Biodiversity in National Parks

Zachary W. Cronkwright

November 16, 2022

Abstract

This projected focused on a fictitious dataset from the United States National Parks Service (NPS). the data showed that the conservation status for a given species is highly related to the frequency at which it was observed in the previous seven days. Analysis also showed that only one species trnasitioned from a conservation status of endangered to in recovery over that same period - that species being Canis lupus. It is my recommendation that the conservation program for this species be adapted to the other endangered species to prevent catastrophic biodiversity loss.

1 Introduction

The National Park Service (NPS) is United States (U.S.) government agency under the direction of the U.S. Department of the Interior. The agency's directive is to provide educational and recreational services related to natural and cultural importance within the U.S. national parks.[1] To provide these services, the NPS employs approximately 20,000 people to maintain 423 national parks.[1, 2]

As a by result of human interaction, the World Wildlife Foundation is reporting an average drop of 69% in observable wildlife population around the world between 1970 and 2018.[3] They report that th U.S. alone has seen a 33% drop in that time frame. To counteract this so that they may continue to provide educational and recreation services to visitors, the NPS conducts regular surveys collecting observational data of the species inhabiting each of the national parks. This project seeks to analyze a fictitious dataset created by Codecademy based on such surveys. The project itself consists of visualizations of the observational and species information data to elucidate tendencies in conservation status and species category within four national parks.

2 Methods

The data were analyzed and visualized using Python 3 within a Jupyter Notebook implementation. The data were imported and stored using the `pandas` library. Data were then analyzed using `pandas` methods. Visualizations of the data were developed using the `matplotlib.pyplot` and `seaborn` Python library.

The dataset was generated by Codecademy using real life NPS observational data. The data was "collected" of a period of the previous seven days in the following national parks:

- Bryce Canyon National Park
- Great Smoky Mountains National Park
- Yellowstone National Park
- Yosemite National Park

The dataset consists of two `.csv` files titled `observations.csv` and `species_info.csv`.

2.1 Observational Data

The observational data is contained within the `observations.csv` dataset. This dataset consist of 23296 records, each with the three following variables:

- `scientific_name`: the two part name given to a species by the scientific community
- `park_name`: the name of the national park the species was observed in
- `observations`: the number of times the species was observed

2.2 Species Information Data

The species information data is contained within the `specie_info.csv` dataset. After cleaning, the file contained 5541 records with the following four variables:

- `category`: the animal category the species belongs to (i.e. mammal, fish, etc.)
- `scientific_name`: the two part name given to a species by the scientific community
- `common_names`: the common name reference for the species
- `conservation_status`: the conservation status as reported by NPS for the given species. Values include:
  - Not at Risk
  - Species of Concern
  - Threatened
  - Endangered
  - In Recovery

After cleaning was completed, `species_info.csv` and `observations.csv` were joined on the `scientific_name` column. The resultant dataset was called `master_data`.

3 Analysis and Results

3.1 Observational Data

The observational data contained within the `observations.csv` file was analyzed with focus directed towards the frequency with which each species was observed. In conducting preliminary exploratory data analysis (EDA) the follow questions arose:

1. The scientific name "Canis lupus" appeared 12 times within this dataset. If one assumes that each species is present in each park, one would then be forced to assume that each species should appear in `observations.csv` four times. What makes Canis lupus different? Are there any other species that appear more in more than four records?

2. How are the observations distributed?

The answer to question 1. is illustrated in Figure 1. Figure 1 shows the number of unique species present in `observations.csv` and how many times they appear. It is clear that the majority of the dataset consists of species that only have one record in each park. However, there are 265 species with two records in each park and four species with three records in each park. Why that is the case is not immediately apparent.

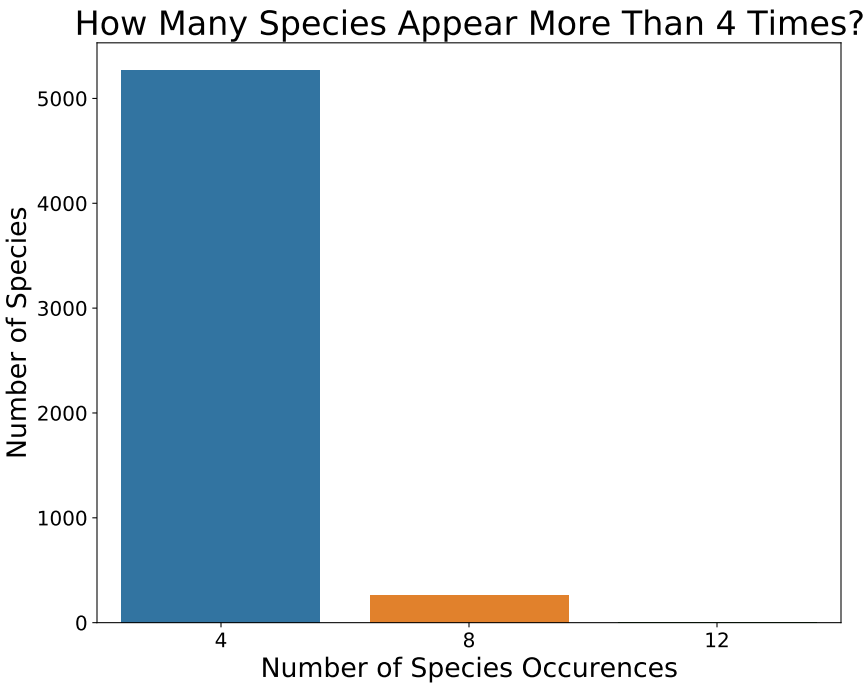

Figure 1: The number of times each unique species appears in observational data.

The second question is answered in Figure 2. When investigating the total population distribution of observations, the bi-modality is obvious. When the data is further divided by national park,

the bi-modality is removed. The dataset consists of observational data from four distinct regions, each with their own approximately normal observations distributions (see Figure 3). Figure 3 also shows the spread of the observations data is quite consistent between each park - each park having a standard deviation of 20 observations and an interquartile range (IQR) of 28 observations. This suggests that the difference between the parks is represented by some constant shift in the number of times a species is observed. That constant is not immediately obvious.

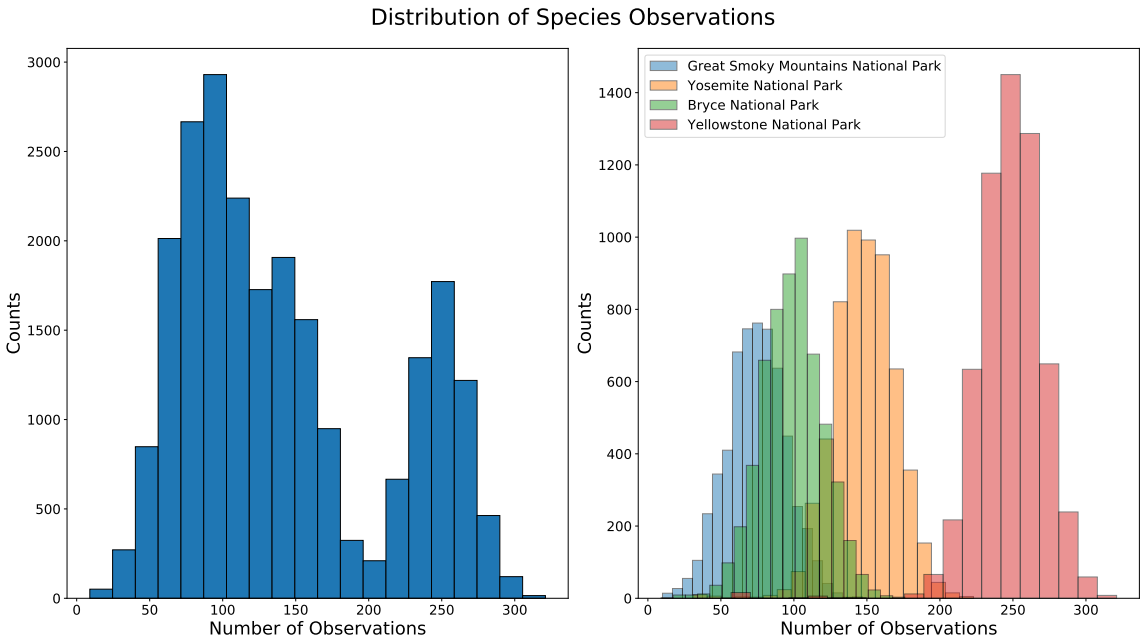

Figure 2: **(left)** Distribution of observations in observational data. **(right)** The same data as in **(left)**, stratified by national park.

3.2 Species Information Data

The information related to each individual species is contained within the `species_info.csv` dataset. Upon initial inspection, it was apparent that many of the species were duplicated. It is important to note that for many of the duplicated records, the set of common names varied. The overwhelming majority of the duplicated scientific names had different sets of common names but the same conservation status. These duplicates were removed.

Two species, Canis lupus (the gray wolf) and Oncorhynchus mykiss (the rainbow trout) appeared twice with duplicated scientific names but different conservation statuses. Canis lupus appeared with the statuses "Endangered" and "In Recovery" suggesting that conservation efforts were successful in bringing this species back. Oncorhynchus mykiss, however, appears with conservation statuses "Not at Risk" and "Threatened". This would suggest that the species population is dwindling, possibly requiring intervention to prevent extinction. Due to the change in conservation status, these two species have been omitted from analysis of the species information data.

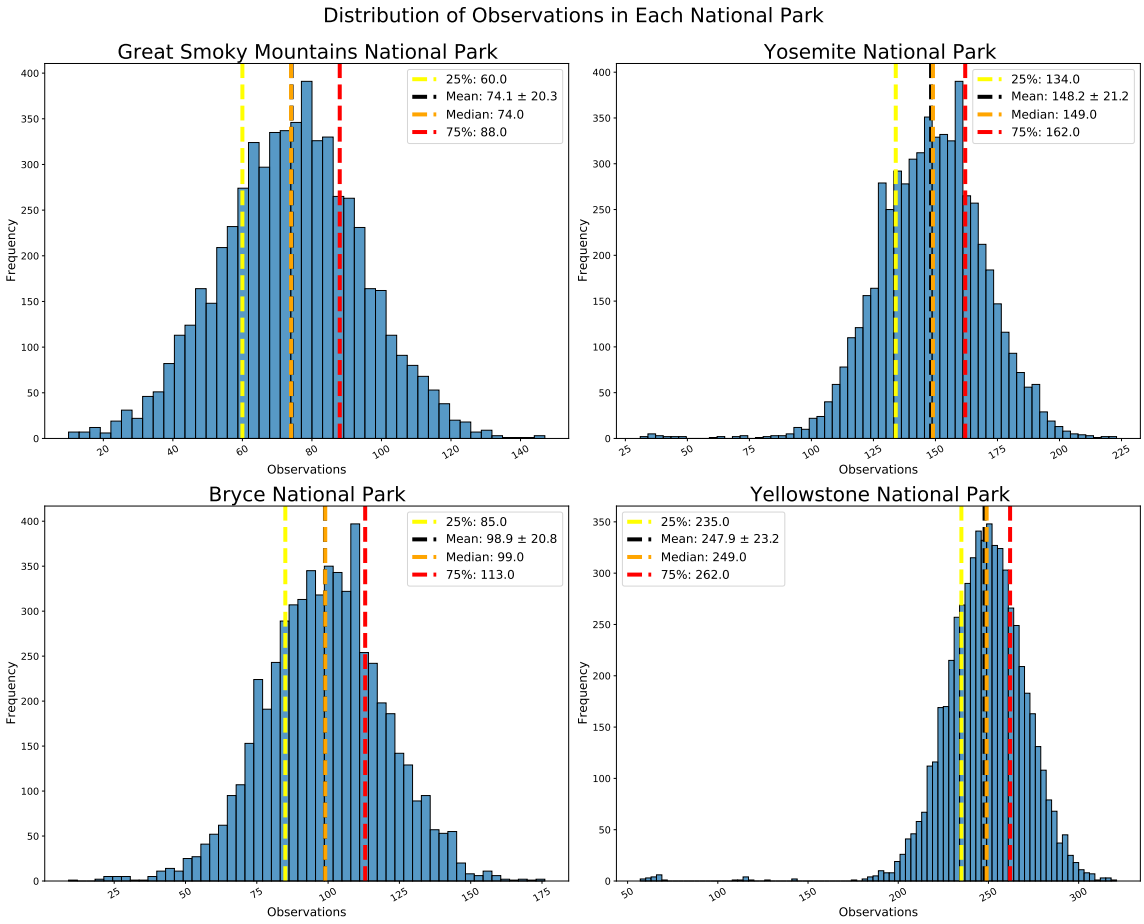

Figure 3: **(upper left)** Distribution of observations in Great Smoky Mountains National Park. **(upper right)** Distribution of observations in Yosemite National Park. **(lower left)** Distribution of observations in Bryce National Park. **(lower right)** Distribution of observations in Yellowstone National Park.

The frequency of each conservation status for each species is presented in Figure 4. The overwhelming majority of species appear are not at risk. Birds and mammals are the only two species present in the represented national parks that have recovering species. The reptile and non-vascular plant categories contain no species with threatened or endangered status. This suggests that these two categories would require less resources than say mammals or fish which display a greater penchant to endangerment. This is likely due to humans propensity to hunt both types for not only food but for sport. Greater conservation efforts for these species may be required.

3.3 Combining Datasets

The two datasets were combined via left join on the `scientific_name` column. With the records for Canis lupus and Oncorhynchus mykiss having been removed from the species information, the category, conservation status and common name information were left empty after the combination. These variables were filled given context from the original species information data. For Canis

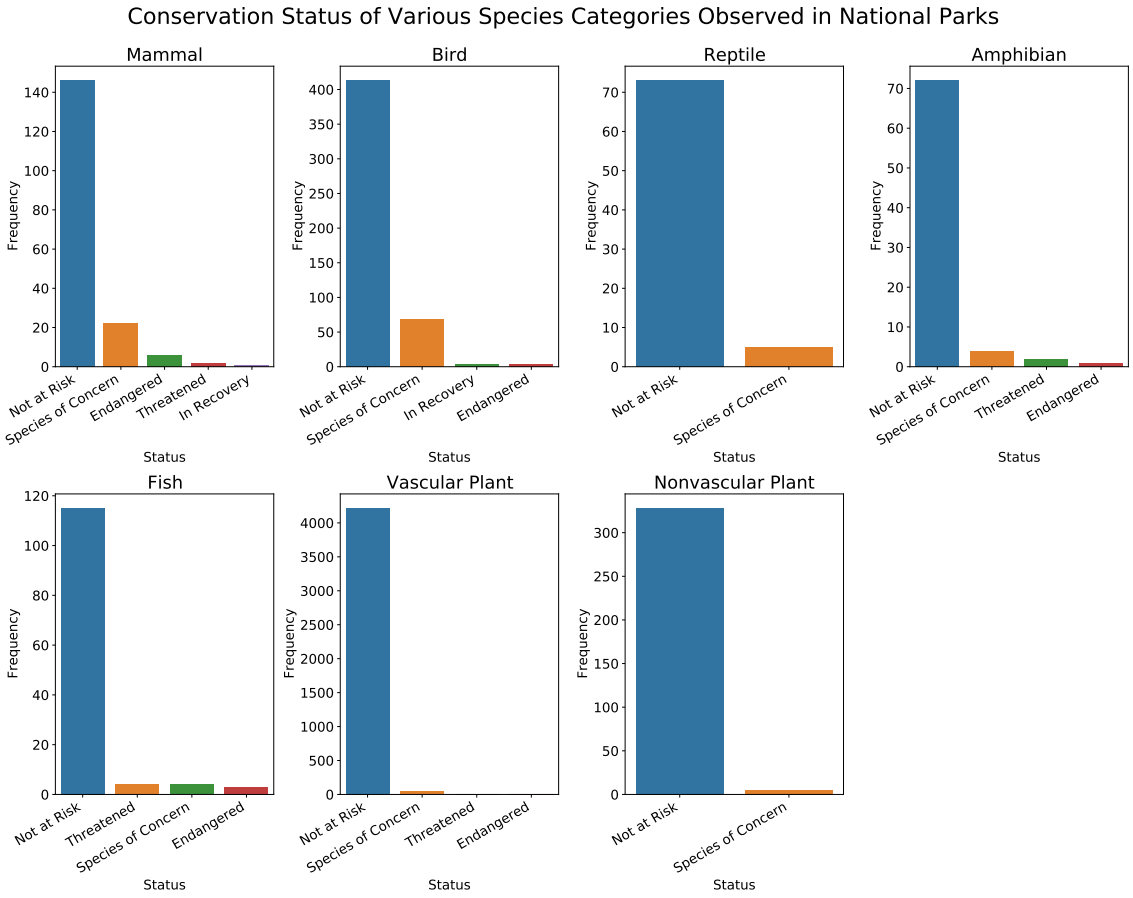

Figure 4: Frequency of species conservation status each species category.

lupus, the greatest number of observations within each park was assumed to related to the "In Recovery" conservation status while the two fewest number of observations were given a status of "Endangered". A similar process was employed for Oncorhynchus mykiss using "Not at Risk" for the greatest number of observations and "Threatened" for the fewest. With these records now cleaned, they were included in the subsequent analyses.

Box plot analysis was employed to elucidate any tendency in the observational data with species category or conservation status. With each park displaying a unique observation distribution, the box plots for each park were visualized on individual axes for ease of reading. Figure 5 depicts the distribution of observations of each species category in each park. It is not so surprising that within each park, the observations distribution is quite consistent for each category, given that the majority of species in each category is not at risk. It is interesting to not that, with the increased number of observations in Yellowstone, several groupings of outliers are observed. The number of groupings appear to be consistent with the number conservation statuses for each category presented in species information data analysis. These groupings are not as obvious for the other parks.

When performing the same analysis using conservation status instead of species category, a

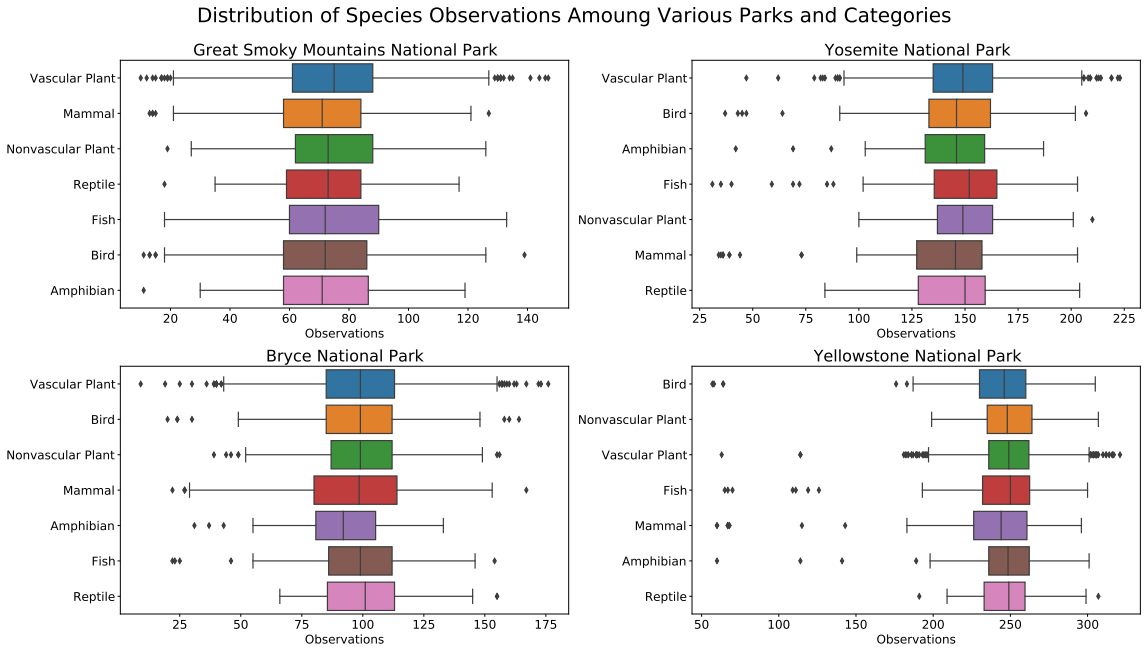

Figure 5: Box plot analysis of observations for each species category present in each of the four national parks

trend in the observational data becomes immediately apparent. As one would guess, those species considered to be not at risk tend to be observed with greater frequency in every park. The order of the remaining statuses, from greatest frequency to lowest frequency of observations is as follows: species of concern, in recovery, threatened, and endangered. This, again is not so surprising. These visualizations suggest that the conservation status given to a species is highly dependant on the frequency in which it was observed. This is most evident when considering the data from Yelllowstone National Park.

Also of note, it appears that a species is considered to be "in recovery" when the frequency of observations increases to a value close to or within the IQR for species of concern. This would explain why Canis lupus appeared to change conservation status. Clearly, the conservation efforts used in the last seven days to prevent the extinction of this species were very effective. With that in mind, I would suggesting using the same program - adapted to the species needs - for the remaining endangered and then threatened species in these parks.

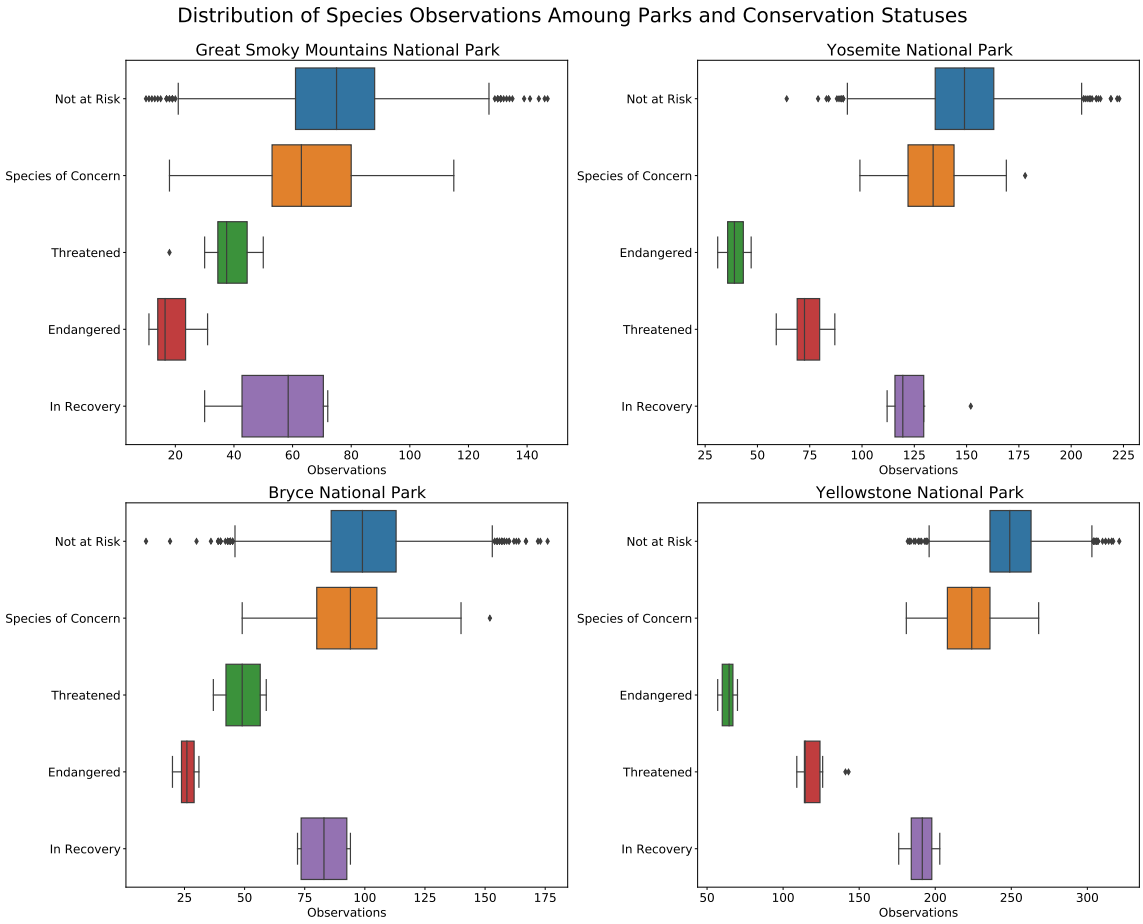

Figure 6: Box plot analysis of observations for each conservation status present in each of the four national parks

4 Conclusions

In this project, data relating to categorical and observation frequency of species present in four national parks were cleaned and analyzed. The vast majority of species present are considered to be not at risk. Only two of the seven species categories present without any threatened or endangered species, that being reptiles and non-vascular plants. This is likely due to the fact that mammals and birds are much more likely to be hunted, while fish and amphibians have to content with rising water temperatures destroying their habitats. Fish will also have to content with fishing.

Of the species that are endangered, only one had transitioned to in recovery status in the last seven day, that being Canis lupus. Whatever conservation efforts had been employed have clearly worked. I suggest adapting this program for other species categories and using it to prevent the extinction of others.

References

[1] National Parks Service. *About Us*. URL: https://www.nps.gov/aboutus/index.htm.

[2] Rocío Lower and Rebecca Watson. *How Many National Parks are There*. URL: https://www.nationalparks.org/connect/blog/how-many-national-parks-are-there.

[3] R.E.A. Almond et al. *Living Plant Report 2022 - Bulding a Nature Positive Society*. WWF, Gland, Switzerland: World Wildlife Foundation, Summer 2022. ISBN: 978-2-88085-316-7.