Hawks: A Data Analysis on 3 Different Hawk Species

Julia Cuellar

Summer 2021

https://github.com/BVUjac8/Hawks.git

Any surprises from your domain from these data?

A GitHub repository of R datasets in the form of CSV files.

Surprisingly no. While performing basic summary statistics of the hawks data set, it clearly showed the minimum for numerical variables, length (908 observations), the 1st quartile for numerical variables, median for numerical variables, class (character for naming column variables), mean for numerical variables, mode (character for categorical variables), third quartile for numerical variables, maximum for numerical variables, and NA's for numerical variables. After reviewing the references, it plainly presented various attributes and features about the three different species of hawks (Cooper's, Red-Tailed, & Sharp-Shinned). Thus, the domain from this data set is just in tune for this data analysis.

The dataset is what you thought it was?

R: Measurements on Three Hawk Species (vincentarelbundock.github.io)

This hawks data set was not what I expected. While performing cleaning of the data set, I noticed that were no nulls which I usually encounter from other data sets that I have worked with. However, from producing the simple summary statistics showcased this exact situation. There was also missing values, but with considering the research questions I proposed, those columns were not handled when employing more exploratory data analysis. Also, while working through the EDA of the hawks data set, it primarily exhibited what I was expecting from this data set.

Have you had to adjust your approach or research questions?

- Which species of hawks grew to maturity (adulthood)?
- Which species of hawks migrated south in the month of November?
- Which species of hawks has a lengthier hallux (killing talon)?
- Does the time of capture have anything to do with the maturity of species of hawks?

Not at the moment. I worked through rudimentary EDA upon the hawks data set as well as showcased various plots from the key variables that I was applying from my research questions. There is three species of hawks, Cooper's, Red-Tailed, & Sharp-Shinned, and out of those three based off the bar plot I created of the species was that the Red-Tailed was the most out the three, thus, they could possibly have a higher chance of reaching the most to maturity. Referred from the references about Sharp-Shinned, this species of hawks will probably be the most who migrated in November. Also, adhering to the references, the Sharp-Shinned hawk is the smallest hawk species and therefore, will not have the longest killing talons, so it is a toss-up between the other two, but most likely Red-Tailed. However, the final research question, I need to figure out how to approach this carefully.

Is your method working?

There is two methods to possibly perform upon the hawks data set: clustering and regression.

Yes, choosing the two machine learning techniques of clustering and regression will definitely unveil which species of hawks grew the most to adulthood, which species of hawks participated in migration during the November month, which species of hawks has the longest hallux or killing talon, yet, for determining whether maturity affects the capture rate of the three different species of hawks, this may be difficult to perform in the modeling step of the data analysis. Hopefully, I can keep the regression method for the last research question.

What challenges are you having?

The only data manipulation that needs to transpire upon this data set is removal of unnecessary columns and fixing the NA's.

I am doing all this analysis in R programming; I am familiar with R programming, but is not as efficient with it compared to Python, hence, still need to learn the various packages from the programming language. Ergo, I needed to harness dplyr and tidyverse for wrangling the hawks data set against removing unneeded columns and replacing NA's in the hallux column. To replace the NA's, you had to make sure you had the correct package first, then the correct format utilizing the replace_na function. After this, yielding to the modeling methods will be a challenge, but after doing preliminary EDA, it should be a lot easier.