Predicting Lemur's Lifespan*

Khushaal Nandwani

November 24, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Possible topics

2 Introduction

We are aiming to predict the lifepspan of a lemur. Specifically, using sex, specis, birth month of lemurs. We have also divided the dataset into two parts: one for lemurs in captivity and one for lemurs in wild. The goal is to realize the success we have achievd by captiving lemurs, see areas of improvement in the pre existing methods.

1. What factors can contribute to lifespan of lemur?

    - sex
    - species
    - birth month (season): some seasons makes stronger. maybe winter?
    - captivity vs wild: captivity makes them use to spoon feeding?
    - litter size: maybe mother is more involved with other kids
    - mother_species and father_species is same or not

3 Data

The data was taken from Cookson (2020), who acquired it from Duke Lemur Center more about which, can be found in Section 3.1. We used R Core Team (2023)

It is important to know about Lemurs and what affects their lifespan because they are the most endangered mammals on the planet. The Duke Lemur Center (DLC) is a global leader

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

in the research, care, and conservation of lemurs. The DLC hosts the most diverse population of lemurs outside their native habitat in Madagascar.

To do so, I chose the following variables from the dataset, which I believe are the most important factors that can affect the lifespan of a lemur:

- `Sex`: The sex of the lemur on birth. It is a cateogrical variable that can be `M` or `F`.
- `Species`: The species of the lemur. It is also a categorical variable and can take one of the following values: `GG`, `COL`, `UL`, `RUF`, `MOH`, `MAC`, `CAT`, `FUL`, `ALB`, `AR`, `VV`, `COQ`, `MED`, `MUR`, `COU`, `TAR`, `PYG`, `ZAZ`, `MON`, `RUB`, `COR`, `SAN`, `FLA`, `MAD`, `POT`. The specific or common names of these species can be found in Appendix A.
- `Genus`: The genus of the lemur. It is a categorical variable and can take one of the following values: `O`, `E`, `G`, `H`, `L`, `V`, `P`, `C`, `M`, `N`, `D`.
- `month_born`: The month in which the lemur was born. It is a categorical variable and can take one of the following values: `1`, `2`, `3`, `4`, `5`, `6`, `7`, `8`, `9`, `10`, `11`, `12`, where `1` corresponds to January, `2` to February, and so on.

Finally, we have the target variable `Age` which is the lifespan of the lemur in years. It is a continuous variable.

The following plots show the distribution of the lemurs in the dataset based on the variables `Sex`, `Species`, `Genus`, and `month_born` for both lemurs in captivity and in the wild.

For `Sex`, we observe that in the wild, the distribution of male and female lemurs is nearly even, with only a slight skew. While in captivity, there are significantly more female lemurs than males. The larger count of female lemurs in captivity could result from conservation breeding programs prioritizing females to ensure the species' survival.

Looking at the distribution of `Species`. We observe that in the wild, the distribution is relatively balanced, with a few exceptions of `MAD`, `CAT` and `MED` which appear to be low in number. In captivity, certain species (e.g., MUR, CAT, MED) are more represented, while others are rare or absent. This can be due to captive environments may focus on saving the most endangered species, as we see the ones least in wild like `CAT` and `MED` are more in captivity.

With the exception of genus `E`, genus representation is balanced in both wild and captive lemurs. `E` might be high in both cases, as it could be a common genus or have a higher conservation priority.

For birth month, births in the wild are concentrated in a few months, suggesting a breeding season. While in captivity, births are more evenly spread across the year, with some peaks in specific months. Seasonal breeding in the wild is influenced by environmental factors like food availability and climate. Captivity can disrupt these natural cycles, as controlled environments and year-round resources allow for more frequent and less seasonal reproduction.

For `Age`, we observe that in the wild, the age distribution of lemurs shows a higher frequency of younger lemurs, with relatively few individuals reaching older ages. In captivity, the age

distribution extends further, with a significant number of lemurs living to older ages. This difference can be attributed to the controlled environments in captivity, which provide consistent food, medical care, and protection from predators, thus increasing longevity. In contrast, the challenges of the wild, such as predation, disease, and fluctuating resources, contribute to shorter lifespans. With there being one exception, we observe signigicant number of captive lemurs below age 1, this can be due to the extensive breeding of captivity or could be a sign of data quality issues.

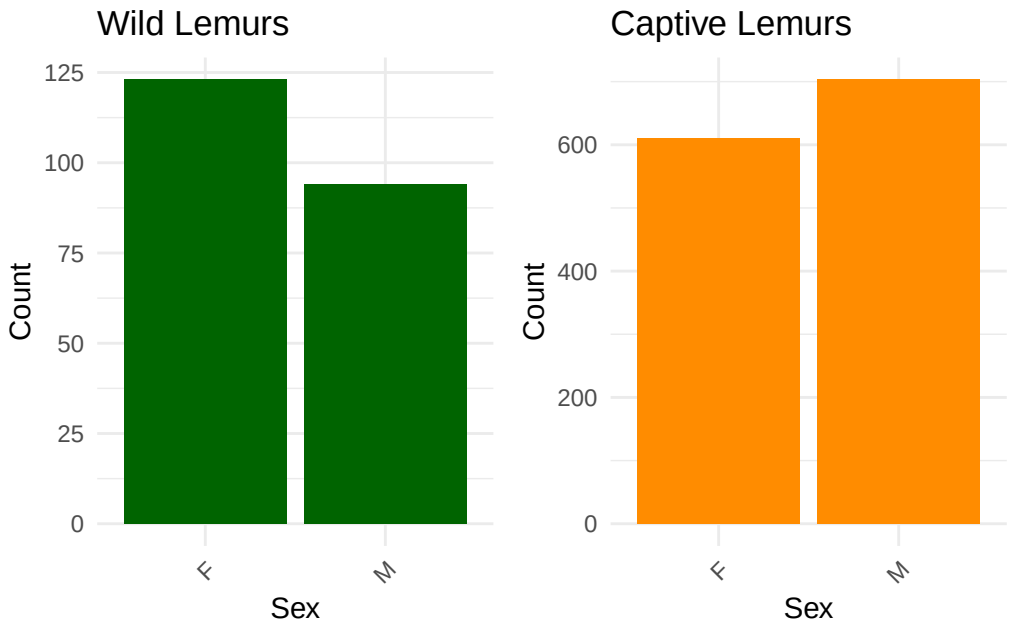

Figure 1: Distribution of Lemurs by Sex

We will also look at the distribution of the target variable `Age` in the dataset. The following plot shows the distribution of the ages of the lemurs in the dataset for both lemurs in captivity and in the wild.

3.1 Data Collection

The data in these sources was acquired and processed by staff at the Duke Lemur Center (DLC).

3.1.1 Data Acquisition

As Zehr et al. (2014) points out, DLC staff collected data about the lemurs according to standard operating procedures and USDA, AZA, and IACUC guidelines. They recorded information about births, deaths, weights, enclosure moves, behaviors, and other significant events

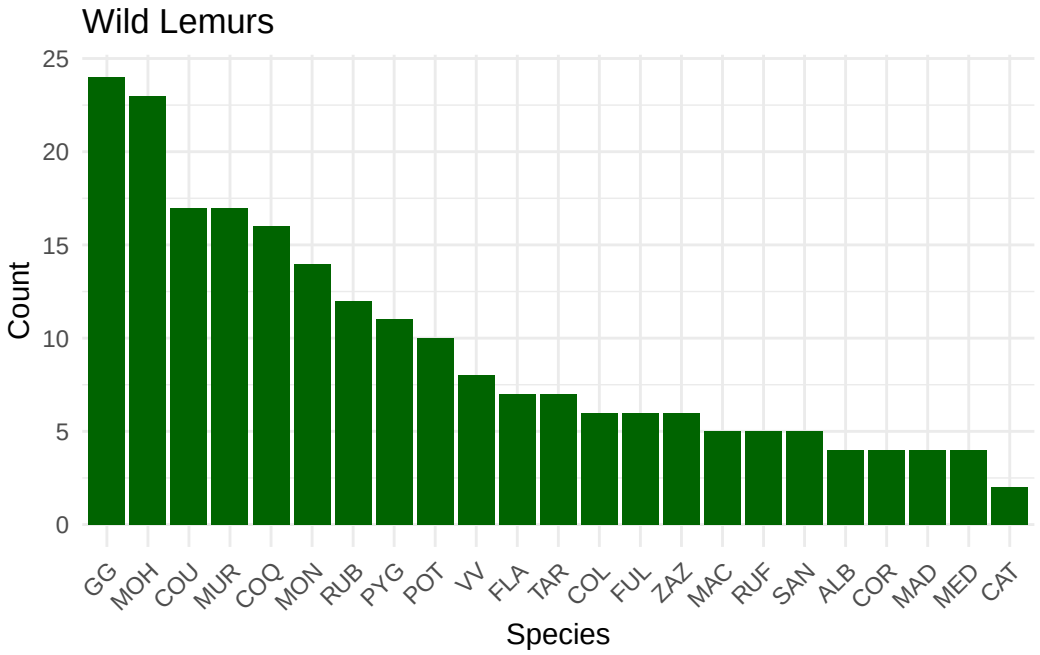

Figure 2

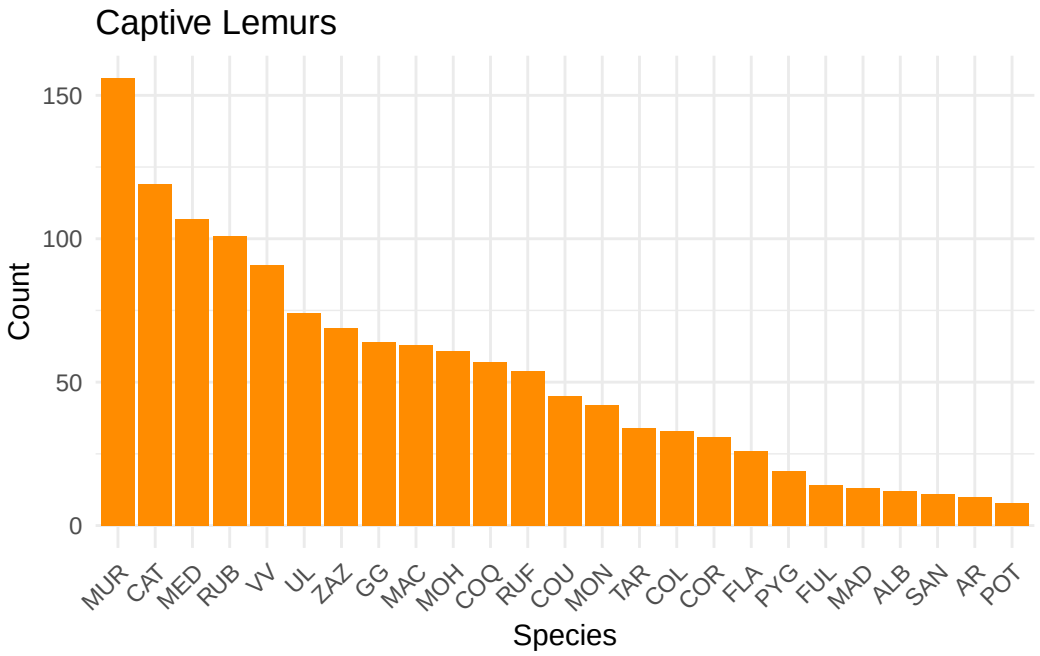

Figure 3

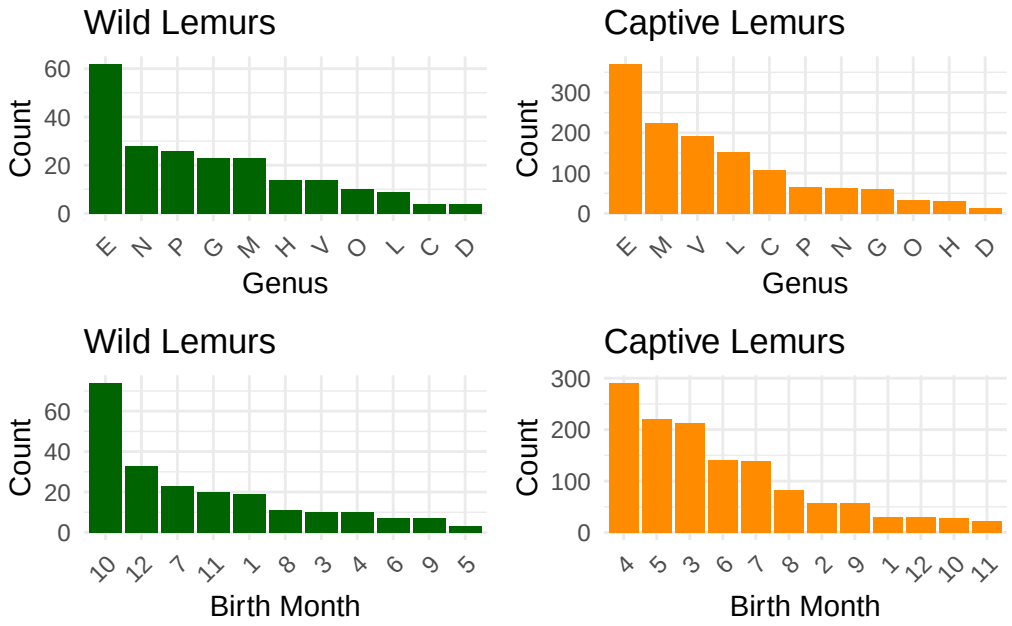

Figure 4

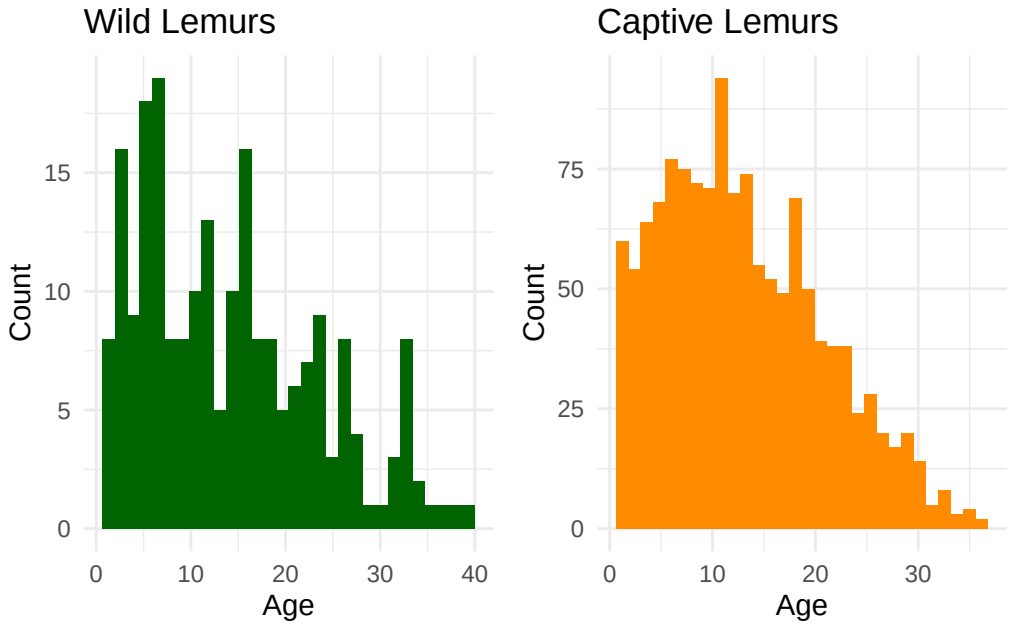

Figure 5: Distribution of ages of lemurs in the dataset for both lemurs in captivity and in the wild.

on a daily basis. Originally, this data was stored in handwritten and typed paper formats. Later, it was computerized.

In the mid-1990s, the DLC started using two databases: the Animal Record Keeping System (ARKS) and MedARKS. These databases allowed the DLC to share information with other organizations through the International Species Information System (ISIS). The DLC is currently transitioning to using the Zoological Information Management System (ZIMS). Data not stored in these databases has been stored in spreadsheets, and the DLC is working on transferring data from older records into these databases.

3.1.2 Data Processing

The DLC used SAS software to build a database for the lemur data. Data from various sources was imported into SAS Enterprise Guide, including ARKS, MedARKS, ZIMS, and spreadsheets. They wrote programs in SAS to extract, match, and join data, calculate new variables, and format the output. They also used tools within SAS Enterprise Guide Projects for calculations and formatting. The DLC uses a unique ID to match data for individual animals, and the taxonomic name for species-related variables. The data was validated by identifying and locating missing data, standardizing codes and text, investigating outliers, and comparing known values to the database output.Data that could not be verified was excluded from the published dataset.

The DLC created two data files from the database: **the DLC Animal List**, which contains single-copy variables for each animal in the colony's history, and **the DLC Weight File**, which contains all weight measurements for each animal. We used the DLC Animal List for this analysis, which was first cleaned by Cookson (2020) and then we clean it as per our needs.

The data in these sources was updated on February 8, 2019. The DLC plans to update the data on a yearly basis.

4 Why Lemur?

Established in 1966 on Duke University's campus in Durham, NC, the Duke Lemur Center (DLC) is a global leader in the research, care, and conservation of lemurs, the planet's most endangered group of mammals. Home to over 200 animals spanning 13 species, the DLC hosts the most diverse population of lemurs outside their native habitat in Madagascar. cite(https://lemur.duke.edu)

5 Model

I used Generalized Linear Model (GLM) framework with a Gaussian family distribution to determine the expected lifespan of a lemur based on the selected variables. GLM is a statistical model that generalizes linear regression to include non-normal distributions. It is used when the dependent variable is not normally distributed or when the relationship between the dependent and independent variables is not linear. The Gaussian family distribution is used when the dependent variable is continuous and normally distributed, which was in my case.

My model will be based on `Sex`, `Species`, `Genus`, and `month_born` as the independent variables and `Age` as the dependent variable.

Mathematically, the model can be represented as:

$$\bar{a} = \beta_0 + \beta_1 \times \texttt{Sex} + \beta_2 \times \texttt{Species} + \beta_3 \times \texttt{Genus} + \beta_4 \times \texttt{month\_born} + \epsilon \tag{1}$$

$$\begin{aligned}
\bar{a}_{\text{wild}} &\sim \text{Normal}(16, 4) \\
\bar{a}_{\text{captive}} &\sim \text{Normal}(26, 4) \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5) \\
\beta_2 &\sim \text{Normal}(0, 2.5) \\
\beta_3 &\sim \text{Normal}(0, 2.5) \\
\beta_4 &\sim \text{Normal}(0, 2.5)
\end{aligned}$$

where,

- $\bar{a}_{\text{wild}}$ is the expected lifespan of a lemur in the wild.
- $\bar{a}_{\text{captive}}$ is the expected lifespan of a lemur in captivity.
- β_0 is the intercept term. It represents the expected lifespan of a lemur when all other variables are zero.
- β_1 is the coefficient for `Sex`. It represents the change in the expected lifespan of a lemur for a one-unit change in `Sex`.
- β_2 is the coefficient for `Species`. It represents the change in the expected lifespan of a lemur for a one-unit change in `Species`.
- β_3 is the coefficient for `Genus`. It represents the change in the expected lifespan of a lemur for a one-unit change in `Genus`.
- β_4 is the coefficient for `month_born`. It represents the change in the expected lifespan of a lemur for a one-unit change in `month_born`.
- ϵ is the error term.

We chose a Gaussian family distribution for the model because the dependent variable `Age` is continuous and normally distributed. From Cape May County, NJ (2024) we found that the expected lifespan of a lemur in wild is upto 18 years, and in captivity about 30 years. So, we set the prior for the expected lifespan of a lemur in the wild to be Normal$(16, 4)$ and in captivity to be Normal$(26, 4)$. We set the priors for the coefficients to be Normal$(0, 2.5)$, because we do not have any prior information about the effect of the independent variables on the dependent variable.

Using moderately broad priors can help regularize the model, reducing the risk of overfitting and make us have more stable estimates, especially in scenarios with limited data.

(**app_model-dataset?**) covers coefficients, standard errors, convergence checks and other model diagnostics for our model.

6 Appendix

A Data Cleaning

I took the following steps to clean the raw data and prepare it for analysis:

1. Loading the Data: The raw data is loaded from the file data/raw_data/animals.csv.
2. Column Selection: The dataset is reduced to relevant columns: `animal_id`, `taxonomic_code`, `sex`, `birth_date`, `death_date`, `birth_type`, `litter_size`, `mother_species`, and `father_species`.
3. Filtering Rows: Rows where death_date is missing are removed to ensure the data includes only animals with complete lifecycle information.
4. Extracting Genus and Species: The taxonomic_code column is split into two new columns: `genus` (first letter) and `species` (remaining characters). The original taxonomic_code column is removed as it is no longer needed.
5. Age Calculation: A new column, `age`, is calculated based on the difference between `death_date` and `birth_date`, measured in years.Lemurs with an age of less than 1 year are removed to focus on mature animals and ignore infant mortality. This was especially common for captive-born animals because excess breeding can lead to high infant mortality rates.
6. Birth Month Extraction: A new column, month_born, is created to indicate the birth month extracted from the birth_date. We needed this information to analyze seasonal effects on lifespan.

7. Splitting Data by Birth Type: The data is split into two subsets based on birth_type: one for animals born in captivity (`data_captive`) and another for those born in the wild (`data_wild`).
8. Column Removal for Wild Data: Since information on mother_species, father_species, and litter_size is unavailable for wild-born animals, these columns are removed from data_wild.
9. Handling Missing Values: Both datasets, data_wild and data_captive, are cleaned by dropping rows with missing values in key columns: sex, species, age, month_born, and genus.
10. Final Output: The cleaned datasets are saved as CSV files: Wild-born animals: `data/analysis_data/wild.csv` Captive-born animals: `data/analysis_data/captive.csv`

This systematic cleaning ensures the dataset is consistent, reliable, and ready for further analysis.

B Analysis Dataset

Hello

C Model Dataset

D Species of Lemurs

The following are the species of lemurs that are present in the dataset:

- `GG`: Genus *Genus*

Cape May County, NJ. 2024. "Ring-Tailed Lemur." https://capemaycountynj.gov/1112/Ring-Tailed-Lemur#:~:text=Captive%2FWild%20Lifespan%3A%20up%20to,May%20being%20prime%20breeding%20time.

Cookson, T. Alexander. 2020. "Duke Lemur Center Dataset." https://github.com/tacookson/data/tree/master/duke-lemur-center.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Zehr, Steven M., Ronald G. Roach, Danielle Haring, Julia Taylor, Fred H. Cameron, and Anne D. Yoder. 2014. "Life History Profiles for 27 Strepsirrhine Primate Taxa Generated Using Captive Data from the Duke Lemur Center." *Scientific Data* 1 (July): 140019. https://doi.org/10.1038/sdata.2014.19.