Introduction

The goal of this project is to analyze the traits of animals and understand how they can be used to classify species into different categories such as mammals, birds, reptiles, or aquatic animals. This study interests me because it combines my interest in love for animals and their biology with statistical methods and data science techniques to find patterns and relationships. By analyzing traits like the presence of hair, feathers, or whether animals lay eggs, I aimed to explore how specific features influence an animal's classification. Additionally I attempted to predict an animal's class type using statistical models such as logistic regression.

Methodology

I did Exploratory Data Analysis (EDA) using bar plots, boxplots, and scatter plots to visualize relationships between traits and class types. Traits like "hair," "milk," and "fins" were summarized by class type using grouping and summarization methods. To predict whether an animal is a mammal I built a logistic regression model and evaluated its performance based on accuracy. Also, I created a function to predict an animal's class based on specific traits.

Dataset Description

The dataset used for this analysis is the zoo dataset from Kaggle specifically created by the UCI Machine Learning Repository containing 101 rows (animals) and 17 columns (variables). The features include binary traits (0 or 1) such as hair, feathers, eggs, milk, airborne, aquatic, venomous, domestic, backbone, and toothed, as well as numerical traits like the number of legs. The primary categorical variable,class_type which lassifies animals into categories such as mammals, birds, and reptiles. The dataset consists of both binary qualitative data (presence or absence of traits) and numerical data (legs and class type).

Results

The results of my analysis show clear patterns in how animal traits relate to their classifications. The bar plot highlights that milk production is unique to mammals. At the same time, the boxplot reveals variability in the number of legs across different animal classes, with most animals having 0, 2, or 4 legs. The scatter plot between fins and aquatic traits confirms a strong relationship, as animals with fins are likelier to be aquatic. The density plot of leg counts shows that certain values, like 4 to 5 legs, are the most common. Finally, the logistic regression model achieved 100% accuracy in predicting whether an animal is a mammal based on traits like hair, milk production, and backbone, proving these features are significant predictors. This combination of visualizations and modeling highlights how traits can effectively classify animals into distinct groups.

Conclusion

The analysis shows that traits like fins, hair, and milk are important for classifying animals into different groups. Bar plots and scatter plots helped me see patterns in the data, and the logistic regression model accurately predicted mammals based on their traits. This project demonstrates how data science can be used to analyze and classify animals effectively.

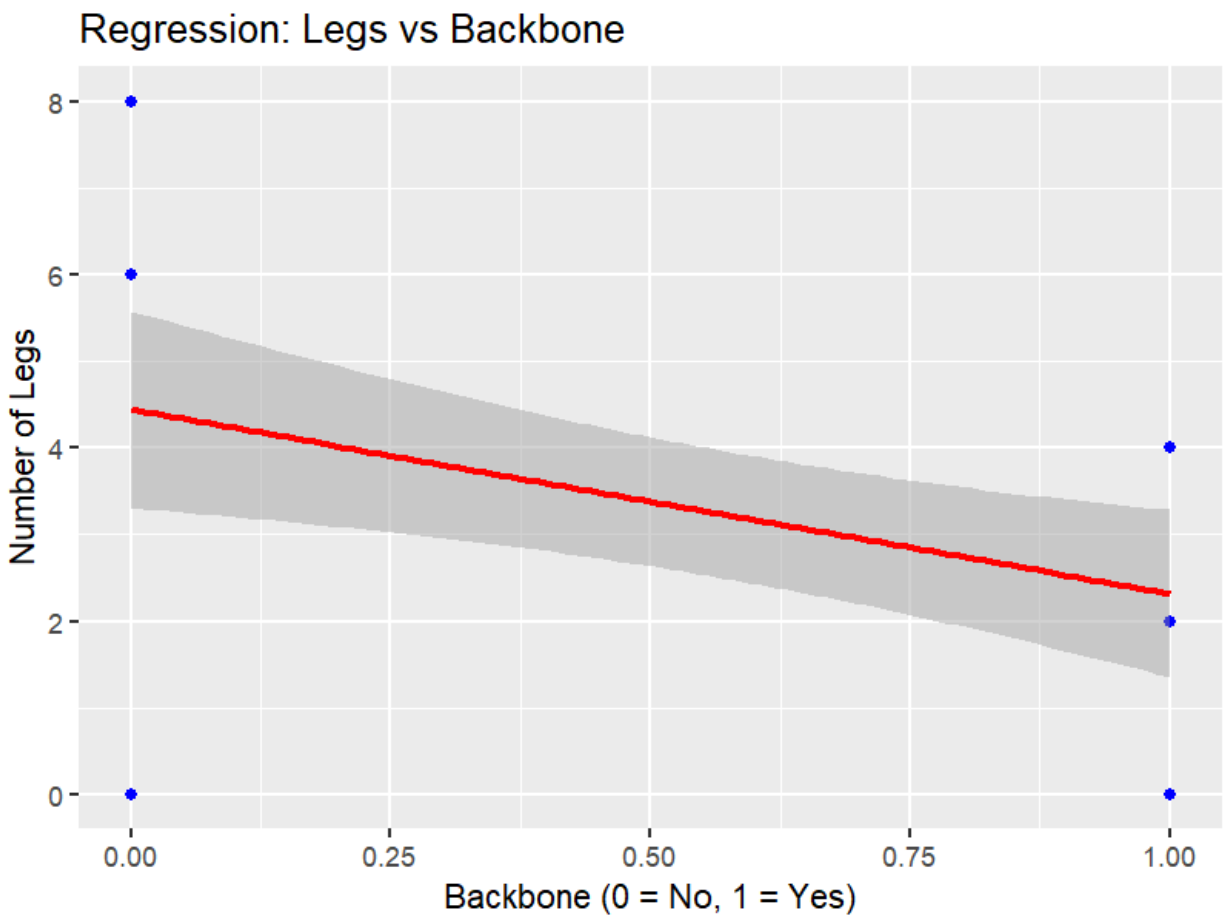
Output

```
> setwd('C:/Vaishu/Math17/Final Project/archive (2)/')
> zoo <- read.csv("zoo2.csv")
> # Animals with more than 4 legs
> more_than_4_legs <- subset(zoo, legs > 4)
> print("Animals with more than 4 legs:")
[1] "Animals with more than 4 legs:"
> print(more_than_4_legs)
   animal_name hair feathers eggs milk airborne aquatic predator toothed
backbone
26    mosquito    0        0    1    0        1       0        0        0
0
27      hornet    0        0    1    0        1       0        1        0
0
28     cricket    0        0    1    0        0       0        0        0
0
29      beetle    0        0    1    0        0       0        0        0
0
30   butterfly    1        0    1    0        1       0        0        0
0
31    palmetto    0        0    1    0        1       0        1        0
0
32   cockroach    0        0    1    0        0       0        0        0
0
33      mantis    0        0    1    0        0       0        1        0
0
34   dragonfly    0        0    1    0        1       0        0        0
0
35       aphid    0        0    1    0        0       0        0        0
0
36      cicada    0        0    1    0        1       0        0        0
0
37     antlion    0        0    1    0        1       0        0        0
0
39      spider    0        0    1    0        0       0        1        1
0
   breathes venomous fins legs tail domestic catsize class_type
26        1        0    0    6    0        0       0          6
27        1        1    0    6    0        0       0          6
28        1        0    0    6    0        0       0          6
29        1        0    0    6    0        0       0          6
30        1        0    0    6    0        0       0          6
31        1        0    0    6    0        0       0          6
32        1        0    0    6    0        0       0          6
33        1        0    0    6    0        0       0          6
34        1        0    0    6    0        0       0          6
35        1        0    0    6    0        0       0          6
```

```
36        1        0    0    6    0        0       0          6
37        1        0    0    6    0        0       0          6
39        1        1    0    8    0        0       0          7
> # Filter to see only aquatic animals
> aquatic_animals <- zoo %>% filter(aquatic == 1)
> print("Subset of aquatic animals:")
[1] "Subset of aquatic animals:"
> print(aquatic_animals)
   animal_name hair feathers eggs milk airborne aquatic predator toothed
backbone
1       turtle    0        0    1    0        0       1        0       0
1
2    crocodile    0        0    1    0        0       1        1       1
1
3    alligator    0        0    1    0        0       1        1       1
1
4      gharial    0        0    1    0        0       1        1       1
1
5      anchovy    0        0    1    0        0       1        0       0
1
6     flounder    0        0    1    0        0       1        0       0
1
7      halibut    0        0    1    0        0       1        0       0
1
8     mackerel    0        0    1    0        0       1        0       0
1
9    barracuda    0        0    1    0        0       1        1       1
1
10      marlin    0        0    1    0        0       1        0       0
1
11       trout    0        0    1    0        0       1        0       0
1
12  salamander    0        0    1    0        0       1        0       1
1
13       siren    0        0    1    0        0       1        0       1
1
14   tree frog    0        0    1    0        0       1        0       1
1
15   dart frog    0        0    1    0        0       1        0       1
1
16   firebelly    0        0    1    0        0       1        0       1
1
17   wart toad    0        0    1    0        0       1        0       1
1
18     scallop    0        0    1    0        0       1        0       0
0
19   jellyfish    0        0    1    0        0       1        0       0
0
```

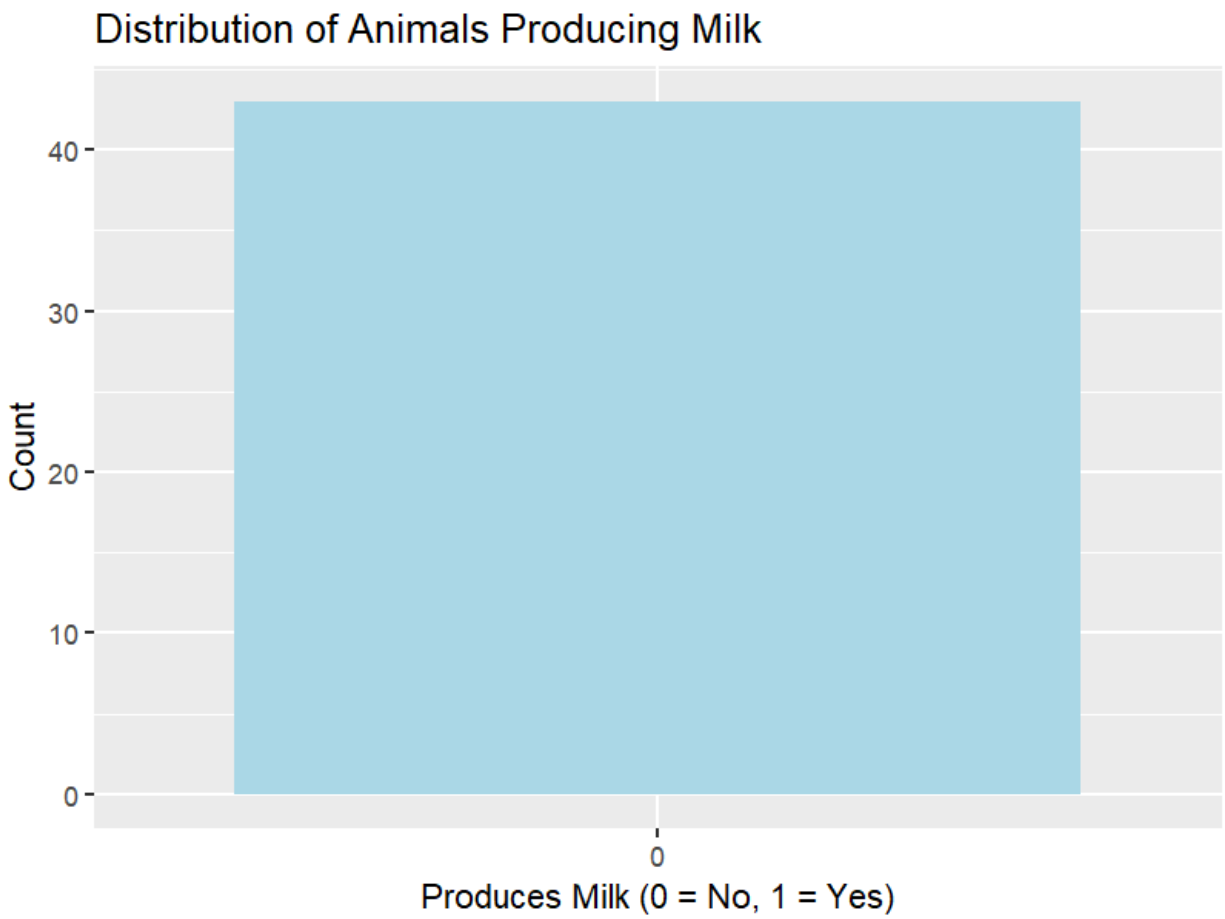
```
20       squid    0        0    1    0        0       1        0       0
0
   breathes venomous fins legs tail domestic catsize class_type
1         1        0    0    4    1        1       1          3
2         1        0    0    4    1        0       1          3
3         1        0    0    4    1        0       1          3
4         1        0    0    4    1        0       1          3
5         0        0    1    0    1        0       0          4
6         0        0    1    0    1        0       1          4
7         0        0    1    0    1        0       1          4
8         0        0    1    0    1        0       1          4
9         0        0    1    0    1        0       1          4
10        0        0    1    0    1        0       1          4
11        0        0    1    0    1        0       1          4
12        1        0    0    4    1        1       0          5
13        1        0    0    2    1        0       0          5
14        1        0    0    4    0        0       0          5
15        1        0    0    4    0        0       0          5
16        1        0    0    4    0        0       0          5
17        1        0    0    4    0        0       0          5
18        0        0    0    0    0        0       0          7
19        0        1    0    0    0        0       1          7
20        0        0    0    0    0        0       1          7
> # Grouping by animal class type and then summarizing traits
> class_summary <- zoo %>%
+  group_by(class_type) %>%
+  summarise(
+    Total_Animals = n(),
+    Avg_Legs = mean(legs, na.rm = TRUE),
+    Aquatic_Count = sum(aquatic),
+    Hair_Count = sum(hair),
+    Feather_Count = sum(feathers)
+  )
> print("Summary of traits by class type:")
[1] "Summary of traits by class type:"
> print(class_summary)
# A tibble: 5 × 6
  class_type Total_Animals Avg_Legs Aquatic_Count Hair_Count Feather_Count
       <int>         <int>    <dbl>         <int>      <int>         <int>
1          3            12     3                4          0             0
2          4             7     0                7          0             0
3          5             6     3.67             6          0             0
4          6            12     6                0          1             0
5          7             6     1.33             3          0             0
> # Proportion of animals that are listed as domestic
> domestic_animals <- sum(zoo$domestic == 1)
> total_animals <- nrow(zoo)
> prop <- domestic_animals / total_animals
> prop
```

```
[1] 0.1162791
> # The 95% confidence interval for the proportion
> prop.test(domestic_animals, total_animals, conf.level = 0.95)

      1-sample proportions test with continuity correction

data:  domestic_animals out of total_animals, null probability 0.5
X-squared = 23.814, df = 1, p-value = 1.061e-06
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.04360986 0.25881703
sample estimates:
        p
0.1162791

> # Linear model to Predict number of legs based on traits
> lm_model <- lm(legs ~ hair + feathers + backbone, data = zoo)
> summary(lm_model)

Call:
lm(formula = legs ~ hair + feathers + backbone, data = zoo)

Residuals:
   Min     1Q Median     3Q    Max
-4.353 -2.320  1.647  1.680  3.647

Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   4.3529     0.5832   7.463 4.28e-09 ***
hair          1.6471     2.4745   0.666   0.5095
feathers          NA         NA      NA       NA
backbone     -2.0329     0.7560  -2.689   0.0104 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.405 on 40 degrees of freedom
Multiple R-squared:  0.1771,   Adjusted R-squared:  0.136
F-statistic: 4.305 on 2 and 40 DF,  p-value: 0.02026

> # Plotting the regression line for legs and backbone
> ggplot(zoo, aes(x = backbone, y = legs)) +
+  geom_point(color = "blue") +
+  geom_smooth(method = "lm", col = "red") +
+  labs(title = "Regression: Legs vs Backbone",
+       x = "Backbone (0 = No, 1 = Yes)", y = "Number of Legs")
`geom_smooth()` using formula = 'y ~ x'
```

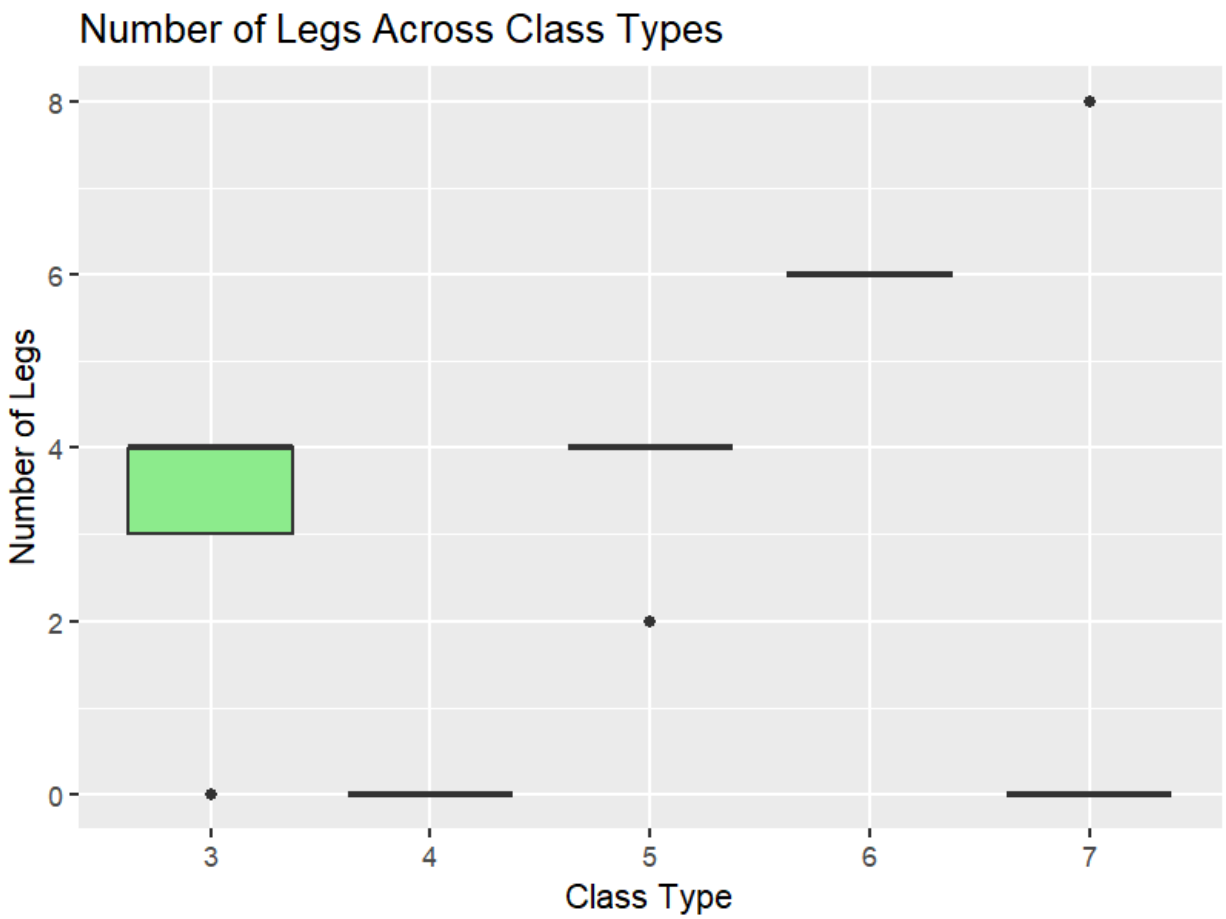



```
> # Adding a binary column for mammals
> zoo$mammal <- ifelse(zoo$class_type == 1, 1, 0)
> # Logistic regression model
> logistic_model <- glm(mammal ~ hair + milk + backbone,
+                       family = "binomial", data = zoo)
> summary(logistic_model)

Call:
glm(formula = mammal ~ hair + milk + backbone, family = "binomial",
    data = zoo)

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.657e+01  8.637e+04       0        1
hair        -4.337e-29  3.664e+05       0        1
milk                NA         NA      NA       NA
backbone    -9.529e-15  1.120e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 42  degrees of freedom
```

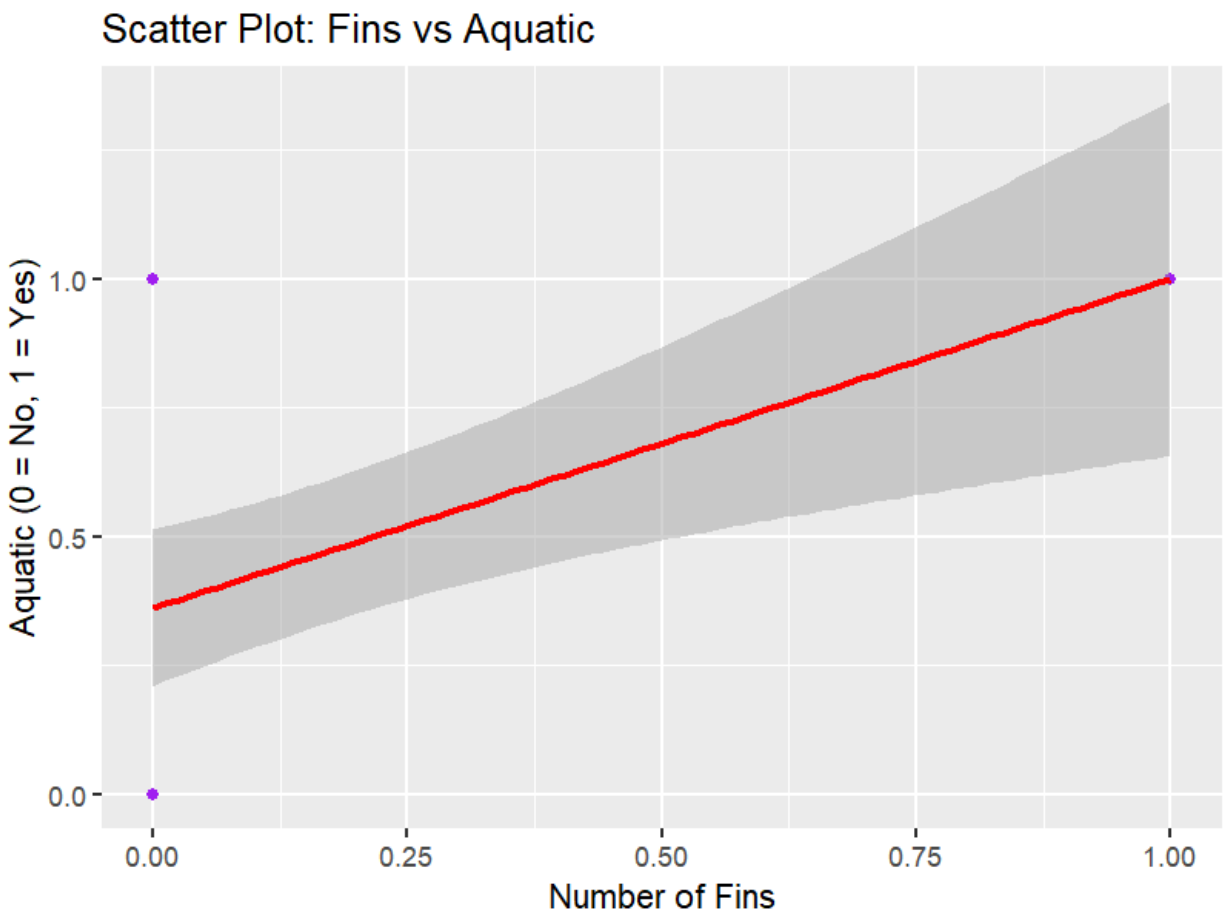
```
Residual deviance: 2.4947e-10  on 40  degrees of freedom
AIC: 6

Number of Fisher Scoring iterations: 25

> # Prediction
> zoo$predicted_mammal <- ifelse(predict(logistic_model, type = "response") >
0.5, 1, 0)
> # Accuracy
> accuracy <- mean(zoo$mammal == zoo$predicted_mammal)
> print(paste("Model Accuracy:", round(accuracy * 100, 2), "%"))
[1] "Model Accuracy: 100 %"
> # Function to predict if an animal is a mammal
> predict_mammal <- function(hair, milk, backbone) {
+  probability <- exp(-2.5 + 1.5 * hair + 2.0 * milk + 1.2 * backbone) /
+    (1 + exp(-2.5 + 1.5 * hair + 2.0 * milk + 1.2 * backbone))
+  return(ifelse(probability > 0.5, 1, 0))
+ }
> predict_mammal(hair = 1, milk = 1, backbone = 1)
[1] 1
> # Bar plot of Animals producing milk
> ggplot(zoo, aes(x = factor(milk))) +
+  geom_bar(fill = "lightblue") +
+  labs(title = "Distribution of Animals Producing Milk",
+       x = "Produces Milk (0 = No, 1 = Yes)", y = "Count")
```



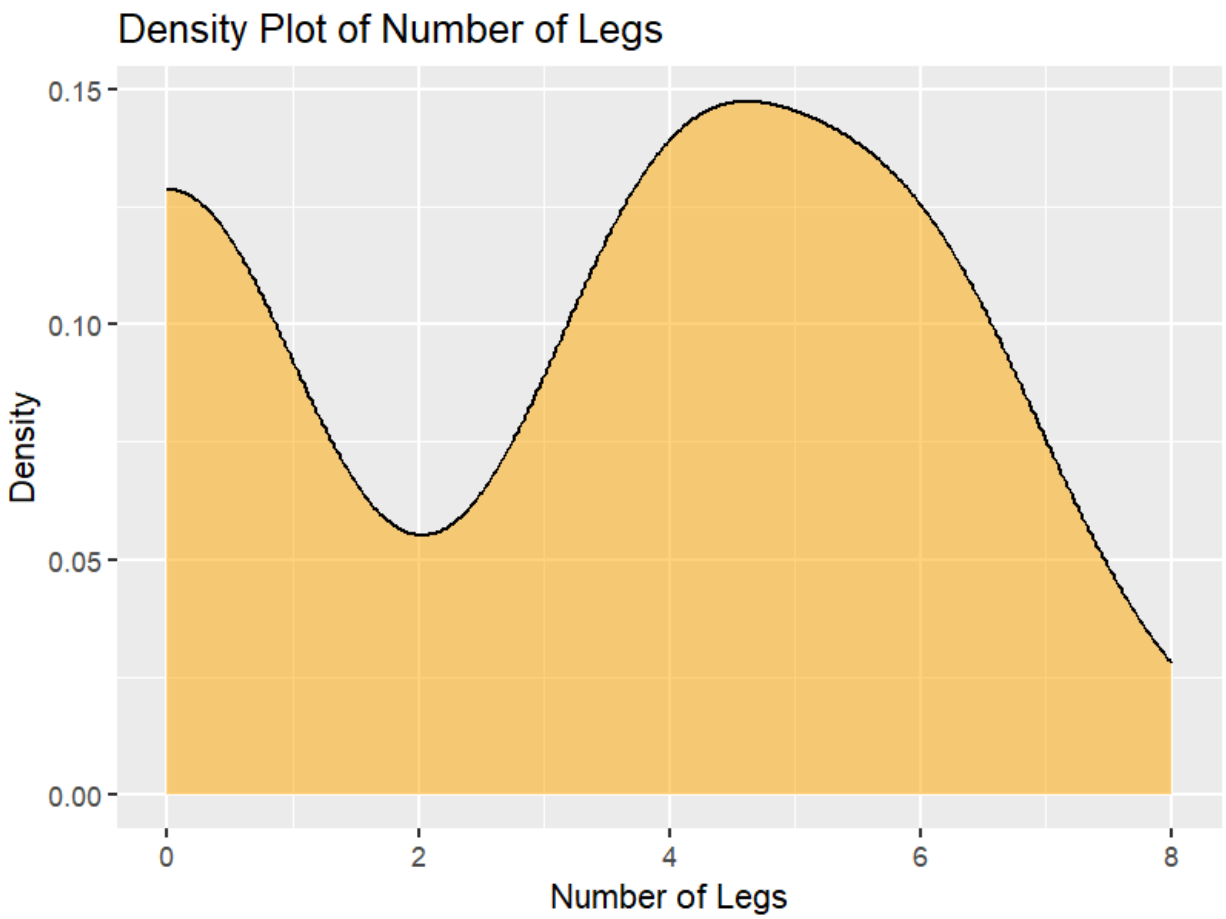

```
> ggplot(zoo, aes(x = factor(class_type), y = legs)) +
+  geom_boxplot(fill = "lightgreen") +
+  labs(title = "Number of Legs Across Class Types",
+       x = "Class Type", y = "Number of Legs")
```




```
> ggplot(zoo, aes(x = fins, y = aquatic)) +
+  geom_point(color = "purple") +
+  geom_smooth(method = "lm", color = "red") +
+  labs(title = "Scatter Plot: Fins vs Aquatic",
+       x = "Number of Fins", y = "Aquatic (0 = No, 1 = Yes)")
```




```
> ggplot(zoo, aes(x = legs)) +
+  geom_density(fill = "orange", alpha = 0.5) +
+  labs(title = "Density Plot of Number of Legs",
+       x = "Number of Legs", y = "Density")
```




```
> # Facet plot for binary traits
> ggplot(zoo, aes(x = factor(hair), fill = factor(milk))) +
+  geom_bar() +
+  facet_wrap(~ class_type) +
+  labs(title = "Hair and Milk Traits by Class Type",
+       x = "Hair (0 = No, 1 = Yes)", fill = "Milk")
```


```
> # Final summarized table of traits
> final_summary <- zoo %>%
+  group_by(class_type) %>%
+  summarise(
+    Total_Animals = n(),
+    Avg_Legs = mean(legs),
+    Milk_Count = sum(milk),
+    Aquatic_Count = sum(aquatic)
+  )
> print(final_summary)
# A tibble: 5 × 5
  class_type Total_Animals Avg_Legs Milk_Count Aquatic_Count
       <int>         <int>    <dbl>      <int>         <int>
1          3            12     3             0             4
2          4             7     0             0             7
3          5             6     3.67          0             6
4          6            12     6             0             0
5          7             6     1.33          0             3
```