Project 4

This is the dataset you will be working with:

```
lemurs <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/20

lemurs
```

```
## # A tibble: 82,609 x 54
##    taxon dlc_id hybrid sex   name      current_resident stud_book dob
##    <chr> <chr>  <chr>  <chr> <chr>     <chr>            <chr>     <date>
##  1 OGG   0005   N      M     KANGA     N                <NA>      1961-08-25
##  2 OGG   0005   N      M     KANGA     N                <NA>      1961-08-25
##  3 OGG   0006   N      F     ROO       N                <NA>      1961-03-17
##  4 OGG   0006   N      F     ROO       N                <NA>      1961-03-17
##  5 OGG   0009   N      M     POOH BEAR N                <NA>      1963-09-30
##  6 OGG   0009   N      M     POOH BEAR N                <NA>      1963-09-30
##  7 OGG   0009   N      M     POOH BEAR N                <NA>      1963-09-30
##  8 OGG   0010   N      M     EEYORE    N                <NA>      1964-05-20
##  9 OGG   0010   N      M     EEYORE    N                <NA>      1964-05-20
## 10 OGG   0014   N      F     ROOLETTE  N                <NA>      1964-10-27
## # ... with 82,599 more rows, and 46 more variables: birth_month <dbl>,
## #   estimated_dob <chr>, birth_type <chr>, birth_institution <chr>,
## #   litter_size <dbl>, expected_gestation <dbl>, estimated_concep <date>,
## #   concep_month <dbl>, dam_id <chr>, dam_name <chr>, dam_taxon <chr>,
## #   dam_dob <date>, dam_age_at_concep_y <dbl>, sire_id <chr>, sire_name <chr>,
## #   sire_taxon <chr>, sire_dob <date>, sire_age_at_concep_y <dbl>, dod <date>,
## #   age_at_death_y <dbl>, age_of_living_y <dbl>, age_last_verified_y <dbl>, ...
```

More information about the dataset can be found here: https://github.com/rfordatascience/tidytuesday/tree/master/data/2021/2021-08-24 and https://www.nature.com/articles/sdata201419.

Question: *Does the weight of a lemur related to the life span? What is the amount of this variation across taxonomies such as EFLA [Blue-eyed black lemur], LCAT [Ring-tailed lemur], VVV [Black and white ruffed lemur] ?What do you observe ?*

Introduction: *The dataset we are working with contains life history records information for over 27 taxonomies that have lived at the Duke Lemur center from 1966 to 2014. Various life-history parameters such as weight, gender, age, number of off-springs and parental information is recorded in the dataset.*

In this project, we will be asking whether there is any relationship between ***weight*** *and* ***life-span*** *of a lemur across specific taxonomies. Moreover, we will also be computing the amount of this variation across taxonomies. Here, we will be working with 3 columns from the dataset*

1. ***taxon****: It lists the taxonomies of the lemur by its taxonomic code. It's a categorical variable. In total, there are 27 taxonomies in this dataset. But we are interested only in EFLA [Blue-eyed black lemur], LCAT [Ring-tailed lemur], VVV [Black and white ruffed lemur]*

2. ***age_max_live_or_dead_y****: It is a quantitative variable which indicates the maximum age the animal could have achieved*

3. ***weight_g****: It is a quantitative variable which indicates the weight of the lemur taken on a specific date in grams.*

Approach: *To clean the data, we will be filtering the* ***taxon*** *variable with the taxomoies we are interested in. Moreover, we will clean the NA values from* ***age_max_live_or_dead_y*** *or* ***weight_g****. Then we will fit the linear model with lm() and then use map to fit these models to groups of data (ie. taxon). Then, we will show the summary of these fit that gives us information about the distribution of the residuals. Amount of the variation can also be derived on the basis of that. Based on the distribution, we will plot the 2 variables using geom_smooth() to see the pattern. geom_smooth() is useful to see the patterns in case of overplotting.*

Analysis:

Cleaning the dataset and statistical modelling

```
#Cleaning the data set
lm_cleaned <- lemurs %>%
  # Filtering the taxons
  filter(taxon %in% c("EFLA", "LCAT", "VVV"))%>%
  filter(!is.na(age_max_live_or_dead_y)) %>%
  filter(!is.na(weight_g))%>%
  mutate(
   taxon = case_when(
      taxon == "EFLA" ~ "Blue-eyed black lemur",
      taxon == "LCAT" ~ "Ring-tailed lemur",
      taxon == "VVV" ~ "Black and white ruffed lemur",
      TRUE ~ NA_character_
    ))

# lm_out <- lm(age_max_live_or_dead_y ~ weight_g, data = lm_cleaned)
# summary(lm_out)

lm_summary <- lm_cleaned %>%
  nest(data = -taxon)%>%
  mutate(
    fit = map(data, ~lm(age_max_live_or_dead_y ~ weight_g, data =.x)),
    glance_out = map(fit, glance)
  ) %>%
  select(taxon, glance_out) %>%
  unnest(cols = glance_out)

lm_summary
```

```
## # A tibble: 3 x 13
##   taxon   r.squared adj.r.squared sigma statistic   p.value    df  logLik    AIC
##   <chr>       <dbl>         <dbl> <dbl>     <dbl>     <dbl> <dbl>   <dbl>  <dbl>
## 1 Black ~     0.131         0.131 10.3       412. 1.71e- 85     1 -10268. 20543.
## 2 Ring-t~     0.233         0.233  7.30     2280. 0             1 -25520. 51047.
## 3 Blue-e~     0.227         0.227  6.90      947. 1.96e-182     1 -10776. 21558.
## # ... with 4 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>,
## #   nobs <int>
```

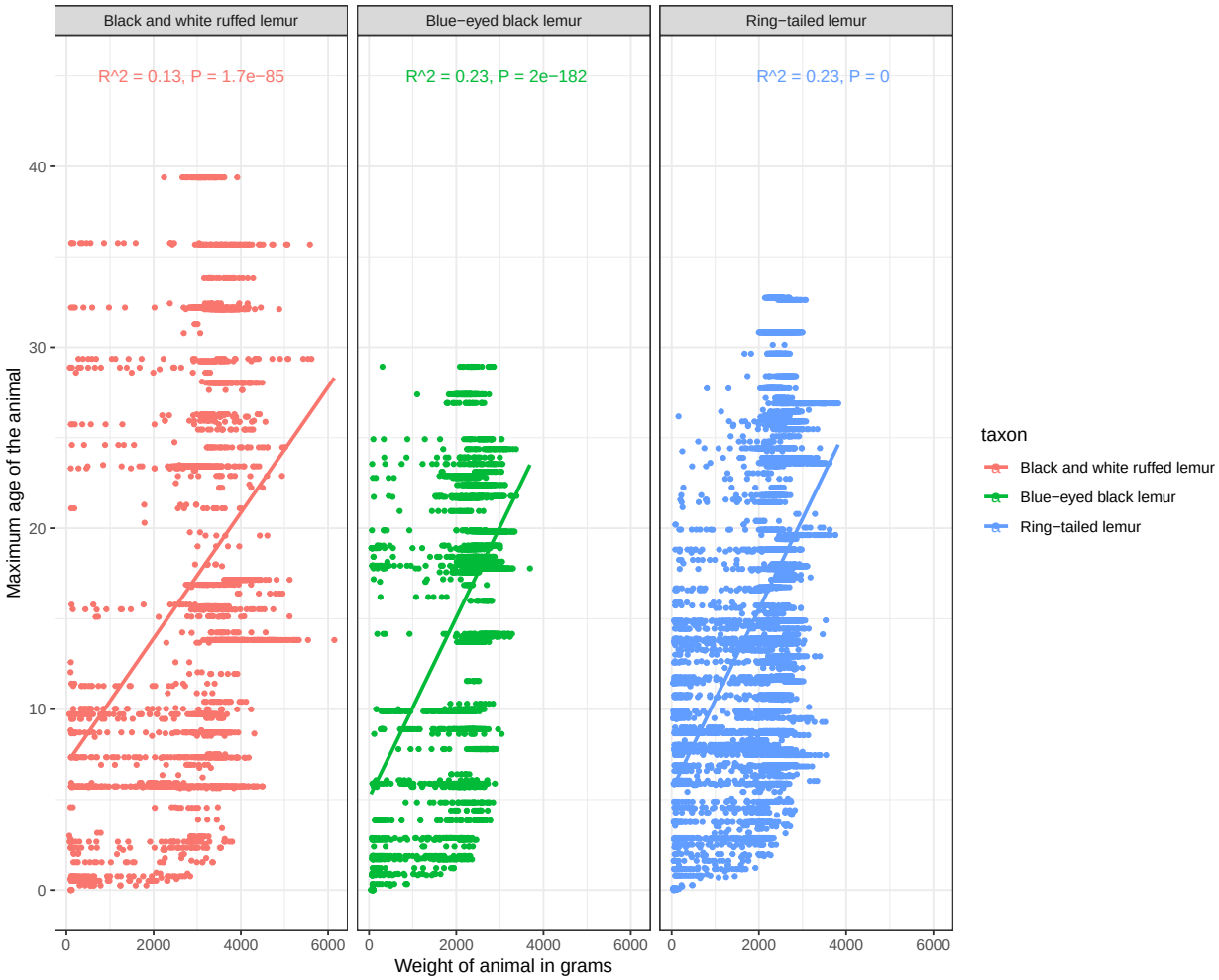
Plot

```
label_data <- lm_summary %>%
  mutate(
    rsqr = signif(r.squared, 2),  # round to 2 significant digits
    pval = signif(p.value, 2),
    label = glue("R^2 = {rsqr}, P = {pval}"),
    weight_g = 5000, age_max_live_or_dead_y = 45 # label position in plot
  ) %>%
```

```
  select(taxon, label, age_max_live_or_dead_y, weight_g)


ggplot(data = lm_cleaned, aes(weight_g, age_max_live_or_dead_y, color = taxon, na.rm = TRUE))+
  geom_point(size = 1)+
  geom_text(
    data = label_data,aes(label = label),
    size = 10/.pt, hjust=1
  )+
  geom_smooth(method = "lm", se = FALSE)+
  xlab("Weight of animal in grams")+
  ylab("Maximum age of the animal")+
  facet_wrap(vars(taxon))+
  theme_bw()
```

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



Discussion: *We can see that the weight and lifespan relationship is not perfectly linear in all the three taxons In case of black and white ruffled lemur, as weight increases, life-span doesn't seem to increase uptil the weight is 2000, but then one can see some increase in lifespan.In case of blue-eyed black lemur, as weight increases, life-span remains constant uptil the weight is 1500, then one can see increase in life-span. Also, there are clusters observed when the lifespan is between 15 to 20 and weight is between 2000 to 3000. In case*

of ring-tailed lemur, as weight increases, life-span remains constant only for a while, but then we can see there is increase in lifespan. One can see that p value is very less in case of all the three taxonomies. So, the relationship can be said as statistically significant. Also, the relationship doesn't vary much among these three taxons as seen by the r-squared values and p-values.