Biodiversity in National Parks

Analyst: Mahya Pourseif

Introduction

In this project we will interpret data from the national park service about endangered species in different parks.

During this project we will analyze, clean up and plot some data as well as pose some questions and seek to find an answer for them. Also investigation will be needed to see if there are any pattern or theme on the types of species that become endangered.

Data Sources:

Both `species_info.csv` and `observations.csv` are provided by https://www.codecademy.com

Project Goals

In this project the perspective will be through a biodiversity analyst for the National Parks Service. The National Park Service wants to ensure the survival of at-risk species, to maintain the level of biodiversity within their parks. Therefore, the main objectives as an analyst will be understanding characteristics about the species and their conservations status, and those species and their relationship to the national parks. Some questions that are posed:

- What is the distribution of conservation status for species?
- Are certain types of species more likely to be endangered?
- Are the differences between species and their conservation status significant?
- Which animal is most prevalent and what is their distribution amongst parks?

Species

The species_info.csv contains information on the different species in the National Parks. The columns in the data set include:

- **category** - The category of taxonomy for each species
- **scientific_name** - The scientific name of each species
- **common_names** - The common names of each species
- **conservation_status** - The species conservation status

Using the column **scientific_name** number of distinct species in the data is **5,541** unique species.

Number of category that are represented in the data are **7** which contains:

Amphibian: 80 species **Bird: 521** species **Fish: 127** species **Mammal: 214** species
Nonvascular Plant: **333** species **Reptile: 79** species **Vascular Plant: 4470** species

The column conservation_status has **5** categories that contains as below:

Species of Concern: 161 species, **Endangered: 16** species, **Threatened:10** species,
In Recovery: 4 species, **nan values:5633** species

Observations

The Observations.csv contains information from recorded sightings of different species throughout the national parks in the past 7 days. The columns included are:

- **scientific_name** - The scientific name of each species
- **park_name** - The name of the national park
- **observations** - The number of observations in the past 7 days

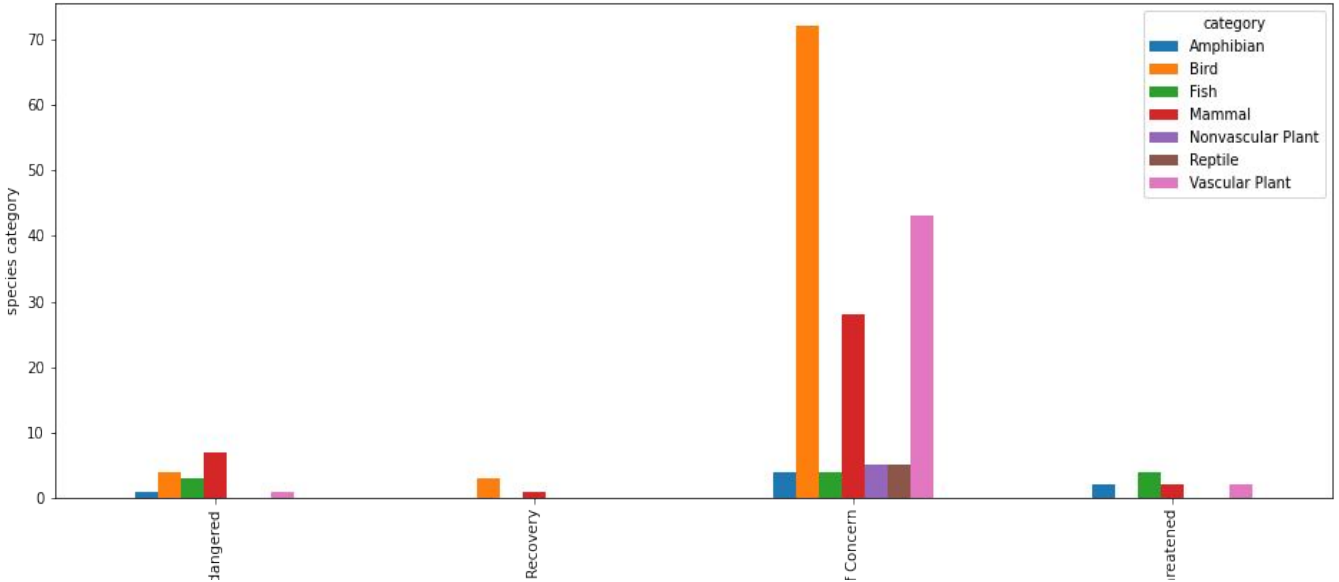
The number of parks that are in the dataset are only **4** national parks that contains: **Great Smoky Mountains National Park, Yosemite National Park, Bryce National Park, Yellowstone National Park**

Here are the total number of observations logged in the parks, there are **3,314,739** sightings in the last 7 days... that's a lot of observations!

.

Analysis

There is a chart below to explore the different categories that are nested in the conservation_status column except for the ones that do not require an intervention.
From the analysis done on the data, for those in the Endangered status, 7 were mammals and 4 were birds. In the In Recovery status, there were 3 birds and 1 mammal, which could possibly mean that the birds are bouncing back more than the mammals. **72** species of **birds**, **43** of **vascular plant** and **28** species of **mammal** are in the species of concern category which are significantly higher than other species.



In conservation

category	not_protected	protected	Percent protected
Amphibian	72	7	8.86
Bird	413	75	15.36
Fish	115	11	8.73
Mammal	146	30	17.04
Nonvascular plant	328	5	1.50
Reptile	73	5	6.41
Vascular plant	4216	46	1.07

The next question is if certain types of species are more likely to be endangered? This can be answered by creating a new column called `is_protected` and include any species that had a value other than `No Intervention`.

It's easy to see that Birds, Vascular Plants, and Mammals have a higher absolute number of species protected.

Absolute numbers are not always the most useful statistic, therefore it's important to calculate the rate of protection that each `category` exhibits in the data. From this analysis, one can see that ~17 percent of mammals were under protection, as well as ~15 percent of birds.

Statistical significant

This section will run some chi-squared tests to see if different species have statistically significant differences in conservation status rates.
The first test will be called `contingency1` and will need to be filled with the correct numbers for mammals and birds.

chi2_contingency(contingency1)
Output: (0.1617014831654557, 0.6875948096661336, 1, array([[24.2519685, 151.7480315],
[10.7480315, 67.2519685]]))

The results from the chi-squared test returns many values, the second value which is 0.69 is the p-value. The standard p-value to test statistical significance is 0.05. For the value retrieved from this test, the value of 0.69 is much larger than 0.05. In the case of mammals and birds there doesn't seem to be any significant relationship between them i.e. the variables independent.

The next pair, is going to test the difference between `Reptile` and `Mammal`.

chi2_contingency(contingency2)
Output: (4.289183096203645, 0.03835559022969898, 1, array([[24.2519685, 151.7480315],
[10.7480315, 67.2519685]]))

This time the p-value is 0.039 which is below the standard threshold of 0.05 which can be take that the difference between reptile and mammal is statistically significant. Mammals are shown to have a statistically significant higher rate of needed protection compared with Reptiles.

Species in Parks

The next set of analysis will come from data from the conservationists as they have been recording sightings of different species at several national parks for the past 7 days.

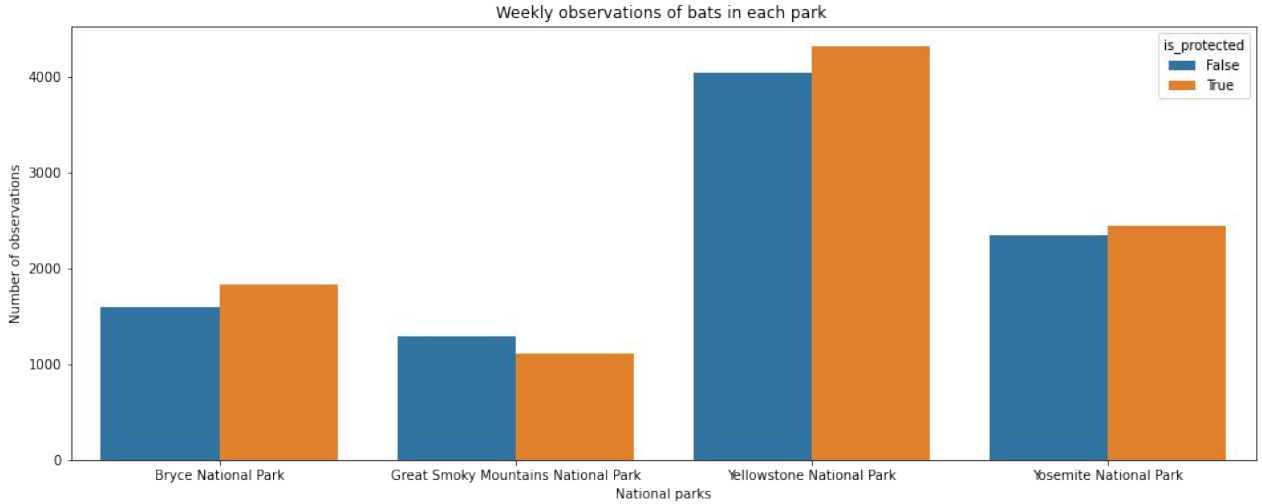
From the analysis of common names from species it seems that **Bats** are the most prevalent mammals in the dataset

Now Let's see how many total bat observations(across all species) were made at each national park.

Park name	Observations
Bryce National Park	*3433*
Great Smoky Mountains National Park	*2411*
Yellowstone National Park	*8362*
Yosemite National Park	*4786*

Now let's see each park broken down by protected bats vs. non-protected bat sightings. It seems that every park except for the Great Smoky Mountains National Park has more sightings of protected bats than not. This could be considered a great sign for bats.

Below is a plot from the output of the last data manipulation. From this chart one can see that Yellowstone and Bryce National Parks seem to be doing a great job with their bat populations since there are more sightings of protected bats compared to non-protected species. The Great Smoky Mountains National Park might need to beef up there efforts in conservation as they have seen more non-protected species.



Conclusions

The project was able to make several data visualizations and inferences about the various species in four of the National Parks that comprised this data set.

This project was also able to answer some of the questions first posed in the beginning:

- What is the distribution of conservation status for species?
  - The vast majority of species were not part of conservation.(5,633 vs 191)
- Are certain types of species more likely to be endangered?
  - Mammals and Birds had the highest percentage of being in protection.
- Are the differences between species and their conservation status significant?
  - While mammals and Birds did not have significant difference in conservation percentage, mammals and reptiles exhibited a statistically significant difference.
- Which animal is most prevalent and what is their distribution amongst parks?
  - the study found that bats occurred the most number of times and they were most likely to be found in Yellowstone National Park.

Further research

This dataset only included observations from the last 7 days which prohibits analyze changes over time. It would be curious to see how the conservation status for various species changes over time. Another piece that is missing is the Area of each park, it can be assumed that Yellowstone National Park might be much larger than the other parks which would mean that it would exhibit more observations and greater biodiversity. Lastly, if precise locations were recorded, the spatial distribution of the species could also be observed and test if these observations are spatially clustered.