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
BSD <- d$Body_mass_male_mean / d$Body_mass_female_mean

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

# 3. Create Defensibility Index
DI <- d$DayLength_km / (sqrt(d$HomeRange_km2 / pi))
```

4

```
# Overall
library(ggplot2)
library(dbplyr)
```

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

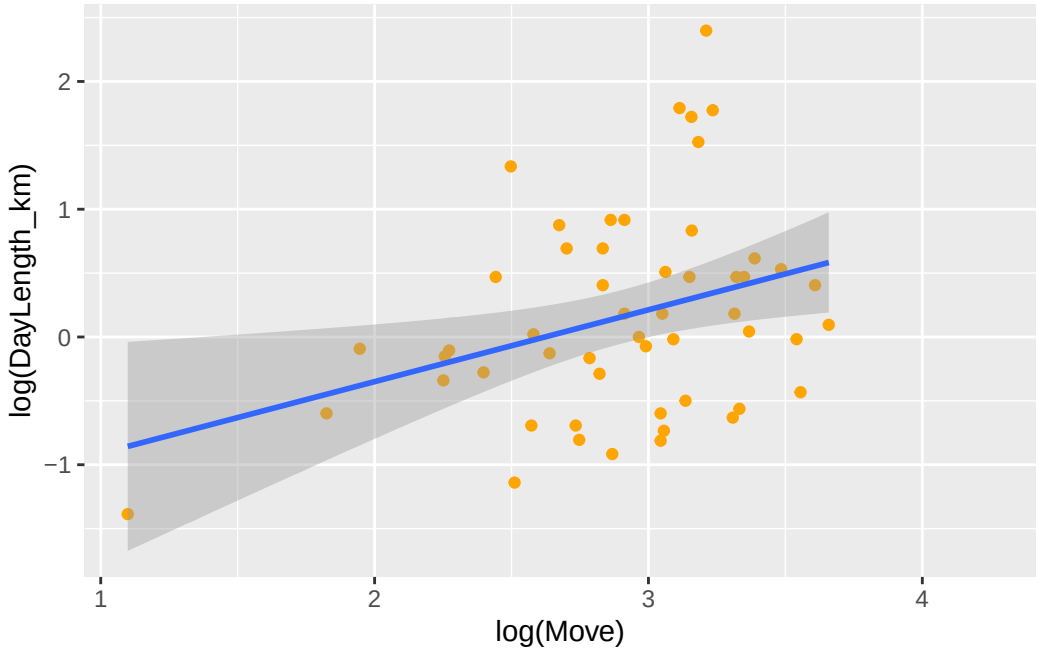
```
Attaching package: 'dbplyr'
```

```
The following objects are masked from 'package:dplyr':

    ident, sql
```

```
p <- ggplot(data = d, aes(x = log(Move), y = log(DayLength_km)))
p <- p + xlab("log(Move)") + ylab("log(DayLength_km)")
p <- p + geom_point( color = "orange", na.rm = TRUE)
p <- p + theme(legend.position = "bottom", legend.title = element_blank())
p <- p + geom_smooth(method = "lm", fullrange =FALSE, na.rm =TRUE)
p
```

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

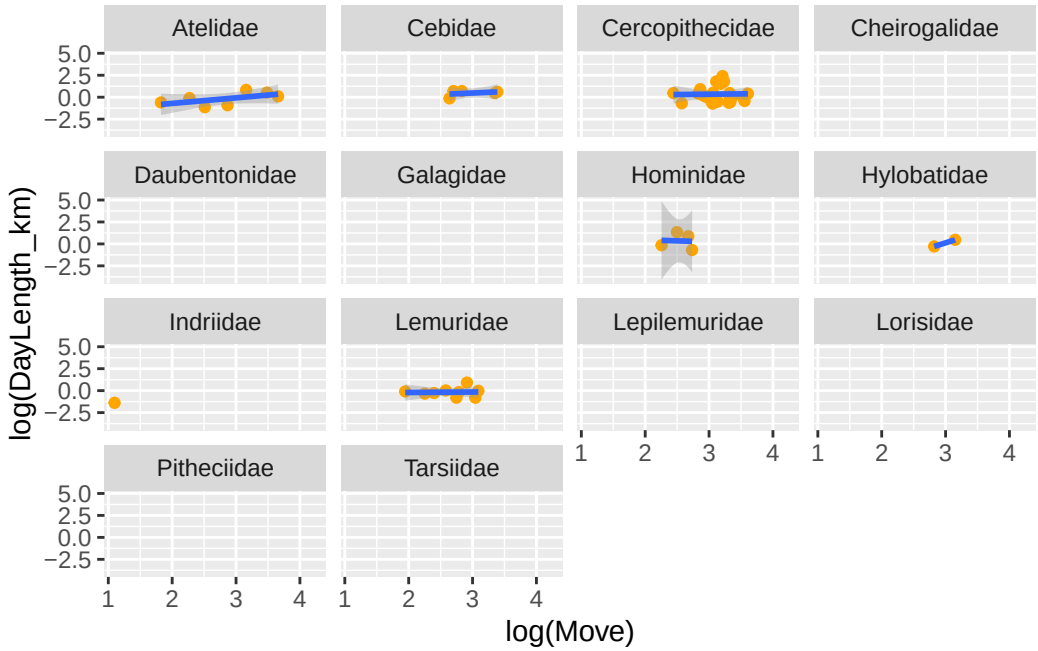



```
# By family
p <- p + facet_wrap(~Family)
     p + theme(legend.position = "none")
```

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

```
Warning in qt((1 - level)/2, df): NaNs produced
```

```
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



Do species that spend more time moving travel farther overall? Looking at the overall trend across all families, the linear regression line with a shaded confidence interval indicates that species that spend more time moving tend to travel farther overall.

How about within any particular primate family? There are positive correlation in species. Atelidae and Cebidae shows a weak positive trend is visible, suggesting that more movement is associated with greater travel distance.

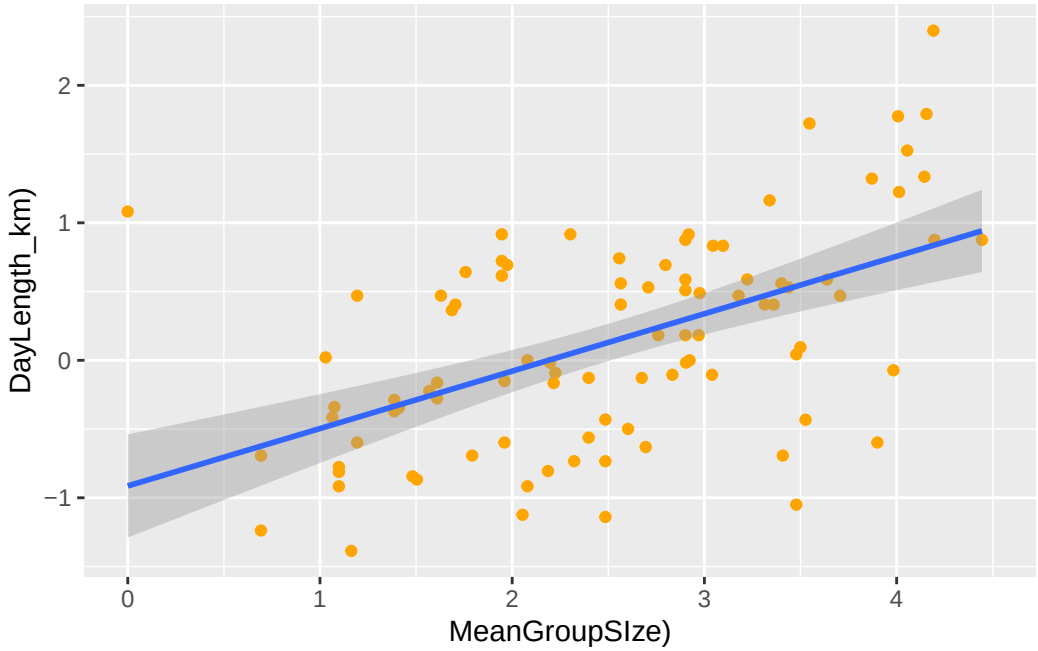
Should you transform either of these variables? Yes, transformated values make clear results for the population distrubition.

5

```
#Overall

library(ggplot2)
p <- ggplot(data = d, aes(x = log(MeanGroupSize), y =log(DayLength_km)))
p <- p + xlab("MeanGroupSIze)") + ylab("DayLength_km)")
p <- p + geom_point( color = "orange", na.rm = TRUE)
p <- p + theme(legend.position = "bottom", legend.title = element_blank())
p <- p + geom_smooth(method = "lm", fullrange =FALSE, na.rm =TRUE)
p
```

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

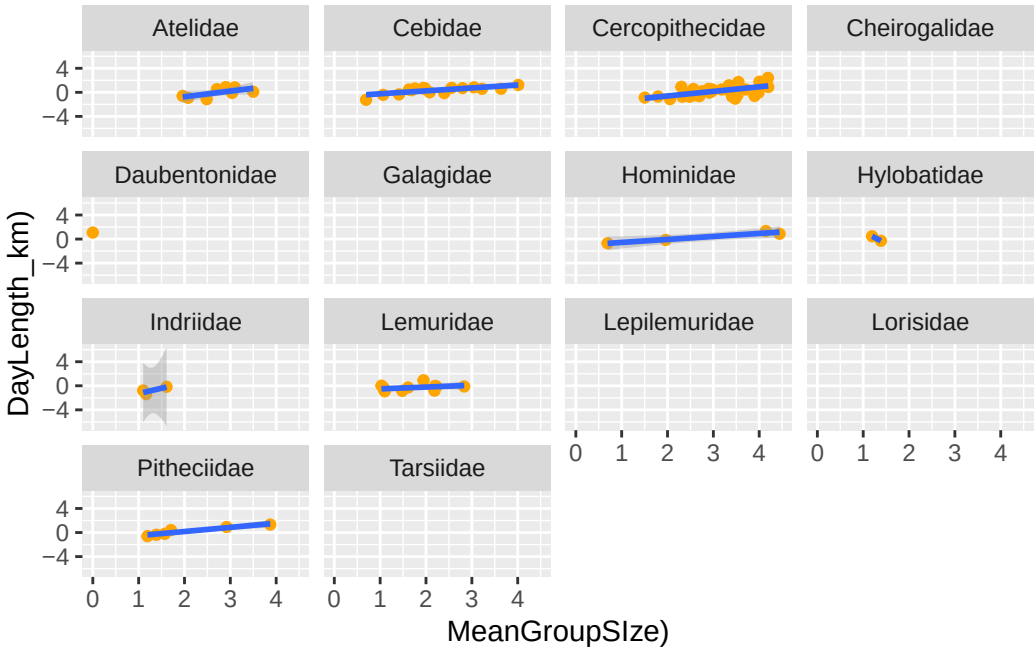



```
#by family
p <- p + facet_wrap(~Family)
     p + theme(legend.position = "none")
```

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

```
Warning in qt((1 - level)/2, df): NaNs produced
```

```
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



Do species in larger groups travel farther overall? Looking at the overall trend across all families, the linear regression line with a shaded confidence interval indicates that species that spend more time moving tend to travel farther overall.

How about within specific primate families? In the Cercopithecidae family, as group size increases, the daily travel distance also tends to increase significantly.

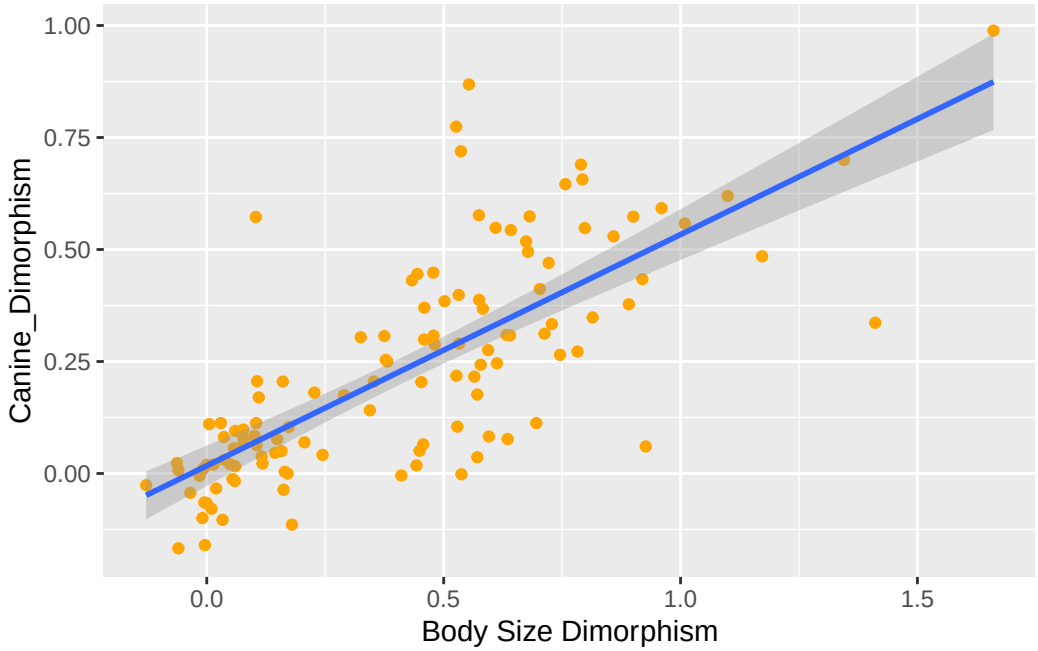
Should we transform either variable? Yes, transformated values make clear results for the population distrubition

6

```
##Overall

library(ggplot2)
p <- ggplot(data = d, aes(x = log(Canine_Dimorphism), y = log(BSD)))
p <- p + xlab("Body Size Dimorphism") + ylab("Canine_Dimorphism")
p <- p + geom_point( color = "orange", na.rm = TRUE)
p <- p + theme(legend.position = "bottom", legend.title = element_blank())
p <- p + geom_smooth(method = "lm", fullrange =FALSE, na.rm =TRUE)
p
```

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

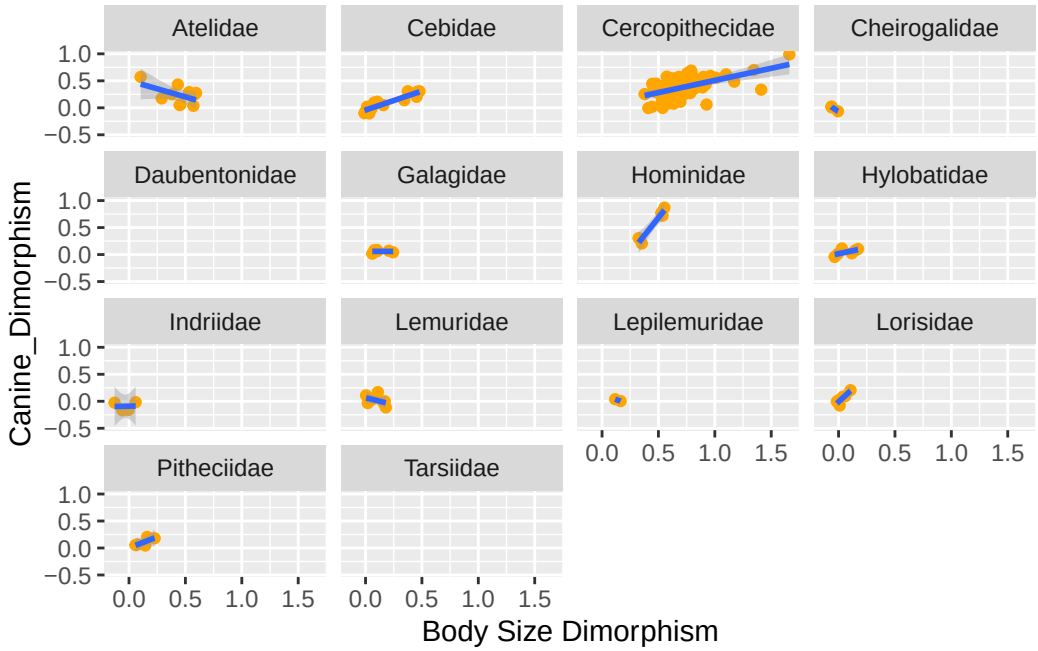



```
#by family
p <- p + facet_wrap(~Family)
     p + theme(legend.position = "none")
```

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

```
Warning in qt((1 - level)/2, df): NaNs produced
```

```
Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning
-Inf
```



Do taxa with greater size dimorphism also show greater canine dimorphism? Yes, taxa with greater size dimorphism generally exhibit greater canine dimorphism.

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: 24
$ 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~
$ diet_strategy           <chr> NA, NA, "frugivore", "folivore", "folivore", "~
```

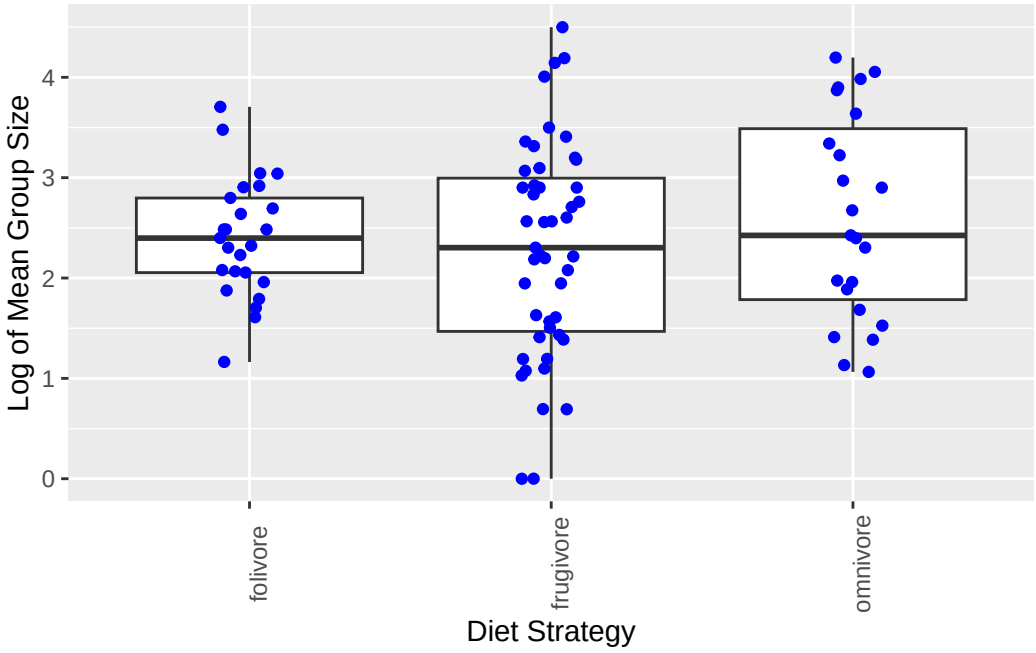
```r
# Cleaning NA
d_clean <- d[complete.cases(d[, c("diet_strategy", "MeanGroupSize")]), ]

head(d_clean[, c("diet_strategy", "MeanGroupSize")])
```

```
# A tibble: 6 x 2
  diet_strategy MeanGroupSize
  <chr>                 <dbl>
1 frugivore              7
2 folivore               8
3 folivore               6.53
4 folivore              12
5 omnivore               6.6
6 omnivore               7.1
```

```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
```



Do frugivores live in larger groups than folivores?

Yes, based on the boxplot frugivores tend to live in larger groups than folivores.

8

```
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.
```