Malaria in Hawaiian Birds

Alyssa Wheeler

```
#load in library
library(tidyverse) #loads ggplot and the others

library(infer) #loads lots of things, but particularly our data

#load in csv file
library(readxl)

Mdata <- read_csv("Malaria prevalence DR file.csv")

## Rows: 2945 Columns: 15
## — Column specification —————————————————————————————————
—————
## Delimiter: ","
## chr (12): Band number, Resample, Species, Age, Sex, Site, Location, Island
, ...
## dbl  (3): Elevation, UTM X, UTM Y
##
## 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 m
essage.
```

File contains data from 2018-2021 Hawaii. Data looks at avian malaria in the Hawaiian Archipelago. Taken from the USGS website.

Citation: Paxton, E.H., 2022, Hawaiian forest bird avian malaria prevalence 2018-2021: U.S. Geological Survey data release, https://doi.org/10.5066/P9F519WO.

Description: Hawaii birds Publication Date: 2023-07-28 Start Date: 2018-12-10 End Date: 2021-06-09

```
#list variables in dataset
ls(Mdata)

##  [1] "Age"         "Band number" "Elevation"   "Island"      "Location"
##  [6] "Prevalence"  "qPCR Run 1"  "qPCR Run 2"  "Resample"    "Sample Date"
## [11] "Sex"         "Site"        "Species"     "UTM X"       "UTM Y"

#list of variables and their types
str(Mdata)

## spc_tbl_ [2,945 × 15] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ Band number: chr [1:2945] "2700-54540" "2700-54539" "2700-54538" "2641-
48296" ...
##  $ Resample   : chr [1:2945] "N" "N" "N" "N" ...
##  $ Species    : chr [1:2945] "Warbling White-eye" "Warbling White-eye" "Wa
```

```
rbling White-eye" "Red-crested Cardinal" ...
##  $ Age        : chr [1:2945] "Adult" "Adult" "Adult" "Adult" ...
##  $ Sex        : chr [1:2945] "U" "U" "U" "U" ...
##  $ Site       : chr [1:2945] "Manana Trail" "Manana Trail" "Manana Trail"
"Manana Trail" ...
##  $ Location   : chr [1:2945] "Ewa Forest Reserve" "Ewa Forest Reserve" "Ew
a Forest Reserve" "Ewa Forest Reserve" ...
##  $ Island     : chr [1:2945] "O'ahu" "O'ahu" "O'ahu" "O'ahu" ...
##  $ Elevation  : num [1:2945] 338 338 338 338 338 338 338 338 338 338 ...
##  $ UTM X      : num [1:2945] 610623 610623 610623 610623 610623 ...
##  $ UTM Y      : num [1:2945] 2370302 2370302 2370302 2370302 2370302 ...
##  $ Sample Date: chr [1:2945] "5/25/2021" "5/25/2021" "5/25/2021" "5/25/202
1" ...
##  $ qPCR Run 1 : chr [1:2945] "35.5211" "37.9423" "36.8564" "Undetermined"
...
##  $ qPCR Run 2 : chr [1:2945] "34.8648" "38.3767" "36.1737" "Undetermined"
...
##  $ Prevalence : chr [1:2945] "positive" "positive" "positive" "negative" .
..
##  - attr(*, "spec")=
##   .. cols(
##   ..   `Band number` = col_character(),
##   ..   Resample = col_character(),
##   ..   Species = col_character(),
##   ..   Age = col_character(),
##   ..   Sex = col_character(),
##   ..   Site = col_character(),
##   ..   Location = col_character(),
##   ..   Island = col_character(),
##   ..   Elevation = col_double(),
##   ..   `UTM X` = col_double(),
##   ..   `UTM Y` = col_double(),
##   ..   `Sample Date` = col_character(),
##   ..   `qPCR Run 1` = col_character(),
##   ..   `qPCR Run 2` = col_character(),
##   ..   Prevalence = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

Categorical Variables: Band number, Resample, Species, Age, Sex, Site, Location, Island, Sample Data, qPCR Run 1, qPRC run 2, Prevalence

Quantitative Variables: Elevation, UTM X, UTM Y

Variable Descriptions (from USGS): Band number: Unique numerical identifier of each bird from metal leg band placed on each sampled bird. In cases where birds were not banded (such as when had pox on legs), unique identifier was the 4 letter bird code and values from 01-08). Resample: Designation of whether birds were sampled more than once at different times over the course of the study (i.e., recaptured in subsequent sampling trips) Species: Common name of forest bird species Age: Age of bird, based on plumage

characteristics Sex: Sex of individual (M= male, F= female, U= unknown) Site: Name of place where sampling occurred Location: General location of sampling site Island: Hawaiian island sample was conducted on Sample Data: Date when individual was sampled in format mm/dd/yyyy qPCR Run 1: CT value (Cycle Threshold) of first qPCR run qPRC run 2: CT value (Cycle Threshold) of duplicate (second) qPCR run Prevalence: Status of qPCR results, summarized over both runs Elevation: Elevation in meters UTM X: Longitudinal UTM coordinate for the sampling site, NAD83 datum and UTM zone 4 (Kaua'i, O'ahu, Maui, Molokai islands) or zone 5 (Hawai'i Island) UTM Y: Latitudinal UTM coordinate for the sampling site, NAD83 datum and UTM zone 4 (Kaua'i, O'ahu, Maui, Molokai islands) or zone 5 (Hawai'i Island)

Purpose: I am examining the these ideas to see how location affects the prevalence of malaria in Hawaii island bird populations.

Data Cleaning and Manipulation:

```
#count for categorical variables
Mdata %>%
count(Age)

## # A tibble: 4 × 2
##   Age          n
##   <chr>    <int>
## 1 Adult     2206
## 2 Juvenile   292
## 3 SubAdult   326
## 4 Unknown    121

Mdata %>%
count(`Band number`)

## # A tibble: 2,895 × 2
##     `Band number`     n
##     <chr>         <int>
##  1 0602-92508        1
##  2 0602-92509        1
##  3 0602-92510        1
##  4 0602-92511        1
##  5 0602-92512        1
##  6 0602-92513        1
##  7 0602-92516        1
##  8 0602-92517        1
##  9 0602-92518        1
## 10 0602-92519        1
## # i 2,885 more rows

Mdata %>%
count(Island)

## # A tibble: 5 × 2
##   Island       n
```

```
##    <chr>    <int>
## 1 Hawai'i  1954
## 2 Kaua'i    122
## 3 Maui      668
## 4 Molokai    38
## 5 O'ahu     163

Mdata %>%
count(Location)

## # A tibble: 32 × 2
##    Location                            n
##    <chr>                           <int>
##  1 ?la'a Forest Reserve               9
##  2 Ewa Forest Reserve                20
##  3 Haleakal? National Park          397
##  4 Hana Highway                      13
##  5 Hawai'i Volcanoes National Park  467
##  6 Hawaiian Paradise Park            68
##  7 Hilo                              21
##  8 Honolulu Watershed Forest Reserve 23
##  9 Honouliuli Forest Reserve         29
## 10 Hualalai                          73
## # i 22 more rows

Mdata %>%
count(Prevalence)

## # A tibble: 2 × 2
##   Prevalence      n
##   <chr>       <int>
## 1 negative     1929
## 2 positive     1016

Mdata %>%
count(`qPCR Run 1`)

## # A tibble: 929 × 2
##     `qPCR Run 1`      n
##     <chr>         <int>
##  1 19.0259           1
##  2 19.3088           1
##  3 21.6523           1
##  4 23.9478           1
##  5 24.7527           1
##  6 24.9547           1
##  7 24.9935           1
##  8 25.0792           1
##  9 25.1494           1
## 10 25.2341           1
## # i 919 more rows
```

```
Mdata %>%
count(`qPCR Run 2`)

## # A tibble: 903 × 2
##    `qPCR Run 2`       n
##    <chr>          <int>
##  1 18.9883            1
##  2 19.2829            1
##  3 21.7934            1
##  4 23.7598            1
##  5 25.0969            1
##  6 25.1361            1
##  7 25.168             1
##  8 25.2024            1
##  9 25.2458            1
## 10 25.2804            1
## # i 893 more rows

Mdata %>%
count(Resample)

## # A tibble: 4 × 2
##    Resample     n
##    <chr>    <int>
## 1 1           46
## 2 2           46
## 3 3            1
## 4 N         2852

Mdata %>%
count(`Sample Date`)

## # A tibble: 251 × 2
##    `Sample Date`      n
##    <chr>          <int>
##  1 1/14/2020         19
##  2 1/16/2020         14
##  3 1/17/2020         17
##  4 1/23/2020         12
##  5 1/24/2020          3
##  6 1/25/2019          2
##  7 1/27/2020          4
##  8 1/28/2020         29
##  9 1/29/2020          5
## 10 1/30/2020          1
## # i 241 more rows

Mdata %>%
count(Sex)
```

```
## # A tibble: 3 × 2
##   Sex       n
##   <chr> <int>
## 1 F       308
## 2 M       653
## 3 U      1984

Mdata %>%
count(Site)

## # A tibble: 66 × 2
##    Site                  n
##    <chr>             <int>
##  1 Ainahou Ranch       134
##  2 Alakai Swamp Trail   31
##  3 Biocomplexity Site    1
##  4 Bird Transect 17      6
##  5 Blair Road            4
##  6 Boy's Home           81
##  7 Byron's Ledge         3
##  8 Cabin                89
##  9 Camp                136
## 10 Charlie Camp         39
## # i 56 more rows

Mdata %>%
count(Species)

## # A tibble: 39 × 2
##    Species                   n
##    <chr>                 <int>
##  1 'Akiapola'au              3
##  2 'Akikiki                  8
##  3 'Akohekohe                1
##  4 'Anianiau                17
##  5 'Apapane                218
##  6 'I'iwi                  171
##  7 African Silverbill        1
##  8 Alawi (Hawai'i Creeper)  18
##  9 Chinese Hwamei            4
## 10 Common Myna               1
## # i 29 more rows

#For Age variable, change Unknown to NA
Mdata=Mdata %>%
  mutate( Age=ifelse(Age=="Unknown", NA, Age))
#check correction
Mdata %>%
count(Age)
```

```
## # A tibble: 4 × 2
##    Age            n
##    <chr>      <int>
## 1 Adult       2206
## 2 Juvenile     292
## 3 SubAdult     326
## 4 <NA>         121
```

#For Band number, rename variable to remove space
```
Mdata=Mdata %>%
  rename(Bandnumber=`Band number`)
```

#For qPCR Run 1 and 2, rename variables to remove spaces
```
Mdata=Mdata %>%
  rename(qPCR1=`qPCR Run 1`)
Mdata=Mdata %>%
  rename(qPCR2=`qPCR Run 2`)
```

#For Sample Date, rename variable to remove space
```
Mdata=Mdata %>%
  rename(Sampledate=`Sample Date`)
```
#check correction
```
Mdata %>%
ls(Mdata)
```

```
##  [1] "Age"        "Bandnumber" "Elevation"  "Island"     "Location"
##  [6] "Prevalence" "qPCR1"      "qPCR2"      "Resample"   "Sampledate"
## [11] "Sex"        "Site"       "Species"    "UTM X"      "UTM Y"
```

#For Sex variable, change Unknown to NA
```
Mdata=Mdata %>%
  mutate(Sex=ifelse(Sex=="U", NA, Sex))
```
#check correction
```
Mdata %>%
count(Sex)
```

```
## # A tibble: 3 × 2
##    Sex        n
##    <chr> <int>
## 1 F        308
## 2 M        653
## 3 <NA>    1984
```

#Summary for quantitative variables
```
Mdata %>%
summary(Mdata)
```

```
##   Bandnumber           Resample            Species              Age
##  Length:2945        Length:2945        Length:2945        Length:2945
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
```

```
##
##
##
##       Sex               Site              Location             Island
##   Length:2945        Length:2945        Length:2945        Length:2945
##   Class :character   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##      Elevation         UTM X             UTM Y            Sampledate
##   Min.   :   2    Min.   :194457   Min.   :2110758   Length:2945
##   1st Qu.: 890    1st Qu.:215998   1st Qu.:2145604   Class :character
##   Median :1240    Median :258282   Median :2185061   Mode  :character
##   Mean   :1231    Mean   :396604   Mean   :2214596
##   3rd Qu.:1662    3rd Qu.:610623   3rd Qu.:2291176
##   Max.   :2000    Max.   :804467   Max.   :2451925
##      qPCR1             qPCR2              Prevalence
##   Length:2945        Length:2945        Length:2945
##   Class :character   Class :character   Class :character
##   Mode  :character   Mode  :character   Mode  :character
##
##
##
```

```r
#For UTM X and Y, rename variables to remove space
Mdata=Mdata %>%
  rename(UTMX=`UTM X`)
Mdata=Mdata %>%
  rename(UTMY=`UTM Y`)
#check correction
Mdata %>%
ls(Mdata)
```

```
##  [1] "Age"        "Bandnumber" "Elevation"  "Island"     "Location"
##  [6] "Prevalence" "qPCR1"      "qPCR2"      "Resample"   "Sampledate"
## [11] "Sex"        "Site"       "Species"    "UTMX"       "UTMY"
```

Description of cleaning process: Most of the cleaning was removing the spaces in between the variable names. The other cleaning done was to change the "Unknown" data to NA.

Variation: TO start looking at the locations, I am first going to find the specifc variables that seem to affect the location or prevalence variable enough to take a closer look.

```r
Mdata %>%
ggplot()+
  geom_bar(aes(x=Age))+
    labs(title="Age of Birds", x="Age", y="count")
```

Age of Birds
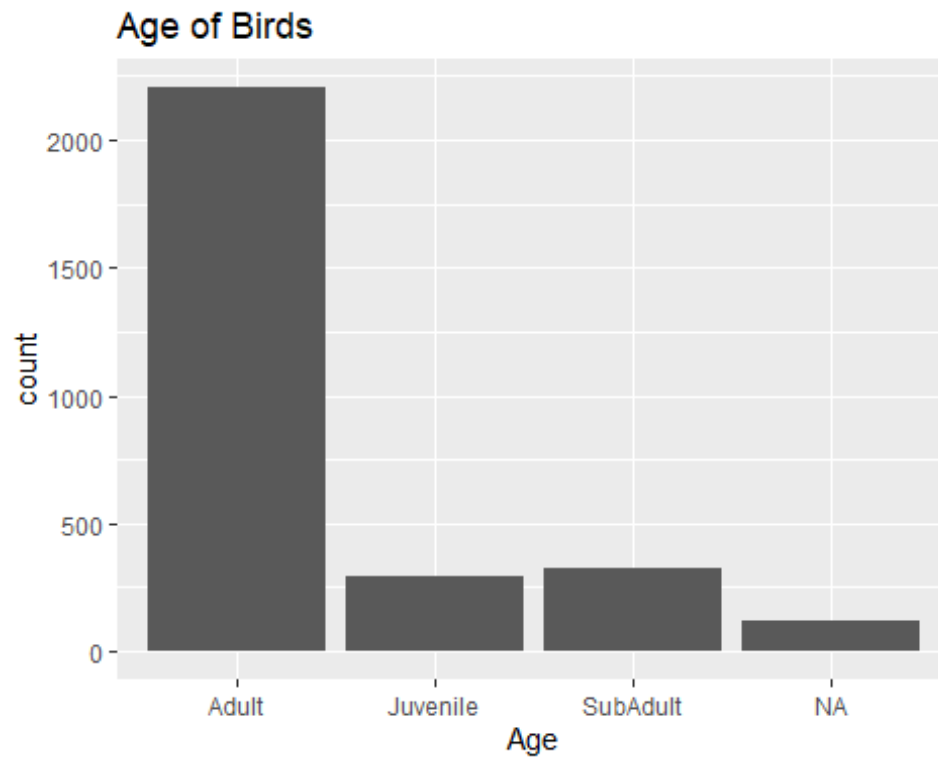
From the bar chart, we can see that the majority of birds in the study were adults. There were not as many that were Juvenile, SubAdult, or unknown.

```r
Mdata %>%
ggplot()+
  geom_bar(aes(x=Island))+
    labs(title="Island Where Birds Found", x="Island", y="count")
```
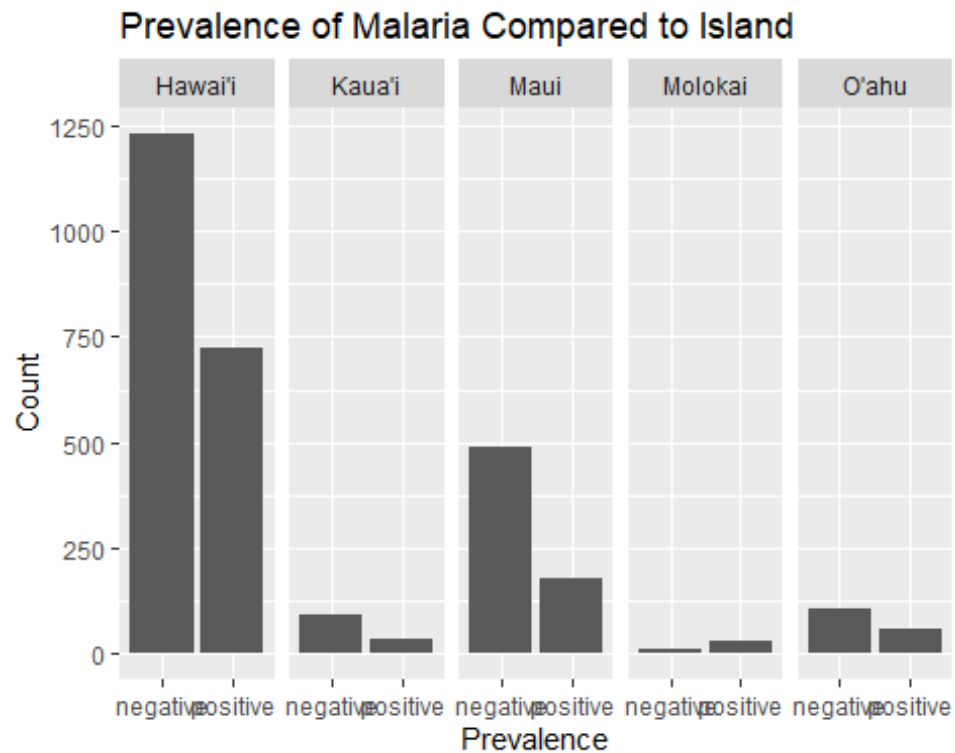
Island Where Birds Found
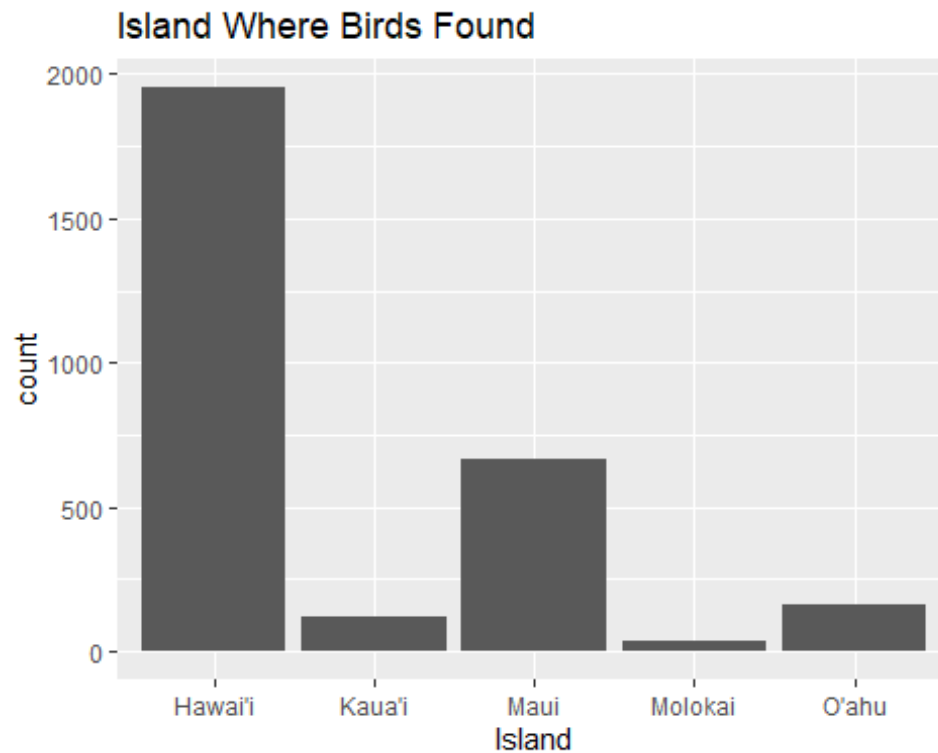
This bar graph shows that most of the birds came from the island of Hawai'i, while the next largest came from Maui. There are only 5 islands overall that had birds in the study.

```
Mdata %>%
ggplot()+
  geom_bar(aes(x=Sex))+
    labs(title="Sex of Birds", x="Sex", y="count")
```

Sex of Birds
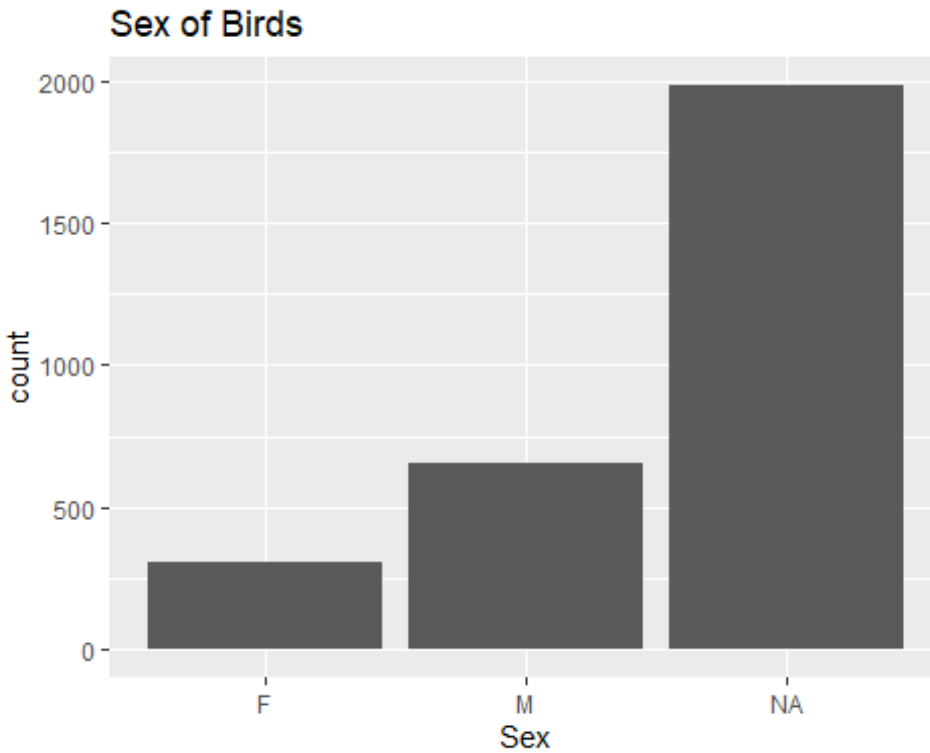
According to the bar chart, the large majority of birds had a sex that was unknown. Of the ones that were known, they were mostly male. Because of the amount of unknown, it would be hard to compare this variable with other variables for accuracy.

```
Mdata %>%
ggplot()+
  geom_bar(aes(x=Prevalence))+
  facet_grid(~Island, scales="free",space="free")+
    labs(title="Prevalence of Malaria Compared to Island", x="Prevalence", y=
"Count")
```

Prevalence of Malaria Compared to Island

| Hawai'i | Kaua'i | Maui | Molokai | O'ahu |

(Count axis: 0, 250, 500, 750, 1000, 1250)

X-axis per facet: negative / positive

X-axis label: Prevalence

There were more positive than negative birds having malaria on the Molokai island. All other islands had more birds test negative for malaria. The Kaua'i island did not have many test positive compared to the negative.

```
Mdata %>%
ggplot()+
  geom_bar(aes(x=Elevation, fill=Island))+
  facet_grid(~Prevalence, scales="free",space="free")+
    labs(title="Island and Elevation Compared to Prevalence of Malaria", x="P
revalence", y="Elevation")
```

Island and Elevation Compared to Prevalence of Mala
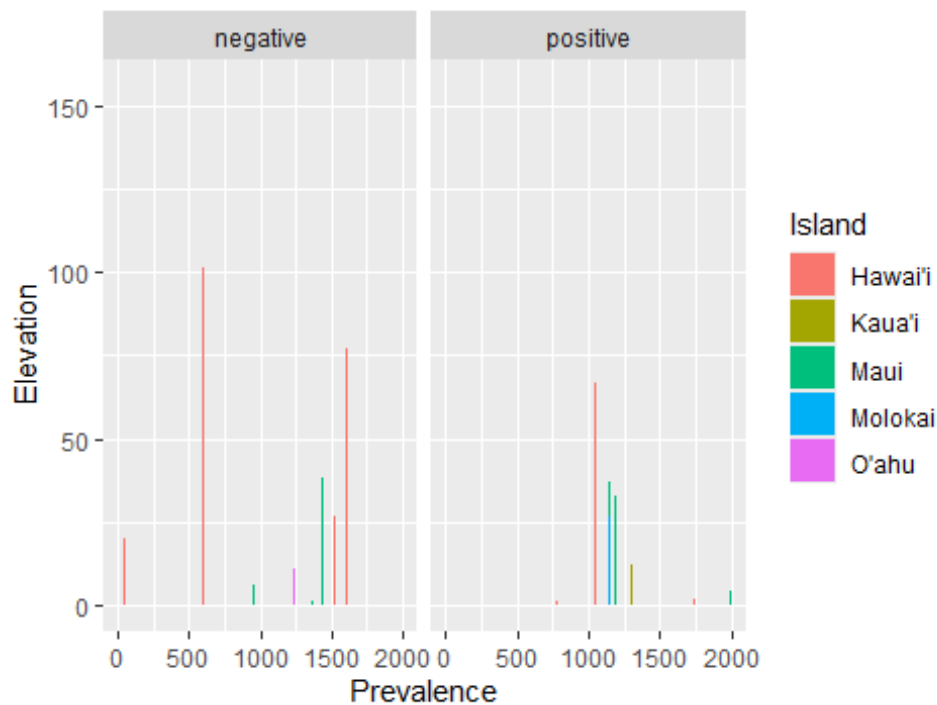
These charts show a few things. The majority of birds are found in Hawai'i and Maui for this study. This graph showed the majority of birds positive for malaria were found at an elevation of 1000-1250 ft. However, the birds that were majority negatives for malaria were found at two different elevations around 600 and around 1500 ft.

```
Mdata %>%
ggplot()+
geom_boxplot (mapping=aes (x=Elevation, y=Prevalence))+
labs(title="Prevalence vs Elevation", x="Elevation (meters)", y="Prevalence")
```

Prevalence vs Elevation
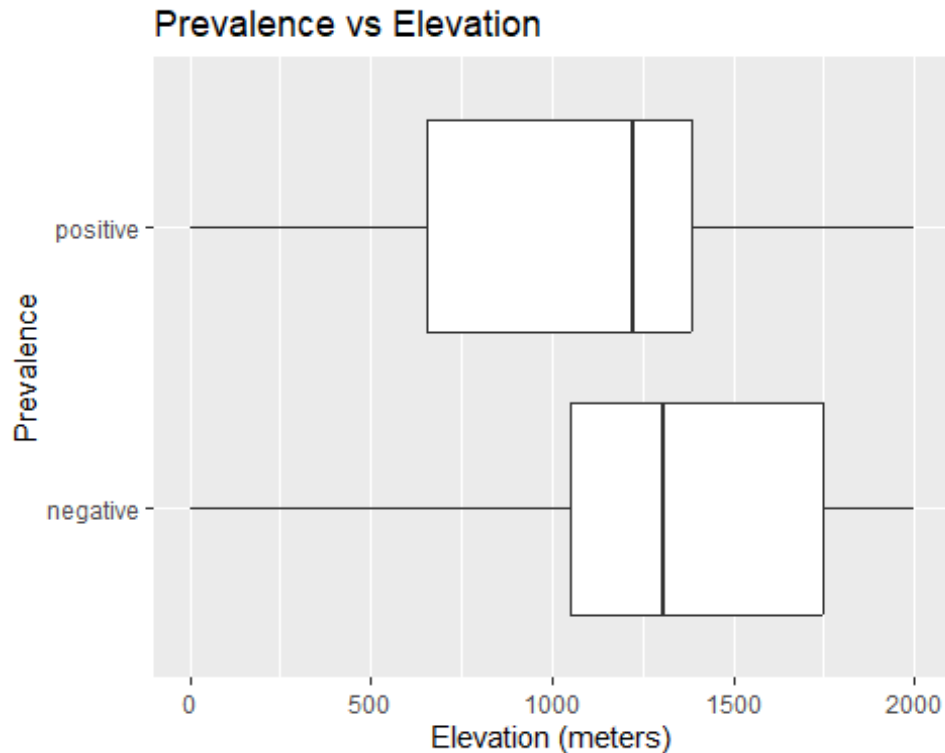
The boxplot shows that the median elevation at which the birds were positive for malaria was around 1200 meters. The median elevation at which birds were negative for malaria was around 1300 meters. Visually, we can see the birds with positive malaria were at a slightly lower elevation than the birds negative for malaria, but there is some overlap in elevation. That is, there were birds found around the similar median elevation for both positive and negative prevalence.

```
Mdata %>%
ggplot()+
geom_point(mapping=aes(x=UTMX, y=UTMY, shape=Island))+
labs(title="Latitudinal and Longitudinal Coordinates with Identified Island",
x="Longitudinal Coordinates", y="Latitudinal Coordinates")
```

Latitudinal and Longitudinal Coordinates with Iden
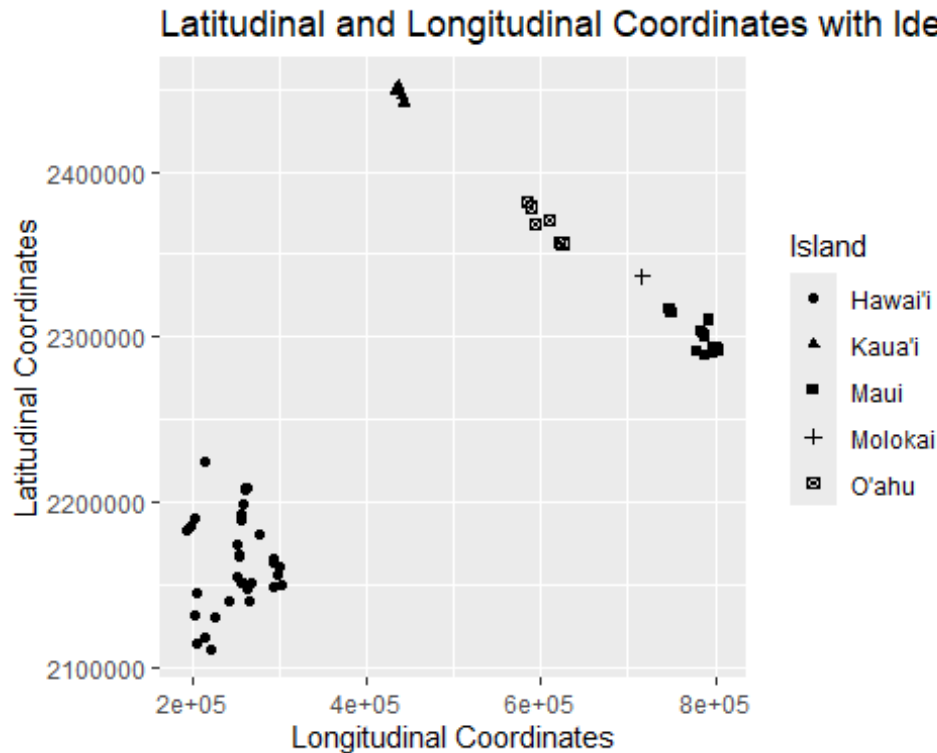
This scatterplot roughly shows a map of where the birds were found because it is looking at coordinates. There seem to be two areas where birds were found for the study. One area is the lower coordinates. The other is a line of coordinates in the top right of the graph. After looking at the two different locations, it seems interesting to group the islands in the top right of the map together because they are close together. It is possible birds from these islands are flying between the islands because they are so close.

```
#Select only variables needed
group_island=Mdata %>%
  select(Island,Prevalence,Elevation,Age,Sex)
group_island

## # A tibble: 2,945 × 5
##     Island Prevalence Elevation Age   Sex
##     <chr>  <chr>          <dbl> <chr> <chr>
##  1 O'ahu   positive         338 Adult <NA>
##  2 O'ahu   positive         338 Adult <NA>
##  3 O'ahu   positive         338 Adult <NA>
##  4 O'ahu   negative         338 Adult <NA>
##  5 O'ahu   negative         338 Adult F
##  6 O'ahu   positive         338 Adult M
##  7 O'ahu   positive         338 Adult M
##  8 O'ahu   negative         338 Adult F
##  9 O'ahu   positive         338 Adult <NA>
## 10 O'ahu   positive         338 Adult <NA>
## # i 2,935 more rows
```

```
#group all islands but Hawaii together
group_island=group_island %>%
  mutate(Island=ifelse(Island=="O'ahu", "MultIsland", Island))
group_island=group_island %>%
  mutate(Island=ifelse(Island=="Molokai", "MultIsland", Island))
group_island=group_island %>%
  mutate(Island=ifelse(Island=="Maui", "MultIsland", Island))
group_island=group_island %>%
  mutate(Island=ifelse(Island=="Kaua'i", "MultIsland", Island))
group_island

## # A tibble: 2,945 × 5
##    Island      Prevalence Elevation Age   Sex
##    <chr>       <chr>          <dbl> <chr> <chr>
##  1 MultIsland positive         338 Adult <NA>
##  2 MultIsland positive         338 Adult <NA>
##  3 MultIsland positive         338 Adult <NA>
##  4 MultIsland negative         338 Adult <NA>
##  5 MultIsland negative         338 Adult F
##  6 MultIsland positive         338 Adult M
##  7 MultIsland positive         338 Adult M
##  8 MultIsland negative         338 Adult F
##  9 MultIsland positive         338 Adult <NA>
## 10 MultIsland positive         338 Adult <NA>
## # i 2,935 more rows

group_island %>%
ggplot()+
  geom_bar(aes(x=Elevation, fill=Island))+
  facet_grid(~Prevalence, scales="free",space="free")+
    labs(title="Grouped Islands and Elevation Compared to Prevalence of Malar
ia", x="Elevation", y="Count")
```

Grouped Islands and Elevation Compared to Prevalen
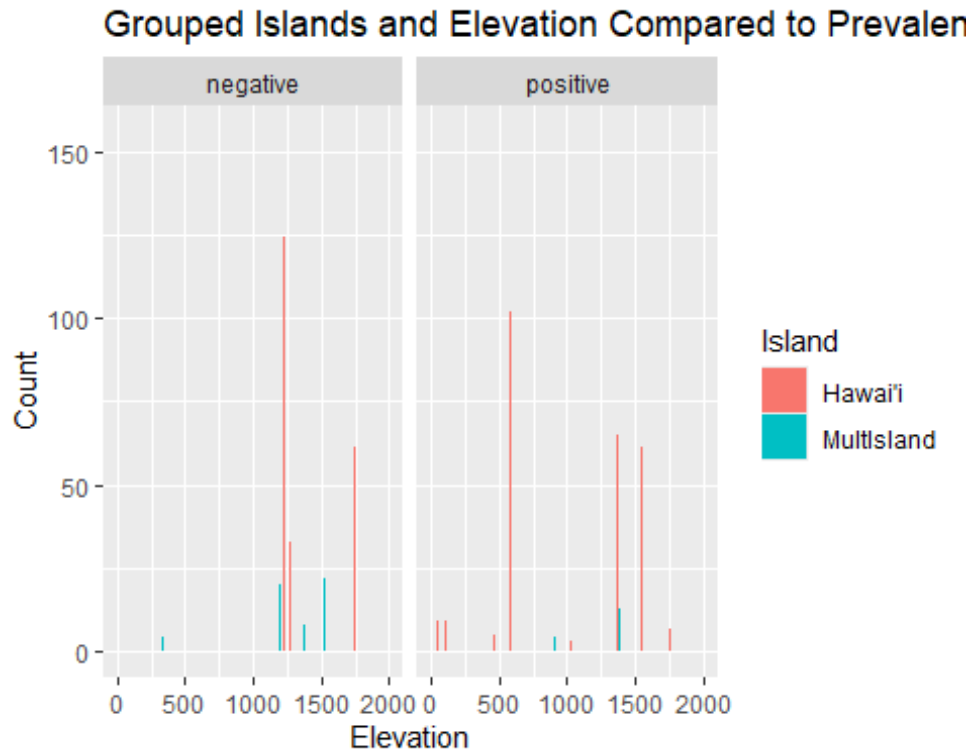
Hawaii still had an overwhelming amount of birds found at specific elevations compared to the other islands. The elevation does not seem to have a specific impact on prevalance of malaria. It seems to be more of a coincidence with where the birds probably are most of the time or possibly where they find the best spots on each island.

```
group_island %>%
ggplot()+
  geom_bar(aes(x=Prevalence))+
  facet_grid(~Island, scales="free",space="free")+
    labs(title="Prevalence Compared to Grouped Islands and Hawaii", x="Preval
ence", y="Count")
```

Prevalence Compared to Grouped Islands and Hawa
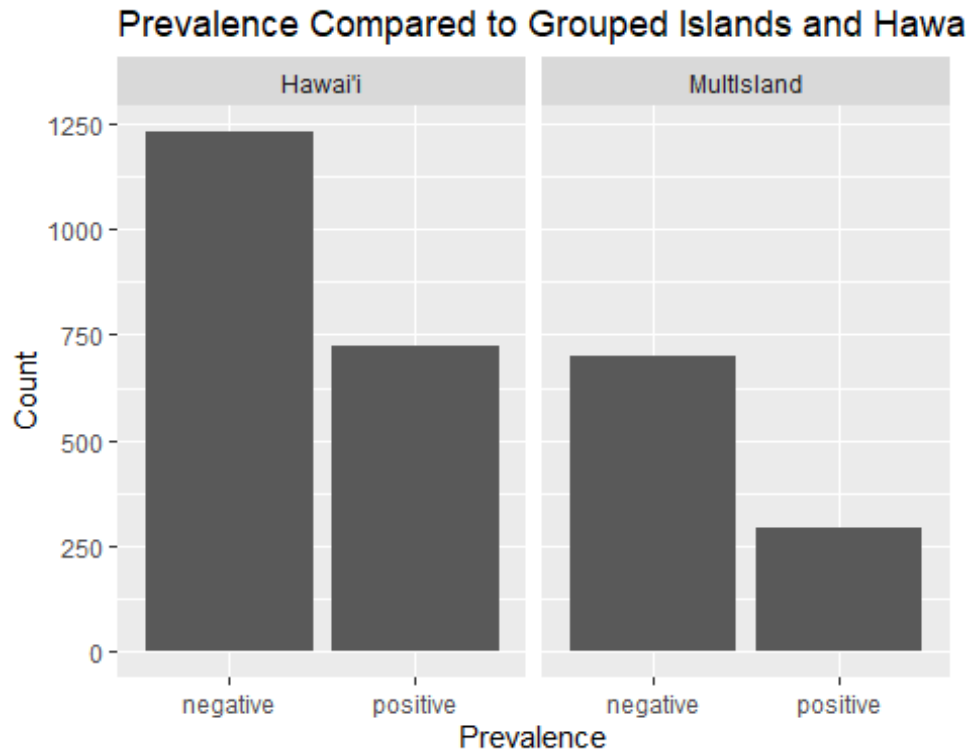
Although there were still more birds found on Hawaii compared to all other islands combined, there still seems to be a trend in the number of birds testing positive for malaria. For both Hawaii and the group of islands, about half of the birds tested positive for malaria. This shows that for however many birds tested about half will test positive for malaria in the location that the birds stay in (grouped islands or one island as in Hawaii).

Conclusion: There are specific elevation points where birds were found, but there doesn't seem to be any correlation between elevation and prevalence for malaria. The majority of birds that were found were from the Hawaii island. There were more positive birds for malaria than negative found on the Hawaii and Maui islands. After looking at the coordinates of where the birds were found compared to the islands, it seemed more likely that the birds should be grouped by either Hawaii or all other islands. This is due to Hawaii being farther from all other islands, and the other islands being relatively close together. It is possible the birds are travelling from island to island in the group of islands because of how close the islands are, and this could cause them to spread malaria more easily between islands. After comparing the grouped islands and Hawaii against the prevalence of malaria, it seems that about half of the birds tested positive for malaria for the area they frequent. This shows that the location plays a role in the prevalence of malaria only when islands are close together. It would be interesting to see how far away islands would have to be to not have an effect on each other. It would also be interesting to track birds between the islands better to verify they are going from island to island or that malaria is being passed from island to island in some way.