STSCI5954

2024-10-28

Practice Performing Exploratory Data Analysis for Classification

Load library

```
library(tidyverse) # Load the tidyverse
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.4     ✓ readr     2.1.5
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.3     ✓ tidyr     1.3.1
## ✓ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✗ dplyr::filter() masks stats::filter()
## ✗ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
e errors
```

Load the data

```
birds <- read_csv("EasternBluebird.csv") # Load data set as birds
```

```
## Rows: 64724 Columns: 23
## ── Column specification ────────────────────────────────────────────────────────
## Delimiter: ","
## dbl (23): LATITUDE, LONGITUDE, ELEV, Shallow_Ocean, CoastShore_lines, Shallo...
##
## ℹ Use `spec()` to retrieve the full column specification for this data.
## ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

Basic review

```
# number of observation
birds %>% count()
```

```
## # A tibble: 1 × 1
##       n
##   <int>
## 1 64724
```

```
# what does each row mean? OR, what is the unit of inference?

# How many sightings in the dataset?
birds %>%count(y,sort=TRUE, na.miss=TRUE)
```

```
## # A tibble: 2 × 3
##       y na.miss     n
##   <dbl> <lgl>   <int>
## 1     0 TRUE    59938
## 2     1 TRUE     4786
```

```
# How many variables?
str(birds)
```

```
## spc_tbl_ [64,724 × 23] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ LATITUDE        : num [1:64724] 35.3 36 36.7 37 37.3 ...
##  $ LONGITUDE       : num [1:64724] -76.6 -78.9 -81.5 -79.5 -80.5 ...
##  $ ELEV            : num [1:64724] 2.24 100.92 939.3 212.17 773.58 ...
##  $ Shallow_Ocean   : num [1:64724] 0 0 0 0 0 ...
##  $ CoastShore_lines: num [1:64724] 0 0 0 0 0 ...
##  $ Shallow_Inland  : num [1:64724] 0 0 0 0 0 ...
##  $ Deep_Inland     : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Moderate_Ocean  : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Deep_Ocean      : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Evergreen_needle: num [1:64724] 40.82 0 0 2.04 0 ...
##  $ Grasslands      : num [1:64724] 2.04 0 0 0 0 ...
##  $ Croplands       : num [1:64724] 0 0 0 0 0 ...
##  $ Urban_Built     : num [1:64724] 0 63.9 0 0 0 ...
##  $ Barren          : num [1:64724] 0 0 0 0 0 ...
##  $ Evergreen_broad : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Deciduous_needle: num [1:64724] 0 0 0 0 0 ...
##  $ Deciduous_broad : num [1:64724] 0 0 100 10.2 100 ...
##  $ Mixed_forest    : num [1:64724] 51 11.1 0 85.7 0 ...
##  $ Closed_shrubland: num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Open_shrubland  : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Woody_savannas  : num [1:64724] 4.08 25 0 2.04 0 ...
##  $ Savannas        : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  $ y               : num [1:64724] 0 0 0 0 0 0 0 0 0 0 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   LATITUDE = col_double(),
##   ..   LONGITUDE = col_double(),
##   ..   ELEV = col_double(),
##   ..   Shallow_Ocean = col_double(),
##   ..   CoastShore_lines = col_double(),
##   ..   Shallow_Inland = col_double(),
##   ..   Deep_Inland = col_double(),
##   ..   Moderate_Ocean = col_double(),
##   ..   Deep_Ocean = col_double(),
##   ..   Evergreen_needle = col_double(),
##   ..   Grasslands = col_double(),
##   ..   Croplands = col_double(),
##   ..   Urban_Built = col_double(),
##   ..   Barren = col_double(),
##   ..   Evergreen_broad = col_double(),
##   ..   Deciduous_needle = col_double(),
##   ..   Deciduous_broad = col_double(),
##   ..   Mixed_forest = col_double(),
##   ..   Closed_shrubland = col_double(),
##   ..   Open_shrubland = col_double(),
##   ..   Woody_savannas = col_double(),
##   ..   Savannas = col_double(),
##   ..   y = col_double()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
# list the variable names
names(birds)
```

```
##  [1] "LATITUDE"         "LONGITUDE"        "ELEV"             "Shallow_Ocean"
##  [5] "CoastShore_lines" "Shallow_Inland"   "Deep_Inland"      "Moderate_Ocean"
##  [9] "Deep_Ocean"       "Evergreen_needle" "Grasslands"       "Croplands"
## [13] "Urban_Built"      "Barren"           "Evergreen_broad"  "Deciduous_needle"
## [17] "Deciduous_broad"  "Mixed_forest"     "Closed_shrubland" "Open_shrubland"
## [21] "Woody_savannas"   "Savannas"         "y"
```

```
# Unit for variable ELEV? Confirm with client
```

Exploratory data analysis to better understand the relationships between our response variable and our predictors. For each predictor, we’ll want to assess whether it’s related to the response variable, and if so, how?

What could we be missing if look at only paired associations - come back to this later

```
# As the amount of croplands increases does it become more likely to see an eastern bluebird are
less likely?

# Association between y (factor or binary) and croplands (continuous numeric)
# make boxplots showing the distribution of croplands according to presence and absence.
```

Define plot parameters and colors

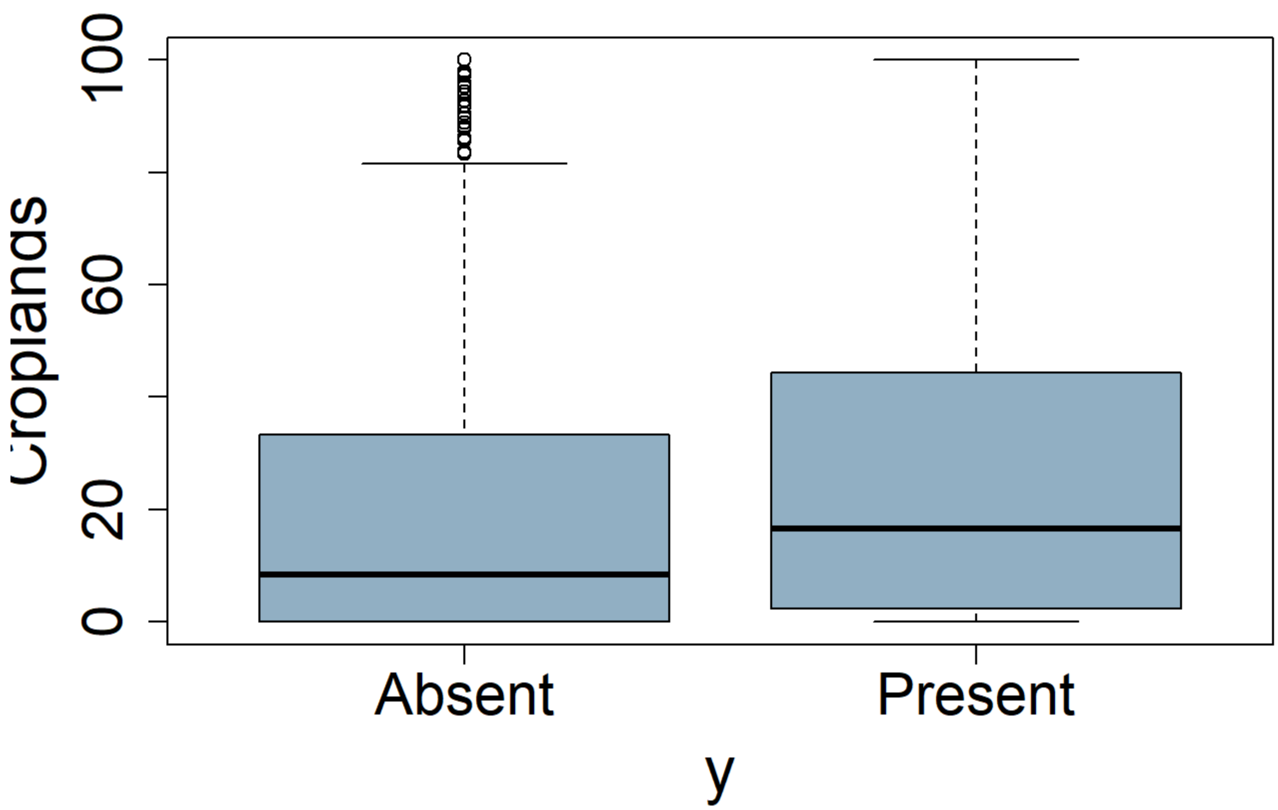
```
par(mar =c(5,5,4,4))
# eCornell Hex Codes:
crimson = '#b31b1b'   # crimson
lightGray = '#cecece' # lightGray
darkGray = '#606366'  # darkGray
skyBlue = '#92b2c4'   # skyblue
gold = '#fbb040'      # gold
ecBlack = '#393f47'   # ecBlack
```

Convert y to a factor

```
birds <- birds %>%
  mutate(y = factor(y, levels = c(0, 1), labels = c("Absent", "Present")))
```

Examine association between y and one predictor variable with a boxplot

```
boxplot(Croplands ~ y, # Plot y on the x-axis to examine the
                       # distribution of the predictor variable
                       # Croplands across different categories of y.
  birds,               # Use data from birds data set
  cex.axis = 1.8,      # Adjust size of axes
  cex.lab = 2,         # Adjust size of tick marks
  col = skyBlue)       # Shade boxes skyBlue
```




```
# We see that the amount of cropland varies both in locations where the birds were found, and in
locations where the birds were not found, but the median amount of cropland is higher in areas w
here the birds were present.
# Calculate the medians
```

Examine association between y and one predictor variable by calculating summary statistics

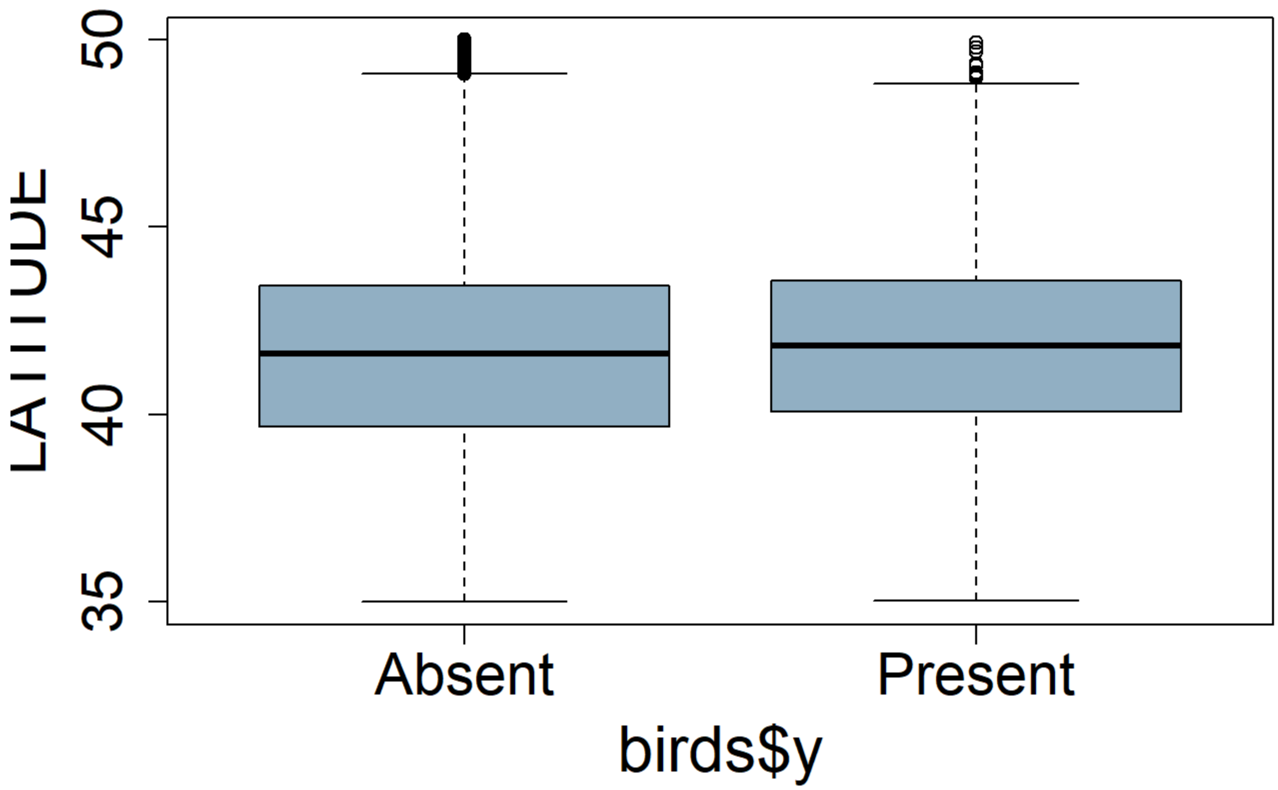
```
birds %>%
group_by(y) %>%
summarise(MedianCroplands = median(Croplands))
```

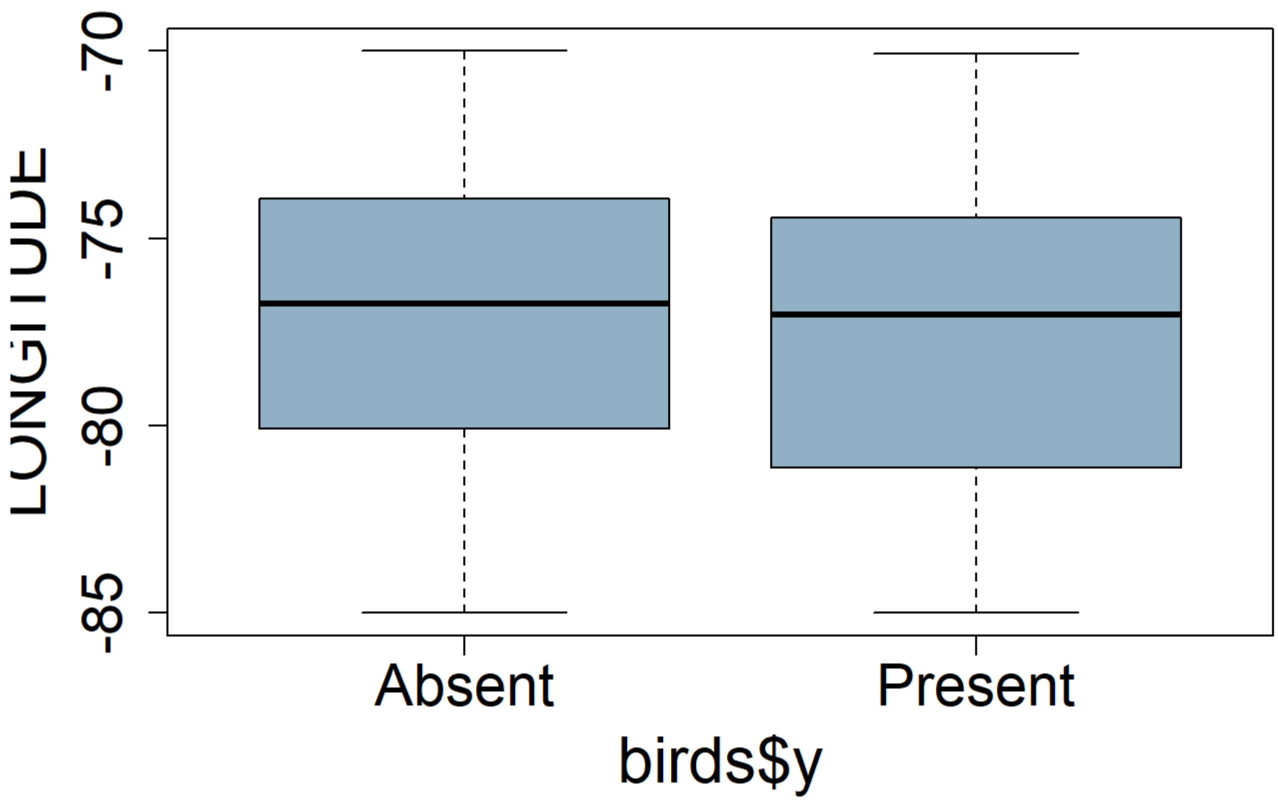
```
## # A tibble: 2 × 2
##   y       MedianCroplands
##   <fct>             <dbl>
## 1 Absent             8.33
## 2 Present           16.7
```

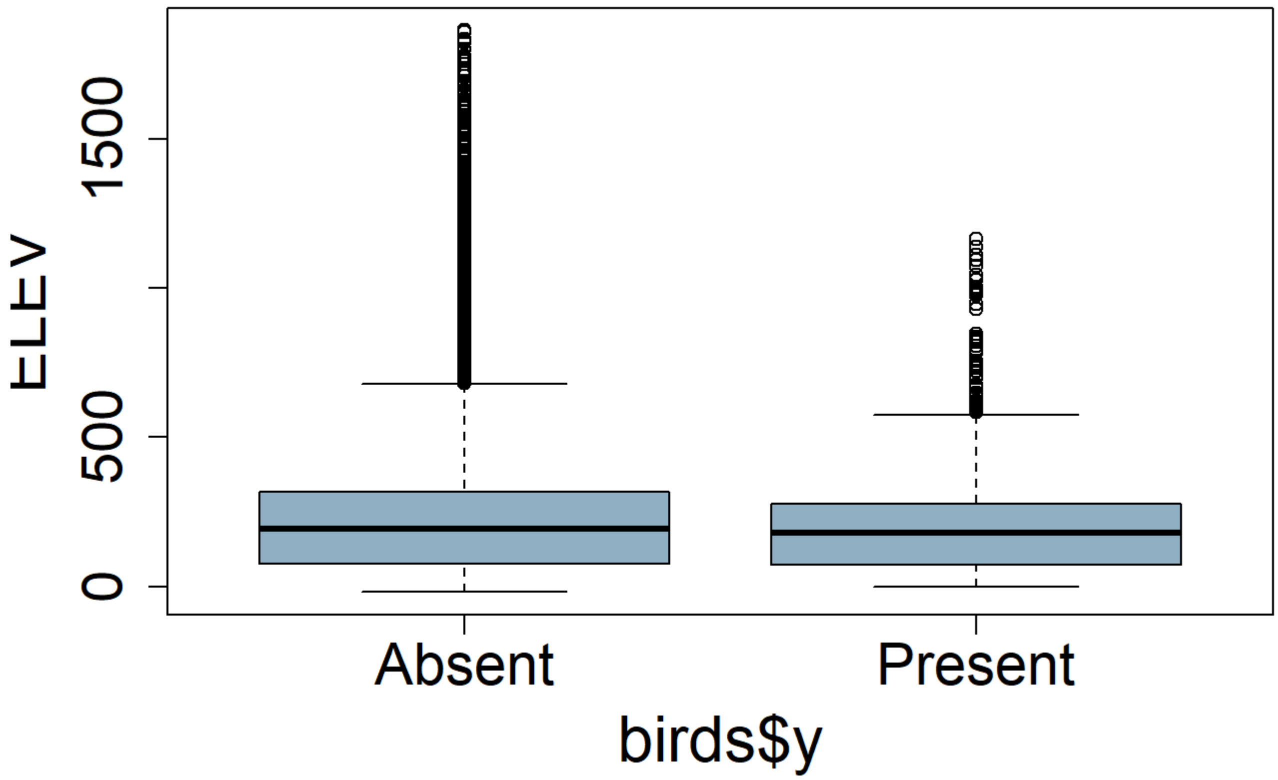
```
# We seen that the median amount of cropland is two times higher in areas where the birds are fo
und, which suggests that croplands may be an important variable and helping us predict presence/
absence.

# Do we need statistical test? Can we test if the medians of two samples are equal or not?
```

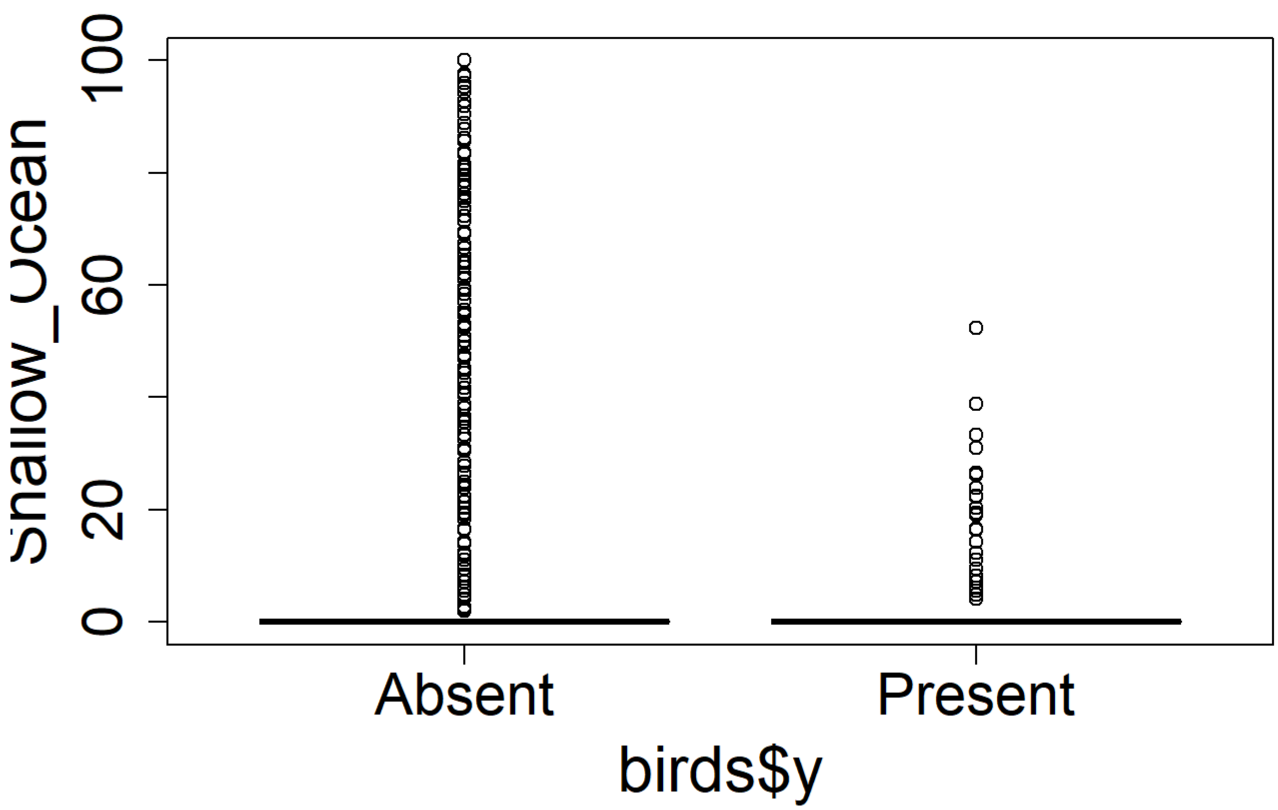
Examine all predictors with a for loop

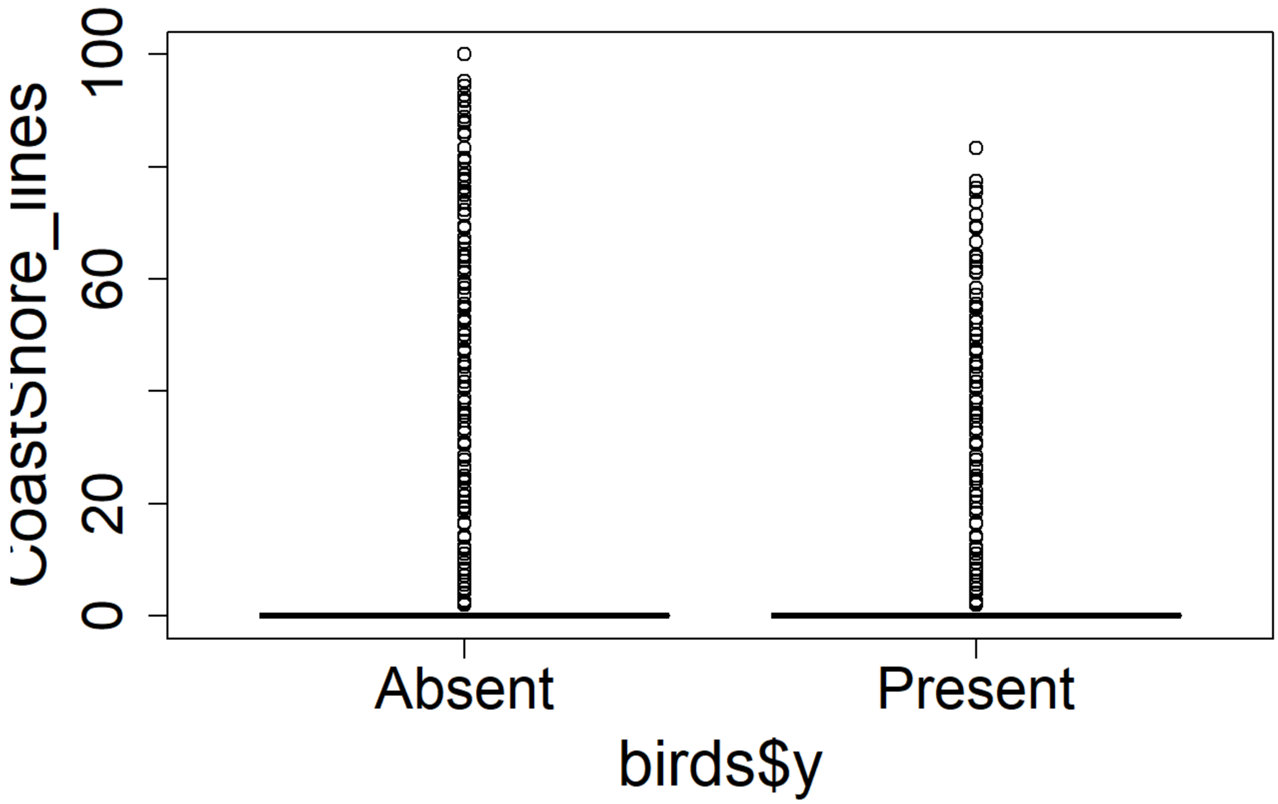
```
for(i in 1:22) { # Loop over columns that contain predictors

  # Extract the values of the ith predictor and save them
  # in the vector predictor_vals:
  predictor_vals <- birds %>% pull(i)

  # Make boxplot showing distribution of predictor_values,
  # broken down by value of y ("Present" or "Absent")
  # Add y-axis label to specify which predictor we're looking at:
  boxplot(predictor_vals ~ birds$y,
          ylab = colnames(birds)[i],
          cex.axis = 1.8,
          cex.lab = 2,
          col = skyBlue)
}
```

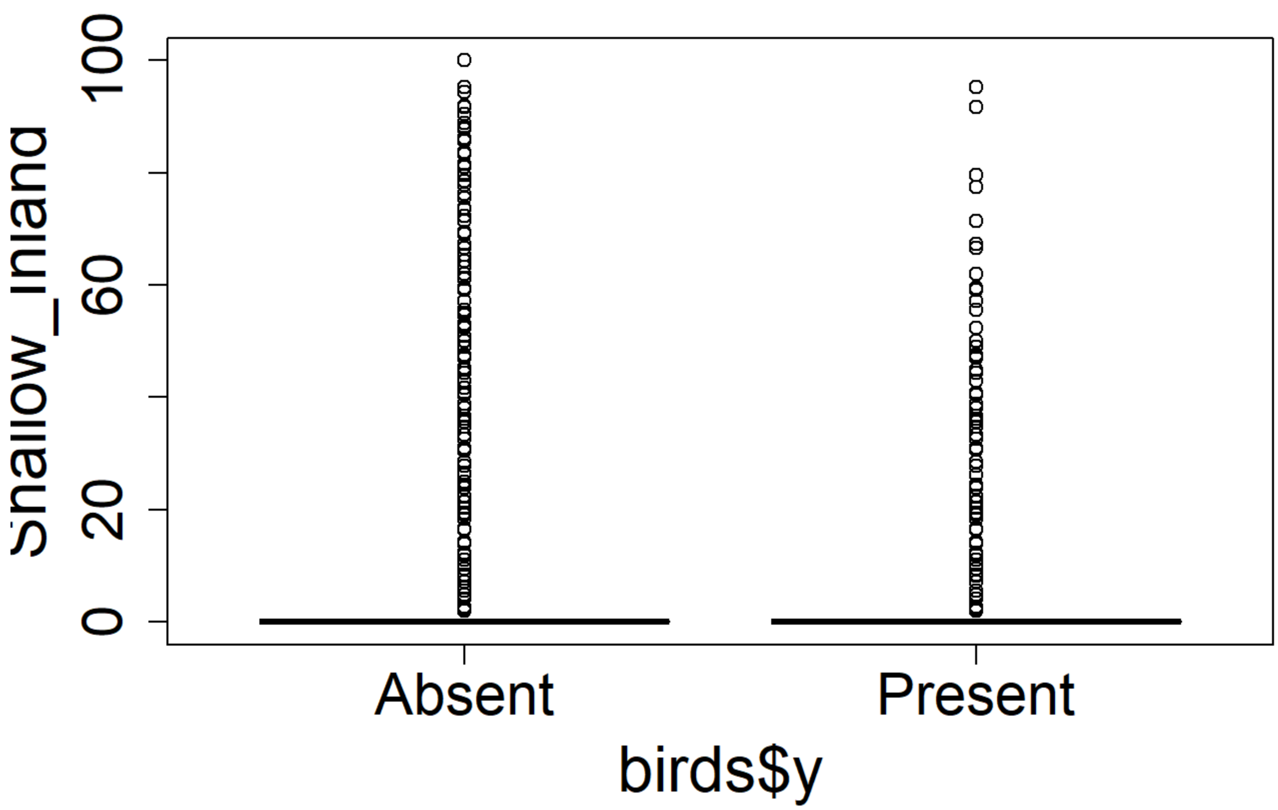

50
45
40
35
Absent
Present
birds$y

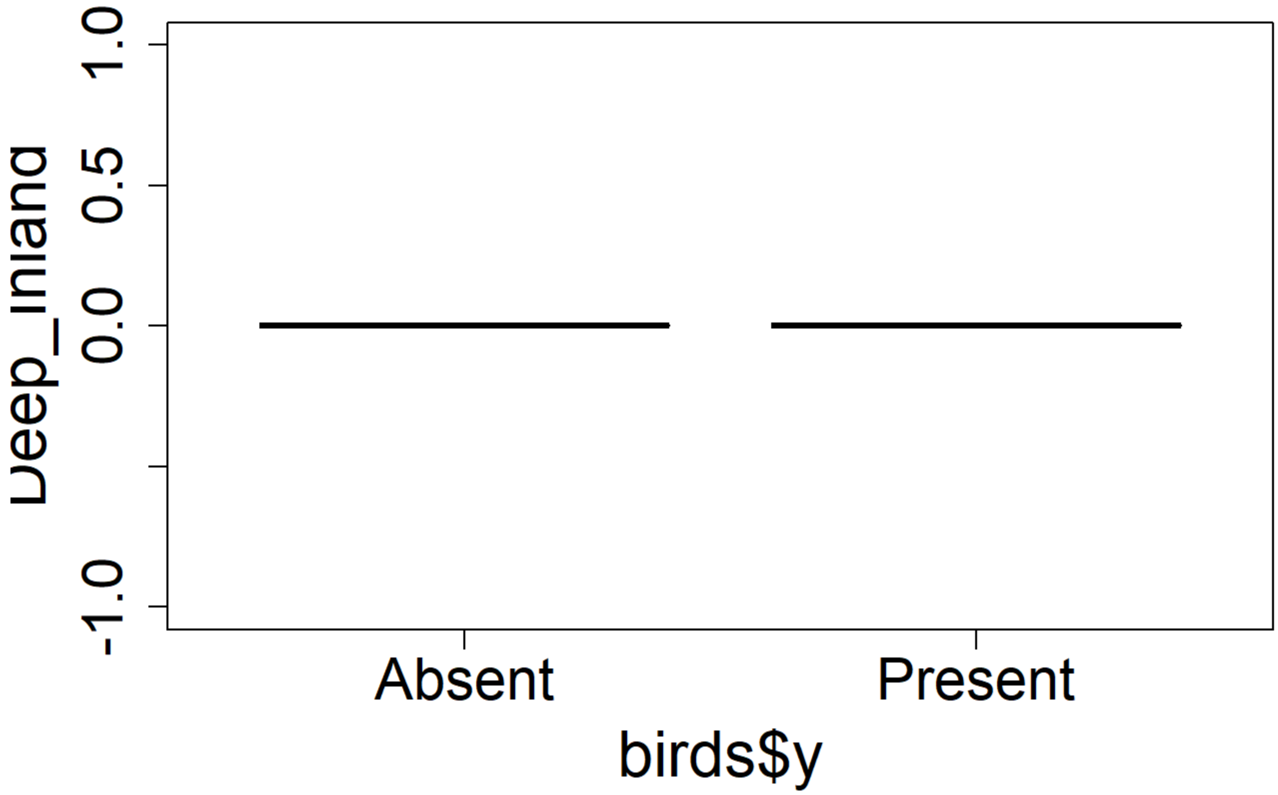

-70
-75
-80
-85
Absent
Present
birds$y


ELEV
0
500
1500
Absent
Present
birds$y

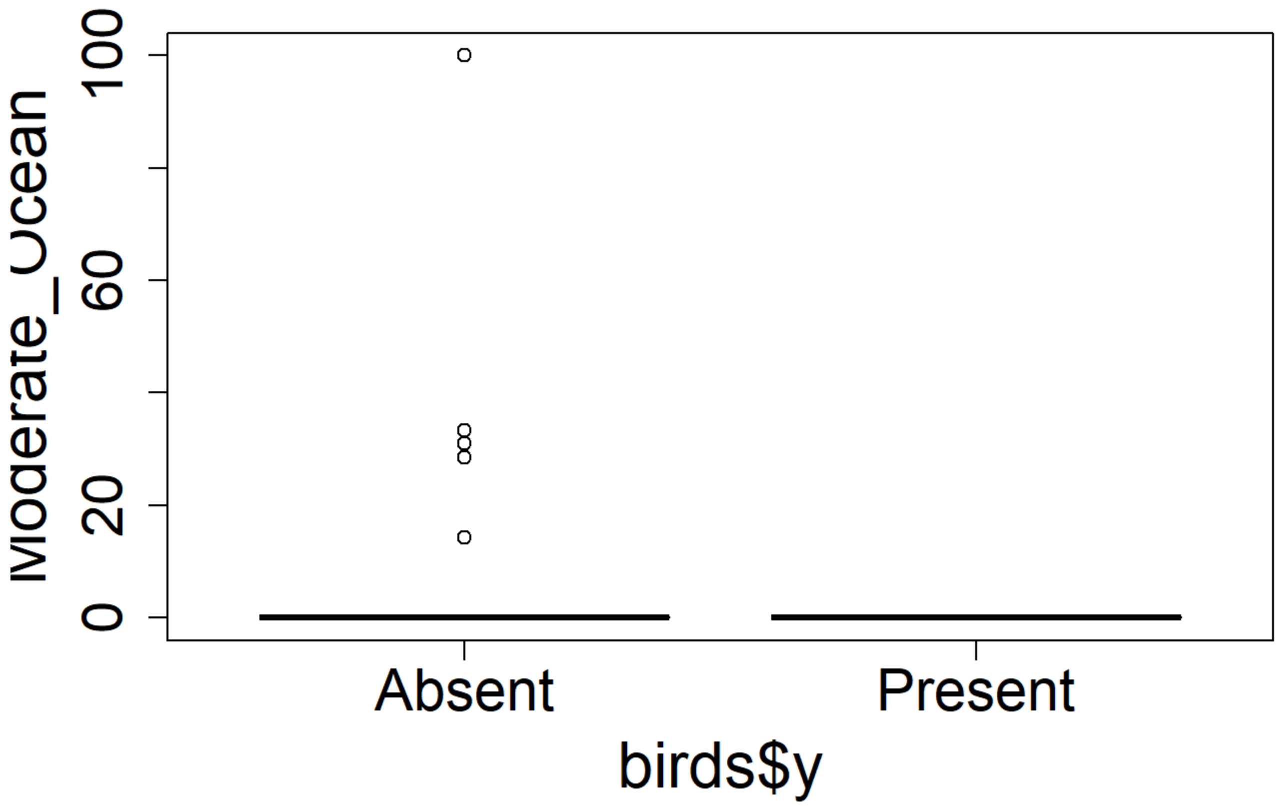

Shallow_Ocean
0
20
60
100
Absent
Present
birds$y


0
20
60
100
Absent
Present
birds$y

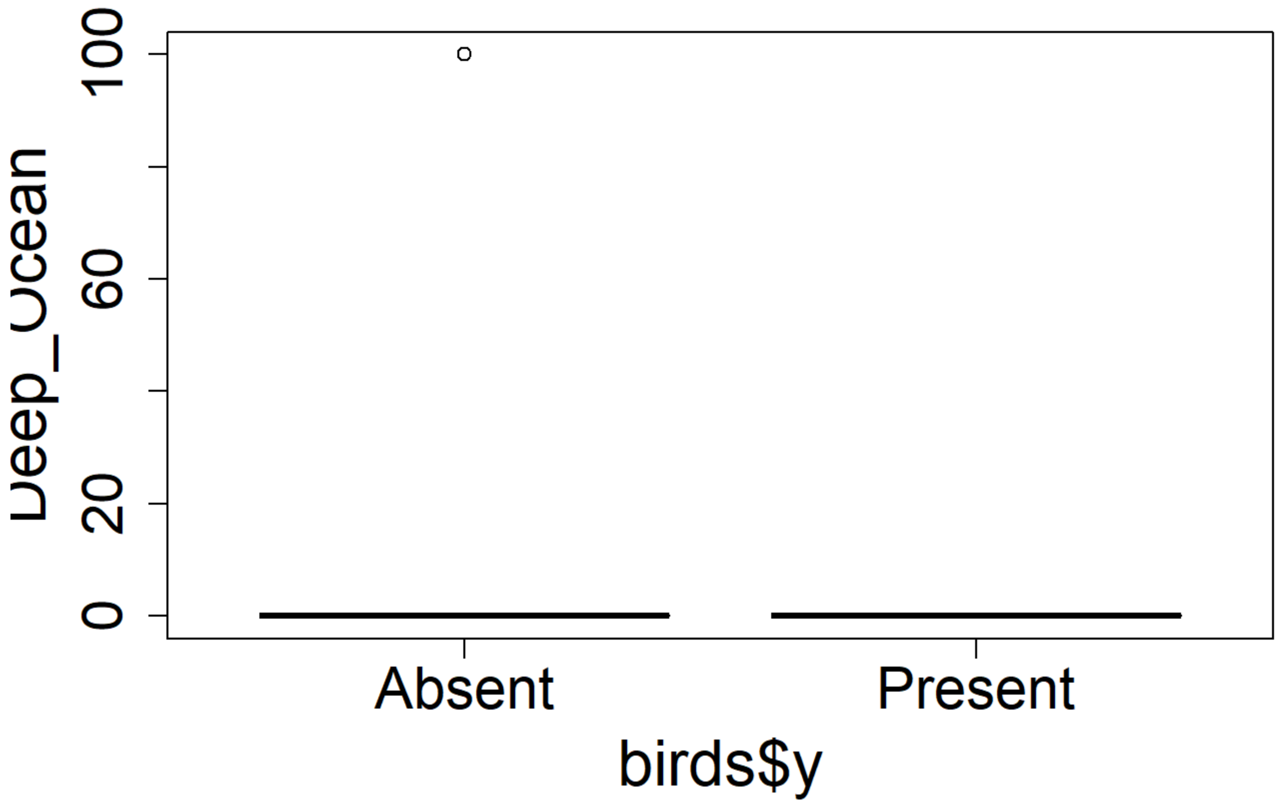

0
20
60
100
Absent
Present
birds$y

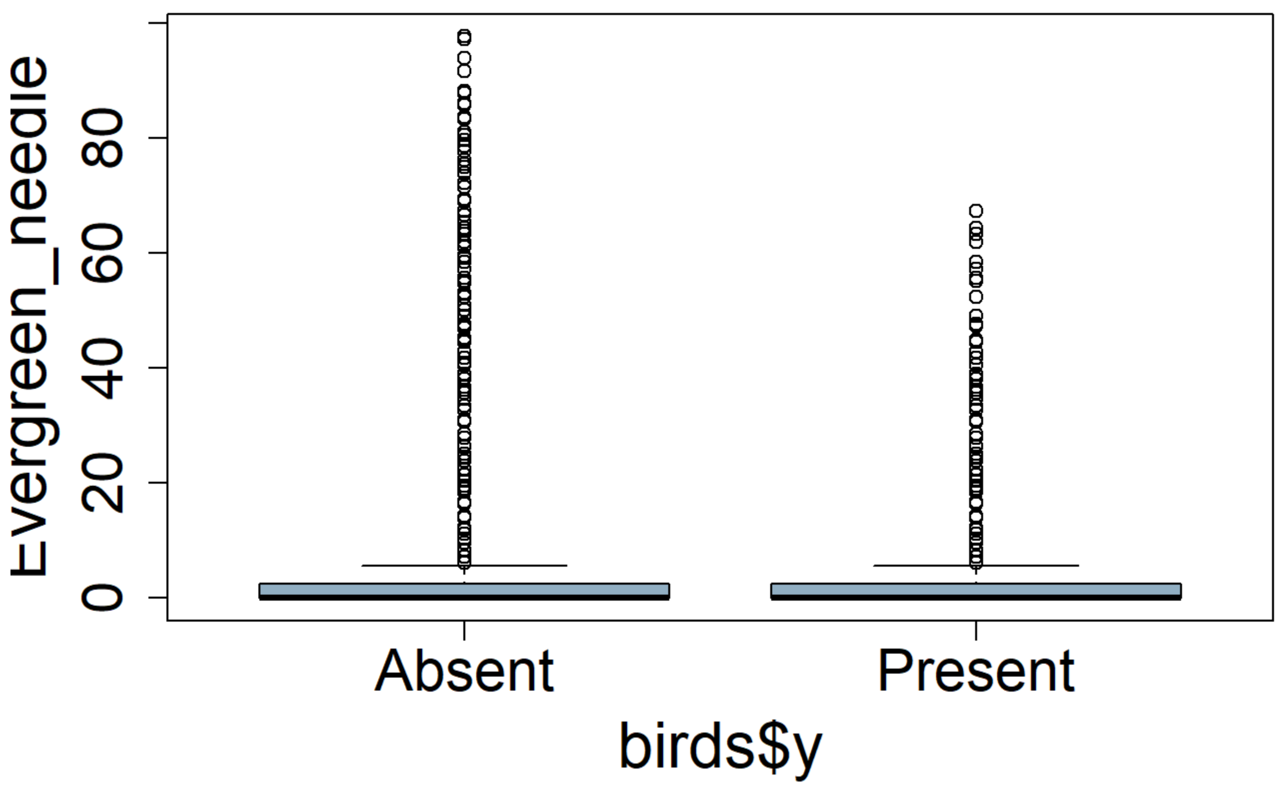


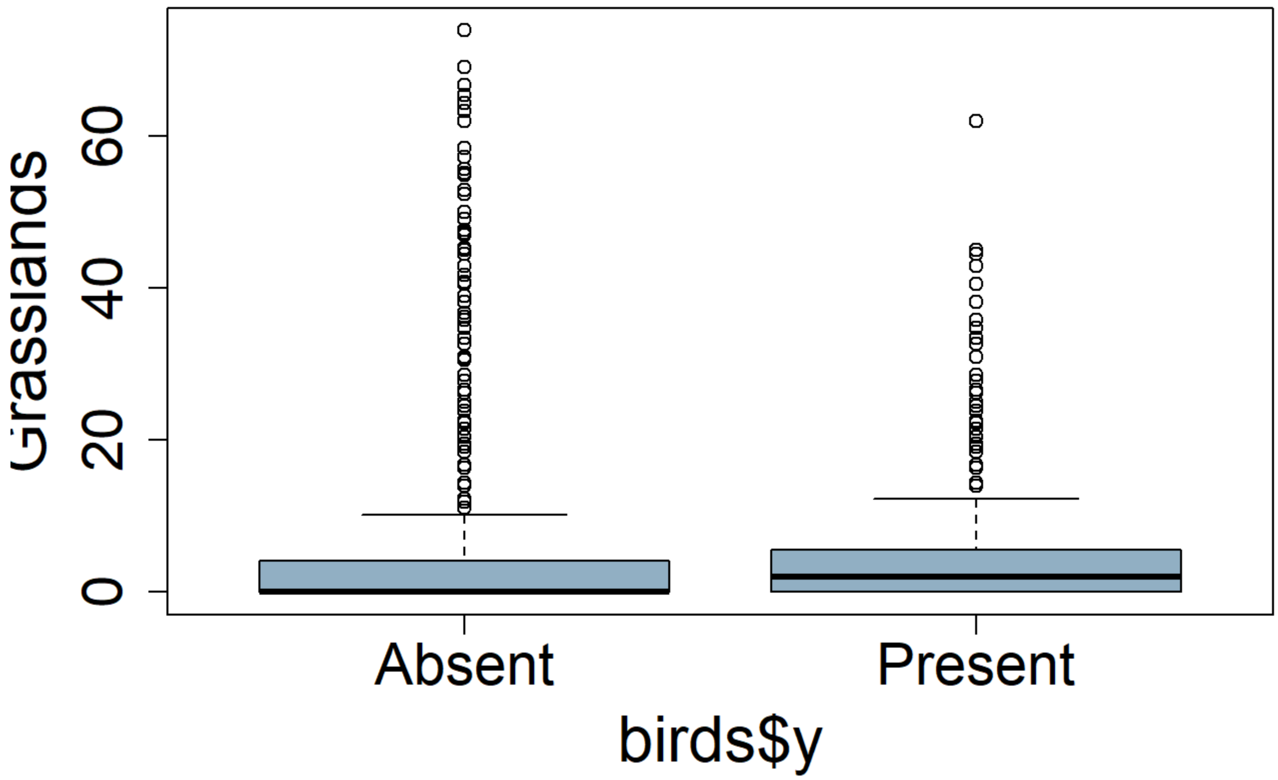
Deep_Inland
1.0
0.5
0.0
-1.0
Absent
Present
birds$y

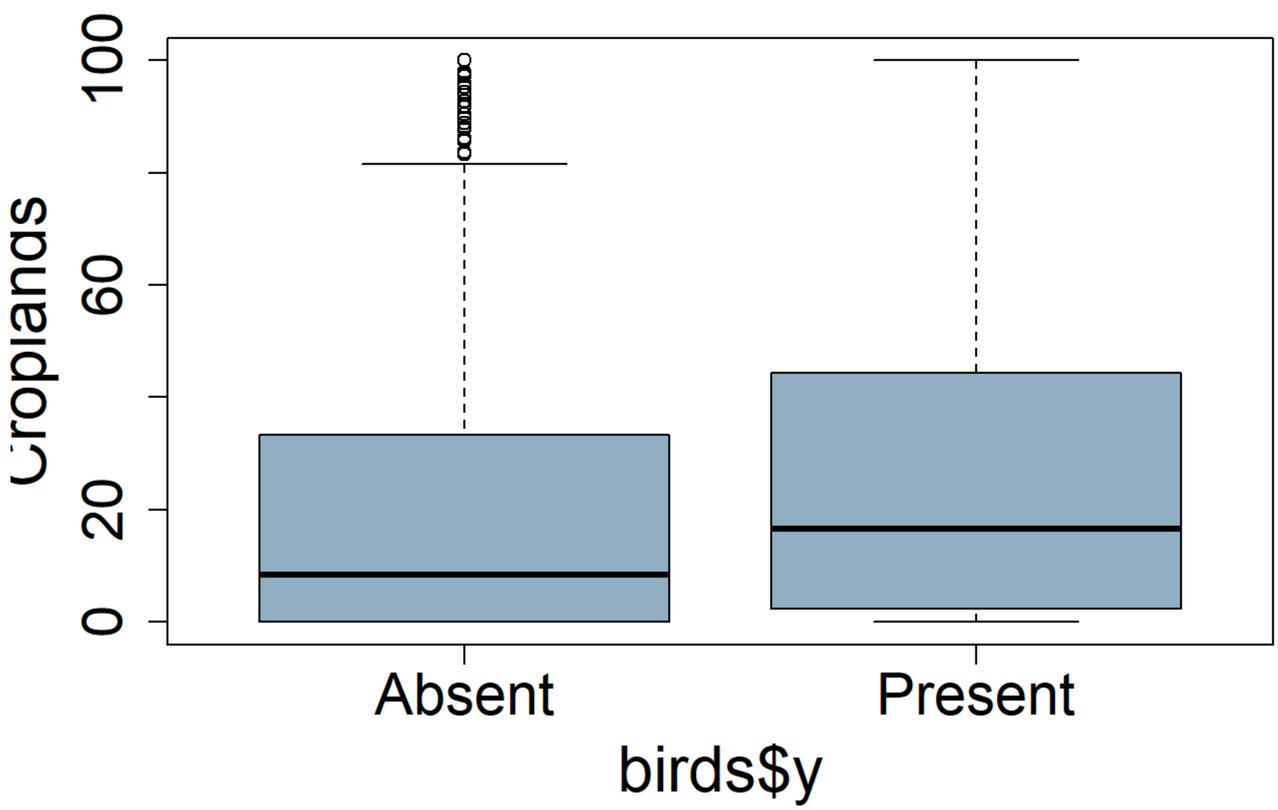


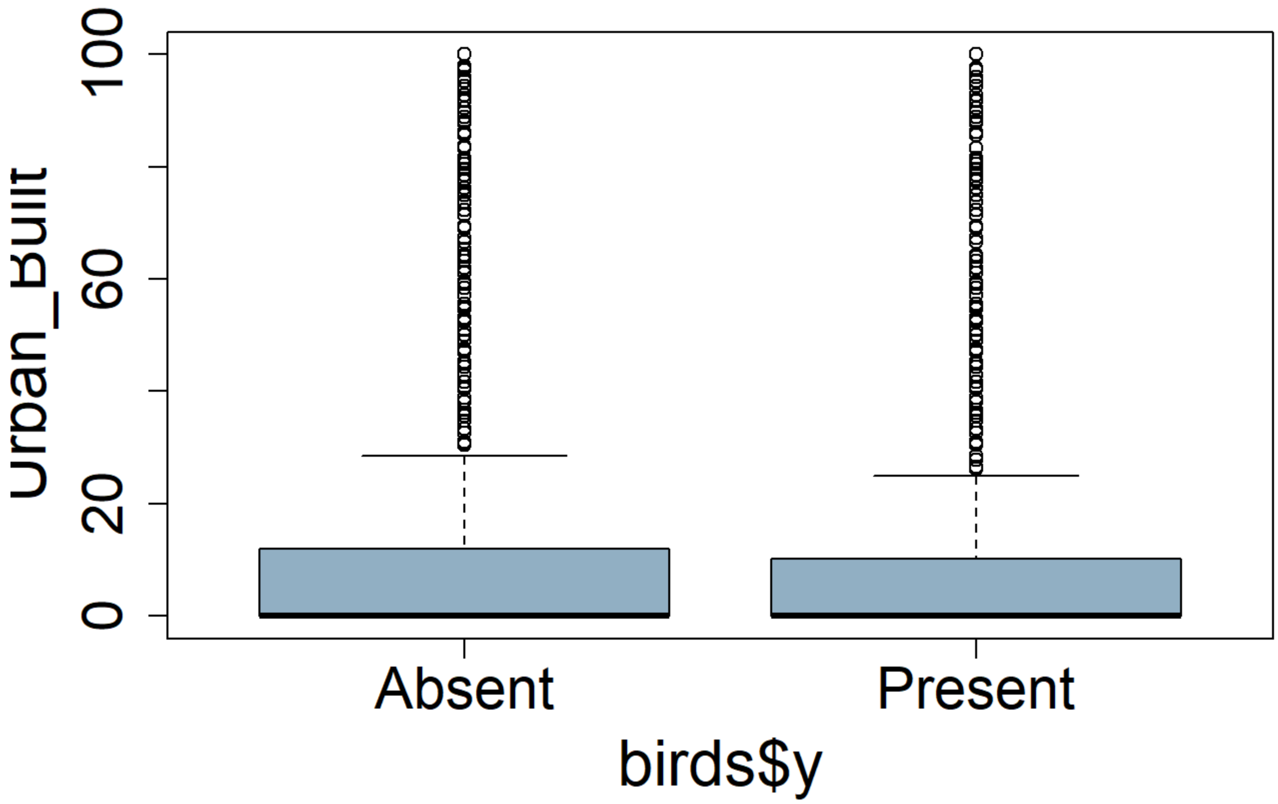
Moderate_Ocean
100
60
20
0
Absent
Present
birds$y

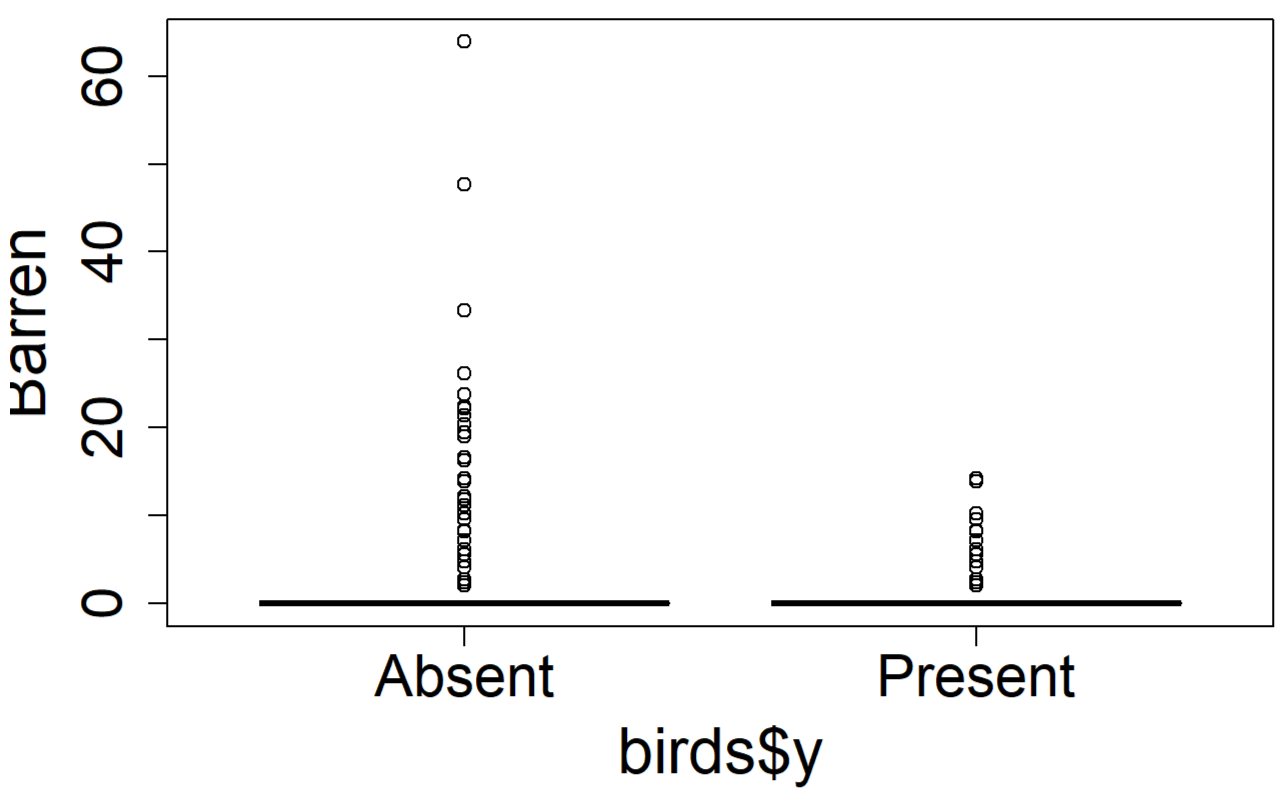

Deep_Ocean
0
20
60
100
Absent
Present
birds$y

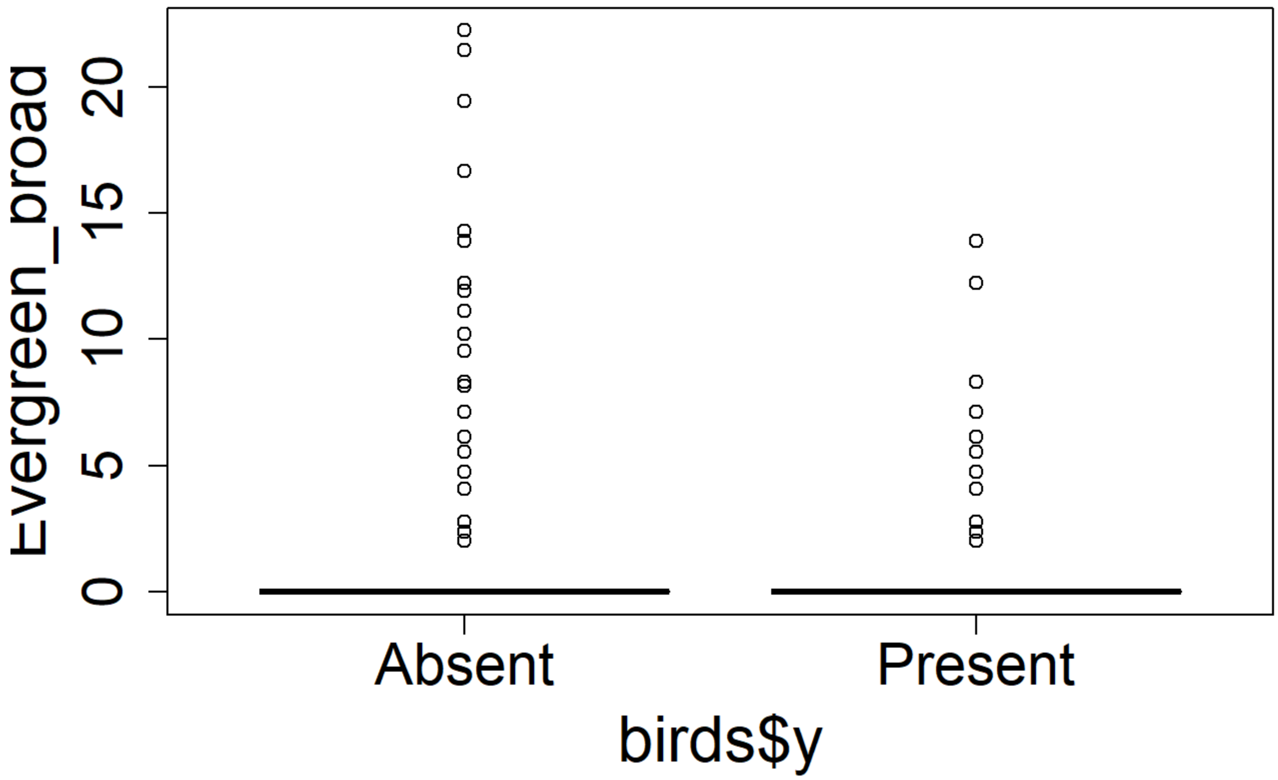

Evergreen_needle
0
20
40
60
80
Absent
Present
birds$y

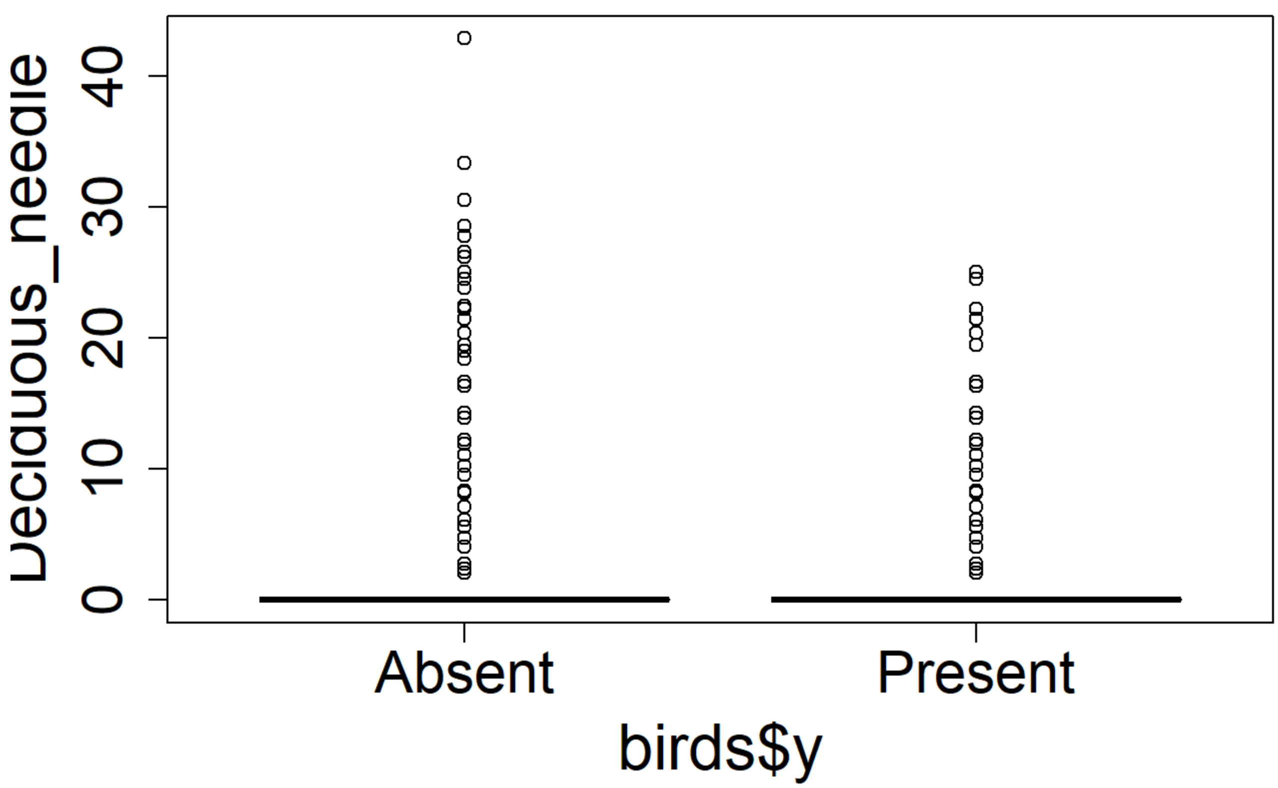

0
20
40
60
Absent
Present
birds$y

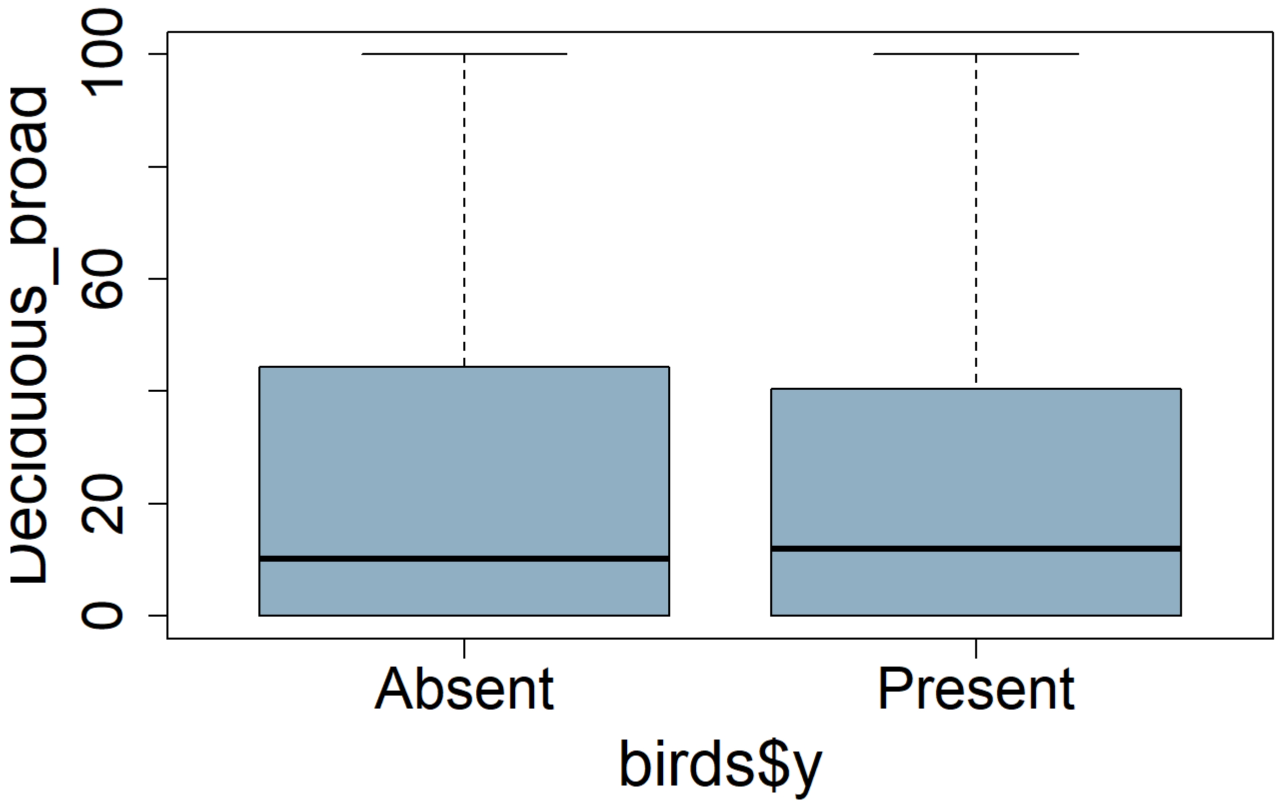

0
20
60
100
Absent
Present
birds$y

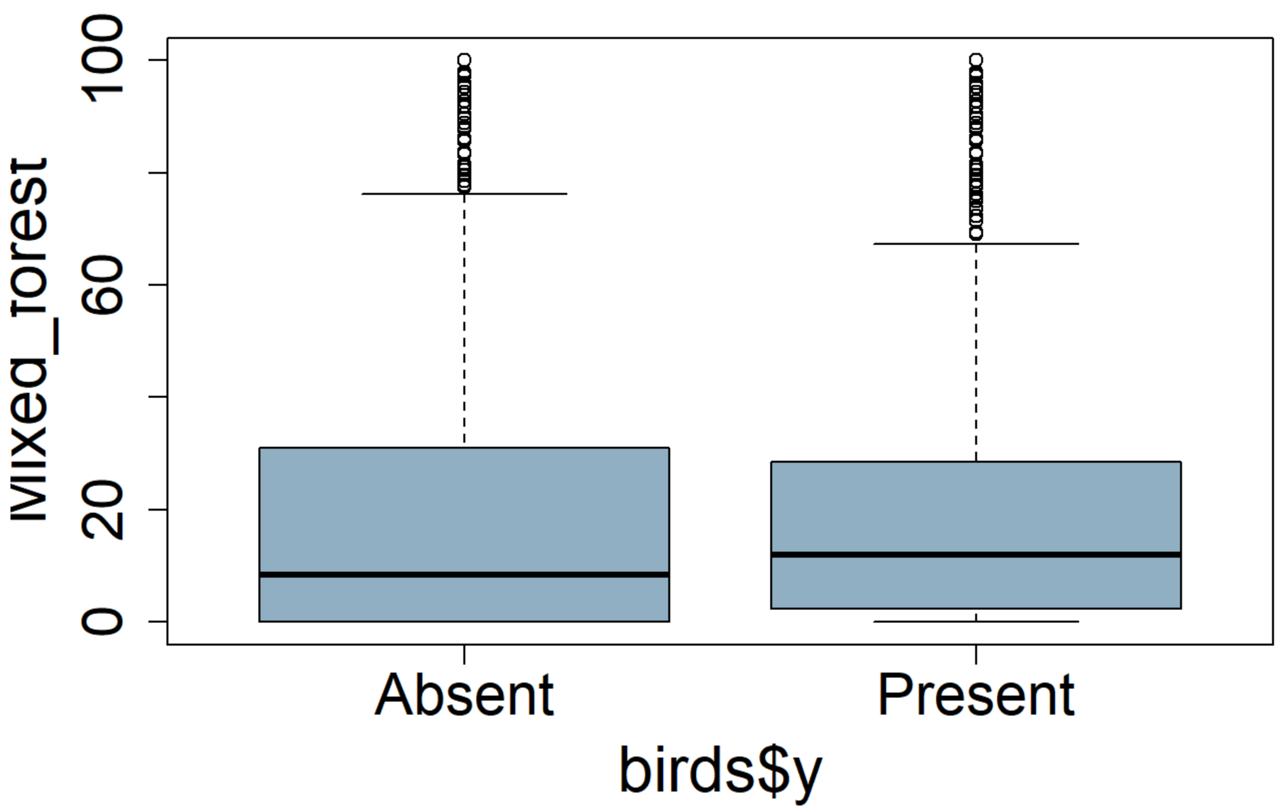

Urban_Built
0
20
60
100
Absent
Present
birds$y

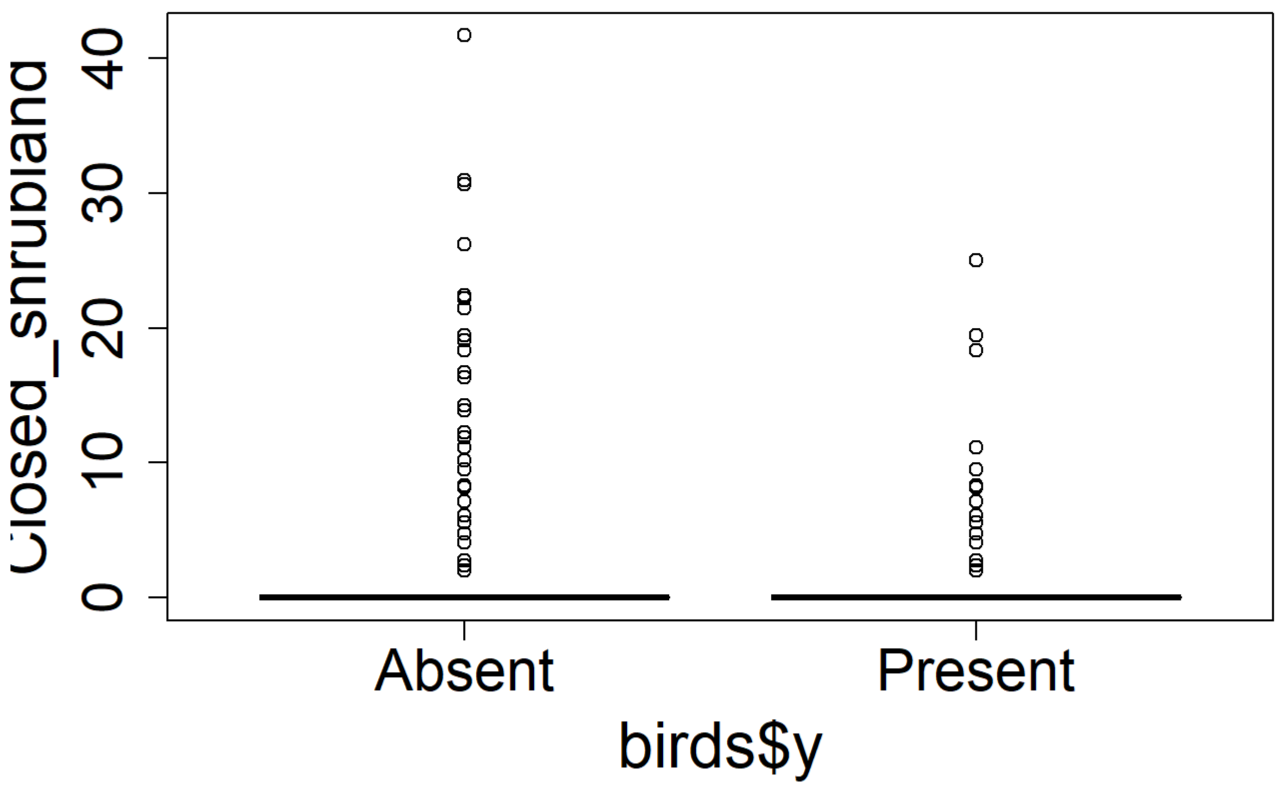

Barren
0
20
40
60
Absent
Present
birds$y

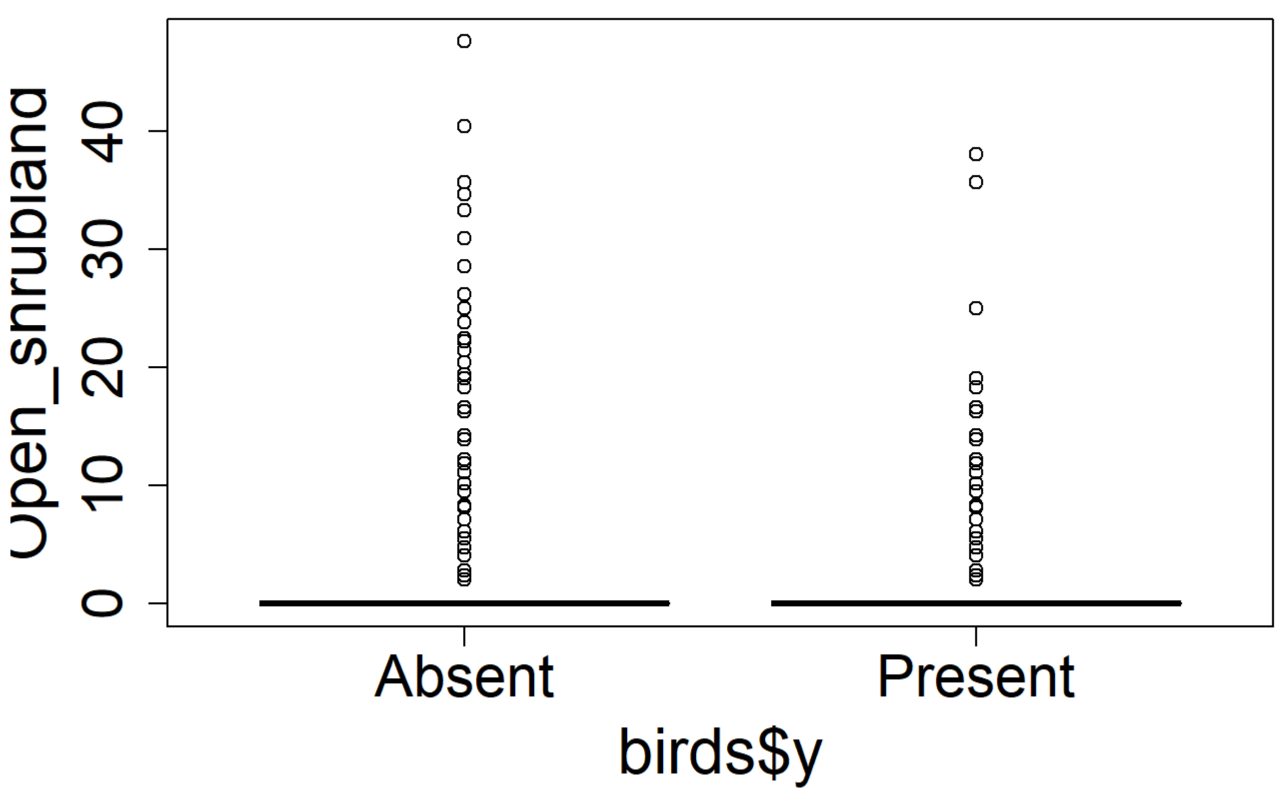

Evergreen_broad
0
5
10
15
20
Absent
Present
birds$y

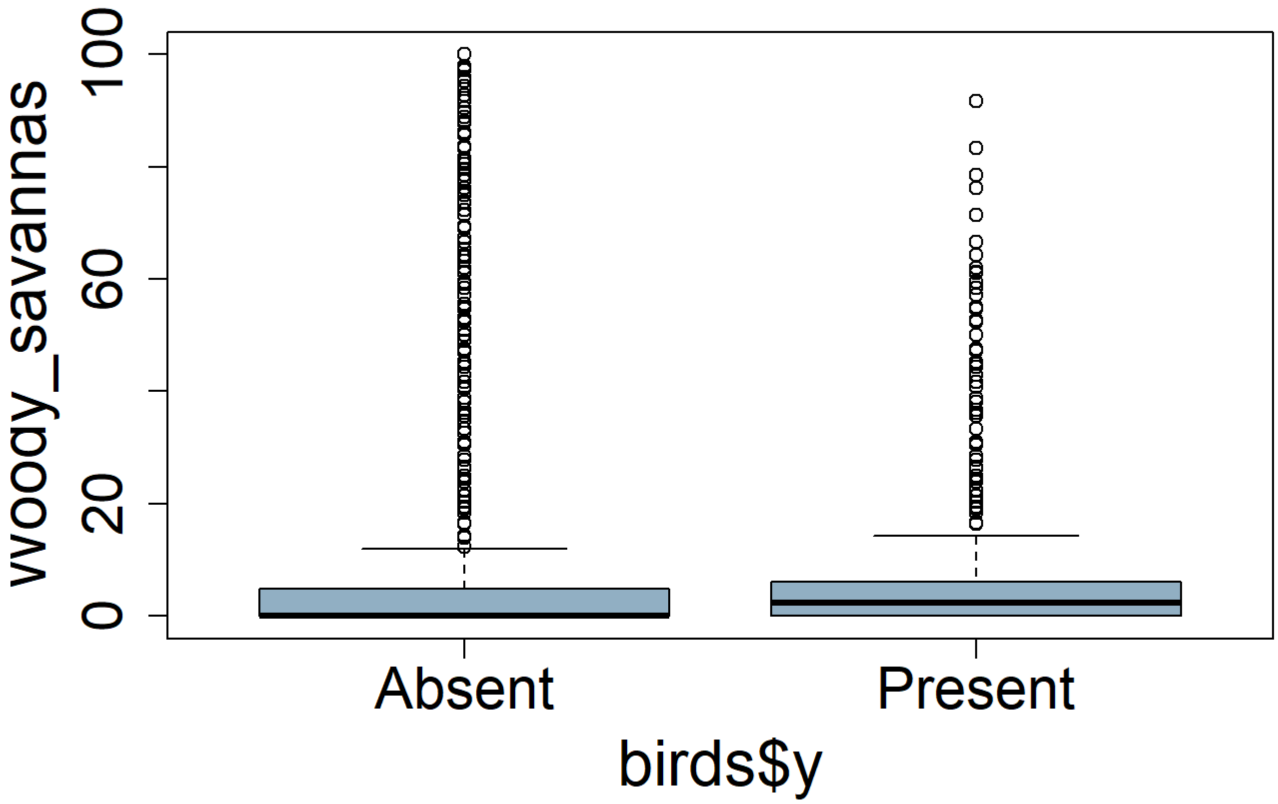

Deciduous_needle
0
10
20
30
40
Absent
Present
birds$y

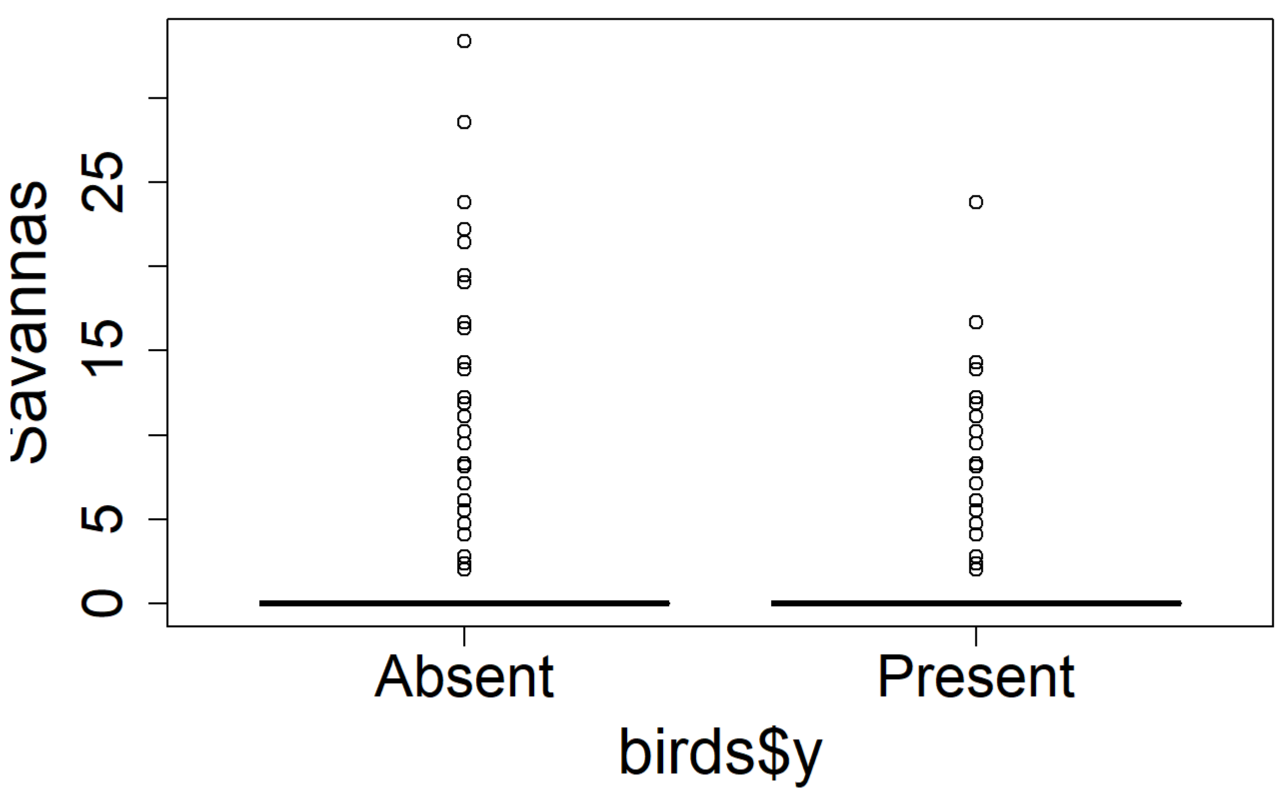

Deciduous_broad
0
20
60
100
Absent
Present
birds$y


Mixed_forest
0
20
60
100
Absent
Present
birds$y


0
10
20
30
40
Absent
Present
birds$y

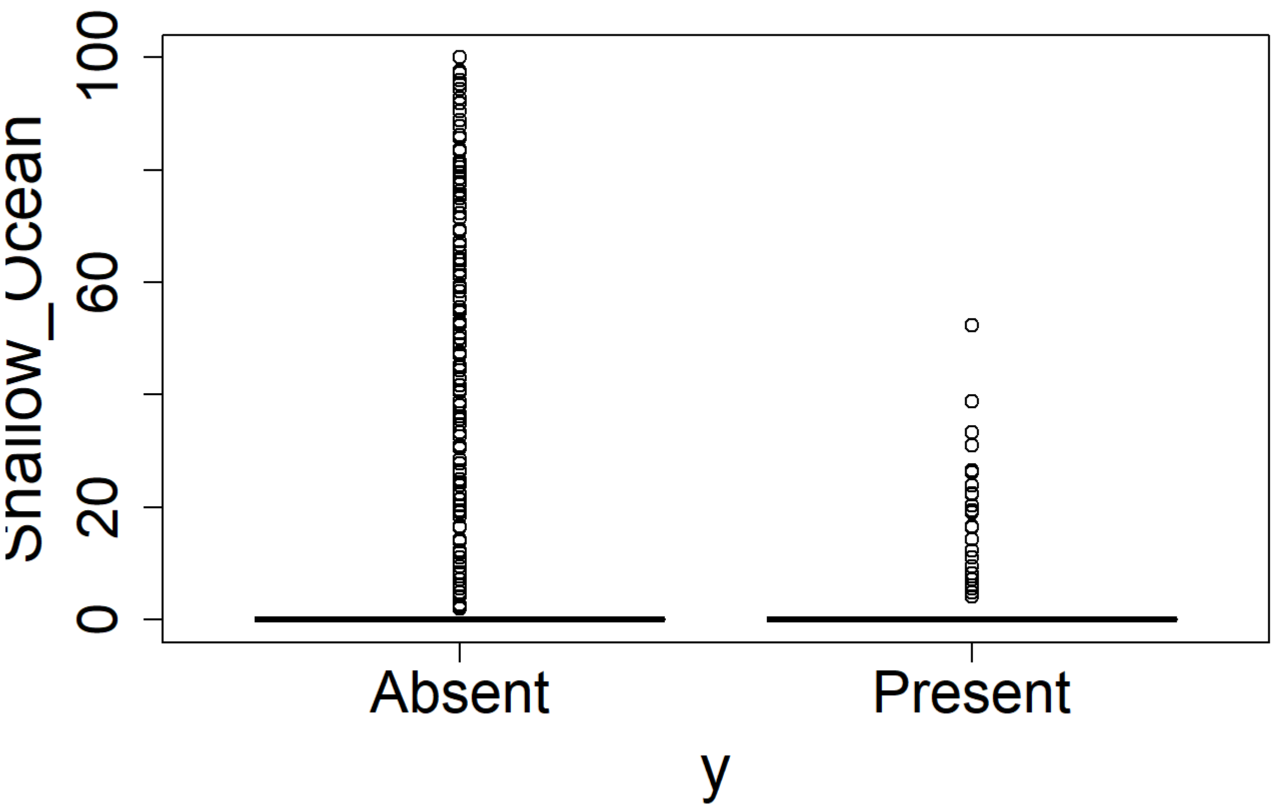

0
10
20
30
40
Absent
Present
birds$y


woody_savannas
100
60
20
0
Absent
Present
birds$y


Savannas
25
15
5
0
Absent
Present
birds$y

Review the box plot for another variable with condensed distribution

```
boxplot(Shallow_Ocean ~ y, # Plot y on the x-axis to examine the
                       # distribution of the predictor variable
                       # Croplands across different categories of y.
  birds,               # Use data from birds data set
  cex.axis = 1.8,      # Adjust size of axes
  cex.lab = 2,         # Adjust size of tick marks
  col = skyBlue)       # Shade boxes skyBlue
```

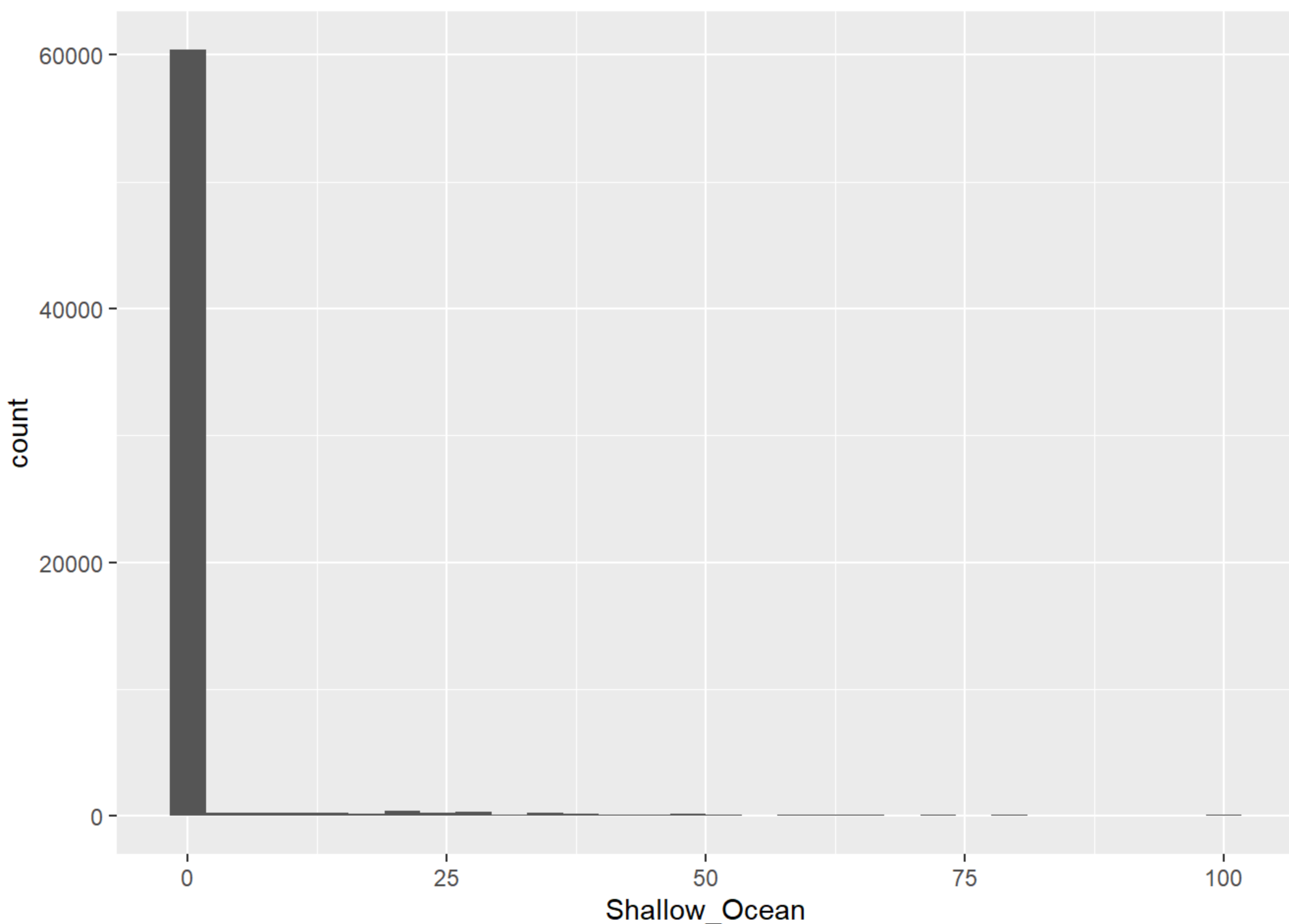



```
birds %>% count(Shallow_Ocean, sort=TRUE) %>% mutate(perc=n*100/sum(n))
```

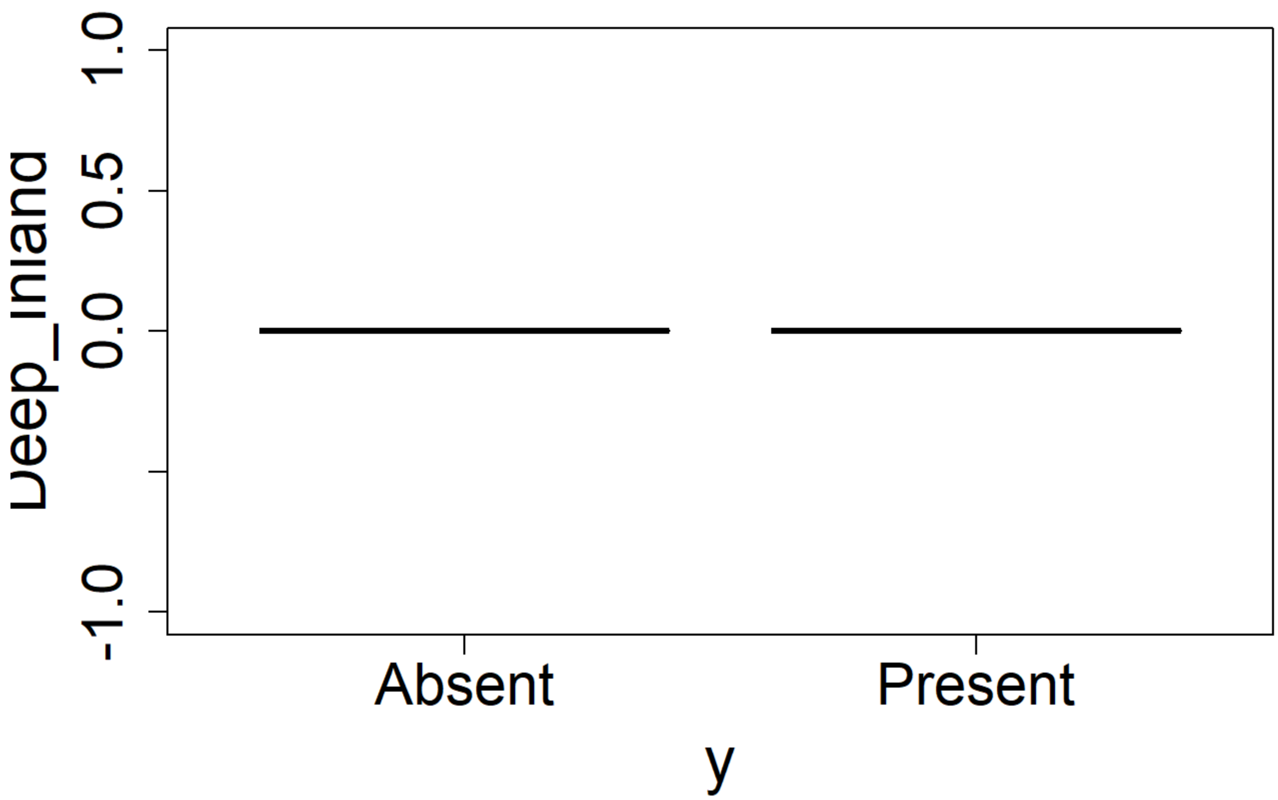
```
## # A tibble: 109 × 3
##    Shallow_Ocean     n   perc
##            <dbl> <int>  <dbl>
##  1          0    60404 93.3  
##  2         33.3    236  0.365
##  3         28.6    202  0.312
##  4         14.3    174  0.269
##  5         22.2    163  0.252
##  6         19.0    156  0.241
##  7         16.7    140  0.216
##  8          9.52   126  0.195
##  9        100      119  0.184
## 10         38.1    115  0.178
## # ℹ 99 more rows
```

```
# 93% values of Shallow_ocean is 0. Remove from dataset?

# Histogram to see the distribution
ggplot(birds, aes(x = Shallow_Ocean)) +
  geom_histogram()
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



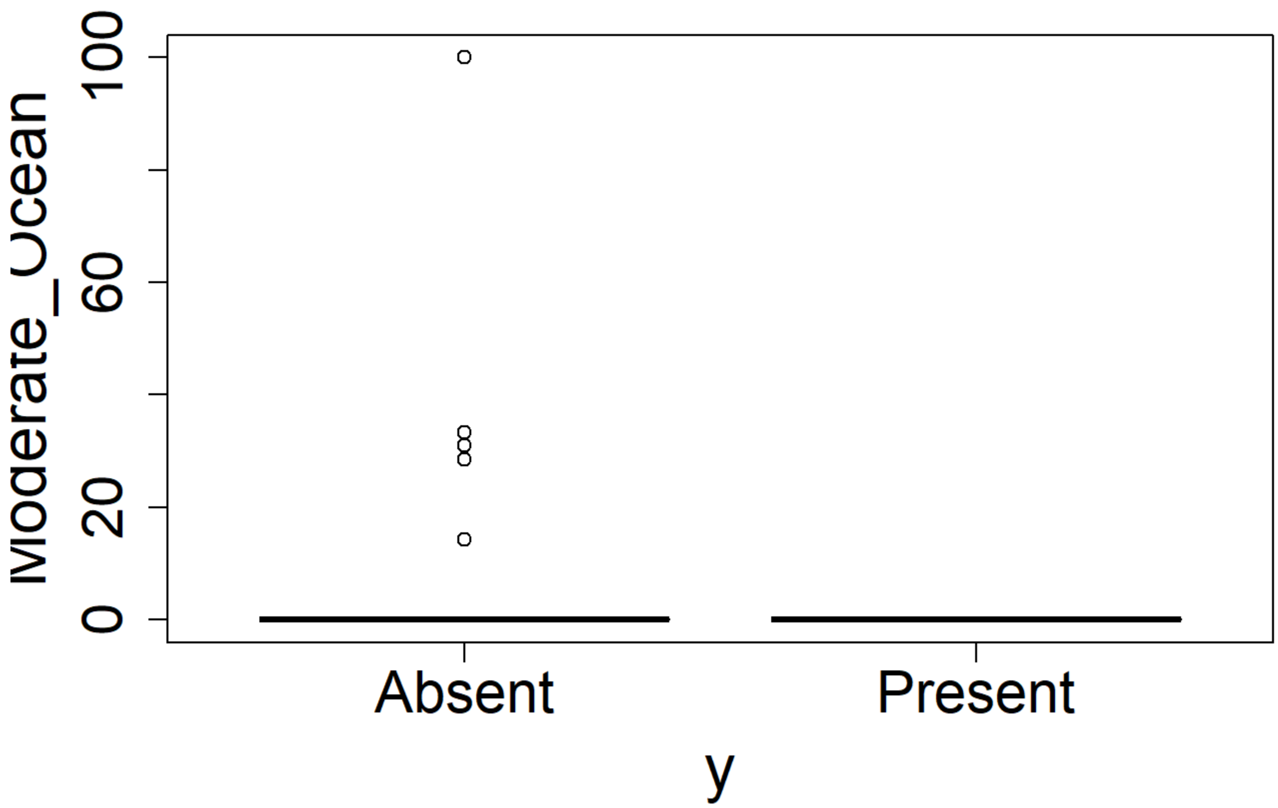
```
boxplot(Deep_Inland ~ y, # Plot y on the x-axis to examine the
                       # distribution of the predictor variable
                       # Croplands across different categories of y.
  birds,               # Use data from birds data set
  cex.axis = 1.8,      # Adjust size of axes
  cex.lab = 2,         # Adjust size of tick marks
  col = skyBlue)       # Shade boxes skyBlue
```




```
birds %>% count(Deep_Inland, sort=TRUE) %>% mutate(perc=n*100/sum(n))
```

```
## # A tibble: 1 × 3
##   Deep_Inland     n  perc
##         <dbl> <int> <dbl>
## 1           0 64724   100
```

```
# 100% values are zero. Remove from dataset?

boxplot(Moderate_Ocean ~ y, # Plot y on the x-axis to examine the
                       # distribution of the predictor variable
                       # Croplands across different categories of y.
  birds,               # Use data from birds data set
  cex.axis = 1.8,      # Adjust size of axes
  cex.lab = 2,         # Adjust size of tick marks
  col = skyBlue)       # Shade boxes skyBlue
```




```
birds %>% count(Moderate_Ocean, sort=TRUE) %>% mutate(perc=n*100/sum(n))
```

```
## # A tibble: 6 × 3
##   Moderate_Ocean     n      perc
##            <dbl> <int>     <dbl>
## 1              0   64706 100.
## 2            100      13   0.0201
## 3             28.6     2   0.00309
## 4             14.3     1   0.00155
## 5             31.0     1   0.00155
## 6             33.3     1   0.00155
```

```
# Only 6 values. Remove from dataset?
# What is many more observations? May try to include as categorical variable
```