Biodiversity Capstone Project

Investigating Protected Species

By: David R. Diaz

Mission

- As a biodiversity analyst working for the National Parks Service I am to analyze raw data about different species in the National Park system. The service requested a data analysis on the conservation status of different species across the National Parks. The following presentation displays my results.

Raw Data

- The National Parks service supplied us with a CSV file titled: species_info.csv. Once loaded into a Data Frame in Pandas we were able to see the data in this file. The file consisted of 5,823 rows of data each of which contained the following columns:

1. Category – describes the type of species (ex: mammal, reptile, bird, etc.)
2. Scientific Name- gives the scientific name of each species

Raw Data

3. Common Names – lists all the common names that the species is known by

4. Conservation Status – lists whether the species is endangered, in recovery, threatened, a species of concern, or in none of these categories.

It is from this raw data that we are able to analyze what the National Park Service would like to know.

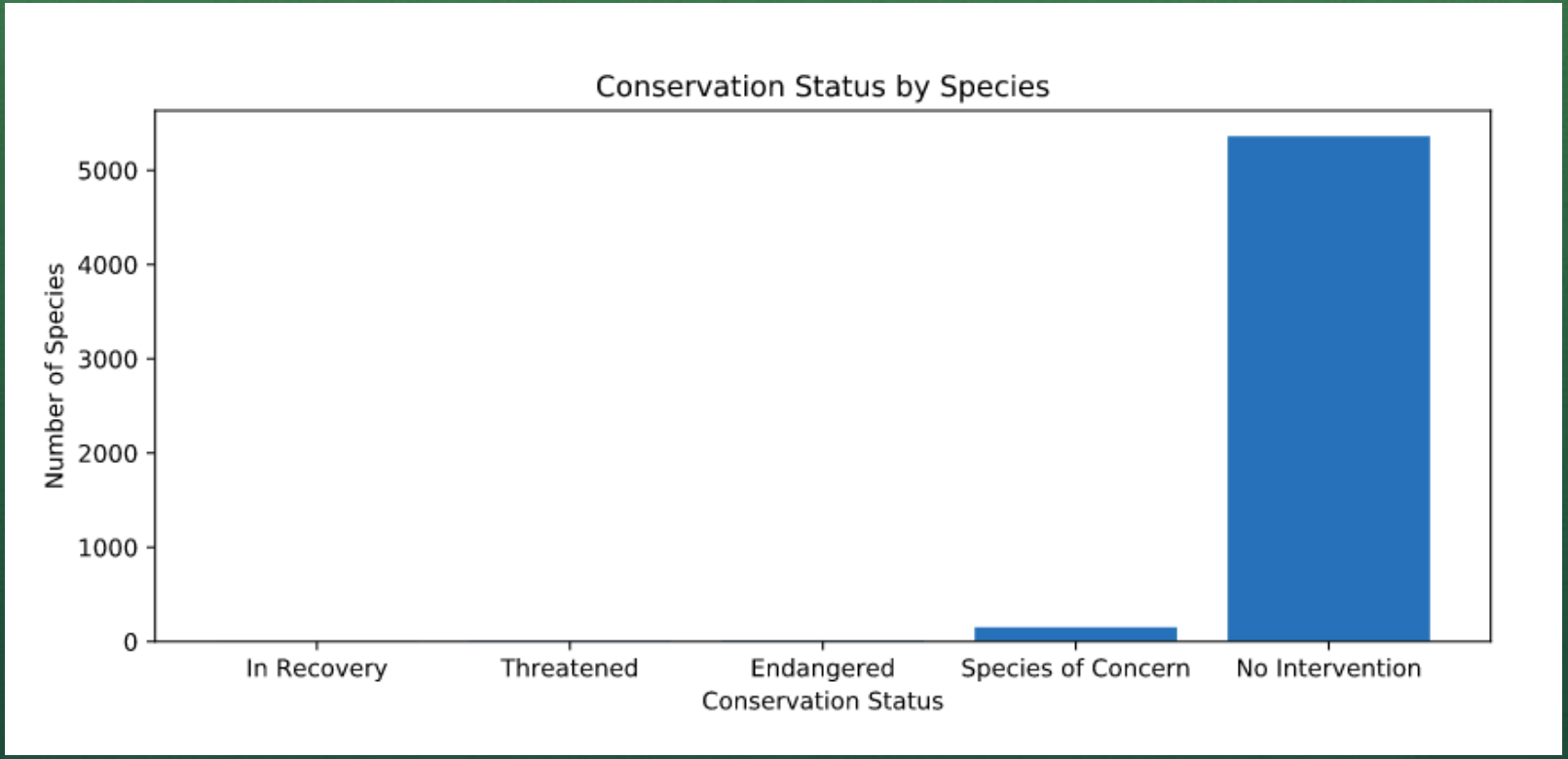
Purpose

- The National Park Service asked for an analysis of the different species based on their conservation status. More specifically, they wanted to know **if certain types of species are more likely to be endangered than other types**. This question can not be answered with the file in its basic list form, therefore we have to manipulate the data.

Data Manipulation

- First order of business is to reorganize the data so that we have a count of how many species fall into each category of conservation status. The following slide illustrates this chart.

Conservation Status by Species



Data Manipulation

- The previous chart clearly shows that the majority of the species in the National Park System are in the category of "No Intervention."
- The chart does not clearly show all the data, however, and does not come close to allowing us to determine if certain types of species are more likely than others to become extinct.
- The next chart more clearly shows the data in the Bar Chart

Conservation Status Count

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10

Data Manipulation

- From here the next step was to break down these numbers even further to see what percentage of each type of species is protected by the National Wildlife system. The following slide illustrates the category of species, whether they are protected or not, and the percentage of each that is protected.

Percent of Protected Species

	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Data Manipulation

- Looking at the percentages it seems that mammals are the most likely to be protected and therefore endangered. But is the data significant? In order to see if mammals are more likely to be endangered, we must perform a statistical analysis.
- There are different statistical tests that can be used with this data, but seeing as it is categorical data with more than 2 pieces of data to analyze, we will perform a Chi ^{2} test.

Chi ^{2} Test

- We test to see if the percentage difference between endangered mammals and endangered birds is significant.
- Our null hypothesis is that the difference is due to chance.
- We then test to see if the percentage difference between endangered mammals and reptiles is significant.
- For significance we need a p-value of less than .05.

Significance Test

- The p-value for the test between mammals and birds was 0.688, which is not significant and therefore is a result of chance.
- However, the p-value for the test between mammals and reptiles was 0.038 which **is** significant.
- A third test comparing the percentages for all types of species resulted in a p-value of 3.4633 e -100.

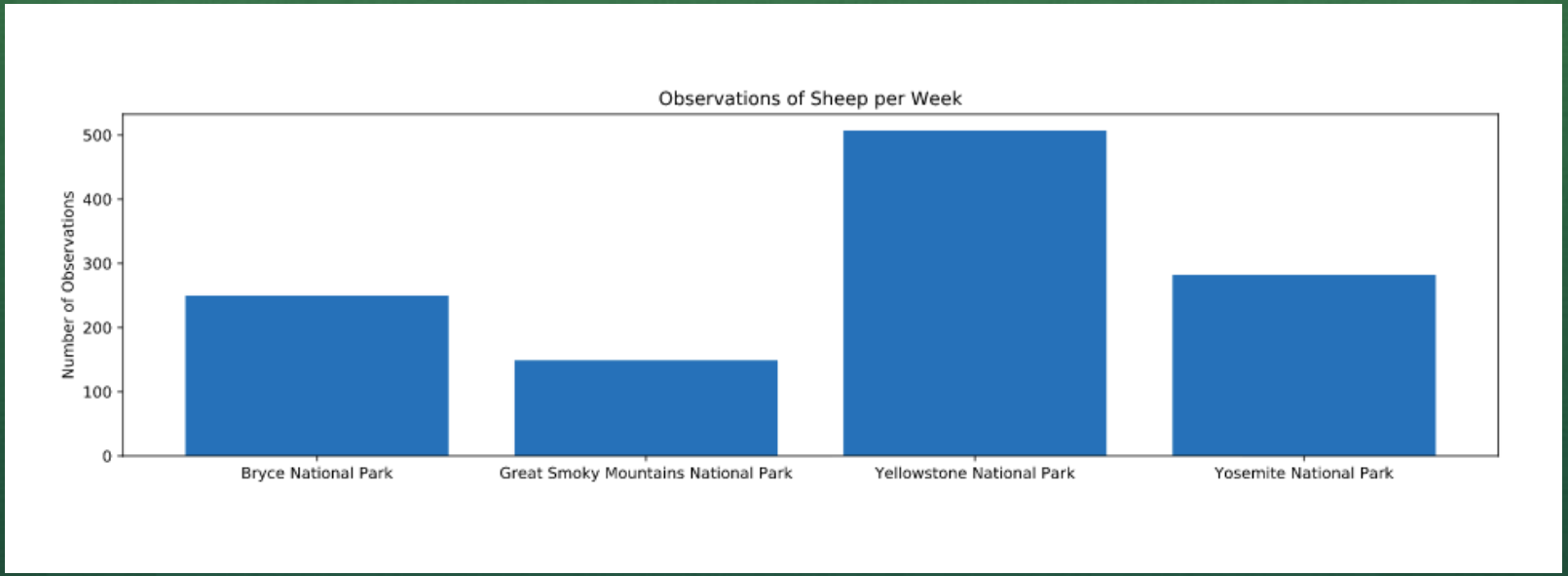
Conclusion

- Based on the data from the significance tests, we can conclude that certain types of species are more likely to be endangered than others.
- Conservationists can use this data to categorize the types of species that are more likely to be endangered and focus their efforts on the species with the highest likelihood of being endangered.

Sheep Raw Data

- The National Parks Service sent over more data to be analyzed. Conservationists have been recording sightings of different species at several national parks over the last week. The data has been sent in a CSV file called observations.csv.
- Specifically they want an analysis of the number if different sheep species observed over the last 7 days and how many were observed in each specific National Park.
- The following charts illustrate the observational data

Weekly Sheep Observation



Sheep Observation per National Park

	park_name	observations
0	Bryce National Park	250
1	Great Smoky Mountains National Park	149
2	Yellowstone National Park	507
3	Yosemite National Park	282

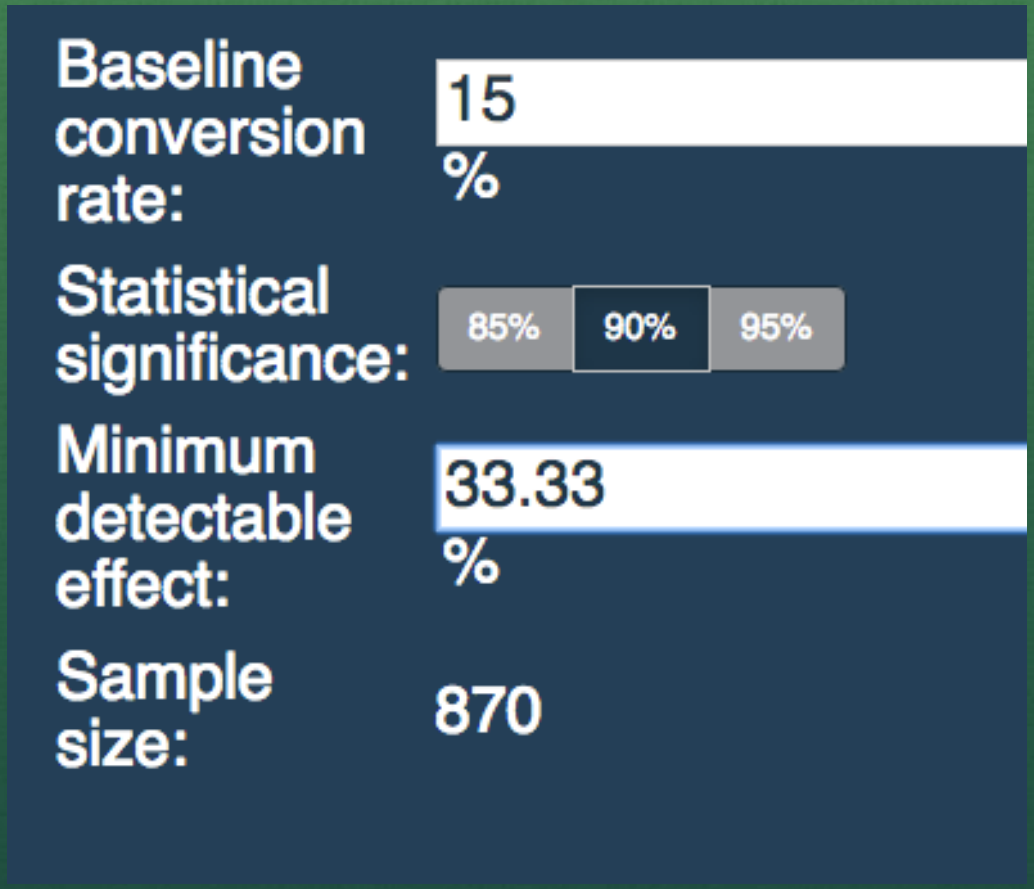
Sheep Disease Analysis

- Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease among the sheep there. The scientists want to test if the program is working and want to be able to detect the reductions of at least 5% points. They would like to know this with confidence.
- The only data they have is that last year, 15% of sheep had the disease at Bryce National Park.
- Based on that percentage and the observations from the previous slide we must calculate the population size of observations that the scientists need.

Sample Size Calculations

- The baseline conversion rate the scientists need is the 15% observed at Bryce last year
- To reduce the amount observed by 5% the minimum detectable effect = 100 x 5 / 15 (*baseline rate*) = 33.33%
- At a 90 % level of significance, we plug the figures into a sample size calculator and get a sample size of 870. The calculator is shown in the following slide.

Sample Size Calculator



Observation Time Needed

- To approximate the observation time necessary for scientists at each National Park to conduct their studies we must take the 870 population size needed and divide by the number of weekly observations recorded for each Park:
- Bryce NP: 870/250 = 3.48 weeks
- Great Smoky NP: 870/149 = 5.84 weeks
- Yellowstone NP: 870/507 = 1.72 weeks
- Yosemite NP: 870/282 = 3.10 weeks

Final Analysis

- Based on two Data Frames with species information and species sightings given to us by the National Parks System, we were able to manipulate the raw data and run statistical analyses to assist both conservationists and scientists better understand the raw data. With this manipulated data, they can go and perform their studies with confidence in the numbers they are using.