

Hoops Longwing Data Analysis

Zane Billings

15 November 2019

In order to start analyzing the Hoops' Longwing sample data, we will first load the **tidyverse** package suite. After loading the packages we need, we can use `readr::read_csv()` to load in the data. But, notice the imported data frame has a useless column at the beginning, which we can easily remove manually.

```
library(tidyverse)

butterfly <- read_csv("hoops_longwing_study.csv")
butterfly <- butterfly[, -1]
```

Now that we have the data imported, we can go ahead and take a quick look at the summary and structure.

```
summary(butterfly)

##   wing_length   wing_width      age   num_offspring
##   Min.      : 7.46   Min.      : 2.730   Min.      : 2.00   Min.      : 8.00
##   1st Qu.:14.19   1st Qu.: 6.670   1st Qu.:12.00   1st Qu.:24.00
##   Median :16.78   Median : 8.135   Median :19.00   Median :28.00
##   Mean    :20.24   Mean    : 8.748   Mean    :22.68   Mean    :27.84
##   3rd Qu.:26.74   3rd Qu.:10.910   3rd Qu.:31.00   3rd Qu.:32.00
##   Max.    :42.18   Max.    :17.390   Max.    :61.00   Max.    :39.00
##   feeding_range   color_peak   num_mates   avg_scale_size
##   Min.      :-0.250   Min.      :357.9   Min.      :-2.000   Min.      :18.27
##   1st Qu.: 2.640   1st Qu.:385.9   1st Qu.: 3.000   1st Qu.:28.02
##   Median : 3.510   Median :392.0   Median : 5.000   Median :32.39
##   Mean    : 5.997   Mean    :392.0   Mean    : 6.212   Mean    :38.43
##   3rd Qu.: 5.990   3rd Qu.:398.1   3rd Qu.: 9.000   3rd Qu.:48.60
##   Max.    :69.880   Max.    :428.3   Max.    :21.000   Max.    :89.43
##   antenna_length  num_spots   population   dispersal_distance
##   Min.      :0.350   Min.      : 2.000   Length:10000   Min.      :21.84
##   1st Qu.:3.140   1st Qu.: 4.000   Class :character   1st Qu.:24.15
##   Median :3.850   Median : 6.000   Mode  :character   Median :24.67
##   Mean    :4.375   Mean    : 5.755               Mean    :24.67
##   3rd Qu.:5.800   3rd Qu.: 7.000               3rd Qu.:25.19
##   Max.    :7.670   Max.    :18.000               Max.    :27.79
##   body_length     sample_id
##   Min.      : 1.000   Length:10000
##   1st Qu.: 5.100   Class :character
##   Median : 6.450   Mode  :character
##   Mean    : 6.773
##   3rd Qu.: 8.480
##   Max.    :14.590

str(butterfly)

## Classes 'tbl_df', 'tbl' and 'data.frame': 10000 obs. of 14 variables:
## $ wing_length      : num 14.1 24.5 21.3 16.2 15.5 ...
## $ wing_width       : num 6.56 11 8.15 5.84 6.72 ...
## $ age              : num 40 38 25 13 43 9 36 23 37 26 ...
## $ num_offspring    : num 33 33 29 23 35 20 35 29 33 30 ...
```

```
## $ feeding_range      : num  10.78 8.58 3.86 3.14 13.07 ...
## $ color_peak         : num  402 387 373 407 399 ...
## $ num_mates          : num   4 8 8 4 3 8 11 3 0 7 ...
## $ avg_scale_size     : num  27.7 41.6 36.2 34.1 29.8 ...
## $ antenna_length     : num   3.05 5.48 4.85 3.68 3.57 5.76 6.29 2.05 2.49 5.54 ...
## $ num_spots          : num   7 4 6 8 7 4 4 9 10 4 ...
## $ population         : chr   "Ternate" "Tidore" "Kayoa" "Ternate" ...
## $ dispersal_distance : num  25.7 24.3 23.1 25.9 25.2 ...
## $ body_length        : num   6.91 8.16 7.03 3.56 7.12 8.42 8.19 5.02 4.36 9.86 ...
## $ sample_id          : chr   "Ter_00001_ZW" "Tid_00002_ZW" "Kay_00003_ZB" "Ter_00004_ZW" ...
```

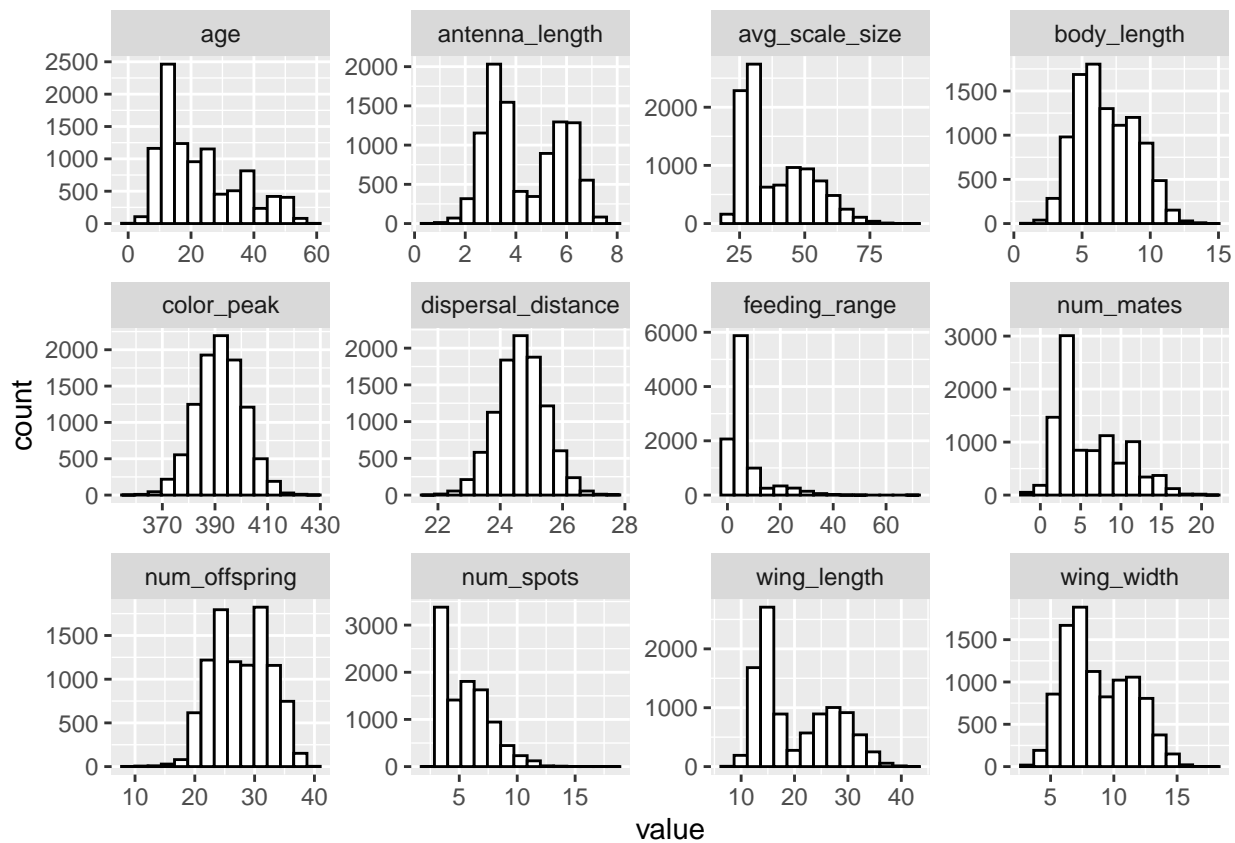
The only real change we need to make is to convert the `population` variable into a factor, since the functions provided in `readr` do not coerce strings to factors by default.

```
butterfly$population <- as.factor(butterfly$population)
summary(butterfly$population)
```

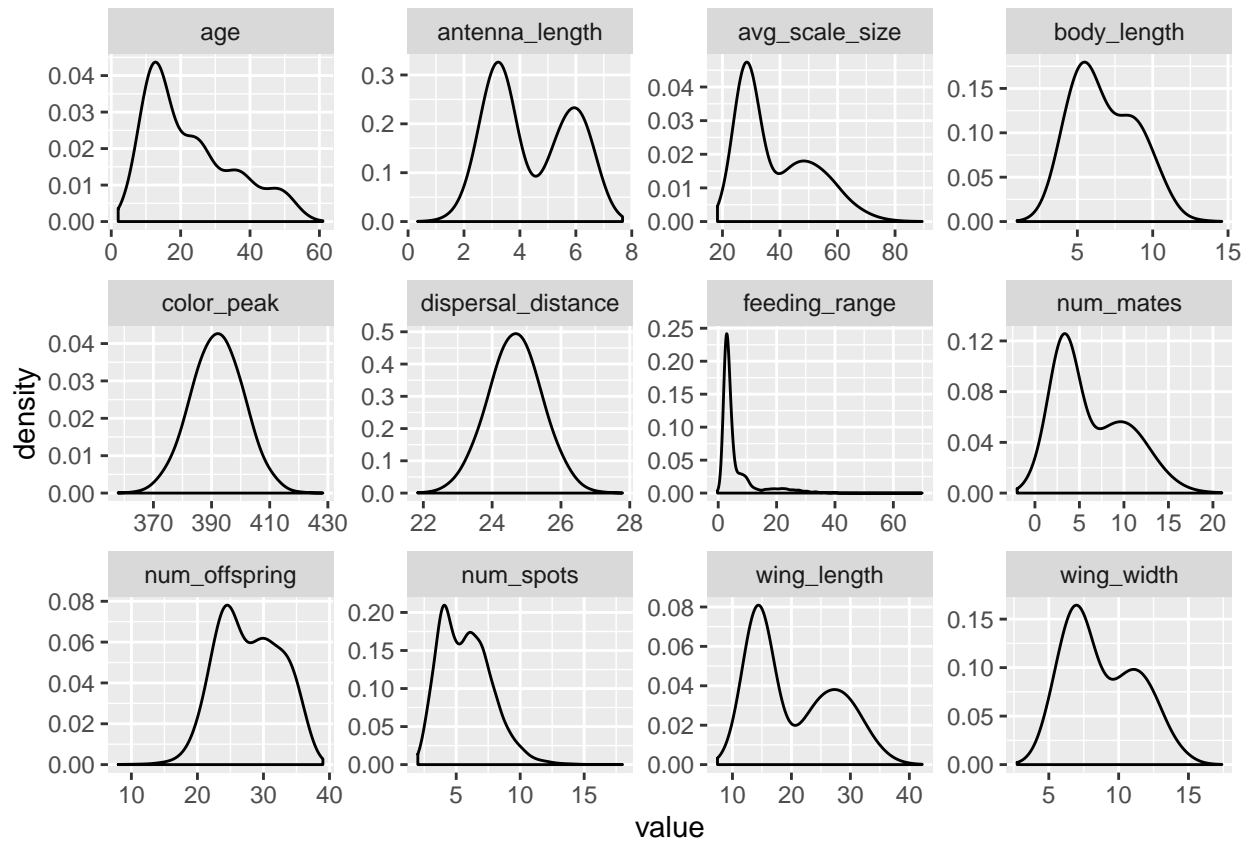
```
##   Kayoa Ternate Tidore
##    1322    5486    3192
```

So, now we can start exploring our data. Let's start by making histograms of all the numeric data.

```
num_bins <- ceiling(log2(nrow(butterfly))) + 1
butterfly %>%
  select(-population) %>%
  gather(key = "field", value = "value", -"sample_id") %>%
  ggplot(aes(x = value)) +
  geom_histogram(col = "black",
                 fill = "white",
                 position = "identity",
                 bins = num_bins) +
  facet_wrap(~field, scales = "free")
```



```
# An alternative way to view the data
butterfly %>%
  select(-population) %>%
  gather(key = "field", value = "value", -"sample_id") %>%
  ggplot(aes(x = value)) +
  geom_density(adjust = 2) +
  facet_wrap(~field, scales = "free")
```

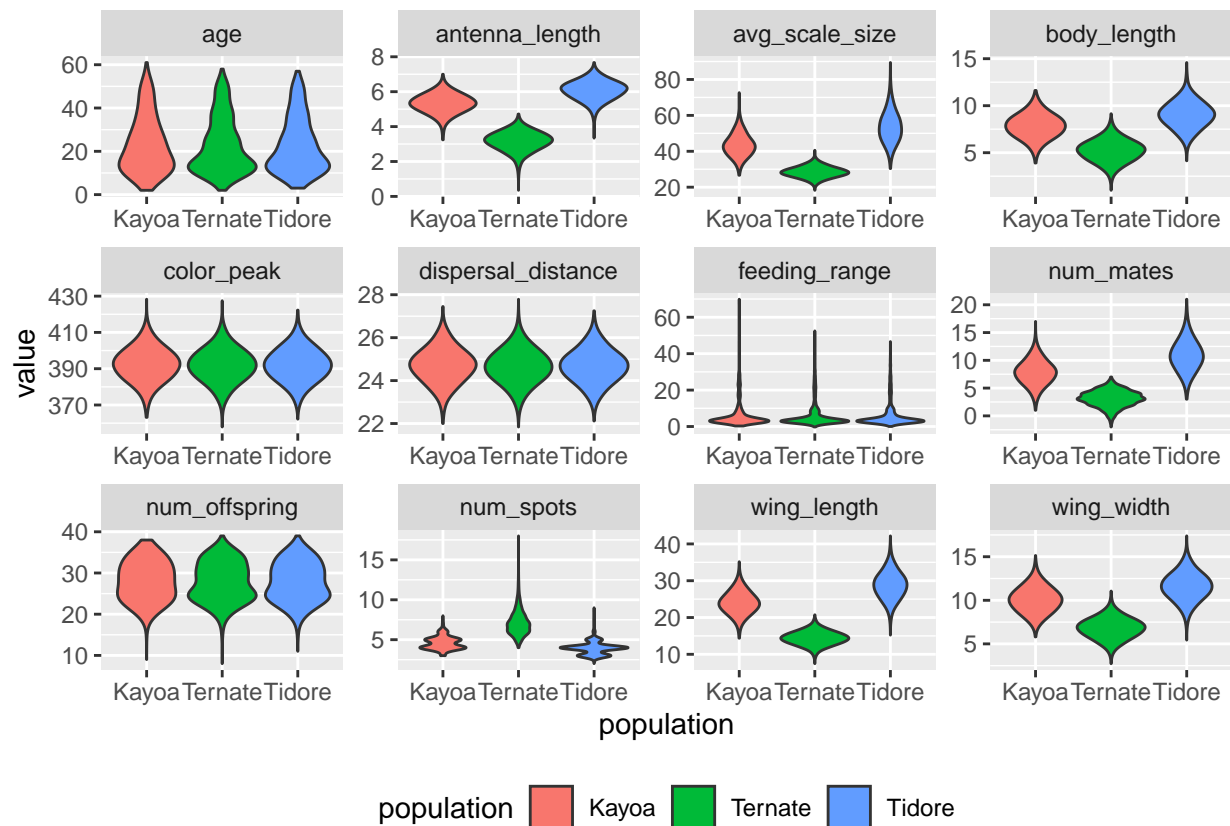


Now we have one categorical variable, so let's look at all of our data stratified by the `population` value.

```

butterfly %>%
  gather(key = "field", value = "value", -c(sample_id, population)) %>%
  ggplot(aes(x = population, y = value, fill = population)) +
  geom_violin(adjust = 2) +
  facet_wrap(~field, scales = "free") +
  theme(legend.position = "bottom")

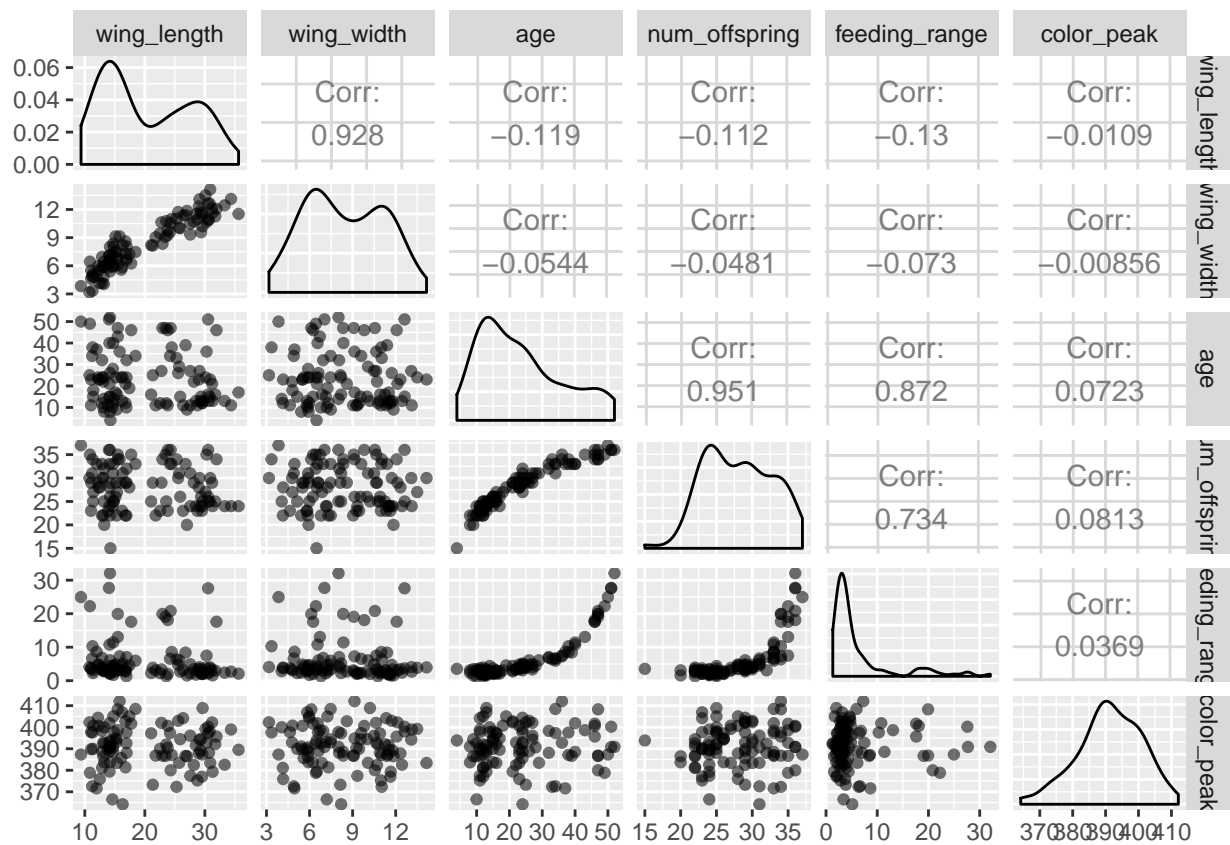
```



Now, using the `GGally` package, we can also make a scatterplot matrix like we did with `graphics::pairs()`.

```
library(GGally)
```

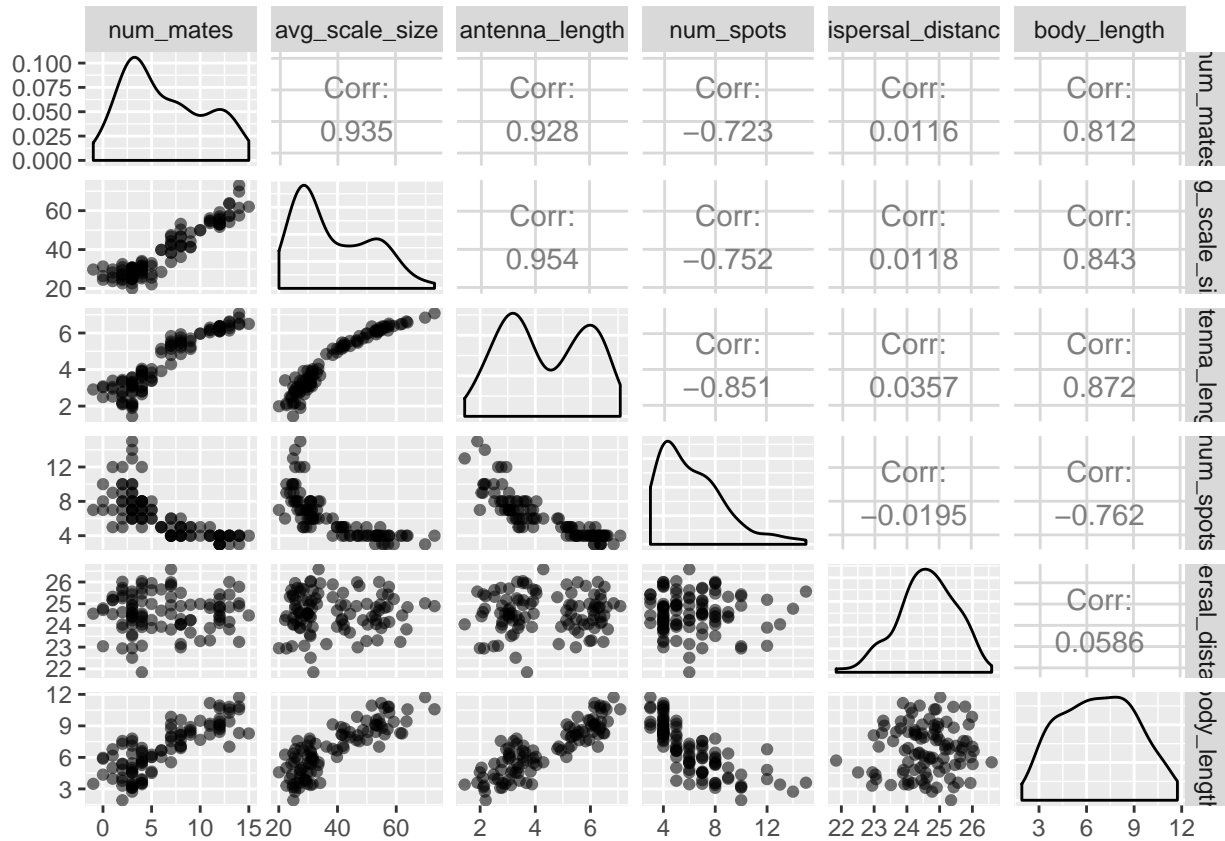
```
butterfly[1:100, 1:6] %>%  
  ggpairs(aes(alpha = 0.2))
```



```

butterfly[1:100, 7:14] %>%
  select(-c(sample_id, population)) %>%
  ggpairs(aes(alpha = 0.2))

```



As you can probably see, this visualization is not ideal when we have a lot of data.

Let's try a correlation table as well. However, note that while a correlation table can give us a good sense of linear relationships, we lose any information we had about nonlinear relationships, which we have to examine visually if we don't have a hypothesis about their existence.

```
library(pander)
butterfly %>%
  select(-c(sample_id, population)) %>%
  cor() %>%
  pander
```

Table 1: Table continues below

	wing_length	wing_width	age	num_offspring
wing_length	1	0.9226	-0.001907	-0.002996
wing_width	0.9226	1	0.001328	-0.001489
age	-0.001907	0.001328	1	0.9413
num_offspring	-0.002996	-0.001489	0.9413	1
feeding_range	-0.004469	-0.00217	0.8559	0.7075
color_peak	0.01542	0.02052	0.00448	0.0009257
num_mates	0.9495	0.8747	-0.002376	-0.003543
avg_scale_size	0.9799	0.903	0.000683	-0.0002265
antenna_length	0.9906	0.9143	-0.002996	-0.00398
num_spots	-0.8294	-0.9372	0.00375	0.005168
dispersal_distance	0.01612	0.02096	0.0007418	-0.003065
body_length	0.8823	0.8511	-0.0002991	-0.002555

Table 2: Table continues below

	feeding_range	color_peak	num_mates
wing_length	-0.004469	0.01542	0.9495
wing_width	-0.00217	0.02052	0.8747
age	0.8559	0.00448	-0.002376
num_offspring	0.7075	0.0009257	-0.003543
feeding_range	1	0.01034	-0.004402
color_peak	0.01034	1	0.009925
num_mates	-0.004402	0.009925	1
avg_scale_size	-0.001537	0.01469	0.942
antenna_length	-0.005578	0.01548	0.9281
num_spots	0.005558	-0.01968	-0.769
dispersal_distance	0.006433	0.9468	0.01057
body_length	-0.0007702	0.01103	0.837

Table 3: Table continues below

	avg_scale_size	antenna_length	num_spots
wing_length	0.9799	0.9906	-0.8294
wing_width	0.903	0.9143	-0.9372
age	0.000683	-0.002996	0.00375
num_offspring	-0.0002265	-0.00398	0.005168
feeding_range	-0.001537	-0.005578	0.005558
color_peak	0.01469	0.01548	-0.01968
num_mates	0.942	0.9281	-0.769
avg_scale_size	1	0.9554	-0.7921
antenna_length	0.9554	1	-0.8447
num_spots	-0.7921	-0.8447	1
dispersal_distance	0.01492	0.01626	-0.02037
body_length	0.8639	0.875	-0.7775

	dispersal_distance	body_length
wing_length	0.01612	0.8823
wing_width	0.02096	0.8511
age	0.0007418	-0.0002991
num_offspring	-0.003065	-0.002555
feeding_range	0.006433	-0.0007702
color_peak	0.9468	0.01103
num_mates	0.01057	0.837
avg_scale_size	0.01492	0.8639
antenna_length	0.01626	0.875
num_spots	-0.02037	-0.7775
dispersal_distance	1	0.01044
body_length	0.01044	1