EDA-challenge.Rmd

2024-02-07

```
library(curl)
```

```
## Using libcurl 8.3.0 with Schannel
```

```
f <- curl("https://raw.githubusercontent.com/difiore/ada-2024-datasets/main/data-wrangling.csv")
d <-  read.csv(f, header = TRUE, sep = ",", stringsAsFactors = FALSE) #Loading "data-wrangling.csv" as
```

```
names(d) #Looking over the variables it contains
```

```
##  [1] "Scientific_Name"         "Family"
##  [3] "Genus"                   "Species"
##  [5] "Brain_Size_Species_Mean" "Body_mass_male_mean"
##  [7] "Body_mass_female_mean"   "MeanGroupSize"
##  [9] "AdultMales"              "AdultFemale"
## [11] "GR_MidRangeLat_dd"       "Precip_Mean_mm"
## [13] "Temp_Mean_degC"          "HomeRange_km2"
## [15] "DayLength_km"            "Fruit"
## [17] "Leaves"                  "Fauna"
## [19] "Canine_Dimorphism"       "Feed"
## [21] "Move"                    "Rest"
## [23] "Social"
```

Step 1

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v ggplot2   3.4.4     v tibble    3.2.1
## v lubridate 1.9.3     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter()     masks stats::filter()
## x dplyr::lag()        masks stats::lag()
## x readr::parse_date() masks curl::parse_date()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

1. Here, I'm creating a new variable named BSD (body size dimorphism) which is the ratio of average male to female body mass.

```
d$BSD <- d$Body_mass_male_mean/d$Body_mass_female_mean
```

2. Here, I'm creating a new variable named sex_ratio, which is the ratio of the number of adult females to adult males in a typical group.

```
d$sex_ratio <- d$AdultFemale/d$AdultMales
```

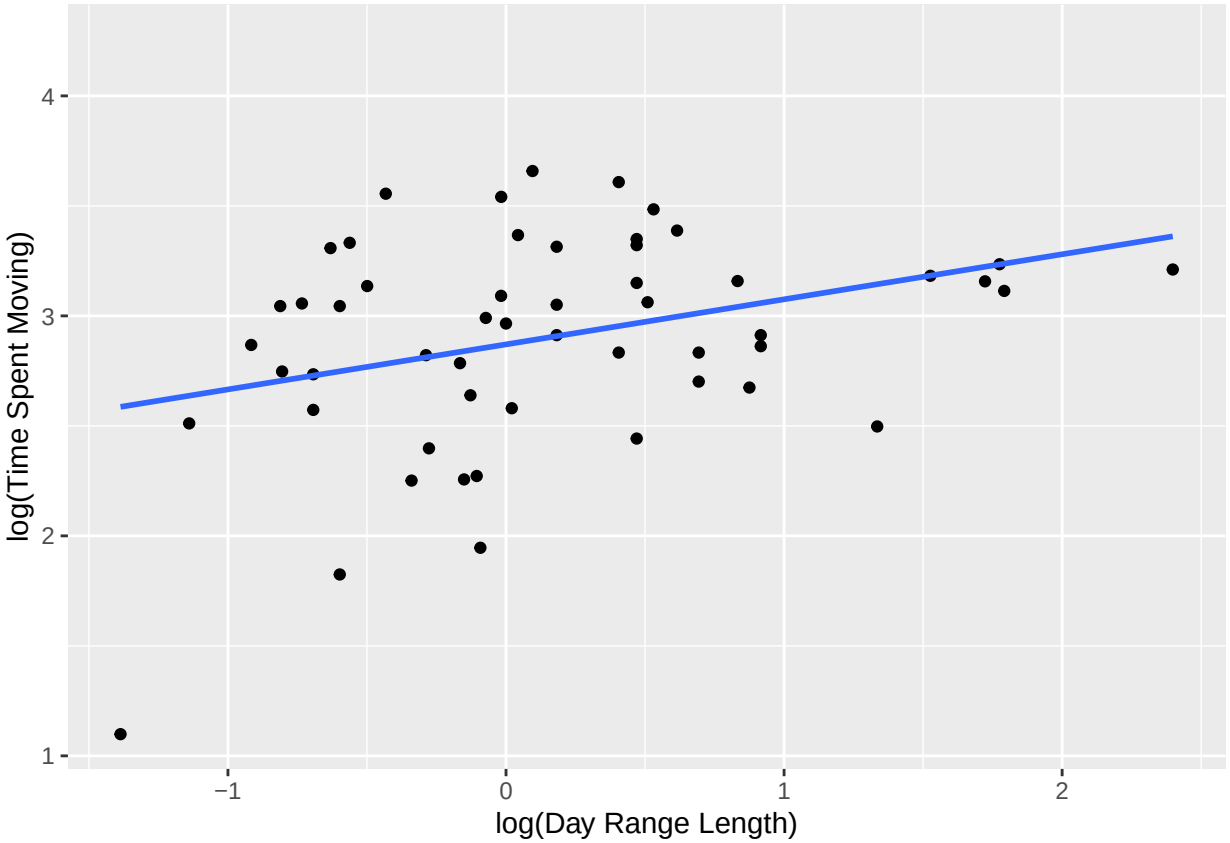
3. Here, I'm creating a new variable named DI (for "defensibility index"), which is the ratio of day range length to the diameter of the home range.

```
d$DI <- d$DayLength_km/(2*(sqrt((d$HomeRange_km2)/pi)))
```

4. Here, I'm plotting the relationship between log(day range length) and log(time spent moving) for these primate species overall.

```
library(dplyr)
library(tidyverse)
library(ggplot2)
p <- ggplot(data = d, aes( #Building plot object
  x = log(DayLength_km), #I am log-transforming the variables to reduce the skew of the distribution
  y = log(Move),
))
p <- p +
  xlab("log(Day Range Length)") +
  ylab("log(Time Spent Moving)") + #Specifying axis labels
  geom_point(na.rm = TRUE) + #Creating my scatterplot
  geom_smooth(method = "lm", se=FALSE, na.rm = TRUE) #Adding a linear regression model
p #Plotting the object
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

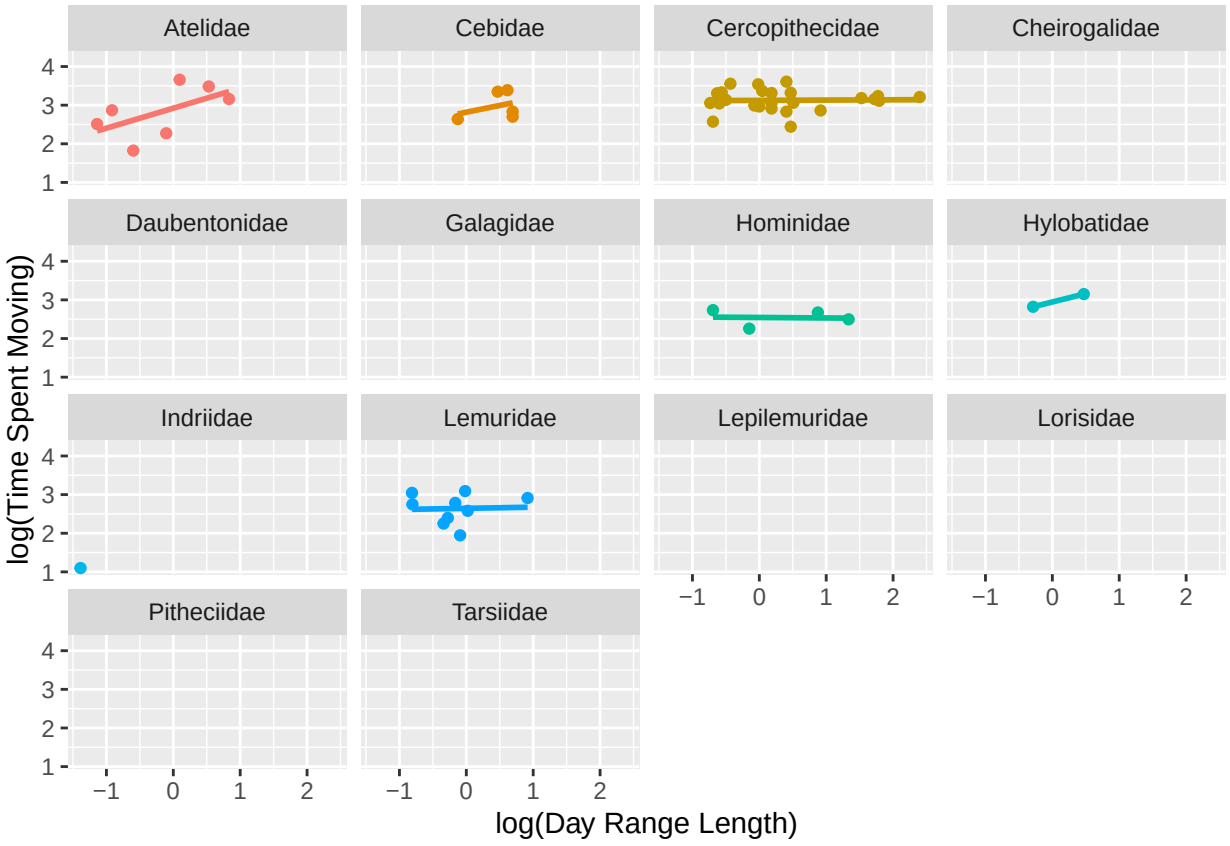


Looking at these primate species overall, there does not appear to be a strong correlation between time spent moving and day range length.

Now, I'm plotting the relationship between log(day range length) and log(time spent moving) for these these primate species by family.

```
p <- ggplot(data = d, aes(
  x = log(DayLength_km),
  y = log(Move),
  color = factor(Family) #Coloring points by family
))
p <- p +
  xlab("log(Day Range Length)") +
  ylab("log(Time Spent Moving)") +
  geom_point(na.rm = TRUE) +
  facet_wrap(~Family, ncol = 4) + #I am wrapping the data by family (14 total) and arranging these subs
  theme(legend.position = "none") + #Because of the above, I do not need a legend to clarify the color
  geom_smooth(method = "lm", fullrange = FALSE, se=FALSE, na.rm = TRUE)#I'm going to add a linear regre
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

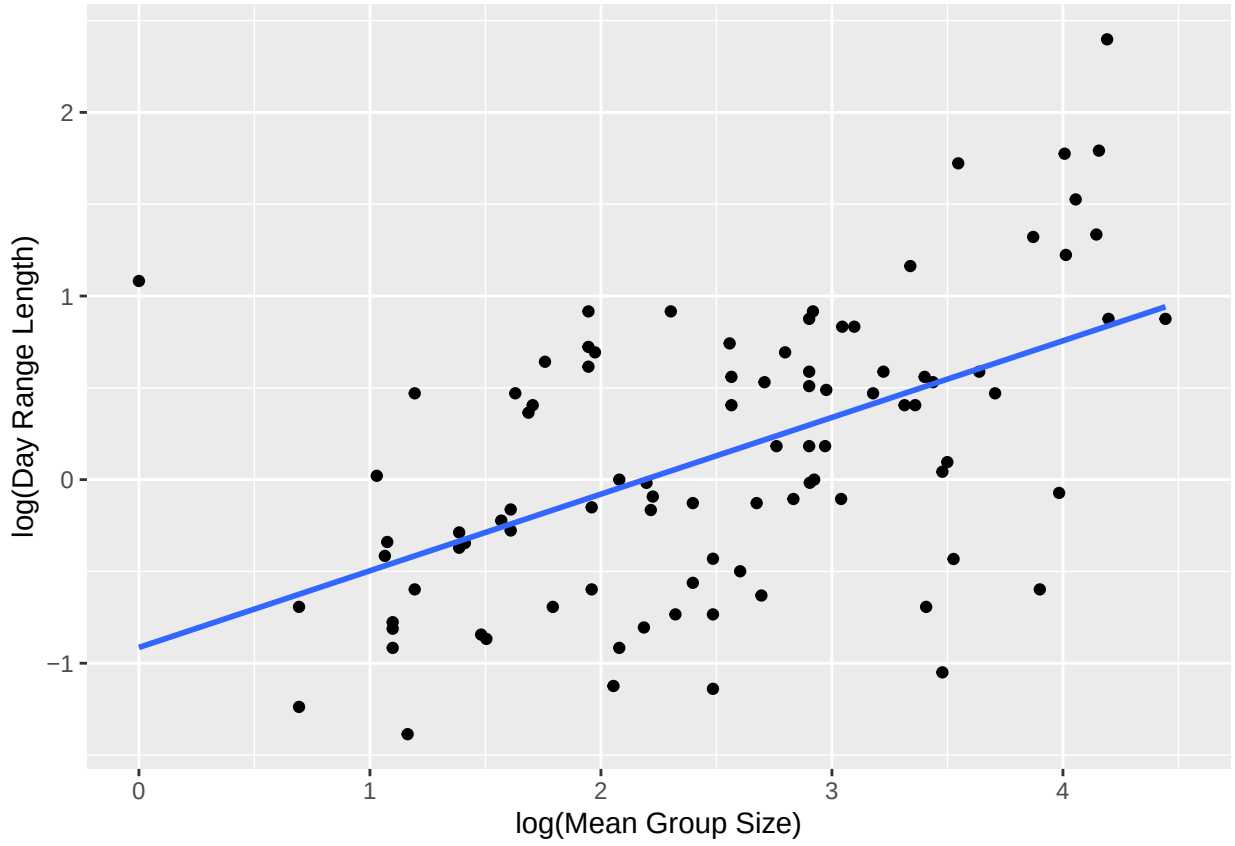


Within Atelidae, Cebidae, and Hylobatidae, it does appear that there could be a trend for species that spend more time moving to travel farther.

5. Here, I'm plotting the relationship between log(day range length) and log(mean group size) for these primate species overall.

```
p <- ggplot(data = d, aes(
  x = log(MeanGroupSize),
  y = log(DayLength_km),
))
p <- p +
  xlab("log(Mean Group Size)") +
  ylab("log(Day Range Length)") +
  geom_point(na.rm = TRUE) +
  geom_smooth(method = "lm", se=FALSE, na.rm = TRUE)
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

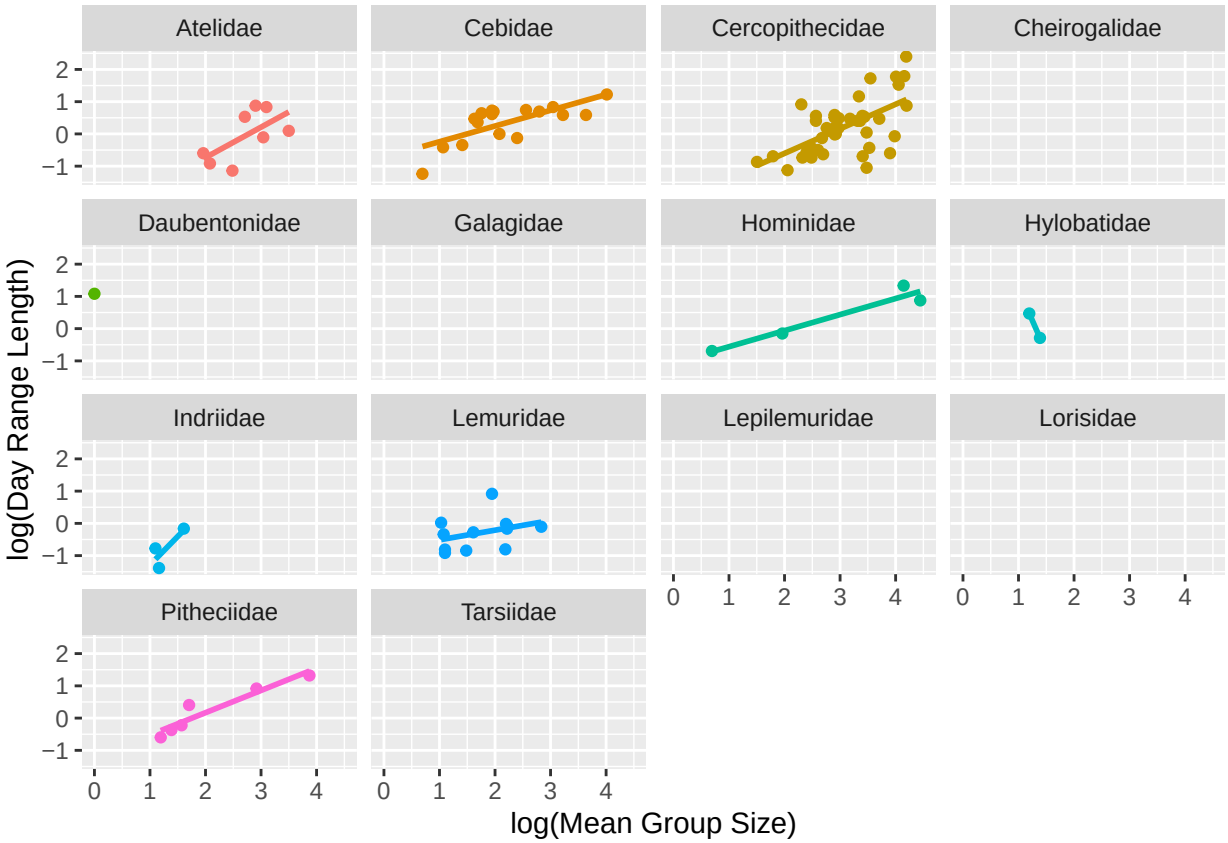


Looking at these primate species overall, there does not appear to be a strong correlation between mean group size and day range length.

Now, I'm plotting the relationship between log(day range length) and log(mean group size) by family.

```
p <- ggplot(data = d, aes(
  x = log(MeanGroupSize),
  y = log(DayLength_km),
  color = factor(Family)
))
p <- p +
  xlab("log(Mean Group Size)") +
  ylab("log(Day Range Length)") +
  geom_point(na.rm = TRUE) +
  facet_wrap(~Family, ncol = 4) +
  theme(legend.position = "none") +
  geom_smooth(method = "lm", fullrange = FALSE, se=FALSE, na.rm = TRUE)
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

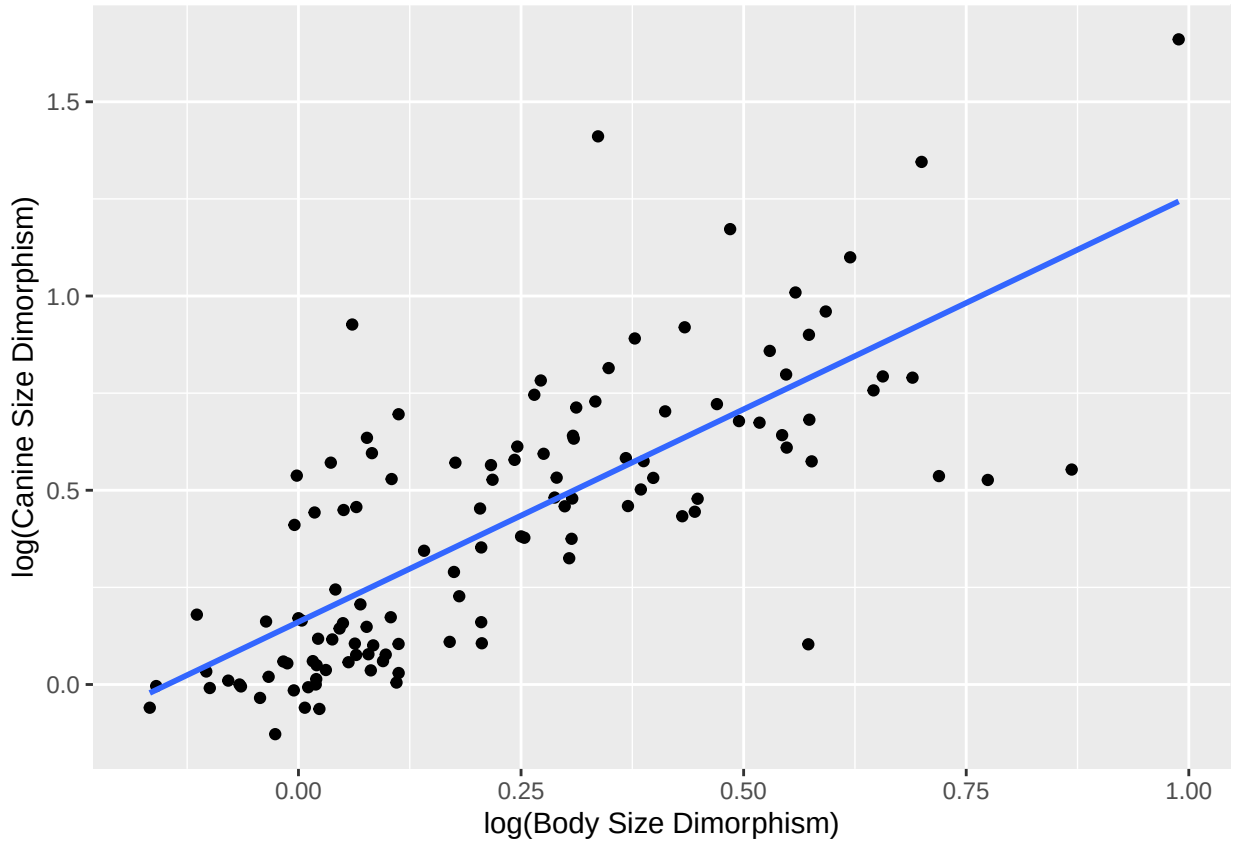


Particularly within Atelidae, Cebidae, Cercopithecidae, Hominidae, and Pitheciidae, it does appear that there could be a trend for species that live in larger groups to travel farther.

6. Here, I'm plotting the relationship between log(body size dimorphism) and log(canine size dimorphism) for these primate species overall.

```
p <- ggplot(data = d, aes(
  x = log(BSD),
  y = log(Canine_Dimorphism),
))
p <- p +
  xlab("log(Body Size Dimorphism)") +
  ylab("log(Canine Size Dimorphism)") +
  geom_point(na.rm = TRUE) +
  geom_smooth(method = "lm", se=FALSE, na.rm = TRUE)
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```

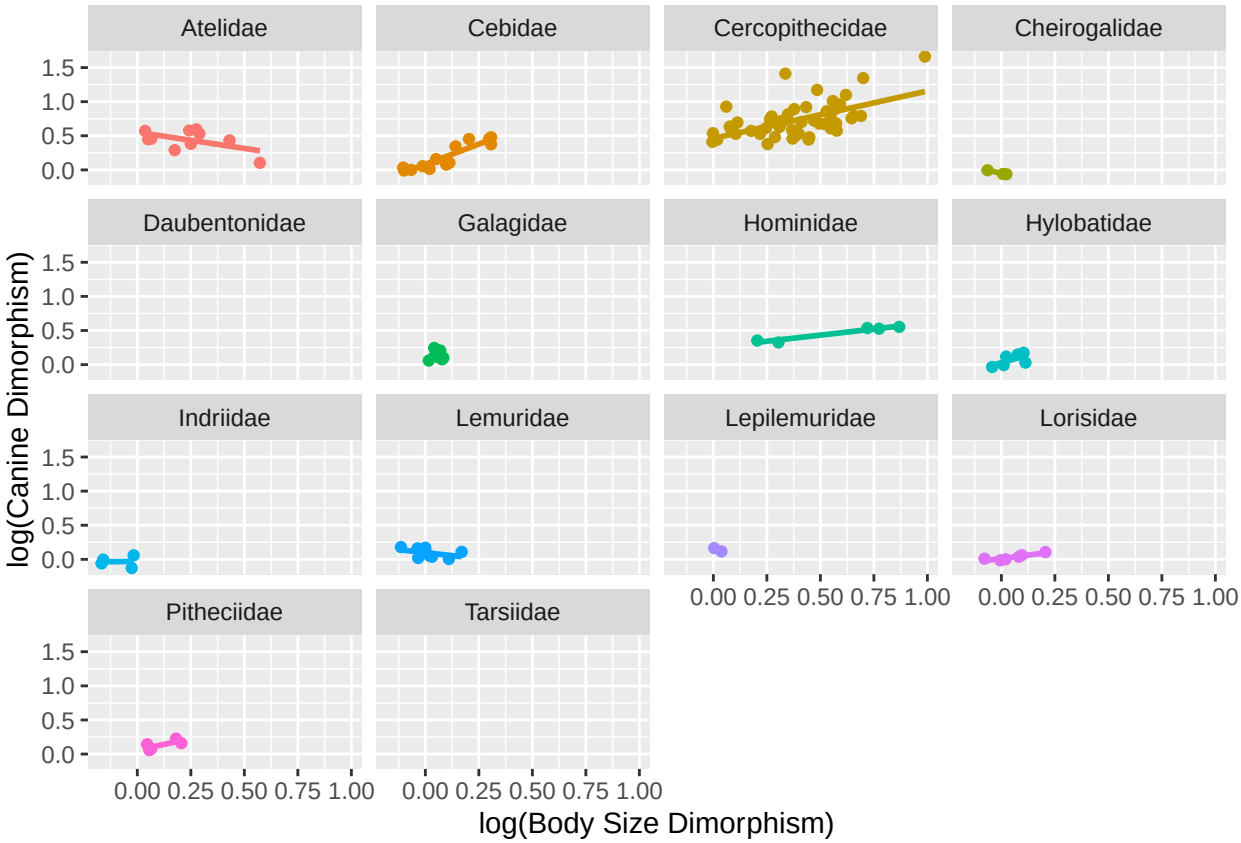


Looking at these primate species overall, it does appear that there could be a trend for species with greater body size dimorphism to also show greater canine size dimorphism.

Now, I'm plotting the relationship between log(body size dimorphism) and log(canine size dimorphism) by family.

```
p <- ggplot(data = d, aes(
  x = log(BSD),
  y = log(Canine_Dimorphism),
  color = factor(Family)
))
p <- p +
  xlab("log(Body Size Dimorphism)") +
  ylab("log(Canine Dimorphism)") +
  geom_point(na.rm = TRUE) +
  facet_wrap(~Family, ncol = 4) +
  theme(legend.position = "none") +
  geom_smooth(method = "lm", fullrange = FALSE, se=FALSE, na.rm = TRUE)
p
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



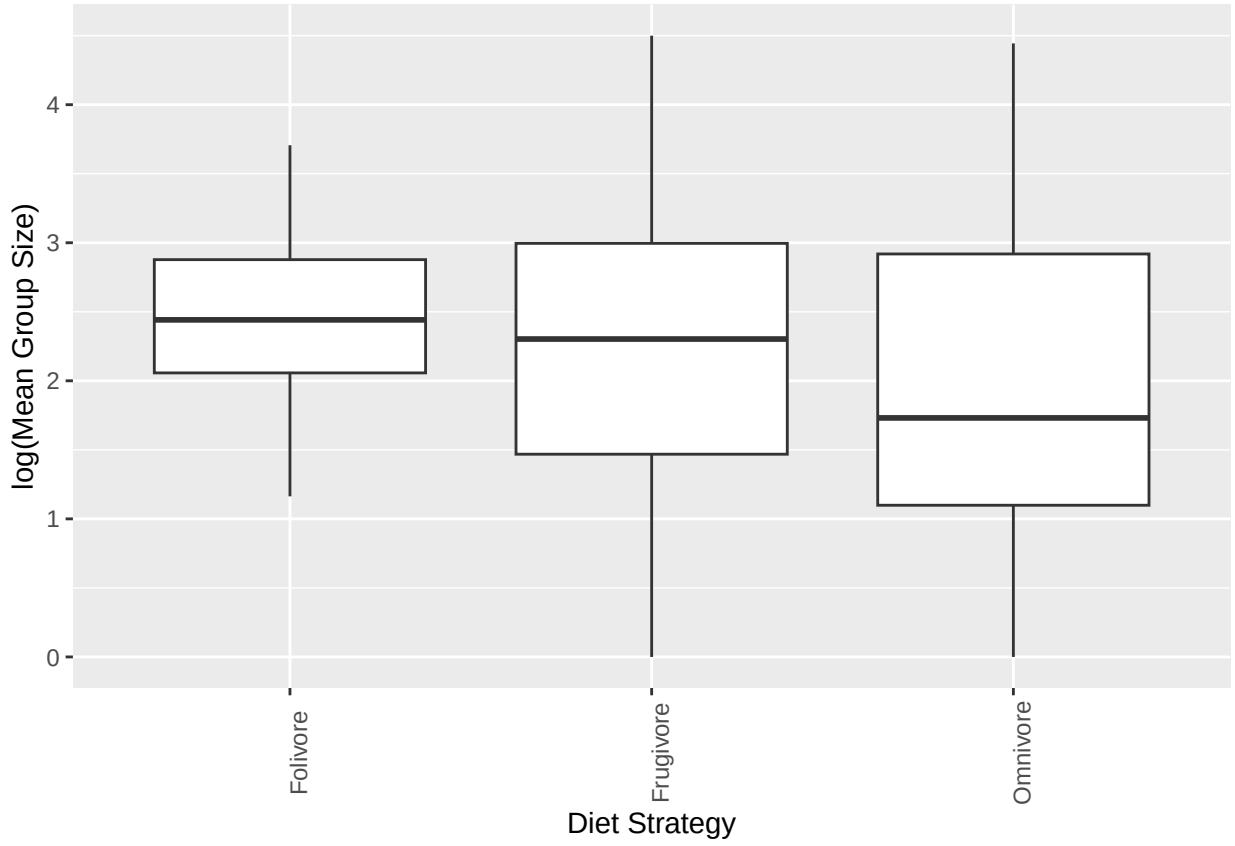
Particularly within Cebidae and Cercopithecidae, it does appear that there could be a trend for species that exhibit greater body size dimorphism to also exhibit greater canine dimorphism.

7. Here, I'm creating a new variable named diet_strategy.

```
d <- d|> mutate(diet_strategy = case_when( #If...
  Fruit > 50.0  ~ "Frugivore", #...fruit makes up >50% of diet, then designate species as "frugivore"
  Leaves > 50.0 ~ "Folivore", #...leaves make up >50% of diet, then designate species as "folivore"
  TRUE ~ "Omnivore", #...neither case is true, designate all others as omnivores
))
```

Now I'm creating boxplots of group size for species with different dietary strategies

```
p <- ggplot(data = d, aes(x = diet_strategy, y = log(MeanGroupSize))) +
  geom_boxplot(na.rm = TRUE) +
  theme(axis.text.x = element_text(angle = 90)) + #Specifying axis labels
  ylab("log(Mean Group Size)") +
  xlab("Diet Strategy")
p
```



The mean group size for all frugivore taxa is less than the mean group size for all folivore taxa.

8.

```
(a <- mutate(d, Binomial = paste(Genus, Species, sep = ",")) |> #creating a new variable, Binomial, whi
  select(Binomial, Family, Brain_Size_Species_Mean, Body_mass_male_mean) |> #Trimming the data frame to
  group_by(Family) |> #Grouping by family
  summarise(avgBrain_Size_Species_Mean = mean(Brain_Size_Species_Mean, na.rm = TRUE), avgBody_mass_male_
  arrange(avgBrain_Size_Species_Mean)) #Arranging by increasing average brain size
```

```
## # A tibble: 14 x 3
##    Family          avgBrain_Size_Species_Mean avgBody_mass_male_mean
##    <chr>                                <dbl>                  <dbl>
##  1 Tarsiidae                             3.26                   3.26
##  2 Cheirogalidae                         4.04                   4.04
##  3 Galagidae                             5.96                   5.96
##  4 Lepilemuridae                         7.27                   7.27
##  5 Lorisidae                             8.67                   8.67
##  6 Lemuridae                            23.1                   23.1
##  7 Cebidae                              23.9                   23.9
##  8 Indriidae                            27.3                   27.3
##  9 Daubentonidae                        44.8                   44.8
## 10 Pitheciidae                          56.3                   56.3
## 11 Atelidae                             80.6                   80.6
## 12 Cercopithecidae                      85.4                   85.4
## 13 Hylobatidae                         101.                   101.
## 14 Hominidae                           410.                   410.
```