Biodiversity of National Parks

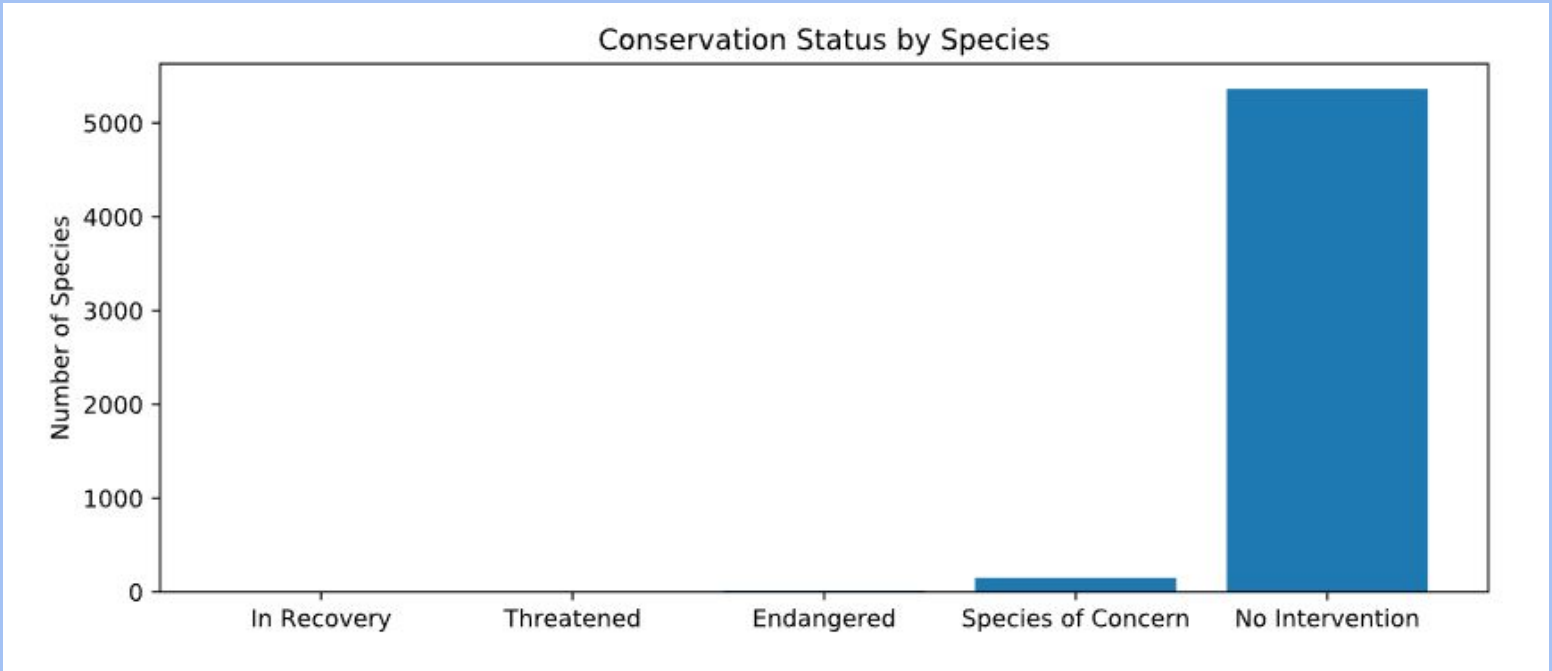
Codecademy Capstone Project
by: Scott Munn

The data provided is an overview of the species present in 4 national parks in the United States.

Listed are 5,541 different species, divided into 7 categories; mammal, bird, reptile, amphibian, fish, and vascular/non-vascular plants.

Of this list, we can see the number of species that are protected under conservation. These are divided as well by their status levels; 'No Intervention', 'In Recovery', 'Species of Concern', 'Threatened', or 'Endangered.'

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



One can infer from the bar graph illustrated above, that there is a massive gap between species in need of protection versus those without any current intervention.

Let us now focus our attention closer upon the first 4 groups and see what we can learn.

If we were to place all the data into a pivot table, we can better look at the overall status of each species and begin to ask our question, “are some species more likely to become endangered than others?”

	category	not_protected	protected	percent_protected
0	Amphibian	73	7	0.087500
1	Bird	442	79	0.151631
2	Fish	116	11	0.086614
3	Mammal	176	38	0.177570
4	Nonvascular Plant	328	5	0.015015
5	Reptile	74	5	0.063291
6	Vascular Plant	4424	46	0.010291

Next we will take all the information provided and create our first test, using a Chi-Square Contingency Test.

To start our test, we first compare **mammals** with **birds** to see if there is any significant difference by using a null hypothesis of 0.05%.

We determine that:

- With a p-value of ~68%, this exceeds our null hypothesis and is not a significant enough result to arrive to a conclusion.

However, we run a second Chi-Square test on **mammals** and **reptiles** instead.

We then find:

- The p-value is ~0.03%, which is below our null hypothesis. Therefore we can conclude there is a statistical difference between these two species.

Based on what we've discovered from our Chi-Square Contingency tests conducted between select species, it can be said with confidence that some are more likely to become endangered than others.

Knowing this, it is recommended that a close eye is kept on mammals and birds, as they are the most threatened out of the species found in our national parks.

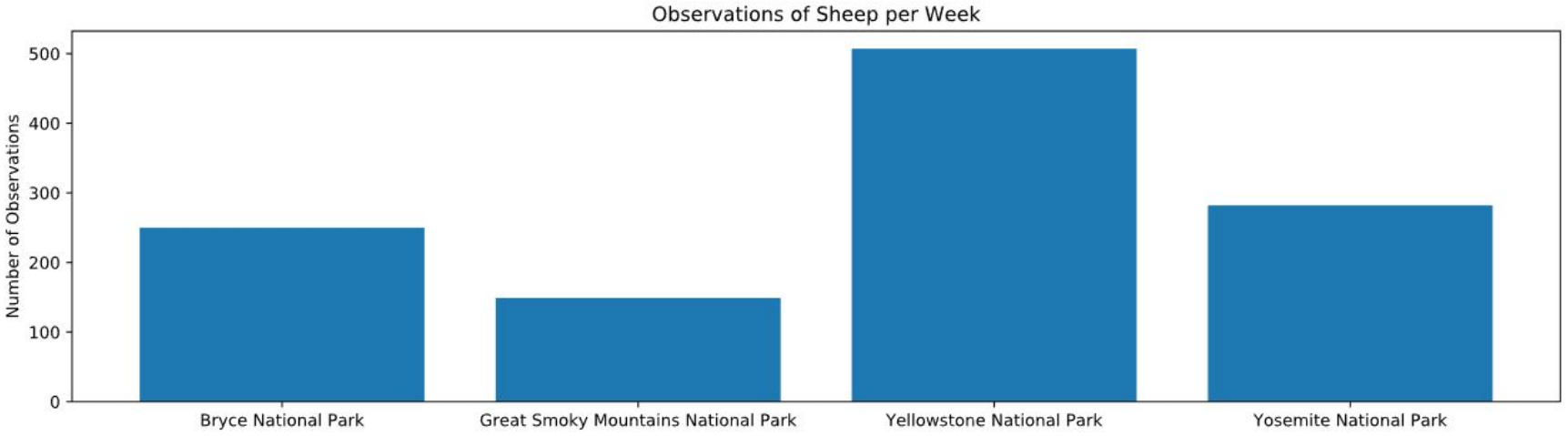
While it is shown that both vascular and nonvascular plants are the least likely to become endangered, special care must be taken to ensure they are not neglected in favor of other species found throughout national parks.

Study on Sheep Observations

Shifting our focus now, let's explore observational data on sheep sightings for national parks.

By merging the recorded sightings with species data, filtering for sheep, we are able to create a table representing what we know.

	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



Using the same information from the previous data frame, we then plot a bar graph showing each of the 4 parks reported sightings of sheep.

Moving on, we can use the number of observations we've gather to help conduct a study for the park rangers at Yellowstone National Park.

The park rangers at Yellowstone National Park want to reduce the rate of sheep afflicted with Foot and Mouth Disease.

Before they begun their program, they learn that as much as 15% of the sheep population in the previous year at Bryce National Park had this disease.

To see if the program has made any improvements, we will conduct an A/B test and check for a difference of at least 5% to ensure there has been a reduction.

In order to receive accurate results, we must first determine the sample size needed to run our test effectively.

Using a sample size calculator, we plug in what we've gathered to get our answer:

- Baseline = 15% (15% of sheep had Foot and Mouth Disease in Bryce National Park)
- Minimum Detectable Effect = 33.33% (100 multiplied by our desired reduction, then divided by the baseline)
- Statistical Difference = 90% (confidence level)
- **Sample Size = 870**

To go a step further, we would like to know just how long it would take for the rangers at Yellowstone National Park to complete their tests.

By taking the sample size and dividing it by the number of observations of sheep at Yellowstone National Park, we learn that it would take roughly **one week** to successfully conduct their test on the sheep population.