Biodiversity for the National Parks

Codecademy Capstone Option #2

Joseph Reed
February 2018

Objective:

Given two data sets perform some data analysis on the conservation statuses of species and investigate if there are any patterns or themes to the type of species that become endangered.

Joseph Reed
February 2018

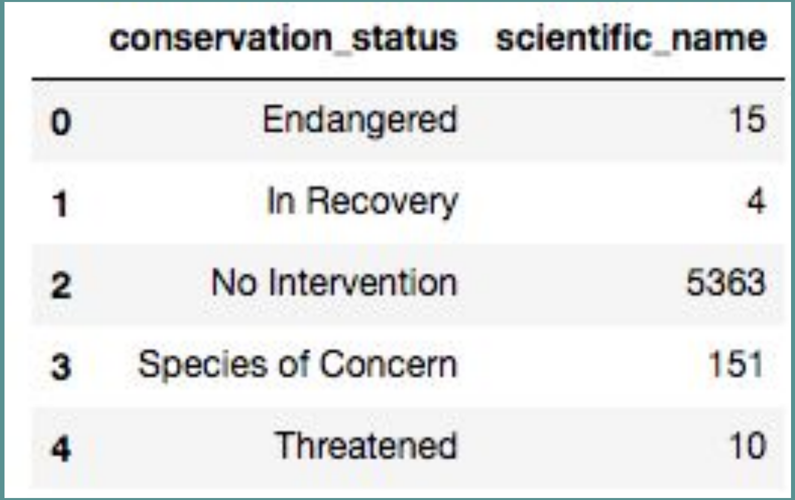
Data Description:

We were provided with a data set that had information on different species in our National Parks. Below are some of the observations made regarding this data:

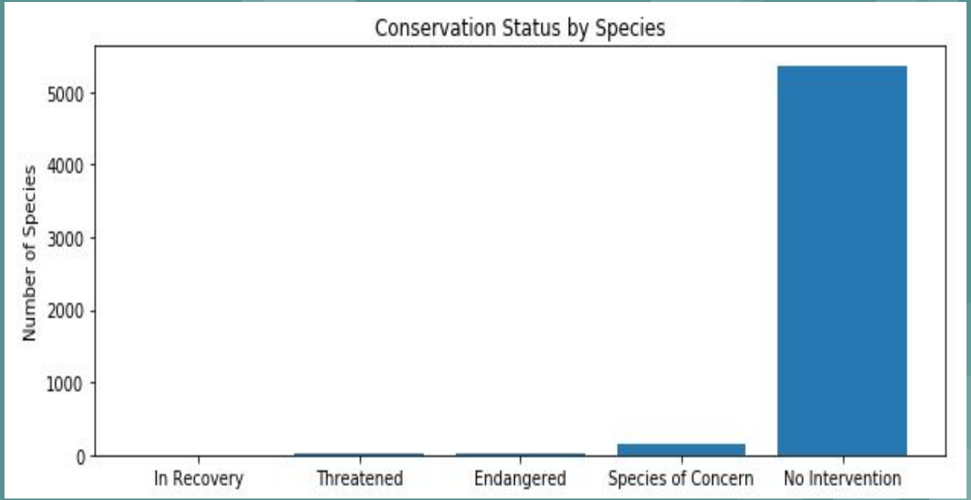
- The overall shape of the data was 5,824 rows with 4 columns.
- We had information on scientific name for each species, common name for each species and species conservation status.
- We had 5,541 unique scientific names in the data set.
- We had the following categories: Mammal, Bird, Reptile, Amphibian, Fish, Vascular Plant & Nonvascular Plant.
- We had the following conservation statuses: Endangered, In Recovery, Species of Concern & Threatened. We also derived a status of No Intervention for rows with a None value in the column.

Data Description:

As you can see a small percent (3.25%) of the species fall into a conservation status requiring intervention with Species of Concern being the largest group. The majority of the species do not require any intervention.



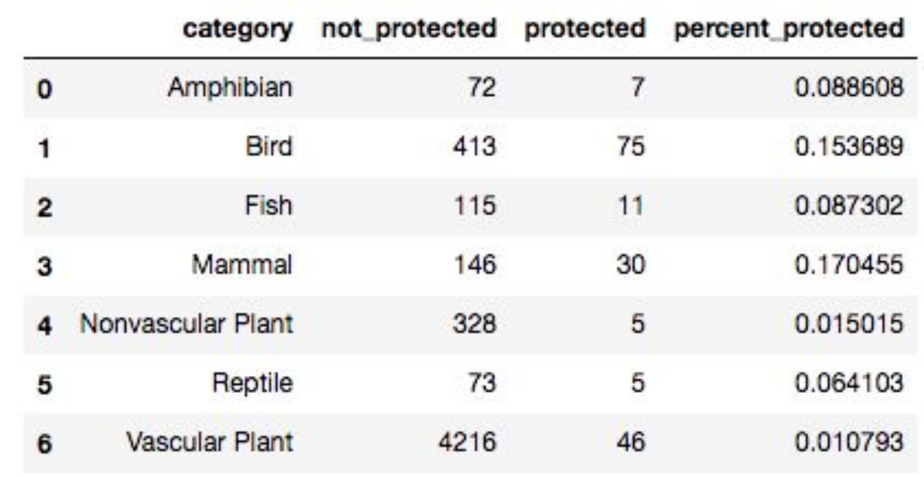
	conservation_status	scientific_name
0	Endangered	15
1	In Recovery	4
2	No Intervention	5363
3	Species of Concern	151
4	Threatened	10



Joseph Reed
February 2018

Data Description:

Next we did some data wrangling to see each category and count how many species in each were in protected status vs non protected status.



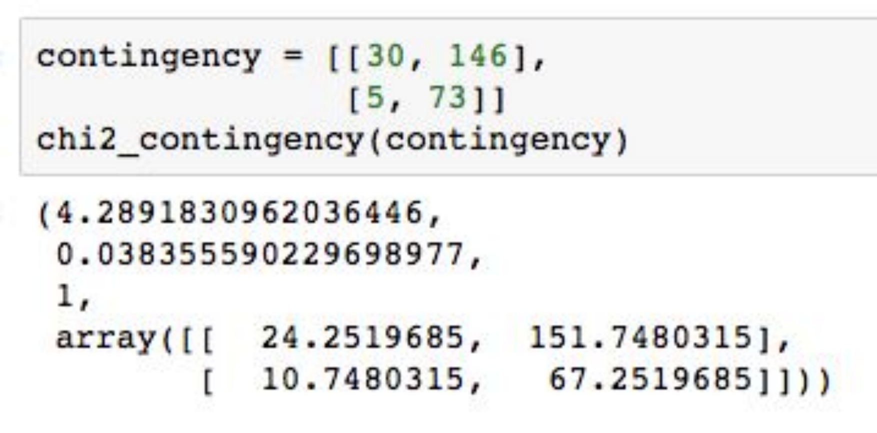
	category	not_protected	protected	percent_protected
0	Amphibian	72	7	0.088608
1	Bird	413	75	0.153689
2	Fish	115	11	0.087302
3	Mammal	146	30	0.170455
4	Nonvascular Plant	328	5	0.015015
5	Reptile	73	5	0.064103
6	Vascular Plant	4216	46	0.010793

Based on the above numbers Mammals are the most likely to be endangered.

Joseph Reed
February 2018

Significance Calculations:

We performed a chi squared test between Mammals and Reptiles to determine that there was a significant difference between the two.



```
contingency = [[30, 146],
               [5, 73]]
chi2_contingency(contingency)
```

```
(4.2891830962036446,
 0.038355590229698977,
 1,
 array([[  24.2519685,  151.7480315],
        [  10.7480315,   67.2519685]]))
```

Recommendation for conservations concerns:

- The categories with the highest number of species that require some level of intervention are Mammal and Bird.

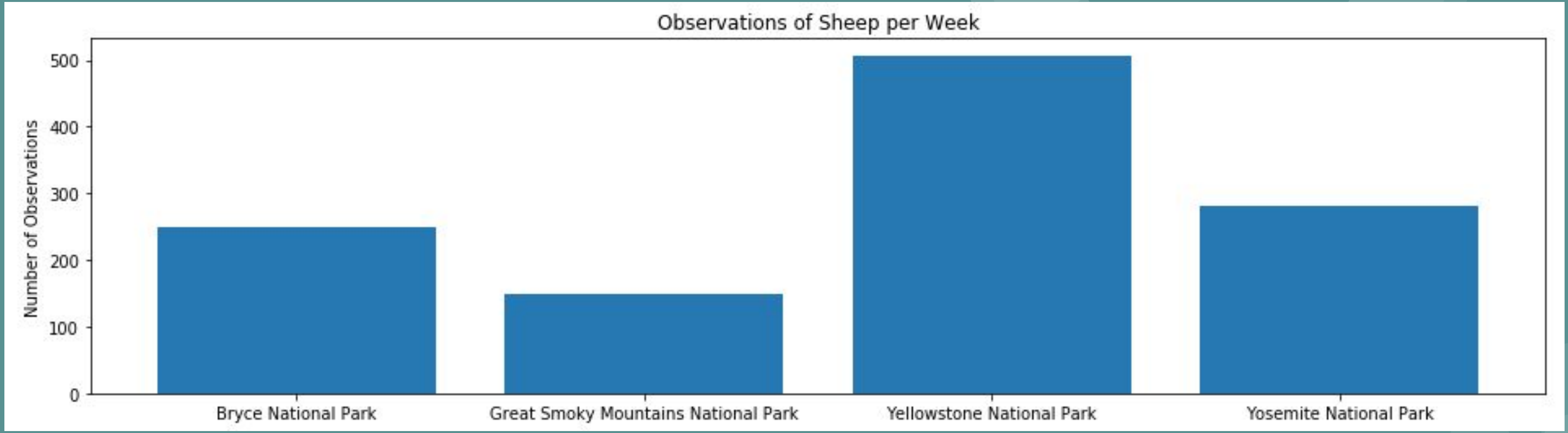
- Amphibians, Fish and Reptiles make up the categories with the middle number of species that require some level of intervention.

- Nonvascular and Vascular plant species have the lowest number of species requiring some level of intervention.

Joseph Reed
February 2018

Sample Size Determination:

Conservationists have been recording sightings of different species at several national parks for the past 7 days.



Joseph Reed
February 2018

Sample Size Determination:

Using the number of sightings we want to determine the sample size needed to verify if a program at Yellowstone National Park to reduce the rate of foot and mouth disease at the park is working. We know that 15% of the sheep at Bryce National Park have foot and mouth disease. We want to be able to detect a reduction of at least 5% with confidence at Yellowstone National Park.

First we determine the minimum detectable effect using the following formula:

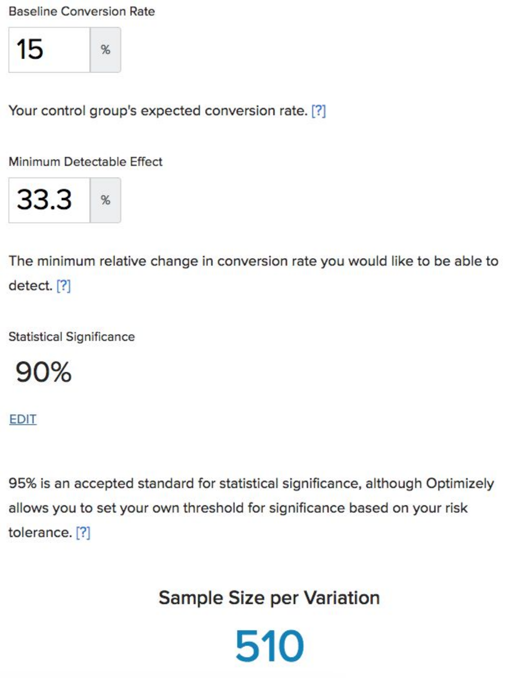
100 * .05 / .15 which gives us 33.33%

With this number and the above information we can now make use of a sample size calculator to determine the sample size needed.

Joseph Reed
February 2018

Sample Size Determination:

Next we utilize a sample size calculate and plug in the following numbers:



Baseline Conversion Rate

15 %

Your control group's expected conversion rate. [?]

Minimum Detectable Effect

33.3 %

The minimum relative change in conversion rate you would like to be able to detect. [?]

Statistical Significance

90%

EDIT

95% is an accepted standard for statistical significance, although Optimizely allows you to set your own threshold for significance based on your risk tolerance. [?]

Sample Size per Variation

510

Joseph Reed
February 2018

Sample Size Determination:

And as you can see we get a sample size of 510. Using this number we can now determine how long we will need to conduct the study.

For Bryce National Park it took 7 days to get 250 sheep sightings so it will take a little over 2 weeks to get our 510 observations.

For Yellowstone National Park it took 7 days to get 507 sheep sightings so it will take 1 week to get our 510 observations.

Joseph Reed
February 2018