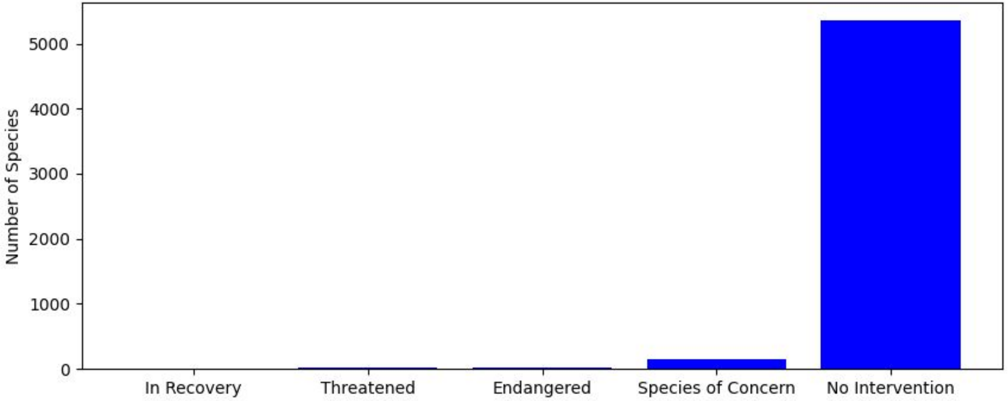
Biodiversity Project

Codecademy Capstone Project - Jackie Pham

The National Parks Service would like you to perform some data analysis on the conservation statuses of these species and to investigate if there are any patterns or themes to the types of species that become endangered. During this project, you will analyze, clean up, and plot data, pose questions and seek to answer them in a meaningful way.

Conservation status by species

conservation_status	quantity
Endangered	15
In Recovery	4
No Intervention	5363
Species of Concern	151
Threatened	10



Are certain types of species more likely to be endangered?

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

It looks like Mammals are more likely to be endangered than Birds, but is it a significant difference?

A significance test to see if this statement is true is necessary. In this test, our **null hypothesis** is that this difference is due to chance.

In this case a Chi-square test will be applied.

	not_protected	protected
Mammal	146	30
Bird	413	75

```
contingency = [[30, 146],[75, 413]]
print(chi2_contingency(contingency))
pval = 0.687
```

pval > 0.05 >> not significant which means we are not able to reject the null hypothesis, i.e this difference could be due to chance.

Let’s do another Chi square test for Mammal and Reptile

	not_protected	protected
Mammal	146	30
Reptile	73	5

```
contingency1 = [[30, 146],[5, 73]]

print(chi2_contingency(contingency1))

pval = 0.038
```

Since the pval of this test < 0.05 (significant) we would be able to reject the null hypothesis, i.e we **cannot** conclude that the difference between the percentages of protected birds and reptile is not significant and is a result of chance.

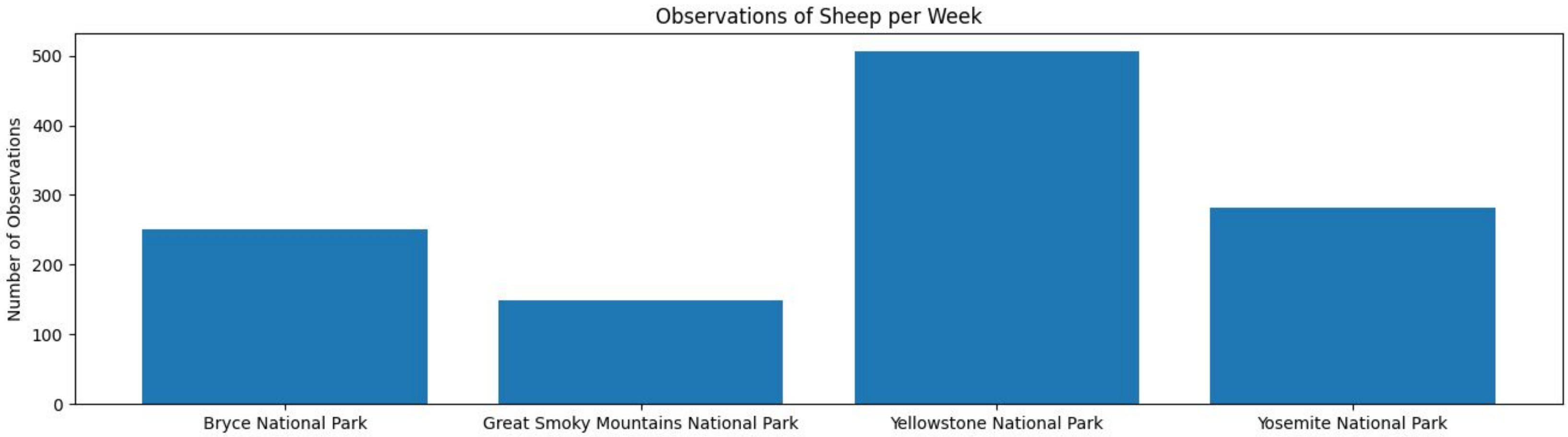
Therefore, we can conclude that certain types of species are more likely to be endangered than others.

Observations DataFrame

A team of ruminant-enthused scientists has been tracking the movements of various species of sheep across different national parks and have asked for your assistance in analyzing the **observation** and **species** DataFrames to help track sheep locations.

Q1. How many total sheep sightings (across all three species) were made at each national park?

park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282


Observations of Sheep per Week
Number of Observations
0
100
200
300
400
500
Bryce National Park
Great Smoky Mountains National Park
Yellowstone National Park
Yosemite National Park

Foot and Mouth Reduction Effort - Sample Size Determination

Park Rangers at Yellowstone National Park have been running a program to reduce the rate of foot and mouth disease at that park. The scientists want to test whether or not this program is working. They want to be able to detect reductions of at least 5 percentage points. For instance, if 10% of sheep in Yellowstone have foot and mouth disease, they'd like to be able to know this, with confidence.

The only information that the scientists currently have is that last year it was recorded that 15% of sheep at Bryce National Park have foot and mouth disease. Using this value and the sample size calculator in the browser window on the right, you will need to calculate the number of sheep that they would need to observe from each park to make sure their foot and mouth percentages are significant. Use the default level of significance (90%).

- Baseline percentage of this sample size determination: baseline = 15

- Calculate “Minimum Detectable Effect”:

    minimum_detectable_effect = 100*5./15

- sample_size_per_variant = 870

- How many weeks would the scientists need to spend at Yellowstone National Park to observe enough sheep?

yellowstone_weeks_observing = sample_size_per_variant/507 = 870/507 = 1.7 weeks

- The scientists also want to repeat their measurements at Bryce National Park. How many weeks will they have to spend there to observe enough sheep?

bryce_weeks_observing = sample_size_per_variant/250 = 870/250 = 3.48 weeks