EDA-challange

STEP1

```
library(tidyverse)
```

```
Warning: package 'tidyverse' was built under R version 4.3.3

Warning: package 'ggplot2' was built under R version 4.3.1

Warning: package 'lubridate' was built under R version 4.3.3

-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.4     v tibble    3.2.1
v lubridate 1.9.4     v tidyr     1.3.0
v purrr     1.0.1
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becor
```

```
f <- "https://raw.githubusercontent.com/difiore/ada-datasets/main/data-wrangling.csv"
d <- read_csv(f, col_names = TRUE )
```

```
Rows: 213 Columns: 23
-- Column specification --------------------------------------------------------
Delimiter: ","
chr  (6): Scientific_Name, Family, Genus, Species, Leaves, Fauna
dbl (17): Brain_Size_Species_Mean, Body_mass_male_mean, Body_mass_female_mea...

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.
```

```
colnames(d)
```

```
 [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"
```

```
glimpse(d)
```

```
Rows: 213
Columns: 23
$ Scientific_Name          <chr> "Allenopithecus_nigroviridis", "Allocebus_tric~
$ Family                   <chr> "Cercopithecidae", "Cercopithecidae", "Atelida~
$ Genus                    <chr> "Allenopithecus", "Allocebus", "Alouatta", "Al~
$ Species                  <chr> "nigroviridis", "trichotis", "belzebul", "cara~
$ Brain_Size_Species_Mean <dbl> 58.02, NA, 52.84, 52.63, 51.70, 49.88, 51.13, ~
$ Body_mass_male_mean      <dbl> 6130.0, 92.0, 7270.0, 6525.0, 5800.0, 7150.0, ~
$ Body_mass_female_mean    <dbl> 3180.0, 84.0, 5520.0, 4240.0, 4550.0, 5350.0, ~
$ MeanGroupSize            <dbl> NA, 1.000, 7.000, 8.000, 6.530, 12.000, 6.600,~
$ AdultMales               <dbl> NA, 1.000, 1.000, 2.300, 1.370, 2.900, 1.925, ~
$ AdultFemale              <dbl> NA, 1.000, 1.000, 3.300, 2.200, 6.300, 2.175, ~
$ GR_MidRangeLat_dd        <dbl> -0.17, -16.59, -6.80, -20.34, -21.13, 6.95, 18~
$ Precip_Mean_mm           <dbl> 1574.0, 1902.3, 1643.5, 1166.4, 1332.3, 1852.6~
$ Temp_Mean_degC           <dbl> 25.2, 20.3, 24.9, 22.9, 19.6, 23.7, 25.1, 25.1~
$ HomeRange_km2            <dbl> NA, NA, NA, NA, 0.030, 0.190, 0.300, 0.100, 0.~
$ DayLength_km             <dbl> NA, NA, NA, 0.400, NA, 0.320, NA, 0.550, NA, N~
$ Fruit                    <dbl> NA, NA, 57.3, 23.8, 5.2, 33.1, 40.8, 40.0, 45.~
$ Leaves                   <chr> NA, NA, "19.1", "67.7", "73", "56.4", "45.1", ~
$ Fauna                    <chr> NA, NA, "0", "0", "0", "0", "0", "0", NA, NA, ~
$ Canine_Dimorphism        <dbl> 2.210, NA, 1.811, 1.542, 1.783, 1.703, 1.109, ~
$ Feed                     <dbl> NA, NA, 13.75, 15.90, 18.33, 17.94, 24.40, 12.~
$ Move                     <dbl> NA, NA, 18.75, 17.60, 14.33, 12.32, 9.80, 6.20~
$ Rest                     <dbl> NA, NA, 57.30, 61.60, 64.37, 66.14, 61.90, 78.~
$ Social                   <dbl> NA, NA, 10.00, 4.90, 3.00, 3.64, 3.80, 2.50, N~
```

1,2,3

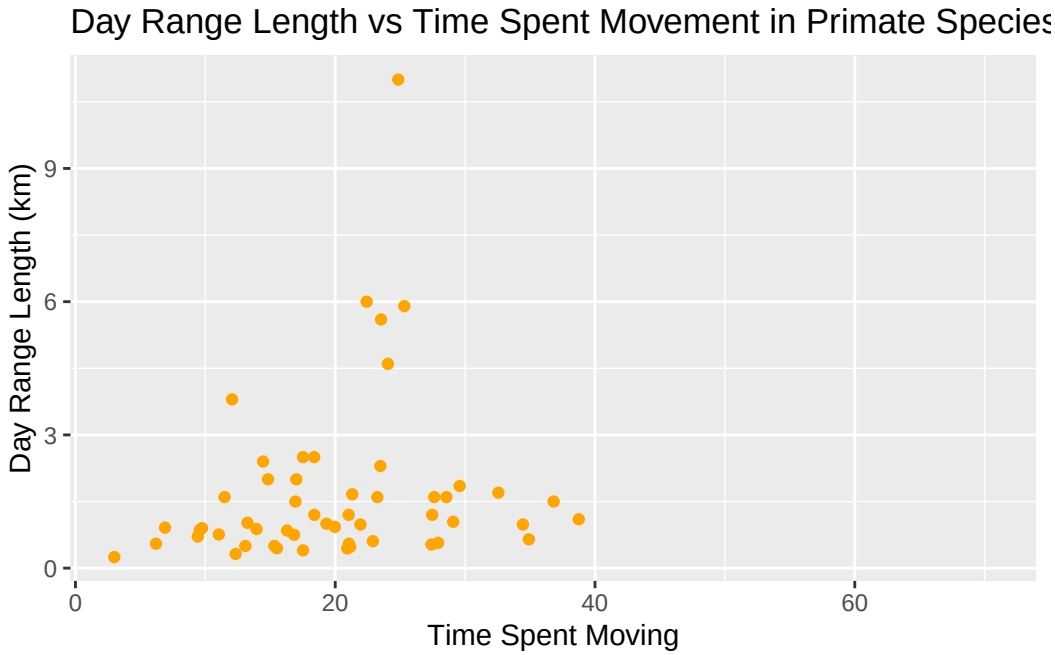
```
# 1. Create BSD
d$BSD <- d$Body_mass_male_mean / d$Body_mass_female_mean

# 2. Create Sex Ratio
d$sex_ratio <- d$AdultFemale / d$AdultMales

# 3. Create Defensibility Index
d$DI <- d$DayLength_km / d$HomeRange_km2

head(d[, c("Scientific_Name", "BSD", "sex_ratio", "DI")])
```

```
# A tibble: 6 x 4
  Scientific_Name               BSD sex_ratio    DI
  <chr>                       <dbl>     <dbl> <dbl>
1 Allenopithecus_nigroviridis  1.93     NA    NA
2 Allocebus_trichotis          1.10      1    NA
3 Alouatta_belzebul            1.32      1    NA
4 Alouatta_caraya              1.54      1.43 NA
5 Alouatta_guariba             1.27      1.61 NA
6 Alouatta_palliata            1.34      2.17  1.68
```

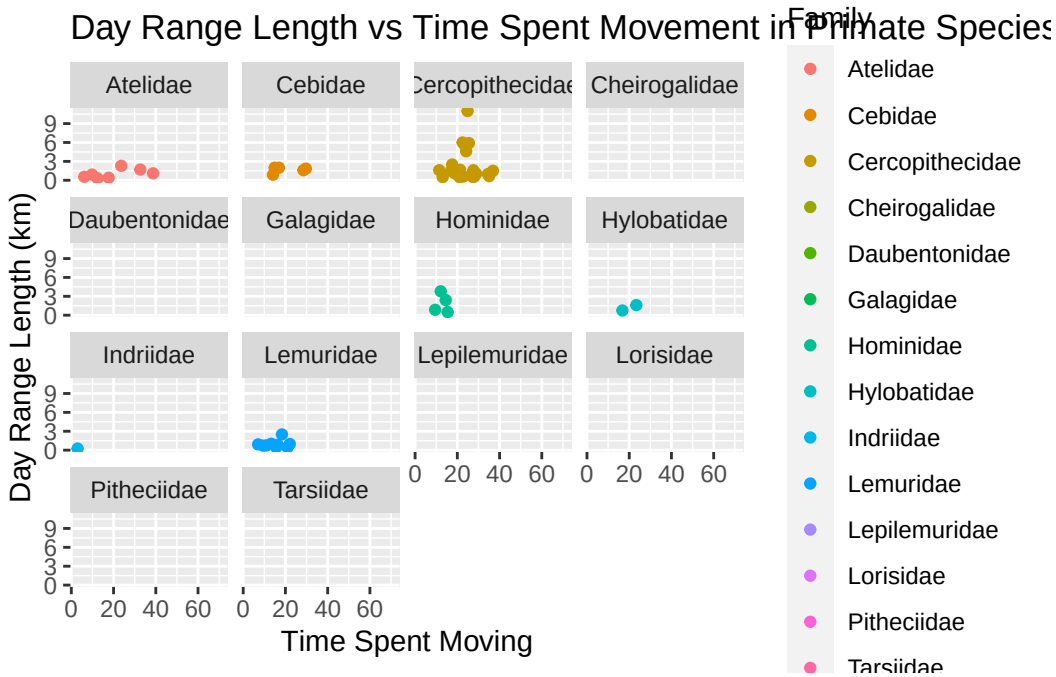
```
#homerange diameter
d$HomeRangeDiameter <- 2 * sqrt(d$HomeRange_km2 / pi)
```

4

```
# Overall
library(ggplot2)

p <- ggplot(d, aes(x = Move, y = DayLength_km)) +
  geom_jitter(color = "orange", width = 0.1, na.rm = TRUE) +
  labs(x = "Time Spent Moving", y = "Day Range Length (km)",
       title = "Day Range Length vs Time Spent Movement in Primate Species")

p
```




```r
# By family
p <- ggplot(d, aes(x = Move, y = DayLength_km, color = Family)) +
  geom_jitter(width = 0.1, na.rm = TRUE) +
  facet_wrap(~ Family) +  # Add the missing + here
  labs(x = "Time Spent Moving", y = "Day Range Length (km)",
       title = "Day Range Length vs Time Spent Movement in Primate Species")

p
```



Do species that spend more time moving travel farther overall?

Looking at the overall trend across all families, it does not seem strong relationship between time spend moving and day range length. Some families, like Cercopithecidae, show a broader spread, with some species having high day range lengths despite a range of movement times.Other families, such as Atelidae and Lemuridae, seem to have relatively short travel distances regardless of movement time.

How about within any particular primate family?

Cercopithecidae and Hominidae appear to have species that exhibit higher day range lengths when they spend more time moving.Lepilemuridae families shows very little variation in movement.It is hard to see a clear trend.

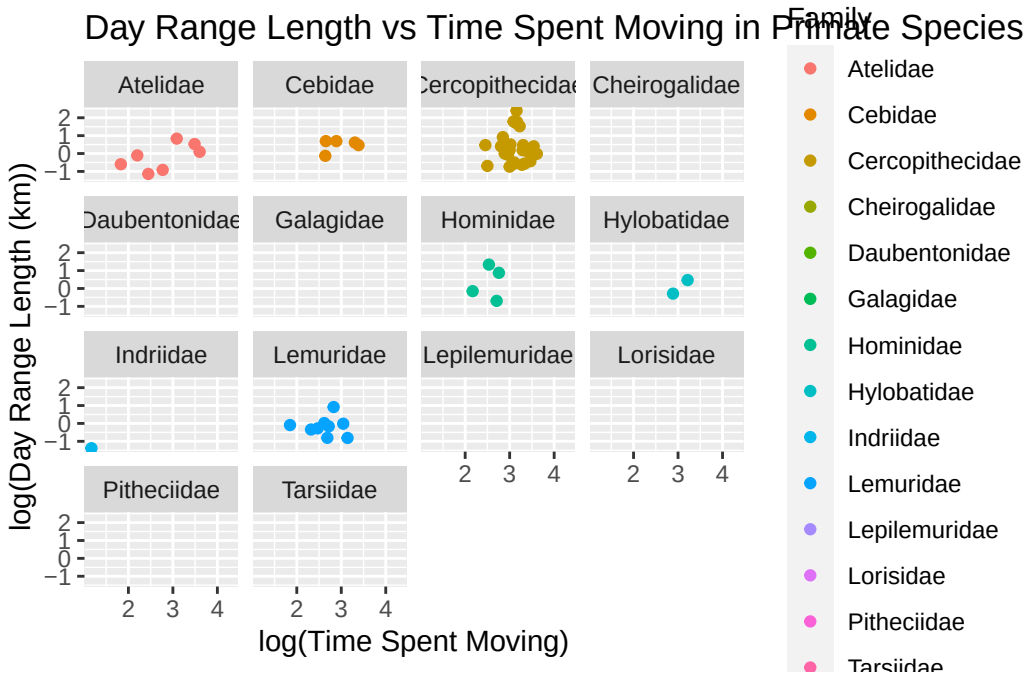
Should you transform either of these variables?

Yes, transformation could be useful, particularly if the variables are skewed.

#Plot-Transformed variables

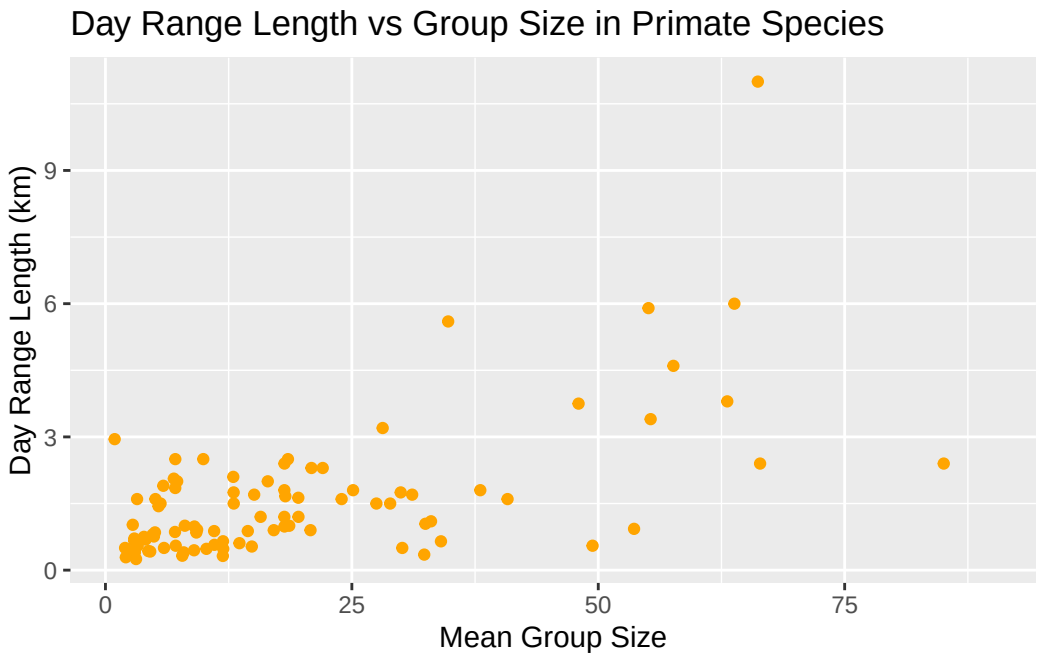
```
library(ggplot2)
p <- ggplot(d, aes(x = log(Move), y = log(DayLength_km), color = Family)) +  #to make differe
  geom_jitter(width = 0.1, na.rm = TRUE) +
  facet_wrap(~ Family) +
  labs(x = "log(Time Spent Moving)",
       y = "log(Day Range Length (km))",
```

```
       title = "Day Range Length vs Time Spent Moving in Primate Species")
p
```

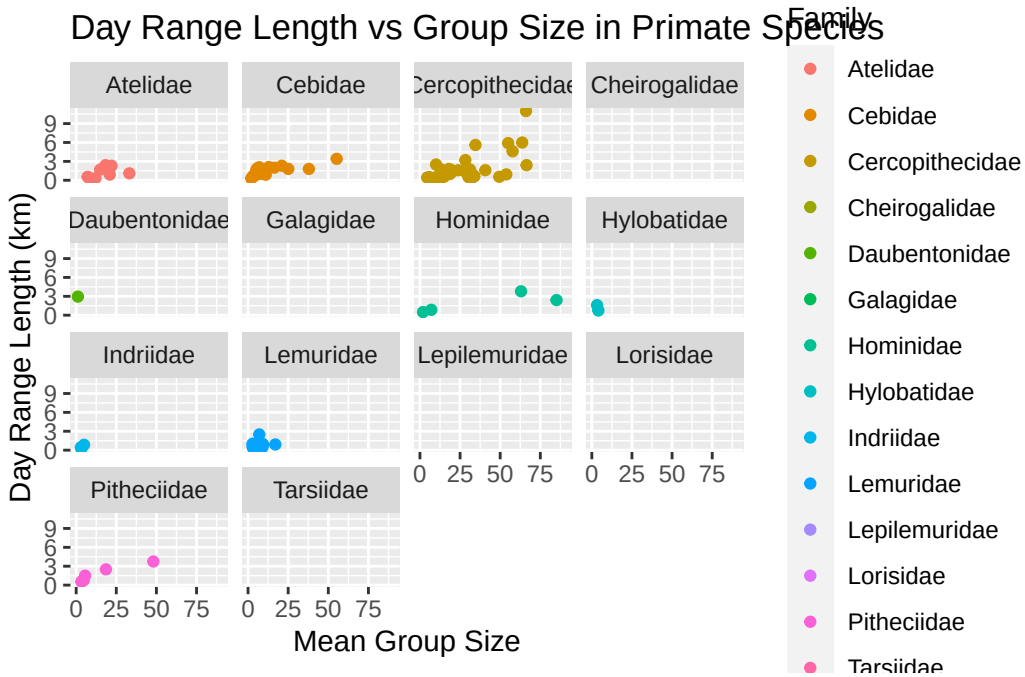


5

```
#Overall

p <- ggplot(d, aes(x = MeanGroupSize, y = DayLength_km)) +
  geom_jitter(color= "orange", width = 0.1, na.rm = TRUE) +
    labs(x = "Mean Group Size",
       y = "Day Range Length (km)",
       title = "Day Range Length vs Group Size in Primate Species")
p
```




```
#by family
p <- ggplot(d, aes(x = MeanGroupSize, y = DayLength_km, color = Family)) +
  geom_jitter(width = 0.1, na.rm = TRUE) +
  facet_wrap(~ Family) +
  labs(x = "Mean Group Size",
       y = "Day Range Length (km)",
       title = "Day Range Length vs Group Size in Primate Species")

p
```



Do species in larger groups travel farther overall?

There seems to be a positive trend, especially in some families such as Cercopithecidae and Hylobatidae, where larger group sizes are associated with longer day ranges.However, Atelidae and Pitheciidae's group size variation does not show strong correlation with day range length.This supports that group size variation mat not fully explain differences in day range alone.

How about within specific primate families?

In Cercopithecidae, there's a noticeable increase in day range length with group size.Hylobatidae, a similar trend is observed, but with fewer data points.For Atelidae, Pitheciidae, and Lemuridae, the relationship appears weaker.

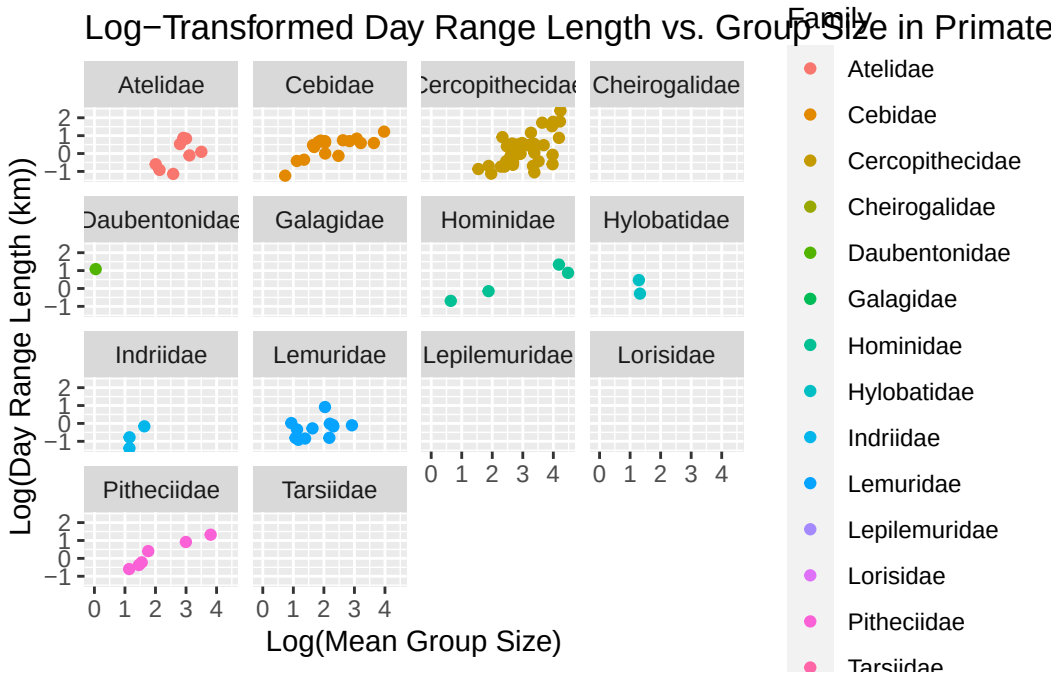
Should we transform either variable?

Day range length (km) is highly right-skewed. A log transformation helps normalize the distribution.

Plot-Transformed variables

```
p <- ggplot(d, aes(x = log(MeanGroupSize), y = log(DayLength_km), color = Family)) +
  geom_jitter(width = 0.1, na.rm = TRUE) +
  facet_wrap(~ Family) +
  labs(x = "Log(Mean Group Size)", y = "Log(Day Range Length (km))",
```

```
       title = "Log-Transformed Day Range Length vs. Group Size in Primate Species")
p
```



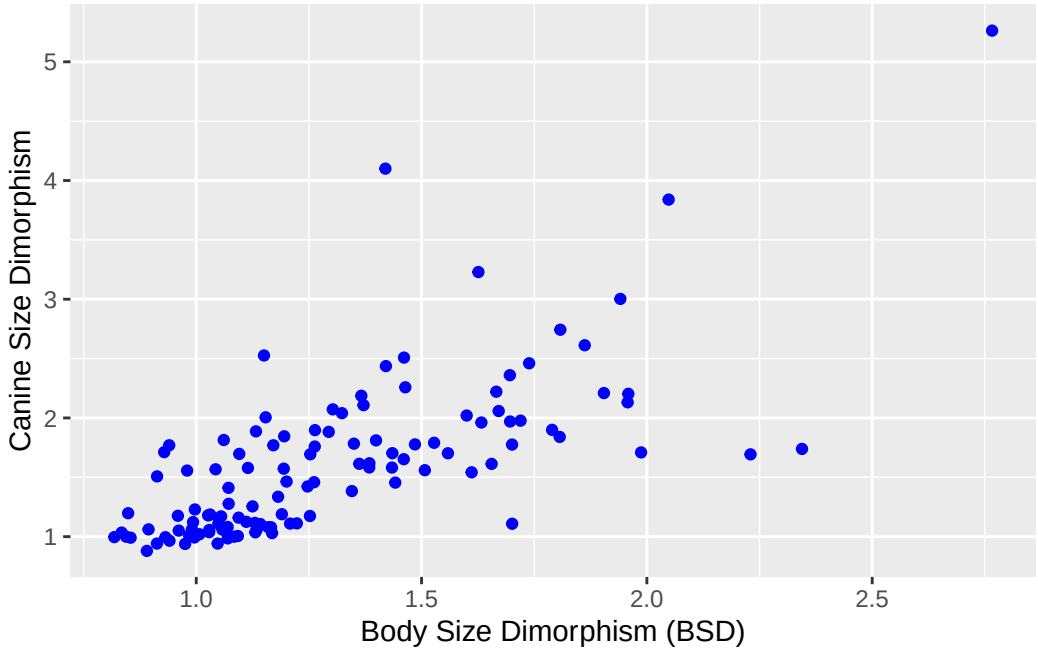
#After log-transforming both variables, the relationship between group size and day range length becomes clearer. There is a positive trend in families like Cercopithecidae and Hylobatidae, suggesting a stronger correlation between group size and movement distance. The transformation makes patterns easier to interpret.

6 #Overall

```
p <- ggplot(d, aes(x = BSD, y = Canine_Dimorphism)) +
  geom_jitter(color = "blue" , width = 0.1, na.rm = TRUE) +
  labs(x = "Body Size Dimorphism (BSD)", y = "Canine Size Dimorphism")
ggsave("overall.png")
```

```
Saving 5.5 x 3.5 in image
```

```
p
```

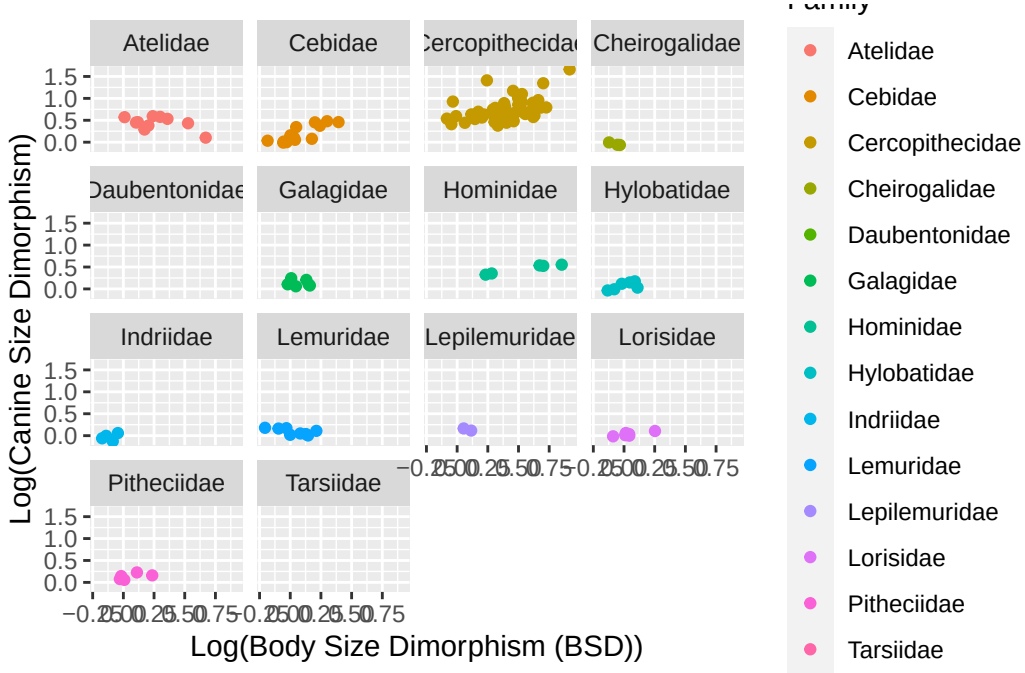


#Plot by family

```
p <- ggplot(d, aes(x = log(BSD), y = log(Canine_Dimorphism), color = Family)) +
  geom_jitter(width = 0.1, na.rm = TRUE) +
  facet_wrap(~ Family) +
  labs(x = "Log(Body Size Dimorphism (BSD))", y = "Log(Canine Size Dimorphism)")

ggsave("family.png")
```

```
Saving 5.5 x 3.5 in image
```

```
p
```



7

```
d <- mutate(d, diet_strategy = ifelse(Fruit >= 50, "frugivore",
                                      ifelse(Leaves >= 50, "folivore",
                                             ifelse(Fruit < 50 & Leaves < 50, "omnivore", NA)


glimpse(d)
```

```
Rows: 213
Columns: 28
$ Scientific_Name         <chr> "Allenopithecus_nigroviridis", "Allocebus_tric~
$ Family                  <chr> "Cercopithecidae", "Cercopithecidae", "Atelida~
$ Genus                   <chr> "Allenopithecus", "Allocebus", "Alouatta", "Al~
$ Species                 <chr> "nigroviridis", "trichotis", "belzebul", "cara~
$ Brain_Size_Species_Mean <dbl> 58.02, NA, 52.84, 52.63, 51.70, 49.88, 51.13, ~
$ Body_mass_male_mean     <dbl> 6130.0, 92.0, 7270.0, 6525.0, 5800.0, 7150.0, ~
$ Body_mass_female_mean   <dbl> 3180.0, 84.0, 5520.0, 4240.0, 4550.0, 5350.0, ~
$ MeanGroupSize           <dbl> NA, 1.000, 7.000, 8.000, 6.530, 12.000, 6.600,~
$ AdultMales              <dbl> NA, 1.000, 1.000, 2.300, 1.370, 2.900, 1.925, ~
$ AdultFemale             <dbl> NA, 1.000, 1.000, 3.300, 2.200, 6.300, 2.175, ~
$ GR_MidRangeLat_dd       <dbl> -0.17, -16.59, -6.80, -20.34, -21.13, 6.95, 18~
$ Precip_Mean_mm          <dbl> 1574.0, 1902.3, 1643.5, 1166.4, 1332.3, 1852.6~
```

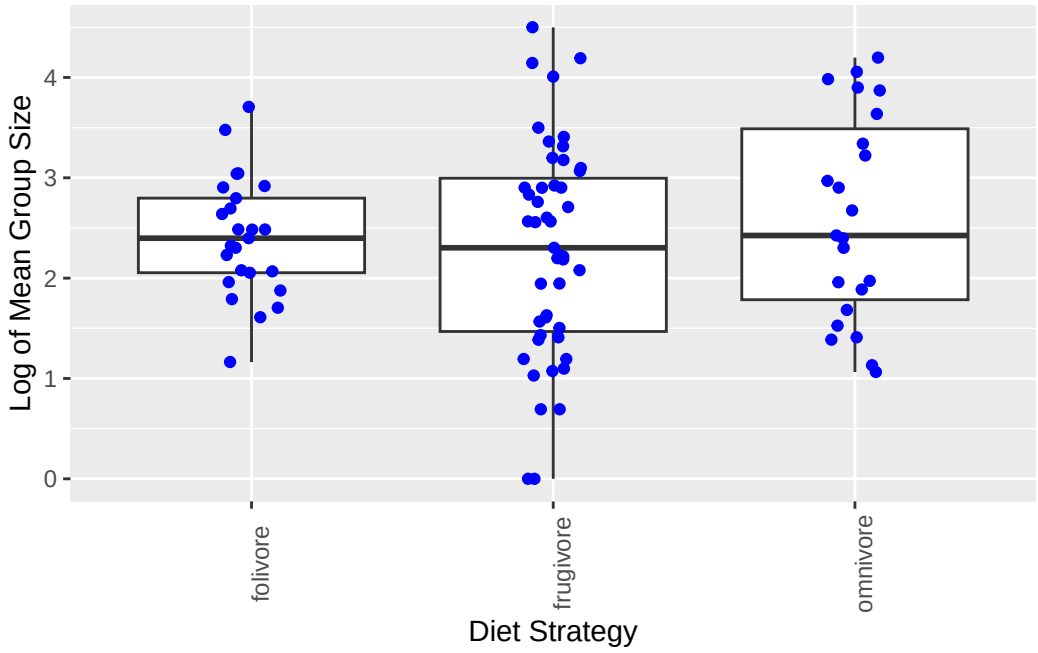
```
$ Temp_Mean_degC          <dbl> 25.2, 20.3, 24.9, 22.9, 19.6, 23.7, 25.1, 25.1~
$ HomeRange_km2           <dbl> NA, NA, NA, NA, 0.030, 0.190, 0.300, 0.100, 0.~
$ DayLength_km            <dbl> NA, NA, NA, 0.400, NA, 0.320, NA, 0.550, NA, N~
$ Fruit                   <dbl> NA, NA, 57.3, 23.8, 5.2, 33.1, 40.8, 40.0, 45.~
$ Leaves                  <chr> NA, NA, "19.1", "67.7", "73", "56.4", "45.1", ~
$ Fauna                   <chr> NA, NA, "0", "0", "0", "0", "0", "0", NA, NA, ~
$ Canine_Dimorphism       <dbl> 2.210, NA, 1.811, 1.542, 1.783, 1.703, 1.109, ~
$ Feed                    <dbl> NA, NA, 13.75, 15.90, 18.33, 17.94, 24.40, 12.~
$ Move                    <dbl> NA, NA, 18.75, 17.60, 14.33, 12.32, 9.80, 6.20~
$ Rest                    <dbl> NA, NA, 57.30, 61.60, 64.37, 66.14, 61.90, 78.~
$ Social                  <dbl> NA, NA, 10.00, 4.90, 3.00, 3.64, 3.80, 2.50, N~
$ BSD                     <dbl> 1.9276730, 1.0952381, 1.3170290, 1.5389151, 1.~
$ sex_ratio               <dbl> NA, 1.000000, 1.000000, 1.434783, 1.605839, 2.~
$ DI                      <dbl> NA, NA, NA, NA, NA, 1.6842105, NA, 5.5000000, ~
$ HomeRangeDiameter       <dbl> NA, NA, NA, NA, 0.1954410, 0.4918491, 0.618038~
$ diet_strategy           <chr> NA, NA, "frugivore", "folivore", "folivore", "~
```

```
# Cleaning NA
d_clean <- d[complete.cases(d[, c("diet_strategy", "MeanGroupSize")]), ]
glimpse(d_clean)
```

```
Rows: 95
Columns: 28
$ Scientific_Name         <chr> "Alouatta_belzebul", "Alouatta_caraya", "Aloua~
$ Family                  <chr> "Atelidae", "Atelidae", "Atelidae", "Atelidae"~
$ Genus                   <chr> "Alouatta", "Alouatta", "Alouatta", "Alouatta"~
$ Species                 <chr> "belzebul", "caraya", "guariba", "palliata", "~
$ Brain_Size_Species_Mean <dbl> 52.84, 52.63, 51.70, 49.88, 51.13, 55.22, 20.6~
$ Body_mass_male_mean     <dbl> 7270.0, 6525.0, 5800.0, 7150.0, 11400.0, 6690.~
$ Body_mass_female_mean   <dbl> 5520.0, 4240.0, 4550.0, 5350.0, 6430.0, 5210.0~
$ MeanGroupSize           <dbl> 7.000, 8.000, 6.530, 12.000, 6.600, 7.100, 3.1~
$ AdultMales              <dbl> 1.000, 2.300, 1.370, 2.900, 1.925, 1.700, 1.00~
$ AdultFemale             <dbl> 1.000, 3.300, 2.200, 6.300, 2.175, 2.200, 1.00~
$ GR_MidRangeLat_dd       <dbl> -6.80, -20.34, -21.13, 6.95, 18.80, 0.68, -17.~
$ Precip_Mean_mm          <dbl> 1643.5, 1166.4, 1332.3, 1852.6, 1341.3, 1823.4~
$ Temp_Mean_degC          <dbl> 24.9, 22.9, 19.6, 23.7, 25.1, 25.1, 24.6, 25.2~
$ HomeRange_km2           <dbl> NA, NA, 0.030, 0.190, 0.300, 0.100, 0.095, 0.1~
$ DayLength_km            <dbl> NA, 0.400, NA, 0.320, NA, 0.550, NA, 0.708, 0.~
$ Fruit                   <dbl> 57.3, 23.8, 5.2, 33.1, 40.8, 40.0, 45.0, 60.0,~
$ Leaves                  <chr> "19.1", "67.7", "73", "56.4", "45.1", "48.1", ~
$ Fauna                   <chr> "0", "0", "0", "0", "0", "0", NA, NA, "15", "0~
$ Canine_Dimorphism       <dbl> 1.811, 1.542, 1.783, 1.703, 1.109, 1.464, NA, ~
```

```
$ Feed                  <dbl> 13.75, 15.90, 18.33, 17.94, 24.40, 12.70, NA, ~
$ Move                  <dbl> 18.75, 17.60, 14.33, 12.32, 9.80, 6.20, NA, NA~
$ Rest                  <dbl> 57.30, 61.60, 64.37, 66.14, 61.90, 78.50, NA, ~
$ Social                <dbl> 10.00, 4.90, 3.00, 3.64, 3.80, 2.50, NA, NA, N~
$ BSD                   <dbl> 1.3170290, 1.5389151, 1.2747253, 1.3364486, 1.~
$ sex_ratio             <dbl> 1.000000, 1.434783, 1.605839, 2.172414, 1.1298~
$ DI                    <dbl> NA, NA, NA, 1.6842105, NA, 5.5000000, NA, 6.74~
$ HomeRangeDiameter     <dbl> NA, NA, 0.1954410, 0.4918491, 0.6180387, 0.356~
$ diet_strategy         <chr> "frugivore", "folivore", "folivore", "folivore~
```

```r
# Boxplot
p <- ggplot(d_clean, aes(x = diet_strategy, y = log(MeanGroupSize))) +
  geom_boxplot() +
  geom_jitter(color = "blue", width = 0.1) +
  labs(x = "Diet Strategy", y = "Log of Mean Group Size") +
  theme(axis.text.x = element_text(angle = 90))

p
```



8

```r
s <- mutate(d, Binomial = paste(Genus, Species, sep = " ")) %>%
     select(Binomial, Family, Brain_Size_Species_Mean, Body_mass_male_mean) %>%
     group_by(Family) %>%
     summarise(
     avg_BrainSize = mean(Brain_Size_Species_Mean, na.rm = TRUE),
     avg_BodyMass = mean(Body_mass_male_mean, na.rm = TRUE)) %>%
     arrange(avg_BrainSize)

s
```

```
# A tibble: 14 x 3
   Family          avg_BrainSize avg_BodyMass
   <chr>                   <dbl>        <dbl>
 1 Tarsiidae                3.26         131
 2 Cheirogalidae            4.04         193.
 3 Galagidae                5.96         395.
 4 Lepilemuridae            7.27         792
 5 Lorisidae                8.67         512.
 6 Lemuridae               23.1         2077.
 7 Cebidae                 23.9         1012.
 8 Indriidae               27.3         3638.
 9 Daubentonidae           44.8         2620
10 Pitheciidae             56.3         1955.
11 Atelidae                80.6         7895.
12 Cercopithecidae         85.4         9543.
13 Hylobatidae            101.          6926.
14 Hominidae              410.         98681.
```