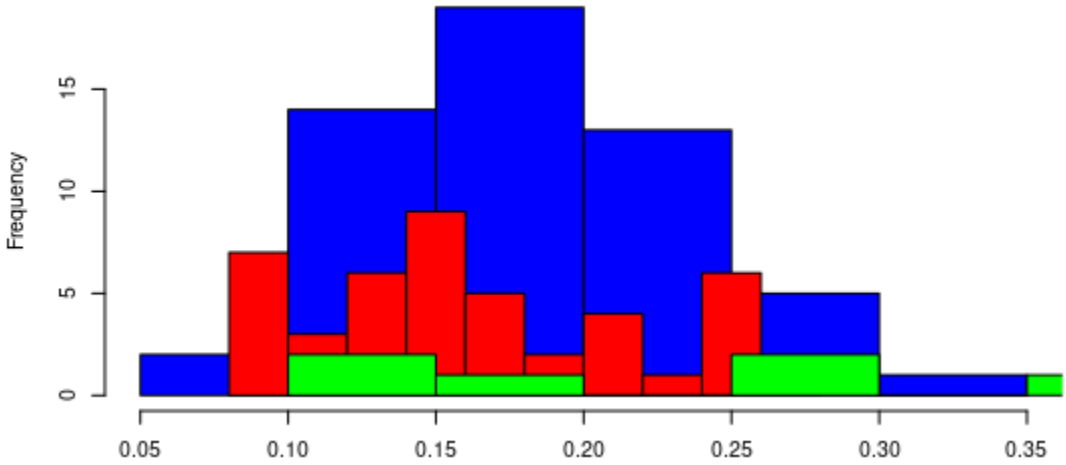
Cicada Dataset Analysis

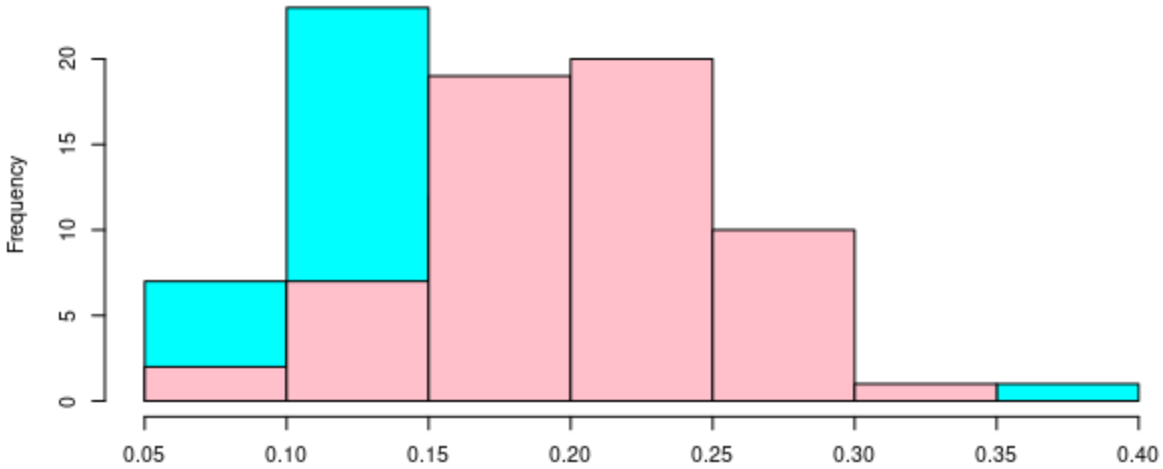
The purpose of this data report is to analyze the Cicada dataset and present key findings and insights derived from the analysis. The main goal of modeling is to try to predict each individual insect's species according to its physical characteristics. The Cicada dataset consists of 105 records of various cicadas. This report aims to provide valuable information about cicada's physical patterns across species, and trends that can be useful for researchers, entomologists, and enthusiasts.

The Cicada dataset was collected by Ginger Rowell and Robert Grammer, Belmont College. The dataset includes various variables such as body weight (BW), body length (BL), wing length (WL), wing width (WW), gender (G), and species (Species). Data preprocessing steps, such as cleaning, normalization, or imputation, were performed to ensure data quality and consistency.

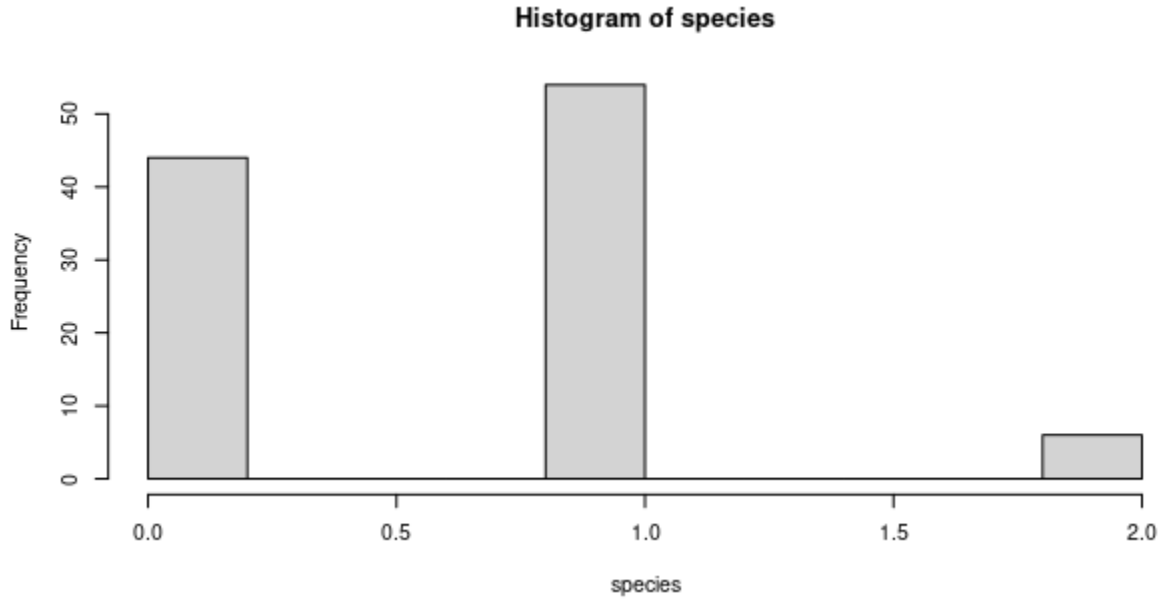
Before proceeding to the modeling, data exploration has to be performed. The first histogram defines the body weight distribution of each cicada across species, where blue color signifies species number 1, red stands for species 0, and green - 2. It is evident from the plot that each species indeed differ in its characteristics.



The next plot shows that female species of insects tend to have bigger body weights, thus some species' gender might be linearly correlated with body weight, thus it is very important to keep this in mind while training the model.



Moreover, we can see that the current dataset is not uniformly distributed for species. There is more than 50 samples of the first species, and a bit less number of species number zero, but the number of recordings of the second species is very small, thus we might not be able to properly project the physical characteristics of species 2, for its lack of data. Nevertheless, we still hope to properly classify species 0 and 1.



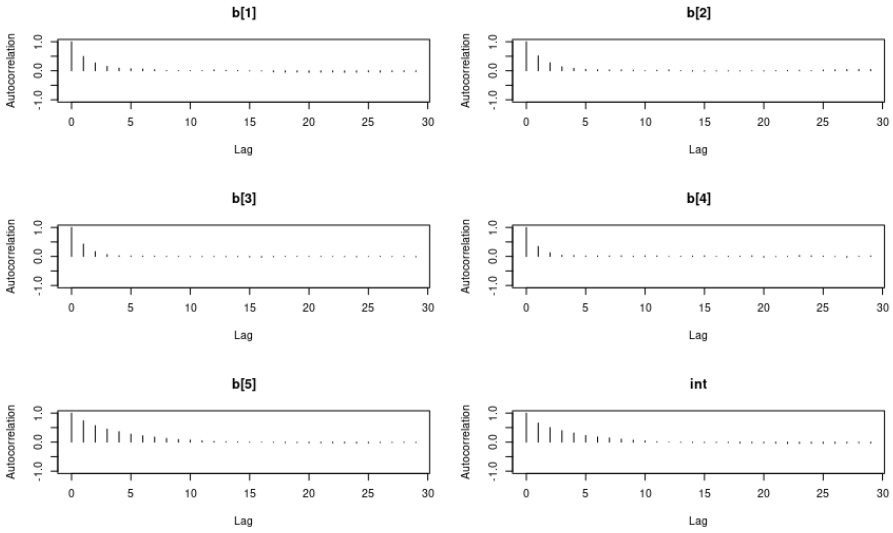
In order to train the model, we decided to exclude species 2, because it's lacking data. The model chosen was Monte Carlo Markov Chain Logistic Regression, which is the most suitable for binary classification. Before the training data was properly scaled in order to prevent this linear model to show coefficients that cannot be interpreted rationally. After training the model on the data, we have seen interesting values of the coefficient. The biggest absolute value of coefficients belongs to Gender (b[5]) being equal to -1.1822, though it had the biggest standard deviation which signifies that gender might be one of the most important feature for species classification, though it tends to have big standard deviation. Intercept here cannot be interpreted as such due to

the fact that the bug with zero weight and zero size does not exist. The second interpretable and most important feature of cicada for classification seem to be the Wing Width (b[3]), and it has a pretty low standard deviation.

```
        Mean     SD Naive SE Time-series SE
b[1] -0.2189 0.2662 0.002173       0.003980
b[2]  0.1867 0.2726 0.002226       0.003993
b[3]  0.3526 0.2736 0.002234       0.003548
b[4]  0.1327 0.2237 0.001826       0.002547
b[5] -1.1822 0.5751 0.004695       0.012927
int   0.7468 0.3413 0.002786       0.007100
```

Autocorrelation assessment showed that the model successfully converged. Standard values for update (1000) and simulating three chains (5000) resulted in convergence.

```
mod_string = " model {
    for (i in 1:length(y)) {
        y[i] ~ dbern(p[i])
        logit(p[i]) = int + b[1]*BW[i] + b[2]*WL[i] + b[3]*WW[i] + b[4]*BL[i] + b[5]*G[i]
    }
    int ~ dnorm(0.0, 1.0/25.0)
    for (j in 1:5) {
        b[j] ~ ddexp(0.0, sqrt(2.0)) # has variance 1.0
    }
} "
```



In order to justify the choice of the model, let's fit another model with simpler architecture. Using basic logistic regression fitted by the Least Square Method resulted in coefficients way different than in the MCMC model. The result again showed colinearity between the Gender variable and its Species. Though, it is not a valid variable to base on, due to the fact that the present dataset lacked enough data on the gender diversity across species of cicadas.

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.8159 -0.3377  0.1550  0.3482  0.9589

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.71307    0.07120  10.015  < 2e-16 ***
BW          -0.08744    0.06149  -1.422  0.15840
WL           0.02292    0.06530   0.351  0.72640
WW           0.08357    0.05800   1.441  0.15302
BL           0.03312    0.05227   0.634  0.52787
G           -0.37812    0.12520  -3.020  0.00327 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4633 on 92 degrees of freedom
Multiple R-squared:  0.1855,    Adjusted R-squared:  0.1412
F-statistic: 4.189 on 5 and 92 DF,  p-value: 0.001785
```

In analyzing the Cicada dataset, one of the key findings was the significance of the width variable in determining the species classification of cicadas. The wing width and gender of cicadas emerged as the most essential variables in accurately differentiating between species 0 and species 1 within the dataset. This finding highlights the importance of considering wing width and gender as crucial factors in species identification and classification for cicadas.

The weight of cicadas can provide valuable insights into their physiological development, nutritional status, and overall health. Differences in weight among cicada species can be attributed to various factors, including variations in diet, environmental conditions, and genetic differences. By incorporating weight as a variable in species classification models or algorithms, researchers and entomologists can improve the accuracy and reliability of species identification.

Accurate species classification is essential for understanding the biodiversity and ecological dynamics of cicadas. Different cicada species may exhibit distinct behaviors, mating patterns, or responses to environmental stimuli. Therefore, correctly identifying and classifying cicada species based on their weight can contribute to a more comprehensive understanding of their ecological roles, population dynamics, and conservation efforts.

In conclusion, the weight of cicadas has been found to be a crucial variable in determining species classification within the Cicada dataset. By considering weight along with other relevant variables, researchers can improve the accuracy and reliability of species identification, leading to a deeper understanding of cicada biodiversity and their ecological significance.