

EDA Challenges

1. Load the dataset data-wrangling and assigning the variable 'd'

```
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      1.3.1
v purrr      1.0.2
```

```
-- 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 become
```

```
f <- "https://raw.githubusercontent.com/difiore/ada-2024-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.
```

```
attach(d)
```

```
names(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"

```

2. Create new variable BSD

```

d$BSD <- d$ Body_mass_male_mean/ d$Body_mass_female_mean
print(d)

```

```

# A tibble: 213 x 24
  Scientific_Name      Family      Genus Species Brain_Size_Species_M~1
  <chr>              <chr>      <chr> <chr>      <dbl>
1 Allenopithecus_nigroviridis Cercopithec~ Alle~ nigrov~      58.0
2 Allocebus_trichotis Cercopithec~ Allo~ tricho~      NA
3 Alouatta_belzebul Atelidae Alou~ belzeb~      52.8
4 Alouatta_caraya Atelidae Alou~ caraya      52.6
5 Alouatta_guariba Atelidae Alou~ guariba      51.7
6 Alouatta_palliata Atelidae Alou~ pallia~      49.9
7 Alouatta_pigra Atelidae Alou~ pigra      51.1
8 Alouatta_seniculus Atelidae Alou~ senicu~      55.2
9 Aotus_azarai Cebidae Aotus azarai      20.7
10 Aotus_brumbacki Cebidae Aotus brumba~      NA
# i 203 more rows
# i abbreviated name: 1: Brain_Size_Species_Mean
# i 19 more variables: Body_mass_male_mean <dbl>, Body_mass_female_mean <dbl>,
# MeanGroupSize <dbl>, AdultMales <dbl>, AdultFemale <dbl>,
# GR_MidRangeLat_dd <dbl>, Precip_Mean_mm <dbl>, Temp_Mean_degC <dbl>,
# HomeRange_km2 <dbl>, DayLength_km <dbl>, Fruit <dbl>, Leaves <chr>,
# Fauna <chr>, Canine_Dimorphism <dbl>, Feed <dbl>, Move <dbl>, ...

```

3. Create new variable sex-ratio

```

d$Sex_ratio <- d$ AdultFemale/ d$ AdultMales
print(d)

# A tibble: 213 x 25
  Scientific_Name      Family      Genus Species Brain_Size_Species_M~1
  <chr>              <chr>      <chr> <chr>      <dbl>
1 Allenopithecus_nigroviridis Cercopithec~ Alle~ nigrov~ 58.0
2 Allocebus_trichotis Cercopithec~ Allo~ tricho~ NA
3 Alouatta_belzebul Atelidae Alou~ belzeb~ 52.8
4 Alouatta_caraya Atelidae Alou~ caraya 52.6
5 Alouatta_guariba Atelidae Alou~ guariba 51.7
6 Alouatta_palliata Atelidae Alou~ pallia~ 49.9
7 Alouatta_pigra Atelidae Alou~ pigra 51.1
8 Alouatta_seniculus Atelidae Alou~ senicu~ 55.2
9 Aotus_azarai Cebidae Aotus azarai 20.7
10 Aotus_brumbaki Cebidae Aotus brumba~ NA
# i 203 more rows
# i abbreviated name: 1: Brain_Size_Species_Mean
# i 20 more variables: Body_mass_male_mean <dbl>, Body_mass_female_mean <dbl>,
# MeanGroupSize <dbl>, AdultMales <dbl>, AdultFemale <dbl>,
# GR_MidRangeLat_dd <dbl>, Precip_Mean_mm <dbl>, Temp_Mean_degC <dbl>,
# HomeRange_km2 <dbl>, DayLength_km <dbl>, Fruit <dbl>, Leaves <chr>,
# Fauna <chr>, Canine_Dimorphism <dbl>, Feed <dbl>, Move <dbl>, ...

```

4. Calculate Diameter of the home range for each species

```

d$home_range_diameter <- d$HomeRange_km2/ 2
d$home_range_area <- pi * (d$home_range_diameter^2)
d$DI <- d$DayLength_km/ d$HomeRange_km2
print(d)

# A tibble: 213 x 28
  Scientific_Name      Family      Genus Species Brain_Size_Species_M~1
  <chr>              <chr>      <chr> <chr>      <dbl>
1 Allenopithecus_nigroviridis Cercopithec~ Alle~ nigrov~ 58.0
2 Allocebus_trichotis Cercopithec~ Allo~ tricho~ NA
3 Alouatta_belzebul Atelidae Alou~ belzeb~ 52.8
4 Alouatta_caraya Atelidae Alou~ caraya 52.6
5 Alouatta_guariba Atelidae Alou~ guariba 51.7
6 Alouatta_palliata Atelidae Alou~ pallia~ 49.9
7 Alouatta_pigra Atelidae Alou~ pigra 51.1
8 Alouatta_seniculus Atelidae Alou~ senicu~ 55.2

```

```

9 Aotus_azarai          Cebidae      Aotus azarai          20.7
10 Aotus_brumbacki      Cebidae      Aotus brumba~        NA
# i 203 more rows
# i abbreviated name: 1: Brain_Size_Species_Mean
# i 23 more variables: Body_mass_male_mean <dbl>, Body_mass_female_mean <dbl>,
#   MeanGroupSize <dbl>, AdultMales <dbl>, AdultFemale <dbl>,
#   GR_MidRangeLat_dd <dbl>, Precip_Mean_mm <dbl>, Temp_Mean_degC <dbl>,
#   HomeRange_km2 <dbl>, DayLength_km <dbl>, Fruit <dbl>, Leaves <chr>,
#   Fauna <chr>, Canine_Dimorphism <dbl>, Feed <dbl>, Move <dbl>, ...

```

5. Create new variable DI (Defensibility Index)

```

d$DI <- d$ DayLength_km/ d$ home_range_diameter
print(d)

```

```

# A tibble: 213 x 28
  Scientific_Name      Family      Genus Species Brain_Size_Species_M~1
  <chr>               <chr>      <chr> <chr>      <dbl>
1 Allenopithecus_nigroviridis Cercopithec~ Alle~ nigrov~      58.0
2 Allocebus_trichotis Cercopithec~ Allo~ tricho~      NA
3 Alouatta_belzebul Atelidae Alou~ belzeb~      52.8
4 Alouatta_caraya Atelidae Alou~ caraya      52.6
5 Alouatta_guariba Atelidae Alou~ guariba      51.7
6 Alouatta_palliata Atelidae Alou~ pallia~      49.9
7 Alouatta_pigra Atelidae Alou~ pigra      51.1
8 Alouatta_seniculus Atelidae Alou~ senicu~      55.2
9 Aotus_azarai Cebidae Aotus azarai      20.7
10 Aotus_brumbacki Cebidae Aotus brumba~      NA
# i 203 more rows
# i abbreviated name: 1: Brain_Size_Species_Mean
# i 23 more variables: Body_mass_male_mean <dbl>, Body_mass_female_mean <dbl>,
#   MeanGroupSize <dbl>, AdultMales <dbl>, AdultFemale <dbl>,
#   GR_MidRangeLat_dd <dbl>, Precip_Mean_mm <dbl>, Temp_Mean_degC <dbl>,
#   HomeRange_km2 <dbl>, DayLength_km <dbl>, Fruit <dbl>, Leaves <chr>,
#   Fauna <chr>, Canine_Dimorphism <dbl>, Feed <dbl>, Move <dbl>, ...

```

6. Create the plot for showing overall relationship between day range length and time spent moving

```

overall_plot <- ggplot(d, aes(x = DayLength_km, y = Move)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +

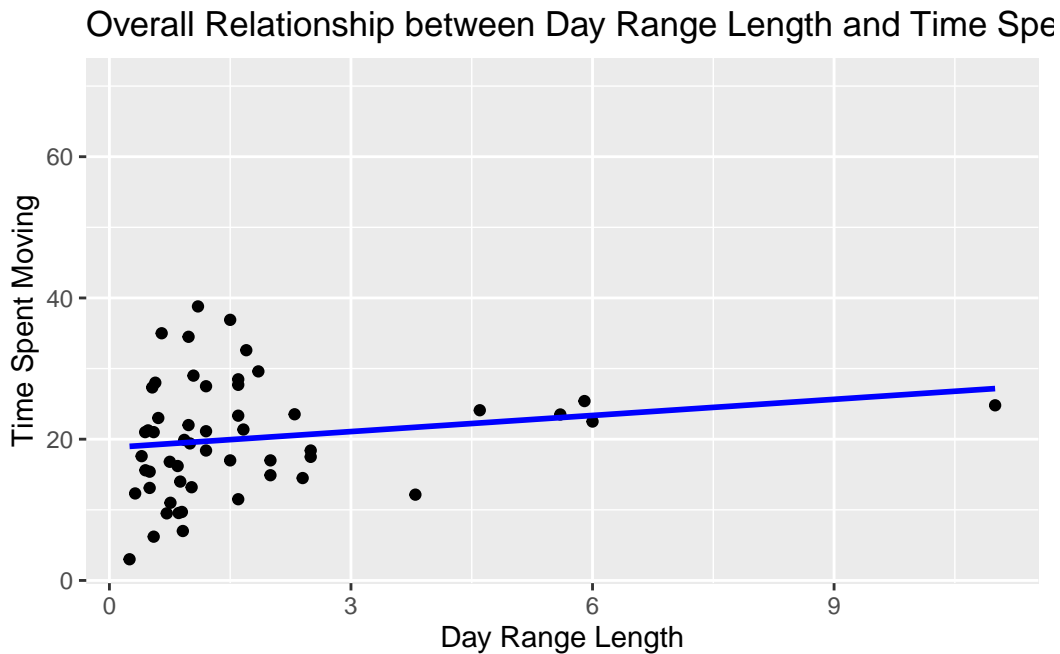
```

```
labs(title = "Overall Relationship between Day Range Length and Time Spent Moving",
     x = "Day Range Length",
     y = "Time Spent Moving")
print(overall_plot)
```

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

Warning: Removed 160 rows containing non-finite values (`stat_smooth()`).

Warning: Removed 160 rows containing missing values (`geom_point()`).



7. Create the plot by family

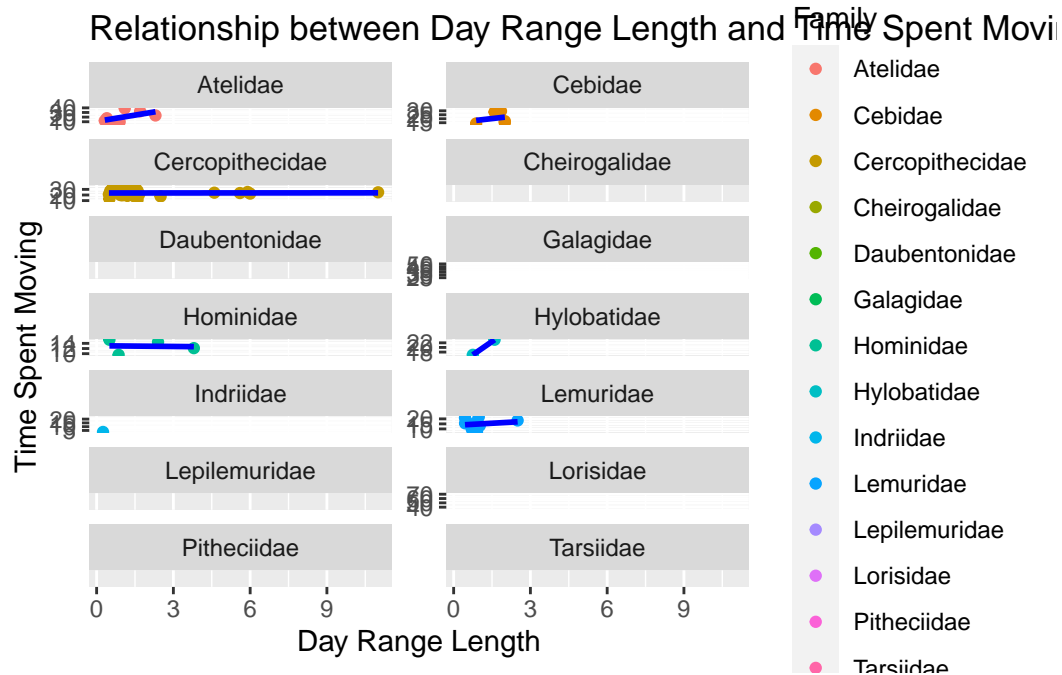
```
family_plot <- ggplot(d, aes(x = DayLength_km, y = Move, color = Family)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  facet_wrap(~Family, scales = "free_y", ncol = 2) +
  labs(title = "Relationship between Day Range Length and Time Spent Moving by Primate",
       x = "Day Range Length",
       y = "Time Spent Moving")

print(family_plot)
```

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

Warning: Removed 160 rows containing non-finite values (`stat_smooth()`).

Warning: Removed 160 rows containing missing values (`geom_point()`).



No, the species that spend more time moving does not travel farther overall. There is no linear relationship between time spent moving and the day range length, overall. However, when we see the relationship by family, Atelidae, cebidae, and Hylobatidae family shows some degree of linear relationship. Yes, I think we should transform either of the variables (logarithmic transformation may improve the linearity).

8. Applying logarithmic transformation to one variable (move)

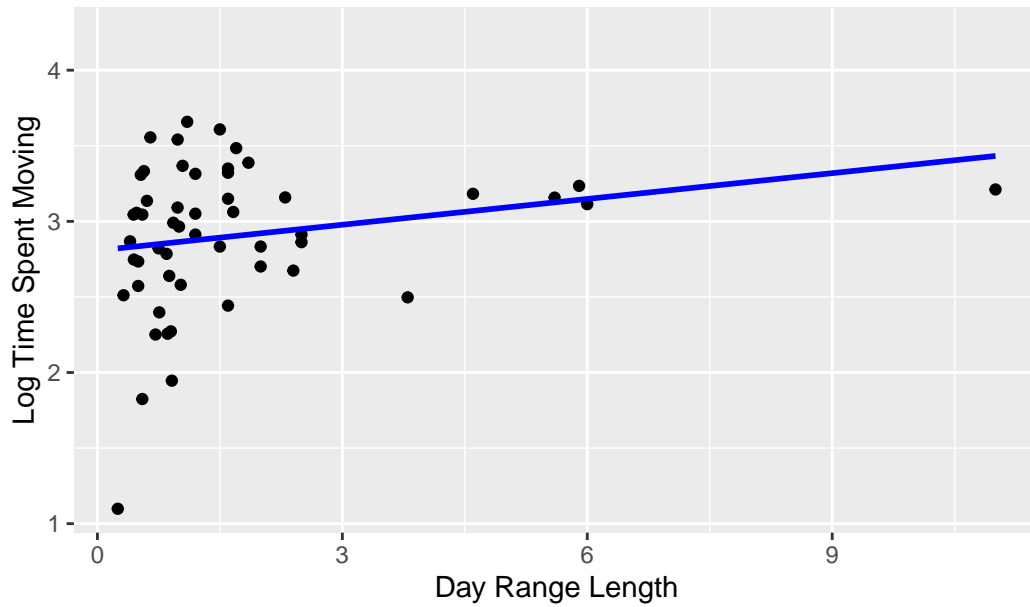
```
d$log_Move <- log(d$Move)
overall_plot <- ggplot(d, aes(x = DayLength_km, y = log_Move)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  labs(title = "Overall Relationship between Day Range Length and Time Spent Moving",
       x = "Day Range Length",
       y = "Log Time Spent Moving")
print(overall_plot)
```

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

```
Warning: Removed 160 rows containing non-finite values (`stat_smooth()`).
```

```
Warning: Removed 160 rows containing missing values (`geom_point()`).
```

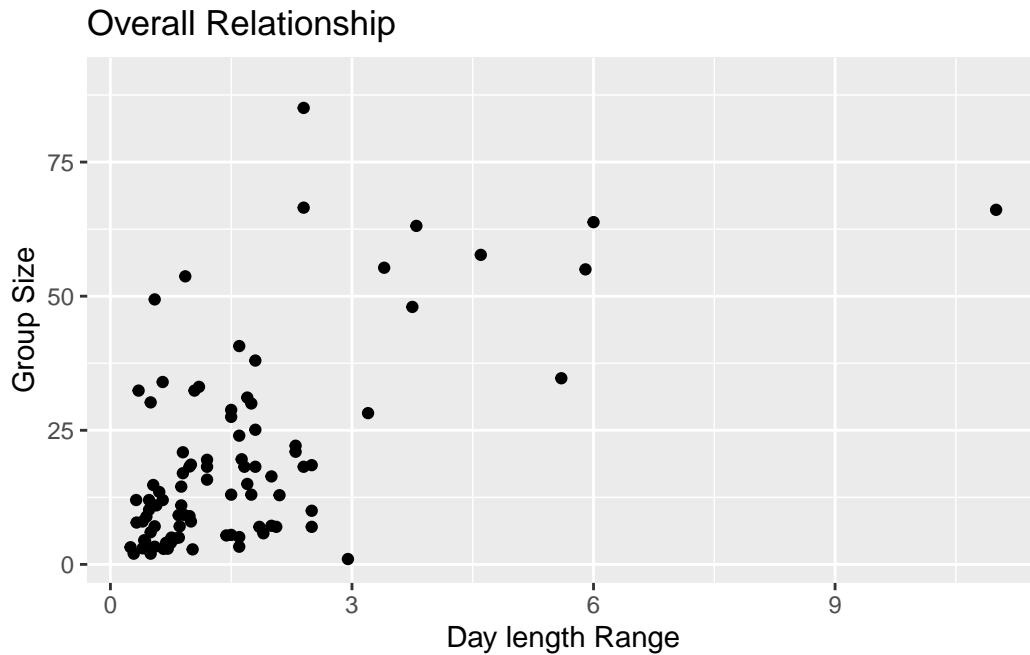
Overall Relationship between Day Range Length and Time Spent Moving



9. Create the plot to show relationship between day range length and time group size overall

```
overall_plot <- ggplot(d, aes(x = DayLength_km, y = MeanGroupSize)) +  
  geom_point() +  
  labs(title = "Overall Relationship",  
        x = "Day length Range",  
        y = "Group Size")  
print(overall_plot)
```

```
Warning: Removed 120 rows containing missing values (`geom_point()`).
```



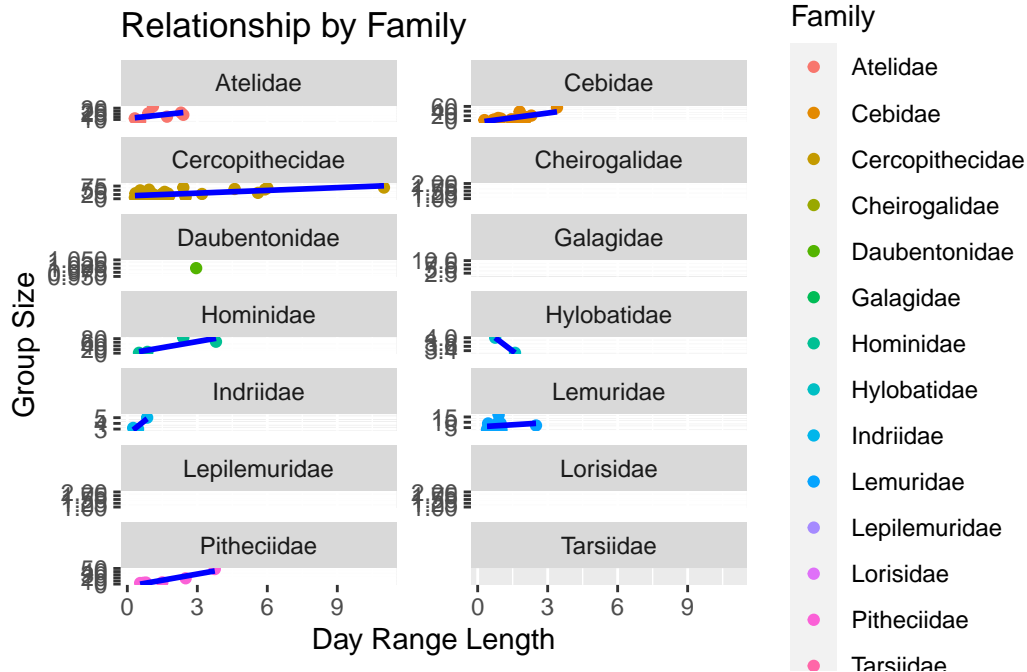
9. Create the plot by family

```
family_plot <- ggplot(d, aes(x = DayLength_km, y = MeanGroupSize, color = Family)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  facet_wrap(~Family, scales = "free_y", ncol = 2) +
  labs(title = "Relationship by Family",
       x = "Day Range Length",
       y = "Group Size")
print(family_plot)
```

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

Warning: Removed 120 rows containing non-finite values (`stat_smooth()`).

Warning: Removed 120 rows containing missing values (`geom_point()`).

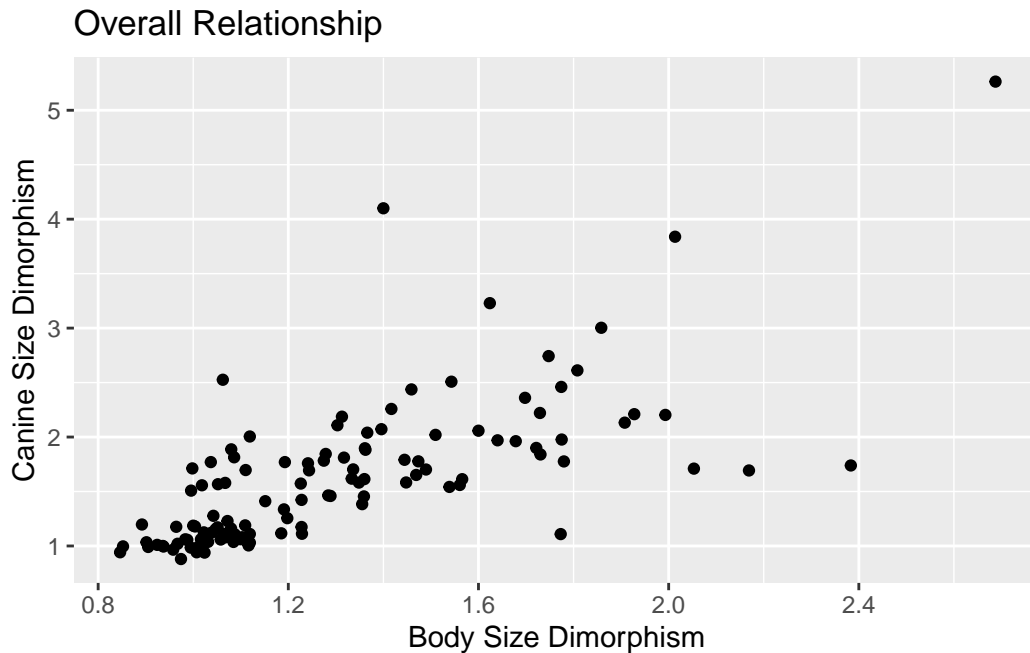


There is some degree of positive linear relationship between Day range length and time group size overall. In this plot, when relationship is analyzed by family, positive linear relationship is seen among Atelidae, Cebidae, cercopithecidae, hominidae, Indriidae. and Pitheciidae, and negative relationship is seen in Hylobatidae family. In my opinion, transformation off data is not required.

10. Plot the relationship between body size dimorphism and canine dimorphism

```
overall_plot <- ggplot(d, aes(x = BSD, y = Canine_Dimorphism)) +
  geom_point() +
  labs(title = "Overall Relationship",
       x = "Body Size Dimorphism",
       y = "Canine Size Dimorphism")
print(overall_plot)
```

Warning: Removed 94 rows containing missing values (`geom_point()`).



Yes, the taxa with greater size dimorphism also show greater canine dimorphism. There is a linear relationship between body size dimorphism and canine size dimorphism

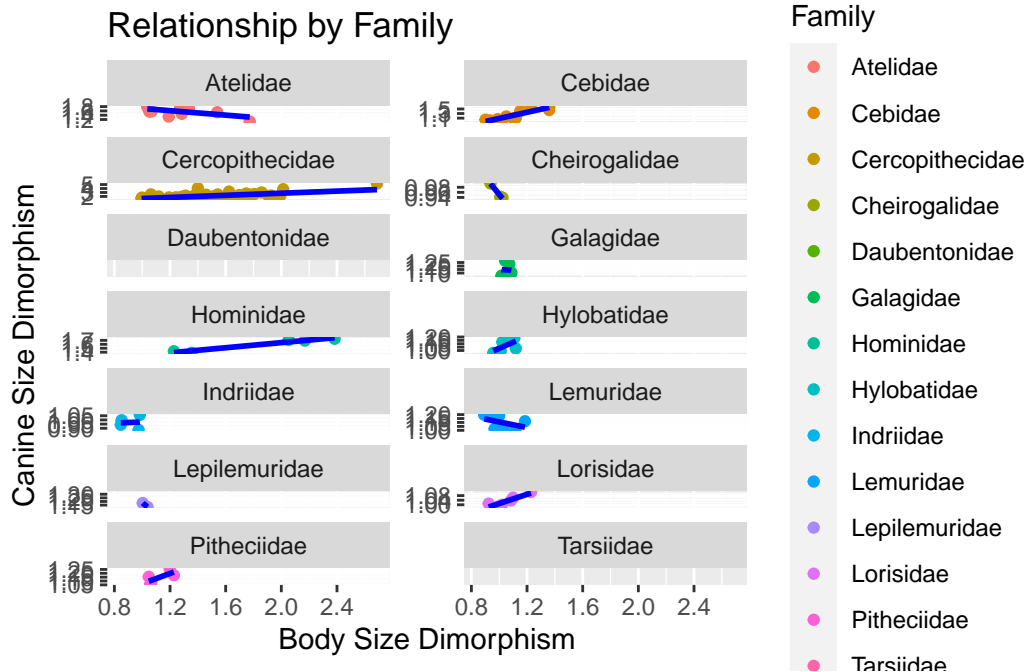
11. Plot the relationship between body size dimorphism and canine dimorphism by family

```
family_plot <- ggplot(d, aes(x = BSD, y = Canine_Dimorphism, color = Family)) +
  geom_point() +
  geom_smooth(method = "lm", se = FALSE, color = "blue") +
  facet_wrap(~Family, scales = "free_y", ncol = 2) +
  labs(title = "Relationship by Family",
       x = "Body Size Dimorphism",
       y = "Canine Size Dimorphism")
print(family_plot)
```

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

Warning: Removed 94 rows containing non-finite values (`stat_smooth()`).

Warning: Removed 94 rows containing missing values (`geom_point()`).



Positive Linear relationship between Body size Dimorphism and canine size Dimorphism is observed among Cebidae, Cercopithecidae, Hominidae, Hylobatidae, Pitheciidae, and Lorisidae family. However, negative linear relationship is observed among Atelidae, Lemuridae, and cheirogalidae family.

12. Create a new variable named **diet_strategy** that is “frugivore” if fruits make up >50% of the diet, “folivore” if leaves make up >50% of the diet, and “omnivore” if neither of these is true.

```
library(dplyr)
library(ggplot2)

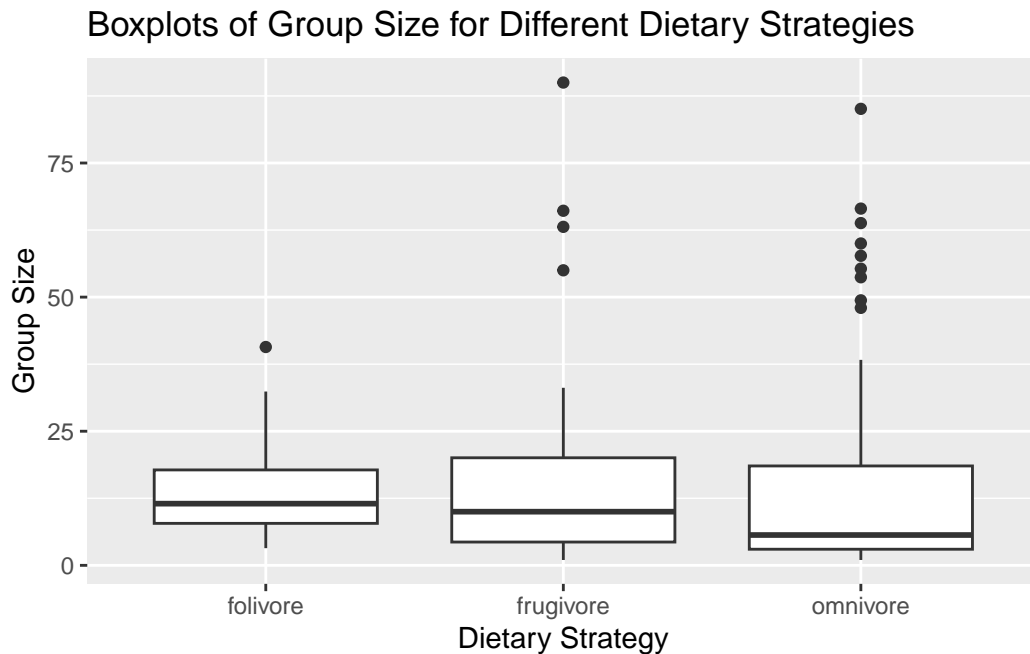
# Creating the new variable diet_strategy
d <- d %>%
  mutate(diet_strategy = case_when(
    Fruit > 50 ~ "frugivore",
    Leaves > 50 ~ "folivore",
    TRUE ~ "omnivore"
  ))

# Updating omnivore category to exclude species where both Fruit and Leaves > 50%
```

```
d <- d %>%
  mutate(diet_strategy = case_when(
    diet_strategy == "omnivore" & (is.na(Fruit) | Fruit <= 50) & (is.na(Leaves) | Leaves <
      TRUE ~ diet_strategy
    ))

# Create boxplots
boxplot_plot <- ggplot(d, aes(x = diet_strategy, y = MeanGroupSize)) +
  geom_boxplot() +
  labs(title = "Boxplots of Group Size for Different Dietary Strategies",
       x = "Dietary Strategy",
       y = "Group Size")
print(boxplot_plot)
```

Warning: Removed 60 rows containing non-finite values (`stat_boxplot()`).



13. In one line of code, using {dplyr} verbs and the forward pipe (%> or |>) operator, do the following:

```
library(dplyr)
library(readr)
```

```
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.

```
D <- d %>%
  mutate(Binomial = paste(Genus, Species, sep = " ")) %>%
  select(Binomial, Family, Brain_Size_Species_Mean, Body_mass_male_mean) %>%
  group_by(Family) %>%
  summarize(
    Avg_Brain_size_species_mean = mean(Brain_Size_Species_Mean, na.rm = TRUE),
    Avg_Body_mass_male_mean = mean(Body_mass_male_mean, na.rm = TRUE)
  ) %>%
  arrange(Avg_Brain_size_species_mean)
D
```

```
# A tibble: 14 x 3
  Family          Avg_Brain_size_species_mean Avg_Body_mass_male_mean
  <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.
```

14. Loading my own dataset “Boxplot.csv” and calculating the summary statistics

```

library (tidyverse)
f <- "Boxplot.csv"
d <- read_csv(f, col_names = TRUE)

Rows: 120 Columns: 4
-- Column specification -----
Delimiter: ","
chr (1): Group
dbl (3): Frequency, Delay, Absorbance

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.

attach(d)
names(d)

[1] "Frequency" "Group"      "Delay"      "Absorbance"

nrow(d)

[1] 120

ncol(d)

[1] 4

variable_names <- names(d)
print(variable_names)

[1] "Frequency" "Group"      "Delay"      "Absorbance"

numeric_variables <- names(d)[sapply(d, is.numeric)]
summary_list <- list()
for (variable in numeric_variables) {
  num_obs <- sum(!is.na(d[[variable]]))
  mean_val <- mean(d[[variable]], na.rm = TRUE)
  sd_val <- sd(d[[variable]], na.rm = TRUE)
  five_num_summary <- summary(d[[variable]], na.rm = TRUE)

  summary_list[[variable]] <- list(
    variable = variable,

```

```

    num_obs = num_obs,
    mean_val = mean_val,
    sd_val = sd_val,
    five_num_summary = five_num_summary
  )
}
for (variable_summary in summary_list) {
  print(paste("Variable:", variable_summary$variable))
  print(paste("Number of observations:", variable_summary$num_obs))
  print(paste("Mean:", variable_summary$mean_val))
  print(paste("Standard Deviation:", variable_summary$sd_val))
  print("Five-Number Summary:")
  print(variable_summary$five_num_summary)
}

```

```

[1] "Variable: Frequency"
[1] "Number of observations: 120"
[1] "Mean: 1810.5"
[1] "Standard Deviation: 689.416816025318"
[1] "Five-Number Summary:"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
1000   1310   1707   1810   2207   2828
[1] "Variable: Delay"
[1] "Number of observations: 97"
[1] "Mean: 126.638350515464"
[1] "Standard Deviation: 62.0801261254426"
[1] "Five-Number Summary:"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
12.25   83.71 120.71 126.64 175.58 283.58     23
[1] "Variable: Absorbance"
[1] "Number of observations: 108"
[1] "Mean: 0.541361111111111"
[1] "Standard Deviation: 0.164624322156487"
[1] "Five-Number Summary:"
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.   NA's
0.1790  0.4255  0.5435  0.5414  0.6615  0.8820     12

```

15. Plotting box-plot for my own dataset "Boxplot.csv"

```

library(tidyverse)
f <- "Boxplot.csv"
d <- read_csv(f, col_names = TRUE)

```

```

Rows: 120 Columns: 4
-- Column specification -----
Delimiter: ","
chr (1): Group
dbl (3): Frequency, Delay, Absorbance

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.

```

```

ggplot(d, aes(x = as.factor(Frequency), y = Absorbance, fill = Group)) +
  geom_boxplot(width = 0.7, outlier.shape = NA, na.rm = TRUE) +
  labs(x = "Frequency in Hz", y = "MEPA Delay Value",
       title = "EHF Loss vs EHF Normal MEPA Delay across frequencies") +
  theme_classic() +
  ylim(0, 1)

```

