Analyze Biodiversity Data From The National Parks Service

Final Presentation

2022

SUMMARY

The purpose of this project is to analyze biodiversity data from the National Park Service, especially in relation to the different species observed in different national park locations. This project will capture, analyse, prepare, display data and attempt to explain the results of the analysis. We will analyze the data and answer questions about which animals and plants in major US national parks are the most protected, and which ones need protection.

We will use data visualization techniques will be employed to understand the data better. Statistical inference will also be used to test if the observed values are statistically significant. Some of the key metrics that will be computed include: Distributions, counts, relationship between species, conservation status of species, observations of species in parks.

Introduction

This goal of this project is to analyze biodiversity data from the National Parks Service, particularly around various species observed in different national park locations.

This project will scope, analyze, prepare, plot data, and seek to explain the findings from the analysis.

Here are a few questions that this project has sought to answer:

- 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?

Data sources

https://www.nps.gov/rlc/researchdatabases.htm, https://www.nps.gov/rlc/researchdatabases.htm

The first file csv in the project 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

The second file .csv contains information from recorded sightings of different species throughout the national parks.. 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

Methodology

Descriptive statistics and data visualization techniques will be employed to understand the data better.

Statistical inference will also be used to test if the observed values are statistically significant;

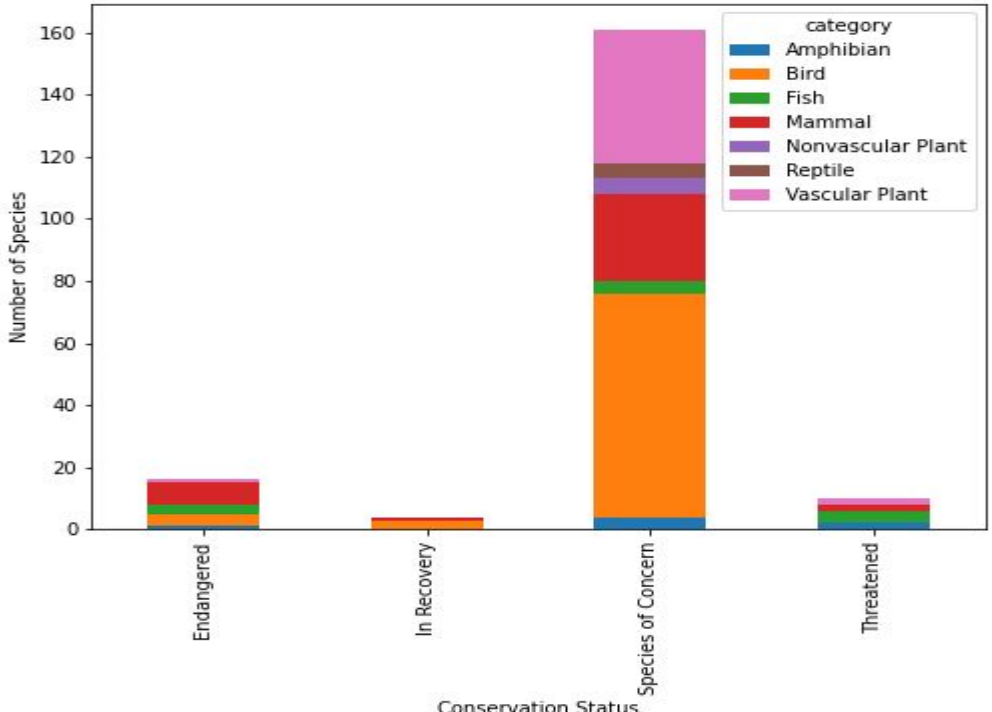
Some chi-squared tests to see if different species have statistically significant differences in conservation status rates.

Analysis

The next step is to check the various categories nested in the conservation_status column, with the exception of those that do not require intervention. Below is a table and diagram for you to study. For those in Endangered status, 7 mammals and 4 birds. There were 3 birds and 1 mammal in In Recovery status, which could mean birds are bouncing back more than mammals.

category	Amphibian	Bird	Fish	Mammal	Nonvascular Plant	Reptile	Vascular Plant
conservation_status							
Endangered	1.0	4.0	3.0	7.0	NaN	NaN	1.0
In Recovery	NaN	3.0	NaN	1.0	NaN	NaN	NaN
Species of Concern	4.0	72.0	4.0	28.0	5.0	5.0	43.0
Threatened	2.0	NaN	4.0	2.0	NaN	NaN	2.0

The next step is to check the various categories nested in the conservation_status column, with the exception of those that do not require intervention. Below is a table and diagram for you to study. For those in Endangered status, 7 mammals and 4 birds. There were 3 birds and 1 mammal in In Recovery status, which could mean birds are bouncing back more than mammals.



From this analysis, one can see that ~17 percent of mammals were under protection, as well as ~15 percent of birds.

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	8.860759
1	Bird	413	75	15.368852
2	Fish	115	11	8.730159
3	Mammal	146	30	17.045455
4	Nonvascular Plant	328	5	1.501502
5	Reptile	73	5	6.410256
6	Vascular Plant	4216	46	1.079305

Statistical Significance

The result from the chi-square test

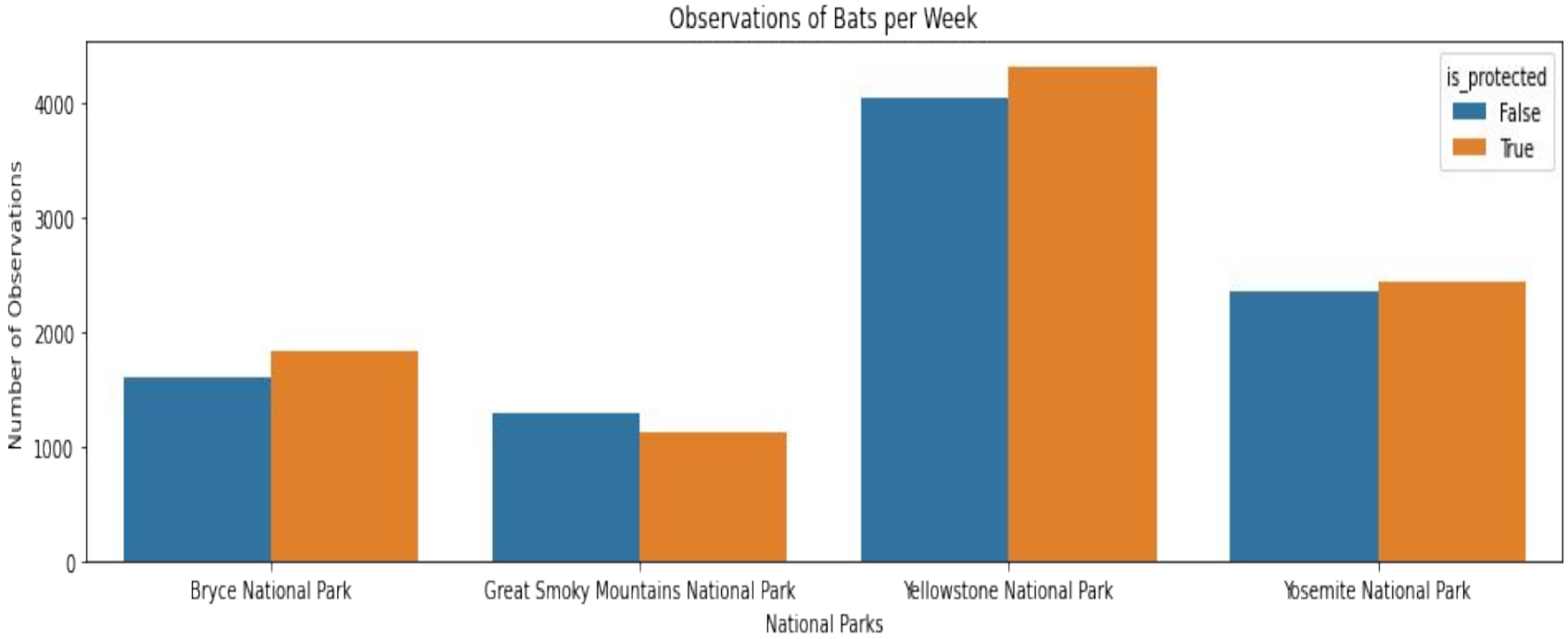
The p-value is 0.039 which is below the standard threshold of 0.05 which can be take that the difference between reptile and mammal is statistically significant. Mammals are shown to have a statistically significant higher rate of needed protection compared with Reptiles.

```
(4.289183096203645, 0.03835559022969898, 1, array([[ 24.2519685, 151.7480315],
       [ 10.7480315,  67.2519685]]))
```

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

	park_name	is_protected	observations
0	Bryce National Park	False	1596
1	Bryce National Park	True	1837
2	Great Smoky Mountains National Park	False	1299
3	Great Smoky Mountains National Park	True	1112
4	Yellowstone National Park	False	4044
5	Yellowstone National Park	True	4318
6	Yosemite National Park	False	2345
7	Yosemite National Park	True	2441

From this chart, you can see that Yellowstone and Bryce National Parks are doing very well with their bat populations, as protected bats are more common compared to unprotected 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.