Project 3

```
library(tidyverse)
library(colorspace)
knitr::opts_chunk$set(echo = TRUE)
```

Joshua Bryer jeb5533

This is the dataset used in this project:

```
# load in data
lemurs <-
readr::read_csv(
'https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2021/2021-08-24/lemur_data.cs
```

Link to the dataset: *https://github.com/rfordatascience/tidytuesday/blob/c45ce93c04857fcdf7988e31b7c452e93a3d9fe1/data/2021/2021-08-24/readme.md*

Part 1

Question: *For the top three most populous taxonomic codes of lemur, which gender lemurs have the highest median weight? and how does the distribution of weights change between the taxonomic codes?*

Introduction: *We are working with the Lemurs dataset, which contains 82,609 observations records from Lemurs. The data comes from a wildlife preservation center named, the Duke Lemur Center.In this dataset, each row corresponds to an observation of a lemur,but not each row represents a distinct lemur.There are fifty-four columns providing information about the lemur's taxonomy,birth, death, the parents, and their offspring. Information about the birth includes their date of birth, and the size of their litter. Information about the lemur includes the sex, the weight, and the taxonomy. Information about the lemur's parents includes their age, their taxonomy.*

To answer the question of Part 1, we will work with three variables, the lemur's taxonomy(column taxon),the lemur's sex (column sex), and the lemur's weight (column weight_g). The lemur's taxonomy is a categorical value presented as a series of letters denoting the first letter of the genus and then the first three letters of the species. The lemur's sex is also a categorical variable and is presented as either male or female.The lemur's weight is shown as a numerical continuous variable and is given in grams.

Approach: *Our approach is to show the difference of median weights between genders for the top three most populous taxonomies using a summary table. Summary tables make it easy to compare medians side-by-side.*

One limitation of the summary tables is that they don't allow us to see the distribution of weights for the taxonomic codes. Therefore, we will visualize the weights of the lemurs in box plots (geom_boxplot()), faceted by the taxonomic codes. Additionally, we will place the box plots on a logarithmic scale as some of the lemur's weight's are much larger than the rest. In combination, this data table and this plot will allow us to answer question one.

Analysis:

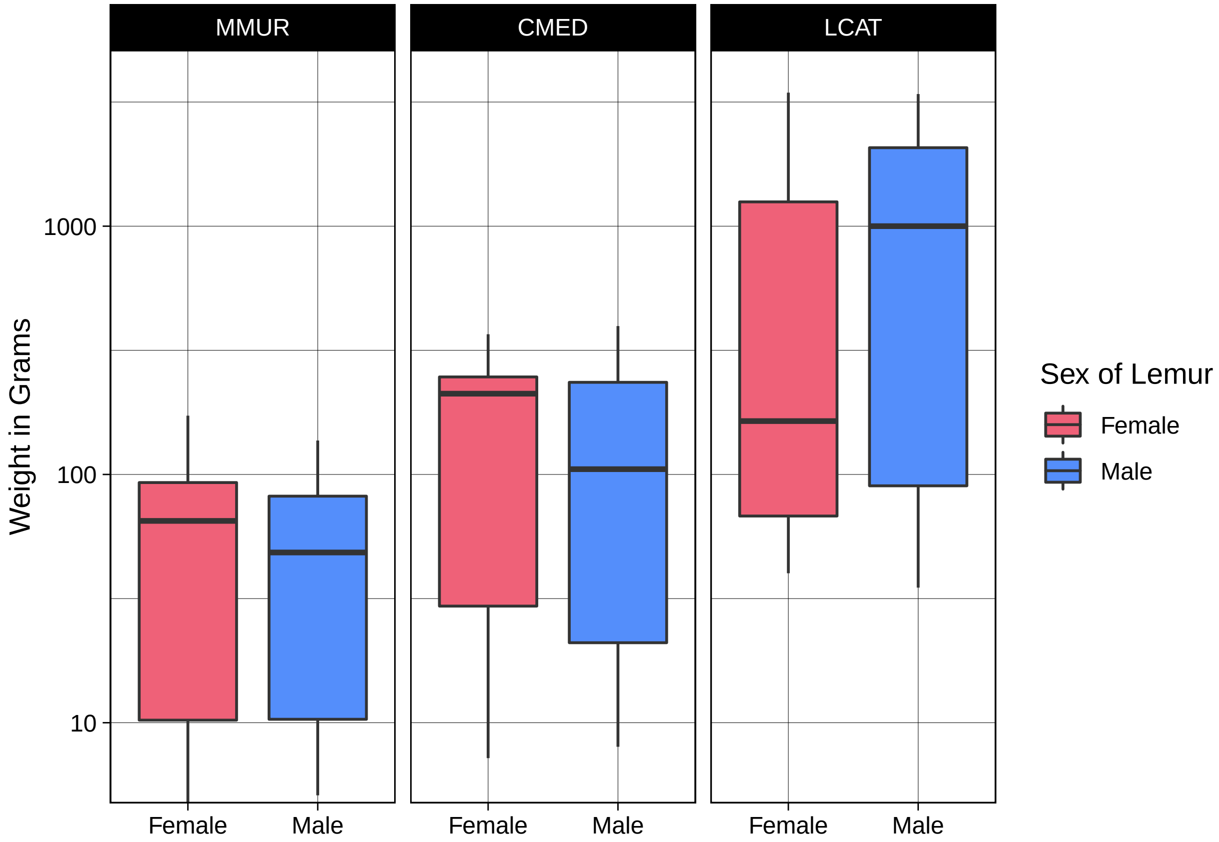
```
#In this code chunk we remove outlier variables, duplicate values of lemurs
#and recode variables in preparation for analysis and visualization
outliers <- boxplot(lemurs$weight_g, plot=FALSE)$out
df<-lemurs
```

```
df<- df[-which(df$weight_g %in% outliers),]
df<-df%>%
  mutate(age_category = recode(age_category, `IJ` = "Juvenile",
                               `young_adult` = "Young Adult",
                               `adult`="Adult"))%>%
  mutate(sex = recode(sex, `F` = "Female", M = "Male"))%>%
  mutate(age_category=fct_relevel(age_category,"Juvenile","Young Adult","Adult"))%>%
  distinct(dlc_id,.keep_all =T)
```

```
#Here we create a data table showing the summary values of lemurs by sex and
#taxonomy
df%>%
  filter(sex!="ND")%>%
   mutate(
    taxon = fct_lump_n(taxon, 3)
   )%>%
  filter(taxon!="Other")%>%
  group_by(sex,taxon) %>%
  summarise(median = median(weight_g), n = n(),max=max(weight_g),min=min(weight_g))
```

```
## # A tibble: 6 x 6
## # Groups:   sex [2]
##   sex    taxon median     n   max   min
##   <chr>  <fct>  <dbl> <int> <dbl> <dbl>
## 1 Female CMED   212.     82  367   7.2
## 2 Female LCAT   164     115 3450  40
## 3 Female MMUR    65      98  172.  4.74
## 4 Male   CMED   105      97  396   8
## 5 Male   LCAT  1000     109 3406  35
## 6 Male   MMUR    48.5   118  137   5.1
```

```
# In this code chunk we create a boxplot showing the distribution of weights
#for the lemurs by sex and taxonomy on a logarithmic scale
df%>%
  filter(sex!="ND")%>%
   mutate(
    taxon = fct_lump_n(taxon, 3)
   )%>%
  filter(taxon!="Other")%>%
  mutate(taxon=fct_relevel(taxon,"MMUR","CMED","LCAT"))%>%
  ggplot(aes(sex,weight_g,fill=sex),na.rm=T)+
  geom_boxplot()+
  facet_wrap(vars(taxon))+
  scale_y_log10(name = "Weight in Grams",
    expand = expansion(mult = c(0, 0.06)))+
  scale_x_discrete(name=NULL,labels=c("Female","Male")
    )+
  theme_linedraw()+
  scale_fill_manual(
    values = c(Male = '#548EFB', Female = '#EF6178'))+
  labs(fill='Sex of Lemur')
```



Discussion: *Sex of the lemurs appears to have a small effect on the median weight. Interestingly, we can see this effect in two directions. Comparing the boxplots in MMUR and CMED and looking at the values in the data table, we see that they have small to moderate different medians, both favoring females to have greater weight. However, a much bigger difference comes from the LCAT group, wherein males(1000g) have a much higher median weight than females(164g). From the boxplots, we see the distribution of weights is varied fairly widely between the taxonomies. For the males, The LCAT group has the largest distribution from 35g to 3400g with the median value of 1000g, while the CMED group have a distribution from 8g to 400g with a median value at 105g and the MMUR group has a much smaller distribution from 5g to 140g with a median value of to 49g. For the females, The LCAT group has the largest distribution from 40g to 3450g with the median value of 164g, while the CMED group have a distribution from 7g to 367g with a median value at 212g and the MMUR group has a much smaller distribution from 5g to 172g with a median value of to 65g.*

These ranges and medians suggest that for the groups MMUR and CMED, females are typically larger than males, but neither are typically big. For the LCAT group, the distribution and center suggest that males are usually larger than females and that they can get to be pretty big.

Part 2

Question: *How does the age category and the weight of the lemur influence its number of offspring? and what is the distribution of number of offspring between the age categories?*

Introduction: *To answer the question of Part 2, we will work with three variables, the lemur's number of known offspring (column n_known_offspring), what age category the lemur belongs to (column age_category), and what is the lemur's weight in grams (column weight_g). The number of known offspring is provided as a discrete numeric value. The age category of the lemur is a categorical nominal variable placing lemurs into the age groups, juvenile, young adult, and adult.Lastly, the weight of the lemur was given as a numeric* *continuous variable in grams*

Approach: *Our approach is to determine if there exists a linear relationship between our response variable number of offspring(n_known_offspring) and our explanatory variables age category(age_category) and weight(weight_g), to do this we will build a linear model and create an output of the model. This is the best most direct way to analyze whether there exists a linear relationship between our variables.*

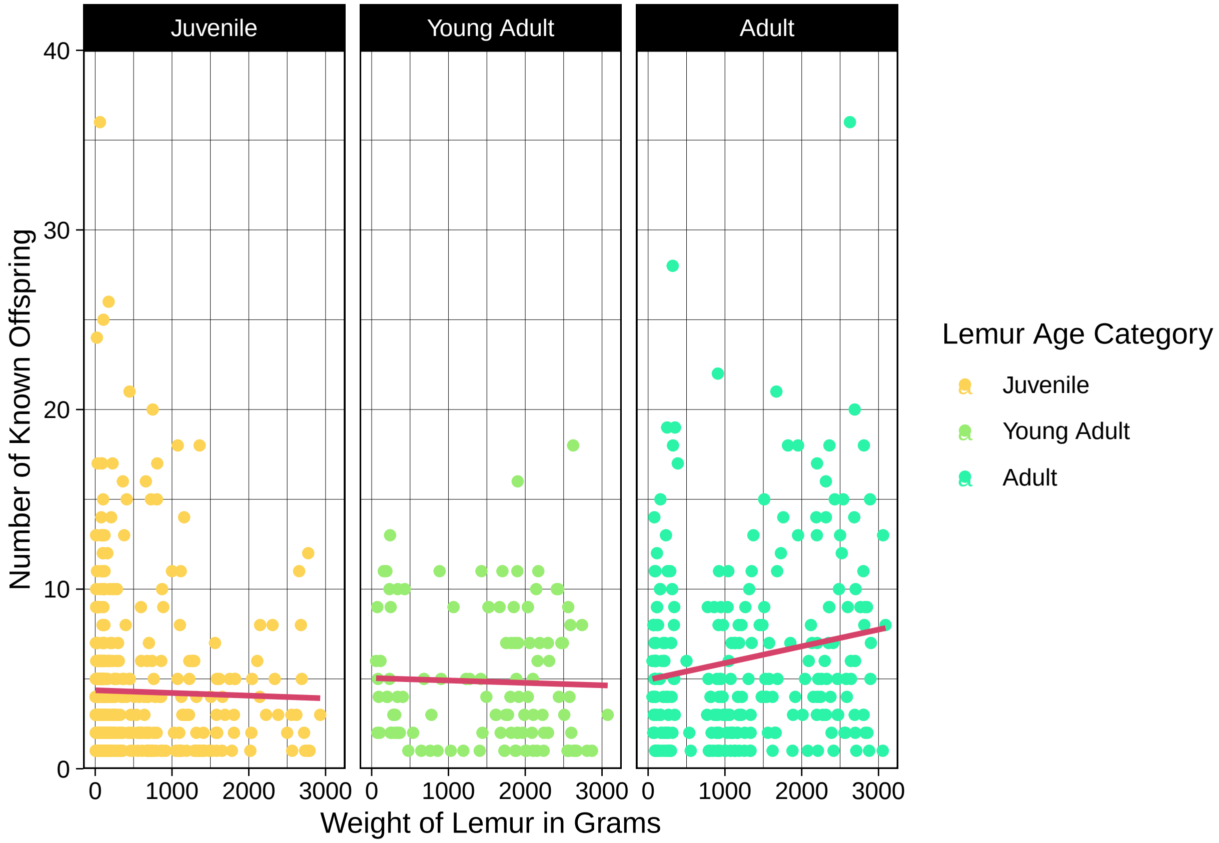
It will not be sufficient to just create a data output as we will not be able to easily comprehend its meaning alone. We will also need to visualize the data on a scatter plot to see the correlation of the variables and their trends. Additionally, to better understand the distribution of lemurs offspring between age groups, we will need to create a density plot between our variables number of offspring and age category. This will allow us to determine how many offspring lemurs have in each age category.

Analysis:

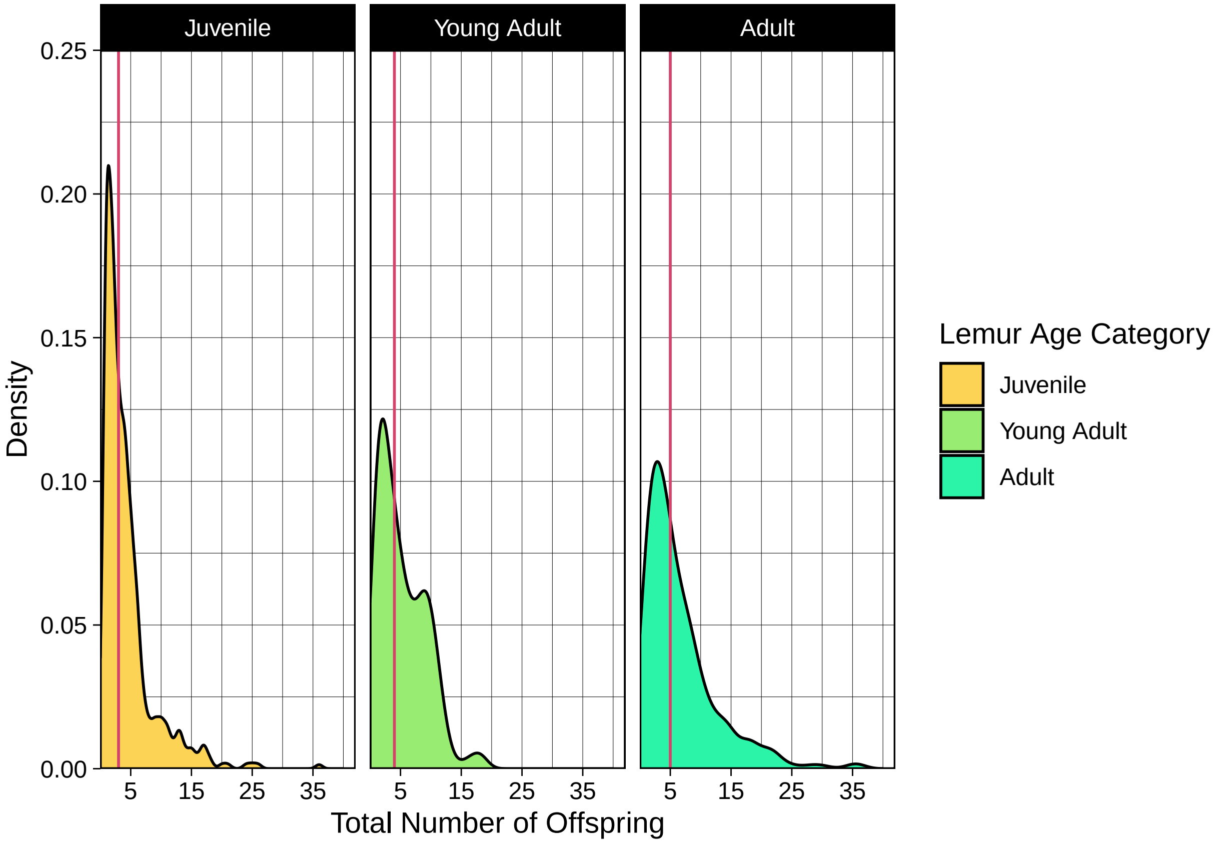
```
#This code fits the variables age category, number of known offspring and
#weight to a linear model then creates a summary table from the model
library(broom)
library(glue)
lm_summary<-df %>%
  nest(data = -age_category)%>%
  mutate(
    fit = map(data, ~lm(n_known_offspring ~ weight_g, data = .x)),
    glance_out = map(fit, glance)
    ) %>%
    select(age_category, glance_out) %>%
    unnest(cols = glance_out)

label_data <- lm_summary %>%
  mutate(
    rsqr = signif(r.squared, 2),
    pval = signif(p.value, 2),
    label = glue("R^2 = {rsqr}, P = {pval}"),
    weight_g = 3500, n_known_offspring = 45
  ) %>%
  select(age_category, label, weight_g, n_known_offspring)

lm_summary
```

```
## # A tibble: 3 x 13
##   age_category r.squared adj.r.squared sigma statistic p.value    df logLik
##   <fct>            <dbl>         <dbl> <dbl>     <dbl>   <dbl> <dbl>  <dbl>
## 1 Adult         0.0634         0.0603   5.70    20.5   8.80e-6     1  -960.
## 2 Young Adult   0.0173         0.00988  3.83     2.33  1.30e-1     1  -369.
## 3 Juvenile      0.000579      -0.00145  4.26     0.285 5.94e-1     1 -1416.
## # ... with 5 more variables: AIC <dbl>, BIC <dbl>, deviance <dbl>,
## #   df.residual <int>, nobs <int>
```

```
#This code plots the linear model into a scatterplot between weight and number
#of offspring faceted by age category and fitted with a trendline
df%>%
  ggplot(aes(weight_g, n_known_offspring,color=age_category)) +
  geom_point(na.rm = T) +
  geom_text(
    data = label_data, aes(label = label),
    size = 10/.pt, hjust = 1,na.rm = T
  ) +
  geom_smooth(method = "lm", se = FALSE,color='#D6436A',na.rm = T)+
  facet_wrap(vars(fct_relevel(age_category,"Juvenile","Young Adult","Adult")))+
  theme_linedraw()+
  scale_color_manual(
    values = c(Juvenile = '#FCD355', `Young Adult` = '#99EC72',Adult='#2BF4A8')
  )+
  scale_y_continuous(name = "Number of Known Offspring",limits = c(0,40),
                     expand = expansion(mult = c(0, 0)))+
  scale_x_continuous(name = "Weight of Lemur in Grams",limits = c(0,3100),
                     expand = expansion(mult = c(0.05, 0.05)))+
  labs(color='Lemur Age Category')
```



```
#Creating faceted density plots to determine what is the lemur's distribution
#of number of offspring between the age categories
df%>%
  group_by(age_category)%>%
  mutate(vline=median(n_known_offspring,na.rm = T))%>%
  ggplot(aes(x=n_known_offspring,fill=age_category)) +
  geom_density(na.rm = T)+
  facet_wrap(~ age_category)+
  theme_linedraw()+
  geom_vline(aes(xintercept=vline),color='#D6436A')+
    scale_fill_manual(
    values = c(Juvenile = '#FCD355', `Young Adult` = '#99EC72',Adult='#2BF4A8'))+
  scale_y_continuous(name = "Density",limits = c(0,0.25),
                     expand = expansion(mult = c(0, 0)))+
  scale_x_continuous(name = "Total Number of Offspring",limits = c(0,40),
                     expand = expansion(mult = c(0, 0.05)),breaks = c(5,15,25,35))+
  labs(fill="Lemur Age Category")
```



Discussion: *From the summary table we are able to see that age category has a significant positive effect on number of offspring. From the scatter plot, we can see that in the adult category weight has a positive relationship on number of offspring as weight increases. In the juvenile and young adult category, there seems to be no effect on number of offspring from increase of weight. This trend suggests that only once the lemurs reach a certain weight do they begin to be able to contend for more breeding access.*

From looking at the density plot between age category and number of offspring, we can see that in the juvenile category most lemurs have a median of about three offspring, but once they get to the young adult category, the median shifts and lemurs have around four offspring. Finally, when lemurs are adults the distribution shifts again, and the median approaches five, but more noticeably there are more lemurs with tens of offspring. The values in the adult category suggest that most lemurs get to breed, but that certain lemurs are able to breed considerably more.