

Project title

Proposal

```
library(tidyverse)
```

Data 1

Introduction and data

- The dataset ‘lemur_data.csv’ comes from Kaggle.com
- The dataset was originally collected from the [2019 data release from the Duke Lemur Center Database](#) by Zehr, SM, Roach RG, Haring D, Taylor J, Cameron FH, Yoder AD.
- This dataset contains information on over 3,500 observations. Each observation represent a lemur, including lemur-information such as ancestry, reproduction, longevity, and body mass (in total 54 columns).

Research question

- What are the top 3 factors that influence the lifespan of lemurs?
- Through our preliminary investigation, we found that the death age of lemurs varies a lot ranging from 0 to 35. Thus, we are interested in researching on what are the determining factors of lemurs’ lifespan. Our hypotheses is that taxon, sex, weight are the top 3 factors that would affect the lifespan of lemurs.
- There are both categorical and quantitative variables involved in our research questions since categorical variables such as taxon and quantitative variables such as weight can all play a role in their lifespan.

Glimpse of data

```
lemur <- read_csv("data/lemur_data.csv")
```

Rows: 82609 Columns: 54

```
-- Column specification -----  
Delimiter: ","  
chr  (19): Taxon, DLC_ID, Hybrid, Sex, Name, Current_Resident, StudBook, Est...  
dbl  (27): Birth_Month, Litter_Size, Expected_Gestation, Concep_Month, Dam_A...  
date  (8): DOB, Estimated_Concep, Dam_DOB, Sire_DOB, DOD, Weight_Date, Conce...
```

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.

```
glimpse(lemur)
```

Rows: 82,609

Columns: 54

```
$ Taxon          <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~  
$ DLC_ID         <chr> "0005", "0005", "0006", "0006", "0009", "000~  
$ Hybrid        <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", ~  
$ Sex           <chr> "M", "M", "F", "F", "M", "M", "M", "M", "M", ~  
$ Name          <chr> "KANGA", "KANGA", "ROO", "ROO", "POOH BEAR", ~  
$ Current_Resident <chr> "N", "N", "N", "N", "N", "N", "N", "N", "N", ~  
$ StudBook      <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~  
$ DOB           <date> 1961-08-25, 1961-08-25, 1961-03-17, 1961-03~  
$ Birth_Month   <dbl> 8, 8, 3, 3, 9, 9, 9, 5, 5, 10, 10, 6, 6, 3, ~  
$ Estimated_DOB <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~  
$ Birth_Type    <chr> "CB", "CB", "CB", "CB", "CB", "CB", "CB", "CB", "C~  
$ Birth_Institution <chr> "Duke Lemur Center", "Duke Lemur Center", "D~  
$ Litter_Size   <dbl> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, ~  
$ Expected_Gestation <dbl> 129, 129, 129, 129, 129, 129, 129, 129, 129, ~  
$ Estimated_Concep <date> 1961-04-18, 1961-04-18, 1960-11-08, 1960-11~  
$ Concep_Month  <dbl> 4, 4, 11, 11, 5, 5, 5, 1, 1, 6, 6, 1, 1, 11, ~  
$ Dam_ID        <chr> "0001", "0001", "0001", "0001", "0001", "000~  
$ Dam_Name      <chr> "WHITE-TAIL", "WHITE-TAIL", "WHITE-TAIL", "W~  
$ Dam_Taxon     <chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~  
$ Dam_DOB       <date> 1959-01-28, 1959-01-28, 1959-01-28, 1959-01~  
$ Dam_AgeAtConcep_y <dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32, 4.~  
$ Sire_ID       <chr> "0002", "0002", "0002", "0002", "0007", "000~
```

\$ Sire_Name	<chr> "BRUISER", "BRUISER", "BRUISER", "BRUISER", ~
\$ Sire_Taxon	<chr> "OGG", "OGG", "OGG", "OGG", "OGG", "OGG", "O~
\$ Sire_DOB	<date> 1959-01-28, 1959-01-28, 1959-01-28, 1959-01~
\$ Sire_AgeAtConcep_y	<dbl> 2.22, 2.22, 1.78, 1.78, 4.32, 4.32, 4.32, 4.~
\$ DOD	<date> 1977-02-07, 1977-02-07, 1974-10-15, 1974-10~
\$ AgeAtDeath_y	<dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38, 10~
\$ AgeOfLiving_y	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ AgeLastVerified_y	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, 14.16, 1~
\$ AgeMax_LiveOrDead_y	<dbl> 15.47, 15.47, 13.59, 13.59, 10.38, 10.38, 10~
\$ N_known_offspring	<dbl> 7, 7, 9, 9, 1, 1, 1, 7, 7, 5, 5, 4, 4, 1, 1, ~
\$ DOB_Estimated	<chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ Weight_g	<dbl> 1086, 1190, 947, 1174, 899, 917, 910, 1185, ~
\$ Weight_Date	<date> 1972-02-16, 1972-06-20, 1972-02-16, 1972-06~
\$ MonthOfWeight	<dbl> 2, 6, 2, 6, 2, 2, 6, 2, 6, 2, 6, 2, 6, 2, 6, ~
\$ AgeAtWt_d	<dbl> 3827, 3952, 3988, 4119, 3061, 3074, 3188, 28~
\$ AgeAtWt_wk	<dbl> 546.71, 564.57, 569.71, 588.43, 437.29, 439.~
\$ AgeAtWt_mo	<dbl> 125.82, 129.93, 131.11, 135.42, 100.64, 101.~
\$ AgeAtWt_mo_NoDec	<dbl> 125, 129, 131, 135, 100, 101, 104, 92, 97, 8~
\$ AgeAtWt_y	<dbl> 10.48, 10.83, 10.93, 11.28, 8.39, 8.42, 8.73~
\$ Change_Since_PrevWt_g	<dbl> NA, 104, NA, 227, NA, 18, -7, NA, 51, NA, 71~
\$ Days_Since_PrevWt	<dbl> NA, 125, NA, 131, NA, 13, 114, NA, 125, NA, ~
\$ Avg_Daily_WtChange_g	<dbl> NA, 0.83, NA, 1.73, NA, 1.38, -0.06, NA, 0.4~
\$ DaysBeforeDeath	<dbl> 1818, 1693, 972, 841, 728, 715, 601, 2086, 1~
\$ R_Min_Dam_AgeAtConcep_y	<dbl> 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0.59, 0.~
\$ Age_Category	<chr> "adult", "adult", "adult", "adult", "adult", ~
\$ Preg_Status	<chr> "NP", "NP", "NP", "NP", "NP", "NP", "NP", "N~
\$ Expected_Gestation_d	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ ConcepDate_IfPreg	<date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ InfantDOB_IfPreg	<date> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ DaysBeforeInfBirth_IfPreg	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ Pct_PregRemain_IfPreg	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
\$ InfantLitSz_IfPreg	<dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~

Data 2

Introduction and data

- The data comes from CORGIS (Collection of Really Great, Interesting, Situated Datasets) website.
- The earthquake data was originally collected on 6/7/2016 from the United States Geological Survey by Ryan Whitcomb.

- Each observation contains a unique earthquake and all the information surrounding the circumstances of the earthquake, including but not limited to: coordinates of where the earthquake occurred, its magnitude, time of earthquake, and the depth of the earthquake.

Research question

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

Glimpse of data

```
earthquakes <- read.csv("data/earthquakes.csv")

glimpse(earthquakes)
```

Rows: 8,394

Columns: 18

```
$ id          <chr> "nc72666881", "us20006i0y", "nc72666891", "nc72666~
$ impact.gap  <dbl> 122.00000, 30.00000, 249.00000, 122.00000, 113.610~
$ impact.magnitude <dbl> 1.43, 4.90, 0.06, 0.40, 0.30, 1.80, 1.00, 2.00, 1.~
$ impact.significance <int> 31, 371, 0, 2, 1, 50, 15, 62, 22, 43, 4, 12, 4, 4,~
$ location.depth <dbl> 15.120, 97.070, 4.390, 1.090, 7.600, 1.300, 2.452,~
$ location.distance <dbl> 0.10340000, 1.43900000, 0.02743000, 0.02699000, 0.~
$ location.full <chr> "13km E of Livermore, California", "58km WNW of Pa~
$ location.latitude <dbl> 37.67233, 21.51460, 37.57650, 37.59583, 39.37750, ~
$ location.longitude <dbl> -121.6190, 94.5721, -118.8592, -118.9948, -119.845~
$ location.name <chr> "California", "Burma", "California", "California",~
$ time.day      <int> 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27, 27~
$ time.epoch    <dbl> 1.469593e+12, 1.469593e+12, 1.469594e+12, 1.469594~
$ time.full     <chr> "2016-07-27 00:19:43", "2016-07-27 00:20:28", "201~
$ time.hour     <int> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1,~
$ time.minute   <int> 19, 20, 31, 35, 41, 52, 53, 58, 3, 4, 9, 13, 17, 1~
$ time.month    <int> 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7, 7,~
$ time.second   <int> 43, 28, 37, 44, 59, 52, 35, 45, 0, 32, 51, 31, 18,~
$ time.year     <int> 2016, 2016, 2016, 2016, 2016, 2016, 2016, 2016, 20~
```

Data 3

Introduction and data

- The data is combined from Niche’s “2023 Best Colleges in America” list (<https://www.niche.com/colleges/search/best-colleges/>) and the U.S. Department of Education.
- State when and how it was originally collected (by the original data curator, not necessarily how you found the data). The Niche data was scraped by Maia on October 17-19 2022.
- Write a brief description of the observations.

Research question

- A well formulated research question. (You may include more than one research question if you want to receive feedback on different ideas for your project. However, one per data set is required.)
- A description of the research topic along with a concise statement of your hypotheses on this topic.
- Identify the types of variables in your research question. Categorical? Quantitative?

Glimpse of data

```
niche_data_500 <- read_csv("data/niche_data_500.csv")
```

```
Rows: 500 Columns: 2
```

```
-- Column specification -----
```

```
Delimiter: ","
```

```
chr (1): college
```

```
dbl (1): rank
```

```
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.
```

```
# load DoE data
```

```
# colleges <- left_join(niche_data_500, ___)
```