Introduction to Data Analysis Capstone

-Codecademy

Sarah Schubert

Species_info.csv

Species_info.csv has 4 columns - Category, Scientific Name, Common Names and Conservation Status.

There were 5541 different species and 7 different categories of species in Species_info.

Most species (5363 of the 5541) don’t require any intervention but the rest require some protection, as shown below.

```
conservation_status
Endangered            15
In Recovery            4
No_Intervention     5363
Species of Concern   151
Threatened            10
```

Significant Calculations

From the species data frame, the percent of each category of species that is being protected was calculated. The results are as follows.

category	False	True	percent_protected
Amphibian	73	7	0.087500
Bird	442	79	0.151631
Fish	116	11	0.086614
Mammal	176	38	0.177570
Nonvascular Plant	328	5	0.015015
Reptile	74	5	0.063291
Vascular Plant	4424	46	0.010291

It was observed that both the mammal and bird categories as well as, to a lesser extent, the mammal and reptile categories are very similar in terms of percent of species being protected. To determine whether or not each set of groups was significantly different, a chi squared test was performed. The p value for the test comparing mammals and birds was 0.4459 - indicating no significant difference between the groups. The p value for the test comparing mammals and reptiles was 0.0234 - indicating that the two groups are significantly different.

Recommendation:

The results above indicate that the categories of species most in need of intervention are mammals and birds. Conservationists concerned about endangered species should, therefore, focus their attention most on these two categories of species.

Sample Size for Foot and Mouth Disease Study

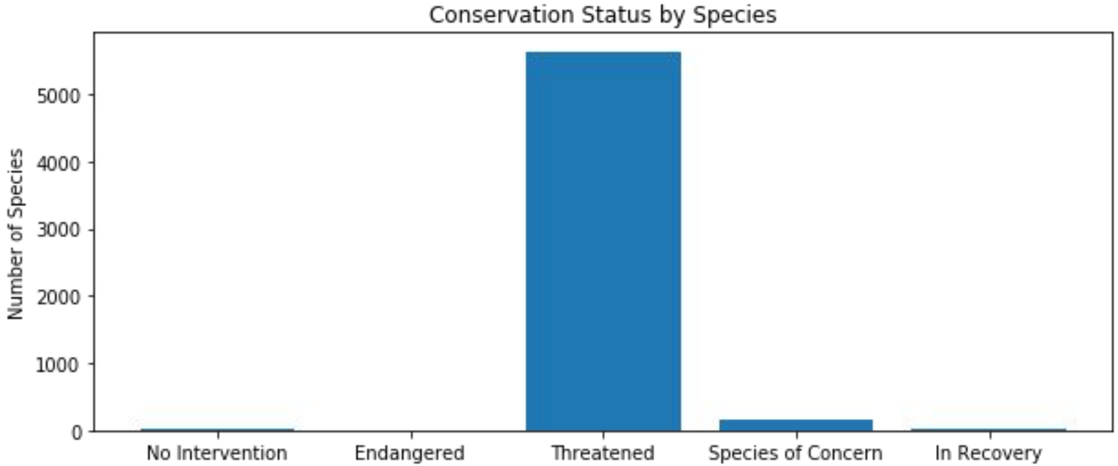
The following table shows the number of sheep observed by scientists at several national parks over the course of a week:

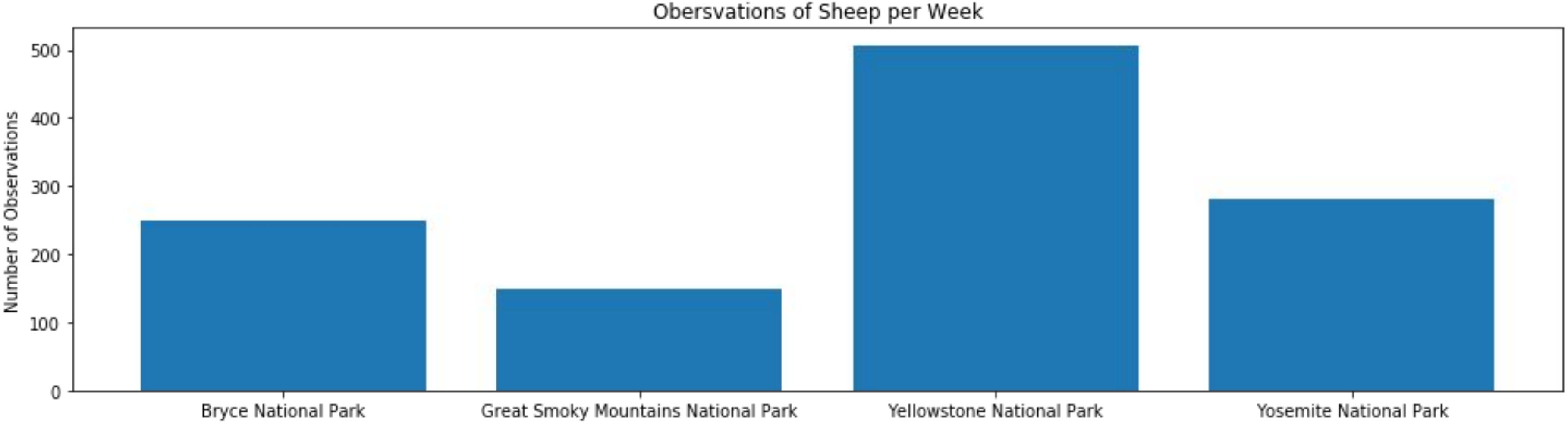
park_name	observations
Bryce National Park	250
Great Smoky Mountains National Park	149
Yellowstone National Park	507
Yosemite National Park	282

In an attempt to determine the success of a program to reduce foot and mouth disease in sheep at Yellowstone National Park, these scientist intend to conduct a study of the sheep at that park. The baseline conversion rate of this test was known to be 15%, based on the rate of the disease at Bryce National Park. In order to detect reductions in the rate of this disease of at least 5 percentage points with a 90% level of significance, the scientists must conduct a study with a minimal sample size of 39000. To achieve this with their current rate of observation at Yellowstone it will take them approximately 77 weeks.

This minimal sample size was determined using the sample size calculator at Optimizely using 15% as the baseline conversion rate, 5% as the minimum detectable effect and a 90% statistical significance.

Graphs


Conservation Status by Species


Obersvations of Sheep per Week