Biodiversity in the National Parks: Data Analysis Capstone Project

Overview of species_info.csv Data File

The species_info.csv data contains four columns of data related to animal species living in national parks:

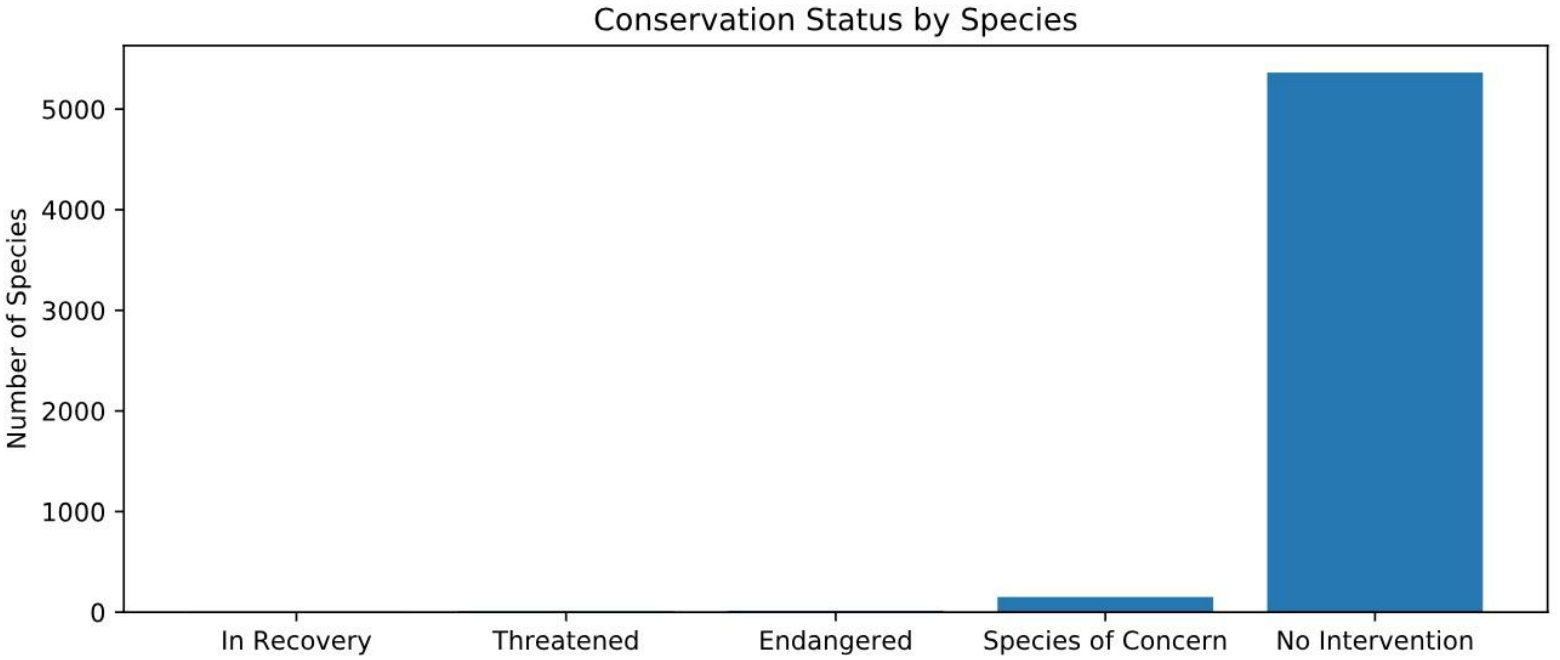
- Category: there are seven types of species (mammal, bird, reptile, amphibian, fish, vascular plant, and nonvascular plant)
- Scientific Name: there are a total of 5541 unique species recorded
- Common Name
- Conservation Status: there are four different conservation statuses (species of concern, endangered, threatened, and in recovery)

Table of the Number of Unique Species which fall into each Conservation Status and a table that also includes the number of Unique Species that did not have a Conservation Status listed (“No Intervention”):

	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	Species of Concern	151
3	Threatened	10

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

According to the data, less than 4% of species in national parks are in need of protection. A bar chart can better visualize this finding:



Research Question:

Are certain types of species more likely to be endangered?

	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

A cursory look at the data shows that some categories of species do appear more likely than others to be endangered (“protected”). However, further analysis is needed to verify if these differences are *statistically* significant or due to chance variations in the data sample.

Hypothesis Testing

Because we want to compare more than one categorical data set, a chi-squared test can be used to test our null hypothesis that there is no significant difference in protection rates between any two categories of species.

Testing shows no significant difference between the protection rates of mammals and birds ($p = .69$). However, there is a statistically significant difference between the rates of mammals and reptiles ($p = .04$). Certain types of species are more likely to be endangered than others.

Recommendation: Based on the data and the chi-squared test, the “protection rate” for mammals is significantly higher than that of reptiles. If the goal of the national parks is to identify high risk animals, mammals should be a category of focus. The national parks may also want to explore why mammals are more likely to be ‘protected’ compared to reptiles.

Tracking Sheep in the National Parks

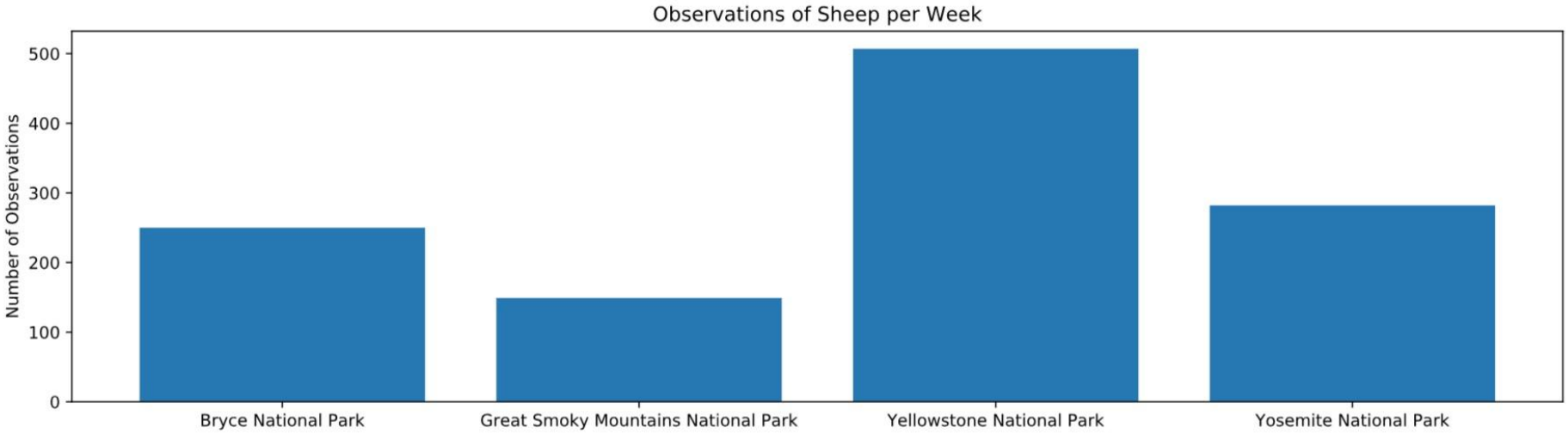
The second dataset from the National Parks Services includes number of observations, scientific names, and national parks. Before analyzing the new dataset, a new dataframe was created to identify which scientific names counted as sheep:

	category	scientific_name	common_names	conservation_status	is_protected	is_sheep
3	Mammal	Ovis aries	Domestic Sheep, Mouflon, Red Sheep, Sheep (Feral)	No Intervention	False	True
3014	Mammal	Ovis canadensis	Bighorn Sheep, Bighorn Sheep	Species of Concern	True	True
4446	Mammal	Ovis canadensis sierrae	Sierra Nevada Bighorn Sheep	Endangered	True	True

This dataframe was then merged with the original observations dataframe from the National Parks Services to show the number of sheep sightings in each 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

This table can also be visualized using a bar chart:



Sample Size for Detecting Foot and Mouth Disease Patterns

To test if the National Park Service's program to reduce foot and mouth disease among sheep is working, they need to know how many observations (the sample size) they have to collect to detect a significant change in the disease rate among the sheep population. Based on the sample size calculator, the rangers would need to collect data on 870 sheep to detect at least a 5% reduction in foot and mouth disease. In Bryce National Park, it would take rangers about 3.5 weeks to observe enough sheep, whereas rangers in Yellowstone would likely have a large enough sample in less than two weeks.