displays a scatterplot of values for length and width. The box in the middle left displays a scatterplot of values for Length and Height, and so on.
- This single plot gives us an idea of the relationship between each pair of variables in our dataset. For example, parameter Length and width seem to be positively correlated with a value of 0.76.

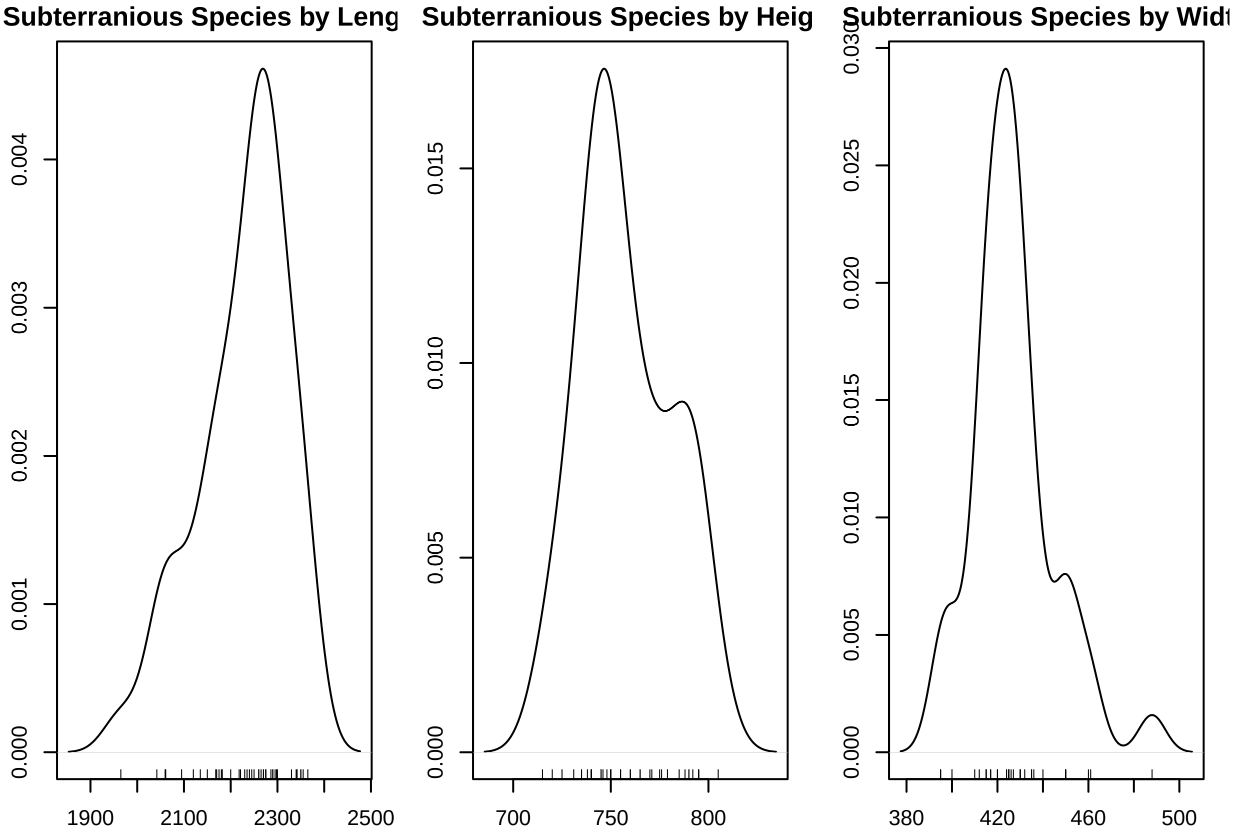


Figure 7: Desnity Plot for Subterranious Species

Density Plot for Subterranious Species

Density curve gives us a good idea of the “shape” of a distribution, including whether or not a distribution has one or more “peaks” of frequently occurring values and whether or not the distribution is skewed to the left or the right.

- For length of subterranious species, density curve is left skewed, then the mean is smaller than the median.
- For height of subterranious species, density curve is right skewed, then the mean is greater than the median.
- For width of subterranious species, density curve has no skew, then the mean is equal to the median.

Formatting and Cleaning Unknwon data (dat.u)

By formatting our unknwon data, we got rid of missing values and typos.

- Firstly, we have checked missing values and found missing values at 21th row. We removed the entire row.
- We have removed typos at 7th and 10th row.

Table 7: First 6 Rows of Unknown data

Skull.Index	Species	Length	Height	Width
1	unknown	2232	821	430
2	unknown	2140	755	405
3	unknown	2295	767	425
4	unknown	2355	842	490
5	unknown	2335	814	481
6	unknown	2355	815	460

```
## Unknwon data has 197 rows and 5 columns.
```

Summary Statistics of Numeric Variables of Unknown data

Table 8: Summary Statistics of Numeric Variables of Unknown data

Length	Height	Width
Min. : 1908	Min. :700.0	Min. :375.0
1st Qu.: 2224	1st Qu.:760.0	1st Qu.:428.0
Median : 2320	Median :790.0	Median :453.0
Mean : 2419	Mean :794.6	Mean :452.9
3rd Qu.: 2407	3rd Qu.:825.0	3rd Qu.:475.0
Max. :23555	Max. :912.0	Max. :545.0

From our summary statistic for the numeric variables, we can say that

- Average value of length suggests that typical length for the unknown species would go around 2419 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (2419) and median (2320) values for the length column in unknown data are close.
- Average value of height suggests that typical height for the unknown species would go around 794.6 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (794.6) and median (790) values for the height column in unknown data are close.
- Average value of width suggests that typical width for the unknown species would go around 452.9 (Scale of 0.01 mm) and we can also say that data is symmetrical as mean (452.9) and median (453) values for the width column in unknown data are close.

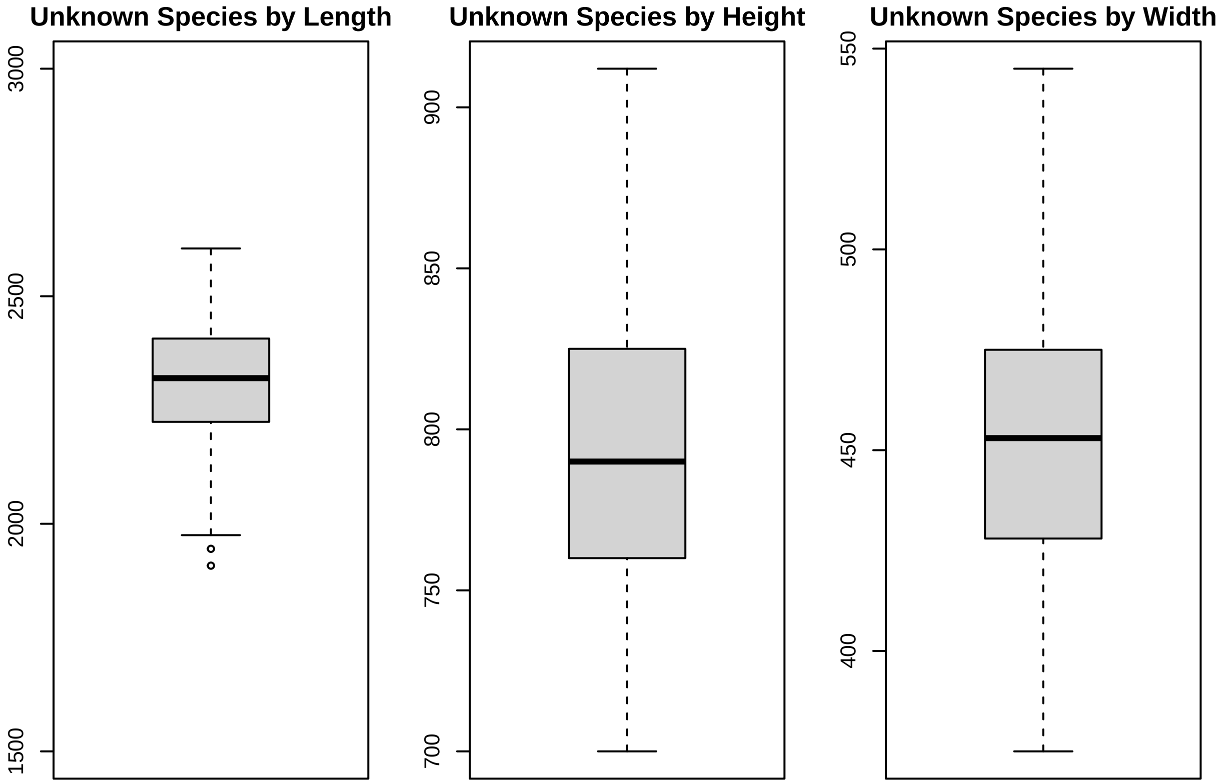


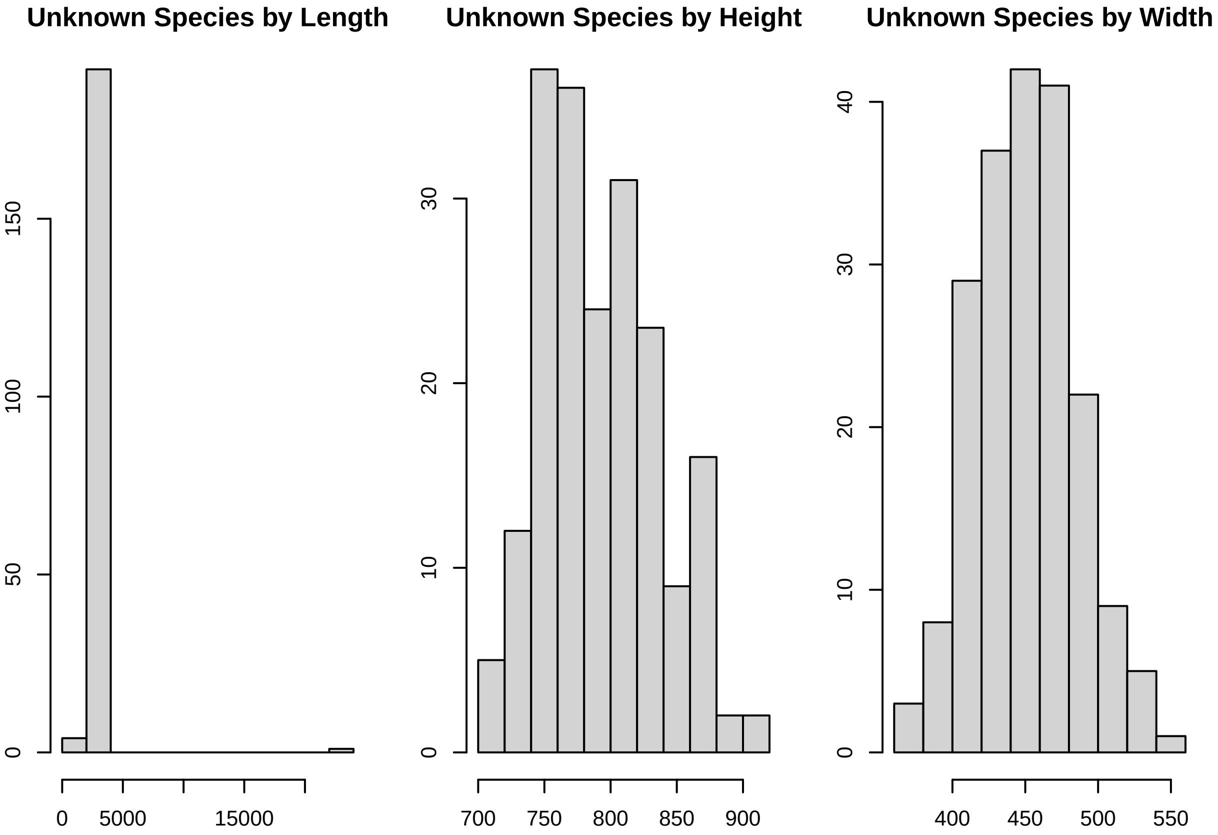
Figure 8: Boxplot for Unknown Species

Individual Boxplot for Unknown Species

From the boxplot depicted above, we can see outliers (identified by asterisks (*)) in length column of the unknown species.

- We can say that data is symmetric in case of height and width for unknown Species.

Individual Histogram for Unknown Species



From the histogram, we can say that:

- The shape of length is undefined.
- For the histogram of the height and width of unknown species, the shape looks like symmetrical.

Pairs Plot for Unknown Species

Table 9: Correlation Matrix for Unknown Data

	Length	Height	Width
Length	1.0000000	0.1078873	0.0506884
Height	0.1078873	1.0000000	0.7977468
Width	0.0506884	0.7977468	1.0000000

All other boxes display a scatterplot of the relationship between each pairwise combination of variables.

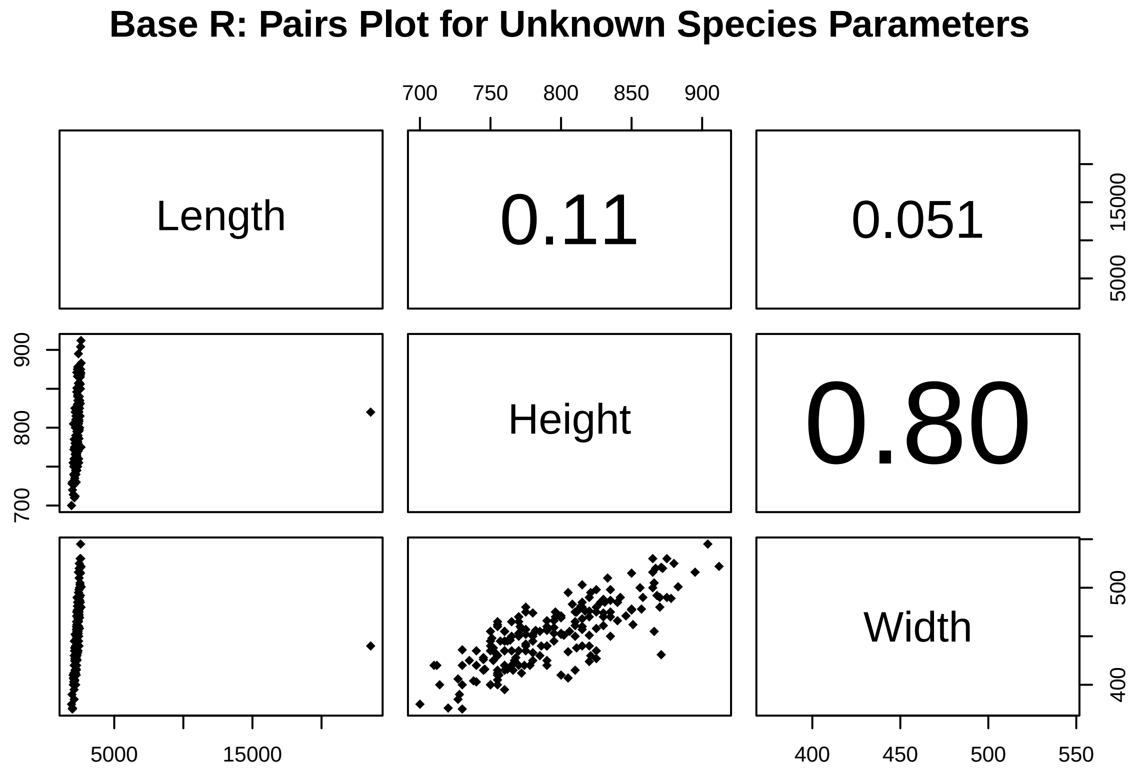
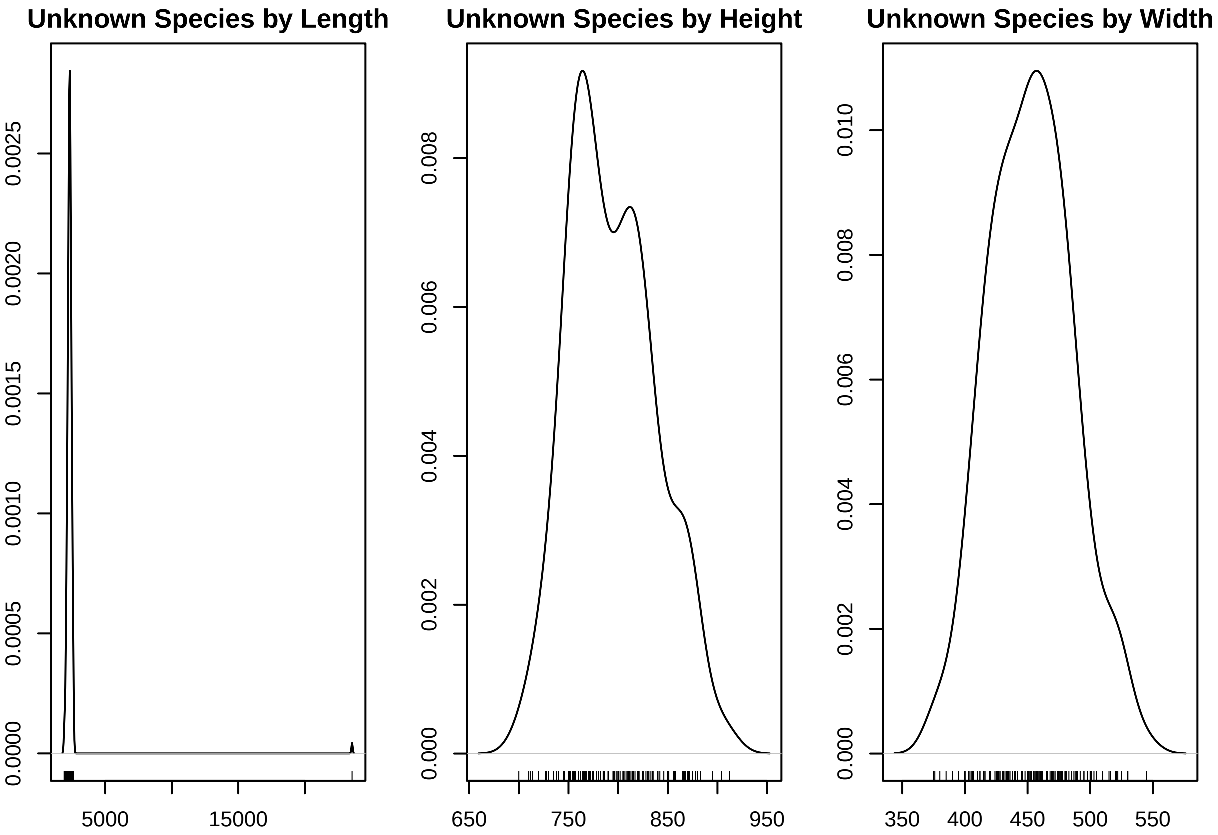


Figure 9: Pairs Plot for Unknown Species

- From the correlation matrix, we can say that Length and height of unknown species are weakly correlated with a value of 0.11.
- Parameter Length and width seem to be positively correlated with a value of 0.80.

Density Plot for Unknown Species



Density curve gives us a good idea of the "shape" of a distribution, including whether or not a distribution has one or more "peaks" of frequently occurring values and whether or not the distribution is skewed to the left or the right.

- For length of unknown species, density curve is narrower and extremely left skewed.
- For height and width of unknown species, density curve has no skew, then it indicates that the mean is equal to the median.

3. Explain your GLM and assess the quality of the fit with the classified observations.

**Answer*

Table 10: Summary of Linear model, Two Variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	77.3115165	18.0061987	4.293606	0.0000176
Length	-0.0168148	0.0061647	-2.727571	0.0063802
Height	-0.0492608	0.0160147	-3.075977	0.0020981

Table 11: Summary of Linear model, Three Variables

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	71.2440515	16.1983635	4.3982253	0.0000109
Length	-0.0014898	0.0075345	-0.1977259	0.8432595
Height	-0.0471862	0.0168347	-2.8029104	0.0050644
Width	-0.0694676	0.0306765	-2.2645172	0.0235423

From our generalized linear model, We can see that the coefficients of height and weight are significant ($p < 0.05$), while the coefficient of length is non-significant considering three variables.

AIC Comparison of Two Models

Table 12: AICs of the two models

Model 1	Model 2
60.41883	56.86992

The Akaike Information Criterion (AIC) provides a method for assessing the quality of our model through comparison of related models. Much like adjusted R-squared, it's intent is to prevent us from including irrelevant predictors.

Our model with 2 variables has AIC of 60.4188 and model with three variables has the smallest AIC(56.87). So, we should select the model with the lowest AIC.

Analysis of Variance

```
## Analysis of Deviance Table
##
## Model 1: Species ~ Length + Height
## Model 2: Species ~ Length + Height + Width
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1        81     54.419
## 2        80     48.870  1   5.5489  0.01849 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

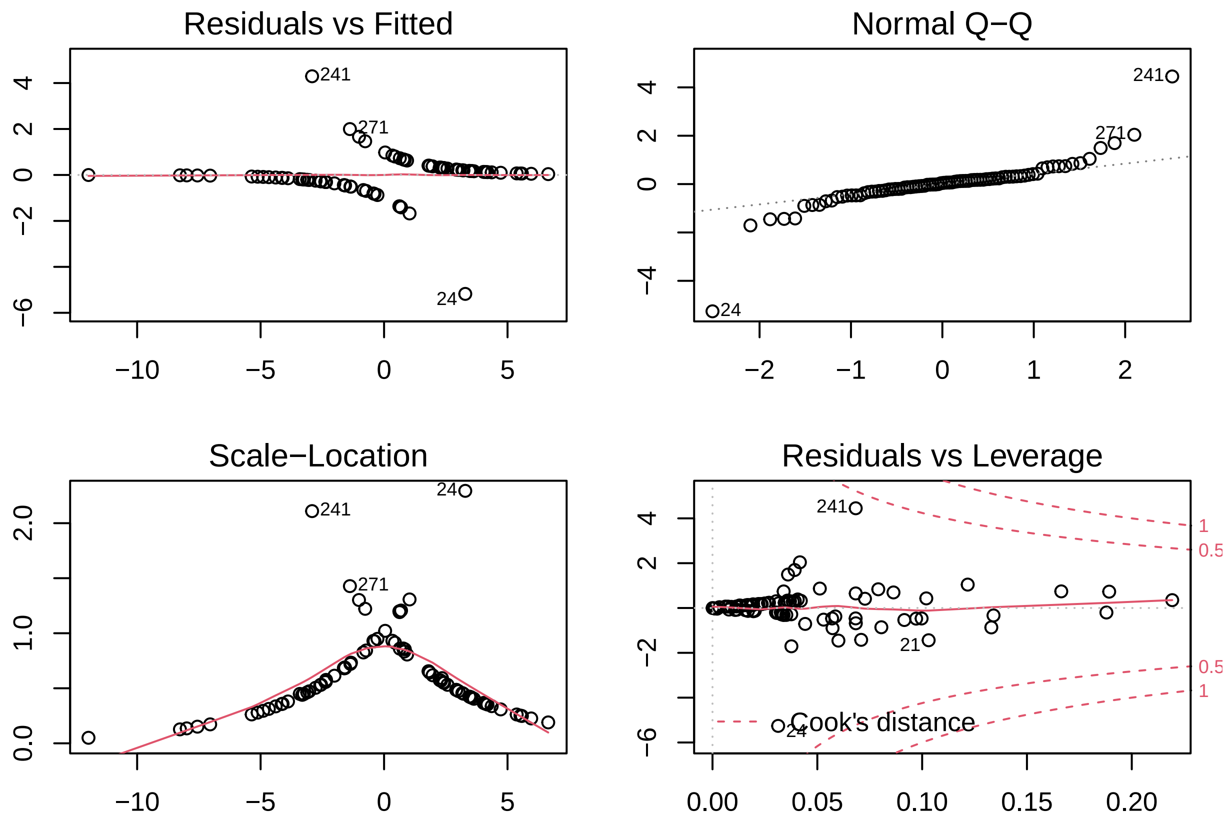
Using the p-value in the ANOVA output, we can determine whether the differences between some of the means are statistically significant. As the p-value is less than the significance level, we reject the null hypothesis and conclude that at least two of the species parameters are different from each other.

Deviance Residual of Model with Three Variables

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.578816 -0.383079  0.061354  0.008953  0.390838  2.436337
```

Since the median deviance residual is close to zero, this means that our model is not biased in one direction (i.e. the out come is neither over- nor underestimated).

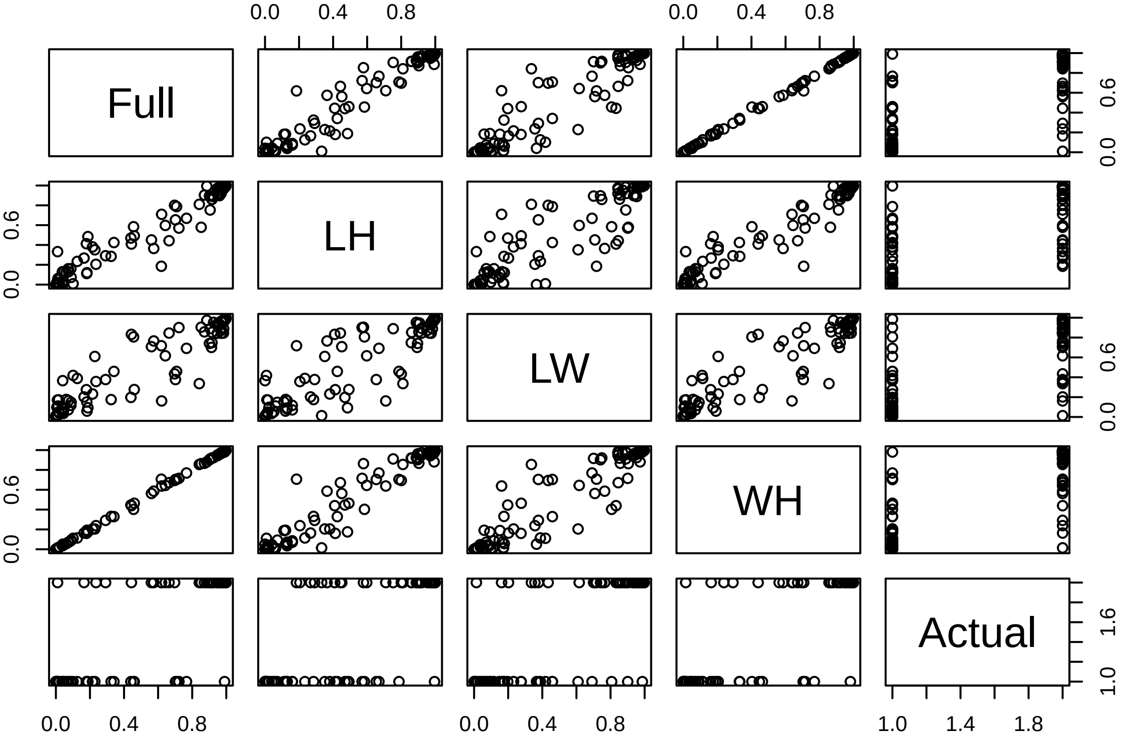
Base R Regression Model Plotting

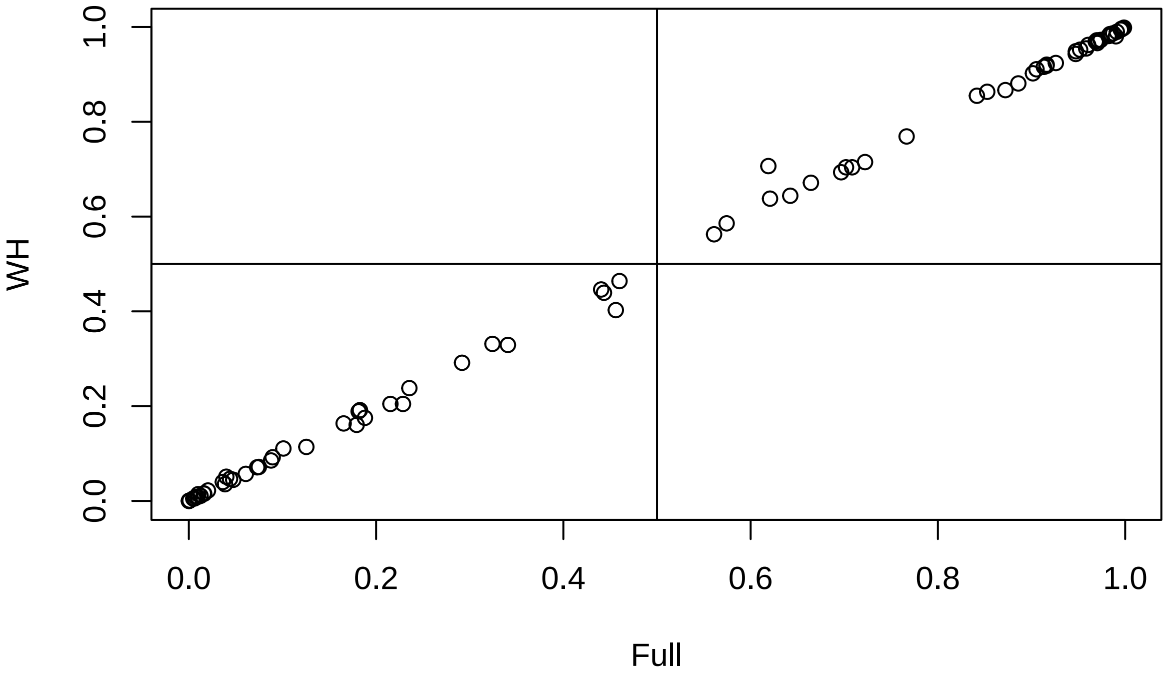


From normal Q-Q plot, we can tell either the data is normally distributed or not. If all the points plotted on the graph perfectly lies on a straight line then we can clearly say that this distribution is Normally distribution because it is evenly aligned with the standard normal variate which is the simple concept of Q-Q plot.

- From our model, we can clearly see that the the data is normally distributed for our model except some outliers like 24,241 and 271.
- We can also say that the two sets come from a population with the same distribution, as the points fall approximately along the reference line drawn at 45-degree angle.
- A residuals vs. leverage plot is a type of diagnostic plot that allows us to identify influential observations in a regression model.We can see that observation #24 and #241 lie closest to the border of Cook's distance, but they don't fall outside of the dashed line. This means there are not any influential points in our regression model.

Cross Validation of Model





The above plot shows that the full model and the model based of width and height are effectively the same for the LOOCV.

```
##       LH
## Actual  1  2
##      1 35  6
##      2  9 34
```

```
##       LW
## Actual  1  2
##      1 36  5
##      2  7 36
```

```
##       WH
## Actual  1  2
##      1 36  5
##      2  5 38
```

This suggests that using the model with just “WH” is better than the models with length in the model.

5. Provide predictions for the unclassified observations.

Table 13: Rows are Predicted Value and Columns are True Value

	Frequency
Multiplex	82
Subterranious	115

```
## [1] "Error For model is  58.3756345177665  %"
```

From our model, we can predict that out of 197 unknown samples, there are 82 multiplex data and 115 subterranious data. The error rate of our model is 58.37%.

Sources

0. Airoldi, J.-P., B. Flury, M. Salvioni (1996) “Discrimination between two species of Microtus using both classified and unclassified observations” Journal of Theoretical Biology 177:247-262

1. Histogram: https://www.statmethods.net/graphs/density.html

2. Feature Plot : https://topepo.github.io/caret/visualizations.html

3. Binwidth: https://stackoverflow.com/questions/14200027/how-to-adjust-binwidth-in-ggplot2

4. Boxplot: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

5. https://www.wellbeingatschool.org.nz/information-sheet/understanding-and-interpreting-box-plots

6. Assumptions of Logistic Regression:https://www.statisticssolutions.com/free-resources/directory-of-statistical-analyses/assumptions-of-logistic-regression/

7. Classification Logistic Regression: https://github.com/cran/sparklyr/blob/c0effdbed11c95e42ea37193b1cfe2516217516b R/ml_classification_logistic_regression.R

8. 601 Final Project: https://github.com/achalneupane/achalneupane/blob/26d43b15758ded9aeb5d3ae36a05926143697a3 achalneupane.github.io/post/Stat_601_FINAL.rmd

9. Logistic Regression Slides: https://courses.washington.edu/b513/handouts/b513_2013_2-2x2.pdf

10. Basic Linear Regression : https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/

11. Confusion Matrix : https://stackoverflow.com/questions/65124061/confusion-matrix-for-a-logistic-model

12. Main Code for Project: https://github.com/AminBaabol/Modern-Applied-Statisitcs-I/blob/c334a0af99caf0dc56c2d4bdaf971aac472436ca/Project/AminBaabol_FinalQ1.Rmd

13. https://rpubs.com/aelhabr/logistic-regression-tutorial

14. Link : https://www.rpubs.com/Quinn_Fargen/STAT701FinalQuinn

15. https://stackoverflow.com/questions/65124061/confusion-matrix-for-a-logistic-model

16. https://www.machinelearningplus.com/machine-learning/logistic-regression-tutorial-examples-r/

17. https://github.com/achalneupane/achalneupane/blob/26d43b15758ded9aeb5d3ae36a05926143697a3c/achalneupane.github.io/post/Stat_601_FINAL.rmd

18. https://github.com/TZstatsADS/spr2017-proj3-group3/blob/8394b05d3176366389265201602a4a18b83d5af7/lib/logistic_Regression_for_all_features.R